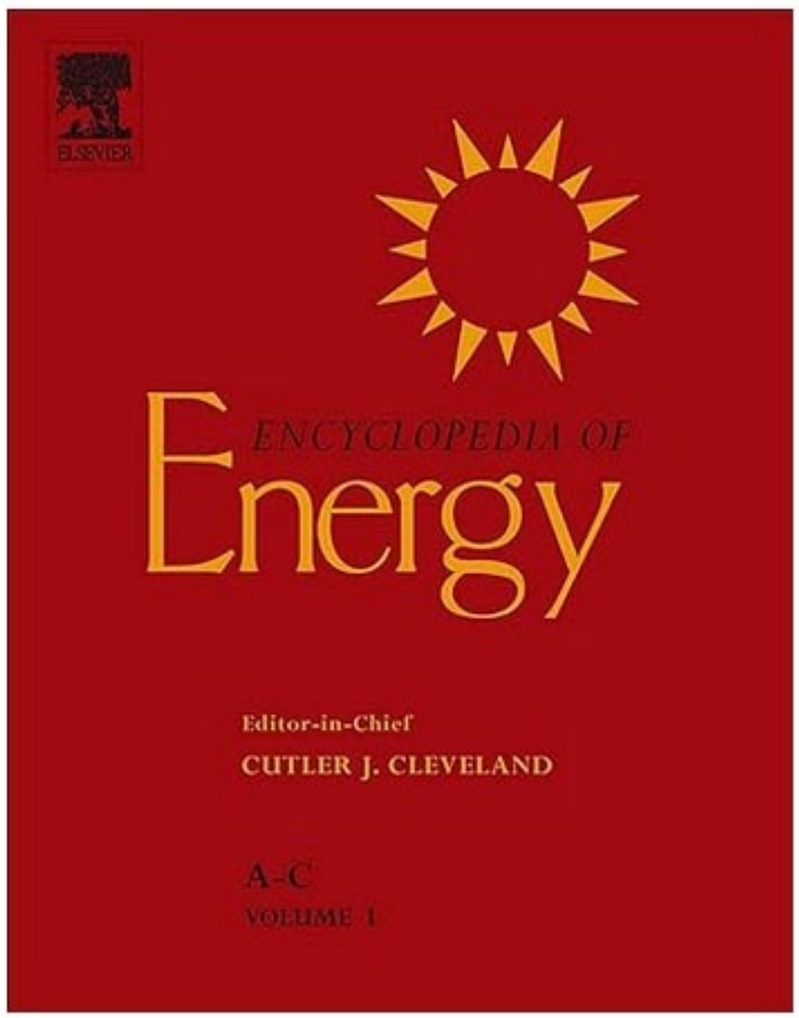
amazon.ca
Hello
Select your address
Books
Search Amazon.ca
All Best Sellers Prime Deals Store New Releases Gift Ideas Home Electronics Books Fashion Health & Household
Books Advanced Search Today's Deals New Releases Amazon Charts Best Sellers & More The Globe & Mail Best Sellers New York Times Best
Multi Test Leads Kit, 21 in-1 Electrical Multimeter Test Lead with Alligator Clips, Test Probe, Spring Grab Voltage Circuit Tester
Books › Professional & Technical › Professional Science



Roll over image to zoom in

Encyclopedia of Energy Hardcover – April 1 2004

by Cutler J. Cleveland (Editor)

4.0 ★★★★☆ 1 rating

Hardcover
from $2,001.81

1 Used from $2,001.81
1 New from $4,733.61

In recent years our usage and understanding of different types of energy has grown at a trem
Cutler Cleveland, and his international team of associate editors have brought together appro
Encyclopedia of Energy. This highly topical reference draws together all aspects of energy, co
the natural, social and engineering sciences.

The Encyclopedia will provide easily accessible information about all aspects of energy, writte

Read more

ISBN-10	ISBN-13	Edition	Publisher
012176480X	978-0121764807	1st	Elsevier Science

Follow the Author

PREFACE

The history of human culture can be viewed as the progressive development of new energy sources and their associated conversion technologies. Advances in our understanding of energy have produced unparalleled transformations of society, as exemplified by James Watt's steam engine and the discovery of oil. These transformations increased the ability of humans to exploit both additional energy and other resources, and hence to increase the comfort, longevity, and affluence of humans, as well as their numbers. Energy is related to human development in three important ways: as a motor of economic growth, as a principal source of environmental stress, and as a prerequisite for meeting basic human needs. Significant changes in each of these aspects of human existence are associated with changes in energy sources, beginning with the discovery of fire, the advent of agriculture and animal husbandry, and, ultimately, the development of hydrocarbon and nuclear fuels. The eventual economic depletion of fossil fuels will drive another major energy transition; geopolitical forces and environmental imperatives such as climate change may drive this transition faster than hydrocarbon depletion would have by itself. There is a diverse palette of alternatives to meet our energy needs, including a new generation of nuclear power, unconventional sources of hydrocarbons, myriad solar technologies, hydrogen, and more efficient energy end use. Each alternative has a different combination of economic, political, technological, social, and environmental attributes.

Energy is the common link between the living and non-living realms of the universe, and thus provides an organizing intellectual theme for diverse disciplines. Formalization of the concept of energy and identification of the laws governing its use by 19th century physical scientists such as Mayer and Carnot are cornerstones of modern science and engineering. The study of energy has played a pivotal role in understanding the creation of the universe, the origin of life, the evolution of human civilization and culture, economic growth and the rise of living standards, war and geopolitics, and significant environmental change at local, regional, and global scales.

The unique importance of energy among natural resources makes information about all aspects of its attributes, formation, distribution, extraction, and use an extremely valuable commodity. The *Encyclopedia of Energy* is designed to deliver this information in a clear and comprehensive fashion. It uses an integrated approach that emphasizes not only the importance of the concept in individual disciplines such as physics and sociology, but also how energy is used to bridge seemingly disparate fields, such as ecology and economics. As such, this *Encyclopedia* provides the first comprehensive, organized body of knowledge for what is certain to continue as a major area of scientific study in the 21st century. It is designed to appeal to a wide audience including undergraduate and graduate students, teachers, academics, and research scientists who study energy, as well as business corporations, professional firms, government agencies, foundations, and other groups whose activities relate to energy.

Comprehensive and interdisciplinary are two words I use to describe the *Encyclopedia*. It has the comprehensive coverage one would expect: forms of energy, thermodynamics, electricity generation, climate change, energy storage, energy sources, the demand for energy, and so on. What makes this work unique, however, is its breadth of coverage, including insights from history, society, anthropology, public policy, international relations, human and ecosystem health, economics, technology, physics, geology, ecology, business management, environmental

science, and engineering. The coverage and integration of the social sciences is a unique feature.

The interdisciplinary approach is employed in the treatment of important subjects. In the case of oil, as one example, there are entries on the history of oil, the history of OPEC, the history of oil prices, oil price volatility, the formation of oil and gas, the distribution of oil resources, oil exploration and drilling, offshore oil, occupational hazards in the oil industry, oil refining, energy policy in the oil industry, the geopolitics of oil, oil spills, oil transportation, public lands and oil development, social impacts of oil and gas development, gasoline additives and public health, and the environmental impact of the Persian Gulf War. Other subjects are treated in a similar way.

This has been a massive and extremely satisfying effort. As with any work of this scale, many people have contributed at every step of the process, including the staff of Academic Press/Elsevier. The project began through the encouragement of Frank Cynar and David Packer, with Frank helping to successfully launch the initiative. Henri van Dorssen skillfully guided the project through its completion. He was especially helpful with integrating the project formulation, production, and marketing aspects of the project. Chris Morris was with the project throughout, and displayed what I can only describe as an uncanny combination of vision, enthusiasm, and energy for the project. I owe Chris a great deal for his insight and professionalism. I spent countless hours on the phone with Robert Matsumura, who was the glue that held the project together. Chris and Robert were ably assisted by outstanding Academic Press/Elsevier staff, especially Nick Panissidi, Joanna Dinsmore, and Mike Early. Clare Marl and her team put together a highly effective and creative marketing plan.

At the next stage, the Editorial Board was invaluable in shaping the coverage and identifying authors. The Board is an outstanding collection of scholars from the natural, social, and engineering sciences who are recognized leaders in their fields of research. They helped assemble an equally impressive group of authors from every discipline and who represent universities, government agencies, national laboratories, consulting firms, think tanks, corporations, and nongovernmental organizations. I am especially proud of the international scope of the authors: more than 400 authors from 40 nations are represented from every continent and every stage of development. To all of these, I extend my thanks and congratulations.

Cutler Cleveland
Boston University
Boston, Massachusetts, United States

FOREWORD

Energy generation and use are strongly linked to all elements of sustainable development: economic, social, and environmental. The history of human development rests on the availability and use of energy, the transformation from the early use of fire and animal power that improved lives, to the present world with use of electricity and clean fuels for a multitude of purposes. This progress built on basic scientific discoveries, such as electromagnetism and the inventions of technologies such as steam engines, light bulbs, and automobiles.

It is thus abundantly clear that access to affordable energy is fundamental to human activities, development, and economic growth. Without access to electricity and clean fuels, people's opportunities are significantly constrained. However, it is really energy services, not energy *per se* that matters. Yet, today some 2 billion people lack access to modern energy carriers.

In addition to the great benefits, the generation, transportation, and use of energy carriers unfortunately come with undesired effects. The environmental impacts are multifaceted and serious, although mostly less evident. Emissions of suspended fine particles and precursors of acid deposition contribute to local and regional air pollution and ecosystem degradation. Human health is threatened by high levels of air pollution resulting from particular types of energy use at the household, community, and regional levels.

Emissions of anthropogenic greenhouse gases (GHG), mostly from the production and use of energy, are altering the atmosphere in ways that are affecting the climate. There is new and stronger evidence that most of the global warming observed over the last 50 years is attributable to human activities. Stabilization of GHG in the atmosphere will require a major reduction in the projected carbon emissions to levels below the present.

Dependence on imported fuels leaves many countries vulnerable to disruption in supply, which might pose physical hardships and economic burdens; the weight of fossil fuel imports on the balance of payments is unbearable for many poorer countries. The present energy system of countries heavily dependent on fossil fuels geographically concentrated in a few regions of the world adds security of supply aspects.

From the issues indicated here it is clear that major changes are required in energy system development worldwide. At a first glance, there appears to be many conflicting objectives. For example, is it possible to sustain poverty alleviation and economic growth while reducing GHG emissions? Can urban areas and transport expand while improving air quality? What would be the preferable trade-offs? Finding ways to expand energy services while simultaneously addressing the environmental impacts associated with energy use represents a critical challenge to humanity.

What are the options? Looking at physical resources, one finds they are abundant. Fossil fuels will be able to provide the energy carriers that the world is used to for hundreds of years. Renewable energy flows on Earth are many thousands of times larger than flows through energy markets. Therefore, there are no apparent constraints from a resource point of view. However, the challenge is how to use these resources in an environmentally acceptable way. The broad categories of options for using energy in ways that support sustainable development are (1) more efficient use of energy in all sectors, especially at the point of end use, (2) increased use of renewable energy sources, and (3) accelerated development and deployment of new and advanced energy technologies,

including next-generation fossil fuel technologies that produce near-zero harmful emissions. Technologies are available in these areas to meet the challenges of sustainable development. In addition, innovation provides increasing opportunities.

Analysis using energy scenarios indicates that it is indeed possible to simultaneously address the sustainable development objectives using the available natural resources and technical options presented. A prerequisite for achieving energy futures compatible with sustainable development objectives is finding ways to accelerate progress for new technologies along the energy innovation chain, including research and development, demonstration, deployment, and diffusion.

It is significant that there already exist combinations of technologies that meet all sustainable development challenges at the same time. This will make it easier to act locally to address pollution problems of a major city or country while at the same time mitigating climate change. Policies for energy for sustainable development can be largely motivated by national concerns and will not have to rely only on global pressures.

However, with present policies and conditions in the marketplaces that determine energy generation and use such desired energy futures will not happen. A prerequisite for sustainable development is change in policies affecting energy for sustainable development. This brings a need to focus on the policy situation and understand incentives and disincentives related to options for options for energy for sustainable development.

Policies and actions to promote energy for sustainable development would include the following:

- Developing capacity among all stakeholders in all countries, especially in the public sector, to address issues related to energy for sustainable development.
- Adopting policies and mechanisms to increase access to energy services through modern fuels and electricity for the 2 billion without.
- Advancing innovation, with balanced emphasis on all steps of the innovation chain: research and development, demonstrations, cost buy-down, and wide dissemination.
- Setting appropriate market framework conditions (including continued market reform, consistent regulatory measures, and targeted policies) to encourage competitiveness in energy markets, to reduce total cost of energy services to end-users, and to protect important public benefits, including the following:
  - Cost-based prices, including phasing out all forms of permanent subsidies for conventional energy (now on the order of $250 billion a year) and internalizing external environmental and health costs and benefits (now sometimes larger than the private costs).
  - Removing obstacles and providing incentives, as needed, to encourage greater energy efficiency and the development and/or diffusion of new technologies for energy for sustainable development to wider markets.
  - Recent power failures on the North American Eastern Seaboard, in California, London (United Kingdom), Sweden, and Italy illustrate the strong dependence on reliable power networks. Power sector reform that recognizes the unique character of electricity, and avoids power crises as seen in recent years, is needed.
- Reversing the trend of declining Official Development Assistance and Foreign Direct Investments, especially as related to energy for sustainable development.

This is a long list of opportunities and challenges. To move sufficiently in the direction of sustainability will require actions by the public and the private sector, as well as other stakeholders, at the national, regional, and global levels. The decisive issues are not technology or natural resource scarcity, but the institutions, rules, financing mechanisms, and regulations needed to make markets work in support of energy for sustainable development. A number of countries, including Spain, Germany, and Brazil, as well as some states in the United States have adopted successful laws and regulations designed to increase the use of renewable energy sources. Some regions, including Latin America and the European Union, have set targets for increased use of renewable energy. However, much remains to be done.

Energy was indeed one of the most intensely debated issues at the United Nations World Summit on Sustainable Development (WSSD), held in Johannesburg, South Africa, in August/September, 2002. In the end, agreement was reached on a text that significantly advances the attention given to energy in the context of sustainable development. This was in fact the first time agreements could be reached on energy at the world level! These developments followed years of efforts to focus on energy as an instrument for sustainable development that

intensified after the United Nations Conference on Environment and Development in 1992.

The United Nations General Assembly adopted the Millennium Development Goals (MDG) in 2000. These goals are set in areas such as extreme poverty and hunger, universal primary education, gender equality and empowerment of women, child mortality, maternal health, HIV/AIDS, malaria and other diseases, and environmental sustainability. However, more than 2 billion people cannot access affordable energy services, based on the efficient use of gaseous and liquid fuels and electricity. This constrains their opportunities for economic development and improved living standards. Women and children suffer disproportionately because of their relative dependence on traditional fuels. Although no explicit goal on energy was adopted, access to energy services is a prerequisite to achieving all of the MDGs.

Some governments and corporations have already demonstrated that policies and measures to promote energy solutions conducive to sustainable development can work, and indeed work very well. The renewed focus and broad agreements on energy in the Johannesburg Plan of Implementation and at the 18th World Energy Congress are promising. The formation of many partnerships on energy between stakeholders at WSSD is another encouraging sign. A sustainable future in which energy plays a major positive role in supporting human well-being is possible!

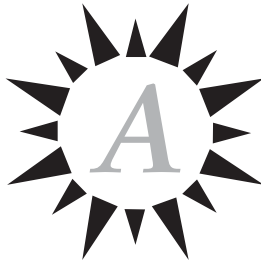
Progress is being made on many fronts in bringing new technologies to the market, and to widening access to modern forms of energy. In relation to energy, a total of 39 partnerships were presented to the United Nations Secretariat for WSSD to promote programs on energy for sustainable development, 23 with energy as a central focus and 16 with a considerable impact on energy. These partnerships included most prominently the DESA-led Clean Fuels and Transport Initiative, the UNDP/World Bank-led Global Village Energy Partnership (GVEP), the Johannesburg Renewable Energy Coalition (JREC), the EU Partnership on Energy for Poverty Eradication and Sustainable Development, and the UNEP-led Global Network on Energy for Sustainable Development (GNESD).

With secure access to affordable and clean energy being so fundamental to sustainable development, the publication of the *Encyclopedia of Energy* is extremely timely and significant. Academics, professionals, scholars, politicians, students, and many more will benefit tremendously from the easy access to knowledge, experience, and insights that are provided here.

Thomas B. Johansson
Professor and Director
International Institute for
Industrial Environmental Economics
Lund University
Lund, Sweden

Former Director
Energy and Atmosphere Programme
United Nations Development Programme
New York, United States

Acid Deposition and Energy Use



JAN WILLEM ERISMAN
Energy Research Centre of the Netherlands
Petten, The Netherlands

1. **Introduction**
2. **Acid Deposition, Its Effects, and Critical Loads**
3. **Processes in the Causal Chain: Emission, Transport, and Deposition**
4. **Emissions from Energy Use**
5. **Abatement and Trends in Emission**
6. **Acid Deposition**
7. **Benefits and Recovery of Ecosystems**
8. **Future Abatement**
9. **Conclusion**

Glossary

acid deposition The removal of acidic or acidifying components from the atmosphere by precipitation (rain, cloud droplets, fog, snow, or hail); also known as acid rain or acid precipitation.

acidification The generation of more hydrogen ions (H^+) than hydroxide ions (OH^-) so that the pH becomes less than 7.

critical level The maximum pollutant concentration a part of the environment can be exposed to without significant harmful effects.

critical load The maximum amount of pollutant deposition a part of the environment can tolerate without significant harmful effects.

deposition Can be either wet or dry. In dry deposition, material is removed from the atmosphere by contact with a surface. In wet deposition, material is removed from the atmosphere by precipitation.

emissions The release of primary pollutants directly to the atmosphere by processes such as combustion and also by natural processes.

eutrophication An increase of the amount of nutrients in waters or soils.

nitrification The conversion of ammonium ions (NH_4^+) to nitrate (NO_3^-).

nonlinearity The observed nonlinear relationship between reductions in primary emissions and in pollutant deposition.

pollutant Any substance in the wrong place at the wrong time is a pollutant. Atmospheric pollution may be defined as the presence of substances in the atmosphere, resulting from man-made activities or from natural processes, causing effects to man and the environment.

Acid deposition originates largely from man-made emissions of three gases: sulfur dioxide (SO_2), nitrogen oxides (NO_x), and ammonia (NH_3). It damages acid-sensitive freshwater systems, forests, soils, and natural ecosystems in large areas of Europe, the United States, and Asia. Effects include defoliation and reduced vitality of trees; declining fish stocks and decreasing diversity of other aquatic animals in acid-sensitive lakes, rivers, and streams; and changes in soil chemistry. Cultural heritage is also damaged, such as limestone and marble buildings, monuments, and stained-glass windows. Deposition of nitrogen compounds also causes eutrophication effects in terrestrial and marine ecosystems. The combination of acidification and eutrophication increases the acidification effects. Energy contributes approximately 82, 59, and 0.1% to the global emissions of SO_2, NO_x, and NH_3, respectively. Measures to reduce acid deposition have led to controls of emissions in the United States and Europe. Sulfur emissions were reduced 18% between 1990 and 1998 in the United States and 41% in Europe during the same period. At the same time, emissions in Asia increased 43%. In areas of the world in which emissions have decreased, the effects have decreased; most notably, lake acidity has decreased due to a decrease in sulfur emissions, resulting in lower sulfate and acid concentrations. However, systems with a long response time have not seen improvement yet and very limited recovery has also been observed. Emissions should therefore be further reduced and sustainable levels should be maintained to decrease the effects and to see recovery. This will require drastic changes in our

 © 2004 Elsevier Inc. All rights reserved.

energy consumption and the switch to sustainable energy sources. Before renewable energy sources can fulfill a large part of our energy needs, the use of zero-emission fossil fuels must be implemented. This should be done in such a way that different environmental impacts are addressed at the same time. The most cost-effective way is to reduce CO_2 emissions because this will result in decreases in SO_2 and NO_x emissions. When starting with SO_2 and NO_x emissions, an energy penalty compensates for the emission reductions.

1. INTRODUCTION

The alarm regarding the increasing acidification of precipitation in Europe and eastern North America was first sounded in the 1960s. Since then, most attention has focused on acid rain's effects, established and suspected, on lakes and streams, with their populations of aquatic life, and on forests, although the list of concerns is far broader: It includes contamination of groundwater, corrosion of man-made structures, and, recently, deterioration of coastal waters.

The processes that convert the gases into acid and wash them from the atmosphere began operating long before humans started to burn large quantities of fossil fuels. Sulfur and nitrogen compounds are also released by natural processes such as volcanism and the activity of soil bacteria. However, human economic activity has made the reactions vastly more important. They are triggered by sunlight and depend on the atmosphere's abundant supply of oxygen and water.

Wood was the first source for energy production. Wood was used for the preparation of food and for heating. After the discovery of fossil fuels in the form of coal as a source of energy, it was used for heating purposes. During the industrial revolution, coal, followed by crude oil and later natural gas, provided a large source of energy that stimulated technological development, resulting in the replacement of animals and people in the production of goods, increased possibilities for mobility, and the enabling of industries to process raw materials into a large variety of products. When the possibilities of energy production and use became apparent, the use of fossil fuels grew exponentially. Figure 1 provides an overview of the worldwide use of fossil fuels since 1900. It shows that the total global energy consumption is closely linked to the growth of the world's population. The share of oil and gas as energy

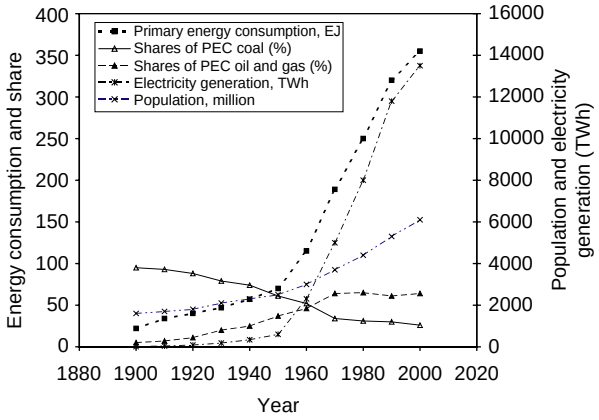


FIGURE 1 The worldwide use of fossil fuels in millions of tons per year. Adapted from Smil (2000).

sources has increased and has become higher than the share of coal.

The first environmental consequences of the use of fossil fuels were the problems of smoke in houses and urban air quality. Human health was affected by the use of coal with high sulfur content, producing sulfur dioxide and carbon monoxide. The history of London clearly shows the ever-increasing problems related to coal use and the struggle to abate this pollution, culminating in the famous London smog in 1952, from which thousands of people died. In the 1950s and 1960s, air pollution became an important issue but was mainly considered on the local scale. One of the abatement options was to increase the stack height in order to take advantage of atmospheric dispersion, leading to dilution of pollutant concentrations before humans could be exposed. This formed the basis of the transboundary nature of acid deposition as we know it today and increased the scale from local to regional and even continental. The use of fossil fuels increased, and the pollutants were transported to remote areas, which had very clean air until that time.

Acid rain was first recognized as a serious environmental problem by Robert Angus Smith, an alkali inspector based in Manchester. He started a monitoring network with approximately 12 sites at which rainfall was collected and analyzed. In 1852, he wrote about the correlation between coal burning and acid pollution in and around the industrial center of Manchester. He identified three types of air: "that with carbonate of ammonia in the fields at a distance … that with sulphate of ammonia in the suburbs, … and that with sulphuric acid, or acid sulphate, in the town." In 1872, he wrote the famous

book, *Air and Rain: The Beginnings of a Chemical Climatology*. He referred to the acidity of rain and air and its effect on humans and buildings. This was the beginning of acidification research, although this was recognized only much later.

Acid deposition is one of the oldest transboundary environmental problems. Several components and thus emission sources contribute to acid deposition, making it a complex problem and difficult to abate. It is one of the most studied environmental problems and there is a large body of literature. This article gives an overview of the state of knowledge of energy production and use and acid deposition. First, an overview is given of acid deposition, the consequences of acid deposition, and the processes involved. The contribution of energy is discussed next, and an estimate of the progress made to decrease the emissions from energy sources is given. Finally, future options to decrease acid deposition and the options for decreasing energy emissions are discussed.

2. ACID DEPOSITION, ITS EFFECTS, AND CRITICAL LOADS

Acid deposition is the deposition of air pollutants to ecosystems leading to acidification and eutrophication, which cause negative effects in these ecosystems. The components contributing to acid deposition are sulfur compounds (SO_2 and SO_4), oxidized nitrogen compounds (NO, NO_2, HNO_2, HNO_3, and NO_3), and reduced nitrogen compounds (NH_3 and NH_4). Some rain is naturally acidic because of the carbon dioxide (CO_2) in air that dissolves with rain water and forms a weak acid. This kind of acid is actually beneficial because it helps dissolve minerals in the soil that both plants and animals need. Ammonia acts as a base in the atmosphere, neutralizing nitric and sulfuric acid, but in soil ammonia it can be converted by microorganisms to nitric acid, producing additional acid in the process.

The first sign of the effects of acid deposition was the loss of fish populations in Canadian, Scandinavian, and U.S. lakes in the early 1960s. It was found that the water in the lakes was acidified to a point at which fish eggs no longer produced young specimens. This was caused by acid deposition; precipitation had introduced so much acid in the lakes that pH levels had declined below the critical limit for egg hatching. A few years later, reports from Canada, the United States, and Germany indicated unusual damage in forests. Trees showed decreased foliage or needles, and this damage even progressed to the point that trees would die, a phenomenon that happened in the border area of former East Germany, Poland, and former Czechoslovakia. This tree dieback was also attributed to acidification of the soil. However, exposure to increased oxidant concentrations was determined to be another cause. In the mid-1980s, nitrogen deposition to terrestrial and aquatic ecosystems was found to cause negative effects through eutrophication. Acidification and eutrophication are both caused by atmospheric deposition of pollutants, and the combination increases the effects, causing too high nutrient concentrations in soil and groundwater. Nitrate and ammonium are beneficial, even essential, for vegetation growth, but in too high concentration they lead to the loss of diversity, especially in oligotrophic ecosystems (i.e., ecosystems adapted to low nutrient availability). This problem has mainly been encountered in central and northwestern Europe.

The effects of acid deposition include changes in terrestrial and aquatic ecosystems showing acidification of soils, shifts in plant community composition, loss of species diversity, forest damage, water acidification and loss of fish populations, reduction in growth yield of sensitive agricultural crops, and damage to cultural heritage and to building materials. The common factor for all these effects is that pollutants, or their precursors, are emitted, transported in the atmosphere, and deposited by way of precipitation (wet deposition) or as gases or particles (dry deposition). In some cases, acid deposition occurs 1000 km or more from where the responsible emissions were generated.

If deposition exceeds sustainable levels, effects can occur depending on the duration of the exposure and the level of exceedance. Sustainable levels can be represented by critical loads and critical levels. The critical load is formally defined by the United Nations Economic Commission for Europe (UNECE) as "a quantitative estimate of exposure to one or more pollutants below which significant harmful effects on sensitive elements of the environment do not occur according to present knowledge." This definition can be applied to a wide range of phenomena and not merely acid rain. Any process in which damage will occur if depositions exceed natural sink processes has a critical load set by those sink processes. The concept of critical loads or levels has been used to provide a scientific basis for the incorporation of environmental impacts in the development of national and international policies to

control transboundary air pollutants, within the framework of the UNECE Convention on the Control of Transboundary Air Pollution. The concept is based on the precautionary principle: Critical loads and levels are set to prevent any long-term damage to the most sensitive known elements of any ecosystem. The critical loads approach is an example of an effects-based approach to pollution control, which considers that the impact of deposited material depends on where it lands. Similar industrial plants may have different pollution impacts depending on the different capacities of the receiving environments to absorb, buffer, or tolerate pollutant loads. The policy goal with effects-based strategies is ultimately to ensure that the abatement on emissions from all plants is adequate so that no absorptive capacities, nor air quality standards, are exceeded.

3. PROCESSES IN THE CAUSAL CHAIN: EMISSION, TRANSPORT, AND DEPOSITION

The causal chain represents the sequence of emissions into the atmosphere, transport and chemical reactions, deposition to the earth surface, and cycling within the ecosystems until eventually the components are fixed or transformed in a form having no effects. This causal chain is shown in Fig. 2. After the causal chain is understood and parameterized in models, it can be used to assess the possibilities for reducing the effects.

3.1 Emission

The use of fossil fuels for energy production is one of the main causes of acid deposition. Fossil fuels contain sulfur compounds, which are emitted into the atmosphere as SO_2 after combustion. Oxidized nitrogen has several sources but is formed mainly by combustion processes in which fuel N is oxidized or atmospheric N_2 is oxidized at high temperatures. These processes are relevant for industries, traffic, and energy production. Among the other sources of oxidized nitrogen, soils are most important. Ammonia is mainly emitted from agricultural sources. Fossil fuels seem less important for ammonia than for the other two pollutants. Through the use of fertilizer, which is produced from natural gas, there is a link with energy. Indirectly, the increased energy use from fossil fuels to produce fertilizer has increased nitrogen emissions from agriculture and contributed to acidification. Energy use first increased mobility. Food is therefore not necessarily produced in the area where it is consumed. Furthermore, relatively cheap nutrient resources (animal feed) are transported from one continent to another, where they are concentrated in intensive livestock breeding areas. Also, fossil fuels, first in the form of coal but now mainly from natural gas, provide the basis to produce fertilizers, which are used to produce the food necessary to feed the ever-growing population throughout the world and to fulfill the need for the growing demands for more luxurious food.

Base cations play an important role because they can neutralize acids in the atmosphere and the soil, limiting the effect of acid deposition. On the other hand, base cations contribute to the particle concentrations in the atmosphere, affecting human health when inhaled. Sources of base cations are energy production, agricultural practices, and road dust.

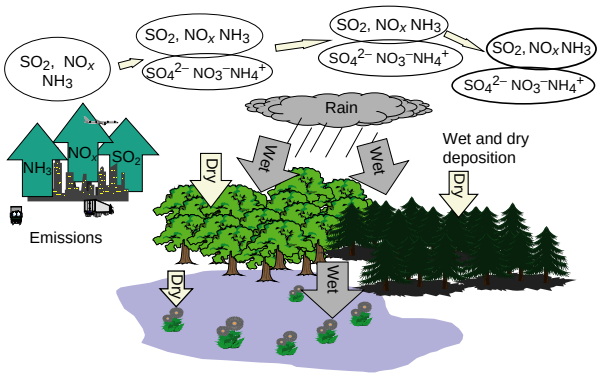


FIGURE 2 Emission, transport, and deposition of acidifying pollutants.

3.2 Transport and Deposition

The sulfur compounds are converted to sulfuric acid and NO_x to nitric acid, and ammonia reacts with these acids to form ammonium salts. Both precursors (SO_2, NO_x, and NH_3) and products (sulfuric acid, nitric acid, and ammonium salts) can be transported over long distances (up to 1500 km), depending on meteorological conditions, speed of conversion, and removal by deposition processes. The gases emitted, SO_2, NO_x, and NH_3, behave differently in the atmosphere. The residence time of NH_3 is relatively short because it is emitted at low level (near the ground), it converts quickly to NH_4^+, and the dry deposition rate is fairly high. NH_3 is transported 100–500 km. The horizontal and vertical concentration gradients are steep, so concentrations and depositions can vary significantly over small distances.

This means that a large proportion of the NH_3 emitted in The Netherlands is also deposited within this country. Once converted into NH_4^+, which has a much lower rate of deposition, the transport distances are much greater (up to 1500 km). SO_2 is mainly emitted into the atmosphere by high sources and can therefore be transported over long distances, despite its relatively high deposition rate. Some NO_x is also emitted by low sources (traffic). However, because of its low deposition rate and relatively slow conversion rate into rapidly deposited gases (HNO_3 and HNO_2), NO_x is transported over relatively long distances before it disappears from the atmosphere. SO_2 is quickly converted to sulfuric acid (H_2SO_4) after deposition, both in water and in soil. NO_x and NH_3 and their subsequent products contribute to the eutrophication and also the acidification of the environment as a result of conversion to nitric acid (HNO_3) in the air (NO_x) or in the soil (NH_3).

Since the three primary gases (SO_2, NO_x, and NH_3) can react and be in equilibrium with each other and with the different reaction products in the atmosphere, there is a strong and complex mutual relationship. If, for example, there were no NH_3 in the atmosphere, SO_2 would be converted less quickly to SO_4^{2-}. The environment, however, would also be "more acid" so that the deposition rate of acidifying compounds would be reduced (poor solubility of these compounds in acid waters). The net impact is difficult to determine, but it is certainly true that if the emission of one of the compounds increases or decreases relative to that of the others, this will also influence the transport distances and deposition rates of the other compounds. This is only partly taken into account in the scenario calculations because such links have not been fully incorporated into the models.

The base cations can neutralize the acidifying deposition and, after deposition to the soil, they act as a buffer both in terms of neutralization and in terms of uptake by plants and the prevention of nutrient deficiencies. This applies especially to the Mg^{2+}, Ca^{2+}, and K^+ compounds. The degree of deposition of base cations is therefore important in determining the critical loads that an ecosystem can bear or the exceedances thereof. Thus, accurate loads are also required for base cations on a local scale. These estimates are not available for The Netherlands nor for Europe.

The nature and size of the load of acidifying substances on the surface depend on the characteristics of the emission sources (height and point or diffuse source), the distance from the source, physical and chemical processes in the atmosphere, and the receptor type (land use, roughness, moisture status, degree of stomatal opening of vegetation, snow cover, etc.). When gases and/or particles are deposited or absorbed directly from the air, we speak of dry deposition. When they reach the surface dissolved in rain or another form of precipitation, we refer to wet deposition. If this occurs in mist or fog, it is cloud/fog deposition. The total deposition is the sum of the dry, wet, and cloud/fog deposition. Throughfall is washed from the vegetation by rain falling on the soil beneath the forest canopy. It is a measure of the load on the vegetation's surface, whereas the total deposition indicates the load on the whole system (soil + vegetation). Net throughfall is the difference between throughfall and wet deposition in the open field; it is a measure of the dry deposition plus cloud/fog deposition when no exchange takes place with the forest canopy (uptake or discharge of substances through leaves, microorganisms, etc.).

To understand the impact of emissions on the effects caused by acid deposition and eutrophication, the whole chain of emissions, transport, atmospheric conversion, dry and wet deposition, and the effects caused by the total deposition loads, including the role of groundwater and soils, must be understood. Wet and dry deposition are not only the processes by which pollutants are transported to the earth surface and humans and ecosystems are exposed to acid deposition and eutrophication but also play an essential role in cleaning of the atmosphere. If removal by wet and dry deposition stopped, the earth's atmosphere would be unsuitable to sustain life in a relatively short period—on the order of a few months.

Since acidifying deposition involves different substances, it is necessary to give these substances a single denominator in order to indicate the total load of acidifying substances. For this purpose, the total load is expressed as potential acid, calculated as follows:

$$2SO_x + NO_y + NH_x (\text{mol } H^+ \text{ha/year}),$$

where SO_x are oxidized sulfur compounds, NO_y are oxidized nitrogen compounds, and NH_x are reduced nitrogen compounds. The concept of potential acid is used because NH_3 is considered to be a potentially acidifying substance. In the atmosphere, NH_3 acts as a base, which leads to the neutralization of acids such as HNO_3 and H_2SO_4. However, the NH_4^+ formed in the soil can be converted to NO_3^- so that acid is still produced via bacterial conversion (nitrification)

according to

$$NH_4^+ + 2O_2 \rightarrow NO_3^- + H_2O + 2H^+.$$

Two moles of acid are finally formed via this process: one originating from the neutralized acid and one originating from NH_3. On balance, as for 1 mol NO_y, 1 mol NH_3 acts maximally to acidify 1 mol H^+ acid. One mole of the bivalent SO_4^{2-} can lead to the formation of 2 mol of H^+. The actual acidification depends on the degree to which NO_3^- and SO_4^{2-} leach out of the soil. Only when this occurs completely is the actual acidification equal to the potential acidification.

4. EMISSIONS FROM ENERGY USE

van Ardenne *et al*. provide a compilation of global sulfur emissions, mainly based on work by different authors. According to their estimates, anthropogenic contributions are on the order of 75% of the total sulfur emissions, 82% of which is related to energy production and use (Fig. 3). Anthropogenic emissions exceeded natural emissions as early as 1950. Important natural sources are volcanoes and DMS emissions from the oceans. Fossil fuel use in industry, fuel production in refineries, and electricity generation are the main activities responsible for anthropogenic emissions. According to EMEP (Cooperative Program for Monitoring and Evaluation of Air Pollutants in Europe) studies, the maximum sulfur emissions in western Europe occurred from 1975 to 1980, at a level of 40 million tons of SO_2. The emissions have decreased since then and are currently at a level of approximately 30 million tons of SO_2.

On a global scale, NO_x emissions are estimated to be equally divided between anthropogenic and biogenic sources. However, anthropogenic contributions are already much higher than the natural emissions in the Northern Hemisphere. Most natural emissions, through biomass burning, occur in agricultural areas of South America and Africa. In western Europe and the United States, traffic is the main source for NO_x, contributing 50% or more of the total NO_x emissions.

Globally, in 1995 the emissions from fossil fuels contributed 82, 59, and 0.1% to the total emissions of SO_2, NO_x, and NH_3, respectively. The contributions of fuel and other anthropogenic sources to the emissions of the three gases are shown in Fig. 4.

5. ABATEMENT AND TRENDS IN EMISSION

Figure 5 shows the change in global emissions of SO_2, NO_x, and NH_3 for the different source categories. The figure shows that there is a steady increase in emissions for the three gases between 1890 and 1995. The contribution of energy emissions is highest during the whole time period for SO_2 and NO_x; for NH_3, energy constitutes only a minor contribution. After World War II, the increase in emission is higher.

After 1990, there seems to be a smaller increase in global emissions. This is mainly due to the decrease

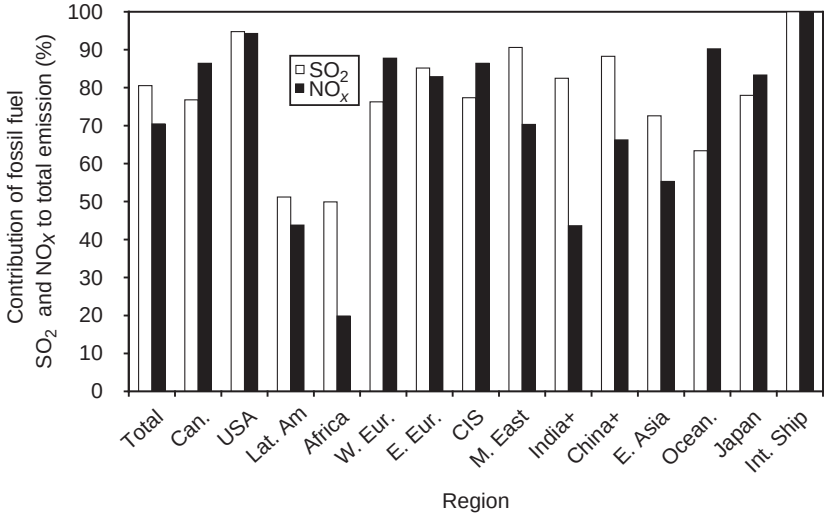


FIGURE 3 Share of fuel emissions of the total global sulfur and nitrogen emissions of 148.5 Tg SO_2 and 102.2 Tg NO_2. Data from van Ardenne *et al*. (2001).

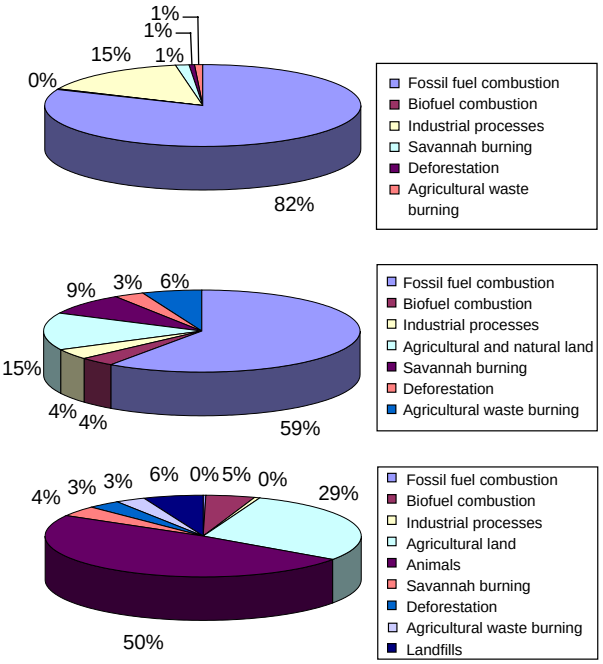


FIGURE 4 Contribution of different sources to the total anthropogenic emission of SO_2 (top of the figure), NO_x (middle), and NH_3 (bottom).

of emissions of SO_2 and, to a lesser extent, NO_x in the United States and Europe. This leveling is a net effect of decreased emissions in the industrialized areas of Europe and the United States and increased emissions in developing countries, especially Asia. Table I shows the estimated changes in emissions between 1990 and 1998 in Europe, the United States, and Asia. Europe shows the most positive trends in the world. Emission trends in the United States between 1990 and 1998 are somewhat upward for the nitrogen components, whereas SO_2 and volatile organic compounds (VOCs) show a downward trend of 18 and 14%, respectively. Emissions of SO_2, NO_x, NH_3, and VOCs show a downward trend of 41, 21, 14, and 24%, respectively. Clearly, policy has been much more successful in Europe than in the United States.

Table I shows that emissions of SO_2 and VOCs in the United States and Europe have decreased, whereas in Asia they have increased. The nitrogen components decreased only in Europe, whereas in the United States they remained about the same. In Asia, NO_x emissions more than doubled, whereas for NH_3 no data are available. There is a distinct difference between the policies in the United States and those in Europe to combat acidification. In the United States, the Clean Air Act is the driving force to reduce emissions, and in Europe the reductions are

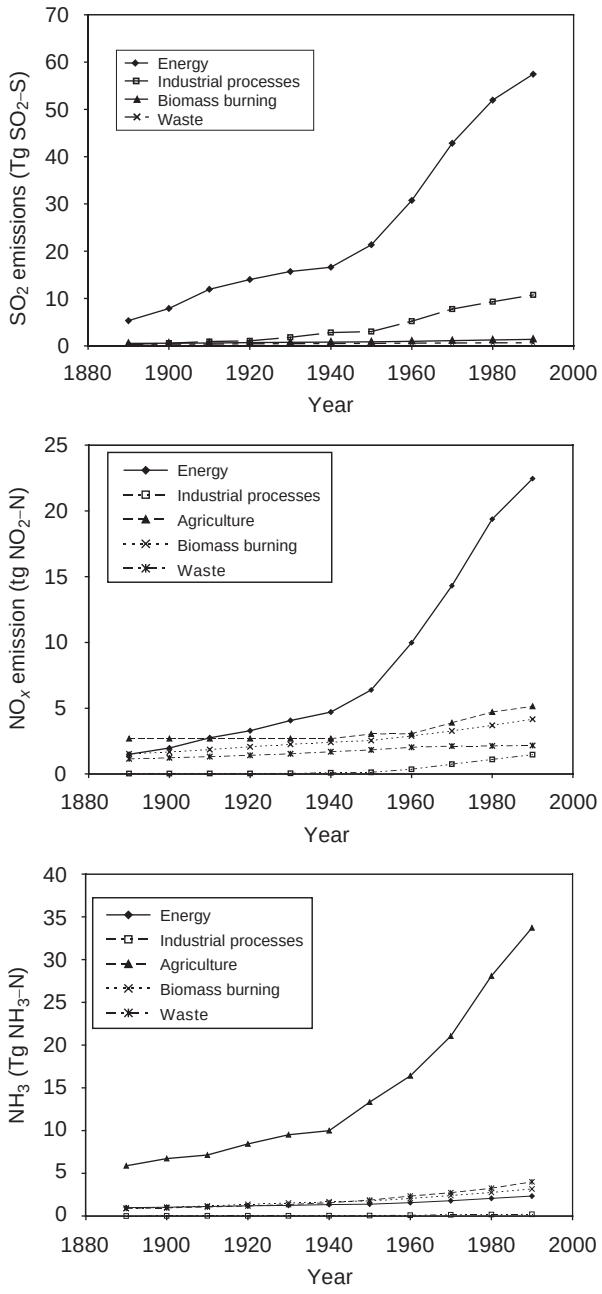


FIGURE 5 Trend in global SO_2, NO_x, and NH_3 emissions of different sources. Data from van Ardenne *et al.* (2001).

achieved through protocols under the Convention of Long-Range Transboundary Air Pollution, with the multipollutant–multieffect protocol signed in 1999 in Gothenburg being the most recent.

5.1 United States

For SO_2, the Acid Rain Program places a mandatory ceiling, or cap, on emissions nationwide from electric

TABLE I

Percentage Changes in Total Emissions of SO_2, NO_x, NH_3, and VOCs between 1990 and 1997 in Europe, the United States, and Asia

Component	Asia[a]	Europe	United States	Global
SO_2	+43	−41	−18	−3
NO_x	+29	−21	+1	+3
NH_3	n.a.[b]	−14	+4	n.a.
VOC	+12	−24	−14	+1

[a] Data from IPCC (2000) for the years 1990 to 1997. Streets *et al.* (2001) report increases of 16% for SO_2 and 53% for NO_x for northeast Asia.

[b] n.a., not available.

utilities and allocates emissions to these pollution sources in the form of allowances, or a permit to emit 1 ton of SO_2. Emissions from large electric power plants have been reduced through phase 1 of the Acid Rain Program. The emission trading has proven to be a cost-effective mechanism and has facilitated 100% compliance by affected sources and stimulated early emission reduction. Phase 2 was started in 2000 and will result in a reduction for a greater number of smaller plants, and emissions from larger plants will be further reduced.

Measures to limit NO_x emissions in the United States have largely focused on decreasing human health effects from tropospheric ozone, of which NO_x is a primary precursor. NO_x abatement has also been achieved as a means to address ecosystem acidification in conjunction with much larger decreases in SO_2 emissions. NO_x emissions from stationary and mobile sources are limited by efforts to comply with various current and future regulations, including National Ambient Air Quality Standards, New Source Performance Standards, Title IV of the 1990 Clean Air Act Amendments, the Ozone Transport Commission NO_x Budget Allowance Trading Program, state-implemented plans for NO_x emissions decreases, Section 126 of the 1990 Clean Air Act Amendments, and Mobile Source Emission Limits.

5.2 Europe

In Europe, the major efforts to decrease the effects of SO_2 and nitrogen emissions have been aimed at decreasing transfers to air, soil, and groundwater. Most of the measures in Europe are focused on decreasing human and plant exposure and decreasing ecosystem loads leading to acidification and eutrophication. Countries agreed to decrease air emissions by signing different protocols developed under the Convention on Long-Range Transboundary Air Pollution, including the first and second sulfur protocols, the first NO_x protocol, and the NMVOC (Non-Methane Volatile Organic Compounds) protocol. The last protocol, the Göteborg Protocol, is unique in the sense that it requires decreases in emissions of four pollutants with the objective of abating three specific effects—acidification, eutrophication, and the effects from tropospheric ozone on human health and vegetation. The protocol, which has been signed by 29 European countries together with the United States and Canada, is based on a gap-closure method aiming at decreasing the spatial exceedances of critical loads and levels in the most cost-efficient way. Critical loads for each European country are defined on the basis of information developed by each country. The agreed upon decreases in emissions for the European Union (EU) member states are expected to lead to overall decreases in the European (except Russia) emissions of 78% for SO_2 and approximately 44% for NO_x during the period 1990 to 2010. The corresponding figure for ammonia is 17%.

The protocols have had a major effect on the emission trends in Europe, especially for SO_2 (Table I). European emissions reductions are being made with the clear objective that environmental loads, exposures, and effects should be decreased—the so-called "effects-based approach," which was initiated under the Second Sulphur Protocol. For NO_x, the main measure to decrease emissions is exhaust gas regulations introduced in the EU countries in approximately 1990, resulting in the application of three-way catalysts in gasoline cars. Even regulations on heavy-duty vehicles have caused emission reductions of NO_x. Furthermore, selective catalytic reduction technologies (SCR) with ammonia or urea as a reductor have been implemented in many combustion plants. In eastern Europe, the main cause of decreases in NO_x emissions is the shutdown of a large number of industrial plants. Despite declining emissions rates, total NO_x emissions have remained steady or even increased during the same period in North America and Europe due to increases in vehicle kilometers/miles traveled, increases in electricity usage, and the sometimes differing regulatory frameworks applied to various sectors. Decreasing total NO_x emissions likely means that continuing technological advances need to be combined with regulatory approaches (e.g., emissions caps).

6. ACID DEPOSITION

Within the causal chain, the activities and resulting emissions form the basis of the effects and thus the abatement should be focused on the emissions. However, the effects are linked primarily with acid deposition. Therefore, the real progress in reducing the effects should be derived from trends in effects or in acid deposition. Erisman *et al.* evaluated the progress in acid and nitrogen policies in Europe. Potential acid deposition, the molar sum of SO_2, NO_x, and NH_3, has decreased 2.5% per year between 1980 and 1999. This is mainly due to the reduction in sulfur emissions, which have decreased 59% averaged throughout Europe. The decrease in emissions has resulted in improved environmental quality. Ambient concentrations of SO_2 have decreased well below critical levels. NO_2 exceedances of critical levels are now only observed near sources and in industrialized areas. Deposition measurements are limited mostly to wet deposition. Trends in measurements show a decrease in wet deposition. However, the decrease is smallest in remote areas, where the most vulnerable ecosystems are located. This is due to nonlinearity in emissions and depositions. Consequently, appreciable emission reductions are still needed to protect ecosystems in remote areas. Although acidity has been decreased to a large extent, nitrogen pollution is still excessive and needs to be reduced to reach critical loads.

In the northeast and mid-Atlantic regions of the United States, where ecosystems are most sensitive to acidic deposition, sulfate levels in precipitation (wet deposition) have declined by up to 25%, mirroring the reductions in SO_2 emissions achieved through the implementation of the Acid Rain Program. No distinct regional trends in nitrate have been detected, consistent with NO_x emissions, which have remained approximately the same. Trends in dry deposition of sulfur and nitrogen are comparable.

Because sulfur emissions decreased more than nitrogen emissions, the relative contribution of nitrogen deposition compared to sulfur deposition has increased. The focus of policies should therefore be on nitrogen emissions, both oxidized and reduced nitrogen.

The unbalanced reduction in emissions can have unexpected consequences. The large decrease in base cation emissions, which was much faster than the decrease in sulfur emissions, has caused precipitation pH to decrease rapidly in several areas. Another example is the nonlinearity in sulfur and nitrogen emission and the respective deposition in remote areas. Nonlinearity in the relationship between the emission and deposition patterns for sulfur and oxidized nitrogen has been detected, which leads to slower than expected reductions in critical loads exceedance in remote areas. There is evidence that NO_3 in rain and wet deposition is increasing at remote, mainly West Coast locations, whereas close to the major sources a small reduction has been detected. This is the result of the nonlinearity.

In relation to acidification, little emphasis is usually placed on the role of deposition of base cations such as Na^+, Mg^{2+}, Ca^{2+}, and K^+. Besides their ability to neutralize acid input, base cations are important nutrient elements for ecosystems. From a variety of measurements of Ca in precipitation in Europe and North America, Hedin *et al.* detected a long-term decline in 1994. The decline in Ca is due to the abatement of particle emissions from coal combustion and industrial sources. The reduction has resulted in a decrease in the neutralizing capacity of the atmosphere, leading to a decrease in pH. Furthermore, because some abatement of SO_2 at industrial sources coincided with a decrease in Ca, the net effect on atmospheric acidity was small; thus, recovery of ecosystems would not proceed as expected. At the time, no data on long-term Ca aerosol concentrations were available. In 1998, Lee *et al.* reanalyzed previously unreported long-term data sets of the atmospheric aerosol composition for the United Kingdom and showed a clear decline in Ca during the period 1970–1995. Paralleling this decline were trends in scandium and arsenic. Scandium is often used as a tracer for noncombustion sources, whereas arsenic is often used as a tracer for combustion. Declines of both metals clearly show a strong decline in fly-ash from industrial and combustion sources; hence, the decline in Ca can be understood.

7. BENEFITS AND RECOVERY OF ECOSYSTEMS

Policies to reduce emissions and deposition are ultimately aimed at decreasing negative effects and/or increasing recovery of affected ecosystems. The response of effects to decreases in emissions depends on the type of effect. Figure 6 shows relative response times to changes in emissions. The effect is fastest for air concentrations and the resulting improvement in visibility, the direct plant response, and acute human health. However, it takes longer for chronic health, aquatic systems, or plant responses to show the results of reduced emissions. It will take several years

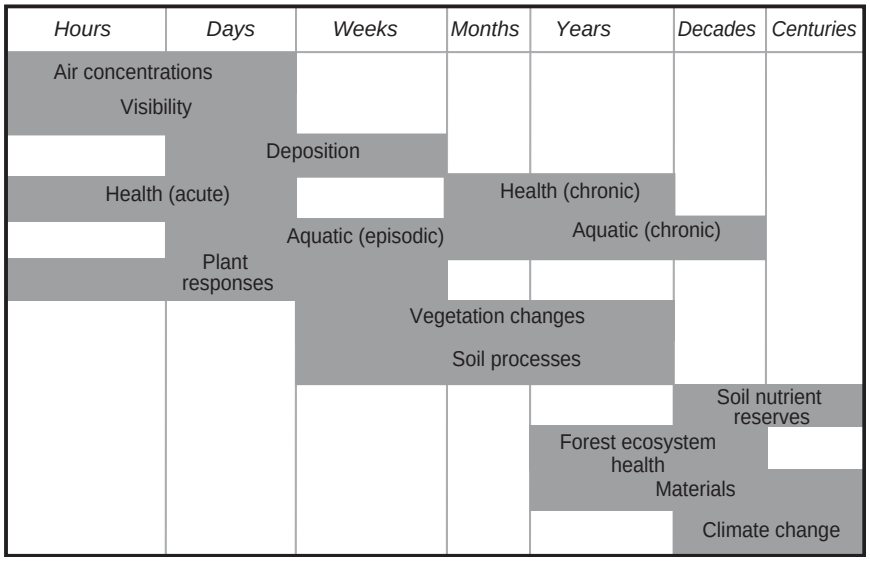


FIGURE 6 Relative response times to changes in emissions. Data from U.S. National Acid Precipitation Assessment Program (2000) and Galloway (2001).

to centuries before the forest ecosystem or soil nutrient status respond to the reductions. This difference in response times has to be taken into account when evaluating the changes in effect parameters, especially because the decrease in emissions started in the 1980s and 1990s.

7.1 Surface Water

Lake ecosystems have improved in northern Europe and North America, where emission reductions occurred. In general, SO_4 concentrations have decreased following the emission trends, but nitrogen concentrations have not shown changes. Data from 98 sites were tested for trends in concentrations over the 10-year period 1989–1998. The (grouped) sites clearly showed significant decreases in SO_4 concentrations. Nitrate, however, showed no regional patterns of change, except possibly for central Europe: Decreasing trends occurred in the Black Triangle. Concentrations of base cations declined in most regions. All regions showed tendencies toward increasing dissolved organic carbon. Recovery from acidification reflected by an increase in surface water acid neutralizing capacity (ANC) and pH was significant in the Nordic countries/United Kingdom region. In central Europe, there was a regional tendency toward increasing ANC, but large spatial differences were found with the low ANC sites showing the largest recovery. Nonforested sites showed clear and consistent signals of recovery in ANC and pH and appropriate (relative to SO_4 trends) rates of base cation declines. Hence, it was concluded that the observed recovery was associated with declining SO_4.

In the most acidic sites in central Europe, improvements in water quality have not yet resulted in improvements in biology. Biological improvements of these sites require considerable improvements in water quality with respect to acidification.

In many lakes in Scandinavia, there is evidence of a small but significant recovery and many species that died because of acidification are returning. The positive signs are mainly observed in lakes and streams with limited acidification. For the most acidified waters, the signs of recovery are still small and unclear.

In a study by Stoddard *et al.*, data for 205 sites in eight regions in North America and Europe between 1980 and 1995 were used to test trends. The data they used were primarily from the International Cooperative Program (ICP) Waters study. They found trends of decreasing SO_4 concentrations in all regions except the United Kingdom and no or very small changes in NO_3. SO_4 levels declined from 0 to -4 μeq/liter/year in the 1980s to -1 to -8 μeq/liter/year in the 1990s. Recovery of alkalinity was associated with the decrease in SO_4, especially in the 1990s.

7.2 Forests

Forested catchments in Europe are studied within the ICP Integrated Monitoring (IM). These sites are located in undisturbed areas, such as natural parks or comparable areas. The network currently covers

50 sites in 22 countries. Fluxes and trends of sulfur and nitrogen compounds were recently evaluated for 22 sites. The site-specific trends were calculated for deposition and runoff water fluxes and concentrations using monthly data and nonparametric methods. Statistically significant downward trends of SO_4 and NO_3 bulk deposition (fluxes and concentrations) were observed at 50% of the sites. Sites with higher nitrogen deposition and lower carbon/nitrogen ratios clearly showed an increased risk of elevated nitrogen leaching. Decreasing SO_4 and base cation trends in output fluxes and/or concentrations of surface/soil water were commonly observed at the ICP IM sites. At several sites in Nordic countries, decreasing NO_3 and H^+ trends (increasing pH) were also observed. These results partly confirm the effective implementation of emission reduction policy in Europe. However, clear responses were not observed at all sites, showing that recovery at many sensitive sites can be slow and that the response at individual sites may vary greatly.

Defoliation and discoloration of stands have been monitored in Europe on a regular 16×16-km grid since 1986—the so-called level 1 program [ICP-Forests (F)]. The network currently covers 374,238 sample trees distributed on 18,717 plots in 31 countries. At approximately 900 sites in Europe, key parameters, such as deposition, growth, soil chemistry, and leaf content, have been monitored since 1994 to study cause–effect relationships and the chemical and physical parameters determining forest ecosystem vitality (level 2). The duration of the level 2 program is too short for trend detection. These data are more relevant for studying cause–effect relationships. Figure 7 shows the annual variation in defoliation for different tree species in Europe as reported in the extended summary report by ICP-F. It is concluded by ICP-F that in all parts of Europe there is defoliation to a various extent. Of the 1999 total transnational tree sample, mean defoliation is 19.7%. Of the main tree species, *Quercus robur* has the highest defoliation (25.1%), followed by *Picea abies* (19.7%), *Fagus sylvatica* (19.6%), and *Pinus sylvestris* (18.9%). The trend in defoliation over 14 years for continuously observed plots shows the sharpest deterioration of *Pinus pinaster* and *Quercus iles* in southern Europe. *Fagus sylvatica* deteriorated in the sub-Atlantic, mountainous (south), and continental regions. *Picea abies* deteriorated in several parts of Europe but improved in the main damage areas of central Europe since the mid-1990s.

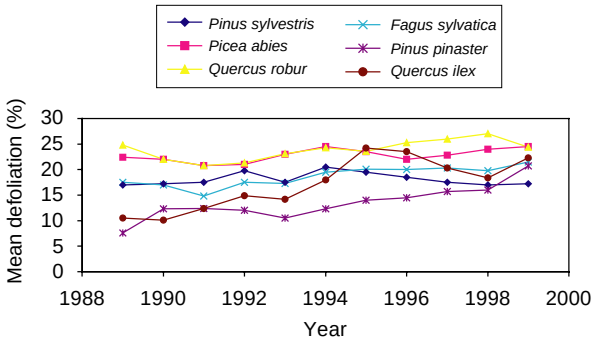


FIGURE 7 Development of mean defoliation of the six most abundant species in Europe. Data from ICP (2000).

Because there are no specific symptoms of individual types of damage, defoliation reflects the impact of many different natural and anthropogenic factors. Weather conditions and biotic stresses are most frequently cited. Several countries refer to air pollution as a predisposing, accompanying, or triggering factor, but the degree to which air pollution explains the spatial and temporal variation of defoliation on the large scale cannot be derived from crown condition assessment alone. Statistical analysis of 262 of the 860 level 2 plots indicates that defoliation is influenced mainly by stand age, soil type, precipitation, and nitrogen and sulfur deposition. These factors explain 30–50% of the variation. Nitrogen deposition correlated with defoliation of spruce and oak, and sulfur deposition was found to correlate with defoliation of pine, spruce, and oak. Calculated drought stress was found to impact nearly all tree species, whereas calculated ozone mainly impacted broadleaves, specifically the common beech trees.

8. FUTURE ABATEMENT

There are basically two ways of reducing acid deposition. Emission control technologies can be attached to smokestacks at power plants and other industries, removing the acid gases before they are emitted into the atmosphere. In coal-fired power plants, sulfur emissions are removed with a "scrubber," in which a limestone slurry is injected into the flue gas to react with the SO_2. The resulting gypsum slurry can eventually be used in other industrial processes. The main problem with scrubbers is that they are expensive, and they decrease the overall operating efficiency of a power plant. The decreased efficiency results in increased emissions of carbon dioxide, a greenhouse gas.

The other alternative to reduce acid deposition is to burn less high-sulfur fossil fuel. This can be

accomplished by switching to alternative sources of energy or improving the efficiency of energy-consuming technologies. Coal-fired power plants can reduce SO_2 emissions by burning coal with a lower sulfur content. Another alternative is for these power plants to switch to fuels with lower acid gas emissions, such as natural gas. Ultimately, the most effective ways to reduce acid rain are the use of renewable energy and improving energy efficiency. Such measures will decrease the use of fossil fuels and therewith reduce the emissions. Renewable energy technologies, such as solar and wind energy, can produce electricity without any emissions of SO_2 or NO_x. There are also many ways to decrease consumption of energy by improving the efficiency of end-use technologies. Both renewable energy and energy efficiency have the added benefit that they also result in reduced emissions of carbon dioxide, the greenhouse gas most responsible for global warming.

It is technically feasible and likely economically possible to further decrease NO_x emissions from fossil fuel combustion to the point at which they become only a minor disturbance to the nitrogen cycle at all scales. Clean electric generation and transportation technologies are commercially available today, or will be commercially available within one or two decades, that have the potential to further decrease NO_x emissions in industrialized countries and to decrease NO_x emissions from projected business-as-usual levels in developing countries. In addition, technologies currently under development, such as renewable energy and hydrogen-based fuel cells, could operate with zero NO_x emissions.

During the 1970s, major attention was given to energy conservation measures designed to avoid depletion of coal, crude oil, and gas reserves. Since then, the known reserves of natural gas have increased from 40,000 to 146,400 billion m^3 in 1998. The current world coal, crude oil, and natural gas reserves are estimated to be 143,400, 500, and 131,800 Mtoe, respectively (Mtoe = million ton oil equivalents). Thus, there appears to be enough fossil fuel available for at least 50–100 years, and it is expected that this will increase as new technologies become available with which to discover and add new sources of fossil fuels. The current world concern about fossil fuel combustion is driven by the apparent necessity to decrease CO_2 emissions. The increase in CO_2 content of the atmosphere and the associated climate change will call for a drastic change in energy production and use. Focusing only on SO_2 and NO_x as pollutants from energy production and use will not lead to drastic changes in the energy systems. In order to abate these emissions from energy production, a multiple-pollutant approach should be used; that developed within the Göteborg Protocol appears to be very promising. Unless yet-to-be developed carbon sequestration technology emerges, decreasing CO_2 emissions will automatically decrease NO_x and SO_2, although this has to be proved. On the contrary, NO_x and SO_2 abatement almost always leads to increased energy use and thus increased CO_2 emissions. These technologies are focused on enhanced combustion, producing less NO_x (lean burn), or using end of pipe technologies, such as SCR or scrubbers. Both technologies decrease energy efficiency, and the majority of SCR technologies require ammonia or urea as a reductor, which is produced from natural gas.

8.1 Consequences of Carbon Management: Addressing Greenhouse Gases for Energy Nitrogen and Sulfur

If the energy system of the world remains based on fossil fuels throughout the 21st century and little is done to target the atmospheric emissions of CO_2, it is plausible that the atmospheric CO_2 concentration may triple its "preindustrial" concentration of approximately 280 parts per million by the year 2100. Strategies to slow the rate of buildup of atmospheric CO_2 are being developed worldwide, and all appear to have a positive effect of reducing the production of energy nitrogen and sulfur. First among these strategies is an increase in the efficiency of energy use throughout the global economy, resulting in less energy required to meet the variety of amenities that energy provides (e.g., mobility, comfort, lighting, and material goods). Greater energy efficiency directly decreases energy nitrogen and sulfur.

Other CO_2 emission-reduction strategies address the composition of energy supply. First, nuclear energy and renewable energy are nonfossil alternatives, resulting in CO_2 emissions only to the extent that fossil fuels are used to produce the nonfossil energy production facilities. Neither nuclear nor renewable energy sources produce energy nitrogen and sulfur, but nuclear waste is another source of concern.

Second, the mix of coal, oil, and natural gas in the energy supply affects CO_2 emissions because they have different carbon intensities (carbon content per unit of thermal energy). Specifically, coal has the highest and natural gas the lowest carbon intensity so

that shifts from coal to oil or natural gas and shifts from oil to natural gas, other factors held constant, reduce the greenhouse effect of the fossil fuel system. The nitrogen and sulfur intensity of fossil fuels (nitrogen and sulfur content per unit of thermal energy) differ in the same way (i.e., on average, the intensity of coal is highest and the intensity of natural gas is lowest). Thus, fuel shifts within the fossil fuel system that reduce greenhouse effects will also reduce fossil nitrogen and sulfur.

Third, many countries throughout the world are investing in research, development, and demonstration projects that explore the various forms of capturing carbon from combustion processes before it reaches the atmosphere and sequestering it on site or off. Several available technologies can be used to separate and capture CO_2 from fossil-fueled power plant flue gases, from effluents of industrial processes such as iron, steel, and cement production, and from hydrogen production by reforming natural gas. CO_2 can be absorbed from gas streams by contact with amine-based solvents or cold methanol. It can be removed by absorption on activated carbon or other materials or by passing the gas through special membranes. However, these technologies have not been applied at the scale required to use them as part of a CO_2 emissions mitigation strategy. The goal is to sequester the carbon in a cost-effective way, for example, as CO_2 injected deep below ground in saline aquifers. This is a relatively new area of research and development, and little attention has been given to the consequences of fossil carbon sequestration for energy nitrogen and sulfur, but decreases are a likely result. For example, to capture and sequester the carbon in coal will require the gasification of coal and the subsequent production of hydrogen and a CO_2 gas stream. Coal nitrogen or sulfur should be amenable to independent management, with results that include extraction as a saleable product or cosequestration below ground with CO_2. The first of these results is one of many "polygeneration" strategies for coal, in which products may include electricity, hydrogen, process heat, hydrocarbons, dimethyl ether, and nitrogen. The long-term goal is to run the economy on noncarbon secondary energy sources, specifically electricity and hydrogen, while sequestering emissions of CO_2.

9. CONCLUSION

The atmosphere functions as a pool and chemical reaction vessel for a host of substances. Many of the most important ones (e.g., oxygen, carbon dioxide, nitrogen, and sulfur compounds) are released by the activity of organisms. Often with the help of the water cycle, they pass through the atmosphere and are eventually taken up again into soil, surface water, and organic matter. Through technology related to energy production and use, humans have added enormously to the atmospheric burden of some of these substances, with far-reaching consequences for life and the environment. The evidence is clearest in the case of acid deposition: gases, particles, and precipitation depositing to the surface causing acidification.

In North America and in some European nations, public concern regarding the effects of acid rain has been transformed into regulations restricting the amount of SO_2 and NO_x released by electric utilities and industries. The result has been a decrease in annual acidic deposition in some areas, especially due to the reduction in sulfates. There is also evidence that when acid deposition is reduced, ecosystems can recover. The chemistry of several lakes has improved during the past 20 years, but full biological recovery has not been observed. Nitrogen has become the main concern because emissions have been reduced much less than those of SO_2. In the future, it will be very important for the industrialized world to transfer its technology and experience to the developing world in order to ensure that the same acid rain problems do not occur as these countries consume more energy during the process of industrialization. Whereas in the past few years the acid deposition decreased in Europe and the United States, an increase has been observed in Asia. It is essential to establish or optimize monitoring programs aimed at following the trends in acid deposition and recovery.

The acid deposition levels in the industrialized areas of the world are well above the critical thresholds of different ecosystems. Because energy emissions are the largest source for acid deposition, abatement is necessary. Multipollutant–multieffect approaches are necessary to ensure that measures are cost-effective and do not create problems for other areas. It has to be determined which effects limit the emissions in different areas most: Is it because of the air quality and the effects on humans, or is it the ecosystem loading? In this way, regional emissions can be optimized and caps can be set to prevent any effects. Furthermore, it is probably most effective to follow CO_2 emission reduction options and to then consider SO_2 and NO_x, instead of the reverse. In most cases, when CO_2 emission is decreased, SO_2

and NO_x are also decreased, whereas if measures to reduce SO_2 and NO_x are taken first, CO_2 is increased because of the energy penalty or additional energy use of abatement options.

SEE ALSO THE FOLLOWING ARTICLES

Air Pollution from Energy Production and Use • Air Pollution, Health Effects of • Biomass: Impact on Carbon Cycle and Greenhouse Gas Emissions • Clean Air Markets • Ecosystem Health: Energy Indicators • Ecosystems and Energy: History and Overview • Greenhouse Gas Emissions from Energy Systems, Comparison and Overview • Hazardous Waste from Fossil Fuels

Further Reading

Breemen, N. van, Burrough, P. A., Velthorst, E. J., Dobben, H. F. van, Wit, T. de, Ridder, T. B., and Reinders, H. F. R. (1982). Soil acidification from atmospheric ammonium sulphate in forest canopy throughfall. *Nature* **299,** 548–550.

Brimblecombe, P. (1989). "The Big Smoke." Routledge, London.

Chadwick, M. J., and Hutton, M. (1990). "Acid Depositions in Europe. Environemental Effects, Control Strategies and Policy Options." Stockholm Environmental Institute, Stockholm, Sweden.

Cowling, E., Galloway, J., Furiness, C., Barber, M., Bresser, T., Cassman, K., Erisman, J. W., Haeuber, R., Howarth, R., Melillo, J., Moomaw, W., Mosier, A., Sanders, K., Seitzinger, S., Smeulders, S., Socolow, R., Walters, D., West, F., and Zhu, Z. (2001). Optimizing nitrogen management in food and energy production and environmental protection: Proceedings of the 2nd International Nitrogen Conference on Science and Policy. *Sci. World* **1**(Suppl. 2), 1–9.

EMEP/MSC-W (1998). Transboundary acidifying air pollution in Europe. Estimated dispersion of acidifying and eutrophying compounds and comparison with observations, MSC-W (Meteorological Synthesizing Centre–West) status report. Norwegian Meteorological Institute, Oslo, Norway.

Erisman, J. W. (2000). Editorial: History as a basis for the future: The environmental science and policy in the next millennium. *Environ. Sci. Pollut.* **3,** 5–6.

Erisman, J. W., and Draaijers, G. P. J. (1995). Atmospheric deposition in relation to acidification and eutrophication. *Stud. Environ. Res.* **63**.

Erisman, J. W., Grennfelt, P., and Sutton, M. (2001). Nitrogen emission and deposition: The European Perspective. *Sci. World* **1,** 879–896.

European Environmental Agency (1999). "Environment in the European Union at the Turn of the Century." European Environmental Agency, Copenhagen.

European Environmental Agency (2000). "Environmental Signals 2000." European Environmental Agency, Copenhagen.

Galloway, J. N. (2001). Acidification of the world: Natural and anthropogenic. *Water Soil Air Pollut.* **130,** 17–24.

Gorham, E. (1994). Neutralising acid rain. *Nature* **367,** 321.

Hedin, L. O., Granat, L., Likens, G. E., Buishand, T. A., Galloway, J. N., Butler, T. J., and Rodhe, H. (1994). Steep declines in atmospheric base cations in regions of Europe and North America. *Nature* **367,** 351–354.

Heij, G. J., and Erisman, J. W. (1997). Acidification research in the Netherlands; Report of third and last phase. *Stud. Environ. Sci.* **69**.

ICP (2000). Forest condition in Europe, 2000 executive report. *In* "UN ECE Convention on Long-Range Transboundary Air Pollution International Co-operative Programme on Assessment and Monitoring of Air Pollution Effects on Forests." Federal Research Centre for Forestry and Forest Products (BFH), Hamburg, Germany.

ICP Waters (2000). The 12-year report: Acidification of surface water in Europe and North America; Trends, biological recovery and heavy metals. Norwegian Institute for Water Research, Oslo.

IPCC (2000). Third assessment report on climate change. UN/FCCC, Bonn, Germany.

Lee, D. S., Espenhahn, S. E., and Baker, S. (1998). Evidence for long-term changes in base cations in the atmospheric aerosol. *J. Geophys. Res.* **103,** 21955–21966.

Matyssek, R., Keller, T., and Günthardt-Goerg, M. S. (1990). Ozonwirkungen auf den verschiedenen Organisationsebenen in Holzpflanzen. *Schweizerische Z. Forstwesen* **141,** 631–651.

Mylona, S. (1993). Trends of sulphur dioxide emissions, air concentrations and depositions of sulphur in Europe since 1880, EMEP/MSC-W Report 2/93. Norwegian Meteorological Institute, Oslo.

Nihlgård, B. (1985). The ammonium hypothesis—An additional explanation for the forest dieback in Europe. *Ambio* **14,** 2–8.

Nilles, M. A., and Conley, B. E. (2001). Changes in the chemistry of precipitation in the United States, 1981–1998. *Water Air Soil Pollut.* **130,** 409–414.

Nilsson, J. (1986). Critical deposition limits for forest soils. *In* "Critical Loads for Nitrogen and Sulphur" (J. Nilsson, Ed.), Miljørapport 1986, Vol. 11, pp. 37–69. Nordic Council of Ministers, Copenhagen.

Nilsson, J., and Grennfelt, P. (1988). "Critical Loads for Sulphur and Nitrogen," Miljørapport 1988, Vol. 15. Nordic Council of Ministers, Copenhagen.

Smil, V. (2000). Energy in the twentieth century. *Annu. Rev. Energy Environ.* **25,** 21–51.

Smith, R. A. (1852). On the air and rain of Manchester. *Manchester Literary Philos. Soc.* **10,** 207–217.

Smith, R. A. (1872). "Air and Rain. The Beginnings of a Chemical Climatology." Longmans, Green, London.

Stoddard, J. L., Jeffries, D. S., Lukewille, A., Clair, T. A., Dillon, P. J., Driscoll, C. T., Forsius, M., Johannessen, M., Kahl, J. S., Kellogg, J. H., Kemp, A., Mannio, J., Monteith, D. T., Murdoch, P. S., Patrick, S., Rebsdorf, A., Skjelkvale, B. L., Stainton, M. P., Traaen, T., van Dam, H., Webster, K. E., Wieting, J., and Wilander, A. (1999). Regional trends in aquatic recovery from acidification in North America and Europe. *Nature* **401,** 575–578.

Streets, D. G., Tsai, N. Y., Akimoto, H., and Oka, K. (2001). Trends in emissions of acidifying species in Asia, 1985–1997. *Water Air Soil Pollut.* **130,** 187–192.

Ulrich, B. (1983). Interaction of forest canopies with atmospheric constituents: SO_2, alkali and earth alkali cations and chloride.

In "Effects of Accumulation of Air Pollutants in Forest Ecosystems" (B. Ulrich and J. Pankrath, Eds.), pp. 33–45. Reidel, Germany.

U.S. National Acid Precipitation Assessment Program (2000). "Acid Deposition: State of Science and Technology Progress Report 2000." U.S. National Acid Precipitation Assessment Program, Washington, DC.

Van Ardenne, J. A., Dentener, F. J., Olivier, J. G. J., Klein Goldewijk, C. G. M., and Lelieveld, J. (2001). A 1×1 resolution data set of historical anthropogenic trace gas emissions for the period 1890–1990. *Global Biogeochem. Cycl.* **15,** 909–928.

WGE (2000). 9th annual report 2000. *In* "UN ECE Convention on Long-Range Transboundary Air Pollution International Co-operative Programme on Integrated Monitoring of Air Pollution Effects on Ecosystems" (S. Kleemola and M. Forsius, Eds.). Working Group on Effects of the Convention on Long-Range Transboundary Air Pollution, Helsinki.

Aggregation of Energy

CUTLER J. CLEVELAND and ROBERT K. KAUFMANN
Boston University
Boston, Massachusetts, United States

DAVID I. STERN
Rensselaer Polytechnic Institute
Troy, New York, United States

1. **Energy Aggregation and Energy Quality**
2. **Economic Approaches to Energy Quality**
3. **Alternative Approaches to Energy Aggregation**
4. **Case Study 1: Net Energy from Fossil Fuel Extraction in the United States**
5. **Case Study 2: Causality in the Energy/GDP Relationship**
6. **Case Study 3: The Determinants of the Energy/GDP Relationship**
7. **Conclusions and Implications**

Glossary

Divisia index A method of aggregation used in economics that permits variable substitution among material types without imposing a priori restrictions on the degree of substitution.

emergy The quantity of solar energy used directly and indirectly to produce a natural resource, good, or service.

energy quality The relative economic usefulness per heat equivalent unit of different fuels and electricity.

energy/real gross domestic product (GDP) ratio (E/GDP ratio) The ratio of total energy use to total economic activity; a common measure of macroeconomic energy efficiency.

energy return on investment (EROI) The ratio of energy delivered to energy costs.

exergy The useful work obtainable from an energy source or material is based on the chemical energy embodied in the material or energy based on its physical organization relative to a reference state. Exergy measures the degree to which a material is organized relative to a random assemblage of material found at an average concentration in the crust, ocean, or atmosphere.

Granger causality A statistical procedure that tests whether (i) one variable in a relation can be meaningfully described as a dependent variable and the other variable as an independent variable, (ii) the relation is bidirectional, or (ii) no meaningful relation exists. This is usually done by testing whether lagged values of one of the variables add significant explanatory power to a model that already includes lagged values of the dependent variable and perhaps also lagged values of other variables.

marginal product of energy The value marginal product of a fuel in production is the marginal increase in the quantity of a good or service produced by the use of one additional heat unit of fuel multiplied by the price of that good or service.

net energy analysis Technique that compares the quantity of energy delivered to society by an energy system to the energy used directly and indirectly in the delivery process.

Investigating the role of energy in the economy involves aggregating different energy flows. A variety of methods have been proposed, but none is accepted universally. This article shows that the method of aggregation affects analytical results. We review the principal assumptions and methods for aggregating energy flows: the basic heat equivalents approach, economic approaches using prices or marginal product for aggregation, emergy analysis, and thermodynamic approaches such as exergy analysis. We argue that economic approaches such as the index or marginal product method are superior because they account for differences in quality among different fuels. We apply economic approaches to three case studies of the U.S. economy. In the first, we account for energy quality to assess changes in the energy surplus delivered by the extraction of fossil fuels from 1954 to 1992. The second and third case studies examine the effect of energy quality on statistical analyses of the relation between energy use and gross domestic product

 © 2004 Elsevier Inc. All rights reserved.

(GDP). First, a quality-adjusted index of energy consumption is used in an econometric analysis of the causal relation between energy use and GDP from 1947 to 1996. Second, we account for energy quality in an econometric analysis of the factors that determined changes in the energy/GDP ratio from 1947 to 1996. Without adjusting for energy quality, the results imply that the energy surplus from petroleum extraction is increasing, that changes in GDP drive changes in energy use, and that GDP has been decoupled from aggregate energy. These conclusions are reversed when we account for changes in energy quality.

1. ENERGY AGGREGATION AND ENERGY QUALITY

Aggregation of primary-level economic data has received substantial attention from economists for a number of reasons. Aggregating the vast number of inputs and outputs in the economy makes it easier for analysts to discern patterns in the data. Some aggregate quantities are of theoretical interest in macroeconomics. Measurement of productivity, for example, requires a method to aggregate goods produced and factors of production that have diverse and distinct qualities. For example, the post-World War II shift toward a more educated workforce and from nonresidential structures to producers' durable equipment requires adjustments to methods used to measure labor hours and capital inputs. Econometric and other forms of quantitative analysis may restrict the number of variables that can be considered in a specific application, again requiring aggregation. Many indexes are possible, so economists have focused on the implicit assumptions made by the choice of an index in regard to returns to scale, substitutability, and other factors. These general considerations also apply to energy.

The simplest form of aggregation, assuming that each variable is in the same units, is to add up the individual variables according to their thermal equivalents (Btus, joules, etc.). Equation (1) illustrates this approach:

$$E_t = \sum_{i=1}^{N} E_{it}, \qquad (1)$$

where E is the thermal equivalent of fuel i (N types) at time t. The advantages of the thermal equivalent approach are that it uses a simple and well-defined accounting system based on the conservation of energy and the fact that thermal equivalents are easily measured. This approach underlies most methods of energy aggregation in economics and ecology, such as trophic dynamics national energy accounting, energy input–output modeling in economies and ecosystems, most analyses of the energy/gross domestic product (GDP) relationship and energy efficiency, and most net energy analyses.

Despite its widespread use, aggregating different energy types by their heat units embodies a serious flaw: It ignores qualitative differences among energy vectors. We define energy quality as the relative economic usefulness per heat equivalent unit of different fuels and electricity. Given that the composition of energy use changes significantly over time (Fig. 1), it is reasonable to assume that energy quality has been an important economic driving force. The quality of electricity has received considerable attention in terms of its effect on the productivity of labor and capital and on the quantity of energy required to produce a unit of GDP. Less attention has been paid to the quality of other fuels, and few studies use a quality-weighting scheme in empirical analysis of energy use.

The concept of energy quality needs to be distinguished from that of resource quality. Petroleum and coal deposits may be identified as high-quality energy sources because they provide a very high energy surplus relative to the amount of energy required to extract the fuel. On the other hand, some forms of solar electricity may be characterized as a

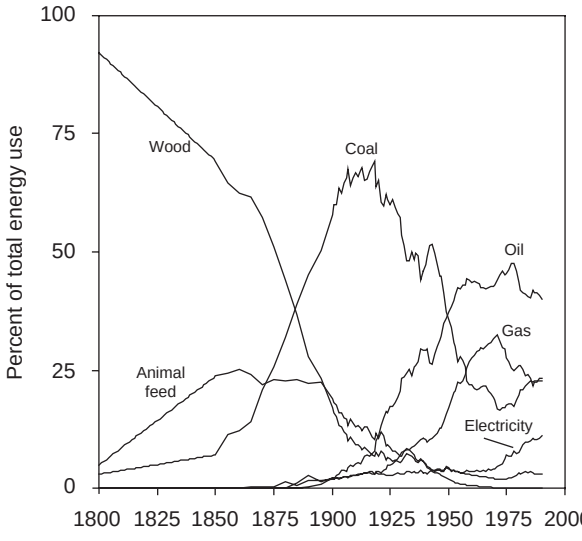


FIGURE 1 Composition of primary energy use in the United States. Electricity includes only primary sources (hydropower, nuclear, geothermal, and solar).

low-quality source because they have a lower energy return on investment (EROI). However, the latter energy vector may have higher energy quality because it can be used to generate more useful economic work than one heat unit of petroleum or coal.

Taking energy quality into account in energy aggregation requires more advanced forms of aggregation. Some of these forms are based on concepts developed in the energy analysis literature, such as exergy or emergy analysis. These methods take the following form:

$$E_t^* = \sum_{i=1}^{N} \lambda_{it} E_{it}, \tag{2}$$

where λ represents quality factors that may vary among fuels and over time for individual fuels. In the most general case that we consider, an aggregate index can be represented as

$$f(E_t) = \sum_{i=1}^{N} \lambda_{it} g(E_{it}), \tag{3}$$

where $f(\,)$ and $g(\,)$ are functions, λ_{it} are weights, the E_i are the N different energy vectors, and E_t is the aggregate energy index in period t. An example of this type of indexing is the discrete Divisia index or Tornquist–Theil index described later.

2. ECONOMIC APPROACHES TO ENERGY QUALITY

From an economic perspective, the value of a heat equivalent of fuel is determined by its price. Price-taking consumers and producers set marginal utilities and products of the different energy vectors equal to their market prices. These prices and their marginal productivities and utilities are set simultaneously in general equilibrium. The value marginal product of a fuel in production is the marginal increase in the quantity of a good or service produced by the use of one additional heat unit of fuel multiplied by the price of that good or service. We can also think of the value of the marginal product of a fuel in household production.

The marginal product of a fuel is determined in part by a complex set of attributes unique to each fuel, such as physical scarcity, the capacity to do useful work, energy density, cleanliness, amenability to storage, safety, flexibility of use, and cost of conversion. However, the marginal product is not uniquely fixed by these attributes. Rather, the energy vector's marginal product varies according to the activities in which it is used; how much and what form of capital, labor, and materials it is used in conjunction with; and how much energy is used in each application. As the price rises due to changes on the supply side, users can reduce their use of that form of energy in each activity, increase the amount and sophistication of capital or labor used in conjunction with the fuel, or stop using that form of energy for lower value activities. All these actions raise the marginal productivity of the fuel. When capital stocks have to be adjusted, this response may be somewhat sluggish and lead to lags between price changes and changes in the value marginal product.

The heat equivalent of a fuel is just one of the attributes of the fuel and ignores the context in which the fuel is used; thus, it cannot explain, for example, why a thermal equivalent of oil is more useful in many tasks than is a heat equivalent of coal. In addition to attributes of the fuel, marginal product also depends on the state of technology, the level of other inputs, and other factors. According to neoclassical theory, the price per heat equivalent of fuel should equal its value marginal product and, therefore, represent its economic usefulness. In theory, the market price of a fuel reflects the myriad factors that determine the economic usefulness of a fuel from the perspective of the end user.

Consistent with this perspective, the price per heat equivalent of fuel varies substantially among fuel types (Table I). The different prices demonstrate that end users are concerned with attributes other than heat content. Ernst Berndt, an economist at MIT, noted that because of the variation in attributes among energy types, the various fuels and electricity are less than perfectly substitutable, either in production or in consumption. For example, from the point of view of the end user, 1 Btu of coal is not perfectly substitutable with 1 Btu of electricity; since the electricity is cleaner, lighter, and of higher quality, most end users are willing to pay a premium price per Btu of electricity. However, coal and electricity are substitutable to a limited extent because if the premium price for electricity were too high, a substantial number of industrial users might switch to coal. Alternatively, if only heat content mattered and if all energy types were then perfectly substitutable, the market would tend to price all energy types at the same price per Btu.

Do market signals (i.e., prices) accurately reflect the marginal product of inputs? Empirical analysis of the relation between relative marginal product and price in U.S. energy markets suggests that this is

TABLE I

U.S. Market Price for Various Energy Types[a]

Energy type	Market price ($/10^6 btu)
Coal	
Bituminous	
Mine-mouth	
Consumer cost	
Anthracite	
Mine-mouth	
Oil	
Wellhead	2.97
Distillate oil	7.70
Jet fuel	4.53
LPG	7.42
Motor gasoline	9.73
Residual fuel oil	2.83
Biofuels	
Consumer cost	
Natural gas	
Wellhead	2.10
Consumer cost	20.34

[a] *Source*. Department of Energy (1997). Values are 1994 prices.

TABLE II

Marginal Product of Coal, Oil, Natural Gas, and Electricity Relative to One Another

	Minimum	Year	Maximum	Year
Oil : coal	1.83	1973	3.45	1990
Gas : coal	1.43	1973	2.76	1944
Electricity : coal	4.28	1986	16.42	1944
Oil : gas	0.97	1933	1.45	1992
Electricity : oil	1.75	1991	6.37	1930
Electricity : gas	2.32	1986	6.32	1930

indeed the case. In the case of the United States, there is a long-term relation between relative marginal product and relative price, and several years of adjustment are needed to bring this relation into equilibrium. The results are summarized in Table II and suggest that over time prices do reflect the marginal product, and hence the economic usefulness, of fuels.

Other analysts have calculated the average product of fuels, which is a close proxy for marginal products. Studies indicate that petroleum is 1.6–2.7 times more productive than coal in producing industrial output, and that electricity is 2.7–18.3 times more productive than coal.

2.1 Price-Based Aggregation

If marginal product is related to its price, energy quality can be measured by using the price of fuels to weight their heat equivalents. The simplest approach defines the weighting factor (λ) in Eq. (2) as

$$\lambda_{it} = \frac{P_{it}}{P_{1t}}, \tag{4}$$

where P_{it} is the price per Btu of fuel. In this case, the price of each fuel is measured relative to the price of fuel type 1.

The quality index in Eq. (4) embodies a restrictive assumption—that fuels are perfect substitutes—and the index is sensitive to the choice of numeraire. Because fuels are not perfect substitutes, an increase in the price of one fuel relative to the price of output will not be matched by equal changes in the prices of the other fuels relative to the price of output. For example, the increase in oil prices in 1979–1980 would cause an aggregate energy index that uses oil as the numeraire to decline dramatically. An index that uses coal as the numeraire would show a large decline in 1968–1974, one not indicated by the oil-based index.

To avoid dependence on a numeraire, a discrete approximation to the Divisia index can be used to aggregate energy. The formula for constructing the discrete Divisia index E^* is

$$\begin{aligned} &\ln E_t^* - \ln E_{t-1}^* \\ &= \sum_{i=1}^{n} \left(\left(\frac{P_{it}E_{it}}{2\sum_{i=1}^{n} P_{it}E_{it}} + \frac{P_{it-1}E_{it-1}}{2\sum_{i=1}^{n} P_{it-1}E_{it-1}} \right) \right. \\ &\quad \left. \times (\ln E_{it} - \ln E_{it-1}) \right) \end{aligned} \tag{5}$$

where P are the prices of the n fuels, and E are the quantities of Btu for each fuel in final energy use. Note that prices enter the Divisia index via cost or expenditure shares. The Divisia index permits variable substitution among material types without imposing a priori restrictions on the degree of substitution. This index is an exact index number representation of the linear homogeneous translog production function, where fuels are homothetically weakly separable as a group from the other factors of production. With reference to Eq. (3), $f(\,) = g(\,) = \Delta\ln(\,)$, whereas λ_{it} is given by the average cost share over the two periods of the differencing operation.

2.2 Discussion

Aggregation using price has its shortcomings. Prices provide a reasonable method of aggregation if the aggregate cost function is homothetically separable in the raw material input prices. This means that the elasticity of substitution between different fuels is not a function of the quantities of nonfuel inputs used. This may be an unrealistic assumption in some cases. Also, the Divisia index assumes that the substitution possibilities among all fuel types and output are equal.

Another limit on the use of prices is that they generally do not exist for wastes. Thus, an economic index of waste flows is impossible to construct.

It is well-known that energy prices do not reflect their full social cost due to a number of market imperfections. This is particularly true for the environmental impact caused by their extraction and use. These problems lead some to doubt the usefulness of price as the basis for any indicator of sustainability. However, with or without externalities, prices should reflect productivities. Internalizing externalities will shift energy use, which in turn will change marginal products.

Moreover, prices produce a ranking of fuels (Table I) that is consistent with our intuition and with previous empirical research. One can conclude that government policy, regulations, cartels, and externalities explain some of the price differentials among fuels but certainly not the substantial ranges that exist. More fundamentally, price differentials are explained by differences in attributes such as physical scarcity, capacity to do useful work, energy density, cleanliness, amenability to storage, safety, flexibility of use, and cost of conversion. Eliminate the market imperfections and the price per Btu of different energies would vary due to the different combinations of attributes that determine their economic usefulness. The different prices per Btu indicate that users are interested in attributes other than heat content.

3. ALTERNATIVE APPROACHES TO ENERGY AGGREGATION

Although we argue that the more advanced economic indexing methods, such as Divisia aggregation, are the most appropriate way to aggregate energy use for investigating its role in the economy, the ecological economics literature proposes other methods of aggregation. We review two of these methods in this section and assess limits on their ability to aggregate energy use.

3.1 Exergy

Other scientists propose a system of aggregating energy and materials based on exergy. Exergy measures the useful work obtainable from an energy source or material, and it is based on the chemical energy embodied in the material or energy based on its physical organization relative to a reference state. Thus, exergy measures the degree to which a material is organized relative a random assemblage of material found at an average concentration in the crust, ocean, or atmosphere. The higher the degree of concentration, the higher the exergy content. The physical units for exergy are the same as for energy or heat, namely kilocalories, joules, Btus, etc. For fossil fuels, exergy is nearly equivalent to the standard heat of combustion; for other materials, specific calculations are needed that depend on the details of the assumed conversion process.

Proponents argue that exergy has a number of useful attributes for aggregating heterogeneous energy and materials. Exergy is a property of all energy and materials and in principle can be calculated from information in handbooks of chemistry and physics and secondary studies. Thus, exergy can be used to measure and aggregate natural resource inputs as well as wastes. For these reasons, Ayres argues that exergy forms the basis for a comprehensive resource accounting framework that could "provide policy makers with a valuable set of indicators." One such indicator is a general measure of "technical efficiency," the efficiency with which "raw" exergy from animals or an inanimate source is converted into final services. A low exergy efficiency implies potential for efficiency gains for converting energy and materials into goods and services. Similarly, the ratio of exergy embodied in material wastes to exergy embodied in resource inputs is the most general measure of pollution. Some also argue that the exergy of waste streams is a proxy for their potential ecotoxicity or harm to the environment, at least in general terms.

From an accounting perspective, exergy is appealing because it is based on the science and laws of thermodynamics and thus has a well-established system of concepts, rules, and information that are available widely. However, like enthalpy, exergy should not be used to aggregate energy and material inputs aggregation because it is one-dimensional. Like enthalpy, exergy does not vary with, and hence does not necessarily reflect, attributes of fuels that determine their economic usefulness, such as energy density, cleanliness, and cost of conversion. The same is true for materials. Exergy cannot explain, for

example, impact resistance, heat resistance, corrosion resistance, stiffness, space maintenance, conductivity, strength, ductility, or other properties of metals that determine their usefulness. Like prices, exergy does not reflect all the environmental costs of fuel use. The exergy of coal, for example, does not reflect coal's contribution to global warming or its impact on human health relative to natural gas. The exergy of wastes is at best a rough first-order approximation of environmental impact because it does not vary with the specific attributes of a waste material and its receiving environment that cause harm to organisms or that disrupt biogeochemical cycles. In theory, exergy can be calculated for any energy or material, but in practice the task of assessing the hundreds (thousands?) of primary and intermediate energy and material flows in an economy is daunting.

3.2 Emergy

The ecologist Howard Odum analyzes energy and materials with a system that traces their flows within and between society and the environment. It is important to differentiate between two aspects of Odum's contribution. The first is his development of a biophysically based, systems-oriented model of the relationship between society and the environment. Here, Odum's early contributions helped lay the foundation for the biophysical analysis of energy and material flows, an area of research that forms part of the intellectual backbone of ecological economics. The insight from this part of Odum's work is illustrated by the fact that ideas he emphasized—energy and material flows, feedbacks, hierarchies, thresholds, and time lags—are key concepts of the analysis of sustainability in a variety of disciplines.

The second aspect of Odum's work, which we are concerned with here, is a specific empirical issue: the identification, measurement, and aggregation of energy and material inputs to the economy, and their use in the construction of indicators of sustainability. Odum measures, values, and aggregates energy of different types by their transformities. Transformities are calculated as the amount of one type of energy required to produce a heat equivalent of another type of energy. To account for the difference in quality of thermal equivalents among different energies, all energy costs are measured in solar emjoules, the quantity of solar energy used to produce another type of energy. Fuels and materials with higher transformities require larger amounts of sunlight to produce and therefore are considered more economically useful.

Several aspects of the emergy methodology reduce its usefulness as a method for aggregating energy and/or material flows. First, like enthalpy and exergy, emergy is one-dimensional because energy sources are evaluated based on the quantity of embodied solar energy and crustal heat. However, is the usefulness of a fuel as an input to production related to its transformity? Probably not. Users value coal based on it heat content, sulfur content, cost of transportation, and other factors that form the complex set of attributes that determine its usefulness relative to other fuels. It is difficult to imagine how this set of attributes is in general related to, much less determined by, the amount of solar energy required to produce coal. Second, the emergy methodology is inconsistent with its own basic tenant, namely that quality varies with embodied energy or emergy. Coal deposits that we currently extract were laid down over many geological periods that span half a billion years. Coals thus have vastly different embodied emergy, but only a single transformity for coal is normally used. Third, the emergy methodology depends on plausible but arbitrary choices of conversion technologies (e.g., boiler efficiencies) that assume users choose one fuel relative to another and other fuels based principally on their relative conversion efficiencies in a particular application. Finally, the emergy methodology relies on long series of calculations with data that vary in quality. However, little attention is paid to the sensitivity of the results to data quality and uncertainty, leaving the reader with little or no sense of the precision or reliability of the emergy calculations.

4. CASE STUDY 1: NET ENERGY FROM FOSSIL FUEL EXTRACTION IN THE UNITED STATES

One technique for evaluating the productivity of energy systems is net energy analysis, which compares the quantity of energy delivered to society by an energy system to the energy used directly and indirectly in the delivery process. EROI is the ratio of energy delivered to energy costs. There is a long debate about the relative strengths and weaknesses of net energy analysis. One restriction on net energy analysis' ability to deliver the insights it promises is its treatment of energy quality. In most net energy

analyses, inputs and outputs of different types of energy are aggregated by their thermal equivalents. This case study illustrates how accounting for energy quality affected calculations for the EROI of the U.S. petroleum sector from 1954 to 1992.

4.1 Methods and Data

Following the definitions in Eq. (2), a quality-corrected EROI* is defined by

$$\mathrm{EROI}_t^* = \frac{\sum_{i=1}^{n} \lambda_{i,t} E_{i,t}^{\mathrm{o}}}{\sum_{i=1}^{n} \lambda_{i,t} E_{i,t}^{\mathrm{c}}}, \qquad (6)$$

where $\lambda_{i,t}$ is the quality factor for fuel type i at time t and E^{o} and E^{c} are the thermal equivalents of energy outputs and energy inputs, respectively. We construct Divisia indices for energy inputs and outputs to account for energy quality in the numerator and denominator. The prices for energy outputs (oil, natural gas, and natural gas liquids) and energy inputs (natural gas, gasoline, distillate fuels, coal, and electricity) are the prices paid by industrial end users for each energy type.

Energy inputs include only industrial energies: the fossil fuel and electricity used directly and indirectly to extract petroleum. The costs include only those energies used to locate and extract oil and natural gas and prepare them for shipment from the wellhead. Transportation and refining costs are excluded from this analysis. Output in the petroleum industry is the sum of the marketed production of crude oil, natural gas, and natural gas liquids.

The direct energy cost of petroleum is the fuel and electricity used in oil and gas fields. Indirect energy costs include the energy used to produce material inputs and to produce and maintain the capital used to extract petroleum. The indirect energy cost of materials and capital is calculated from data for the dollar cost of those inputs to petroleum extraction processes. Energy cost of capital and materials is defined as the dollar cost of capital depreciation and materials multiplied by the energy intensity of capital and materials (Btu/$). The energy intensity of capital and materials is measured by the quantity of energy used to produce a dollar's worth of output in the industrial sector of the U.S. economy. That quantity is the ratio of fossil fuel and electricity use to real GDP produced by industry.

4.2 Results and Conclusions

The thermal equivalent and Divisia EROI for petroleum extraction show significant differences

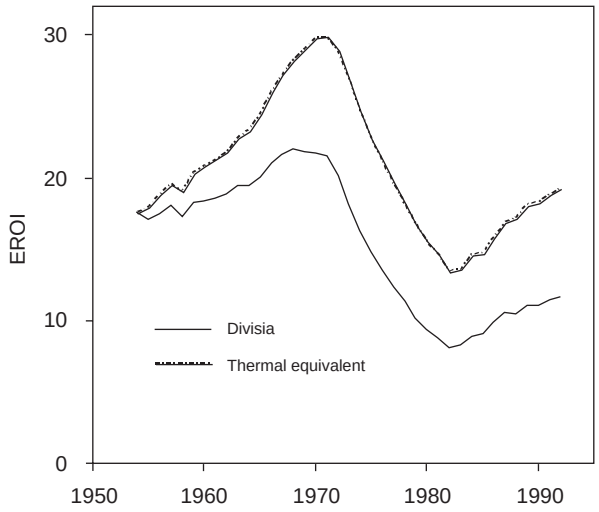


FIGURE 2 Energy return on investment (EROI) for petroleum extraction in the United States, with energy inputs and outputs measured in heat equivalents and a Divisia index.

(Fig. 2). The quality-corrected EROI declines faster than the thermal-equivalent EROI. The thermal-equivalent EROI increased by 60% relative to the Divisia EROI between 1954 and 1992. This difference was driven largely by changes in the mix of fuel qualities in energy inputs. Electricity, the highest quality fuel, is an energy input but not an energy output. Its share of total energy use increased from 2 to 12% during the period; its cost share increased from 20 to 30%. Thus, in absolute terms the denominator in the Divisia EROI is weighted more heavily than in the thermal-equivalent EROI. The Divisia-weighted quantity of refined oil products is larger than that for gas and coal. Thus, the two highest quality fuels, electricity and refined oil products, comprise a large and growing fraction of the denominator in the Divisia EROI compared to the thermal-equivalent EROI. Therefore, the Divisia denominator increases faster than the heat-equivalent denominator, causing EROI to decline faster in the former case.

5. CASE STUDY 2: CAUSALITY IN THE ENERGY/GDP RELATIONSHIP

One of the most important questions about the environment–economy relationship regards the strength of the linkage between economic growth and energy use. With a few exceptions, most analyses ignore the effect of energy quality in the assessment

of this relationship. One statistical approach to address this question is Granger causality and/or cointegration analysis. Granger causality tests whether (i) one variable in a relation can be meaningfully described as a dependent variable and the other variable as an independent variable, (ii) the relation is bidirectional, or (iii) no meaningful relation exists. This is usually done by testing whether lagged values of one of the variables add significant explanatory power to a model that already includes lagged values of the dependent variable and perhaps also lagged values of other variables.

Although Granger causality can be applied to both stationary and integrated time series (time series that follow a random walk), cointegration applies only to linear models of integrated time series. The irregular trend in integrated series is known as a stochastic trend, as opposed to a simple linear deterministic time trend. Time series of GDP and energy use are usually integrated. Cointegration analysis aims to uncover causal relations among variables by determining if the stochastic trends in a group of variables are shared by the series so that the total number of unique trends is less than the number of variables. It can also be used to test if there are residual stochastic trends that are not shared by any other variables. This may be an indication that important variables have been omitted from the regression model or that the variable with the residual trend does not have long-term interactions with the other variables.

Either of these conclusions could be true should there be no cointegration. The presence of cointegration can also be interpreted as the presence of a long-term equilibrium relationship between the variables in question. The parameters of an estimated cointegrating relation are called the cointegrating vector. In multivariate models, there may be more than one such cointegrating vector.

5.1 Granger Causality and the Energy GDP Relation

A series of analyses use statistical tests to evaluate whether energy use or energy prices determine economic growth, or whether the level of output in the United States and other economics determine energy use or energy prices. Generally, the results are inconclusive. Where significant results are obtained, they indicate causality from output to energy use.

One analysis tests U.S. data (1947–1990) for Granger causality in a multivariate setting using a vector autoregression model of GDP, energy use, capital, and labor inputs. The study measures energy use by its thermal equivalents and the Divisia aggregation method discussed previously. The relation among GDP, the thermal equivalent of energy use, and the Divisia energy use indicates that there is less "decoupling" between GDP and energy use when the aggregate measure for energy use accounts for qualitative differences (Fig. 3). The multivariate methodology is important because changes in energy use are frequently countered by substitution with labor and/or capital and thereby mitigate the effect of changes in energy use on output. Weighting energy use for changes in the composition of the energy input is important because a large portion of the growth effects of energy is due to substitution of higher quality energy sources such as electricity for lower quality energy sources such as coal (Fig. 1).

Bivariate tests of Granger causality show no causal order in the relation between energy and GDP in either direction, regardless of the measure used to qualify energy use (Table III). In the multivariate model with energy measured in primary Btus, GDP was found to "Granger cause" energy use. However, when both innovations—a multivariate model and energy use adjusted for quality—are employed, energy Granger causes GDP. These results show that adjusting energy for quality is important, as is considering the context within which energy use is occurring. The conclusion that energy use plays an important role in determining the level of economic

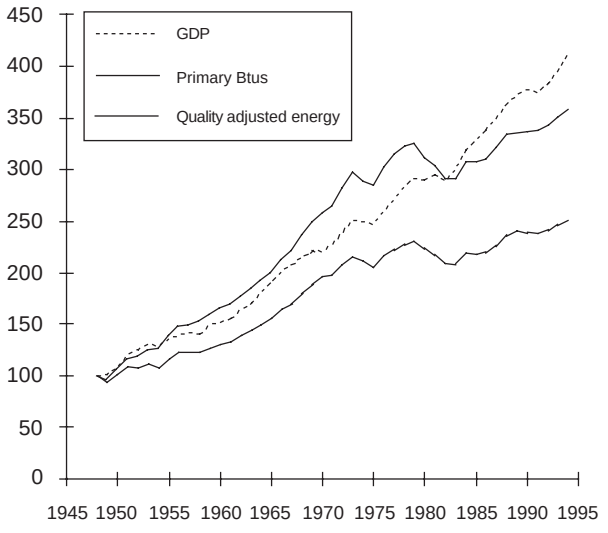


FIGURE 3 Energy use and GDP in the United States, with energy use measured in heat equivalents and a Divisia index. From Stern (1993).

TABLE III

Energy GDP Causality Tests for the United States, 1947–1990[a]

	Bivariate model		Multivariate model	
	Primary Btus	Quality-adjusted energy	Primary Btus	Quality-adjusted energy
Energy causes GDP	0.8328	0.9657	0.5850	3.1902
	0.4428	0.4402	0.5628	0.3188E−01
GDP causes Energy	0.3421	0.7154	9.0908	0.8458
	0.7125	0.5878	0.7163E−03	0.5106

[a] The test statistic is an *F* statistic. Significance levels in italics. A significant statistic indicates that there is Granger causality in the direction indicated.

activity is consistent with results of price-based studies of other energy economists.

6. CASE STUDY 3: THE DETERMINANTS OF THE ENERGY/GDP RELATIONSHIP

One of the most widely cited macroeconomic indicators of sustainability is the ratio of total energy use to total economic activity, or the energy/real GDP ratio (E/GDP ratio). This ratio has declined since 1950 in many industrial nations. There is controversy regarding the interpretation of this decline. Many economists and energy analysts argue that the decline indicates that the relation between energy use and economic activity is relatively weak. This interpretation is disputed by many biophysical economists, who argue that the decline in the E/GDP ratio overstates the ability to decouple energy use and economic activity because many analyses of the E/GDP ratio ignore the effect of changes in energy quality (Fig. 1).

The effect of changes in energy quality (and changes in energy prices and types of goods and services produced and consumed) on the E/GDP ratio can be estimated using Eq. (7):

$$\frac{E}{\text{GDP}} = \alpha + \beta_1 \ln\left(\frac{\text{natural gas}}{\text{E}}\right) + \beta_2 \ln\left(\frac{\text{oil}}{\text{E}}\right) + \beta_3 \ln\left(\frac{\text{primary electricity}}{\text{E}}\right) + \beta_4\left(\frac{\text{PCE}}{\text{GDP}}\right) + \beta_5(\text{product mix}) + \beta_6 \ln(\text{price}) + \varepsilon, \quad (7)$$

where E is the total primary energy consumption (measured in heat units); GDP is real GDP; primary electricity is electricity generated from hydro, nuclear, solar, or geothermal sources; PCE is real personal consumption expenditures spent directly on energy by households; product mix measures the fraction of GDP that originates in energy-intensive sectors (e.g., chemicals) or non-energy-intensive sectors (e.g., services); and price is a measure of real energy prices.

The effect of energy quality on the E/GDP ratio is measured by the fraction of total energy consumption from individual fuels. The sign on the regression coefficients β_1–β_3 is expected to be negative because natural gas, oil, and primary electricity can do more useful work (and therefore generate more economic output) per heat unit than coal. The rate at which an increase in the use of natural gas, oil, or primary electricity reduces the E/GDP ratio is not constant. Engineering studies indicate that the efficiency with which energies of different types are converted to useful work depends on their use. Petroleum can provide more motive power per heat unit of coal, but this advantage nearly disappears if petroleum is used as a source of heat. From an economic perspective, the law of diminishing returns implies that the first uses of high-quality energies are directed at tasks that are best able to make use of the physical, technical, and economic aspects of an energy type that combine to determine its high-quality status. As the use of a high-quality energy source expands, it is used for tasks that are less able to make use of the attributes that confer high quality. The combination of physical differences in the use of energy and the economic ordering in which they are applied to these tasks implies that the amount of economic activity generated per heat unit diminishes as the use of a high-quality energy expands. Diminishing returns on energy quality is imposed on the model by specifying the fraction of the energy budget from petroleum, primary electricity, natural gas, or oil in natural

logarithms. This specification ensures that the first uses of high-quality energies decrease the energy/GDP ratio faster than the last uses.

The regression results indicate that Eq. (7) can be used to account for most of the variation in the E/GDP ratio for France, Germany, Japan, and the United Kingdom during the post-World War II period and in the United States since 1929. All the variables have the sign expected by economic theory and are statistically significant, and the error terms have the properties assumed by the estimation technique.

Analysis of regression results indicate that changes in energy mix can account for a significant portion of the downward trend in E/GDP ratios. The change from coal to petroleum and petroleum to primary electricity is associated with a general decline in the E/GDP ratio in France, Germany, the United Kingdom, and the United States during the post-World War II period (Fig. 4). The fraction of total energy consumption supplied by petroleum increased steadily for each nation through the early 1970s. After the first oil shock, the fraction of total energy use from petroleum remained steady or declined slightly in these four nations. However, energy mix continued to reduce the E/real GDP ratio after the first oil shock because the fraction of total energy use from primary electricity increased steadily. The effect of changes in energy mix on the E/GDP ratio shows no trend over time in Japan, where the fraction of total energy consumption supplied by primary electricity declined through the early 1970s and increased steadily thereafter. This U shape offsets the steady increase in the fraction of total energy use from petroleum that occurred prior to 1973.

These regression results indicate that the historical reduction in the E/GDP ratio is associated with shifts in the types of energies used and the types of goods and services consumed and produced. Diminishing returns to high-quality energies and the continued consumption of goods from energy-intensive sectors such as manufacturing imply that the ability of changes in the composition of inputs and outputs to reduce the E/real GDP ratio further is limited.

7. CONCLUSIONS AND IMPLICATIONS

Application of the Divisia index to energy use in the U.S. economy illustrates the importance of energy quality in aggregate analysis. The quality-corrected index for EROI indicates that the energy surplus delivered by petroleum extraction in the United States is smaller than indicated by unadjusted EROI. The trend over time in a quality-adjusted index of total primary energy use in the U.S. economy is significantly different, and declines faster, than the standard heat-equivalent index. Analysis of Granger causality and cointegration indicates a causal relationship running from quality-adjusted energy to GDP but not from the unadjusted energy index. The econometric analysis of the E/real GDP ratio indicates that the decline in industrial economies has been driven in part by the shift from coal to oil, gas, and primary electricity. Together, these results suggest that accounting for energy quality reveals a relatively strong relationship between energy use and economic output. This runs counter to much of the conventional wisdom that technical improvements and structural change have decoupled energy use from economic performance. To a large degree, technical change and substitution have increased the use of higher quality energy and reduced the use of lower quality energy. In economic terms, this means that technical change has been "embodied" in the fuels and their associated energy converters. These changes have increased energy efficiency in energy extraction processes, allowed an apparent decoupling between energy use and economic output, and increased energy efficiency in the production of output.

The manner in which these improvements have been largely achieved should give pause for thought. If decoupling is largely illusory, any increase in the cost of producing high-quality energy vectors could have important economic impacts. Such an increase might occur if use of low-cost coal to generate electricity is restricted on environmental grounds, particularly climate change. If the substitution process cannot continue, further reductions in the E/GDP ratio would slow. Three factors might limit future substitution to higher quality energy. First, there are limits to the substitution process. Eventually, all energy used would be of the highest quality variety—electricity—and no further substitution could occur. Future discovery of a higher quality energy source might mitigate this situation, but it would be unwise to rely on the discovery of new physical principles. Second, because different energy sources are not perfect substitutes, the substitution process could have economic limits that will prevent full substitution. For example, it is difficult to imagine an airliner running on electricity. Third, it is likely that supplies of petroleum, which is of higher quality than coal, will decline fairly early in the 21st century.

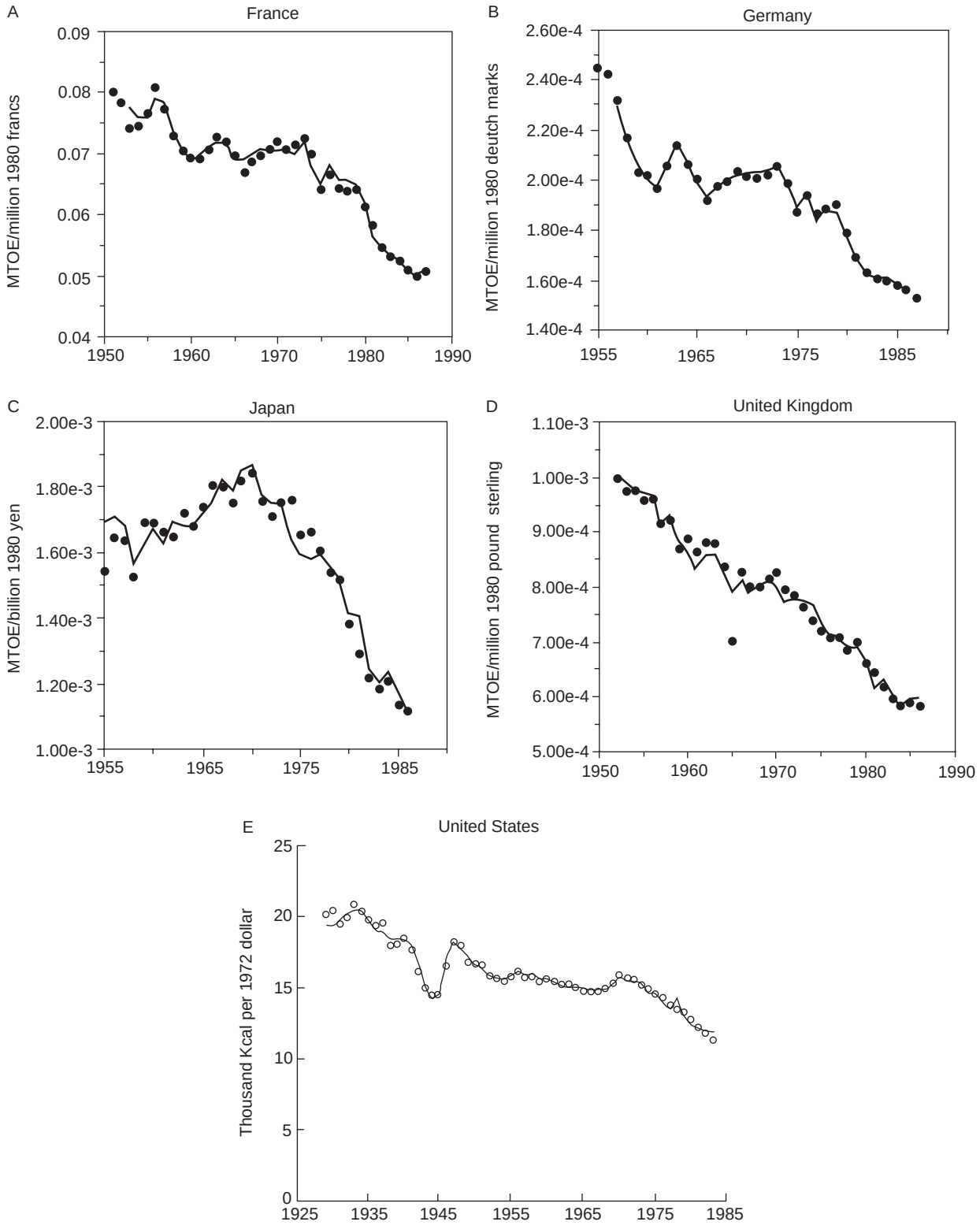


FIGURE 4 The results of a regression analysis (Eq. (7) that predicts the E/GDP ratio as a function of fuel mix and other variables. •, actual values; solid lines, values predicted from the regression equation; ○, actual values for the United States. From Kaufmann (1992) and Hall *et al*. (1986).

Finally, our conclusions do not imply that one-dimensional and/or physical indicators are universally inferior to the economic indexing approach we endorse. As one reviewer noted, ecologists might raise the problem of Leibig's law of the minimum, in which the growth or sustainability of a system are constrained by the single critical element in least supply. Exergy or mass are appropriate if the object of analysis is a single energy or material flux. Physical units are also necessary to valuate those flows. Integrated assessment of a material cycle within and between the environment and the economy is logical based on physical stocks and flows. However, when the question being asked requires the aggregation of energy flows in an economic system, an economic approach such as Divisa aggregation or a direct measure of marginal product embody a more tenable set of assumptions than does aggregation by one-dimensional approaches.

SEE ALSO THE FOLLOWING ARTICLES

Economic Thought, History of Energy in • Emergy Analysis and Environmental Accounting • Entropy and the Economic Process • Exergy • Exergy Analysis of Energy Systems • Exergy: Reference States and Balance Conditions • Net Energy Analysis: Concepts and Methods • Thermodynamics and Economics, Overview

Further Reading

Ayres, R., and Martiñas, K. (1995). Waste potential entropy: The ultimate ecotoxic? *Econ. Appliqueé* **48,** 95–120.

Berndt, E. (1990). Energy use, technical progress and productivity growth: A survey of economic issues. *J. Productivity Anal.* **2,** 67–83.

Cleveland, C. J. (1992). Energy quality and energy surplus in the extraction of fossil fuels in the U.S. *Ecol. Econ.* **6,** 139–162.

Cleveland, C. J., Costanza, R., Hall, C. A. S., and Kaufmann, R. (1984). Energy and the U.S. economy: A biophysical perspective. *Science* **255,** 890–897.

Cottrell, W. F. (1955). "Energy and Society." McGraw-Hill, New York.

Darwin, R. F. (1992). Natural resources and the Marshallian effects of input-reducing technological changes. *J. Environ. Econ. Environ. Management* **23,** 201–215.

Gever, J., Kaufmann, R., Skole, D., and Vorosmarty, C. (1986). "Beyond Oil: The Threat to Food and Fuel in the Coming Decades." Ballinger, Cambridge, UK.

Hall, C. A. S., Cleveland, C. J., and Kaufmann, R. K. (1986). "Energy and Resource Quality: The Ecology of the Economic Process." Wiley Interscience, New York.

Hamilton, J. D. (1983). Oil and the macroeconomy since World War II. *J. Political Econ.* **91,** 228–248.

Kaufmann, R. K. (1992). A biophysical analysis of the energy/real GDP ratio: Implications for substitution and technical change. *Ecol. Econ.* **6,** 35–56.

Odum, H. T. (1996). "Environmental Accounting." Wiley, New York.

Rosenberg, N. (1998). The role of electricity in industrial development. *Energy J.* **19,** 7–24.

Schurr, S., and Netschert, B. (1960). "Energy and the American Economy, 1850–1975." Johns Hopkins Univ. Press, Baltimore.

Stern, D. I. (1993). Energy use and economic growth in the USA: A multivariate approach. *Energy Econ.* **15,** 137–150.

Aircraft and Energy Use

JOOSUNG J. LEE, STEPHEN P. LUKACHKO, and IAN A. WAITZ
Massachusetts Institute of Technology
Cambridge, Massachusetts, United States

1. **Introduction**
2. **Economic Growth, Demand, and Energy Use**
3. **Energy Use, Emissions, and Environmental Impact**
4. **Trends in Energy Use**
5. **Energy Consumption in an Aircraft System**
6. **Historical Trends in Technological and Operational Performance**
7. **Technological and Operational Outlook for Reduced Energy Use in Large Commercial Aircraft**
8. **Industry Characteristics, Economic Impacts, and Barriers to Technology Uptake**

Glossary

bypass ratio The ratio of air passed through the fan system to that passed through the engine core.
contrail The condensation trail that forms when moist, high-temperature air in a jet exhaust, as it mixes with ambient cold air, condenses into particles in the atmosphere and saturation occurs.
drag The aerodynamic force on an aircraft body; acts against the direction of aircraft motion.
energy intensity (E_I) A measure of aircraft fuel economy on a passenger-kilometer basis; denoted by energy used per unit of mobility provided (e.g., fuel consumption per passenger-kilometer)
energy use (E_U) A measure of aircraft fuel economy on a seat-kilometer basis (e.g., fuel consumption per seat-kilometer).
great circle distance The minimum distance between two points on the surface of a sphere.
hub-and-spoke system Feeding smaller capacity flights into a central hub where passengers connect with flights on larger aircraft that then fly to the final destination.
lift-to-drag ratio (**L/D**) A measure of aerodynamic efficiency; the ratio of lift force generated to drag experienced by the aircraft.
load factor The fraction of passengers per available seats.
radiative forcing A measure of the change in Earth's radiative balance associated with atmospheric changes; positive forcing indicates a net warming tendency relative to preindustrial times.
structural efficiency (**OEW/MTOW**) The ratio of aircraft operating empty weight (OEW) to maximum takeoff weight (MTOW); a measure of the weight of the aircraft structure relative to the weight it can carry (combined weights of structure plus payload plus fuel).
thrust A force that is produced by engines and propels the aircraft.
thrust specific fuel consumption (**SFC**) A measure of engine efficiency as denoted by the rate of fuel consumption per unit thrust (e.g., kilograms/second/Newton).
turbofan engine The dominant mode of propulsion for commercial aircraft today; a turbofan engine derives its thrust primarily by passing air through a large fan system driven by the engine core.

An aircraft is composed of systems that convert fuel energy to mechanical energy in order to perform work—the movement of people and cargo. This article describes how aircraft technology and operations relate to energy use. Historical trends and future outlook for aircraft performance, energy use, and environmental impacts are discussed. Economic characteristics of aircraft systems as they relate to energy use are also presented.

1. INTRODUCTION

The first powered passenger aircraft were developed at the turn of the 20th century. Since then, there has been rapid growth in aviation as a form of mobility and consequently significant growth in energy use. In 2002, aviation accounted for 3 trillion revenue passenger-kilometers (RPKs), or approximately 10% of world RPKs traveled on all transportation modes and 40% of the value of world freight shipments. Among all modes of transport, demand for air travel has grown fastest. If, as expected, strong growth in

 © 2004 Elsevier Inc. All rights reserved.

air travel demand continues, aviation will become the dominant mode of transportation, perhaps surpassing the mobility provided by automobiles within a century. This evolution of transportation demand also suggests an increase in per-person energy use for transportation. Minimizing energy use has always been a fundamental design goal for commercial aircraft. However, the growth of air transportation renders ever-increasing pressures for improvements in technology and operational efficiency to limit environmental impacts.

In the analysis presented here, trends in aviation transportation demand, energy use, and associated environmental impacts are examined (Sections 2–4). In Sections 5 and 6, aircraft systems from an energy conversion perspective are introduced and key performance parameters of aircraft technology and operation are discussed. A technology and operational outlook for reduced aircraft energy use is presented in Section 7, followed by a summary of industry characteristics and economic impacts that affect energy use of individual aircraft and the fleet as a whole in Section 8.

2. ECONOMIC GROWTH, DEMAND, AND ENERGY USE

On a per capita basis, rising demand for mobility is well correlated with growth in gross domestic product (GDP)—a measure of national economic activity—across a wide variety of economic, social, and geographic settings. One reason for this may be found in the roughly constant shares of income and time people dedicate to transportation. A fixed budget for travel leads to an increase in total travel demand per capita (e.g., RPK per capita) in approximate proportion to income. In addition, a person spends an average of 1.0–1.5 hours/day traveling. One key implication of such invariance is that as demand for movement increases, travelers tend to shift toward faster modes of transportation. Consequently, continuing growth in world population and income levels can be expected to lead to further demand for air travel, both in terms of market share and RPKs. As a result, high-speed transportation, in which aviation is anticipated to be the primary provider, will play an increasingly important role and may account for slightly more than one-third of world passenger traffic volume within the next 50 years. In general, among industry and government predictions, growth in passenger air transportation has been typically projected to be between 3 and 6%/year as an average over future periods of 10–50 years.

Aviation fuel consumption today corresponds to 2–3% of the total fossil fuel use worldwide, more than 80% of which is used by civil aviation operation. Energy use in the production of aircraft is relatively minor in comparison to that consumed in their operation. Although the majority of air transportation demand is supplied by large commercial aircraft, defined as those aircraft with a seating capacity of 100 or more, smaller regional aircraft have emerged as an important component of both demand and energy use within air transportation. For example, in the United States, although regional aircraft currently perform under 4% of domestic RPKs, they account for almost 7% of jet fuel use and for 40–50% of total departures. Future growth in demand for regional aircraft RPKs could be up to double the rate for large commercial aircraft. Cargo operations account for some 10% of total revenue ton-kilometers and fuel use within the aviation sector. Economic activity, as measured by world GDP, is the primary driver for the air cargo industry growth. World air cargo traffic is expected to grow at an average annual rate of over 6% for the next decade.

3. ENERGY USE, EMISSIONS, AND ENVIRONMENTAL IMPACT

The growth in air transportation volume has important global environmental impacts associated with the potential for climate change. On local to regional scales, noise, decreased air quality related primarily to ozone production and particulate levels, and other issues, such as roadway congestion related to airport services and local water quality, are all recognized as important impacts. In this section, the focus is on emissions-related impacts; because of its relative importance, some additional detail on the aviation role in climate change is provided.

The total mass of emissions from an aircraft is directly related to the amount of fuel consumed. Of the exhaust emitted from the engine core, 7–8% is composed of carbon dioxide (CO_2) and water vapor (H_2O); another 0.5% composed of nitrogen oxides (NO_x), unburned hydrocarbons (HC), carbon monoxide (CO), and sulfur oxides (SO_x); there are other trace chemical species that include the hydroxy family (HO_x) and the extended family of nitrogen

compounds (NO_y), and soot particulates. Elemental species such as O, H, and N are also formed to an extent governed by the combustion temperature. The balance (91.5–92.5%) is composed of O_2 and N_2.

Emissions of CO_2 and H_2O are products of hydrocarbon fuel combustion and are thus directly related to the aircraft fuel consumption, which in turn is a function of aircraft weight, aerodynamic design, engine design, and the manner in which the aircraft is operated. Emissions of NO_x, soot, CO, HC, and SO_x are further related to details of the combustor design and, to some extent, to postcombustion chemical reactions occurring within the engine. These emissions are thus primarily controlled by the engine design, but total emissions can be reduced through improvements in fuel efficiency. Such emissions are therefore typically quoted relative to the total amount of fuel burned as an emission index (e.g., grams of NO_x/kilogram of fuel). A host of minor constituents exist in very small, trace amounts.

The climate effects of aviation are perhaps the most important of the environmental impacts, both in terms of economic cost and the extent to which all aspects of the aviation system, operations, and technology determine the impact. Because a majority of aircraft emissions are injected into the upper troposphere and lower stratosphere (typically 9–13 km in altitude), resulting impacts on the global environment are unique among all industrial activities. The fraction of aircraft emissions that is relevant to atmospheric processes extends beyond the radiative forcing effects of CO_2. The mixture of exhaust species discharged from aircraft perturbs radiative forcing two to three times more than if the exhaust was CO_2 alone. In contrast, the overall radiative forcing from the sum of all anthropogenic activities is estimated to be a factor of 1.5 times CO_2 alone. Thus the impact of burning fossil fuels at altitude is approximately double that due to burning the same fuels at ground level. The enhanced forcing from aircraft compared with ground-based sources is due to different physical (e.g., contrails) and chemical (e.g., ozone formation/destruction) effects resulting from altered concentrations of participating chemical species and changed atmospheric conditions. However, many of the chemical and physical processes associated with climate impacts are the same as those that determine air quality in the lower troposphere.

Estimates of the radiative forcing by various aircraft emissions for 1992 offered by the Intergovernmental Panel on Climate Change (IPCC) and the 1999 projections from Penner *et al.* for the year 2050 are shown in Fig. 1. The estimates translate to 3.5% of the total anthropogenic forcing that occurred in 1992 and to an estimated 5% by 2050 for an all-subsonic fleet. Associated increases in ozone levels are expected to decrease the amount of ultraviolet radiation at the surface of the earth. Future fleet composition also impacts the radiative forcing estimate. A supersonic aircraft flying at 17–20 km would have a radiative forcing five times greater than a subsonic equivalent in the 9- to 13-km range. It is important to note that these estimates are of an uncertain nature. Although broadly consistent with

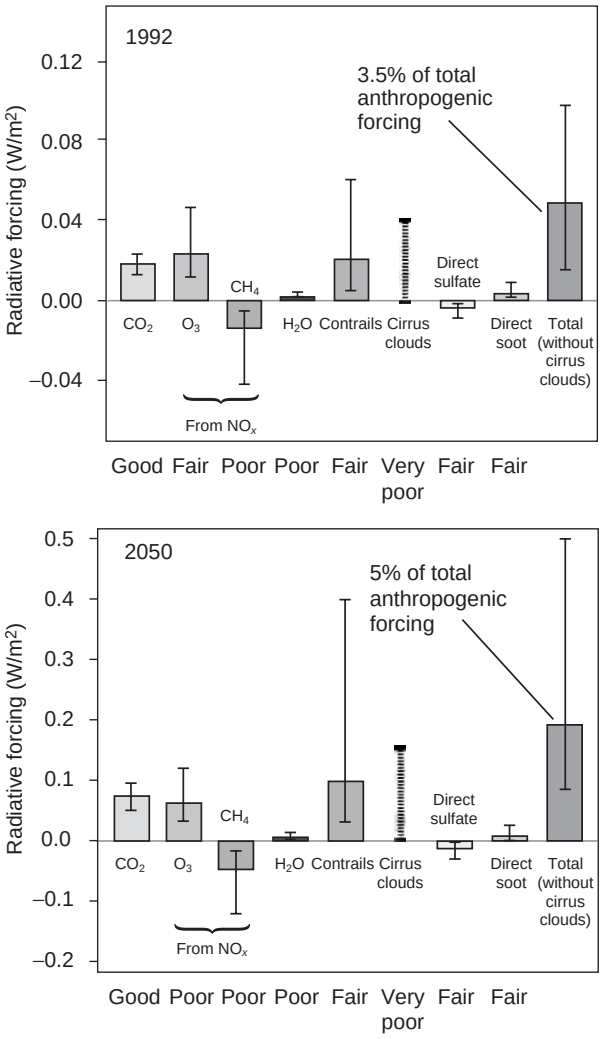


FIGURE 1 Radiative forcing estimated for 1992 (0.05 W/m^2 total) and projected to 2050 (0.19 W/m^2 total). Note differences in scale. Note also that the dashed bars for aviation-induced cirrus cloudiness describe the range of estimates, not the uncertainty. The level of scientific understanding of this potential impact is very poor and no estimate of uncertainty has been made. Cirrus clouds are not included in the total radiative forcing estimate. Reproduced from Penner *et al.* (1999), with permission.

these IPCC projections, subsequent research reviewed by the Royal Commission on Environmental Protection (RCEP) in the United Kingdom has suggested that the IPCC reference value for the climate impact of aviation is likely to be an underestimate. In particular, although the impact of contrails is probably overestimated in Fig. 1, aviation-induced cirrus clouds could be a significant contributor to positive radiative forcing; NO_x-related methane reduction is less than shown in Fig. 1, reducing the associated cooling effect, and growth of aviation in the period 1992–2000 has continued at a rate larger than that used in the IPCC reference scenario.

4. TRENDS IN ENERGY USE

Fuel efficiency gains due to technological and operational change can mitigate the influence of growth on total emissions. Increased demand has historically outpaced these gains, resulting in an overall increase in emissions over the history of commercial aviation. The figure of merit relative to total energy use and emissions in aviation is the energy intensity (E_I). When discussing energy intensity, the most convenient unit of technology is the system represented by a complete aircraft. In this section, trends in energy use and E_I are elaborated. In the following section, the discussion focuses on the relation of E_I to the technological and operational characteristics of an aircraft.

Reviews of trends in technology and aircraft operations undertaken by Lee *et al.* and Babikian *et al.* indicate that continuation of historical precedents would result in a future decline in E_I for the large commercial aircraft fleet of 1.2–2.2%/year when averaged over the next 25 years, and perhaps an increase in E_I for regional aircraft, because regional jets use larger engines and replace turboprops in the regional fleet. When compared with trends in traffic growth, expected improvements in aircraft technologies and operational measures alone are not likely to offset more than one-third of total emissions growth. Therefore, effects on the global atmosphere are expected to increase in the future in the absence of additional measures. Industry and government projections, which are based on more sophisticated technology and operations forecasting, are in general agreement with the historical trend. Compared with the early 1990s, global aviation fuel consumption and subsequent CO_2 emissions could increase three- to sevenfold by 2050, equivalent to a 1.8–3.2% annual rate of change. In addition to the different demand growth projections entailed in such forecasts, variability in projected emissions also originates from different assumptions about aircraft technology, fleet mix, and operational evolution in air traffic management and scheduling.

Figure 2 shows historical trends in E_I for the U.S. large commercial and regional fleets. Year-to-year variations in E_I for each aircraft type, due to different operating conditions, such as load factor, flight speed, altitude, and routing, controlled by different operators, can be $\pm 30\%$, as represented by the vertical extent of the data symbols (Fig. 2A). For large commercial aircraft, a combination of technological and operational improvements led to a reduction in E_I of the entire U.S. fleet of more than 60% between 1971 and 1998, averaging about 3.3%/year. In contrast, total RPK has grown by 330%, or 5.5%/year over the same period. Long-range aircraft are $\sim 5\%$ more fuel efficient than are short-range aircraft because they carry more passengers over a flight spent primarily at the cruise condition. Regional aircraft are 40–60% less fuel efficient than are their larger narrow- and wide-body counterparts, and regional jets are 10–60% less fuel efficient compared to turboprops. Importantly, fuel efficiency differences between large and regional aircraft can be explained mostly by differences in aircraft operations, not technology.

Reductions in E_I do not always directly imply lower environmental impact. For example, the prevalence of contrails is enhanced by greater engine efficiency. NO_x emissions also become increasingly difficult to limit as engine temperatures and pressures are increased—a common method for improving engine efficiency. These conflicting influences make it difficult to translate the expected changes in overall system performance into air quality impacts. Historical trends suggest that fleet-averaged NO_x emissions per unit thrust during landing and takeoff (LTO) cycles have seen little improvement, and total NO_x emissions have slightly increased. However, HC and CO emissions have been reduced drastically since the 1950s.

5. ENERGY CONSUMPTION IN AN AIRCRAFT SYSTEM

Energy intensity can be related to specific measures of technological and operational efficiency in the air transportation system. The rest of this article takes a more detailed look at trends in these efficiencies and

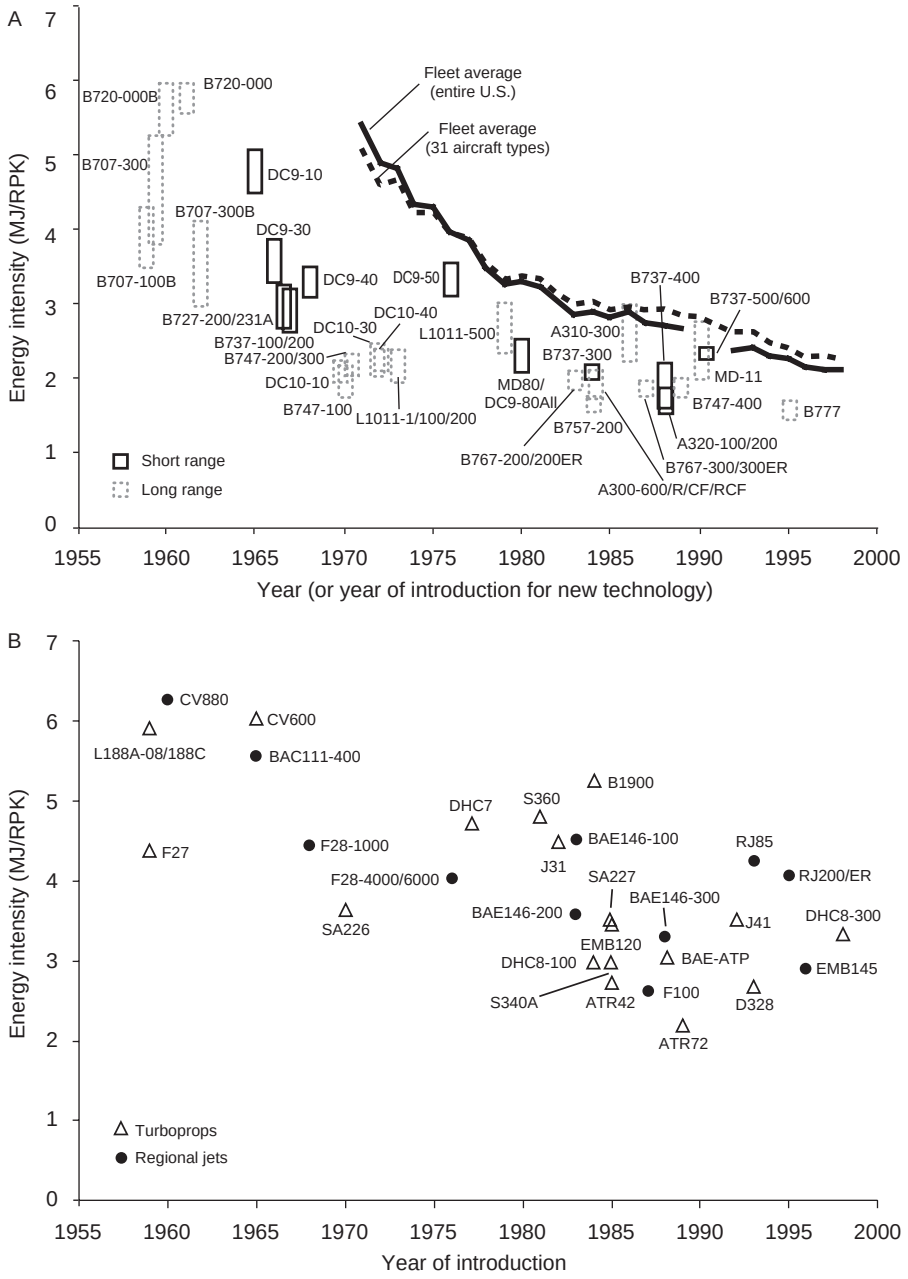


FIGURE 2 (A) Historical trends in energy intensity of the U.S. large commercial fleets. Individual aircraft E_I based on 1991–1998 operational data with the exception of the B707 and B727, which are based on available operational data prior to 1991. Fleet averages were calculated using a revenue passenger-kilometer (RPK) weighting. Data were not available for the entire U.S. fleet average during 1990 and 1991. Reproduced from Lee *et al.* (2001), with permission. (B) Historical trends in energy intensity of the U.S. regional fleets.

options for controlling energy use. The first step is a simplified description of the energy conversion within an aircraft engine. An aircraft engine converts the flow of chemical energy contained in aviation fuel and the air drawn into the engine into power (thrust multiplied by flight speed). Overall engine efficiency is defined by the ratio of power to total fuel energy flow rate. Only one-fourth to one-third of fuel energy is used to overcome drag and thus propel the aircraft. The remaining energy is expelled as waste heat in the engine exhaust. A parameter that is closely related to the overall engine efficiency is the specific fuel consumption (SFC). When judging the efficiency of an aircraft system, however, it is more relevant to

consider work in terms of passengers or payload carried per unit distance. Energy intensity is an appropriate measure when comparing efficiency and environmental impact to other modes. E_I consists of two components—energy use, E_U, and load factor, α, as described by Eq. (1). Energy use is energy consumed by the aircraft per seat per unit distance traversed and is determined by aircraft technology parameters, including engine efficiency. E_U observed in actual aircraft operations reflects operational inefficiencies such as ground delays and airborne holding. The fleet average E_U is of interest because it is the fleet fuel efficiency that determines the total energy use. Load factor is a measure of how efficiently aircraft seats are filled and aircraft kilometers are utilized for revenue-generating purposes. Increasing load factor leads to improved fuel consumption on a passenger-kilometer basis.

$$E_I = \frac{\mathrm{MJ}}{\mathrm{RPK}} = \frac{\mathrm{MJ}}{\mathrm{ASK}} \Big/ \frac{\mathrm{RPK}}{\mathrm{ASK}} = \frac{E_U}{\alpha}, \tag{1}$$

where MJ is megajoules of fuel energy, RPK is revenue passenger-kilometers, ASK is available seat-kilometers, and α is load factor. To show E_I as a function of the engine, aerodynamic, and structural efficiencies of an aircraft system as well as load factor, it is necessary to have a model of aircraft performance. Because a major portion of aircraft operation is spent at cruise, the Breguet range (R) equation, which describes aircraft motion in level, constant-speed flight, is a relevant model. In the Breguet range equation [Eq. (2)], engine thrust is balanced by drag, and lift balances aircraft weight. Propulsion, aerodynamic, and structural characteristics are represented by three parameters: SFC, lift-to-drag ratio (L/D), and structural weight ($W_{\mathrm{structure}}$). Given these technological characteristics as well as other operability parameters, including the amount of payload (W_{payload}) and fuel on board (W_{fuel}), the Breguet range equation can be used to determine maximum range for a level, constant-speed flight.

$$R = \frac{V(\mathrm{L/D})}{g \cdot \mathrm{SFC}} \times \ln\left(1 + \frac{W_{\mathrm{fuel}}}{W_{\mathrm{payload}} + W_{\mathrm{structure}} + W_{\mathrm{reserve}}}\right), \tag{2}$$

where g is the gravitational acceleration constant, W_{reserve} is the reserve fuel, and V is flight speed. By rearranging Eq. (2), a relationship between aircraft energy use and technology parameters can be derived as shown in Eq. (3). As implied by Eq. (3), aircraft system efficiency improves with lower SFC, greater L/D, and lighter structural weight.

$$E_U \equiv \alpha E_I \equiv \frac{1}{\eta_U}$$

$$E_U = \frac{Q W_{\mathrm{fuel}}}{S} \cdot \frac{g \cdot \mathrm{SFC}}{V(\mathrm{L/D})} \times \frac{1}{\ln\left(\dfrac{W_{\mathrm{fuel}}}{W_{\mathrm{payload}} + W_{\mathrm{structure}} + W_{\mathrm{reserve}}}\right)}, \tag{3}$$

where η_U is fuel efficiency (e.g., seat-kilometers/kilogram of fuel consumption), Q is the lower heating value of jet fuel, and S is the number of seats.

6. HISTORICAL TRENDS IN TECHNOLOGICAL AND OPERATIONAL PERFORMANCE

6.1 Technological Performance

As shown in Eq. (3), engine, aerodynamic, and structural efficiencies play an important role in determining the energy intensity of an aircraft. Engine efficiency in large commercial aircraft, as measured by the cruise SFC of newly introduced engines, improved by approximately 40% over the period 1959–1995, averaging an annual 1.5% improvement. Most of this improvement was realized prior to 1970, with the introduction of high-bypass turbofan engines. However, as bypass ratios have increased, engine diameters have also become larger, leading to an increase in engine weight and aerodynamic drag. Other routes to engine efficiency improvement include increasing the peak pressure and temperature within the engine, which is limited by materials and cooling technology, and improving engine component efficiencies. Aerodynamic efficiency in large commercial aircraft has increased by approximately 15% historically, averaging 0.4%/year for the same period. Better wing design and improved propulsion/airframe integration, enabled by improved computational and experimental design tools, have been the primary drivers. Historical improvements in structural efficiency are less evident. One reason is that over the 35-year period between the introduction of the B707 and the B777, large commercial aircraft have been constructed almost exclusively of aluminum and are currently about 90% metallic by weight. Composites are used for a limited number of components. Another reason is that improvements in aircraft structural efficiency have been

largely traded for other technological improvements, such as larger, heavier engines and increased passenger comfort.

6.2 Operational Performance

Infrastructure characteristics also impact efficiency. In particular, delays on the ground and in the air can increase energy intensity. Extra fuel is burned on the ground during various non-flight operations, and hours spent in the air (airborne hours) do not account for more than 75–90% of the total operational hours of the aircraft (block hours). The ratio of airborne to block hours can be treated as ground-time efficiency, η_g. Similarly, non-cruise portions of the flight, poor routing, and delays in the air constitute inefficiencies related to spending fuel during the flight beyond what would be required for a great circle distance trip at constant cruise speed. This inefficiency can be measured by the ratio of minimum flight hours to airborne hours, η_a. Minimum flight hours are calculated with the assumption that all aircraft fly the entire route at Mach 0.80 and at an altitude of 10.7 km (no climbing, descending, or deviation from the minimum distance, the great circle route). Minimum flight hours represent the shortest time required to fly a certain stage length and reveal any extra flight time due to nonideal flight conditions. The product of η_g and η_a gives the flight time efficiency, η_{ft}. Both η_a and η_g increase with stage length. The lower η_{ft} associated with short-range aircraft is related to the more than 40% of block time spent in non-cruise flight segments. Long-range aircraft operate closer to the ideal as total flight time efficiency approaches 0.9. The impact of operational differences on E_U is evident in Fig. 3, which shows the variation of E_U with stage length for turboprop and jet-powered aircraft (both regional and large jets) introduced during and after 1980. Aircraft flying stage lengths below 1000 km have E_U values between 1.5 and 3 times higher compared to aircraft flying stage lengths above 1000 km. Regional aircraft, compared to large aircraft, fly shorter stage lengths, and therefore spend more time at airports taxiing, idling, and maneuvering into gates, and in general spend a greater fraction of their block time in non-optimum, non-cruise stages of flight. Turboprops show a pattern distinct from that of jets and are, on average, more efficient at similar stage lengths. The energy usage also increases gradually for stage lengths above 2000 km because the increasing amount of fuel required for increasingly long stage lengths leads to a heavier aircraft and a higher rate of fuel burn.

Aircraft E_I is also improved through better utilization (e.g., load factor) and greater per-aircraft capacity (e.g., number of seats). Historically, the load factor on domestic and international flights operated by U.S. carriers climbed 15% between 1959 and 1998, all of which occurred after 1970 at an average of 1.1%/year. Figure 4 shows historical load factor evolution for both U.S. large commercial and regional aircraft. Load factor gains have been attributed to deregulation in the United States and global air travel liberalization, both of which contributed to the advent of hub-and-spoke systems. As airlines have sought greater route capacity, the

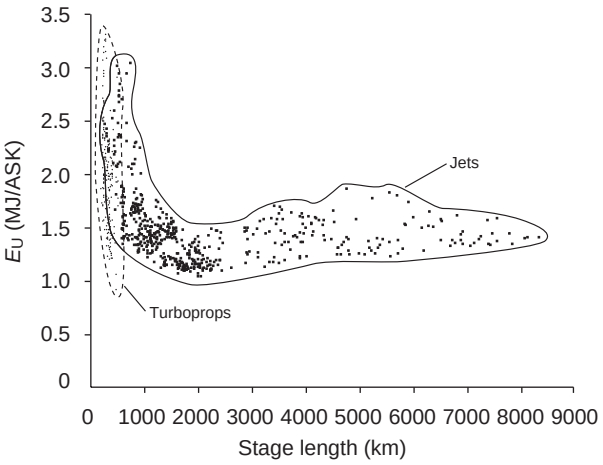


FIGURE 3 Variation of E_U with stage length. ASK, Available seat-kilometers.

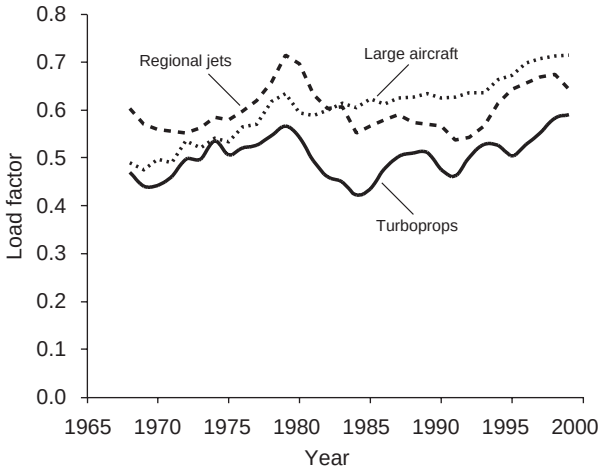


FIGURE 4 Historical trends in load factor for the U.S. large commercial and regional fleets.

average number of seats has also increased, by 35% between 1968 and 1998, or from 108 to 167 seats (an average of 1.4%/year), most of which occurred prior to 1980.

7. TECHNOLOGICAL AND OPERATIONAL OUTLOOK FOR REDUCED ENERGY USE IN LARGE COMMERCIAL AIRCRAFT

The outlook for future reductions in energy use is necessarily based on the potential for increased technological and operational efficiencies. In this section, the outlook for such improvements in large commercial aircraft over the next quarter century is examined.

Engine efficiencies may be improved by between 10 and 30% with further emphasis on moving more mass through engines that operate at higher temperatures and higher pressures. A continuation of the historical trend would lead to a 10% increase in L/D by 2025, and further improvements in the reduction of parasitic drag may extend these savings to perhaps 25%. However, the technologies associated with these improvements have weight and noise constraints that may make their use difficult. For example, the lack of historical improvement in structural efficiency suggests that weight reductions will be offset by added weight for other purposes (e.g., engines, entertainment). However, weight represents an area wherein major improvements may be found without the constraints that may hinder improvement in engine or aerodynamic efficiency. If lighter weight, high-strength materials can be substituted into the predominantly metallic aircraft structures of today, the potential exists for 30% savings through the use of composites to the extent they are currently employed for some military applications.

Although some studies suggest that non-optimum use of airspace and ground infrastructure will be reduced through congestion control, such improvements may only maintain current efficiencies because air traffic growth remains strong. Historical trends for η_g, η_a, and η_{ft} also show constant air traffic efficiencies since 1968. Improved scheduling and equipment commitment can improve load factor, but congestion and low load factor during early morning/late evening flights may limit improvements to the historical rate. This represents a continuation of recent historical trends. At this rate of improvement, about 0.2%/year, the worldwide average load factor is expected to reach around 0.77 by 2025. For an individual aircraft, a seating arrangement that utilizes a small amount of floor space per seat is beneficial to E_I. This trend also applies to the fleet as a whole. It has been estimated that the average number of seats per aircraft may grow by 1.0% annually over the next 20 years.

Based on this outlook, a 25–45% reduction in E_U would be possible by 2025. This is equivalent to a change in E_U of 1.0–2.0%/year, i.e., 40–75% of the average rate over the previous 35 years. In terms of E_I, the addition of load factor results in an estimated improvement of 1.2–2.2%/year, or a 30–50% reduction in E_I by 2025. As was shown in Fig. 2, over the period 1971–1985, airlines found profitable an average 4.6%/year reduction in fleet average E_I on an RPK basis, which translates into a slower 2.7% improvement on a seat-kilometer basis when the contribution of load factor is removed (E_U). Over the period 1985–1998, however, the rate of change was slower, at approximately 2.2% in E_I and 1.2%/year in E_U. Fleet average projections in the literature suggest a 1.3–2.5% annual change in fleet average E_I and a 0.7–1.3%/year change in E_U. These studies are consistent with recent historical trends.

Beyond the evolution of the current aircraft platform, hydrogen and ethanol have been proposed as alternative fuels for future low-emission aircraft. Hydrogen-fueled engines generate no CO_2 emissions at the point of use, may reduce NO_x emissions, and greatly diminish emissions of particulate matter. However, hydrogen-fueled engines would replace CO_2 emissions from aircraft with a threefold increase in emissions of water vapor. Considering uncertainties over contrails and cirrus cloud formation, and the radiative impact of water vapor at higher altitudes (Fig. 2), it is not clear whether use of hydrogen would actually reduce the contribution of aircraft to radiative forcing. In addition, several issues must be resolved before a new fuel base is substituted for the existing kerosene infrastructure. The actual usefulness of such alternative fuels requires a balanced consideration of many factors, such as safety, energy density, availability, cost, and indirect impacts through production. Renewable biomass fuels such as ethanol have much lower energy density than does kerosene or even hydrogen, requiring aircraft to carry more fuel. They would again increase water vapor emissions from aircraft in flight. Hence, kerosene is likely to remain the fuel for air travel for the foreseeable future.

8. INDUSTRY CHARACTERISTICS, ECONOMIC IMPACTS, AND BARRIERS TO TECHNOLOGY UPTAKE

Although reducing energy intensity tends to reduce overall emissions, factors inherent to air transportation can act to counter the potential benefits. Reductions in emissions are hindered by the relatively long life span and large capital and operating costs of individual aircraft, and the resulting inherent lag in the adoption of new technologies throughout the aviation fleet. In improving the performance of technologies that are adopted, trade-offs are inevitable. For example, increasing the efficiency of new engines may potentially increase NO_x emissions as a result of higher peak engine temperatures. Further, the impact of any efficiency improvements is diminished by fuel wasted in airborne or ground travel delays or in flying partially empty aircraft. Perhaps most importantly, we do not know the cost of change. In this section, some industry characteristics and the economic impact of introducing energy-saving aircraft technologies are examined.

The lag in technology introduction is apparent in Fig. 2. It has typically taken 15–20 years for the U.S. fleet to achieve the same fuel efficiency as that of newly introduced aircraft. Apart from in-use aircraft performance improvements, the rate of improvement in the average E_I is determined by the gradual adoption of new, more fuel-efficient aircraft into the existing fleet. This process of technology uptake depends on various cost factors and market forces. The models described below consider the limitations on this process imposed by cost factors. In assessing future aviation fuel consumption and emissions, it is important to consider the time delay between technology introduction and its full absorption by the world fleet.

Figure 5 shows the relationship between how much it costs to carry a passenger 1 km, in terms of aircraft operation (direct operating cost/revenue passenger-kilometer) and the fuel efficiency of the aircraft (here in RPK/kilogram) for 31 selected aircraft types during the period 1968–1998. In addition to fuel efficiency, stage length has a strong influence on operating costs of regional (very short stage length) flights. The direct operating cost (DOC)–fuel efficiency relationship is indicative of the use of technological advances for the purposes of lowering operating costs. However, reductions in the future cost stream are purchased through higher capital and investment costs. That is, airlines are willing to pay a higher acquisition cost if they can gain from savings in DOC, mainly through lower fuel and maintenance costs during the lifetime of aircraft. The plot of aircraft price-per-seat versus DOC/RPK in Fig. 6 shows that aircraft price is inversely proportional to DOC.

The DOC–fuel efficiency and price–DOC relationships imply a potential constraint for energy use and emissions reduction in the aviation sector. If the relative changes in DOC and price with respect to technological improvements occur at historically accepted levels, airlines will continue to adopt newer and more efficient technologies at a higher price, balanced by the promise of sufficient future revenue.

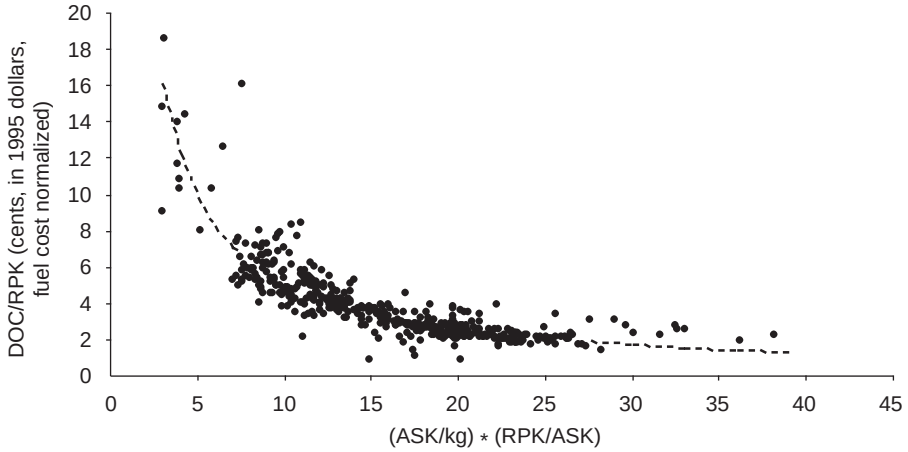


FIGURE 5 Direct operating cost (DOC) and fuel efficiency relationship. The DOC here is composed of crew, fuel, and maintenance costs. RPK, revenue passenger-kilometers; ASK, available seat-kilometers. Reproduced from Lee *et al.* (2001), with permission.

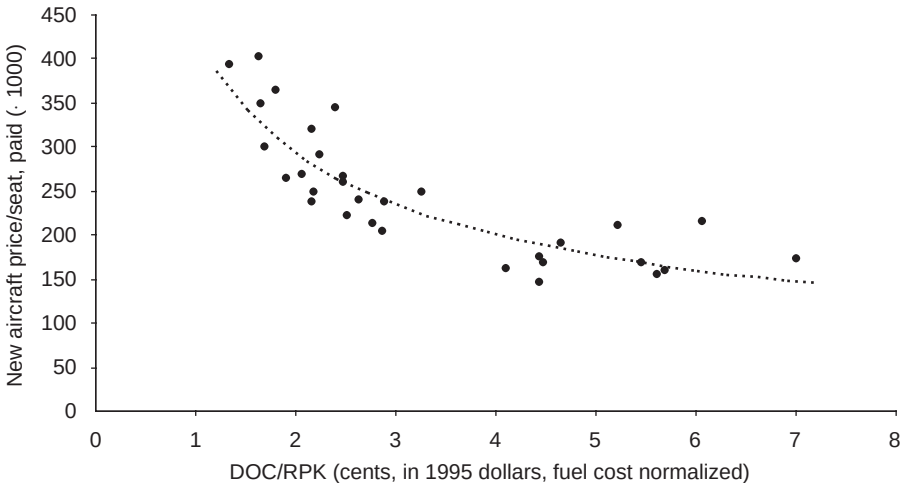


FIGURE 6 Direct operating cost (DOC) and price relationship. Reported prices are average market values paid in then-year (1995) dollars for new airplanes at the time of purchase. All prices were adjusted to 1995 dollars using gross domestic product deflators. RPK, revenue passenger-kilometers. Reproduced from Lee *et al.* (2001), with permission.

However, it is unclear whether future technologies can be delivered at an acceptable price-to-DOC ratio. If the price is too high, airlines may not choose to pay more for energy-saving technologies, in which case further improvements in energy use for the aviation sector may be limited.

SEE ALSO THE FOLLOWING ARTICLES

Internal Combustion Engine Vehicles • Marine Transportation and Energy Use • Transportation and Energy, Overview • Transportation and Energy Policy • Transportation Fuel Alternatives for Highway Vehicles

Further Reading

Anderson, J. D. (2000). "Introduction to Flight." 4th ed. McGraw-Hill, Boston.

Babikian, R., Lukachko, S. P., and Waitz, I. A. (2002). Historical fuel efficiency characteristics of regional aircraft from technological, operational, and cost perspectives. *J. Air Transport Manage.* **8**(6), 389–400.

Balashov, B., and Smith, A. (1992). ICAO analyses trends in fuel consumption by world's airlines. *ICAO J.* August, 18–21.

Greene, D. L. (1992). Energy-efficiency improvement potential of commercial aircraft. *Annu. Rev. Energy Environ.* **17,** 537–573.

International Civil Aviation Organization (ICAO). (2002). 2001 annual civil aviation report. *ICAO J.* August, 12–20.

Kerrebrock, J. L. (1992). "Aircraft Engines and Gas Turbines." 2nd ed. MIT Press, Cambridge, Massachusetts.

Lee, J. J., Lukachko, S. P., Waitz, I. A., and Schafer, A. (2001). Historical and future trends in aircraft performance, cost and emissions. *Annu. Rev. Energy Environ.* **26,** 167–200.

National Research Council, Aeronautics and Space Engineering Board. (1992). "Aeronautical Technologies for the Twenty-First Century." National Academies, Washington, D.C.

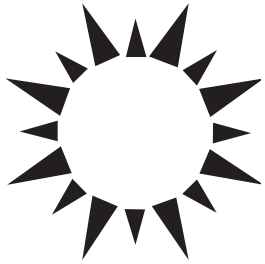
Penner, J. E., Lister, D. H., Griggs, D. J., Dokken, D. J., and McFarland, M. (eds.). (1999). "Aviation and the Global Atmosphere: A Special Report of the Intergovernmental Panel on Climate Change." Cambridge Univ. Press, Cambridge, U.K.

Royal Commission on Environmental Pollution (RCEP). (2002). "Special Report. The Environmental Effects of Civil Aircraft in Flight." RCEP, London.

Schafer, A., and Victor, D. (2000). The future mobility of the world population. *Transp. Res. A* **34**(3), 171–205.

U.S. Federal Aviation Administration (FAA). (2000). "FAA Aerospace Forecasts Fiscal Years 2000–2011." FAA, Washington, D.C.

Air Pollution from Energy Production and Use



J. SLANINA
Netherlands Energy Research Foundation
Petten, The Netherlands

1. **Introduction**
2. **Emissions**
3. **The Atmosphere, Meteorology, and Transport**
4. **Chemical Conversion in the Troposphere and Stratosphere**
5. **Deposition**
6. **Lifetimes of Pollutants**
7. **Impact of Air Pollutants**
8. **Abatement Measures**

Glossary

acid deposition The total effects caused by the deposition of acids and acid-generating substances on ecosystems.

aerosol The system of small particles suspended in air. The term is also loosely used to describe the particles as such. Aerosol particles range from 0.003 to 100 μm in aerodynamic diameter. They are partly emitted as particles (e.g., by dust storms) and partly generated by conversion of gaseous emissions in the atmosphere.

anthropogenic emissions Caused, directly or indirectly, by human activities. The emission of sulfur dioxide due to use of fossil fuels is an example of a direct cause of emissions, and the emission of nitrogen oxides from farmland as a function of fertilizer application is an example of an indirect cause.

eutrophication The total effects on ecosystems, terrestrial as well as aqueous, due to high input of nutrients, mainly nitrogen and phosphorus compounds.

global air pollution Increased air pollutant concentrations of long-lived (>5 years) compounds, which affect air quality on a global scale regardless of the location of the emissions.

local air pollution Increased air pollutant concentrations and the effect caused by this increase near sources of emission, up to distances of a few tens of kilometers.

natural emissions Occur without any interaction of human activities. Generation of sea spray aerosols by waves or the emission of sulfur dioxide by volcanoes are examples.

oxidants The products of the reactions of volatile organic compounds with nitrogen oxides, which impact human health, agriculture, and ecosystems. The most important oxidant is ozone, but hydrogen peroxide and peroxy acetyl nitrate are other representatives.

regional air pollution Increased air pollutant concentrations and the effects thereof caused by emissions at distances up to 1000 km.

The increased use of fossil energy since the industrial revolution, and especially since 1950, has been the major cause of increased emissions of air pollutants and, correspondingly, many environmental problems. Emissions due to the use of energy are major sources of sulfur dioxide, nitrogen oxides, carbon dioxide, and soot and constitute a large contribution of methane, nonmethane volatile organic compounds, and heavy metals. Depending on conversion due to atmospheric chemical reactions, on meteorological transport, and on deposition processes, air pollution can be transported from hundreds of kilometers (ammonia) to several thousands kilometers (aerosols) on a truly global scale (CO_2 and CFCs). The adverse effects of emissions due to use of energy range from very local to regional and global. In cities, traffic can cause very high concentrations of nitrogen oxides and carbon monoxide, but also secondary products such as ozone and aerosols, especially under conditions of stagnant air. The adverse effects of ozone on human health at concentrations higher than $200\,\mu g\,m^{-3}$ are well documented. Investigations indicate that aerosol concentrations of $50–100\,\mu g\,m^{-3}$ are harmful and in The Netherlands (16 million inhabitants) cause the same number of deaths as traffic accidents per year, approximately 1500. On a regional scale (1000 km), acid deposition and photochemical smog (ozone) cause well-documented harm to vegetation and human health. Eutrophication (too high concentrations

 © 2004 Elsevier Inc. All rights reserved.

of nutrients such as phosphate, ammonium, and nitrate) is also a regional problem, quite manifest in Western Europe. Loss of visibility due to backscatter of light by aerosols is also largely a regional phenomenon. A truly global problem is the changes in the radiative balance of the earth due to increased concentrations of greenhouse gases and aerosols.

1. INTRODUCTION

Air pollution is difficult to define because many air pollutants (at low concentrations) are essential nutrients for sustainable development of ecosystems. Therefore, air pollution can be defined as a state of the atmosphere that leads to exposure of humans and/or ecosystems to such high levels or loads of specific compounds or mixtures thereof that damage is caused. With very few exceptions, all compounds that are considered air pollutants have both natural and man-made origins, mainly from energy production and use. Also, air pollution due to energy use is not a new phenomenon; it was forbidden in medieval times to burn coal in London while Parliament was in session. Air pollution problems have dramatically increased in intensity and scale due to the increased use of fossil fuel since the industrial revolution.

All reports on air pollution in the 19th and early 20th centuries indicated that the problems were local, in or near the industrial centers and the major cities. Even the infamous environmental catastrophes in the area of Liege in the 1930s and in London in the 1950s were essentially local phenomena. In the London smog episode, stagnant air accumulated such extremely high sulfur dioxide and sulfuric acid concentrations of approximately 1900 and 1600 $\mu g\,m^{-3}$, respectively—some 20 times the current health limit—that 4000 inhabitants died as a result. The main causes were the emissions from coal stoves used for heating and the fact that all emissions were trapped in a layer of air probably only a few hundred meters high, with no exchange of air within the city.

During the second half of the 20th century, the effects of air pollution due to energy use and production were detected on regional (>500 km), continental, and even global scales. In approximately 1960, the first observed effects from acid deposition were observed on regional and continental scales. Fish populations in lakes in Scandinavia and North America declined as the lakes were acidified by acid deposition to such a degree that fish eggs would not hatch and no young fish were produced. Approximately 10 years later, damage to forests, the loss of vitality of trees, also contributed to acid deposition. Smog episodes in cities such as Los Angeles were reported during the same period. Reactions of volatile organic compounds (VOCs) and nitrogen oxides, emitted by traffic, produced high concentrations of ozone and peroxides, which are harmful for humans and ecosystems. During the same period, high oxidant concentrations (the complex mixture of ozone, peroxides, and other products of the reactions of organics and nitrogen oxides) were becoming increasingly more frequent in Europe during stagnant meteorological conditions. Also, severe eutrophication (damage and changes to ecosystems due to the availability of large amounts of nutrients) occurred in Europe and the United States. Deposition of ammonium and nitrates (partly caused by fossil energy use) was shown to contribute substantially to high nutrient concentrations in soil and groundwater, leading to large-scale dying off of fish in the United States and Europe. It also caused extremely high nitrate concentrations in groundwater, with the result that a large part of the superficial groundwater in The Netherlands is now unfit for human consumption.

Since 1990, the increased concentrations of radiative active substances (compounds that alter the radiative balance of the earth—greenhouse gases, aerosols, and water in liquid form as clouds) and the resulting climatic change have received a lot of attention. Greenhouse gases absorb long-wave infrared radiation emitted from the earth, thereby retaining heat in the atmosphere and increasing the total radiative flux on the surface of the earth. Aerosols and clouds reflect incoming short-wave sunlight and influence the optical properties of clouds toward more reflection of sunlight; hence, increasing aerosol concentration leads to a decrease in the radiative flux on the surface.

The destruction of stratospheric ozone by chlorofluorocarbon compounds (CFCs) is one of the few environmental problems not related to energy use. Epidemiological research has demonstrated the effects of aerosols on the respiratory tract (inducing asthma and bronchitis), and a large part of the ambient aerosol is caused by emissions due to energy use and production.

This timing of air pollution problems could give the impression that there were sudden increases in air pollutant concentration, but this is probably not correct, as can be demonstrated in the case of ozone. By carefully characterizing old methodologies, it has been possible to reconstruct ozone concentrations in the free troposphere (the air not directly influenced by processes taking place on the earth surface) during

the past 125 years (Fig. 1). The ozone concentrations in Europe slowly increased at a rate of 1% or 2% per year from 10 parts per billion (ppb) to more than 50 ppb as a result of the use of fossil energy. It is well documented that the effects of ozone start at levels of approximately 40 ppb (ppb is a mixing ratio of one molecule of ozone in 1 billion molecules of air). Therefore, it is not surprising that the effects of ozone were detected in the 1970s because the background concentration of continental ozone was 30 ppb and additional oxidant formation would increase the ozone concentrations locally or regionally. However, the increase in the continental background concentrations, mainly caused by fossil fuel, had been occurring for a long period.

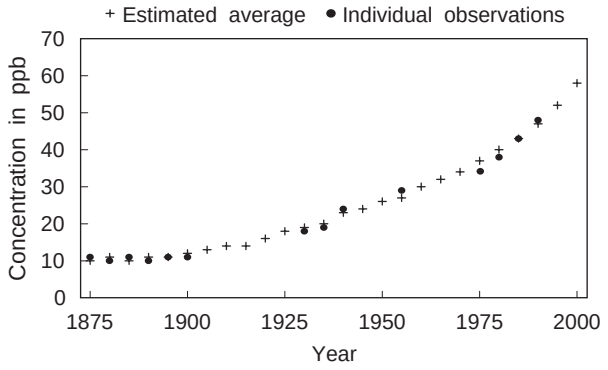


FIGURE 1 Ozone concentration in Europe since 1870. After Voltz-Thomas and Kley (1988).

In general, effects of pollution, and those of air pollution, are a function of the degree of transgression of the limits over which effects can be expected. Figure 2 shows that the use and production of energy is the major cause of most of the current environmental problems, with the exceptions of eutrophication and stratospheric ozone loss.

2. EMISSIONS

Most pollutants are emitted by both natural and anthropogenic sources. Natural sources are not influenced by man or man-induced activities. Volcanoes are a good example. Many emissions are biogenic (i.e., produced by living organisms), but these emissions are often influenced by man's activities. N_2O is a greenhouse gas that is largely emitted during nitrification and denitrification processes (the conversion of ammonium to nitrate and nitrate to N_2 and ammonium, respectively) that occur in the soil. However, the largest N_2O emissions are observed where nitrogen-containing fertilizer is applied in agriculture. The ratio between total anthropogenic, energy-related and natural emissions is very important in order to assess the impact of the use of fuel.

The SO_2 emissions from the use of fossil fuel (Table I) are by far the largest anthropogenic source and are much higher compared to natural emissions

	Energy	Others
Local air pollution		
-Aerosols affect human health	Major	Industry/agriculture
-Aerosols reduce visibility	Major	Industry/agriculture
Regional air pollution		
-Acid deposition	Major	Industry/agriculture
-Eutrophication	Minor	Agriculture
-Oxydant formation	Major	Industry
Global air pollution		
-Climatic change	Major	Industry/land use
-Stratospheric ozone loss	Minor	Consumer/industry
-Antarctic ozone hole		

FIGURE 2 Impact of energy use on the environment.

TABLE I

Global Natural and Anthropogenic Emissions of SO_2 and NO_x

	Emissions (Tg/year)	
	SO_2[a]	No_x[b]
Industrial and utility activities	76	22
Biomass burning	2.2	6
Volcanoes	9.3	
Lightning		5
Biogenic emissions from land areas	1.0	15
Biogenic emissions from oceans	24	
Total anthropogenic emissions	78.2	27
Total natural emissions	34.3	21
Total emissions	112.5	48

[a] Data from Albritton and Meira Filho (2001).
[b] Data from Graedel and Crutzen (1993).

TABLE II

Global Emissions of VOCs

	Anthropogenic (Tg/year)	Natural (Tg/year)
Fossil fuel use	45	
Biomass burning, etc.	45	
Solvents	15	
Oceans		25
Isoprene by vegetation		350
Terpenes by vegetation		480
Total	105	855

TABLE III

Overview of the Emissions of Greenhouse Gases

	CO_2 (Pg C/year)	CH_4 (Tg/year)	N_2O (Tg N/year)
Fossil fuel	5.5	100	
Biomass burning	1.6	40	
Ruminants		80	
Waste treatment		80	
Rice fields		60	
Total anthropogenic	7.1	360	4.4 (sources not well quantified)
Natural		120	6
Wetlands/ continents			
Oceans		10	2
Other		20	
Total natural	C cycles very large	150	8

and exceeded these natural emissions as early as 1950. On a global scale, the use of fossil fuel may contribute approximately 60% of acid deposition.

The emissions of NO_x due to fossil fuels are on the same order of magnitude as the natural emissions. NO_x emissions, together with the emissions of VOCs, are responsible for the formation of oxidants such as ozone and peroxide acetyl nitrate (PAN). Some NO_x sources, especially the natural ones, are uncertain, but the contribution of natural sources to ambient VOCs is not very well-known either. An overview of the emissions of VOCs on the global scale is given in Table II.

Assuming that ozone formation is generally NO_x limited, energy use is responsible for at least 50% of oxidant formation on the global scale. The actual contribution of fossil fuel consumption to acid deposition and ozone is probably much higher because these environmental problems are most pronounced in areas where high emissions by fossil fuel use occur.

The estimates of VOC emissions, especially for the natural/biogenic emissions, are quite uncertain but make clear that the natural sources are predominant. The contribution of fossil fuels is 45% of the total anthropogenic emissions. Biomass burning does not include the use of biomass for energy but denotes the burning of forests and savannahs for agricultural purposes.

If it is assumed (not completely true) that the ambient concentrations vary linearly with the emissions and that only natural emissions are responsible for concentrations before the year 1800, the ratio of natural versus anthropogenic sources can be derived from old trapped air (e.g., in glaciers and polar ice masses). The ratio of natural and anthropogenic contributions varies considerably for the different greenhouse gases: It is highest for N_2O, approximately a factor 10; for carbon dioxide, approximately a factor of 3; for methane, approximately a factor of 1; and lowest for CFCs, a factor of 0, because there are no known natural sources. An overview of the emissions of greenhouse gases is given in Table III. Natural sources for CO_2 are not mentioned because large carbon cycles (e.g., decay of biomass in fall and uptake during leaf formation and growth in spring) occur, which are in equilibrium.

Emissions due to fossil fuel use contribute approximately 80% of the increased carbon dioxide (the most important greenhouse gas, responsible for

50% of the greenhouse effect) and approximately 20% of the increase in methane, which is thought to contribute approximately 20% of the greenhouse effect. Obviously, fossil fuel use also contributes to N_2O emissions (e.g., emissions of cars equipped with three-way catalysts), but this contribution is not well-known.

From the overview provided in this section, it is clear that the use of fossil fuel is the basis of most environmental problems.

3. THE ATMOSPHERE, METEOROLOGY, AND TRANSPORT

Emitted pollutants are diluted by mixing with ambient air, and they are transported by the general displacement of the air as a function of meteorological conditions. Nearly all pollutants are converted to other species by means of atmospheric reactions. The effect of emissions of pollutants is strongly dependent on dilution, transport, and conversion.

Vertical and horizontal mixing is a function of the lifetime of pollutants. If pollutants have a short lifetime, they are not transported very far and vertical mixing in the atmosphere is very limited. If they have a long lifetime, mixing on a global scale occurs.

The basis of all transport of air masses in the atmosphere is movements of air from areas with high pressure to areas with lower pressure. Unequal heating of the earth's surface is the main reason for these pressure differences. The solar heat flux is much higher in the tropics compared to the polar regions. Therefore, heat is transferred by air movement as well as ocean currents (such as the warm Gulf Stream) from the equator to the poles. Also, the characteristics of the surface (sand is heated very quickly, whereas ocean surfaces are not) are very important and lead to phenomena that differ very much in scale, from a local see breeze to the distribution of high and low pressures over oceans and continents. Variation in the earth's rotation speed from the poles to the equator prevents the straightforward movement of air from low- to high-pressure areas and induces the so-called Coriolis force, resulting in circular movements of air around high-pressure and low-pressure areas. The friction induced by mountains and smaller objects, such as cities or forests, is another factor that influences wind direction and speed. All these factors together result in an average wind pattern as shown in Fig. 3.

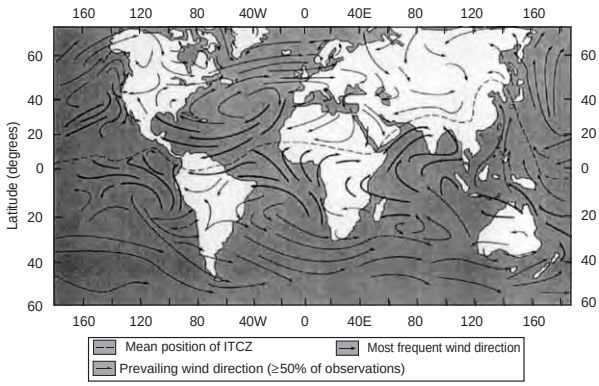


FIGURE 3 Average surface winds (geostrophic winds).

The average wind direction are obviously very important for transport of pollutants: The western winds at the coast in Europe will transport the emissions of the United Kingdom to The Netherlands and Germany, whereas a relatively small amount of German emissions will reach the United Kingdom. All winds converge near the equator, and this phenomenon delays transport from one hemisphere to the other. Consequently, whereas mixing in either hemisphere takes approximately 1 year, a 5-year period is required for global mixing.

In the troposphere (Fig. 4), the temperature decreases as a function of adiabatic expansion (temperature in a gas decreases if the pressure is lowered without heat exchange, approximately 1°C/100 m) and less heat transport from the surface. Rising heated air (called sensible heat) is not the only factor for vertical heat transport. At the surface, over water bodies, and due to evaporation from vegetation, large amounts of water vapor are added to the air. This evaporation takes up much energy, which is freed as water vapor is condensed to water droplets in clouds at higher altitude. However, latent heat transport becomes weaker higher in the troposphere because cooling results in less water vapor in the air.

The heated surface of land, together with turbulence created by objects such as trees and houses, creates a well-mixed layer—the mixing layer or planetary boundary layer. This layer varies from a few hundred meters during the night to 1.5 km during the day under conditions of solar heating. Emissions into this layer are dispersed quickly, especially during the day. If a warm layer is formed over cold air in the troposphere, it can act like a barrier for vertical dispersion: Air masses moving up under these conditions are colder and heavier than this layer and cannot pass this layer. This situation is

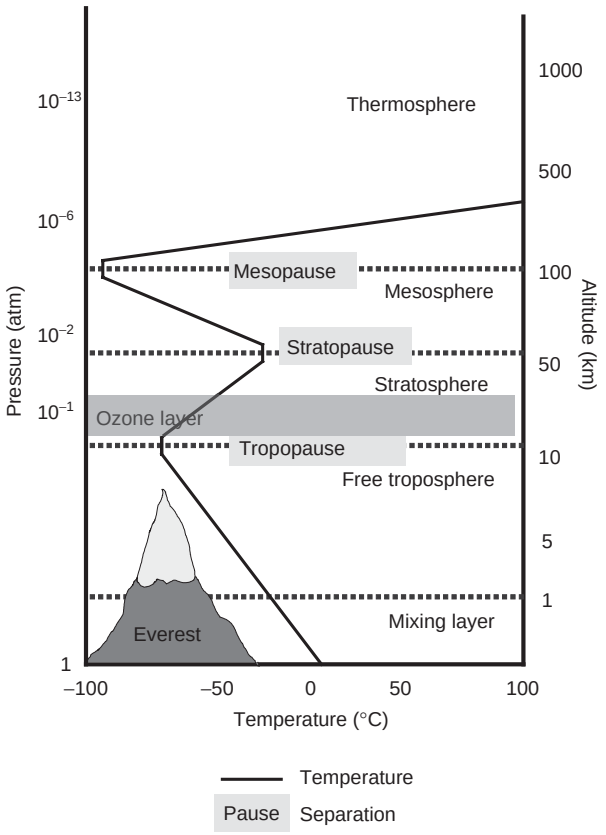


FIGURE 4 Vertical structure of the atmosphere.

called inversion (temperature increasing with height), and all emissions are trapped under this warm layer of air. A strong inversion was an important factor in the infamous London fog episode and is also important for photochemical smog in cities such as Los Angeles, Athens, and Mexico City.

Pollutants are transported by turbulence to the free troposphere, the part of the troposphere not in contact with the earth surface. Movement of air masses in this layer is the main factor for long-range transport of air pollutants, as discussed later. Temperatures decrease as a function of altitude in the troposphere until a minimum of -50 to $-80^{\circ}C$ is reached at the tropopause. The height of the tropopause varies with latitude. Due to much higher solar fluxes, the tropopause is at a height of 17 km in the tropics and approximately 10 km in the polar regions. At greater height, the atmosphere is heated by absorption of ultraviolet (UV) light by oxygen and ozone,

$$O_2 + h\nu \rightarrow 2\ O^{\bullet}\ (\text{wavelength} < 240\ \text{nm})$$
$$O^{\bullet} + O_2 \rightarrow O_3$$
$$\text{Net } 3\ O_2 + h\nu \rightarrow 2\ O_3,$$

and also destroyed again:

$$O_3 + h\nu \rightarrow O^{\bullet} + O_2\ (\text{wavelength} < 1140\ \text{nm})$$
$$O^{\bullet} + O_3 \rightarrow 2\ O_2$$
$$\text{Net } 2\ O_3 + h\nu \rightarrow 3\ O_2,$$

where $O^{\bullet}$ is an oxygen atom with an unpaired electron, a radical.

The net result is that an ozone concentration on the order of 5000 ppb is encountered, whereas at the surface concentration levels of approximately 10–50 ppb are found. The warm layer above the tropopause is a barrier to exchange between the troposphere and stratosphere. Hence, complete vertical mixing over the troposphere and stratosphere takes many years.

Short-lived compounds (a few hours) are transported a few hundred kilometers, mostly in the mixing layer. Compounds with lifetimes of a few weeks are transported in the free troposphere and reach distances of 1000 km. It takes several months to 1 year for a compound emitted at the surface to reach the stratosphere. In some areas, less time is required; cold fronts in the northern Atlantic mix the atmosphere to a height of 20 km, which leads to relatively rapid exchange (weeks) locally between the troposphere and stratosphere. Compounds with lifetimes of 5 years or longer are globally mixed. An overview of all mixing processes, including mixing in the ocean, is given in Fig. 5.

Some air pollutants (e.g., ozone) have such a long lifetime that intercontinental transport occurs. Ozone originating in the United States, for instance, can reach Europe with the predominant western winds. Oxidants of European origin contribute to the oxidant concentrations in Asia. Currently, there is concern regarding the Asiatic contribution to ozone problems in the United States. Even though the contribution is relative small, less than 10% of the average concentrations according to recent estimates, it will become significant as precursor concentrations increase in one region while stringent abatement reduces the emissions of these precursors in others.

4. CHEMICAL CONVERSION IN THE TROPOSPHERE AND STRATOSPHERE

The competition between meteorological transport and losses due to conversion and removal by deposition determines how far pollutants are transported. The main atmospheric reactant is the OH

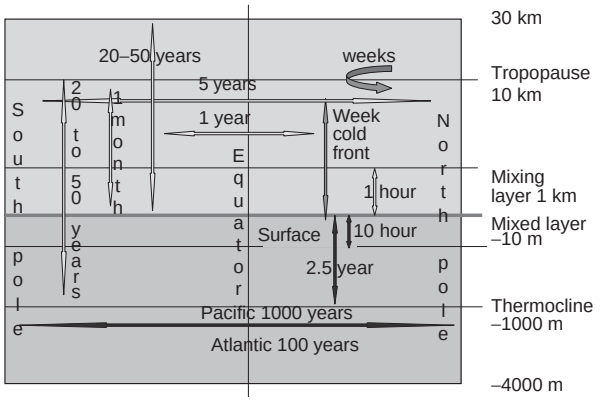


FIGURE 5 Transport times for different scales.

radical, generated by reaction of ozone, water, and short-wave solar radiation:

$$O_3 \overset{\text{Uv rad}}{\Rightarrow} O_2 + O$$

$$H_2O + O^{\bullet} \Rightarrow 2\ OH^{\bullet}.$$

The dot following OH is used to indicate that the species is a radical; it has only one electron instead of the electron pair, which normally constitutes a chemical bond.

The main emissions due to the use of fossil fuels are sulfur and nitrogen oxides and organic compounds. Sulfur and nitrogen oxides are oxidized in the atmosphere to sulfates and nitrates, and organic compounds are transformed into carbon dioxide and water. Some of the emitted organic compounds and the intermediate products of atmospheric conversion are toxic and affect human health(e.g., benzene, toluene, and formaldehyde).

SO_2 is oxidized to sulfuric acid (H_2SO_4) via gas phase reactions and in cloud droplets. The gas phase conversion starts with reaction with the OH radical, leading to the formation of sulfuric acid:

$$SO_2 + OH^{\bullet} \Rightarrow HOSO_3 + O_2 \Rightarrow HO^{\bullet} + SO_3$$

$$SO_3 + H_2O \Rightarrow H_2SO_4.$$

The speed of gas phase conversion can vary from a few tenths of a percent to a few percent per hour, depending on OH radical concentrations and other factors. The conversion in the water phase starts with uptake of SO_2 in droplets and the formation of sulfurous acid:

$$SO_2 \Leftrightarrow SO_{2(aq)} + H_2O \Leftrightarrow H_2SO_3 \Leftrightarrow H^+ + HSO_3^-.$$

Thus, SO_2 uptake is dependent on the next equilibria. If the H^+ formed by dissociation of sulfurous acid is removed (e.g., by reaction with ammonia or carbonates), more sulfur dioxide is dissolved. If, however, high H^+ concentrations are present (e.g., due to uptake of nitric or hydrochloric acid or the formation of sulfuric acid), the uptake of SO_2 is inhibited.

Next, H_2SO_3 is oxidized:

$$H_2SO_3 + O_3/H_2O_2/O_2 \Rightarrow H_2SO_4 \Rightarrow 2H^+ + SO_4^{2-}.$$

Oxidation by oxygen is only possible when suitable catalysts, such as Fe(III) or MN(II), are present. The reaction with ozone is fast at low H^+ concentrations but slow in acid conditions. Sulfur dioxide reacts fast with hydrogen peroxide independent of the pH of the droplets. The pH of the cloud droplets determines the uptake of SO_2 in the droplets; lower pH means that the equilibrium is displaced toward SO_2.

Nitrogen dioxide is oxidized only in the gas phase because both NO and NO_2 are very sparingly soluble in water. The basic reaction in daytime is

$$NO_2 + OH^{\bullet} \Rightarrow HNO_3.$$

The reaction rate of NO_2 with OH radicals in the summer is higher compared to that of SO_2.

A very important series of reactions involving OH radicals, nitrogen oxides, and reactive organic compounds leads to the formation of oxidants. Again, the initial reaction involves OH radicals:

$$RH + OH^{\bullet} \Rightarrow RO + H(A),$$

where RH is a organic aliphatic or olefenic compound, such as ethane, butane, or propene:

$$RO + OH^{\bullet} \Rightarrow RO_2 + H^{\bullet}.$$

Here, a peroxide is formed, such as PAN:

$$H^{\bullet} + O_2 \Rightarrow HO_2^{\bullet}$$

$$HO_2^{\bullet} + NO \Rightarrow OH^{\bullet} + NO_2.$$

The role of NO is to convert HO_2 to OH radicals:

$$NO + RO_2 \Rightarrow NO_2 + RO.$$

NO, NO_2, O_3, and the total light flux are in equilibrium:

$$O_3 + NO \Rightarrow NO_2 + O_2$$

$$NO_2 + h\nu \Rightarrow NO + O^{\bullet}(<380\ \text{nm})$$

$$O^{\bullet} + O_2 \Rightarrow O_3$$

Net:

$$RO_2 + h\nu + O_2 \rightarrow RO + O^{\bullet} + O_3.$$

This means that the hydrocarbon compound is oxidized, ultimately, to carbon dioxide and water; that ozone and OH radicals are formed; but that nitrogen oxide is formed again. Therefore, nitrogen oxide acts as a catalyst but is not really a reactant.

The next important observation is that radicals are actually formed during oxidant formation as long as sufficient NO_x is present. The positive aspect is that the oxidation by OH radicals of organic compounds cleans the atmosphere; in this way, many toxic organic compounds are eliminated.

Species that are not destroyed by reactions with the OH radical or the longer wave UV radiation encountered in the lower layers of the atmosphere ultimately pass the tropopause and enter the stratosphere.

5. DEPOSITION

Deposition is the process that brings air pollutants down to vegetation, soil, and water surfaces and exposes ecosystems to acid deposition. Deposition of acid substances and eutrophication (excessive deposition of nutrients, especially nitrogen compounds) damage ecosystems. Deposition, together with conversion, is a sink for pollutants and, hence, a factor that determines the range of transport.

Three main processes for deposition can be distinguished:

1. Wet deposition: Compounds are transported to the surface by means of precipitation, rain or snow.
2. Occult deposition: Compounds dissolved in fog or cloud droplets are deposited on the surface by means of deposition of these cloud or fog droplets.
3. Dry deposition: Deposition of compounds to the surface directly, not involving the liquid phase.

Wet deposition is the main deposition process in "background" areas—those regions where low concentrations of gases and aerosols in the atmosphere occur, such as Amazonia in Brazil, northern Norway, and the Tarim Basin in the Gobi desert.

Occult deposition is important at locations frequently exposed to fog and clouds (e.g., some hills in Scotland and the Po Valley in Italy) but a minor deposition mechanism at most places. Dry deposition is the dominate process in areas characterized by high air concentrations of gases and aerosols because, in general, dry deposition fluxes increase linearly with concentration.

Whereas wet deposition in northwestern Norway (a background area) is responsible for more than 80% of the deposition load (load is deposition flux integrated over a time period, typically expressed in units of moles per hectare per year) and occult deposition contributes more than 40% of the pollution load over Great Dunn Fell in Scotland (which is covered with low-altitude clouds 50% of the time), dry deposition is responsible for more than 75% of the total deposition load in The Netherlands (typical for European industrial countries that are moderately polluted).

5.1 Wet Deposition

Wet deposition is directly linked with cloud formation. Cloud droplets are formed on aerosol particles. If no particles are present, very high supersaturation of water vapor is needed to form droplets. Particles with a diameter of 0.05–2 μm, consisting of water-soluble material, are the favorite cloud condensation nuclei—the particles on which droplets are formed. This pathway is very important for the uptake of aerosols in clouds and precipitation. Water-soluble gases are also scavenged by cloud droplets.

NO_2 and NO are examples of compounds that are not readily incorporated in the water phase due to limited solubility. Thus, they are not scavenged effectively. SO_2 is scavenged as a function of the pH of the droplet, and the resulting sulfite is oxidized to sulfate.

Maps of wet deposition are generally directly based on measurements: Precipitation is sampled in collectors of wet deposition monitoring networks and analyzed in the laboratory. Figure 6 provides an overview of the wet deposition of sulfate (corrected for the sulfate produced by sea salt) per hectare per year in Europe.

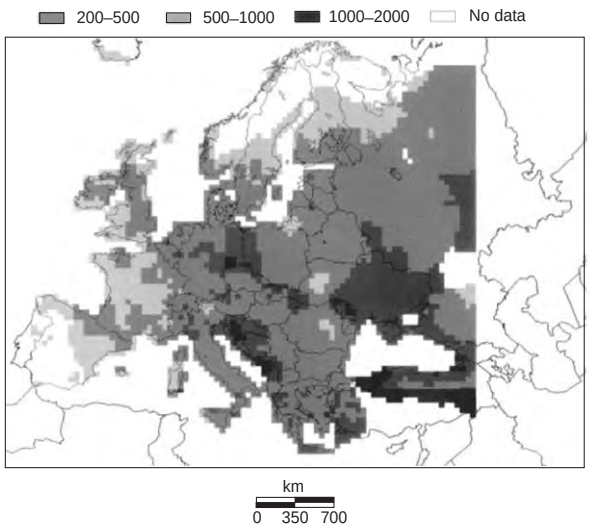


FIGURE 6 Wet deposition of sulfate in Europe.

5.2 Dry Deposition

Aerosol particles and gases are also deposited directly on vegetation and soil without prior uptake in the water phase. The deposition flux is calculated as the product of vertical deposition velocity (Vd) and ambient concentration:

$$F_x = \mathrm{Vd}_x \cdot \mathrm{C}_x,$$

where F_x is the dry deposition flux of compound X, Vd_x is the deposition velocity of compound X, and C_x is the concentration of compound X.

The deposition velocity, the vertical speed of compounds directed downward, is dependent on a number of factors:

- Transport in troposphere occurs by means of turbulence in the form of eddies. If the troposphere is turbulent (e.g., due to high wind speed or strong mixing due to solar heating), the vertical transport of gaseous compounds and aerosol in the boundary layer is quite fast
- Each surface dampens turbulent transport, and very near the surface the turbulent transport is completely stopped.
- Transport through this stagnant layer is only possible by means of diffusion (in the case of gases) or by impaction or interception (for aerosols).
- The compound or aerosol particle reaches the surface next. It can adhere to the surface or bounce back to the atmosphere. Very polar gases, such as HNO_3, NH_3, or HCl, are adsorbed strongly by nearly all surfaces, so they are retained quantitatively. The same is true for particles, which are in fact very small droplets. Other compounds are not adsorbed and can leave the surface again.

The dry deposition load (deposited flux over a specified time period per unit area) is a linear function of the concentration. Thus, dry deposition loads are dominant in areas with higher air pollution concentration.

Vd_x levels vary depending on circumstances. The surface of vegetation in Western Europe is often covered with a water layer due to precipitation or dew. Sulfur dioxide dissolves easily in water, especially at a pH of 7 or higher, so the combination of high ambient ammonia concentrations and the frequent presence of water layers means that sulfur dioxide is easily taken up by the surface of vegetation. In southern Spain, on the other hand, water layers are rarely present on vegetation. Thus, the Vd in The Netherlands is at least twice the Vd for SO_2 in Spain, so the acid deposition fluxes in The Netherlands are two times higher at the same ambient concentration.

Because deposition fluxes and Vd are difficult to measure, dry deposition fluxes are generally modeled, and the models are parameterized and validated using dry deposition flux measurements. The flux is either derived from variations in the concentration as a function of vertical wind speed or from the gradient in concentration of the compound of interest as a function of height.

5.3 Modeling of Acid Deposition

Models are used to derive the concentrations as well as wet and dry deposition fluxes based on descriptions of emissions, atmospheric transport, and conversion and deposition processes. One of the major challenges in atmospheric chemistry is to formulate these models and to provide the necessary data to derive descriptions and to validate the model results. This complicated exercise is shown in Fig. 7.

Emissions and meteorological data are used to develop concentrations maps (Fig. 7, left). Based on high-resolution land use maps (Fig. 7, right) and meteorological data, the vertical deposition velocity, Vd, is derived for every compound. Next, the dry deposition flux is calculated by multiplying the concentration with the vertical deposition velocity. An example of the results is provided in Fig. 8, which shows the NO_y load in Europe in moles/ha/year (NO_y is the sum of HNO_3, HNO_2, NO_2, and NO).

The wet deposition flux is added to derive the total deposition flux (Fig. 6). The total deposition (Fig. 9) is used to derive the impact of the deposition and to calculate the loss of compounds by deposition, necessary to calculate lifetime and transport of emissions. The largest deposition loads are found near the regions where most fossil fuels are used, but some of the emissions are transported up to 1000 km.

6. LIFETIMES OF POLLUTANTS

Conversion in the troposphere and stratosphere and removal by deposition processes determine the fate and the lifetime of pollutants. A simplified approach is used here. Analogous to a resistance model, the overall lifetime can be expressed as

$$1/\tau_{(\mathrm{resultant})} = 1/\tau_{\mathrm{w}} + 1/\tau_{\mathrm{d}} + 1/\tau_{\mathrm{c}},$$

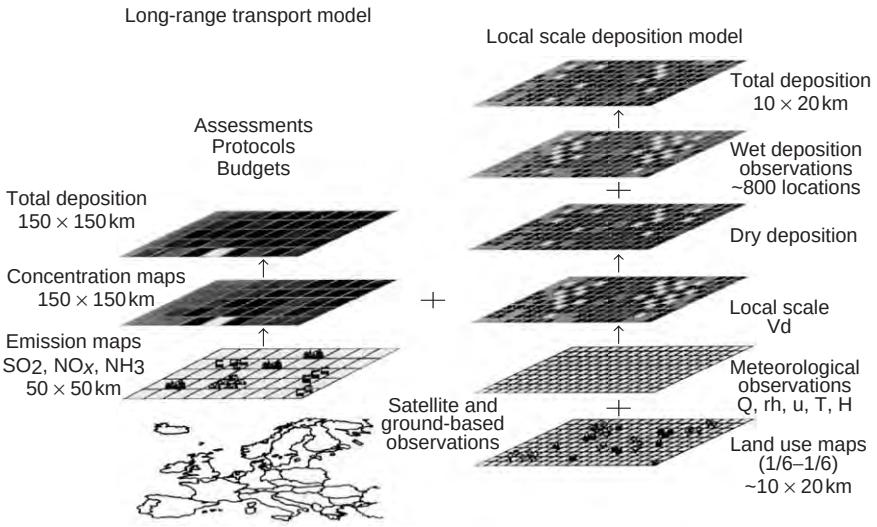


FIGURE 7 Scheme for calculation of deposition loads by modeling.

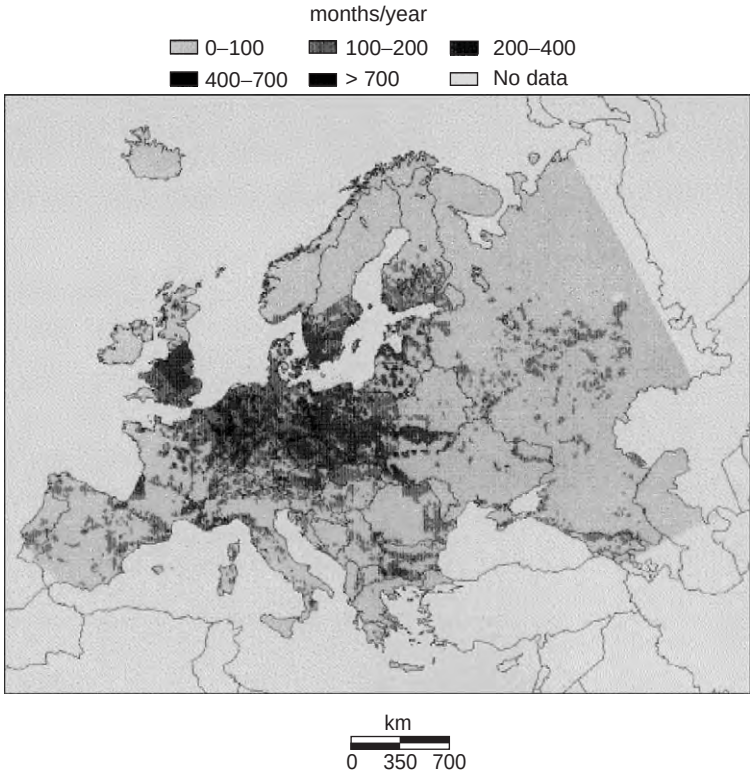


FIGURE 8 Dry deposition load of NO_y (sum of HNO_3, HNO_2, NO_2, and NO) in moles per hectare per year for Europe.

where $\tau_{(resultant)}$ denotes the total lifetime, τ_c is lifetime as a function of conversion, τ_d is lifetime as a function of removal by dry deposition, and τ_w lifetime as a function of removal by wet deposition. An overview of lifetimes of some important pollutants is given in Table IV. In addition to estimates of τ_c, τ_w, and τ_d, Table IV also provides estimates of the total lifetime [$\tau_{(resultant)}$] and of the transport distance. If the lifetime of a compound is so long that a global distribution is observed, this is indicated by "global." "Stratosphere" means that the compound is not destroyed by OH radicals in the

troposphere but only by short-wave UV radiation in the stratosphere.

Very small aerosol particles, as generated by combustion engines, have very short lifetimes because they more or less behave like gas molecules, including transport by Brownian diffusion and rapid coagulation. The large ones are removed by gravity; they fall out of the atmosphere. But particles of approximately 1 μm, which contain most anthropogenic-produced sulfates, nitrates, and organics, live quite long and are generally removed by cloud formation only.

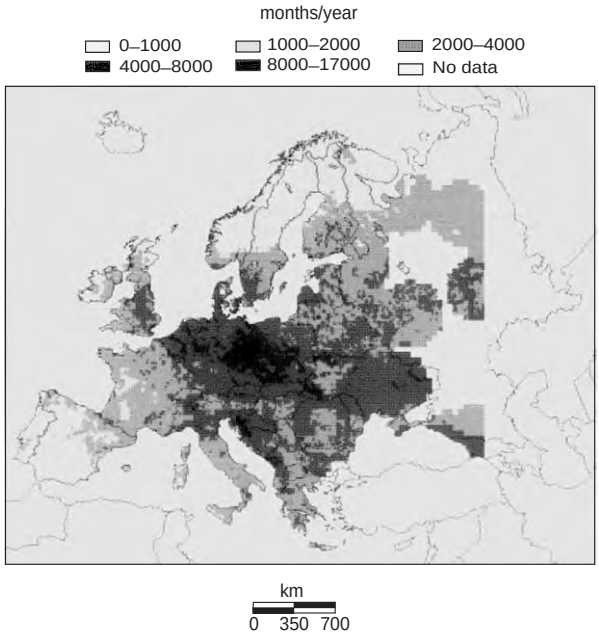


FIGURE 9 Total acid deposition in Europe in moles of acid rain per hectare per year.

For sulfur dioxide, deposition processes and conversion are of equal weight, whereas nitrogen oxides are removed by conversion. The transport distance of ammonia is very limited due to short deposition and conversion lifetimes. The reactivity of different classes of organic compounds varies considerably; hence, there are major differences in lifetimes and transport distance.

Conversion of ozone is important near sources because it reacts rapidly with NO, such as emitted by traffic. Far from sources, dry deposition is the main loss term.

Carbon dioxide is taken up by vegetation (e.g., in spring) and again released as litter decomposes (e.g., in fall), under condition that the vegetation is in steady state. The total cycle of carbon, including vegetation and ocean uptake and release, is so large that the lifetime of emitted CO_2 is estimated to be 100–200 years.

Methane is first converted to CO, and next to CO_2 in the reaction with OH radicals. Uptake of methane occurs in soils, but this probably constitutes only a small fraction compared with conversion to CO and CO_2. N_2O and CFCs are not destroyed in the troposphere but are decomposed in the stratosphere.

7. IMPACT OF AIR POLLUTANTS

The effects of energy-related air pollutants on human health, ecosystems, and essential requirements such

TABLE IV

Lifetimes and Transport Distances of Air Pollutants

Compound	τ_c	τ_w	τ_d	τ (resultant)	Distance (km)
Aerosol					
<0.1 μm	Minutes–hours	—	Hours	Hour	10
>10 μm	—	Large	Hours	Hours	100
(≈ 1 μm)	—	Week	Week	Week	1000
Sulfur dioxide	Days	Days	Days	Days	1000
Nitrogen oxides	Days	—	Days	Days	1000
Ammonia	Hours	Hours	Hours	Hours	250
VOC reactive	Hours–days	Large	—	Hours–days	200–1000
Ozone	Hours–days	—	Days	Hours	200–1000
Carbon dioxide	—	—	200 Years	200 Years	Global
Methane	10 Years	—	Large	12 Years	Global
N_2O	Stratosphere	—	—	200 Years	Global
CFCs	Stratosphere	—	—	100–200 Years	Global

as drinking water vary widely in scale, ranging from very local to global. Here, an overview is given as a function of scale, starting with local effects and proceeding to regional impacts and global issues.

7.1 Local Pollution and Effects

Local pollution is generally caused by too high ambient concentrations as found near sources. Historic pollution problems, such as the London smog periods in the 1950, fall into this category. Most local air pollution problems in cities in developed nations are associated with traffic or, rather, the use of combustion engines in traffic. Cities in the United States and Europe, but also those in Mexico and east Asia, such as Beijing and Delhi, have problems with CO and NO_x concentrations that are too high, directly traffic related. In developing countries in which sulfur-rich coal is used and in which many industries are located in towns, local SO_2 concentrations also present environmental problems. Most developed countries have successfully eliminated this environmental hazard.

Countries and also multinational institutions, such as the European Union, have developed different air quality standards to ensure that the population is not affected by emissions in cities. These standards vary between different countries; an example is given in Table V for nitrogen dioxide. Some cities have limited exchange of ambient air due to the fact that they are surrounded by mountains. In this stagnant air, emissions from traffic can cause high concentrations of nitrogen oxides and reactive organic compounds, and oxidants can form, with high ozone concentrations as the result. Ozone affects human health and also vegetation. This is reflected in the air quality standards for ozone. Table VI includes different ozone air quality standards for protection of vegetation and human health. In some cities in developing countries, the limit at which the population should be warned (e.g., to stay indoors as much as possible) is frequently exceeded.

Epidemiological investigations have provided evidence that particulate matter, aerosols, induce increased mortality, ranging from 4% for U.S. cities to 2% for Western European cities and less than 1% for Eastern European cities for a level of 100 $\mu g\,m^{-3}$. In addition to increased mortality (which for a country such as The Netherlands amounts to 1500–4000 acute deaths in a population of approximately 16 million, equal to the total number of deaths caused by traffic incidents and approximately 10,000 per year due to chronic effects), an increased incidence of afflictions of the respiratory tract (asthma and bronchitis) has been found (e.g., amounting to 30,000 extra cases of bronchitis and asthma in The Netherlands). No mechanisms for this increased mortality have been established, but research indicates that very small particles (mainly emitted by traffic and other fossil fuel use) are probably responsible.

TABLE V

Air Quality Standards for NO_2

Value ($\mu g\,m^{-3}$)	Average time	Remarks
135	98% (1 hour)[a]	The Netherlands standard
175	99.5% (1 hour)	The Netherlands standard
25	50% (1 hour)	Target value, The Netherlands
80	98% (1 hour)	Target value, The Netherlands
200	98% (1 hour)	EU standard
50	50% (1 hour)	Target value, EU
135	98% (1 hour)	Target value, EU

[a] Ninety-eight percent indicates that the standard may be exceeded only 2% of the time. Target value is the air quality standard that is seen as desirable and that probably will be implemented in the future because the necessary technical means have been developed to reach this standard.

TABLE VI

Ozone Air Quality Standards[a]

Value ($\mu g\,m^{-3}$)	Average time	Remarks
240	1 hour	NL[b] standard (not to exceed 2 days/year)
160	8 hours	NL standard (not to exceed 5 days/year)
100	Growing season	March 1–September 1, 9 am to 5 pm
240	1 hour	NL target value
160	8 hours	NL target value
100	Growing season	NL target, March 1–September 1, 9 am to 5 pm
120	1 hour	NL long-term goal
50	Growing season	NL long-term goal
110	8 hours	EU standard (health protection)
200	1 hour	EU standard (protection of vegetation)
65	24 hours	EU standard (protection of vegetation)
180	1 hour	EU information of civilians
360	1 hour	EU warning of civilians

[a] See Table V footnote for description.
[b] NL, The Netherlands.

Aerosols scatter light, and this scattering is very effective if the diameter of the aerosol particle is on the same order as the wave length of the scattered light. This backscattering process is detected by the human eye as haze, limiting visibility. Aerosols with a diameter of 0.1 to approximately 2 μm scatter visible solar light effectively and hence impede visibility. In some polluted cities, aerosols limit visibility to 500 m, especially under conditions of high humidity. In Fig. 9, the left picture was taken when aerosol concentrations where approximately 20 $\mu g\,m^{-3}$. Visibility is quite good and probably more than 30 km. The picture on the right was taken during a dust storm. Dust storms contain large concentrations of not only large particles but also small particles, which reduce visibility to less than 1 km.

7.2 Regional Problems

Regional problems are caused by emissions within a distance of 1000 km; acid deposition, eutrophication, and regional ozone formation fall into this category. The first signs of the effects of acid deposition were encountered when losses of fish populations in Canadian, Scandinavian, and U.S. lakes were reported in the early 1960s. A few years later, there were reports of unusual damage to forests from Canada, the United States, and Germany. Trees were observed to have less foliage or needles than normal, and this damage even progressed to the point that trees died, a phenomenon that occurred on the border of former East Germany, Poland, and former Czechoslovakia. In the case of tree dieback, the cause–effect relation was much more complicated than that for the loss of fish in lakes. In fact, much later it was determined that not only acid deposition but also increased oxidant concentrations (e.g., ozone, PAN, and hydrogen peroxide) contribute to tree damage.

A third type of effect of acid deposition was reported in the mid-1980s. The same deposition processes that were responsible for the high loads (load is a time-integrated flux) of acid compounds were also the cause of high nutrient concentrations in soil, groundwater, and surface waters. Nitrate and ammonium are beneficial, even essential, for development in vegetation, but in excessively high concentrations they lead to the loss of biodiversity, especially in oligotrophic (adapted to low nutrient availability) ecosystems. This problem originated in The Netherlands, Belgium, and parts of Germany, Denmark, and southern Sweden but is now clearly affecting large areas of the United States, Canada, and China. In most countries, agricultural activities play a much larger role in eutrophication than fossil fuel use.

The common factor in all these cases is that pollutants, or their precursors, are emitted, transported in the atmosphere, and deposited by either dry or wet deposition. In some cases, acid deposition takes place 1000 km or more from the place where the responsible emissions originated. The most important acid-forming species are sulfur dioxide (which is partially converted to sulfuric acid in the atmosphere), nitrogen oxides (converted to nitric acid during atmospheric transport), and ammonia. Ammonia may act as a base in the atmosphere, neutralizing nitric and sulfuric acid, but in soil or groundwater it is largely converted by microorganisms to nitric acid, producing additional acid in the process.

Both precursors (SO_2 and NO_x) and products (sulfuric acid, nitric acid, and ammonium) can be transported large distances (up to 1500 km) depending on meteorological conditions, the speed of conversion, and removal by deposition processes. Ammonia, on the other hand, is transformed quickly in ammonium by reactions with sulfuric acid and nitric acid and deposited rapidly in the form of both wet and dry deposition; it is generally transported 250 km or less.

Oxidant problems are not only a local but also a regional phenomenon. An example is the buildup of high concentrations of oxidants, mainly ozone, in Europe when the weather is dominated by large high-pressure systems. The air within these high-pressure systems, which often have a diameter of 1000 km, is not exchanged quickly. So precursor concentrations of NO_x and reactive organic compounds increase due to ongoing emissions mainly related to fossil fuel, and very high ozone concentrations, often exceeding 150 $\mu g\,m^{-3}$, are the result. These ozone concentrations affect not only human health but also vegetation (Fig. 10).

The loss of production of wheat due to ozone starts at a threshold concentration of approximately 40 $\mu g\,m^{-3}$ and increases linearly with the ozone concentration. (Fig. 11). Not only do ozone problems occur in Europe and the United States, but there is also increasing concern in developing countries. For instance, high ozone concentrations occur in Hong Kong and the Chinese province of Guangdong, which includes cities such as Guangzhou and Shenzhen.

Aerosols generally derive from both local and regional sources. The backscatter by aerosols is receiving much attention because backscatter of solar radiation is a very important but uncertain factor in determining the balance of incoming and

FIGURE 10 Visibility on the campus of Peking University. (Left) Photograph taken when the aerosol concentration was approximately 20 μg m^{-3}. (Right) Photograph taken during a dust storm, with a high concentration of small particles.

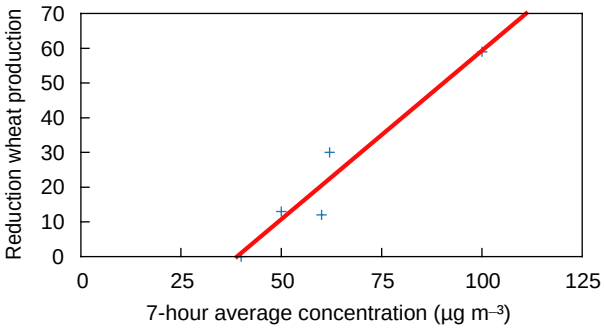


FIGURE 11 Reduction of wheat production as a function of daytime ozone concentrations.

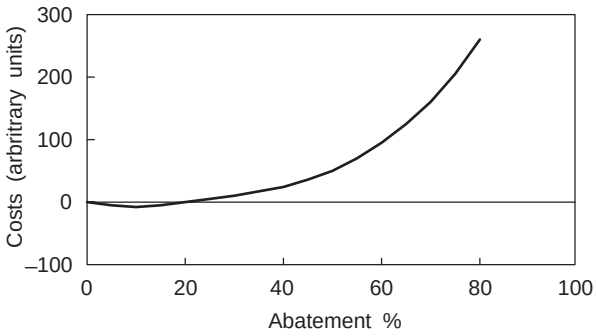


FIGURE 12 Costs of abatement in relative cost units as a function of reduction (abatement).

outgoing radiation of the earth and hence in assessments of climatic change.

7.3 Pollution on a Global Scale

Pollutants that are not converted in the troposphere or deposited rapidly and have lifetimes of 10 years or more can have an impact on a global scale. The greenhouse effect and the Antarctic ozone hole are caused by long-lived air pollutants. Because climatic change is discussed elsewhere in this encyclopedia and the destruction of the stratospheric ozone, leading to the Antarctic ozone hole, is not primarily caused by the use of fossil fuels, these subjects are not treated here.

8. ABATEMENT MEASURES

In view of the major environmental problems caused by the use of fossil fuel, on the one hand, and economic development, on the other hand, it is very important to evaluate the effects of possible abatement measures. The costs of abatement measures increase exponentially with the degree of reduction (Fig. 12).

The scientific basis for abatement measures is obviously emission–effect relations: How are undesirable environmental effects related to specific emissions, and how will changes in these emissions influence the effects of air pollution? It is very important to have as detailed as possible descriptions of emission–effect relations in order to have a scientific and solid fundament for environmental measures. Not only is this necessary to assess the cost-effectiveness of measures but also national consensus of economic groups and international consensus between countries are very dependent on the acceptance of emission–effect relations.

It should be kept in mind that smaller emission reductions, in general, are not very costly. Problems arise when very large emission reductions have to be made. The first hurdle is to reach a generally accepted view on emission–effect relations. Then, a national or international consensus must be reached and so a basis is constituted for reaching agreements. Examples are the protocols under the aegis of UN-EMEP

(Cooperative Programme for Monitoring and Evaluation for Long Range Transport in Europe) for abatement of acid deposition, eutrophication, and impact of oxidants in Europe or the Montreal Protocol, which stringently minimizes emissions of CFCs.

These emission–effect relations are generally so complex that such a description is seldom based on observations only, and mathematical models are needed to obtain these relations. The first stage of this process is to develop a description of the critical limits or critical loads (load = integrated deposition fluxes). Critical concentrations and loads are the maximum concentrations and deposition fluxes above which damage is likely to occur. These descriptions are based on information on the effects of acid deposition, eutrophication, or ozone on ecosystems, taking into account the buffer capacity of the soils and the sensitivity of ecosystems. The result is a map of critical loads in Europe, as shown in Fig. 13 for acid deposition. This description of loads of acid deposition (Fig. 9) is then compared to the description of critical loads, and the transgression (the term exceedence is generally used) can be derived (Fig. 14). The difference between the actual deposition load and the critical load is the central steering mechanism for European environmental policy. These descriptions are the internationally accepted basis of the European protocols for abatement of acid deposition, eutrophication, and the impact of oxidants. If a quantitative and validated scientific description is generally accepted, then a good scientific fundament is present for abatement measures. Therefore, EMEP must describe all processes from emissions, natural and anthropogenic, transport and chemical conversion, and deposition to effects and also provide an assessment of the costs for abatement in order to obtain scientifically defendable conclusions.

Next, individual countries must meet the targets using appropriate measures, optimized to their specific conditions. The following measures have proven to be very effective:

- Not only will the use of wet desulfurization plants reduce the emissions of sulfur dioxide of, for

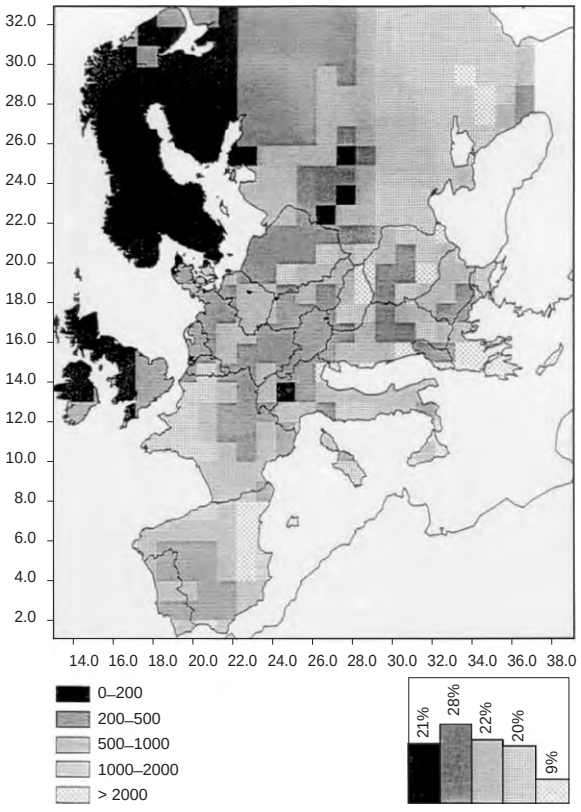


FIGURE 13 Critical loads for acid deposition in moles of acid per hectare per year. The very dark areas are extremely sensitive; damage occurs at loads as low as 200 mol/ha/year. The light areas can accept deposition loads of at least 2000 mol/ha/year without unacceptable damage.

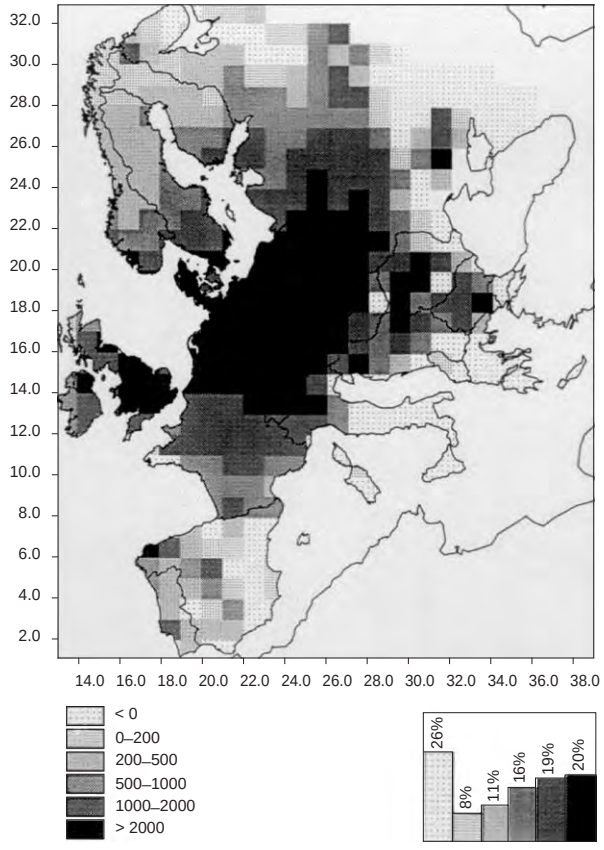


FIGURE 14 Exceedence of critical loads in Europe. The light areas denote no or little transgressions, and the dark areas indicate regions with serious acid deposition problems.

example, electricity-generation plants, by approximately 95% but these installations will also reduce emissions of compounds such as hydrochloric acid and hydrofluoric acid by the same amount.

- The three-way catalyst, now in general use, reduces the emissions of nitrogen oxides and reactive volatile organic compounds by 90–95%, depending on traffic conditions.
- Low NO_x burners, burning fossil fuel in two stages, can reduce NO_x emissions by power plants and industry by 70% or more. Low NO_x burners are even applied in household central heating systems using natural gas.
- Increasing the efficiency of the use of fossil fuels is a very effective way to decrease most emissions.

A very important issue for which such a general acceptance of the scientific description of emission–effect relations has not been reached is the reduction of emissions of greenhouse gases (Kyoto Protocol). On the one hand, the costs of such reductions play an important role, and costs of abatement measures increase exponentially at higher reduction rates. On the other hand, the perceived uncertainty in the assessment of the environmental effects is seen as very important:

- The measures necessary to reach the agreed abatement are judged to be very expensive by a number of countries.
- The effects of climatic change are potentially very damaging and far-reaching. However, the uncertainties regarding regional impacts as well as estimates of the time before these changes are detected give countries reasons to defer measures.
- Uncertainties exist with regard to sources and sinks of greenhouse gases, especially regarding the role of vegetation in the release and sequestration (uptake) of carbon in vegetation and soil.
- Because most of the current increases in ambient carbon dioxide concentrations must be allotted to the use of fossil fuel by developed countries, undeveloped countries voice quite strongly the opinion that they should be exempt from measures that could hamper their economic development. Some developed countries counter this argument by pointing out that developing countries will be responsible in the near future for a large percentage of greenhouse gas emissions and that measures will not be effective unless developing countries also take abatement measures.
- The main impact of climatic change will be experienced, in the opinion of many scientists, by developing countries, whereas most of the costs must be covered by developed countries. This obviously is a very complicating factor.

Although part of the necessary solutions to the debate must be provided by atmospheric chemistry research, particularly information on the fluxes of different sources and sinks, reduction in the use of fossil fuel seems to be unavoidable in order to mitigate the current environmental problems.

SEE ALSO THE FOLLOWING ARTICLES

Acid Deposition and Energy Use • Air Pollution, Health Effects of • Clean Air Markets • Ecosystems and Energy: History and Overview • Gasoline Additives and Public Health • Hazardous Waste from Fossil Fuels • Indoor Air Quality in Developing Nations • Indoor Air Quality in Industrial Nations • Thermal Pollution

Further Reading

Albritton, D. L., Meira Filho, L. G., and the IPCC. (2001). Technical summary. www.ipcc.ch.

EMEP Summary Report. (2000). Transboundary acidification and eutrophication in Europe, CCC and MSC-W Research Report No. 101. Norwegian Meteorological Institute, Oslo.

Finlayson-Pitts, B. J., and Pitts, J. N. (2000). "Chemistry of the Upper and Lower Atmosphere." Academic Press, New York.

Graedel, T. E., and Crutzen, P. J. (1993). "Atmospheric Change; An Earth System Perspective." Freeman, New York.

Jacob, D. J. (1999). "Introduction to Atmospheric Chemistry." Princeton Univ. Press, Princeton, NJ.

Slanina, J. (ed.). (1997). "Biosphere–Atmosphere Exchange of Trace Gases and Pollutants in the Troposphere." Springer-Verlag, Berlin.

Volz-Thomas, A. and Kley, D. (1988). *Biogeochemistry* **23**, 197–215.

Air Pollution, Health Effects of

JONATHAN LEVY
Harvard University School of Public Health
Boston, Massachusetts, United States

1. **Historical Air Pollution Episodes**
2. **Types of Studies Used to Evaluate Health Impacts**
3. **Key Pollutants and Health Outcomes**
4. **Health Effects of Indoor Air Pollution**
5. **Conclusions**

Glossary

cohort study An epidemiological study in which a group of individuals is enrolled and tracked over time to evaluate whether more highly exposed individuals are more likely to develop disease or to die prematurely.

confounder A variable independently associated with both the exposure and the health outcome in an epidemiological study, and could therefore be responsible for the observed relationship.

epidemiological study An observational study of human populations (either occupational or general); designed to determine the association between exposure and disease.

inhalation unit risk The excess lifetime cancer risk estimated to result from continuous lifetime exposure to a substance at a concentration of 1 $\mu g/m^3$ in the air.

particulate matter (PM) A solid or liquid suspended in the air; typically defined by the aerodynamic diameter of the particle and categorized as fine or respirable particles ($PM_{2.5}$, particles less than 2.5 μm in diameter), inhalable particles (PM_{10}, particles less than 10 μm in diameter), coarse particles ($PM_{2.5–10}$, particles between 2.5 and 10 μm in diameter), or total suspended particles (TSPs; suspended particles of all sizes).

reference dose An estimate (with uncertainty spanning perhaps an order of magnitude) of the daily exposure to the human population (including sensitive subgroups) that is likely to be without an appreciable risk of adverse effects during a lifetime; used for the evaluation of noncancer risks of toxic chemicals.

time-series study An epidemiological study in which changes in pollution levels over time are correlated with changes in health outcomes over time, capturing the short-term effects of pollution on health.

toxicological study A controlled laboratory experiment in which a small number of animals is given a gradient of doses of compounds with suspected health impacts, and the relationship between dose and disease rate is estimated.

Understanding the health impacts of air pollution is a crucial step in the determination of appropriate energy policy. Although it has been well established that large increases in pollution exposures can lead to substantial increases in death and disease, it is somewhat more complex to quantify the precise impacts of an incremental change in exposures to a specific pollutant at the levels commonly faced by individuals. This is in part because there is an array of different air pollutants in outdoor and indoor air; pollutants can come from different sources and have different types of health effects, thus different forms of evidence are used to draw conclusions regarding pollutant health impacts. Broadly, air pollution causes health impacts through a pathway in which emissions lead to concentrations of key pollutants in the air, human activity patterns influence the exposures people face from these concentrations, and characteristics of the exposed populations help determine the types and magnitudes of health impacts.

1. HISTORICAL AIR POLLUTION EPISODES

Although there have been a number of documented events whereby extreme levels of air pollution led to elevated death and disease rates, three episodes occurring near the middle of the 20th century are among the most well documented and were the first pollution events studied in great detail. In December 1930, an atmospheric inversion led to a thick pollution fog settling over the Meuse Valley in

 © 2004 Elsevier Inc. All rights reserved.

Belgium. During a 4-day period, 60 deaths were attributed to the pollution (10 times more than normal), and many individuals suffered from respiratory distress, with the elderly or those with respiratory or cardiovascular disease at greatest risk. Levels of a number of pollutants were increased during this episode, with extremely high concentrations of sulfur-based compounds. In the second event, in October 1948, a similar set of circumstances (temperature inversion in an area with substantial industrial activity) led to an air pollution episode in Donora, Pennsylvania. Nearly half of the residents of the town became ill, with 20 deaths attributed to the multiday episode. Finally, perhaps the most damaging and well-known pollution episode occurred in December 1952 in London, where a temperature inversion and associated elevated pollution levels led to 4000 deaths during the episode and to another 8000 deaths during the subsequent months (demonstrating a lagged or cumulative effect of exposure). As previously, elderly individuals and those with respiratory and cardiovascular disease were at greatest risk of health impacts.

These and other similar episodes provided clear evidence of the effects of extremely high levels of air pollution, which generally occurred when significant emissions from low stacks were trapped near the ground due to atmospheric conditions. Because of increased emission controls and taller stacks in many settings at the start of the 21st century, the pollution levels seen in the early studies are currently rarely seen in either developed or developing countries. Evidence from additional sources of information is therefore needed to understand health impacts of lower levels of air pollution, as well as to determine precisely which pollutants are responsible for mortality and morbidity impacts.

2. TYPES OF STUDIES USED TO EVALUATE HEALTH IMPACTS

The evidence used to determine the magnitude of mortality and morbidity effects of air pollution comes from two general types of studies. Toxicological studies are controlled laboratory experiments in which relatively high doses of compounds are administered to animals and rates of diseases (often cancer) are estimated. As controlled experiments, toxicological studies avoid some of the issues in observational studies, wherein a number of potential causal factors can be changing at the same time. However, in order to determine health effects in humans, results from toxicological studies must be extrapolated in at least two important ways—results from animals must be applied to humans, and results from high doses necessary to see effects in a small number of animals must be used to draw conclusions about lower doses typically seen in humans. Because of these often substantial uncertainties, epidemiological studies, when they are available and interpretable, are generally preferred for quantifying the health impacts of air pollution, with toxicological studies providing supporting evidence for a causal effect.

Epidemiological studies are observational studies of human populations; they are designed to show the relationship between exposure to an agent or set of agents and the development or exacerbation of disease. Although the focus on human populations and observed exposures removes some of the extrapolation issues associated with toxicological studies, epidemiological studies run a greater risk of yielding results that do not reflect a true causal relationship between exposure and disease (because correlation does not necessarily imply causation). This would occur if there were confounders, which are variables independently associated with both the exposure and the disease in question, that could actually be responsible for the observed relationship. For an epidemiological study to adequately determine the health effects of air pollution, it must use statistical techniques or a restrictive study design to remove the possibility of confounding by nonpollution variables. Also, unlike toxicological studies, which generally focus on a single pollutant or a defined list of pollutants, epidemiological studies measure health effects that could be associated with a number of pollutants that are often correlated with one another in time or space. Epidemiological studies must also be able to provide accurate exposure estimates, which can be more difficult than in toxicological studies, in which the doses can be measured in a controlled setting, particularly for large populations, for long-term exposures, or for pollutants that vary greatly in time or space.

There are three primary types of epidemiological studies used to understand the health effects of air pollution, each of which has different strengths and different issues related to potential confounders. The first is the cross-sectional study, in which disease rates in different locations are correlated with pollution levels in those settings, looking at a single point in time. Although these studies can generally be

conducted using only publicly available information, they can suffer both from confounding issues and from the ecological fallacy, which is the incorrect assumption that relationships evaluated from large groups are applicable to individuals. A classic example of the ecological fallacy involved a study that showed that whereas there was a correlation between race and illiteracy rates across states, this relationship was drastically reduced when considering individuals rather than states.

In terms of confounding issues, individual behaviors with important linkages to health (such as smoking status or occupation) cannot be incorporated into a cross-sectional study, making it difficult to interpret the causal association between air pollution and health. Therefore, although many of the original air pollution epidemiological studies were cross-sectional because of the availability of information, these studies are conducted less often at present.

Air pollution health effects are also determined through time-series studies, which are studies that compare changes in daily or multiday air pollution levels with changes in daily or multiday health outcomes (such as emergency room visits or deaths). These studies therefore yield information about the influence of air pollution on health over a relatively short number of days (known as acute health effects). Time-series studies have the strength of having a relatively limited set of potential confounding variables, because any confounder must be correlated with both air pollution levels and health impacts on a temporal basis. For example, cigarette smoking could only confound the relationship in a time-series study if more people smoked on high-pollution days than on low-pollution days, which is unlikely. However, cigarette smoking could act as an effect modifier, a variable that influences the relationship between exposure and disease but does not distort the true relationship between exposure and disease. Plausible confounders include weather variables such as temperature and relative humidity, which can affect both pollutant formation and population health, as well as air pollutants other than those studied or measured. Because time-series studies rely largely on publicly available information and are informative about the relationship between air pollution and health, the majority of air pollution epidemiological studies are time-series studies.

Despite the numerous advantages of time-series studies, they cannot provide information about the effects of longer term exposures to air pollution (known as chronic health effects). Time-series studies can expand the time window under consideration beyond periods of days, but seasonal trends in health risks and statistical limitations of time-series studies make it impossible to evaluate the effects of daily exposure beyond about a couple of months. Information about long-term health risks is primarily taken from retrospective or prospective cohort studies, in which a group of individuals is tracked for a significant period of time (often decades) to evaluate whether their air pollution exposures are linked to higher risks of mortality or chronic disease. Because individual-level data about nonpollution risk factors can be collected, confounders can be dealt with in a more substantial way than in cross-sectional studies. However, there remain more plausible confounders for a cohort study than for a time-series study (including lifestyle factors such as smoking, diet, or occupation, which can theoretically be spatially correlated with air pollution exposures). Statistical methods are used to evaluate the independent effects of air pollution given the levels of other risk factors, which addresses this issue to an extent. The long time horizon to conduct a cohort study implies that fewer of these studies can be done, but the long-term exposure focus means that the health impacts estimated in cohort studies would be expected to be greater than the health impacts in time-series studies (which do not capture the effects of long-term exposures).

In addition, it should be noted that other types of studies provide evidence about the health effects of air pollution. Human chamber studies, in which people are exposed to a pollutant for a short period of time in a controlled laboratory setting, can help corroborate relationships observed in epidemiological studies (particularly for short-term and reversible health effects). Other epidemiological study designs can also be informative, such as case-crossover studies (in which individuals act as their own controls, comparing exposures when the health outcome of concern occurred with exposure levels during a comparable point in time) or panel studies (in which subjects are monitored intensively over a short period of time to evaluate the effects of exposure on outcomes such as daily symptoms or lung function).

Regardless of the type of toxicological or epidemiological information, a crucial final step in determining the health effects of an air pollutant is to evaluate whether the evidence is sufficient to infer a causal relationship rather than a simple association. Although proper statistical control of confounders is an important first step, inferences about

causality require evidence that is outside of the domain of individual studies. To make this evaluation more systematic, Sir Austin Bradford Hill proposed nine "causal criteria" that could be applied. These include strength of the association, consistency across studies or techniques, specificity of the health effect, temporality (exposure preceding disease), biological gradient (higher exposures leading to greater health risks), biological plausibility, coherence with known facts about the disease, supporting experimental evidence, or analogy with other exposures. Although these provide a useful basis for analysis, Hill believed that these should not be interpreted as rules, all of which must be followed, but rather as a framework in which evidence about causality can be interpreted.

3. KEY POLLUTANTS AND HEALTH OUTCOMES

A number of air pollutants have been connected with human health impacts in epidemiological and/or toxicological studies. These pollutants are conventionally divided into two categories:

1. Criteria pollutants: particulate matter (PM), ozone (O_3), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), carbon monoxide (CO), and lead (Pb), as defined by the United States Environmental Protection Agency (U.S. EPA) in the Clean Air Act. These pollutants are generally ubiquitous and have a range of health impacts at typical ambient levels.
2. Toxic air pollutants: also known as "hazardous air pollutants," substances considered to cause cancer or to lead to other potential noncancer effects that might include reproductive, developmental, or neurological impacts. The U.S. EPA maintains a list of hundreds of toxic air pollutants that are anticipated to impact human health.

As might be expected, there is substantially more epidemiological evidence for criteria air pollutants than for toxic air pollutants, given the generally higher levels and shorter term effects (as well as the existence of monitoring networks necessary for most epidemiological investigations). In general, criteria air pollutants have been associated with respiratory or cardiovascular outcomes ranging from symptom exacerbation to premature death, with some limited evidence for lung cancer effects. Toxic air pollutants have largely been associated with cancer through a combination of toxicological and occupational epidemiology studies, along with miscellaneous neurological or reproductive end points. Another important difference between the two pollutant categories is that although it is assumed that cancer risks from air toxics have no threshold (meaning that any level of exposure would lead to some risk of cancer), criteria air pollutants or noncancer risks from air toxics are assumed to have thresholds (levels below which no population health impacts would occur) unless it can be demonstrated otherwise.

3.1 Particulate Matter

Of the criteria pollutants, particulate matter has received the greatest attention within the scientific community to date. PM can be simply defined as any solid or liquid substance suspended in the air. As such, PM can include many different substances, ranging from sea salt, to elemental carbon directly emitted by diesel vehicles, to sulfate particles secondarily formed from SO_2 emissions. Air quality regulations have focused historically on the size of particles rather than on the chemical composition, given the complexity of particulate matter, the difficulty of measuring numerous constituents, and the physiological importance of particle size.

Originally, regulations in the United States and elsewhere focused on total suspended particles (TSPs), which included airborne particles of all sizes. However, the largest particles seemed unlikely to cause significant health impacts, because they remain suspended in the air for brief periods of time and cannot travel into the lungs (see Fig. 1). In 1987, the U.S. EPA modified the particulate matter National Ambient Air Quality Standard (NAAQS) to focus on PM_{10}, defined as the fraction of particles less than 10 μm in aerodynamic diameter. PM_{10} has also been referred to as inhalable particles, because these particles are sufficiently small to enter the respiratory system. The size fraction of interest was further constrained by the U.S. EPA in 1997, when they proposed to change the focus of the NAAQS to $PM_{2.5}$, the fraction of particles less than 2.5 μm in aerodynamic diameter. These smaller particles, known as fine particles or respirable particles, are most able to reach the lower portions of the lungs, where gas exchange occurs.

Historically, a large amount of the evidence of particulate matter health effects has come from epidemiological studies, in part because of the regulatory emphasis on total particulate matter rather than chemical constituents. A number of time-series studies have evaluated the relationship

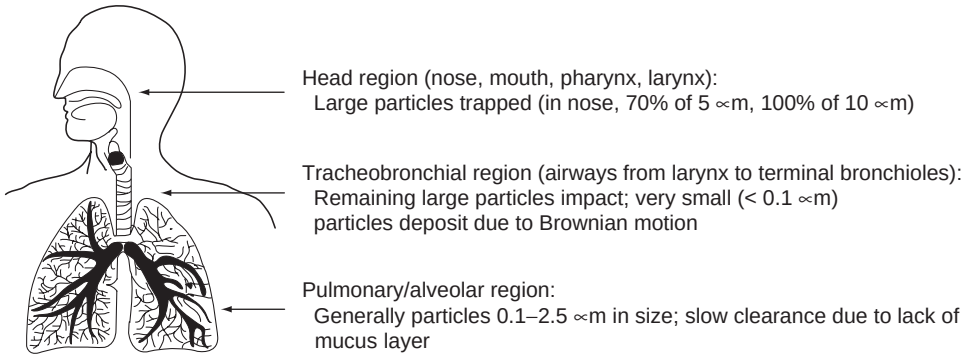


FIGURE 1 Relationship between particle size and site of particle deposition in the lung.

between daily changes in particulate matter concentrations and daily changes in number of deaths, generally within a few days of the pollution increment. These studies have been conducted in both developed and developing country settings and in cities with a range of pollution types or levels, but they have largely demonstrated a consistent relationship between PM and mortality even when the effects of weather or other air pollutants are taken into account. Across the numerous studies that evaluated the relationship between PM_{10} and daily mortality, estimates generally have been on the order of a 0.5–1% increase in deaths per 10 $\mu g/m^3$ increase in daily average PM_{10} concentrations (although with selected studies yielding either nonsignificant or greater effects). These estimates have been similar in cities with different levels of other air pollutants, different sources of particulate matter, and different correlations between other pollutants and particulate matter, lending support to the hypothesis that particulate matter rather than other pollutants or confounders is responsible for this observed relationship. As anticipated, in most studies, the effects were greater for respiratory or cardiovascular causes of death than for other causes of death.

A crucial issue for interpretation of time-series air pollution studies is whether they represent a significant loss of life expectancy or whether the deaths are simply a function of "harvesting," whereby individuals who would have otherwise died the next day are dying one day earlier due to acute pollution exposure. If this were the case, then the death rate might increase above the expected level on one day but drop below the expected level the next day, with no net increase in deaths over the week following the pollution increase. Studies that have investigated the harvesting question have found that the relationship between PM and mortality does not vanish as the time period after the exposure increases, indicating that harvesting is unlikely.

A final question from the time-series mortality literature is whether a threshold exists, and, if so, at what concentration. Studies have generally not found evidence of a population threshold within the range of ambient concentrations measured, although a threshold could exist at levels lower than those observed in the epidemiological studies.

In contrast, there have been relatively fewer cohort mortality studies in the published literature, given the resource intensity of these investigations. Although multiple cohort mortality studies exist, two studies published in the mid-1990s were among the earliest completed and provided an important basis for subsequent regulatory decisions and assessments of the health effects of air pollution. The Six Cities Study tracked 8111 white adults who lived in six cities in the eastern half of the United States for 15–17 years. In each city, at the start of the study, the research team set up a centrally located monitoring station, where concentrations of both particulate matter of various sizes and gaseous pollutants were measured. As a cohort study, this study was able to collect detailed information about individual behaviors that could influence health, such as smoking, obesity, and occupational exposures.

Controlling for these and other plausible confounders, the investigators found that mortality rates were significantly higher in the high-pollution cities compared with the low-pollution cities, with the strongest associations found for three different measures of particulate matter (PM_{10}, $PM_{2.5}$, and sulfate particles). The relationship for SO_2 was weaker than the PM relationships, and there was no association between mortality and ozone. Air pollution was significantly associated with death from cardiovascular and respiratory disease, was

positively associated with lung cancer (although not statistically significantly so), and was not associated with other causes of death. The relative risk for mortality was reported to be 1.26 for an 18.6 $\mu g/m^3$ increase in $PM_{2.5}$ concentrations (95% confidence interval: 1.08, 1.47), with no apparent threshold at the pollution levels documented in the study (between 11.0 and 29.6 $\mu g/m^3$ of $PM_{2.5}$).

To determine if the conclusions from the Six Cities Study were robust, the American Cancer Society cohort study was conducted, in which individuals enrolled in the American Cancer Society Cancer Prevention Study II were matched with the nearest ambient air pollution monitoring data. This cohort contained over 500,000 United States subjects living in all 50 states, with detailed individual risk factor data collected. As in the Six Cities Study, the American Cancer Society study found a significant increase in mortality risks as levels of both $PM_{2.5}$ and sulfate particles increased across cities. Cardiopulmonary mortality was significantly elevated for both pollutants, whereas lung cancer mortality was significantly associated with sulfate particles but not $PM_{2.5}$. The relative risk for all-cause mortality was 1.17 for a 24.5 $\mu g/m^3$ increase in median annual $PM_{2.5}$ concentrations (95% confidence interval: 1.09, 1.26), slightly lower than the relative risk from the Six Cities Study. No threshold for mortality risks was apparent across the range of fine particle concentrations in the study (between 9.0 and 33.5 $\mu g/m^3$ of median annual $PM_{2.5}$ concentrations).

Additional analyses of these two cohorts emphasized the robustness of the findings across different statistical formulations and additional confounders. A detailed reanalysis of both studies by the Health Effects Institute concluded that the findings did not appear to be confounded by non-air-pollution variables and that there was an association between mortality and $PM_{2.5}$ and sulfate particles, as well as with gaseous SO_2. A follow-up to the American Cancer Society study used more monitoring data and more years of population follow-up on the original cohort, along with a refined statistical approach and an increased number of behavioral confounders. The estimated effect of air pollution on mortality was similar to the earlier findings, but the follow-up study documented a statistically significant association between PM exposure and lung cancer (potentially due to the increased sample size given more years of follow-up).

Thus, these two seminal cohort mortality studies documented an effect of air pollution on mortality that is somewhat larger than the effect from the time-series studies (even when both types of studies consider the same particle size) and was associated in large part with particulate matter. Furthermore, the cohort mortality studies likely represent a greater loss of life expectancy potentially related to induction rather than exacerbation of disease (such as lung cancer, which cannot plausibly be related to air pollution in a time-series study). Although inherent limitations of epidemiology make it difficult to disentangle the effects of individual pollutants, and cohort studies have numerous behavioral confounders that must be addressed statistically, the cohort mortality studies supported the time-series mortality studies regarding the role of particulate matter and demonstrated a potentially greater effect for fine particles than for larger particles.

Along with the mortality effects, the epidemiological literature for PM has documented morbidity outcomes with a wide range of severity. Hospital admissions and emergency room visits for cardiovascular or respiratory causes have been studied extensively in time-series investigations, in part due to the publicly available information and the relative ease of defining the health end points. Other acute health end points that have been associated with particulate matter exposure include upper and lower respiratory symptoms, asthma attacks, and days with restricted activities. Long-term particulate matter exposure has also been associated with the development of chronic bronchitis, defined as inflammation of the lining of the bronchial tubes on a chronic basis, with the presence of mucus-producing cough most days of the month for at least 3 months of the year for multiple years, without other underlying disease. These morbidity end points are consistent with the types of premature deaths found in the mortality studies and are therefore supportive of the relationship between particulate matter exposure and respiratory and cardiovascular health.

3.2 Ozone

In contrast with the health evidence for PM, the primary studies documenting ozone health effects have been a combination of toxicological, human chamber, and epidemiological studies. This is in part because, unlike PM, ozone is a single substance with well-understood chemical properties. Ozone is an oxidant gas that is poorly water soluble. Ozone is therefore able to travel throughout the respiratory tract, reacting with molecules on the surface of the lung and leading to pulmonary edema, inflammation, and the destruction of epithelial cells that line the

respiratory tract. These and other physical changes to the lung, many of which are transient but some of which may be permanent, could plausibly lead to a number of health outcomes documented in the epidemiological literature, given sufficient ozone exposure. As in any epidemiological analysis, the crucial question is what level of ozone exposure is sufficient to induce health effects (or whether population health effects are found for any level of ozone exposure) and what the magnitude of the relationship is between ozone exposure and various health outcomes.

As anticipated, epidemiological and human chamber studies have shown that ozone has an array of acute effects on the respiratory system. Short-term exposures to ozone have been associated with significant but reversible decreases in lung function as well as increases in pulmonary inflammation. As would be expected, these physiological changes have resulted in increases in both mild (respiratory symptoms and restrictions of activity) and severe (emergency room visits and hospitalizations) morbidity outcomes, depending on the health of the individual and the magnitude of the exposure.

Ozone has also been associated with premature death in some time-series studies; however, because high-ozone days tend to correspond with hot and humid days (which are associated with higher mortality rates), it can be difficult to determine the relative contributions of air pollution and weather. Ozone may also be correlated with particulate matter and other air pollutants in some settings, impairing the ability to determine the independent role of each pollutant. Finally, ozone does not penetrate indoors as readily as PM, making indoor exposures lower and outdoor monitor-based health evidence harder to interpret. Regardless, considering a subset of published studies that controlled both for other air pollutants (primarily PM) and for the nonlinear effects of weather on health, there appears to be an association between daily average ozone concentrations and premature death, with an approximate 1% increase in deaths per 10 parts per billion (ppb) increase in daily average ozone. Because fewer studies have focused on ozone mortality in detail than have focused on particulate matter mortality, both the existence and the magnitude of this effect are somewhat uncertain.

Ozone may also be associated with chronic, nonreversible health outcomes. Children who exercise outdoors in high-ozone areas have been shown to have increased risk of asthma development, demonstrating potential long-term effects of ozone exposure as well as the importance of breathing rate and time spent outdoors (because ozone levels indoors are low, given the reactivity of ozone). Ozone exposure has also been linked with decreased lung function in young adults who lived for long periods of time in cities with high ozone levels. Although the Six Cities Study and the American Cancer Society study did not document an influence of long-term exposure to ozone on premature death, there is both epidemiological and physiological evidence supporting chronic pulmonary inflammation and other chronic health effects due to low-level ozone exposure.

3.3 Other Criteria Air Pollutants

Although PM and ozone are the two pollutants studied most extensively in many past epidemiological investigations (especially studies of mortality risk), there is also evidence of health effects of other criteria air pollutants. It has been well established that extremely high levels of carbon monoxide (CO) can lead to damage to the central nervous system, eventually leading to death at high enough concentrations. This is because hemoglobin preferentially reacts with CO over oxygen (forming carboxyhemoglobin), thereby reducing the oxygen-carrying capacity of the blood. Symptoms at lower levels of carboxyhemoglobin in the blood (below 30%) include headache, fatigue, and dizziness, with coma, respiratory failure, and subsequent death occurring at levels above 60%.

There has also been evidence that low levels of exposure to CO (corresponding to typical ambient levels in some settings) can lead to cardiovascular impairment. Time-series investigations have correlated cardiovascular hospital admissions with exposure to CO even when controlling for other pollutants, although the degree to which these effects are independent of the effects of other pollutants is difficult to determine. However, controlled human chamber studies have demonstrated that carboxyhemoglobin levels as low as 2% are associated with cardiovascular effects such as anginal pain or changes in electrocardiogram measures, providing support for the epidemiological findings.

Similarly, there is clear evidence that exposure to extremely high levels of nitrogen dioxide can lead to respiratory distress, with symptoms such as cough, shortness of breath, and chest tightness occurring given short-term exposures on the order of 1 part per million (ppm; a level rarely seen in outdoor air but occasionally found in poorly ventilated indoor settings containing combustion sources).

These effects occur for reasons similar to those involved in ozone health impacts, with the insoluble nature of nitrogen oxides coupled with their oxidative properties leading to damage to epithelial cells and other forms of cytotoxicity.

At lower levels of air pollution, epidemiological studies have documented linkages between NO_2 exposures and outcomes such as decreased lung function or increased respiratory infections. A number of studies have investigated the relationship between NO_2 exposure and respiratory illness in children, often using the presence of gas stoves as a proxy for elevated exposures. These studies generally found that acute lower respiratory symptoms in children increased with incremental increases in NO_2 concentrations, down to fairly low levels. Children and individuals with preexisting respiratory disease have been most strongly linked with respiratory outcomes associated with NO_2 exposure.

As mentioned previously, there was an association between sulfur dioxide and premature mortality in the American Cancer Society cohort study. However, it was difficult to determine the relative importance of gaseous SO_2 versus $PM_{2.5}$ or secondary sulfate particles, and this finding has generally not been interpreted as a definitive causal relationship, given the lack of supporting evidence for the lethality of low-level SO_2 exposure. Sulfur dioxide has been established as an irritant to the eyes and upper respiratory tract at high concentrations (multiple ppm). The site of the documented respiratory tract effect differs somewhat from the sites hypothesized for ozone or nitrogen oxides. This is because SO_2 is highly water soluble and can therefore react in the upper respiratory tract, making that the site of its irritant effects.

Considering acute SO_2 exposures at typical ambient concentrations, high-risk populations include children with mild to moderate asthma and others with preexisting respiratory disease. Short-term exposures to SO_2 concentrations on the order of 0.25–0.5 ppm have been associated with bronchoconstriction and subsequent decreases in lung function and increases in respiratory symptoms, for those with asthma or other respiratory diseases. In agreement with these findings, some epidemiological studies have associated respiratory hospital admissions or emergency room visits with SO_2, although the evidence supporting these more severe outcomes has not been substantial. However, it is clear that sulfur dioxide and its related secondary by-products have potential health impacts at higher pollution levels.

Finally, although the health effects of airborne lead have been drastically reduced in much of the developed world due to the use of unleaded fuels, outdoor airborne lead can have significant health impacts near industrial sites (such as lead smelters) or in countries that still use a significant portion of leaded fuel. Although extremely high levels of blood lead have been associated with severe central nervous system problems, the more prevalent health impacts, which have been documented at blood lead levels as low as 10 μg/dl or lower, include incremental increases in blood pressure and incremental decreases in the IQ of children. Although both of these health end points are only incrementally affected by lead at generally seen blood lead levels, the population health implications of small shifts in blood pressure or IQ can be significant.

3.4 Toxic Air Pollutants

Given the numerous toxic air pollutants (189 compounds are listed in Section 112 of the 1990 U.S. Clean Air Act Amendments) with both cancer and noncancer health effects, the anticipated health effects of each of these pollutants cannot be described in detail here. However, the severity of the carcinogenic effects of a toxic air pollutant can generally be described in one of three ways—its cancer classification, its inhalation unit risk, or its population cancer risk.

Because evidence about carcinogenicity can arise from a number of sources, the U.S. EPA, the International Agency for Research on Cancer (IARC), and other similar organizations have developed methods to categorize the weight of evidence for carcinogenicity. As indicated in Table I, both the U.S. EPA and IARC evaluate evidence from animal and human studies, as well as available information about the mechanism of action of the compound, to draw conclusions about the likelihood that the substance is carcinogenic. In general, few substances have been categorized as known human carcinogens (EPA Category A, IARC Category 1), because this requires statistically significant epidemiological evidence, which is difficult to obtain, given the need for studies that follow exposed populations for decades, accurately characterize long-term exposures, and adequately control for potential confounders for cancer development (such as smoking). Similarly, few substances have been determined to be noncarcinogens, because most substances tested (which is only a subset of those substances to which humans are exposed) have some positive evidence of carcinogenicity in some test or species.

Once the likelihood of carcinogenicity has been established, the inhalation unit risk is calculated, given evidence from animal and human studies. The

TABLE I

Cancer Classification Systems by the U.S. Environmental Protection Agency (EPA) and International Agency for Research on Cancer (IARC)

Category	Description	Animal evidence	Human evidence
U.S. EPA			
A	Known carcinogen	Any	Sufficient
B1	Probable carcinogen	Sufficient	Limited
B2	Probable carcinogen	Sufficient	Inadequate or none
C	Possible carcinogen	Limited	Inadequate or none
D	Not classifiable	Inadequate	Inadequate or none
E	Noncarcinogen	None	None
IARC			
1	Known carcinogen	Any Sufficient	Sufficient Strong relevant mechanism
2A	Probable carcinogen	Sufficient	Limited Inadequate + strong relevant mechanism
2B	Possible carcinogen	< Sufficient Sufficient	Limited Inadequate
3	Not classifiable	Inadequate or limited	Inadequate
4	Probable noncarcinogen	Negative	Negative Inadequate

inhalation unit risk refers to the excess lifetime cancer risk estimated to result from continuous lifetime exposure to the given substance at a concentration of $1\,\mu g/m^3$ in the air. For most compounds, the inhalation unit risk must be calculated from animal studies that use much higher doses than are commonly found for human populations. Statistical techniques based in part on hypothesized mechanisms of action for cancer development are used to determine appropriate inhalation unit risks at low doses, but this analytical process contains multiple steps that contribute uncertainty to the final quantitative estimate.

As is clear from the definition of inhalation unit risk, the population cancer risk from a toxic air pollutant will be a function of both the inhalation unit risk and the level of population exposure to the substance. Investigations by the U.S. EPA provide some indication of the outdoor air toxics that contributed most substantially to cancer risk in the United States in 1990, based on estimated inhalation unit risks and modeled outdoor concentrations of the pollutants. In total, five substances contributed more than 75% of the total estimated cancer risk from outdoor air toxics—polycyclic organic matter, 1,3-butadiene, formaldehyde, benzene, and chromium. Along with their relatively high inhalation unit risks and ambient concentrations, these substances are all considered probable or known human carcinogens.

Although this captures some of the most substantial health effects from air toxics, there are a few important omissions. The analytical approach just described focused on cancer risks; however, many air toxics also have noncancer risks that can be substantial. A prominent example is mercury, for which there is substantial evidence of neurological impairment (such as impaired motor skills or cognitive function) but limited evidence of carcinogenic effects. For these noncancer effects, the general analytical approach is to determine whether an adverse health effect is likely to occur, rather than quantifying a unit risk factor. This is done by determining a reference dose from animal or human studies, i.e., an estimate (with uncertainty spanning perhaps an order of magnitude) of the daily exposure to the human population (including sensitive subgroups) that is likely to be without an appreciable risk of adverse effects during a lifetime. As indicated by the careful language in the definition, this reference dose is not meant to be a definitive bright line, in part because of the number of uncertainties embedded in its calculation (i.e., extrapolation from short-term studies of genetically homogeneous rats to health effects from long-term exposures in heterogeneous human populations).

IV. HEALTH EFFECTS OF INDOOR AIR POLLUTION

The information discussed thus far was obtained from studies that largely focused on the effects of

pollution generated outdoors (such as from power plants or motor vehicles) on population health, which includes exposure during time spent outdoors as well as exposure indoors from pollution penetrating from the outdoors. However, most of these studies did not consider indoor-generated pollution. Indoor-generated air pollution can be extremely important in both developing and developed country settings, both because of the amount of time people spend at home and because, in a confined space, pollution levels can quickly reach extremely high levels when a highly emitting fuel is burned indoors.

The uncontrolled combustion of biofuels, ranging from wood, to animal waste, to crop residues, to coal, is prevalent among poorer households in developing countries, with about 50% of the global population relying on these energy sources for heating and cooking. Numerous pollutants are emitted at high rates from biofuel combustion, including particulate matter, carbon monoxide, nitrogen oxides, and carcinogens such as formaldehyde or benzo[*a*]pyrene. The high emission rates into confined spaces result in extremely high indoor concentrations, with daily average levels as high as 50 ppm of CO and 3000 μg/m^3 of PM_{10} (versus U.S. EPA primary health standards of 9 ppm of CO and 150 μg/m^3 of PM_{10}). As a result, rates of multiple diseases are elevated in homes burning biofuels, including acute respiratory infections, chronic obstructive pulmonary disease, cancer, and low birth weight. It has been estimated that indoor air pollution is responsible for nearly 3 million deaths per year globally, approximately two-thirds of which occur in rural areas of developing countries.

Although the health effects of indoor air quality and its linkage with energy are predominantly a developing country story, the developed world is not immune to these problems. Homes that have been tightened to improve energy efficiency can have reduced ventilation rates and subsequent increases in indoor pollution concentrations. As already mentioned, the use of gas stoves can increase indoor concentrations of nitrogen dioxide, which has been associated with increased respiratory disease in children. In addition, for many air toxics, exposure to indoor sources can exceed exposure to outdoor sources, due to the use of cleaning products, off-gassing from building materials, and other emission sources.

V. CONCLUSIONS

Although the evidence supporting the health impacts of air pollution is varied and evolving over time, some general conclusions can be drawn based on existing information:

1. At current levels of air pollution in most developed and developing country settings, there is evidence of both mortality and morbidity effects. However, extreme episodes are rare and the relative risk of health effects of air pollution is low enough to necessitate incorporating evidence from a number of types of studies (including epidemiological and toxicological studies) to determine the magnitude of the effect and the likelihood of causality.
2. For criteria air pollutants, the most extensive evidence is related to particulate matter, which has been associated with premature death and cardiopulmonary morbidity due to both short-term and long-term exposures. However, uncertainties exist regarding the components of particulate matter causally linked with health outcomes and the relative contributions of other criteria pollutants to the public health burden of air pollution. The fact that long-term cohort studies have been conducted only in developed countries also contributes uncertainty in extrapolating the findings to locations with much higher levels of pollution.
3. In most settings, a subset of toxic air pollutants contributes a substantial portion of the population cancer risk from outdoor air toxic emissions, with polycyclic organic matter, 1,3-butadiene, formaldehyde, benzene, and chromium among the most substantial contributors in the United States at present.
4. Indoor air pollution, compared to outdoor air pollution, could contribute more significantly to health impacts, given the amount of time spent indoors and the higher concentrations of combustion-related pollutants in indoor settings with limited ventilation. This problem is most severe in rural settings in developing countries, where the direct burning of biofuels has been associated with respiratory disease and other health end points.

SEE ALSO THE FOLLOWING ARTICLES

Acid Deposition and Energy Use • *Air Pollution from Energy Production and Use* • *Clean Air Markets* • *Climate Change and Public Health: Emerging Infectious Diseases* • *Gasoline Additives and Public Health* • *Indoor Air Quality in Industrial Nations* • *Radiation, Risks and Health Impacts of* • *Thermal Pollution*

Further Reading

Bruce, N., Perez-Padilla, R., and Albalak, R. (2000). Indoor air pollution in developing countries: A major environmental and public health challenge. *Bull. World Health Org.* **78,** 1078–1092.

Dockery, D. W., Pope, C. A. 3rd, Xu, X., Spengler, J. D., Ware, J. H., Fay, M. E., Ferris, B. G. Jr., and Speizer, F. E. (1993). An association between air pollution and mortality in six U.S. cities. *N. Engl. J. Med.* **329,** 1753–1759.

Hill, A. B. (1965). The environment and disease: Association or causation? *Proc. R. Soc. Med.* **58,** 295–300.

Holgate, S. T., Samet, J. M., Koren, H. S., and Maynard, R. L. (Eds.). (1999). "Air Pollution and Health." Academic Press, San Diego.

Krewski, D., Burnett, R. T., Goldberg, M. S., Hoover, K., Siemiatycki, J., Jarrett, M., Abrahamowicz, M., and White, W. H. (2000). "Particle Epidemiology Reanalysis Project. Part II: Sensitivity Analyses." Health Effects Institute, Cambridge, MA.

Levy, J. I., Carrothers, T. J., Tuomisto, J., Hammitt, J. K., and Evans, J. S. (2001). Assessing the public health benefits of reduced ozone concentrations. *Environ. Health Perspect.* **109,** 1215–1226.

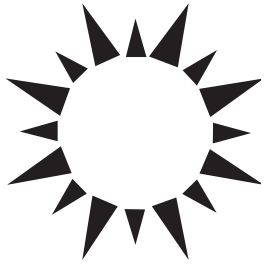
Pope, C. A. III, Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., Ito, K., and Thurston, G. D. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *J. Am. Med. Assoc.* **287,** 1132–1141.

Pope, C. A. III, Thun, M. J., Namboodiri, M. M., Dockery, D. W., Evans, J. S., Speizer, F. E., and Heath, C. W. Jr. (1995). Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults. *Am. J. Respir. Crit. Care Med.* **151,** 669–674.

Thurston, G. D., and Ito, K. (2001). Epidemiological studies of acute ozone exposures and mortality. *J. Expo. Anal. Environ. Epidemiol.* **11,** 286–294.

Woodruff, T. J., Caldwell, J., Cogliano, V. J., and Axelrad, D. A. (2000). Estimating cancer risk from outdoor concentrations of hazardous air pollutants in 1999. *Environ. Res. (Sect. A)* **82,** 194–206.

Alternative Transportation Fuels: Contemporary Case Studies



S. T. COELHO and J. GOLDEMBERG
University of São Paulo
São Paulo, Brazil

1. Introduction
2. General Overview
3. Case Studies: Ethanol Use Worldwide
4. Flex-Fuel Vehicles

Glossary

biodiesel An ester that can be made from substances such as vegetable oils and animal fats. Biodiesel can either be used in its pure state or blended with conventional diesel fuel derived from petroleum.

ethanol Ethyl alcohol (CH_3CH_2OH); one of a group of chemical compounds (alcohols) composed of molecules that contain a hydroxyl group (OH) bonded to a carbon atom. Ethanol is produced through the fermentation of agricultural products such as sugarcane, corn, and manioc. Most of the ethanol produced in the world is from sugarcane, mainly in Brazil. In the United States, ethanol is made from corn.

gasohol A mixture of 20–26% anhydrous ethanol (99.6° Gay-Lussac and 0.4% water) and gasoline; used in most vehicles in Brazil.

methyl *tert*-butyl ether (MTBE) A chemical compound that is manufactured by the chemical reaction of methanol and isobutylene. MTBE is produced in very large quantities (over 3.18×10^6 liters/day in the United States in 1999) and is almost exclusively used as a fuel additive in motor gasoline. It is one of a group of chemicals commonly known as "oxygenates" because they raise the oxygen content of gasoline. At room temperature, MTBE is a volatile, flammable, and colorless liquid that dissolves rather easily in water.

Organization for Economic Cooperation and Development (OECD) An international body composed of 30 member nations. Its objective is to coordinate economic and development policies of the member nations.

transesterification The reaction of a fat or oil with an alcohol in the presence of a catalyst to produce glycerin and esters or biodiesel. The alcohol, which carries a positive charge to assist in quick conversion, is recovered for reuse. The catalyst is usually sodium hydroxide or potassium hydroxide.

Alternative transportation fuels are those fuels not derived from petroleum. They include not only those derived from renewable sources but also natural gas and its derived fuels. A general overview of the existing status of development of alternative fuels for transportation throughout the world is presented in this article. Long-term experiences, such as the Brazilian Alcohol Program and several programs in the United States, as well as those in other countries, are also discussed.

1. INTRODUCTION

The perspective that envisions exhaustion of the world's oil reserves is not the main reason for researching and developing alternative fuels; rather, it is for environmental and strategic reasons that alternative fuel technology is of importance. For example, big cities such as São Paulo, Mexico City, and Delhi (among others) currently face serious environmental problems due to the pollutant emissions from vehicles. Some countries have proposed and enacted solutions for these problems. Ambient concentrations of lead in the São Paulo metropolitan region dropped more than 10 times from 1978 to 1991, far below the air quality standard maximum, due to the mandatory use of either an ethanol/gasoline blend or straight ethanol in all cars, legislated through the Brazilian Alcohol Program (further details are discussed in Section 3.1). Strategic aspects of alternative fuel development are also significant due to the fact that most oil reserves are

 © 2004 Elsevier Inc. All rights reserved.

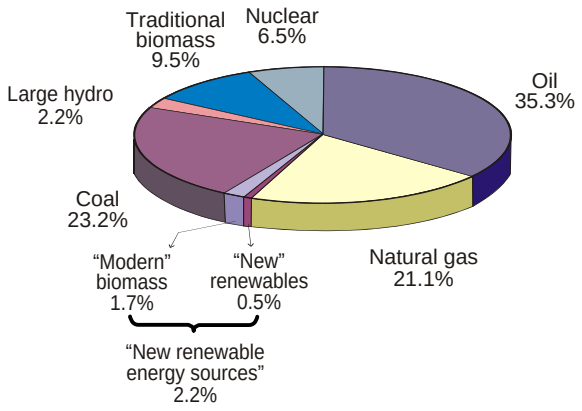


FIGURE 1 World consumption of primary energy and renewables, by energy type, 1998.

in the Middle East, a region facing complex political conflicts. The current trends show that the world will continue to depend on fossil fuels for decades; however, because the largest share of the world's oil resources is concentrated in regions with potential or active political or economic instabilities, alternative fuels ease the complexity of worldwide distribution of necessary energy resources. Nuclear energy plants, although an alternative to fossil fuels, are also concentrated only in a few countries and nuclear technology raises numerous concerns on physical security grounds.

Organization for Economic Cooperation and Development (OECD) countries, which account for 80% of the world economic activity, are quite dependent on oil imports, with a 63% share of global oil consumption (expected to rise to 76% in 2020). Asian Pacific countries are expected to increase the external dependence of their energy requirements to 72% by 2005. Compared to fossil and nuclear fuels, renewable energy resources are more evenly distributed, although only 2.2% of the world energy supply in 1998 was from new renewable sources. New renewable sources include modern biomass, small hydropower, geothermal energy, wind energy, solar (including photovoltaic) energy, and marine energy. Natural gas accounts for 21.1% of this supply (Fig. 1).

2. GENERAL OVERVIEW

The alternative fuels being used in the world today include biodiesel, electricity, ethanol, hydrogen, methanol, natural gas, propane, and solar energy. In 1999, the U.S. Department of Energy finalized an amendment to the Energy Policy Act of 1992 that added certain blends of methyltetrahydrofuran (MeTHF), ethanol, and light hydrocarbons to the list of approved "alternative fuels." These liquid products for spark-ignited engines have come to be known under the registered trademark, "P-series." The discussion here is limited to the fuels already being commercialized.

2.1 Biodiesel

Biodiesel is an ester that can be made from substances such as vegetable oils and animal fats. Biodiesel can either be used in its pure state or blended with conventional diesel fuel derived from petroleum. Vegetable oil was used as a diesel fuel as early as 1900, when Rudolf Diesel demonstrated that a diesel engine could run on peanut oil. For vegetable oils to be used as fuel for conventional diesel engines, the oils must be further processed, primarily because of their high viscosity. Transesterification (production of the ester) of vegetable oils or animal fats, using alcohol in the presence of a catalyst, is the most popular process. For every 100 units of biodiesel fuel produced using this method, there are 11 units of glycerin as a by-product. Glycerin is used in such products as hand creams, toothpaste, and lubricants. Another biodiesel production process in limited use involves cold-pressed rapeseed oil, but no glycerin by-product is produced. Alternatively, unprocessed vegetable oils can be used in modified diesel engines. Such engines have limited production and are therefore more expensive, although their numbers are increasing in Europe.

The main benefits of biodiesel can be categorized as strategic (increased energy self-sufficiency for oil-importing countries), economic (increased demand for domestic agricultural products), and environmental (biodegradability and improved air quality, particularly lower sulfur emissions and almost null carbon balance). Exhaust emission improvements include substantial reduction in carbon monoxide, hydrocarbons, and particulates, although the production of nitrogen gases can be similar to that of regular diesel fuel (depending of the diesel quality). The United States, New Zealand, Canada, and several European Union countries have conducted extensive tests of biodiesel in trucks, cars, locomotives, buses, tractors, and small boats. Testing has included the use of pure biodiesel and various blends with conventional diesel. Among the developing countries, the Biodiesel Program started in Brazil in 2002 was notable for its goal to replace part of the diesel consumption in the Brazilian transportation

sector. Presently, the major disadvantage of biodiesel is its high production cost. As discussed in Section 3.1, in the case of Brazil, biodiesel forecasted costs are higher than diesel costs.

2.1.1 Biodiesel in the United States

Interest in biodiesel in the United States was stimulated by the Clean Air Act of 1990, combined with regulations requiring reduced sulfur content in diesel fuel and reduced diesel exhaust emission. The Energy Policy Act of 1992 established a goal of replacing 10% of motor fuels with petroleum alternatives by the year 2000 (a goal that was not reached) and increasing to 30% by the year 2010.

2.1.2 Biodiesel in Europe

Two factors have contributed to an aggressive expansion of the European biodiesel industry. Reform of the Common Agricultural Policy to reduce agricultural surpluses was of primary importance. This policy, which provides a substantial subsidy to non-food crop production, stimulated the use of land for non-food purposes. Secondarily, high fuel taxes in European countries normally constitute 50% or more of the retail price of diesel fuel. In 1995, Western Europe biodiesel production capacity was 1.1 million tons/year, mainly produced through the transesterification process. This added over 80,000 tons of glycerin by-products to the market annually, creating created a large surplus. Germany thus decided to limit the production of biodiesel using the transesterification process. When it is not possible to market the glycerin by-product, one method of disposal of the excess is incineration; however, this creates an environmental risk and results in additional costs. Germany is now focusing on biodiesel production using the cold-pressed rapeseed method to avoid the problem of excess glycerin.

2.1.3 Biodiesel in Japan

In early 1995, Japan decided to explore the feasibility of biodiesel by initiating a 3-year study. Biodiesel plants using recycled vegetable oils collected in the Tokyo were planned; 10% of federal vehicles were expected to use alternative fuels to set an example for the private automotive and fuel industries. The 3-year study indicated production costs for biodiesel in Japan are 2.5 times that of petroleum diesel. The program has only recently reach its objectives.

2.1.4 Biodiesel in Canada

In the early 1990s, Canadian canola production increased in response to higher market prices relative to cereal grains and increased grain handling and transportation costs. Canola production peaked in 1994 and 1995, limited by suitable land base and crop rotational requirements. However, higher yield due to new *Brassica juncea* varieties and improved chemical weed control may further increase production in the medium term. There is a potential for the use of lower quality canola oils derived from overheated or frost-damaged seed without any ill effects on biodiesel quality.

2.1.5 Biodiesel in Brazil

In 1998, several initiatives were implemented in Brazil, aiming to introduce biodiesel into the Brazilian energy matrix. The initiatives included (1) tests performed in South Brazil, using the so-called B20 blend (20% ester and 80% diesel oil), in specific fleets of urban buses, (2) the building of a small-scale pilot plant for biodiesel production from fat and palm oil (largely produced in North Brazil), and (3) laboratory-scale production and tests of biodiesel using soybean oil/sugarcane ethanol. The Brazilian federal government subsequently decided to establish a work group of specialists from all involved sectors, creating the National Biodiesel Program in 2002. This program will mainly analyze the use of surplus of soybean oil, which is produced on a large scale in Brazil and is presently facing, export barriers.

The economic competitiveness of biodiesel and diesel oil has been evaluated; studies in Brazil show that biodiesel production costs are higher than diesel costs (Table I).

TABLE I

Production Cost for Methyl Ester (1 Ton) from Soy Oil in Brazil[a]

	Amount			
Input	**Kilograms**	**Tons**	**Price (R $/ton)**	**Cost**
Soy oil	1015	1.015	169.64	172.19
Methanol[b]	140	0.14	93.57	13.10
Catalyzer[c]	12	0.012	125.00	1.50
Input cost				186.79
Production cost (115% of input cost)				401.59
Total cost (U.S. $/ton)				588.38

[a] Data from the Brazilian Reference Center on Biomass (CENBIO). Exchange rate, Brazilian real (R)/U.S. dollar, $2.8/1 (July 2002). Diesel price in São Paulo pump stations is around R $1.00/liter (about U.S. $350/ton).

[b] Methanol cost based on prices at pump stations in California (U.S. $0.88–1.10/gallon).

[c] Catalyzer cost based on Brazilian market price.

2.2 Electricity

Electricity is not, technically speaking, a fuel, but it is used in existing alternative technologies for powering transportation. Electricity can be used to power electric and fuel cell (discussed later) vehicles. When used to power electric vehicles (EVs), electricity is stored in an energy storage battery, which must be replenished by plugging the vehicle into a recharging unit. The electricity for recharging the batteries can come from the existing power grid (from large hydroelectric or thermoelectric power plants), or from distributed renewable sources, such as biomass, small hydro, solar, or wind energy. The main benefits to electric-powered vehicles are the lower pollutant emissions at the point of use, although the emissions generated in the electricity production process at the power plants or from the fuel reform reaction (when the fuel cells use hydrogen produced by the reform reaction, as discussed later) can be indirectly attributed to EVs. Economic aspects of using EVs include the high initial capital cost, which can be partially offset by the lower maintenance costs. When compared to the cost of gasoline, the cost of an equivalent amount of fuel for an EV is lower. Maintenance costs for EVs are lower because EVs have fewer moving parts to service and replace.

2.3 Ethanol

Ethanol (ethyl alcohol, CH_3CH_2OH) is one of a group of chemical compounds (alcohols) with molecules that contain a hydroxyl group (OH) bonded to a carbon atom. Ethanol is produced through the fermentation of agricultural products such as sugarcane, corn, and manioc, among others. Most ethanol produced worldwide is from sugarcane, mainly in Brazil. In the United States, ethanol is made from corn.

Ethanol is used as a high-octane fuel in vehicles. More than 4 million cars run on pure, hydrated ethanol in Brazil, and all gasoline in the country is blended with anhydrous ethanol (20–26% ethanol), as a result of a government program to make ethanol from sugarcane, in place since the 1970s. In the United States, there is a similar program being started and the number of vehicles using ethanol is increasing.

Ethanol makes an excellent motor fuel: it has a motor octane number that exceeds that of gasoline and a vapor pressure that is lower than that of gasoline, which results in lower evaporative emission. Ethanol's flammability in air is also much lower than that of gasoline, which reduces the number and severity of vehicle fires. Anhydrous ethanol has lower and higher heating values of 21.2 and 23.4 megajoules (MJ)/liter, respectively; for gasoline the values are 30.1 and 34.9 MJ/liter. Because ethanol in Brazil is produced from sugarcane, it has the lowest production cost in the world. This is due not only to high agricultural and industrial productivity levels, but also to the extremely favorable energy balance of the alcohol production. In the United States, ethanol is produced from corn and represents a large consumption of fossil fuels, with much lower energy balance, despite the existing controversy among specialists. Table II shows a comparison of the ethanol energy balance in Brazil (from sugarcane) and the United States (from corn). Ethanol can also be produced from cellulose feedstock (corn stalks, rice straws, sugarcane bagasse, etc.), through a process still under development.

In the United States, because of support from corn-growing states and the U.S. Departments of Energy and Agriculture, ethanol-fueled vehicle production is increasing. Auto manufacturers began in 1997 to produce cars and pickup trucks that could use either ethanol or gasoline. These flexible fuel (or flex-fuel) vehicles are discussed in Section 4.

2.4 Hydrogen

Hydrogen (H_2) is a gas that has considerable potential as an alternative fuel for transportation but, at this point, little market presence. The most important use of hydrogen is expected to be in electric fuel cell vehicles in the future. Fuel cell cars are in development by most major manufacturers, but hydrogen still lacks a wide distribution infrastructure. The current emphasis is on the use of hydrogen to supply the fuel cells that power electric vehicles. Fuel cells produce electricity. Similar to a battery, a fuel cell converts energy produced by a chemical reaction directly into usable electric power. However, unlike a battery, a fuel cell needs an external fuel source—typically hydrogen gas—and generates electricity as long as fuel is supplied, meaning that it never needs electrical recharging. Inside most fuel cells, oxygen and hydrogen from a fuel tank combine (in an electrochemical device) to produce electricity and warm water. As a simple electrochemical device, a fuel cell does not actually burn fuel, allowing it to operate pollution free. This also makes a fuel cell quiet, dependable, and very fuel efficient.

An impediment to the use of fuel cells is the hydrogen production and storage. The predominant

TABLE II

Energy Balance of Ethanol[a] from Sugarcane and from Corn

Product	Source	Fossil fuel consumption in ethanol production (MJ/liter of ethanol)[b]	Final energy balance[c]
Ethanol from sugarcane[d]	Macedo (2000)	1.89[h]	11.2:1
Ethanol from corn			
Existing plants[e]	Pimentel (1991)	42.6	Negative
Laboratory tests[f]	Pimentel (1991)	25.6	Negative
Existing plants[e]	Unnasch (2000)	11.8	1.8:1
With allocation to coproducts[g]	Unnasch (2000)	7.19	2.9:1

[a] Lower ethanol energy value: 21.2 MJ/liter.

[b] Fossil fuel consumption in ethanol production corresponds to diesel oil use in the agricultural phase and during transport, to coal and natural gas use in corn-based ethanol plants, and to natural gas use for fertilizer production.

[c] Final energy balance corresponds to the low energy content in 1 liter of ethanol divided by the total fossil fuel consumption to produce 1 liter of ethanol.

[d] Includes bagasse surplus production (Brazil); see footnote b.

[e] Large plants in the United States.

[f] Using membrane technology.

[g] Considering the coproduction of corn oil and animal feed products, the allocation of energy inputs corresponds to 1096 kcal/liter of ethanol.

[h] In sugarcane-origin ethanol plants, there is no fossil fuel consumption in the plant (all fuel consumed in the plant is sugarcane bagasse, the by-product of sugarcane crushing); fossil fuel consumption corresponds to the agricultural phase and fertilizer production.

method of making hydrogen today involves using natural gas as a feedstock. Petroleum-based fuels, including gasoline and diesel, can also be used, but this may compromise a major objective behind alternative fuels, i.e., to reduce oil consumption. Hydrogen production is through the "reform reaction" of an existing fuel (natural gas, methanol, ethanol, or naphtha), a chemical reaction that "extracts" the hydrogen from the fuel, producing a gas mixture of carbon monoxide (CO) and hydrogen. The hydrogen must be separated from the CO to be fed into the fuel cell. This reaction can be performed in a stationary system (and the vehicle will carry a high-pressure hydrogen storage tank) or in an on-board reformer/fuel cell system (and the vehicle will carry a conventional fuel tank to feed the system). Both possibilities are now under study in several countries. According to some specialists, ethanol should be the fuel of choice for fuel cells due to its lower emissions and because it is produced from a renewable source (biomass).

Though fuel cells have been widely publicized in recent years, they are not new: the first one was produced in 1839. Fuel cells powered the Gemini spacecraft in the 1960s, continue to power the Space Shuttle, and have been used by the National Aeronautics and Space Administration (NASA) on many other space missions. Although their operation is simple, they have been quite expensive to make. Extensive research and development have promised the widespread use of fuel cells in the near future. All major auto companies have fuel cell-powered vehicles in the works, and a nascent fuel cell industry is growing rapidly. By the kilowatt, fuel cells still cost more today than do conventional power sources (from $1000 upto $5000/kW, U.S. dollars), but an increasing number of companies are choosing fuel cells because of their many benefits, and large-scale production is expected to make fuel cell costs decline.

According to the U.S. National Renewable Energy Laboratory, hydrogen prices vary widely, depending on the transport distance and the type of hydrogen (from 17 to 55 cents/100 cubic feet), but cost figures for hydrogen use as transportation fuel are not yet available.

2.5 Methanol

Methanol (or methyl alcohol) is an alcohol (CH_3OH) that has been used as alternative fuel in flexible fuel vehicles that run on M85 (a blend of 85% methanol and 15% gasoline). However, methanol is not commonly used nowadays because car manufacturers are no longer building methanol-powered vehicles. Methanol can also be used to make methyl-*tert*-butyl ether (MTBE), an oxygenate that is blended with gasoline to enhance octane and reduce pollutant emissions. However, MTBE

production and use have declined due to the fact that MTBE contaminates groundwater. In the future, methanol may be an important fuel, in addition to ethanol, to produce the hydrogen necessary to power fuel cell vehicles; such a process is now under development.

Methanol is predominantly produced by steam reforming of natural gas to create a synthesis gas, which is then fed into a reactor vessel in the presence of a catalyst to produce methanol and water vapor. Although a variety of feedstocks other than natural gas can and have been used, today's economics favor natural gas. Synthesis gas refers to combinations of carbon monoxide (CO) and hydrogen (H_2). Similar to ethanol vehicles, methanol-powered vehicles emit smaller amounts of air pollutants, such as hydrocarbons, particulate matter, and NO_x, than do gasoline-powered vehicles. However, the handling of methanol is much more dangerous compared to ethanol due to its high negative impacts on health.

World demand for methanol is around 32 million tons/year and is increasing modestly by about 2–3%/year, but with significant changes in the industry profile. Since the early 1980s, larger plants using new, efficient low-pressure technologies have replaced less efficient small facilities. Demand patterns have been changing in Europe; methanol was once blended into gasoline (when its cost was around one-half of that of gasoline), but now is not competitive with lower oil prices. Offsetting this was the phasing out of leaded gasoline in developed countries, mandating the use of reformulated gasoline. The United States promoted the use of MTBE derived from methanol. The United States produces almost one-quarter of the world's supply of methanol, but there is a significant surplus of methanol in the world. Figure 2 shows the world methanol supply, indicating this significant excess capacity.

The largest market for methanol in the United States is for the production of MTBE; there are nearly 50 U.S. plants. It is estimated that 3.3 billion gallons of MTBE was used in 1996 for blending into clean, reformulated gasoline serving 30% of the U.S. gasoline market. MTBE displaces 10 times more gasoline than all other alternative vehicle fuels combined.

Methanol prices in the United States have varied significantly since 1989; they doubled by 1994 and returned to the 1993 levels by 1996. In 2002, methanol prices reached 64 cents/gallon, slightly higher than European prices (61.7 cents/gallon) and Asian prices (55.9 cents/gallon).

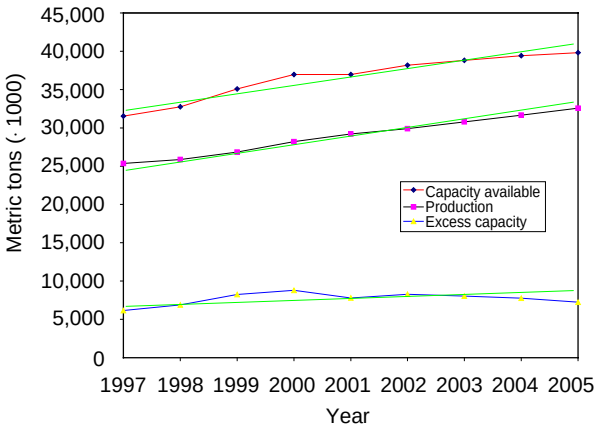


FIGURE 2 World methanol supply. Data from the Methanol Institute (www.methanol.org).

2.6 Natural Gas

Natural gas is a mixture of hydrocarbons, mainly methane (CH_4), but it also contains hydrocarbons such as ethane (C_2H_4) and propane (C_3H_8) and other gases such as nitrogen, helium, carbon dioxide, hydrogen sulfide, and water vapor. It is produced either from gas wells or in conjunction with crude oil production (associated natural gas). Natural gas is consumed in the residential, commercial, industrial, and utility markets. The interest in natural gas as an alternative fuel comes mainly from its clean burning properties, with extremely low sulfur and particulate contents. Also, carbon emissions from natural gas are lower compared to those from other fossil fuels, such as oil and coal. Because of the gaseous nature of natural gas, it must be stored on board a vehicle in either a compressed gaseous state (compressed natural gas) or in a liquefied state (liquefied natural gas). Natural gas is mainly delivered through pipeline systems.

Natural gas reserves are more evenly distributed worldwide than are those of oil. According to BP Global, reserves at the end of 2000 totaled 5304 trillion cubic feet, with 37.8% of the reserves located in the former Soviet Union, 35% in the Middle East, 6.8% in Asia Pacific, 6.8% in North America, 4.6% in South and Central America, and 3.5% in Europe. By contrast, crude oil reserves are heavily concentrated in the Middle East, which alone held 65.3% of oil reserves at the end of 2000. Global proved natural gas reserves available in 2000 would last more than 60 years at the 2000 rate of production. Although production and consumption are likely to increase over time, past trends suggest that reserves will increase as well, especially with

improvements to the technology employed to find and produce natural gas.

Consumption of natural gas for fuel increased more than consumption of any other fuel in 2000, with global consumption rising by 4.8%, the highest rate since 1996. This was driven by growth in consumption of 5.1% in the United States and Canada, which together represent more than 30% of world demand. Chinese consumption increased by 16%, although China still represents only 1% of world consumption. In the former Soviet Union, growth of 2.9% was the highest for a decade.

Gas production rose by 4.3% worldwide, more than double the average of the last decade. This growth is mainly due developments in the United States and Canada, where production rose by 3.7%, and the fastest since 1994. Growth exceeded 50% in Turkmenistan, Nigeria, and Oman, and 10% in 11 other countries. Russia, the second largest producer, saw output decline by 1.1%. North America produced and consumed in 2000 more than 70 billion cubic feet (bcf)/day. The former Soviet union produced about 68 bcf/day and consumed about 53 bcf/day. Europe produced less than 30 bcf/day but consumed about 1.5 times more. The balance is practically even in South and Central America (with a little less than 10 bcf/day) and is almost even in Asia Pacific (about 30 bcf/day; consumption higher than production) and the Middle East (about 20 bcf/day; production a little higher than consumption). Africa produced more than 10 bcf/day but consumed less than half of it.

In the U.S. market, natural gas prices are about $1.5/million British thermal units (MMBtu) above oil spot prices, which ranged from $2.1 to $3.5/MMBtu by the end of 2002 (according to the U.S. Department of Energy; information is available on their Web site at www.eia.doe.gov).

2.7 Propane

Liquefied petroleum gas (LPG) consists mainly of propane (C_3H_8) with other hydrocarbons such as propylene, butane, and butylene, in various mixtures. However, in general, this mixture is mainly propane. The components of LPG are gases at normal temperatures and pressures. Propane-powered vehicles reportedly have less carbon build-up compared to gasoline- and diesel-powered vehicles.

LPG is a by-product from two sources: natural gas processing and crude oil refining. When natural gas is produced, it contains methane and other light hydrocarbons that are separated in a gas-processing plant. The natural gas liquid components recovered during processing include ethane, propane, and butane, as well as heavier hydrocarbons. Propane and butane, along with other gases, are also produced during crude refining as a by-product of the processes that rearrange and/or break down molecular structure to obtain more desirable petroleum compounds.

The propane market is a global market. Approximately 1.3 billion barrels of propane are produced worldwide. Although the United States is the largest consumer of propane, Asian consumption is growing fast. According to the World LP Gas Association, during 1999 China achieved a growth rate of over 20% in their consumption of propane, largely in the residential/commercial sector. Other notable increases were recorded in India, Iran, and South Korea, which are rebounding from the Asian economic crisis.

2.8 Solar Energy

Solar energy technologies use sunlight to produce heat and electricity. Electricity produced by solar energy through photovoltaic technologies can be used in conventional electric vehicles. Using solar energy directly to power vehicles has been investigated primarily for competition and demonstration vehicles. Solar vehicles are not available to the general public, and are not currently being considered for production. Pure solar energy is 100% renewable and a vehicle run on this fuel emits no pollutants.

3. CASE STUDIES: ETHANOL USE WORLDWIDE

3.1 The Brazilian Experience

The Brazilian Alcohol Program (PROALCOOL) to produce ethanol from sugarcane was established during the 1970s, due to the oil crises, aiming to reduce oil imports as well as to be a solution to the problem of the fluctuating sugar prices in the international market. The program has strong positive environmental, economic, and social aspects, and has become the most important biomass energy program in the world. In 1970, some 50 million tons of sugarcane were produced, yielding approximately 5 million tons of sugar. In 2002, sugarcane production reached 300 million tons, yielding 19 million tons of sugar and 12 billion liters of alcohol (ethanol). In 2002, the total land area

covered by sugarcane plantations in Brazil was approximately 4.2 million hectares (60% in the state of São Paulo, where sugarcane has replaced, to a large extent, traditional coffee plantations). The average productivity of sugarcane crops in Brazil is 70 tons/hectare, but in São Paulo State there are mills with a productivity of 100 tons of cane per hectare.

Ethanol production costs were close to $100/barrel in the initial stages of the program in 1980. After that, they fell rapidly, to half that value in 1990, due to economies of scale and technological progress, followed by a slower decline in recent years. Considering the hard currency saved by avoiding oil importation through the displacement of gasoline by ethanol, it is possible to demonstrate that the Alcohol Program has been an efficient way of exchanging dollar debt for national currency subsidies, which were paid by the liquid fossil fuel users.

The decision to use sugarcane to produce ethanol in addition to sugar was a political and economic one that involved government investments. Such decision was taken in Brazil in 1975, when the federal government decided to encourage the production of alcohol to replace gasoline, with the idea of reducing petroleum imports, which were putting great constraints in the external trade balance. At that time, sugar price in the international market was declining very rapidly and it became advantageous to shift from sugar to alcohol production. Between 1975 and 1985, the production of sugarcane quadrupled and alcohol became a very important fuel used in the country. In 2002, there were 321 units producing sugar and/or alcohol (232 in central–south Brazil and 89 in northeast Brazil). An official evaluation of the total amount of investments in the agricultural and industrial sectors for production of ethanol for automotive use concluded that in the period 1975–1989, a total of $4.92 billion (in 2001 U.S. dollars) was invested in the program. Savings on oil imports reached $43.5 billion (2001 U.S. dollars) from 1975 to 2000.

In Brazil, ethanol is used in one of two ways: (1) as an octane enhancer in gasoline in the form of 20–26% anhydrous ethanol (99.6° Gay-Lussac and 0.4% water) and gasoline, in a mixture called gasohol, or (2) in neat-ethanol engines in the form of hydrated ethanol at 95.5° Gay-Lussac. Increased production and use of ethanol as a fuel in Brazil were made possible by three government actions during the launching of the ethanol program. First, it was decided that the state-owned oil company, Petrobrás, must purchase a guaranteed amount of ethanol; second, economic incentives were offered to agro-

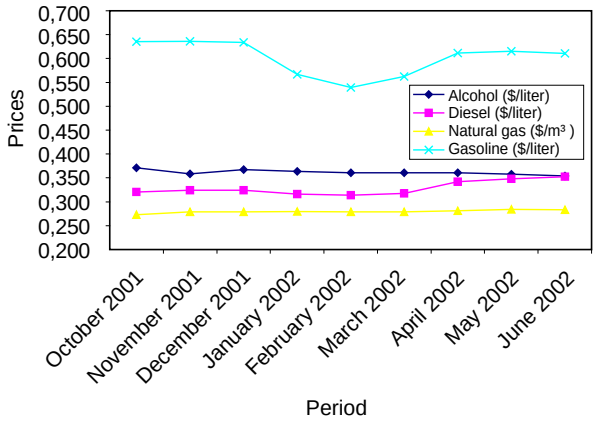


FIGURE 3 Transportation fuel prices in Brazil (in U.S. dollars).

industrial enterprises willing to produce ethanol, in the form of loans with low interest rates, from 1980 to 1985; third, steps were taken to make ethanol attractive to consumers, by selling it at the pump for 59% of the price of gasoline. This was possible because the government at that time set gasoline prices.

The subsidies for ethanol production have been discontinued and ethanol is sold for 60–70% of the price of gasoline at the pump station, due to significant reduction of production costs. These results show the economic competitiveness of ethanol when compared to gasoline. Considering the higher consumption rates for net-ethanol cars, ethanol prices at the station could be as much as 80% of gasoline prices. Fig. 3 shows a comparison of the different transportation fuels in Brazil.

3.1.1 Impact of Alcohol Engines on Air Pollution

All gasoline used in Brazil is blended with 25% anhydrous ethanol, a renewable fuel with lower toxicity, compared to fossil fuels. In addition to the alcohol–gasoline (gasohol) vehicles, there is a 3.5 million-vehicle fleet running on pure hydrated ethanol in Brazil, 2.2 million of which are in the São Paulo metropolitan region. Initially, lead additives were reduced as the amount of alcohol in the gasoline was increased, and lead was completely eliminated by 1991. Aromatic hydrocarbons (such as benzene), which are particularly toxic, were also eliminated and the sulfur content was reduced as well. In pure ethanol cars, sulfur emissions were eliminated, which has a double dividend. The simple substitution of alcohol for lead in commercial gasoline has dropped the total carbon monoxide, hydrocarbon, and sulfur emissions by significant numbers. Due to the ethanol blend, ambient lead concentrations in the São Paulo metropolitan region dropped from 1.4 $\mu g/m^3$ in 1978

to less than 0.10 μg/m^3 in 1991, according to CETESB (the Environmental Company of São Paulo State), far below the air quality standard maximum of 1.5 μg/m^3. Alcohol hydrocarbon exhaust emissions are less toxic, compared to those from gasoline. They present lower atmospheric reactivity and null greenhouse emissions balance (9.2 million tons of carbon dioxide avoided in 2000 due only to the gasoline replacement by ethanol).

One of the drawbacks of the use of pure ethanol is the increase in aldehyde emissions as compared to gasoline or gasohol use. It can be argued, however, that the acetaldehyde from alcohol use is less detrimental to human health and to the environment, compared to the formaldehyde produced when gasoline is used. Total aldehyde emissions from alcohol engines are higher than emissions from gasoline engines, but it must be noted that these are predominantly acetaldehydes. Acetaldehyde emissions produce fewer health effects compared to the formaldehydes emitted from gasoline and diesel engines. Aldehyde ambient concentrations in São Paulo present levels substantially below the reference levels found in the literature. In addition, carbon monoxide (CO) emissions have been drastically reduced: before 1980, when gasoline was the only fuel in use, CO emissions were higher than 50 g/km; they went down to less than 5.8 g/km in 1995.

3.1.2 Social Aspects of the Brazilian Alcohol Program

Social considerations are the real determinants of the impact of the Brazilian alcohol program. Presently, ethanol production generates some 700,000 jobs in Brazil, with a relatively low index of seasonal work. The cost of creating a single job in the ethanol agroindustry is around $15,000 (U.S. dollars), according to recent studies by the São Paulo Sugarcane Agroindustry Union (UNICA). In Brazil, job generation in most other industries requires higher investments, as shown in Fig. 4.

Although Brazilian workers have low incomes relative to workers in developed countries, there has been a significant increase in agricultural mechanization, particularly in the São Paulo region. There are two main reasons for this trend: social changes such as urban migration have reduced the number of workers available to the sugarcane industry, promoting mechanization, and mechanization has proved cheaper than hand labor. Thus, the delicate balance between mechanization and the number and quality of new jobs created by the ethanol industry is likely to remain a key issue for several years.

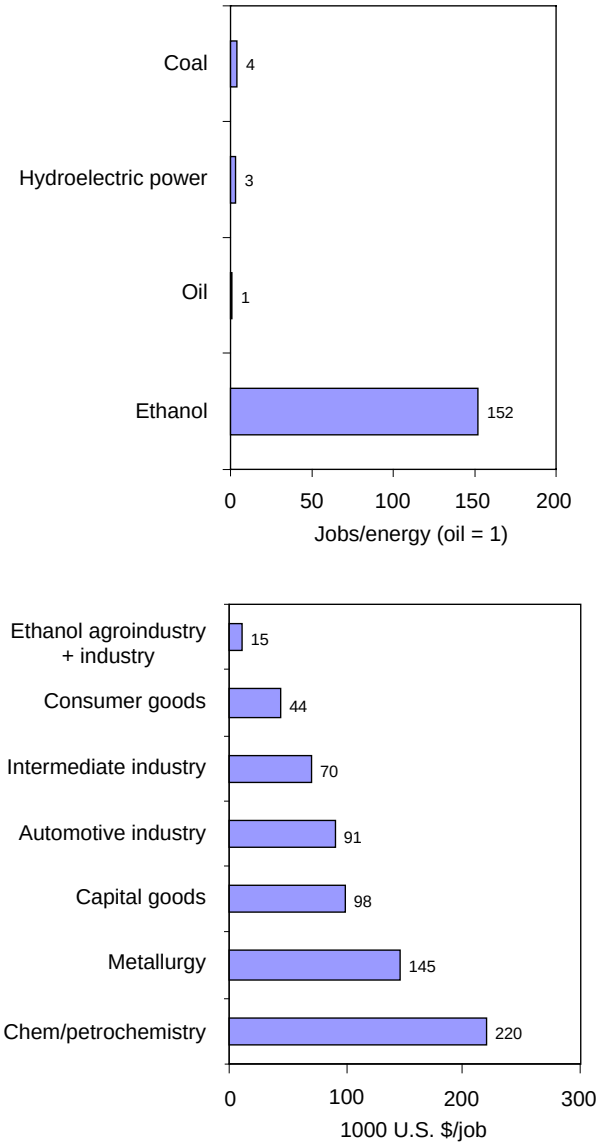


FIGURE 4 Employment numbers (2001) for the Brazilian Ethanol Program. Data from UNICA (2002).

3.1.3 Outlook for the Brazilian Alcohol Program

The use of alcohol as a fuel in Brazil was an extraordinary success until 1990. In 1988, sales of ethanol-powered cars represented 96% of the market; by the end of the decade, 4.5 million automobiles were sold. Since then, however, there has been a precipitous drop in the sale of new ethanol-powered cars, due to the following reasons:

- The price of pure alcohol was set in 1979 at 64.5% of the price of gasoline but increased gradually to 80%, reducing the economic competitiveness.

- The tax on industrialized products (IPI), initially set lower for alcohol-fueled cars, was increased in 1990, when the government launched a program of cheap "popular cars" (motors with a cylinder volume up to 1000 cm^3) for which the IPI tax was reduced to 0.1%. According to the local manufacturers, the "popular cars" could not be easily adapted to use pure alcohol because this would make them more expensive and would take time. Competition between manufacturers required an immediate answer to benefit from the tax abatement.
- Significant lack of confidence in a steady supply of alcohol and the need to import ethanol/methanol to compensate for a reduction in local production due to the increase on sugar exports (high sugar prices in international market).

As a result, the sales of pure ethanol cars dropped almost to zero in 1996. However, in 2001, this scenario started to change: 18,335 units of pure ethanol car were sold, showing an increase in consumer interest. Use of blended ethanol–gasoline has not been affected due the existence of a federal regulation that requires addition of ethanol in all gasoline sold in Brazil.

Other countries adopt special taxes (mainly on gasoline) that are quite substantial (around 80% in Italy and France, around 60% in the United Kingdom, Belgium, Germany, and Austria, and around 30% in the United States). In addition, ethanol as an additive is made viable in some countries by differentiated taxes and exemptions. In the United States, federal taxes on gasoline are 13.6 cents/gallon (3.59 cents/liter). The addition of 1 gallon of ethanol leads to a credit of 54 cents/gallon, or $26.68/barrel (U.S. dollars). Some states have also introduced tax exemptions that are the equivalent of up to $1.10/gallon, or $46.20/barrel. The rationale for these subsidies and exemptions is the same as was used in Brazil in the past, such as the social benefits of generating employment and positive environmental impacts.

An interesting option regarding the future for the ethanol program is its expansion. To guarantee ethanol market expansion, there are two distinct alternatives: increase national demand or create an international alternative liquid fuel market. However, international trade of ethanol still faces several difficulties regarding import policies of developed countries. Finally, it is also important to consider the effect the use of ethanol has in the balance of payments made by Brazil. Since the beginning of the alcohol program in 1975, and until 2000, at least 216 million m^3 of ethanol have been produced, displacing the consumption of 187 million m^3 of gasoline. This means that around $33 billion (U.S. 1996 dollars) in hard currency expenditures were avoided.

The international alternative liquid fuel market is currently an interesting option if developed countries decide effectively to limit CO_2 emissions to satisfy the goals established in the 1992 Rio de Janeiro conference, in the Climate Convention adopted in the United Nations World Summit for Sustainable Development. There are also reasons to develop an alternative to fossil fuels for economic and strategic reasons while promoting better opportunities for farmers. Food production efficiency has improved so much in the past 30 years that around 40–60 million hectares (Mha) of land is now kept out of production to maintain food product prices at a reasonable level able to remunerate farmers in the OECD countries. With the introduction of biomass-derived liquid fuels, a new opportunity exists for utilization of such lands and preservation of rural jobs and employment in industrialized countries. Not all industrial countries have free agricultural areas to produce biofuels, and even those that do will not be able to satisfy fully the demand for biofuel, at a reasonable price and in an environmentally sound way. The total automobile fleet in Western Europe exceeds 200 million units, demanding more than 8 million barrels/day (bl/day) of gasoline. To satisfy such a level of demand with a 10% ethanol blend, production of 0.8 million bl/day is necessary, of which probably two-thirds would be imported. This means an international market of 530,000 bl/day for ethanol from sugarcane, which could be shared mainly by a few developing tropical countries, such as Brazil, India, Cuba, and Thailand. With the creation of such a market, some developing countries could redirect their sugarcane crops to this market instead of to sugar. Sugar is a well-established commodity with a declining real average price in the international market. This is not the case for gasoline, for which the average real price is increasing.

Finally it is worthwhile to add that the international fuel market option has the potential to increase ethanol demand to a level above the level presented here, because (depending on the seriousness of environmental problems) ethanol blends of up to 25% may be marketed in the future, as is already happening in Brazil, as well as the possibility that such a practice will be followed by the diesel oil industry. Brazil's sugarcane industry is recovering from a significant decrease in production in recent

TABLE III

Comparative Prices of Transportation Fuels in Brazil and the United States[a]

	Prices[b]			
	Brazil		United States	
Fuel	Low	High	Low	High
Gasoline	20.92	24.69	11.49	20.38
Natural gas	7.63	7.94	6.21	6.48
Diesel	8.67	9.75	8.81	10.41
Ethanol[c]	16.39	17.76	17.92	21.90
Methanol	na[d]	na[d]	12.93	16.14

[a] Data from Brazilian Petroleum National, the Alternative Fuels Data Center, the U.S. Gas Price Watch, and the Ministry of Science and Technology. Calculations by the Brazilian Reference Center on Biomass.

[b] All prices in U.S. dollars ($/m^3); average prices taken between January and July, 2002. Exchange rate, Brazilian real (R)/U.S. dollar, 2.8/1.

[c] Brazilian ethanol from sugarcane and U.S. ethanol from corn.

[d] na, Not available.

years; in 2002, a sugarcane surplus allowed Brazil to increase ethanol exports. Table III shows comparative prices of gasoline, natural gas, diesel oil, methanol, and ethanol in both Brazil and the United States. The low prices of gasoline in the United States are remarkable compared to Brazilian prices. It also must be noted that ethanol prices in Brazil, in contrast to the United States, do not include any subsidies.

3.2 The U.S. Experience

In 2000, 19.7 million barrels of crude oil and petroleum products were consumed in the United States per day (25% of world production); more than half of this amount was imported. The United States has the lowest energy costs in the world. In 2001, ethanol-blended fuels represented more than 12% of motor gasoline sales, with 1.77 billion gallons produced from corn. Corn is used as the principal feedstock, in contrast to practices in Brazil, where all ethanol is produced from sugarcane, which considerably less expensive than corn. The U.S. Congress established the Federal Ethanol Program in 1979 to stimulate rural economies and reduce the U.S. dependence on imported oil. In 1992, the Energy Policy Act established a goal of replacing 10% of motor fuels with petroleum alternatives by the year 2000, increasing this to 30% by the year 2010.

In 2002, the U.S. Senate passed an energy bill that includes a provision that triples the amount of ethanol to be used in gasoline in the next 10 years. This policy, combined with regulations requiring reduced sulfur content in diesel fuel and reduced diesel exhaust emission, also was designed to foster an interest in biodiesel. Ethanol blended fuel is now marketed throughout the United States as a high-quality octane enhancer and oxygenate, capable of reducing air pollution and improving vehicle performance. The federal administration denied California an oxygenate waiver in 2002, thus ethanol fuel production is expected to grow strongly. According to the Renewable Fuels Association, by the end of 2003, U.S. annual ethanol production capacity will reach 3.5 billion gallons (13.25 billion liters).

California currently uses 3.8 billion gallons/year of MTBE, compared with total U.S. use of 4.5 billion gallons/year. MTBE, like ethanol, is an oxygenate that allows lower pollutant emissions from gasoline. Under federal law, in urban areas with the worst pollution (as in many California cities), gasoline must contain at least 2% oxygen by weight. This requirement applies to about 70% of the gasoline sold in California. The use of MTBE will be discontinued by 2003 in California, being replaced by ethanol, which does not present risks of contamination of water supplies.

In the United States, fuel ethanol production increased from about 150 million gallons in 1980 to more than 1700 million gallons in 2001, according to the Renewable Fuels Association. In the period 1992–2001, the U.S. demand for fuel ethanol increased from 50 to 100 thousand barrels/day, whereas MTBE demand increased from 100 to 270 thousand barrels/day. According to the Renewable Fuels Association, ethanol production in the United States affords several advantages:

- Ethanol production adds $4.5 billion to U.S. farm income annually.
- More than 900,000 farmers are members of ethanol production cooperatives; these cooperatives have been responsible for 50% of new production capacity since 1990.
- Ethanol production provides more than 200,000 direct and indirect jobs.
- Ethanol production reduces the U.S. negative trade balance by $2 billion/year.

In June 2002, ethanol and MTBE prices in the United States were practically equivalent ($1/gallon). Gasoline prices were $0.8/gallon.

3.3 Other Countries

3.3.1 Europe

In the European Union, currently only about 0.01 billion liters of ethanol is used as fuel. In 1994, the European Union decided to allow tax concessions for the development of fuel ethanol and the other biofuels, and as result a number of ethanol projects have been announced in The Netherlands, Sweden, and Spain. France has one of the most developed fuel alcohol programs in the European Union. A 1996 law requires the addition of oxygenate components to fuel, and ethanol was given an early exemption from the gasoline excise tax.

The European Commission (EC) is currently drafting a directive that could force member states to require 2% biofuels in all motor fuels. The directive, if eventually confirmed by the European Parliament and Council, could be enforced as early as 2005. France, Austria, and Germany have already experimented with biodiesel, and other countries, including Sweden, Spain, and France, use ethanol or ethyl *tert*-butyl ether (ETBE) in gasoline. Sweden is the first country in Europe that has implemented a major project introducing E85 fuel and vehicles. Approximately 300 vehicles have been in use in Sweden since the mid-1990s and 40 E85 fueling stations have been established from Malmo, in the south, to Lulea, in the north. In 1999, a consortium composed of the Swedish government, many municipal administrations, companies, and private individuals contacted over 80 different car factories inquiring about the production of an ethanol-fueled vehicle for use in Sweden. A major criterion was that it should not be much more expensive than a gasoline-powered vehicle. Car manufacturers are presently marketing these vehicles. The ethanol fuel will be produced and distributed by all major gasoline companies in Sweden. The production of the blend, 4–5% ethanol in gasoline from the first large-scale plant, will occur at the oil depots in Norrköping, Stockholm, and Södertälje. In the region near these depots, all unleaded 95-octane gasoline will contain ethanol. In 1998, Swedish gasoline consumption was 5.4 million m^3; a 5% ethanol blend means consumption of 270,000 m^3 ethanol annually, i.e., more than five times the capacity of the plant in Norrköping.

3.3.2 Australia

There is a plan under development in Australia that will enable Australia to produce 350 million liters of ethanol annually by 2010. The increase in production, equivalent to approximately 7% of the current petrol market, would total 1% of the liquid fuel market. The Agriculture Minister also intends to ensure that biofuels contribute to 2% of Australia's transportation fuel use by 2010. Under the new plan, a capital subsidy of A$0.16 for each liter of new or expanded biofuel production capacity constructed would be presented to investors. This is equivalent to a subsidy of about 16% on new plant costs. At least five new ethanol distilleries are expected to be set up under the new program, generating a possibility of 2300 construction and 1100 permanent jobs.

The Australian Environment Ministry recently announced that the federal government is trying to determine an appropriate level of ethanol in fuel and would not enforce mandatory levels of ethanol. The federal government has been asked by Australian sugarcane growers to create a plan that would develop a variable and sustainable ethanol industry throughout the country. A producers' working group would assist the government in the production of the plan. The sugarcane growers have also announced plans to launch an educational campaign that will focus on the benefits of ethanol. The cane growers hope to win the support of other farmers throughout the country. Manufacturers announced they would build a 60-million gallon/year ethanol plant in Dalby, Australia, making it the country's largest. The project is expected to be completed by the end of 2003. The plant will process sorghum, wheat, and maize into ethanol that will be blended with petrol sold through 40 or more of the company's service stations.

3.3.3 New Zealand

The Queensland Government has announced that its vehicle fleet will switch to petrol blended with ethanol. Although the switch was welcomed by the Queensland Conservation Council, concerns were expressed about the environmental risks and problems associated with sugarcane farming. The Council noted that they would work to ensure that sugarcane growing is environmentally sound. The Cane Growers' Association believes it is unlikely that the sugar industry will expand due to the government's decision. The Association will not encourage the planting of more cane, because it is likely that the ethanol will be made from the by-product molasses or that it will be sourced out of cereal crops.

3.3.4 South Africa

There are four major ethanol producers in South Africa. The largest ones produce up to 400 million

liters/year of ethanol from coal. This capacity was developed during the 1950s to reduce South Africa's dependence on oil imports during the apartheid era. Synthetic ethanol production can top 400 million liters/year, but usually fluctuates with demand on the world market in general, and on the Brazilian market in particular. Since the mid-1990s, the efforts to introduce coal-derived ethanol into gasoline have repeatedly failed because of petroleum and automobile industry complaints about the low quality of the ethanol. Coal-derived alcohols are not pure ethanol. The original alcohol additive contained only 65% ethanol, which caused significant engine problems. In 1990, the quality issues seem to have been resolved through development of an 85% ethanol blend, and the coal-derived ethanol is used in South Africa as a 12% blend with gasoline. South Africa also produces ethanol from natural gas, about 140 million liters/year, and it is quite expensive. In addition, there are other plants that use molasses as a feedstock.

3.3.5 Thailand

In 2000, Thailand launched a program to mix 10% ethanol with gasoline. Ethanol would be produced from molasses, sugarcane, tapioca, and other agricultural products. The goal was the production of 730 million liters/year by 2002. One of the major sugar groups in Thailand announced that it would invest Bt800 in an ethanol plant in 2002. This group would apply for a license to produce ethanol as an alternative fuel for automobiles. On approval from the National Ethanol Development Committee, the building of the plant was scheduled to begin immediately and it was expected to take 12 to 18 months to complete. The group expects to use molasses, a by-product supplied by the group's sugar mills, to produce 160,000 liters of ethanol/day.

3.3.6 Japan

In order to reduce automobile emissions, Japan is considering introducing a policy that will support the use of blending ethanol with gasoline. The policy is a result of pressures to cut greenhouse gas emissions that lead to global warming. As the second largest consumer of gasoline in the world, Japan has no extra agricultural produce to use for fuel output. Thus, the use of biofuels could create a big export opportunity for ethanol-producing countries. The trading house of Mitsui & Co. is backing ethanol use in Japan. Mitsui estimates Japan's ethanol market at approximately 6 million kiloliters/year at a blending ratio of 10%. Due to an ample supply of low-cost gasoline produced by imported crude oil, the Japanese have yet to use ethanol as a fuel. Japan's interest in green energy increased with the 1997 Kyoto Protocol, which went into effect in 2002. The Kyoto Protocol aims to cut carbon dioxide emissions by 5.2% by 2012.

In order to implement mandatory use of ethanol in Japan, the Environment Ministry must win the support of the oil-refining industry as well as the energy and transportation arms of the government.

3.3.7 Malawi

Malawi has very favorable economic conditions for ethanol. Like Zimbabwe, Malawi was on the forefront of fuel alcohol development; Malawi has blended ethanol with its gasoline continuously since 1982, and has thereby eliminated lead. Because of high freight costs, the wholesale price of gasoline is about 56 cents/liter retail. Moreover, Malawi's molasses has a low value because the cost of shipping it to port for export typically exceeds the world market price. Malawi's Ethanol Company Ltd. produces about 10–12 million liters/year, providing a 15% blend for gasoline.

3.3.8 China

China is also interested in introducing an alcohol program. A pilot program was introduced in the Province of Jilin in 2001, and, in 2002, a Chinese delegation from the Province of Heilonjiand visited Brazil. The Chinese government is concerned about the increase in oil consumption in the country (around 7–7.5%/year between 2000 and 2005) and the alcohol–gasoline blend appears to be an interesting option due to job generation and to the potential for reducing the pollution in large Chinese cities. China finds itself short on fuel and producing only 70% of the nation's demand, and the country is also faced with a sagging economy for farmers. In order to fight both problems, China is considering a new program to launch its first ethanol plant. Despite the numerous projected benefits, the program is not expected to be implemented right away. Although ethanol is environmentally friendly, it is still expensive to produce in China and there are difficulties in transporting it. China hopes that the cost will come down in the next 5 years, making ethanol a viable option for the country.

Despite no laws mandating the use of greener fuels, China has continued to phase out petroleum products since 1998. China is encouraging the use of alternative fuels in their "Tenth Five-Year Plan," which includes trial fuel ethanol production (2001–2005).

3.3.9 India

India is one of the largest producers of sugarcane in the world (300 million tons of sugarcane, similar to Brazilian production) and has vast potential to produce ethanol fuel, which would significantly reduce air pollution and the import of petroleum and its products. The Indian distillery industry has an installed capacity of over 3 billion liters (Brazilian production is in average 13 billion liters of alcohol per year), but operates only at 50% capacity. Also, the Indian sugar industry currently faces an uncertain future, with high stock valuations, unattractive export balances, and narrow operating margins. There are also the issues related to the balance of payments due to oil imports. Therefore, there is a growing interest in ethanol as major business opportunity, mainly to produce anhydrous ethanol to be used in a 5% ethanol–gasoline blend.

The New Delhi government is currently reviewing the excise duty structure on ethanol. The government hopes to make ethanol more attractive to oil companies that blend ethanol with petrol. Government officials are currently concerned about India's sugar farmers, who have found it difficult to sell their crop. As the demand for ethanol-blended petrol increases, so will the demand for sugarcane increase. India produced 18.5 million tons of sugar in the past year. It is expected that, if a trial run proves successful, ethanol-blended petrol (5% ethanol) will be sold throughout the country. The government plans to free all market pricing of petrol.

4. FLEX-FUEL VEHICLES

Flex-fuel vehicles are vehicles that can operate with multiple fuels (or fuel blends). Such technology was created 1980 and there are around 2 million flex-fuel vehicles in the United States today. The main fuels used include gasoline and several alternative fuels, such as pure ethanol (already used in Brazil in automotive vehicles) and blends of ethanol and gasoline. Already in use in Brazil is an ethanol–gasoline; blend at a percentage of 20–26% ethanol. Another blend of ethanol–gasoline (E85), with 85% ethanol, is used in the United States; a blend of methanol–gasoline also used in the United States has 85% methanol. The methanol–gasoline blend has a limited potential for widespread use, because most automobile manufacturers do not build fuel systems compatible with methanol blends. Flex-fuel vehicles have a small processor placed inside the fuel system; the processor detects the fuel blend being used and automatically adjusts the ignition time and the mixture of air and fuel. The greatest advantage of the flex-fuel vehicles is that they can operate with regular gasoline when alternative fuels are not available or are not economically competitive.

Flex-fuel cars in the United States are built to utilize natural gas, pure gasoline, and gasoline blended with a small percentage of ethanol. In Brazil, flex-fuel motors have to be built to accept a much larger percentage of ethanol; the larger percentage of ethanol does negatively affect the life span of the fuel tank and other parts of the engine system, and the design requirements and shortened system life span add to the overall expense of the vehicles.

Acknowledgments

The authors acknowledge the contribution of specialists from the Brazilian Reference Center on Biomass (CENBIO), the University of São Paulo, and the Environmental Company of São Paulo State (CETESB).

SEE ALSO THE FOLLOWING ARTICLES

Biodiesel Fuels • Biomass for Renewable Energy and Fuels • Ethanol Fuel • Hybrid Electric Vehicles • Hydrogen, End Uses and Economics • Internal Combustion (Gasoline and Diesel) Engines • National Energy Policy: Brazil • Natural Gas Processing and Products • Solar Fuels and Materials • Transportation Fuel Alternatives for Highway Vehicles

Further Reading

Goldemberg, J. (2002). "The Brazilian Energy Initiative." World Summit on Sustainable Development, São Paulo, June 2002.

Macedo, I. (2000). O ciclo da cana-de-açúcar e reduções adicionais nas emissões de CO_2 através do uso como combustível da palha da cana. Inventário de Emissões de Gases Efeito Estufa, Report to the International Panel on Climate Change (IPCC), Ministério de Ciência e Tecnologia, March 2000, Brazil.

Moreira, J. R., and Goldemberg, J. (1997). The Alcohol Program, Ministério de Ciência e Tecnologia, June 1997, Brazil.

Pimentel, D. (1991). Ethanol fuels: Energy security, economics and the environment. *J. Agricult. Environ. Ethics* **4**, 1–13.

Pioneira (ed.). (2002). *Energia e Mercados*, March 16, 2002.

São Paulo Sugarcane Agroindustry Union (UNICA). (2001). *Informação Unica* **5**(43), October 2001.

Unnasch, S. (2000). Fuel cycle energy conversion efficiency. Report prepared for the California Energy Commission, June 2000. California Energy Commission, Sacramento.

Winrock International India (WII). (2001). "Ethanol, Sustainable Fuel for the Transport Sector," Vol. 3, October. WII, New Delhi.

Aluminum Production and Energy

H.-G. SCHWARZ
University of Erlangen
Erlangen, Germany

1. **Aluminum Process Chain and Production Figures**
2. **Specific Final Energy Consumption**
3. **Specific Primary Energy Consumption**
4. **Energy as Cost Factor**
5. **Summary**

Glossary

alumina (**aluminum oxide**) Made from bauxite via the Bayer process. Alumina is smelted using the Hall–Héroult process to produce primary aluminum.

aluminum Second most abundant metallic element in the earth's crust after silicon. It is a comparatively new industrial metal that has been produced in commercial quantities for slightly more than 100 years. It is made from alumina via the Hall–Héroult process (primary aluminum). In addition, aluminum is produced from secondary raw materials—that is, from old and/or new scrap (secondary aluminum). Aluminum is used in a wide range of applications, such as automobiles, airplanes, buildings tools, and packaging.

aluminum alloy Aluminum can be alloyed with other materials to make an array of metals with different properties. The main alloying ingredients are silicon, zinc, copper, and magnesium. Aluminum alloys are classified into two groups: cast aluminum and wrought aluminum. Cast aluminum cannot be reshaped plastically and must therefore be cast. The content of alloying ingredients is relatively high. In contrast, wrought aluminum can be reshaped plastically. The content of alloying ingredients is low for wrought aluminum.

bauxite Naturally occurring, heterogeneous material composed primarily of one or more aluminum hydroxide minerals plus various mixtures of silica, iron oxide, titania, and other impurities in trace amounts. Bauxite is converted to alumina via the Bayer process. Bauxite occurs mainly in tropical and subtropical areas, such as Africa, the West Indies, South America, and Australia.

Bayer process Wet chemical caustic leach process to convert bauxite to alumina.

Hall–Héroult process Process of electrolytic reduction in a molten bath of natural and synthetic cryolite to convert alumina to aluminum. The process is energy intensive. The specific electricity consumption is 13–17 MWh per ton of primary aluminum.

recycling share and quota Recycling share is the share of secondary aluminum production of the total aluminum production of 1 year (aluminum, secondary aluminum). In contrast, recycling quota is the share of the actual recycled secondary material of the theoretical available secondary material. It describes how much aluminum is recovered at the end of the use phase of products containing aluminum. The calculation of the recycling quota for one product is based on its aluminum content, lifetime, the consumption for one reference year, and the quantity of recycled aluminum. The total recycling quota of a metal is a weighted average of the quotas of the different products containing the metal. The lifetime is very different for the different aluminum-consuming sectors. The range is 6 months for containers and packaging to approximately 20–40 years for building and construction materials. The average lifetime is approximately 15 years. Only an average lifetime of less than 1 year leads to an identicalness of recycling share and total recycling quota.

secondary aluminum (**recycling aluminum**) Aluminum produced from secondary raw materials (i.e., from old and/or new scrap). Depending on the type of scrap and the desired product quality, different types of furnaces for melting aluminum scrap are used. Scrap for the production of casting alloys is commonly melted in rotary furnaces under a layer of liquid melting salt. Producers of wrought alloys prefer open-hearth furnaces in varying designs (aluminum alloys).

Aluminum is a comparatively new industrial metal that has been produced in commercial quantities for slightly more than 100 years. A necessary precondition was the (independent) discovery of molten-salt

 © 2004 Elsevier Inc. All rights reserved.

electrolysis by American Charles Martin Hall and Frenchman Paul Toussaint Héroult in the late 19th century. World production in 1900 totaled 7339 tons and took place in only four countries. The United States was the leading producer, with a production share of 44%, followed by Switzerland (34%), France (14%), and the United Kingdom (8%). One hundred years later, primary aluminum production has multiplied: In 2000, approximately 24.5 million tons of primary aluminum was produced in 43 countries. The United States remained the world's largest producer and accounted for 15% of the world's production, followed by the former Soviet Union (15%), China (12%), and Canada (10%). Since the inception of the aluminum industry, metal has been recovered from scrap. After World War II, so-called secondary production expanded, especially since the 1960s. In 2000, approximately 8.5 million tons of secondary aluminum was produced. Secondary production accounted for one-fourth of the total global aluminum consumption of 33 million tons.

Aluminum is commonly seen as coagulated energy. In fact, specific energy requirements for primary production of aluminum are high. In the mid- to late 1990s, the specific worldwide average electricity consumption for electrolysis was 15.5 MWh per ton of primary aluminum. The electricity supply for electrolysis represented more than three-fourths of total specific primary energy consumption of approximately 200 GJ per ton of primary aluminum. The specific primary energy consumption of primary aluminum production is at least eight times higher than the consumption caused by the primary production of iron and at least seven times higher than that for the primary production of copper.

The absolute energy consumption figures are also impressive. In 2000, the global production of 24.5 million tons of primary aluminum required 380 TWh of electric power for electrolysis, which represented 2.8% of the total global power generation. The primary energy consumption of 4.90 EJ for global primary aluminum production accounted for 1.2% of the total worldwide primary energy consumption.

This article focuses on the remarkable role of the production factor "energy" in aluminum production. First, the process chain of aluminum production is discussed and the production figures for the Year 2000 are given. Second, the determinants of specific energy consumption and primary energy consumption are presented. Finally, the importance of cost factor energy is analyzed.

1. ALUMINUM PROCESS CHAIN AND PRODUCTION FIGURES

1.1 Survey of Process Chain

The process of primary aluminum production takes place in three stages (Fig. 1). Initially, bauxite is extracted by mining. Then, it is converted to alumina via the Bayer process. According to the worldwide average, 2.3 tons of bauxite is necessary to produce 1 ton of alumina. Primary aluminum is produced via the Hall–Héroult process, a process of electrolytic reduction in a molten bath of natural and synthetic cryolite. Two tons of alumina is necessary to produce 1 ton of primary aluminum.

The primary aluminum produced is alloyed in the cast house; here, cast aluminum and wrought aluminum are produced. The content of alloying ingredients is relatively high for cast aluminum. Cast aluminum cannot be reshaped plastically and must therefore be cast. In contrast, wrought aluminum can be reshaped plastically and its content of alloying ingredients is low. The alloys are converted to castings and semis (sheet plates and profiles). These are used for the production of final products. Figure 2 provides an overview of the consumption of aluminum by domestic end users in the United States, the world's most important aluminum consumer.

Old scrap and new scrap are the initial products of secondary aluminum production. Old scrap consists of

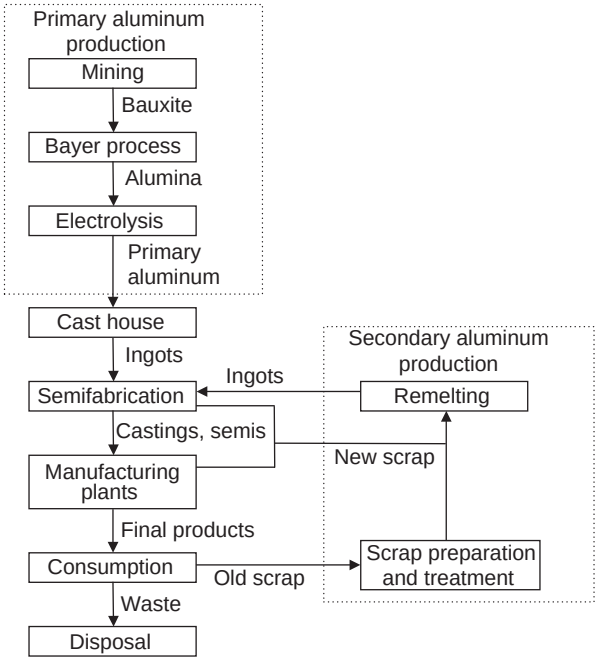


FIGURE 1 Material flow of aluminum production.

products made from aluminum that at the end of their service lives enter the recycling chain as secondary raw materials. New scrap (production scrap) arises during the processing of semifabricated products, mostly clean and sorted, in the form of processing residues (e.g., turnings and residues from punching or stamping). Secondary aluminum production results in predominantly cast aluminum but, increasingly, also wrought aluminum for further processing.

Currently, the alloy elements (e.g., silicon, zinc, copper, and magnesium) within the aluminum scrap cannot be cost-effectively removed. Therefore, in contrast to primary production, the composition of the initial product (scrap) determines to a great extent the possible range of applications. Nevertheless, secondary aluminum is a substitute for primary aluminum, at least as long as cast aluminum, which is predominantly produced, can be sold profitably on the market and therefore the input of secondary aluminum can replace the input of primary aluminum. Thus, it seems reasonable to regard an increase in the recycling share as an important instrument to reduce specific energy consumption in aluminum production.

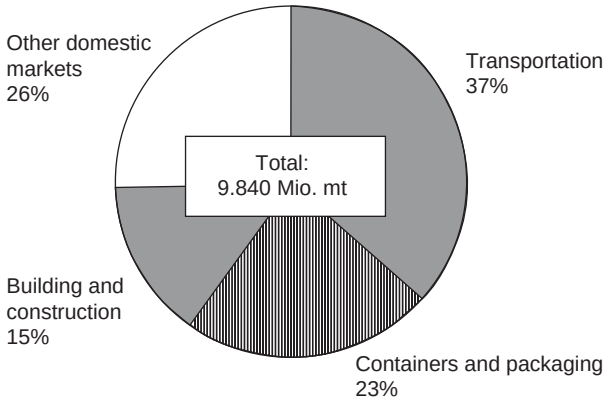


FIGURE 2 Consumption of primary and secondary aluminum by domestic end users in the United States in 2000. Mio mt., million metric tons. From U.S. Geological Survey (2000). "U.S. Geological Survey Minerals Yearbook." U.S. Geological Survey, Reston, VA.

1.2 Primary Aluminum Production

1.2.1 Mining

Bauxite is the most common aluminum ore. Approximately 98% of primary aluminum production is based on bauxite. In the former Soviet Union, aluinite and kaolinite are also used in small proportions for aluminum production. Bauxite is a naturally occurring, heterogeneous material composed primarily of one or more aluminum hydroxide minerals plus various mixtures of silica, iron oxide, titania, and other impurities in trace amounts. The principal aluminum hydroxide minerals found in varying proportions within bauxite are gibbsite (trihydrated) and boehmite and diaspore (both monohydrated). The content of aluminum oxide is 30–60%. Bauxite is generally extracted by open cast mining from strata, typically 4–6 m thick under a shallow covering of topsoil and vegetation (Table I).

Bauxite occurs mainly in tropical and subtropical areas. In 2000, Australia was the leading worldwide bauxite producer with a production share of 39%, followed by Guinea (13%), Brazil (10%), and Jamaica (8%) (Table II). Bauxite reserves are estimated to be 23,000 million tons. They are concentrated in areas that currently dominate production: Australia (share of 24%), Guinea (24%), Brazil (12%), and Jamaica

TABLE I

Characteristics of Important Bauxite Deposits[a]

Bauxite deposits	Reserves (million metric tons)	Content of Al_2O_3 (%)	Average ore thickness (m)	Average surface thickness (m)
Weipa (Australia)	230 reliable + 150 likely	55	2.4	0.8
Gove (Australia)	150 reliable	50	3.5	0.8
Boddington (Australia)	330 reliable + 140 likely	31	6	0.9
Boké (Guinea)	300 reliable	56	7	2
Trombetas (Brazil)	780 reliable	49	6	8
Los Pijiguaos (Venezuela)	177 reliable + 500 likely	49	5	0.5
Linden/Berbice (Guyana)	500 reliable + likely	55–61	3–10	20–60
Dry Harbour Mountain (Jamaica)	110 reliable	48	8	0.3

[a] Adapted with permission from Mori and Adelhardt (2000).

TABLE II

Geography of Aluminum Production and Consumption in 2000[a]

	Primary production (%)			Secondary aluminum production (%)	Aluminum total consumption (%)
	Bauxite	Alumina	Primary aluminium		
Europe	9	22	33	31	>21
France	0	1	2	3	3
Germany	0	2	3	7	8
Italy	0	2	3	7	8
Norway	0	0	4	3	2
United Kingdom	0	0	1	3	3
Former USSR	6	11	15	0	n.a.[b]
Asia	13	11	21	17	>15
China	6	8	12	0	n.a.
India	5	0	3	0	n.a.
Japan	0	2	0	14	12
Korea	0	0	0	1	3
Africa	13	1	5	1	1
Guinea	13	1	0	0	0
South Africa	0	0	3	1	1
America	26	33	34	50	37
Brazil	10	7	5	3	2
Canada	0	2	10	1	3
Jamaica	8	7	0	0	0
Suriname	3	4	0	0	0
United States	0	9	15	41	29
Australia and Oceania	39	31	9	2	2
Australia	39	31	7	1	1
World (1000 Metric tons)	137,547	50,916	24,485	8490	32,401

[a] Adapted with permission from the World Bureau of Metal Statistics (2001), "Metal Statistics." World Bureau of Metal Statistics, Ware, UK.

[b] n.a., not applicable.

(9%). With a global production of 135 million tons in 2000, the static range of coverage is approximately 170 years. Only 4% of mined bauxite is not converted to alumina, and this is used, among other things, in the chemical and cement industry.

1.2.2 Alumina Production

Due to high specific transport costs of bauxite, alumina production often takes place near bauxite mines. In 2000, Australia manufactured 31% of the 50.9 million tons of alumina produced worldwide, follow by the United States (9%), China (8%), and Jamaica (7%) (Table II). Approximately 10% of the alumina produced is not converted to aluminum and is used, among others things, in the ceramic industry.

Approximately 98% of the alumina produced worldwide is manufactured via the Bayer process, which consists of digestion and calcination. Initially, the bauxite is washed and ground. During digestion, the bauxite is dissolved in caustic soda (sodium hydroxide) at high pressure (40–220 bar) and at high temperature (120–300°C). The resulting liquor contains a solution of sodium aluminates and undissolved bauxite residues containing iron, silicon, and titanium. These residues sink gradually to the bottom of the tank and are removed. They are known as red mud. The clear sodium aluminate solution is pumped into a huge tank called a precipitator. Aluminum hydroxide is added to seed the precipitation of aluminum hydroxide as the liquor cools. The

aluminum hydroxide sinks to the bottom of the tank and is removed. It is then passed through a rotary or fluidized calciner at up to 1250°C to drive off the chemically combined water. The result is alumina (aluminum oxide), a white powder.

1.2.3 Electrolysis

The basis for all primary aluminum smelting plants is the Hall–Héroult process (Table II). Alumina is dissolved in an electrolytic bath of molten cryolithe (sodium aluminum fluoride) at 960°C inside a large carbon- or graphite-lined steel container known as a pot. An electric current is passed through the electrolyte at low voltage (3.9–4.7 V) and at very high amperage (100–320 kA). The electric current flows between a carbon anode (positive), made of petroleum coke and pitch, and a cathode (negative), formed by the thick carbon or graphite lining of the pot. The anodes are used up during the process when they react with the oxygen from the alumina. Molten aluminum is deposited at the bottom of the pot and is siphoned off periodically (Fig. 3).

1.3 Secondary Aluminum Production

Secondary aluminum production is concentrated in Western industrial countries, the main aluminum consumers. In 2000, the United States was the world's leading producer with a production share of 41%, followed by Japan (14%), Italy (7%), and Germany (7%) (Table II). Secondary production consists of two process steps: scrap preparation, mixing, and charging, on the one hand, and the processes within the remelter/refiner (remelting, refining, and salt slag preparation), on the other hand.

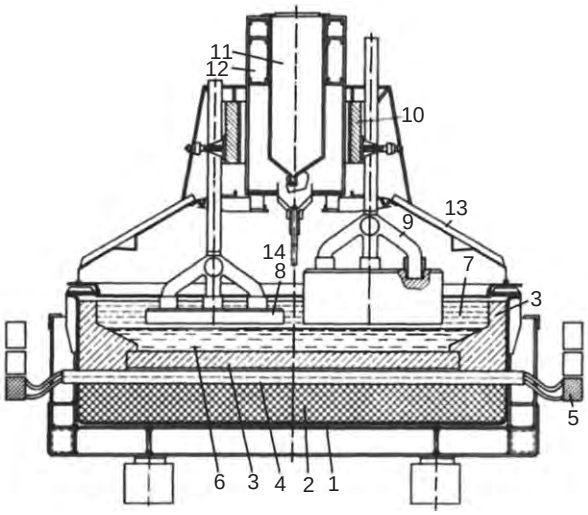


FIGURE 3 Profile of a modern electrolytic cell. 1, steel shell; 2, insulation; 3, cathode (carbon in base and sides; 4, iron cathode bar; 5, cathodic current entry; 6, liquid aluminum; 7, electrolyte; 8, prebaked anodes (carbon); 9, steel nipple; 10, anode beam; 11, aluminum oxide hopper; 12, gas suction unit; 13, detachable covering; 14, crust crusher. From Kammer, C. (1999). "Aluminium Handbook. Volume 1: Fundamentals and Materials." Aluminium-Verlag, Düsseldorf. Reproduced with permission.

1.3.1 Scrap Preparation, Mixing, and Charging

Scraps recovered are treated according to their quality and characteristics. Common treatment processes are sorting, cutting, baling, or shredding. Free iron is removed by magnetic separators. The different (prepared) scrap types are selected and mixed in such a way that their chemical composition is as close as possible to that of the required alloy.

1.3.2 Melting, Refining, and Salt Slag Preparation

Depending on the type of scrap and the desired product quality, different types of furnaces for melting aluminum scrap are used. Scrap for the production of casting alloys is commonly melted in rotary furnaces under a layer of liquid melting salt. A company producing casting alloys from old and new scrap is commonly called a refiner. Producers of wrought alloys prefer open hearth furnaces in varying designs. These furnaces are normally used without salt. Wrought aluminum from mainly clean and sorted wrought alloy scrap is produced in a remelter.

The alloy production in rotary furnaces is followed by a refining process. The molten alloy is fed into a holding furnace (converter) and purified through the addition of refining agents.

After the melting process, the liquid melting salt used in rotary furnaces is removed as salt slag. In the past, the salt slag was land filled. Today, the salt slag is prepared as a rule. The aluminum and the salt within the salt slag are recovered.

2. SPECIFIC FINAL ENERGY CONSUMPTION

2.1 Survey

Table III shows the specific final energy consumption of world aluminum production in the mid- to late 1990s. The electricity consumption for primary production is almost 35 times higher than the consumption for secondary production. The fuel consumption is more than 2.5 times higher. Electrolysis accounts for the main portion (96%) of electricity requirements,

TABLE III

Specific Final Energy Consumption of World Aluminum Production, Mid- to Late 1990s[a]

	Primary production		Total production (75% PA + 25% SA)[b]		Secondary production		
	Electricity (MWh/t_{pa})	Fuels (GJ/t_{pa})	Electricity (MWh/t_{ai})	Fuels (GJ/t_{ai})		Electricity (MWh/t_{sa})	Fuels (GJ/t_{sa})
Transport	<0.1	1.6				0.0	1.0
Processes							
Mining	<0.1	1.0			Scrap preparation and treatment	0.05	2.5
Refining	0.5	27.1			Remelting plant	0.42	8.4
Smelting	15.5	0.0			Remelting	0.32	4.5
					Salt slag treatment	0.09	2.1
					Refining	<0.01	1.8
Main physical inputs							
Caustic soda production	0.1	0.2					
Anode product	<0.1	0.9					
Total	16.2	30.8	12.3	26.1		0.47	11.9

[a] Adapted with permission from Schwarz (2000) and Krone (2000).
[b] PA, primary aluminum; SA, secondary aluminum.

and alumina production accounts for the main portion of fuel consumption (88%) of primary production. The processes within the refiner/remelter dominate in terms of electricity (89%) and fuel consumption (71%) for secondary production.

2.2 Recycling Share

In addition to the specific final energy consumption for the different process steps, the share of secondary production in total production (recycling share) is of great importance for the specific final energy consumption for aluminum production. This is due to the very different final energy requirements for primary and secondary production. Table IV shows the development of total, primary, and secondary production as well as the development of the recycling share since 1950. Aluminum production expanded at high rates in the 1950s and 1960s. Since the 1970s, there has been a noticeable slowdown in growth rates. The recycling share began to increase again in the 1960s and 1970s. In 2000, the recycling share was 26%. This percentage seems to be rather low, but one has to take into consideration the lifetime of aluminum consumer goods—an estimated 15 years. Compared to a total production of 19.9 million tons in 1985, secondary production of 8.5 million tons in 2000 represents a recycling quota of approximately 43%.

(Optimistic) projections estimate a further expansion of aluminum production. In 2030, a total production of 50 million tons is expected, with a recycling share of 44–48% (Table IV). There are two reasons for the projected increase in recycling share: the predicted growth in the recycling quota and the expected continued slowdown in the growth rates of total production. The presented scenarios assume a future growth rate of 1.4% per year compared with 2.6% in the 1980s and 1990s. The decline in the growth rate automatically results in an increase in the recycling share.

If a fixed level of final energy consumption is assumed, a duplication of the recycling share from approximately 25 to 50% would lead to a decrease in specific electricity consumption of 32% and a decrease in fuel consumption of 18%.

TABLE IV

Primary, Secondary, and Total Aluminum Production, 1950–2030(p)[a]

	1950	1960	1970	1980	1990	2000	2030p
Primary production (million tons)	1.5	4.5	10.3	16.1	19.4	24.5	26–28
Secondary production (million tons)	0.4	0.8	2.3	3.7	6.1	8.4	22–24
Total (million tons)	1.9	5.3	12.6	19.7	25.5	32.9	50
Recycling share (%)	22	14	18	19	24	26	44–48

Growth rates (% per year)	**1951–1960**	**1961–1970**	**1971–1980**	**1981–1990**	**1991–2000**	**2001–2030p**
Primary production	11.7	8.5	4.6	1.9	2.4	0.2–0.4
Secondary production	5.8	11.9	4.6	5.3	3.3	3.2–3.5
Total	10.6	9.0	4.6	2.6	2.6	1.4

[a] Adapted with permission from the World Bureau of Metal Statistics, "Metal Statistics" (various volumes), World Bureau of Metal Statistics, Ware, UK; and Altenpohl and Paschen (2001).

2.3 Specific Final Energy Consumption in Primary Production

2.3.1 Transport, Mining, and Alumina Production

The final energy consumption for transport is of little importance. Transport accounts for approximately 5% of the total fuel consumption for primary production. The importance of final energy consumption for bauxite extraction is also minimal. Three percent of the total fuel consumption for primary production is due to mining. Diesel is used as fuel for trucks, scrappers, and excavators. The belt conveyers are responsible for the very low electricity consumption.

Substantial amounts of thermal energy are required for alumina production—for digestion as well as for calcinations. Autoclaves, in almost equal shares of low and high temperature, were predominantly (98%) used for digestion of bauxite in the mid- to late 1990s. Tube reactor use was very low (2%). High-temperature autoclaves and the tube reactor are able to digest all types of bauxite, whereas low-temperature autoclaves can only digest trihydrated bauxite economically. The digestion of monohydrated bauxite would cause a great loss of alumina. On average, the tube reactor offers the best value for energy requirements (and consumption of caustic soda), followed by low-temperature autoclaves. Autoclaves and tube reactors can be combined with a rotary kiln or fluidized bed for calcinations. A fluidized bed is considered the superior technology with higher energy efficiencies. Therefore, rotary kilns are increasingly being replaced by fluidized beds. In the late 1990s, approximately three-fourths of worldwide capacity was based on the fluidized bed.

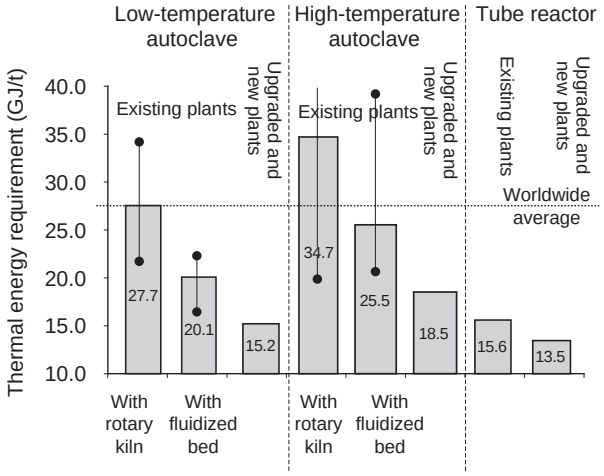


FIGURE 4 Average worldwide specific thermal energy consumption of existing alumina-producing plants (inclusive range for 16 world regions) and for upgrading and new plants (GJ/t primary aluminum). From Schwarz (2000). Reproduced with permission.

Figure 4 provides a review of thermal energy requirements for the different technologies for existing plants on a worldwide average and for upgraded and new plants. Upgraded and new plants require approximately 50% of the thermal energy necessary for existing plants. The upgrading of existing plants and the building of new plants mean that a noticeable decline in specific final energy requirement can be estimated for the next few years. Estimates indicate a decrease in specific thermal energy consumption of approximately 30% in the 15 years from 1995 to 2010.

2.3.2 Electrolysis

The Hall–Héroult process has been the basis for all primary smelting plants for more than 100 years. Indeed, cell design has been improved since the beginning. Modern electrolysis cells work at very high amperage. They use prebaked anodes and are point fed. In addition, process computers are used. Table V shows typical parameters of modern electrolysis cells for the years 1895, 1950, and 2000. At least in the next few years, it is not expected that greater process changes (e.g., the use of inert anodes) will prove to be economically beneficial.

Usually, existing electrolysis cells are characterized in terms of the kinds of anodes used (Soederberg and prebaked technology) and the method of feeding alumina to the smelter (side-worked, center-worked, and point-feeder technology). Soederberg technology uses self-baking anodes. Liquid electrode mass in a steel tank is put into the electrolytic bath. Prebaked technology uses prebaked anodes in which the anodes are baked in a separate furnace. The main advantage of prebaked anodes is their homogeneous structure resulting from better conditions during baking. Due to the homogeneous structure, prebaked anodes have lower electrical resistance and lower specific consumption of anodes in comparison to self-baking anodes.

Side-worked means that the alumina is replenished from the side of the furnace. The crust over the electrolyte is opened by hammer crushers approximately every 3 h until hundreds of kilograms of alumina have been fed into the furnace. Center-worked plants open the crust between the anode lines with beam crushers. The intervals of feeding new alumina can be less than 1 h. Point feeding is the most modern technology. The voltage of the cell is monitored continuously. The measured values allow assessment of how much additional alumina is necessary. The crust is opened by a ram and alumina is added automatically to the smelting pot. This operation is repeated every few minutes. Only a few kilograms of alumina is put into the pot each time to ensure optimal concentration of alumina.

In the mid- to late 1990s, slightly more than 30% of smelting capacity was based on (mostly side-worked) Soederberg technology. The modern point-feeder prebaked (PFPB) technology accounted for more than 40% of smelting capacity. The older center-worked prebaked and side-worked prebaked technologies represented the remaining 30% of smelting capacity. Figure 5 shows the electricity consumption for the different technologies for existing plants as a worldwide average and for upgraded and new plants. The electricity consumption for new plants is only 14% lower than the average consumption for existing plants. Due to the upgrading of old plants with PFPB and the building of new PFPB plants, a moderate decline in specific electricity consumption of electrolysis can be expected. A 10% decline is estimated to occur from 1995 to 2010.

TABLE V

Typical Parameters of Aluminum Reduction Cells—1895, 1950, and 2000[a]

Parameter	1895[b]	1950	2000
Current rating (kA)	9	50–60	300–325
Aluminum production (kg Al/day)	59	385	2475
Unit energy (DC kWh/kg Al)	31	18.5–19.0	12.9–13.5
Potroom worker hours (per ton of Al)	n.a.[c]	5–8	1.7
Interval of Al_2O_3 additions (min)	n.a.	80–240	0.7–1.5

[a] Adapted with permission from Morel (1991) and Æye *et al.* (1999).

[b] Data for a plant of the Pittsburgh Reduction Company (Niagara Falls).

[c] n.a., not applicable.

2.4 Specific Final Energy Consumption in Secondary Production

2.4.1 Transport and Scrap Preparation

Transport accounts for approximately 8% of the total fuel consumption for secondary production due to decentralized scrap recovery and the resulting transport distances. Scrap recovery represents 21% of total

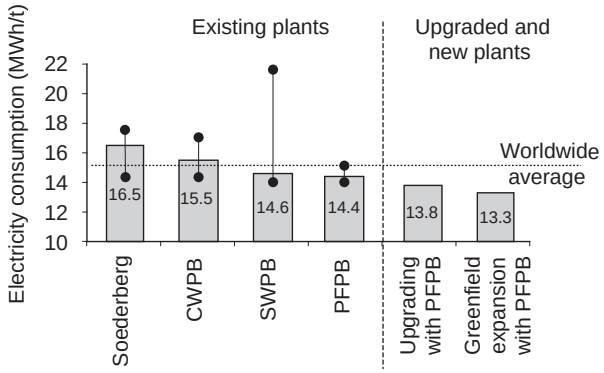


FIGURE 5 Average worldwide specific electricity consumption of existing smelters (inclusive range for 16 world regions) and for upgrading and new plants (MWh/t primary aluminum). From Schwarz (2000). Reproduced with permission.

TABLE VI

Final Energy Consumption of Preparation and Treatment of Different Important Scrap Types in Germany, Mid- to Late 1990s[a]

Scrap	Processes	Thermal energy (GJ/t_{SA})	Electric energy (MWh/t_{SA})
Old cars	Disassembly, shredder, sink-float-separation	—	0.13
Packaging	Sorting, with part pyrolysis, mechanical treatment	0.097	0.07
Building and mixed scrap	Shredder, with part DesOx-granulating, sink-float-separation	—	0.12
Electrical	Shredder, with part sink-float-separation	—	0.18

[a] Adapted with permission from Wolf (2000), "Untersuchungen zur Bereitstellung von Rohstoffen für die Erzeugung von Sekundäraluminium in Deutschland. ["Studies on the Preparation of Raw Materials for Secondary Aluminum Production in Germany"]. Shaker, Aachen, Germany.

fuel consumption and 11% of total electricity consumption for secondary production. Different treatment processes are used depending on the initial products (scrap types). These treatment processes have very different energy requirements. Table VI presents the specific final energy consumption of scrap preparation for different scrap types in Germany.

2.4.2 Melting, Refining, and Salt Slag Preparation

Melting is responsible for 38% of the overall specific fuel consumption and 69% of overall specific electricity consumption for secondary consumption. Additionally, the requirements for refining (15 and 1%, respectively) and salt slag preparation (18 and 19%, respectively) have to be considered.

These data are only valid as averages. As previously mentioned, depending on the type of scrap and the desired product quality, different types of furnaces for melting aluminum scrap are used. The selection of the most appropriate melting process is determined by the metal content of the scarp (oxide content), type and content of impurity (annealing loss), geometry of the scrap, frequency of change in alloy composition, and operating conditions (Fig. 6).

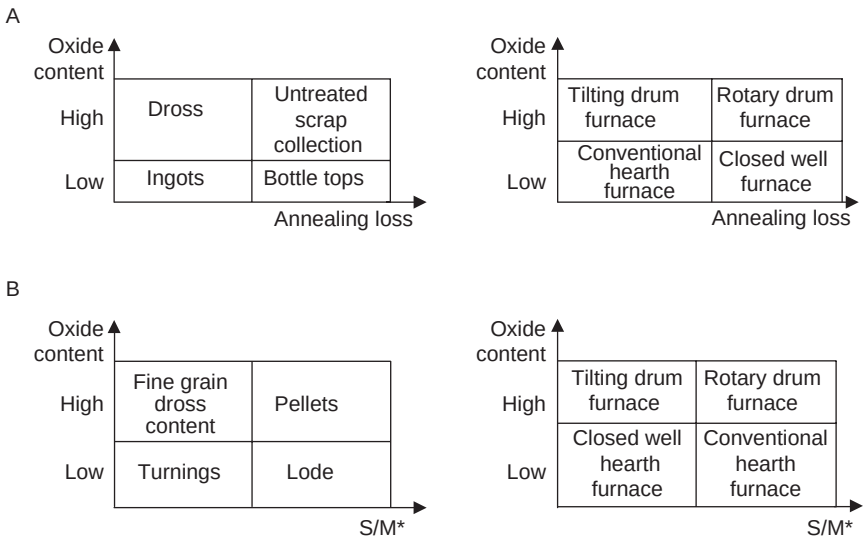


FIGURE 6 Classification of scrap types and furnace selection criteria by (A) oxide content and annealing loss and (B) oxide content and ratio of surface area to intermediate part size. *Surface area: mass of scrap pieces. From Boin, U., Linsmeyer, T., Neubacher, F., and Winter, B. (2000). "Stand der Technik in der Sekundäraluminiumerzeugung im Hinblick auf die IPPC-Richtlinie" ["Best Available Technology in Secondary Aluminum Production with Regard to the IPCC Guideline"]. Umweltbundesamt, Wien, Austria. Reproduced with permission.

TABLE VII

Typical Final Energy Consumption of Modern Secondary Aluminum Furnaces[a]

Furnace type	Scrap	Typical yield of metal (%)	Assumed yield of metal (%)	Electricity consumption (MWh/t_{SA})	Thermal energy consumption (GJ/t_{SA})
Induction	Pigs, ingots, turnings	95–99	98	0.70–0.93	—
Closed-well	Clean scrap, organically contaminated scrap	88–95	92	—	2.52–4.30
Rotary drum (static)	Turnings, pellets, shredder scrap, new scrap	75–92	74	—	4.41–4.70
Rotary drum (static + O_2 burner)	Turnings, pellets, shredder scrap, new scrap	75–92	74	—	1.87–2.05
Rotary drum (tilting)	Dross, pellets	50–80	78	—	2.67

[a] Adapted with permission from Boin *et al.* (2000). "Stand der Technik in der Sekundäraluminiumerzeugung im Hinblick auf die IPPC-Richtlinie" ["Best Available Technology in Secondary Aluminum Production with Regard to the IPCC-Guideline"]. Umweltbundesamt, Wien, Austria.

The appropriate melt aggregate for organically contaminated scrap with a low oxide content is the closed-well hearth furnace. The higher the fine and/or oxide content, the more appropriate it becomes to use a rotary drum furnace. In secondary melting plants, induction furnaces are only used in individual cases due to the necessity for clean, oxide-free scrap. Hearth and rotary drum furnaces are heated by fuel oil and natural gas. Only induction furnaces use electricity for melting. Table VII gives an indication of the very different final energy consumption for the most commonly used melting furnaces.

3. SPECIFIC PRIMARY ENERGY CONSUMPTION

3.1 Survey

Table VIII shows the specific primary energy consumption for global aluminum production in the mid- to late 1990s. If, as is usually the case, the regional energy carrier mix of electricity production is used for all processes, including electrolysis, to convert electricity consumption into primary energy consumption, the specific primary energy consumption of primary production is more than 11 times higher than the requirements for secondary production. If the contract mix is used for electrolysis, the consumption of primary production is nine times higher. Electrolysis accounts for the main portion (80 and 75%, respectively) of primary energy consumption for primary production, the processes within the refiner/remelter dominate (75%) in terms of the primary energy consumption for secondary production.

3.2 Determinants, Especially the Overall Efficiency of Electricity Production

The overall efficiency of both fuel supply and electricity production must be determined to convert final energy consumption into primary energy consumption. In our case (Table VIII), the same overall efficiencies of fuel supply (diesel, 88%; oil, 90%; natural gas, 90%; and coal, 98%) are assumed worldwide. The overall efficiency of electricity production depends on the efficiencies of power plants, the mix of energy carriers for electricity production, and the distribution losses. The allocation of the primary energy carriers to electricity production is usually determined by the regional mix of energy carriers. The regional mix is based on the average shares of electricity production of energy carriers within a region. In our case, the overall efficiencies of electricity production for 16 world regions were calculated. The resulting world overall efficiency of electricity production is different for the separate process stages (e.g., mining and alumina production) and depends on the associated production shares of the regions, which are used as weighted factors.

Additionally, the contract mix of energy carriers is used for the calculation of the primary energy consumption for electrolysis. The contract mix reflects the mix of energy carriers according to the

TABLE VIII

Specific Primary Energy Consumption of World Aluminum Production, Mid- to Late 1990s[a]

Primary production (GJ/t_{PA})		Total production (75% PA + 25% SA)[b](GJ/t_{ai})	Secondary production (GJ/t_{SA})	
Transport	2.0			1.1
Processes				
Mining	1.3		Scrap preparation and treatment	3.3
Refining	34.2		Remelting plant	13.3
Smelting	159.8[c] (120.1)[d]		Remelting	8.1
			Salt slag treatment	3.2
			Refining	2.0
Main physical inputs				
Caustic soda production	2.2			
Anode product	1.4			
Total	200.9[c] (161.1)[d]	155.1[c](125.3)[d]		17.7

[a] Adapted with permission from Schwarz (2000) and Krone (2000).
[b] PA, primary aluminum; SA, secondary aluminum.
[c] Regional energy carrier mix for electrolysis.
[d] Contract energy carrier mix for electrolysis.

contracts of the smelters with utilities. Whereas the regional mix provides a purely statistical allocation of smelters to power-generating units, the contract mix permits a more causal allocation. The higher the electricity demand of a smelter, the more electricity has to be supplied by the contractual generating units. Despite this advantage, the calculations based on the contract mix for electrolysis are seen merely as additional information (and therefore the data are given in brackets; see Table VIII). The reason is simple; for a consistent calculation of primary energy consumption for total aluminum production based on contract mix, it would be necessary to assess the contract mix of energy carriers for all production stages in all regions. This is simply not possible because of a lack of data.

Table IX shows the share of hydropower for the 16 world regions based on the regional mix and on the contract mix for electrolysis. In most regions, a larger share of electricity from hydropower is used for the contract mix. This is not surprising because low-price electricity from hydropower is one of the deciding factors for the siting of electrolysis. The larger share of hydropower leads mostly to a higher overall efficiency of electricity production because of the relatively high efficiency of hydroelectric power plants (80% is assumed). The efficiencies of power generation based on the other energy carriers are much lower and in the mid- to late 1990s were quite different for the individual regions (the following are assumed: natural gas, 22–33%; oil, 23–41%; coal, 24–36%; and nuclear power plants, 33%).

For most regions, the lower overall efficiency of electricity production of the regional mix leads to a higher primary energy consumption for electrolysis (Table IX). This is also true on a worldwide average. The specific primary energy consumption for electrolysis based on a regional mix is 160 GJ per ton of primary aluminum, with an overall efficiency of 35% and a share of hydropower of 29% of total electricity production. Based on contract mix, the data are 120 GJ per ton of primary aluminum, 46% efficiency, and 60% share of hydropower.

3.3 Future Primary Energy Consumption

The specific primary energy consumption of world aluminum production should decrease noticeable in the future. One main reason is the expected increase in the recycling share. An increase in the recycling share from 25 to 50% as projected for 2030 would lead to a decrease of 30% in specific primary consumption. In addition, there should be a decrease in final energy consumption for primary and secondary production and an improvement in overall efficiency of electricity production. Estimates for electrolysis, the deciding energy-consuming process, present an interesting picture: A 26% decrease in

TABLE IX

Regional Primary Energy Consumption for Electrolysis and Regional Overall Efficiency and Share of Hydropower for Electricity Production[a]

	Regional mix			Contract mix		
	Primary energy consumption electrolysis (GJ/t_{PA})	Overall efficiency (%)	Share of hydropower (%)	Primary energy consumption electrolysis (GJ/t_{PA})	Overall efficiency (%)	Share of hydropower (%)
Europe						
Germany	175.0	32	6	173.7	32	8
EU without Germany	142.0	36	17	125.6	41	29
Rest of Western Europe	81.3	67	86	68.1	80	10
Former USSR	233.3	28	21	118.8	55	80
Rest of Eastern Europe	212.4	26	2	139.8	39	53
Asia						
China	195.9	28	19	195.0	28	18
India	190.0	29	20	194.4	28	20
Near and Middle East	180.3	30	5	131.4	42	50
Rest of Asia	148.7	36	10	67.1	80	100
Africa	152.1	35	15	139.1	38	29
America						
Brazil	75.1	78	95	68.1	79	100
Canada	65.9	51	61	103.5	80	100
United States	163.4	34	12	127.0	44	44
Rest of Latin America	153.7	36	30	69.5	79	100
Australia and Oceania	150.8	35	11	139.1	38	23
World[b]	159.8	(35)	(29)	120.1	(46)	(60)

[a] Adapted with permission from Schwarz (2000).
[b] Regional primary aluminium production shares as weighted factor.

primary energy consumption is estimated for the years 1995–2010. Interestingly, the main reason for this decline is not the decrease in specific electricity consumption, which accounts for less than 40% of the decline. The main reason is the improvement in overall efficiency of electricity production globally, which accounts for the remaining 60% or more. This improvement in worldwide overall efficiency is the result of increased efficiency of fossil-fired power plants, especially in the former Soviet Union, China, and the United States. These are regions with high primary aluminum world production shares and with high shares of fossil fuels for electricity production. In addition, the relocation of primary aluminum production to hydropower-rich regions accounts for the increase in worldwide overall efficiency.

4. ENERGY AS COST FACTOR

4.1 Survey

Table X presents the average cost structure of world aluminum production. Less reliable are the data for secondary production (only the processes within the refiner/remelter). These are based on rough estimations. However, it is clear that the share of energy costs in total operating costs of secondary production is very low in comparison with the costs of all processes of primary production. The energy cost shares of alumina and primary aluminum production are approximately the same. Nevertheless, the energy costs of electrolysis are more important with respect to choosing location.

4.2 Primary Production

4.2.1 Mining and Alumina Production

Energy costs represent 11% of total operating costs for bauxite mining. Labor is the dominant cost factor, with a share of more than 40%. Deposits with bauxite, which is easy to extract (poor surface and good ore thickness) and of high quality (e.g., high aluminum oxide content), are preferred for mining. The conditions of extraction influence the necessary labor input and energy requirements. The higher the quality of the bauxite, the higher the attainable price. Furthermore, general political conditions play a major role. An exorbitant bauxite levy, for example, hindered the bauxite industry in Jamaica for many years.

Energy costs are relatively high for alumina production. On worldwide average, they account for 25% of total operating costs. Bauxite costs are the dominant cost factor. They represent more than one-third of the total operating costs. Bauxite costs are heavily dependent on the location of alumina production. They vary considerably between regions. For example, in Latin America the cost of bauxite per ton of alumina was approximately $40 in the mid- to late 1990s, whereas in Western Europe (with the exception of Greece) the cost was $80. Approximately half of this difference is due to high transport costs, which amounted to $15–20 per ton for transport to Western Europe. The longer the transport distances are (and the lower the alumina content of the bauxite), the higher the specific transport costs. This explains the trend for locating refineries close to mines (and also why only bauxite with a high alumina content is traded between regions). The bauxite-producing countries (with a production >2 million tons in 2000) increased their alumina production by approximately 36% from 1990 to 2000, whereas that in countries with no significant bauxite production declined by approximately 14%. The production share of countries with significant mining activities increased from 64% in 1990 to 74% in 2000.

The location of alumina production is determined by transport costs for bauxite. The energy costs play only a minor role in this respect. On the one hand, upgraded and new plants have almost the same technical parameters, including energy requirements for alumina production. On the other hand, the energy carriers coal, oil, and natural gas can increasingly be obtained in similar conditions worldwide, with the result that energy costs (at least for upgraded and new plants) do not differ much between regions.

TABLE X

World Average Total Operating Costs for Primary and Secondary Aluminium Production, Mid- to Late 1990s[a] (USD/t_{output})

	Primary production			
Total operating costs (%)	Mining	Refining	Smelting	Secondary production[b]
	15	150	1350	(1400–1500)
Main raw material	—	34	33	(>65)
Labour	43	13	10	(>8)
Electricity	—	3	24	(1)
Other energy	11	22	—	(1–2)
Caustic soda	—	13	—	—
Other operating costs	46	15	33	(<25)

[a] Adapted with permission from Schwarz (2000).
[b] Remelting, inclusive salt slag treatment and refining.

4.2.2 Electrolysis

For smelting, alumina costs were dominant, with a share of 33% of total operating costs. Nevertheless, electricity costs, which were responsible for one-fourth of the total costs, are more important with regard to determining where to locate smelters because the variation in price of electricity is much greater than that of alumina. This is not surprising because electricity is only tradable to a limited extent. The necessary net infrastructure and the distribution losses limit regional interchange. Typically, primary aluminum smelters that operate with low-electricity tariffs obtain power from nearby hydroelectric power stations. Smelters that operate in regions with a high potential for cheap hydroelectric power have a natural competitive advantage. The abundance of energy is practically "stored" in aluminum and therefore becomes tradable.

There is a moderate trend of shifting smelting capacity to regions with low electricity prices. The building of new plants and capacity expansions of existing plants are usually done in these regions. Nevertheless, regions with high electricity prices for smelting, such as in the European Union, still command a relatively large share of world aluminum production (approximately 10% in 2000). The status of old plants in disadvantageous regions is usually not endangered because the high initial investment costs of these plants can be characterized as sunk costs due to the high degree of factor immobility and the limited scope of alternative uses of plants and machinery in the smelting industry. Therefore, old capacity will only be superseded by new capacity when the specific total costs of new capacity are lower than the specific operating costs of old capacity.

4.3 Secondary Aluminum Production

Energy costs for secondary aluminum production are of little relevance. The share of energy costs in total operating costs is 3% maximum. The cost of the scrap containing aluminum is the dominant factor ($>65\%$), followed by labor costs ($>8\%$). The profitability of aluminum recycling is determined by the difference (margin) between the scrap price and the attainable price of the different alloys. The prices for secondary and primary aluminum are interdependent. Assuming that initially both markets are in equilibrium, if the price of secondary aluminum were to increase, secondary aluminum would be increasingly substituted by primary aluminum. The decreasing demand for secondary aluminum would lead to a declining demand for scrap containing aluminum and therefore to lower scrap prices until the initial market equilibrium is again attained.

The major aluminum-consuming countries are still the major secondary aluminum producers. Obviously, the close proximity to initial products (scrap) and purchasers is still the decisive factor with regard to location. In the future, many expect that producers in South Asia, South America, and the Gulf states will gain production shares because of lower labor, environmental, and other costs.

5. SUMMARY

The determinants of specific primary energy consumption for world aluminum production are the recycling share, final energy requirements of the relevant process steps, and the overall efficiencies of fuel supply and electricity production. If, as expected, recycling share increases, final energy consumption decreases, and overall efficiencies improve, primary energy consumption will decline noticeably in the next few years.

Market processes are responsible for the expected growth in recycling share and the decrease in final energy requirements. The importance of energy costs in primary production makes both secondary production (and therefore the increase in recycling share) and efforts to reduce final energy requirements (especially for alumina production and primary aluminum smelting) attractive.

One question remains: What is the future absolute primary energy consumption of world aluminum production? Let us assume an increase in world aluminum production from 33 million tons in 2000 to 50 million tons in 2030 (an increase of 52%). Unchanged absolute primary energy consumption would require a decrease in specific primary energy consumption of 34%. Assuming an increase in the recycling share from 26 to 46–48% (Table IV), this would lead to a 22–26% decrease in specific consumption. In this scenario, it is very likely that the increase in the recycling share, together with the decrease in final energy requirements of the relevant process steps and improvements in overall efficiencies of fuel supply and electricity production, will lead to a decline in specific consumption of $>34\%$ and therefore to a decreasing absolute primary energy requirement for world aluminum production. The question is whether the slowdown in growth rates of primary production will in fact be as strong as assumed here.

SEE ALSO THE FOLLOWING ARTICLES

Cement and Energy • Cost–Benefit Analysis Applied to Energy • External Costs of Energy • Plastics Production and Energy • Steel Production and Energy

Further Reading

Altenpohl, D., and Paschen, P. (2001). Secondary aluminium is the raw material highlight of the 21st century. *Aluminium* **77**(1/2), 8–13.

Bielfeldt, K., and Grojtheim, K. (1988). "Bayer and Hall–Héroult Process—Selected Topics." Aluminium-Verlag, Düsseldorf.

Kammer, C. (1999). "Aluminium Handbook. Volume 1: Fundamentals and Materials." Aluminium-Verlag, Düsseldorf.

Krone, K. (2000). "Aluminiumrecycling. Vom Vorstoff zur fertigen Legierung" ["Aluminum Recycling. From Initial Product to Aluminum Alloy"]. Vereinigung deutscher Schmelzhütten, Düsseldorf.

Æye, H. A., Mason, N., Peterson, R. D., Rooy, E. L., Stevens McFadden, F. J., Zabreznik, R. D., Williams, F. S., and Wagstaff, R. B. (1999). Aluminum: Approaching the new millennium. *J. Metals* **51**(2), 29–42.

Morel, P. (1991). "Histoire technique de la production d'aluminium" ["History of Technologies in Aluminum Production"]. Presses Universitaires de Grenoble, Grenoble, France.

Mori, G., and Adelhardt, W. (2000). "Stoffmengenflüsse und Energiebedarf bei der Gewinnung ausgewählter mineralischer Rohstoffe, Teilstudie Aluminium" ["Material Mass Flows and Energy Consumption Caused by the Production of Selected Mineral Raw Materials, Section: Aluminum"]. Bundesanstalt für Geowissenschaften und Rohstoffe, Hannover, Germany

Pawlek, P. (1999). 75 Years development of aluminium electrolysis cells. *Aluminium* **75**(9), 734–743.

Schwarz, H.-G. (2000). "Grundlegende Entwicklungstendenzen im weltweiten Stoffstrom des Primäraluminiums" ["Fundamental Trends in the Global Flow of Primary Aluminum"]. Forschungszentrum, Zentralbibliothek, Jülich.

Schwarz, H.-G., Kuckshinrichs, W., and Krüger, B. (2000). Studies on the global flow of primary aluminium. *Aluminium* **76**(1/2), 60–64.

Schwarz, H.-G., Briem, S., and Zapp, P. (2001). Future carbon dioxide emissions in the global flow of primary aluminium. *Energy* **26**(8), 775–795.

Thonstad, J., Fellner, P., Haarberg, G.M., Híves, J., Kvande, H., and Sterten, A. (2001). "Aluminium Electrolysis. Fundamentals of the Hall–Héroult Process." Aluminium-Verlag, Düsseldorf.

Aquaculture and Energy Use

M. TROELL
The Beijer Institute and Stockholm University
Stockholm, Sweden

P. TYEDMERS
Dalhousie University
Halifax, Nova Scotia, Canada

N. KAUTSKY
Stockholm University and The Beijer Institute
Stockholm, Sweden

P. RÖNNBÄCK
Stockholm University
Stockholm, Sweden

1. Introduction
2. Overview of Aquaculture
3. Energy Analysis in Aquaculture
4. Energy Performance of Aquaculture
5. Biophysical Efficiency and Sustainability

Glossary

aquaculture The farming of aquatic organisms, including fish, mollusks, crustaceans, and aquatic plants. Farming implies some form of intervention in the rearing process to enhance production, such as regular stocking, feeding, and protection from predators. Farming also implies individual or corporate ownership of the stock being cultivated.

embodied energy density The accumulated energy use per area of production. Can include both direct and indirect industrial energy inputs as well as fixation of solar energy in ecological systems.

emergy The sum of the available energy (exergy) of one kind previously required directly and indirectly through input pathways to make a product or service.

energy intensity (EI) The accumulated energy inputs required to provide a given quantity of a product or service of interest. In the current context, energy intensity is expressed as the total joules of energy required to produce a live-weight tonne of fish, shellfish, or seaweed.

fossil fuel equivalents An expression of auxiliary energy inputs from the economy, embodied in goods and services (e.g., 1 unit of electricity = 3–4 units of fossil fuels).

industrial energy analysis (IEA) The sum of the energy (fossil fuel equivalents) needed for manufacturing machinery and products used as inputs to the system, and the fuel energy needed for operation.

life-support systems The parts of the earth that provide the physiological necessities of life, namely, food and other energies, mineral nutrients, oxygen, carbon dioxide, and water; the functional term for the environment, organisms, processes, and resources interacting to provide these physical necessities.

solar energy equivalents A unit making it possible to combine ecological and economic energy requirements, based on the energy from solar fixation. It takes about 10 times of solar fixation per 1 unit of fossil fuel.

In this article describing the various techniques used to assess the energy performance of aquaculture, the discussion focuses on the results of different energy analyses that have been undertaken. The main focus is on industrial energy analysis, i.e., embodied fossil fuel energy. However, analyses within a wider ecosystem context are also discussed. An analysis of the broad range of aquaculture practices is not within the scope of this general overview. Instead, the focus is on some selected culture systems, representing intensities and methodologies dominating global production.

 © 2004 Elsevier Inc. All rights reserved.

1. INTRODUCTION

The need for indicators to assess sustainable development of aquaculture has been high on the research agenda for the past decade. Useful indicators preferably focus on both biophysical and socioeconomic dimensions. An example of a biophysical indicator used in aquaculture is the measure of the dependence of aquaculture on different forms of energy. Even though energy consumption (i.e., of auxiliary energy) within the aquaculture sector is small compared with global economic activities, and energy analyses generally provide only a partial tool for assessing sustainability of an industry, it is of interest to increase our understanding of aquaculture's dependence on an external finite resource. In addition, energy dependence provides a basis on which we can evaluate the efficiency characteristics of different types of aquaculture production systems and processes; thus, this indicator can be used to measure the extent to which alternative options move toward efficiency goals implied by sustainability. Further, the energy costs of production not only involve sustainability issues of ecosystem resource efficiency and nonrenewable resource depletion, but also the potential cost to future societies through environmental change from pollution and global climate change.

2. OVERVIEW OF AQUACULTURE

2.1 Definition and Forms

Aquaculture is the farming of aquatic organisms, including fish, mollusks, crustaceans, and aquatic plants. By definition, aquaculture implies some sort of intervention in the rearing process to enhance production. From an industrial perspective, the objective is to maximize production of some demanded species to achieve maximum economic benefits. This is the objective for most recently developed operations, but not necessarily for the many traditional systems whose outputs are primarily directed toward rural and household subsistence consumption. Aquaculture is a highly diverse activity, and species choice is important from an energy perspective because different species have different characteristics that limit the circumstances under which they can be cultured.

Farming takes place in fresh, brackish, and marine waters and encompasses two main subsectors, commercial/industrial- and rural/subsistence-based operations. The former usually target species generating high income, and the latter target species largely used for subsistence consumption or local markets. Culturing may be practiced either as monoculture or polyculture. Traditional freshwater farming systems in China have been, and to a large extent still are, practiced as polycultures using low stocking densities. By contrast, modern-day operations are usually monoculture systems employing high stocking densities. Freshwater pond farming accounts for the majority of global cultured finfish production. Only a small proportion of freshwater production is conducted in land-based tanks and these are almost exclusively found in developed countries. The most common open-water culture technologies, conducted primarily in marine waters, include either cages or pens for finfish, and rafts, long lines, or poles for seaweed and mussel culture.

Farming practices can be classified into extensive, semiintensive, and intensive farming. The classification is often based on resource (e.g., feed) and energy use, but rearing densities, maintenance costs or inputs, chemical use, etc. may also be considered. Intensive culture systems include the following general characteristics:

- High stocking/rearing densities.
- High dependence on exogenous direct and indirect energy inputs, with the latter in the form of feed and chemicals.
- Low inputs of labor energy.
- Production of relatively high-valued species.

Intensive culture is typically conducted in tanks, ponds, and in open-water cages. In most of these systems, with the possible exception of unlined earthen ponds, capital investments are usually high. Carnivorous finfish species such as salmon, seabass, groupers, eel, halibut, and cod are typically farmed intensively and depend almost exclusively on commercially prepared feeds or aquafeeds. This is also true for some omnivorous penaeid shrimp species (e.g., tiger shrimps). In contrast, extensive and semiintensive systems in which most detrivorous, herbivorous, and omnivorous fish species (carps, milkfish, tilapia, etc.) are cultured, rely to varying degrees on food produced within the culture environment. However, many of these systems are fertilized, using either industrial fertilizers or organic by-products from agriculture, to stimulate primary production within the pond environment. In recent years, many of these traditional extensive systems have been moving along a path toward greater

intensification and compound aquafeed usage in order to increase production.

2.2 Importance of Aquaculture

A wide range of plants and animals are farmed globally, with more than 220 different species of finfish and shellfish being produced. Although most aquaculture is undertaken to produce food, other applications include fodder, fertilizers, extracted products for industrial purposes (e.g., seaweed compounds used in manufacturing processes, food processing, and microbiology), goods for decorative purposes (shell, pearls, and ornamental fish), and recreation (sport fishing, ornamentals, etc.). The demand for seafood is steadily increasing worldwide, but most of the world's fishing areas have reached their maximal potential for capture fisheries production (Fig. 1). It is therefore logical that aquaculture, currently one of the fastest growing food production sectors in the world ($\sim$10% yr^{-1}), will be increasingly called on to compensate for expected shortages in marine fish and shellfish harvests. Currently, aquaculture accounts for approximately 30% of total food fish production globally, of which roughly 90% originates from Asia. Production histories of some major aquaculture species groups from 1970 to 2000 are presented in Fig. 1.

Geographically, China contributes more than two-thirds of total aquaculture production (mainly from traditional freshwater cultures). Europe, North America, and Japan collectively produce just over one-tenth of global output but consume the bulk of farmed seafood that is traded internationally.

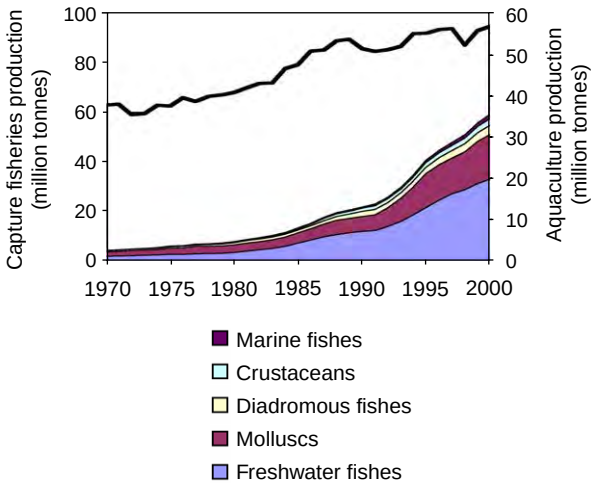


FIGURE 1 Production development of global capture fisheries (thick line, including both marine and freshwater fishes) and major aquaculture groups from 1970 to 2000. Figures for aquaculture (but not for capture fisheries) are cumulative.

On a tonnage basis, extensive and semiintensive freshwater aquaculture dominates global fish production (56% by weight). Currently, within these systems about 80% of carp and 65% of tilapia worldwide are farmed without the use of modern compound feeds. It should be noted, however, that during the 1990s, the more capital- and energy-intensive production of shrimp and marine (including diadromous) fish species increased significantly.

3. ENERGY ANALYSIS IN AQUACULTURE

3.1 Forms of Energy Used in Aquaculture

3.1.1 Ecosystem Support

From an ecological perspective, most aquaculture systems involve the redirection and concentration of energy and resource flows from natural ecosystems to the culture environment. Consequently, very few systems depend solely on the solar radiation incident on the production facility. This is easily appreciated when one considers the relatively small surface area of many high-density culture systems that are exposed to the sun. Exceptions, however, include extensive culture of autotrophic species such as seaweeds. Similarly, filter-feeding organisms such as mussels are dependent on phytoplankton produced in adjacent waters that are brought to the farm by currents. For most farming systems, though, productivity is enhanced by supplementing auxiliary energy and resource inputs produced in other ecosystems (Fig. 2). For example, a fish cage or shrimp pond is dependent on large ecosystem support areas outside the farm border (i.e., a culture system's "ecological footprint" is much larger than the area occupied by the farm), where solar energy is being transferred through food webs to produce feed. Currently, fish meal and fish oil remain the dominant ingredients in most compound feeds for carnivorous finfish and marine shrimps. As a result, many aquaculture systems that utilize compounded feeds containing fish meal and fish oil actually appropriate more fish biomass as inputs to the feed than is represented by the farmed product produced. And as will be shown, this relative dependence on capture fisheries products is also important when discussing energy consumption, because the energy associated with providing the feed is often the main input to many aquaculture systems.

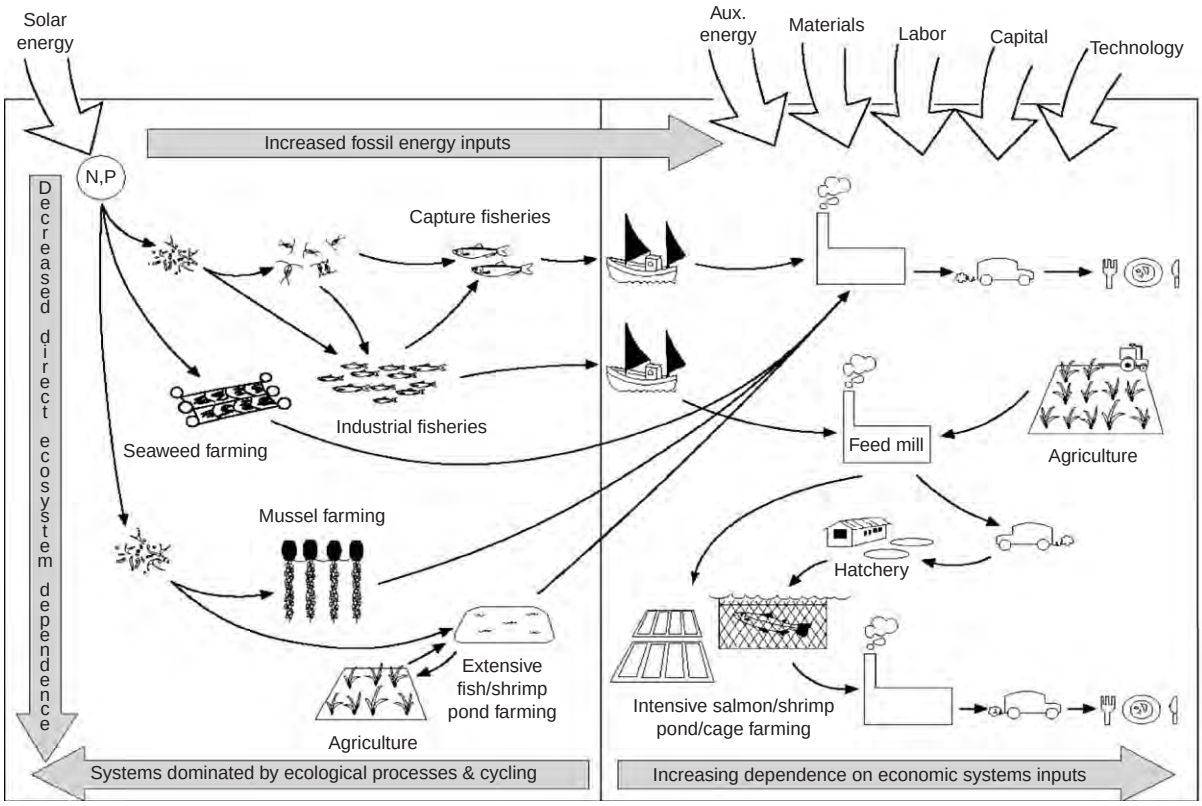


FIGURE 2 Simplified energy and matter flow model for various types of seafood production from solar fixation by plants in the natural system and until the product reaches the consumer. In natural ecosystems, as well as in extensive farming systems, solar energy and recirculation of nutrients constitute the limiting factor for biological production. Due to energy losses in each transfer between trophic levels, the harvest per area will decrease when a higher level of the food chain is exploited. Extensive seaweed and mussel farming may require additional inputs of materials for support structures (buoys, ropes, etc.), and extensive fish and shrimp production usually requires ponds and inputs of seed, fertilizers, and/or agricultural by-products to maintain production. In contrast, intensive cultures of carnivorous species such as salmon and shrimp require high industrial energy intensity (fossil fuel input per weight of product), mainly due requirements for formulated feed made from fish- and crop-derived products, sophisticated cages and ponds, energy for aeration and pumping, inputs of hatchery-reared smolts/fry, and medication. Moreover, each step in the feed production chain from fisheries and crop harvesting, processing, transportation, etc. requires fossil energy, materials, labor, technology, and capital to a varying degree. Consequently, it is clear that there are significantly more energy-demanding and pollution-generating steps involved in intensive shrimp/salmon farming than in extensive aquaculture or capture fisheries.

3.1.2 Direct Energy Inputs

Direct energy inputs to aquaculture operations encompass a range of activities, including the collection/production of juveniles, general system operations, and the harvesting, processing, and transport of the product. A first-order approximation of the industrial energy dependence of a culture system can be obtained by considering the intensity of the operation, the degree of mechanization, the type or means of production, and the quality/quantity of feed inputs. Provisions of feed and juveniles for stocking are the main material inputs, and fuels and electricity are needed for various grow-out operations.

Energy inputs associated with harvesting, processing, and transporting the various feed components are implicit in the reliance of intensive forms of aquaculture on remote ecosystems (Fig. 2). For example, fishing fleets and reduction plants consume large amounts of fossil and electrical energy in order to provide fish meal and fish oil. Similarly, a wide range of direct and indirect energy inputs are required by the agriculture industry to supply plant- and livestock-derived components of feed pellets. This includes the indirect energy cost for fuel and labor, together with the direct energy-demanding processes. Further energy inputs to aquaculture operations are associated with either the production of juveniles in hatcheries or their collection from the wild. Hatchery production is also dependent on energy-consuming activities for the provision of

resources such as feed and broodstock (adult individuals, either wild captured or reared entirely within the culture environment).

By comparison, most extensive culture systems receive no formulated feeds. Instead, inputs to these systems typically comprise agricultural products or wastes (e.g., supplemented manure) and chemical fertilizers. This typically results in smaller industrial energy inputs, particularly if crop residuals and manure, essentially by-products of local agricultural production, are treated as biophysically free with only marginal energy inputs associated with further processing. Similarly, feed inputs to semiintensive systems may consist, either wholly or in part, of fish from local fisheries. In these cases, the unprocessed fish may be used directly or supplied as feed after minor transformation (i.e., as moist pellets). Thus, utilization of locally available fisheries resources may involve less fossil fuel energy, particularly if drying- and transportation-related energy inputs are limited. However, depending on the fishing technology used and the relative abundance of local harvestable stocks, fuel inputs may be surprisingly high when compared with more modern, large-scale reduction fisheries. Moreover, because raw fish products can spoil quickly if unused, energy inputs to a culture system that utilizes these products will increase as a result of any waste that occurs. Alternatively, fish processing wastes can, to some extent, replace fish protein inputs from whole fish. Such a substitution might again imply "no energy cost" for extraction, but only for transportation and any secondary processing that is required.

The methodological appropriateness of allocating a zero biophysical cost for by-products from fish processing and agriculture production is debatable. It can be argued that treating the energy costs associated with making a given by-product available as free, simply because they are designated as "wastes," is arbitrary and potentially misleading. From this perspective, the energy inputs associated with any production process, regardless of how its many outputs may be regarded, should be fully included in any energy analysis. Alternatively, for some products the energy costs could be based on the proportion to the relative weights of product and by-product. However, such an accounting system may not be applicable to all products (e.g., manure production). A third possible accounting convention might be based on the relative prices of coproducts. This approach only works, however, if a market for the by-product exists, or if alternatives to the by-product exist, from which an opportunity cost could be estimated. The degree to which these alternative approaches capture the true energy costs of different inputs to aquafeeds is open for debate. In practice, however, to date, most, if not all, of these techniques have been used by various analysts when evaluating the energy costs of aquaculture.

Accounting for labor inputs is also a complicated issue within the context of energy analysis. In most aquaculture analyses, labor has been excluded, with the justification that its contribution to the energy profile of the system is negligible when compared to fossil fuel inputs. Again, this may be accurate for some types of systems but not for others. Methods that have been used to account for the energy costs of labor either incorporate the nutritional energy content of the food required to sustain the requisite labor inputs, or, alternatively, incorporate the industrial energy needed to provide the food. It is, however, questionable if these methods, based as they are on standard energy conversion values, should be applied widely, because there are large discrepancies with respect to energy production costs of food and consumption levels between developed and developing countries. To address these context-specific concerns, some analysts incorporate the total industrial energy associated with wages paid for labor inputs, using the industrial energy to gross national product (GNP) ratio for the country within which labor is supplied.

3.1.3 Indirect (Embodied) Energy Inputs

Fixed capital inputs may include different forms of enclosures (tanks, nets, etc.), feeders, aerators, pumps, boats, vehicles, etc. The energy embodied in building and maintaining these capital inputs may be substantial, particularly in intensive systems. Methodologically, a common practice for estimating the energy inputs associated with fixed capital has been to use input–output analysis of the country's economy. The resulting energy equivalents provide, after accounting for depreciation, estimates of the related energy costs per year. However, not all studies of aquaculture systems have accounted for fixed costs in this way. Instead of applying an annualized energy cost over a defined time period, some studies treat fixed costs as a single lump energy input. Alternatively, some studies have simply excluded fixed costs completely from the analysis. To what extent these alternative approaches affect the overall energy analysis depends on the characteristics of the system being analyzed. Generally, the fixed energy cost in intensive systems is insignificant compared to the overall energy profile, due to the high embodied

energy cost for feed (Table I). In culture systems that depend predominantly on naturally available productivity and with low inputs of supplemented feed, the energy embodied in capital infrastructure may play a more important role in the overall industrial energy analysis (IEA) (Table I). The variation in the methodological approaches used and the rigor with which they are applied between different studies can make the direct comparison of their results difficult. For example, in some cases, total energy inputs are reported as a single value without any specification of the various inputs. In other instances, only one form of energy may be accounted for (for example, only the electrical energy consumption in the grow-out phase is included, whereas the energy inputs associated with feed, hatchery rearing, fixed infrastructure, harvesting, etc. are left out). In other seemingly more comprehensive analyses, a specific source of energy input may be completely overlooked (for example, not accounting for the energy consumed within the fishery sector that is required to sustain raw feed materials). Similarly, energy inputs to hatchery operations may be left out.

3.2 Analytical Techniques

Relatively few energy analyses of aquaculture systems have been conducted to date. A variety of analytical approaches have been used, the most frequent being quantification of direct and indirect auxiliary (industrial) energy inputs (e.g., electric energy, oil, gas, or coal) dissipated in the myriad supporting economic processes. These apply regular process and input–output analysis normally using land, labor, and capital as independent primary factors of economic production. This method sets the "system boundary" at the point of material extraction. Thus, the method is limited in the sense that it does not consider the impact of extraction on the future availability of resources, and does not value the wider resource base on which the production depends.

In a second approach, the boundaries of the system are expanded to include processes and the energy transformed by renewable resources. In doing so, the focus of an analysis shifts to an understanding of the solar energy dependency of economic activities. In the context of aquaculture, solar energy enters mainly through agriculture and fisheries, together with the energy embodied in fuels and raw materials (Fig. 2). The most highly developed form of this approach has given rise to the concept of "emergy." This method includes all forms of energy such as material and human inputs and environmental services (including energy of hydrological and atmospheric processes) being converted into their solar energy equivalents. Thus, it evaluates the work done by the environment and by the human economy on a common basis.

Not surprisingly, different approaches result in different kinds of information. Ideally, it is necessary to consider the total energy use from pond to plate, expressed per unit of edible produced protein. Within the context of sustainability, it can be argued that an analysis that also captures the work by nature, i.e., the energy flows through dynamic ecosystems, more accurately reflects dependencies on the overall resource base. All economic activities depend on these life-supporting systems. Often, however, this support is taken for granted and is not included in any efficiency calculations.

4. ENERGY PERFORMANCE OF AQUACULTURE

4.1 Extensive Systems

Extensive culture systems generally have relatively low dependence on direct auxiliary and indirect embodied energy inputs due to farming practice and species choice. The culture of species groups such as mussels, carp, and tilapia (Table I) have generally low energy intensity (EI) input per unit of production and also per edible protein energy return. These organisms are able to utilize natural primary production (phytoplankton or macrophytes) or low-quality agriculture by-products and thereby require little if any artificial feed. The various means for enhancing natural production, i.e., applying fertilizers or organic inputs (e.g., manure), are in most cases not costly from an energy perspective. Human labor inputs are typically high in countries where extensive culturing methods dominate, both at the farm site and in activities generating different inputs to the farm (Table I).

The classification into extensive culture systems, based on energy consumption, may not always be as straightforward as it might seem. For example, the same type of culture system may depend on very different forms and quantities of energy inputs depending on where the system is situated (e.g., whether in industrialized or nonindustrialized countries). For some activities, fossil fuels may be substituted for human labor. Similarly, the energy embodied in various material inputs that could be

TABLE I

Direct and Indirect Energy Inputs and Resulting Energy Intensities of Semiintensive Shrimp Pond Farming, Intensive Salmon Cage Farming, Semiintensive Carp + Tilapia Pond Polyculture, Indian Carp Polyculture, Intensive Catfish Pond Farming, Semiintensive Tilapia Pond Farming, and Mussel Long-Line Culture[a]

	Species Farmed													
Parameter	**Shrimp**		**Salmon**		**Polyculture (carp recirc)[b]**		**Polyculture (Indian carp)**		**Catfish**		**Tilapia**		**Mussel**	
System Characteristics														
Site area (ha)	10		0.5		1		1.0		1.0		0.005		2	
Production/yr(t)	40		200		4		4.3		3.4		0.058		100	
Production/yr/ha(t)	4		400		4		4.3		3.4		12		50	
Energy Inputs[c]	EI	%	EI	%	EI	%	EI[d]	%	EI	%	EI	%	EI	%
Fixed Capital														
Structures/equipment	2500	2	5940	6	3556	7	2114	8	11,691	10	0	0	2700	58
Operting Inputs														
Fertilizer, chemicals	22,750	15	0	0	9316	18	633	2	369	0	1000	3	0	0
Seed	18,750	12	2970	3	15	0	0	0	11,076	10	0	0	0	0
Feed	58,250	37	78,210	79	15,451	31	287	1	86,389	75	23,280	97	0	0
Electricity, fuel	54,250	35	11,880	12	21,928	44	26,176	97	5415	5	0	0	1900	42
Total	156,750		99,000		50,265		27,096		114,940		24,000		4600	
Labor Inputs[e]														
Person-days/t	n.d		10.0		6.5		66.7		3.5		172.4		5.0	
Percentage	6.4[f]		14		n.d		n.d		n.d		35[g]		63	
Energy Ratio														
Edible product(MJ/kg)[b]	275	***	142		84	***	45	***	192	***	40		12	
Edible protein(MJ/kg)	784	**	688		272		135	**	575	**	199		116	

[a] Data from Larsson *et al.* (1994) (shrimp), Stewart (1994) (salmon), Bardach (1980) (recirc. Asian carp), Singh and Pannu (1998) (Indian carp), Waldrop and Dillard (1985) (catfish), and Stewart (1994) (tilapia and mussels).

[b] One-third of water recirculated.

[c] Energy intensities expressed in kilojoules/kilogram (wet weight).

[d] Calculating with 20 yrs of deprivation time for pond construction.

[e] Labor inputs expressed as person-days per tonne of production and as a percentage of the total capital cost; n.d., no data available.

[f] Due to lack of data this value is based on % labor energy consumption (calculated form GNP and wages) of total IE inputs.

[g] Labor for pond construction dominating input.

[b] Per kilogram of wet weight.**, Calculating with 20% protein of product wet weight; ***, calculating with 60% of product being edible.

used to construct a culture system may differ due to variations in the energy associated with the input extraction, processing, and transport. To illustrate the potential challenges associated with an energy-based classification scheme, consider mussel farming, widely regarded as an extensive culture system in that it involves purely filter-feeding organisms that grow well without supplemented feeds or fertilizers. Although indirect energy inputs may be relatively small, the forms and quantities of direct energy for harvesting, sorting, and cleaning can vary widely. For example, mussel farming in the developed world is typically more dependent on industrial energy inputs for these operations. In contrast, in developing countries, industrial energy inputs are often low because most operational work is carried out by hand, because labor is relatively abundant and inexpensive. Similarly, the farming of herbivorous fish species in which agricultural inputs are used to enhance production may, in more industrialized countries, result in higher energy consumption, because the underlying agricultural production system is more fossil fuel dependent. Even though the discussion here includes no case requiring change in the general application of the intensity classification, it is necessary to be aware that site-specific factors can significantly influence the energy analysis, even if the same types of systems and species are being used.

Increased pressure for expansion of global aquaculture production, together with increased scientific knowledge concerning the optimization of growth performance, has resulted in the incremental intensification of many extensive culture systems. This intensification process typically entails the increased use of artificial feeds containing fish meal. Not only does this transformation reduce the total global supply of fish potentially available for human consumption, but it also increases the industrial energy intensity of the culture system. For example, relatively small increases in the use of artificial feeds, in the context of extensive systems, can quickly dominate the total industrial energy inputs to the system Table I.

4.2 Intensive Systems

It becomes obvious from comparing extractive species, such as seaweeds and mussels, with carnivorous species, such as salmon, that the inherent or natural trophic level of the species being farmed plays a significant role in the energy intensity of the culture system. However, Table I also reveals that there are other production factors that can play a significant role in the overall industrial energy costs of a given system. These include intensity level, dependence on artificial feed, degree of mechanization, and environmental control needed in the husbandry system.

Direct and indirect energy inputs to operational activities can represent more than 90% of the total EI inputs to intensive culture systems. Of this, the energy required to provide feed is usually the single most important item (Fig. 3A). Other inputs of importance, however, include direct electricity and fuel inputs, along with the energy required to supply juveniles. In herbivorous/omnivorous fish culture systems, the energy required to supply fertilizer and green fodder can also be substantial (Fig. 3B). In contrast, the embodied and direct construction-related energy costs associated with the fixed capital infrastructure of many intensive farming systems are relatively insignificant (Fig. 3A, Table I). Similarly, regardless of type of culture system, labor in Western countries is generally negligible in comparison to fossil fuel inputs.

It is no surprise that intensive land-based and open-water aquaculture activities differ significantly with respect to their energy costs associated with

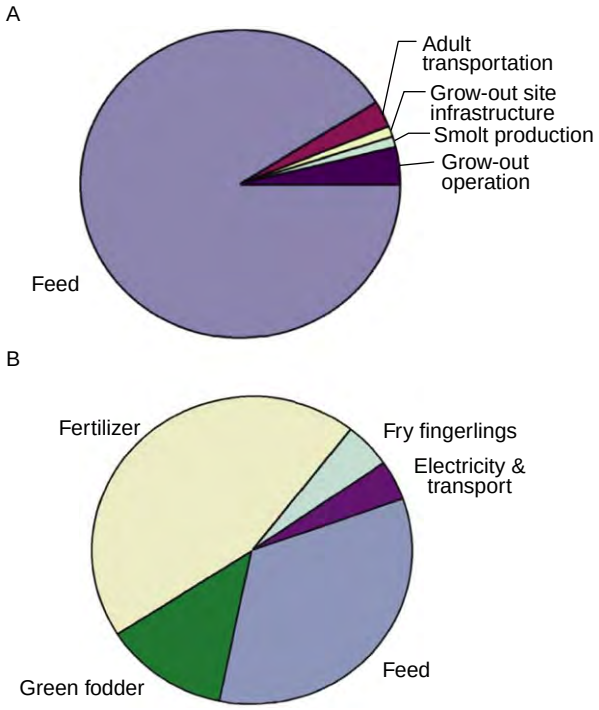


FIGURE 3 Source and percentage of industrial energy inputs to produce (A) Atlantic salmon (intensive cage farming) (from Tyedmers, 2000) and (B) silver and bighead carps (semi-intensive ponds) (from Jingsong and Honglu, 1989).

water quality maintenance (i.e., for circulation and water aeration). Thus, cages in open water benefit from natural water circulation that removes waste metabolites and reoxygenates the water. Conversely, tank and pond systems on land can have high energy requirements for replacing, circulating, and aerating the water. Consequently, a semiintensive shrimp pond farm can consume as much as five times more energy for pumping and aeration activities compared to a salmon cage farm energy consumption for building and maintaining the cage structures (Table I). Although feed-related energy inputs typically account for a large proportion of the overall energy profile of an intensive culture system, the absolute scale of these inputs is very sensitive to the specific components of the feed being used. In general, fish- and livestock-derived inputs to feeds typically have much higher associated energy costs than do crop-derived components. Moreover, even within the context of a specific type of feed input, say fish meal and oils, energy inputs can vary widely between different sources and through time as conditions change. In the case of these fish-derived inputs, major sources of energy input variations result from differences in the fishing methods used, and in the relative abundance of the fish stocks being targeted. Similarly, the energy costs of agriculture products can also vary depending on the specific crops produced and the agriculture system they are derived from. This is of specific importance when aquaculture systems in industrialized and nonindustrialized countries are being compared.

To illustrate the relative energy costs associated with providing the major forms of feed inputs, Fig. 4 shows the energy inputs associated with producing a typical salmon feed used in Canada. By comparing Figs. 4A and 4B, we see that although over 50% of the mass of the feed is derived from fish, these inputs only represent about 35% of the total energy inputs. What is more interesting to note, however, is that although livestock-derived inputs represent only 16% of the feed by mass, they account for well over 40% of the total energy costs of this feed. In contrast, crop-derived inputs comprise over 25% of the feed but only account for approximately 5% of the total energy cost.

5. BIOPHYSICAL EFFICIENCY AND SUSTAINABILITY

Aquaculture is becoming an increasingly important component of the food production sector. As such, it is of interest to compare its energy performance with other food production systems. Table II provides an overview of the energy performance of selected food production systems, including fisheries and terrestrial crop and animal production, using a ratio of total industrial energy invested in the system relative to the edible protein energy return. By this measure, we find that the energy performance of intensive shrimp and salmon aquaculture is of a magnitude similar to that of shrimp and flatfish fisheries, as well as feedlot beef production. Seaweed, mussels, and some extensive fish aquaculture systems are comparable to sheep and rangeland beef farming. Interestingly, intensive rainbow trout cage culture has a surprisingly low energy ratio, similar to that of tilapia and mussel farming. Lobster and oyster farming systems in tanks represent the most energy-demanding aquaculture system for which data are available. It is important to note, however, in the case of both of these systems, the energy data represent experimental production systems from the late 1970s and early 1980s. As far as any

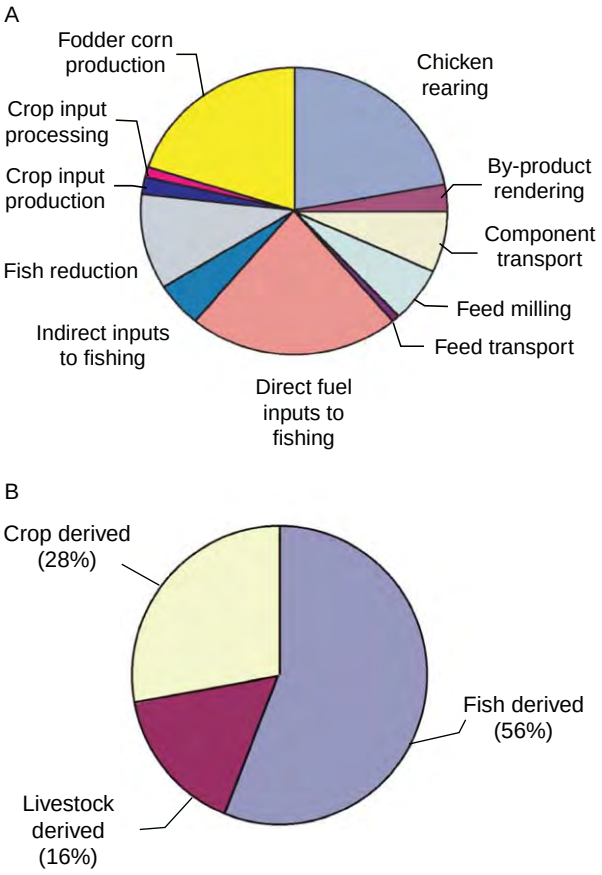


FIGURE 4 (A) Breakdown of industrial energy inputs to produce a generic salmon feed; (B) the major components of this feed. From Tyedmers (2000).

TABLE II

Ranking of Industrial Energy Inputs Excluding Labor per Protein Output in Aquaculture Compared with Capture Fisheries and Agriculture[a]

Food type	Industrial energy input/protein energy output (J/J)
Herring (purse, seine, NE Atlantic)	2–3
Vegetable crops	2–4
Carp farming (ponds, Indonesia, China, Israel)	1–9
Tilapia culture (extensive ponds, Indonesia)	8
Seaweed culture (West Indies)	7–5
Sheep farming	10
Rangeland beef farming	10
Cod fisheries (trawl and long line, North Atlantic)	10–12
Salmon fisheries (purse seine, gillnet, troll, NE Pacific)	7–14
Milk (USA)	14
Mussel culture (long line, Northern Europe)	10–20
Tilapia culture (semiintensive ponds, Africa)	18
Rainbow trout culture (cages, Finland, Ireland)	8–24
Catfish culture (ponds, USA)	25
Eggs (USA)	26
Broiler farming	34
Shrimp culture (Semiintensive Ponds, Ecuador)	40
Atlantic salmon culture (cages, Canada–Sweden)	40–50
Shrimp fisheries	17–53
Flatfish (trawl, NE Atlantic)	53
Beef feed-lot (USA)	53
Lobster fisheries (trawl, NE Atlantic, NW Pacific)	38–59
Seabass culture (Cages, Thailand)	67
Shrimp culture (intensive ponds, Thailand)	70
Oyster culture (tanks, USA)	136
Lobster culture (tanks, USA)	480

[a] Data for oyster farming and carp tank farming from the late 1970s; lobster farming data from the early 1980s. All other data mainly from the 1990s and subsequent years, with some few exceptions from late 1980s. Data from Tyedmers (2001), Watanabe and Okubo (1989), Tyedmers (2000), Larsson *et al.* (1994), Folke and Kautsky (1992), Pimentel *et al.* (1996), Seppàlà *et al.* (2001), Stewart (1994), Berg *et al.* (1996), Smith (1990), Ackefors (1994), Rawitscher (1978), Bardach (1980), and Sing and Pannu (1996).

potential changes in performance from experimental systems in the 1970s and 1980s to now, it would be expected that the energy performance of these systems has improved. This is due to decreased fish meal/oil content in the feed, to increased feeding efficiency (methods, feed characteristics, etc.), and probably also to increased technological efficiency (in machinery production, increased machinery efficiency, etc.).

The data in this table should not be discussed in isolation from other environmental considerations, i.e., various negative externalities, but instead should be seen as adding important pieces of information to such discussions. By expressing energy performance on the basis of protein energy output, an important aspect of food production is captured. However, it is necessary to acknowledge that fish make particularly valuable contributions to the human diet because they provide complete protein and are low in saturated and high in polyunsaturated fats compared to other animal protein sources. A further complicating factor, and one not usually addressed when using maximum protein yield as a basis for comparing energy performance, is the degree to which potentially available protein is actually utilized. This issue is potentially most problematic when comparisons are being made between food production systems set in different cultural contexts. For example, a catfish produced and consumed in the United States may actually result in a smaller portion of the available protein of the fish being consumed when compared to catfish consumed in many Asian countries, where a much larger proportion of the animal may be regarded as edible.

It has been argued that without a clear recognition of the aquaculture industry's large-scale dependency and impact on natural ecosystems and traditional societies, the aquaculture industry is unlikely to either develop to its full potential or continue to supplement ocean fisheries. Thus, only by defining and quantifying linkages between the aquaculture operation and its supporting systems, together with socioeconomic considerations, will it be possible to ensure sustainable development. The extent to which measurement of energy inputs to aquaculture operations contributes to our understanding of its sustainability is an important issue for further consideration. Does less need for industrial energy inputs imply increased sustainability? If a reduction in dependence on fossil fuels is considered the most important or only indicator of sustainability, the answer is clearly yes. No doubt, reduced pressure on valuable finite resources is important. However, from a more holistic perspective on aquaculture, it may not be the most

important factor for achieving sustainability. In other words, although energy analysis can certainly help select forms and types of culture systems that reduce demand for industrial energy, it cannot capture a wider range of negative environmental or socioeconomic externalities and benefits that result from production. For example, a comparison of EI inputs between a system that depends on renewable energy resources and a system more dependent on industrial energy does not reflect whether the extraction of renewable resources is being performed in a sustainable way (i.e., sources, extraction methods, socioeconomic conflicts). To do this requires additional tools and analysis. Industrial energy analysis should instead be viewed as a measure of one aspect of sustainability–not as a complete tool for evaluating sustainability. Its narrow focus on nonrenewable resource use, and in particular fossil fuel use, functionally sets the analytical boundaries at the point of material extraction. It does not value the wider resource base on which the activity depends.

Comparing results from different energy analyses of aquaculture systems is also not as straightforward as the literature may indicate. A common problem is the lack of detail regarding what has been included in the analysis. Frequently, information is missing regarding (a) how transportation has been addressed (e.g., including also transport from farm to processing), (b) detailed descriptions of the feed components used, (c) how the energy required to provide various feed components has been calculated, (d) if inputs to hatchery operations prior to grow-out are included, (e) how fixed infrastructure costs been accounted for, if at all, (f) what if any harvest-related energy costs are included (processing, storage, packing, and transportation), and (g) the size of the operation that EI usages are based on.

Further, there is a general problem when comparing EI inputs between different studies due to spatial and temporal variations in how raw materials are obtained and processed (different sources, extraction processes, distances and means of transport, and processing technology). This makes it difficult to generalize about a specific culture system. Also, comparisons of farming systems in industrialized countries with developing countries may result in a skewed picture because labor is much cheaper in developing countries, resulting in less dependence on industrial energy-consuming technologies. Finally, information regarding the size of the production unit being studied is important given that larger culture units may require less energy input per unit of production—essentially giving rise to "energy economies of scale".

Acknowledgments

The authors thank Alan Stewart, Craig Tucker, and Peter Edwards for help with data access, and Robert Kautsky for help with illustrations.

SEE ALSO THE FOLLOWING ARTICLES

Ecological Footprints and Energy • Ecological Risk Assessment Applied to Energy Development • Ecosystem Health: Energy Indicators • Emergy Analysis and Environmental Accounting • Fisheries and Energy Use • Industrial Agriculture, Energy Flows in • Livestock Production and Energy Use • Ocean, Energy Flows in • Sustainable Development: Basic Concepts and Application to Energy • Wetlands: Impacts of Energy Development in the Mississippi Delta

Further Reading

Ackefors, H., Huner, J. V., and Konikoff, M. (1994). Energy use in aquaculture production. *In* "Introduction to the General Principles of Aquaculture" (R. E. Gough, Ed.), pp. 51–54. Food Products Press, New York.

Andersen, O. (2002). Transport of fish from Norway: energy analysis using industrial ecology as the framework. *J. Cleaner Prodn.* **10,** 581–588.

Bardach, J. (1980). Aquaculture. *In* "Handbook of Energy Utilisation in Agriculture" (D. Pimentel, Ed.), pp. 431–437. CRC Press, Boca Raton.

Berg, H., Michélsen, P., Troell, M., and Kautsky, N. (1996). Managing aquaculture for sustainability in tropical lake Kariba, Zimbabwe. *Ecol. Econ.* **18,** 141–159.

Boyd, C. E., and Tucker, C. S. (1995). Sustainability of channel catfish production. *World Aquacult.* **26**(3), 45–53.

Costa-Pierce, B. A. (ed.). (2002). "Ecological Aquaculture." Blackwell Science, Oxford, UK.

Edwardson, W. (1976). The energy cost of fishing. *Fish. News Int.* **15**(2), 2.

Folke, C., and Kautsky, N. (1992). Aquaculture with its environment: prospects for sustainability. *Ocean Coast. Manage.* **17,** 5–24.

Larsson, J., Folke, C., and Kautsky, N. (1994). Ecological limitations and appropriation of ecosystem support by shrimp farming in Columbia. *Environ. Manage.* **18**(5), 663–676.

Mathews, S. B., Mock, J. B., Willson, K., and Senn, H. (1976). Energy efficiency of Pacific salmon aquaculture. *Progress. Fish-Cultur.* **38**(2), 102–106.

Naylor, L., Goldburg, R. J., Primavera, J., Kautsky, N., Beveridge, M., Clay, J., Folke, C., Lubchenco, J., Mooney, H., and Troell, M. (2000). Effect of aquaculture on world fish supplies. *Nature* **405,** 1017–1024.

Pimentel, D., Shanks, R. E., and Rylander, J. C. (1996). Bioethics of fish production: energy and the environment. *J. Agricult. Environ. Ethics* **9**(2), 144–164.

Rawitscher, M. A. (1978). Energy cost of nutrients in the American diet. Ph.D. thesis (unpubl.). University of Connecticut, Storrs, Connecticut.

Seppälä, J., Silvenius, F., Grönroos, J., Mäkinen, T., Silvo, K., and Storhammar, E. (2001). "Rainbow Trout Production and the Environment. The Finnish Environment" [in Finnish]. Suomen ympäristö 529. Technical report.

Sing, S., and Pannu, C. J. S. (1998). Energy requirements in fish production in the state of Punjab. *Energy Convers. Manage.* **39,** 911–914.

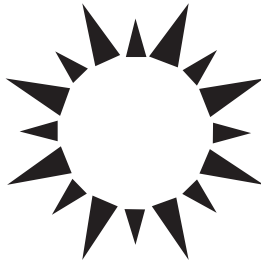
Stewart, J. A. (1995). "Assessing Sustainability of Aquaculture Development." Ph.D. thesis. University of Stirling, Stirling, Scotland.

Tyedmers, P. (2000). Salmon and sustainability: The biophysical cost of producing salmon through the commercial salmon fishery and the intensive salmon culture industry. Ph.D. dissertation. University of British Columbia, Vancouver, Canada.

Tyedmers, P. (2001). Energy consumed by north atlantic fisheries. Fisheries impacts on North Atlantic ecosystems: Catch, effort and national/regional datasets (D. Zeller, R. Watson, and D. Pauly, Eds.). *Fisheries Centre Res. Rep.* **9**(3), 12–34.

Watanabe, H., and Okubo, M. (1989). Energy input in marine fisheries of Japan. *Bull. Jpn. Soc. Scientif. Fisheries/Nippon Suisan Gakkaishi* **53**(9), 1525–1531.

Arid Environments, Impacts of Energy Development in



DAVID FAIMAN
Ben-Gurion University of the Negev
Beer-Sheva, Israel

1. **Arid Lands as a Key to Energy Stability**
2. **The Desert Ecosystem**
3. **Danger from Unbridled Development of Deserts**
4. **Technological Milestones for Safe Development of Arid Zones**
5. **Potential of Photovoltaic Technology**
6. **Other Available Technologies**
7. **Prospects and Conclusions**

Glossary

desertification The conversion to true desert, as a result of human activities, of dryland located on the margin between desert and fertile regions.

ecosystem The totality of interdependent living organisms in a given physical environment.

megawatt (MW) Generally, a unit of power ($= 10^6$ W). When used in the context of a photovoltaic system, megawatt units refer to the maximum power output of the photovoltaic modules under specific (and somewhat artificial) test conditions. For this reason, the power ratings of photovoltaic and nonphotovoltaic power plants are not directly comparable.

performance ratio The annual energy output of a photovoltaic system, at a particular site, divided by the power rating of its photovoltaic modules. This figure of merit is, to a large extent, independent of the specific kind of photovoltaic modules employed.

photovoltaic module A panel assembled from a number of individual photovoltaic cells connected in series and parallel so as to provide the required voltage and current specifications of the module. Photovoltaic cells use the energy of incoming light to generate an electric current via the so-called photovoltaic effect. A single silicon photovoltaic cell, with typical dimensions 10×10 cm, would typically give a voltage of 0.5 V, and a current of 2 A when exposed to strong sunlight.

solar thermal A method of generating power from solar energy; in contrast to the photovoltaic method, solar thermal produces heat as an intermediate stage (which may then be used to drive a turbogenerator).

terawatt-hour (TWh) A unit of electrical energy equal to 10^{12} watt-hours, or 3.6×10^{-3} exajoules (EJ).

There is a need to develop the world's arid areas; dual and related incentives to do so involve both raising living standards in the developing world and reducing fossil fuel consumption in the developed world. The focus here is on reasons why this may be possible and proposals on how it may come about. A number of possible impacts of development of arid regions are emphasized, although the list of such impacts cannot be exhaustive at the present time, given the enormous area of Earth's arid ecosystems and the huge range of variables that must be considered in proposing major changes. The purpose here is to outline the nature of the principal impacts and how it might be possible to control them.

1. ARID LANDS AS A KEY TO ENERGY STABILITY

The world is suffering increasingly from environmental problems, many of which, if not all, may be traced to the uncontrolled use of energy and related resources by an ever-increasing global population. On the one hand, the industrially developed part of the world consumes fossil fuel to generate the energy needs of its inhabitants and industries, and to provide transportation both on the ground and in the air. At the other end of the social spectrum, the developing world suffers from an acute dearth of the energy needed for development and, not unnaturally, clamors to right this imbalance. If this situation is not to lead to a globally catastrophic conclusion, it will be

 © 2004 Elsevier Inc. All rights reserved.

necessary to solve a number of urgent energy-related problems during the next few generations. First, the standard of living among all people on Earth must be raised to a high enough level to promote political stability. Second, the worldwide consumption of energy and the concomitant production of waste products must level off at a rate commensurate with environmental stability. Third, worldwide population growth must be reduced to zero.

Of course, each of these problems, though easy to state, encompasses a veritable Pandora's box of social and technical subproblems. As such, even to approach their solution will be a far from simple matter. However, it is clear that a start must be made, and for all three problems the energy development of the world's deserts may hold an important key. First, much of the world's poverty is located in arid regions. Second, solar energy is locally available for the development of those regions. Third, the deserts receive sufficient solar energy that, if it is harnessed, the deserts could provide nonpolluting power station replacements for the rest of the world. Fourth, there is sufficient available land area within the world's deserts to accommodate a large population increase during the time it will take to achieve the ultimately desired goal of global environmental stability.

Naturally, and this is often forgotten, there are many possible impacts (both positive and negative) of energy development in arid environments. It is therefore of vital importance to promote the positive impacts while minimizing the negative ones.

2. THE DESERT ECOSYSTEM

Certain regions on Earth are arid due to the natural circulation of the atmosphere. Warm, moist air, rising from the equatorial belt, cools and releases its precipitation in the tropics, and the resulting dry air returns to Earth, forming two desert regions at approximate latitudes of $\pm 30^{\circ}$. As a result of the dearth of rainfall in these regions, they have little natural vegetation and are consequently sparsely populated both by animals and humans. In spite of an apparent almost complete absence of life in deserts, these regions actually support a wealth of species, both floral and faunal. However, in absolute numbers, their abundances are low and they coexist as a very frail ecosystem. In fact, so frail is the desert ecosystem that even though the species have evolved to be able to withstand arid conditions, extended periods of drought periodically bring about extinctions. It is also possible, and even likely, that additional extinction of species in the arid regions comes about as a result of anthropogenic activities elsewhere on Earth. For example, unnatural damage to the ozone layer, caused by the worldwide use of chlorofluorocarbons (CFCs) and other chemicals, can result in the exposure of living species in the desert belts to elevated and deadly levels of solar ultraviolet-B (UV-B) radiation. Also, sparse as the rainfall in these regions is, the introduction of externally caused acid rain may also be expected to present a serious threat to the natural desert ecosystem.

Naturally, the encroachment of humans on the arid regions, albeit, in the past, in relatively small numbers, also causes serious perturbations to the natural desert ecosystem; desertification, the conversion of marginal desert areas to true deserts, is probably the most notorious example of the impact of human encroachment. It is apparent that today, although the deserts of the world are still relatively undeveloped, energy development elsewhere on Earth can have a negative impact on arid environment ecosystems. Of course, with the massive development of deserts that inevitably may be expected to occur, their natural ecosystems will vanish into history. However, whatever replaces them will clearly need to be taken into consideration because, as we are learning from the phenomenon of global warming, Nature has a tendency to fight back.

In the light of the relatively limited knowledge we possess today on desert ecosystems, it is obvious that there are some rather nontrivial questions that science would do well to try and answer, before unbridled technological development occurs in the arid regions. Specifically, it is important to ask what the respective effects are on a desert ecosystem of introducing various kinds of technology, what should be the criteria for assessing whether the results are good or bad, and what kinds of resulting ecosystems could be considered desirable.

3. DANGER FROM UNBRIDLED DEVELOPMENT OF DESERTS

In addition to being concerned about desert ecosystems, the entire global ecosystem must also be considered. There is already evidence that anthropogenic activities outside arid regions can have a negative impact on the desert ecosystem. However, the reverse can also be true. If the development of arid regions was traditionally limited in the past by the nature of their aridity, modern technology has proved that this need no longer be the case. In fact, the real limiting

factor is not water but, rather, energy. After all, were enough energy to be available, then, following the Saudi Arabian example, seawater could be desalinated and transported large distances, allowing food to be grown in even the most arid of deserts. So, suppose, for example, that solar energy could suddenly become available for the practical development of arid regions. The next step is to examine what the overall consequences could be for the world at large.

The first effect would be a massive improvement in the local standard of living. For example, a recent study employing matrices of input/output economic factors for Morocco calculated some of the effects of constructing a single photovoltaic module manufacturing facility in the Sahara Desert. The facility would, by supposition, be capable of producing 5 MW of photovoltaic modules per year. The authors of the study concluded that such a manufacturing facility would create more than 2500 new jobs. In order to place this result in its correct perspective, it is important to point out that the study was part of a scenario in which five such factories would be constructed initially, to be followed by four much larger factories, each with an annual module production of 50 MW. The combined enterprise would provide (a) enough electricity for operating the successive factories and (b) enough photovoltaic modules to enable the construction of a 1.5-GW photovoltaic electric power generating plant, a so-called very large solar photovoltaic plant (VLS-PV). The entire project would take, by calculation, 43 years to complete. The resulting annual electricity output would be 3.5 TWh, at a cost perhaps as low as $0.05/kWh (U.S. dollars)—enough electricity to provide nearly 1 million people with living standards comparable to those enjoyed in today's industrially developed countries. Of course, the concomitant cost reduction that would accompany such a large-scale manufacture of photovoltaic panels would also lead to their more widespread use on an individual family basis in outlying desert areas. Such improved living conditions may be expected to result in a corresponding population growth fueled by the migration of poor people into this new paradise on Earth. Housing and feeding these people should present little problem thanks to the plentiful availability of energy. This much is clearly desirable, because the large-scale relief of poverty may be expected to lead to greater political stability. But what would be achieved at the global environmental level?

Although the local generation of electric power from solar energy would be accompanied by a negligible generation of carbon dioxide, this is not necessarily the case for the associated growth in infrastructure that would result. In the first place, surface and air transport, both within the newly developed arid region and between it and the other developed parts of the world, would result in the release of some 3 kg of CO_2 for each 1 kg of fuel that was burned. If we take octane (C_8H_{18}) as being a typical constituent of motor fuel, then the complete combustion of one molecule of this hydrocarbon produces eight molecules of CO_2, or, in mass units, 1 kg of octane produces 3.1 kg of carbon dioxide. Second, there would be a corresponding increase in the worldwide production of construction materials and consumer goods, transported from places of manufacture where solar energy might not be available at prices that could compete with fossil fuel. Third, the benefits of newfound prosperity in the arid regions would likely lead to lower infant mortality and fewer deaths from disease, resulting in larger, longer lived families, hence increased energy consumption. Clearly, all of these energy-related side effects could lead to a serious overall increase in CO_2 production even though the original wherewithal—solar-generated electricity—is environmentally benign. Of course, solar-generated electricity may lead to a net per capita decrease in CO_2 production, but unless world population growth declines at a sufficiently rapid rate, this will still lead to an absolute increase in CO_2 production.

4. TECHNOLOGICAL MILESTONES FOR SAFE DEVELOPMENT OF ARID ZONES

In order avoid the various spin-off problems of arid region development, a number of parallel technological developments will have to reach fruition:

1. Solar energy (particularly photovoltaic) technology will have to reach an appropriate level of mass-producibility to enable large-area systems to be constructed in all major desert locations.
2. High-temperature (actually, ambient-temperature) superconductive cables will need to be developed, so as to enable the lossless transmission of solar-generated power from desert regions to the other, more industrially developed, parts of the world. This will enable large numbers of fossil-fueled power plants to be made redundant. Of course, environmentally clean electricity could then become a valuable export commodity for desert dwellers.

3. Hydrogen fuel cells and hydrogen-burning combustion engines will have to reach an appropriate level of development and scale of mass-production, to enable the replacement of today's hydrocarbon-based transportation fuel by solar-generated hydrogen. In this way, the present CO_2 waste product (and many others) would be replaced by water vapor. (H_2O is, of course, the major greenhouse gas in Earth's atmosphere, being some three to four times more potent than CO_2. However, it is important to realize that water vapor released by hydrogen combustion engines is generated from water in the first place, so it merely constitutes a part of the natural hydrological cycle of nature. It is here assumed that hydrogen is generated via the electrolysis of water, using solar-generated electricity.)

4. Nonpolluting air transport will need to be developed. This could be based (in situations where time considerations are not of pressing urgency) on the return of a more streamlined and safety-conscious form of hydrogen-based airship technology. Alternatively, new, hydrogen-fueled jet engines will need to be developed.

With the possible exception of pollution-free aeroengines, the technologies itemized here are already under intensive study.

5. POTENTIAL OF PHOTOVOLTAIC TECHNOLOGY

At this stage, it is worth pausing to examine some numbers. Table I lists the total world consumption of primary energy for the year 2001, according to geographic region. Also shown is the potential for photovoltaic power generation of the principal deserts in those regions. A number of clarifying comments may facilitate a better understanding of the tabulated data. First, primary energy consumption includes only commercially traded fuels. It excludes many of the important local fuels such as wood, peat, and animal waste, for which consumption statistics are hard to document. Second, the consumption of primary energy is given in the original source in units of tonnes of oil equivalent (TOE, which has here been converted to Système International units at the conversion rate of 1 TOE = 42 GJ). This begs the important question of how the primary energy is actually used. For low-grade heating purposes, the available energy from each TOE is comparatively high, but it is much lower if used for transportation fuel or electricity production. Thus, each of the tabulated entries for primary consumption should really be broken down into usage categories before comparison with solar availability. On the other hand, the tabulated photovoltaic potential of various desert areas assumes the generation of alternating current (AC) electricity via the use of photovoltaic cells. This is a relatively low-efficiency method (typically 10%) of harnessing solar energy compared with using solar collectors to generate thermal energy (with typical efficiency of 50%). Thus, the tabulated photovoltaic potential is an underestimate of the total solar potential of these deserts. Finally, the tabulated photovoltaic potential has been calculated assuming the use of 14% efficient photovoltaic modules covering each desert surface with a 0.5 spacing factor. What this means is that, in effect, only 50% of the total ground area is covered by photovoltaic panels, the remaining area being needed for access, for ancillary system components, and to prevent mutual shading among the panels (Fig. 1). A system performance ratio of 0.7 has been adopted as being typical of the range of values indicated by the detailed computer studies.

TABLE I

Regional Comparison of Primary Energy Consumption in 2001[a]

Geographic region	Primary energy consumption for the year 2001 (EJ)	Estimated photovoltaic energy capability of local deserts (EJ)
North America	110.9	458.6
South and Central America	19.0	209.2
Europe	79.6	—
Former Soviet Union	39.9	190.4
Middle East	16.7	1115.3
Africa	11.8	4005.4
Asia Pacific	105.5	1013.4
World total	383.2	6992.3

[a]Data from British Petroleum Company (2002) and the estimated photovoltaic electricity production capability of deserts in those regions (Kurokawa, 2003).

Turning now to Table I, we may distinguish between two sets of energy needs: the need to provide energy for the developing world and the need to replace the present usage of fossil fuel in the developed world by renewable energy. These two needs may be vividly contrasted on a map of the world (Fig. 2), on which light areas represent those regions that require energy for development and dark areas represent regions in which today's heavy energy

FIGURE 1 An array of photovoltaic panels at kibbutz Samar in the Negev Desert. Note the spacing between panels, necessary to prevent mutual shading and to allow access for cleaning. Photo courtesy of Bryan Medwed.

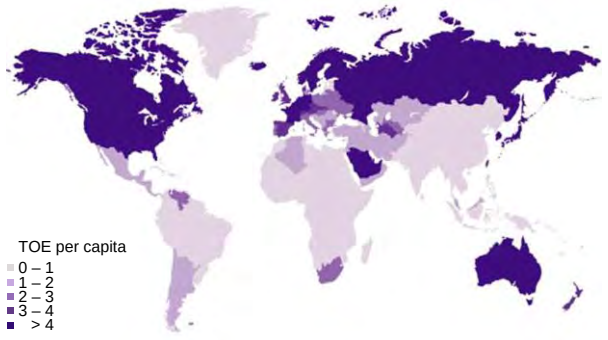


FIGURE 2 World map showing primary energy consumption per capita (2001). TOE, Tonnes of oil equivalent. From British Petroleum Company (2002).

users reside. Regarding the former, one may notice, from Table I, the startling difference between energy consumption and solar availability in both Africa and the Middle East. It is not impossible that much of the instability that has traditionally been associated with these regions could be alleviated by the systematic development of solar energy. Take just one of the Middle East's present-day complex issues, the Israeli–Palestinian conflict, and consider it purely within the context of solar energy development. Although Israel's Negev Desert ($12{,}000\,km^2$) is minuscule by world, or even Middle Eastern, standards, it occupies approximately 43% of the total land area of the state but houses only a negligible proportion of the country's 6 million inhabitants. Most of that population enjoys a high standard of living due, in large measure, to 9 GW of fossil-fueled electrical generating capacity. On the other hand, the Palestinians have only a poorly developed electrical infrastructure and a correspondingly low standard of living. If, however, a mere 1% of the Negev Desert were to be covered with photovoltaic panels, they could generate the electrical equivalent of a 2- to 3-GW fossil-fueled power plant, enough electrical energy for 2–3 million people. This small example from the Negev provides a vivid indication of the enormous clean energy potential that resides in the vast expanses of desert that cover the entire Middle East.

As regards Africa, there are two promising features in Table I. First, we note that Europe does not, at first glance, appear to have any desert energy-generating potential. However, that continent is only a short distance, by superconducting cable, from the Sahara, where the solar electricity generation potential is clearly enormous. Therefore, a mere 2% of Africa's photovoltaic potential could replace the equivalent of 100% of Europe's present consumption of primary fuel. This would require an area of desert only 400×400 km. Superconducting cable links to the European electricity grid could be made from Morocco to Spain or from Libya to Italy. Thus, solar-generated electricity could constitute an important future export commodity for the African continent. Second, the Sahara Desert also contains sufficient wasteland and solar irradiance to provide energy for all the developmental needs of the African continent in addition to those of Europe. The Sahara Desert is thus a perfect place for the large-scale development of solar energy,

Continuing with the need to replace fossil fuel consumption in the developed world, Table I indicates that North America, the world's major energy consumer, "may" not have sufficient desert area to enable solar energy to provide for all of its future needs. The word "may" is placed in quotation marks because, modulo the previously mentioned caveats, the photovoltaic potential of the North American deserts is a factor of four times higher than their present energy usage. However, it is clearly unfeasible to use all or most of these arid areas for such a purpose. This touches on a far less quantifiable impact of energy development in arid regions than has been hitherto discussed, related, of course, to the problem of severely changing the natural ecosystem, and to all of those scientific questions that need to be studied. But it is also a question of aesthetics. Many people justifiably regard deserts as places of exquisite, unspoiled beauty, sanctuaries for periodic retreat from the hustle and bustle of everyday life. Thus, covering the entire desert, or even 25% of it, with solar collectors would probably be regarded as socially unacceptable. Therefore, to

the various ecological questions suggested in Section II, there should be added a number of important sociological questions that will need to be answered before the energy development of arid regions can take place within the developed world.

6. OTHER AVAILABLE TECHNOLOGIES

The need to develop certain renewable technologies for widespread adoption both in arid regions and worldwide was mentioned in Section 4. These technologies were power transmission cables that would superconduct at ambient temperatures, hydrogen fuel cells and hydrogen combustion engines for surface and air travel, and low-cost photovoltaic solar electricity generators. In Section 5 the discussions elaborated the promise of very large-scale photovoltaic systems for two reasons. On the one hand, such discussions serve as a useful indicator as to the manner in which arid regions could serve as power stations of the future, and what some of the associated developmental impacts might be. But on the other hand, photovoltaics represents a technology that has been in existence for approximately half a century and is, consequently, relatively mature, albeit costly compared to fossil fuel alternatives. Nevertheless, it is important to remember that there are a number of alternative renewable technologies that have also been successfully demonstrated at the multimegawatt scale, and which may also be expected to play an important role in the development of arid regions.

Solar-thermal power stations have been built on the scale of tens of megawatts apiece, and some 350 MW of total electrical generating capacity have been operating in the California desert for the past 15–20 years. These power plants employ long rows of troughlike parabolic mirrors in order to concentrate sunlight onto a central tube. Oil flowing within the tube is heated to high temperatures and is used to convert water into steam that, in turn, drives a turbogenerator. Although the economic climate that prevailed in California when these systems were first constructed favored their use for electricity generation, it is important to realize that the thermal energy these systems generate as an intermediate stage represents a far more efficient use of solar energy. In these systems, the oil is heated to approximately 350°C, at an efficiency close to 50%. This heat is then used to power a steam turbine at an efficiency of about 30%. Thus, the combined solar-to-electrical efficiency is only around 15%. It was previously mentioned that for many situations in which thermal energy is required (e.g., industrial processes), the use of photovoltaic electricity would be needlessly inefficient. It is understandable, therefore (Fig. 3), how, in addition to providing electricity, such rows of parabolic trough reflectors could provide solar-powered thermal energy for the industries of desert towns in the future.

Wind power represents another technology that has been successfully demonstrated on a large scale (Fig. 4). This technology could usefully supplement the employment of solar energy in arid zones because its generating capacity is not limited to the daylight hours. In fact, at certain locations and with careful system design, it may be possible to use the same parcels of land for power generation using both solar and wind technologies.

Of course, as the variety of energy-producing technologies increases, so too, to a certain extent,

FIGURE 3 Solar-thermal electric power plant at Kramer Junction, California, United States. Photo courtesy of Solel Solar Systems Ltd.

FIGURE 4 Wind turbines at Jaisalmer, India. Photo courtesy of Safrir Faiman.

will be the variety of additional impacts that will have to be studied. The possibility of oil leakage and fire hazard will be among topics of relevance if solar-thermal systems are to be employed on a very large scale, as will the possible effects of wind turbines on bird life and human hearing.

7. PROSPECTS AND CONCLUSIONS

Historically, technological advances have occurred piecemeal, being driven, by and large, by motives of greed. It is only comparatively recently that the importance of the larger picture has come to be realized: namely, the effects of technological development on the environment. Hand in glove with such technological advances, important social issues are also more appreciated today than in the past.

The development of deserts might be allowed to come about driven purely by the profit motive. For example, photovoltaic manufacturers will naturally be interested in pursuing the potentially large markets that exist in the world's desert regions. With capital investment from the outside, such actions will certainly improve local living standards. However, as was indicated previously, if this occurs too rapidly, on a large enough scale, it could lead to an alarming increase of induced CO_2 production at the global level, as the industrial countries increase their output of consumer goods. Ideally, what is needed is an orchestrated approach by industry and government among the wealthier nations. The first priority should be to raise living standards in the developing world by the provision of environmentally clean electricity. Not only will this help stabilize many of those regions in which poverty is rife, but it will also help bring on that time when all the world's population can, together, discuss what is best for everybody. Of course, as the use of consumer goods increases among the newly wealthy parts of the world, it may be necessary for the previously developed nations to take ever more stringent measures to check their own usage of energy. This will not necessarily mean reducing living standards, but it will mean heavier investment into energy efficiency and research into new methods of alternative energy usage. Mention was made of the importance of hydrogen fuel cells and the need to develop hydrogen combustion engines for surface and air transport. But, no doubt, many other developments may be expected to occur as the research proceeds.

SEE ALSO THE FOLLOWING ARTICLES

Aircraft and Energy Use • Fuel Cell Vehicles • Global Energy Use: Status and Trends • Hydrogen, End Uses and Economics • Photovoltaic Energy: Stand-Alone and Grid-Connected Systems • Photovoltaics, Environmental Impact of • Population Growth and Energy • Solar Thermal Energy, Industrial Heat Applications • Solar Thermal Power Generation

Further Reading

British Petroleum Company (BP). (2002). "BP Statistical Review of World Energy," 51st Ed. BP Corporate Communication Services, London. Available at the corporate Web site at http://www.bpamoco.com.

Johansson, T.B., Kelly, H., Reddy, A.K.N., and Williams, R.H. (eds.). (1993). Renewable Energy: Sources for Fuels and Electricity. Island Press, Washington, D.C.

Kurokawa, K. (ed.). (2003). "Energy from the Desert: Feasibility of Very Large Scale Photovoltaic Power Generation (VLS-PV) Systems." James & James, London.

Rabl, A. (1985). "Active Solar Collectors and Their Applications." Oxford Univ. Press, New York.

Batteries, Overview

ELTON J. CAIRNS
Lawrence Berkeley National Laboratory and
University of California, Berkeley
Berkeley, California, United States

1. How Batteries Work
2. History
3. Types of Batteries and Their Characteristics
4. Applications
5. Environmental Issues
6. Future Outlook

Glossary

battery An assemblage of cells connected electrically in series and/or parallel to provide the desired voltage and current for a given application.
cell An electrochemical cell composed of a negative electrode, a positive electrode, an electrolyte, and a container or housing.
electrode An electronically conductive structure that provides for an electrochemical reaction through the change of oxidation state of a substance; may contain or support the reactant or act as the site for the electrochemical reaction.
electrolyte A material that provides electrical conduction by the motion of ions only; may be a liquid, solid, solution, polymer, mixture, or pure substance.
primary cell A cell that cannot be recharged and is discarded after it is discharged.
secondary cell A rechargeable cell.

Batteries are an important means of generating and storing electrical energy. They are sold at a rate of several billions of dollars per year worldwide. They can be found in nearly all motor vehicles (e.g., automobiles, ships, aircraft), all types of portable electronic equipment (e.g., cellular phones, computers, portable radios), buildings (as backup power supplies), cordless tools, flashlights, smoke alarms, heart pacemakers, biomedical instruments, wristwatches, hearing aids, and the like. Batteries are so useful and ubiquitous, it is difficult to imagine how life would be without them. Batteries, strictly speaking, are composed of more than one electrochemical cell. The electrochemical cell is the basic unit from which batteries are built. A cell contains a negative electrode, a positive electrode, an electrolyte held between the electrodes, and a container or housing. Cells may be electrically connected to one another to form the assembly called a battery. In contrast to the alternating current available in our homes from the electric utility company, batteries deliver a direct current that always flows in one direction. There are a few different types of batteries: Primary batteries can be discharged only once and then are discarded; they cannot be recharged. Secondary batteries are rechargeable. Forcing current through the cells in the reverse direction can reverse the electrochemical reactions that occur during discharge. Both primary and secondary batteries can be categorized based on the type of electrolyte they use: aqueous, organic solvent, polymer, ceramic, molten salt, and so on.

1. HOW BATTERIES WORK

The electrical energy produced by batteries is the result of spontaneous electrochemical reactions. The driving force for the reactions is the Gibbs free energy of the reaction, which can be calculated easily from data in tables of thermodynamic properties. The maximum voltage that a cell can produce is calculated from this simple relationship:

$$E = -\Delta G/nF, \qquad (1)$$

where E is the cell voltage, ΔG is the Gibbs free energy change for the cell reaction (joules/mole), n is the number of electrons involved in the reaction (equivalents/mole), and F is the Faraday constant (96487 coulombs/equivalent).

It is clear from Eq. (1) that a large negative value for the Gibbs free energy of the cell reaction is desirable if we wish to have a significant cell voltage.

 © 2004 Elsevier Inc. All rights reserved.

In practical terms, cell voltages of 1 to 2 V are achieved when using aqueous electrolytes, and cell voltages up to approximately 4 V are achieved when using nonaqueous electrolytes. When larger voltages are required, it is necessary to connect a number of cells electrically in series. For example, 12-volt automotive batteries are composed of six 2-volt cells connected in series.

An important property of batteries is the amount of energy they can store per unit mass. This is the specific energy of the battery, usually expressed in units of watt-hours of energy per kilogram of battery mass (Wh/kg). The maximum value of the specific energy is that which can be obtained from a certain mass of reactant materials, assuming the case that any excess electrolyte and terminals have negligible mass. This is called the theoretical specific energy and is given by this expression:

$$\text{Theoretical Specific Energy} = -\Delta G/\Sigma Mw, \quad (2)$$

where the denominator is the summation of the molecular weights of the reactants.

It can be seen from Eq. (2) that the theoretical specific energy is maximized by having a large negative value for ΔG and a small value for ΣMw. The value of ΔG can be made large by selecting for the negative electrode those reactant materials that give up electrons very readily. The elements with such properties are located on the left-hand side of the periodic chart of the elements. Correspondingly, the positive electrode reactant materials should readily accept electrons. Elements of this type are located on the right-hand side of the periodic chart. Those elements with a low equivalent weight are located toward the top of the periodic chart. These are useful guidelines for selecting electrode materials, and they help us to understand the wide interest in lithium-based batteries now used in portable electronics. Equations (1) and (2) give us a useful framework for representing the theoretical specific energy and the cell voltage for a wide range of batteries, as shown in Fig. 1. The individual points on the plot were calculated from the thermodynamic data. These points represent theoretical maximum values. The practical values of specific energy that are available in commercial cells are in the range of one-fifth to one-third of the theoretical values. As expected, the cells using high equivalent weight materials (e.g., lead, lead dioxide) have low specific energies, whereas those using low equivalent weight materials (e.g. lithium, sulfur) have high specific energies. The lines in Fig. 1 represent the cell voltages and simply represent the relationships given by Eqs. (1) and (2).

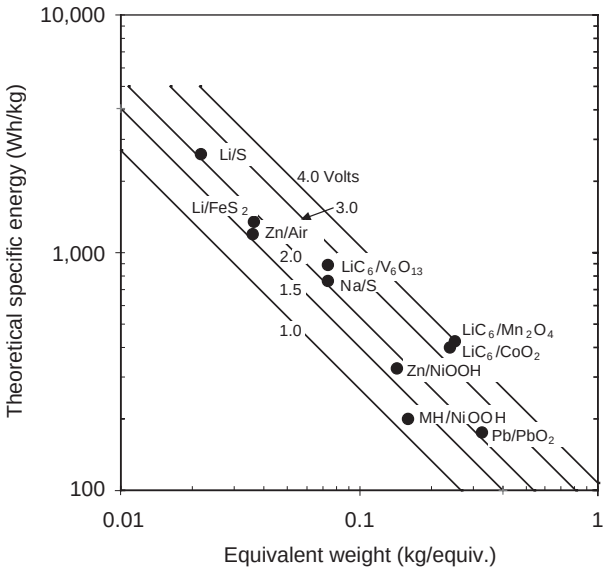


FIGURE 1 Theoretical specific energy for various cells as a function of the equivalent weights of the reactants and the cell voltage.

The details of the electrochemical reactions in cells vary, but the principles are common to all. During the discharge process, an electrochemical oxidation reaction takes place at the negative electrode. The negative electrode reactant (e.g., zinc) gives up electrons that flow into the electrical circuit where they do work. The negative electrode reactant is then in its oxidized form (e.g., ZnO). Simultaneously, the positive electrode reactant undergoes a reduction reaction, taking on the electrons that have passed through the electrical circuit from the negative electrode (e.g., MnO_2 is converted to MnOOH). If the electrical circuit is opened, the electrons cannot flow and the reactions stop.

The electrode reactions discussed in the preceding can be written as follows:

$$Zn + 2OH^- = ZnO + H_2O + 2e^-$$

and

$$2MnO_2 + 2H_2O + 2e^- = 2MnOOH + 2OH^-.$$

The sum of these electrode reactions is the overall cell reaction:

$$Zn + 2MnO_2 + H_2O = ZnO + 2MnOOH.$$

Notice that there is no net production or consumption of electrons. The electrode reactions balance exactly in terms of the electrons released by the negative electrode being taken on by the positive electrode.

2. HISTORY

Volta's invention of the Volta pile in 1800 represents the beginning of the field of battery science and engineering. The pile consisted of alternating layers of zinc, electrolyte soaked into cardboard or leather, and silver. Following Volta's report, many investigators constructed electrochemical cells for producing and storing electrical energy. For the first time, relatively large currents at high voltages were available for significant periods of time. Various versions of the pile were widely used. Unfortunately, there was a lack of understanding of how the cells functioned, but as we now know, the zinc was electrochemically oxidized and the native oxide layer on the silver was reduced. The cells were "recharged" by disassembling them and exposing the silver electrodes to air, which reoxidized them. Inevitably, many other electrochemical cells and batteries were developed.

John F. Daniell developed a two-fluid cell in 1836. The negative electrode was amalgamated zinc, and the positive electrode was copper. The arrangement of the cell is shown in Fig. 2. The copper electrodes were placed in (porous) porcelain jars, which were surrounded by cylindrical zinc electrodes and placed in a larger container. A copper sulfate solution was placed in the copper electrode's compartment, and sulfuric acid was put in the zinc electrode compartment.

The electrode reactions were as follows:

$$Zn = Zn^{+2} + 2e^{-}$$

and

$$Cu^{+2} + 2e^{-} = Cu.$$

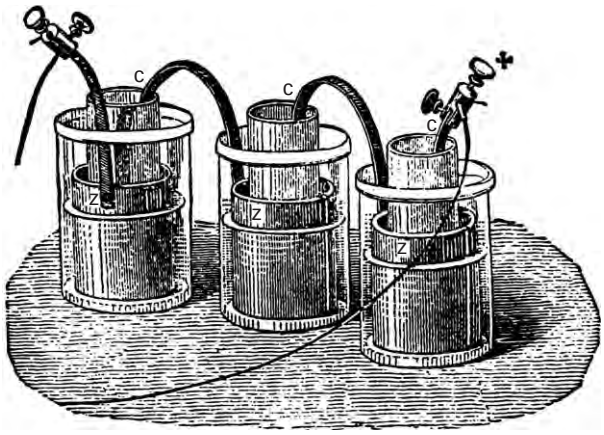


FIGURE 2 A set of three Daniell cells connected in series.

Sir William R. Grove, a lawyer and inventor of the fuel cell, developed a two-electrolyte cell related to the Daniell cell in 1839. Grove used fuming nitric acid at a platinum electrode (the positive) and zinc in sulfuric acid (the negative). Variants on this formulation were popular for a number of years. Of course, all of these cells were primary cells in that they could be discharged only once and then had to be reconstructed with fresh materials.

Gaston Planté invented the first rechargeable battery in 1860. It was composed of lead sheet electrodes with a porous separator between them, spirally wound into a cylindrical configuration. The electrolyte was sulfuric acid. These cells displayed a voltage of 2.0 V, an attractive value. During the first charging cycles, the positive electrode became coated with a layer of PbO_2. The charging operation was carried out using primary batteries—a laborious process. Because of the low cost and the ruggedness of these batteries, they remain in widespread use today, with various evolutionary design refinements.

The electrode reactions during discharge of the lead–acid cell are as follows:

$$Pb + H_2SO_4 = PbSO_4 + 2H^{+} + 2e^{-}$$

and

$$PbO_2 + H_2SO_4 + 2H^{+} + 2e^{-} = PbSO_4 + 2H_2O.$$

Notice that sulfuric acid is consumed during discharge and that the electrolyte becomes more dilute (and less dense). This forms the basis for determining the state of charge of the battery by measuring the specific gravity of the electrolyte. The theoretical specific energy for this cell is 175 Wh/kg, a rather low value compared with those of other cells.

Waldemar Jungner spent much of his adult life experimenting with various electrode materials in alkaline electrolytes. He was particularly interested in alkaline electrolytes because there generally was no net consumption of the electrolyte in the cell reactions. This would allow for a minimum electrolyte content in the cell, minimizing its weight. Jungner experimented with many metals and metal oxides as electrode materials, including zinc, cadmium, iron, copper oxide, silver oxide, and manganese oxide.

In parallel with Jungner's efforts in Sweden, Thomas Edison in the United States worked on similar ideas using alkaline electrolytes and many of the same electrode materials. Patents to these two inventors were issued at nearly the same time in 1901.

During the period since the beginning of the 20th century, a wide variety of cells have been investigated, with many of them being developed into commercial

products for a wide range of applications. Representative cells are discussed in the next section.

3. TYPES OF BATTERIES AND THEIR CHARACTERISTICS

Batteries are usually categorized by their ability to be recharged or not, by the type of electrolyte used, and by the electrode type. Examples of various types of electrolytes used in batteries are presented in Table I.

Primary (nonrechargeable) cells with aqueous electrolytes are usually the least expensive and have reasonably long storage lives. Historically, the most common primary cell is the zinc/manganese dioxide cell used for flashlights and portable electronics in sizes from a fraction of an amp-hour to a few amp-hours. Various modifications of this cell have been introduced, and now the most common version uses an alkaline (potassium hydroxide) electrolyte. The materials in this cell are relatively benign from an environmental point of view. The electrode reactions in an alkaline electrolyte may be represented as follows:

$$Zn + 2OH^- = ZnO + H_2O + 2e^-$$

and

$$2MnO_2 + 2H_2O + 2e^- = 2MnOOH + 2OH^-.$$

The cell potential is 1.5 V, and the theoretical specific energy is 312 Wh/kg. Practical cells commonly yield 40 to 50 Wh/kg.

Another common primary cell is the zinc/air cell. It uses an aqueous potassium hydroxide electrolyte, a zinc negative electrode, and a porous catalyzed carbon electrode that reduces oxygen from the air. The overall electrode reactions are as follows:

$$Zn + 2OH^- = ZnO + H_2O + 2e^-$$

and

$$\tfrac{1}{2}O_2 + H_2O + 2e^- = 2OH^-.$$

TABLE I

Electrolyte Types and Examples

Aqueous electrolyte	H_2SO_4, KOH
Nonaqueous electrolyte	
Organic solvent	Li salt in ethylene carbonate-diethyl carbonate
Solid	
Crystalline	$Na_2O \times 11Al_2O_3$
Polymeric	Li salt in polyethylene
Molten salt (high temperature)	LiCl–KCl

The sum of these reactions gives this overall cell reaction:

$$Zn + \tfrac{1}{2}O_2 = ZnO.$$

The cell voltage is 1.6 V, and the theoretical specific energy is 1200 Wh/kg. Practical specific energy values of 200 Wh/kg have been achieved for zinc/air cells. This cell has a very high energy per unit volume, 225 Wh/L, which is important in many applications. The materials are inexpensive, so this cell is very competitive economically. This cell is quite acceptable from an environmental point of view.

The zinc/mercuric oxide cell has the unique characteristic of a very stable and constant voltage of 1.35 V. In applications where this is important, this is the cell of choice. It uses an aqueous potassium hydroxide electrolyte and is usually used in small sizes. The electrode reactions are as follows:

$$Zn + 2OH^- = ZnO + H_2O + 2e^-$$

and

$$HgO + H_2O + 2e^- = Hg + 2OH^-.$$

The high equivalent weight of the mercury results in the low theoretical specific energy of 258 Wh/kg. Because mercury is toxic, there are significant environmental concerns with disposal or recycling.

The zinc/silver oxide cell has a high specific energy of about 100 Wh/kg (theoretical specific energy = 430 Wh/kg), and because of this, the cost of the silver is tolerated in applications that are not cost-sensitive. Potassium hydroxide is the electrolyte used here. The electrode reactions are as follows:

$$Zn + 2OH^- = ZnO + H_2O + 2e^-$$

and

$$AgO + H_2O + 2e^- = Ag + 2OH^-.$$

This gives the following overall cell reaction:

$$Zn + AgO = Ag + ZnO.$$

The cell voltage ranges from 1.8 to 1.6 V during discharge. Because of the high density of the silver oxide electrode, this cell is quite compact, giving about 600 Wh/L.

Primary (nonrechargeable) cells with nonaqueous electrolytes have been under development since the 1960s, following the reports of Tobias and Harris on the use of propylene carbonate as an organic solvent for electrolytes to be used with alkali metal electrodes. It has long been recognized that lithium has very low equivalent weight and electronegativity (Table II). These properties make it very attractive

TABLE II

Some Properties of Lithium and Zinc

	Lithium	Zinc
Equivalent weight (g/equiv.)	6.94	32.69
Reversible potential (V)	−3.045	−0.763
Electronegativity	0.98	1.65
Density (g/cm^3)	0.53	7.14

for use as the negative electrode in a cell. Because lithium will react rapidly with water, it is necessary to use a completely anhydrous electrolyte. There are many organic solvents that can be prepared free of water. Unfortunately, organic solvents generally have low dielectric constants, making them poor solvents for the salts necessary to provide electrolytic conduction. In addition, organic solvents are thermodynamically unstable in contact with lithium, so solvents that react very slowly and form thin, conductive, protective films on lithium are selected.

Perhaps the most common lithium primary cell is that of Li/MnO_2. It has the advantage of high voltage (compared with aqueous electrolyte cells), 3.05 V, and a high specific energy ($\sim$170 Wh/kg). The electrode reactions are as follows:

$$Li = Li^+ + e^-$$

and

$$MnO_2 + Li^+ + e^- = LiMnO_2.$$

The electrolyte is usually $LiClO_4$ dissolved in a mixture of propylene carbonate and 1,2-dimethoxyethane. In general, the specific power that can be delivered by these cells is less than that available from aqueous electrolyte cells because the ionic conductivity of the organic electrolyte is much lower than that of aqueous electrolytes.

An interesting cell is the one using fluorinated carbon as the positive electrode material. Because this material is poorly conducting, carbon and titanium current collection systems are used in the positive electrode. The electrode reaction is as follows:

$$(CF)_n + nLi^+ + ne^- = nLiF + nC.$$

The electrolyte commonly is $LiBF_4$ dissolved in gamma butyrolactone.

An unusual primary cell is that of lithium/thionyl chloride ($SOCl_2$). The thionyl chloride is a liquid and can act as the solvent for the electrolyte salt ($LiAlCl_4$) and as the reactant at the positive electrode. This cell can function only because of the relatively stable thin protective film that forms on the lithium electrode, protecting it from rapid spontaneous reaction with the thionyl chloride. The cell reaction mechanism that is consistent with the observed products of reaction is the following:

$$4Li + 2SOCl_2 = 4LiCl + 2SO'$$

and

$$2SO' = SO_2 + S,$$

where SO′ is a radical intermediate that produces SO_2 and sulfur. This cell provides a high cell voltage of 3.6 V and a very high specific energy but can be unsafe under certain conditions. Specific energies of up to 700 Wh/kg have been achieved compared with the theoretical value of 1460 Wh/kg.

Table III summarizes the characteristics of several representative primary cells.

Primary cells with molten salt electrolytes are commonly used in military applications that require a burst of power for a short time (a few seconds to a few minutes). These batteries are built with an integral heating mechanism relying on a chemical reaction to provide the necessary heat to melt the electrolyte (at 400–500°C). A typical cell uses lithium as the negative electrode and iron disulfide as the positive electrode. The following are example electrode reactions:

$$Li = Li^+ + e^-$$

and

$$FeS_2 + 4Li = 2Li_2S + Fe.$$

The molten salt electrolyte (a mixture of alkali metal chlorides) has a very high conductivity, allowing the cell to operate at very high specific power levels in excess of 1 kW/kg.

Rechargeable cells with aqueous electrolytes have been available for more than 140 years, beginning with the Planté cell using lead and sulfuric acid as discussed previously. Modern Pb/PbO_2 cells and batteries have received the benefit of many incremental improvements in the design and optimization of the system. Current versions are very reliable and inexpensive compared with competitors. Depending on the design of the cell, lifetimes vary from a few years to more than 30 years. Specific energy values of up to approximately 40 Wh/kg are available.

Alkaline electrolyte systems are available with a variety of electrodes. Perhaps the most common alkaline electrolyte cell is the Cd/NiOOH cell. It offers very long cycle life (up to thousands of charge–discharge cycles) and good specific power (hundreds

TABLE III

Properties of Some Lithium Primary Cells

	Open circuit voltage (V)	Practical specific energy (Wh/kg)	Theoretical specific energy (Wh/kg)
Li/MnO_2	3.0	280	750
$Li/(CF)_n$	2.8	290	2100
$Li/SOCl_2$	3.6	450–700	1460
Li/FeS_2	1.7	220	760

of W/kg), albeit with a modest specific energy (35–55 Wh/kg). The cell voltage, 1.2 V, is lower than most and can be somewhat of a disadvantage. The electrode reactions are as follows:

$$Cd + 2OH^- = Cd(OH)_2 + 2e^-$$

and

$$2NiOOH + 2H_2O + 2e^- = 2Ni(OH)_2 + 2OH^-.$$

Of course, the reverse of these reactions takes place on recharge.

A very robust rechargeable cell is the Edison cell, which uses iron as the negative electrode, nickel oxyhydroxide as the positive electrode, and an aqueous solution of 30 w/o potassium hydroxide as the electrolyte. During the early years of the 20th century, Edison batteries were used in electric vehicles. They proved themselves to be very rugged and durable, although they did not have high performance. The iron electrode reaction is as follows:

$$Fe + 2OH^- = Fe(OH)_2 + 2e^-.$$

The positive electrode reaction is the same as for the Cd/NiOOH cell described earlier.

The NiOOH electrode has proven itself to be generally the best positive electrode for use in alkaline electrolytes and has been paired with many different negative electrodes, including Cd, Fe, H_2, MH, and Zn. Recently, the metal hydride/nickel oxyhydroxide cell (MH/NiOOH) has been a significant commercial success and has captured a large fraction of the market for rechargeable cells. It offers a sealed, maintenance-free system with no hazardous materials. The performance of the MH/NiOOH cell is very good, providing up to several hundred watts/kilogram peak power, up to about 85 Wh/kg, and several hundred cycles. The negative electrode operates according to the following reaction:

$$MH + OH^- = M + H_2O + e^-.$$

The very high theoretical specific energy and low materials cost of the zinc/air cell make it attractive for consumer use. There have been many attempts to develop a rechargeable zinc/air cell, but with limited success due to the difficulties of producing a high-performance rechargeable air electrode.

The electrode reactions during discharge for this cell are as follows:

$$Zn + 2OH^- = ZnO + H_2O + 2e^-$$

and

$$O_2 + 2H_2O + 4e^- = 4OH^-.$$

During recent years, the cycle life of the air electrode has been improved to the point where more than 300 cycles are now feasible. Development efforts continue. In the meantime, various versions of a "mechanically rechargeable" zinc/air cell have been tested. These cells have provision for removing the discharged zinc and replacing it with fresh zinc. This approach avoids the difficulties of operating the air electrode in the recharge mode but creates the need for recycling the spent zinc to produce new zinc electrode material.

The status of some rechargeable cells with aqueous electrolytes is shown in Table IV.

Rechargeable cells with nonaqueous electrolytes have been under development for many years, although they have been available on the consumer market for only a decade or so. All of the types of nonaqueous electrolytes shown in Table I have been used in a variety of systems.

Organic solvent-based electrolytes are the most common of the nonaqueous electrolytes and are found in most of the rechargeable lithium cells available today. A typical electrolyte consists of a mixture of ethylene carbonate and ethyl-methyl carbonate, with a lithium salt such as $LiPF_6$ dissolved in it. Various other solvents, including propylene carbonate, dimethyl carbonate, and diethyl carbonate, have been used. Other lithium salts that have been used include lithium perchlorate, lithium hexafluoro arsenate, and lithium tetrafluoro borate. All of these combinations of solvents and salts yield electrolytes that have much lower conductivities than do typical aqueous electrolytes. As a result, the electrodes and electrode spacing in these lithium cells are made very thin to minimize the cell resistance and maximize the power capability.

Lithium is difficult to deposit as a smooth compact layer in these organic electrolytes, so a host material, typically carbon, is provided to take up the

TABLE IV

Rechargeable Aqueous Battery Status

System	Cell voltage (V)	Theoretical specific energy (Wh/kg)	Specific energy (Wh/kg)	Specific power (W/kg)	Cycle life	Cost (dollars/kWh)
Pb/PbO_2	2.10	175	30–45	50–100	>700	60–125
Cd/NiOOH	1.20	209	35–55	400	2000	>300
Fe/NiOOH	1.30	267	40–62	70–150	500–2000	>100
H_2/NiOOH	1.30	380	60	160	1000–2000	>400
MH/NiOOH	1.20	200	60–85	200+	500–600	>400
Zn/NiOOH	1.74	326	55–80	200–300	500+	150–250 (est.)
Zn/Air	1.60	1200	65–120	<100	300	100 (est.)

Note. est., estimate.

Li on recharge and deliver it as Li ions to the electrolyte during discharge. Many types of carbon, both graphitic and nongraphitic, have been used as the lithium host material. In addition, a variety of intermetallic compounds and metals have been used as Li host materials. All of the commercial Li rechargeable cells today use a carbon host material as the negative electrode.

The positive electrodes of Li cells are usually metal oxide materials that can intercalate Li with a minimal change in the structure of the material. The most common positive electrode material used in commercial Li cells is $LiCoO_2$. This material is very stable toward the removal and reinsertion of Li into its structure. It can be cycled more than 1000 times with little change in capacity.

The electrode reactions during discharge for this cell are as follows:

$$xLi(C) = xLi^+ + C + xe^-$$

and

$$Li_{1-x}CoO_2 + xLi^+ + xe^- = LiCoO_2.$$

Excellent engineering of this cell has resulted in a specific energy of more than 150 Wh/kg and a peak specific power of more than 200 W/kg, making it very desirable for applications requiring high specific energy. This cell has been manufactured extensively first in Japan and now in Korea and China, and it has been used worldwide.

Even though the $Li(C)/CoO_2$ cell has many desirable features, it has some limitations and disadvantages. The use of Co represents an environmental hazard because Co is toxic. It is also too expensive to use in large batteries suitable for low-cost applications such as electric and hybrid vehicle propulsion. In addition, there are safety problems associated with overcharge of this cell. Significant overcharge can cause the cell to vent and burn. Operation at elevated temperatures can also cause fire. Special microcircuits mounted on each cell provide control over the cell, preventing overcharge and overdischarge.

During the past several years, there has been a large effort to replace the $LiCoO_2$ with a less expensive, more environmentally benign material that can withstand overcharge and overdischarge while retaining the excellent performance of $LiCoO_2$. Many metal oxide materials have been synthesized and evaluated. Some of them are promising, but none has surpassed $LiCoO_2$ so far. Popular candidate materials include lithium manganese oxides such as $LiMn_2O_4$, $LiMnO_2$, and materials based on them, with some of the Mn replaced by other metals such as Cr, Al, Zn, Ni, Co, and Li. Other candidates include $LiNiO_2$, $LiFePO_4$, and variants on these, with some of the transition metal atoms replaced by other metal atoms in an attempt to make the materials more stable to the insertion and removal of the Li during discharge and charge. Progress is being made, and some Mn-based electrode materials are appearing in commercial cells.

Some of the instabilities that have been observed are traceable to the electrolyte and its reactivity with the electrode materials. This has resulted in some research on more stable solvents and Li salts. The Li salts must be stable to voltages approaching 5 V and must be very soluble and highly conductive. The solvents must have a wide temperature range of liquidity so that cells can be operated from very cold to very warm temperatures (-40 to $+60^\circ C$ for some applications).

The emphasis on lower cost electrode materials has continued, and some encouraging results have been obtained for materials based on $LiFePO_4$. This

material is very stable toward repeated insertion and removal of Li and has shown hundreds of cycles with little capacity loss. One problem with $LiFePO_4$ is its low electronic conductivity, making full use and high rate operation difficult, even with enhanced current collection provided by added carbon.

Some unusual electrode materials are under investigation in the quest to reduce cost. These include elemental sulfur and organo-disulfide polymers. The Li/S cell has been investigated several times during the past 20 years or more. From a theoretical specific energy point of view, it is very attractive at 2600 Wh/kg. The cost of sulfur is extremely low, and it is environmentally benign. A significant difficulty with the sulfur electrode in most organic solvents is the formation of soluble polysulfides during discharge. The soluble polysulfides can migrate away from the positive electrode and become inaccessible, resulting in a capacity loss.

The overall reaction for the sulfur electrode is as follows:

$$S + 2e^- = S^{-2}.$$

And the overall cell reaction is as follows:

$$2Li + S = Li_2S.$$

Recent work on this system has involved isolating the soluble polysulfides with an ionically conductive membrane or using solvents that have a low solubility for the polysulfides that are formed during discharge. These approaches have resulted in a large increase of cycle life, up to several hundred cycles.

The organo–disulfides offer a lower specific energy than does elemental sulfur, but they do not exhibit the same relatively high solubility. Long cycle lives have been demonstrated for these electrodes. They are usually coupled with a lithium metal negative electrode, which has a short cycle life. Current efforts are under way to improve the cycle life of the lithium metal electrode.

The status of some rechargeable cells with nonaqueous electrolytes is presented in Table V. The performance of several rechargeable batteries is shown in Fig. 3.

There are a few rechargeable batteries with molten salt or ceramic electrolytes that operate at elevated temperatures, but these have found only specialty (demonstration) applications such as stationary energy storage and utility vehicle propulsion and are not yet in general use. Examples of these are

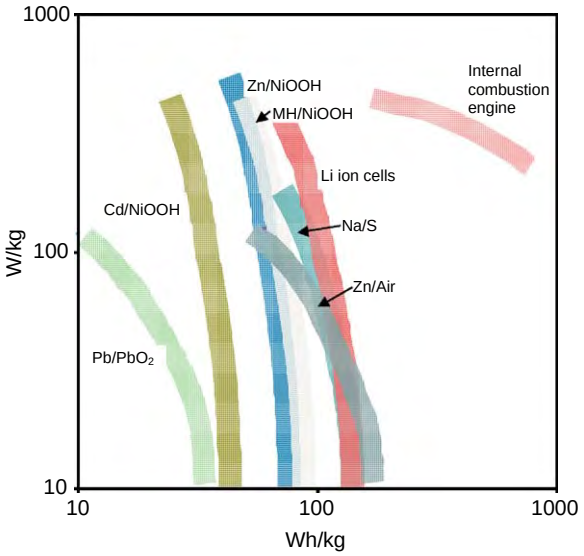


FIGURE 3 Specific power, watts/kilogram (W/kg), as a function of specific energy, watt-hours/kilogram (Wh/kg), for various cells. A curve for an internal combustion engine is included for reference.

TABLE V

Nonaqueous Rechargeable Battery Status

System	Cell voltage (V)	Equivalent weight (kg)	Theoretical specific energy (Wh/kg)	Specific energy (Wh/kg)	Specific power (W/kg)	Cost (dollars/kWh)
Li/CoO_2	3.60	0.1693	570	125	>200	>>100
Li/Mn_2O_4	4.00	0.1808	593	150	200	~200
$Li/FePO_4$	3.5	0.151	621	120	100(est.)	~200
Li/V_6O_{13}	2.40	0.0723	890	150	200	>200
Li/TiS_2	2.15	0.1201	480	125	65	>200
Li/SRS	3.00	0.080	1000	200(est.)	400(est.)	N.A.
Li/S	2.10	0.0216	2600	300	200	<100

Note. est., estimate; N.A., not available.

TABLE VI

High-Temperature Battery Status

System	Cell voltage (V)	Theoretical specific energy (Wh/kg)	Specific energy (Wh/kg)	Specific power (W/kg)	Cycle life	Cost (dollars/kWh)	Temperature (°C)
Na/S	2.08	758	180	160	500–2000	>100	350
$Na/NiCl_2$	2.58	787	90	90–155	600	>300	350
$LiAl/FeS_2$	1.80	514	180–200	>200	1000	N.A.	400

Note. N.A., not available.

the lithium/iron sulfide cell with a molten alkali halide electrolyte, the sodium/sulfur cell with a beta-alumina ceramic ($Na_2O \cdot Al_2O_3$) electrolyte, and the sodium/nickel chloride cell with both molten salt and beta-alumina electrolytes. These cells operate at 350 to 400°C. The characteristics of these cells are shown in Table VI.

4. APPLICATIONS

Current applications of batteries are very wide-ranging. It is difficult to identify many technologies that do not rely on either primary or secondary batteries. Batteries are part of our everyday lives and are essential to many fields with which most people do not come into contact.

Consumer applications of batteries include flashlights, portable electronics (e.g., cell phones), radios, televisions, MP3 players, personal digital assistants, notebook computers, portable tools, power garden tools, burglar alarms, smoke detectors, emergency lighting, remote control devices for television, stereos, electronic games, remote-controlled toys, and other entertainment equipment. Medical applications include heart pacemakers, medical diagnostic devices (e.g., glucose monitors), pumps for administering drugs, and nerve and muscle stimulators. Automobiles depend on batteries for starting, lighting, ignition, and related functions. Golf carts, utility vehicles, and forklift trucks represent a variety of propulsion-power applications.

There are many government and military applications with stringent requirements for batteries, including communications radios and phones, portable lasers for range finding, field computers, night vision equipment, sensors, starting for all types of vehicles, power sources for fuses, timers, various missiles, aircraft starting and electrical systems, electrically powered vehicles (e.g., submarines), and off-peak energy storage for utility power.

The space program depends critically on batteries for all manner of power supplies for communication, operation of the electrical systems in shuttle craft and space stations, power for the small vehicles and sensors deployed on the moon and planets, life support systems, power for space suits, and many others.

The batteries that are used in these applications come in many sizes, shapes, and types. The sizes of individual cells range from a couple of millimeters or less to a couple of meters. Depending on the application, many cells are connected in series-parallel arrangements to provide the required voltage and current for the necessary time.

Future applications for batteries are certain to include those listed here as well as a variety that are difficult to imagine at this time. We are now seeing the appearance of electric hybrid vehicles with a battery as part of the propulsion system and an electric motor. These vehicles are much more efficient than the standard vehicles and deliver approximately 50 miles per gallon. In this application, metal hydride/nickel oxide cells and lithium cells are used. We are certain to see many more of these hybrid vehicles with relatively large batteries providing much of the propulsion power. These batteries store a few kilowatt-hours of energy. We have also seen a small number of all-battery electric vehicles on the road. They use larger batteries of up to 20 kWh storage capability. As batteries of higher specific energy and lower cost become available, more electrically powered vehicles (both battery and hybrid) will be sold.

Systems for the remote generation and storage of electrical energy depend on batteries. These include solar- and wind-powered systems for remote buildings. We can expect to see ever-smaller cells being used as power sources on microcircuit chips in computers and other electronic devices. Biomedical applications will expand with a wide variety of battery-powered sensors, stimulators, transmitters, and the like. A battery-powered artificial heart is a distinct possibility.

Future household applications include a battery-powered robot capable of performing routine household tasks. As computer-controlled homes become more common, batteries will be used in many sensors, transmitters, and receivers, either because they are mobile or to be independent of the power grid. As fuel cells are introduced, it is likely that in most applications the fuel cell will be accompanied by a battery to store energy and meet peak power requirements, lowering the total cost of the system for a given peak power requirement.

5. ENVIRONMENTAL ISSUES

Some types of batteries contain hazardous materials that must be kept out of the environment. Examples of these are the lead in lead–acid batteries and the cadmium in cadmium/nickel oxide batteries. As batteries continue to experience growth in their use, developers and manufacturers must be cognizant of environmental issues relating to the materials used in the batteries. For example, more than 80% of the lead in a new lead–acid battery has been recycled. With the use of clean technology in the recycling process, this hazardous material can be used and reused safely.

6. FUTURE OUTLOOK

We can expect to see continuing growth in the variety of electrochemical couples available and the sizes and shapes of cells to satisfy a growing variety of applications, from microwatt-hours to megawatt-hours of energy. The world market for batteries is a multi-billion-dollar per year market and has been growing very rapidly, especially during the past decade or so. This growth will continue, especially as the developing countries install their communications systems and more extensive infrastructures for electrical power and transportation. Lower cost batteries will be in demand, as will batteries for various specialty applications. Batteries capable of fitting into oddly shaped spaces will be needed, and cells with polymer electrolytes can satisfy this unique requirement. Such batteries are being introduced now.

As batteries show an increase in specific energy in the range of 300 to 400 Wh/kg, battery-powered electric cars will become quite attractive. Batteries could power buses and other vehicles with ranges of 200 to 300 miles.

In the development of new cells, it will be important to continue to reduce the equivalent weight of the electrodes. At this time, the highest theoretical specific energy of any cell being developed is that of Li/S at 2600 Wh/kg. If a lithium/oxygen, lithium/phosphorus, or lithium/fluorine cell could be developed, this would lead to an even higher specific energy.

SEE ALSO THE FOLLOWING ARTICLES

Batteries, Transportation Applications • *Electrical Energy and Power* • *Electric Motors* • *Storage of Energy, Overview*

Further Reading

Linden, D., and Reddy, T. B. (eds.). (2002). "Handbook of Batteries," 3rd ed. McGraw–Hill, New York.

Socolow, R., and Thomas, V. (1997). The industrial ecology of lead and electric vehicles. *J. Industr. Ecol.* **1**, 13.

Vincent, C. A., Bonino, F., Lazzari, M., and Scrosati, B. (1984). "Modern Batteries: An Introduction to Electrochemical Power Sources." Edward Arnold, London.

Batteries, Transportation Applications

MICHAEL M. THACKERAY
Argonne National Laboratory
Argonne, Illinois, United States

1. **Battery Systems for Electric Transportation**
2. **Performance Requirements**
3. **Lead–Acid Batteries**
4. **Nickel (Alkaline) Batteries**
5. **High-Temperature Batteries**
6. **Lithium Batteries**
7. **Regulations for Transporting Batteries**

Glossary

battery A direct-current voltage source consisting of two or more electrochemical cells connected in series and/or parallel that converts chemical energy into electrical energy.
capacity The current output capability of an electrode, cell, or battery over a period of time, usually expressed in ampere-hours.
conductivity A measure of the ability of a material to conduct an electrical current, equal to the reciprocal of resistivity, $S = 1/R$.
current The rate of transfer of electrons, positive ions or negative ions, measured in amperes.
electric vehicle A vehicle that is powered solely by an electrochemical power source, such as a battery or fuel cell.
electrochemical cell A device that converts chemical energy into electrical energy by passing current between a negative electrode and a positive electrode through an ionically conducting electrolyte.
electrode A conductor by which an electrical current enters or leaves an electrochemical cell.
electrolyte Any liquid or solid substance that conducts an electric current by the movement of ions.
energy density The amount of energy per unit volume, usually expressed in watt-hours per liter.
fuel cell An electrochemical cell in which electrical energy is produced by the electrochemical oxidation of a fuel, usually hydrogen.
hybrid electric vehicle A vehicle that is powered by a combination of power sources such as a battery with an internal combustion engine or a fuel cell.
impedance The effective resistance to the flow of electric current.
power The rate at which electrical energy can be supplied, measured by the mathematical product of voltage and current, usually expressed in watts.
power density The amount of power per unit volume, usually expressed in watts per liter.
specific capacity The amount of capacity per unit mass, usually expressed in ampere-hours per kilogram.
specific conductivity The conductivity of a material per unit length, usually expressed in siemens per centimeter.
specific energy The amount of energy per unit mass, usually expressed in watt-hours per kilogram.
specific power The amount of power per unit mass, usually expressed in watts per kilogram.
voltage The potential difference between two electrodes, measured in volts.

Batteries provide a means for storing energy and as such they have become an indispensable entity in modern-day life. They are used for powering a myriad of devices, both large and small. The use of batteries as a source of energy for electrically powered vehicles was recognized early. Electric cars powered by nickel–iron batteries were manufactured during the early 20th century, but this mode of transportation could not withstand or compete against the rapid and very successful development of the internal combustion engine. Nevertheless, electric transportation continued to exist for specialized applications, such as mining and underground vehicles. Although the nickel–iron battery was extremely rugged, long-lived, and durable and played a significant role from its introduction in 1908 through the 1970s, it lost market share to the industrial lead–acid batteries. Lead–acid batteries have continued to dominate these

 © 2004 Elsevier Inc. All rights reserved.

specialized transportation sectors, which now include applications such as forklift trucks, delivery vehicles, and golf carts.

The oil crisis in the early 1970s, when the world experienced a temporary shortage of petroleum fuels, sparked a renewed interest in finding alternative sources of power for mass transport. At the same time, there was a growing awareness that electric vehicles (EVs) with "zero gas" emissions and hybrid electric vehicles (HEVs) with "low gas" emissions would significantly reduce the levels of pollution in urban areas. Since then, several new battery technologies with significantly superior performance properties have emerged—for example, nickel–metal hydride batteries, high-temperature molten-salt systems with sodium or lithium electrodes, and, most recently, nonaqueous lithium ion and lithium–polymer systems. These developments in battery technology have been driven by the need for improved high-energy, high-power systems and by environmental, safety, and cost issues.

1. BATTERY SYSTEMS FOR ELECTRIC TRANSPORTATION

1.1 Historical Background

Batteries have been known for two centuries. The first significant discovery has been credited to Volta who, between 1798 and 1800, demonstrated that a "voltaic pile" (an electrochemical cell) in which two unlike metal electrodes (silver and zinc), separated by a liquid electrolyte, provided an electrochemical potential (the cell voltage). At the same time, he demonstrated that when the area of the electrodes was increased, the cell could deliver a significantly higher current. In so doing, he unequivocally proved that there was a strong relationship between chemical and electrical phenomena. Scientists were quick to capitalize on this discovery, with the result that during the 19th century great progress was made in advancing electrochemical science. Many notable names appear in the chronology of battery development during this period. For example, Humphrey Davy (1807) showed that a battery could be used to electrolyze caustic soda (NaOH); Michael Faraday (1830) developed the basic laws of electrochemistry; the copper–zinc Daniell cell was discovered in 1836; William Grove (1839) demonstrated the principle of a fuel cell; Planté (1859) discovered the lead–acid battery; LeClanché (1868) introduced the zinc–carbon "dry" cell; and, by the turn of the century, Junger (1899) had patented the nickel–cadmium cell and Thomas Edison (1901) had patented other nickel-based electrochemical couples, notably nickel–iron. These pioneering discoveries of battery systems that used aqueous electrolytes laid the foundation for many of the battery systems in use today and opened the door to the discovery and development of alternative battery chemistries, particularly the nonaqueous sodium- and lithium-based systems that revolutionized battery technology toward the end of the 20th century.

1.2 Electrochemical Reactions and Terminology

Batteries consist of an array of electrochemical cells, connected either in series to increase the battery voltage or in parallel to increase the capacity and current capability of the battery. The individual cells consist of a negative electrode (the anode) and a positive electrode (the cathode) separated by an ionically conducting electrolyte. During discharge, an oxidation reaction occurs at the anode and ionic charge is transported through the electrolyte to the cathode, where a reduction reaction occurs; a concomitant transfer of electrons takes place from the anode to the cathode through an external circuit to ensure that charge neutrality is maintained. For secondary (rechargeable) batteries, these processes are reversed during charge.

The voltage of an electrochemical cell (E^0_{cell}) is determined by the difference in electrochemical potential of the negative and positive electrodes, i.e.,

$$E^0_{cell} = E^0_{positive\ electrode} - E^0_{negative\ electrode}.$$

From a thermodynamic standpoint, the cell voltage can also be calculated from the free energy of the electrochemical reaction (ΔG):

$$\Delta G = nFE^0_{cell},$$

where n is the number of electrons transferred during the reaction and F is Faraday's constant (96,487 coulombs). During the discharge of electrochemical cells, it is a general requirement that if maximum energy and power are to be derived from the cell, particularly for heavy-duty applications such as electric vehicles, then the voltage of the cell should remain as high and as constant as possible. This requirement is never fully realized in practice because the internal resistance (impedance) of cells and polarization effects at the electrodes lower the practical voltage and the rate at which the electrochemical reactions can take place. These limitations

are controlled to a large extent not only by the cell design but also by the type of electrochemical reaction that occurs at the individual electrodes and by the ionic conductivity of the electrolyte. For example, the following electrochemical reactions take place at the individual electrodes of a sodium–nickel chloride cell:

$$\text{Anode}: \ 2Na \rightarrow 2Na^+ + 2e^-$$

$$\text{Cathode}: \ NiCl_2 + 2e^- \rightarrow Ni + 2Cl^-$$

$$\text{Overall reaction}: 2Na + NiCl_2 \rightarrow 2NaCl + Ni.$$

At the anode, the sodium provides a constant potential throughout the reaction. The cathode also operates as a constant potential electrode because it contains three components (*C*: Na, Ni, and Cl) and three phases (*P*: $NiCl_2$, NaCl, and Ni). The Gibbs phase rule,

$$F = C - P + 2,$$

dictates that the number of degrees of freedom (*F*) for such an electrode reaction is two (temperature and pressure), thereby keeping the potential invariant. Therefore, sodium–nickel chloride cells provide a constant potential (under thermodynamic equilibrium) throughout charge and discharge.

In contrast, lithium ion cells operate by insertion reactions, whereby lithium ions are electrochemically introduced into, and removed from, host electrode structures during discharge and charge. For a Li_xC_6/$Li_{1-x}CoO_2$ cell with $x_{max} \approx 0.5$, the electrode reactions are

$$\text{Anode}: \ Li_xC_6 \rightarrow xLi^+ + C_6 + xe^-$$

$$\text{Cathode}: \ xLi^+ + Li_{1-x}CoO_2 + xe^- \rightarrow LiCoO_2$$

$$\text{Overall reaction}: Li_xC_6 + Li_{1-x}CoO_2 \rightarrow LiCoO_2 + C_6.$$

The lithiated graphite electrode operates by a series of staging reactions, each of which constitutes a two-component, two-phase electrode that provides a constant potential a few tens of millivolts above that of metallic lithium. On the other hand, the $Li_{1-x}CoO_2$ electrode is a two-component (Li and CoO_2) and single-phase ($Li_{1-x}CoO_2$) system. Therefore, according to the Gibbs phase rule, the number of degrees of freedom for the cathode reaction is three, and the potential of the $Li_{1-x}CoO_2$ electrode varies with composition (x) as it is discharged and charged. Lithium ion cells therefore do not operate at constant potential; they show a sloping voltage profile between the typical upper and lower limits of operation—4.2 and 3.0 V, respectively.

The calendar (shelf) life and cycle life of cells, which depend on the chemical stability of the charged and discharged electrodes in the electrolyte environment, are also important characteristics that have to be taken into account when considering battery systems for electric vehicles. For transportation applications, it is necessary for the battery to be able to store a large quantity of energy per unit mass and also to deliver this energy at rapid rates, particularly during the acceleration of the vehicles. The specific energy of the battery in watt-hours per kilogram (Wh/kg), which is proportional to the distance that a vehicle will be able to travel, is defined by the product of the battery voltage, V, and the electrochemical capacity of the electrodes, the unit of which is ampere-hour per kilogram (Ah/kg). The volumetric energy density is given in terms of watt-hours per liter (Wh/liter), which is an important parameter when the available space for the battery is a prime consideration. The specific power, which defines the rate at which the battery can be discharged, is given as watts per kilogram (W/kg) and the power density as watts per liter (W/liter). The power parameters are largely defined by the kinetics of the electrochemical reaction, the surface area and thickness of the electrodes, the internal resistance of the individual cells, and the size and design of the cells. Therefore, it is clear that in order to be able to travel large distances and for vehicles to be able to quickly accelerate, the batteries should be as powerful and light as possible.

Tables I and II provide a list of the battery systems that have received the most attention during the past 25 years and hold the best promise for transportation applications. Table I defines the chemistry of each system and provides theoretical voltages, capacities, and specific energies based on the masses of the electrochemically active components of the cell reactions. Table II provides typical performance characteristics that have been reported specifically for EV or HEV applications. The order in which the systems are given approximates the chronology of their development.

1.3 Electric Vehicles vs Hybrid Electric Vehicles

Pure EVs are limited by the range they can travel before the batteries become depleted and need to be recharged. Vehicle range is determined by the specific energy of the batteries. Table I shows that the practical specific energy of heavy-duty batteries varies widely from ~30 Wh/kg for lead–acid to ~150 Wh/kg for advanced lithium batteries. Although lead–acid batteries are relatively inexpensive and have a rugged construction, the range they

TABLE I

Battery Systems and Theoretical Electrochemical Properties[a]

System	Negative electrode (anode)	Positive electrode (cathode)	Open circuit voltage of reaction (V)	Theoretical capacity (Ah/kg)	Theoretical specific energy (Wh/kg)
Lead–acid	Pb	PbO_2	2.1	83	171
Nickel–cadmium	Cd	NiOOH	1.35	162	219
Nickel–metal hydride	Metal hydride alloy	NiOOH	1.35	~178[b]	~240[b]
Sodium–sulfur	Na	S	2.1–1.78(2.0)[c]	377	754
Sodium–metal chloride	Na	$NiCl_2$	2.58	305	787
Lithium ion	Li_xC_6	$Li_{1-x}CoO_2$	4.2–3.0(4.0)[c]	79[d]	316[d]
Lithium–polymer	Li	VO_x	~3.0–2.0(2.6)[c]	~340	~884

[a] Theoretical values of capacity and specific energy are based on the masses of electrode and electrolyte components that participate in the electrochemical reaction. For example, for lead–acid cells the masses of Pb, PbO_2, and H_2SO_4 are included in the calculation.
[b] Values are dependent on the composition of the selected metal hydride alloy.
[c] Average voltages for the reaction in parentheses.
[d] Calculations are based on $x_{max} = 0.5$.

TABLE II

Typical Performance Characteristics of Various Battery Systems[a]

System	Specific energy (Wh/kg)	Energy density (Wh/liter)	Specific power (W/kg)	Power density (W/liter)	No. of cycles
Lead–acid					
EV	35–50	90–125	150–400	400–1000	500–1000
HEV	27	74	350	900	?
Nickel–cadmium					
EV	60	115	225	400	2500
Nickel–metal hydride					
EV	68	165	150 (80% DOD)	430 (80% DOD)	600–1200
HEV	50	125	750 (50% DOD)	>1600 (50% DOD)	?
Sodium–sulfur					
EV	117	171	240	351	800
Sodium–metal chloride					
EV	94	147	170	266	>1000
Lithium–ion[b]					
EV battery module	93	114	350	429	800 (40°C)
Lithium–polymer					
EV battery module	155	220	315	447	600

[a] Abbreviations used: EV, electric vehicle; HEV, hybrid electric vehicle; DOD, depth of discharge of the battery.
[b] $C/LiMn_2O_4$ system.

offer (on average, 50 miles between charges) limits their application to utility fleets, such as forklift trucks, delivery vehicles, and golf carts, for which short driving distances are acceptable. (Because lead–acid batteries are heavy, much of the energy is used to transport the mass of the battery.) For more energetic and advanced systems, such as nickel–metal hydride (68 Wh/kg), sodium–nickel chloride (94 Wh/kg),

lithium ion (93 Wh/kg), and lithium–polymer batteries (155 Wh/kg), greater attention has to be paid to factors such as safety and cost. Although high-energy density battery systems that provide 100 Wh/kg can offer a typical range of 100 miles or more, this is still too short a distance for pure EVs to be considered attractive, particularly in the United States, where driving distances can be considerably longer. In the absence of built-in infrastructures to cater to the demands of recharging EVs and the time that it may take to charge them, it seems that this mode of consumer transportation will be initially restricted to urban travel and introduced particularly in cities that are severely overpolluted by gas emissions from internal combustion vehicles.

Currently, HEVs, in which a battery and an internal combustion engine (or a fuel cell) are coupled, are the most attractive option for reducing the demands on fossil fuels and for optimizing gasoline consumption. Such systems still make use of the current infrastructure of the gasoline industry and do not require independent battery charging stations. In HEVs, the battery is used predominantly to assist in acceleration; the gasoline engine takes over during cruising, when the battery can be recharged. Energy from regenerative braking can also be used to charge the batteries. Unlike pure EVs in which the batteries may become completely discharged, which can negatively impact cycle life, batteries for HEVs are subject to only shallow charge and discharge reactions during operation, thereby enabling a very high number of cycles and significantly longer calendar lifetimes.

2. PERFORMANCE REQUIREMENTS

The growing concern about pollution led the California Air Resources Board in the early 1990s to mandate that 2% of all vehicles sold in the state of California should be zero-emission vehicles by 1998 and 10% by 2003. Although these mandates have not been successfully implemented, they led to the creation of the U.S. Advanced Battery Consortium (USABC), which consists of U.S. automobile manufacturers, electric utility companies, and government authorities, to support and manage the development of batteries for pure EVs. The emphasis of the development efforts has been redirected toward developing high-power batteries for HEVs through the Program for a New Generation of Vehicles, now called the FreedomCAR Partnership. Selected guidelines and performance requirements established by the USABC in 1996 for mid- and long-term EV batteries and by Freedom CAR in 2001 for HEV batteries are provided in Tables III and IV, respectively. HEV batteries are classified into two types: those for "power-assist" operation, where the battery is always used in conjunction with the internal combustion engine, and those for "dual-mode" operation, where the battery can provide energy and power to the vehicle either on its own or together with the engine. The energy requirement for dual mode (15 Wh/kg) is considerably higher than that for power assist (7.5 Wh/kg) so that the battery can provide sufficient energy and power to the vehicle on occasions when it is used alone.

TABLE III

Selected USABC Mid-Term and Long-Term Battery Performance Goals for Electric Vehicles (1996)

Performance criteria	Mid-term (5 years)	Long term (10 years)
Specific energy (Wh/kg)	80	200
Energy density (Wh/liter)	135	300
Specific power (W/kg)	150	400
Power density (W/liter)	250	600
Life (years)	5	10
Cycle life (cycles, 80% depth of discharge)	600	1000
Operating temperature (°C) (environment)	−30 to 65	−40 to 85
Recharge time (h)	<6	3–6

3. LEAD–ACID BATTERIES

Lead–acid batteries have existed as commercial products for more than 100 years and have found application in a large number of diverse systems. The most notable application has been in the automotive industry for starting internal combustion engines and for providing the energy to power electrical equipment in vehicles, such as lights, windows, and displays. These batteries are commonly called SLI batteries (starting, lighting, and ignition) and are capable of providing the high currents required for initially cranking the engine. Other applications include photovoltaic energy storage, standby energy storage (particularly for providing backup emergency power and uninterrupted power supply for strategic equipment, such as lighting, telephone, and communication systems), and transportation devices such as underground mining conveyances, submarines (particularly for silent maneuverability when submerged),

TABLE IV

Selected FreedomCAR Battery Performance Goals for Hybrid Electric Vehicles (2002)[a]

Performance criteria	Power assist	Dual mode
Total available energy over DOD range in which power goals are met (Wh/kg)	7.5 (at C/1 rate)[b]	15 (at 6-kW constant power)
Pulse discharge power (W/kg)	625 (18 s)	450 (12 s)
Calendar life (years)	15	15
Cycle life (cycles, for specified SOC increments)	300,000	3750
Operating temperature (°C)	−30 to 50	−30 to 50

[a] Abbreviations used: DOD, depth of discharge of the battery; SOC, state of charge.
[b] C/1 rate, 1-h rate.

golf carts, milk and postal delivery trucks, and small electrically powered cars and trucks. Lead–acid batteries constitute approximately 40% of the world's total battery sales, which can be attributed to their well-developed and robust technology and significant cost advantage.

3.1 General Chemistry, Construction, and Operation

Lead–acid batteries consist of a metallic lead (Pb) negative electrode, a lead dioxide (PbO_2) positive electrode, and a sulfuric acid electrolyte. The overall cell reaction is

$$Pb + PbO_2 + 2H_2SO_4 \rightarrow 2PbSO_4 + 2H_2O.$$

The voltage of lead–acid cells on open circuit is approximately 2 V; a standard 12-V (SLI) battery therefore consists of six individual cells connected in series. During discharge, lead sulfate ($PbSO_4$) is produced as both Pb and PbO_2 electrodes, with water (H_2O) as a by-product of the reaction. Because the specific gravity (SG) or density of H_2SO_4 differs markedly from that of H_2O, the SG of the electrolyte can be conveniently used to monitor the state of charge of the cells and battery and to identify "dead" cells. The SG of lead–acid cells depends on the type of battery construction; fully charged cells typically have an SG of approximately 1.3, whereas the SG of fully discharged cells is typically 1.2 or lower.

Lead–acid batteries are assembled in the discharged state from electrodes that are manufactured by reacting PbO, Pb, and sulfuric acid to form "tribasic" ($3PbO{\cdot}PbSO_4$) and "tetrabasic" ($4PbO{\cdot}PbSO_4$) salts, which are either pasted onto flat lead grids or compacted into porous, tubular electrode containers with a central current-collecting rod, and then they are cured under controlled conditions of humidity and temperature. A formation (initial charge) process is required to convert the salts to Pb and PbO_2 at the negative and positive electrodes, respectively. Red lead (Pb_3O_4), which is more conductive than PbO, is sometimes added to the positive electrode to promote the formation of PbO_2.

Because Pb is a soft metal (melting point, 327°C), small amounts of additives, such as antimony, calcium, and selenium, have been used to increase the stress resistance of the lead grids. Antimony, in particular, significantly increases the mechanical strength of the grids but also increases their electrical resistance. An additional disadvantage of antimony is that toxic stibene (SbH_3) can form during charge, which precludes the use of lead–acid batteries with lead–antimony electrodes in poorly ventilated areas, such as in underground mining operations and submarines.

Charging of lead–acid batteries can also be hazardous. The voltage of lead–acid batteries (2 V) is higher than that required for the electrolysis of water, with the result that hydrogen and oxygen are released while batteries are being charged. Improvements have been made with the introduction of valve-regulated lead–acid ("maintenance-free") batteries in which evolved oxygen is recombined with lead at the negative electrode and in which evolved hydrogen, which can be minimized by the addition of tin to the lead grids, is vented. Nevertheless, hydrogen evolution can still pose a safety hazard in some lead–acid battery designs. Another unavoidable and inherent safety hazard of lead–acid batteries is the potential risk of spillage of the sulfuric acid electrolyte, which is corrosive and can cause severe chemical burns.

3.2 Transportation Applications

Despite their relatively low energy density, lead–acid batteries have found widespread application in numerous transportation devices, as previously noted, where the range of the vehicle is not a prime

consideration. Disadvantages that will ultimately count against lead–acid batteries for transportation applications in the face of alternative emerging technologies are their relatively low specific energy and decay during long-term storage if kept in a discharged state. Nonetheless, great strides have been made by General Motors during the past few years to develop a compact, limited-production electric vehicle, known as EV1, which is now in its second generation. The estimated driving range is 55–95 miles, depending on driving conditions, which is made achievable through an aerodynamic, lightweight design and regenerative braking of the battery. The greatest advantage that lead–acid batteries have over competitive technologies is cost; they are as much as 10 times less expensive than advanced nickel–metal hydride, lithium ion, and lithium–polymer batteries that are being developed for high-performance vehicles.

4. NICKEL (ALKALINE) BATTERIES

Several nickel-based batteries with a NiOOH positive electrode and an alkaline (KOH) electrolyte were developed during the 20th century, including the nickel–iron (Edison), nickel–zinc, nickel–cadmium, and nickel–metal hydride systems. Although both nickel–iron and nickel–zinc systems commanded significant attention as possible candidates for electric vehicles, they fell from favor because of insurmountable problems. For example, despite their ability to provide a high number of deep discharge cycles, nickel–iron batteries were limited by low power density, poor low-temperature performance, and poor charge retention and gas evolution on stand. A major limitation that has slowed the development of nickel–zinc batteries is the solubility of the zinc electrode in the alkaline electrolyte. The severe shape change of the zinc electrode that occurs during cycling, and the formation of dendrites, which can cause internal short circuits, are problems that have not been entirely overcome. Moreover, corrosion of the zinc electrode generates hydrogen gas. Because of these limitations and the lack of major interest in these technologies today, the nickel–iron and nickel–zinc systems are not discussed further.

4.1 Nickel–Metal Hydride and Nickel–Cadmium Batteries

Nickel–metal hydride (NiMH) cells were developed during the last quarter of the 20th century as an environmentally acceptable alternative to nickel–cadmium (NiCad) cells; the toxic cadmium (negative) electrode was replaced with a metal hydride electrode. During this period, the rapid growth of NiMH technology was fueled by the need for advanced batteries for early consumer electronics devices, such as cellular phones and laptop computers.

Both NiMH and NiCad systems use an alkaline (KOH) electrolyte and a NiOOH positive electrode, and they both have a cell voltage of 1.35 V. NiMH cells can therefore be substituted directly for NiCad cells in many applications. The electrochemical reaction that occurs in a NiCad cell is

$$Cd + 2NiOOH + 2H_2O \rightarrow Cd(OH)_2 + 2Ni(OH)_2.$$

Water is consumed during the reaction to form $Cd(OH)_2$ and $Ni(OH)_2$ at the negative and positive electrodes, respectively; however, in practice, the concentration of the alkaline electrolyte is such that it does not change much during the reaction. The insolubility of the electrodes in the electrolyte at all states of charge and discharge contributes to the long cycle and standby life of NiCad cells, even though the cells lose capacity on storage and require charging to be restored to their full capacity. Particular advantages of NiCad cells compared to NiMH cells are lower cost, higher rate capability, and better low-temperature operation. However, the toxicity of cadmium and the lower specific energy and energy density delivered by practical NiCad batteries compared to NiMH batteries (Table II) are major factors preventing further growth of the NiCad battery market.

For NiMH cells, the overall electrochemical reaction is

$$MH + NiOOH \rightarrow M + Ni(OH)_2.$$

During discharge, hydrogen is released from the MH electrode into the electrolyte, which is inserted into the NiOOH electrode in the form of protons (H^+ ions) with the concomitant reduction of the nickel ions from a trivalent to divalent state to form $Ni(OH)_2$; water is therefore not consumed during the overall reaction. Because the electrolyte is water based, an inherent problem of MiNH cells is that electrolysis of water can occur during charge to form hydrogen and oxygen at the negative and positive electrodes, respectively. In sealed NiMH cells, however, the inherent dangers of overcharge are reduced because the oxygen evolved at the positive electrode can diffuse through the electrolyte and permeable separator to the MH electrode, where it can react to form water and thus reduce a pressure

increase in the cell:

$$2MH + 1/2O_2 \rightarrow 2M + H_2O.$$

An additional advantage of this reaction, which depletes the negative electrode of hydrogen, is that the negative electrode never becomes fully charged, thereby suppressing the evolution of hydrogen gas. Nickel–metal hydride cells are therefore constructed with a "starved-electrolyte" design, which promotes and exploits the advantages of the oxygen-recombination reaction.

The MH negative electrodes of NiMH batteries have complex formulae; they are based on the overall composition and structure types of AB_5 or AB_2 alloys. AB_5 alloys contain several rare earth metals (Misch metals) and tend to be preferred over the AB_2-type alloys even though they have slightly lower specific and volumetric capacities. The NiOOH positive electrodes commonly contain minor quantities of Co and Zn to improve the electronic conductivity of the NiOOH electrode (Co) and to control oxygen evolution and microstructural features of the electrode (Co and Zn).

NiMH batteries have made a significant contribution to improving the range of pure EVs and to the fuel consumption of HEVs. Replacement of lead–acid batteries with NiMH batteries in the General Motors electric vehicle EV1 increased the driving range from 55–95 to 75–130 miles. NiMH batteries were introduced into the first generation of modern-day HEVs by Toyota (the Prius) and Honda (the Insight) in the late 1990s.

Despite the very significant progress that has been made, the future of NiMH batteries is uncertain, particularly in the consumer electronics industry, in which they have lost significant market share to lithium–ion batteries, which offer a significantly higher cell voltage and superior specific energy (Tables I and II).

5. HIGH-TEMPERATURE BATTERIES

The possibility of using high-temperature batteries with molten electrodes or molten salts was first seriously explored in the early 1960s. Early attempts to construct an electrochemical lithium–sulfur cell with a LiCl–KCl molten salt electrolyte (melting point, 352°C) were unsuccessful because both electrodes were also molten at the operating temperature of the cell (400°C). In order to overcome the problems of electrode containment, researchers replaced the two liquid electrodes with a solid lithium alloy and a metal sulfide. This approach resulted in the development of a highly successful primary LiAl/FeS_x battery ($1 < x < 2$), which has been used predominantly in military devices such as missile systems. Attempts to develop a viable rechargeable system for EVs were thwarted by several problems, such as cycle life, corrosion, and cost, with the result that the USABC abandoned its support for the project in 1995 and opted for the emerging lithium battery technologies. Nevertheless, research on lithium–iron sulfide batteries continues, but for operation at lower temperature (90–135°C) using a metallic lithium negative electrode with a polymer electrolyte. Time will tell whether this new approach will be successful.

Two other high-temperature batteries were developed in the latter part of the 20th century, namely, the sodium–sulfur battery and the sodium–nickel chloride battery, both of which use molten sodium as the positive electrode. These advanced sodium batteries were made possible by the discovery in 1967 of the sodium ion-conducting solid electrolyte, β-alumina. This solid electrolyte, which has a composition between $Na_2O \cdot 5.33Al_2O_3$ and $Na_2O \cdot 11Al_2O_3$, has an anomalously high sodium ion specific conductivity of 0.2 S/cm at 300°C, comparable to that of liquid electrolytes (e.g., the specific conductivity of 3.75 M H_2SO_4 at room temperature is 0.8 S/cm). The structure of β-alumina consists primarily of Al_2O_3 spinel blocks, which are structurally related to the mineral "spinel" $MgAl_2O_4$. These spinel blocks are separated by oxygen-deficient layers, in which sodium ions can diffuse rapidly when an electrochemical potential difference exists across the ceramic electrolyte. The conventional design of the β-alumina solid electrolyte for sodium batteries is tubular. The β-alumina tubes are brittle and must be thin to permit the electrochemical reaction to occur at an acceptable rate. The manufacture of high-quality tubes, which requires a sintering temperature of 1600°C to ensure a high-density ceramic, is a complex process and great care must be taken to minimize flaws that might lead to fracture of the tubes during battery operation.

5.1 Sodium–Sulfur Batteries

In most conventional sodium–sulfur cells, which have the configuration Na/β-alumina/S, the molten sodium electrode is contained within the β-alumina tube, typically called the central sodium design. The molten sulfur electrode is contained on the outside between the tube and the inner walls of the metal container,

which are usually coated with chromium to resist attack from the highly corrosive sulfur electrode. Cells can also be constructed with a central sulfur design, but this design provides lower power and is less tolerant to freeze–thaw cycling (i.e., cooling and heating of the battery during which the electrodes are allowed to solidify and melt repeatedly).

During discharge of Na/β-alumina/S cells, the sodium reacts electrochemically with sulfur to form a series of polysulfide compounds according to the general reaction

$$2Na + xS \rightarrow Na_2S_x$$

over the range $3 \leq x \leq 5$. The reaction can be considered to take place in two steps. The initial reaction produces Na_2S_5, which is immiscible with the molten sulfur. This reaction is therefore a two-phase process; it yields a constant cell voltage of 2.08 V. Thereafter, the sulfur electrode discharges in a single-phase process to form Na_2S_x compositions between Na_2S_5 and Na_2S_3, during which the cell voltage decreases linearly. The open circuit voltage at the end of discharge is 1.78 V. The composition Na_2S_3 represents the fully discharged composition because further discharge yields Na_2S_2, which is a solid, insoluble product that leads to an increase in the internal resistance of the cell. Moreover, it is difficult to recharge a Na_2S_2 electrode. In practice, therefore, it is common for the end voltage of sodium–sulfur cells to be restricted to between 1.9 and 1.78 V not only because it prevents over-discharge in local areas of the cells but also because it reduces the severity of the corrosion reactions that occur for Na_2S_x products with lower values of x.

Because sodium and sulfur are relatively light elements, the sodium–sulfur battery offers a high theoretical specific energy (754 mAh/g). However, the robust battery construction, heating systems, and thermal insulation that are required for efficient and safe operation place a penalty on the specific energy that can be delivered in practice (117 Wh/kg). The molten state of the electrodes and high-temperature operation ensure an acceptable rate capability; a specific power of 240 W/kg for electric vehicle batteries has been reported (Table I).

One of the major disadvantages of sodium–sulfur cells is that they are inherently unsafe; any fracture of the β-alumina tube that allows the intimate contact of molten sodium and molten sulfur can result in a violent reaction that is difficult to control. Furthermore, damaged cells that fail in an open-circuit mode need to be electrically bypassed to ensure continued operation of the battery. These limitations, the need for hermetic seals in a very corrosive environment, and the difficulty of meeting cost goals were among the major reasons why support for the development of sodium–sulfur batteries for electric vehicle applications was terminated worldwide in the mid-1990s. Nevertheless, considerable development of this technology is under way, particularly in Japan, for stationary energy storage applications.

5.2 Sodium–Nickel Chloride Batteries

The sodium–nickel chloride battery, commonly referred to as the Zebra battery, had its origins in South Africa in the late 1970s. The early research on these batteries capitalized on the chemistries and designs of two high-temperature battery systems, sodium–sulfur and lithium–iron sulfide, that had been made public several years before. The first generation of Zebra cells used an iron dichloride electrode with the configuration,

$$Na/\beta\text{-alumina}/FeCl_2,$$

in which the sodium electrode and β-alumina electrolyte components mimic the sodium–sulfur cell and the iron dichloride electrode acts as a substitute for the FeS_x electrode in the LiAl/KCl, LiCl/FeS_x cell. Because $FeCl_2$ is a solid, a molten salt electrolyte must be added to the positive electrode compartment of Na/β-alumina/$FeCl_2$ cells. $NaAlCl_4$ is an ideal candidate for this purpose because it has a low melting point (152°C) and because it dissociates into two stable ionic species, Na^+ and $AlCl_4^-$, thereby providing good Na^+ ion conductivity at elevated temperatures. Moreover, it is probable that the chloride ions of the molten electrolyte play a significant role in the dechlorination and rechlorination reactions that occur at the iron electrode during discharge and charge.

The electrochemical reaction of a Na/β-alumina/$FeCl_2$ cell occurs in two distinct steps:

$$6Na + 4FeCl_2 \rightarrow Na_6FeCl_8 + 3Fe$$
$$2Na + Na_6FeCl_8 \rightarrow NaCl + 4Fe.$$

The overall reaction is

$$8Na + 4FeCl_2 \rightarrow NaCl + 4Fe.$$

The initial reaction provides an open-circuit voltage of 2.35 V; the second reaction occurs only a few tens of millivolts lower. The intermediate phase that is formed, Na_6FeCl_8, has a defect rock salt structure in which one iron divalent atom replaces two Na atoms in the NaCl structure. Moreover, despite a volume increase, the closely packed arrangement of the

chloride atoms of the $FeCl_2$ structure remains essentially unaltered during the transitions to Na_6FeCl_8 and NaCl. Crystallographically, the reaction at the positive electrode is therefore a relatively simple one. Moreover, all the compounds formed during charge and discharge are essentially insoluble in the $NaAlCl_4$ electrolyte, provided that the electrolyte is kept basic [i.e., slightly NaCl rich (or $AlCl_3$ deficient)]. This prevents ion-exchange reactions from occurring between the Fe^{2+} in the electrode and Na^+ in the β-alumina structure, which would severely degrade the conductivity of the solid electrolyte membrane. The simplicity of the electrochemical reaction and the insolubility of the metal chloride electrode in the molten salt electrolyte are key factors that contribute to the excellent cycle life of Zebra cells.

The second generation of Zebra cells used a nickel chloride electrode because it provided a slightly higher cell voltage (2.58 V), resulting in cells with a higher power capability. Unlike $FeCl_2$ electrodes, $NiCl_2$ electrodes do not form any intermediate compositions but react directly with sodium to form NaCl:

$$2Na + NiCl_2 \rightarrow 2NaCl + Ni.$$

Zebra cells are usually designed with the positive electrode housed inside the β-alumina tube. Therefore, during charge, sodium is electrochemically extracted from the NaCl/Ni composite electrode to form $NiCl_2$; the metallic sodium is deposited on the outside of the tube and fills the void between the tube and inner wall of the cylindrical metal container. Although the power capability of the early Zebra cells was limited to approximately 100 W/kg, significantly improved performance was achieved by changing the shape of the cylindrical β-alumina tube to a fluted-tube design, thereby increasing the surface area of the solid electrolyte considerably. Fully developed Zebra batteries now have a specific energy of approximately 100 Wh/kg and a specific power of 170 W/kg.

Zebra cells have several major advantages over sodium–sulfur cells:

1. Zebra cells have a higher operating voltage (2.58 vs 2.08 V).
2. Unlike sodium–sulfur cells, which are assembled in the charged state with highly reactive sodium that requires special handling conditions, Zebra cells can be assembled in the discharged state by blending common salt with nickel powder. (Note that state-of-the-art Zebra electrodes contain a small proportion of iron, which significantly improves the power capability of the cells toward the end of discharge.)
3. From a safety standpoint, if the β-alumina tubes fracture, the molten $NaAlCl_4$ electrolyte is reduced according to the reaction

$$NaAlCl_4 \rightarrow 4NaCl + Al,$$

   resulting in solid products; the deposited aluminum metal also provides an internal short circuit that allows for failed cells to be electrically bypassed.
4. Zebra cells can be safely overdischarged by electrochemical reduction of the electrolyte at 1.58 V (the reaction is identical to that given for point 3), provided that the reaction does not go to completion when the electrode solidifies completely.
5. There is less corrosion in Zebra cells.
6. Zebra cells offer a wider operating temperature range (220–450°C) compared with sodium–sulfur cells (290–390°C).
7. Zebra cells are more tolerant to freeze–thaw cycles.

Zebra batteries have been tested extensively in EVs by Daimler-Benz in Europe, for example, in a Mercedes A Class sedan. These batteries have met all the USABC mid-term performance goals listed in Table III, except for operating temperature. Although development work on Zebra batteries is ongoing, the market for pure EVs for which Zebra batteries are best suited is lacking, particularly because of the greater attention that is currently being paid to HEVs. Also, high-temperature batteries have been criticized because they require sophisticated thermal management systems for both heating and cooling. Nonetheless, because they are temperature controlled, these systems have been shown to work exceptionally well in both extremely hot and extremely cold climates.

6. LITHIUM BATTERIES

6.1 Lithium Ion Batteries

Although primary, nonaqueous, lithium batteries that operate at room temperature have been in existence since the 1970s for powering small devices such as watches and cameras, it was the introduction of rechargeable lithium ion batteries by Sony Corporation in 1990 that brought about a revolution in the battery industry. These batteries

were introduced by Japan in anticipation of the boom in electronic devices such as cellular phones and laptop computers, which took the world by storm during the next decade and was responsible for the technology "dot-com" industry.

Lithium ion batteries derive their name from the electrochemical processes that occur during charge and discharge of the individual cells. During charge of these cells, lithium ions are transported from a positive host electrode to a negative host electrode with concomitant oxidation and reduction processes occurring at the two electrodes, respectively; the reverse process occurs during discharge.

State-of-the-art lithium ion cells consist of a graphite negative electrode and a $LiCoO_2$ positive electrode. The electrolyte consists of a lithium salt, such as $LiPF_6$, dissolved in an organic solvent, typically a blend of organic carbonates, such as ethylene carbonate and diethyl carbonate. The cell reaction is

$$Li_xC_6 + Li_{1-x}CoO_2 \rightarrow C_6 + LiCoO_2$$

Lithium ion cells, like Zebra cells, are conveniently assembled in the discharged state, which avoids the necessity of having to handle metallic lithium. The electrochemically active lithium is therefore initially contained in the $LiCoO_2$ positive electrode. Lithium extraction from $LiCoO_2$ occurs at a potential of approximately 4 V vs a metallic lithium (or lithiated graphite) electrode. This voltage is three times the voltage of a nickel–metal hydride cell. Therefore, one lithium ion cell provides a voltage equivalent of three nickel–metal hydride cells in series. Four-volt lithium ion cells are high-energy power sources. From a safety standpoint, it is fortunate that a protective passivating layer forms at the lithiated graphite–electrolyte interface, which prevents the Li_xC_6 electrode from reacting continuously and vigorously with the electrolyte. The high voltage and the light electrodes give lithium ion cells a high theoretical specific energy ($\sim$400 Wh/kg). However, despite this advantage, these high-voltage cells contain highly reactive electrodes, particularly when they are fully charged. If cells are overcharged, the possibility of exothermic reactions of a highly delithiated and oxidizing $Li_{1-x}CoO_2$ electrode and an overlithiated and reducing Li_xC_6 electrode (which may contain metallic lithium at the surface of the lithiated graphite particles) with a flammable organic electrolyte poses a severe safety hazard. In the event of thermal runaway and the buildup of gas pressure, cells can vent, rupture, and catch fire. Individual lithium ion cells must therefore be protected electronically from becoming overcharged, which contributes significantly to the cost of lithium ion cells. With built-in microprocessors, millions of small lithium ion cells, now considered safe, are produced (mainly by Japanese manufacturers) and sold every month. For larger scale applications, such as EVs and HEVs, in which considerably more energy is contained in the batteries, the safety issue is of paramount importance and continues to receive much attention.

Compared with all other systems, lithium ion batteries are the most versatile in terms of their chemistry. They can offer a diverse range of electrochemical couples because there are a large number of structures that can act as host electrodes for lithium. For example, various metals, such as Al, Si, Sn, and Sb, and intermetallic compounds, such as SnSb, Cu_6Sn_5, and Cu_2Sb, have received considerable attention because they react with lithium, either by lithium insertion or by a combination of lithium insertion into and metal displacement from a host structure at a potential above that of a Li_xC_6 electrode. These metal/intermetallic electrodes should therefore be potentially safer than Li_xC_6 electrodes. The electrochemical performance of these metal and intermetallic electrodes is limited by large crystallographic volume increases during the uptake of lithium and by a large irreversible capacity loss that commonly occurs during the initial cycle of the lithium cells.

Although $LiCoO_2$ has been the positive electrode of choice for more than a decade, lithium extraction is limited to approximately 0.5 Li per formula unit, which corresponds to a practical specific capacity of 137 mAh/g. Moreover, cobalt is a relatively expensive metal. Therefore, there has been a great motivation in the lithium battery community to find alternative positive electrodes to increase the capacity and specific energy of lithium ion cells and to reduce their cost. In particular, much effort has been focused on developing alternative lithium–metal oxide electrodes that are isostructural with $LiCoO_2$, such as $LiNi_{0.8}Co_{0.15}Al_{0.05}O_2$, which has a higher capacity (170–180 mAh/g) without compromising the cell voltage significantly. A $LiMn_{0.33}Ni_{0.33}Co_{0.33}O_2$ system has been developed that offers an electrode capacity in excess of 200 mAh/g. These advances bode well for the next generation of lithium ion batteries.

Other positive electrode materials have also commanded attention. One example is $LiMn_2O_4$, which has a spinel-type structure. This electrode is considerably less expensive than $LiCoO_2$ and is more stable to the extraction of lithium, a reaction that

also occurs at approximately 4 V vs a metallic lithium electrode. Moreover, the Mn_2O_4 spinel framework provides a stable three-dimensional interstitial space for lithium diffusion and is more tolerant to high-rate discharge and charge compared with the layered electrode structures. Although $LiMn_2O_4$ has an inferior capacity compared to that of $LiCoO_2$, the spinel remains a very attractive low-cost and high-rate electrode for HEV batteries, which are subjected to only shallow levels of charge and discharge during operation. Spinel electrodes suffer from solubility problems if the temperature of the battery reaches 40–50°C as a result of trace quantities of hydrofluoric acid that can be generated in the cell by hydrolysis of commonly used fluorinated lithium salts such as $LiPF_6$. It has been reported that this problem can be overcome by using nonfluorinated lithium salts that contain boron.

Another possible alternative positive electrode to $LiCoO_2$ is the family of lithium transition metal phosphates, which have an olivine-type structure, that is, $LiFePO_4$. This low-cost material, although it provides a relatively low theoretical electrode capacity (170 mAh/g), is significantly more stable to lithium extraction than layered and spinel transition metal oxide electrode structures. Unlike the layered $LiMO_2$ (M = Co, Ni, and Mn) and spinel LiM_2O_4 (M = Mn and Ni) electrodes that, on lithium extraction, operate off either a $M^{4+/3+}$ or a $M^{4+/2+}$ couple at 4 V or higher, the $LiFePO_4$ electrode provides a significantly more stable $Fe^{3+/2+}$ couple at an attractive potential of 3.5 V. $Li_xC_6/Li_{1-x}FePO_4$ cells therefore show very little capacity decay and provide exceptionally good cycling stability. An intrinsic limitation of phosphate electrodes is that they are inherently poor electronic and ionic conductors, two essential features for a high-performance insertion electrode. Significant progress has been made in improving the electronic conductivity of these electrodes, for example, by applying a coating of electronically conducting carbon to the surface of the phosphate particles.

Spinel and olivine electrode structures can be exploited by selecting various transition metals as the electrochemically active component to tailor the potential of the electrode and hence the voltage of the cell. For example, in the family of spinel compounds, $Li_4Ti_5O_{12}$ (alternatively, in spinel notation, $Li[Li_{0.33}Ti_{1.67}]O_4$) provides a constant 1.5 V vs Li^0 when reacted with lithium to the rock salt composition $Li_7Ti_5O_{12}$. Therefore, $Li_4Ti_5O_{12}$ can be used as a negative electrode against a high-potential positive electrode such as $LiMn_{1.5}Ni_{0.5}O_4$ (4.5 V vs Li^0) to yield a 3.0-V cell. Lithium ion cells with metal oxide negative and positive electrodes should be inherently safer compared with state-of-the-art $Li_xC_6/LiCoO_2$ cells, but they are limited by relatively low voltages and specific energy. With respect to the family of olivine electrodes, replacement of Fe in $LiFePO_4$ by Mn and Co yields lithium ion cells with significantly higher voltages of 4.1 and 4.8 V, respectively.

With respect to electrolytes, considerable progress has been made in developing "polymer-gel" systems in which a liquid electrolyte component is immobilized in a polymer matrix. Although the Li^+ ion conductivity of these gel electrolytes is slightly inferior to that of pure liquid electrolyte systems, the ability to manufacture cells with laminated electrodes and electrolytes has provided much flexibility in the design of lithium ion cells. Moreover, cells with polymer-gel electrolytes are construed to be safer than those containing liquid electrolytes; they can also be easily constructed with lightweight packaging materials such as aluminum foil, thereby significantly improving the energy density of the cells.

Lithium ion battery technology has not reached maturity, and there is little doubt that significant advances will continue to be made in the years to come with regard to improving energy, power, and safety. Although the emphasis on the immediate development of a pure EV has been reduced, the future of lithium ion batteries for HEVs and in the shorter term for other selected modes of transportation, such as electric bikes and scooters, remains bright. In this respect, the first generation of lithium ion batteries containing a manganese spinel electrode has been developed for both EV and HEV applications by Shin-Kobe in Japan.

6.2 Lithium–Solid Polymer Electrolyte Batteries

Batteries that contain a metallic lithium negative electrode coupled via solid polymer electrolyte to a transition metal oxide positive electrode (usually a vanadium oxide, for example, V_2O_5, V_6O_{13}, and LiV_3O_8, commonly denoted VO_x) have been under development since the mid-1970s. A major advantage of using lithium instead of a lithiated graphite electrode is that a metallic lithium electrode offers a substantially higher capacity (3863 mAh/g) than LiC_6 (372 mAh/g). However, the performance of

these batteries is compromised by the relatively low room-temperature ionic conductivity of polymer electrolytes, which necessitates an elevated operating temperature, typically 80–120°C. Another limitation of these batteries is that dendritic growth of lithium may occur during the charging of the battery, leading to electrolyte penetration and internal short circuits. This problem has not been entirely overcome, thereby restricting the cycle life of this type of battery. Despite this limitation, excellent progress has been made during the past few years, and development efforts are ongoing. Battery modules are being evaluated in demonstration EVs and HEVs, such as the Ford Motor Company's Th!nk *city* vehicle and the Honda Insight.

7. REGULATIONS FOR TRANSPORTING BATTERIES

International regulations exist for transporting hazardous materials by land, sea, and air. Batteries that contain highly reactive elements, such as lithium and sodium, therefore fall under the scope of these regulations. In particular, since the early 1990s, the regulations for lithium ion, lithium–polymer, sodium–sulfur, and sodium–metal chloride batteries have been continually revised to allow their transport, particularly for large and heavy-duty applications, such as EVs and HEVs. The "United Nations Recommendations on the Transport of Dangerous Goods" provides regulatory guidance for the transport of dangerous goods both internationally and domestically. Regulations stipulate that lithium cells and batteries of any size can be shipped if they pass a set of safety tests described under the section titled "Manual of Tests and Criteria." Because both sodium–sulfur and sodium–metal chloride batteries operate at elevated temperatures, the regulations also define specific conditions for the transportation of cold and hot sodium batteries.

Acknowledgments

I thank Ralph Brodd, Dennis Dees, and Gary Henriksen for useful discussions. Financial support from the U.S. Department of Energy, Offices of FreedomCAR & Vehicle Technologies, is gratefully acknowledged. This article was created under Contract No. W-31-109-ENG-38.

SEE ALSO THE FOLLOWING ARTICLES

Batteries, Overview • Electric Motors • Fuel Cells • Fuel Cell Vehicles • Fuel Cycle Analysis of Conventional and Alternative Fuel Vehicles • Hybrid Electric Vehicles • Internal Combustion Engine Vehicles

Further Reading

Besenhard, J. O. (ed.). (1999). "Handbook of Battery Materials," Wiley-VCH, Weinheim.

Linden, D., and Reddy, T. B. (eds.). (2002). "Handbook of Batteries," 3rd ed. McGraw-Hill, New York.

McNicol, B. D., and Rand, D. A. J. (1984). "Power Sources for Electric Vehicles." Elsevier, Amsterdam.

Morris, C. (ed.). (1992). "Dictionary of Science and Technology." Academic Press, London.

Nakai, K., Aiba, T., Hironaka, K., Matsumura, T., and Horiba, T. (2000). Development of manganese type lithium ion battery for pure electric vehicles (PEV). *In* "Proceedings of the 41st Japan Battery Symposium, The Electrochemical Society of Japan, November 20–22, Nagoya."

Ovshinsky, S. R., Fetcenko, M. A., Young, K., Fierro, C., Reichman, B., and Koch, J. (2002). High performance NiMH battery technology. *In* "Proceedings of the 19th International Seminar and Exhibit on Primary and Secondary Batteries, Florida Educational Seminars, Inc., March 11–14, Fort Lauderdale."

Pistoia, G. (ed.). (1994). "Lithium Batteries, New Materials, Developments and Perspectives." Elsevier, Amsterdam.

Rand, D. A. J., Woods, R., and Dell, R. M. (1998). "Batteries for Electric Vehicles." Research Studies Press, Taunton, UK.

St.-Pierre, C., Rouillard, R., Belanger, A., Kapfer, B., Simoneau, M., Choquette, Y., Gastonguay, L., Heiti, R., and Behun, C. (1999). Lithium–metal–polymer battery for electric vehicle and hybrid electric vehicle applications. *In* "Proceedings of EVS16 Symposium, China Electrotechnical Society, Beijing."

Sudworth, J., and Tilley, R. (1985). "The Sodium/Sulfur Battery." Chapman & Hall, London.

United Nations (1997). "UN Recommendations on the Transport of Dangerous Goods." 10th rev. ed. United Nations Sales Section, New York.

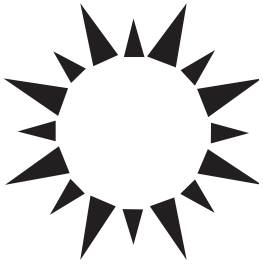
Van Schalkwijk, W. A., and Scrosati, B. (eds.). (2002). "Advances in Lithium-Ion Batteries." Kluwer Academic/Plenum, New York.

Vincent, C. A., and Scrosati, B. (1997). "Modern Batteries." 2nd ed. Arnold, London.

Yamanaka, K., Hata, K., Noda, T., Yamaguchi, K., and Tsubota, M. (2001). Development of 36 V valve regulated lead–acid batteries for hybrid vehicle application. *In* "Proceedings of the 42nd Japan Battery Symposium, The Electrochemical Society of Japan, November 21–23, Yokohama."

Bicycling

CHARLES KOMANOFF
Komanoff Energy Associates
New York, New York, United States



1. **Introduction**
2. **Bicycle Development**
3. **The Bicycle During the Auto Age**
4. **The Bicycle and Human Power**
5. **Bicycle Variety**
6. **Bicycles Around the World**
7. **Bicyclist Safety and Danger**
8. **Bicycle Policies**
9. **Bicycle Prospects**

Glossary

automobile Any self-guided, motorized passenger vehicle used for land transport, usually with four wheels and an internal combustion engine.
bicycle Any pedal-driven, two-wheeled conveyance propelled by human power.
bike lane A linear portion of a roadway demarcated for the predominant or exclusive use of bicycles.
bike path A linear path or pathway physically separated from conventional motorized roadways for the predominant or exclusive use of bicycles.
efficiency For transport vehicles, ratio of distance traversed to energy consumed.
gear For bicycles, relationship between pedaling and movement, often expressed as distance traveled in one revolution of the pedals (metric gear ratio).
sprawl Land-use patterns characterized by low population density, high automobile density, extensive road network, and little or no opportunity to travel via transit, cycling, or walking.
traffic calming Any roadway design feature or environmental intervention intended to reduce speeds and volumes of vehicular traffic.

A bicycle is any pedal-driven, two-wheeled conveyance propelled by human power. Bicycles revolutionized transport during the 1800s, becoming the first mass-produced personal transportation device, only to be literally pushed aside by automobiles during the 20th century. Nevertheless, the bicycle remains the world's most numerous transport vehicle, outnumbering cars by a two-to-one margin. Bicycles are the predominant transport vehicle in China and are a staple of urban transit throughout Northern Europe. They are the most energy-efficient machine of any kind in widespread use and are three orders of magnitude less energy-consuming than automobiles per distance traversed. Moreover, bicycles conserve not just on a per-trip basis but also at the social level by encouraging the substitution of proximity for distance and adding to the efficiency advantage of dense urban settlements over sprawling, suburbanized land-use patterns. In addition, cycling provides the opportunity to obtain healthful physical activity in the course of daily life while enhancing personal autonomy vital to mental health. Accordingly, preserving and indeed expanding the bicycle's role in urban transport is increasingly viewed as a global priority for social cohesion, urban viability, oil conservation, and protection of the climate against greenhouse gases.

1. INTRODUCTION

The bicycle—any pedal-driven, two-wheeled conveyance propelled by human power—is the world's most numerous transport vehicle and the most energy-efficient machine in widespread use. The bicycle was also the first mass-produced personal transportation device and created the preconditions for the development of the automobile a century ago. It is ironic, then, that the automobile, wherever it has been widely adopted, has largely driven the bicycle from the roads, causing global energy use to skyrocket along with a multitude of other social ills.

Consider that to travel 1 km on flat terrain, a cyclist operating a sturdy "roadster" bicycle at 10 mph

 © 2004 Elsevier Inc. All rights reserved.

expends a mere 60,000 joules (14.5 kilocalories). To cover the same distance with a typical U.S. passenger car rated at 20 mpg requires 115 million joules or nearly 2000 times more energy.

To be sure, the car/bicycle energy ratio varies with the number of people carried (cars' "load factors" average more than 1, but so do bicycles' load factors in developing countries), with vehicle weight and efficiency (small cars use 33–50% less fuel than so-called light trucks, and similarly lightweight "10-speed" bicycles require less energy than do roadster bikes), and with energy losses in processing gasoline and obtaining food. Still, under most combinations of assumptions, bicycles can cover a given distance using one-thousandth of the fuel that automobiles use.

Moreover, as social theorist Ivan Illich observed, the distance one travels is a product of the dominant mode of transport. Bicycles go hand in hand with short distances and urban density, whereas cars' voracious need for space both demands and feeds suburban sprawl. In short, bicycles serve proximity, whereas cars create distance.

Partly as a result of this vicious circle, cars have come to account for nearly 30% of world petroleum consumption and to produce nearly 15% of emissions of carbon dioxide, the most prominent greenhouse gas. Thus, expanding the role of bicycles vis-à-vis automobiles seems to be an obvious prescription for a world riven by conflict over petroleum and facing ecological upheaval from climate change.

Yet despite concerted efforts by thousands of cycle advocates in scores of countries, cycling appears to be losing ground, or at best running in place, in most nations' transport mix. Bikes outnumber cars worldwide, but they are used for fewer "trips" overall and probably cover less than one-tenth as many person-miles as do autos. How can this quintessentially human scale and efficient machine flourish again during the 21st century?

2. BICYCLE DEVELOPMENT

The essential elements of the bicycle are two wheels in line connected by a chain drive mechanism, with a rider simultaneously balancing and pedaling. Some bicycle historians attribute the first sketch of such a device to Leonardo da Vinci or, as argued by author David Perry, to an assistant in Leonardo's studio. A drawing with two large eight-spoked wheels, two pedals, a chain drive, saddle supports, a frame, and a tiller bar appears in the *Codex Atlanticus,* a volume of Leonardo's drawings and notations dating from around 1493 and assembled during the 16th century. However, the drawing was not discovered until 1966, during a restoration, leaving its authenticity open to question.

European inventors produced a number of self-propelled, hand- or foot-powered conveyances over the subsequent several centuries, all employing four wheels for stability. A practical two-wheeled device was first produced by the German Karl von Drais in 1816 and was patented 2 years later. His 40-kg Laufmaschine ("running machine") was propelled not by a mechanical drive but rather by pushing the feet against the ground, like a hobby horse, yet was capable of 13- to 14-kph speeds on dry firm roads. Similar machines with foot-operated drive mechanisms, using treadle cranks, appeared in Scotland during the late 1830s. True bicycles—two-wheel vehicles with pedal cranks located on the hub of the drive wheel—finally emerged in 1863 in France and spread quickly throughout Europe and to the United States.

Because these "pedal velocipedes" lacked gearing, each turn of the pedals advanced them only a distance equal to the circumference of the front wheel. To achieve high speeds, designers resorted to enormous front-drive wheels, reaching diameters of 4 feet for ordinary use and up to 6 feet for racing models. Although ungainly and difficult to operate, these "high-wheelers" proliferated and led to important innovations such as tangent-spoked wheels to resist torque, tubular diamond-shaped frames to absorb stress, hand-operated "spoon" brakes that slowed the wheels by pushing against them, and hubs and axles with adjustable ball bearings.

The final two steps in the evolution of the modern bicycle came during the 1880s: gearing to allow the use of smaller, more manageable wheels and chain-and-sprocket drives that transferred the drive mechanism to the rear wheel. These advances led to the so-called safety bicycle, the now-familiar modern design in which the cyclist sits upright and pedals between two same-sized wheels—the front for steering and the rear for traction.

As noted by Perry, this modern machine revolutionized cycling and is widely considered the optimal design. Innovations making bicycles safer, easier to use, and more comfortable followed in quick succession; these included pneumatic tires, "free-wheels" allowing coasting, multispeed gearing, and lever-actuated caliper brakes operating on rims rather than on tires. The safety bike transformed bicycling from a sport for athletic young men to a transport vehicle for men, women, and children

alike. By 1893, safety bicycles had replaced velocipedes, and by 1896, Americans owned more than 4 million bicycles—1 per 17 inhabitants.

The bicycle boom of the late 1800s swept through the industrialized world, and bicycle manufacture became a major industry in Europe and America. One census found 1200 makers of bicycles and parts, along with 83 bicycle shops, within a 1-mile radius in lower Manhattan. The pace of invention was so frenetic that during the mid-1890s the United States had two patent offices: one for bicycles and another for everything else. The lone urban traffic count in the United States to include bicycles, taken in Minneapolis, Minnesota, in 1906 after bicycling levels had peaked, found that bicycles accounted for more than one-fifth of downtown traffic—four times as much as did cars.

This "golden age of bicycling" proved to be short-lived. Following the classic pattern of corporate capitalism, a wave of mergers and buy-outs consolidated small shops run by enthusiasts and financed on the cheap into factories whose assembly-line efficiencies entailed high fixed costs. Overproduction followed, and then came market saturation, price wars, stock manipulations, and bankruptcies. Never universally popular, particularly in dense urban areas where swift and stealthy bicycles frightened pedestrians, the bicycle industry found its public image tarnished.

Of course, reversals of fortune were standard fare in late-19th century capitalism, and many industries, particularly those employing advanced technology, bounced back. Unfortunately for the bicycle business, on the heels of the shakeout in bike manufacture came the automobile.

3. THE BICYCLE DURING THE AUTO AGE

The bicycle catalyzed development of the car. A number of technical advances essential to the fledgling automobile industry, from pneumatic tires and ball bearings to factory-scale production engineering and a functional network of paved urban roads, owe their emergence to bicycles. No less important, the bicycle's ethos of independent, self-guided travel helped to split open the railroad-based paradigm of travel as mass transport along a fixed linear track.

But once the car took hold, it imposed its own ideology, one antithetical to bicycles. For one thing, cars used the same roads as did bicycles (the very roads that were paved as a result of bicyclists' campaigning), and through incessant noise and fumes, superior speed, and sheer physical force, cars literally pushed bicyclists aside. What is more, as recounted by social historian Wolfgang Sachs, the engine-driven car proved to be a more alluring cultural commodity than did the self-propelled bicycle. Although the bicycle leveraged bodily energy and broadened the individual's arena of direct activity many times over, the substitution of mechanical power for muscular exertion conveyed a sense of joining the leisure class and became equated with progress.

Thus, the bicycle's "defect of physicality," as Sachs termed it, put it at a disadvantage compared with the new technologies of internal combustion, electric motor drive, and flying machines. Rather than defend their right to cycle, the masses aspired to abandon the bicycle in favor of the auto. And abandon it they did, as fast as rising affluence and the advent in 1908 of the mass-produced, affordable car, Henry Ford's Model T, permitted. Although reliable data are lacking, by the end of the 1920s, bicycles probably accounted for only 1 to 2% of U.S. urban travel, an order-of-magnitude decline in just a few decades.

A similar devolution occurred in Europe during the long boom after World War II, albeit less steeply and with important exceptions. However, even now, bicycles outnumber cars by a two-to-one margin around the world, primarily due to economics. Cars cost roughly 100 times as much to buy as do bicycles and require fuel as well as maintenance, putting them out of reach of a majority of the world's people.

4. THE BICYCLE AND HUMAN POWER

According to data compiled by Vance A. Tucker of Duke University, a walking human consumes approximately 0.75 calorie of energy per gram of body weight for each kilometer traveled. This is less than the rate for birds, insects, and most mammals but is more than that for horses and salmon. However, atop a bicycle, a human's energy consumption falls fivefold, to a rate of roughly 0.15 calorie per gram per kilometer. As S. S. Wilson noted, "Apart from increasing his unaided speed by a factor of three or four, the cyclist improves his efficiency rating to No. 1 among moving creatures and machines."

Wilson attributed the bicycle's high efficiency mainly to its effective use of human muscles.

TABLE I
Cyclist Performance Factors

Factor	Average	Sport	Pro
Resting heart rate (breaths/min)	70	50	30
Anaerobic threshold (breaths/min)	160	175	185
Maximum heart rate (breaths/min)	180	190	195
Blood volume (ml)	10	25	50
Lung capacity (L)	5	6.5	8
VO_2 maximum (O_2 uptake, ml/kg-minute)	40	60	85
Thrust (Foot-pounds)	15	30	55
Cadence (rpm)	70	90	100
Watts	50	200	500
Calories (kcal)	135	750	2,150
Speed (kph)	15	30	50

Whereas a walker expends energy raising and lowering the entire body as well as accelerating and decelerating the lower limbs, the cyclist's sitting posture relieves the leg muscles of their supporting function. Because the cyclist's feet rotate smoothly at a constant speed and the rest of the body is still, the only reciprocating parts of the cyclist's body are the knees and thighs. Even the acceleration and deceleration of the legs is optimized given that one leg is raised by the downward thrust of the other. Wind resistance, the main constraint on the racing cyclist (because it varies with the square of the cyclist's velocity relative to the wind), is less significant at ordinary utilitarian speeds.

According to Perry, during a 1-h ride, an average person on a touring bike covering 15 km burns approximately 135 calories for an average power output of 50 W; over the same hour, a professional racing cyclist covers 50 km, burning 2150 calories and producing approximately 500 W (0.67 horsepower) (Table I).

"It is because every part of the design must be related to the human frame," wrote Wilson, "that the entire bicycle must always be on a human scale." He concluded, "Since the bicycle makes little demand on material or energy resources, contributes little to pollution, makes a positive contribution to health, and causes little death or injury, it can be regarded as the most benevolent of machines."

5. BICYCLE VARIETY

There is an enormous variety of bicycle types, reflecting human ingenuity, technical evolution, and a broad range of design criteria such as comfort, roadworthiness, speed, durability, and economy. As noted by Perry, a 22-pound road bicycle contains some 1275 parts in two dozen functional systems (e.g., wheels, chains, derailleurs, crank sets). Bicycle components require an unusual mix of lightness and durability, rigidity, and flexibility to provide a range of functions such as steering, braking, balancing, and climbing over a variety of terrains in various weather conditions. Moreover, in much of the world, bicycles must withstand the stress of carrying several passengers or cargoes weighing several hundred kilograms.

Most bikes fit into one of five broad categories:

- *Safety bicycles*. These bikes are the standard design established during the late 19th century and still employed throughout Asia and Africa. They have wide upright handlebars, medium-width tires, heavy construction for stability and durability, and up to three internal hub gears.
- *Racing or touring (10-speed or road) bicycles*. These bikes have the lightweight aerodynamic design initially developed for racing during the early 20th century and widely used today for sport. They have narrow "dropped" handlebars and skinny tires and achieve 10 gears through a double front chain ring and a 5-speed rear derailleur.
- *Mountain bikes*. These bikes are a recent (circa 1980) design, adding lightweight racing and touring components to the classic upright posture safety bicycle, with wide knobby tires for off-road use and a triple chain ring for steep inclines. A "hybrid" variant with slimmer tires is widely used in industrial countries for urban commuting.
- *Human-powered vehicles (or HPVs)*. HPVs are an entire class of machines combining aspects of bicycles, tricycles, and even cars developed by engineers and enthusiasts to "push the envelope" of nonmotorized transportation. They include recumbent bicycles in which riders recline against a backrest and the pedals are placed far forward. Using streamlined "fairing" to minimize wind resistance, HPVs have achieved 60-min speeds of 82 kph versus a maximum of 56 kph for standard racing bikes.
- *Utility cycles*. These bikes, with dedicated compartments and/or trailers for carrying large and heavy loads, are common in Asia and also are used in industrial nations in settings ranging from factory floors to urban food delivery. Pedicabs conveying passengers in separate compartments are widely used in China, Bangladesh, and parts of Africa, although authorities in Indonesia and elsewhere

have confiscated millions in forced motorization campaigns.

6. BICYCLES AROUND THE WORLD

Along with 1.2 billion bicycles, the world's 6.1 billion people possess 600 million motorized passenger vehicles (cars and light trucks), making for roughly 1 bike per 5 persons and 1 automobile per 10 persons. However, only a handful of countries actually show this precise two-to-one ratio because most of the world's motor vehicles are in the industrial nations, whereas most bicycles are in the developing world.

Although precise data are not available, it is likely that the average car is driven approximately 10,000 miles per year, whereas the average bicycle probably logs fewer than 500 miles per year. Based on these rough figures, the world's bicycles collectively travel less than one-tenth as many miles as do cars, although their share of trips is somewhat larger.

Three countries or regions are of particular interest: China, the world's most populous nation and still a bicycling stronghold, in spite of policies designed to encourage car use; the United States, the *ne plus ultra* of automobile use; and Northern Europe, where public policy restraining automobile use has brought about a bicycle renaissance amid affluence (Table II).

4.1 China

Large-scale bicycle manufacture and use have been a centerpiece of Chinese industrialization and urbanization since shortly after the 1949 revolution. By the 1980s, production for both domestic use and export had reached 40 million bikes per year, outnumbering total world car output. Today, China's 1.3 billion people own a half-billion bicycles, 40% of the world's total, and the bicycle is the mainstay of urban transportation throughout the country. Not just individuals but also whole families and much cargo are conveyed on bicycles, bike manufacture and servicing are staples of China's economy, and mass urban cycling is an indelible part of China's image in the world. Traffic controllers in the largest cities have counted up to 50,000 cyclists per hour passing through busy intersections; in comparison, a four-lane roadway has a maximum throughput of 9000 motor vehicles per hour.

This is now changing, perhaps irreversibly, as China invests heavily in both automobiles and mechanized public transport. Although domestic auto sales in 2002 numbered just 800,000 versus bicycle sales of 15 to 20 million, the auto sector is growing rapidly at 10 to 15% per year. "Bicycle boulevards" in Beijing and other major cities have been given over to cars, with bikes excluded from 54 major roads in Shanghai alone. The safety and dignity enjoyed by generations of Chinese are beginning to crumble under the onslaught of motorization.

Although the number of bicycles in China is still growing, sales have dropped by one-third since the early 1990s. With cyclists increasingly forced onto buses and subways, the bicycle's share of trips in Beijing and Shanghai has fallen precipitously to 40% and 20%, respectively, from more than 50% a decade or so ago. Indeed, the incipient conversion of the world's premier bicycle nation into a car-cum-transit society is eerily reminiscent of America a century ago.

TABLE II

Bicycles and Automobiles in Selected Countries and Regions (circa 2000)

Country or region	Bicycles	Autos	Bicycle/auto ratio	Bicycles per 1000	Autos per 1000
China	500,000,000	18,000,000	28	392	14
India	60,000,000	10,000,000	6	59	10
Japan	60,000,000	40,000,000	1.5	472	315
Germany	60,000,000	40,000,000	1.5	732	488
The Netherlands	12,000,000	6,000,000	2	750	375
United States	120,000,000	180,000,000	0.7	421	632
Argentina	5,000,000	5,000,000	1	135	135
Africa	40,000,000	20,000,000	2	50	25
World totals	1,200,000,000	600,000,000	2	198	99

The difference is that China's population is an order of magnitude larger than that of the United States in 1900, making the fate of cycling in China a matter of global moment. Profligate burning of fossil fuels is recognized to threaten humanity through global climate change, and a simple calculation demonstrates that if China were to match the U.S. per capita rate of auto use, world carbon dioxide emissions would rise by one-quarter, a catastrophic defeat in the effort to limit greenhouse gases.

6.2 United States

Like no other society in history, the United States is dominated spatially, economically, and psychologically by automobiles. Registered autos outnumber bikes by a two-to-one ratio, but more important, more than 90% of individuals' transportation trips are by car. A mere one-hundredth as many, or 0.9%, are by bicycle, and a majority of these are for recreational riding rather than for "utilitarian" transport. Some reasons follow:

- *Car culture*. Under America's cultural triumvirate of mobility, physical inactivity, and speed, the car has been enshrined as the norm and cycling is consigned to the margins. In turn, the perception of cycling as eccentric or even deviant contributes to an unfavorable climate for cycling that tends to be reinforcing.
- *Sprawling land use*. A majority of Americans live in suburbs, increasingly in the outermost metropolitan fringe whose streets and roads are suitable only for motorized vehicles. Cities are more conducive to cycling, with smaller distances to cover and congestion that limits motor vehicle speeds, but they too are engineered around autos, and few have marshaled the fiscal and political resources to ensure safety, much less amenity, for cyclists.
- *Subsidized driving*. Low gasoline taxes, few road tolls, and zoning codes requiring abundant free parking are a standing invitation to make all journeys by car, even short trips that could be walked or cycled.
- *Poor cycling infrastructure*. Spurred by federal legislation letting states and localities apply transportation funds to "alternative" modes, the United States has invested $2 billion in bicycle facilities since the early 1990s. Nevertheless, provision of on-street bike lanes, separated bike paths, cycle parking, and transit links has been haphazard at best and is often actively resisted.
- *Cycling danger*. A bike ride in the United States is three times more likely to result in death than is a trip in a car, with approximately 800 cyclists killed and 500,000 injured annually. Not surprisingly, the prospect of accident and injury is a powerful impediment to bicycling in the United States. Moreover, cycling's actual risks are compounded by cultural attitudes that attribute cycle accidents to the supposedly intrinsic perils of bicycles, unlike motorist casualties, which are rarely considered to imply that driving as such is dangerous.

These inhibiting factors are mutually reinforcing. Particularly with America's pressure group politics, the lack of broad participation in cycling severely limits support for policies to expand it. The lack of a consistent visible presence of cyclists on the road exacerbates the inclination of drivers to see cyclists as interlopers and to treat them with active hostility. "Feedback loops" such as these keep cycling levels low.

6.3 Northern Europe

There is one region in which bicycling coexists with affluence and automobiles. Despite high rates of car ownership, more than a half-dozen nations of Northern Europe make at least 10% of urban trips by bike, surpassing the U.S. mode share at least 10-fold.

The highest bike share, and the steadiest, is in The Netherlands, with 26% of urban trips in 1978 and 27% in 1995. Cycling's modal share rose sharply in Germany during the same period, from 7 to 12%—still less than half the Dutch level but impressive given Germany's rapid growth in auto ownership and use. Also notable is cycling's high share of trips made by seniors: 9% in Germany and 24% in The Netherlands.

Robust bicycling levels in Northern Europe are not accidental but rather the outcome of deliberate policies undertaken since the 1970s to reduce oil dependence and to help cities avoid the damages of pervasive automobile use. Not only are road tolls, taxes, and fees many times higher than those in the United States, but generously funded public transit systems reduce the need for cars, increasing the tendency to make short utilitarian trips by bicycle. Particularly in Germany, Denmark, and The Netherlands, comprehensive cycle route systems link "traffic-calmed" neighborhoods in which cycling alongside cars is safe and pleasant.

Indeed, the same kind of feedback loops that suppress cycling in the United States strongly nurture

it in Northern Europe. Both density and bicycling are encouraged by policies ranging from provision of transit and cycle infrastructures to "social pricing" of driving; and Northern European states refrain from subsidizing sprawl development. Not only do a majority of Europeans live in cities as a result, but population densities in urban areas are triple those in the United States; correspondingly, average trip distances are only half as great, a further inducement to cycle.

7. BICYCLIST SAFETY AND DANGER

The bicycle's marvelous economy and efficiency have negative corollaries. First, unlike three- or four-wheeled conveyances, bikes are not self-balancing. Continuous motion is required to keep them upright, and they can tip and crash due to road defects, mechanical failure, or operator error. Second, crash mitigation measures such as seat belts, air bags, and crumple zones that have become standard in automobiles are not feasible for bicycles; only helmets offer a modicum of protection, and perhaps less than is commonly believed.

The exposed position of cyclists on the road makes them vulnerable to motor vehicles. Surpassing bicycles several-fold in velocity and at least 100-fold in mass, automobiles have approximately 1000 times more kinetic energy to transfer to a bicycle in a collision than is the case vice versa. Not surprisingly, although most of the total injury-accidents to bicyclists occur in falls or other bike-only crashes, severe injuries and fatalities are due mostly to being hit by cars. Approximately 90% of bicycle fatalities (95% for child cyclists) in the United States have motorist involvement, and the percentages elsewhere are probably as high.

Ironically, bicycles pose little danger for other road users and far less than do automobiles. In the United States, fewer than 5 pedestrians die each year from collisions with bikes, whereas 5000 are killed by motor vehicles. Motor vehicle users exact an enormous toll on themselves as well as on each other. Worldwide, total road deaths are estimated at 1 million people each year, with millions more becoming disabled in accidents. Based on "disability-adjusted life years," a statistic incorporating permanent injuries and the relative youth of victims, road deaths were ranked by the World Health Organization as the world's ninth-leading health scourge in 1990 and were predicted to rank third by 2020.

7.1 Worldwide Differences

Bicycle safety policies differ widely around the world, as illustrated in the regions profiled earlier. China and other developing countries are too poor to invest in bicycle safety programs, but until recently they were also too poor for the cars that make such programs necessary. Historically, cyclists in China and other Asian nations have been able to rely on their sheer numbers to maintain their rights-of-way. In the United States, bicycle safety measures focus on changing cyclist behavior, primarily increasing helmet use (especially by children) and training cyclists to emulate motor vehicles through "effective cycling" programs. Little effort is made to address the nature and volume of motor vehicle traffic or the behavior of drivers toward cyclists.

In contrast, bicycle safety in Europe is promoted holistically as part of policies to encourage widespread cycling and universal road safety. Germany and The Netherlands promote both through provision of elaborate and well-maintained cycling infrastructures, urban design oriented to cycling and walking rather than to motor traffic, disincentives for and restrictions on car use, and enforcement of traffic regulations that protect pedestrians and cyclists.

The European safety model appears to be validated by low fatality rates. Despite stable or rising cycling levels from 1975 to 1998, cycle fatalities fell 60% in Germany and The Netherlands. The U.S. decline was less than half as great (25%) and may have been largely an artifact of reduced bicycling by children. Currently, bicycle fatalities per kilometer cycled are two to three times lower in Germany and The Netherlands than in the United States, and pedestrian fatalities per kilometer walked are three to six times less. Perhaps most tellingly, although Germany and The Netherlands have higher per-kilometer fatality rates for auto users than does the United States, their overall per capita rates of road deaths are lower—by one-third in Germany and by one-half in The Netherlands—in large part because cars are used less in both countries.

7.2 Cycle Helmets

Helmets have become an intensely polarized subject in bicycling during recent years. Many observers trace the origins of the debate to a 1989 epidemiological study in Seattle, Washington, associating helmet use with an 85% reduction in brain and head injuries. The authors subsequently employed better statistical methods and scaled back their results

considerably, to a mere 10% reduction in severe injuries when body and not just head trauma is taken into account. But the initially reported connection between helmet use and injury reduction sparked campaigns in the United States and Australia to compel helmet use by child cyclists (later extended to skateboards, roller skates, and scooters) and to promote helmet wearing by adult cyclists.

Whether these campaigns have been useful or not is difficult to say. Child cycling fatalities have decreased in the United States, but that may be because fewer children now cycle. Adult fatalities have risen, possibly due to growth in the more dangerous kinds of motor traffic such as sport utility vehicles and drivers' use of mobile phones. Nevertheless, although links between helmet promotion and cycling injury prevention are inconclusive, the "helmet paradigm" is firmly established in U.S. policy.

In Europe, where injury prevention is subordinated to the larger goal of health promotion and where social responsibility is emphasized alongside individual accountability, helmets are considered irrelevant or even counterproductive to health. The influential British social scientist Mayer Hillman contends that cardiovascular and other physiological and psychological gains from cycling far outweigh the rider's crash risk, even in unsatisfactory present-day road environments. Therefore, Hillman's paradigm of cycle encouragement holds that cycling is so beneficial to individuals and society that no interferences should be tolerated, not even the inconvenience and unattractiveness of a helmet or the subliminal message that helmets may send about the dangers of cycling.

7.3 Safety in Numbers

Anecdotal evidence has long suggested that the per-cyclist rate of bicycle–motor vehicle crashes declines as the amount of cycling on a road or in a region increases. This "safety in numbers" effect is thought to occur because as cyclists grow more numerous and come to be an expected part of the road environment, motorists become more mindful of their presence and more respectful of their rights. The implication is that adding more cyclists to the road makes it less likely that a motorist will strike an individual cyclist and cause serious injury. Conversely, removing cyclists from the traffic stream raises the risk to those who continue to cycle.

This safety in numbers effect offers a plausible explanation for the fact that per-kilometer cycling fatality rates in Germany and The Netherlands are four times less than that in the United States, even though cycling percentages are more than 10 to 20 times higher in these European countries. Now, time-series estimates of this effect, although preliminary and site specific, are pointing intriguingly toward a "power law" relationship of approximately 0.6 between cyclist numbers and cyclist safety. According to this relationship, the probability that a motorist will strike an individual cyclist on a particular road declines with the 0.6 power of the number of cyclists on that road. Say the number of cyclists doubles. Because 2 raised to the 0.6 power is 1.5, each cyclist would be able to ride an additional 50% without increasing his or her probability of being struck. (The same phenomenon can be expressed as a one-third reduction in per-cyclist crash risk per doubling in cycling volume given that the reciprocal of 1.5 is 0.67.)

The implications for cycling are profound. Countries that have based safety promotion on cyclist behavior modification (e.g., the United States) might reconstruct safety in a social context. One consequence would be to deemphasize helmet use in favor of jump-starting broader participation in cycling so as to stimulate a "virtuous circle" in which more cycling begets greater safety, which in turn encourages more cycling. In addition, countries such as China might reconsider policies that threaten to erode large-scale cycling, lest safety in numbers in reverse leads to a downward spiral as in the current U.S. situation, where bike riding is limited to small numbers of enthusiasts and to others who have no alternatives.

8. BICYCLE POLICIES

Policies to support and "grow" bicycling fall into three categories: cycling infrastructure, cyclists' rights, and disincentives to driving.

Cycling infrastructure policies aim to attract cycle trips by providing "facilities" such as on-street bicycle lanes, off-street bicycle paths (e.g., "greenways") with separate rights-of-way, bicycle parking, and integration with the metropolitan or regional transit system. Constructing and maintaining such facilities has proven to be politically popular in some states and localities in the United States and absorbed most of the $2 billion spent on bicycle programs from 1992 to 2002.

Cyclists' rights initiatives seek to improve the legal standing of cycling and, thus, to make cycling safer and more socially validated. It is believed that exerting closer authority over driver conduct through the legal system, police enforcement, and cultural shifts would directly reduce the threat to bicyclists and so encourage cycling.

Disincentives to driving are policies to make driving less attractive economically and logistically and, therefore, to reduce the level of motor traffic. Measures falling under the rubric of "social pricing" of automobiles include gasoline taxes, "carbon" taxes on fossil fuels, road pricing (fees on each kilometer driven), and restructuring auto insurance and local road taxes to pay-per-use.

That these three kinds of initiatives are complementary is seen by examining Germany and The Netherlands, which have used all three to maintain and restore bicycling since the early 1970s. No single approach is sufficient, and each supports the others by increasing opportunities and rewards for cycling and establishing a social context in which cycling is "valorized" as appropriate and praiseworthy rather than viewed as deviant behavior.

9. BICYCLE PROSPECTS

Strong societal currents are pushing bicycling forward, yet equally mighty forces are suppressing it. Much hangs in the balance, both for billions of the earth's peoples who may wish to master their own mobility through cycling and for our planet's ecological and political well-being.

Each trip cycled instead of driven conserves gasoline and stops the addition of climate-altering carbon dioxide to the atmosphere. The potential effects are large given that passenger vehicles account for nearly one-third of global petroleum consumption and generate more than one-eighth of carbon dioxide emissions. "Green cars" are no more than a palliative; a world in which everyone drove at the U.S. per capita rate would emit more carbon dioxide than is currently the case, even if autos could be made to be five times more efficient.

Thus, sustaining the earth's climate and political equilibrium requires robust alternatives to the American model of one car per journey. Moreover, bicycles conserve several times over, and not just at the per-trip level, by encouraging the substitution of proximity for distance and adding to the efficiency advantage of dense urban settlements over sprawling suburbanized land-use patterns. Therefore, ecological imperatives are a potent reason to maintain bicycling in China and other developing countries as well as to foster it in automobile-dependent societies such as the United States.

Health promotion is a major rationale as well. As noted, road traffic accidents are or soon will be among the world's half-dozen leading causes of death and disability. Sedentary lifestyles, including substitution of motorized transport for walking and cycling, are also recognized as a cause of fast-growing obesity and of ill health generally. In contrast, cycling provides the opportunity to obtain physical activity in the course of daily life while enhancing personal autonomy and aiding mental health.

Yet motorization itself is a powerful suppressant to cycling, and not just in the often-lethal competition between cars and bikes. Just as pernicious is the automobile's grip on transportation's "mind-share"—the automatic equating of mobility with motor vehicles and of motor vehicles with success—that leaves bicyclists, both individually and institutionally, on the outside looking in.

Large-scale cycling seems reasonably assured in the countries of Northern Europe that view it as integral to national objectives of reducing greenhouse gases, sustaining urban centers, and promoting health and self-guided mobility. Preserving mass cycling in China and developing it in the United States will probably require dethroning the automobile as the symbol and engine of prosperity and sharply reducing its enormous financial and political power—a tall order.

The bicycle—a pinnacle of human efficiency and an icon of vernacular culture for more than a century—will survive. Whether it will again flourish may make a difference in how, and whether, humanity itself survives.

SEE ALSO THE FOLLOWING ARTICLES

Alternative Transportation Fuels: Contemporary Case Studies • *Development and Energy, Overview* • *Fuel Cycle Analysis of Conventional and Alternative Fuel Vehicles* • *Global Energy Use: Status and Trends* • *Internal Combustion Engine Vehicles* • *Lifestyles and Energy* • *Motor Vehicle Use, Social Costs of* • *Passenger Demand for Travel and Energy Use* • *Vehicles and Their Powerplants: Energy Use and Efficiency*

Further Reading

Carlsson, C. (ed.). (2002). "Critical Mass: Bicycling's Defiant Celebration." AK Press, Edinburgh, UK.

Herman, M., Komanoff, C., Orcutt, J., and Perry, D. (1993). "Bicycle Blueprint: A Plan to Bring Bicycling into the Mainstream in New York City." Transportation Alternatives, New York.

Hillman, M. (1992). "Cycling: Towards Health and Safety." British Medical Association, London.

Illich, I. (1974). "Energy and Equity." Harper & Row, New York.

Jacobsen, P. (2003). Safety in numbers: More walkers and bicyclists, safer walking and bicycling. *Injury Prevention* **9**, 205–209.

Komanoff, C., and Pucher, J. (2003). Bicycle transport in the U.S.: Recent trends and policies. *In* "Sustainable Transport: Planning for Walking and Cycling in Urban Environments" (R. Tolley, Ed.). Woodhead Publishing, Cambridge, UK.

Lowe, M. D. (1989). "The Bicycle: Vehicle for a Small Planet." Worldwatch Institute, Washington, DC.

McShane, C. (1994). "Down the Asphalt Path: The Automobile and the American City." Columbia University Press, New York.

Perry, D. B. (1995). "Bike Cult." Four Walls Eight Windows, New York.

Pucher, J. (1997). Bicycling boom in Germany: A revival engineered by public policy. *Transport. Q.* **51**(4), 31–46.

Pucher, J., and Dijkstra, L. (2000). Making walking and cycling safer: Lessons from Europe. *Transport. Q.* **54**(3), 25–50.

Rivara, F. P., Thompson, D. C., and Thompson, R. S. (1997). Epidemiology of bicycle injuries and risk factors for serious injury. *Injury Prevention* **3**(2), 110–114.

Sachs, W. (1992). "For Love of the Automobile." University of California Press, Berkeley.

Tolley, R. (ed.). (2003). "Sustainable Transport: Planning for Walking and Cycling in Urban Environments." Woodhead Publishing, Cambridge, UK.

Whitt, F. R., and Wilson, D. G. (1989). "Bicycling Science." MIT Press, Cambridge, MA.

Wilson, S. S. (1973). Bicycle technology. *Sci. Am.* **228**, 81–91.

Biodiesel Fuels

LEON G. SCHUMACHER
University of Missouri, Columbia
Columbia, Missouri, United States

JON VAN GERPEN
Iowa State University
Ames, Iowa, United States

BRIAN ADAMS
University of Missouri, Columbia
Columbia, Missouri, United States

1. **Biodiesel Fuels and Their Origins**
2. **Chemical Properties**
3. **Physical Properties**
4. **Biodiesel Production**
5. **Advantages of Biodiesel**
6. **Disadvantages**
7. **Biodiesel Storage and Use**
8. **Biodiesel Economic Considerations**

Glossary

aromatics Any unsaturated hydrocarbon containing resonance stabilized carbon-to-carbon bonds characterized by benzene-type ring structure.

biodiesel A renewable energy fuel from vegetable or animal derived oil (triglyceride) that has been chemically modified to reduce its viscosity. It can be used in any concentration with petroleum-based diesel fuel in existing diesel engines with little or no modification. Biodiesel is not the same thing as raw vegetable oil. It is produced by a chemical process, which removes the glycerol from the oil.

distillation curve A measure of volatility of a fluid. More specifically, it is the temperature of the fluid (fuel) when a specific volume of the fluid has evaporated.

feedstock Raw or processed material that is chemically reacted to produce biodiesel. Note that the raw material must first be processed before the oil can be used to produce biodiesel.

gas chromatograph A chemical is heated to a gaseous state, and then the gas is passed through a cylinder and the different chemicals that make up the gas "stick" to the wall of the cylinder at different intervals. This information is then used to determine the chemical makeup of the gas.

olefin Any unsaturated hydrocarbon containing one or more pairs of carbon atoms linked by a double bond.

triglyceride A naturally occurring ester formed from glycerol and one to three fatty acids. Triglycerides are the main constituent of vegetable and animal derived fats and oils.

The United States depends heavily on imported oil to fuel its transportation infrastructure. The use of alternative fuel derived from plant oils was examined by researchers in the mid-1970s to determine if internal combustion engines could be fueled from sources other than petroleum. The initial research on pure vegetable oils as a replacement for petroleum diesel fuel was met with mostly negative results. Researchers determined that transesterification of these plant- and animal-derived oils reduced the viscosity of the oil without any other significant changes to the oil. Since the new fuel was bio-derived and was used to fuel a diesel engine, the name "biodiesel" was selected to refer to the new fuel. This article focuses more specifically on how biodiesel fuel was developed, its chemical and physical properties, advantages, disadvantages, and how biodiesel is used and stored. The article concludes by reviewing the economic issues associated with its use.

1. BIODIESEL FUELS AND THEIR ORIGINS

1.1 What Is Biodiesel?

Biodiesel is made from a number of feedstocks including vegetable oil, tallow, lard, and waste

 © 2004 Elsevier Inc. All rights reserved.

cooking oils (yellow grease). "Biodiesel is defined as mono-alkyl esters of long chain fatty acids derived from vegetable oils or animal fats which conform to American Society of Testing Materials (ASTM D5453) International specifications for use in diesel engines." Biodiesel contains no petroleum, but it can be blended at any level with petroleum diesel to create a biodiesel blend. Feedstocks can be transesterified to make biodiesel using an alcohol that has been mixed with a catalyst such as potassium hydroxide or sodium hydroxide. The most commonly used alcohol for transesterification is methanol. Methanol reacts easily and is less expensive to use than most other alcohols.

Soybean derived biodiesel is the most commonly used biodiesel in the United States. The most commonly used feedstock for biodiesel production in Europe is rapeseed. Biodiesel is biodegradable, nontoxic, and essentially free of sulfur and aromatics. Biodiesel is considered a renewable resource due to the fact that it is derived from products that can be grown and produced domestically.

1.2 What Biodiesel Is Not

In 1898 when Rudolph Diesel's compression ignition engine was demonstrated at the World's Exhibition in Paris, it ran on virgin peanut oil. Some document this event as the first use of biodiesel. However, biodiesel is not vegetable oil or animal fats. Biodiesel refers to the alkyl esters produced from a transesterification reaction between the oil or fat and an alcohol. Others refer to mixtures of biodiesel and petroleum diesel fuel as "biodiesel." This mixture is referred to as a biodiesel blend and is commonly designated in much the same way as a blend of gasoline and alcohol (E85). For example, B100 is 100% biodiesel; B20 is a blend of 20% biodiesel and 80% diesel fuel.

1.3 Beginnings

Vegetable oils were transesterified prior to the mid-1800s. Transesterification is the process of reacting a triglyceride molecule with an excess of alcohol in the presence of a catalyst (KOH, NaOH, $NaOCH_3$, etc.) to produce glycerol and fatty esters as shown in the chemical reaction in Fig. 1.

Companies such as Proctor & Gamble have used this process to make soap for years. According to Knothe, the earliest known use of alkyl esters for fuel appears in a Belgian patent granted in 1937 to G. Chavanne.

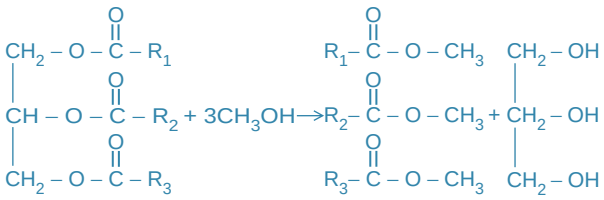


FIGURE 1 Chemical structure of biodiesel.

Methyl and ethyl esters are essentially by-products of this process. Ethyl esters are made using ethanol and methyl esters are made using methanol. Ethanol is made from grain such as corn. The methanol is either wood based or derived from natural gas (methane).

Peanut oil, hemp oil, corn oil, and tallow are typically transesterified in the soap-making process. Soybeans, industrial or edible rapeseed (or its cousin, canola oil), corn, recycled fryer oil, lard, and tallow are common resources for the complex fatty acids and their by-products.

It is important to find other sources of oil to enhance our ability to produce biodiesel. Research has been conducted concerning oil production from algae according to sources at the National Renewable Energy Laboratory (NREL). This oil source could have yields greater than any feedstock now known. Most biodiesel researchers worldwide agree that it would be difficult to replace more than 10% of the diesel fuel that is used for transportation purposes with biodiesel.

2. CHEMICAL PROPERTIES

Researchers have determined that the following properties characterize the chemical properties of biodiesel: fatty acid content, aromatics, olefins, paraffins, carbon, hydrogen, oxygen, and sulfur content, acid neutralization number, iodine number, and Conradson carbon residue number.

Note the chemical structure of a triglyceride (Fig. 2). R_1, R_2, and R_3 represent the hydrocarbon chain of the fatty acid elements of the triglyceride. There is a three-carbon chain called the glycerol backbone that runs along the left side of the molecule. Extending away from this backbone are the three long fatty acid chains. The properties of the triglyceride and the biodiesel fuel will be determined by the amounts of each fatty acid that are present in the molecules.

Fatty acids are designated by two numbers: the first number denotes the total number of carbon

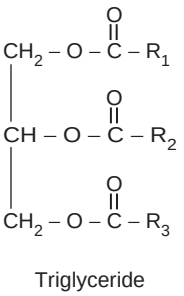


FIGURE 2 Chemical structure of a triglyceride.

Chain	Percentage by weight
C14:0	0.278
C16:0	10.779
C18:0	4.225
C18:1	20.253
C18:2	54.096
C18:3	9.436
C20:0	0.395

FIGURE 3 Fatty acid profile for soybean derived biodiesel.

atoms in the fatty acid, and the second is the number of double bonds. For example, 18:1 designates oleic acid, which has 18 carbon atoms and one double bond. A typical sample of soybean oil based biodiesel would have the fatty acid profile shown in Fig. 3.

Biodiesel is essentially free of sulfur and aromatics. This is an advantage for biodiesel because sulfur poisons catalytic converter technology that is used to reduce engine exhaust emissions. The sulfur levels in biodiesel by ASTM D5453 are found to be as low as 0.00011% by mass (1 ppm), where petroleum diesel is often no lower than 0.02% (200 ppm). The lack of aromatic hydrocarbons is also an advantage for biodiesel, as many of these compounds are believed to be carcinogenic. The test procedure normally used to measure aromatics in petroleum fuel (ASTM D1319) should not be used to determine the aromatics of biodiesel. This analytical procedure mistakenly identifies the double bonds commonly found in biodiesel as the resonance stabilized bonds normally associated with aromatics.

Paraffins are hydrocarbon compounds that are normally associated with petroleum diesel fuel. These compounds help increase the cetane value of the diesel fuel. However, they also typically increase cold flow problems of petroleum diesel fuel. Olefins are hydrocarbons that contain carbon-carbon double bonds. Molecules having these types of bonds are called unsaturated. Biodiesel from common feedstocks is usually 60 to 85% unsaturated. Some olefins are present in petroleum-based diesel fuel. However, the amount of olefins is usually small as they contribute to fuel oxidation.

The carbon content of biodiesel is nearly 15% lower than petroleum diesel fuel on a weight basis. Conversely, biodiesel has approximately 11% oxygen, on a weight basis, while petroleum diesel has almost no oxygen. Very little differences exist between biodiesel and petroleum diesel fuel concerning the weight percentage of hydrogen.

The neutralization number is used to reflect the acidity or alkalinity of an oil. This number is the weight in milligrams of the amount of acid (hydrochloric acid [HCL]) or base (potassium hydroxide [KOH]) required to neutralize one gram of the oil, in accordance with ASTM test methods. If the neutralization number indicates increased acidity (i.e., high acid number) of an oil, this may indicate that the oil or biodiesel has oxidized or become rancid. Biodiesel is allowed to have a neutralization number up to 0.8mg KOH/g.

The iodine value is a measure of the unsaturated fatty acid content of biodiesel, and reflects the ease with which biodiesel will oxidize when exposed to air. The iodine value for petroleum diesel fuel is very low, but the iodine value of biodiesel will vary from 80 to 135.

A weighed quantity of fuel is placed in a crucible and heated to a high temperature for a fixed period to determine the Conradson carbon residue of a fuel. The crucible and the carbonaceous residue is cooled in a desiccator and weighed. The residue that remains is weighed and compared to the weight of the original sample. This percentage is reported as the Conradson carbon residue value. This procedure provides an indication of relative coke forming properties of petroleum oils. No real differences are to be expected when comparing biodiesel with petroleum diesel fuel (0.02 versus 0.01).

3. PHYSICAL PROPERTIES

The following properties reflect the physical properties of biodiesel: distillation curve, density/specific gravity, API gravity, cloud point, pour point, cold filter plug point, flash point, corrosion, viscosity, heat of combustion, and cetane number.

Each fuel has its own unique distillation curve. Some compare this to a fingerprint. This curve tells a chemist which components are in the fuel, their

molecular weights by identifying their boiling points and their relative amounts. The temperature at which a specific volume boils off is used to establish this curve (0, 10, 50, 90, and 100%). The range of boiling points for biodiesel is much narrower than petroleum diesel fuel. For example, the initial boiling point of petroleum diesel fuel is 159°C. and the end boiling point is 336°C. The initial boiling point of biodiesel is 293°C. and the end boiling point is 356°C. Thus, the distillation range for diesel fuel is 177° versus 63°C for biodiesel fuel.

The specific gravity of a product is the weight of the product compared to an equal volume of water. The specific gravity of biodiesel is slightly higher than petroleum diesel fuel. For example, the specific gravity of No. 2 petroleum diesel fuel is approximately 0.84 (7.01 pounds per gallon). The specific gravity of biodiesel is approximately 0.88 (7.3 pounds per gallon).

ASTM test procedures are used to determine cloud and pour point of biodiesel. Cloud point is the temperature at which the first wax crystals appear. Pour point is the temperature when the fuel can no longer be poured. According to ASTM specifications for diesel fuel, the cloud points of petroleum diesel fuel are determined by the season (temperature) that the fuel is used. For example, ASTM standards essentially require that the fuel distributor treat number 2 petroleum diesel fuel with cold flow improvers (CFI) for winter operation. Cold weather operation with 100% biodiesel also forces the distributor to use a CFI. However, the CFI of choice for petroleum diesel fuel is usually not as effective when used with biodiesel. The cloud point for biodiesel is higher than petroleum diesel fuel (1.6°C for biodiesel compared with −9.4° to −17.7°C for diesel fuel). Chemical companies (Lubrizol, Octell Starreon) are experimenting with CFI chemicals that work effectively with biodiesel.

An alternative method used to determine how the fuel will perform during cold weather operation is the cold filter plugging point. The cold filter plugging point is the lowest temperature at which the fuel, when cooled under specific conditions, will flow through a filter during a given period of time. The filter is a defined wire mesh. This procedure indicates the low temperature operability of a fuel with/without cold flow improver additives when cooled below the cloud point temperature. The cold filter plugging point for fuels has become an important issue due to the reduction in the size of the openings in the fuel filters (2–5 microns versus 45 microns).

The flash point of a fuel is defined as the temperature at which the air/fuel vapor mixture above the product will ignite when exposed to a spark or flame. A sample is heated and a flame is passed over the surface of the liquid. If the temperature is at or above the flash point, the vapor will ignite and a detectable flash (unsustained or sustained flame) will be observed. The flash point of biodiesel is substantially higher (159°C versus 58°C) than for petroleum diesel fuel. Biodiesel is categorized as a combustible fuel, not a flammable fuel, as is the standard classification for petroleum diesel fuel. Biodiesel is a safer fuel to transport due to its higher flash point.

As fuel deteriorates it becomes acidic. Copper is particularly susceptible to corrosion by the acids in the fuel. As a result, test procedures have been developed to detect the fuel's corrosiveness to copper. A polished copper strip is immersed in a heated sample of fuel. After a prescribed period of time, the strip is removed and examined for evidence of corrosion. A standardized system is used to assign a value between 1 and 4. This number is assigned based on a comparison with the ASTM Copper Strip Corrosion Standards. No differences have been detected using this methodology between petroleum diesel fuel and biodiesel.

The viscosity of a fluid is a measure of its resistance to flow. The greater the viscosity, the less readily a liquid will flow. Test procedures are used to measure the amount of time necessary for a specific volume of fuel to flow through a glass capillary tube. The kinematic viscosity is equal to the calibration constant for the tube multiplied by the time needed for the fuel to move through the tube. The viscosity of biodiesel is approximately 1.5 times greater than petroleum diesel fuel (4.01 versus 2.6 cSt@40°C).

The heat of combustion is the amount of energy released when a substance is burned in the presence of oxygen. The heat of combustion, also known as the heating value, is reported in two forms. The higher, or gross, heating value assumes that all of the water produced by combustion is in the liquid phase. The lower, or net, heating value assumes the water is vapor. The lower heating value for biodiesel is less than for number 2 diesel fuel (37,215 kJ/kg versus 42,565 kJ/kg).

The cetane number reflects the ability of fuel to self-ignite at the conditions in the engine cylinder. In general, the higher the value, the better the performance. The cetane number of biodiesel varies depending on the feedstock. It will be 48 to 52 for soybean oil based biodiesel and more than 60 for recycled greases. This is much higher than the typical

TABLE I

ASTM Biodiesel Specification—D6751 versus ASTM LS #2 Diesel Fuel—D975

Property	ASTM Method	Limits D6751	Units D6751	Limits D975	Units D975
Flash point	D3	130.0 min.	°C	52.0 min.	°C
Water and Sediment	D2709	0.050 max.	% vol.	0.050 max.	% vol.
Carbon residue, 100% sample	D4530	0.050 max.	% mass	N/A	N/A
Ramsbottom carbon residue	D524	N/A	N/A	0.35	% mass
Sulfated ash	D874	0.020 max.	% mass	N/A	N/A
Ash	D482	N/A	N/A	0.01	% mass
Kinematic viscosity, 400°C	D445	1.9–6.0	mm2/s	1.9–4.1	mm2/s
Sulfur	D5453	0.05 max.	% mass	N/A	N/A
Sulfur	D2622	N/A	N/A	0.05 max.	% mass
Cetane	D613	47 min.		40 min.	
Cloud point	D2500	By customer	°C	By customer	°C
Copper strip corrosion	D130	No. 3 max.		No. 3 max.	
Acid number	D664	0.80 max.	mg KOH/gm	N/A	N/A
Free glycerol	D6584	0.020 max.	% mass	N/A	N/A
Total glycerol	D6584	0.240 max.	% mass	N/A	N/A
Phosphorus content	D4951	0.001	% mass	N/A	N/A
Distillation temperature, 90% recovered	D1160	360 max.	°C	N/A	N/A
Distillation temperature, 90% recovered	D86	N/A	N/A	338 max.	°C

43 to 47 observed for petroleum diesel fuel. The cetane number is typically estimated for petroleum diesel fuel using the cetane index. However, due to the fact that this index was developed for diesel fuel and not biodiesel, this ASTM test procedure provides faulty information for biodiesel. A cetane engine must be used to determine the cetane number for biodiesel. See Table I.

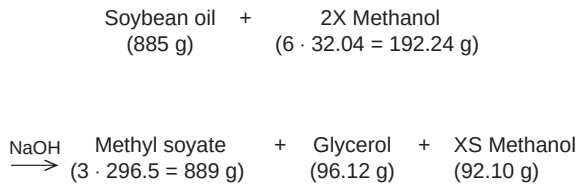


FIGURE 4 Basic transesterification reaction for soybean oil.

4. BIODIESEL PRODUCTION

4.1 Introduction

Biodiesel consists of the monoalkyl esters of fatty acids derived from vegetable oils or animal fats. It is most commonly produced through a process known as transesterification, which is a chemical reaction where an alkoxy group of an ester is exchanged with that of another alcohol to form the new ester product. The chemical reaction with methanol is shown schematically in Fig. 1.

This reaction is reversible so to force the equilibrium in the direction of the products, from 60% to 200% excess methanol is added. The reaction requires a catalyst and strong bases such as potassium hydroxide and sodium hydroxide. The actual reaction using 100% excess methanol is shown in Fig. 4.

The alkali catalysts usually result in the reaction proceeding to completion in 4 to 8 hours at ambient conditions and in 1 hour at 60°C. In general, the catalyst is dissolved in the methanol before addition to the oil to prevent direct contact between the concentrated catalyst and the oil. Since the methanol is only slightly soluble in the soybean oil, agitation is required during the early part of the reaction. The reaction proceeds through a sequence of steps involving the removal of fatty acid chains from the triglyceride to produce diglycerides, monoglycerides, and, ultimately, free glycerol. When the reaction has proceeded to the point where substantial amounts of di- and monoglycerides have been produced, agitation is less important. To further drive the equilibrium to products, the reaction is often conducted in steps. During the first step, only a portion, say 80%, of the methanol and catalyst are added. The reaction proceeds substantially to equilibrium and then the

reactants are allowed to settle so the resulting glycerol can be removed. Then, the remaining methanol and catalyst are added and the reaction continued. Removal of the glycerol forces the reaction to the product side and since the alkali catalyst is selectively attracted to the glycerol, the presence of the glycerol can limit the speed of the reaction.

The most important characteristics of the triglyceride feedstocks are low water content (preferably less than 0.2%) and low free fatty acids (less than 0.5%). High free fatty acid (FFA) feedstocks can be processed but pretreatment to remove the FFAs or to convert them to biodiesel is required. This pretreatment will be described later. Water should be excluded from the reaction because it contributes to soap formation as the fatty acid chains are stripped from the triglycerides. The soap sequesters the alkali catalyst and inhibits the separation of the glycerol from the biodiesel. Excessive soap also contributes to the formation of emulsions when water washing is used at a later stage of the production process.

Another important characteristic of the feedstock is the presence of saturated and polyunsaturated triglycerides. Excessive levels of saturated fatty acid chains can produce biodiesel with a high pour point making it difficult to use at low temperatures. High levels of polyunsaturates can provide poor oxidative stability requiring the resulting fuel to be treated with an antioxidant.

While most biodiesel is made using methanol, because of its low price (and quick conversion), other alcohols, such as ethanol and isopropanol, can also be used. Higher alcohols provide superior cold flow properties but are generally more difficult to produce, requiring higher temperatures, lower levels of water contamination, and more complex alcohol recycling due to the formation of azeotropes.

As mentioned earlier, strong alkalis such as sodium hydroxide and potassium hydroxide are common catalysts. These bases form the corresponding methoxides when dissolved in methanol. Water is also formed in this reaction and is probably responsible for some soap formation although not enough to inhibit the transesterification reaction. A more desirable option is to use sodium methoxide formed using a water-free process. This catalyst is available as a concentrated solution in methanol (25% and 30%), which is easier to use because it is a liquid. Research is underway to develop heterogeneous catalysts for biodiesel production that would minimize soap formation and provide cleaner glycerol. Since most of the catalyst ends up as a contaminant in the glycerol, either as soap or free alkali, a heterogeneous catalyst would simplify glycerol refining. Research is also being conducted to find reaction conditions that do not require a catalyst. However, these conditions are at very high temperature ($>250^{\circ}C$) and appear to produce undesirable contaminants in the biodiesel.

Figure 5 shows a schematic of a biodiesel production process. Oil enters the reactor where it is mixed with methanol and catalyst. Usually, the catalyst has been mixed with the methanol before contacting the oil to prevent direct contact of the concentrated catalyst and the oil to minimize soap formation. The reactor can be either a batch process or, as is more common with larger plants, a continuously stirred tank reactor (CSTR) or plug flow reactor. When a CSTR is used, it is common to use more than one stage to ensure complete reaction. After the reaction is complete, the glycerol is separated from the biodiesel. This separation can be accomplished with a gravity decanter or using a centrifuge. The unreacted methanol will split between the biodiesel and glycerol giving about 1 to 3% methanol in the biodiesel and 30 to 50% in the glycerol. The methanol in the biodiesel should be recovered for reuse. It may be as much as half of the excess methanol. This is usually accomplished by a vacuum flash process, but other devices such as falling film evaporator have also been used. The methanol-free biodiesel is then washed with water to remove residual methanol, catalyst, soap, and free glycerol. If the biodiesel contains excessive soap, this washing process can be problematic, as the soap will cause an emulsion to form between the water and the biodiesel. To minimize the formation of emulsions, a strong acid is sometimes added to the biodiesel to split the soap into free fatty acids (FFA) and salt. Without the soap, water consumption is greatly reduced and the salt, methanol, catalyst, and glycerol are removed with as little as 3 to 10% water. The water should be heated to $60^{\circ}C$ to assist in the removal of free glycerol and should be softened to minimize the transfer of calcium and magnesium salts to the biodiesel. The final step in the process is to heat the biodiesel to remove water that may be dissolved in the biodiesel or entrained as small droplets. This is accomplished with a flash process.

Also shown in the diagram is the preliminary processing of the co-product glycerol. This glycerol contains virtually all of the catalyst and a considerable amount of soap and unreacted methanol. Usually the glycerol will be acidulated to split the soaps into FFA and salt. The FFAs are not soluble in the glycerol and rise to the top where they can be

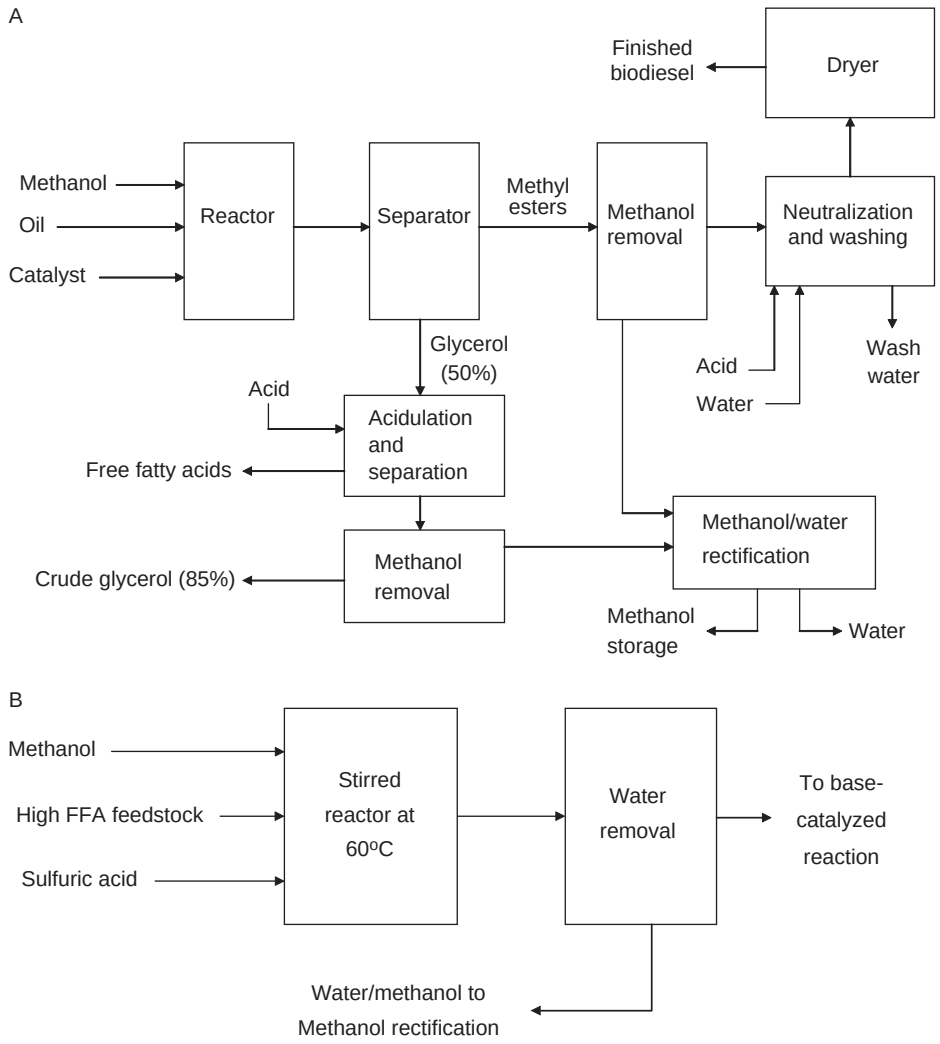


FIGURE 5 (A) Process flow schematic for biodiesel production. (B) Acid-catalyzed pretreatment process.

decanted and returned to the biodiesel process after pretreatment. After FFA removal, the methanol is removed by a flash process or by a thin-film evaporator leaving a crude glycerol product that is 80 to 90% pure. The balance will be salts, residual FFAs, water, and phosphotides, color-bodies, and other contaminants from the oil.

Although water is not deliberately added to the process until after the methanol has been removed, small amounts of water will enter the system as a contaminant in the oil, alcohol, and catalyst. This water will tend to accumulate in the methanol, so before it can be returned to the process, the methanol should undergo fractional distillation.

Biodiesel plants can use either batch or continuous flow processing. Batch processing is most common in small plants of less than 4 million liters/year. Batch processing provides the ability to modify the process for variations in feedstock quality. Continuous flow requires greater uniformity in the feedstock quality, generally requires 24 hour operation, 7 days per week, increases labor costs, and is most suitable for larger operations of greater than 40 million liters/year.

4.2 Pretreatment

Small amounts of FFAs can be tolerated by adding enough alkali catalyst to neutralize the FFAs while still leaving enough to catalyze the reaction. When an oil or fat has more than 3 to 5% FFA, the amount of soap formed with the use of an alkali catalyst will be large enough that separation of the glycerol from the oil may not be possible. In addition, when excess

alkali is added, the FFAs are lost and not available for conversion to biodiesel.

An alternative is to convert the FFAs to methyl esters using an acid catalyst such as sulfuric acid with a large excess of methanol (>20:1 molar ratio based on the FFA content) (Fig. 5B). The acid-catalyzed reaction is relatively fast (1 hour at 60°C), converting the FFAs to methyl esters with water produced as a by-product. The water eventually stops the reaction before all of the FFAs have been converted and must be removed from the system, either by decanting with the methanol or by vaporization. After the FFAs have been converted to methyl esters, an alkali catalyst can be used to convert the triglycerides. Acid catalysis can be used to complete the transesterification reaction, but the time required is prohibitive.

4.3 Product Quality

Modern diesel engines require high-quality fuels. The fuel injection system, which is often the most expensive element of the engine, can be damaged by fuel contaminants. Water and solid particles are the largest problem.

The contaminants most frequently found in biodiesel are the products of incomplete reaction and residual alcohol, catalyst, and free glycerol. Incompletely reacted biodiesel will contain monoglycerides, diglycerides, and triglycerides. These compounds are usually detected using a gas chromatograph and then the glycerol portion is summed to yield a total glycerol quantity for the fuel. ASTM standards require that the total glycerol be less than 0.24%. This means that more than 98% of the original glycerol portion of the triglycerides feedstock must be removed. Excessive amounts of monoglycerides, especially for saturated compounds, may precipitate from the fuel and plug fuel filters.

If the biodiesel is not washed with water, it may contain some unreacted alcohol. The amount will usually be small enough that it does not adversely affect the operation of the engine, but it can lower the flash point of the fuel to where it must be considered flammable and accorded the same safety requirements as gasoline. The residual catalyst can cause excessive ash formation in the engine. Free glycerol can separate from the fuel and collect in the bottom of storage tanks. This glycerol layer can extract mono- and diglycerides from the biodiesel and produce a sludge layer that may plug filters and small passages in the fuel system.

5. ADVANTAGES OF BIODIESEL

Table II shows the changes that were observed in the regulated exhaust emissions of three diesel engines that were tested to produce emissions characterization data for the U.S. Environmental Protection Agency's Fuels and Fuel Additives registration program. The reductions in unburned hydrocarbons (HC) and carbon monoxide (CO) are dramatic although these specific pollutants are not generally a concern with diesel engines. The particulate matter (PM) reductions are also quite striking. Oxides of nitrogen (NO_x) were found to increase with the use of biodiesel. The increase varied depending on the engine tested, but it is clear that biodiesel may produce a NO_x increase of 5 to 13%. The reasons for this increase are still under investigation but appear to be a combination of several effects, including biodiesel's higher speed of sound and isentropic bulk modulus and the tendency of many engine fuel injection systems to advance the injection timing when greater volumes of fuel are injected. Due to biodiesel's lower energy content, a typical test protocol may demand a higher fuel flow rate when biodiesel is used, causing an inadvertent timing advance and resulting NO_x increase.

A comprehensive Life-Cycle Inventory of biodiesel conducted by the National Renewable Energy Laboratory showed that biodiesel provided 3.2 units of fuel energy for every unit of fossil energy consumed in its life cycle. Further, although some fossil-based CO_2 is released during biodiesel production, mainly from the methanol consumed, the net production of CO_2 is reduced by 78%.

6. DISADVANTAGES

6.1 Economics

One of the largest factors preventing the adoption of biodiesel is cost. The feedstock costs for biodiesel

TABLE II

Changes in Regulated Emissions with Biodiesel

Engine	HC	CO	PM	NO_x
Cummins N-14	−95.6	−45.3	−28.3	+13.1
DDC S-50	−83.3	−38.3	−49.0	+11.3
Cummins B5.9	−74.2	−38.0	−36.7	+4.3

Derived from Sharp *et al.*

tend to be high in comparison to the cost of petroleum diesel fuel. The end result is that the cost of biodiesel is higher than that of petroleum diesel fuel. In addition, transportation costs are significantly greater for biodiesel due to the fact that the transportation infrastructure for biodiesel is in its infancy. The costs associated with the production and transportation of biodiesel fuels are discussed in more detail in the economics section of this article.

6.2 NO_x and Other Exhaust Emissions

Biodiesel produces more NO_x emissions than diesel fuel. If B100 is used, NO_x production may be increased by 13%. If a B20 blend is used, NO_x production is only increased by 2%, and the engine will typically satisfy the EPA engine exhaust emissions requirements under the Clean Air Act. To meet the EPA emissions requirements in 2006, engine manufacturers will likely use exhaust after-treatment technology that will reduce NO_x emissions. The low sulfur levels in biodiesel fuels make them a good candidate for use with the exhaust after-treatment technologies that are available. Even though biodiesel fuels produce more NO_x emissions, they have been shown to reduce carbon monoxide, particulate matter, unburned hydrocarbons, and other pollutants.

6.3 Fuel Quality

Many problems associated with biodiesel stem from poor fuel quality from the supplier. Most often this is related to the completeness of the production reaction. The ASTM has developed a quality standard for biodiesel. At this point in time, fuel manufacturer compliance with the standard is voluntary. Generally, it is a good idea to ensure that the biodiesel manufacturer sells biodiesel that meets or exceeds the ASTM specifications.

6.4 Energy Content

Biodiesel fuels contain about 12.5% less energy per unit of weight than petroleum diesel fuel (37,215 kJ/kg vs. 42,565 kJ/kg). However, since biodiesel has a higher density, the energy content per unit of volume is only 8% less. As a result, the fuel economy of the diesel engine that is powered with biodiesel tends to be slightly less than when powered with petroleum diesel fuel.

6.5 Cold Weather

The cloud point and cold filter plugging point are much higher for biodiesel than diesel fuel. This means that the fuel will not work in the engine as well as diesel fuel at lower temperatures. Additives can be used to reduce the cloud point and CFPP of biodiesel fuel. The cold-weather properties of biodiesel can also be improved by using a lower level blend of biodiesel fuel (i.e., B5 instead of B20). Additionally, the fuel may be blended with number 1 diesel instead of number 2 diesel to improve the cold-weather properties.

6.6 Material Compatibility

Biodiesel fuel will react with some plastics and some metals in a negative manner. The plastics that seem to be compatible with biodiesel include nylon, teflon, and viton. When in contact with nonferrous metals, such as copper and zinc, biodiesel fuel can cause precipitates to form. Some of these materials can be found in fuel tank liners, fuel lines, transfer pump diaphragms, injector seals, and injection pump seals (among others).

6.7 Solvency

Biodiesel can also act as a solvent. This creates some problems when used in existing systems. The biodiesel can dissolve existing residues in fuel tanks and lines and carry them to the fuel system. Generally, after a few tanks of fuel have been used, these problems tend to be reduced.

6.8 Stability

The oxidative stability of biodiesel fuel is a major factor in determining the allowable storage time for biodiesel fuel. The iodine number can be used to estimate the oxidative stability of the fuel before any stabilizers are added. Typically biodiesel fuels can be stored for up to 6 months without problems. If biodiesel fuel needs to be stored longer, antioxidants can be added to the fuel to improve the stability. If the fuel is not stabilized, biodiesel can form gums and sediments that clog filters or form deposits on fuel system components, including fuel pumps and injectors.

Additionally, as with diesel fuel, some climatic conditions promote biological growth (such as algae), in the fuel. If this occurs, the problem can be treated with a biocide. Reduction in water

contamination also reduces the amount of biological growth in the fuel since the algae grows on the water.

6.9 Warranties

Most engine manufacturers do not warranty their engines for use with a specific fuel. Consumers with engine problems that can be traced to the fuel are directed to their fuel supplier. Many engine manufacturers have developed policy statements for biodiesel that allow the use of up to 5% biodiesel but indicate that more experience is needed before fueling with higher level blends. Most require that biodiesel meet the ASTM standard. The best practice is to check with the engine and vehicle manufacturer before using biodiesel.

7. BIODIESEL STORAGE AND USE

7.1 Blending

Blending of biodiesel is not recommended if the temperature of either fuel is below 4.4°C. Low temperatures impact how easily the biodiesel mixes with petroleum diesel fuel. In most situations, splash blending works effectively (i.e., splashing or pouring the biodiesel into the diesel fuel), as biodiesel mixes readily with petroleum diesel fuel. Once mixed, the biodiesel tends to remain blended.

If splash blending biodiesel in a large tank, the biodiesel should be introduced after the diesel fuel has been placed in the tank or the blend should be prepared before placing the blended fuel into storage. This is due to the fact that the biodiesel is heavier than diesel fuel and will essentially rest at the bottom of the tank until some type of agitation is provided. Bulk plants or terminals may also use pumps but will most likely rely on electronic injection or in-line blending to prepare the required blend.

7.2 Transportation

Biodiesel, due to its high flash point, is *not* considered flammable. The fuel is considered combustible, just as is vegetable oil (feedstock). As such, the transportation of "neat" biodiesel may be handled in the same manner as vegetable oil (Code of Federal Regulations 49 CFR 171-173). This is not the case for a low-level blend or a B20 blend. These blends exhibit flash point tendencies that essentially mirror diesel fuel. Blends of biodiesel should be handled in the same manner as petroleum diesel fuel.

7.3 Storage Tanks

Storage tanks for biodiesel can be constructed from mild steel, stainless steel, fluorinated polyethylene, fluorinated polypropylene and teflon. Biodiesel, like petroleum diesel fuel, should be stored in a clean, dry, dark environment. In the event that the container selected is made from polyethylene or polypropylene, the container should be protected from sunlight.

Some authors suggest that aluminum is suitable for use as a storage tank. However, nonferrous metals, such as aluminum, tend to react unfavorably with biodiesel by shortening the shelf life of the fuel. Much is the same for tin and zinc. Concrete lined tanks, varnish lined tanks, or tanks lined with PVC cannot be used to store biodiesel. Biodiesel reacts with each of these products, breaking down the chemical structure of each.

As with any fuel, steps must be taken to prevent water from entering the tank. Algae can grow in biodiesel just as it does with petroleum diesel fuel.

Measures should be taken to ensure that the biodiesel will flow in cold weather. This is often accomplished by mixing the fuel with either number 1 or number 2 diesel fuel. Cold flow improvers (CFI) can also be added to enhance the cold flow characteristics of biodiesel.

7.4 Material Compatibility

Essentially the same materials that are used to construct a biodiesel storage tank can be used with biodiesel (stainless steel, mild steel, viton, some forms of teflon, and fluorinated polyethylene/polypropylene). Rubber elastomers cannot be used, as pure biodiesel will dissolve the rubber. The effect is lessened with lower percentage blends, but little research has been conducted to determine the long-term material compatibility of biodiesel blends.

7.5 Safety

Biodiesel is nontoxic, biodegradable, and much less irritating to the skin than petroleum diesel. However, the same safety rules that pertain to petroleum diesel fuel also apply to the use of biodiesel. The following list summarizes several of these issues:

- Store in closed, vented containers between 10°C and 50°C.
- Keep away from oxidizing agents, excessive heat, and ignition sources.
- Store, fill, and use in well-ventilated areas.

- Do not store or use near heat, sparks, or flames; store out of the sun.
- Do not puncture, drag, or slide the storage tank.
- A drum is not a pressure vessel; never use pressure to empty.
- Wear appropriate eye protection when filling the storage tank.
- Provide crash protection (i.e., large concrete-filled pipe near the storage tank).

8. BIODIESEL ECONOMIC CONSIDERATIONS

The cost of biodiesel consists of five major components: the cost of the feedstock, the cost of the biodiesel, the price of glycerol by-product, and availability of biodiesel. We will examine each of these components.

8.1 Cost of the Feedstock

The cost of the feedstock (oil/tallow/lard, etc.) will vary from one season to the next. The cost will also vary from one crop to the next (soybeans versus rapeseed) as well as from one year to the next, depending on supply and demand. For example, soybean oil has fluctuated from 14.15 cents/pound to 25.8 cents/pound during 1995–2002. Peanut oil has varied from 27 to 49 cents/pound during the same time period. At the same time, lard has sold for 12.5 to 23.02 cents/pound. During 2001–2002, the price of cottonseed oil varied from 14.4 to 22.3 cents per pound and corn oil varied from 11.38 to 23.02 cents/pound. During 2001–2002, the high for corn was in December and the low was in May. The high for cottonseed oil occurred in September and the low occurred in October. Some biodiesel producers have been able to purchase used restaurant grease for as little as 3 to 4 cents/pound (McDonalds, etc.), while some have been paid to haul the waste oil away (transportation was the only feedstock cost).

8.2 Cost of Biodiesel

Economies of scale are a factor when buying biodiesel. Since biodiesel is not yet available on the pipeline network, most deliveries are made by truck or rail. Transportation costs can be a significant portion of the product cost.

The most significant contributor to the cost of the fuel is the cost of the oil itself. As noted in the feedstock section, the price of the oil has varied from as little as 3 to 4 cents/pound to a high of 25.8 cents/pound. If soybeans were used as the feedstock, the cost could range from $1.73 to $3.10/gallon. If a less expensive feedstock were used, the price range would even be greater anywhere from $0.50 to $63.10/gallon. (See Fig. 6.)

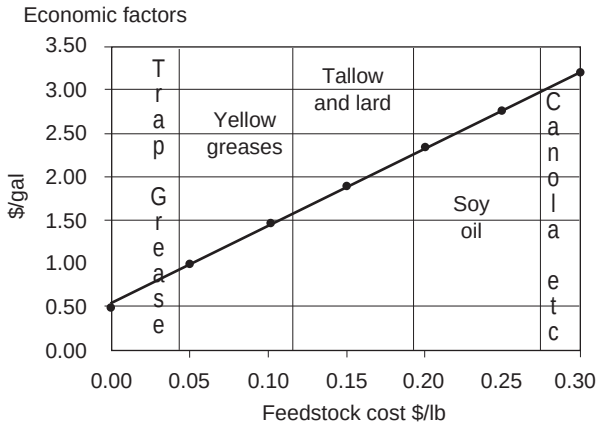


FIGURE 6 Production cost per gallon. Reprinted the from National Renewable Energy Laboratory, Kansas Cooperative Development Center/KSU.

The next highest cost when producing biodiesel is the cost to convert the feedstock from a pure oil to a transesterified fuel. Some researchers report that this amounts to approximately 30% of the total cost (as nearly 70% of the cost is tied up in the raw materials [soybean oil feedstock, methanol, catalyst]).

The conversion costs for biodiesel can range from $0.30 to $0.60/gallon. One study reported that it cost $0.58/gallon for transesterification and $0.33/gallon for overhead. A credit of $0.39/gallon was applied for the resale of the glycerol, bringing the total for transesterification to $0.52/gallon.

8.3 Price of Glycerol

As noted in the previous paragraph, the glycerol is a valuable by-product that must be disposed of after transesterification (credit = $0.39/gallon). Glycerol is used in pharmaceuticals, cosmetics, toothpaste, paints, and other products. Most researchers believe that a flood of glycerol on the market (which would result from increased biodiesel production) would reduce the net credit. Alternative uses for the glycerol would develop and these uses should help maintain a solid price for the glycerol. Presently, refined glycerol prices stand in the United States at approximately $0.65/pound.

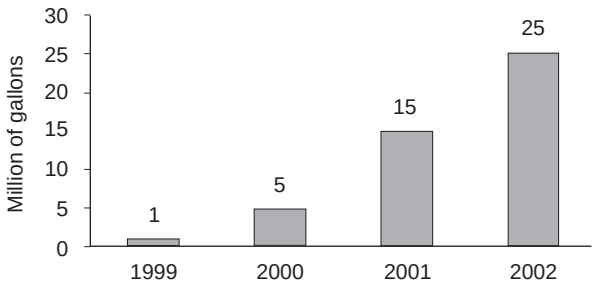


FIGURE 7 U.S. biodiesel production. Reprinted from the National Biodiesel Board, Kansas Cooperative Development Center/KSU.

8.4 Availability

Most economists are of the opinion that biodiesel has the capability to replace about 2 to 5% of the diesel fuel used for transportation in the United States. Although this may seem small in comparison to the volume of diesel fuel used, this would indicate that we would need to produce roughly 6 billion gallons of biodiesel each year. However, Brazil has become a major exporter in the soybean market, and large carryover stocks of soybeans have resulted. The estimated soybean oil carryover for the year 2002 was 1.2 million metric tons. The United States produces 25 million gallons of biodiesel a year. (See Fig. 7.)

SEE ALSO THE FOLLOWING ARTICLES

Alternative Transportation Fuels: Contemporary Case Studies • Biomass, Chemicals from • Biomass for Renewable Energy and Fuels • Ethanol Fuel • Fuel Cycle Analysis of Conventional and Alternative Fuel Vehicles • Internal Combustion (Gasoline and Diesel) Engines • Life Cycle Analysis of Power Generation Systems • Renewable Energy, Taxonomic Overview

Further Reading

Ash, M., and Dohlman, E. (2003). Bumper South American soybean crop buoys world consumption growth. *In* "Oil Crops Outlook," a USDA report in February of 2003. Found at www.ers.usda.gov.

"Biodiesel: Fueling Illinois." (2003). Illinois Soybean Association. Bloomington, IL. Found at www.ilsoy.org.

Coltrain, J. (2002). Biodiesel: Is it worth considering? Kansas State University. Found at www.agecon.ksu.edu.

National Renewable Energy Laboratory (2003). Kansas Cooperative Development Center. Department of Agricultural Economics. Kansas State University. Manhattan, KS 66506.

Sharp, C. A., S.A. Howell, and J. Jobe (2000).The effect of biodiesel fuels on transient emissions from modern diesel engines, part I: Regulated emissions and performance. Society of Automotive Engineers Paper 2000-01-1967.

Sheehan, J., Carnobreco, V., Duffield, J., Graboski, M., and Shapouri, H. (1998). Urban bus operation. *In* "Life Cycle Inventory of Biodiesel and Petroleum Diesel for Use in an Urban Bus," pp. 171–188. Found at www.nrel.gov.

Sigmon, M. (1997). Demonstrations and economic analysis of biodiesel in fleet vehicle and marine applications. For the City of Chicago and Department of Energy. Found at www.biodiesel.org.

"Standard Specification for Diesel Fuel Oils." (2002). ASTM International. West Conshohocken, PA. Found at www.astm.org.

Biomass, Chemicals from

DOUGLAS C. ELLIOTT
Pacific Northwest National Laboratory
Richland, Washington, United States

1. **Historical Developments**
2. **The "Biorefinery" Concept**
3. **Biomass-Derived Chemical Products**
4. **The Future Biorefinery**

Glossary

biomass Material generated by living organisms; typically the cellular structure of a plant or animal or chemical products generated by the organism.

biorefinery A factory incorporating a number of processing steps, including pretreatments, separations, and catalytic and biochemical transformations, for the production of chemical and fuel products from biomass.

carbohydrate Chemicals composed of carbon, hydrogen, and oxygen and typically having a hydrogen-to-oxygen ratio of 2:1. Representatives include sugars, such as glucose or xylose, starches, and cellulose or hemicelluloses.

catalyst An additive to a chemical reaction that increases the rate of reaction without being consumed in the reaction.

chemurgy The philosophy of using agriculturally derived feedstocks for production of chemical products.

fermentation The production of simple chemical products by microbial action on a complex feedstock, often a carbohydrate.

hydrogenation The chemical addition of hydrogen into a molecule, usually performed in the presence of a metal catalyst. Special types of hydrogenation include hydrodeoxygenation, in which oxygen is removed from a molecule by the reaction of hydrogen, and hydrogenolysis, in which a molecule is broken into smaller molecules by the reaction of hydrogen.

hydrolysis A chemical reaction in which water is chemically combined into a reactant, and in the process the reactant is broken down into smaller molecules.

Refined processing methods and advanced technologies for biomass conversion enable low-value byproducts or waste materials to be transformed into value-added products. Such products include the same chemicals produced from petroleum that are used, for example, as plastics for car components, food additives, clothing fibers, polymers, paints, and other industrial and consumer products. Other replacement products that can be produced from biomass include different chemicals with similar or better properties. Currently, 17% of the volume of the products derived from petroleum in the United States is chemicals. If that petroleum, used for chemical synthesis, were displaced by biomass, it would thus be available for the energy market. In 1998, the U.S. Department of Energy stated as its goal the production of at least 10% of the basic chemical building blocks from plant-derived renewables by 2020 (a fivefold increase) and 50% by 2050 (a subsequent fivefold increase). The chemicals and products discussed here are capable of reducing the 17% petroleum utilization number and making a significant contribution toward meeting those goals.

The discussion on chemicals from biomass is limited to large-scale commodity products for which a noticeable impact on the energy market would be made by displacing petroleum. Specialty chemical products whose limited market would have minimal impact on energy markets are not addressed, nor is the use of biomass for material products such as cellulosic fiber for paper or lignocellulosic material for construction materials. Using biomass to produce a vast array of small-market specialty chemicals based on unique biochemical structures is another part of the overall concept of a sustainable economy but is outside the scope of this article. Instead, it focuses on the chemical structure of various commodity products and their fabrication from biomass. A history of the use of biomass as the basis for chemical products is included for background.

1. HISTORICAL DEVELOPMENTS

Biomass has been used to meet the needs of civilization since prehistory. Several biomass chemical

 © 2004 Elsevier Inc. All rights reserved.

products can be traced to a time when neither the chemical mechanistic transformation process nor the chemical identity of the product were understood. These early products include chemicals still widely used, such as ethanol and acetic acid. A more scientific understanding of chemistry in the 19th century led to the use of biomass as a feedstock for chemical processes. However, in the 20th century, the more easily processed and readily available petroleum eventually changed the emphasis of chemical process development from carbohydrate feedstock to hydrocarbon feedstock. Through major investments by industry and government, utilization of petroleum as a chemical production feedstock has been developed to a high level and provides a wide array of materials and products. Now, the expansion of synthetic chemical products from petroleum provides the framework in which to reconsider the use of renewable biomass resources for meeting society's needs in the 21st century. However, only since the 1980s has interest returned to the use of biomass and its chemical structures for the production of chemical products; therefore, much research and development of chemical reactions and engineered processing systems remains to be done (see Table I)

1.1 Early Developments

The earliest biomass uses often involved food products. One of the first was ethanol, which was limited to beverage use for centuries. Ethanol's origin is lost in antiquity, but it is generally thought to be an accidental development. Only with an improved understanding of chemistry in the mid-1800s was ethanol identified and its use expanded for solvent applications and as a fuel. Acetic acid, the important component in vinegar, is another accidental product based on fermentation of ethanol by oxidative organisms. A third fermentation product from antiquity, lactic acid, also an accidental product of oxidative fermentation, is commonly used in sauerkraut, yogurt, buttermilk, and sourdough.

Another old-world fermentation method is the production of methane from biomass, a multistep process accomplished by a consortium of bacteria, which is an important part of the cyclical nature of carbon on Earth. These processes, now generally referred to as anaerobic digestion, are well-known in ruminant animals as well as in stagnant wet biomass. The methane from these processes has only been captured at the industrial scale for use as a fuel since the 1970s.

Additional early uses of biomass involved its structural components. The fibrous matrix (wood) was first used in constructing shelter. Later, the fibrous component was separated for paper products. As a result, two major industries, the wood products industry and the pulp and paper industry, were developed at sites where biomass was available. Biomass also comprised materials used for clothing. Cotton fibers, consisting mainly of cellulose, and flax fiber are examples.

At the beginning of the 20th century, cellulose recovered from biomass was processed into chemical products. Cellulose nitrate was an early synthetic polymer used for "celluloid" molded plastic items and photographic film. The nitrate film was later displaced by cellulose acetate film, a less flammable and more flexible polymer. Cellulose nitrate lacquers

TABLE I

Petroleum-Derived Products in the United States 2001[a]

	Net production (thousands of barrels per day)	Percentage of net production
Finished motor gasoline(total)	8560	44.6
Finished aviation gasline(total)	22	0.1
Jet fuel	1655	8.6
Kerosene	72	0.4
Distillate fuel oil	3847	20.0
Residual fuel oil	811	4.2
Asphalt and road oil	519	2.7
Petroleum coke	437	2.3
Chemical products	3283	17.1
Liquefied petroleum gases (net[b])	1778	
Ethane/ethylene	695	
Propane/Propylene	1142	
Butane/Butylene	−5	
Isobutane/Isobutylene	−28	
Pentanes plus	−32	
Naphtha for petrochemicals	258	
Other oils for petrochemicals	312	
Special naphthas	41	
Lubricants	153	
Waxes	18	
Still gases	670	
Miscellaneous products	59	
Total net production	19,206	

[a] Source: Energy Information Administration, Petroleum Supply Annual 2001, Vol. 1, Table 3.

[b] Net, products minus natural gas liquids input.

retained their market, although acetate lacquers were also developed. Cellulose acetate fiber, called Celanese, grew to one-third of the synthetic market in the United States by 1940. Injection-molded cellulose acetate articles became a significant component of the young automobile industry, with 20 million pounds produced in 1939. Today, cellulose-based films and fibers have largely been supplanted by petroleum-based polymers, which are more easily manipulated to produce properties of interest, such as strength, stretch recovery, or permeability.

Wood processing to chemicals by thermal methods (pyrolysis), another practice from antiquity, was originally the method used for methanol and acetic acid recovery but has been replaced by petrochemical processes. One of the earliest synthetic resins, the phenol-formaldehyde polymer Bakelite, was originally produced from phenolic wood extractives. However, petrochemical sources for the starting materials displaced the renewable sources after World War I. More chemically complex products, such as terpenes and rosin chemicals, are still recovered from wood processing but only in a few cases as a by-product from pulp and paper production.

1.2 Chemurgy

In the early 20th century, the chemurgy movement blossomed in response to the rise in the petrochemical industry and in support of agricultural markets. The thrust of the movement was that agricultural products and food by-products could be used to produce the chemicals important in modern society. A renewable resource, such as biomass, could be harvested and transformed into products without depleting the resource base. Important industrialists, including Henry Ford, were participants in this movement, which espoused that anything made from hydrocarbons can be made from carbohydrates.

Supply inconsistencies and processing costs eventually turned the issue in favor of the petrochemical industry. The major investments for chemical process development based on hydrocarbons were more systematically coordinated within the more tightly held petrochemical industry, whereas the agricultural interests were too diffuse (both geographically and in terms of ownership) to marshal much of an effort. Furthermore, the chemical properties of petroleum, essentially volatile hydrocarbon liquids, allowed it to be more readily processed to chemical products. The large scale of petroleum refining operations, developed to support fuel markets, made for more efficient processing systems for chemical production as well.

2. *THE "BIOREFINERY" CONCEPT*

The vision of a new chemurgy-based economy began evolving in the 1980s. The biorefinery vision foresees a processing factory that separates the biomass into component streams and then transforms the component streams into an array of products using appropriate catalytic or biochemical processes, all performed at a scale sufficient to take advantage of processing efficiencies and resulting improved process economics. In essence, this concept is the modern petroleum refinery modified to accept and process a different feedstock and make different products using some of the same unit operations along with some new unit operations. The goal is to manufacture a product slate, based on the best use of the feedstock components, that represents higher value products with no waste. Overall process economics are improved by production of higher value products that can compete with petrochemicals. Some biomass processing plants already incorporate this concept to various degrees.

2.1 Wood Pulp Mill

The modern wood pulp mill represents the optimized, fully integrated processing of wood chips to products. The main process of pulping the wood allows the separation of the cellulosic fiber from the balance of the wood structure. The cellulose has numerous uses, ranging from various paper and cardboard types to chemical pulp that is further processed to cellulose-based chemicals. The lignin and hemicellulose portions of the wood are mainly converted to electrical power and steam in the recovery boiler. In some cases, the lignin is recovered as precipitated sulfonate salts, which are sold based on surfactant properties. In other cases, some lignin product is processed to vanillin favoring. The terpene chemicals can be recovered, a practice limited primarily to softwood processing plants in which the terpene fraction is larger.

2.2 Corn Wet Mill

In the corn wet mill, several treatment and separation processes are used to produce a range of food and chemical products. Initial treatment of the dry corn kernel in warm water with sulfur dioxide results in a softened and protein-stripped kernel and a protein-rich liquor called corn steep liquor. The corn is then milled and the germ separated. The germ is pressed and washed with solvent to recover the corn oil. Further milling and filtering separate the cornstarch

from the corn fiber (essentially the seed coat). The starch is the major product and can be refined to a number of grades of starch product. It can also be hydrolyzed to glucose. The glucose, in turn, has various uses, including isomerization to fructose for high-fructose corn syrup; hydrogenation to sorbitol; or fermentation (using some of the steep liquor for nutrients) to a number of products, such as ethanol, lactic acid, or citric acid. The balance of the corn steep liquor is mixed with the corn fiber and sold as animal feed.

2.3 Cheese Manufacturing Plant

Cheese manufacturing has evolved into a more efficient and self-contained processing operation in which recovering the whey by-product (as dried solids) is the industry standard instead of disposing of it, which was the practice for many centuries. The whey is a mostly water solution of proteins, minerals, and lactose from the milk following coagulation and separation of the cheese. To make the whey more valuable, ultrafiltration technology is used in some plants to separate the protein from the whey for recovery as a nutritional supplement. Further processing of the ultrafiltration permeate is also used in a few plants to recover the lactose for use as a food ingredient. Developing methods for recovering the mineral components is the next stage in the development process.

3. BIOMASS-DERIVED CHEMICAL PRODUCTS

For production of chemicals from biomass, there are several processing method categories, which are used based on the product desired: fermentation of sugars to alcohol or acid products; chemical processing of carbohydrates by hydrolysis, hydrogenation, or oxidation; pyrolysis of biomass to structural fragment chemicals; or gasification by partial oxidation or steam reforming to a synthesis gas for final product formation. Various combinations of these processing steps can be integrated in the biorefinery concept. Table II illustrates the wide range of biomass-derived chemicals. These products include the older, well-known fermentation products as well as new products whose production methods are being developed. Development of new processing technology is another key factor to making these products competitive in the marketplace.

3.1 Modern Fermentation Products

Fermentation products include ethanol, lactic acid, acetone/butanol/ethanol, and citric acid. Ethanol and lactic acid are produced on a scale sufficient to have an impact on energy markets. Ethanol is primarily used as a fuel, although it has numerous potential derivative products as described later. Lactic acid is new as an industrial chemical product; the first lactate derivative product, poly lactic acid, came on stream in 2002.

3.1.1 Ethanol

In 1999, ethanol was produced at 55 plants throughout the United States at a rate of 1.8 billion gallons per year. The feedstock for the fermentation ranges from corn starch-derived glucose, wheat gluten production by-product starches, and cheese whey lactose to food processing wastes and, until recently, wood pulping by-products. Although some of this ethanol is potable and goes into the food market, the vast majority is destined for the fuel market for blending with gasoline. However, other possible chemical products derived from ethanol include acetaldehyde, acetic acid, butadiene, ethylene, and various esters, some of which were used in the past and may again become economically viable.

3.1.2 Lactic Acid

Lactic acid, as currently utilized, is an intermediate-volume specialty chemical used in the food industry for producing emulsifying agents and as a nonvolatile, odor-free acidulant, with a world production of approximately 100 million pounds per year in 1995. Its production by carbohydrate fermentation involves any of several *Lactobacillus* strains. The fermentation can produce a desired stereoisomer, unlike the chemical process practiced in some areas of the world. Typical fermentation produces a 10% concentration lactic acid broth, which is kept neutral by addition of calcium carbonate throughout the 4- to 6-day fermentation. Using the fermentation-derived lactic acid for chemical products requires complicated separation and purification, including the reacidification of the calcium lactate salt and resulting disposal of a calcium sulfate sludge by-product. An electrodialysis membrane separation method, in the development stage, may lead to an economical process. The lactic acid can be recovered as lactide by removing the water, which results in the dimerization of the lactic acid to lactide that can then be recovered by distillation.

TABLE II
Chemicals Derived from Biomass

Chemical	Derivative chemicals	Uses
Fermentation products		
Ethanol		Fuel
Lactic acid	Poly lactic acid	Plastics
Acetone/butanol/ethanol		Solvents
Citric acid		Food ingredient
Carbohydrate chemical derivatives		
Hydrolysate sugars	Fermentation feedstocks	Fermentation products
Furfural, hydroxymethylfurfural	Furans, adiponitrile	Solvents, binders
Levulinic acid	Methyltetrahydrofuran, δ-amino lactic acid, succinic acid	Solvents, plastics, herbicide/pesticide
Polyols	Glycols, glycerol	Plastics, formulations
Gluconic/glucaric acids		Plastics
Pyrolysis products		
Chemicals fractions	Phenolics, cyclic ketones	Resins, solvents
Levoglucosan, levoglucosenone		Polymers
Aromatic hydrocarbons	Benzene, toluene, xylenes	Fuel, solvents
Gasification products		
Synthesis gas	Methanol, ammonia	Liquid fuels, fertilizer
Tar chemicals		Fuels
Development fermentations		
2,3-Butanediol/ethanol		Solvents
Propionic acid		Food preservative
Glycerol, 1,3-propanediol		C3 plastics, formulations
3-Dehydroshikimic acid	Vanillin, catechol, adipic acid	Flavors, plastics
Catalytic/bioprocessing		
Succinic acid	Butanediol, tetrahydrofuran	Resins, solvents
Itaconic acid	Methyl-1,4-butanediol and -tetrahydrofuran (or methyltetrahydrofuran)	Resins, solvents
Glutamate, lysine	Pentanediol, pentadiamine	C5 plastics
Plant-derived		
Oleochemicals	Methyl esters, epoxides	Fuel, solvents, binders
Polyhydroxyalkanoates		Medical devices

Poly lactic acid (PLA) is the homopolymer of biomass-derived lactic acid, usually produced by a ring-opening polymerization reaction from its lactide form. It is the first generic fiber designated by the Federal Trade Commission in the 21st century and the first renewable chemical product commercially produced on an industrial scale. Cargill–Dow, which produces PLA, opened its plant rated at 300 million pounds per year capacity in April 2002.

Other chemical products that can be derived from lactic acid include lactate esters, propylene glycol, and acrylates. The esters are described as nonvolatile, nontoxic, and biodegradable, with important solvent properties. Propylene glycol can be produced by a simple catalytic hydrogenation of the lactic acid molecule. Propylene glycol is a commodity chemical with a 1 billion pound per year market and can be used in polymer synthesis and as a low-toxicity antifreeze and deicer. Dehydration of lactate to acrylate appears to be a straightforward chemical process, but a high-yield process has yet to be commercialized.

3.1.3 Acetone/Butanol/Ethanol

Acetone/butanol/ethanol fermentation was widely practiced in the early 20th century before being displaced by petrochemical operations. The two-step bacterial fermentation is carried out by various species of the *Clostridium* genus. The solvent product ratio is typically 3:6:1, but due to its

toxicity, the butanol product is the limiting component. The yield is only 37%, and final concentrations are approximately 2%. The high energy cost for recovery by distillation of the chemical products from such a dilute broth is a major economic drawback. Alternative recovery methods, such as selective membranes, have not been demonstrated.

3.1.4 Citric Acid

More than 1 billion pounds per year of citric acid is produced worldwide. The process is a fungal fermentation with *Aspergillus niger* species using a submerged fermentation. Approximately one-third of the production is in the United States, and most of it goes into beverage products. Chemical uses will depend on the development of appropriate processes to produce derivative products.

3.2 Chemical Processing of Carbohydrates

3.2.1 Hydrolysate Sugars

Monosaccharide sugars can be derived from biomass and used as final products or further processed by catalytic or biological processes to final value-added products. Disaccharides, such as sucrose, can be recovered from plants and used as common table sugar or "inverted" (hydrolyzed) to a mixture of glucose and fructose. Starch components in biomass can readily be hydrolyzed to glucose by chemical processes (typically acid-catalyzed) or more conventionally by enzymatic processes. Inulin is a similar polysaccharide recovered from sugar beets that could serve as a source of fructose. More complex carbohydrates, such as cellulose or even hemicellulose, can also be hydrolyzed to monosaccharides. Acid-catalyzed hydrolysis has been studied extensively and is used in some countries to produce glucose from wood. The processing parameters of acidity, temperature, and residence time can be controlled to fractionate biomass polysaccharides into component streams of primarily five- and six-carbon sugars based on the relative stabilities of the polysaccharides. However, fractionation to levels of selectivity sufficient for commercialization has not been accomplished.

The development of economical enzymatic processes for glucose production from cellulose could provide a tremendous boost for chemicals production from biomass. The even more complex hydrolysis of hemicellulose for monosaccharide production provides a potentially larger opportunity because there are fewer alternative uses for hemicellulose than for cellulose.

3.2.2 Furfural, Hydroxymethylfurfural, and Derived Products

More severe hydrolysis of carbohydrates can lead to dehydrated sugar products, such as furfural from five-carbon sugar and hydroxymethylfurfural from six-carbon sugar. In fact, the careful control of the hydrolysis to selectively produce sugars without further conversion to furfurals is an important consideration in the hydrolysis of biomass when sugars are the desired products. However, furfural products are economically valuable in their own right. The practice of processing the hemicellulose five-carbon sugars in oat hulls to furfural has continued despite economic pressure from petrochemical growth, and it is still an important process, with an annual domestic production of approximately 100 million pounds. Furfural can be further transformed into a number of products, including furfural alcohol and furan. Even adiponitrile was produced from furfural for nylon from 1946 to 1961.

3.2.3 Levulinic Acid and Derived Products

Levulinic acid (4-oxo-pentanoic acid) results from subsequent hydrolysis of hydroxymethylfurfural. It is relatively stable toward further chemical reaction under hydrolysis conditions. Processes have been developed to produce it from wood, cellulose, starch, or glucose.

Levulinic acid is potentially a useful chemical compound based on its multifunctionality and its many potential derivatives. In the latter half of the 20th century, its potential for derivative chemical products was reviewed numerous times in the chemical literature. Some examples of the useful products described include levulinate esters, with useful solvent properties; δ-aminolevulinic acid, identified as having useful properties as a herbicide and for potentially controlling cancer; hydrogenation of levulinic acid, which can produce γ-valerolactone, a potentially useful polyester monomer (as hydroxyvaleric acid); 1,4-pentanediol, also of value in polyester production; methyltetrahydrofuran, a valuable solvent or a gasoline blending component; and diphenolic acid, which has potential for use in polycarbonate production.

3.2.4 Hydrogenation to Polyols

Sugars can readily be hydrogenated to sugar alcohols at relatively mild conditions using a metal catalyst. Some examples of this type of processing include sorbitol produced from glucose, xylitol produced from xylose, lactitol produced from lactose, and maltitol produced from maltose. All these sugar

alcohols have current markets as food chemicals. New chemical markets are developing; for example, sorbitol has been demonstrated in environmentally friendly deicing solutions.

In addition, the sugar alcohols can be further processed to other useful chemical products. Isosorbide is a doubly dehydrated product from sorbitol produced over an acid catalyst. Isosorbide has uses in polymer blending, for example, to improve the rigidity of polyethylene terephthalate in food containers such as soda pop bottles.

Hydrogenation of the sugar alcohols under more severe conditions can be performed to produce lower molecular-weight polyols. By this catalytic hydrogenolysis, five- or six-carbon sugar alcohols have been used to produce primarily a three-product slate consisting of propylene glycol, ethylene glycol, and glycerol. Although other polyols and alcohols are also produced in lower quantities, these three are the main products from hydrogenolysis of any of the epimers of sorbitol or xylitol. Product separations and purifications result in major energy requirements and costs. Consequently, controlling the selectivity within this product slate is a key issue in commercializing the technology. Selectivity can be affected by processing conditions as well as catalyst formulations.

3.2.5 Oxidation to Gluconic/Glucaric Acid

Glucose oxidation can be used to produce six-carbon hydroxyacids—either the monoacid, gluconic, or the diacid, glucaric. The structures of these chemicals suggest opportunities for polyester and polyamide formation, particularly polyhydroxypolyamides (i.e., hydroxylated nylon). Oxidation is reportedly performed economically with nitric acid as the catalyst. Specificity to the desired product remains to be overcome before these chemicals can be produced commercially.

3.3 Potential Pyrolysis Products

Pyrolysis of wood for chemicals and fuels production has been practiced for centuries; in fact, it was likely the first chemical production process known to humans. Early products of charcoal and tar were used not only as fuels but also for embalming, filling wood joints, and other uses. Today, pyrolysis is still important in some societies, but it has largely been displaced by petrochemical production in developed countries. At the beginning of the 20th century, important chemical products were methanol, acetic acid, and acetone. Millions of pounds of these products were produced by wood pyrolysis until approximately 1970, when economic pressure caused by competition with petroleum-derived products became too great. Due to new developments in "flash" pyrolysis beginning in the 1980s, there are new movements into the market with wood pyrolysis chemicals. However, these products are specialty chemicals and not yet produced on a commodity scale.

3.3.1 Chemical Fractions

Most new development work in pyrolysis involves separating the bulk fractions of chemicals in the pyrolysis oil. Heavy phenolic tar can be separated simply by adding water. A conventional organic chemical analytical separation uses base extraction for acid functional types, including phenolics. Combining the water-addition step and the base extraction with a further solvent separation can result in a stream of phenolics and neutrals in excess of 30% of the pyrolysis oil. This stream has been tested as a substitute for phenol in phenol–formaldehyde resins commonly used in restructured wood-based construction materials, such as plywood and particle board. This application allows for the use of the diverse mixture of phenolic compounds produced in the pyrolysis process, which typically includes a range of phenols, alkyl-substituted with one- to three-carbon side chains, and a mix of methoxyphenols (monomethoxy from softwoods and a mix of mono- and dimethoxy phenols from hardwood) with similar alkyl substitution. In addition, there are more complex phenolics with various oxygenated side chains, suggesting the presumed source—lignin in the wood.

The lower molecular-weight nonphenolics comprise a significant but smaller fraction of pyrolysis oil as produced in flash processes. Methanol and acetic acid recovery has potential. Hydroxyacetaldehyde, formic acid, hydroxyacetone, and numerous small ketones and cyclic ketones, which may have value in fragrances and flavorings, also have potential.

3.3.2 Levoglucosan and Levoglucosenone

Cellulose pyrolysis can result in relatively high yields of levoglucosan or its subsequent derivative levoglucosenone. Both are dehydration products from glucose. Levoglucosan is the 1,6-anhydro product, and levoglucosenone results from removal of two of the three remaining hydroxyl groups of levoglucosan with the formation of a conjugated olefin double bond and a carbonyl group. Levoglucosenone's chiral nature and its unsaturated character suggest uses for this compound in pharmaceuticals and in polymer formation.

3.3.3 Aromatic Hydrocarbon Compounds

The large fraction of phenolic compounds in pyrolysis oil suggests the formation of aromatics for chemical or fuel use. Catalytic hydrogenation (in this case, hydrodeoxygenation) can be used for removing the hydroxyl group from the phenolic compounds to produce the respective phenyl compound. For example, hydrogenation of phenol or methoxyphenol (guaiacol) would give benzene, and methyl guaiacol would give toluene. This chemistry was studied in the 1980s, when aromatics were being added to gasoline to improve octane number. Due to increased restriction on aromatics in gasoline to meet emission guidelines, this process concept was shelved. Production of such hydrocarbons (and other saturated cyclic hydrocarbons) remains an option for chemical production from wood pyrolysis; however, the economics are a drawback due to the extensive amount of hydrogenation required to convert the oxygenate to a hydrocarbon.

3.4 Potential Gasification Products

Whereas the pyrolysis process results in a product that retains much of the chemical functional character of the biomass feedstock, gasification operates under more severe process conditions and results in products of a relatively simple chemical structure but with wide applications. Gasification usually entails high-temperature partial oxidation or steam reforming of the biomass feedstock with typical products of hydrogen, carbon monoxide and dioxide, and light hydrocarbons. Other gasification processes using catalysis or bioconversion methods can also result in a primarily methane and carbon dioxide product gas.

3.4.1 Synthesis Gas to Produce Chemicals

Synthesis gas produced from biomass can be used for chemical production in much the same way as synthesis gas from steam reforming of natural gas or naphtha. Certain process configurations can be used with proper optimization or catalysis to control the important ratio of hydrogen to carbon monoxide in the synthesis gas. The synthesis gas can then be used in catalyzed reactions to produce methanol or hydrocarbons or even ammonia. Methanol can subsequently be converted by further catalytic steps to formaldehyde, acetic acid, or even gasoline.

3.4.2 Tar Chemicals

By-products formed in incomplete gasification include an array of organics ranging from the pyrolysis chemicals to further reacted products, including aromatics and particularly polycyclic aromatics, if a sufficient amount of time at high temperature is applied. As a function of temperature and time at temperature, the composition of the tar can vary from a highly oxygenated pyrolysis oil produced at lower temperature and short residence time to a nearly deoxygenated polycyclic aromatic hydrocarbon produced at high temperature. Because of these high-temperature tar components, including four- and five-aromatic ring structures, mutagenic activity can be significant.

3.5 Developmental Fermentations

New fermentation processes for additional biomass-derived chemical products are in various developmental stages. These processes range from improvements to known fermentations to new chemical products produced in genetically engineered organisms.

3.5.1 Acetic Acid

Acetic acid is a major organic chemical commodity product, with an annual domestic demand in excess of 6 billion pounds. The acetic acid currently available for chemical uses is petrochemically derived. Acetic acid fermentation, known from antiquity to form vinegar, is an *Acetobacter* partial oxidation of ethanol. More recently developed is the homofermentative conversion of glucose to acetic acid by "acetogenic" *Clostridium* bacteria. In some cases, 3 mol of acetate can be produced from each mole of six-carbon sugar feedstock, and concentrations of acetate of up to 1.5% can be accumulated. This type of fermentation effectively increases the theoretical yield of acetic acid from glucose by 50% and provides a more reasonable pathway for renewable resources.

3.5.2 Propionic Acid

All chemical production of propionic acid, best known as a food preservative, is currently derived from petroleum as a coproduct or by-product of several processes. However, propionic acid can readily be fermented with propionibacterium. Domestic production is approximately 220 million pounds per year and increasing only 1 or 2% annually. The production of esters for use as environmentally benign solvents represents a small but growing market.

3.5.3 Extremophilic Lactic Acid

As an economical production process, lactic acid fermentation from glucose has led the way into the marketplace for chemicals production. Although the

lactic acid process used for PLA production is based on conventional fermentation, the recovery of the lactide product for polymerization eliminates the neutralization and reacidification steps typically required. Efforts to improve the process are focused on the development of new organisms that can function under acidic conditions and produce higher concentrations of lactic acid. These organisms, called extremophiles, are expected to operate at a pH less than 2 in a manner similar to citric acid fermentations. Other work to develop higher temperature-tolerant organisms may also result in reduced production costs.

3.5.4 2,3-Butanediol/Ethanol

For this fermentation, a variety of bacterial strains can be used. An equimolar product mix is produced under aerobic conditions, whereas limited aeration will increase the production of the 2,3-butanediol to the level of exclusivity. Either optically active 2,3-butanediol or a racemic mixture can be produced, depending on the bacterial strain. Similarly, the final product concentration, ranging from 2 or 3 to 6–8%, depends on the bacterial strain. Recovering the high-boiling diol by distillation is problematic.

3.5.5 Succinic Acid

Succinate production from glucose and other carbohydrates has been well characterized for numerous bacterial organisms, including several succinogenes and the *Anaerobiospirillum succiniproducens*. The biological pathways for these organisms are similar and usually result in significant acetate production. The pathways also include a carbon dioxide insertion step such that potential theoretical yield is greater than 100% based on glucose feedstock. Mutated *Escherichia coli* has also been developed for succinic acid production, with yields as high as 90% shown on the pilot scale and a final succinic acid concentration of 4 or 5%. As with lactic acid production, maintaining the pH of the fermentor at approximately neutral is vital. The resulting disposal of mineral by-products from neutralization and reacidification is problematic. Recent advances in membrane separation methods or acid-tolerant fermentation organisms are important for the commercial development of succinic acid production.

3.5.6 Itaconic Acid

Itaconic acid is succinic acid with a methylene group substituted onto the carbon chain. It is typically produced in a fungal fermentation at relatively small scale and high cost ($1.2–2/lb). It is feasible that the economics of this fermentation could be improved in a manner similar to that for citric acid fermentation. As a result, itaconic acid could become available for use in polymer applications as well as for further processing to value-added chemical products.

3.5.7 Glycerol and 1,3-PDO

Bacterial fermentation of glycerol to 1,3-propanediol (PDO) has been studied as a potential renewable chemical production process. The glycerol feedstock could be generated as by-product from vegetable oil conversion to diesel fuel (biodiesel). The anaerobic conversion of glycerol to PDO has been identified in several bacterial strains, with *Clostridium butyricum* most often cited. The theoretical molar yield is 72%, and final product concentration is kinetically limited to approximately 6 or 7%. Genetic engineering of the bacteria to include a glucose to glycerol fermentation is under development and would result in a single-step bioprocess for PDO directly from glucose. The PDO product is a relatively new polyester source, only recently economically available from petroleum sources. Its improved properties as a fiber include both stain resistance and easier dye application.

3.5.8 3-Dehydroshikimic Acid

This interesting building block has been identified as a potential renewable resource. It is a hydroaromatic intermediate in the aromatic amino acid biosynthetic pathway. Its formation in mutated *E. coli* and its recovery in significant yields have been reported. It can be produced from six-carbon sugars with a theoretical yield of 43%, whereas a 71% yield is theoretically possible from five-carbon sugars based on a different metabolic pathway. Further processing of 3-dehydroshikimic acid can lead to protocatechuic acid, vanillin, catechol, gallic acid, and adipic acid.

3.6 Catalytic/Bioprocessing Combinations

Combining bioprocessing systems with catalytic processing systems provides some important opportunities for biomass conversion to chemicals. Although fermentation of biomass feedstocks to useful chemicals can be a direct processing step, in many cases the fermentation product can, in turn, be transformed into a number of value-added chemical products. Thus, the fermentation product acts as a platform chemical from which numerous final products can be derived. Some of these combined bioprocessing platform chemicals and catalytically derived families of value-added chemicals are described here.

3.6.1 Succinic Acid Derivatives

Two branches of the succinic acid family tree have been described. The first, involving hydrogenation, is similar to hydrogenation of maleic anhydride performed in the petrochemical industry based on butane oxidation. The hydrogenation products from succinic acid include γ-butyrolactone, 1,4-butanediol (BDO), and tetrahydrofuran (THF). Currently, 750 million pounds of both BDO and THF are produced per year. Lactone and THF have solvent markets, whereas BDO is an important polyester resin monomer (polybutylene terephthalate). THF is also used for Spandex fiber production.

The second pathway for succinic acid chemical production is through the succinamide to form pyrrolidones. *N*-methylpyrrolidone is an important low-toxicity, environmentally benign solvent with a growing annual market of more than 100 million pounds, displacing chlorinated hydrocarbon solvents. 2-Pyrrolidone can form the basis for several polymers, including polyvinyl pyrrolidone.

3.6.2 Itaconic Acid Derivatives

The structural similarity of itaconic acid to that of succinic acid suggests that it could be used to produce similar families of chemical products. Typically, the products would be methylated versions of the succinic-derived products, such as 2-methyl-1,4-butanediol, 3-methyltetrahydrofuran, or 3-methyl-*N*-methylpyrrolidone. Therefore, the final products would include methylated polyesters or methylated Spandex.

3.6.3 Glutamic Acid and Lysine

Glutamic acid and lysine are major chemical products, with glutamate sold as the sodium salt, monosodium glutamate, and lysine used as an important animal feed supplement. Both compounds are produced industrially by fermentation.

Glutamate is produced as L-glutamic acid at 680 million pounds per year, with 85% of the production in Asia. In the first half of the 20th century, glutamate was produced by hydrolysis of protein, primarily wheat gluten. The current fermentation process, using molasses or another cheap glucose source, had replaced protein hydrolysis processing by 1965. Because the current cost of glutamate is more than \$1.4/lb, it is too expensive for use as a platform chemical, and its use for chemical production is dependent on improved fermentation, recovery method modification, and integration of catalytic processing with the existing fermentation process.

Worldwide lysine production is only approximately one-third that of glutamate. Thus, the cost of lysine is approximately twice that of glutamate. Consequently, chemical production based on glutamate rather than lysine appears to be more likely.

Glutamate and lysine are interesting as renewable chemical feedstocks because each provides both multifunctionality and a five-carbon backbone. The five-carbon-based polymers are not widely used and are less well-known or understood since there is no easy way to produce five-carbon petroleum products. Processing steps can be envisioned for direct conversion of glutamate either to a five-carbon diacid by deamination (or thereafter to 1,5-pentanediol) for polyester production or to a five-carbon terminal amine/acid for nylon production. Lysine could be decarboxylated to produce a five-carbon diamine or could be deaminated to the amine/acid (the same as that for glutamate) for nylon production.

3.7 Plant-Derived Chemicals

One strategy for the use of biomass for chemicals production is the recovery of chemical products produced directly by plants. This strategy can involve either recovery from existing plants, such as the oils from certain seed crops, or recovery from genetically modified plants in which the chemical is produced specifically for harvesting.

3.7.1 Oleochemicals

The oil recovered from certain seed crops has been an important food product for many centuries. The growth of scale of the processing industry in the case of corn and soybeans has resulted in a reduction of the cost of the oil to the point that it can be considered for use in commodity chemical products. The use of vegetable oil for fuel as the transesterified methyl esters of the fatty acids derived from the triglycerides, known as biodiesel, has made some market inroads but primarily to meet alternative fuel use requirements and it is not based on demand. An important market breakthrough for renewable resources was the use of soybean oils in printing inks, instituted in the 1990s, which has been growing steadily. They are also used in toners for photocopy machines and in adhesive formulations. Other uses, such as plasticizers or binders based on epoxide formation from the unsaturated fatty acids, have also been investigated.

3.7.2 Polyhydroxyalkanoates

Polyhydroxyalkanoate (PHA) production has been demonstrated by both fermentative and plant growth methods. PHA is a biodegradable polymer with

applications in the biomedical field. Originally, PHA was limited to polyhydroxybutyrate (PHB), but it was found to be too brittle. A copolymer (PHBV) of 75% hydroxybutyrate and 25% hydroxyvalerate is now a commercial product with applications in medical devices. However, the fermentative process has some potentially significant drawbacks that, when considered on a life cycle basis, could result in poorer performance of PHA production from a renewable perspective than that of petroleum-derived polystyrene. To address these drawbacks, which include high energy requirements for cell wall rupture and product recovery, plant-based production of PHA has been investigated using genetic modification to engineer the PHB production pathway into *Arabidopsis* as the host plant. Alternatively, PHBV might be produced by other organisms that can use multiple sugar forms from a less-refined biomass feedstock than corn-derived glucose.

4. THE FUTURE BIOREFINERY

The future biorefinery (Fig. 1) is conceptualized as an optimized collection of the various process options described previously. In addition to the chemical processing steps for producing value-added chemicals, it will likely incorporate components from existing biorefinery-type operations, such as wood pulp mill, wet corn mill, and cheese manufacturing processes.

Carbohydrate separation and recovery will likely involve both the five- and six-carbon compounds. These compounds will respond to separations by selective hydrolysis because they have different activities in either chemical or enzymatic hydrolyses. Any oil or protein components liberated through this processing will be considered as potential high-value chemicals. The lignin component, the least developed of the biomass fractions, is often viewed as a fuel to drive the processing systems, but it is also a potential source of aromatic chemical products. Mineral recovery is most likely to be important for maintaining agricultural productivity, with the return of the minerals to the croplands as fertilizer.

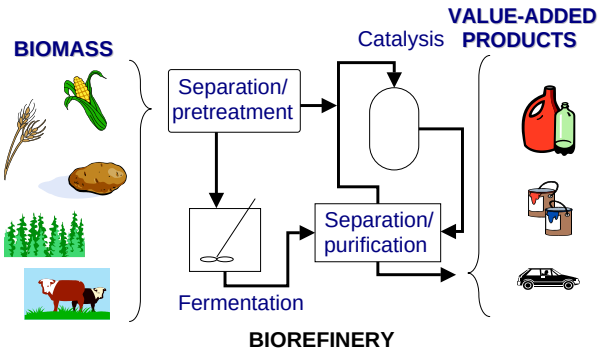


FIGURE 1 Biorefinery concept for production of value-added chemicals from biomass.

In order to achieve a true biorefinery status, each biomass component will be used in the appropriate process to yield the highest value product. Therefore, a combination of processing steps, tailored to fit the particular operation, will be used to produce a slate of chemical products, which will be developed in response to the market drivers to optimize feedstock utilization and overall plant income. In this manner, the return on investment for converting low-cost feedstock to high-value chemical products can be maximized.

The future biorefinery will likely have a major impact on society as fossil-derived resources become scarce and more expensive. Indeed, use of biomass is the only option for maintaining the supply of new carbon-based chemical products to meet society's demands. The biorefinery can have a positive environmental impact by deriving its carbon source (via plants) through reducing atmospheric levels of carbon dioxide, an important greenhouse gas. With appropriate attention to the biomass-derived nitrogen and mineral components in the biorefinery processes, a life cycle balance may be achieved for these nutrient materials if their recovery efficiencies can be maintained at high levels.

SEE ALSO THE FOLLOWING ARTICLES

Biomass Combustion • Biomass for Renewable Energy and Fuels • Biomass Gasification • Biomass: Impact on Carbon Cycle and Greenhouse Gas Emissions • Biomass Resource Assessment

Further Reading

Arntzen, C. J., Dale, B. E., *et al.* (2000). "Biobased Industrial Products." National Academy Press, Washington, DC.

Bozell, J. J. (ed.). (2001). "Chemicals and Materials from Renewable Resources," ACS Symposium Series No.784. American Chemical Society, Washington, DC.

Bozell, J. J., and Landucci, R. (eds.). (1993). "Alternative Feedstocks Program, Technical and Economic Assessment: Thermal/Chemical and Bioprocessing Components." National Renewable Energy Laboratory, Golden, CO.

Donaldson, T. L., and Culberson, O. L. (1983). "Chemicals from Biomass: An Assessment of the Potential for Production of Chemical Feedstocks from Renewable Resources, ORNL/TM-8432." Oak Ridge National Laboratory, Oak Ridge, TN.

Executive Steering Committee, Renewables Vision 2020 (1999). "The Technology Roadmap for Plant/Crop-Based Renewable Resources 2020," DOE/GO-10099-706. U.S. Department of Energy, Washington, DC.

Fuller, G., McKeon, T. A., and Bills, D. D. (eds.). (1996). "Agricultural Materials as Renewable Resources—Nonfood and Industrial Applications, ACS Symposium Series No.647." American Chemical Society, Washington, DC.

Goldstein, I. S. (ed.). (1981). "Organic Chemicals from Biomass." CRC Press, Boca Raton, FL.

Morris, D., and Ahmed, I. (1992). "The Carbohydrate Economy: Making Chemicals and Industrial Materials from Plant Matter." Institute for Local Self-Reliance, Washington, DC.

Pierre, St., L.E., and Brown, G. R. (eds.). (1980). "Future Sources of Organic Raw Materials: CHEMRAWN I." Pergamon, Oxford, UK.

Rowell, R.M., Schultz, T.P., Narayan, R. (eds.). (1992). "Emerging Technologies for Materials and Chemicals from Biomass," ACS Symposium Series No.476. American Chemical Society, Washington, DC.

Saha, B. C., and Woodward, J. (eds.). (1997). "Fuels and Chemicals from Biomass, ACS Symposium Series No.666." American Chemical Society, Washington, DC.

Scholz, C., and Gross, R. A. (eds.). (2000). "Polymers from Renewable Resources—Biopolyesters and Biocatalysis, ACS Symposium Series No.764." American Chemical Society, Washington, DC.

Biomass Combustion

ANDRÉ P. C. FAAIJ
Utrecht University
Utrecht, The Netherlands

1. **Introduction**
2. **Basics of Biomass Combustion: Principles and Fuels**
3. **Key Technological Concepts: Performance and Status**
4. **The Role of Biomass Combustion to Date: Different Contexts and Markets**
5. **Future Outlook and Closing Remarks**

Glossary

bubbling fluidized bed/circulating fluidized bed (BFB/CFB) Boiler designs in which the fuel is fed to a furnace where an inert bed material (like sand) is kept in motion by blowing (combustion) air from underneath.

co-combustion and co-firing Co-combustion or co-firing implies the combined combustion of different fuels (e.g., biomass and coal) in one furnace.

combined heat and power generation (CHP) Combustion facilities that produce both electricity (e.g., by driving a steam turbine and generator) and heat (e.g., process steam or district heating).

excess air Amount or ratio of air needed in surplus of oxygen theoretically needed for stochiometric complete combustion; excess air ratios can differ substantially between differ boiler concepts for effective combustion.

heating value (lower heating value [LHV]/higher heating value [HHV]) Energy content of a fuel released at complete combustion. Higher heating value gives the energy content for dry fuel. For the lower heating value, the energy needed to evaporate water present in the fuel and formed during combustion is substracted from the HHV.

municipal solid waste (MSW) Heterogeneous mixture of organic fractions, plastics, paper, and so on, as collected in urban areas from households and the service sector.

Biomass combustion means burning fuels from organic origin (e.g., wood, organic wastes, forest, and agriculture residues) to produce heat, steam, and power. Biomass combustion is responsible for over 95% of the production of secondary energy carriers from biomass (which contributes some 10 to 15% of the world's energy needs). Combustion for domestic use (heating, cooking), waste incineration, use of process residues in industries, and state-of-art furnace and boiler designs for efficient power generation all play their role in specific contexts and markets. This chapter gives an overview of the role, performance, technological concepts, and markets of biomass combustion.

1. INTRODUCTION

Biomass combustion is as old as civilization. Wood and other biomass fuels have been used for cooking and industrial processes, like iron production, for thousands of years.

Energy supplies in the world are dominated by fossil fuels (some 80% of the total use of more than 400 EJ per year). About 10 to 15% (or 45 ± 10 EJ) of this demand is covered by biomass resources (see Table I). On average, in the industrialized countries biomass contributes some 9 to 13% to the total energy supplies, but in developing countries this is as high as one-fifth to one-third. In quite a number of countries, biomass covers even over 50 to 90% of the total energy demand. It should be noted, though, that a large part of this biomass use is noncommercial and used for cooking and space heating, generally by the poorer part of the population. This also explains why the contribution of biomass to the energy supply is not exactly known; noncommercial use is poorly mapped. In addition, some traditional use is not sustainable because it may deprive local soils of needed nutrients, cause indoor and outdoor pollution, and result in poor health. It may also contribute to greenhouse gas (GHG) emissions and affect ecosystems. Part of this use is commercial though—that is, the household fuel wood in industrialized countries and charcoal and firewood in urban and

 © 2004 Elsevier Inc. All rights reserved.

TABLE I
World Primary Energy Use (1998)

	ExaJoule	ExaJoule	%
Fossil fuels		320	80
Oil	142		
Natural Gas	85		
Coal	93		
Renewables		56	14
Large hydropower	9		
Traditional biomass	38		
'Modern' renewables	9		
Nuclear		26	6
Total		402	100

industrial areas in developing countries—but there are almost no data on the size of those markets. An estimated 9 ± 6 EJ is covered by this category.

Poor statistics on bioenergy use even occur for industrial countries because of differences in definitions and fuels in or excluded (such as MSW and peat). Modern, commercial energy production from biomass (such as in industry, power generation, or transport fuels) makes a lower, but still significant contribution (some 7 EJ/year in 2000), and this share is growing. By the end of the 1990s, an estimated 40 GWe biomass-based electricity production capacity was installed worldwide (good for 160 TWh/year) and 200 GW heat production capacity (>700 TWh/year).

Nevertheless, by far the largest contribution of renewables to the world's energy supply comes from biomass, and combustion plays the key role as conversion route for producing useful energy from it. Besides domestic and industrial use of biomass via combustion, waste incineration is another major market where biomass combustion technology is widely applied, and a diversity of concepts has emerged over time.

In the past, a considerable part of the combustion capacity for electricity production (e.g., at the sugar and paper and pulp industry for utilization of process residues) was installed for solving disposal problems of biomass residues. The use of low-cost fuels has resulted in plants with low efficiencies. In the past decade or so, attention has begun to shift toward more efficient biomass combustion power plants, which is partly explained by increased utilization of more expensive biomass resources (such as agricultural and forest residues, which are more costly to collect and transport) and environmental and energy policies in various countries that stimulate the use of low- or zero-carbon energy sources. A range of advanced combustion concepts and technologies has been and is being developed over time, such as fluid bed technology, fully automated plants, flexible fuel concepts capable of dealing with both fossil and a diversity of biomass fuels, and a variety of co-combustion options. Another generic trend is that such plants are built with increasing capacities, thus resulting in economies of scale.

Section 2 goes into the basics of biomass combustion in relation to the properties of biomass fuels. Section 3 gives an overview of technological concepts deployed for various applications. Section 4 discusses the role biomass combustion plays in a number of example countries and sector. This article concludes with a look at the future of biomass combustion in Section 5.

2. BASICS OF BIOMASS COMBUSTION: PRINCIPLES AND FUELS

2.1 Basics: The Process of Biomass Combustion

Combustion can ideally be defined as a complete oxidation of the fuel. The actual full oxidation of the biomass fuel consist of several basic steps: drying, followed by pyrolysis, gasification, and finally full combustion. Biomass is never 100% dry, and any water in the fuel will be evaporated. The higher the moisture content, the more energy is needed to evaporate water. Given that many biomass resources have high moisture contents, this is an important parameter in the overall system performance. At moisture contents over 60%, it is generally impossible to maintain the combustion process. The heating value of a (biomass) fuel can be defined by the higher heating value, which is basically the energy content on dry basis, and the lower heating value, which subtracts the energy needed for the evaporation of water during the combustion process. Lower heating values are calculated by the following formula:

$$LHV_{wet} = HHV^{*}_{dry}(1 - W) - E^{*}_{w}(W + H^{*}m_{H_2O}),$$

in which HHV is the energy content of the dry biomass released during complete combustion, E_w is the energy required for evaporation of water (2.26 MJ/kg), W is the moisture content, H is the hydrogen content (weight percent of wet fuel), and

m_{H2O} is the weight of water created per unit of hydrogen (8.94 kg/kg).

This formula explains that the LHV can drop below zero at given moisture contents. Drying prior to the combustion process (e.g., with waste heat) is a possible way to raise the heating value to acceptable levels and this is sometimes deployed in practice. The heating value and moisture content of the fuel therefore also heavily affect the net efficiency of the combustion process in question. Some boiler concepts deploy condensation of the water in present the flue gases, which enables recovery of the condensation energy from the combustion process. This is an expensive option though.

Drying is followed by pyrolysis, thermal degradation in absence of oxygen and devolatization, at which tars, charcoal, and combustible gases are formed. The subsequent gasification results in formation of combustible gases (e.g., through char oxidation). Finally, at complete combustion, all gases are converted to CO_2 and H_2O and a range of contaminants, which partly depend on fuel composition and partly on the conditions of the combustion process. Incomplete combustion can, for example, lead to emissions of CO, soot, methane, polycyclic aromatic hydrocarbons (PAH), and volatile organic components (VOC), which is typical for lower temperature combustion and poor combustion air control. Dust emissions are caused by the mineral fraction (e.g., sand) of the fuel. Fuel bound nitrogen as well as thermal conversion of molecular nitrogen can be converted to NO_x, in particular at higher temperatures. N_2O (a strong greenhouse gas) can be formed in small amounts during combustion as well, depending on process conditions. Sulfur in the fuel is partly converted to SO_x. Both NO_x and SO_x contribute to acidification. SO_2 can also lead to corrosion in downstream equipment. The same is true for chlorine, which results in HCl emissions. Presence of chlorine may, under specific conditions, also lead to formation small amounts of the highly toxic dioxins and furans. Alkali metals (in particular K and Na) also contribute to corrosion problems. Those metals, as well as the silicon content and Ca, influence ash melting points as well. In case heavy metals are present, these can partly evaporate and end up in the flue gases as well. In particular, cadmium and mercury are volatile and evaporate already at lower combustion temperatures.

Combustion temperatures can vary between 800 and 2000°C and are influenced by moisture and excess air ratio. The latter can typically be controlled by the design of the combustion equipment—that is, with various air feeding steps (primary and secondary air) and boiler or furnace geometry (e.g., to create specific primary and secondary combustion zones, dominated by one of the basic processes described above). The higher the air ratio, the lower the maximum (adiabatic) temperature, but also the more complete the combustion.

The control and performance of the entire combustion process (including costs, efficiency, and emission profile) is always a trade-off. On the one hand, there are various technical variables such as design of the equipment, materials used, air and fuel feeding methods, and control strategies; on the other hand, a number of process variables (e.g., heat transfer, residence times, excess air, insulation) and fuel properties (moisture, mineral fraction, composition) must all be balanced to obtain the desired performance. Clearly, with bigger facilities, more advanced design, materials, and equipment such as emission control devices become more feasible; in general, better quality fuels (clean wood, pellets, etc.) are used for firing smaller combustion installations (i.e., below some 1 MWth) and poorer quality fuels (such as straw, waste fuels) for bigger installations.

2.2 Characteristics of Biomass Resources

Biomass resources that can be used for energy are diverse. A distinction can be made between primary, secondary, and tertiary residues (and wastes), which are available already as a by-product of other activities and biomass that is specifically cultivated for energy purposes:

- Primary residues are produced during production of food crops and forest products. Examples are residues from commercial forestry management and straw from various food crops (cereals and maize). Such biomass streams are typically available in the field and need to be collected to be available for further use.
- Secondary residues are produced during processing of biomass for production of food products or biomass materials. Such residues typically are available in larger quantities at processing facilities, such as the food- and beverage industry, saw- and paper mills, and the like.
- Tertiary residues become available after a biomass derived commodity has been used, meaning a diversity of waste streams is part of this category, varying from the organic fraction of municipal solid waste (MSW), waste and demolition wood, sludges, and so on.

Properties like moisture content, mineral fraction, density, and degree of contamination (e.g., heavy and alkali metals, nitrogen, sulfur, chlorine) differ widely depending on the source of the biomass.

Biomass is a highly reactive fuel compared to coal and has a much higher oxygen content, higher hydrogen/carbon ratio, and higher volatile content. The bulk composition of biomass in terms of carbon, hydrogen, and oxygen (CHO) does not differ much among different biomass sources though; typical (dry) weight percentages for C, H, and O are 45 to 50%, 5 to 6%, and 38 to 45%, respectively. The volatile content (lighter components that are particularly released during the pyrolysis stage) can vary between some 70 and 85%. The latter is typically higher for nonwoody and 'younger' greener biomass sources. CHO shares can be different for atypical fuels as sludges or MSW (the higher C content is due to biological breakdown and plastic fraction). Table II summarizes some characteristics for a range of biomass resources in terms of indicative ranges. Clearly, the diversity of characteristics (also in relation to price) of biomass resources is a challenge for combustion processes, and in practice a wide diversity of technical concept is deployed to convert different feedstocks to energy carriers for different markets and capacities.

Characteristics of the biomass are influenced by the origin of the biomass, but also by the entire supply system preceding any conversion step. Depending on the combustion process technology chosen, various pretreatment steps such as sizing (shredding, crushing, chipping) and drying are needed to meet process requirements. Furthermore, certain fuel properties exclude certain biomass fuels for specific combustion options, partially for technical and partially for environmental reasons.

Storage, processing (e.g., densification) before transport, and pretreatment influence moisture content, ash content (e.g., by means of sieving), particle size, and sometimes even degree of contamination (an example being straw washing, which reduces the salt content (chlorine and alkali content).

2.3 Power Generation

The resulting hot gases from the combustion can be used for a variety of applications such as the following:

- Smaller scale heat production for direct space heating, cooking, and water heating.
- Larger scale heat production for district and process heating (such as furnaces and drying).

TABLE II

Some Example Values on the Composition and Characteristics of a Selection of Biomass Fuels

Biomass feedback	Moisture content (percentage wet basis)	Ash content (percentage dry basis)	Heating value (LHVGJ/ wet ton)	Nitrogen (mg/kg, dry basis)	Sulfur (mg/kg dry basis)	Chlorine (mg/kg, dry basis)	Heavy metals	Costs (indications) (Euro/GJ)
Wood (chips)	30–50	1–2%	8–13	900–2000	70–300	50–60	Very low	From 1 to 5, when cultivated
Bark	40–60	3–5%	7–10	3000–4.500	300–550	150–200	Very low	From some 0.5 to 1
Straw	15	10%	14–15	3000–5000	500–1100	2500–4000	Low	From zero to high market prices
Grass	15–60	10%	6–15	4000–6000	200–1400	500–2000	Low	As residue low, cultivated Miscanthus: 1.5 to 4
Demolition wood	10–15	1%	15–17	300–3000	40–100	10–200	Low to very high	From negative to market prices
Organic domestic waste	50–60	20%	4–5	1000–4000	400–600	100–1000	Low	Negative prices due to tipping fees to zero
Sewage sludge	Over 70 in wet form	40%	1–10 (when dried)	2000–7000	600–2000	50–200	Very high	Strongly negative prices due to processing costs, up to zero.

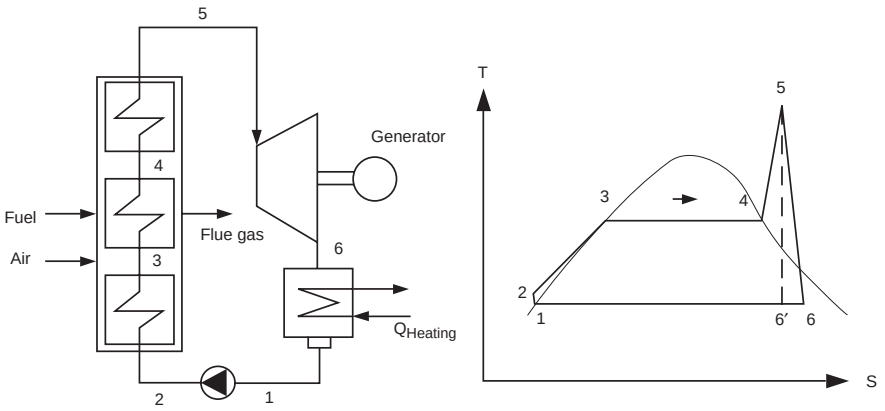


FIGURE 1 Basic principle and T/S diagram for a steam cycle including cogeneration. Steps from 1 to 6 cover pressure increase (1–2), heating up to evaporation temperature (2–3), evaporation in the boiler (3–4), superheating of vapor (4–5), expansion in the turbine (5–6) and finally condensation (6–1).

- Raising (high temperature) steam for industrial processes or subsequent electricity production using a steam turbine or steam engines.

Various devices can be used to convert heat from combustion to power. The most important and widely deployed are steam turbines (in the range of 0.5 to 500 MWe). Furthermore, steam engines for capacities below 1 MWe or so are used. Alternatives are the use of closed gas turbines (hot air turbine), Stirling engines, and an Organic Rankine Cycle, but those are either under development or to date only used in a few specific projects.

Steam turbines are mature, proven technology, which, when applied at larger scales, can reach high efficiencies when high pressure and steam is available. Steam pressures deployed in steam turbines range from 20 to 200 bars for superheated steam. Raising such high pressures with biomass fuels can be problematic when corrosive components (such as chlorine, sulfur, and alkali metals) are present in high concentrations in the flue gas. This is a reason why, for example, waste incinerators generally maintain low steam pressures (with resulting lower efficiencies).

Combined heat and power generation (CHP) configurations and delivery of process steam are made easy by the use of back-pressure turbines, a technique widely deployed in the cane-based sugar and paper-and-pulp industry for processing residues and supplying both power and process steam to the factory. Steam turbines (and the required steam generator) are expensive on a smaller scale, however, and high-quality steam is needed. Superheating steam (for high efficiency) is generally only attractive at larger scales. The same is true for efficiency-increasing measures like feed water preheating and steam reheating. For a condensing plant (implying exhausted steam that exits the turbine is cooled to ambient temperatures), a simple process scheme and T/S, temperature/entropy diagram, is depicted in (Fig. 1). Typical net overall electrical efficiencies of power plants range from less than 20% electrical efficiency at smaller scales to over 40% for very large systems (e.g., obtained with co-firing in a coal plant).

Steam engines play a much smaller role in current markets. Advantages are their flexibility (part load operation) and suitability for smaller scales. Electrical efficiencies are limited to 6% to up to 20% for the most advanced designs.

2.4 Ash

Ash remains after complete combustion, containing the bulk of the mineral fraction of the original biomass, but so does unburned carbon (which strongly depends on the technology deployed). High alkali and Si content typically give low ash melting temperatures (while Mg and Ca increase ash melting temperature). Ash sintering, softening, and melting temperatures can differ significantly among biofuels, and this characteristic is essential in determining temperature control to avoid sintering (e.g., on grates or in fluid beds) or slagging (resulting in deposits on, e.g., boiler tubes). Typically, grasses and straw have mineral fractions giving low melting temperatures.

Potential utilization of ash is also influenced by contaminants (such as heavy metals) and the extent to which the ash is sintered or melted. In case clean biomass is used and the ash contains much of the minerals (such as phosphor) and important trace

elements, it can be recycled to forest floors, for example. This is demonstrated in Sweden and Austria. Heavy metals typically concentrate in fly ash, which itself presents the possibility of removing a smaller ash fraction with most of the pollutants. This is the practice in Austria, particularly for removing cadmium.

Co-firing can give specific issues, since biomass additions to, for example, coal firing can change ash characteristics (which are of importance for reuse in cement production, for example). More important is that biomass ash can result to increased fouling and reducing the fusion temperatures of coal ash. The extent to which those mechanisms become serious issues depends on the fuel properties, both of the coal and biomass, as well as the specific boiler design.

3. KEY TECHNOLOGICAL CONCEPTS: PERFORMANCE AND STATUS

As argued, a wide range of combustion concepts is developed and deployed to serve different markets and applications. To structure this range, a distinction can first be made between small-scale domestic applications (generally for heating and cooking) and larger scale industrial applications. First, domestic applications will be discussed, distinguishing between the use of stoves in (in particular in developing countries) and the use of more advanced furnaces, boilers, and stoves utilized in industrialized countries. For industrial applications, a basic distinction is made between standalone systems and co-firing.

3.1 Domestic Biomass Use in Developing Countries

The aforementioned (and largely noncommercial) use of biomass for cooking and heating in particular in developing countries is applied with simple techniques, varying from open fires to simple stoves made from clay or bricks. Typically, the efficiency of such systems is low, to less than 10% when expressed as energy input in wood fuel versus actual energy needed for cooking food. Improvements in the energy efficiency of cooking in particular can therefore lead to significant savings of fuel demand. Improved stoves, the use of high-pressure cookers, and in some cases the shift to liquid or gaseous fuels (like bio-ethanol, fischer-tropsch liquids, or biogas) can have a major positive impact in both overall energy efficiency and emission control.

The use of agricultural residues is common in regions were forest resources are lacking. The use of straw or dung is typically found in countries like China, India, and parts of Africa. Wood is definitely preferred over agricultural residues because the latter have poor combustion properties, low density complicating transport, and cause a lot of smoke when used.

3.2 Modern Biomass Combustion for Domestic Applications

Typical capacities for combustion systems in domestic applications for production of heat and hot water range from only a few kWth to 25 kWth. This is still a major market for biomass for domestic heating in countries like Austria, France, Germany, Sweden, and the United States. Use of wood in open fireplaces and small furnaces in houses is generally poorly registered, but estimated quantities are considerable in countries mentioned. Use of wood in classic open fireplaces generally has a low efficiency (sometimes as low as 10%) and generally goes with considerable emissions of, for example, dust and soot. Technology development has led to the application of strongly improved heating systems, which are, for example automated, have catalytic gas cleaning, and make use of standardized fuel (such as pellets). The efficiency benefit compared to, for example, open fireplaces is considerable: open fireplaces may even have a negative efficiency over the year (due to heat losses through the chimney), while advanced domestic heaters can obtain efficiencies of 70 to 90% with strongly reduced emissions. The application of such systems is especially observed in, for example, Scandinavia, Austria, and Germany. In Sweden in particular, a significant market has developed for biomass pellets, which are fired in automated firing systems.

Concepts for production of heat vary from open fireplaces to fireplace inserts, heat storing stoves (where a large mass of stone or tiles is used to store heat and again gradually dissipate that heat for a period of up to several days) to catalytic combustion systems, and use of heat exchangers that avoid losses of the produced heat through the chimney. The transfer of heat takes place by convection and direct radiation.

Boilers are used for hot water production (for either space heating or tap water). Basic concepts cover over-fire, under-fire, and downdraught boilers, which primarily differ with the direction of the combustion air flow through the fuel. Such systems

are usually linked to water heating systems with a capacity of some 1 to 5 m^3, including the possibility for temporal heat storage. Equipment is commercially available and deployed with advanced emission control systems. All systems mentioned generally rely on good quality wood, like logs, which are not too wet and uncontaminated.

Pellet burners are also available and the way combustion is performed shows some similarities with oil firing, since the pellets are crushed prior to combustion, resulting in a constant flow of fine particle to the burner, which is easily controlled. Fully automated systems are on the market that rely on a large storage of pellets in, for example, a basement which is filled once every half year or so. Control is steered by the heat demand of the hot water system. Infrastructure with pellet distribution is present in countries like Sweden.

Wood chip stoves can have similar characteristics in terms of automation, but require more voluminous storage and an oven or firebox for combustion of the chips. Those are called stoker burners. Temperatures can rise up to 1000°C when good quality, dry fuel is used.

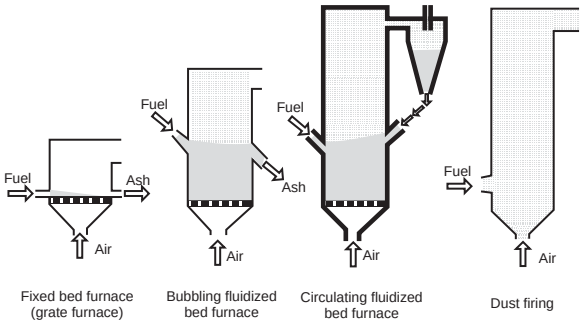


FIGURE 2 Basic combustion concepts for industrial applications.

3.3 Biomass Combustion for Industrial Applications: Stand Alone

Biomass combustion concepts exceeding a capacity of some 1 MWth are generally equipped with automated feeding and process control. Basic combustion concepts include fixed bed or grate furnaces, as well as bubbling fluidized bed (BFB) and circulating fluidized bed (CFB) furnaces. Furthermore, dust firing is deployed, but this is less common (see Fig. 2). Table III summarizes some of the main characteristics of the combustion concepts discussed.

- *Fixed bed.* Fixed bed furnaces (e.g., stoker furnaces) are typically suited for smaller capacities (e.g., up to 6 MWth). Grate-fired systems are deployed in a wide scale range. They can be divided

TABLE III

Some Main Characteristics of Key Combustion Concepts Currently Deployed

Combustion concept	Key advantages	Key disadvantages	Key applications
Underfeed stokers	Low cost at smaller capacities, easy control	Needs high-quality fuel with respect to composition and size	<6 MWth heating; buildings, district heating
Grate furnaces	Relatively cheap and less sensitive to slagging	Different fuels cannot be mixed for certain grate-type; nonhomogeneous combustion conditions (resulting in higher emissions)	Wide range; widely applied, especially for waste fuels and residues; typical capacity some 20–30 MWe; waste incinerators up to 1–2 Mtonne capacity/year
Dust combustion	High efficiency, good load control and low Nox	Small particle size required, wearing of lining	Less common (especially suited for sawdust)
BFB furnaces	Highly flexible for moisture contents, low Nox, high efficiency and decreased flue gas flow	High capital and operational costs, sensitive to particle size (requires pretreatment), high dust load in flue gas, some sensitivity for ash slagging; erosion of equipment	>20 MWth; deployed for flexible fueled plants and more expensive fuels (e.g., clean wood)
CFB furnaces	High fuel flexibility with respect to moist and other properties, low Nox, easy adding additives (e.g., for emission control), high efficiency and good heat transfer	High capital and operational costs, sensitivity to particle size (requires pre-treatment), high dust loads in flue gas, inflexible for part load operation; very sensitive to ash slagging, erosion of equipment	>30 MWth; deployed for flexible fueled plants and more expensive fuels; of CFB and BFB combined, some 300 installations build worldwide

into co-, cross-, and counter-current fired systems, which principally differ in the direction of the combustion air-flow in relation to the biomass flow. While the fuel is moving over the grate, the combustion process develops starting with drying, devolatization, formation of charcoal, and finally ash on the grate. Typically, primary and secondary combustion chambers are found, leading to staged combustion. Grate furnaces are very flexible with respect to fuel properties (e.g., moisture and size). Depending on the fuel properties, different grate types are deployed. Traveling grates are used meaning the grate is working like a transport belt, while the biomass is gradually converted along the way. Ash typically falls through the grate and is removed with collection systems. Ash removal can be both wet and dry. Horizontal, vibrating, and inclined grates are also applied. Typically, MSW incineration (which is a very inhomogeneous fuel) makes use of inclined grates, which result in longer residence times of the fuel particles. Another distinction can be made between air and water cooled grates. The latter is typically used for dryer fuels that give a higher combustion temperature; the first is deployed for wetter fuels.

- *Bubbling fluidized bed.* In a BFB, the fuel is fed to a furnace where an inert bed material (like sand) is kept in motion by blowing (combustion) air from underneath. Temperatures are typically 800 to 900°C, and bed temperatures are homogeneous. Low excess air is needed, minimizing flue gas flows. Homogenous conditions allow for good emission control and adding of additives to the bed (like limestone) can be deployed to reduce emissions further. Fuel bed (FB) technology is flexible for different fuel properties. In contrast, this technology is inflexible to particle size, resulting in the need for more extensive pretreatment.
- *Circulating fluidized bed.* Key characteristics are rather similar to BFB technology, but fluidizing velocities of the bed are higher, leading to higher turbulence. This results in better heat transfer (and higher efficiency), but requires a particle collector (or cyclone) to feed bed material that is blown out of the furnace back to the bed. Dust loads in the flue gas are high therefore. Due to high capital and operation costs, both BFB and CFB become attractive at larger scales (i.e., over 30 MWth). At such capacities, the advantages for smaller downstream equipment (boiler and flue gas) cleaning start to pay off.
- *Other concepts.* Some concepts that do not really fit the division given here are the dust combustion (e.g., deployed for dry fine fuels as sawdust), the rotating cone furnace (involving a rotating vessel in which the fuel is fed; in particular suited for waste fuels), and the whole tree energy concept. This design is capable of firing compete logs, which are dried over a longer period time in a covered storage using waste heat from the combustion process, and the resulting net energy efficiency is high. The system has not yet been commercially applied though.

3.4 Co-combustion

Co-combustion involves the use of biomass in existing fossil fuel fired power plants next to coal, wastes, peat, or other fuels. Co-combustion can be categorized in direct co-firing (directly fed to the boiler), indirect co-firing (fuel gas produced via gasification is fed to the boiler), and parallel combustion (use of biomass boiler and use of the resulting steam in an existing power plant).

Direct co-firing can be done by mixing with coal, separate pretreatment and handling of the biomass fuel and subsequent joint firing with coal, separate biomass fuel burners, and the use of biofuel as a reburn fuel (for NO_x reduction), but the latter is not commercially applied to date.

The advantages of co-firing are apparent: the overall electrical efficiency is high due to the economies of scale of the existing plant (usually around 40% in the case of pulverized coal boilers of some 500 to 1000 MWe), and investments costs (in particular for pretreatment) are low to negligible when high-quality fuels as pellets are used. Also, directly avoided emissions are high due to direct replacement of coal. Combined with the fact that many coal-fired power plants in operation are fully depreciated, this makes co-firing usually an attractive GHG mitigation option. In addition, biomass firing generally leads to lowering sulfur and other emissions, although ash utilization and increased fouling can become serious problems in some cases.

3.5 Environmental Performance and Emission Control

Both the biomass properties and composition and the combustion process influence the environmental performance of combustion processes. Section 2 listed the main contaminants. Table IV summarizes some generic data on major emissions of key combustion options for West European conditions.

TABLE IV

Generic Emission Data for Key Biomass Combustion Options

Combustion option	NO_x (mg/MJ)	Particulates (mg/MJ)	Tar (mg/MJ)	CO (mg/MJ)	UHC as CH_4 (mg/MJ)	PAH (μg/MJ)
FB boiler	170	2	n.m.	0	1	4
Grate-fired boiler	110	120	n.m.	1850	70	4040
Stoker burner	100	60	n.m.	460	4	9
Modern wood stove	60	100	70	1730	200	30
Traditional wood stove	30	1920	1800	6960	1750	3450
Fireplace	n.m.	6050	4200	6720	n.m.	110

Values concern wood firing only and West European conditions (and emissions standards). N.m., not measured.

Those data clearly reflect the better emission profiles of FB technology (except fro NOx) and the poor performance of small-scale, traditional combustion.

Depending on applicable emission standards, various technical measures may need to be taken to meet those standards. Measures can be divided into primary measures (covering design and control of the combustion process and technology) and secondary measures that remove contaminants after the actual combustion. These typically include particle control (e.g., by means of cyclones, filters, and scrubbers), SO_2 removal (additives as limestone and scrubbing), and No_x reduction measures, which include primary measures as staged combustion but also selective (Non-) catalytic reduction. Generally, secondary measures add significantly to the total capital and operational costs, which becomes even more apparent at smaller scales.

3.6 Efficiency and Economics

Table V shows a generic overview of energy efficiency and costs of the main conversion options discussed. For comparison, gasification for electricity production is also included.

Overall economy of a plant or system strongly depends on fuel prices (and quality), emissions standards, load factors, and economic factors as interest rates and feed-in tariffs for power and heat (when applicable). Clearly, as will also be discussed in the next section, various countries installed very different policy measures, financial instruments, and standards that sometimes resulted in very specific markets for certain technologies. The next section discusses a number of countries and markets and the specific deployment of biomass combustion technology to date.

4. THE ROLE OF BIOMASS COMBUSTION TO DATE: DIFFERENT CONTEXTS AND MARKETS

This section discusses some main developments in the role of biomass combustion in various countries, contexts, and sectors, covering domestic heating, district heating and CHP, large-scale combustion and industrial use, waste incineration, and co-firing. Table VI gives an overview of some countries in the bioenergy field, as well as their specific policies, resources, and the role of combustion for production of power and heat.

It is important to understand the interlinkages between biomass resource availability and the use of specific technologies. The biomass resource base is a determining factor for which technologies are attractive and applied in relation to its characteristics (as discussed in Section 2) but also in relation to the growing biomass market. To take Europe as an example, organic wastes (in particular MSW) and forest residues represent the largest available potential and deliver the largest contribution to energy production from organic (renewable) material. Market developments, natural fluctuations in supply (e.g., due to weather variations), as well as the economics and various other factors (such as accessibility of areas) of bioenergy all influence the technical and economic potential of such resources, and it is impossible to give exact estimates of the potential. However, *available* resources are increasingly utilized. When the use of bioenergy expands further in the future (as projected by many scenario studies and policy targets), such larger contributions of biomass to the energy supply can only be covered by dedicated production systems (e.g., energy crops). Over time, several stages can be observed in biomass

TABLE V

Global Overview of Current and Projected Performance Data for the Main Conversion Routes of Biomass to Power and Heat, and Summary of Technology Status and Deployment in the European Context

Conversion option		Typical capacity range	Net efficiency (LHV basis)	Investment cost ranges (Euro/kW)	Status and deployment in Europe
Combustion	*Heat*	Domestic 1–5 MWth	From very low (classic fireplaces) up to 70–90% for modern furnaces	~100/kWth 300–700/kWth for larger furnaces	Classic firewood use still widely deployed in Europe, but decreasing. Replacement by modern heating systems (i.e., automated, flue gas cleaning, pellet firing) in, for example, Austria, Sweden, Germany, ongoing for years
	CHP	0.1–1 MWe	60–90% (overall) ~10% (electrical) 80–100% (overall) 15–20% (electrical)		Widely deployed in Scandinavia countries, Austria, Germany, and to a lesser extent France; in general increasing scale and increasing electrical efficiency over time
	Stand-alone	20–100's MWe	20–40% (electrical)	2.500–1600	Well-established technology, especially deployed in Scandinavia; various advanced concepts using fluid bed technology giving high efficiency, low costs, and high flexibility commercially deployed; mass burning or waste incineration goes with much higher capital costs and lower efficiency; widely applied in countries like the Netherlands and Germany
	Co-combustion	Typically 5–20 MWe at existing coal fired stations; higher for new multifuel power plants.	30–40% (electrical)	~250 + costs of existing power station	Widely deployed in many EU countries; interest for larger biomass cofiring shares and utilization of more advanced options (e.g., by feeding fuel gas from gasifiers) is growing in more recent years
Gasification	*Heat*	Usually smaller capacity range around 100's kWth	80–90% (overall)	Several 100's/kWth, depending on capacity	Commercially available and deployed; but total contribution to energy production in the EU is very limited
	CHP gas engine	0.1–1 MWe	15–30%	3.000–1.000 (depends on configuration)	Various systems on the market; deployment limited due to relatively high costs, critical operational demands, and fuel quality
	BIG/CC	30–100 MWe	40–50% (or higher; electrical efficiency)	5.000–3.500 (demos) 2.000–1.000 (longer term, larger scale)	Demonstration phase at 5–10 MWe range obtained, rapid development in the 1990s has stalled in recent years; first generation concepts prove capital intensive

Based on a variety of literature sources. Due to the variability of data in the various references and conditions assumed, all cost figures should be considered as indicative. Generally they reflect European conditions.

utilization and market developments in biomass supplies. Different countries seem to follow these stages over time, but clearly differ in the stage of development (see also Table V):

1. Waste treatment. Waste treatment, such as MSW and the use of process residues (paper industry, food industry), onsite at production facilities is generally the starting phase of a developing bioenergy market and system. Resources are available and often have a negative value, making utilization profitable. Utilization is usually required to solve waste management problems.
2. Local utilization. Local utilization of resources from forest management and agriculture is generally next. Such resources are more expensive to collect and transport, but usually still economically attractive. Infrastructure development is needed.
3. Biomass market development on regional scale. Larger scale conversion units with increasing fuel flexibility are deployed, increasing average transport distances and further improved economies of scale. Increasing costs of biomass supplies make more energy-efficient conversion facilities necessary as well as feasible. Policy support measures such as feed-in tariffs are usually needed to develop into this stage.
4. Development of national markets with an increasing number of suppliers and buyers, creation of a marketplace, increasingly complex logistics. There is often increased availability, due to improved supply systems and access to markets. Price levels may therefore even decrease.
5. Increasing scale of markets and transport distances, including cross border transport of biofuels; international trade of biomass resources (and energy carriers derived from biomass). In some cases, conflicts arise due to profound differences in national support schemes (subsidies, taxes, and environmental legislation; an example is the export of waste wood from the Netherlands and Germany to Sweden). Biomass is increasingly becoming a globally traded energy commodity.
6. Growing role for dedicated fuel supply systems (biomass production largely or only for energy purposes). There are complex interlinkages with existing markets for pulp- and construction wood, use of agricultural residues, and waste streams. So far, dedicated crops are mainly grown because of agricultural interests and support (subsidies for farmers, use of set-aside subsidies), which concentrates on oil seeds (like rapeseed) and surplus food crops (cereals and sugar beet). Sweden has largest experience with SRC-Willow. There is interest in agricultural crops such as sweet sorghum in Southern Europe. Perennial grasses, poplar, and plantation forest receive attention in various countries.

4.1 Domestic Heating

Combustion on a domestic scale for heat production (space heating, cooking) is widely applied, as argued. In various developing countries in Asia and Africa, biomass use accounts for 50% to even 90% of the national energy supply, which is almost exclusively deployed through often inefficient combustion. Improved technologies, often meaning relatively simple stove designs, are available and could make a huge difference in fuel wood consumption and emissions. It is clear though that biomass is not a preferred fuel once people can afford more modern energy carriers, such as kerosene, gas, or electricity. Typically, developing regions follow the so-called fuel ladder, which starts with biomass fuels (first agricultural residues, then wood, charcoal, kerosene, propane, gas, and finally electricity). Given the inefficiency of direct biomass use for cooking in particular, the future use of available biomass resources in developing countries may well be the production of modern energy carriers as power and fuels.

In industrialized countries, where more modern combustion technology for domestic heating is available, production of heat with biomass is not increasing as fast when compared to biomass-based power generation or, more recently, production of fuels for the transport sector. This is illustrated by development in the European Union where in 1990 the production of heat from biomass amounted about 1500 PJ rising to over 1800 PJ in 1999 (an increase of 2% per year), while electricity production from biomass amounted 15 TWhe in 1990 and rose up to 46 TWhe in 1999 (an increase of 9% per year). This growth is almost exclusively realized by application of modern biomass combustion technology. Furthermore, despite the modest role of biofuels in energy terms, the production and use of biofuels rapidly increased over the past 10 years. Biodiesel production increased from 80 ktonne in 1993 up to 780 ktonnes in 2001. Germany and France produce the bulk, with minor contributions from Italy and Austria. Ethanol production in the EU increased from 48 up to 216 ktonne in the same period.

TABLE VI

Global Overview of Bioenergy Use, Policy, and Developments Over Time in a Selection of Countries

Country	Status of bioenergy, main achievement and goals	Main policy instruments deployed	Biomass supplies; use and potential	Heat and power
Austria	Biomass accounts for 11% of the national energy supply. Forest residues are used for (district) heating, largely in systems of a relatively small scale. Some of the policy measures deployed are province oriented.	Financial support for individual heating systems and gasifiers. Subsidies for farmers producing biomass (as rape). Tax reduction biofuels and tax on mineral fuels.	Mainly forest residues and production of rape for biodiesel.	Heat production declined end of nineties due to declining number of classic wood boilers. Wood boilers for domestic heating still important. About one-fifth consists of modern concepts. Bioelectricity steady grower.
Denmark	Running program for utilisation of 1.2 million tonnes of straw as well as the utilization of forest residues. Various concepts for cofiring biomass in larger scale CHP plants, district heating, and digestion of biomass residues.	Long-term energy program; energy crops to contribute from 2005. Subsidies for bioenergy projects up to 50% are possible. Feed in tariffs guaranteed for 10 years	Use of straw and wood residue, waste, and biogas production (digestion) and a significant role for energy crops in 2010 (total ~150 PJ).	1990–2000, heat production more or less constant, electricity production increased 10-fold.
Finland	Obtains 20% of its primary energy demand from biomass (one third being peat). Especially the pulp and paper industry makes a large contribution by efficient residue and black liquor utilization for energy production. Strong government support for biomass; a doubling of the contribution is possible with available resources.	Exemption from energy tax. Financial support for forest owners for producing forest residues. Targets for waste to energy production.	Forest residues, black liquor (paper industry; two-thirds of the total) and to a lesser extent peat play a key role. Further use of forest residues likely to result in more expensive supplies in the future.	In 2010 over 1000 MWe biomass electricity generation capacity should be installed. Heat production increased about 60% between 1990 and 2000, electricity some 70%.
France	There is no specific bioenergy policy in France, but the role of biofuels (ethanol and biodiesel) and biomass for heating is significant: some 400 PJ heat, 500 PJ of fuel and 10 PJe are produced. Growth rates have been very low, however, in recent years. Biomaterials explicitly included in the strategy. Electricity has a low priority. Biomass contributes about two-thirds of the total energy production for renewables (in turn, 6% of the total).	100% tax exemption for biodiesel; 80% for bioethanol. Investment subsidies, R&D. Feed-in tariffs waste incineration and landfill gas; probably to be extended to other options.	Mainly focus on available resources, residues, and wastes for heat production. Surplus cereal production and production rape. Specific attention for increasing the use of biomaterials.	Major role for domestic heating, stable market, no real technological developments. Smaller contribution for collective and industrial heating systems. Modern boiler concepts for heat production key are pursued. Furthermore, considerable biogas utilization.
Germany	Renewables contribute some 3% to the total energy supply. The contribution of biomass is to increase from about 300 PJ (fuel input) to over 800 PJ in 2010. Waste policies relevant due to stringent targets in reuse and prohibiting landfilling. Targets for biofuels are ambitious; about 10-fold more compared to current levels.	Tax exemption on biofuels. Subsidies up to 100% and support for higher fuel costs. Various R&D programmes and support measures. Important role for the ministry of agriculture.	Waste and residue utilisation of key importance. Rapeseed production to biodiesel.	About a doubling of bioelectricity production between 1990 and now (total some 19 Pje). Similar rate for heat (now over 200 PJ). Significant technical developments in smaller scale heat and electricity production concepts (e.g., gasification based).

Netherlands	Biomass and waste contribute over 50 PJ in 2000. Target for 2007 in 85 PJ, going up to about 150 PJ in 2020. In long-term strategy formulation biomass envisaged to play a key role in the energy supply (600–1000 PJ in 2040).	Feed-in tariffs for various technologies (differentiating between scales and fuels). Tax measures and subsidies for investments.	Waste and residues main resources. Import of biomass current practice for green electricity production. Import a key role to play for meeting long-term targets.	Waste incineration and cofiring in coal-fired power stations play a prominent role. Widespread application of digestion. Electricity production to increase.
New Zealand	Renewables cover some 30% of the national energy supply; bioenergy and waste over 6%. Natural resource base (forest in particular) only partly used.	National strategy on renewable energy, but so far not covering quota or energy taxes.	Woody biomass (forest residues) and black liquor account for the bulk of currently used resources.	25 PJ in wood processing industries, 21 PJ in pulp and paper industry. Growth foreseen in domestic heat markets.
Spain	Contribution of biomass in total to increase from some 160 PJ (about 50% of total renewable energy) at present to 430 PJ in 2010 (about two-thirds of renewable energy). RES contribute some 6% in 2000 with targets for 2010 set at 30%.	Tax deduction for investments (10%) and direct subsidies (up to 30%) and discount on interest (up to 5 points) on bioenergy projects. Reduced rate excise duty.	Use of waste and residues but also major attention for energy crops, both classic (rape, cereals) as new options such as sweet sorghum.	Ambitious targets for increased electricity production and biogas production and use. Heat production constant.
Sweden	Biomass accounts for 17% of the national energy demand. Use of residues in the pulp and paper industry and district heating (CHP) and the use of wood for space heating and dominant. Biomass projected to contribute 40% to the national energy supply in 2020.	Taxation and administrative measures, most notably a CO_2 tax and energy tax. Subsidies for CHP facilities. Tax exemption (energy, environmental, and fees) for biofuels; no direct subsidies.	Wood fuels (forest residues) dominate the supplies (about 70% of the total). Wood market expanded over time with logistics covering the whole country and international trade; 14.000 ha SRC-Willow.	Bioelectricity not a very fast grower. Heat production steady increase from 160 PJ in 1990 to some 270 PJ in 2000.
United Kingdom	Renewable energy in total contributes about 1% to the total energy supply of the United Kingdom. Biomass accounts for about two-thirds of that. Rapid growth in electricity production from biomass and waste is observed. Wastes but also larger scale use of energy crops to play main role.	Bioenergy part of general GHG mitigation policy. Waste policy of major relevance. Innovative bids for biofuel projects supported. Duty reduction on biofuels. No direct subsidies. Support for farmers growing SRC.	MSW large part of total current supplies. Energy crops (SRC-Willow in particular) seen as important for longer term.	MSW incineration accounts for about 50% of total energy from biomass. Electricity production increased eight-fold between 1990 and 2000 (up to 4500 GWhe). Heat production doubled to almost 40 PJ.
United States	Approximately 9000 MWe biomass fired capacity installed; largely forest residues. Production of 4 billion litres of ethanol In total some 2600 PJ supplied from biomass (over 3% of primary energy use)	Past decade emphasis on efficient power generation. For 2020 bioenergy is identified to become major cost effective supplier of heat and power. The biorefinery concept (combined production of products, fuels, and power) deserves considerable attention.	Forestry residues, waste, and agricultural residues are all important. Considerable use of (surplus) corn for ethanol production. Various crop production systems tested and considered.	Electricity production and CHP cover some 1600 PJ of biomass use. Mainly through combustion of residue and agricultural waste processing. An additional 600 PJ for domestic heating. Increase in installed capacity has stalled.

4.2 District Heating and CHP

The application of biomass fired district heating is widely applied in Scandinavian countries and Austria. In Scandinavia, biomass fired CHP really took off in the 1980s as a result from national climate and energy policies. In the first stages, retrofits of existing coal-fired boilers were popular. Over time, the scale of CHP systems shows an increasing trend, with apparent advantages as higher electrical efficiencies and lower costs. This was also combined with a developing biomass market, allowing for more competitive and longer distance supplies of biomass resources (especially forest residues). During the 1990s, Denmark deployed a major program for utilizing straw. Various technical concepts were developed and deployed, such as the so-called cigar burners combined with efficient straw baling equipment, transport, and storage chains. Other innovations were needed to deal with the difficult combustion characteristics of straw, such the high alkali and chlorine content. This led to complex boiler concepts, such as those involving two-stage combustion, but also new pretreatment techniques such as straw washing. Austria, another leading country in deploying biomass-fired CHP, focused on smaller scale systems on village level, generally combined with local fuel supply systems. All countries mentioned have colder climates, making CHP economically attractive. Furthermore, involvement of local communities has proven important. Municipalities and forest owners are often the owners of the CHP-plants. Energy costs of those systems are usually somewhat higher. Local societal support is generally strong though, especially due to the employment and expenditures that benefit the local community. High labor costs also led to high degrees of automation though, with unmanned operation typical for many of the newer facilities. In Eastern European countries, district heating is widely deployed (but with little use of biomass). Opportunities in Central and Eastern Europe come from the energy system requiring further modernization and replacement of outdated power (and heat) generating capacity. Retrofitting and repowering schemes for coal-fired power and CHP generation capacity allowing for co-combustion of biomass streams are already successful in various Central and Eastern European countries. In the past 5 years or so, in particular Scandinavian companies have become active in this field in the context of joint implementation. Clearly, due to the combination of available district heating networks and abundant forest resources, the potential for biomass fired CHP in countries like Poland, the Baltic region, and Russia is high.

4.3 Larger Scale Combustion and Industrial Use for Power Generation and Process Heat

Larger scale combustion of biomass for the production electricity (plus heat and process steam) is applied commercially worldwide. Many plant configurations have been developed and deployed over time. Basic combustion concepts include pile burning, various types of grate firing (stationary, moving, vibrating), suspension firing, and fluidized bed concepts. Typical capacities for standalone biomass combustion plants (most using woody materials, such as forest residues, as fuel) range between 20 and 50 MWe, with related electrical efficiencies in the 25 to 30% range. Such plants are only economically viable when fuels are available at low costs or when a carbon tax or feed-inn tariff for renewable electricity is in place. Advanced combustion concepts have penetrated the market. The application of fluid bed technology and advanced gas cleaning allows for efficient and clean production of electricity (and heat) from biomass. On a scale of about 50 to 80 MWe, electrical efficiencies of 30 to 40% are possible. Finland is on cutting edge of the field with development and deployment of BFB and CFB boilers with high fuel flexibility, low costs, high efficiency, and deployed at large scale. One of the latest plants realized in Finland has a capacity of 500 MWe and is co-fired with a portfolio of biomass fuels, partly supplied by water transport.

Thailand and Malaysia are two countries in Asia that are active in deploying state-of-the-art combustion technology (typically plants of 20 to 40 MWe) for utilizing forest, plantation, and agricultural residues. The potential in the southeast Asian region for such projects is large and fits in the rapid growth of the electricity demand in the region giving ample room for investments.

Two large industrial sectors offer excellent opportunities to use available biomass resources efficiently and competitively worldwide. Those are the paper and pulp industry and the sugar-industry (particularly the one using sugar cane as feed). Traditionally, those sectors have used biomass residues (wood waste and bagasse) for their internal energy needs, which usually implies inefficient conversion to low pressure steam and some power. Plants are usually laid out to meet internal power needs only. Older plants often use process steam to drive steam

turbines for direct mechanical drive of equipment (such as the sugar mills). Although in general the power production potential is much larger, most plants are not equipped with power generation units to export a surplus power. This situation is particularly caused by the absence of regulations that ensure reasonable electricity tariffs for independent power producers, thus making it unattractive for industries to invest in more efficient power generation capacity. However, liberalization of the energy markets in many countries removed this barrier (or is close to implementation). Combined with a need to reduce production costs and the need for modernization of, often old, production capacity, this provides a window of opportunity for these sectors worldwide.

In the sugar industry, the improved use of bagasse for energy production in sugar mills has gained some momentum. In Africa, Asia (e.g., through a World Bank supported project in Indonesia), and Latin America, retrofits have been carried out or are considered at many mills. Retrofits generally include the installation of efficient boilers and in some cases energy efficiency measures in the mills themselves. In combination, such retrofits could easily double the electricity output of a sugar mill, although projects are always site and plant specific.

Power output and exports can be considerably increased at low investment costs. As examples, in Nicaragua, electricity production from available bagasse by using improved boilers could meet the national electricity demand. Countries like Cuba, Brazil, and others offer similar opportunities. The other key sector applying biomass combustion for power generation is the paper-and-pulp industry. Combustion of black liquor, a main residue from the pulping process containing various chemicals, and, to a lesser extent, wood residues are the key fuels. Conventional boilers for combined production of power and process steams and recovery of pulping chemicals (so-called Tomlinson boilers) constitute common technology for the paper-and-pulp sector. More advanced boilers or gasification technology (BIG/CC) could offer even further efficiency gains and lower costs, such as when applied for converting black liquor, since recovery of pulping chemicals is easier. Generally, the power generation is competitive to grid prices, since the biomass is available anyway.

4.4 Waste Incineration

Waste incinerators, combined with very stringent emission standards, were widely deployed starting in the 1980s in countries like Germany and the Netherlands. Typical technologies deployed are large-scale (i.e., around 1 Mtonne capacity per plant per year) moving grate boilers (which allow mass burning of very diverse waste properties), low steam pressures and temperatures (to avoid corrosion), and extensive flue gas cleaning. Typical electrical efficiencies are between 15 to over 20%. Mass burning became the key waste-to-energy technology deployed in Europe, but it is also relatively expensive with treatment costs in the range of 50 to 150 Euro/tonne (off-set by tipping fees). Emission standards have a strong impact on the total waste treatments costs. In the Netherlands, flue gas cleaning equipment represents about half of the capital costs of a typical waste to energy plant. The more advanced BFB and CFB technology may prove to be a more competitive alternative in the short to medium term, because emission profiles are inherently better. On the other hand, more complex pretreatment (e.g., production of refuse derived fuel from municipal solid waste, basically an upgrading step) is required.

4.5 Co-combustion

Co-combustion of biomass, in particular in pulverized coal-fired power plants, is the single largest growing conversion route for biomass in many OECD countries (e.g., in the United States, Scandinavia, Spain, Germany, and the Netherlands, to name a few). Generally, relatively low co-firing shares are deployed with limited consequences for boiler performance and maintenance. For some countries, biomass co-firing also serves as a route to avoid the need for investing in additional flue gas cleaning equipment that would be required by national emission standards for 100% coal firing.

Because many plants in mentioned countries are now equipped with some co-firing capacity, interest for higher co-firing shares (e.g., up to 40%) is rising. Technical consequences for feeding lines and boiler performance, for example, as well as ash compositions (ash is generally used and changing characteristics can cause problems in established markets) are more severe though and development efforts focus on those issues. This is a situation where gasification becomes attractive as co-firing option, since feeding fuel gas avoids many of aforementioned consequences. Two examples of such configurations are Lahti in Finland (60 MWth gasification capacity for a 200 MWe and 250 MWth CHP plant, using natural gas, biomass, and other solid fuels) and AmerGas in the Netherlands (over 80 MWth gasification capacity for a 600 MWe coal-fired power station). Power

plants capable of firing, for example, natural gas or coal with various biomass streams, have been built in Denmark (e.g., the Avedore plant) and Finland with the benefit of gaining economies of scale as well as reduced fuel supply risks. In Denmark straw is a common fuel. The chlorine and alkaline-rich straw caused problems in conventional combustion systems through increased corrosion and slagging. In multi-fuel systems, however, straw can be used for raising low-temperature steam, after which the steam is superheated by fossil fuels. This approach nearly eliminates the problems mentioned.

5. FUTURE OUTLOOK AND CLOSING REMARKS

To date, the use of biomass for energy production in the world (which is considerable with some 45 EJ) is dominated by combustion. In terms of net energy produced, gasification, pyrolysis, fermentation (for ethanol production), extraction (e.g., biodiesel), and other options play a marginal role. Although various technological options mentioned may have better economic and efficiency perspectives in the long term, in general, it is likely that biomass combustion will have a major role to play for a long time to come. Currently, biomass combustion is the workhorse for biomass as a primary and renewable energy source. Looking ahead, many technological improvements can be deployed and it is likely the principle of technological learning will lead to further improvements in energy efficiency, environmental performance, reliability, and flexibility and economy.

Better stoves in developing countries can have a tremendous positive impact on both efficiency improvements of wood fuel use as well as improving indoor health quality and resulting improvement of health, in particular for women who are generally responsible for preparing meals and maintaining the household. Simply reducing the need for fuel wood collection can save large amounts of time in many situations and adds to increased well-being and welfare as well.

Possibilities for further improvements of economy, energy efficiency, emissions, and availability for biomass combustion at larger scales include optimized and novel boiler designs lowering emissions, better materials, such as for steam generators and new power-generation concepts (closed turbine, etc.), and Stirling engines. Economies of scale have been discussed as a key development, as well as advanced co-firing concepts. With respect to co-firing, the distinction between combustion and gasification is becoming less relevant, since biomass gasifiers are now successfully deployed to deliver fuel gas to larger boilers, which are fired with coal or other fuels. Furthermore, the organization and optimization of the entire fuel supply chain (including combining different fuels and biomass pretreatment in direct relation to the conversion technology) has improvement potential in most current markets.

Clearly, new, advanced technologies will replace the older generation combustion technology, both in domestic and industrial markets. Furthermore, specific markets (such as those that affect fuels, economic conditions, and the energy system) require specific solutions, and dedicated designs are required to find optimal solutions. Another key trend, as described, is the increasing scale of bioenergy projects and the bioenergy market. The increased availability of biomass, which is also more expensive, requires conversion units to be more efficient and reach lower cost. Deploying larger scale conversion units is therefore a likely continuing trend, which is already observed in various markets. In addition, advanced co-firing systems have a key role to play, also a trend from the 1990s. Increased fuel flexibility, reducing supply risks, and allowing for cost minimization while at the same time allow for gaining economies of scale are important pathways for bioenergy as a whole.

In turn, biomass markets become increasingly intertwined, and specific national approaches (which have led to many innovations over time) may be replaced by increased international collaboration between countries and companies, resulting in products and systems for the world market. Nevertheless, new and improved technologies in the field of gasification (such as biomass integrated gasification combined cycle [BIG/CC] technology) and production of biofuels (such as methanol, fischer-tropsch liquids, and hydrogen, as well as bio-ethanol production from ligno-cellulosic biomass) have the potential over the long term to make the use of (more expensive and cultivated) biomass competitive or at least more attractive in efficiency and economic terms than direct combustion for power and heat. This also implies increased competition among different options. At this stage, it is likely various technologies will have their own specific markets and development trajectories for some time to come.

Furthermore, developments in energy systems as a whole, like increased deployment of renewables (such as wind energy), efficient use of low carbon fuels (natural gas), or CO_2 removal and storage can affect the attractiveness of biomass use for power

generation once power production becomes less carbon intensive. Similar reasoning holds for increased use of biomaterials, which may have the combined benefit of saving fossil fuels in production of reference materials (e.g., plastics from mineral oil versus bioplastics) and which can be converted to energy when turned to waste. Over time, this may shift the attractiveness of applying biomass for heat and power to production of transportation fuels or bio-materials.

SEE ALSO THE FOLLOWING ARTICLES

Biodiesel Fuels • Biomass, Chemicals from • Biomass for Renewable Energy and Fuels • Biomass Gasification • Biomass: Impact on Carbon Cycle and Greenhouse Gas Emissions • Biomass Resource Assessment • Combustion and Thermochemistry • Peat Resources • Renewable Energy, Taxonomic Overview • Waste-to-Energy Technology • Wood Energy, History of

Further Reading

The Bio-energy Agreement of the International Energy Agency Web site, found at www.ieabioenergy.com, and more specifically the Task on Biomass Combustion and Co-firing, found at www.ieabioenergy-task32.com.

Broek, R. van den, Faaij, A., and van Wijk, A. (1996). Biomass combustion for power generation. *Biomass & Bioenergy* **11**(4), 271–281.

Dornburg, V., and Faaij, A. (2001). Efficiency and economy of wood-fired biomass energy systems in relation to scale regarding heat and power generation using combustion and gasification technologies. *Biomass & Bioenergy* Vol. 21(2), 91–108.

Faaij, A. (2003). Bio-energy in Europe: Changing technology choices. Prepared for a special of the *J. Energy Policy on Renewable Energy in Europe,* August.

Faaij, A., van Wijk, A., van Doorn, J., Curvers, A., Waldheim, L., Olsson, E., and Daey-Ouwens, C. (1997). Characteristics and availability of biomass waste and residues in the netherlands for gasification. *Biomass and Bioenergy* **12**(4), 225–240.

Harmelinck, M., Voogt, M., Joosen, S., de Jager, D., Palmers, G., Shaw, S., and Cremer, C. (2002). "PRETIR, Implementation of Renewable Energy in the European Union until 2010." Report executed within the framework of the ALTENER program of the European Commission, DG-TREN. ECOFYS BV, 3E, Fraunhofer-ISI, Utrecht, The Netherlands, plus various country reports.

Hillring, B. (2002). Rural development and bioenergy—experiences from 20 years of development in sweden. *Biomass and Bioenergy* **23**(6), 443–451.

Hoogwijk, M., Faaij, A., van den Broek, R., Berndes, G., Gielen, D., and Turkenburg, W. Exploration of the ranges of the global potential of biomass for energy. *Biomass and Bioenergy* **25**(2), 119–133.

Loo van, S., and Koppejan, J. (eds.). (2002). "Handbook Biomass Combustion and Co-firing." Twente University Press, Enschede, The Netherlands.

Nikolaisen, L., *et al.* (eds.). (1998). Straw for energy production. Centre for Biomass Technology, Denmark, available at www.videncenter.dk.

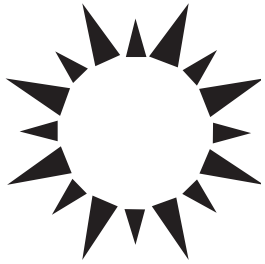
Serup, H., *et al.* (eds.) (1999). Wood for energy production. Centre for Biomass Technology, Denmark, available at www.videncenter.dk.

Solantausta, Y., Bridgewater, T., and Beckman, D. (1996). "Electricity Production by Advanced Biomass Power Systems." VTT Technical Research Centre of Finland, Espoo, Finland (report no. 1729).

Turkenburg, W. C., and Faaij, A., *et al.* (2001). Renewable energy technologies. "World Energy Assessment of the United Nations," UNDP, UNDESA/WEC, Chapter 7, September. UNDP, New York.

U.S. Department of Energy, Office of Utility Technologies (1998). "Renewable Energy Technology Characterizations." Washington, DC, January.

Biomass for Renewable Energy and Fuels



DONALD L. KLASS
Entech International, Inc.
Barrington, Illinois, United States

1. **Fundamentals**
2. **Biomass Conversion Technologies**
3. **Commercial Biomass Energy Markets and Economics**
4. **Environmental Impacts**

Glossary

barrels of oil equivalent (boe) The total energy content of a non-petroleum-based product or fuel in GJ divided by 5.904–6.115 GJ/boe.
biodiesel The methyl or ethyl esters of transesterified triglycerides (lipids, fats, cooking greases) from biomass.
biofuel A solid, gaseous, or liquid fuel produced from biomass.
biogas A medium-energy-content gaseous fuel, generally containing 40 to 80 volume percent methane, produced from biomass by methane fermentation (anaerobic digestion).
biomass All non-fossil-based living or dead organisms and organic materials that have an intrinsic chemical energy content.
biorefinery A processing plant for converting waste and virgin biomass feedstocks to energy, fuels, and other products.
gasohol A blend of 10 volume percent ethanol and 90 volume percent gasoline.
independent power producer (IPP) A nonutility generator of electricity, usually produced in a small capacity plant or industrial facility.
integrated biomass production conversion system (IBPCS) A system in which all operations concerned with the production of virgin biomass feedstocks and their conversion to energy, fuels, or chemicals are integrated.
landfill gas (LFG) A medium-energy-content fuel gas high in methane and carbon dioxide produced by landfills that contain municipal solid wastes and other waste biomass.
methyl *t*-butyl ether (MTBE) An organic compound used as an oxygenate and octane-enhancing additive in motor gasolines.
oxygenated gasoline Gasolines that contain soluble oxygen-containing organic compounds such as fuel ethanol and MTBE.
quad One quad is 10^{15} (1 quadrillion) Btu.
refuse-derived fuel (RDF) The combustible portion of municipal solid wastes.
tonnes of oil equivalent (toe) The total energy content of a non-petroleum-based product or fuel in GJ divided by 43.395–44.945 GJ/toe.

The world's energy markets rely heavily on the fossil fuels coal, petroleum crude oil, and natural gas as sources of thermal energy; gaseous, liquid, and solid fuels; and chemicals. Since millions of years are required to form fossil fuels in the earth, their reserves are finite and subject to depletion as they are consumed. The only natural, renewable carbon resource known that is large enough to be used as a substitute for fossil fuels is biomass. Included are all water- and land-based organisms, vegetation, and trees, or virgin biomass, and all dead and waste biomass such as municipal solid waste (MSW), biosolids (sewage) and animal wastes (manures) and residues, forestry and agricultural residues, and certain types of industrial wastes. Unlike fossil fuel deposits, biomass is renewable in the sense that only a short period of time is needed to replace what is used as an energy resource.

1. FUNDAMENTALS

1.1 The Concept

The capture of solar energy as fixed carbon in biomass via photosynthesis, during which carbon dioxide (CO_2) is converted to organic compounds, is the key initial step in the growth of virgin biomass

 © 2004 Elsevier Inc. All rights reserved.

and is depicted by the following equation:

$$CO_2 + H_2O + \text{light} + \text{chlorophyll} \rightarrow (CH_2O) + O_2. \quad (1)$$

Carbohydrate, represented by the building block (CH_2O), is the primary organic product. For each gram mole of carbon fixed, about 470 kJ (112 kcal) is absorbed.

The upper limit of the capture efficiency of the incident solar radiation in biomass has been estimated to range from about 8% to as high as 15%, but under most conditions in the field, it is generally less than 2% as shown in Table I. This table also lists the average annual yields on a dry basis and the average insolation that produced these yields for a few representative biomass species.

The global energy potential of virgin biomass is very large. It is estimated that the world's standing terrestrial biomass carbon (i.e., the renewable, above-ground biomass that could be harvested and used as an energy resource) is approximately 100 times the world's total annual energy consumption. The largest source of standing terrestrial biomass carbon is forest biomass, which contains about 80 to 90% of the total biomass carbon (Table II). Interestingly, marine biomass carbon is projected to be next after the forest biomass carbon in terms of net annual production, but is last in terms of availability because of its high turnover rates in an oceanic environment.

The main features of how biomass is used as a source of energy and fuels are schematically illustrated in Fig. 1. Conventionally, biomass is harvested for feed, food, fiber, and materials of construction or is left in the growth areas where natural decomposition occurs. The decomposing biomass or the waste products from the harvesting and processing of biomass, if disposed on or in land, can in theory be partially recovered after a long period of time as fossil fuels. This is indicated by the dashed lines in the figure. The energy content of biomass could be diverted instead to direct heating applications by collection and combustion. Alternatively, biomass and any wastes that result from its processing or consumption could be converted directly into synthetic organic fuels if suitable conversion processes were available. Another route to energy products is to grow certain species of biomass such as the rubber tree *(Hevea braziliensis),* in which high-energy hydrocarbons are formed within the species by natural biochemical mechanisms, or the Chinese tallow tree *(Sapium sebiferum),* which affords high-energy triglycerides in a similar manner. In these cases, biomass serves the dual role of a carbon-fixing apparatus and a continuous source of high-energy organic products without being consumed in the process. Other biomass species, such as the herbaceous guayule bush *(Parthenium argentatum)* and the gopher plant *(Euphorbia lathyris),* produce hydrocarbons too, but must be harvested to recover them. Conceptually, Fig. 1 shows that there are several pathways by which energy products and synthetic fuels can be manufactured.

Another approach to the development of fixed carbon supplies from renewable carbon resources is to convert CO_2 outside the biomass species to synthetic fuels and organic intermediates. The ambient air, which contains about 360 ppm by volume of CO_2, the dissolved CO_2 and carbonates in the oceans, and the earth's large carbonate deposits, could serve as renewable carbon resources. But since CO_2 is the final oxidation state of fixed carbon, it contains no chemical energy. Energy must be supplied in a chemical reduction step. A convenient method of supplying the required energy and of simultaneously reducing the oxidation state is to reduce CO_2 with hydrogen. The end product, for example, can be methane (CH_4), the dominant component in natural gas and the simplest hydrocarbon known, or other organic compounds. With all components in the ideal gas state, the standard

$$CO_2 + 4H_2 \rightarrow CH_4 + H_2O \quad (2)$$

enthalpy of the process is exothermic by −165 EJ (−39.4 kcal) per gram mole of methane formed. Biomass can also serve as the original source of hydrogen via partial oxidation or steam reforming to yield an intermediate hydrogen-containing product gas. Hydrogen would then effectively act as an energy carrier from the biomass to CO_2 to yield a substitute or synthetic natural gas (SNG). The production of other synthetic organic fuels can be carried out in a similar manner. For example, synthesis gas (syngas) is a mixture of hydrogen and carbon oxides. It can be produced by biomass gasification processes for subsequent conversion to a wide range of chemicals and fuels as illustrated in Fig. 2. Other renewable sources of hydrogen can also be utilized. These include continuous water splitting by electrochemical, biochemical, thermochemical, microbial, photolytic, and biophotolytic processes.

The basic concept then of using biomass as a renewable energy resource consists of the capture of solar energy and carbon from ambient CO_2 in growing biomass, which is converted to other fuels (biofuels, synfuels, hydrogen) or is used directly as a

TABLE I

Examples of Biomass Productivity and Estimated Solar Energy Capture Efficiency

Location	Biomass community	Annual yield dry matter (t/ha-year)	Average insolation (W/m^2)	Solar energy capture efficiency (%)
Alabama	Johnsongrass	5.9	186	0.19
Sweden	Enthrophic lake angiosperm	7.2	106	0.38
Denmark	Phytoplankton	8.6	133	0.36
Minnesota	Willow and hybrid poplar	8–11	159	0.30–0.41
Mississippi	Water hyacinth	11.0–33.0	194	0.31–0.94
California	*Euphorbia lathyris*	16.3–19.3	212	0.45–0.54
Texas	Switchgrass	8–20	212	0.22–0.56
Alabama	Switchgrass	8.2	186	0.26
Texas	Sweet sorghum	22.2–40.0	239	0.55–0.99
Minnesota	Maize	24.0	169	0.79
New Zealand	Temperate grassland	29.1	159	1.02
West Indies	Tropical marine angiosperm	30.3	212	0.79
Nova Scotia	Sublittoral seaweed	32.1	133	1.34
Georgia	Subtropical saltmarsh	32.1	194	0.92
England	Coniferous forest, 0-21 years	34.1	106	1.79
Israel	Maize	34.1	239	0.79
New South Wales	Rice	35.0	186	1.04
Congo	Tree plantation	36.1	212	0.95
Holland	Maize, rye, two harvests	37.0	106	1.94
Marshall Islands	Green algae	39.0	212	1.02
Germany	Temperate reedswamp	46.0	133	1.92
Puerto Rico	*Panicum maximum*	48.9	212	1.28
California	Algae, sewage pond	49.3–74.2	218	1.26–1.89
Colombia	Pangola grass	50.2	186	1.50
West Indies	Tropical forest, mixed ages	59.0	212	1.55
Hawaii	Sugarcane	74.9	186	2.24
Puerto Rico	*Pennisetum purpurcum*	84.5	212	2.21
Java	Sugarcane	86.8	186	2.59
Puerto Rico	Napier grass	106	212	2.78
Thailand	Green algae	164	186	4.90

Note. Insolation capture efficiency calculated by author from dry matter yield data of Berguson, W., *et al*. (1990). "Energy from biomass and Wastes XIII" (Donald L. Klass, Ed.). Institute of Gas Technology, Chicago; Bransby, D. I., and Sladden, S. E. (1991). "Energy from Biomass and Wastes XV" (Donald L. Klass, Ed.). Institute of Gas Technology, Chicago; Burlew, J. S. (1953). "Algae Culture from Laboratory to Pilot Plant," Publication 600. Carnegie Institute of Washington, Washington, DC; Cooper, J. P. (1970). "Herb." *Abstr. m*, **40**, 1; Lipinsky, E. S. (1978). "Second Annual Fuels from Biomass Symposium" (W. W. Shuster, Ed.), p. 109. Rensselaer Polytechnic Institute, Troy, New York; Loomis, R. S., and Williams, W. A. (1963). *Crop. Sci.* **3**, 63; Loomis, R. S., Williams, W. A., and Hall, A. E. (1971). *Ann. Rev. Plant Physiol.* **22**, 431; Rodin, I. E., and Brazilevich, N. I. (1967). "Production and Mineral Cycling in Terrestrial Vegetation." Oliver & Boyd, Edinburgh, Scotland; Sachs, R. M., *et al*. (1981). *Calif. Agric.* **29**, July/August; Sanderson, M. A., *et al*. (1995). "Second Biomass Conference of the Americas," pp. 253–260. National Renewable Energy Laboratory, Golden, CO; Schneider, T. R. (1973). *Energy Convers*. 13, 77; and Westlake, D. F. (1963). *Biol. Rev.* **38**, 385.

source of thermal energy or is converted to chemicals or chemical intermediates.

The idea of using renewable biomass as a substitute for fossil fuels is not new. In the mid-1800s, biomass, principally woody biomass, supplied over 90% of U.S. energy and fuel needs, after which biomass energy usage began to decrease as fossil fuels became the preferred energy resources. Since the First Oil Shock of 1973–1974, the commercial utilization of biomass energy and fuels has increased slowly but steadily. The contribution of biomass energy to U.S. energy consumption in the late 1970s was more than

TABLE II

Estimated Distribution of World's Biomass Carbon

	Forests	Savanna and grasslands	Swamp and marsh	Remaining terrestrial	Marine
Area (10^6 km^2)	48.5	24.0	2.0	74.5	361
Percentage	9.5	4.7	0.4	14.6	70.8
Net C production (Gt/year)	33.26	8.51	2.70	8.40	24.62
Percentage	42.9	11.0	3.5	10.8	31.8
Standing C (Gt)	744	33.5	14.0	37.5	4.5
Percentage	89.3	4.0	1.7	4.5	0.5

Note. Adapted from Table 2.2 in Klass, D. L. (1998). "Biomass for Renewable Energy, Fuels, and Chemicals." Academic Press, San Diego, CA.

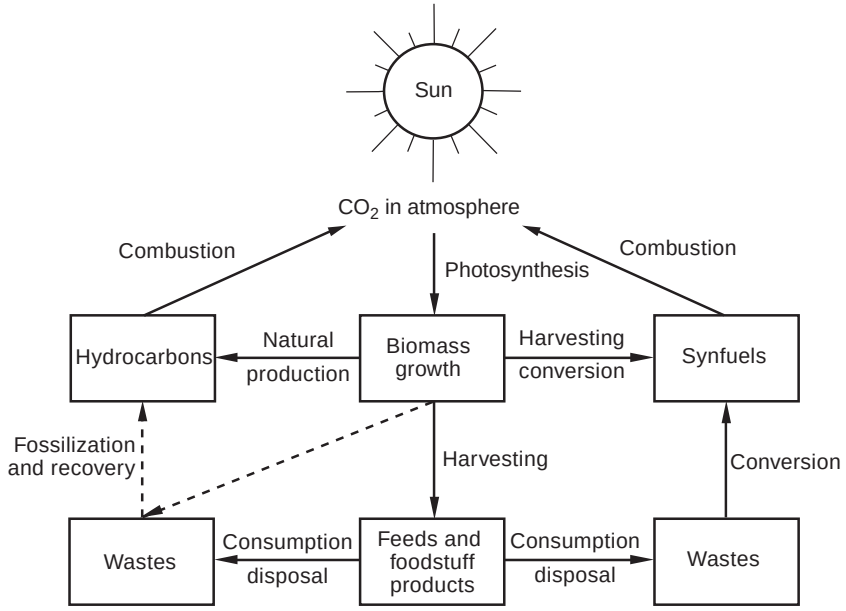


FIGURE 1 Main features of biomass energy technology. From Klass (1998).

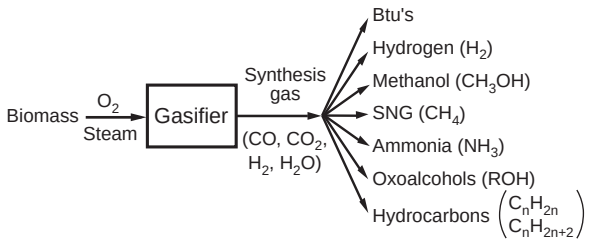


FIGURE 2 Chemicals from syngas by established processes. From Klass (1998).

850,000 barrels of oil equivalent per day (boe/day), or more than 2% of total primary energy consumption at that time. In 1990, when total U.S. primary energy consumption was about 88.9 EJ (84.3 quad), virgin and waste biomass resources contributed about 3.3% to U.S. primary energy demand at a rate of about 1.4 Mboe/day, as shown in Table III. By 2000, when total primary energy consumption had increased to 104.1 EJ (98.8 quad), virgin and waste biomass resources contributed about 23% more to primary energy demand, 1.60 Mboe/day, although the overall percentage contribution was about the same as in 1990 (Table IV).

According to the United Nations, biomass energy consumption was about 6.7% of the world's total energy consumption in 1990. For 2000, the data compiled by the International Energy Agency (IEA) from a survey of 133 countries indicate that biomass' share of total energy consumption, 430 EJ (408 quad), for these countries is about 10.5% (Table V). Although the IEA cautions that the quality and

TABLE III

Consumption of Biomass Energy in United States in 1990

Biomass resource	EJ/year	boe/day
Wood and wood wastes		
Industrial sector	1.646	763,900
Residential sector	0.828	384,300
Commercial sector	0.023	10,700
Utilities	0.013	6,000
Total	2.510	1,164,900
Municipal solid wastes	0.304	141,100
Agricultural and industrial wastes	0.040	18,600
Methane		
Landfill gas	0.033	15,300
Biological gasification	0.003	1,400
Thermal gasification	0.001	500
Total	0.037	17,200
Transportation fuels		
Ethanol	0.063	29,200
Other biofuels	0	0
Total	0.063	29,200
Grand total	2.954	1,371,000
Percentage primary energy consumption	3.3	

Note. From Klass, D. L. (1990). *Chemtech* **20**(12), 720–731; and U.S. Department of Energy (1991). "Estimates of U.S. Biofuels Consumption DOE/EIA-0548," October. Energy Information Administration, Washington, DC.

TABLE IV

Consumption of Biomass Energy in United States in 2000

Biomass resource	EJ/year	boe/day
Wood	2.737	1,270,100
Waste	0.570	264,800
Alcohol fuels	0.147	68,000
Total	3.454	1,602,900

Note. Adapted from Energy Information Administration (2002). *Monthly Energy Review,* August, Table 10.1. Washington, DC. Wood consists of wood, wood waste, black liquor, red liquor, spent sulfite liquor, wood sludge, peat, railroad ties, and utility poles. Waste consists of MSW, LFG, digester gas, liquid acetonitrile waste, tall oil, waste alcohol, medical waste, paper pellets, sludge waste, solid by products, tires, agricultural by-products, closed-loop biomass, fish oil, and straw. Alcohol fuels consist of ethanol blended into motor gasoline.

reliability of the data they compiled on biomass may be limited, which makes comparison between countries difficult, and that the proper breakdown between renewables and nonrenewables is often not available, it is clear that a significant portion of global energy consumption is based on biomass resources. It is also evident that the largest biomass energy consumption occurs among both industrialized and developing countries. Some countries meet a large percentage of their energy demands with biomass resources such as Sweden, 17.5%; Finland, 20.4%; Brazil, 23.4%; while many other countries in South America, Africa, and the Far East use biomass energy resources that supply much higher percentages of total energy demand. As expected, most countries in the Middle East where large proved crude oil and natural gas reserves are located and where dedicated energy crops might be difficult to grow meet their energy and fuel needs without large contributions from biomass resources.

The IEA reports that the share of energy consumption for renewables in 2000 was 13.8% of total energy consumption, of which 79.8% is combustible renewables and waste, most of which is biomass, and that the balance of 20.2% consists of hydroelectric power, 16.5%, and 3.7% other renewables.

Despite some of the inconsistencies that can occur because of data reliability, specific comparisons for total and biomass energy consumption in Table VI, and for total electricity and biomass-based electricity generation in Table VII, are shown for the eleven largest energy-consuming countries including the United States. The data for seven countries in these tables are for 2000; data for four countries are for 1999. The United States is the largest energy consumer, but China and India are the largest biomass energy consumers. Primary biomass solids as described in the footnote for Table VI are the largest biomass resources for these countries as well as the other countries listed in this table. In the case of electricity generation from biomass, the United States, Japan, and Germany are the largest producers of the 11 countries listed in Table VII, and the biomass resource most utilized is primary biomass solids for the United States and Japan, while Germany uses much less of that resource. It is surprising that China and India are each reported to use "0" biogas for power generation, since it is well known that each these countries operate millions of small-scale and farm-scale methane fermentation units, while many major urban cities utilize the high-methane fuel gas produced during wastewater treatment by anaerobic digestion. The lack of data is probably the cause of this apparent inconsistency.

It is noteworthy that some energy analysts have predicted that the end of seemingly unlimited petroleum crude oil and natural gas resources is in

TABLE V

Total Energy Consumption and Biomass' Share of Total Consumption for 133 Countries in 2000

		Total consumption		Biomass' share total consumption		
Region	Country	(Mtoe)	(EJ/year)	(%)	(EJ/year)	(Mboe/day)
North America	Canada	251	10.9	4.5	0.491	
	Cuba	13.2	0.573	21.1	0.121	
	Dominican Republic	7.8	0.34	17.4	0.059	
	Haiti	2	0.09	75.4	0.068	
	Jamaica	3.9	1.5	12.1	0.182	
	Mexico	153.5	6.664	5.2	0.347	
	Panama	2.5	0.11	18.1	0.020	
	Trinidad and Tobago	8.7	0.38	0.4	0.002	
	United States	2300	99.85	3.4	3.395	
	Subtotal:	2742.6	120.41	3.9	4.685	2.100
South America	Argentina	61.5	2.67	4.4	0.117	
	Bolivia	4.9	0.21	14.7	0.031	
	Brazil	183.2	7.953	23.4	1.861	
	Chile	24.4	1.06	17.4	0.184	
	Colombia	28.8	1.25	18.3	0.229	
	Ecuador	8.2	0.36	8.5	0.031	
	El Salvador	4.1	0.18	34.0	0.061	
	Guatemala	7.1	0.31	54.5	0.169	
	Honduras	3	0.1	44.0	0.044	
	Netherlands Antilles	1.1	0.048	0.0	0.000	
	Nicaragua	2.7	0.12	51.6	0.062	
	Paraguay	3.9	0.17	58.2	0.099	
	Peru	12.7	0.551	17.6	0.097	
	Uruguay	3.1	0.13	13.7	0.018	
	Venezuela	59.3	2.57	0.9	0.023	
	Subtotal:	408.0	17.7	17.7	3.026	1.356
Europe	Albania	1.6	0.069	3.6	0.002	
	Austria	28.6	1.24	10.9	0.135	
	Belgium	59.2	2.57	1.2	0.031	
	Bosnia and Herzegovina	4.4	0.19	4.2	0.008	
	Bulgaria	18.8	0.816	3.1	0.025	
	Croatia	7.8	0.34	4.8	0.016	
	Cyprus	2.4	0.10	0.4	0.000	
	Czech Republic	40.4	1.75	1.5	0.026	
	Denmark	19.5	0.847	8.8	0.075	
	Finland	33.1	1.44	20.4	0.294	
	France	257.1	11.16	4.5	0.502	
	Germany	339.6	14.74	2.5	0.369	
	Gibraltar	0.2	0.009	0.0	0.000	
	Greece	27.8	1.21	3.7	0.045	
	Hungary	24.8	1.08	1.5	0.016	
	Iceland	3.4	0.15	0.0	0.000	
	Ireland	14.6	0.634	1.2	0.008	
	Italy	171.6	7.449	4.9	0.365	
	Luxembourg	3.7	0.16	0.8	0.001	

continues

Table V continued

Region	Country	Total consumption (Mtoe)	Total consumption (EJ/year)	Biomass' share total consumption (%)	Biomass' share total consumption (EJ/year)	Biomass' share total consumption (Mboe/day)
	Macedonia	2.8	0.12	7.7	0.009	
	Malta	0.8	0.03	0.0	0.000	
	Netherlands	75.8	3.29	2.3	0.076	
	Norway	25.6	1.11	5.3	0.059	
	Poland	90	3.9	4.5	0.176	
	Portugal	24.6	1.07	8.3	0.089	
	Romania	36.3	1.58	7.9	0.125	
	Russia	614	26.7	1.1	0.294	
	Slovak Republic	17.5	0.760	0.5	0.004	
	Slovenia	6.5	0.28	6.5	0.018	
	Spain	124.9	5.422	3.6	0.195	
	Sweden	47.5	2.06	17.5	0.361	
	Switzerland	26.6	1.15	6.0	0.069	
	United Kingdom	232.6	10.10	0.9	0.091	
	Fed. Rep. of Yugoslavia	13.7	0.595	1.8	0.011	
	Former Yugoslavia	35.1	1.52	1.8	0.027	
	Subtotal:	2432.9	105.6	3.3	3.522	1.579
Former USSR	Armenia	2.1	0.091	0.0	0.000	
	Azerbaijan	11.7	0.508	0.1	0.001	
	Belarus	24.3	1.05	4.1	0.043	
	Estonia	4.5	0.20	11.1	0.022	
	Georgia	2.9	0.13	2.5	0.003	
	Kazakhstan	39.1	1.70	0.1	0.002	
	Kyrgystan	2.4	0.10	0.2	0.000	
	Latvia	3.7	0.16	22.4	0.036	
	Lithuania	7.1	0.31	8.7	0.027	
	Republic of Moldova	2.9	0.13	2.0	0.003	
	Tajikistan	2.9	0.13	0.0	0.000	
	Turkmenistan	13.9	0.169	0.0	0.000	
	Ukraine	139.6	6.060	0.2	0.012	
	Uzbekistan	50.2	2.18	0.0	0.000	
	Subtotal:	307.3	13.34	1.1	0.149	0.067
Africa	Algeria	29.1	1.26	0.3	0.004	
	Angola	7.7	0.33	74.5	0.246	
	Benin	2.4	0.10	75.5	0.076	
	Cameroon	6.4	0.28	78.4	0.220	
	Congo	0.9	0.04	65.6	0.026	
	Cote d'lvoire	6.9	0.30	60.9	0.183	
	Egypt	46.4	2.01	2.8	0.056	
	Eritrea	0.7	0.03	70.9	0.021	
	Ethiopia	18.7	0.812	93.1	0.756	
	Gabon	1.6	0.072	59.2	0.043	
	Ghana	7.7	0.33	68.8	0.227	
	Kenya	15.5	0.239	76.1	0.182	
	Libya	16.4	0.712	0.8	0.006	
	Morocco	10.3	0.447	4.3	0.019	
	Mozambique	7.1	0.31	92.7	0.287	

continues

Table V continued

Region	Country	Total consumption (Mtoe)	Total consumption (EJ/year)	Biomass' share total consumption (%)	Biomass' share total consumption (EJ/year)	Biomass' share total consumption (Mboe/day)
	Namibiae	1	0.04	16.8	0.007	
	Nigeria	90.2	3.92	80.2	3.144	
	Senegal	3.1	0.13	55.8	0.073	
	South Africa	107.6	4.671	11.6	0.542	
	Sudan	16.2	0.703	86.9	0.611	
	United Rep. of Tanzania	15.4	0.669	93.6	0.626	
	Togo	1.5	0.065	67.7	0.044	
	Tunisia	7.9	0.34	15.7	0.053	
	Zambia	6.2	0.27	82.2	0.222	
	Zimbabwe	10.2	0.443	54.8	0.243	
	Subtotal:	437.1	19.00	41.7	7.917	3.548
Middle East	Bahrain	6.4	0.28	0.0	0.000	
	Iran	112.7	4.893	0.7	0.034	
	Iraq	27.7	1.20	0.1	0.001	
	Israel	20.2	0.877	0.0	0.000	
	Jordan	5.2	0.23	0.1	0.000	
	Kuwait	20.9	0.907	0.0	0.000	
	Lebanon	5.1	0.22	2.5	0.006	
	Oman	9.8	0.43	0.0	0.000	
	Qatar	15.7	0.682	0.0	0.000	
	Saudi Arabia	105.3	4.571	0.0	0.000	
	Syria	18.4	0.799	0.1	0.001	
	Turkey	77.1	3.35	8.4	0.281	
	United Arab Emirates	29.6	1.28	0.1	0.001	
	Yemen	3.5	0.15	2.2	0.003	
	Subtotal:	457.6	19.87	1.6	0.327	0.147
Far East	Bangladesh	18.7	0.812	40.8	0.331	
	Brunei	2	0.09	0.9	0.001	
	China	1142	49.58	18.7	9.271	
	Taiwan	83	3.6	0.0	0.000	
	Hong Kong (China)	15.5	0.673	0.3	0.002	
	India	501.9	21.79	40.2	8.760	
	Indonesia	145.6	6.321	32.6	2.061	
	Japan	524.7	22.78	1.1	0.251	
	North Korea	46.1	2.00	1.1	0.022	
	South Korea	193.6	8.405	5.8	0.487	
	Malaysia	49.5	2.15	5.1	0.110	
	Myanmar	12.5	0.543	73.3	0.398	
	Nepal	7.9	0.34	85.2	0.290	
	Pakistan	64	2.8	37.6	1.053	
	Philippines	42.4	1.84	22.5	0.414	
	Singapore	24.6	1.07	0.0	0.000	
	Sri Lanka	8.1	0.35	52.8	0.185	
	Thailand	73.6	3.20	19.4	0.621	
	Vietnam	37	1.6	61.2	0.979	
	Subtotal:	2992.7	129.92	19.4	25.236	11.311

continues

Table V continued

Region	Country	Total consumption (Mtoe)	Total consumption (EJ/year)	Biomass' share total consumption (%)	Biomass' share total consumption (EJ/year)	Biomass' share total consumption (Mboe/day)
Oceania	Australia	110.2	4.784	4.9	0.234	
	New Zealand	18.6	0.807	6.5	0.052	
	Subtotal:	128.8	5.591	5.1	0.286	0.128
	Total:	9907	430.1	10.5	45.148	20.236

Note. Total energy consumption in Mtoe for each country listed was compiled by the International Energy Agency (2002). IEA's data for total energy consumption were converted to EJ/year (for 2000) in this table using a multiplier of 0.043412. The multiplier for converting EJ/year to Mboe/day is 0.4482×10^6. For each country, the IEA reported the total share of renewables as a percentage of the total consumption and as a percentage of the total consumption excluding combustible renewables and waste (CRW). Since CRW is defined to contain 97% commercial and noncommercial biomass, the percentage share of biomass for each country listed here is calculated as the difference between the percentage of total consumption and the percentage of CRW.

TABLE VI

Total Energy Consumption, Total Biomass Energy Consumption, and Biomass Energy Consumption by Biomass Resource in EJ/Year for United States and Top 10 Energy-Consuming Countries

Country	Total	Total biomass	Renewable MSW	Industrial wastes	Primary biomass solids	Biogas	Liquid biomass
United States	99.85	3.373	0.308	0.166	2.616	0.143	0.140
China*	47.25	9.244	0	0	9.191	0.054	0
Russia*	26.18	0.326	0	0.111	0.216	0	0
Japan	22.78	0.242	0.044	0	0.198	0	0
India*	20.84	8.596	0	0	8.596	0	0
Germany	14.74	0.366	0.076	0.045	0.213	0.024	0.007
France	11.16	0.496	0.079	0	0.399	0.008	0.011
Canada	10.9	0.487	0	0	0.487	0	0
United Kingdom	10.10	0.093	0.012	0.002	0.036	0.035	0
South Korea	8.405	0.092	0.065	0.015	0.007	0.002	0
Brazil*	7.80	1.862	0	0	1.547	0	0.322

Note. The energy consumption data for each country listed here were adapted from the International Energy Agency (2002). The data presented in Mtoe were converted to EJ/year using a multiplier of 0.043412. The data for those countries marked with an asterisk are for 1999; the remaining data are for 2000. Data reported as "0" by the IEA are shown in the table (see text). The sum of the energy consumption figures for the biomass resources may not correspond to total biomass energy consumption because of rounding and other factors (see text). The nomenclature used here is IEA's, as follows: biomass consists of solid biomass and animal products, gas/liquids from biomass, industrial waste, and municipal waste any plant matter that is used directly as fuel or converted into fuel, such as charcoal, or to electricity or heat. Renewable MSW consists of the renewable portion of municipal solid waste, including hospital waste, that is directly converted to heat or power. Industrial waste consists of solid and liquid products such as tires, that are not reported in the category of solid biomass and animal products. Primary biomass solids consists of any plant matter used directly as fuel or converted into other forms before combustion, such as feedstock for charcoal production. This latter category includes wood, vegetal waste including wood wastes and crops used for energy production. Biogas consists of product fuel gas from the anaerobic digestion of biomass and soild wastes—including landfill gas, sewage gas, and gas from animal wastes—that is combusted to produce heat or power. Liquid biomass includes products such as ethanol.

sight. Irreversible shortages of these fossil fuels are expected to occur before the middle of the 21st century because their proved reserves have been projected to be insufficient to meet demands at that time. Supply disruptions are expected to start first with natural gas. This is illustrated by using a reserves availability model to plot global proved natural gas reserves and five times the proved reserves versus year as shown in Fig. 3. Presuming this model provides results that are more valid over the long term than reserves-to-consumption ratios, the trend in the curves indicates that shortages of natural gas would be expected to occur in the early years of the 21st century and then begin to cause

TABLE VII

Total Electricity Generation, Total Biomass-Based Electricity Generation, and Biomass-Based Electricity Generation by Biomass Resource in TWh/Year for United States and Top 10 Energy-Consuming Countries

Country	Total	Total biomass	Renewable MSW	Industrial wastes	Primary biomass solids	Biogas
United States	4003.5	68.805	15.653	6.552	41.616	4.984
China*	1239.3	1.963	0	0	1.963	0
Russia*	845.3	2.075	0	2.045	0.030	0
Japan	1081.9	16.518	5.209	0	11.309	0
India*	527.3	0	0	0	0	0
Germany	567.1	10.121	3.688	3.946	0.804	1.683
France	535.8	3.290	1.995	0	0.949	0.346
Canada	605.1	7.379	0	0	7.379	0
United Kingdom	372.2	4.360	0.695	0	0.700	2.556
South Korea	292.4	0.396	0.361	0	0	0
Brazil*	332.3	8.519	0	0	8.519	0

Note: The electricity generation data for each country listed here were adapted from the International Energy Agency (2002). The data for those countries marked with an asterisk are for 1999; the remaining data are for 2000. Data reported as "0" by the IEA are shown in the table (see text). The data compiled by the IEA is defined as "electricity output." The sum of the electricity generation figures for the biomass resources may not correspond to total biomass electricity generation because of rounding and other factors (see text). The nomenclature used here is IEA's, as follows: biomass consists of solid biomass and animal products, gas/liquids from biomass, industrial waste and municipal waste, and any plant matter that is used directly as fuel or converted into fuel, such as charcoal, or to electricity or heat. Renewable MSW consists of the renewable portion of municipal solid waste, including hospital waste, that is directly converted to heat or power. Industrial waste consists of solid and liquid products, such as tires, that are not reported in the category of solid biomass and animal products. Primary biomass solids consists of any plant matter used directly as fuel or converted into other forms before combustion, such as feedstock for charcoal production. This latter category includes wood, vegetal waste including wood wastes and crops used for energy production. Biogas consists of product fuel gas from the anaerobic digestion of biomass and soild wastes—including landfill gas, sewage gas, and gas from animal wastes—that is combusted to produce heat or power. Liquid biomass includes products such as ethanol.

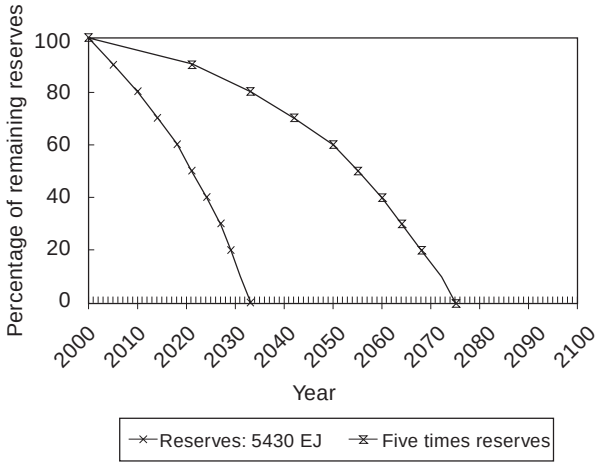


FIGURE 3 Global natural gas reserves remaining at annual growth rate in consumption of 3.2%. From Klass (2003). *Energy Pol.* **31**, 353.

serious supply problems in the next 20 to 30 years. Large-scale price increases for fossil fuels are probable because of what has been called the first derivative of the law of supply and demand, the law of energy availability and cost. This eventuality coupled with the adverse impacts of fossil fuel consumption on the environment are expected to be the driving forces that stimulate the transformation of virgin and waste biomass and other renewable energy resources into major resources for the production of energy, fuels, and commodity chemicals.

1.2 Biomass Composition and Energy Content

Typical organic components in representative, mature biomass species are shown in Table VIII, along with the corresponding ash contents. With few exceptions, the order of abundance of the major organic components in whole-plant samples of terrestrial biomass is celluloses, hemicelluloses, lignins, and proteins. Aquatic biomass does not appear to follow this trend. The cellulosic components are often much lower in concentration than the hemicelluloses as illustrated by the data for water hyacinth *(Eichhornia crassipes)*. Other carbohydrates

TABLE VIII

Organic Components and Ash in Representative Biomass

Biomass type	Marine	Fresh water	Herbaceous	Woody	Woody	Woody	Waste
Name	Giant brown kelp	Water hyacinth	Bermuda grass	Poplar	Sycamore	Pine	RDF
Component (dry wt %)							
Celluloses	4.8	16.2	31.7	41.3	44.7	40.4	65.6
Hemicelluloses		55.5	40.2	32.9	29.4	24.9	11.2
Lignins		6.1	4.1	25.6	25.5	34.5	3.1
Mannitol	18.7						
Algin	14.2						
Laminarin	0.7						
Fucoidin	0.2						
Crude protein	15.9	12.3	12.3	2.1	1.7	0.7	3.5
Ash	45.8	22.4	5.0	1.0	0.8	0.5	16.7
Total	100.3	112.5	93.3	102.9	102.1	101.0	100.1

Note. All analyses were performed by the Institute of Gas Technology (Gas Technology Institute). The crude protein content is estimated by multiplying the nitrogen value by 6.25. RDF is refuse-derived fuel (i.e., the combustible fraction of municipal solid waste).

and derivatives are dominant in marine species such as giant brown kelp *(Macrocystis pyrifera)* to almost complete exclusion of the celluloses. The hemicelluloses and lignins have not been found in *M. pyrifera*.

Alpha-cellulose, or cellulose as it is more generally known, is the chief structural element and major constituent of many biomass species. In trees, it is generally about 40 to 50% of the dry weight. As a general rule, the major organic components in woody biomass on a moisture and ash-free basis in weight percent are about 50 cellulosics, 25 hemicellulosics, and 25 lignins. The lipid and protein fractions of plant biomass are normally much less on a percentage basis than the carbohydrate components. The lipids are usually present at the lowest concentration, while the protein fraction is somewhat higher, but still lower than the carbohydrate fraction. Crude protein values can be approximated by multiplying the organic nitrogen analyses by 6.25. The sulfur contents of virgin and waste biomass range from very low to about 1 weight percent for primary biosolids. The sulfur content of most woody species of biomass is nil.

The chemical energy content or heating value is of course an important parameter when considering energy and fuel applications for different biomass species and types. The solid biomass formed on photosynthesis generally has a higher heating value on a dry basis in the range of 15.6 to 20.0 MJ/kg (6,700 to 8,600 Btu/lb), depending on the species. Typical carbon contents and higher heating values of the most common classes of biomass components are shown on a dry basis in Table IX. The higher the carbon content, the greater the energy value. It is apparent that the lower the degree of oxygenation, the more hydrocarbon-like and the higher the heating value. When the heating values of most waste and virgin biomass samples are converted to energy content per mass unit of carbon, they usually

TABLE IX

Typical Carbon Content and Heating Value of Selected Biomass Components

Component	Carbon (wt %)	Higher heating value (MJ/kg)
Monosaccharides	40	15.6
Disaccharides	42	16.7
Polysaccharides	44	17.5
Crude proteins	53	24.0
Lignins	63	25.1
Lipids	76–77	39.8
Terpenes	88	45.2
Crude carbohydrates	41–44	16.7–17.7
Crude fibers	47–50	18.8–19.8
Crude triglycerides	74–78	36.5–40.0

Note. Adapted from Klass, D. L. (1994). "Kirk-Othmer Encyclopedia of Chemical Technology," 4th ed., vol. 12, pp. 16–110. John Wiley & Sons. New York. Carbon contents and higher heating values are approximate values for dry mixtures; crude fibers contain 15 to 30% lignins.

fall within a narrow range. The energy value of a sample can be estimated from the carbon and moisture analyses without actual measurement of the heating values in a calorimeter. Manipulation of the data leads to a simple equation for calculating the higher heating value of biomass samples and also of coal and peat samples. One equation that has been found to be reasonably accurate is

$$\text{Higher heating value in MJ/dry kg} = 0.4571\ (\%\text{C on dry basis}) - 2.70. \quad (3)$$

2. BIOMASS CONVERSION TECHNOLOGIES

2.1 Processes

The technologies include a large variety of thermal and thermochemical processes for converting biomass by combustion, gasification, and liquefaction, and the microbial conversion of biomass to obtain gaseous and liquid fuels by fermentative methods. Examples of the former are wood-fueled power plants in which wood and wood wastes are combusted for the production of steam, which is passed through a steam turbine to generate electricity; the gasification of rice hulls by partial oxidation to yield a low-energy-value fuel gas, which drives a gas turbine to generate electric power, and finely divided silica coproduct for sale; the rapid pyrolysis or thermal decomposition of wood and wood wastes to yield liquid fuel oils and chemicals; and the hydrofining of tall oils from wood pulping, vegetable oils, and waste cooking fats to obtain high-cetane diesel fuels and diesel fuel additives. Examples of microbial conversion are the anaerobic digestion of biosolids to yield a relatively high-methane-content biogas of medium energy value and the alcoholic fermentation of corn to obtain fuel ethanol for use as an oxygenate and an octane-enhancing additive in motor gasolines.

Another route to liquid fuels and products is to grow certain species of biomass that serve the dual role of a carbon-fixing apparatus and a natural producer of high-energy products such as triglycerides or hydrocarbons. Examples are soybean, from which triglyceride oil coproducts are extracted and converted to biodiesel fuels, which are the transesterified methyl or ethyl esters of the fatty acid moieties of the triglycerides having cetane numbers of about 50, or the triglycerides are directly converted to high-cetane value paraffinic hydrocarbon diesel fuels having cetane numbers of about 80 to 95 by catalytic hydrogenation; the tapping of certain species of tropical trees to obtain liquid hydrocarbons suitable for use as diesel fuel without having to harvest the tree; and the extraction of terpene hydrocarbons from coniferous trees for conversion to chemicals. A multitude of processes thus exists that can be employed to obtain energy, fuels, and chemicals from biomass. Many of the processes are suitable for either direct conversion of biomass or conversion of intermediates. The processes are sufficiently variable so that liquid and gaseous fuels can be produced that are identical to those obtained from fossil feedstocks, or are not identical but are suitable as fossil fuel substitutes. It is important to emphasize that virtually all of the fuels and commodity chemicals manufactured from fossil fuels can be manufactured from biomass feedstocks. Indeed, several of the processes used in a petroleum refinery for the manufacture of refined products and petrochemicals can be utilized in a biorefinery with biomass feedstocks. Note also that selected biomass feedstocks are utilized for conversion to many specialty chemicals, pharmaceuticals, natural polymers, and other higher value products.

2.2 Integrated Biomass Production-Conversion Systems

The energy potential of waste biomass, although of significant importance for combined waste disposal, energy-recovery applications, is relatively small compared to the role that virgin biomass has as an energy resource. The key to the large-scale production of energy, fuels, and commodity chemicals from biomass is to grow suitable virgin biomass species in an integrated biomass-production conversion system (IBPCS) at costs that enable the overall system to be operated at a profit. Multiple feedstocks, including combined biomass–fossil feedstocks and waste biomass, may be employed. Feedstock supply, or supplies in the case of a system that converts two or more feedstocks, is coordinated with the availability factor (operating time) of the conversion plants. Since growing seasons vary with geographic location and biomass species, provision is made for feedstock storage to maintain sufficient supplies to sustain plant operating schedules.

The proper design of an IBPCS requires the coordination of numerous operations such as biomass planting, growth management, harvesting, storage, retrieval, transport to conversion plants, drying, conversion to products, emissions control,

product separation, recycling, wastewater and waste solids treatment and disposal, maintenance, and transmission or transport of salable products to market. The design details of the IBPCS depend on the feedstocks involved and the type, size, number, and location of biomass growth and processing areas needed. It is evident that a multitude or parameters are involved. In the idealized case, the conversion plants are located in or near the biomass growth areas to minimize the cost of transporting biomass to the plants, all the nonfuel effluents of which are recycled to the growth areas (Fig. 4). If this kind of plantation can be implemented in the field, it would be equivalent to an isolated system with inputs of solar radiation, air, CO_2, and minimal water; the outputs consist of the product slate. The nutrients are kept within the ideal system so that addition of external fertilizers and chemicals is not necessary. Also, the environmental controls and waste disposal problems are minimized.

It is important to understand the general characteristics of IBPCSs and what is required to sustain their operation. Consider an IBPCS that produces salable energy products at a rate of 10,000 boe/day from virgin biomass. This is a small output relative to most petroleum refineries, but it is not small for an IBPCS. Assume that the plant operates at an availability of 330 day/year at an overall thermal efficiency of converting feedstock to salable energy products of 60%, a reasonable value for established thermochemical conversion technologies. Equivalent biomass feedstock of average energy content of 18.60 GJ/dry tonne would have to be provided at the plant gate to sustain conversion operations at a rate of 5291 dry tonne/day, or a total of 1,746,000 dry tonne/year. This amount of feedstock, at an average biomass yield of 25 dry tonne/ha-year, requires a biomass growth area of 69,840 ha (270 square miles), or a square area 26.4 km (16.4 miles) on each edge. For purposes of estimation, assume the product is methanol and that no coproducts are formed. The total annual methanol production is then approximately 1.237 billion liters/year (327 million gallons/year). Fifty-four IBPCSs of this size are required to yield 1.0 quad of salable methanol energy per year, and the total growth area required is 3,771,400 ha (14,561 square miles), or a square area 194.2 km (120.7 miles) on each edge. Again exclusive of infrastructure and assuming the conversion facilities are all centrally located, the growth area is circumscribed by a radial distance of 101.4 km (68.1 miles) from the plants.

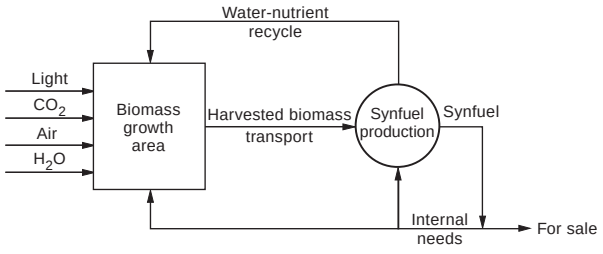


FIGURE 4 Idealized biomass growth and manufacturing system. From Klass (1998).

This simplistic analysis shows that the growth areas required to supply quad blocks of energy and fuels would be very large when compared with conventional agricultural practice, but that 10,000-boe/day systems are not quite so large when compared with traditional, sustainable wood harvesting operations in the forest products industry. The analysis suggests that smaller, localized IBPCSs in or near market areas will be preferred because of logistics and product freight costs, and multiple feedstocks and products will have advantages for certain multiproduct slates. For example, commercial methanol synthesis is performed mainly with natural gas feedstocks via synthesis gas. Synthesis gas from biomass gasification used as cofeedstock in an existing natural gas-to-methanol plant can utilize the excess hydrogen produced on steam reforming natural gas. Examination of hypothetical hybrid methanol plants shows that they have significant benefits such as higher methanol yields and reduced natural gas consumption for the same production capacity.

Sustainable virgin biomass production at optimum economic yields is a primary factor in the successful operation of IBPCSs. Methodologies such as no-till agriculture and short-rotation woody crop (SWRC) growth have been evaluated and are being developed for biomass energy and combined biomass energy-coproduct applications. Most of the IBPCSs that have been proposed are site specific—that is, they are designed for one or more biomass species, in the case of a multicropping system, for specific regions. Field trials of small IBPCSs or modules of IBPCSs are in progress in the United States, but no full-scale systems have yet been built. Some of the large, commercial forestry operations for tree growth, harvesting, and transport to the mills can be considered to be analogous in many respects to the biomass production phase of managed IBPCSs. The growth, harvesting, and transport of corn to fermentation plants for fuel ethanol manufacture in the U.S. Corn Belt is perhaps the

closest commercial analog to an IBPCS in the United States. Most of the other large IBPCSs that have been announced for operation outside the United States are either conceptual in nature or have not been fully implemented.

The historical development of IBPCSs shows that large-scale biomass energy plantations must be planned extremely carefully and installed in a logical scale-up sequence. Otherwise, design errors and operating problems can result in immense losses and can be difficult and costly to correct after construction of the system is completed and operations have begun. It is also evident that even if the system is properly designed, its integrated operation can have a relatively long lag phase, particularly for tree plantations, before returns on investment are realized. The financial arrangements are obviously critical and must take these factors into consideration.

3. COMMERCIAL BIOMASS ENERGY MARKETS AND ECONOMICS

3.1 Some Examples

The United States only has about 5% of the world's population, but is responsible for about one-quarter of total global primary energy demand. The markets for biomass energy in the United States are therefore already established. They are large, widespread, and readily available as long as the end-use economics are competitive.

To cite one example, petroleum crude oils have been the single largest source of transportation fuels since the early 1900s in the United States. Because of undesirable emissions from conventional hydrocarbon fuels, the U.S. Clean Air Act Amendments of 1990 included a requirement for dissolved oxygen levels in unleaded gasoline of at least 2.7 weight percent during the four winter months for 39 so-called nonattainment areas. The Act also required a minimum of 2.0 weight percent dissolved oxygen in reformulated gasoline in the nine worst ozone nonattainment areas year-round. The largest commercial oxygenates for gasolines in the United States have been fermentation ethanol, mainly from corn, and MTBE, which is manufactured from petroleum and natural gas feedstocks. Oxygenated gasolines are cleaner burning than nonoxygenated gasolines, and the oxygenates also serve as a replacement for lead additives by enhancing octane value in gasoline blends.

U.S. motor gasoline production was about 473 billion liters (125 billion gallons) in 2000, during which total U.S. gasohol production was about 61.7 billion liters (16.3 billion gallons) if it is assumed that all domestically produced fuel ethanol from biomass feedstocks in 2000, 6.17 billion liters (1.63 billion gallons), was blended with motor gasolines as gasohol. MTBE has been a major competitor of fuel ethanol as an oxygenate and octane-improving additive for unleaded gasolines since the phase-out of leaded gasolines began in the 1970s and 1980s. Without reviewing the detailed reasons for it other than to state that leakage from underground storage tanks containing MTBE-gasoline blends has polluted underground potable water supplies, federal legislation is pending that would eliminate all MTBE usage in gasolines and establish a renewables energy mandate in place of an oxygenate mandate. The provisions of this mandate are expected to include the tripling of fuel ethanol production from biomass by 2012. MTBE would be prohibited from use in motor gasoline blends in all states within a few years after enactment of the legislation. If the mandate does not become federal law, the replacement of MTBE by fuel ethanol is still expected to occur because many states have already announced plans to prohibit MTBE usage. Other states are exploring the benefits and logistics of removing it from commercial U.S. gasoline markets.

The cost of fermentation ethanol as a gasoline additive has been reduced at the pump by several federal and state tax incentives. The largest is the partial exemption of $0.053/gallon for gasohol-type blends ($0.53/gallon of fuel ethanol), out of a federal excise gasoline tax of $0.184/gallon, and a small ethanol producers tax credit of $0.10/gallon. The purpose of these incentives is to make fermentation ethanol-gasoline blends more cost competitive. Without them, it is probable that the market for fuel ethanol would not have grown as it has since it was re-introduced in 1979 in the United States as a motor fuel component. With corn at market prices of $2.00 to $2.70/bushel, its approximate price range from 1999 to 2002, the feedstock alone contributed 20.3 to 27.4 cents/liter ($0.769 to $1.038/gallon) to the cost of ethanol without coproduct credits. In 1995–1996, when the market price of corn was as high as $5.00/bushel, many small ethanol producers had to close their plants because they could not operate at a profit.

An intensive research effort has been in progress in the United States since the early 1970s to improve the economics of manufacturing fermentation ethanol using low-grade, and sometimes negative-cost feedstocks such as wood wastes and RDF, instead of corn

and other higher value biomass feedstocks. Significant process improvements, such as large reductions in process energy consumption, have been made during this research, but the target production cost of $0.16/liter ($0.60/gallon) has not yet been attained. It is believed by some that fuel ethanol production will exhibit even larger increases than those mandated by legislation, possibly without the necessity for tax incentives, when this target is attained.

Some cost estimates indicate that fuel ethanol as well as the lower molecular weight C_3 to C_6 alcohols and mixtures can be manufactured by thermochemical non-fermentative processing of a wide range of waste biomass feedstocks at production costs as low as 25 to 50% the cost of fermentation ethanol from corn. The C_3+ alcohols and mixtures with ethanol also have other advantages compared to ethanol in gasoline blends. Their energy contents are closer to those of gasoline; their octane blending values are higher; the compatibility and miscibility problems with gasolines are small to nil; excessive Reid vapor pressures and volatility problems are less or do not occur; and they have higher water tolerance in gasoline blends, which facilitates their transport in petroleum pipelines without splash blending. Splash blending near the point of distribution is necessary for fuel ethanol-gasoline blends.

Another example of a commercial biomass-based motor fuel is biodiesel fuel. It is utilized both as a cetane-improving additive and a fuel component in diesel fuel blends, and as a diesel fuel alone. Biodiesel is manufactured from the triglycerides obtained from oil seeds and vegetable oils by transesterifying them with methanol or ethanol. Each ester is expected to qualify for the same renewable excise tax exemption incentive on an equal volume basis as fuel ethanol from biomass in the United States. Unfortunately, the availability of biodiesel is limited. The main reason for the slow commercial development of biodiesel is the high production cost of $75 to $150/barrel caused mainly by the relatively low triglyceride yields per unit growth area, compared to the cost of conventional diesel fuel from petroleum crude oils.

In Europe, where the costs of motor fuels including diesel fuel are still significantly higher than those in the United States, the commercial scale-up of biodiesel, primarily the esters from the transesterification of rape seed triglycerides, has fared much better. Production is targeted at 2.3 million tonnes (5.23 billion liters, 1.38 billion gallons) in 2003, and 8.3 million tonnes (18.87 billion liters, 4.98 billion gallons) in 2010. Several European countries have established zero duty rates on biodiesel to increase usage and reduce greenhouse gas emissions from diesel-fueled vehicles.

Still another example of the commercial application of biomass energy is its use for the generation of electricity. U.S. tax incentives have been provided to stimulate and encourage the construction and operation of biomass-fueled power generation systems. Most of them are operated by independent power producers or industrial facilities, not utilities. The installed, nonutility electric generation capacity fueled with renewables in the United States and the utility purchases of electricity from nonutilities generated from renewable resources including biomass in 1995 by source, capacity, and purchases are shown in Table X. Note that the sums of the biomass-based capacities—wood and wood wastes, MSW and landfills, and other biomass—and purchases are about 43% and 57% of the totals from all renewable energy resources.

Unfortunately, several of the federal tax incentives enacted into law to stimulate commercial power generation from biomass energy have expired or the qualifying conditions are difficult to satisfy. In 2002, there were no virgin biomass species that were routinely grown as dedicated energy crops in the United States for generating electricity. There are many small to moderate size power plants, however, that are fueled with waste biomass or waste biomass-fossil fuel blends throughout the United States. These plants are often able to take credits such as the tipping fees for accepting MSW and RDF for disposal via power plants that use combustion or gasification as the means of energy recovery and disposal of these wastes, the federal Section 29 tax credit for the

TABLE X

Installed U.S. Nonutility Electric Generation Capacity from Renewable Energy Resources and Purchases of Electricity by Utilities from Nonutilities by Resource in 1995

Renewable resource	Capacity (GW)	Purchases (TWh)
Wood and wood wastes	7.053	9.6
Conventional hydro	3.419	7.5
MSW and landfills	3.063	15.3
Wind	1.670	2.9
Geothermal	1.346	8.4
Solar	0.354	0.8
Other biomass	0.267	1.5
Total	17.172	46.0

Note. Adapted from Energy Information Administration (1999). Renewable Energy 1998: Issues and Trends, DOE/EIA-0628(98), March, Washington, DC.

conversion to electricity of the LFG collected from landfills, the equivalent cost of purchased power generated from biogas in a wastewater treatment plant for on-site use, or the sale of surplus power to utilities by an IPP at the so-called avoided cost, which is the cost the utility would incur by generating the power itself.

3.2 Advanced Technologies

The research programs funded by the public and private sectors in the United States to develop renewable energy technologies since the First Oil Shock have led to numerous scientific and engineering advances for basically all renewable energy resources. Some of the advanced biomass-related technologies are listed here. Many of them have already been or will be commercialized.

- The development of hybrid trees and special herbaceous biomass species suitable for use as dedicated energy crops in different climates.
- Advanced plantation designs for the managed multicropping of virgin biomass species in integrated biomass production-conversion systems.
- Advanced biorefinery system designs for the sustained production of multiple product slates.
- Practical hardware and lower cost installation methods for recovering LFG from sanitary landfills for power generation and mitigation of methane emissions.
- Safety-engineered, unmanned LFG-to-electricity systems that operate continuously.
- High rate anaerobic treatment processes for greater destruction of pathogens and biosolids in wastewaters at higher biogas yields and production rates.
- Zero-emissions waste biomass combustion systems for combined disposal-energy recovery and recycling.
- Genetically engineered microorganisms capable of simultaneously converting all pentose and hexose sugars from cellulosic biomass to fermentation ethanol.
- Catalysts for thermochemical gasification of biomass feedstocks to product gases for conversion to preselected chemicals in high yields.
- Processes for the thermochemical conversion of waste and virgin biomass feedstocks to ethanol and lower molecular weight alcohols and ethers.
- Close-coupled biomass gasification-combustion systems for the production of hot water and steam for commercial buildings and schools.
- Advanced biomass gasification processes for the high-efficiency production of medium-energy-content fuel gas and power.
- Short-residence-time pyrolysis processes for the production of chemicals and liquid fuels from biomass.
- Catalytic processes for the direct conversion of triglycerides and tall oils to “super cetane” diesel fuels and diesel fuel additives having cetane numbers near 100.

3.3 Economic Impacts and Barriers

When full-scale, well-designed IBPCSs are in place in industrialized countries and are supplying energy, organic fuels, and commodity chemicals to consumers, conventional fossil fuel production, refining, and marketing will have undergone major changes. Numerous economic impacts are expected to occur. Biomass energy production and distribution will be a growth industry, while the petroleum and gas industries will be in decline. Because of the nature of IBPCSs, employment in agriculture and forestry and the supporting industries will exhibit significant increases over many different areas of the country. Unlike petroleum refineries, which are geographically concentrated in relatively few areas, and are therefore dependent on various long-distance modes of transporting refined products to market, the biomass energy industry will be widely dispersed in rural areas. Most IBPCSs will incorporate their own biorefineries. The transport distances of refined products to market will be relatively short, and the logistics of supplying energy demands will change. It is apparent that there will be many national and international impacts of the Renewable Energy Era.

The regional economic impact of biomass energy alone is illustrated by an assessment for the U.S. Southeast from which it was concluded that industrial wood energy generated 71,000 jobs and 1 billion dollars of income annually. It was estimated in another study that 80 cents of every dollar spent on biomass energy in a given region stays in the region, while almost all expenditures on petroleum products leave the region. Still another assessment conducted for the state of Wisconsin in the Midwest Corn Belt indicates the economic impacts of shifting a portion of Wisconsin’s future energy investment from fossil fuels to biomass energy. This study assumed a 75% increase in the state’s renewable energy use by 2010: 775 MW of new electric generating capacity to supply electricity to 500,000 Wisconsin homes and 379 million liters

(100 million gallons) per year of new ethanol production to supply gasohol to 45% of Wisconsin's automobiles. This scenario generated about three times more jobs, earnings, and sales in Wisconsin than the same level of imported fossil fuel usage and investment and was equivalent to 62,234 more job-years of net employment, $1.2 billion in higher wages, and $4.6 billion in additional output. Over the operating life of the technologies analyzed, about $2 billion in avoided payments for imported fuels would remain in Wisconsin to pay for the state-supplied renewable resources, labor, and technologies. Wood, corn, and waste biomass contributed 47% of the increase in net employment.

Nationwide, the projected economic impacts of biomass energy development are substantial. In 2001, with petroleum crude oil imports at about 9.33 million barrels/day, consumption of biomass energy and fuels corresponds to the displacement of 1.64 Mboe/day, or 17.6% of the total daily imports. This effectively reduces expenditures for imported oil, and beneficially impacts the U.S. trade deficit. Since agricultural crops and woody biomass as well as industrial and municipal wastes are continuously available throughout the United States, biomass also provides a strategic and distributed network of renewable energy supplies throughout the country that improve national energy security.

Conservatively, the energy and fuels available in the United States for commercial markets on a sustainable basis from virgin and waste biomass has been variously estimated to range up to about 15 quad per year, while the energy potentially available each year has been estimated to be as high as 40 quad. This is made up of 25 quad from wood and wood wastes and 15 quad from herbaceous biomass and agricultural residues. Utilization of excess capacity croplands of up to 64.8 million hectares (160 million acres) estimated to be available now for the growth of agricultural energy crops could open the way to new food, feed, and fuel flexibility by providing more stability to market prices, by creating new markets for the agricultural sector, and by reducing federal farm subsidy payments. Based on the parameters previously described for one-quad IBPCSs, this acreage is capable of producing about 17 quad of salable energy products from herbaceous feedstocks each year. Other opportunities to develop large IBPCSs exist in the United States using federally owned forest lands. Such IBPCSs would be designed for sustainable operations with feedstocks of both virgin and waste wood resources such as thinnings, the removal of which would also reduce large-scale forest fires that have become commonplace in the dryer climates, particularly in the western states.

Because of the multitude of organic residues and biomass species available, and the many different processing combinations that yield solid, liquid, and gaseous fuels, and heat, steam, and electric power, the selection of the best feedstocks and conversion technologies for specific applications is extremely important. Many factors must be examined in depth to choose and develop systems that are technically feasible, economically and energetically practical, and environmentally superior. These factors are especially significant for large-scale biomass energy plantations where continuity of operation and energy and fuel production are paramount. But major barriers must be overcome to permit biomass energy to have a large role in displacing fossil fuels.

Among these barriers are the development of large-scale energy plantations that can supply sustainable amounts of low-cost feedstocks; the risks involved in designing, building, and operating large IBPCSs capable of producing quad blocks of energy and fuels at competitive prices; unacceptable returns on investment and the difficulties encountered in obtaining financing for first-of-a-kind IBPCSs; and the development of nationwide biomass energy distribution systems that simplify consumer access and ease of use. These and other barriers must ultimately be addressed if any government decides to institute policies to establish large-scale biomass energy markets.

Without IBPCSs, biomass energy will be limited to niche markets for many years until oil or natural gas depletion starts to occur. The initiation of depletion of these nonrenewable resources may in fact turn out to cause the Third Oil Shock in the 21st century.

4. ENVIRONMENTAL IMPACTS

Several environmental impacts are directly related to biomass energy production and consumption. The first is obviously the environmental benefit of displacing fossil fuel usage and a reduction in any adverse environmental impacts that are caused by fossil fuel consumption. In addition, the use of a fossil fuel and biomass together in certain applications, such as electric power generation with coal and wood or coal and RDF in dual-fuel combustion or cocombustion plants, can result in reduction of undesirable emissions. The substitution of fossil fuels and their derivatives by biomass and biofuels also helps to conserve depletable fossil fuels.

Another beneficial environmental impact results from the combined application of waste biomass disposal and energy recovery technologies. Examples are biogas recovery from the treatment of biosolids in municipal wastewater treatment plants by anaerobic digestion, LFG recovery from MSW landfills, which is equivalent to combining anaerobic digestion of waste biomass and LFG "mining," and the conversion of MSW, refuse-derived fuel (RDF), and farm, forestry, and certain industrial wastes, such as black liquor generated by the paper industry, to produce heat, steam, or electric power. Resource conservation and environmental benefits certainly accrue from such applications.

Another environmental impact is more complex. It concerns the growth and harvesting of virgin biomass for use as dedicated energy crops. By definition, sustainable, biomass energy plantations are designed so that the biomass harvested for conversion to energy or fuels is replaced by new biomass growth. If more biomass is harvested than is grown, the system is obviously not capable of continued operation as an energy plantation. Furthermore, the environmental impact of such systems can be negative because the amount of CO_2 removed from the atmosphere by photosynthesis of biomass is then less that that needed to balance the amount of biomass carbon removed from the plantation. In this case, virgin biomass is not renewable; its use as a fuel results in a net gain in atmospheric CO_2. Energy plantations must be designed and operated to avoid net CO_2 emissions

TABLE XI

Estimated Annual Global Carbon Dioxide and Carbon Exchanges with the Atmosphere

	Carbon dioxide		Carbon equivalent	
Source or sink	To atmosphere (Gt/year)	From atmosphere (Gt/year)	To atmosphere (Gt/year)	From atmosphere (Gt/year)
Terrestrial				
Cement production	0.51		0.14	
Other industrial processes	0.47		0.13	
Human respiration	1.67		0.46	
Animal respiration	3.34		0.91	
Methane emissions equivalents	1.69		0.46	
Natural gas consumption	3.98		1.09	
Oil consumption	10.21		2.79	
Coal consumption	8.15		2.22	
Biomass burning	14.3		3.90	
Gross biomass photosynthesis		388		106
Biomass respiration	194		53	
Soil respiration and decay	194		53	
Total terrestrial	432	388	118	106
Oceans				
Gross biomass photosynthesis		180		49
Biomass respiration	90		25	
Physical exchange	275	202	75	55
Total oceans	365	382	100	104
Total terrestrial and oceans	797	770	218	210

Note. The fossil fuel, human, and animal emissions were estimated by Klass (1998). Most of the other exchanges are derived from exchanges in the literature or they are based on assumptions that have generally been used by climatologists. It was assumed that 50% of the terrestrial biomass carbon fixed by photosynthesis is respired and that an equal amount is emitted by the soil. The total uptake and emission of carbon dioxide by the oceans were assumed to be 104 and 100 Gt C/year (Houghton, R. A., and Woodwell, G. M. (1989). *Sci. Am.* **260**(4), 36) and biomass respiration was assumed to emit 50% of the carbon fixed by photosynthesis. The carbon dioxide emissions from cement production and other industrial processes are process emissions that exclude energy-related emissions; they are included in the fossil fuel consumption figures.

to the atmosphere. A few biomass plantations are now operated strictly to offset the CO_2 emissions from fossil-fired power plants, particularly those operated on coal. Sometimes, the fossil-fired power plant and the biomass plantation are geographically far apart. It is important to emphasize that established IBPCSs that utilize dedicated energy crops will normally involve the harvesting of incrementally new virgin biomass production.

Finally, there is the related issue of the causes of increasing concentrations of atmospheric CO_2, which is believed to be the greenhouse gas responsible for much of the climatic changes and temperature increases that have been observed. Most climatologists who have studied the problem portray atmospheric CO_2 buildup to be caused largely by excessive fossil fuel usage. Some assessments indicate that biomass contributes much more to the phenomenon than formerly believed, possibly even more than fossil fuel consumption. Because terrestrial biomass is the largest sink known for the removal of atmospheric CO_2 via photosynthesis, *the accumulated loss in global biomass growth areas with time and the annual reduction in global CO_2 fixation capacity* are believed by some to have had a profound adverse impact on atmospheric CO_2 buildup. The population increase and land use changes due to urbanization, the conversion of forest to agricultural and pasture lands, the construction of roads and highways, the destruction of areas of the rainforests, large-scale biomass burning, and other anthropological activities appear to contribute to atmospheric CO_2 buildup at a rate that is much larger than fossil fuel consumption. This is illustrated by the estimated annual global CO_2 exchanges with the atmosphere shown in Table XI. Despite the possibilities for errors in this tabulation, especially regarding absolute values, several important trends and observations are apparent and should be valid for many years. The first observation is that fossil fuel combustion and industrial operations such as cement manufacture emit much smaller amounts of CO_2 to the atmosphere than biomass respiration and decay and the physical exchanges between the oceans and the atmosphere. The total amount of CO_2 emissions from coal, oil, and natural gas combustion is also less than 3% of that emitted by all sources. Note that human and animal respiration are projected to emit more than five times the CO_2 emissions of all industry exclusive of energy-related emissions. Note also that biomass burning appears to emit almost as much CO_2 as oil and natural gas consumption together. Overall, the importance of the two primary sinks for atmospheric CO_2—terrestrial biota and the oceans—is obvious. No other large sinks have been identified.

Somewhat paradoxically then, it is logical to ask the question: How can large-scale biomass energy usage be considered to be a practical application of virgin biomass? The answer, of course, has already been alluded to. At a minimum, all virgin biomass harvested for energy and fuel applications must be replaced with new growth at a rate that is at least equal to the rate of removal. Even more desirable is the creation of additional new biomass growth areas, most likely forests, because they are the largest, long-lived, global reserve of standing, terrestrial biomass carbon. New biomass growth in fact seems to be one of the more practical routes to remediation of atmospheric CO_2 buildup.

SEE ALSO THE FOLLOWING ARTICLES

Alternative Transportation Fuels: Contemporary Case Studies • Biodiesel Fuels • Biomass, Chemicals from • Biomass Combustion • Biomass Gasification • Biomass: Impact on Carbon Cycle and Greenhouse Gas Emissions • Biomass Resource Assessment • Ethanol Fuel • Forest Products and Energy • Renewable Energy, Taxonomic Overview • Waste-to-Energy Technology

Further Reading

"Bioenergy '96, Proceedings of the Seventh National Bioenergy Conference," Vols. I–II. (1996). The Southeastern Regional Biomass Energy Program (and previous and subsequent biennial proceedings published by the U.S. Regional Biomass Energy Program). Tennessee Valley Authority, Muscle Shoals, AL.

Bisio, A., Boots, S., and Siegel, P. (eds.). (1997). "The Wiley Encyclopedia of Energy and the Environment, Vols. I–II." John Wiley & Sons, New York.

Bridgwater, A. V., and Grassi, G. (eds.). (1991). "Biomass Pyrolysis Liquids Upgrading and Utilization." Elsevier Science, Essex, United Kingdom.

Chartier, P., Beenackers, A. A. C. M., and Grassi, G. (eds.). (1995). "Biomass for Energy, Environment, Agriculture, and Industry, Vols. I-III (and previous and subsequent biennial books)." Elsevier Science, Oxford, United Kingdom.

Cross, B. (ed.). (1995). "The World Directory of Renewable Energy Suppliers and Services 1995." James & James Science, London, United Kingdom.

Cundiff, J. S., *et al.* (eds.) (1996). "Liquid Fuels and Industrial Products from Renewable Resources." American Society of Agricultural Engineers, St. Joseph, MI.

Directory of U.S. "Renewable Energy Technology Vendors: Biomass, Photovoltaics, Solar Thermal, Wind." (1990). Biomass Energy Research Association, Washington, DC.

"First Biomass Conference of the Americas: Energy, Environment, Agriculture, and Industry," Vols. I–III (1993, 1942). NREL/CP-200-5768, DE93010050 (and subsequent biennial books). National Renewable Energy Laboratory, Golden, CO.

Hogan, E., Robert, J., Grassi, G., and Bridgwater, A. V. (eds.). (1992). "Biomass Thermal Processing. Proceedings of the First Canada/European Community Contractors Meeting." CPL Press, Berkshire, United Kingdom.

Klass, D. L. (ed.). (1981). "Biomass as a Nonfossil Fuel Source, American Chemical Society Symposium Series 144." American Chemical Society, Washington, DC.

Klass, D. L. (ed.). (1993). "Energy from Biomass and Wastes XVI (and previous annual books)." Institute of Gas Technology, Chicago, IL.

Klass, D. L. (1998). "Biomass for Renewable Energy, Fuels, and Chemicals." Academic Press, San Diego, CA.

Klass, D. L., and Emert, G. H. (eds.). (1981). "Fuels from Biomass and Wastes. Ann Arbor Science." The Butterworth Group, Ann Arbor, MI.

Myers, R. A. (ed.). (1983). "Handbook of Energy Technology & Economics." John Wiley & Sons, New York.

Nikitin, N. I. (1966, translated). "The Chemistry of Cellulose and Wood." Academy of Sciences of the USSR, translated from Russian by J. Schmorak, Israel Program for Scientific Translations.

Wilbur, L. C. (ed.). (1985). "Handbook of Energy Systems Engineering, Production and Utilization." John Wiley & Sons, New York.

Biomass Gasification

AUSILIO BAUEN
Imperial College London
London, United Kingdom

1. Introduction
2. Fundamentals of Gasification
3. Gasification and Gasifier Types
4. Activities Characteristic of Biomass Gasification Systems for Heat and Electricity Generation
5. Economics of Biomass Gasification for Heat and Electricity Generation
6. Constraints and Issues Affecting the Development of Biomass Gasification Systems

Glossary

biomass Organic matter of vegetable and animal origin.
gasification Thermochemical process that converts a solid or liquid hydrocarbon feedstock of fossil or renewable origin to a gaseous fuel.

Gasification is a thermochemical process that has been exploited for more than a century for converting solid feedstocks to gaseous energy carriers. The first gasifier patent was issued in England at the end of the 18th century and producer gas from coal was mainly used as lighting fuel throughout the 19th century. At the turn of the 20th century, the main use of producer gas, obtained essentially from coal, switched to electricity generation and automotive applications via internal combustion engines. The use of producer gas was gradually supplanted by the use of higher energy density liquid fuels and as a result confined to areas with expensive or unreliable supplies of petroleum fuels. Efforts in gasification technology research have persisted, however, driven mainly by the need for cleaner and more efficient electricity generation technologies based on coal. In the past decade, biomass and municipal solid waste gasification have attracted increasing interest.

1. INTRODUCTION

Biomass gasification allows the conversion of different biomass feedstocks to a more convenient gaseous fuel that can then be used in conventional equipment (e.g., boilers, engines, and turbines) or advanced equipment (e.g., fuel cells) for the generation of heat and electricity. The conversion to a gaseous fuel provides a wider choice of technologies for heat and electricity generation for small- to large-scale applications. Furthermore, electricity generation from gaseous fuels is likely to be more efficient compared to the direct combustion of solid fuels. Efficiency is a particularly important issue for biomass systems because of the possible energy and cost implications of the production and transport of biomass fuels, which are generally characterized by a low energy density. The upgrading of biomass feedstocks to gaseous fuels is also likely to lead to a cleaner conversion. In addition to the production of heat and electricity, the product gas could be used to produce transport fuels, such as synthetic diesel or hydrogen.

The coupling of biomass gasification with gas and steam turbines can provide a modern, efficient, and clean biomass system for the generation of heat and electricity. A number of small- to large-scale biomass gasification systems integrated with power generation equipment are being developed and commercialized. Their market penetration will depend on a number of factors:

- Successful demonstration of the technology
- Economic competitiveness with other energy conversion technologies and fuel cycles
- Environmental performance of the biomass fuel cycle as well as the influence of environmental factors on decision making
- Other socioeconomic factors (e.g., energy security and independence, job creation, and export potential of the technology)

 © 2004 Elsevier Inc. All rights reserved.

This article provides an overview of the state-of-the-art of biomass gasification with a focus on heat and electricity generation. The product gas, however, could also be used for the production of a range of transport fuels.

2. FUNDAMENTALS OF GASIFICATION

Thermochemical processing of biomass yields gaseous, liquid, and solid products and offers a means of producing useful gaseous and/or liquid fuels. Gasification is a total degradation process consisting of a sequence of thermal and thermochemical processes that converts practically all the carbon in the biomass to gaseous form, leaving an inert residue. The gas produced consists of carbon monoxide (CO), hydrogen (H_2), carbon dioxide (CO_2), methane (CH_4), and nitrogen (N_2) (if air is used as the oxidizing agent) and contains impurities, such as small char particles, ash, tars, and oils. The solid residue will consist of ash (composed principally of the oxides of Ca, K, Na, Mg, and Si) and possibly carbon or char. Biomass ash has a melting point of approximately 1000°C; thus, it is important to keep the operating temperature below this temperature to avoid ash sintering and slagging. At temperatures higher than 1300°C, the ash is likely to melt and may be removed as a liquid.

The following sequence is typical of all gasifiers: drying, heating, thermal decomposition (combustion and pyrolysis), and gasification. Table I illustrates the reactions occurring in gasifiers. In directly heated gasifiers, the energy (heat) necessary for the endothermic reactions is provided by combustion and partial combustion reactions within the gasifier. In indirectly heated gasifiers, the heat is generated outside the gasifier and then exchanged with the gasifier. Gasifiers can operate at low (near-atmospheric) or high (several atmospheres) pressure.

The Badouard, steam–carbon, and the methanization reactions are favored with increasing temperature and decreasing pressure; the hydrogasification reaction is favored with decreasing temperature and increasing pressure; the water–gas shift reaction is favored at low temperatures but is independent of pressure. Temperature and pressure operating conditions as well as residence time are key factors in determining the nature of the product gas. Data from existing gasifiers indicate that the heating value of the product gas varies little between pressurized and atmospheric gasification for similar operating temperatures. Gas quality is usually measured in terms of CO and H_2 quantity and ratio.

TABLE I
Reactions Occurring in Gasifiers

Reactions	Enthalpy of reaction (kJ/mol)
Heterogeneous reactions	
Combustion	
$C + \frac{1}{2}O_2 = CO$	−123.1
$C + O_2 = CO_2$	−405.9
Pyrolysis	
$4C_nH_m = mCH_4 + (4n - m)C$	Exothermic
Gasification	
$C + CO_2 = 2CO$ (Badouard)	159.7
$C + H_2O = CO + H_2$ (steam–carbon)	118.7
$C + 2H_2 = CH_4$ (hydrogasification)	−87.4
Homogeneous reactions	
Gas phase reactions	
$CO + H_2O = CO_2 + H_2$ (water–gas shift)	−40.9
$CO + 3H_2 = CH_4 + H_2O$ (methanization)	−206.3

2.1 Effect of Feedstock Properties

Water vapor is an essential component of gasification reactions. However, a high moisture content of the feedstock has an adverse effect on the thermal balance. A low ash content improves the thermal balance, reduces the occlusion and loss of carbon in the residue, and reduces operating problems due to sintering and slagging. Sintering and slagging depend on the gasifier temperature and, in the case of biomass, are likely to be related to the presence of SiO_2, which possesses the lowest melting point among the ash components. The size of the fuel particles will affect the rate of heat and mass transfer in the gasifier. Elements such as sulfur and chlorine lead to the formation of corrosive gas components such as H_2S and HCl. Alkali metals are also a major concern with regard to corrosion, especially when combined with sulfur. Nitrogen in the feedstock leads to the formation of ammonia (NH_3), which can act as a major source of NO_x emissions when combusted in engines or gas turbines.

2.2 Effect of Operating Parameters

The operating temperature will determine the equilibrium composition of the gas. High operating

temperatures increase to different extents the intrinsic rate of all chemical and physical phenomena and result in leaner gases. Gasifier temperatures should be sufficiently high to produce noncondensable tars in order to avoid problems in downstream conversion equipment. Condensable tars must be avoided if the product gas is to be used in engine or gas turbine applications and they will have to be cracked or removed prior to the operation of engines or gas turbines. Exceedingly high temperatures ($>1000^{\circ}C$) may lead to ash sintering and slagging. High operating pressures increase the absolute rate of reaction and the heat and mass transfer, and they shift the gas equilibrium composition in favor of CH_4 and CO_2. The air factor (air flow rate into the gasifier) is a key regulating parameter of a fluidized bed gasifier using air as the gasifying agent. Excessive air factors lead to low heating values (nitrogen dilution and excessive combustion). Insufficient air factors lead to low reactor temperatures and low rates of gasification.

3. GASIFICATION AND GASIFIER TYPES

Three main types of gasification can be distinguished based on the gasifying agent and the way in which the heat for the gasification reactions is provided: directly heated air gasification, directly heated oxygen gasification, and indirectly heated gasification.

In the first two cases, the injected gasifying agent burns part of the feedstock to provide the heat necessary to gasify the remainder of the feed in an air-poor environment. Air gasification leads to a product gas rich in nitrogen (50–65%) and consequently low in calorific value (4–8 MJ/Nm^3). Small-scale gasifiers are usually of the air gasification type, but air gasification may also be the choice for larger scale gasification systems. Oxygen gasification requires an oxygen-producing plant, which increases costs and energy consumption but leads to a producer gas low in nitrogen content and of medium calorific value (10–18 MJ/Nm^3). Steam can be added to both air and oxygen gasification processes to act as a thermal moderator and as a reagent in the gasification process, and it enhances the calorific value of the gas. Oxygen and steam gasification are not required for biomass gasification but are used for the gasification of less reactive fuels such as coal.

Indirectly heated gasifiers do not require air or oxygen input because the heat necessary for gasification is generated outside the gasifier. For example, the bed material can be heated in a separate reactor by burning the char from the gasifier, as in the Battelle process, or the heat can be generated by pulse burners and transferred by in-bed tubes, as in the MTCI process. Indirectly heated gasifiers produce a medium calorific value gas, and steam can be input to the gasifier in order to favor the gasification reaction.

The oxygen or indirectly heated gasification route may be preferred in cases in which the dilution of the product gas with nitrogen may have adverse effects on its downstream uses, for example, in relation to the production of other fuels, possibly for transport applications.

Gasifier design influences the gaseous product with respect to gas composition, condensed liquids, suspended solids, and water content. A number of reactor types are available: fixed bed, fluidized bed, circulating fluidized bed, entrained flow, and molten bath. There are numerous examples of small-scale fixed bed biomass gasification applications throughout the world, mainly coupled with engines for electricity generation. Larger gasifiers, of the fluidized bed type, are in use in niche industrial applications, and there is little operating experience with gasifiers coupled to gas turbines.

4. ACTIVITIES CHARACTERISTIC OF BIOMASS GASIFICATION SYSTEMS FOR HEAT AND ELECTRICITY GENERATION

In general, the conversion of biomass to heat and electricity via gasification involves the following steps: biomass storage and transport, size reduction, drying, feeding, gasification, fuel gas cleaning and cooling, power generation, flue gas cleaning, and ash disposal or recycling. Most of these activities, apart from power generation and flue gas cleaning, are also typical of gasification systems in which the product gas is destined to other end uses (e.g., transport fuel production). This section focuses on equipment and activities typical of medium- to large-scale systems.

4.1 Storage, Transfer, and Pretreatment of the Feedstock

4.1.1 Storage

Biomass storage is required to ensure the continuous operation of the facility. To limit the space required for storage at the plant site, biomass must be stored

in relatively high piles. Two main problems associated with fuel storage are decomposition and self-heating. Self-heating increases the rate of decomposition and fire risk, and it encourages the growth of thermophilic fungi whose spores can cause a respiratory condition in humans similar to farmers lung. Some small biomass losses may occur at the storage stage, but they are likely to be negligible. For intermediary storage of the fuel between the pretreatment (e.g., drying and sizing) and gasification stage, storage silos may be used.

4.1.2 On-Site Biomass Transfer

On-site biomass transfer occurs between storage facilities, pretreatment equipment, and gasification equipment. Such transport technology is readily available, and cranes and enclosed conveyor belt systems may be used. The material flow characteristics of the biomass fuel need to be considered in the design of the on-site biomass transfer and intermediary storage system in order to avoid flow problems.

4.1.3 Size Control

Biomass particle size affects gasification reaction rates and gas composition. Since size control operations are expensive and energy intensive, there is a trade-off, in terms of cost and energy, between particle size reduction and reactor design and the yield and characteristics of the product gas. In practical terms, the size of the feedstock particles is dependent on the biomass requirements imposed by the adopted gasification system.

In the case of wood chips, the acceptable size of the chips fed to the gasifier is usually 2–5 cm, and size reduction may be achieved by using crushers. Size reduction generally makes drying, transfer, and intermediate storage of the biomass easier.

4.1.4 Drying

The moisture content of the feedstock affects the gas composition and the energy balance of the process since gasification is an endothermic process. Water vapor, however, is an essential component of gasification reactions. Therefore, there is a trade-off between the extent of fuel drying and the quality of product gas. Drying of the feedstock to a moisture content of approximately 15% is commonly adopted.

Fuel drying is likely to be the most energy-intensive activity in the gasification process. Important contributions can be made to the energy balance by using flue gases or steam to dry the biomass. The heat used for drying does not have to be high temperature, and a low temperature level is actually desired because it will prevent the evaporation of undesirable organic components. Direct heating systems, in which the heating medium (e.g., flue gas) is in direct contact with the fuel to be dried, based on a rotary drum or fluidized bed design may be used. The systems are also likely to be open, meaning that the heating medium is discharged to the atmosphere.

Drying activities will result in dust emissions. In most cases, a simple baghouse filter is employed to satisfactorily reduce dust emissions. However, in the case of a large drying facility, considerable quantities of water vapor (likely containing significant quantities of organic compounds) could result in addition to dust. A wet gas scrubber followed by a flue gas condensation system may be used to clean the flue gas (mainly of dust and organic compounds such as terpenes) and to recover the heat from the water vapor present in the flue gas. The condensed water requires a biological treatment before it can be discharged to the sewage system. Condensed organic compounds from the fuel-drying activity possess a fuel value; hence, the energy can be recovered from their combustion. Particular care is required in the design of drying installations to avoid fire and explosion risks.

4.2 Feeding the Biomass

Physical properties of the biomass fuel, such as size and density, affect the performance of feeding systems and have often been a source of technical problems. The choice of a feeder depends on the gasifier design, mainly the pressure against which it has to operate. Although some simple small-scale gasifiers may have an open top through which the biomass is fed, screw feeders are commonly used in low-pressure closed-vessel gasifiers.

A screw feeder, in which the screw forms a compact, pressure-retaining plug from the feedstock in the feed channel, is suitable for atmospheric gasifiers. For pressurized gasifiers, a lock-hopper feeder or a lock-hopper/screw-piston feeder is required. The lock-hopper feeder uses a screw feeder but is more complex than a simple screw feeder because of the need to pressurize the feedstock prior to its input into the gasifier. Pressurization of the feedstock is generally achieved using lock-hopper devices, which may require large quantities of inert gases (e.g., liquid nitrogen). Care is required in the design of feeding systems to limit blockages as well as fire and explosion risks.

4.3 Circulating Fluidized Bed Gasification

Air-blown circulating fluidized bed gasifiers are of interest because they produce a good quality, low calorific value (LCV) gas (4–6 MJ/Nm^3) and possess a very high carbon conversion efficiency while allowing high capacity, good tolerance to variations in fuel quality, and reliable operation. The high and homogeneously distributed temperatures and the use of particular bed materials, such as dolomite, favor tar cracking. Successful tar cracking can also be achieved using secondary circulating fluidized bed reactors. Also, successful tests on catalytic tar cracking have been performed, for example, by introducing nickel compounds into the gasifier. Sulfur control is made easier because of the significant reduction that can be achieved by adding limestone or dolomite to the gasifier bed. However, biomass feedstocks are not likely to require sulfur control because of their very low sulfur content, which is likely to meet turbine requirements indicated in Table II. The high fluid velocity entrains large amounts of solids with the product gas that are recycled back to the gasifier via the cyclones to improve conversion efficiency. Carbon conversion efficiencies for circulating fluidized bed gasifiers can reach approximately 98%.

A fluidized bed gasifier consists of a plenum chamber with a gas distributor, a refractory lined steel shell enclosing the granular bed and the freeboard zone, and a feeding device. The bed material consists of a clean and graded fraction of heat-resistant material, generally sand, alumina, limestone, dolomite, or fly ash. It is fluidized by the upward stream of the gasifying agent rising in the form of bubbles, and the continuous motion causes a mixing of the solids and results in a fairly uniform temperature distribution. High superficial velocities of the gas cause elutriation of solid particles leading to a circulating bed design.

The gas composition and heating values do not differ significantly between pressurized and atmospheric operation for similar operating temperatures. The LCV gas (4–6 MJ/Nm^3) consists mainly of inert nitrogen and carbon dioxide gases and the combustible gases carbon monoxide, hydrogen, and methane. Contaminant concentrations (e.g., sulfur, chlorine, alkali and heavy metals) depend on the quality and composition of the biomass and the bed material used.

Pressurized gasifiers require a more complex and costly feeding system (i.e., lock-hopper feeder and inert gas) compared to near-atmospheric gasifiers. Also, pressurized systems require a higher degree of process control. However, if the product gas is to be used for power generation in a gas turbine, pressurized gasifiers considerably reduce the product gas compression requirements. Pressurization is obtained by compressing the gasifying air using the turbine compressor. This results in the compression of a much lower volume of air compared to the product gas volume that would otherwise have to be compressed prior to combustion in the gas turbine. Also, in pressurized gasification systems using hot gas cleanup equipment, there is less chance that tars will condense and cause damage to equipment; thus, they need not be removed from the gas, which allows use of the energy content of tars present in the product gas. Generally, the reduced energy requirement of hot gas cleanup and reduced compression needs of pressurized systems may lead to efficiency gains on the order of 3% compared to systems based on atmospheric gasification and wet gas scrubbing. Hot gas cleanup may also be applied to atmospheric gasification systems.

TABLE II

Notional Gas Turbine Fuel Specifications

Gas heating value (LHV, MJ/Nm^3)	4–6
Gas inlet temperature (°C)	<425
Gas hydrogen content (vol. %)	10–20
Alkali metals (Na + K) (ppmw)	<0.1
H_2S (ppmv)	<100
HCl (ppmw)	<0.5
Naphthalene and tars (ppmv)	<100
Particulates (99% below 10 μm) (ppmw)	<1
Vanadium (ppmw)	<0.1
Combinations: Alkali metals + sulfur (H_2S) (ppmw)	<0.1

4.4 Product Gas Cleaning

The product gas contains a series of impurities (e.g., organic compounds, alkali metals, ammonia, char, and ash). These need to be removed to varying degrees depending on the downstream conversion process and on potential environmental impacts.

Gasification systems in which the product gas is used in engines generally comprise a series of cyclones and physical filters (e.g., baghouse filter). Gasification systems in which the product gas demands more stringent cleanup (e.g., for use in gas turbines) generally comprise catalytic or thermal tar cracking, cyclones, and high-efficiency hot or cold gas cleanup equipment. Physical filtration, in the

form of hot gas filtration systems using ceramic or sintered metal filters, is likely to offer a simpler and less costly option than wet gas scrubbing. It also considerably reduces the water consumption and liquid effluents of the plant. However, unlike gas scrubbing, hot gas filtration systems are not a fully demonstrated and commercial technology. The removal of some polluting and corrosive compounds (e.g., ammonia and chlorine) and of excessive water vapor from the product gas may be problematic when using hot gas filters. However, preliminary testing using hot gas cleanup systems indicates that environmental and gas turbine fuel requirements can be met with hot gas filtration systems. These results are very much fuel dependent.

4.5 Heat and Electricity Generation

The product gas can be burned in boilers to generate heat and raise steam, in internal combustion engines to generate electricity and heat at small to medium scale (from a few kilowatts to a few megawatts), and in gas turbines to generate electricity (Brayton cycle) and heat at small to large scale. In large-scale systems using gas turbines, the exhaust gas from the gas turbine can be used to raise steam in a heat recovery steam generator to generate additional electricity using a steam turbine (Rankine cycle), resulting in combined cycle operation. In a combined heat and power plant, designed for district heating, the flue gas from the combustion of the product gas goes through a heat exchange system to raise the temperature of a heat transport fluid, generally water, circulating in a district heating system. Residual heat in the flue gas can be used to dry the biomass prior to its discharge to the atmosphere.

Factors such as scale, technical performance, capital costs, efficiencies, and emissions will determine the preferred generating technology. Also, the relative demand for heat and electricity will influence the technology choice.

4.5.1 Combustion in Engines

Engines require input of a clean gas to minimize engine wear, avoid tar deposition, and reduce coking in the engine, particularly in the valve area. The gas temperature should be as low as possible to inject a maximum amount of energy into the cylinders.

Engines have the advantage that they can be run on a variety of fuels or fuel combinations with relatively minor adjustments. Gas engines for power generation are commercially available for small- to large-scale applications (from tens of kilowatts to tens of megawatts). Their efficiencies are estimated to range between 25 and 40% depending on the capacity.

There is considerable experience with coupling boilers and engines to gasifiers at small to medium scale ($<10\,MW_e$). The provision of a product gas of suitable quality is the major problem leading to frequent maintenance and shortened engine lifetimes. Engines are commercially available, there is much experience with running engines on a wide variety of gases, and they are generally more tolerable to contaminants than turbines. However, turbines hold promise of higher efficiencies (especially at scales allowing for combined cycle operation), lower costs, and cleaner operation. Also, turbines are likely to be more suited to combined heat and power application because of higher grade heat generation compared to engines.

4.5.2 Combustion in Gas Turbines

Gas turbines cover a wide range of electrical capacities ranging from a few hundred kilowatts to tens of megawatts, with recent developments in microturbines of a few tens to a few hundred kilowatts capacity. However, there is little experience with operating gas turbines in product gas from biomass gasification. Gas turbines have more stringent fuel requirements but may have higher efficiency and lower emissions compared to reciprocating engines. Due to the stringent gas quality requirements of gas turbines, care must be taken in considering the quality and composition of the feedstock, the gasification process, and the gas cleaning system used. The main areas of concern are contamination of the fuel gas by alkali metals, tars, sulfur, chlorine, particulates, and other trace elements; the fuel gas heating value and therefore its composition and volume; the flame properties of the gas within the combustion chamber; and the presence in the fuel gas of fuel-bound nitrogen leading to NO_x formation during combustion. The minimum allowable gas heating value depends on turbine design (heating value affects air-to-fuel ratio and therefore affects the inlet requirements based on total mass flow). Care must be taken to ensure that the fuel gas is supplied to the turbine at a temperature greater than the gas dew point temperature in order to avoid droplet formation. Also, the fuel gas must be unsaturated with water to avoid the formation of acids that could result mainly from the presence of H_2S and CO_2.

Gas turbine fuel requirements are shown in Table II. Table III provides an indicative comparison of fuel gas requirements for different applications.

TABLE III
Product Gas Requirements

Application	Fuel gas temperature	Maximum particulates	Maximum tars	Contaminants
Boilers	High to use sensible heat	Low–moderate	Moderate	Low–moderate
Engine	As low as possible	Very low	Very low	Low
Turbine	As high as possible	None	None	None–low

The strict gas specifications result in low emission levels for most pollutants. Thermal NO_x formation for LCV gas combustion may achieve very low levels (less than 10 ppmv), particularly compared to emissions from natural gas-fired turbines, in which low NO_x burners achieve emissions of approximately 25 ppmv. Low emissions result from the lower combustion temperature and a leaner combustion.

The following factors contribute to the achievement of gas turbine fuel requirements:

1. Biomass feedstock properties:
   - Low (10–15%) moisture content so as not to hamper efficient gasification process and adversely affect the gas heating value.
   - Low ash feedstock to reduce filtering demand and potential slagging.
   - Low alkalinity feedstock to reduce fouling.
2. Gasification process:
   - Choice of gasifying agent (air, oxygen, or steam) influences heating value and can also reduce or eliminate formation of problem tars.
   - Choice of gasifier type influences carryover of particulates.
   - Higher operating temperatures vaporize alkali metals, which must be condensed and filtered out [e.g., vaporization occurs in circulating fluidized beds (900–1100°C) but not in bubbling fluidized beds (600–800°C)], and tars are also vaporized and cracked to a certain extent (e.g., circulating fluidized beds crack more tar compounds than do bubbling fluidized beds).
3. Gas cooling and cleaning system:
   - Special bed materials and catalysts in the gasifier or in a separate reactor can be used for sulfur and chlorine removal, to avoid the formation of ammonia, and for tar cracking.
   - Gas cooling is necessary to protect gas turbine components from heat damage, and the degree of cooling depends on the gas cleaning technique adopted (e.g., hot gas filter or wet gas scrubbing).
   - Hot gas filter systems remove particulates and alkali metals.
   - Tars may not need to be removed from the gas if the temperature remains higher than the gas condensation temperature.
   - Wet gas scrubbing removes most fuel gas contaminants (ammonia is not removed by hot gas filters but can be washed out of the gas by an acidic solution in a wet gas scrubber).

Pressurized gasifiers are well suited for hot gas cleanup (below 600°C) prior to direct combustion in the turbine. At temperatures below 600°C, alkali metals precipitate onto the particulates present in the gas and are removed by the gas cleaning system. Tar cracking is not required for pressurized gasification systems because of the elevated gas temperature (the tars are likely to be in noncondensable form). In the case of hot gas cleaning, fuel-bound nitrogen, in the form of ammonia, may cause concern regarding NO_x emissions since approximately 60% of the ammonia in the gas may be converted to NO_x during combustion. Atmospheric gasifiers generally require tar cracking or removal, cold (<200°C) gas cleaning, and compression of the gas prior to turbine combustion.

4.5.2.1 Combined Cycle Operation Single cycle efficiencies of approximately 36% and combined cycle efficiencies of 47% or higher (up to approximately 52%) are typical of natural gas-fueled plants. Integrated biomass gasification and single or combined cycle efficiencies, relative to the energy content of the fuel input, will not match the efficiencies of gas turbines fueled with natural gas because combustion temperatures are lower and part of the energy content in the biomass fuel is dissipated in the production of the fuel gas. Table IV shows notional gas and steam turbine generating efficiencies and Table V provides an indication of biomass integrated gasification/combined cycle (BIG/CC) efficiencies.

4.5.3 Comparison with Biomass Direct Combustion

Gasification has a number of advantages compared to direct combustion, particularly at the relatively

TABLE IV

Notional Generating Equipment Efficiencies[a]

Equipment	Efficiency (%)
GT running on NG	36
GT running on LCV gas	31
ST generating system ($>50\,MW_e$)	35
ST generating system ($<50\,MW_e$)	15–35

[a] Abbreviations used: GT, gas turbine; NG, natural gas; LCV, low calorific value; ST, steam turbine.

TABLE V

Notional BIG/CC Electrical Efficiencies

	Capacity (%)	
	$<100\,MW_e$	$>100\,MW_e$
Atmospheric system	43	50
Pressurized system	46	53

small scales typical of biomass-to-energy systems. Engines and single or combined cycle gas turbines are likely to have higher electrical conversion efficiencies compared to steam cycles. At small scale, steam turbines have lower efficiencies compared to engines and gas turbines, and at a larger scale gasification offers the possibility of combined cycle operation. Also, direct combustion and steam turbine systems are characterized by important economies of scale, possibly making gasification systems coupled with engines and gas turbines a more likely economically viable option at small to medium scale (from a few hundred kilowatts to a few tens of megawatts).

4.5.4 Comparison with Coal Gasification

Coal has a number of advantages and disadvantages with respect to biomass. On the one hand, it is easier and less costly to handle and process; it is easier to feed, particularly in pressurized systems; and it may require less complex supply logistics. On the other hand, coal has a much lower level of volatiles than biomass (typically 30% compared to more than 70% for biomass); its char is significantly less reactive than biomass char, implying greater reactor sizes, residence times, and the use of oxygen as a gasifying agent; it possesses a higher ash content than biomass (the pro and it often contains significant quantities of sulfur, requiring more complex and costly gas cleaning systems. Coal also has a major environmental disadvantage with respect to biomass in that it produces large CO_2 emissions, contributing to global warming.

4.6 Waste Disposal

Wastewater and solid waste result from power plant operation and require treatment prior to disposal or recycling. Liquid effluents resulting, for example, from wet gas scrubbing and flue gas condensation must be treated. The wastewater treatment is conventional, although oxygenated organics such as phenols (derived from tars) and ammonia may create problems. Systems using hot gas filtration reduce liquid effluents and may potentially reduce costs and environmental impacts.

Solid residues consist of inert ash. The ash is likely to be disposed of conventionally through landfilling or recycled, for example, as a soil nutrient. Care must be taken, however, because in some cases dust from fly ash may be toxic due to the absorption of chemicals such as benzo[*a*]pyrene. Also, heavy metal concentrations in the ash may in some cases be too high for direct recycling of the ash. Research programs in Sweden have obtained promising results with regard to ash recycling from wood gasification and combustion systems. In most cases, ash can be recycled without any processing other than wetting and crushing it.

5. ECONOMICS OF BIOMASS GASIFICATION FOR HEAT AND ELECTRICITY GENERATION

The economics of biomass gasification for electricity and heat generation is the principal factor influencing its future market penetration. The Costs of semi-commercial gasification systems integrated with engines at a scale of approximately $300\,kW_e$ are estimated to be approximately $\$1800/kW_e$. Costs could be reduced by approximately 20% with continued commercialization. Larger scale BIG/CC systems are in the demonstration stage and characterized by high costs in the range of \$5000–6,000/kW_e. However, commercialization of the technology is expected to reduce systems costs to $\$1500/kW_e$ for plants with capacities between 30 and $60\,MW_e$.

Capital costs alone do not provide sufficient basis for comparison between different technologies and energy sources. Variable costs, particularly fuel costs, are of key importance. These are also affected by the

efficiency of the plants. Biomass fuel costs can vary considerably depending on type and location. For biomass energy to be competitive, an indicative cost of biomass of $2/GJ or less is often cited. However, biomass costs may range from negative (i.e., in the case of waste products) to significantly in excess of the suggested cost.

6. CONSTRAINTS AND ISSUES AFFECTING THE DEVELOPMENT OF BIOMASS GASIFICATION SYSTEMS

Gasification systems for power generation are in the semicommercial or demonstration stage, and their market development will depend on a number of factors ranging from technical developments to energy market structure and government policies with regard, for example, to energy security and the environment.

A number of technical issues must be addressed with regard to gasification systems. Demonstration projects are valuable in this respect and technical problems do not appear to be a major obstacle to the development of the systems. More operating experience is required, particularly with different types of biomass fuels.

Gasification systems are characterized by relatively high investment costs and, in some cases, high biomass fuel costs. However, gasification offers advantages in terms of scale, efficiency, and emissions that make it an attractive renewable energy technology. Technology development, replication, learning by doing, and competition will lead to cost reductions. The identification of opportunities for early market penetration is key for the successful commercialization of the technology.

Efficient biomass supply logistics and low-cost biomass resources are very important in the development of biomass energy systems. The issue of security of fuel supply (i.e., ensuring a fuel supply for the lifetime of the plant) needs to be addressed and is of particular importance in the case of energy crops.

The regulatory environment is also a key factor in the successful development of biomass gasification facilities. Regulations that facilitate decentralized power generation are necessary. Also, regulations or incentives aimed at reducing emissions may favor integrated gasification systems, which are likely to be characterized by low emission levels of regulated pollutants and no CO_2 emissions from the conversion of biomass.

Biomass gasification may also provide a source of clean transport fuels ranging from synthetic diesel to hydrogen.

SEE ALSO THE FOLLOWING ARTICLES

Biodiesel Fuels • Biomass, Chemicals from • Biomass Combustion • Biomass for Renewable Energy and Fuels • Biomass: Impact on Carbon Cycle and Greenhouse Gas Emissions • Biomass Resource Assessment • Combustion and Thermochemistry • Renewable Energy, Taxonomic Overview

Further Reading

Bridgwater, A. V. (1995). The technical and economic feasibility of biomass gasification for power generation. *Fuel* **74**(5), 631–653.

Bridgwater, A. V., and Evans, G. D. (1993). "An Assessment of Thermochemical Conversion Systems for Processing Biomass and Refuse." Energy Technology Support Unit, UK.

Brown, A. E., and van den Heuvel, E. J. M. T. (1996). "Producer Gas Quality Requirements for IGCC Gas Turbine Use—A State-of-the-Art Review." Biomass Technology Group, Enschede, The Netherlands.

Consonni, S., and Larson, E. D. (1996). Biomass gasifier/aeroderivative gas turbine combined cycles: Part B—Performance calculations and economic assessment. *Trans. ASME* **118**.

International Energy Agency Bioenergy Program. www.ieabioenergy.com.

Prabhu, E., and Tiangco, V. (1999). The Flex-Microturbine for biomass gases—A progress report. In "Biomass: A Growth Opportunity in Green Energy and Value-Added Products. Proceedings of the 4th Biomass Conference of the Americas, August 29–September 2, Oakland CA" (R. Overend, *et al.*, Eds.). Elsevier.

Stassen, H. E. (1995). Small-scale biomass gasifiers for heat and power: A global review, World Bank Technical Paper No. 296. World Bank, Washington, DC.

Sydkraft. (1998). Värnamo demonstration plant: A demonstration plant for biofuel-fired combined heat and power generation based on pressurised gasification—Construction and commissioning 1991–1996. Sydkraft/Elforsk/Nutek, Sweden.

Wilen, C., and Kurkela, E. (1997). Gasification of biomass for energy production—State of the technology in Finland and global market perspectives, Research Note No. 1842. Technical Research Centre of Finland (VTT), Espoo, Finland.

Biomass: Impact on Carbon Cycle and Greenhouse Gas Emissions

CARLY GREEN and KENNETH A. BYRNE
University College Dublin
Dublin, Ireland

1. **Biomass and Carbon Cycling**
2. **Climate Change and International Policy**
3. **Bioenergy and CO_2 Emissions**
4. **Bioenergy Sources**
5. **Managing Biomass as a Carbon Sink**
6. **Potential Role of Wood Products**
7. **Increasing Global Importance**
8. **Concluding Remarks**

Glossary

afforestation Direct, human-induced conversion of land that has not been forested for a period of at least 50 years to forested land through planting, seeding, or the human induced promotion of natural seed sources.

autotrophic respiration Carbon loss as a result of internal plant metabolism, which typically amounts to half of that fixed through photosynthesis.

bioenergy Renewable energy produced from biomass. Biomass is organic nonfossil material of biological origin, for example, forest residues, wood and wood wastes, agricultural crops and wastes, municipal and industrial wastes. Biomass may be used directly as a fuel or processed into solid, liquid, or gaseous biomass fuels.

biological carbon sequestration The uptake and storage of carbon in the biosphere. Trees and other plants absorb carbon dioxide through photosynthesis, release the oxygen, and store the carbon in their biomass.

carbon dioxide (CO_2) A naturally occurring gas, produced through respiration, decomposition, and combustion. The rise in atmospheric CO_2 concentration is attributed to burning of fossil fuels and deforestation. It is the principal anthropogenic greenhouse gas that affects the earth's temperature and it is the reference gas against which other greenhouse gases are indexed.

carbon sink Any process, activity or mechanism that removes CO_2 from the atmosphere.

carbon source Any process or activity that releases CO_2 to the atmosphere.

CO_2 equivalent A measure used to compare the emissions of different greenhouse gases based upon their global warming potential (GWP). The CO_2 equivalent for a non-CO_2 greenhouse gas is calculated by multiplying the quantity of the gas by the associated GWP. For example, the GWP for carbon dioxide is 1 and for methane is 21. Therefore 1 ton of methane is equivalent to 21 tons of carbon dioxide.

CO_2 flux The rate of transfer of CO_2 between carbon reservoirs. Positive values are fluxes to the atmosphere; negative values represent removal from the atmosphere.

coppice A method of regenerating a stand in which trees in the previous stand are cut and the majority of new growth is associated with vigorous sprouting from stumps, roots, or root suckers.

gigatonne C (GtC) The unit of measure for stock and fluxes of carbon in the global carbon cycle. 1 GtC = 1000 million tonnes of carbon.

greenhouse gas (GHG) Atmospheric gases, both natural and anthropogenic, that absorb and reemit infrared radiation. Greenhouse gases include water vapor, carbon dioxide (CO_2), methane (CH_4), nitrous oxide (N_2O), halogenated fluorocarbons (HCFCs), ozone (O_3), perfluorinated carbons (PFCs), and hydrofluorocarbons (HFCs).

net primary production (NPP) The difference between photosynthesis and autotrophic respiration and is referred to as net primary production (NPP). It constitutes the total annual growth increment (both above and below ground) plus the amounts grown and shed in senescence (the growth phase from full maturity to death), reproduction or death of short-lived individuals, plus the amounts consumed by herbivores.

reference energy system Used as a base case in a life-cycle assessment, a reference energy system should consider how electricity, heat, or transportation fuels would have been produced in the absence of the bioenergy system.

reservoir A component of the climate system where a greenhouse gas is stored.

thinning Removal of trees to reduce stand density primarily to concentrate growth on fewer stems, thereby increasing production of high-value timber.

 © 2004 Elsevier Inc. All rights reserved.

Large amounts of carbon are stored in the earth's biosphere as biomass. Humans have long made use of the various forms of biomass carbon for energy; however, measures to mitigate the enhanced greenhouse effect have stimulated renewed interest in its utilization. The role of biomass in modern energy production and its carbon sink potential can significantly influence the contemporary global carbon cycle resulting in potential reductions in greenhouse gas (GHG) emissions to the atmosphere. The various application of biomass in this context and its potential contribution to stabilizing GHG emissions into the future is explored.

1. BIOMASS AND CARBON CYCLING

Biomass accumulates in the earth's biosphere in vegetation following the process of photosynthesis. Powered by energy from sunlight, plants use carbon dioxide (CO_2) from the atmosphere and water from the soil to manufacture carbohydrates, a storable energy source. Oxygen (O_2) and water vapor are released to the atmosphere as by-products. CO_2 is cycled back to the atmosphere as the plant material decays, is respired or is burned, and therefore becomes available again for photosynthesis.

The average annual amount of solar energy potentially available at the earth's surface is estimated at $180\,Wm^{-2}$. The proportion available for photosynthesis varies both temporally and spatially, being dependent on latitude and season. The photosynthetic conversion efficiency of land plants is estimated at 3.3 to 6.3%, varying between plant species. Taking these constraints into consideration, the energy stored in biomass globally is substantial with photosynthetic carbon fixation comprising a major component of the global carbon cycle.

1.1 The Global Carbon Cycle

The global carbon cycle is made up of carbon reservoirs (stocks) and the dynamic transfer of carbon between them (fluxes). The complex system is summarized in a simplified carbon cycle (Fig. 1).

The total carbon stored in the earth's surface is estimated at more than 650,000,000 GtC. The majority is stored in glacial deposits and minerals, fossil fuels, and the deep oceans. Atmospheric carbon accounts for a further 750 GtC. Due to their heterogeneity, biomass carbon stocks are difficult to measure empirically. The carbon stored in soil (which includes detritus, soil organic matter, and inert carbon) are 1400 GtC with plants accounting for a further 550 GtC.

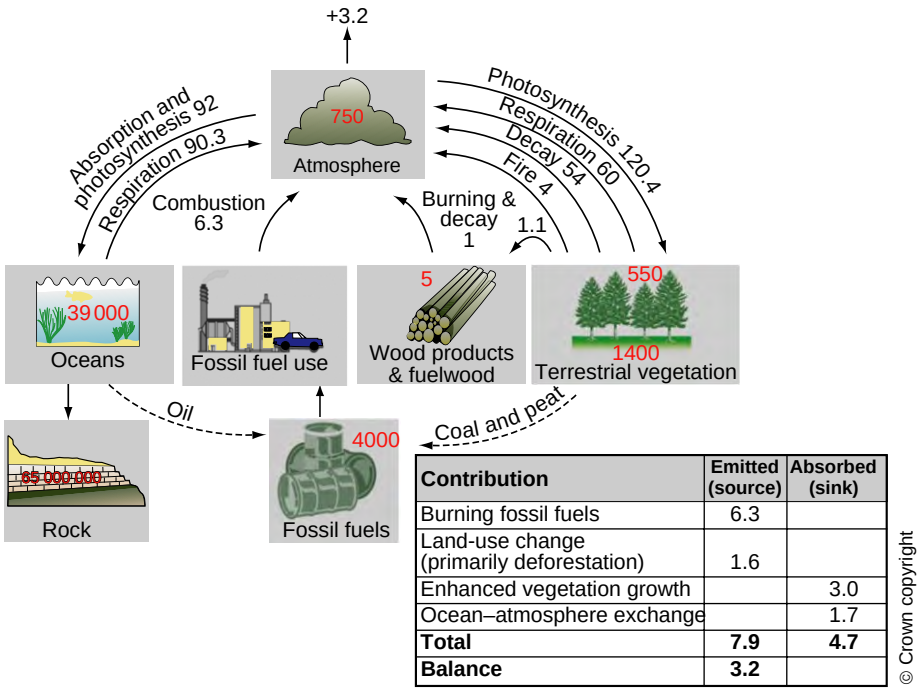
Various natural and anthropogenic processes stimulate CO_2 fluxes between the reservoirs. The overall net flux of the global carbon cycle indicates that carbon is accumulating in the atmosphere at an average rate of $3.2\,GtCyr^{-1}$. Natural carbon fluxes are driven primarily by plant photosynthesis, respiration, and decay as well as oceanic absorption and release of CO_2. Anthropogenic processes such as burning fossil fuels for energy, as well as deforestation contribute a significant net positive flux to the atmosphere.

The largest anthropogenic flux within the global carbon cycle is caused by the anthropogenic burning of fossil fuels. During the 1990s, this source was reportedly $6.3\,GtCyr^{-1}$ and is considered the main cause of large increases in atmospheric CO_2 concentrations over the past 100 to 150 years. The largest natural flux is experienced between the atmosphere and the ocean where the ocean acts as a net sink of $1.7\,GtCyr^{-1}$. The terrestrial biosphere is a slightly smaller net sink. Incorporating both natural processes and human-induced activities, the flux from atmosphere to biosphere is estimated at $1.4\,GtCyr^{-1}$, representing the annual net accumulation of carbon in the biosphere.

1.2 Net Primary Production

Net primary production (NPP) is a measure of the annual productivity of the plants in the biosphere. The 550 GtC of carbon in the global reservoir of plant biomass has a corresponding NPP of $60\,GtC\,yr^{-1}$, which indicates that globally the average carbon residence time on land is approximately 9 years with an average biomass production (NPP) of $4\,tCha^{-1}\,yr^{-1}$. However, average NPP, as reported by the Intergovernmental Panel on Climate Change (IPCC) (Table I) and calculated residence time varies greatly at ecosystem level and is dependent on plant species and management, ranging from 1 year for croplands to approximately 15 years for forests.

Production forests are comparatively stable ecosystems, experiencing a longer growth cycle than food and energy crops. Food and energy crops are usually harvested at the end of a relatively short rapid growth phase, leading to a higher average NPP per unit area. In contrast to stable forest ecosystems, the majority of the NPP associated with energy and



Contribution	Emitted (source)	Absorbed (sink)
Burning fossil fuels	6.3	
Land-use change (primarily deforestation)	1.6	
Enhanced vegetation growth		3.0
Ocean–atmosphere exchange		1.7
Total	**7.9**	**4.7**
Balance	**3.2**	

© Crown copyright

FIGURE 1 The simplified global carbon cycle; carbon reservoirs (GtC) represented within the boxes and associated fluxes ($GtCyr^{-1}$) indicated by solid arrows between the reservoirs. The broken lines represent long-term (millennia) movement of carbon within the system. A simplified summary of the global carbon budget is also given. Values are for the 1990s. The most important natural fluxes in terms of the atmospheric CO_2 balance are the land-atmosphere exchange ($+1.4\,GtCyr^{-1}$), which includes land-use change and in enhanced vegetation growth, and the physical ocean-atmosphere exchange ($+1.7\,GtCyr^{-1}$). Fossil fuel burning ($-6.3\,GtCyr^{-1}$) represents the largest anthropogenic release of CO_2 to the atmosphere. The net result of these processes is an increase in atmospheric CO_2 concentration of $+3.2\,GtCyr^{-1}$. From Broadmeadow and Matthews (2003).

TABLE I

Terrestrial Carbon Stock Estimates, NPP and Carbon Residence Time Globally Aggregated by Biome

Biome	Area (10^9 ha)	Global carbon stocks (GtC)	Carbon density (tC/ha)	NPP (GtC/yr)	Residence time[a]
Tropical forests	1.75	340	194	21.9	15.5
Temperate forests	1.04	139[b]	134	8.1	17.2
Boreal forests	1.37	57	42	2.6	22.1
Tropical savannas and grasslands	2.76	79	29	14.9	5.4
Temperate grasslands and Mediterranean shrublands	1.78	23	13	7.0	3.3
Deserts and semi-deserts	2.77	10	4	3.5	3.2
Tundra	0.56	2	4	0.5	4.5
Croplands	1.35	4	3	4.1	1.0
Total	14.93[c]				

Source. Modified from Mooney *et al*., 2001.

[a] Average carbon residence time = (carbon density × area)/NPP.

[b] Based on mature stand density.

[c] 1.55×10^9 ha of ice are included.

food crops ends up in products exported from the site. Soil carbon density is generally lower and may become depleted as a result.

1.3 The Available Biomass Resource

In determining the extent of the biomass resource, an understanding of the global allocation to land use is required. As the values shown in Table I are globally aggregated by biome, they do not identify competing land use functions such as areas for human settlement, recreation, and conservation of biodiversity, and ecosystem function. Estimates of the possible contribution of biomass in the future global energy supply vary from 100 to 400 EJ yr^{-1} by 2050. The variation in estimates can be largely attributed to land availability and energy crop yield levels. In terms of GHG emission benefits, such an increase in bioenergy production could lead to an overall reduction of between 2 and 6.2 GtC per year as a result of the direct substitution of fossil fuels. Projections suggest that annual GHG emissions associated with fossil fuel combustion will exceed 11.4 GtC by the year 2050 if consumption continues in accordance with trends. Therefore biomass can potentially provide GHG emissions avoidance in the order of 17.5 to 54% by 2050.

In contrast, the potential carbon sink through vegetation management has been estimated at between 60 and 87 GtC by 2050. This is equivalent to the sequestration of 14 to 20% of the average fossil fuel emissions for the period 2000 to 2050.

2. CLIMATE CHANGE AND INTERNATIONAL POLICY

Climate change caused by the enhanced greenhouse effect is one of the most significant global environmental issues. Increased emissions of GHG to the atmosphere, most notably CO_2, are considered the main cause of global climate change. Increasing energy consumption, a reliance on fossil fuels to meet these needs, and deforestation related to land use change are the main sources of increasing atmospheric CO_2. The ability of biomass to make a significant contribution to the stabilization of atmospheric CO_2 concentrations through its utilization as an energy source and the preservation of terrestrial biomass stocks has lead to its inclusion in global policy developments.

2.1 Global Environmental Policy

Global concern over the increase in atmospheric GHG concentrations and the acknowledgment that human activity is leading to climate change led to nations recognizing the need for international effort. The United Nations Framework Convention on Climate Change (UNFCCC) was agreed at the Rio Earth Summit in 1992 and entered into force in 1994. Parties agreed to report on GHG emissions and to take action to mitigate climate change by limiting those emissions and to protect and enhance GHG sinks and reservoirs. The Kyoto Protocol to the UNFCCC, developed in 1997, set the relatively short-term target for developed countries and countries with economies in transition to stabilize their combined GHG emissions, in CO_2 equivalents, to 5.2% below 1990 levels during the first commitment period 2008–2012. Actions that enhance carbon stock in biomass and soil, such as reforestation and altered management in forestry and agriculture, can be used as credits to offset GHG emissions to meet commitments under the Kyoto Protocol.

Biomass can address the issue of global climate change in three ways: (1) bioenergy produced from biomass can replace fossil fuels, directly reducing emissions; (2) biomass can sequester carbon as it is growing, increasing the terrestrial carbon stock; and (3) biomass products can substitute for more energy intensive building products, reducing the emissions associated with production of these materials and additionally increasing the carbon stock in wood products.

Offsetting fossil fuel emissions with sequestration in biomass and reducing fossil fuel emissions through substitution of renewable bioenergy are likely to be key factors in the ability of countries to achieve their reduction commitments under the Kyoto Protocol.

2.2 Carbon Neutrality

Bioenergy produced from biomass is sometimes called a carbon-neutral energy source, because the same quantity of carbon released when the biomass is burned is sequestered again when the crop or forest is regrown (Fig. 2). Referring to bioenergy as carbon neutral or having zero net emissions may be misleading; there are emissions associated with producing the biomass, such as from fossil fuel used in cultivation, harvest, processing and transport, and in manufacture and construction of fuel conversion technology. Then there are less obvious emissions, such as the fossil fuel emissions in producing fertilizer and

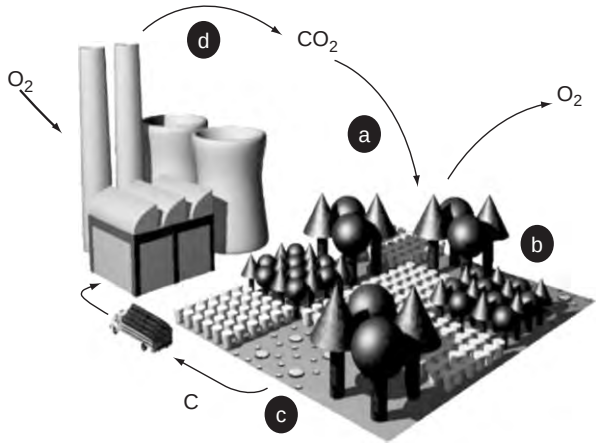


FIGURE 2 Illustration of the apparent carbon neutrality of a bioenergy system. (a) CO_2 is captured by the growing crops and forests; (b) oxygen (O_2) is released and carbon (C) is stored in the biomass of the plants; (c) carbon in harvested biomass is transported to the power station; (d) the power station burns the biomass, releasing the CO_2 captured by the plants back to the atmosphere. Considering the process cycle as a whole, there are no net CO_2 emissions from burning the biomass. From Matthews and Robertson (2002).

herbicide. Although it is clearly not a zero-emissions process, the greenhouse impact of bioenergy is much lower than that of fossil fuel energy sources.

The production, processing, transport, and conversion of fossil fuels also requires additional fossil fuel inputs; the significant difference between fossil fuels and biomass in terms of GHG emissions is the ability of biomass to "recycle" the emissions released in fuel conversion acting as a closed loop process.

GHG emissions and removals are reported to the United Nations Framework Convention on Climate Change (UNFCCC) by sector. Fuel use is reported separately from agricultural activities and forestry removals. The emissions associated with particular forestry and agricultural operations are not explicit in national GHG inventories; nevertheless, they are reported in consumption figures. Recommendations for future reporting suggest that full life cycle accounting be undertaken for projects, so that credits will be earned only for net removals. Therefore policy is being formulated that will make such auxiliary emissions completely transparent.

3. BIOENERGY AND CO_2 EMISSIONS

Substituting fossil fuels with renewable energy sources offers a significant opportunity for reducing GHG emissions. Renewable bioenergy has a number of benefits over other renewable technologies such as wind and solar as it is capable of supplying the full range of energy markets (i.e., heat, electricity and transport fuel).

3.1 Role of Bioenergy

Biomass has the longest history of any energy source and still provides approximately 44 $EJyr^{-1}$ to meet 11% of the worlds primary energy needs. Most of this is utilized in developing countries where traditional low efficiency applications of fuel wood for cooking and heating account for 38 EJ. The remaining 6 EJ is utilized in the developed countries for heat and power generation with modern technologies.

In the past, the most common way to capture energy from biomass was through combustion. However, advancements in technology mean that more efficient and cleaner methods for energy conversion are readily available. Solid, liquid, and gaseous biofuels can replace fossil fuels in almost every application.

The extent and way that biofuels are utilized varies greatly. Only 3% of the total primary energy needs in industrialized counties are supplied by biomass compared, with over 40% in the developing world. A comparison of the contribution that biomass makes to various countries' energy requirements, Table II shows a wide range, from 1% in Japan, Ireland, and The Netherlands to 94% in Tanzania and Ethiopia. Of the industrialized countries, Finland, Sweden, and Austria source relatively high levels of their energy requirements from bioenergy, which can be largely attributed to the use of wood chips, industrial wood waste, and straw as fuel for district heating.

3.2 Bioenergy Life Cycles

The quantification of the actual reduction in GHG emissions resulting from the substitution of fossil fuels with biomass requires a complete lifecycle assessment (LCA). A systematic framework for estimating the net GHG emissions from bioenergy systems and comparing them against the fossil fuel reference system that it would replace has been developed (Fig. 3). The major considerations of the life cycle assessment approach to quantifying the greenhouse impacts of bioenergy are as follows:

- *Carbon stock dynamics*. Changes in the carbon stock of plants, plant debris, and soils result from

TABLE II

Bioenergy Consumption and Resulting Greenhouse Gas Benefits

Country	Total energy consumption (TJ)	Bioenergy consumption (TJ)	Energy supply share (%)	Emission avoidance ($GtCO_2$)[a]
Austria	1162	104	9	7
Japan	21558	312	1	20
Sweden	2174	348	16	22
Finland	1385	246	18	18
Ireland	523	7	1	0
Netherlands	3136	41	1	3
Australia	4255	219	5	14
United States	90527	2841	3	183
Brazil	7203	1693	24	109
China	46010	8720	19	562
Egypt	1657	52	3	3
India	19302	8072	42	520
Mozambique	321	290	90	19
Tanzania	597	561	94	36
Ethiopia	717	675	94	43

[a] Emissions factor = 64.4 $kgCO_2$/GJ. Values are for 1997. (From World Resources Institute, 2003.)

biomass growth and harvest. This necessitates time-dependent analysis, as such changes might extend over long periods of time until a new equilibrium is reached.

- *Trade-offs and synergies*. The trade-off between competing land uses, namely biomass production, food production, and carbon sequestration, should be considered in determining the benefits of bioenergy. A timber plantation, thinned to maximize value of wood products, where thinning residues are used for bioenergy, is an example of synergy.
- *Leakage*. The use of biomass fuels does not always avoid the use of fossil fuels to the full extent suggested by the amount of bioenergy actually used and therefore the reduction in fossil fuel use that can be attributed to the project will be reduced. This is commonly referred to as leakage.
- *Permanence*. Irreversible GHG mitigation offered by the system is required to permanently offset the CO_2 emissions. The greenhouse mitigation benefit from substituting bioenergy for fossil fuel use is irreversible. In contrast, the mitigation benefit of reforestation will be lost if the forest biomass is reduced by harvest or natural disturbances.
- *Emissions factors*. The quantity of GHG produced per unit of fossil fuel energy consumed. This factor influences the net benefit of a bioenergy project: if bioenergy displaces natural gas, the benefit is lower than if it displaces coal, as coal has a higher emissions factor than natural gas.
- *By-products*. The emissions and offsets associated with both products and by-products of a biomass production system must be considered.
- *Efficiency*. In determining the quantity of fossil fuel energy displaced by a quantity of biomass feedstock, it is important to know the energy output per unit of that feedstock.
- *Upstream and downstream GHG emissions*. Auxiliary (energy inputs) emissions resulting from the bioenergy and reference systems should be considered, such as from fossil fuels used in production/extraction, processing, and transport of the feedstock.
- *Other greenhouse gases*. The emissions of other GHG associated with biomass and fossil fuel chains, such as CH_4 and N_2O, expressed in CO_2 equivalents, should be included in the calculation of net greenhouse mitigation benefits.

LCA assessments have been applied to a variety of fuel systems to generate an understanding of their contribution to atmospheric CO_2 concentrations. For example, full fuel cycle CO_2 emissions per GWh for coal have been calculated as 1142 t compared to 66 to 107 t for biomass.

4. BIOENERGY SOURCES

Biomass used directly or indirectly for the production of bioenergy can be sourced from any organic matter,

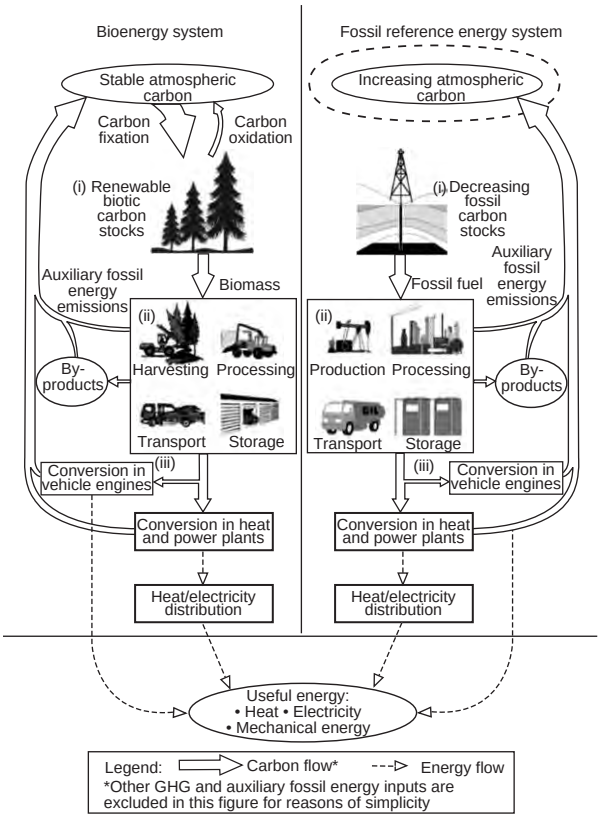


FIGURE 3 Life cycle assessment (LCA) approach for calculation of GHG balances. The components of the comparison between the (a) Bioenergy system and (b) fossil reference system to reach the (c) useful energy stage are shown. The system boundary incorporates the (i) carbon flows related to the fuel source *in situ*, (ii) carbon flows involved in supplying the fuel for energy conversion including any by-products or auxiliary energy requirements, and (iii) carbon and energy flows associated with conversion technologies to useful energy. Reproduced from IEA Bioenergy (2002).

of plant or animal origin, and can be categorized into two types: dependent or dedicated biomass resources. Dependent biomass resources are by-products of other activities (e.g., wood residues and recovered wood waste, straw, forestry residues, livestock manure, sugarcane, and recoverable waste cooking oil). Dedicated biomass resources are energy crops grown specifically for fuel (e.g., short rotation tree crops, coppice, herbaceous grasses, and whole crop cereals).

4.1 Bioenergy from Dependent Resources

Agricultural and forestry residues provide the largest proportion of biomass used for the production of bioenergy. Some estimates suggest that globally available biomass in the form of recoverable residues represents about 40 $EJyr^{-1}$, enough to meet 10% of the total present energy use of 406 $EJyr^{-1}$. However, realizing this potential is limited by factors such as ease and cost of recovery and environmental concerns relating to sustainable land use practices.

In developed countries, sawmill and pulp mill residues comprise the largest proportion of residues in the existing bioenergy mix. Industries make use of recovered operational waste such as sawdust and off cuts in timber mills and black liquor in pulp and paper production to supply process energy needs. Additionally, techniques such as densification can produce a usable fuel from sawmill sawdust and bark as well as shavings from furniture manufacturing.

Substantial opportunities exist in the forestry sector. Forest management residues including branches, tree tops, and small diameter stems from thinning and harvest of commercial plantations are a potentially large source of biomass, which is underutilized. Full utilization of thinnings, which may represent approximately 25% of the biomass produced, permits energy recovery from a resource that may otherwise decay on the forest floor. In both cases, the carbon in the biomass is returned to atmosphere, but where bioenergy is produced, greenhouse mitigation benefit is obtained from displacement of fossil fuel emissions. Technological advancements are being made in this area, making use of bundling technology for the densification of harvest residues. This process, largely being developed in Finland, involves recovery of the residues and bundling at source to produce an easily transportable and storable fuel.

Used vegetable oils and animal fats can be reprocessed into biodiesel. Biodiesel has similar physical properties to conventional diesel to which it can be blended or alternatively used in its pure form. Biodiesel is commercially available at service stations in countries such as the United States, Brazil, France, Germany, and Austria. In 1996, the International Energy Agency (IEA) reported that 21 countries produced some 591,000 ton of biodiesel. Direct substitution of diesel oil would result in the mitigation of over 500,000 ton of CO_2 emissions. As approximately 20% of global CO_2 emissions can be attributed to the transportation sector, further development of biofuels, including biodiesel, can contribute significantly to reduction in GHG emissions from this sector.

Wet waste streams such as domestic, agricultural, and industrial sludge and animal manure emit

methane (CH_4), as a product of anaerobic decomposition of organic matter. CH_4 has a greenhouse warming potential 21 times higher than that of CO_2 and therefore, avoidance of CH_4 emissions produces significant greenhouse benefits. Anaerobic digestion is a waste management and energy recovery process highly suited to wet waste streams. The bacterial fermentation of organic material produces biogas (65% CH_4, 35% CO_2), which is captured and utilized to produce heat and power. Although anaerobic digestion of sewage, industrial sludge and wastewater are fully commercialized, systems involving the processing of animal manure and organic wastes are still being developed. Technology can generate approximately 200 kilowatt-hours (kWh) of electricity from 1 ton of organic waste. If this was used to replace electricity from coal production, direct GHG emissions could be reduced by approximately 220 $kgCkWh^{-1}$.

The energy contribution of animal sludge and municipal solid waste globally was estimated at 5300 to 6300 MW in 1995, 95% of which can be attributed to Asia. Globally, anaerobic digestion of organic wastes is predicted to increase to 8915 to 20130 MW in 2010. This projection is largely attributed to the developed world increasing its contribution.

Energy production from solid municipal waste offers a potentially sustainable approach to a growing environmental problem. Such sources include urban wood waste (i.e., furniture, crates, pallets) and the biomass portion of garbage (i.e., paper, food, green waste.). The biogas emitted when waste decomposes in landfill can be extracted for use in the generation of electricity. Maximizing the potential of such waste streams is hampered by logistical issues relating to recovery, negative public perception of "incineration," and opposition from environmental groups to the establishment of a market for waste.

4.2 Bioenergy from Dedicated Resources

The future development of energy crops, to the level at which they would replace residues as the major bioenergy fuel source, will be largely dependent on regional factors such as climate and local energy requirements and emission factors, which will determine their environmental and financial viability.

Energy production per hectare is determined by crop yields. This is in turn dependent on climate, soil, and management practices in addition to the number of harvests per year, the fraction of the biomass harvested, and the usable fraction of the harvest. The conversion efficiency to a useable energy form is a further important consideration.

Generally perennial crops produce higher net energy yields than annual crops, due to the lower management inputs, and therefore lower production costs, both economic and ecological. Management techniques, fertilizer application, and genetic improvements may increase the productivity by a factor of 10 compared with no inputs; however, this can result in increased environmental problems. Eutrophication of water can be associated with fertilizer runoff; soil erosion with intensive cultivation, and concerns regarding consequences for biodiversity with genetically improved crops.

There is sometimes a misconception that it takes more energy to produce biomass for bioenergy than is generated from the biomass itself. However, it has been repeatedly shown that such systems are capable of producing significant energy ratios (Table III).

For example, the energy production per hectare of woody biomass from a commercial plantation is 20 to 30 times greater than the energy input for agricultural operations such as fertilizer and harvesting. However, liquid fuel production from energy crops may experience conversion rates that are substantially less. In Europe and the United States for example, achieving a high-energy output per unit input requires careful management at the production stage. Ethanol produced in the U.S. yields about 50% more energy than it takes to produce, whereas in Brazil where ethanol is produced from sugarcane, higher yields are possible largely as a result of greater rates of biomass growth and lower auxiliary energy inputs.

4.3 Modernizing Bioenergy

Estimation of the future technical potential of biomass as an energy source is dependent on assumptions with respect to land availability and productivity as well as conversion technologies. With the emergence of energy crops as the major source of biomass fuel, land use conflicts, especially in relation to food production, may arise. However, with efficient agricultural practices, plantations and crops could supply a large proportion of energy needs, with residues playing a smaller role without compromising food production or further intensifying agricultural practices.

Considering future projections of increasing energy and food demands associated with increasing standard of living and population, pressure on the availability of land is likely to increase. In the European Union (EU) it has been proposed that

TABLE III

Current and Feasible Biomass Productivity, Energy Ratios, and Energy Yields for Various Crops and Conditions

Crop and conditions	Yield (dry t $ha^{-1}yr^{-1}$)	Energy ratio[a]	Net energy yield[b] (GJ $ha^{-1}yr^{-1}$)
Short rotation woody crops (willow, hybrid poplar; United States, Europe)			
Current	10–12	10:1	180–200
Feasible	12–15	20:1	220–260
Miscanthus/ switchgrass			
Current	10–12	12:1	180–200
Feasible	12–15	20:1	220–260
Sugarcane (Brazil, Zambia)	15–20	18:1	400–500
Wood (commercial forestry)	1–4	20/30:1	30–80
Sugar beet (northwest Europe)			
Current	10–16	10:1	30–100
Feasible	16–21	20:1	140–200
Rapeseed (including straw yields; northwest Europe)			
Current	4–7	4:1	50–90
Feasible	7–10	10:1	100–170

[a] Energy ratio = Unit of bioenergy produced per unit of fossil fuel input. Fuel for processing and transporting the biomass is excluded, except for the ration of sugarcane in Brazil and Zambia.

[b] Net energy yield = Energy output − Energy input (i.e., agricultural operations, fertilizer, and harvest operations).

From World Energy Assessment (2002).

biomass could contribute an additional 3.8 EJ annually by 2010, compared with the contribution of 1.9 EJ. Energy crops are expected to provide this extra capacity on 4% of the total land area of the EU. If all of this 1.9 EJ replaced energy generated from coal in Western Europe, reductions of net CO_2 in the order of 50 $MtCyr^{-1}$ or more than 7% of the current anthropogenic CO_2 emissions from the region could be expected.

Heat and power generation through combustion is still the most common method of generating bioenergy, using the same technology used in power plants burning solid fossil fuels. Co-firing biomass with coal in existing electricity generation plants is becoming more widespread and may be one of the most promising near-term options for increasing market share of bioenergy. With low capital input, fuels such as waste wood, crop debris, forestry residues, and agricultural wastes can be mixed with coal to reduce net GHG emissions. The technique is being developed in the European Union and the United States where biomass has been successfully co-fired with coal at around 10% without compromising efficiency. If 10% of the 50,000 PJ of electricity generated each year from coal were to be replaced with biomass in only 10% of installations, approximately 350 $MtCO_2yr^{-1}$ emitted to the atmosphere could be offset. Co-firing offers additional environmental benefits associated with reduced NO_x and SO_x emissions from conventional generating plants.

Technological advancements in the production of liquid biofuels such as pyrolysis are at the pilot stage, and new processes for the production of ethanol from woody biomass have been developed, however, they are not yet commercial. In developing countries, efficiency of conversion technologies is also being increased. For example, in India, traditional cooking fires are being replaced with improved higher efficiency cooking stoves, and biogas produced from anaerobic digestion of animal and other wastes is now used for cooking, lighting, and power generation. Modernizing biomass production and bioenergy conversion technologies will transform this traditional energy source to one that is feasible and widely acceptable and ultimately offer significant GHG reductions.

Although the majority of energy sourced from biomass is converted through combustion, and some through liquification to alcohol, the adoption of new technologies such as gasification can improve the power production efficiency of biomass. Increasing efficiencies in conversion and the global economy will determine to what extent biofuels can displace CO_2 emissions into the future.

5. MANAGING BIOMASS AS A CARBON SINK

Another possible role for biomass as a means for mitigating atmospheric GHG is through carbon sequestration in the growing biomass and soil. The rapid rates of assimilation of CO_2 by vegetation provide opportunities for sequestering carbon in the relatively short term. Forests can capture and retain a more significant stock of carbon in biomass (40 to

120 $tCha^{-1}$) compared with grassland (7 to 29 $tCha^{-1}$) or cropland (3 $tCha^{-1}$), underlying the importance of the role of forests preserved as a carbon reservoir.

Afforestation and reforestation activities have a role in reducing atmospheric CO_2 concentrations. The Kyoto Protocol requires countries to report such activities and allows carbon stock increases due to post-1990 afforestation and reforestation to be offset against emissions. If the forest is maintained, it does not continue to sequester CO_2 indefinitely; however, it provides a perpetual sink for the quantity of carbon represented by the difference in average carbon density between the previous and the forested land use. Utilizing forests to sequester carbon provides some relatively short-term solutions to rising atmospheric CO_2 concentrations.

5.1 Carbon Dynamics in Forest Ecosystems

The accumulation of carbon within a forest growth cycle can be considered in four stages. The initial establishment stage involves low carbon accumulation and may even experience net carbon loss (particularly from soil) as a result of site preparation and low biomass inputs. A rapid uptake of carbon is then experienced during the second phase, known as the full vigor stage, which subsequently levels off as the stand reaches the mature stage. Finally, the forest reaches old growth and the carbon is in steady state with accumulation associated with new growth balanced by mortality and disturbances.

Modeling management for forest conservation in the United Kingdom indicates that the maximum accumulation of carbon, and the subsequent removal of CO_2 from the atmosphere, over the life cycle of the stand is approximately 200 $tCha^{-1}$ (Fig. 4a). Forest stands managed for commercial production experience periodic harvesting and generally have lower carbon stocks than stands that are not harvested. Under such conditions, commercial stands rarely experience growth phases past the full vigor phase. Over a typical 50-year rotation, commercially managed Sitka spruce stands yield average carbon accumulation rates of 70 $tCha^{-1}$ (Fig. 4b). Although these figures indicate that the carbon stock increases even under periodic harvesting, the longer term C stock depends on a balance between the impacts of harvesting and the rate of forest regeneration. Consideration of the long-term average C stocks in these two examples indicates that even though young, fast-growing trees experience high sequestration rates, over the longer term, average carbon stocks are higher in old growth forests.

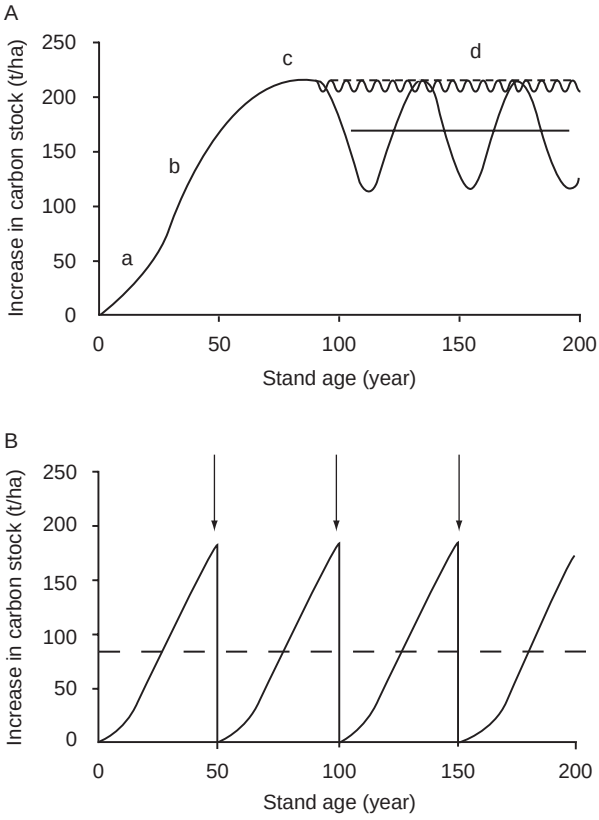


FIGURE 4 (A) Carbon accumulation in a newly created stand of trees managed as a carbon sink, based on an average stand of Sitka Spruce in Britain. Four stages of growth are shown: (a) establishment phase, (b) full-vigor phase, (c) mature phase, (d) long-term equilibrium phase. Following an initial increase in carbon stock due to the establishment of the stand, carbon stocks neither increase nor decrease because accumulation of carbon in growing trees is balanced by losses due to natural disturbances and oxidation of dead wood onsite. Two examples of carbon dynamics with low (dotted line) and high (dashed line) long-term equilibrium stocks are illustrated. Carbon dynamics in soil, litter, and coarse woody debris are ignored. (B) Carbon accumulation in a newly created commercial forest stand. Periodically the stand is felled (indicated by vertical arrows) and subsequently replanted. Over several rotations, carbon stocks neither increase nor decrease because accumulation of carbon in growing trees is balanced by removals due to harvesting. In practice, a forest usually consists of many stands, all established and harvested at different times. Averaged over a whole forest, therefore, the accumulation of carbon stocks is more likely to resemble the time-averaged projection shown as a dashed line. Carbon dynamics in soil, litter, coarse woody debris, and wood products are ignored. From Matthews and Robertson (2002).

Based on these two different management examples, between 5000 and 14,000 ha of newly established forests would be required to sequester 30 years of CO_2

emissions from a 30 MW power station burning fossil fuels. In contrast, approximately 11,250 ha of dedicated energy crops could supply such a power plant and indefinitely replace fossil fuels, providing permanent GHG emission reductions.

There may be potential to promote carbon sinks globally over the next 100 years or more to the extent of sequestering 90 GtC, but increasing carbon stocks in vegetation will ultimately reach ecological or practical limits and other measures will need to be adopted. Forests have a finite capacity to remove CO_2 from the atmosphere. The provisions of the Kyoto Protocol with respect to sinks can be seen as a valuable incentive to protect and enhance carbon stock now, while possibly providing the biomass resources needed for continued substitution of fossil fuels into the future.

5.2 Permanence

If a forest is harvested and not replanted (deforestation) or is permanently lost as a result of fire or disease, then the carbon reservoir is lost. Whatever safeguards may be in place to protect against future loss of stored carbon, it is impossible to guarantee indefinite protection, and therefore any emissions reductions are potentially reversible and therefore not guaranteed to be permanent.

For a sink to offset the emissions from the factory, the forest would have to be maintained in perpetuity. The issue of permanence has become increasingly prominent in discussions on measures to reduce national and international GHG emissions. These discussions focus around the use of accounting systems to trace land areas ensuring that any loss in carbon storage is identified and reported. This raises the issue of intergenerational equity, where land use choice may be foreseen as a liability to future generations.

In contrast, the GHG benefits resulting from the utilization of biomass as an energy source are irreversible. Even if the bioenergy system operates for a fixed period, the offset effects are immediate and are not lost even if the energy system reverts back to fossil fuels.

6. POTENTIAL ROLE OF WOOD PRODUCTS

To date, most discussion and research relating to the various roles of biomass in mitigating CO_2 emissions has been focused around its use as a fuel or as a sink. However, full utilization of the potential of biomass products, particularly from woody biomass, may provide significant opportunities.

Wood products themselves are a carbon pool and can be a sink if increasing demand for wood products results in increased carbon storage. The potential carbon sink in wood products is estimated to be small compared to the carbon sink in living vegetation and biomass, or compared to the potential of wood products to displace fossil fuel consumption; however, utilizing the full range of options may enable such products to play a significant role.

Indirectly, biomass could displace fossil fuels by providing substitute products for energy intensive materials such as steel, aluminum, and plastics (Fig. 5). Although, product substitution will be dependent on practical and technically feasible applications at a local to regional level, there is a large potential to replace a range of materials in domestic and industrial applications.

The construction industry is one example where maximizing the use of biomass materials may provide GHG emissions reductions of between 30 and 85% in the construction of housing for example. However, over its life cycle, house-heating requirements contribute 90% of the total associated GHG emissions, therefore the use of bioenergy for domestic heating may provide a more significant role.

Disposal options of wood products at the end of their usable life, such as landfill or combustion, can influence CO_2 emissions. Calculations suggest that a

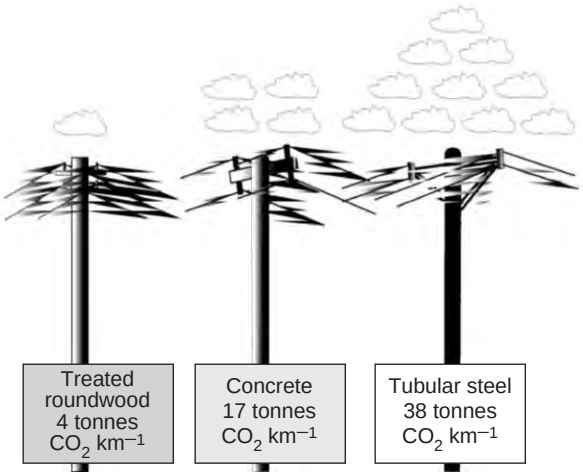


FIGURE 5 Biomass can be used as product substitutes such as in the construction of transmission lines. In CO_2-equivalents, 1 km using treated round wood would emit 4 tonnes $CO_2\,km^{-1}$, compared to concrete 17 tonnes $CO_2\,km^{-1}$ and 38 tonnes $CO_2\,km^{-1}$ for tubular steel. This analysis was based on a 60-year life cycle and includes the impact of disposal. From Matthews and Robertson (2002).

significant sink may exist should wood products be disposed to landfill. Such products decompose to a very limited extent in landfill with over 90% of the carbon apparently stored permanently. Alternatively, wood products could be collected and used as a combustible "carbon neutral" fuel.

7. INCREASING GLOBAL IMPORTANCE

The IPCC considers that increasing the use of biomass across all scenarios is important in the long-term mitigation of greenhouse gas emissions. Increasing the share of biomass in the total energy market to between 25 to 46% by 2100 could be possible. In the biomass intensive scenario (46%), cumulative emissions between 1990 and 2100 could be as low as 448 GtC compared with 1300 GtC in the business-as-usual case.

As discussed in Section 5, vegetation, particularly forests, can be managed to sequester carbon. Through global afforestation and management, the forestry sector could have the potential to sequester up to 90 GtC over the 21st century with policy developments encouraging such activity. However, each use need not compete and preventative and reactive measures to mitigating atmospheric CO_2 emissions could be jointly applied.

7.1 Possible Synergistic Opportunities

Synergistic opportunities exist, often on a regional scale, between bioenergy, wood production, and management as a carbon sink. For example, establishing a forest or crop on former degraded land can provide a source of biomass, which can be harvested for bioenergy or wood products. As a direct result of this establishment, the aboveground, belowground, and soil carbon stock may be increased to create a sink. Alternatively, management approaches aimed at increasing productivity, such as thinning a commercial forest stand, may provide specified amounts of fuel wood for bioenergy. The wood harvested from the stand could be used and reused in a number of products, being available as a fuel at the end of its useful life. Depending on the objectives, the choice of biomass crop and application is likely to be regionally specific based on environmental, economic, and social factors (Fig. 6).

In the production of biomass, high yields are important; therefore fast-growing crops and short rotation forestry are best suited, which in some cases also have a small carbon sink capacity by increasing soil carbon. Such crops are more suited to providing pulp, wood chip, and some timber production. Residues provide some options for bioenergy and a small sink capacity. However, if the intention were long-term storage of the carbon stock, maximizing the carbon ultimately attained and retaining it indefinitely, rather than the rapid capture of carbon in the initial stages, then enduring tree species and the promotion of regeneration and succession would be preferable.

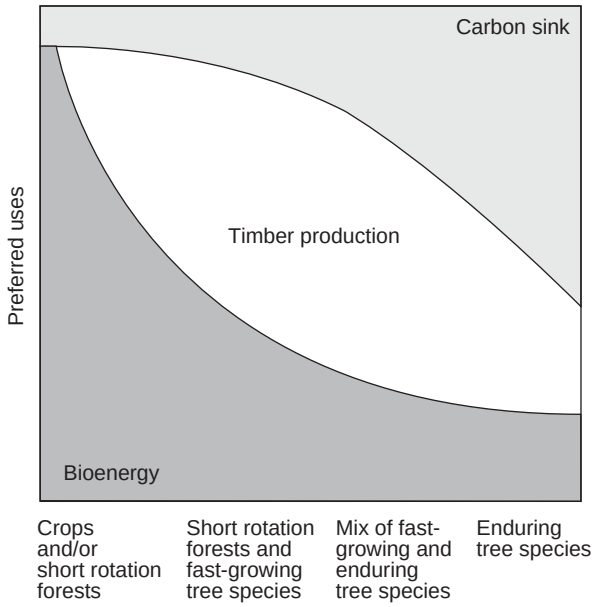


FIGURE 6 The choice of crops, type of tree species, and management regime can be selected to achieve a mix of bioenergy production, wood production, and carbon sink. From Matthews and Robertson (2002).

Although sequestration of carbon in biomass sinks may be promoted under climate change policy, a preventative approach, reducing total GHG emissions by substituting fossil fuels either directly with bioenergy or indirectly as an alternative to energy intensive materials, could yield more stable and less problematic reductions.

7.2 Balance of Benefits

The future potential of biomass to provide a substantial contribution to reducing GHG emissions will depend on a number of environmental, technical, social, and economic factors.

Land availability and competition may be the largest limiting factor to maximizing the potential of biomass. Offsetting CO_2 emissions using biomass, either through carbon storage or directly by fuel

substitution, requires large areas of land. Competition, especially for food production, is likely to occur and vary widely from region to region. Large areas of degraded land may be available in developing countries, and through sustainable planning, such areas could be restored as energy crops while still encouraging biodiversity and enhancing the local environment.

Biomass is often seen as a fuel of the past and one that is associated with poverty and poor health as traditional uses have not been sustainable. However, when managed in a sustainable manner ensuring that all issues related to its practical exploitation are carefully considered, bioenergy can contribute significantly to the development of renewable energy markets in both the developed and less developed countries.

Bioenergy is competing in a heavily subsidized market. Unless cheap or negative cost wastes and residues are used, bioenergy is rarely competitive. For large-scale energy crops to make a substantial contribution to the energy market and subsequently to reducing GHG emissions, they need to be competitive. Carbon and energy taxes may encourage this transition. More efficient conversion technologies may also contribute through the provision of higher conversions at lower cost. Further development of combined heat and power as well as gasification technologies will be crucial. The development of dedicated fuel supply systems, including higher yields, greater pest resistance, better management techniques combined with reduced inputs, and improvements in harvesting, storage, and logistics will be required to lower costs and raise productivity without contributing to nutrient and organic matter depletion.

8. CONCLUDING REMARKS

As energy demands grow and pressure from international conventions and protocols lead to the requirement for lower GHG emissions, biomass may provide an important part of an integrated renewable energy future, through direct supply or as a substitute for energy intensive materials. Biomass will also provide options for increasing the biospheric sink capacity, helping to reduce atmospheric concentrations of CO_2 in the coming decades.

Forests, forest resources and energy crops can affect the global cycling of carbon in several ways. Carbon dynamics in a number of areas such as carbon storage in terrestrial reservoirs, fossil fuel displacement, auxiliary energy requirements associated with biomass management and conversion processes and the various applications of biomass products can influence the concentration of atmospheric CO_2. To fully appreciate the effect that biomass can have in terms of the carbon cycle and GHG emissions, a detailed assessment of global, regional, and local conditions taking a life-cycle approach should determine whether it is better utilized as a sink, as fuel, or as long-lived timber products.

The success of biomass, in the future of renewable energy and sustainable production, will be strongly influenced by policy and economic incentives to stimulate reductions in GHG emissions. Evaluating the options of carbon management through the use of biomass will require a number of considerations including prospective changes in land use, sustainable management practices, forest and crop productivity rates, as well as the manner and efficiency with which biomass is used. Regardless of the focus, biomass has a major role in the global energy system of the future.

Acknowledgments

Financial support for this study was provided by the International Energy Agency (IEA) Implementing Agreement Task 38—Greenhouse Gas Balances of Biomass and Bio-energy Systems. We are grateful to Annette Cowie and to Gregg Marland for reviewing the manuscript, Robert Matthews for assisting in the preparation of the figures, and for the many useful comments provided by the National Task Leaders of IEA Bioenergy Task 38.

SEE ALSO THE FOLLOWING ARTICLES

Biomass, Chemicals from • Biomass Combustion • Biomass for Renewable Energy and Fuels • Biomass Gasification • Biomass Resource Assessment • Carbon Sequestration, Terrestrial • Climate Change and Energy, Overview • Diet, Energy, and Greenhouse Gas Emissions • Greenhouse Gas Emissions, Alternative Scenarios of • Greenhouse Gas Emissions from Energy Systems, Comparison and Overview

Further Reading

Berndes, G., Hoogwijk, M., and van de Broek, R. (2003). The contribution of biomass in the future global energy supply: A review of 17 studies. *Biomass and Bioenergy* **25**(1), 1–28.

Broadmeadow, M., and Matthews, R. (2003). The Carbon Dynamics of Woodland in the UK. Information Note. UK Forestry Commission, Edinburgh.

Gustavsson, L., Karjalainen, T., Marland, G., Savolainen, I., Schlamadinger, B., and Apps, M. (2000). Project based

greenhouse-gas accounting: Guiding principles with a focus on baselines and additionality. *Energy Policy* **28,** 935–946.

Hall, D. O., and Scrase, J. I. (1998). Will biomass be the environmentally friendly fuel of the future? *Biomass and Bioenergy* **15**(4 and 5), 357–367.

Hoogwijk, M., Faaij, A., van den Broek, R., Berndes, G., Gielen, D., and Turkenburg, W. (2003). Exploration of the ranges of the global potential of biomass for energy. *Biomass and Bioenergy* **25**(2), 119–123.

IEA Bioenergy Task 38 (2002). "Greenhouse Gas Balances of Biomass and Bioenergy Systems." http://www.joanneum.ac.at/iea-bioenergy-task38/publications/

IPCC (2001). Climate change 2001. IPCC Third Assessment Report. Intergovernmental Panel on Climate Change.

IPCC (2002). "Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change" (J. T. Houghton, Y. Ding, D. J. Griggs, M. Noguer, P. J. van der Linden, X. Dai, K. Maskell, and C. A. Johnson, Eds.). Cambridge, UK, and New York, NY.

Kheshgi, H. S., Prince, R. C., and Marland, G. (2000). The Potential of Biomass Fuels in the Context of Global Climate Change: Focus on Transportation Fuels. *Annual Rev. Energy and the Environment* **25,** 199–244.

Kirschbaum, M. (2003). To sink or burn? A discussion of the potential contributions of forests to greenhouse gas balances through storing carbon or providing biofuels. *Biomass and Bioenergy* **24**(4 and 5), 297–310.

Matthews, R., and Robertson, K. (2002) Answers to ten frequently asked questions about bioenergy, carbon sinks and their role in global climate change. IEA Bioenergy Task 38. http://www.joanneum.ac.at/iea-bioenergy-task38/publications/faq/

Mooney, H., Roy, J., and Saugier, B. (eds.) (2001). "Terrestrial Global Productivity: Past, Present, and Future." Academic Press, San Diego, CA.

Schlamadinger, B., and Marland, G. (1996). The role of forest and bioenergy strategies in the global carbon cycle. *Biomass and Bioenergy* **10,** 275–300.

Schlamadinger, B., and Marland, G. (2001). The role of bioenergy and related land use in global net CO_2 emissions. *In* "Woody Biomass as an Energy Source: Challenges in Europe" (P. Pelkonen, P. Hakkila, T. Karjalainen, and B. Schlamadinger, Eds.) EFI Proceedings No. 39, pp. 21–27. European Forest Institute, Finland.

United Nations Development Programme (UNDP) (2000). World Energy Assessment: Energy and the Challenge of Sustainability. United Nations Publications, New York.

World Resources Institute (2003). Earth Trends. The Environmental Information Portal. Found at http://earthtrends.wri.org

Biomass Resource Assessment

MARIE E. WALSH
Oak Ridge National Laboratory
Oak Ridge, Tennessee, United States

1. **Introduction**
2. **Approaches to Biomass Resource Assessment**
3. **Example of an Integrated Supply Curve Approach for Biomass Resource Assessment**
4. **Discussion of Limitations and Conceptual Differences in Biomass Resource Assessments**
5. **Summary and Conclusions**

Glossary

agricultural residues Include the stems, leaves, cobs, husks, some roots, and chaff of agricultural crops such as corn, wheat, and other small grains as well as rice, sugarcane, and oilseed crops.

animal wastes Mostly manure (dung), but may include carcasses and industrial processing wastes (whey from milk separation).

biomass resource assessment An estimate of the quantities of biomass resources available by location and price levels given existing and potential physical, technical, environmental, economic, policy, and social conditions.

biomass supply curve An estimate of the quantity of biomass resources that could be available as a function of the price that can be paid for them; can be static (model variables are independent and do not vary in response to changes in each other) or dynamic (model variables change in response to changes in each other) as well as deterministic (there is no variability in model parameters) or stochastic (model parameters display a distribution of values).

energy crops Include a number of grass resources (herbaceous crops) such as switchgrass (*Panicum virgatum*), *Miscanthus,* reed canary grass, *Arundo donax,* and bamboo, as well as fast-growing short-rotation trees such as hybrid poplar (*Populus spp*), *Eucalyptus,* and willows (*Salix spp*).

forest industry residues Include those residues collected from forest operations such as logging residues, forest thinning, and removal of cull, dead, and dying material; also may include those residues produced at primary mills (e.g., saw mills, pulp mills, veneer mills) and secondary mills (e.g., furniture and cabinet makers).

fossil fuels Include coal, oil, and natural gas and comprise approximately 86% of the primary energy used in the world.

other miscellaneous biomass resources Include captured gases from decomposing plant and animal materials (biogas) that occur in landfills, sewage treatment plants, and feedlots; food processing wastes (e.g., olive pits, nut shells), other fibers (e.g., cotton gin, clothing remnants), and the wood component of the municipal solid waste and construction and demolition waste streams; also may include oils from oilseed crops (e.g., soybeans, rapeseed) and industry (e.g., restaurant cooking fats) and starch derived from agricultural food and feed crops (e.g., corn, wheat, other small grains).

Biomass resource assessments estimate the quantities of resources available by location and price levels given existing and potential physical, technical, environmental, economic, policy, and social conditions. Factors that affect the quantities and prices of biomass resources include physical limitations (e.g., soil and water limitations for energy crops, access to forests in remote areas), technical limitations (e.g., efficiency of harvest machinery to collect the resource, energy crop yields), environmental limitations (e.g., quantities of agricultural residues needed to control soil erosion and soil carbon for long-term productivity, wildlife diversity needs), economic considerations (e.g., competition for land between energy crops and traditional crops, alternative uses for biomass resources), policy considerations (e.g., crop subsidies, renewable portfolio standards that require a certain percentage of energy generation to come from biomass, emission regulations), and social factors (e.g., ethical values, aesthetic values, NIMBY [not in my back yard]).

 © 2004 Elsevier Inc. All rights reserved.

1. INTRODUCTION

Biomass resources are widely used throughout the world, mostly for home cooking and heat but also for industrial uses such as heat, steam, and electricity. Estimates indicate that for the year 2000, 11% (~ 1095 million tons oil equivalent) of the world's total primary energy supply (TPES) came from biomass. Regional percentages of TPES ranged from a low of 0.3% in the Middle East to a high of 48% in Africa, with biomass accounting for 18% of TPES in China, 31% in the rest of Asia, and 17% in Latin America. Biomass as a percentage of TPES in non-Organization for Economic Cooperation and Development (OECD) Europe, the former Soviet Union, and OECD countries accounted for an estimated 5, 1, and 3%, respectively.

Biomass resources include a wide range of plant and animal materials available on a renewable or recurring basis. Most assessments focus on a few key resources such as forest industry residues, agricultural residues, and animal wastes—the most widely available and used resources in the world. Many assessments also evaluate potential new energy crops such as grass crops (herbaceous resources, e.g., switchgrass [*Panicum virgatum*], *Miscanthus,* reed canary grass, *Arundo donax,* bamboo) and fast-growing short-rotation trees (e.g., hybrid poplar [*Populus spp*], *Eucalyptus,* willows [*Salix spp*]). Forest industry residues include those collected from forest operations such as logging residues, forest thinning, and removal of cull, dead, and dying material. Additional forest industry residues include those produced at primary mills (e.g., saw mills, pulp mills, veneer mills) and secondary mills (e.g., furniture and cabinet makers). Agricultural residues include the stems, leaves, cobs, husks, some roots, and chaff of agricultural crops such as corn (stover), wheat, and other small grains (straw) as well as rice, sugarcane, and oilseed crops. Animal wastes are mostly manure (dung).

In addition to these biomass resources, food processing wastes (e.g., olive pits, nut shells); other fibers (e.g., cotton gin, clothing remnants); the wood component of municipal solid waste, construction and demolition waste, and yard trimmings; oils from oilseed crops (e.g., soybeans, rapeseed) and industry (e.g., restaurant cooking fats); and starch from agricultural food and feed crops (e.g., corn, wheat, other small grains) are available. Captured gases from decomposing plant and animal materials (i.e., biogas) in landfills, sewage treatment plants, and feedlots are also considered to be biomass resources.

The focus on energy crops (particularly trees), wood from forest and mill operations, agricultural residues, and animal manure in most studies is due not only to their worldwide availability but also to their assumed uses as steam, heat, and/or electricity. These resources are readily combustible and are well suited to these uses. However, biomass resources can be used to produce numerous other products, such as liquid and gaseous transportation fuels (e.g., ethanol, methanol, compressed natural gas, bio-diesel, hydrogen for fuel cells), and bioproducts currently produced from fossil fuels (e.g., coal, oil, natural gas), such as organic chemicals (e.g., plastic precursors, acetic acid, lactic acid, ethylene), new composite fiber materials (e.g., wood-plastic, wood-fiberglass), lubricants, pharmaceuticals, cosmetic ingredients, coatings (e.g., lacquers, varnishes, paints), and printing inks, among others. The key biomass resources can be used to produce many of these products as well, but some of the described products use biomass resources other than energy crops, forest industry residues, agricultural residues, or animal dung.

Interest in bioenergy and bioproducts is increasing because biomass is an abundant renewable resource that, when used in high-efficiency systems, can reduce greenhouse gas emissions (e.g., SO_x, N_2O, CO_2) relative to fossil fuel systems. Its use can also enhance economic development, particularly in rural areas, and can provide increased farm income, improve soil quality, be a viable option (e.g., energy crops) for production on degraded and abandoned lands, and offer the greatest flexibility (among renewable resources) in terms of the products that can be produced. This flexibility provides opportunities to establish highly diverse bio-based industries that can displace imported oil from politically unstable regions, enhancing energy security for some countries.

Three basic approaches are used for biomass resource assessments: an inventory approach that estimates quantities only, a quantity and associated cost approach that estimates either an average or range of quantities with an associated average or range of production and/or collection costs, and a supply curve approach that estimates biomass quantities as a function of the price that can be paid. The supply curve approach is generally the best suited to conduct integrated assessments that evaluate changes in physical, technical, environmental, economic, policy, and social factors. The potential availability of biomass resources for bioenergy and bioproducts generally decreases as one goes from the inventory approach to the supply curve approach.

The complexity of the models (integrated supply curves) and data limitations become more severe as

the analysis becomes more geographically disaggregated, that is, as it goes from the global level to the local level. However, biomass resource quantities and prices are influenced significantly by local conditions (e.g., soil quality and weather for agricultural residues, forest residues, energy crops) and the low energy density of biomass resources (i.e., energy available vs volume and/or weight of the resource) that substantially increase transportation costs as distance increases. Conducting biomass resource assessments at the smallest geographic unit feasible is preferable.

2. APPROACHES TO BIOMASS RESOURCE ASSESSMENT

The inventory, quantity and associated cost, and supply curve approaches can be applied to all biomass resources. However, throughout this article, one resource category—energy crops—is used to demonstrate each approach. Hypothetical examples are presented to facilitate simplicity and transparency; however, the costs and yields used in the examples are within the yield and production cost ranges that can be realistically achieved.

2.1 Inventory Approach

The inventory approach is the most common method used to evaluate biomass resources. This approach generally estimates physical quantities only and does not include costs or prices that can be paid. It is used for local, regional, national, and global analysis. The degree of sophistication ranges from simple local surveys and/or observation of local use, to GIS satellite imaging, to solar radiation combined with soil type, water and nutrient availability, and weather conditions to estimate biomass growth potential. Lack of available data (especially in developing countries) is a major limitation and constrains the types and geographic scale on which assessments can be made. Inventory approaches are highly useful in establishing theoretical maximum quantities of biomass resources.

Although most inventory studies estimate the maximum physical quantities of biomass, some adjust these quantities to estimate available resources by including physical, technical, economic, and/or environmental factors. Physical limitations include areas where temperature, soil, and water conditions limit crop production opportunities. Technical considerations include limits to collection machinery efficiency and physical inaccessibility such as forest harvest at high altitudes. Environmental constraints include leaving a portion of agricultural and forest residues to maintain soil quality and reduce erosion, although most studies use very general rather than detailed local approximations to make this adjustment. Economic constraints generally include subtracting quantities of biomass currently being used.

A study by Hall and colleagues illustrates this approach. They estimated reasonable current yields of short-rotation woody crops that can be produced in the midwestern United States. Theoretical maximum quantities based on photosynthetic efficiency (a function of the amount of visible light that can be captured, converted to sugars, and stored) are first estimated. Losses from plant respiration are subtracted. Biomass production rates are estimated next by using the energy contained in the visible light that reaches an area (e.g., square meters) and multiplying by the photosynthetic efficiency. For short-rotation trees, this results in approximately 104 dry metric tons per hectare per year (dMT/ha/year) (where dry is 0% moisture). The theoretical maximum is then adjusted for temperature, and this reduces biomass yields to 60 dMT/ha/year. Adjustments for the parts of the trees that will be used for energy (e.g., trunk, branches) further reduce biomass yields to 35 dMT/ha/year. Given additional adjustments for the light intercepted during the growing season, and assuming sufficient water, nutrients, and pest and disease control, expected yields of 10 to 15 dMT/ha/year are reasonable. Genetic improvements could potentially double these yields. Total biomass is estimated as the expected yields multiplied by the land available to support these yield levels. Results from selected inventory studies are shown in Table I. Units are those reported by the authors of the study, with no attempt to adjust to a common measure.

2.2 Quantity and Associated Cost Approach

The quantity and associated cost approach includes an inventory approach component but extends the assessment to include biomass cost estimates in addition to physical quantities. This approach has been used in local to global assessments and varies substantially in complexity and sophistication. Studies range from examining one biomass resource (e.g., an energy crop) to evaluating

TABLE I

Selected Biomass Quantities: Inventory Approach

Reference number	Biomass resources included	Geographic region	Biomass quantities
1	Forest industry residues, agricultural residues, wood energy crops, animal wastes	Global	266.9EJ/Year
2	Forest industry residues, agricultural residues, animal wastes, biogases	Non-OECD countries	879.6 billion tons oil-equivalent
3	Forest industry residues, agricultural **residues**	Latin America, Caribbean	664,225 Kboe
4	Forest industry residues, agricultural residues, energy crops	Asia	2.51 billion tons
5	Forest industry residues(fuel wood)	World	472.3 million tons oil-equivalent

Sources. (1) Hall, D., Rosillo-Calle, F., Williams, R., and Woods, J. (1993). Biomass for energy: supply prospects. *In* "Renewable Energy: Sources for Fuels and Electricity" (T. Johansson, H. Kelly, A. Reddy, and R. Williams, Eds.), pp. 593–651. Island Press, Washington, DC. (2) Denman, J. (1998). Biomass energy data: System methodology and initial results. *In* "Proceedings of Biomass Energy: Data, Analysis, and Trends" Paris, France. (3) Hernandez, G. (1998). Biomass energy in Latin America and the Caribbean: Estimates, analysis, and prospects. *In* "Proceedings of Biomass Energy: Data, Analysis, and Trends." Paris, France. (4) Koopmans, A., and Koppejan, J. (1998). "Agriculture and Forest Residues: Generation, Utilization, and Availability," Food and Agricultural Organization Regional Wood, Energy Development Programme in Asia. (5) World Energy Council, (2001). "Survey of Energy Resources." London (www.worldenergy.org)

multiple resources (e.g., energy crops, forest residues, agricultural residues, animal wastes) in the geographic area examined.

As a first step, these studies create an inventory of available biomass resources using techniques similar to those in the inventory approach and often include some physical, technical, economic, and environmental adjustments. Biomass resource quantities are estimated generally as an average amount or a range. This approach can be applied to all of the biomass resources, but energy crops are used to illustrate it.

The average approach estimates the amount of resource available as the number of hectares that can be used to produce energy crops times the average yield per hectare. That is, if 10,000 ha are available in a region and the average yield is 10 dMT/ha, the total quantity of biomass available is 100,000 dMT. The range approach generally estimates available quantities as a minimum and a maximum (with a minimum expected yield of 5 dMT/ha and a maximum of 15 dMT/ha, for 10,000 ha, energy crop quantities range from 50,000 to 150,000 dMT).

Biomass resource costs are estimated to correspond to the average, minimum, and maximum quantities. For the average approach, the costs of management and production practices associated with achieving the expected yield level are estimated. If expected yields are 10 dMT/ha, estimated cost is U.S. $40/dMT, and available land is 10,000 ha, the cost of supplying 100,000 dMFT of the energy crops is U.S. $40/dMT. For the range approach, if expected yields are 5 to 15 dMT/ha and corresponding production costs are U.S. $50 and $30/dMT, respectively, the cost of supplying 50,000 to 150,000 dMT is U.S. $30 to $50/dMT.

Production and/or collection costs generally include all variable cash costs, but the extent to which fixed costs and owned resource costs are included varies substantially by study. For energy crops, variable cash costs include items such as seeds (e.g., grass crops); cuttings (e.g., tree crops); fuel, oil, and lubrication; machinery repair; fertilizers and pesticides; hired labor; custom harvest; hired technical services; and interest associated with purchasing inputs used during production but not repaid until the crop is harvested. Fixed costs include general overhead (e.g., bookkeeping, maintaining fences and roads, equipment storage), property taxes, insurance, management, and either rental payments for leasing land or interest for purchasing land. Owned resource costs include a producer's own labor, the value of owned land, machinery depreciation costs, and either the interest rate for loans to purchase machinery or the forgone earnings that could be obtained if capital to purchase machinery was invested in other opportunities. Future changes in technologies (e.g., machinery efficiency, crop productivity) require additional assumptions as to how costs will change over time. No consensus international production costs methodology exists, but some countries and regions have standard recommended methodologies

(e.g., the American Agricultural Economics Association for methodology, the American Society of Agricultural Engineers for methodology and machinery data). Table II presents results from selected studies. Units are those reported by the study authors, with no attempt to adjust to a common measure.

2.3 Supply Curve Approach

In the supply curve approach, biomass resource quantities are estimated as a function of the price that can be paid for the resources. Estimates of several such price/quantity combinations can be used to construct the supply curve. Supply curve estimates can use simple approaches based largely on the quantity and associated cost approach for several quantity/cost combinations. More sophisticated approaches generally use an optimization approach (either profits are maximized or costs to produce a given quantity of resource are minimized). Integrated supply curve analyses also use optimization approaches but generally conduct the analyses within a dynamic framework (i.e., allow variables to change in response to changes in other variables), allocate land in the case of energy crops or biomass resources to different end-use products based on relative economics, and often try to incorporate as many physical, technical, environmental, economic, policy, and social factors as is reasonably feasible. In general, the simpler approaches are used to estimate global supply curves, whereas the more integrated approaches are used for smaller geographic scales (e.g., a national level) due to the complexity of the models and data limitations. Most supply curve analyses estimate supply curves independently for each biomass resource, although attempts to integrate multiple biomass resources that interact with each other simultaneously are being pursued. An increasing number of integrated supply curve analyses are either under way or planned.

For the energy crops example in the previous section, expected yields are 5, 10 (average), and 15 dMT/ha, with respective production costs of U.S. $50, $40, and $30/dMT. A simple supply curve can be estimated by increasing the number of yield/production cost points (e.g., yields of 5.0, 7.5, 10.0, 12.5, and 15.0 dMT/ha and corresponding production costs of U.S. $50, $45, $40, $35, and $30/dMT). At this point, a typical return to the producer (profit) and/or a typical transportation cost (for defined distances) may be added to the production cost estimates but are excluded from this example. The land available for energy crop production is further delineated to include that which can support each production level (of the 10,000 ha available, yields of 5.0 dMT/ha can be achieved on 1000 ha, 7.5 dMT/ha on 2000 ha, 10.0 dMT/ha on 4000 ha, 12.5 dMT/ha on 2000 ha, and 15.0 dMT/ha on 1000 ha). Thus, at U.S. $50, $45, $40, $35, and $30/dMT, quantities of energy crops that can be produced are 5000, 15,000, 40,000, 25,000, and 15,000 dMT, respectively. The

TABLE II

Selected Biomass Quantities and Costs: Quantity and Associated Cost Approach

Reference number	Geographic region	Biomass resources	Quantity	Cost
1	United States	Secondary mill residues	1.22 million dry tons/year	$U.S. 20/dry ton
2	Finland	Forest	3.6 cubic m/year	FIM 12.5/GJ
			8.8 cubic m/year	FIM 15.3/GJ
3	Belarus	SRWC	12 tons/ha	40 Euros/ton[a]
			6 tons/ha	100 Euros/ton[a]
4	Nicaragua	Eucalyptus	12.7 dMT/ha/year	$1.7/GJ[b]

Sources. (1) Rooney, T. (1998). "Lignocellulose Feedstock Resource Assessment," NEOS Corporation, Lakewood, CO.
(2) Malinen, J., Pesonen, M., Maatta, T., and Kajanus, M. (2001). Potential harvest for wood fuels (energy wood) from logging residues and first thinnings in southern Finland. *Biomass and Bioenergy* **20,** 189–196.
(3) Vandenhove, H., Goor, F., O' Brien, S., Grebenkov, A., and Timofeyev, H. (2002). Economic viability of short rotation coppice for energy production for reuse of Caesium-contaminated land in Belarus. *Biomass and Bioenergy* **22,** 421–431.
(4) van den Broek, R., van den Burg, T., van Wijk, A., and Turkenburg, W. (2000). Electricity generation from eucalyptus and bagasse by sugar mill in Nicaragua: A comparison with fuel oil electricity generation on the basis of costs, macroeconomic impacts, and environmental emissions. *Biomass and Bioenergy* **19,** 311–335.

[a] Price paid to farmer (not production cost).
[b] Case study for one plant.

supply curve is constructed by ordering the costs from lowest to highest value (U.S. $30 to $50/dMT) and using cumulative quantities at each price level. Thus, for U.S. $35/dMT, the total quantity is that at U.S. $30/dMT (15,000 dMT) plus that between U.S. $30 and $35/dMT (25,000 dMT) for a total of 40,000 dMT. Figure 1 illustrates the supply curve for this example.

3. EXAMPLE OF AN INTEGRATED SUPPLY CURVE APPROACH FOR BIOMASS RESOURCE ASSESSMENT

A joint U.S. Department of Agriculture and U.S. Department of Energy project conducted by Oak Ridge National Laboratory (ORNL) and the University of Tennessee (UT) is used to illustrate the integrated supply curve approach. Work completed is presented, and possible extensions to the analysis are discussed.

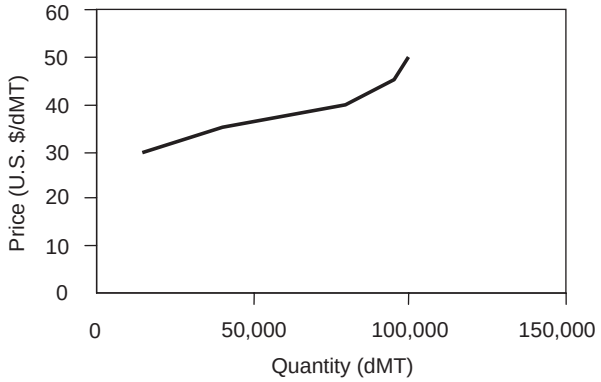


FIGURE 1 Example supply curve for energy crops: simple supply curve approach.

3.1 Model Description

A dynamic agricultural sector model (POLYSYS) is used to estimate the potential for energy crop production and its impacts on U.S. agriculture. POLYSYS is structured as a system of interdependent modules simulating crop supply and demand, livestock supply and demand, and farm income. Figure 2 provides a schematic of the POLYSYS model.

POLYSYS can simulate impacts on the U.S. agricultural sector resulting from changes in market, policy, technical, economic, or resource conditions. It is anchored to published baseline projections and estimates deviations from the baseline for planted and harvested acres, yields, production quantities, exports, variable costs, demand quantities for each use, market price, cash receipts, government payments, and net income.

The crop supply module is composed of 305 independent regions with relatively homogenous production characteristics (agricultural statistical districts [ASDs]). Cropland in each ASD that can be brought into crop production, shifted to production of a different crop, or moved out of production is first identified The extent to which lands can shift among crops is a function of whether the profit for

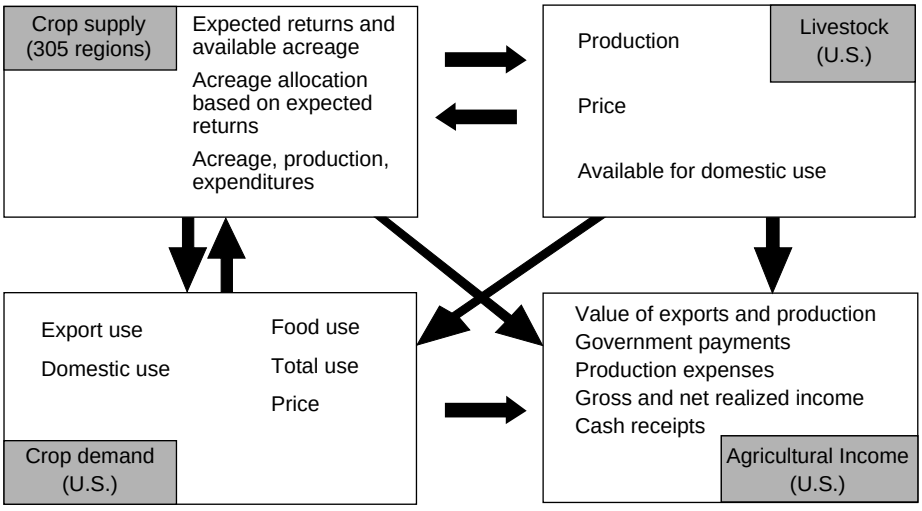


FIGURE 2 Schematic of the POLYSYS model showing the linkages among the submodules. From Ray, D., De La Torre Ugarte, D., Dicks, M., and Tiller, K. (1998). "The POLYSYS Modeling Framework: A Documentation" (unpublished report). Agricultural Policy Analysis Center, University of Tennessee, Knoxville. (http://apacweb.ag.utk.edu/polysys.html.)

each crop is positive, negative, or a mixture for the preceding 3 years, with larger land shifts permitted for crops with negative profits. The amount of idle and pasture lands that can be shifted is limited to 20% of the land in an ASD in any given year. Once the land that can be shifted is determined, the supply module allocates available cropland among competing crops based on maximizing returns above costs (profits).

The crop demand module estimates demand quantities and market prices for each crop and use (e.g., food, feed, industrial). Crop demand is a function of each crop's own price, prices of other crops that compete for the same end-use markets, and nonprice variables (e.g., quality, policy). The livestock module uses statistically derived equations that interact with the demand and supply modules to estimate production quantities and market prices. The primary link between the livestock and crop sectors is through livestock feed demand. Livestock quantities are a function of current and previous years' production and feed prices. The income module uses information from the supply, demand, and livestock modules to estimate cash receipts, production costs, government outlays, net returns to farmers, and net income.

Each module is self-contained but works interdependently to perform a multiperiod simulation. The demand module generates market prices used to form price expectations in the supply module that are used to allocate land to crops by ASD and to form national supply quantities. National supply quantities are then returned to the demand module to estimate prices, demand quantities, and surplus quantities that are carried over to the next year (i.e., carryover stocks). The livestock sector is linked to the crop sector through feed grains where livestock quantities affect feed grain demand and prices and feed grain prices and supplies affect livestock production decisions. Livestock and crop production quantities and prices are used by the income module.

POLYSYS includes the major U.S. crops, that is, corn, soybeans, wheat, grain sorghum, oats, barley, cotton, rice, alfalfa, and other hay. The livestock sector includes beef, pork, lamb and mutton, broiler chickens, turkeys, eggs, and milk. Food, feed, industrial, and export demands are considered as well as carryover stocks. The model includes all U.S. lands classified as cropland (174.7 million ha, 431.4 million acres) currently in crop production, idled, in pasture, or in the Conservation Reserve Program (CRP).

Crop production costs are developed using detailed field operation schedules and include variable input costs (e.g., seed, fertilizer, pesticides, custom operations, fuel and lube, repairs, paid labor, drying, irrigation, technical services, interest), fixed costs (e.g., insurance and taxes, storage, general overhead), capital returns on machinery, and the value of the producer's own labor and management. The analysis uses recommended methodology.

3.2 Modification of Model to Include Energy Crops

POLYSYS was modified to include three energy crops: switchgrass (*Panicum virgatum*), hybrid poplar (*Populus spp.*), and willow (*Salix spp.*). Switchgrass is a native perennial grass and a major species of the U.S. tall grass prairies. Hybrid poplar is a cross between two or more *Populus spp.* that include aspens and cottonwoods. Willows used for energy are improved varieties of native shrubs (not landscaping ornamentals). Other trees and grasses can be used and may be preferred under some conditions; however, many of these alternatives have similar expected yields and production practices, so switchgrass, poplar, and willow can serve as reasonable representations of other energy crops. U.S. cropland identified as suitable for energy crop production (given current development status) and used in the analysis is 149 million ha (368 million acres).

The production location, yields, and management practices of each energy crop differ across each ASD in POLYSYS. Switchgrass is planted for 10 years and is harvested annually as large round bales. Varieties appropriate to each region are assumed, and fertilizer (quantity and type) varies by region. Hybrid poplar is planted for 6- to 10-year rotations (depending on region) and is harvested once at the rotation end as whole tree chips. Fertilizer applications vary by region. Willows are planted in a 22-year rotation, harvested every third year with regrowth from the stump following harvest (coppicing), and delivered as whole tree chips. Fertilizer is applied during the year following harvest. Herbicide applications for all crops are limited to the first 2 years of production. All assumed management practices are based on research results, demonstration of commercial field experience where available, and expert opinion.

Achievable energy crop yields on land currently producing traditional agricultural crops are used as the base yields in the analysis. Energy crop yields on idled and pasture lands are assumed to be 85% of the base yields. Production practices used on idle and pasture lands are similar to those on cropped lands except for additional herbicide applications to kill

existing cover and increased tillage to prepare the sites for planting.

Energy and traditional crop production costs are estimated using the POLYSYS budget generator model except for hybrid poplar and willow harvesting costs. These costs are estimated using the ORNL BIOCOST model. Because of the multiyear production rotations of energy crops, annual profits for energy and traditional crops are discounted using a 6.5% real discount rate (a net present value approach). A common planning horizon of 40 years is used to account for the different rotation lengths of the energy crops.

Cropland idled or in pasture is assumed to be so due to economic reasons; at given prices, production costs, and yields, converting the land to pasture or idling it is the most economic use. To return these lands to production, the present value profits of crops must be higher than the most profitable crop under the baseline assumptions. To account for possible inertia to keep the land in its current use and/or the value of pasture land in livestock operations, a premium of 10% above the baseline present value returns for idled lands and 15% for pasture lands is required.

Energy crops compete for land with both traditional crops and each other; at an equivalent energy price (same cost/gigajoule [U.S. $/GJ]) for each energy crop, land is allocated to the most profitable energy crop given expected yields and production costs. Switchgrass is the most profitable energy crop, followed by hybrid poplar and willow in most ASDs; thus, nearly all (99%) of the currently planted, idle, and pasture land shifted to energy crop production goes to switchgrass production.

The analysis estimates that at a price of U.S. $2.44/GJ (U.S. $44/dMT [U.S. $40/dry ton]), 11.7 million ha (29 million acres) producing 120.4 million dMT (132.5 million dry tons) of energy crops can be more profitable than existing agricultural uses for cropped, idled, and pasture lands. The biomass produced is equivalent to 2.016 exajoules (2.05 quads) of primary energy and, given existing conversion rates, can produce 44.7 billion L (11.8 billion gallons) of ethanol or 216 billion kilowatt-hours of electricity equivalent to approximately 5% of current U.S. production levels. Traditional crop prices increase by 9 to 14% depending on crop. Net farm income increases by U.S. $6.0 billion/year (U.S. $2.3 billion from energy crops and the rest from higher traditional crop prices). Figure 3 summarizes estimated quantities of energy crops that are more profitable than alternative land uses for several price levels. Figure 4 shows the regional distribution of energy crop production at U.S. $2.44/GJ.

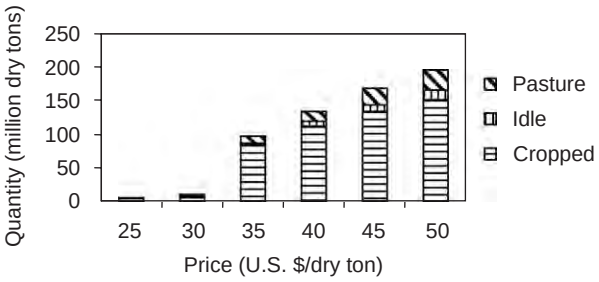


FIGURE 3 Energy crop production on U.S. cropped, idle, and pasture acres at selected crop prices.

3.3 Analysis of the Conservation Reserve Program

Modifications to the CRP have been suggested as a means to introduce energy crops to agriculture and reduce their prices to end users. The CRP, established in the 1985 Farm Bill, sets asides environmentally sensitive lands under 10- to 15-year contracts. CRP lands are planted to conservation crops, such as perennial grasses and trees, and farmers receive an annual rental payment. Harvest is prohibited except under emergency conditions. Policy changes in this agricultural program were analyzed with the modified POLYSYS model.

Enrolled lands considered in the analysis (12.1 million ha [29.7 million acres]) were adjusted for geographic suitability and environmental sensitivity. Lands that are enrolled as buffer strips, classified as wetlands, or critical to watershed management, or that provide critical habitat in wildlife conservation priority areas were excluded. CRP land available for energy crop production in the analysis amounts to 6.8 million ha (16.9 million acres).

Two energy crop management practices were analyzed: one to achieve high biomass production (production scenario) and one to achieve high wildlife diversity (wildlife scenario). The wildlife scenario uses fewer fertilizer and chemical inputs than does the production scenario, and the former restricts switchgrass harvest to alternating years. Cover crops to minimize erosion are required during the establishment years for all energy crops. Yields of energy crops produced on CRP lands are adjusted by an index of traditional crop yields obtained on CRP lands prior to being enrolled in the program.

When existing CRP contracts expire, farmers can renew their current contracts or produce energy crops under modified contracts (the analysis is simplified by not enrolling new lands in the program or by switching existing CRP lands back to traditional crop production). In exchange for being

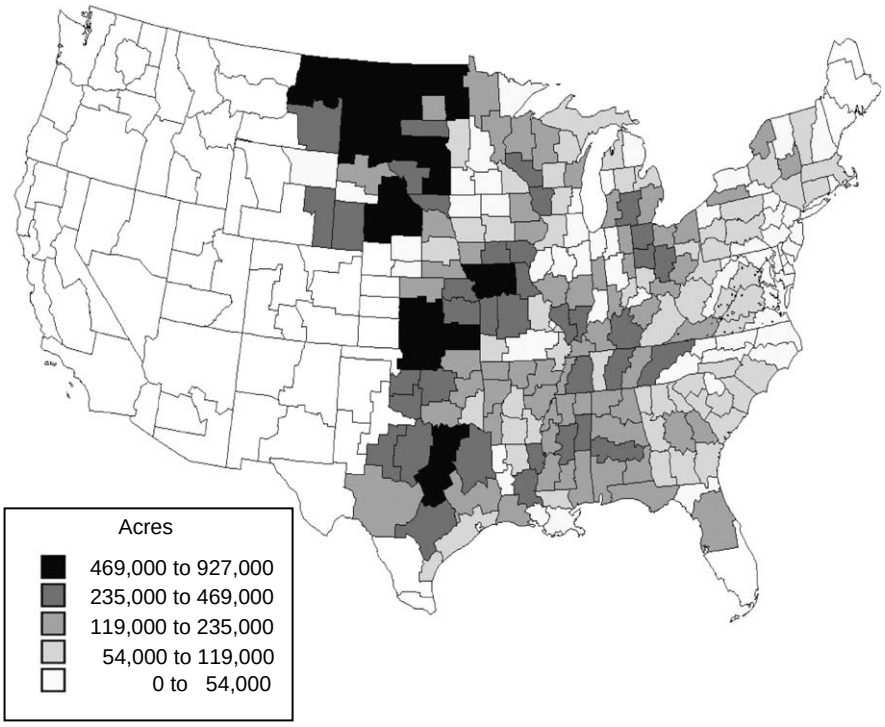


FIGURE 4 Distribution of energy crops on cropland currently in traditional crop production, idled, or in pasture at a price of U.S. $2.44/GJ (U.S. $40/dry ton). From De La Torre Ugarte, D., Walsh, M., Shapouri, H., and Slinsky, S. (2003). "The Economic Impacts of Bioenergy Crop Production on U.S. Agriculture," Agricultural Economic Report No. 816. U.S. Department of Agriculture, Washington, DC.

allowed to produce and harvest energy crops on CRP lands, 25% of the current rental rate is forfeited. Thus, the decision to allocate CRP lands to energy crop production is reduced to a comparison of their present value profits and the present value of the forgone CRP rental payments.

Under the wildlife scenario at U.S. $2.44/GJ, CRP lands are split between switchgrass (annual production of 1.0 million dMT) and hybrid poplar (35.1 million dMT when harvested). Under the production scenario, switchgrass dominates, with 5.2 million ha (12.91 million acres) producing 50.3 million dMT annually. The U.S. Congress used the analysis to establish a pilot program (based on the wildlife scenario) to allow production and harvest of energy crops on CRP lands (Public Law 106-78, section 769). Figure 5 shows the distribution of CRP lands under the production scenario at U.S. $2.44/GJ.

3.4 Potential Extensions of the Integrated Supply Curve Analysis

The POLYSYS model provides a solid decision framework to anchor additional analyses. For the analysis described previously, a deterministic enterprise version of POLYSYS was used. A stochastic crop rotation version also exists, and future plans are to incorporate energy crops into this version. Because POLYSYS is an agricultural sector model, additional agricultural-based biomass resources (e.g., corn stover, wheat straw) can be added. The removal of agricultural residues raises many soil quality issues such as erosion, carbon levels, moisture, and long-run productivity. POLYSYS was originally designed to be able to evaluate environmental constraints and impacts and includes several major soil types, tillage, and management options by crop rotation and ASD. Work is currently under way to estimate the amounts of agricultural residues that must be left on the field to maintain soil erosion and carbon at determined levels. The work evaluates needed residue levels considering grain yields, crop rotation, tillage and other management practices, soil type, topology, and climate. Response curves developed from this analysis will be incorporated into POLYSYS, allowing biomass supply curves for agricultural residues, energy crops, and traditional crops to be estimated simultaneously.

The supply curves from POLYSYS, as well as the underlying data used to produce them, can be used

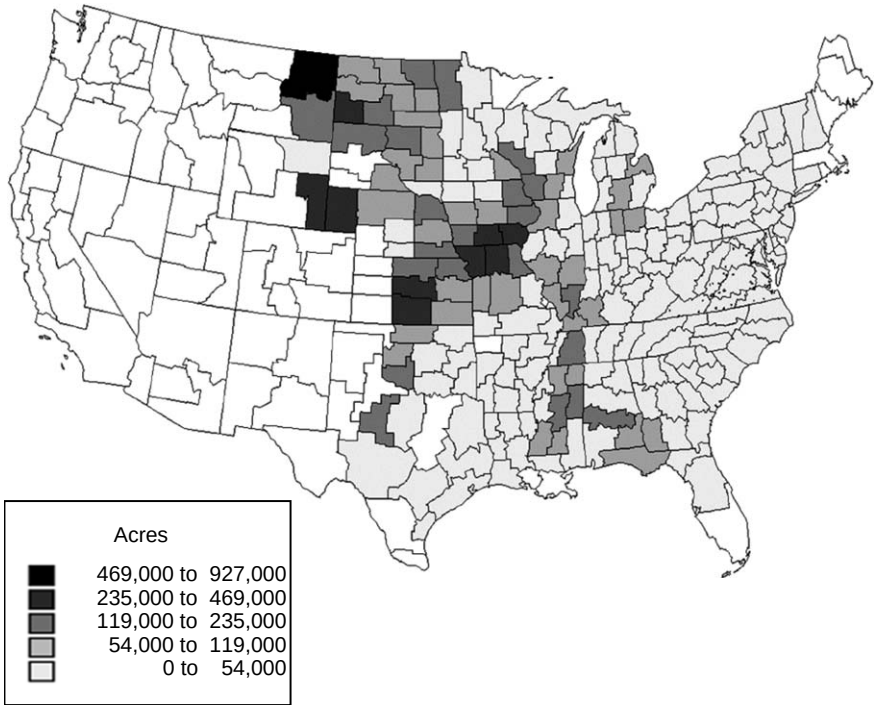


FIGURE 5 Distribution of energy crops on Conservation Reserve Program (CRP) acres at a price of U.S. $2.44/GJ (U.S. $40/dry ton) and assuming the production CRP scenario. From De La Torre Ugarte, D., Walsh, M., Shapouri, H., and Slinsky, S. (2003). "The Economic Impacts of Bioenergy Crop Production on U.S. Agriculture," Agricultural Economic Report No. 816. U.S. Department of Agriculture, Washington, DC.

with other models to conduct additional analyses, including developing industry-level bioenergy and bioproduct expansion curves, life cycle analyses, and economic impact assessments. Product expansion curves can be estimated using a transportation and siting model to identify location and delivered biomass cost given user facility demand levels. Addition of conversion costs and market prices for the product(s)—multiple products can be examined simultaneously—permits estimation of the size, location, and cost of the industry as a whole (industry expansion curves). Life cycle analyses can be used to estimate the potential to displace fossil fuels, energy balances, impact on greenhouse gas emissions, and impact on other air pollutants relative to existing products (e.g., ethanol vs gasoline as a transportation fuel). Economic impacts (e.g., jobs, value added, total product output, farm income) of a national bioenergy or bioproduct industry can be estimated using appropriate economic models. Once the entire system is established, the impacts of changes in technology, economic and environmental policy, and regulatory conditions can be evaluated. Figure 6 illustrates how the component pieces relate to each other.

4. DISCUSSION OF LIMITATIONS AND CONCEPTUAL DIFFERENCES IN BIOMASS RESOURCE ASSESSMENTS

International consensus regarding the definition of biomass resource assessment, the terminology and methodologies used to conduct assessments, and the units used to present results does not currently exist. Disciplinary training, historical conditions, and cultural differences are among the factors that complicate reaching a consensus.

This article defines biomass resource assessment as estimating the quantity of biomass that can be realistically supplied by price and location under existing and potential physical, technical, environmental, economic, policy, and social conditions. This definition includes the inventory, average quantity and cost, and simple to integrated supply curve approaches. However, not all biomass analysts concur with this view, and a number would limit their definition to inventory analysis only.

Terminology to define biomass resource assessment approaches is not standard. For example, the

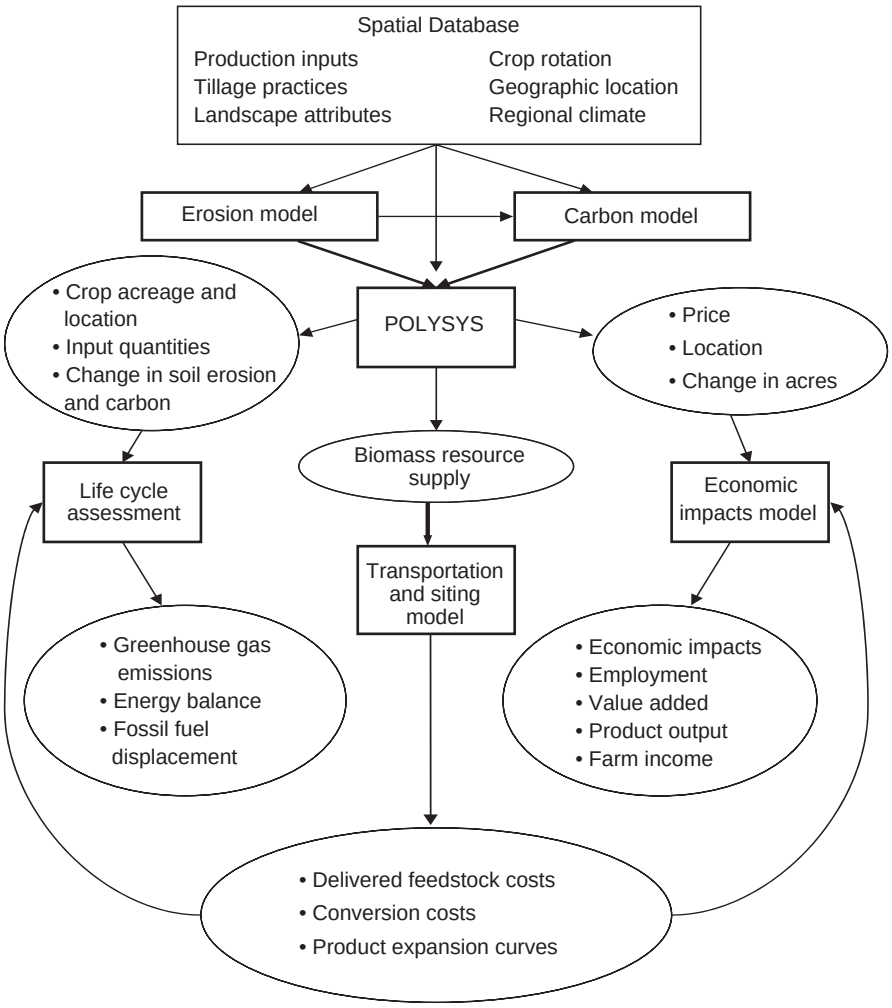


FIGURE 6 Schematic of an integrated biomass system. The figure illustrates the relationship among potential extensions of the integrated supply curve analysis (POLYSYS modeling).

International Institute for Applied Systems Analysis defines assessments in terms of theoretical, technical, and economic approaches. This article combines the theoretical and technical approaches into the inventory approach and divides the economic approach into the quantity and associated cost and supply curve approaches. Recommended methodologies are available for selected components of biomass resource assessments in some regions, but an international consensus of approaches flexible enough to be applied to different situations for all components does not exist.

As illustrated in Tables I and II, units to express results vary substantially by study. The various units used are appropriate to the purpose of each study but make it difficult to compare studies. With respect to biomass resource assessments, dry metric tons are applicable to most solid biomass resources. Inclusion of this measure would facilitate comparisons across studies. It is important to note that trying to develop one standard for all situations might not be an appropriate goal. Rather, a set of standards applicable to different situations may be better. Alternative methods might be needed to accommodate data limitations.

Inclusion/exclusion of biomass resources currently used for low-efficiency energy or nonenergy purposes is a major conceptual difference among biomass resource assessments. Most studies exclude biomass resources currently used that can reduce available quantities to as little as 10% of those produced in some studies. The rationale for exclusion is that these alternative uses are more economically attractive relative to energy use. However, the relative attractiveness is based on current technologies, assumed energy product, and existing economic, market,

policy, and regulatory conditions. Under different conditions, the relative attractiveness can differ radically. Some studies assume that currently used resources are still available for energy use, but at higher prices. This approach allows greater flexibility to evaluate biomass potential under alternative situations. Studies using this approach show that under plausible technology, product, market, economic, policy, and regulatory conditions, prices paid for biomass can be sufficient to divert these resources from existing uses to bioenergy and bioproduct uses.

A second major conceptual difference is the value of idled land used to produce energy crops. Most studies assume zero to low value for idled lands. This assumption is reasonable for severely degraded lands that are physically unable to support most forest or food crop production such as previously forested areas that were clear cut, planted to food crops until the soils were depleted, and then abandoned. Alternative uses for these lands are severely limited. However, in many industrialized nations, lands are idle due to low profits or policy and can be returned to production under different conditions. Under these circumstances, the assumption of zero to low value is questionable; valuing idle lands based on their best alternative uses is more appropriate.

A third conceptual difference involves whether to include environmental factors as constraints to biomass quantities that can be collected or as impacts that occur after collection. This difference is particularly apparent in the way in which available agricultural residues are estimated, particularly with respect to soil erosion and carbon. Most studies either ignore these issues altogether or view them as impacts resulting from collection. Some studies acknowledge the erosion concerns and try to account for them, albeit usually in simple ways (e.g., 50% of residues must remain). Soil carbon implications are rarely considered. In the United States, studies are under way to estimate the quantities of agricultural residues that can be removed and still maintain soil erosion and carbon at defined levels under different crop rotation, soil type, tillage and management practices, crop yields, and geo-climatic conditions.

In addition to conceptual differences and no standard definitions and methodologies, biomass resource assessments are significantly affected by data availability. In many cases, data are unavailable to support even simple inventory approaches, much less the extremely data-intensive integrated supply curve approaches. Also, the complexity and difficulty of creating integrated dynamic and stochastic models can be severely limiting.

For those interested in finding out more about biomass resource assessments, bioenergy, and bioproducts, good starting points include biomass-related journals, Web sites at institutions conducting biomass research, and biomass conference proceedings.

5. SUMMARY AND CONCLUSIONS

Biomass resource assessments estimate the quantities of resources available by location and price levels given existing and potential physical, technical, environmental, economic, policy, and social conditions. Three basic approaches are used to estimate biomass resources: the inventory approach, the quantity and associated cost approach, and the supply curve approach. The inventory approach estimates physical quantities of biomass resources only, with no consideration of price. The quantity and associated cost approach typically estimates a minimum–maximum range or average quantities along with corresponding production and/or collection costs. The supply curve approach estimates biomass quantities as a function of the price that can be paid. Each is inclusive of the previous approach and results in fewer available resources and higher prices.

International consensus regarding the definition of biomass resource assessment, biomass terminology, methodology, and units to express results is lacking. Major differences in conceptual issues lead to substantially different ways in which to conduct biomass resource assessments and to substantially different results of the analyses.

Data availability is a major constraint limiting the types of biomass resource assessments that can be done. Integrated supply curve approaches incorporating as many physical, technical, environmental, economic, policy, and social factors as are reasonably feasible are extremely data intensive and involve creating complex models. However, they provide maximum flexibility to examine changes in any given variable, and if they are conducted in a dynamic framework, they allow variables within a model to change in response to each other. An example of an integrated supply curve analysis in the United States, with proposed extensions, was presented.

SEE ALSO THE FOLLOWING ARTICLES

Biomass, Chemicals from • *Biomass Combustion* • *Biomass for Renewable Energy and Fuels* • *Biomass*

Gasification • Biomass: Impact on Carbon Cycle and Greenhouse Gas Emissions • Forest Products and Energy • Renewable Energy, Taxonomic Overview • Waste-to-Energy Technology

Further Reading

Biomass and Bioenergy. (n.d.). Elsevier Science, New York.

Hall, D. O., Rosillo-Calle, F., Williams, R. H., and Woods, J. (1993). Biomass for energy: Supply prospects. *In* "Energy Sources for Fuels and Electricity" (T. Johansson, H. Kelly, A. Reddy, and R. Williams, Eds.), pp. 593–651. Island Press, Washington, DC.

Overend, R., and Chornet, E. (eds.). (1999). "Proceedings of the Fourth Biomass Conference of the Americas: Biomass—A Growth Opportunity in Green Energy and Value-Added Products." Elsevier Science, Oakland, CA.

"Proceedings of Bioenergy 2000: Moving Technology into the Marketplace." (2000). Northeast Regional Biomass Program, Buffalo, NY.

"Proceedings of Biomass Energy: Data, Analysis, and Trends Conference." (1998). International Energy Agency, Paris.

"Proceedings of the First World Conference on Biomass for Energy and Industry." (2001). James & James, Sevilla, Spain.

Walsh, M., De La Torre Ugarte, D., Shapouri, H., and Slinsky, S. (2003). Bioenergy crop production in the United States: potential quantities, land use changes, and economic impacts on the agricultural sector. *Environ. Resource Econ.* **24,** 313–333.

Bottom-Up Energy Modeling

JAYANT SATHAYE and ALAN H. SANSTAD
Lawrence Berkeley National Laboratory
Berkeley, California, United States

1. **Goals, Context, and Use: Bottom-Up Approaches**
2. **Structure of an Energy Sector Bottom-Up Assessment**
3. **Typical Models for a Bottom-Up Energy Sector Assessment**
4. **Key Challenges in the Bottom-Up Modeling Approach**
5. **Data Sources for Bottom-Up Analysis**
6. **Developing Scenarios for Use in a Bottom-Up Assessment**
7. **Results from a Bottom-Up Approach**

Glossary

base year The year for which the inventory is to be taken. In some cases (such as estimating CH_4 from rich production), the base year is simply the last year of a number of years over which an average must be taken.

bottom-up modeling A modeling approach that arrives at economic conclusions from an analysis of the effect of changes in specific parameters on narrow parts of the total system.

carbon tax A tax on fossil fuels based on the individual carbon content of each fuel. Under a carbon tax, coal would be taxed the highest per MBtu, followed by petroleum and then natural gas.

demand-side management The planning, implementation, and monitoring of utility activities designed to encourage customers to modify their pattern of electricity usage.

discount rate The rate at which money grows in value (relative to inflation) if it is invested.

dynamic In the field of modeling, a dynamic model includes intertemporal relations between variables. A model that does not include such relations is called static.

energy forms and levels Primary energy is energy that has not been subjected to any conversion or transformation process. Secondary energy (derived energy) has been produced by the conversion or transformation of primary energy or of another secondary form of energy. Final energy (energy supplied) is the energy made available to the consumer before its final conversion (i.e., before utilization). Useful energy is the energy made usefully available to the consumer after its final conversion (i.e., in its final utilization).

energy intensity The amount of energy required per unit of a particular product or activity.

energy services The service or end use ultimately provided by energy. For example, in a home with an electric heat pump, the service provided by electricity is not to drive the heat pump's electric motor but rather to provide comfortable conditions inside the house.

engineering approach A particular form of bottom-up modeling in which engineering-type process descriptions (e.g., fuel efficiency of end-use devices) are used to calculate a more aggregated energy demand. This term is particularly used in contrast to econometric models.

exogenous variables Variables determined outside the system under consideration. In the case of energy planning models, these may be political, social, or environmental, for example.

feedback When one variable in a system (e.g., increasing temperature) triggers changes in a second variable (e.g., cloud cover), which in turn ultimately affects the original variable (i.e., augmenting or diminishing the warming). A positive feedback intensifies the effect. A negative feedback reduces the effect.

fossil fuel Coal, petroleum, or natural gas or any fuel derived from them.

general equilibrium analysis An approach that considers simultaneously all the markets in an economy, allowing for feedback effects between individual markets. It is particularly concerned with the conditions that permit simultaneously equilibrium in all markets and with the determinants and properties of such an economy-wide set of equilibrium.

income elasticity The expected percentage change in the quantity demand for a good given a 1% change in income. An income elasticity of demand for electricity of 1.0 implies that a 1% increase in income will result in a 1% increase in demand for electricity.

input-output analysis A method of investigating the interrelationship between the branches of a national economy in a specific time period. The representation, in the form of a matrix table, is called an input-output table. An input-output analysis allows the changes in total demand in related industrial branches to be estimated.

least-cost planning In energy planning, the practice of basing investment decisions on the least costly option

 © 2004 Elsevier Inc. All rights reserved.

for providing energy services. It is distinguished from the more traditional approach taken by utilities, which focuses on the least costly way to provide specific types of energy, with little or no consideration of less costly alternatives that provide the same energy service at lower costs.

linear programming A practical technique for finding the arrangement of activities that maximizes or minimizes a defined criterion subject to the operative constraints. For example, it can be used to find the most profitable set of outputs that can be produced from a given type of crude oil input to a given refinery with given output prices. The technique can deal only with situations where activities can be expressed in the form of linear equalities or inequalities and where the criterion is also linear.

macroeconomics The study of economic aggregates and the relationships between them. The targets of macroeconomic policy are the level and rate of change of national income (i.e., economic growth), the level of unemployment, and the rate of inflation. In macroeconomics, the questions about energy are how its price and availability affect economic growth, unemployment, and inflation, and how economic growth affects the demand for energy.

marginal costs In linear programming, this term has the very specific meaning of change of the objective function value as a result of a change in the right-hand-side value of a constraint. If, for example, the objective is to minimize costs, and if the capacity of a particular energy conversion facility, such as a power plant, is fully utilized, the marginal cost in the linear planning sense expresses the (hypothetical) reduction of the objective function value (i.e., the benefit) of an additional unit of capacity.

market clearing The economic condition of supply equaling demand.

optimization model A model describing a system or problem in such a way that the application of rigorous analytical procedures to the representation results in the best solution for a given variable(s) within the constraints of all relevant limitations.

price elasticities The expected percentage change in quantity demand for a good given a 1% change in price. A price elasticity of demand for electricity of -0.5 implies that a 1% increase in price will result in a half percent decrease in demand for electricity.

renewable energy Energy obtained from sources that are essentially inexhaustible (unlike, for example, the fossil fuels, of which there is a finite supply). Renewable sources of energy include wood, waste, wind, geothermal, and solar thermal energy.

retrofit To update an existing structure or technology by modifying it, as opposed to creating something entirely new from scratch. For example, an old house can be retrofitted with advanced windows to slow the flow of energy into or from the house.

scenario Coherent and plausible combination of hypotheses, systematically combined, concerning the exogenous variables of a forecast.

sensitivity analysis A method of analysis that introduces variations into a model's explanatory variables to examine their effects on the explained.

simulation model Descriptive model based on a logical representation of a system and aimed at reproducing a simplified operation of this system. A simulation model is referred to as static if it represents the operation of the system in a single time period; it is referred to as dynamic if the output of the current period is affected by evolution or expansion compared with previous periods. The importance of these models derives from the impossibility of excessive cost of conducing experiments on the system itself.

top-down modeling A modeling approach that proceeds from broad, highly aggregated generalizations to regionally or functionally disaggregated details.

Two general approaches have been used for the integrated assessment of energy demand and supply: the so-called bottom-up and top-down approaches. The bottom-up approach focuses on individual technologies for delivering energy services, such as household durable goods and industrial process technologies. For such technologies, the approach attempts to estimate the costs and benefits associated with investments in increased energy efficiency, often in the context of reductions in greenhouse gas (GHG) emission or other environmental impacts. The top-down method assumes a general equilibrium or macroeconomic perspective, wherein costs are defined in terms of losses in economic output, income, or gross domestic product (GDP), typically from the imposition of energy or emissions taxes.

1. GOALS, CONTEXT, AND USE: BOTTOM-UP APPROACHES

The fundamental difference between the two approaches is in the perspective taken by each on consumer and firm behavior and the performance of markets for energy efficiency. The bottom-up approach assumes that various market "barriers" prevent consumers from taking actions that would be in their private self-interest—that is, would result in the provision of energy services at lower cost. These market barriers include lack of information about energy efficiency opportunities, lack of access to capital to finance energy efficiency investment, and misplaced incentives that separate responsibilities

for making capital investments and paying operating costs. In contrast, the top-down approach generally assumes that consumers and firms correctly perceive, and act in, their private self-interest (are utility and profit maximizers) and that unregulated markets serve to deliver optimal investments in energy efficiency as a function of prevailing prices. In this view, any market inefficiencies pertaining to energy efficiency result solely from the presence of environmental externalities that are not reflected in market prices.

In general, an assessment carried out using the bottom-up approach will very likely show significantly lower costs for meeting a given objective (e.g., a limit on carbon emissions) than will one using a top-down approach. To some extent, the differences may lie in a failure of bottom-up studies to accurately account for all costs associated with implementing specific actions. Top-down methods, on the other hand, can fail to account realistically for consumer and producer behavior by relying too heavily on aggregate data, as noted by Krause *et al.* In addition, some top-down methods sacrifice sectoral and technology detail in return for being able to solve for general equilibrium resource allocations. Finally, Boero *et al.* noted that top-down methods often ignore the fact that economies depart significantly from the stylized equilibria represented by the methods. Each approach, however, captures costs or details on technologies, consumer behavior, or impacts that the other does not. Consequently, a comprehensive assessment should combine elements of each approach to ensure that all relevant costs and impacts are accounted for.

The two approaches have been used in the development of national energy plans or policies that require identification and analysis of different actions that governments could take to encourage adoption of energy technologies and practices. Based on such analyses, policymakers can decide which options not only satisfy specific policy objectives but are also within institutional, political, and budget constraints. Typically, the analytic process will follow a series of steps, each of which produces information for decision makers. The manner in which these steps are performed will reflect each country's resources, objectives, and decision-making process.

Both approaches have been used extensively in the assessment of costs of climate change mitigation. The earlier literature dating to 1970s focused primarily on evaluation of the energy and particularly the petroleum sector, and categorized approaches into sectoral models, industry-market models, energy system models, and energy economy models. The first three approaches would be referred to as bottom-up approaches and the last one as the top-down approach today. From the late 1980s to date, much of the attention in the application of these two energy-sector approaches has focused on climate change mitigation, that is, the long-term stabilization of climate change and the short-term reduction of GHG emissions. The following sections focus on the evolution and use of energy-sector bottom-up approaches as they have been applied for mitigation assessment, particularly the short-term reduction of GHG emissions.

2. STRUCTURE OF AN ENERGY SECTOR BOTTOM-UP ASSESSMENT

The energy sector comprises the major energy demand sectors (industry, residential and commercial, transport, and agriculture) and the energy supply sector (resource extraction, conversion, and delivery of energy products). GHG emissions occur at various points in the sector, from resource extraction to end use, and, accordingly, options for mitigation exist at various points.

The bottom-up approach involves the development of scenarios based on energy end uses and evaluation of specific technologies that can satisfy demands for energy services. One can compare technologies based on their relative cost to achieve a unit of GHG reduction and other features of interest. This approach gives equal weight to both energy supply and energy demand options. A variety of screening criteria, including indicators of cost-effectiveness as well as noneconomic concerns, can be used to identify and assess promising options, which can then be combined to create one or more mitigation scenarios. Mitigation scenarios are evaluated against the backdrop of a baseline scenario, which simulates the events assumed to take place in the absence of mitigation efforts. Mitigation scenarios can be designed to meet specific emission reduction targets or to simulate the effect of specific policy interventions. The results of a bottom-up assessment can then be linked to a top-down analysis of the impacts of energy sector scenarios on the macroeconomy.

Energy-sector bottom-up assessments require physical and economic data about the energy system, socioeconomic variables, and specific technology options, and GHG emissions if these are targeted. Using these data, a model or accounting system of the energy sector is designed to suit local circumstances.

The manner in which an assessment is performed reflects each country's resources, objectives, and decision-making process as well as the type of modeling approach employed. Figure 1 depicts the basic steps of a typical mitigation assessment and how they relate to one another. Some of the steps are interlinked, so they are not necessarily sequential, and require iterations.

An initial step is to assemble data for the base year on energy demand by sector, energy supply by type of energy source, and energy imports and exports. The disaggregated energy data is normalized to match the national energy supply totals for the base year. One then calibrates base year emissions with the existing GHG inventory, as needed. The analyst also assembles data for the base year on the technologies used in end-use sectors and in energy supply.

The data for the base year is used as a starting point for making projections of future parameters and developing integrated scenarios of energy demand and supply. On both the demand and supply side, one identifies and screens potential technology options to select those that will be included in the analysis. The screening is guided by information from an assessment of energy resources, as well as the potential for energy imports and exports.

Once the list of technologies has been made manageable by the screening process, the analyst characterizes potential technology options in end-use sectors and in energy supply with respect to costs, performance, and other features. On the demand side, this characterization will assist in projecting end-use energy demands in the various sectors. Projecting energy demand also requires one to project activity levels in each subsector for indicators such as tons of various industrial products, demand for travel and freight transport, and number of urban and rural households. These projections are based on the assumptions for growth in key parameters such as GDP and population. Assumptions about sectoral policies with respect to energy pricing and other factors are also important in developing projections of energy demand.

The data from the energy demand and supply analyses are then entered into an energy sector model or accounting framework that allows for integrated analysis of the various options that can meet energy requirements. This analysis calculates costs and impacts over the time horizon considered, the results are reviewed for reasonableness, and uncertainty is taken into consideration. This step involves combining technology options to meet the objectives of each scenario. The selection of technologies may be made directly by the analyst or performed by the model (as with an optimization model).

The baseline scenario projects energy use and emissions over the time horizon selected, reflecting the development of the national economy and energy system under the assumption that no policies are introduced to reduce GHG emissions. The baseline scenario must include sufficient detail on future energy use patterns, energy production systems, and technology choices to enable the evaluation of specific policy options. An alternative baseline

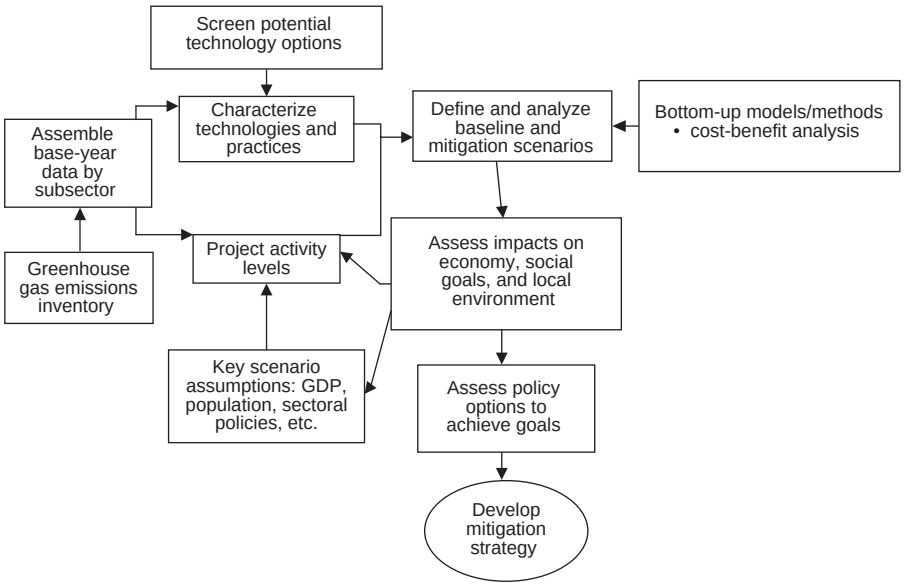


FIGURE 1 Structure of a bottom-up assessment.

scenario can be developed if desired (for example, to reflect different assumptions about GDP growth).

Policy scenarios can be defined to meet particular emission reduction targets, to assess the potential impact of particular policies or technologies, or to meet other objectives. The comparisons of policy and baseline scenarios reveal the net costs and impacts of the policy options. The results are assessed with respect to reasonableness and achievability, given barriers to implementation and the policy instruments that might be used, such as taxes, standards, or incentive programs.

For both baseline and policy scenarios, the analyst assesses the impacts on the macroeconomy, social goals (such as employment), and the national environment. One approach is to integrate bottom-up assessment with a macroeconomic model. Decision analysis methods that allow for consideration of multiple criteria may also be appropriate.

After scenarios have been analyzed and options have been ranked in terms of their attractiveness (on both quantitative and qualitative terms), it is desirable to conduct a more detailed evaluation of policies that can encourage adoption of selected options. Such an evaluation can play an important role in the development of a national strategy. The latter step requires close communication between analysts, policymakers, and other interested parties.

3. TYPICAL MODELS FOR A BOTTOM-UP ENERGY SECTOR ASSESSMENT

Bottom-up assessments of the energy sector typically use an accounting or modeling framework to capture the interactions among technologies and to ensure consistency in the assessment of energy, emission, and cost impacts. Accounting and modeling methods can vary greatly in terms of their sophistication, data intensiveness, and complexity. This section provides an overview of key concepts and capabilities of some models that have been used for this purpose in energy/environmental studies in developing and industrialized countries.

As discussed earlier, it is common to divide energy models into two types, so-called bottom-up and top-down, depending on their representation of technology, markets, and decision making. In practice, there is a continuum of models, each combining technological and economic elements in different ways. At one extreme are pure bottom-up energy models, which focus on fuels and energy conversion or end-use technologies and treat the rest of the economy in an aggregated fashion. At the other extreme are pure top-down models, which treat energy markets and technologies in an aggregated manner and focus instead on economy-wide supply-demand relations and optimizing behavior. Between these two cases are a number of models that combine elements of both extremes with various degrees of emphasis and detail.

The description of the future varies among the models. Some models can only analyze a snapshot year and compare this to another year, without any representation of the transition between them. Dynamic models, on the other hand, allow for time-dependent descriptions of the different elements of the energy system. While the snapshot models enable great detail in the representation of the system, dynamic models allow for representation of technology capacity transfer between time periods and thus time-dependent capacity expansion, time-dependent depletion of resources, and abatement costs as they vary over time.

In dynamic modeling, information about future costs and prices of energy is available through two diametrically different foresight assumptions. With myopic foresight, the decisions in the model are made on a year-by-year basis, reflecting the assumption that actors expect current prices to prevail indefinitely. Assuming perfect foresight, the decisions at any year are based on the data for the entire time horizon. The model thus reflects the activities of market participants as if they use the model itself to predict prices.

Table I summarizes the key design features of some of the models. Most of the models listed in Table I can be used to integrate data on energy demand and supply. The models can use this information for determining an optimal or equilibrium mix of energy supply and demand options. The various models use cost information to different degrees and provide for different levels of integration between the energy sector and the overall economy.

3.1 Energy Accounting Models

Energy accounting models reflect an engineering or input-output conception of the relations among energy, technology, and the services they combine to produce. This view is based on the concept of energy services that are demanded by end users. Schematically, this can be represented as follows:

$$\text{Energy inputs} > \text{Technology} > \text{Energy services.}$$

TABLE I

Key Design Features of Types of Bottom-Up Energy Sector Models

Model characteristics	Energy accounting	Engineering optimization	Iterative equilibrium	Hybrid
Energy supply representation	Process analysis	Process analysis	Supply curve	Process analysis
Energy demand representation	Exogenous	Exogenous	Exogenous	Utility maximization
Multiperiod	Yes	Yes	Yes	Yes
Consumer/producer foresight	Not applicable	Perfect/myopic	Myopic	Perfect/myopic
Solution algorithm	Accounting	Linear or nonlinear optimization	Iteration	Nonlinear optimization

For policy purposes, the significance of this approach is that a given type and level of energy service can be obtained through various combinations of energy inputs and technologies. In particular, holding the service constant while increasing the energy efficiency of the technology allows decrease in the required level of energy input. In a range of cases, when other factors are held equal, this lowers the overall cost of the energy service. With accounting models, the evaluation and comparison of policies is performed by the analyst external to the model itself.

These models are essentially elaborations on the following accounting identity describing the energy required for satisfying a given level of a specific energy service:

$$E = AI,$$

where E indicates energy, A indicates activity, and I indicates intensity. With multiple end uses, aggregate energy demand is simply the sum as follows:

$$E = \text{Sum of } (A_i I_i).$$

Accounting models are essentially spreadsheet programs in which energy flows and related information such as carbon emissions are tracked through such identities. The interpretation of the results, and the ranking of different policies quantified in this manner, is external to the model and relies primarily on the judgment of the analyst.

Note that these calculations assume that a number of factors are held constant, including energy service level and equipment saturations. Also, these expressions represent a quasi-static view; in actual practice, such calculations would be performed over time paths of costs, activities, intensities, and prices developed in scenario construction. Finally, it is easy to see that factors for carbon savings from the shift to efficient technologies can be easily included in such calculations.

In energy accounting models, macroeconomic factors enter only as inputs in deriving demand-side projections—that is, there is no explicit representation of feedback from the energy sector to the overall economy. While different models contain different levels of detail in representing the supply sector, supply-demand balancing in this type of model is accomplished by back calculation of supply from demand projections.

In contrast to optimization models, accounting models cannot easily generate a least-cost mitigation solution. They can be used to represent cost-minimizing behavior estimated by the analyst, however. They tend to require less data and expertise and are simpler and easier to use than optimization models.

3.2 Engineering Optimization Models

In engineering optimization models, the model itself provides a numerical assessment and comparison of different policies. These models are linear programs in which the most basic criterion is total cost of providing economy-wide energy services under different scenarios; when this criterion is used, the structure of this type of model as used in mitigation analysis can be represented schematically as follows:

Minimize total cost of providing energy
and satisfying end-use demand subject to

1. Energy supplied, energy demanded
2. End-use demands satisfied
3. Available resource limits not exceeded

In addition to the overall optimization structure of these models, perhaps the key distinction between these and the accounting models is that, within the model structure itself, trade-offs are made among different means of satisfying given end-use demands for energy services.

The intertemporal structure of these linear programming models varies. Some are constructed to perform a target year analysis: the model is

first parameterized and run for the base year, then the procedure is repeated for a single designated future year (typically 2005, 2010, or 2020). Others perform a more elaborate dynamic optimization, in which time paths of the parameters are incorporated and the model generates time paths of solutions.

In engineering optimization models, macroeconomic factors enter in two ways. First, they are used to construct forecasts of useful energy demands. Second, they can be introduced as constraints. For example, the overall cost minimization can be constrained by limits on foreign exchange or capital resources. In both cases, the models do not provide for the representation of feedbacks from the energy sector to the overall economy.

Supply and demand are balanced in engineering optimization models by the presence of constraints, as indicated earlier. The engineering detail and level of disaggregation used in both the supply and demand side are at the discretion of the user, and in practice these vary widely among models.

This type of model allows several means of analyzing GHG emissions and the effects thereupon of various policy options. For example, as an alternative to minimizing energy costs, criteria such as minimizing carbon output subject to the constraints can be employed. In addition, an overall cap on carbon emissions can be entered as a constraint in the model and the cost minimization performed with this restriction. Each such approach allows the comparison of different policy intervention.

3.3 Iterative Equilibrium Models

These models incorporate the dynamics of market processes related to energy via an explicit representation of market equilibrium—that is, the balancing of supply and demand. These are used to model a country's total energy system and do not explicitly include an economy model integrated with the energy system model. Thus, macroeconomic factors enter the model exogenously, as in the previous model types discussed. (That is, demands for energy services are derived from macroeconomic drivers rather than being obtained endogenously.) These models thus occupy an intermediate position between engineering, energy-focused models, and pure market equilibrium models.

The methodology employed to solve the model is a process network wherein individual energy processes are represented with standard model forms, with application specific data, and linked together as appropriate. Prices and quantities are then adjusted iteratively until equilibrium is achieved. This iterative approach makes it much easier to include noncompetitive-market factors in the system than in the optimization approach.

3.4 Hybrid Models

In hybrid models, the basic policy measure is the maximization of the present value of the utility of a representative consumer through the model planning horizon. Constraints are of two types: macroeconomic relations among capital, labor, and forms of energy, and energy system constraints. The model generates different time paths of energy use and costs and macroeconomic investment and output. The energy submodel contains representations of end-use technologies, with different models containing different levels of detail. Schematically, this type of model can be represented as follows:

Maximize (discounted) utility of consumption

subject to

1. Macroeconomic relations among output, investment, capital, labor, and energy
2. Energy system and resource constraints (as in engineering optimization models)

The constraints in this case are also dynamic: they represent time paths for the model variables.

In this type of model, energy demands are endogenous to the model rather than imposed exogenously by the analyst. In addition, this optimization structure indicates the difference we noted earlier in the way the different models incorporate macroeconomic data. Specifically, in accounting and engineering optimization models, these data—on GDP, population growth, capital resources, and so on—enter essentially in the underlying constructions of the baseline and policy scenarios. In the hybrid model, however, such data enter in the macroeconomic relations (technically, the aggregate production function) as elasticities and other parameters. Within this model framework, changes in energy demand and supply can feed back to affect macroeconomic factors. It should be noted that, despite their inclusion of engineering optimization subcomponents, these models typically do not contain as much detail on specific end-use technologies as many purely engineering models.

4. KEY CHALLENGES IN THE BOTTOM-UP MODELING APPROACH

A number of key challenges arise in the bottom-up modeling approach. These include (1) incorporating efficiency versus equity; (2) aggregation over time, regions, sectors, and consumers; (3) representing decision rules used by consumers; (4) incorporating technical change; (5) capturing facility retirement dynamics; (6) avoiding extreme solutions; and (7) accounting for carbon flows.

4.1 Incorporating Efficiency versus Equity

None of the models discussed earlier provide explicitly for making trade-offs between efficiency and equity. Different models, however, have different implications for the analyst's consideration of this important issue. Nonoptimization models do not themselves choose among different policies in an explicit way but can allow for the ranking of policies according to criteria specified by the analyst, including considerations of equity. Engineering optimization models, since they focus on least-cost solutions to the provision of energy services, leave to the analyst the judgment of how to trade-off the importance of energy with that of other economic and social priorities. The models that optimize the utility of a representative consumer, in a sense, constrain consideration of the issue the most. Embedded in this modeling structure is a view of the economic system that equates social optima with competitive economic equilibria; the appropriateness of this perspective in the application at hand must be weighed carefully.

4.2 Aggregation over Time, Regions, Sectors, and Consumers

Perhaps the most fundamental formulation issue is the level of aggregation at which costs and benefits are calculated. Economic efficiency is generally insured if discounted net benefits are maximized in the aggregate; any desired income redistribution is handled subsequently. Decision makers, however, are fundamentally interested in how the costs and benefits fall on various income, industry, and regional groups. Coupled with the relative emphasis of the analysis on equity versus efficiency is the desired level of disaggregation of the model by region, time periods, industry, and income group. Obviously, the level of disaggregation must be sufficient to allow reporting of results at the desired level, but in some cases the projection of an aggregate variable can be improved by some level of disaggregation to capture the heterogeneity in decision making objects on the part of the different groups. These decision rules themselves are critical elements of the models and range from minimizing discounted future costs (or maximizing benefits) over a 40- or 50-year time horizon to picking investments that come close to minimizing costs based on conditions for a single year only.

4.3 Representing Decision Rules Used by Consumers

Accounting models contain no explicit representation of consumer decision making. In practice, however, their use often reflects the view that certain market barriers constrain consumers from making optimal decisions with respect to energy efficiency. At the other extreme, the use of the representative consumer in the optimization models rests on strong assumptions regarding consumer behavior, serving primarily to ensure mathematical and empirical tractability. Key among these are perfect foresight and essential homogeneity among consumers or households.

4.4 Incorporating Technical Change

Another set of key assumptions about inputs are those made about the costs and efficiencies of current and future technologies, both for energy supply and energy use. Most analysts use a combination of statistical analysis of historical data on the demand for individual fuels and a process analysis of individual technologies in use or under development to represent trends in energy technologies. At some point these two approaches tend to look quite similar though, as the end-use process analysis usually runs out of new technology concepts after some years or decades, and it is then assumed that the efficiency of the most efficient technologies for which there is an actual proposed design will continue to improve as time goes on. Particularly important, but difficult, here is projecting technological progress. Almost all top-down models have relied on the assumption of "exogenous" or "autonomous" technical change, that is, productivity trends (relating to energy as well as other factors of production) that are a function of time only, and in particular not affected by either changes in relative factor prices or by policy

interventions. Bottom-up approaches, by contrast, assume (usually implicitly) that technological progress can be accelerated by policy intervention. This difference constitutes another significant contrast between the two approaches. There has been an acceleration of research, among top-down modelers, on incorporating technological change that is "endogenous"—that is, responsive to price changes, policies, or both. This research may eventually contribute to a partial reconciliation of the two approaches in the treatment of technological progress.

4.5 Capturing Facility Retirement Dynamics

Most modeling approaches focus on investments in new energy producing and consuming equipment, which is typically assumed to have a fixed useful lifetime. In scenarios where conditions change significantly (either through external factors or explicit policy initiatives), it may be economic to retire facilities earlier or later than dictated purely by physical depreciation rates. This endogenous calculation of facility retirement dates can be handled analytically in most models, but it represents a major increase in data and computational requirements.

4.6 Avoiding Extreme Solutions

Another typical problem, particularly with models that assume optimizing behavior on the part of individual economic agents, is the danger of knife edge results, where a small difference in the cost of two competing technologies can lead to picking the cheaper one only. This is generally handled by disaggregating consumers into different groups who see somewhat different prices for the same technology (e.g., coal is cheaper in the coal fields than a thousand miles away), modeling the decision process as somewhat less than perfect, or building appropriate time lags into the modeling structure.

4.7 Accounting for Carbon Flows

Finally, estimating carbon flows for a given energy system configuration can be complicated. It is more accurate to measure emissions as close to the point of combustion as possible so types of coal and oil product can be distinguished and feedstocks (which do not necessarily produce carbon emissions) can be netted out. However, a point-of-use model requires far more data and computation than the models described here, which aggregate several fuel types and use average carbon emissions factors for each fossil fuel.

5. DATA SOURCES FOR BOTTOM-UP ANALYSIS

Regardless of the approach taken and analysis tool used, the collection of reliable data is a major and relatively time-consuming aspect of bottom-up analysis. To keep data constraints from becoming a serious obstacle to the analysis, two points are essential. First, the bottom-up model should be sufficiently flexible to adapt to local data constraints. Second, the data collection process should be as efficient as possible. Efficiency can be maximized by focusing the detailed analysis on sectors and end uses, where the potential for GHG mitigation is most significant, and avoiding detailed data collection and analysis in other sectors.

Data collection generally begins with the aggregate annual energy use and production figures typically found on a national energy balance sheet. The remaining data requirements depend largely on (1) the disaggregated structure of the analysis, (2) the specific technology and policy options considered, and (3) local conditions and priorities.

Table II shows the typical types of data needed for a bottom-up approach to mitigation analysis. They tend to fall within five general categories: macroeconomic and socioeconomic data, energy demand data, energy supply data, technology data, and emission factor data. The full listing of potential data requirements may appear rather daunting. In practice, however, much of the data needed may already be available in the form of national statistics, existing analytical tools, and data developed for previous energy sector studies. The development and agreement on baseline projections of key variables, the characterization of mitigation options relevant to local conditions, and, if not already available, the compilation of disaggregated energy demand data are typically the most challenging data collection tasks facing the analyst.

In general, emphasis should be placed on locally derived data. The primary data sources for most assessments will be existing energy balances, industry-specific studies, household energy surveys, electric utility company data on customer load profiles, oil company data on fuel supply, historical fuel price series maintained by government departments, vehicle statistics kept by the transportation department,

TABLE II

Data Sources for a Bottom-Up Mitigation Analysis

Data categories	Types of data	Common data sources
Macroeconomic variables		
Aggregate driving variables	GDP/value added, population, household size	National statistics and plans; macroeconomic studies
More detailed driving variables	Physical production for energy intensive materials; transportation requirements; agricultural production and irrigated area; changes in income distribution, etc.	Macroeconomic studies; transport sector studies; household surveys, etc.
Energy demand		
Sector and subsector totals	Fuel use by sector/subsector	National energy statistics, national energy balance, energy sector yearbooks (oil, electricity, coal, etc.)
End-use and technology characteristics by sector/subsector	Energy consumption breakdown by enduse and device: e.g., energy use characteristics of new versus existing building stock; vehicle stock; breakdown by type, vintage, and efficiencies; or simpler breakdowns	Local energy studies; surveys and audits; studies in similar countries; general rules of thumb from end-use literature
Response to price and income changes (optional)	Price and income elasticities	Local econometric analyses; energy economics literature
Energy supply		
Characteristics of energy supply, transport, and conversion facilities	Capital and O&M costs, performance (efficiencies, unit intensities, capacity factors, etc.)	Local data, project engineering estimates, Technical Assessment Guide; IPCC Technology Characterization Inventory
Energy prices		Local utility or government projections; for globally traded energy products
Energy supply plans	New capacity online dates, costs, characteristics	National energy plans; electric utility plans or projections; other energy sector industries (refineries, coal companies, etc.)
Energy resources	Estimated, proven recoverable reserves of fossil fuels; estimated costs and potential for renewable resources	Local energy studies
Technology options		
Technology costs and performance	Capital and O&M costs, performance (efficiencies, unit intensities, capacity factors, etc.)	Local energy studies and project engineering estimates; technology suppliers; other mitigation studies
Penetration rates	Percent of new or existing stock replaced per year; overall limits to achievable potential	
Emission factors	Kg GHG emitted per unit of energy consumed, produced, or transported	National inventory assessments; IPCC Inventory Guidelines IPCC Technology Characterization Inventory

and so on. The main thrust of the data collection effort is not so much on collecting new primary data but on collating secondary data and establishing a consistent data set suitable for analysis using the model of choice.

Where unavailable, particularly in developing country analyses, local data can be supplemented with judiciously selected data from other countries. For example, current and projected cost and performance data for some mitigation technologies (e.g., high-efficiency motors or combined cycle gas units) may be unavailable locally, particularly if the technologies are not presently in wide use. For this purpose, technology data from other countries can provide indicative figures and a reasonable starting point. For data on energy use patterns, such as the fraction of electricity used for motor drive in the textile industry, the use of extemal data can be somewhat more problematic. In general, it may be possible to use estimates and general rules of thumb

suggested by other country studies, particularly data from other countries with similar characteristics.

6. DEVELOPING SCENARIOS FOR USE IN A BOTTOM-UP ASSESSMENT

A bottom-up assessment typically compares the energy and environmental consequences of one future scenario against another. The scenarios may be composed of widely different views of the world (e.g., a rapidly growing economic world versus a slowly growing one). These types of scenarios allow companies to prepare for operations in either type of future. Often, governments though may wish to analyze the consequences of their policy or programmatic actions and in such a case it is desirable to compare a policy scenario against a reference one. The latter is often referred to as a business-as-usual or baseline scenario. Developing these scenarios is a complex process. It requires the combining of both analyses with some judgment of the future evolution of variables that are likely to affect the energy and environment system.

6.1 Developing a Baseline Scenario

Developing a baseline scenario that portrays social, demographic, and technological development over a 20- to 40-year or longer time horizon can be one of the most challenging aspects of a bottom-up analysis. The levels of projected future baseline emissions shape the amount of reductions required if a specific policy scenario target is specified and the relative impacts and desirability of specific mitigation options. For instance, if many low-cost energy efficiency improvements are adopted in the baseline scenario, this would yield lower baseline emissions and leave less room for these improvements to have an impact in a policy scenario.

Development of a baseline scenario begins with the definition of scenario characteristics (e.g., business as usual). Changes in exogenous driving variables are then specified and entered into the model, which is run to simulate overall energy use and emissions over the time horizon selected. The baseline scenario is evaluated for reasonableness and consistency and revised accordingly. Uncertainty in the evolution of the baseline scenario can be reflected through a sensitivity analysis of key parameters such as GDP growth.

The procedure will vary somewhat depending on the modeling approach used and the nature of a baseline scenario. In an optimization model, the use of different technologies is to a certain degree decided within the model, dependent on how much one wants to constrain the evolution of the baseline scenario. For example, the analyst might choose to construct a baseline scenario in which the energy supply system closely reflects or extrapolates from published plans. Alternately, the analyst might choose to give the model more flexibility to select a future energy supply system based on specific criteria. If one is using an optimization model, it is necessary to introduce certain constraints if one wishes to force the model toward a solution that approximates a business-as-usual future.

The rate of economic growth and changes in domestic energy markets are among the most important assumptions affecting projected baseline emissions. Official government GDP projections may differ from other macroeconomic projections. In terms of domestic energy markets, the removal of energy price subsidies could greatly affect fuel choice and energy efficiency, and thus baseline emissions and the impacts of mitigation options.

A preparatory step in developing a baseline scenario is to assemble available forecasts, projections, or plans. These might include national economic development plans, demographic and economic projections, sector-specific plans (e.g., expansion plans for the iron and steel industry), plans for transport and other infrastructure, studies of trends in energy use (economy wide, by sector, or by end use), plans for investments in energy facilities (electricity expansion plans, new gas pipelines, etc.), studies of resource availability, and projections of future resource prices. In short, all studies that attempt to look into a country's future—or even the future of a region—may provide useful information for the specification of a baseline scenario. However, it is unlikely that every parameter needed to complete the baseline scenario will be found in national documents or even that the documents will provide a consistent picture of a country's future. As with much of the modeling process, the judgment of the analyst in making reasonable assumptions and choices is indispensable.

6.1.1 Developing Projections of Energy Demand

In bottom-up approaches, projections of future energy demand are based on two parameters for each subsector or end use considered: a measure of the activity that drives energy demand and a measure of the energy intensity of each activity, expressed in energy units per unit of activity.

Measures of activity include data on household numbers, production of key industrial products, and demand for transport and services. Activity may be measured in aggregate terms at the sectoral level (e.g., total industrial value added, total passenger-km or ton-kin) and by using indicators at the subsector level. These two measures need not be identical. For example, total industrial value added is a common indicator for aggregate activity for the industrial sector, but for specific subsectors such as steel or cement one often uses tons of production as a measure of activity. In general, physical measures of activity are preferable, but they are not appropriate in all cases (such as in light industry, where there is no aggregate measure of physical production).

In bottom-up approaches, future values for driving activities are exogenous—that is, based on external estimates or projections rather than being estimated by the model itself. Future values can be drawn from a variety of sources or estimated using various forecasting methods. Estimates of the future levels of activity or equipment ownership depend on local conditions and the behavioral and functional relationships within each sector.

Projections of the future development of energy intensities in each subsector or end use can be expressed in terms of final energy or useful energy. When the choice of end-use options is conducted within the model, however, the energy demand should be given in terms of useful energy to allow the model to select among technologies for meeting the energy requirements.

Projections should start from the base year values, such that the sum of the product of energy intensity and activity level in each subsector add up to total final or useful energy use in the base year. In energy statistics, the data are normally presented in terms of final energy. If the useful energy refers to the amount of energy required to meet particular demands for energy services. It is typically estimated by multiplying final energy consumption by the average conversion efficiency of end-use equipment (e.g., of oil-fired boilers). Applying the concept of useful energy is more difficult in the transport sector, although one can use the estimated conversion efficiency of generic engine types in various vehicle classes. If the projections are to be given in useful energy units, the statistical data should be converted using estimated base year efficiencies for each end use.

Ideally, the projections of energy intensities are based in part on historical developments. To the extent statistical data on a disaggregated level are available, they give limited information for economies that have undergone significant structural changes or changes in taxation/subsidies on energy. Even if reliable historical data are available, assumptions on the development of the energy intensities have to be made using careful judgment about the status of existing end-use technologies and future efficiency improvements that should be included in the projections.

It is important to distinguish between improvements included in the exogenous demand projections and improvements that are explicitly included as technology options in the model. For example, future improvements of building insulation standards can either be directly included in the demand projections as an improvement of intensity for space heating or modeled as technology options that can be chosen individually in the assessment, depending on their attractiveness in the different scenarios. The distinction is especially important in optimization models. One can assume that the insulation standards will be implemented (e.g., through use of regulations) or allow the model to choose the implementation. In the latter case, the insulation option will be implemented if its cost is less than the cost of supplying the heat using available options for space heating.

Once the initial baseline scenario is prepared, it is reviewed to assess whether it presents a comprehensive and plausible future for the country in light of real-world constraints. Some specific questions might include the following:

- Can the indicated growth rates in energy demand be sustained over the study period? Is it a reasonable rate, given recent experience in the country and region?
- Is the level of capital investment needed to sustain the indicated levels of industrial growth likely to be forthcoming?
- Will the country be able to afford the bill for fuel imports that is implied by the baseline scenario?
- Will the capital investment needed for energy supply system expansion be available, given competition for limited financial resources?
- Is the indicated increase in transportation use plausible, given current and planned transportation infrastructure?
- Are the emission factors in use appropriate for future technologies?

Answers to these types of questions might indicate the need for adjustments to the baseline scenario or sensitivity analyses of key parameters.

6.2 Developing Policy Scenarios

The process of developing a policy scenario or scenarios involves establishing a scenario objective and combining specific options in an integrated scenario. Integrated scenario analysis is essential for developing accurate and internally consistent estimates of overall cost and emissions impacts since the actual emissions reduction from employing a specific option can depend on the other options included in a scenario. For instance, the level of reduction in GHG emission associated with an option that saves electricity is dependent on the electricity generation resources whose use would be avoided (e.g., coal, oil, hydro, or a mix). In reality, the type of electricity generation that the efficiency option would avoid will change over time, and, if lower GHG emitting electricity resources are also introduced in the scenario, the GHG savings of the efficiency option may be reduced. Integrated scenario analyses are intended to capture these and other interactive effects.

Where using an optimization model, the difference in input data for the baseline scenario and the policy scenario(s) is typically less than in an accounting model, where the choice of technologies is exogenous to the model. An optimization model chooses from the whole set of available technologies to satisfy the given constraints.

6.2.1 Objectives for a Policy Scenario

Several objectives are possible for designing a policy scenario. The objective depends on political and practical considerations. Types of objectives include the following:

- *Emission reduction targets.* For example: 12.5% reduction in CO_2 emissions by 2010, and 25% reduction by 2030, from baseline levels. An alternative is to specify reductions from base year levels, which avoids making the amount of reduction dependent on the specification of the baseline scenario.
- *Identification of options up to a certain cost per ton of energy or emissions reduction.* The energy or emissions reduction given by the resulting technology mix would reflect the level of reduction that could be achieved at a certain marginal cost.
- *No-regrets scenario.* This scenario is a common variant of the previous type of objective, where the screening threshold is essentially zero cost per tonne of energy or GHG reduced.
- *Specific technology options or packages of options.* Examples of this type of scenario might be a natural gas scenario, a renewable energy scenario, or a nuclear scenario.

7. RESULTS FROM A BOTTOM-UP APPROACH

Following are some key questions that an analyst would seek to answer from a bottom-up energy sector assessment:

1. What is the economic cost of providing energy services in a baseline or policy scenario? What is the incremental cost between scenarios?

1a. What are the capital and foreign exchange implications of pursuing alternative scenarios?

2. What is the economic cost of pursuing particular mitigation (policy and technology) options, such as high efficiency lighting or a renewable technology, and what are its local and global environmental implications?
3. What are the costs of reducing emissions to a predetermined target level? (Target may be annual or cumulative).
4. What is the shape of the marginal cost curve for reducing carbon emissions?

4a. How do alternative technologies rank in terms of their carbon abatement potential?

Any one of the three types—accounting, optimization, and iterative equilibrium—of models discussed here can address questions 1 and 2. Question 1a is important for developing country planners, since these two cost components are often insufficient in these countries.

Question 3 is easiest to address using an optimization model. Question 4 requires that energy supply and demand be evaluated in an integrated manner. A demand-side mitigation measure may change the energy supply configuration measure, which will affect the GHG emissions of the energy system. The optimization and iterative models are capable of capturing the integrated effect and deriving the changes in environmental emissions. The accounting models may not capture the changes in the energy supply mix and the consequent GHG emissions and thus may show higher emissions reduction for a mitigation measure. Only the optimization models can calculate marginal costs directly. In the other models, an approximation of the marginal cost curve can be constructed by performing a large number of carefully selected runs.

SEE ALSO THE FOLLOWING ARTICLES

Economics of Energy Demand • Economics of Energy Supply • Modeling Energy Markets and Climate Change Policy • Modeling Energy Supply and Demand: A Comparison of Approaches • Multicriteria Analysis of Energy • National Energy Modeling Systems

Further Reading

IPCC (2001). Third Assessment Report, Chapter 8.

Krause, F., Baer P., and DeCanio, S. (2001). Cutting Carbon emissions at a Profit: Opportunities for the US. IPSEP, El Cerrito CA.

LBNL and ORNL (2000). Clean Energy Futures for the U.S.

Sathaye, J., and Ravindranath, N. (1998). Climate change mitigation in the energy and forestry sectors of developing countries. *Annual Rev. Energy and Environment* **23,** 387–437.

Business Cycles and Energy Prices

STEPHEN P. A. BROWN, MINE K. YÜCEL, and JOHN THOMPSON
Federal Reserve Bank of Dallas
Dallas, Texas, United States

1. **Definition of Business Cycles and the Role of Energy Prices**
2. **Understanding the Basic Response**
3. **Amplification of the Basic Response**
4. **Asymmetry of the Response**
5. **A Weakening Response**
6. **Conclusions and Implications for Policy**

Glossary

aggregate channels Changes in economic activity that work through changes in aggregate supply or demand.
aggregate economic activity The sum total of economic activity in a country, which is commonly measured by gross domestic product (GDP).
allocative channels Changes in economic activity that work through changes in the composition of output.
asymmetry A process in which aggregate economic activity is reduced more by an oil price increase than it is increased by an oil price decrease.
business cycles Fluctuations in aggregate economic activity.
consumption smoothing The process of adjusting consumption spending to lifetime income by holding consumption relatively constant when there are short-term fluctuations in income.
core inflation A measure of inflation that is thought to provide a better signal of underlying inflationary pressure because it excludes food and energy prices, which are quite volatile.
federal funds rate The interest rate at which commercial banks borrow from each other's reserves that are on deposit with the Federal Reserve System.
net oil price An oil price series that captures no price decreases and only those increases that take the price of oil higher than it has been during the past year.
neutral monetary policy The conduct of monetary policy in such a way that it has no effect on economic activity. According to Robert Gordon of Northwestern University, a neutral monetary policy leaves nominal GDP unchanged. According to Milton Friedman, who is retired from the University of Chicago, a neutral monetary policy requires that the money supply be unchanged. Many other economists regard a constant federal funds rate as neutral monetary policy. In the absence of aggregate supply shocks, all three definitions of neutrality are consistent with each other. Aggregate supply shocks can cause a divergence between these three measures of monetary neutrality.
oil price shock A sharp change in the price of oil.
putty–clay technology Technology in which the energy-to-output, capital-to-output, and labor-to-output ratios can be varied over the long term as capital is purchased and installed but cannot be changed in the short term because they are embedded in the capital stock.
real interest rate The market interest rate minus expected inflation.
sticky wages and prices The inability of market forces to change wages and prices. This stickiness usually occurs when market forces would reduce nominal wages and prices.

Oil price shocks have figured prominently in U.S. business cycles since the end of World War II, although the relationship seemed to weaken during the 1990s. In addition, the economy appears to respond asymmetrically to oil price shocks; increasing oil prices hurt economic activity more than decreasing oil prices help it. This article focuses on an extensive economics literature that relates oil price shocks to aggregate economic activity. It examines how oil price shocks create business cycles, why they seem to have a disproportionate effect on economic activity, why the economy responds asymmetrically to oil prices, and why the relationship between oil prices and economic activity may have weakened. It also addresses the issue of developing energy policy to mitigate the economic effects of oil price shocks.

 © 2004 Elsevier Inc. All rights reserved.

1. DEFINITION OF BUSINESS CYCLES AND THE ROLE OF ENERGY PRICES

Despite occasional periods of notable recession, the U.S. economy has generally traveled an expansionary path throughout its history. In fact, U.S. gross domestic product (GDP) has averaged a 3.6% annual rate of increase since 1929. Population increases account for some of the growth, but productivity gains have also been essential to long-term national economic growth.

Although the U.S. economy has generally expanded over time, its upward trend has not been free of interruption. Shocks have occasionally hit the U.S. economy, disrupting the expansionary forces and creating business cycles. Since World War II, oil price shocks have played a significant role in U.S. business cycles. In fact, increasing oil prices preceded 9 of the 10 recessions that occurred from the end of World War II through 2002 (Fig. 1). Research conducted by James Hamilton strongly suggests that the recessions that followed sharply rising oil prices were not the result of other business cycle variables, such as aggregate demand shocks or contractionary monetary policy.

By examining the influence of oil price shocks on U.S. economic activity in the six decades since the end of World War II, economists have found that oil price shocks seem to have a disproportionate effect on economic activity. In addition, the economy appears to respond asymmetrically to oil price movements—that is, the gains in economic activity that follow oil price declines are not commensurate with the losses in economic activity that follow oil price increases. In addition, oil price shocks seem to have had less effect on economic activity during the 1990s than in previous decades.

These observations suggest a number of questions. What basic factors account for the negative influence that sharply rising oil prices have on U.S. economic activity? Why do sharply rising oil prices have a disproportionate effect in weakening economic activity? Why doesn't the economy respond as favorably to declining oil prices as it responds unfavorably to increasing oil prices? Why has the economy become less sensitive to oil price shocks? How does our understanding of the effect of oil price shocks on economic activity influence desired energy policies?

2. UNDERSTANDING THE BASIC RESPONSE

The oil price shock of 1973 and the subsequent recession, which (at the time) was the longest of the post-World War II recessions, gave rise to many studies on the effects of oil price increases on the economy. The early research documented and sought to explain the inverse relationship between oil price increases and aggregate economic activity. Subsequent

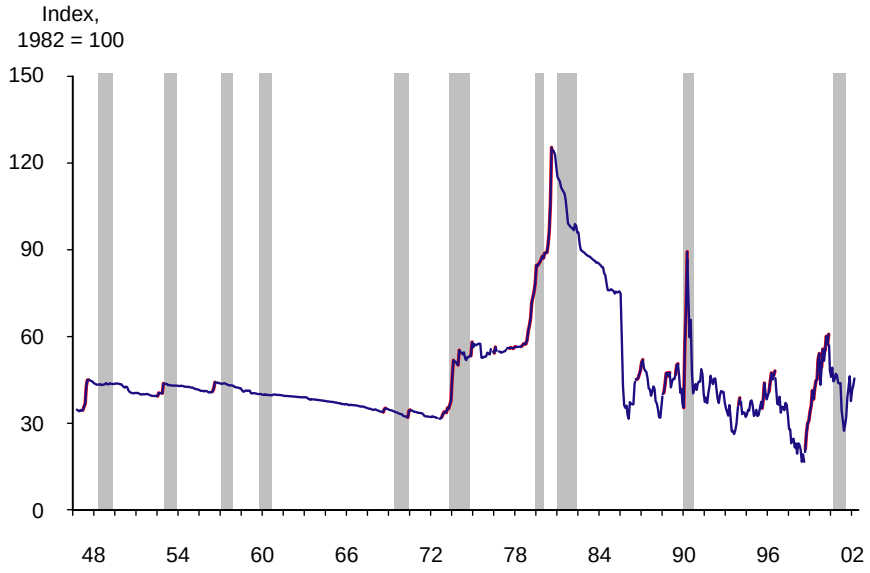


FIGURE 1 Real oil prices and U.S. business cycles. Episodes of sharply rising oil prices have preceded 9 of the 10 post-World War II recessions in the United States (bars).

empirical studies confirmed an inverse relationship between oil prices and aggregate economic activity for the United States and other countries. The latter research included James Hamilton's finding that other economic forces could not account for the negative effect that rising oil prices had on U.S. economic activity.

Economists have offered a number of explanations for why rising oil prices hurt aggregate U.S. economic activity. The most basic explanation is the classic supply shock, in which rising oil prices are indicative of the reduced availability of an important input to production. Another explanation is that rising oil prices result in income transfers from oil-importing nations to oil-exporting nations, which reduces U.S. aggregate demand and slows economic activity.

The remaining explanations sought to attribute the effects of oil price shocks to developments in the financial markets. One is simply that the monetary authorities responded to rising oil prices with a contractionary monetary policy that boosted interest rates. Another is that rising oil prices led to increased money demand as people sought to rebalance their portfolios toward liquidity. A failure of the monetary authority to meet growing money demand with an increased money supply boosted interest rates. In either case, rising interest rates retarded economic activity.

Sorting through the explanations that economists have offered for the inverse relationship between oil prices and economic activity requires a comparison between how the economy has responded to oil price shocks historically and how the various explanations suggest the economy should respond. In the United States, past episodes of rising oil prices have generally resulted in declining GDP, a higher price level, and higher interest rates (Table I). As discussed later, of the various explanations that have been offered for why rising oil prices hurt economic activity, only the classic supply-side effect can account for declining GDP, a rising price level, and higher interest rates.

TABLE I

Expected Responses to Increasing Oil Price

	Real GDP	Price level	Interest rate
Historical record	Down	Up	Up
Classic supply shock	Down	Up	Up
Aggregate demand shock	Down	Down	Down
Monetary shock	Down	Down	Up
Real balance effect	Down	Down	Up

2.1 A Classic Supply Shock

Many economists consider an oil price shock to be illustrative of a classic supply shock that reduces output. Elevated energy prices are a manifestation of increased scarcity of energy, which is a basic input to production. With reduced inputs with which to work, output and labor productivity are reduced. (In more mild cases, the growth of output and productivity are slowed.) In turn, the decline in productivity growth reduces real wage growth and increases the unemployment rate.

If consumers expect such a rise in oil prices to be temporary, or if they expect the short-term effects of output to be greater than the long-term effects, they will attempt to smooth their consumption by saving less or borrowing more. These actions boost the real interest rate. With reduced output and a higher real interest rate, the demand for real cash balances declines and the price level increases (for a given rate of growth in the monetary aggregate). Therefore, higher oil prices reduce real GDP, boost real interest rates, and increase the price level (Table I). The expected consequences of a classic supply shock are consistent with the historical record.

2.2 Income Transfers and Reduced Aggregate Demand

When oil prices increase, there is a shift in purchasing power from the oil-importing nations to the oil-exporting nations. Such a shift is another avenue through which oil price shocks could affect U.S. economic activity. Rising oil prices can be thought of as being similar to a tax that is collected from oil-importing nations by oil-exporting nations. The increase in oil prices reduces purchasing power and consumer demand in the oil-importing nations. At the same time, rising oil prices increase purchasing power and consumer demand in the oil-exporting nations. Historically, however, the increase in consumer demand occurring in oil-exporting nations has been less than the reduction in consumer demand in the oil-importing nations. On net, world consumer demand for goods produced in the oil-importing nations is reduced.

Reduced consumer demand for goods produced in the oil-importing nations increases the world supply

of savings, which puts downward pressure on interest rates. Lower interest rates could stimulate investment, partially offsetting the lost consumption spending and partially restoring aggregate demand. The net result, however, is a reduction in aggregate demand.

The reduction in aggregate demand puts downward pressure on the price level. Economic theory suggests that real prices will continue to decline until aggregate demand and GDP are restored to preshock levels. However, if nominal prices are sticky downward, as most economists believe, the process of adjustment will not take place, and aggregate demand and GDP will not be restored—unless unexpected inflation increases as much as GDP growth declines.

The reduction in aggregate demand necessitates a lower real price level to yield a new equilibrium. If the real price level cannot decline, consumption spending will decline more than investment spending increases. Consequently, aggregate demand and output are reduced. With nominal prices downward sticky, the only mechanism by which the necessary reduction in real prices can occur is through unexpected inflation that is at least as great as the reduction in GDP growth. Domestically, the necessary change in inflation can be accomplished through a monetary policy that holds the growth of nominal GDP constant. A monetary policy that allows nominal GDP to decline will not generate enough unexpected inflation to restore aggregate demand and output.

To the extent that income transfers and reduced aggregate demand account for the aggregate effects of oil price shocks, monetary policy would have to allow nominal GDP to decline. The aggregate effects would be lower interest rates, reduced real GDP, and inflation that decreased or increased by less than the reduction in real GDP growth (Table I). These effects are inconsistent with the historical record for the United States, which shows that interest rates rise, real GDP declines, and the price level increases by as much as real GDP declines. Such a record seems to indicate that the effects of income transfers are negligible at best.

2.3 Monetary Policy

Monetary policy was prominent among the early explanations of how oil price shocks affected aggregate economic activity, but it was gradually supplanted by real business cycle theory, which attributed the effects to classic supply shocks rather than monetary policy. Nevertheless, an apparent breakdown in the relationship between oil and the economy during the 1980s and 1990s led researchers to question the pure supply shock theory of real business cycle models and to revisit other channels through which oil could affect the economy, such as changes in monetary policy.

However, it is difficult to explain the basic effects of oil price shocks on aggregate economic activity using monetary policy. Restrictive monetary policy will result in rising interest rates, reduced GDP growth, reduced inflation, and reduced nominal GDP growth (Table I). This is inconsistent with the historical record.

2.4 The Real Balance Effect

The real balance effect was one of the first explanations that economists offered for the aggregate economic effects of an oil price. In this theory, an increase in oil prices led to increased money demand as people sought to rebalance their portfolios toward liquidity. The failure of the monetary authority to meet growing money demand with an increased money supply would boost interest rates and retard economic growth, just like a reduction in the money supply. Adjustment would put downward pressure on the price level. The effects would be higher interest rates, lower GDP, and a reduced price level, inconsistent with the historical record (Table I).

2.5 Sorting through the Basic Theories

Of the explanations offered for the inverse relationship between oil price shocks and GDP growth, the classic supply-side shock argument best explains the historical record (Table I). The classic supply shock explains the inverse relationship between oil prices and real GDP and the positive relationship between oil price shocks and measured increases in inflation and interest rates. Income transfers that reduce aggregate demand can explain reduced GDP but cannot explain rising interest rates. Neither monetary policy nor the real balance effect can explain both slowing GDP growth and increased inflationary pressure.

3. AMPLIFICATION OF THE BASIC RESPONSE

A number of economists have recognized that basic supply-shock effects can account for only a portion of the intense effect that oil price shocks have on

aggregate economic activity. Consequently, additional explanations for the intensity of the response are important. Among the possible explanations are restrictive monetary policy, adjustment costs, coordination externalities, and financial stress.

3.1 Monetary Policy

Monetary policy can influence how an oil price shock is experienced. When the monetary authorities hold the growth of nominal GDP constant, the inflation rate will accelerate at the same rate at which real GDP growth slows. To the extent that the market is slow to adjust to monetary surprises, a more accommodative monetary policy, which is accomplished by reducing interest rates, would temporarily offset or partially offset the losses in real GDP while it increased inflationary pressure. A counterinflationary (restrictive) monetary policy, which is accomplished by increasing interest rates, would temporarily intensify the losses in real GDP while it reduced inflationary pressure.

In the case of counterinflationary monetary policy, adjustment could be lengthy. If nominal wages and prices are sticky downward, real wages and prices would fail to decrease as is necessary to clear the markets. Consequently, unemployment would increase, aggregate consumption would decline, and GDP growth would be slowed beyond that which would arise directly from the supply shock.

If wages are nominally sticky downward, the reduction in GDP growth will lead to increased unemployment and a further reduction in GDP growth, unless unexpected inflation increases as much as GDP growth decreases. The initial reduction in GDP growth is accompanied by a reduction in labor productivity. Unless real wages decrease by as much as the reduction in labor productivity, firms will lay off workers, which will generate increased unemployment and exacerbate GDP losses. Therefore, if nominal wages are sticky downward, the only mechanism by which the necessary wage reduction can occur is through unexpected inflation that is at least as great as the reduction in GDP growth.

Several lines of research assert that restrictive monetary policy accounts for much of the decline in aggregate economic activity following an oil price increase. Douglas Bohi reasoned that energy-intensive industries should be most affected if a classic supply shock explains the principal effects of oil price shocks. Using industry data for four countries, Bohi found no relationship between the industries affected and their energy intensity. He also found inconsistent effects of oil price shocks across countries and time. Asserting that monetary policy was tightened in each of the countries he examined, Bohi concluded that much of the negative impact of higher oil prices on output must be due to restrictive monetary policy.

Similarly, Ben Bernanke, Mark Gertler, and Mark Watson showed that the U.S. economy responds differently to an oil price shock when the federal funds rate is constrained to be constant than when it is unconstrained. In their unconstrained case, a sharp increase in oil prices leads to a higher federal funds rate and a reduction in real GDP. With the federal funds rate held constant, they found that an oil price increase leads to an increase in real GDP. Defining neutral monetary policy as holding the federal funds rate constant, they found that monetary policy has tightened in response to increased oil prices, and they conclude that this monetary tightening accounts for the fluctuations in aggregate economic activity.

Despite some findings to the contrary, other research casts doubt on the idea that monetary policy accounts for much of the response of aggregate economic activity to oil price shocks. James Hamilton and Ana Maria Herrera show that the findings of Bernanke *et al.* are sensitive to specification. Using a longer lag time, Hamilton and Herrera found that oil price shocks have a substantially larger direct effect on the real economy. Furthermore, with the longer lag time, Hamilton and Herrera found that even when the federal funds rate is kept constant, an oil price shock still yields a sizeable reduction in output, which implies that monetary policy has little effect on easing the real consequences of an oil price shock.

Hamilton and Herrera's findings are consistent with research conducted by others that showed that counterinflationary monetary policy was only partly responsible for the real effects of oil price shocks from 1970 to 1990. Some researchers agree that monetary policy has become more restrictive following an oil price shock but conclude that oil price shocks have a stronger and more significant impact on real activity than monetary policy.

Brown and Yücel argue that U.S. monetary policy likely has had no role in aggravating the effects of past oil price shocks. Using a specification similar to that of Bernanke *et al.*, Brown and Yücel found that oil price shocks lead to an increased federal funds rate, reduced real GDP, and an increased price level. The increase in the price level is approximately the same as the reduction in real GDP, which means that nominal GDP is unchanged. Such a finding conforms to Robert Gordon's definition of monetary neutrality,

which is achieved when monetary policy is adjusted to hold nominal GDP constant.

Brown and Yücel also criticize the assertion that an unchanged federal funds rate necessarily represents a neutral monetary policy. They found that holding the federal funds rate constant (in a counterfactual experiment) after an oil price shock boosts real GDP, the price level, and nominal GDP, which is consistent with Gordon's definition of accommodative monetary policy. In short, monetary policy can cushion the real effects of an oil price shock but at the expense of accelerating inflation.

The use of the federal funds rate to gauge the effects of monetary policy can be faulty when the economy is adjusting to a supply shock. As explained previously, the attempt by consumers to smooth consumption in response to a basic supply shock accounts for the higher interest rates. In a market with rising interest rates, a policy of holding the federal funds rate constant will accelerate money growth, contributing to short-term gains in real GDP and increased inflationary pressure.

In summation, oil price shocks create the potential for a monetary policy response that exacerbates the basic effects of an oil supply shock. The research assessing whether monetary policy has amplified the basic effects of a supply shock is contradictory, but the most compelling evidence suggests monetary policy has had a relatively small or no role in amplifying the basic effects in the United States. Furthermore, interest rates are not a good way to assess the effect of monetary policy when there is a supply shock, and measured in other ways, monetary policy has remained neutral in response to past oil price shocks.

3.2 Adjustment Costs

Adjustment costs are another way in which the basic response to rising oil prices might be amplified. Adjustment costs can arise from either capital stock that embodies energy technology or sectoral imbalances. In either case, adjustment costs can further retard economic activity.

To a great extent, the technology that a firm chooses for its production is embedded in the capital equipment it purchases. (Economists refer to this characteristic of technology and capital as "putty–clay" because the firm can vary its energy-to-output, capital-to-output, and labor-to-output ratios over the long term as it purchases capital but not in the short term.) With production technology embedded in the capital stock, a firm must change its capital stock to respond to rising energy prices. The consequences are slow adjustment and a disruption to economic organization when energy prices increase, with stronger effects in the short term compared to the long term.

In a similar way, changes in oil prices can also create sectoral imbalances by changing the equilibrium between the sectors. For example, rising oil prices require a contraction of energy-intensive sectors and an expansion of energy-efficient sectors. These realignments in production require adjustments that cannot be achieved quickly. The result is increased unemployment and the underutilization of resources.

3.3 Coordination Problems

Coordination problems are a potential outgrowth of sectoral imbalances. Coordination problems arise when individual firms understand how changing oil prices affect their output and pricing decisions but lack enough information about how other firms will respond to changes in oil prices. As a consequence, firms experience difficulty adjusting to each other's actions and economic activity is further disrupted when oil prices increase.

3.4 Uncertainty and Financial Stress

Uncertainty about future oil prices increases when oil prices are volatile, and such uncertainty reduces investment. When firms are uncertain about future oil prices, they find it increasingly desirable to postpone irreversible investment decisions. When technology is embedded in the capital, the firm must irreversibly choose the energy intensity of its production process when purchasing its capital. As uncertainty about future oil prices increases, the value of postponing investment decision increases, and the net incentive to invest decreases. In addition, uncertainty about how firms might fare in an environment of higher energy prices is likely to reduce investor confidence and increase the interest rates that firms must pay for capital. These two effects work to reduce investment spending and weaken economic activity.

3.5 Sorting through the Explanations

A consensus about why the effects of rising oil prices are so strong has not been reached. Recent research casts doubt on the idea that monetary policy is the primary factor amplifying the aggregate economic

effects of oil price shocks. No empirical research has attempted to sort through the other effects—adjustment costs, coordination problems, and financial stress. In fact, these effects are consistent with observed facts and are mutually reinforcing rather than mutually exclusive. All three effects may be at work.

4. ASYMMETRY OF THE RESPONSE

Prior to the 1980s, the large shocks in oil prices were increases. The 1980s brought the first major decrease in oil prices, and it gradually became evident that U.S. economic activity responded asymmetrically to oil price shocks. That is, rising oil prices hurt U.S. economic activity more than declining oil prices helped it. Although all but one of the post-World War II recessions followed sharp rises in oil prices, accelerated economic activity did not follow the sharp price declines of the 1980s and 1990s.

4.1 The Discovery of Asymmetry

Initially, the weak response of economic activity to oil price decreases was seen as a breakdown in the relationship between oil price movements and the economy, and researchers began to examine different oil price specifications in their empirical work to reestablish the relationship between oil price shocks and economic activity. Knut Mork found that when he grouped oil price changes into negative and positive oil price changes, oil price increases had more effect on economic activity than oil price decreases. Later research conducted by others found that oil price increases had a significant effect on economic activity, whereas oil price decreases did not. Similar asymmetry was also found at a more detailed industry level, and further research established that economic activity in seven industrialized countries responded asymmetrically to oil price movements.

Hamilton further refined the analysis by creating what he called a "net oil price." This measure of oil prices attempts to capture episodes of oil price increases that are outside normal market fluctuations. This measure reflects only those increases in oil prices that take the price higher than it has been in the past 12 months. (Technically, the net oil price series is the price of oil for all periods in which the price is higher than it has been during the past 12 months and zero in all other months.) Hamilton's research and subsequent research by others found a statistically significant and stable negative relationship between the net oil price series and output, whereas various series constructed to reflect episodes of sharp oil price declines do not seem to have any explanatory power. In a related vein, Steven Davis and John Haltiwanger constructed an oil price series that combined asymmetry with persistence and also found an asymmetric relationship between oil price shocks and economic activity.

4.2 Understanding Asymmetry

Classic supply-side effects cannot explain asymmetry. Operating through supply-side effects, reductions in oil prices should help output and productivity as increases in oil prices hurt economic output and productivity. Accordingly, economists have begun to explore the channels through which oil prices affect economic activity. Monetary policy, adjustment costs, coordination problems, uncertainty and financial stress, and asymmetry in the petroleum product markets have been offered as explanations. Of these, adjustment costs, coordination problems, and financial stress seem most consistent with the historical record.

4.3 Monetary Policy and Asymmetry

Monetary policy may contribute to an asymmetric response in aggregate economic activity to oil price shocks in two ways. Monetary policy may respond to oil price shocks asymmetrically. Another possibility is that nominal wages are sticky downward but not upward, and monetary policy is conducted in such a way that nominal GDP declines when oil prices are rising and nominal GDP increases when oil prices are declining.

When oil prices increase, real wages must decline to clear markets and restore equilibrium. If real wages do not decline, the economic displacement will be greater. When oil prices decline, real wages must increase to clear markets and restore equilibrium. If real wages do not increase, the gains in economic activity will be greater.

If nominal wages are sticky downward, an increase in unexpected inflation that is at least as great as the decline in real GDP is necessary to yield the necessary reduction in real wages. If the monetary authority maintains a neutral monetary policy (nominal GDP is unchanged), prices will increase as much as real GDP declines, and real wages will adjust sufficiently. If the monetary authority conducts policy such that nominal GDP declines, however, prices will increase by less than

the amount that real GDP declines, and because nominal wages are sticky downward, real wages will not adjust sufficiently to restore equilibrium.

Because nominal wages can adjust upward freely, however, unexpected disinflation is not required for adjustments in real wages to occur, and the conduct of monetary policy is not as important. If the monetary authority conducts policy such that nominal GDP increases, prices will increase by more than the real GDP declines, but because nominal wages adjust upward freely, real wages will increase enough to restore equilibrium. Consequently, a monetary policy that boosts nominal GDP when oil prices fall would have less effect on real activity than a policy that decreases nominal GDP when oil prices rise.

John Tatom provided early evidence that monetary policy responded asymmetrically to oil price shocks by showing that the economy responded symmetrically to oil price shocks if the monetary policy is taken into account. In a later contribution, Peter Ferderer showed that monetary policy cannot account for the asymmetry in the response of real activity to oil price shocks. Nathan Balke, Stephen Brown, and Mine Yücel found that the Federal Reserve's response to oil price shocks does not cause asymmetry in real economic activity. In their model, the asymmetry does not disappear, and is in fact enhanced, when monetary policy (as measured by either the federal funds rate or the federal funds rate plus expectations of the federal funds rate) is held constant.

4.4 Adjustment Costs and Coordination Problems

James Hamilton contributed the idea that adjustment costs may lead to an asymmetric response to changing oil prices. Rising oil prices retard economic activity directly, and decreasing oil prices stimulate economic activity directly. The costs of adjusting to changing oil prices retard economic activity whether oil prices are increasing or decreasing. Consequently, rising oil prices result in two negative effects on economic activity that reinforce each other. On the other hand, declining oil prices result in both positive and negative effects, which tend to be offsetting.

As described previously, these adjustment costs can be the result of the energy-to-output ratio being embedded in the capital or the result of sectoral imbalances. Coordination problems may reinforce adjustment costs whether oil prices are increasing or decreasing.

4.5 Uncertainty and Financial Stress

Uncertainty and financial stress may also contribute to an asymmetric response in aggregate economic activity to oil price shocks. As explained by Peter Ferderer, uncertainty about future oil prices is reflected in increased interest rates and reduced investment demand, which adversely affect aggregate economic activity. In addition, if the energy-to-output ratio is embedded in the capital stock, firms will find it increasingly desirable to postpone irreversible investment decisions until they are more certain about future oil prices.

Volatile oil prices contribute to oil price uncertainty and weaker economic activity, whether oil prices are increasing or decreasing. As is the case for adjustment costs, uncertainty and financial stress augment the basic supply-side effect when oil prices are increasing and offset the basic supply-side effect when oil prices are decreasing. The result is that aggregate economic activity responds to oil price shocks asymmetrically.

4.6 Asymmetry in Petroleum Product Prices

Petroleum product prices are another avenue through which the economy may respond asymmetrically to fluctuations in crude oil prices. A considerable body of research shows that petroleum product prices respond asymmetrically to fluctuations in oil prices, and most of the price volatility originates from crude oil prices rather than product prices. It is a relatively short leap to suggest that asymmetry in the response of product prices may account for the asymmetry between crude oil prices and aggregate economic activity.

Hillard Huntington found that the economy responds symmetrically to changes in petroleum product prices, but that petroleum product prices respond asymmetrically to crude oil prices. The consequence is an asymmetric relationship between crude oil prices and aggregate economic activity. These findings need further examination, but substantial research shows that the asymmetric response of aggregate economic activity to oil price shocks arises through channels that cannot be explained by the asymmetric response of product prices alone.

4.7 Sorting through the Explanations of Asymmetry

Although asymmetry is now fairly well accepted, relatively few studies have attempted to distinguish

empirically through what channels oil price shocks might yield an asymmetric response in aggregate economic activity. The available research seems to rule out the likelihood that asymmetry is the result of monetary policy, but the findings are consistent with explanations of asymmetry that rely on adjustment costs, coordination problems, or uncertainty and financial risk.

Prakash Loungani found that oil price shocks lead to a reallocation of labor across sectors, which increases the unemployment rate, whether oil prices are increasing or decreasing. These sectoral shifts are consistent with adjustment costs and coordination problems, but they do not preclude uncertainty and financial risk.

Steven Davis and John Haltiwanger assessed the channels through which oil price shocks affect economic activity. By examining job elimination and creation data, they found that allocative channels (e.g., changes in the desired distributions of labor and capital) contribute to the asymmetric response of aggregate economic activity to oil price shocks. Again, sectoral shifts are consistent with adjustment costs and coordination problems, but they do not preclude uncertainty and financial risk. In addition, Davis and Haltiwanger showed that the allocative effects are relatively strong and reinforce the basic supply-side effect when oil prices are increasing and cancel the basic supply-side effect when prices are declining.

Balke, Brown, and Yücel found that monetary policy responds asymmetrically because the economy has responded asymmetrically to oil price shocks and that monetary policy cannot account for the asymmetry. They also found that output and interest rates respond asymmetrically to oil price shocks, with the shocks transmitted through asymmetric movements in market interest rates. Such interest rate movements are consistent with several explanations for asymmetry. Asymmetric movements in the interest rates may be indicative of the increased uncertainty and financial risk that result from oil price shocks. Alternatively, movements in market interest rates may be a reflection of expectations that oil price shocks will necessitate costly adjustments in the future.

5. A WEAKENING RESPONSE

By the mid-1990s, the quantitative relationship between oil prices and economic activity seemed fairly robust and reasonably well understood, even if the exact channels through which the effects operated were not known with certainty. During the latter half of the 1990s, however, the relationship seemed to weaken. In the late 1990s and early 2000s, rising oil prices had less effect on economic activity than previous research suggested might have been expected.

Rising oil prices led to increases in the unemployment rate from the early 1970s through the early 1990s, and unemployment declined with oil prices from 1982 through 1990 and in the late 1990s (Fig. 2). Nonetheless, the relationship began to weaken in the late 1990s. Although oil prices were relatively strong, the unemployment rate continued to decline. One explanation is that high world oil prices were a result of a strong world economy, and strong demand rather than a supply shock accounted for rising oil prices.

The data also show a weaker relationship between rising oil prices and core inflation in the late 1990s (Fig. 3). Mark Hooker reevaluated the oil price–inflation relationship and found that since approximately 1980, oil price changes have not seemed to affect core measures of inflation. Prior to 1980, however, oil price shocks contributed substantially to core inflation in the United States. One possibility is that U.S. monetary policy in the Volcker and Greenspan eras was significantly less accommodative to oil price shocks than it had been under their predecessors, and thus monetary policy no longer contributed to high inflation.

Nonetheless, the relationship between oil prices and interest rates was relatively unchanged through the 1990s. Rising oil prices led to higher interest rates, which is the expected consequence of supply shocks that have greater short-term effects than

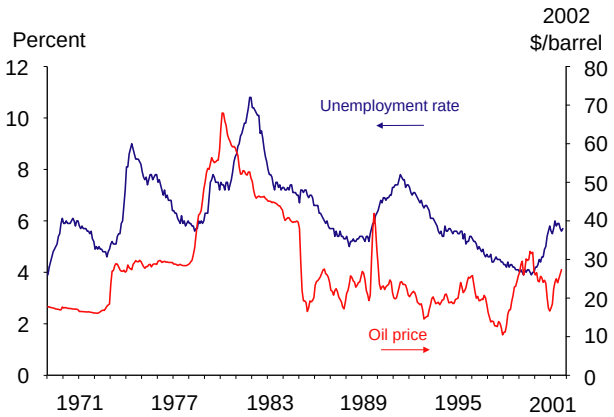


FIGURE 2 Real oil prices and U.S. unemployment. Unemployment declined with oil prices from 1982 to 1990 and in the late 1990s, but the relationship began to weaken in the late 1990s.

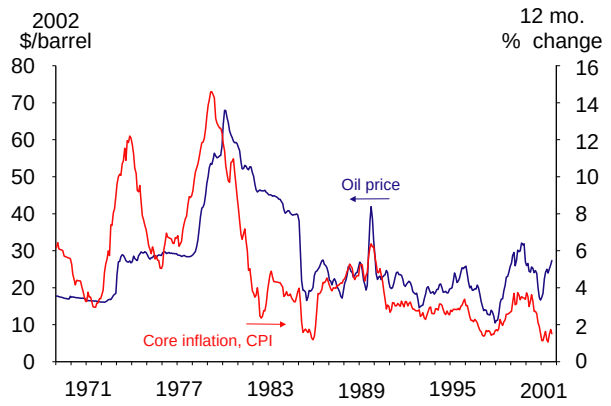


FIGURE 3 Real oil prices and core inflation. Prior to 1980, real oil prices and core inflation were more closely related.

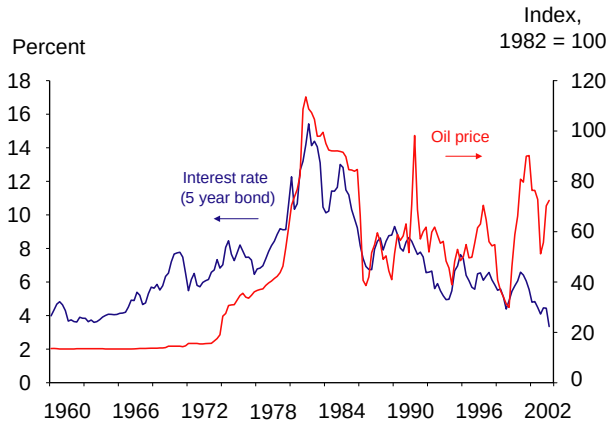


FIGURE 4 Real oil prices and U.S. interest rates. Oil prices and market interest rates moved closely together until mid-2000.

long-term effects (Fig. 4). Brown and Yücel have shown that some of the increases in the U.S. federal funds rate that occurred in 1999 and 2000 may have been part of a general increase in market interest rates that resulted from higher oil prices. Since approximately mid-2000, however, interest rates have not increased with oil prices. This changing relationship may be further evidence of an economy that is becoming less sensitive to oil price shocks.

5.1 Factors Contributing to a Weakening Relationship

Several factors have likely contributed to a weakening relationship between oil prices and economic activity during the past three decades, including a reduced energy-to-GDP ratio, the fact that oil price increases were the result of increased demand rather

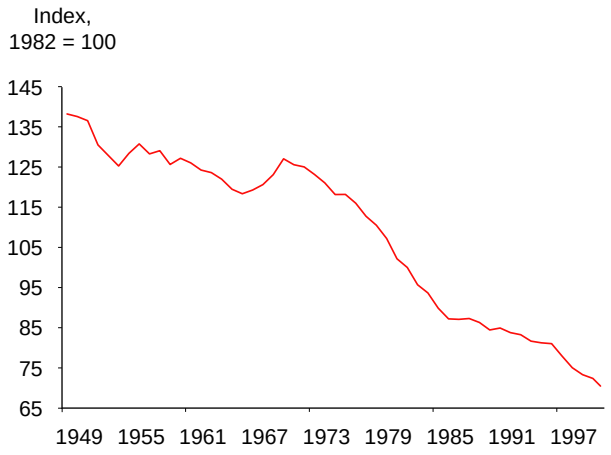


FIGURE 5 Decrease in U.S. energy-to-GDP ratio. The energy-to-GDP ratio has decreased since the end of World War II. Higher energy prices contributed to strong reductions in the energy-to-GDP ratio from the mid-1970s to the mid-1980s.

than oil supply shocks, and prior experience with oil price shocks. In addition, the 1990s boom was marked by strong productivity gains that may have simply obscured the relationship between oil prices and aggregate economic activity.

5.2 A Reduced Energy-to-GDP Ratio

Although it represents a continuing trend, the energy consumption-to-GDP ratio declined from the 1970s to the early 2000s (Fig. 5). On the basis of this decline, Brown and Yücel estimate that the U.S. economy may have been approximately one-third less sensitive to oil price fluctuations in 2000 than it was in the early 1980s and approximately half as sensitive as it was in the early 1970s.

5.3 Increased Demand Rather Than a Supply Shock

Another factor that could have contributed to a weakening relationship between oil prices and economic activity was the fact that rising oil prices in the late 1990s were largely the result of economic expansion. Price increases that are the result of increased demand rather than supply shocks may be less disruptive to economic activity.

Due to increased energy efficiency, U.S. oil consumption increased only moderately during the 1990s, but oil consumption in the industrialized nations increased steadily (Fig. 6). World oil consumption was further boosted by the dramatic gains in oil

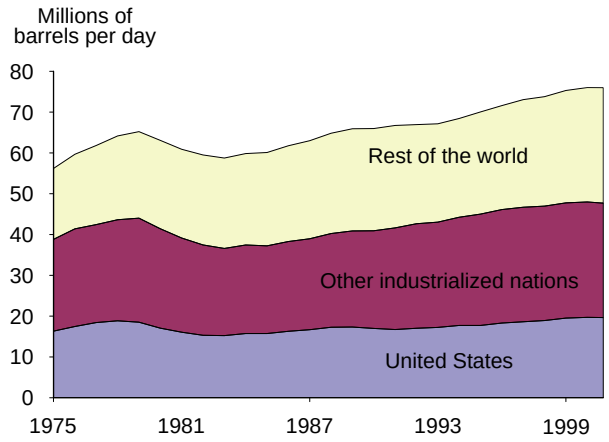


FIGURE 6 Increase in world oil consumption. United States oil consumption increased moderately in the 1990s. Oil consumption in the industrialized countries combined increased steadily during the 1990s. The largest increase in oil consumption occurred in the newly industrializing economies, such as China and Korea.

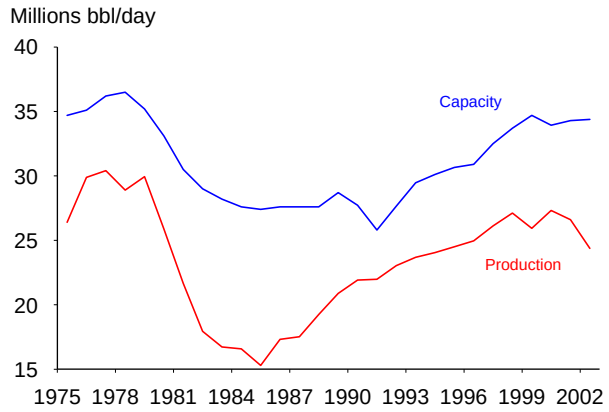


FIGURE 7 Production and capacity of the Organization of Petroleum Exporting Countries (OPEC). OPEC operated fairly close to capacity in the 1990s.

consumption outside countries belonging to the Organization for Economic Cooperation and Development, with the strongest gains in consumption occurring in Asian countries, such as China and Korea.

In addition, world capacity to produce oil did not keep pace with growing consumption throughout much of the 1990s. Consequently, the gap between capacity and production of the Organization of Petroleum Exporting Countries was relatively small throughout most of the 1990s (Fig. 7). With demand growing faster than supply, oil prices increased in the latter part of the decade, although the real increase was relatively small by 1980s standards (Fig. 8). In 2000, world oil demand increased sharply (Fig. 9), which led to greater consumption and an increase in

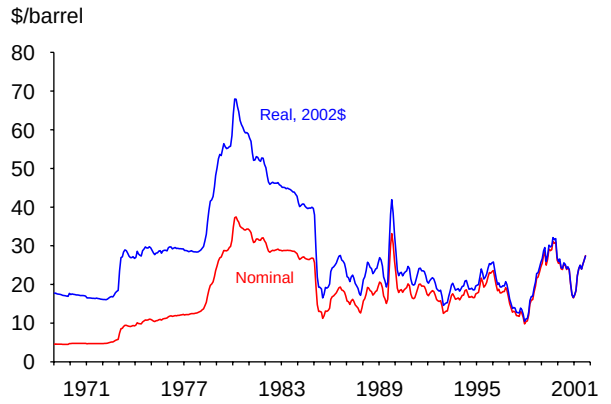


FIGURE 8 Real and nominal oil prices. Real oil prices are adjusted for inflation, whereas nominal oil prices are not. The real oil price indicates that the oil price gains in 2000 were moderate by historical standards.

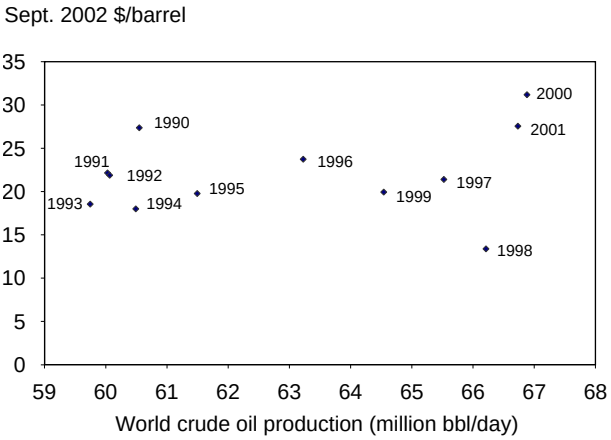


FIGURE 9 Oil demand increased sharply in 2000. The plotted points show the market prices and quantities of oil for the years 1991–2001. Increases in quantity and price are indicative of increased demand.

prices. Because the increase in oil prices was the result of strong demand due to economic expansion, the usual decline in economic activity did not follow.

5.4 Prior Experience with Energy Price Shocks

Prior experience with oil price shocks may have also contributed to the muted response of U.S. economic activity to oil price shocks. Such experience may have resulted in reduced adjustment costs, coordination problems, and uncertainty and financial stress. In addition, monetary authorities have increased experience with oil price shocks, which may have resulted in reduced inflationary pressure caused by oil price shocks.

6. CONCLUSIONS AND IMPLICATIONS FOR POLICY

Economic research provides considerable evidence that rising oil prices contribute to real GDP losses. Economic theory suggests that a number of channels may account for this phenomenon. The most basic is a classic supply-side effect in which rising oil prices are indicative of the reduced availability of a basic input to production. Other channels include reduced aggregate demand and contractionary monetary policy. Of these channels, the empirical evidence is most consistent with a classic supply-side effect. This effect best explains the reduced aggregate output, increased price level, and higher interest rates.

Given the current understanding of how oil price shocks affect economic activity, analysts are relatively well positioned to prescribe the proper fiscal and monetary policy responses to oil price shocks. Also, given that rising oil prices reduce potential GDP and the reductions in aggregate demand are likely to be negligible, a fiscal policy response seems unnecessary. Regarding monetary policy, a neutral response to rising oil prices, which consists of holding nominal GDP constant, will neither aggravate the GDP losses nor offset them. Such a policy will also lead to an increase in the price level that is the same as the loss in real GDP.

In addition to taking a neutral stance, monetary policy can shape the aggregate effects of an oil price shock. If the monetary authority is willing to accept higher inflation, it can temporarily boost real GDP through an expansionary policy. If the monetary authority wants to lessen the inflationary consequences of rising oil prices, it can tighten policy, which will temporarily aggravate the losses in real GDP.

Economists have variously suggested that monetary policy, adjustment costs, coordination problems, and increased uncertainty and financial stress account for an amplification of the basic supply-side effect when oil prices are increasing and a lessening of the aggregate response when oil prices are decreasing. Although research indicates that the conduct of U.S. monetary policy has not resulted in such effects, monetary policy could have such effects, and the issue is not completely resolved. Researchers have provided evidence that adjustment costs, coordination problems, or increased uncertainty and financial stress amplify basic supply-side effects when oil prices are increasing. They have not been able to reliably distinguish between these effects.

Because economic research has not been able to distinguish between the contributions of adjustment costs, coordination problems, and uncertainty and financial stress, analysts are less able to prescribe the best course of action for energy policy. In particular, research has not determined whether the private sector is capable of providing the optimal level of insurance against price shocks. Given the possibility of coordination problems across industries and the asymmetric response of aggregate economic activity to oil price shocks, an energy policy that counters movements in international oil prices seems justified. Nonetheless, considerably more research must be conducted before economists can provide sound guidance regarding the extent to which policymakers should go to reduce an economy's vulnerability to oil price shocks.

SEE ALSO THE FOLLOWING ARTICLES

Economics of Energy Demand • *Economics of Energy Supply* • *Energy Futures and Options* • *Inflation and Energy Prices* • *Innovation and Energy Prices* • *Oil Price Volatility* • *Prices of Energy, History of* • *Stock Markets and Energy Prices*

Further Reading

Balke, N. S., Brown, S. P. A., and Yücel, M. K. (2002). Oil price shocks and the U.S. economy: Where does the asymmetry originate? *Energy J.* **23**(3), 27–52.

Bernanke, B. S., Gertler, M., and Watson, M. (1997). Systematic monetary policy and the effects of oil price shocks. *Brookings Papers Econ. Activity* **1997**(1), 91–157.

Brown, S. P. A., and Yücel, M. K. (1999). Oil prices and U.S. aggregate economic activity: A question of neutrality. *Econ. Financial Rev.* [Federal Reserve Bank of Dallas], Second Quarter, 16–23.

Brown, S. P. A., and Yücel, M. K. (2002, Second quarter). Energy prices and aggregate economic activity: An interpretative survey. *Q. Rev. Econ. Finance* 42(2).

Davis, S. J., and Haltiwanger, J. (2001). Sectoral job creation and destruction responses to oil price changes and other shocks. *J. Monetary Econ.* **48**, 463–512.

Ferderer, J. P. (1996). Oil price volatility and the macroeconomy: A solution to the asymmetry puzzle. *J. Macroecon.* **18**, 1–16.

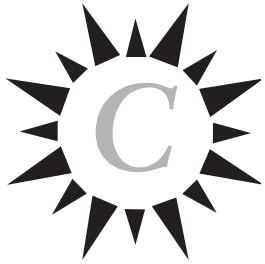
Hamilton, J. D. (1983). Oil and the macroeconomy since World War II. *J. Political Econ.* **91**, 228–248.

Hamilton, J. D., and Herrera, A. M. (2004). Oil shocks and aggregate macroeconomic behavior: The role of monetary policy. *J. Money Credit Banking* **35**(6).

Huntington, H. G. (2003). Oil security as a macroeconomic externality. *Energy Econ.* **25**, 119–136.

Mork, K. A. (1994). Business cycles and the oil market. *Energy J.* **15**, 15–38.

Carbon Capture and Storage from Fossil Fuel Use



HOWARD HERZOG and DAN GOLOMB
Massachusetts Institute of Technology
Cambridge, Massachusetts, United States

1. Introduction
2. Carbon Sources
3. Capture Processes
4. CO_2 Storage
5. Economics
6. Alternate Approaches

Glossary

carbon capture The separation and entrapment of CO_2 from large stationary sources.

carbon sequestration The capture and secure storage of carbon that would otherwise be emitted to or remain in the atmosphere.

carbon sources Large stationary sources of CO_2, e.g., fossil-fueled power plants, cement manufacturing and ammonia production plants, iron and nonferrous metal smelters, industrial boilers, refineries, and natural gas wells.

CO_2 storage The injection of CO_2 into geologic or oceanic reservoirs for timescales of centuries or longer.

One of the approaches for mitigating potential global climate change due to anthropogenic emissions of CO_2 and other greenhouse gases is to capture CO_2 from fossil fuel-using sources, and to store it in geologic or oceanic reservoirs. The capture technologies are described, and their efficiencies, cost, and energy penalties are estimated. Storage capacities and effectiveness are estimated, as well as transportation costs and possible environmental impacts.

1. INTRODUCTION

Carbon sequestration can be defined as the capture and secure storage of carbon that would otherwise be emitted to, or remain, in the atmosphere. The focus of this article is the removal of CO_2 directly from industrial or utility plants and subsequently storing it in secure reservoirs. This is called carbon capture and storage (CCS). The rationale for carbon capture and storage is to enable the use of fossil fuels while reducing the emissions of CO_2 into the atmosphere, and thereby mitigating global climate change. The storage period should exceed the estimated peak periods of fossil fuel exploitation, so that if CO_2 reemerges into the atmosphere, it should occur after the predicted peak in atmospheric CO_2 concentrations. Removing CO_2 from the atmosphere by increasing its uptake in soils and vegetation (e.g., afforestation) or in the ocean (e.g., iron fertilization), a form of carbon sequestration sometimes referred to as enhancing natural sinks, is only addressed briefly.

At present, fossil fuels are the dominant source satisfying the global primary energy demand, and will likely remain so for the rest of the century. Fossil fuels supply over 85% of all primary energy; nuclear power, hydroelectricity, and renewable energy (commercial biomass, geothermal, wind, and solar energy) supply the remaining 15%. Currently, nonhydroelectric renewable energy satisfies less than 1% of the global energy demand. Although great efforts and investments are made by many nations to increase the share of renewable energy to satisfy the primary energy demand and to foster conservation and efficiency improvements of fossil fuel usage, addressing climate change concerns during the coming decades will likely require significantly increasing the contributions from carbon capture and storage.

2. CARBON SOURCES

Pathways for carbon capture derive from three potential sources (see Fig. 1). Several industrial

 © 2004 Elsevier Inc. All rights reserved.

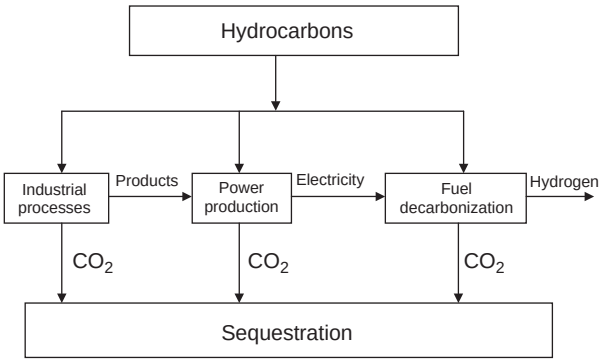


FIGURE 1 Sources of CO_2 for sequestration, as industrial by-product, captured from power plants, or as by-product of future fuel decarbonization plants.

processes produce highly concentrated streams of CO_2 as a by-product. Although limited in quantity, these by-products make a good capture target, because the captured CO_2 is integral to the total production process, resulting in relatively low incremental capture costs. For example, natural gas ensuing from wells often contains a significant fraction of CO_2 that can be captured and stored. Other industrial processes that lend themselves to carbon capture are ammonia manufacturing, fermentation processes, and hydrogen production (e.g., in oil refining).

By far the largest potential sources of CO_2 today are fossil-fueled power production plants. Power plants emit more than one-third of the CO_2 emissions worldwide. Power plants are usually built in large centralized units, typically delivering 500–1000 MW of electrical power. A 1000-MW pulverized coal-fired power plant emits between 6 and 8 Mt/yr of CO_2, an oil-fired single-cycle power plant emits about two-thirds of that, and a natural gas combined-cycle power plant emits about one-half of that.

Future opportunities for CO_2 capture may also arise from decarbonization, i.e., producing hydrogen fuels from carbon-rich feedstocks, such as natural gas, coal, and biomass. The CO_2 by-product will be relatively pure and the incremental costs of carbon capture will be relatively low. The hydrogen can be used in fuel cells and other hydrogen fuel-based technologies, but there are major costs involved in developing a mass market and infrastructure for these new fuels.

3. CAPTURE PROCESSES

CO_2 capture processes from power production fall into three general categories: (1) flue gas separation, (2) oxyfuel combustion in power plants, and (3) precombustion separation. Each of these technologies carries both an energy and an economic penalty. The efficiencies and economics of several technologies are discussed in Section 5.

3.1 Flue Gas Separation

Currently, flue gas separation and CO_2 capture are practiced at about a dozen facilities worldwide. The capture process is based on chemical absorption. The captured CO_2 is used for various industrial and commercial processes, e.g., urea production, foam blowing, carbonation of beverages, and dry ice production. Because the captured CO_2 is used as a commercial commodity, the absorption process, although expensive, is profitable because of the price realized for the commercial CO_2.

Chemical absorption refers to a process in which a gas (in this case, CO_2) is absorbed in a liquid solvent by formation of a chemically bonded compound. When used in a power plant to capture CO_2, the flue gas is contacted with the solvent in a packed absorber column, where the solvent preferentially removes the CO_2 from the flue gas. Afterward, the solvent passes through a regenerator unit, where the absorbed CO_2 is stripped from the solvent by counterflowing steam at 100–120°C. Water vapor is condensed, leaving a highly concentrated (over 99%) CO_2 stream, which may be compressed for commercial utilization or storage. The lean solvent is cooled to 40–65°C and is recycled into the absorption column. The most commonly used absorbent for CO_2 absorption is monoethanolamine (MEA). The fundamental reaction for this process is

$$C_2H_4OHNH_2 + H_2O + CO_2 \leftrightarrow C_2H_4OHNH_3^+ + HCO_3^-$$

During the absorption process, the reaction proceeds from left to right; during regeneration, the reaction proceeds from right to left. Cooling and heating of the solvent, pumping, and compression require power input from the power plant thermal cycle, derating the thermal efficiency (heat rate) of the power plant. A schematic of a chemical absorption process for power plant flue gas is depicted in Fig. 2.

In order to reduce the capital and energy costs and the size of the absorption and regenerator (stripper) columns, new processes are being developed. One example is the membrane-absorption process, whereby a microporous membrane made of poly(tetrafluoroethylene) separates the flue gas from the

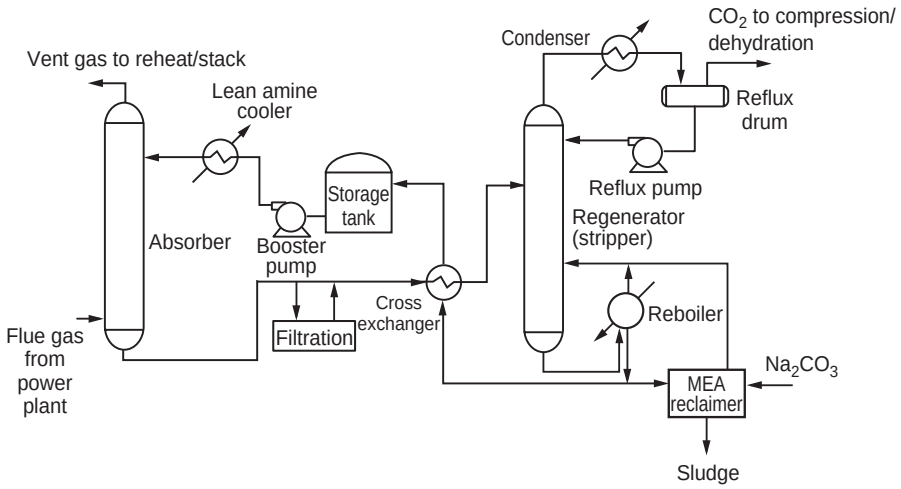


FIGURE 2 Process flow diagram for the amine separation process. MEA, monoethanolamine.

solvent. The membrane allows for greater contacting area within a given volume, but by itself the membrane does not perform the separation of CO_2 from the rest of the flue gases. It is the solvent that selectively absorbs CO_2. The use of a gas membrane has several advantages: (a) high packing density, (b) high flexibility with respect to flow rates and solvent selection, (c) no foaming, channeling, entrainment, and flooding, all common problems in packed absorption towers, (d) ease of transport of the unit, e.g., offshore, and (e) significant savings in weight.

It is possible to design a once-through scrubbing process (i.e., no regeneration step). For example, the CO_2 could be scrubbed from flue gas with seawater, returning the whole mixture to the ocean for storage. However, to date these approaches are not as practical compared to those using a regenerable solvent. In the seawater scrubbing example, the large volumes of water that are required result in large pressure drops in the pipes and absorber.

Other processes have been considered to capture CO_2 from power plant and industrial boiler flue gases, e.g., membrane separation, cryogenic fractionation, and adsorption using molecular sieves. Generally, these processes are less energy efficient and more expensive compared to the absorption methods.

3.2 Oxyfuel Combustion

When a fossil fuel (coal, oil, and natural gas) is combusted in air, the fraction of CO_2 in the flue gas ranges from 3 to 15%, depending on the carbon content of the fuel and the amount of excess air necessary for the combustion process. The separation of CO_2 from the rest of the flue gases (mostly N_2) by chemical or physical means is capital and energy intensive. An alternative is to burn the fossil fuel in pure or enriched oxygen. In such a fashion, the flue gas will contain mostly CO_2 and H_2O. A part of the flue gas needs to be recycled into the combustion chamber in order to control the flame temperature. From the nonrecycled flue gas, water vapor can be readily condensed, and the CO_2 can be compressed and piped directly to the storage site.

Of course, the separation process has now shifted from the flue gas to the intake air: oxygen must be separated from nitrogen in the air. The air separation unit (ASU) alone may consume about 15% of a power plant's electric output, requiring a commensurate increase of fossil fuel to be consumed for achieving the rated electric output of the plant. In the ASU, air is separated into liquid oxygen, gaseous nitrogen, argon, and other minor ingredients of air. The latter are marketable by-products of the oxyfuel plant. Pilot scale studies indicate that the oxyfuel method of capturing CO_2 can be retrofitted to existing pulverized coal (PC) plants.

3.3 Precombustion Capture

Capturing CO_2 before combustion offers some advantages. First, CO_2 is not yet diluted by the combustion air. Second, the CO_2-containing stream is usually at elevated pressure. Therefore, more efficient separation methods can be applied, e.g., using pressure-swing absorption in physical solvents, such as methanol or poly (ethylene glycol) (commercial brands are called Rectisol and Selexol). Precombustion capture is usually applied in integrated coal gasification combined cycle power plants. This

process includes gasifying the coal to produce a synthesis gas composed of CO and H_2, reacting the CO with water (water–gas shift reaction) to produce CO_2 and H_2, capturing the CO_2, and sending the H_2 to a turbine to produce electricity. Because the primary fuel sent to the gas turbine is now hydrogen, some can be bled off as a fuel for separate use, such as in hydrogen fuel cells to be used in transportation vehicles. One of the biggest barriers to this pathway is that currently electricity generation is cheaper in pulverized coal power plants than in integrated coal gasification combined cycle plants. The precombustion process could be utilized when natural gas is the primary fuel. Here, a synthesis gas is formed by reacting natural gas with steam to produce CO_2 and H_2. However, it is unproved whether precombustion capture is preferable to the standard postcombustion capture for the case of using natural gas.

Worldwide, gasification facilities currently exist; most do not generate electricity, however, but produce synthesis gas and various other by-products of coal gasification. In these facilities, after the gasification stage, CO_2 is separated from the other gases, such as methane, hydrogen, or a mix of carbon monoxide and hydrogen. The synthesis gas or hydrogen is used as a fuel or for chemical raw material, e.g., for liquid-fuel manufacturing or ammonia synthesis. The CO_2 can also be used as a chemical raw material for dry ice manufacturing, carbonated beverages, and enhanced oil recovery (EOR). For example, the Great Plains Synfuel Plant, near Beulah, North Dakota, gasifies 16,326 metric tons per day of lignite coal into 3.5 million standard cubic meters per day of combustible syngas, and close to 7 million standard cubic meters of CO_2. A part of the CO_2 is captured by a physical solvent based on methanol. The captured CO_2 is compressed and 2.7 million standard cubic meters per day is piped over a 325-km distance to the Weyburn, Saskatchewan, oil field, where the CO_2 is used for enhanced oil recovery.

4. CO_2 STORAGE

Following the capture process, CO_2 needs to be stored, so that it will not be emitted into the atmosphere. Several key criteria must be applied to the storage method: (a) the storage period should be prolonged, preferably hundreds to thousands of years, (b) the cost of storage, including the cost of transportation from the source to the storage site, should be minimized, (c) the risk of accidents should be eliminated, (d) the environmental impact should be minimal, and (e) the storage method should not violate any national or international laws and regulations.

Storage media include geologic sinks and the deep ocean. Geologic storage options include deep saline formations (subterranean and subseabed), depleted oil and gas reservoirs, formations for enhanced oil recovery operations, and unminable coal seams. Deep ocean storage approaches include direct injection of liquid carbon dioxide into the water column at intermediate depths (1000–3000 m), or at depths greater than 3000 m, where liquid CO_2 becomes heavier than seawater, so CO_2 would drop to the ocean bottom and form a "CO_2 lake." In addition, other storage approaches are proposed, such as enhanced uptake of CO_2 by terrestrial and oceanic biota, and mineral weathering. The latter approaches refer to the uptake of CO_2 from the atmosphere, not CO_2 that has been captured from emission sources. Finally, captured CO_2 can be used as a raw material by the chemical industry. However, the prospective amounts of CO_2 that can be utilized are but a very small fraction of CO_2 emissions from anthropogenic sources. Table I lists the estimated worldwide capacities for CO_2 storage in the various media. As a comparison to the storage capacities, note that current global anthropogenic emissions amount to close to 7 Gt C per year (1 Gt C = 1 billion metric tons of carbon equivalent = 3.7 Gt CO_2).

4.1 Geologic Storage

Geological sinks for CO_2 include depleted oil and gas reservoirs, enhanced oil recovery operation formations, unminable coal seams, and deep saline

TABLE I

The Worldwide Capacity of Potential CO_2 Storage Reservoirs[a]

Sequestration option	Worldwide capacity (Gt C)[b]
Ocean	1000–10,000 +
Deep saline formations	100–10,000
Depleted oil and gas reservoirs	100–1000
Coal seams	10–1000
Terrestrial	10–100
Utilization	Currently < 0.1 Gt C/yr

[a] Ocean and land-based sites together contain an enormous capacity for storage of CO_2. The world's oceans have by far the largest capacity for carbon storage. Worldwide total anthropogenic carbon emissions are ~7 Gt C per year (1 Gt C = 1 billion metric tons of carbon equivalent).

[b] Orders of magnitude estimates.

formations. Together, these can hold hundreds to thousands of gigatons of carbon (Gt C), and the technology to inject CO_2 into the ground is well established. CO_2 is stored in geologic formations by a number of different trapping mechanisms, with the exact mechanism depending on the formation type.

4.1.1 Depleted Oil and Gas Reservoirs

Though a relatively new idea in the context of climate change mitigation, injecting CO_2 into depleted oil and gas fields has been practiced for many years. The major purpose of these injections was to dispose of "acid gas," a mixture of CO_2, H_2S, and other by-products of oil and gas exploitation and refining. In 2001, nearly 200 million cubic meters of acid gas was injected into formations across Alberta and British Columbia at more than 30 different locations. Acid gas injection has become a popular alternative to sulfur recovery and acid gas flaring, particularly in Western Canada. Essentially, acid gas injection schemes remove CO_2 and H_2S from the produced oil or gas stream, compress and transport the gases via pipeline to an injection well, and reinject the gases into a different formation for disposal. Proponents of acid gas injection claim that these schemes result in less environmental impact than alternatives for processing and disposing of unwanted gases. In most of these schemes, CO_2 represents the largest component of the acid gas, typically up to 90% of the total volume injected for disposal. Successful acid gas injection requires a nearby reservoir with sufficient porosity, amply isolated from producing reservoirs and water zones. Historically, depleted and producing reservoirs have proved to be extremely reliable containers of both hydrocarbons and acid gases over time.

4.1.2 Enhanced Oil Recovery Operations

Carbon dioxide injection into geological formations for enhanced oil recovery (EOR) is a mature technology. In 2000, 84 commercial or research-level CO_2 EOR projects were operational worldwide. The United States, the technology leader, accounts for 72 of the 84 projects, most of which are located in the Permian Basin. Combined, these projects produced 200,772 barrels (bbl) of oil per day, a small but significant fraction (0.3%) of the 67.2 million bbl per day total of worldwide oil production that year. Outside the United States and Canada, CO_2 floods have been implemented in Hungary, Turkey, and Trinidad.

In most CO_2 EOR projects, much of the CO_2 injected into the oil reservoir is only temporarily stored. This is because the decommissioning of an EOR project usually involves the "blowing down" of the reservoir pressure to maximize oil recovery. This blowing down results in CO_2 being released, with a small but significant amount of the injected CO_2 remaining dissolved in the immobile oil. The Weyburn Field in southeastern Saskatchewan, Canada, is the only CO_2 EOR project to date that has been monitored specifically to understand CO_2 storage. In the case of the Weyburn Field, no blowdown phase is planned, thereby allowing for permanent CO_2 storage. Over the anticipated 25-year life of the project, it is expected that the injection of some 18 million tons of CO_2 from the Dakota Gasification Facility in North Dakota will produce around 130 million BBL of enhanced oil. This has been calculated to be equivalent to approximately 14 million tons of CO_2 being prevented from reaching the atmosphere, including the CO_2 emissions from electricity generation that is required for the whole EOR operation.

4.1.3 Unminable Coal Seams

Abandoned or uneconomic coal seams are another potential storage site. CO_2 diffuses through the pore structure of coal and is physically adsorbed to it. This process is similar to the way in which activated carbon removes impurities from air or water. The exposed coal surface has a preferred affinity for adsorption of CO_2 rather than for methane, with a ratio of 2:1. Thus, CO_2 can be used to enhance the recovery of coal bed methane (CBM). In some cases, this can be very cost-effective or even cost free, because the additional methane removal can offset the cost of the CO_2 storage operations. CBM production has become an increasingly important component of natural gas supply in the United States during the past decade. In 2000, approximately 40 billion standard cubic meters (scm) of CBM was produced, accounting for about 7% of the nation's total natural gas production. The most significant CBM production, some 85% of the total, occurs in the San Juan basin of southern Colorado and northern New Mexico. Another 10% is produced in the Black Warrior basin of Alabama, and the remaining 5% comes from rapidly developing Rocky Mountain coal basins, namely, the Uinta basin in Utah, the Raton basin in Colorado and New Mexico, and the Powder River basin in Wyoming. Significant potential for CBM exists worldwide. A number of coal basins in Australia, Russia, China, India, Indonesia, and other countries have also been identified as having a large CBM potential. The total

worldwide potential for CBM is estimated at around 2 trillion scm, with about 7.1 billion tons of associated CO_2 storage potential.

4.1.4 Deep Saline Formations

Deep saline formations, both subterranean and subseabed, may have the greatest CO_2 storage potential. These reservoirs, which are the most widespread and have the largest volumes, are very distinct from the more familiar reservoirs used for freshwater supplies. Research is currently underway in trying to understand what percentage of these deep saline formations could be suitable CO_2 storage sites.

The density of CO_2 depends on the depth of injection, which determines the ambient temperature and pressure. The CO_2 must be injected below 800 m, so that it is in a dense phase (either liquid or supercritical). When injected at these depths, the specific gravity of CO_2 ranges from 0.5 to 0.9, which is lower than that of the ambient aquifer brine. Therefore, CO_2 will naturally rise to the top of the reservoir, and a trap is needed to ensure that it does not reach the surface. Geologic traps overlying the aquifer immobilize the CO_2. In the case of aquifers with no distinct geologic traps, an impermeable cap rock above the underground reservoir is needed. This forces the CO_2 to be entrained in the groundwater flow and is known as hydrodynamic trapping. Two other very important trapping mechanisms are solubility and mineral trapping. Solubility and mineral trapping involve the dissolution of CO_2 into fluids, and the reaction of CO_2 with minerals present in the host formation to form stable, solid compounds such as carbonates. If the flow path is long enough, the CO_2 might all dissolve or become fixed by mineral reactions before it reaches the basin margin, essentially becoming permanently trapped in the reservoir.

The first, and to date only, commercial-scale project dedicated to geologic CO_2 storage is in operation at the Sleipner West gas field, operated by Statoil, located in the North Sea about 250 km off the coast of Norway. The natural gas produced at the field has a CO_2 content of about 9%. In order to meet commercial specifications, the CO_2 content must be reduced to 2.5%. At Sleipner, the CO_2 is compressed and injected via a single well into the Utsira Formation, a 250-m-thick aquifer located at a depth of 800 m below the seabed. About 1 million metric tons of CO_2 have been stored annually at Sleipner since October 1996, equivalent to about 3% of Norway's total annual CO_2 emissions. A total of 20 Mt of CO_2 is expected to be stored over the lifetime of the project. One motivation for doing this was the Norwegian offshore carbon tax, which was in 1996 about $50 (USD) per metric ton of CO_2 (the tax was lowered to $38 per ton on January 1, 2000). The incremental investment cost for storage was about $80 million. Solely on the basis of carbon tax savings, the investment was paid back in about 1.5 years. This contrasts to most gas fields worldwide, where the separated CO_2 is simply vented into the atmosphere.

Statoil is planning a second storage project involving about 0.7 Mt per year of CO_2 produced at the Snohvit gas field in the Barents Sea off northern Norway; injection will be into a deep subsea formation.

4.1.5 Environmental and Safety Concerns

Fundamentally, a geologic storage system can be broken down into two general subsystems, namely, operational and *in situ*. The operational subsystem is composed of the more familiar components of CO_2 capture, transportation, and injection, which have been successfully deployed in the previously discussed applications. Once CO_2 is injected in the reservoir, it enters an *in situ* subsystem in which the control of CO_2 is transferred to the forces of nature. Years of technological innovation and experience have given us the tools and expertise to handle and control CO_2 in the operational subsystem with adequate certainty and safety; however, that same level of expertise and understanding is largely absent once the CO_2 enters the storage reservoir. Direct environmental and human health risks are of utmost concern. As such, researchers are now conducting studies to evaluate the likelihood and potential impacts associated with leaks, slow migration and accumulation, and induced seismicity.

4.2 Ocean Storage

By far, the ocean represents the largest potential sink for anthropogenic CO_2. It already contains an estimated 40,000 Gt C (billion metric tons of carbon) compared with only 750 Gt C in the atmosphere and 2200 Gt C in the terrestrial biosphere. Apart from the surface layer, deep ocean water is unsaturated with respect to CO_2. It is estimated that if all the anthropogenic CO_2 that would double the atmospheric concentration were injected into the deep ocean, it would change the ocean carbon concentration by less than 2%, and lower its pH by less than 0.15 units. Furthermore, the deep waters of the ocean are not hermetically separated from the atmosphere.

Eventually, on a timescale of 1000 years, over 80% of today's anthropogenic emissions of CO_2 will be transferred to the ocean. Discharging CO_2 directly to the ocean would accelerate this ongoing but slow natural process and would reduce both peak atmospheric CO_2 concentrations and their rate of increase.

In order to understand ocean storage of CO_2, some properties of CO_2 and seawater need to be elucidated. For efficiency and economics of transport, CO_2 would be discharged in its liquid phase. If discharged above about 500 m depth (that is, at a hydrostatic pressure less than 50 atm), liquid CO_2 would immediately flash into a vapor and bubble up back into the atmosphere. Between 500 and about 3000 m, liquid CO_2 is less dense than seawater, therefore it would ascend by buoyancy. It has been shown by hydrodynamic modeling that if liquid CO_2 were released in these depths through a diffuser such that the bulk liquid breaks up into droplets less than about 1 cm in diameter, the ascending droplets would completely dissolve before rising 100 m. Because of the higher compressibility of CO_2 compared to seawater, below about 3000 m liquid CO_2 becomes denser than seawater, and if released there, would descend to greater depths. When liquid CO_2 is in contact with water at temperatures less than 10°C and pressures greater than 44.4 atm, a solid hydrate is formed in which a CO_2 molecule occupies the center of a cage surrounded by water molecules. For droplets injected into seawater, only a thin film of hydrate forms around the droplets.

There are two primary methods under serious consideration for injecting CO_2 into the ocean. One involves dissolution of CO_2 at middepths (1500–3000 m) by injecting it from a bottom-mounted pipe from shore or from a pipe towed by a moving CO_2 tanker. The other is to inject CO_2 below 3000 m, where it will form a "deep lake." Benefits of the dissolution method are that it relies on commercially available technology and the resulting plumes can be made to have high dilution to minimize any local environmental impacts due to increased CO_2 concentration or reduced pH. The concept of a CO_2 lake is based on a desire to minimize leakage to the atmosphere. Research is also looking at an alternate option of injecting the CO_2 in the form of bicarbonate ions in solution. For example, seawater could be brought into contact with flue gases in a reactor vessel at a power plant, and that CO_2-rich water could be brought into contact with crushed carbonate minerals, which would then dissolve and form bicarbonate ions. Advantages of this scheme are that only shallow injection is required (>200 m) and no pH changes will result. Drawbacks are the need for large amounts of water and carbonate minerals.

Discharging CO_2 into the deep ocean appears to elicit significant opposition, especially by some environmental groups. Often, discharging CO_2 is equated with dumping toxic materials into the ocean, ignoring that CO_2 is not toxic, that dissolved carbon dioxide and carbonates are natural ingredients of seawater, and, as stated before, atmospheric CO_2 will eventually penetrate into deep water anyway. This is not to say that seawater would not be acidified by injecting CO_2. The magnitude of the impact on marine organisms depends on the extent of pH change and the duration of exposure. This impact can be mitigated by the method of CO_2 injection, e.g., dispersing the injected CO_2 by an array of diffusers, or adding pulverized limestone to the injected CO_2 in order to buffer the carbonic acid.

5. ECONOMICS

Carbon capture and storage costs can be considered in terms of four components: separation, compression, transport, and injection. These costs depend on many factors, including the source of the CO_2, transportation distance, and the type and characteristics of the storage reservoir. In this section, costs associated with capture from fossil fuel-fired power plants and with subsequent transport and storage are considered. In this case, the cost of capture includes both separation and compression costs because both of these processes almost always occur at the power plant.

5.1 Cost of Capture

Technologies to separate and compress CO_2 from power plant flue gases exist and are commercially available. However, they have not been optimized for capture of CO_2 from a power plant for the purpose of storage. The primary difference in capturing CO_2 for commercial markets versus capturing CO_2 for storage is the role of energy. In the former case, energy is a commodity, and all we care about is its price. In the latter case, using energy generates more CO_2 emissions, which is precisely what we want to avoid. An energy penalty can be calculated as $(x-y)/x$, where x is the output in kilowatts of a reference power plant without capture and y is the output in kilowatts of the same plant with capture. The calculation requires that the same fuel input be used in both cases. For example, if the power plant output is reduced by 20%

because of the capture process ($y = 0.8x$), the process is said to have an energy penalty of 20%. We can account for the energy penalty by calculating costs on a CO_2 avoided basis. As shown in Fig. 3, due to the extra energy required to capture CO_2, the amount of CO_2 emissions avoided is always less than the amount of CO_2 captured. Therefore, capturing CO_2 for purposes of storage requires more emphasis on reducing energy inputs than is the case in traditional commercial processes.

Based on the results of major economic studies available in the literature adjusted to a common economic basis, Fig. 4 summarizes the present cost of electricity (COE) from three types of CO_2 capture power plants: integrated gasification combined cycle(IGCC), pulverized coal-fired single cycle (PC), and natural gas combined cycle (NGCC). The mean and range (± 1 standard deviation) are shown for each capture plant, along with a typical COE for a no-capture plant. This results in an increase in the cost of electricity of 1–2 ¢/kWh for an NGCC plant, 1–3 ¢/kWh for an IGCC plant, and 2–4 ¢/kWh for a PC plant.

The energy penalties for each of these processes have also been estimated. The energy penalty for an NGCC plant is about 16%, whereas for a PC plant it is 28%. Each of these plants uses the amine solvent process (see Section 3). The energy penalty for a PC plant is greater than for an NGCC plant because coal has a larger carbon content compared to gas. The major energy losses are associated with energy required to blow the flue gas through the amine absorption column, the heat required to strip off the CO_2 and regenerate the amine, and the energy required to compress the CO_2. The energy penalty for an IGCC plant is 14%, actually less than for a PC plant despite its use of coal. This is because the high CO_2 partial pressure in the IGCC plant allows the use of an energy-efficient physical absorption process instead of the chemical absorption process. However, some of these gains are offset by the energy loss associated with converting the coal into CO_2 plus H_2.

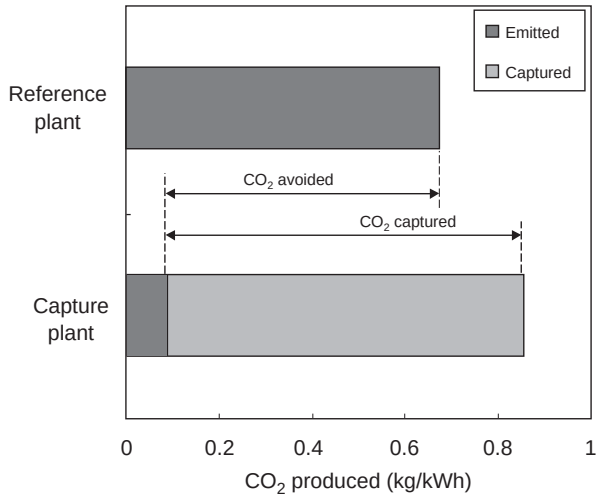


FIGURE 3 Graphical representation of avoided CO_2. The avoided emissions are simply the difference of the actual emissions per kilowatt-hour of the two plants. Note that due to the energy penalty, the emissions avoided are always less than the captured CO_2.

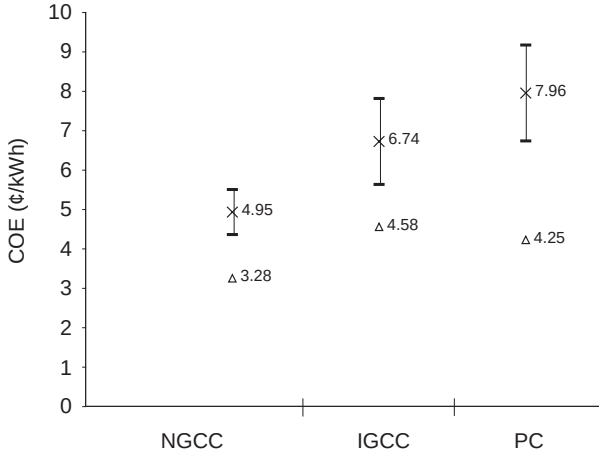


FIGURE 4 The cost of electricity (COE) with capture for various types of power plants. NGCC, Natural gas combined cycle; IGCC, integrated coal gasification combined cycle; PC, pulverized coal. The triangles represent a reference plant with no capture. The cost of electricity for CO_2 capture plants is based on a survey of the literature and is shown as a mean and a range of 1 standard deviation.

5.2 Cost of Transport

Figure 5 shows the cost of transporting CO_2 in large quantities by pipeline. Costs can vary greatly because pipeline costs depend on terrain, population density, etc. Economies of scale are realized when dealing with over 10 million metric tons per year (equivalent to about 1500 MW of coal-fired power). This cost is about \$0.50/t/100 km, compared to truck transport of \$6/t/100 km.

5.3 Cost of Injection and Storage

Figure 6 summarizes the cost of the various carbon storage technologies on a greenhouse gas-avoided basis. The points on the graphs are for a typical base case, and the bars represent the range between representative high- and low-cost cases. The ranges reflect the range of conditions found in the various reservoirs (depth, permeability, etc.), the distance

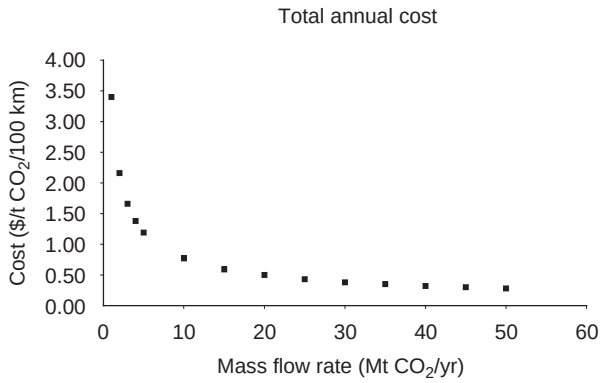


FIGURE 5 Total annual cost (capital and operation and maintenance) for CO_2 transport via pipeline as a function of CO_2 mass flow rate.

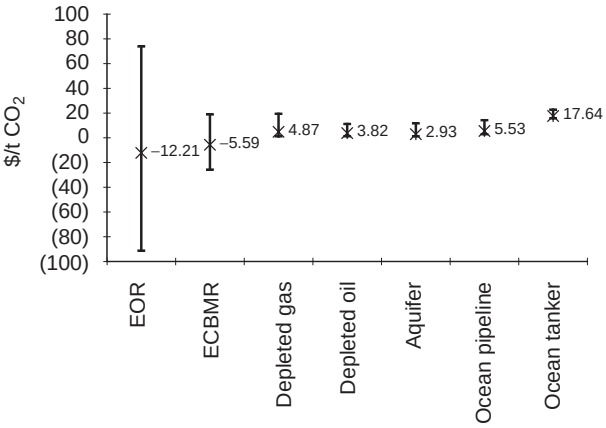


FIGURE 6 Range of costs for various carbon storage technologies on a greenhouse gas-avoided basis. EOR, Enhanced oil recovery; ECBMR, enhanced coal bed methane recovery.

between source and sink (a range of 0–300 km here), and the by-product prices (i.e., oil and gas). Excluding the more expensive ocean tanker option, the typical base case costs for CO_2 storage (transport + injection) without oil or gas by-product credit is in the range of \$3–5.50/t CO_2 (\$11–20/t C). The overall cost range can be characterized as \$2–15/t CO_2 (\$7–55/t C). With a by-product credit for the gas or oil, the credit will offset the storage costs in many instances. For example, in the base EOR case, one can afford to pay \$12.21/t CO_2 and still break even (i.e., the costs equal the by-product credit).

5.4 Overall Costs

Economic models of the general economy (i.e., a general equilibrium model) can be used to estimate the market carbon price required for adaptation of CCS technologies in the electric power industry. Carbon prices must be established through government policy, such as a tax or a cap-and-trade system. Assuming the costs and technology level outlined in the preceding section, carbon prices must reach \$100/t C in order for CCS technologies to start being adopted by the power industry on a significant scale (>5% market penetration). As the carbon price increases, CCS technologies will be adapted more quickly and achieve larger market penetration.

CCS technologies can be adopted at carbon prices much less than \$100/t C. These targets of opportunity will either have very inexpensive capture costs (from nonpower sources such as natural gas processing and ammonia production) or will be able to claim a by-product credit (e.g., EOR). All the commercial-scale CO_2 storage projects either in operation (Sleipner, Weyburn) or planned (Snovit by Statoil in the North Sea and In Salah by BP in Algeria) can be classified as targets of opportunity. Finally, new technologies can reduce the costs associated with CCS.

6. ALTERNATE APPROACHES

In the previous sections, we addressed the technologies for separating CO_2 from fossil fuel streams before or after combustion and storing the captured CO_2 in geologic or oceanic sinks. In this section, we briefly identify some alternative approaches that have been proposed for CO_2 capture and/or storage. The topics that we have chosen to include in this section are ones that have received significant publicity and/or funding. Their inclusion is in no way an endorsement, just as the exclusion of any approach is not a rejection. The enhanced uptake of CO_2 by the terrestrial biosphere (e.g., afforestation) is currently a subject of intensive debate, but this approach falls outside the scope of this article.

6.1 Capture by Microalgae

The concept is to grow algae in artificial ponds, add the necessary nutrients, and fertilize the ponds with CO_2 from flue gas. Under these conditions, it is possible to enhance the growth of microalgae, harvest the algal biomass, and convert it to food, feed, or fuel. At present, about 5000 tons of food- and feed-grade microalgae biomass is produced annually in large open pond systems. As such, this approach cannot be considered as a sequestration method because the CO_2 will be returned to the atmosphere on digestion and respiration of the food or feed. What is even worse, when used as a feed for

ruminating animals, some of the ingested carbon may be converted to methane, which, compared to carbon dioxide, is a stronger greenhouse gas. But if the biomass is converted to biofuel and subsequently combusted, then it replaces fossil fuel, and thus the commensurate emission of fossil fuel-generated CO_2 is avoided. However, for this approach to be viable as a greenhouse gas control method, it is necessary to significantly lower the cost from today's level. Despite some intensive efforts, primarily from Japan, little progress has been made toward this goal.

6.2 Ocean Fertilization

It has been hypothesized that by fertilizing the ocean with limiting nutrients such as iron, the growth of marine phytoplankton will be stimulated, thus increasing the uptake of atmospheric CO_2 by the ocean. The presumption is that a portion of the phytoplankton will eventually sink to the deep ocean. Researchers have targeted high-nutrient/low-chlorophyll (HNLC) ocean regions, specifically the eastern Equatorial Pacific, the northeastern Subarctic Pacific, and the Southern Ocean.

Four major open-ocean experiments have been conducted to test the "iron hypothesis," two in the Equatorial Pacific (IRONEX I in 1993 and IRONEX-II in 1995) and two in the Southern Ocean (SOIREE in 1999 and EISENEX in 2000). These experiments, funded through basic science programs (not sequestration programs), show conclusively that phytoplankton biomass can be dramatically increased by the addition of iron. However, although a necessary condition, it is not sufficient to claim iron fertilization will be effective as a CO_2 sequestration option.

The proponents of iron fertilization claim very cost-effective mitigation, on the order of \$1–10/t C, but critical scientific questions remain unanswered. Although iron increases uptake of CO_2 from the atmosphere to the surface ocean, it needs to be exported to the deep ocean to be effective for sequestration. No experiments have yet attempted to measure export efficiency, which is an extremely difficult value to measure (some people claim that it cannot be measured experimentally). In addition, there are concerns about the effect on ecosystems, such as inducing anoxia (oxygen depletion) and changing the composition of phytoplankton communities.

6.3 Mineral Storage

Several minerals found on the surface of the earth uptake CO_2 from the atmosphere, with the formation of carbonates, thus permanently storing the CO_2. Such minerals are calcium and magnesium silicates. For example, the following reaction occurs with serpentine, a magnesium silicate:

$$Mg_3Si_2O_5(OH)_4 + 3CO_2(g) = 3MgCO_3 + 2SiO_2 + 2H_2O$$

Although the reaction is thermodynamically favored, it is extremely slow in nature (characteristic time on the order of a 100,000 years). The challenge is to speed up the reaction in order to be able to design an economically viable process. Many reaction pathways have been explored to varying degrees. Although some have shown progress, none has yet resolved all the issues necessary to make a commercial process.

6.4 Nonbiological Capture from Air

The terrestrial biosphere routinely removes CO_2 from air, primarily through photosynthesis. It has been suggested that CO_2 can also be removed from air via nonbiological means. Although some concept papers have been published, no viable methods to accomplish this goal have been proposed. The problem is that the partial pressure of CO_2 in air is less than 0.0004 atm, compared to about 0.1 atm in flue gas and up to 20 atm in synthesis gas. The difficulty in capture increases as the partial pressure of CO_2 decreases. Therefore, it is questionable whether CO_2 can be captured from air with acceptable energy penalties and costs. If so, it almost surely will take development of a capture process very different from those that exist today.

6.5 Utilization

CO_2 from fossil fuel could be utilized as a raw material in the chemical industry for producing commercial products that are inert and long-lived, such as vulcanized rubber, polyurethane foam, and polycarbonates. Only a limited amount of CO_2 can be stored in such a fashion. Estimates of the world's commercial sales for CO_2 are less than 0.1 Gt C equivalent, compared to annual emissions of close to 7 Gt C equivalent. It has been suggested that CO_2 could be recycled into a fuel. This would create a market on the same scale as the CO_2 emissions. However, to recycle CO_2 to a fuel would require a carbon-free energy source. If such a source existed, experience suggests that it would be more efficient

and cost-effective to use that source directly to displace fossil fuels rather than to recycle CO_2.

SEE ALSO THE FOLLOWING ARTICLES

Biomass: Impact on Carbon Cycle and Greenhouse Gas Emissions • Carbon Sequestration: Terrestrial • Carbon Taxes and Climate Change • Climate Change and Energy, Overview • Climate Change: Impact on the Demand for Energy • Greenhouse Gas Abatement: Controversies in Cost Assessment • Greenhouse Gas Emissions, Alternative Scenarios of • Greenhouse Gas Emissions from Energy Systems, Comparison and Overview

Further Reading

Herzog, H. J. (2001). What future for carbon capture and sequestration? *Environ. Sci. Technol.* **35,** 148A–153A.

Herzog, H., Drake, E., and Adams. E. (1997). CO_2 capture, reuse and storage technologies for mitigating global climate change: A white paper. Energy Lab. Rep., DOE Order DE-AF22-96PC01257. Massachusetts Institute of Technology, Cambridge, MA.

U.S. Department of Energy. (1993). The capture, utilization and disposal of carbon dioxide from fossil fuel-fired power plants. DOE/ER-30194. Department of Energy, Washington, D.C.

U.S. Department of Energy. (1999). Carbon sequestration research and development. DOE/SC/FE-1. National Technical Information Service, Springfield, VA.

Williams, D. J., Durie, R. A., McMullan, P., Paulson, C. A. J., and Smith, A. Y. (eds.). (2001). Greenhouse gas control technologies. Proceedings of the Fifth International Conference on Greenhouse Gas Control Technologies, CSIRO Publ., Collingwood, Australia.

Carbon Sequestration, Terrestrial

R. LAL
The Ohio State University
Columbus, Ohio, United States

1. Introduction
2. Carbon Sequestration
3. Processes of Terrestrial Carbon Sequestration
4. Enhancing Terrestrial Carbon Sequestration
5. Ancillary Benefits
6. Restoration of Degraded Ecosystems and Terrestrial Carbon Sequestration
7. Potential and Challenges of Soil Carbon Sequestration
8. Climate Change and Terrestrial Carbon Sequestration
9. Nutrient Requirements for Terrestrial Carbon Sequestration
10. The Permanence of Carbon Sequestered in Terrestrial Ecosystems
11. Terrestrial Carbon Sequestration and Global Food Security
12. Conclusions

Glossary

ancillary benefits Additional benefits of a process.
anthropogenic Human-induced changes.
biomass/biosolids Material of organic origin.
carbon sequestration Transfer of carbon from atmosphere into the long-lived carbon pools such as vegetation and soil.
humus The decomposed organic matter in which remains of plants and animals are not identifiable.
mean residence time The average time an atom of carbon spends within a pool; calculated by dividing the pool by flux.
soil organic matter Sum of all organic substances in the soil.
soil quality Ability of a soil to perform functions of interest to humans such as biomass production, water filtration, biodegradation of pollutants, and soil carbon sequestration.
terrestrial carbon Carbon contained in the vegetation and soil.

Atmospheric concentration of carbon dioxide (CO_2) and other greenhouse gases (GHGs) has increased drastically since the industrial revolution and is still increasing with attendant risks of increase in global temperature. Thus, the need to identify options for mitigating atmospheric concentration of CO_2 is widely recognized. The Kyoto Protocol, ratified by several countries but not by the United States, mandates that industrialized nations reduce their net emissions with reference to the 1990 level by an agreed amount and recognizes a wide range of options to minimize the risks of global warming. The choice of specific options depends on the economic conditions, industrial infrastructure, land and water resources, agricultural conditions, and the like.

1. INTRODUCTION

Specific options of mitigating climate change can be grouped broadly into two categories: adaptive responses and mitigative responses (Fig. 1). Adaptive responses involve identification of options for sustainable management of natural resources (e.g., terrestrial ecosystems, water resources and wetlands, agricultural ecosystems, deserts) under changed climate with projected alterations in precipitation, frequency of extreme events, growing season duration, and incidence of pests and pathogens affecting health of human and other biota. Mitigative responses involve controlling atmospheric concentration of carbon dioxide (CO_2) either by removing it from the atmosphere or by avoiding its emission into the atmosphere. Important among mitigative strategies are reducing and sequestering emissions. Emissions reduction may involve improving energy use/conversion efficiency, developing alternatives to fossil fuel or developing carbon (C)-neutral fuel sources, using space reflectors to reduce the amount of radiation reaching the earth, absorbing CO_2 from the atmosphere using special extractants, and sequestering C in long-lived pools. Both adaptive and mitigative strategies are important to reducing

 © 2004 Elsevier Inc. All rights reserved.

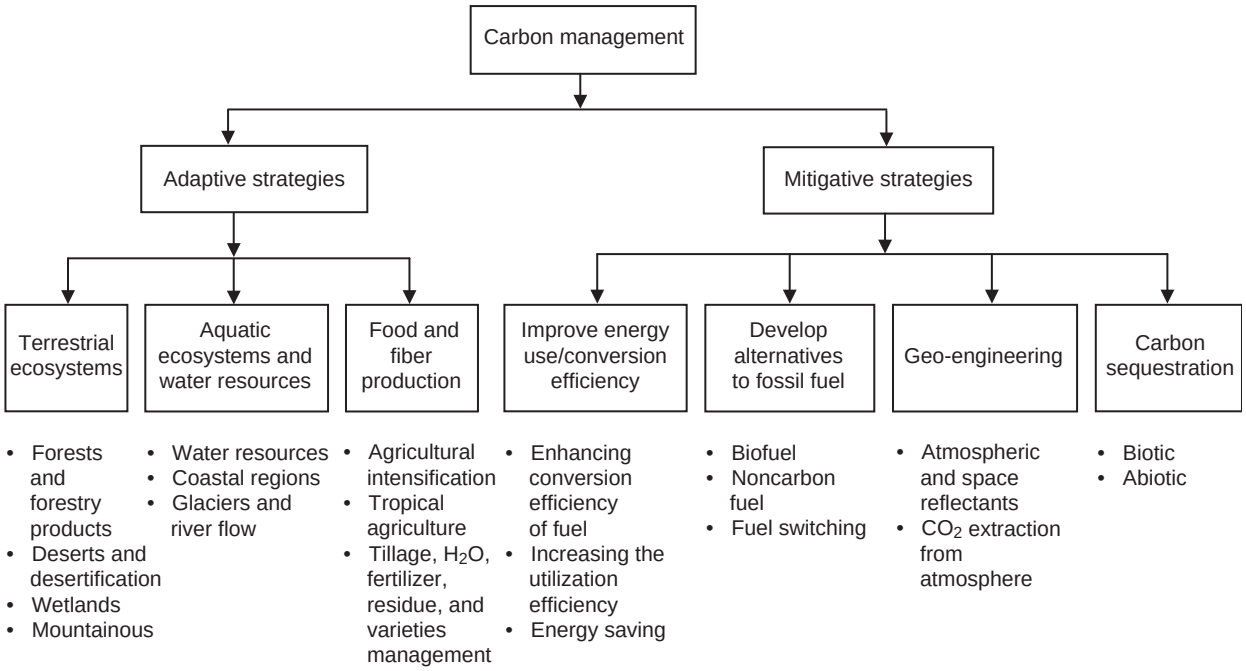


FIGURE 1 Carbon management strategies to mitigate the greenhouse effect.

the rate of enrichment of atmospheric concentration of CO_2. This article describes the principles, opportunities, and challenges of C sequestration in terrestrial pools composed of soils and vegetation.

2. CARBON SEQUESTRATION

Carbon sequestration implies capture and secure storage of C that would otherwise be emitted to or remain in the atmosphere. The objective is to transfer C from the atmosphere into long-lived pools (e.g., biota, soil, geologic strata, ocean). It is the net removal of CO_2 from the atmosphere or prevention of CO_2 emissions from terrestrial ecosystems. Strategies of C sequestration can be grouped under two broad categories: biotic and abiotic (Fig. 2). Biotic strategies are based on natural process of photosynthesis and transfer of CO_2 from atmosphere into vegetative, pedologic, and aquatic pools through mediation via green plants. Abiotic strategies involve separation, capture, compression, transport, and injection of CO_2 from power plant flue gases and effluent of industrial processes deep into ocean and geologic strata. The strategy is to use engineering techniques for keeping industrial emissions from reaching the atmosphere.

Terrestrial C sequestration is a natural process. It involves transfer of atmospheric CO_2 through photosynthesis into vegetative and pedologic/soil C pools. The terrestrial pool is the third largest among the five global C pools (Fig. 3). It includes 2300 petagrams (Pg, where 1 Pg $= 10^{15}$ g $= 1$ billion metric tons) of C in soil to 1-m depth and 600 Pg in the vegetative pool, including the detritus material. Thus, the terrestrial C pool is approximately four times the size of the atmospheric pool and 60% the size of the fossil C pool. It is also a very active pool and is continuously interacting with atmospheric and oceanic pools.

The natural processes of terrestrial C sequestration are already absorbing a large fraction of the anthropogenic emissions. For example, oceanic uptake is about 2 Pg C/year, and unknown terrestrial sinks absorbed a total of 2.8 Pg C/year during the 1990s. Thus, 4.8 Pg/year out of a total anthropogenic emission of 8 Pg/year (or 59%) was absorbed by natural sinks during the 1990s. Therefore, it is prudent to identify and enhance the capacity of natural processes.

Carbon sequestration, both biotic and abiotic, is an important strategy for mitigating risks of global warming. It can influence the global C cycle over a short period and reduce the equilibrium level of atmospheric CO_2 until the alternatives to fossil fuel take effect. A hypothetical scenario in Fig. 4 shows that atmospheric concentration of CO_2 by 2100 may be 700 parts per million in terms of volume (ppmv)

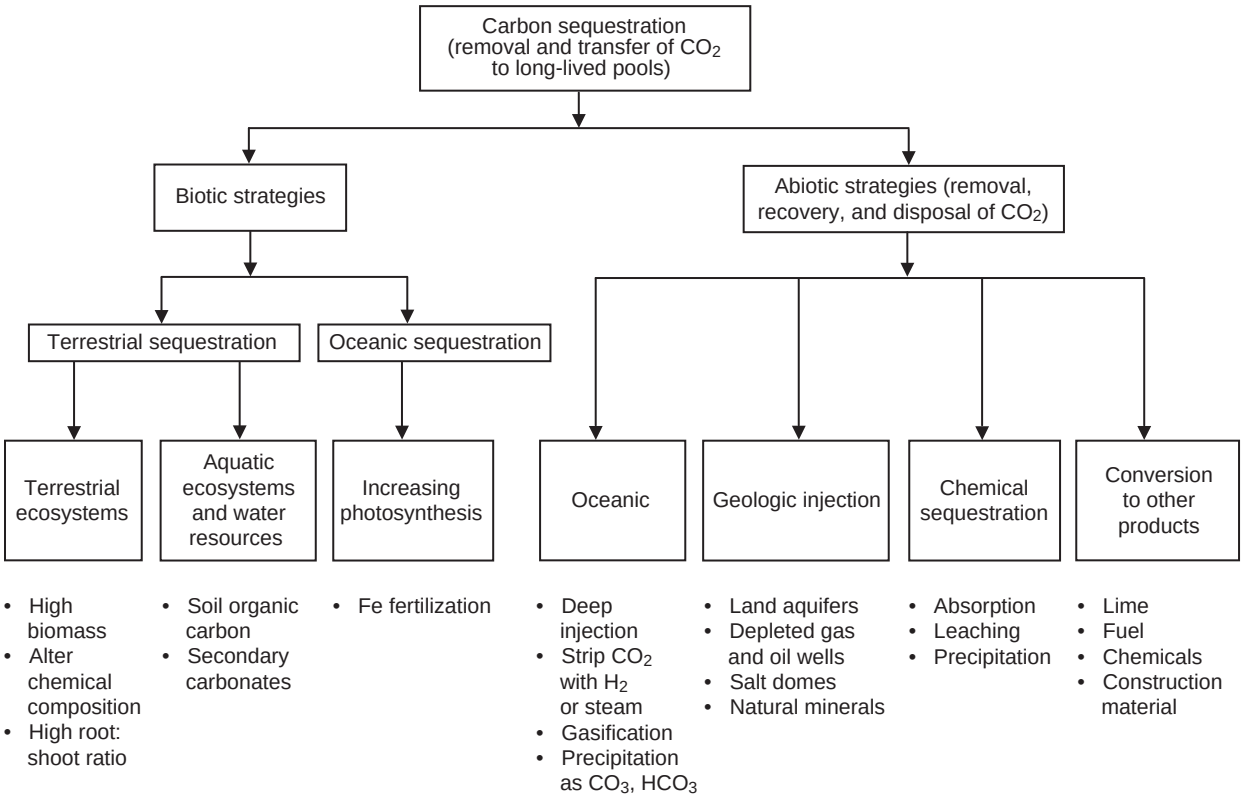


FIGURE 2 Carbon sequestration strategies.

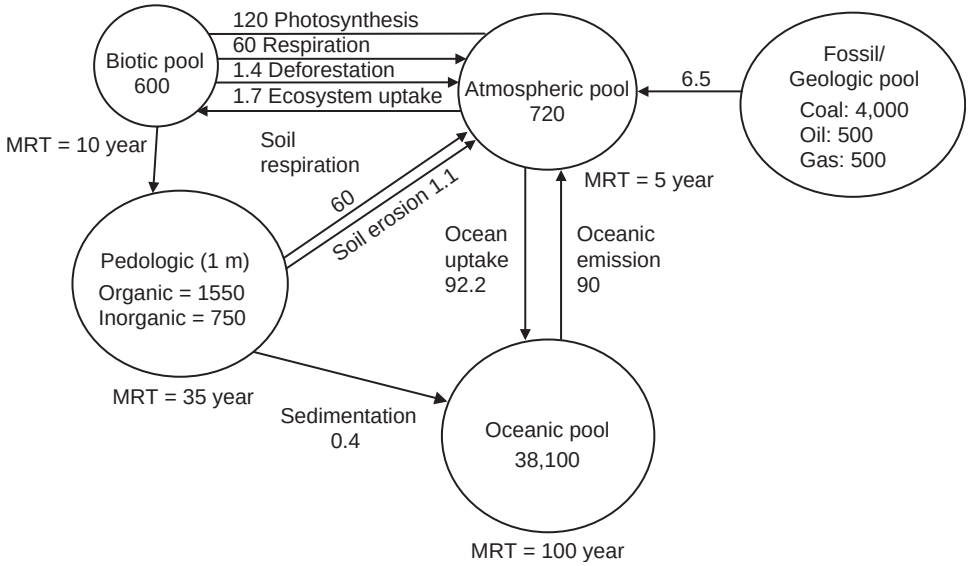


FIGURE 3 Global C pools and fluxes between them. All numbers are in petagrams (Pg = 10^{15} g = 1 billion metric tons). MRT, mean residence time.

with business as usual and only 550 ppmv by adoption of several C sequestration strategies. Feasibility of a C sequestration strategy depends on numerous factors: (1) sink capacity of the pool (e.g., soil, vegetation, geologic strata), (2) fate of the C sequestered and its residence time, (3) environmental impact on terrestrial or aquatic ecosystems, and (4) cost-effectiveness of the strategy.

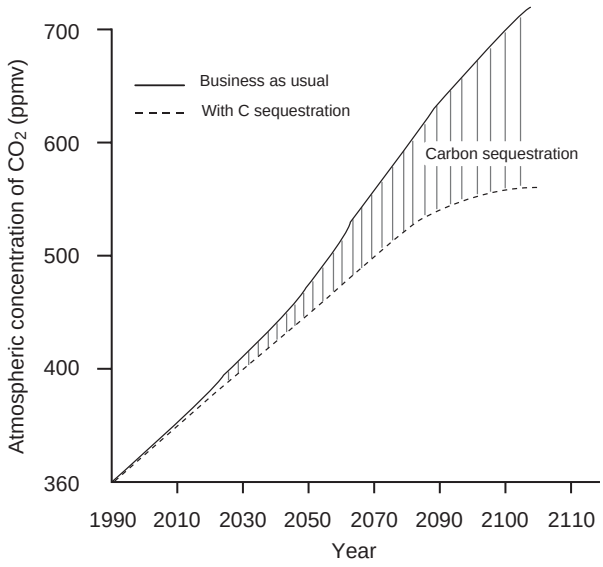


FIGURE 4 A hypothetical scenario about the impact of C sequestration by biotic and abiotic strategies on atmospheric concentration of CO_2.

The distribution of C pool in different biomes shows the high amount of terrestrial C in forest and savanna ecoregions and in northern peat lands (Table I). The terrestrial C pool in managed ecosystems (e.g., cropland, permanent crops) differs from others in terms of the biomass C pool that is nearly negligible because it is harvested for human and animal consumption. The data in Table II show the potential of terrestrial C sequestration at 5.7 to 10.1 Pg C/year, of which the potential of agricultural lands at 0.8 to 0.9 Pg C/year includes only soil C. Similar estimates for croplands and world soils have been made by others, yet there are numerous uncertainties in these estimates.

3. PROCESSES OF TERRESTRIAL CARBON SEQUESTRATION

The terrestrial sequestration involves biological transformation of atmospheric CO_2 into biomass and soil humus. Currently, the process leads to sequestration of approximately 2 Pg C/year. With improved management and natural CO_2 fertilization effect under the enhanced CO_2 levels, the potential of terrestrial C sequestration may be much larger (Table II). Principal among management options include the following:

- Conversion of marginal croplands to biomass energy crop production

TABLE I

Distribution of Terrestrial C Pool in Various Ecosystems

Ecosystem	Area (Mha)	NPP (Pg C/year)	Biomass C (Pg)	Soil C to 1 m (Pg)
Forest				
Tropical	1480	13.7	244.2	123
Temperate	750	5.0	92.0	90
Boreal	900	3.2	22.0	135
Savanna				
Temperate woodlands	200	1.4	16.0	24
Chaparral	250	0.9	8.0	30
Tropical	2250	17.8	65.9	263
Temperate grassland	1250	4.4	9.0	295
Tundra, arctic & alpine	950	1.0	6.0	121
Deserts	3000	1.5	7.2	191
Lake and streams	200	0.4	0.0	0
Wetlands	280	3.3	12.0	202
Peat lands	340	0.0	0.0	455
Cultivated and permanent crops	1480	6.3	3.0	117
Perpetual ice	1550	0.0	0.0	0
Urban	200	0.2	1.0	10
Total	15,080	59.1	486.4	2056

Source. Adopted from Amthor and Huston (1998), Intergovernment Panel on Climate Change (2000), and U.S. Department of Energy (1999).

- Restoration of degraded soils and ecosystems
- Agricultural intensification on cropland and pastures
- Afforestation and forest soil management
- Restoration of wetlands

If the rate of C sequestration by adopting these options can be increased by 0.2% per year over 25 years, it can transfer 100 Pg C from the atmosphere into the terrestrial ecosystem through natural processes of biological transformation.

4. ENHANCING TERRESTRIAL CARBON SEQUESTRATION

The natural processes governing terrestrial C sequestration are outlined in Fig. 5. The vegetative C pool can be enhanced by improving biomass production,

TABLE II

Global Potential of Terrestrial Carbon Sequestration

Biome	Potential of C sequestration (Pg C/year)
Agricultural lands	0.85–0.90
Biomass croplands	0.50–0.80
Grass lands	0.50
Range lands	1.2
Forest lands	1.0–3.0
Urban lands	—
Deserts and degraded lands	0.8–1.3
Terrestrial sediments	0.7–1.7
Boreal peat lands and other wetlands	0.1–0.7
Total	5.65–10.1

Source. U.S. Department of Energy (1999).

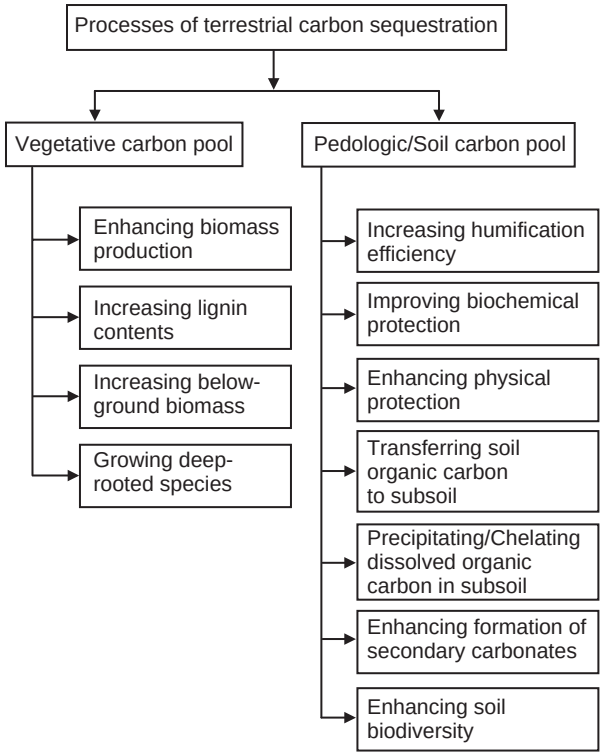


FIGURE 5 Processes of enhancing terrestrial carbon sequestration.

increasing the lignin content that reduces the rate of decomposition, and diverting biomass into a deep and prolific root system. Afforestation, conversion of agriculturally marginal soils into restorative land uses, choice of fast-growing species with deep root systems, and adoption of recommended management practices (RMPs) for stand, soil fertility, and pest management. In addition to the return of biomass to soil, there are several processes that moderate the soil C pool. The humification efficiency, the proportion of biomass converted to the stable humic fraction, depends on soil water and temperature regimes and the availability of essential elements (e.g., N, P, S, Zn, Cu, Mo, Fe). The amount and nature of clay, essential to formation and stabilization of aggregates, strongly influence the sequestration and residence time of C in soil. Effectiveness of chemical, biochemical, and physical processes of soil C sequestration depends on formation of stable micro-aggregates or organo–mineral complexes. Soil biodiversity, activity and species diversity of soil fauna such as earthworms and termites, plays an important role in bioturbation, C cycle, and humification of biosolids. Transfer of humus into the subsoil, precipitation of dissolved organic C (DOC), leaching and precipitation of carbonates, and formation of secondary carbonates are important processes that enhance soil C pool and increase its residence time.

Conversion of natural ecosystems to agricultural ones (e.g., cropland, grazing land) depletes soil organic C (SOC) pool by as much as 50 to 80%. The magnitude and rate of depletion are generally greater in warm climates than in cool climates, in coarse-textured soils than in heavy-textured soils, and in soils with high antecedent SOC pool than in low antecedent SOC pool. Therefore, most agricultural soils now contain lower SOC pool than their potential capacity for the specific ecoregion. Conversion to a restorative land use and adoption of RMPs can enhance SOC pool. Feasible RMPs of agricultural intensification include using no-till farming with frequent incorporation of cover crops (meadows) in the rotation cycle, using manure and other biosolids as a component of integrated nutrient management strategy, and improving pasture growth by planting appropriate species and controlling stocking rate and the like. Restoring eroded/degraded soils and ecosystems is extremely important to enhancing the terrestrial C pool.

5. ANCILLARY BENEFITS

There are numerous ancillary benefits of terrestrial C sequestration. Enhancing vegetative/ground cover and soil C pool reduces the risks of accelerated runoff, erosion, and sedimentation. The attendant decline in dissolved and suspended loads reduces

risks of eutrophication of surface waters. Similarly, an increase in soil C pool decreases the transport of agricultural chemicals and other pollutants into the ground water. There are notable benefits of increased activity and species diversity of soil fauna and flora with an increase in SOC pool. In addition to enhancing soil structure, in terms of aggregation and its beneficial impacts on water retention and transmission, an increase in soil biodiversity also strengthens elemental/nutrient cycling. Soil is a living membrane between bedrock and the atmosphere, and its effectiveness is enhanced by an increase in soil C pool and its quality.

An increase in terrestrial C sequestration in general, and in soil C sequestration in particular, improves soil physical, chemical, and biological quality (Fig. 6). Improvements in soil physical quality lead to decreased risks of accelerated runoff and erosion, reduce crusting and compaction, and improve plant-available water capacity. Improvements in soil chemical quality lead to increased cation/anion exchange capacity, elemental balance, and favorable soil reaction. Improvements in soil biological quality lead to increased soil biodiversity and bioturbation, elemental/nutrient cycling, and biodegradation and bioremediation of pollutants. An integrative effect of improvement in soil quality is to increase biomass/agronomic productivity. On a global scale, soil C sequestration can have an important impact on food security.

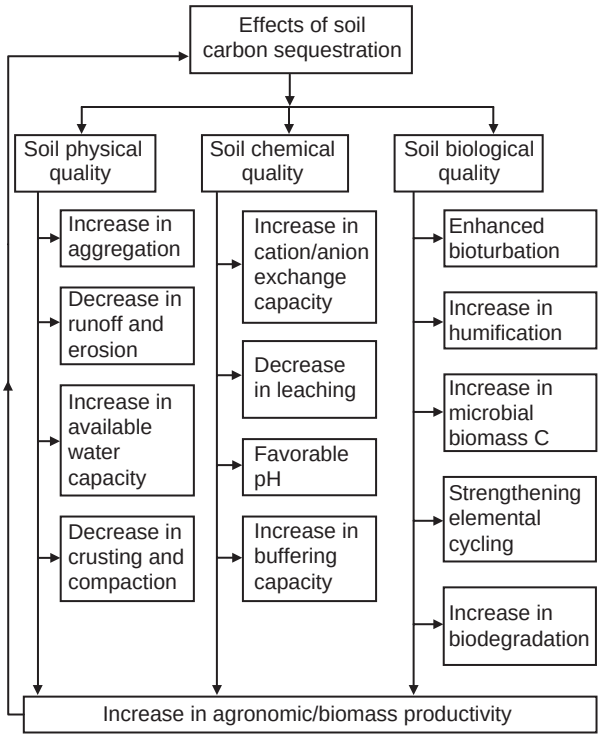


FIGURE 6 Favorable effects of soil carbon sequestration on soil quality and agronomic/biomass productivity.

6. RESTORATION OF DEGRADED ECOSYSTEMS AND TERRESTRIAL CARBON SEQUESTRATION

Accelerated soil erosion, soil degradation by other degradative processes (e.g., salinization, nutrient depletion, elemental imbalance, acidification), and desertification are severe global issues. Accelerated erosion by water and wind is the most widespread type of soil degradation. The land area affected by water erosion is 1094 million hectares (Mha), of which 751 Mha is severely affected. The land area affected by wind erosion is 549 Mha, of which 296 Mha is severely affected. In addition, 239 Mha is affected by chemical degradation and 83 Mha is affected by physical degradation. Thus, the total land area affected by soil degradation is 1965 Mha or 15% of the earth's land area. With a sediment delivery ratio of 13 to 20%, the suspended sediment load is estimated at 20×10^9 metric tons (Mg)/year. Soil erosion affects SOC dynamics by slaking and breakdown of aggregates, preferential removal of C in surface runoff or wind, redistribution of C over the landscape, and mineralization of displaced/redistributed C. The redistributed SOC is generally light or labile fraction composed of particulate organic carbon (POC) and is easily mineralized. As much as 4 to 6 Pg C/year is transported globally by water erosion. Of this, 2.8 to 4.2 Pg C/year is redistributed over the landscape and transferred to depressional sites, 0.4 to 0.6 Pg C/year is transported into the ocean, and 0.8 to 1.2 Pg C/year is emitted into the atmosphere. The historic loss of SOC pool from soil degradation and other anthropogenic processes is 66 to 90 Pg C (78 ± 12 Pg C), compared with total terrestrial C loss of 136 ± 55 Pg C. Of the total SOC loss, that due to erosion by water and wind is estimated at 19 to 32 Pg C (26 ± 9 Pg C) or 33% of the total loss.

There is also a strong link between desertification and emission of C from soil and vegetation of the dryland ecosystems. Desertification is defined as the diminution or destruction of the biological potential of land that ultimately can lead to desert-like conditions. Estimates of the extent of desertification are wide-ranging and are often unreliable. The land

area affected by desertification is estimated at 3.5 to 4.0 billion ha. The available data on the rate of desertification is also highly speculative and is estimated by some to be 5.8 Mha/year.

Similar to other degradative processes, desertification leads to depletion of the terrestrial C pool. The historic loss of C due to desertification is estimated at 8 to 12 Pg from the soil C pool and 10 to 16 Pg from the vegetation C pool. Thus, the total historic C loss due to desertification may be 18 to 28 Pg C.

Therefore, desertification control and restoration of degraded soils and ecosystems can reverse the degradative trends and sequester a large fraction of the historic C loss. The potential of desertification control for C sequestration is estimated at 0.6 to 1.4 Pg C/year. Management of drylands through desertification control has an overall C sequestration potential of 1.0 Pg C/year. These estimates of high C sequestration potential through restoration of degraded/desertified soils and ecosystems are in contrast to the low overall potential of world soils reported by some.

7. POTENTIAL AND CHALLENGES OF SOIL CARBON SEQUESTRATION

Of the two components of terrestrial C sequestration—soil and vegetation—the importance of soil C sequestration is not widely recognized. The Kyoto Protocol has not yet accepted the soil C sink as offset for the fossil fuel emission. Yet soil C sequestration has vast potential and is a truly win–win situation.

Assuming that 66 Pg C has been depleted from the SOC pool, perhaps 66% of it can be resequestered. The global potential of soil C sequestration in the world's croplands is estimated at 0.6 to 1.2 Pg C/year. This potential can be realized over 50 years or until the sink capacity is filled.

Some have questioned the practicability or feasibility of realizing the potential of soil C sequestration. The caution against optimism is based on the hidden C costs of fertilizers, irrigation, and other input that may produce no net gains by soil C sequestration. However, it is important to realize that SOC sequestration is a by-product of adopting RMPs and agricultural intensification, which are needed to enhance agronomic production for achieving global food security. The global population of 6.2 billion is increasing at a rate of 73 million persons (or 1.3%) per year and is projected to reach 7.5 billion by 2020 and 9.4 billion by 2050. Consequently, the future food demand will increase drastically, and this will necessitate adoption of restorative measures to enhance soil quality and use of off-farm input (e.g., fertilizers, pesticides, irrigation, improved varieties) to increase productivity. Thus, fertilizers and other inputs are being used not for soil C sequestration but rather for increasing agricultural production to meet the needs of the rapidly increasing population. Soil C sequestration is an ancillary benefit of this inevitability and global necessity.

There is also a question of the finite nature of the terrestrial C sink in general and that of the soil C sink in particular. The soil C sink capacity is approximately 50 to 60 Pg C out of the total capacity of the terrestrial C sink of 136 Pg. In contrast, the capacity of geologic and terrestrial sinks is much larger. However, soil C sequestration is the most cost-effective option in the near future, and there are no adverse ecological impacts. Engineering techniques for abiotic (e.g., deep injection of CO_2 into geologic strata and ocean) strategies remain a work in progress.

The question of permanence also needs to be addressed. How long can C sequestered in soil and biota remain in these pools? Does an occasional plowing of a no-till field negate all of the gains made so far in soil C sequestration? These are relevant questions, and the answers may vary for different soils and ecoregions.

Although ratified by neither the United States nor Russia, the Kyoto Protocol demands that C stored in terrestrial ecosystems be assessed by transparent, credible, precise, and cost-effective methods. Questions have often been raised about the assessment of soil C pool and fluxes on soilscape, farm, regional, and national scales. It is important to realize that standard methods are available to quantitatively assess SOC pool and fluxes on scales ranging from the pedon or soilscape level and then extrapolated to the ecosystem, regional, and national levels. Soil scientists and agronomists have studied changes in soil organic matter in relation to soil fertility and agronomic productivity since 1900 or earlier. They have adapted and improved the analytical procedures for assessing soil C sequestration for mitigating climate change since the 1980s. New procedures are being developed to assess soil C rapidly, precisely, and economically. Promising among new analytical procedures are accelerator mass spectrometry (AMS), pyrolysis molecular beam mass spectrometry (Py-MBMS), and laser-induced breakdown spectroscopy (LIBS). The LIBS technique presumably has good detection limits, precision, and accuracy.

8. CLIMATE CHANGE AND TERRESTRIAL CARBON SEQUESTRATION

The increase in global temperature during the 20th century has been estimated at 0.6°C and is projected to be 1.4 to 5.8°C by 2100 relative to the global temperature in 1990. The projected increase in global temperature is attributed to enrichment of the atmospheric concentration of CO_2 (which has increased from 280 to 365 ppm and is increasing at the rate of 0.4% or 1.5 ppm/year), CH_4 (which has increased from 0.80 to 1.74 ppm and is increasing at the rate of 0.75% or 0.015 ppm/year), and N_2O (which has increased from 288 to 311 ppb and is increasing at the rate of 0.25% or 0.8 ppb/year).

The implications of rapidly changing atmospheric chemistry are complex and not very well understood. The projected climate change may have a strong impact on terrestrial biomes, biomass productivity, and soil quality. It is estimated that for every 1°C increase in global temperature, the biome or vegetational zones (e.g., boreal forests, temperate forests, wooded/grass savannas) may shift poleward by 200 to 300 km. There may be a change in species composition of the climax vegetation within each biome. In northern latitudes where the temperature change is likely to be most drastic, every 1°C increase in temperature may prolong the growing season duration by 10 days. With predicted climate change, the rate of mineralization of soil organic matter may increase with adverse impacts on soil quality.

There is a general consensus that CO_2 enrichment will have a fertilization effect and will enhance C storage in the vegetation (both above and below ground) for 50 to 100 years. Numerous free-air CO_2 enrichment (FACE) experiments have predicted increases in biomass production through increased net primary productivity (NPP), particularly in high latitudes. It was predicted that with an increase in temperature, tundra and boreal biomass will emit increasingly more C to the atmosphere while the humid tropical forests will continue to be a net sink. Soils of the tundra biome contain 393 Pg C as SOC and 17 Pg C as soil inorganic C (SIC) (410 Pg total), which is 16.4% of the world soil C pool. Soils of the boreal biome contain 382 Pg of SOC and 258 Pg of SIC (640 Pg total), which is 25.6% of the world soil C pool. Together, soils of the tundra and boreal (taiga) biomes contain 772 Pg or 42% of the world soil C pool. Some soils of these regions, especially histosols and cryosols, have been major C sinks in the historic past. The projected climate change and some anthropogenic activities may disrupt the terrestrial C cycle of these biomes and render them a major source of CO_2, CH_4, and N_2O. There exists a real danger that an increase in global temperature may result in a long-term loss of the soil C pool.

There are numerous knowledge gaps and uncertainties regarding the impact of climate change on terrestrial C pool and fluxes. Principal knowledge gaps and uncertainties exist regarding fine root biology and longevity, nutrient (especially N and P) availability, interaction among cycles of various elements (e.g., C, N, P, H_2O), and below-ground response.

9. NUTRIENT REQUIREMENTS FOR TERRESTRIAL CARBON SEQUESTRATION

Carbon is only one of the building blocks of the terrestrial C pool. Other important components of the biomass and soil are essential macroelements (e.g., N, P, K, Ca, Mg) and micronutrients (e.g., Cu, Zn, B, Mo, Fe, Mn). A total of 1 Mg (or 1000 kg) of biomass produced by photosynthesis contains approximately 350 to 450 kg of C, 5 to 12 kg of N, 1 to 2 kg of P, 12 to 16 kg of K, 2 to 15 kg of Ca, and 2 to 5 kg of Mg. Therefore, biomass production of 10 Mg/ha/year requires the availability of 50 to 120 kg of N, 10 to 20 kg of P, 120 to 160 kg of K, 20 to 150 kg of Ca, and 20 to 50 kg of Mg. Additional nutrients are required for grain production. Nutrient requirements are lesser for wood and forages production than for grain and tuberous crops.

Biomass (e.g., crop residues, leaf litter, detritus material, roots) is eventually converted into SOC pool through humification mediated by microbial processes. The humification efficiency is between 5 and 15%, depending on the composition of the biomass, soil properties, and climate. Soil C sequestration implies an increase in the SOC pool through conversion of biomass into humus. Similar to biomass production through photosynthesis, humification of biomass into SOC involves availability of nutrients. Additional nutrients required are N, P, S, and other minor elements. Assuming an elemental ratio of 12:1 for C:N, 50:1 for C:P, and 70:1 for C:S, sequestering 10,000 kg of C into humus will require 25,000 kg of residues (40% C), 833 kg of N, 200 kg of P, and 143 kg of S. Additional nutrients may be available in crop residue or soil or may be supplied through fertilizers and amendments. Nitrogen may

also be made available through biological nitrogen fixation (BNF) and atmospheric deposition. It is argued that sequestration of soil C can be achieved only by the availability of these nutrients.

Nutrient recycling, from those contained in the biomass and in the subsoil layers, is crucial to terrestrial C sequestration in general and to soil C sequestration in particular. For soil C sequestration of 500 kg C/ha/year, as is the normal rate for conversion from plow till to no till in Ohio, sequestration of 500 kg of C into humus (58% C) will require about 14 Mg of biomass (40% C and humification efficiency of 15%). This amount of residue from cereals will contain 65 to 200 kg of N, 8 to 40 kg of P, 70 to 340 kg of K, 22 to 260 kg of Ca, and 10 to 130 kg of Mg. In comparison, the nutrients required for sequestration of 500 kg of C into humus are 42 kg of N, 10 kg of P, and 7 kg of S. Indeed, most of the nutrients required are contained in the biomass. Therefore, soil C sequestration is limited mostly by the availability of biomass. However, the production of biomass is limited by the availability of nutrients, especially N and P.

10. THE PERMANENCE OF CARBON SEQUESTERED IN TERRESTRIAL ECOSYSTEMS

The mean residence time (MRT) of C sequestered in soil and biota depends on numerous factors, including land use, management system, soil type, vegetation, and climate. The MRT is computed as a ratio of the pool (Mg/ha) to the flux (Mg/ha/year). On this basis, the global MRT of C is 5 years in the atmosphere, 10 years in the vegetation, 35 years in the soils, and 100 years in the ocean. However, there is a wide range of MRT depending on the site-specific conditions and characteristics of the organic matter. For example, the MRT of soil organic matter is less than 1 year for the labile fraction, 1 to 100 years for the intermediate fraction, and several millennia for the passive fraction. The MRT also depends on management. Conversion from plow till to no till sequesters C in soil, and the MRT depends on continuous use of no till. Even occasional plowing can lead to emission of CO_2.

Biofuel production is another important strategy. The C sequestered in biomass is released when the biosolids are burned. However, C released from biofuel is recycled again and is in fact a fossil fuel offset. There is a major difference between the two. The fossil fuel combustion releases new CO_2 into the atmosphere, and burning of biofuels merely recycles the CO_2.

11. TERRESTRIAL CARBON SEQUESTRATION AND GLOBAL FOOD SECURITY

An important ancillary benefit of terrestrial C sequestration in general, and of soil C sequestration in particular, is the potential for increasing world food security. The projected global warming, with a likely increase in mean global temperature of $4 \pm 2^{\circ}C$, may adversely affect soil temperature and water regimes and the NPP. The increase in soil temperature may reduce soil organic matter pool with an attendant decrease in soil quality, increase in erosion and soil degradation, and decrease in NPP and agronomic yields. In contrast, the degradative trends and downward spiral may be reversed by an increase in soil C sequestration leading to improvement in soil quality and an increase in agronomic/biomass productivity. There already exists a large intercountry/interregional variation in average yields of wheat (*Triticum aestivum*) and corn (*Zea mays*). Yields of wheat range from a high of 6.0 to 7.8 Mg/ha in European Union countries (e.g., United Kingdom, Denmark, Germany, France), to a middle range of 2.4 to 2.7 Mg/ha (e.g., United States, India, Romania, Ukraine, Argentina, Canada), to a very low range of 1.0 to 2.2 Mg/ha (e.g., Pakistan, Turkey, Australia, Iran, Russia, Kazakhstan). Similarly, average yields of corn range from a high of 8 to 10 Mg/ha (e.g., Italy, France, Spain, United States), to a middle range of 4 to 7 Mg/ha (e.g., Canada, Egypt, Hungary, Argentina), to a low range of 1 to 2 Mg/ha (e.g., Nigeria, Philippines, India). A concern regarding projected global warming is whether the agroecological yield may be adversely affected by the change in soil temperature and moisture regimes and the threat of increase in food deficit in Sub-Saharan Africa and elsewhere in developing countries.

The adverse impact of decline in SOC pool on global food security is also evident from the required increase in crop yield in developing countries to meet the demands of increased population. With no possibility of bringing new land area under cultivation in most developing countries, the yields of cereals in developing countries will have to increase from 2.64 Mg/ha in 1997–1998 to 4.4 Mg/ha by 2025 and to 6.0 Mg/ha by 2050. Can this drastic increase in yields of cereals in developing countries

be attained by agricultural intensification in the face of projected global warming? An important factor that can help to attain this challenging goal is improvement in soil quality through terrestrial and soil C sequestration. Improvements in soil C pool can be achieved by increasing below-ground C directly through input of root biomass. The importance of irrigation, fertilizer and nutrient acquisition, erosion control, and use of mulch farming and conservation tillage cannot be overemphasized.

12. CONCLUSIONS

Terrestrial C sequestration is a truly win–win scenario. In addition to improving the much needed agronomic/biomass production to meet the needs of a rapidly increasing world population, terrestrial C sequestration improves soil quality, reduces sediment transport in rivers and reservoirs, enhances biodiversity, increases biodegradation of pollutants, and mitigates the risks of climate change. Besides, it is a natural process with numerous ancillary benefits. Although it has a finite capacity to absorb atmospheric CO_2, it is the most desirable option available at the onset of the 21st century. The strategy of terrestrial C sequestration buys us time until C-neutral fuel options take effect.

SEE ALSO THE FOLLOWING ARTICLES

Biomass for Renewable Energy and Fuels • Biomass: Impact on Carbon Cycle and Greenhouse Gas Emissions • Carbon Capture and Storage from Fossil Fuel Use • Carbon Taxes and Climate Change • Clean Coal Technology • Greenhouse Gas Abatement: Controversies in Cost Assessment • Greenhouse Gas Emissions, Alternative Scenarios of • Greenhouse Gas Emissions from Energy Systems, Comparison and Overview

Further Reading

Amthor, J. S., and Huston, M. A. (1998). "Terrestrial Ecosystem Responses to Global Change: A Research Strategy (ORNL/TM-1998/27)." Oak Ridge National Laboratory, Nashville, TN.

Eswaran, H., Vanden Berg, E., Reich, P., and Kimble, J. M. (1995). Global soil carbon resources. *In* "Soils and Global Change" (R. Lal, J. M. Kimble, E. Levine, and B. A. Stewart, Eds.), pp. 27–43. CRC Press, Boca Raton, FL.

Halmann, M. M., and Steinberg, M. (1999). "Greenhouse Gas Carbon Dioxide Mitigation: Science and Technology." Lewis Publishers, Boca Raton, FL.

Hao, Y., Lal, R., Owens, L., Izaurralde, R. C., Post, M., and Hothem, D. (2002). Effect of cropland management and slope position on soil organic carbon pools at the North Appalachian Experimental Watersheds. *Soil Tillage Res.* **68,** 133–142.

Himes, F. L. (1998). Nitrogen, sulfur, and phosphorus and the sequestering of carbon. *In* "Soil Processes and the Carbon Cycle" (R. Lal, J. M. Kimble, R. F. Follett, and B. A. Stewart, Eds.), pp. 315–319. CRC Press, Boca Raton, FL.

Intergovernment Panel on Climate Change. (1996). "Climate Change 1995: Impacts, Adaptations, and Mitigation of Climatic Change—Scientific–Technical Analyses." Cambridge University Press, Cambridge, UK.

Intergovernment Panel on Climate Change. (2000). "Land Use, Land Use Change, and Forestry: Special Report." Cambridge University Press, Cambridge, UK.

Intergovernment Panel on Climate Change. (2001). "Climate Change: The Scientific Basis." Cambridge University Press, Cambridge, UK.

Lal, R. (1995). Global soil erosion by water and carbon dynamics. *In* "Soils and Global Change" (R. Lal, J. M. Kimble, E. Levine, and B. A. Stewart, Eds.), pp. 131–141. CRC/Lewis Publishers, Boca Raton, FL.

Lal, R. (1999). Soil management and restoration for C sequestration to mitigate the greenhouse effect. *Prog. Environ. Sci.* **1,** 307–326.

Lal, R. (2001). Potential of desertification control to sequester carbon and mitigate the greenhouse effect. *Climate Change* **15,** 35–72.

Lal, R. (2001). World cropland soils as a source or sink for atmospheric carbon. *Adv. Agronomy* **71,** 145–191.

Lal, R. (2003a). Global potential of soil C sequestration to mitigate the greenhouse effect. *Critical Rev. Plant Sci.* **22,** 151–184.

Lal, R. (2003b). Soil erosion and the global carbon budget. *Environ. Intl.* **29,** 437–450.

Mainguet, M. (1991). "Desertification: Natural Background and Human Mismanagement." Springer-Verlag, Berlin.

Oberthür, S., and Ott, H. E. (2001). "The Kyoto Protocol: International Climate Policy for the 21st Century." Springer, Berlin.

Oldeman, L. R. (1994). The global extent of soil degradation. *In* "Soil Resilience and Sustainable Land Use" (D. J. Greenland and I. Szabolcs, Eds.), pp. 99–118. CAB International, Wallingford, UK.

Oldeman, L. R., and Van Lynden, G. W. J. (1998). Revisiting the GLASOD methodology. *In* "Methods for Assessment of Soil Degradation" (R. Lal, W. H. Blum, C. Valentine, and B. A. Stewart, Eds.), pp. 423–440. CRC Press, Boca Raton, FL.

Schlesinger, W. H. (1999). Carbon sequestration in soils. *Science* **286,** 2095.

U.S. Department of Energy. (1999). "Carbon Sequestration: Research and Development." U.S. DOE, National Technical Information Service, Springfield, VA.

Carbon Taxes and Climate Change

MARC CHUPKA
The Brattle Group
Washington, DC, United States

1. **The Policy Issue**
2. **Environmental Taxes and Conventional Taxes**
3. **Carbon Taxes as Environmental Policy**
4. **Carbon Taxes as Fiscal Instruments**
5. **Carbon Tax Implementation**
6. **Conclusion**

Glossary

distortionary tax A conventional tax levied on a good or service that has an economic burden in excess of the direct revenue generated.

environmental tax A tax levied on pollution or polluting behavior designed to reduce the amount of environmental damage.

equity The "fairness" of the distribution of tax burden or incidence.

incidence The measurement of the tax burden that accounts for shifts away from the nominal taxpayer to other parties through changes in prices and quantities transacted.

revenue neutrality Changes in tax structures or rates that increase revenues from one source and decrease revenues from another source by an equal amount.

revenue recycling Utilizing environmental tax revenues as a source of funds for reducing other taxes.

Carbon taxes, which are levies on fossil fuel in proportion to the carbon content, have been analyzed as a policy instrument to reduce carbon dioxide (CO_2) emissions. Economists have long recognized the economic advantages of emission taxes compared to more widely used instruments of environmental control such as standards. Environmental tax approaches also generate revenue that can be utilized to reduce distortionary taxes on labor, income, or consumption. In order to appreciably reduce CO_2 emissions, carbon taxes would be set at levels high enough to generate substantial revenues and significantly alter macroeconomic outcomes. This article explores some of the conceptual issues raised by imposing carbon taxes as instruments of environmental control and as tools of fiscal policy.

1. THE POLICY ISSUE

The bulk of greenhouse gas emissions from industrialized economies occurs as carbon dioxide (CO_2) released from burning fossil fuels, such as coal, petroleum products, and natural gas. Concern regarding the potential for global climate change from the accumulation of greenhouse gases in the atmosphere has spawned an assessment of policy tools to reduce greenhouse gas emissions. Among the vast array of potential policies to reduce CO_2 emissions, one frequently proposed and analyzed option would levy a tax on CO_2 emissions or, as a practical equivalent, tax fossil fuels in proportion to their carbon content. This is often called a carbon tax.

Analyses of the potential impacts of carbon taxes encompass a diverse set of economic disciplines, including environmental economics, energy economics, public finance, macroeconomic growth, and monetary and fiscal policy. The span of analytic approaches is necessary because a carbon tax is both an instrument of environmental control, with advantages and disadvantages relative to other potential measures; and a tax, which needs to be evaluated relative to alternative methods of raising revenues. The diversity of analyses also reflects the sheer magnitude of a carbon tax policy: A tax high enough to appreciably reduce aggregate CO_2 emissions could significantly alter energy prices, change patterns of economic activity, and generate a substantial tax

 © 2004 Elsevier Inc. All rights reserved.

revenue stream. These impacts are large enough to affect macroeconomic outcomes such as consumption, investment, fiscal balances, inflation, interest rates, and trade.

2. ENVIRONMENTAL TAXES AND CONVENTIONAL TAXES

The concept of an environmental tax is grounded in the basic intuition that an economic activity that imposes external cost (i.e., the social cost is greater than the private cost) will be pursued at a level higher than is socially optimal. Therefore, economic welfare can be improved by curtailing the activity, such as the production of a good that imposes pollution costs on individuals or other firms. A tax on the offending activity will discourage it, and if set optimally the tax will reduce the level of the activity to the point at which marginal social cost is equivalent to marginal social benefit. Casting the climate change issue in terms of external costs of fossil fuel consumption motivates the imposition of a carbon tax as a potential policy response. Assuming that CO_2 emissions cause economic harm, then a tax on CO_2 emissions represents a classic externality tax.

The assessment of carbon taxes raises issues similar to those that arise in the traditional analysis of conventional taxes (e.g., commodity excise taxes or income taxes), including efficiency, incidence, equity, and administrative burden. However, environmental taxes are fundamentally different from conventional taxes in one key respect: They are intended to discourage harmful behavior. Conventional taxes, in contrast, are levied on activities or transactions that normally create value or enhance economic welfare, such as providing labor or capital. To the extent that conventional taxes lower the returns from labor or investment, and thereby discourage employment, saving, or capital formation, they reduce overall economic welfare.

The distinction between conventional and environmental taxes is colloquially referred to as the difference between taxing economic "goods" and taxing economic "bads." The degree to which traditional economic analysis frameworks are modified to account for this distinction can significantly affect how one interprets the results of carbon tax analyses. The fundamental insight that environmental taxes can improve economic welfare while conventional taxes can reduce economic welfare also has motivated recent analyses of a "tax shift" in which the revenues from environmental taxes are used as a substitute for revenues from conventional distortionary taxes.

3. CARBON TAXES AS ENVIRONMENTAL POLICY

Evaluations of carbon taxes as an environmental policy instrument compare carbon taxes to other ways of reducing CO_2 emissions. These include such targeted measures as efficiency standards, including establishing appliance efficiency standards or fuel economy standards for vehicles, providing tax credits for specific investments, removing subsidies, or providing consumer information. An alternative economywide intervention into fossil fuel markets—tradable CO_2 permits—is related to a carbon tax and will be discussed in greater detail. The primary criteria used to judge the relative desirability of carbon taxes are the cost-effectiveness of CO_2 reductions, the predictability of CO_2 reductions, overall economic efficiency (including such dynamic impacts such as induced technology improvement), incidence (who ultimately pays the tax), and equity (fairness of the ultimate tax burden).

3.1 Carbon Taxes vs Direct Controls or Standards

Economists have long favored the use of economic incentives such as effluent taxes rather than various command-and-control instruments to reduce pollution because taxes are much more likely to induce efficient abatement responses from affected entities. If levied directly on the offending behavior, effluent taxes enable a broader set of economic adjustments that serve to reduce the costs of abatement.

Given the complexity of energy use in the economy, economists generally find that carbon taxes would produce more cost-effective CO_2 emission reductions than a comparable slate of more targeted interventions. This is primarily due to the inability of policymakers to accurately assess the marginal costs of carbon abatement in various sectors and across the myriad economic decisions that affect aggregate CO_2 emission levels. Firms and consumers responding to tax-enhanced price signals will reduce CO_2 emissions in the most efficient manner possible, assuming the absence of other significant market barriers or distortions. Subtle but widespread changes in behavior involving various

substitutions—fuel choice in electricity generation, more investment in energy efficient technology, and reduced demand for energy services—all contribute to cost-effective carbon abatement.

Not all analyses of energy-related policies indicate the advantages of carbon taxes over other instruments. For example, "bottom-up" engineering analysis generally assumes that market barriers inhibit efficient outcomes, and therefore energy prices may not induce cost-effective abatement decisions compared to policies that might remove these barriers or simply mandate specific actions such as efficiency standards. The degree to which these market barriers exist and can be overcome by various policy measures is a matter of dispute among policy analysts.

In addition, emission reductions from efficiency standards can partially be eroded by offsetting behavioral adjustments, such as "snap back" or "rebound" effects and "new source bias." The former occurs when a mandated performance standard reduces energy-related variable costs and thus stimulates higher energy service demand, for example, when efficient air conditioners encourage homeowners to lower thermostats and fuel-efficient cars encourage more driving. The latter can occur if a performance standard raises the price of new technology, encouraging the continued use or repair of older, less efficient versions in order to avoid high replacement costs. This can retard capital replacement rates. If properly implemented, carbon taxes generally do not create offsetting incentives. Therefore, unless significant market barriers or distortions exist, carbon taxes are more likely to induce least-cost abatement.

3.2 Setting the Level of Carbon Taxes

Economic theory suggests that carbon taxes should be set at the level of the external cost of the activity in order to align marginal social cost to marginal social benefit. However, estimating the external cost of 1 ton of CO_2 or carbon emitted into the atmosphere is extremely difficult and controversial, particularly since the long-term nature of the problem raises significant discounting and intergenerational equity issues.

Instead, most analyses and recent climate policy proposals, including the Kyoto Protocol, would target a particular level of CO_2 emissions. This represents a challenge for and highlights a key disadvantage to the carbon tax approach: Namely, the inherent uncertainty regarding the emission reduction expected from a given level of tax. This uncertainty arises from the same lack of information regarding abatement costs and opportunities that contributes to the general cost-effectiveness of the carbon tax approach. Different models and analytic approaches used to estimate the emission impacts of carbon taxes demonstrate an exceptionally wide range of emission reduction results. Although carbon taxes would induce efficient abatement, for example, as measured on a \$/ton of CO_2 reduced basis, uncertainty regarding market responses implies significant difficulty in predicting the resulting emission levels. To the extent that specific emission levels are desired, alternative instruments (e.g., capping CO_2 emissions and distributing tradable permits or allowances) may be preferred.

The primary trade-off between using price instruments such as carbon taxes and quantity targets such as emission caps emerges when supply and demand (or costs and benefits of reductions) are uncertain. Uncertainty in supply or demand of an activity subject to a tax or quantity restriction will manifest itself as relatively variable quantity outcomes under a tax policy and relatively variable price outcomes under a quantity restriction. In the present context, a carbon tax will result in uncertain CO_2 emission levels even though the overall costs of abatement are minimized. A CO_2 emission cap will yield an unpredictable price for CO_2 permits (and uncertainty in overall abatement costs) even while the cost per unit of emission reduced is still minimized through the tradable permit scheme. The inherently large uncertainty regarding the marginal benefit from CO_2 reductions would generally favor a tax approach over an emission cap policy.

3.3 Carbon Taxes, CO_2 Permits, and Revenues

Under emission "cap-and-trade" policies, the overall level of CO_2 is determined in advance, and the equivalent amount of permits to emit are distributed or auctioned to market participants. Under certain conditions, the resultant price of the CO_2 permit can be shown to be exactly the carbon tax that would achieve the same level of reductions, a result that holds whether or not permits are allocated freely or auctioned to market participants. The primary difference between a free allocation and an auction is the distribution of economic burden (incidence) and, in the case of the auction, a revenue stream comparable to the carbon tax revenues.

The revenues expected under a carbon tax and an auctioned permit system (even if set equivalently in

terms of expected value of CO_2 emissions) would also be uncertain in ways that could differ between the two policies. In simple terms, revenues are tax rate (or permit price) multiplied by carbon quantity, and the variability in revenues under each policy depends on the relative range of uncertainty of the two terms of the equation. If the likely range of emissions (in percentage terms) were larger under the carbon tax than the likely range of permit prices (in percentage terms) under the permit auction, then the permit auction revenues would be more predictable. Conversely, carbon tax revenues would be more predictable if the range of emissions were smaller under the tax than the range of likely prices under the permit auction. Because the revenues from carbon taxes (or permit auctions) represent an opportunity to offset some of the economic burden of the policy through the reduction of other taxes, the relative predictability of revenues may be important in the selection of instrument choice.

3.4 Long-Term Efficiency and Technical Change

The economic literature on technical change generally concludes that the private market economy underinvests in R&D because the returns for successful R&D are not fully appropriated by those funding the research. This is the classic rationale for publicly funded R&D, especially at the basic research level. The presence of external costs creates an even larger divergence between social benefit of R&D and the private returns. For example, the private returns to an energy-saving innovation (to the extent that they accrue to the entity that undertook the R&D effort that led to the improvement) would not include the additional social benefit of reducing CO_2 emissions.

Because climate change is an inherently long-term problem, a key determinant of the cost of climate change policy is the potential ability of technology improvement to reduce the costs of shifting toward less carbon-intensive economic activities. A comparative evaluation of climate change policies, therefore, needs to consider the incentives created for research, development, and demonstration of new technologies. Alternative policy approaches may differ in their potential to encourage development of new, lower emitting technologies.

Carbon taxes may have advantages in the long term over other policy instruments to reduce CO_2 emissions because the taxes provide a continual incentive to reduce the costs of carbon abatement. This incentive would increase the private return on R&D into carbon-reducing technologies. Other policy instruments such as targeted standards would create more limited long-term dynamic incentives for improved technology.

4. CARBON TAXES AS FISCAL INSTRUMENTS

Because a carbon tax would collect revenue, it is also important to analyze and compare it with other taxes designed for the same purpose. Traditional environmental tax analysis did not concern itself with the collected revenues since the problems addressed generally were small in relationship to the economy. Because carbon taxes would collect a significant amount of revenue, the ultimate disposition of that revenue becomes a key determinant of the overall economic impact of the carbon tax policy. Therefore, economic analysis of carbon taxes must include explicit and realistic assumptions regarding the disposition of the revenue. There are several choices: Explicitly target revenues for particular public investments, reduce government fiscal deficits, or use carbon tax revenues to offset other taxes. The latter option is sometimes referred to revenue recycling and opens up the prospect for shifting taxes to create net economic gains.

Studies of revenue recycling typically examine the use of proceeds from carbon taxes to offset lower revenues from reducing or eliminating other taxes (i.e., a concept of revenue neutrality). Revenue neutrality is defined as a change in the tax code or spending policy that does not impact the net fiscal balances of government. In the context of carbon taxes, a revenue neutral shift would reduce other taxes by the amount of revenues expected from imposing a carbon tax.

Revenue neutrality does not imply welfare neutrality: Since conventional taxes on labor or investment impose economic penalties beyond their direct cost, there exists an opportunity to use carbon tax revenues to enhance economic efficiency. This is sometimes referred to as the double dividend hypothesis. In its strongest and most simple form, the double dividend hypothesis observes that a carbon tax that enhances economic welfare (e.g., when set at the level of external cost of CO_2 emissions) generates revenue that can be used to lower distortionary taxes on labor or investment. In this way, an environmental tax that cost-effectively reduces the economic damage from pollution (the first dividend) can also serve to increase

economic welfare in another way when the revenues are used to reduce taxes that otherwise would impair economic welfare (the second dividend).

Although carbon tax revenues combined with some decrease in taxes that inhibit investment could, in theory, enhance overall economic outcomes, analyses of double dividend are quite complex. They must simultaneously account for the impact of carbon taxes on emissions and the interactions with all relevant existing features of the tax system to estimate the impact of reducing distortionary taxes. Some studies indicate the potential for this double dividend of decreased environmental harm from imposing carbon taxes and enhanced economic efficiency from using the collected revenues to reduce distortions from conventional taxes. Such results, however, depend on the specific taxes reduced by the carbon tax revenues, which preexisting tax distortions are measured, and the methods used to estimate overall economic welfare.

5. CARBON TAX IMPLEMENTATION

The ubiquitous use of fossil fuels in modern economies makes imposition of a tax directly on measured CO_2 emissions completely infeasible. The direct sources of CO_2, including homes, vehicles, buildings, and factories, are simply too numerous. The nature of fossil fuel combustion, however, suggests an indirect practical alternative that is equivalent to an actual emission tax. The carbon content of commercially available fossil fuels, such as coal, oil and petroleum products, and natural gas, is relatively easy to determine, does not vary significantly within defined classifications, and is essentially proportional to the CO_2 emissions released when combusted. Therefore, a direct tax on fossil fuels in proportion to carbon content will mimic the imposition of a tax on CO_2 emissions.

On a $/Btu basis, a tax on coal would be approximately twice that of natural gas, reflecting its nearly double carbon content, and that for oil would be approximately midway between the two. A carbon tax is a differential excise tax that is independent of the market price of these fuels (i.e., not an *ad valorem* tax based on a percentage of price). Because the tax would not change with the underlying market price, this construction assumes that no other taxes or subsidies affect the value of the carbon tax. Such an assumption does not necessarily reflect the reality of existing differential taxes and subsidies on fossil fuels in most industrialized economies.

5.1 Carbon Tax Base

Between initial extraction and end use, fossil fuels are widely transported, refined or converted into higher valued products such as vehicle fuels and electricity, and subjected to multiple commercial transactions. The issue of where to impose a carbon tax becomes one of balancing administrative efficiency and ensuring complete coverage in order to avoid introducing distortions that would arise from taxing some fossil fuels and not others. Similar issues arise with regard to the distribution of tradable carbon permits under a cap-and-trade policy. Generally, this leads to proposals to tax fossil fuels "upstream" near the point of extraction, importation, or initial commercial transaction. Thus, taxing coal at the mine or processing plant, oil at the wellhead or refinery, and natural gas at the collection field or processing level represents points in the domestic value chain at which the fewest commercial entities would be subject to tax liabilities while ensuring universal coverage. Carbon taxes on imported fuels could be levied at the point of importation, such as harbors and pipelines. Finally, nonfuel uses, such as oil and natural gas used as chemical production feedstocks, can be exempted or subject to tax rebates. Likewise, fossil fuel exports would generally remain untaxed because their disposition does not entail CO_2 emissions from the exporting country.

5.2 Carbon Tax Incidence

Traditional analysis suggests that carbon tax incidence will shift both backward (reducing net prices received by extraction industries) and forward into energy and product prices (increasing prices for fossil fuel users). The degree of forward incidence depends on the elasticity of demand, the cost share of taxed fossil fuels, and pricing rules and regulations in downstream markets. To the extent that intermediate users (e.g., electric generators) and/or end users (e.g., consumers) reduce their fossil fuel use in response to the carbon tax, incidence is avoided and emissions are reduced.

Electricity markets provide a good illustration of the complex responses and incidence impacts. Assuming a supply sector composed of coal- and natural gas-fired generators and nuclear and renewables, a carbon tax on utility fuels would have a variety of impacts. Under least-cost dispatch, coal-fired generation would decline, whereas natural gas generation may increase (i.e., substitute for coal) or

decrease (i.e., nonfossil generation would substitute for coal and overall generation would decline as end-use consumers curtail electricity demand because of higher prices). The degree to which fuel price increases translate into end-use electric rate increases depends on wholesale trading rules and retail regulation. Finally, end users may respond to increased prices by reducing the demand for electricity services. These short-term responses will reduce emissions and shift the tax incidence.

In the longer term, the composition of investment in the generation sector would be influenced by the carbon tax, and electricity consumers will invest in more efficient equipment or substitute other energy forms for electricity. Manufacturers that use power and fuels will experience cost increases that they will attempt to pass on in higher product prices. However, the degree to which manufacturers can shift the tax burden forward depends on the elasticity of demand for their products. All these responses simultaneously will shift tax incidence throughout the economy and determine the amount of CO_2 reduction.

One interesting dimension of carbon tax policy concerns goods in international trade. Unless border adjustments are made (e.g., exempting exports from the estimated carbon tax embedded in the product price and levying taxes on imported products in proportion to their estimated carbon content), a domestic carbon tax can significantly impact trade balances and shift incidence across borders.

5.3 Macroeconomic Impacts

Studies that have estimated the economywide impact of carbon taxes are too numerous to cite or even usefully summarize here. To put the discussion into perspective, a carbon tax or tradable permit scheme to reduce U.S. CO_2 emissions by 10–20% from projected levels approximately a decade in the future is typically estimated to cost a few percent of gross domestic product under generally unfavorable assumptions regarding revenue recycling. Revenue recycling that encourages capital formation tends to lower these figures and in some cases shows a net economic benefit over time. These assessments assume a baseline of no action (i.e., there is no presumed economic savings compared with pursuing CO_2 reductions with less efficient policy instruments).

Some general observations are in order. First, as implied previously, the overall outcomes depend significantly on the assumed mechanism for revenue recycling. Second, the degree of substitutability (between fossil and non-fossil fuels, among different products, between capital and energy, etc.) assumed by the model will determine the amount of CO_2 reduction expected by a given carbon tax and the measured economic impacts. Third, the model structure, temporal horizon, and the manner in which expectations of future economic conditions affect current behavior will influence the estimated level of a carbon tax needed to achieve a given CO_2 reduction target level and the types of economic impacts estimated.

Short-term macroeconomic models often do not represent detailed substitutions in energy markets while tending to emphasize transitional imbalances, such as unemployment or myopic monetary policy reactions, that might flow from a carbon tax. Thus, these models typically estimate high carbon taxes needed to reduce CO_2 emissions and commensurately large economic costs. Despite these limitations, such models shed important light on some of the issues involved in a rapid shift toward a less carbon-intensive economy. Longer term general equilibrium models often provide richer substitution possibilities and assume that the economy will rapidly adjust to the new energy prices. Thus, smaller carbon taxes suffice to reduce CO_2 emissions and the tax policy inflicts less economic harm. Although these models generally gloss over the potential short-term dislocations from imposing an aggressive carbon tax policy, they are helpful in showing how the economy might eventually adjust to a less carbon-intensive economic structure with modest long-term economic penalties. Given the extraordinary complexity involved with estimating how billions of economic decisions might be influenced by carbon taxes and revenue recycling policies, it is probably wise to examine the results of different models to gain understanding of how actual policies might affect real economic outcomes.

5.4 Equity

Environmental policies of this magnitude, especially ones that involve tax structures, naturally raise equity issues. Although economics as a discipline can only make limited moral or ethical judgments, some concepts of equity have economic dimensions. A few observations are offered here.

First, there is a tension between the measurement of contemporaneous cost burden of CO_2 reductions and the potential for significant climate change damages that might be borne by future generations. These do not easily reduce to arguments over

discount rates; perhaps ethicists or philosophers are better equipped to analyze these issues.

Second, carbon taxes are typically considered to be regressive since in most industrialized economies energy costs represent a higher portion of expenditures or income for lower income families. However, a complete analysis must consider both the impact of existing tax structure and expenditure patterns, and possibly the impact on various income classes of alternative policies to reduce CO_2 emissions.

Finally, equity issues are especially sensitive to revenue recycling assumptions. In many economic analysis frameworks, reducing existing distortionary taxes on capital provides a greater boost to economic growth than equivalent reductions on labor taxes or devising a more neutral transfer based on expanding the personal exemption (e.g., shielding the initial portion of income from income tax). In any case, the revenues from a carbon tax might be used to offset perceived inequities in the current tax system. The equity implications of any overall carbon tax and revenue recycling scheme must be evaluated carefully and completely before drawing conclusions.

6. CONCLUSION

Although any policy prescription to combat climate change has advantages and disadvantages, on balance, carbon taxes appear to have several distinct advantages over other approaches to deal with climate change. Economic incentive approaches, such as taxes or tradable permits, will induce cost-effective abatement responses. The incidence of such policies is complicated but generally operates to reduce CO_2 emissions in cost-effective ways in both the short and long term. The revenues generated from carbon tax (or permit systems) can be used to reduce distortions or inequities or both in current tax schemes; such revenue recycling will lower and possibly eliminate the economic cost associated with reducing CO_2 emissions.

SEE ALSO THE FOLLOWING ARTICLES

Biomass for Renewable Energy and Fuels • Biomass: Impact on Carbon Cycle and Greenhouse Gas Emissions • Carbon Capture and Storage from Fossil Fuel Use • Carbon Sequestration, Terrestrial • Climate Change and Energy, Overview • Climate Change: Impact on the Demand for Energy • Energy Efficiency and Climate Change • Greenhouse Gas Abatement: Controversies in Cost Assessment • Greenhouse Gas Emissions from Energy Systems, Comparison and Overview • Taxation of Energy

Further Reading

Bohm, P., and Russell, C. (1985). Comparative analysis of alternative policy instruments. *In* "Handbook of Natural Resource and Energy Economics" (A. V. Kneese and J. L. Sweeney, Eds.), Vol. 1, pp. 395–460. North-Holland, Amsterdam.

Bovenberg, A. L., and de Mooij, R. A. (1994). Environmental levies and distortionary taxation. *Am. Econ. Rev.* **84,** 1085–1089.

Bovenberg, A. L., and Goulder, L. H. (1996). Optimal environmental taxation in the presence of other taxes: general equilibrium analyses. *Am. Econ. Rev.* **86,** 985–1000.

Boyd, R., Krutilla, K., and Viscusi, W. K. (1995). Energy taxation as a policy instrument to reduce CO_2 emissions: A net benefit analysis. *J. Environ. Econ. Management* **29,** 1–24.

Dewees, D. N. (1983). Instrument choice in environmental policy. *Econ. Inquiry* **21,** 53–71.

Gaskins, D., and Weyant, J. (eds.). (1995). "Reducing Global Carbon Emissions: Costs and Policy Options." Energy Modeling Forum, Stanford, CA.

Goulder, L. H. (1995). Environmental taxation and the "double dividend": A reader's guide. *Int. Tax Public Finance* **2,** 157–183.

Goulder, L. H., and Schneider, S. H. (1999). Induced technological change and the attractiveness of CO_2 abatement policies. *Resour. Energy Econ.* **21**(3/4), 211–253.

Goulder, L. H., Parry, I. W. H., and Burtraw, D. (1997). Revenue-raising vs. other approaches to environmental taxation: The critical significance of pre-existing tax distortions. *Rand J. Econ.* **28,** 708–731.

Goulder, L. H., Parry, I. W. H., Williams, R. C., and Burtraw, D. (1999). The cost effectiveness of alternative instruments for environmental protection in a second-best setting. *J. Public Econ.* **72,** 329–360.

Grübler, A., Nakicenovic, N., and Nordhaus, W. D. (eds.). (2002). "Technological Change and the Environment." Resources for the Future, Washington, DC.

Parry, I. W. H. (1995). Pollution taxes and revenue recycling. *J. Environ. Econ. Management* **29,** S64–S77.

Parry, I. W. H. (1997). Environmental taxes and quotas in the presence of distorting taxes in factor markets. *Resour. Energy Econ.* **19,** 203–220.

Parry, I. W. H., and Bento, A. M. (2000). Tax deductions, environmental policy, and the "double dividend" hypothesis. *J. Environ. Econ. Management* **39,** 67–96.

Parry, I. W. H., Williams, R. C., and Goulder, L. H. (1999). When can carbon abatement policies increase welfare? The fundamental role of distorted factor markets. *J. Environ. Econ. Management* **37,** 52–84.

Pizer, W. A. (1999). The optimal choice of climate change policy in the presence of uncertainty. *Resour. Energy Econ.* **21**(3/4), 255–287.

Repetto, R., and Austin, D. (1997). "The Costs of Climate Protection: A Guide for the Perplexed." World Resources Institute, Washington, DC.

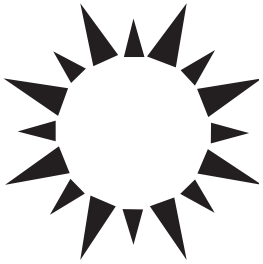
Repetto, R., Dower, R. C., Jenkins, R., and Geoghegan, J. (1992). "Green Fees: How a Tax Shift Can Work for the Environment and the Economy." World Resources Institute, Washington, DC.

Sterner, T. (2002). "Policy Instruments for Environmental and Natural Resource Management." Resources for the Future, Washington, DC.

Terkla, D. (1984). The efficiency value of effluent tax revenues. *J. Environ. Econ. Management* **11**(2), 107–123.

Cement and Energy

ERNST WORRELL
Lawrence Berkeley National Laboratory
Berkeley, California, United States



1. Introduction
2. Cement
3. Cement Production
4. Energy Use in Cement Making
5. Energy Conservation Opportunities in Cement Making

Glossary

blended cement Cement containing less clinker than typical portland cement and in which the clinker is replaced with other cementious materials, such as blast furnace slag or fly ash.
calcination The reaction of limestone at high temperatures that produces calcium oxide and carbon dioxide.
cement A powder that consists of clinker and other materials and that can be mixed with sand and water to make concrete.
clinker A hard substance produced in a high-temperature process that consists of a mixture of calcined limestone and other minerals.
rotary kiln A rotating tube in which limestone and other ingredients are heated to high temperatures to calcine the limestone and produce clinker.

Cement is one of the most important building materials in the world. It is mainly used for the production of concrete. Cement is produced in almost every country, although a few countries dominate the global production (e.g., China, India, the United States, and Japan). Cement production is a highly energy-intensive process. The cement industry consumes more than 2% of world energy use and contributes approximately 5% of global anthropogenic CO_2 emissions. Cement is made in two basic steps: the making of clinker from limestone and the mixing of clinker with other materials. Energy is mainly consumed in the first production step because of the energy-intensive calcination process. Energy conservation opportunities include energy efficiency improvement, new processes, application of waste fuels, and increased use of additives in cement making.

1. INTRODUCTION

Cement is one of the most important building materials in the world. It is mainly used for the production of concrete. Concrete is a mixture of inert mineral aggregates (e.g., sand, gravel, crushed stones, and cement). Cement consumption and production are closely related to construction activity and therefore to general economic activity. Global cement production is estimated at 1650 million tonnes, and it is one of the most produced materials in the world.

Due to the importance of cement as a construction material and the geographic abundance of the main raw material (i.e., limestone), cement is produced in virtually all countries. The widespread production is also due to the relative low price and high density of cement, which limits ground transportation due to the relatively high cost. Ground deliveries generally do not exceed distances of more than 150–200 km. Bulk sea transport is possible but very limited. Generally, the international trade (excluding plants located on borders) is limited compared to the global production. The major producers are large countries (e.g., China and India) or regions and countries with a developed economy, such as the European Union, the United States, and Japan.

Cement production is a highly energy-intensive process. Energy consumption by the cement industry is estimated to be approximately 2% of the global primary energy consumption or approximately 5% of the total global industrial energy consumption. Due to the dominant use of carbon-intensive fuels such as coal in clinker making, the cement industry is a major source of CO_2 emissions. In addition to energy consumption, during clinker making CO_2 is emitted from the calcining process. Due to both

 © 2004 Elsevier Inc. All rights reserved.

emission sources and the emissions from electricity production, the cement industry is a major source of carbon emissions, and it should be considered in the assessment of carbon emission-reduction options.

The high temperatures in the combustion process also contribute to relatively high NO_x emissions, whereas sulfur in the coal may be emitted as SO_x. Furthermore, the cement industry is a source of particulate matter emissions from both quarrying and material processing. However, in modern cement plants NO_x emissions are reduced by the use of precalciners, and filters reduce particulate and SO_x emissions.

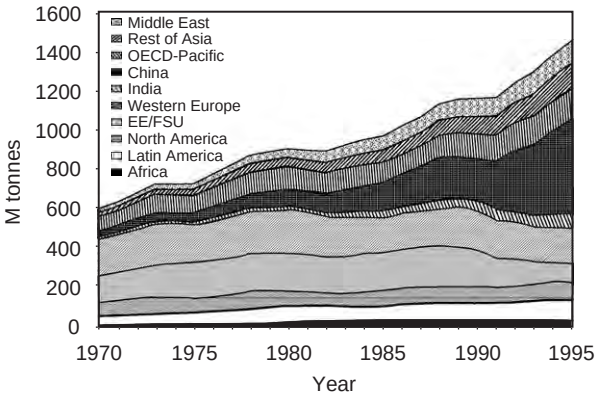


FIGURE 1 Historical production trends for cement in 10 regions of the world.

2. CEMENT

Cement is an inorganic, nonmetallic substance with hydraulic binding properties. Mixed with water, it forms a paste that hardens due to the formation of hydrates. After hardening, the cement retains its strength. In the common cement types (e.g., portland), the hardening is mainly due to the formation of calcium silicate hydrates. Cement is made from a mixture of raw materials, mainly calcium oxide, silicon dioxide, aluminum oxide, and iron (III) oxide. These compounds react during a heating process, forming the intermediate "clinker." This clinker is ground to a desired fineness together with additives to form cement. There are numerous cement types due to the use of different sources of calcium oxide and different additives to regulate properties. The most important sources of calcium oxide are limestone, blast furnace slag, and fly ash.

Global cement production increased from 594 Mt in 1970 to 1650 Mt in 2001. Consumption and production of cement are cyclical, concurrent with business cycles. Historical production trends for 10 world regions are shown in Fig. 1. The regions with the largest production levels are China, Europe, the Organization for Economic Cooperation and Development (OECD)-Pacific region, the rest of Asia, and the Middle East.

As a region, China clearly dominates current world cement production, manufacturing 595 Mt in 2001, more than twice as much as the next largest region. Cement production in China has increased dramatically, with an average annual growth rate of 12%. In many respects, China's cement industry is unique due to the large number of plants, the broad range of ownership types, and the variety of production technologies. By late 1994, there were more than 7500 cement plants throughout the country. Chinese plants tend to be small, approximately one-10th the size of the average plant in the United States. In India, cement production has increased at an average annual rate of 6.6%. Currently, India is the second largest cement producer in the world, producing approximately 100 Mt annually. In 1995, the OECD-Pacific region produced 154 Mt of cement, predominantly in Japan and South Korea. South Korean cement production increased at a high rate to 52 Mt in 2001. The rest of Asia experienced a high average annual growth of 8% between 1970 and 1995, increasing to 130 Mt in 1995. The largest producing countries in this region are Thailand, Indonesia, and Taiwan.

Cement production in North America increased slightly to 91 Mt in 2001. Recent economic growth has led to increased cement demand. Brazil (40 Mt) and Mexico (35 Mt) dominate cement production in Latin America; together they are responsible for more than 50% of the production in this region.

Cement production in Western Europe was relatively stable. The largest cement-producing countries in this region are Germany (40 Mt), Italy (36 Mt), and Spain (30 Mt). In the Eastern Europe/former Soviet Union region, cement production increased at an average rate of 2.3% per year between 1970 and 1988. After the breakup of the Soviet Union and the major restructuring that began in that region in 1988, production levels decreased by 13% per year on average between 1990 and 1995. Countries of the former Soviet Union with the highest production levels are the Russian Federation, Ukraine, and Uzbekistan.

Production of cement in the Middle East also increased rapidly, averaging 7% per year. The largest cement-producing countries in this region are

Turkey, Egypt, Iran, and Saudi Arabia. Africa showed relatively high growth, with an increase in production to approximately 50 Mt at an average annual rate of 4 to 5%. The largest cement-producing African countries are South Africa, Algeria, and Morocco, although none are among the top 20 cement-producing countries worldwide.

3. CEMENT PRODUCTION

3.1 Mining and Quarrying

The most common raw materials used for cement production are limestone, chalk, and clay. Most commonly, the main raw material, limestone or chalk, is extracted from a quarry adjacent or very close to the plant. Limestone and chalk provide the required calcium oxide and some of the other oxides, whereas clay, shale, and other materials provide most of the silicon, aluminum, and iron oxides required for portland cement. Limestone and chalk are most often extracted from open-pit quarries, but underground mining can be employed. The collected raw materials are selected, crushed, ground, and proportioned so that the resulting mixture has the desired fineness and chemical composition for delivery to the pyroprocessing systems (Fig. 2). It is often necessary to increase the content of silicon oxides and iron oxides by adding quartz sand and iron ore. The excavated material is transported to a crusher. Normally, first a jaw or gyratory crusher, followed by a roller or hammer mill, is used to crush the limestone. The crushed material is screened and stones are returned. At least 1.5 tons of raw material is required to produce 1 ton of portland cement.

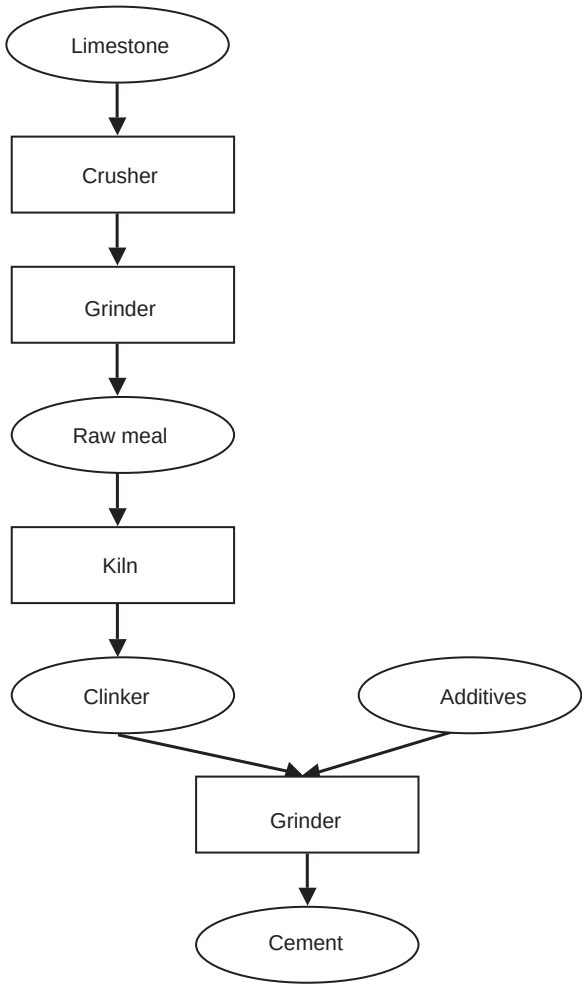


FIGURE 2 Simplified process schematic for cement making.

3.2 Kiln Feed Preparation

Raw material preparation is an electricity-intensive production step generally requiring approximately 25–35 kWh/tonne raw material, although it can require as little as 11 kWh/tonne. The raw materials are further processed and ground. The grinding differs with the pyroprocessing process used. The raw materials are prepared for clinker production into a "raw meal" by either dry or wet processing. In dry processing, the materials are ground into a flowable powder in ball mills or in roller mills. In a ball (or tube) mill, steel-alloy balls are used to decrease the size of the raw material pieces in a rotating tube. Rollers on a round table perform this task of comminution in a roller mill. The raw materials may be further dried from waste heat from the kiln exhaust, clinker cooler vent, or auxiliary heat from an air heater before pyroprocessing. The moisture content in the (dried) feed of the dry kiln is typically 0.5% (0–0.7%).

When raw materials contain more than 20% water, wet processing may be preferable. In the wet process, raw materials are ground with the addition of water in a ball mill to produce a slurry typically containing 24–48% water. Various degrees of wet processing [e.g., semiwet (moisture content of 17–22%] are used to reduce fuel consumption in the kiln.

3.3 Clinker Production (Pyroprocessing)

Clinker is the main constituent of portland cement. Clinker consists of calcium oxide and other mineral

oxides (iron, aluminum, and silicon) and has cementious activity when reacted with water. Clinker is produced by pyroprocessing in large kilns. These kiln systems evaporate the free water in the meal, calcine the carbonate constituents (calcination), and form cement minerals (clinkerization). In the chemical transformation, calcium carbonate is transformed into calcium oxide, leading to a reduction in the original weight of raw materials used. Furthermore, cement kiln dust may be emitted to control the chemical composition of the clinker. Clinker production is the most energy-intensive stage in cement production, accounting for more than 90% of total industry energy use and virtually all the fuel use.

The main kiln type used throughout the world is the large-capacity rotary kiln. In these kilns, a tube with a diameter up to 8 m is installed at a 3–4° angle and rotates one to three times per minute. The ground raw material, fed into the top of the kiln, moves down the tube toward the flame. In the sintering (or clinkering) zone, the combustion gas reaches a temperature of 1800–2000°C. Although many different fuels can be used in the kiln, coal is the primary fuel in most countries.

In a wet rotary kiln, the raw meal typically contains approximately 36% moisture. These kilns were developed as an upgrade of the original long dry kiln to improve the chemical uniformity in the raw meal. The water is first evaporated in the kiln in the low-temperature zone. The evaporation step requires a long kiln. The length-to-diameter ratio may be up to 38, with lengths up to 230 m. The capacity of large units may be up to 3600 tonnes of clinker per day. Fuel use in a wet kiln can vary between 5.3 and 7.1 GJ/tonne clinker. The variation is due to the energy requirement for the evaporation and, hence, the moisture content of the raw meal. Originally, the wet process was preferred because it was easier to grind and control the size distribution of the particles in a slurry form. The need for the wet process was reduced by the development of improved grinding processes.

In a dry kiln, feed material with much lower moisture content (0.5%) is used, thereby reducing the need for evaporation and reducing kiln length. The dry process was first developed in the United States and consisted of a long dry kiln without preheating or with one-stage suspension preheating. Later, multistage suspension preheaters (i.e., a cyclone) or shaft preheaters were developed. Additionally, precalciner technology has recently been developed in which a second combustion chamber has been added to a conventional preheater that allows for further reduction of kiln energy requirements. The typical fuel consumption of a dry kiln with four- or five-stage preheating varies between 3.2 and 3.5 GJ/tonne clinker. A six-stage preheater kiln can theoretically use as low as 2.9–3.0 GJ/tonne clinker. The most efficient preheater, precalciner kilns, use approximately 2.9 GJ/tonne clinker. Kiln-dust bypass systems may be required in kilns in order to remove alkalis, sulfates, or chlorides. Such systems lead to additional energy losses since sensible heat is removed with the dust. Figure 3 depicts the typical specific energy consumption for different kiln types in use throughout the world.

Once the clinker is formed, it is cooled rapidly to ensure the maximum yield of alite (tricalcium silicate), an important component for the hardening properties of cement. The main cooling technologies are either the grate cooler or the tube or planetary cooler. In the grate cooler, the clinker is transported over a reciprocating grate passing through a flow of air. In the tube or planetary cooler, the clinker is cooled in a countercurrent air stream. The cooling air is used as combustion air for the kiln.

In developing countries, shaft kilns are still in use. These are used in countries with a lack of infrastructure to transport raw materials or cement or for the production of specialty cements. Today, most vertical shaft kilns are in China and India. In these countries, the lack of infrastructure, lack of capital, and power shortages favor the use of small-scale local cement plants. In China, this is also the consequence of the industrial development pattern, in which local township and village enterprises have been engines of rural industrialization, leading to a

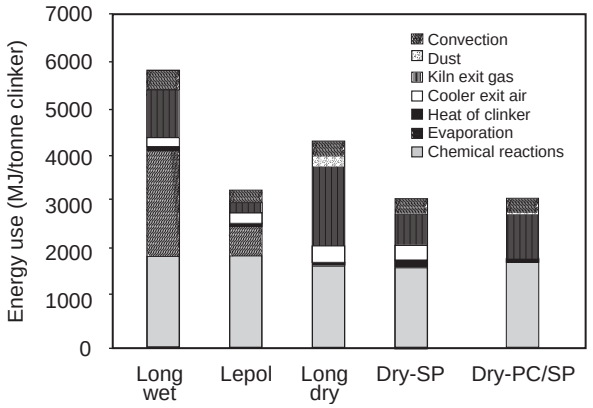


FIGURE 3 Energy consumption and losses for the major kiln types: wet process, Lepol or semiwet, long dry, dry process with four-stage suspension preheating (SP), and dry process with four-stage suspension preheating and precalcining (PC/SP).

substantial share of shaft kilns for the total cement production. Regional industrialization policies in India also favor the use of shaft kilns, in addition to the large rotary kilns in major cement-producing areas. In India, shaft kilns represent a growing portion of total cement production, approximately 10% of the 1996 production capacity. In China, the share is even higher, with an estimated 87% of output in 1995. Typical capacities of shaft kilns vary between 30 and 180 tonnes of clinker per day.

The principle of all shaft kilns is similar, although design characteristics may vary. The shaft kiln is most often cone shaped. The pelletized material travels from top to bottom, through the same zones as those of rotary kilns. The kiln height is determined by the time needed for the raw material to travel through the zones and by operational procedures, pellet composition, and the amount of air blown. Shaft kilns can achieve reasonably high efficiency through interaction with the feed and low energy losses through exhaust gases and radiation. Fuel consumption in China varies between 3.7 and 6.6 GJ/tonne clinker for mechanized kilns, with the average estimated at 4.8 GJ/tonne clinker. In India, presumably, shaft kilns use between 4.2 and 4.6 GJ/tonne clinker. The largest energy losses in shaft kilns are due to incomplete combustion. Electricity use in shaft kilns is very low.

3.4 Finish Grinding

After cooling, the clinker is stored in the clinker dome or silo. The material-handling equipment used to transport clinker from the clinker coolers to storage and then to the finish mill is similar to that used to transport raw materials (e.g., belt conveyors, deep bucket conveyors, and bucket elevators). To produce powdered cement, the nodules of cement clinker are ground. Grinding of cement clinker, together with additives (gypsum and/or anhydrite; 3–5% to control the setting properties of the cement) can be done in ball mills, roller mills, or roller presses. Combinations of these milling techniques are often applied. Coarse material is separated in a classifier and returned for additional grinding.

Power consumption for grinding depends on the required fineness of the final product and the additives used (in intergrinding). Electricity use for raw meal and finish grinding heavily depends on the hardness of the material (limestone, clinker, and pozzolan extenders) and the desired fineness of the cement as well as the amount of additives. Blast furnace slags are more difficult to grind and hence use more grinding power, between 50 and 70 kWh/tonne for a 3500 Blaine (expressed in cm^2/g). (Blaine is a measure of the total surface of the particles in a given quantity of cement, or an indicator of the fineness of cement. The higher the Blaine, the more energy required to grind the clinker and additives to the desired fineness.) Traditionally, ball or tube mills are used in finish grinding, but many plants use vertical roller mills. In ball and tube mills, the clinker and gypsum are fed into one end of a horizontal cylinder and partially ground cement exits the other end. Modern ball mills may use 32–37 kWh/tonne for cements with a Blaine of 3500.

Modern state-of-the-art concepts are the high-pressure roller mill and the horizontal roller mill (e.g., Horomill), which use 20–50% less energy than a ball mill. The roller press is a relatively new technology and is more common in Western Europe than in North America. Various new grinding mill concepts are under development or have been demonstrated.

Finished cement is stored in silos, tested and filled into bags, or shipped in bulk on bulk cement trucks, railcars, barges, or ships. Additional power is consumed for conveyor belts and packing of cement. The total consumption for these purposes is generally low and not more than 5% of total power use. Total power use for auxiliaries is estimated to be approximately 10 kWh/tonne clinker. The power consumption for packing depends on the share of cement packed in bags.

4. ENERGY USE IN CEMENT MAKING

The theoretical energy consumption for producing cement can be calculated based on the enthalpy of formation of 1 kg of portland cement clinker, which is approximately 1.76 MJ. This calculation refers to reactants and products at 25°C and 0.101 MPa. In addition to the theoretical minimum heat requirements, energy is required to evaporate water and to compensate for the heat losses. Heat is lost from the plant by radiation or convection and, with clinker, emitted kiln dust and exit gases leaving the process. Hence, in practice, energy consumption is higher. The kiln is the major energy user in the cement-making process. Energy use in the kiln basically depends on the moisture content of the raw meal. Figure 3 provides an overview of the heat requirements of different types of kilns. Most electricity is consumed in the grinding of the raw materials and finished cement. Power consumption for a rotary kiln

is comparatively small, generally approximately 17–23 kWh/tonne clinker (including the cooler and preheater fans). Additional power is consumed for conveyor belts and packing of cement. Total power use for auxiliaries is estimated at approximately 10 kWh/tonne clinker. Table I summarizes the typical energy consumption for the different processing steps and processes used.

There is a relatively large difference in the energy consumption per tonne of cement produced in different countries. International comparisons of energy intensity have shown that part of the difference can be explained by the types of cement produced. However, a large part of the difference is due to the use of less energy-efficient equipment. For example, in Eastern Europe and the former Soviet Union countries, as well as the United States, a relatively large part of the clinker is produced in inefficient wet process kilns. The differences in energy efficiency, fuels used, and cement types produced also affect CO_2 emissions. Figure 4 shows the CO_2 emissions for cement production in various regions of the world.

TABLE I

Energy Consumption in Cement-Making Processes and Process Types[a]

Process step	Fuel use (GJ/tonne product)	Electricity use (kWh/tonne product)	Primary energy (GJ/tonne cement)[b]
Crushing			
Jaw crusher		0.3–1.4	0.02
Gyratory crusher		0.3–0.7	0.02
Roller crusher		0.4–0.5	0.02
Hammer crusher		1.5–1.6	0.03
Impact crusher		0.4–1.0	0.02
Raw meal grinding			
Ball mill		22	0.39
Vertical mill		16	0.28
Hybrid systems		18–20	0.32–0.35
Roller press, Integral		12	0.21
Roller press, pregrinding		18	0.32
Clinker kiln			
Wet	5.9–7.0	25	6.2–7.3
Lepol	3.6	30	3.9
Long dry	4.2	25	4.5
Short dry suspension preheating	3.3–3.4	22	3.6–3.7
Short dry, preheater and precalciner	2.9–3.2	26	3.2–3.5
Shaft	3.7–6.6	n.a.[d]	3.7–6.6
Finish grinding[c]			
Ball mill		55	0.60
Ball mill/separator		47	0.51
Roller press/ball mill/separator		41	0.45
Roller press/ separator/ball mill		39	0.43
Roller press/ separator		28	0.31

[a] Specific energy use is given per unit of throughput in each process. Primary energy is calculated per tonne of cement, assuming portland cement (containing 95% clinker), including auxiliary power consumption.

[b] Primary energy is calculated assuming a net power-generation efficiency of 33% (LHV).

[c] Assuming grinding of Portland cement (95% clinker and 5% gypsum) at a fineness of 4000 Blaine.

[d] n.a., not applicable.

5. ENERGY CONSERVATION OPPORTUNITIES IN CEMENT MAKING

Improvement in energy efficiency reduces the costs of producing cement. Improvement may be attained by applying more energy-efficient process equipment and by replacing old installations with new ones or shifting to completely new types of cement-production processes. By far the largest proportion of energy consumed in cement manufacture consists of fuel that is used to heat the kiln. Therefore, the greatest gain in reducing energy input may come from improved fuel efficiency. The dry process is generally more energy efficient than the wet process. The processes are exchangeable to a large extent but may be limited by the moisture content of the available raw material. The main opportunities for improving efficiency in the kiln are the conversion to more energy-efficient process variants (e.g., from a wet process to a dry process with preheaters and precalciner), optimization of the clinker cooler, improvement of preheating efficiency, improvement of burners, as well as improvement of process control and management systems. Electricity use can be reduced by improved grinding systems, high-efficiency classifiers, high-efficiency motor systems, and process control systems. Table II provides an overview of the different energy-efficient technologies available to the cement industry.

Several studies have demonstrated the existence of cost-effective potentials for energy efficiency

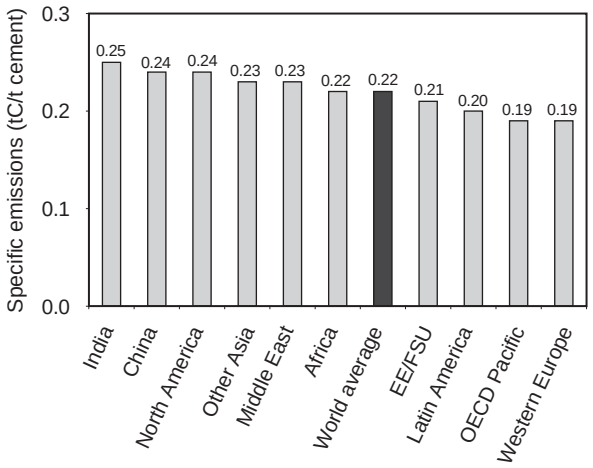


FIGURE 4 Carbon intensity of cement production in different regions.

improvement in the cement industry. In China, various programs have developed technologies to improve the efficiency of shaft kilns by increasing mechanization, insulation, bed distribution, as well as control systems. They demonstrated an energy efficiency improvement potential of 10–30% for all shaft kilns. A study of the Indian cement industry found a technical potential for energy efficiency improvement of approximately 33% with currently commercially available technology. Incorporating future technologies, the energy savings potential is estimated to be approximately 48%. However, the economic potential for energy efficiency improvement is estimated to be 24% of total primary energy use (using a discount rate of 30%). Focusing on commercially available technology, 29 energy-efficient technologies were identified that could be adopted to some extent by the U.S. cement industry. Together, these have a technical potential for energy efficiency improvement of 40%. However, the economic potential (using a discount rate of 30%) is estimated to be only 11% due to the high capital costs and low energy costs in the United States.

The production of clinker is the most energy-intensive step in the cement-manufacturing process and causes large process emissions of CO_2. In blended cement, a portion of the clinker is replaced with industrial by-products, such as coal fly ash (a residue from coal burning) or blast furnace slag (a residue from ironmaking), or other pozzolanic materials (e.g., volcanic material). These products are blended with the ground clinker to produce a homogeneous product—blended cement. Blended cement has different properties compared to portland cement (e.g., setting takes longer but ultimate strength is higher).

The current application of additives in cement making varies widely by country and region. Whereas in Europe the use of blended cements is quite common, it is less common in North America. The relative importance of additive use is indicated by the clinker:cement (C:C) ratio of cement production in a specific country. Portland cement has a C:C ratio of 0.95, whereas blast furnace slag cement has a C:C ratio as low as 0.35. Countries such as the United States, Canada, and the United Kingdom have high C:C ratios, indicating the dominance of portland cement, whereas countries such as Belgium, France, and the former Soviet Union have lower C:C ratios, indicating the relatively larger use of blended cements. Because no international sources collect clinker production data, it is not possible to accurately estimate the current practices in all cement-producing countries. The major barriers to further application of blended cements do not seem to be supply or environmental issues but rather existing product standards and specifications and building codes.

The potential for application of blended cements depends on the current application level, the availability of blending materials, and standards and legislative requirements. The global potential for CO_2 emission reduction by producing blended cement is estimated to be at least 5% of total CO_2 emissions from cement making (56 Mt CO_2) but may be as high as 20%. The potential savings will vary by country and region. The potential for energy savings for 24 countries in the OECD, Eastern Europe, and Latin America was estimated to be up to 29%, depending on the availability of blending materials and the structure of the cement market. The average emission reduction for all countries (producing 35% of world cement in 1990) was estimated to be 22%.

The costs of blending materials depend on the transportation costs and may be \$15–30 per ton for fly ash and approximately \$24 per ton for blast furnace slag. Shipping costs may significantly increase the price, depending on distance and shipping mode. The prices are still considerably lower than the production costs of cement, estimated to be approximately \$36 per ton (1990) in the United States.

Additives such as fly ash contain high concentrations of heavy metals, which may leach to the environment in unfavorable conditions. No negative environmental effects of slag and fly ash addition in cement have been found. The use of nonferrous slags seems to be limited to slag contents of 15% by mass. However, fly ash and blast furnace slag may be

TABLE II

Energy-Efficient Practices and Technologies in Cement Production

Raw materials preparation	
Efficient transport systems (dry process)	
Slurry blending and homogenization (wet process)	
Raw meal blending systems (dry process)	
Conversion to closed-circuit wash mill (wet process)	
High-efficiency roller mills (dry cement)	
High-efficiency classifiers (dry cement)	
Fuel preparation: roller mills	
Clinker production (wet)	Clinker production (dry)
Energy management and process control	Energy management and process control
Seal replacement	Seal replacement
Kiln combustion system improvements	Kiln combustion system improvements
Kiln shell heat loss reduction	Kiln shell heat loss reduction
Use of waste fuels	Use of waste fuels
Conversion to modern grate cooler	Conversion to modern grate cooler
Refractories	Refractories
Optimize grate coolers	Heat recovery for power generation
	Low-pressure drop cyclones for suspension preheaters
Conversion to preheater, precalciner kilns	Optimize grate coolers
Conversion to semidry kiln (slurry drier)	Addition of precalciner to preheater kiln
Conversion to semiwet kiln	Long dry kiln conversion to multistage preheater kiln
Efficient kiln drives	Long dry kiln conversion to multistage preheater, precalciner kiln
Oxygen enrichment	Efficient kiln drives
	Oxygen enrichment
Finish grinding	
Energy management and process control	
Improved grinding media (ball mills)	
High-pressure roller press	
High-efficiency classifiers	
General measures	
Preventative maintenance (insulation, compressed air system)	
High-efficiency motors	
Efficient fans with variable speed drives	
Optimization of compressed air systems	
Efficient lighting	
Product and feedstock changes	
Blended cements	
Limestone cement	
Low-alkali cement	
Use of steel slag in kiln (CemStar)	
Reducing fineness of cement for selected uses	

considered hazardous wastes under environmental legislation in some countries, limiting the use of fly ash to specified companies. In the United States, fly ash falls under the Resource Conservation and Recovery Act, which gives states the jurisdiction to define fly ash as a hazardous waste. In practice, the state regulation varies widely throughout the United States, limiting the reuse of fly ash.

SEE ALSO THE FOLLOWING ARTICLES

Aluminum Production and Energy • External Costs of Energy • Greenhouse Gas Emissions from Energy Systems, Comparison and Overview • Plastics Production and Energy • Steel Production and Energy

Further Reading

Alsop, P. A., and Post, J. W. (1995). "The Cement Plant Operations Handbook." Tradeship, Dorking, UK.

Cembureau (1996). "World Cement Directory." Cembureau, Brussels.

Cembureau (1997). "Best Available Techniques for the Cement Industry." Cembureau, Brussels.

Duda, W. H. (1985). "Cement Data Book, International Process Engineering in the Cement Industry." 3rd ed. Bauverlag, Wiesbaden, Germany.

Feng, L., Ross, M., and Wang, S. (1995). Energy efficiency of China's cement industry. *Energy* **20,** 669–681.

Hargreaves, D. (ed.). (1996). "The Global Cement Report," 2nd ed. Tradeship, Surrey, UK.

Hendriks, C. A., Worrell, E., Price, L., Martin, N., and Ozawa Meida, L. (1999). The reduction of greenhouse gas emissions from the cement industry, Report No. PH3/7. IEA Greenhouse Gas R&D Program, Gloucestershire, UK.

U.S. Geological Survey (various years). "Minerals Yearbook." U.S. Geological Survey, Washington, DC (http://minerals.er.usgs.gov/minerals).

Von Seebach, H. M., Neumann, E., and Lohnherr, L. (1996). State-of-the-art of energy-efficient grinding systems. *ZKG Int.* **49,** 61–67.

Worrell, E., Martin, N., and Price, L. (2000). Potentials for energy efficiency improvement in the U.S. cement industry. *Energy* **25,** 1189–1214.

Worrell, E., Price, L., Martin, N., Hendriks, C. A., and Ozawa Meida, L. (2001). Carbon dioxide emissions from the global cement industry. *Annu. Rev. Energy Environ.* **26,** 203–229.

City Planning and Energy Use

HYUNSOO PARK and CLINTON ANDREWS
Rutgers University
New Brunswick, New Jersey, United States

1. **Scope of City Planning**
2. **Siting Energy Facilities**
3. **Financing Energy Infrastructure**
4. **Influence of Land Use Patterns on Energy**
5. **Influence of Urban/Architectural Design on Energy Policies and Practices**
6. **Centralized vs Distributed Energy Supplies**
7. **Integrated vs Segmented Energy Planning**
8. **Energy for the World's Burgeoning Megacities**
9. **Conclusion**

Glossary

alternative dispute resolution (**ADR**) Any techniques, such as arbitration, mediation, early neutral evaluation, and conciliation, for the purpose of consensus building before litigation.

combined heat and power (**CHP**) Systems that generate electricity and thermal energy simultaneously in an integrated manner, so as to capture wasted energy.

distributed energy resources (**DER**) A variety of small, modular electric generation units located near the end user; used to improve the operation of the electricity network by complementing central power.

smart growth A development strategy emphasizing mixed land uses with convenient transportation choices in existing communities, thereby alleviating sprawl and preserving open spaces.

City planning is a future-oriented activity that shapes land use and the built environment by means of designs, regulations, and persuasion. The scope of city planning includes a range of interdependent decisions at the nexus of the private and public spheres: transportation and utility infrastructure investments, allowable land uses, and the locations and characteristics of housing, retail stores, offices, schools, hospitals, factories, and even energy-producing facilities. Although planners rarely focus explicitly on energy issues, planning decisions influence energy use and production in profound, long-lasting ways. This article focuses mostly on the U.S. case.

1. SCOPE OF CITY PLANNING

Although urban design has deep historical roots, modern city planning began with the Industrial Revolution and the associated rapid urbanization that led to a degradation of the local quality of life. City planning was a response to the growing ugliness and unhealthfulness of cities, and it became a movement to reform an old paradigm of cities, which in many countries had been based on laissez-faire principles. In the Anglo-American world, which paced the rest of northern Europe, early signs of the city planning impetus in the late 19th and early 20th centuries were the English Garden City movement led by Ebenezer Howard, the U.S. City Beautiful movement that started in Chicago, and the zoning movement that attempted to separate residential and industrial land uses. Decentralization advocate Frank Lloyd Wright, and "tower in a park" advocate Le Corbusier, utopians of very different stripes, provided intellectual fodder to planning debates.

After World War II, the agenda of United States city planners changed periodically in accordance with evolving economic, political, and ideological forces. Scientific planning, emphasizing optimization and rationalistic analysis, dominated during the 1950s and has never disappeared. Advocacy planning, emphasizing empowerment of disenfranchised slum dwellers, gained modest influence in the 1960s and 1970s. Equity planning, with a focus on improved fairness in land use decisions, became a central tenet of planning from the late 1960s onward. In the 1970s, environmental planning became an important component of city planning. Starting in the late 1980s, city planning increasingly emphasized participatory and communicative planning processes.

 © 2004 Elsevier Inc. All rights reserved.

Collaboration, community visioning, smarter growth, and sustainable development are the watchwords of the planning profession at the beginning of the 21st century.

Planning practice begins with how to regulate land use in urban areas. Zoning and subdivision regulations are basic tools in determining land use patterns. Land use planning is inextricably related to transportation planning, which analyzes the travel behavior of individuals and aggregates on current and prospective networks. Urban designers apply these tools to give cities unity and coherence, with clear and legible structures and hierarchies of centers. In the United States, however, low-density development in suburban areas during the post-World War II period created "sprawl." Planners devised growth management strategies to control sprawl, and starting in the 1990s, a "smart growth" movement advocated an integrated form of regional growth management to restore the community vitality of center cities and older suburbs. City planners increasingly rely on public decision-making processes to organize the built environment, and recent efforts at communicative planning encourage citizens' direct participation in policy-making processes, implementation, and monitoring of policies, so as to minimize conflicts and negative externalities.

The city planning profession, however, has paid little attention to energy use, even though most cities experienced significant energy crises in the 1970s. Most planners assume that cities are, above all, places to consume energy rather than produce it. They are increasingly being shown to be wrong.

2. SITING ENERGY FACILITIES

2.1 Regulations and Institutions

In the United States context, most land use and facility siting decisions take place at the local level. The federal government has reserved siting authority for only a few types of energy facilities: pipelines and nuclear waste dumps are the chief examples. Most states delegate facility siting tasks to municipalities, although statewide bodies have been created in a few states. Statewide decisions about siting facilities are made by a siting board, an energy facility siting council, or an energy commission. The members of the siting institutions vary from state to state; for instance, the members of the Energy Commission in California consist of four professionals and one citizen, and the members of the Energy Facility Siting Council in Oregon consist of seven geographically representative citizens appointed by the governor. These institutions issue a certification for the construction and operation of energy facilities.

The siting certification processes also vary from state to state. Generally, in the beginning of the process, these siting institutions or subcommittees oversee the submission of plans by the utilities. Then a potential applicant submits a notice of intent, which describes the proposed facilities and which allows the office to collect public inputs through public hearings and to identify laws, regulations, and ordinances. At the next stage, the applicants submit the application to the office and then the siting institution decides whether to issue a site certification.

General criteria for siting facilities are based on standards determined by each state. The environmental standards include considerations of noise, wetlands, water pollution, and water rights; soil erosion and environmentally sensitive areas are also considered. Structural standards consider public health and facility safety. The siting institutions consider the applicant's organizational, managerial, and technical expertise; the applicants are required to have financial assurance. Siting institutions must also evaluate socioeconomic effects such as expected population increases, housing, traffic safety, and so on. However, deciding where to site facilities is not easy due to the divergent interests of multiple stakeholders, and the process of siting facilities therefore often brings about locational conflicts.

2.2 Locational Conflicts

Siting facilities can be understood as an exercise in siting locally undesirable land uses (LULUs), many of which cause locational conflicts and thereby place government officials and policy makers in opposition to grassroots "not in my back yard" (NIMBY) groups. Locational conflicts can escalate from local disputes into broad protests against structural inequities. Major issues driving locational conflicts include health and safety concerns as well as environmental impacts. In addition, facilities may be sited in neighborhoods whose residents are poor or people of color, becoming an issue of environmental justice. Opponents of constructing facilities sometimes adopt grassroots techniques, such as protests, civil disobedience, and initiatives. Many conflicts lead to judicial review, promoting high transaction costs.

New approaches to the resolution of locational conflicts have emerged since the 1970s to reduce

the high transaction costs. Alternative dispute resolution (ADR) is a method based on consensus building; this is also called principled negotiation or consensus-based negotiation. In their 1981 book, *Getting to Yes*, Roger Fisher and William Ury emphasize the importance of principled negotiation that applies scientific merit and objectivity to the problem in question. Planners engaged in siting facilities now routinely consider both efficiency and equity/fairness criteria.

2.3 Balancing Efficiency and Equity/ Fairness Considerations

By and large, the criteria to evaluate the benefits and costs of siting energy facilities will be determined by three factors: risk from the facilities, fairness in siting facilities, and productivity or efficiency of the facilities after the decision on siting the facility. It is desirable to reduce risks related to health and safety issues. There is a fairness issue behind the risk issue. The NIMBY syndrome usually results from an unequal distribution of costs and benefits. Sites of energy facilities are correlated to a variety of economic factors, such as natural resources, land prices, and urban infrastructures. In his 1992 piece in the *Virginia Environmental Law Journal*, Robert Collin argues that generally racial and ethnic minorities and low-income groups are more likely to be exposed to hazardous waste and pollution because they are more likely to live near those treatment facilities. Thus, the balancing of efficiency and fairness becomes a major argument in siting energy facilities. Participation by the affected parties and negotiation are standard components of most current United States energy facility siting processes. During the negotiation process, each party reviews the plan to site facilities and pays close attention to the type and amount of compensation awarded to the community and the criteria used for siting facilities. If the compensation is not sufficient, renegotiation may be necessary.

3. FINANCING ENERGY INFRASTRUCTURE

3.1 Public or Private Ownership

Energy infrastructures include many components: generation, transmission, and distribution of electricity; physical networks of oil and natural gas pipelines; oil refineries; and other transportation elements such as marine and rail transportation. Historically, industrialized countries have financed the energy sector privately. Wood, coal, oil, natural gas, and electricity were produced and transported mostly by private firms well into the first decades of the 20th century. Thereafter, as network energy utilities became ubiquitous, public involvement increased. The new Soviet Union made electrification a national project; financially unstable private utilities of the United States begged for and received government regulation following World War I, and New Deal regional and rural policies made access to electricity a public goal justifying public investment and ownership. Before the Great Depression of the early 1930s, several holding companies controlled more than 75% of all U.S. generation. After the Great Depression, holding companies proved to be financially unstable and caused the price of electricity to increase, so support grew for government ownership of utilities, especially hydroelectric power facilities. Since the energy crisis in the early and late 1970s, the structure of electricity industry has changed; non-utilities have been added in the electricity market and the vertical structure has been unbundled. Because of the uncertainty over the future structure of the industry and recovery of investment costs in electricity infrastructures, financing electricity infrastructure has been in difficulty. In the United States as of 1999, federally owned utilities controlled 10.6% of the total generation, cooperatives controlled 5.2%, publicly owned utilities controlled 13.5%, investor-owned utilities controlled 70.7%, and nonutilities controlled 18.9%. The combination of private and public ownership differs from region to region. There are about 5000 power plants in the United States.

As electricity was transformed from a novelty to a necessity during the mid-20th century, many countries worldwide nationalized their electricity sectors. Only since the 1980s have privatization and liberalization reversed this trend. In the developing world, lack of public funds has forced many countries to turn to private, often transnational, financiers for energy sector investments.

In the oil and natural gas sectors, historically, major energy companies have possessed a vertically integrated oil and natural gas infrastructure, involving oil and natural gas exploration, development, and production operations and petroleum refining and motor gasoline marketing. Also, independent oil and natural gas companies take part in segments of vertically integrated structures as producers, petroleum refiners, and providers of transmission pipelines. Oil and natural gas pipelines may run across state boundaries or federal lands in order to link supply locations to market areas in cities. In the

natural gas industry, there has been a structural change since the issue of Order 636 by the Federal Energy Regulatory Commission (FERC), which no longer permitted gas pipeline companies to engage in the sale of natural gas. Thus, several pipeline companies have been consolidated under single corporate umbrellas, such as El Paso Merchant Energy Company, the Williams Company, Duke Energy Corporation, and the now defunct Enron Corporation. Several of these companies have acquired interstate pipeline companies.

3.2 Vertical Disintegration

In the United States electricity sector, the industry structure throughout the 20th century consisted of three vertical components: generation, transmission, and distribution. With increasing demand, robust transmission networks, availability of new generating technologies, and an ideological shift, restructuring in electricity and natural gas industries has become significant. In the electricity sector, regulatory reform has changed the structure of the electricity industry; according to Order 888 issued by FERC, all utilities can have access to the U.S. transmission systems; in addition, as already mentioned, nonutilities can provide electricity through the transmission system. To reduce the costs resulting from constructing new power plants, the federal government is trying to remove the transmission constraints between regions, some of which are traceable to the 1935 Public Utility Holding Company Act (PUHCA) limiting interstate utility operations. Horizontal market power is replacing vertical integration as the dominant utility business strategy.

In the natural gas sector, business operations are going in a direction parallel to that of the electricity industry. In his 2001 report, *Natural Gas Transportation—Infrastructure Issues and Operational Trends*, James Tobin discusses that, to take a better position to handle the large growth in natural gas demand, natural gas companies have consolidated operations through major mergers.

3.3 Municipalization

In the early 20th century, local governments in the United States played key roles by awarding competing bidders electric service franchises. In fact, however, some local officials received bribes in return for granting franchises to utilities, and franchise holders imputed this burden to citizens. Citizens rebuked utility companies for causing the rates to be high, and called for local governments to take over some local electricity systems. Today, a significant minority of municipalities control electricity distribution to their residents, and a smaller number also generate and transmit power. In parallel, in the 20th century, strong state regulations were introduced to prevent corruption and guarantee the fair pricing among investor-owned utilities. From historical experience, municipalization has both strengths and weaknesses.

Nevertheless, the recent restructuring of the electricity sector might open new opportunities for public financial involvement. Municipal governments can serve as aggregators and use their market power to buy reasonably priced electricity on the open market. They can reduce rates for their residential customers or meet expectations of clean air and water by offering clean energy options. They can also use various renewable energy sources to meet their citizens' demand (this will be further discussed in Section 6).

4. INFLUENCE OF LAND USE PATTERNS ON ENERGY

4.1 Effects of Density, Grain, and Connectivity

Land use patterns can be parsimoniously characterized in four dimensions: degree of centralization or decentralization (urban form), ratio of population or jobs to area (density), diversity of functional land uses such as residential and industrial (grain), and extent of interrelation and availability of multiple modes of circulation for people and goods among local destinations (connectivity). The use of resources per capita diminishes as urban form becomes more centralized, density goes up, grain becomes finer, and connectivity shrinks. Metropolitan land use patterns in the United States after World War II show increased energy use due to increasing regional populations, decentralization, decreasing density, rougher grain, and increased connectivity.

Throughout the 19th century, most people in the United States lived in small towns and villages. With the advent of the 20th century, many people moved to industrial cities for jobs, a trend that peaked in the 1920s and was compounded by overseas immigration. Following the interruptions of the Great Depression and World War II, the pent-up demand for housing was met by a conscious process of suburbanization. Achieving the dream of home ownership became feasible for many Americans as new federal mortgage guarantee policies removed

TABLE I

Intensity of Land Use in Global Cities, 1990[a]

	Metropolitan density		Central city density[b]		Inner-area density		Outer-area density	
City	Pop.[c]	Jobs	Pop.	Jobs	Pop.	Jobs	Pop.	Jobs
American average[d]	14.2	8.1	50.0	429.9	35.6	27.2	11.8	6.2
Australian average[e]	12.2	5.3	14.0	363.6	21.7	26.2	11.6	3.6
Canadian average[f]	28.5	14.4	37.9	354.6	43.6	44.6	25.9	9.6
European average[g]	49.9	31.5	77.5	345.1	86.9	84.5	39.3	16.6
Asian average[h]	161.9	72.6	216.8	480.1	291.2	203.5	133.3	43.5

[a] From "Sustainability and Cities" by Peter Newman and Jeffrey Kenworthy. Copyright © 1999 by Peter Newman and Jeffrey Kenworthy. Adapted by permission of Island Press, Washington, D.C. Density expressed in persons per hectare and jobs per hectare.

[b] Central business district.

[c] Population.

[d] Average of Sacramento, Houston, San Diego, Phoenix, San Francisco, Portland, Denver, Los Angeles, Detroit, Boston, Washington, Chicago, and New York.

[e] Average of Canberra, Perth, Brisbane, Melbourne, Adelaide, and Sydney.

[f] Average of Winnipeg, Edmonton, Vancouver, Toronto, Montreal, and Ottawa.

[g] Average of Frankfurt, Brussels, Hamburg, Zurich, Stockholm, Vienna, Copenhagen, Paris, Munich, Amsterdam, and London.

[h] Average of Kuala Lumpur, Singapore, Tokyo, Bangkok, Seoul, Jakarta, Manila, Surabaya, and Hong Kong.

financial barriers to home ownership, as developers such as William Levitt perfected mass production of affordable housing units on greenfield sites, and as federal dollars poured into road building. Discrimination in lending and housing markets prevented most black Americans from participating in this exodus. Yet, by 1960, the suburban lifestyle was the conventional land use practice in the United States.

Suburban lifestyle has brought about sprawling, low-density suburban communities; according to the 1990 census, from 1970 to 1990, the density of urban population in the United States decreased by 23%. From 1970 to 1990, more than 30,000 square miles (19 million acres) of once-rural lands in the United States became urban areas, an area equal to one-third of Oregon's total land area. From 1969 to 1989, the population of the United States increased by 22.5%, and the number of miles traveled by that population ("vehicle miles traveled") increased by 98.4%.

Anthony Downs defines the term "sprawl" as (1) unlimited outward extension, (2) low-density residential and commercial settlements, (3) leapfrog development, (4) fragmentation of powers over land use among many small localities, (5) dominance of transportation by private automotive vehicles, (6) no centralized planning or control of land use, (7) widespread strip commercial development, (8) great fiscal disparities among localities, (9) segregation of types of land use in different zones, and (10) reliance mainly on the trickle-down, or filtering, process to promote housing to low-income households. The impacts of urban sprawl have caused increasing traffic congestion and commute times, air pollution, inefficient energy consumption and greater reliance on foreign oil, inability to provide adequate urban infrastructures, loss of open space and habitat, inequitable distribution of economic resources, and the loss of a sense of community.

When land use patterns in the United States are compared to other countries, they show significant differences. In comparisons of metropolitan density, U.S. cities are of low density in residential and business areas whereas European cities are three to four times denser. Asian cities are 12 times denser compared to American cities (Table I).

The average energy use for urban transportation in American cities is 64.3 gigajoules (GJ) of fuel per capita compared to 39.5 GJ in Australia, 39.2 GJ in Canada, 25.7 GJ in Europe, and 12.9 GJ in Asia. Energy use in American cities is five times more than in Asian cities. In addition, European cities consume two times more energy compared to Asian cities. This shows that energy use in transportation is closely related to land use patterns and income levels (Table II).

4.2 Environmental and Public Health Implications

Low-density development and urban sprawl are correlated with high-level air pollutant emissions from transportation. Emission rates of the greenhouse gas carbon dioxide in U.S. cities are higher than those of

TABLE II

Transportation Energy Use per Capita in Global Regions, 1990[a]

City	Private transportation			Public transportation			Total transportation energy (MJ)	Total Transportation energy/$ of GRP[b] (MJ/$)
	Gasoline (MJ)	Diesel (MJ)	Private % (of total)	Diesel (MJ)	Electricity (MJ)	Public % (of total)		
American average[c]	55,807	7764	99%	650	129	1%	64,351	2.38
Australian average[d]	33,562	4970	98%	764	159	2%	39,456	1.96
Canadian average[e]	30,893	6538	97%	1057	163	3%	39,173	?
European average[f]	17,218	7216	95%	604	653	5%	25,692	0.83
Asian average[g]	6311	5202	89%	1202	148	11%	12,862	3.81

[a] From "Sustainability and Cities" by Peter Newman and Jeffrey Kenworthy. Copyright © 1999 by Peter Newman and Jeffrey Kenworthy. Adapted by permission of Island Press, Washington, D.C. Use expressed in megajoules.

[b] GRP, gross regional product is the measure of all goods and services produced in the regional urban area.

[c] Average of Sacramento, Houston, San Diego, Phoenix, San Francisco, Portland, Denver, Los Angeles, Detroit, Boston, Washington, Chicago, and New York.

[d] Average of Canberra, Perth, Brisbane, Melbourne, Adelaide, and Sydney.

[e] Average of Winnipeg, Edmonton, Vancouver, Toronto, Montreal, and Ottawa.

[f] Average of Frankfurt, Brussels, Hamburg, Zurich, Stockholm, Vienna, Copenhagen, Paris, Munich, Amsterdam, and London.

[g] Average of Kuala Lumpur, Singapore, Tokyo, Bangkok, Seoul, Jakarta, Manila, Surabaya, and Hong Kong.

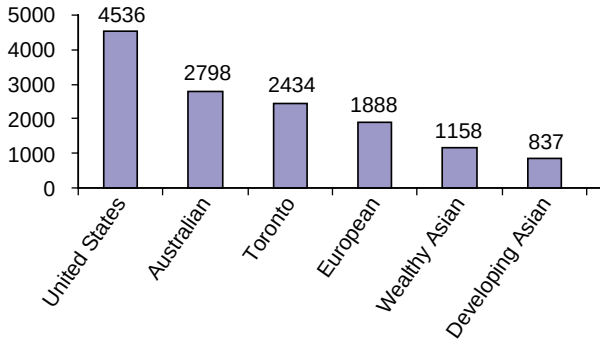


FIGURE 1 According to 1990 statistics, rates of annual CO_2 emissions (shaded bars, in kilograms/person) in the United States are 2.4 and 5.4 times higher than those of European cities and developing Asian cities, respectively. From "Sustainability and Cities" by Peter Newman and Jeffrey Kenworthy. Copyright © 1999 by Peter Newman and Jeffrey Kenworthy. Adapted by permission of Island Press, Washington, D.C.

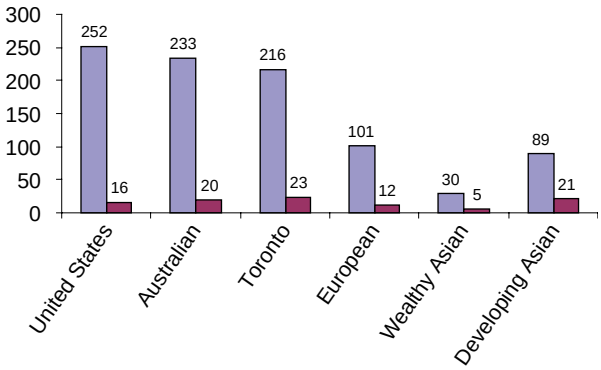


FIGURE 2 Rates of annual NO_x, SO_2, CO, and volatile hydrocarbon smog-related emissions (light shaded bars, in kilograms/person) in the United States are 2.5 and 2.8 times higher than those of European and Asian cities, respectively, in 1990. Volatile particulates are also shown (dark shaded bars, in grams/passenger-kilometers). From "Sustainability and Cities" by Peter Newman and Jeffrey Kenworthy. Copyright © 1999 by Peter Newman and Jeffrey Kenworthy. Adapted by permission of Island Press, Washington, D.C.

other cities (Fig. 1). Health problems are caused by smog-related emissions, involving nitrogen oxides (NO_x), sulfur dioxide (SO_2), carbon monoxide (CO), volatile hydrocarbons (VHCs), and volatile particulates (VPs). Per capita emission rates in the United States and Australia are higher than those in European and Asian cities (Fig. 2).

Los Angeles (LA) is famous for its smog caused by automobile emissions. The analysis in the 1996 report of the LA Air Quality Management District is that average ozone levels in LA, the South Coast Region, are twice the federal health standard. Ozone concentrations have exceeded standards on as many as 98 days per year. Because of the frequency of occurrence, the smog in this region is especially harmful to aged people and children. The symptoms caused by smog are usually aching lungs, wheezing, coughing, and headache.

Ozone also hurts the respiratory system's ability to fight infection. Other cities have similar air pollution problems; LA is no longer exceptional, especially among the growing cities of the American sunbelt.

4.3 Can Technological Innovations in Transportation Offset Impacts of Sprawling Land Use Patterns?

Optimists hope that innovations in transportation technology will solve the problems associated with sprawling land use patterns. Traffic congestion, air pollution, and burgeoning energy consumption are each the target of specific research initiatives. The U.S. government supports research to develop vehicles, such as hybrid electric and fuel cell cars, that will use energy sources (i.e., compressed natural gas, biodiesel fuel, ethanol, and hydrodiesel electricity) less harmful, compared to fossil fuels, to the environment. The government also promotes fuel efficiency through the Corporate Average Fuel Economy (CAFE) standards. In addition, with federal funds, some cities are introducing Intelligent Transportation Systems (ITSs), which increase the effectiveness of existing roadways by giving drivers real-time information about the best travel routes. However, many analysts argue that transportation energy efficiency and ITSs will not reduce the number of vehicles in use and air pollution emissions. Thus, technological innovations should be accompanied by nontechnological solutions. The Department of Energy suggests new land use planning strategies that will improve energy efficiency and will protect natural corridors and open space. These include transit-oriented design, mixed-use strategies, urban growth boundaries, infill development, greenways, brownfields redevelopment, transfer of development rights, open-space protection, urban forestry, land trusts, agricultural land protection, and solar access protection.

5. INFLUENCE OF URBAN/ ARCHITECTURAL DESIGN ON ENERGY POLICIES AND PRACTICES

Urban form at the metropolitan scale strongly affects transportation-related energy consumption, but smaller scale urban and architectural design decisions also influence energy use. Site layouts and building material choices affect microclimate and can create heat island effects.

Heat island effects arise from multiple sources. First, urbanized areas have residential, commercial, and industrial zones that displace trees and shrubs to varying extents. Trees and shrubs have the capacity to control their own temperature by releasing moisture, resulting in a natural cooling effect known as evapotranspiration. The displacement of trees and shrubs in the most built-up areas removes beneficial natural cooling. In addition, impervious surfaces, dumping excess heat from air-conditioning systems, and air pollution cause the ambient temperatures to rise, increasing energy use due to the large demand for air conditioning.

Site layouts and physical features of buildings and other impervious materials in a city have an influence on microclimate changes such as sunlight access, wind speed, temperature, and noise. Sunlight striking dark, impervious surfaces becomes sensible heat, causing the temperatures to rise. Building locations influence wind speed, sometimes impeding and other times redirecting the flow of wind on a site. Solar access, prevailing winds, tree location, and topographic modifications change microclimate. Wind facilitates good air circulation, thereby reducing build-up of heat. Effectively narrowing a street with street trees can reduce summer temperatures by 10°F. A modified site layout (street orientation, building placement, location on slope, and landscaping) can reduce the energy consumption of an ordinary residence by 20%.

Architects and, to a lesser extent, planners can influence building design and materials choices at the micro level. Energy efficiency in building design can be improved by maximizing solar access, by minimizing infiltration but taking full advantage of natural ventilation, by creating non-window spaces as buffers on north walls, and by utilizing natural convection and passive solar designs, for example. High-performance lighting and maximization of natural light by solar access can also improve energy efficiency, reducing energy costs. Institutionally, in the United States, the Energy Policy and Conservation Act (EPC Act) of 1975 and the Energy Policy Act (EP Act) of 1992 have improved the energy efficiency of household and commercial building appliances. Advanced building design techniques and associated technologies and appliances available today can cut in half the energy consumption of buildings.

6. CENTRALIZED VS. DISTRIBUTED ENERGY SUPPLIES

6.1 Implications for Planners

Energy supply, particularly electricity supply, needs to be analyzed in its regional context. It has long

been assumed that increasing returns to scale result from increasing efficiency and technological specialization by means of large-scale supply. Thus, the single-minded pursuit of centralized energy supply led to a North American electric power system that was interconnected on a multistate basis and included gigawatt-scale generating plants. Centralized energy supply also brought with it certain inefficiencies associated with monopoly power, because utilities had few incentives to be operationally efficient and innovative. The recent restructuring efforts have been inspired by expectations of increased dynamic efficiency.

Another weakness of a centralized energy supply (characterized as "brittle" by Lovins) is the vulnerability of large power plants and transmission lines to disruptions, both natural and man-made, accidental and intentional. From the transmission system-related 1965 Northeast blackout in the United States, to the generation system-related 1986 Chernobyl accident in the former Soviet Union, to the utility distribution system disruptions caused by the 1993 and 2001 World Trade Center attacks in New York, to the 2003 cascading outages in the Northeast, there is accumulating evidence that the redundancies built into the large-scale grid provide inadequate reliability and security.

The introduction of smaller scale distributed energy (DE) power plants is one potential solution to these problems. Community-based energy supply falls within the planners' geographic sphere of influence. If this paradigm catches hold, energy planning could become a basic task for planners.

6.2 History of Centralization

In the United States, the electric power supply grew from the neighborhood scale in the 1880s to the municipal scale by the turn of the century. During the early 1900s, integrated large holding companies emerged, and interconnection of municipal systems proceeded apace, driven by a desire to increase load diversity and reliability of supply. Interstate transmission networks were entering use by the late 1920s. By that time, as mentioned earlier, 16 holding companies supplied most of the U.S. generation. Because of their nontransparent, unstable financial structure, the PUHCA was passed in 1935. This act mandated multistate holding companies to be subject to the state authority, and forbade a parent holding company from taking out loans from an operating utility. The state-based reorganization through PUHCA did not cause any serious problems until the 1960s. Following the 1965 Northeast blackout, the North American Electric Reliability Council (NERC) was formed to coordinate utility responses to regional disruptions and to avoid cascading failures.

Two oil crises in the 1970s and the restructuring of other regulated industries (such as telecommunication) brought about a rethinking of the electricity industry. During this period, technological innovations in gas turbines reduced generation costs and optimal generating-plant sizes, and thus made the large-scale power plants designed to exploit economies of scale no longer necessary. In addition, PUHCA exacerbated problems by placing the siting of transmission lines under state rather than federal authority, thereby constraining power flows from states with power surpluses to other states experiencing shortfalls. As a result, the vertically integrated structure is being unbundled, competition has been introduced in the electricity market, and a greater variety of new power generation techniques and resources are entering use. In particular, the DE technologies are gaining a toehold.

Most DE stations are located close to their ultimate customers. Supply-side distributed energy resources (DERs) include wind turbines, natural gas reciprocating engines, microturbines, photovoltaics, and fuel cells (Fig. 3). Demand-side DERs, for example, include schemes to reduce peak electricity demand and designs for high-efficiency buildings

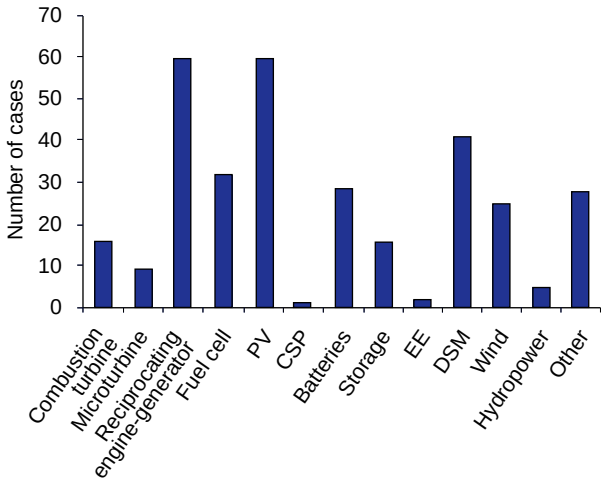


FIGURE 3 In a study of 275 distributed energy resources installations, reciprocating engines, solar electric power (PV, photovoltaic), and demand-side management (DSM) systems are the three most commonly used distributed energy resources technologies. In some cases, more than one technology was used per installation. CSP, Concentrating solar power; EE, energy efficiency. Reprinted from *DER Technologies*, the Office of Energy Efficiency and Renewable Energy, DOE.

and advanced motors and drives for industrial applications.

6.3 Special Case of District Energy Systems

Although centralized energy systems locate generators remote from consumers, some distributed energy systems locate generators on the consumer's premises, occupying valuable floor space and imposing significant installation and maintenance costs. District heating and/or cooling systems lie in-between these extremes. These systems produce steam, hot/chilled water, and electricity at a central location and distribute them to nearby buildings (Fig. 4).

A district heating or cooling system can reduce capital costs and save valuable building space, and can use not only conventional resources such as coal, oil, and natural gas, but also renewable fuels such as biomass and geothermal, thereby increasing fuel diversity. In particular, combined heat and power (CHP; also known as cogeneration), which reuses waste heat after producing electricity, can increase the overall process energy efficiency to more than 70%. Steam or hot water can be transformed to chilled water for refrigeration by means of absorption chiller technology.

City planning decisions strongly influence the economic feasibility of district heating and cooling systems. Such systems are most viable in compact downtown areas with high load density and a diversity of uses that support 24-hour operations. They are not viable in dispersed suburban settings.

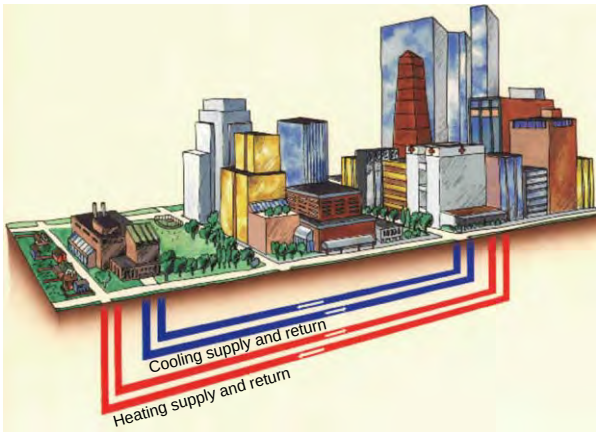


FIGURE 4 District energy systems distribute steam, hot/chilled water, and electricity from a central power plant to nearby buildings. With the technology of combined heat and power, they use 40% of input energy for electricity production and 40% for heating and cooling. Reprinted from *What Is District Energy?*, with permission of the International District Energy Association (2001).

6.4 Codes and Covenants Affecting Distributed Fossil Energy Systems and Distributed Renewable Energy Systems

The United States Public Utility Regulatory Policies Act of 1978 encouraged cogeneration and renewable (solar photovoltaic, wind) energy production. However, siting the facilities and interconnecting them with the grid involve significant problems, still unsolved today. There is no uniform standard code for interconnection between grids, and utilities have little incentive to modernize technical requirements for grid interconnection. Federal and state regulatory agencies share responsibility for setting the rules governing the development of distributed generation. The California Energy Commission argues that the main issues include paying off the utilities' stranded investments, siting and permitting hurdles, interconnection to the grid, environmental impacts, and transmission system scheduling and balancing.

At the state level, supportive policy-makers could revise the rules on siting facilities, financial incentive programs, interconnection, net-metering programs, and air quality standards to encourage the development of the DE market. At the local level, existing building codes, zoning ordinances, and subdivision covenants often forbid distributed energy system installations. For example, many communities have regulations that restrict homeowners' opportunities to install solar energy systems on their roof. Thus, a partnership among community groups, local governments, and developers is needed to address these restrictions before DERs can move significantly forward.

6.5 Prospects for Local Self-Sufficiency

The United States Department of Energy (DOE)sets forward one long-term vision of DER: "the United States will have the cleanest and most efficient and reliable energy system in the world by maximizing the use of affordable distributed energy resources." In the short term, DOE sets forward the plan to develop the technologies for the next-generation DE systems and to remove regulatory barriers. In the medium term, the goals are concentrated on reducing costs and emissions and on improving energy efficiency. In the long term, the target is to provide clean, efficient, reliable, and affordable energy generation and delivery systems. California is one of the few states to set a target: it has a plan to increase DE generation by 20% of state electricity generation. Yet municipal energy self-sufficiency is

an unlikely prospect anywhere in the world during the next 20 years, if it is even desirable. More likely is an infiltration of DE technologies into a national grid that also contains many large central power plants. DE supply will depend on private investment in the DE industry. To encourage private investment, government will continue to have a role in setting market rules and helping investors overcome barriers associated with siting facilities, grid interconnection standards, and financing.

7. INTEGRATED VS. SEGMENTED ENERGY PLANNING

7.1 History of Debate

In his 2001 book, *Urban Development*, Lewis Hopkins argues that planning, when done well, is very much a big-picture exercise designed to improve coordination among distinct, yet interdependent, decision arenas. Good urban plans coordinate various private land use and public infrastructure system decisions, for example. In the energy arena, planning can have a similar integrating function.

7.1.1 Economies of Scope

A primary driver of integrated energy planning is the potential for economies of scope. This economic argument states that in certain circumstances, the average total cost decreases as the number of different goods produced increases. For instance, a company can produce both refrigerators and air conditioners at a lower average cost than what it would cost two separate firms to produce the same goods, because it can share technologies, facilities, and management skills in production, thereby reducing costs. In the energy industry, for over a century, many firms have realized economies of scope by integrating natural gas and electricity distribution and sales. Governmental energy planners have pursued similar economies since at least the 1970s by planning for the energy sector as a whole rather than keeping electricity, gas, oil, and other sectoral plans apart.

7.1.2 Integrated Resource Planning Movement of the 1980s/Early 1990s

The energy supply and financing crises of the 1970s inspired some U.S. utility regulators to demand a new kind of planning from regulated utilities. Rather than use only supply-side options and treat demand as exogenous, utility planners were directed to consider both supply- and demand-side options using a methodology known as least-cost planning, and later called integrated resource planning (IRP). In the United States, more than 30 state Public Utility Commissions (PUCs) adopted IRP procedures at the end of 1980s. The EP Act, passed in 1992, encouraged all electric utilities to exploit integrated resource planning.

Under IRP, utilities attempted to shape future energy supply and demand. In the process of planning, they considered factors such as energy efficiency and load-management programs, environmental and social aspects, costs and benefits, public participation, and uncertainties. In addition, demand-side resources were given the same weight as supply-side resources. Demand-side options included consumer energy efficiency, utility energy conservation, renewables installed on the customer side of the meter, and pricing signals. Supply-side options consisted of conventional power plants, non-utility-owned generation, power purchases from other suppliers, and remote renewables (Table III).

However, many states began to think about deregulating or restructuring their gas and electric power industries shortly after the EP Act was passed in 1992. This meant that state governments could no longer regulate planning practices and prices so closely, because utility companies were situated under competition in the energy supply market. Under restructuring, IRP became a strategic tool for the utility rather than a public planning process. Additionally, Kevin Kelly, in his 1995 chapter in the book *Regulating Regional Power Systems*, points out that there was a discrepancy between regional least-cost planning at the federal level and state least-cost planning at the state level. The need for public energy planning remained in the United States, but by the mid-1990s it no longer carried the force of regulation.

7.1.3 Sustainable Communities

Since the 1992 Earth Summit in Rio de Janeiro, and accelerating through the turn of the millennium, the term "sustainable development" has been a visible agenda item in every social sphere. Sustainable development does not simply regard development as growth; as Mark Roseland puts it in his 1998 book, *Toward Sustainable Communities*, it seeks to improve the well being of current and future generations while minimizing the environmental impact. There are different expressions of this idea: sustainable cities, sustainable communities, eco-cities, green communities, ecovillages, green villages,

TABLE III

Demand-Side and Supply-Side Options

Options	Example
Demand Side	
Consumer energy efficiency	Home weatherization, energy-effcient appliances for lighting, heating, air conditioning, water heating, duct repair, motor, refrigeration, energy-efficient construction programs, appliance timers and control, thermal storage, and geothermal heat pumps
Utility energy conservation	Load management, high-efficiency motors, and reduced transmission and distribution losses
Rates	Time-of-use, interruptible, and revenue decoupling
Renewables	Solar heating and cooling, photovoltaics, passive solar design, and daylighting
Supply side	
Conventional power plants	Fossil fuel, nuclear, life extensions of existing plants, hydro/pumped storage, repowering, and utility battery storage
Non-utility-owned generation	Cogeneration, independent power producers, and distributed generation
Purchase	Requirement transactions, coordination transmissions, and competitive bidding
Renewables	Biomass, geothermal, solar thermal, photovoltaics, and wind

Source. U.S. Department of Energy (2001).

econeighborhoods, and so on. One of the fundamental concepts is to minimize our consumption of essential natural capital, including energy resources.

In this sense, sustainable communities might perform IRP and could even give more weight to demand-side management than to supply options. However, in doing so, some argue that they push public policies counter to market forces, which currently make supply investments more profitable. Yet IRP could be performed from the consuming community's point of view, as a guide to aggregate expenditures on energy services. Some municipal electric utilities (in Sacramento, California, for example) have adopted this approach and they have maintained stable energy bills while shifting toward renewable energy supplies and more efficient energy usage.

Restructuring of energy industries is replacing regulated, firm-level IRP processes. Energy supply service is becoming more diversified under deregulation. Now communities are well placed to take on the challenge of integrative planning, and a very few are already becoming broad community businesses, extending the concept of the community energy utility to include recycling and reusing waste products within the community (Table IV).

7.2 Industrial Complexes and Ecoindustrial Parks

Integrated planning can clearly extend beyond energy into other realms. The Kalundborg industrial complex in Denmark is frequently cited as an

TABLE IV

Comparison of Conventional and Community Energy Utilities[a]

Parameter	Conventional energy utility	Community energy utility
Generating plant	Large scale and remote from customers	Small scale and locally based near customers
Customer base	Very large, with tens of thousands of customers	Relatively small, usually a few thousand customers and could be much smaller
Legal structure	Public limited company, subsidiary to multinational parent company	Variety of legal structures, including joint venture companies, cooperatives, and charities
Control	Main control lies with the multinational parent, but shareholders also influence decisions through their desire to maximize dividends	Day-to-day control could be in the hands of a commercial management; energy end users have a substantial stake in the utility; local authority may also be a stakeholder as an agent for the community
Technology	Conventional steam electricity, nuclear power, conventional hydroelectricity, etc.	Combustion turbine, microturbine, reciprocating engine-generator, fuel cell, photovoltaic, wind, etc.

[a] Adapted from Houghton (2000).

exemplar of industrial symbiosis that uses energy cascades and closed-loop materials cycling among several firms to yield economic benefits with less pollution, more efficient use of resources, and less need for environmental regulatory supervision, compared to traditional arrangements.

Firms in such ecoindustrial parks can share their environmental management infrastructure and increase ecoefficiency by rationalizing and optimizing aggregate materials and energy flows. Power plants, especially CHP systems, make natural anchor tenants. City planners specializing in industrial parks are just beginning to focus on such possibilities, although the examples and ideas have been around for decades.

7.3 Optimizing across Multiple Energy Sources

Energy planning must consider several factors: energy resources are unevenly distributed across Earth, technological innovations can disrupt existing equilibria, political and economic boundaries and regulations affect the use of energy resources, culturally and socioeconomically mediated human behavior can influence energy consumption patterns, and different places have different climates. Optimizing the use of energy resources will require different strategies in different places to reflect specific local conditions. For instance, a higher population density will make CHP more practical, but it may militate against other devices, such as passive solar design, due to the increased overshadowing.

7.4 What the Future Holds

Integrated energy planning remains an attractive proposition for firms and for governments, although deregulation of electricity and gas markets has disrupted previous regulatory drivers of this practice. Its scope may intersect in a significant way with that of city planning if current interest in distributed generation, industrial ecology, and sustainable communities persists.

8. ENERGY FOR THE WORLD'S BURGEONING MEGACITIES

8.1 Challenges of Scale and City–Hinterland Linkages

One of the trends at the beginning of the 21st century is the emergence of big cities. Currently, there are 30–35 cities in the world whose population size is more than 5 million. With the growing concern for the environmental impact of cities, people are beginning to pay attention to city systems from the perspective of sustainability. In his 2000 book, *Green Urbanism*, Timothy Beatley considers the ecological footprints of cities, which consume energy, materials, water, land, and food to support them, and then emit wastes. Urban growth and development have triggered the expansion of urban infrastructures, increasing energy consumption. For instance, in their 1997 *Nature* article, "The Value of the World's Ecosystem Service and Natural Capital," Robert Costanza and colleagues estimate that the energy consumed in urban areas is produced in coastal areas, forests, grasslands, and wetlands, and its total value is at least $721 billion per year. There is no primary energy production in urban areas. Rather, urban systems consume energy produced from the environment. The United Kingdom's International Institution of Environment and Development estimates that the entire productive land of the UK is necessary to maintain the population of London, whose population size is 7 million. The goal of managing urban areas is shifting from ensuring the unconstrained consumption of energy resources to encouraging their efficient use.

8.2 Cities as Systems

Cities are extremely complex combinations of physical subsystems, including those of urban infrastructure, various housing and other structures, transportation, communication, water transport, geology, ecosystem, solid waste, food and water distribution, economic zones, and demographics. In addition, social, political, economic, and cultural networks are interconnected, and create these built environments. To maintain and manage complex city systems, many agencies are involved. In the United States, these include the Department of Energy, Department of Defense, Department of Transportation, Environmental Protection Agency, Department of Housing and Urban Development, Federal Emergency Management Agency, United Nations, state and local planning agencies, and others. Yet none of these government agencies controls or plans cities to any significant extent: cities are self-organizing systems. Their top-down steering capacity is minimal. With continued urbanization worldwide, and with the development of high technology and the growing concerns for the limit of natural resources, there are severe challenges for cities. Adequate

energy supplies rank high among these challenges. In the developing world, financial and environmental constraints will dictate more diversity in energy supplies, and more common use of demand management tools, than has been the case in the industrialized world to date.

8.3 Spaces of Flows

Cities are connected to one another in global transportation and communication networks that allow rapid flows of information, capital, people, and goods from one place to another. Under global capitalism, a hierarchy of cities exists in which a few cities (e.g., New York, London, and Tokyo) serve as corporate command centers while others specialize in manufacturing, back-office operations, market niches, and tourism, or serve primarily local needs. Cities likewise withdraw resources from and deliver products and residues to their hinterlands. In his 2001 book, *The Rise of the Network Society*, Manuel Castells refers to a shift from spaces of places to spaces of flows, arguing that cities are nodes through which flows occur. This conception implies that cities are influenced by a larger systemic logic, and that the nature of the flows and transformations the city imposes on those flows is relevant to city planners. The flows slow to a trickle in the absence of adequate energy supplies, and environmentally insensitive transformations of energy flows in urban areas are choking places such as Bangkok and Mexico City. Efforts by planners and local officials worldwide to ensure reliable, secure urban energy infrastructures have increased since the terrorist acts of September 11, 2001.

9. CONCLUSION

City planning evolved as a reaction to the ugliness and unhealthfulness of urban areas following the industrialization and urbanization in the late 19th century. In the 20th century, city planning broadened its scope and developed its tools, including land use planning, zoning, transportation planning, and growth management. Energy planning has not been the major agenda in city planning.

Since the oil crises in the 1970s and with increasing public awareness of the environmental impacts of fossil fuel consumption, there has been a fundamental change in energy policy. The security of the energy supply, the price of energy, and its environmental impacts have formed a volatile tripartite political agenda, and technically innovative, alternative energy resources have begun emerging.

Policy levers exist at several levels. For instance, although transit-oriented development is located in macro-level (or metropolitan) planning, walkable street development belongs to meso-level (or community) planning, and parking requirements play out at the micro (site) level of planning. At the meso (community) level are the placement of technological innovations such as distributed energy resources. At the micro (site) level, design decisions can alter the energy efficiency of building layouts and materials, lighting, and appliances. However, some innovations are restricted by current regulations and energy policies at each government level. There are especially severe energy-related challenges for city planning in the world's growing megacities.

Planning spans the gap between innovative technologies of energy production and their implementation. Hopkins argues that planning is a complicated and integrated process interconnecting regulation, collective choice, organizational design, market correction, citizen participation, and public sector action. Planners struggle with the vagueness of what their goals are, how their processes perform, and who can be involved in the planning process.

In sum, energy planning could be one of the major agendas in city planning, and it could conceivably be incorporated into city planning, although it has not been to date. Energy planning and city planning intersect at both the community and metropolitan levels, and thus planners will be involved in all attempts to provide more reliable, affordable, and environmentally sound energy for the world's cities.

SEE ALSO THE FOLLOWING ARTICLES

Ecological Footprints and Energy • Economic Growth and Energy • Heat Islands and Energy • Land Requirements of Energy Systems • Population Growth and Energy • Suburbanization and Energy • Urbanization and Energy

Further Reading

Andrews, C. J. (2002). Industrial ecology and spatial planning. *In* "A Handbook of Industrial Ecology" (R. U. Ayres and L. W. Ayres, Eds.), pp. 476–487. Edward Elgor, Cheltenham, UK.

Barton, H. (ed.). (2000). "Sustainable Communities the Potential for Eco-neighbourhoods." Earthscan Publ. Ltd., London.

Beatley, T. (2000). "Green Urbanism: Learning From European Cities." Island Press, Washington, D.C.

Calthorpe, P., and Fulton, W. (2001). "The Regional City." Island Press, Washington, D.C.
Campbell, S., and Fainstein, S. (eds.). (2000). "Readings in Planning Theory." Blackwell Publ., Cambridge, Massachusetts.
Castells, M. (2001). "The Rise of the Network Society." Blackwell Publ., Cambridge, Massachusetts.
Hopkins, L. D. (2001). "Urban Development: the logic of making plans." Island Press, Washington, D.C.
Newman, P., and Kenworthy, J. (1999). "Sustainability and Cities." Island Press, Washington, D.C.

Clean Air Markets

ALEXANDER E. FARRELL
University of California, Berkeley
Berkeley, California, United States

1. Introduction
2. Concepts
3. Examples
4. Looking Ahead

Glossary

banking The ability to save excess allowances in a Cap-and-Trade system from one period in order to use or trade them in a subsequent period.
cap-and-trade An emission trading program in which the government creates a fixed number of allowances, auctions or allocates these to regulated sources, and requires regulated sources to surrender allowances to cover actual emissions.
command-and-control regulation A form of regulation that is highly specific and inflexible, such as an emission standard (e.g., tons/year) or a performance standard (e.g., kg/unit of output).
discrete emission reduction A credit for a one-time reduction in pollution in the past below a regulatory standard. Often measured in mass (e.g., tons).
emission allowance The currency used in Cap-and-Trade programs. Allows the emission of one unit (e.g., ton) of pollutant.
emission reduction credit A credit for permanently reducing pollution below a regulatory standard or baseline. Often measured in mass emission rate (e.g., tons/year)
emission trading A generic term for environmental regulation that uses an allowances (or credit) for emissions (or emission reductions) that can be bought and sold.
market-based regulation Any regulatory method that use taxes, subsidies, tradable instruments (e.g., emissions allowances and emission reduction credits), or similar methods in stead of, or as a complement to command-and-control regulations.
verified emission reduction A credit for greenhouse gas emission reduction that has been verified by a third party but is not necessarily acknowledged by a government.

Government regulation can apply economic forces to achieve policy goals, as well as include mandatory standards and prohibitions. Such approaches are increasingly being applied to energy systems, especially for the control of air pollution and greenhouse gases, and the promotion of renewable energy. Most common are emission trading programs, which create clean air markets. This article discusses key concepts associated with clean air markets and the major implementation challenges associated with them. Several examples are used to illustrate these ideas.

1. INTRODUCTION

The use of market-based instruments (MBIs) for environmental policy is growing rapidly in regulating energy production and will likely be central in the still-evolving climate change policy regime. (See Farber in this volume for an overview.) Several features make MBIs attractive relative to more traditional command and control (CAC) regulation, but MBIs need to be designed and implemented carefully. This entry discusses some basic concepts and challenges of MBI and examines important examples of MBIs that feature air pollution allowances or credits, since these are of most importance in energy.

The most obvious benefit of MBIs is that they can help reduce the cost of environmental protection, which they do by providing significant flexibility to regulated organizations (typically firms). In the United States, programs that use tradable MBIs such as credits and allowances (e.g., emission trading) have enabled environmental protection while saving billions of dollars in pollution control costs. One important reason for the increasing interest in MBIs is that as environmental controls have become tighter and tighter, they have also become more expensive, making cost-effectiveness ever more important. The extreme case is climate change, which is likely to require the deployment of energy

 © 2004 Elsevier Inc. All rights reserved.

supply systems that are largely free of fossil fuels over the next several decades. Attempting such a massive change too fast or with CAC regulation alone would raise the cost and make it very difficult to achieve in democracies.

MBIs give firms incentives for environmental performance, rather than specifying emission rates or technical standards, as CAC regulations do. The basic logic behind MBIs is that firms differ in how much it costs them to control emissions, so if a high-cost company can pay a low-cost company to control emissions on its behalf, the net result is a savings. This approach assumes emissions from all sources are equivalent (i.e., uniform mixing); when this is not the case, additional rules can be added. Importantly, the use of MBIs can make environmental protection look more like an ordinary business issue to managers and can allow them to apply risk management tools (often in the form of financial derivatives). In addition, MBIs can be (but are not always) easier to implement since government often does not have use intrusive inspections or detailed permitting processes that require considerable industry knowledge, which may be hard for government to obtain. However, careful (and often expensive) monitoring may be needed, both at the plant level and in the marketplace, to ensure MBIs perform as expected. Finally, MBIs typically provide significant incentives for innovation, both managerially and technologically. However, the size and shape of these benefits vary greatly with the details of MBI design and implementation.

2. CONCEPTS

There are several types of MBIs, including taxes, subsidies, and different varieties of tradable instruments. Others include liability rules and information.

2.1 Taxes and Subsidies

The simplest MBI is a tax on pollution, which will give firms incentives to reduce emissions and to innovate to find new, cheaper ways to do so. Subsidies can work similarly, but, of course, they give positive incentives rather than negative. Economists point out that by making polluters pay for the damage they produce, and thus internalizing what they call externalities, emission taxes can increase economic efficiency. However, emission taxes place a cost on all firms, no matter how clean, and so are unpopular with firms. It is also difficult to achieve a specific environmental outcome with taxes, since the government will never know ahead of time what tax would be needed to achieve a desired level of emissions and requires governments to have detailed knowledge of regulated industries as well as the ability to forecast well. Two exceptions are carbon (or energy) taxes, which sometimes replace a previous energy tax, and auxiliary taxes in emission trading programs.

2.2 Emission Reduction Credits (ERC)

In places with poor air quality and a CAC regulatory framework, new, large sources (like a new power plant) or sources undergoing expansions are required to offset their new emissions. The instruments used for this purpose are emission reduction credits (ERCs), which are project based, rely on permanent reductions in emissions from specific sources compared to a baseline. For instance, ERCs can be created when a polluting facility goes beyond regulatory requirements in pollution control (over-control), permanently slows output (and thus emissions), or shuts down. Emission credits can be mass based (e.g., tons) or rate based (e.g., pounds per day), depending on the specifics of the underlying regulatory program to which the ERC has been added. However, the company that is trying to create emission credits must get government approval, which can be a slow process and raise transaction costs.

Typically, ERC systems are add-ons to preexisting CAC regulations, designed to provide flexibility and reduce costs. CAC regulations typically specify specific emission rates or control technologies and require regulated sources to submit detailed compliance plans explaining how they will meet these requirements. Regulators approve these plans and inspectors periodically verify compliance. Compliance planning is a significant burden for some firms since it is costly and limits their flexibility should market or technical conditions change. It also tends to reduce innovation since it is risky and time consuming to get approval for new technologies.

In the United States, ERC systems have had limited success, but they are not used much elsewhere. They have played a role in programs for federal air pollution control and are used by a few states. One problem is convincing environmental regulators of the size of emission reductions, because it is often impossible to measure emissions directly and because of disputes over appropriate baselines. Firms must prove that the ERCs are “surplus” or

"additional"—that the actions that create the ERCs would not have taken place anyway. A further disincentive in ERC programs is the detailed oversight of ERC creation and trading.

An important feature of ERC systems is that they are entirely voluntary (although the underlying regulations are not), which, in practice, has meant that the incentives to create allowances and put them on the market have been weak. Firms generally do not go out of their way to create ERCs since they do not see themselves as being in the ERC business, do not want to invite greater scrutiny by regulators, and prefer to invest capital in their core businesses. When firms do create excess allowances, they often want to keep them to support possible expansions of their own facilities in the future rather than selling them into a market where potential competitors might buy them. Unfortunately, this problem can limit the number of available emission credits, making it hard for new entrants to buy credits to support new businesses. Additionally, the limited number of programs, the terms and conditions associated with ERCs, and the relatively small markets for them have hindered the development of risk management tools for them (e.g., futures and options), limiting their utility. One of the reactions to this problem has been for local governments to obtain ERCs as part of the process approving their creation, which can be used to foster growth (and job creation) by giving them to new companies entering the area.

2.3 Discrete Emission Reductions (DERs)

DERs are created by reducing emissions relative to a baseline and are project based. However, DERs are temporary and can only be used once. DERs can be generated by installation of pollution-control equipment, installation of control equipment with a higher-than-required efficiency, or prior to a compliance date for additional control or a process change. Importantly, DERs are wholly past reductions and are quantified after they have been created. In the United States, several states have rules that allow DERs, but they have different mechanisms for granting credits. For example, some states certify DERs, while others allow for self-certification with third-party verification. A similar term, verified emission reductions (VERs), is sometimes use to describe one-time reductions in greenhouse gas (GHG) emissions, but a key difference is that VERs are created *before* the rules for emission trading are in place. This makes their future validity uncertain.

2.4 Cap-and-Trade Systems (C/T)

The best known MBI is probably the emission allowance used in cap-and-trade (C/T) systems, which, as the name implies, create a permanent limit on emissions, a key virtue for environmental advocates. In a C/T system, the government defines the regulated sources and the total amount of pollution that they can emit, the "cap," over a set period of time, usually 1 year. Typically, the cap is set in mass units (e.g., tons), is lower than historical emissions, and declines over time. The government creates allowances equal to the size of the cap and then distributes them to the regulated sources, a process called allocation. All C/T existing systems allocate allowances based on historical emissions, which is problematic for new entrants, as discussed later.

The government then requires regulated facilities to surrender emission allowances equal to the emissions of the facilities on a periodic basis (sometimes called true-up or covering emissions). The government also sets standards for emissions monitoring, establishes rules for how allowances may be used, and defines enforcement. These are crucial choices, not just details. One particularly important decision is if and how allowances can be saved (or banked) from one period to another, which is discussed in detail later.

Because the allocation to each firm is smaller than its previous emissions, regulated firms have four basic options: (1) control emissions to exactly match their allocation, (2) undercontrol and buy allowances to cover their emissions, (3) overcontrol and then sell their excess, or (4) overcontrol and bank allowances for use in future years (when even fewer allowances will be allocated). The reason companies might buy or sell allowances is that facilities will have different emission control costs or they might change operations so that they needed more (or fewer) allowances. Companies with higher costs will be able to save money by undercontrolling and buying allowances from those with lower costs, which make money by overcontrolling and selling.

The government regulates the trading of emissions allowances differently in various C/T systems. The government usually acts as the accountant for C/T systems by establishing a registry for participants. Usually, participants are required to report the size of transactions and the names of buyer and seller. This can be facilitated by creating a serial number for each allowance. However, there is often no requirement that market participants disclose the price at which a sale was made, nor any requirement that they inform

government of the trade in a timely manner. This lack of information can limit the transparency of the market as participants may delay reporting trades in order to conceal strategic information. Brokerage and consulting firms complete the picture by providing services to market participants, including markets in derivative commodities, and by increasing transparency by providing information (including price information) about the markets. Simplicity in market design and competition among brokers has tended to keep transaction costs low (up to a few percent of allowance prices) in emission allowance markets.

Several key features of C/T systems are worth noting. First, a cap on total emissions means that as an economy grows, new emissions-control technologies or emission-free production processes will be needed. Some observers worry that a fixed emission cap is a limit to economic growth; but so far the evidence shows no such effect. Second, the problem of getting allowances to new entrants can be minimized if an active, liquid market develops or if a small number of allowances are set aside by the government for this purpose. Third, the standardization of C/T allowances authorizes larger emissions markets and permit brokers to offer derivative securities based on them. This has proved important since the ability to use derivative securities like options and futures greatly enhances flexibility and reduces risk.

2.5 Open Market Trading

There have been attempts to allow DERs to be used in C/T systems, a concept called open market trading. Advocates of this approach typically look to create ERCs in the mobile source sector (by buying and scrapping old vehicles or paying for upgrades to cleaner vehicles) to sell to stationary sources. Although the open market trading approach has been attempted several times, it has usually failed, often due to disagreements over credit certification requirements or because this approach could nullify the cap. For example, a prominent open market trading program in New Jersey collapsed in late 2002 after years of work developing it. It is likely that the costs of adequate monitoring and verification for the use of DERs in C/T systems may be high enough to eliminate the value of doing so.

2.6 Renewable Energy Credits (RECs)

An addition to clean air markets is the renewable energy credit (REC), which is created by the production of electricity from renewable sources. RECs can be used to implement a renewable portfolio standard (RPS) cost-effectively. In such a program, RECs are created when renewable electricity is generated while companies that sell electricity are required to surrender RECs equal to a percentage of their sales as defined by the RPS. The logic here is similar to other MBIs—firms will vary a lot in their cost of generating renewable energy; an urban distribution company may find it more difficult to generate renewable energy, for instance, than a rural electric cooperative that has abundant land on which to site wind turbines.

2.7 Implementation Challenges

A number of significant challenges exist in the implementation of clean air market programs. Many of these are discussed in the cases that follow the descriptions presented next.

2.7.1 Environmental Goals

Ensuring that the environmental goals of a policy that uses an MBI (such as that the cap in a C/T program is not exceed) is important, since these are the essential reason for implementing such policies. A key threat is leakage, the migration of regulated activities to locations or sectors that are not controlled by the cap. For instance, if emissions from only from large electricity generators were regulated, the replacement of large power plants with smaller ones could constitute leakage. Migration of polluting industries from one country to another is a major concern of environmentalists and labor alike.

Another important issue is how MBIs might affect ambient air pollution and health risks since most pollutants are not uniformly mixed. A chief concern is over uneven spatial distribution of emissions and health risks (sometimes called the hotspots problem), which poorly designed emission trading systems may ignore or even exacerbate. A similar term is directionality, which is used when there are identifiable upwind and downwind locations. Another concern is that environmental equity goals may ignored by emissions trading systems. While such concerns have stopped some MBI proposals, they have been addressed in the design of others. With one minor exception (a part of the RECLAIM program, discussed later), there is no evidence that emission trading programs have caused environmental equity problems.

2.7.2 Regulatory and Price Uncertainty

Uncertainty is possibly the most important concern of firms in programs that use tradable instruments, and it comes from two basic sources. The first is that regulators will not be able to develop all the rules needed to operate a C/T or ERC program, or that they will change the rules. The second is whether allowances will be available and, if so, at what price. This uncertainty comes from the fact that most mangers (especially in small and medium firms) have little experience with allowance markets or resources to devote to the issue, while having enough allowances (or credits) may be crucial to continued operation. The implication of these uncertainties is that some firms regulated by MBIs may be unwilling to rely on entering the market as either buyer or seller, which makes concerns about a lack of allowances on the market self-fulfilling. However, it is not clear if emissions markets have more or fundamentally different uncertainties than do other markets (such as petroleum or computer chips) that have exhibited both dramatic price and regulatory changes in the past. If not, then the risk management tools developed for these markets can be adapted to manage uncertainty.

2.7.3 Banking

The ability to use allowances or credits left over from previous periods in the future is called banking and can be highly contentious. Note that DERs always involve banking, since they are created by retrospective analyses of one-time emission reductions. The ability to bank allowances is one of the major sources of flexibility in an emission trading program since it allows a firm to time its capital expenditures for process changes or emission controls. Banking also allows for the cheapest emissions reductions to occur first and for time to apply research to lower the costs of the more expensive cases.

2.7.4 Allowance Allocation

The creation of a C/T system produces a new type of asset, the allowance, which can have considerable value. For instance the SO_2 allowances given to the U.S. electric power industry each year has a value of more than $1.5 billion. However, a cost-free allocation based on historic emissions (called grandfathering) fails to internalize the external costs of pollution and give a weaker incentive for innovation. It also gives existing firms an inherent advantage over new entrants, who must buy allowances, often from potential competitors. Another important problem with cost-free allocation is that it adds to uncertainty about the market price of allowances.

Although all C/T systems in use so far in the United States have used grandfathered allocations, in principle allowances could be auctioned or distributed in other ways (e.g., distributed based on current production). Firms typically resist auctions, which will always cost more money, in effect arguing that their previous emissions entitles them to the new assets. This approach will also reduce the role of emissions brokers, since there would be less need for interfirm sales after such an auction.

One way to deal with the new entrant and price uncertainty problems is for the government to set a small portion of allowances aside for sale to new entrants and for early auctions to aid in what brokers call price discovery by providing evidence of what prices firms are willing to pay for allowances. Auctions could also be used if prices rise to levels that create significant hardships for a few market participants. These goals can also be accomplished with relatively small set-asides. A slightly different technique that has been proposed under the term "relief valve" is to have government create and sell as many allowances as firms wanted to buy for use that period at a predetermined value. The activation of such a safety valve would preclude banking for that year. This would break the cap of a C/T system for that year, obviously, but it would limit the costs of control and would have the desirable properties of an emissions tax.

2.7.5 Monitoring

Most emission trading programs are applied to stationary sources such as electric power plants, refineries, and manufacturers, partly because it is usually feasible to monitor the emissions of these sources accurately. Estimates of emissions are allowed in some cases, for instance, when calculating DER and ERC quantities since this always involves estimating a counterfactual baseline. Larger sources also have economies of scale in managing an emissions allowance account. Good monitoring and accounting practices reduce concerns about fraudulent (or simply flawed) sales of emission allowances and help ensure their environmental integrity. Fossil carbon in fuels is already accounted for and traded routinely in fuels markets, which may make it easier to monitor the withdrawal of fossil carbon from the ground than its release (as carbon dioxide, CO_2) to the atmosphere. However, such an upstream approach is probably politically impossible, especially internationally

2.7.6 Double Counting and Additionality

It is important for ERCs that the activities that create them are not already required by some law or regulation (additionality) and that they are not counted twice in two different ERC programs (double counting).

2.7.7 Market Transparency

It is not always easy for regulated firms to observe the operation of emission markets in detail, reducing their confidence in them. Unlike trading in securities (e.g., stocks, bonds, and futures contracts), emission trading markets are not regulated uniformly and often not very closely. In new or very small programs, emissions markets may be characterized by bilateral trades, generally put together by a consulting company and approved by government. Large emissions markets typically use computer systems that a handful of brokerage firms use, on which bids from regulated companies (which can number in the dozens to hundreds) to buy and sell allowances are posted. Regulated companies often do not have direct access to these markets, but work through brokers (who usually also trade in electricity, fuel, and sometimes other commodities). A few specialized newsletters survey brokers on market data and publish the results. In many cases, the environmental regulator will post some transaction information on the Internet, although just what information they put up varies.

Problems with market transparency include the lack of regulatory oversight of brokers (from which most of the price information comes), the potential for bilateral transactions to go unreported for long periods, and sometimes a difficulty in obtaining market information from government. These problems add to price uncertainty.

2.7.8 Incentives for Credit Creation

The creation of an ERC or C/T system does not automatically lead firms to reduce emissions for the purpose of selling credits or allowances. Managers may not see themselves as having any particular advantage in creating allowances, relative to their competitiveness in their product markets, or may simply not have the resources to devote to studying the problem. Those they do create may be considered vital to future expansion, so the firm may be unwilling to sell them.

2.7.9 Buyer/Seller Liability

There is some concern about what happens if an allowance or credit trade occurs and the traded instruments are later found to be invalid: Who is liable, the buyer or seller? What if the seller has gone bankrupt or disappeared? Questions associated with this issue are most important in international arenas, in ERC and DER programs, and in C/T systems without serialized allowances.

2.7.10 Enforcement

MBI programs typically include enforcement provisions. Enforcement for ERCs and DERs is based on the underlying CAC regulations, requiring action and judgment by the regulator, and often involves the courts. In contrast, C/T systems tend to have more automatic enforcement provisions, including such features as fines based on the cost of allowances, a requirement to surrender the next period with a multiple of the gap between actual emissions and allowances held at true-up. Such provisions reduce uncertainty and have led to extremely high compliance in the large C/T programs in the United States, where noncompliance has been limited to a handful of minor oversights, with one important exception (RECLAIM) discussed later.

3. EXAMPLES

Four emission trading systems for controlling air pollution in the United States will be examined, plus two climate change programs in Europe. This is by no means the entire set of emission trading programs in use. Others include pilot programs in Canada, China, and Europe, and regional efforts like the Houston/Galveston Area Mass Emission Cap & Trade program.

3.1 State ERC Programs in the United States

The original introduction of MBIs was the result of a U.S. Environmental Protection Agency (EPA) requirement that new facilities wishing to locate in areas with unacceptable air quality "offset" their emissions. The only way to do this was to obtain offsets from existing firms. This suddenly turned the legacy of emitting pollution into an asset that could be sold to firms entering or expanding in an area. This was late expanded to include a number of similar provisions called bubbling, netting, and banking, that improve the efficiency of CAC regulation. The banking provision enables many ERC markets.

State air pollution agencies typically manage these systems with guidance and oversight from the EPA.

Environmental groups have had an important role in ensuring the environmental integrity of these programs. Pollutants in these programs include volatile organic compounds (VOCs), nitrogen oxides (NO_X), sulfur dioxide (SO_2), and others. Price and availability for ERCs vary greatly by pollutant, location (ERCs cannot be transferred between different urban areas), and sometimes by other specific conditions attached to ERCs. These programs have created significant intrafirm trades (e.g., bubbling), but not many market transactions since they are somewhat cumbersome and have poor incentives for the generation and sale of credits.

3.2 U.S. Acid Rain Program (Title IV)

The best-known C/T system is the EPA's Acid Rain Program for SO_2 emissions from coal-fired power plants. Key features include the strict monitoring provisions for both SO_2 and NO_X; the national scope of the program; the relatively deep cuts in emissions (50%); completely unrestricted trading and banking; a small auction program in the early years of the program; and, most important, its success. Facilities regulated by the Acid Rain Program must still comply with health-based CAC SO_2 regulations that prevent hotspots from developing, although these restrictions have not affected the market.

The Acid Rain Program has been a success in several ways. First, substantial emission reductions have occurred. From 1990 to 2002, SO_2 emissions from regulated sources declined by about one-third. During the less-stringent Phase I (1995–1999), regulated sources overcontrolled and banked more than a year's worth of allowances, which they began to use in Phase II. This bank will be empty in about 2010.

Second, the program has greatly reduced the cost of SO_2 control compared to command-and-control polices. In the first 5 years, emissions trading reduced compliance costs by about one-third to one-half, and estimates of the savings range from $350 million to $1400 million. Allowance prices have thus been lower than forecast, ranging from $66/ton to about $200/ton. This is not to say that the cost of SO_2 control has been cheap. In 1995, annual costs were about $726 million, and capital costs for scrubbers in Phase I alone are estimated at $3.5 billion. However, most of these savings are not due to trading of allowances per se, but from the flexibility in compliance that allowed firms to find their own least-cost approach. An important development was that Midwestern power plants designed to burn high-sulfur local coal were adapted to burn low-sulfur Western fuel (from the Powder River Basin) just as it was becoming cheaper due to railroad deregulation.

Third, the SO_2 market has been a success. Prices in this market are relatively reliable because there are up to several dozen trades each day, resulting in from 20,000 to 100,000 allowances trading hands each week and considerable market information. All vintages of allowances are priced the same because there are no restrictions on banking, which helps smooth the operation of the market. An important feature of this market was auctions of set-aside allowances starting in 1992 and 1993, several years before the first compliance year. Although the design of these early auctions market has been criticized, they were extremely valuable because they helped with price discovery in an untested and highly uncertain market. The revenue from these auctions was returned to the regulated firms in proportion to their allowance allocation. These auctions also allowed for new entrants and the public to participate in the market. (Several school and public interest groups bought allowances in order to retire them.) This system also had an attractive program for early emission reductions, while some state programs encouraged early investment in emission control equipment (scrubbers), which made some allowances excess and created natural sellers. All of this helped the industry build up a bank of allowance before the first compliance year, reducing regulatory and price uncertainty.

Several important findings about the operation of allowance markets and industries regulated by a C/T system have emerged from the Acid Rain Program. First, market participation and compliance strategies have evolved, from an autarkic approach toward a greater and greater reliance on the market. Second, allowance prices have shown considerable volatility, and the lower bound of emissions prices have been shown to be equal to the marginal cost of operating emission control devices (scrubbers). Third, the relatively few units that did install scrubbers increased their utilization and lowered their emissions beyond original design specifications as a result of the incentives in the allowance market.

3.3 Ozone Transport Commission NO_x Budget

The first multilateral C/T system to go into effect is the Ozone Transport Commission (OTC) NO_X budget program, covering eleven states in the Eastern

Seaboard and the District of Columbia. The NO_x budget applies to power plants and large industrial facilities and covers emissions from May through September. There are more than 470 individual sources in the program, with about 100 owners. Beginning in 1999, the NO_x budget will eventually reduce emissions by 65 to 75% through a three-phase approach. Importantly, the OTC NO_X budget is not a centrally organized system (like the Acid Rain Program), but rather the result of coordinated laws and rules in each state, although the federal EPA provides some assistance (e.g., by tracking the allowances).

The states in the OTC were able to coordinate their individual laws and regulations to produce a successful program for a number of reasons, the most important of which were that they had similar interests in NO_X control, they had a history of cooperation in this area, and they engaged in a multiyear technical/political negotiation process that they all felt was fair. The result of this process was a model rule that could be adopted by each state with modifications to account for local differences without endangering the political or environmental integrity of the agreement.

This process illustrates how complex regulatory uncertainty can be. The development of the NO_X budget model rule began in late 1994, at which point the level and timing of emission reductions was set. The technical/political negotiation process that followed was designed to determine if emission trading would be used or not. In early 1996, the model rule was published and the OTC states began to develop their own rules, which were all final by the end of 1998 to allow the NO_X budget to start in May 1999 (except for Maryland, which entered a year late). Regulated sources felt this was actually rather short-fused, because the rules were not in place in several states until less than a year was left before the start of the first ozone season, while contracting for engineering and constructing NO_X control technologies can take several years. Regulators disagree with this assessment, noting that if the emission trading program had not come together, similar command-and-control regulation would have been the default. Further, much of the delay in developing state rules was due to the need to address the concerns of regulated industries, which then complained about the uncertainty.

The OTC NO_x budget has some distinctive features. First, it had no early auctions or other methods for price discovery before it went into effect. Second, the banking of allowances in the NO_x budget can be restricted in a way that can reduce the face value (in tons) of banked allowances, but firms do not know if or how much of a discount will be applied until after they have banked the allowances. This creates uncertainty since facilities cannot be sure exactly what their banked allowances will be worth in the future. Thus, banked allowances sell at a discount.

These factors led to a significant spike in allowance prices in the OTC NO_x budget, up to more than $7000/ton, far above the cost of control for any regulated sources. Prices stayed high for several months, but by July they had fallen back to the predicted range and by the end of the year fell to around $1000/ton. This volatility had nothing to do with the cost of emission controls, but was the result of the supply and demand changes in the allowance market. Importantly, neither state governments nor the regulated firms abandoned the allowance market by changing the rules or seeking regulatory relief in the courts. Since that first, difficult year, the OTC NO_x budget has matured; emission reductions have been substantial, compliance failures have been trivial, a large bank of NO_X allowances has been built up, and price discovery for allowances in the next, more stringent phase of the is now underway.

Subsequently, a failed attempt was made to develop a multilateral C/T program larger set of eastern states. This failure suggests several important factors are necessary. Probably the most important factor is that potential participants recognize that they have similar interests in controlling pollution cost-effectively. Secondly, they must be able to develop a process for creating and enforcing the C/T system that cannot be used against their interests.

3.4 California RECLAIM

The last U.S. example is California's Regional Clean Air Incentives Market (RECLAIM), a C/T program for SO_2 and NO_X emissions from industrial sources such as power plants, refineries, and metal fabricators. An important feature of RECLAIM is that it does not permit banking, since government regulators felt this would compromise its environmental integrity. When RECLAIM was first implemented in 1994, the cap was generous, allowing for an increase in emissions over historical levels for many sources, but it declined steadily each year, aiming at an overall reduction of about 75% by 2003. An example of an environmental equity problem in a C/T program is the mobile source opt-in provisions of the RECLAIM program. Residents living near several refineries were

able to have these provisions overturned because they would have allowed the refineries to avoid reduce smog-forming emissions, which were also toxic pollutants.

For the first several years, the RECLAIM market functioned quite well, with readily available allowances at low prices. However, emissions in 1993 to 1998 did not decline nearly as fast as the cap due to a failure of many (but not all) participants to install emission control equipment. Although the state regulatory agency amply warned participants of a looming problem, many firms were unwilling to take appropriate actions because of a failure to consider future emission allowance markets and a belief that the government would bail them out in case of serious problems.

The result was a breakdown of the market and in response a temporary abandonment of the MBI approach by the state government. By early 2000 it had become clear to even the most shortsighted that emissions would exceed allocations, which was a problem because RECLAIM had no banking provision, and prices for NO_X allowances skyrocketed to over $40,000/ton. Electricity companies, which were making record profits at the time, could afford these prices, but other companies in the RECLAIM market could not. Thus, the RECLAIM cap was broken, and several firms were significantly out of compliance and paid record fines. This is the only failure of a C/T system to date. Facing significant political pressure, the state regulatory agency decided to essentially go back to a CAC approach for electric power plants by requiring them to submit compliance plans. This is particularly important in that most cost savings in C/T systems come from the ability to innovate in compliance strategy, not from buying or selling allowances. In addition, state regulators separated power companies from the rest of the RECLAIM market and subjected them to a high tax for emissions not covered by allowances. For other participants, RECLAIM proceeds as before and allowance prices have moderated.

Several key lessons emerge from the RECLAIM experience. First, because they force firms to gather more information and make more decisions, MBIs may be *more* difficult for firms to understand and manage than CAC programs, even if they have lower costs. This is especially true for smaller companies, some of which may even have increases in monitoring costs required by RECLAIM that were greater than the savings in control costs. Second, in some cases the optimal strategy may be noncompliance, placing more emphasis on the design of penalties. Third, emission markets are no different from others; they are volatile (especially when it is not possible to store the commodity, like electricity).

3.5 Denmark CO_2 Program

The first emission trading program to address climate change was a C/T system for CO_2 emissions adopted by Denmark in 1999 to achieve a national GHG emission reduction target of 5% in 2000 (relative to 1990) and 20% by 2005. The Danish program covers the domestic electricity sector (electricity imports and exports are treated separately), which is made up of eight firms, although two account for more than 90% of all emissions. Allocations were made on a modified historical basis and are not serialized. Banking is limited and there is some uncertainty about the validity of allowances beyond 2003. This program includes a tax of about $5.5/ton for emissions that are not covered by an allowance, which means the integrity of the cap is not guaranteed. It is possible to use VERs as well as credits created through provisions of the Kyoto Protocol (discussed later). Monitoring is accomplished through an analysis of fuel consumption—in-stack monitors are not required.

The C/T program was adopted to regulate the last major unregulated set of CO_2 emitters so that Denmark could reach its domestic CO_2 goals cost-effectively while also gaining experience with emission trading. In the first 2 years of the program, about two dozen trades were made in the $2 to $4/tonne range. This has resulted in more than half a million allowances changing hands. Some of these trades have been swaps of Danish allowances for VER credits, and one trade of Danish allowances for U.K. allowances (discussed later), the first instance of a trade of two government-backed MBIs. However, due to the small size of a market, it is unlikely that an "exchange" for Danish CO_2 allowances will emerge; instead most trades will probably be bilateral and the Danish system will probably be superseded by a European Union (EU) system.

3.6 United Kingdom Climate Change Levy and Emission Trading System

The first economy-wide GHG control policy was announced in November 2000 by the United Kingdom, and it contained a combination of MBIs, including taxes, subsidies, and ERCs. This program is designed both to achieve significant emission

reductions and to provide U.K. organizations experience in emission trading, not least so that the City of London might become a leader in this activity.

The first part of this policy is the climate change levy (CCL) on fossil fuel energy supply for industrial users that came into effect in April 2001 and typically adds 15 to 20% to the cost of energy. The CCL has moderately increased the demand for energy efficiency, and the revenue it generates has been recycled into a fund that subsidizes consulting and capital purchases to reduce energy use. Moreover, it provides an important basis for voluntary participation and credit generation in the ERC program, possibly overcoming one of the key problems associated with past ERC programs.

The second component is the Emission Trading System (ETS), an ERC program open to reductions in all GHGs (measured in CO_2-equivalent, or CO_2e). To join the program, firms must enter into a climate change levy agreement (CCLA) in which they voluntarily accept an emissions cap in return for an 80% reduction in their climate change levy until 2013. Companies that adopt such a target may sell credits generated by exceeding their target.

The last component, direct entry, is a £215m ($310 million) subsidy that the U.K. government made available through an auction for voluntary actions by eligible firms to reduce GHG emissions in 2002–2006 from a 1998–2000 baseline and generate ETS credits. This auction is designed to obtain the maximum emission reduction and provide some price discovery by allowing eligible companies to bid a fixed amount of emission reduction for a given price and selecting (through repeated rounds) the price that yielded the maximum emission reduction given the government's budget. Most electricity and heat production were not eligible for this program, which made bids for reductions of non-CO_2 GHGs more likely and created a natural set of buyers (energy companies) and sellers (manufacturers and end users) for ETS credits. This auction was held via the Internet over 2 days in March 2002 and resulted in 34 bidders (of 38) winning subsidies at the level of about $22/ton-$CO_2$e. Over half of the emissions will be non-CO_2e GHGs.

Organizations can also join the ETS through more traditional means by creating ETS credits through a specific project that meets all the necessary monitoring and verification requirements. Over time overseas-sourced ERCs (similar to those in anticipated in the Kyoto Protocol, as discussed later) may be allowed, but the U.K. government is going slowly to reduce uncertainty in the early stages of the market. By the end of 2002, more than 400 companies had opened accounts on the U.K. registry and about 1 million credits had been exchanged in several hundred transactions. Prices on this market are in the range of $5 to $10/ton-$CO_2$e.

4. LOOKING AHEAD

4.1 Multipollutant and Toxics

By 1999, European nations adopted an international treaty that deals with multiple pollutants and multiple impacts holistically, and many multipollutant emissions control proposals have been made in the United States. The reason for including multiple pollutants in one law is that this would greatly reduce the regulatory uncertainty for companies (especially electric power producers) that have been subject over the past 30 years to a long sequence of ever tighter control requirements on different pollutants. Such an approach is very expensive, especially for capital designed to last several decades. Capital investment in emission control retrofits and new facilities will be much less risky and probably less expensive for firms if they can count on flexible MBIs due to the flexibility in compliance strategy and risk management opportunities. Thus, industry may seek to replace uncoordinated and rigid CAC regulation with MBIs that will remain fixed for an extended period (a concept called the safe harbor). Typically, public and environmental groups press for more stringent regulations.

Major issues in considering multipollutant emission include whether or not to include toxic pollutants, such as mercury (Hg), and GHGs, such as CO_2, as well as what level of controls and predictability to include. In the United States, the term "3-P" applies to legislation that includes SO_2, NO_X, and Hg, and "4-P" applies to legislation that adds CO_2. Importantly, different environmental problem will almost always have different levels of consensus, both scientific and political, at any given time so the length of time a safe harbor may be different for various pollutants. Another is that there is no consensus exists about whether it is appropriate to allow emission trading to apply to toxics, since the issue of hotspots may be significant. Finally, achieving the emission reductions envisioned in 3-P and 4-P control will be serious technological challenges, especially for coal-fired electricity generation plants. The interaction of technologies to control all four emissions is highly complex. For

instance, controlling SO_2 and NO_X tends to reduce power plant efficiency and thus increase Hg and CO_2 emissions, but some SO_2 and NO_X emission control technologies may also control Hg. The flexibility of MBIs will be even more important in multipollutant approaches since different facilities will likely have extremely varied least-cost compliance options.

4.2 Carbon Dioxide and the Kyoto Protocol

Probably the largest application of MBIs may be to control GHGs under the United Nations Framework Convention on Climate Change and the subsequent Kyoto Protocol (KP). The Framework Convention is an agreement of virtually all nations of the world to avoid "dangerous interference to the climate system," without committing signatories to any specific actions. Subsequently, negotiations produced the KP, which provides specific GHG emission caps for more than three dozen industrialized countries. These caps, combined with relatively lenient provisions for biological sequestration (e.g., credits for regrowing forests) and the decisions by the United States and Australia, create the background for international emissions trading in the near term. The original goal of the KP was to reduce emissions from industrialized countries to 5% below 1990 levels during 2008–2012. With the more recent provisions and participation, emission reductions of 1 to 2% are expected. Nonetheless, this is a substantial change from business as usual, which would have resulted in an increase of several percent.

The KP contains both C/T and ERC features. The C/T component addresses the trading of "assigned amounts" among industrialized countries, while the ERC components are mechanisms designed to allow trading between industrialized countries covered by an emissions cap and other countries that are not. These are called joint implementation (JI) and the clean development mechanism (CDM), and, like all ERC instruments, they are project specific. By the early 2000s, the rules for these provisions had still not been agreed upon, creating great regulatory uncertainty.

Despite uncertainties about future baselines, monitoring requirements, time frames, and other factors, many companies already began to engage in CO_2 emissions trading in 1996–2002. Most of these are VERs that have sold for $2 to $20/ton of CO_2. More than 1 billion such units have been traded. Although it is not clear that regulators will accept these credits in the future, there seem to be two values to these activities. First, companies that participate are learning how to measure their own CO_2 emissions and how to structure deals in this market. Second, although they are unofficial, these early CO_2 emission trades may be counted at least partially against future compliance obligations.

Although there is not yet agreement on emission trading rules for the KP as a whole, in October 2001 the European Union (EU) produced a directive to establish a framework for a mandatory EU-wide C/T system beginning in 2005 in order to implement the KP. Initially, only approximately 46% of estimated EU CO_2 emissions in 2010 were to be included. Between 4000 and 5000 installations from the following activities will be covered: electricity and heat generation (greater than 20 megawatts), petroleum refining, and the manufacture of iron, steel, cement clinker, ceramics, glass, and pulp and paper. Excluded sectors include oil and gas production, solid waste incineration, and chemical manufacturing, as well as transportation and residential energy consumption. The proposed system would allow for trading between companies, to be tracked by national governments that would establish registries for CO_2 emission allowances. A trade across a national boundary would require corresponding entries (one addition and one subtraction) in the two national registries involved. Banking from one year to the next is allowed, including from one commitment period to another. The EU CO_2 C/T system is also important because it sets a key precedent for international CO_2 emission trading. The EU would create the largest CO_2 emission market in the world, and other countries that wanted access to that market would have to follow the procedures set down by the EU. Expectations are that EU CO_2 credits would cost $2.5 to $10/ton before 2005 and $5 to $20/ton by 2010.

SEE ALSO THE FOLLOWING ARTICLES

Acid Deposition and Energy Use • Air Pollution from Energy Production and Use • Air Pollution, Health Effects of • Carbon Taxes and Climate Change • Fuel Economy Initiatives: International Comparisons • Greenhouse Gas Abatement: Controversies in Cost Assessment • Hazardous Waste from Fossil Fuels • Market-Based Instruments, Overview • Modeling Energy Markets and Climate Change Policy • Taxation of Energy

Further Reading

Ellerman, A. D., Joskow, P. L., and Harrison, D. (2003). "Emissions Trading in the U.S: Experience, Lessons, and Considerations for Greenhouse Gases." Pew Center on Global Climate Change: Arlington, VA.

Ellerman, A. D., *et al.* (2000). "Markets for Clean Air: the U.S. acid rain program." Cambridge University Press, Cambridge, UK.

Farrell, A. E. (2001). Multi-lateral emission trading: Lessons from inter-state NO_x control in the United States. *Energy Policy* **29**(13), 1061–1072.

Foster, V., and Hahn, R. W. (1995). Designing more efficient markets: Lessons from Los Angeles smog control. *J. Law and Economics* **38**(1), 19–48.

Grubb, M., Vrolijk, C., and Brack, D. (1999). "The Kyoto Protocol: A Guide and Assessment." Earthscan, London.

Israels, K., *et al.* (2002). "An Evaluation of the South Coast Air Quality Management District's Regional Clean Air Incentives Market—Lessons in Environmental Markets and Innovation." U.S. Environmental Protection Agency Region 9, San Francisco.

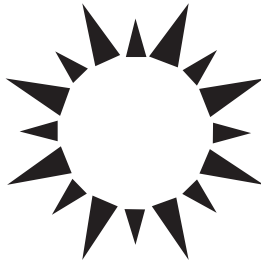
Raufer, R. (1996). Market-based pollution control regulation: implementing economic theory in the real world. *Environmental Policy and Law* **26**(4), 177–184.

Solomon, B., and Lee, R. (2000). Emissions trading systems and environmental justice: why are some population segments more exposed to pollutants than others? *Environment* **42**(8), 32–45.

Stavins, R. (2003). Experience with market-based environmental policy instruments. *In* "The Handbook of Environmental Economics" (K. Goren-Maler and J. Vincent, Eds.). North-Holland/Elsevier Science: Amsterdam.

Clean Coal Technology

MILDRED B. PERRY
U.S. Department of Energy, National Energy Technology Center
Pittsburgh, Pennsylvania, United States



1. **Impact of Clean Coal Technology**
2. **Background of Development**
3. **International Clean Coal Program**
4. **United States Program**
5. **Benefits and Future of Clean Coal Technology**

Glossary

acid rain When coal is burned, the sulfur and nitrogen in the coal are converted to sulfur dioxide (SO_2) and nitrogen oxides (NO_x). In the atmosphere, these substances are converted to sulfuric acid and nitric acid, which dissolve in raindrops to form acid rain.

carbon dioxide sequestration The storage of carbon dioxide (CO_2) in underground reservoirs or in deep-sea locations to prevent its release into the atmosphere. CO_2 can also be sequestered by enhancing the growth of forests and other vegetation.

clean coal technology Processes developed to reduce the emissions of pollutants from coal combustion; include technologies to remove pollutants and technology to improve efficiency, which also decreases emissions.

clean coal technology by-products Useful substances created during clean coal production; include sulfur and sulfuric acid, which have many applications; gypsum, used in the manufacture of wallboard; fly ash, used in the manufacture of cement; and aggregate, used in the construction industry.

flue gas desulfurization A clean coal technology consisting of a device, called a scrubber, fitted between a power plant's boiler and its stack, designed to remove the sulfur dioxide in the flue gas.

fluidized bed combustion (FBC) A process in which pulverized or granulated fuel and air are introduced into a fluidized bed of sand or some other material, where combustion takes place. Fluidized beds are classified as bubbling or circulating, depending on whether the bed material remains in place or is transported out of the vessel by the fluidizing gas, recovered, and returned to the bed. Plants may operate either at atmospheric pressure (AFBC) or at elevated pressure (PFBC).

fly ash Fine particles of ash that are entrained with the flue gas when coal is burned in a furnace.

greenhouse gases (GHGs) Gaseous compounds that act to trap heat in the atmosphere, thus tending to raise global temperature. The main GHG is carbon dioxide (CO_2). Other important GHGs are methane (CH_4) and nitrous oxide (N_2O).

gypsum A mineral having the chemical composition $CaSO_4 \cdot 2H_2O$, used in the manufacture of wallboard.

integrated gasification combined cycle (IGCC) A conversion process in which coal is first gasified to form a synthesis gas (a mixture of hydrogen and carbon monoxide), which is used as fuel for a combined-cycle power plant (gas turbine/steam turbine).

low-NO_x burner A burner that is designed to burn fuel in such a way so as to reduce the amount of NO_x produced.

NO_x The symbol for nitrogen oxides, consisting of a mixture of mainly NO and NO_2. NO_x is formed when coal is burned, partly from oxidation of the nitrogen in the coal and partly from reaction of nitrogen and oxygen in the combustion air.

overfire air The air introduced into a furnace above the combustion zone with the objective of completing combustion; used in conjunction with low-NO_x burners and reburning.

reburning A technique for reducing NO_x emissions consisting of introducing a hydrocarbon fuel into a reducing zone above the combustion zone to react with the NO_x and convert it to N_2.

repowering Rebuilding or replacing major components of a power plant as an alternative to building a new plant, often used to increase plant capacity.

scrubber A device designed to remove (scrub) sulfur oxides from flue gas, usually through contact with a limestone slurry to form calcium sulfate (gypsum).

selective catalytic reduction (SCR) A process installed downstream of a boiler, consisting of a catalytic reactor, through which the flue gas passes. A reducing agent, typically ammonia, is introduced into the flue gas upstream of the reactor and reacts with the NO_x in the flue gas to form nitrogen and water.

sulfur oxides When a fuel containing sulfur is burned, most of the sulfur in the fuel is converted to SO_2, with a small fraction being converted to sulfur trioxide (SO_3). In the atmosphere, sulfur oxides are converted to

sulfuric acid, which is washed from the air and falls to the ground as acid rain, or converted to sulfate, which is a fine particulate.

synthesis gas (syngas) A mixture of hydrogen and carbon monoxide, used as an intermediate in the production of a number of chemicals. Syngas can also be used as a fuel.

trace element An element that appears in very small quantity (typically less than 10 parts per million) in coal. Because coal is a sedimentary rock, during its formation, it accumulated mineral matter, which is the source of most trace elements.

water–gas shift reaction The catalytic reaction of H_2O with CO to form H_2 and CO_2, used to increase the hydrogen content of synthesis gas.

Clean coal technologies can have a major impact on the world's economy by allowing the continued use of coal in an environmentally acceptable manner, thus reducing dependence on more expensive petroleum and natural gas, while avoiding economic problems caused by natural gas price fluctuations. Although natural gas is expected to fuel many new power plants and industrial facilities to meet growing demand over the next two decades, it is critical to maintain fuel flexibility, thus enabling the use of low-cost indigenous fuels, such as coal.

1. IMPACT OF CLEAN COAL TECHNOLOGY

Coal is one of the world's most abundant fossil fuels. It is also widely distributed, and many countries with little or no petroleum reserves have coal resources. Unfortunately, in addition to carbon and hydrogen, most coals contain sulfur, nitrogen, minerals, and chlorine, in various concentrations. Coals also contain trace elements, such as mercury, lead, arsenic, cadmium, selenium, thorium, and uranium, in very small concentrations, generally only a few parts per million. When coal is burned, the trace elements may be released to the environment in various, sometimes harmful, forms. In the past, coal combustion has contributed to acid rain, smog, forest die off, eutrophication of lakes, and other environmental problems. Because such impacts are no longer acceptable, new methods to reduce harmful emissions from plants utilizing coal are needed. Technology to accomplish this is generally referred to as clean coal technology (CCT) and includes technology designed either to remove or to reduce emissions, as well as technology designed to improve power plant efficiency. Efficiency improvements reduce pollution by decreasing the amount of coal that must be consumed to produce a given amount of power. Less coal burned means less pollution.

2. BACKGROUND OF DEVELOPMENT

Concern over environmental problems in the 20th century prompted a number of countries to address the problem of acid rain, resulting mainly from emissions of SO_2. In the United States in 1980, an independent organization, the National Acid Precipitation Assessment Program (NAPAP), was created, to coordinate acid rain research activities and periodically report to Congress on all aspects of the acid rain issue. In 1982, Canada proposed to reduce SO_2 emissions by 50%. In 1984, the United States, Poland, and the United Kingdom joined the "30% Club," a group of 18 European nations whose goal was to cut emissions of SO_2 by 30% or more by 1997. A similar agreement, the Helsinki Protocol, was later signed by 21 nations. In 1986, Canada and the United States assessed the international issues associated with transboundary air pollution and recommended actions for solving them, particularly with respect to acid rain. This entailed substantial research and financial commitments. Because of these efforts, SO_2 emissions were markedly reduced. For example, by the 1990s, emissions had dropped in most of Western Europe by 40% and in West Germany by 70%. Unfortunately, by that time, many forests and lakes worldwide had been damaged. Because of coal's importance to the economy, the United States government has been involved in the development and promotion of CCTs for over 20 years. In 1986, in response to regulatory requirements, the U.S. Department of Energy (DOE) initiated its original Clean Coal Technology Program, which consisted of five rounds of project funding. This program sponsored projects in the following areas: environmental control devices, advanced electric power generation, coal processing for clean fuels, and industrial applications.

3. INTERNATIONAL CLEAN COAL PROGRAM

Coal continues to be a vital fuel resource in numerous countries around the world. In order to maintain the viability of coal production, combustion, and conversion, a variety of international CCT research

programs have been established to further enhance coal's utilization efficiency and environmental acceptability. Advanced technologies increase the amount of energy gained from each ton of coal consumed while reducing emissions and waste. This ongoing international research ensures that advances that focus on the needs of producers and end-users are continually being made.

Because of concern for coal's impact on the environment, many governments have been instrumental in funding, at least partially, the development of CCTs. Some of the more significant international programs and their goals are discussed in the following sections.

3.1 International Energy Agency Coal Research Clean Coal Center

The International Energy Agency (IEA) was established in 1974 within the framework of the Organization for Economic Cooperation and Development (OECD). A basic aim of the IEA is to foster cooperation among the IEA participating countries in order to increase energy security through diversification of energy supply and to promote energy conservation and cleaner and more efficient use of energy. This is achieved, in part, through a program of collaborative research and development of which IEA Coal Research, established in 1975, is by far the largest and the longest established project. IEA Coal Research is governed by representatives of its member countries. The IEA Clean Coal Center supports projects related to efficient coal supply and use, including technologies for reducing emissions from coal combustion.

3.2 United States

For more than 20 years, the U.S. DOE has had a very active program in the development and promotion of CCTs. This program is summarized in Section 4.

3.3 European Commission

The European Commission Clean Coal Technology Program encompasses projects performed in a series of research programs that promote the use of clean and efficient technologies in industrial plants using solid fuels, with an overall goal of limiting emissions, including CO_2, and encouraging the deployment of advanced clean solid fuel technologies, in order to achieve improved operations at affordable cost. A wide range of technologies is supported, including coal-cleaning plants for upgrading coal; handling, storage, and transport facilities; combustion and conversion plants; and waste disposal.

3.4 The United Kingdom

The U.K. Cleaner Coal Technology Program directly implements U.K. government policy for research and development (R&D) and technology transfer efforts related to environmentally friendly coal utilization technologies. This program was initiated through the Department of Trade and Industry (DTI) as a part of the government's Foresight Initiative. The DTI has established a 6-year collaborative program of activities that link R&D with technology transfer and export promotion activities. The overall aim of the program is to provide an incentive for U.K. industries to develop cleaner coal technologies and obtain an appropriate share of the growing world market for these technologies. The program objectives are to assist industry in meeting the technology targets for advanced power generation, to encourage fundamental coal science research in universities, to encourage the development of an internationally competitive clean coal industry and promote U.K. expertise and know-how in the main export markets, and to examine the potential for developing the U.K. coal bed methane resource and underground coal gasification technology.

3.5 The Netherlands

The Clean Fossil Fuels unit of the Energy Research Center of The Netherlands, the largest research center in The Netherlands concerned with energy, develops concepts and technology for cleaner and more efficient use of fossil fuels. The unit contributes to the transition toward sustainable energy based on three basic programs: climate-neutral energy carriers, emission reduction technologies (including CO_2 sequestration), and energy and environmental research and assessment. The climate-neutral energy supply program investigates decarbonization of fossil fuels, CO_2 capture technologies for power generation and hydrogen production systems, and novel hydrogen production technologies. The emissions reduction program is developing technologies, such as catalytic reduction of NO_x and N_2O, aimed at reducing the emissions of hazardous compounds to the environment. The objective of the energy and environmental research and assessments program is quantification of human influences on the environment as a result of fossil fuel use.

3.6 Australia

Several Cooperative Research Centers (CRCs) in Australia deal directly with the enhanced utilization of coal for power generation to decrease environmental impact and increase conversion efficiency. The CRC for Clean Power from Lignite has as its principal objective the development of technologies to reduce greenhouse gas emissions from lignite-fired power stations while enhancing Australia's international competitiveness by lowering energy costs. The technologies under development relate to both current power stations (pulverized coal-fired boilers) and to high-efficiency advanced cycles. Programs are in four areas: coal–water beneficiation, coal combustion and gasification, fluid bed process development, and ash and deposit formation.

The CRC for Coal in Sustainable Development, an unincorporated joint venture consisting of 18 participating organizations from the black coal-producing and -using industries and research organizations in Queensland, New South Wales, and Western Australia, has as one of its key goals to conduct research and development to improve the environmental performance of current technologies and reduce risks inherent in adopting emerging CCTs, such as pressurized fluidized bed combustion (PFBC) and the integrated gasification combined cycle (IGCC) process.

3.7 Japan

Since its establishment in June 1989, the Center for Coal Utilization, Japan (CCUJ) has been actively promoting the development of CCTs to utilize coal efficiently and cleanly, as well as the development of technical skills in the field of coal utilization. The activities of CCUJ have been extended to encompass sustainable economic growth and improvement of the global environment for the benefit of all industries that use coal, including power generation and the steel and cement industries. Efforts have been focused on development of more economical and efficient coal utilization systems with reduced emissions of CO_2, SO_2, NO_x, and particulates. CCUJ is promoting technical innovation in coal utilization and international cooperation in order to disseminate CCT information overseas, especially in Asian countries, in order to achieve a stable energy supply, sustainable economic growth, and an improved global environment. Some of the projects being sponsored are advanced PFBC, advanced coke making, coal gasification, and integrated coal gasification/fuel cell combined cycle (IGFC).

The New Energy and Industrial Technology Development Organization (NEDO) established the NEDO Clean Coal Technology Center in 1992 to carry out research and development of next-generation coal utilization technologies to reduce the amount of CO_2 and other emissions released to the atmosphere. CCT development areas include basic technology, next-generation coal utilization technology, low-emissions technologies, coal gasification to produce hydrogen for fuel cell applications, and coal liquefaction.

3.8 China

In April 1997, China's State Council approved the implementation of CCT in China. In August 1998, the State Electric Corporation established a long-term plan for controlling acid rain and SO_2 emissions during the next three Five-Year Plans, covering the period from 2001 to 2015. The Chinese CCT program covers key areas spanning coal processing, coal combustion, coal conversion, pollution control, and waste treatment. Power generation technologies being investigated are circulating fluidized bed combustion (CFBC), pressurized fluidized bed combustion, and integrated coal gasification combined cycle.

4. UNITED STATES PROGRAM

Coal accounts for over 94% of the proved fossil energy reserves in the United States and supplies the bulk of the low-cost, reliable electricity vital to the nation's economy and global competitiveness. Almost 90% of the coal mined in the United States is consumed in power plants to produce electricity. In 2000, over half of the U.S. electricity was produced with coal, and projections by the U.S. Energy Information Agency (EIA) predict that coal will continue to dominate electric power production well into the first quarter of the 21st century. However, if coal is to continue to be a major energy source for the United States, it must be used in an environmentally acceptable manner. The twin marketplace drivers of the need to continue burning coal and the simultaneous need to protect the environment led to the initiation in 1986 of the CCT Program in the United States. The CCT Program was established to demonstrate the commercial feasibility of CCTs to respond to a growing demand for a new generation of advanced coal-based units, characterized by enhanced operational, economic, and environmental performance. The CCT Program represents an investment of over $5.2 billion in

advanced coal-based technology. Industry and state governments provided an unprecedented 66% of the funding. Of 38 CCT projects selected in five separate solicitations, 29 projects, valued at $3.5 billion, address electric power generation applications (18 environmental control projects and 11 advanced power projects). The other nine projects involve coal processing for clean fuels and industrial applications.

Ten guiding principles have been instrumental in the success of the U.S. CCT Program:

- Strong and stable financial commitment for the life of a project.
- Multiple solicitations spread over a number of years.
- Demonstrations conducted at commercial scale in actual user environments.
- A technical agenda established by industry, not the government.
- Clearly defined roles of government and industry.
- A requirement for at least 50% cost-sharing.
- An allowance for cost growth, but with a ceiling and cost-sharing of the increase.
- Industry retention of real and intellectual property rights.
- A requirement for industry to commit to commercialization of the technology.
- A requirement to repay the government's cost-share from commercialization profits.

Typical clean coal technology projects sponsored in the initial U.S. CCT Program are described in the following sections. These examples, all completed between 1992 and 2003, illustrate the wide range of technologies developed to permit coal to be used in a more environmentally acceptable fashion. Information on the entire CCT Program is covered extensively in the Clean Coal Compendium on the Internet.

4.1 Environmental Control Devices—SO_2 Control

Pure Air's advanced flue gas desulfurization (AFGD) process (Fig. 1) was installed on a 528-MWe (MWe = megawatt electrical, as distinct from thermal) boiler at Northern Indiana Public Service Company's Bailly Generating Station. This process removes SO_2 from flue gas by contact with a limestone slurry in a scrubber vessel. Although installed on a large power plant, the scrubber occupied a relatively small plot area, because it incorporated three functions in one vessel: prequenching of the hot flue gas, absorbing SO_2 in a limestone slurry, and oxidizing the resultant sludge to gypsum. Because of the unit's high

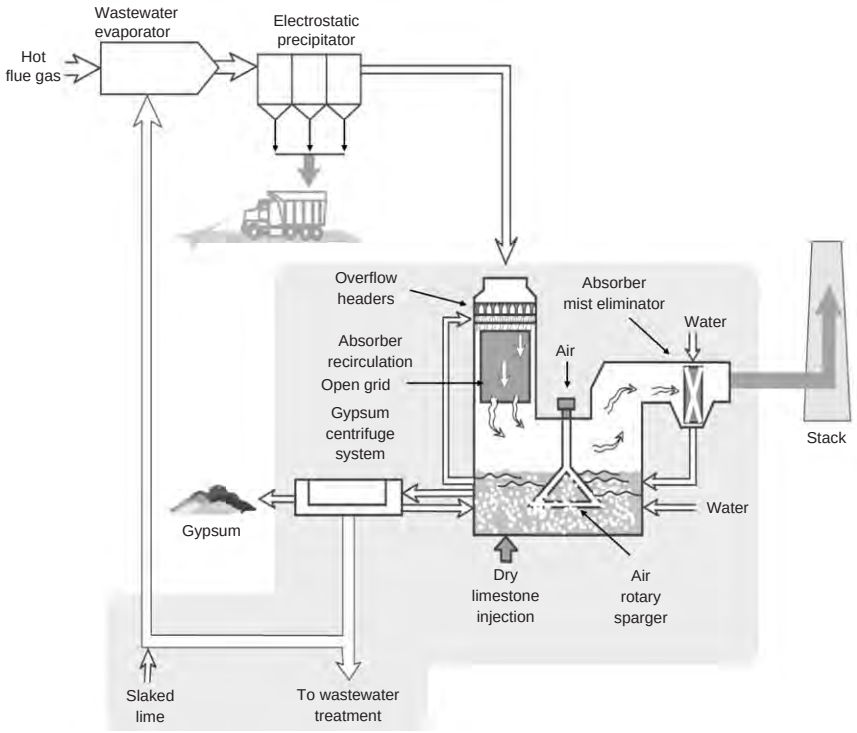


FIGURE 1 Pure Air's advanced flue gas desulfurization process.

availability of over 99%, no spare scrubber vessel was required. The AFGD unit successfully removed over 95% of the SO_2 in the flue gas and produced a wallboard-grade gypsum that was sold to a commercial producer, who used the gypsum for the manufacture of wallboard. Thus, this technology not only prevented sulfur from entering the atmosphere, but also generated a valuable by-product at the same time.

In another project, Chiyoda Corporation's Chiyoda Thoroughbred-121 (CT-121) AFGD process (Fig. 2), which uses a unique jet bubbling reactor (JBR), was installed at Georgia Power Company's Plant Yates. In this scrubber, the flue gas is injected below the surface of the limestone slurry in the JBR. Air is also injected into the slurry to oxidize the calcium sulfite initially formed into calcium sulfate. Over 90% SO_2 removal was achieved, and limestone utilization was over 97%. To prevent buildup of gypsum in the reactor, a slipstream is removed from the scrubber and sent to a collection pond. The scrubber also proved to be an effective particulate removal device.

4.2 Environmental Control Devices—NO_x Control

4.2.1 Selective Catalytic Reduction

A project at Gulf Power Company's Plant Christ consisted of testing a series of selective catalytic reduction (SCR) catalysts (Fig. 3). The test facility consisted of nine small reactors, each of which was loaded with a different SCR catalyst and fed a slipstream of flue gas from one of the plant's boilers. SCR catalyst is typically supported on the surface of ceramic honeycombs or plates, which are loaded into the reactor. Ammonia was injected into the flue gas before it passed over the catalyst. Ammonia injection needs to be carefully controlled to prevent the escape of unreacted ammonia, known as ammonia slip. The ammonia reacted with NO_x to form nitrogen gas and water. Although SCR had been used in Germany and Japan, it had not been proved to be applicable when burning higher sulfur U.S. coals. This study showed that SCR is capable of removing up to 90% of the NO_x in flue gas and that the process can be effective with U.S. coals.

4.2.2 Low-NO_x Burners

Several projects have demonstrated the effectiveness of low-NO_x burners for reducing NO_x emissions. These include demonstrations on wall-fired, tangentially fired, and cell burner-fired furnaces. Low-NO_x burners reduce NO_x by staged combustion. Figure 4 is an example of low-NO_x burners installed on a wall-fired boiler. Such burners are designed to mix the fuel initially with only part of the combustion air. This reduces NO_x formation both because the combustion temperature is reduced and because the concentration of oxygen is lower. The rest of the air, referred to as overfire air, is introduced at a higher

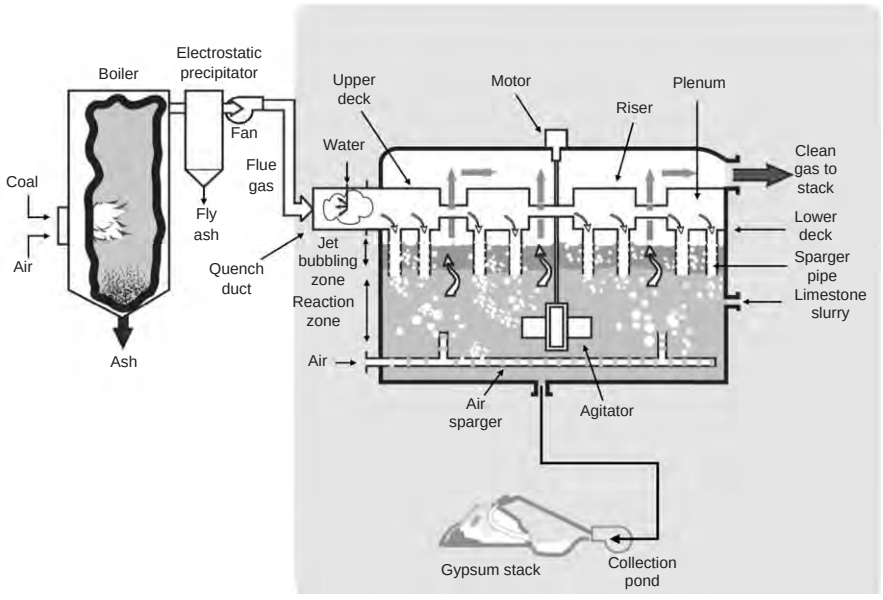


FIGURE 2 Chiyoda Corporation's Chiyoda Thoroughbred-121 advanced flue gas desulfurization process using the jet bubbling reactor.

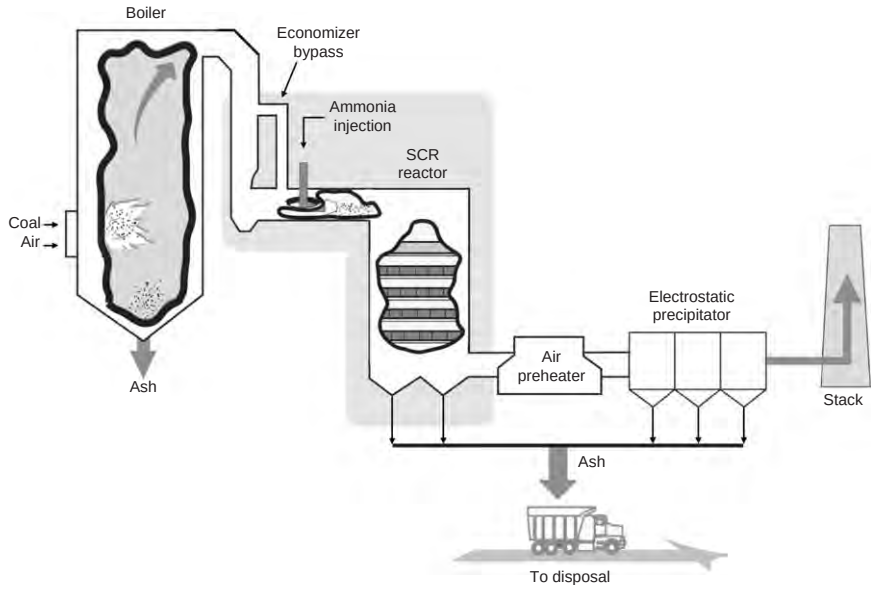


FIGURE 3 Selective catalytic reduction (SCR) installation at a power plant.

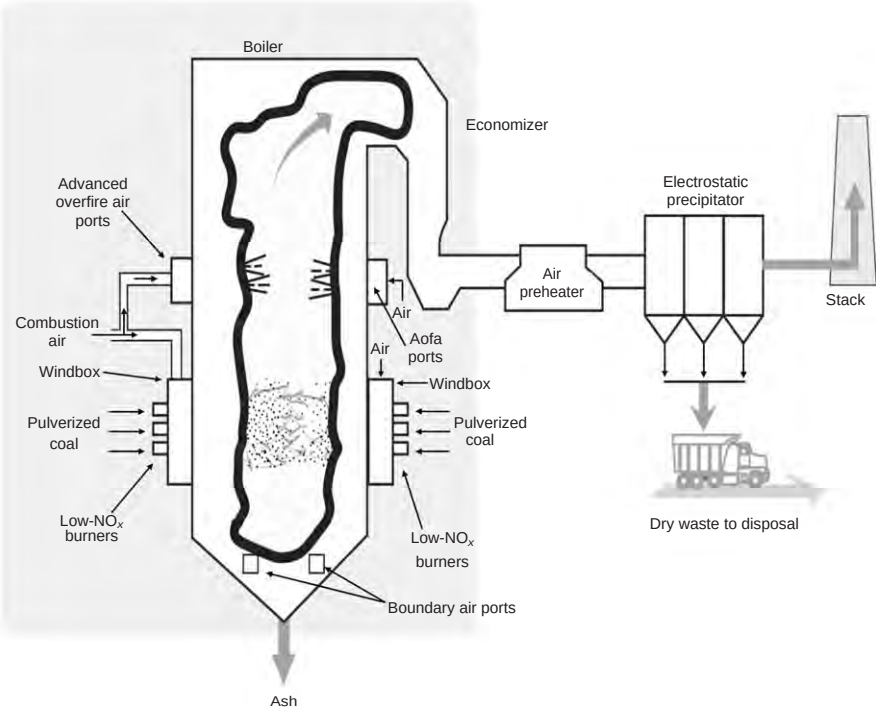


FIGURE 4 Low-NO_x burners with advanced overfire air (AOFA) installation on a wall-fired boiler.

elevation in the furnace to complete combustion. Low-NO_x burners typically reduce NO_x by 50–60%, which is generally not sufficient to meet current NO_x emissions regulations. Therefore, low-NO_x burners may need to be supplemented with SCR or another auxiliary process. Low-NO_x burners can result in increased unburned carbon in the ash, which may make the ash unsalable for cement production.

4.2.3 Reburning

Another NO_x reduction technology demonstrated in CCT projects is reburning (Fig. 5). In reburning, a

hydrocarbon fuel, such as natural gas or finely pulverized coal, is injected into the furnace above the burners to create a reduction zone, where hydrocarbon fragments react with NO_x to form nitrogen gas. Overfire air is introduced above the reburning zone to complete combustion. Reburning, which typically achieves 50–60% NO_x reduction, is not as effective as SCR.

4.3 Environmental Control Devices—Combined SO_2 and NO_x Removal

In the SNOX (a trademarked name) process (Fig. 6), demonstrated at Ohio Edison's Niles Station in Niles, Ohio, flue gas first passes though a high-efficiency fabric filter baghouse to minimize fly ash fouling of downstream equipment. Ammonia is then added to

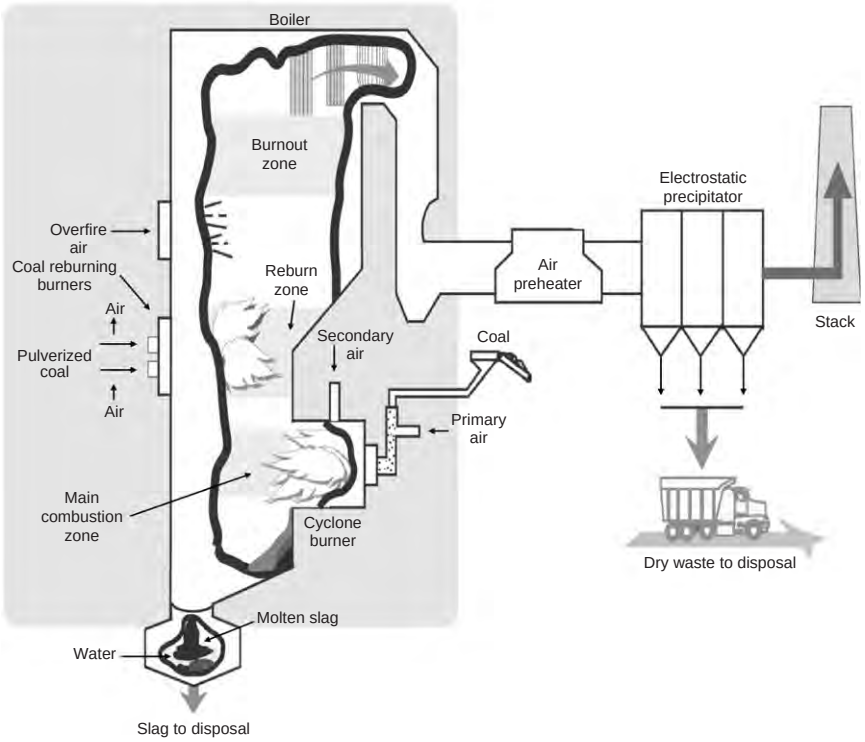


FIGURE 5 Coal reburning installation on a boiler with a cyclone burner.

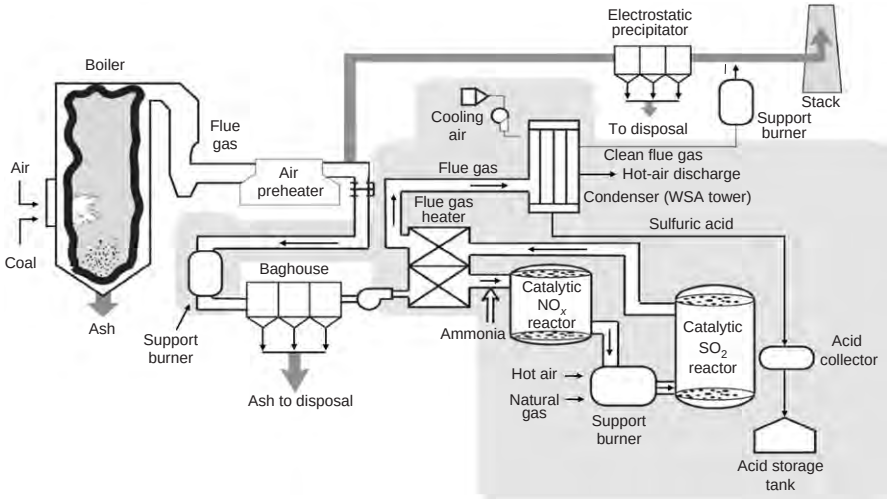


FIGURE 6 Haldor Topsoe's SNOX advanced flue gas cleanup system. WSA, Wet gas sulfuric acid.

the flue gas, which passes to a catalytic reactor where the ammonia reacts with NO_x to form nitrogen gas and water. The next step is a catalytic reactor, which oxidizes SO_2 to SO_3. The presence of this reactor oxidizes any unreacted ammonia to N_2 and water, thus preventing the formation of ammonium sulfate, which can form deposits that foul downstream equipment. The oxidation catalyst also virtually eliminates CO and hydrocarbon emissions. The gas then flows to a condenser for the recovery of sulfuric acid and then to the stack. Sulfur removal is about 95% and NO_x removal averages 94%. The sulfuric acid exceeds federal specifications for Class I acid. The fabric filter removes over 99% of the particulates.

Another demonstration combining SO_2 and NO_x removal was the SOx–NOx-Rox Box (SNRB) project at Ohio Edison's R.E. Burger Plant. The SNRB process combines the removal of SO_2, NO_x, and particulates in one unit, a high-temperature fabric filter baghouse (Fig. 7). SO_2 removal is effected by injecting a calcium- or sodium-based sorbent into the flue gas upstream of the baghouse. Ammonia is also injected into the flue gas to react with NO_x and to form N_2 and water over the SCR catalyst, which is installed inside the bags in the baghouse. Particulate removal occurs on the high-temperature fiber filter bags. Sulfur removal with calcium-based sorbents was in the 80–90% range at a Ca/S ratio of 2.0. NO_x removal of 90% was achieved.

4.4 Advanced Electric Power Generation

4.4.1 Integrated Gasification Combined Cycle

A project at PSI Energy's Wabash River Generating Station demonstrated the use of IGCC for the production of clean power from coal (Fig. 8). In IGCC, the coal is first partially combusted with air or oxygen to form a low to medium heating-value fuel gas, whose main combustible constituents are carbon monoxide and hydrogen, with some methane and heavier hydrocarbons. After being cleaned, the fuel gas is burned in a combustion turbine. The hot turbine flue gas is sent to a heat recovery steam generator (HRSG), which produces steam for a conventional steam turbine. A major advantage of IGCC is that, in the gasification process, the sulfur, nitrogen, and chlorine in the coal are converted to hydrogen sulfide, ammonia, and hydrogen chloride, respectively. It is much easier to remove these materials from the fuel gas than it is to remove SO_2 and NO_x from the flue gas of a conventional coal-fired boiler.

In the Wabash project, the gasifier used was Global Energy's E-Gas Technology. This is a two-stage gasifier, with a horizontal stage followed by a vertical stage. The ash in the coal is removed from the bottom of the gasifier as molten slag, which is quenched before being sent to disposal or sales. This has proved to be a very clean plant, with sulfur removal above 99% and NO_x levels at 0.15 lb/million Btu. Particulate levels were below detectable

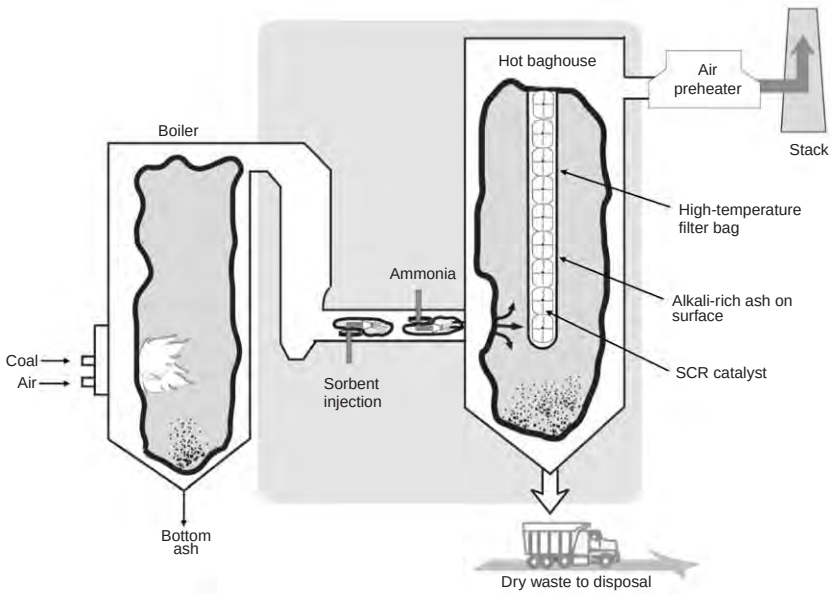


FIGURE 7 The Babcock & Wilcox Company's SOx–NOx-Rox Box process. SCR, Selective catalytic reduction.

limits. The recovered hydrogen sulfide was converted to liquid sulfur for sale.

Another IGCC demonstration was Tampa Electric Company's Polk Power Station Unit No. 1. This project used a Texaco gasifier with a radiant fuel gas cooler installed in the reactor vessel below the gasifier (Fig. 9). For this project, the recovered hydrogen sulfide was converted to sulfuric acid for sale. This plant also proved to be very clean, with SO_2, NO_x, and particulate emissions well below regulatory limits for the plant site.

4.4.2 Fluidized Bed Combustion

Foster Wheeler's atmospheric circulating fluidized bed (ACFB) combustion system (Fig. 10) was installed at the Nucla Station in Nucla, Colorado. In the combustion chamber, a stream of air fluidizes and entrains a bed of coal, coal ash, and sorbent (limestone). Relatively low combustion temperatures limit NO_x formation. Calcium in the sorbent combines with SO_2 to form calcium sulfite and sulfate. Solids exiting the combustion chamber with the flue gas are removed in a hot cyclone and recycled to the

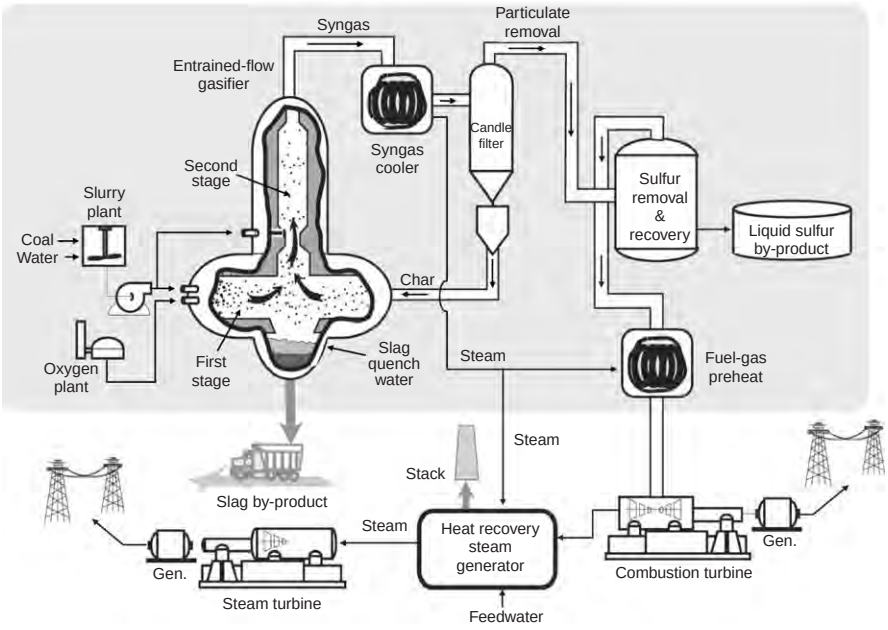


FIGURE 8 Integrated gasification combined cycle installation using Global Energy's E-Gas (syngas) technology.

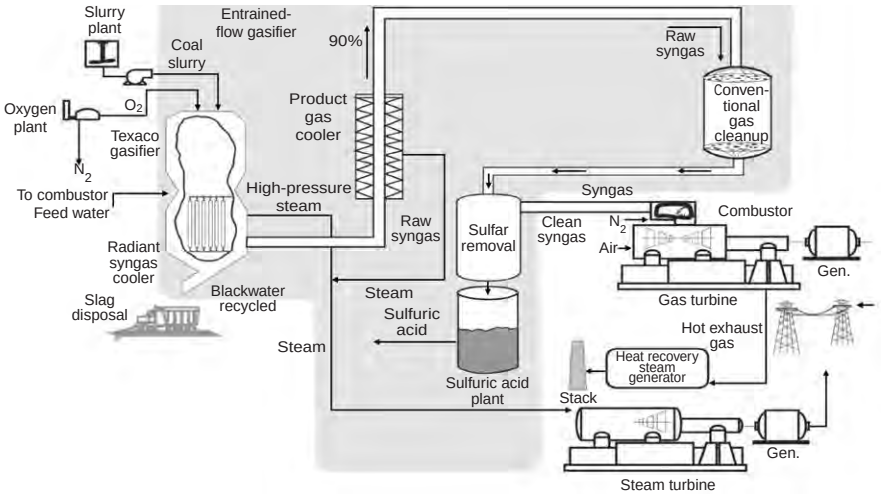


FIGURE 9 Integrated gasification combined cycle installation based on Texaco's pressurized, oxygen-blown, entrained-flow gasifier technology.

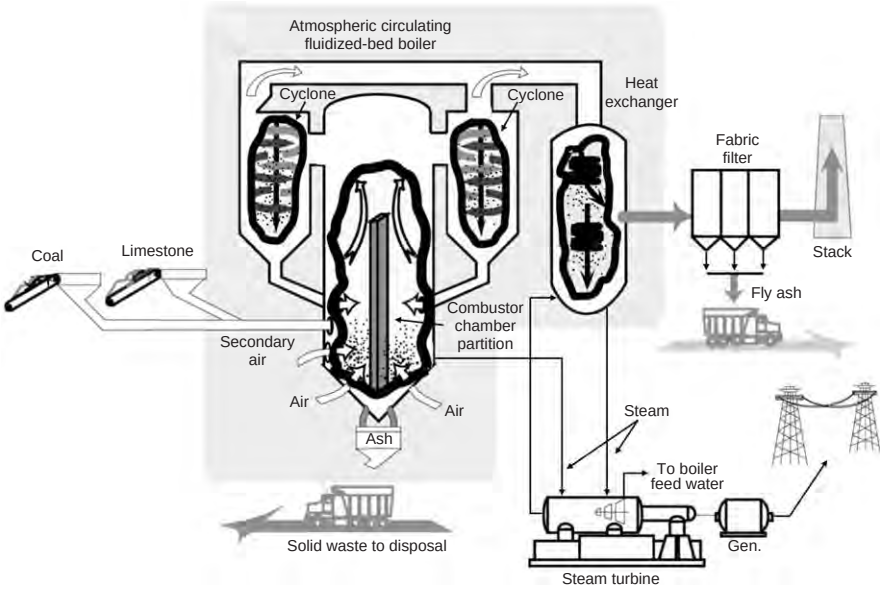


FIGURE 10 Foster Wheeler's atmospheric circulating fluidized bed combustion system.

combustion chamber. Continuous circulation of coal and sorbent improves mixing and extends the contact time of solids and gases, thus promoting high carbon utilization and high sulfur capture efficiency. Heat in the flue gas exiting the hot cyclone is recovered in the economizer. Flue gas passes through a baghouse, where particulate matter is removed. This project was the first repowering of a U.S. power plant with ACFB technology and demonstrated this technology's ability to burn a wide range of coals cleanly and efficiently.

4.5 Coal Processing for Clean Fuels

4.5.1 Methanol Production

A project at Eastman Chemical Company's Kingsport, Tennessee, facility demonstrated Air Products and Chemicals' liquid-phase methanol (LPMEOH) process (Fig. 11). This process converts synthesis gas (a mixture of CO and H_2) into methanol according to the reaction $CO + 2H_2 \rightarrow CH_3OH$. The process differs from conventional gas-phase processes in that instead of pelleted catalyst in beds or tubes, the LPMEOH process uses a finely divided catalyst suspended in a slurry oil. Use of a slurry reactor results in a uniform temperature throughout the reactor and improves heat removal by the internal heat exchanger. Thus, much higher per-pass conversions can be achieved. Furthermore, because the water–gas shift capabilities of the catalyst promote the conversion of CO to H_2, low H_2/CO ratio synthesis gas, typical of that produced by a coal gasifier, can be fed to the process. After some initial problems with catalyst poisons in the coal-derived synthesis gas were overcome, the LPMEOH process worked very well. All the methanol produced was used by Eastman Chemicals as an intermediate in the production of other chemicals.

One possibility for use of the LPMEOH process would be in conjunction with an IGCC system. In this concept, part of the synthesis gas from the gasifier in the IGCC unit would be diverted to the LPMEOH process for conversion to methanol. The amount diverted would depend on the load requirement of the power plant. This would permit the gasifier, the most expensive section of the IGCC unit, to operate at full load regardless of the power demand. The methanol produced could be sold or used as fuel in a combustion turbine during high electric power demand periods. During the demonstration project, tests showed that the LPMEOH process can rapidly ramp feed rate up and down, thus indicating its compatibility with the IGCC system concept.

4.6 Industrial Applications

4.6.1 Granulated Coal Injection into a Blast Furnace

One of the most expensive and pollution-prone aspects of blast furnace operation is the production of coke. To reduce coke usage, blast furnaces may inject natural gas or pulverized coal directly into the blast furnace. In a demonstration project at Bethlehem Steel's Burns Harbor Plant in Indiana, it was

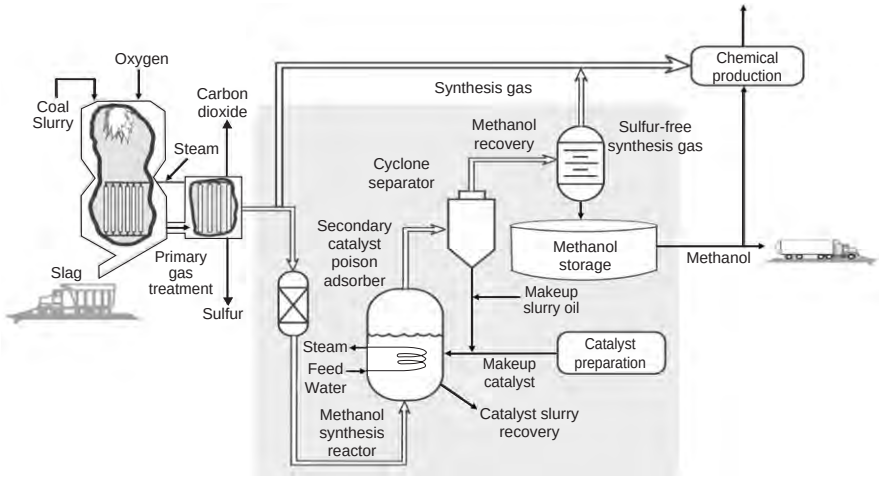


FIGURE 11 Air Products and Chemicals' liquid-phase methanol process.

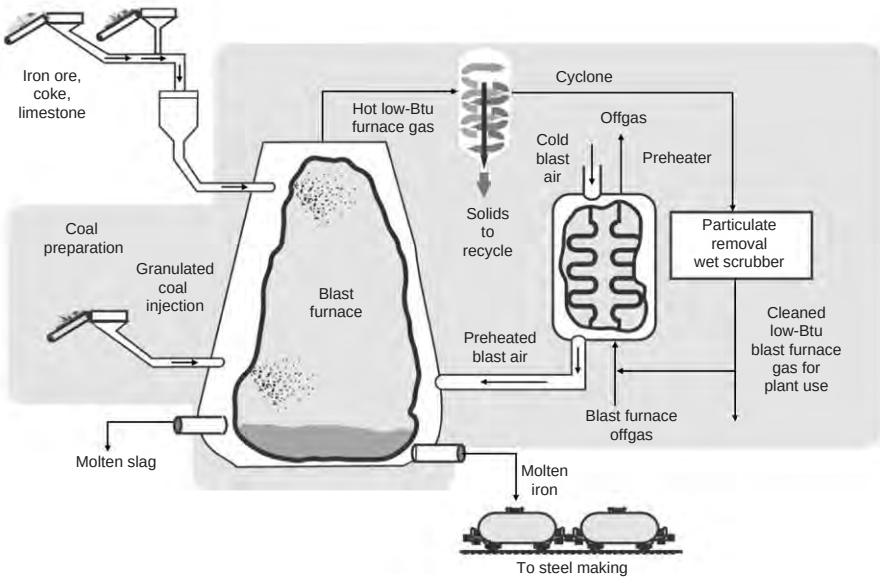


FIGURE 12 British Steel and Clyde Pneumatic blast furnace granular coal injection process.

shown that granular coal injection (Fig. 12) worked as well as pulverized coal injection and required about 60% less grinding energy. The coal replaced coke on almost a one-for-one basis. All the coke cannot be replaced because of the need to preserve porosity in the blast furnace, but enough coke can be displaced to improve economics significantly. Sulfur and ash in the injected coal end up in the slag removed from the furnace.

4.6.2 Cement Kiln Flue Gas Recovery Scrubber

A cement kiln flue gas recovery scrubber project was located at Dragon Products Company's coal-fired cement kiln in Thomaston, Maine. One of the problems at this plant was that about 10% of the kiln product was cement kiln dust (CKD), an alkali-rich waste that, because of its potassium content, cannot be blended with the cement product. Another problem was SO_2 in the kiln exhaust gas. The innovative cement kiln flue gas recovery scrubber (Fig. 13) demonstrated in this project solved both problems simultaneously by slurrying the CKD with water and using the slurry to scrub the kiln exhaust gas. The potassium in the CKD reacts with SO_2 in the kiln gas in the reaction tank to form potassium sulfate, which is water soluble. After liquid/solid separation, the low-potassium solid can be blended with the cement product, and potassium sulfate can

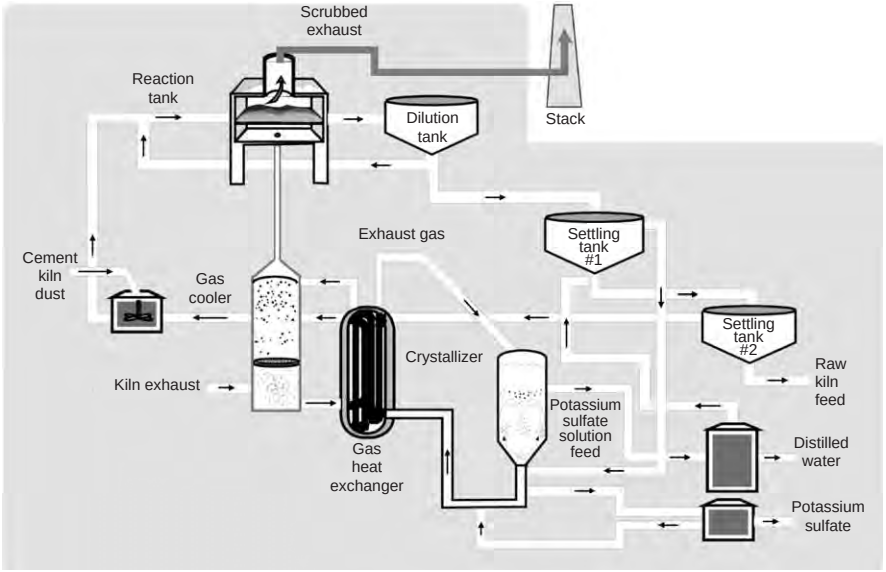


FIGURE 13 Passamaquoddy Technology recovery scrubber for treating cement kiln flue gas.

be recovered from the solution for sale as fertilizer. This project reduces SO_2 emissions by about 90%, while simultaneously reducing raw material requirements by about 10%, significantly improving economics.

4.7 Follow-On CCT Initiatives

The Clean Air Act of 1970 and subsequent amendments, especially the Acid Rain Program provisions of the 1990 amendments, and implementation of the original CCT Program have brought about major reductions in emissions of acid gases (SO_2 and NO_x) and particulates from coal-fired plants.

However, coal-fired power plants are increasingly being required to further cut emissions. Renewed concerns about fine particulates and their precursors (nitrogen and sulfur oxides), trace elements (especially mercury), and ozone (and its NO_x precursor) have created new requirements for cleaner plants.

Building on the success of the original CCT Program, the U.S. government has promoted two further solicitations, the Power Plant Improvement Initiative (PPII) in 2000 and the Clean Coal Power Initiative (CCPI) in 2002. PPII was established to promote the commercial-scale demonstration of technologies that would assure the reliability of the energy supply from existing and new electric-generating facilities. PPII is geared toward demonstrations of advanced near-term technologies to increase efficiency, lower emissions, and improve the economics and overall performance of coal-fired power plants. Most PPII projects focus on technologies enabling coal-fired power plants to meet increasingly stringent environmental regulations at the lowest possible cost. Proposed technologies must be mature enough to be commercialized within a few years, and the cost-shared demonstrations must be large enough to show that the technology is commercially viable.

The CCPI, a 10-year program, was initiated to further develop, improve, and expand the use of CCTs to increase national energy security and the reliability of the electric power supply and reduce costs while improving the environment. Candidate technologies are demonstrated at full scale to ensure proof of operation prior to widespread commercialization.

5. BENEFITS AND FUTURE OF CLEAN COAL TECHNOLOGY

The clean coal technology development effort has provided, and will continue to provide, significant economic, environmental, and health benefits. Economic benefits arise in a number of areas. The CCT Program has been instrumental in the commercialization of technologies such as AFBC and IGCC. The program has also demonstrated a variety of new options for the control of sulfur oxides, nitrogen oxides, and particulate emissions from electric power plants operating on coal. Considerable economic activity is generated through the sale, design,

construction, and operation of these new technologies. Furthermore, these new power generation and pollution control options will permit coal to continue to be used as a fuel while minimizing environmental impacts. Because coal is the cheapest fossil fuel, its continued use will save billions of dollars that can be invested in other economic activities.

CCT benefits to the environment are obvious. Decreasing SO_2 and NO_x emissions reduces acid rain, which in turn reduces acidification and eutrophication of lakes and damage to forests and other vegetation. It also reduces damage to structures made of steel, limestone, concrete, and other materials. CCTs have reduced the amount of pollutants emitted by fossil fuel-fired power plants. For example, between 1970 and 2000, emissions of sulfur and nitrogen pollutants from the average U.S. coal-fired power plant have declined by 70 and 45%, respectively. This has enabled coal use to more than double while allowing the United States to meet its clean air objectives.

Another major benefit is in the area of human health. Decreases in smog precursors are very beneficial to the health of human beings, particularly people with respiratory problems, such as asthma. Decreases in emissions of mercury and other air toxics are expected to result in fewer cancers and other diseases. These benefits should decrease medical costs by tens, if not hundreds, of billions of dollars over the next couple of decades.

Another benefit of CCT efforts has been international cooperation. Because pollution does not respect national borders, all nations must be concerned, and this has fostered international efforts. Technologies developed in the United States and elsewhere are available internationally, and global cooperation on international energy issues is ongoing. However, new environmental constraints will require the development of new clean coal technologies. At the time the United States initiated its first CCT program, there were no regulations on mercury emissions or fine particulates (those smaller than 2.5 μm in diameter). NO_x and SO_2 are precursors to the atmospheric formation of fine particulates and, therefore, will likely undergo more stringent regulation. Further regulation of these pollutants will lead to the development of new clean coal technologies.

Although not currently regulated in the United States , there is growing concern about the buildup of carbon dioxide in the atmosphere. Decreasing CO_2 emissions from coal-burning plants will require the development of improved CO_2 capture technologies, as well as the development of CO_2 sequestration techniques. Of perhaps more importance in the short term is to improve the efficiency of power plants. Decreasing the amount of fuel that has to be burned to produce a given amount of electricity directly reduces the amount of CO_2 produced. An area that will see significant development will be advanced plant control systems, such as Electric Power Research Institute's Generic NO_x Control Intelligence System (GNOCIS), demonstrated at Gulf Power Company's Plant Hammond in the United States. Such systems allow a power plant to operate at or near optimum conditions, thus improving efficiency and lowering emissions.

In the United States, it is projected that reliance on fossil fuels will increase from the present level of 85% to 90% by 2020 under current price and usage trends. The use of fossil fuels to produce electricity is also expected to rise from the current 67% to 78% by 2020. Energy consumption in the developing world (Asia, Africa, the Middle East, and Central and South America) is expected to more than double by 2020, with the highest growth rates expected in Asia and Central and South America. The next generation of coal-fired power plants is emerging to meet this growth. These systems offer the potential to be competitive with other power systems from a cost and performance standpoint, while having improved environmental performance.

The Clean Coal Power Initiative, the newest CCT program in the United States, is aimed at technologies that will focus on development of multipollutant (SO_2, NO_x, and Hg) control systems. In the past, environmental regulations have been issued incrementally—generally, one pollutant was regulated at a time. This approach has proved to be costly to the electric power industry. The new multipollutant control approach should achieve equivalent reductions in pollutants at much lower cost. The CCPI will also focus on the development of high-efficiency electric power generation, such as gasification, advanced combustion, fuel cells and turbines, retrofitting/repowering of existing plants, and development of new plant technologies.

Because biomass combustion provides a means for recycling carbon, composite fuels consisting of coal and waste biomass can offset the release of greenhouse gases during combustion. Throughout U.S. coal-producing regions are large amounts of coal residing in waste ponds. Separation processes to recover this coal economically could make it available for power generation and, concurrently, eliminate existing and potential environmental problems associated with these waste sites. This type of waste coal is particularly applicable to fluidized bed combustion.

The examples discussed here illustrate the wide range of technologies included under the clean coal technology umbrella. As regulatory and economic drivers change, an even wider range of activities will need to be included in the future.

SEE ALSO THE FOLLOWING ARTICLES

Acid Deposition and Energy Use • Air Pollution, Health Effects of • Air Pollution from Energy Production and Use • Carbon Sequestration, Terrestrial • Clean Air Markets • Coal, Chemical and Physical Properties • Greenhouse Gas Emissions from Energy Systems, Comparison and Overview • Hazardous Waste from Fossil Fuels

Further Reading

Center for Coal Utilization, Japan (CCUJ). Program and project information is accessible on the Internet (http://www.ccuj.or.jp/index-e.htm). (Accessed May 20, 2003.)

Clean Coal Engineering & Research Center of Coal Industry (CCERC), China (http://www.cct.org.cn/eng/index.htm).

Cooperative Research Centre for Coal in Sustainable Development, Clean Coal Technology in Australia: Industry and Policy Initiatives (D .J. Harris and D. G. Roberts) (http://www.nedo.go.jp/enekan/ccd2.pdf). (Accessed September 9, 2003.)

Energy Information Administration. (2003). Annual Energy Outlook 2003 with Projections to 2025, January 2003. Energy Information Administration, Washington, D.C.

Energy, Security, and Environment in Northeast Asia (ESENA) Project. Program and project information is accessible on the Internet (http://www.nautilus.org/esena/). (Accessed January 7, 2000.)

European Commission Clean Coal Technology Programme. Program and project information is accessible on the Internet (http://europa.eu.int/comm/energy/en/pfs_carnot_en.html and http://www.euro-cleancoal.net/). (Accessed July, 2000.)

IEA Coal Research Clean Coal Centre. Program and project information is accessible on the Internet (http://www.ieacoal.org.uk/). (Accessed November 26, 2003.)

New Energy and Industrial Technology Development Organization (NEDO). Program and project information related to clean coal technology is accessible on the Internet (http://www.nedo.go.jp/enekan/eng/2-1.html). (Accessed September 9, 2003.)

UK Cleaner Coal Technology Program. Program and project information is accessible on the Internet (www.dti.gov.uk/cct). (Accessed March 14, 2003.)

U.S. Department of Energy. (2002). Clean Coal Technology Demonstration Program, Program Update 2001, July 2002. U.S. Department of Energy, Washington, D.C.

U.S. Department of Energy Clean Coal Power Initiative. Program and project information is accessible on the Internet (http://www.fossil.energy.gov/programs/powersystems/cleancoal/). (Accessed December 1, 2003.)

U.S. Dept. of Energy, Clean Coal Technology Compendium. Program and project information is available on the Internet (http://www.lanl.gov/projects/cctc). (Accessed October 31, 2003.)

U.S. Department of Energy Clean Coal Technology Program. Program and project information is accessible on the Internet (http://www.fossil.energy.gov/programs/powersystems/cleancoal/86-93program.html). (Accessed November 6, 2003.)

U.S. Department of Energy Power Plant Improvement Initiative. Program and project information is accessible on the Internet (http://www.lanl.gov/projects/ppii/). (Accessed October 20, 2003.)

Climate Change and Energy, Overview

MARTIN I. HOFFERT
New York University
New York, New York, United States

KEN CALDEIRA
U.S. Department of Energy, Lawrence Livermore National Laboratory
Livermore, California, United States

1. **Introduction: Global Warming as an Energy Problem**
2. **Efficiency**
3. **Fossil Fuels and CO_2 Emission Reduction**
4. **Renewables**
5. **Energy Carriers and Off-Planet Energy Production**
6. **Fission and Fusion**
7. **Geoengineering**
8. **Conclusion**

Glossary

carbon intensity (i) Carbon intensity of economic productivity. Amount of carbon dioxide (CO_2) produced or emitted per unit economic output, as might be expressed, for example, in moles C per dollar. (ii) Carbon intensity of energy. Amount of CO_2 produced or emitted per unit energy produced or consumed, as might be expressed in moles C per joule. Care must be taken to distinguish between carbon intensity of economic productivity and carbon intensity of energy. In the case of carbon intensity of energy, care must be taken to clarify whether the reference is to primary or secondary energy. If mass units are used to measure CO_2, care must be taken to clarify whether the reference is to the mass of CO_2 or the mass of carbon in CO_2.

carbon sequestration The storage of carbon in a location or form that is isolated from the atmosphere. CO_2 can be separated from the combustion products of fossil fuels or biomass and stored in geologic formations or the ocean. CO_2 can be sequestered by removal from the atmosphere through photosynthesis or engineered processes. Carbon can be stored as CO_2, elemental carbon, carbonate minerals, or dissolved inorganic carbon. Some authors reserve the term sequestration to indicate removal from the atmosphere and use the word storage to refer to removal from industrial point sources, such as fossil-fuel-fired power plants.

climate sensitivity The change in global mean temperature that results when the climate system, or climate model, attains a new steady state in response to a change in radiative forcing, typically expressed in units of degree Celsius warming per watts per meter squared change in radiative forcing or degree Celsius warming per doubling of atmospheric CO_2 concentration. The new steady state is typically assumed to include equilibration of rapidly responding components of the climate system, such as clouds and snow cover, but not more slowly responding components, such as ice sheets and ecosystems.

economics A branch of behavioral biology dealing with the allocation of scarce resources by *Homo sapiens*, one of the many organisms found on Earth.

energy conservation A change in demand for services leading to a reduction in energy demand. The change in demand services may increase, decrease, or have no effect on the value of services provided. Typically, however, energy conservation attempts to contribute to a decline in energy intensity.

energy efficiency Useful energy output divided by energy input. The meaning of useful energy can vary with context. Energy input may mean primary or secondary energy sources, depending on context.

energy intensity Consumption of primary energy relative to total economic output (gross domestic, national, or world product), as may be expressed, for example, in joules per dollar. Changes in energy efficiency and structural and behavioral factors all contribute to changing energy intensity.

geoengineering (i) Approaches to climate stabilization involving alteration of the planetary energy balance, for example, through the scattering or reflection to

 © 2004 Elsevier Inc. All rights reserved.

space of sunlight incident on the earth. (ii) Any attempt to promote climate stabilization involving planetary-scale interference in natural systems (e.g., large-scale storage of CO_2 in the ocean interior).

primary energy Energy embodied in natural resources (e.g., coal, crude oil, sunlight, and uranium) that has not undergone any anthropogenic conversion or transformation.

radiative forcing The perturbation to the energy balance of the earth–atmosphere system following, for example, a change in the concentration of CO_2 or a change in the output of the sun. The climate system responds to the radiative forcing so as to reestablish the energy balance. A positive radiative forcing tends to warm the surface and a negative radiative forcing tends to cool the surface. It is typically measured in watts per meter squared.

secondary energy Energy converted from natural resources (e.g., coal, crude oil, sunlight, and uranium) into a form that is more easily usable for economic purposes (e.g., electricity and hydrogen). Typically, there are energy losses in the conversion from primary to secondary energy forms.

Stabilizing climate is an energy problem. Setting ourselves on a course toward climate stabilization will require a buildup within the following decades of new primary energy sources that do not emit carbon dioxide (CO_2) to the atmosphere in addition to efforts to reduce end-use energy demand. Mid-century CO_2 emissions-free primary power requirements could be several times what we now derive from fossil fuels ($\sim 10^{13}$ W), even with improvements in energy efficiency. In this article, possible future energy sources are evaluated for both their capability to supply the massive amounts of carbon emissions-free energy required and their potential for large-scale commercialization. Possible candidates for primary energy sources include terrestrial solar, wind, solar power satellites, biomass, nuclear fission, nuclear fusion, fission–fusion hybrids, and fossil fuels from which carbon has been sequestered. Non-primary-power technologies that could contribute to climate stabilization include conservation, efficiency improvements, hydrogen production, storage and transport, superconducting global electric grids, and geoengineering. For all, we find severe deficiencies that currently preclude cost-effective scale-up adequate to stabilize climate. A broad range of intensive research and development is urgently needed to develop technological options that can allow both climate stabilization and economic development.

1. INTRODUCTION: GLOBAL WARMING AS AN ENERGY PROBLEM

1.1 The Carbon Dioxide Challenge

More than 100 years ago, Svante Arrhenius suggested that CO_2 emissions from fossil fuel burning could raise the infrared opacity of the atmosphere enough to warm Earth's surface. During the 20th century, the human population quadrupled and humankind's energy consumption rate increased 16-fold, mainly from increased fossil fuel combustion. In recent years, observational confirmations of global warming from the fossil fuel greenhouse have accumulated, and we have come to better understand the interactions of fossil fuel burning, the carbon cycle, and climate change.

Atmospheric CO_2 concentration has been increasing during the past few centuries mainly from fossil fuel emissions (Fig. 1A). The buildup of atmospheric CO_2 so far has tracked energy use closely. Carbon cycle models recovering the observed concentration history (1900–2000) have most of the emitted CO_2 dissolved in the oceans (primarily as bicarbonate ions, the largest sink). As long as carbon emissions keep rising, approximately half the fossil fuel CO_2 will likely remain in the atmosphere, with the rest dissolving in the seas. Forests produce a two-way (photosynthesis–respiration) flux with residuals that temporarily add or subtract CO_2 but cancel in the long term unless the "standing crop" of biosphere carbon changes.

Stabilizing CO_2 concentration in the atmosphere at any specified level will require progressively more stringent emission reductions. However, even a total cessation of emissions would not restore preindustrial levels for more than millennia. The technology challenge of the century may be to reduce fossil fuel CO_2 emissions enough to prevent unacceptable climate change while population, energy use, and the world economy continue to grow.

1.2 Climate

A doubling of atmospheric CO_2 is estimated to produce 1.5–4.5°C surface warming based on paleoclimate data and models with uncertainties mainly from cloud–climate feedback. Unabated fossil fuel burning could lead to between two and three times this warming.

Climate change from the fossil fuel greenhouse is driven by radiative forcing—the change in solar

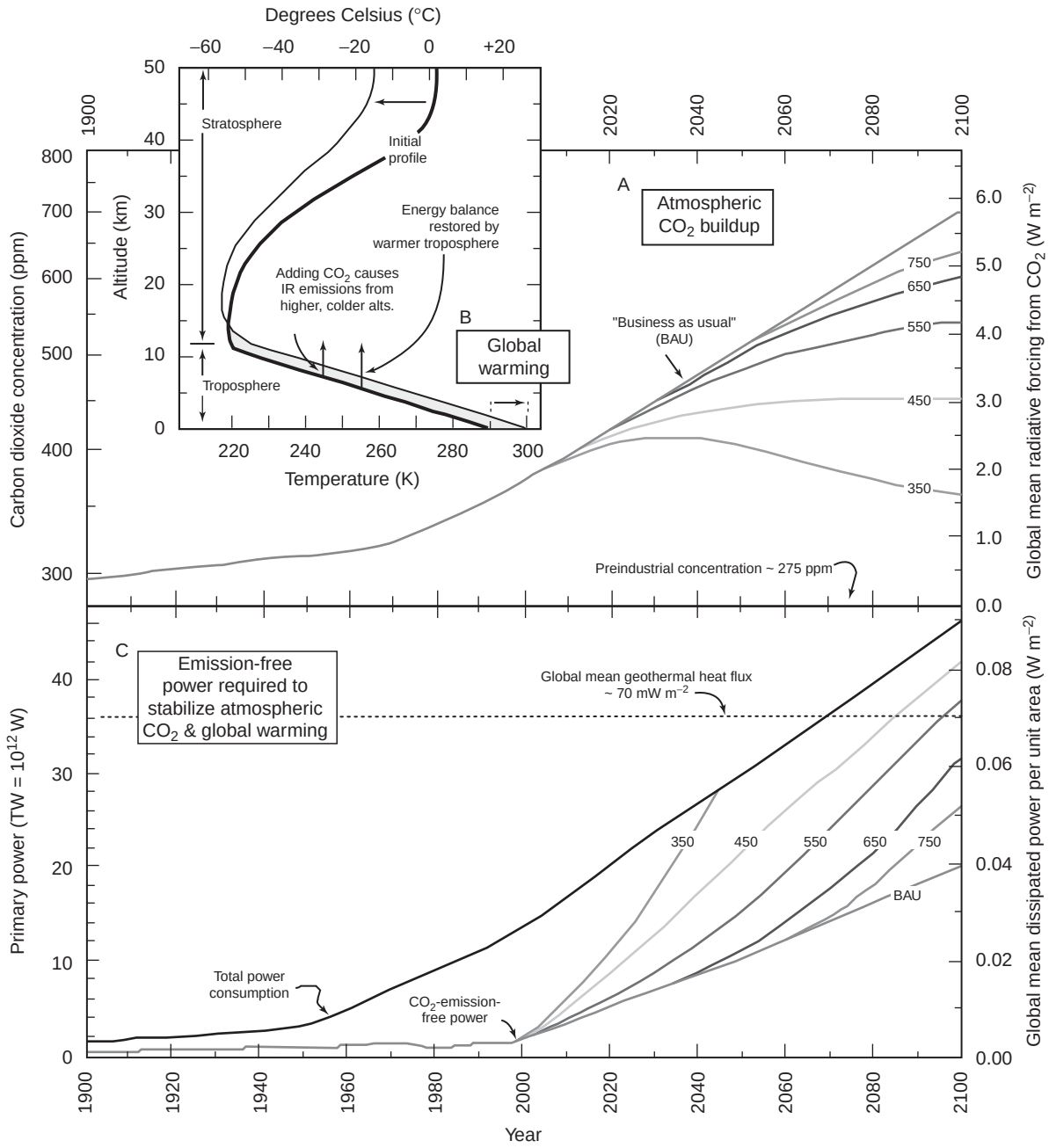


FIGURE 1 The trapping of outgoing long-wave radiation by CO_2 may mean that we need to produce vast amounts of power without releasing CO_2 to the atmosphere. (A) Atmospheric CO_2 concentration (left-hand scale) and radiative forcing (right-hand scale). The increase in CO_2 thus far is mainly from fossil fuel emissions. Future concentrations (2000–2100) are for a business as usual (BAU) emission scenario and the Wigley–Richels–Edmonds scenarios that stabilize atmospheric CO_2 (eventually) at concentrations in the range of 350–750 parts per million (ppm). (B) The lower atmosphere is rendered more opaque by IR-trapping greenhouse gases. Maintenance of the planetary energy balance requires that the earth's CO_2 radiate to space at earth's effective temperature from higher in the atmosphere. This results in surface warming. (C) Total and CO_2 emission-free power. Curves prior to 2000 are historical; curves from 2000 onward are BAU and CO_2-stabilization projections from a carbon cycle/energy model.

heating less infrared cooling to space per unit surface area relative to a preindustrial reference (Fig. 1B). Global warming by the "greenhouse effect" requires that (i) the lower atmosphere is rendered more opaque by infrared-trapping greenhouse gases, and (ii) the troposphere maintains (by convective

adjustments) a nearly constant vertical temperature lapse rate ≈ 6 K/km.

The energy balance for Earth implies a subfreezing temperature for blackbody cooling to space, $T_{\text{eff}} \approx 255$ K (-18°C). These infrared photons do not emanate, on average, from the surface but from a midtroposphere altitude of approximately 5.5 km. Natural greenhouse gases (H_2O, CO_2, and O_3) maintained a global surface temperature hospitable to life over Earth history that was 33 K warmer than the T_{eff} today: $T_{\text{surface}} \approx 255\,\text{K} + (5.5\,\text{km} \times 6\,\text{K km}^{-1}) \approx 288$ K (15°C). More CO_2 in the atmosphere from fossil fuel burning makes the troposphere radiate from even higher, colder altitudes, reducing infrared cooling. Solar heating, now partly unbalanced by infrared cooling, generates greenhouse radiative forcing [$\approx 4\,\text{W m}^{-2}$ for doubling CO_2 to 550 parts per million (ppm)], driving the troposphere toward warmer temperatures, which depend on the climate sensitivity[= (temperature change)/(radiative forcing)].

Most of the infrared flux comes from the troposphere, but the center of the 15-μm CO_2 vibration band is so strongly absorbing that the atmosphere remains opaque near these wavelengths into the stratosphere. Here, temperature increases with altitude, and more CO_2 produces cooling.

Surface warming accompanied by stratospheric cooling from CO_2 increases was predicted in the 1960s and subsequently observed (ozone depletion combined with tropospheric soot might produce similar effects). Also predicted and observed is heat flow into the sea. Instrumental and paleotemperature observations are generally consistent with an emergent CO_2 greenhouse modulated by solar variability and by volcanic and anthropogenic aerosols. A marked global warming increase is predicted to occur in the following decades, the magnitude of which will depend on climate sensitivity and on our ability to limit greenhouse gas emissions as the global population and economy grow.

It is not known what concentrations will trigger the "dangerous anthropogenic interference with the climate system" that the United Nations' 1992 Rio de Janeiro Framework Convention on Climate Change aimed to prevent. A common reference case is doubling preindustrial CO_2, projected to produce global warming of 1.5–4.5 K, which is comparable in magnitude, but opposite in sign, to the global cooling of the last ice age 20,000 years ago. Evidently, dramatic increases in primary power from non-CO_2-emitting sources, along with improvements in the efficiency of primary energy conversion to end uses, could be needed to stabilize CO_2 at "only" twice preindustrial levels (Fig. 1C). This article assesses the ability of a broad range of advanced technologies to achieve CO_2 and climate stabilization targets.

1.3 Energy

Carbon dioxide is not a pollutant in the sense that it emerges from trace elements peripheral to the main business of energy. It is a product of combustion vital to how civilization is powered. The way to reduce CO_2 emissions significantly without seriously disrupting civilization is to identify and implement revolutionary changes in the way energy is produced, distributed, and lost to inefficiencies.

Factors influencing CO_2 emissions are evident in the Kaya identity, which illustrates the quandary of emission reductions with continued economic growth. The Kaya identity expresses the carbon emission rate, $\dot{C}$, as the product of population, N, per capita gross domestic product, GDP/N, primary energy intensity, E/GDP, and the carbon emission factor, C/E:

$$\dot{C} = N \times (\text{GDP}/N) \times (E/\text{GDP}) \times (\text{C}/E).$$

Few governments intervene in their nation's population growth, whereas continued economic growth is policy in virtually every country on Earth; thus, the first two terms of the equation are normally off limits to climate change policy. Their product is the growth rate of GDP. No nation advocates reducing its GDP to reduce carbon emissions. Stabilizing CO_2 with 2 or 3% per year global GDP growth requires comparable or greater reductions in the last two terms of the equation: the energy intensity and the carbon emission factor.

Figure 1C shows primary power required to run the global economy projected out 100 years for continued economic growth assuming a "business as usual" economic scenario. Figure 1C also shows the power needed from CO_2 emission-free sources to stabilize atmospheric CO_2 at a range of concentrations. Curves prior to 2000 are historical; curves from 2000 onward are business as usual and CO_2-stabilization projections run through a carbon cycle/energy model. Note the massive transition of the global energy system implied by all scenarios. Even business as usual (incorporating no policy incentives to limit emissions) ramps to 10 TW emission free by 2050. Emission-free primary power required by midcentury increased to 15 TW to stabilize at 550 ppm and to ≥ 30 TW to recover recent concentrations of 350 ppm (including energy to remove

CO_2 previously emitted). All scenarios assume aggressive continuing energy efficiency improvements and lifestyle changes such that global energy intensity (*E*/GDP) declines 1.0% per year during the entire 21st century. The main concern is CO_2, not heat dissipated by energy use. Projected to pass mean geothermal heating this century, heat dissipated by humankind's primary power consumption is approximately 1/100 the radiative forcing from the fossil fuel greenhouse.

Technologies capable of producing 100–300% of current total primary power levels, but CO_2 emission free, could be needed by midcentury (Fig. 1C). These do not exist operationally or at the pilot plant stage today. One need only consider that Enrico Fermi's "atomic pile" is more distant in time than the Year 2050, whereas nuclear power today represents less than 5% of primary power (17% of electricity) globally.

Can we produce enough emission-free power in time? We assess a broad range of advanced technologies for this job, along with pathways to their commercialization. Our goal is to explore possibilities that can revolutionize the global energy system. As Arthur C. Clarke stated, "The only way to discover the limits of the possible is to venture a little past them into the impossible." However, we cannot pick winners and losers, and we make no claim that our proposals are exhaustive. What is clear is that the future of our civilization depends on a great deal of research investment in energy options such as those considered here.

2. EFFICIENCY

2.1 Energy Efficiency and Energy Intensity

Reducing CO_2 emissions is sometimes considered synonymous with improvements in "efficiency." Energy efficiency is the ratio of useful energy output to energy input. Progress has been made with regard to improving energy efficiencies in building, manufacturing, and transportation end-use sectors of the United States and other nations. Technologies include compact fluorescents, low-emissivity window coatings, combined heat and power systems (cogeneration), and hybrid vehicles (internal combustion plus electric power). Improvements in energy efficiency may be adopted on purely economic grounds—the so-called "no regrets" strategy. However, cost-effectiveness is more important to market share than efficiency as such; for example, amorphous photovoltaic solar cells are less efficient, but cheaper per unit of electrical energy output, than crystalline cells.

Energy intensity (*E*/GDP, energy used per unit GDP generated) subsumes efficiency, lifestyle, and structural economic factors. Typically, the energy intensity of developing nations starts low, peaks during industrialization (when energy goes to infrastructure building rather than GDP), and then declines as service sectors dominate postindustrial economies. This final stage, modeled as a sustained percent decrease, can predict large emission reductions over century timescales. Current U.S. policy aims to continue the domestic trend of declining carbon intensity [C/GDP = (*E*/GDP) × (C/*E*)]. However, even wealthy nations such as the United States that are on track to service economies consume goods manufactured elsewhere. A decline in global energy intensity adequate to stabilize the climate may be unrealistic given the massive industrialization, driven by globalization of capital and markets, in countries such as China, India, and Indonesia.

The stock of primary energy includes fossil fuels (coal, oil, and gas—a fraction of the organic carbon produced by plants from sunlight hundreds of millions of year ago), fission fuels (uranium and thorium—elements with metastable nuclei blown out of distant supernovae before the solar system formed in limited supply as high-grade ore deposits), and fusion fuels (cosmically abundant deuterium [D] and helium-three [^{3}He] relics of the big bang—D in seawater and ^{3}He in the lunar regolith and outer planet atmospheres). This energy is stored in chemical and nuclear bonds. "Renewables" are primary energy in natural fluxes (solar photons; wind, water, and heat flows). Converting this energy to human end uses always involves losses to heat dissipation—losses that engineers have in many cases already spent enormous effort reducing (Fig. 2). The efficiency of internal combustion engines, for example, is 18–23% after more than 100 years of development (Fig. 3).

Energy intensity reductions are often treated as a surrogate for energy efficiency improvements. With the "evolving economy" factor held constant, a decline of a few percent per year in *E*/GDP implies very large energy efficiency increases over long times: a factor of $(1.01)^{100} = 2.7$ at 1% per year and $(1.02)^{100} = 7.2$ at 2% per year after 100 years. Such increases will be physically impossible if they exceed thermodynamic limits.

Efficiency improvements and alternate energy sources need to be pursued in combination. It is a false dichotomy to pit efficiency against alternate

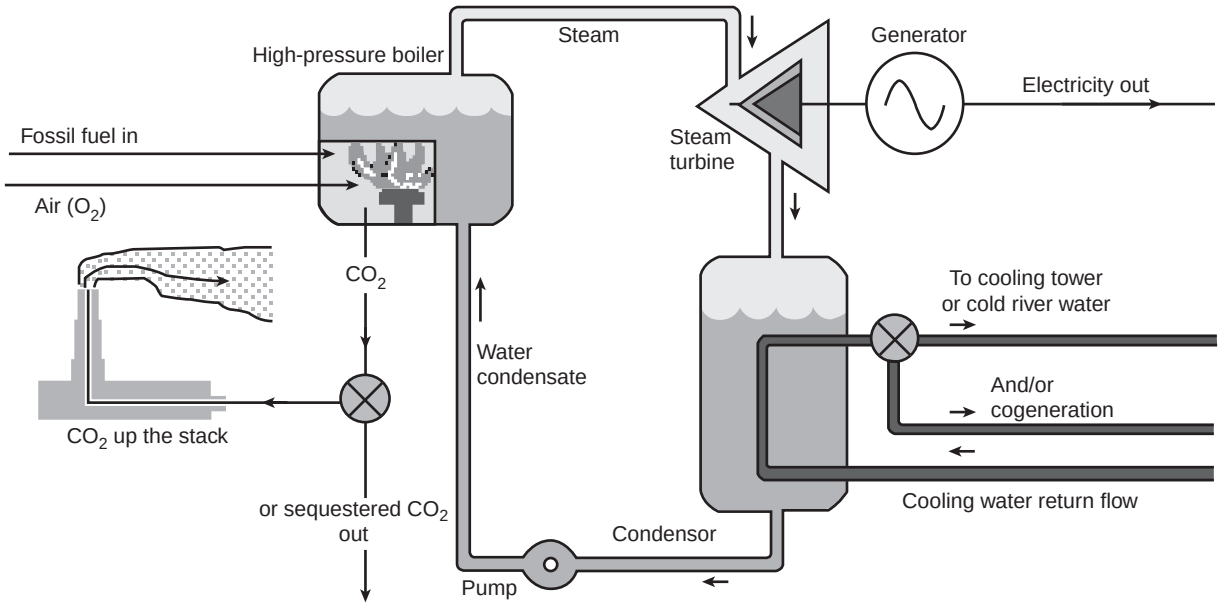


FIGURE 2 The world's electricity is generated mainly by heat engines employing hot water as the working fluid. Actual thermal efficiencies for power plant steam cycles are 39–50%, with primary energy-to-electricity efficiencies of 31–40%, roughly consistent with the conversion factor 3 kW (primary) $\approx$ 1 kW_e (electrical). Because small temperature increases cause exponentially large increases in pressure (and hoop stress) in boilers, it is unlikely that turbine inlet temperature and turbine efficiency will increase significantly (and CO_2 emissions decline) without breakthroughs in materials strong enough to prevent explosions.

energy. The trade-off in which more of one requires less of the other to achieve a given CO_2 stabilization target was analyzed by Hoffert and colleagues in 1998. We have assumed a century-long 1% per year decrease in global *E*/GDP to estimate human primary power needed in Fig. 1C. Some believe even greater rates of decline are possible, perhaps as much as 2% per year. Our analysis indicates that even 1% per year will be difficult. For example, an effective strategy in the transportation sector would be to increase energy efficiency by a path leading to CO_2 emission-free hydrogen from renewable sources as the energy carrier.

2.2 The Transportation Sector

A frequently cited opportunity for emission reductions involves improving the efficiency of the transportation sector, which is responsible for 25% of CO_2 emissions in the United States, of which approximately half is by private cars. Figure 3 compares the efficiency of fossil-fueled cars of similar mass, aerodynamics, and driving cycles for four different propulsion systems: internal combustion (IC), battery–electric, IC–electric hybrid, and fuel cell–electric. A twofold improvement seems reasonable for a transition from the IC engine to fuel cell–electric. Analyses of transportation sector efficiency should consider the entire fuel cycle, "wells to wheels." Smaller cars requiring less fuel per passenger-kilometer, mass transit, and improvements in manufacturing efficiency can also reduce CO_2 emissions. According to some estimates, the manufacturing of a car can create as much emissions as driving it for 10 years.

In the United States, sport utility vehicles are driving emissions higher, whereas economic growth of China and India has led 2 billion people to the cusp of transition from bicycles to personal cars. (Asia already accounts for $>$80% of the growth in petroleum consumption.) CO_2 emissions from transportation depend on fuel carbon-to-energy ratios (*C*/*E*), fuel efficiency, vehicle mass, aerodynamic drag, and patterns of use. Storability and high energy density of liquid hydrocarbons have fostered crude oil derivatives as dominant motor and aviation fuels and massive capital investment worldwide in oil exploration, oil wells, supertankers, pipelines, refineries, and filling stations. Gasoline-powered IC technology today embodies a sophisticated and difficult to displace infrastructure. CO_2 emission-neutral fuels derived from biomass (e.g., ethanol) are limited in availability by low-yield photosynthesis and competition with agriculture and biodiversity for land. Natural gas

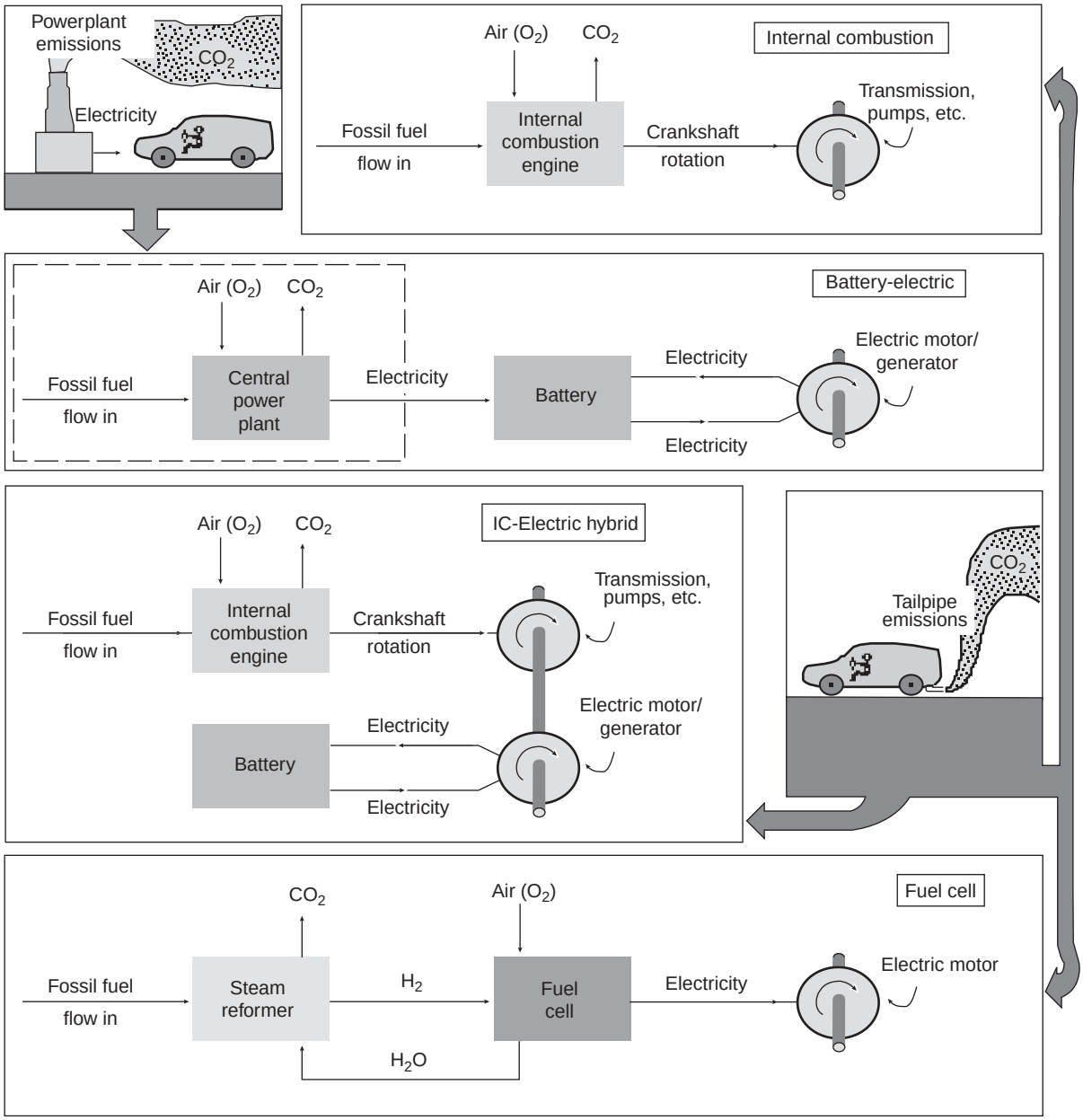


FIGURE 3 The transportation sector presents unique challenges for CO_2 emission reductions. CO_2 emissions from transportation depend on fuel carbon-to-energy ratios (C/E), fuel efficiency, vehicle mass, aerodynamic drag, and patterns of use. Fuel efficiencies [η = (mean torque × angular velocity)/(fossil fuel power in)] of the systems shown are as follows: internal combustion (IC), $\eta = 18$–23%; battery–electric, $\eta = 21$–27% (35–40% central power plant, 80–85% charge–discharge cycles, and 80–85% motor); IC–electric hybrid, $\eta = 30$–35% (higher efficiency from electric power recovery of otherwise lost mechanical energy); and fuel cell–electric, $\eta = 30$–37% (75–80% reformer, 50–55% fuel cell, and 80–85% motor). Cost-effective efficiencies of 40% may be possible with additional emission reductions per passenger-kilometer from lighter cars.

(methane, CH_4) is a possible transitional fuel, lower in CO_2 emissions than gasoline, but it requires high-pressure onboard storage (accompanied by cooling if stored as liquid natural gas).

Current batteries are prohibitively heavy for normal driving patterns unless combined with IC engines as hybrids. Fuel cells have high power densities; utilize various fuels, electrolytes, and electrode materials; and operate at various temperatures. The clear favorite for automobiles is the proton exchange membrane (PEM) cell—a compact unit converting H_2 to electricity at a relatively cool

350 K. Hydrogen (plus oxygen from the atmosphere with water as the reaction product) emits zero CO_2 from IC engines and fuel cells. A seamless transition to fuel cells is fostered by extraction of hydrogen from gasoline in a reformer, which is a miniature onboard chemical processing plant that does emit CO_2.

Fuel efficiencies [η = (mean torque × angular velocity)/(fossil fuel power in)] of the systems shown are as follows: IC, $\eta = 18$–23%; battery–electric, $\eta =$ 21–27% (35–40% central power plant, 80–85% charge–discharge cycles, and 80–85% motor); IC–electric hybrid, $\eta = 30$–35% (higher efficiency from electric power recovery of otherwise lost mechanical energy); and fuel cell–electric, $\eta = 30$–37% (75–80% reformer, 50–55% fuel cell, and 80–85% motor). Cost-effective efficiencies of 40% may be possible with additional emission reductions per passenger-kilometer from lighter cars.

A scenario aimed at sustainable CO_2 emission-free vehicles would first replace IC engines with gasoline-powered fuel cells. In the second phase, a network of filling stations dispensing H_2 derived from fossil fuel with CO_2 sequestration is deployed as onboard hydrogen storage (metal hydride tanks, carbon nanotubes, etc.) becomes safe and cost-effective. In the third and presumably sustainable phase, H_2 is generated by water electrolysis from renewable or nuclear primary sources (Figs. 4–6). This scenario, and others emphasizing mass transit, requires technology breakthroughs from basic and targeted research. Also critical for CO_2 and climate stabilization are sequencing and diffusion of transportation sector transitions worldwide.

Fossil fuel	Energy content [TW-yr]	Carbon content [GtC]	(E_{fuel}/C) [TW-yr/GtC]	(E/C) [TW-yr/GtC]	Sequestration rate [GtC/yr]
Gas	1200	570	2.1	1.9 – 1.6	5 – 6
Oil	1200	750	1.6	1.4 – 1.2	7 – 8
Coal	4800	3690	1.3	1.2 – 1.0	9 – 10

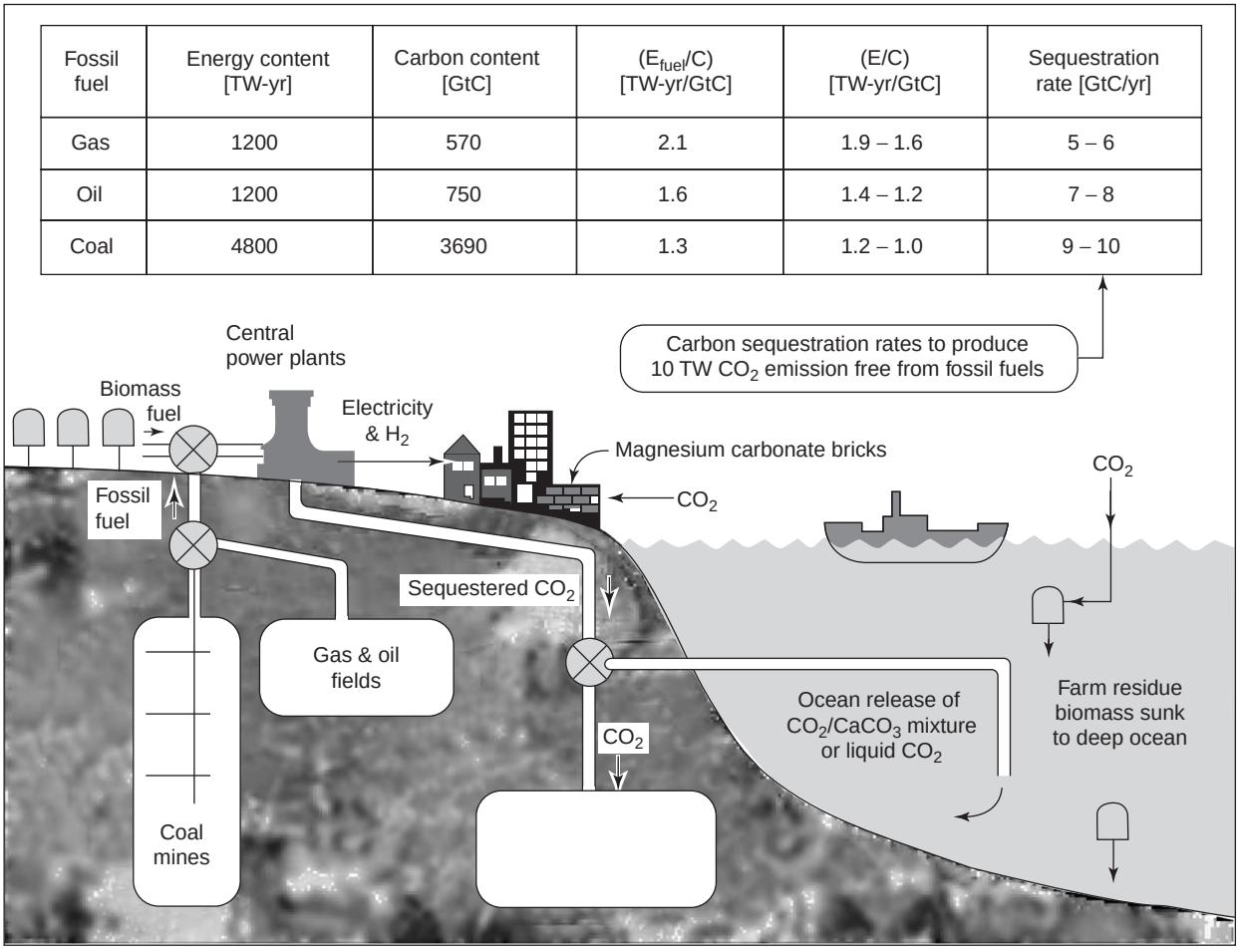


FIGURE 4 Separating and sequestering the CO_2 or solid carbon by-product of fossil fuel or biomass oxidation in subsurface reservoirs while producing electricity in central power plants and/or hydrogen may have the potential to reduce anthropogenic CO_2 emissions and/or atmospheric concentrations. More CO_2 must be sequestered (and more fossil fuel consumed) per unit energy available for human power consumption than would be emitted as CO_2 to the atmosphere without sequestration, some of which may return later to the atmosphere from leaky reservoirs.

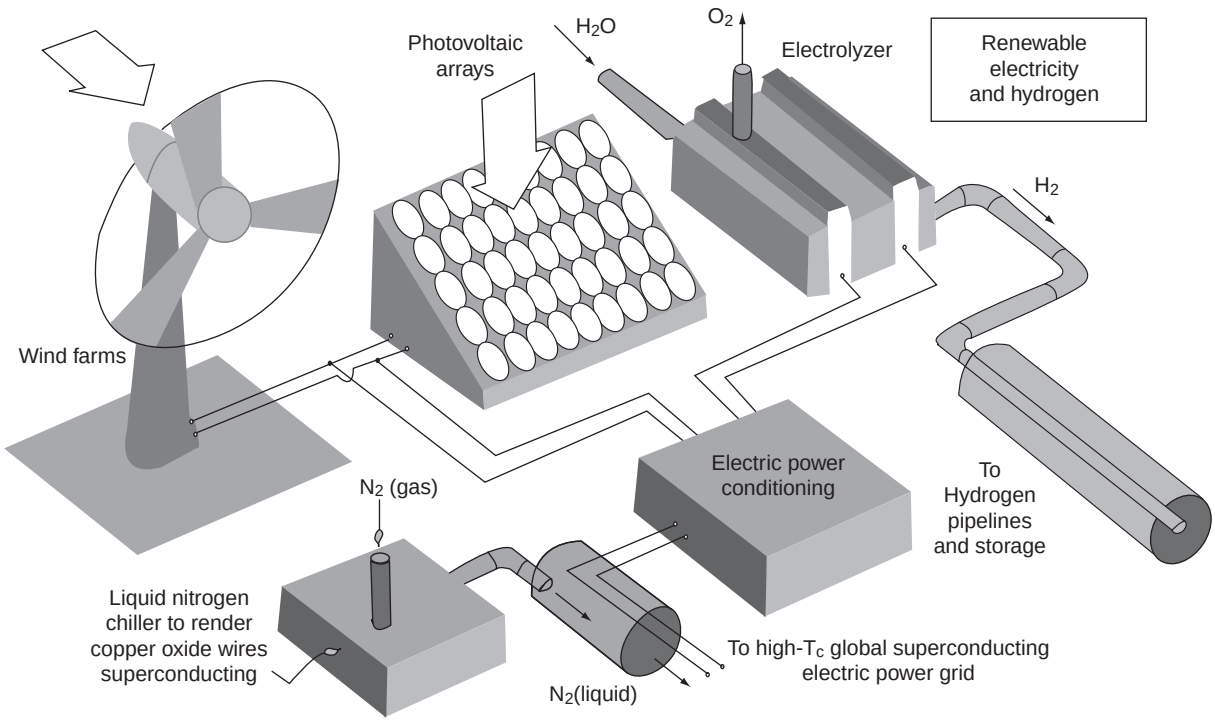


FIGURE 5 Renewable energy at Earth's surface. Greatest potential is from solar and wind with electric currents or H_2 as energy carriers.

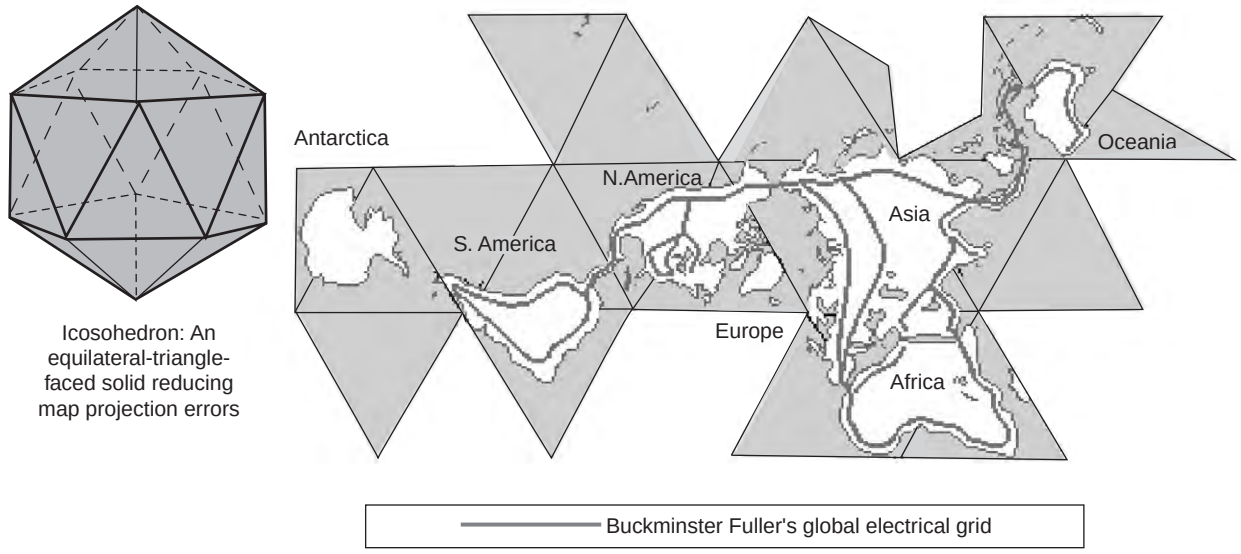


FIGURE 6 An alternative energy carrier is the low I^2R loss global electric grid envisioned by R. Buckminster Fuller. Modern computer technology may permit grids smart enough to match supply from intermittent renewable sources to baseload demands worldwide.

3. FOSSIL FUELS AND CO_2 EMISSION REDUCTION

Reductions in carbon intensity, C/*E*, the carbon emitted per unit of energy generated, reflect the degree to which societies decarbonize their energy sources. The long-term trend has been a shift from coal to oil to natural gas—hydrocarbons with decreasing C/H ratios emitting progressively less CO_2 per joule. However, the increasing use of clean

low-carbon fuels is not sustainable without somehow disposing of excess carbon because it opposes the trend in the abundance of fossil fuels, with coal resources being the most abundant followed by oil and gas (Fig. 4).

Energy in recoverable fossil fuels ("reserves" plus "resources" from conventional and unconventional sources) is ~4800 TW-year for coal and ~1200 TW-year each for oil and gas (1 Gtoe = 44.8 EJ = 1.42 TW-year; natural gas includes liquids but excludes methane hydrates). Total energy in recoverable fossil fuels is in the range 6400–8000 TW-year, of which two-thirds is coal. Coal resources could roughly double if energy in shales could be extracted cost-effectively.

Separating and sequestering the CO_2 or solid carbon by-product of fossil fuel or biomass oxidation in subsurface reservoirs while producing electricity in central power plants and/or hydrogen may have large potential for CO_2 emission reductions. Industrial hydrogen is made primarily by partial (only the carbon is oxidized) oxidation of methane (net: $CH_4 + O_2 \rightarrow CO_2 + 2H_2$). Some energy loss is unavoidable for H_2 produced from fossil fuels even without CO_2 sequestration. Chemical energy transferred to hydrogen is typically 72% from natural gas, 76% from crude oil, and 55–60% from coal.

Primary energy available after sequestration ($E = E_{fuel} - E_{seq}$) per unit carbon content depends on the energy-to-carbon ratio of the fuel (E_{fuel}/C) and the ratio of sequestration energy to fuel energy (E_{seq}/E_{fuel}): $(E/C) = (E_{fuel}/C)[1-(E_{seq}/E_{fuel})]$. Carbon sequestration rates tabulated in Fig. 4 needed to produce 10 emission-free terawatts by midcentury from gas, oil, or coal $[(10\,TW)/(E/C)]$ are in the 5–10 GtC/year range, depending on the fossil fuel mix and assuming $(E_{seq}/E_{fuel}) = 10$–25%—an extrapolation of current technology judged necessary for cost-effectiveness. This would be a massive carbon transfer akin to the global extraction of fossil fuels in reverse. More CO_2 must be sequestered (and more fossil fuel consumed) per unit energy available for human power consumption than would be emitted as CO_2 to the atmosphere without sequestration, some of which may return later to the atmosphere from leaky reservoirs.

Potential carbon sequestration reservoirs include oceans, trees, soils, depleted natural gas and oil fields, deep saline reservoirs, coal seams, and solid mineral carbonates (Fig. 4). The advantage of sequestration is that it is compatible with existing fossil fuel infrastructure, including CO_2 injections for enhanced recovery from existing oil and gas fields and capture of CO_2 from power plant flue gas. Challenges include cost and energy penalties of capture and separation (more fuel is needed per joule than freely venting combustion products), understanding the long-term fate of CO_2 sequestered in various reservoirs, and investigating environmental impacts.

The simplest sequestration technology is growing trees. CO_2 uptake only works during net biomass growth. At best, it could mitigate 10–15% of emissions over a limited time. Approximately 2500 Gt C is organically fixed in land plants and soil. Carbon cycle models indicate that absorption by trees and soil exceeded emissions by 2.3 ± 1.3 Gt C per year in the 1990s. Model simulations matching observations are consistent with CO_2 uptake at midlatitudes; however, some models predict these lands will reverse from sinks to sources by midcentury as climate warms. The advantage of tree and soil sequestration is that they do not require separation of combustion products or more fuel. Although not a long-term solution, tree and soil uptake can buy time at low cost with current technology—time for learning, capital turnover, research, and development of advanced power technologies.

A biological sequestration scheme that might store fossil fuel CO_2 longer than trees and soils consists of sinking agricultural residues in bales to deep and/or anoxic zones of the oceans. This is conceptually similar to storing trees in underground vaults and to fertilizing high-latitude ocean plankton with iron. An early proposal by Cesare Marchetti being explored today is CO_2 disposal in the deep sea. Although most fossil fuel CO_2 emitted to the atmosphere dissolves in the oceans on decadal to century timescales, some persists in the atmosphere much longer because of a shift in the carbonate equilibrium of the air/sea system. For a given emission scenario, ocean injections can significantly decrease peak atmospheric CO_2 levels, although all cases eventually diffuse some CO_2 back to the atmosphere.

Carbon dioxide dissolves in water and slowly corrodes carbonate sediment, resulting in the storage of fossil fuel carbon primarily as bicarbonate ions dissolved in the ocean. Left to nature, the net reaction $CO_2 + H_2O + CaCO_3 \rightarrow Ca^{2+} + HCO_3^-$ would occur on the timescale of ~6000 years. Back diffusion to the atmosphere and pH impacts of ocean CO_2 disposal could be greatly diminished by accelerating carbonate mineral weathering reactions that would otherwise slowly neutralize the oceanic acidity produced by fossil fuel CO_2.

A potentially far-reaching removal scheme involves reacting CO_2 with the mineral serpentine to sequester carbon as a solid in magnesium carbonate "bricks." The core idea is to vastly accelerate silicate rock weathering reactions, which remove atmospheric CO_2 over geologic timescales. Atmospheric CO_2 is sequestered on geologic timescales by weathering calcium and magnesium silicate rocks. The net reaction with calcium silicate is schematically represented as $CO_2 + CaSiO_3 \rightarrow SiO_2 + CaCO_3$. Nature (rainfall, runoff, biology, and tectonics) precipitates the calcium carbonate product as seashell sediment and coral reefs. (Atmospheric CO_2 would be depleted in $\sim$400,000 years were it not replenished by CO_2 emissions from volcanoes and hydrothermal vents.)

Carbon dioxide sequestration is a valuable bridge to renewable and/or nuclear energy. If emission-free primary power at 10–30 TW levels is unavailable by midcentury, then enormous sequestration rates could be needed for atmospheric CO_2 stabilization (Fig. 4). In light of these requirements, research and demonstration investments are required now for sequestration technology to be available at the required scale when needed.

4. RENEWABLES

4.1 Biomass

Biomass energy farms with tree regrowth can produce energy with zero net CO_2 emissions. Unfortunately, photosynthesis is too inefficient ($\sim$1%) for bioenergy to be the answer to CO_2 stabilization. Producing 10 TW from biomass would require at least 10% of Earth's land area, comparable to all human agriculture. Competition for land with agriculture and biological diversity is a likely showstopper for bioenergy as a dominant factor in climate stabilization, although it could be a notable contributor in some locales.

4.2 Wind and Solar

Solar and wind with electric currents or H_2 as energy carriers have great potential because their planetary-mean areal power densities are greater than those for biomass, hydroelectricity, ocean, and geothermal power. Considerable effort has been expended in recent years on the suite of renewable energy systems—solar thermal and photovoltaic (PV), wind, hydropower, biomass, ocean thermal, geothermal, and tidal. With the exception of hydroelectricity, these collectively represent $<1\%$ of global energy production. Limited by low areal power densities and intermittency of supply, renewables nonetheless have the potential to power human civilization. However, they require substantial investment in critical enabling technologies before they can be real CO_2 emission-free energy options at the global scale (Figs. 5 and 6).

Solar flux at Earth's orbital distance perpendicular to the sun's rays is $S_o \sim 1370\,W\,m^{-2}$. Sun power at the surface averaged over day–night cycles is reduced by the fraction of Earth's disk to surface areas ($f_1 = 0.25$), cloud fraction ($f_2 \sim 0.5$), and fraction of sunlight reaching the surface under clear skies ($f_3 \sim 0.8$): $S_{surf} = f_1 f_2 f_3 S_o \sim 0.1\,S_o \sim 140\,W\,m^{-2}$. [The factor of 10 solar flux reduction is a major argument for putting solar collectors in space.] Commercial photovoltaic arrays are 10–15% efficient, producing a long-term mean power $\sim$10–20 $W_e\,m^{-2}$ (the theoretical peak for single bandgap crystalline cells is $\sim$24%; it is higher for multi-bandgap cells and lower for the most cost-effective amorphous thin-film cells).

Wind power per surface area depends on the fraction of wind turbine disk to surface area minimizing interference from adjacent turbines ($\sim$0.1), air density ($\rho \sim 1.2\,kg\,m^{-3}$), and wind speed ($U \sim 5\,m\,s^{-1}$): $P_{wind} = 0.1 \times (1/2)\rho U^3 \sim 15\,W\,m^{-2}$. There are typically large local variations. Wind turbine efficiencies of 30–40% yield $\sim 5\,W_e\,m^{-2}$ (theoretical peak is the Betz limit of $\sim$59.3%; actual efficiency depends on wind spectrum and "cut-in" speeds).

The capital costs of PV arrays and wind turbines are rapidly decreasing. Locally, cost-effective wind power is available at many sites that tend to be remote and offshore high-wind locations. Distribution to consumers is limited by long-distance transmission costs. It is argued that the annualized total consumption of electric energy within the United States could be supplied by PV arrays covering a fraction of Nevada's relatively clear-sky deserts. It is similarly argued that the annualized consumption of global electric energy could be supplied by connected arrays of photovoltaics in the deserts of Asia, North Africa, the Americas, and Australia. However, arrays of terrestrial photovoltaics intended to be the sole supplier of large-scale commercial power face daunting problems, including the continental and global distribution of power and uncertainties in the supply of sunlight at the surface of Earth and the total energy storage requirements.

A challenge for demand-driven global energy systems is intermittency from variable clouds and winds. Solar and wind are intermittent, spatially dispersed sources unsuited to baseload electricity without breakthroughs in transmission and storage. Meeting urban electricity demand over day–night cycles and during overcast periods with locally sited PV arrays requires prohibitively massive pumped-storage or lead–acid batteries. A solar–hydrogen economy with electrolyzers, storage tanks, and pipelines could be self-sufficient in principle. However, solar–hydrogen system costs are so high that many advocates of H_2 as an energy carrier promote fossil fuel or nuclear as the primary energy source.

5. ENERGY CARRIERS AND OFF-PLANET ENERGY PRODUCTION

5.1 Hydrogen

A continued historical trend of decreasing carbon content of fuels points to the combustion fuel with zero carbon emissions: hydrogen. However, hydrogen does not exist in underground gas fields or Earth's atmosphere. It must be made from chemical processes requiring energy. If hydrogen is made from fossil fuels (as is commercial H_2 by steam-reforming natural gas), even more carbon or CO_2 must be disposed of than would be produced by burning these fuels directly.

A vision of coal power liberated from environmental constraints incorporates CO_2 capture and sequestration. In one such scheme, coal and/or other fossil fuel feedstocks along with biomass and some waste materials are gasified in an oxygen-blown gasifier, and the product is cleaned of sulfur and reacted with steam to form hydrogen and CO_2. After heat extraction, the CO_2 is sequestered from the atmosphere. The hydrogen can be used as a transportation fuel, or it could be oxidized in a high-temperature fuel cell, with reacted hot gases driving gas turbines or steam generators to make additional electricity. Electrochemical fuel cells are more efficient than heat engines subject to thermodynamic limits (steam turbines, gas turbines, and IC engines). This is somewhat offset by their need for hydrocarbon reforming to H_2 plus CO_2. Like conventional fossil fuel power plants, fuel cells are converters of chemical energy to electricity, not sources of primary power.

Alternately, hydrogen can be produced by electrolysis of water powered by renewable or nuclear sources. Decarbonization and fuel switching to natural gas and eventually H_2 will not mitigate global warming. The underlying problem is the need for non-CO_2-emitting primary power sources during the next 50 years capable of generating 10–30 TW.

Figure 5 depicts electricity routed directly to consumers via grid or employed to make hydrogen from water. (Photoelectrochemical cells combining PV and electrolysis cells in one unit are also being studied.) Water electrolysis employs H_2O as both feedstock and electrolyte (hydroxyl ions, OH^-, present in low concentrations in water are the charge carriers). Commercial electrolyzers employ various electrolytes (and ions) to raise ionic conductivity in alkaline water (K^+ and OH^-), solid polymer PEM (H^+), and ceramic solid oxide (O_2^-) cells. When a potential of approximately 1.5 V is applied, reactions at the electrodes evolve hydrogen and oxygen ($2H_2O + 2e^- \rightarrow H_2 + 2OH^-$ and $2OH^- \rightarrow (1/2)O_2 + H_2O + 2e^-$). Electrolyzers can operate reversibly as fuel cells like rechargeable batteries but with higher power densities because reactants and products flow to and from electrodes as opposed to diffusing though a solid matrix. Power densities per electrode area range from $2\,kW_e\,m^{-2}$ (commercial alkaline water electrolysis) to $20\,kW_e\,m^{-2}$ (PEM cells). The latter require platinum catalysts whose loading has been driven down by intensive research to $5 \times 10^{-3}\,kg\,Pt\,m^{-2}$ without significant loss of performance. However, the high power densities of PEM cells require $5 \times 10^{-4}\,kg\,Pt/kW$ assuming 50% overall efficiency. In other words, PEM cells (electrolyzers and/or fuel cells) need 5000 metric tonne Pt to produce and/or consume H_2 at our canonical 10 TW—30 times today's annual global platinum production rate. Another hurdle for renewable hydrogen is the need for 1 liter per day of chemically pure water per kilowatt, which may be an issue in deserts. Big bang hydrogen is the most abundant element in the universe, but it is too chemically reactive to exist freely on Earth and too light, even in its stable molecular form (H_2), to have been retained by the atmosphere. Hydrogen is a secondary fuel, and water splitting to get H_2 must be powered by primary fuels or energy fluxes in nature.

Renewable electric power systems could in principle evolve from natural gas-powered fuel cells intended to provide grid-independent electricity and cogenenerated heating/cooling to residential and commercial complexes. This path would replace reformers in a later phase with H_2 generated by fuel cells run backward as electrolyzers powered by local PV arrays and/or wind turbines (Fig. 5). With H_2

stored locally, dependence on grids and pipelines could be avoided. Such independence has perceived advantages, particularly during utility outages such as those experienced in California. A natural gas fuel cell path to renewables has the advantage of increasing fossil fuel energy efficiency on the way to renewable power, although it may require internalizing the costs of air capture of its emitted CO_2. Electricity costs from decentralized fuel cells are still many times more than those from utility grids, owing largely to precious metal catalysts employed by commercial reformers and fuel cell stacks.

5.2 Advanced Electric Grids

Advanced technology global electrical grids may be more promising for renewables. With current transmission lines, even if costs per kilowatt-hour of PV arrays and turbines declined drastically and production rates were increased to millions per year like mass-produced automobiles, the grids of the United States and the world could not handle the load-management demands. Power must flow where needed, when needed. Existing grids are hub-and-spoke networks designed for central power plants tens to hundreds of kilometers from users. These need to be reengineered into "smart grids"—a global Internet and superhighway system for electricity.

Expanding such grids to the global scale of tens of thousands of kilometers and undersea links between continents evokes the global electrical grid proposed by R. Buckminster Fuller (Fig. 6). Fuller envisioned electricity produced worldwide "wheeled" between day and night hemispheres and pole-to-pole to balance supply and demand with minimum power loss. Fuller's global grid had relatively short links between continents. A major technology issue may be maintaining the undersea grid segments.

Long-distance transmissions might be accomplished initially by high-voltage direct current lines using conventional aluminum/copper conductors. A more revolutionary technology exploits the discovery of "high-temperature" superconducting ceramic copper oxide wires. The electrical resistance of high-temperature superconductors vanishes below the boiling point of liquid nitrogen (77 K), which is much easier to produce from chilled air than from the 4 K liquid helium previously needed. Nitrogen-cooled cables have been fabricated as transmission lines. Long-term implications of high-temperature superconductors transmission lines for renewables and nuclear are profound.

Routing electricity over tens of thousands of kilometers loss-free by "smart grids" could permit solar energy to flow from the Sahara to sub-Saharan Africa, China, and India and nuclear energy to politically unstable regimes. Modern computer technology may permit grids smart enough to match supply from intermittent renewable sources to base-load demands worldwide. The ultimate electric utility deregulation would have buyers and sellers worldwide creating the market price of a kilowatt-hour on the grid.

5.3 Wireless Power Transmission and Solar Power Satellites

Approximately 40% of Earth's population is "off the grid," mainly in developing countries. Wireless power transmission envisioned by Nikola Tesla a century ago is feasible today. Microwave beams can propagate power efficiently along lines-of-sight over long distances. Orbiting microwave reflectors could form the basis of a global electric grid (Fig. 7).

An advanced technology path to electrification is the solar power satellite (SPS) proposed by Peter Glaser (Fig. 7). Solar flux is ~ 10 times higher in space outside Earth's shadow cone than the long-term average at the surface of spinning, cloudy Earth, and power from space can be beamed by microwave efficiently through cloudy skies to the surface where it is needed.

Rockets routinely put PV-powered satellites in orbit, and the Moon, visited decades ago by astronauts, can provide a stable platform and materials. NASA and the Department of Energy studied a reference solar power satellite design in the 1970s with a solar array the size of Manhattan in geostationary orbit (35,000 km above the equator) beaming power to a 10×13-km rectenna on the surface at 35° latitude with 5 GW_e output to a grid. Today, the capital investment to build even one appears too mind-boggling to proceed, even if launch costs decline enough to make electricity cost-effective at 3 or 4¢ $(kW_e\text{-hr})^{-1}$. New ideas include satellite constellations providing integrated electric power and communication/Internet services with lower capital costs to "first power."

A solar power satellite in a lower orbit can be smaller and cheaper than one in geostationary orbit because it does not spread its beam as much, but it does not appear fixed in the sky and has a shorter duty cycle (the fraction of time power is received at a given surface site). When a line-of-sight exists the

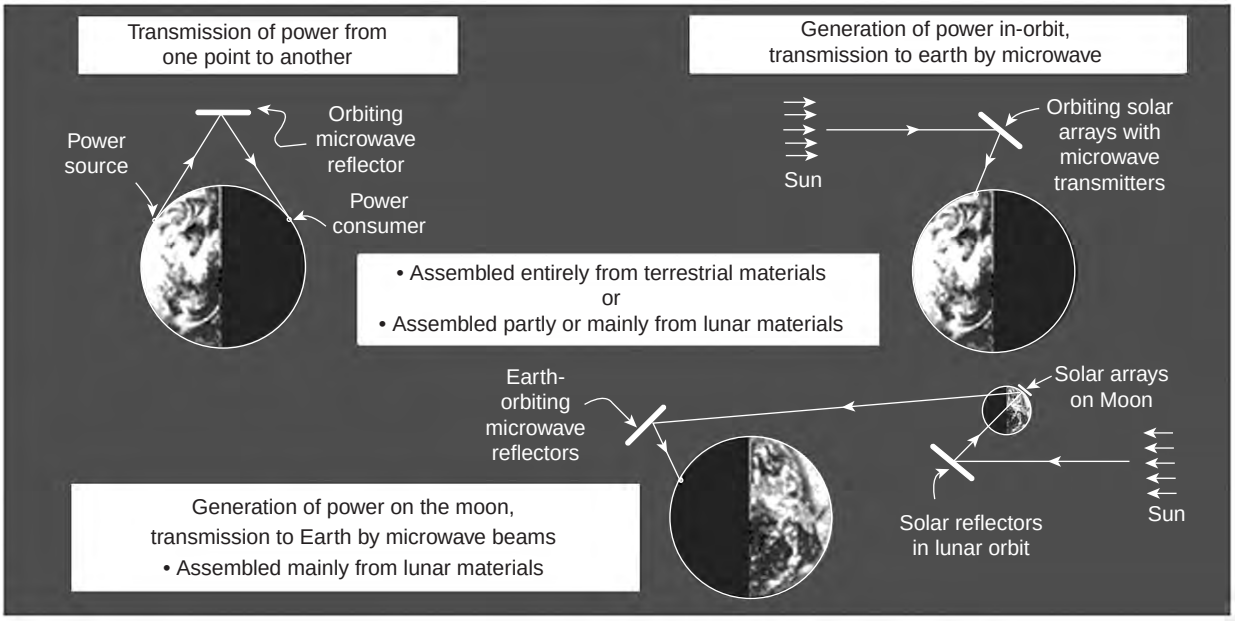


FIGURE 7 Capturing and controlling sun power in space. Microwave beams can propagate power efficiently along lines-of-sight over long distances. The power relay satellite, solar power satellite, and lunar power system all exploit unique attributes of space (high solar flux, lines-of-sight, lunar materials, and shallow gravitational potential well of the moon). In principle, a ring of these in geostationary orbit could power human civilization.

beam from a phased-array transmitter can be slued to track a surface rectenna. The flat transmitter array of aperture D_t is steered electronically with phase data from a computer homing on a pilot beam from the surface. Power is collected on the surface by a rectenna of aperture D_r of quarter-wave antenna and Shottky diode elements. Most of the rectenna is empty space, permitting vegetation and grazing below (Fig. 8). Biota and land use impacts could be minimal. A rectenna can produce the same power with 1/10 the areal footprint as PV panels for microwave intensities comparable to those routinely experienced by cell phone users (>100 million worldwide). Theoretically, transmission efficiency, $\eta_t = (\text{DC power out})/(\text{EM power transmitted})$, is $>90\%$ when $D_t D_r/(\lambda d) \geq 2$, where d is distance and with $\eta_t = 60\%$ a reasonable target. Sunsats bypassing fossil-fueled central power stations offers a CO_2 emission-free path to developing nation electrification analogous to bypassing telephone lines with cell phones.

Important progress has been made in SPS enabling technologies: solar power generation (already at 30% efficiency and increasing), wireless power transmission, intelligent systems, and others. A major issue is the scale needed for net power generation. Beam diffraction drives component sizes up for distant orbits. (The "main lobe" where power is concentrated spreads by a small angle even in coherent diffraction-limited beams.) Unlike the case for fusion, no basic scientific breakthroughs are needed. Dramatic reductions in launch costs that NASA is pursuing independently could accelerate the technology.

A proposed short-term demonstration project would beam solar energy from low orbit. Relatively small equatorial satellites could supply electricity to developing nations within a few degrees of the equator. This appears feasible, but it would require a large number of satellites for orbits below the earth's radiation belts for continuous power. Papua New Guinea, Indonesia, Ecuador, and Columbia on the Pacific Rim, as well as Malaysia, Brazil, Tanzania, and the Maldives, have agreed to participate in a demonstration project.

One cost-reduction strategy is the building of solar power satellites from materials mined and launched from the moon. This requires less energy than launches from Earth but implies a long-term commitment to revisit the moon and exploit lunar resources. For power generation and transmission, geostationary orbit 36,000 km above the equator has the advantage of presenting a fixed point in the sky but has disadvantages as well. Alternate locations are 200- to 10,000-km altitude satellite constellations, the moon, and the earth–sun L_2 Lagrange point.

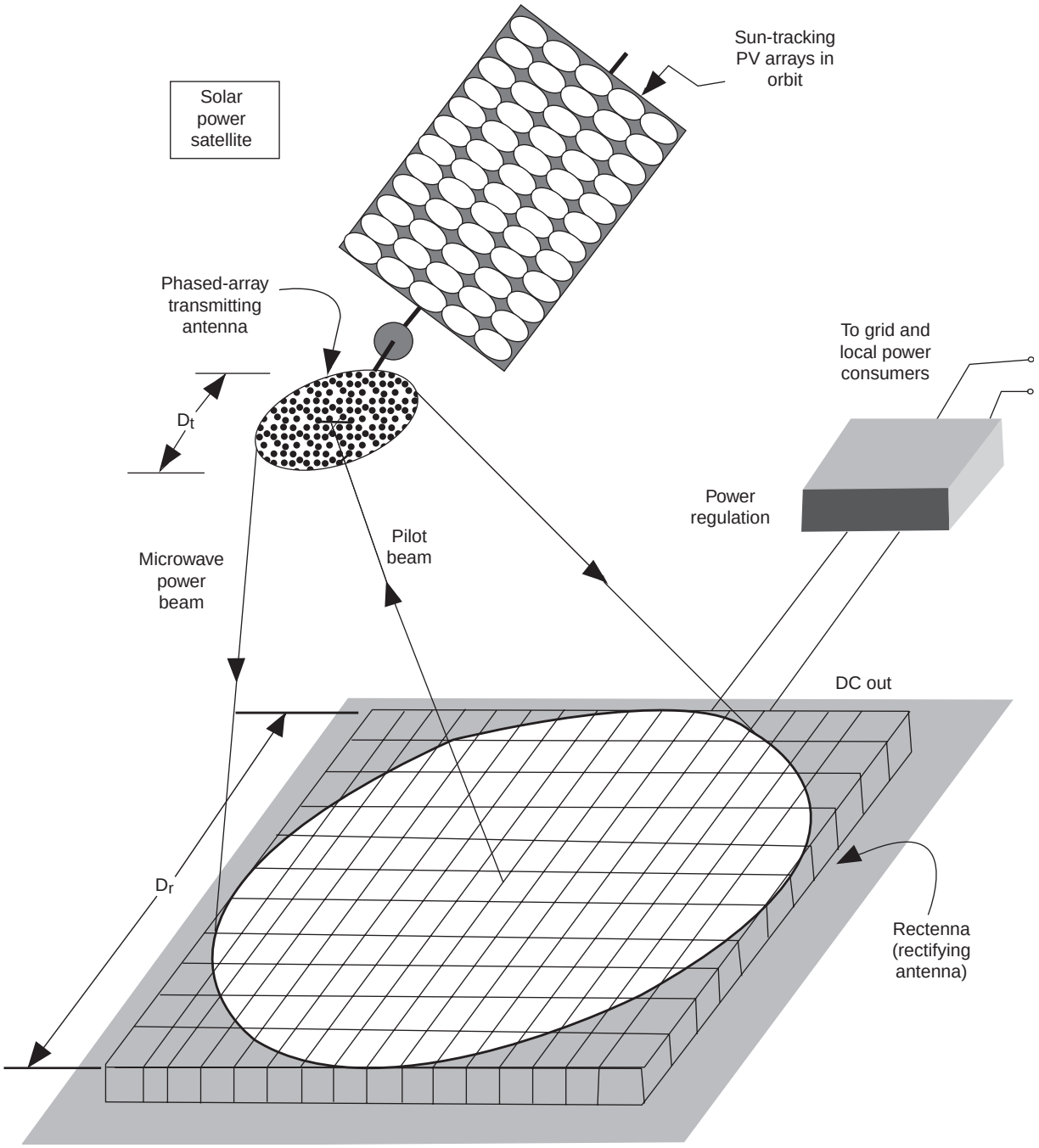


FIGURE 8 When a line-of-sight exists, the beam from a phased-array transmitter can be slued to track a surface rectenna. Most of the rectenna is empty space, permitting vegetation and grazing below. Biota and land use impacts could be minimal. A rectenna can produce the same power with 1/10 the areal footprint as photovoltaic (PV) panels for microwave intensities comparable to those routinely experienced by cell phone users.

With adequate research investments, space solar power could be matured and demonstrated in the next 15–20 years, with a full-scale pilot plant (producing $\sim$1–3 GW_e for a terrestrial market) in the 2025–2040 timeframe. Solar power beamed from space has the potential to deliver baseload power to global markets remote from electrical transmission lines and/or hydrogen pipelines. It could thus be an effective complement to terrestrial renewable electricity for developing nations. In the long term, space

solar power could help stabilize global climate, provide a transition from military projects for aerospace industries, create new "orbital light and power" industries, and foster international cooperation.

6. FISSION AND FUSION

6.1 Fission

Nuclear electricity is fueled by fissioning the U-235 isotope—0.72% of natural uranium, usually slightly enriched in fuel rods. Neutrons liberated from fissioning nuclei (two or three per fission event) are moderator slowed to foster chain reactions in further uranium nuclei. Approximately 200 MeV of energy is released in each fission and appears as kinetic energy of the fission fragments. Reactors operate near steady states (criticality) in which heat is transferred by coolants to steam turbogenerators. Current nuclear reactor technology provides CO_2 emission-free electricity while posing unresolved problems of waste disposal and nuclear weapons proliferation.

Otto Hahn and Fritz Strassmann (working with Lisa Meitner) discovered that bombarding natural uranium with neutrons moving at a few eV split the nucleus, releasing a few hundred million eV: $^{235}U + n \rightarrow$ fission products $+ 2.43n + 202$ MeV. The ^{235}U isotope, 0.72% of natural uranium, is often enriched to 2 to 5% in reactor fuel rods. The ~ 500 nuclear power plants worldwide today are technology variants of ^{235}U thermal reactors: the light water reactor (LWR, in both pressurized and boiling versions); heavy water (CANDU); graphite-moderated, steam-cooled (RBMK), like the plant at Chernobyl; and gas-cooled graphite. A total of 85% are LWRs, based on Hyman Rickover's choice for the first nuclear submarine and commercial power reactor. The first electricity-generating commercial reactor (Shippingsport, PA) went online on December 2, 1957—18 years after the Hahn–Strassmann paper.

The LWR, primarily a "burner" powered by uranium slightly enriched in ^{235}U in zirconium alloy-clad fuel rods, employs H_2O as both coolant and working fluid (Fig. 9). Power is controlled by moderator rods that increase fission rates when lowered into the core. LWRs and their variants, and a smaller number of reactors employing D_2O (heavy water), produce CO_2 emission-free power (18% of global electricity; $<5\%$ primary power). However, they are susceptible to loss-of-coolant meltdowns and their efficiencies are limited by 600 K steam temperatures exiting the boiler. (Boiler pressures are set lower in nuclear plants [≤ 160 bar] than in fossil-fueled plants because of radioactivity danger from explosions.)

The "passively safe" helium-cooled, pebble-bed reactor (Fig. 9) is theoretically immune to meltdowns. Helium coolant exits the core at 1200 K, driving a 40% efficient gas turbine/generator. Its graphite-coated fuel pebbles are designed not to melt even if coolant flow stops.

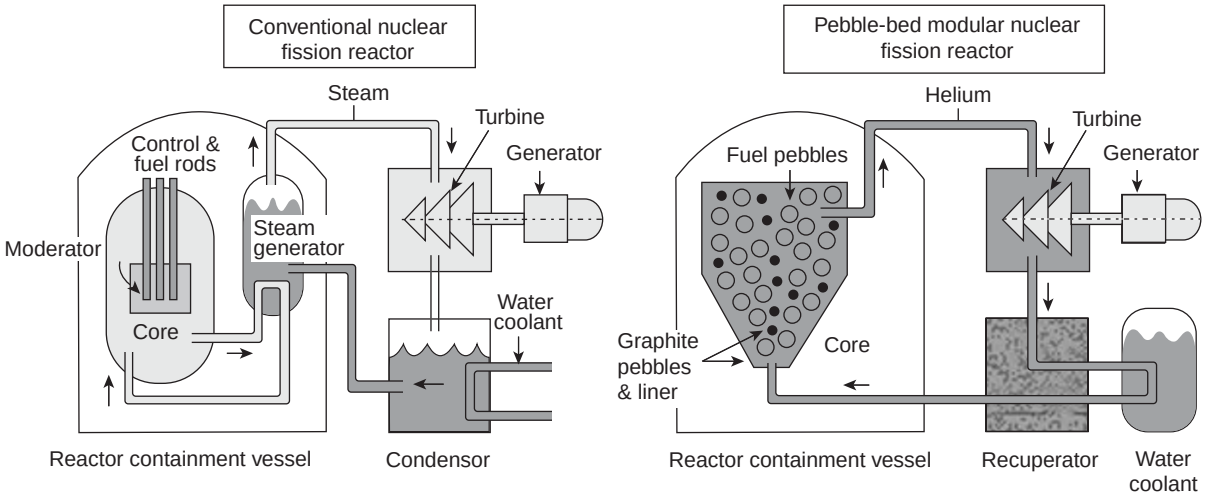


FIGURE 9 Fission. The light water reactor (LWR), primarily a burner powered by uranium slightly enriched in ^{235}U in zirconium alloy-clad fuel rods, employs H_2O as both coolant and working fluid. Power is controlled by moderator rods that increase fission rates when lowered into the core. They are susceptible to loss-of-coolant meltdowns and have efficiencies limited by 600 K steam temperatures exiting the boiler. The passively safe helium-cooled, pebble-bed reactor shown is theoretically immune to meltdowns. Helium coolant exits the core at 1200 K, driving a 40%-efficient gas turbine/generator.

6.2 Fuel Supply and Breeders

The overriding hurdle for fission as an emission-free power source is the mind-boggling scale of power buildup required: "Once through" reactors burning ^{235}U at 10 TW will exhaust identified ore deposits in 5–30 years, and mining the oceans for uranium requires seawater processing at flow rates more than five times global river runoff. Breeding fertile ^{232}Th and ^{238}U can increases fissile fuels $\sim$100 times, but a massive commitment to neutron production is needed to make fuel fast enough. Moreover, diversion of nuclear fuel by terrorists and rogue regimes and disposal of high-level reactor wastes are unresolved issues.

Current best estimates of U-235 primary energy in crustal ores are 60–300 TW/year. Current estimates of uranium in proven and ultimately recoverable resources are 3.4 and 17 million tonnes, respectively. Deposits containing 500–2000 ppmw are considered recoverable. Vast amounts of uranium in shales at concentrations of 30–300 ppmw are unlikely to be recoverable because of high processing and environmental costs. The fission energy per tonne ^{235}U is

$$\frac{202\ \text{MeV}}{235\ \text{amu} \times 1.661 \times 10^{-27}\ \text{kg/amu}} \times \frac{1.602 \times 10^{-13}\text{J}}{1\ \text{MeV}} \times \frac{10^3\ \text{kg}}{1\ \text{tonne}} = 8.3 \times 10^{16} \frac{\text{J}}{\text{tU-235}}.$$

The energy content of the U-235 isotope in natural uranium is therefore $0.0072 \times 8.3 \times 10^{16}$ J t$^{-1} \sim 5.7 \times 10^{14}$ J t^{-1}, and the energy in proven uranium reserves and ultimately recoverable resources is $\sim$61 and 300 TW/year, respectively (1 TW/year = 31.5 EJ). At 10 TW, reserves and ultimate resources of fissionable uranium last only 6 and 30 years, respectively. However, the recoverable uranium ore may be underestimated due to lack of exploration incentives.

What about the seas? Japanese researchers propose harvesting dissolved uranium with organic particle beds immersed in flowing seawater. The oceans contain 3.2×10^{-6} kg of dissolved uranium per cubic meter—a U-235 energy density of 1.8 million J/m^3. Multiplying by the oceans' huge volume (1.37×10^{18} m^3) gives large numbers: 4.4 billion tonnes of uranium and 80,000 TW/year in U-235. Even with 100% U-235 extraction, the volumetric flow rate to make reactor fuel at the 10-TW rate is five times the outflow of all Earth's rivers, which is 1.2×10^6 m^3 s^{-1}. By comparison, the flow rate through a hypothetical seawater extraction plant to yield ^{235}U fuel at 10 TW is

$$10\ \text{TW} \times \frac{1.37 \times 10^{18}\text{m}^3}{8 \times 10^4\ \text{TW year}} \times \frac{1\ \text{year}}{3.15 \times 10^7\text{s}} = 5.4 \times 10^6 \text{m}^3\text{s}^{-1},$$

approximately five times global river runoff.

Fissionable fuels could be made by breeding Pu-239 and/or U-233, which do not exist in nature. They must be bred from their feed stocks, U-238 ($\sim$99% of uranium) and Th-232 ($\sim$100% of thorium), using suitable neutron sources ($^{238}\text{U} + n \rightarrow {}^{239}\text{Pu}$; $^{232}\text{Th} + n \rightarrow {}^{233}\text{U}$). Fission energy per unit mass is comparable to U-235 ($^{239}\text{Pu} + n \rightarrow$ fission products $+ 2.9n + 210$ MeV; $^{233}\text{U} + n \rightarrow$ fission products $+ 2.5n + 198$ MeV). We estimate that breeding could increase fission energy resources by factors of 60 and 180 for Pu-239 and U-233, respectively. Breeding rates depend on feedstock mining rates and availability of neutrons.

Breeders might be more attractive if fuel cycles safely transmuting high-level wastes could be developed. The conclusion is inescapable that breeders have to be built very soon for fission to significantly displace CO_2 emissions. If one goes the breeder route, then thorium is a preferable feedstock to uranium because U-233 is more difficult than plutonium to separate from nuclear wastes and divert to weapons.

6.3 Fusion

The best hope for sustainable carbon emission-free nuclear power may not be fission but, rather, fusion. The focus so far has been on the D–tritium (T) reaction ($\text{D} + \text{T} \rightarrow {}^4\text{He} + n + 17.7$ MeV). Controlled fusion is difficult. Temperatures in the 30- to 300-million K range are needed, hence the skepticism about "cold fusion." To ignite the plasma, energy is normally input by neutral particle beams or lasers. In a reactor, energy would emerge as neutrons penetrating reactor walls to a liquid lithium blanket ultimately driving steam turbines.

The most successful path to fusion has been confining a D + T plasma with complex magnetic fields and heating it to ignition by neutral particle beams in a toroidal near-vacuum chamber (tokamak; Fig. 10). Classically, the 100- to 300-million K plasma should be contained by the magnetic fields, but in the real world "anomalous" diffusion (turbulence) has taken decades to suppress. "Breakeven" is when input power equals the power of fusion

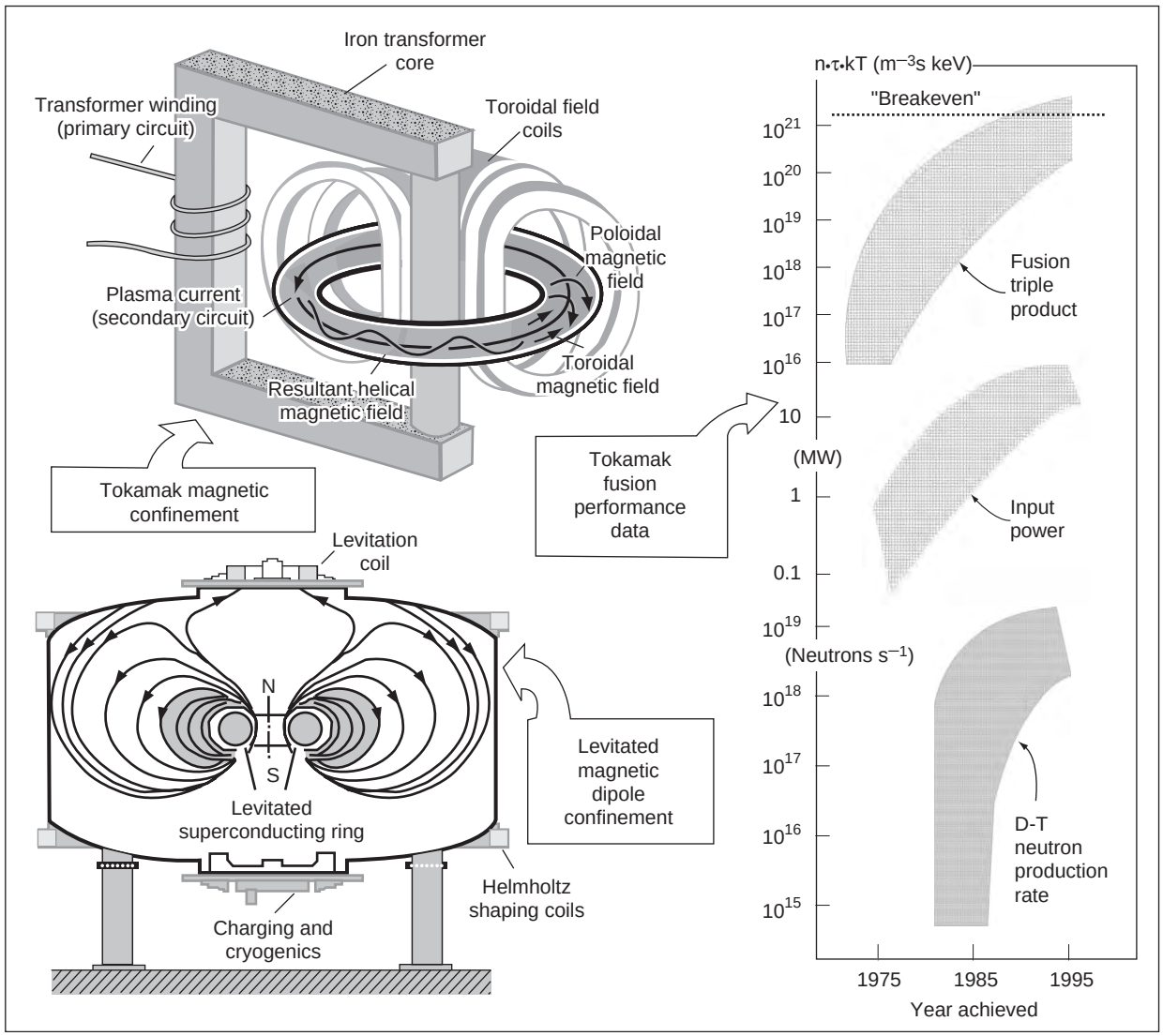


FIGURE 10 Fusion. The most successful path to fusion has been confining a deuterium + tritium plasma with complex magnetic fields and heating it to ignition by neutral particle beams (not shown) in a toroidal near-vacuum chamber (tokamak). Breakeven is when input power equals the power of fusion neutrons hitting the innermost reactor wall. This requires the fusion triple product (number density × confinement time × temperature) to equal a critical value. Despite progress, engineering breakthroughs are needed for fusion electricity to be viable. Experimental work on advanced fusion fuel cycles is being pursued at the National Ignition Facility and on simpler magnetic confinement schemes such as the levitated dipole experiment shown.

neutrons hitting the innermost reactor wall. This requires the fusion triple product (number density × confinement time × temperature) to equal a critical value. Recent tokamak performance improvements have been dramatic, capped by near breakeven. Commercial plants would surround the torus wall with neutron-absorbing molten lithium blankets where tritium is bred and heat exchanged with turbogenerators—the latter feeding electricity back to heaters and magnets, with excess power to consumers. Despite progress, engineering breakthroughs are needed for fusion electricity to be viable. Short-term experiments are advocated on accelerated breeding of fissile fuels from the high neutron flux in D–T fusion blankets. Fission–fusion hybrids could impact CO_2 stabilization by midcentury, even if pure fusion is delayed. (Current nuclear weapons, for example, use both fission and fusion nuclear reactions to create powerful explosions.) Experimental work includes simpler magnetic confinement schemes such as the levitated dipole experiment shown in Fig. 10.

Currently, there is no fusion-generated electricity. However, researchers have very nearly achieved breakeven [$Q =$ (neutron or charged particle energy out)/(energy input to heat plasma) $= 1$] with tokamak magnetic confinement proposed in the 1950s by Igor Tamm and Andrei Sakharov. Reactor power balances indicate that $Q = 1$ for the D–T reaction requires the product of number density n, confinement time τ, and temperature T to satisfy the Lawson criteria: $n \times \tau \times kT \sim 1 \times 10^{21}\,\mathrm{m^{-3}\,s\,keV}$, where $k = 0.0862\,\mathrm{keV/MK}$. Experimental results from the TFTR and JET tokamaks are within a factor of two of this value. The Lawson criteria for breakeven with D–D and D^{-3}He reactions requires a factor of 10 or higher triple product; net power-generating reactors require even higher Qs and triple products.

Tritium has to be bred from lithium ($n + {}^6\mathrm{Li} \rightarrow {}^4\mathrm{He} + \mathrm{T} + 4.8\,\mathrm{MeV}$). Reactors burning D–T mixtures are thus lithium burners. Lithium in crustal ores bred to tritium could generate 16,000 TW/year, approximately twice the thermal energy in fossil fuels. This is no infinite power source, although more lithium might be mined from the sea. Deuterium in the oceans is virtually unlimited, whether utilized in the D–T reaction or the more difficult to ignite D–D reactions ($\rightarrow {}^3\mathrm{He} + n +$ 3.2 MeV and $\rightarrow \mathrm{T} + p + 4.0\,\mathrm{MeV}$). D–T tokamaks give the best plasma confinement and copious neutrons.

New devices with either a larger size or a larger magnetic field strength are required for net power generation. Demonstrating net electric power production from a self-sustaining fusion reactor could be a breakthrough of overwhelming importance. A more near-term idea is fission–fusion hybrids. These could breed fissionable fuel from thorium mixed with lithium in tokamak blankets. Each 14-MeV fusion neutron, after neutron multiplication in the blanket, can breed one ^{233}U atom (burned separately in a nuclear power plant) and one T atom. Since $^{233}\mathrm{U} + n$ fission releases $\sim 200\,\mathrm{MeV}$, the power output of a fission–fusion plant could be ~ 15 times that of a fusion plant alone. A fusion breeder could support many more satellite reactors than a fission breeder. Fusion is energy poor but neutron rich, whereas fission is energy rich but neutron poor: Their marriage might breed fissionable fuel fast enough to help stabilize the fossil fuel greenhouse. Our analysis indicates that a fission–fusion hybrid plant based on existing tokamak technology but requiring major experimental confirmation might begin breeding fuel for 300-MW reactors in 10–15 years. However, the fission–fusion breeder path will also require (actually and perceived) safe ^{233}U fission reactor and fuel cycles by midcentury.

More difficult to ignite than D–T, there is renewed interest in the D–^{3}He reaction ($\mathrm{D} + {}^3\mathrm{He} \rightarrow {}^4\mathrm{He} + p + 18.3\,\mathrm{MeV}$), which yields charged particles directly convertible to electricity with little radioactivity (some neutrons may be produced by side reactions). Studies of D–^{3}He and D–D burning in inertial confinement fusion targets are being conducted at Lawrence Livermore National Laboratory. Computer modeling suggests that fast-ignited D–T spark plugs can facilitate ignition. Moreover, targets can be designed to release the vast majority of their fusion energy as high-energy charged particles leading to applications such as high-specific-impulse propulsion for deep-space missions and direct energy conversion for power generation. Experiments are also under way testing dipole confinement by a superconducting magnet levitated in a vacuum chamber. This method of plasma confinement was inspired by satellite observations of energetic plasma within the magnetospheres of the outer planets, and it may be a prototype for D–^{3}He fusion since theories predict low heat conduction at high plasma pressure. Rare on Earth, ^{3}He may be cost-effective to mine from the moon and is even more abundant in (and potentially recoverable from) atmospheres of gas giant planets. Deuterium in seawater and outer planet ^{3}He could provide global-scale power for a longer period than from any source besides the sun.

Fission and fusion offer several options to combat global warming. However, they can only become operational by midcentury if strategic technologies are pursued aggressively now through research and development and, in the case of nuclear fission, if outstanding issues of high-level waste disposal and weapons proliferation are satisfactorily resolved.

7. GEOENGINEERING

No discussion of global warming mitigation would be complete without mentioning geoengineering, which is also called climate engineering or planetary engineering on Earth and terraforming on other planets. CO_2 capture and sequestration having entered the mainstream, geoengineering today refers mainly to altering the planetary radiation balance to affect climate. Roger Revelle and Hans Suess recognized early on that mining and burning fossil fuels to power our industrial civilization was an

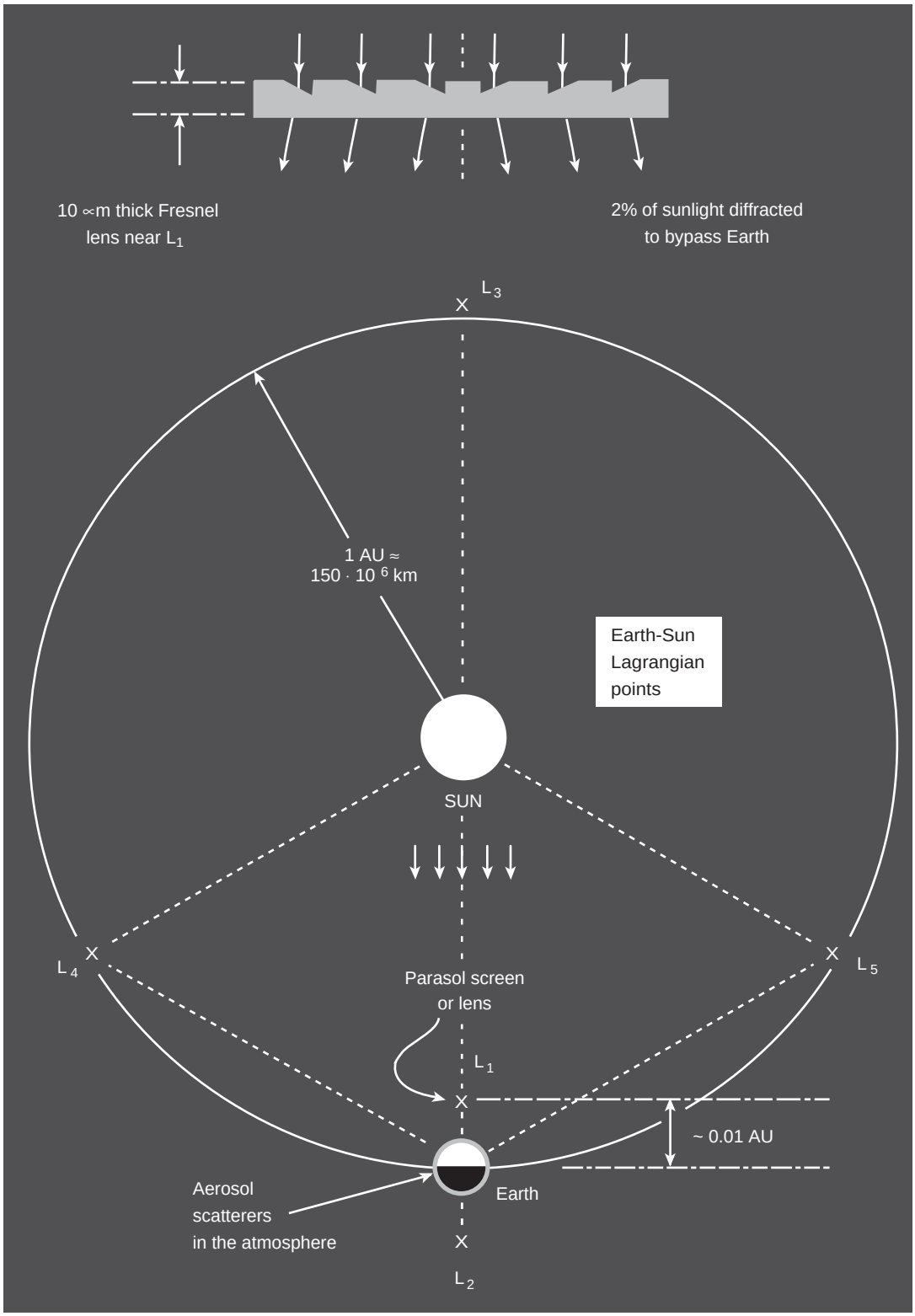


FIGURE 11 Space-based geoengineering. A 2000-km-diameter parasol near L1 could deflect 2% of incident sunlight, roughly compensating for the radiative forcing from CO_2 doubling. Deployment of stratospheric aerosols with optical properties engineered to compensate for adverse climatic change is also proposed, although continuous injections of upper-atmosphere particles to maintain 2% sunblocking could have unintended adverse consequences.

unprecedented technological intervention in the carbon cycle:

> *Human beings are now carrying out a large-scale geophysical experiment of the kind that could not have happened in the past nor be reproduced in the future. Within a few centuries we are returning to the atmosphere and oceans the concentrated organic carbon stored in sedimentary rocks over hundreds of millions of years.*

Geoengineering employs technologies to compensate for the inadvertent global warming produced by fossil fuel CO_2 and other greenhouse gases. An early idea by Soviet climatologist Mikhail Budyko would put layers of reflective sulfate aerosol in the upper atmosphere to counteract greenhouse warming. Variations on the sunblocking theme include injecting submicron dust to the stratosphere in shells fired by naval guns and increasing cloud cover by seeding and shadowing Earth by objects in space. Perhaps most ambitious is James Early's proposed 2000-km-diameter mirror of 10-μm glass fabricated from lunar materials at the L_1 Lagrange point of the sun–earth system illustrated in Fig. 11. The mirror's surface would look like a permanent sunspot, deflect 2% of solar flux, and roughly compensate for the radiative forcing of a CO_2 doubling. Climate model runs indicate that the spatial pattern of climate would resemble that without fossil fuel CO_2. Among other things, this scheme would require renegotiating existing treaties relating to human activities in space.

The Lagrange interior point L_1, one of the five libration points corotating with the earth–sun system, provides an opportunity for negative radiative forcing opposing global warming. However, solar parasols are currently more interesting as thought experiments than as real technology options. Deployment of stratospheric aerosols with optical properties engineered to compensate for adverse climatic change has also been proposed, although continuous injections of upper-atmosphere particles to maintain 2% sunblocking could have unintended adverse consequences. A better investment may be space power, which can produce 10 TW_e (equivalent to 30 TW primary) with $<3\%$ of the disk area needed for a $2 \times CO_2$ compensating parasol or aerosol layer and provide long-term renewable energy for Earth.

It seems prudent to pursue geoengineering research as an insurance policy should global warming prove worse than anticipated and conventional measures to stabilize CO_2 and climate fail or prove too costly. However, large-scale geophysical interventions are inherently risky and need to be approached with caution.

8. CONCLUSION

Even as reliable evidence for global warming accumulates, the fundamental dependence of our civilization on the oxidation of coal, oil, and gas for energy makes it difficult to respond appropriately. This condition will likely become more acute as the world economy grows and as ever larger reductions in CO_2-emitting energy relative to growing total energy demand are needed to stabilize atmospheric CO_2. Energy is critical to our aspirations for global prosperity and equity. If Earth continues to warm, as predicted, the people of Earth may turn to advanced, non-CO_2-emitting energy technologies for solutions. Combating global warming by radical restructuring of the global energy system may be a major research and development challenge of the century. It is likely to be a difficult but rewarding enterprise—one that could lead to sustainable global-scale power systems compatible with planetary ecosystems.

It is unlikely that a single "silver bullet" will solve this problem. We have identified a portfolio of promising advanced energy technologies here, many of which are radical departures from our current fossil fuel system. There are doubtless other good ideas. Inevitably, many concepts will fail, and staying the course will require leadership. Consider how John F. Kennedy voiced the goals of the Apollo program: "We choose to go to the moon in this decade, and to do the other things, not because they are easy, but because they are hard. Because that goal will serve to organize and measure the best of our energies and skills."

Stabilizing global climate change is not likely to be easy. At the very least, it will require political will, targeted research and development, and international cooperation. Most of all, it will require the recognition that, although regulation can play a role, the fossil fuel greenhouse is an energy problem that cannot be simply regulated away. Solutions need to be invented, tested, and implemented—an effort that could "organize and measure the best of our energies and skills."

SEE ALSO THE FOLLOWING ARTICLES

Biomass for Renewable Energy and Fuels • Biomass: Impact on Carbon Cycle and Greenhouse Gas Emissions • Carbon Capture and Storage from Fossil Fuel Use • Carbon Sequestration, Terrestrial • Carbon

Taxes and Climate Change • Climate Change and Public Health: Emerging Infectious Diseases • Climate Change: Impact on the Demand for Energy • Climate Protection and Energy Policy • Energy Efficiency and Climate Change • Greenhouse Gas Emissions from Energy Systems, Comparison and Overview

Further Reading

Department of Energy "The Vision 21 Energy Plant of the Future." Available at http://www.fe.doe.gov.

Fowler, T. K. (1997). "The Fusion Quest." Johns Hopkins Univ. Press, Baltimore.

Glaser, P. E., Davidson, F. P., and Csigi, K. I. (eds.). (1997). "Solar Power Satellites." Wiley–Praxis, New York.

Hoffert, M. I., Caldeira, K., Jain, A. K., Haites, E. F., Harvey, L. D. D., Potter, S. D., Schlesinger, M. E., Schneider, S. H., Watts, R. G., Wigley, T. M. L., and Wuebbles, D. J. (1998). Energy implications of future stabilization of atmospheric CO_2 content. *Nature* **395,** 881–884.

Hoffert, M. I., Caldeira, K., Benford, G., Criswell, D. R., Green, C., Herzog, H., Katzenberger, J. W., Kheshgi, H. S., Lackner, K. S., Lewis, J. S., Manheimer, W., Mankins, J. C., Marland, G., Mauel, M. E., Perkins, L. J., Schlesinger, M. E., Volk, T., and Wigley, T. M. L. (2002). Advanced technology paths to global climate stability: Energy for a greenhouse planet. *Science* **295,** 981–987.

Johansson, T. B., Kelly, H., Reddy, A. K. N., and Williams, R. H. (eds.). (1993). "Renewable Energy." Island Press, Washington, DC.

Metz, B., Davidson, O., Swart, R., and Pan, J. (eds.). (2001). "Climate Change 2001: Mitigation: Contribution of Working Group III to the Third Assessment Report of the Intergovernmental Panel on Climate Change." Cambridge Univ. Press, New York.

Nakicenovic, N., and Swart, R. (eds.). (2001). "Emissions Scenarios: A Special Report of Working Group III of the Intergovernmental Panel on Climate Change." Cambridge Univ. Press, New York.

Nakicenovic, N., Grübler, A., and McDonald, A. (1998). "Global Energy Perspectives." Cambridge Univ. Press, New York.

National Laboratory Directors (1997). "Technology Opportunities to Reduce U.S. Greenhouse Gas Emissions." U.S. Department of Energy, Washington, D.C. Available at http://www.ornl.gov.

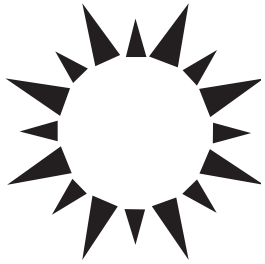
National Research Council (1996). "Nuclear Wastes: Technologies for Separations and Transmutations." National Academy Press, Washington, DC.

Reichle, D., *et al.* (1999). "Carbon Sequestration Research and Development." Office of Fossil Energy, Department of Energy, Washington, DC. Available at http://www.ornl.gov.

Smil, V. (1999). "Energies: An Illustrated Guide to the Biosphere and Civilization." MIT Press, Cambridge, MA.

Watts, R. G. (2002). "Innovative Energy Systems for CO_2 Stabilization." Cambridge Univ. Press, New York.

Climate Change and Public Health: Emerging Infectious Diseases



PAUL R. EPSTEIN
Harvard Medical School Center for Health and the Global Environment
Boston, Massachusetts, United States

1. **Background on Climate Change**
2. **Climate and Infectious Disease**
3. **Pest Control: One of Nature's Services**
4. **Climate Change and Biological Responses**
5. **Sequential Extremes and Surprises**
6. **Case Study: West Nile Virus**
7. **Discontinuities**
8. **Conclusion**
9. **Next Steps**

Glossary

climate change The heating of the inner atmosphere, oceans, and land surfaces of the earth. The warming is associated with more intense extreme weather events and the altered timing, intensity, and distribution of precipitation.

emerging infectious diseases Those diseases new to medicine since 1976, plus old diseases undergoing resurgence and redistribution.

Climate restricts the range of infectious diseases, whereas weather affects the timing and intensity of outbreaks. Climate change scenarios project a change in the distribution of infectious diseases with warming and changes in outbreaks associated with weather extremes, such as flooding and droughts. The hypothesis of this article is that the ranges of several key diseases or their vectors are already changing in altitude due to warming, along with shifts in plant communities and the retreat of alpine glaciers. In addition, more intense and costly weather events create conditions conducive to outbreaks of infectious diseases, as heavy rains leave insect breeding sites, drive rodents from burrows, and contaminate clean water systems. Conversely, drought can spread fungal spores and spark fires (and associated respiratory disease). In addition, sequences of extremes can destabilize predator/prey interactions, leading to population explosions of opportunistic, disease-carrying organisms. Advances in climate forecasting and health early warning systems may prove helpful in catalyzing timely, environmentally friendly public health interventions. If climate change continues to be associated with more frequent and volatile and severe weather events, we have begun to see the profound consequences that climate change can have for public health and the international economy.

> *Epidemics are like signposts from which the statesman of stature can read that a disturbance has occurred in the development of his nation—that not even careless politics can overlook.*
>
> *—Rudolf Virchow (1848)*

1. BACKGROUND ON CLIMATE CHANGE

The climate system can remain stable over millennia due to interactions and feedbacks among its basic components: the atmosphere, oceans, ice cover, biosphere, and energy from the sun. Harmonics among the six orbital (Milankovitch) cycles (e.g., tilt and eccentricity) of the earth about the sun, as revealed by analyses of ice cores and other "paleothermometers" (e.g., tree rings and coral cores), have governed the oscillations of Earth's climate between ice ages and warm periods until the 20th

 © 2004 Elsevier Inc. All rights reserved.

century. To explain the global warming during the 20th century of approximately 1°C, according to all studies reviewed by the Intergovernmental Panel on Climate Change, one must invoke the role of heat-trapping greenhouse gases (GHGs), primarily carbon dioxide (CO_2), oxides of nitrogen, methane, and chlorinated hydrocarbons. These gases have been steadily accumulating in the lower atmosphere (or troposphere), out to about 10 km, and have altered the heat budget of the atmosphere, the world ocean, land surfaces, and the cryosphere (ice cover).

For the past 420,000 years, as measured by the Vostok ice core in Antarctica, CO_2 has stayed within an envelope of between 180 and 280 parts per million (ppm) in the troposphere. Today, the level of CO_2 is 370 ppm and the rate of change during the past century surpassed that observed in ice core records. Ocean and terrestrial sinks for CO_2 have presumably played feedback roles throughout millennia. Today, the combustion of fossil fuels (oil, coal, and natural gas) is generating CO_2 and other GHGs, and the decline in sinks, primarily forest cover, accounts for 15–20% of the buildup.

1.1 Climate Stability

As important as the warming of the globe is to biological systems and human health, the effects of the increased extreme and anomalous weather that accompanies the excess energy in the system may be even more profound. As the rate of warming accelerated after the mid-1970s, anomalies and wide swings away from norms increased, suggesting that feedback, corrective mechanisms in the climate system are being overwhelmed. Indeed, increased variability may presage transitions: Ice core records from the end of the last ice age ($\sim$10,000 years ago) indicate that increased variability was associated with rapid change in state.

Further evidence for instability comes from the world ocean. Although the ocean has warmed overall in the past century, a region of the North Atlantic has cooled in the past several decades. Several aspects of global warming are apparently contributing.

Recent warming in the Northern Hemisphere has melted much North Polar ice. Since the 1970s, the floating North Polar ice cap has thinned by almost half. A second source of cold fresh water comes from Greenland, where continental ice is melting at higher elevations each year. Some meltwater is trickling down through crevasses, lubricating the base, accelerating ice "rivers," and increasing the potential for sudden slippage. A third source of cold fresh water is rain at high latitudes. Overall ocean warming speeds up the water cycle, increasing evaporation. The warmed atmosphere can also hold and transport more water vapor from low to high latitudes. Water falling over land is enhancing discharge from five major Siberian rivers into the Arctic, and water falling directly over the ocean adds more fresh water to the surface.

The cold, freshened waters of the North Atlantic accelerate transatlantic winds, and this may be one factor driving frigid fronts down the eastern U.S. seaboard and across to Europe and Asia in the winters of 2002–2004. The North Atlantic is also where deep-water formation drives thermohaline circulation, the "ocean conveyor belt," considered key to climate stabilization. In the past few years, the northern North Atlantic has freshened, and since the 1950s the deep overflow between Iceland and Scotland has slowed by 20%.

The ice, pollen, and marine fossils reveal that cold reversals have interrupted warming trends in the past. The North Atlantic Ocean can freshen to a point at which the North Atlantic deep-water pump—driven by sinking cold, salty water that is in turn replaced by warm Gulf Stream waters—can suddenly slow. Approximately 13,000 years ago, when the world was emerging from the Last Glacial Maximum and continental ice sheets were thawing, the Gulf Stream abruptly changed course and shot straight across to France. The Northern Hemisphere refroze—for the next 1300 years—before temperatures increased again, in just several years, warming the world to its present state.

Calculations (of orbital cycles) indicate that our hospitable climate regime was not likely to end due to natural causes any time soon. However, the recent buildup of heat-trapping greenhouse gases is forcing the climate system in new ways and into uncharted seas.

1.2 Hydrological Cycle

Warming is also accelerating the hydrological (water) cycle. As heat builds up in the deep ocean, down to 3 km, more water evaporates and sea ice melts. During the past century, droughts have lasted longer and heavy rainfall events (defined as >5 cm/day) have become more frequent. Enhanced evapotranspiration dries out soils in some regions, whereas the warmer atmosphere holds more water vapor, fueling more intense, tropical-like downpours elsewhere. Prolonged droughts and intense

precipitation have been especially punishing for developing nations.

Global warming is not occurring uniformly. It is occurring twice as fast as overall warming during the winter and nighttime, a crucial factor in the biological responses, and the winter warming is occurring faster at high latitudes than near the topics. These changes may be due to greater evaporation and the increased humidity in the troposphere because water vapor is a natural greenhouse gas and can account for up to two-thirds of all the heat trapped in the troposphere. Warming nights and winters along with the intensification of extreme weather events have begun to alter weather patterns that impact the ecological systems essential for regulating the vectors, hosts, and reservoirs of infectious diseases.

Other climate-related health concerns include temperature and mortality, especially the role of increased variability in heat and cold mortality; synergies between climate change and air pollution, including CO_2 fertilization of ragweed and excess pollen production, asthma, and allergies; travel hazards associated with unstable and erratic winter weather; and genetic shifts in arthropods and rodents induced by warming. This article focuses on climate change and emerging infectious diseases.

2. CLIMATE AND INFECTIOUS DISEASE

Climate is a key determinant of health. Climate constrains the range of infectious diseases, whereas weather affects the timing and intensity of outbreaks. A long-term warming trend is encouraging the geographic expansion of several important infections, whereas extreme weather events are spawning "clusters" of disease outbreaks and a series of surprises. Ecological changes and economic inequities strongly influence disease patterns. However, a warming and unstable climate is playing an ever-increasing role in driving the global emergence, resurgence, and redistribution of infectious diseases.

The World Health Organization reports that since 1976, more than 30 diseases have appeared that are new to medicine, including HIV/AIDS, Ebola, Lyme disease, Legionnaires' disease, toxic *Escherichia coli*, a new hantavirus, and a rash of rapidly evolving antibiotic-resistant organisms. Of equal concern is the resurgence of old diseases, such as malaria and cholera. Declines in social conditions and public health programs underlie the rebound of diseases transmitted person to person (e.g., tuberculosis and diphtheria). The resurgence and redistribution of infections involving two or more species—mosquitoes, ticks, deer, birds, rodents, and humans—reflect changing ecological and climatic conditions as well as social changes (e.g., suburban sprawl).

Waves of infectious diseases occur in cycles. Many upsurges crest when populations overwhelm infrastructures or exhaust environmental resources, and sometimes pandemics can cascade across continents. Pandemics, in turn, can affect the subsequent course of history. The Justinian plague emerged out of the ruins of the Roman Empire in the 6th-century AD and arrested urban life for centuries. When the plague, carried by rodents and fleas, reappeared in the repopulated and overflowing urban centers of the 14th century, it provoked protests and helped end feudal labor patterns. In "Hard Times," Charles Dickens describes another transitional period in the overcrowded 19th-century England:

> *Of tall chimneys, out of which interminable serpents of smoke trailed themselves forever and ever, and never got uncoiled … where the piston of the steam-engine worked monotonously up and down like the head of an elephant in a state of melancholy madness.*

These conditions bred smallpox, cholera, and tuberculosis. However, society responded with sanitary and environmental reform, and the epidemics abated. How will our society respond to the current threat to our health and biological safety?

2.1 An Integrated Framework for Climate and Disease

All infections involve an agent (or pathogen), host(s), and the environment. Some pathogens are carried by vectors or require intermediate hosts to complete their life cycle. Climate can influence pathogens, vectors, host defenses, and habitat.

Diseases carried by mosquito vectors are particularly sensitive to meteorological conditions. These relationships were described in the 1920s and quantified in the 1950s. Excessive heat kills mosquitoes. However, within their survivable range, warmer temperatures increase their reproduction and biting activity and the rate at which pathogens mature within them. At 20°C, falciparum malarial protozoa take 26 days to incubate, but at 25°C they develop in 13 days. *Anopheline* mosquitoes, which are carriers of malaria, live only several weeks. Thus, warmer temperatures permit parasites to mature in time for the mosquito to transfer the infection.

Temperature thresholds limit the geographic range of mosquitoes. Transmission of *Anopheline*-borne falciparum malaria occurs where temperatures exceed 16°C. Yellow fever (with a high rate of mortality) and dengue fever (characterized by severe headaches and bone pain, with mortality associated with dengue hemorrhagic fever and dengue shock syndrome) are both carried by *Aedes aegypti*, which is restricted by the 10°C winter isotherm. Freezing kills *Aedes* eggs, larvae, and adults. Thus, given other conditions, such as small water containers, expanding tropical conditions can increase the ranges and extend the season with conditions allowing transmission.

Warm nights and warm winters favor insect survival. Fossils from the end of the last ice age demonstrate that rapid, poleward shifts of insects accompanied warming, especially of winter temperatures. Insects, notably Edith's checkerspot butterflies today, are superb paleothermometers, outpacing the march of grasses, shrubs, trees, and mammals in response to advancing frost lines.

In addition to the direct impacts of warming on insects, volatile weather and warming can disrupt co-evolved relationships among species that help to prevent the spread of "nuisance" species.

3. PEST CONTROL: ONE OF NATURE'S SERVICES

Systems at all scales have self-correcting feedback mechanisms. In animal cells, errors in structural genes (mismatched base pairs) resulting from radiation or chemicals are "spell-checked" by proteins propagated by regulatory genes. Malignant cells that escape primary controls must confront an ensemble of instruments that comprise the immune surveillance system. A suite of messengers and cells also awaits invading pathogens—some that stun them, and others, such as phagocytes, that consume them.

Natural systems have also evolved a set of pheromones and functional groups (e.g., predators, competitors, and recyclers) that regulate the populations of opportunistic organisms. The diversity of processes provides resistance, resilience, and insurance, whereas the mosaics of habitat—stands of trees near farms that harbor birds and nectar-bearing flowers that nourish parasitic wasps—provide generalized defenses against the spread of opportunists. Against the steady background beat of habitat fragmentation, excessive use of toxins, and the loss of stratospheric ozone—all components of global environmental change—climate change is fast becoming a dominant theme, disrupting relationships among predators and prey that prevent the proliferation of pests and pathogens.

4. CLIMATE CHANGE AND BIOLOGICAL RESPONSES

Northern latitude ecosystems are subjected to regularly occurring seasonal changes. However, prolonged extremes and wide fluctuations in weather may overwhelm ecological resilience, just as they may undermine human defenses. Summer droughts depress forest defenses, increasing vulnerability to pest infestations. Also, sequential extremes and shifting seasonal rhythms can alter synchronies among predators, competitors, and prey, releasing opportunists from natural biological controls.

Several aspects of climate change are particularly important to the responses of biological systems. First, global warming is not uniform. Warming is occurring disproportionately at high latitudes, above Earth's surface and during winter and nighttime. Areas of Antarctica, for example, have already warmed over 1°C this century, and the temperatures within the Arctic Circle have warmed 5.5°C in the past 30 years. Since 1950, the Northern Hemispheric spring has occurred earlier and fall later.

Although inadequately studied in the United States, warm winters have been demonstrated to facilitate overwintering and thus northern migration of the ticks that carry tick-borne encephalitis and Lyme disease. Agricultural zones are shifting northward but not as swiftly as are key pests, pathogens, and weeds that, in today's climate, consume 52% of the growing and stored crops worldwide.

An accelerated hydrological cycle—ocean warming, sea ice melting, and rising atmospheric water vapor—is demanding significant adjustments from biological systems. Communities of marine species have shifted. A warmer atmosphere also holds more water vapor (for each 1°C warming 6%), insulates escaping heat, and enhances greenhouse warming. Warming and parching of Earth's surface intensifies the pressure gradients that draw in winds (e.g., winter tornadoes) and large weather systems.

Elevated humidity and lack of nighttime relief during heat waves directly challenge human (and livestock) health. These conditions also favor mosquitoes.

4.1 Range Expansion of Mosquito-Borne Diseases

Today, half of the world's population is exposed to malaria on a daily basis. Deforestation, drug resistance, and inadequate public health measures have all contributed to the recent resurgence. Warming and extreme weather add new stresses. Dynamic models project that the warming accompanying the doubling of atmospheric CO_2 will increase the transmission capacity of mosquitoes in temperate zones, and that the area capable of sustaining transmission will increase from that containing 45% of to world's population to that containing 60%, although recent statistical modeling projects less of a change. All these analyses rely on average temperatures, rather than the more rapid changes in minimum temperatures being observed, and thus may underestimate the biological responses.

In addition, historical approaches to understanding the role of temperature and infectious disease have argued that the relationships do not hold for periods such as the medieval warm period and the little ice age. It is important to note, however, that the changes in CO_2 and temperature, and their rates of change, during the 20th century are outside the bounds of those observed during the entire Holocene (the past 10,000 years).

Some of these projected changes may be under way. Since 1976, several vector-borne diseases (VBDs) have reappeared in temperate regions. *Anopheline* mosquitoes have long been present in North America and malaria circulated in the United States in the early 20th century. However, by the 1980s, transmission was limited to California after mosquito control programs were instituted. Since 1990, small outbreaks of locally transmitted malaria have occurred during hot spells in Texas, Georgia, Florida, Michigan, New Jersey, New York, and Toronto. Malaria has returned to South Korea, areas of southern Europe, and the former Soviet Union. Malaria has recolonized the Indian Ocean coastal province of South Africa, and dengue fever has spread southward into northern Australia and Argentina.

These changes are consistent with climate projections, although land clearing, population movements, and drug and pesticide resistance for malaria control have all played parts. However, a set of changes occurring in tropical highland regions are internally consistent and indicative of long-term warming.

4.2 Climate Change in Montane Regions

In the 19th century, European colonists sought refuge from lowland "mal arias" by settling in the highlands of Africa. These regions are now getting warmer. Since 1970, the height at which freezing occurs (the freezing isotherm) has increased approximately 160 m within the tropical belts, equivalent to almost 1°C warming. These measurements are drawn from released weather balloons and satellites.

Plants are migrating to higher elevations in the European Alps, Alaska, the U.S. Sierra Nevada, and New Zealand. This is a sensitive gauge because a plant shifting upward 500 m would have to move 300 km northward to adjust to the same degree of global warming.

Insects and insect-borne diseases are being reported at high elevations in eastern and central Africa, Latin America, and Asia. Malaria is circulating in highland urban centers, such as Nairobi, and rural highland areas, such as those of Papua New Guinea. *Aedes aegypti*, once limited by temperature to approximately 1000 m in elevation, has recently been found at 1700 m in Mexico and 2200 m in the Colombian Andes.

These insect and botanical trends, indicative of gradual, systematic warming, have been accompanied by the hardest of data: the accelerating retreat of summit glaciers in Argentina, Peru, Alaska, Iceland, Norway, the Swiss Alps, Kenya, the Himalayas, Indonesia, Irian Jaya, and New Zealand. Glaciers in the Peruvian Andes, which retreated 4 m annually in the 1960s and 1970s, were melting 30 m per year by the mid-1990s and 155 m per year by 2000. Many small ice fields may soon disappear, jeopardizing regional water supplies critical for human consumption, agriculture, and hydropower.

Highlands, where the biological, glacial, and isotherm changes are especially apparent, are sensitive sentinel sites for monitoring the long-term impacts of climate change.

4.3 Extreme Weather Events and Epidemics

Although warming encourages the spread of infectious diseases, extreme weather events are having the most profound impacts on public health and society. The study of variability also provides insights into the stability of the climate regime.

A shift in temperature norms alters the variance about the means, and high-resolution ice core

records suggest that greater variance from climate norms indicates sensitivity to rapid change. Today, the enhanced hydrological cycle is changing the intensity, distribution, and timing of extreme weather events. Large-scale weather patterns have shifted. Warming of the Eurasian land surface, for example, has apparently intensified the monsoons that are strongly associated with mosquito- and water-borne diseases in India and Bangladesh. The U.S. southwest monsoons may also have shifted, with implications for disease patterns in that region.

Extremes can be hazardous for health. Prolonged droughts fuel fires, releasing respiratory pollutants. Floods foster fungi, such as the house mold *Stachybotrys atra*, which may be associated with an emerging hemorrhagic lung disease among children. Floods leave mosquito-breeding sites and also flush pathogens, nutrients, and pollutants into waterways, precipitating water-borne diseases (e.g., *Cryptosporidium*).

Runoff of nutrients from flooding can also trigger harmful algal blooms along coastlines that can be toxic to birds, mammals, fish, and humans; generate hypoxic "dead zones"; and harbor pathogens such as cholera.

4.4 The El Niño/Southern Oscillation

The El Niño/Southern Oscillation (ENSO) is one of Earth's coupled ocean–atmospheric systems that apparently helps to stabilize the climate system by undulating between states every 4 or 5 years. ENSO events are accompanied by weather anomalies that are strongly associated with disease outbreaks over time and with spatial clusters of mosquito-, water-, and rodent-borne illnesses. The ENSO cycle also affects the production of plant pollens, which are directly boosted by CO_2 fertilization, a finding that warrants further investigation as a possible contributor to the dramatic increase in asthma since the 1980s.

Other climate modes contribute to regional weather patterns. The North Atlantic Oscillation is a seesaw in sea surface temperatures (SSTs) and sea level pressures that governs windstorm activity across Europe. Warm SSTs in the Indian Ocean (that have caused bleaching of more than 80% of regional coral reefs) also contribute to precipitation in eastern Africa. A warm Indian Ocean added moisture to the rains drenching the Horn of Africa in 1997–1998 that spawned costly epidemics of cholera, mosquito-borne Rift Valley fever, and malaria, and warm SSTs catalyzed the southern African deluge in February 2000.

Weather extremes, especially intense precipitation, have been especially severe for developing nations, and the aftershocks ripple through economies. Hurricane Mitch, nourished by a warmed Caribbean, stalled over Central America in November 1998 for 3 days, dumping precipitation that killed more than 11,000 people and caused more than $5 billion in damages. In the aftermath, Honduras reported 30,000 cases of cholera, 30,000 cases of malaria, and 1000 cases of dengue fever. The following year, Venezuela suffered a similar fate, followed by malaria and dengue fever. In February 2000, torrential rains and a cyclone inundated large areas of southern Africa. Floods in Mozambique killed hundreds, displaced hundreds of thousands, and spread malaria, typhoid, and cholera.

Developed nations have also begun to experience more severe and unpredictable weather patterns. In September 1999, Hurricane Floyd in North Carolina afforded an abrupt and devastating end to an extended summer drought. Prolonged droughts and heat waves are also afflicting areas of Europe. In the summer of 2003, five European nations experienced about 35,000 heat-associated deaths, plus wildfires and extensive crop failures. Extreme weather events are having long-lasting ecological and economic impacts on a growing cohort of nations, affecting infrastructure, trade, travel, and tourism.

The 1990s was a decade of extremes, each year marked by El Niño or La Niña (cold) conditions. After 1976, the pace, intensity, and duration of ENSO events quickened, and extremes became more extreme. Accumulating heat in the oceans intensifies weather anomalies; it may be modifying the natural ENSO mode. Understanding how the various climate modes are influenced by human activities, and how the modes interact, is a central scientific challenge, and the results will inform multiple sectors of society. Disasters such as the $10 billion European heat waves and Hurricane Floyd suggest that the costs of climate change will be borne by all.

5. SEQUENTIAL EXTREMES AND SURPRISES

5.1 Hantavirus Pulmonary Syndrome

Extremes followed by subsequent extremes are particularly destabilizing for biological and physical

systems. Light rains followed by prolonged droughts can lead to wildfires, and warm winter rains followed by cold snaps beget ice storms. Warm winters also create snowpack instability, setting the stage for avalanches, which are triggered by heavy snowfall, freezing rain, or strong winds.

Polar researchers suspect that melting at the base of the Greenland ice sheet may be sculpting fault lines that could diminish its stability. Shrinking of Earth's ice cover (cryosphere) has implications for water (agriculture, hydropower) and for albedo [reflectivity] that influences climate stability.

The U.S. Institute of Medicine report of 1992 on emerging infectious diseases warned that conditions in the United States were ripe for the emergence of a new disease. What it did not foresee was that climate was to play a significant role in the emergence and spread of two diseases in North America: the hantavirus pulmonary syndrome in the Southwest and the West Nile virus in New York City.

5.1.1 Hantavirus Pulmonary Syndrome, U.S. Southwest, 1993

Prolonged drought in California and the U.S. southwest from 1987 to 1992 reportedly reduced predators of rodents: raptors (owls, eagles, prairie falcons, red-tailed hawks, and kestrels), coyotes, and snakes. When drought yielded to intense rains in 1993 (the year of the Mississippi floods), grasshoppers and piñon nuts on which rodents feed flourished. The effect was synergistic, boosting mice populations more than 10-fold, leading to the emergence of a "new," lethal, rodent-borne viral disease: the hantavirus pulmonary syndrome (HPS). The virus may have already been present but dormant. Alterations in food supplies, predation pressure, and habitat provoked by sequential extremes multiplied the rodent reservoir hosts and amplified viral transmission.

Controlled experiments with rabbits demonstrate such synergies in population dynamics. Exclusion of predators with protective cages doubles their populations. With extra food, hare density triples. With both interventions, populations increase 11-fold.

By summer's end, predators apparently returned (indicating retained ecosystem resilience) and the rodent populations and disease outbreak abated. Subsequent episodes of HPS in the United States have been limited, perhaps aided by early warnings. However, HPS has appeared in Latin America, and there is evidence of person-to-person transmission.

6. CASE STUDY: WEST NILE VIRUS

West Nile virus (WNV) was first reported in Uganda in 1937. WNV is a zoonosis, with "spillover" to humans, which also poses significant risks for wildlife, zoo, and domestic animal populations. Although it is not known how WNV entered the Western Hemisphere in 1999, anomalous weather conditions may have helped amplify this *Flavivirus* that circulates among urban mosquitoes, birds, and mammals. Analysis of weather patterns coincident with a series of U.S. urban outbreaks of St. Louis encephalitis (SLE) (a disease with a similar life cycle) and four recent large outbreaks of WNV revealed that drought was a common feature. *Culex pipiens*, the primary mosquito vector (carrier) for WNV, thrives in city storm drains and catch basins, especially in the organically rich water that forms during drought and the accompanying warm temperatures. Because the potential risks from pesticides for disease control must be weighed against the health risks of the disease, an early warning system of conditions conducive to amplification of the enzootic cycle could help initiate timely preventive measures and potentially limit chemical interventions.

6.1 Background on WNV

WNV entered the Western Hemisphere in 1999, possibly via migratory or imported birds from Europe. Although the precise means of introduction is not known, experience with a similar virus, SLE, as well as the European outbreaks of WNV during the 1990s suggests that certain climatic conditions are conducive to outbreaks of this disease. Evidence suggests that mild winters, coupled with prolonged droughts and heat waves, amplify WNV and SLE, which cycle among urban mosquitoes (*Culex pipiens*), birds, and humans.

SLE and WNV are transmitted by mosquitoes to birds and other animals, with occasional spillover to humans. *Culex pipiens* typically breeds in organically rich standing water in city drains and catch basins as well as unused pools and tires. During a drought, these pools become even richer in the organic material that *Culex* needs to thrive. Excessive rainfall flushes the drains and dilutes the pools. Drought conditions may also lead to a decline in the number of mosquito predators, such as amphibians and dragonflies, and encourage birds to congregate around shrinking water sites, where the virus can circulate more easily. In addition, high temperatures accelerate

the extrinsic incubation period (period of maturation) of viruses (and parasites) within mosquito carriers. Thus, warm temperatures enhance the potential for transmission and dissemination. Together, these factors increase the possibility that infectious virus levels will build up in birds and mosquitoes living in close proximity to human beings.

6.2 Outbreaks of St. Louis Encephalitis in the United States

SLE first emerged in the city of St. Louis in 1933, during the dust bowl era. Since 1933, there have been 24 urban outbreaks of SLE in the United States. SLE as an appropriate surrogate for study because of its similarity to WNV and because of the significant number of SLE outbreaks in the United States, along with accurate weather data (i.e., the Palmer Severity Drought Index [PSDI], a measure of dryness that is a function of precipitation and soil moisture compared with 30 years of data in the same location). The PSDI ranges from −4 (dry) to +4 (wet). From 1933 to the mid-1970s, 10 of the 12 urban SLE outbreaks—regionally clustered in Kentucky, Colorado, Texas, Indiana, Tennessee, and Illinois—were associated with 2 months of drought. (One of the remaining outbreaks was associated with 1 month of drought.) After the mid-1970s, the relationship shifted and outbreaks were associated with anomalous conditions that included droughts and heavy rains.

Note that outbreaks of SLE during the 1974–1976 period and after show a variable pattern in relation to weather. Once established in a region, summer rains may boost populations of *Aedes japonicus* and other *Aedes* spp. that function as "bridge vectors," efficiently carrying virus from birds to humans. The roles of "maintenance" (primarily bird-biting mosquitoes) and bridge vectors in WNV transmission are under study.

6.3 Outbreaks of WNV

6.3.1 Romania, 1996

A significant European outbreak of WNV occurred in 1996 in Romania, in the Danube Valley and in Bucharest. This episode, with hundreds experiencing neurological disease and 17 fatalities, occurred between July and October and coincided with a prolonged drought (May–October) and excessive heat (May–July). Human cases in Bucharest were concentrated in blockhouses situated over an aging sewage system in which *C. pipiens* were breeding in abundance.

6.3.2 Russia, 1999

A large outbreak of WNV occurred in Russia in the summer of 1999 following a drought. Hospitals in the Volgograd region admitted 826 patients; 84 had meningoencephalitis, of which 40 died.

6.3.3 United States, 1999

In the spring and summer of 1999, a severe drought (following a mild winter) affected the northeastern and mid-Atlantic states. The prolonged drought culminated in a 3-week July heat wave that enveloped the Northeast. Then the pendulum swung in the opposite direction, bringing torrential rains at the end of August (and, later, Hurricane Floyd to the mid-Atlantic states). *Culex* spp. thrived in the drought months; *Aedes* spp. bred in the late summer floodwaters. In the New York outbreak, 7 people died, and, of the 62 people who suffered neurological symptoms and survived, the majority reported chronic disabilities, such as extreme muscle weakness and fatigue.

6.3.4 Israel, 2000

WNV was first reported in Israel in 1951, and sporadic outbreaks followed. Israel, a major stopover for migrating birds, usually receives little precipitation from May to October. In 2000, the region was especially dry as drought conditions prevailed across southern Europe and the Middle East, from Spain to Afghanistan. Between August 1 and October 31, 2000, 417 cases of serologically confirmed WNV were diagnosed in Israel, and there were 35 deaths. *Culex pipiens* was identified as a vector.

6.3.5 United States, 2002

In 2002, much of the western and midwestern United States experienced a severe spring and summer drought. Lack of snowpack in the Rockies (warming winters leading to more winter precipitation falling as rain) contributed. Forest fires burned more than 7.3 million acres, and haze and respiratory disease affected several Colorado cities. There was also an explosion of WNV cases, with human or animal WNV being documented in 44 states and the District of Columbia, reaching to California. Drought conditions were present in June in Louisiana, the first epicenter of WNV in 2002. Widespread drought conditions and heat waves may have amplified WNV and contributed to its rapid spread throughout the continental United States.

(Health officials have also become convinced that WNV can be transmitted via organ transplant and blood transfusion.)

Of greatest concern, however, WNV spread to 230 species of animals, including 138 species of birds and 37 species of mosquitoes. Not all animals fall ill from WNV, but the list of hosts and reservoirs includes dogs, cats, squirrels, bats, chipmunks, skunks, rabbits, and reptiles. Raptors (e.g., owls and kestrels) have been particularly affected; WNV likely caused thousands of birds of prey to die in Ohio and other states in July 2002. Some zoo animals have died.

Note that the population impacts on wildlife and biodiversity have not been adequately evaluated. The impacts of the decline in birds of prey could ripple through ecological systems and food chains and could contribute to the emergence of disease.

Declines in raptors could have dramatic consequences for human health. These birds of prey are our guardians because they prey on rodents and keep their numbers in check. When rodent populations explode—when floods follow droughts, forests are clear-cut, or diseases affect predators—their legions can become prolific transporters of pathogens, including Lyme disease, leptospirosis, plague, hantaviruses and arenaviruses such as Lassa fever and Guaranito, Junin, Machupo, and Sabia, associated with severe hemorrhagic fevers in humans.

As of March 12, 2003, the Centers for Disease Control and Prevention reported the following for 2002:

Laboratory confirmed human cases nationally: 4156
WNV-related deaths: 284
Most deaths: Illinois (64), Michigan (51), Ohio (31), and Louisiana (25)

There were also 14,045 equine cases in 38 states reported to the U.S. Department of Agriculture APHIS by state health officials as of November 26, 2002. WNV has been associated with illness and death in several other mammal species, including squirrel, wolf, and dog in Illinois and mountain goat and sheep. The largest number of equine cases was reported in Nebraska. Because of the bird and mammal reservoirs for WNV, there is the potential for outbreaks in all eastern and Gulf States and into Canada in the future.

6.3.6 United States, 2003

In the summer of 2003, cases of WNV concentrated in Colorado, the Dakotas, Nebraska, Texas, and Wyoming—areas that experienced prolonged spring drought (and extensive summer wildfires) in association with anomalous conditions in the Pacific Ocean. The eastern part of the United States (where a cold, snowy winter occurred in association with the North Atlantic freshening and North Atlantic High along with continued warming of tropical waters) experienced a relatively calm summer/fall in relation to WNV. (Both the Pacific and the Atlantic Oceans were in anomalous states beginning in the late 1990s. The state of the Pacific created perfect conditions for drought in many areas of the world.)

6.4 Public Health Implications

Multimonth drought, especially in spring and early summer, was found to be associated with urban SLE outbreaks from its initial appearance in 1933 through 1973 and with recent severe urban outbreaks of WNV in Europe and the United States. Each new outbreak requires introduction or reintroduction of the virus, primarily via birds or wildlife, so there have been seasons without SLE outbreaks despite multimonth drought. Spread of WNV and sporadic cases may occur, even in the absence of conditions amplifying the enzootic cycling. In Bayesian parlance, drought increases the "prior probability" of a significant outbreak once the virus becomes established in a region. Other factors, such as late summer rains that increase populations of bridge vectors, may affect transmission dynamics.

Further investigation and modeling are needed to determine the role of meteorological factors and identify reservoirs, overwintering patterns, and the susceptibility of different species associated with WNV. The migration path of many eastern U.S. birds extends from Canada across the Gulf of Mexico to South America, and WNV has spread to Mexico, Central America, and the Caribbean.

Factors other than weather and climate contribute to outbreaks of these two diseases. Antiquated urban drainage systems leave more fetid pools in which mosquitoes can breed, abandoned tires and swimming pools are ideal breeding sites, and stagnant rivers and streams do not adequately support healthy fish populations to consume mosquito larvae. Such environmental vulnerabilities present opportunities for environmentally based public health interventions following early warnings of conducive meteorological conditions.

State plans to prevent the spread of and contain WNV have three components:

1. Mosquito surveillance and monitoring of dead birds

2. Source (breeding site) reduction though larviciding (*Bacillus sphaericus* and Altocid or methoprene) and neighborhood cleanups
3. Pesticide (synthetic pyrethrins) spraying, when deemed necessary.

The information covering predisposing climatic conditions and predictions of them may be most applicable for areas that have not yet experienced WNV but lie in the flyway from Canada to the Gulf of Mexico. Projections of droughts (e.g., for northeast Brazil during an El Niño event) could help focus attention on these areas, enhancing surveillance efforts (including active bird surveillance), public communication, and environmentally friendly, public health interventions. They may also help set the stage for earlier chemical interventions once circulating virus is detected.

Finally, in terms of the public perception and concerns regarding the risks of chemical interventions, understanding the links of WNV to climatic factors and mobilizing public agencies, such as water and sewage departments, to address a public health threat may prove helpful in garnering public support for the combined set of activities needed to protect public health.

The WNV may have changed because it took an unusually high toll on birds in the United States. Alternatively, North American birds were sensitive because they were immunologically naive. However, the unexpected outbreak of a mosquito-borne disease in New York City and rapid spread throughout the nation in 2002 also serve as a reminder of the potential for exponential spread of pests and pathogens, and that pathogens evolving anywhere on the globe—and the social and environmental conditions that contribute to these changes—can affect us all.

7. DISCONTINUITIES

Climate change may not prove to be a linear process. Polar ice is thinning and Greenland ice is retreating, and since 1976 several small stepwise adjustments appear to have reset the climate system. In 1976, Pacific Ocean temperatures warmed significantly; they warmed further in 1990 and cooled in 2000. The intensity of ENSO has surpassed the intensity it had 130,000 years ago during the previous warm interglacial period. Cold upwelling in the Pacific Ocean in 2000 could portend a multidecadal correction that stores accumulating heat at intermediate ocean layers. Meanwhile, two decades of warming in the North Atlantic have melted Arctic ice, plausibly contributing to a cold tongue from Labrador to Europe and enhancing the Labrador Current that hugs the U.S. east coast. Such paradoxical cooling from warming and ice melting could alter projections for climate, weather, and disease for northern Europe and the northeastern United States. It is the instability of weather patterns that is of most concern for public health and society.

Winter is a blessing for public health in temperate zones, and deep cold snaps could freeze *C. pipiens* in New York City sewers, for example, reducing the risk of WNV during those cold winters. Thus, the greatest threat of climate change lies not with year-to-year fluctuations but with the potential for a more significant abrupt change that would alter the life-support systems underlying our overall health and well-being.

8. CONCLUSION

The resurgence of infectious diseases among humans, wildlife, livestock, crops, forests, and marine life in the final quarter of the 20th century may be viewed as a primary symptom of global environmental and social change. Moreover, contemporaneous changes in greenhouse gas concentrations, ozone levels, the cryosphere, ocean temperatures, land use, and land cover challenge the stability of our epoch, the Holocene—a remarkably stable 10,000-year period that followed the retreat of ice sheets from temperate zones. The impacts of deforestation and climatic volatility are a particularly potent combination creating conditions conducive to disease emergence and spread. Given the rate of changes in local and global conditions, we must expect synergies and new surprises.

Warming may herald some positive health outcomes. High temperatures in some regions may reduce snails, the intermediate hosts for schistosomiasis. Winter mortality in the Northern Hemisphere from respiratory disease may decline. However, the consequences of warming and wide swings in weather are projected to overshadow the potential health benefits.

The aggregate of air pollution from burning fossil fuels and felling forests provides a relentless destabilizing force on the earth's heat budget. Examining the full life cycle of fossil fuels also exposes layers of damages. Environmental damage from their mining, refining, and transport must be added to direct health effects of air pollution and acid precipitation.

Returning CO_2 to the atmosphere through their combustion reverses the biological process by which plants drew down atmospheric carbon and generated oxygen and stratospheric ozone, helping to cool and shield the planet sufficiently to support animal life.

9. NEXT STEPS

Solutions may be divided into three levels. First-order solutions to the resurgence of infectious disease include improved surveillance and response capability, drug and vaccine development, and greater provision of clinical care and public health services.

Second is improved prediction. Integrating health surveillance into long-term terrestrial and marine monitoring programs—ecological epidemiology—can benefit from advances in satellite imaging and climate forecasts that complement fieldwork. Health early warning systems based on the integrated mapping of conditions, consequences, and costs can facilitate timely, environmentally friendly public health interventions and inform policies.

Anticipating the health risks posed by the extreme conditions facing the U.S. East Coast in the summer of 1999 could have (a) enhanced mosquito surveillance, (b) heightened sensitivity to bird mortalities (that began in early August), and (c) allowed treatment of mosquito breeding sites, obviating large-scale spraying of pesticides.

The third level is prevention, which rests on environmental and energy policies. Restoration of forests and wetlands, "nature's sponges and kidneys," is necessary to reduce vulnerabilities to climate and weather. Population stabilization is also necessary, but World Bank data demonstrate that this is a function of income distribution.

Development underlies most aspects of health and it is essential to develop clean energy sources. Providing basic public health infrastructure—sanitation, housing, food, refrigeration, and cooking—requires energy. Clean energy is needed to pump and purify water and desalinate water for irrigation from the rising seas. Meeting energy needs with nonpolluting sources can be the first step toward the rational use of Earth's finite resources and reduction in the generation of wastes, the central components of sustainable development.

Addressing all these levels will require resources. Just as funds for technology development were necessary to settle the Montreal Protocol on ozone-depleting chemicals, substantial financial incentives are needed to propel clean energy technologies into the global marketplace. International funds are also needed to support common resources, such as fisheries, and for vaccines and medications for diseases lacking lucrative markets.

Human and ecological systems can heal after time-limited assaults, and the climate system can also be restabilized, but only if the tempo of destabilizing factors is reduced. The Intergovernmental Panel on Climate Change calculates that stabilizing atmospheric concentrations of greenhouse gases requires a 60% reduction in emissions.

Worldviews can also shift abruptly. Just as we may be underestimating the true costs of "business as usual," we may be vastly underestimating the economic opportunities afforded by the energy transition. A distributed system of nonpolluting energy sources can help reverse the mounting environmental assaults on public health and can provide the scaffolding on which to build clean, equitable, and healthy development in the 21st century.

SEE ALSO THE FOLLOWING ARTICLES

Air Pollution, Health Effects of • *Climate Change and Energy, Overview* • *Climate Change: Impact on the Demand for Energy* • *Climate Protection and Energy Policy* • *Development and Energy, Overview* • *Energy Efficiency and Climate Change* • *Environmental Injustices of Energy Facilities* • *Gasoline Additives and Public Health* • *Greenhouse Gas Emissions from Energy Systems, Comparison and Overview*

Further Reading

Albritton, D. L., Allen, M. R., Baede, A. P. M., *et al.* (2001). "IPCC Working Group I Summary for Policy Makers, Third Assessment Report: Climate Change 2001: The Scientific Basis." Cambridge Univ. Press, Cambridge, U.K.

Daszak, P., Cunningham, A. A., and Hyatt, A. D. (2000). Emerging infectious diseases of wildlife—Threats to biodiversity and human health. *Science* **287,** 443–449.

Epstein, P. R. (1999). Climate and health. *Science* **285,** 347–348.

Epstein, P. R., Diaz, H. F., Elias, S., Grabherr, G., Graham, N. E., Martens, W. J. M., Mosley-Thompson, E., and Susskind, E. J. (1998). Biological and physical signs of climate change: Focus on mosquito-borne disease. *Bull. Am. Meteorol. Soc.* **78,** 409–417.

Harvell, C. D., Kim, K., Burkholder, J. M., Colwell, R. R., Epstein, P. R., Grimes, J., Hofmann, E. E., Lipp, E., Osterhaus, A. D. M. E., Overstreet, R., Porter, J. W., Smith, G. W., and Vasta, G. (1999). Diseases in the ocean: Emerging pathogens, climate links, and anthropogenic factors. *Science* **285,** 1505–1510.

Karl, T. R., and Trenbath, K. (2003). Modern climate change. *Science* **302,** 1719–1723.

Levitus, S., Antonov, J. I., Boyer, T. P., and Stephens, C. (2000). Warming of the world ocean. *Science* **287,** 2225–2229.

McCarthy, J. J., Canziani, O. F., Leary, N. A., Dokken, D. J., and White, K. S. (eds.). (2001). "Climate Change 2001: Impacts, Adaptation & Vulnerability Contribution of Working Group II to the Third Assessment Report of the Intergovernmental Panel on Climate Change (IPCC)." Cambridge Univ. Press, Cambridge, UK.

McMichael, A. J., Haines, A., Slooff, R., and Kovats, S. (eds.). (1996). "Climate Change and Human Health." World Health Organization, World Meteorological Organization, and United Nations Environmental Program, Geneva, Switzerland.

National Research Council, National Academy of Sciences (2001). "Abrupt Climate Change: Inevitable Surprises." National Academy Press, Washington, DC.

Petit, J. R., Jouze, J., Raynaud, D., Barkov, N. I., Barnola, J. M., Basile, I., Bender, M., Chapellaz, J., Davis, M., Delaygue, G., Delmotte, M., Kotlyakov, V. M., Legrand, M., Lipenkov, V. Y., Lorius, C., Peplin, L., Ritz, C., Saltzman, E., and Stievenard, M. (1999). Climate and atmospheric history of the past 420,000 years from the Vostok Ice Core, Antartica. *Nature* **399,** 429–436.

Rosenzweig, C., Iglesias, A., Yang, X. B., Epstein, P. R., and Chivian, E. (2001). Climate change and extreme weather events: Implications for food production, plant diseases, and pests. *Global Change Hum. Health* **2,** 90–104.

Climate Change: Impact on the Demand for Energy

TIMOTHY J. CONSIDINE
The Pennsylvania State University
University Park, Pennsylvania, United States

1. Impact of Climate Change on Energy Demand
2. Recent Climate Trends
3. Climate Impacts on Energy Demand
4. Sensitivity of Energy Demand to Temperature Changes
5. Numerical Simulation of Climate Impacts
6. Conclusion

Glossary

cooling degree days This measure is the maximum of zero or the difference between the average daily temperature at a given location and 65°F. For example, if average daily temperature is 85°, there are 20 heating degree days.

elasticity The proportionate change in one variable in response to a proportionate change in a causal factor. For example, an own-price elasticity of demand equal to −0.10 indicates that a 1% change in price reduces consumption by 0.1%.

heating degree days This measure is the maximum of zero or the difference between 65°F and the average daily temperature at a given location. For example, if average daily temperature is 35°, there are 30 heating degree days.

Although the focus of many policy studies of climate change is on establishing the causal links between anthropogenic systems, emissions of greenhouse gases, and climate, the line of causation also runs the other way. Short-term fluctuations in climate conditions, particularly in the temperate zones on the planet, affect energy consumption. If the popular expectation that the climate will become warmer becomes a reality, we can expect winters and summers that are warmer than those of the past. Warmer summers induce people to use more air-conditioning, which requires more electricity and greater consumption of fuels used to generate that electric power. On the other hand, warmer winters reduce heating requirements and the associated consumption of natural gas, distillate fuel oil, propane, and other heating fuels. The net effect of these two opposing forces on total annual energy demand depends on a number of factors, including local climate conditions, technology, and the behavioral response of consumers to weather conditions.

1. IMPACT OF CLIMATE CHANGE ON ENERGY DEMAND

Establishing the impact of climate change on energy demand requires a measure of heating and cooling requirements. In the United States, this measure is a degree day, which is defined in terms of an absolute difference between average daily temperature and 65°F, which is an arbitrary benchmark for household comfort. Heating degree days are incurred when outside temperatures are below 65°F, generally during the winter heating season from October through March. Cooling degree days occur when temperatures exceed 65°F, often during the summer cooling season from April through September. Degree days are intended to represent the latent demand for heating and air-conditioning services based on temperature departures from the comfort level of 65°F.

Heating and cooling degree days are reported by the National Oceanic and Atmospheric Administration at a weekly frequency for 50 cities throughout the United States. State, regional, and national population weighted averages of heating and cooling degree days are also reported and commonly used by the utility industry and energy commodity trading community.

Clearly, a higher number of degree days—heating, cooling, or both—should be associated with greater

 © 2004 Elsevier Inc. All rights reserved.

fuel consumption. Estimating how energy consumption rises for every degree day requires some form of model that links weather conditions to energy consumption. There are two basic modeling approaches. One approach uses engineering models that estimate energy consumption by multiplying a utilization rate for an energy-consuming durable by a fixed energy efficiency rate. For example, to compute the reduction in space heating energy from a warmer winter, an engineering approach would involve estimating the impact on the utilization rate and then multiplying the fixed unit energy efficiency by this new utilization rate. The main advantage of this approach is its high level of detail, although the accuracy is often in doubt given the limited data available to compute key parameters, such as utilization rates and unit energy efficiencies, particularly for older equipment.

Another approach develops statistical relationships often specified by economists on the basis of behavioral models that assume energy consumers, such as households and businesses, maximize their satisfaction conditional on heating and cooling degree days, relative fuel prices, income or sales levels, and the technical characteristics of energy-consuming durable equipment. Although this approach is more abstract than engineering models, it captures the behavioral dimensions of energy demand. Econometric models of energy demand are able to estimate the separate effects of price, income, weather variations, and other factors affecting energy demand.

The sensitivities of energy demand to price, weather, and other factors are expressed in terms of elasticities, defined as the percentage change in energy consumption associated with a given percentage change in one of these causal factors. The term associated is used to convey that the relationship is estimated based on a statistical model, which by definition allows a random error that reflects a number of uncertainties stemming from the inherent vagaries of human behavior. Such elasticities are a critical component in models of economic growth that are used to estimate the net social benefits of policies to mitigate or control greenhouse gas emissions. If a warmer climate is likely and if this results in a net reduction in energy demand, then the net social cost of greenhouse gas pollution would be lower by the energy expenditure savings. Determining the relevance and size of this potential effect has broad implications for climate change policies. Moreover, the feedback of climate change on energy demand also indirectly affects carbon emissions. Over the long term, this feedback effect may be an important element in understanding the global carbon cycle. In addition, future climate change agreements may need to allow some variance in carbon emissions due to normal climate fluctuations.

Many utilities and government agencies recognize the link between weather conditions and energy demand. Detecting the effects of climate trends on weather conditions is not without controversy, and the following section explains some of the nuances of detecting trends in heating and cooling degree days. Next, the formulation and construction of econometric models of energy demand that incorporate weather conditions are described in more detail. Estimates of price, income, and weather elasticities from these models are described along with numerical simulations of the impacts of climate change on U.S. energy demand. Finally, a discussion of the policy implications of the findings is presented.

2. RECENT CLIMATE TRENDS

Climatologists generally agree that there is accumulating evidence that a warming trend has been occurring since the mid-1960s. There are rather distinct seasonal and regional variations in the lower 48 states of the United States. Livezey and Smith determined that the average national warming trend has been 0.015°F per year. Since 1964, this implies that average annual temperature has increased by approximately one-half of one degree.

Nevertheless, there are some dissenting views, such as that of Balling, who argues that temperature has not changed in any meaningful way since 1932. His analysis tests for linear trends in temperature and finds no statistically significant trends in the data. Nonlinear trends, however, are possible and the evidence presented in Figs. 1 and 2 suggests that this indeed is the case. The smooth lines are predicted temperatures from a simple nonlinear statistical trend model in which degree days are specified as a function of a constant, the time period, and the time period squared. The hypothesis that these latter two terms are individually or jointly zero could not be rejected. The probability levels, or the chances that this hypothesis is wrong, are zero for cooling degree days and slightly less than 2% for the heating degree day trend terms. Notice that the trend for cooling degree days is downward from 1932 to approximately 1970 and then upward since then. Heating degree days display the inverse pattern, increasing during the early part of the sample but then decreasing during the 1980s and 1990s.

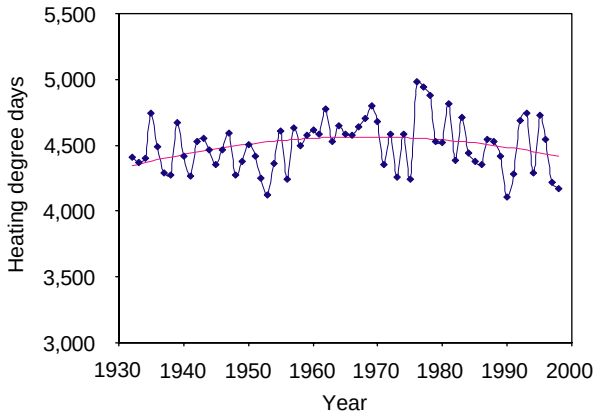


FIGURE 1 Actual and fitted nonlinear trend in heating degree days.

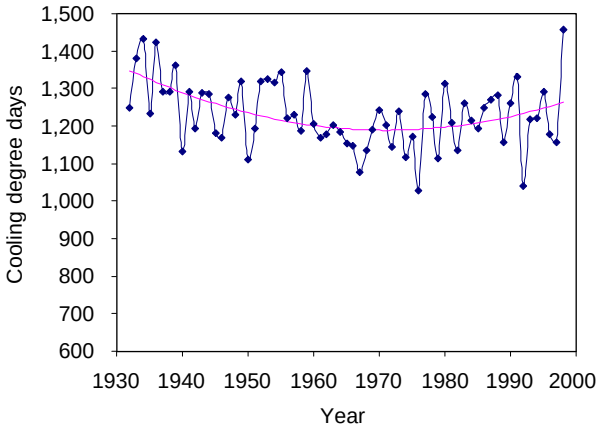


FIGURE 2 Actual and fitted nonlinear trend in cooling degree days.

The highly volatile nature of temperature as illustrated in Figs. 1 and 2 is cause for some concern because statistical models with a relatively small number of observations are prone to be heavily influenced or skewed by extreme observations, such as the severe cold during the winter of 1976–1977. Far more observations are required to place a high degree of confidence in detecting linear or nonlinear trends in temperature. With these caveats in mind, temperature trends during the past 30 years are consistent with the global warming hypothesis.

Morris found that climate trends vary by region and by season. Most regions of the eastern United States had average or slightly below average temperatures. In contrast, the western and Great Plains states experienced warmer temperatures. The western and upper Great Plains regions experienced the most warming during the winter months. Average temperatures during the fall are actually lower for most of the eastern and Midwest regions of the United States. This finding suggests that studies linking energy demand to temperature or degree days should capture some of these regional and seasonal patterns. The analysis presented next explicitly captures the seasonal dimensions. Although a regional energy demand model is not developed here, the findings of Enterline suggest that regional aggregation is not a serious problem.

3. CLIMATE IMPACTS ON ENERGY DEMAND

Energy is consumed by various segments of the economy, including households, commercial establishments, manufacturing enterprises, and electric power generators. Only a portion of total energy demand is sensitive to temperature. Consumption of gasoline and jet fuel has strong seasonal components but is not very sensitive to heating and cooling degree days. On the other hand, consumption of electricity, natural gas, propane, and heating oil is quite sensitive to weather conditions, particularly temperature. In addition, the derived demand for primary fuels used by electric utilities, such as coal, natural gas, and residual fuel oil, is also sensitive to weather because the demand for electricity changes with variations in temperature and other atmospheric conditions.

The challenge for understanding the importance of climate on energy demand is to identify and estimate the separate effects that the determinants of energy demand have on energy consumption. The demand for energy depends on a number of factors, including the price of energy relative to prices for other goods and services, household income and demographic features, output and technology adoption by businesses, and weather conditions. Using history as a guide, the challenge is to determine to what extent trends in energy consumption are influenced by trends in these basic drivers of energy demand.

Considine develops energy demand models that estimate the separate effects that these factors have on energy consumption trends. These models are derived from economic principles and then tested using econometric methods. For example, businesses are assumed to minimize energy expenditures subject to exogenously determined constraints, such as output levels, fuel prices, and weather conditions.

This optimization model provides a basis to formulate a set of energy demand equations that

relate energy consumption and these exogenous factors. Moreover, this approach allows a consistent representation of substitution possibilities among fuels. For instance, if the demand for natural gas increases with an increase in fuel oil prices, suggesting that natural gas and fuel oil are substitutes, then the demand for fuel oil would increase with higher prices for natural gas. This consistency between demand equations is known as symmetry and follows from the mathematical conditions associated with economic optimization.

Another important property of demand systems is that if all fuel prices increase the same percentage, then total energy expenditures would increase the same amount and relative fuel use would not change. This property, known as zero-degree homogeneity in demand, essentially means that only relative prices matter in determining energy demand.

Given these considerations regarding the role of prices in determining energy demand, identifying the effect of climate on energy demand requires measurement of two components of the climate signal: the fixed seasonal effect and a random weather shock. Fixed seasonal effects include the length of day and other fixed monthly effects and can be represented by zero-one or dummy variables. For instance, household energy consumption in the United States may have a January effect, increasing a fixed amount each year as day length or consumer habits change in fixed seasonal ways. Energy demand during other months of the year may have similar fixed monthly or seasonal effects. The second effect, associated with random weather shocks, represents that portion of energy consumption directly associated with departures from normal temperatures. One good measure for this effect is heating and cooling degree day deviations from their 30-year means. The monthly deviation of degree days from its corresponding monthly mean measures the extent to which temperature is above or below normal. Therefore, for example, heating degree day deviations 30° below normal would imply a proportionately larger increase in energy demand than heating degree day deviations 10° below normal. This approach is similar to the study of natural gas demand by Berndt and Watkins.

The demand models developed by Considine assume that monthly energy consumption depends on economic forces, technological factors, fixed seasonal effects, and random weather shocks. Energy demand systems are estimated for four sectors of the U.S. economy: residential, commercial, industrial, and electric utilities. A two-step modeling approach is adopted. The first step determines the level of energy consumption in each sector, and the second determines the mix of fuels to satisfy that total demand.

In the first step, Considine specifies that aggregate energy demand is a simple log-linear function of real energy price, income or output, fixed seasonal effects, heating and cooling degree days, and a time trend as a proxy for exogenous technological change. This relationship is also dynamic so that shocks to demand during one period affect predictions of demand in future periods.

The second step, determining the fuel mix, assumes that the fuels in each sector's aggregate energy price index are weakly separable from other factors of production. This means that the marginal rate of substitution between two fuels is independent of the rate at which aggregate energy substitutes with other goods or factors of production. This specification is reasonable because substitution possibilities between energy and other factors of production are likely to be very limited over a month time span.

The combinations of fuels used in each sector are modeled using energy expenditure systems specified from a logistic function, which ensures adding-up and positive cost shares. Considine and Mount show how to impose constraints consistent with economic theory. Given the nature of the model, however, Considine imposes these conditions locally—either at a point with linear parameter restrictions or at each point in the sample using an iterative estimation procedure. The results presented next impose these conditions at the mean cost shares, which simplifies model simulation for policy analysis and forecasting.

4. SENSITIVITY OF ENERGY DEMAND TO TEMPERATURE CHANGES

Here, the sensitivity of energy demand to climate is measured two ways. The first method uses elasticities that provide simple summary measures of how departures from normal temperatures affect energy consumption. The second approach, reported in the following section, uses econometric simulation to estimate how climate changes affect energy demand.

The heating and cooling degree day elasticities of energy demand by sector are shown in Table I. The elasticities are interpreted as the percentage change in fuel consumption for a percentage change in heating or cooling degree days. For instance, total U.S. residential natural gas consumption increases 0.33%

TABLE I

Heating and Cooling Degree Day Elasticities of Energy Demand (with Probability Values)

	Residential	Commercial	Industrial	Electric utility
		Heating degree days		
Natural gas	0.333	0.296	0.114	0.05
Distillate oil	0.262	0.277	0.906	
Residual fuel oil			0.388	0.606
Coal			0.287	0.106
Electricity	0.148	0.08	0.055	
		Cooling degree days		
Natural gas	−0.022	0.038	−0.055	0.154
Distillate oil	−0.065	0.046	0.439	
Residual fuel oil			0.233	0.355
Coal			0.026	0.071
Electricity	0.141	0.059	−0.032	

for every 1% increase in heating degree days. With the exception of natural gas in the electric utility sector, all heating degree day elasticities are larger than the cooling degree day elasticities. In other words, heating degree days have a proportionately greater impact on energy consumption than cooling degree days. Overall, this finding suggests that global warming would reduce energy consumption because higher fuel use associated with hotter summers is offset by lower fuel use due to warmer winters.

This trade-off is clearly illustrated by the heating and cooling degree day elasticities for the residential sector reported in Table I. In addition to the previously natural gas elasticity, the heating degree day elasticities for distillate fuel oil and electricity are 0.262 and 0.148, respectively. In contrast, the cooling degree day elasticities for natural gas, distillate oil, and electricity are −0.022, −0.065, and 0.141, respectively. Higher cooling degree days reduce residential natural gas and heating oil consumption due to the reduced heating requirements during the spring and fall. Note that the heating and cooling degree day elasticities for electricity in the residential sector are approximately the same, suggesting that winter peaks from demands for electrical resistance heating are as sensitive to summer peak air-conditioning requirements. The commercial sector provides a similar set of elasticities, except that the cooling degree day elasticities are positive but extremely small for natural gas and distillate fuel oil.

Industrial demand for distillate and residual fuel oil is quite sensitive to temperature. For instance, for every percent change in heating degree days, industrial distillate use increases 0.9%. Residual fuel oil and coal demand increases with temperature, increasing 0.39 and 0.29%, respectively, for each percent change in heating degree days. Natural gas consumption is the next most temperature-sensitive fuel in the industrial sector, with a heating degree day eleasticity of 0.11. Electricity consumption is the least sensitive among the industrial fuels.

The demand for primary fuels in electric power generation is also quite sensitive to temperature. Like the other sectors, heating degree day elasticities are greater than the cooling degree day elasticities. Residual fuel oil consumption increases 0.66 and 0.36% for every 1% change in heating and cooling degree days, respectively. The natural gas cooling degree day elasticity is larger than the heating oil counterpart, which reflects the relatively greater use of natural gas to meet summer cooling demands. Note that the degree day elasticities assume that the demand for electricity is fixed. To capture the induced impact of degree days on primary fuel demand by electric utilities, a full model simulation is required that links the demand for electricity in the residential, commercial, and industrial sectors with electric power generation and fuel use.

5. NUMERICAL SIMULATION OF CLIMATE IMPACTS

To determine the effects of past climate trends on energy consumption, the econometric equations providing the degree day elasticities reported previously are combined into an econometric simulation model. The endogenous variables determined by the model include energy demand in the residential, commercial, and industrial sectors of the U.S. economy and the derived demand for primary fuels used in electric power generation. The demand and supply of electricity are determined in the model. The exogenous or predetermined variables include real personal disposable income, retail sales, industrial production, energy prices, and heating and cooling degree days. Like the degree day elasticities reported previously, the equations provide estimates of price and income elasticities of demand, which were reported by Considine in 2000. Hence, the model provides a tool for estimating the contributions of each determinant of energy demand, including climate, prices, and income.

To identify the effects of past climate trends on historical energy consumption, two simulations are required. The first simulation provides a baseline using the 30-year means for heating and cooling degree days, which gives an estimate of what energy demand would have been under average climate conditions. The second simulation uses actual degree days, which yields an estimate of predicted energy demand associated with actual weather. All other exogenous variables, such as energy prices and income, are the same in the two simulations. Consequently, the changes from the base simulation reported in Table II represent those changes in energy demands associated with deviations of weather conditions from their 30-year means.

As discussed previously, the climate in North America has been getting warmer since the early 1980s. Indeed, cumulative cooling degree days are 2.2% higher than normal and heating degree days are 1.65% lower. The impacts of these changes in

TABLE II

Historical Impacts of Weather on Energy Demand in Percentage Changes from Predictions Using Average Weather

		Fuel consumption					
Year	Cooling degree days	Electricity	Natural gas	Distillate	Residual	Coal	Propane
		Cooling season, April–September					
1983	6.57	0.79	1.15	2.03	6.01	1.35	1.16
1984	0.36	0.18	0.48	0.92	1.33	0.38	0.71
1985	−2.22	−0.47	−0.97	−1.26	−3.03	−0.94	−1.00
1986	3.58	0.10	−0.48	−0.28	0.59	0.14	−0.61
1987	8.23	0.31	−0.48	0.11	2.36	0.69	−0.93
1988	8.96	0.80	0.33	0.66	4.84	1.44	−0.14
1989	−3.91	−0.30	0.00	−0.17	−1.51	−0.56	0.16
1990	3.43	0.23	0.03	0.37	1.68	0.50	−0.05
1991	11.11	0.50	−0.63	0.21	3.56	1.10	−1.09
1992	−13.69	−1.03	−0.20	−0.97	−6.12	−2.18	0.31
1993	3.15	0.52	0.52	0.39	2.84	0.90	0.46
1994	1.56	−0.02	−0.36	−0.30	0.48	0.00	−0.41
1995	8.20	1.07	0.97	1.56	6.21	2.14	0.71
1996	0.16	0.06	0.23	0.36	0.49	0.20	0.31
1997	−7.82	−0.16	0.97	0.54	−1.22	−0.36	1.52
		Fuel consumption					
Year	Heating degree days	Electricity	Natural gas	Distillate	Residual	Coal	Propane
		Heating season, October–March					
1983–1984	3.16	0.17	0.88	2.18	2.43	0.60	1.64
1984–1985	−1.08	−0.04	0.00	0.33	−0.16	−0.04	−0.24
1985–1986	−3.24	−0.18	−1.15	−1.41	−1.28	−0.40	−1.20
1986–1987	−3.72	−0.41	−1.44	−2.04	−2.85	−0.79	−1.70
1987–1988	−0.18	−0.13	−0.22	−0.10	−0.88	−0.23	0.02
1988–1989	−1.06	0.02	−0.80	−0.99	−1.05	−0.05	−0.71
1989–1990	−4.55	−0.47	−1.86	−1.48	−3.01	−0.70	−1.50
1990–1991	−8.92	−1.01	−2.93	−3.65	−6.40	−2.00	−3.90
1991–1992	−7.88	−1.10	−3.12	−3.80	−7.01	−2.08	−3.66
1992–1993	2.05	0.27	0.40	0.64	1.32	0.48	0.53
1993–1994	4.63	0.64	1.64	2.54	3.84	1.21	2.21
1994–1995	−8.44	−1.23	−3.08	−3.53	−6.87	−2.48	−3.88
1995–1996	3.09	0.38	1.05	1.22	2.53	0.90	1.29
1996–1997	−3.62	−0.33	−1.63	−1.36	−2.66	−0.80	−1.63

temperature are illustrated in Table II, which presents the percentage changes in energy demand from the predictions using average and actual degree days during the cooling and heating seasons.

There is an unambiguous, positive relationship between heating degree days and energy demand (Table II). With warmer winters, the demand for energy declines. Notice the string of warm winters during the late 1980s and early 1990s and the associated reductions in energy consumption. For example, the winters of 1990–1991 and 1991–1992 were approximately 8% warmer than average. As a result, the consumption of natural gas was approximately 3% lower than it would have been under normal weather conditions. Distillate and residual fuel oil use was almost 4% and more than 6% lower than normal, respectively.

The effects of cooling degree days are more ambiguous because some months of the cooling season also contain heating degree days, which may offset the effects associated with cooling degree days. Nevertheless, there are several years with warmer than normal summers and slightly higher energy use. During the entire period, energy demand under the base simulation with actual weather is 0.2% lower than that under the alternative simulation using the 30-year mean. This suggests that a warmer climate would slightly lower energy consumption because lower fuel use associated with reduced heating requirements offsets higher fuel consumption to meet increased cooling demands.

Simulating the effects of a 3-month shock to heating and cooling degree days provides a more controlled experiment. First, consider the effects of a colder than normal winter with heating degree days 10% more than the 30-year mean for 3 months. All fuel demands increase (Fig. 3). Higher residential and commercial consumption is the principal reason for the approximately 7–10% increase in natural gas consumption. In addition, electric utility fuel use, particularly oil, increases due to the simulated 2–4% increase in electricity consumption. The large estimated elasticity of electric utility oil use with respect to electricity generation accounts for most of the increase in residual fuel consumption. This result may reflect the practice of electric utilities to use oil-fired generators to service peak power demands.

The last simulation estimates the effects of a 3-month, 10% increase in cooling degree days. The impacts on fuel demands and carbon emissions are shown in Fig. 4. As expected, electricity demand increases, leading to an increase in the demand for coal and oil. Natural gas consumption increases slightly due primarily to higher consumption in the electric utility sectors. Carbon emissions increase 1.5–3.7%.

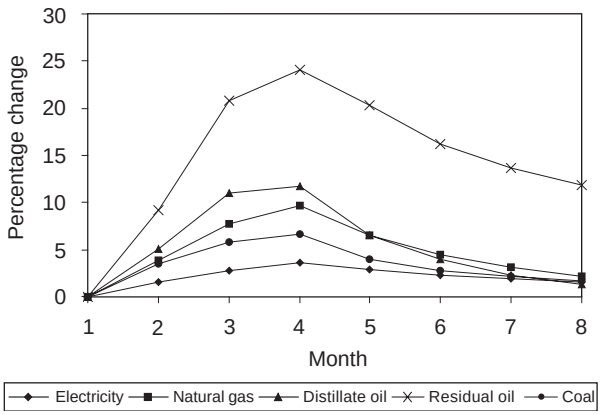


FIGURE 3 Simulated effects of a 10% colder than normal winter on fuel consumption.

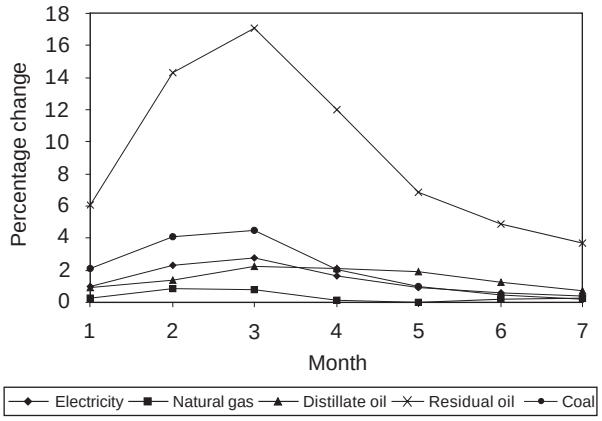


FIGURE 4 Simulated effects of a 10% warmer than normal winter on fuel consumption.

6. CONCLUSION

Overall, warmer climate conditions on balance slightly reduce U.S. energy demand. The estimated impact, however, is very small and pales in comparison to the 7% reduction from 1990 emission levels for the United States under the Kyoto Protocol. Short-term variations in temperature can have sizeable impacts on energy demand. Petroleum consumption is particularly sensitive to temperature, primarily through the induced effect of weather on electricity demand and the fuels in electric

power generation. These short-term weather-related variations in energy demand suggest that any future international agreements on carbon emissions trading should have provisions that adjust compliance rules based on the variance in energy consumption due to weather fluctuations.

SEE ALSO THE FOLLOWING ARTICLES

Carbon Taxes and Climate Change • Climate Change and Energy, Overview • Climate Protection and Energy Policy • Energy Efficiency and Climate Change • Modeling Energy Markets and Climate Change Policy

Further Reading

Balling, R. C. (1999). Analysis of trends in United States degree days, 1950–1995. http://www.greeningearthsociety.org/Articles/1999/trends.htm.

Berndt, E. R., and Watkins, G. C. (1977). Demand for natural gas: Residential and commercial markets in Ontario and British Columbia. *Can. J. Econ.* **10,** 97–111.

Considine, T. J. (1990). Symmetry constraints and variable returns to scale in logit models. *J. Business Econ. Statistics* **8**(3), 347–353.

Considine, T. J. (2000). The impacts of weather variations on energy demand and carbon emissions. *Resource Energy Econ.* **22,** 295–312.

Considine, T. J., and Mount, T. D. (1984). The use of linear logit models for dynamic input demand systems. *Rev. Econ. Statistics* **66,** 434–443.

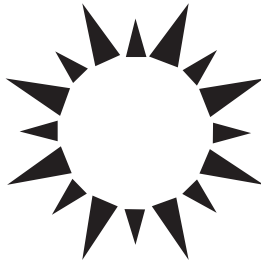
Enterline, S. (1999). The effect of weather and seasonal cycles on the demand for electricity and natural gas, M.S. thesis, Department of Energy, Environmental, and Mineral Economics, Pennsylvania State University, University Park.

Hansen, J., Sato, M., Glascoe, J., and Ruedy, R. (1998). A common sense climate index: Is climate changing noticeably? *Proc. Natl. Acad. Sci. USA* **95,** 4118–4120.

Livezey, R.E., and Smith, T. (1999). Covariability of aspects of North American climate with global sea surface temperatures on interannual and interdecadal timescales. *J. Climate*, January, 289–302.

Morris, M. (1999). The impact of temperature trends on short-term energy demand. http://www.eia.doe.gov/emeu/steo/pub/special/weather/temptrnd.html.

Climate Protection and Energy Policy



WILLIAM R. MOOMAW
The Fletcher School of Law and Diplomacy, Tufts University
Medford, Massachusetts, United States

1. **Impact of Climate Change**
2. **Strategies for Energy Sources**
3. **Energy Strategies for Addressing Climate Change**
4. **Energy for Buildings**
5. **Energy for Industry**
6. **Energy for Transportation**
7. **Energy for Agriculture**
8. **Energy for Electric Power**
9. **Alternative and Integrated Strategies**
10. **Implications for the Future**

Glossary

climate The 30- to 40-year average of weather measurements, such as average seasonal and annual temperature; day and night temperatures; daily highs and lows; precipitation averages and variability; drought frequency and intensity; wind velocity and direction; humidity; solar intensity on the ground and its variability due to cloudiness or pollution; and storm type, frequency, and intensity.

fossil fuels Energy-rich substances that contain carbon, hydrogen, and smaller amounts of other materials that have formed from long-buried plants. Fossil fuels, which include petroleum, coal, and natural gas, provide 85% of the energy that powers modern industrial society.

greenhouse effect The trapping of heat by greenhouse gases that allow incoming solar radiation to pass through the earth's atmosphere but prevent most of the outgoing infrared radiation from the surface and lower atmosphere from escaping into outer space. This process occurs naturally and has kept the earth's temperature approximately 33°C warmer than it would be otherwise.

greenhouse gases Any gas that absorbs infrared radiation in the atmosphere. Greenhouse gases include water vapor, carbon dioxide, methane, nitrous oxide, ozone, hydrochlorofluorocarbons, perfluorocarbons, and hydrofluorocarbons. Whereas the first three gases occur naturally and are increased by human activity, the latter gases are primarily synthesized by humans.

radiation Energy that is transmitted through space in the form of electromagnetic waves and their associated photons. Examples include visible light, ultraviolet light, and infrared light; also called radiant heat.

radiative forcing A measure of heat trapping by gasses and particles in the atmosphere (watts per square meter) that is directly proportional to the temperature increase at the earth's surface.

The earth's climate is determined by the interaction between the radiation received from the sun and the distribution and transformation of that energy by the atmosphere, oceans, land-based (terrestrial) ecosystems, and ice and snow and also by the differential absorption and reflection of solar energy by clouds, land, water, and snow. The one constant of climate over millions of years is its variability. The geological record reveals that at various times the earth has been either warmer or colder, and sometimes wetter or dryer, than it is now. The reason for the intense interest in climate and climate change is that there is evidence that the release of heat-trapping greenhouse gases from the energy sector, industry, and land use changes is directly and indirectly altering the earth's climate system. This article describes the technologies, policies, and measures that are proposed and that are being implemented to mitigate climate change for the energy sector.

1. IMPACT OF CLIMATE CHANGE

Since 1990, every 5 years a group of approximately 2000 natural and physical scientists, economists,

 © 2004 Elsevier Inc. All rights reserved.

social scientists, and technologists assemble under the auspices of the United Nations-sponsored Intergovernmental Panel on Climate Change (IPCC). These scientists spend 3 years reviewing all of the information on climate change and produce a voluminous report following a public review by others in the scientific community and by governments. "Climate Change 2001" is 2665 pages long and contains more than 10,000 references. Information for this article is based on this report, the references in it, as well as recent studies to ensure accurate, up-to-date information.

Climate is defined as the 30- to 40-year average of weather measurements, such as average seasonal and annual temperature; day and night temperatures; daily highs and lows; precipitation averages and variability; drought frequency and intensity; wind velocity and direction; humidity; solar intensity on the ground and its variability due to cloudiness or pollution; and storm type, frequency, and intensity.

Understanding the complex planetary processes and their interaction requires the effort of a wide range of scientists from many disciplines. Solar astronomers carefully monitor the intensity of the sun's radiation and its fluxuations. The current average rate at which solar energy strikes the earth is 342 watts per square meter (W/m^2). It is found that $168\,W/m^2$ reaches the earth's surface mostly as visible light: $67\,W/m^2$ is absorbed directly by the atmosphere, $77\,W/m^2$ is reflected by clouds, and $30\,W/m^2$ is reflected from the earth's surface.

A number of trace gases in the atmosphere are transparent to the sun's visible light but trap the radiant heat that the earth's surface attempts to emit back into space. The net effect is that instead of being a frozen ball averaging $-19^\circ C$, Earth is a relatively comfortable $14^\circ C$. This difference of $33^\circ C$ arises from the natural greenhouse effect. Human additions of greenhouse gases appear to have increased the temperature an additional $0.6 \pm 0.2^\circ C$ during the 20th century. The transmission of visible light from the sun and the trapping of radiant heat from the earth by gases in the atmosphere occur in much the same way as the windows of a greenhouse or an automobile raise the temperature by letting visible light in but trap outgoing radiant heat. The analogy is somewhat imperfect since glass also keeps the warm inside air from mixing with the cooler outside air.

The temperature rise at the earth's surface is directly proportional to a measure of heat trapping called radiative forcing. The units are watts per square meter, as for the sun's intensity. The radiative forcing from human additions of carbon dioxide since the beginning of the industrial revolution is $1.46\,W/m^2$. Methane and nitrous oxide additions have provided relative radiative forcings of 0.48 and $0.15\,W/m^2$, respectively. Other gases have individual radiative forcings of less than $0.1\,W/m^2$. The total radiative forcing of all greenhouse gases added by human activity to the atmosphere is estimated to be $2.43\,W/m^2$. This should be compared to the $342\,W/m^2$ that reaches the earth from the sun. Hence, the greenhouse gases added by human activity are equivalent to turning up the sun's power by an extra 0.7%, which is enough to cause the global temperature to increase by the observed $0.6^\circ C$ ($1.1^\circ F$).

Carbon dioxide is removed from the atmosphere by dissolving in the ocean, by green plants through photosynthesis, and through biogeological processes, such as coral reef and marine shell formation. Approximately half of the carbon dioxide released today will remain in the atmosphere for a century, and one-fourth may be present 200 years from now. Some of the synthetic, industrial greenhouse gases have half-lives of hundreds, thousands, and, in at least one case (carbon tetrafluoride), tens of thousands of years. These long residence times are effectively irreversible changes in the composition of the atmosphere and in climate parameters on the timescale of human generations. This is why some argue that the continued release of greenhouse gases is not sustainable since it can adversely affect future generations. Even without complete information about the consequences of increasing greenhouse gases for climate change, many invoke the precautionary principle and urge the world to move toward reducing and eliminating them.

The post-ice age period in which human civilization has evolved is known as the Holocene. It has long been assumed that this represented an extended, stable, and generally warm climate period. Temperatures were slightly higher than those of today at the beginning of the Holocene and gradually declined until recently. Measurements from Greenland have shown that the general changes were subject to sudden shifts in temperature on timescales of a decade or a century. It is believed that these rapid temperature changes occurred as the result of sudden alterations in oceanic and atmospheric circulation.

Direct land and ocean instrumental temperature measurements began in 1861. These records show two periods of global temperature rise, from 1910 to 1945 and from 1976 to the present. In between, from 1946 to 1975, there was little change in global average temperature, although some regions cooled

while others warmed. The total rise for land-based measurements from 1861 to 2000 is estimated by the IPCC to be $0.6^{\circ}C \pm 0.2^{\circ}C$. The warming rate across the entire earth's surface during the two periods has been $0.15^{\circ}C$ per decade. Sea-surface temperatures are increasing at approximately one-third this rate, and it has been found that the heat content of the top 300 m of the ocean increased significantly during the second half of the 20th century. The lower atmosphere up to 8 km has also warmed, but at approximately one-third the land rate. Daily minimum temperatures are rising at $0.2^{\circ}C$ per decade, approximately twice the rate of daily maximum temperatures. This has extended the frost-free season in northern areas, and indeed satellite data confirm that total plant growth is increasing at latitudes above $50^{\circ}N$.

There has been much discussion of whether the rise in the instrumental record during the 20th century is simply a return from a little ice age that occurred between the 15th and 19th centuries to conditions of a medieval warming period that predated it. An analysis of surrogate measurements of temperature utilizing data from tree rings, coral reefs, and ice cores that extends back 1000 years clearly shows that the temperature rise of the 20th century for the Northern Hemisphere lies outside the variability of the period. In fact, the current rate of increase, $0.6^{\circ}C$ per century, reverses a downward trend of $-0.02^{\circ}C$ per century during the previous 900 years. The statistical analysis clearly shows that there is less than a 5% probability that the recent temperature increase is due to random variability in the climate system.

The IPCC concludes that the 20th century was the warmest in the past 1000 years, and the 1990s were the warmest decade in that period. The year 1998 may have been the warmest single year, $0.7^{\circ}C$ above the 40-year average from 1961 to 1990 and nearly $1.1^{\circ}C$ warmer than the projected average temperature extrapolated from the previous 900-year trend.

By the end of the 19th century, scientists had identified the major greenhouse gases in the atmosphere. In 1897, Svente Arhenius, a distinguished Swedish chemist, carried out the first calculation of the effect of greenhouse gases on the temperature of the earth. His estimate, made without a computer or even a good calculator, was surprisingly close to what we know today. Arhenius also performed a thought experiment in which he recognized that the industrial revolution was moving carbon from under the ground in the form of coal and oil and putting it into the atmosphere as carbon dioxide. What would happen, he asked, if human activity were to double the concentration of carbon dioxide in the atmosphere? His answer was approximately twice the upper limit of the range estimated by today's best climate models of $1.5–4.5^{\circ}C$. This is called the climate sensitivity. The actual increase in temperature depends on the amount of greenhouse gas released into the atmosphere. It is estimated that this will amount to approximately a doubling of carbon dioxide equivalent from preindustrial levels some time before 2100.

Systematic measurements of atmospheric composition begun in 1959 by Charles David Keeling in Hawaii have confirmed that carbon dioxide is indeed increasing along with the increase in fossil fuel consumption. Air trapped in glacial ice confirms that carbon dioxide in the atmosphere has risen one-third from 280 parts per million (ppm) in preindustrial times to 373 ppm in 2002.

Human activities are influencing climate in several ways. The most direct is through the release of greenhouse gases.

The greatest share of greenhouse gases comes from the combustion of fossil fuels, such as coal, oil, and natural gas, which release carbon dioxide when burned. Coal releases nearly twice as much carbon dioxide as does natural gas for each unit of heat energy released. Oil releases an amount in between.

Methane released from coal mining constitutes the largest source of anthropogenic methane. Methane releases during natural gas leaks during drilling and transportation are also significant. Long-burning coal mine and peat fires also contribute significant amounts of carbon dioxide, methane, and many air pollutants to the atmosphere.

Land clearance, deforestation, and forest fires release carbon dioxide (23% of total human emissions) and methane.

Land-use change alters the reflectivity or albedo of the land. Deforestation in the northern forests usually increases reflectivity by replacing dark evergreen trees with more white-reflecting snow. In the tropics, less solar energy is absorbed in most cases after forests are removed. Urbanization usually increases the absorption of solar energy. The net effect of land-use change is thought to offset warming by approximately $-0.2 \pm 0.2\ W/m^2$.

Combustion of fossil fuels and clearing of land also contribute dust, particulates, soot, and aerosol droplets. These either reflect sunlight away from the earth or absorb radiation. Some aerosols are also believed to increase the amount of cloud cover.

Agriculture releases carbon dioxide through fossil fuel energy use and from the oxidation of soils. This sector is the major producer of nitrous oxide from the bacterial breakdown of nitrogen fertilizer. Rice culture and livestock are also major producers of methane.

Industrial processes release a variety of greenhouse gases into the atmosphere either during manufacturing or through the use of products.

Air pollution that results in ozone and other gases near the earth's surface traps heat, whereas the depletion of the stratospheric ozone in the upper atmosphere allows more heat to escape to space. The pollution near the earth's surface is the larger effect.

The waste sector releases large quantities of methane from sewage and industrial wastewater, animal feedlots, and landfills.

2. STRATEGIES FOR ENERGY SOURCES

The two primary heat-trapping gases that are associated with energy are carbon dioxide from fossil fuel combustion and methane leakages from natural gas production, transport, and use. Ground-level ozone in smog is also an important secondary pollutant from the energy sector that is a significant heat-trapping gas. As temperatures rise on sunny days, chemical reactions in the atmosphere accelerate to produce more ozone, which in turn traps additional heat.

In the following sections, the sources of energy-related carbon dioxide are summarized by sector. It is customary to report only the carbon content of carbon dioxide that represents 12/44 or 27.3% of the mass of carbon dioxide. World total energy use and carbon dioxide increased rapidly between 1971 and 1995. Energy use increased by 67% to 319 exajoules (10^{18} J), whereas carbon dioxide emissions from energy use increased 54% during this period to 5.5 gigatons of carbon (GtC) (10^9 metric tons). The annual average growth rate between 1990 and 1995 was 0.7% for energy and 1.0% for carbon dioxide. In 1995, industrial countries released 49% of carbon dioxide from energy, with an average annual growth rate of 0.6% per year. Countries with economies in transition (i.e., the former Soviet Union and those of Eastern Europe) were well below their peak of 1990 with just 16% of emissions, an increase of 18% since 1971. Emissions were declining at a rate of 5.1% annually. Developing countries accounted for 35% of emissions, with growth rates ranging from 2.0% for Africa to 6.1% for Asia. There is considerable discussion among governments on responsibility for past contributions of greenhouse gases and the future potential for countries in which fossil fuel use is growing rapidly.

3. ENERGY STRATEGIES FOR ADDRESSING CLIMATE CHANGE

An important insight from Amory Lovins is to recognize that it is not energy we need but, rather, the services that energy can supply in differing forms of heat and work. Hence, in designing technologies, policies, and measures to reduce energy-related greenhouse gases, one must consider not only the present source of emissions and how to reduce them but also alternative ways to deliver the desired energy services.

A simple identity has been developed by Ehrlich and Holdren and extended by Kaya that simplifies this task:

$$\begin{aligned} CO_2\ (\text{tons}) &= (\text{population}) \\ &\times (\text{economic value/capita}) \\ &\times (\text{energy use/economic value}) \\ &\times (CO_2/\text{energy used}). \end{aligned}$$

Since greenhouse gas reduction policies do not address either population or wealth (economic value per capita), the major focus is on the last two factors. If one can either find a way to create large economic value with little energy use (factor 3) or use an energy source that produces little CO_2, then the total amount of CO_2 will be small. Note that using no energy or using a zero CO_2 energy technology drives the emissions to zero. Hence, strategies to control CO_2 focus on technological choice, energy efficiency, and the policies and practices that will promote them.

The strategy for reducing emissions from the energy sector consists of five basic options:

1. Increase the efficiency of providing the energy service by finding a less energy-demanding technology or practice.
2. Provide the energy in a more efficient manner.
3. Shift the technology to one that produces fewer or zero greenhouse gases.
4. Shift the fuel to one that produces lower or zero greenhouse gases.

5. Capture and store carbon dioxide and capture and burn methane for fuel.

In the following sections, we examine the opportunities for reducing climate-altering greenhouse gases by economic sector utilizing these strategies. In addition to being able to provide the appropriate energy services, alternative technologies and practices must be economically affordable and socially and politically acceptable.

4. ENERGY FOR BUILDINGS

Buildings use energy directly for heating and cooling and indirectly through electricity for lighting, appliances, office equipment, food storage and preparation, and motors in elevators, pumps, and ventilating systems. Globally, buildings account for 34.4% of energy and 31.5% of energy-related carbon dioxide emissions. The IPCC found many examples of opportunities for substantially reducing the energy use and carbon dioxide emissions from buildings, often at net negative cost. For example, in the United States, new refrigerators use only one-fourth as much electricity as they did in 1971. Improving windows, insulation, and heat exchangers and reducing air and moisture leaks in the building envelope combined with appropriate siting in an integrated building design have made it possible to lower the heating, cooling, and lighting requirements by an average of 40%, with demonstrated cases of 60–71% for residential and commercial buildings, respectively. The cost of achieving a 40% reduction is estimated to be \$3/GJ (10^9) saved in the United States compared to an average cost for building energy of \$14/GJ. Because of the rapid growth of appliance purchases in developing countries, there is enormous potential for major savings in electricity use at much lower cost than building additional power plants. For example, China is the largest producer, consumer, and exporter of refrigerators and air conditioners.

In the case of lighting, the shift from incandescent lamps, which still dominate because of their low initial cost and simplicity, to direct replacement compact fluorescent lamps lowers electricity use by a factor of four. Although the initial cost of compact fluorescent lamps is substantially higher than that of incandescant lamps, they last 10 times longer, resulting in a life cycle cost saving of a factor of two or more. However, compact fluorescents have not achieved universal acceptance because of their different color balance, less concentrated light for use in spotlighting, and higher initial cost. The impressive energy savings of compact fluorescent lamps may be eclipsed by solid-state lighting, such as light-emitting diodes. Because of their ability to be directed, it is possible to provide a solid-state reading light that uses only 1 W to produce the same illumination on a page as a 75-W incandescent. Light-emitting diodes are replacing traffic lights, producing brighter, easier to see signals with an 85% reduction in electricity and a 10 times longer life than incandescents.

It is estimated by the IPCC that there could be cost-effective savings in the building sector using available technology that would lower carbon dioxide emissions by 27% by 2010, 31% by 2020, and 52% by 2050. It is concluded that policies such as improved building codes will need to be put into place to achieve these cost-effective goals.

5. ENERGY FOR INDUSTRY

The industrial sector is the largest user of energy (41%) and releases the largest amount of energy-related carbon dioxide (42.5%) as well as other greenhouse gases. What is unusual in this sector is that for developed countries energy use has been growing slowly at only 0.9% per year, whereas carbon dioxide releases from all of the Organisation for Economic Cooperation and Development industrial countries combined was 8.6% lower in 1995 than in 1971. Developing country emissions amount to only 36.4% of total industrial carbon dioxide releases. Industrial emissions from countries with economies in transition are down from their peak in 1990 by 28%. The major growth has come from developing countries in Asia, where industrial carbon dioxide emissions now exceed those of the developed industrial countries. The IPCC estimates that there is the potential to reduce carbon dioxide emissions through energy efficiency gains by 10–20% by 2010 and by twice that amount by 2020. An additional 10% reduction could come from more efficient use of materials, including recycling, by 2010 and three times that amount by 2020.

Major industrial facilities and refineries are beginning to build fossil-fueled combined heat and electric power facilities that substantially reduce carbon dioxide emissions by 40–50%. The largest industrial combined heat and power facility in the world is the 1000-MW gas-fired facility run by Dow Chemical Company in the United States. The paper industry of Scandnavia and North America produces

more than half of its steam and electricity from biomass waste, a renewable resource.

Energy required to produce 1 ton of iron or steel is actually less in some developing countries, such as Brazil or Korea, than it is the United Kingdom or the United States, but China and India remain substantially less efficient. Clearly, choosing improved technology can make a major difference as nations industrialize and as the economies in transition reindustrialize. With 75% of its energy being consumed in the industrial sector, it is important that China is improving its energy/gross domestic product intensity by an estimated 4% per year. There were even reports during the late 1990s that Chinese coal consumption and carbon dioxide emissions had actually declined. Japan continues to have the most energy-efficient industrial sector in the world, and industries in countries such as the United States, Canada, and several European nations have tried to match Japanese levels in specific industries.

6. ENERGY FOR TRANSPORTATION

The transport sector consumes just 21.6% of global primary energy and produces 22.0% of fossil fuel carbon dioxide. However, it is the fastest growing sector, with annual growth in carbon dioxide emissions of 2.4% per year. The world automobile fleet has been growing at a rate of 4.5% per year and now exceeds 600 million light vehicles. Unfortunately, the gains in efficiency that characterized the 1980s and early 1990s have begun to reverse.

On the technology side, two Japanese manufacturers have had hybrid gasoline electric cars on the market since 1997, approximately 5 years earlier than was anticipated by the IPCC report of 1996. These vehicles raise fuel economy by 50–100% relative to similar, gasoline-powered cars and provide a substantial clean air bonus as well without sacrificing performance. The second-generation hybrids introduced in the fall of 2003 boast a 22% improvement in fuel economy and a 30% decrease in emissions from already low levels. Europe diesel autos comprise 40% of the market and provide fuel economy gains over gasoline engines in the range of 30% or more. Fuel cell technology has also advanced rapidly: Several manufacturers have announced the introduction of fuel cell-powered vehicles by 2005, approximately 10–20 years earlier than was anticipated only recently. These innovations are expensive, and it remains to be seen if the price targets can be met.

In addition to new engines, there is the potential for major weight reduction with alternative materials. A prototype composite vehicle using a fuel cell has achieved fuel economy of 2.5 liters/100 km (100 miles per gallon), but the cost of a production model is uncertain. It is estimated by IPCC that the potential for carbon dioxide emission reductions from the transport sector is in the range of 10–30% by 2010, with reductions of twice that amount possible by 2020.

The major problem is that the amount of vehicle kilometers traveled continues to increase and outpace any technological gains in efficiency. European governments have kept fuel prices high through taxes on fuel and have negotiated a voluntary, improved fuel economy standard with manufacturers. In contrast, in the United States, the trend is toward larger, less fuel-efficient vehicles and the exploitation of loopholes that undermine the legal fuel economy standards.

The use of vehicle fuels derived from biomass could also reduce net carbon dioxide emissions substantially. Brazil has led the way with its extensive use of ethanol from the sugarcane industry to provide 40% of vehicle fuel mostly through a large fraction blend with gasoline. The Brazilian industry is the most efficient in the world, producing 10–12 times the energy value of alcohol fuel as it comsumes in fossil fuels. In contrast, the U.S. corn-based ehanol industry produces only approximately 1.3 times the energy content of biomass ethanol as it comsumes in fossil fuels. Approximately 140 countries produce sugarcane. Josè Morera of CENBIO proposed that this could become the basis for a major transportation fuel export industry that would boost rural development in poor tropical countries, provide transport fuel domestically, and lower balance of payments problems. Biodiesel from seed crops also has the potential to lower net carbon dioxide emissions and improve air quality by providing a zero-sulfur fuel for trucks and buses. The cost of Brazilian ethanol in recent years has been slightly higher than the price of petroleum-based gasoline, but when oil prices rise it becomes cheaper. The price of biodiesel is approximately twice as high as that of diesel fuel, but it is essentially a small-volume, handcrafted product in the United States and the price could decrease significantly in the future.

In addition to improving technological performance of vehicles, there is potential for alternative transport systems that build on more effective land-use planning, and there are alternatives to transporting people and goods. Low-energy, modern communications technology can substitute for travel to

meetings and may be able to reduce travel for entertainment and other activities.

7. ENERGY FOR AGRICULTURE

Agriculture is responsible for only approximately 4% of energy-related carbon dioxide. Economies in transition are the heaviest users of fossil fuels in agriculture, followed by the developing countries of east Asia and the industrial nations. Other developing countries use very little fossil fuels. Rapid growth rates of 10% or higher in China and Latin America are offset by declines of 5% in economies in transition for an annual global growth rate in carbon dioxide of only 0.6%. Agriculture, however, is responsible for approximately 20% of global warming because of heavy emissions of methane from rice farming, livestock, and animal waste. Nitrous oxide is released from agricultural lands by bacterial denitrification of nitrogen fertilizer and the decomposition of animal waste, accounting for 65–80% of total human emissions of this gas. It is technically feasible to lower emissions of methane and nitrous oxide from the agriculture sector by 10–12% by 2010 and by approximately twice that amount by 2020.

Biogas (methane) from anaerobic manure and crop residue decomposition has long been used in India and China for on-farm fuel production. Recently, this technique has been introduced in Europe and North America as a substitute for natural gas and to assist in the management of animal waste and crop residues. Significant additional quantities of both methane and carbon dioxide are associated with rapid rates of deforestation and land-use change to produce new agricultural lands in developing countries. An unknown amount of soil carbon is also oxidized to carbon dioxide, especially from peat soils when they are exposed to air. Nontilling methods and other land management techniques can simultaneously reduce these sources of greenhouse gases and lower emissions from energy use.

Agriculture can also produce energy crops such as wood or sugar that can be burned directly or gasified or liquefied. Brazil has been at the forefront in converting sugar into ethanol for transport fuel, with the production of 10 units of energy for each unit of fossil fuel input. For U.S. corn production of ethanol for fuel, the conversion efficiency is only approximately 1.5. Brazilian industry representatives have realized that sugar constitutes only approximately one-third of the energy value of sugarcane and efforts are being made to gasify the rest of the plant to run gas turbines for electric power generation. Energy plantations can often utilize land that is unproductive for agricultural crops. The burning of biomass releases about as much carbon dioxide per unit of energy as does coal, but if energy crops are grown sustainably, the growing fuel crop absorbs as much carbon dioxide while growing as it will release when burned.

8. ENERGY FOR ELECTRIC POWER

The electric power sector is heavily dependent on coal and is responsible globally for 2.1 GtC per year, which is 38.2% of global energy and 37.5% of global carbon dioxide energy-related emissions. The emissions from this sector are incorporated into the buildings, industrial, transport, and agricultural sectors in calculating their contributions.

In many ways, electric power offers the largest range of possible technological options for reducing greenhouse gases. However, because of the rapid growth in electricity use, emission-reduction potential is estimated to be just 3–9% by 2010 and possibly two to four times that by 2020.

Efficient gas turbines have become the fossil fuel technology of choice since they achieve more than 50% conversion efficiencies from heat to electricity, have very low initial capital costs, and produce relatively few air pollutants. Natural gas also produces only approximately half the carbon dioxide emissions per unit of heat produced as does coal. Work is proceeding on coal gasification that could make use of this technology in an integrated fashion. Emissions from all fossil fuel power plants could be reduced by recovering a portion of the half to three-fourths of the heat that is simply dumped into the environment by utilizing combined heat and power systems. Fuel cells and microturbines promise a revolution of highly efficient distributed power systems that could change electric power generation by providing both electricity and heat on site with significant reductions in transmission and distribution losses.

Renewable energy also holds significant promise. On a percentage basis, wind turbines represent the fastest growing energy source, growing at a rate of more than 25% per year during the late 1990s; by the end of 2002, wind accounted for approximately 31 GW of installed power or 0.1% of installed electric capacity worldwide. The distribution is very uneven. Denmark currently meets 15% of its electric

needs from wind. Germany is the global leader, with one-third of the world's total wind power, followed by Spain and the United States. India leads the developing countries in part because of Danish foreign assistance and investment. The potential for wind is particularly great in North America, Africa, and China and also in Russia and other economies in transition, and it substantially exceeds the total current use of electricity worldwide. The direct conversion of solar energy to electricity represents a small fraction of annually added energy, but conversion efficiency continues to increase and costs continue to decrease. Promising innovations such as solar roof tiles could enhance the potential for distributed roof-top power systems.

Burning or, even better, gasifying biomass would also substantially reduce net carbon dioxide to the atmosphere since the carbon is returned to growing plants on a sustained basis. Wood and wood waste in the forest products industry and bagass from sugarcane are among the leading biomass fuels used for electricity and steam production. Geothermal systems also make significant contributions in Central America, the Philippines, the United States, and Iceland. Hydropower continues to grow in many developing countries, including China, India, and Brazil, but few major projects are under development elsewhere.

A major effort has begun to explore the use of the hydrogen from fossil fuels and biomass for fuel cells and gas turbines. The carbon dioxide produced when hydrogen is removed could be sequestered in depleted oil and gas fields, injected into underground aquifers, or possibly injected into the deep ocean. This carbon dioxide storage option would greatly extend the use of fossil fuels without adding to the atmospheric loading of carbon dioxide. The economic and technical feasibility is being studied along with possible environmental consequences of carbon storage.

9. ALTERNATIVE AND INTEGRATED STRATEGIES

To stabilize carbon dioxide concentrations at current levels will require a reduction in annual emissions of 75–85%. To return atmospheric concentrations to preindustrial levels will require a shift to an energy system that produces zero emissions. If, in addition, the goal is to fully stabilize atmospheric concentrations of the other 40% of other greenhouse gases that contribute to global warming, it will be necessary to reduce the release of methane, nitrous oxide, and other industrial gases by 75–100%.

There are two alternative approaches for addressing climate change. The first is a gradualist approach that adjusts around the edges of existing technologies to improve efficiency or involves relatively modest shifts in technology and fuels. This approach can slow the rise of carbon dioxide and global warming but cannot by itself halt it. Nevertheless, it can effect significant reductions while awaiting major transformations of the energy sector. In the 2001 IPCC analysis, it was estimated that sufficient technologies currently exist to reduce energy-related carbon dioxide by 15% during each of the first two decades of the 21st century at zero or net negative cost. In other words, investing in these technologies would pay back in energy savings at reasonable social discount rates. An additional reduction of 15% could be achieved at a cost of between zero and $100 per metric tonne of carbon—the equivalent of 25¢ per gallon of gasoline or 6¢ per liter. Other practices, such as minimizing the use of lighting and appliances (turning off the lights, TV, computer, or copy machine, driving vehicles for maximum fuel economy, etc.), can yield substantial additional reductions of 10–50%.

The second strategy is to combine energy efficiency of existing technologies with a concerted effort to replace each existing technology with an alternative that emits little or no carbon dioxide. Such a strategy requires policies that shift the replacement of existing technologies to very low- or zero-emission technologies during the turnover of capital stock or as new additional facilities are built. For example, the light vehicle fleet is largely replaced every 8–12 years in industrial countries. After a new efficient auto technology, such as gasoline– or diesel–electric hybrids or fuel cell vehicles, is introduced, it might take a decade for it to penetrate the market depending on the cost and policies encouraging or mandating the shift. Hence, it might be possible to essentially replace the entire light vehicle fleet in approximately 35 years. For lighting, the change to compact fluorescent lighting and eventually to solid-state lighting could occur within a shorter time of approximately 10 years. Other end-use appliances such as office equipment are replaced on a 3- to 5-year cycle, whereas home appliances such as refrigerators and washers can be replaced every 15 years with devices that are twice as efficient.

Electric power plants have an operational lifetime of 40–50 years, so it would probably take a century to replace them all with some combination of

zero-emitting renewable technologies (wind, solar, biomass, and geothermal), fuel cells, combined heat and power, a new generation of nuclear power, or a comprehensive system of carbon dioxide capture and storage technology. Transforming the electrical system to one of multiple, clean, distributed generators utilizing a combination of smaller combined heat and power turbines and fuel cells plus renewables could reduce current electric power generation emissions in half in less than 50 years. Such a distributed power system could also be much more robust and less likely to be disrupted by accidents, storms, or sabotage by terrorists.

Buildings are less permanent than they appear. Low-cost commercial structures often have a lifetime of 25 years or less, as does some housing. Homes and office buildings typically exceed 50–75 years. However, these are upgraded several times during their lifetime, and the addition of new replacement windows, improved insulation and moisture barriers, and advanced heating and cooling systems and controls can substantially lower energy use by 25–50%. Building rooftops are also ideal locations for solar hot water and photovoltaic electricity systems that can be either part of new construction or retrofitted to existing structures.

Hydrogen represents an energy carrier that can either be burned to provide heat or to fuel a gas turbine or it can react chemically in a fuel cell to produce electricity directly. The advantage of hydrogen is that it produces only water as it releases energy. It can be produced from water, fossil fuels, or biomass or solid waste. Producing hydrogen from any of these sources is very energy intensive. To achieve the benefits of zero carbon dioxide emissions from hydrogen on a life cycle basis will require the capture and storage of any carbon released when fossil fuels are the source. An advantage of obtaining hydrogen from biomass or organic solid waste is that if carbon capture and storage is used, there is a net removal of carbon dioxide from the atmosphere.

Finally, there are high-tech, long-term future scenarios that grip the imagination of some. Carbon dioxide capture and storage will require not only modifying all existing coal- and gas-burning power plants, and many industrial facilities, but also an enormous infrastructure of pipes and storage locations. It will take at least half a century to implement on a sufficient scale to make a major difference, but it does allow the current fossil fuel infrastructure to remain largely intact. Hopes for commercially available fusion power have been dashed numerous times, and it, like solar power satellites, remains in the realm of technological futurism. Costs are unknown but are likely to be quite high. Although such technologies may someday be a reality, waiting for these radical technological transformations when more modest technological changes to reduce greenhouse gases are available in the short to midterm does not appear to be a prudent strategy.

10. IMPLICATIONS FOR THE FUTURE

Reducing energy sector greenhouse gas emissions by the large amounts needed to restrain severe climate disruption requires nothing short of a dramatic shift in the delivery of energy services. Although there are advocates for different technologies, it appears that there is no single option that can provide those services in an acceptable manner at a cost that people are willing to pay. Since it is extremely difficult to pick winning future technology, it appears that a portfolio approach of multiple low- and zero-emission technologies is the most likely to succeed. Improving the efficiency of end-use devices, including vehicles, appliances, motors, and lighting, can not only reduce emissions using conventional fossil fuel energy supplies but also substantially lower the cost and many environmental impacts. Doubling the efficiency of lighting and appliances reduces the cost of zero-emission electricity supply whether photovoltaics, wind, fuel cells, or nuclear power plants provide the electricity. If fossil fuels continue to be utilized and carbon capture and storage is used, improved efficiency will lower the amount of carbon dioxide that must be removed, transported, and stored.

Because of the large current investment in existing capital stock of vehicles, buildings, appliances, and industrial facilities, it is essential to ensure that each generation of replacement technology provides substantially lower emissions. Also, since fossil fuels currently provide 85% of the world's energy, there is a strong "lock-in" effect that is technological, economic, and political. To shift course, major policy initiatives are needed that will require a combination of regulations and economic incentives. Until the cost of climate change and other environmental damage is included in the price of fossil fuels, it will be difficult to effect the revolution that is needed. In developing countries in which capital stock is just beginning to be accumulated, the opportunity exists to introduce low and zero carbon dioxide-emitting technology

from the start of industrialization. Unfortunately, in most cases, governments and industries in these countries continue to opt for older, less efficient energy technology that has a lower initial cost. Companies and governments in industrial countries can best address this through a combination of technological sharing and investments. There are many low-cost opportunities to provide energy services to the people of developing countries with zero or low carbon dioxide emissions. Since climate change is truly global, it is in everyone's interest to reduce emissions throughout the world.

Since the shift to fossil fuels began in earnest more than 150 years ago, it may take a century or more to complete the next industrial revolution. A 2% annual reduction in carbon dioxide and methane emissions will lead to a net reduction in greenhouse gases necessary to stabilize gas concentrations during this century. It will require multiple policies and measures designed to deliver energy services in new ways if the world is to meet this reduction and avoid a dangerous increase in carbon monoxide levels to 550 ppm.

SEE ALSO THE FOLLOWING ARTICLES

Biomass: Impact on Carbon Cycle and Greenhouse Gas Emissions • Carbon Capture and Storage from Fossil Fuel Use • Carbon Sequestration, Terrestrial • Carbon Taxes and Climate Change • Climate Change and Energy, Overview • Climate Change: Impact on the Demand for Energy • Climate Change and Public Health: Emerging Infectious Diseases • Energy Efficiency and Climate Change • Greenhouse Gas Emissions from Energy Systems, Comparison and Overview

Further Reading

American Council for an Energy Efficient Economy Web site. at www.aceee.org.Fletcher School of Law and Diplomacy, Tufts University, Multilaterals treaty Web site at http://www.fletcher.tufts.edu.

Hawkins, P., Lovins, A., and Lovins, H. (1999). "Natural Capitalism: Creating the Next Industrial Revolution." Little, Brown, Boston.

Intergovernmental Panel on Climate Change (2001). "Climate Change 2001." Cambridge Univ. Press, Cambridge, UK.

Intergovernmental Panel on Climate Change Web site at http://www.ipcc.ch.

Romm, J. J. (1999). "Cool Companies." Island Press, Washington, DC.

Sawin, J. (2003). "Charting a New Energy Future State of the World 2003." Norton, New York.

Tata Energy Research Institute, India, Web site as http://www.teriin.org.

Tufts Climate Initiative, Tufts University, Web site at www.tufts.edu.

United Nations Development Program, United Nations Department of Economic and Social Affairs, World Energy Council (2000). "World Energy Assessment: Energy and the Challenge of Sustainability." United Nations Development Program, New York.

U.S. Environmental Protection Agency Energy Star Program Web site at http://www.energystar.gov.

Coal, Chemical and Physical Properties

RICHARD G. LETT
U.S. Department of Energy, National Energy Technology Laboratory
Pittsburgh, Pennsylvania, United States

THOMAS C. RUPPEL
National Energy Technology Laboratory Site Support Contractor, Parsons
South Park, Pennsylvania, United States

1. Introduction
2. Coal Sampling and Analysis
3. Coal Classification
4. Composition
5. Physical Properties
6. Chemical Properties
7. Coal Databases

Glossary

anthracite High-rank coal, usually possessing a bright luster and uniform texture, that is high in carbon content.
aromatic carbon Carbon atoms in unsaturated six carbon ring structures, exemplified by benzene, or in unsaturated poly-condensed ring structures.
ash The solid (oxidized) residue remaining after coal is subjected to specific combustion conditions.
bituminous coal Hard black coal of intermediate rank.
coalification The complex process of chemical and physical changes undergone by plant sediments to convert them to coal.
fixed carbon The solid residue, other than ash, obtained by heating coal under standardized conditions.
lignite/brown coal Nonfriable, low-rank coal, dark brown to nearly black in appearance.
mineral matter The inorganic compounds or minerals associated with a coal as opposed to the organic material.
rank The stage of geologic maturity or degree of coalification of a coal.
volatile matter The material, less moisture, evolved by heating coal in the absence of air under rigidly controlled conditions.

Coal is a sedimentary rock composed primarily of organic material, mineral matter, and water. It is formed from the partially decomposed remains of plant material that was originally deposited in a swampy environment and later covered by other sediments. The original plant material is transformed under primarily anaerobic conditions by a combination of microbial action, pressure, and heat over geological time periods in a process termed coalification. Peat is usually considered to be the precursor to coal.

1. INTRODUCTION

Coal is found widely distributed and occurs in seams layered between other sedimentary rocks. Owing to the diversity of the original organic debris and deposition environments and subsequent metamorphic processes, coal compositions and properties vary widely. Properties may vary considerably not only from seam to seam but also with location within a given seam. Since coal is heterogeneous, distinctions must be made between properties of some small size fraction of a coal of scientific interest and properties of representative bulk coal samples that are used in commercial activities.

The organic portion of coal is heterogeneous, but under microscopic examination certain regions can be grouped by their physical form and reflectivity. Depending on appearance, these are usually placed into one of three general maceral groups: vitrinite, exinite (liptinite), or inertinite. Vitrinite is typically the most prevalent maceral in bituminous coals and originates largely from the metamorphosis of woody tissue. Exinite is composed of resinous and other hydrogen rich materials. Inertinite is the most aromatic maceral with properties similar to charcoal and has the highest carbon and lowest hydrogen content.

 © 2004 Elsevier Inc. All rights reserved.

The properties of coals vary with the degree of geological maturity, a concept embodied in the use of the term coal rank. The rank of a coal is a qualitative indication of the geological maturity or degree of coalification. Low-rank coals usually date from the Tertiary or Mesozoic age. The bituminous coals and anthracite are usually from the Carboniferous age. Although it is convenient to view the coal maturation process as progressing in the orderly sequence of peat to brown coal or lignite, to bituminous coal, and ultimately to anthracite, in actuality this is somewhat of an oversimplification and the ultimate rank depends on the severity of the transformation conditions. The low-rank coals tend to be soft and friable with a dull, earthy appearance and a high oxygen and moisture content. Higher rank coals are usually harder and stronger with a black vitreous luster. An increase in rank is attended by an increase in carbon and energy contents and a decrease in moisture and oxygen functionalities.

2. COAL SAMPLING AND ANALYSIS

Standardized laboratory procedures for coal characterization and evaluation have been developed by the American Society for Testing and Materials (ASTM), the International Organization for Standardization (ISO), and the standardization bodies of other nations. Owing to the heterogeneity of coal, these include specific procedures for collecting representative samples and sample preparation. Two procedures, the proximate analysis (ASTM D3172; ISO/CD 17246) and the ultimate analysis (ASTM D3176, ISO/CD 17247), deserve special mention because of their wide use and applicability. A sulfur forms analysis (ASTM D2492, ISO 157), which determines pyritic and sulfate sulfur and calculates organic sulfur by difference, is often performed in conjunction with the ultimate analysis.

The proximate analysis is the simplest and most common form of coal evaluation and constitutes the basis of many coal purchasing and performance indices. It describes coal in terms of moisture, volatile matter, ash and fixed carbon, which is calculated by difference. The moisture may be determined by establishing the loss in weight when heated under rigidly controlled conditions on the as received sample and on an air-dried sample. For low-rank coals, particularly for classification purposes, an equilibrium moisture may be determined by equilibrating the moisture in a sample under standard temperature and humidity conditions. The volatile matter represents the loss of weight, corrected for moisture, when the coal sample is heated to 950°C in specified apparatus under standard conditions. Ash is the residue remaining after complete combustion of the coal organic material and oxidation of the mineral matter at a specified heating rate to a specified maximum temperature (700°C to 750°C). Results are conventionally presented on an as-received or moisture-free basis. Typical results for a suite of U.S. coals of differing rank are shown in Table I.

The ultimate analysis expresses the composition of coal in terms of the percentage of carbon, hydrogen, nitrogen, sulfur, oxygen, and ash, regardless of their origin. Standard procedures are used for the determination of carbon, hydrogen, nitrogen, sulfur, and ash, and oxygen is calculated by difference. Results can be expressed on several bases, depending on the manner in which moisture, sulfur, and ash are treated. It is common to report the results on an as-received basis, a moisture-free basis, and on a moisture- and ash-free basis. One alternative is

TABLE I

Proximate Analyses of Selected Coal Samples U.S. Department of Energy Coal Sample Bank, Coal and Organic Petrology Laboratories, The Pennsylvania State University.

	Percentage moisture		Percentage ash	Percentage volatile matter	Percentage fixed carbon
Coal	As received	Equilibrium	Moisture-free basis		
Lykens #3 (DECS-21) anthracite	3.99	2.10	11.15	4.51	84.34
Pocahontas #3 (DECS-19) low volatile bituminous	1.04	4.87	4.60	18.31	77.09
Pittsburgh (DECS-23) high volatile C bituminous	2.00	2.50	9.44	39.42	51.14
Illinois #6 (DECS-24) high volatile C bituminous	13.20	11.59	13.39	40.83	45.78
Rosebud (DECS-10) subbituminous	21.58	23.98	12.56	41.67	45.77
Beulah (DECS-11) lignite	33.88	35.22	9.56	56.08	34.36

TABLE II

Ultimate and Sulfur Forms Analyses of Selected Coal Sample U.S. Department of Energy Coal Sample Bank, Coal and Organic Petrology Laboratories, The Pennsylvania State University.

	Percentage carbon	Percentage hydrogen	Percentage nitrogen	Percentage oxygen	Percentage pyritic S	Percentage sulfate S	Percentage organic S
Coal	Moisture and mineral-matter free basis				Moisture-free basis		
Lykens #3 (DECS-21) anthracite	91.53	4.06	0.81	3.60	0.06	0.00	0.43
Pocahontas #3 (DECS-23) 19 low volatile bituminous	90.61	4.94	1.14	3.30	0.21	0.00	0.52
Pittsburgh (DECS-23) high volatile A bituminous	84.64	5.82	1.54	8.00	2.23	0.01	1.63
Illinois #6 (DECS-24) high volatile C bituminous	80.06	5.56	1.38	12.99	2.64	0.25	2.64
Rosebud (DECS-10) subbituminous	79.69	4.30	1.07	14.94	0.68	0.02	0.47
Beulah (DECS-11) lignite	74.10	4.51	1.01	20.37	0.31	0.03	0.39

illustrated by Table II, which presents the total sulfur and sulfur forms analysis separately on a moisture-free basis. Carbon, hydrogen, nitrogen, and oxygen may then be presented on a moisture- and mineral matter-free basis.

3. COAL CLASSIFICATION

Coals are classified to systematically represent and order their properties and to facilitate commercial use. Many systems of classification exist as each nation with domestic coal resources tended to develop its own scheme for its particular applications. However, all coal classification schemes use a measure of rank as one of the parameters to classify coal. No single measured property or empirically determined characteristic is usually sufficient to classify coals.

The ASTM coal classification system (ASTM D388) in Table III is used extensively in North America and many other parts of the world. This classification is applicable to coals that are composed mainly of vitrinite. It divides coals into four classes—lignite, subbituminous, bituminous, and anthracite—and differentiates groups within these classes. The primary parameters used to classify coals are the fixed carbon (dry, mineral matter-free basis) for the higher rank coals and gross calorific value (moist, mineral matter-free basis) for the lower rank coals. The agglomerating character is used to differentiate between adjacent groups in certain classes.

The National Coal Board (NCB) coal classification, last revised in 1964, is still used in the United Kingdom and has many of the features of classification systems developed outside North America. Bituminous and anthracite coals are classified primarily on the basis of volatile matter (dry, mineral matter-free basis) and coking properties. The different classes of coal are designated by a three-figure numerical code. Lignites and brown coals are outside the range of the NCB classification scheme.

An attempt by the Coal Committee of the Economic Commission for Europe to develop an international standard resulted in a rather elaborate scheme for the International Classification of Hard Coals by Type. Hard coal is defined as coal having a calorific value of more than 23.86 MJ/kg on a moisture- and ash-free basis. This system classifies hard coals into nine classes (akin to rank) based on dry, ash-free volatile matter content and moisture, ash-free calorific value. These classes are further divided into groups according to their caking properties. The groups are divided into subgroups according to their coking properties. A three-figure code number is then used to express the coal classification. The first figure indicates the class of the coal, the second figure the group, and the third figure the subgroup. A new international coal classification standard is now being developed under the sponsorship of the International Organization for Standardization.

An International Standard for Classification of Brown Coals and Lignites (ISO 2950) was published

TABLE III

ASTM Classification of Coal by Rank (ASTM D388)[a]

Class/group	Fixed carbon limits (dry, mineral-matter-free basis) percentage		Volatile matter limits (dry, mineral-matter-free basis), percentage		Gross calorific value limits (moist,[b] MJ/kg)		Agglomerating character
	Equal or greater than	Less than	Greater than	Equal or less than	Equal or greater than	Less than	
Anthracitic							
Meta-anthracite	98	—	—	2	—	—	Nonagglomerating
Anthracite	92	98	2	8	—	—	
Semianthracite[c]	86	92	8	14	—	—	
Bituminous							
Low-volatile bituminous coal	78	86	14	22	—	—	Commonly agglomerating[e]
Medium-volatile bituminous coal	69	78	22	31	—	—	
High-volatile A bituminous coal		69	31	—	32.6[d]	—	
High-volatile B bituminous coal	—	—	—	—	30.2[d]	32.6	
High-volatile C bituminous coal	—	—	—	—	26.7	30.2	
					24.4	26.7	Agglomerating[e]
Subbituminous							
Subbituminous A coal	—	—	—	—	24.4	26.7	Nonagglomerating
Subbituminous B coal	—	—	—	—	22.1	24.4	
Subbituminous C coal	—	—	—	—	19.3	22.1	
Lignitic							
Lignite A	—	—	—	—	14.7	19.3	
Lignite B	—	—	—	—	—	14.7	

[a] This classification dose not apply to certain coals, discussed in Section 1 of D388.

[b] Moist refers to coal containing its natural inherent moisture but not including visible water on the surface of the coal.

[c] If agglomerating, classify in low-volatile group of the bituminous class.

[d] Coals having 69% or more fixed carbon on the dry, mineral-matter-free basis shall be classified according to fixed carbon, regardless of gross calorific value.

[e] It is recognized that there may be nonagglomerating varieties in these groups of the bituminous class, and that there are notable exceptions in the high-volatile C bituminous group.

in 1974. This applies to coal having a calorific value of less than 23.86 MJ/kg on a moisture- and ash-free basis. This system defines six classes based on the percentage of total moisture (ash-free run-of-mine coal) and further divided into five groups according to tar yield (dry, ash-free basis). A two-digit code is then applied, the first digit indicating the class and the last digit the group. This system is similar to one published in 1957 by the Coal Committee of the Economic Commission of Europe that used a four-digit code.

4. COMPOSITION

4.1 Coal Structure

The organic portion of coal consists primarily of carbon, hydrogen, and oxygen with lesser amounts of nitrogen and sulfur. Owing to the heterogeneous character of coals, it is not possible to attribute a single unique molecular structure to a coal. Based on solubility, swelling behavior, viscoelastic properties,

and evidence from other physical and chemical analyses, a prevailing view is that the coal matrix is basically a three-dimensionally cross-linked structure. However, coals are not composed of regularly repeating monomeric units, but a complex mixture of structural entities. Adsorbed or embedded but not chemically bound in this matrix is potentially extractable or volatile material of a wide range of molecular structures and molecular weights. Further complications are introduced by the admixture of mineral matter to various degrees with the other coal components.

Coals have significant aromatic character, with the fraction of aromatic carbon increasing with coal rank from 0.4 to 0.5 for low-rank coals to 0.6 to 0.7 for middle rank bituminous coals to over 0.9 for anthracite. However, contrary to some evidence from early studies, more recent studies indicate the average aromatic ring size in coals may be rather modest until the anthracite stage is approached. This information suggests the condensed units in bituminous coals are often no more than four rings with two- and three-ring structures predominating. The aromatic units in low-rank coals tend to be even smaller. These structures may be joined to hydroaromatic groups and cross-linked by methylene, ether, or similar groups. Aliphatic substituents on these structures are usually of short carbon length, predominately methyl or ethyl.

Low-rank coals are rich in oxygen functionalities, particularly carboxylate, phenolic, carbonyl, and furan groups or saturated ether groups alpha or beta to aromatic moieties. Much of the ion-exchangeable hydrogen in the oxygen functionalities of low-rank coals is typically replaced by alkali or alkaline earth cations, particularly calcium. With increasing coal rank there is a general loss of organic oxygen content with the remaining oxygen tending to be of the phenolic or furan type.

Definitive information on the chemical forms of nitrogen in coal is rather meager. Available evidence indicates the nitrogen exists primarily in pyrrolic or pyridinic structures with pyrrolic nitrogen being the most abundant. The fraction of pyridinic nitrogen appears to increase with increasing rank.

The organic sulfur, particularly in higher rank coals, is largely in heterocyclic structures. It is predominately of the thiophenic type. There is indirect evidence for the presence of disulfides and aliphatic sulfides in some low-rank coals based largely on the analysis of select samples of unusual coals of very high sulfur content.

4.2 Extractable Material

Coals contain a substantial amount of material that is not tightly bound to the insoluble organic matrix and extractable by appropriate solvents at mild temperatures ($<200^{\circ}C$) below the onset of significant thermal decomposition. Coal extracts vary considerably in yield and composition, depending on the coal and its preparation, the solvent, and the extraction conditions. Specific solvents may extract as much as 10 to 20% of low- and middle-rank coals; primary amines (pyridine) are particularly good extraction agents for bituminous coals. Extract yields fall rapidly with the high-rank coals; only a few percent of a low volatile bituminous coal may be extractable and even less of anthracite.

Pyridine extracts are a complex mixture of organic components with a broad distribution of molecular weights from a few hundred to more than a thousand irrespective of the parent coal. These extracts are much more amenable to many analysis techniques than the whole coal. Since they appear to have many of the chemical compositional features of the solid organic phase, extracts have often been studied to provide insight into the possible nature of the chemical entities and functionalities that make up the insoluble coal matrix. The material extracted initially or at low temperatures with poorer extraction agents is usually hydrogen rich and, particularly for brown coals and lignites, often contains material derived from resins and waxes that maybe atypical of the constitution of the coal as a whole.

4.3 Moisture

The moisture content of coal is difficult to precisely specify and control because it exists in several forms in coal, yet it is a parameter that is of great interest in determining commercial value and suitability for many applications. The total (as-received) moisture of a coal sample is usually considered to be all of the water that is not chemically combined (e.g., as water of hydration in minerals). This includes moisture on the external surfaces of coal particles as well as water (inherent moisture) condensed in the pore structure of the coal. Standard coal analysis procedures are often referenced to a moist coal basis (i.e., including the moisture remaining in a coal sample after bringing the sample to a standard condition before analysis). Neither the total nor the air-dried moisture contents from the proximate analysis of coal include water that is chemically combined or water associated with minerals. The bed moisture is the

moisture that exists as an integral part of the virgin coal seam and is approximately the amount of water a coal will hold when fully saturated at nearly 100% relative humidity.

The natural bed moisture content of coals generally decreases with increasing coal rank. It can vary from less than 5% in higher rank anthracite and bituminous coals to more that 40% in low-rank lignite and brown coals. In contrast to higher rank coals, brown coals and lignites can lose water very rapidly upon exposure to the atmosphere.

4.4 Mineral Matter

The mineral matter in coal originates from several sources including sedimentary material deposited with the coal, minerals transported by water, and inherent material in the plant precursors. The majority of the mineral matter consists of minerals from one of four groups: aluminosilicates (clays), carbonates, sulfides, and silica (quartz). The most common clays are kaolinite, illite, and mixed-layer illite-montmorillonite. Pyrite is the dominant sulfide mineral in most coals. Sulfates are found in varying amounts depending on the extent of weathering. Likewise, small amounts of elemental sulfur detected in some coals also appear to be the result of pyrite weathering. There is a wide range of mineral compositions for the carbonate minerals in coals depending on the origin. Commonly found carbonate minerals are calcite, siderite, dolomite, and ankerite. The amounts of quartz are also dependent on the origin of the coals. A multitude of accessory minerals can be identified at low concentrations indicative of the particular geological environment of the coal seam.

Owing to the difficulty of directly determining the mineral matter content of coal, it is customary practice to estimate it by calculation using empirical relationships involving more readily determined properties. Often used is the Parr formula, or some variation thereof, that relates the mineral matter to the ash content determined by standard procedures with a correction for pyritic sulfur.

4.5 Trace and Minor Elements

From the geochemical standpoint, the makeup of coal mineral matter is similar to that of the earth's crust, and it contains most of the elements on the periodic table. Although minor and trace element concentrations vary considerably between coals and from sample to sample, the elemental survey analysis in Table IV illustrates the general level of most of these elements found in one small sample of a bituminous coal from West Virginia. The most prevalent inorganically bound minor elements are those associated with the dominant minerals. Interest in the trace element content of coal stems largely from their potential release, either in volatile form or as fine particulate from coal use, and the resulting environmental impact. However, there is nothing particularly toxicologically unique about the trace element composition of most coals.

Actual trace element emissions are somewhat variable depending on coal source, combustion or utilization process, and pollution abatement equipment. Among the volatile elements considered most hazardous are Be, F, As, Se, Cd, Pb, and Hg. Power plants equipped with modern gas cleanup systems efficiently prevent most of these elements from reaching the atmosphere. An exception is mercury (Hg), where the majority may bypass electrostatic emission controls. Although present at only levels of the order of 0.1 μg/g in coals, owing to the huge tonnages of coal used by the power industry, coal use is a significant factor in the total Hg being introduced into the environment by human activity. Development of practical control technology to capture and prevent the release of mercury is a high R&D priority. In developing countries where coal is burned without emission controls or where briquettes may still be used for domestic heating, certain elements such as As and F may present localized problems.

Although considerable information exists pertaining to the concentrations of trace elements in representative coals, it is much more difficult to obtain direct information on the modes of occurrence of trace elements. On a rather rudimentary level, elements are often classified by degree of organic or inorganic association based on the analysis of fractions from physical separation procedures. Such classifications may be highly misleading and may be more a reflection of particle size rather than chemical speciation. However, in most coals many of the elements of potential environmental concern including Co, Ni, Cu, Zn, As, Se, Mo, Cd, Hg, and Pb tend to be primarily associated with the sulfide (chalcophile) minerals.

5. PHYSICAL PROPERTIES

5.1 Porosity

The porosity of coal may be characterized by its pore volume and density, surface area, and pore size

TABLE IV

Semi-Quantitative Trace and Minor Element Survey of West Virginia (Ireland Mine) hvAb Coal

H																
					Concentrations (μg/g)[a]											
Li	Be				Fe =1.74% (atomic absorption)							B	C	N	O	F
1.5	INT											230				≤60
Na	Mg											Al	Si	P	S	Cl
300[b]	900[b]											9500[b]	22000[b]	80		520[b]
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br
1000[b]	800[b]	INT	600[b]	12	12	24	REF	3.2	22	7.6	21	4.6	5	7	0.4	18[b]
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I
14	90	3.9	15	5.1	1.1		≤0.4	≤0.2	≤0.4	≤0.1	≤0.5	REF	1.2	0.7	≤0.2	0.8
Cs	Ba	La	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At
1.1	38	4.4	0.7	INT	1.3	REF						2.3	14	0.6		
Fr	Ra	Ac	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu
			11	1.4	5.4		1.2	0.2	1	0.2	1.2	0.2	0.6	0.3	0.4	0.05
			Th	Pa	U											
			1.9		0.6											

[a]Determined by spark source mass spectrometry on low-temperature ash renormalized to whole coal basis unless otherwise indicated. Volatile elements such as halogens, S, and Sb are largely lost in ashing procedure.
[b]Determined by X-ray fluorescence on whole coal.
Note. REF, reference; Fe is used as an internal reference; In and Re are added as additional low-concentration references; INT, intereference.
Source: National Energy Technology Laboratory unpublished data.

distribution. These can be determined by a variety of experimental techniques, however, the value obtained is procedure and technique dependent. The porosity of coal has great significance in coal mining, preparation, and utilization. The inherent moisture resides in these pores. In its natural state, the pore structure in coal contains significant amounts of adsorbed methane that is both a mining hazard and in some cases a potentially recoverable resource. The ease of diffusion of methane is determined by the pore volume and pore size distribution. A proposed strategy for mitigating greenhouse gas emissions by sequestration of CO_2 in unmineable coal seams is also dependent on the porosity and adsorption capacity of the coal.

5.1.1 Pore Size

Coals are highly porous solids with pores that vary in size from cracks of micron dimensions to small voids, which are even closed to helium at room temperature. The open pore volume of a coal, calculated as the difference between the particle density (mass of a unit volume of the solid including pores and cracks) and helium density, may comprise as much as 10 to 20% of the coal volume. Total pore volumes derived using CO_2 densities tend to be even higher.

It is common practice to classify pores greater than 50 nm diameter as macropores, pores with diameters in the range 2 nm to 50 nm as mesopores (transitional pores), and pores less than 2 nm as micropores. Conventionally, the physical adsorption of gases or fluids is used to provide measurements of the pore size distribution of coals as well as a total pore volume. The pore size distribution of the macro- and mesopores is often determined by mercury porosimetry techniques. The micropore volume is then often estimated as the difference between the pore volume calculated from helium or CO_2 adsorption data and the pore volume from mercury porosimetry. The porosity in low-rank coals (carbon content $<75\%$) is predominately due to macropores, the porosity of high volatile bituminous coals

(carbon content 75 to 84%) is largely associated with meso- and micropores, while the porosity of higher rank coals such as low volatile bituminous and anthracite is predominately due to micropores.

5.1.2 Surface Area

The internal surface area of coals is rather high and primarily located in the micropores. The most commonly used determination methods are based on gas adsorption; CO_2 adsorption at $-78^{\circ}C$ has been used extensively. Generally CO_2 surface areas of coals are between $100\,m^2/g$ to $300\,m^2/g$, with values near $200\,m^2/g$ being typical.

The measured surface area of coals, particularly lignites and brown coals, can be significantly affected by sampling and sample preparation procedures. Implementation of gas adsorption procedures requires removal of the water present in micropores without changing the coal structure. Unless suitable procedures are employed, the structure of low-rank coals can shrink and be irreversibly altered upon drying.

5.1.3 Density

The density of coal is usually determined by fluid displacement. Owing to the porous nature of coal and interactions with the coal matrix, the measured density depends upon the fluid employed. The true density is usually determined by helium displacement with the assumption (not entirely valid) that helium will penetrate all of the pores without interacting chemically. The density of the organic phase usually lies between $1.3\,g/cm^3$ to $1.4\,g/cm^3$ for low-rank coals, has a shallow minimum at about $1.3\,g/cm^3$ for mid-rank bituminous coals, and then increases rapidly with rank and carbon content to as high as $1.7\,g/cm^3$ for anthracite. Other fluids often used to determine apparent densities are methanol and water. The methanol density is consistently higher than the helium density. The water density is similar to the helium density for many coals, but interpretation is complicated by the fact that water interacts with oxygen functionalities present. It may also induce swelling, particularly in low-rank coals, opening up porosity that may be closed to helium. The particle density is conventionally determined by mercury displacement.

Most of the mineral constituents in coal have densities about twice that of the organic phase. A notable exception is pyrite which has a density of about $5\,g/cm^3$. These density differences are exploited in physical coal cleaning procedures.

5.2 Optical and Spectral Properties

5.2.1 Reflectance

The reflectance is the proportion of directly incident light, conventionally expressed as a percentage, that is reflected from a plane polished surface, usually under oil-immersion, and under specified conditions of illumination. The reflectance of coal is rather low and is usually dominated by the vitrinite reflectance. For coals of the same rank, the liptinite maceral group shows the lowest reflectance and the inertinite group the highest reflectance. There is a continuous increase in vitrinite reflectance with carbon content and coal rank from less than 1% for low-rank coals to more than 4% for very high-rank coals. This is usually ascribed to the growth of aromatic clusters and greater structural ordering. Consequently, the reflectance is widely used as an index of coal rank. Procedures for polished specimens have been automated and standardized (ASTM D2798, ISO 7404/1-5) using polarized green light of wavelength 546 nm.

From reflectance measurements in two different media, two additional optical constants, the refractive index (n) and absorption (extinction) coefficient (k) can be derived. For vitrinites, refractive indexes measured range from 1.68 for vitrinite of 58% carbon content to 2.02 for vitrinite of 96% carbon content.

5.2.2 Absorption Spectra

The ultraviolet and visible absorption is unspecific and increases regularly from longer to shorter wavelengths. This is consistent with the ultraviolet and visible absorption being mainly due to electronic transitions in condensed aromatic systems. In contrast, in the infrared region, several major diagnostic absorption bands can be identified in the spectra of properly prepared coal samples despite a relatively high diffuse background. These bands can be associated with vibrations involving various structural groups in the coal.

5.3 Thermal Properties

5.3.1 Calorific Value

The calorific value (heating value) is a direct indication of the energy available for the production of steam and probably the most important parameter for determining the commercial usefulness of coal. It is also used as a rank classification parameter, particularly for low-rank coals. It is usually determined by the complete combustion of a specified quantity of coal in standardized calorimetric procedures (ASTM D5865; ISO 1928). The calorific

value of coal varies with rank in a manner that can be predicted from the changes in chemical composition. A typical value on a moist, mineral matter free basis for bituminous coal or anthracite is about 30 MJ/kg. Values for low-rank coals tend to be less, even on a moisture free basis, because of their higher oxygen content.

5.3.2 Heat Capacity/Specific Heat

The heat capacity of coal and its products is important in the design and development of any process that includes the thermal processing of coal and is also important for the complete understanding of the fundamental properties of coal. It is usually tabulated as the specific heat, that is, the dimensionless ratio of the heat capacity of the material of interest to the heat capacity of water at 15°C. The specific heat of coal usually increases with its moisture content, carbon content, and volatile matter content. The values for specific heats of various coals typically fall into the general range of 0.25 to 0.35. They can be estimated using a semi-empirical relationship between the specific heat and the elemental analysis.

5.3.3 Thermal Conductivity

The thermal conductivity is useful in estimating the thermal gradients in coal combustion furnaces and coal conversion reactors. It is difficult to assign a single value to the thermal conductivity of in place coal in a coal bed because in many cases the thermal conductivity parallel to the bedding plane appears to be higher than the thermal conductivity perpendicular to the bedding plane. The thermal conductivity of anthracite is of the order of 0.2 W/(m. K) to 0.4 W/(m. K), and the thermal conductivity of bituminous coal is typically in the range 0.17 W/(m. K) to 0.29 W/(m. K). The thermal conductivity of coal generally increases with an increase in apparent density as well as with volatile matter content, ash content, and temperature.

However, it is pulverized coal that is of interest for many applications. The thermal conductivity of pulverized coal is lower than that of the corresponding in place coal. For example, the thermal conductivity of pulverized bituminous coal usually falls in the range 0.11 W/(m. K) to 0.15 W/(m. K).

5.4 Mechanical Properties

5.4.1 Strength and Elasticity

At room temperature, bituminous and higher rank coals behave as brittle solids, whereas brown coals and lignites often display brittleness only when the rates of stressing are extremely high. The strength of coal specimens is commonly studied by means of a compression test, because the results have direct application to estimation of the load bearing capacities of pillars in coal mines. Measurement results are dependent on the orientation of the measurement to the bedding plane of the coal. In common with many other rocks, bituminous and higher rank coals exhibit a stress-strain curve in compression with an initial nonlinear portion associated with closure of cracks followed by an essentially linear elastic range. The breaking stress decreases with increasing specimen size. A dependence of both compression and impact strength on rank has been noted with a minimum for coals with 88 to 90% carbon on a moisture- and ash-free basis.

One of the most used parameters associated with the mechanical properties of coal is Young's modulus, a measure of the compressive or tensile strength of the coal relative to the strain or deformation before fracture. Measurements of Young's modulus in compression and in bending have been made for bituminous coal and anthracite, the mean values being about 3.4×10^9 Pa to 4.3×10^9 Pa, respectively. Measurement of the elastic properties of bulk coals are closely concerned with the presence of cracks or flaws and these must be taken into account in interpreting the results of stress-strain measurements and derived elastic constants.

5.4.2 Grindability

The grindability of coal, or the ease with which it may be ground fine enough for use as a pulverized fuel, is of considerable practical importance. It is a composite physical property usually evaluated by empirical tests; the Hardgrove grindability procedure and index (ASTM D409, ISO 5074) is commonly used. In general, the easier-to-grind coals are in the medium- and low-volatile bituminous ranks. They are easier to grind than coal of the high volatile bituminous, subbituminous, and anthracite ranks. High moisture contents, such as may be encountered in low-rank coals, lead to difficulty in grinding.

5.4.3 Friability

The friability is another empirically evaluated composite coal property of great significance in coal mining and beneficiation. Standard tests (ASTM D440, ASTM D441) provide a measure of the tendency of coal toward breakage on handling. Low-rank coals with high moisture content tend to be the least friable. The friability typically increases

with coal rank, reaching a maximum for low volatile bituminous coals and then decreasing for anthracite.

5.5 Electrical Properties

5.5.1 Electrical Conductivity (Resistivity)

The electrical conductivity of most coals is rather low. This parameter is usually expressed as its reciprocal, the resistivity or specific resistance. The resistivity decreases with increase in both rank and temperature. Dry bituminous coals have a specific resistance between 10^{10} and 10^{14} ohm cm at room temperature and can be classed as semiconductors. There is a rapid decrease in resistivity as the carbon content of coals becomes greater than 87% owing to an increase in the three-dimensional ordering of the carbon atoms that facilitates freer movement of electrons from one region to another. Anthracites have a specific resistance between 1 and 10^5 ohm cm, which is orientation dependent. The moisture content of coals strongly affects measurement results, since the specific resistance of water is about 10^6 ohm cm.

5.5.2 Dielectric Constant

The dielectric constant is a measure of the electrostatic polarizability of the electrons in coal. It is strongly dependent on the water content and indeed has been used to estimate the water content of low-rank coal. For carefully dried coal, the dielectric constant varies with coal rank. The value is approximately 5 for lignite, drops to a minimum of approximately 3.5 for bituminous coal, and then rises abruptly with increasing carbon content to a value of 5 for anthracite. Maxwell's relation that equates the dielectric constant of a non-polar material to the square of the refractive index only holds for coal at the minimum dielectric constant. The decrease in dielectric constant from lignite to bituminous coal may be attributed to the loss of polar functional groups. The increase in the dielectric constant as the carbon content increases toward anthracite has been attributed to the presence of polarizable electrons in condensed aromatic systems that increase in size with coal rank.

5.6 Magnetic Properties

5.6.1 Magnetic Susceptibility

The organic phase of coal is slightly diamagnetic—that is, there is a net repulsion when placed in a magnetic field. There are also traces of paramagnetism, at least partially due to stable free radicals. Some of the associated mineral matter may be ferromagnetic necessitating chemical removal before measurements. Correction for the paramagnetic contribution is possible by establishing the effect of temperature, then extrapolating to infinite temperature where paramagnetism essentially vanishes. Magnetized susceptibilities of many demineralized coals have been found to lie between -4×10^{-5} m^3/kg and -5×10^{-5} m^3/kg. The susceptability of one anthracite has been found to be -9.15×10^{-5} m^3/kg and exhibited anisotropy parallel and perpendicular to the bedding layers.

Only a slight dependence of the true diamagnetic susceptibility of coals on carbon content has been demonstrated. The susceptibility can be related to the calculated sums of the atomic susceptibilities of the component elements, with modification for carbon in various ring structures. From these results, estimates of the average number of aromatic rings per structural unit can be made.

5.6.2 Magnetic Resonance

Stable free radicals or other paramagnetic species containing unpaired electrons are present in all coals. These free radicals will interact with applied magnetic fields and can be studied by electron spin resonance spectroscopy which measures the absorption of electromagnetic radiation, usually in the microwave region, by a material in an appropriately strong external magnetic field. The conventional electron spin resonance spectrum of coal consists of a single broad line without resolvable hyperfine structure, but information regarding the overall concentration of free radicals and their general nature can still be derived. The free radical concentrations in coal increase approximately exponentially with increasing carbon content up to a carbon content of about 94%, after which a rapid decrease in the number of free spins is observed. Typical free spin concentrations range from 1 to 6×10^{18} g^{-1} for lignites to 5.3×10^{19} g^{-1} for an anthracite. These free radicals appear to be primarily associated with the aromatic structures in the coal.

Coals also are composed of many atomic nuclei (e.g., ^{1}H and ^{13}C) that have a nonzero nuclear spin and the resulting nuclear magnetic moments will also interact with applied magnetic fields and can be studied by nuclear magnetic resonance techniques employing radio frequency radiation. In particular, ^{13}C nuclear magnetic resonance has provided a great deal of useful information on the chemical types and environment of carbon atoms in solid coals.

6. CHEMICAL PROPERTIES

6.1 Thermal Changes

Coal undergoes a variety of physical and chemical changes when it is heated to elevated temperatures. Adsorbed water and adsorbed gases such as methane and carbon dioxide will appear as a product of thermal treatment below 100°C. Low-rank coals, such as lignites, begin to evolve carbon dioxide by thermal decarboxylation as temperatures are raised above 100°C.

Important internal molecular rearrangements (e.g., condensation and other chemical reactions) begin to occur at temperatures as low as 175°C to 200°C. These result in alteration of the original molecular structures without generating major amounts of volatile reaction products.

With the exception of anthracite coals, massive weight losses begin to occur at temperatures between 350°C to 400°C from the evolution of volatile matter in the form of tars and light oils. These result from extensive thermally induced molecular fragmentation and concurrent stabilization and recombination reactions. A solid residue of higher carbon and aromatic content than the original coal remains. As temperatures are raised above ~550°C only light gases, primarily H_2 and CO accompanied by smaller amounts of CH_4 and CO_2, are emitted.

Of special commercial importance are medium-rank caking coals that soften in the temperature range 400°C to 500°C, go through a transient plastic stage, and agglomerate to form a solid semicoke residue. Upon heating to even higher temperatures, the semicoke shrinks, hardens, and evolves additional light gases, and eventually forms strong metallurgical coke at 950°C to 1050°C. Prime coking coals are most typically vitrinite-rich bituminous coals with volatile contents of 20 to 32%. In addition to the standard coal assays, a number of empirical laboratory tests have been developed to assess the commercial coking potential of a coal. These tests usually involve determination of the swelling, plasticity, or agglomeration characteristics upon heating at a fixed rate to a specified temperature and correlation with coke quality.

6.2 Oxidation

6.2.1 Weathering

With the exception of anthracite, freshly mined coals are susceptible to oxidation at even ambient temperatures. The oxidation of coal is an exothermic process and the rate of reaction increases markedly with temperature. It is also a function of the surface area and promoted by moisture. Under appropriate conditions, spontaneous ignition of coal can occur. Care must be exercised in the mining, storage, and shipment of many coals, such as lower rank coals from the western United States, to maintain conditions that will prevent the occurrence of spontaneous ignition. Extensive weathering may also negatively impact the heating value and caking characteristics of a coal.

6.2.2 Combustion

Most mined coal is burned to produce steam for electric power generation. An obvious property of interest is the heating value, which determines the amount of steam that can be generated and is determined by a standardized calorimetric test procedure. Numerous other properties determined by standard tests are of interest in selecting a coal for a particular combustion system. The moisture content of the coal affects handling characteristics and freight costs as well as tending to lower the flame temperature. The amount of volatile matter affects the ease of combustion and flame length. In general, high-rank coals are more difficult to ignite than low-rank coals. The amount of ash governs the loss of sensible heat in ash rejection and affects ash disposal and transport costs. The sulfur content of the feed coal is of importance in meeting environmental emission requirements.

6.2.3 Gasification (Partial Oxidation)

If supplied with less than the stoichiometric amount of oxygen for complete combustion at elevated temperatures, usually above 700°C, coals will yield a gaseous product rich in CO and H_2 that is suitable for use either as a source of energy or as a raw material for synthesis of chemicals, liquid fuels, or other gaseous fuels. Industrial gasifiers commonly utilize a mixture of steam and air or oxygen with the amount of oxygen being generally one-fifth to one-third the amount theoretically required for complete combustion. The distribution and chemical composition of the products will depend not only on the coal feedstock and supplied atmosphere, but on actual gasifier conditions such as temperature, pressure, heating rate, and residence time.

Gasification technology has been adapted to a wide range of coals from brown coals to anthracites. However, reactivity tends to decrease with increase in coal rank. Alkaline elements such as Ca may also have catalytic properties that increase the char

gasification reactivity. Depending on the design of the gasifier system, ease of grindability and a low ash content are usually desirable. Characteristics of the resulting coal ash are very important, particularly the ash fusion temperature. The composition of the gasifier slag impacts gasifier refractory life. Coals with low sulfur and halogen content may be preferred to minimize downstream corrosion.

Gasification technologies, particularly Integrated Gasification Combined Cycles (IGCC), have become increasingly attractive as an option to improve power generation efficiency while decreasing atmospheric emissions and providing the possibility of cogeneration of hydrogen or other ultra clean fuels and chemicals with electricity.

6.3 Hydrogenation/Hydrolysis

Although the organic portion of coal will undergo hydrogenation and hydrolysis reactions at relatively low temperatures ($\sim$100°C) in certain strong chemical reducing systems (e.g., metal/amines), most large-scale practical applications utilize reaction with H_2. The latter requires conditions of some severity, that is, elevated temperature and pressure to proceed at reasonable rates. Compared to other hydrocarbons, coal may be considered a hydrogen deficient natural product with an H/C ratio of about 0.8 versus 1.4 to 1.8 for most hydrocarbon oils and a ratio of 4.0 for methane. The addition of hydrogen is a means of not only saturating the hydrogen deficient products of the thermal breakdown of the coal structure, but also terminating side reactions that would lead to higher molecular weight products and solid residues. It also promotes the removal of nitrogen, oxygen, and sulfur as NH_3, H_2O, and H_2S.

The direct hydrogenation of coal has been extensively investigated as a means of producing liquid fuels or petroleum substitute liquids. Modern multistage process concepts involve subjecting coal slurried with recycled process derived vehicle under a hydrogen-rich atmosphere to elevated temperature and pressure. First stage temperatures are often 400°C to 450°C and pressures $\sim$14 MPa to promote coal dissolution. Subsequent stages are typically catalytic and at slightly lower temperatures to promote hydrogenation and upgrading of the first stage product. Depending on the coal, these processes can convert one ton of coal into as much as 4.5 barrels of refinery feedstock. However, although technically feasible, direct hydrogenation of coal is not at the moment practiced on a large commercial scale owing to the availability of economic and proven alternatives.

6.4 Miscellaneous Reactions

Other than the direct reaction with oxygen and hydrogen, the organic material in coal will react with a variety of chemical reagents. Owing to the complexity and heterogeneity of the coal structure, such reactions tend to be somewhat nonselective and yield a complex mixture of products that complicates detailed analysis and interpretation. Reagent accessibility and the extent of reaction are also potential problems. These reactions are typically carried out on a laboratory scale and are now of most interest from the standpoint of gaining insight into the fundamental structure of coal or for analytical applications.

The organic coal structure will react in a rather nonselective manner with many common oxidants such as nitric acid, peroxides, performic acid, permanganate, dichromate, and hypochlorites. Mild oxidation tends to result in an increase in high molecular weight alkali soluble material (humic acids), particularly in the case of low-rank coals. Upon further oxidation, these degrade to lower molecular weight oxygenated components and ultimately to carbon dioxide.

A number of acid or base catalyzed reactions have been used to partially depolymerize or increase the solubility of coals to provide products more amenable to analysis. Typical examples are acid catalyzed reactions such as treatment with BF_3/phenol, reductive alkylation using an alkali metal/tetrahydrofuran, Friedel-Crafts alkylation or acylation, and base hydrolysis. Although these reactions have provided some insight into the chemical structures in coals, the exact relationship of the ultimate reaction products to their precursors in coal is not easily established.

The organic structure of coal will undergo addition and substitution reactions with halogens producing a complex mixture of products that depend on the coal type and extent of halogenation.

Certain chemical functional groups, particularly those containing oxygen such as phenolic OH and carboxyl, can be quantified by procedures based upon derivatization reactions with selective chemical reagents.

7. COAL DATABASES

Extensive tabulations of the properties of coals are maintained by various government agencies and other organizations in the major coal producing nations. In the mid-1970s, the U.S. Geological Survey (USGS) initiated a project to create a

comprehensive national coal information database. The original goal was to obtain and characterize at least one sample per coal bed from every geographic quadrangle (approximately 50 square miles) underlain by coal in the United States. The database, now known as the National Coal Resources Data System, is the largest publicly available database of its kind. Readily accessible are comprehensive analyses of a subset of more than 13,000 samples of coal and associated rocks from every major coal-bearing basin and coal bed in the United States. Also available from the USGS is coal quality information on selected coals of the Former Soviet Union compiled in cooperation with the Vernadsky State Geologic Museum and the Russian Academy of Sciences.

The Premium Coal Sample Program implemented at Argonne National Laboratory resulted in the availability of samples of eight highly uniform, premium (unexposed to oxygen) coals for laboratory researchers investigating coal structure, properties, and behavior. The suite of coals include a lignite, subbituminous, high-volatile, medium-volatile, and low-volatile bituminous, as well as a liptinite-rich, an inertinite-rich, and a coking coal. Associated with this program is a comprehensive database of information developed from these representative coals and an extensive bibliography of published research carried out using them.

The Energy Institute at the Pennsylvania State University maintains the Penn State Coal Sample Bank and Database, which still has more than 1100 coal samples available for distribution. Particularly useful is the subset maintained as the Department of Energy Coal Sample Bank and Database. This comprises a suite of 56 carefully sampled coals stored in argon in foil/polyethylene laminate bags under refrigeration and an associated computer accessible database of analyses.

The British Coal Utilization Research Association (BCURA) maintains a coal sample bank containing 36 well-characterized coal samples ranging from lignite to anthracite. These are primarily coals from the United Kingdom, but also now include a number of internationally traded, primarily power station coals.

The Center for Coal in Sustainable Development (CCSD) also maintains a coal bank of 14 representative Australian coals that have been analyzed by standard procedures and are available for research purposes.

SEE ALSO THE FOLLOWING ARTICLES

Clean Coal Technology • Coal Conversion • Coal, Fuel and Non-Fuel Uses • Coal Industry, Energy Policy in • Coal Industry, History of • Coal Mine Reclamation and Remediation • Coal Mining, Design and Methods of • Coal Mining in Appalachia, History of • Coal Preparation • Coal Resources, Formation of • Coal Storage and Transportation • Markets for Coal

Further Reading

Argonne National Laboratory. "Users Handbook for the Argonne Premium Coal Sample Program." Found at www.anl.gov.

ASTM Committee D-5 on Coal and Coke (2003). Gaseous fuels, coal and coke."Annual Book of ASTM Standards," Vol. 05.06. American Society for Testing and Materials. West Conshohocken, PA.

Berkowitz, N. (1985). "The chemistry of coal. Coal Science and Technology, 7." Elsevier, New York.

The British Coal Utilization Research Association (BCURA) Coal Bank. Found at www.bcura.org.

International Organization for Standardization (ISO). Technical Committee 27, Subcommittee 5 (Solid Mineral Fuels). Found at www.iso.ch.

NAS-NRC Committee on Chemistry of Coal (1963). "Chemistry of Coal Utilization, Supplementary Volume" (H. H. Lowry, Ed.), Chapters 1–6. John Wiley & Sons, New York.

NAS-NRC Committee on Chemistry of Coal (1981). "Chemistry of Coal Utilization, Second Supplementary Volume" (M. A. Elliott, Ed.), Chapters 1–8. John Wiley & Sons, New York.

National Research Council Committee (1945). "Chemistry of Coal Utilization" (H. H. Lowry, Ed.), Vol. 1. John Wiley & Sons, New York.

The Pennsylvania State University Energy Institute, Coal and Organic Petrology Laboratories, Penn State Coal Sample Bank and Database. Found at www.ems.psu.edu.

U.S. Geological Survey, National Resources Data System, U.S. Coal Quality Database. Found at www.energy.er.usgs.gov.

van Krevelen, D. W. (1961). "Coal." Reprinted 1981 as "Coal Science and Technology," Vol. 3. Elsevier, Amsterdam, The Netherlands.

Coal Conversion

MICHAEL A. NOWAK
U.S. Department of Energy, National Energy Technology Laboratory
Pittsburgh, Pennsylvania, United States

ANTON DILO PAUL and RAMESHWAR D. SRIVASTAVA
Science Applications International Corporation, National Energy Technology Laboratory Pittsburgh, Pennsylvania, United States

ADRIAN RADZIWON
Parsons, National Energy Technology Laboratory
South Park, Pennsylvania, United States

1. Combustion
2. Pyrolysis and Coking
3. Gasification
4. Liquefaction

Glossary

gasification Partial oxidation of coal to produce a mixture consisting primarily of carbon monoxide and hydrogen.
hydrogenation Chemical addition of hydrogen to an unsaturated molecule.
integrated gasification combined cycle A process that integrates a combustion turbine fed by a coal gasifier and waste heat recovery to produce steam for a steam turbine or process heat.
NO_x Chemical shorthand for a mixture of oxides of nitrogen; the components of NO_x are acid rain precursors and also participate in atmospheric ozone chemistry.
SO_2 Chemical shorthand for sulfur dioxide; SO_2 is an acid rain precursor.

Chemical conversion of coal can be divided into broad categories: combustion to provide process heat and steam and chemical conversion to liquid or gaseous fuels and chemicals. This article discusses the combustion of coal for the purposes of making electricity and the conversion of coal into solid, gaseous, or liquid products that can be used in metallurgical processing or converted into finished fuels and chemicals.

1. COMBUSTION

There are several technologies currently in use to burn coal. Whichever technology is used, the basic chemistry of combustion is the same. Although coal is used in a wide range of processes, the vast majority is simply burned to obtain heat. The majority of that heat is used to produce steam for electric power production. The balance is used to provide process steam and steam for space heating. This is the case worldwide, more so in the United States. The following are the basic important reactions that summarize the combustion process:

$$C + O_2 \rightarrow CO_2$$
$$H_2 + 1/2O_2 \rightarrow H_2O$$
$$S + O_2 \rightarrow SO_2$$
$$N\,(\text{fuel-bound}) + O_2 \rightarrow NO_x$$
$$N_2\,(\text{atmospheric}) + O_2 \rightarrow NO_x$$

The first two reactions are important because they account for nearly all the heat released in the combustion process. The last three are important since they produce pollutants that must be controlled if coal is to be burned in an environmentally acceptable manner. In addition to SO_2 and NO_x, particulate matter is also a pollutant that needs to be controlled, with the current emphasis being on particles that are less than 2.5 μm in diameter. Particulate matter is readily controlled with electrostatic precipitators or fabric filters. Development work continues to improve the performance of these devices and to allow them to operate at higher temperatures. Work is also being carried out on

 © 2004 Elsevier Inc. All rights reserved.

high-temperature ceramic filters. A few coal-fired plants also use Venturi scrubbers to remove particulates, but this technology is little used today. SO_2 is typically controlled with calcium- or sodium-based sorbents. It can be injected into the boiler or ductwork, wet or dry, and removed in the particulate control system. The sorbent can also be used in the form of a slurry or solution and used to contact the flue gas in a system installed downstream of the boiler. These wet systems typically show the greatest removal rates. There are also several technologies (NOXSO, copper oxide, and SNOX) that can remove the SO_2 and recover elemental sulfur or sulfuric acid, which can be sold. These processes also provide NO_x control. Some wet SO_2 removal systems that use a calcium-based reagent produce a marketable gypsum by-product. NO_x is controlled by several techniques or combinations of techniques. One set of NO_x control technologies involves carrying out the initial, hottest part of the combustion process in an air-deficient zone and then introducing additional air to complete the combustion process after the reaction products have cooled to temperatures that are less conducive to NO_x formation. This can be accomplished by specially designed burners or operating burners in an air-deficient mode and injecting overfire air into the cooler region. Another technique, reburning, entails operating the burners with normal coal:air ratios and then injecting additional fuel downstream of the burners, thus creating a reducing zone in which NO_x is destroyed. Overfire air is then injected to complete the combustion reactions at lower temperature. NO_x is also lowered by using selective noncatalytic reduction (SNCR) or selective catalytic reduction (SCR). In both cases, a reducing agent, such as ammonia or urea, is injected into the flue gas to react with the NO_x to form N_2 and water vapor. In the SNCR process, the reagent is injected into the boiler, where conditions are optimal for the reaction to proceed. The SCR process uses a catalyst bed at lower temperatures to carry out the reaction. SCR is both the most expensive and the most effective of the NO_x control strategies. Although there are no requirements to control mercury and control technologies are still under development, it is anticipated that controls may become necessary.

As stated previously, there are several technologies in use to burn coal. The first step toward modern coal combustion (Table I) systems was the stokers. Of those commonly in use today, the underfed stoker is the oldest. It uses a fixed grate and coal is forced into the bottom of the furnace by a mechanical ram.

TABLE I

Time Line of Coal Combustion Technology

300 BC	Greeks record use of coal combustion
1700s	First steam engines
1820s	Stoker boiler
1881	First steam electric plant
1920s	Pulverized coal boiler
1920–present	Atmospheric fluidized bed combustor
	Pulverized coal boiler to 38% efficiency
	Supercritical boilers
	Cyclone combustors (1950s)
	Magnetohydrodynamics
	Pressurized fluidized bed combustor (1980s)
	Environmental controls for No_x and SO_2

As additional coal is fed into the furnace, the coal is forced up and onto the grate. As burning proceeds, the ash is forced off the grate and removed. These units are inefficient, limited in size, and have relatively high maintenance costs. These have been largely replaced by traveling grate and spreader stokers. Both have continuous grates that slowly move through the furnace. In the traveling grate, coal is fed at the front end, and the grate moves front to back, where the ash is discharged. The feed system maintains a supply of coal at the end of the grate outside the combustion zone. The coal is carried into the furnace as the grate moves and the quantity is controlled by a fixed plate a few inches above the bed. Ash is discharged from the grate at the back of the furnace. In the spreader stoker, the grate moves back to front and the ash is discharged below the feed mechanism. The feed mechanism consists of a small conveyor that feeds the coal to a device (reciprocating or rotary) that essentially tosses the coal outward over the grate. The spreader and traveling grate stokers can be made larger than the underfed but are still limited in size compared to more modern boilers. This is especially important when environmental controls are needed since reverse economies of scale make such investments more costly on a unit capacity basis.

The workhorse of today's coal-fired generation is the pulverized coal boiler (Fig. 1). The coal is typically pulverized so that 70% will pass through a 200-mesh (74-μm opening) screen. These units are usually supplied as wall-fired or tangentially fired units. A few are also built with the burners mounted on top with the flame directed downward. The burners of a wall-fired boiler are mounted on one or

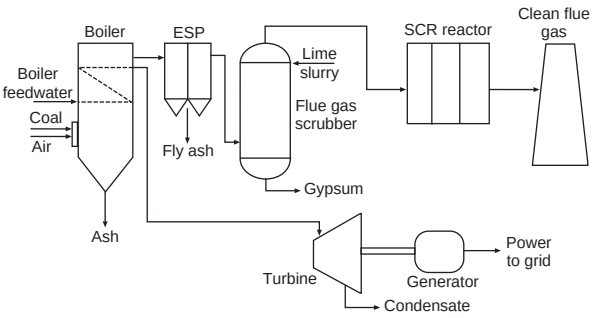


FIGURE 1 Block diagram of a typical pulverized coal-fired power plant.

two (opposing) walls and direct the coal and air horizontally into the boiler. The burners of a tangentially fired boiler are normally mounted on the corners of the boiler and the flame is directed to a point near, but not at, the center of the boiler to impart a swirling motion to the gases that promotes good mixing of the air and fuel. All burners are designed to provide the desired degree of mixing the air and coal. Although the burners of wall- and tangentially fired boilers are basically horizontal, they can be inclined to allow better control of the overall combustion operation. In low-NO_x burners, the design is such that there is some delay in mixing all the air with the fuel to inhibit NO_x formation. These boilers can either be wet- or dry-bottom depending on whether the ash is removed as a molten slag or as dry solid.

Cyclone boilers, which came into use after pulverized coal-fired boilers, have combustors mounted on the outside of the boiler. They inject the air tangentially into the combustor to create the cyclonic motion. This cyclonic motion promotes intense combustion and causes the ash to accumulate on the walls of the combustor, from which it is removed as a molten slag. These burners use a coarser coal than the pulverized coal boilers. Cyclone burners are mounted so that they are sloped toward the boiler. Cyclone boilers are seldom considered for new installations since the intense combustion also promotes NO_x formation.

In recent years, fluidized bed boilers have become more widely used. Fluidized bed systems offer several advantages over pulverized coal and cyclone boilers, including higher thermal efficiency, better environmental performance, the ability to burn a much wider range of coals and other fuels (including waste coal), and a relatively low combustion temperature that inhibits NO_x formation. If necessary, they can be equipped with SCR or SNCR to achieve extremely low NO_x levels. In these boilers, combustion takes place in a fluidized bed of sand or ash. The combustion air is injected through a distribution plate or bubble caps at the bottom of the boiler to maintain fluidization in the bed. Coal and limestone (for SO_2 removal) are injected into the bed at a point above the distribution plate. With no airflow (slumped bed) the bed is simply a bed of solid particles. As airflow increases, the bed becomes fluidized. If airflow increases, the air moves upward as air bubbles. This is referred to as a bubbling fluidized bed. If airflow is further increased, more solids are entrained, removed in an internal cyclone, and returned to the bed, which is now a circulating fluidized bed. Steam is generated in tubes mounted on the walls and tube banks immersed in the bed, with additional heat recovery taking place downstream. This arrangement results in excellent heat transfer characteristics. The first fluidized bed combustors to enter the market operated at atmospheric pressure. Next came the pressurized fluid bed combustors, which can operate at pressures up to 20 atm. In addition to generating power from the steam, additional power is generated by passing the hot, pressurized combustion products (after particulate removal) through an expansion turbine to drive the air compressor and generate additional electric power. The next version of fluidized bed is a pressurized fluidized bed coupled with a partial gasifier that produces a fuel gas that is used in a gas turbine for power generation. The char from the partial gasifier fuels the fluidized bed combustor, which operates in the same manner as when coal is fed directly to the combustor. By careful integration of the subsystems, very high thermal efficiencies can be achieved.

2. PYROLYSIS AND COKING

Although combustion is an exothermic reaction involving stoichiometric amounts of oxygen and achieving flame temperatures as high as 1650°C, coal pyrolysis is an endothermic process. When coal is heated, it undergoes thermal decomposition that results in the evolution of gases, liquids, and tars, leaving behind a residue known as char. Coal pyrolysis is an important process for making metallurgical coke. Coal pyrolysis is a very old technique based on relatively simple technology that dates back to the 18th century. Most pyrolysis systems used in the late 1800s to the early 1900s were located in Europe, where the objective was the

production of a smokeless fuel (char) for domestic heating and cooking purposes. It was soon realized in the manufacture of char for fuel that the coal tar fraction contained valuable chemical products that could be separated through refining. However, as inexpensive petroleum appeared in the mid-1900s, interest in coal pyrolysis and its by-products faded.

Coal pyrolysis processes are generally classified as low temperature (<700°C), medium temperature (700–900°C), or high temperature (>900°C). Coal undergoes many physical and chemical changes when heated gradually from ambient temperature to approximately 1000°C. Low-temperature changes include loss of physically sorbed water, loss of oxygen, and carbon–carbon bond scission. At higher temperatures (375–700°C), thermal destruction of the coal structure occurs, as reflected by the formation of a variety of hydrocarbons, including methane, other alkanes, polycyclic aromatics, phenols, and nitrogen-containing compounds. In this temperature range, bituminous coals often soften and become plastic (thermoplastic) to varying degrees. At temperatures between 600 and 800°C, the plastic mass undergoes repolymerization and forms semicoke (solid coke containing significant volatile matter). At temperatures exceeding 600°C, semicoke hardens to form coke with the evolution of methane, hydrogen, and traces of carbon oxides. Pyrolysis of coal is essentially complete at approximately 1000°C.

During pyrolysis, the yield of gaseous and liquid products can vary from 25 to 70% by weight, depending on a number of factors, including coal type, pyrolysis atmosphere, heating rate, final pyrolysis temperature, and pressure. Although certain operating conditions may increase product yield, achieving these conditions may result in increased costs. Coal rank is the predominant factor in determining pyrolysis behavior. Higher rank coal, particularly high volatile A bituminous coals, produce the highest yield of tar. However, higher rank coals give products that tend to be more aromatic (i.e., they have a lower hydrogen:carbon ratio).

Other factors that can improve pyrolysis yields are lower pyrolysis temperatures, utilization of smaller particle size coal, and reduced pressure or employing a reducing (i.e., hydrogen) atmosphere. A reducing atmosphere also improves the yield of liquid and lighter products. Heating rate also affects yield distribution, with rapid heating providing a higher liquid:gas ratio. Pyrolysis in the presence of other gases, particularly steam, has been investigated and is reported to improve liquid and gas yields, but little information is available.

Process conditions also affect the product char. At temperatures higher than 1300°C, the inorganic components of coal can be separated from the products as slag. Rapid heating, conducted in a reactive atmosphere, produces a char with higher porosity and reactivity. The use of finer coal particles, lower temperatures resulting in longer process times, and the complexities of using vacuum, pressure, or hydrogen all add to the cost of coal pyrolysis.

The uses of coal pyrolysis liquids can be divided into two broad categories: (i) direct combustion, requiring little or no upgrading, and (ii) transportation fuels and chemical, requiring extensive upgrading. Much attention with regard to low-temperature tar processing has been devoted to hydroprocessing techniques, such as hydrotreating and hydrocracking, with the primary objectives of reducing viscosity, reducing polynuclear aromatics, and removing heteroatoms (sulfur, nitrogen, and oxygen) to produce usable fuels and chemicals. The cost of hydrogen is the primary impediment to tar upgrading. The tar fraction can be used as a source for chemicals, such as phenolics, road tars, preservatives, and carbon binders, but these uses do not constitute a large enough market to support a major coal pyrolysis industry.

Probably the most common application of coal pyrolysis is in the production of coke. The production of metals frequently requires the reduction of oxide-containing ores, the most important being production of iron from various iron oxide ores. Carbon in the form of coke is often used as the reducing agent in a blast furnace in the manufacture of pig iron and steel. The blast furnace is basically a vertical tubular vessel in which alternate layers of iron ore, coke, and limestone are fed from the top. Coal cannot be fed directly at the top of a blast furnace because it does not have the structural strength to support the column of iron ore and limestone in the furnace while maintaining sufficient porosity for the hot air blast to pass upward through the furnace.

Not all coals can produce coke that is suitable for use in a blast furnace. The property that distinguishes coking coals is their caking ability. Various tests are performed to identify suitable coal for conversion to coke through pyrolysis. Frequently, several coals are blended together to achieve the necessary coal properties to produce a suitable coke. Commercial coke-making processes can be divided into two categories: nonrecovery coke making and by-product coke making.

In nonrecovery coke plants, the volatile components released during coke making are not recovered

but, rather, are burned to produce heat for the coke oven and for auxiliary power production. One of the earliest nonrecovery units was the beehive oven, which for many years produced most of the coke used by the iron and steel industry. With these ovens, none of the by-products produced during coking were recovered. Because of their low efficiency and pollution problems, beehive ovens are no longer in use in the United States.

Nonrecovery coking takes place in large, rectangular chambers that are heated from the top by radiant heat transfer and from the bottom by conduction through the floor. Primary air for the combustion of evolved volatiles is controlled and introduced through several ports located above the charge level. Combustion gases exit the chamber through down comers in the oven walls and enter the floor flue, thereby heating the floor of the oven. Combustion gases from all the chambers collect in a common tunnel and exit via a stack that creates a natural draft for the oven. To improve efficiency, a waste heat boiler is often added before the stack to recover waste heat and generate steam for power production.

At the completion of the coking process, the doors of the chamber are opened, and a ram pushes the hot coke (approximately 2000°F) into a quench car, where it is typically cooled by spraying it with water. The coke is then screened and transported to the blast furnace.

The majority of coke produced in the United States comes from by-product coke oven batteries. By-product coke making consists of the following operations: (i) Selected coals are blended, pulverized, and oiled for bulk density control; (ii) the blended coal is charged to a number of slot type ovens, each oven sharing a common heating flue with the adjacent oven; (iii) the coal is carbonized in a reducing atmosphere while evolved gases are collected and sent to the by-product plant for by-product recovery; and (iv) the hot coke is discharged, quenched, and shipped to the blast furnace.

After the coke oven is charged with coal, heat is transferred from the heated brick walls to the coal charge. In the temperature range of 375–475°C the coal decomposes to form a plastic layer near the walls. From 475 to 600°C, there is marked evolution of aromatic hydrocarbons and tar, followed by resolidification into semicoke. At 600–1100°C, coke stabilization occurs, characterized by contraction of the coke mass, structural development of coke, and final hydrogen evolution. As time progresses, the plastic phase moves from the walls to the center of the oven. Some gas is trapped in the plastic mass, giving the coke its porous character. When coking is complete, the incandescent coke mass is pushed from the oven and wet or dry quenched prior to being sent to the blast furnace. Modern coke ovens trap the emissions released during coke pushing and quenching so that air pollution is at a minimum.

Gases evolving during coking are commonly termed coke oven gas. In addition to hydrogen and methane, which constitute approximately 80 volume percent of typical coke oven gas, raw coke oven gas also contains nitrogen and oxides of carbon and various contaminants, such as tar vapors, light oil vapors (mainly benzene, toluene, and xylene), naphthalene, ammonia, hydrogen sulfide, and hydrogen cyanide. The by-product plant removes these contaminants so that the gas can be used as fuel. The volatiles emitted during the coking process are recovered as four major by-products: clean coke oven gas, coal tar, ammonium sulfate, and light oil. Several processes are available to clean coke oven gas and separate its constituents.

In the past, many products valuable to industry and agriculture were produced as by-products of coke production, but today most of these materials can be made more economically by other techniques. Therefore, the main emphasis of modern coke by-product plants is to treat the coke oven gas sufficiently so that it can be used as a clean, environmentally friendly fuel. Coke oven gas is generally used in the coke plant or a nearby steel plant as a fuel.

3. GASIFICATION

Coal was first gasified in England by William Murdock in 1792, and the world's first coal gas company was chartered in England in 1812. Coal gas was first produced in the United States in 1816 in Baltimore, and by 1850 more than 55 commercial coal gasification plants in the United States were generating gas for lighting and heating. During the late 1800s and early 1900s, a large number of coal gasifiers operated commercially in the United States and Europe to produce industrial and residential fuel gas.

Most of the early gasifiers were moving bed units, charged with sized coal and blown with steam and air to generate "producer gas" (150 Btu/scf). Operating these moving bed gasifiers in a cyclic mode (blowing first with air to heat the coal, followed by contact with steam to produce "water gas")

increased the heating value of the product gas to 300 Btu/scf. The heating value was further increased to approximately 500 Btu/scf by cofeeding oi1 with steam to produce "carbureted water gas," which contained hydrocarbons in addition to H_2 and CO. An early gasification process, still in use today, was that developed by Lurgi.

Extensive process development was carried out in the United States in the late 1940s to mid-1950s, prompted by a concern that natural gas reserves were limited. Recent interest in coal gasification has been driven by the potential of integrated gasification combined cycle (IGCC) facilities to increase the efficiency of power production while reducing emissions.

The initial step in coal gasification involves grinding and/or pretreatment of the coal to put it into a form suitable for injection into the gasifier. In the gasifier, the coal is heated in the presence of a reactive gas (air, oxygen, and steam), the nature of which depends on the product desired. Reactions occurring during gasification of coal can be divided into three groups: pyrolysis reactions (thermal decomposition of the coal), gasification reactions (gas–solid reactions), and gas–gas reactions. The following are major reactions:

Pyrolysis reaction

$$C_xH_y \rightarrow (x - y/4)C + y/4CH_4$$

Gas–solid reactions

$$C + 1/2O_2 \rightarrow CO$$
$$C + O_2 \rightarrow CO_2$$
$$C + H_2O \rightarrow CO + H_2$$
$$C + CO_2 \rightarrow 2CO$$
$$C + 2H_2 \rightarrow CH_4$$

Gas–gas reactions

$$CO + H_2O \rightarrow H_2 + CO_2$$
$$CO + 3H_2 \rightarrow CH_4 + H_2O$$

The first reaction that occurs is pyrolysis or devolatilization. Depending on the type of gasifier, condensable hydrocarbons may be collected as a by-product or may be completely destroyed. Combustion reactions are the fastest gasification reactions and are highly exothermic. Oxidation reactions occur rapidly with essentially complete oxygen utilization. The primary combustion products are CO and CO_2. The gas–solid reactions illustrate the gasification of char by reaction with various gases. In addition to the carbon–steam reaction, steam undergoes a side reaction called the water gas shift reaction. Because coal chars are highly microporous, most of the gasification reactions take place inside the char particles.

The reducing nature of the product gas causes most heteroatoms (sulfur, nitrogen, and chlorine) to appear in reduced form as hydrogen sulfide, ammonia, and hydrogen chloride. In most cases, these materials are scrubbed from the product gas before it is burned or reacted. Ammonia and HCl are very water soluble and easily removed by a water wash, and several processes are available for removal of H_2S. Typical gasifier raw gas has a composition of 25–30% H_2, 30–60% CO, 5–15% CO_2, 2–30% H_2O, and 0–5% CH_4. Other components include H_2S, COS, NH_3, N_2, Ar, and HCN, depending on the composition of the coal, the purity of the oxygen used, and the operating conditions.

Gasification processes can be classified into three major types: moving bed (countercurrent flow), fluidized bed (back-mixed); and entrained-flow (not backmixed). Moving bed gasifiers consist of a downward-moving bed of coal in contact with a countercurrent flow of gas moving upward through the bed. As it moves down the bed, the coal undergoes the following processes: drying, devolatilization, gasification, combustion, and ash cooling. Moving bed gasifiers can be operated at atmospheric or higher pressure with either air or oxygen as the oxidant, with ash removed either dry or molten, and with or without stirrers to prevent agglomeration. At the top of the bed, the hot upward-flowing gases dry the coal. As the coal moves down, its temperature increases until at approximately 315–480°C pyrolysis occurs, generating gases, oils, and tars. As the resulting char moves further down the bed, it is gasified by reaction with steam, carbon dioxide, and hydrogen to produce a mixture of carbon monoxide, hydrogen, methane, carbon dioxide, unreacted steam, and a small amount of other gases. Below this zone, char is combusted by reaction with oxygen.

The ratio of steam to oxygen in the gasifier controls the temperature in the combustion zone. If dry-bottom operation is desired, sufficient steam is added to offset the exothermic oxidation reactions with endothermic steam–carbon reactions to stay below the ash fusion temperature. Slagging gasifiers operate at a higher temperature, and ash is removed in a molten state and then quenched in a water bath. Moving bed gasifiers require sized coal for proper operation, typically in a range 0.25–2 in.

In a fluidized bed gasifier, reactant gases are introduced through a distributor at the bottom of

the bed at a velocity sufficient to suspend the feed particles. The result is a bed of highly mixed solids in intimate contact with the gas phase. The high degree of mixing leads to a uniform temperature throughout the bed and results in reaction rates that are typically higher than those experienced in moving bed gasifiers.

Exit gas temperature for a fluidized bed gasifier is higher than that for a moving bed gasifier, resulting in further pyrolysis of intermediate hydrocarbon products so that the product gas contains a much lower concentration of tar/oil. Unconverted char and ash are removed as dry solids. If strongly caking coals are used, pretreatment is required. Fluidized bed gasifiers can be operated at atmospheric or higher pressure. The fluidizing gas is a mixture of steam with either air or oxygen.

In an entrained-flow gasifier (Fig. 2), a mixture of finely ground coal entrained in a reactant gas flows through the reactor, with little or no backmixing.

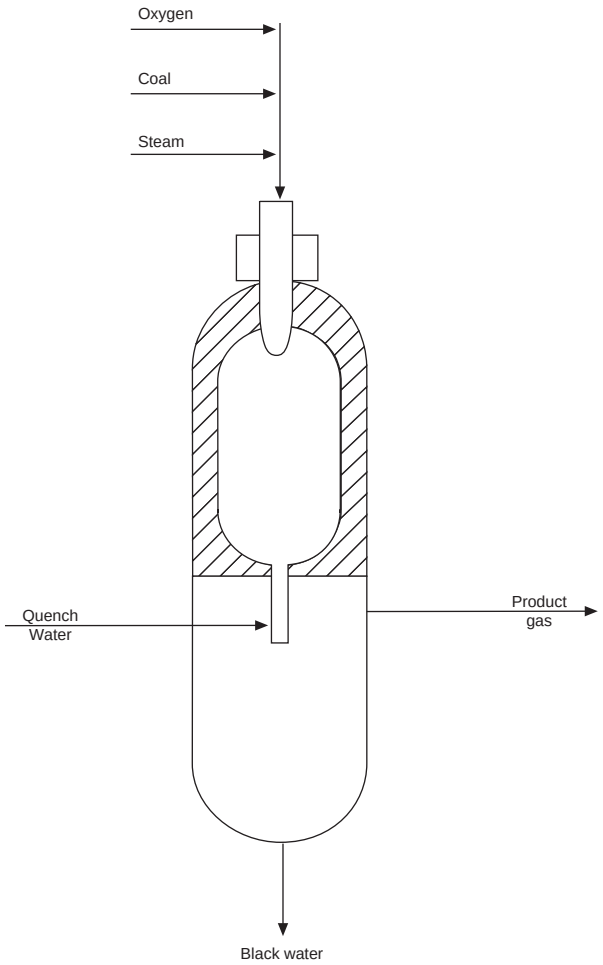


FIGURE 2 Configuration for Texaco coal gasification process.

This type of gasifier may be either single stage or two stage. In general, high temperatures (650–960°C) are used to achieve complete gasification of the coal in a mixture with steam and oxygen or air.

Entrained-flow gasifiers can handle all coals, including strongly caking coals, without pretreatment. The high temperature of operation produces a gas free of methane, oils, and tar. In two-stage gasifiers, the incoming coal is first entrained with reactant gases to produce syngas and char; the resultant char is gasified further in a second stage, which may or may not be entrained. Because the more reactive incoming coal can be gasified at a lower temperature than the less reactive char, staged operation achieves better overall thermal efficiency without sacrificing higher throughput. Entrained-flow gasifiers can be operated at atmospheric or higher pressure, and ash may be removed either dry or molten.

Many different gasifiers have been developed, at least through the demonstration stage, by a variety of organizations, both public and private. However, not all these gasifiers have achieved commercial success, and improved processes now supercede some technologies that were widely used in the past. In 1999, the Texaco, Shell, and Lurgi (dry ash) processes accounted for more than 75% of the installed and planned coal gasification capacity, with Texaco being the leader with almost 40% of installed capacity.

High reliability, acceptable capital and operating costs, and minimal environmental impact make coal gasifiers attractive for utility applications. Numerous studies confirm that gasifiers coupled with a gas turbine/steam turbine combined cycle represent one of the most promising technologies for future coal-based power generation. IGCC technologies offer the potential for high efficiencies with low pollutant emissions. High efficiencies are achieved in IGCC operation by combining efficient combustion turbines with a steam turbine bottoming cycle.

A major goal of power production is minimal environmental impact. Since the product gas from IGCC systems is purified before combustion, burning this clean fuel results in low pollutant emissions. Ash leaving the system is usually in the form of molten slag that is water quenched to form benign vitreous material suitable for a variety of uses or for disposal in a nonhazardous landfill. On balance, coal gasification systems are environmentally superior to alternative coal utilization technologies and can meet rigorous environmental standards for SO_2, NO_x, and particulates. Also, because of their increased efficiency, IGCC plants emit less CO_2, the major greenhouse gas, per unit of electricity generated.

4. LIQUEFACTION

Coal liquefaction is the process of making a liquid fuel from coal. The fundamental difference between coal, a solid, and liquid fuels is that the liquid fuels have a higher hydrogen:carbon ratio. Liquid fuels have lower ash contents and are easier to upgrade (i.e., to remove unwanted impurities such as nitrogen and sulfur). There are three major methods of producing liquid fuels from coal: pyrolysis, direct liquefaction, and indirect liquefaction.

Pyrolysis is a thermal process in which coal is heated to produce gases, liquids, and tars, leaving behind a hydrogen-depleted char. The gases, liquids, and tars are usually further treated (refined) to produce conventional liquid fuels. In direct coal liquefaction, finely divided coal is mixed with a solvent and heated under pressure in a hydrogen atmosphere and in the presence of a catalyst. After separation from ash and unconverted coal, the resulting liquids are usually further refined to produce finished products. Indirect liquefaction is a two-step process. The first step involves gasification of coal to produce a mixture of hydrogen and carbon monoxide, called synthesis gas. In the second step, the synthesis gas is heated in the presence of a catalyst to make hydrocarbons, with water being the major by-product. The second step most frequently considered is called Fischer–Tropsch synthesis. The hydrocarbon product is then refined into the desired finished products.

The conversion of coal to liquid fuels is not practiced commercially in the United States. Both direct and indirect liquefaction were practiced in Germany during World War II and were used to fuel the German war effort. Research and development efforts have been ongoing, at various levels of effort, since the 1940s to make coal liquefaction a commercially viable enterprise. The only commercial liquefaction process for manufacturing fuel is conducted in South Africa by Sasol using the Fischer–Tropsch process. Sasol's primary products are high-value chemicals in addition to fuels. Eastman Chemical operates a coal conversion plant in Kingsport, Tennessee, where each day coal is converted into 260 tons of methanol to supply Eastman's acetate synthesis plants. In 2002, China announced plans to construct a direct coal liquefaction plant.

Direct coal liquefaction can be traced back to the work of the German chemist Friedrich Bergius, who, in approximately 1913, began studying the chemistry of coal. In his studies, he attempted the hydrogenation of coal using temperatures exceeding 350°C and hydrogen under high pressure. The first commercial-scale direct liquefaction plant was constructed by I. G. Farben in Leuna in 1927. The Germans constructed several more plants during the 1930s. Some fundamental but no large-scale research was conducted in the 1940s–1960s. It was not until the 1970s that interest in direct coal liquefaction was revived.

As coal is heated in a suitable solvent, it is rapidly converted to soluble entities that have high molecular weight and an average elemental composition very similar to that of the starting feed coal. The primary chemical reaction of direct coal liquefaction is hydrogenation of the coal to break the coal into smaller molecules and increase the hydrogen: carbon ratio. Addition of hydrogen also removes the heteroatoms nitrogen, sulfur, and oxygen by converting them to ammonia, hydrogen sulfide, and water, respectively. Some combination of iron, nickel, cobalt, and molybdenum catalysts can be used to lower the required temperature and pressure and improve the rate and extent of hydrogenation.

The direct liquefaction process (Fig. 3) employs a coal ground to a fine powder, solvent, hydrogen, and catalyst employed at elevated pressures (1000–2500 psig) and temperatures (400–450°C). The solvent mixed with the coal forms a pumpable paste.

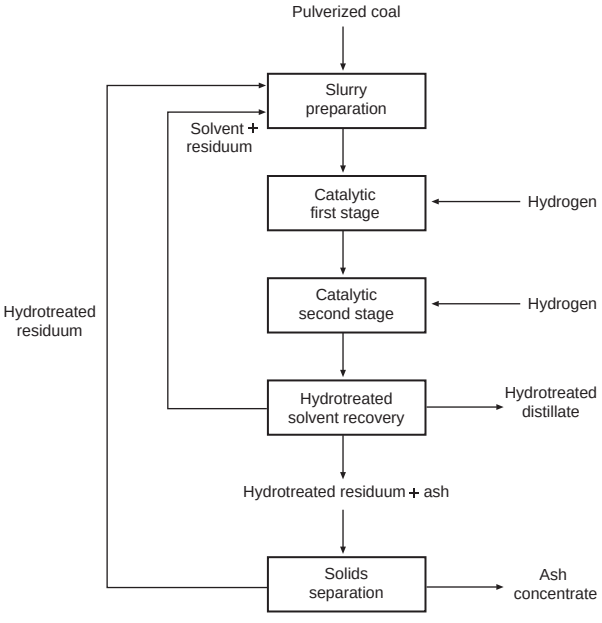


FIGURE 3 Bock diagram of advanced direct coal liquefaction process.

Intimate contact between solid coal and solvent is easier to achieve than that between coal and gaseous hydrogen; therefore, in addition to serving as vehicle and solvent, the coal liquefaction solvent also serves as a hydrogen shuttle or donor solvent. After transferring hydrogen to the coal, the hydrogen-depleted donor solvent reacts with the gaseous hydrogen, regenerating itself. Processes can be designed so that they rely primarily on dissolution, use one-stage or two-stage hydrogenation with or without catalyst addition, and use various schemes for recovery and regeneration of solvent.

If dissolution is the primary mechanism for converting coal, the product is really an ash-free solid that, if desired, can undergo further processing. When hydrogenation schemes are employed, the resulting liquid products are rich in aromatic and olefinic materials, and after separation from mineral matter and unconverted coal they need to be upgraded to finished products through conventional refinery operations such as distillation and hydrotreating. The exact product slate will depend on the type of coal used, solvent employed, and specific operating conditions.

Indirect liquefaction is the production of fuels or chemicals from synthesis gas derived from coal. Synthesis gas is a mixture of hydrogen and carbon monoxide. A hydrogen:carbon monoxide ratio of approximately 2:1 is required for hydrocarbon production. The by-product is water. In coal liquefaction, the synthesis gas is produced by coal gasification.

Franz Fischer and Hans Tropsch first reported the hydrogenation of carbon monoxide in 1925. By-products to the straight-chain hydrocarbons were a mixture of alcohols, aldehydes, and acids. Nine Fischer–Tropsch plants operated in Germany during World War II. Japan also had Fischer–Tropsch plants during World War II. In 1953, South Africa, well endowed with coal resources but lacking significant petroleum sources, began construction of a Fischer–Tropsch plant at Sasolburg, near Johannesburg. The plant is operated by the South African Coal Oil and Gas Corporation, commonly known as Sasol. In 1973, Sasol began construction of a second plant, called Sasol-2, at Secunda. Eventually, a third plant, Sasol-3, came online in 1989. No commercial Fischer–Tropsch plants have been built since, although there is considerable interest in manufacturing fuels from synthesis gas obtained by the steam reforming of methane.

The Fischer–Tropsch reaction occurs over a catalyst, typically iron or cobalt based. The Fischer–Tropsch reaction is basically a chain polymerization reaction followed by hydrogenation. The reactions are conducted at pressures ranging from 1 to 30 atm and at 200–350°C. The growth of the linear chains follows an Anderson–Schultz–Flory (normal) distribution and can be described by the term α. When $\alpha = 0$, methane is the product. As α approaches 1, the product becomes predominantly high-molecular wax. Iron catalysts may be employed when the gasification step produces a synthesis gas with insufficient hydrogen because iron is known to promote the water gas shift reaction, thus synthesizing additional hydrogen *in situ*.

The Fischer–Tropsch reaction is highly exothermic. The first commercial units used fixed bed shell and tube reactors. Because of the excess heat produced, the diameter of the tubes is limited. Tube packing is also cumbersome. By using fluidized beds, tubes could be made larger and temperature control still maintained. However, the fluidized bed requires a higher operating temperature (340°C) and deposition of products on the catalysts disrupts fluid bed operations. A newer concept, the slurry bubble contact reactor (SBCR), uses finely divided catalyst dispersed in inert liquid while reactive gases are introduced from the bottom of the reactor. Heat exchanger tubes provide temperature control. The SBCR offers high throughput and simplicity of design. Sasol commissioned a commercial SBCR (2500 bbl/day) in 1993.

The resulting hydrocarbon products need to be separated from oxygenated by-products and further refined into finished fuels or petrochemical feedstocks. When wax is a predominant product, the wax can be cracked to produce high-quality diesel fuels and jet fuels. Because the gasification process allows for ready removal of contaminants from the synthesis gas product, the resulting Fischer–Tropsch products have sulfur levels below the detection limits. Olefinic products can be used in place of petrochemical feedstocks.

As suggested by the discussion on Fischer–Tropsch synthesis, synthesis gas can also be converted to oxygenated liquids, including methanol, higher alcohols, and ketones. In the presence of a catalyst, often a zinc–chromium or copper–zinc catalyst, synthesis gas can be converted to methanol. The coal conversion plant at Kingsport, Tennessee, combines coal gasification with methanol synthesis in a SBCR. Methanol can be used as a fuel or as a chemical feedstock. It can also be dehydrated and polymerized over a zeolite catalyst to produce gasoline-like products.

SEE ALSO THE FOLLOWING ARTICLES

Clean Coal Technology • Coal, Chemical and Physical Properties • Coal, Fuel and Non-Fuel Uses • Coal Industry, Energy Policy in • Coal Industry, History of • Coal Mine Reclamation and Remediation • Coal Mining, Design and Methods of • Coal Mining in Appalachia, History of • Coal Preparation • Coal Resources, Formation of • Coal Storage and Transportation • Markets for Coal

Further Reading

Alpert, S. B., and Wolk, R. H. (1981). Liquefaction processes. *In* "Chemistry of Coal Utilization" (M. A. Elliott, Ed.), 2nd Suppl. Vol., pp. 1919–1990. Wiley, New York.

Bartok, W., and Sarofim, A. F. (eds.). (1993). "Fossil Fuel Combustion—A Source Book." Wiley, New York.

Bartok, W., Lyon, R. K., McIntyre, A. D., Ruth, L. A., and Sommerlad, R. E. (1988). Combustors: Applications and design considerations. *Chem. Eng. Prog.* **84,** 54–71.

Berkowitz, N. (1979). "An Introduction to Coal Technology." Academic Press, New York.

Brown, T., Smith, D., Hargis, R., and O'Dowd, W. (1999). Mercury measurement and its control: What we know, have learned, and need to further investigate. *J. Air Waste Management Assoc.* **49,** 628–640.

Derbyshire, F. (1986). Coal liquefaction. In "Ullmann's Encyclopedia of Industrial Chemistry," Vol. A7. VCH, New York.

El Sawy, A., Gray, D., Talib, A., and Tomlinson, G. (1986). A techno-economic assessment of recent advances in direct coal liquefaction, Sandia Contractor Report No. SAND 86–7103. Sandia National Laboratory, Albuquerque, NM.

Elliott, M. A. (ed.). (1981). "Chemistry of Coal Utilization." Wiley–Interscience, New York.

Gavalas, G. R. (1982). Coal pyrolysis. *In* "Coal Science and Technology," Vol. 4. Elsevier, Amsterdam.

Howard, J. B. (1981). Fundamentals of coal pyrolysis and hydropyrolysis. *In* "Chemistry of Coal Utilization" (M. A. Elliott, Ed.), 2nd Suppl. Vol., pp. 665–784. Wiley, New York.

Juntgen, H., Klein, J., Knoblauch, K., Schroter, H., and Schulze, J. (1981). Liquefaction processes. *In* "Chemistry of Coal Utilization" (M. A. Elliott, Ed.), 2nd Suppl. Vol., pp. 2071–2158. Wiley, New York.

Khan, M. R., and Kurata, T. (1985). "The feasibility of mild gasification of coal: Research needs, Report No. DOE/METC-85/4019, NTRS/DE85013625." Department of Energy, Washington, DC.

Penner, S. S. (1987). Coal gasification: Direct applications and synthesis of chemicals and fuels: A research needs assessment, Report No. NTIS-PR-360. Office of Energy Research, Washington, DC.

Probstein, R. F., and Hicks, R. E. (1982). "Synthetic Fuels." McGraw-Hill, New York.

Richardson, F. W. (1975). "Oil from Coal." Noyes Data, Park Ridge, NJ.

Schobert, H. H. (1987). "Coal: The Energy Source of the Past and Future." American Chemical Society, Washington, DC.

Seglin, L., and Bresler, S. A. (1981). Low-temperature pyrolysis technology. *In* "Chemistry of Coal Utilization" (M. A. Elliott, Ed.), 2nd Suppl. Vol., pp. 785–846. Wiley, New York.

SFA Pacific (1993). Coal gasification guidebook: Status, applications and technologies, Research Project No. 2221-39. EPRI, Palo Alto, CA.

Singer, J. G. (ed.). (1991). "Combustion Fossil Power—A Reference Book on Burning and Steam Generation." Combustion Engineering, Windsor, Ontario, Canada.

Solomon, P.R., and Serio, M. A. (1987). Evaluation of coal pyrolysis kinetics. *In* "Fundamentals of Physical Chemistry of Pulverized Coal Combustion" (J. Lahaye and G. Prado, Eds.). Nijhoff, Zoetermeer.

Stultz, S. C., and Kitto, J. B. (1992). "Steam—Its Generation and Use." Babcock & Wilcox, Barberton, OH.

U.S. Department of Energy (1989). "Coal liquefaction: A research and development needs assessment, DOE Coal Liquefaction Research Needs (COLIRN) Panel Assessment Final Report, Vol 2, DOE/ER-0400." U.S. Department of Energy, Washington, DC.

U.S. Department of Energy, Morgantown Energy Technology Center (1991). "Atmospheric Fluidized Bed Combustion—A Technical Source Book," Publication No. DOE/MC/1435-2544. U.S. Department of Energy, Morgantown Energy Technology Center, Morgantown, WV.

U.S. Department of Energy, National Energy Technology Laboratory (1987). "Clean Coal Technology Compendium." U.S. Department of Energy, National Energy Technology Laboratory, Pittsburgh, PA.

Coal, Fuel and Non-Fuel Uses

ANTON DILO PAUL
Science Applications International Corporation
Pittsburgh, Pennsylvania, United States

1. Coal Use History
2. Current Fuel Use of Coal
3. Environmental Issues Associated with Coal Use
4. Current Non-Fuel Uses of Coal
5. Coal Combustion By-Products
6. Modern Coal Plants
7. The Future of Coal Use

Glossary

acid rain Precipitation containing small amounts of sulfuric acid and nitric acid, which derive from sulfur and nitrogen oxides emitted during coal combustion.

boiler A device that utilizes the heat from a furnace to convert water into high-pressure steam. When steam production is the sole objective, the boiler is usually integrated with the furnace and the two terms are used synonymously in industry parlance.

coal combustion emissions Gaseous pollutants released during the combustion of coal.

coal combustion products (CCPs) Residual material from the combustion of coal. These materials were disposed of as waste in the past, but are now used in other industries.

cogeneration The use of coal (or other fuel) to produce electricity, whereby the residual heat is made available for external use.

furnace/hearth A device that burns coal efficiently using forced air circulation inside a closed container.

greenhouse gases (GHGs) Various naturally occurring atmospheric gases (particularly carbon dioxide) that are also emitted during coal combustion. Excess quantities are suspected to cause changes in weather patterns and an overall warming of the earth.

steam engine/steam turbine A device that converts the heat energy in steam into mechanical energy.

The use of coal ranges from simple combustion for heat to complex partial oxidation to produce heat, gaseous/liquid fuels, and chemical feedstocks. Heat produced from coal combustion was traditionally used for home cooking and heating. From the dawn of the Industrial Revolution until about the 1950s, coal was the primary fuel to supply process heat for industrial processes and transportation engines. Apart from its use as a fuel, coal provided chemical raw materials that were the starting blocks for many household and industrial chemical products. Unrestrained coal use during this period, however, soon led to severe air, land, and water pollution. Environmental legislation commenced in the 1970s in the United States to limit the emissions of air pollutants from coal. Prior to this, oil began to replace coal as a transportation fuel, and natural gas competed with coal as a heating fuel. These factors significantly reduced coal use, but the world still consumes about 5 billion tons/year, largely for the generation of baseload electric power. With new coal combustion technologies, coal use today is environmentally clean and shows promise of becoming even cleaner in the future.

1. COAL USE HISTORY

1.1 Home Heating and Cooking

Documented coal use dates back to the 1700s, when coal was used as a heating fuel, but historical evidence indicates that coal was used in 100–200 AD by the Romans during their occupation of England. Until about the 18th century, cooking took place in an open hearth with firewood as fuel. Hearth cooking was hard work and messy; trees had to be chopped into firewood and brought indoors and large amounts of fine-sized ash had to be hauled outdoors. Coal was a welcome relief, because its high heating value per unit mass (7000–13,000 Btu/lb) compared to wood meant that less fuel was needed. However, coal is hard to ignite and, without sufficient air supply, produces a smoky flame, but it can be burned inside a closed container with a controlled airflow to produce a much hotter flame

 © 2004 Elsevier Inc. All rights reserved.

than is produced by wood fuel. Using this principle, the coal-fired pot-bellied stove made its appearance in almost every home and building during the 1800s and early 1900s. Coal stoves not only provided heat for cooking, but also warmed the room in which they were situated. They were often hooked up to a hot-water tank to provide running hot water as well as steam for steam baths.

1.2 Coal Use during the Industrial Revolution

The Industrial Revolution of the 18th and 19th centuries greatly expanded the use of coal as a fuel to run the new industries, particularly for the manufacture of iron products. Prior to the Industrial Revolution, wood charcoal, or char, was the source of heat in industrial hearths. Wood char provided more heat compared to wood, but even so, char hearths rarely produced enough heat to melt iron. Instead, these hearths could only heat iron until it became sufficiently malleable to be hammered into shapes, yielding wrought-iron products. On the other hand, coal hearths provided enough heat to melt iron into its liquid form. Liquid iron could then be poured into special molds to manufacture cast-iron products—a faster way of producing iron goods. Coal hearths speeded operations in other heat-intensive industries, such as the smelting of iron ore in a blast furnace or the production of steel from pig iron. Regions where coal and iron ore were present together, for example, South Wales in England, flourished during the Industrial Revolution because of the abundance of both raw materials in one location.

1.3 Coke and Gas Production

Another historic use of coal was for street lighting. Coal was heated in ovens with minimal air to produce "town gas" or "manufactured gas." The main components of town gas are carbon monoxide and hydrogen. The gas was piped throughout a town in underground pipes and burned in street lamps for illumination. Town gas was common in many European cities, notably London, even up to the 1950s. The residue from the production of town gas is coke, which is nearly pure carbon (except for the ash). Coke and the by-products of coke manufacture established an important role in the iron/steel and chemical manufacturing industry from 1850 to 1950. (For further discussion on this topic, see Section 4.1)

Early coke-making processes used "beehive" ovens that vented unburnt volatile matter in coal into the atmosphere. Such coke-making practices produce severe air pollution and are nonexistent today in the United States and other developed countries. The more common practice is to recover all emissions from a coke oven and distill them into different fractions. These fractions contain many chemicals (including naphthalene, ammonia, methanol, and benzene) that form the basis or feedstock for many consumer products. Aspirin (acetylsalicylic acid), road tar, creosote, fertilizers, and disinfectants are a few examples of products traditionally manufactured from coal. A comprehensive depiction of the many products that can be produced from coal is seen in the "coal tree," which may be accessed on the Appalachian Blacksmiths' Association web site at http://www.appaltree.net/aba/coalspecs.htm.

1.4 The Rise and Fall of "King" Coal

The invention of the steam engine by James Watt in the late 1700s greatly expanded the market for coal, because it provided a method for converting the heat energy in coal into mechanical energy. The golden era of the railroads dawned in the early 1900s as coal-fired steam locomotives and steam wagons quickly replaced horse-drawn wagons and carriages. In the locomotive's boiler, a coal fire heated water and converted it to steam. The energy in the steam moved back and forth a piston that was attached to the locomotive's wheels, providing motive power. Coal fueled many forms of transport, from giant locomotives that hauled goods and passengers hundreds of miles across cities, to steam wagons and streetcars that transported the same cargo across a few city blocks. Compared with traveling on horseback or stagecoach, train travel was comfortable and fast—passengers relaxed in Pullman cars heated by pot-bellied stoves, dined on meals cooked over coal-fired stoves, and traveled in fast trains pulled by steam locomotives.

Stationary coal-fired steam engines provided mechanical power to drive factory machinery through an elaborate system of pulleys. The availability of mechanical power led to the development of machine tools that rapidly increased industrial productivity. The construction industry boomed as steam shovels moved tons of earth for new buildings and steamrollers smoothed bumpy roads into highways. The Panama Canal, the Hoover Dam, the Empire States Building, and the first few U.S. highways could not have been built during their time without these coal-fired machines.

From the reciprocating steam engine evolved the steam turbine, in which steam impinges on the vanes

on a shaft to produce rotary motion instead of moving a piston back and forth. The steam turbine ran smoother than the reciprocating engine and was more suited for marine propulsion and electric power generation. Shipbuilders replaced masts and sails with smoke funnels and built larger and faster ships geared toward transatlantic travel. The first crossing of the Atlantic Ocean solely under steam power occurred in 1838 when two British steamships, Great Western and Sirius, made the trip in 15 days, compared to 30–45 days in a sail ship. This started the era of the luxurious ocean liners, including the famed Titanic, which were powered by coal carried in gigantic bunkers within the ship's hull.

Coal also fueled large furnaces for glass making, to process wood pulp, to produce iron and steel, and to heat lime and cement kilns. During the late 1800s to the 1950s, coal was the world's primary heating and transportation fuel and the backbone of several interdependent industries. Steam shovels helped mine coal and mineral ores that were transported by steam trains to blast furnaces and coke ovens. Coal-fired furnaces processed the ores into metals and finished metal goods. These goods were transported to their final destination by coal-powered trains and steam wagons.

Coal's reign as a premium fuel diminished with the discovery of crude oil. By refining crude oil, liquid fuels such as gasoline and diesel became available to power internal combustion engines. Diesel–electric locomotives, automobiles, trucks, and buses soon became the common mode of transportation, replacing steam-driven vehicles and eventually the use of coal as a direct transportation fuel. Oil and natural gas became the preferred fuels for heating and cooking in residential and commercial buildings. By the late 1970s, coal's use as a transportation fuel had diminished significantly in most developed countries, but it remained as a fuel for generating electric power, particularly in countries with large minable coal reserves. Fig. 1 shows the historic change in coal use for transportation and residential and industrial heating in the United States since 1949.

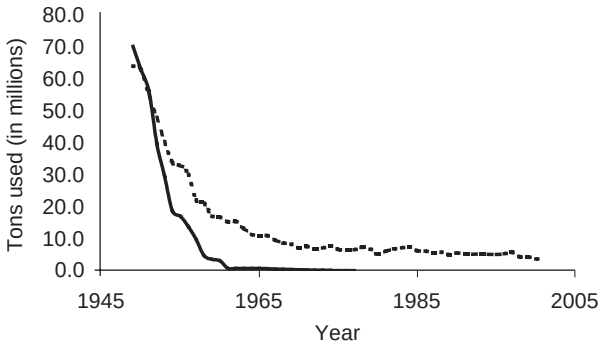


FIGURE 1 Decrease in coal use in the United States. Solid lines, transportation use; dashed lines, residential and commercial use. Data from Energy Information Agency, Annual Energy Review (2001), Section 7—Coal.

2. CURRENT FUEL USE OF COAL

2.1 Generation of Electric Power

The largest use of coal today is in the generation of electric power by electric utilities. According to the U.S. Energy Information Administration, about 65% of coal mined in the world is used for power generation. The next largest use of coal is in the production of coke for the iron and steel industries. Coal is still used for industrial heating and even commercial and residential heating in certain countries.

When coal is used to generate electric power, it is combusted in large furnaces. The heat generated is transferred to tubes inside the furnace. Water in these tubes is converted to high-pressure steam, which is piped to steam turbines that convert the thermal energy in the steam into mechanical (rotational) energy. An electric power generator connected to the turbine converts the mechanical energy into electric power, which is transmitted over high-voltage lines to a power network and is sold to industrial, commercial, and residential users. The spent (low-pressure) steam from the turbines is condensed into water and is returned to the boiler. Fig. 2 is a simplified schematic depicting the use of coal for electric power generation.

The conversion of coal to electric power can be viewed as a convenient use of coal. It is easier to burn coal continuously in a single large furnace than it is to burn it intermittently in many small home stoves. Electric power obtained from converting the heat energy derived from a large coal furnace provides a method for distributing coal energy to multiple users. In homes that depend on coal for heating and cooking, electricity generated from coal provides the same energy, but without the dust, soot, and physical labor associated with firing and stoking a coal fire. For transport systems, electricity generated from coal can move electric trains in a far more efficient and less polluting manner compared to burning coal in several steam locomotives. Hence, coal still plays a role in the transportation sector today, but in an indirect manner through the generation of electricity.

2.2 Coproduction of Heat and Electricity

After utility company power production, the next largest use of coal is for the production of heat and electricity for commercial buildings and industrial processes. Rather than purchase electricity from a central grid, an industrial plant or building owner may choose to operate a boiler that supplies both heat and electric power. The concept of cogeneration, or combined heat and power (CHP) production, is similar to the use of coal for electric power production, except that the spent (low-pressure) steam from the turbines is piped to an industrial plant or building before being condensed and returned to the boiler. Residual heat in the low-pressure steam supplies process heat for industrial application or temperature control—for example, to maintain a greenhouse at constant temperature so that vegetables and flowers can be grown during cold seasons. Cogeneration plants are common in Europe, where they are used for district heating and generating electricity. In the United States, cogeneration plants are common on industrial sites, hospitals, and college and university campuses. Table I illustrates coal consumption in the United States by end-use application.

2.3 Home Heating and Cooking

Coal is still used to a small extent for home heating and cooking. In the homes of more affluent nations,

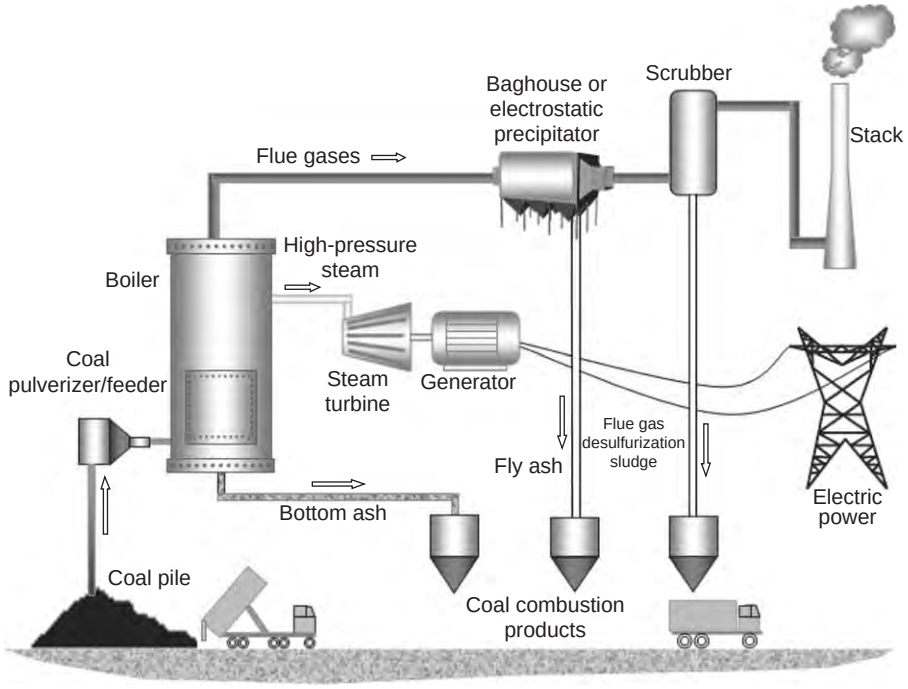


FIGURE 2 Schematic diagram illustrating coal use for electric power generation.

TABLE I

U.S. Coal Consumption by End-Use Sector, 1996–2002[a]

Year	Electric utilities	Coke plants	Other industrial	Residential and commercial	Other power products	Total
1996	874,681	31,706	71,689	6006	22,224	1,006,306
1997	900,361	30,203	71,515	6463	21,603	1,030,114
1998	910,867	28,189	67,439	4856	26,941	1,038,292
1999	894,120	28,108	64,738	4879	52,691	1,044,536
2000	859,335	28,939	65,208	4127	12,3285	1,080,894
2001	806,269	26,075	63,631	4127	15,0637	1,050,470

[a] In thousands of short tons. Data adapted from the Energy Information Agency Review (2001), Coal Consumption Statistics.

coal is used for recreational or nostalgic reasons, rather than for routine heating and cooking. Coal stoves designed for this purpose are aesthetically attractive and are engineered for efficient operation with near-zero indoor air pollution. In contrast, coal used for cooking and heating in developing countries is burned in crude stoves with inadequate ventilation. With no chimneys or hoods to vent the smoke, emissions inside homes can quickly reach levels that cause respiratory diseases. Even when stoves are vented, smoke emissions significantly increase local air pollution. In rural South Africa, for example, improperly designed or damaged stoves, coupled with poor ventilation, produce high risks to health and safety, including high levels of carbon monoxide. Robert Finkelman from the U.S. Geological Survey has reported that over 3000 people in the Guizhou Province of southwest China suffer from severe arsenic poisoning. Coal from this region contains high levels of toxins, including arsenic, which accumulates in chili peppers and other foodstuffs that are dried over unvented coal stoves. Polycyclic aromatic hydrocarbons, released during unvented coal combustion have also been cited as a primary cause for lung cancer. The World Bank, the World Health Organization, and the U.S. Geological Survey are among the many organizations that are educating developing and less developed nations about the dangers of unvented coal combustion.

3. ENVIRONMENTAL ISSUES ASSOCIATED WITH COAL USE

3.1 Emissions from Coal Combustion

As discussed in the previous section, coal emissions from poorly designed stoves are hazards to human health. However, coal emissions from large furnaces are equally hazardous. From the time of the industrial revolution until about the 1970s, little attention was paid to the air pollution created by coal combustion. Meanwhile, coal-fired power plants and industrial boilers spewed out tons of gaseous and particulate pollutants into the atmosphere. During combustion, the small amounts of sulfur and nitrogen in coal combine with oxygen to form sulfur dioxide (SO_2), sulfur trioxide (SO_3), and the oxides of nitrogen (NO_x). If these gases are not captured, they are released into the atmosphere through the power plant's stack. They combine with water vapor in the upper atmosphere to form droplets of sulfuric and nitric acids, which return to the earth's surface as acid rain. Acid rain is blamed for killing fish in streams and lakes, damaging buildings, and upsetting the ecological balance.

Flue gas from a coal plant's stack contains tiny particles of mineral matter (fly ash) and unburned carbon (soot). When fog and smoke are trapped beneath a cold air mass, they form smog, which can sometimes be deadly. For example, two killer smogs occurred in London, England in December of 1892 and 1952, lasting 3–5 days and resulting in over 5000 deaths. Menon and co-workers studied climate changes caused by coal fires; they found that soot from these fires absorbs sunlight and heats the surrounding air, but blocks sunlight from reaching the ground. Heated air over a cool surface makes the atmosphere unstable, creating heavy rainfall in some regions and droughts in others.

Smog and air pollution from coal-fired plants are responsible for many health problems observed in people who live in highly industrialized areas. Examples of highly industrialized areas include regions of Poland, the Czech Republic, and Germany, comprising the Black Triangle, one of Europe's most heavily industrialized and polluted areas; England's Black Country, a 19th century industrialized area west and north of Birmingham; and in the United States, Pittsburgh, Pennsylvania, which supplied most of the steel needed for World War II. Choking coal dust and toxic gas emissions from coke plants, blast furnaces, and coal-fired electric power and district heating plants once plagued the residents of these areas with respiratory problems, heart disease, and even hearing losses. Fortunately, most of the coal-based heavy industries that polluted these areas have been cleaned up using modern coal-burning and emission control technologies.

3.2 Mitigation of Coal Combustion Emissions

A combination of legislation and technology has helped clean up many of the world's coal-burning plants. Both developed and developing countries have adopted increasingly stringent environmental regulations to govern emissions from coal-fired power plants. In the United States, all coal-fired power plants built after 1978 must be equipped with postcombustion cleanup devices to capture pollutants before they escape into the atmosphere. Cyclones, baghouses, and electrostatic precipitators filter out nearly 99% of the particulates. Flue gas scrubbers use a slurry of crushed limestone and water to absorb sulfur oxides from flue gas. The limestone

reacts with the sulfur dioxide to form calcium sulfate, which may be used to produce wallboard. Staged combustion and low-NO_x burners are used to burn coal to minimize NO_x formation. Another strategy, selective catalytic reduction, reacts ammonia with NO_x over a catalyst to produce nonpolluting nitrogen and water vapor.

Conventional coal-fired power plants capture pollutants from the flue gas after it leaves the boiler. Circulating fluidized bed (CFB) combustors capture most of the pollutants before they leave the furnace. Crushed coal particles and limestone circulate inside the CFB combustor, suspended by an upward flow of hot air. Sulfur oxides released during combustion are absorbed by the limestone, forming calcium sulfate, which drops to the bottom of the boiler. The CFB combustor operates at a lower temperature (1400°F) compared to pulverized coal (PC) boilers (2700°F), which also helps reduce the formation of NO_x.

Precombustion coal cleaning is another strategy to reduce sulfur emissions by cleaning the coal before it arrives at the power plant. Sulfur in coal is present as pyrite (FeS_2), which is physically bound to the coal as tiny mineral inclusions, and as "organic sulfur," which is chemically bound to the carbon and other atoms in coal. Pyrite is removed in a coal preparation plant, where coal is crushed into particles less than 2 inches in size and is washed in a variety of devices that perform gravity-based separations. Clean coal floats to the surface, whereas pyrite and other mineral impurities sink. Additional cleaning may be performed with flotation cells, which separate coal from its impurities based on differences in surface properties. Precombustion removal of organic sulfur can be accomplished only by chemical cleaning. So far, chemical cleaning has proved to be too costly, thus flue gas scrubbers are often required to achieve near-complete removal of sulfur pollutants.

The tightening of environmental regulations is likely to continue throughout the world. In the United States, for example, by December 2008, it is anticipated that coal-fired power plants will have to comply with maximum emission levels for mercury. Emissions of mercury and other trace metals, such as selenium, are under increasing scrutiny because of suspected adverse effects on public health.

3.3 Cofuel Applications

Coal is sometimes combusted with waste material as a combined waste reduction/electricity production strategy. The disposal of waste from agriculture and forestry (biomass), municipalities, and hospitals becomes costly when landfill space is limited. Some wastes, particularly biomass, are combustible, but their low energy density (compared with coal) limits their use as an electricity production fuel. Blending coal with these fuels provides an economical method to produce electric power, reduce waste, and decrease emissions. Most wood wastes, compared to coal, contain less fuel nitrogen and burn at lower temperatures. These characteristics lead to lower NO_x formation. In addition, wood contains minimal sulfur ($<0.1\%$ by weight) and thus reduces the load on scrubbers and decreases scrubber waste.

Numerous electric utilities have demonstrated that 1–8% of woody biomass can be blended with coal with no operational problems. Higher blends may also be used, but require burner and feed intake modifications as well as a separate feed system for the waste fuel. Cofiring in fluidized bed boilers may avoid some of these drawbacks, but the economics of cofiring are not yet sufficiently attractive to make it a widespread practice.

4. CURRENT NON-FUEL USES OF COAL

4.1 Coke Manufacture

Select coals are used to manufacture coke; the high pure carbon content, porous structure, and high resistance to crushing of coke have made it a necessary material in the production of molten iron (or hot metal, in iron/steel industry parlance). Coke also serves as a fuel in blast furnaces, as well as serving several nonfuel functions. In a typical blast furnace, iron ore, coke, and other materials, including sinter and fluxes, are loaded in layers. Air, preheated to 900°C, is blasted through the bottom of the furnace through nozzles called tuyeres. Oxygen in the air reacts with coke to form carbon monoxide and generate heat. Carbon monoxide reduces iron ore to hot metal, which flows downward and is tapped off at the bottom of the blast furnace. Coke provides physical support for the materials in a blast furnace by maintaining sufficient permeability for the upward flow of gases and the downward flow of hot metal.

During the past century, the coke industry grew concurrently with the growth of the iron and steel industry. Large integrated steel mills operated coke-making plants, blast furnaces, and basic oxygen furnaces that converted iron ore into steel. However, coke manufacture is a highly polluting operation,

and coke cost is one of the major expenses of operating a blast furnace. Consequently, methods of producing steel without coke are under development and some are in operation today. The most straightforward approach is the recycling of steel using electric arc furnaces (also known as minimills) that melt and process scrap steel for reuse, thereby dispensing with the need for fresh steel. Direct reduced iron (DRI) is another process that reduces coke need by directly reducing iron ore without the use of coke. Another process is the direct injection of granulated coal into a blast furnace. Such a system is in operation at Bethlehem Steel's Burns Harbor Plant in the United States as well as in other locations throughout the world.

4.2 Coal-Derived Solid Products

The most common solid product derived from coal is activated carbon. This product is used as a filter medium for the purification of a wide range of products, such as bottled water, prepared foods, and pharmaceuticals. Activated carbon plays an important role in the treatment of contaminated ground water by absorbing toxic elements. Impregnated activated carbons are used for more complex separations, such as the removal of radioactive isotopes, poisonous gases, and chemical warfare agents.

Activated carbon is produced from premium coals, or a form of coal known as anthracite, but other materials, including coconut shells, palm-kernel shells, and corncobs, may also be used. Coal, or other feed material, is first heated in a controlled atmosphere to produce a char, which is then thermally or chemically activated to yield a highly porous final product. Surface areas of activated carbons range from 500 to 1400 m^2/g, with macropores ($>0.1\,\mu m$ in diameter) and micropores ($0.001–0.1\,\mu m$ in diameter). Macropores provide a passageway into the particle's interior, whereas micropores provide the large surface area for adsorption. The starting material and the activation process greatly influence the characteristics and performance of the finished activated carbon. Coal, for example, produces highly macroporous activated carbons and is the preferred base material for many impregnated activated carbons.

4.3 Coal-Derived Liquid Products

Coal-derived liquid products have been traditionally obtained as by-products of coke manufacture. These products contain chemicals that form the basis or feedstock for many consumer products. In a coke plant, the volatile components emanating from coal are collected as unpurified "foul" gas, which contains water vapor, tar, light oils, coal dust, heavy hydrocarbons, and complex carbon compounds. Condensable materials, such as tar, light oils, ammonia, and naphthalene, are removed, recovered, and processed as chemical feedstocks. However, both the availability and the market for coke by-products are declining. The highly polluting nature of coke manufacture requires costly environmental controls to comply with tightening environmental regulations, and some steelmakers have found it more economical to close down their coke plants and import coke from countries with less stringent environmental regulations. Consequently, the availability of coking by-products has diminished significantly in the United States and in Europe, thereby limiting the production of liquids from coal. More importantly, products from petroleum refining are a more cost-effective method to obtain the various chemical feedstocks that were formerly obtained from coal.

Another liquid product from coal is obtained via the direct or indirect liquefaction of coal. In direct liquefaction, coal is heated under pressure with a solvent and a catalyst to produce a liquid, which can then be refined into other liquids, including transportation fuels (gasoline, diesel fuel, and jet fuel), oxygenated fuel additives (methanol, ethers, and esters), and chemical feedstocks (such as olefins and paraffin wax). With indirect liquefaction, coal is first gasified in the presence of oxygen and steam to generate a gas containing mostly carbon monoxide and hydrogen (i.e., syngas). Next, syngas is cleaned of its impurities and converted into liquid fuels and chemical feedstocks.

The conversion of coal into liquid products is not widely practiced on a commercial scale. For most countries, it is more economical at present to use crude oil to produce liquid fuels than to produce them from coal; hence, coal liquefaction does not constitute a large part of coal use today. The exceptions are two commercial coal liquefaction operations, namely, the Sasol Chemical Industries plant in South Africa and the Eastman Chemical plant in the United States, that use coal liquefaction to produce high-value chemicals.

4.4 Carbon Products

Coal is being tested to produce carbon products, using both coal and the liquid products derived from coal liquefaction. Carbon products are becoming

increasingly common over a wide spectrum of applications, ranging from sports equipment to engineered materials for fighter aircraft and space shuttles. In addition, carbon materials such as carbon fibers and carbon foams are finding new applications and are becoming cost-effective alternatives to steel in automotive and aircraft components as well as in many other industries.

Starting feedstocks for carbon products are presently obtained as by-products from the oil-refining industry and the manufacture of coke. These include impregnating pitches, mesophase pitches, needle cokes, and isotropic cokes. However, several processes are under investigation to produce such materials directly from coal. Chemical solvents, including *N*-methylpyrrolidone and dimethyl sulfoxide, are used to dissolve the carbon in coal. The resulting gel-like substance is filtered and separated from the mineral matter present in coal, after which it is hydrogenated and processed to produce binder and impregnation pitches, anode and needle cokes, fibers, foams, specialty graphites, and other carbon products. The Consortium for Premium Carbon Products from Coal (CPCPC) is one organization that promotes work on the production of high-value carbon products from coal.

5. COAL COMBUSTION BY-PRODUCTS

Coal used in an electric power-producing application generates several by-products, including fly ash, bottom ash, and flue gas desulfurization sludge. Coal combustion products (CCPs) were historically considered waste and were dumped in landfills. However, with the growing scarcity of landfill space, CCPs have found applications as low-cost additives in the manufacture of cement, roofing tiles, structural fills, and fertilizers.

CCPs contain mineral elements similar to those in the earth's crust, including silicon dioxide, aluminum oxide, iron oxide, calcium oxide, and trace amounts of sulfur trioxide, sodium oxide, and potassium oxide. These elements can act as replacement for natural materials and offer the environmental benefit of avoiding the need to mine additional natural resources. The three most widely used CCPs today are the residue from flue gas desulfurization (FGD), commonly known as FGD sludge, fly ash, and bottom ash.

Flue gas desulfurization residue is produced in flue gas scrubbers by the reaction of lime with sulfur oxides. It is chemically similar, but not identical in its physical properties, to natural gypsum. Because of this, it has limited application as a raw material in the manufacture of wallboard. It has also found application as a soil conditioner and fertilizer. It has been stabilized with fly ash and used as roadbase material in highway construction.

Fly ash refers to the fine particles suspended in the flue gas that are captured by electrostatic precipitators and baghouses. Fly ash can be used as a direct replacement for Portland cement or as an additive in the manufacture of cement. A challenge in the use of fly ash in cement manufacture is the presence of unburned carbon, which is present in the fly ash of most boilers that use staged combustion. Carbon makes fly ash unusable as a cement additive because it creates color problems and reduces concrete durability under freeze and thaw conditions. Research efforts are underway to remedy this problem.

Bottom ash is the coarse, hard, mineral residue that falls to the bottom of the furnace. Bottom ash and fly ash have similar chemical compositions, but bottom ash is of larger size and lacks cementitious properties. It is used as an aggregate or filler material in the construction industry. Another form of bottom ash, produced by older, slagging combustors, is hard and glassy and has found a market as a low-silica blasting grit.

6. MODERN COAL PLANTS

Coal use today is no longer evocative of dirty power plants with polluting black smoke billowing from their smokestacks. Many of these plants have been transformed through technology to operate more efficiently and with significantly lower emissions. Some fire coal with other waste materials and others produce both electric power and heat. Cases of plant retrofits and their new performance statistics are documented by various institutions, including the Energy Information Administration (http://www.eia.doe.gov) and the World Coal Institute (http://www.wci-coal.com). The following examples highlight clean coal use throughout the world:

- The Belle Vue power plant in Mauritius cofires coal with bagasse to generate electricity in an environmentally clean fashion. Bagasse, the residue after extracting juice from sugar cane, is available in large quantity in Mauritius. Instead of being a waste product incurring disposal costs, bagasse is now cofired with coal to produce electric power.

- The Turów power plant located in Europe's formerly polluted Black Triangle now operates as a clean power plant using flue gas scrubbers, low-NO_x burners, and electrostatic precipitators.
- A coal-fired cogeneration plant in Northern Bohemia now provides clean electric power to the Skoda automobile production plant and steam for district heating of the town of Mladá Boleslav.
- A 1970 coal plant in Ulan Bator, Mongolia was refurbished in 1995 with higher efficiency boilers and improved emission controls. The plant now provides electric power and heat for district heating with significantly lower pollution emissions and failure rates compared to its pre-1995 operation.
- The 35-year-old Northside Generating Station in Jacksonville, Florida in the United States was renovated under the U.S. Clean Coal Technology Program and is now ranked as the cleanest burning coal power plant in the world. Using circulating fluidized bed combustors and additional pollution controls, the plant cuts down on emissions by nearly 98% and generates 2.5 times more electric power.
- The Mirant-Birchwood Power Plant Facility in Virginia in the United States produces 238 MW of electric power as well as process steam, which is used to heat local greenhouses. In addition, it converts 115,000 tons of flue gas scrubber ash into 167,000 tons of lightweight aggregate for use in the manufacture of lightweight masonry blocks or lightweight concrete.

7. THE FUTURE OF COAL USE

7.1 Coal Use Trends

According to the U.S. Energy Information Administration, the world consumes around 5 billion short tons of coal annually, but coal use has been in a period of slow growth since the 1980s. This is largely because of growing environmental concerns, particularly in the developed countries. Coal used for coke manufacture is expected to decline due to advances in steel production technology, steel reuse via electric arc furnaces, and the replacement of steel by plastics and carbon composites.

Coal use is expected to decline in Europe and the former Soviet Union over the next 20 years, largely due to the growing use of natural gas. Other factors contributing to reductions in coal use include growing environmental concerns and continuing pressure on member countries by the European Union to reduce domestic coal production subsidies. In contrast, coal use is expected to grow in Asia, especially in China and India. These countries have vast reserves of coal with limited access to other sources of energy, creating a situation for increased coal use to fuel their rapidly growing industrial sector. China has also shown interest in coal liquefaction and is planning to build their first direct coal liquefaction plant in the coal-rich province of Shanxi.

7.2 New Environmental Challenges

In the quest for a cleaner, healthier world, there is increasing pressure for a cleaner operation of coal-using operations. Tighter environmental regulations that require further reductions in the emissions of SO_2, NO_x, air toxics, and fine particulates are likely to be enacted over the next several years. Compliance with these increasingly stringent restrictions on emissions could be costly for coal users and lead to a reduced coal use. The most significant emerging issue against coal use is climate change associated with the production of greenhouse gases (GHGs). Although still hotly debated, many scientists believe GHGs are responsible for observed changes in the world's climate. Coal-fired plants, along with oil and natural gas-fired power plants, will require significant reductions in CO_2 emissions to mitigate the impact of GHGs. The challenge to continued coal use is the development of technologies that can operate efficiently and economically within tighter environmental regulations.

Coal gasification is a growing technology to use coal with near-zero emission of pollutants. It also provides the flexibility to generate a wide range of products from coal, including electricity, gaseous and liquid fuels, chemicals, hydrogen, and steam. Coal (or any carbon-based feedstock) is reacted with steam at high temperatures and pressures under carefully controlled conditions to partially oxidize the coal. The heat and pressure break apart chemical bonds in coal's molecular structure and form a gaseous mixture, not unlike the gas mixture generated by the indirect liquefaction of coal (discussed earlier). The hot gases are sent to a special filter that removes nearly 99% of sulfur pollutants and GHGs. The clean, coal-derived gas is similar to natural gas and can be used for electric power generation or converted to clean liquid fuels and chemical feedstocks. Pollutants containing nitrogen and sulfur are processed into chemicals and fertilizers and residual solids are separated and sold as coal combustion products.

Coal gasification is a more efficient way to generate electricity compared to conventional coal-fired power

plants. Instead of burning coal in a boiler to produce steam that drives a steam turbine/generator, coal gasification uses a special cycle, an integrated gasification combined cycle (IGCC), whereby the hot, high-pressure coal-derived gas is first used to power a gas turbine/generator. The exhaust from the turbine still contains sufficient heat to be sent to a boiler to produce steam for a steam turbine/generator. This dual generation of electric power represents an increased extraction of heat from coal. An IGCC plant operates with an efficiency of 45–55%, compared to 33–35% for most conventional coal-fired power plants. Higher efficiencies translate into better economics and inherent reductions in GHGs. Pressurized fluidized bed combustion (PFBC) boilers and IGCC will likely be used in future coal-fired power plants.

7.3 The Vision 21/FutureGen Plan

Vision 21/FutureGen represents a new approach to coal use that is being developed by the U.S. Department of Energy and private businesses. Unlike the single-fuel, single-product (usually electricity) plant, a Vision 21/FutureGen plant would produce a range of products—electric power, liquid fuels, chemical feedstocks, hydrogen, and industrial process heat. It also would not be restricted to a single fuel, but could process a wide variety of fuels such as coal, natural gas, biomass, petroleum coke, municipal waste, or mixtures of these fuels. The Vision 21/FutureGen concept envisions a suite of highly advanced technologies in modular form that can be interconnected and customized to meet different market requirements. Such modules include clean combustion/gasification systems, advanced gas turbines, fuel cells, and chemical synthesis technologies, coupled with the latest environmental control and carbon dioxide sequestration technologies.

The Vision 21/FutureGen plant may use a fuel cell (instead of a turbine/generator) to convert coal into electricity. The syngas produced from coal gasification is processed via an advanced water–gas shift reactor to separate hydrogen, and the carbon monoxide is converted to carbon dioxide and sequestered. The hydrogen is supplied to a fuel cell to produce electric power in a clean and efficient manner.

A Vision 21/FutureGen plant will be nearly emissions free, and any waste generated would be sequestered and recycled into products. Plant operating efficiencies are expected to be greater than 60% for coal-based systems and 75% for natural gas-based systems. Further details on the Vision 21/FutureGen plant are available by browsing through the U.S. Department of Energy's Web site at http://www.fossil.energy.gov/.

SEE ALSO THE FOLLOWING ARTICLES

Acid Deposition and Energy Use • Clean Coal Technology • Coal, Chemical and Physical Properties • Coal Conversion • Coal Industry, Energy Policy in • Coal Industry, History of • Coal Mine Reclamation and Remediation • Coal Mining, Design and Methods of • Coal Preparation • Coal Storage and Transportation • Cogeneration • Greenhouse Gas Emissions from Energy Systems, Comparison and Overview • Markets for Coal

Further Reading

Ashton, T. S. (1999). "The Industrial Revolution, 1760–1830." Oxford University Press, Oxford.

Bell, M.L., Davis, D.L. (2001). Reassessment of the lethal London fog of 1952: Novel indicators of acute and chronic consequences of acute exposure to air pollution. *Environ. Health Perspect.* **109** (Suppl. 3), 389–394.

Finkelman, R.E., Skinner, H.C., Plumlee, G.S., Bunnell, J.E., (2001). Medical geology. Geotimes (Nov. 2001), pp. 20–23 Available at www.geotimes.org.

Menon, S., Hansen, J., Nazarenko, L., and Luo, Y. (2002). Climate effects of black carbon aerosols in China and India. *Science* **297**, 2250–2253.

Paul, A.D., Maronde, C.P. (2000). The present state of the U.S. coke industry. Proc. 8th Int. Energy Forum (ENERGEX 2000), Las Vegas, pp. 88–93. Technomic Publ. Co., Lancaster, Pennsylvania.

Plasynski, S., Hughes, E., Costello, R., and Tillman, D. (1999). "Biomass Co-firing: A New Look at Old Fuels for a Future Mission." Electric Power '99, Baltimore, Maryland.

Wyzga, R.E., Rohr, A.C. (2002). Identifying the components of air pollution/PM associated with health effects. In "Air Quality III, Proceedings." Energy and Environmental Research Center, University of North Dakota.

Coal Industry, Energy Policy in

RICHARD L. GORDON
The Pennsylvania State University
University Park, Pennsylvania, United States

1. **Land Disturbance from Coal Mines**
2. **Health and Safety Practice**
3. **Air Pollution**
4. **Rent Seeking and the Coal Industry**
5. **The Search for Appropriate Charges for Land**
6. **The Case of Coal Leasing: An Overview**
7. **Coal Protectionism**
8. **Promoting Coal Utilization Technology**
9. **Electricity Regulation and Coal**
10. **Notes on the Economics of Regulation**
11. **Regulation and Public Land**
12. **Conclusion**

Glossary

air pollution The harmful discharges from combustion and other human activities.
due diligence A legal requirement for adequate performances such as time limits on implementing a mineral lease.
economic rents Income in excess of (economic) costs, often due to access to superior-quality resources.
externalities The effects of economic activities on bystanders.
fair market value An estimate of the probable selling price of an asset being purchased or sold by a government.
lease bonus A fixed advanced payment for mineral rights.
public utilities Firms regulated by the government because of concern regarding lack of competition.
publicness The simultaneous provision of commodities to everyone.
royalties Payments associated with production levels; usually a percentage of selling price but can be set as an amount per unit of production.
severance taxes Tax payments associated with production levels; usually a percentage of selling price but can be set as an amount per unit of production.
subsidy Government aid payments.
surface mining Extraction by uncovering surface material.
underground mining Extraction by digging under the surface cover.

Throughout the world, coal is profoundly affected by public policies. In almost every case, policies are directed at the consequences of coal production and use. Rarely is the promotion or even the limitation of coal use the primary objective. The best known policies affecting coal are those seeking to control air pollution from coal burning and water pollution and land disturbance from coal mining. Special, separate regulations govern efforts to protect the health and safety of coal. United States government public land policy includes complex rules for leasing coal resources on public lands. Since coal use is increasingly concentrated in electricity generation, regulatory activities in the electric power sector profoundly affect coal. This article reviews these policies.

1. LAND DISTURBANCE FROM COAL MINES

Mining disturbs the surface to some extent whether or not it occurs on the surface. This may involve damages with widespread effects. For example, abandoned mines are considered eyesores and may contain material that causes water pollution. Some sites may comprise ecosystems valuable for contributions beyond the marketable goods that can be produced. Socially beneficial biota may be present.

Although coal mining states had imposed controls, Congress decided that federal action was imperative. Thus, a national law, the U.S. Surface Mining Control and Reclamation Act of 1977, was passed. The law sets specific reclamation goals and requires the enforcing agency to create even more specific implementation rules. The basic goals are to restore the land to the "approximate original contour," restore the land to support its prior use or better ones, segregate topsoil so it can be replaced, and designate some land off limits to surface mining.

 © 2004 Elsevier Inc. All rights reserved.

The law established an Office of Surface Mining (OSM) in the Department of the Interior (DOI) and created principles for guiding reclamation. OSM was to supervise the development of state programs to implement the law or, if a state chose not to develop a program, regulate mines in that state. The law covered reclamation of both active and abandoned coal mines. Taxes were levied on surface and underground coal mines to finance reclamation of abandoned mines. For operating mines, the law required the establishment of permitting systems under which the right to mine was granted only when extensive data, including the mining and reclamation plans and information on the impacts of mining, were submitted. Mine inspections were also mandatory. The law had complex provisions prohibiting or discouraging surface mining on certain types of land, such as "prime agricultural land."

Since this law produces benefits that are largely difficult to quantify, impact studies are limited. Neither the impacts on the industry nor the benefits of control are accurately known. Eastern surface mining started declining at approximately the time that the act was enacted, but surface mining growth continued in the western United States, albeit at a markedly lower rate. The act was not the only force at work and its effects are difficult to isolate.

2. HEALTH AND SAFETY PRACTICE

The traditional concern on the production side is worker safety. After a severe mine disaster in 1968, regulation was revised in the Coal Mine Health and Safety Act of 1969. Safety regulation was made more stringent and regulations of health effects were added. Moreover, administration was transferred from the U.S. Bureau of Mines of the DOI to a newly created Mine Safety and Health and Safety Administration, which was initially a part of the DOI but was transferred in 1977 to the Department of Labor.

Regulation has involved both setting rules for operating mines and federal inspection of the mines. The 1969 act both tightened the rules and increased the frequency of inspection. New rules required that no mine work occur under unsupported roofs; improved ventilation; use of curtains and watering to ensure reduction of "respirable dust" levels and increase the flow of air to the "face" (i.e., where mining was occurring); and increases in the monitoring of mines to determine levels of methane, dust, and other dangerous material. The number of inspections increased greatly. In the early years, many mines had daily inspections because of special problems or because they were so large that only daily visits sufficed to ensure that the required tasks could be completed.

Simultaneously with the passage of the Act, underground coal mining productivity began declining from a peak of 1.95 short tons per worker hour in 1969 to a low of 1.04 short tons in 1978. Then a reversal occurred so that by 2000, the level was up to 4.17 short tons.

Much effort was devoted to determining the role of the Act in the initial productivity declines. Evaluation was hindered by the standard problems of multiple, interrelated influences on which data were unavailable. In particular, the passage of the Act coincided with (and perhaps was a major cause of) an influx of inexperienced miners. As a result of the Act, more labor use in mines appears to have been necessary. The need to increase inspections contributed to the need for new workers because experienced miners were recruited as inspectors. The effects that this had on mine productivity are unclear, but whatever they were, they have been substantially reversed.

3. AIR POLLUTION

Concerns about air pollution from coal began with widespread coal use. The initial and still relevant focus was on the visible effects of spewing soot (particulates) from coal. Coal has a substantial but variable particle content. At best, the content is much greater than that of oil or gas.

Coal contains a variety of other contaminants. The most notable is sulfur. Sulfur contents of coals are also widely variable, and the lowest sulfur coals may have considerably less sulfur than many crude oils. During combustion, nitrogen oxides (NO_x), carbon dioxide (CO_2), and carbon monoxide form. In the 20th century, the problem of largely invisible pollution from SO_x and NO_x emerged, as did concern regarding smaller, invisible particulates. Recently, attention has turned to CO_2. It is the main greenhouse gas suspected of contributing to increased temperatures in the atmosphere (global warming).

Pollution control can be performed in several ways. Naturally less polluting fuels can be used, precombustion cleanup can be performed, and the pollutants can be captured after combustion. However, precombustion treatment is not an economically

feasible option with coal. Available coal washing and sorting technologies can only modestly reduce particle and sulfur content.

More radical transformation is technically possible but economically unattractive. Prior to the development of a natural gas distribution network, coal gasification provided the input for local gas distribution systems. Other technologically established possibilities are to produce a higher quality gas than obtained from the earlier gasifiers, synthesize petroleum products, or simply produce a liquid that would compete successfully with oil and gas. All these processes would have the further benefit of facilitating the removal of pollutants. To date, the costs have proven too great to permit competition with crude oil and natural gas. Thus, emphasis is on changing the type of fuel and postcombustion cleanup. Changing to oil and gas lessens particulate and SO_x emissions. Another approach to SO_x is to use lower sulfur coal.

An alternative is to trap the particles and sulfur before discharge into the air and deposit the trapped wastes on nearby lands. The latter approach is widely used. Particulate control is invariably effected by postcombustion methods. Two standard methods are available—a mechanical filter known as a bag house or devices called electrostatic precipitators that trap the particles by electrical charges. Particle control techniques are long extant and have low control costs. Thus, visible discharges are largely eliminated. Little coal burning occurs in facilities too small to control particulates. Even the best particulate control devices, however, cannot capture all the smallest particles, and many pre-1960 regulations did not require the greatest control that was physically possible.

Sulfur oxide control proved more expensive and more difficult to perfect than expected when control rules were imposed. The wastes from these cleanup facilities can cause disposal problems. The reduction in sulfur and particulate emission associated with a switch to oil or gas is associated with increases in nitrogen oxides and unburnt hydrocarbons.

Similarly, the lowest cost, low-sulfur coal in the United States is subbituminous from Wyoming, which has a higher ash content than eastern coal. Thus, lowering sulfur by changing fuel may result in increased particulate emissions. Some observers argue that the close association between particulate and SO_x emissions precludes definitive assignment of impacts. SO_x may be blamed for effects that are actually caused by the accompanying particulates, and coal-source shifting to fuel lower in SO_x may be harmful.

Complex air pollution control programs have been developed. United States air pollution laws have resulted in multitiered regulations that are periodically tightened. The practice involves increasing the number of regulated pollutants. In the major cases, a two-phased approach prevails. One set of rules determines air-quality objectives, and another controls emissions from different sources.

The overall goals specify the allowable concentrations in the atmosphere of the pollutants with which the law is concerned. States must establish implementations plans (SIPs) to ensure that they meet these goals. Failure to do so results in a region being labeled a nonattainment area.

In addition, a federal court ordered that the preamble to the law that called for maintaining air quality require "prevention of significant deterioration" (PSD) in areas in compliance with the rules. Widespread failure to meet SIP goals and the vagueness of the PSD mandate were treated in the 1977 Clean Air Act amendments. The country was separated into nonattainment areas in which actions must be taken to comply and PSD regions in which increases in pollution were to be limited to amounts specified by the law. Continued nonattainment resulted in further remedial legislation in 1990. This effectively divided the country into nonattainment areas legally obligated to come into compliance with the goals and PSD areas. Attaining compliance is difficult, and the major amendments in 1977 and 1990 to the Clean Air Act contained extensive provisions to stimulate compliance further.

Similarly, ever more complex rules have been developed to determine how major polluters should limit pollution. Newer facilities have long been subjected to more severe new source performance standards. The new source rules originally simply imposed stricter emission controls on new facilities. This was justified by the fact that the cost of incorporating controls into a newly designed plant was less than that of adding controls to an old plant. The disincentives to adding and operating new plants were ignored. (An additional factor that became controversial was precisely defining how much refurbishment of an old facility was consistent with maintaining its status as an existing plant.)

The initial rules resulted in considerable changes in the type of fuel used. Western coal was heavily used in such producing states as Illinois and Indiana and in states such as Minnesota that previously relied on Illinois coal. Mainly in response to this change, the rules were changed in the 1977 Clean Air Act amendments to restrict how emissions were reduced.

The amendments required use of best available control technology for preventing the discharge of pollution. The available option was employing scrubbers, a cleanup technique that emphasizes capture of pollutants between burning and discharge of waste up the smokestacks. Since some cleanup was necessary, the advantage of shifting to a cleaner fuel was diminished.

Under the 1990 Clean Air Act amendments, a new program targeted the reduction of emissions in the most heavily polluting existing electric power plants. Reflecting a long-standing debate about "acid rain," the 1990 amendments instituted a complex program for reducing sulfur and nitrogen oxide emissions from existing sources.

This two-phased program initially imposed rollbacks on the units (specifically named in the law) at 111 power plants that exceeded both a critical size and level of sulfur oxide emissions. By January 1, 2000, electric power plants were required to undertake a 10 million ton reduction (from 1980 levels) of sulfur dioxide and a 2 million ton reduction in nitrogen oxide emissions. The named plants had to undertake the reductions in the first phase of the program. The law encouraged cheaper abatement by allowing the polluters opportunities to buy offsets to their activities. The provisions thus involve a paradox. Congress first extended the command and control approach by specifying limits on specific plants. Then a concession to market forces was added.

The compliance choice for existing plants, in principle, is limited by the constraints on the plants that must make changes and the supply of lower sulfur coals. Thus, boiler design may prevent the use of some types of coal and result in loss of capacity when other coals are used. At the time of enactment, those modeling the impacts debated the availability and usability of low-sulfur coal from different states. Factors such as proximity to coal fields, railroads, and waterways affect the ease of procurement from alternative suppliers. The nature of available receiving facilities is another influence. Some plants were built near coal mines to which they are connected by conveyer belts. Others are served by private rail lines and in one case a slurry pipeline. Some lack land to add large new facilities.

In practice, implementation involved fuel changes rather than increased scrubber use. Wyoming coal was shipped farther than expected (e.g., Alabama and Georgia). Emissions trading proved effective in lowering compliance costs.

Global warming, an increase in temperature produced by discharge of carbon dioxide from fossil fuel burning, became prominent in the late 1980s. It is contended that emission of carbon dioxide into the atmosphere will inevitably and undesirably raise atmospheric temperatures.

There is an enormous amount of literature on the effect of global warning and what it might mean. First, questions arise about the exact extent of warming that will occur. Second, uncertainties prevail about the physical effects of this warming. Third, numerous estimates have been made about the economic effects. Should the United States commit to massive reductions in greenhouse gas emissions, it may result in severe problems for the coal industry. Coal burning produces more greenhouse gas than the use of other fuels.

4. RENT SEEKING AND THE COAL INDUSTRY

Another profound influence on the coal industry is government policy toward the ownership and taxation of coal resources. Ownership and taxation are not inherently linked but have become so in practice. Ownership and taxation are alternative ways to extract income. The sovereign can and has used its ownership, taxing, and regulatory authorities to affect ownership, utilization, and disposition of the resulting income. Disputes occur perennially.

An important consideration is that private and public owners can and do divide and separately sell mineral and surface rights. Some governments reserve all mineral rights for the state. The U.S. government adopted a practice of retaining mineral rights to some lands it sold to private owners. Separate ownership of the land's subsurface and the surface separates an indivisible property and guarantees conflicts. The subsurface cannot be exploited without affecting the rights of the surface property owner. Defining and, more critically, implementing an acceptable system of surface owner compensation are a major problem of government or private retention of mining rights. In particular, regulations strong enough to protect surface owners from misuse of sovereign power may constitute de facto privatization of the mineral rights to surface owners.

Starting with the 1920 Mineral Leasing Act, Congress moved toward what was to prove a standard policy of extending access on a rental rather than ownership basis. The 1920 act established leasing as the means of access to fossil fuels (oil, gas, and coal) and fertilizer minerals. Periodic revisions of the leasing act have tightened rules of access.

The increase in federal environmental legislation has also greatly influenced public land administration. One pervasive influence was the requirement in the 1969 National Environmental Policy Act that the federal government examine the environmental impacts of its major actions. The result was a thriving adversarial process in the 1970s. Those opposed to federal actions successfully delayed actions by challenges to the federal evaluation process. However, this tactic subsequently became irrelevant with regard to energy but apparently still affects may other government decisions.

5. THE SEARCH FOR APPROPRIATE CHARGES FOR LAND

Policymakers and economists have long debated the method of charging for rights of access to those who extract minerals. The concerns focus on economic rents—the extra profits of a superior resource. Many alternative terms exist for this extra income, such as excess profits, surplus, and windfalls. Whatever the name, it consists of the differences between revenues and the social costs, including the necessity to earn the required rate of return on investment.

For a mineral deposit, there are rent-generating advantages with a better location, higher quality material, easier mining conditions, and similar attractive features. This applies more broadly to all natural resources. Thus, a charge on these rents allows "the public" to share in the bounty of naturally superior resources without distorting resource allocation. Therefore, policies that transfer rents to the public are often deemed desirable. However, a paradox prevails. The charges can be imposed using a classic textbook example of a tax with no effects but, in practice, the charges employ a method that follows an equally classic textbook approach that clearly discourages desirable activity.

Economic analyses of rents concentrate on securing as much government revenue as possible without inefficiently reducing production. A transfer of rent has no effect on effort if the terms are independent of any action of the payer. The relevant classic textbook case is a lump-sum charge, a payment that is a fixed amount that is not alterable by the buyer's actions. The preferable approach is to impose a charge only at the time of transfer of rights to avoid the temptation subsequently to impose distortionary charges. The arrangement may be an outright sale or a lease in perpetuity with neither control over end use nor any ability to alter charges.

A lease bonus, a single payment made when the lease is implemented, is the quintessence of such a preset total payment. The bonus, by definition, cannot be altered by any action of the firm, and no subsequent decisions can be affected by having paid a bonus. In economic jargon, the bonus is a sunk cost. That fixed unalterable decision at the time of sale is administratively feasible is demonstrated by the history of U.S. public land sales and by the millions of outright sales that occur in the private sector.

In the case of land grants, competitive bidding to pay a fixed sum before production begins would capture rents. The maximum anyone would pay is the (present-discounted) value of the expected rents. Vigorous competition would force payments of the maximum.

Rent transfers thus seem ideal for tax collectors. The rents are excess profits that can be taxed without affecting output, and the transfers are seen as fair because the rents, in principle, provide rewards unnecessary to produce the effort.

In practice, deliberate decisions are universally made not to realize the potential for efficient transfer of income. As economics textbooks routinely warn, basing payments on output or its sale creates a disincentive to produce. Since the severance taxes used on minerals are sales taxes by another name and royalties have the same impacts, the analysis applies to royalties and severance taxes. Funds that consumers wanted to give producers are diverted to the imposer of the tax. This revenue transfer discourages producers from acting. The amount ultimately purchased by consumers declines because of these disincentives. This is a violation of the central economic principle that every expansion of output that costs less than its value to consumers should occur. The technical language is that all outputs with marginal costs (the cost of expanding output) less than price should be undertaken. Therefore, the charges imposed by federal land agencies, other governments throughout the world, and private landlords differ greatly in practice from ideal rent taxes and take a form that eliminates the differences from reliance on established tax collection agencies.

Moreover, royalties and severance taxes also create difficult collection problems and thus increased administrative cost. Requiring more payments means more compliance efforts by government and land users. DOI administration can and has encountered efforts to conceal production. A percentage of revenue royalty can be difficult to

administer because of lack of satisfactory data on market prices. Lump-sum payments eliminate the expenses of monitoring.

The imposition of output-related charges is justified by claiming that because of imperfect competition, the defects of linking the payments to activity are outweighed by the income gains.

A further complication is that U.S. fossil fuel leasing policy involves both bonus payments and royalties. The drain of royalties will lower the amount of an initial payment. At a minimum, this hinders the evaluation of the adequacy of bonus bids. The residual after payment of royalties is necessarily less than the present value of untaxed rent. It is a residual of a residual—namely, the rent.

6. THE CASE OF COAL LEASING: AN OVERVIEW

The U.S. government owns many of the nation's most economically attractive coal resources. In particular, it is the dominant owner of coal in such western states as Wyoming, Montana, Colorado, Utah, North Dakota, and New Mexico. Most U.S. coal output growth since 1970 has occurred in these states. The coal leasing law and its administration changed little from 1920 to 1971. Leases were freely granted under the provisions of the 1920 Mineral Leasing Act.

These policies allowed satisfaction of most desires for leases. Coal leasing proceeded modestly until the late 1960s, when it grew rapidly. At least retrospectively, the 1960s leasing surge is explicable by anticipation of the expansion of western output that actually occurred.

Starting in the late 1960s, both the legislative and administrative processes changed radically. Congress passed a series of laws directly or indirectly affecting coal leasing. Federal coal policy became so polarized that a hiatus in leasing has prevailed for most of the period from 1971 to today. In 1971, the DOI became concerned about an acceleration of leasing without a concomitant increase in output. The result is that coal leasing ceased from 1971 to 1981, resumed briefly in 1981, was stopped in 1983, and its resumption remained indefinite as of 2003. The resumption in 1981 was the first test of 1976 revisions in the laws governing coal leasing. Enforcement proved problematic.

The Coal Leasing Amendment Act of 1976 radically altered coal leasing policy. President Ford vetoed the act, but the veto was overridden. The central components of the Act were that it (i) required competitive bidding on all leases; (ii) required the receipt of fair market value on all leases; and (iii) required a royalty of at least 12.5% of the selling price on all surface-mined coal, with the Secretary of the Interior given discretion to set a lower rate on underground-mined coal.

Many ancillary provisions were provided, but a complete review is not critical here. The provision that proved to have the greatest practical relevance was that requiring forfeiture of any lease that was not developed within 10 years (the due diligence requirement). Such requirements are a response to the output-retarding effects of heavy royalties. This accepts the belief in lease limits criticized later in this article. Such diligence requirements are either redundant or harmful. In the most favorable circumstances, the requirement will leave the decision about whether and when to operate unchanged. Cases can arise in which it is more profitable to start too soon than to lose the lease. When surrender is preferable to operating and occurs, no guarantee exists that reoffer will occur before the optimal starting time.

Diligence requirements are also disincentives to arbitrage in bidding and to prevention of mining when superior private uses exist. If lease lengths were unlimited, a leaseholder who believed that other uses were more profitable than mining could refrain forever from mining the land.

The 1983 moratorium was inspired by charges that some lease auctions had failed to secure fair market value. The Commission on Fair Market Value for Federal Coal Leasing was created to evaluate the charges and proposed solutions. Given that the problem was inadequate data to support the decisions, the commission proposed devising more elaborate evaluation procedures. (The DOI could find only one public record of a sale of coal reserves, a defect that later critics, including the commission, could not remedy.) These were created, but large-scale leasing never resumed. The reluctance to resume leasing occurred because the amount of coal already leased appeared sufficient to meet predicted consumption levels for several more decades.

As noted previously, coal leasing ceased despite all the safeguards imposed. This suggests that land management has become so controversy bound that even when stringent safeguards are in place, the critics remain dissatisfied. Critics of the experience, such as Nelson, Tarlock, and Gordon, argue that unrealistic expectations governed the adopted processes.

7. COAL PROTECTIONISM

Efforts to promote coal prevail but are problematic. Two are prominent. Given the drawbacks of coal use, persistent although not extensive efforts are made to develop better ways to use coal. The term clean coal technology characterizes the desired results. Governments in many countries, most notably in Western Europe and Japan, have sought to prevent or limit the decline of uneconomic coal industries.

The United States also protected coal from imported oil but on a much smaller scale. Until the oil price increase of the 1970s, the United States maintained oil import quotas. These were imposed largely to support a domestic oil price program that state agencies had established in the 1930s. However, proponents also recognized the benefits of such controls to the coal industry. Nevertheless, one of the first breaks in the policy was the 1969 change that effectively removed the restrictions on imports of residual fuel oil into the eastern United States, which was the main competition for coal.

On a much more modest scale, the anthracite-producing region of Pennsylvania was represented in the U.S. Congress by two well-placed representatives who regularly secured government support for assistance programs. The most effective was the requirement of export of anthracite for use by the U.S. military in Germany. Otherwise, the effort was limited to studies. This culminated at the start of the Carter administration. A commission was created with the stated goal of significantly stimulating anthracite production.

However, the available options were unattractive. Charles Manula, a professor of mining engineering at The Pennsylvania State University, indicated that the best prospect for increasing production was to expose and remove the coal left behind in abandoned mines. Although the commission nominally supported this proposal, it received no further support. The drawback was that the approach created much land disturbance to produce little more coal than the output of one western strip mine.

Much more extensive and sustained programs prevailed in Western Europe and Japan. For more than 45 years, resistance to rapid rationalization of the coal industry has distorted European energy decision making. The difficulties started long before. Among the problems caused by World War I was a disruption of European coal industries that worsened due to the effects of the Great Depression and World War II. Europeans agonized about how to revive the industry.

After World War II, initially it was widely believed that coal was a mismanaged industry that could thrive with the right management. It was believed that a vigorous rebuilding program would restore the coal industry to its former might. Britain and France nationalized the coal industries. They operated on a break-even basis, which means rents subsidized outputs. The Germans and the Belgians preferred reliance on a heavily regulated and subsidized, largely privately owned coal industry.

The Dutch had developed their coal industry early in the century as a nationalized venture. The Belgians recognized that the large number of small old mines in the south were beyond saving but seemingly had to be protected for political reasons.

A turning point occurred in 1957. In the aftermath of the 1956 Suez Canal crisis, optimism about European coal was eliminated. Middle East oil could no longer be dismissed as a transitory energy source. Conversely, the contention that sufficient effort would make coal more competitive was demonstrated to be false by the pervasive failure of massive revival efforts. Oil could be sold for less than coal for many decades. Following typical patterns of aid to distressed industries, the responses were programs to slow the retreat. Consolidation was proposed and slowly effected.

Even the oil shocks of 1973–1974 and 1979–1980 failed to help. The situation for European coal was so bad that competition with higher oil prices was still not feasible. Moreover, ample supplies of cheaper coal were available elsewhere. Even, as in coking coal, when coal use was efficient, given the ample supplies available in the United States, it was cheaper to import. Thus, a policy allegedly designed to limit dependence on unreliable supplies from the Middle East actually mostly harmed reliable suppliers from first the United States and then, as the world coal industry developed, Australia, South Africa, and Colombia.

Major national differences caused each of the coal-producing countries to adopt largely independent policies. The countries had important differences regarding both goals and programs. The fervor of German devotion to coal and the willingness to impose regulations were particularly intense. The Germans erected an elaborate structure of defenses and were the most reluctant to contract. The critical factor in Britain was the ability of the mineworkers union successfully to make resistance politically untenable until the union overreached in the Thatcher years.

The impact of the differences among countries is most clearly seen by comparing production trends.

The Dutch discovered a gas field that could displace coal and produce for export. In 1974, the Dutch were the first to exit the coal industry. The main adjustment was that the coal mining company was made a participant in gas production.

The French and the Belgians moved toward substantial but slow contraction. The French contraction continues; Belgian production ended in 1992. The preservation programs in France, Britain, Spain, and Belgium were largely ones of direct subsidies to coal mines. (Japan took a similar approach.)

British aid was largely direct but reinforced with a contract between the coal industry (originally the National Coal Board and then renamed British Coal before privatization) and the former Central Electricity Generating Board. This contract yielded higher prices than those of competing imports. Similarly, before its privatization, British Steel apparently also had a cozy relationship with British Coal.

During the Thatcher years, there was an impetus for radical reform. First, in 1984, the leader of the mineworkers union chose to implement a strike without an authorizing vote. The combination of incomplete support from the members and strong resistance by the Central Electricity Generating Board resulted in defeat for the union.

After the failure of the strike, a restrained move was made toward reducing assistance. British Steel shifted to imported coal after it privatized in 1988. Electric power restructuring in 1990 led to radical reductions in utility purchase of British coal. The splitting of the Central Electric Generating Board into three companies, two of which are privatized fossil fuel-using organizations, created pressures to alter procurement practices. The new companies were more aggressive about fuel procurement. The increase in gas-using independent power producers added competition for coal. In this changing climate, the Major government effected privatization of coal. All these factors accelerated the industry's decline. Output declined from 116 million tonnes in 1983, the year before the strike, to 30 million tonnes in 2002.

The Germans had the most complex and intractable of the systems. Previously, support was financed by funds from general federal and state revenues, revenues from a special tax on electricity consumers, and forcing consumers to pay more than market prices. Aid was dispensed in several forms: direct subsidies of specific current activities of the coal industry, programs to promote the use of German hard coal in coking and electricity generation, and assumption of the cost burdens (mostly miner pensions) from past mining. (The last policy is standard among Western European coal-assistance programs.) The electricity tax paid for direct and indirect costs of coal use but was ultimately banned by a German court. A long-term contract committing domestic electric utilities to buy preset, modestly increasing tonnages in Germany at unfavorable prices prevailed from the mid-1980s to 1995. A move to direct subsidies similar to those elsewhere was made in response to the court ruling and resistance to renewing the coal purchase contract.

Since the coal crisis, the Germans have periodically set output goals presumably for many years. As costs mount, a reduction is undertaken. The goals set in 1973 called for attainment in 1978 of 83 million tonnes. The 1974 revision at the height of concern over the oil shocks called for 90 million in 1980; actual 1980 output was 94 million. A 1981 report seemed to advocate maintaining output at approximately 90 million tonnes. A 1987 review called for a 15 million tonne decrease to 65 million by 1995. A 1991 revision set goals of reducing output from approximately 80 million tonnes in 1989 to 50 million tonnes by 2001 and 43 million tonnes by 2010. Dissenters on the advisory committee called for a cut to 35–40 million tonnes. Output in 2002 was 27 million tonnes.

German aid averaged more than $100 per tonne starting in the 1990s. In contrast, nominal dollar prices of coal delivered to Europe were approximately $45 per tonne for coal for power plants and less than $60 per tonne for coal used to make coke meeting iron-making requirements.

8. PROMOTING COAL UTILIZATION TECHNOLOGY

As noted previously, in principle there are numerous ways to improve coal utilization, ranging from better ways to employ solid coal to transforming the coal into a different fuel. Programs to develop these technologies have prevailed with variable vigor for many years. However, neither transformation technologies nor new ways to use coal have proved to be commercial successes.

The U.S. government has developed a series of programs in this area. In the early 1970s, the Office of Coal Research was combined into the Energy Research and Development Administration largely composed of the research arm of the former Atomic Energy Commission. This, in turn, became the main component of the Department of Energy. In 1979, the Carter administration decided to expand its

energy initiatives and in particular to encourage synthetic fuel development from coal. This policy was proposed even though research up to that point suggested that synthetic fuels from coal were not economical. A government-owned Synthetic Fuels Corporation was established, funded a few projects including a large lignite gasification facility in North Dakota, and was abolished in the 1980s. What was left was a much smaller clean coal technology program in the Department of Energy.

9. ELECTRICITY REGULATION AND COAL

Extensive, convoluted government involvement in the activities of electricity generation, transmission, and distribution existed long before the 1970s. However, the most profound effects on coal occurred with 1970s energy turmoil. The 1970s witnessed increasing prices in energy markets and policy changes. State regulators independently adopted measures to increase control over utilities and responded to federal mandates to alter practice.

The initial response to frequent requests for rate increases necessitated by the changing economic climate was multifaceted resistance. Denial of requests was supplemented by numerous forms of increased examination of realized and anticipated actions. Both major cost elements—construction and fuel purchases—were scrutinized. Management audits were also initiated.

The mass of federal energy regulations during the period from 1973 to 1992 had many elements that at least potentially had significant direct or indirect impacts on electric power. Two proved central. The 1978 Public Utility Regulatory Policy Act required states to consider many kinds of changes in regulation. However, the only major response was to the requirement to establish a mechanism to encourage utilities to purchase power from nonutility generators (NUGs) deemed worthy, mainly those using renewable resources (other than traditional large hydroelectric facilities) and "cogeneration" (electricity production by industrial companies in the facilities providing energy to their operations). Utilities were to be required to purchase from these sources when they were cheaper or, in the jargon that was developed, when the avoided cost was lower. However, the states were granted considerable discretion about how these costs would be estimated.

Given that the process involved choosing among investment options, estimates of lifetime costs had to be made. Given the climate of the time, estimates in states partial to NUGs incorrectly assumed substantial, steady increases in the cost of conventional sources of power. Utilities became committed to contracts at higher prices than were justified by prevailing market conditions. Again imitating prior experience (with fuel procurement contracts), the move to the economically appropriate task of contract renegotiation is only slowly emerging.

Small plants had several features that made them economically attractive. Their short lead time and low capital investment characteristics made them more flexible in an era of low and uncertain growth than the traditional large plants, with their high unit and total capital costs and long lead times. To the extent that suppliers would assume the risk of these ventures, they were even more attractive. NUGs provided an increasing portion of capacity additions. At least part of the industry saw independent generation and perhaps even the adoption of the proposal for totally independent generation as desirable.

Under prior law, almost all subsidiary–parent relationships were treated as public utility holding companies and subject to stringent controls. Revisions enacted in 1992 exempted operations that were incidental to the main activities of the firm. Thus, the power-producing subsidiary of a gas pipeline, an engineering firm, or even an electric company located elsewhere in the country no longer would be subject to stringent regulation.

The evidence suggests that from fairly modest beginnings in the mid-1980s, additions of NUG capacity increased while utilities lowered their own additions. These relied heavily on gas turbines often linked to a conventional boiler that uses the waste heat this is called a combined cycle. Simultaneously, there was cessation of building of conventional large coal-fired plants.

In the 1990s, several states tried to deal with these problems by altering how they regulated electric utilities. This generally involved deregulating wholesale transactions and lessening controls on rates to final consumers. The details differed greatly among states. One major difference related to ownership of generation. At one extreme (New York, Maine, and Massachusetts), divesting of all generation was required. At the other extreme, some states (e.g., Texas and Pennsylvania) made the sales voluntary.

These requirements and voluntary actions increased the role of at least nominally independent producers. Both established electric utilities from other states and specialists in independent power plants purchased the divested plants. Two forms of

nominal independence were allowed. Several independent power producers purchased integrated electric utilities and were allowed to treat the acquired generating plants as independent production. Conversely, regulators agreed to treat new generating subsidiaries of some established integrated utilities as independent producers.

The ability to sustain this pattern is doubtful. Although gas-supply potential has been underestimated, it is unlikely that the gas industry can economically serve its traditional markets plus a large part of electricity generation and transportation. In 2002, the electric power industry consumed 29 quadrillion BTUs of coal and nuclear energy and 5.7 quadrillion BTUs of natural gas. Total U.S. natural gas consumption was only 23 quadrillion BTUs.

10. NOTES ON THE ECONOMICS OF REGULATION

From an economic perspective, regulation is marred by a persistent tendency to adopt clumsy ways to attain goals. Often, the goals are disputed. The concerns are based on the limits that economic analysis places on the role of government. Standard economic theory employs the concept of market failure to delineate the reasons why government action might be desirable.

The critical concern is what economists call the public-good-supply problem, provision of services whose existence makes them available to everyone. Examples include national defense, police protection, and pollution control. As a vast literature indicates, enormous problems arise in determining and efficiently satisfying the demands for such goods.

Prior to 1960, it was widely believed that causing an "externality," an impact on others than the decisionmaker, also inherently required intervention. However, Ronald Coase published an enormously influential article on the subject. He showed that side effects could be and were privately corrected when the number of affected parties was sufficiently small to make negotiation feasible. It was only when the number of participants became too large that the private solution broke down. Implicitly, impacts became a public good and their control an example of the problem of public good provision.

As is rarely noted, Coase qualified his conclusions. Throughout his writings on economic policy, Coase stressed that governments have limited capacity for action. Therefore, it is unclear when it is preferable for the government to be the financer. In a follow-up, he illustrated his point by showing that a classical example of the need for government action, provision of lighthouses, was invalid. In England, lighthouses were originally successfully privately provided. The government stepped in only after the private success. Coase's implicit point is that too much stress is given problems of ensuring that all beneficiaries participate in private programs. The resulting inadequacy is still clearly preferable to failing to secure any government aid and might even be superior to a halfhearted government effort.

A more critical issue is why governments so often intervene when no clear publicness prevails. Key examples relate to the mass of regulations affecting labor relations, such as coal mine health and safety laws. The stated rationale is that the bargaining is unfair and governments must intervene to offset this defect. This may in some cases be a form of publicness. An alternative view is that such policies are simply examples of favoritism toward the politically influential. The issue is why a supposedly effective labor union could not have secured the benefits in the national negotiations it conducted.

The resulting problem is establishing the desirability of specific policies. First, the reality of a problem must be established. Second, it must be ascertained whether the impacts justify intervention. Finally, the best way to attain the regulatory goals must be determined. The first area involves expertise in many disciplines because environmental impacts may affect human health, damage property, cause unsightliness, and alter climate. People outside these disciplines must somehow evaluate the credibility of the often tentative and conflicting claims that are made.

The next essential step of appraisal of the desirability of action involves the economic approach of cost–benefit analysis. The problem is standard in applied economics. The methodology is well established; the data for implementation are severely deficient. An analyst must associate values to the imperfect impact estimates developed in other disciplines and relate them to flawed estimates of compliance costs. Thus, the actual estimates are uncertain.

Economics also produces a preference for decentralized decision making. Economists have long disparaged the presumption of politicians to decide what is preferable for others. Thus, among economists, the tendency toward detailed instructions that characterizes environmental regulation is censured as command and control regulation.

The starting point of the argument is that politicians have considerable difficulty determining the proper overall goals of public policy and great

problems determining the degree and nature of response to the goals. Therefore, if regulation is to occur, governments should adopt systems in which broad goals are set and the maximum possible flexibility is given with regard to implementation.

In particular, economists suggest two basic approaches: charges for damages or permits to damage. The underlying principle is that economically pollution control is a commodity whose benefits and cost vary with the amount of effort exerted. Thus, the standard economic technique of supply and demand analysis applies. An optimal level of control exists. At that level, the price that the public is willing to pay for that quantity of control attained has a value equal to the marginal cost of the control. Thus, as with all commodities, a market-clearing price is associated with the optimum level of control.

This viewpoint simply formalizes realities behind actual regulation. Total control is never attempted because of the high cost of perfection. Indeed, perfection may be technologically impossible (an infinite cost). Thus, denunciations of economic analyses as immoral are apparent denials of the realities. What really seems to be involved is concern that the limits of knowledge will cause less stringent controls than are justified.

In any case, the key is the association of a price with the market-clearing level of control. With sufficient information, both the market-clearing level of control and the associated price are knowable. This forms the basis for the tax and marketable permits basis for regulation. Under the tax system, every polluter is charged a tax equal to the value of efficient abatement. The permit system instead sets a total limit on pollution and allocates shares in that limit. In both systems, the affected parties decide how much and in what way to abate.

The previously noted problems of inadequate information imply that charges, transferable rights, and command and control all will fail to attain the unknowable efficient level of control. To make matters worse, the magnitude of the resulting policy defects depends critically on the magnitude of the error in the policy goals. Therefore, no *a priori* presumptions are possible about which policy fails least.

11. REGULATION AND PUBLIC LAND

Government land ownership is politically popular. However, the economists working on the subject are skeptical about this acquiescence. The standard economic arguments for intervention have little or no relevance. Briefly, most U.S. federal land use is by private individuals for ordinary commercial activities, such as grazing, forestry, and mineral extraction. Thus, the usual reasons for preferring private ownership prevail.

Another basic problem is that, as noted previously, while access charges in principle can be levied without affecting production, the standard practice is to adopt charge systems that discourage production. The view that not to charge is to cause irreparable losses to the Treasury is overstated. Any business established on public lands, its suppliers, and its employees will pay their regular taxes. The greater the efficiency of the firm, the more regular taxes will be paid. Thus, the normal tax system is at least a partial corrective to the failure directly to charge. It is unclear that this approach produces less tax revenue than a royalty system.

Moreover, the concept of rent transfer to government must coexist with a rival policy approach to rents—transfer to consumers. Although this objective has not arisen for U.S. coal, it was a major consideration with regard to European coal (and with regard to 1970s U.S. policy for oil and natural gas).

Another consideration is that the concept of optimal lease timing is invalid. More precisely, you can sell too late but never too soon. Market value is determined by expected profits from optimal development. The development schedule is determined by the trends in the markets for the output of the resource. The value is determined by the existence of the mineral, not its ownership. The land sale or lease transfers control without altering mineral existence and value.

Any private firm making a lease or purchase before it is efficient to use the property will wait until the socially most desirable exploitation date. Failure to realize this basic principle results in invalid concerns about adequate payments and calls to limit leasing. It is possible only to lease too late. Resources may not be made available to a qualified operator until after the most desirable time for exploitation. Moreover, the case does not depend on perfect markets. Delaying leasing cures no market failure. At best, time is gained that might be used to eliminate the failures.

Another problem is that politicians and critics of leasing impose unattainable standards of proof on the acceptance of offers of initial payments. Fears of giveaways prevail. The intrinsic drawbacks of rent taxation are aggravated by the insistence on detailed estimates of whether selling prices are sufficient.

Concerns also exist about the feasibility of administering the system and ensuring public faith in the integrity of the system.

In public land policy debates, leases or sales of public land are conditional on payment of "fair market value." Fair market value more generally is the legal term used to define the appropriate price to be paid in the purchase and sale of property by government agencies. Its receipt is a perennial issue. The term seems simply to mean what the asset would sell for in a competitive market. The U.S. government policy manual on valuation, "Uniform Appraisal Standards for Federal Land Acquisitions," implies such an interpretation, as does at least one Supreme Court decision.

Economic principles indicate that vigorous competition for leases will guarantee receipt of fair market value. For skeptics, some assurance such as price data may be needed. Although oil and gas leasing continues because its long history produced extensive information, the effort to establish competitive bidding for coal leases was undone by the absence of data.

12. CONCLUSION

The prior discussion suggests the diverse and profound ways that regulation affects coal. Clearly, many questions arise about these actions and the available information is inadequate to resolve the concerns. What is clear is the underlying belief that coal is such an unattractive fuel that its production and use must be tightly controlled. Thus, although the 1970s were years of advocacy of increased coal use, the 1990s saw advocacy of discouraging coal consumption.

SEE ALSO THE FOLLOWING ARTICLES

Air Pollution from Energy Production and Use • Clean Coal Technology • Coal, Chemical and Physical Properties • Coal, Fuel and Non-Fuel Uses • Coal Industry, History of • Equity and Distribution in Energy Policy • Markets for Coal • Natural Gas Industry, Energy Policy in • Transportation and Energy Policy

Further Reading

Coase, R. H. (1988). "The Firm, the Market and the Law." Univ. of Chicago Press, Chicago.

Ellerman, A. D., Joskow, P. L., Schmalensee, R., Montero, J-P., and Bailey, E. M. (2000). "Markets for Clean Air: The U.S. Acid Rain Program." Cambridge Univ. Press, Cambridge, UK.

Gordon, R. L. (1988). Federal coal leasing: An analysis of the economic issues, Discussion Paper No. EM88-01. Resources for the Future, Energy and Materials Division, Washington, DC.

Gordon, R. L. (2001). Don't restructure electricity: Deregulate. *Cato J.* **20**(3), 327–358.

McDonald, S. (1979). "The Leasing of Federal Lands for Fossil Fuels Production." Johns Hopkins Univ. Press for Resources for the Future, Baltimore.

Mueller, D. C. (2003). "Public Choice III." Cambridge Univ. Press, Cambridge, UK.

Nelson, R. H. (1983). "The Making of Federal Coal Policy." Duke Univ. Press, Durham, NC.

Parker, M. J. (2000). "Thatcherism and the Fall of Coal." Oxford Univ. Press for the Oxford Institute for Energy Studies, Oxford.

Samuelson, P. A. (1954). The pure theory of public expenditure. *Rev. Econ. Statistics* **36**(4), 387–389. [Reprinted in Stiglitz, J. E. (Ed.). (1966). "The Collected Papers of Paul A. Samuelson," Vol. 2, pp. 1223–1225. MIT Press, Cambridge; and Arrow, K. J., and Scitovsky, T. (Eds.) (1969). "Readings in Welfare Economics," pp. 179–182. Irwin, Homewood, IL]

Samuelson, P. A. (1969). Pure theory of public expenditures and taxation. *In* "Public Economics" (Margolis, J., and Guitton, H., Eds.), pp. 98–123. St. Martins, New York. [Reprinted in Merton, R. C. (1972). "Collected Scientific Papers of Paul A. Samuelson," Vol. 3, pp. 492–517. MIT Press, Cambridge, MA]

Tarlock, A. D. (1985). The making of federal coal policy: Lessons for public lands from a failed program, an essay and review. *Natural Resources J.* **25**, 349–371.

U.S. Commission on Fair Market Value Policy for Federal Coal Leasing (1984). "Report of the Commission." U.S. Government Printing Office, Washington, DC.

U.S. Congress, Office of Technology Assessment (1984). "Acid Rain and Transported Air Pollutants: Implications for Public Policy." U.S. Government Printing Office, Washington, DC.

U.S. National Acid Precipitation Assessment Program (1991). "Acidic Deposition: State of Science and Technology." U.S. Government Printing Office, Washington, DC.

Coal Industry, History of

JAAK J. K. DAEMEN
University of Nevada
Reno, Nevada, United States

1. Introduction
2. Pre- and Early History
3. Middle Age and Renaissance
4. Precursors to the Industrial Revolution
5. Industrial Revolution
6. 19th Century
7. 20th Century
8. The Future
9. Coal and Culture

Glossary

coal A black or brown rock that burns; a solid combustible rock formed by the lithification of plant remains; a metamorphosed sedimentary rock consisting of organic components; a solid fossil hydrocarbon with a H/C ratio of less than 1, most commonly around 0.7 (for bituminous coal); a solid fossil fuel; fossilized lithified solar energy.

Coal Age The historical period when coal was the dominant fuel, from late 18th through middle 20th century; name of a trade journal devoted to the coal industry (ceased publication summer 2003).

Coalbrookdale A town in England's Black Country, symbol of the Industrial Revolution; site of first iron production using coke produced from coal; center of early steam engine and railroad development; celebrated in paintings by Williams, de Loutherbourg, Turner, and others, damned in poems by Anna Seward and others; part of the Ironbridge Gorge UNESCO World Heritage Site; a British national monument.

coal field Region in which coal deposits occur.

coal gas Fuel rich gas produced by partial oxidation (burning) of coal (also producer gas).

coal liquefaction Method for producing liquid fuels (e.g., gasoline, diesel) from coal.

coal preparation Treatment of coal to prepare it for any particular use; improve the quality of coal to make it suitable for a particular use by removing impurities, sizing (crushing, screening), and special treatments (e.g., dedusting); upgrading (beneficiation) of coal to a more uniform and consistent fuel (or chemical feedstock) of a particular size range, with specified ash, sulfur, and moisture content.

Industrial Revolution Controversial term to design the period when modern industrial society developed, initially applied primarily to Britain, later to other countries as well. From about 1760 to about 1850.

International Energy Agency (IEA) Autonomous body within OECD (Organisation for Economic Co-operation and Development) to implement an international energy program. Publishes reports on the coal industry, many of which are posted on its Web page.

water gas Gas produced by the interaction of steam on hot coke, used for lighting (primarily during the 19th through early 20th century) and as fuel (well into 20th century).

In a narrow sense, the coal industry can be considered as the coal mining industry, including coal preparation. More broadly it includes major users, such as steam generators for electric power, coking for steel production, and, historically, coal as a feedstock for chemicals and as transportation fuel, especially steam locomotives and ships, as well as the wide range of applications of steam engines. The transportation of coal, given its bulk and low value, can be a major cost of its use. Coal exploration and geology are part of the front end of the coal mining cycle.

1. INTRODUCTION

Coal has been mined for centuries. Until shortly before the Industrial Revolution, its use was local, and it did not make a significant contribution to the overall energy consumption of the world. Coal use increased rapidly in the two centuries before the Industrial Revolution and has continued to grow ever since, with occasional temporary down dips. Coal was the dominant fuel during the 19th century and the first half of the 20th century. The rise of modern

 © 2004 Elsevier Inc. All rights reserved.

society is intimately intertwined with the growth of the coal industry. Developments driven by coal with a major impact on technological progress include the steam engine, the railroad, and the steamship, dominant influences on the formation of modern society. For over a century, coal was the major energy source for the world. Its relative contribution declined over the second half of the 20th century. In absolute terms its contribution continues to increase.

The environmental disadvantages of coal have been recognized for centuries. Efforts to reduce these disadvantages accelerated over the last third of the 20th century. Because coal remains the largest and most readily available energy source, its use is likely to continue, if not increase, but will have to be supported by improved environmental control.

2. PRE- AND EARLY HISTORY

Where coal was mined and used initially is not clear. Secondary sources vary as to when coal use may have started in China, from as early as 1500 BC to as late as well after 1000 AD. It has been stated that a Chinese coal industry existed by 300 AD, that coal then was used to heat buildings and to smelt metals, and that coal had become the leading fuel in China by the 1000s. Marco Polo reported its widespread use in China in the 13th century. Coal may have been used by mammoth hunters in eastern central Europe.

The Greeks knew coal, but as a geological curiosity rather than as a useful mineral. Aristotle mentions coal, and the context implies he refers to "mineral" or "earth" coal. Some mine historians suggest he referred to brown coal, known in Greece and nearby areas. Theophrastus, Aristotle's pupil and collaborator, used the term ανθραζ(anthrax), root of anthracite. Although Theophrastus reported its use by smiths, the very brief note devoted to coal, compared to the many pages dealing with charcoal, suggest, as does archaeological evidence, that it was a minor fuel. The term is used with ambiguous meanings, but at least one was that of a solid fossil fuel. Theophrastus describes spontaneous combustion, still a problem for coal storage and transportation.

Coal was used in South Wales during the Bronze Age. The Romans used coal in Britain, in multiple locations, and in significant quantities. After the Romans left, no coal was used until well into the second millennium. Lignite and especially peat, geological precursors to coal, were used earlier and on a larger scale in northern and western Europe. Pliny, in the first century AD, mentions the use of earth fuel by inhabitants of Gaul near the Rhine mouth, to heat their food and themselves. All indications are that he describes the use of peat in an area currently part of The Netherlands, where peat still was used as a fuel nearly twenty centuries later. Romans observed coal burning near St. Étienne, later a major French coal mining center. The Romans carved jet, a hard black highly polishable brown coal, into jewelry. Jet was used as a gemstone in Central Europe no later than 10000 BC.

The recorded history of coal mining in India dates from the late 18th century. Place and river names in the Bengal-Bihar region suggest that coal may have been used, or at least that its presence was recognized, in ancient times.

Coal use was rare, even in most parts of the world where it was readily accessible in surface outcrops, until well into the second millennium, at which time its use became widespread, be it on a small scale.

The Hopis mined coal at the southern edge of Black Mesa, in northern Arizona, from about the 13th through the 17th century AD. Most was used for house fuel, some for firing pottery. Coal was mined by a primitive form of strip mining, conceptually remarkably similar to modern strip mining. At least a few underground outcrop mines were pursued beyond the edge of the last (deepest) strip, a practice reminiscent of modern auger mining. Gob stowing was used to prevent or control overburden collapse.

3. MIDDLE AGE AND RENAISSANCE

Coal and iron ore formed the basis of the steel industry in Liège in present Belgium. Coal mining rights were granted in Liège in 1195 (or 1198?) by the Prince Bishop. At about the same time, the Holyrood Abbey in Scotland was granted coal mining rights. A typical medieval situation involved abbeys or cloisters driving technology. In France, a 13th-century real estate document established property limits defined by a coal quarry. The earliest reliable written documentation in Germany appears to be a 1302 real estate transaction that included rights to mine coal. The transaction covered land near Dortmund in the heart of the Ruhr.

Coal use started before it was documented in writing. Religious orders tended to keep and preserve written materials more reliably than others. It seems reasonable to postulate that coal use began no later than the 12th century in several West European countries. Documented evidence remains anecdotal

for several more centuries but indicates that coal use increased steadily from the 12th century on. England led in coal production until late in the 19th century. Contributing to this sustained leadership were that wood (and hence charcoal) shortages developed earlier and more acutely there than in other countries, there was ready access to shipping by water (sea and navigable rivers), and large coal deposits were present close to the surface.

By the 13th century, London imported significant amounts of coal, primarily sea-coal, shipped from Newcastle-upon-Tyne. The use of coal grew steadily, even though it was controversial, for what now would be called environmental impact reasons. Smoke, soot, sulfurous odors, and health concerns made it undesirable. From the 13th century on, ordinances were passed to control, reduce, or prevent its use. None of these succeeded, presumably because the only alternatives, wood and charcoal, had become too expensive, if available at all. Critical for the acceptance of coal was the development of chimneys and improved fire places. By the early 1600s, sea-coal was the general fuel in the city. The population of the city grew from 50,000 in 1500 to more than 500,000 by 1700, the coal imported from less than 10,000 tons per year to more than 500,000 tons.

Early coal use was stimulated by industrial applications: salt production (by brine evaporation), lime burning (for cement for building construction), metal working, brewing, and lesser uses. By the late Middle Ages, Newcastle exported coal to Flanders, Holland, France, and Scotland.

By the early 16th century, coal was mined in several regions in France. In Saint Étienne, long the dominant coal producer, coal was the common household fuel, and the city was surrounded by metals and weapons manufacturing based on its coal. Legal and transportation constraints were major factors in the slower development of coal on the European continent compared to England. In the latter, surface ownership also gave subsurface coal ownership. In the former, the state owned subsurface minerals. Notwithstanding royal incentives, France found it difficult to promulgate coal development on typically small real estate properties under complex and frequently changing government regulations. Only late in the 17th century did domestic coal become competitive with English coal in Paris, as a result both of the digging of new canals and of the imposition of stiff tariffs on imported coal. Coal mining remained artisanal, with primitive exploitation "technology" of shallow outcrops, notably in comparison with the by this time highly developed underground metal mining.

The earliest coal mining proceeded by simple strip mining: the overburden was stripped off the coal and the coal dug out. As the thickness of the overburden increased, underground mines were dug into the sides of hills, the development of drift mines. Mining uphill, strongly preferred, allowed free water drainage and facilitated coal haulage. Some workings in down-dipping seams were drained by driving excavations below the mined seams.

Shafts were sunk to deeper coal formations. When coal was reached, it was mined out around the shaft, resulting in typical bell pits. Coal was carried out in baskets, on ladders, or pulled out on a rope. Where necessary, shafts were lined with timber.

In larger mines, once the coal was intersected by the shaft, "headings" were driven in several directions. From these, small coal faces were developed, typically about 10 ft wide, usually at right angles from the heading, and pillars, blocks of coal, were left in between these so-called bords, the mined-out sections. Pillar widths were selected to prevent roof, pillar, and overburden failure, which could endanger people and could lead to loss or abandonment of coal. Each working face was assigned to a hewer, who mined the coal, primarily with a pick. The hewer first undercut the coal face: he cut a groove along the floor as deep as possible. He then broke out the overhanging coal with picks, wedges, and hammers. Lump coal was hand loaded into baskets, usually by the hewer's helper. The baskets were pushed, dragged, or carried to and sometimes up the shaft. This haulage commonly was performed by women and children, usually the family of the hewer. The hewer operated as an independent contractor. He was paid according to the amount of coal he and his family delivered to the surface. Once established as a hewer, after multiple years of apprenticeship, the hewer was among the elite of workers in his community.

In deeper or better equipped mines, the coal baskets or corves were hoisted up the shaft using a windlass, later horse-driven gins. These also were used for hoisting water filled baskets or buckets, to dewater wet mines. By the end of the 17th century, shaft depths reached 90 ft, occasionally deeper. Encountering water to the extent of having to abandon pits became frequent. Improving methods to cope with water was a major preoccupation for mine operators.

Ventilation also posed major challenges. Larger mines sank at least two shafts, primarily to facilitate

operations. It was recognized that this greatly improved air flow. On occasion shafts were sunk specifically to improve ventilation. During the 17th century, the use of fires, and sometimes chimneys, was introduced to enhance air updraft by heating the air at the bottom of one shaft. In "fiery" mines, firemen, crawling along the floor, heavily dressed in wetted down cloths, ignited and exploded gas accumulations with a lighted pole. This practice was well established in the Liège basin by the middle of the 16th century. As mines grew deeper, and production increased, explosions became more frequent, on occasion killing tens of people.

In North America, coal was reported on Cape Breton in 1672. Some coal was mined for the French garrison on the island, some was shipped to New England before 1700. Coal was found by Joliet, Marquette, and Father Hennepin, in 1673, along the Illinois river. It is possible that some was used at that time by the Algonquins in this area.

4. PRECURSORS TO THE INDUSTRIAL REVOLUTION

The best sun we have is made of Newcastle coal.
—Horace Walpole, 1768

Whether or not one subscribes to the (controversial) concept of a first industrial revolution, the use of coal increased significantly from about the middle of the 16th century. A prime cause was the shortage of firewood, acute in England, noticeable in France. In Britain coal production increased from less than 3 million tons in 1700 to more than 15 million tons in 1800 (and then doubled again to more than 30 million tons per year by 1830). Driving were the growing industrial uses of coal, in particular the technology that made possible the smelting and forging of iron with coal. Steam power, developed to dewater mines, created a voracious demand for coal.

Land positions and legal statutes facilitated the development of coal. The frequent closeness of iron ore, limestone, and coal deposits stimulated the iron industry. Canals provided the essential low cost water transportation for inland coal fields. Tramways and turnpikes fed canal transport and led to railroads and highways.

Major technical advances were made in mining technologies: dewatering, haulage, and ventilation, challenges to mining that continue to this day and presumably always will.

Thomas Savery's *Miners Friend*, patented in 1698, promised the use of steam for a mine dewatering pump. Thomas Newcomen's steam-driven pump fulfilled the promise. The first Newcomen steam pump, or "fire-engine" or "Invention to Raise Water by Fire," was installed at a coal mine in 1712. Although expensive and extraordinarily inefficient by modern standards or even by comparison with Watt's steam engine of 60 years later, 78 Newcomen engines operated by 1733. The first Boulton and Watt engine, based on Watt's patent, was installed at a coal mine in 1776. By the end of the century, Boulton and Watt engines also found increasing use for shaft hoisting. For most of the century, horse power had been the dominant shaft hoisting method.

Underground haulage improved. Wheels were mounted on the baskets previously dragged as or on sledges. Planks were placed on the floor and were replaced by wooden and eventually iron rails—forerunners of the railroad. Horses replaced boys for pulling, at least in those mines where haulageways were large enough.

Dewatering became practical to great depths. Haulage improved greatly. Coping with explosive gases proved difficult. Fires at shaft bottoms remained the dominant method to induce airflow. Numerous ingenious devices were invented to course air along working faces, as air coursing was pursued vigorously. But explosions remained a major and highly visible hazard, resulting in an increasing cost in lives.

The physical coal winning—hewer with pick—remained unchanged. Improvements in production were associated with better lay-outs of the workings. Most mines continued to operate the room and pillar method, leaving over half the coal behind. An exception was the development of the longwall method. Here, a single long face was mined out entirely. As the coal was mined, the roof was supported with wooden props. The gob, the mined out area, was back-filled with small waste coal. This method allowed a nominal 100% recovery or extraction but resulted in surface subsidence. It now has become, in a highly mechanized form, the dominant underground coal mining method in the United States, Australia, and Europe.

The European continent lagged far behind Britain in all aspects of coal industry development: production, technology, transportation, use, and legal framework. By 1789, France produced somewhat more than 1 million tons per year, about one-tenth of that produced in Britain. The Anzin Coal Company, formed after considerable government prodding to

try to increase coal production, mined 30% of all French coal. France now recognized the growing threat posed by British industrialization.

Although by 1800 some 158 coal mines operated in the Ruhr, production that year barely exceeded 250,000 tons. The first steam engine at a coal mine in the region became operational only in 1801.

In North America, wood remained widely available at low cost, as well as water power in the more developed areas (e.g., New England). There was no need to use coal, and little was used. Major deposits were discovered and would prove extremely significant later on.

Coal was known to exist near Richmond, Virginia, by 1701, and some was mined by 1750. Some was sold to Philadelphia, New York and Boston by 1789. In 1742, coal was discovered in the Kanawha Valley, West Virginia. The Pocahontas seam, a major coal source for many decades, was mapped in Kentucky and West Virginia in 1750. An extensive description of coal along the Ohio River was made in 1751. George Washington visited a coal mine on the Ohio River in 1770. Thomas Jefferson recorded coal occurrences in the Appalachian mountains.

5. INDUSTRIAL REVOLUTION

We may well call it black diamonds. Every basket is power and civilization. For coal is a portable climate. It carries the heat of the tropics to Labrador and the polar circle; and it is the means of transporting itself whithersoever it is wanted.
—Emerson, ***Conduct of Life,*** *1860*

Coal fueled the Industrial Revolution, arguably the most important development in history. It transformed the world. While the term "Industrial Revolution" is controversial, it remains helpful to identify that period when fundamental changes developed in the basic economic structure, initially in Great Britain, shortly thereafter on the European and North American continents, and eventually throughout much of the world. The transition to the predominant use of fossil fuels, initially coal, was a major aspect and driving force of industrialization.

In Britain, the Industrial Revolution usually is considered the period from about 1760 through about 1840. In Belgium it started about 1830; in France, the United States, and Germany it started within the following decades. Other West European countries followed well before the end of the 19th century. Russia and Japan initiated their industrial revolutions by the end of the century.

The reasons why Britain took the lead are complex and involve all aspects of society. Readily accessible coal and iron ore in close proximity to each other was a major factor. Mine dewatering needs led to the development of the steam engine and mine haulage to railroads, two core technologies of the Industrial Revolution.

Belgium had coal and iron ore of good quality in close proximity and the necessary navigable river system. In France and Germany, coal and iron ore were far apart, and in France they were not of a quality that could be used for steel production until late in the century, when its metallurgy was understood. France and Germany lacked the water transportation necessary for the development of a heavy industry (Fig. 1). Steam locomotives and steam ships permanently altered travel and perceptions of space and time. Both were major users of coal and required that coal be provided at refueling points all around the world, starting worldwide coal trade and transportation. For many decades, Britain dominated this coal shipping and trading. Coal bunkering ports were major way stations toward building the British Empire.

Steam converted the Mississippi into the commercial artery it still is, although it is now diesel fueled. Railroads opened up the Midwest, the West, and later Canada. As the railroads expanded, so did the coal mines feeding them. Anthracite built eastern Pennsylvania, bituminous coal built Pittsburgh.

Coal gas and water gas dramatically, often traumatically, increased available lighting and changed living habits. Initially used in manufacturing plants, to allow around the clock work, gas light slowly conquered streets and homes and turned Paris into the "city of light."

Social, political, and cultural revolutions accompanied the Industrial Revolution. Working and living conditions of the lower classes became a societal issue and concern and industrial societies accumulated the wealth needed to address such concerns. Dickens, Marx, Engels, and many others identified social problems and proposed solutions either within or outside the existing political structure.

Coal mining was marked by difficult labor relations from the beginning and continues to be so in most parts of the world where it is a significant economic activity. Responsibility of the government in the social arena found its expression in multiple studies and reports and numerous labor laws. A focal point was child and female labor. The age at which children were allowed to start working underground was gradually raised. In Belgium, women miners were

FIGURE 1 *The Coal Wagon.* Théôdore Géricault (1821). Transportation cost always has been a major item in coal use. Road haulage was and is extremely expensive. A significant advantage for early British industrial development was its ready access to sea, river, and canal haulage of coal. Courtesy of the Fogg Art Museum, Harvard University Art Museums Bequest of Grenville L. Winthrop. Gray Collection of Engravings Fund. © President and Fellows of Harvard College. Used with permission.

more adamant and more successful in their opposition to being banned from underground and worked there legally until late in the 19th century (Fig. 2).

Although science and the scientific approach were well developed by this time, they had little influence on the early technological developments. The lack of geological understanding made the search for deeper coal uncertain and expensive. Geological sciences grew rapidly toward the end of this period and would soon make major contributions to finding coal and understanding the complexities of coal—even though its origin and formation remained a matter of controversy for decades. While the official geological community had little interaction with coal mining, William Smith, author of the first geological map of Britain, a masterpiece of geological mapping, was involved in coal mines, although his prime engineering work was for canals, mainly coal shipping canals. "Strata Smith" was a founder of stratigraphy because of his recognition of the possibility to use fossils to correlate strata. The first paper published by James Hutton, the founder of modern geology, was a coal classification. Hutton devised a method to allow customs agents and shippers to agree on a coal classification to set customs fees. While Hutton recognized the usefulness of the vast amount of geological information disclosed by coal mining, his later interest was strictly scientific. He did present a theory for coal formation based on metamorphosis of plant materials. While coal mining influenced the early development of geology, its impact was less than that of metal mining.

During the 1840s, the leading British geologist Charles Lyell visited several coal mining areas in North America. He proposed igneous activity combined with structural disturbance as the mechanism of formation of the anthracites in northeastern Pennsylvania, then the booming heart of the U.S. Industrial Revolution. He used the creep of mine roof and floor as an analog for the slow large deformations of rock formations.

As in Britain, much early U.S. geological mapping was for surveys for canals built during the beginning of the U.S. Industrial Revolution. Because much of the initial geological mapping was utilitarian, including looking for building stone and metal ore, at the time when the coal industry was developing, finding and mapping coal was a major objective of early geologists. This included the First Geological Survey of Pennsylvania, and the mapping of the upper Midwest by David Dale Owen, first state geologist of Indiana.

Although the connection between coal and geology may seem obvious, and it was far less so then than now, even less obvious might be relations between Industrial Revolution technology and fundamental geology. Yet Hutton, good friend of James Watt, was influenced by the *modus operandi* of the steam engine in his understanding of the uplifting of mountains.

FIGURE 2 *The Bearers of Burden*. Vincent Van Gogh (1881). The first bearer carries a safety lamp, suggesting she came from underground, which still was legal in the Belgian Borinage coal mining region when Van Gogh worked there in the 1870s. Collection Kröller-Müller Museum, Otterlo, The Netherlands. Photograph and reproduction permission courtesy of the Kröller-Müller Foundation.

Among the members of the Oysters Club in Edinburgh that included James Watt and James Hutton was Adam Smith. *The Wealth of Nations* addresses many strengths and weaknesses of the coal industry, while laying the theoretical economic basis for the free market economy that allowed it and the Industrial Revolution to thrive. Adam Smith identified the need for a coal deposit to be large enough to make its development economically feasible, the need for it to be located in the right place (accessible to markets, i.e., on good roads or water-carriage), and the need for it to provide fuel cheaper than wood (i.e., to be price competitive). Also, as Smith stated, "the most fertile coal-mine regulates the price of coals at all other mines in its neighbourhood." He discussed the appropriate purchase price and income of a coal mine and the economic differences between coal and metal mines—in sum, the basics of mineral and fuel economics.

The French engineer-scientist Coulomb reported on steam engines based on his observations during a trip to Britain and pushed French industrialists toward the use of coal. It was recognized that steam engines remained extremely inefficient. Scientific investigations of the performance of the engines evolved toward the science of thermodynamics. In 1824, Carnot published a theoretical analysis of the heat engine, a founding publication of thermodynamics, although the term was introduced only 25 years later by Lord Kelvin in his *Account of Carnot's Theory*. From 1850 on there was considerable interaction between the development of the steam engine and of thermodynamics, as demonstrated by the involvement in both of Rankine, one of the Scottish engineers-scientists who drove the technological developments of the Industrial Revolution.

6. 19TH CENTURY

The 19th century was the century of King Coal. Coal production grew dramatically in most countries that became industrialized, and coal use grew in all of them. While Britain continued to lead in coal production, the growth rate in the United States was so fast that by the end of the century its production exceeded Britain's (Table I). In 1900, coal exports accounted for 13.3% of total British export value. Coal trading was greatly liberated during the first half of the century. Direct sales between coal producers and consumers (e.g., gas producers) became legal. In this free and open domestic trading environment coal flourished. Internationally, the Empire encouraged and protected its domestic industry, supporting domestic production for export.

TABLE I

Annual Coal Production for Some Countries, in Millions of Tonnes[a] (Mt) per Year

Year	World	United States	United Kingdom	Belgium	Germany	France	Japan	China	Australia	South Africa	Indonesia	Poland[c]	Russia/USSR/FSU[d]	Colombia	India
1700			2.5			<0.1									
1800		0.1	>10 (15?)		1	1									
1815			22.6		<1.3	0.9									
1830		0.9	22.8 (30.9)	2.3	1.8	1.9									
1850		6.4	69.5	5.8	4.2	5									
1860	135	13.2	80.0	9.6	12.3	8.3									
1870									0.9			8	0.4		0.4
1871		42.5	117.4	13.8	24	13.2									
1880	332	64.9	147.0	16.9	47	19.4	0.9		1.6	<0.1		14	2		1
1890	512	143.2	181.6	20.4	70.2	26.1	2.6								
1900	700	244.2	225.2	23.5	109.3	33.4	7.4	<1	6.5	0.9	0.2	31	10		6.2
1910	1160	454.6	264.4	25.5	151.1	38.3	15.7	4.2							
1913	1341	517.2	287.4	22.9	277.4[b]	40.9	21.3	14.0	12.6	8.9	0.5	36	32.2	<0.1	16.5
1920	1320	597.1	233.2	22.4	131.4	24.3	30.5	19.5	13.2	10.4	1.1	32	7.6		18.2
1930	1414	487.1	262.1	27.4	142.7	53.9	33.4	26.4	11.5	12.2	1.7	34.0	36.2		21.9
1932	1124	326.1	212.1	21.4	104.7	46.3	28.1	28.0	11.3	9.9		28.8	53.7		22.0
1940	1497	462.0	227.9	25.6	240.1	39.3	57.3	46.5	11.8	17.2	2.0	77.1	148.7	0.5	29.9
1950	1508	509.4	220	27.3	113.8	52.5	38.5	41.1	16.5	26.5	0.8	78.0	185.2	1.0	32.8
1952	1496	484.2	228.5	30.4	143.7	52.4	43.4	66.6	19.4	28.1	1.0	84.4	215.0	1.0	36.9
1960	1991	391.5	197.8	22.5	148.0	56.0	57.5	397.2	21.9	38.2	0.7	104.4	374.9	2.6	52.6
1970	2208	550.4	147.1	11.4	118.0	37.8	40.9	354.0	45.4	54.6	0.2	140.1	206.7	2.75	73.7
1980	2810	710.2	130.1	8.0	94.5	20.2	18.0	620.2	72.4	115.1	0.3	193.1	553.0	4.2	113.9
1990	3566	853.6	94.4	2.4	76.6	11.2	8.3	1050.7	158.8	174.8	10.5	147.7	527.7	21.4	211.7
2000	3639	899.1	32.0	0.4	37.4	4.4	3.1	1171.1	238.1	225.3	78.6	102.2	321.6	37.1	309.9

[a] 1 tonne = 1000 kg = 2204.6 lb = 1.102 (short) tons

[b] Production for pre–World War I territory, sum of hard coal and brown coal (lignite).

[c] There are considerable differences in data, especially before 1920, depending on territory considered—that is, on how changed boundaries (one of which has intersected, in different ways, Upper Silesia, the major coal producing area) affected coal production. Given is an exceedingly simplified summary of very approximate production data for the area currently (2003) Poland.

[d] FSU = Former Soviet Union (data from 1980 through 2000). Data from Coal Information (2001), "International Energy Agency," Paris, France (2001); "Energy Information Adminstration," U.S. Department of Energy, Washington, D.C.; "Minerals Yearbook, U.S. Department of the Interior," Washington, D.C., (1918), (1923), (1934), (1950); "Historical Statistics of the United States to 1957: A Statistical Abstract Supplement," Washington, D.C. (1960); B. R. Mitchell, "International Historical Statistics, Europe (1750–1988)," 3rd ed. Stockton Press, New York, NY (1992); B. R. Mitchell, "International Historical Statistics, Africa, Asia & Oceania," 2nd Rev. ed., Stockton Press, New York, NY (1995); K. Takahashi (1969), "The Rise and Development of Japan's Modern Economy," The Jiji Tsushinsha (The Jiji Press, Ltd.), Tokyo; A. L. Dunham (1955), "The Industrial Revolution in France (1815–1848)," Exposition Press, New York; B. R. Mitchell (1962), "Abstract of British Historical Statistics," Cambridge at the University Press; W. W. Lockwood (1954), "The Economic Development of Japan," Princeton University Press, Princeton, NJ; A. S. Milward and S. B. Saul (1973), "The Economic Development of Continental Europe," Rowman and Littlefield, Totowa, NJ; W. Ashworth (1975), "A Short History of the International Economy Since 1850," Longman, London; M. Gillet (1973), "Les Charbonnages du Nord de la France au XIX^{e} Siècle," Mouton, Paris; R. Church (1986), "The History of the British Coal Industry, Vol. 3 (1830–1913): Victorian Pre-eminence," Clarendon Press, Oxford; V. Muthesius (1943), "Ruhrkohle (1893–1943)," Essener Verlagsanstalt, Essen, Germany; B. R. V. Mitchell (1980), "European Historical Statistics (1750–1975)," *Facts on File*, New York; M. W. Flinn (1984), "The History of the British Coal Industry, Vol. 2 (1700–1830): The Industrial Revolution," Clarendon Press, Oxford; J. A. Hodgkins (1961), "Soviet Power," Prentice-Hall, Englewood Cliffs, NJ; H. N. Eavenson (1935), "Coal Through the Ages," AIME, New York; S. H. Schurr and B. C. Netschert (1960), "Energy in the American Economy (1850–1975)," The Johns Hopkins Press, Baltimore; "A History of Technology," T. I. Williams, Ed. (1978), Clarendon Press, Oxford; "COAL, British Mining in Art (1680–1980)," Arts Council of Great Britain, London; T. Wright, Growth of the modern chinese coal industry, "Modern China," Vol. 7, No. 3, July 1981, 317–350; R. L. Gordon (1970), "The Evolution of Energy Policy in Western Europe," Praeger, New York; J. S. Furnivall (1939) (1967), "Netherlands, India, A Study of Plural Economy," Cambridge at the University Press; D. Kumar, Ed. (1983) "The Cambridge Economic History of India," Cambridge University Press; Z. Kalix, L. M. Fraser, and R. I. Rawson (1966), "Australian Mineral Industry: Production and Trade (1842–1964)," Bulletin No. 81, Bureau of Mineral Resources, Geology and Geophysics, Commonwealth of Australia Canberra; P. Mathias and M. M. Postan, Eds., "The Cambridge Economic History of Europe," Vol. VII, Part I, Cambridge University Press, Cambridge (1978); N. J. G. Pounds, "The Spread of Mining in the Coal Basin of Upper Silesia and Southern Moravia," *Annals of the Association of American Geographers*, Vol. 48, No. 2, 149–163 (1958). Most data prior to 1900 must be considered as subject to large uncertainties and to significant differences between different sources.

Capital was needed to develop the mines—for example, to sink shafts, construct surface and underground operating plant (e.g., pumps, hoisting engines, fans), build loading, processing, and transportation facilities. Coal mine ownership slowly shifted from private and partnerships to stock-issuing public corporations. Employment in British coal mining grew from about 60,000 to 80,000 in 1800 to nearly 800,000 in 1900.

In 1800, the United States produced barely over a 100,000 tons of soft coal and almost certainly much less anthracite. Although many Pennsylvania anthracite outcrops were well known by then and were within hauling distance from Philadelphia, Boston, and New York City, no satisfactory method for burning anthracite had been developed.

The American Industrial Revolution started around 1820, in parallel with the growth of the anthracite industry in northeastern Pennsylvania. Canals were built to provide transportation but were soon superseded by railroads. Anthracite had been promoted and used somewhat as a domestic fuel late in the 18th century. It did not become accepted until efficient fire grates became available, well into the 19th century, and until the price of wood fuel had risen significantly, particularly in the large cities. Following intense development of burning technologies, anthracite became the fuel of choice for public buildings and was used in steam engines (stationary, e.g., for coal mine pumping, manufacturing, and mobile: locomotives and boats) and iron production. The commercialization of improved and easier to use stoves made anthracite the home heating fuel of choice for well over the next century.

Early mining of anthracite was easy, as it was exposed in many outcrops. Many small operators could start mining it with little or no capital, manpower, or knowledge required. Once underground, the geological complexity of the intensely folded and faulted deposits required operating knowledge that led to consolidation in the industry. The drift mines and shallow shaft mines (rarely deeper than 30 ft) started in the 1810s were followed in the 1830s by slope mines. Breast and pillar mining was common: coal pillars were left in between the breasts (faces, rooms) that were mined out. The pillars carried the weight of the overburden. Horses and mules pulled cars running on rails. Black powder was used to shoot the coal (break out the coal). Wooden props (timber posts) provided safety by holding up the immediate roof.

Contract mining was standard. Each miner worked a breast, assisted by one or two helpers who loaded the coal and installed timber. The miner was paid on a piece rate and paid his helpers. The miner worked as an independent, with minimal supervision. Death and injury rates were high, predominantly from roof falls, not from the explosions that occurred all too often. By the end of the century the fatality rate in U.S. coal mining approached 1000 per year.

Over the course of the 19th century, numerous safety lamps were developed, typically designed with a protective wire mesh screen and glass to prevent the flame from igniting an explosion. The Davy lamp is the most famous, although it was not the one most widely used. Whether these lamps improved safety, (i.e., reduced accidents) or whether they simply allowed miners to work in more gaseous (i.e., more dangerous) conditions remains controversial.

Bituminous or soft coal mines operated with room and pillar mining, similar to the anthracite mines, but predominantly in flat or nearly flat beds—not the steeply dipping beds of the anthracite region. In 1871, virtually all coal was still undercut by hand and shot with black powder. Coal preparation was almost nonexistent. Animal haulage was universal.

Major efforts were started to mechanize coal mining. Cutter machines, designed to replace the manual undercutting, were patented, but it took several more decades before they performed well and became accepted. Late in the century, electric locomotives were introduced. They quickly gained widespread acceptance, replacing animal haulage. In 1871, most mines that used artificial ventilation—and many did not—used furnaces. By the end of the century, mechanical fans dominated. They were driven by steam engines, as were the pumps that dewatered wet mines.

By the end of the century, a pattern of difficult labor relations was well established in coal fields around the world. The first strike in the Pennsylvania anthracite region took place in 1849. Labor actions and organizations started in Scotland, England, and Wales in the 18th century. Even though by the early 19th century miners' wages were substantially higher than those in manufacturing, the relentless demand for labor, driven by the rapidly increasing demand for coal, facilitated the growth of labor movements, notwithstanding the dominant political power of the coal producers. The high fatality and injury rate gave impetus to a labor emphasis on improving safety.

The 1892 first national miners strike in Britain stopped work from July through November. Two men were killed in riots. In the 1890s the UMWA (United Mine Workers of America) was trying to

organize. General strikes were called in 1894 and 1897. In 1898, an agreement was reached between the UMWA and the main operators in Illinois, Ohio, Indiana, and western Pennsylvania. Missing from this list are the anthracite region of northeastern Pennsylvania and West Virginia.

By the end of the century, coal had broadened its consumption base. In 1900, the world produced 28 million tons of steel, the United States, 11.4 million tons. Coking coal for steel had become a major coal consumer. To improve steel quality, steel producers tightened specifications on coke and thereby on the source coal.

Coal gas was introduced early in the century, and had become the major light source in both large and small cities by midcentury. Electric light was introduced by the end of the century. Electric power generation became a major coal user only by the middle of the 20th century, however. Both the conversion to gas light (from candles) and the later one to electric light (from gas) required adjustments on the part of the users. For both changes a major complaint was the excessive brightness of the new lights. (Initially electric light bulbs for domestic use were about 25 W.)

The heavy chemical industry received a major boost when it was discovered that coal tar, a waste by-product of coke and gas production, was an excellent feedstock for chemicals, notably organic dyes. Discovered in Britain, shortly after mid-century, Germany dominated the industry by the end of the century and did so until the first world war. The industry became a textbook example of the application of research and development (R&D) for the advancement of science, technology, and industry.

An important step during the 19th century was the development of coal classifications. Coal is complex and variable. It can be classified from many points of view. Geologically, coal classification is desirable to bring order in a chaotic confusion of materials (macerals) with little resemblance to the mineralogy and petrography of other rocks. Chemical classification is complicated by the fact that the material is a highly variable, heterogeneous mixture of complex hydrocarbons. From the users and the producers point of view, the buyer and the seller, some agreement needs to be reached as to what is the quality of the delivered and the received product. Depending on the use, "quality" refers to many different factors. It includes calorific value (i.e., how much heat it generates), moisture content, and chemical composition (e.g., how much carbon or hydrogen it contains). Impurities are particularly important (e.g., ash and sulfur content and the behavior of the coal during and after burning or coking).

From 1800 to 1889, world production of coal increased from 11.6 to 485 million tons. In 1899, coal provided 90.3% of the primary energy in the United States. It was indeed the century of King Coal.

7. 20TH CENTURY

Well into the second half of the 20th century, coal remained the world's major primary energy source. Throughout most of the century, the relative importance of coal declined—that is, as a fraction of total energy production coal decreased. Worldwide coal use increased steadily, but the major production centers shifted. The coal industry changed fundamentally.

Two world wars changed the world and the coal industry. During both wars most industrial countries pushed their steel production, and hence their coal production, as high as possible. Due to the wartime destructions, both wars were followed by severe coal shortages. In response, coal-producing nations pushed hard to increase coal production. In both cases, the demand peak was reached quickly and subsided quickly. A large overcapacity developed on both occasions, resulting in steep drops in coal prices, in major production cutbacks, and in severe social and business dislocations. After the second world war, two major changes impacted coal: railroads converted from steam to diesel and heating of houses and buildings switched to fuel oil and natural gas. (Diesel replaced coal in shipping after World War I.) In the United States, railroads burned 110 million tons of coal in 1946, 2 million tons in 1960. Retail deliveries, primarily for domestic and commercial building heating, dropped from 99 million tons in 1946 to less than 9 million tons in 1972. One hundred fifty years of anthracite mining in northeastern Pennsylvania was ending.

A remarkable aspect of the coal user industry is the growth in efficiency in using coal. Railroads reduced coal consumption per 1000 gross ton-miles from 178 lbs in 1917 to 117 lbs in 1937. One kWh of electrical power consumed 3.5 lbs of coal in 1917, 1.4 lbs in 1937. A ton of pig iron required 2,900 lbs of coking coal in 1936, down from 3500 lbs in 1917. Improvements in use efficiency continued until late into the century. Coal consumption per kWh of electric power dropped from 3.2 lbs in 1920 to 1.2 lbs in 1950, and to 0.8 lbs in the 1960s. After that efficiency decreased somewhat due to inefficiencies

required to comply with environmental regulations. Super efficient steam generating and using technologies in the 1990s again improved efficiencies. Even more dramatic was the efficiency improvement in steel production (i.e., the reduction in coal needed to produce steel). Concurrent with the loss of traditional markets came the growth in the use of coal for generating electrical power, the basis for the steadily increasing demand for coal over the later decades of the century and for the foreseeable future.

Major shifts took place in worldwide production patterns. Britain dropped from first place to a minor producer. Early in the century, the United States became the largest coal producer in the world and maintained that position except for a few years near the very end of the century when the People's Republic of China became the largest producer. Most West European countries and Japan reached their peak production in the 1950s, after which their production declined steeply. In the much later industrialized eastern European countries, in Russia (then the Soviet Union), and in South Korea, the peak was reached much later, typically in the late 1980s.

In Australia, India, and South Africa, coal production increased over most of the century, with major growth in the last few decades. The large production growth in China shows a complex past, with major disruptions during the 1940s (World War II) and the 1960s (cultural revolution). The recent entries among the top coal producers, Colombia and Indonesia, grew primarily during the 1980s and 1990s.

Worldwide production patterns have changed in response to major transportation developments. Large bulk carrier ships reduced the cost of shipping coal across oceans. While some international coal trading existed for centuries (e.g., from Newcastle to Flanders, Paris, and Berlin and later from Britain to Singapore, Cape Horn, and Peru), only during the second half of the 20th century did a competitive market develop in which overseas imports affect domestic production worldwide. Imports from the United States contributed to coal mine closures in Western Europe and Japan during the 1950s. Imports from Australia, South Africa, Canada, Poland, and the United States contributed to the demise of coal mining in Japan, South Korea, and most of Western Europe.

Inland, unit trains haul coal at competitive cost over large distances: Wyoming coal competes in Midwestern and even East Coast utility markets. It became feasible to haul Utah, Colorado, and Alberta coking coal to West Coast ports and ship it to Japan and South Korea.

Coal mining reinvented itself over the 20th century. A coal hewer from 1800 would readily recognize a coal production face of 1900. A coal miner from 1900 would not have a clue as to what was going on at a coal face in 2000.

The most obvious and highly visible change is the move from underground to surface mining (Fig. 3). Large earthmoving equipment makes it possible to expose deeper coal seams by removing the overburden. Although large-scale surface mining of coal started early in the century, by 1940 only 50 million tons per year was surface mined, barely over 10% of the total U.S. production. Not until the 1970s did surface production surpass underground production.

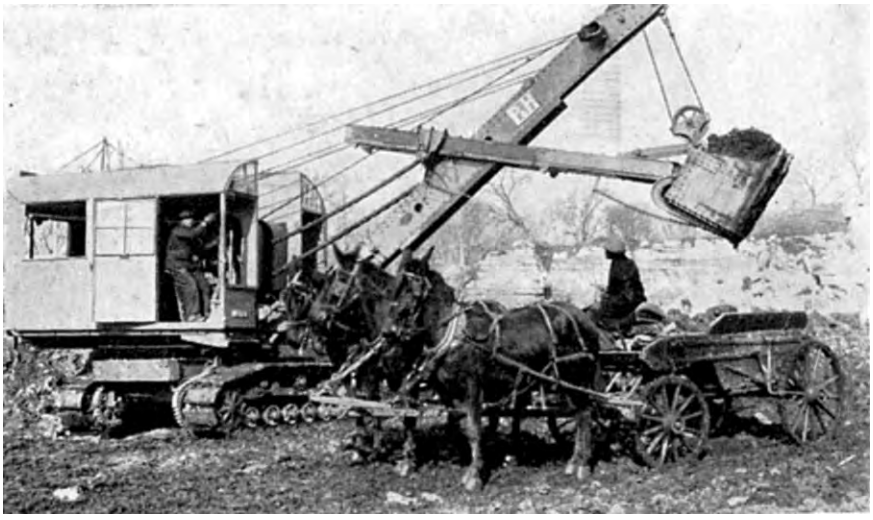


FIGURE 3 Early mechanized surface coal mining. A 1920s vintage P&H shovel loading in a coal wagon pulled by three horses. Photograph courtesy of P&H Mining Equipment, A Joy Global Inc. Company, Milwaukee, WI. Used with permission.

By 2000, two-thirds of U.S. coal production was surface mined.

Room and pillar mining dominated underground U.S. coal mining until very late in the century. Early mechanization included mechanical undercutting and loading. Conventional mining, in which the coal is drilled and blasted with explosives, decreased steadily over the second half of the century and was negligible by the end of the century. Continuous mining grew, from its introduction in the late 1940s (Fig. 4), until very late in the century, when it was overtaken by longwall mining (Fig. 5). Modern mechanized longwall mining, in which the coal is broken out mechanically over the entire face, was developed in Germany and Britain by the middle of the century. Geological conditions made room and pillar mining impractical or even impossible in many European deposits. In the last two decades of the century, American (and Australian) underground coal mines adopted longwalling, and greatly increased its productivity. In conjunction with the increased production arose a serious safety problem: coal is being mined so fast that methane gas is liberated at a rate difficult to control safely with ventilation systems.

Technological advances depend on equipment manufacturers. Surface mining equipment size peaked in the 1960s and 1970s. The largest mobile land-based machine ever built was the *Captain*, a Marion 6360 stripping shovel that weighed 15,000 tons. *Big Muskie*, the largest dragline ever built, swung a 220 cu yd bucket on a 310 ft boom. The demise of the stripping shovel came about because coal seams sufficiently close to the surface yet deep enough to warrant a stripping shovel were mined out. While the stripping shovel was exceedingly productive and efficient, its large cost required that it operate in a deposit that could guarantee a mine life of at least 10 to 20 years. Capital cost for a stripping shovel was markedly higher than for a dragline.

Large draglines shipped during the 1980s were mostly in the 60 to 80 cu yard bucket size range. A few larger machines (120 to 140 cu yd) were build in the 1990s. By the end of the century, the conventional mine shovel reached a bucket size approaching that of all but the largest stripping shovels ever built.

Worldwide research was conducted in support of the coal industry. The U.S. Bureau of Mines was established in 1910 and abolished in 1996. Its mission changed, but it always conducted health

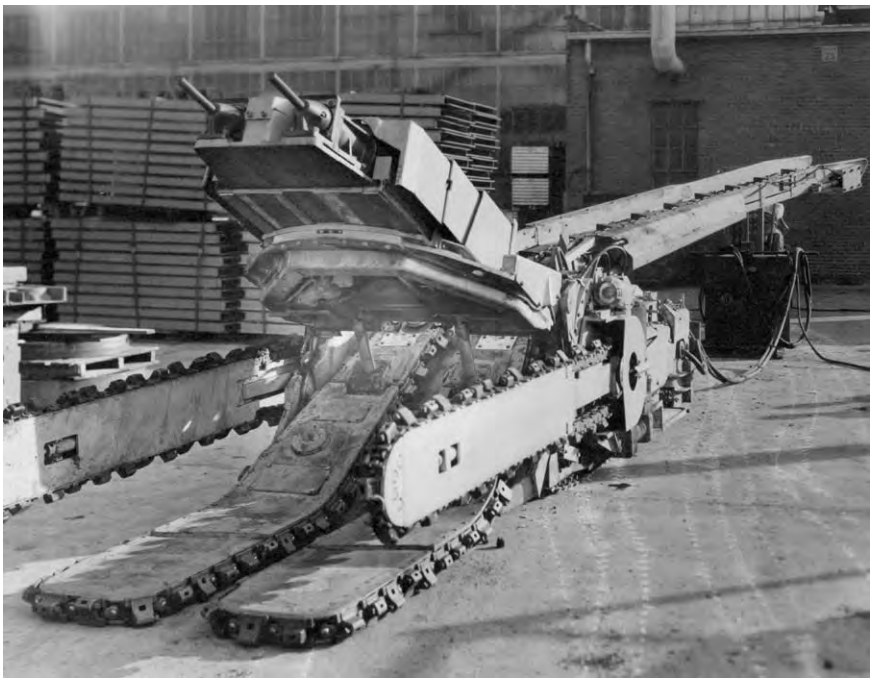


FIGURE 4 Early attempt at mechanized underground coal mining. The Jeffrey 34F Coal Cutter, or Konnerth miner, introduced in the early 1950s. The machine was designed to mechanize in a combined unit the most demanding tasks of manual coal mining: breaking and loading coal. Photograph and reproduction courtesy of Ohio Historical Society, Columbus, OH. Used with permission.

FIGURE 5 A major, highly successful advance in mechanizing underground coal mining: replacing manual undercutting by mechanical cutting. Jeffrey 24-B Longwall Cutter. Photograph and reproduction courtesy of Ohio Historical Society, Columbus, OH. Used with permission.

and safety research. The Bureau tested electrical equipment for underground coal mines, permissible explosives, designed to minimize the chances of initiating a gas or dust explosion, and improved ground control. The Bureau produced educational materials for health and safety training. In 1941, Congress authorized Bureau inspectors to enter mines. In 1947, approval was granted for a federal mine health and safety code. The 1969 Coal Mine Health and Safety Act removed the regulatory authority from the Bureau, and transferred it to the Mine Safety and Health Administration (MSHA).

Organizations similar to the Bureau were established in most countries that produced coal. In Britain, the Safety in Mines Research and Testing Board focused on explosions, electrical equipment, and health, the latter particularly with regard to dust control. In West Germany, the Steinkohlenbergbauverein was known for its authoritative work in ground control, especially for longwalls. CERCHAR in France, INICHAR in Belgium, and CANMET in Canada studied coal mine health and safety. In the Soviet Union and the People's Republic of China, highly regarded institutes ran under the auspices of their respective National Academy of Science.

Over the course of the 20th century, the classification of coal took on ever more importance, resulting in a proliferation of classification methods. Early in the century, when international coal trade was not common and user quality specifications less comprehensive, national and regional classifications were developed. As international coal trade grew, over the second half of the century the need arose for classification schemes that could be applied worldwide.

In situ coal gasification has been demonstrated and could be developed if economics made it attractive. Conceptually simple, a controlled burning is started in a coal seam to produce gas containing CO, H_2, CH_4, and higher order hydrocarbons. The complexity of the fuel, the variability of the deposits, and the potential environmental impacts complicate implementation.

You can't dig coal with bayonets.
—John L. Lewis, president, UMWA, 1956

You can't mine coal without machine guns.
—Richard B. Mellon, American industrialist

Difficult labor relations plagued coal mining through much of the century in most free economy countries that mined coal. In many parts of the world, coal miners formed the most militant labor

unions. The West Virginia mine wars, lasting for most of the first three decades of the century, were among the most prolonged, violent, and bitter labor disputes in U.S. history. In Britain, the number of labor days lost to coal mine strikes far exceeded comparative numbers for other industries. Intense violent labor actions, frequently involving political objectives, have recurred throughout much of the century in Britain, France, Germany, Belgium, Australia, Canada, and Japan. During the last few decades of the century, strikes in Western Europe, Poland, Japan, and Canada were driven largely by mine closure issues. Strikes in Russia, the Ukraine, and Australia dealt primarily with living and working conditions. Coal miners in Poland, Serbia, and Rumania were leaders, or at least followers, in strikes with primarily political objectives. The last major strikes in Britain also had a strong political component, although pit closure concerns were the root cause. In the United States, the last two decades of the century were remarkably quiet on the labor front, especially compared to the 1970s.

The structure of the coal mining industry changed significantly over the course of the 20th century. In the United States during the 1930s, many family-owned coal mining businesses were taken over by corporations. Even so, the historical pattern of coal mining by a large number of small producers continued until late in the century. Production concentration remained low compared to other industries and showed an erratic pattern until late in the century. In 1950, the largest producer mined 4.8% of the total, in 1970 11.4%, in 1980 7.2%. In 1950, the largest eight producers mined 19.4% of the total; in 1970, 41%; in 1980, 29.5%. High prices during the energy crisis of the 1970s facilitated entry of small independents. The top 50 companies produced 45.2% of the total in 1950, 68.3% in 1970, 66.3% in 1980, confirming the significant reduction of the small producers.

During the 1970s, oil and chemical companies took over a significant number of coal companies because they believed widely made claims during that decade of an impending depletion of oil and gas reserves. As the hydrocarbon glut of the 1980s and 1990s progressed, most of these coal subsidiaries were spun off and operated again as independent coal producers.

Toward the end of the century, there was significant consolidation of large coal producers, domestically and internationally. Even so, the industry remained characterized by a relatively large number of major producers. In 2000, the 10 largest private companies controlled barely over 23% of the world production. In the United States the largest producer, Peabody, mined 16% of U.S. coal, the second largest one, Arch Coal, 11%. Coal remained highly competitive, domestically and internationally.

8. THE FUTURE

Coal resources are the largest known primary energy resource. Supplies will last for centuries, even if use grows at a moderate rate. Reserves are widely distributed in politically stable areas. The many uses of coal, from electrical power generation to the production of gaseous or liquid fuels and the use as petrochemical feedstock, make it likely that this versatile hydrocarbon will remain a major raw material for the foreseeable future.

The mining and especially the use of coal will become more complicated and hence the energy produced more expensive. Coal mining, coal transportation, and coal burning have been subjected to ever more stringent regulations. This trend toward tighter regulations will continue. A major environmental factor that will affect the future of coal is the growing concern about global warming. While technologies such as CO_2 capture and sequestration are being researched, restrictions on CO_2 releases will add significantly to the cost of producing energy, in particular electricity, from coal.

Predictions for the near future suggest a modest, steady increase in coal production. The main competition in the next few decades will be from natural gas. Natural gas reserves have risen steadily over the past 30 years, in parallel with the increased demand and use, and hence the increased interest in exploration for gas. Natural gas is preferred because it is richer in hydrogen, poorer in carbon, and hence the combustion products contain more steam rather than CO_2. If, in the somewhat more distant future, the predictions of a reduction in supply of natural gas and oil were to come through—and for over a century such predictions have proved "premature"—coal might once again become the dominant fossil fuel. The rise in demand for electric power seems likely to continue in most of the world to reach reasonable living standards and in the developed world for such needs as electric and fuel cell–driven vehicles and continued growth in computers and electronics in general.

To make coal acceptable in the future, steps need to be taken at all phases of the coal life cycle, from production through end use. A major focus in the

production cycle is minimizing methane release associated with mining. Methane (CH_4) is a greenhouse gas. It also is a main cause of coal mine explosions. Great strides have been made in capturing methane prior to and during mining. In gassy seams, it now is collected as a fuel. In less gassy seams, especially in less technologically sophisticated mines, it remains uneconomical and impractical to control methane releases. Extensive research is in progress to reduce methane releases caused by coal mining.

Other environmental problems associated with mining coal include acid mine drainage, burning of abandoned mines and waste piles, subsidence, spoil pile stability issues, and mine site restoration and reclamation. Technological remedies exist, but their implementation may need societal decisions for regulatory requirements.

Coal preparation is critical for improving environmental acceptability of coal. Super clean coal preparation is feasible. Technically, virtually any impurity can be removed from coal, including mercury, which has drawn a great deal of attention over the last few years. Coal transportation, particularly in ocean going vessels, has modest environmental impacts, certainly compared to oil.

Coal users carry the heaviest burden to assure that coal remains an acceptable fuel. Enormous progress has been made in reducing sulfur and nitrogen oxide emissions. Capturing and sequestering CO_2 will pose a major challenge to the producers of electric power.

More efficient coal utilization contributes to the reduction in power plant emissions. Modern power plants run at efficiencies of about 37%. During the 1990s, power plants have come on stream that run at over 40%. It is likely that 50% can be achieved by 2010. Increasing efficiency from 37 to 50% reduces by one-third the coal burned to generate electricity and reduces by one-third gas (and other) emissions.

Coal has been attacked for environmental reasons for over seven centuries. With ups and downs, its use has grown over those seven centuries because of its desirable characteristics: low cost, wide availability, ease of transport and use. It will be interesting to see whether it can maintain its position as a major energy source for another seven centuries or whether more desirable alternatives will indeed be developed.

9. COAL AND CULTURE

Given the pervasive influence of coal and in particular of its uses on the fundamental transformations of societies, especially during the 19th century, it is not surprising to see coal reflected in cultural contexts. Pervasive was the sense of progress associated with industrial development, the perception of dreadful social problems associated with the industrial progress, and the disintegration of an older world.

Zola's *Germinal* remains the major novel rooted in coal mining, a classic in which have been read different, contradictory meanings from revolutionary to bourgeois conservative. D. H. Lawrence grew up in a coal mining town, and it and its collieries pervade several of his masterpieces. Again, these incorporate deep ambiguities with respect to coal mining: admiration for the male camaraderie in the pits, the daily dealing with a hostile dangerous environment, the solidarity in the mining community, and the stifling constraints of it all. Similar ambiguities are found in Orwell's *The Road to Wigan Pier* and in the reactions to it. It was published by the Left Book Club. These publishers inserted an apologetic introduction, justifying why they published this book, frequently so critical of the left. The book includes a widely quoted description of underground mining practices in Britain in the 1930s, as well as sympathetic descriptions of life in the mining communities—not deemed sufficiently negative by many on the left.

Richard Llewellyn's popular *How Green Was My Valley* introduces a small coal mining community in South Wales through a voluptuous description of rich miner's meals. (Orwell, as described in his personal diary, published long after *Wigan Pier*, similarly was struck by the rich miners meals.)

Lawrence's descriptions of coal mine towns in the early 20th century gained wide distribution through films made of his works, several of which were successful commercially. The critically acclaimed and commercially successful movie adaptation by John Ford of *How Green Was My Valley* introduced a wide audience to a coal mining community in South Wales. While justifiably criticized as sentimental, the book and film offer a sympathetic view of a community that often has felt looked down upon, examples of nostalgic flashback descriptions of mining communities often found in the regional literature of mining districts.

Joseph Conrad, in his 1902 story "Youth," fictionalized, although barely, the loss at sea, to spontaneous combustion, of a barque hauling coal from Newcastle to Bangkok. The self-ignition of coal, particularly during long hauls, was a major problem during the late 19th century and continues to pose headaches for those who haul or store coal over extended periods of time. "Youth" shows that

Conrad, an experienced coal hauling sailor, was thoroughly familiar with spontaneous combustion.

Upton Sinclair, a leading American muckraker of the early 20th century, based *King Coal* on the long and bitter 1913–1914 Colorado coal strikes. Neither a critical nor a commercial success (the sequel *The Coal War* was not published until over half a century later and then decidedly for academic purposes), it describes the recurring problem that has plagued coal forever: difficult labor relations.

Van Gogh sketched and painted female coal bearers during his stay in the Borinage, the heart of early coal mining in Belgium. Also originating in the Borinage were the sculptures by Constantin Meunier, including many of miners. The tour of his work through Eastern and Midwestern cities (1913–1914) brought to the American public a visual art that recognized the contribution, hardships, and strengths of blue collar workers.

More visible than coal mining, visual celebrations of the Industrial Revolution focused on the offspring of coal mining and its customers, railroads and the steam locomotive. Among the better known ones is *Rain, Steam and Speed*, by J. W. Turner, the painter of the British industrial landscape, immortalized in *Keelmen Heaving in Coals by Night, Coalbrookdale by Night*, and his paintings of steamships.

The opening up of the American landscape and space, or the intrusion into the landscape, was depicted by Cole's 1843 *River in the Catskills*, his *The Oxbow*, Melrose's *Westward the Star of Empire Takes Its Way*, Bierstadt's *Donner Lake*, and especially George Inness' *The Lackawanna Valley*, Henry's *The First Railway on the Mohawk and Hudson Road*, and Durand's *Progress*. During the 19th century, Currier & Ives' prints popularized progress and penetration of the wilderness thanks to the railroads and the steam engine. Walt Whitman's *Ode to a Locomotive* summarized the widely held view that a continent was being conquered and opened up thanks to the railroad and its most visible emblem, the locomotive's steam.

Nikolai Kasatkin, a leading Russian realist and one of the Itinerants (peredvizhniki—wanderers, travelers), lived several months in the Donetsk coal basin at a time (late 1800s) when Russian coal production was increasing rapidly. One of his most famous paintings, *Miners Changing Shift* (Fig. 6), is often referenced but rarely reproduced. His *Poor People Collecting Coal in an Abandoned Pit* and *Woman Miner*, both far more lyrical than the subject might suggest, also are far more readily accessible. Closer to socialist realism is Deineka's *Before Descending into the Mine*. One of the better known Soviet railroad celebrations is *Transport Returns to Normal* by Boris Yakovlev.

Charles Sheeler and Edward Hopper continue the American tradition of painting railroads in land- and cityscapes. American primitive John Kane, born in Scotland, includes coal barges on the river in his

FIGURE 6 *Coal Miners: Shift Change*. Nikolai A. Kasatkin (1895). Photograph and reproduction courtesy of Tretyakov State Gallery, Moscow.

exuberant *Monongahela River Valley* and *Industry's Increase*. Probably the only one among these artists who worked as a coal miner ("the best work I knew and enjoyed"), Kane published his autobiography *Sky Hooks*, one of the sunnier recollections among the many written by people who grew up in mining communities. More representative of the bleak conditions during the 1930s may be Ben Shahn's sad and haunting *Miners' Wives*.

In France, Théodore Géricault painted the horse haulage of coal, common well into the 19th century, in *Coal Wagon*. Manet, Monet, Caillebotte, Seurat, and Pissarro recorded their impressions of railroads, railroad stations, steamboats, and their impact on landscape and society. The responses ranged from reluctant acceptance, at best, by Pissarro, to the enthusiastic celebrations by Monet of railroads, steam engines, and in particular the Gare St. Lazare, the first railroad station in France and Paris. Monet's *Men Unloading Coal*, a rather darker picture than most Monets, illustrates the manual unloading of coal from river barges late in the 19th century.

SEE ALSO THE FOLLOWING ARTICLES

Coal, Chemical and Physical Properties • Coal, Fuel and Non-Fuel Uses • Coal Industry, Energy Policy in • Coal Mining, Design and Methods of • Coal Mining in Appalachia, History of • Coal Resources, Formation of • Electricity Use, History of • Manufactured Gas, History of • Markets for Coal • Natural Gas, History of • Nuclear Power, History of • Oil Industry, History of

Further Reading

Berkowitz, N. (1997). "Fossil Hydrocarbons: Chemistry and Technology." Academic Press, San Diego, CA.

Bryan, A. M., Sir. (1975). "The Evolution of Health and Safety in Mines." Ashire, London.

Gregory, C. E. (2001). "A Concise History of Mining." A. A. Balkema, Lisse, The Netherlands.

Jones, A. V., and Tarkenter, R. P. (1992). "Electrical Technology in Mining: The Dawn of a New Age." P. Peregrinus, London, in association with the Science Museum.

Peirce, W. S. (1996). "Economics of the Energy Industries." 2nd ed. Praeger, Westport, CT, London.

Pietrobono, J. T. (ed.). (1985). "Coal Mining: A PETEX Primer." Petroleum Extension Service, The University of Texas at Austin, Austin, TX, in cooperation with National Coal Association, Washington, DC.

Shepherd, R. (1993). "Ancient Mining." Elsevier Applied Science, London and New York.

Stefanko, R. (1983). "Coal Mining Technology." Society of Mining Engineers of AIME, New York.

Thesing, W. B. (ed.). (2000). "Caverns of Night: Coal Mines in Art, Literature, and Film." University of South Carolina Press, Columbia, SC.

Trinder, B. (ed.). (1992). "The Blackwell Encyclopedia of Industrial Archaeology." Blackwell, Oxford, UK.

Coal Mine Reclamation and Remediation

ROBERT L. P. KLEINMANN
National Energy Technology Laboratory
Pittsburgh, Pennsylvania, United States

1. **Introduction**
2. **Potential Environmental Problems Associated with Coal Mining**
3. **Land Reclamation**
4. **Mine Water Remediation**
5. **Other Environmental Aspects**

Glossary

acid mine drainage (**AMD**) Acidic water that drains or discharges from a mine; the acidity is caused by the oxidation of iron sulfide (FeS_2), or pyrite, which also causes the water to have elevated concentrations of sulfate and dissolved iron; also known as acid rock drainage (ARD), because the condition can result from non-mining-related pyrite exposure.

approximate original contour (**AOC**) A requirement that the final topographic configuration of backfilled and reclaimed surface-mined land approximate premining slopes and the general aspect of the premine topography.

bonding Mining companies are often required to post a bond to ensure that mined land will be properly reclaimed. The bond is typically released in phases, as remediation proceeds, either by recognizing successfully completed reclamation for part of the permitted area or by reclamation stages. In the latter case, most of the bond money is released when backfilling and regrading is completed, but some is kept for up to 5 years to ensure that revegetation remains adequate and self-sustaining and that water quality requirements are met. In the past, bond monies, when forfeited by companies that went bankrupt, were never adequate to reclaim a mine site and to treat the contaminated water, so government agencies are now requiring higher bonding levels.

coal refuse The waste left when raw coal is prepared or "cleaned" for market; usually contains a mixture of rock types, but is typically elevated in pyrite, causing it to be a source of acid mine drainage if not handled carefully. In the past, this material was simply piled up; nowadays, disposal requires compaction and stabilization measures.

fugitive dust The silt, dust, and sand that become airborne, carried away from a mine property by the wind.

mountaintop removal An often profitable but controversial method of surface mining in mountainous terrain where the coal seams lie beneath the ridge tops. All of the overburden is removed, exposing 100% of the coal for recovery. The controversy is due to the fact that the initial spoil must be placed in valley or head-of-hollow fills, though most of the subsequent spoil is placed on the level land left after mining proceeds. The final surface configuration, compared to the original landscape, has less relief.

Office of Surface Mining (**OSM**) A United States federal agency created by the Surface Mine Control and Reclamation Act of 1977, with responsibilities that include overseeing state enforcement of reclamation regulations, collecting fees per ton of coal mined, disbursing those funds (within the limits provided annually by Congress) to the states for reclamation of abandoned mined lands, and providing technical assistance and guidance to the states.

remining Reclamation of previously mined land by private companies, which extract the coal that was left behind by earlier mine operators; water treatment liabilities are not acquired by the new operator unless the new activity makes the water quality worse.

subsidence When fracturing and collapse of strata extend to the land surface; often associated with underground excavations.

Modern coal mines often cause environmental problems, but it must be recognized that these problems are much less extensive and less severe than those associated with mines that were abandoned before regulatory controls went into effect. The progress that has occurred over time is due in

large part to the efforts of the environmental activists and the environmental regulations that they inspired, but it is also due, in part, to advances in technology that have made cost-effective environmental progress possible. Such technology can be used to avoid and prevent, or to reclaim and remediate, environmental problems.

1. INTRODUCTION

Any intensive use of the earth's resources carries with it potential environmental consequences. Mining is no exception. Historically, mining companies have extracted the earth's resources wherever economics made it feasible, secure in their knowledge that their products were essential to society. Prior to the passage of the Surface Mining Control and Reclamation Act in 1977, 1.1 million acres (445,000 ha) of coal-mined land in the United States was left unreclaimed and over 10,500 miles (16,900 km) of streams and rivers were adversely affected in the Appalachian region alone. At about the same time, 55,000 ha of mined land was left derelict in Great Britain, though not all of that was due to coal extraction. Although numbers are not readily available for other countries, it can safely be said that leaving land unreclaimed was common around the world during the time period prior to environmental legislation in the United States (Fig. 1).

However, over the past 25 years, a new generation of miners has taken over the industry. Most of these miners grew up with an environmental awareness, and have bought into the philosophy of environmentally responsible mining. Coal mining operations today are designed to comply with environmental regulations and to minimize adverse environmental impacts; exceptions exist, however, and serve to inspire environmental activists and conservation groups to resist and reject mining wherever it is proposed. This societal tension serves as a backdrop to the dynamics of issuing or rejecting mining permit applications.

FIGURE 1 Unreclaimed mine site; vegetation is largely volunteer growth.

2. POTENTIAL ENVIRONMENTAL PROBLEMS ASSOCIATED WITH COAL MINING

2.1 Surface Mines

The most obvious environmental effect of surface mining is the disruption of the land surface; surface mining is generally more visually obtrusive compared to underground mining. In most cases, coal occurs at depth and is covered by soil and rock that must be removed to allow access to the coal, leaving large holes and piles of removed material. Even when the coal outcrops on the land surface, removing the coal leaves a hole in the ground. At prelegislation mines, the land was sometimes left in a disrupted state, but mining companies are now required to separate topsoil from the overburden material, to fill in the excavation as mining proceeds, to cover the regraded land with topsoil, and then to revegetate the land surface. Moreover, in most countries, the regraded and reclaimed land has to be left in a final condition, the approximate original contour (AOC), that approximates the topography that existed before mining. Also, in the United States, bond money must be posted before mining is permitted; before the bonds can be released, the land must pass certain sustainable revegetation requirements, based on the premining soil conditions. For example, if the land was originally prime farmland, the reclaimed land must be productive enough to qualify as prime farmland (Fig. 2). Finally, the reclaimed land surface must be stable; measures must be taken to avoid erosion.

Given the fact that coal has been extracted, it might seem problematic to refill the excavated pit completely. In practice, the problem is typically the reverse; once the consolidated overburden rock is disrupted and broken, it cannot be placed in the same volume of space that it occupied before mining. Excess material (often called "spoil" or "fill") must be placed so as to be stable, either on level ground created during the mining process or in nearby valleys. At some sites, a pit is intentionally left and allowed to flood, providing a pond or small lake.

FIGURE 2 Properly regraded and reclaimed mined land.

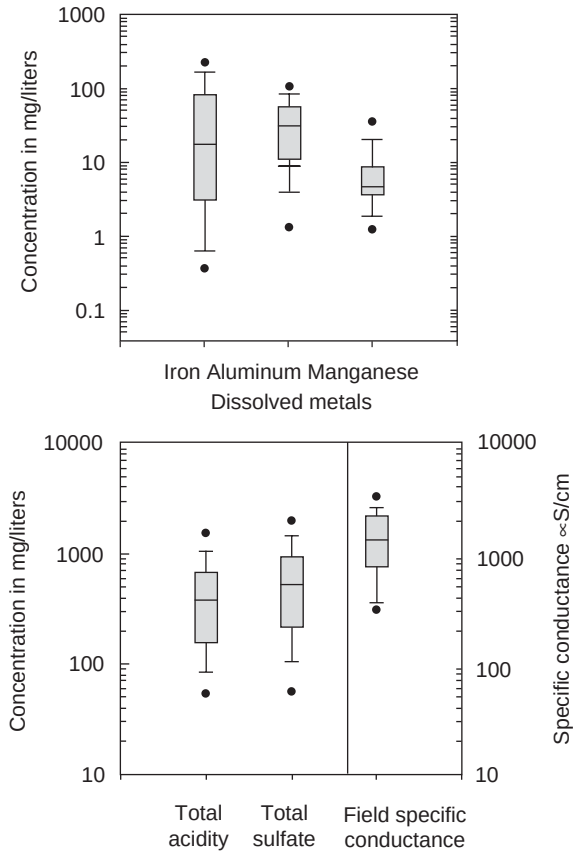


FIGURE 3 Range of contaminant concentrations in typical acidic coal mine drainage.

This avoids the expense of hauling back spoil material already placed elsewhere, but the issue of water quality in such ponds, and indeed at all mine sites, is a key concern.

At many coal mines, iron sulfide (pyrite) associated with the coal and the overburden strata oxidizes on exposure; this produces water with elevated concentrations of sulfate, acidity, and dissolved metals. If there is sufficient alkalinity present (typically in the form of limestone), the alkalinity will neutralize the acidity. In such cases, the environmental impact on water quality may be relatively minor. However, if there is insufficient alkalinity, the water becomes acidic. This type of water is known both as acid mine drainage (AMD), in most coal-mining regions and especially in the eastern United States, or as acid rock drainage (ARD), in most metal ore-mining regions. The term ARD is also preferred by those who like to point out that the same water quality results from such rock exposure in road cuts and construction projects (e.g., the Halifax Airport). ARD can typically be recognized, wherever it occurs, by the color of the water, i.e., red or yellow-orange, which is caused by dissolved and suspended iron (Fig. 3).

The prediction of postmining water quality is a key component in obtaining a mining permit. This is typically determined by analyzing rock cores for pyrite and alkalinity-producing rock strata. Then, using one of several procedures, a prediction is made about whether the water will be acidic. If it appears that it will probably generate AMD, or have other adverse hydrologic consequences, the regulatory agency may deny permission to mine or require the mining company to undertake special measures to decrease the likelihood of AMD. For example, the mining company may be required to handle pyritic strata selectively, and place it in such a manner that it will be less exposed to air and/or water, or to mix it with the more alkaline rock strata, or to import additional alkalinity to the site.

AMD can contaminate many miles of stream down-gradient from a mine, sometimes rendering it toxic to aquatic life. By regulation, any site that generates water not meeting regulatory standards (typically pH 6–9, no net acidity, iron less than 3 mg/liter, and manganese less than 2 mg/liter) must treat the water before it can be discharged. The water treatment must continue for as long as the untreated water fails to meet these discharge criteria. However, it should be noted that at surface mines, after the pyritic rock is buried, AMD typically moderates. Acid salts formed during exposure to the atmosphere continue to dissolve, but once these are dissipated, acid generation begins to decrease. It may take decades, but water quality, at even the worst of these sites, does improve.

In the United States, AMD is a major problem in Pennsylvania, northern West Virginia, western Maryland, and eastern Ohio, but also occurs at many mine

sites in other states. Elsewhere in the world, acid drainage is often associated with coal mining, though the extent of the problem varies with local geology, site conditions, mining methods, etc. In the western United States, AMD is less likely to be a problem, due in part to the lower concentrations of pyrite associated with the strata there and in part due to the fact that the climate is drier. In fact, the lack of adequate precipitation can make reclamation difficult, and also introduces the problem of sodic spoils. Such materials are exposed during mining and contain elevated concentrations of sodium salts, which dissolve and add to the salinity of the soil and the downstream waterways.

Other environmental problems associated with surface mines include fugitive dust (airborne particles that blow away in the wind), ground vibrations and noise associated with blasting, loss or conversion of wildlife habitat, and loss of aesthetics (visual resources). The first two are temporary problems, but the latter two can be temporary or permanent, depending on the eventual land use. However, if an area is viewed as important to the life cycle of an endangered species, permit denial is almost certain. In contrast, mining companies sometimes take advantage of land disturbance and reclamation to enhance wildlife activity; for example, in the United States, an exception was granted to the AOC requirement to allow a mining company to establish an appropriate habitat for a bird that was native to the area but declining in population.

The lack of aesthetics generally attributed to mined and reclaimed land, though it varies with the individual site, generally refers to the fact that mined land is typically reclaimed in a bland and uniform manner. The lack of visual contrast, rather than the actual disturbance, is what is noticed. This is one, of many, objections that citizens often make about mountaintop removal, which is currently the most controversial form of permitted surface mining, and deserves to be specifically mentioned here. Mountaintop removal, or mountaintop mining, is used in mountainous areas such as southern West Virginia to extract multiple seams of coal over large areas. The excess overburden is placed in what were previously valleys, with French drains constructed to handle the intermittent stream flow that might have been there previously, though major drainage patterns are preserved. The original rugged ridge and valley appearance of the land is converted to a more sedate topography, creating usable flat land where before there was none, and typically reducing the hazard of storm-related flooding; local inhabitants, although they may appreciate the jobs brought to the area, find their neighborhood forever changed (some would say ruined). Local biota is of course affected as well. There have been numerous court fights attempting to end the practice of mountaintop removal, and it is not yet clear how the issue will finally be resolved, but an interesting and so far unexplained finding that elevated levels of selenium have been found downstream of West Virginia mountaintop removal operations may turn out to be significant.

2.2 Underground Mines

Although underground mine operations are not as visible as surface mining, their overall environmental impact can be greater than that of the typical surface mine. A key environmental problem is subsidence. Underground mines are large cavities in the rock, and depending on the strength of the intervening strata, the depth of the mine, and the type of mining and roof support, the rock walls can fail, causing cracks and land collapse at the surface. Typically, coal seams at depths greater than about 200 feet are extracted by underground mining methods rather than by surface mining, with the exact depth principally based on the relative amount of coal and overburden. However, before improved technology made surface mining so affordable, the trade-off occurred at much shallower depths; some abandoned underground mines are only 35 feet below the land surface.

In longwall mines, the subsidence is induced as mining proceeds. Most of the subsidence occurs within a few months after mining. Changes in surface elevation can be significant, affecting highways, waterways, etc., though mining beneath such features may be restricted. The extensive fractures and settling also disrupt aquifers, though deeper aquifers typically recover once the void spaces fill with water. In many European countries, where mining is centrally planned and directed, subsidence above longwall mines is typically anticipated and planned for (for example, in how buildings are constructed) years ahead of time. Elsewhere, where mining proceeds based on the decisions of mining companies and local and regional regulatory agencies, subsidence may be anticipated, but is rarely coordinated with other regional development activities. In contrast, more traditional room-and-pillar mining may resist subsidence for decades, failing only as pillars, left to support the overburden rock, erode and then collapse. The lateral extent of the subsidence area may not be as great as above longwall mines, but because the collapse is more localized, there is a

greater risk of differential subsidence (for example, one corner of a house subsides more than the rest of the house, severely damaging or destroying it (Fig. 4). The ability to predict the extent of subsidence has improved over the past decade, but this is still a very inexact science.

The effect of subsidence on streams and waterways can be subtle or profound. In the worst cases, subsidence fractures reach the streambed and partially or totally drain the stream, diverting the water underground. Such fractures can be detected using terrain conductivity, and inexpensively sealed using an appropriate grout, but until that is done, the stream may go dry, and mining may be impeded or even temporarily halted. Even where waterways are not affected by subsidence fractures, the changes in slope will cause profiles to change, causing erosion in some areas and sediment deposition in others, and locally affecting stream biota.

Depending on the coal seam and location, AMD generation in underground mines can exceed that at surface mines. Because pyrite forms in a swampy environment, coal seams often contain more pyrite than do the overlying strata. At a surface mine, virtually all of the coal is removed, but in an underground mine, significant amounts of coal remain behind to provide roof support; if the coal is pyritic, it is problematic. In addition, alkalinity that may be present in the overburden strata is not as exposed to dissolution as it is when it is disrupted by surface mining. Finally, an underground mine is essentially a void, and behaves like a well; water flows through the surrounding rock into the mine void. When mining ceases and the water is no longer being pumped, the mine begins to flood and the water table rises. Thus, the volume of AMD that eventually discharges to the surface can be very great, completely overwhelming the buffering capacity of the receiving waterways. If the coal seam is completely inundated, the large mine pool will gradually improve in quality once all of the acid salts are washed away, because inundated pyrite does not continue to oxidize to a significant extent. However, coal seams typically slope (or "dip"); this may mean that only part of the seam is underwater, and the rest is continually exposed to the atmosphere, creating an ideal acid-generating environment that can continue to produce AMD for centuries.

FIGURE 4 An abandoned coal refuse pile that leaches acidic drainage; many of these waste piles can be converted into a resource by burning the material in a fluidized bed combustion unit.

2.3 Refuse Piles

Mined coal often contains other rock types (e.g., shale and clay associated with the coal or lying immediately above or below the coal seam) and impurities (such as pyrite). Such coal is considered "high ash," because the impurities remain behind in the ash after combustion. To improve the coal value, such coal is typically crushed and "cleaned" or "washed" in a preparation plant. This process separates the coal from the waste material, which is transported to a disposal area. This coal refuse is typically very high in pyrite, low in alkalinity, and quite reactive because it has been crushed to a relatively fine grain size. Today, there are strict procedures on how the material must be compacted and covered; even so, AMD is a common problem. Abandoned piles of prelegislation coal refuse, which frequently lie in or near waterways, are problematic (Fig. 5). Such piles can generate extremely acidic mine drainage and sometimes catch on fire. However, many of these old waste piles contain quite a bit of coal, given the fact that coal cleaning procedures were not always as efficient as they are nowadays. Such piles can be converted into a resource, because the material can be burned in a fluidized bed combustion (FBC) unit.

2.4 The Environmental Legacy Associated with Abandoned and Orphaned Mines

Currently, site reclamation is planned for during the permitting process and is incorporated into the mining operation. However, this was not always the case. Many mine sites were legally abandoned in an unreclaimed or poorly reclaimed condition because mining was completed before environmental

FIGURE 5 A house damaged by subsidence caused by the collapse of pillars in an abandoned underground mine.

FIGURE 6 Excavation and extinguishment of a burning coal refuse pile.

regulations went into effect. These abandoned mines are scars on the landscape and cause most of the water pollution attributed to mining. These old mines are considered abandoned because, in most countries, no one is legally required to reclaim the land or to treat the water. A similar problem occurs at mine sites that are or were operated by companies that have gone bankrupt. In theory, "bonds" (money held in escrow or guaranteed by a third party, required in many countries) posted by the companies will pay for reclamation and water treatment; however, at most sites, the amount of money required to reclaim the site exceeds the required bonds. In the United States, forfeiting a bond means that the company can no longer mine, so it is serious, but sometimes the costs of environmental compliance exceed the resources that a company has available. Several orphan underground mines are located near the Pennsylvania–West Virginia borders. These mines are gradually flooding with AMD, and are projected to begin discharging contaminated water to tributaries of the Monongahela River within the next 2 years. The anticipated contaminant load (contaminant concentration x flow) is such that it would have a major impact on water quality for many miles of the Monongahela River downstream; state and federal agencies are scrambling, trying to figure out what to do.

Regardless of whether a mine is abandoned or orphaned, the land remains inadequately reclaimed and the water that discharges from the site is typically not treated. In addition to the potential environmental problems associated with active mines, and already discussed, abandoned and orphan mines often have additional problems that can affect the health and safety of people who live in the area. For example, open pits with steep cliffs inevitably attract children, as do unsealed, unsafe underground mines. Unreclaimed waste rock piles and mine dumps pose a risk of slope failure and landslides. Another potential problem (discussed in detail in the next section) is mine fires, which are more typically viewed as a health and safety problem. However, at abandoned mines, such fires can smolder for decades, generating toxic gases that can flow through cracks into basements, and can accelerate subsidence events (Fig. 6).

So how are these problems dealt with? In the United States, the Surface Mining and Reclamation Act of 1977 (as extended and amended by the Abandoned Mine Reclamation Act of 1990 and the Energy Policy Act of 1992) did more than require mine operators to maintain environmental standards during mining and reclamation. It also imposed a fee of 35 cents/ton of coal mined by surface methods, 15 cents/ton of coal mined underground, and 10 cents/ton of lignite. These fees go into a fund managed by the Office of Surface Mining (OSM) in the U.S. Department of the Interior. This money can be used by the states and the OSM to address problems at abandoned mine sites. Most of these funds have gone to remediate sites that are hazardous to people and property, but increasingly funds are also being used to remediate environmental problems. These funds cannot be used to resolve problems at orphan mines.

In addition, regional and national governments have often provided additional funds for environmental remediation of abandoned sites. In particular, Pennsylvania should be singled out for its unique approach; it encourages the formation of watershed associations, which can apply for funds to remediate

abandoned mines. Through the efforts of these enthusiastic volunteers, Pennsylvania accomplishes much more remediation than would be otherwise possible.

3. LAND RECLAMATION

Under natural conditions, a landscape represents a balance of geomorphic processes; this dynamic stability is disrupted by mining. In this context, the goal of reclamation is the reestablishment of the steady state. In the United States, regulations tightly control the reclamation process at active operations, dictating the slope (AOC), the extent of revegetation, and the rate at which reclamation and revegetation must proceed. Failure to comply results in loss of bond money and/or fines. Other countries (e.g., Canada) have more flexibility built into their regulatory structure. They consider the imposition of AOC inappropriate at many sites—for example, in areas where flat land would be desirable; enhanced flexibility allows such aspects to be negotiated. In the United States, exceptions can be granted by the OSM, but are atypical. Instead, operators have learned to operate within the limits of the regulation. For example, they may choose to reduce erosion potential on steep slopes by installing terraces and manipulating infiltration rates.

When reclamation agencies are attempting to reclaim derelict or abandoned operations, no attempt is typically made to restore the land to AOC. Instead, the primary intent is to remove potential hazards (e.g., extinguish potentially dangerous mine fires, seal mine openings); secondarily, the intent is to bury acid-forming material, and, finally, to create a stable surface that will resist erosion with minimal maintenance. Funds for reclamation of abandoned sites are quite limited and represent a small fraction of what would be necessary to reclaim all abandoned mine sites.

Many abandoned sites have developed a natural, if sometimes sparse, vegetative cover of volunteer species. In addition, an unusual incentive has recently developed that may encourage companies to reclaim abandoned sites that have not naturally revegetated. As pressure gradually grows on companies to lower carbon emissions, proposals have been made that credit should be given for revegetating barren lands, because this would sequester carbon from the atmosphere. Arid deserts and barren abandoned mine sites are the two most likely types of places for such an effort, if and when it becomes codified.

Another aspect of mined land remediation is subsidence control. Because subsidence events above old room-and-pillar operations commonly cluster, when subsidence begins in an area, state and OSM emergency personnel are quickly mobilized. Typically, they attempt to backfill, inject, or stow, hydraulically or pneumatically, as much solid material into the mine void as possible, figuring that reducing the void volume with fine solids (sand, fly ash, etc.) will both decrease the amount of subsidence that will occur and reinforce pillars that might otherwise fail. However, because the mine voids that they are injecting the material into are often flooded, it is sometimes difficult to ascertain if these efforts make much of a difference. An alternative option pursued by many landowners in undermined areas is government-subsidized subsidence insurance. This insurance fund pays property damage after subsidence has occurred.

4. MINE WATER REMEDIATION

Before the passage of regulations dictating mined land reclamation and mine water discharge standards, streams and rivers down-gradient of mine sites were often contaminated with high levels of suspended and dissolved solids. In the eastern United States, AMD was also a major problem. Nowadays, streams and rivers near active mine sites have much less of an impact. Sediment ponds are constructed to collect suspended solids and if the mine water does not meet regulations, chemicals [typically lime, $Ca(OH)_3$] are added to neutralize acidity and precipitate dissolved metals. Consequently, the remaining sources of contaminated water discharging to streams and rivers in mined areas are generally from abandoned or orphan mines. However, before moving on to water remediation at such sites, the techniques used to prevent, ameliorate, or remediate mine water problems at active operations need to be addressed. As alluded to previously, a powerful tool in preventing AMD is the permitting process. Variation exists in how much certainty of AMD generation a regulatory agency requires before a permit is denied. In the United States, Pennsylvania is the most conservative, and can legitimately boast that it has improved the accuracy of its permitting decisions from about 50% in the 1980s to 98%; however, it should be noted that this definition of success is based on AMD generation at permitted operations (following reclamation) and does not include mine sites that had permits denied and might not have produced AMD if allowed to operate.

In contrast, adjacent states (and many areas outside of the United States) permit a higher percentage of mines to operate, but generally require that measures be taken to reduce the risk of acid generation. In addition to the selective handling of overburden and the importation of alkalinity from off-site mentioned earlier, minimizing exposure of the pyrite to either oxygen or water can decrease the amount of acidity generated. Restricting exposure to oxygen generally means that the pyritic material must be placed beneath the eventual water table, which is difficult to do at many surface mines. Alternatively, the operator can attempt to place the pyritic material in an environment where it is dry most of the time, well above the potential water table and capped by material of low permeability. Compaction can also be used to reduce permeability, though not on the final soil surface, where compaction makes it difficult to establish a vegetative cover. Water can be diverted around the mine site by intercepting overland flow with ditches, and by constructing drainage ways along the final highwall to intercept groundwater and prevent it from flowing through the overburden material.

At many sites, mining companies have been given special permission to remine old abandoned mines to improve water quality. The mining companies harvest coal left behind by the old operations. For example, old room-and-pillar mines sometimes left as much as 50% of the coal behind to support the roof rock. Some of these old mines are relatively shallow and can be inexpensively surface mined using modern machinery. Because the coal pillars are sources of acid generation, removing them generally improves water quality. Similarly, old surface mines could not economically remove as much overburden as is now possible. The exposed highwall can now be economically mined with modern machinery. The mining companies are required to reclaim the land to current standards, but their water discharge requirements typically only require them to at least meet the water quality that existed before the remining operation. In most cases, the water quality improves and the land is reclaimed, at no cost to the public.

A more exotic approach of controlling AMD involves the inhibition of iron-oxidizing bacteria. These ubiquitous bacteria normally catalyze pyrite oxidation; inhibiting them reduces acid generation significantly. In practice, however, it is difficult to do this. The only cost-effective approach that has been developed involves the use of anionic surfactants (the cleansing agents in most laundry detergents, shampoos, and toothpaste); their use selectively inhibits the iron-oxidizing bacteria by allowing the acidity that they generate to penetrate through their cell walls. This approach has been used effectively to treat pyritic coal refuse, reducing acid generation 50–95%. Slow-release formulations have been developed for application to the top of the pile before topsoil is replaced and the site is revegetated.

FIGURE 7 Mine water can be directed through a specially constructed wetland to passively improve the water quality.

Discharge criteria must be met, and if the at-source control measures are not completely effective, some form of water treatment is required. Sometimes water treatment is required only during the mining and reclamation operation, and water quality improves soon after reclamation is completed. Sometimes the water quality remains poor. Chemical treatment is simple and straightforward, but is expensive to maintain for long periods.

Passive and semipassive water treatment technologies, though by no means universally applicable, are an option at many sites. These techniques developed as a result of observations at natural and volunteer wetlands, which were observed to improve the quality of mine drainage. The simplest of these techniques, appropriate for near-neutral mine water that is contaminated with iron, involves the construction of shallow ponds, planted with plants that will tolerate the water, such as *Typha* (commonly called cattails in North America) (Fig. 7). The iron in the water oxidizes and precipitates in this constructed wetland instead of in the stream, allowing natural stream biota to repopulate. These constructed wetlands cannot be viewed as equivalent to natural wetlands; they have been designed to optimize water treatment, and any associated

ecological benefits, though they may be significant, have to be viewed as secondary.

Water that is acidic can be neutralized by the inexpensive addition of alkalinity. Several passive methods have been developed, generally using limestone and/or sulfate-reducing bacteria. Alkaline waste products (e.g., steel slag) have also been used. Water quality, site considerations, and flow dictate which approach is most cost-effective at a given location.

The realization that passive techniques can be used to treat low to moderate flows of mine water has made it possible for state reclamation agencies and local watershed associations to remediate mine water at abandoned and orphan mines without a long-term financial commitment for chemicals. However, highly contaminated AMD or high flows still cannot be cost-effectively treated passively. But the technology is continuing to evolve. Semipassive techniques, such as windmills and various water-driven devices, are being used to add air or chemical agents to mine water, extending and supplementing the capabilities of the passive treatment technologies.

Contaminated water in abandoned or orphan underground mines represents the ultimate challenge because of the large volume of water that must be treated. One option currently being explored in the United States is whether or such mine pools can be used as cooling water for a power plant. Locating new power plants is becoming more difficult because of the water needed for cooling purposes. A large mine pool could be a good resource for such an operation. The power plant would have to treat the mine water chemically, but this cost could be partially subsidized by the government, which would otherwise have to treat the water chemically or allow the mine discharge to flow untreated into the local streams and rivers.

5. OTHER ENVIRONMENTAL ASPECTS

The issue of aesthetics has already been introduced. In many ways, it can be the most challenging obstacle to popular acceptance of mining in an area, because, to many, anything less than complete restoration of the land is unacceptable. However, mine operators have learned to minimize the aesthetic impact of mining by avoiding extremely visible sites, by using screens (e.g., trees left in place as a visual barrier or revegetated topsoil stockpiles placed in locations where they hide the mine pit from view), by minimizing the duration of impact, and by ensuring that regrading and revegetation proceeds as quickly as possible.

Finally, the issue of ultimate land use of the reclaimed land must be considered. Nowadays, this is actually addressed as part of the permitting process. Some mine operators have successfully created planned industrial sites, some have created highly productive farmland, but much of the mined land is reclaimed for wildlife use. Costs for seed and plantings are relatively low, and plans are often coordinated with local conservation groups to reclaim the land in an appropriate manner for wildlife use. To be effective in this regard, these reclaimed lands must be physically connected to other areas where such wildlife currently exists; isolated fragments of habitat will not serve the desired purpose of restoring wildlife use disrupted by mining activity and road construction. With regulatory approval, ponds can be left in locations where they will serve migratory bird populations or provide watering areas for permanent wildlife populations.

SEE ALSO THE FOLLOWING ARTICLES

Coal, Chemical and Physical Properties • Coal Industry, Energy Policy in • Coal Mining, Design and Methods of • Coal Mining in Appalachia, History of • Coal Preparation • Coal Storage and Transportation • Nuclear Power Plants, Decommissioning of • Uranium Mining: Environmental Impact

Further Reading

Berger, J. J. (1990). "Environmental Restoration." Island Press, Washington, D.C.

Brady, K. B. C., Smith, M. W., and Schueck, J. (eds.). (1998). "Coal Mine Drainage Prediction and Pollution Prevention in Pennsylvania." Pennsylvania Dept. of Environmental Protection, Harrisburg, Pennsylvania.

Brown, M., Barley, B., and Wood, H. (2002). "Minewater Treatment." IWA Publ., London, England.

Kleinmann, R. L. P. (ed.). (2000). "Prediction of Water Quality at Surface Coal Mines National Mine Land Reclamation Center." West Virginia University, Morgantown, WV.

Marcus, J. J. (ed.). (1997). "Mining Environmental Handbook." Imperial College Press, London, England.

PIRAMID Consortium. (2003). Engineering Guidelines for the Passive Remediation of Acidic and/or Metalliferous Mine Drainage and Similar Wastewaters, 151 pp., accessible at http://www.piramid.info.

Skousen, J., Rose, A., Geidel, G., Foreman, J., Evans, R., and Hellier, W. (1998). "Handbook of Technologies for Avoidance and Remediation of Acid Mine Drainage." National Mine Land

Reclamation Center, West Virginia University, Morgantown, West Virginia.
Watzlaf, G. W., Schroeder, K. T., Kleinmann, R. L. P., Kairies, C. L., and Nairn, R. W. (2003). "The Passive Treatment of Coal Mine Drainage." U.S. Department of Energy (CD available on request), Pittsburgh, Pennsylvania.
Younger, P. L., Banwart, S. S., and Hedin, R. S. (2002). "Mine Water." Kluwer Academic Publ., London, England.

Coal Mining, Design and Methods of

ANDREW P. SCHISSLER
The Pennsylvania State University
University Park, Pennsylvania United States

1. Introduction
2. Surface Mining Methods and Design
3. Underground Methods and Design

Glossary

area mining A surface system that removes coal in areas.
contour mining A surface system that removes coal following the contours of hillsides.
extraction ratio The ratio of coal mined divided by the coal originally in place before mining commenced, expressed as a percentage.
full extraction mining An underground system that removes the coal in large areas causing the void to be filled with caved rock.
interburden The rock that occurs in between coal seams.
overburden The rock that occurs over coal seams.
partial extraction mining An underground mining system that removes coal in small areas in which the roof rock remains stable.
raw in place tons The total weight of coal contained within coal seams of a reserve, including in-seam impurities.
raw recoverable tons The total weight of coal, including in-seam impurities that would be mined by a specific underground or surface mine plan applied to a reserve.
salable tons The total weight of coal mined and sold. The impurities are removed by surface coal processing methods applied at the mine.
strip mining A surface system of mining that removes coal in rectangular and parallel pits.

The design and methods of coal mining is made up of the use of land, labor, and capital to remove, mine, or extract coal from a reserve, utilizing a planned and systematic approach called a mine plan. A reserve of coal is represented by a seam of coal, which is a layered deposit of the mineral. A coal reserve is usually described by the quantity of weight, expressed in tons contained within a specified boundary in plan. Tonnage for a reserve can be quantified several ways that include raw in-place tons, raw recoverable tons, and salable tons.

1. INTRODUCTION

1.1 Coal Geology

Coal seams originate and are formed by the consolidation of biota, minerals, and natural chemicals through geologic time. Compression, heat, sedimentation, erosion, and chemical energy are agents of the coal formation process.

Multiple coal seams can exist in a reserve, with each coal seam separated by sedimentary rock interburden. Coal, and the rock above and below the coal seam, is primarily layered sedimentary rock in origin, as opposed to igneous or metamorphic rock. The noncoal layers are usually sandstone, limestone, shale, siltstone, and mudstone. The layers can be very symmetric and consistent in thickness, or they can appear in the reserve in nonregular patterns such as sandstone channels, which can exist above the coal seam or replace the coal seam by scouring during geologic formation. Sandstone channels can also create vertical downward displacement of the coal seam by differential compaction. Chemical and heat processes during mineral formation can also produce small pockets of nonsedimentary rock within the coal seam and surrounding rock such as iron pyrite, mica, and other minerals. These minerals, depending on their hardness, can interrupt the mining process as most coal mining methods are designed to extract soft mineral having uniaxial compressive strength less than 6000 psi to 8000 psi.

 © 2004 Elsevier Inc. All rights reserved.

The processes of the geologic environment in which the coal was formed determines the end characteristics of the coal seam, such as thickness, seam pitch, energy value (generally quantified as BTU per lb), nonvolatile content, moisture content, depth below the surface, structure (faults, folds, rock fractures), the presence of an igneous intrusives called dikes, and characteristics of the rock immediately above and below the coal seam. All of these characteristics affect the mine design and methods. Layered coal seams and surrounding rock thickness range from less than 1 ft to over 150 ft. Figure 1 shows a typical cross section of coal seams interspersed in layers of sandstone and shale.

1.2 Mining Methods

Coal is mined by three methods: surface, underground, or *in situ*. The percentage of coal produced by each mining method can be identified from production statistics kept by the Energy Information Administration (EIA), a U.S. agency. According to the EIA, in 2002 745 surface mines produced 736 million tons (67% of total U.S. production) and 654 underground mines produced 357 million tons (33% of total U.S. production). For practical purposes, coal produced from *in situ* mining methods is nil.

Coal mine design and methods are chosen based on the reserve characteristics. Where the coal seam is near the ground surface, surface mining methods prevail over underground methods. Generally, surface mining methods are less expensive than underground methods when the strip ratio is less than an economic limit. Strip ratio is the quantity of overburden (expressed in cubic yards) that has to be removed to uncover 1 ton of coal. Companies change mining method from surface to underground when the strip ratio from an existing surface mine increases due to seam pitch, which in turn increases production cost. At the economic limit, the cost of production from surface methods equals underground methods.

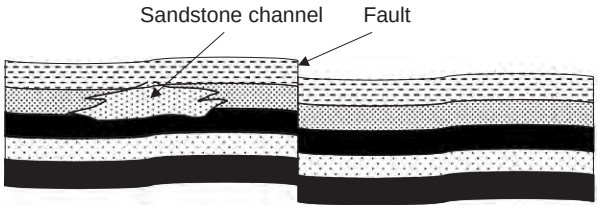


FIGURE 1 Typical cross section of shale, siltstone, coal, sandstone, and coal reserve.

2. SURFACE MINING METHODS AND DESIGN

The surface mining method of coal involves three ordered steps: overburden removal, coal seam removal, and reclamation. The first step in surface mining is to remove the rock over the coal seam called the overburden (Fig. 2).

Overburden is first drilled and fractured with explosives. The overburden is then removed and displaced to a previously mined pit, where the coal has already been removed, or to other disposal areas. Other disposal areas can include adjacent hollows, or topographic lows. After the overburden is removed, the coal seam that is now exposed is removed by digging methods and hauled or conveyed to surface facilities that process the coal for shipment. Downstream transportation such as railroads or overland trucks hauls the coal to markets. The third step in surface mining is reclamation. Reclamation in surface mining consists of placing the overburden rock into the hole, pit, or hollow, and recovering the rock with topsoil that was removed and stockpiled before the overburden removal began. Vegetation is replanted in the topsoil to complete the reclamation process.

Overburden removal, coal seam removal, and reclamation occur in order, and simultaneously within the life of a mine or reserve. In other words, a mine does not remove all of the overburden in a deposit before removing the coal. Surface mine design involves the planning of a series of pits *within* the overall deposit boundary that allow overburden removal, followed by coal seam removal, which is then followed by reclamation. The planning process of these sequential operations is called operations planning or sequencing.

The primary surface mine design parameter and driver of unit production cost is the strip ratio. Strip ratios for minable coal reserves range from 1:1 to over 35:1. A decrease in the strip ratio will decrease

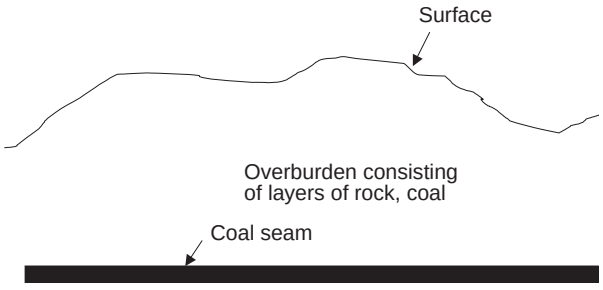


FIGURE 2 Step 1 in the surface mining process, overburden removal (in cross section).

the cost of producing the coal. Underground mining methods are employed when the amount of overburden that would have to be removed by surface mining methods is cost prohibitive (i.e., the strip ratio is beyond an economic limit).

Surface mine design is determined by the stripping ratio, the overall shape of the reserve from a plan view, and the type of rock. Three main surface methods are utilized, with each having the same common steps of overburden removal, coal seam removal, and reclamation. The three surface mining methods are contour stripping, area stripping, and strip mining.

2.1 Contour Stripping

Coal seams and associated overburden can outcrop on the edges of hills and mountains. Figure 3 shows a cross-section of a typical contour mine plan. Contour surface mining takes advantage of the low strip ratio coal reserves adjacent to the outcrop. Contour mining equipment includes a drill, loader, bulldozers, and trucks. The steps in contour mining begin with drilling and blasting a triangle shaped portion of overburden (in cross section) above the coal to be mined. The fractured overburden is then loaded into rubber-tired trucks. The loading equipment utilizes buckets such as front-end loaders, front-loading shovels, and backhoes. The coal seam is now uncovered and is loaded with the same type of bucket-equipped machinery as loaded the overburden. Some coals require drilling and blasting as the breakout force required to remove the coal without fragmentation by explosives is beyond the capabilities of the bucket-equipped machinery. Dumping the overburden into the pit begins reclamation. The overburden is usually hauled around the active pit where coal is being removed. The overburden is dumped and regraded to replicate as much as possible the original contour or slope of the hillside. Bulldozers grade and smooth the overburden in preparation for topsoil application. Topsoil is placed on the smoothed overburden and replanted with vegetation. Contour stripping gets its name because the mining operation follows the contour of where the coal outcrop meanders along the hillside (Fig. 4).

The advantages of contour mining include the ability to control strip ratio during the mining process. The strip ratio is controlled by the height of the triangle portion of overburden removed. Controlling the ratio allows the unit cost of production to be largely controlled. The disadvantages of contour mining include the difficulty of maintaining the stability of steep, reclaimed hillsides and the limited width for effective operations on the coal seam bench.

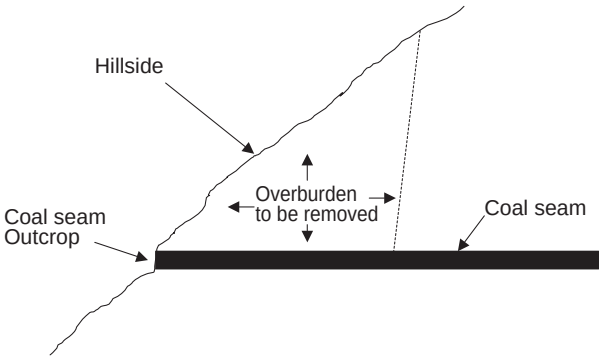


FIGURE 3 Contour mine design, surface mining method (in cross section).

2.2 Area Stripping

Area stripping is a surface mining method that is usually employed in combination with strip mining. Area stripping utilizes truck and shovels to remove the overburden down to a predetermined depth in the layers of rock and coal. The strip mining method continues to mine the overburden below the depth at which the area was area-stripped. Draglines or continuous excavators, which are used to strip the overburden from the lower seams, are large pieces of equipment and require large, level areas from which to work. Draglines and continuous excavators have dig depth limitations. Area stripping allows the draglines and continuous excavators to operate within their depth limitations. When area stripping is used in combination with draglines and continuous excavators, area stripping is called prestripping. The purpose of prestripping is to provide a flat surface for the dragline or excavator on which to set, move, and operate. The depth of overburden rock removed by area stripping is governed by dig depth capacity of the equipment utilized for subsequent strip mining.

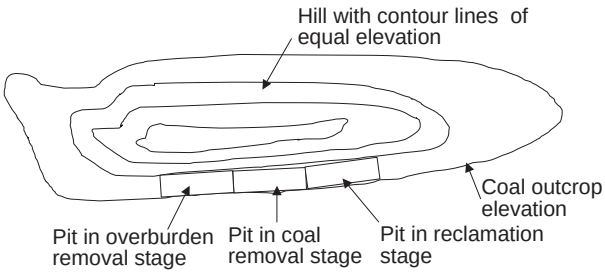


FIGURE 4 Contour mine design (in plan view).

Usually, the prestripping mine plan is arranged in rectangular blocks of overburden to be removed that coincide with pits to be mined by stripping methods underneath the prestripping area (Fig. 5).

Area stripping application includes the removal of overburden in reserves that are not conducive to strip mining because the lack of rectangular areas (in plan). The machinery used for area stripping such as shovels, front-end loaders, and trucks allow for flexibility in the mine design layout.

2.3 Strip Mining

Strip mining is employed in coal reserves where the overburden is removed in rectangular blocks in plan view called pits or strips. The pits are parallel and adjacent to each other. Strip mining is fundamentally different from contour or area mining on how the overburden is displaced, called spoil handling. In contour or area stripping, the overburden is hauled with different equipment than what digs or removes the overburden. In strip mining, the overburden is mined and moved by the same equipment: draglines or continuous excavators. The movement of overburden in strip mining is called the casting process.

The operating sequence for each pit includes drilling and blasting, followed by overburden casting, then coal removal. Some overlap exists in operational steps between pits. Draglines and continuous excavators move or displace the overburden from the active pit to the previous pit that has had the coal removed.

The primary planning mechanism used in strip mining is the range diagram, which is a cross-sectional plan of the shape of the pit in various stages of mining. The range diagram allows the dragline or continuous excavator equipment characteristics of dig depth, reach, and physical size to be placed on the geologic dimensions of depth to seams (overburden), and depth between seams (interburden). By comparing machinery specifications with dimensional characteristics of the geology, the mine designer can plan the pit width and dig depth (Fig. 6).

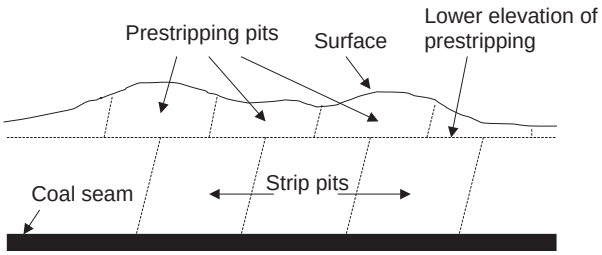


FIGURE 5 Prestrip surface mining method (in cross section).

As the dragline or continuous excavator moves the overburden to the adjacent empty pit where the coal has been removed, the rock swells in volume. Earth or rock increases in volume, called the swell factor, when the material is removed from its *in situ* or in-ground state and placed into a pit or on the surface. The range diagram allows the mine planner to identify the equipment dump height required to keep the displaced overburden (spoil) from crowding the machinery and mining operations. In certain cases of mining multiple coal seams from one pit, a coal seam can provide the boundary between the prestrip and strip elevations.

In a relatively new technique that originated in 1970s to early 1980s, explosives are used to move or throw the overburden into the previous pit in a process called cast blasting. The difference in the quantity of explosives required to fragment rock in place versus fragment and cast or throw the rock across the active pit and into the previous pit is cost-effective. Many surface strip mines use explosives to move overburden in addition to the primary swing equipment (dragline or continuous excavator), displacing up to 35% of the overburden by cast blasting. When cast blasting is used, the dragline may excavate from the spoil side of the pit, sitting on the leveled, blasted overburden.

Surface mine design principles emanate from the operational characteristics of surface mining, which are drilling and blasting, spoil handling, coal removal, and haulage. Except in a few circumstances, overburden in surface mining requires the rock to be fractured by explosives to allow it to be excavated. The goal of drill and blast design is to optimize rock fracturing, which optimizes digging productivity. Fracturing is optimized by using the correct amount of explosive per cubic yard of overburden employed in the drill hole spacing in plan view. The amount of explosive in weight per cubic yard of overburden is called the powder factor. Drill and blast design is

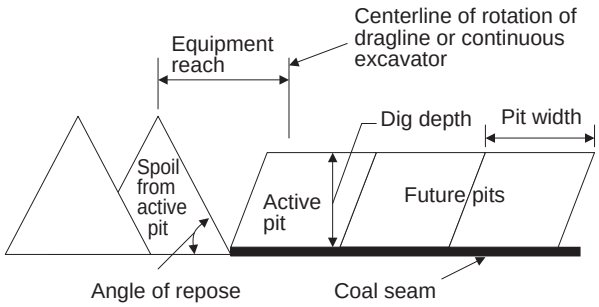


FIGURE 6 Range diagram for the strip mining method (in cross section).

accomplished by empirical methods and by experience. The drill hole layout and powder factor change when cast blasting is utilized.

Spoil handling design is of critical importance, as this function is usually the most expensive cost element in surface mining. When the surface mining method utilizes trucks, spoil handling is designed to minimize the overall haul distance for logical units of spoil volume, which may be driven by pit layout, topography, or area stripping requirements. Mine plan alternatives are evaluated to minimize the distance that spoil volumes are moved from the beginning centroid of mass to the ending centroid of mass. Spoil handling design goals for strip mining surface methods that utilize draglines and continuous excavators also include the minimization of spoil haulage distance. For the dragline, the average swing angle is identified by evaluating alternative mine plan layouts. The goal is to minimize the swing angle, which maximizes productivity.

The goals of coal removal and haulage design in surface mining include minimizing the distance coal is hauled from pits to surface processing and loadout facilities in near term years, locating haul road ramps out of the pits to minimize interference with overburden removal, and engaging excavation practices and equipment that minimize coal dilution by mining noncoal rock floor.

Surface mining has two design parameters that affect mine cost, which are minimizing rehandle and maximizing pit recovery. Rehandle occurs when overburden is handled twice and sometimes multiple times during excavation and spoil placement. Having 0% rehandle of the original inplace overburden is not achievable because of inherent design requirements of surface mining such as ramps into the pit and mining conditions such as sloughing ground that covers the coal. Simulating alternative mine plans and anticipating where overburden will be placed can minimize rehandle. Rehandle can more than double the cost of mining portions of the overburden.

The goal of coal pit recovery is to obtain as close to 100% as possible. One method to maximize pit recovery is to minimize drill and blast damage to the top of the coal. Drill and blast damage is reduced by stopping the drill holes from touching the coal seam or by placing nonexplosive material in each drill hole, called stemming. Pit recovery is also maximized by matching the pit width with the characteristics of the machinery used to extract the coal. Again, the range diagram as a planning tool is used in this evaluation.

3. UNDERGROUND METHODS AND DESIGN

Underground coal mining methods fall in two categories: full extraction and partial extraction. Full-extraction underground mining creates a sufficiently large void by coal extraction to cause the roof rock to cave into the void in a predictable and systematic manner. Partial-extraction systems disallow caving by mining reduced areas which sustains the roof rock above the tunnel in a stable or uncaved *in situ* state. When coal seams are completely below the ground surface, vertical shafts and tunnels driven on angles to the seam below called slopes are mined through the overburden rock to access the seam. When the edges of coal seams are exposed to the surface, called outcrops, access to the main reserve can be obtained by the creation of tunnels in the coal seam. Coal outcrops occur when sedimentary rock and coal is layered within a hill or mountain and the rock and coal outcrops above the surface ground level of drainages that define the topography. The main coal reserve is divided up into areas or panels (in plan view) that allow orderly extraction of the coal by the chosen equipment and panel mining method.

Tunnels excavated in coal are called entries. Entries are mined to gain access to the reserve and the production panels. Entries are excavated adjacent to one another in sets of entries. The sets of entries have additional tunnels, called crosscuts that are normally excavated at right angles to the entries. The entries and crosscuts form pillars of coal. The entire set of entries, crosscuts, and pillars, are referred to as mains, submains, or gateroads, depending on the degree of proximity to production panels. Entries excavated in production panels are sometimes called rooms.

3.1 Panel-Mining Methods

The methods to mine panels of coal in an underground mine layout are partial extraction using continuous miners to mine rooms, partial extraction using conventional mining methods to mine rooms, full extraction using continuous miners to mine rooms and pillars, full extraction using longwalls to extract a block of coal outlined by entries around the panel perimeter, and other underground methods.

3.2 Partial Extraction with Continuous Miners

Continuous miners are machines that mechanically mine or extract the coal without the use of

explosives. Continuous miners consist of a rotating drum having carbide-tipped steel bits that strike the coalface. As the bits strike the face, tension cracks are created on the coal surface. These cracks separate blocks of coal. The blocks of coal are propelled out of the face by the bits and drop into a portion of the continuous miner called the gathering pan. The action of the bits extracting the coal by the creation of tension cracks is called mechanical mining. Coal is conveyed through the continuous miner and is loaded onto vehicles that haul the coal to be further transloaded onto conveyors, which convey the coal out of the mine to surface processing facilities.

Within panels, continuous miners and associated equipment mine entries and crosscuts leaving the pillars of coal. Panels are arranged adjacent to one another, usually with the long axis of the panel dimension 90° to the main or submain entries (Fig. 7). Panel size is determined by the size and shape of the overall reserve in plan view.

3.3 Partial Extraction with Conventional Methods

In partial extraction with conventional methods, explosives fragment the coal at the face in contrast to continuous miners where bits or tools mechanically mine the coal. In conventional mining, holes are drilled in the coalface allowing explosives to be loaded into the holes. The explosives fracture and loosen the coal allowing loading and removal of the coal into the haulage system of the mine. Rubber-tired scoops load the coal. Thus, entries and crosscuts are created by the conventional method. Panel layout for conventional methods is the same as for continuous mining.

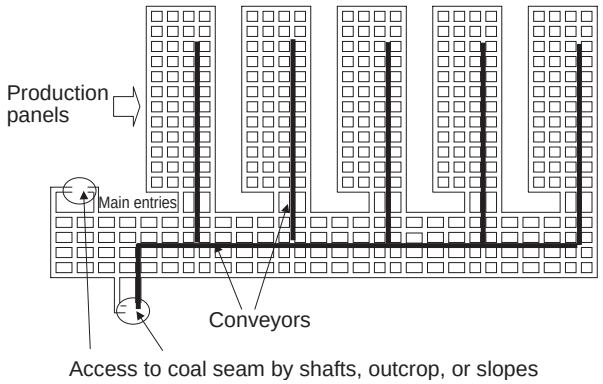


FIGURE 7 Partial extraction with continuous miners (plan view).

3.4 Full Extraction Using Continuous Miners

Full extraction using continuous miners is the same as partial extraction except that the pillars that are surrounded by entries and crosscuts in the panels are mined and removed. By removing the pillars, a sufficiently large area is created that causes the roof rock to become unstable and fall into the void. The caving process in full extraction techniques is an integral part of the mining method. Figure 8 shows the general method of individual pillar extraction.

The stumps shown in Fig. 8 are small coal pillars that are left after the extraction process. The stumps are necessary to provide roof support for the pillar during mining, but are not sufficiently large to inhibit the caving process.

3.5 Full Extraction Using Longwalls

The longwall method utilizes a machine group that consists of shields, face conveyor, stage loader, and shearer. This group of equipment is installed along one entry that bounds the short dimension of a rectangular block of coal. The block of coal is entirely surrounded by entries that outline the perimeter, and this block area is called a longwall panel. The longwall equipment extracts the coal from this rectangular panel. Each longwall panel is separated from an adjacent panel by a group of entries called the gateroad system. These entries, as

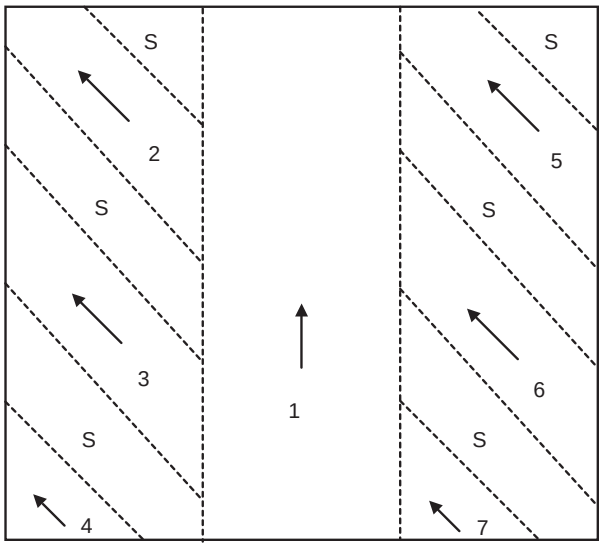


FIGURE 8 Typical pillar extraction sequence (plan view).

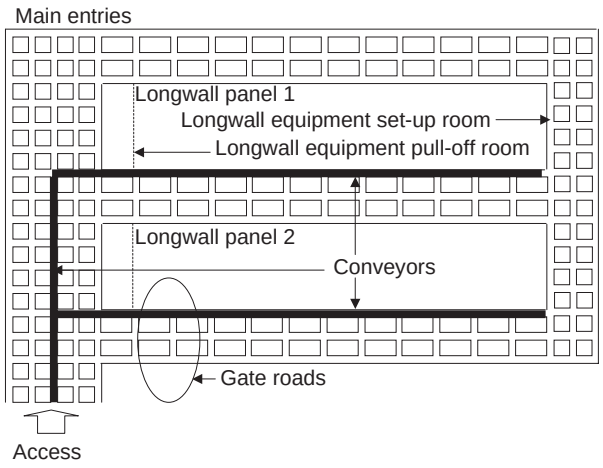


FIGURE 9 Longwall mine plan (plan view).

with the mains and submains, are mined with continuous miners. Once the longwall extracts the coal from the rectangular block, the equipment is disassembled at the pull-off room and moved and reassembled at the end of the next panel of coal in the startup room (Fig. 9).

Access to a coal seam for the longwall mining method is accomplished by outcrop access, shafts, or slopes. Gateroad systems can consist of one to five entries. Figure 9 portrays an example mine plan consisting of two longwall panels. The number of longwall panels contained within a mine plan is determined by the shape of the reserve and by safety considerations. The larger the reserve, the more longwall panels can be planned parallel to each other. Longwall panels are sometimes grouped together into subsets called districts. Longwall districts within an overall mine plan are separated by large blocks of coal called barrier pillars. Barrier pillars compartmentalize the reserve. A compartmentalized reserve allows worked out areas to be sealed, which increases safety and efficiency with respect to handling water, gas, and material.

Longwall panels range in size (in plan view), from 500 ft wide by 2000 ft long to over 1200 ft wide by 20,000 ft long. Size is determined by the reserve and the availability of capital to buy equipment. Some large underground mines have as much as 50 longwall panels separated into six or more districts.

3.6 Other Underground Methods

Other underground methods still incorporate entries, crosscuts, and pillars, but differ from previous methods by the method to extract the coal from the working face. They are hand loading and hydraulic mining. Hand loading is using human-powered pick and shovels to remove and load coal from a face into haulage equipment. This method is not used in the United States.

Hydraulic mining utilizes the force of water sprayed under high pressure to extract coal. Hydraulic mining utilizes continuous miners to mine entries to gain access to production panels. Once a production panel is established, high-pressure water cannons are used to break the coal off the face. The coal-water mixture falls into a metal sluice trough that carries the coal to further haulage equipment, usually conveyors. Production panels in a hydraulic mine are oriented to utilize gravity to carry the coal-water mixture away from the face and in the troughs. Hand loading and hydraulic mining can be either partial or full extraction depending on the extent of pillar removal in the panels.

3.7 *In Situ* Methods

In situ methods extract the coal with no human contact inside the coal seam, such as miners going underground to attend machinery. These methods include augering, highwall mining, and leaching or dissolving the coal by heat, pressure, or chemical means followed by capture of the media.

Augering consists of extracting coal from an outcrop by the repeated and parallel boring of holes into the seam. The machine, called an auger, is literally a large drill that bores 1.5 ft to 5 ft diameter holes up to 600 ft deep into the coal seam. Augering sometimes occurs in tandem with the surface contour mining method. Contour surface mining creates a flat bench on which to operate the auger.

Highwall mining is similar to augering. Highwall mining bores rectangular holes into a coal seam from an outcrop using a remote-controlled continuous miner connected to an extensible conveyor, or haulage machine. Highwall mining also can work in tandem with contour mining.

3.8 Underground Mine Design Parameters

As introduced, underground mining methods fall into two groups: full extraction and partial extraction. Key mine design considerations for underground mining include roof control, ventilation, seam pitch, extraction ratio, and industrial engineering as it pertains to maximizing safety and productivity.

3.9 Roof Control

Roof control is defined as the systematic design and installation of coal pillars and artificial structures to support the roof, floor, and sides of entries to protect the miner from falling and propelled rock. Roof, floor, and entry sides are subjected to three main forces that cause instability and require that support be installed in underground coal mines. These forces are gravity, mining-induced forces, or forces derived tectonics and mountain building. Gravity primarily loads underground workings with vertical stress. The general rule for U.S. underground mining is that gravity stress equals 1.1 psi per ft of depth of overburden. This value is derived from an assumed weight of sedimentary rock that overlays coal seams of 158.4 lb per ft^3.

Mining-induced forces are caused by the transference of overburden load from the void to pillars adjacent to the void's perimeter. As the void increases in size during full extraction, the weight transferred to the abutment pillars increases until the rock caves. In partial extraction mine design, the void area is limited to the spans across the entry roof. The roof rock in the entry remains stable. In full extraction mining, the caved rock is called gob or goaf, which is broken rock. As the gob consolidates and forms the ability to support weight, some load that was originally transferred to the pillars is reaccepted and transferred back to the gob. This process is called gob formation.

Tectonic forces in mining are caused by plate tectonics, regional mountain building forces, stress that resides in structural geology such as faults and slips, and other natural geologic processes. The result of tectonic force and stress in underground coal mining is increased stress in all directions, particularly in the horizontal plane, or the plane parallel to the bedding (layers of rock). Coal mining can reactivate stresses along structural geologic features. The primary coal mine design criteria for high horizontal stress is that the long axis of a production panel rectangle (in plan view) be oriented 15° to 45° off the maximum horizontal stress. Other than high horizontal stress, other forces from tectonic origin are difficult to incorporate into design because of the uncertainty of the force's magnitude, direction, and dynamics.

The primary design criteria in roof control in underground coal mining is to control the roof overhead where miners work and to provide abutment support in the form of pillars and blocks of coal on the edges of the excavation or void area. Pillars are designed by empirical methods and by mathematical stress modeling. Empirical formulas have been derived from actual mine case histories in all coal producing regions of the world. Most of these formulas have three primary variables to determine pillar strength: *in situ* coal strength, pillar height, and pillar width.

For pillar design in the United States, the vertical stress, assumed to be 1.1 psi per foot of depth, plus other mining induced stresses caused by load transfer would be compared against pillar strength. A pillar size would be chosen that has a satisfactory safety factor. The safety factor is strength divided by load. Each empirical method and mathematical stress modeling approach has recommended design pillar safety factors. Pillar design needs to recognize the variability of geology and the nonhomogeneous nature of coal. Rock mechanics design for underground coal mines requires a thorough design analysis of the local site geology and the comparison of mining conditions at neighboring mines. Civil engineering structures have more uniform design methodology because materials used in buildings, bridges, roads, and other structures are composed of homogeneous materials and the forces demanded on a structure are more predictable.

The main fixture used for primary support in underground coal mining is the roof bolt. A roof bolt is a steel rod from 0.5 in. to more than 1.5 in. diameter and from 2 ft to over 10 ft long. Roof bolts are installed on cycle, meaning that the bolts are installed after the entry or tunnel is excavated a specific distance. This distance the entry is advanced is called the depth-of-cut. The depth-of-cut of an entry, which ranges from 2 ft to over 40 ft, is determined by the ability of the roof rock to remain in place before being bolted; that is, to be stable. Roof bolts are anchored in a hole drilled into the roof by resin grout or by mechanical systems. Roof bolts systems bind the layered roof rock together by forming a beam or by suspending the rock layers from a zone of competent anchorage in the geologic cross section.

Roof is also supported by what is called secondary support. Secondary support is installed after the entry is advanced and after roof bolts are installed. Secondary support is generically called cribbing, which is named after wooden blocks stacked Lincoln-log style to form a vertical, free-standing structure that is wedged between the roof and floor of the entry. Other forms of cribbing or secondary support, in addition to wood blocks, include columns of cementaceous grout, steel structures, wooden posts, and wooden block structures (other than Lincoln-log stacked). Secondary support is used to fortify the

entries around a full-extraction area to assist in roof support when ground load is transferred from the void to abutment pillars. The distance the load is transferred is a function of the strength and thickness characteristics of the rock over the void. The presence of strong and massive sandstones in the overburden can resist caving over long spans. The cantilevering nature of the sandstone beam can create large stress increases in the abutment areas, which requires special mine designs be implemented.

3.10 Ventilation

Ventilation, along with roof control, is critical to the safe and efficient extract of coal in underground coal mining. Ventilation is necessary to provide air to all areas of an underground coal mine, except those areas that have been intentionally sealed, to dilute mine gases, dilute airborne dust, and provide fresh air for miners. A coal mine is ventilated by main mine fans located on the surface, which either push air through the mine called a blowing system or pull air through the mine called an exhausting system. Ventilated air moves through the mine in the entries that comprise independent circuits of air called splits. The main fan that pumps air though the mine is designed to deliver a minimum quantity of air to each area of the mine where miners and equipment extract the coal. The fan size and horsepower are designed to overcome friction losses that accumulate because of turbulent airflow against the rubbing surface of the entry roof, floor, and sides. The fan creates a pressure differential that causes air to flow from areas of high pressure to areas of low pressure. Air pressure differential can also be created by atmospheric variations caused by weather patterns and seasonal and daily temperature fluctuations. These pressure variations can move air through mines, particularly in abandoned mines or idled mines that are missing mechanical ventilation devices (fans). This form of ventilation is called natural ventilation. Ventilation design for mines with regard to the size of main fans must recognize natural ventilation.

Ventilation circuits within underground coal mines can be a single split or a double split of air. Double split is often called fishtail ventilation (Fig. 10).

As shown in Fig. 10, air movement is directed to the working faces in the entries. The stoppings shown are called ventilation control devices. Stoppings are walls, usually constructed of cinder block and mortar, built in the crosscuts to separate air circuits or splits of air. Without the stoppings, air flow would short circuit and mix with other air

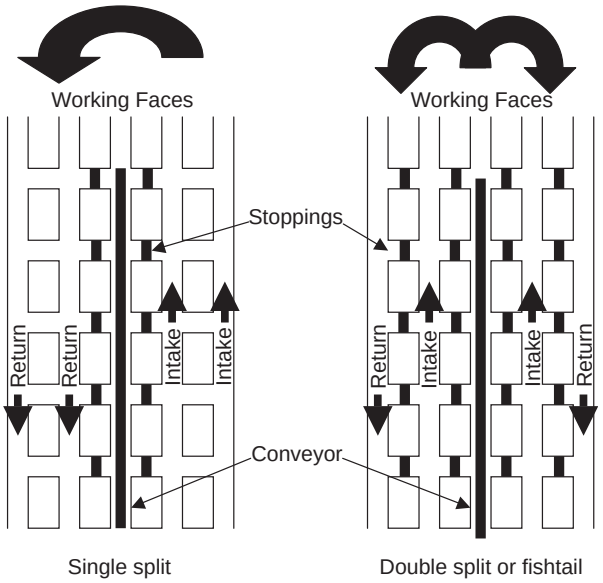


FIGURE 10 Single-and double-split ventilation showing air flow (plan view).

before reaching the working faces. Intake air is fresh air delivered to the working face that has not been used to ventilate other working faces where coal is produced. Return air is air that has been used to ventilate the working face. Return air contains coal dust and mine gases, particularly methane that has been liberated from the coal during production. Sometimes two different airways cross at an entry and a crosscut. A ventilation control device called overcast or undercast separates the two different types of air (i.e., intake and return).

Figures 7, 9, and 10 illustrate the entries containing the conveyor. The entry containing the conveyor usually contains a minimum of airflow and is not used to ventilate the working face. The airflow velocity is lessened in this entry to keep coal dust from becoming airborne.

Some mines utilize conveyor air to ventilate the working face. This practice occurs in longwall mines with long panels. Long panels require parallel intake or return air splits to increase air quantity to the working face and minimize resistance.

Ventilation design for a coal mine is based choosing a mine fan or fans, as in the case of a mine with multiple access openings, that can overcome the resistance to airflow and maintain the required air quantity at the working faces. Mine resistance varies with air quantity and obeys basic empirical laws that have friction factors based on measured data. Increasing the number of parallel entries of the spilt reduces resistance in that split.

3.11 Seam Pitch

Mines are designed to incorporate seam pitch. Rarely are coal seams completely flat. Geologic processes of coal formation, deposition, and structure cause seams to have pitch, which can vary from 0 degrees to vertical. Partial- and-full extraction mining methods can efficiently handle up to approximately 10 degrees of pitch. An underground mine plan is designed to take advantage of pitch when possible for water handling, pumping, material haulage, and gas migration.

3.12 Extraction Ratio

The extraction ratio is defined as the volume of coal mined divided by the total volume of coal within a reserve. Coal is a nonrenewable resource. Once it is combusted to make steam for electrical power generation, steel production, or other uses, all that remains is ash and gaseous substances. Maximizing the volume of coal mined from a given reserve is important for the overall energy cycle and to minimize the cost of production. With the exception of in situ mining systems, partial extraction underground coal mining methods range in extraction ratio from 20% to 50%, and full-extraction mining methods have extraction ratios that range from 50% to over 85%. Achieving a 100% volumetric extraction ratio is difficult for three reasons. First, pillars that are necessary for roof support in the plan cannot be mined. Second, coal is sometimes left in place on the floor or roof of an entry to improve the quality of the coal extracted. The immediate top or bottom layer of a coal seam can have a greater content of noncombustible material than the central portion of the coal seam. Finally, underground coal mining methods and machinery have a maximum practical mining height of approximately 14.5 ft.

Surface mining methods cannot achieve 100% extraction. Coal is damaged from drilling and blasting and cannot be recovered. Also, overburden dilution in pit recovery operations reduces recovery.

3.13 Industrial Engineering

Underground coal mining, as in any producing enterprise, requires an infusion of industrial engineering to maximize safety and productivity. Coal mine design practice has shown that safety and productivity are mutually inclusive goals.

Underground mining methods are a mix of continuous and batch processes. For example, while the continuous miner is extracting the coal from the face of the entry, the process is in continuous mode. The continuous miner has to stop mining frequently to allow loaded shuttle cars to switch out with empty shuttle cars, to advance the face ventilation system, and to allow roof bolts to be installed. These stoppages cause the face operations to be an overall batch process. Once the shuttle cars dump the coal onto a conveyor belt, the material is continuously transported outside to a stockpile. The coal is then fed through a processing facility that crushes the coal and removes impurities to customer specifications in continuous processing. Underground coal mine design is optimized when the batch processes that occur in underground mining are engineered to become continuous processes. An example of an effort in this area has been the introduction of continuous haulage equipment, which in certain conditions replace the shuttle car: a batch process.

Mine surveying plays a key role in both surface and underground coal mining. Surveying is used to direct the mining advancement as planned, to monitor performance giving feedback on issues of underground extraction ratio and surface pit recovery, to define overburden rehandle, and to perform checks against weighed coal production for inventory control.

SEE ALSO THE FOLLOWING ARTICLES

Clean Coal Technology • *Coal, Chemical and Physical Properties* • *Coal Conversion* • *Coal, Fuel and Non-Fuel Uses* • *Coal Industry, Energy Policy in* • *Coal Industry, History of* • *Coal Mine Reclamation and Remediation* • *Coal Mining in Appalachia, History of* • *Coal Preparation* • *Coal Resources, Formation of* • *Coal Storage and Transportation* • *Markets for Coal* • *Oil and Natural Gas Drilling*

Further Reading

Hartman, H. L. (ed.). (1992). "SME Mining Engineering Handbook". 2nd ed. pp. 1298–1984. Society of Mining, Metallurgy, and Exploration, Littleton, CO.

Hartman, H., and Mutmansky, J. (2002). "Introductory Mining Engineering." 2nd ed. John Wiley & Sons, Hoboken, NJ.

Mark, C. (1990). "Pillar Design Methods for Longwall Mining." U S Bureau of Mines Information Circular 9247, Washington, DC. pp. 1–53.

Peng, S. S. (1978). "Coal Mine Ground Control." John Wiley & Sons, New York.

Pfleider, E. P. (1972). "Surface Mining." American Institute of Mining Metallurgical and Petroleum Engineers, New York.

Woodruff, S. D. (1966). "Methods of Working Coal And Metal Mines," Volume 3. Pergamom Press, Oxford.

Coal Mining in Appalachia, History of

GEOFFREY L. BUCKLEY
Ohio University
Athens, Ohio, United States

1. Introduction
2. Historical Overview
3. Breaking Ground
4. Life in the Company Town
5. Working in a Coal Mine

Glossary

Appalachian Regional Commission Federal agency established in 1965 "to serve the needs of 21,000,000 people residing in one of the most economically distressed regions of the country."

black lung The common name for coal workers' pneumoconiosis, a respiratory disease caused by the inhalation of coal dust.

longwall mining A method of mining in which long sections of coal—up to 1000 feet across—are mined and deposited directly onto a conveyor system with the help of shields that support the roof and advance the longwall rock sections.

mountaintop removal A surface mining technique in which the top of a mountain is blasted away to expose a seam of coal; debris is then deposited in a valley fill.

rock dusting White, powdered limestone is sprayed on the roofs, bottoms, and ribs of a mine to mitigate the explosive qualities of coal dust.

roof bolting A process in which holes are drilled into the roof of a mine; long bolts coated with glue are then screwed into the holes to support the roof.

scrip Nonlegal tender issued to workers in place of cash; generally, redeemable only at the company-owned store.

timbering The use of wood for constructing a roof support in an underground mine.

tipple An elevated structure, located near a mine entrance, that receives coal from mine cars or conveyors and from which coal is dumped or "tipped," washed, screened, and then loaded into railroad cars or trucks.

welfare capitalism An industrial strategy in which company flexibility and benevolence are exercised in an attempt to maximize production and promote social control of the workforce.

yellow-dog contracts Agreements that prohibit employees from joining or supporting labor unions.

Appalachia—as defined by the United States Appalachian Regional Commission (ARC)—is a sprawling region encompassing approximately 400 counties in 13 eastern states: New York, Pennsylvania, Ohio, Maryland, West Virginia, Virginia, Kentucky, Tennessee, North Carolina, South Carolina, Georgia, Alabama, and Mississippi. Rich in natural resources, such as timber, coal, oil, iron ore, gold, and copper, and located within relatively easy reach of major industrial centers along the Atlantic seaboard, this expansive area supplied the raw materials that fueled America's Industrial Revolution. With vast deposits of bituminous ("soft") coal underlying approximately 72,000 square miles from Pennsylvania to Alabama, and an additional 484 square miles of anthracite ("hard") coal concentrated in northeastern Pennsylvania, exploitation of this one resource uniquely shaped the region.

1. INTRODUCTION

Opening the Appalachian coalfields was no easy feat. Nor was it achieved overnight. For aspiring "captains of industry," there were great financial risks involved. Substantial capital investment was required to survey and purchase coal lands, to construct mines and provide for the needs of a large workforce, and to establish a reliable and cost-effective means of delivering this bulky and relatively low-value good to market. For many, the road to success led instead to financial ruin. For those

 © 2004 Elsevier Inc. All rights reserved.

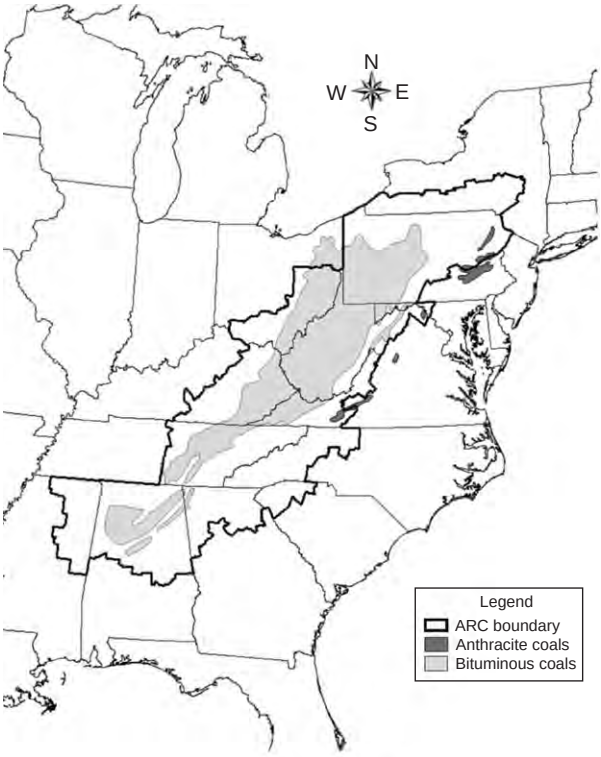


FIGURE 1 Location of mineral resources in Appalachia. ARC, Appalachian Regional Commission. Adapted from Drake (2000).

charged with the task of extracting "black diamonds" from the earth, there were great personal risks involved. Mining is and always has been hazardous work. Historically, periodic layoffs and the threat of black lung added to the worries of the miner and his family. In the final analysis, it must be concluded that coal mining in Appalachia has had as profound an impact on the region's inhabitants as it has had on the region's forest and water resources. Although much diminished, King Coal's influence is still discernible today—both in the region's historic mining districts and in those areas where underground and surface mining are still being carried out (Fig. 1).

2. *HISTORICAL OVERVIEW*

In his opening address to the White House Conference on Conservation in 1908, President Theodore Roosevelt reminded those in attendance that coal's ascendancy was a relatively recent occurrence: "In [George] Washington's time anthracite coal was known only as a useless black stone; and the great fields of bituminous coal were undiscovered. As steam was unknown, the use of coal for power production was undreamed of. Water was practically the only source of power, save the labor of men and animals; and this power was used only in the most primitive fashion." Over the course of the next century, however, coal's utility would be demonstrated and its value to industry affirmed. By 1900, the year before Roosevelt took office, coal accounted for nearly 70% of the national energy market in the United States. In short, it had become America's fuel of choice and Appalachia was the nation's primary producer.

2.1 Fueling the Industrial Revolution

During the early years of the 19th century, coal was used primarily to fire blacksmith forges and as a domestic heating fuel. In ensuing years, it was valued as a fuel for salt and iron furnaces, brick and pottery kilns, and steam engines. Although annual coal production increased significantly between 1800 and 1860—from a mere 100,000 tons to 20,000,000 tons—it was the post-Civil War years that first witnessed truly explosive growth in coal production and consumption in the United States. In 1885, the annual production figure had climbed to 110,000,000 tons. Stimulated by growth in domestic consumption, as well as increased use as a fuel for steamship travel, coal was now being used to stoke the furnaces of an emerging steel industry and to provide the motive power for steam locomotives. By 1900, production topped 243,000,000 tons, making the United States the world's leading producer. With respect to this tremendous growth, historian Duane A. Smith notes: "Without mining—from coal to iron to gold—the United States could not have emerged as a world power by the turn of the century, nor could it have successfully launched its international career of the twentieth century."

It would be difficult to overstate the importance of Appalachian coal in propelling America into the industrial age. Prior to World War II, virtually all coal mined in the United States came from the Appalachian fields. In 1925, for example, the Appalachian region produced 92% of the total amount of coal mined in the United States, with Pennsylvania alone accounting for approximately one-third of this output. Between 1880 and 1930, Pennsylvania consistently ranked as the leading producer of both bituminous and anthracite coal. By the mid-1920s, however, it had become apparent that all was not well with the coal industry in Appalachia.

2.2 The Decline of King Coal

Several years in advance of the stock market crash of 1929, the coal industry in general and miners in particular felt the first tremors of the Great Depression. In Appalachia, overexpansion stimulated by World War I, interstate competition with midwestern states, and competition between "union" and "nonunion" mines led to market gluts, falling prices, spiraling unemployment, and labor unrest. Although mining had weathered economic downturns in the past, most notably in the mid-1890s, nothing compared with the depression that beset the industry during the 1920s and 1930s. West Virginia, which lost 33,000 coal jobs as production fell from 146 million tons in 1927 to 83.3 million tons in 1932, was hit particularly hard.

In subsequent years, the rising popularity of new energy sources—petroleum, natural gas, and electricity—challenged coal's dominance of the domestic heating market. Although World War II provided a boost to the region's coal-based economy, it was only temporary. When railroads completed the conversion from coal to diesel after the war, the industry received another blow, contributing further to the industry's recent history of boom and bust. The coal operators' response was to accelerate the process of mechanization begun in earnest during the 1930s and to opt for surface mining over underground mining wherever feasible. Economic instability coupled with widespread utilization of new labor-saving machinery, typified most noticeably by the introduction of enormous electric-powered shovels to surface operations, forced many miners and their families to leave the region altogether in the post-World War II era.

Since the 1960s, the only major growth market for coal has been to supply electric power producers and for export abroad. With respect to the former, coal remains an important player. Today 52% of U.S. electricity is generated by coal-fired plants. However, increased production from western states starting in the late 1950s, but gaining momentum especially during the 1980s and 1990s, has further eroded Appalachia's share of coal markets. Compliance with environmental laws since the 1970s, including the Clean Air Act and its amendments, the Clean Water Act, and the Surface Mining Control and Reclamation Act (SMCRA), not to mention various state-level regulations, has also contributed to a rise in coal production costs. By 1999, more than half of the coal mined in the United States was coming from states located west of the Mississippi River, and far and away the largest single producer was Wyoming, not Pennsylvania. Nevertheless, coal mining has managed to remain viable in some portions of the Appalachian region due in large part to controversial mining techniques such as mountaintop removal.

3. BREAKING GROUND

Given the increase in demand for coal spurred on by industrialization, the existence of coal so near to the Atlantic seaboard aroused early interest among local boosters, land speculators, foreign investors, and others wishing to purchase mineral lands and mine them for profit. Before large-scale coal-mining operations could commence, however, certain obstacles had to be surmounted. First, coal lands had to be surveyed and evaluated, and the rights to the coal purchased or leased. Equally important, a reliable and cost-effective means of transporting the coal to eastern and Great Lakes markets, as well as to the steel mills of the North and South, had to be financed and developed. As historian Crandall A. Shifflett has pointed out: "All of the land and coal was worthless… if it could not be gotten out of its mountain redoubts. There was no use to open mines if the coal could not be transported to distant markets."

3.1 Railroads Usher in the Industrial Era

As early as the 1820s, coal mining had become an important, albeit seasonal, activity in some remote locations. During the first decade of the 19th century, for example, small quantities of coal, along with lumber and a variety of agricultural goods, were sent down the Potomac River via flatboat to Georgetown. On arrival, the large rafts were dismantled and their crews walked home. Such trade could only take place during the spring season when floodwaters raised the level of the river. Without significant investment in transportation improvements, it was clear that coal could not be mined profitably on a large scale.

Although coal and coke (fuel derived from coal) were slow to replace wood and charcoal in the manufacture of iron prior to the Civil War, coal found an early use in blacksmith shops and as a domestic heating fuel. After about 1840, coal gradually began to replace charcoal in the furnaces of the burgeoning iron industry. By this time, it was also being recognized as an important source of power for steam engines. After the Civil War, it would play an indispensable role in the manufacture of steel.

Initially, it was the anthracite fields of northeastern Pennsylvania that attracted great attention. Once problems associated with the burning of hard coal were resolved, production rose sharply. Between 1830 and 1860, production of Pennsylvania anthracite, much of it destined for the Northeast, increased more than 40-fold. As Andrew Roy put it in 1903, "the 'black rocks' of the mountains, heretofore deemed useless, at once rose in public estimation." Miners' villages sprouted wherever new mines were opened as English, Welsh, Irish, Germans, and, in later years, eastern and southern Europeans entered the region in search of high wages and a new beginning in America.

Before America's Industrial Revolution could "take off," another revolution—this one in the transportation sector—was needed. As the country grew in geographic extent, government officials became increasingly aware that improvements in communication and trade depended largely on the development of new, more efficient modes of travel. In 1808, Albert Gallatin, Secretary of the Treasury under Thomas Jefferson, envisioned a system of roads and canals crisscrossing the country and financed by the federal government. Although the Gallatin Plan was never carried out, great changes were in the works. The spectacular success of New York's Erie Canal served as a catalyst. Indeed, the completion of the Erie Canal in 1825 ushered in a period of canal building in the United States as eastern cities competed with one another in a race to expand their hinterlands and capture market share. Government officials and businessmen in Pennsylvania, Maryland, Ohio, and Virginia soon unveiled plans for canal systems they hoped would allow them to remain competitive with New York. In the end, none of the canal systems to the south could match the financial success of the Erie Canal and its branch connections. Difficult terrain and competition from a new form of transportation, the railroad, ensured that the reign of the canal would be relatively short (Fig. 2).

On the eve of the Civil War, large parts of northern Appalachia had established outlets for coal via rivers, canals, and railroads. In addition to the anthracite fields of eastern Pennsylvania, bituminous coal from the great "Pittsburgh" seam in western Pennsylvania, western Maryland, northern West Virginia, and Ohio was now being tapped. In contrast, central and southern Appalachia remained relatively unaffected by the transportation revolution that had seized the north. After the Civil War, an already impressive network of railroads in the Appalachian north was expanded. As these railroads, and to a lesser extent canals, penetrated the mountains, operations were expanded in western Maryland, western Pennsylvania, northern West Virginia, and southeastern Ohio, contributing to the growth of the iron industry and, especially after the introduction of the Bessemer hot-blast furnace in the 1860s, the nascent steel industry. Following the steady progress of the railroads, new fields were opened up in southern West Virginia, northeastern Tennessee, and northern Alabama, and finally, by the 1890s, southwestern Virginia and eastern Kentucky. As was the case with Pennsylvania's anthracite fields, the opening of mines in these areas was accompanied by a flurry of coal town construction and an influx of immigrant labor. As historian Richard Drake reminds us, in addition to carrying a great deal of coal, rail lines such as the Baltimore and Ohio, Chesapeake and Ohio, Norfolk and Western, Louisville and Nashville, and the Southern Railway, to name but a few, figured prominently in the growth of major industrial centers concentrated in the northern and southern extremities of the region. Of particular importance was the development of Pittsburgh and its industrial satellites—Johnstown, Altoona, Morgantown, and Fairmont—and the "A.B.C." triangle of the Deep South—Atlanta, Birmingham, and Chattanooga.

FIGURE 2 Rodgers locomotive on turntable at Corning, Ohio, ca. 1910. Reprinted from the Buhla Collection, Ohio University, Archives and Special Collections, with permission.

3.2 Clearing the Way

Long before the new transportation lines reached their destinations, however, land speculators were clearing the way for this initial wave of industrialization. During the first half of the 19th century, corporations interested in mining coal and iron ore

and cutting timber, and backed by capital from centers of commerce such as New York, Baltimore, Boston, and London, dispatched geologists and engineers into the mountains to measure the size and evaluate the quality of coal and iron ore deposits. When favorable reports were returned, they began accumulating property rights, a process that was often mediated by land speculators. Some of these speculators and businessmen were outsiders who entered the region as representatives of mining interests headquartered in the east. Others were local residents who recognized the value of coal and accumulated property rights based on the assumption that the arrival of the railroad would cause real estate values to soar. Still others were opportunistic politicians—men like Henry Gassaway Davis, Johnson Camden, and Clarence Watson—who used their power and influence in the chambers of the U.S. Senate and in various state capitols to finance railroads, purchase coal lands, and amass personal fortunes. Invariably the result was the same. By the time the railroad arrived, the coalfields were controlled for the most part by outside interests eager to begin the mining process.

One of the best known of these early entrepreneurs was Pike County, Kentucky native John C. C. Mayo. During the 1880s and 1890s, Mayo traveled throughout eastern Kentucky purchasing land or mineral rights from local farmers. Using the infamous "Broadform Deed," which permitted the mountain farmer to retain rights to the surface while authorizing the owner of the mineral rights to use any means necessary to extract the minerals, Mayo acquired thousands of acres of land that he eventually sold to companies such as the giant Consolidation Coal Company. As late as the 1980s, coal companies in Kentucky could count on the courts to accept the binding nature of Broadform Deeds signed in the late 19th and early 20th centuries. However, rather than extract the coal using underground mining techniques, as was the practice during the 19th and early 20th centuries, companies were now employing surface mining techniques. Essentially, the Broadform Deeds permitted coal operators to destroy the surface to reach the mineral deposits. The stability of many rural communities was sorely tested. In 1988, the citizens of Kentucky passed an amendment to the state constitution requiring coal companies to obtain permission from surface owners before strip mining land. Five years later, the Kentucky Court of Appeals accepted the constitutionality of this amendment.

4. LIFE IN THE COMPANY TOWN

A vexing problem that confronted coal operators in the early stages of mine development was a shortage of labor. Because exploitable seams of coal were often found in sparsely populated areas, operators were forced to recruit and import labor. Although some coal operators used prison labor, as in northern Alabama, or drew from the pool of displaced mountain farmers, by far the greater number chose to lure white labor from Europe or, in the years leading up to World War I, to enlist black migrants from the South. Thus, in addition to mine development and railroad construction, coal companies had to provide housing and other amenities for newly arrived miners and their families.

These company towns, as they came to be known, played an especially important role in central and southern Appalachia during the 19th and early 20th centuries. In eastern Kentucky, southern West Virginia, and southwestern Virginia during the 1920s, for example, an estimated two-thirds to three-fourths of all miners and their families resided in such communities. In southern West Virginia, the figure approached 98%. Beginning in the 1930s, the company town's usefulness began to fade. A number of factors contributed to this decline, including the diffusion of the automobile, which greatly enhanced worker mobility; the introduction of labor-saving devices, which made workers redundant; and, later, the adoption of surface-mining techniques, which further reduced the need for a large labor force. No longer profitable, by the 1950s and 1960s most mining firms had sold or otherwise disposed of company housing.

4.1 A Distinctive Feature of the Mining Landscape

There can be little doubt that the company town is one of the most distinctive features associated with Appalachia's historic mining districts. Although some scholars have argued that company towns were a necessity, at least at the outset, and that many companies treated their workers with respect and fairness, others contend that these carefully planned and controlled settlements allowed companies to exert an objectionable level of power over miners and other residents. Considering the extent to which some companies dominated local law enforcement, monitored educational and religious institutions, curtailed personal liberties, and influenced worker

behavior, their power was pervasive. As Shifflett reminds us, however, some miners' families, especially the first generation to leave hardscrabble farms in rural Appalachia, may have favored life in the company town, especially the more durably constructed "model" towns: "Contrary to the views that the coal mines and company controlled towns led to social fragmentation, disaffection, and alienation of the workforce, many miners and their families found life in them to be a great improvement over the past. Miners viewed the company town, not in comparison to some idyllic world of freedom, independence and harmony, but against the backdrop of small farms on rocky hillsides of Tennessee, Kentucky, and Virginia, or a sharecropper's life in the deep South states."

Company towns in Appalachia generally shared several distinguishing features (Fig. 3). First, the town bore the stamp of the company in every conceivable way. Everything in the town, from the houses and the company store to the schools, churches, hospitals, and the recreational facilities, if such "frills" even existed, was provided by the company. Second, housing generally conformed to a standard design. In general, one worker's house was identical to the next with respect to style and construction. A third trait was economy of construction. Companies typically limited their investment in towns because they knew they were not likely to evolve into permanent settlements. Renting houses to workers ensured that even a substantial investment in housing would be recovered quickly. Because companies generally preferred to hire married men with families over single men, believing they were less inclined to quit their jobs, single- and two-family dwellings were far more common in the coalfields than were large boarding houses. A fourth feature was that the company houses and other structures were usually laid out in a standard gridlike pattern, with long, narrow lots lined up along the railroad tracks. Finally, there was an outward manifestation of socioeconomic stratification and segregation according to racial and ethnic group expressed in the landscape. With respect to the former, sharp distinctions could be found in the size and quality of housing occupied by miners and company officials. Regarding the latter, historian Ronald Lewis notes that some companies sought a "judicious mix" of whites, blacks, and immigrants as a means of maintaining order and fending off the advances of union organizers. Thus it was not unusual to find distinct "neighborhoods"—an immigrant town, a Negro town, and an "American" town—existing side by side in the same town, each with its own separate facilities and social and cultural life.

One of the most controversial features of the typical coal mining town was the company store. Indeed, it represented both the best and worst of company town life. Often doubling as a meeting hall, lodge, and recreation facility, while also housing the company offices, the store carried a great variety of goods, including food and clothing, mining supplies, tools, and durable household goods. It also served as an important informal gathering place for both miners and managers alike. Given the remote location of many towns, the company store was often the only commercial retail establishment to which the mining population had regular access. In some cases, coal companies prohibited the establishment of other stores; in other cases, they penalized miners for making purchases in neighboring towns. Customer loyalty was achieved through the issuance of company scrip. Paying wages at least partially in scrip ensured that a portion of a worker's paycheck would be funneled back to the company. Although company officials claimed that such an arrangement protected families, critics have argued that companies charged higher prices in an attempt to keep the cost of coal production down. It is not surprising, then, that charges of debt peonage, coercion, price gouging, and differential pricing were leveled at the company store.

Some companies encouraged the planting of vegetable and flower gardens to improve the appearance of towns and to supplement food supplies. With the men at work in the mines, primary responsibility for the planting and care of the gardens rested on the

FIGURE 3 Railroad station, Fleming, Kentucky, ca. November 1915. Reprinted from the Smithsonian Institution, National Museum of American History, with permission.

shoulders of women and children. While they contributed significantly to the miners' larder, the gardens may have served another purpose. By encouraging miners to grow much of their own food, company officials could justify paying lower wages. Such a strategy enabled companies located at a considerable distance from major coal markets to keep the price of coal low and remain competitive with companies located nearer to centers of industry and population.

4.2 Culturally Diverse Communities

Coal towns in Appalachia often exhibited a high degree of cultural diversity. During the first half of the 19th century, immigrants to the minefields typically came from England, Wales, Scotland, Ireland, and Germany. By the end of the century, large numbers were being recruited from southern and eastern Europe, a trend that continued well into the first two decades of the 20th century (Fig. 4). When immigration from Europe plummeted during the war years, the door was opened wide for large numbers of African-Americans to enter the mines. Discouraged by low wages in the mines of northern Alabama or in search of seasonal work when there was a lull in activities on the family farm, this group made up a significant proportion of the workforce in places such as southern West Virginia and eastern Kentucky from approximately 1910 to the 1930s.

Cultural heterogeneity in the company town manifested itself in many and varied ways. It was not uncommon to find safety instructions posted at the mine opening written in several languages. Nor was it uncommon to find a diversity of religions represented in a single company town. Although some have argued that building churches and paying ministers' salaries allowed company officials to exercise even greater control over residents of the company town, others, citing low attendance, have downplayed the importance of religion in these communities.

FIGURE 4 Miners in front of Oakdale or Jumbo Mine No. 311, Ohio, 1887–1907. Reprinted from the Buhla Collection, Ohio University, Archives and Special Collections, with permission.

Although some coal operators looked on their creations with a certain measure of pride, believing them superior to anything one might find in outlying areas, living and working conditions, in fact, varied considerably from one location to the next. The time and money an operator was willing to invest in the development of a town depended on several factors, including the projected life of the mine, the number of houses that needed to be built, and the amount of capital the company had at its disposal. Towns constructed of durable materials and offering a range of amenities were clearly meant to last. Ramshackle housing, on the other hand, was a sure sign that a company's investment in a mine site was limited. According to contemporary theories of corporate paternalism and welfare capitalism, well-built towns equipped with electricity and indoor plumbing, and offering a wide range of amenities, such as churches, recreation buildings, meeting halls, ballparks, and stores, attracted a more dependable and loyal breed of miner—one who, in the end, would reject the temptation to join a union. Because the company maintained ownership of housing, the threat of eviction was never far from the mind of the miner. In the years following the First World War, companies fought off the advances of the unions by equating union membership with communist sympathy or radicalism. Companies employing a high number of foreign-born workers kept their workforce in line by equating loyalty to the company with patriotism and "Americanism." Eager to prove they were good American citizens and avoid persecution, recent immigrants were often reluctant to participate in union activities.

5. WORKING IN A COAL MINE

For the most part, coal miners during the 19th and early 20th centuries worked side by side in a dark, damp, and often dangerous environment, regardless of their cultural or ethnic background (Fig. 5). Given the stressful nature of their work and the dangers they faced, it was absolutely essential that safety and teamwork take precedence over all other matters in the mines. Dust and gas accumulations, roof falls,

FIGURE 5 Miner loading car at Mine No. 26, West Virginia. Reprinted from the Smithsonian Institution, National Museum of American History, with permission.

and heavy equipment malfunctions and accidents were just a few of the hazards miners dealt with on a daily basis. Under such conditions, even the most safety-conscious miner could be caught off guard. If a miner was fortunate enough to avoid a serious lost-time injury over the course of his career, black lung, a condition caused by the inhalation of coal dust, could cut his life short. Aboveground, miners went their separate ways. As Robert Armstead recently detailed in an autobiographical account, segregation and Jim Crow were a fact of life when the shift was over.

The coal mining industry in Appalachia has witnessed a great many changes since the days when miners toiled underground with picks and shovels and hand-loaded coal into wagons for transport to the surface. Although many of these changes have had a positive effect on the lives of miners and their families, e.g., improved safety, wage increases, and retirement benefits, to name just a few, others, such as the introduction of labor-saving equipment, have had the effect of putting miners out of work. Indeed, the effect that mechanization had on miners and mining communities during the middle years of the 20th century was far reaching.

5.1 Social Relations among Miners

Although relations between whites, blacks, and European immigrants were sometimes strained, especially when blacks or immigrants were brought in as strikebreakers, antagonism appears to have been the exception rather than the rule, at least on the job. According to Ronald Eller, "[a] relatively high degree of harmony existed between the races at a personal level. Working side by side in the mines, the men came to depend upon each other for their own safety, and the lack of major differences in housing, pay, and living conditions mitigated caste feelings and gave rise to a common consciousness of class." If integration typified relations in the mines, segregation characterized relations on the surface. When the workday was over, blacks and whites showered in separate facilities (if such facilities were, in fact, available) and then walked or drove home to their own neighborhoods. Foreign-born miners and their families were often subjected to similar, albeit more subtle, forms of discrimination.

Although small numbers of women worked in small "country bank" mines, they were generally discouraged from engaging in such work. In some states, such as Ohio, they were prohibited by law from coal mining work unless they actually owned the mine. Their absence is frequently attributed to the superstitious belief that a woman's presence in the mines was sure to bring bad luck. Starting in the early 1970s, small numbers of women were hired to join the ranks of male coal miners. Hiring peaked in 1978 and then tailed off again during the 1980s. The historical record seems to support geographer Richard Francaviglia's statement on the matter: "mining has traditionally been 'man's work', and no amount of neutralized language can…conceal that fact."

5.2 Mechanization

Coal mining was one of the last major industries in America to mechanize production. From the beginning of the 19th century through the first two decades of the 20th century, the manner in which coal was extracted from underground mines and brought to the surface changed very little. Before the advent of the mechanical loading machine in the 1920s, miners typically broke coal from the mine face using picks and wedges and shoveled it by hand into coal cars. Loaded cars were then pushed to the surface by miners or pulled by horses or other animals. By the end of the 19th century, miners were "undercutting" the seam and using drills and explosives to increase productivity. Considering that a miner's pay was directly linked to the amount of coal he mined and the number of cars he filled, new methods of mining that saved labor and increased productivity were quickly adopted.

As demand for coal increased over time, companies sought to overcome bottlenecks in the production process. When overexpansion and competition began

to cut into profits, and unionization forced the cost of coal production to go up, the drive to mechanize underground mining gained momentum. The logical starting point was to replace the hand-loading system with one that emphasized the use of mechanical loading machines. From the 1930s to the 1970s, the introduction of mechanical loaders, cutting machines, continuous miners, and longwall mining changed forever the way coal would be mined.

Another aspect of mining transformed by mechanization was in the area of haulage. During the pick and hand-loading era, miners shoveled coal into wagons and pushed them to the surface. Drift mines were generally constructed so that they pierced the hillside at a slight upward angle to allow for drainage and to facilitate the transport of loaded cars to the tipple. As mines grew in size and the distance to the tipple increased, companies turned to animal power as the principal means of haulage. When larger coal cars came on the scene, mechanical and electrical haulage, including conveyors, became increasingly necessary.

From the vantage point of the coal operator, mechanization offered several advantages. It allowed for more easily loaded coal, reduced the need for explosives, and lowered timbering costs. Most important, mechanization permitted greater amounts of coal to be mined. Mechanization had its drawbacks, however, at least from the perspective of the miner, for the introduction of laborsaving equipment greatly reduced the need for a large workforce. The introduction of the mechanical coal-loading machine, for instance, reduced the need for coal miners by approximately 30% industrywide between 1930 and 1950. According to Ronald Lewis, black miners bore the brunt of layoffs during this period because they were disproportionately employed as hand loaders. The continuous miner, which combined cutting, drilling, blasting, and loading functions in one machine, had a similar impact. After the introduction of the continuous miner, the number of miners working in West Virginia was cut by more than half between 1950 and 1960. The impact of mechanization also shows up clearly in production and employment figures for Ohio. Between 1950 and 1970, coal production in this state climbed steadily despite a 90% decrease in the number of underground mines and an 83% decline in the size of the labor force. Starting in the 1970s, longwall mining cut even further into employee rolls.

Mechanization presented miners with other problems as well. In mining's early days, miners walked to and from the mine face and relied on natural ventilation to prevent the buildup of mine gas. As mines expanded in size, miners had to walk greater distances to get to work, and problems with ventilation and illumination arose. To facilitate the movement of men, electric trolleys were introduced to the mines. Dust and gas problems were alleviated somewhat by the installation of mechanical fans. Meanwhile, illumination was improved when safety lamps and electric lights were substituted for candles. While solving some problems, these and other solutions contributed to others. In the words of one authority on the matter, now workers could be "crushed, run over, squeezed between cars, or electrocuted on contact with bare trolley wires." In addition, some of the new machinery generated sparks and produced tremendous amounts of ignitable coal dust.

5.3 Health and Safety

Coal mining has always been and continues to be a hazardous occupation. Indeed, fatalities and serious nonfatal injuries occur still. The risks miners face today, however, pale in comparison to the perils miners faced prior to World War II. Accidents resulting from roof falls, ignition of mine gas and coal dust, the handling of mechanical haulage equipment, and the operation of electric-powered equipment of all types contributed to a high accident rate in the United States. Increased demand, fierce competition, advanced technology, negligence on the part of miners, and a poorly trained immigrant workforce have all been blamed for the high accident and fatality rates. Coal operators must share a portion of the blame as well. Opposed to unions and wary of burdensome safety regulations, operators often stood in the way of meaningful safety reform.

Initially, enactment of mine safety legislation rested with the states. In 1870, Pennsylvania passed the first substantive mine safety law. Over the next 25 years, several states passed mine inspection laws of their own. By the mid-20th century, 29 states had mine safety laws on the books. Unfortunately, these laws were often poorly enforced. Although the U.S. Bureau of Mines had been created in 1910, it was primarily an "information-gathering" agency. Authority to inspect underground mines was not conferred until 1941. Enforcement powers were not granted until 1953. Continued high fatality rates eventually sparked an effort to pass comprehensive federal legislation. Proponents of federal legislation had to wait until 1969 before such measures were enacted.

By far, the greatest number of deaths was caused by roof falls. Nevertheless, it was the ignition of methane gas and coal dust that captured the attention of the news media. Wherever miners used candles or open-flame safety lamps for illumination and black powder to break coal from the face, and where the mine was deep, "gassy," and poorly ventilated, conditions were ripe for an explosion. During the 10-year period from 1891 to 1900, 38 major explosions rocked American mines, resulting in 1006 deaths. Some of the worst explosions occurred during the first decade of the 20th century, including the worst one in American history, the dreadful Monongah mine disaster. On December 6, 1907, 362 miners employed by the Fairmont Coal Company of West Virginia were killed when the company's No. 6 and No. 8 mines blew up. Previously, Andrew Roy had noted that "more miners are annually killed by explosions in West Virginia, man for man employed, or ton for ton mined, than in any coal producing State in the Union or any nation in the world." It was clear that much needed to be done to improve mine safety in the United States.

Discovering the cause of mine explosions was of paramount importance to officials at the fledgling U.S. Bureau of Mines. Several causes were identified. First, as mines grew deeper, providing fresh air to miners and dispersing mine gas became more of a challenge. Given that miners typically used candles or open-flame safety lamps to light their way, this was a particularly serious problem. There was also a preference on the part of the miners to use greater amounts of explosives to "shoot off the solid," that is, to blast without first undermining the coal. As with the open flame, explosives were another source of ignition. Finally, there was the widely held belief that coal dust alone could not set off an explosion. Disproving this belief was particularly important (one that Europeans had confirmed by a much earlier date) given the increasing amounts of dust being produced by new mining machinery. The adoption of new technologies eventually reduced the risk of explosion. Improved ventilation, electric cap lamps, rock dusting, safer mining machinery, first-aid and mine rescue training, and passage of legislation permitting federal inspection of mines combined to improve conditions considerably. Although the frequency of mine explosions diminished considerably after the 1930s, the Farmington, West Virginia disaster of 1968, in which 78 miners were killed, served as a reminder that mine explosions were still a potential threat. With respect to roof falls, more effective supervision, systematic timbering, and roof bolting would, in time, reduce the frequency of these accidents.

5.4 Coal Mine Unionism

No history of coal mining in Appalachia would be complete without mentioning organized labor. Given the conditions that coal miners often had to work under, it is not surprising that they were among the first workers in the United States to organize. Given the power and control coal companies wielded throughout the better part of the 19th and early 20th centuries, it is also not surprising that coal operators fought the unions at every turn. Proponents of corporate paternalism and welfare capitalism believed that company benevolence could effectively stave off the unions, but others believed that coercion and intimidation would more effectively produce the desired results.

Although coal mine unionism in Appalachia can be traced to the early 1840s, the modern era of unionism began in 1890 when two groups, the National Progressive Union of Miners and Mine Laborers and the National Assembly Number 135 of Miners of the Knights of Labor, joined forces in Columbus, Ohio to form the United Mine Workers of America (UMWA). After winning recognition from Central Competitive Field Operators in 1897, the UMWA set out to stake a middle ground between ineffectual "company" unions and more radical unions, such as the National Miners Union. Under the able leadership of John Mitchell, the UMWA won strikes and concessions from the coal operators and built membership. Withdrawing from what they saw as a temporary truce with union organizers during the war years, coal operators sought to roll back the gains the UMWA had made during the first two decades of the 20th century. Citing a 1917 U.S. Supreme Court decision in which yellow-dog contracts forbidding union membership were ruled legal, coal operators set out to reclaim lost ground. An attempt by new union leader, John L. Lewis, to organize the southern coalfields during the middle of the decade was crushed. With demand for coal down and nonunion mines undercutting the cost of production in union mines, UMWA membership plummeted during the 1920s.

The 1930s, and more specifically President Franklin D. Roosevelt's New Deal legislation, breathed new life into the beleaguered union. Passage of the National Industrial Recovery Act (1933), the Bituminous Coal Code, or "Appalachian Agreement" (1933), the National Labor Relations, or "Wagner,"

Act (1935), and the first and second Bituminous Coal Conservation, or "Guffey," acts (1936 and 1937) shifted the advantage to the unions once again. Particularly important was the fact that the union had finally won the right to be recognized as a collective bargaining agent. Among the union's "victories" at this time was approval of an 8-hour workday, a 40-hour workweek, and higher wages. A minimum work age of 17 was also set, ending, in theory at least, the industry's long-standing practice of utilizing child labor. In addition, miners were no longer forced to use company scrip, to shop only at the company store, or to rent a company house, all mechanisms by which some companies attempted to keep up with the rising cost of production.

The gains of the 1930s and 1940s were replaced by the uncertainties of the 1950s and 1960s. The terms of the Love–Lewis agreement, signed in 1950 by Bituminous Coal Operators Association president George Love and UMWA president Lewis, reflect the ever-changing fortunes of the miners and their union. Although the agreement provided a high wage and improved health benefits for miners, it prevented the union from opposing any form of mechanization. The adoption of new technologies, including augering and stripping, resulted in layoffs and a decline in membership. Although certain gains were made, e.g., portal-to-portal pay, improved insurance and pension benefits, and passage of black lung legislation, the 1970s and 1980s saw many workers return to nonunion status.

Sometimes, relations between management and labor took a violent turn. Students of Appalachia know well the price that was paid in human life in places such as Matewan, West Virginia, "bloody" Harlan County, and at Blair Mountain, where 3000 UMWA marchers clashed on the battlefield with an army of West Virginia State Police, company mine guards, and assorted others representing the interests of the coal operators. In more recent times, violence has flared up in places such as Brookside, Kentucky. In 1969, charges of fraud and the murder of a UMWA presidential candidate tarnished the reputation of the union.

Although unions, the UMWA in particular, can take justifiable pride in improving the lot of miners, some scholars have argued that the higher labor costs associated with these victories ended up increasing the cost of coal production and accelerating the move toward mechanization. Thus, an impressive and hard-earned package of benefits was eventually passed down to an increasingly smaller pool of beneficiaries.

SEE ALSO THE FOLLOWING ARTICLES

Clean Coal Technology • Coal, Chemical and Physical Properties • Coal, Fuel and Non-Fuel Uses • Coal Industry, History of • Coal Mine Reclamation and Remediation • Coal Mining, Design and Methods of • Coal Storage and Transportation

Further Reading

Armstead, R. (2002). "Black Days, Black Dust: The Memories of an African American Coal Miner." The University of Tennessee Press, Knoxville.

Corbin, D. (1981). "Life, Work, and Rebellion in the Coal Fields: The Southern West Virginia Miners, 1880–1922." University of Illinois Press, Urbana.

Crowell, D. (1995). "History of the Coal-Mining Industry in Ohio." Department of Natural Resources, Division of Geological Survey, Columbus, Ohio.

Drake, R. (2000). "A History of Appalachia." The University Press of Kentucky, Lexington.

Eller, R. (1982). "Miners, Millhands, and Mountaineers: Industrialization of the Appalachian South, 1880–1930." The University of Tennessee Press, Knoxville.

Francaviglia, R. (1991). "Hard Places: Reading the Landscape of America's Historic Mining Districts." University of Iowa Press, Iowa City.

Gaventa, J. (1980). "Power and Powerlessness, Quiescence and Rebellion in an Appalachian Valley." University of Illinois Press, Urbana.

Harvey, K. (1969). "The Best-Dressed Miners: Life and Labor in the Maryland Coal Region, 1835–1910." Cornell University, Ithaca, New York.

Hennen, J. (1996). "The Americanization of West Virginia: Creating a Modern Industrial State, 1916–1925." University Press of Kentucky, Lexington.

Lewis, R. (1987). "Black Coal Miners in America: Race, Class, and Community Conflict, 1780–1980." University Press of Kentucky, Lexington.

Roy, A. (1903). "A History of the Coal Miners of the United States: From the Development of the Mines to the Close of the Anthracite Strike of 1902 Including a Brief Sketch of British Miners." J. L. Trauger Printing Co., Columbus, Ohio.

Salstrom, P. (1994). "Appalachia's Path to Dependency: Rethinking a Region's Economic History, 1730–1940." University Press of Kentucky, Lexington.

Shifflett, C. (1991). "Coal Towns: Life, Work, and Culture in Company Towns of Southern Appalachia, 1880–1960." The University of Tennessee Press, Knoxville.

Thomas, J. (1998). "An Appalachian New Deal: West Virginia in the Great Depression." University Press of Kentucky, Lexington.

Williams, J. (1976). "West Virginia and the Captains of Industry." West Virginia University Library, Morgantown.

Coal Preparation

PETER J. BETHELL
Bethell Processing Solutions LLC
Charleston, West Virginia, United States

GERALD H. LUTTRELL
Virginia Polytechnic Institute & State University
Blacksburg, Virginia, United States

1. **Introduction**
2. **Coal Sizing**
3. **Coal Cleaning**
4. **Solid-Liquid Separation**
5. **Miscellaneous**

Glossary

ash A measure of coal purity that consists of the noncombustible residue that remains after coal is completely burned.
coking coal A coal product having properties that make is suitable for consumption by the metallurgical market for the manufacture of steel.
cut point The density corresponding to a particle that has an equal probability of reporting to either the clean coal or reject product of a density-based separator.
cut size The particle size that has an equal probability of reporting to either the oversize or undersize product of a sizing device such as a screen or hydraulic classifier.
heavy media An aqueous suspension of micron-sized particles (usually magnetite) that is used to create an artificial, high-density fluid that is passed through various coal washing processes to separate coal from rock based on differences in density.
middling A composite particle containing both organic and inorganic matter.
open area The percentage of the total cross-sectional area of a screening surface that is open for particles to pass. A larger open area improves screen capacity and efficiency.
organic efficiency An indicator of separator performance (usually expressed as a percentage) that is calculated by dividing the actual clean coal yield by the theoretical maximum yield that could be achieved at the same ash content according to washability analysis.
steam coal A coal product having properties that make it suitable for consumption by the power industry for the production of electricity.
throw The maximum displacement created by the rotary vibrator of a screening mechanism.
washability analysis A laboratory procedure used to assess the potential cleanability of coal that uses dense liquids (usually organic) to partition a particulate sample into various density fractions (also called float-sink tests).
washing A generic term used to describe the process by which mined material is treated by various particulate separation processes to remove unwanted impurities from the valuable organic matter (also called *cleaning*).
yield An indicator of separation performance (usually reported as a percentage) that is calculated by dividing the clean coal tonnage by the feed coal tonnage.

Coal preparation is the process in which mined coal is upgraded to satisfy size and purity specifications that are dictated by a given market. The upgrading, which occurs after mining and before shipping to market, is achieved using low-cost unit operations that include screening, cleaning, and dewatering. The process is typically required for economic reasons that are driven by a desire to improve utilization properties, reduce transportation costs, or minimize environmental impacts.

1. INTRODUCTION

1.1 Why Process Coal?

Mined coal is a heterogeneous mixture of both organic (carbonaceous) and inorganic (mineral) matter. The inorganic matter is comprised of a wide range of mineral components including shale, slate, and clay. During combustion, these impurities reduce the heating value of the coal and leave behind an ash residue as waste. The impurities also influence the suitability of the coal for high-end uses such as the

 © 2004 Elsevier Inc. All rights reserved.

manufacture of metallurgical coke or generation of petrochemicals. Some of the mineral impurities, such as pyrite, form gaseous pollutants when burned that are potentially harmful to the environment. The impurities may represent a significant portion of the mined coal weight and, as a result, need to be removed to minimize the cost of transporting the coal to the customer. For these reasons, a tremendous incentive exists for the removal of inorganic impurities prior to the downstream utilization of the coal. Coal purchasing agreements commonly include limitations on ash (mineral matter), sulfur, and moisture.

1.2 Brief History

Little attention was given to coal preparation in the early industrial years. Hand sorting of rock and sizing by screening were the major methods of preparing mined coals for market. Most of the fine (<0.5 mm) coal was either discarded or sold for very low prices. More recently, increased emphasis has been placed on coal preparation because of the depletion of higher quality reserves and increased levels of out-of-seam dilution resulting from mechanized mining. Consumers in the steam (utility) and metallurgical (coking) coal markets have also placed new demands on coal producers to provide consistent high-quality products. The combined influence of these factors has made coal preparation far more important than it was in the past. There are estimated to be more than 2000 plants operating worldwide with capacities ranging from as little as 100 ton per hour to more than several thousand ton per hour. In terms of annual production, coal preparation facilities produced just over 1.3 billion tons of the 3.5 billion tons of coal sold by the top 10 coal-producing countries in 1998. The United States produced more than one-third of this tonnage (35%), followed by China (20%), South Africa (14%), and Australia (13%). The average plant capacity in North America is approximately 750 ton per hour.

1.3 Typical Systems

Despite the perception of coal preparation as a simple operation, it has become far more complicated than most people realize. Modern coal preparation plants can be as complex and sophisticated as processing facilities used in the chemical industries. However, for simplicity, plant flowsheets can be generically represented by a series of sequential unit operations for sizing, cleaning, and dewatering (Fig. 1). This sequence of operations, which is commonly called a circuit, will be repeated several times since the processes employed in modern plants each have a limited range of applicability in terms of particle size. Modern plants may include as many as four separate processing circuits for treating the coarse (plus 50 mm), intermediate (50×1 mm), fine (1×0.15 mm), and ultrafine (minus 0.15 mm) material. Although some commonalities exist, the selection of what number of circuits to use, which types of unit operations to employ, and how they should be configured are highly subjective and largely dependent on the inherent characteristics of the coal feedstock.

2. COAL SIZING

2.1 Background

Run-of-mine coal produced by mechanized mining operations contains particles as small as fine powder and as large as several hundred millimeters. Particles too large to pass into the plant are crushed to an appropriate upper size or rejected where insufficient recoverable coal is present in the coarse size fractions. Rotary breakers, jaw crushers, roll crushers, or sizers are used to achieve particle size reduction. Crushing may also improve the cleanability of the coal by liberating impurities locked within composite particles (called middlings) containing both organic and inorganic matter. The crushed material is then segregated into groups having well-defined maximum and minimum sizes. The sizing is achieved using various types of equipment including screens, sieves, and classifying cyclones. Screens are typically employed for sizing coarser particles, while various combinations of fine coal sieves and classifying cyclones are used for sizing finer particles. Figure 2 shows the typical sizes of particles that can be produced by common types of industrial sizing equipment.

2.2 Screening Principles

Screens are mechanical sizing devices that use a mesh or perforated plate to sort particles into an undersize product (particles that pass through the openings in the screen) and an oversize product (particles that are retained on the screen surface). In coal preparation, screens are used for a wide range of purposes including the extraction of large rock and trash prior to treatment and the production of optimum size

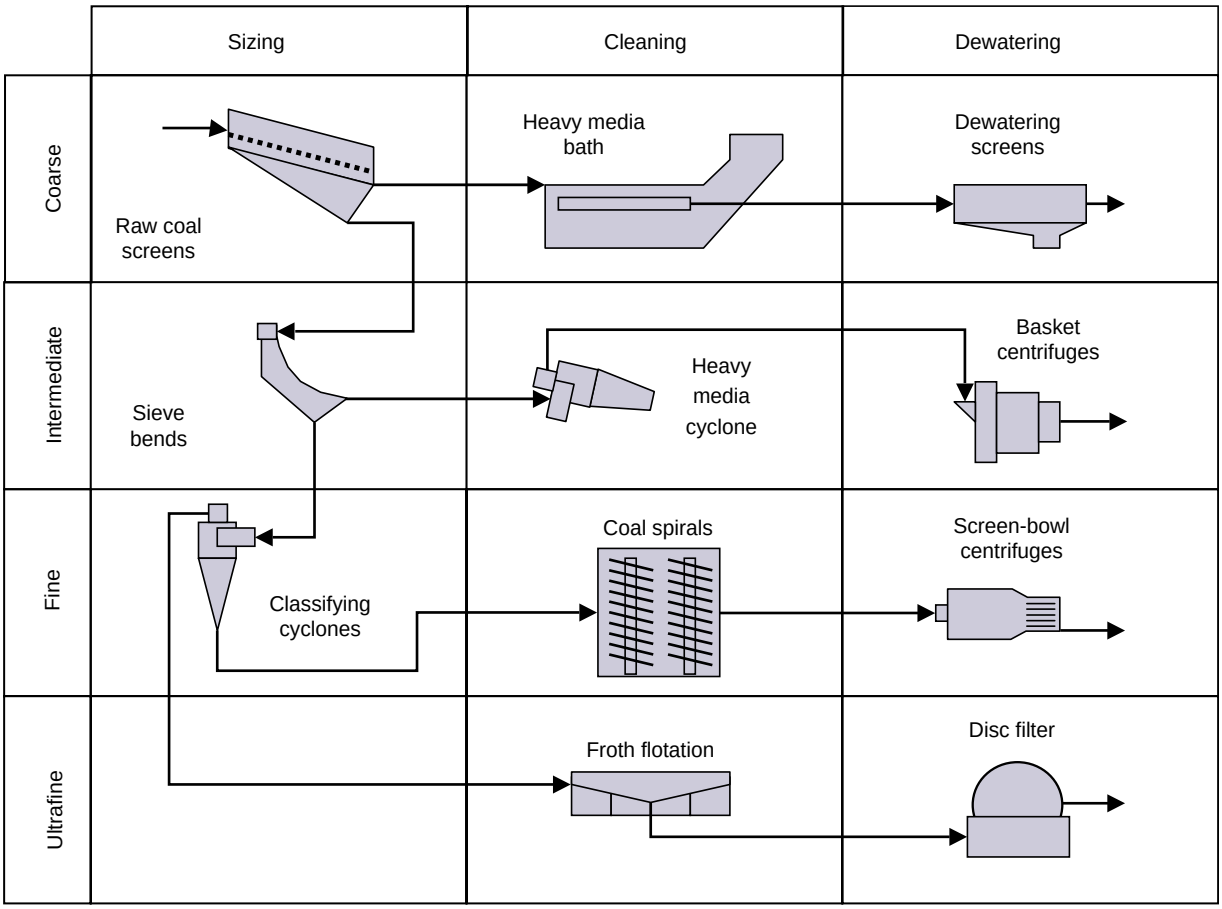


FIGURE 1 Conceptual preparation plant flowsheet subdivided according to unit operation and particle size.

fractions for separation devices. In addition, screens are commonly employed for the recovery of heavy media suspensions and the separation of water from solid particles. The efficiency of the screening operation is influenced by the material properties of the feed and the operating and design characteristics of the machine, particularly its size.

2.3 Types of Screens

Several types of screens are commonly used in coal preparation. These include the grizzly, vibrating screens and high-frequency screens. A grizzly is a series of evenly spaced parallel bars that are positioned in the direction of material flow. This simple device is used only for gross sizing of very coarse particles (>50 mm) and is generally applied only for the protection of equipment from large oversized material.

Vibrating screens are the most common equipment for sizing particles and are used in almost every aspect of coal preparation. These units utilize an out-of-balance rotating mechanism to create a vibrating motion to sort particles and to move material along the screen surface (called a deck). The three common versions of vibrating screens are the incline screen, horizontal screen, and banana screen. The operating characteristics of these units are compared in Table I. An incline screen uses a combination of gravity flow and machine movement to convey material along the screen by mounting the screen deck 15 to 20 degrees from horizontal. Incline screens are available in single-, double-, and triple-deck arrangements with dimensions up to 3 m (10 ft) wide and 7.3 m (24 ft) long. For a horizontal screen, material is conveyed without the aid of gravity, thus a higher speed and throw are required. These units may be as large as 3 m (10 ft) wide and 7.3 m (24 ft) long and are available in single-, double-, and triple-deck arrangements. Finally, banana screens are designed so that the slope of the screen surface changes along the length of the unit (Fig. 3). The slope at the feed end will typically be 30

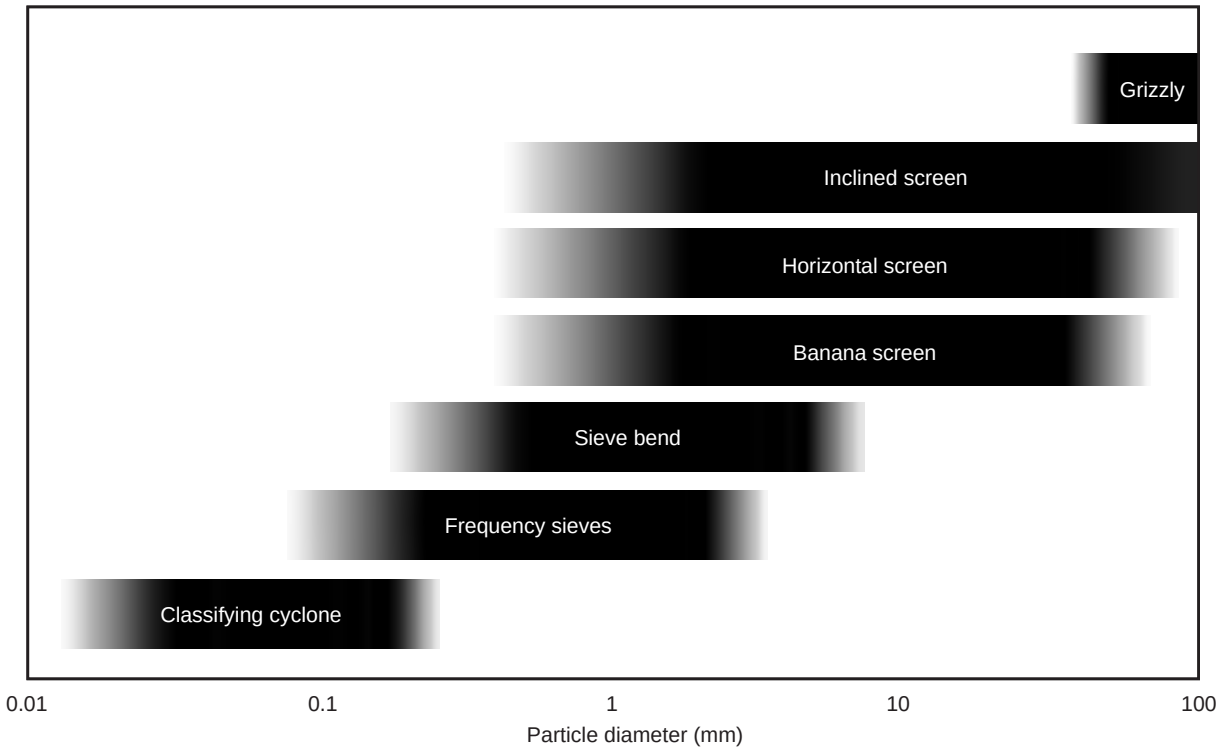


FIGURE 2 Typical ranges of applicability for various sizing equipment.

TABLE I

Typical Operating Characteristics of Vibrating Screens

	Incline	Horizontal	Banana	High frequency
Application	Raw coal sizing scalping	Raw coal sizing prewet drain and rinse dewatering	Raw coal sizing deslime drain and rinse dewatering	Dewatering fines recovery
Incline (degrees)	15–25	Horizontal	Variable	Opposing
Motion	Circular	Linear	Linear	Linear
Speed (rpm)	750–800	930	930	1200
Throw	9.6–11.2 mm (0.38–0.44 inch)	11.2 mm (0.44 inch)	11.2 mm (0.44 inch)	6.4 mm (0.25 inch)
Travel rate	0.5–0.6 m/s (100–120 fpm)	0.2–0.23 m/s (40–45 fpm)	0.9–1.0 m/s (180–200 fpm)	0.13–0.18 m/s (25–35 fpm)

to 35 degrees, while the slope at the discharge end can be 0 to 15 degrees. The steep initial slope enhances the ability of fines to pass since the higher rate of travel reduces the depth of the particle bed. The shallower slope at the discharge end of the screen increases screening efficiency by slowing the travel rate and providing more opportunities for near-size particles to pass. Commercial units are available with dimensions exceeding 3.7 m (12 ft) wide and 7.3 m (24 ft) long in single- and double-deck arrangements. Banana screens are becoming increasingly popular since they offer increased capacity and lower capital and operating costs compared to conventional horizontal and inclined screens.

High-frequency screens are normally used for dewatering fine coal or refuse (Fig. 4). These screens are designed to retain feed slurry between two opposing inclines. The screen motion transports dewatered cake up a gradual slope to the discharge end of the screen. The screen vibrator, which may be an unbalanced rotor or linear exciter, operates at a higher speed and with a shorter thrown than vibrating

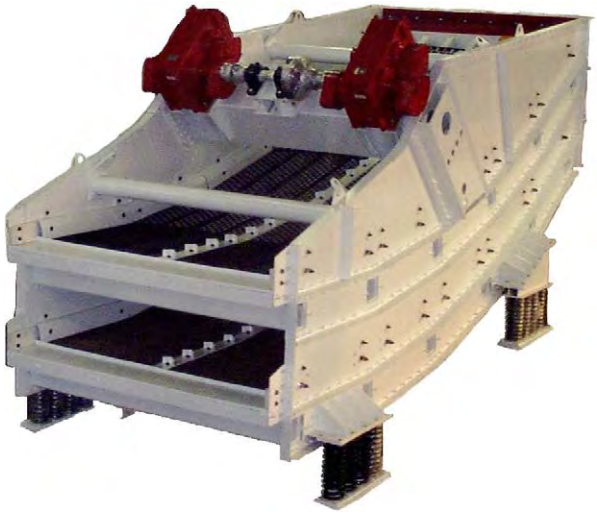


FIGURE 3 Production-scale banana (multislope) screen. Courtesy of Innovative Screen Technologies.

FIGURE 4 Production-scale high frequency screen. Courtesy of Innovative Screen Technologies.

screens. The operating characteristics are beneficial for the efficient dewatering of fine solids. Commercial units are now available in single-deck configurations that exceed 2.4 m (8 ft) wide and 3.7 m (12 ft) long.

2.4 Screen Surfaces

A wide variety of screen surfaces are commercially available for coal sizing. The most common types of media include punch plate, woven wire, rubber, urethane, and profile wire (Fig. 5). The selection of the most appropriate media is dictated by type of

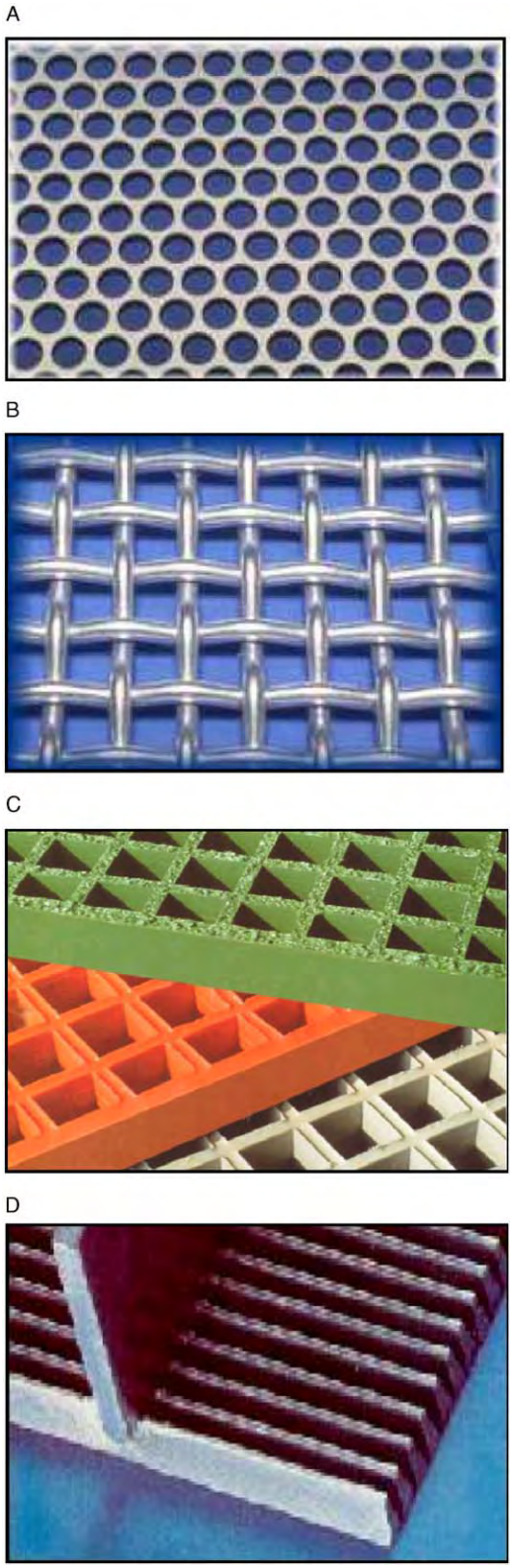


FIGURE 5 Common types of screening media. (A) punch plate, (B) woven wire, (C) polyurethane, and (D) profile wire.

application and considerations associated with wear and maintenance. Punch plate typically consists of steel sheets that are perforated with square, round, or rectangular holes. This type of media is used for primary sizing. Woven wire media consists of fine wires that are woven together to form a mesh pattern with the desired hole size. Also, the woven mesh must be strong and wear resistant (which requires larger wires) and still possess sufficient open area for the efficient passage of particles (which requires smaller wires). The final selection is usually based on an economic tradeoff between replacement frequency and sizing efficiency. Both rubber and urethane media can be manufactured in modular panels that can be easily installed and replaced. These materials offer a longer wear life and are less likely to become plugged than other types of screening media. Urethane panels can be manufactured with openings ranging from 0.15 to 100 mm. The major concern with this type of media is the loss of open area that normally occurs when they are used to replace woven wire media. In cases where fine (3×0.5 mm) sizing is performed, profile wire may be used to reduce problems with plugging and to increase the open area of the screen. The wires in this type of media are constructed with tapered cross sections (usually an inverted triangle or trapezoid) that allow particles to freely pass once they have cleared the top surface of the screen media. This type of screening media improves open area and screen capacity but typically has a shorter wear life.

2.5 Fine Coal Sieves

Sizing of fine coal is difficult due to the increased likelihood of plugging and the large number of particles that must pass through the screen surface. Fine sizing is typically performed either utilizing banana screens or sieve bends in conjunction with a horizontal screens. A sieve bend (Fig. 6) consists of a curved panel constructed from inverted trapezoidal bars (profile or wedge wire) placed at right angles to fluid flow. The capacity of the sieve is relatively high since the leading edge of the trapezoidal bars "slice" material from the flowing stream. Although the screen has no moving parts, passage of solids is also greatly enhanced by the centrifugal force created by the fluid flow along the curved sieve surface. Because of the unique geometry, the size retained by the sieve is approximately three-fourths the bar spacing (e.g., 0.5 mm spacing retains 0.38 mm particles). Sieve bends are commonly used for fine particle dewatering and are often used ahead of vibrating screens. A variation on the sieve bend theme incorporates the use of high-frequency vibration.

FIGURE 6 Production-scale sieve bend for fine coal sizing. Courtesy of Conn-Weld Industries.

2.6 Classification

Classification is used in coal preparation for fine sizing where conventional screening becomes impractical or too expensive. Classification is usually applied to dilute suspensions of <1 mm diameter particles. This technique is generally less efficient than screening or sieving since the separation can be influenced by particle shape and density, as well as particle size. The most common type of classification device used in the coal industry is the classifying cyclone (Fig. 7). In this device, dilute slurry is pumped under pressure to multiple parallel units (banks) of cyclones. The rotating flow field within the cyclone creates a centrifugal force that increases the settling rate of the particles. Larger particles are forced to the wall of the cyclone where they are carried to the discharge port (apex) at the bottom of the cyclone. Finer particles exit out the top of the cyclone (vortex finder). Classifying cyclones are commonly applied to size (cut) at 0.10 to 0.15 mm and represent the only practical option for sizing ultrafine particles (at a cut of 0.045 mm). Smaller

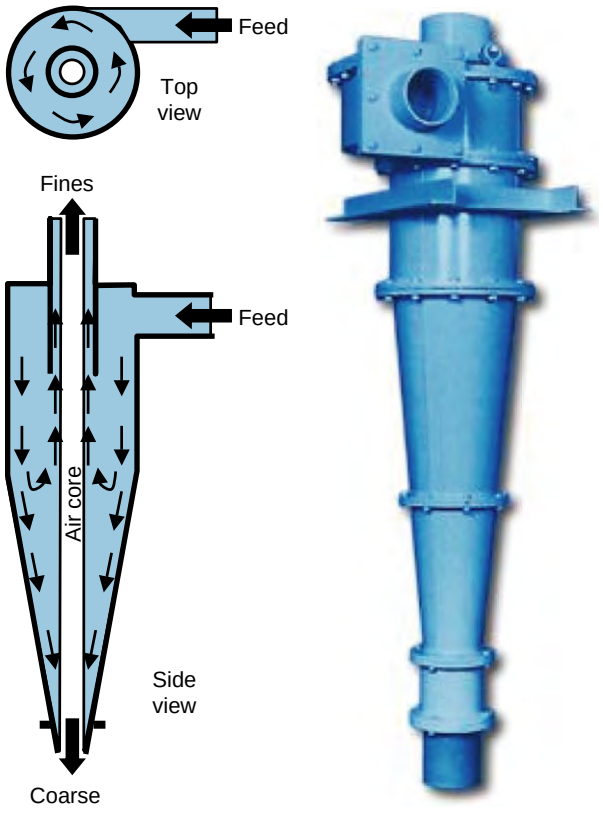


FIGURE 7 Illustration of a classifying cyclone. Courtesy of Krebs Engineers.

cyclones are typically used to size finer particles since these units provide a higher centrifugal force. Classifying cyclones have become very popular in the coal industry since they offer a high unit capacity, operational flexibility, and low maintenance requirements.

3. COAL CLEANING

3.1 Background

One of the primary goals of coal preparation is to separate the valuable carbonaceous material (organic matter) from contaminants in the surrounding host rock (mineral matter). The separation is typically accomplished using low-cost processes that exploit differences in the physical properties that vary according to mineralogical composition. Some of the common properties that are used to separate coal and rock include size, density, and surface wettability. The effectiveness of the separation depends largely on the characteristic properties of the feed material and the operational efficiency of the separator. The effectiveness of most separators is limited to a relatively narrow range of particles sizes (Fig. 8) due to limitations associated with processing efficiency and equipment geometry. Several different types of cleaning processes are incorporated into the plant flowsheet to treat the wide range of particle sizes present in the run-of-mine coal. Density separators, such as heavy media baths and cyclones, are commonly used to treat coarse and intermediate sized particles. Froth flotation, which separates particles based on differences in surface wettability, is the only practical method for treating very fine coal (<0.15 mm). Particles in the intermediate size range between 0.15 mm and 2 to 3 mm may be treated by a variety of processes including water-only cyclones, spirals, teeter bed separators, or combinations of these units. Each of these units is described in greater detail in the following sections.

3.2 Coal Washability

Most of the physical separation processes used in coal preparation take advantage of the differences in the density of coal and rock. Coal preparation terminology commonly interchanges density and specific gravity (a dimensionless parameter defined as the density of a substance divided by the density of water). The specific gravity (SG) for pure coal is approximately 1.3, while the SG for common coal contaminants such as shale and sandstone are in the 2.3 to 2.7 SG range. The density of pyrite ($SG = 4.8$) is significantly higher. Unfortunately, mined coal does not consist of pure coal and pure rock. Instead, middlings material consisting of composite particles of varying composition are also present to a greater or lesser extent. A typical distribution of the composite material is shown in Table II. There is an approximate linear relationship between the density of the composite particle and its ash (mineral matter) content. The SG at which preparation plants operate is adjusted to provide operating levels which satisfy customer quality constraints.

3.3 Coarse Coal Cleaning (Plus 12.5 mm)

One of the oldest forms of coarse coal cleaning is jigging. Jigging is the process of particle stratification due to alternate expansion and compaction of a bed of particles by a vertical pulsating fluid flow. A single jigging vessel can treat large particles (up to 200 mm) at capacities approaching 680 tonnes (750 tons) per hour. Common versions include the baum and batac designs. The baum jig, which is most common, uses a

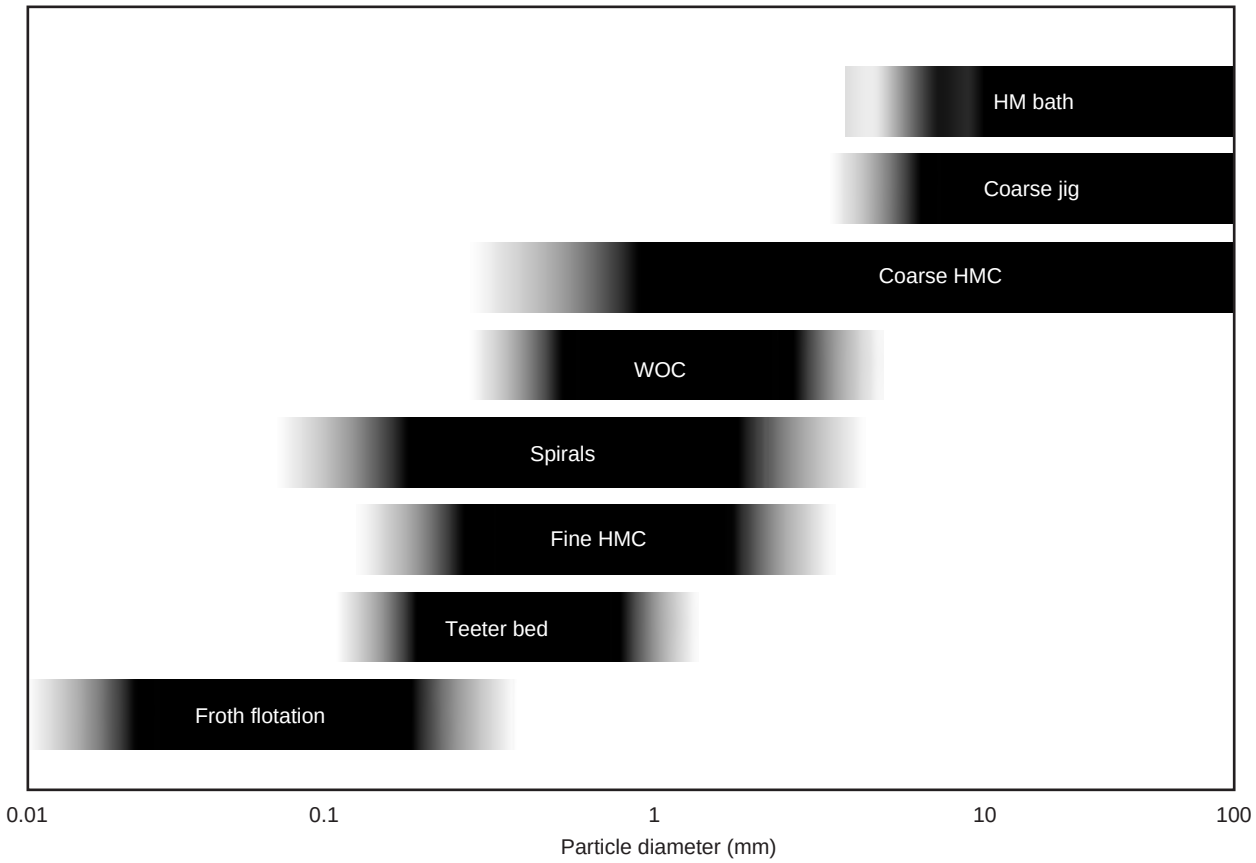


FIGURE 8 Typical ranges of applicability for various cleaning equipment.

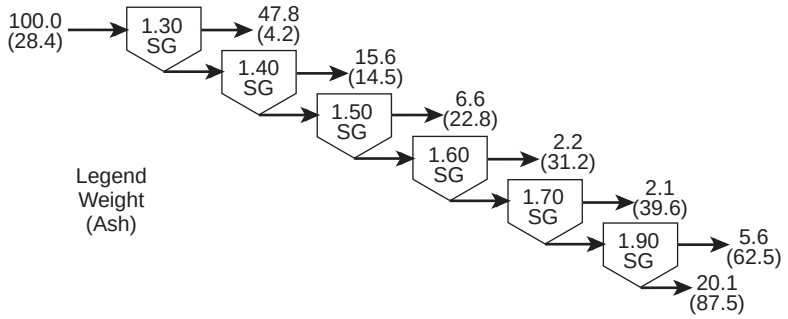
U-shaped vessel that is open on one size and closed on the other (Fig. 9). Water is pulsed through the particle bed atop the perforated plate by sequentially injecting and releasing air in and out of the enclosed chamber. A batac jig is a variation of this design that uses an air chamber below the perforated plant to trap air and create the pulsing action. Jig installations require specific conditions in order to be effective and are significantly less efficient than competitive heavy media processes. As a result, the use of jigs for coal cleaning has declined rapidly. However, jigs can treat a wider size range than single heavy media processes and provide an obvious advantage in applications where capital is limited.

The most common process for upgrading coarse coal (>12.5 mm) is the heavy media bath. A common type of heavy media bath is the Daniels vessel (Fig. 10). This unit consists of a large open vessel through which an aqueous suspension of finely pulverized magnetite (SG = 5.2) is circulated. The magnetite creates a suspension with an artificial density that is between that of pure coal and pure rock. Lighter coal particles introduced into media float to the surface of the vessel where they are transported by the overflowing media to a collection screen. Waste rock, which is much denser, sinks to the bottom of the vessel where it is collected by a series of mechanical scrapers (called flights). The flights travel across the bottom of the vessel and eventually drag the waste rock over a discharge lip at one end of the separator. The unit is highly flexible, since the quality of the clean coal product can be readily adjusted by varying the density of the media from 1.3 to 1.7 SG by controlling the amount of magnetite introduced into the suspension. Sharp separations are possible even in the presence of large amounts of middlings, since SG can be controlled very precisely (as close as 0.005 SG units in some cases) using online instrumentation. A wide variety of different designs of heavy media baths exist, with the major differences being the mechanisms used to discharge the clean coal and waste rock products. Common examples include heavy media cones, drums, and several variations of deep and shallow baths. Drums, usually Wemco or Teska, are widely

TABLE II

Example of Coal Washability Data

Specific gravity		Individual		Cumulative float		Cumulative sink	
Sink	Float	Weight (%)	Ash (%)	Weight (%)	Ash (%)	Weight (%)	Ash (%)
	1.30	47.80	4.20	47.80	4.20	100.00	28.38
1.30	1.40	15.60	14.50	63.40	6.73	52.20	50.52
1.40	1.50	6.60	22.80	70.00	8.25	36.60	65.88
1.50	1.60	2.20	31.20	72.20	8.95	30.00	75.35
1.60	1.70	2.10	39.60	74.30	9.81	27.80	78.85
1.70	1.90	5.60	62.50	79.90	13.51	25.70	82.05
1.90		20.10	87.50	100.00	28.38	20.10	87.50
		100.00	28.38				



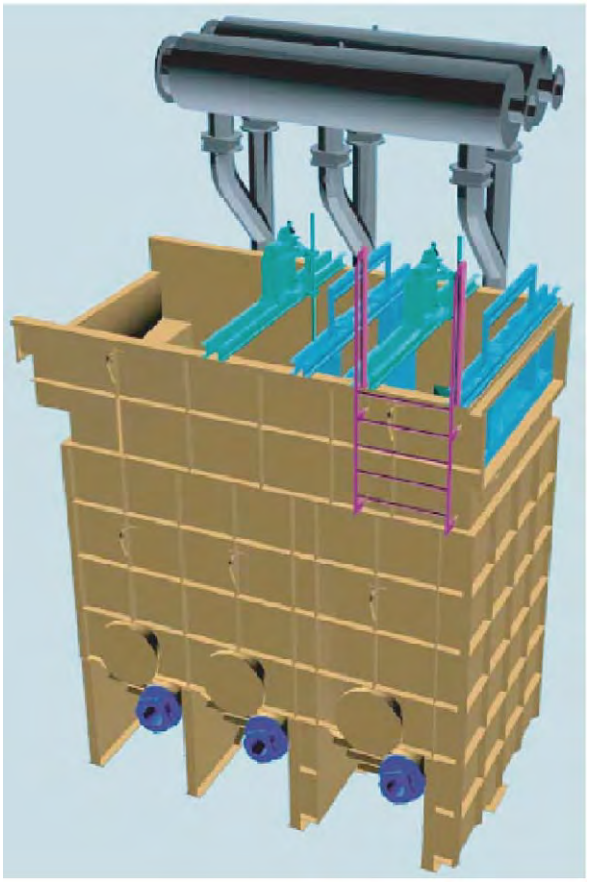
used in South Africa and to a certain extent in Australia. This is where coals are difficult to wash and very high efficiency levels are required.

The primary disadvantage of a heavy media process is the ancillary equipment that must be installed to support the operation of the separator (Fig. 11). A typical circuit requires that the clean coal and refuse products pass over drain-and-rinse screens to wash the media from the surfaces of the products and to dewater the particles. The media that is recovered from the drain section is circulated back to the separator by means of a pump, while media from the rinse section is diluted by spray water and must be concentrated before being fed back to the separator. Fortunately, magnetite is magnetic and can be readily recovered from a dilute suspension using magnetic separators. A rotating drum magnetic separator is commonly used for this purpose (Fig. 12). These devices are highly efficient and generally recover $>99.9\%$ of the magnetite to the magnetic product. For cost reasons, losses of magnetite must be kept as low as possible. Typical losses fall in the range of about 0.45 kg of magnetite loss per tonne (1 lb per ton) of feed coal treated. A portion of the circulating media from the drain section may also be diverted (bled) to the magnetic separator so that the buildup of contaminants such as ultrafine clay can be prevented. Such a build increases the viscosity of the circulating media and adversely impacts the separation efficiency.

Heavy media processes make use of a variety of instrumentation to monitor and control the performance of the separator. In most cases, this consists of a gamma (nuclear) density gauge or other device that continuously monitors the specific gravity of the circulating media. Automated controls are used to add water or magnetite so as to maintain a constant density in the circulating media.

3.4 Intermediate Coal Cleaning (12.5×1 mm)

Heavy media cyclones, which were originally developed by the Dutch State Mines, are the most effective method for treating coal particles in the size range between 1.0 and 12.5 mm. These high-capacity devices (Fig. 13) make use of the same basic principle as heavy media baths—that is, an artificial media is used to separate low-density coal from high-density rock. In this case, however, the rate of separation is greatly increased by the gravitational effect created by passing media and coal through



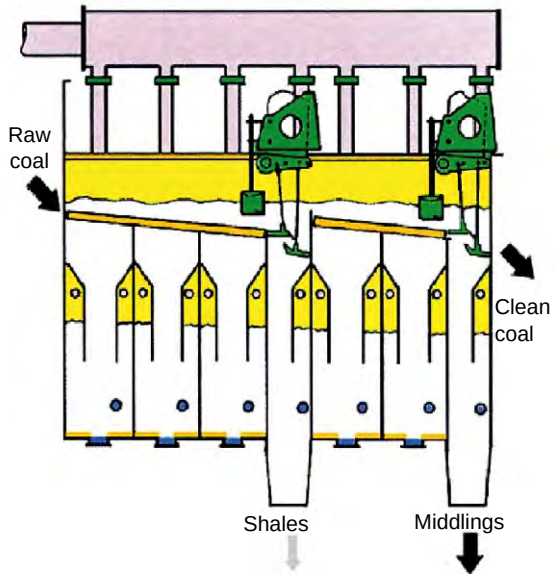


FIGURE 9 Illustration of a baum jig. Courtesy of Bateman.

one or more cyclones. Heavier particles of rock sink in the rotating media and are pushed toward the wall of the cyclone where they are eventually

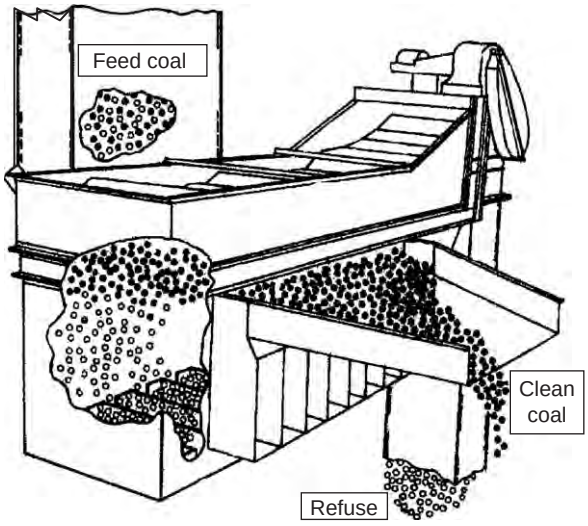


FIGURE 10 Illustration of a heavy media bath. Courtesy of Peters Equipment.

rejected out the bottom. Lighter particles of coal rise to the surface of the rotating media and are carried with the bulk of the media flow out the top of the cyclone. Major advantages of heavy media cyclones include a high separation efficiency, high unit capacity, and ease of control. As with baths, heavy media cyclones have the disadvantage of higher costs associated with the ancillary operations required to support the separator. Several different configurations are used to feed heavy media cyclones. These include gravity flow, pump fed, and wing tanks. Of these, the pump fed configuration (Fig. 14) offers the greatest operational flexibility. The circuit is similar in function to that employed by a heavy media bath.

Advances in cyclone technology and improved wear-resistant materials (ceramic linings) have allowed the upper particle size limit to increase in many applications to 60 to 75 mm. Massive

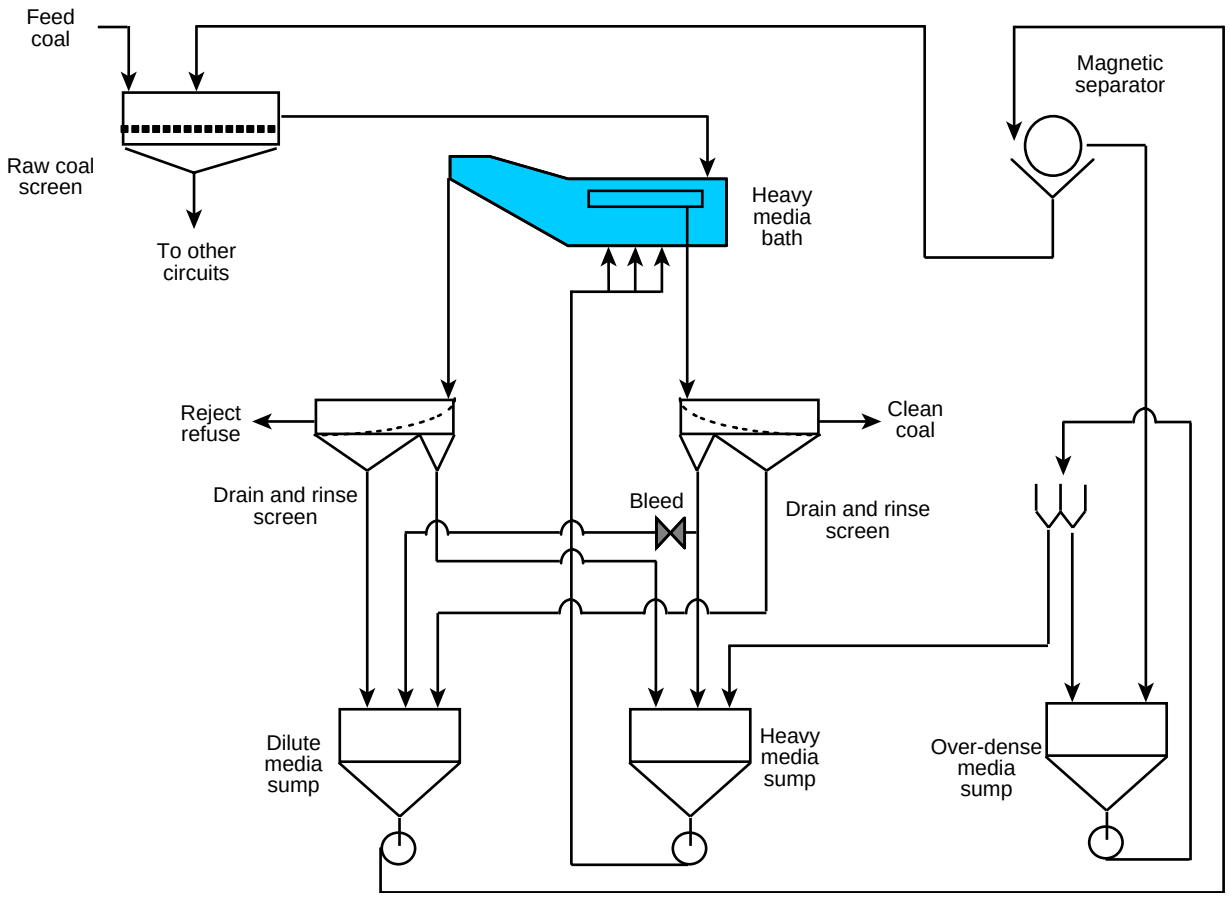


FIGURE 11 Simplified schematic of a heavy media bath circuit.

increases in throughput capacity have also been realized through the use of larger cyclones having diameters of 1.2 m and more. A single unit of this size is capable of treating more than 360 to 410 tonnes per hour (400 to 450 tons per hour) of feed coal. Unfortunately, increased diameter is accompanied with poorer efficiency, particularly in the finer size fractions. The ability to increase both throughput and the upper particle size for cyclones makes it possible to eliminate the heavy media bath in many new plants. This capability is particularly useful for those plant treating very friable coals containing minimal quantities of larger (>50 mm) particles. The simplified circuitry also has significant benefits in terms of capital cost and ease of maintenance. For high reject or hard coals, significant advantages are still present for the bath/cyclone circuit. This results from the reduced necessity for crushing, reduced wear in the plant, and improved total product moisture content. If low density separations are required (<1.4 SG), then the two circuit plant incorporating both baths and cyclones may be necessary since low density separations are easy in baths but difficult in cyclones.

Another option for cleaning coarse coal is a large coal dense media separator, known as a larcodem. This unit was originally developed to handle coal feeds in the size range between 0.6 and 100 mm. The device has the major advantage of being able to treat in one process what was traditionally processed in two separate heavy media devices (heavy medium bath and heavy medium cyclone). Larcodems can also clean far more efficiently what was previously fed to jigging operations. Capital and operating costs for circuits incorporating Larcodems are considerably less that would be the case for traditional heavy medium designs. Despite these advantages, Larcodems have not been widely used in the United States due to lower fine coal cleaning efficiency and the ability of the new generation of cyclones to handle similar top-size material.

FIGURE 12 Illustration of a drum-type magnetic separator. Courtesy of Eriez Manufacturing.

3.5 Fine Coal Cleaning (1 × 0.15 mm)

A wide variety of options are available for upgrading fine coal in the size range between 0.15 and 1 mm. Common methods include water-only cyclones, spirals, and teeter-bed separators. A water-only cyclone (WOC) is a water-based process that employs no extraneous media to separate rock from coal. Water-only cyclones were introduced in the early 1950s for the treatment of fine coal (< 5 mm). Water-only cyclones are similar to classifying cyclones, but typically have a somewhat stubby wide-angled conical bottom and are usually equipped with a long axially mounted vortex finder. Separations occur due to differences in the settling rates of coal and rock in the centrifugal field within the cyclone body. The separation is also enhanced by the formation of autogenous media created by the natural fines already present in the feed slurry. These units are often employed in two stages or in combination with other water-based separators to improve performance. Water-only cyclones do offer some advantages for selected applications including high throughput capacity, operational simplicity, and low operating and maintenance costs.

A separator that has grown tremendously in popularity is the coal spiral (Fig. 15). A spiral separator consists of a corkscrew shaped conduit (normally four turns) with a modified semicircular cross section. The most common units are constructed from polyurethane-coated fiberglass.

FIGURE 13 Illustration of a heavy media cyclone bank. Courtesy of Krebs Engineers.

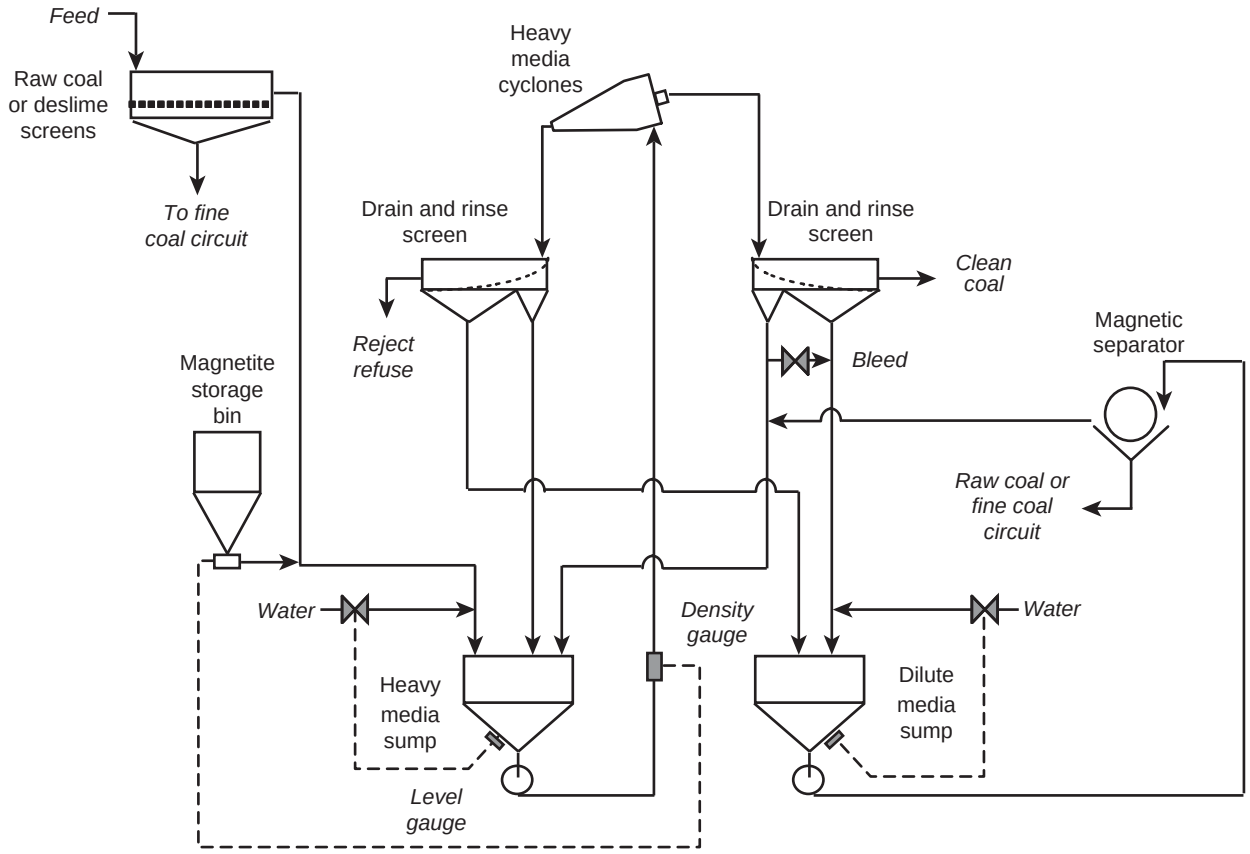


FIGURE 14 Simplified schematic of a pump fed heavy media cyclone circuit.

During operation, feed slurry is introduced to the top of the spiral and is permitted to flow by gravity along the helical path to the bottom of the spiral. Particles in the flowing film are stratified due to the combined effects of differential settling rates, centrifugal forces, and particle-particle contacts. These complex mechanisms force lighter coal particles to the outer wall of the spiral, whereas heavier particles are forced inward to the center of the spiral. The segregated bands of heavy to light materials are collected at the bottom of the spiral. Adjustable diverters (called splitters) are used to control the proportions of particles that report to the various products. A three-product split is usually produced, giving rise to three primary products containing clean coal product, middlings, and refuse. Because of the low unit capacity 1.8 to 3.6 tonnes per hour (2 to 4 tons per hour), spirals are usually arranged in groups fed by an overhead radial distributor. To save space, several spirals (two or three) may be intertwined along a single central axis. In practice, the middlings product should be retreated using a secondary set of spirals to maximize efficiency. The secondary circuit, which is often 10 to 20% of the size of the primary circuit, enables any misplaced material from the primary circuit to be captured. Circuits have been designed to use spirals in treating wider size ranges (e.g., 0.1×3.0 mm), but performance has been poor and efficient spiral cleaning at present would appear to be limited to the 0.15×1.5 mm size range. The main advantages of the spiral circuits are comparatively low capital and operating costs coupled with ease of operation. A major drawback of spiral circuitry is the inability to make a low gravity separation since spirals normally cut in the 1.7 to 2.0 SG range. A consistent volumetric feed rate and minimum oversize (>2 mm) and undersize (<0.1 mm) in spiral feed are also essential requisites to good operation. Spirals have been successfully utilized in combination with water-only cyclones to improve the efficiency of separating fine coal. Latest spiral technologies incorporate primary and secondary cleaning in a single unit. Clean coal and middlings being produced in the upper section are reprocessed in an extended spiral totaling seven turns.

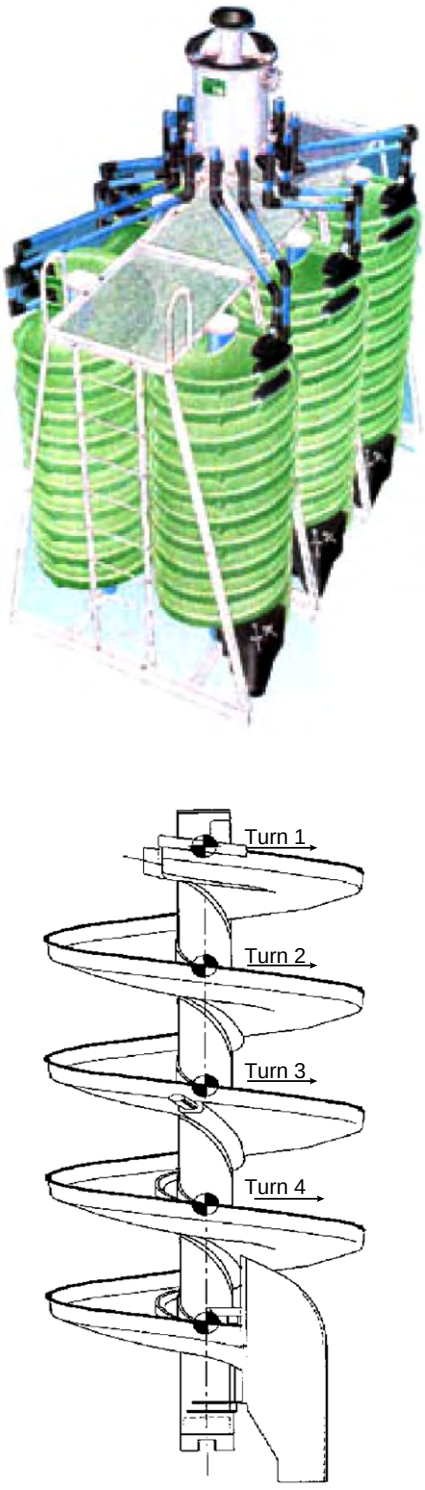


FIGURE 15 Illustration of a coal spiral. Courtesy of Multotec.

Heavy media cycloning of fine coal is practiced, but usually only where a low SG of separation is required. High efficiencies and low impurity levels

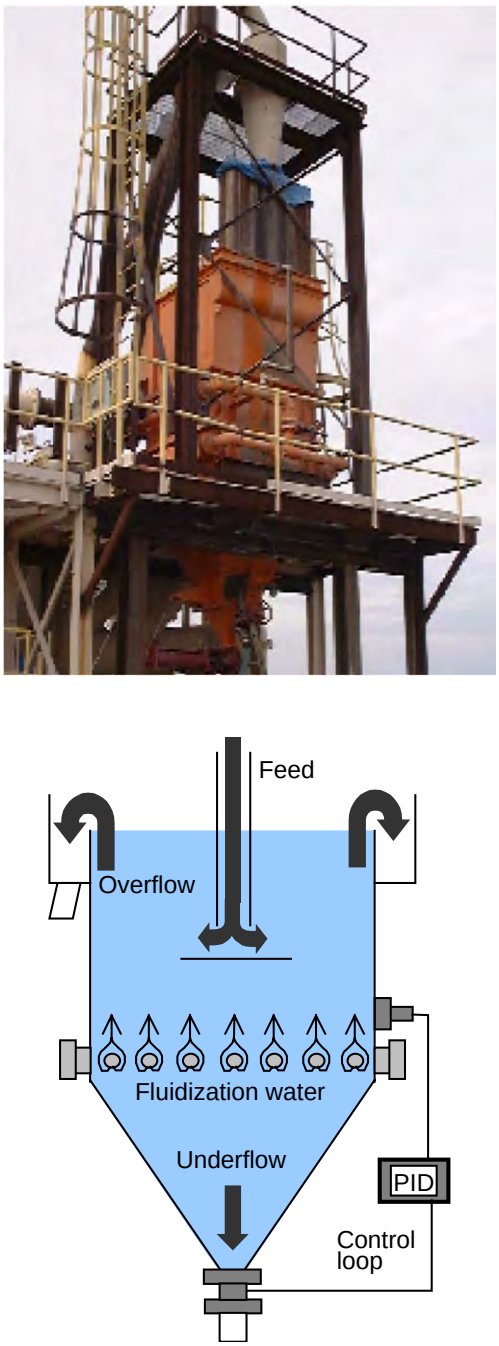


FIGURE 16 Illustration of a teeter-bed separator. Courtesy of Eriez Manufacturing.

(ash) are possible. Operating and capital costs are high.

Teeter bed separators (Fig. 16) are gaining popularity, particularly in Australia. Separation is achieved as a result of the different setting rates of coal and rock in water currents. These units have high capacity and can make lower SG cut points than

spirals. The major unit disadvantage is the limited size range which can be efficiently treated (4:1 top size to bottom size).

3.6 Ultrafine Coal Cleaning (<0.15 mm)

The most widely accepted method for upgrading ultrafine coal is froth flotation. Froth flotation is a physicochemical process that separates particles based on differences in surface wettability. Flotation takes place by passing finely dispersed air bubbles through an aqueous suspension of particles (Fig. 17). A chemical reagent, called a frother, is normally added to promote the formation of small bubbles. Typical addition rates are in the order of 0.05 to 0.25 kg of reagent per tonne (0.1 to 0.5 lb per ton) of coal feed. Coal particles, which are naturally hydrophobic (dislike water), become selectively attached to air bubbles and are carried to the surface of the pulp. These particles are collected from a coal-laden froth bed that forms atop the cell due to the frother addition. Most of the impurities that associate with coal are naturally hydrophilic (like water) and remain suspended until they are discharged as dilute slurry waste. Another chemical additive, called a collector, may be added to improve the adhesion between air bubbles and coal particles. Collectors are commonly hydrocarbon liquids such as diesel fuel or fuel oil. In the United States, flotation is typically performed on only the finest fractions of coal (<0.1 mm), although coarse particle flotation (<0.6 mm) is practiced in Australia where coal floatability is high and the contaminants in the process feed low. The removal of clay slimes (<0.03 mm) is carried out ahead of some flotation circuits to minimize the carryover of this high ash material into the froth product. The ultrafine clay slimes are typically removed using large numbers of small diameter (15 mm or 6 inch) classifying cyclones (Fig. 18).

Most of the industrial installations of flotation make use of mechanical (conventional) flotation machines. These machines consist of a series of agitated tanks (four to six cells) through which fine coal slurry is passed. The agitators are used to ensure that larger particles are kept in suspension and to disperse air that enters down through the rotating shaft assembly. The air is either injected into the cell using a blower or drawn into the cell by the negative pressure created by the rotating impeller. Most commercial units are very similar, although some variations exist in terms of cell geometry and impeller design. Industrial flotation machines are now available with individual cell volumes of 28.3 m^3 (1000 ft^3) or more. Coal flotation is typically performed in a single stage with no attempt to reprocess the reject or concentrate streams. In some cases, particularly where high clay concentrations are present, advanced flotation processes such as column cells have been used with great success. Conventional flotation cells allow a small amount of clay slimes to be recovered with the water that reports to the froth product. A column cell (Fig. 19) virtually eliminates this problem by washing the clay slimes from the froth using a countercurrent flow of wash water. This feature allows columns to produce higher quality concentrate coals at the same coal recovery.

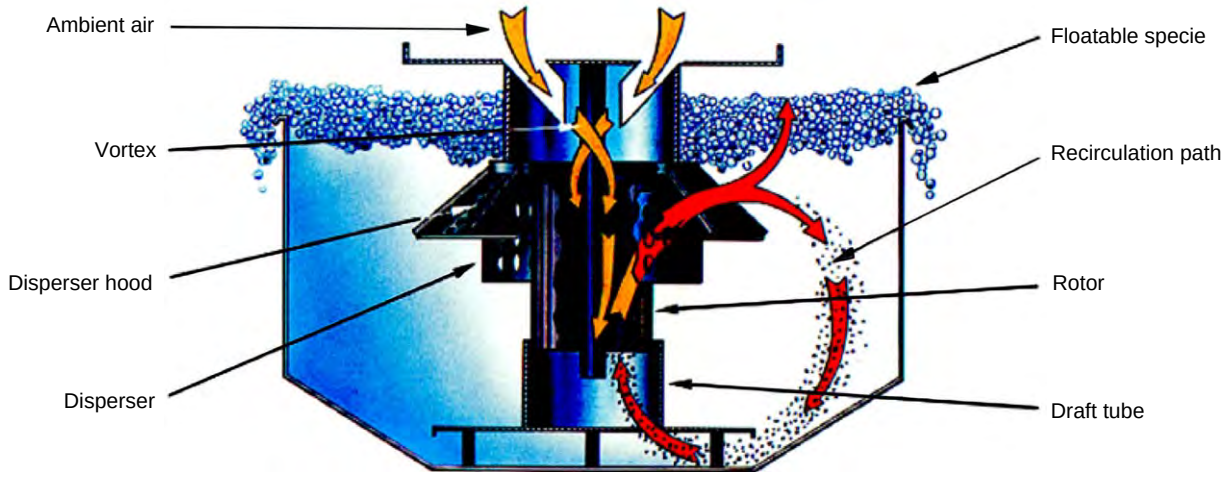


FIGURE 17 Illustration of a conventional flotation cell. Courtesy of Wemco.

FIGURE 18 Bank of small diameter classifying cyclones. Courtesy of Multotec.

4. SOLID-LIQUID SEPARATION

4.1 Background

Practically all of the processes employed in coal preparation use water as the medium. After treatment, the coal must be dewatered. The purpose of dewatering is to remove water from the coal to provide a relatively dry product. A coal product with an excessively high moisture content possesses a lower heating value, is subject to freezing, and is more costly to transport.

4.2 Mechanical Dewatering

A variety of methods are used to mechanically dewater the different sizes of particles produced by coal cleaning operations (Fig. 20). Dewatering of coarse coal is predominantly carried out using screens or centrifugal dryers. For screens, the prime force involved in dewatering is gravity. This is a relatively simple process for particles larger than approximately 5 mm. Dewatering screens have the advantage of being simple, effective, and inexpensive. Disadvantages include blinding and poor performance for smaller particles. Finer particles, which have a much higher surface area, tend to have correspondingly higher moisture content and are typically dewatered by centrifugal methods.

Centrifugal dryers are comprised of a rotating screen or basket, an entrance for feed slurry, a power drive for rotation, and a mechanism for solids removal. Centrifugal dewatering equipment utilizes sedimentation and filtration principles that are amplified by centrifugal forces. Centrifuges can be used to dewater particles from 50 mm to a few microns. However, no single unit can effectively treat this entire size range (Fig. 20). For coarser fractions, two popular units are the vibratory centrifuge and the screen-scroll centrifuge. Both are available in horizontal or vertical configurations. The primary difference in these units is the mode by which solids are transported. The vibrating centrifuge uses a reciprocating motion to induce the flow of solids across the surface of the dewatering basket (Fig. 21a). The screen-scroll centrifuge uses a screw-shaped scroll that rotates at a slightly different speed than the basket to positively transport the solids through the unit (Fig. 21b). The vibratory centrifuge works best for coarser feeds up to 30 to 50 mm. Both the screen-scroll and the vibratory centrifuges are appropriate for intermediate sizes, although the screen-scroll centrifuge typically provides lower product moistures. For finer material (<1 mm), both the screen-scroll centrifuge and another popular design, called a screen-bowl centrifuge, may be used. The screen-bowl centrifuge is a horizontal unit that consists of a solid bowl section and a screen section

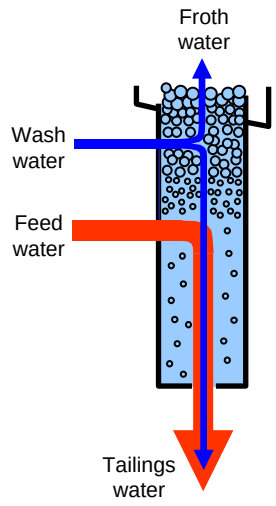


FIGURE 19 Illustration of column flotation technology.

(Fig. 22). Feed slurry is introduced into the bowl section where most of the solids settle out under the influence of the centrifugal field. A rotating scroll transports the settled solids up a beach and across the screen section where additional dewatering occurs before the solids are discharged. Solids that pass through the screen section are typically added back to the machine with the fresh feed. This unit is capable of providing low product moistures, although some ultrafine solids are lost with the main effluent that is discarded from the solid bowl section. This process has become dominant in the United States for drying combined fine and ultrafine coal due to low operating cost and ease of maintenance.

If high coal recovery is desirable, then higher-cost filtration processes may be used to dewater fine coal. Filtration involves the entrapment of fine solids as a cake against a porous filtering media. Traditionally, flotation concentrates have been dewatered using some form of vacuum filtration. These units are capable of maintaining high coal recoveries (>97%) while generating product moisture contents in the 20 to 30% range. One of the most popular types of vacuum filters is the disc filter (Fig. 23). This device consists of a series of circular filtration discs that are constructed from independent wedge shaped segments. At the beginning of a cycle, a segment rotates down into the filter tub filled with fine coal slurry. Vacuum is applied to the segment so that water is drawn through the filter media and a particle cake is formed. The resultant cake continues to dry as the segment rotates out of the slurry and over the top of the machine. The drying cycle ends when the vacuum is removed and the cake is discharged using scrapers and a reverse flow of compressed air.

More recently, horizontal belt vacuum filters have become increasingly popular for dewatering fine coal (particularly in Australia). This device consists of a continuous rubber carrier belt supported between two end rollers (Fig. 24). Filter fabric is placed along the top surface of the belt to serve as filter media. Coal slurry is introduced at one end of the moving belt and is dewatered by the vacuum that is applied along the entire length of the underside. As the belt reaches the end roller, the vacuum is removed and the dewatered cake is discharged. A significant advantage of this design is the ability to clean the filter cloth on the return side using high-pressure sprays. Other advantages include the ability to control the cake thickness (by varying belt speed) and the independent control of cake formation and drying cycle times.

4.3 Thermal Drying

Thermal drying is used when mechanical dewatering is incapable of producing a coal moisture content that is acceptable for market. Although many types of thermal dryers have been tried including indirect heating dryers, fluidized bed dryers have become prominent in drying coals of bituminous rank, often producing product moistures in the 5 to 6% range. Fuel used is most often coal either in a pulverized form or as a stoker product. As far as lower rank coal drying is concerned, numerous thermal drying

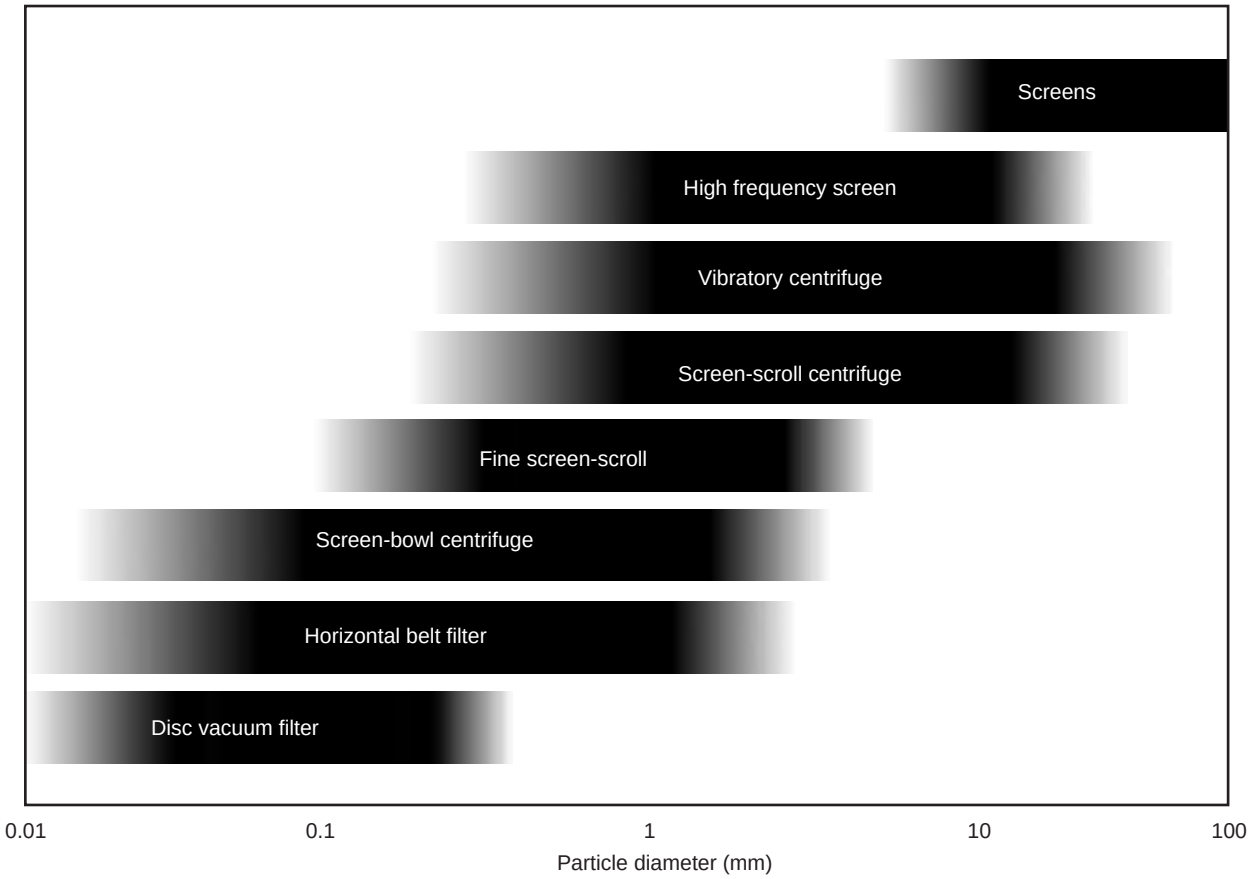


FIGURE 20 Typical ranges of applicability for various dewatering equipment.

processes have been developed, several of which include some degree of surface pyrolysis to reduce the tendency of the coal to reabsorb water or to be subject to spontaneous combustion. Briquetting is also utilized in several processes to deal with the fines inevitably generated in the drying process. Advantages of thermal drying include the ability to achieve very low moistures and the potential to improve other coal properties (e.g., coking). Disadvantages include high operating and capital costs, fugitive dust emissions, and safety hazards associated with potential fires and explosions.

4.4 Thickening

The process of cleaning coal is a wet one, using large quantities of water (i.e., thousands of gallons per minute). As a result, vast quantities of dirty (fine rock-laden) water are produced. Beneficiation processes require clean water. Consequently, dirty water is processed, solids removed, and clean water recycled for further use in the plant. This process is called thickening. A thickener is simply a large tank in which particles (usually flocs) are allowed to settle, thereby producing a clean overflow and a thickened underflow (Fig. 25). Thickeners are typically fed dilute slurry (<5 to 10% solids) that is rejected from the fine coal cleaning circuits. The clarified water that overflows along the circumference of the thickener is recovered and reused as process water. The thickened underflow, which typically contains 20 to 35% solids, is transported to the center of the thickener by a series of rotating rakes. The thickened sludge is typically pumped to an appropriate disposal area or is further dewatered prior to disposal. Chemical additives are usually introduced ahead of the thickener to promote the aggregation of ultrafine particles so that settling rates are increased. The pH levels are also monitored and controlled.

Conventional thickeners, which are typically 15 to 60 m (50 to 200 ft) diameter, require constant monitoring to ensure that overflow clarity and underflow density is maintained and the rake mechanism is not overloaded. A torque sensor on

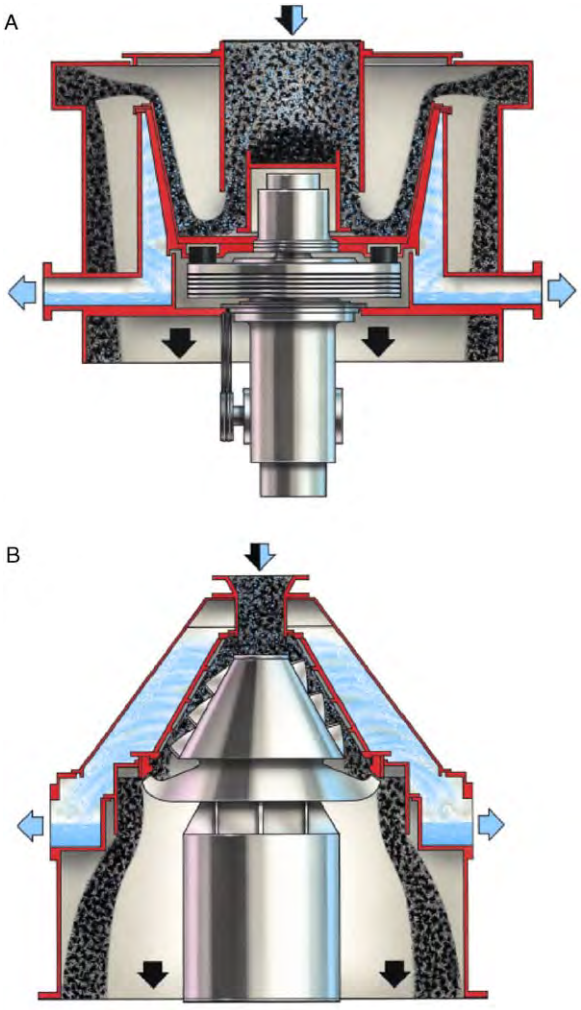


FIGURE 21 Illustrations of (A) vibratory and (B) screen-scroll centrifugal dryers. Courtesy of Centrifugal Machines, Inc.

the central drive is typically used to monitor rake loading, while a nuclear density gauge is used to monitor underflow density.

5. MISCELLANEOUS

5.1 Process Control

A state-of-the-art coal processing facility typically uses a programmable logic controller (PLC) as the heart of the plant control system. The PLC is a computer based controller that acquires data (input) representing the status of all the units and control loops in the plant, makes decisions based on the data input and programmed instructions, and sends data (output) to the plant units reflecting the decisions.

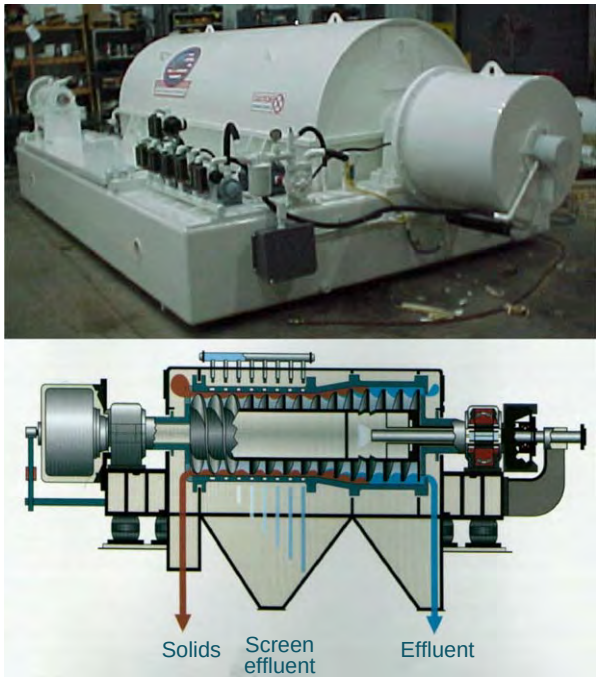


FIGURE 22 Illustration of a screen-bowl centrifuge. Courtesy of Decanter Machine, Inc.

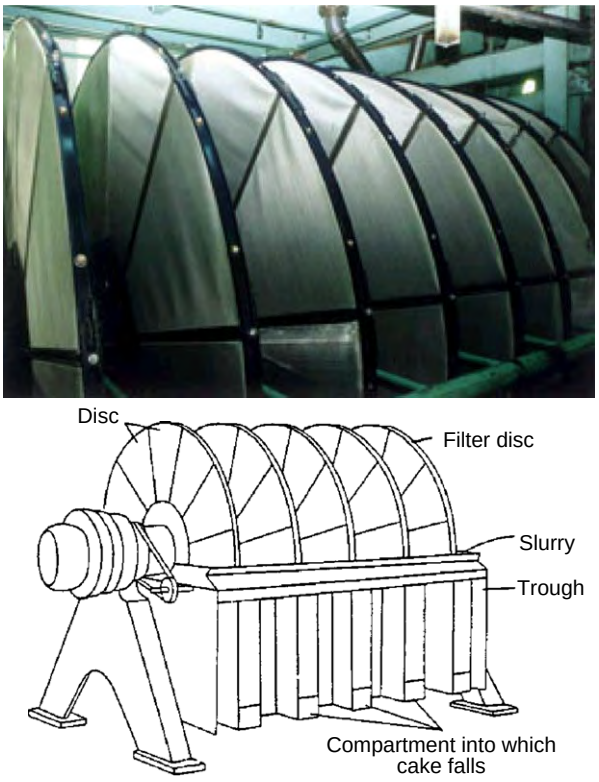


FIGURE 23 Illustration of a conventional disc filter. Courtesy of Peterson Filters.

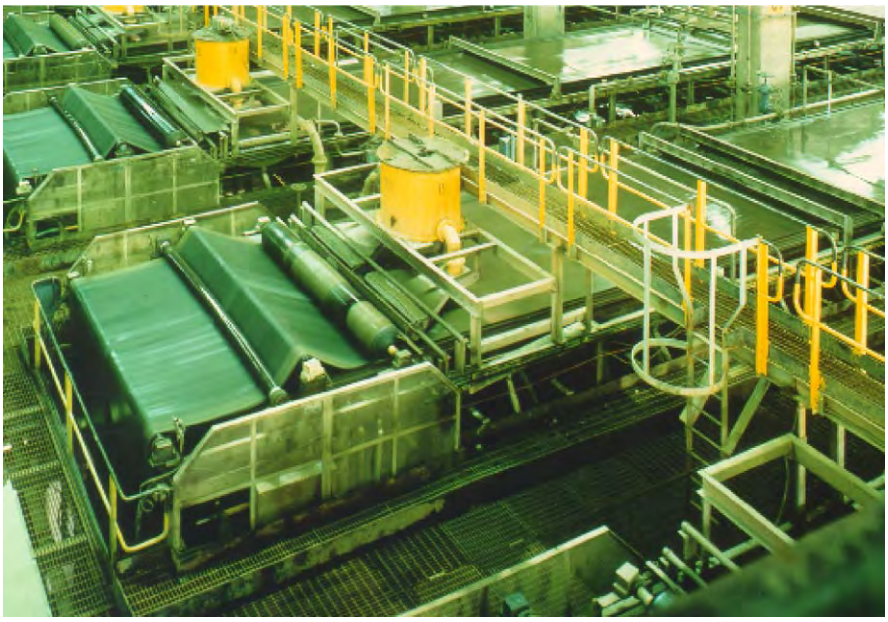


FIGURE 24 Production-scale horizontal belt filter. Courtesy of Delkor Pty, Ltd.

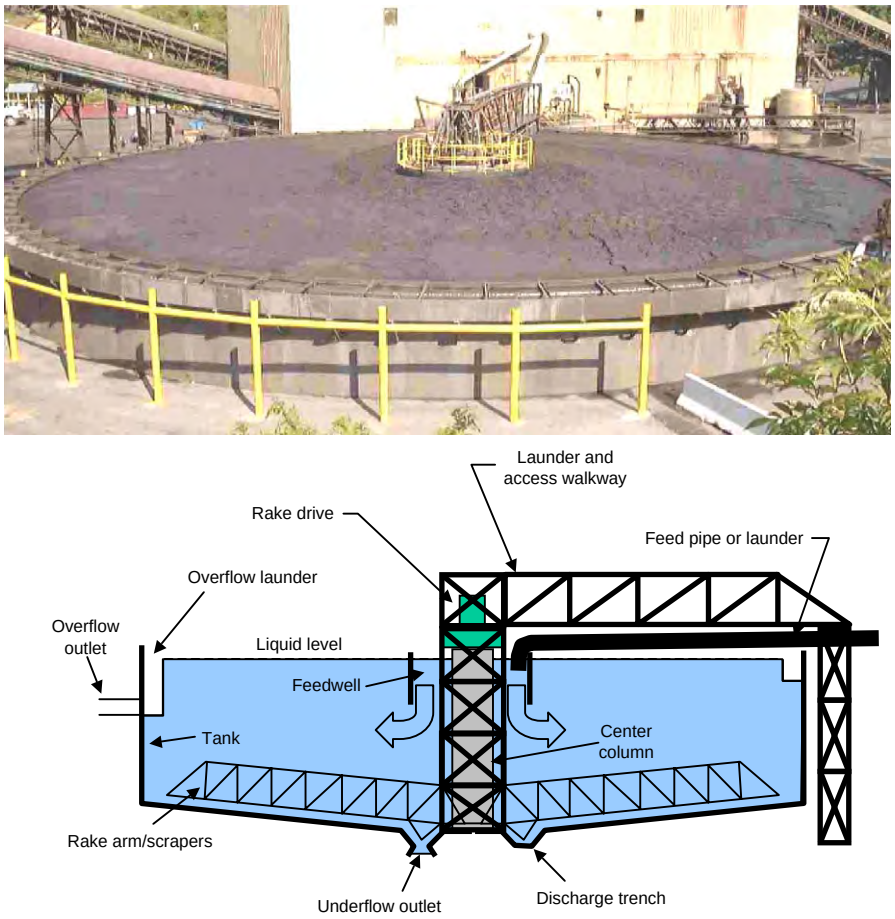


FIGURE 25 Illustration of a refuse thickener.

Inputs include a wide variety of data such as the electrical status of motor control circuits, operating/position status of moving equipment, electrical and mechanical faults, and operating status of control loops (levels, temperatures, flow rates, pressure, etc.). Field devices that provide input to the PLC include electronic transmitters for levels, temperature, pressure, position, and flow rates. Nuclear devices are used for determining online measurements for specific gravity and solids content of slurry streams, for mineral and elemental analysis of process streams, and for flow rates. Digital (microprocessor-based) equipment is becoming available that provides faster communication, less hard wiring, and more reliable data. Digital equipment is also being used to replace standard motor control circuits, which provides better monitoring for those circuits. Outputs include various instructions to field devices such a start/stop signals to motors, open/close or position signals to process equipment (valves, gates, etc.), and signals to the control loops (valves, diverters, feeders, gates, etc.). Field devices that translate PLC output instructions into control actions include standard relay based or digital motor control circuits; electric or electro-pneumatic actuators for valves, gates, and so on; and electronic or digital variable speed controllers for feeders, pumps, conveyors, and so on.

The PLC in conjunction with a full color graphical interface (typically based on a personal computer) provides the operator with all relevant data concerning the operation of the facility. The capabilities of the system includes real time monitoring and control. A major advantage of the PLC system in is the ability to modify the program quickly to change interlocks, startup/showdown procedures. Two other techniques being used to enhance monitoring and control are fiber optics communications and radio frequency telemetry systems. The use of fiber optics for data transmission over long distances provides faster communications and immunity to electrical (lightning) interference. Radio frequency telemetry (direct and satellite relay) provides communications where it is impractical to install long runs of wires and conduits for remote monitoring and control.

During the past decade major enhancements in online analysis of coals have occurred. A variety of ash analyzers are now commercially available for monitoring the quality of the clean coal products produced by a preparation plant. These analyzers are typically mounted on a conveyor belt for the through-the-belt analysis of process streams or at the coal loadouts where the product streams are intercepted. These units can be used to monitor product quality, but more usually in the blending of coals of different qualities. This is done to meet customer quality specifications. They can use the principle of measuring the attenuation of a beam of nuclear radiation as it passes through a process stream. There are also slurry ash analyzers available, including one that uses submersible probes and another than is similar to a nuclear density gauge.

Several commercial monitors are also available of online determination of product moisture content. The principle of measuring the attenuation of a beam of nuclear or microwave radiation as it passes through a process stream is utilized in moisture determination. Accuracy for the microwave analyzers decreases as the variability in the size consists of the process stream increase.

5.2 Waste Disposal

Coal washing processes produce large volumes of waste rock that must be discarded into refuse piles or impoundments. Refuse piles are designed to receive coarse particles of waste rock that can be easily dewatered. This material is relatively easy to handle and can be transported by truck or belt haulage systems to the disposal area. In contrast, impoundments are designed to retain fine waste rock that is difficult to dewater. The waste is normally pumped from a thickener to the impoundment as slurry containing water, coal fines, silt, clay, and other fine mineral particulates. In most cases, the slurry is retained behind a embankment (earthen dam) constructed from compacted refuse material. The impoundment is designed to have a volume that is sufficiently large to ensure that fine particles settle by gravity before the clear water at the surface is recycled back to the plant for reuse. In some cases, chemical additives may be used to promote settling and to control pH. Impoundments, like any body of water contained behind a dam, can pose a serious environmental and safety risk if not properly constructed, monitored, and maintained. Potential problems include structural failures, seepage/piping, overtopping, acid drainage, and black water discharges. In the United States, strict engineering standards are mandated by government agencies to regulate the design and operation of impoundments. In some countries, combined refuse systems are employed in which both coarse and fine refuse are

disposed together as a solid product or alternatively as a pumped slurry.

SEE ALSO THE FOLLOWING ARTICLES

Clean Coal Technology • Coal, Chemical and Physical Properties • Coal Conversion • Coal, Fuel and Non-Fuel Uses • Coal Industry, Energy Policy in • Coal Industry, History of • Coal Mine Reclamation and Remediation • Coal Mining, Design and Methods of • Coal Mining in Appalachia, History of • Coal Resources, Formation of • Coal Storage and Transportation • Markets for Coal • Oil Refining and Products

Further Reading

Bethell, P. J. (1998). "Industrial Practice of Fine Coal Processing" (Klimpel and Luckie, Eds.). Society for Mining, Metallurgy, and Exploration (SME), Littleton, CO, September, pp. 317–329.

Coal Preparation in South Africa. (2002). "South African Coal Processing Society." Natal Witness Commercial Printers (Pty) Ltd., Pietermaritzburg, South Africa.

Laskowski, J. S. (2001). "Coal Flotation and Fine Coal Utilization: Developments in Mineral Processing." Elsevier Science Publishing, Amsterdam, The Netherlands.

Luttrell, G. H., Kohmuench, J. N., Stanley, F. L., and Davis, V. L. (1999). "Technical and Economic Considerations in the Design of Column Flotation Circuits for the Coal Industry." SME Annual Meeting and Exhibit, Denver, Colorado, March, Preprint No. 99–166.

Wood, C. J. (1997). "Coal Preparation Expertise in Australia: In-Plant Issues and the Potential Impact of Broader Applications," Proceedings, Coal Prep '97, Lexington, KY, pp. 179–198.

Coal Resources, Formation of

SHREEKANT B. MALVADKAR, SARAH FORBES, and GILBERT V. MCGURL
U.S. Department of Energy, National Energy Technology Laboratory
Morgantown, West Virginia, United States

1. Origin of Coal
2. Formation of Coal
3. Describing and Classifying Coal
4. Constituents of Coal and Their Impacts
5. Compositional Analysis and Properties of Coal
6. Frontiers of Coal Research

Glossary

ASTM Abbreviation for American Society for Testing and Materials.
coalification Geochemical process by which buried peat is converted into increasingly higher ranking coal.
coal rank American Society for Testing and Materials-adopted coal classification system, accepted in the United States, that includes brown coal, lignite, subbituminous, bituminous, and anthracite as increasingly higher coal ranks.
lithology Characterization of the structural aspects of the coal bed and the coal texture obtained from the macro- and microscopic observation of the coal.
macerals The altered plant remains, classified into three groups: vitrinite, liptinite, and inertinite.
peatification Biochemical conversion of plant material by anaerobic bacterial action.
petrography Systematic quantification of different bands of material in coal and their characteristics by microscopic study.
petrology Study of the origin, composition, and technological application of different bands of material in coal.

Coal is defined by the American Society for Testing and Materials (ASTM) as "brown to black combustible sedimentary rock composed principally of consolidated and chemically altered plant remains." Coal, formed by biological and geological processes over millions of years, is a heterogeneous and nonuniform material containing macromolecular organic compounds, mineral matters, elements, moisture, gases, and oily and tarry material. Its properties are the result of types of plants, geological age, and biological and geological processes to which plants are subjected.

1. ORIGIN OF COAL

An understanding of coal origin and factors influencing coal formation is important because the demand for electricity, a large fraction of which comes from coal-fired power plants, is increasing. Locating coal sources, assessing their extent and quality, and using coal in an environmentally sound manner are aided by such an understanding.

1.1 Influence of Age

The first plants to grow on land made their appearance toward the end of the Silurian period. These were the earliest vascular plants, leafless shoots that transported nutrients and water throughout the plant. Extensive forests existed by the middle of the Devonian period from which the oldest coal comes. The environment was warm and humid, allowing club mosses, scouring rushes, and horsetails to flourish.

1.1.1 Carboniferous Period

During this period, under a worldwide tropical/subtropical climate with long growing seasons and moderate to heavy rainfall, plants extended their geographic coverage and increased in size. A forest could cover 20,000 to 25,000 square miles, and more plant species existed than do today. Ground pines and club mosses were 30 to 75 feet tall. Giant forerunners of today's horsetail and scouring rushes were abundant. Forests were in swampy areas located on low-lying flat land near shallow seas or rivers. Periodic flooding or drought conditions altered the nature of the forests and led to the formation of anthracite or bituminous coal.

1.1.2 Permian Period

A cooler climate and seasonal changes characterized this period. Except in China, deposition leading to formation of coal in the North declined. In the South, important deposits leading to bituminous coal were laid down in South Africa and Australia.

1.1.3 Cretaceous Period

During this period, the climate was so mild that vegetation grew even in the polar regions, leading to the formation of coalfields there. A major shift in vegetation occurred, namely, plants emerged with seeds having protective coverings. Because of its relatively young age, the coal of this period is of low rank.

1.1.4 Tertiary Period

Deposition of plant materials continued during this period. Ashes, willows, and maples were among the common plants. The younger age of these deposits led to lignite formation.

A strong relationship is found between the rank of coal and the geological age of plant deposition. For example, Texas lignite deposits are from the Tertiary Ecocene period, whereas subbituminous coal in the Big Horn basin and eastern Wyoming was formed during the Paleocene epoch and bituminous coal in western Pennsylvania and eastern Ohio came from the Carboniferous period. However, age alone does not determine the rank. For example, brown coal and lignite found near Moscow, Russia, were formed from plant deposits of the Carboniferous period. However, the plants were not buried deeply. Thus, the temperatures and pressures to which the plant materials were subjected were not high enough to form high-ranking coals.

1.2 Influence of Geology

In spite of seeming abundance in various regions of the world, coal beds form only a very small percentage of the total thickness of the overall sedimentary deposits termed "coal measures" in coal-bearing areas. In a simplified description of coal formation, there are two sequential processes: peat formation (peatification) by biological processes and conversion of peat to coal (coalification) by geological processes. Geological activities also play an important role during the peatification phase. The composition, thickness, hardness, and regularity of the strata both below and above coal beds are created by geological processes over millions of years. These characteristics must be considered in assessing coal mineability and its economics and in selecting mining approaches. For example, in underground mining, the weight and strength of the overlying material ("roof") have a significant bearing on the underground support structure needed for safe operation. Similarly, the "seat" rock on which a coal bed lies may become unduly soft or sticky when wet or excessively plastic when under pressure, causing problems in mining operation.

Under some geological conditions, peat-forming material accumulated at or close to the place of plant growth (autochthonous origin). Under different conditions such as along beaches, along river streams, or when flooding occurs, plant material was moved to a different location, where it was deposited (allochthonous origin) possibly on top of existing plants. Coals formed by the allochthonous mechanism are relatively rare and have less commercial value because of their higher mineral contents.

During peatification, lasting 10,000 to 15,000 years, biochemical activities predominated. Plants and partially decomposed plants were subjected to varying environments (e.g., flooding, drought), addition of plants from other locations, and so on. The extent to which plants were buried under water had a large effect on the type of biochemical activity (e.g., anaerobic vs aerobic fermentation). Some geological changes destroyed swamps and halted the formation and accumulation of peat. Long periods of drought promoted aerobic action, leading to inferior coalfields.

The aerobic process, if it proceeds unimpeded, destroys plant material completely. The anaerobic process leads to peat, and eventually to various ranks of coal, depending on the prevailing geological conditions. If lightning strikes during peatification, fires set by it lead to the formation of charcoal. The charcoal may be incorporated in the coal to a lithotype form of coal called fusain, also termed mineral charcoal or mother-of-coal. An alternate theory for fusain formation also has a geological basis: a drier environment is claimed to lead to the formation of fusain. Cannel coal is theorized to be formed in aerated water, which decomposes all but the spores and pollen.

In stagnant water, the concentrations of acids increase as microbial activities proceed. These acids tend to kill bacteria and short-circuit peatification. In rushing water, this mechanism is less important because acids formed by microbial activities are removed or diluted. Water itself acts as a shield against attack by atmospheric oxygen. Thus, a freshwater environment is most favorable for coal formation.

Coalification primarily involves the cumulative effects of temperature and pressure over millions of years. The sediment covering the peat provides the pressure and insulation, and heat comes from the earth's interior. (Temperature increases ~4–8°C each 100 m.) The change from bituminous coal to anthracite coal involves significantly higher pressure (as in mountain building activity under tectonic plate movements) and higher temperatures than does the formation of bituminous coal. Pressure affects the physical properties of coal (e.g., hardness, strength, porosity). Higher temperatures can be the result of several factors: (1) relative proximity of the earth's magma to formative coal seams, (2) close approach, or contact with, igneous intrusions/extrusions, (3) volcanic activity, and (4) depth of burial. Tectonic plate movement itself creates additional heat. Volcanic activity near the peat or partially formed coal can have a negative impact such as sulfur, mercury, and other environmentally undesirable mineral matters being incorporated into the coal bed. Rapid coalification may lead to parts of coalfields to have natural coke.

The kinds of sediments surrounding the coals are also important. For example, clays do not resist pressure well and deform relatively easily. Thus, a peat deposit in contact with clays reflects the results of pressure more. Although rare, geological processes introduce in some coal beds relatively narrow, vertical, or steeply slanting veins of clay, sometimes partially pyritized.

Several geological phenomena have an impact on the coal bed placement underground:

- *Partings*. Partings are formed when plants buried in a swamp get covered with layers of mud, silt, or clay material followed by additional plant accumulation trapping nonplant materials in between two coal seams. Often, the partings are not uniform and, instead, form a wedge shape that is progressively thicker toward the source of deposited material. Partings impair the mining procedures and, especially if pyritized, decrease the quality of the mineable coal.
- *Concretions*. Formations, perhaps several feet in diameter, that impede mining operations are formed during the overlaying of nonplant material (lenticular masses composed of pyrite, calcite, or siderite or their combinations).
- *Heterogeneity*. During coal bed formation, valleys are formed by the downward erosion caused by nearby streams, occasionally leading to exposure of existing coal beds during their formative stages. The beds are then covered with additional plants or a different type of sedimentation.
- *Geological stresses*. These stresses induce folding and sharp deformation and lead to specific dislocation ("faults"). Such dislocations may range up to hundreds of feet vertically and thousands of feet laterally.

Throughout coalification, formative coal beds are subjected to the action of ground waters. Most of the mineral constituents in the coal bed, such as pyretic material and gypsum, are brought in by ground water either as suspension or by ion exchange. During the bituminous coal stage, the beds develop vertical or nearly vertical fractures that are often filled with coatings of pyrite, calcite, kaolinite, or other minerals deposited by circulating ground water. Similar deposition occurs during peatification, but with a difference. Peat tends to be highly porous and acts as a filter. Secondly, living organisms within peat interact with dissolved minerals such as sodium, potassium, and calcium to exchange ions.

1.3 Influence of Microbes

Microbial activity leads to peat formation (peatification). Microbial activity continues for many years even after the geochemical phase gets under way so long as pressure and temperature are not excessive.

Microbial activity can be of fungal, aerobic, or anaerobic type. The first two need oxygen and cause complete plant decay. If the plant is submerged in water, anaerobic activity dominates. Decomposition of plant material by microbes can be classified as shown in Table I. For the formation of a commercially viable product (coal), peatification or putrefaction is required. Attacks of microbes, oxygen, dissolved oxygen, enzymes, and water-borne chemicals cause degradation of plant material, forming new organic macromolecules so that the original plant structure is not recognizable. Methane and carbon dioxide are also generated under these conditions, hence the term "marsh gas" for methane.

Different parts of plant material have different levels of resistance to microbial activities. Also, their products have different levels of toxicity. For example, cellulose, a major constituent of cell walls, and protein are more susceptible to microbial attacks than are plant constituents such as lignin, tanins, fats, waxes, and resins. Humic acids and other humic substances are important peatification products. Woody material passes through a plastic gel-like

TABLE I

Decomposition of Plant Material via Microbial Actions

Type of action	Type of product
Disintegration (aerated)	Gases, liquids, limited resins
Mouldering (aerobic bacteria and fungi)	Inertinite
Peatification (anaerobic/aerobic bacteria and fungi)	Vitrinite
Putrefaction (anaerobic bacteria)	Exinite

Source. Mehta, A. (Project Manager) (1986). "Effects of Coal Quality on Power Plant Performance and Costs," Vol. 4: "Review of Coal Science Fundamentals," (EPRI CS-283). Electric Power Research Institute, Palo Alto, CA.

stage before solidifying to huminite and eventually vitrinite. Formation of other maceral structures, such as liptinite and inertinite, is also the result of microbial activities.

2. FORMATION OF COAL

2.1 Petrology, Petrography, and Lithotypes

Careful visual examination of a coal sample shows that it is layered; that is, it consists of bands of different materials with distinct entities distinguishable by optical characteristics. Coal petrology is the study of the origin, composition, and technological application of these materials, whereas coal petrography involves systematic quantification of their amounts and characteristics by microscopic study. Lithological characterization of coal pertains to (1) the structural aspects of the coal bed and (2) the coal texture obtained from the macro- and microscopic observation of the coal.

The texture of the coal is determined by the type, grain, and distribution of its macro- and microscopic components. Banded coals composed of relatively coarse, highly lustrous vitrain lenses (6 mm or more in thickness) are labeled as coarsely textured. As the thickness narrows, the texture becomes fine-banded and then microbanded. Coals with homogenous texture are rare and regarded as nonbanded or canneloid coals. The lithotype nomenclature is often not suitable for lignite due to the presence of plant pieces.

Some geologists consider coal to be a rock that has microscopically identifiable altered plant remains and minerals. The altered plant remains are called macerals, which are classified into three groups: vitrinite, liptinite, and inertinite. Vitrinite is derived from coalified plant tissues and gels. Liptinite is formed from plants' protective coatings, reproductive organs, and resinous cell fillings. Inertinite is produced from oxidation of vitrinite and liptinite precursors. Lithotype classification is partly independent of coal rank. In spite of the similarity of naming conventions, macerals are not minerals because they do not have definite chemical compositions and are not crystalline.

The quantities of various types of macerals in coal can be estimated petrographically. Bulk chemical data can be used to assess the influence of macerals on coal composition. Petrographic analysis involves the use of a microscope to measure optical properties. A widely accepted method is called vitrinite reflectance.

2.1.1 The Stopes–Heerlen System

The Stopes–Heerlen system has been codified by the International Committee for Coal Petrology. In this system, coal is classified either as hard coal or brown/lignite coal on the basis of visual appearance. The hard coals are further divided into banded/humic coals and saporpolic coal (nonbanded coal with large concentrations of algae or spores). The coal types are further divided into four lithotypes based on visual observation of features greater than 3 mm:

- *Vitrain:* bright, black, brittle
- *Clarain:* semi-bright, black, finely stratified
- *Durain:* dull, black–gray, black, hard, rough
- *Fusain:* charcoal-like, silky, black, fibrous, friable

The lithotypes are further divided into microlithotypes, which must be at least 50 microns wide and contain at least 5% of the minor components. The combination of individual macerals in a particular coal sample identifies it and determines whether it is mono-, bi-, or trimaceral. Table II provides a list of the combinations of macerals found in coal and the names of their resultant hybrids. The optical classification of brown and lignite coals has recently been standardized along lines similar to those used for hard coals.

2.1.2 Impact of Lithotype on Utilization Potential

Lithotype descriptions are primarily used to characterize coal beds in geological investigations. This includes the study of lateral variabilities in thickness and structure. They are used as the principal basis for coal resource and reserve determination, but not to predict specific coal properties such as ash content and coal rank.

A wide range of petrographic applications have been developed to assess coals for their specific use,

TABLE II
Characteristics of Hard Coal

Group maceral	Maceral	Submaceral	Maceral variety	Cryptomaceral	Occurrence (percentage)	Reflectance class
Vitrinite	Telinite	Telinite 1	Cordeltetelinite		75–98	T0–T21
			Fungotelinite			
			Xylotelinite			
		Telinite 2	Lepidophytotelinite			
			Sigillariotelinite			
	Collinite	Telocollinite		Cryptotelenite		C0–C21
		Gelocellinite		Cryptocorpocollinite		
		Desmocollinite		Cryptogelocollinite		
		Corpocollinite		Cryptovitrodetrinite		
	Vitrodetrinite					
Exinite or liptinite	Sporinite		Tenuisporinite		0–10	St0–St15
			Crassisporinite	Cryptoexosporinite		
			Microsporinite	Cryptointosporinite		
			Macrosporinite			
	Cutinite				0–1	Ct0–Ct15
	Resinite				0–5	R0–R15
	Suberinite					
	Alginite				0–1	At0–At15
	Liptodetrinite					
	Fluorinite					
	Bituminite					
	Exudatinite					
Inertinite	Micrinite				0–10	M0–M18
	Macrinite				0–10	
	Semifusinite				0–20	SF00–SF21
	Fusinite	Pyrofusinite			0.1–10	
		Degradofusinite				
	Sclerotinite			Plectenchyminite	0–1	
				Corposclerotinite		
				Pseudocorposclerotinite		
	Inertodetrinite				Common	

Sources. Mehta, A. (Project Manager), (1986). "Effects of Coal Quality on Power Plant Performance and Costs," Vol. 4: "Review of Coal Science Fundamentals," (EPRI CS-283). Electric Power Research Institute, Palo Alto, CA; Simon, J. A. Coal. *In* "Encyclopedia of Energy," 2nd ed. McGraw–Hill, New York; Vorres, K. S. Coal. *In* "Encyclopedia of Chemical Technology", (J. Kroschwitz, Ed.), 4th ed. John Wiley and Sons, New York.

predict potential problems in their application, and solve problems when they are encountered. Petrographic data are used for determination of seam correlations, geological history, and ease of coal oxidation as well as for measurements of rank, interpretation of petroleum formation from coal deposits, and prediction of coke properties. For example, constituents of seams can be measured over considerable distances apart, and these measurements can then be used to establish profiles for entire seams or coal basins. The vitrinite reflectance histogram can provide a quantitative measure of the ease of coal burning.

A vital role is played by petrographic analyses of various coals under consideration as potential feed material to form metallurgical coke. The coking behavior of the coal depends on the rank, maceral constituents, and their relative proportions. Sometimes, as many as eight different coals have to be

blended, maintaining a balance between "reactive" and "inert" maceral constituents, to attain the desired coking properties. Vitrinite has good plasticity and swelling properties and produces an excellent coke, whereas inertinite is nearly inert and does not soften on heating. Exinite becomes extremely plastic and is nearly completely distilled as asphaltenes (tar). By carefully controlling the blends of coals based on their petrological compositions, behavior during carbonization can be controlled.

Petrographic analyses are also useful for the evaluation of coals for liquefaction and gasification. Bituminous and lower ranking coals that are rich in exinite and vitrinite macerals are desirable for liquefaction. Based on microscopic analysis of dispersed "coaly" materials in shales and other sedimentary rocks, geologists can identify favorable areas for oil and gas exploration.

Extensive work has been carried out to correlate elemental composition, functional forms in which elements are present, and physical properties of coal with maceral forms and their proportions in the coal.

2.2 Factors Influencing Formation of Coal

The net result of peatification and coalification is reduction in hydrogen, oxygen, and free moisture and a corresponding increase in fixed carbon content of the original plant material.

2.2.1 Components of Plants

A perspective of the biochemical degradation leads to three plant components:

a. Structural components meant for the plant's rigidity and form include cellulose (polymerized glucose), hemicellulose (polymerized xylose, or wood sugar), lignin (polymerized *n*-propyl benzene), and calcium pectate.

b. Stored food includes carbohydrates, starch, fats and oil, and proteins. Carbohydrates tend to decompose earlier during peatification.

c. Specialty materials are meant to provide various services to the plant. These materials, functioning to protect the plant, include resin, cutin (protection of leaf surfaces), waxes (moisture retention), and sporopollenin (protective outer layer of spores and pollen). They are relatively resistant to biochemical degradation.

2.2.2 Biochemical Phase (Peatification)

Although bacteria and fungi attack all parts of plants simultaneously, the first ones to succumb are carbohydrates, starch, and calcium pectate. This is followed by decomposition of hemicellulose and then decomposition of cellulose. Lignin is more resistant, specialty materials are even more so, and sporopollenin is the most resistant.

Brackish or saline water often contains enough sulfur to support bacterial activities that lead to conversion of sulfate ions to hydrogen sulfide or elemental sulfur. These two by-products, instead of escaping formative peat, may be incorporated into preexisting pyrites or into the organic structures in the peat.

As peatification proceeds, more plants may be dumped on top of the previously buried plant material, compacting the peat-like material underneath, squeezing out some of the trapped water, and increasing its density. Degradation of cellulose under bacterial attack opens up cell walls and helps to disintegrate the large pieces of plant tissue. The structure of coal itself changes drastically during this phase, increasing carbon content from 50 to 60%.

2.2.3 Geochemical Phase (Coalification)

The geochemical phase begins when sedimentation such as silt, mud, sand, and clay is placed on top of the peat or partially formed peat. The weight of the sedimentation compresses the peat and buries it deeper as more sedimentation is piled up.

Transformation of peat into brown coal, lignite, subbituminous, bituminous, and (finally) anthracite is a sequential and irreversible process. The primary factors that determine coalification progress are temperature, pressure, and duration of coalification to which peat is subjected. With a shorter time, coalification does not proceed beyond the rank of brown coal or lignite. Incomplete coalification produces coal that has recognizable plant fragments.

Higher temperatures are critical for the metamorphosis of peat to increasingly higher ranks of coal. The coalification of brown coals and lignite occurs at temperatures less than 100°C. These coals retain much water and volatile matter and more of the functional groups of the original plant. Temperatures in the range of 100 to 150°C yield bituminous coals through the loss of moisture, aliphatic and oxygenated groups, and hydrogen. In the process, there is a decrease in the pore structure and an increase in the carbon and heat content of the coal. Low- and medium-volatile coals are produced by further exposure to temperatures in the range of 150 to 180°C. Anthracites are produced at temperatures greater than 180°C, with further loss of hydrogen and organic functional groups. Loss of hydrogen during this stage actually decreases slightly the

heating value of coal. Anthracite formation also increases the fine pore structure of the coal. Anomalous coal of higher ranking has been produced during relatively short coalification periods when peat is subjected to magma intruding near it.

2.3 Structure of Coal

Because coal is the result of the partial decay and deposition of vegetation, it consists mostly of organic material. In addition, inorganic matter is incorporated in the coal bed, some of which becomes an integral part of the organic material in coal. Coal also contains free water (inherent moisture). The percentage of inherent moisture ranges from 40 to 50% in lignite to less than 5% in anthracite. Coal composition and structure vary substantially as the rank changes from lignite to anthracite. As rank increases, fixed carbon increases and oxygen and inherent moisture/volatile matter decrease.

In general, rank and maceral composition are viewed as substantially fixed for a particular coalfield, certainly within a particular coal seam. The structure of the coal seam/bed is itself highly irregular, depending on its geological history. Thus, the height of the seam and its horizontal spread are often unpredictable.

2.3.1 Bonding within a Macromolecule

A variety of macromolecules with different molar masses, proportions of functional groups, and structural features exist in any coal piece. These molecules consist of planar fused ring clusters linked to nonpolar hydroaromatic structures. The overall structure is complex, irregular, and open. Entanglement between macromolecules occurs, as does cross-linking of hydrogen-bonded species. These different molecular shapes and sizes in coal lead to irregularities in packing and porosity in coal and make it amorphous. In the extracted material of the coal, the average molar mass is 1000 to 3000 for 5 to 50% fraction. Unextractable coal contains even larger macromolecules. The magnetic resonance technique was applied to coals of various ranks to generate the intramolecular functional group data given in Table III.

The structure of vitrinites, the major maceral groups in bituminous and anthracite coals, has been analyzed and found to have the following characteristics:

a. The organic molecules contain a number of small aromatic clusters, with each typically having one to four fused benzene rings. The average number of clusters depends on the rank, increasing slowly up to 90% fixed carbon and then rapidly for higher carbon content. (Lignites typically have 0.60 aromaticity, whereas anthracite has 0.95.)

b. The aromatic cluster linkage is, in part, by hydroaromatic, alicyclic, and ring structures, which also contain six carbons. On dehydrogenation, these ring structures can become part of the cluster through molecular rearrangement, increasing the average cluster size.

TABLE III

Carbon Structural Distribution: Fraction of Carbon Type

	Lignite	Subbituminous	High-volatile bituminous	Medium-volatile bituminous	Low-volatile bituminous
sp^2 hybridized carbon	0.61	0.63	0.72	0.81	0.86
sp^2 hybridized carbon in aromatic ring	0.54	0.55	0.72	0.81	0.86
Carbon in carbonyl	0.07	0.08	0.00	0.00	0.00
Protonated aromatic	0.26	0.17	0.27	0.28	0.33
Nonprotonated aromatic	0.28	0.38	0.45	0.53	0.53
Phenolic or phenolic ether carbon	0.06	0.08	0.06	0.04	0.02
Alkylated aromatic carbon	0.13	0.14	0.18	0.20	0.17
Aromatic bridgehead carbon	0.09	0.16	0.22	0.29	0.34
sp^3 hybridized carbon	0.39	0.37	0.28	0.19	0.14
CH or CH_2 aliphatic carbon	0.25	0.27	0.16	0.09	0.08
CH_3 aliphatic carbon	0.14	0.10	0.12	0.10	0.06
Oxygen-bound aliphatic carbon	0.12	0.10	0.04	0.02	0.01

Source. Vorres, K. S. Coal. *In* "Encyclopedia of Chemical Technology" (J. Kroschwitz, Ed.), 4th ed. John Wiley and Sons, New York.

c. Clusters can also form linkages employing short groups such as methylene, ethylene, and ether oxygen.

d. A large fraction of hydrogen sites on the aromatic and hydroaromatic rings are substituted by methyl or larger aliphatic groups.

e. Oxygen is usually present in phenolic and ether groups, substituting for hydrogen on the aromatic ring.

f. In a small fraction of the aromatic and hydroaromatic rings, oxygen replaces carbon in the ring, making the ring heterocyclic. (Some of these rings are five-membered.)

g. Compared to oxygen in phenolic groups, a lesser fraction of oxygen appears in the carbonyl structure (quinone form on aromatic and ketone form on hydroaromatic rings). In both cases, oxygen is apparently hydrogen bonded to adjacent hydroxyl groups.

h. Nitrogen, which is less abundant in vitrinite than is oxygen, is usually present as a heteroatom in a ring structure or as a nitrile.

i. Most of the sulfur, especially if its content exceeds 2%, is in the form of inorganic material. Organic sulfur occurs in both aliphatic and aromatic forms. Sulfide and disulfide groups are involved in linking clusters.

2.3.2 Bonding between Macromolecules

Individual macromolecules in coal structure are bonded to each other by a variety of forces and bonds to form a solid, three-dimensional coal structure. Aromatic clusters are linked by connecting bonds through oxygen, methylene or longer aliphatic groups, or disulfide bridges. The proportions of the various functional groups change with rank. Acid groups decrease early in the rank progression and other groups follow, leaving ether groups to act as cross-linking groups. Another type of linkage involves hydrogen bonds that, for example, hold hydroxy and keto groups together in solids.

3. DESCRIBING AND CLASSIFYING COAL

Coal samples from around the world show differences in appearance and properties—soft, moist, and brownish; very hard, glossy, and black solid; or numerous variations in between. Coal is primarily used for electric power generation. Other coal uses include drilling fluids, briquettes, and formation of coke for iron smelting. For more than a century, it supplied hundreds of chemical intermediates to prepare medicines, dyes, paints, and perfumes. These coal applications require different physical, chemical, and other properties of coal.

A number of classification systems have been developed to facilitate the production, sale, and use of coal and to satisfy the needs of major stakeholders. They seek a systematic way by which to assess performance and economic value of coals in particular applications and to facilitate assessing potential operational problems.

3.1 The Ranks of Coal

In the United States in 1938, the American Society for Testing and Materials (ASTM) adopted classification of coals "by rank" as a standard means of coal specifications (D388), parameters of which are provided in Table IV. Lower ranking coals are classified by their calorific values on a moisture- and mineral-free basis. Higher ranking coals are specified in terms of percentage of fixed carbon ($\geq 69\%$) or volatile matter ($\leq 31\%$) on a moisture- and mineral-free basis. ("Volatile matter" is defined by the ASTM D3175 procedure.) Many properties of coal vary in a fairly regular way with rank. This allows one to judge the properties of a particular coal on the basis of its rank alone.

Anthracite, bituminous, subbituminous, and lignite are ranks of coal in descending order. Peat itself is not considered coal. Less mature lignite is called brown coal. A further distinction is made within bituminous and subbituminous coals on the basis of volatile matters and agglomerating properties. Lignite and subbituminous are low-ranking coals, whereas bituminous and anthracite are high-ranking coals.

3.1.1 Descriptions of Coals by Their Ranks

3.1.1.1 Brown Coal Geologically least mature, brown coal has high moisture, sometimes more than 60%. On drying, it tends to disintegrate. Although susceptible to spontaneous ignition, it has a low heating value ($\sim$3000 BTU/pound as mined).

3.1.1.2 Lignite Geologically not mature, with easily recognizable plant parts present, lignite has less than 75% carbon on a moisture- and ash-free basis and moisture as high as 40 to 42%. Because of its high volatile matter, it ignites easily, but with a smoky flame. It has a low heating value and tends to disintegrate on drying.

TABLE IV
Classification of Coals by Rank

	Fixed carbon (percentage)[a]		Volatile matter (percentage)[a]		Gross calorific value (kJ/kg)[b]		
	≥	<	≥	<	≥	<	Agglomerating character
Coal rank							
Anthracite							
Meta-anthracite	98			2			Nonagglomerating
Anthracite	92	98	2	8			Nonagglomerating
Semi-anthracite[c]	86	92	8	14			Nonagglomerating
Bituminous							
Low volatile	78	86	14	22			Commonly agglomerating[e]
Medium volatile	69	78	22	31			Commonly agglomerating[e]
High volatile							
A		69	31		32,500[d]		Commonly agglomerating[e]
B					30,200[d]	32,500	Commonly agglomerating[e]
C					26,700	30,200	Commonly agglomerating[e]
					24,400	26,700	Agglomerating
Subbituminous							
A					24,400	26,700	Nonagglomerating
B					22,100	24,400	Nonagglomerating
C					19,300	22,100	Nonagglomerating
Lignite							
A					14,600	19,300	Nonagglomerating
B						14,600	Nonagglomerating

Source. Vorres, K. S. Coal. *In* "Encyclopedia of Chemical Technology" (J. Kroschwitz, Ed.), 4th ed. John Wiley and Sons, New York.
[a] Dry, mineral-matter-free basis.
[b] Moist, mineral-matter-free basis (i.e., contains inherent moisture)
[c] If agglomerating, classify in low-volatile group of the bituminous class.
[d] Coals having 69% or more fixed carbon on the dry, mineral matter-free basis are classified according to fixed carbon, regardless of gross calorific value.
[e] There may be nonagglomerating varieties in the groups of bituminous class, and there are notable exceptions in high-volatile C bituminous group.

3.1.1.3 Subbituminous Coal ("Black Lignite") Geologically more mature, as manifested by the absence of woody texture, subbituminous coal has a tendency to disintegrate and spontaneously ignite like lignite, but it burns more cleanly. Since the 1970s, environmental concerns have increased interest in low-sulfur subbituminous coal.

3.1.1.4 Bituminous Coal Bituminous coal is black and often appears to be banded with alternating layers of glossy and dull black colors. It has a higher heating value and lower moisture and volatile matters than does subbituminous coal. It has little tendency to disintegrate or ignite spontaneously.

3.1.1.5 Anthracite With very low volatile matter, sulfur, and moisture, anthracite burns with a hot clean flame and no smoke or soot. Stable in storage, it can be handled without forming dust. Hardest and densest of all the coal ranks, anthracite is jet black and has a high luster.

Although not part of the standard ASTM classification, some coals, with 25 to 50% ash by weight on a dry basis, are termed "impure coal." The mineral matter is introduced during deposition, mostly as clay or as fine-grained detrital mineral matter (bone coal) or later by secondary mineralization (mineralized coal). Other deposits in which mineral matter predominates, such as oil shale, coaly carbonaceous shale, and bituminous shale, are also not classified as coal.

3.2 The International Systems for Coal Classification

Many countries have official national organizations that develop and maintain standards for analyzing coal and classification schemes that are used in the coal trade and compilation of coal resource and reserve data. In the United Kingdom, the British Standards Organization serves this purpose, as does the ASTM in the United States.

3.2.1 National Coal Board Classification for British Coals

Proposed in 1946 by the Department of Scientific and Industrial Research, this classification is based on (1) volatile matter on a dry, mineral-free basis and (2) coking power, as measured by the Gray–King assay. Classification of lower ranking coals is based primarily on the Gray–King assay. Because this classification is applied to coals having less than 10% ash, high-ash coals need to be first cleaned by methods such as a float–sink separation.

3.2.2 International Classification (Hard Coal)

The rapid increase in the international coal commerce since 1945 necessitated an international system of coal classification. In 1956, the Coal Committee of the European Economic Community agreed on a system designated as the International Classification of Hard Coal by Type. Among the parameters considered are volatile matter and gross calorific value on a moisture- and ash-free basis. Table V provides various national classifications and their relationships to the classes of the international system.

In coal classification, most countries use volatile matter as the primary criterion, with a secondary criterion being coking properties. Criteria employed by some countries are as follows:

- Belgium: none
- Germany and The Netherlands: appearance of coal button
- France and Italy: Free Swelling Index (FSI)
- Poland: Roga Index
- United Kingdom: Gray–King assay

The International Organization for Standardization (ISO) based in Geneva, Switzerland, formed a committee to develop international standards. These standards, which incorporate both coal rank and coal facies parameter, replaced the UN Economic Commission of Europe's classification for hard coal.

A three–digit classification of the international system combines rank (class), caking (group), and coking (subgroup) parameters into a three-code system, as shown in Table VI. The first digit indicates the class or rank (higher digits for lower ranks). Coals having less than 33% volatile matter are divided into classes 1 to 5, and those with higher volatile matter are divided into classes 6 to 9. The second digit is based on the caking properties such as the Roga Index and the FSI. The third digit defines a subgroup based on coking properties.

3.2.3 International Classification (Brown Coal and Lignite)

These types of coals are defined as those having less than 23,860 kJ/kg heating values. Their classification employs a four-digit code. The first two digits (class parameter) are for total moisture content of freshly mined coal on an ash-free basis. The third and fourth digits are defined by the tar yield on a dry, ash-free basis.

4. CONSTITUENTS OF COAL AND THEIR IMPACTS

4.1 Organic Components in Coal

The functional groups within coal contain primarily elements C, H, O, N, and S. The significant oxygen-containing groups found in coals are carbonyl, hydroxyl, carboxylic acid, and methoxy. The nitrogen-containing groups include aromatic nitriles, pyridines, carbazoles, quinolines, and pyrroles. The sulfur-containing groups are thiols, dialkyl and aryl-alkyl thioethers, thiophenes, and disulfide.

The relative amounts of various functional groups depend on coal rank and maceral type. In vitrinites of mature coals, the principal oxygen-containing functional groups are phenolic and conjugated carbonyls as in quinones. In exinites, oxygen is present in functional groups such as ketones (unconjugated carbonyls).

The aromaticity of macromolecules in coal increases with coal rank. Calculations based on several coal models indicate that the number of aromatic carbons per cluster varies from 9 for lignite to 20 for low-volatile bituminous coal, and the number of attachments per cluster varies from 3 for lignite, to 4 for low-volatile bituminous coal, to 5 for subbituminous through medium-volatile bituminous coal.

4.2 Inorganic Constituents in Coal

To determine inorganic constituents in coal, a variety of instrumental techniques are available, including

TABLE V

The International System and Corresponding National Systems of Coal Classes

International system			National classifications							
Class number	Volatile matter (percentage)	Calorific value (kJ/g)[a,b]	Belgium	Germany	France	Italy	The Netherlands	Poland	United Kingdom	United States
0	0–3					Anthraciti Speciali		Meta-antracyt		Meta-anthracite
1A	3.0–6.5		Maigre	Anthrazit	Anthracite	Anthraciti Comuni	Anthraciet	Antracyt	Anthracite	Anthracite
1B	6.5–10.0							Polantracyt		
2	10–14		$\frac{1}{4}$ Gras	Magerkohle	Maigre	Carboni Magri	Mager	Chudy	Dry steam	Semianthracite
3	14–20		$\frac{1}{2}$ Gras	Esskohle	Demigras	Carboni Semigrassi	Esskool	Polkoksowy meta-koksowy	Coking steam	Low-volatile bituminous
4	20–28		$\frac{3}{4}$ Gras	Fettkohle	Gras à Courte flamme	Carboni Grassi Corta fiamaa	Vetkool	Orto-koksowy	Medium volatile coking	Medium-volatile bituminous
5	28–33		Gras	Gaskohle	Gras Proprement dit	Carboni Grassi Media fiamma		Gazowo koksowy		High-volatile bituminous A
6	>33(33–40)	32.4–35.4				Carboni da gas	Gaskool			
7	>33(32–44)	30.1–32.4			Flambant gras	Carboni grassi da vapore	Gasvlamkool	Gazowy	High volatile	High-volatile bituminous B
8	>33(34–46)	25.6–30.1		Gas flammkohle	Flambant scc	Carboni Secchi	Vlamkkol	Gazowoplomienny		High-volatile bituminous C
9	>33(36–48)	<25.6						Plomienny		Subbituminous

Source. Vorres, K. S. Coal. *In* "Encyclopedia of Chemical Technology" (J. Kroschwitz, Ed.), 4th ed. John Wiley and Sons, New York.

[a] Calculated to standard moisture content.

[b] To convert to BTU/pound from kJ/g, multiply by 430.2.

TABLE VI

International Classification of Hard Coals

	Value	Volatile matter (percentage; dry ash-free basis)	Heating value (MJ/kg; moist ash-free basis)
First digit: coal class	0	0–3	
	1A	>3–6.5	
	1B	>6.5–10	
	2	>10–14	
	3A	>14–16	
	3B	>16–20	
	4	>20–28	
	5	>28–33	
	6	>33–41	>32.45
	7	>33–44	32.45–30.15
	8	>35–50	30.15–25.55
	9	>42–50	25.55–23.78

	Value	Free swelling index	Roga index
Second digit: coal group	0	0–0.5	0–5
	1	1–2	>5–20
	2	2.5–4	>20–45
	4	>4	>45

	Value	Gray–King coke type	Dilatometer, maximum dilatation (percentage)
Third digit: coal subgroup	0	A	Nonsoftening
	1	B–D	Contracting only
	2	E–G	<0
	3	G1–G4	>0–50
	4	G5–G8	>50–140
	5	>G8	>140

Source. Damberger, H. H. Coal. *In* "Encyclopedia of Science and Technology," 7th ed. McGraw–Hill, New York.

X-ray diffraction, infrared spectroscopy, differential thermal analysis, electron microscopy, and petrographic analysis. Typically, the coal sample has to be first prepared by low-temperature "ashing" to remove the organic material. Analyses have shown the presence of nearly all of the elements in coal, at least in trace quantities. Certain elements (e.g., Ge, Be, B, Sb) are found primarily with organic matter, whereas others (e.g., S, P, Na, K, Ca, Mg) are found in both inorganic and organic components of coal. The primary elements forming inorganic constituents are Al, Si, Ca, Mg, Na, and S, with the secondary elements being Zn, Cd, Mn, As, Mo, and Fe.

The presence of inorganic constituents in coal is a result of several separate processes. These include (1) natural occurrence in plants as, for example, silica in cell walls of scouring rushes, (2) incorporation by ion exchange (cations linking with carboxylic acid groups or formation of chelating complexes), and (3) incorporation by mechanisms such as filtration. Some come from the detrital minerals deposited during peatification, whereas others come from secondary minerals that crystallize from water that percolates through the coal seams. During the early stages of coal development, minerals can be deposited in the peat either allogenically (minerals carried and deposited unaltered from outside the peat bed) or authigenically (minerals altered during the deposition in the peat bed).

4.3 Minerals Mixed with Coal

Table VII provides a list of different minerals that have been found in coal. Mineral matter in coal can be classified as inherent mineral matter, included mineral matter, and excluded mineral matter. Inherent mineral matter is bound to organic matter in the coal's structure. Included mineral matter is that which forms an inorganic entity within pulverized coal particles. Excluded mineral matter is that which is external to the coal seam, taken in accidentally.

Mineral matter occurs in coal seams as layers, nodules, fissure fillings, rock fragments, and small particles finely distributed throughout. These are classified as inherent (plant derived), detrital, syngenetic, or epigenetic. Inherent minerals contribute less than 2% to the coal ash.

Detrital minerals are those that were washed into the peat swamp by water streams or blown in by wind. The various clay minerals are the most common detrital minerals. Other common detrital minerals include quartz, feldspar, garnet, apatite, zircon, muscovite, epidote, biotite, augite, kyanite, rutile, staurolite, topaz, and tourmaline. The secondary minerals are kaolinite, calcite, and pyrite.

Syngenetic minerals are formed *in situ* from colloidal, chemical, and biochemical processes as organic matter accumulates in swamps. Pyrite is the most common type, and in some coals certain clay materials are syngenetic in origin. Epigenetic minerals are formed during later stages of coalification. These include calcite, dolomite, ankerite, siderite,

TABLE VII
Minerals Found in Coals

Clay minerals		**Chloride minerals**	
Montmorillonite	Common	Halite	Common
Halloysite	Rare	Sylvite	Rare
Illite-Sericite	Abundant	Bischofite	Rare
Kaolinite	Dominant	**Silicate minerals**	
Mixed layer	Common	Quartz	Dominant
Clay minerals/smectite		Tourmaline	Rare
Sulfide minerals		Staurolite	Rare
Pyrite (isometric)	Common	Sanidine	Rare
Chalcopyrite	Rare	Hornblende	Rare
Millerite	Rare	Biotite	Rare
Sphalerite	Rare	Garnet	Rare
Pyrrhotite	Rare	Epidote	Rare
Arsenopyrite	Rare	Orthoclase	Rare
Galena	Rare	Topaz	Rare
Marcasite (orthorhombic)	Rare	Zircon	Rare
Carbonate minerals		Kyanite	Rare
Calcite	Common	Albite	Rare
Siderite	Rare	Augite	Rare
Dolomite	Common	Feldspar	Common
Witherite	Rare	**Oxide and hydroxide minerals**	
Ankerite (ferroan dolomite)	Common	Hematite	Rare
Sulfate minerals		Limonite	Rare
Barite	Rare	Lepidocrocite	Rare
Bassanite	Rare	Magnetite	Rare
Rozenite	Rare	Goethite	Rare
Roemerite	Rare	Rutile/Anatase	Rare
Sideronatrite	Rare	Diaspore	Rare
Gypsum	Common	**Phosphate minerals**	
Jarosite	Rare	Apatite	Rare
Melanterite	Rare		
Mirabilite	Rare		
Anhydrite	Rare		
Szomolnokite	Rare		
Coquimbite	Rare		
Kieserite	Rare		

Source. Mehta, A. (Project Manager). (1986). "Effects of Coal Quality on Power Plant Performance and Costs," Vol. 4: "Review of Coal Science Fundamentals," (EPRI CS-283). Electric Power Research Institute, Palo Alto, CA.

pyrite, marcasite, sphalerite, barite, certain clay minerals (especially kaolinite), gypsum, carbonates, and chlorides. They commonly precipitate from solutions as they percolate through the peat or along joints and pores in the coal. Sometimes, they are coarsely crystalline or in the shape of nodules.

4.3.1 Impact of Mineral Matter on Coal Use

Mineral matters in coal cause major problems in coal-fired power plants. Hard minerals cause erosion and wear of equipment during handling, crushing, and grinding. Ash, produced by mineral matters, needs to be disposed of in an environmentally safe manner, adding to the power generation cost. High-temperature ash discharge as a molten slag involves substantial loss of sensible heat. Slagging and fouling on boiler and superheater walls reduce heat transfer efficiency, increase fan power requirements, and in extreme cases block passages in boilers. Major air pollution is caused by mineral matters such as sulfur and mercury. Minerals in coal can behave as poisons for a variety of coal conversion processes. For

example, titanium compounds in coal act as poisons in coal gasification and liquefaction.

Mineral matters do not always have a negative impact on coal use. A high sulfur content of feed coal in the synthetic fuels process is deemed to be best for catalysis. A high calcium carbonate content in combustion promotes self-cleaning of emissions. The presence of sodium in flue gas enhances the performance of electrostatic precipitators.

To assess the potential adverse effects of minerals in the coals, it is insufficient to depend on the "ultimate and trace element" analyses. The forms in which various elements are present in coal are also important. For example, sulfur in coal can have 30 different forms. Organically bound sulfur exists in a variety of functional forms. Selenium occurs in pyrite and sphalerite or as organically bound selenium, lead selenide, and ion-exchangeable, water-soluble selenium.

In general, one needs to know four factors about mineral matters in the coal to assess their effects on coal use:

- Chemical composition of the minerals
- Particle size distribution of mineral matters
- Percentage compositions of minerals
- Extent of crystallanity of the minerals

4.3.2 Metallurgical Coke

The quality of metallurgical coke depends on ash yield and composition and the sulfur content of the coal in addition to its rank, maceral composition, and density. Phosphorus and vanadium in coal are undesirable in the coke-forming process. Depending on the maceral and other properties, cokes of various types and strength are generated. Therefore, to produce cokes of a specific strength to support the iron ore in a blast furnace, coal blending is necessary.

4.3.3 Fouling

The buildup of sintered ash deposits ("fouling") on the heat exchange surfaces of coal-fired boilers is one of the most costly problems. Fouling drastically reduces boiler efficiency and also promotes corrosion and erosion. The size and strength of these deposits depend on the configuration and operation of the boiler and the mineral contents of the coal. Although sodium is often cited as the primary cause of fouling, the formation of sintered ash deposits is the result of the interaction of many inorganic constituents. Research indicates that ash yield and sodium, calcium, and magnesium concentrations in ash are the most important factors influencing the fouling behavior of low-rank coal.

4.3.4 Slagging

Slag is a molten ash deposited on boiler walls in zones subjected to radiant heat transfer. Minerals that have low melting points tend to stick to the boiler walls and accumulate to form massive fused deposits. Many slags reach a point at which the viscosity begins to rise rapidly with very small temperature changes. Above this "temperature of critical viscosity," slag becomes too viscous to flow and drain from the equipment.

The iron content of the coal ash has a major influence on the ash fusion temperature and the slagging potential of a coal. Iron sulfides tend to promote slagging more than do other iron-bearing minerals. Illite, a clay mineral, also may play a role in the slagging process by acting as an initiator.

4.3.5 Corrosion

Corrosion is enhanced by nonsilicate impurities in coal that volatilize and collect on the metal walls. Low-temperature corrosion can be caused by acidic gases (chlorine and sulfur oxides) or condensing acids (sulfuric acid), leading to most severe damage of the boiler's coolest parts. Therefore, the flue gas temperature needs to be above the acid dew point to minimize sulfur corrosion effects. High-temperature corrosion is caused by oxidation, sulfidation, or chlorination of the metal. Sodium and potassium sulfates play an important role in the corrosion process. High levels of chlorine (>1000 ppm) and (to a lesser extent) fluorine are considered to be potentially corrosive. At high levels ($>0.6\%$), chlorine, normally associated with sodium and complex sulfate deposits, initiates deposition on superheater tubes and also initiates stress corrosion cracking of tubes.

Although the elements arsenic, lead, boron, and (perhaps) zinc are typically present in trace quantities, they also enhance corrosion under certain conditions. Under high-firing temperature conditions, phosphorus at levels greater than 0.03% leads to phosphate deposits.

4.3.6 Erosion and Abrasion

Erosion and abrasion lead to wear. They cause wear in boilers and in mining and grinding equipment. Abrasive minerals in coal include quartz and pyrite. Particle size and shape affect the extent of erosion and abrasion. Coal technologists use the silicon/aluminum ratio as an indicator of abrasiveness, based on the assumption that the higher the ratio, the greater the quartz content. Ratios greater than 3 are deemed abrasive.

4.3.7 Environmental Impact

There are many ways by which coal use can negatively affect the environment. Coal mining involves blasting, dredging, and excavation that leads to air pollution by particulate matter being discharged into the atmosphere. Waste material dumped in local streams contaminates them. Coal dust explosions, mining accidents, and black lung disease among miners are some of the major health, safety, and hazard issues of mining. During coal transportation in open railroad cars, some coal particles fly off and spread over the land.

The 1990 Amendments to the Clean Air Act (CAAA) cited 12 elements in the Hazardous Air Pollutant List (P, Sb, As, Be, Cd, Cr, Co, Pb, Mn, Hg, Ni, and Se), several of which are present in coal. Many of the same elements (As, Be, Cd, Cr, Pb, Hg, and Se) appear in the U.S. Environmental Protection Agency's (EPA) list of potential pollutants in drinking water.

The EPA found that most coal wastes are not hazardous as defined by the Resources Conservation and Recovery Act. However, it reported the presence of groundwater contamination in the vicinity of waste disposal sites. At one site, nearby drinking water wells were contaminated with vanadium and selenium. The release of selenium from coal combustion wastes has caused extensive fish kills at two sites in North Carolina and at one site in Texas. Small amounts of selenium can be potentially toxic to livestock. Although coals in the Powder River Basin contain less than l ppm selenium, coal mining there releases selenium in sufficient quantity to exceed the permissible groundwater selenium concentration limit for livestock drinking water.

In power plants, mineral matters are collected as bottom or fly ash. Sometimes, they are sluiced up for environmentally sound disposal. To collect fly ash, electrostatic precipitators, bag houses, or cyclones are employed. Roughly 20 million tons of ash, or 25% of all particulate matter discharged into the air from all sources, is generated each year. Much of the ash is discarded as waste, is returned to the mines, or is dumped in the ocean. Ash has been used to make concrete. It is also a potential raw material to make aluminum. Some ashes contain valuable quantities of gallium, gold, and silver, making them attractive revenue sources.

4.3.7.1 Sulfur Pollution Sulfates in coal, constituting approximately 15% of the sulfur, are retained as ash. Pyritic and organically bound sulfur is combusted to form sulfur oxides, which in the absence of emission controls escape to the atmosphere. Sulfur oxides and sulfurous and sulfuric acids are corrosive and harmful to the health of humans, animals, and plants. Acid rains have caused deforestation and damaged buildings and monuments.

To comply with increasingly stringent government regulations, power plants have had to resort to a variety of actions. They include (1) switching from coal to alternate fuels, (2) switching from high-sulfur to low-sulfur coal, (3) blending coals, (4) beneficiating coal to lower its sulfur content, and (5) desulfurizing flue gas. Low-sulfur subbituminous coal from eastern Wyoming has come into high demand to lower sulfur emissions.

4.3.7.2 Mercury Pollution Mercury in flue gases can be present simultaneously as elemental mercury, oxidized mercury, and/or "particulate mercury," the form of which has not been determined by current methods. No reliable methods that can control mercury regardless of its form are yet available. The leading candidate is injection of activated carbon. Currently, U.S. power plants are exempt from having to control mercury pollution resulting from their operations. With the 2002 "Clear Skies" initiative, the U.S. government proposed to reduce mercury emissions from power plants by approximately 70% by 2018. The EPA is also separately considering mercury regulations under the CAAA. Thus, some mercury emission regulations can be expected, but their timing and stringency are not known.

4.3.7.3 Radionuclides Pollution Less known is the fact that coal contains trace quantities of radionuclides such as uranium, polonium, thorium, and radium. They are removed by wet scrubbers and electrostatic precipitators, and they tend to concentrate in the ash that is disposed. The use of ash in building materials has led to radon problems in residential and commercial buildings in some cases. Based on a U.S. Department of Energy's (DOE) comprehensive study showing emission values at or below the crustal values of the earth, the EPA concluded that levels of radionuclide emissions are not harmful to the ecosystem or humans and exempted power plants from reporting them.

5. COMPOSITIONAL ANALYSIS AND PROPERTIES OF COAL

5.1 Sampling Issues

Assessment of coal resources and their efficient use require knowing the coal properties in sufficient

detail. The sample must be representative and large enough for all planned testing procedures and their replications. To ensure integrity of the analyses, the sample must be properly handled. The most widely accepted authority on sampling methods is U.S. Geological Survey (USGS) bulletin 1111-B. The procedures for taking a sample, reducing the particle size of the sample, and separating a smaller portion for later analysis are in ASTM procedures D2234, D2013, and BS 1017.

5.2 Proximate Analysis (ASTM D3172)

In this analysis, a coal sample is gently heated at 105°C in an inert atmosphere, and loss of weight is attributed to "moisture" in the coal (D3173). However, this loss is also due to the evolution of gaseous matter trapped in coal.

When loss of weight at 105°C stabilizes, coal is further heated at 750°C, again under inert atmosphere. The loss in weight under these conditions is assigned to volatile matter consisting of gases, organic oil, and tar, which normally is not further analyzed (D3175).

In the last step of proximate analysis, the remaining char is burned off, leaving behind ash (D3175, D388). By subtraction of moisture, volatile matter, and ash masses from the original sample mass, the fixed carbon mass is obtained (D3174).

Although not part of the standard proximate analysis, low-temperature oxidation of fixed carbon (and organic combustible material) at 120 to 150°C using oxygen-rich plasma is employed to ensure that the original minerals remain essentially unaltered. The resulting ash is then subjected to a wide range of analyses such as X-ray diffraction.

Proximate analysis is easy and quick to perform and is of great practical value in making empirical predictions of the coal combustion behavior. The volatile matter/fixed carbon ratio predicts how the coal will burn, that is, the flame characteristics and ease of ignition. Because high-rank coals contain less volatile matter than do low-rank coals, the former are more difficult to ignite and have a shorter flame than do the latter. This makes low-rank coals more suited for cyclone burners in which rapid intense combustion leads to maximum carbon use and minimum smoke emissions. Ash data generated by the proximate analysis allow engineers to predict the frequency and quantity of ash disposal. The proximate analysis data can be used to (1) establish the ranks of coals, (2) aid in buying and selling decisions, and (3) evaluate the need for beneficiation and the like.

5.3 Ultimate Analysis (ASTM D3176)

This is an elemental analysis limited to carbon, hydrogen, sulfur, nitrogen, and oxygen. Carbon and hydrogen are measured by burning the coal and collecting the resulting carbon dioxide and water (D3178).

Because of increasingly stringent environmental restrictions on SO_x emissions, particular interest has developed in the total sulfur part of the ultimate analysis. Sulfur determination (the Eschka method) (D3177) is based on conversion of all types of sulfur in the coal (e.g., pyrite, sulfates, sulfides, organically bound sulfur) to sodium sulfate. A coal sample is ignited with magnesium oxide and sodium carbonate. The resulting sodium sulfate is then converted to water-insoluble barium sulfate, which is filtered, dried, and weighed. There are two additional sulfur determination methods. In the bomb-washing method, sulfur is precipitated as barium sulfate from the oxygen bomb calorimeter washings. In the high-temperature combustion method, the sample is burned in a tube furnace and sulfur oxides are collected and quantified by titration. Forms of sulfur are determined by the D2492 method.

Nitrogen is usually determined by first liberating elemental nitrogen from the sample by the digestion method (Kjeldahl–Gunning), converting the elemental nitrogen catalytically to ammonia, and (finally) titrating the resulting ammonia with acid (D3179).

There are no relatively simple analytical methods for oxygen determination. The most commonly used method requires a neutron source such as a nuclear reactor. Neutrons are bombarded on the sample to convert oxygen into nitrogen-16, a very unstable isotope of nitrogen. Nitrogen-16 decays rapidly by emitting beta and gamma rays that are counted. Because most laboratories do not have access to a neutron source, oxygen content is calculated by subtracting the masses of other elements from the mass of the sample (D3176).

Chorine may also be included in the ultimate analysis. It is determined by either the bomb calorimeter or the Eschka method (D2361, D4208).

5.4 Heating Value or Calorific Value

The heating value of coal is measured by an adiabatic calorimeter (D2015). The procedure gives the "gross heating value." The "net heating value" is based on the assumption that all of the water resulting from

combustion remains in vapor form (D407). The heating value of a coal determines the amount of coal a power plant needs.

5.5 Other Chemical Analysis Methods

Since the 1970s, environmental concerns have led to increasingly restrictive levels of permissible emissions of some of the trace elements in coal. A procedure has been adopted by USGS, outlined in Table VIII, to analyze trace mineral matters. MAS, ICP/AES, and ICP/MS analyses are performed on ash produced at 525°C.

5.6 Physical, Mechanical, and Other Properties of Coal

5.6.1 Physical Properties

Most of the physical properties of coal discussed here depend on the coal orientation and history of the piece of coal. Properties may vary from one coal piece to another coming from the same seam or mine.

The specific electrical conductivity of dry coal is in the semiconductor zone, 10^{10} to 10^{14} ohm.cm, increasing with rank and temperature. Moisture in coal also enhances conductivity.

The dielectric constant also varies with rank, with a minimum at about 88% carbon. Polar groups in low-rank coal and higher electrical conductivity in high-rank coal contribute to their dielectric constants.

Magnetic susceptibility measurements indicate that the organic part of coal is diamagnetic, with traces of paramagnetic behavior resulting from the presence of free radicals or unpaired electrons.

Thermal conductivity and diffusivity of coal depend on pore and crack structure. Conductivity ranges from 0.23 to 0.35 $W/(cm \cdot K)$, and diffusivity ranges from 1×10^{-3} to 2×10^{-3} cm^2/s. At 800°C, these ranges increase to 1 to 2 $W/(cm \cdot K)$ and $(1–5) \times 10^{-2}$ cm^2/s, respectively. The increases are due to radiation across pores and cracks at higher temperatures. The thermal conductivity of coal is influenced by moisture content, the chemical and physical nature of the coal during reaction, and convective and radiant heat transfer within the particle. Below 400°C, thermal conductivity of coal has been measured in the range of $(5–8) \times 10^{-4}$ $cal/(cm \cdot s \cdot K)$.

The specific heat of coal is estimated to be in the range of 1 to 2 $J/(g \cdot K)$ at 20°C to 0.4 $J/(g \cdot K)$ at 800°C. The specific heat is affected by the oxidation of the coal.

5.6.2 Mechanical and Other Properties of Coal

These properties have a much more direct bearing on the operations of power plants or other coal processing units.

5.6.2.1 Density The true density of coal, which increases with carbon content or rank, is most accurately obtained by measuring helium displacement. Values of 1.4 to 1.6 g/ml have been reported for greater than 85% carbon coal. The bulk density (D291), typically about 0.8 g/ml, depends on the void fraction. It affects coal's storage, hauling requirements, and sizing for its conveyance.

5.6.2.2 Porosity Fine porosity of coal restricts the access of oxygen to the internal structure. This retards combustion rates and depresses burnout in power plant boilers. Porosity is the result of the random arrangement of organic macromolecules within coal. Total porosity of coal ranges from 0.02 to 0.2%, with no apparent trends with carbon

TABLE VIII

Methods Employed by U.S. Geological Survey

Elements	Methods employed
Hg, F	Wet chemical analysis
Cl, P	X-ray diffraction
Se	Atomic absorption spectroscopy hydride ASS
Na, Be, Cu, Li, Th, Zn, Co, Mn, Ni, V, Cr, Sc, Sr, Y	Induced coupled plasma/atomic emission spectroscopy (acid digestion)
Si, K, Ba, Al, B, Ca, Ti, Zr, Mg, P, and Iron oxides	Induced coupled plasma/atomic emission spectroscopy (sinter)
Ag, As, Au, Bi, Cd, Cs, Ga, Ge, Mo, Nb, Pb, Sb, Sn, Te, Tl, U	Induced coupled plasma/mass spectroscopy (acid digestion)
Ce, Dy, Er, Eu, Gd, Hf, Ho, La, Nd, Pr, Sm, Tb, Tm, W, Yb, Ta	Induced coupled plasma/mass spectroscopy (sinter)
S	MAS

Source. Stanton, R. W., and Finkelman, R. B. Coal quality. *In* "Energy Technology and the Environment" (A. Bisio and S. Boots, Eds.), Vol. 1. John Wiley and Sons, New York.

content. Porosity is often subdivided into macropore ($d = 300\text{A}^\circ$), intermediate pores ($d = 12$–300A°), and micropores ($d < 12\text{A}^\circ$). Micropore volume appears to be low for low-rank coals and high for coals with carbon contents between 74 and 80% and greater than 90%. This is one reason why low-rank coals burn more easily.

5.6.2.3 Equilibrium Moisture Content Equilibrium or inherent moisture (D1412) is the moisture-holding capacity of coal at 30°C in an atmosphere of 97% relative humidity. The organic reflux moisture test (ISO R348) has been used to determine the residual moisture in the prepared sample.

Total moisture or "as received" moisture is determined by a two-step procedure (D3302, D3173) involving surface moisture measurement, reduction of gross sample, and determination of residual moisture.

Higher total moisture leads to a lower heating value of the coal and lower flame temperature, and it lowers the plant efficiency. It also causes higher haulage costs per million BTUs and causes the power plant to purchase more coal. A number of processes have been patented to dry coal at the mine site before its shipment, but none of them has been commercialized, due partly to excessive capital requirements.

5.6.2.4 Ash Fusibility (D1857) The ash fusibility test provides the "melting point" of ash. Although empirical, the procedure requires strict adherence to ensure reproducible results. Ash fusibility data are critical to power plants because they are used to assess operational difficulties encountered in using the coal (e.g., slagging, loss of heat transfer rate, frequency of boiler shutdown).

Ash fusibility under oxidizing conditions tends to be higher than that under reducing or neutral conditions. In the procedure, the condition of the starting cone of ash is reported as it undergoes several physical changes as temperature increases. These include (1) initial (cone tip) deformation, (2) softening (height/base = 1.0), (3) hemispherical shape formation (height/base = 0.5), and (4) fluid (puddle) formation.

A related criterion, namely T_{250}, the temperature at which slag has a viscosity of 250 poise, depends on base/acid and silica/alumina ratios. The lower the base/acid ratio and the higher the silica/alumina ratio, the higher the T_{250} of the ash.

5.6.2.5 Ash Analysis This term is applied to the analysis of the major elements commonly found in coal and coal ash (D2795). Oxides of Si, Al, Fe, Ti, Ca, Mg, Na, K, and P are measured. Phosphorus oxide has importance in the coke used in steelmaking processes. Two other ASTM methods for ash composition based on atomic absorption are D3682 and D3683.

The ASTM ash analysis is widely used to predict slagging and fouling. However, the way by which ash is produced in the boiler is considerably different from the way by which it is produced by the ASTM procedure.

5.6.2.6 Free Swelling Index This index is determined by quickly heating a coal in a nonrestraining crucible (D720). The values are assigned from 0 to 9, with noncaking, nonswelling coals being assigned the value 0. The higher the FSI value, the greater the swelling property.

The FSI is not recommended for predicting coal behavior in a coke oven. The index is not deemed important for pulverized coal-fired boilers. Some units (e.g., retort stokers) form coke in their normal operation. For certain coals, the FSI values reduce while in storage, indicating coal deterioration and partial oxidation.

Gray–King and ROGA tests, employed in Europe, serve a function similar to that of the FSI test in the United States.

5.6.2.7 Washability A large part of mined coal is cleaned to reduce its mineral contents before it is sold. Washability tests are performed to predict the effectiveness of coal cleaning. This is a float–sink test in heavy liquids that generates various gravity samples that are analyzed for their heating values, ash and sulfur contents, and so on. Sometimes, efforts to lower ash content by washing are uneconomical.

5.6.2.8 Strength and Hardness These are important coal properties in mine stability and comminution. Various tests are available to estimate coal's compressive and shearing strengths and hardness. On the Mohs hardness scale, lignites range from 1 to 3, bituminous from 2.5 to 3.0, and anthracite from 3 to 4.

5.6.2.9 Grindability The Hardgrove Grindability Index (HGI) method is used to determine the relative ease of coal pulverization in comparison with a series of standard coals (D409). The data can be used to determine power and sizing of grinding equipment and its expected wear rate. For low-rank coals, subbituminous and lignite, the HGI is found to depend on coal moisture, which decreases as grinding proceeds.

5.6.2.10 Friability The breaking of coal into dust during handling, transporting, and pulverizing causes explosion hazards, control of which requires special sensors and control equipment. The most friable coal contains approximately 90% carbon. This point also corresponds to the maximum HGI and minimum Vickers microhardness values. Charcoal-like fusain tends to be soft and friable.

5.6.2.11 Slaking Slaking causes coal disintegration during storage due to coal drying. This is an important phenomenon for lignite and subbituminous coals but not for high-rank coals, as lignite and subbituminous coals have high moisture content.

5.6.2.12 Spontaneous Ignition Temperature of a coal pile builds up slowly if it is not turned over. This triggers spontaneous ignition, which is a serious hazard requiring monitoring and preventive measures. Coal particle size distribution and its oxygen content have a great influence on spontaneous ignition. Other important factors include pyrite content, oxygen circulation in the pile, moisture, and rank of coal.

5.6.2.13 Coal Petrography/Vitrite Reflectance Test Extensive use has been made of vitrite reflectance measurements (D2797, D2798) and reflectograms along with the physical microscopic components (D2797, D2799) to assess coking coals. Vitrinite reflectance limits data for various coal rank classes are given in Table IX.

5.6.2.14 Gieseler Plasticity Plasticity measurement (D2639) gives the coal's tendency to fuse or soften and reach a fluid state on heating during simulated coking stages.

5.6.2.15 Audibert Arni Dilatometer/Sole Heated Oven Determination The dilatometer measures the contraction and expansion of coal during carbonization (ISO R349). The corresponding ASTM test is called Sole Heated Oven Determination (D2014). These data are important in predicting whether coke can be removed during production without damaging the equipment.

5.6.2.16 Other ASTM Tests for Coal Table X lists other ASTM tests that are relevant for coal characterization.

TABLE IX

Vitrinite Reflectance Limits Taken in Oil versus Coal Rank Classes

Coal rank	Maximum reflectance (percentage)
Subbituminous	<0.47
High-volatile bituminous	
C	0.47–0.57
B	0.57–0.71
A	0.71–1.10
Medium-volatile bituminous	1.10–1.50
Low-volatile bituminous	1.50–2.05
Semianthracite	2.05–3.00 (estimated)
Anthracite	$>$ 3.00 (estimated)

Source. Vorres, K. S. Coal. *In* "Encyclopedia of Chemical Technology" (J. Kroschwitz, Ed.), 4th ed. John Wiley and Sons, New York.

6. FRONTIERS OF COAL RESEARCH

6.1 Advanced Instrumental Methods of Measurements

The advanced measurements listed here overcome some of the limitations of the standard methods by providing more detailed and accurate characterization of specific coal. These methods can be grouped into (1) composition, (2) combustion characteristics, and (3) tests for improved correlations.

6.1.1 Composition

Table XI lists some of the advanced methods that have been used with some success, but much work still needs to be done to ensure their reliability and to assess the limits of their applicability. This is a

TABLE X

Other ASTM Tests for Coal

Test	ASTM method	Parameters	Plant performance
Calorific value	D3286, D2015	—	General combustion
Fineness	D197	75–1180 micron sieve	Burner
Sieve analysis	D410	3/8- to 8-inch sieve	Coal handling/ storage
Drop–Shatter test	D440	Friability	Pulverization
Tumbler test	D441	Friability	Pulverization
Dustiness	D547	Dustiness indices	Handling
Sulfur in ash	D1757	Sulfur	SO_x pollution

TABLE XI

Advanced Coal Composition Measurements

Method	Parameters
Electron microscopy	Ultra-fine granular structure
	Pore structure
X-ray diffraction	Size distribution of aromatic ring systems
	Diameter and thickness of lamellae
	Mean bond length
	Mineral phase evaluation
Mass spectroscopy	Molar mass and formulae
	Analysis to determine types of hydrocarbons
	Carbon number distribution
	Trace elements
Nuclear magnetic resonance	H_{ar}/H and C_{ar}/C ratios
	Distribution of isotopic tracers
Electron spin resonance	Stable free radicals
	Other species with unpaired electrons
Infrared absorption	Functional groups (e.g., OH, CH_{ar})
	Minerals
Ultraviolet–visible absorption	Difficult to apply to whole coal samples but can be used to identify components in extracts
Electron spectroscopy for chemical analysis	Elemental distribution and chemical state of surface
Optical emission spectroscopy	Elemental analysis
	Single-element and survey analysis
	Major, minor, and trace element analysis for metals and semi-metals
	Applies to coal ash only
X-ray fluorescence	Inorganic element analysis
Atomic absorption	Precise quantitative determinations of metals and semi-metals
	Single element analysis for minor and trace constituents
Mossbauer spectroscopy	Iron compounds and association with sulfur
Scanning electron microscope automatic image analysis	Size, shape, and composition of minerals in coal
Neutron activation analysis	Elemental analysis
Spark source mass spectroscopy	Semiquantitative elemental analysis of coal ash
Inductively coupled plasma spectroscopy	Quantitative multi-element analysis of coal ash
Electron microprobe analysis	Elemental analysis of micrometer-size areas
Thermometric techniques	Chemical characterization of coal ash
Optical microscopy	Identification of minerals in coal or coal ash
Ion microprobe mass analysis	Isotopic characterization of minerals or macerals
Extended X-ray analysis fine structure	Chemical environment of an atom

Sources. Mehta, A. (Project Manager). (1986). "Effects of Coal Quality on Power Plant Performance and Costs," Vol. 3: "Review of Coal Quality and Performance Diagnostics" (EPRI CS-4283). Electric Power Research Institute, Palo Alto, CA; Stanton, R. W., and Finkelman, R. B. Coal quality. *In* "Energy Technology and the Environment" (A. Bisio and S. Boots, Eds.), Vol. 1. John Wiley and Sons, New York.

dynamic area because new methods and applications are being developed continually. The methods are used to analyze microstructure, trace elements, organic material structure, and mineral matter composition and structure. They make it possible to characterize coal with more details and better accuracy. However, they require expensive instrumentation and highly skilled personnel to operate.

6.1.2 Combustion Characteristics

Although proximate analysis provides key coal data, it provides no data on devolatalization and combustion characteristics. These data are provided by

thermogravimetric analyses. Four important methods needed to design boilers and assess performance of existing boilers are the following:

- Burning profile (Babcock & Wilcox)
- Volatile release profile (Babcock & Wilcox)
- Thermogravimetric analysis of char (Combustion Engineering)
- Drop tube furnace system (Combustion Engineering)

6.1.3 Tests for Improved Correlations

Procedures have been developed to improve the correlation of coal properties with coal's performance in power plant systems. These include analysis of physical separation of pulverized coal (pulverizer wear and slagging), analysis of alkalies in coal ash (fouling), sintering strength of fly ash (fouling), free quartz index (pulverizer wear), and continuous coal grindability (pulverizer capacity).

6.2 Active Areas of Coal Research

Active coal research areas worldwide include the following:

- Advances in pollution controls for NO_x, SO_x, hazardous air pollutants (HAPs), and particulate matter
- Development of economically viable mercury emissions control technologies
- Advances in coal liquefaction and gasification
- Advances in power generation systems
- Studies in greenhouse gases sequestration
- Advanced studies to correlate physical structure, petrography, and properties
- Advanced processing studies (e.g., management of coal use by-products, coal chemicals, pyrolysis, and coal–oil–water interaction)
- Studies of coal gasification fundamentals

6.3 Carbon Sequestration

Major activities are under way worldwide to find economically viable means of sequestration of carbon dioxide, a major greenhouse gas. Although its emissions from fossil fuel power plants are only approximately 39% of its total emissions in the United States (32% from coal-fired plants), sequestration of these power plant emissions is the focus of research and development and analysis because they are localized. CO_2 use for enhanced resource recovery is being considered along with the use of underground saline formations, possibly oceans, and mineral carbonation for its sequestration. The United States has not signed the International Climate Change agreement (Kyoto Protocol). Nevertheless, its government has undertaken a 10-year demonstration project, "FutureGen," to create a coal-based zero-emission power plant.

6.4 Computational Methods

In support of experimental research programs, much work continues to be carried out in coal-related computational methods, including the following:

- Constitutive laws of multiphase flow
- Process simulation and modeling (e.g., black liquor gasification, circulating and spouted beds, spontaneous combustion, oxidative pyrolysis, coal–biomass co-firing, char gasification)
- Computational fluid dynamics in support of process design and optimization
- Three-dimensional modeling of subsurface conditions (e.g., simulation of coal bed for methane recovery)
- Engineering design models (e.g., hydroclone, circulating fluidized bed, transport reactors)
- Modeling for pollution control (e.g., NO_x formation)

SEE ALSO THE FOLLOWING ARTICLES

Clean Coal Technology • Coal, Chemical and Physical Properties • Coal Conversion • Coal, Fuel and Non-Fuel Uses • Coal Industry, Energy Policy in • Coal Industry, History of • Coal Mine Reclamation and Remediation • Coal Mining, Design and Methods of • Coal Mining in Appalachia, History of • Coal Preparation • Coal Storage and Transportation • Markets for Coal • Oil and Natural Gal Liquids: Global Magnitude and Distribution • Peat Resources

Further Reading

Damberger, H. H. (1992). Coal. *In* "Encyclopedia of Science and Technology," 7th ed. McGraw–Hill, New York. [The following subsections of this article have been excerpted with permission from McGraw–Hill: 1.2 (Influence of Geology) and 2.1 (Petrology, Petrography, and Lithotypes).]

Mehta, A. (Project Manager) (1986). "Effects of Coal Quality on Power Plant Performance and Costs," Vol. 3: "Review of Coal Quality and Performance Diagnostics" (EPRI CS-4283). Electric Power Research Institute, Palo Alto, CA. [The following subsections of this article have been excerpted with permission

from EPRI: 5.6.2.4 (Ash Fusibility [D1857]), 5.6.2.16 (Other ASTM Tests for Coal), and 6.1.2 (Combustion Characteristics).]

Mehta, A. (Project Manager) (1986). "Effects of Coal Quality on Power Plant Performance and Costs," Vol. 4: "Review of Coal Science Fundamentals" (EPRI CS-283). Electric Power Research Institute, Palo Alto, CA. [The following subsections of this article have been excerpted with permission from EPRI: 2.1 (Petrology, Petrography, and Lithotypes), 2.2.3 (Geochemical Phase [coalification]), and 4.3 (Minerals Mixed with Coal).]

Perry, M. B. (Conference Chairman), and Ekmann, J. M. (Program Chairman). (2001). "Proceedings of the 11th International Conference on Coal Science." San Francisco, CA.

Savage, K. I. (1977). Coal: Properties and Statistics. *In* "Energy Technology Handbook" (D. M. Considine, Ed.). McGraw–Hill, New York. [The following subsections of this article have been excerpted with permission from McGraw–Hill: 1.2 (Influence of Geology), 5.1 (Sampling Issues), 5.6.2.3 (Equilibrium Moisture Content), and 5.6.2.6 (Free Swelling Index).]

Schobert, H. H. (1987). "Coal: The Energy Source of the Past and Future." American Chemical Society, Columbus, OH. [The following subsections of this article have been excerpted with permission from the American Chemical Society: 1.1 (Influence of Age), 2.2.1 (Components of Plants), 3.1.1 (Descriptions of Coals by Their Ranks), 4.2 (Inorganic Constituents in Coal), 4.3 (Minerals Mixed with Coal), 5.2 (Proximate Analysis [ASTM D3172]), and 5.3 (Ultimate Analysis [ASTM D3176]).]

Simon, J. A. (1981). Coal. *In* "Encyclopedia of Energy," 2nd ed. McGraw–Hill, New York. [The following subsections of this article have been excerpted with permission from McGraw–Hill: 1.2 (Influence of Geology) and 4.3 (Minerals Mixed with Coal).]

Stanton, R. W., and Finkelman, R. B. (1995). Coal quality. *In* "Energy Technology and the Environment" (A. Bisio and S. Boots, Eds.), vol. 1. John Wiley, New York. [The following subsections of this article have been excerpted with permission from John Wiley and Sons: 4.3.1 (Impact of Mineral Matter on Coal Use) and 4.3.7 (Environmental Impact).]

Vorres, K. S. (1992). Coal. *In* "Encyclopedia of Chemical Technology" (J. Kroschwitz, Ed.), 4th ed. John Wiley, New York. [The following subsections of this article have been excerpted with permission from John Wiley and Sons: 1.2 (Influence of Geology), 2.1 (Petrology, Petrography, and Lithotypes), 2.3.1 (Bonding within a Macromolecule), 3.2 (The International Systems for Coal Classification), and 5.6.1 (Physical Properties).]

Coal Storage and Transportation

JAMES M. EKMANN and PATRICK H. LE
U.S. Department of Energy
Pittsburgh, Pennsylvania, United States

1. Coal Transport
2. Coal Preparation
3. Stockpiles
4. Handling Coal at the End-Use Site
5. Storing Coal at the End-Use Site

Glossary

coal preparation A process that includes not only crushing and breaking methods (coal pretreatment) but also all the handling and treatment methods necessary to remove impurities and prepare coal for commercial use.
coal pretreatment A process consisting of breaking, crushing, and screening run-of-mine coal to produce coal with a predetermined top size.
coal screening A method to separate coal particles by passing them through a series of screens, of decreasing size.
coal sizing A process to separate raw coal into various particle sizes, either by screening or by other classification methods.
coal spontaneous combustion The process whereby coal ignites on its own, because the rate of heat generation by the oxidation reaction exceeds the rate of heat dissipation.
coal storage The storage process to accomplish various objectives: (1) to reclaim coal promptly and economically, (2) to carry-out blending/homogenization in order to achieve coal with uniform or desirable characteristics, and (3) to store various coals whose price and market demand vary seasonally, thus maximizing profits.
dead storage of coal The concept that coal is not immediately accessible because it is in a compacted pile to prevent weatherization and some means is needed to reclaim the coal.
live storage of coal The concept whereby stored coal is immediately accessible by the principal means of reclaiming in coal piles or by gravity flow in silos.
run-of-mine (ROM) coal A heterogeneous material that is often wet and contaminated with rock and/or clay and is unsuitable for commercial use without treatment.
unit train A railroad technique to efficiently transport coal on a predetermined schedule, by using a dedicated train and the associated equipment for loading and unloading.

This article deals with the coal management system that encompasses coal preparation, storage, and transportation. Coal must first be transported from the mines to the point of consumption, often an electric power plant. Railroads are the predominant form of transportation, sometimes in cost-effective "unit trains." Barges and ships are the other major mode of coal transportation, with trucks seeing limited use due to higher costs. At the preparation plants, the coal undergoes preparation processes that remove rocks, noncombustibles, and other impurities from run-of-mine coal and also crush and size the coal for its end use. After preparation, coal is usually stored. Since even prepared coal is not homogenous, various methods are employed to ensure that the coal properties remain fairly uniform when burned. Also, coal will weather and oxidize, changing its properties, and in certain cases, coal can self-ignite, unless precautions are taken.

1. COAL TRANSPORT

Coal has relatively low energy content per unit weight and competes on the market with other fuels as a low-cost boiler fuel. Consequently, a low-cost and efficient delivery system based on many alternatives is essential to maintain its competitiveness.

Based on statistics from Consol Energy, the modes of coal transportation in the United States are distributed as follows: 62% by rail, 12% by truck, 16% by barge and/or vessel, and 10% by conveyor belt or slurry pipeline; these numbers were presented at a coal and power training course in July 2000.

1.1 Rail Transport

Railroads are very well positioned to transport U.S. coals from the mines to end-use destinations, such as utility and industrial plants. Their dominance as the major carrier of coal is expected to continue. The U.S. railway system has 126,682 miles of Class 1 track, 47,214 miles of Class 2 and 3 track, and 1.2 million freight cars to carry all commodities, according to data from Consol Energy presented by Robert J. Schneid at the above-mentioned coal and power training course.

The trend in the industry is toward a consolidation through state ownership or a merger of privately owned systems. Dominant companies include the following: in the West, Union Pacific and Burlington Northern Santa Fe, and in the East, CSX and Norfork Southern. Table I gives an insight into the railroad consolidation of U.S. companies, with Stan Kaplan of PA Consulting Group as the source of information in year 2002.

Several strategies have been considered to improve railroad services and profitability: (1) cost cutting and productivity improvements; (2) merger; (3) differential pricing; (4) diversification; (5) new services; and (6) investment.

Strategies (1), (2), and (4) have exhausted their potential. Diversification is being abandoned. Cost cutting may be reaching a point of diminishing returns. Mergers have not proved to be a financial success. As a result, the focus of the railroad industry is on improved intermodal and other truck-competitive services, but not for coal as a high-priority service. The tension between demands for shipment at low rates with high-quality services and objectives to improve profitability will persist.

TABLE I

Consolidation of U.S. Railroad Companies

Old companies	New company
Seaboard Coast line	CSX
Chessie System	
Conrail	
Conrail	Norfolk Southern (NS)
Norfolk & Western	
Southern Railroad	
Missouri Kansas Texas	Union Pacific (UP)
Denver & Rio Grande Western	
Southern Pacific	
Union Pacific	
Missouri Pacific	
Chicago North Western	
Santa Fe	Burlington
Burlington Northern	Northern Santa Fe
Kansas City Southern	Kansas City Southern
Illinois Central Gulf	Canadian Pacific

Source. Adapted from Kaplan (2002).

In general, railroads have offered three major types of service to transport coal: (1) multiple-carload, (2) trainload, and (3) unit trains. Another less important type of service is single-carload rates for low volumes of smaller mines. Multiple-carload rates ranging from as few as 5 or 10 cars to as many as 100 cars are usually significantly lower than single-carload rates. Trainload rates are characterized by large minimum-volume or car requirements, since they move from origin to destination as one unit. Not meeting the annual volume requirements could lead to a higher penalty rate. Unit train rates usually have the lowest rates possible; the difference between them and trainload rates is that unit train rates use dedicated equipment and operate on a regular schedule.

The unit train (approximately 100 to 150 cars) came into existence in the early 1960s in response to other competitive alternatives in the transportation industry. Whereas most unit trains consist of 100 cars, they can also operate with as few as 20 cars. For a train unit of 100 to 150 cars, the payload would range from 10,000 to 15,000 tons and require four to six locomotives per train rated at 4000 to 5000 hp each. If lightweight aluminum cars are used, a unit train of 100 to 120 cars, carrying 121 tons (110 tonnes) each, could move more than 14,000 tons (12,700 tonnes) in total.

In terms of motive power for unit trains, five or six earlier diesel locomotive models (3000 hp) would be required to move a train of 100 to 125 cars. In the 1990s, distributed power became the most efficient method of transporting coal by rail. In this system, a remote-controlled engine is placed in the middle of the train. In the early 2000s, two or three modern diesel electric locomotives (6000 hp) based on microprocessor and alternating current motor technology could do the work of five earlier models. In addition, advances in control technology, called microwave centralized traffic control systems, have facilitated the safe movement of trains with minimal between-train clearance, therefore leading to significant increases in moving capacity.

1.1.1 Loading and Unloading Railcars

Typical coal railcars can carry 80 to 125 tons per car and the average load is 100 to 110 tons. Two types of

railcars are in service: the solid bottom gondola and the hopper car.

A rotary dump system can handle gondola and hopper cars, whereas the undertrack pit can handle only hopper cars with bottom-opening doors.

The selection of the unloading method requires an evaluation of several trade-offs, such as the following: (1) gondola cars are the least expensive, need minimal maintenance, and can be completely unloaded with minimum coal hang-up but do require an expensive rotary dumper with the capability to accurately position the cars; and (2) hoppers (or bottom-dump cars) are more expensive and require more maintenance than the gondolas, due to the gate mechanism, but they can be unloaded more quickly on a relatively inexpensive track hopper. In addition, hopper cars are more difficult to unload under freezing conditions.

In addition to loading and unloading the coal, the issue of windage losses is frequently overlooked. Windage losses, i.e., coal blown out by wind from open-top cars, cause economic loss and environmental concern. One solution is a lid system, such as a flip-top lid, to control windblown loss, reduce dust, and prevent additional moisture from penetrating the coal, thus reducing the potential for freezing. Frozen coal is a troublesome problem related to unloading coal because the exposed top surface, the size plates of the car, and the surfaces of the sloped hoppers could be frozen. Several methods to thaw the cars are available, ranging from bonfires and oil-burning pans to sophisticated sheds. Treating the car sides and bottoms of the cars with oil can minimize freezing and ease the coal unloading process.

In terms of loading systems, three types of trackage are commonly used for unit trains. These are (1) the loop or balloon track, (2) the single track that is situated above and below the loading station, and (3) the parallel tracks that allow loading on both tracks. The loading system consists of locomotive, car haul, tripper conveyor, and hydraulic ram. The locomotive moves the cars in one pass under the loading station. The car-haul system moves two strings of cars in opposite directions on parallel tracks under two loading stations. The tripper conveyor system moves two-way loading chutes over two stationary parallel strings of cars. The hydraulic ram system moves the cars into loading position, after the locomotive has been removed.

In terms of unloading systems, there are two major types: the rotary dump and the undertrack pit. The rotary dump system handles the rollover type of car (hopper or gondola design) and the undertrack pit handles bottom-up cars.

1.2 Transport by Barge and Ship—Water Transport

The U.S. Great Lakes and inland waterway systems are a very valuable asset for moving coal and other commodities, as shown in Table II. In addition, the U.S. river network is connected to the Gulf Intracoastal Waterway, allowing highly efficient barge and lake carrier services. The Great Lakes facilitate shipments to the export market but have seasonal limitations. The inland waterway system is characterized by open top barges and the size of the tow. The tow is a function of the width and depth of the respective river and the configurations of its locks and dams.

1.2.1 Barge Transportation—Loading and Unloading

Barge transportation offers a lower cost alternative to railways for the majority of power plants and coke ovens that are located on inland waterways. In this waterway system, 25,000 miles of navigable rivers, lakes, and coastlines and a system of 192 locks and dams are ideal means for moving coal over long distances. Consequently, barge transportation could be considered competitive because it requires less fuel, the infrastructure is already in place or needs only to be updated, and barges carry large loads. A jumbo barge can carry an amount of coal equivalent to 15 railcars or 58 truck loads.

In addition, barge transportation is a good example of an intermodal carrier. With mines near a navigable waterway, several combinations are possible: (1) for short distances, trucks or conveyors

TABLE II

U.S. Riverways

Characteristics
• River network with portion highly navigable (depth of 9 ft or more)
• Connects with the Gulf Intracoastal Waterway
• Spans the country from Pennsylvania and West Virginia to the east and Minnesota and Wisconsin to the north
• Alabama and Florida on the Gulf Coast to the southeast
• Louisiana and Texas on the Gulf coast to the southwest
• Great Lakes allows heavy freight distribution
— From Illinois, Wisconsin, and Minnesota to the west and to Ohio, Pennsylvania, and New York to the east
— Access for exports to Canada through the St. Lawrence Seaway to overseas customers (Europe and Mediterranean region)

Source. Adapted from Lema (1990).

are used to transport coal to a barge terminal (for instance, up to 10 miles for conveyors and 30 or more miles for trucks, away from the mine); and (2) for longer distances, railroads can be used to transport coal to the river terminal (for instance, over 100 miles away from the mine).

However, the economics of barge transportation is affected by the circuitry of the waterways, their nature, the weather, and market conditions. Barge size is subject to the size, number, and condition of the locks. Low water levels and frozen rivers can halt shipment. In shipment, coal competes with other commodities to be transported and backhaul is a good solution to reduce coal rates.

To increase productivity, barge size has increased from a standard 1000-ton capacity to the jumbo-sized barge (9 ft draft) capable of carrying 1400 to 1800 tons. Usually, the jumbo-sized barge would carry 1800 tons, which enables the tow to move approximately 72,000 tons of coal. This is comparable to the loading of five unit trains. Towboats have pushed up to 48 jumbo barges on the Lower Mississippi River where there are no locks, but 40 barges are more common. In other instances, 15 jumbo barges (as maximum) are more common for a single lock, such as the 1200 ft lock on the Ohio River.

In the United States, five categories of loading plants are used, depending on the loading requirements. These categories are as follows:

1. The simple dock, which is used only to a small extent, mostly by small operators and for temporary facilities. It depends on a relatively stable water level. The trucks dump directly into the barges or into a fixed chute.
2. The stationary-chute system has declined in popularity; it is relatively economical and has the same disadvantages as the simple dock system (sensitive to water level fluctuations).
3. The elevating-boom system is commonly used; it is very flexible and adjustable to variations in water levels. It can handle high capacities. The common features of this system found in most plants are the hinged boom, the pantograph trim chute, and the extensive dock area.
4. The spar-barge system is suitable for rivers with fluctuating levels. It has two essential components: the floating spar barge and the shuttle. The floating spar-barge (as floating dock) contains the loading boom and the barge-positioning equipment. The shuttle or retracting loading belt leads to the spar-barge.
5. The traveling-tripper system is the method used to stockpile coal at large power plants; it is applied to load barges. The barges are held stationary by a barge-mooring system. The traveling tripper (as a movable loading point), with a reversible traverse shuttle belt and telescoping chutes, discharges coal in the barges. This barge loading method can load eight barges at a time, four on each side.

Regarding barge unloading, coal operators usually receive barge sizes and types that will accommodate their barge unloading systems. Earlier, barge unloading was performed using whirler-type cranes with clamshell buckets. Most barge coal is still unloaded by clamshell buckets that are mounted on either stationary straight-line-type towers or on moving tower barge unloaders; in the latter case, the tower is mounted on rails, it straddles the dock belt, and the barges remain stationary.

1.2.2 Ocean Transport

The U.S. export coal market has primarily been exporting coking coal for use in steel production overseas, but the market for steam coal has developed at a fast rate. Consequently, modern vessels and modern or upgraded docking facilities (e.g., larger and faster processing capabilities) are needed.

In addition, several possible schemes to transport steam coal from the mine to overseas markets have been considered. These schemes are based on coupling ocean transport with pipeline or railroad transport and moving the coal in either dry or slurry form. The first scheme involves transportation from the mine to the preparation plant (size reduction, storage, and pumping), to the pipeline, and to the ship, and from the ship to the pump station/pipeline and to the dewatering plant/thermal drying and to the end-use. The second scheme is similar to the first scheme, except that the coal is dewaterized before being loaded into the ship and on the other shore the coal is turned into slurry again before being shipped through a pipeline. In the third scheme, the coal is reduced in size and washed before it is loaded on the railway. The railroad brings the coal to a maritime terminal, where on the other shore it is again transported by rail to a grinding plant and to the end-use destination.

Coal is exported from the United States through several major coal loading facilities located in Baltimore (Maryland), Philadelphia (Pennsylvania), Mobile (Alabama), Jersey City (New Jersey), Charleston (South Carolina), New Orleans (Louisiana), Newport News (Virginia), and Norfolk (Virginia).

The last two facilities have undergone extensive modernization and new construction to satisfy the increasing demands for export coal. Coal is now shipped in modern vessels, such as Panamax vessels (60,000 dwt [deadweight tonnage]) and Capesize carriers (200,000 dwt or greater).

1.3 Truck Transport

For short distances, trucks play an important role in transporting the coal (or coke to foundries) to a stop station before the coal is shipped to its final destination or to mine-mouth power plants. Due to routing flexibility and low capital investments, trucks can economically move coal approximately 100 miles (at most) one way in relatively small batches. Trucks are used predominantly for short-distance hauls of 50 miles or less. The size of the payload varies according to state gross vehicle weight limitations; for instance, the payload is 20 to 40 tons per truck east of the Mississippi River and 40 to 60 tons per truck west of the Mississippi River.

Basically, there are two types of off-highway trucks: the rear-dump truck and the bottom-dump truck. In a conventional rear-dump truck, the body is mounted on the chassis and lifted up by means of a hydraulic hoist system. The bottom-dump truck, as indicated by its name, dumps material through the bottom and is not well suited to hilly or rough-terrain operation.

Rear-dump trucks have been tested in sizes up to 350 tons; however, 200 tons is considered a large truck and 100 tons is viewed as a standard off-highway truck. Typically, rear-dump trucks carry approximately 25 to 50 tons. Bottom-dump trucks are normally in the range of 35 to 180 tons.

Trucks of up to 85 tons are usually equipped with diesel engines having power shift transmission and torque converters. In the range of 85 to 100 tons, the diesel engine of these trucks could transmit power to the drive wheels by either mechanical or electrical means. At 100 tons and over, the diesel engine drives a generator that produces electricity for electric traction motors that move the wheels.

1.4 Transport by Slurry Pipelines

Pipelines carrying solids such as coal, iron concentrate, and copper concentrate follow the same process: (1) the solids are ground to a fine size, (2) mixed with a liquid vehicle such as water, and (3) pumped through a buried steel pipeline.

Although advantages related to economics, the environment, and simplicity of operation for gas and liquid pipelines are applicable to solid-carrying pipelines, only a few coal pipelines have been built: two in the United States (108 and 272 miles long) and one in the former Soviet Union (38 miles long).

Few coal pipelines have been built because several barriers exist, including long distances, high volumes required, inadequate rail service, and long-term commitment (20 years).

Slurry pipe technology is not new:

1. There are conventional coal water slurries (CWS) and coal water fuels (CWF). In the CWS case, pipelines ship coal at a ratio of approximately 50:50 coal and water by weight in a turbulent flow. In addition to water, other types of fluids can be considered vehicle fluids, such as oil, liquid carbon dioxide, or methanol. The carbon dioxide can be used for enhanced oil recovery, after separation from coal. In the CWF case, fuels are a mixture of coal and water, flowing in a laminar regime, and dry coal weight varies from 60 to 75% depending on the coal type and on the stabilizing additives.
2. In the slurry preparation process, coal slurry suitable for transportation is produced by performing size reduction (crushing and grinding) and chemically treating (corrosion inhibition and thinning) it, if needed, to improve the final characteristics of the product. Two parameters are important for controlling the slurry: density and the product top size. The slurry concentration affects the friction loss of the pipeline and the critical velocity of the flow. The top size is directly related to potential pipeline plugs caused by large settling particles (safety screens are used as countermeasures).
3. To be suitable for transportation, a compromise is made between pumpability and dewatering requirements. If particle size is fine, pumpability is enhanced, but if particle size is too coarse, dewatering at the destination is affected because the slurry must be pumped at higher flow rates to maintain a suspension.
4. To inhibit corrosion in the pipeline, chemical agents, such as sodium dichromate, are added to form a film on the wall, or oxygen scavengers, such as sodium sulfite, are added to prevent attack by dissolved oxygen.
5. A test loop is normally included in the system as the final control. All slurry must pass through it before entering the pipeline. Any increase above the specified limits of pressure drop would indicate the formation of a bed of fast-settling particles in the test loop.
6. For dewatering slurries, continuous vacuum filters and centrifuges are used. For the first U.S.

pipeline, dewatering was performed by thickening, vacuum filtration, and thermal drying.

7. Although the water requirements for coal slurry pipelines are significant, they normally are at approximately 15% of the required power plant makeup water; therefore, the slurry water is commonly used as part of the makeup water in conventional practice.

8. Some advanced transport modes have been considered based on pneumatic and container concepts, e.g., capsules in pipelines. The capsules may be rigid (cylinders or spheres of the cast material), semirigid (flexible container with the material inside), or paste plugs (material mixed with a liquid not miscible with the carrier liquid). These concepts have not yet been developed for commercial application.

1.4.1 Commercial Projects of Coal Slurry Pipelines

Consolidation Coal Company built the first major U.S. coal slurry pipeline from Cadiz, Ohio, to Cleveland, Ohio. This 108-mile pipeline was successful but was mothballed in 1963 after 6 years of operation. The second successful U.S. pipeline is the 273-mile-long Black Mesa pipeline built in 1970 in Arizona. The pipeline has a 5 million ton/year capacity and transports pumped coal slurry (47% solids by weight) through an 18 in. pipe. The Black Mesa pipeline is available 99% of the time. After being "dewatered" by centrifuges, the coal cake from the pipeline is used as fuel in two 750 MW power plants in southern Nevada.

1.5 Economics of Coal Transport Alternatives

Cost comparisons of coal transport alternatives on identical bases are not available. Assuming the best conditions is not appropriate because transport costs are strongly influenced by several factors, such as terrain conditions, sunk cost elements (such as existing railway infrastructure), and inflation.

In the past, several studies have provided some trends:

1. A study performed by Ebasco has indicated that pipelines are most economical for long-distance and high-quantity applications. In this study, the different forms of transporting coal energy have been considered, e.g., by rail, pipeline, rail–barge combination, and extra high voltage power lines, the assumptions being transportation distances of 500–1500 miles and power plant sizes of 1600–9000 MW. If water transport is available, the combination of rail–water transport is competitive, assuming that existing tracks are utilized.

2. The U.S. Department of the Interior conducted a study comparing costs based on the premise that 25 million tons of coal slurry was transported for 1000 miles. It led to the following generalizations:

- Water transport is almost always the most inexpensive way to move coal; it usually involves coupling it with another transport mode depending on the distance between mine and riverway—rail, pipeline, and trucks are possible.
- Truck transport is not competitive with rail or pipeline for high tonnage and long distances (e.g., more than 50 miles and 1 million tons per year). Trucks are the preferable mode for short distances from mine mouth to a terminal (preparation plant or en route to a final destination).

3. A U.S. Department of Energy-sponsored study by Argonne National Laboratory, involving Williams Technologies Inc., identified the potential for U.S. coal exports to the Asian Pacific Basin. High inland transportation costs combined with a relatively low heating value of western coal had resulted in coal prices exceeding world market levels. In a portion of this study, Williams considered how to lower the transportation costs of western coal to the Pacific Basin. The first target is a price of less than $2.05 per MBtu and the second target is $2.9 to $3.25 per MBtu. Four U.S. coal regions/states considered as sources were Four Corners, Utah, Powder River Basin, and Alaska. For Four Corners and Utah, the coal water slurry systems and the coal water fuel systems would go to Stockton, California, and Los Angeles, California; for Powder River Basin, pipeline delivery will go to Puget Sound. For Alaska, two routes were evaluated: (1) from Nenana to Cook Inlet and (2) from Beluga to Tyonek Dock. The findings of the study indicated that Four Corners and Utah, being close to existing ports, provide opportunities for both conventional coal slurries and coal–water fuels, whereas Powder River Basin offers potential for coal–water fuels.

In summary, railroads and slurry pipelines are usually the choices for high tonnage and long distances. Normally, railroad distances are longer than pipeline distances; railroads outperform pipelines when no track must be built, but if new trackage is needed, pipelines are more economical. If inflation is considered, the percentage of transport tariffs subject to inflation is higher for railroad than for pipeline.

2. COAL PREPARATION

Coal preparation is defined as the process of removing the undesirable elements from run-of-mine (ROM) coal by employing separation processes to achieve a relatively pure, uniform product. These separation processes are based on the differences between the physical and surface properties of the coal and the impurities. Coal preparation would include not only sizing (crushing and breaking), but also all the handling and treating techniques necessary to prepare the coal for market. Table III gives a general overview of the characteristics of different coal types offered in the market.

ROM coal is usually a heterogeneous material with no size definition and can consist of pieces ranging from fine dust to lumps; it is often wet and contains rock and/or clay, making it unsuitable for commercial use, such as in combustion. The objective of coal preparation, in terms of combustibility, is to improve the quality of the coal by removing these extraneous noncombustible materials leading to:

1. Ash reduction (lower particulates, reduced ash handling at power stations)
2. Sulfur reduction (lower SO_x emission at power plants)
3. Carbon consumption and NO_x emission reduction (at power plants)
4. Reduced transportation costs per unit of heat
5. A guaranteed consistent heat value (improved plant efficiency) to meet customers' requirements
6. For optimum calorific value (heating value) consistent grindability, minimal moisture, and ash variability, such as in the area of power generation
7. Economical design and operation of steam-generating equipment

TABLE III

Characteristics of Coal Types

Coal type	Characteristics
Lignite	4500 to 7000 Btu/lb (2500 to 3890 kcal/kg) High moisture (25%) with ash
Low-grade bituminous	5000 to 9000 Btu/lb (2780 to 5000 kcal/kg) High ash (+20%) with moisture (10%)
Thermal bituminous	8000 to 12,000 Btu/lb (4450 to 6675 kcal/kg) Medium ash (10 to 12%) with moisture (9%)
High-grade bituminous	+12,000 Btu/lb (+7000 kcal/kg) Low ash (<7%) with low moisture (<8%)

Source. Adapted from Sharpe (2002).

2.1 Size Reduction

Pretreatment, or size reduction, is a common term used to include the processes of breaking, crushing, and screening ROM coal to provide a uniform coal feed with a certain predetermined top size. The number of stages in the size reduction process will depend on the end use of the coal. For instance, in power generation, the number of stages is more than the number needed for coking coal. In breaking the coal, four types of equipment are available: rotary breakers; roll crushers; hammer mills (ring mills), and impactors.

Before describing these four systems, the following terms frequently used in pretreatment should be defined:

1. The top size (or upper limit) of the crushed material is the smallest sieve opening that will retain a total of less than 5% of the sample (ASTM D431);
2. The nominal top size is the smallest sieve opening that will retain a total of 5 to 20% of the sample;
3. The reduction ratio is defined as the ratio between feed top size and product top size;
4. The Hardgrove grindability index is the index determined after passing 50 g of 16×30-mesh dried coal in a standardized ball-and-race mill and sieve this sample through a 200-mesh sieve (ASTM D409) to determine the amount of material going through it.

The four types of equipment for size reduction of run-of-mine coal ([2] and [3]) are described as follows:

1. The rotary breaker causes the coal to break by lifting it to a given height and dropping it against a hard surface. It is a cylindrical unit (operating at low speed) that subjects the coal particles on the liner to undergo centrifugal motion; the liners of the rotary breakers are screen plates that allow peripheral discharge of the properly sized material along the length of the rotary breaker.

2. The roll crusher is the workhorse of coal size reduction to shear and/or compress the material. It is compactly engineered in the form of a rotating roll and a stationary anvil (single roll crusher) or two equal-speed rolls (double roll crusher) rotating in opposite directions.

3. The hammer mill is the most commonly used coal pulverizer. In this device, feed coal is impacted

by rotating hammers and further by grid breaker plates. Usually a high portion of fines is produced. A ring-type hammer mill (rings instead of hammers) would minimize the amount of fines in the product.

4. The impactor impacts the coal, which is then projected against a hard surface or against other coal particles. The rotor-type impact mill uses rotors to reduce the size of the material; the shattered material rebounds in the rotor path repetitively until the product is discharged from the bottom.

The sizing process usually follows the crushing process. The three major reasons for sizing the coal are to separate coal into various sizes for marketing, to feed washing and dewatering units, and to recover the solids used to control the specific gravity in these washing devices. Sizing of the coal is performed by screening or classification.

The most common screening method includes bringing each particle against an opening or aperture where the particle will either pass through or be retained. Several screen configurations have been developed, such as the following: (1) the Dutch sieve bend is a fixed screen with no moving parts and is situated at a level lower than the feed, (2) the Vor-Siv, developed in Poland, and (3) the loose-rod deck, developed in the United States by Inland Steel Corp., which consists of steel rods set at a suitable pitch to create the screening aperture.

Although most sizing of coal is performed with screening devices, classifiers are also in use. Classification refers to achieving size separation by exploiting different flow rates through a fluid. It includes dry classification and wet classification. Dry classification uses air as the ambient fluid and involves higher fluid volumes and higher velocities than those found in wet classification.

2.2 Cleaning Methods

Coal preparation has been evolving for more than a century and the core of most of the process development work is in the area of separating the impurities from the coal, utilizing the physical difference in specific gravity between the coal and the rock impurities. An overview of the different coal cleaning systems is provided in Table IV.

2.3 Coal Dewatering and Drying Systems

After the cleaning process, the coal retains a significant amount of moisture, which can have negative effects on its transport and handling and its calorific value. Therefore, there is a need to reduce the moisture content by either dewatering or drying. Dewatering means mechanically separating water from the coal and drying means separation by thermally evaporating water.

Table V outlines the different technologies used in the dewatering process or in the drying process of coal.

2.4 Arrangement of a Coal Preparation Facility

Given all the above-mentioned techniques for preparing the coal, it is useful to elaborate on the flow of coal within a coal preparation facility through the following stages:

1. Size reduction of coal
2. Sizing of the coal to match the proper feed sizes with the various types of coal-washing equipment
3. Washing the coal
4. Dewatering the products
5. Recycling the process water (treatment/reuse)
6. Blending the product streams into storage
7. Loading out the marketable products

Table VI gives an example of the flow of coal through a coal preparation facility and the equipment involved in the subprocesses.

3. STOCKPILES

3.1 Stacking and Reclaiming of Stockpiles

Coal pile management, an integral and vital element in the coal chain, is carried out at coal mines, coal preparation plants, the transshipment sites for export/import of coal, and end-user sites; end users usually include power plants, iron and steel plants, coking plants, and cement plants. Since these "stockyards" tie up capital with no return on investment, the owners of these plants are interested in optimizing coal inventories to minimize this capital and addressing the issues of stockpiling of coals. The significant issues are stock size, stock pile turnover period, and availability of cheaper coals on the international market. All of these issues involve the best strategy for blending coals with the accuracy required by the customer and just-in-time delivery. Coal pile management includes examining why stockyards are used and how they are designed/erected and operated in compliance with environmental regulations.

TABLE IV

Coal Cleaning Systems

Beneficiation method		Systems
Jigs		• Predominant washing technique throughout the world, as pulse water separators • Two jigs known as Tacub jig (with uniform air distribution underneath the screen) and Batac jig (with improved air valves and control mechanism in addition)
Cyclones	Dense-medium cyclonic separator	• First patent in 1958[a] using metal chloride salts, as first dense-medium cyclone • Dense-medium cyclones effect sharper separation between coal and impurity over a wider specific gravity range than jigs • Several variations (Japanese Swirl cyclone, three Russian cyclone models, British Vorsyl cyclone, and the Dynawhirlpool cyclone)
	Water-only cyclones	Not regarded as dense-medium cyclone that has a much greater included angle in the cone—up to 120°—and a longer vortex finder • Constant velocity separators • Do not achieve a separation as sharp as dense-medium cyclones; nevertheless, provide clean coal products free of high-ash materials in all but the finest sizes • Simplicity/small space requirements/economical operation attributes lead to adoption of water-only cyclones as a rougher device for decreasing the load on downstream cleaning units
Concentrating tables		• One of the oldest most widely used cleaning devices for washing coals • Configurations of two-deck and four-deck systems • Potential to remove pyritic sulfur down to a certain size
Froth flotation (aeration) cells[a]	Flotation air	• Coal particles adhere to air bubbles and form a froth a top the pulp • Froth skimmed off and separated from the refuse that remains in the pulp
Oil agglomeration		• A process of removing fine particles from suspension by selective wetting and agglomeration with oil • Preferential wetting of the coal by the oil; if air admitted, bubbles attach to the oil–wet coal surfaces, leading to increased buoyancy and rapid separation of heavier, unoiled mineral particles
Sand suspension process	Chance process[a]	• Patented in 1917, first applied to bituminous coal in 1925 (specific gravity separations of 1.45 to 1.60 commonly used) • One of the most successful heavy-medium cleaning processes; involves a large inverted conical vessel with sand in suspension in an upward current of water; increasing or decreasing the amount of sand in suspension causes a change in density of the fluid
Magnetite process	Conklin process[a]	• Used powdered magnetite in 1922 • After modification, as Tromp process in 1938 for magnetic medium
Pneumatic process	Air process	• Use air as separating medium

Source. Adapted from Sharpe (2002) and from Deurbrouck and Hucko (1981).
[a] Studies on flotation reagents reported in 1973 and 1976.

With the mounting pressure imposed by economic constraints associated with using smaller stockyards, blending various coals to the accuracy demanded by the customers, and addressing the just-in-time concept, the main functions of the coal pile are very clearly defined: (1) to serve as a buffer capacity to accommodate the differences between supply and demand, thereby acting as a strategic stock against short- and long-term interruptions, and (2) to provide a uniform feedstock of the required quality by accurate blending of coals available on the market.

In the case of a power plant or a process industry, stockpiling involves two important types of storage, namely, live and dead storage. Live storage is defined as "coal within reach" that can be retrieved by using available equipment or by gravity flow. In contrast, dead storage would require using

TABLE V

Technologies for Dewatering of Drying Coal

Method	Technology	Characteristics of technology
Dewatering	Screens	• Dutch State Mines sieve bend is a stationary wedge bar screen with the bars oriented at right angles and across the flow stream; advantageous feature is that the screen generally does not blind • Polish Vor-Siv is a stationary device to handle high volumes of solid in-water slurry; it combines certain characteristics of cyclones, sieve bends, and cross-flow screens • Conventional screens handle less flow volume than Vor-Siv per unit surface area
	Cyclones	• Normally processes a pulp with a low percentage of solids (5 to 20%) and produces a relatively thick suspension by removing a portion of water; solids discharged at the bottom, and water and slimes exit at the top of the unit
	Centrifuges	• Two main types available: solid bowl centrifuges and screen bowl centrifuges; there are many variations of these two types
	Filters	• Most often used is the vacuum disk filter; for best performance, solids in the feed should be coarse • Occasionally used is the drum-type filter; able to operate under varying load resulting from changes in feed rate • Also done is heating the feed slurry (steam filtration) by covering a conventional vacuum filter with a hood and applying steam under the hood to facilitate additional water drainage from the filter cake
	Emerging techniques	• Shoe rotary press developed by Centre de Recherche Industrielle du Quebec • Electro-acoustic apparatus developed by Battelle Columbus Laboratories • Tube filter press developed by English China Clays Ltd. • Membrane pressure filter developed by National Coal Board (England)
Thermal drying	Direct drying	• Before 1958, screen, vertical tray cascade, multilouvers, and suspension dryers dominated the market of thermal drying • After 1958, fluidized bed dryers gained rapid favor over the previous dryer techniques • After 1967, there was a decline in the use of fluidized bed dryer due to increased costs because of stricter emission regulations
	Indirect drying	• Eliminates the potential problem of excessive fines carryover of direct drying systems • Based on contacting the coal with a heated surface; two designs: Torus (or Thermal Disk) and Joy Holo-Flite (variation of a helicoid screw) with oil as heating medium • Another concept (Bearce Dryer) is to heat metallic balls and mix the balls with the wet coal as it travels down a sloping/rotating cylinder

Source. Adapted from Deurbrouck and Hucko (1981) and from Parek and Matoney (1991).

a secondary or an auxiliary means to reclaim this coal. When a coal arrives at a power plant, a decision is made either to add it to live storage or send it to long-term storage in the reserve stockpile; reserve stockpiles ensure against disruptions of deliveries or are opportunities for taking advantage of low prices on the spot market. Buffer stockpiles are sometimes available at power plants and act as a buffer between live and reserve stockpiles. Coals of a similar type can be stacked in the pile, but different coal types/grades are normally stored separately.

3.1.1 Features of Coal Stockpiles

Live and dead storage of large quantities of coal is usually carried out by storing the coal in open stockpiles on the ground, unless stringent environmental considerations require a roof over the stockpile, or enclosures. In this case, silos and bunkers are the alternatives to open piles, but these methods would be more expensive and capacity-limited due to the feasible size of the bunkers or silos.

For an efficient management of the stockpiles, the stockyard layout would satisfy the following requirements:

1. Optimize the land use (tons/acre) consistent with storage and environment constraints;
2. Maintain easy access to stored coal;
3. Maximize the load distance efficiency factor;
4. Position coal such that stacking and reclaiming rates can be achieved with the least effort;

TABLE VI

An Example of a Coal Preparation Plant

Coal preparation facility		
Step	**Interim products**	**Equipment involved**
Run-of-mine storage		• Ground piles • Bunkers • Raw coal blending
Size reduction		• Crushers and rotary breakers
Sizing/dedusting/desliming		• Cylindrical screens • Grizzlies • Vibrating screens
Cleaning method for various types of coal	Option (1) Pneumatic cleaning	• Refuse stream • Product as dry product
	Option (2) Coarse coal washers (+ 1/4 inch)	• Heavy media vessels • Jigs • Refuse stream
	Option (3) Fine coal washers (1/4 inch × 28 mesh)	• Heavy media cyclones • Fine coal jigs • Tables • Water-only cyclones • Refuse stream
	Option (4) Ultrafine coal washers (−28 mesh)	• Water-only cyclones • Flotation cells • Cleaned with coarse coal • Oil agglomeration
Dewatering	Option 2	• Vibrating screens • Centrifuges • Stream going to water clarification • Magnetite recovery circuit (with dryers)
	Option 3	• Sieve bends • Vibrating screens centrifuges, vibrating basket-solid screen bowls • Magnetite recovery circuit (with dryers)
	Option 4	• Raped sieve bend • High-speed screens • Filters • Solid bowl centrifuge
Water clarification (return to circuit use)	Options 2 and 4	• Thickeners and cyclones for option 4 • Refuse dewatering for option 4 with filters, solid bowl centrifuges, filter presses, and ponds
Thermal drying	Options 3 and 4	• Fluid bed dryer • Fine coal dryer
Dry product		• From Option 1
Product load-out		Coming from all the options

Source. Adapted from Speight (1994).
Note. See Fig. 6.1. of Speight (1994), for further details.

5. Attain the desired level of remote control of the equipment and the desired degree of automation of the plant;
6. Meet the homogenization and blending requirements if accuracy is requested;
7. Maintain/improve the uniformity, integrity, and quality of the stored coal;
8. Maximize equipment capability and reliability and minimize manpower requirements;
9. Provide a safe and dependable system consistent with environmental regulations, such as potential stock fires, dust emission, noise levels, and water drainage;

10. Minimize the overall costs (dollars per ton of coal handled) in terms of operation and capital costs.

Characteristics of the land location (in terms of size, shape, and load-bearing capacity) play an important role in defining the choice for the overall storage system of an open pile. These will determine the pile and height of the coal that can be stored.

Several measures must be considered in preparing the ground for a new coal pile:

1. The topsoil must be removed and a stabilizing layer of crushed rock or similar material be laid down;
2. Adequate site drainage must be in place;
3. Utility pipes and conduits should not be run under the pipes;
4. Long stockpiles should have their narrow side facing the prevailing winds to reduce dusting and oxidation.

In addition to the kidney-shaped stockpiles, the most commonly found shapes are the following:

1. Longitudinal beds (or linear beds): the beds are located either side-by-side or end-on. For instance, in the case of two beds, one pile is stacked and the other is reclaimed.
2. Circular beds: they are used for environmental reasons, requiring covering of the stockpiles; many small circular piles are preferred over a single large one for reducing the risk of a breakdown of the reclaimer system and for convenience of maintenance.

In today's business climate, the end user has no plausible reason to invest in excessive coal piles. The drawback of a circular bed is that it has only a limited storage capacity and cannot be increased once built. However, the circular beds eliminate the nonstandard end cones that are formed in the linear beds; these end cones are not representative of all layers in the stockpile; not reclaiming these end cones could cause short-term grade variations when these ends are processed. But the linear beds can store more coal, provided that space is available. The circular beds, which have limited pile capacity and short layer length, will have relatively smaller input lot size but will have many different lots. Capital operating and maintenance costs of circular beds are less than those of linear beds, but the costs for civil engineering work are higher for circular beds.

3.1.2 Methods of Stacking and Reclaiming Coal Piles

3.1.2.1 Stacking Coal During stacking, a stockpile is built with layers of coal, creating a pile with a triangular cross section; the layers may come from the same coal type or from different coal types. In this stacking process, several issues need to be addressed adequately, such as the following:

1. Minimize coal size segregation; the segregation could lead to spontaneous combustion and could contribute to flowslides;
2. Avoid dropping coal from excessive heights, thereby lessening the formation of dust;
3. Take into consideration coals that are more friable, that is, could lead to more formation of fines that, when wet, could cause pluggage of chutes.

Taking these issues into consideration, there are several main methods for stacking coal based on different components and arrangements. These methods are further based on the capability of the different types of stackers. The stacker runs on rails laid along the total length of the bed. The capability of the stacker will partly determine the ability to switch from one stacking method to another. The stacking equipment for longitudinal stockpiles is designed to reflect the different amounts of flexibility, the varying ability to blend/process the coal types, and the range of investments. To address these features, there are several types of stackers: (1) stacker with rigid boom; (2) stacker with luffing boom; (3) stacker with rigid and retractable boom; (4) stacker with luffing, rigid, and retractable boom; (5) stacker with slewing boom; (6) stacker with luffing and slewing boom.

In the cone shell method, a pile is formed by stacking coal in a single cone from a fixed position. When this cone is full, the stacking system moves to a new position to form a new cone against the shell of the previous cone. This method is not efficient for blending coals and is used in cases where homogenization is not required. The strata method stacks the coal in horizontal layers and will create alternating layers when two or more coals are to be blended; each layer will extend from one end of the bed to the other end and the whole bed should be completed before reclaiming is performed from the end of the pile. The skewed chevron is the strata method with the difference that the layers are inclined and reclaiming is carried out from the side of the bed.

The chevron, windrow, and chevron–windrow systems are all intended to have coal reclaiming carried out from one end of the bed and they can achieve a high blending efficiency. The chevron system is the simplest system since it requires only a stacker with one discharge point: the stacker moving along the central axis of the pile. The drawback is that segregation of material will lead to fine particles depositing in the central part of the pile and coarse particles depositing on the surface and at the bottom of the pile. The windrow and chevron–windrow methods will require more expensive stackers with a slewable boom allowing multiple discharge points; both of these methods minimize particle segregation. The Chevcon method (continuous chevron) is a combination of the cone shell and chevron methods and utilizes a circular rather than a linear bed. This circular bed has a round base with a ring-shaped pile where coal is stacked at one end and reclaimed from the other end.

3.1.2.2 Coal Reclaiming The technology of equipment for reclaiming coal has evolved from a discontinuous process originally to a continuous process, with the emergence of bucket-wheel, bridge-type, and portal scraper reclaimers.

In North America, the first bucket-wheel stacker reclaimer was installed in the Potomac Electric Power Plant at Chalk Point (Maryland) in 1963. Prior to this reclaiming method, coal stockpiles at large power plants relied on gravity systems that moved coal on a belt conveyor in a tunnel beneath the pile. The bucket-wheel stacker reclaimer that operates in the United States has two additional features not available in the early European version: (1) a movable counterweight instead of a fixed one (safety hazard eliminated) and (2) complete automation (labor-saving feature).

These reclaimers have become popular and make up approximately 90% of the total coal applications. The ability to perform homogenization/blending efficiently varies from one type to another. They reclaim different coal types stored in separate beds in a sequence as defined by the blending process required to meet a customer's expectations. This key feature is helpful at transshipment facilities tasked with blending imported coals.

A description of the types of reclaimers in light of their key features and functionality is given below:

1. The bucket-wheel recalimer is designed with a bucket wheel to reclaim the coal and operates from the side of the bed. This type of reclaimer is suitable only for operation in the open (due to the counterweight); it is deployed in stockyards with high throughput. Its homogenization/blending efficiency is not good. The bucket wheel is often combined with a stacker.
2. The gantry-type reclaimer may have multiple bucket wheels to improve the homogenization/blending efficiency.
3. The drum-type reclaimer uses buckets attached to a long rotating drum to reclaim the coal across the entire width of the pile, resulting in an improvement of the homogenization/blending efficiency.
4. The bridge-type scraper is designed with a scraper chain to reclaim coal across the front face of the longitudinal pile. This reclaimer has a very good blending/homogenization efficiency, but it cannot jump over stockpiles (whereas side-face reclaimers can).
5. The portal, semiportal, and cantilever scraper systems reclaim coal from the side of the bed. As side-face reclaimers, they are suitable for removing coal with signs of imminent spontaneous combustion.
6. The portal bridge scraper combines the advantages of the side- and front-face scraper reclaimers. A special version has been developed to reclaim circular beds; this special version is based on a bridge-type scraper and a cantilever scraper reclaimer, both mounted on a central column.
7. Front-end loaders and bulldozers are used for smaller stockpiles to stack and reclaim coals.
8. The gravity reclaim is utilized in some stockpiles. This system uses hoppers (located in a tunnel below the stockpile) and feeders (to control the discharge into the conveyors) to reclaim coal. Vibratory stockpile reclaimers would improve the reclaiming process. The use of multiple hopper systems (allowing intersection of flow channels) would improve the reclaiming performance. Blending efficiency within the stockpile is poor.

In addition to the reclaimers, some other types of mobile handling equipment play a role in stacking/reclaiming coal at a smaller volume than the above-mentioned reclaimers do. Front-end loaders and bulldozers are used to compact the coal and to move coal from reserve pile to live pile.

3.1.3 Bins and Silos

Coal can be stored in silos or bunkers as an alternative to open piles on the ground or in areas with enclosures. However, this method of storing coal in silos or bunkers is more expensive and constrained by the feasible size of the units.

Silos and bunkers are either steel or concrete, usually have round cross sections, and can have single or multiple outlet cones. An internal lining of stainless steel is a big advantage when storing high-sulfur coals to prevent corrosion. Regarding structural integrity of the foundations, loads should not be offset on the foundations when eccentrically charging or eccentrically withdrawing coal from these units.

A silo with a large diameter and a flat bottom will have multiple outlets. Bunkers with a height-to-diameter ratio of 2:1 to 3:1, designed to be self-cleaning, will have a cone bottom (with a suitable angle, e.g., a suitable flow factor) and a single outlet.

These coal storage devices do offer unique benefits such as: (1) controlling the dust and moisture content of the coal, (2) controlling runoff from rainwater, and (3) providing blending capacities by having several units feeding a single reclaim conveyor.

3.1.4 Trends in Management of Coal Piles

Traditionally, the coal stockyard for U.S. power plants contained approximately one-fourth of the plant's annual burn, such as the 90- to 120-day storage level at several American Electric Power plants, but that level in the United States has declined to much lower coal inventories. In the United States, stockpile levels have decreased to a level of approximately 50 days since mid-1993.

In comparison, some power plants have even kept 7- to 10-day coal inventories when their plants were located in the vicinity of their mines, such as the case of the South African Kendal mine-mouth power plant.

With a just-in-time delivery system, some power plants have coal stockpiles of only 5–8 days, such as those in the Netherlands or in U.S. centralized plants. Relative to just-in-time delivery, the decision to buy is driven more by demand than lower on-the-spot coal prices. As such, the decision could be subject to increases in coal prices due to regional coal shortages. These shortages could affect several utilities drawing on the same pool of mines. Consequently, before reducing its coal inventory, a power plant would need to carefully consider many factors such as: (1) transport options (rail, barge, truck, and ROM conveyor) in terms of reliability; (2) number of supply sources; (3) backup sources, such as other power plants; (4) reduced burn rate; and (5) delivery time for coal from supplier to the power plant (in terms of days).

Another trend in management of coal stockpiles is outsourcing; this concept is suitable for mine-mouth plants or power plants served by a single supplier.

3.2 Sampling, Analysis, and Auditing

With larger, more expensive shipments of various compositions, coal must be sampled and analyzed before storage for the following reasons:

1. Sampling is performed during loading operations at terminals or at power plants to ensure that the coal meets the contract specification (penalty and bonus clauses related to coal properties).
2. Sampling leads to knowledge about the coal consignment; this information allows the coal operators to store coal in different piles—coals of similar grade are stockpiled together and blending of different types of coal can be performed based on this knowledge.
3. Sampling during the reclamation process helps evaluate changes in properties during the storage period.

Given the above-mentioned reasons, a good sampling technique is critical for obtaining a truly representative sample of the coal consignment and of paramount importance for the analysis of the coal. The usual procedure is to divide the sample into three parts—one for the supplier, one for the customer, and one for independent analysis in case of a contract dispute.

3.2.1 International Standards for Coal Sampling Techniques

Coal is one of the most difficult materials to sample because it has the tendency to segregate by size and mass. The ideal objective of a sampling would be to have every particle size represented—ash content and heating value may be different for each particle size.

Other factors in the sampling process are the use of analytical results, the availability of sampling equipment, the quantity of the sample, and the required degree of precision according to standards (ASTM D2234).

Since biased results can be introduced into the sampling procedure, certain precautions to reduce the sources of bias are used, such as: (1) selecting the most suitable location for sampling purposes; (2) using sampling equipment that conforms to necessary specifications; and (3) taking special precautions when performing specific sampling procedures, such as sampling for total moisture content or size.

Table VII gives an overview of the international standards on sampling and sampling procedures. These standards are applicable only for sampling coal from moving streams.

In general, the standards specify the number and weight of the increments to be taken for each sampling unit. The increment is defined as a small portion of the coal lot collected in a single operation of the sampling device; the higher the number of increments, the greater the accuracy. A greater number of increments is recommended when sampling blended coal. The individual increments can be evenly spaced in time or in position, or randomly spaced, and taken out of the entire lot. The increments are then combined in a gross sample; the gross sample is then crushed and divided, according to standard procedures, into samples for the analysis. In the analysis part, standards differ in detail and often do not take into account the characteristics of the coal being sampled. These coal characteristics would include particle size distribution, density, particle shape of the coal, and the distribution of the parameter of interest. Consequently, sampling models have been developed to supplement these standards.

3.2.2 Sampling Systems and Bias Testing

Sampling in a stockpile *in situ* presents an enormous challenge in terms of obtaining a representative sample given the following:

1. Since the coal is subject to weather/oxidation and to size segregation, its quality is always different at the various locations of the pile (top, sides, or in the middle).
2. With many additions of different coals to the pile, the pile may contain coal zones that vary markedly.
3. Because sampling coal in the middle of the pile is difficult, sampling from a moving stream during the stacking or reclaiming process is preferable; however, if sampling of stockpiles *in situ* is necessary, the sample should be taken by drilling or augering into the pile.
4. Drill sampling and auger sampling have limited applications. The drill sampling method is used to penetrate the full depth of the pile at each point sampled and extract the whole column of coal. However, this technique does not collect all of the fine material at the bottom of the pile, because of the drill or auger design. Because this technique can

TABLE VII

International Standards on Sampling Coal from Moving Streams

Organizations	Standards
American Society for Testing and Materials (ASTM)	• D2013 Preparation of samples for analysis • D2234 Collection of a gross sample • D4702 Inspection of coal-sampling systems for conformance with current ASTM standards • D4916 Mechanical auger sampling
Standards Australia (AS)	• AS 4264.1 Coal and coke—Sampling; with Part 1 of sampliing procedures for higher rank coal • AS 4264.3 Coal and coke—Sampling; with Part 3 of sampling procedures for lower rank coal • AS 4264.4 Coal and coke—Sampling; with Part 4 of determination of precision and bias
British Standards Institution (BSI)	• BS 1017: Part 1 as sampling of coal and coke; methods for sampling of coal revised to adopt ISO/DIS 13909
International Organization for Standardization (ISO)	• ISO 1988 Hard coal—Sampling • ISO 5069/1 Brown coals and lignite—Principles of sampling; Part 1 of sampling for determination of moisture content and general analysis • ISO 9411-1 Solid mineral fuels—Mechanical sampling from moving streams; Part 1: Coal • ISO/DIS 13909-1 Hard Coal and coke—Mechanical sampling; Part 1: General Introduction • ISO/DIS 13909-2 Hard coal and coke—Mechanical sampling; Part 2 of coal—Mechanical sampling from moving streams • ISO/DIS 13909-3 Hard coal and coke—Mechanical sampling; Part 3 of coal—Mechanical sampling from stationary lots • ISO/DIS 13909-4 Hard coal and coke—Mechanical sampling; Part 4 of coal—preparation of test samples • ISO/DIS 13909-7 Hard coal and coke—Mechanical sampling; Part 7 of methods for determining the precision of sampling • ISO/DIS 13909-8 Hard coal and coke—Mechanical sampling; Part 8 of methods of testing for bias

Source. Adapted from Carpenter (1999).

break the particles, it is not recommended when collecting samples for size analysis and bulk density determination.

As mentioned above, standards specify the method and the number of increments to be collected. The method of collection is particularly important in terms of sampling location and tool. Standard scoop tools allow every coal particle size to have an equal chance of being sampled; spades are inappropriate and give a sample that is mainly fines. Regarding sampling location, coal can be collected while it is transported on the belt conveyors or to and from the stockpile.

The most accurate sampling method is the stopped-belt or reference method (a full cross section of the flow would be obtained). If stopping the belt is impractical or uneconomical, sampling could be done manually either from the moving belt or from a falling column of coal at a certain point. For high tonnage rates, sampling should be done by mechanical samplers. Some standards, such as BS 1017 and ISO 1988, stipulate that manual sampling should not be exercised in cases where the nominal coal top size is above 63 mm or when the coal flow rate is greater than 100 tons per hour.

Mechanical samplers provide a representative sample that is difficult to attain with manual sampling. The advantage of mechanical samplers is that statistics of precision and bias are applicable. However, these mechanical samplers cannot be applied to all situations, such as sampling within a stockpile or at facilities with small tonnages. The common characteristics of the mechanical samplers are as follows: (1) they provide a full cross-sectional representation instead of a partial cross-sectional sample, (2) their design and operation are covered in various standards, (3) their functionality is to avoid separation of the various coal densities and/or sizes by minimizing disturbances of the coal, (4) they require sufficient capacity of coal to perform sampling without any loss or spillage, (5) they require material to flow freely through all stages of the sampling system without clogging or losing material, and finally (6) the standards specify the design and the number of increments to be collected and the size of the cutter opening (typically three times the coal top size) and its speed.

In terms of categories, these mechanical samplers are divided into two types:

1. Cross-belt samplers (sweep arm or hammer samplers) sweep a cross section of the coal on the moving conveyor (or belt) into the hopper. For the sake of accuracy, these devices need to be adjusted so that no coal fines are left on the belt.
2. Cross-stream (or falling-stream or cross-cut) cutter samplers collect a cross section of a freely falling stream of coal. The location of this sample collection is typically a gap at a transfer point, such as between two conveyor belts.

All sampling systems should be checked for bias, because systematic errors may be introduced. These systematic errors generally are caused by a loss or gain in the mass of increments during collection or by cyclical variations of coal quality at time intervals coinciding with systematic sampling time. Bias testing is discussed in certain standards, such as ISO/DIS 13909 Part 8 for bias testing of mechanical samplers and in ASTM D4702 with guidelines for inspecting cross-cut, sweep arm, and auger mechanical sampling systems.

3.2.3 On-line Analyzers

Obtaining representative samples of coal from large stockpiles is burdensome and time-consuming when laboratory analysis of these samples based on standards is applied. The results do not reflect the variations in coal quality. Real-time information about these variations would help stockyard operators manage their assets much more efficiently. In response to this problem, on-line analyzers have been developed to continuously determine the coal quality and in real time.

Advantages of on-line analyzers over conventional analyzers are obvious when it comes to the following: (1) determination of coal quality in real time, thereby allowing the performance of trend and variability analyses; (2) simplification of sampling requirements in the context of large coal quantities to be analyzed; and (3) determination of coal quality in tons per hour instead of grams per day.

On the downside, on-line analyzers are expensive because they are too site-specific and application-specific. Their performance is related strongly to the initial installation, calibration, maintenance, and application environment. Analysis of coals beyond the initial calibration will not produce the same accuracy. Calibration drifts over time result in the need for recalibration of the analyzer.

International standards (such as ISO CD 15239, ASTM, and AS 1038.24) do not provide particular test methods because of the multitude of on-line analyzers and their interaction with sampling/analysis systems. On-line analyzers have gradually reached

the level of precision and reliability to be used in several commercial applications:

1. Monitoring the incoming coal to determine whether it meets specified requirements, as is done at the Herdrina plant in South Africa;
2. Sorting and segregating coal into different stockpiles (as is done at the Rotowaro coal handling facility in New Zealand);
3. Blending coal from different stockpiles to meet customer requirements (the Rotowaro coal handling facility);
4. Monitoring coal during reclamation to check for the desired specifications.

On-line analyzers available on the market have been developed based on various technologies, such as: (1) electromagnetic principles, (2) X-ray fluorescence, and (3) capacitance/conductance methods. Table VIII lists the application areas for these on-line analyzers based on the principle of operation.

All of these on-line coal analysis techniques are capable of meeting ASTM measurement standards for a predetermined range of coals. Table IX gives an overview of the four main on-line measurement techniques that are in use.

3.2.4 Stockpile Audits

Coal pile management also involves coal stock audits; the intent is to adjust the book inventory. In this task, the stockpile tonnage (calculated by measurements of volume and density), as the measured inventory, is to be reconciled with the book inventory.

In the book inventory, recorded weights of coal going in and out of the stockpile are documented.

Consequently, accurate weighting systems are important to gain better control over the stockpile and eliminate overpayment; it is recommended that the systems are properly situated, installed, maintained, and regularly calibrated. By knowing more precisely how many tons of coal are in the stockpile, one can write off coal used more quickly and/or can reduce inventories that are too large and thereby achieve significant savings.

Both the book inventory and the measured inventory are subject to measurement errors. Whereas inventory is subject to errors coming from volume and density measurements, several sources of errors are associated with the book inventory: (1) weight errors associated with the stockpile; (2) unaccounted changes in coal moisture in the stockpile; (3) losses due to spontaneous combustion; (4) losses due to runoff and wind erosion; and (5) losses due to contamination of the coal at the base of the pile in contact with the underlying strata.

Because stockpiles have various shapes and sizes, measurements must be used to determine volume. Both ground and aerial surveys are used to determine volume.

In an aerial survey using a photogrammetry method, two independent photogrammetrists using the same photograph on stockpiles between 50,000 and 500,000 tons can obtain accuracies of $\pm 5\%$.

For ground surveys, an electro-optical distance ranging system has been developed. It uses a theodolite (based on a laser) for distance measurement. The theodolite applies numerous point measurements to provide three-dimensional information about the surface of the pile. Using density contour information, specialized computer software computes the mass of the stockpile.

For density determination, the methods used would include volume displacement procedures and nuclear surface and depth density determination. Among these methods, a nuclear depth density gauge is probably the most dominant method, although the procedures have not yet been standardized. The gauges take horizontal and vertical density measurements at different locations and depths within the pile. However, these gauges require calibration for

TABLE VIII

Applications Areas for On-line Analyzers

Principle of operation	Suitable for measuring	Energy source
Prompt gamma neutron activation analysis	S, Btu, Ash, H, C, Si, Fe, N, Cl, Al, Ti, Ca	Californium-252
Gamma-energy transmission	Ash	Barium-133, cesium-137, americium-241
Gamma-energy backscatter	Ash	Americium-241
Microwave energy attenuation	Moisture	Microwave generator

Source. Adapted from Connor (1987).

TABLE IX

Overview of Four Main On-line Measurement Techniques

On-line analyzer technique	Characteristics
Natural gamma systems	• Requires no radioactive source • Measures the gamma emission from the conveyed coal • Combines this measurement with a measurement of the weight of the load to calculate the ash content • Not the most accurate system but the least expensive
Dual-energy gamma-ray transmission systems	• Combines measurements of the intensity of two narrow beams of high-and low-intensity gamma rays passing vertically on the conveyor belt to determine the ash content; coal must be mixed well • Instrumentation to split the beam and with multiple detectors has been developed to determine ash content across the full belt width • Varying chemical composition, especially iron content, can cause inaccuracies; accuracy is better for low-ash coals • Triple-energy gamma transmission systems developed for improving accuracy
Prompt gamma neutron activation analysis (PGNAA)	• Provides an elemental analysis of coal through measurement of the gamma radiation emitted when coal is exposed to a neutron source (such as californium-252) • Extra radiation shielding required for safety precaution and sophisticated signal-processing equipment needed to interpret the gamma-energy signals • Energy of the gamma ray is characteristic of the emitting element; gamma rays detected and collected in a spectrum and later correlated to coal composition • Interpretation of the spectrum is a complicated process; each manufacturer has a corresponding deconvolution process • Carbon, hydrogen, sulfur, nitrogen, and chlorine are measured directly • Ash content measured indirectly by its elements (mainly silicon, iron, calcium, aluminum potassium, and titanium) • Calculation of heating value possible if moisture analyzer is present • Instrumentation with multiple sodium iodide detectors developed to deal with coals from multiple sources • PGNAA is the most accurate analyzer but is also significantly more expensive than gamma analyzers
Microwave moisture meters	• Determine moisture content by measurement of energy attenuation and phase shift of microwaves through coal—Moisture of coal absorbs energy by a "back and forth" rotation of the water molecules in tune with the microwave frequency • Often incorporated in dual-energy gamma-ray transmission or on PGNAA analyzers to enable calculation of heating value • Measurement of coal density or inference of coal density required, usually provided by a high-energy source similar to that in gamma-energy attenuation ash analyzers • Not applicable for measuring frozen coal because ice and coal matter have similar energy-absorbing properties

Source. Adapted from Carpenter (1999) and from Connor (1987).

each type of coal and therefore encounter problems when measuring a stockpile with different types of coals. In addition, their accuracy depends on penetrating the coal without disturbing the *in situ* density and they are expensive to acquire.

3.3 Coal Deterioration

During storage in open stockpiles, most coals are subject to weathering and atmospheric oxidation; these processes alter the properties and structure of the stored coal and could render the coal less valuable for its intended use. Assessing the effects of weathering and atmospheric oxidation of the stored coal would mean examining the following key factors: (1) heating value losses, (2) handling related to coal moisture content, (3) coal cleaning (coal flotation), (4) combustion behavior, and (5) coking behavior.

3.3.1 Heating Value Losses

Oxidation can reduce the heating value of coal. Coal is affected by several factors, such as the characteristics (composition, rank, and particle size) of the

stored coal, the storage conditions (temperature and moisture content in the stockpile, storage time, location in the depth of the pile, and ventilation pattern through the pile), and the climatic conditions. The oxidation rate is coal-type-dependent. Low-rank coals are more susceptible to weathering and atmospheric oxidation than high-rank coals. Smaller particles, given their increased surface area, tend to oxidize more rapidly than larger particles. The heating value of the coal decreases with an increase in ash content. Also, the temperature and moisture content in the stockpile affect the rate of coal oxidation and hence the heating value. In a stockpile, the heating value losses decrease with depth; these losses are highest at the surface of the pile where coal comes in intimate contact with the atmosphere. Local hot spots can lower the heating value, due to the airflow through the pile. Compacted coal piles have less heating value losses than uncompacted piles.

Climatic conditions cause coal to lose or gain moisture. Climatic factors, such as solar intensity and rainfall frequency, cause heating value losses. Coal with ice formation on the surface can reduce the oxidation rate.

3.3.2 Coal Handling—Coal Moisture Content

The performance of the coal handling equipment is strongly affected by the moisture content of the coal. In open stockpiles, weather conditions (heavy rainfall or freezing) can increase the coal moisture content or freeze the coal so that equipment failures occur in belt conveyors, chutes, and bins. In hot and dry climates, through moisture loss by evaporation, the stored coal becomes more friable and dustiness increases. A high percentage of fines in combination with high moisture content can lead to a high potential for pluggage. In all these cases, weathering and oxidation increase the friability of the coals and the production of more fines.

Several practices at the stockyard should be avoided. Others are beneficial to control the coal moisture content of the pile. The practices that lead to increased moisture content and that should be avoided are the following: (1) creating large stock piles with little, if any, runoff; (2) allowing "ponding" on the surface of the pile; (3) moving coal through low areas with standing water during the reclaiming process; and (4) having pile levels lower than groundwater levels.

The practices that are beneficial to controlling the coal moisture content include the following: (1) stacking the pile with a maximum height over the smallest area, consistent with stable side slopes; (2) avoiding flat tops for the piles; (3) compacting the surface and the side slopes of the stockpile to enhance runoff; (4) developing a good strategy for efficient drainage; and (5) reducing the moisture content of the incoming coal to the maximum level practical at coal preparation plants.

Other practices include adjusting the equipment to handle wet coal by inserting plastic liners in the chutes and by using intermittent air blasts and vibrators. Chemicals should not be added since they rarely can be applied in sufficient quantities and mixed well enough to be effective. In addition, these chemical treatments affect the end-use behavior of the coal.

3.3.3 Coal Cleaning—Flotation

Coal cleaning processes, such as flotation, are negatively impacted by weatherized or oxidized coal. No method for efficiently cleaning the coal has been found.

Acidic oxidation products on the surface of the coal have dissolved in the process water and lowered the pH. Consequently, this has led to a reduction in flotation recovery and equipment wear. In response, wear-resistant material has been used for surfaces exposed to these adverse elements and magnesium hydroxide is added to raise the pH of the process water. Blending the oxidized coal (up to 20%) with fresh coal to dilute these adverse effects has been practiced.

Another adverse effect of the oxidation process is that oxidation causes a change in the wettability of coal that consequently affects the surface-based separation processes (flotation and oil agglomeration).

The performance of these separation processes decreases with the storage time. As remedies to improve the performance of these coals, several measures have been carried out:

1. neutralizing these adverse acidic oxidation products with alkalis;
2. using surface conditioners (or collectors);
3. using cationic collectors (amines);
4. using low-concentration electrolytes (iron chlorides);
5. grinding the stored coal to −28-mesh (standard flotation size) before subjecting it to flotation;
6. washing the stored coal with water to remove the dissolved metal ions prior to the flotation process.

3.3.4 Impact on Combustion Behavior

Coal stockpiles stored in the open for a long period of time undergo changes that affect their combustion

behavior and consequently affect boiler operations. The NO_x emission levels, carbon burnout, and slagging behavior could be impacted by changes in swelling index, maceral composition, and chlorine. Moisture increase in the stored coal would have significant consequences in the boiler plant in terms of the ability to handle the coal, mill capacity, and boiler efficiency.

In addition to lowering the heating value of the stored coal, atmospheric oxidation of the coal causes a delay in devolatization, leading to a change in coal reactivity. Stored coals with high free swelling index (FSI) have shown a significant loss in the FSI value and hence a loss in reactivity. A decrease in FSI and reactivity also means the appearance of oxidized vitrine, which is found to produce compact char structures requiring long combustion times.

The loss-on-ignition (LOI) directly related to char burnout increases significantly when weathered coals are fired. Any increase of unburned carbon in the fly ash, as the consequence of LOI, would affect the sale of the fly ash. To ensure a satisfactory burnout of the char, a longer residence time in the furnace burning zone is needed.

In general, to minimize the above-mentioned negative effects on the performance of the boiler (power plant), several measures have been put into practice: (1) modifying the plant operating parameters; (2) blending the oxidized coals with unoxidized coals; and (3) minimizing the storage time of sensitive coals.

3.3.5 Coking of Coal

Weathering and oxidation have several negative effects on the coking of coal: (1) reduction in coke stability; (2) issues of coal bulk density (related to particle size distribution and moisture content); (3) reduced coke production; (4) overheated charges; (5) carbon deposits; (6) oven damage, such as to coke oven walls; (7) generation of fines; (8) increased coke reactivity; and; (9) decreased coking rate.

Bulk density decreases with decreases in particle size distribution and increases with additional moisture over 8%. To produce a good metallurgical coke, two properties are important: good thermoplastic and caking properties. Deterioration of these properties due to weathering/oxidation of the coal in open stockpiles would affect coke production. Alvarez *et al.* presented the influence of natural weathering on coal in two papers in 1995 and 1996. The study was performed on a typical blend of 13 different coals in a 100-ton open stockpile in Spain during a 1-year period. Four of the component coals (in piles of 50–60 tons) were stored without grinding (unlike the blend) and were exposed to weathering for a period of 5 to 6 months. With mild weathering conditions, results based on the test parameters of the free swelling index, Gieseler maximum fluidity, and Audibert-Arnu dilation demonstrate that there was no significant change in free swelling index for the blend for each of the four coals, except for one coal after 138 days, and the Gieseler maximum fluidity factor (as the most sensitive indicator for loss of thermoplastic properties and oxidation) decreases over time for the blend and the four coals.

Another result of the study was that weathering/oxidation does not necessary lead to deterioration of coke quality. Mild weathering and oxidation can help produce a coke of improved quality in high-volatility and high-fluidity coals and in some medium-volatility coals.

Other studies indicated that the free swelling index generally is not sensitive to slight-to-moderate weathering, although there are exceptions. Studies also suggest that moderate caking coals (FSI between 3 and 5) are more prone to lose their caking properties than coals with higher FSIs.

With the increased trend to use pulverized coal injection into the tuyeres of steel blast furnaces and to blend coals, coke use has declined. Because coke strength and reactivity are especially important, steel works are very interested in having the right coke blend.

3.4 Spontaneous Coal Combustion

Spontaneous coal combustion is an important concern due to its effect on safety, the environment, economics, and coal handling. Economic losses occur because of fires in the coal pile and subsequent unsuitability of the coal for its intended use. Even if spontaneous combustion does not happen in coal piles, fires can occur later in confined spaces, such as on ships or in railroad cars. Ship fires, such as those reported on coal shipments from New Orleans, Louisiana, to the Far East in 1993, have led to the practice of not loading coal at temperatures greater than 104°F (40°C).

Although coal producers cannot always control the factors that govern spontaneous combustion, they can control management of the coal piles.

The tendency of coal to heat spontaneously in storage is related to its tendency to oxidize. This oxidation process is closely related to the rank of the coal (the higher the rank, the less potential for oxidization), the size of the coal in the pile, the

stacking method, the temperature at which the coal is stored, the external heat additions, the amount/size of pyrite present in the coal, the moisture content, and the storage conditions (in terms of oxidation heat release and heat dissipation by ventilation).

In lower rank coals, the amount of carbon is low and the amount of oxygen is high. In the higher rank coals, the reverse is true. Hydrogen remains relatively constant. Sulfur in the coal is present in the form of iron-free sulfur, organic sulfur, pyrite, and sulfate in various percentages, ranging from less than 0.5 to 8%.

Spontaneous combustion is the result of oxidation of the coal when heat generated by complex chemical reactions exceeds the rate of heat dissipation. Of the coal types, subbituminous coals and lignite require special handling to prevent spontaneous combustion. The lower the rank of the coal, the more easily it can absorb oxygen. In addition, the low-rank coals are higher in moisture, oxygen content, and volatile matter, all of which contribute to rapid oxidation. Also, coal can gain or lose moisture during storage. When dried or partially dried coal absorbs moisture, heat is generated (heat of wetting) and the temperature of the coal can increase significantly if this heat is not removed. Consequently, spraying water to extinguish fires in coal piles is not the solution, but instead chemical dust suppressants could be used.

Other factors that can contribute to coal oxidation are coal porosity, particle size, segregation during stockpile construction, and mineral content and composition. A more porous coal is more susceptible to oxidation because of the larger number of exposed areas. The effects of particle size are linked to coal porosity; a decrease in the particle size provides a greater surface, and hence a higher oxidation effect.

Regarding particle size segregation, differently sized coals should not be stacked. During construction of the pile, size segregation can occur in conical piles when there is a range of particle sizes; the outer area of the pile contains large-size particles and the middle of the pile has fine particles. This will lead to a so-called "chimney effect," whereby air passes up freely from the outer to the inner areas of the pile, creating the potential for heat production. In terms of mineral content and composition of the coal, some mineral elements can accelerate the spontaneous heating process. Pyrite, sodium, and potassium (in the form of organic salts) can act as catalysts for spontaneous combustion.

3.4.1 The Five Stages of Spontaneous Combustion

Although opinion differs on the exact mechanism of oxidation, the process of oxidation occurs in five steps:

1. Coal starts to oxidize slowly until a temperature of approximately 120°F is reached.
2. The coal–oxygen reaction rate increases at an increasing rate and the coal temperature increases until the temperature reaches 212 to 280°F.
3. At approximately 280°F, carbon dioxide and most of the water are given off.
4. The liberated carbon dioxide increases rapidly until a temperature of 450°F is attained; at this stage, spontaneous combustion may occur.
5. At 660°F, the coal ignites and burns vigorously.

For pyrite, the chemical reaction of FeS_2 with water and oxygen leads to the formation of iron sulfate ($FeSO_4$) and sulfuric acid (H_2SO_4) and is an exothermic reaction. In an open pile, pyrite, through its transformation to bulkier materials, is sometimes responsible for causing slacking and forming fines. Consequently, coal operators are reluctant to store high-sulfur coals for extended periods of time, but coals with low sulfur content are no guarantee for safe storage either.

3.4.2 Safe Storage of Coal through Prevention

Coal producers and users must know the self-heating propensity of stored coals. Practices for long-term storage include the following:

1. The storage area should be dry and on solid ground or sloped to allow good drainage of any pockets of water.
2. The area should be cleared of any debris and combustible materials with low-ignition temperatures, such as wood, oil-soaked rags, paper, and chemicals.
3. The area should be void of any buried steam lines, sewer lines, hot water lines, and other heat sources—stockpiles should not be added to on hot, sunny days.
4. The formation of conical piles should be avoided because larger pieces tend to roll away, creating a chimney effect for the fines situated at the center; for conical piles, removal should occur according to the principle of "first-in, first-out."
5. The windrow method is preferred over the chevron method for stacking coals that are prone to spontaneous combustion.

6. Air must be able to circulate freely in the pile or the pile must be packed in 1 ft layers or less to ensure the absence of air circulation (either restrict the entry of oxygen into the pile or provide adequate ventilation to remove the generated heat).
7. Upright structural supports in or adjacent to the storage piles could create chimney effects.
8. Particle-size segregation should be avoided.
9. Stoker-sized coals may be treated with oil before storage to slow the absorption of oxygen and moisture. Road tar, asphalt, coal tar, and various surfactant dust-control binders have been used as a seal; among these, asphalt is unsatisfactory in some applications.
10. Seals of compacted fine coal may be used to minimize the undesirable formation of fines if the coal has a tendency to slack. Chemical inhibitors have been used to reduce the reactivity of coal; these inhibitors include phenol, aniline, ammonium chloride, sodium nitrate, sodium chloride, calcium carbonates, and borates.
11. Good access should be available around the perimeter of the pile to allow the removal of hot spots.

3.4.3 Detection and Control

Early detection of conditions that cause spontaneous combustion is critical for safe storage of the coal and elimination of conditions, such as subsurface voids, that can endanger operators. The cost and effort to monitor stockpiles are generally small compared to the value of the coal piles. Spontaneous combustion is often detected visually by smoke or steam coming out of the pile or by the melting of ice or snow at certain locations in the pile. In hot weather, a hot area is identified by the lighter color of dry coal surfaces due to the escaping heat, a situation that requires immediate action. To detect/monitor the coal self-heating process, two means are available: (1) temperature measurements, through metal rods, thermocouples, and infrared detectors; and (2) gas analysis, such as monitoring the carbon dioxide, hydrocarbon, and hydrogen concentrations emitted.

Once self-heating is under way, two basic control options are available:

1. Removing the hot coal and using it immediately; if it is too hot for the transport equipment, it should be cooled first.
2. Removing the hot coal, cooling it by spreading into thin layers, and repiling and recompacting it; the coal can be left overnight to cool or can be sprayed with water. However, wet coal is difficult to handle.

Water is generally not recommended for extinguishing fires in coal piles. Dry ice (solid carbon dioxide) has been used successfully to control spontaneous coal combustion in an enclosed or covered area. However, dry ice is not recommended in an open area because wind would disperse it.

3.5 Dust Emissions

Fugitive dust emissions are considered monetary losses and pose environmental problems in the areas surrounding a stockyard. Coal dust poses health risks and damages equipment. Environment regulations are becoming more stringent regarding emissions of particulates such as PM_{10} and $PM_{2.5}$.

Dust emissions can be caused by wind erosion in a coal pile and stacking and reclaiming operations of coal piles.

The magnitude of these coal dust emissions depends on the following: (1) the coal characteristics (moisture content, particle shape, and particle size distribution); (2) the stockpile arrangement (layout and construction); and (3) the local climatic conditions (wind velocity, sun, and rainfall).

From these above-mentioned factors, the moisture content of the coal plays an important role in defining the level of dust generated. This dust emission potential is viewed as severe when the surface moisture of the coal lies between 0 and 4%, mild if it is between 4 and 8%, and low at 8% or above. In addition to this moisture factor, the volume of fines is another factor affecting dust emission. Small-size coal piles with a high volume of fines will produce higher dust emission levels than coal piles with larger particles; coal particles greater than 1 mm in size are not likely to become airborne. The next factor affecting wind-borne dust emission is the particle size distribution (which is related to the friability of the coal). Low-rank coals are generally more friable and consequently more prone to generate dust. Very high rank coals are also susceptible to dust emission.

Regarding the stockpile arrangement and its layout, an appropriate selection of the stacking and the reclaiming method, when combined with a series of control measures, can help lower dust emissions.

3.5.1 Methods for Dust Emission Control and Monitoring

Several methods are available to address dust suppression: (1) wind control (stockpile layout and construction with fences/berms); (2) wet suppression systems (water, chemical wetting agents, and foams);

(3) chemicals binders and agglomerating agents; (4) sealants (chemicals or vegetation); and (5) stockpile enclosure (complete or partial).

Table X gives an overview of the applicability of these dust control methods.

These dust control methods are applied during stacking and reclaiming of coal and applicable to active and inactive (long-term) storage piles. In the stacking process, controlling the size and the particle distribution of the stock piles is important. Conical piles are subject to particle segregation; minimizing the drop height by fitting appropriate stacking chutes can decrease particle fragmentation and consequently reduce dust emissions. Among reclaiming systems, gravity reclamation (through bottom reclaimers underneath the pile) typically generates the least dust, but safety precautions must be exercised due to fire and explosion hazards related to dust generation.

The orientation of the stockpiles can help minimize the dust emission level. This orientation can include presenting the smallest surface area (narrowest side) of the piles to the prevailing wind (wind direction according to climatic conditions), taking advantage of any natural shelters, or setting berms and fences. In addition to acting as wind breaks, berms (earth embankments) would also reduce noise pollution and diminish the visual adverse impact of the piles on the landscape. Fences, if erected upwind, can reduce the wind speed and turbulence; if placed downwind, they can prevent dust transport but can also create dead zones where dust can be deposited before being redispersed.

3.5.2 Dust Suppressants

Several dust control and suppressing agents are available for use on coal piles, including water, wetting agents, foams, binders, agglomerating agents, and chemicals. These agents are as follows:

1. Water is the simplest, cheapest method in the short-tem, but not always the most effective. Water cannons are used to wet the coal. There are many drawbacks, such as short-lived effect, reduced heating value of the coal with excessive water content (compromising contract moisture specification), and accelerated surface weathering of the coal (the coal breaks down). Also, as a consequence, handling of wet coal and water runoff must be taken into consideration.

2. Wetting agents are added to reduce the amount of water used; they reduce the surface tension of water to enable fine water droplets and to enable the ability of water to wet the coal. The concentration of wetting agents ranges from 0.05 to 2% in spray water. A substantial number of wetting agents is offered on the market, but the appropriateness of the wetting agent to a certain type of coal must be established in laboratory or field tests and needs to satisfy environmental requirements (biodegradability and aqueous toxicity).

3. Foams are produced from water, air, and a foaming agent. If properly applied, they are better at controlling the moisture content of coal than water or a wetting agent.

4. Agglomerating agents (also called binders) bind the particles to one another or to larger particles; they include oils, oil products (such as asphalt), lignins, resins and polymer-based products. Oil and oil products are suitable when freezing problems are encountered. Safety (fire) and environmental issues (groundwater) need to be considered; polysulfides may be toxic to aquatic organisms. The binders that claim to be nontoxic are Dustcoat series and Coalgard (oil emulsion).

5. Sealants, as crusting or coating agents, are sprayed over the stockpile with a water cannon to form a water-insoluble crust. This method is suitable only for long-term stockpiles and requires regular inspection; reapplication of the crusting agent is necessary once the surface of the pile is broken for reclaiming. The crusting agents include polymer-based products (polyvinyl acetates, acrylics, and latex), resins, lignosulfonates, lignins, waste oils,

TABLE X

Applicability of Dust Control Methods

	Wind control	Wet suppression	Agglomerating agents	Sealants	Enclosures
Stacking		X	X		
Reclaiming		X	X		
Active piles	X	X	X		X
Long-term piles			X	X	

Source. Adapted from Carpenter (1999).

and bitumens. Lignins and waste oil have drawbacks in terms of water solubility and heavy metal contamination. Other novel approaches as sealant solutions are applications of grass or fly ash to the pile.

6. A complete or partial enclosure of piles is a solution to limit dust formation. Circular stockpiles are cheaper to cover than longitudinal piles. Safety measures would include ventilation to inhibit fire and explosion (dusts and methane). In addition, provision of ample openings to allow access for moveable handling equipment should be considered.

Monitoring of dust levels around a stockyard not only helps determine the effectiveness of dust control practices, but also helps to confirm that ambient air quality standards have been satisfied.

A successful monitoring of coal dust would require adequate measurement points (taking into consideration the direction of the prevailing wind), knowledge of the meteorological conditions (wind speed, temperature, humidity, and rainfall), and coal-handling operations. Although instrumentation to measure dust concentration or dust deposition is available, particulate measurements are generally known to be difficult to standardize.

3.6 Runoff and Flowslides

3.6.1 Coal Pile Runoff

In light of more stringent environmental regulations, rising costs of water management, and disposal of waste products, stockyard operators are encouraged to incorporate systems that will address runoff and flowslide of coal piles.

Coal pile runoff occurs when water seeps through the pile or runs off the ground surface. Several elements could generate this runoff: rain, snow, piles left to drain to meet moisture content requirements, and underground streams beneath the pile.

The amount of rainfall has the greatest impact on the amount of runoff; soil permeability under the pile also has an impact but to a lesser extent. Factors influencing the coal pile runoff include the following: the composition and size distribution of the coal; the drainage patterns in terms of contact time between coal and infiltrating water; and the amount of water seeping through the pile.

In general, runoff is characterized by high levels of solids in suspension, a variable pH, and heavy metal content. The composition of this effluent can vary widely. Acidic leachates show high levels of many metals. The pH of these leachates also depends on the presence of the neutralizing elements in the coal. Thus, bituminous coals produce runoff that is acidic; subbituminous coals produce neutral to alkaline runoff.

3.6.2 Runoff Collection

Site evaluation and site preparation are necessary steps to limit soil contamination caused by runoff. The potential site is evaluated in terms of its soil characteristics, bedrock structure, drainage patterns, and climatic conditions (precipitation records and potential flooding). Site preparation would involve compacting the soil, layout, and the construction of drainage ditches around the perimeter of the piles; the ground of the piles would be slightly sloped toward the drainage ditches and sloped away from the reclaiming area. The drainage system is designed to collect the runoff, collect rainwater from the entire stockyard, and handle the storm water. Several ways to prevent runoff and groundwater contamination include the installation of an impenetrable layer under the stockpile by using the following: (1) a plastic liner with compacted sand as the protected layer (Ashdod coal terminal, Israel); (2) a reinforced polypropylene membrane layer with fly ash/bottom ash mixed with an activator as the protective layer (Morgantown, West Virginia, Generating Station); and (3) an impermeable membrane protected by a cement-treated base overlain by rejected waste coal as the protective barrier (Los Angeles, California, Export Terminal).

3.6.3 Runoff Treatment

For recovery of the coal fines, coal pile runoff is directed to settling ponds located close to the corners of the stockyard and away from the stacking and reclaiming areas. Another alternative would be to install presettling ponds ahead of the settling ponds so that the bulk of the fines would be recovered in the primary ponds. This would reduce the amount of fines collected in the settling ponds, leading to a less frequent cleanup of these ponds. Different types of coal produce different types of runoff. In certain effluents (runoff), additives are used to remove the suspended solids and to modify the pH. Examples of these additives are alum (added to the runoff of the Powder River Basin [PRB] subbituminous coal at the U.S. St. Clair Power Plant) and cationic polymer flocculant (added to the runoff at Bolsover Works, United Kingdom). In addition, filter presses, other mechanical equipment, and chemical processes for wastewater treatment may be involved. Bioremedia-

tion techniques can be applied to produce biodegradable by-products and recover metals.

3.6.4 Coal Pile Flowslides

Instability in coal piles results from two types of slope failure: shallow slipping and deep-seated sliding, which is the type that causes flowslides.

In the first case, wetting of the surface layer can lead to erosion gullies and to shallow slipping with small flows depositing saturated coal at the toe of the slope. In the second case, deep-seated sliding causes a major flowslide within a short time. The stockpile shows a flat final slope and a temporary steep scarp that subsequently slumps back to the angle of repose. This slope failure results in economic loss associated with cleanup costs, loss of production, damage to equipment, and danger to personnel. Significant conditions that could lead to flowslides include the following:

1. Saturation of the stockpile base due to the infiltration of heavy rainfall, leading to the potential of a deep-seated slip;
2. Redistribution of moisture within the coal at placement (there is a threshold moisture content below which no saturated zone develops);
3. Loosely stacked coal (prone to structural instability);
4. Particle size distribution (migration of fine coal particles under the influence of water flow causes local water saturation and reduction in shear strength).

Failures can happen with relatively fresh coal (placed within the previous 2 weeks), and a significant coal pile height increases the risk of collapse.

Ideally, strategies to prevent and control flowslides would mean modifying the stacking method and the overall stockpile installation, but in reality operational and other factors may preclude major modifications to the existing operations. Consequently, a more practical approach is to adopt safety precautions and adjust the height of the stockpiles as prevention measures:

1. Excluding pedestrians and mobile equipment from stockpiles thought to be prone to flowslides or from coal piles with modified characteristics, particularly in the week after heavy rain or placement of wet coal.
2. Minimizing stockpile height during the rainy season; in case of busy shipping schedules and heavy rain, access roads should be closed to personnel; if a coal slide occurs, it should be cleaned up and restacked.
3. Reclaiming the coal from the toe of the pile by front-end loaders should be avoided because of the hazard; as coal dries out and adhesion between particles becomes weaker, the unstable sides of the pile may collapse.

To control flowslides, experience has shown that:

1. Controlling the moisture content of the pile can decrease the risk of flowsliding.
2. Compacting the entire coal pile, or selected areas of it, can reduce the tendency for the saturated coal to suddenly lose strength and to flow.
3. Building drainage slopes can facilitate surface runoff.
4. Constructing the pile with an appropriate profile and top facilitates runoff (with convex longitudinal profile and no flat-topped center, or slightly concave cross-sectional profile, or slightly crowned top for compacted piles).
5. Sealing the coal pile is useful in heavy rain areas but is not practical for working piles.
6. Enclosing the pile keeps the coal dry but is an expensive option and not applicable at coal blending terminals.
7. Draining the stockpile toe is beneficial for preventing minor instabilities that could lead to an overall failure.

4. HANDLING COAL AT THE END-USE SITE

4.1 Pneumatic Transport of Solids

Pneumatic transport of particles occurs in almost all industrial applications that involve powder and granular materials. The purpose of pneumatic transport is to protect the products from the environment and protect the environment from the products. Although this is not an energy-efficient method of transport because it requires power to provide the motive air or gas, it is easy and convenient to put into operation.

Five components are included in a pneumatic system: conveying line, air/gas mover, feeder, collector, and controls.

Pneumatic systems are broken down into three classifications: pressure system, vacuum system, and pressure/vacuum system. Their modes of transport are categorized as dilute, strand (two-phase), or

dense phase and take into account the characteristics of the particles in terms of the following:

1. Material size and size distribution;
2. Particle shape;
3. Overall force balance of the particle/gas system, taking into account the acceleration, the drag forces, the gravity, and the electrostatic forces; terminal velocity is usually the characterizing term for particles—the higher the terminal velocity value, the greater the size and/or density of the particle;
4. The pressure drop and the energy loss for operational costs; the pressure loss calculations would take into account the voidage, the particle velocity, and the friction factors;
5. Acceleration of the flow at several locations of the pneumatic system (bend or connection, outlet or inlet);
6. Saltation (deposition of particles at the bottom surface of the pipe);
7. Pickup velocity, defined as the feed point of the solids and the velocity required to pick up particles from the bottom of the pipe;
8. Compressibility, which is of paramount importance for long distances and high-pressure systems;
9. Bends, creating significant pressure loss for systems with several bends and a relatively short length (<300 ft);
10. A dense phase that includes all types of flow, except dilute-phase flow (a lightly loaded flow); and
11. Choking conditions (in a vertical direction).

4.2 Pneumatic Transport of Coal

Three different types of air movers are commonly used for pneumatic transport: fans, blowers, and compressors. The fans are the most inexpensive movers, the blowers are the workhorse of pneumatic transport, and the compressors are used for pressure ratios over 2:1.

When any of these movers are used in coal-fired steam power plants, primary air fans are used to supply air to the mills. There are many varieties of feeders, such as rotary feeders with different design variations (rotor, blades, and casing), gate lock feeders in pressurized systems, and screw feeders. In addition, filters and cyclones are available for separating, dividing, or redirecting the flow.

4.3 Transport Properties of Slurries or Wet Solids

Understanding and accurately predicting the flow behavior of solid–liquid mixtures through the slurry fuel-handling system is key to the successful operation of these handling systems. Slurries are solid–liquid mixtures. By definition, slurries should contain sufficient liquid to wet the surface of every particle, small and large. Therefore, particle size and the relative density between solids and liquids are important and highlight four types of slurry flow regimes: homogenous; heterogeneous; saltation; and moving bed.

These flow regimes are investigated in laboratory-scale and pilot-scale rheological tests to elucidate slurry behavior, e.g., its viscosity under certain flow conditions. In these tests, various geometries and instruments can be used to measure the rheological properties of the slurry. These properties are expressed in terms of shear stress (force/area), shear rate, or strain rate (velocity/distance). Equations or charts are developed to distinguish between Newtonian and non-Newtonian fluids. Virtually all slurries display non-Newtonian behavior and are represented commonly by shear stress characteristics, such as velocity gradient curves (pseudo-plastic, plastic, Newtonian, and dilatant curves). The Bingham plastic is the most common curve; it is actually characterized by a straight line with an intercept on the shear stress axis. This helps in designing the handling systems and the operating parameter ranges. For heterogeneous slurries, the design of this type of transport is based almost entirely on correlations developed from experience.

Pipelines are used to transport slurries over long distances and these slurries must be stable enough to withstand the demands of transport, readily separable after reaching their destination, and cost-competitive with other forms of transport.

For pipelines, operating at lower velocities would result in less pressure loss per length of pipe (smaller pumps) and greater pipeline lifetime (less erosion). The minimum operating velocity is the lowest velocity at which all the particles stay in suspension.

As expected, for a given particle size distribution, the greater the solids loading, the more difficult it is to pump the slurry. However, adding wood pulp fibers to the slurry has considerably reduced the drag because of changes in the turbulent flow regimes. For instance, in aqueous slurries of 20% coal with and without wood pulp fibers flowing at 15 and 60 ft/s, the drag has been reduced up to 30% over this range of velocities by the addition of 0.8% fibers.

In addition to the type of drag-reducing additives that help increase the solid loading in the slurry (to 70% solids with additives, compared to 55% solids without additives), there are other additives that will help to retard particle settling. These can increase the

shelf-life of the slurry from a few days to several months.

4.4 Slurry Transport of Coal

A particle size must be created to transport coal slurry, requiring a balance between pumping and dewatering characteristics. If the slurry is too fine, the pumpability is good but the process of separating water from solids could be difficult. If the slurry is too coarse, it is not homogenous and higher pumping power is needed. In short, two parameters need to be controlled: the density of the slurry and the top particle size. The first parameter, in terms of slurry concentration, has an important impact on friction loss and the critical velocity of the slurry. The second parameter prevents pluggage of pipelines by controlling the coal particles. In addition, corrosion inhibitors, such as sodium dichromate, will protect the pipe wall, and other additives, such as sodium sulfite, will act as oxygen scavengers.

In commercial pipelines, a so-called critical concentration range exists where the slurry provides maximum support for the particles without an untenably high critical velocity (for the pumping stations). At the destination, the objective is to remove the water and recover the solids with minimum moisture content by using centrifuges or filters.

Coal slurries find applications in power generation. The IEA Coal Research report of 1993 gives an insight into the impact of the coal on the power station performance. In conventional steam power plants, coal slurry technology has been commercialized. One example is the 273-mile Black Mesa coal pipeline that provides all the fuel requirements (approximately 5 million tons/year) to two 750 MW generating units in southern Nevada.

In pressurized fluidized bed combustion, coal and sorbent (usually lime or limestone) are injected into a fluidized bed as a water-mixed paste using concrete pumps or pneumatically as a dry suspension via lock hoppers. In the first case, the coal could be crushed by roll crushers to a certain size distribution, such as a top size of 6 mm, and the sorbent could be crushed through hammer mills to a certain top size, such as 3 mm. After that, the coal and sorbent are mixed to form a pumpable slurry. Optimizing the paste moisture content at 20–30% is an acceptable compromise relative to boiler efficiency. At the Tidd Plant in Brilliant, Ohio, for example, the coal paste had a nominal value of 25% weight water.

Regarding atmospheric circulating fluidized bed combustion, coal slurry represents a preferable solution when using low-grade coals and coal wastes. Three options are available: a dilute slurry ($>40\%$ water), a dense slurry ($<40\%$ water), or a nominally dry material ($\approx 12\%$ water). Coal slurry has also been used in integrated gasification combined cycles, such as at the Wabash power station and at the Tampa Electric Company's Polk power station.

TABLE XI

Example of Coal Storage Requirements at a Coal-Fired Power Plant

	Coal			
	Lignite	**Subbituminous**	**Midwestern bituminous**	**Eastern bituminous**
Heat to turbine	4591×10^6 kJ/h	4591×10^6 kJ/h	4591×10^6 kJ/h	4591×10^6 kJ/h
Boiler efficiency	83.5%	86.2%	88.5%	90.5%
Coal heat input	5499×10^6 kJ/h	5326×10^6 kJ/h	5185×10^6 kJ/h	5073×10^6 kJ/h
Coal HHV	14.135 MJ/kg	19.771 MJ/kg	26.284 MJ/kg	33.285 MJ/kg
Coal flow rate	429 tons/h	297 tons/h	217 tons/h	168 tons/h
	Design storage requirements, tons			
Bunker (12 h)	5148	3564	2604	2016
Live (10 days)	102,960	71,280	52,080	40,320
Dead (90 days)	926,640	641,520	468,720	362,880
	Storage time for eastern bituminous plant design[a]			
Bunker	4.7 h	6.8 h	9.3 h	12.0 h
Live	3.9 days	5.7 days	7.7 days	10.0 days
Dead	35 days	51 days	69 days	902 days

Source. Adapted from Folsom *et al.* (1986).

[a] Equivalent storage time for a plant designed for eastern bituminous coal but fired with the coals listed.

5. STORING COAL AT THE END-USE SITE

5.1 Safe Storage of Crushed Coal

Crushed coal storage can be divided into two types: live storage (e.g., active storage with a short residence time) and dead storage (e.g., reserve storage). A common practice is to directly transfer part of a shipment to live storage and divert the remainder to reserve storage.

Usually, the storage components are sized to a certain capacity equivalent to a fixed period of firing at full load. The length of time for each storage type is as follows: 12 h for in-plant storage in bunkers, 10 days for live storage, and more than 90 days for reserve storage. These time periods are usually set by the operating procedures of the plant (power station capacity, heat rate, and coal heating value) and by other constraints (e.g., stocking strategies). If the actual heating value of the coal changes and if the steam generator is expected to operate at a certain fixed heat input, the storage capacity and the quantities delivered to the plant must increase. As a result, silos or bunkers represent an important link to provide an interrupted coal flow to the pulverizers. Typically at coal-fired power plants, a number of round silos with conical bottoms are stationed on each side of the boiler (steam generator).

Silos and bunkers are constructed of either concrete or steel and are generally round. Loading and unloading of coal from the silo are carefully controlled to prevent dangerous offset loading on the foundations. The coal is supplied to the mills and pulverized continuously and a mixture of pulverized coal/air is pneumatically transported to the burners. A single mill can supply several burners (tangentially fired systems and wall-fired systems). Table XI gives an example of coal storage requirements at a coal-fired power plant in the 1980s.

TABLE XII

Preferred Range of Coal Properties for Mills

Properties	Mill type		
	Low speed	Medium speed	High speed
Maximum capacity	100 tons/h	100 tons/h	30 tons/h
Turndown	4 : 1	4 : 1	5 : 1
Coal feed top size	25 mm	35 mm	32 mm
Coal moisture	0–10%	0–20%	0–15%
Coal mineral matter	1–50%	1–30%	1–15%
Coal quartz content	0–10%	0–3%	0–1%
Coal fibre content	0–1%	0–10%	0–15%
Hardgrove grindability index	30–50, 80	40–60	60–100
Coal reactivity	Low	Medium	Medium

Source. Adapted from Sligar (1985).

5.2 Milling of Pulverized Coal

At the mill, fine grinding of coal involves producing coal particle sizes such that 70% or more are finer than 75 μm (200-mesh) to ensure complete

TABLE XIII

Impact of Coal Properties on Mill Performance

Mill performance	Low-speed and medium-speed mills
Drying	• Moisture affects primary air requirements and power consumption of mills • Volatile matter affects the susceptibility to mill fires • Total moisture causes a −3% throughput for a 1% moisture increase for low-speed mill; a −1.5% throughput for 1% moisture increase above 12% moisture for medium-speed mill
Mill throughput	• Total moisture (see above) • −1% throughput for 1-unit reduction in HGI[a] for both mill types (caution when using HGI for coal blend behavior) • Raw coal top size: −3% throughput for 5 mm increase in top size for low-speed mill; no loss in throughput below 60 mm top size for medium-speed mill • Pulverized fuel size distribution: reduction of fraction going through <75 μm mesh screen by 0.35% for a 1% increase in throughput for low-speed mill; reduction by 0.9% for a 1% increase in throughput for medium-speed mill
Wear	• Mineral matter and size distribution influence the operation and maintenance of both types of mills

Source. Adapted from Lowe (1987).
[a] *Note.* HGI, Hardgrove grindability index.

combustion and minimization of ash and carbon on the heat exchanger surfaces.

Three types of mills, according to speed, are available: low-speed mills of ball/tube design; medium-speed mills of vertical spindle design; and high-speed mills with a high-speed rotor.

Table XII gives the preferred range of coal properties applicable to the types of mills and Table XIII shows the impact of coal properties on the performance of the low-speed and medium-speed mills.

SEE ALSO THE FOLLOWING ARTICLES

Clean Coal Technology • Coal, Chemical and Physical Properties • Coal Conversion • Coal, Fuel and Non-Fuel Uses • Coal Industry, Energy Policy in • Coal Industry, History of • Coal Mine Reclamation and Remediation • Coal Mining, Design and Methods of • Coal Preparation • Coal Resources, Formation of • Hydrogen Storage and Transportation • Markets for Coal • Natural Gas Transportation and Storage

Further Reading

Alvarez, R., Barriocanal, C., Casal, M. D., Diez, M. A., González, A. I., Pis, J. J., and Canga, C. S. (1996). Coal weathering studies. *In* "Ironmaking Conference Proceedings, Pittsburgh, PA, March 24–27, 1996." Vol. 55.

Alvarez, R., Casal, M. D., Canga, C. S., Diez, M. A., González, A. I., Lázaro, M., and Pis, J. J. (1995). Influence of natural weathering of two coking coals of similar rank on coke quality. *In* "Coal Science, Proceedings of the 8th International Conference on Coal Science, Oviedo, Spain, September 10–15, 1995." Elsevier Science, Amsterdam/New York.

Brolick, H. J., and Tennant, J. D. (1990). Innovative transport modes: Coal slurry pipelines. *In* "Fuel Strategies—Coal Supply, Dust Control, and By-product Utilization, The 1990 International Joint Power Generation Conference, Boston, MA, October 21–25, 1990." National Coal Association, Washington, DC.

Carpenter, A. M. (1999). "Management of Coal Stockpiles." IEA Coal Research, October 1999. International Energy Agency, London, UK.

Chakraborti, S. K. (1995). American Electric Power's coal pile management program. *Bulk Solids Handling* **15**, 421–428.

Chakraborti, S. K. (1999). An overview of AEP's coal handling systems. *Bulk Solids Handling* **19**, 81–88.

Craven, N. F. (1990). Coal stockpile bulk density measurement. *In* "Coal Handling and Utilisation Conference, Sydney, NSW, Australia, June 19–21, 1990."

Department of Energy (1990). "Innovative Alternative Transport Modes for Movement of U.S. Coal Exports to the Asian Pacific Basin," DOE Report DOE/FE-61819-H1, Prime Contractor, Argonne National Laboratories; Subcontractors, William Technologies Inc., The Fieldston Co., and TransTech Marine Company, March 1, 1990.

Deurbrouck, A. W., and Hucko, R. E. (1981). Coal preparation. *In* "Chemistry of Coal Utilization, Secondary Volume," Chap. 10 (M. Elliott, Ed.). John Wiley & Sons.

Dong, X., and Drysdale, D. (1995). Retardation of the spontaneous ignition of bituminous coal. *In* "Coal Science, Proceedings of the 8th International Conference on Coal Science, Oviedo, Spain, September 10–15, 1995", Vol. 1, pp. 501–504. Elsevier Science, Amsterdam, the Netherlands/ New York.

Folsom, B. A., Heap, M. P., Pohl, J. H., Smith, J. L., and Corio, M. R. (1986). Effect of coal quality on power plant performance and costs. *In* "Review of Coal Quality Impact Evaluation Procedures," EPRI-CS-4283-Vol. 2 (February 1986). EPRI Technical Information Services.

Kaplan, S. (2002). "Coal Transportation by Rail—Outlook for Rates and Service, Energy Information Administration, March 12, 2002," PA Consulting Group, Washington, D.C.

Klinzing, G. E. (1995). Pneumatic conveying. *In* "NSF Short Course on Fluid/Particle Processing," June 5–9, 1995, University of Pittsburgh, Pittsburgh, PA.

Lema, J. E. (1990). Attributes of selected transportation modes. *In* "Fuel Strategies—Coal Supply, Dust Control, and By-product Utilization, The 1990 International Joint Power Generation Conference, Boston, MA, October 21–25, 1990." National Coal Association, Washington, DC.

Leonard, J. W., III, and Hardinge, B. C. (eds.). (1991). "Coal Preparation," 5th ed. Society for Mining, Metallurgy and Exploration, Inc., Littleton, CO.

Lowe, A. (1987). Important coal properties for power generation. *In* "Geology and Coal Mining Conference," pp. 188–196. Sydney, Australia, October 13–5, 1987.

Meikle, P. G., Bucklen, O. B., Goode, C. A., and Matoney, J. T. (1991). Post preparation/storage and loading. *In* "Coal Preparation" (J. W. Leonard, III and B. C. Hardinge, Eds.), 5th ed., Chap. 9.

O'Connor, D. C. (1987). On-line coal analysis: A new challenge. *In* "Coal Sampling, Fundamentals and New Applications–Belt Conveyor Systems." The 1987 Joint Power Generation Conference, October 4–8, 1987.

Parek, B. K., and Matoney, J. P. (1991). Dewatering. *In* "Coal Preparation," 5th ed., Chap. 8 (Leonard, J. W., III, and Hardinge, B. C., Eds.). Society for Mining, Metallurgy, and Exploration Inc., Littleton, CO.

Schneid, R. J. (2000). "Coal Transportation, Coal and Power Industry Training Course," Presented to U.S. Department of State Foreign Service Officers and Other U.S. Government Employees, July 25, 2000. Consol Energy Inc.

Scott, D. H., and Carpenter, A. M. (1996). "Advanced Power Systems and Coal Quality, IEA Coal Research, May 1996." International Energy Agency, London, UK.

Sharpe, M. A. (2001). "Coal Preparation, Coal and Power Industry Training Course, Presented to the U. S. Department of State Foreign Service Officers, July 25, 2001." CLI Corp, Canonsburg, PA.

Skorupska, N. M. (1993). "Coal Specifications—Impact on Power Station Performance, IEA Coal Research, January 1993." International Energy Agency, London, UK.

Sligar, J. (1985). Coal pulverizing mill section. *In* "Course on the Characterization of Steaming Coals," pp. 10.1–10.23. University of Newcastle, Institute of Coal Research, Newcastle, NSW, Australia.

Speight, J. G. (1994). Preparation and transportation. *In* "The Chemistry and Technology of Coal," 2nd ed., Chap. 6, p. 121. Dekker, New York.

Sujanti, W., and Zhang, D.-K. (1999). Low-temperature oxidation of coal studied using wire-mesh reactors with both steady-state and transient methods. *Combustion Flame* **117,** 646–651.

Thomson, T. L., and Raymer, F. B. (1981). Transportation, storage and handling of coal. *In* "Chemistry of Coal Utilization" (M. A. Elliott, Ed.), Chap. 9, pp. 523–570. Wiley, New York.

Walker, S. (1999). "Uncontrolled Fires in Coal and Coal Wastes, CCC/16, IEA Coal Research, April 1999." International Energy Agency, London, UK.

Wildman, D., Ekmann, J., and Mathur, M. (1995). Slurry transport and handling. *In* "NSF Short Course on Fluid/Particle Processing," June 5–9, 1995, Pittsburgh Energy Technology Center, Pittsburgh, PA.

Zumerchik, J. (2001). Coal transportation and storage. *In* "Macmillan Encyclopedia of Energy," Vol. 1, (J. Zumerchik, Ed.). Macmillan Co., New York.

Cogeneration

DOUG HINRICHS
Clean Energy Consultant
Silver Springs, Maryland, United States

1. Time Line
2. Technical Characterization
3. Applications and Markets
4. Regulatory and Rate Structure Issues
5. U.S. Federal Government Support

Glossary

absorption chiller An absorption chiller transfers thermal energy from the heat source to the heat sink through an absorbent fluid and a refrigerant. The absorption chiller accomplishes its refrigerative effect by absorbing and then releasing water vapor into and out of a lithium bromide solution.

cogeneration The simultaneous production of electric and thermal energy in on-site, distributed energy systems; typically, waste heat from a prime mover is recovered and used to heat/cool/dehumidify building space. Technically, if cooling is an end result, the process is called trigeneration.

combined heat and power Synonymous with cogeneration, combined heat and power (CHP) is more commonly used in the United States than in Europe or China, where cogeneration is more common. The term CHP applies more to the industrial sector, in which "waste heat" may be used in applications other than heating or cooling building space.

desiccant A solid or liquid material with an affinity for absorbing water molecules.

desiccant dehumidifier Desiccant dehumidifiers use recovered heat from the prime mover to dry building space using a desiccant material. The two general types are based on the desiccant used: solid (dry desiccant) and liquid (liquid desiccant).

distributed generation In contrast to centralized utility electricity generation, distributed generation involves small-scale prime movers that can be located at or near the building or facility where the energy is used to provide greater power reliability and reduced emissions.

district energy District energy systems produce steam, hot water, or chilled water at a central plant and then pipe that energy out to buildings in the district for space heating, domestic hot water heating, and air-conditioning. A district energy system can obviate the need for individual buildings to have their own boilers or furnaces, chillers, or air conditioners. A district energy system can serves many customers from one location; use a variety of conventional fuels, such as coal, oil, and natural gas; and transition to use renewable fuels, such as biomass, geothermal, and combined heat and power.

"first-generation" packaged cooling, heating, and power systems On August 8, 2001, Energy Secretary Abraham announced awards to seven industry teams for research, development, and testing of first-generation packaged cooling, heating, and power systems. In doing so, the U.S. Department of Energy is helping manufacturers work together to integrate their individual cogeneration components into easier to use packages and to make "plug-and-play" packages more readily available in the commercial and industrial marketplace; objectives include reduction of engineering time/cost at the installation site.

fuel cell Device for producing electricity using a chemical process that combines hydrogen and oxygen rather than conventional combustion processes with electric generators.

prime mover The electric and thermal energy generator; typically an on-site distributed generator. For cogeneration purposes, microturbines, reciprocating engines, and industrial turbines are most commonly used; other prime movers include fuel cells, concentrating solar power systems, and Sterling engines.

Public Utilities Regulatory Policy Act (PURPA) Congress passed PURPA in 1978. The nation had recently survived several energy crises resulting from the escalation of crude oil prices coupled with curtailments of production. Policymakers belatedly realized that fuel was dependent on the forces of international interests and beyond the control of governmental officials and industries in the United States. Public utility regulators and Congress determined that steps to avoid future energy crises should include both focused conservation programs and the development of alternative sources for producing electricity. PURPA was a major response to these crises and started the power industry on the road to competition by ending promotional rate

 © 2004 Elsevier Inc. All rights reserved.

structures, establishing unconventional qualifying facilities, and encouraging conservation. Unintentionally, PURPA started the competitive process because it enabled nonutility generators to produce power for use by customers attached to a utility's grid. PURPA broke the stranglehold on power companies' previous monopoly in the generation function because now any unregulated cogenerator or renewable energy producer could sell electricity into the power grid, and regulated utilities were unable to dictate terms. Of particular note regarding the cogeneration industry, PURPA promoted cogeneration as an energy-efficient technology.

thermally activated technologies Thermally activated technologies (TATs) represent a diverse portfolio of equipment that transforms heat for a useful purpose, such as energy recovery, heating, cooling, humidity control, thermal storage, or bottoming cycles. TATs are the essential building blocks for cogeneration systems.

trigeneration The simultaneous production of electric and thermal energy in on-site distributed energy systems; typically, waste heat from a prime mover is recovered and used to heat and cool/dehumidify building space. Technically, if cooling is an end result, the process is called trigeneration.

waste heat The portion of the energy input to a mechanical process that is rejected to the environment.

The efficiency of central electricity generation has not improved much in the past 40 years, with the exception of an increase in natural gas utilization and the use of combined cycle plants by central plants. Currently, power plants are approximately 33% efficient, meaning that only one-third of the energy is used to generate new power. The remaining two-thirds is lost, vented as "waste heat" to the environment that, along with carbon/nitrogen/sulfur-based emissions from coal or oil, causes significant ecosystem degradation. Cogeneration, or combined heat and power, is a clear alternative to central electricity generation: Using smaller scale generation technologies that primarily use natural gas at or near a building, campus, or industrial facility, cogeneration systems take advantage of waste heat by using it to heat, cool, or dehumidify building space, with overall system efficiencies in the 70–90% range. (Technically, an energy system that makes heat, power, and cooling is considered trigeneration. In this article, cogeneration includes trigeneration.) This doubling or tripling of efficiency from cogeneration or trigeneration has far-reaching implications for the possibility of a more sustainable energy paradigm in commercial and industrial sectors in industrialized countries.

1. TIME LINE

1.1 Historical

Cogeneration is not new. It first appeared in the late 1880s in Europe and in the early 20th century in the United States, when most industrial plants generated their own electricity using coal-fired boilers and steam-turbine generators. Many of these plants used the exhaust steam for industrial processes. According to COGEN Europe, an advocacy group based in Belgium, cogeneration produced as much as 58% of the total power from on-site industrial power plants in the United States in the early 1900s.

As a result of improvements in the cost and reliability of electricity generated by the separate electric power industry as well as increasing regulation, the electric generation capacity at most cogeneration facilities was abandoned in favor of more convenient purchased electricity. A few industries, such as pulp and paper and petroleum refining, continued to operate their cogeneration facilities, in part driven by high steam loads and the availability of by-product fuels.

In the mid-1900s, the pendulum began to swing in the opposite direction with the advent of central electric power plants and reliable utility grids, which allowed industrial plants to begin purchasing lower cost electricity. As a result, on-site industrial cogeneration declined significantly to only 15% of total U.S. electrical generation capacity by 1950, and it then declined to approximately 4% by 1974.

The first dramatic rise in fuel costs and uncertainty in the fuel supply in the 1970s triggered an upswing in cogeneration, particularly in large industrial applications that required vast quantities of steam. In recent years, smaller cogeneration systems have begun to make inroads in the food, pharmaceutical, and light manufacturing industries, commercial buildings, and on university campuses.

In the late 1970s interest in cogeneration in the United States was renewed in response to the Public Utilities Regulatory Policy Act of 1978 (PURPA), which included measures to promote cogeneration as an energy-efficient technology. With the passage of PURPA, multiple megawatt projects at large pulp and paper, steel, chemical, and refining plants were undertaken. Although PURPA created an enormous incentive for these large-scale projects that helped increase cogeneration capacity significantly, the act may have caused utilities to discourage beneficial projects because of the regulatory mandate. However, these attitudes are changing: Forward-thinking

utilities are actively pursuing new cogeneration projects as part of their business strategy. PURPA has played a critical role in moving cogeneration into the marketplace by addressing many barriers that were present in the early 1980s.

During the 1990s, approximately 9500 miles of new high-voltage transmission lines were constructed (an approximately 7% increase)—short of the need—which represented a new opportunity for efficient distributed generation (DG) and cogeneration use. The U.S. Department of Energy (DOE) estimates that more than 390 GW of new generating capacity will be needed by 2020 to meet growing U.S. energy demand and offset power lost from retired power plants. However, obtaining approval for construction of new transmission lines and acquiring right-of-ways is very difficult.

According to Tom Casten, founder of Trigen and currently chairman and CEO of Private Power, "The public simply detests new transmission lines, and no one wants them in their back yard. Most recent U.S. power problems were caused by lack of adequate transmission and distribution (T&D), and nobody close to the industry believes enough new transmission can be built." Distributed cogeneration systems require no new construction of T&D lines. Tom Casten cites industry forecasts that estimate that the United States will need 137,000 MW of new capacity by 2010. According to Casten, meeting this demand will require $84 billion for new power plants and $220 billion for new T&D, for a total of $304 billion. Meeting the same demand with DG will require $168 billion for new plants but $0 for T&D. Casten notes, "A distributed energy future would save $136 billion of capital investment and reduce the cost of new power by about 3¢/kW."

1.2 Present

The Energy Information Administration (EIA) reports that as of 2000, cogeneration accounted for approximately 7.5% of electricity generation capacity and almost 9% of electricity generated in the United States. Private Power studied data from an unpublished EIA survey and other EIA data and determined that total DG was 10.7% of U.S. totals in 2000.

In stark contrast, in The Netherlands, more than 40% of electricity is obtained from combined heat and power (CHP) systems, and in Denmark, approximately 60% is obtained from these systems. In the United Kingdom, cogeneration's share of electric power production has doubled in the past decade, with additional growth targeted by the government.

2. TECHNICAL CHARACTERIZATION

2.1 Efficiency Gains

According to the U.S. Combined Heat and Power Association, cogeneration technologies produce both electricity and steam from a single fuel at a facility located near the consumer. These efficient systems recover heat that normally would be wasted in an electricity generator, and they save the fuel that would otherwise be used to produce heat or steam in a separate unit.

Cogeneration offers dramatic advantages in efficiency and much lower air pollution than conventional technologies (Fig. 1). A wide variety of cogeneration technologies generate electricity and meet thermal energy needs (direct heat, hot water, steam, and process heating and/or cooling) simultaneously at the point of use. Cogeneration can result in system efficiencies of 70–90%, compared with the national average of 30% efficiency for conventional electricity plants and perhaps 50% for conventional thermal applications.

The environmental benefits of cogeneration are apparent. Cogeneration systems in the United States:

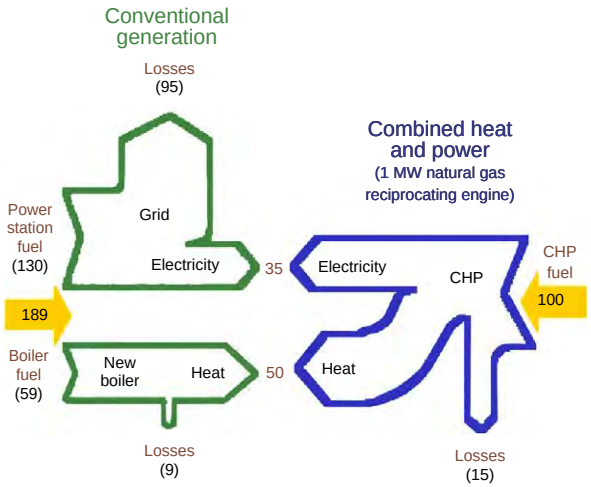


FIGURE 1 This diagram compares the typical fuel input needed to produce 35 units of electricity and 50 units of heat using conventional separate heat and power. For typical electric and thermal efficiencies, cogeneration is nearly twice as efficient.

- Decrease energy use by approximately 1.3 trillion Btus/year.
- Reduce NO_x emissions by 0.4 million tons/year.
- Reduce sulfur dioxide (SO_2) emissions by more than 0.9 million tons/year.
- Prevent the release of more than 35 million metric tons of carbon equivalent into the atmosphere.

2.2 Technology Advances

Nearly all U.S. buildings and industries use thermal energy from boilers for space heating, hot water, steam systems, and direct process heating applications. Most cogeneration systems installed in the United States are used for industrial applications, but there is increasing usage by U.S. commercial and institutional building owners and district energy systems worldwide. Newer thermally activated technology applications include recycling waste heat for cooling and humidity control, which has significant energy and economic potential.

Recent advances in technology have resulted in the development of a range of efficient and versatile systems for industrial and other applications. Improvements in electricity generation technologies, particularly advanced combustion turbines and engines, have allowed for new configurations that reduce size but increase output. Distributed, on-site power generation technologies, such as microturbines, reciprocating engines, industrial turbines, and fuel cells, are becoming more effective at generating electricity and producing recoverable thermal energy.

Recovered heat from these generation technologies can be used to power thermally activated technologies, control humidity with desiccant dehumidifiers, and heat buildings with steam or hot water in cogeneration systems.

2.3 Prime Movers

Prime movers generate electricity and heat, or electric and thermal energy. They include several technologies. First, microturbine generator units are composed of a compressor, combustor, turbine, alternator, recuperator, and generator. In a simple cycle turbine (without a recuperator), compressed air is mixed with fuel and burned under constant pressure conditions. The resulting hot gas is allowed to expand through a turbine to perform work.

Recuperated units use a heat exchanger (recuperator or regenerator) that recovers some of the heat from the turbine exhaust and transfers it to the incoming air stream for combustion in the turbine. By using recuperators that capture and return waste exhaust heat, existing microturbine systems can reach 25–30% cycle efficiency. The incorporation of advanced materials, such as ceramics and thermal barrier coatings, could further improve their efficiency by enabling a significant increase in engine operating temperature.

Microturbines offer many advantages over other technologies for small-scale power generation, including the ability to provide reliable backup power, power for remote locations, and "peak shave." Other advantages include less maintenance and longer lifetimes because of a small number of moving parts, compact size, lighter weight, greater efficiency, lower emissions, and quicker starting. Microturbines also offer opportunities to use waste fuels such as landfill gas.

Second, gas-fired reciprocating engines are the fastest selling, least expensive distributed generation technology in the world. Commercially available natural gas versions of the engine produce power from 0.5 kW to 10 MW, have efficiencies between 37 and 40%, and can operate down to NO_x levels of 1 g per horsepower-hour. When properly treated, these rugged engines can run on fuel generated by waste treatment (methane) and other biofuels. By using recuperators that capture and return waste exhaust heat, reciprocating engines can also be used in cogeneration systems in buildings to achieve energy-efficiency levels approaching 80%.

Gas-fired reciprocating engines offer many advantages over other technologies for small-scale power generation, including the ability to provide highly reliable, inexpensive backup power; provide power for remote locations; and generate on-site power during peak periods when utility charges are highest.

Reciprocating engines require fuel, air, compression, and a combustion source to function. Depending on the ignition source, they generally are classified into two categories: spark-ignited engines, typically fueled by gasoline or natural gas, and compression-ignited engines, typically fueled by diesel oil.

Third, tremendous efficiency strides with regard to industrial gas turbines have been made in recent years. Combustion turbines are a class of electricity-generation devices that produce high-temperature, high-pressure gas to induce shaft rotation by impingement of the gas on a series of specially designed blades. Industrial gas turbines are used in many industrial and commercial applications ranging from 1 to 20 MW.

Because gas turbines are compact, lightweight, quick starting, and simple to operate, they are used widely in industry, universities and colleges, hospitals, and commercial buildings to produce electricity, heat, or steam. In such cases, simple cycle gas turbines convert a portion of input energy from the fuel to electricity and use the remaining energy, normally rejected to the atmosphere, to produce heat. This waste heat may be used to power a separate turbine by creating steam. The attached steam turbine may generate electricity or power a mechanical load. This is referred to as a combined cycle combustion turbine since two separate processes or cycles are derived from one fuel input to the primary turbine.

These reliable distributed generation technologies are becoming the mainstay of peak shaving equipment. When needed, this equipment can be started immediately and can quickly generate full capacity.

The waste heat generated by the turbine can be used to generate steam in a heat recovery steam generator (HRSG), which can power a steam turbine, heat living space, or generate cooling using steam-driven chillers. The advantages of these types of systems are inexpensive electrical power and better reliability since the user may be independent from the grid. These systems can be started even if the grid has failed.

Advanced materials, such as ceramics, composites, and thermal barrier coatings, are some of the key enabling technologies under development to improve the efficiency of distributed generation technologies. Efficiency gains can be achieved with materials such as ceramics, which allow a significant increase in engine operating temperature. The increased operating temperature also lowers greenhouse gas and NO_x emissions.

Two successful applications of HRSG systems are at Princeton University and the Massachusetts Institute of Technology (MIT). Each school deploys a gas turbine for electrical generation to meet the majority of the campus's needs. The exhaust heat from these turbines drives an HRSG. The steam from the HRSG provides heating for campus living spaces and cooling through steam-driven chillers. Frequently, summer peak-time operation of electrical chillers is expensive. Producing steam on-site enables the use of steam-driven chillers. Operating steam-driven chillers acts to peak shave the user's high demand operations, thus saving money. The electricity generated by the gas turbine may also be used to operate conventional electric chillers. Since the user generates this electricity, the price may be lower than that of the local utility and does not have a peak or demand charge associated with it.

Fourth, fuel cells were discovered in 1839 when scientist William Grove accidentally let a hydrolysis experiment run backwards. He was surprised to learn that electron flows were created when hydrogen and oxygen combined. NASA made this technology well-known with its use in space shuttles.

Fuel cells consist of two electrodes (cathode and anode) around an electrolyte. The generation process involves introducing hydrogen to one side of the fuel cell (the anode), where it breaks apart into protons and electrons. The electrode conducts protons but not electrons. The protons then flow through while the electrons travel through the external circuit and provide electrical power. The electrons and protons are reunited at the other end of the fuel cell (the cathode). When combined with oxygen from the air, fuel cells produce water and heat in a process that is practically silent, nearly emission free, and involves no moving parts.

Fuel cells are categorized by the kind of electrolyte they use. Electrolyte types used in the building sector include phosphoric acid, molten carbonate, solid oxide, and proton exchange membrane. High-temperature molten carbonate and solid oxide fuel cells are undergoing full-scale demonstration with some existing applications. Manufacturers are actively pursuing this technology for widespread, general use in commercial buildings, homes, and transportation.

Although price and performance data on reciprocating engines and turbines are fairly well established, data for fuel cells are based on a limited number of demonstration projects. As a result, comparisons of price and performance should be interpreted with caution.

The price and performance of engines, turbines, and fuel cells are summarized in Table I.

Finally, concentrating solar electric generators concentrate solar power during the high solar radiation periods onto tubes of water (or oil) to make usable steam, which can then be used to make electricity in a steam generator or to power a chiller. Some concentrating solar electric generators use a hybrid renewable/fossil fuel system to diversify their energy source.

California's Mojave Desert is home to the world's largest concentrating solar power facility. The Solar Electric Generating System plants have a combined capacity of 354 MW and are configured as hybrids. The nine hybrid plants generate more than 75% of the total power output from the sun, with natural gas providing the balance.

TABLE I

Cost and Performance of Distributed Generation Technologies for Cogeneration Systems[a]

Technology	Reciprocating engine	Turbine and microturbine	Fuel cell
Size	30 kW–8 MW	30 kW–20 + MW	100–3000 kW
Installed cost ($/kW)[b]	300–1500	350–1500	2000–5000
Electricity efficiency (LHV)	28–42%	14–40%	40–57%
Overall efficiency[c]	~80–85%	~85–90%	~80–85%
Variable O&M ($/kWh)	0.0075–0.02	0.004–0.01	0.002–0.05
Footprint (ft^2/kW)	0.22–0.31	0.15–0.35	0.9
Emissions (lb/kWh unless otherwise noted)			
Diesel			
NO_x	0.22–0.025	3–50 ppm	<0.00005
CO	0.001–0.002	3–50 ppm	<0.00002
Natural gas			
NO_x	0.0015–0.037		
CO	0.004–0.006		
Fuels	Diesel, natural gas, gasoline, digester gas, biomass and landfill gas; larger units can use dual-fuel (natural gas-diesel) or heavy fuels	Natural gas, diesel, kerosene, naphtha, methanol, ethanol, alcohol, flare gas, digester gas, biomass and landfill gas	Natural gas, propane, digester gas, biomass and landfill gas (potentially)

[a] *Source*. Resource Dynamics Corporation (2002).
[b] Cost varies significantly based on siting and interconnection requirements as well as unit size and configuration.
[c] Assuming cogeneration.

2.4 Thermally Activated Technologies

Thermally activated technologies (TATs) take advantage of the waste heat from prime movers. There are several types of TATs.

First, absorption chillers provide cooling to buildings by using heat. This seemingly paradoxical but highly efficient technology is most cost-effective in large facilities with significant heat loads. Absorption chillers not only use less energy than conventional equipment but also cool buildings without the use of ozone-depleting chlorofluorocarbons.

Unlike conventional electric chillers, which use mechanical energy in a vapor compression process to provide refrigeration, absorption chillers primarily use heat energy with limited mechanical energy for pumping. These chillers can be powered by natural gas, steam, or waste heat.

An absorption chiller transfers thermal energy from the heat source to the heat sink through an absorbent fluid and a refrigerant. The absorption chiller accomplishes its refrigerative effect by absorbing and then releasing water vapor into and out of a lithium bromide solution. The process begins as heat is applied at the generator and water vapor is driven off to a condenser. The cooled water vapor then passes through an expansion valve, where the pressure reduces. The low-pressure water vapor then enters an evaporator, where ambient heat is added from a load and the actual cooling takes place. The heated, low-pressure vapor returns to the absorber, where it recombines with the lithium bromide and becomes a low-pressure liquid. This low-pressure solution is pumped to a higher pressure and into the generator to repeat the process.

Absorption chiller systems are classified by single-, double-, or triple-stage effects, which indicate the number of generators in the given system. The greater the number of stages, the higher the overall efficiency. Double-effect absorption chillers typically have a higher initial cost, but a significantly lower energy cost, than single-effect chillers, resulting in a lower net present worth. Triple-effect systems are underdevelopment.

Chillers employ cogeneration thermal output during cooling periods when heating uses are limited to domestic hot water loads or zonal heating, which may be small in many building types. These units involve a complex cycle of absorbing heat from the cogeneration system to create chilled water. The

waste heat from the cogeneration system is used to boil a solution of refrigerant/absorbent. Most systems use water and lithium bromide for the working solution. The absorption chiller then captures the refrigerant vapor from the boiling process and uses the energy in this fluid to chill water after a series of condensing, evaporating, and absorbing steps are performed. This process is essentially a thermal compressor, which replaces the electrical compressor in a conventional electric chiller. In so doing, the electrical requirements are significantly reduced, requiring electricity only to drive the pumps that circulate the solution by single-effect chillers (Table II).

Second, engine-driven chillers (EDCs) are essentially conventional chillers driven by an engine instead of an electric motor. They employ the same thermodynamic cycle and compressor technology as electric chillers but use a gas-fired reciprocating engine to drive the compressor. As a result, EDCs can provide cooling economically where gas rates are relatively low and electric rates are high. EDCs also offer better variable-speed performance, which yields improved partial load efficiencies (Table III).

Third, desiccant dehumidifiers use thermal energy to dehumidify and to some extent control the indoor air quality of building space. With the recent push toward building humidity control, manufacturers of desiccant dehumidification systems have developed new packages designed to treat building ventilating air efficiently. A desiccant dehumidifier uses a drying agent to remove water from the air being used to condition building space. Desiccants can work in concert with chillers or conventional air-conditioning systems to significantly increase energy system efficiency by allowing chillers to cool low-moisture air. Desiccants can run off the waste heat from distributed generation technologies, with system efficiency approaching 80% in cogeneration mode.

The desiccant process involves exposing the desiccant material (e.g., silica gel, activated alumina, lithium chloride salt, and molecular sieves) to a high relative humidity air stream—allowing it to attract and retain some of the water vapor—and then to a lower relative humidity air stream, which draws the retained moisture from the desiccant. In some applications, desiccants can reduce cooling loads and peak demand by as much as 50%, with significant cost savings.

In stand-alone operations, a solid (dry) desiccant removes moisture from the air; a heat exchanger and evaporative cooler then chill the air. A desiccant dehumidifier wheel, composed of lightweight honeycomb or corrugated matrix material, dries the air as the wheel rotates through the supply air before reaching the building. Once the desiccant material reaches its saturation point, heat must be added or the material must be replaced to regenerate the moisture-absorbing capability. Natural gas, waste heat, or solar energy can be used to dry moisture that

TABLE II

Cost and Performance of Single-Effect, Indirect-Fired Absorption Chillers[a]

Tons	Cost ($/ton)	Electric use (kW/ton)	Thermal input (Mbtu/ton)	Maintenance cost ($/ton annual)
10–100	700–1200	0.02–0.04	17–19	30–80
100–500	400–700	0.02–0.04	17–18	20–50
500–2000	300–500	0.02–0.05	17–18	10–30

[a] *Source*. Resource Dynamics Corporation (2002).

TABLE III

Cost and Performance of Engine-Driven Chillers[a]

Tons	Cost ($/ton)	Electric use (kW/ton)	Thermal input (Mbtu/ton)	Maintenance cost ($/ton annual)
10–100	800–1050	0.05–0.07	9–12	45–100
100–500	650–950	0.01–0.05	8–11	35–75
500–2000	450–750	0.003–0.01	7–8	25–60

[a] *Source*. Resource Dynamics Corporation (2002).

is either absorbed or collected on the surface to regenerate the material.

Liquid absorption dehumidifiers spray the air with a desiccant solution that has a lower vapor pressure than that of the entering air stream. The liquid has a vapor pressure lower than water at the same temperature, and the air passing over the solution approaches this reduced vapor pressure. The conditioner can be adjusted so that it delivers air at the desired relative humidity. Liquid desiccant materials include lithium chloride, lithium bromide, calcium, chloride, and triethylene glycol solutions (Table IV).

Finally, hybrid power systems combine different power generation devices or two or more fuels for the same device. When integrated, these systems overcome limitations inherent in either one. Hybrid power systems may feature lower fossil fuel emissions as well as continuous power generation when intermittent renewable resources, such as wind and solar, are unavailable.

Solar concentrators/chiller systems use sunlight, which is plentiful in many regions of the country and is a load follower for afternoon cooling demand, to power an absorption cooling in low- to medium-temperature-driven systems, with either nontracking or tracking solar energy collectors and a two-stage absorption chiller. The resultant hybrid systems can produce an average cooling season solar fraction of at least 60% (40% of the cooling load supplied by natural gas) so that the net primary source fuel COP for the fossil fuel used is 3.0. Therefore, solar/gas absorption cooling systems will consume one-third the average primary source energy used for electric air-conditioning in the United States.

The widespread use of solar/thermal systems would have a significant impact on reducing the demand for electric power and for extending natural gas reserves. Every 1000 tons of installed solar/natural gas absorption air-conditioning equipment relieves the electric power grid of an average of 1 MW of demand.

3. APPLICATIONS AND MARKETS

New cogeneration project growth throughout the United States has been curtailed because of natural gas and oil prices and price volatility, making spark spread an unreliable cogeneration potential indicator. Natural gas price volatility and historic high prices have dampened new cogeneration natural gas-fueled projects in general. Market factors causing these price excursions are complex and not fully understood. Concurrently, there has been a similar increase in oil prices.

Because of the uncertainty in fuel prices and availability, spark spread is a difficult dimension to quantify or qualify, and industry growth should take precedence over other standard measures of cogeneration economic feasibility such as spark spread—the cost differential between electricity generated from central plants vs distributed generation—according to the American Council for an Energy-Efficient Economy (ACEEE).

Cogeneration systems can be especially useful in areas of the country in which development is constrained due to poor air quality. This is particularly significant in older, industrial cities such as Chicago. In air emission zones in California, the Northeast, and the Midwest, new development can create the need for emissions offsets. Cogeneration systems can meet this requirement, especially when using output-based emission standards that capture these system's high fuel efficiencies.

3.1 Utilities

Grid reliability could be a strong driver for cogeneration growth; for example, New York City's power system needs an additional 7100 MW to enhance reliability, according to the Center for Business Intelligence. With utility restructuring, access to downtown generation capacity from a cogeneration

TABLE IV

Cost and Performance of Desiccant Dehumidification Systems[a]

SCFM	Cost ($/SCFM)	Thermal input (hourly Btu/SCFM)	Maximum latent removal (hourly Btu/SCFM)
1500–5000	8–18	30–100	30–60
5000–10,000	6–11	30–100	30–60
10,000 +	6–9	30–100	30–60

[a] *Source*. Resource Dynamics Corporation (2002).

facility may be more attractive and less problematic to the local electric utility in terms of projected load loss.

Addressing the need for distributed energy systems to interconnect to the electric utility grid, and to obtain supplemental and backup power, is critical to the success of cogeneration projects throughout the United States. Including multiple, dispersed generating and cogeneration units throughout the grid is a relatively new concept—one that has to be incorporated into the existing technical, regulatory, and institutional framework.

Some utilities are considering investing in the cogeneration value proposition to not only help them meet air quality objectives in their service territories but also to respond to customer service demands. Some utilities understand that cogeneration systems can significantly lower power plant, offer affordable incremental power costs, help optimize natural gas resources, hold gas costs down, and may open the door to new thermal energy business.

Also, if regulators do not permit, or if utilities disincent, cogeneration competition on an interconnected basis, end users will be forced to install the on-site systems on an "islanded" basis, leaving electric utilities out of the energy business. Now is an opportune time for utilities to embrace cogeneration as part of a national energy security campaign and as part of their own strategic business plans.

3.2 Industrial

Cogeneration has been most successful in large industrial applications that require large amounts of steam, but it is also well suited to use in hospitals, laundries, and health clubs. Since 1980, approximately 50,000 MW of cogeneration capacity has been built in the United States.

Cogeneration in the large to medium industrial market is typically found in the process industries, such as petroleum refining, pulp and paper, and chemicals. These systems have installed electricity capacities >25 MW (often hundreds of megawatts) and steam generation rates measured in hundreds of thousands of pounds of steam per hour. This sector represents the largest share of the current installed cogeneration capacity in the United States and is the segment with the greatest potential for short-term growth. Some facilities of this type are merchant power plants using combined cycle configurations. They are owned by an independent power producer that seeks an industrial customer for its waste stream and sells excess electricity on the wholesale market.

Thousands of boilers provide process steam to a broad range of small industrial U.S. manufacturing plants. These boilers offer a large potential for adding new electricity generation between 50 kW and 25 MW by either modifying boiler systems to add electricity generation (e.g., repowering existing boilers with a combustion turbine) or replacing the existing boiler with a new cogeneration system. Small manufacturers represent an important growth segment during the next decade.

Various industrial markets, including petrochemical, food products, bioproducts, and biotech, rank as the highest priority markets for future industry growth. The petrochemical market is the largest cogeneration market sector nationally, representing approximately 40% of existing cogeneration capacity. Bulk chemicals are in economic decline domestically. Some sectors of the chemicals industry are continuing to experience significant growth, including pharmaceuticals, specialty chemicals, and bioproducts including ethanol and biofeedstocks.

Food products manufacturing is a fast growing and stable market with demonstrated opportunity for cogeneration. The ACEEE has identified food products as one of the most geographically dispersed industry groups, with a significant presence in almost all states. In contrast to wood products and bulk chemicals industries, which are in economic decline domestically, the food products industry is among the fastest growing industry groups.

The growing biotech industry has overtaken the computer and semiconductor industry as a leader in projected economic growth. Both sectors have similar power reliability and thermal management needs. The biotech market has not experienced the same slowdown that has been seen in the computer-related markets.

3.3 District Energy

District energy systems are a growing market for cogeneration because these systems significantly expand the amount of thermal loads potentially served by cogeneration. District energy also has a major added benefit of reducing the requirement for size and capital investment in production equipment due to the diversity of consumer loads. These systems also tend to use larger and more efficient equipment and can take advantage of such things as thermal energy storage that are not economically effective on a small scale. Additionally, district energy systems aggregate thermal loads, enabling more cost-effective cogeneration. District energy systems may be installed

at large, multibuilding sites, such as universities, hospitals, and government complexes. District energy systems can also serve as merchant thermal systems providing heating (and often cooling) to multiple buildings in urban areas.

According to the International District Energy Association (IDEA), there are three primary markets for district energy: colleges and universities, downtowns, and airports. The college and university market seems to be the most promising due to several factors. Colleges and universities are installing cogeneration in response to campus load growth, asset replacement of aging boiler capacity, and the favorable economics from fuel efficiency improvements with cogeneration.

Frequently, large campuses centralize heating and cooling and have large electrical loads. Many options are available for satisfying this electrical, heating, and cooling demand. A common way is to size a boiler for the entire campus's heating load. Large facilities may be able to negotiate attractive fuel contracts that lower heating costs and may choose to oversize the boiler, enabling the use of excess steam to power a steam turbine for electrical generation or a steam-driven chiller. This cheap, on-site secondary power source (steam) may make the generation of electricity or cooling less expensive than conventional methods such as purchasing from the grid and using conventional electric chillers.

The University of North Carolina at Chapel Hill, the University of Texas at Austin, and Cornell University represent examples of this situation. In each case, the universities have large boilers making steam for a variety of uses. Each campus ties the boilers to a steam turbine and the campus's heating system. The steam turbine generates electricity for campus use, reducing the dependency on the local grid and saving operating dollars for the campus. This cheap electricity may also be used to operate electric chillers for cooling, further reducing cooling costs. The remainder of the steam is used for heating of domestic water and all the campus's office, classroom, and dormitory spaces. Steam generated by the boilers may also be used by steam-driven chillers for cooling campus living space.

IDEA reports that college and university markets have significant pressures for "greening the campus," especially in the Northeast. Princeton, MIT, Rutgers, and the New Jersey Higher Education Partnership for Sustainability are exemplary, but there are other promising college and university examples throughout the country.

IDEA also notes that many colleges and universities conduct critical research and therefore need reliable power to avoid losing months or years of research and to assure students and the community that test animals are being properly cared for. Cogeneration could provide both power reliability and space conditioning.

Major urban centers are also a very promising market for adding cogeneration to existing district energy systems. Many district energy steam plants were originally cogeneration facilities that generated both power and steam when owned by the local electric utility. There is a growing need for local grid support in light of utility divestiture of generating capacity coupled with solid market growth in downtown district energy systems.

Airports represent another promising opportunity; these facilities are often in NO_x nonattainment areas and face significant emissions pressure from both regulators and the community. With large space conditioning and electrical load with long hours of operation, airports are often well suited to add cogeneration to their district energy systems.

3.4 Federal Government

The federal government is the largest energy consumer in the United States, with new mandates to meet increased demand, reduce peak operating costs, enhance energy security, and improve the reliability of electric power generation through DG and cogeneration. The Federal Energy Management Program was created to reduce government costs by advancing energy efficiency, water conservation, and the use of solar and other renewable energies. Executive Order 13123, "Greening the Government through Efficient Energy Management,: specifies that federal facilities shall use combined cooling, heat, and power systems when lifecycle costs indicate that energy reduction goals will be achieved. This market sector resembles the district energy sector, although the two sectors have different drivers and budgetary pressures.

According to the ACEEE, the largest cogeneration potential at federal facilities has been identified at military facilities, many of which are managed by various agencies within the Department of Defense.

3.5 Commercial/Institutional

The United States consumed more than 94 quadrillion Btus of energy in 2000. Of this total,

commercial buildings accounted for more than 16 quadrillion Btus—equivalent to the amount of energy of gasoline consumed in the United States in 1 year. The growth of the economy, as well as the nation's increasing population, is leading to greater numbers of larger and more energy-intensive homes and commercial buildings, resulting in increased energy consumption in this sector.

Several reports indicate CHP potential and interest in the large commercial office buildings, hospital/health care, supermarket, hotel/motel, restaurant, and large retail markets. Some markets appear to have strong drivers to adopt CHP; for example, grocery stores have significant dehumidification loads to keep water from condensing on freezer displays, and hospitals and health care facilities have significant reliability needs. Schools appear to be a prime market for desiccant dehumidifiers, which control humidity and therefore reduce the likelihood of mold and bacteria growth.

In 2002, the Federal Communications Commission (FCC) adopted a plan that will give consumers access to digital programming over television by requiring off-air digital television (DTV) tuners on nearly all new TV sets by 2007. By enacting a 5-year rollout schedule that starts with larger, more expensive TV sets, the FCC is minimizing the costs for equipment manufacturers and consumers but creating new demands on power reliability for TV stations.

Commercial brownfield sites, or "real property, the expansion, redevelopment, or reuse of which may be complicated by the presence or potential presence of a hazardous substance, pollutant, or contaminant," are an appealing market for cogeneration. Redeveloping contaminated brownfield sites removes serious environmental hazards. Cogeneration can play an important role in the effort. Suzanne Watson of the Northeast-Midwest Institute notes that "with the right new technology fixes, there is the potential to create developments that improve the natural environment. Cogeneration is one of these technology solutions."

The DOE is helping manufacturers work together to integrate their individual CHP components into easier to use packages and to make "plug-and-play" packages more readily available in the commercial marketplace. On August 8, 2001, Energy Secretary Abraham announced awards to seven industry teams for research, development, and testing of "first-generation packaged cooling, heating, and power systems.

3.6 Residential

The DOE and the U.S. Department of Housing and Urban Development (HUD) are currently encouraging residential cogeneration units for individual or aggregated loads, which could be achieved in the multifamily markets under HUD jurisdiction.

There is significant potential for cogeneration in the privately owned assisted and public housing multifamily markets, where electric and thermal loads are already aggregated. According to HUD, the assisted housing market has a greater number of properties with less than 50 units, as opposed to the public housing market (20 units was used as a reasonable cogeneration entry point for both markets). In the assisted housing market, there are approximately 109,000 HUD properties with 20–49 units and 53,000 properties with more than 50 units. In the public housing market, there are approximately 4535 HUD properties with 20–49 units and 7154 properties with more than 50 units.

4. REGULATORY AND RATE STRUCTURE ISSUES

4.1 Standardizing Interconnect

Addressing the need for distributed energy systems to interconnect to the electric utility grid, and to obtain supplemental and backup power, is critical to the success of cogeneration projects. Until recently, the U.S. electric system was based on central station generation, tied to customers through the use of transmission lines. Including multiple, dispersed generating, and cogeneration units throughout the grid is a relatively new concept—one that has to be incorporated into the existing technical, regulatory, and institutional framework.

The DOE has been working for several years with industry through the Institute of Electrical and Electronics Engineers (IEEE) to develop a uniform national standard for interconnection. The IEEE P1547, "Standard for Distributed Resources Interconnected with Electric Power Systems," will provide interconnection guidelines for performance, operation, testing, safety, and maintenance. It is hoped that many utilities and state utility regulators will adopt it, giving their customers more power generation choices.

The Federal Energy Regulatory Commission has a rule making under way to establish standards and practices for interconnection of DG systems. The comment period for the proposed rule closed in early 2002.

In some cases, cogeneration projects would be able to proceed if the interconnection procedures for generators were standardized through a process of law making or rule making with the state simplifying, or at least setting, a recognized standard for the process. Both Texas and California have developed state-level interconnection standards (CEC 2001 and Texas 2000). These standards represent a compromise between utilities and generators. Although the guidelines differ in some specifics and main motivation, both create a relatively rapid and standard procedure for all generators seeking to connect to the grid.

The Texas Public Utility Regulatory Act of 1999 granted all utility customers access to DG. On February 4, 1999, the Texas Public Utility Commission (PUCT) adopted interconnection standards (PUCT 1999). PUCT continued to investigate DG and in May 2002 published a guidebook to interconnection in the state.

California's decision to expedite the streamlining of interconnection procedures resulted from their well-publicized energy problems in 2000 and 2001. Although there are many reasons why California had to resort to rolling blackouts, increasing DG is a way to remove pressure from the grid and reduce the likelihood that the event will be repeated.

The California Energy Commission (CEC) completed a state DG plan (CEC 2002), which included a plan for creating standardized guidelines for interconnection to the grid. In 2002, the CEC's Rule 21 Interconnection Working Group completed this guideline. Soon thereafter, the California Public Utilities Commission adopted the standards. Individual utility versions of the interconnection standards can be found at www.energy.ca.gov.

4.2 Output-Based Standards

Implementation of output-based standards, pollution allowances per kilowatt-hour, or useful electrical and thermal energy recognizes that cogeneration systems achieve significant reductions in environmental emissions due to their much higher efficiencies of fuel conversion relative to conventional separate heat and power systems. These standards would measure environmental emissions, or the rate at which emissions are discharged by the technology to the environment per unit of total usable energy produced by the technology as opposed to emissions per unit of fuel consumed. These standards should value electric and nonelectric output equally and be independent of fuel source. They should not discriminate against different ownership situations.

A promising development that the U.S. Environmental Protection Agency (EPA) is sponsoring is "demand response" run by regional transmission organizations. End users can bid back to the grid pool to reduce their electricity use by using on-site power or energy efficiency at set rates.

4.3 Time-of-Use Pricing

Time-of-use pricing is a rate structure intended to encourage customers to shift their energy use away from peak hours when wholesale electricity prices are highest. Real-time pricing can be readily used in large commercial and industrial facilities to manage thermal loads, making cogeneration more cost-effective. Chilled water and ice thermal energy storage, for example, are effective means to shift load to off-peak periods.

5. U.S. FEDERAL GOVERNMENT SUPPORT

The U.S. federal government is supporting cogeneration at the highest administration levels as well as through various agency programs and association support.

5.1 Administration

Released in May 2001, the Bush administration's National Energy Plan recognized the important role that cogeneration can play in meeting national energy objectives and maintaining comfort and safety of commercial and office buildings. Section 3–5 states:

> *A family of technologies known as combined heat and power (CHP) can achieve efficiencies of 80% or more. In addition to environmental benefits, CHP projects offer efficiency and cost savings in a variety of settings, including industrial boilers, energy systems, and small, building scale applications. At industrial facilities alone, there is potential for an additional 124,000 MW of efficient power from gas-fired CHP, which could result in annual emissions reductions of 614,000 tons of NO_x emissions and 44 million tons of carbon equivalent. CHP is also one of a group of clean, highly reliable distributed energy technologies that reduce the amount of electricity lost in transmission while eliminating the need to construct expensive power lines to transmit power from large central power plants.*

5.2 U.S. Environmental Protection Agency

Currently, power plants are responsible for two-thirds of the nation's annual SO_2 emissions, one-fourth of

the NO_x emissions, one-third of the mercury emissions, and one-third of the CO_2 emissions, a leading greenhouse gas. These emissions contribute to serious environmental problems, including global climate change, acid rain, haze, acidification of waterways, and eutrophication of critical estuaries. They also contribute to numerous health problems, such as chronic bronchitis and aggravation of asthma, particularly in children.

The EPA understands that resolving these problems must start with pollution prevention, which equates to using fewer energy resources to produce goods and services. The National Energy Plan includes four specific recommendations to promote efficient cogeneration, three of which were directed to the EPA for action:

1. Promote cogeneration through flexible environmental permitting.
2. Issue guidance to encourage development of highly efficient and low-emitting cogeneration through shortened lead times and greater certainty.
3. Promote the use of cogeneration at abandoned industrial or commercial sites known as brownfields.

As a follow-up to these recommendations, the EPA joined with 18 *Fortune* 500 companies, city and state governments, and nonprofits in February 2002 in Washington, DC, to announce the EPA Combined Heat and Power Partnership. Cogeneration advances cogeneration as a more efficient, clean, and reliable alternative to conventional electricity generation.

As another follow-up to the National Energy Plan, the EPA is continuing work on a guidance document that clarifies how source determinations should be made for district energy and cogeneration projects under the New Source Review and Title V permitting regulations. The EPA intends for this guidance to help permitting officials and project developers streamline the development and permitting of these projects nationwide.

Through multiple programs, the DOE and its network of national laboratories have been working with manufacturers, end users, and other government offices to support increased use of cogeneration technologies.

5.3 U.S. Department of Energy

The DOE's Office of Distributed Energy Resources coordinates the CHP Initiative to raise awareness of the energy, economic, and environmental benefits of cogeneration and highlight barriers that limit its increased implementation. The initiative supports a range of activities, including regional, national, and international meetings; industry dialogues; and development of educational materials.

The DOE is also helping manufacturers work together to integrate their individual cogeneration components into easier to use packages and to make plug-and-play packages more readily available in the commercial and industrial marketplace. On August 8, 2001, Energy Secretary Abraham announced awards to seven industry teams for research, development, and testing of first-generation packaged cooling, heating, and power systems.

Integrated energy systems (IES) is a new approach to integrating all types of energy technologies into a building's energy system, including DG, cogeneration, HVAC, doors, windows, distribution systems, controls, insulation, building materials, lighting, and other building equipment. The link between building design and energy use is key to IES. As a DOE National User Facility, Oak Ridge National Laboratory (ORNL) provides manufacturers access to the user facility to help them evaluate and optimize the performance of IES and develop IES certification protocols.

The IES test center at the University of Maryland is designed to show building owners and managers how to save energy without sacrificing comfort in their buildings. The center is testing various cogeneration systems in an office building on campus. Phil Fairchild, program manager for Cooling, Heating, and Power for ORNL, explains,

> *What we're trying to do at the test center is to put together different pieces of equipment, use recoverable energy, and in doing so, understand the science of integration so that we can assist or better advise manufacturers about how to integrate the equipment in the future. If we reach our goals, manufacturers will have a better incentive, in terms of efficiency gains, to offer packaged or modular systems to the commercial building sector.*

The DOE is supporting a series of regional CHP application centers to provide application assistance, technology information, and education to architects and engineering companies, energy services companies, and building owners in eight states. The Midwest CHP application center was the first, established in March 2001 as a partnership between the DOE, the University of Illinois at Chicago's Energy Resources Center, the Gas Technology Institute, and ORNL. The center's educational material includes a Web site, accredited training classes, engineering tool kits, case studies, and application screening tools.

The DOE's Industrial Technologies Program (ITP) coordinates a combustion program with the industrial combustion community. The program promotes research and development of combustion systems used to generate steam and heat for vital manufacturing processes, heat-processing materials from metals to chemical feedstocks, and change the mechanical and chemical properties of materials and products.

The Steam Challenge is a public–private initiative sponsored by ITP. The goal of Steam Challenge is to help industrial customers retrofit, maintain, and operate their steam systems more efficiently and profitably. Steam Challenge provides steam plant operators with tools and technical assistance while promoting greater awareness of the energy and environmental benefits of efficient steam systems.

SEE ALSO THE FOLLOWING ARTICLES

Coal, Fuel and Non-Fuel Uses • Conservation of Energy, Overview • District Heating and Cooling • Energy Efficiency, Taxonomic Overview • Fuel Cells • Industrial Energy Efficiency • Solar Thermal Power Generation • Thermal Energy Storage

Further Reading

Garland, P., Hinrichs, D., and Cowie, M. (2003). Integrating CHP into commercial buildings. *Cogeneration On-Site Power Production*, May/June.

Hinrichs, D. (2002). CHP in the United States: Gaining momentum. *Cogeneration On-Site Power Production*, July/August.

Jimison, J., Thornton, R., Elliott, N., and Hinrichs, D. (2003). Emerging drivers for CHP market development in the U.S. *Cogeneration On-Site Power Production*, July/August.

Oak Ridge National Laboratory (2002). "Analysis of CHP Potential at Federal Sites." Oak Ridge National Laboratory, Oak Ridge, TN.

Onsite Sycom Energy (2000). "The Market and Technical Potential for Combined Heat and Power in the Commercial/Institutional Sector." Onsite Sycom Energy.

Onsite Sycom Energy (2000). "The Market and Technical Potential for Combined Heat and Power in the Industrial Sector." Onsite Sycom Energy.

Resource Dynamics Corporation (2002). "Integrated Energy Systems (IES) for Buildings: A Market Assessment—Final Report." Resource Dynamics.

U.S. Combined Heat and Power Association, U.S. Department of Energy, and U.S. Environmental Protection Agency (2001). "National CHP Roadmap: Doubling Combined Heat and Power Capacity in the United States by 2010." U.S. Combined Heat and Power Association, U.S. Department of Energy, and U.S. Environmental Protection Agency.

Combustion and Thermochemistry

SHEN-LIN CHANG[†]
Argonne National Laboratory
Argonne, Illinois, United States

CHENN QIAN ZHOU
Purdue University Calumet
Hammond, Indiana, United States

1. **Introduction**
2. **Combustion Elements**
3. **Thermochemistry**
4. **Hydrodynamics and Mixing**
5. **Chemical Reactions**
6. **Radiation Heat Transfer**
7. **Tools for Combustion Analysis**
8. **Pollution Control Techniques**
9. **Concluding Remarks**

Glossary

adiabatic flame temperature The highest flame temperature that is achievable under the condition of no external heat transfer.
combustion The burning of fuel and oxidant to produce heat and/or work.
computational fluid dynamics (CFD) Computer program for calculating fluid dynamics properties of a flow.
diffusive flame Separate fuel and oxidant inlets before burning takes place.
flammability Combustion reaction criteria.
gasification Incomplete fuel-rich combustion that produces CO and/or H_2.
ignition Initial conditions for sustained combustion.

Much solar energy is stored in hydrocarbons. Early humans, seeking protection and comfort, learned to convert hydrocarbon energy to fire to survive. Combustion, the most important form of energy conversion, is a chemical reaction between a hydrocarbon fuel and an oxidant. Combustion generates significant amounts of light, heat, and/or mechanical work. Furnaces and combustors have been developed to generate electricity and manufacture tools and equipment. Combustion engines have enabled transportation in every region of Earth and even into space. In recent years, concerns over energy availability and the environmental impacts of combustion processes have attracted more and more attention. To use energy resources efficiently and responsibly has become a main goal of many societies. To achieve this goal, knowledge of combustion science has to be learned and applied.

1. INTRODUCTION

Combustion includes thermal, hydrodynamic, and chemical processes. Combustion starts with the mixing of fuel and oxidant, and sometimes in the presence of additional compounds or catalysts. The fuel/oxidant mixture can be ignited with a heat source; chemical reactions between fuel and oxidant then take place and the heat released from the reactions makes the process self-sustaining. Instability may occur, inhibiting or strengthening combustion. Combustion products include heat, light, chemical species, pollutants, mechanical work, and plasma. Sometimes, a low-grade fuel, e.g., coal, biomass, and coke, can be partially burned to produce a higher grade fuel, e.g., methane. The partial burning process is called gasification. Various combustion systems, e.g., furnaces, combustors, boilers, reactors, and engines, have been developed to utilize combustion heat, chemical species, and

[†]Deceased.

 © 2004 Elsevier Inc. All rights reserved.

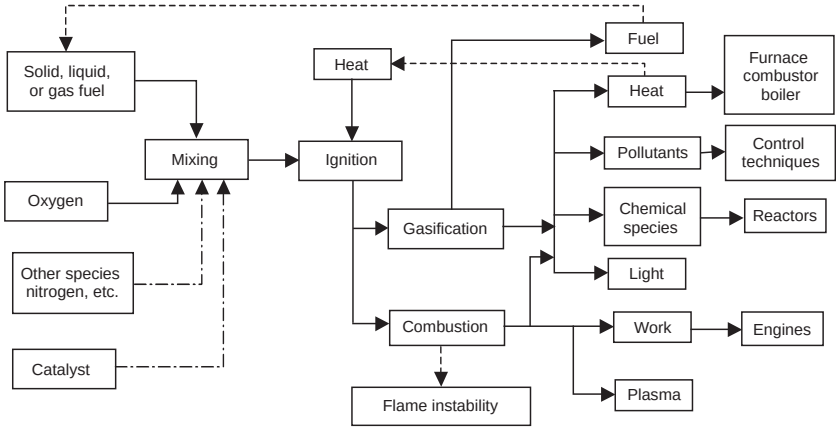


FIGURE 1 Typical combustion processes.

work. Figure 1 shows a flow diagram of typical combustion processes.

Advanced measurement/control devices have been traditionally used as tools for analysis of combustion systems. The analysis can lead to lower energy use and lower pollutant emissions by a system. In recent years, the computational analyses, including computer models of kinetics and fluid dynamics, have been used as an added tool for combustion analysis. The focus here is to assess the overall process of analyzing a combustion system.

A computational combustion analysis starts with the identification of the participating elements and their thermochemical properties and relations. This is followed by derivation of the governing equations and development of the required phenomenological models, including turbulent mixing, droplet evaporation, particle devolatization, radiation heat transfer, and interfacial interactions. Then, the governing equations are solved numerically on a computer. The results are validated with experimental data. Thus, the validated computer code can be used to evaluate the combustion system and to control combustion products.

2. COMBUSTION ELEMENTS

The elements (or species) involved in a combustion process include reactants, inert and intermediate species, products, and catalysts. Reactants generally include a fuel and an oxidant. Fuel can be in solid, liquid, or gas phase. Solid fuels include coal and biomass. Liquid fuels include hydrogen, gasoline, diesel, kerosene, and jet fuels. Gas fuels include natural gas, methane, and propane. The oxidant is mostly oxygen but may include F_2O, F_2O_2, NF_3, ClF_3, ClF_3, ClO_3F, O_3, and N_2F_4 in some rocket applications. The fuel and the oxidant can also be formed as a monopropellant in rocket applications. Catalysts are sometimes used to enhance combustion.

Inert species are not directly involved in the combustion process, but at high temperatures these species can be dissociated or ionized. These species include the nitrogen in the air and the ash in coal. Major combustion products include CO_2 and H_2O. Because these species can emit and absolve radiation energy, they are regarded as greenhouse gases that are related to the global warming issue. Many intermediate species, such as OH, CO, CH, are present during combustion. Combustion products also include some minor species, i.e., NO_x, SO_x, unburnt hydrocarbons (UHCs), and soot. These species are regarded as pollutants. NO_x and SO_x are believed to cause acid rain, UHCs play a role in smog, and soot chemicals, including polynuclear aromatics, may be carcinogenic.

In a combustion reaction, there can be so many combustion elements that it is impossible to analyze them all. Therefore, many combustion elements are lumped into one category for the convenience of analysis. For example, coal consists of thousands of hydrocarbons, but in most coal combustion analyses, the hydrocarbons are grouped into four categories: moisture, volatiles, char, and ash. The moisture is water, the volatiles represent all of the hydrocarbons that can be vaporized during combustion, the char includes all hydrocarbons that cannot be vaporized, and the ash is inert material, mostly metals.

3. THERMOCHEMISTRY

Assume that Reaction (1) represents a combustion reaction of I lumped reactants and J lumped products:

$$\sum_{i=1}^{I} n_i \text{ Reactants} \rightarrow \sum_{j=1}^{J} n_j \text{ Products.} \qquad (1)$$

In the reaction, the n_i values on the reactant side and n_j values on the product side are called the stoichiometric coefficients. The coefficients can be on a molar or a mass basis. On a molar basis, the unit is moles; on a mass basis, the unit is kilograms.

A computational combustion analysis calculates the heat, work, and/or species generated from a combustion process, which includes the combustion reaction, fluid dynamics, and heat transfer. Thermodynamic and chemical (thermochemical) properties of the combustion elements are needed to calculate the heat, work, and species. The thermochemical properties most often used in a combustion analysis include temperature, pressure, density, enthalpy, specific heat, viscosity, diffusivity, thermal conductivity, Gibbs free energy, molecular weight, species concentration, heat of combustion, latent heat, and equilibrium constant.

3.1 Equations of State

The thermochemical properties of a combustion species can be obtained by direct measurements or theoretical calculation. These properties are not all independent. Thermodynamic properties of a combustion species are generally functions of temperature and pressure. In most practical applications, these properties can be expressed in terms of temperature only. Sometimes, even constant property values are used. Most properties of a combustion species have been determined and tabulated in tables and other references produced by the joint interagency (Army, Navy, NASA, and Air Force) committee, JANNAF (formerly JANAF, before NASA was included). Properties of a mixture are determined by adding the properties of its components according to their concentrations. The partial pressure of a component is defined as the total pressure times its molar concentration.

Other than direct measurement, the thermochemical properties of members of a grouped ("lumped") combustion element (or species) can be obtained by adding the properties of its major constituents according to their proportion. For example, a fuel composed of 90% methane (CH_4) and 10% ethane (C_2H_6) becomes a lumped species $C_{1.1}H_{4.2}$. Its molecular weight is 17.1 kg/mol, calculated by adding 90% of 16 kg/mol (molecular mass of methane) and 10% of 30 kg/mol (molecular mass of ethane). Similarly, other properties of the lumped species can be obtained from those of the chemical species.

3.2 Chemical Equilibrium

At constant temperature and pressure, a combustion reaction such as Reaction (1) will reach an equilibrium state. The equilibrium constant K_p is defined as

$$K_p = \prod_{j=p} P_j^{n_j} / \prod_{i=r} P_i^{n_{ij}} \qquad (2)$$

in which the P_i values are partial pressures of the reactants (r) and products (p). The equilibrium constants of many combustion reactions have been determined and are tabulated in most combustion textbooks.

A combustion process generally consists of many reactions. The equilibrium species concentrations can be calculated from the equilibrium constants and the initial species concentrations. Bittker and Scullin developed a general chemical kinetics computer (GCKP) program to perform equilibrium calculations of hydrogen/air combustion and many other combustion processes.

3.3 Adiabatic Flame Temperature

Combustion of a fuel produces heat. When all of the heat is used to heat the products, without heat transfer from external sources, the final temperature is called the adiabatic flame temperature. The temperature can be determined by balancing the heat of combustion, the heat of dissociation, and the total enthalpy of the combustion products. At a combustion temperature higher than 2000 K, a small portion of the combustion products CO_2 and H_2O dissociates into intermediate species, such as CO, O, and OH, and the dissociation uses up a significant amount of the heat of combustion. Thus, the energy balance equation needs to be solved in conjunction with the equilibrium calculation to obtain an adiabatic flame temperature. Assuming the initial temperature and pressure are under standard conditions, the adiabatic flame temperature of a methane/air mixture is about 2200 K. Chang and Rhee developed a set of empirical formulations that can be used to estimate adiabatic flame temperature of lean fuel/air mixtures. The adiabatic flame temperature is an important thermochemical property in a

combustion analysis because it is the upper bound of the temperature.

4. HYDRODYNAMICS AND MIXING

A fuel and an oxidant must mix to react. The mixing process is a part of flow dynamics. Gas phase mixing is mainly due to turbulence; liquid injection in droplet sprays is commonly used to enhance the mixing between liquid droplets and gas, and some solid fuels are pulverized to increase the mixing between the particles and gas.

4.1 Governing Equations

Combustion generates heat, work, and product species. In a computational combustion analysis, heat is calculated from the temperature and velocity; work is calculated from pressure and density; and the species are calculated from concentrations and velocities. The governing equations of these properties are derived from fundamental thermodynamic, hydrodynamic, and chemistry principles. The derivation can be on a fixed mass basis or a control volume basis. In the following discussion, a control volume approach is used.

The principles of mass, species, and momentum conservation are used to determine pressure, concentration, and velocity. The first law of thermodynamics (or conservation of energy) is used to determine temperature. The equation of state is used to determine density. In deriving the governing equations, many derived properties are created. For example, enthalpy is used in the energy equation and specific heat is used in the caloric equations of state. Equation (3) is a typical energy equation derived on a Cartesian coordinate system:

$$\sum_{i=1}^{3} \frac{\partial}{\partial x_i}\left(\rho u_i h + \frac{\mu_e}{\sigma_h}\frac{\partial h}{\partial x_i}\right) = S_{\text{comb}} + S_{\text{rad}} + S_{\text{int}}, \quad (3)$$

in which x_i is a coordinate, ρ is density, u_i is velocity components, h is enthalpy, μ_e is effective viscosity, σ_h is a turbulent scale factor for the enthalpy, S_{comb} is the combustion heat release rate, S_{rad} is net radiation heat flux, and S_{int} is heat transfer from the liquid or solid phase. Effective diffusivity is used in the momentum, energy, and species equations. It is the sum of both laminar and turbulent diffusivities. The source terms of the governing equations need to be determined from various models. In the energy equation (3), there are three source terms. The combustion heat release rate needs to be calculated from combustion kinetics, radiation heat flux needs to be solved from a radiative transport equation, and the last term needs to be determined from an interfacial interactions model.

4.2 Turbulent Mixing

Diffusive flames, commonly used in a furnace or a combustor, are mostly turbulent. Turbulence is understood to be the small eddies that enhance mixing and combustion, but understanding the physics of the turbulence is still far from complete. An enclosure type of the $k-\varepsilon$ turbulence model is commonly used for combustion analysis. The model expresses the turbulent viscosity μ_t in terms of two turbulent parameters, kinetic energy k and dissipation rate ε,

$$\mu_e = \mu + \mu_t = \mu + C_\mu \rho k^2/\varepsilon,$$

in which C_μ is an empirical constant.

Lauder and Spalding proposed two additional transport equations to solve for these two turbulent parameters. Turbulent diffusivities for the enthalpy and species equations are also expressed in terms of the turbulent viscosity with a scaling factor. In a multiphase flow, solid particles or liquid droplets tend to inhibit eddy motion and reduce turbulent diffusion and the adjustment of the turbulent viscosity is needed. In addition, statistical turbulence models are used to describe the turbulence mixing effects. In this approach, probability density functions are chosen to represent the fluctuation of the flow properties. Large eddy and direct numerical simulations have been introduced with the hope of improving the description of turbulent flow.

4.3 Liquid Spray and Droplet Vaporization

Liquid fuel can be burned in pools or sprays. The burning processes start with the vaporization of the liquid fuel. Next, the fuel vapor is mixed with oxygen. Finally, the combustion reaction takes place. The pool burning is mostly fire. Because of limited contact area, the mixing of fuel vapor and oxygen is poor. Therefore, combustion is mostly fuel rich and a lot of smoke is produced. A fuel injector breaks up the liquid into tiny droplets and the high pressure makes droplets penetrate deep into the gas flow. Empirical models have been developed to describe the droplet size and number distributions off the tip of an injector. The penetration enhances the mixing of the fuel and the oxygen and the high ratio of surface area to volume speeds up the evaporation process.

Liquid fuel vaporizes in two modes. At low temperature, vaporization occurs due to the diffusion of mass. At a temperature higher than the boiling point, the liquid boils and the vaporization rate is controlled by the heat transfer rate from the surroundings to the liquid. In most combustion systems, the heat transfer is convective. Empirical correlations have been developed to calculate the interfacial heat transfer rate from other flow properties, such as, velocity, conductivity, and viscosity.

Liquid droplets are generally treated in a Lagrangian or an Eulerian approach. The Lagrangian approach takes one droplet at a time. It traces the motion of a single droplet in the flow and takes the average effect of all the droplets. Zhou and Yao also included the group effects on the spray dynamics. The Eulerian approach treats a group of droplets as a continuum. Governing equations are derived for each droplet group for solving the flow field.

4.4 Devolatization of Solid Fuel

Pulverized coal is used in many utility boilers. Due to the small particle size, pulverized coal can be transported pneumatically. Coal particles contain four major components: moisture, volatiles, char, and ash. When heated, the moisture is dried and the volatiles are devolatized. The volatile vapor mixes with oxygen and the mixture burns. The char left on the particles is burned with oxygen when contacted. The drying and devolatilization processes are controlled by the heat transfer from the gas to the particles. Similar to the liquid vaporization, the heat transfer is convective. Empirical correlations are used to determine the interfacial heat transfer rate. Similar Lagrangian and Eulerian approaches can be applied for particles.

5. CHEMICAL REACTIONS

The mixture of fuel and oxidant is burned in chemical reactions. A combustion reaction is the effective collisions of reactants (and catalyst), breaking the chemical bonds of the reactants, forming the products, and releasing heat. Three types of reactions are significant in a combustion analysis: ignition, heat release, and pollutant formation.

5.1 Ignition and Flammability

Fuel and oxidant can be self-ignited (autoignition) or by an external heat source. For example, diesel engine combustion is autoignition and gasoline engine combustion is ignited by a spark. At an ignition temperature, the combustion reactions generate heat and the heat is lost to the surroundings by conduction, convection, and radiation heat transfer. If the heat generation is larger than the heat loss, the combustion reactions are sustained.

There are mixture ratios of fuel and oxidant that the combustion will not sustain after the ignition source is removed. For premixed laminar flame, there exist flammability limits that are frequently defined as percentage fuel by volume in the mixture. The lower limit is the leanest mixture that will allow steady flame, and the upper limit represents the richest mixture. Flammability limits in both air and oxygen are found in the literature. For example, the lower flammability limit of methane is about 5% by volume.

5.2 Combustion Reactions

Assume that fuel is a hydrocarbon C_mH_{4n}. The combustion of the fuel and oxygen is shown in Eq. (4):

$$C_mH_{4n} + (m+n)O_2 \xrightarrow{k} mCO_2 + 2nH_2O. \quad (4)$$

The extent of reaction, ξ, is defined as the molar concentration of the consumed fuel. For each mole of ξ consumed, the sum $(m+n)$ moles of oxygen is consumed and m moles of CO_2 and $2n$ moles of H_2O are produced. The reaction rate $d\xi/dt$ of the combustion reaction can be expressed in an Arrhenius formulation:

$$\frac{d\xi}{dt} = -\frac{d}{dt}[C_mH_{4n}] = -\frac{1}{m+n}\frac{d}{dt}[O_2] = \frac{1}{m}\frac{d}{dt}[CO_2] = \frac{1}{2n}\frac{d}{dt}[H_2O], \quad (5a)$$

$$\frac{d\xi}{dt} = k_{30}\exp(-E/RT)[C_mH_{4n}][O_2]^{(m+n)}, \quad (5b)$$

in which [ ] represents molar concentration of a species, t is reaction time, k_0 is a preexponential constant, E is an activation energy, and R is the universal gas constant. The preexponential constant and the activation energy are empirical constants extracted from experimental data. Thus, the species consumption/generation rates can be obtained from Eqs. (5a) and (5b) and the heat release rate becomes

$$dq_{comb}/dt = (d\xi/dt)H_o, \quad (6)$$

in which H_o is the heat of combustion.

5.2.1 Gasification Reactions

Gasification is an incomplete combustion of the fuel. In a fuel-rich combustion, the intermediate species such as CO becomes a major product species. If water is added to the reaction, hydrogen is also produced. Because both CO and H_2 can be burnt later with oxygen and can release heat, the gasification has been used to convert low-grade fuels, such as, coal and biomass, to a higher grade gaseous fuel, CO and H_2. The gasification process is now also used to convert diesel or gasoline fuel to hydrogen for fuel-cell applications.

5.2.2 Heterogeneous Reactions

Heterogeneous reactions include char burning with oxygen and gaseous reactions on catalyst surfaces. Char burning is regarded as similar to a combustion reaction. Empirical constants are extracted from experimental data. The catalytic reaction is generally treated as a third-body reaction and an Arrhenius formula is used for analysis.

5.3 Pollution Reactions

In addition to the major heat release reaction, many intermediate species are produced in numerous combustion reactions. Pitz and Wesbrook have compiled hundreds of combustion reactions and the list is still growing. Some of the intermediate species attract more public attention because they are pollutants. Pollution is defined as the contamination of the environment. All matter emitted from combustion chambers that alters or disturbs the natural equilibrium of the environment must be considered as a pollutant. Pollutants of concern include particular matter, the sulfur oxides, unburned and partially burned hydrocarbons, nitrogen oxides, carbon monoxide, and carbon dioxide.

5.3.1 Soot Kinetics

Soot is formed in the fuel-rich region of a combustion flow. Because soot is a strong contributor to the radiation heat transfer, a soot model is needed to calculate soot concentration for the radiation heat transfer calculation. The following simplified soot model assumes that soot (st) is generated in the fuel-rich flame by the cracking of the fuel and is burnt up by oxygen. The soot formation/oxidation rate is expressed as

$$df_{\mathrm{st}}/dt = A_{\mathrm{O2}} f_{\mathrm{fu}} \exp(-E_2/RT) - A_{\mathrm{O3}} f_{\mathrm{st}} f_{\mathrm{ox}} T^{1/2} \exp(-E_3/RT). \quad (7)$$

5.3.2 NO_x Kinetics

NO_x is formed by combustion in the presence of nitrogen. There are three types of NO_x formation: thermal, prompt, and fuel. Thermal NO_x is formed in high-temperature reactions, prompt NO_x comes from reactions with intermediate species, and fuel NO_x comes from the nitrogen in the fuel, e.g., coal. The Gas Research Institute (GRI) has compiled a long list (over 300) of reactions related to NO_x formation. However, lumped reduce mechanisms with many less reactions are used in practice.

5.3.3 SO_x Formation

Many coals contain significant amount of sulfur. The sulfur reacts with oxygen during combustion and produces SO_x. Sulfur is also contained in many lubricants. In diesel combustion, SO is formed and becomes a major source of soot particle nucleation.

6. RADIATION HEAT TRANSFER

At combustion temperature, radiation is an important if not dominating mode of heat transfer. In a combustion process, radiatively participating species include soot and gaseous species such as CO_2 and H_2O. Heat is emitted, absorbed, and scattered in waves of various wavelengths. Assuming the scattering effect in the combustion flow is negligible, the radiative heat transfer becomes the balance of emissive and absorption powers. For each wavelength, local net radiation heat power is obtained by integrating the absorption of the incoming radiation from all other locations and subtracting the emitted power. A radiative transport equation is obtained by integrating the net radiation heat power over all wavelengths, Eq. (8):

$$q_r(x, y, z) = \int_0^\infty \left[\oint \kappa'_\lambda e'_{b\lambda}(T') e^{-\int \kappa'' dl''} d\nu' - \kappa_\lambda e_{b\lambda}(T) \right] d\lambda, \quad (8)$$

in which λ is wavelength, κ is volume absorptivity, l is optical length, ν is a control volume, and $e_{b\lambda}(T)$ is the blackbody radiation function. The spectral volumetric absorptivities of these media must be determined from gas and soot radiation models.

6.1 Gas Radiation Model

The H_2O species has five strong radiation bands, centered at wavelengths of 1.38, 1.87, 2.7, 6.3, and

20 μm, and the CO_2 species has six bands, at 2.0, 2.7, 4.3, 9.4, 10.4, and 15 μm. A wideband model was introduced to calculate total band absorptance of these species. For each band, species concentrations, pressure, and temperature are used to determine a set of semiempirical parameters: the integrated band intensity, the bandwidth parameter, and the line width parameter. A semiempirical correlation is then used to calculate the total band absorptance from these parameters.

6.2 Soot Radiation Model

If the scattering is negligible, the Rayleigh-limit expression of the soot volume absorptance $\kappa_{s\lambda}$ can be used:

$$\kappa_{s\lambda} = \frac{36n^2k(\pi/\lambda)f_v}{[n^2(1+k^2)+2]^2 + 4n^4k^2}. \qquad (9)$$

Soot volume absorptance is proportional to the soot volume fraction f_v and inversely proportional to the wavelength. The optical refraction indices, n and k, are weak functions of wavelength that can be derived from classical electromagnetic theory.

7. TOOLS FOR COMBUSTION ANALYSIS

Measurement and testing have long been the ways to analyze and evaluate combustion systems. However, computational software has now become an additional tool for combustion analysis.

7.1 Computational Fluid Dynamics

Combustion flow calculations adopt a control volume approach to convert the governing equations to algebraic equations on a staggered, discretized grid system. The algebraic equations are solved iteratively with proper boundary conditions. Exit flow conditions have a strong impact on the convergence of the calculation. Because most flows are not fully developed, the exit flow conditions that satisfy global mass balance appear to improve numerical convergence the most.

The major species flow calculation uses Patankar's SIMPLER computational scheme to solve the pressure-linked momentum equations. A major species flow calculation is considered to have converged if the local and global mass balances are smaller than a set of predetermined criteria. The subspecies flow is solved using the flow properties calculated from the major species flow calculation. Free from the interactions of the pressure and velocity fluctuations, the calculation of the partially decoupled subspecies transport equations becomes very stable numerically.

Chang and colleagues developed an advanced methodology to solve the radiative transfer equation (RTE) in a Cartesian coordinate system. The methodology guarantees the integrity of the energy balance between emitting and absorbing powers. The RTE solution routine divides the radiation heat flux into many bands of wavelength. The blackbody radiation function can be discretized in the wavelength domain by using a closed form solution. For each wavelength band, a calculation is performed to balance the emitted and absorbed energy. For each node in the computational fluid dynamics (CFD) grid system, energy emitted is calculated. This emitted energy is then traced along the optical paths of all angles. The energy absorbed in every node (including the wall surface node) is calculated and added to the absorption energy of the initial node.

7.2 Kinetics Calculations

Many schemes have been proposed for the number of equations and reaction rates that may be used to simulate combustion reactions. One of examples is the GCKP program developed at the National Aeronautics and Space Administration (NASA). Another example is the CHEMKIN program, a widely used general-purpose package for solving chemical kinetics problems. CHEMKIN was developed at Sandia National Labs and provides a flexible and powerful tool for incorporating complex chemical kinetics into simulations of fluid dynamics. The latest version includes capabilities for treating multifluid plasma systems that are not in thermal equilibrium. These features allow researchers to describe chemistry systems that are characterized by more than one temperature, in which reactions may depend on temperatures associated with different species; i.e., reactions may be driven by collisions with electrons, ions, or charge-neutral species.

Both thermodynamic and kinetic databases are needed for the kinetics calculations. The GRI-Mech combustion model is a widely used database of detailed chemical kinetic data. The latest version is an optimized mechanism designed to model natural gas combustion, including NO formation and reburn chemistry, containing 325 reactions and 53 species. It provides sound basic kinetics and also furnishes the best combined modeling predictability of basic combustion properties.

7.3 Measurements

Measurements of flame properties can be made by the insertion of probes into a flame, by optical methods applied to emissions from the flames, or by laser probing. The techniques for probing in flames include temperature measurement by thermocouples, velocity measurement by pressure impact probes and particle-tracking photography, and species concentration measurements using suction probes followed by gas analysis.

The optical methods for making measurements in flames include interferometry, schlieren, and photographic methods. All of the probing methods may cause some form of disturbance to the fluid flow, heat transfer, and species concentrations. Optical methods provide information that is integrated over the optical length of the measurement and the spatial resolution is poor. Lasers provide noninvasive probing of flames and high spatial resolution of measurements. The measurement of temperature and species concentration in flames can be carried out by laser spectroscopy. Laser anemometers are widely used to measure velocity components, turbulent characteristics, and particle size in flames. Various advanced combustion sensors have been developed with the potential to enable direct measurement of key combustion parameters for improved combustion control.

8. POLLUTION CONTROL TECHNIQUES

Control of pollutant emissions is a major factor in the design of modern combustion systems. Pollution control techniques can be mainly classified as two types. One involves process modification to prevent or reduce the formation of pollutant emissions. The other is postcombustion treatment to remove or destroy pollutants from combustion before being emitted into the ambient air.

8.1 Process Modification

The techniques to modify combustion for reducing pollutant emissions include fuel changes and combustion modifications. Switching vehicles from gasoline to compressed natural gas, propane, or ethanol can reduce vehicular air pollutant emissions. Adding oxygenated compounds to motor fuels can change the combustion reaction to lower CO emissions significantly. Using low-sulfur fuels can reduce sulfur dioxides emissions.

The principle of combustion modification is widely used in NO_x emission control. Modifying burner configuration or operation conditions can change distributions of temperature, oxygen, and nitrogen to reduce the formation of NO_x. Such techniques include using low excess air, staged combustion, low NO_x burners, water injection, oxyfuel combustion, exhaust gas recirculation (EGR), and reburn. These techniques can be economical and effective. However, in some applications, combustion system modification alone is unable to reduce pollutant emission levels below legislated standards, and postcombustion treatment may be required.

To reduce SO_2 emissions, fluidized bed combustion, in which limestone particles are used as bed materials, is an alternative way to burn coal. Another combustion modification alternative for SO_2 reduction that is in the demonstration stage is the combined-cycle power plant.

8.2 Postcombustion Treatment

Postcombustion treatment involves installation of pollution control devices downstream of the combustor. For example, selective noncatalytic reduction (SNCR) and selective catalytic reduction (SCR) processes are widely used for NO_x reduction. These techniques involve a nitrogen-containing additive, either ammonia, urea, or cyanuric acid, that is injected and mixed with flue gases to effect chemical reduction of NO to N_2, without or with the aid of a catalyst. Other postcombustion treatment techniques utilize wet or dry limestone scrubbers and spray dryers to reduce SO_2 emissions; cyclones, electrostatic precipitators (ESPs), filters, and scrubbers for particle removal; or three-way catalytic converters for the control of NO_x, CO, and hydrocarbon (HC) emissions from automobiles.

To stabilize and ultimately reduce concentrations of the greenhouse gas CO_2, carbon sequestration (carbon capture, separation, and storage or reuse) is a major potential effective tool for reducing carbon emissions from fossil fuel combustion.

9. CONCLUDING REMARKS

This article is intended to guide a reader through the overall process of analyzing a combustion system, especially using the newly added computational tools, and to advocate the good use of combustion

analyses to convert combustion processes for cleaner and more efficient energy production.

SEE ALSO THE FOLLOWING ARTICLES

Acid Deposition and Energy Use • Greenhouse Gas Emissions from Energy Systems, Comparison and Overview • Heat Transfer • Temperature and Its Measurement

Further Reading

Bittker, D. A., and Scullin, V. J. (1972). "General Chemical Kinetics Computer Program for Static and Flow Reactions, with Application to Combustion and Shock-tube Kinetics," NASA TN D-6586. NASA, Washington, D.C.

Chang, S. L., and Rhee, K. T. (1983). Adiabatic flame temperature estimates of lean fuel/air mixtures. *Combust. Sci. Technol.* **35,** 203–206.

Chang, S. L., Golchert, B., and Petrick, M. (2000). Numerical analysis of CFD-coupled radiation heat transfer in a glass furnace, No. 12084. *Proc. 34th Natl. Heat Transfer Conf.*, Pittsburgh, Pennsylvania.

Glassman, I. (1977). "Combustion." Academic Press, San Diego.

Kee, R. J., Rupley, F. M., Meeks, E., and Miller, J. A. (1996). "CHEMKIN-III: A Fortran Chemical Kinetics Package for the Analysis of Gasphase Chemical and Plasma Kinetics," Rep. SAND96-8216. Sandia National Laboratories, Livermore, CA.

Lauder, B. E., and Spalding, D. B. (1974). The numerical computation of turbulent flows. *Comput. Meth. Appl. Mech. Eng.* **3,** 269–289.

Patankar, S. V. (1980). "Numerical Heat Transfer and Fluid Flow." Hemisphere, Washington, D.C.

Pitz, W. J., and Westbrokk, C. K. (1986). Chemical kinetics of the high pressure oxidation of *n*-butane and its relation to engine knock. *Combust. Flame* **63,** 113–133.

Smith, G. P., Golden, D. M., Frenklach, M., Moriarty, N. W., Eiteneer, B., Goldenberg, M., Bowman, C. T., Hanson, R. K., Song, S., Gardiner, W. C., Jr., Lissianski, V. V., and Qin, Z. (2002). GRI-Mech 3.0. Available on the Internet at http://www.me.berkeley.edu/.

Stull, D. R., and Prophet, H. (1971). "JANAF Thermochemical Tables," 2nd ed., NSRDS-NBS 37. U.S. Dept. of Commerce, National Reference Standard Data System-National Bureau of Standards, Washington, D.C.

Turns, S. R. (2000). "An Introduction to Combustion: Concepts and Applications." McGraw Hill, New York.

Zhou, Q., and Yao, S. C. (1992). Group modeling of impacting spray dynamics. *Int. J. Heat Mass Transfer* **35,** 121–129.

Commercial Sector and Energy Use

J. MICHAEL MACDONALD
Oak Ridge National Laboratory
Oak Ridge, Tennessee, United States

1. **Definition and Extent of Commercial Sector**
2. **Magnitude and Significance of Commercial Energy Use**
3. **Measuring Energy Performance**
4. **Performance Rating Systems**
5. **Energy Efficiency Variation and Improvement**
6. **Sectoral Data and Modeling**
7. **Sectoral Views Formed by Models**
8. **Driving Toward Energy Efficiency**
9. **Future Energy Performance and Use**

Glossary

annual total energy The sum of all energy used from all sources in a year.
British thermal unit (Btu) Generically, the amount of energy or heat required to raise the temperature of 1 lb of water (about 0.5 quart or 0.5 liter) 1 degree Fahrenheit (equals about 1055 joules).
commercial sector The portion of buildings in a nation or the world including all buildings that are not residential, industrial, or agricultural.
electricity losses Energy lost in generation, transmission, and distribution of electricity.
empirical Obtained through physical measurement or observation.
energy intensity Annual total energy divided by some normalizing factor, usually gross floor area of a building.
energy performance An empirical value indicating the energy efficiency of one commercial building compared to other, usually similar, commercial buildings.
Energy Star rating system Energy performance rating systems developed by the United States Environmental Protection Agency for specific U.S. commercial building types, which as of June 2003 included offices, hospitals, primary and secondary schools, grocery stores, and hotels and motels, with adjustment capabilities for special spaces such as computer centers and parking garages.
exajoule (EJ) 10^{18} joules.
joule Watt-second, the energy to maintain 1 W for 1 sec.
kBtu 1000 British thermal units.
microdata Detailed survey data arranged in an electronic file for computer analysis.
normalization A method to adjust a quantity, in this case energy use, to account for allowable differences in operational or other factors, such as worker density, hours of operation, and personal computer density.
quads A quadrillion British thermal units; 1 quadrillion is 10 raised to the fifteenth power.
rating scale A numerical scale matched to a range of values of interest, such as normalized annual energy use.
regression An analysis method for statistically determining functional relationships between quantities that are correlated, often using empirical data.
standards Authoritative or legally fixed bases for comparison, valuation, or compliance.
subsectors Subsets of a sector; office buildings, for example, represent a subsector of the commercial sector.

Energy use in commercial buildings is complicated to understand, due to the wide range of building uses and ownership, variations in the size and complexity of energy systems, differences in energy system operation, and other factors. Due to this complexity, as national economies grow toward more specialized services and enterprise management, the share of national energy use held by the commercial sector also tends to grow relative to other sectors. Increased understanding of commercial building energy performance, probably through performance ratings and certifications, is needed to help reduce commercial sector energy growth relative to other sectors in advancing economies.

1. DEFINITION AND EXTENT OF COMMERCIAL SECTOR

The commercial sector is defined typically in terms of including everything else that other sectors do not

 © 2004 Elsevier Inc. All rights reserved.

include. The energy use breakout of national economies related to buildings typically covers residential and commercial buildings; where commercial buildings are typically all buildings that are not residential, industrial, or agricultural. Sometimes this broader grouping of commercial buildings is separated further into institutional and commercial, or governmental and commercial. Thus the commercial sector consists of buildings used by businesses or other organizations to provide workspace or gathering space or offer services. The sector includes service businesses, such as shops and stores, hotels and motels, restaurants, and hospitals, as well as a wide range of facilities that would not be considered commercial in a traditional economic sense, such as public schools, specialized governmental facilities, and religious organizations. Many other types of buildings are also included.

The wide range of building uses is one factor contributing to the complexity of the commercial sector. In the United States, a major national survey of commercial buildings, examining an extensive set of their characteristics and their energy use, is conducted every 4 years. The latest available survey, for 1999, includes in the detailed microdata over 40 different types of building uses that can categorize commercial buildings (Table I). Commercial buildings in the United States, where perhaps the most specialized services and enterprise management exist, are estimated to number 4–5 million, with a total gross floor area of over 6 billion m^2 (over 65 billion ft^2). The approximate breakout of commercial buildings (Table I), showing the range of building uses, the number of buildings, and total floor area estimates, provides an informative starting point for understanding the commercial sector in the United States and elsewhere. Although comparison of these estimates with other, more detailed estimates for subsectors, such as schools, would show some discrepancy in estimates, the scope and size of the commercial sector are well illustrated. For even the least developed countries, the same general range of facilities that comprise the commercial sector will exist, although the aggregate size relative to other sectors typically is smaller.

2. MAGNITUDE AND SIGNIFICANCE OF COMMERCIAL ENERGY USE

The amount of energy consumed in the commercial sector often must be estimated as a fraction of energy use in the combined residential and commercial sectors; national energy use in buildings is often tracked within the major sectors, categorized as industrial, transportation, and "other," with residential and commercial buildings aggregated and accounting for most of the energy use in this "other" sector. Thus some quick checks on world total energy consumption are useful. The units used to sum world energy use are not easily comprehended by most people, so the important knowledge to retain is the relative values. Total world energy consumption in the year 2000 was about 395 quads (1×10^{15} Btu), or about 420 EJ (exajoules). Fuel processing, nonenergy use of fuels, and other losses reduce the total final energy consumption in the major energy-using sectors to about 270 quads (280 EJ).

Because the importance of energy use in the industrial and building sectors can be misunderstood if losses associated with generation and distribution of electricity are not included, comparisons that show both totals are useful. The estimated sectoral breakouts (Table II), without accounting for electricity losses, are 87 quads for industry, 71 quads for transport, and 109 quads for "other," which is primarily residential and commercial buildings. Adding approximate electricity losses (Table II) brings the totals to 122 quads for industry, 73 quads for transport, and 156 quads for "other," for a total of about 350 quads, or 370 EJ.

For the world overall in the year 2000, commercial sector energy use is approximately 30% of the "other" energy use, which amounts to a little over 30 quads (35 EJ) when electricity losses are not included. This energy use represents about 12% of the approximately 270 quads of total final energy consumption for the world. When electricity losses are included, commercial sector energy use is about 45 quads (50 EJ).

Although commercial sector energy use is only 12% of the world total, as economies develop, energy use in this sector tends to rise relative to other sectors and is one of the most difficult to reduce, due to complexities of systems, building ownership, and building uses. The rise in energy use relative to other sectors appears to result from the need for increasingly sophisticated facilities to handle activities in this sector as national economies advance, as well as from a concurrent rise in income within the sector relative to the cost of facilities.

The history of the commercial sector relative to the residential sector in the United States provides an example of this pattern of change. Specialization in services and enterprise management grew significantly in the United States throughout the last

TABLE I

U.S. Commercial Buildings, 1999; Approximate Breakout

Building use	No. of buildings (thousands)	Floor area (billions) (m^2)	Floor area (billions) (ft^2)	Building use	No. of buildings (thousands)	Floor area (billions) (m^2)	Floor area (billions) (ft^2)
Administrative/professional office	503	0.83	8.92	Nursing home/assisted living	25	0.06	0.68
Auto dealership/showroom	56	0.04	0.41	Other education	40	0.04	0.40
Auto service/auto repair	210	0.13	1.42	Other food sales or service	129	0.04	0.39
Bank/financial	128	0.09	0.97	Other health care	22	0.03	0.31
Clinic/outpatient health	37	0.03	0.34	Other lodging	4	0.02	0.17
College/university	25	0.11	1.18	Other office	25	0.03	0.37
Courthouse/probation office	3	0.02	0.23	Other public assembly	21	0.06	0.63
Doctor/dentist office	111	0.10	1.03	Other public order and safety	11	0.02	0.20
Dormitory/fraternity/sorority	35	0.17	0.80	Other retail	88	0.04	0.42
Dry cleaner/laundromat	58	0.03	0.30	Other service	118	0.08	0.82
Elementary/middle/high school	230	0.61	6.52	Post office/postal center	25	0.03	0.33
Enclosed mall	3	0.15	1.64	Preschool/daycare	32	0.05	0.51
Entertainment (theater/sports arena/nightclub)	35	0.07	0.78	Recreation (gymnasium/bowling alley/health club)	88	0.11	1.22
Fire station/police station	52	0.04	0.48	Refrigerated warehouse	14	0.08	0.87
Government office	33	0.11	1.21	Religious worship	307	0.32	3.40
Grocery store/food market	49	0.06	0.62	Repair shop	68	0.05	0.51
Hospital/inpatient health	7	0.15	1.65	Restaurant (bar/fast food/cafeteria)	345	0.17	1.84
Hotel	28	0.17	1.81	Social meeting center/convention center	132	0.11	1.22
Jail/reformatory/penitentiary	6	0.02	0.26	Store	389	0.36	3.92
Laboratory	27	0.04	0.42	Strip shopping center	131	0.37	3.94
Library/museum	28	0.05	0.52	Vacant	253	0.18	1.90
Motel/inn/resort	61	0.10	1.04	Other	75	0.07	0.80
Nonrefrigerated warehouse	589	0.89	9.56	**Total**	4654	6.22	67.00

TABLE II

World Sectoral Energy Consumption in 2000[a]

Sector	Without electric losses		With electric losses	
	Quads	EJ	Quads	EJ
Industry	87	92	122	129
Transport	71	75	73	77
Other	109	115	156	165
Total	267	282	351	370

[a] Values are approximate and can also vary depending on the detail in accounting for different types of losses or fuel processing. Sectors may not add to totals due to rounding.

half of the 20th century. Allowing some time to pass after World War II, so that wartime effects and rationing-induced behaviors can be discounted, energy use in the commercial sector was about 50% of residential energy use in the year 1960. This ratio grew to about two-thirds by 1980, and in the year 2000 was over 83% (Fig. 1). Residential sector energy use in the United States was 20% of total national energy use in both 1960 and 2000, whereas commercial sector use increased from 10 to 17% of the total. Thus, commercial sector energy use was about 30% of sectoral energy use in the building or "other" category in the United States in 1960, but by 2000 had grown to 45% of the building sector total.

Variations among countries and location are dramatic. Although the commercial sector in the United States accounts for about 17% of all energy use, in China it accounts for about 5%. Commercial buildings in rural areas typically are less complicated

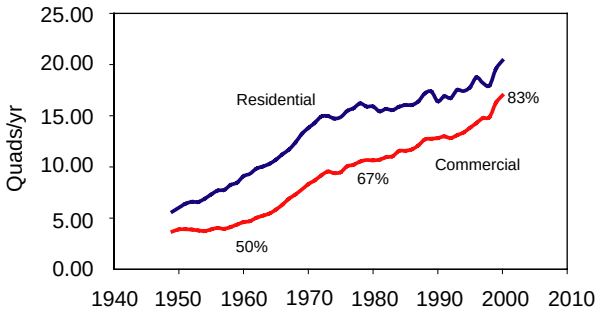


FIGURE 1 Comparison of commercial sector and residential energy use in the United States from 1960 to 2000; the ratio of commercial to residential energy use grew from 50 to 83%. Electricity losses are included.

and use less energy, as compared to those in metropolitan areas; over 80% of commercial sector energy use is in metropolitan areas.

3. MEASURING ENERGY PERFORMANCE

Interest in rating the real-life energy performance of buildings has increased in recent years, and the real-life efficiency performance rating of buildings is important for any sustainable energy future. The ability to compare the energy performance of one commercial building with that of another is important for determination of national and international energy efficiency because comparison allows meaningful measurements of potential relative improvements. This ability also may allow different classes of buildings to be analyzed together (e.g., offices and hospitals).

The European Union has been examining requirements for improving the energy performance of residential and commercial buildings, because a large potential improvement in energy performance has been determined to exist. Among the requirements examined is the establishment of a general framework for a common methodology for calculating the integrated energy performance of buildings.

The United States Environmental Protection Agency (EPA) has established an empirical energy performance rating system for some commercial building types, the Energy Star rating system, whereby a normalized energy performance rating scale is developed. The energy use of a specific building is normalized based on the factors in the method, and the normalized energy is compared to the performance rating scale. Buildings scoring in the top 25% on the scale have energy performance level that makes them eligible for consideration of award of an Energy Star label.

Commercial building energy performance, or energy efficiency, is often measured to a certain degree by building energy experts, and even many nonexperts, without using any real standards. To judge how well a specific building is doing, however, energy performance measurement should involve a comparison of building energy use to some type of standard, which in the past has typically been the energy use of other, similar buildings. The challenge over the years has been to determine a true standard for comparison and to determine what a "similar" building is. Because the historical methods of comparison had known limitations, building energy experts developed their own sense of what constitutes an energy-efficient building. This expert sense is based on experience with similar buildings, the types of activities within specific buildings, and any history of achieving reductions in energy use in comparable buildings. This expert knowledge has gaps and is not easily transferable, because it is usually based on several years of experience concerning expected patterns of energy use for different buildings and impacts of schedules, uses, geographic location, and system configurations. This expert knowledge is used to "adjust" the measure of the performance of a commercial building to provide a more informed measure of performance. However, this knowledge is ad hoc, with multiple practitioners probably arriving at differing assessments of the same building. In the end, the result is essentially a subjective expert opinion, albeit possibly a very good one, but also possibly not.

Five generic classes of building energy data analysis methods have been identified as useful in measuring the energy performance of commercial buildings:

1. Annual total energy and energy intensity comparisons.
2. Linear regression and end-use component models.
3. Multiple regression models.
4. Building simulation programs.
5. Dynamic thermal performance models.

All of these analytical approaches can be used to develop building energy performance measurement methods, but the most effective current approach in use today, based on results achieved, involves the third approach, multiple regression models. When calculating commercial building energy performance using multiple regression models, the effects of many factors can be modeled, potentially factoring out

influences such as the number of people in a building or occupant density. The Energy Star rating system develops its performance rating scales using multiple regression models.

The limitations of the other methods include their inability to cover wide ranges of buildings without an inordinate amount of data. Some of the other methods require large volumes of data to develop empirical results. In the following discussion of performance rating systems, both simple annual total energy intensity comparisons (Method 1 above) and multiple regression method information will be addressed; the first is useful both as an example and as a well-understood quantity, whereas multiple regression analysis illustrates the current state-of-the-art methodology.

4. PERFORMANCE RATING SYSTEMS

Many people confuse building simulation and energy audits or energy assessments with energy performance ratings. Energy performance ratings are less detailed and provide much less information regarding potential causes of specific energy performance. Instead, what is provided is a true indication of overall energy performance relative to similar buildings. Highly technical assessments, including calibrated simulations, are a tremendous tool for diagnosing root causes of specific building energy performance. But these approaches typically provide only very limited information about performance relative to other buildings or relative to any ranking scale based on performance of similar buildings, and their complexity makes them impractical for extensive use in rating performance.

Performance rating can be done many ways. The EPA Energy Star rating system for buildings uses a percentile rating scale of 1 to 100 for a particular building type, with a rating of 75 or greater required to qualify for an Energy Star label. The energy performance rating of 1 to 100 can be obtained simply by using the rating tool. The rating scale is developed from a regression analysis of energy use versus key characteristics of the class of buildings against which energy use is to be normalized. A simple and straightforward way of quantifying and comparing building energy performance is accomplished by using the annual total energy and energy intensity data. Annual total energy is the sum of the energy content of all fuel used by the building in one year. Energy intensity (also called energy use intensity or index, EUI) is the total energy used divided by the total floor area. It would also be possible to examine annual energy or energy intensities for individual fuels.

The strength of the total energy and energy intensity comparisons is their ease of use and widespread familiarity. However, knowledge is lacking regarding causes of variation that have been observed and the relative impacts of factors such as schedules, functional uses, and density of use on the energy performance. This general approach to rating commercial building energy performance is of interest for quick comparison of one building's energy use from one year to another or quick comparisons of many buildings, but information to adjust for at least some of the wide variation typically observed across a data set with many buildings is lacking.

In cases in which a performance rating system is desired for a specific type of building in a relatively homogeneous climate region, rating scales based on total energy or energy intensity by floor area have some practicality. In the Czech Republic, "labels" of actual, measured energy performance (energy intensity) have been studied, tested, and are now required for apartment buildings. The European Union is likely to require energy performance certificates for buildings by the year 2010. These certificates may have to be renewed every 5 years. With reasonably small ranges of climatic differences, certificates for specific types of buildings based on simple energy intensity values can be a moderately reasonable approach to an energy performance rating system. However, even with a common building type and common climate, there are other variations in key factors that should typically be considered. The basic annual energy intensity accounts for floor area, which has been found to be the most important factor to use in normalizing energy use. If multiple climates must be considered, adjustments for climatic variation, such as normalization for heating degree-days and cooling degree-days, should be included in an energy performance rating system.

Beyond floor area and climate, there will typically be variations in other important factors based on building use, e.g., hospitals and offices will have different normalization factors. Decisions on such factors can be, and at times are, arbitrary. Also, differing policy perspectives can strongly influence consideration of what parameters should be evaluated for normalization of energy performance. Rating systems of the more sophisticated, multifactor type typically consist of parameters for normalization of energy use, a normalization equation or

calculation method, and a normalized distribution of energy use. The normalized distribution is typically matched to some scale, often a percentile scale of 1 to 100, to allow a simplified rating result to be obtained. After the energy use of a building is normalized for the factors in the method, the normalized result is compared to the rating scale to determine a rating.

A generic rating system developed for the entire commercial sector in the United States included normalization factors (Table III) that adjusted for floor area, climate, amount of building cooled and heated, worker density, personal computer density, extent of food service and education/training facilities, hours open, and average adjustments for specific types of building space uses. These factors were found to account for over 70% of the variation in energy use in the entire combined U.S. commercial sector, with its wide range of building types. The percentile scale for this particular rating system (Fig. 2) was based on the building energy use index that included electricity losses.

TABLE III

Normalization Factors in a Generic Energy Performance Rating System for All U.S. Commercial Buildings

Floor area (ft^2)
Annual cooling degree days (base 65°F)
Annual heating degree days (base 65°F)
Floor area cooled is less than 25% (yes or no)
Floor area cooled is 50 to 75% (yes or no)
Floor area cooled is 75 to 100% (yes or no)
Floor area heated is 75 to 100% (yes or no)
Personal computers per 1000 ft^2
Workers per 1000 ft^2
Food seats per 1000 ft^2
Educational seats per 1000 ft^2
Hours per week open
Fraction of area that is laboratory
Fraction of area that is nonrefrigerated warehouse
Fraction of area that is food sales (grocery)
Fraction of area that is outpatient health care
Fraction of area that is refrigerated ware house
Fraction of area that is worship space
Fraction of area that is public assembly
Fraction of area that is educational space
Fraction of area that is restaurant
Fraction of area that is inpatient health care
Fraction of area that is strip shopping mall
Fraction of area that is larger shopping mall
Fraction of area that is service

Rating system analyses for specific types of buildings in the United States have found that about 90% or more of the variation in energy use in a specific building type can be normalized based on reasonable factors. As an example, for offices, in addition to floor area, climate, fraction of building heated or cooled, worker density, personal computer density, and hours of operation shown in the generic model (Table III), the number of floors was also a factor found to be important for normalization. Additional examples of the extensive information on the Energy Star rating system for specific building types can be found on the Energy Star Web site (www.energystar.gov).

Another factor that might be considered important for energy performance rating systems is the unit price of energy in a particular location. The energy unit price is significant; analyses comparing all-electric buildings with those that are not indicate an important statistical difference between the energy use distributions of these two categories, suggesting that they should not be treated as equivalent unless some adjustment is made for the difference. Study has shown that this disparity between all-electric buildings and other buildings can be accounted for fairly well by including electricity losses in total energy use as a surrogate for the typical energy unit price differential and other factors related to remote efficiency losses. But the average unit price of energy also appears to adjust fairly well for these differences between all-electric and other buildings, as well as introducing an adjustment for the local economic incentive to be efficient.

Building energy performance rating systems are important tools that offer reasonably quick building energy performance assessment without rigorous

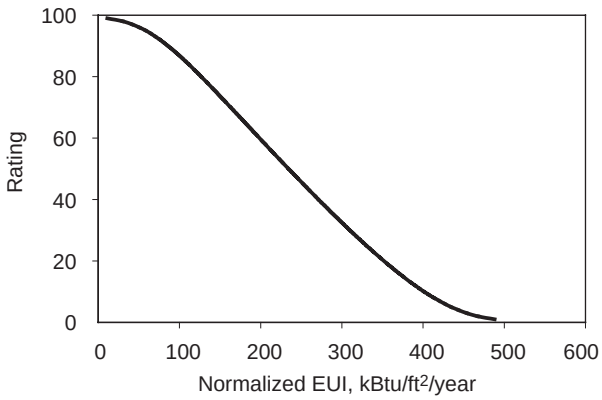


FIGURE 2 Rating scale for the entire U.S. commercial sector. Normalized annual energy use intensity (EUI) matched to percentile scale. Electricity losses are included.

evaluation. In addition, energy performance ratings provide an empirical statement of energy performance not available with other methods, even those that are more rigorous and complicated. Because documenting building energy performance has been determined to be important for many nations, understanding and improving systems for performing such ratings appear to be important for continued progress in commercial building energy efficiency.

5. ENERGY EFFICIENCY VARIATION AND IMPROVEMENT

Energy performance ratings tell what the energy performance of a building is, but if the energy performance of a building is to be improved, the causes of lower than desired performance must be understood, and methods of achieving improved performance must be determined. Causes of variation in energy performance among commercial buildings are understood to a degree, but much remains to be learned.

Estimates of potential improvements in energy efficiency of commercial buildings, i.e., the potential to reduce energy use, have, over many years, indicated that a lot has been and still could be accomplished. Issues of economic incentive and resource allocation influence the estimated savings values, but reductions of 20–40%, on average for the entire sector, appear reasonably achievable if a great enough need exists. In the European Union, a potential savings of 40% has been presented as possible. A savings of 20% of the worldwide annual commercial sector energy use of 30 quads (35 EJ) is 6 quads/yr (7 EJ/yr), which is about 2% of total world energy use. Unfortunately, energy costs are often not large enough, relative to other costs of running a business or organization, to receive major attention, so efficiency improvements have a lower priority.

Research has indicated many reasons why energy efficiency varies so much in commercial buildings. The causes of variation in efficiency can be categorized as variations in: efficiency of operation, efficiency of systems, and efficiency of equipment. Of these three, about half of the potential improvement for the sector would result from operational improvements, with the remainder from equipment and system improvements.

Many studies have shown the importance of operational improvements, with typical savings of 10–20% possible in a wide range of buildings. Other studies show significant savings from equipment and systems improvements. The United States Federal Energy Management Program is responsible for achieving reductions in annual energy intensity for most U.S. government buildings. Significant reductions have been achieved through attention to all three areas of efficiency improvement: operation, equipment, and systems. From 1975 to 1985, a reduction in annual energy intensity of about 20% was achieved. After 1985, additional goals were required by Executive Order and other means, and the annual energy intensity of U.S. government buildings is on track to reduce annual energy intensity an additional 30% by the year 2005 relative to the year 1985 (Fig. 3).

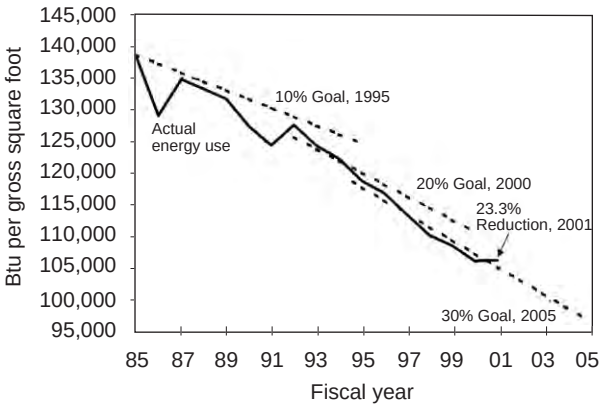


FIGURE 3 Energy efficiency progress for government buildings in the United States. Electricity losses are not included in the annual energy intensity.

The potential for improvement in energy efficiency in the commercial sector is large, if the desire to improve is there, as witnessed by the progress in U.S. government buildings. European estimates also show a large savings potential. If the sources of energy inefficiency in commercial buildings can be reduced through attention to the three major areas of efficiency improvement, about a 2% reduction in world energy use appears achievable.

6. SECTORAL DATA AND MODELING

Extensive data are collected on energy use and sectoral characteristics throughout the world. Sectoral data covering energy use of major economic energy sectors are available for many countries, and world data are available from the International

Energy Agency and from some major nations. The commercial sector, as defined here, is often called by other names, such as "general" or "tertiary," indicating the "everything else" nature of the sector, including all buildings other than residential, industrial, or agricultural.

As an example of data for a nation with one of the most specialized commercial sectors in the world, the United States collects extensive energy data for this sector, including consumption by fuel type, prices by fuel type, and expenditures by fuel type for the overall sector. The special sampling survey of 5000 to 6000 buildings that is conducted every 4 years provides detail on the annual consumption of each fuel and extensive characteristics data, allowing additional extensive analyses to be performed and supporting national modeling of energy use in this sector.

Energy data provide a historical record and are used to forecast trends into the future. Several organizations model world energy use, although often not with the commercial sector separated. Many individual nations also track historical energy use and forecast future energy use. The National Energy Modeling System (NEMS), used in the United States for energy use forecasts and other analyses, provides an example of an extensive sectoral and national modeling system. The NEMS Commercial Sector Demand Module is a simulation tool based on economic and engineering relationships; it models commercial sector energy demands, with breakout detail at the geographic level of nine census divisions, using 11 distinct categories of commercial buildings.

Projections of future energy use involve selections of equipment for the major fuels of electricity, natural gas, and distillate fuel, and for the major services of space heating, space cooling, water heating, ventilation, cooking, refrigeration, and lighting. The equipment choices are made based on an algorithm that uses life-cycle cost minimization constrained by factors related to commercial sector consumer behavior and time preference premiums. The algorithm also models demand for the minor fuels of residual oil, liquefied petroleum gas, coal, motor gasoline, and kerosene. The use of renewable fuel sources (wood, municipal solid waste, and solar energy) is also modeled. Decisions regarding the use of distributed generation and cogeneration technologies are performed using a separate cash-flow algorithm.

The NEMS Commercial Module generates midterm (20- to 30-year) forecasts of commercial sector energy demand. The model facilitates policy analysis of energy markets, technological development, environmental issues, and regulatory development as they interrelate with commercial sector energy demand. Input to this model is quite extensive, and in addition to the sectoral energy data includes building lifetime estimates, economic information such as demand elasticities for building services and market forecasts for certain types of equipment, distributed electricity generation/cogeneration system data, energy equipment market data, historical energy use data, and short-term energy use projections from another model. The primary output of the modeling process is a forecast of commercial sector energy consumption by fuel type, end use, building type, census division, and year. The module also provides forecasts of the following parameters for each of the forecast years:

- Construction of new commercial floor space by building type and census division.
- Surviving commercial floor space by building type, year of construction, and census division.
- Equipment market shares by technology, end use, fuel, building type, and census division.
- Distributed generation and cogeneration of electricity and fuels used.
- Consumption of fuels to provide district services.
- Nonbuilding consumption of fuels in the commercial sector.
- Average efficiency of equipment mix by end use and fuel type.

The NEMS Commercial Module interacts with and requires input from other modules in NEMS. This relationship of the Commercial Module to other components of NEMS is depicted schematically in Fig. 4. Not shown are the other sectoral modules and the central controlling module in NEMS that integrates and reconciles national level results for all sectors.

In contrast to such extensive sectoral models that are used to provide forecasts of energy use, the energy performance rating systems presented previously provide another type of model for the commercial sector; they allow additional understanding of current energy use and the influence of certain operating parameters of buildings. The type of information and the understanding of the commercial sector generated by different modeling approaches differ in perspective and ability to understand potential for improvements in energy efficiency. Effects of these differing perspectives are presented next.

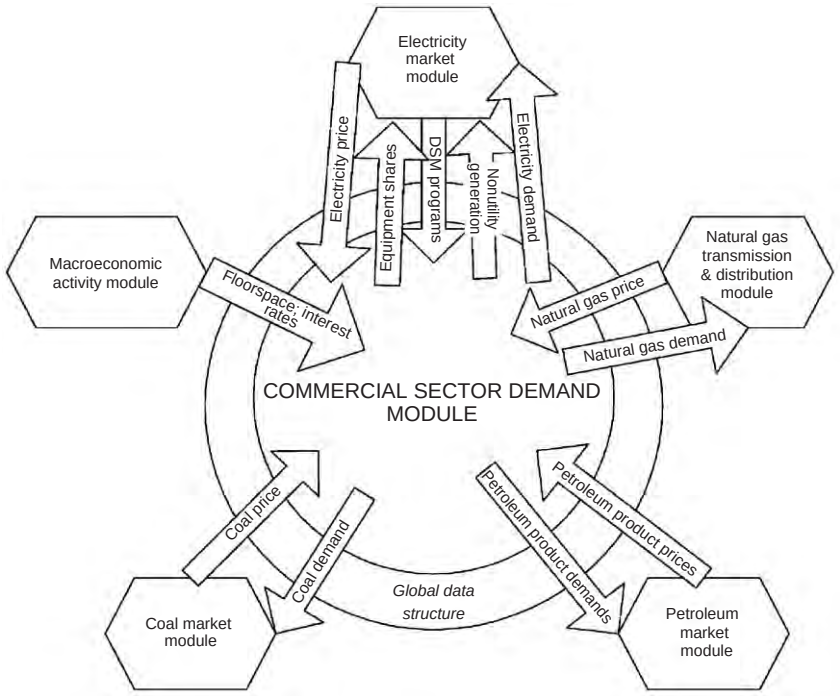


FIGURE 4 Relationship of the commercial sector and other National Energy Modeling System modules. DSM, demand-side management.

7. SECTORAL VIEWS FORMED BY MODELS

Energy models such as NEMS can be called economic-engineering models. Such models use engineering data and analysis results to feed into and partially interact with an economic model of commercial sector changes and demand for energy. Because changes in the efficiency of buildings and the energy use patterns of buildings tend to take many years to evolve, such economic-engineering models often do a good job of representing the energy use of a large group of buildings, including the entire commercial sector in a country. These models can also forecast the end uses of energy in buildings (Table IV) and are usually good at doing so.

Although NEMS forecasts about a 50% increase in U.S. commercial sector energy use between the years 2001 and 2025, the energy intensity of these buildings remains almost constant over this forecast period. This information indicates that growth in commercial sector energy use is attributed almost exclusively to growth in floor area, and that any efficiency improvements are modeled as offset by growth in end uses of energy not affected by the efficiency improvements. Clearly, with little change in annual energy intensity over a 25-year period, a tendency to limit certain types of change can be seen in this modeling approach. In addition, although the model provides a breakout of energy according to end uses such as heating, cooling, lighting, and seven other uses (Table IV), data on the impacts of the density of workers or occupants, schedule of operations, and density of personal computers in buildings are not modeled and are not known. To forecast total energy use, this type of normalization of energy is not required, because it represents a different way of looking at the sector, and normalized energy is not the desired output.

Economic-engineering models are also capable of forecasting impacts of new energy technologies and more efficient operations on energy use, but new energy technologies and impacts of those technologies on new buildings are more capably modeled in NEMS than are improvements in operations, which must typically be treated as changes in annual energy intensity. Unfortunately, this characteristic means that impacts of improvements in energy system operations are not understood and cannot be estimated well with this modeling approach.

Detailed engineering simulation models such as DOE-2, Energy Plus, and BLAST, which have the capabilities to model energy use in commercial buildings, also have difficulty modeling improvements

TABLE IV

NEMS Commercial Sector Energy End-Use Baseline and Forecast[a]

	Year						
Use	2000	2001	2005	2010	2015	2020	2025
Space heating	2.13	1.95	2.22	2.26	2.3	2.37	2.42
Space cooling	1.35	1.39	1.35	1.37	1.39	1.41	1.43
Water heating	1.15	1.12	1.21	1.26	1.3	1.35	1.39
Ventilation	0.56	0.55	0.55	0.56	0.56	0.56	0.57
Cooking	0.38	0.37	0.39	0.41	0.42	0.44	0.45
Lighting	3.34	3.31	3.54	3.73	3.81	3.87	3.89
Refrigeration	0.69	0.69	0.71	0.73	0.75	0.77	0.78
Office equipment (PC)[b]	0.5	0.52	0.6	0.74	0.85	0.95	1.05
Office equipment (non-PC)	0.98	0.99	1.12	1.44	1.8	2.21	2.69
Other uses	6.13	6.56	6.89	7.65	8.54	9.6	10.6
Total	17.2	17.44	18.59	20.15	21.72	23.52	25.33

[a] In quads/year; electricity losses are included.
[b] PC, personal computer.

in energy system operations, because the simulation routines are all set up to model systems and components that work correctly. Simulating improvements in operations often requires knowing the answer first, and then tricking the simulation program into calculating the correct results. Again, the limitations of the models influence decisions about appropriate energy efficiency improvements for buildings. Limitations in ability to model impacts of operational improvements on energy efficiency lead to a view of the commercial sector that essentially ignores the potential of such improvements. A problem with this situation is that policy officials typically also do not receive information on the potential of operational improvements and lack an understanding of the importance of improving operations in commercial buildings. Because improvements in operations represent about half of the potential energy savings that could be achieved in the commercial sector, acceptable means of modeling impacts and potential efficiency benefits would be helpful. Until acceptable means of modeling are developed, the benefits of operational improvements must continue to be determined empirically.

Improved understanding of commercial sector energy use, and the potential of operational improvements for saving energy in commercial buildings, would result from the ability to model the effects of operational improvements on forecasts of energy use in the commercial sector. Improved understanding of the potential for energy efficiency improvements appears possible through modeling of normalized energy use for the sector, whereby adjustments for a few key factors known to cause variation in energy use in commercial buildings, together with performance rating values, could allow analysis and pursuit of scenarios of improvement toward higher efficiency ratings. Advances in modeling such as these could be important for the future, if the percentage increase in commercial sector energy use in advancing economies becomes a challenge for world energy efficiency and climate impacts.

8. DRIVING TOWARD ENERGY EFFICIENCY

As the need for energy efficiency becomes more pronounced, the drive toward efficiency in the commercial sector will be impeded by its complicated mix of building sizes and uses, the complicated systems often used in commercial buildings, and the relative lack of understanding of operations factors impacting energy use and how to achieve efficiency.

In the United States, commercial energy use has increased from 10 to 17% or more of national energy use between the years 1960 and 2000. A significant reason for this increase is the low cost of energy relative to the other costs of conducting business, but the difficulty in understanding energy systems and energy use in commercial buildings is also an important contributing factor. Policy officials often have difficulty understanding discussions of the needs for improvements in commercial buildings.

As economies advance and commercial sector energy use begins to grow relative to other sectors, an improved understanding of methods of measuring commercial energy performance, and the means of achieving efficiency improvements in this sector, will be important in any drive toward efficiency.

One warning sign of the need to increase understanding of energy performance is an increase in the use of air conditioning in commercial buildings. When air conditioning use increases, energy system complexity and indoor space quality issues also increase significantly. If air conditioning use is increasing, any proposal for increased efficiency that relies heavily on thermal insulation should be treated warily, because insulation optimization becomes more difficult, and other system complexities tend to become much more important.

A warning should be given overall for energy standards for buildings, as they currently exist around the world, because, despite the existence and use of these standards for many years, the effect of standards has been only moderate in most cases. The shortcoming of existing standards is that they rely too heavily on simulation of expected performance, without conducting true empirical studies to verify the effectiveness of what the standards achieve. One major reason such empirical studies have been conducted in only limited and mostly ineffective fashion is that an empirical method of measuring energy performance of commercial buildings, although still adjusting for legitimate building use differences, has only recently been established in concept, only for certain building types, and not with the stated intent of being a performance standard (Energy Star label for buildings). Interestingly, however, if such energy performance standards existed, energy standards currently in use, with their typically complicated requirements, would not necessarily be needed any longer as standards.

Use of energy performance certificates may be necessary to overcome the difficulty users, occupants, code and policy officials, and owners have in understanding commercial energy use and performance. Without certification of energy performance, the complexity of systems and uses makes understanding energy performance by anyone other than an expert, and even by some experts, difficult. However, without increased understanding of the most appropriate means to normalize energy use for legitimate differences in building function and use, energy performance certificates may offer an unsatisfactory solution, due to inequities that will be obvious to many, if reasonable normalizations are not applied.

Increased energy efficiency in the commercial sector is an important piece of the national efficiency strategy in advanced economies, in which the priority for efficiency must be increased. Complexities in commercial buildings have made progress in energy efficiency for this sector less than desirable in notable cases. Fortunately, methods and knowledge needed to increase success in this sector have begun to be developed, and better solutions can be offered in the near future to help increase energy efficiency in the commercial sectors of countries where the need is pressing.

9. FUTURE ENERGY PERFORMANCE AND USE

Current forecasts call for solid growth in world energy use over the next 20 years, potentially increasing 60% above current use. With the forces in place to keep energy use patterns the same, a safe, conservative assumption would be that the commercial sector will contribute about 12% to final total energy consumption in the year 2020. If world energy use grows to 600 quads, or 630 EJ, by the year 2020, and total final consumption in the energy-using sectors is about 400 quads, or 420 EJ, final consumption in the commercial sector, at 12%, would be about 50 quads/yr (or about 50 EJ), without electricity losses included. This energy use would be two-thirds more than the energy use in the year 2000.

Without significant changes in energy performance of commercial buildings, this scenario of 50 quads of commercial energy use in the year 2020 is likely to occur, absent major world upheaval. If important progress on improving the energy performance of commercial buildings in the advanced economies can be made, potentially a reduction of 2–4 quads or more in commercial sector worldwide use in 2020 (a 4–8% reduction) could be achieved. Such an achievement would require that energy performance certifications become the norm, that operational standards increase significantly, and that 25–40% of buildings in the advanced economies see significant improvements in their performance.

Methods for certifying the energy performance of commercial buildings have been developed to the conceptual and practical applications stages. However, these methods are in their infancy. Energy policy and research attention to the commercial sector have been lacking relative to the growth

observed. Without increased attention to improving commercial sector energy efficiency, energy use growth relative to other sectors will continue to make this sector a challenge when decreased energy use and emissions are sought.

SEE ALSO THE FOLLOWING ARTICLES

District Heating and Cooling • Industrial Ecology • Industrial Energy Efficiency • Industrial Energy Use, Status and Trends • International Comparisons of Energy End Use: Benefits and Risks • Obstacles to Energy Efficiency • Thermal Comfort

Further Reading

Energy Information Administration. (2001). Model Documentation Report: Commercial Sector Demand Module of the National Energy Modeling System. DOE/EIA-M066(2002). Department of Energy, Washington, D.C.

Energy Information Administration. (2002). International Energy Outlook 2002. DOE/EIA-0484(2002). Department of Energy, Washington, D.C.

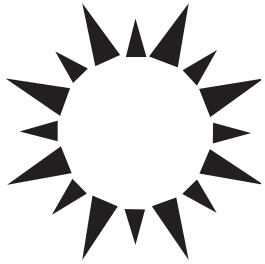
Energy Information Administration. (2003). Annual Energy Outlook 2003. DOE/EIA-0383(2003). Department of Energy, Washington, D.C.

Energy Star Web site. (2003). Available at http://www.energystar.gov/.

European Commission. (2001). Proposal for a directive of the European Parliament and of the Council on the energy performance of buildings. COM(2001) 226 final, 2001/0098 (COD). EC, Brussels.

International Energy Agency. (2000). Key World Energy Statistics Pamphlet. IEA, Paris.

Complex Systems and Energy



MARIO GIAMPIETRO
Istituto Nazionale di Ricerca per gli Alimenti e la Nutrizione
Rome, Italy

KOZO MAYUMI
University of Tokushima
Tokushima, Japan

1. **Introduction**
2. **The Energetics of Human Labor: An Epistemological Analysis of the Failure of Conventional Energy Analysis**
3. **The Problematics Related to Formal Definitions of Energy, Work, and Power in Physics and Their Applicability to Energy Analysis**
4. **The Metaphor of Prigogine Scheme to Characterize Autocatalytic Loops of Energy Forms in Hierarchically Organized Adaptive Systems**
5. **Conclusion**

Glossary

complex system A system that allows one to discern many subsystems, depending entirely on how one chooses to interact with the system.

dissipative systems All natural systems of interest for sustainability (e.g., complex biogeochemical cycles on this planet, ecological systems, human systems when analyzed at different levels of organization and scales beyond the molecular one); self-organizing open systems, away from thermodynamic equilibrium.

epistemological complexity Complexity that is in play every time the interests of the observer (the goal of the mapping) are affecting what the observer sees (the formalization of a scientific problem and the resulting model).

hierarchical systems Systems that are analyzable into successive sets of subsystems; when alternative methods of description exist for the same system.

hierarchy theory A theory of the observer's role in any formal study of complex systems.

scale The relation between the perception of a given entity and its representation.

The purpose of this article is to examine key epistemological issues that must be addressed by those willing to apply energy analysis to complex systems and particularly to the sustainability of adaptive systems organized in nested hierarchies. After a general introduction, sections 2 and 3 provide a critical appraisal of conventional energy analysis. Such a criticism is based on concepts developed in the field of hierarchy theory. Section 4 suggests an alternative view of the metaphorical message given by Prigogine's scheme in nonequilibrium thermodynamics—a message that can be used to develop innovative and alternative approaches for energy analysis. Section 5 presents the conclusions.

1. INTRODUCTION

The revolution entailed by complex systems theory is finally forcing many scientists to acknowledge that it is impossible to have a unique substantive representation of the reality. Different observers with various "scale-dependent" measuring schemes and goals (e.g., humans using either a microscope, the naked eye, or a telescope for different purposes) will perceive different aspects of a particular ontological entity. These different views do reflect the special relation established between the observer and the components of the entity observed. This predicament gets worse when dealing with hierarchically organized adaptive systems. These systems are, in fact, organized on different levels (e.g., cells, organs, individual humans, households, communities, countries). This implies that their identity is associated with the ability to express various patterns simultaneously on different scales. This implies that when doing energy analysis on these systems, wholes, parts, and their contexts require the parallel use of nonequivalent descriptive domains. That is, the

 © 2004 Elsevier Inc. All rights reserved.

definitions of parts, wholes, and contexts referring to different levels are not reducible to each other. To make things more difficult for modelers, dissipative systems are becoming in time, meaning that their identities change in time. Their hierarchical nature entails that the pace of evolution of the identities of the parts will be different from that of the identity of the whole.

The disciplinary field of energy analysis seems to have ignored the epistemological challenge posed to the analysts by the fact that energy assessments must be a result of different perceptions and representations of the reality that cannot be reduced to each other. This problem is particularly serious considering that hierarchically organized adaptive systems, by definition, are operating across multiple scales. Not only do the energetic descriptions have to be based on the use of nonequivalent descriptive domains, but they also have to be updated in time. Thus, scientists dealing with the sustainability of complex dissipative systems adapting in time must be able to (1) individuate relevant causal relations and relevant scales in their modeling efforts, (2) use different selections of variables for representing changes in relevant attributes, and (3) continuously check the validity of these selections.

The main epistemological implication of studying systems organized across hierarchical levels and scales is that there is a virtual infinite universe of possible ways of perceiving and representing them. For example, there is virtually an infinite number of possible combinations of views of a human that can be generated when looking at him or her using microscopes, ultrasound scanners, X-ray machines, blood tests, conventional cameras, radars, and telescopes simultaneously. Any formal representation—the information space used to make models in terms of variables and formal systems of inference—must be finite and closed. This represents a huge challenge for scientists.

The main message of the new scientific paradigm entailed by complexity can be expressed using two crucial points. When dealing with the sustainability of complex adaptive systems, first, scientists must accept dealing with an unavoidable nonreducibility of models defined on different scales and reflecting different selections of relevant attributes. This implies acknowledging the impossibility of reaching an uncontested agreement on a substantive and unique description of the reality that can be assumed to be the "right" one. Second, scientists must accept dealing with incommensurability among the priorities used by nonequivalent observers when deciding how to perceive and represent nested hierarchical systems operating across scales. Nonequivalent observers with different goals will provide logically independent definitions of the relevant attributes to be considered in the model. This implies the impossibility of having a substantive and agreed-on definition of what should be considered the "most useful" narrative to be adopted when constructing the model.

This new paradigm associated with complexity, which is rocking the reductionist building, is the offspring of a big epistemological revolution started during the first half of 19th century by classical thermodynamics (e.g., by the works of Carnot and Clausius) and continued during the second half of the 20th century by nonequilibrium thermodynamics (e.g., by the works of Schroedinger and Prigogine). Both revolutions used the concept of "entropy" as a banner. The equilibrium thermodynamics represented a first bifurcation from mechanistic epistemology by introducing new concepts such as irreversibility and symmetry breaking when describing real-world processes. Then, the introduction of the new nonequilibrium paradigm implied a final departure from reductionist epistemology in that it implies the uncomfortable acknowledgment that scientists can work only with system-dependent and context-dependent definitions of entities. Put differently, food is a "high-quality" energy input for humans but not for cars. Saying that 1 kg of rice has an "energy content" of a given amount of megajoules can be misleading because this "energy input" is not an available energy input for driving a car that is out of gasoline. In the same way, the definition of what should be considered "useful energy" depends on the goals of the system that is expected to operate within a given associative context. For example, fish are expected to operate under the water to get their food, whereas birds cannot fly to catch prey below the ground.

Classical thermodynamics first, and nonequilibrium thermodynamics later on, gave a fatal blow to the mechanist epistemology of Newtonian times, introducing brand-new concepts such as disorder, information, entropy, and negentropy. However, we cannot replace the hole left by the collapse of Newtonian mechanistic epistemology just by using these concepts, as if they were "substantive" concepts that are definable in a strictly physical sense as well as context independent. For example, the concept of "negative entropy" is not a substantive concept. Rather, it is a "construction" (an artifact) associated with the given identity of a dissipative system that is perceived and represented as operating at a given

point in space and time. Therefore, the concept of negative entropy reflects the perception and representation of the "quality" of both energy inputs and energy transformations that are associated with a particular pattern of dissipation (or typology of metabolism). But this quality assessment is valid only for the specific dissipative system considered that is assumed to operate within a given associative context or an admissible environment. These can seem to be trivial observations, but they are directly linked to a crucial statement we want to make in this article. It is not possible to characterize, using a "substantive formalism," qualitative aspects of energy forms, and this is applicable to all conceivable dissipative systems operating in the reality. That is, it is impossible to define a "quality index" that is valid in relation to all of the conceivable scales and when considering all of the conceivable attributes that could be relevant for energy analysis. Conventional variables used in mechanics can be relevant attributes to define the stability of a dissipative process, but there are situations in which unconventional attributes, such as smell and color, can become crucial observable qualities (e.g., in ecology in relation to the admissibility of the expected associative context). This is the norm rather than the exception for complex adaptive systems. In this case, it is not always easy to quantify the relevance of these attributes in terms of an energy quality index.

To make things more challenging, we must bear in mind the fact that not only the observed system but also the observer is becoming something else in time. That is, an observer with different goals, experiences, fears, and knowledge will never perceive and represent energy forms, energy transformations, and a relative set of "indexes of quality" in the same way as did a previous one.

The consequences of this fact deserve attention. This entails that, on the one hand, various concepts derived from thermodynamic analysis used to generate energy indexes of quality (e.g., output/input energy ratios, exergy-based indexes, entropy-related concepts, embodied assessments associated with energy flows) do carry a powerful metaphorical message. On the other hand, like all metaphors, they always require a semantic check before their use. That is, concepts such as efficiency, efficacy, and other quality indexes are very useful in a discussion about sustainability issues. The problem is that they cannot and should not be used to provide normative indications about how to deal with a sustainability predicament in an algorithmic way. That is, policies cannot be chosen or evaluated on the basis of the maximization of a single parameter (e.g., efficiency, power, rate of supply of negentropy) or on the basis of the application of a set of given rules written in protocols to a given simplified model of the reality.

The wisdom of Carnot, well before the complexity revolution, led him to similar conclusions in 1824 that can be found in the closing paragraph of his *Reflections on the Motive Power of Fire, and on Machines Fitted to Develop That Power*:

> *We should not expect ever to utilize in practice all the motive power of combustibles. The attempts made to attain this result would be far more harmful than useful if they caused other important considerations to be neglected. The economy of the combustible is only one of the conditions to be fulfilled in heat-engines. In many cases it is only secondary. It should often give precedence to safety, to strength, to the durability of the engine, to the small space which it must occupy, to small cost of installation, etc. To know how to appreciate in each case, at their true value, the considerations of convenience and economy which may present themselves; to know how to discern the more important of those which are only secondary; to balance them properly against each other in order to attain the best results by the simplest means; such should be the leading characteristics of the man called to direct, to co-ordinate the labours of his fellow men, to make them co-operate towards a useful end, whatsoever it may be.*

2. THE ENERGETICS OF HUMAN LABOR: AN EPISTEMOLOGICAL ANALYSIS OF THE FAILURE OF CONVENTIONAL ENERGY ANALYSIS

Attempts to apply energy analysis to human systems have a long history. Pioneering work was done by, among others, Podolinsky, Jevons, Ostwald, Lotka, White, and Cottrel. However, it was not until the 1970s that energy analysis became a fashionable scientific exercise, probably due to the oil crisis surging during that period. During the 1970s, energy input/output analysis was widely applied to farming systems and national economies and was applied more generally to describe the interaction of humans with their environment by Odum, Georgescu-Roegen, and Pimentel, among others. At the IFISA workshop of 1974, the term "energy analysis" (as opposed to "energy accounting") was officially coined. The second energy crisis during the 1980s was echoed by the appearance of a new wave of interesting work by biophysical researchers and a second elaboration of their own original work by the "old guard." However, quite remarkably, after less

than a decade or so, the interest in energy analysis declined outside the original circle. Indeed, even the scientists of this field soon realized that using energy as a numeraire to describe and analyze changes in the characteristics of agricultural and socioeconomic systems proved to be more complicated than one had anticipated.

We start our critical appraisal of the epistemological foundations of conventional energy analysis using one of the most well-known case studies of this field: the attempt to develop a standardized tool kit for dealing with the energetics of human labor. This has probably been the largest fiasco of energy analysis given the huge effort dedicated to this by the community of energy analysts.

Looking over the vast literature on the energetics of human labor, one can find agreement about the need to know at least three distinct pieces of information, which are required simultaneously, to characterize indexes of "efficiency" or "efficacy":

1. *The requirement (and/or availability) of an adequate energy input needed to obtain the conversion of interest.* In the case of human labor, a flow of energy contained in food is energy carriers compatible with human metabolism.
2. *The ability of the considered converter to transform the energy input into a flow of useful energy to fulfill a given set of tasks.* In this case, a system made up of humans has to be able to convert available food energy input into useful energy at a certain rate, depending on the assigned task.
3. *The achievement obtained by the work done, that is, the results associated with the application of useful energy to a given set of tasks.* In this case, this has to do with the usefulness of the work done by human labor in the interaction with the context.

If we want to use indexes based on energy analysis to formalize the concept of performance, we have to link these three pieces of information to numerical assessments based on "observable" qualities (Fig. 1). For this operation, we need at least four nonequivalent numerical assessments related to the following:

1. *Flow of a given energy input that is required and consumed by the converter.* In the case of the study of human labor, it is easy to define what should be considered food (energy carriers) for humans: something that can be digested and then transformed into an input for muscles. If the converter were a diesel engine, food would not be considered an energy input. This implies that the energy input can be characterized and defined only in relation to the given identity of the converter using it.

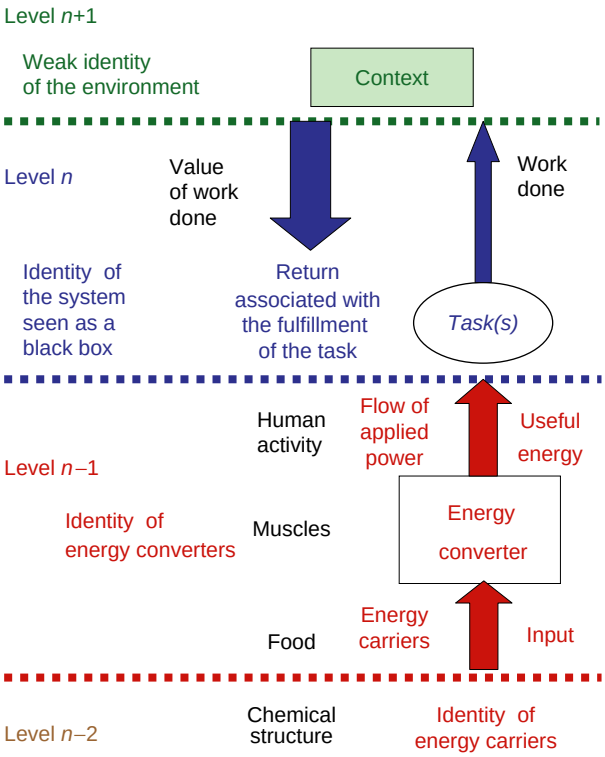


FIGURE 1 Relevant qualities for characterizing the behavior of an energy converter across hierarchical levels. From Giampietro, M. (2003). "Multi-Scale Integrated Analysis of Agroecosystems." CRC Press, Boca Raton, FL. Used with permission.

2. *The power level at which useful energy is generated by the converter.* This is a more elusive "observable quality" of the converter. Still, this information is crucial. As stated by Odum and Pinkerton in 1955, when dealing with the characterization of energy converters, we always have to consider both the pace of the throughput (the power) and the output/input ratio. A higher power level tends to be associated with a lower output/input ratio (e.g., the faster one drives, the lower the mileage of one's car). On the other hand, it does not make any sense to compare two output/input ratios if they refer to different throughput rates. In fact, we cannot say that a truck is less efficient than a small motorbike on the basis of the information given by a single indicator, that is, by the simple assessment that the truck uses more gas per mile than does the motorbike. However, after admitting that the power level is another piece of information that is required to assess the performance of an energy converter, it becomes very difficult to find a standard definition of "power level" applicable to "complex

energy converters" (e.g., to compare human workers to light bulbs or economic sectors). This is especially true when these systems are operating on multiple tasks. Obviously, this is reflected in an impossibility to define a standard experimental setting to measure such a value for nonequivalent systems. Moreover, power level (e.g., 1000 horsepower of a tractor vs 0.1 horsepower of a human worker) does not map onto either energy input flows (how much energy is consumed by the converter over a given period of time used as a reference [e.g., 1 year]) or how much applied power is delivered (how much useful energy has been generated by a converter over a given period of time used as reference [e.g., 1 year]). In fact, these two different pieces of information depend on how many hours the tractor or human worker has worked during the reference period and how wisely the tractor or human worker has been operating. This is reflected by the need to introduce a third piece of information to the analysis.

3. *Flow of applied power generated by the conversion.* The numerical mapping of this quality clearly depends heavily on the previous choices about how to define and measure power levels and power supply. In fact, an assessment of the flow of applied power represents the formalization of the semantic concept of "useful energy." Therefore, this is the measured flow of energy generated by a given converter (e.g., an engine) that is used to fulfill a specified task (e.g., water pumped out of the well). To make things more difficult, the definition of usefulness of such a task can be given only when considering the hierarchical level of the whole system to which the converter belongs (e.g., if the water is needed to do something by the owner of the pump). Put differently, such a task must be defined as "useful" by an observer operating at a hierarchical level higher than the level at which the converter is transforming energy input into useful energy. What is produced by the work of a tractor has a value that is generated by the interaction of the farm with a larger context (e.g., the selling of products on the market). That is, the definition of "usefulness" refers to the interaction of the whole system—the farm, which is seen as a black box (to which the converter belongs as a component) with its context. Therefore, the assessment of the quality of "usefulness" requires a descriptive domain different from that used to represent the conversion at the level of the converter (gasoline into motive power in the engine). This introduces a first major epistemological complication: the definition of usefulness of a given task (based on the return that this task implies for the whole system) refers to a representation of events on a given hierarchical level (the interface $n/n+1$) that is different from the hierarchical level used to describe and represent (assess indexes of efficiency of) the conversion of the energy input into useful energy (the interface $n-1/n$).

4. *Work done by the flow of applied power, that is, what is achieved by the physical effort generated by the converter.* Work is another very elusive quality that requires a lot of assumptions to be measured and quantified in biophysical terms. The only relevant issue here is that this represents a big problem regarding energy analysis. Even if assessments 3 and 4 use the same measurement unit (e.g., megajoules), they are different in terms of the relevant observable qualities of the system. That is, assessment 3 (applied power) and assessment 4 (work done) not only do not coincide in numerical terms but also require different definitions of descriptive domain. In fact, the same amount of applied power can imply differences in achievement due to differences in design of technology and know-how when using it. The problem is that it is impossible to find a "context-independent quality factor" that can be used to explain these differences in substantive terms.

The overview provided in Fig. 1 should already make clear to those familiar with epistemological implications of hierarchy theory that practical procedures used to generate numerical assessments within a linear input/output framework cannot escape the unavoidable ambiguity and arbitrariness implied by the hierarchical nature of complex systems. A linear characterization of input/output using the four assessments discussed so far requires the simultaneous use of at least two nonequivalent and nonreducible descriptive domains. Therefore, a few of these four assessments are logically independent, and this opens the door to an unavoidable degree of arbitrariness in the problem structuring (root definitions of the system). Any definition or assessment of energy flows, as input and/or output, will in fact depend on an arbitrary choice made by the analyst about what should be considered the focal level n, that is, what should be considered as a converter, what should be considered as an energy carrier, what should be considered as the whole system to which the converter belongs, and what has to be included and excluded in the characterization of the environment when checking the admissibility of boundary conditions and the usefulness of work done. A linear representation in energy analysis forces the analyst to decide "from scratch," in a

situation of total arbitrariness, on the set of formal identities adopted in the model to represent energy flows and conversions using the finite set of variables used to describe changes in the various elements of interest. This choice of a set of formal identities for representing energy carriers, converters, and the black box—the finite set of characteristics related to the definition of an admissible environment—is then translated into the selection of the set of epistemic categories (variables) that will be used in the formal model. It is at this point that the capital sin of the assumption of linear representation becomes evident. No matter how smart the analyst may be, any assessment of energy input (the embodied energy of the input) or energy output (what has been achieved by the work done) will be unavoidably biased by the arbitrary preanalytical choice of root definitions. In terms of hierarchy theory, we can describe this fact as follows.

The unavoidable preliminary triadic filtering needed to obtain a meaningful representation of the reality is at the basis of this impasse. For this filtering, we have to select (1) the interface between the focal level n and the lower level $n-1$ to represent the structural organization of the system and (2) the interface between the focal level n and higher level $n+1$ to represent the relational functions of the system. Ignoring this fact simply leads to a list of disorganized nonequivalent and nonreducible assessments of the same concepts. A self-explanatory example of this standard impasse in the field of energy analysis is given in Table I, which reports an example of several nonequivalent assessments of the energetic equivalent of 1 h of labor found in the literature.

That is, every time we choose a particular hierarchical level of analysis for assessing an energy flow (e.g., an individual worker over a day), we are also selecting a space–time scale at which we will describe the process of energy conversion (e.g., over a day, over a year, over the entire life). This, in turn, implies a nonequivalent definition of the context or environment and of the lower level where structural components are defined. Human workers can be seen as individuals operating over a 1-h time horizon where muscles are the converters, or they can be seen as citizens of developed countries where machines are the converters. The definitions of identities of these elements must be compatible with the identity of energy carriers such that individuals eat food, whereas developed countries eat fossil energy. This implies that, whatever we choose as a model, the various identities of energy carriers, parts, whole, and environment have to make sense in their reciprocal conditioning. Obviously, a different choice of hierarchical level considered as the focal level (e.g., the worker's household) requires adopting a different system of accounting for input and output.

3. THE PROBLEMATICS RELATED TO FORMAL DEFINITIONS OF ENERGY, WORK, AND POWER IN PHYSICS AND THEIR APPLICABILITY TO ENERGY ANALYSIS

In the previous case study, we have actually dealt the problematics related to formal definitions of energy, power, and work when dealing with an energetic assessment of human labor. Because this is a point carrying quite heavy epistemological implications, we would like to elaborate a bit more on it. To be able to calculate substantive indexes of performance based on concepts such as efficiency (maximization of an output/input ratio) and efficacy (maximization of an indicator of achievement in relation to an input) within energy analysis, we should be able to define, first, three basic concepts—"energy," "work," and "power"—in substantive terms. That is, we should be able to agree on definitions that are independent of the special context and settings in which these three concepts are used. However, if we try to do that, we are getting into an even more difficult situation.

3.1 Energy

Much of the innate indeterminacy of energy analysis, especially when applied to complex systems, has its roots in the problematic definition of energy in physics. As Feynman and colleagues pointed out in 1963, we have no knowledge of what energy is, and energy is an abstract thing in that it does not tell us the mechanism or the reasons for the various formulas. In practice, energy is perceived and described in a large number of different forms: gravitational energy, kinetic energy, heat energy, elastic energy, electrical energy, chemical energy, radiant energy, nuclear energy, mass energy, and so on. A general definition of energy, without getting into specific context and space–time scale-dependent settings, is necessarily limited to a vague expression such as "the potential to induce physical transformations." Note that the classical definition of energy in conventional physics textbooks, "the potential to do

TABLE I

Nonequivalent Assessments of Energy Requirement for 1 h of Human Labor

		Energy input per hour of labor		
Method[a]	Worker system boundaries[b]	Food energy input[c] (MJ/h)	Exosomatic energy input[d] (MJ/h)	Other energy forms across scales[e]
1	Man	0.5[f]	Ignored	Ignored
1	Woman	0.3[g]	Ignored	Ignored
1	Adult	0.4[h]	Ignored	Ignored
2	Man	0.8[f]	Ignored	Ignored
2	Woman	0.5[g]	Ignored	Ignored
2	Adult	0.6[h]	Ignored	Ignored
3	Man	1.6[f]	Ignored	Ignored
3	Woman	1.2[g]	Ignored	Ignored
3	Adult	1.3[h]	Ignored	Ignored
4	Man	2.5[f]	Ignored	Ignored
4	Woman	1.8[g]	Ignored	Ignored
4	Adult	2.1[h]	Ignored	Ignored
5	Household	3.9[i]	Ignored	Ignored
6	Society	4.2[j]	Ignored	Ignored
7a	Household	3.9[i]	39 (food system)[k]	Ignored
7b	Society	4.2[j]	42 (food system)[k]	Ignored
8	Society	Ignored	400 (society)[l]	Ignored
9	Society	Ignored	400 (society)[l]	2×10^{10}MJ[m]

Source. From Giampietro, M. (2003). "Multi-Scale Integrated Analysis of Agroecosystems." CRC Press, Boca Raton, FL. Used with permission.

[a] The nine methods considered are (1) only extra metabolic energy due to the actual work (total energy consumption minus metabolic rate) in 1 h; (2) total metabolic energy spent during actual work (including metabolic rate) in 1 h; (3) metabolic energy spent in a typical work day divided by the hours worked in that day; (4) metabolic energy spent in 1 year divided by the hours worked in that year. (5) Same as method 4 but applied at hierarchical level of household; (6) same as method 4 but applied at the level of society; (7) besides food energy, also including exosomatic energy spent in food system per year divided by the hours of work in that year (7a at household level, 7b at society level); (8) total exosomatic energy consumed by society in 1 year divided by the hours of work delivered in that society in that year; and (9) assessing the megajoules of solar energy equivalent of the amount of fossil energy assessed with method 8.

[b] The "systems delivering work" considered are typical adult man (Man), typical adult woman (Woman), average adult (Adult), typical household, and an entire society. Definitions of "typical" are arbitrary and serve only to exemplify methods of calculation.

[c] Food energy input is approximated by the metabolic energy requirement. Given the nature of the diet and food losses during consumption, this flow can be translated into food energy consumption. Considering also postharvest food losses and preharvest crop losses, it can then be translated into different requirements of food (energy) production.

[d] We report here an assessment referring only to fossil energy input.

[e] Other energy forms, acting now or in the past, that are (were) relevant for the current stabilization of the described system even if operating on space–time scales not detected by the actual definition of identity for the system.

[f] Based on a basal metabolic rate for adult men of BMR = 0.0485 W + 3.67 MJ/day (W = weight = 70 kg) = 7.065 MJ/day = 0.294 MJ/h. Physical activity factor for moderate occupational work (classified as moderate) 2.7 × BMR. Average daily physical activity factor 1.78 × BMR (moderate occupational work). Occupational workload: 8 h/day considering work days only; 5 h/day average workload over the entire year, including weekends, holidays, and absence.

[g] Based on a basal metabolic rate for adult women of BMR = 0.0364 W + 3.47 (W = weight = 55 kg) = 5.472 MJ/day = 0.228 MJ/h. Physical activity factor for moderate occupational work 2.2 × BMR. Average daily physical activity factor (based on moderate occupational work) 1.64 × BMR. Occupational workload: 8 h/day considering work days only; 5 h/days average work load over the entire year, including weekends, holidays, and absence.

[h] Assuming a 50% gender ratio.

[i] A typical household is arbitrarily assumed to consist of one adult male (70 kg, moderate occupational activity), one adult female (55 kg, moderate occupational activity), and two children (male of 12 years and female of 9 years).

[j] This assessment refers to the United States. In 1993, the food energy requirement was 910,000 TJ/year and the work supply was 215 billion h.

[k] Assuming 10 MJ of fossil energy spent in the food system per 1 MJ of food consumed.

[l] Assuming a total primary energy supply in 1993 in the United States (including the energy sector) of 85 million TJ divided by a work supply of 215 billion h.

[m] Assuming a transformity ratio for fossil energy of 50 million EMJ/joule.

work," refers to the concept of "free energy" or "exergy," and this is another potential source of confusion. Both of these concepts require a previous formal definition of work and a clear definition of operational settings to be applied.

3.2 Work

The similar ambiguous definition of energy is also found when dealing with a general definition of work, a definition applicable to any specific space–time scale-dependent settings. The classical definition found in dictionaries refers to the ideal world of elementary mechanics: "work is performed only when a force is exerted on a body while the body moves at the same time in such a way that the force has a component in the direction of the motion." Others express work in terms of "equivalent to heat," as is done in thermodynamics: "the work performed by a system during a cyclic transformation is equal to the heat absorbed by the system." This calorimetric equivalence offers the possibility to express assessments of work in the same unit as energy, that is, in joules. However, the elaborate description of work in elementary mechanics and the calorimetric equivalence derived from classical thermodynamics are of little use in real-life situations. Very often, to characterize the performance of various types of work, we need inherently qualitative characteristics that are impossible to quantify in terms of "heat equivalent." For example, the work of a director of an orchestra standing on a podium cannot be described using the previously given definition of elementary mechanics, nor can it be measured in terms of heat. An accounting of the joules related to how much the director sweats during the execution of a musical score will not tell anything about the quality of his or her performance. "Sweating" and "directing well" do not map onto each other. Indeed, it is virtually impossible to provide a physical formula or general model defining the quality (or value) of a given task (what has been achieved after having fulfilled a useful task) in numerical terms from within a descriptive domain useful in representing energy transformations based on a definition of identity for energy carriers and energy converters. In fact, formal measurements leading to energetic assessments refer to transformations occurring at a lower level (e.g., how much heat has been generated by muscles at the interface level $n/n-1$), whereas an assessment of performance (e.g., how much the direction has been appreciated by the public) refers to interactions and feedback occurring at the interface level $n/n+1$). This standard predicament applies to all real complex systems and especially to human societies.

3.3 Power

The concept of power is related to "the time rate of an energy transfer" or, when adopting the theoretical formalization based on elementary mechanics, to "the time rate at which work is being done." Obviously, this definition cannot be applied without a previous valid and formal mapping of the "transfer of energy" in the form of a numerical indicator or of "doing the work." At this point, it should be obvious that the introduction of this "new" concept, based on the previous ones, does not get us out of the predicament experienced so far. Any formal definition of power will run into the same epistemological impasse discussed for the previous two concepts given that it depends on the availability of a valid definition of them in the first place.

It should be noted, however, that the concept of power introduces an important new qualitative aspect—a new attribute required to characterize the energy transformation—not present in the previous two concepts. Whereas the concepts of energy and work, as defined in physics, refer to quantitative assessment of energy without taking into account the time required for the conversion process under analysis, the concept of power is, by definition, related to the rate at which events happen. This introduces a qualitative dimension that can be related either to degree of organization of the dissipative system or to the size of the system performing the conversion of energy in relation to the processes that guarantee the stability of boundary conditions in the environment. To deliver power at a certain rate, a system must have two complementing but distinct relevant features. First, it must have an "adequate organized structure" to match the given task. In other words, it must have the capability of doing work at a given rate. For example, an individual human cannot process and convert into power 100,000 kcal of food in a day. Second, it must have the capability of securing a sufficient supply of energy input for doing the task. For example, to gain an advantage from the power of 100 soldiers for 1 year, we must be able to supply them with enough food. In short, "gasoline without a car" (the first feature) or "a car without gasoline" (the second feature) is of no use.

This is the point that exposes the inadequacy of the conventional output/input approach. In hierarchically organized adaptive systems, it is not

thinkable to have an assessment of energy flows without properly addressing the set of relevant characteristics of the process of transformation that are level and scale dependent in their perception and representation. The analyst must address the existence of expected differences in the perception and representation of concepts such as energy, power, and work done on different hierarchical levels and associated with the multiscale identity of the dissipative system in the first place. This means that the three theoretical concepts—energy, work, and power—must be simultaneously defined and applied within such a representation in relation to the particular typology of dissipative system (engines of cars are different from muscles in humans).

At this point, if we accept the option of using nonequivalent descriptive domains in parallel, we can establish a reciprocal entailment on the definitions of identity of the various elements (converter, whole system, energy carriers, and environment) to characterize the set of energy transformations required to stabilize the metabolism of a dissipative system.

In this case, we have two couples of self-entailment among identities:

Self-Entailment 1: The identities adopted for the various converters (on interface level n and level $n-1$) referring to the power level of the converter *define/are defined by* the identities adopted for the energy carriers (on interface level $n-1$ and level $n-2$).

Self-Entailment 2: The identities of the set of energy forms considered to be useful energy on the focal level, that is, characterization of the autocatalytic loop from within (on interface between level $n-1$ and level n) in relation to their ability, validated in the past, to fulfill a given set of tasks, *define/are defined by* the compatibility of the identities of the whole system in relation to its interaction with the larger context, that is, characterization of the autocatalytic loop from outside (on the interface between level n and level $n+1$). This compatibility can be associated with the determination of the set of useful tasks in relation to the availability in the environment of the required flow of input and sink capacity for wastes.

When using these two set of relations and identities, which depend on each other for their definitions, simultaneously, we can link nonequivalent characterizations of energy transformations across scales (Fig. 2).

Bridge 1: conversion rates represented on different levels must be compatible with each other. This implies a constraint of compatibility between the definition of identity of the set of converters defined at level $n-1$ (triadic reading: $n-2/n-1/n$) and the definition of the set of tasks for the whole defined at level $n+1$ (triadic reading: $n-1/n/n+1$). This constraint addresses the ability of the various converters to generate "useful energy" (the right energy form applied in the specified setting) at a given rate that must be admissible for the various tasks. This bridge deals with qualitative aspects of energy conversions.

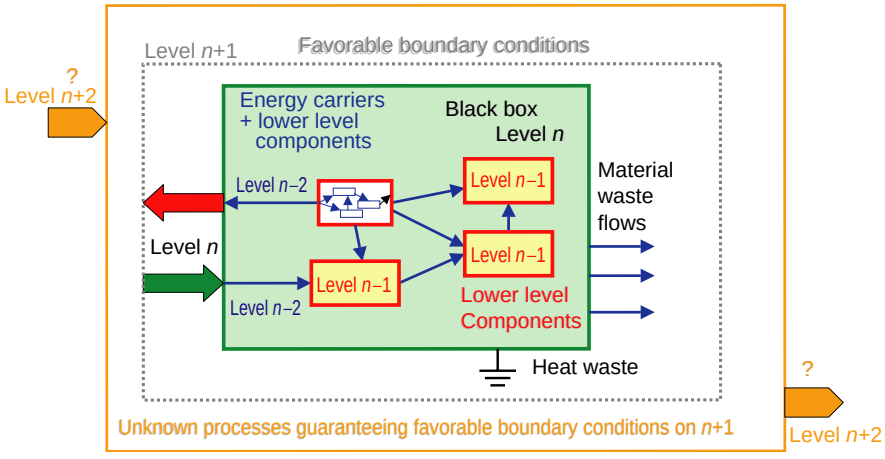


FIGURE 2 Hierarchical levels that should be considered when studying autocatalytic loops of energy forms. From Giampietro, M. (2003). "Multi-Scale Integrated Analysis of Agroecosystems." CRC Press, Boca Raton, FL. Used with permission.

Bridge 2: The flow of energy input from the environment and the sink capacity of the environment must be enough to cope with the rate of metabolism implied by the identity of the black box. This implies a constraint of compatibility between the size of the converters defined at level $n-1$ (triadic reading: $n-2/n-1/n$) and the relative supply of energy carriers and sink capacity related to processes occurring at level $n+1$. The set of identities of the input (from the environment to the converters) and output (from the converters to the environment) is referring to a representation of events belonging to level $n-2$. In fact, these energy carriers will interact with internal elements of the converters to generate the flow of useful energy and will be turned out into waste by the process of conversions. However, the availability of an adequate supply of energy carriers and of an adequate sink capacity is related to the existence of processes occurring in the environment that are needed to maintain favorable conditions at level $n+1$. Differently put, the ability to maintain favorable conditions in the face of a given level of dissipation can be defined only by considering level $n+1$ as the focal one (triadic reading: $n/n+1/n+2$).

This means that when dealing with quantitative and qualitative aspects of energy transformations over an autocatalytic loop of energy forms, we have to bridge at least five hierarchical levels (from level $n-2$ to level $n+2$).

Unfortunately, by definition, the environment (processes determining the interface between levels $n+1$ and $n+2$) is something we do not know enough about. Otherwise, it will become part of the modeled system. This means that when dealing with the stability of "favorable boundary conditions," we can only hope that they will remain the way they for as long as possible. On the other hand, the existence of favorable boundary conditions is a must for dissipative systems. That is, the environment is and must generally be assumed to be an "admissible environment" in all technical assessments of energy transformations.

Therefore, the existence of favorable boundary conditions" (interface level $n+1/n+2$) is an assumption that is not directly related to a definition of usefulness of the tasks (interface level $n/n+1$) in relation to the issue of sustainability. The existing definition of the set of useful tasks at level n simply reflects the fact that these tasks were perceived as useful in the past by those living inside the system. That is, the established set of useful tasks was able to sustain a network of activities compatible with boundary conditions (*ceteris paribus* at work). However, this definition of usefulness for these tasks (what is perceived as good at level n according to favorable boundary conditions at level $n+1$) has nothing to do with the ability or effect of these tasks in relation to the stabilization of boundary conditions in the future. For example, producing a given crop that provided an abundant profit last year does not necessarily imply that the same activity will remain so useful next year. Existing favorable boundary conditions at level $n+1$ require the stability of the processes occurring at level $n+2$ (e.g., the demand for that crop remains high in the face of a limited supply, natural resources such as nutrients, water, soil, and pollinating bees will remain available for the next year). This is information we do not know, and cannot know enough about, in advance. This implies that analyses of "efficiency" and "efficacy" are based on data referring to characterizations and representations relative to identities defined only on the four levels $n-2$, $n-1$, n, and $n+1$ (on the *ceteris paribus* hypothesis and reflecting what has been validated in the past). Because of this, they are not very useful in studying coevolutionary trajectory of dissipative systems. In fact, they (1) do not address the full set of relevant processes determining the stability of favorable boundary conditions (they miss a certain number of relevant processes occurring at level $n+2$) and (2) deal only with qualitative aspects (intensive variables referring to an old set of identities) but not quantitative aspects such as the relative size of components (how big is the requirement of the whole dissipative system—an extensive variable assessing the size of the box from the inside—in relation to the unknown processes that stabilize the identity of its environment at level $n+2$). As observed earlier, the processes that are stabilizing the identity of the environment are not known by definition. This is why, when dealing with coevolution, we have to address the issue of "emergence." That is, we should expect the appearance of new relevant attributes (requiring the introduction of new epistemic categories in the model, that is, new relevant qualities of the system so far ignored) to be considered as soon as the dissipative system (e.g., human society) discovers or learns new relevant information about those processes occurring at level $n+2$ that were not known before.

4. THE METAPHOR OF PRIGOGINE SCHEME TO CHARACTERIZE AUTOCATALYTIC LOOPS OF ENERGY FORMS IN HIERARCHICALLY ORGANIZED ADAPTIVE SYSTEMS

This section applies the rationale of hierarchical reading of energy transformations to the famous scheme proposed by Prigogine to explain the biophysical feasibility of dissipative systems in entropic terms. As Schroedinger showed, living systems can escape the curse of the second law thanks to their ability to feed on "negentropy." Differently put, living systems are, and must be, open systems that can preserve their identity thanks to a metabolic process.

This original idea has been further developed by the work of the school of Prigogine in nonequilibrium thermodynamic. These researchers introduced the class of dissipative systems in which the concept of entropy is associated with that of self-organization and emergence. In this way, it becomes possible to better characterize the concept of metabolism of dissipative systems. The ability to generate and preserve in time a given pattern of organization depends on two self-entailing abilities: (1) the ability to generate the entropy associated with the energy transformations occurring within the system and (2) the ability to discharge this entropy into the environment at a rate that is at least equal to that of internal generation. As observed by Schneider and Kay in a famous article in 1994, the possibility to have life, self-organization, and autocatalytic loops of energy forms is strictly linked to the ability of open dissipative systems to generate and discharge entropy into an admissible environment. Actually, the more such systems can generate and discharge entropy, the higher the complexity of the patterns they can sustain.

Dissipative systems have to maintain a level of entropy generation and discharge that is admissible in relation to their identity of metabolic systems. Differently put, the pattern of energy dissipation associated with their expected identity in terms of an ordered structure of material flows across hierarchical levels must be compatible with a set of favorable boundary conditions within the environment. Using the vocabulary developed by Schroedinger and Prigogine, this requires compensating the unavoidable generation of entropy associated with internal irreversibility (dS_i) by importing an adequate amount of negentropy (dS_e) from their environment. The famous relation proposed by Prigogine to represent this idea can be related as

$$dS_T \Leftrightarrow dS_i + dS_e \quad (1)$$

Relation (1) indicates that the identity of the system defined at interface level $n/n-1$ must be congruent or compatible with the identity of the larger dissipative system defined at interface level $n+1/n$. In other words, the two flows of entropy internally generated (dS_i) and "imported/exported" (dS_e) must be harmonized with change in entropy of the system (dS_T) in which the identity of the first is embedded.

Let us now try to interpret, in a metaphorical sense, the scheme proposed by Prigogine to characterize autocatalytic loops of energy forms within nested metabolic systems.

First, dS_i refers to the representation of the mechanism of internal entropy production that is associated with the irreversibility generated to preserve system identity. This assessment necessarily must be obtained using a representation of events related to a perception obtained within the black box on interface level $n-2/n-1/n$. This can be done by using a mapping of an energy form that makes it possible to represent how the total input used by the black box is then invested among the various parts (the internal working of the metabolism).

Second, dS_e refers to the representation of imported negative entropy. The term "negentropy" entails the adoption of a mechanism of mapping that has to be related to the previous choices made when representing dS_i. This term requires establishing a bridge between the assessment of two forms of energy: (1) one used to describe events within the black box in terms of dS_i (the mapping used to assess the effect of internal mechanisms associated with irreversibility) and (2) another used to describe how the metabolism of the larger system (the environment) providing favorable boundary conditions is affected by the activity of the black box. This implies using the mapping of a second energy form. This second energy form, reflecting the interference of the black box over the metabolism of the environment, has to be effectively linkable to the first energy form.

The first energy form is used to represent relevant mechanisms inside the black box. The second energy form is used to represent the effects of the interaction of the black box with its associative context. It represents the view of the metabolism of the black box from the outside.

Third, dS_T translates into a required compatibility, in quantitative and qualitative terms, of the overall interaction of the black box with the environment according to the particular selection of mapping of the two energy forms. Whenever $dS_T < 0$, the system can expand its domain of activity, either by growing in size (generating more of the same pattern, so to speak) – or by moving to another dissipative pattern (developing emergent properties). On the contrary, whenever $dS_T > 0$, the system has to reduce its activity. It can remain alive, at least for a while, by sacrificing a part of its activities. In this case, a certain level of redundancy can be burned, providing a temporary buffer, to get through critical moments and cyclic perturbations (disappearance of expected favorable conditions).

As noted earlier, it is not possible to obtain a simultaneous check in formal terms—and in a substantive way—of the relation among the three *dS*s defined on different descriptive domains. On the other hand, an autocatalytic loop of different energy forms can be used to study the forced congruence across levels of some of the characteristics of the autocatalytic loop defined on different descriptive domains (Fig. 3).

In Fig. 3A, we resume the key features of the energetic analysis of dissipative systems. In particular, we provide a different perspective on the discussion related to the representation of hierarchically organized dissipative systems (Fig. 1). That is, the network represented in Fig. 2 can be represented by dividing the components described at level $n-1$ into two compartments: (1) those that do not interact directly with the environment (aggregated in the compartment labeled "indirect" in Fig. 3A) and (2) those that interact directly with the environment by gathering environmental inputs and providing those useful tasks aimed at such a gathering (aggregated in the compartment labeled "direct" in Fig. 3A). Thanks to the existence of

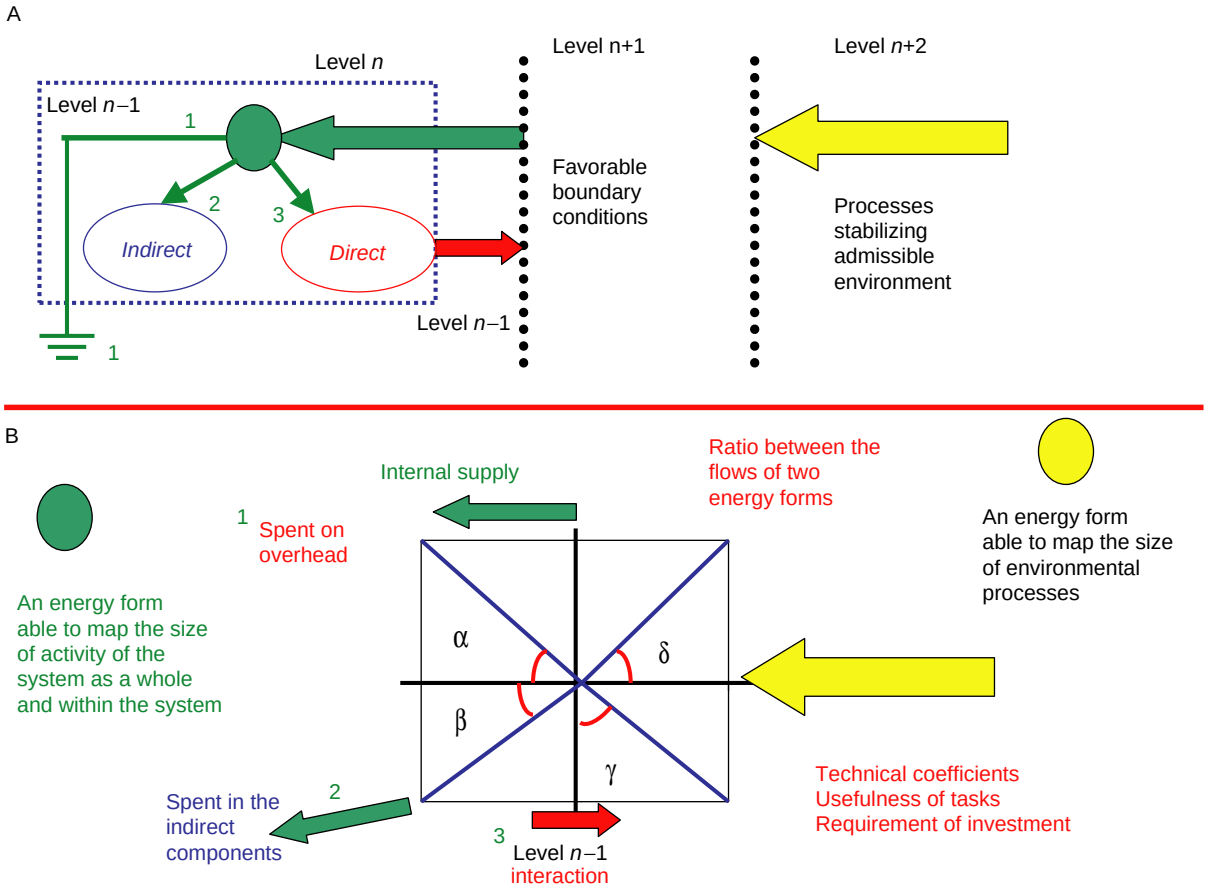


FIGURE 3 Congruence across levels of characteristics of the autocatalytic loop defined over nonequivalent descriptive domains. From Giampietro, M. (2003). "Multi-Scale Integrated Analysis of Agroecosystems." CRC Press, Boca Raton, FL. Used with permission.

favorable conditions (perceived at level $n+1$), the whole system (at level n) can receive an adequate supply of required input. This input can then be expressed in an energy form that is used as reference to assess how this total input is then invested over the lower level compartment. Therefore, we can expect that this total input will be dissipated within the black box in three flows (indicated by the three arrows in Fig. 3A): a given overhead (1) for the functioning of the whole, (2) for the operation of the compartment labeled "indirect," and (3) for the operation of the compartment labeled "direct." The favorable conditions perceived at level $n+1$, which make possible the stability of the environmental input consumed by the whole system, in turn are stabilized due to the existence of some favorable gradient generated elsewhere (level $n+2$), and this is not accounted for in the analysis. This available gradient can be exploited thanks to the tasks performed by the components belonging to the direct compartment (energy transformations occurring at level $n-1$). The return between the energy input made available to the whole system (at level n) per unit of useful energy invested by the direct compartment into the interaction with the environment will determine the strength of the autocatalytic loop of energy generated at level $n-1$ by the exploitation of a given energy resource. This integrated use of representations of energy transformation across levels is at the basis of a new approach that can be used to describe the sustainability of complex adaptive systems that base their identity on metabolic processes. In particular, three very useful concepts are (1) a mosaic effect across hierarchical levels, (2) impredicative loop analysis, and (3) narratives capable of surfing in complex time. These concepts are required to provide coherent representation of identities of energy forms generating an autocatalytic loop over five contiguous hierarchical levels (from level $n-2$ to level $n+2$). These three concepts can be used to structure the information space in a way that makes it possible to perform a series of congruence checks on the selected representation of the autocatalytic loop of energy forms.

An example of this congruence check is shown in Fig. 3. It is possible to establish a relation among the representation of parts, whole, and interaction with the environment across scales using a four-angle figure that combines intensive and extensive variables used to represent and bridge the characterization of metabolic processes across levels. The two angles on the left side (α and β) refer to the distribution of the total available supply of energy carriers (indicated on the upper part of the vertical axis) among the three flows of internal consumption. The angle α refers to the fraction of the total supply that is invested in overhead. The angle β refers to the fraction of the total supply that is spent on the operation of internal components. The value of the lower part of the vertical axis at this point refers to the amount of primary energy (the mapping related to the internal perception of energy form) that is invested in the direct interaction with the environment. The two angles on the right (γ and δ) refer to a characterization of the interaction of the system with the environment.

The selection of mapping (what set of variables have to be used in such a representation) must fulfill the double task of making it possible to relate the perception and representation of relevant characteristics of what is going on inside the black box to characteristics that are relevant to studying the stability of the environment. That is, this choice does not have the goal of establishing a direct link between a model studying the dynamics inside the black box and models studying relevant dynamics in the environment. To reiterate, this is simply not possible. This analytical tool simply makes it possible to establish bridges among nonequivalent representations of the identity of parts and wholes.

An example of three impredicative loop analyses is presented in Fig. 4. These formalizations reflect three logically independent ways of looking at an autocatalytic loop of energy forms according to the scheme presented in Fig. 3. It is very important to note that these three formalizations cannot be directly linked to each other because they are constructed using logically independent perceptions and characterizations. However, they share a meta-model used for the semantic problem structuring that made it possible to tailor in the formalization (when putting numerical assessment in it) according to the system considered.

5. CONCLUSION

The existence of both multiple levels of organization in the investigated system and different useful perspectives for looking at the reality entails the use of multiscale analysis when dealing with the sustainability of dissipative systems organized in nested hierarchies. This translates into a severe epistemological challenge for energy analysis. Innovative

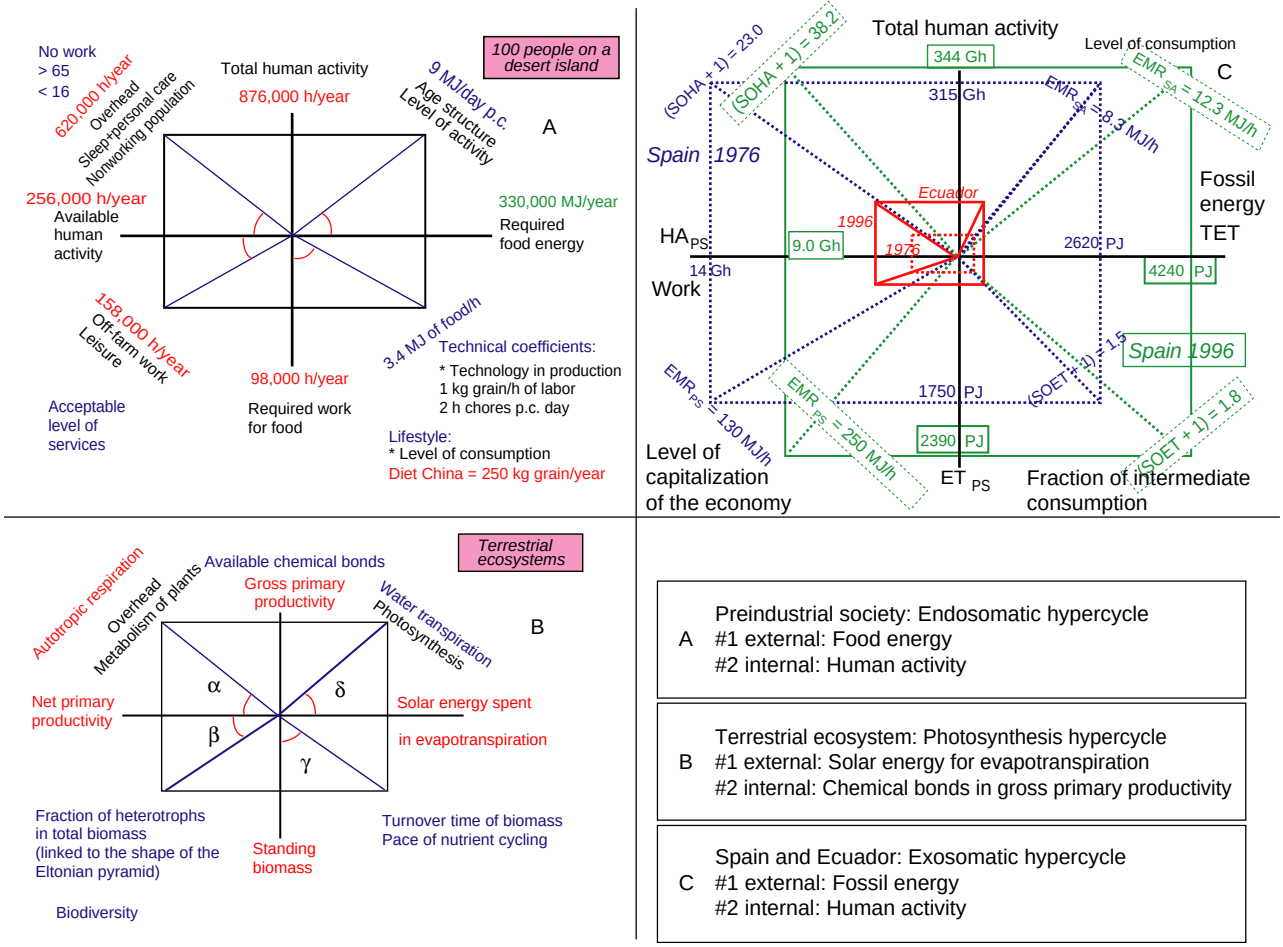


FIGURE 4 Autocatalytic loop of energy forms. From Giampietro, M. (2003). "Multi-Scale Integrated Analysis of Agroecosystems." CRC Press, Boca Raton, FL. Used with permission.

analytical methodologies can be developed taking advantage of principles derived from complex system theory and nonequilibrium thermodynamics. However, this innovative approach requires accepting two negative side effects.

First, the same autocatalytic loop can be represented in different legitimate ways (by choosing different combinations of formal identities). That is, there is no such thing as the "right model" when dealing with an analysis of the sustainability of an autocatalytic loop of energy. Rather, all models are incomplete. Analysts have the role of helping in the societal discussion about how to select narratives and models that are useful.

Second, there is an unavoidable degree of ignorance associated with any formal representation of an autocatalytic loop of energy forms. In fact, the very set of assumptions that make it possible to represent the system as viable (e.g., about the future stability of the relation between investment and return, about the future admissibility of the environment on level $n+2$, about the relevance of the selected set of attributes) guarantees that such a representation is affected by uncertainty. Both the observer and the observed system are becoming "*ceteris* are never *paribus*" when coming to the representation of the evolution in time of autocatalytic loops. The concept of entropy, in this case, translates into a sort of "yin yang" predicament for the analysis of complex adaptive systems. The more a dissipative system is successful in defending and imposing its identity on the environment, the sooner it will be forced to change it.

In short, any representation in energy terms of the predicament of sustainability for a complex dissipative system is (1) just one of many possible alternatives and (2) unavoidably affected by uncertainty, ignorance, and certain obsolescence.

SEE ALSO THE FOLLOWING ARTICLES

Human Energetics • Input–Output Analysis • Multicriteria Analysis of Energy • Thermodynamics, Laws of • Work, Power, and Energy

Further Reading

Allen, T. F. H., and Starr, T. B. (1982). "Hierarchy." University of Chicago Press, Chicago.

Giampietro, M. (2003). "Multi-Scale Integrated Analysis of Agroecosystems." CRC Press, Boca Raton, FL.

Hall, C. A. S., Cleveland, C. J., and Kaufmann, R. (1986). "Energy and Resource Quality." John Wiley, New York.

Herendeen, R. A. (1998). "Ecological Numeracy: Quantitative Analysis of Environmental Issues." John Wiley, New York.

Odum, H. T. (1996). "Environmental Accounting: Energy and Decision Making." John Wiley, New York.

O'Neill, R. V. (1989). Perspectives in hierarchy and scale. *In* "Perspectives in Ecological Theory" (J. Roughgarden, R. M. May, and S. Levin, Eds.), pp. 140–156. Princeton University Press, Princeton, NJ.

Pimentel, D., and Pimentel, M. (eds.). (1996). "Food, Energy, and Society," 2nd ed. University of Colorado Press, Niwot.

Prigogine, I. (1978). "From Being to Becoming." W. H. Freeman, San Francisco.

Prigogine, I., and Stengers, I. (1981). "Order out of Chaos." Bantam Books, New York.

Schneider, E. D., and Kay, J. J. (1994). Life as a manifestation of the second law of thermodynamics. *Math. Comp. Modelling* **19,** 25–48.

Smil, V. (1991). "General Energetics: Energy in the Biosphere." John Wiley, New York.

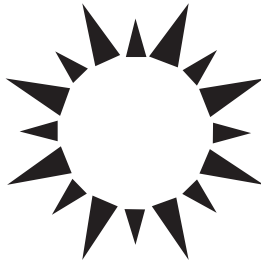
Rosen, R. (2000). "Essays on Life Itself." Columbia University Press, New York.

Rotmans, J., and Rothman, D. S. (eds.). (2003). "Scaling Issues in Integrated Assessment." Swets & Zeitlinger, Lissen, Netherlands.

Salthe, S. N. (1985). "Evolving Hierarchical Systems: Their Structure and Representation." Columbia University Press, New York.

Schrödinger, E. (1967). "What Is Life?" Cambridge University Press, Cambridge, UK.

Computer Modeling of Renewable Power Systems



PETER LILIENTHAL
U.S. Department of Energy, National Renewable Energy Laboratory
Golden, Colorado, United States

THOMAS LAMBERT
Mistaya Engineering, Inc.
Calgary, Alberta, Canada

PAUL GILMAN
U.S. Department of Energy, National Renewable Energy Laboratory
Golden, Colorado, United States

1. Introduction
2. Renewable Resource Modeling
3. Component Modeling
4. System Modeling
5. Models for Renewable Power System Design
6. Conclusions

Glossary

beam radiation Solar radiation that travels from the sun to Earth's surface without any scattering by the atmosphere. Beam radiation is also called direct radiation.

clearness index The fraction of solar radiation striking the top of Earth's atmosphere that is transmitted through the atmosphere to strike Earth's surface.

declination The latitude at which the sun's rays are perpendicular to the surface of Earth at solar noon.

diffuse radiation Solar radiation that has been redirected by atmospheric scattering.

extraterrestrial radiation Solar radiation striking the top of Earth's atmosphere.

global radiation The sum of beam radiation and diffuse radiation; sometimes called total radiation.

life-cycle cost The summation of all appropriately discounted costs, both recurring and nonrecurring, associated with a system or component of a system during its life span.

sensitivity analysis An investigation into the extent to which changes in certain inputs affect a model's outputs.

spinning reserve Surplus electrical generating capacity that is operating and able to respond instantly to a sudden increase in the electric load or to a sudden decrease in the renewable power.

Renewable power systems produce electricity from one or more renewable resources, and often incorporate nonrenewable power sources or energy storage devices for backup. Computer models of such systems are mathematical equations or algorithms programmed into a computer that can simulate system behavior or costs. Renewable power system designers use computer modeling to predict performance and explore different design options.

1. INTRODUCTION

Recent improvements in the economics of renewable power sources have made them competitive with conventional power sources in many situations. But the design of renewable power systems is made difficult by the novelty of the technologies, the intermittency and geographic diversity of the renewable resources on which they depend, and the large number of possible system configurations. Computer models, which are simplified mathematical representations of real systems, can help overcome these challenges. Models allow the designer to simulate the long-term performance and economics of renewable power systems, explore different design options, and test the effects of input assumptions.

This article deals with the modeling of solar, wind, small hydro, and biomass power systems suitable for small-scale (less than 1 MW) electric power generation, both stand-alone and grid-connected. The focus will be on long-term performance and economic modeling rather than dynamic modeling, which is concerned with short-term system stability and response. The material is structured to first address the modeling of each of the renewable resources, then the renewable technologies, then the power systems in which resources and technologies may be combined with each other and with more conventional power sources. In the final section, there is a description of the types of models currently available, and examples of each.

2. RENEWABLE RESOURCE MODELING

Characteristics of the renewable resources influence the behavior and economics of renewable power systems. Resource modeling is therefore a critical element of system modeling. In this section, the focus is on those aspects of each renewable resource that are important for modeling renewable power systems.

2.1 Solar Resource

Solar power technologies convert solar radiation into electricity. The goal of solar resource modeling is to estimate the amount of solar radiation available for conversion at some location, and how that amount varies over time. Four factors determine the amount of radiation striking a plane on Earth's surface: the amount of extraterrestrial radiation, the fraction of that radiation being transmitted through the atmosphere, the orientation of the plane, and shading of the plane by nearby objects such as buildings or mountains.

Extraterrestrial radiation varies in an entirely predictable fashion as Earth rotates on its axis and revolves around the sun. For any time between sunrise and sunset, the following equation gives the solar radiation incident on a horizontal plane at the top of the atmosphere:

$$G_o = G_{sc}\left(1 + 0.033\cos\frac{360n}{365}\right) \times (\cos\phi\cos\delta\cos\omega + \sin\phi\sin\delta)$$

where G_{sc} is the solar constant (normally taken as 1367 W/m^2), n is the day of the year, ϕ is the latitude (north positive), ω is the hour angle (zero at solar noon, increasing at 15° per hour throughout the day), and δ is the declination, given by

$$\delta = 23.45\sin\left(360\frac{284+n}{365}\right)$$

Integrating the extraterrestrial radiation over 1 hour gives the energy per unit area received over that hour. Integrating from sunrise to sunset gives the energy per unit area received daily. Summing this daily total for each day of the month gives the monthly total.

Atmospheric scattering and absorption reduce the amount of solar radiation striking Earth's surface to some fraction of the extraterrestrial radiation. This fraction, called the clearness index, varies stochastically in time, due principally to the fluctuating amount of water vapor and water droplets contained in the atmosphere. The equation for the clearness index is as follows:

$$k_t = G/G_o$$

where G is the radiation incident on a horizontal plane on the surface of Earth and G_o is the radiation incident on a horizontal plane at the top of the atmosphere. Atmospheric scattering affects the character as well as the amount of surface radiation. Under very clear conditions, the global radiation may be as much as 80% beam radiation and only 20% diffuse radiation. Under very cloudy conditions, the global radiation may be almost 100% diffuse radiation. Duffie and Beckman provide correlations that relate this diffuse fraction to the clearness index. The diffuse fraction is of interest to modelers for two reasons. First, the effect of surface orientation differs for beam and diffuse radiation, because beam radiation emanates only from the direction of the sun whereas diffuse radiation emanates from all parts of the sky. Second, some solar conversion devices (those that focus the sun's rays) can only make use of beam radiation.

Researchers have developed several algorithms for calculating the beam and diffuse radiation striking a surface with a particular slope (the angle between the surface and the horizontal) and azimuth (the direction toward which the surface faces), as well as means of accounting for shading from nearby objects (details can be found in the book by Duffie and Beckman, *Solar Engineering of Thermal Processes*). Most measured solar resource data are reported as global radiation on a horizontal surface, averaged for each hour, day, or month of the year. Models can use the above-mentioned algorithms to calculate the corresponding clearness indices, resolve the global radiation into its beam and diffuse components, and calculate the radiation striking a particular surface. The National Aeronautics and Space Administration

(NASA) Surface Solar Energy Web site provides satellite-measured monthly average radiation values for any location worldwide at a 1° resolution. Numerous other sources provide monthly or hourly data for more limited areas at higher spatial resolution. Several authors, including Mora-López and Sidrach-de-Cardona, have proposed algorithms for the generation of synthetic hourly radiation data from monthly averages.

2.2 Wind Resource

Wind turbines convert the kinetic energy of the wind into electricity. The goal of wind resource modeling is to describe the statistical properties of the wind at a particular location in order to estimate the amount of power a wind turbine would produce over time at that location. The wind power density, or power per unit area, depends on the air density ρ and the wind speed v according to the following equation:

$$P/A = (1/2)\rho v^3$$

The air density depends on pressure and temperature according to the ideal gas law. Air pressure varies with elevation above sea level, but does not vary significantly in time. Temperature does vary significantly with time at many locations, but because wind power density varies with the cube of the wind speed, the most significant factor in the variation of wind power with time is the fluctuating wind speed.

Driven by complex large-scale atmospheric circulation patterns and local topographic influences, wind speeds at a particular location typically exhibit seasonal, daily, and short-term fluctuations. Short-term fluctuations (on the order of minutes or seconds) can affect a wind turbine's structural loads and power quality, but they do not significantly affect long-term production or economics. The seasonal and daily patterns of wind speed are more difficult to predict than are those of solar radiation, because the wind speed is not driven by any well-defined phenomenon comparable to the solar extra-terrestrial radiation. Accurate determination of seasonal and daily patterns typically requires anemometer measurements spanning months or years.

A defining characteristic of a wind regime is its long-term distribution of wind speeds. Long-term wind speed observations typically conform well to the two-parameter Weibull distribution, the probability distribution function of which is given by

$$f(v) = \frac{k}{c}\left(\frac{v}{c}\right)^{k-1} \exp\left[-\left(\frac{v}{c}\right)^{k}\right]$$

where v is the wind speed, c is the scale parameter (in the same units as v), and k is the dimensionless shape factor. The following equation relates the two parameters c and k to the mean wind speed:

$$\bar{v} = c\Gamma\left(\frac{1}{k} + 1\right)$$

where Γ is the gamma function. The Rayleigh distribution is a special case of the Weibull distribution, with $k = 2$. The Weibull distribution provides a convenient way to characterize the wind resource, because a mean wind speed and a Weibull k factor are sufficient to quantify the energy potential. Figure 1 shows the Weibull probability distribution function for a mean wind speed of 6 m/s and three values of the shape factor k. Lower values of k correspond to broader distributions. Weibull distributions fitted to measured wind data typically have k values between 1.0 and 3.0. Stevens and Smulders have determined several methods for fitting a Weibull distribution to measured data (see Fig. 1).

The wind speed tends to increase with height above ground, an important factor when using wind measurements taken at one height to estimate the production of a wind turbine at a different height. Modelers typically assume that the wind speed's increase with height follows either a logarithmic profile or a power law profile. The logarithmic profile is given by

$$\frac{v(z_1)}{v(z_2)} = \frac{\ln(z_1/z_0)}{\ln(z_2/z_0)}$$

where $v(z_1)$ and $v(z_2)$ are the wind speeds at heights z_1 and z_2, and the parameter z_0 is the surface roughness length. The power law profile is given by

$$\frac{v(z_1)}{v(z_2)} = \left(\frac{z_1}{z_2}\right)^{\alpha}$$

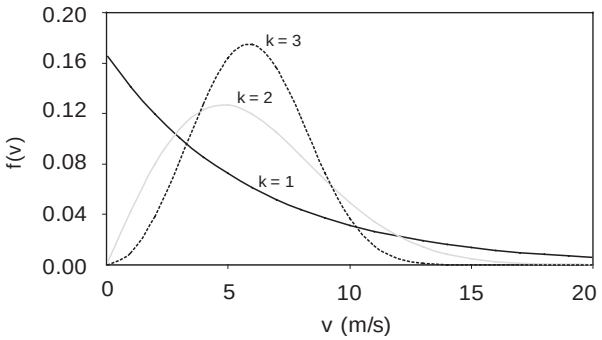


FIGURE 1 The Weibull distribution with a mean wind speed of 6 m/s and three values of k.

where α is the power law exponent. Manwell, McGowan, and Rogers have provided guidance on the selection of surface roughness length and the power law exponent, as well as a more complete examination of wind resource modeling.

2.3 Hydro Resource

Hydropower systems convert the energy of flowing water into electricity, usually by diverting a portion of a watercourse through a penstock (also called a pressure pipe) to a turbine. Hydro resource modeling aims to estimate the amount of power that a hydropower system could extract from a particular watercourse.

The ability of falling water to produce power depends on the flow rate and the head (the vertical distance through which it falls) according to the following equation:

$$P = \rho g h \dot{Q}$$

where ρ is the density of water, g is the gravitational acceleration, h is the head, and $\dot{Q}$ is the flow rate. Therefore the head (which is determined by local topography) and the available flow rate define the hydro resource at any particular location. The flow rate often varies seasonally and may in some cases vary according to the time of day, although it generally does not display the intermittency inherent to the solar and wind resources. Modelers sometimes characterize a watercourse's flow rate by computing its flow duration curve, an example of which appears in Fig. 2. This example shows that the flow rate exceeds 5 m^3/s about 37% of the time, and exceeds 10 m^3/s about 13% of the time. Flow duration curves can be created using long-term flow measurements or synthesized based on precipitation data within the catchment basin. Because most small hydro systems do not incorporate storage, the critical value is the firm flow, usually defined as the flow rate available 95 or 100% of the time. Designers typically size hydro systems based on the firm flow, or some portion thereof. The Web site www.microhydropower.net contains information on many aspects of small hydro systems, including guidance on the measurement of head and flow.

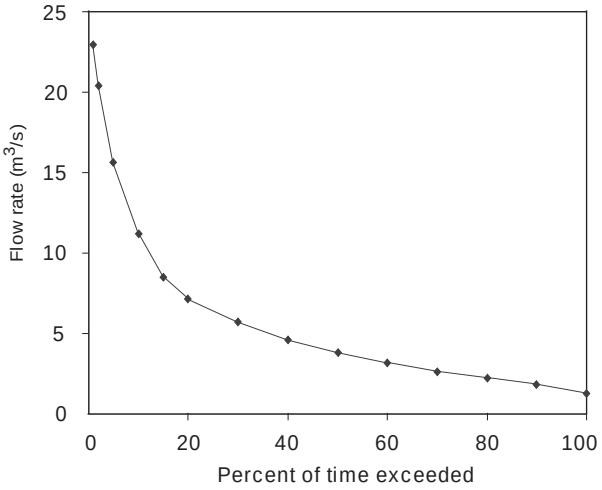


FIGURE 2 Sample flow duration curve.

2.4 Biomass Resource

Biomass power systems produce electricity from the chemical energy contained in organic matter. Many different types of biomass are suitable for power production, including wood waste, agricultural residue, animal waste, and energy crops. The goal of biomass resource modeling is to quantify the availability, price, and physical properties of the feedstock.

Three factors distinguish the modeling of the biomass resource from that of the other renewable resources. First, the amount of available biomass depends largely on human action, because the feedstock must be harvested, transported, and, in the case of dedicated energy crops, even sowed. So, although the gross amount of biomass depends on natural factors such as growing season and rainfall, the intermittency and unpredictability that characterize the other renewable resources do not apply to the biomass resource. Second, the biomass feedstock is often easily stored, further reducing intermittency and increasing flexibility. Third, the feedstock is often converted to a liquid or gaseous fuel, which can be stored and transported. As a result of these factors, biomass resource modeling typically focuses not on availability, but rather on the properties of the feedstock (such as moisture content) and the costs of delivering it in a usable form. McKendry has provided a thorough overview of the biomass resource.

3. COMPONENT MODELING

The characteristics of renewable technologies relevant to their modeling are briefly outlined here. Because renewable power systems often include some form of backup power to compensate for the intermittent nature of the renewable resources,

fossil-fueled generators, energy storage devices, and grid connection are also covered here.

3.1 Photovoltaic Modules

Photovoltaic (PV) modules convert solar radiation directly to direct current (DC) electricity, with sizes ranging from a few watts to hundreds of kilowatts. The output current of a photovoltaic module increases linearly with increasing incident global radiation, and decreases linearly with increasing module temperature. At any fixed temperature and radiation value, current (and hence power) varies nonlinearly with voltage, as shown in Fig. 3.

The peak of the power curve in Fig. 3 is called the maximum power point. The voltage corresponding to the maximum power point varies with radiation and temperature. If the photovoltaic module is directly connected to a DC load (or a battery bank), then the voltage it experiences will often not correspond to the maximum power point, and performance will suffer. A maximum power point tracker is a solid-state device that decouples the module from the load and controls the module voltage so that it always corresponds to the maximum power point, thereby improving performance. The complexity of modeling the power output from photovoltaic modules depends on the presence of the maximum power point tracker. In the absence of a maximum power point tracker, the modeler must account for the effects of radiation, temperature, and voltage. Duffie and Beckman have provided algorithms to do so. A maximum power point tracker eliminates the nonlinear dependence on voltage and allows the modeler to use the following equation for module efficiency:

$$\eta = \eta_{\text{ref}} + \mu(T_{\text{c}} - T_{\text{ref}})$$

where T_{c} is the module temperature, T_{ref} is the reference temperature, η_{ref} is the module efficiency at the reference temperature, and μ is the efficiency temperature coefficient. The efficiency temperature coefficient is a property of the module and is typically a negative number. From the efficiency, a modeler can calculate the power output of a photovoltaic module using the following equation:

$$P = \eta G_{\text{T}} A$$

where η is the efficiency, G_{T} is the global (total) solar radiation incident on the surface of the module, and A is the area of the module.

The life-cycle cost of a photovoltaic subsystem is dominated by initial capital cost. The lack of moving parts (with the rare exception of tracking systems) results in long lifetimes and near-zero operation and maintenance costs. Economic models of photovoltaic modules commonly assume 20- to 30-year lifetimes and neglect operation and maintenance costs altogether.

3.2 Wind Turbines

Wind turbines generate alternating or direct current (AC or DC) power, in sizes ranging from a few watts to several megawatts. Each wind turbine has a characteristic power curve that describes its power output as a function of the wind speed at its hub height. Figure 4 shows an example. Power curves typically display a cut-in wind speed below which the output is zero, then a section wherein the output varies roughly with the cube of the wind speed, then a section wherein the output peaks and declines due

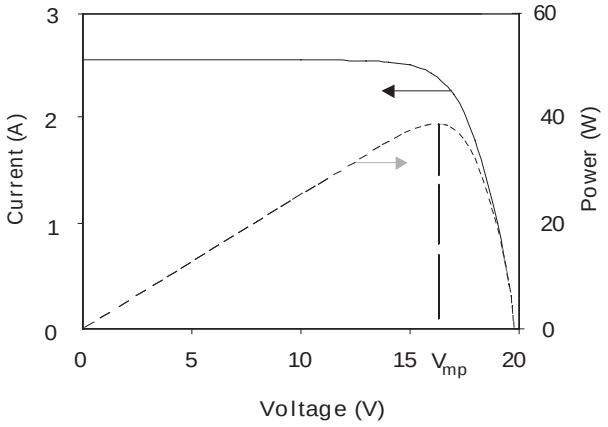


FIGURE 3 Current and power versus voltage for a typical photovoltaic module.

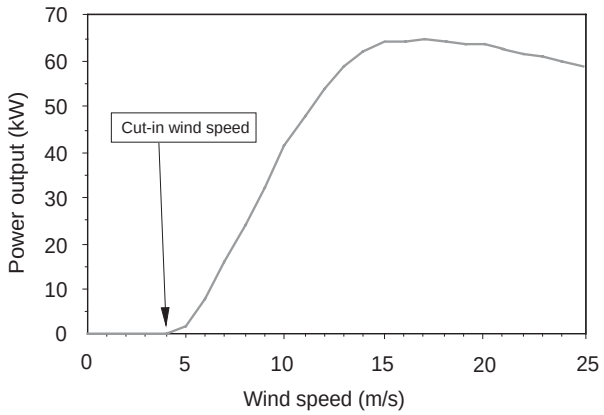


FIGURE 4 A wind turbine power curve.

to limitations designed into the turbine. The power curve describes turbine performance at a reference air density, typically corresponding to standard temperature and pressure. For a turbine operating at a different air density, the modeler can calculate its power output using the following equation:

$$P = P_{pc}(\rho/\rho_{ref})$$

where P_{pc} is the power output predicted by the power curve, ρ is the actual air density, and ρ_{ref} is the reference air density. Because the wind turbine output is so nonlinear with wind speed, and the wind speed is so variable, a modeler must exercise care when estimating the output of a wind turbine over long periods. The direct use of annual, monthly, or even daily average wind speeds causes unacceptably large inaccuracies. Time series data can be used directly only if its averaging interval is 1 hour or less. For longer term averages, the modeler must consider the frequency distribution of wind speeds.

Initial capital costs (including installation) dominate the life-cycle cost of wind turbines, but maintenance costs do occur. Economic models of wind turbines often assume a 15- to 25-year lifetime and an annual operation and maintenance cost of 1–3% of capital cost.

3.3 Small Hydropower Systems

Small hydropower systems generate AC or DC electricity from flowing water. The following equation gives the output power of a hydropower system:

$$P = \eta_{turb}\eta_{gen}\rho g h_{eff}\dot{Q}$$

where P is the output power in watts, η_{turb} is the efficiency with which the hydro turbine converts water power into shaft power, η_{gen} is the efficiency with which the generator converts shaft power into electrical power, ρ is the density of water in kg/m^3, g is the gravitational acceleration in m/s^2, h_{eff} is the effective head (the available head minus the head loss due to turbulence and friction) in meters, and $\dot{Q}$ is the volumetric flow rate through the turbine in m^3/s.

The turbine efficiency depends on the type of turbine and the flow rate. Hydro turbines fall into two broad categories, impulse turbines and reaction turbines. In an impulse turbine, water flows through a nozzle and emerges as a high-velocity jet, which strikes the turbine blades and causes them to rotate. In a reaction turbine, the flowing water surrounds the turbine blades, which are shaped such that the flowing water creates lift forces, causing rotation. Most small hydro systems use impulse turbines such as the Pelton, Turgo, or crossflow, but some low-head applications use reaction turbines such as the Francis, propeller, or Kaplan. Figure 5 shows typical efficiency curves. Although the impulse turbines exhibit near-constant efficiencies throughout their operating range, the reaction turbines (excluding the Kaplan) are much less efficient at lower flow rates. The Kaplan turbine, an adjustable-blade propeller turbine, exhibits an efficiency curve similar to that of the impulse turbines (see Fig. 5).

Friction and turbulence in the penstock cause a reduction in the water pressure upstream of the turbine. This pressure loss can be expressed as a head loss. The effective head can then be calculated by subtracting this head loss from the available head (the vertical drop from the penstock inlet to the turbine). Modelers can calculate the head loss using the standard techniques of fluid mechanics (Darcy friction factor, Moody diagram) or estimate it using rules of thumb.

Small hydro systems rarely incorporate storage reservoirs, because the economic and ecological costs of reservoirs often exceed their value in compensating for variation in the flow rate and adding a measure of control. Most run-of-river systems (ones that do not incorporate storage) have no means of restricting the flow rate though the turbine, so the flow rate in the preceding equation is simply equal to the flow through the penstock, which cannot exceed the stream flow.

Like photovoltaic modules and wind turbines, small hydro systems have high initial costs and low operating costs. The capital cost of a hydro system, which includes the cost of the required civil works plus that of the turbine and related machinery,

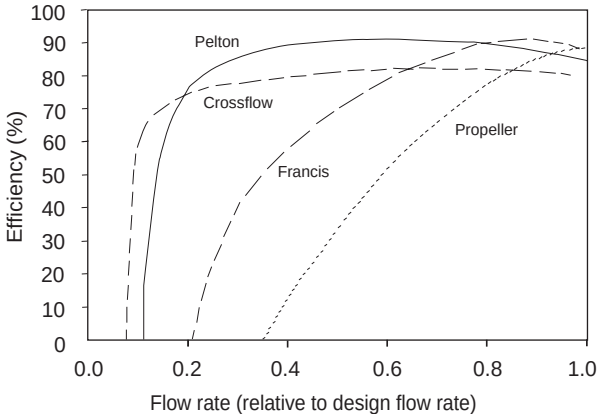


FIGURE 5 Efficiencies curves of common small hydro turbines. The Pelton and crossflow are impulse turbines; the Francis and propeller are reaction turbines. From Paish (2002), with permission.

depends very strongly on the local topography. Sites with high head and low flow generally have more favorable economics than do those with low head and high flow, because the required size of the machinery depends most strongly on the flow rate. Hydro systems are very long-lived, and lifetimes of 30 to 50 years are not uncommon.

3.4 Biomass Power Systems

The conversion of biomass into electricity typically makes use of either a thermochemical process (gasification, pyrolysis, or direct combustion), or a biochemical process (fermentation or anaerobic digestion). With the exception of direct combustion, each of these processes results in a gaseous or liquid fuel that can be used like any conventional fuel (sometimes in combination with a conventional fuel) in an internal combustion engine, microturbine, or fuel cell. In a direct combustion process, heat generated by burning biomass feedstock drives a thermal cycle. From a modeling perspective, these biomass conversion technologies are more akin to conventional forms of electrical production than to the other renewable power technologies. In particular, their ability to produce power on demand and their relatively high operating costs distinguish them from solar, wind, and hydropower technologies. The differences between biomass power and conventional power lie in technical issues (such as fuel handling and combustion characteristics) and the possibly seasonal availability of the biomass-derived fuel.

3.5 Fuel-Fired Generators

The diesel generator, an electrical generator driven by a reciprocating diesel engine, is the most common source of electric power in remote locations. Sizes range from a few kilowatts to 1 MW. Gasoline- and propane-fueled generators are also common below about 25 kW. Microturbines have become commercially available in recent years in the 25- to 250-kW size range, and fuel cells may become widespread in years to come.

Their ability to produce power on demand and their fast response to a fluctuating electric load make conventional generators valuable in stand-alone systems. When used in combination with renewable power sources, a generator can provide backup power during times of insufficient renewable output. For constant-speed reciprocating engine generators as well as microturbines, the rate of fuel consumption is well approximated by the following first-order equation:

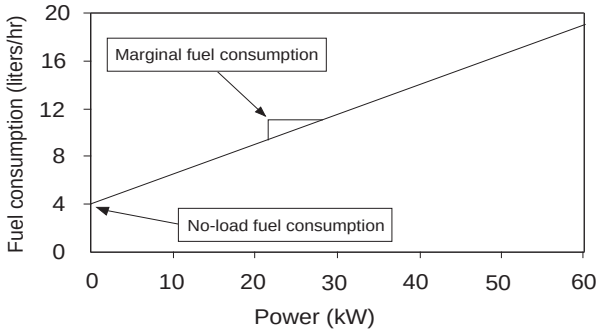


FIGURE 6 Typical generator fuel consumption curve.

$$F = F_0 + F_1 \cdot P$$

where F is the rate of fuel consumption, F_0 and F_1 are constants (the no-load fuel consumption and the marginal fuel consumption, respectively), and P is the output power of the generator. This relation is shown in Fig. 6. Skarstein and Uhlen found that, for diesel generators, the value of F_0 scales linearly with generator size. The fuel consumption of variable-speed reciprocating engine generators is better approximated with a higher order function, although they still exhibit a no-load fuel consumption.

Most models assume that a generator will need replacement (or a major overhaul) after a certain number of operating hours. To avoid engine damage, constant-speed diesel generators typically do not operate below about 30% of their rated capacity, and most other generator types have similar restrictions. Compared to renewable power sources, conventional generators have a low capital cost per kilowatt capacity, and a high operating cost per kilowatt-hour of electricity produced. Importantly, the operating cost per kilowatt-hour decreases with increasing output power. This operating cost has two components: maintenance and fuel. The maintenance cost per hour of operation is essentially independent of the output power; so on a per-kilowatt-hour basis it is inversely proportional to output power. Due to the aforementioned shape of the fuel consumption curve, the fuel cost per kilowatt-hour also decreases with output power. Generators are therefore most economically efficient when running at full capacity, a fact that has implications for power system design and operation.

3.6 Energy Storage Technologies

Energy storage devices can play an important role in stand-alone renewable power systems. In systems

without generators, they can provide backup power for intermittent renewable power sources. In systems with generators, they can save fuel and help avoid inefficient generator operation by serving the load during times of low electric demand, where the generator is least efficient. By buffering against the fluctuations in renewable power output, energy storage can also reduce the frequency of generator starts. Beyer, Degner, and Gabler found that in high-penetration wind/diesel systems (where the installed wind capacity exceeds the average load), even a very small amount of storage dramatically reduces the frequency of diesel starts.

Batteries, in particular lead-acid batteries, remain the predominant energy storage device. Many competing battery types (nickel–cadmium, nickel–metal hydride, lithium ion, sodium–sulfur, metal–air, flow batteries) surpass the lead-acid battery in one or more aspect of performance such as cycle life, round-trip efficiency, energy density, charge and discharge rate, cold-weather performance, or required amount of maintenance. In most applications, however, their lower cost per kilowatt-hour of capacity makes lead-acid batteries the optimal choice. Alternatives such as flywheels, ultracapacitors, or hydrogen storage may become commercially successful in the future, but are rare presently. In this section, the focus is exclusively on batteries.

For technoeconomic modeling, the three most important aspects of battery performance are efficiency (or losses), charge and discharge capacity, and lifetime. The simplest way to model the efficiency is to assume a constant energetic efficiency, meaning that some fixed percentage of the energy absorbed by the battery can later be retrieved. A more accurate alternative is to model the battery as a voltage source in series with a resistance. This "internal resistance" causes a difference in voltage between charging and discharging, and losses can be calculated either based on that voltage difference (charge is conserved, so 1 amp-hour of energy enters the battery at a high voltage and leaves at a lower voltage) or, equivalently, by calculating the I^2R losses directly.

The charge and discharge capacities of a battery are a measure of how quickly the battery can absorb or release energy. The simplest battery models assume unlimited charge and discharge capacities. More sophisticated models account for the fact that both the battery's state of charge and its recent charge and discharge history affect its charge and discharge capacities. The kinetic battery model, proposed by Manwell and McGowan and illustrated in Fig. 7, does this by modeling the battery as a two-tank system. The energy in one tank is available for immediate use, but the energy in the other is chemically bound. A conductance separates the two tanks and limits the rate that energy can move from one to the other. Some models also account for self-discharge that occurs within batteries, although the rate of self-discharge is typically small compared to the rates of discharge typically experienced by batteries in power systems. Regardless of the method chosen to calculate charge and discharge capacity, performance models should account for the fact that most batteries do not tolerate being discharged below a certain critical state of charge. For lead-acid batteries, this minimum state of charge ranges from about 30 or 40% for deep-discharge varieties to as much as 80% for common car batteries.

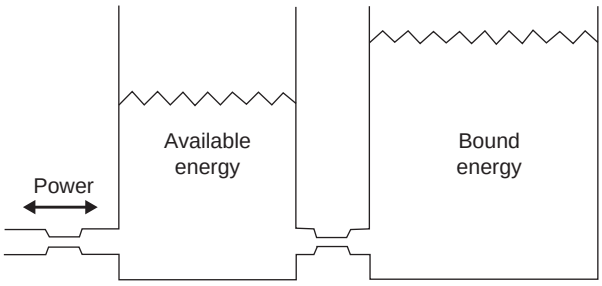


FIGURE 7 Kinetic battery model.

Spiers and Rasinkoski proposed three types of factors limiting the lifetime of a battery. Primary factors include the cycle life (governed by the amount of energy cycled through the battery) and the float life (governed by internal corrosion). Secondary factors include effects such as sulfation of lead-acid batteries caused by extended periods at low states of charge. Catastrophic factors include manufacturing defects, misuse, and freezing. Because secondary factors normally have little effect and catastrophic failures can be prevented, simple battery lifetime models typically account only for the primary factors. Spiers and Rasinkoski suggest calculating the float life and the cycle life, and assuming that the battery lifetime will be equal to the lesser of the two. The float life can be either assumed constant or calculated based on the average ambient temperature. Spiers and Rasinkoski suggest a simple formula for this calculation. The cycle life can be modeled most simply by assuming that a fixed amount of energy can cycle through the battery before it needs replacement, a good assumption for most lead-acid batteries. Alternatively, one could determine the number of cycles of different depths experienced by

the battery per year, and then refer to the battery lifetime curve (cycles to failure versus depth of discharge) to calculate the resulting years of lifetime. Downing and Socie have proposed two algorithms for counting cycles. The effect of temperature on the storage capacity, charge/discharge capacity, and internal resistance may be significant for certain modeling applications. Perez has offered insight into this and other aspects of battery performance.

3.7 AC/DC Converters

Renewable power systems commonly comprise a mixture of AC and DC loads and components, and hence require AC/DC conversion. Solid-state inverters and rectifiers typically perform this conversion, although some systems use rotary converters. For technoeconomic modeling, the important performance characteristics of a conversion device are its maximum capacity and its efficiency. Another characteristic important for modeling an inverter is whether it can operate in parallel with an AC generator; those that cannot are sometimes called "switched inverters."

Sukamongkol *et al.* have proposed that the power flow through solid-state inverters may be modeled by the following equation:

$$P_O = \alpha(P_I - \beta)$$

where P_O is the output power, P_I is the input power, and α and β are constants. The value of β, often called the "standing losses," is small compared to the capacity of the inverter. Many models therefore ignore standing losses and assume a constant inverter efficiency. The same treatment applies to rectifiers. Solid-state inverters typically exhibit efficiencies near or above 90%, with rectifier efficiencies slightly lower. Rotary converters typically have slightly lower efficiencies, compared to solid-state converters.

3.8 Grid Connection

Distributed generation systems supply electricity to grid-connected energy consumers and either store the excess or sell it back to the utility. Such systems may also supply heat, in which case they are called cogeneration or combined heat and power (CHP) systems. The utility rate structure (or tariff structure) often determines the economic viability of distributed generation systems. This rate structure typically consists of an energy charge (on the number of kilowatt-hours purchased per billing period) and a demand charge (on the peak demand within the billing period). Both may vary according to season, time of day, quantity, or other factors. Some utilities allow the sale of excess electricity back to the grid, usually at the utility's avoided cost or according to a net metering scheme whereby the customer pays for net monthly or annual consumption. Utilities may also assess an interconnection charge (a one-time fee for connecting a power system to the grid) or standby charges (monthly or annual fees for providing backup power).

4. SYSTEM MODELING

Several characteristics of renewable power technologies combine to complicate the process of designing renewable power systems. The principal virtue of renewable power sources, that they do not require fossil fuel, makes them suitable for many remote and off-grid applications. Such "stand-alone" systems require careful design because of the need to balance electrical supply and demand. Models can help by simulating the performance of a system to determine its viability, robustness, and stability.

The intermittent and seasonal nature of renewable resources often makes it necessary (or at least advantageous) to combine renewable power sources with each other, with conventional power sources such as diesel generators, and with energy storage technologies such as batteries. Hence renewable power systems frequently comprise several dissimilar components, often both AC and DC. Models can help both in simulating the complex behavior of a composite system and in evaluating the many different possible combinations of component types and sizes.

Renewable power sources typically have high capital costs and low operating costs, in contrast to most conventional (fossil-fueled) power sources. Any fair comparison of the relative economics of conventional and renewable sources of energy must therefore consider operating costs in addition to initial capital costs. This requires more sophisticated economic modeling than may be necessary when simply choosing between conventional power sources.

The design of renewable power systems involves many variables, some inherently uncertain. Computer models can help the system designer take this uncertainty into consideration. A model may, for example, allow a designer to assess the technical and economic implications of an increase in fuel price,

growth in electric demand, or an inaccurate estimate of wind speed.

4.1 Temporal Variability

Renewable power systems experience two kinds of temporal variability. Seasonal or long-term variability results from the cyclic annual pattern of the renewable resources. Short-term variability, which applies in particular to systems comprising solar or wind power, results from the hour-to-hour, even second-to-second, intermittency in the solar radiation, wind speed, and electric load. Both types of variability influence power system design.

The seasonal variability of resources often makes it impossible to rely on a single renewable power source year-round without backup power from a fuel-fired generator or utility grid. If stream flow declines sharply during the winter, for example, a diesel generator may be needed to supplement the output of the hydro turbine for part of the year. The degree to which different resources complement each other also affects the optimal combination of power sources. For example, a combination of wind and solar power will more likely be optimal if wind speeds peak during the winter (so that increased wind power output makes up for decreased solar power output) than if they peak during the summer.

Short-term variability has wide-ranging implications for system design and operation. It is the principal factor in determining the optimal amount of energy storage, and an important consideration when sizing dump loads (which manage excess renewable power) and AC/DC converters. Short-term variability also plays a role in the control of backup generators, which must modulate their output in response to fluctuations in the renewable output and electric demand.

4.2 Component Selection and Sizing

A power system designer must choose which power sources the system should comprise, decide whether to include energy storage, and select an appropriate size for each system component. This process of selecting the best system configuration from a long list of possibilities is an optimization problem well suited to computer modeling. The objective is typically to minimize the cost of energy or the total system cost, subject to any number of constraints. For stand-alone systems, the principal constraint is normally that the system must supply some fraction of the annual electric demand, often 100%. Other constraints may limit such parameters as fuel usage, emissions, or generator run time.

The optimal system design varies widely, depending on the magnitude and character of the electric demand, the quality of the renewable resources, and several economic factors. Relevant economic factors may include component costs, fuel price, interest rate, and, in the case of grid-connected systems, the utility rate structure. Models allow the designer to explore the many different possible system configurations in search of the optimum. Some models perform the optimization procedure by simulating multiple different system configurations and comparing them based on user-specified criteria. Others simply simulate one system at a time, in which case the designer can perform the optimization process manually.

4.3 System Operation and Control

For some renewable power systems, the designer must also consider the method of operation. A small photovoltaic/battery or wind/battery system needs no complicated control logic, other than appropriate constraints on battery charging and discharging. The system simply charges the battery with excess renewable power and discharges the battery when necessary. Similarly straightforward control logic would apply to a grid-connected photovoltaic system or a low-penetration wind/diesel system in which the diesel always operates and the wind power simply saves fuel. But a high-penetration wind/diesel system, in which the wind power may occasionally exceed the electric load, needs control logic to determine if and when the diesel can safely shut down without risking a loss of load. Spinning reserve is an important consideration in such decision-making. The manner in which the controller makes dispatch decisions can affect both the economics and the optimal design of the power system.

Control strategies are also important for any system comprising more than one dispatchable power source. A stand-alone power system that includes both a generator and a battery bank is one example. Two common control strategies for such a system are load following and cycle charging. Under the load-following strategy, whenever the generator operates, it produces only enough power to serve the load (surplus renewable power charges the batteries). Under the cycle-charging strategy, whenever the generator operates, it produces excess power to charge the battery bank. Barley and Winn have presented a thorough analysis of these two strategies

and their variations, and proposed a design chart to help select the better of the two, depending on the configuration and economics of the system.

Other examples of systems comprising more than one dispatchable power source include a stand-alone system containing multiple diesel generators, or a grid-connected biomass-fueled generator. In either case, the system must depend on some control strategy to decide how to operate various power sources that may differ in cost, fuel source, efficiency, thermal output, noise, or emissions. Modeling can help the system designer explore the consequences of different control strategies.

4.4 Economic Modeling

Economics frequently figure prominently in system design. Even when nonmonetary factors are involved, economic figures of merit can help in selecting among design options. To account properly for both initial capital costs and recurring operating costs, economic models should consider the time value of money. One approach is to calculate the present value of all cash flows (annual fuel payments, maintenance costs, grid power costs, component replacement costs, etc.) that occur within the project lifetime. This yields the system's total life-cycle cost. An equivalent approach is to annualize all one-time costs and add them to the recurring annual costs to calculate the system's total annualized cost. This number divided by the total energy delivered annually to the load yields the levelized cost of energy. For system retrofits, when an investment in renewable energy results in future cost savings, the payback period and internal rate of return are also useful figures of merit.

When calculating life-cycle cost it is helpful to make use of two economic concepts, the present value of a future cash flow and the present value of a sequence of equal annual cash flows. The present value of a cash flow occurring N years in the future is found by multiplying by the present worth factor (PWF), given by

$$PWF = 1/(1+i)^N$$

where i is the annual discount rate. Inflation can be factored out by using the real annual discount rate, given by

$$i = (i' - f)/(1+f)$$

where i' is the nominal interest rate and f is the inflation rate.

The present value of a sequence of equal annual cash flows is equal to the annual cash flow divided by the capital recovery factor (CRF). This factor is given by

$$CRF = i(1+i)^N/[(1+i)^N - 1]$$

where N is the number of years over which the annual cash flows occur. The capital recovery factor is also used to calculate the equivalent annualized value of a one-time cash flow (for a more complete discussion of these factors and economic modeling in general, see DeGarmo *et al.* or any introductory economics text).

4.5 Sensitivity Analysis

Regardless of how carefully a designer models the cost and performance of a power system, uncertainty in key inputs invariably causes actual and modeled results to differ. In a wind/diesel system for example, if the actual average wind speed turns out to be lower than the estimate used during modeling, the system will likely consume more fuel than predicted, hence the life-cycle cost will be greater than predicted. Similarly, an underestimate of the diesel fuel price over the project lifetime will also result in an underestimate of life-cycle cost.

A modeler can take this uncertainty into consideration by performing multiple simulations under different input assumptions. For example, the modeler could analyze the system cost and performance using several different values of fuel price to establish the sensitivity of key outputs to that input. The results of such a sensitivity analysis are often plotted as a spider graph such as the one shown in Fig. 8. This example shows how the cost of energy varies with changes in particular input variables. The modeler calculated the cost of energy assuming five different values of fuel price: a best estimate, plus two values above and two values below that best estimate. The modeler did the same for two more input variables, the average wind speed and the average solar radiation. The graph shows that the cost of energy increases with increasing fuel price, and decreases with increasing wind speed or solar radiation. Changes in wind speed have a more dramatic effect on the cost of energy than do changes in solar radiation.

For a more formal treatment of uncertainty, a modeler can make use of the techniques of decision analysis, a field devoted to decision-making in the presence of uncertainty. The canonical decision analysis approach is to assign probabilities to multiple

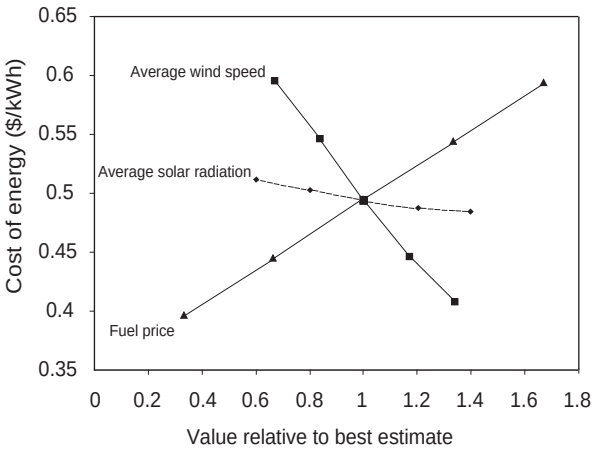


FIGURE 8 Spider graph showing results of a sensitivity analysis for a small photovoltaic/wind/diesel system.

input scenarios, simulate each scenario, and then calculate the resulting probability-weighted average value of any output variable. This "expected value" represents the best estimate of that output variable given the uncertainty in the inputs. When applied to the power system design problem, decision analysis can help select a system that is optimal not just for a fixed (and often dubious) set of input assumptions, but rather with consideration of a wide range of possible scenarios and their attendant risks. None of the models presented in the next section has a built-in decision analysis capability, but any one of them could help inform the decision analysis process by simulating input scenarios.

5. MODELS FOR RENEWABLE POWER SYSTEM DESIGN

Many available computer models simulate the performance and economics of renewable power systems. Each model uses representative load and resource data to predict the performance of the system, typically over a 1-year period. Performance parameters of interest may include the amount of energy produced by a renewable power source, the amount of fuel consumed by a generator, the amount of energy cycled through the batteries, or the amount of energy purchased from the grid. It is common to extrapolate these performance data over many years to calculate economic figures of merit such as life-cycle cost or cost of energy.

Models can be categorized as statistical models or time series models according to the technique used to predict system performance under highly variable conditions of load and renewable resources. Statistical models consider large time intervals (monthly or annual) and rely on statistical techniques to predict the outcome of the short-term variability occurring within each time interval. Time series models approach the issue of variability by dividing the year into time steps of 1 hour or less and simulating the behavior of the system at each time step.

5.1 Statistical Models

Statistical models usually account for seasonal variability by simulating system performance separately for each month of the year. But to account for short-term variability, they depend on some form of statistical manipulation. For example, a common technique when simulating a stand-alone PV/battery system is to calculate the loss-of-load probability each month. This is the probability that the system will experience a supply shortage at some time during the month, a scenario that may occur after the load exceeds the PV output long enough to drain the battery bank completely. A model can calculate the loss-of-load probability using information about the variability of the solar resource and its correlation with electric load. If the loss-of-load probability exceeds some user-specified critical value, the system is unacceptable.

Compared to time series models, statistical models tend to be simpler to use and quicker to provide results. Their simplicity stems partly from their limited input requirements; they require monthly or annual average load and resource data, rather than the hourly data required by most time series models. They also tend to simulate system performance and cost using less detailed algorithms that require fewer, more easily obtainable inputs. This input simplicity makes statistical models relatively easy to learn. It also means that a designer can quickly assemble the necessary load, resource, and component data to analyze a particular system. A further advantage of statistical models over time series models is their generally faster computation speed; most produce outputs virtually instantaneously. The combination of faster assembling of inputs and faster production of outputs makes statistical models ideal for quick system comparisons and cost estimates.

The simplicity of statistical models, however, comes at the cost of reduced accuracy. Because they rely on simplifying assumptions, statistical models sometimes overlook important aspects of real system behavior. For example, when modeling a system

comprising diesel generators, a statistical model may assume that the generator's fuel consumption is linear with power output, meaning the generator's efficiency is constant. Although greatly simplifying system modeling, this assumption neglects the fact that generators are much less efficient at low loads, a fact that has important implications for the control and economics of the system. Their reduced accuracy means that statistical models often cannot reliably model the behavior of complex systems such as high-penetration wind/diesel systems. In such systems, details for which a simplified model cannot account may have a substantial impact on both performance and economics.

Statistical models also tend to allow less flexibility in the configuration of the system. Because of their inability to simulate complicated systems, for example, statistical models typically allow a designer to simulate at most one renewable power source and one dispatchable generator. They typically cannot analyze systems comprising multiple renewable power sources, multiple generators, or other advanced features such as cogeneration. An additional drawback of statistical models is that despite using fewer and generally simpler inputs, a small number of their inputs can be difficult to understand and to estimate. A statistical model may, for example, require the user to specify the degree to which the electric load correlates with the solar radiation, or the fraction of the wind turbine output that will have to be dumped as excess. Time series models require neither of these inputs; the solar and load correlation is unnecessary because the model has access to the hourly solar and load data, and a time series simulation can determine the wind turbine's excess energy fraction.

In summary, statistical models offer simplicity and speed at the cost of accuracy and flexibility. They are suitable for prefeasibility analyses of many types of renewable power systems. In fact, certain uncomplicated systems (such as PV/battery or grid-connected PV) may require no further modeling. However, in the design of more complex systems, particularly those requiring sophisticated control strategies or careful balancing of electrical supply and demand, statistical modeling should be supplemented with time series modeling.

Several statistical models for renewable power system design are currently available and may be accessed on the Internet. PVSYST, created by the University of Geneva, can simulate PV/battery and grid-connected PV systems. Orion Energy's NSol! can simulate these systems as well as PV/generator/battery systems. RETScreen, created by Natural Resources Canada, can simulate these systems as well as grid-connected wind, low-penetration wind/diesel, and small hydro systems.

5.2 Time Series Models

The premise of time series modeling is that by simulating system performance at a sufficiently fine temporal resolution, the effects of short-term variability can be modeled directly. Time series models typically perform annual simulations using a 1-hour time step, and assume that the time-dependent variables (electric load and renewable resources) remain constant within that time step. Shorter time steps lead to improved accuracy but increased computation time. That, along with the difficulty of obtaining higher resolution resource and load data, explains why most models use a 1-hour time step.

At the core of time series modeling is the energy balance performed each time step. In this energy balance, the model calculates the available renewable power, compares it to the electric load, and simulates how the system would deal with any surplus or deficit. In the case of insufficient renewable power, the model may draw power from the grid, dispatch one or more generators, or discharge the battery bank to avoid unmet load. In the case of an oversupply of renewable power, the model may sell power to the grid, charge the battery bank, serve an optional load, or dissipate the excess power in a dump load. When making such dispatch decisions, time series models typically refer to some preset operating strategy, and many models allow the comparison of several different strategies so that the user can determine which is optimal for a particular application.

The division of the year into many short intervals allows time series models to simulate even complex systems accurately. As a result, time series models allow the modeler to simulate a wide variety of system types, sometimes comprising multiple renewable power sources, multiple generators, battery storage, and grid connection. With this flexibility, a modeler can evaluate many more design options than is possible with a statistical model.

Hourly simulations require hourly load and resource data. The limited availability of such data can hinder the use of time series models, although several algorithms exist for synthesizing hourly load, solar radiation, and wind speed data from more readily available monthly or annual average data. Virtually all time series models, for example, allow

the user to construct hourly load data from typical daily load profiles. Algorithms that synthesize hourly solar and wind data typically require monthly averages plus a few parameters. Some models integrate such algorithms directly into their interfaces, but external resource data generators can produce data suitable for any time series model.

The computation time required to simulate annual system performance with a time series model depends strongly on whether the model tracks the voltage of the DC components. This DC bus voltage can fluctuate widely over time, in contrast to the AC bus voltage, which is fixed by the system controller. Because of the feedback effect between the DC bus voltage and the performance of batteries, photovoltaic modules, and DC wind turbines, calculating the DC bus voltage requires an iterative solution procedure within each time step. Models that track the DC bus voltage therefore typically require many seconds or a few minutes of computation time per simulation. Models that simply track power flows without calculating voltage do not require this iteration, and so can simulate systems much faster.

Several time series models for renewable power system design are available and may be accessed on the Internet. PV-DesignPro, a product of Maui Solar Energy Software Corporation, can model stand-alone or grid-connected systems comprising PV modules, wind turbines, batteries, and a generator. PV*SOL, created by Valentin EnergieSoftware, can model stand-alone and grid-connected systems comprising PV modules, batteries, and a generator. PV*SOL uses a statistical modeling approach to perform a quick initial simulation, then a time series modeling approach for a more accurate final simulation. HOMER, a product of the National Renewable Energy Laboratory, can model systems comprising PV modules, wind turbines, small hydro, biomass power, batteries, hydrogen storage, grid connection, and as many as three generators, serving any combination of AC, DC, and thermal loads. HOMER also performs optimization and sensitivity analysis. Hybrid2, created by the National Renewable Energy Laboratory and the University of Massachusetts, can model stand-alone systems comprising PV modules, wind turbines, batteries, and as many as eight generators, with a time step as short as 1 minute. Hybrid2 also models a wide variety of different system control and load management strategies. Though it is a time series model, Hybrid2 incorporates some aspects of statistical modeling to account for the short-term variability occurring within each time step.

6. CONCLUSIONS

In the design of renewable power systems, significant complexity arises due to the intermittent nature of renewable power sources, the large number of possible system configurations, the uncertainty in key input parameters, and the need to compare different design options on the basis of life-cycle cost. Computer modeling can help the system designer to account for the many factors affecting the performance and economics of renewable power systems, to evaluate diverse design options, and to explore multiple input scenarios.

Various computer models are currently available for use in renewable power system design. Statistical models, offering speed and ease of use, are well suited to the type of prefeasibility analyses commonly performed early in the design process. Time series models, generally simulating renewable power systems of a wider variety and with a higher accuracy than statistical models, are suitable for more precise modeling later in the design process.

SEE ALSO THE FOLLOWING ARTICLES

Biomass for Renewable Energy and Fuels • National Energy Modeling Systems • Neural Network Modeling of Energy Systems • Renewable Energy and the City • Renewable Energy Policies and Barriers • Renewable Energy, Taxonomic Overview • System Dynamics and the Energy Industry

Further Reading

Barley, C. D., and Winn, C. B. (1996). Optimal dispatch strategy in remote hybrid power systems. *Solar Energy* **58,** 165–179.

Beyer, H. G., Degner, T., and Gabler, H. (1995). Operational behaviour of wind diesel systems incorporating short-term storage: an analysis via simulation calculations. *Solar Energy* **54,** 429–439.

Clemen, R. T. (1991). "Making Hard Decisions: An Introduction to Decision Analysis." PWS-Kent, Boston.

DeGarmo, E. P., Sullivan, W. G., and Bontadelli, J. A. (1988). "Engineering Economy." 8th ed. Macmillan, New York.

Downing, S. D., and Socie, D. F. (1982). Simple rainflow counting algorithms. *Int. J. Fatigue* **4,** 31–39.

Duffie, J. A., and Beckman, W. A. (1991). "Solar Engineering of Thermal Processes." 2nd ed. Wiley, New York.

Hunter, R., and Elliot, G. (1994). "Wind–Diesel Systems." Cambridge Univ. Press, Cambridge, Massachusetts.

Manwell, J. F., and McGowan, J. G. (1993). Lead acid battery storage model for hybrid energy systems. *Solar Energy* **50,** 399–405.

Manwell, J. F., McGowan, J. G., and Rogers, A. L. (2002). "Wind Energy Explained." Wiley, New York.

McKendry, P. (2002). Energy production from biomass (Part 1): Overview of biomass. *Bioresource Technol.* **83,** 37–46.

McKendry, P. (2002). Energy production from biomass (Part 2): Conversion technologies. *Bioresource Technol.* **83,** 47–54.

Mora-López, L. L., and Sidrach-de-Cardona, M. (1998). Multiplicative ARMA models to generate hourly series of global irradiation. *Solar Energy* **63,** 283–291.

Paish, O. (2002). Small hydro power: Technology and current status. *Renew. Sustain. Energy Rev.* **6,** 537–556.

Perez, R. (1993). Lead-acid battery state of charge vs. voltage. *Home Power* **36,** 66–69.

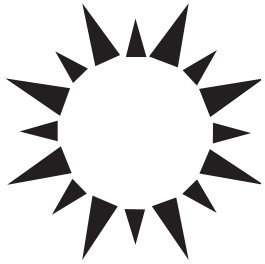
Skarstein, O., and Uhlen, K. (1989). Diesel considerations with respect to long-term diesel saving in wind/diesel plants. *Wind Engineer.* **13,** 72–87.

Spiers, D. J., and Rasinkoski, A. A. (1996). Limits to battery lifetime in photovoltaic applications. *Solar Energy* **58,** 147–154.

Stevens, M. J. M., and Smulders, P. T. (1979). The estimation of the parameters of the Weibull wind speed distribution for wind energy utilization purposes. *Wind Engineer.* **3,** 132–145.

Sukamongkol, Y., Chungpaibulpatana, S., and Ongsakul, W. (2002). A simulation model for predicting the performance of a solar photovoltaic system with alternating current loads. *Renewable Energy* **27,** 237–258.

Conservation Measures for Energy, History of



JOHN H. GIBBONS and HOLLY L. GWIN
Resource Strategies
The Plains, Virginia, United States

1. Introduction
2. Decade of Crisis: 1970s
3. Decade of No- and Low-Cost Options: 1980s
4. Decade of Globalization: 1990s
5. Decades Ahead

Glossary

end-use efficiency Substituting technological sophistication for energy consumption through (1) obtaining higher efficiency in energy production and utilization and (2) accommodating behavior to maximize personal welfare in response to changing prices of competing goods and services.

energy conservation The wise and thoughtful use of energy; changing technology and policy to reduce the demand for energy without corresponding reductions in living standards.

energy intensity The amount of energy consumed to produce a given economic product or service; often measured as the ratio of the energy consumption (E) of a society to its economic output (gross domestic product, GDP), measured in dollars of constant purchasing power (the E/GDP ratio).

energy services The "ends," or amenities, to which energy is the "means," e.g., space conditioning, lighting, transportation, communication, and industrial processes.

externalities The environmental, national security, human health, and other social costs of providing energy services.

least-cost strategy A strategy for providing individual and institutional energy consumers with all of the energy services, or amenities, they require or want at the least possible cost. It includes the internal economic costs of energy services (fuel, capital, and other operating costs) as well as external costs.

The history of energy conservation reflects the influence of technology and policy on moderating the growth of demand for energy. It traces the evolution of concepts of conservation from curtailment of energy use to least-cost strategies through end-use efficiency. Through understanding the past efforts to address the need for energy conservation, the future potential of the role for energy conservation in the global economy is better understood.

1. INTRODUCTION

Just as beauty is in the eye of the beholder, energy conservation means different things to different people. Concerned about deforestation in Pennsylvania, Benjamin Franklin took heed of his own advice ("a penny saved is a penny earned") by inventing the Franklin stove. In his *Account of the New Invented Fireplaces*, published in 1744, Franklin said "As therefore so much of the comfort and conveniency of our lives for great a part of the year depend on this article of fire; since fuel is become so expensive, and (as the country is more clear'd and settle'd) will of course grow scarcer and dearer; any new proposals for saving the wood, and for lessening the charge and augmenting the benefit of fire, by some particular method of making and managing it, may at least be thought worth consideration." The Franklin stove dramatically improved the efficiency of space heating. In Franklin's time, however, others viewed energy consumption as a signal of economic progress and activity. This view was epitomized by company letterheads picturing a factory with chimneys proudly billowing clouds of smoke!

Viewed as a matter of how much energy it takes to run an economy—often referred to as a society's energy intensity—energy conservation has been with us since the middle of the 20th century. Between 1949 and 2000, the amount of energy required to

 © 2004 Elsevier Inc. All rights reserved.

produce a (1996 constant) dollar of economic output fell 49%. Much of this improvement occurred by the combined effects of improved technology (including increased use of higher quality fuels in new technologies, such as high-efficiency gas turbines) and perceptions of higher energy costs after the oil shocks of the 1970s, when energy intensity began to fall at an average rate of 2% per year. (Fig. 1).

Although largely oblivious to long-term improvements in energy intensity, consumers undoubtedly felt the oil shocks, and came to equate energy conservation with waiting in long lines to purchase gasoline. For many people, even to the present day, energy conservation connotes curtailment or denial of the services energy provides, e.g., turning down the heat on a cold winter's day, turning off the lights, or driving lighter cars at slower speeds. This viewpoint came into sharp focus one night during the 1973 oil embargo, when one author (Gibbons), director of the newly established Federal Office of Energy Conservation, fielded a call from an irate Texan who proclaimed that "America didn't *conserve* its way to greatness; it *produced* its way to greatness!"

For purposes of this article, we accept and adapt the term "wise and thoughtful use" for energy conservation. This term implies that, when technically feasible and economically sound, a society will use technology and/or adopt social policy to reduce energy demand in providing energy services. It also implies that society's effort to bring energy supply and energy demand into balance will take into full account the total costs involved. Market prices for energy are generally less than its total social cost due to various production subsidies and nonmarket costs, or externalities, such as environmental impacts and security requirements. This approach has been labeled a "least-total cost" strategy for optimizing the investment trade-off between an increment of supply versus an increment of conservation.

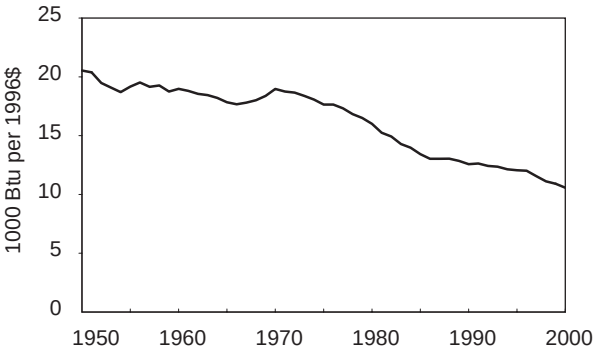


FIGURE 1 Energy use per dollar of gross domestic product. From the U.S. Energy Information Administration, *Annual Review of Energy 2000*.

2. DECADE OF CRISIS: 1970s

Events of the late 1960s highlighted many of the external costs of energy production and consumption. Many of the worsening problems of air and water pollution were directly attributable to energy use. The growth in the power of the Organization of the Petroleum Exporting Countries (OPEC) and increasing tensions in the Middle East contributed to a growing wariness of import dependence. High projected electric demand growth exacerbated worries not only about coal and the environment, but also about the sheer amount of capital investment that was implied. Gas curtailments began to show up along with spot shortages of heating oil and gasoline. And, surprising to many in the energy field, there were growing doubts about the nuclear option.

2.1 Rethinking Energy Demand

Those forces converged to trigger work on improved and alternative energy sources as well as on the dynamics and technologies of demand. Public support for analysis of conservation strategies came at first not from the "energy agencies" but from the National Science Foundation (NSF). In 1970, Oak Ridge National Laboratory (ORNL) researchers (who were well-sensitized to the vexing problems of coal and fission) seized on the opportunity provided by the NSF to investigate demand dynamics and technologies. The owner/operator of ORNL, the Atomic Energy Commission (AEC), was not keen on the idea but accepted the new direction for research as "work for others"!

It quickly became apparent that despite declining energy prices in previous years, energy efficiency had been slowly improving. Technology relevant to efficient use was advancing so impressively that more efficiency was economically attractive in virtually every sector of consumption. Further examination at ORNL and elsewhere showed that such improvements would accelerate more rapidly if energy prices were to rise and/or if consumers became better informed. For example, it was demonstrated that cost-effective technical improvements in refrigerators could double the energy efficiency of household refrigerators for a very modest price increment (Fig. 2). Clearly, major environmental benefits would also accrue from more conservation.

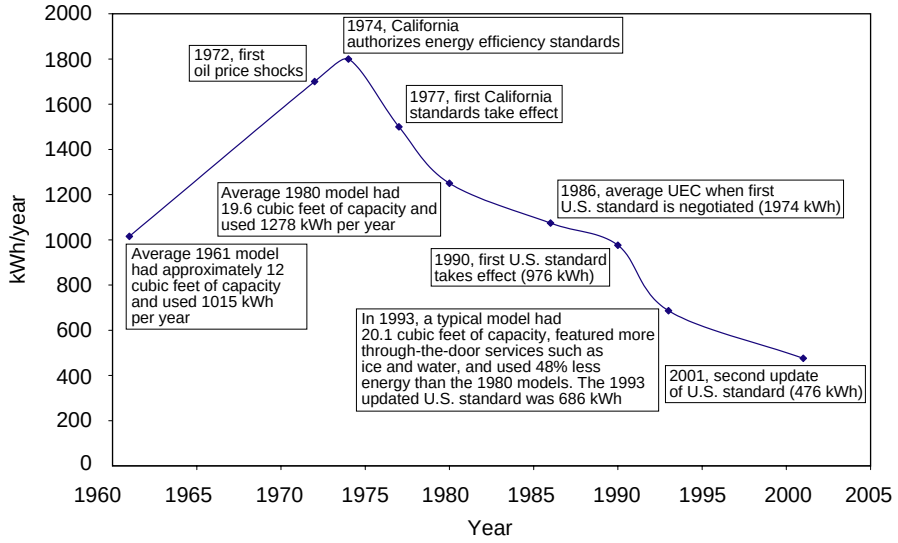


FIGURE 2 Changes in energy demand of refrigerators in the United States since 1961. UEC, unit energy consumption (kWh/year). From the Lawrence Berkeley National Laboratory.

The environmental concerns of the 1960s prompted the first serious efforts to project future energy demand. Scenario analyses, prepared by many different authors throughout the 1970s, produced a wide range of estimates, helping to shape a diverse range of views on the importance of conservation. Eleven forecasts of consumption made during the 1970s turned out to be remarkably consistent but ludicrously high, even for 1985 (ranging from 88 to 116 quads—the actual number was 75 quads). Prior to the oil embargo of 1973, most projections of U.S. energy consumption for the year 2000 fell in the range of 130–175 quads. The actual number turned out to be about 100 quads. Some of the difference can be attributed to lower than anticipated economic growth, but the dominant difference is due to the unexpected role of conservation.

Limits to Growth, published in 1972, was the first analysis to capture the public's attention. This computerized extrapolation of then-current consumption trends described a disastrous disjuncture of the growth curves of population and food supply and of resource use, including energy, and pollution. One scenario climaxed in a collapse mode, the sudden and drastic decline in the human population as a consequence of famine and environmental poisoning. This apocryphal report correctly made the point that exponential growth in finite systems is unsustainable, but its macroeconomic model failed because it did not consider responses to resource prices or advances in technologies.

A Time to Choose: America's Energy Future, a group of scenarios commissioned by the Ford Foundation and published in 1974, became notable, and controversial, for its "zero energy growth" (ZEG) scenario. This ZEG scenario clearly was meant to show that energy growth could technically be decoupled from economic growth. ZEG called for a leveling off in energy demand at about 100 quads per year shortly after 1985. In addition to those economically feasible measures taken by 1985, reduced demand would be effected by means of energy consumption taxes and other constraints on use. Taxes would increase the price of fuels so that substitution of energy-saving capital investments could be made economical while simultaneously encouraging shifts to less energy-intensive services. *A Time to Choose* also included a "technical fix" scenario that estimated 1985 consumption at 91 quads by coupling a microeconomic model with analysis of the engineering potential for conservation as a response to increased energy price and government policies. In 1976, Amory Lovins published a far more controversial and attention-grabbing scenario called "Energy Strategy: The Road Not Taken" in the journal *Foreign Affairs*. Lovins advocated a "soft energy path" based on three components: (1) greatly increased efficiency in the use of energy, (2) rapid deployment of "soft" technologies (diverse, decentralized, renewable energy sources), and (3) the transitional use of fossil fuels. His assertion that the "road less traveled" could bring America to greatly reduced energy demand (less than 50 quads per

year in 2025) is both venerated and despised to this day.

These private sector analyses competed with scenarios developed or commissioned by government agencies for the attention of energy policy makers throughout the 1970s. *Guidelines for Growth of the Electric Power Industry*, produced by the Federal Power Commission (FPC) in 1970, was a classic study of energy demand based largely on an extrapolation of past trends. The FPC had credibility because of its 1964 success in predicting 1970 electricity demand, but its forecast of a doubling of energy demand from 1970 to 1990 was dramatically far from the mark. The Atomic Energy Commission made more accurate estimates in *The Nation's Energy Future*, published in 1973, but it mistakenly forecast U.S. independence from imported oil by 1980—no doubt caused or at least influenced by President Nixon, who was determined to achieve energy independence before the end of his term.

In 1975, the U.S. National Academy of Sciences established a Committee on Nuclear and Alternative Energy Strategies (CONAES). The CONAES Panel on Demand and Conservation published a path-breaking paper in the journal *Science* in 1978: "U.S. Energy Demand: Some Low Energy Futures." The CONAES scenarios not only incorporated different assumptions for gross domestic product (GDP) and population growth, but also, for the first time, incorporated varied assumptions about energy prices. The lowest CONAES estimate, based on a 4% annual real price increase coupled with strong nonmarket conservation incentives and regulations, came in at only 58 quads for 2010. The highest estimate, based on a 2% annual real price increase coupled with a rapidly growing economy, came in at 136 quads for 2010. This study clearly illustrated the importance in scenario analysis of assumptions about everything from income, labor force, and the rate of household formation, to worker productivity, energy productivity, energy price, income, and energy substitutability in scenario analysis. It also highlighted the importance of capital stock turnover in energy futures, in that the largest technical opportunities for increased efficiency derive from new stock, where state-of-the-art technology can be most effectively employed. The CONAES project had a major influence on all of the energy modeling work that followed.

2.2 Oil Shocks

In October 1973, the United States helped Israel beat back the Yom Kippur invasion by Egypt and Syria. OPEC (dominated by Arab oil producers) retaliated with an oil embargo on the United States and a 70% oil price increase on West European allies of the United States. Homeowners were asked to turn down their thermostats and gas stations were requested to limit sales to 10 gallons per customer. Illumination and speed limits were lowered, and the traffic law allowing a right turn on a red light came into use. Gasoline lines, conspiracy theories, and recession followed in rapid succession. As Adlai Stevenson once observed, "Americans never seem to see the handwriting on the wall until their backs are up against it." We perceived the wall behind our backs and responded accordingly—for a brief spell. American humor also surfaced. Some wonderful cartoons depicted innovative ways to save energy—for example, by sleeping with a chicken, which has a normal body temperature of 107°F, and public interest highway signs appeared proclaiming "Don't be fuelish!"

OPEC lifted the embargo in 1974, but political unrest in the Middle East again precipitated a major oil price increase in 1979. When anti-Western Islamic fundamentalists gained control of Iran, oil production in that nation dropped off dramatically and U.S. crude oil prices tripled. These price increases also led to skyrocketing inflation and recession, and rekindled interest in energy efficiency.

2.3 Policy Responses

The need to curtail energy consumption in response to U.S. vulnerability to oil supply disruptions dominated energy policy in the 1970s. In the summer of 1973, President Nixon, anticipating possible shortages of heating oil in the coming winter, and summer gasoline shortfalls, established the Office of Energy Conservation to coordinate efforts to cut energy consumption in federal agencies by about 5%, raise public attention to the need for increased efficiency, encourage private industry to give greater attention to saving energy, and devote more federal research and development to efficiency. After the first oil shock, President Nixon urged voluntary rationing and pushed for a commitment to U.S. energy independence by 1985. President Carter emphasized the importance of conservation in the effort to achieve energy independence, but appeared to equate conservation with wearing a sweater in a chilly White House. Early actions by our political leaders certainly contributed to a negative association of conservation with sacrifice. But they also contributed to an atmosphere conducive to serious public and private sector actions to incorporate efficient end-use

technologies in the U.S. economy. Congress enacted legislation to create standards for Corporate Average Fuel Economy (CAFE) and building energy performance. They also created a cabinet-level Department of Energy (DOE) by merging the functions of preexisting agencies scattered throughout government. In response to an analysis by the congressional Office of Technology Assessment, Congress explicitly included energy conservation in the DOE's new mandate.

3. DECADE OF NO- AND LOW-COST OPTIONS: 1980s

The combination of the oil price shocks with increasing public concern about the environmental consequences of energy use prompted policy makers to take a serious look at policy tools for decreasing energy demand growth during the 1980s. Faced with higher oil prices, industrial and commercial oil consumers also sought out cost-effective investments in energy conservation during this decade. A plethora of no- and low-cost options were ripe for the picking, and progressive industries did just that, with great success at the bottom line.

3.1 Policy Tools

At the beginning of the 1980s, energy analysts converged on a set of principles for a comprehensive energy policy that had conservation at its core. These principles included the following ideas:

1. Consider the production and use of energy as means to certain ends, not as goals in and of themselves. Remember always that, given time and the capital for adjustment, energy is a largely substitutable input in the provision of most goods and services.
2. Application of technical ingenuity and institutional innovation can greatly facilitate energy options.
3. Energy decisions, like other investment decisions, should be made using clear signals of comparative total long-run costs, marginal costs, and cost trends.
4. It is important to correct distorted or inadequate market signals with policy instruments; otherwise, external costs can be ignored and resources can be squandered. This correction includes internalizing in energy price and/or regulation, to the extent possible, the national security, human health, and environmental costs attributable to energy.
5. Investment in both energy supply and utilization research and development is an appropriate activity for both the public and private sectors, because costs and benefits accrue to both sectors.
6. There are other, generally more productive, ways (for example, assistance with insulation or lighting retrofits, fuel funds, or even refundable tax credits) to assist underprivileged citizens with their energy needs, rather than subsidizing energy's price to them.
7. In a world characterized by tightly integrated economies, we need to increase our cognizance of world energy resource conditions and needs with special regard for international security as well as concern for the special needs of poor nations.

By the mid-1980s, the national energy situation had begun to reflect some of these ideas. CAFE standards reached their peak (27.5 miles per gallon) in 1985. Congress created the Strategic Petroleum Reserve (SPR) to respond to energy supply emergencies. Oil and gas price deregulation was virtually complete, but that was only the first step in internalizing the costs of energy consumption. Energy prices still did not reflect the manifold environmental costs of production and use, the costs of defending Middle East oil production and shipping lanes, the costs of purchasing and storing oil to meet emergencies, or the impacts of U.S. competition for scarce oil resources on developing nations. As the price of imported oil decreased, U.S. policy makers were lulled into complacency about the nation's energy strategy. A dearth of energy research and development expenditures symbolized this declining interest. Before the end of the decade, however, Congress passed the National Appliance Energy Conservation Act, which authorized higher appliance efficiency standards and efficiency labeling that have significantly impacted the market for energy-efficient appliances.

3.2 Private Sector Activities

Immediately following the oil shocks, private sector efforts to save energy focused on changing patterns of energy use within the existing infrastructure, such as lowering thermostats. Most actions involved investments in technology, however—either retrofits of existing technology, such as insulating existing homes, or new investments in technology, such as energy-efficient new construction or autos with improved mileage. Later in the 1980s, energy efficiency gains accrued as incidental benefits to

other, much larger investments aimed at improving competitiveness of U.S. products in world markets.

All in all, energy efficiency investments in the 1970s and the 1980s turned out to be generally easier and much more cost-effective than finding new sources of energy. Nongovernmental institutions helped to keep the conservation ball of progress rolling, especially in the residential and commercial sectors, with information to increase awareness of cost-effective conservation opportunities, rebates, and prizes. In the first decade after the oil shocks, industry cut energy requirements per unit of output by over 30%. Households cut energy use by 20% and owners and operators of commercial buildings cut energy use per square foot by more than 10%. Transportation energy efficiency steadily improved as CAFE requirements slowly increased fleet average mileage. However, political maneuvering by car companies resulted in excepting "light trucks," including sport utility vehicles (SUVs), from the mileage requirements for passenger cars. This has resulted in a major slow down of improvement in fleet performance.

Some surprising partnerships formed to promote energy efficiency. For example, the Natural Resources Defense Council designed a Golden Carrot Award to encourage production of a superefficient refrigerator. In response, 24 major utilities pooled their resources to create the $30 million prize. Whirlpool won the competition and collected its winnings after manufacturing and selling at least 250,000 units that were at least 25% more efficient than the 1993 efficiency standard required. By the end of the 1980s, energy efficiency improvements lost momentum. One reason was the decline in oil prices: corrected for inflation, the average price of gasoline in 1987 was half that in 1980. Natural gas and electricity prices also fell, and the price-driven impetus for consumers and businesses to conserve was diminished.

3.3 Lessons from Experience

By the end of the decade, conservation advocates were urging policy makers to take several lessons from the experiences of 1970 and 1980s:

1. First and foremost, that energy conservation worked. Nothing contributed more to the improved American energy situation than energy efficiency. By the late 1980s, the United States used little more energy than in 1973, yet it produced 40% more goods and services. According to one estimate, efficiency cut the nation's annual energy bill by $160 billion.
2. Oil and gas price controls are counterproductive. Price increases of the 1970s and 1980s sparked operating changes, technology improvements, and other conservation actions throughout the United States.
3. Technological ingenuity can substitute for energy. A quiet revolution in technology transformed energy use. Numerous efficient products and processes—appliances, lighting products, building techniques, automobiles, motor drives, and manufacturing techniques—were developed and commercialized by private companies. Energy productivity rose as more efficient equipment and processes were incorporated into buildings, vehicles, and factory stock.
4. Complementary policies work best. The most effective policies and programs helped overcome barriers to greater investment in conservation—barriers such as the lack of awareness and low priority among consumers; lack of investment capital; reluctance among some manufacturers to conduct research and to innovate; subsidies for energy production; and the problem of split incentives, exemplified by the building owner or landlord who does not pay the utility bills but does make decisions on insulation and appliance efficiency.
5. The private sector is the primary vehicle for delivering efficiency improvements, but government–industry cooperation is essential.

Many activists advocated adoption of a new national goal for energy intensity in order to regain momentum toward a secure energy future. The United States achieved a 2.7% annual rate of reduction in energy intensity between 1976 and 1986. The DOE was predicting, however, that U.S. energy intensity would decline at less than 1% per year through the rest of the 20th century. This is the rate of improvement now widely assumed as part of "business as usual." Recognizing that economic structural change had also influenced this rapid improvement and that many of the least expensive investments in efficiency had already been made, the present authors and others, including the American Council for an Energy Efficient Economy, nonetheless recommended a goal of reducing the energy intensity of the U.S. economy by at least 2.5% per year into the 21st century. For the United States, even with an apparently disinterested Administration, it is now proposed that energy intensity over the next

decade be improved by 18% (roughly the average yearly gain over the past two decades).

4. DECADE OF GLOBALIZATION: 1990s

The United States went to war for oil at the beginning of the 1990s. Oil prices initially spiraled upward, temporarily reinvigorating interest in conservation investments. That interest dwindled as quickly as prices fell. Still, after an inauspicious beginning, during the economic boom of the late 1990s, with major investments in new capital stocks, energy intensity dropped at an average rate of over 2% per year.

The role of carbon emissions in global climate change began to dominate the debate over energy in the 1990s. Although U.S. total energy consumption far outstripped any other country's, consumption was projected to grow at much higher rates in the developing countries, and U.S. policy makers shifted their attention to U.S. participation in international cooperation on energy innovation. In 1993, the U.S. government and the U.S. automobile industry forged an unprecedented alliance under the leadership of President Clinton and Vice President Gore. The partnership included seven federal agencies, 19 federal laboratories, and more than 300 automotive suppliers and universities and the United States Council for Automotive Research, the precompetitive research arm of Ford, DaimlerChrysler, and General Motors. The Partnership for a New Generation of Vehicles (PNGV) supported research and development of technologies to achieve the program's three research goals: (1) to significantly improve international competitiveness in automotive manufacturing by upgrading manufacturing technology; (2) to apply commercially viable innovations resulting from ongoing research to conventional vehicles, especially technologies that improve fuel efficiency and reduce emissions; and (3) to develop advanced technologies for midsized vehicles that deliver up to triple the fuel efficiency of today's cars (equivalent to 80 miles per gallon), without sacrificing affordability, performance, or safety. The research plan and the program's progress were peer-reviewed annually by the National Research Council. PNGV made extraordinary progress toward achieving its aggressive technical goals. In March 2000, PNGV unveiled three concept cars demonstrating the technical feasibility of creating cars capable of getting 80 miles per gallon. All three cars employ some form of hybrid technology that combines a gasoline- or diesel-powered engine with an electric motor to increase fuel economy. The three major automakers also confirmed their commitment to move PNGV technology out of the lab and onto the road by putting vehicles with significant improvements in fuel economy into volume production and into dealers' showrooms. Work continues on technologies that might contribute to the full achievement of goals for the 2004 prototype.

In 1999, the President's Committee of Advisors on Science and Technology (PCAST) issued recommendations for improving international cooperation on energy conservation. Reflecting a growing national consensus, they reported that efficient energy use helps satisfy basic human needs and powers economic development. The industrialized world depends on massive energy flows to power factories, fuel transport, and heat, cool, and light homes, and must grapple with the environmental and security dilemmas these uses cause. These energy services are fundamental to a modern economy. In contrast, many developing-country households are not yet heated or cooled. Some 2 billion people do not yet have access to electric lighting or refrigeration. The Chinese enjoy less than one-tenth as much commercial energy per person as do Americans, and Indians use less than one-thirtieth as much. Raising energy use of the world population to only half that of today's average American would nearly triple world energy demand. Such growth in per capita energy use coupled with an increase by over 50% in world population over the next half-century and no improvement in energy efficiency would together increase global energy use more than four times. Using conventional technologies, energy use of this magnitude would generate almost unimaginable demands on energy resources, capital, and environmental resources—air, water, and land.

Energy-efficient technologies can cost-effectively moderate those energy supply demands. For example, investments in currently available technologies for efficient electric power use could reduce initial capital costs by 10%, life-cycle costs by 24%, and electricity use by nearly 50%, compared to the current mix of technologies in use. Modest investments in efficient end-use technologies lead to larger reductions in the need for capital-intensive electricity generation plants. Conversely, when unnecessary power plants and mines are built, less money is available for factories, schools, and health care.

In the language of economics, there is a large opportunity cost associated with energy inefficiency. It is both an economic and environmental imperative that energy needs be satisfied effectively. Incorporating energy efficiency measures as economies develop can help hold energy demand growth to manageable levels, while reducing total costs, dependence on foreign sources of energy, and impacts on the environment. Ironically, energy is most often wasted where it is most precious. Developing and transition economies, such as in China, the world's second largest energy consumer, use much more energy to produce a ton of industrial material compared to modern market economies. Developing-nation inefficiency reflects both market distortions and underdevelopment. Decades of energy subsidies and market distortions in developing and transition economies have exacerbated energy waste. Energy efficiency requires a technically sophisticated society. Every society in the world, including the United States, has at one time encouraged energy inefficiency by controlling energy prices, erecting utility monopolies, subsidizing loans for power plant development, and ignoring environmental pollution. Many nations, including the United States, continue these wasteful practices. These subsidies and market distortions can seriously delay technological advances, even by decades, compared to best practice. The worldwide shift to more open, competitive markets may reduce some of the distortions on the supply side, but will do little to change the inherent market barriers on the demand side.

China, India, Brazil, Russia, and other developing and transition economies are building homes and factories with outdated technology, which will be used for many decades. In cold northern China and Western Siberia, for example, apartments are built with low thermal integrity and leaky windows and are equipped with appliances half as efficient as those available in the United States. Developing nations thus tend to build in excessive energy costs, lock out environmental protection, and diminish their own development potential. There are important exceptions. Progress can be both rapid and significant. Over recent decades, China has, through energy-sector reform and pursuit of least-cost paths via energy efficiency opportunities, made unprecedented progress in energy intensity reduction, faster than any developing nation in history. China has held energy demand growth to half of its rate of overall economic growth over the past two decades (Fig. 3). This example demonstrates that economic development can proceed while restraining energy demand growth; indeed, by employing cost-effective energy-efficient technologies, funds are freed for investment in other critical development needs. Energy efficiency can thus be a significant contributor to social and economic development.

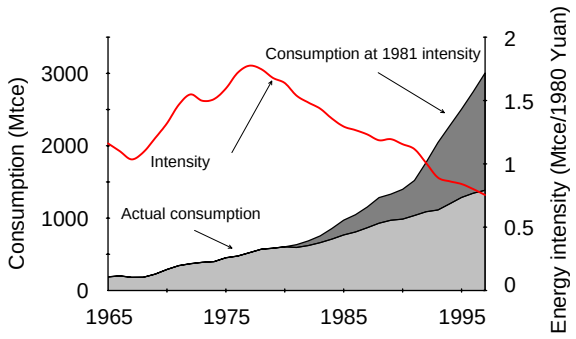


FIGURE 3 Energy intensity and consumption in China. Mtce, Million tons of coal equivalent. From Advanced International Studies Unit, Battelle Pacific Northwest National Laboratory.

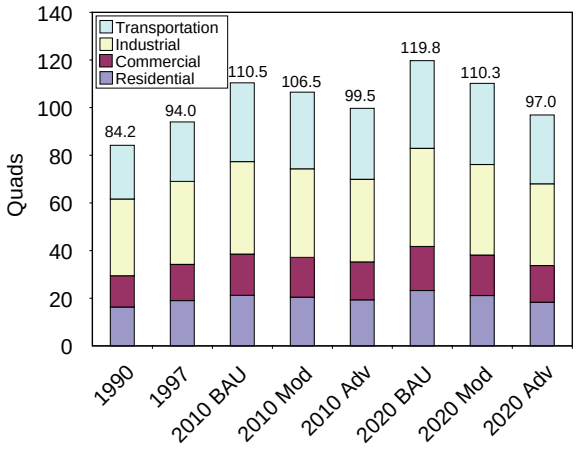


FIGURE 4 Primary energy consumption by sector. BAU, Business-as-usual model, compared to moderate (Mod) and advanced (Adv) conservation models. From Interlaboratory Working Group (2000).

5. DECADES AHEAD

Energy projections indicate differences in U.S. energy demand in 2020 between the high and low scenarios of approximately 25 quads (Fig. 4). The difference between these scenarios is substantially due to differing assumptions about development and deployment of energy-efficient technologies. The differences also reflect the plausible range of key consumption-driving factors that, operating over 20 years, can have a profound influence. These factors

include economic growth rate and the (rising) urgency to cut emissions of greenhouse gases in order to slow global climate change.

5.1 Cultural Barriers to Conservation

Sadly, there remains a persistent struggle between advocates of energy supply and conservation advocates. On the supply side, the energy industry is massive and concentrated, wielding great political influence (e.g., obtaining tax subsidies, allowing unhealthy levels of environmental emissions). On the conservation side, the relevant industry is diffusely spread over the entire economy, and energy efficiency is seldom the primary concern. Furthermore, decisions regarding the energy efficiency of appliances (e.g., heating, ventilation and air-conditioning systems and refrigerators) are frequently made (e.g., by developers, not consumers) on the basis of minimum first cost rather than life-cycle cost, thus generally favoring lower efficiency.

Two streams of progress over the past 50 years have combined to accentuate the attention given to energy conservation. First, steady and ubiquitous advances in technology (such as computers and various special materials) have enabled remarkable improvements (lower cost, higher efficiency) in productivity, including energy. Second, there has been a growing awareness of major direct and indirect externalities associated with energy production and conversion. Although such externalities (e.g., health, ecological stress, climate change, and air and water pollution) are now widely recognized, they still are not generally incorporated into national economic accounts. Instead, various standards, regulations, and policies have been developed or proposed to reflect the external costs by acting as a "shadow price," in order to level the playing field of price signals.

Despite its profound impacts, the conservation revolution has resulted in only a fraction of its ultimate potential. Here are some of the reasons:

1. Lack of leadership. Whereas an energy consumer can make short-term adjustments to save energy (driving slower, driving fewer miles and traveling less, adjusting the thermostat, or turning off lights), these mostly have the unpopular effect (other than saving money) of curtailing the very energy services that we seek. Sadly, such heroic short-term curtailment measures are too often posed as the essence of conservation rather than emergency measures. On national television, President Carter advocated energy conservation as a moral imperative, but delivering his speech while wearing a sweater and sitting by an open fire connoted sacrifice and denial of comfort. Worse yet, President Reagan referred to conservation as being hot in summer and cold in winter! In the short run, curtailment is about all anyone can do, but if we stop there, we lose by far the most impressive opportunities. After the shocks of the 1970s subsided, attention shifted to the more substantial but also more challenging task of capturing longer term opportunities.

2. Time lags. The largest opportunities for energy conservation derive from incorporating more efficient technologies into new capital investments (e.g., cars, buildings, appliances, and industrial plants). These logically accrue at a rate that reflects depreciation of existing capital stock, ranging from years to many decades. Just as for the case of major transitions in energy systems, the major opportunities for conservation inherently require decades to be fully achieved.

3. Discount rates. With a few exceptions, capital purchase decisions tend to favor minimum first cost, rather than minimum life-cycle cost. Generally, the minimum first-cost model corresponds to the least energy-efficient model. Thus the "least-cost" solution is typically not taken by most consumers and this slows the rate at which efficiency is gained.

5.2 Remaining Potential for Conservation of Energy Resources

Liquid fuels, so attractive because they can be stored and transported with relative ease, loom as a global challenge because the main supply (petroleum) is very unevenly distributed geographically and is often concentrated in insecure locations. Technologies for petroleum conservation are steadily improving, from elegant, new methods of exploration and production to highly efficient combustion in motors and turbines. Thus the petroleum resource base, although headed for depletion in this century, can be stretched (conserved) by a variety of technologies. Advances in high-temperature and low-corrosion materials now enable a variety of combustion turbines that can be employed to generate electricity with 60% or better thermodynamic efficiency—double the best-practice several decades ago.

Probably the greatest challenge for liquid fuel conservation for the future is in the transportation sector. Technology, the auto industry, public policies, and consumer preference all play important roles. The industry is most profitable in making large,

heavy and powerful cars and SUVs; consumers in the United States presently have a love affair with the image of power, ruggedness, and convenience with little regard for efficiency. And the record shows that U.S. public policy is to keep fuel price below total cost. Advances in automotive efficiency such as hybrid-electric/advanced diesel and fuel cells could help alleviate the situation. For example, over the past decade, work on streamlining, electrical power control systems, high-compression/direct-injection engines, electric drive trains, etc. has resulted in the recent market entry of hybrid cars that are very popular and have about 50% better mileage; in addition, rapid advances in fuel cell technology promise within a decade or so more conservation as well as lower environmental emissions. However, there seems to be little reason to hope for major gains in efficiency without major increases in fuel price or its equivalent shadow price (performance regulation such as CAFE).

Electricity generation is getting very efficient and the technologies that use electricity are also advancing. The world is still electrifying at a rapid pace. Electricity can now be generated from diverse energy sources and with diverse sizes of generators. Source diversity means inherent resilience of the system and capability for disaggregated generation means that "waste heat" from generators can be more readily utilized (cogeneration). Electricity is such a high-quality and versatile energy form that it is likely it will continue to increase in market share. Thus further advances in the efficiency of conversion and use of electricity should find ready markets.

The historic succession of energy sources over the past 150 years has led to progressively higher hydrogen-to-carbon (H/C) ratios for the fuels (e.g., coal, followed by oil, followed by natural gas). Global climate protection argues that we need to move with dispatch toward even higher H/C ratios in the coming decades. This implies a combination of removal (sequestering) of carbon from combustion releases, increased carbon recycle (e.g., renewable fuels from biomass), direct use of solar energy, and increased use of nuclear energy. Under this scenario, hydrogen becomes the most likely candidate for the derived fuel. This underscores the promise of high-efficiency fuel cells for energy conversion to electricity.

The amount of conservation achieved through advancing technology has been outstanding over the past half-century. Many more gains are possible, but we can also discern limits dictated by limits imposed by natural law. Carnot's thermodynamic rule for efficiency of heat engines is one limit. On the other hand, fuel cells give an opportunity to bypass Carnot. Success in this area could effectively enable a doubling of the useful work out of a unit of energy compared to an internal combustion engine, and with basically no environmental emissions. Success beckons but will require many technical advances in special materials and also in providing an acceptable primary energy source—probably hydrogen—all at a much lower production cost than now. The artificially low current price of gasoline constitutes a serious barrier to market entry of the new technologies.

Human ingenuity in the form of technological innovation will continue to provide new options to society. Energy conservation, which derives from sophisticated and elegant technologies, has every chance to continue paying big dividends to technologically capable societies. Conservation enables humanity to receive more "goods" along with fewer "bads." But most things have limits. Ultimately, we must recognize the folly of trying to follow an exponential path of population growth and resource consumption in its present form. We must equip ourselves with new tools, including conservation, and move with dispatch on a journey toward sustainability in the 21st century. There is little time to spare.

SEE ALSO THE FOLLOWING ARTICLES

Alternative Transportation Fuels: Contemporary Case Studies • Economics of Energy Efficiency • Energy Efficiency, Taxonomic Overview • Internal Combustion Engine Vehicles • International Comparisons of Energy End Use: Benefits and Risks • Obstacles to Energy Efficiency • Oil Crises, Historical Perspective • OPEC Market Behavior: 1973–2003 • Passenger Demand for Travel and Energy Use • Transportation and Energy, Overview

Further Reading

Chandler, W. (2000). "Energy and Environment in the Transition Economies." Westview Press, Boulder and Oxford.

Chandler, W., Geller, H., and Ledbetter, M. (1988). "Energy Efficiency: A New Agenda." GW Press, Springfield, Virginia.

Demand and Conservation Panel of the Committee on Nuclear and Alternative Energy Systems. (1978). U.S. Energy Demand: Some Low Energy Futures(14 April 1978). *Science* **200,** 142–152.

Gibbons, J. (1997). "This Gifted Age: Science and Technology at the Millennium." AIP Press, Woodbury, New York.

Gibbons, J., and Chandler, W. (1981). "Energy: The Conservation Revolution." Plenum Press, New York and London.

Gibbons, J., Blair, P., and Gwin, H. (1989). Strategies for energy use. *Sci. Am.* **261,** 136–143.

Interlaboratory Working Group. (2000). "Scenarios for a Clean Energy Future." Oak Ridge National Laboratory and Lawrence National Laboratory, Oak Ridge, Tennessee, and Berkeley, California.

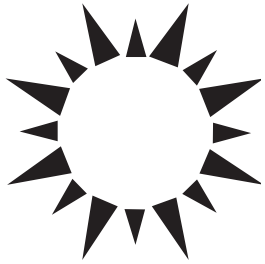
President's Committee of Advisors on Science and Technology. (1997). "Federal Energy Research and Development for the Challenges of the 21st Century." Office of Science and Technology Policy, Washington, D.C.

President's Committee of Advisors on Science and Technology. (1999). "Powerful Partnerships: The Federal Role in International Cooperation on Energy Innovation." Office of Science and Technology Policy, Washington, D.C.

Sant, R. (1979). "The Least-Cost Energy Strategy: Minimizing Consumer Costs through Competition." Carnegie-Mellon University Press, Pittsburgh, Pennsylvania.

U.S. Congress, Office of Technology Assessment. (2004). "OTA Legacy" (includes reports on Conservation). U.S. Govt. Printing Office, Washington, D.C. See also http://www.wws.princeton.edu/ota/.

Conservation of Energy Concept, History of



ELIZABETH GARBER
State University of New York at Stony Brook
Stony Brook, New York, United States

1. Motion and Mechanics before 1800
2. Caloric Theory and Heat Engines
3. Equivalence of Heat and Work
4. Conservation of Energy
5. Energy Physics
6. Energy in 20th Century Physics
7. Energy Conservation beyond Physics

Glossary

caloric theory The theory that heat is a substance that flows into or out of a body as it is heated or cooled; this theory was replaced by the notion of the conservation of energy around 1850.

conservation of energy The notion that in a closed system, energy cannot be created or destroyed but can only be converted into other forms.

energy The capacity of a body to do work; it comes in various forms that can be converted into each other.

ether A medium that was assumed to fill space to explain the propagation of light, heat, and electromagnetic waves; it was discarded during the early 20th century.

heat engine Any device that takes heat and converts it into mechanical work.

mechanical equivalent of heat A constant that expresses the number of units of heat in terms of a unit of work.

mechanical theory of heat The theory that heat consists of the motions of the particles that make up a substance.

perpetual motion The fallacious theory that a machine, producing work, can be operated without any external source of power; it was discussed by philosophers and sold to kings but has been ignored by engineers since ancient times.

vitalism The theory that life depends on a unique force and cannot be reduced to chemical and physical explanations; this theory has been discarded.

work done Measured by force multiplied by the distance through which its point of application moves.

The modern principle of the conservation of energy emerged during the middle of the 19th century from a complex of problems across the sciences, from medicine to engineering. It was quickly accepted as a principle of nature and was applied to problems in mechanics and electromagnetism, and then in other sciences, even when it was not the most useful approach. It survived challenges even into the 20th century.

1. MOTION AND MECHANICS BEFORE 1800

Some ancient philosophers and most engineers assumed that perpetual motion was impossible, but nobody examined the efficiency of machines. The latter did not become central to engineering until the 18th century, after exhaustive metaphysical and methodological debates about motion and mathematics as necessary components of a comprehensive philosophy of nature. Natural philosophers and mathematicians argued over the conditions under which particular conservation laws were valid, and through these debates motion and collisions drew more attention from engineers and mathematicians.

In Rene Descartes's natural philosophy, the universe consisted of matter, whose quantity was measured by size, and motion. Collisions between particles was the mechanism for change, and the total amount of motion in the universe remained constant, although motion could be lost or gained in each collision. Descartes also assumed that the natural motion of bodies was linear—not circular—and was a vector. In collisions, the quantity of motion, size multiplied by velocity, could be transferred from one particle to another. The force of a

 © 2004 Elsevier Inc. All rights reserved.

particle was its size multiplied by the square of its velocity. Although Descartes modeled his philosophical method on geometry, very little mathematics was contained within it. It was left to later generations to put his philosophy of nature into mathematical form.

One of the most successful was Isaac Newton, who began as a follower of Descartes and ended as his scientific and philosophical rival. In his explorations of Descartes's system, Newton became dissatisfied with Cartesian explanations of cohesion and other phenomena. He eventually accepted the reality of the concept of "force," banned by Descartes as "occult," as acting on micro and macro levels. Newton's universe was as empty as Descartes's was full, and God was necessary to keep His vast creation going. In such a universe, conservation laws and collisions were not of central importance. Yet Newton introduced a measure of force, momentum, that became part of the debates on conservation laws in mechanics.

One of the first entries in this debate was that of the Cartesian Christian Huyghens, who examined circular motion and the collisions of perfectly "hard" (perfectly elastic) bodies. He demonstrated that in such collisions, the sum of the magnitude of the bodies multiplied by the square of the velocity was a constant. However, for Huyghens, this expression was mathematically significant but not physically significant.

This pattern was repeated as 18th century continental mathematicians debated the measure of force and the quantities conserved in collisions. Gottfried Leibniz argued that Descartes's measure of force was insufficient. According to Leibniz, a body dropped a certain distance acquired just enough force to return to its original position, and this was a truer measure of force. Descartes had confused motive force with quantity of motion. Force should be measured by the effect it can produce, not by the velocity that a body can impress on another. In perfectly hard body collisions "vis viva," living force ($\frac{1}{2}mv^2$) was conserved. Central to subsequent debates were the experimental results of Giovanni Poleni and Wilhelm s'Gravesande, who measured the depths of the indentations left by metal balls dropped at different speeds into clay targets. If balls of different masses but the same size were dropped, the indentations remained the same. Lighter balls left the same indentations if they were dropped from heights proportional to 1/masses. s'Gravesande's experiments also showed that the force of a body in motion was proportional to v^2, that is, to its vis viva. He also did some elegant experiments in which hollow metal balls were filled with clay to investigate collisions of "soft" bodies. The results only fueled debates about the conditions over which measure of force was appropriate, mv or $\frac{1}{2}mv^2$.

s'Gravesande began his experiment to prove Newton's theories were correct, but the experiments seemed to Cartesians to support their point of view. Despite this evidence, s'Gravesande remained a Newtonian; for him, $\frac{1}{2}mv^2$ remained a number of no physical significance. Mathematicians in these debates over force included Madame de Chatelet, who published a French translation of Newton's *Principia* just 2 years after she had published *Institutiones* to spread Leibniz's ideas among French mathematicians. The subject of vis viva became a part of ongoing debates in the salons and Paris Academie des Sciences over Newtonian or Cartesian natural philosophy and the accompanying politics of each of those institutions.

Nothing was resolved, and mathematicians eventually moved on to other problems. They were inclined to use whatever physical principles seemed the most appropriate in their quest to develop the calculus using mechanical problems of increasing sophistication. Mechanics, or problems within it, became the instrument to display mathematical expertise and brilliance rather than explorations into physical problems. A perfect illustration of this is the work of Jean Louis Lagrange. By considering a four-dimensional space and using calculus, Lagrange reduced solid and fluid dynamics to algebra. He began with the fundamental axiom of virtual velocity for mass particles,

$$\sum m\frac{d\vec{v}}{dt}\cdot\delta\vec{r} = \sum \vec{F}\cdot\delta\vec{r},$$

where m is mass, v is velocity, $\delta\vec{r}$ is virtual displacement, and $\vec{F}$ is the applied force acting on a particle. Lagrange then introduced generalized coordinates, $q_1 \ldots q_n$, and generated the equations,

$$\frac{\partial T}{\partial q_i} - \frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_i}\right) - \frac{\partial \Phi}{\partial q_i} = 0,$$

where T is the vis viva and Φ is a function. Thus, he succeeded in making rational mechanics as abstract, general, and algebraic as possible. No diagrams, geometry, or mechanical reasoning graced the pages of his text. It consisted "only of algebraic operations."

2. CALORIC THEORY AND HEAT ENGINES

During the 18th century, French engineers developed entirely different, yet mathematical, approaches to mechanics. Mathematicians had used problems on the construction of bridges, the construction of pillars, and the design of ships to explore the calculus. The education of French engineers began to include some calculus to solve engineering problems, especially in those schools of engineering within branches of the military. Here they developed an independent approach to the understanding of structures and the operations of machines.

Out of this tradition emerged Lazare Carnot, whose 1783 treatise on the efficiency of machines also began with the principle of virtual velocities. From this, he developed a relationship among the moment of activity, force multiplied by distance (work), and vis viva. He concluded that to get the maximum "effect" from the cycles of a machine, it was necessary to avoid inelastic impact in the working of its parts. During the 1820s, his son Sadi turned the same analytical approach toward the new "machine" that was revolutionizing the economy of Great Britain, the steam engine. As Carnot noted, removing this source of power would "ruin her prosperity" and "annihilate that colossal power." By this time, steam engines had been used in Britain for more than a century as pumps for mines. More recently, through the development by James Watt of the separable condenser and a governor, the steam engine became a new driving force for machinery. By the 1830s, the use of steam power and associated machinery was integrated into American and British cultures as a measure of the "civilized" states of societies.

Carnot wanted to establish the conditions under which maximum work could be produced by a heat engine. He stripped the heat engine down to its essential elements and operations. Through a series of examples, he demonstrated that successive expansions and contractions from heating and cooling—not heat alone—produced the motive power. Steam was not unique; any substance that expanded and contracted could be a source of power. Gases generally afforded the most expansion and contraction on heating and cooling. Carnot thought of heating and then cooling as the "reestablishment of equilibrium" in caloric and the substance of heat as the source of motive power. Therefore, he began to search for the general conditions that would produce the maximum motive power, and in principle, this could be applied to any heat engine.

Sadi Carnot used the same approach as his father and other French engineers in analyzing machines by looking at infinitesimal changes. In this case, the temperatures of the heat source and sink differed only slightly from that of the working substance. Carnot examined the changes in temperature and pressure of the working substance in a complete cycle of the engine, again a familiar approach in French engineering. Initially, the body is in contact with the heat source, and its piston rises and its volume expands at a constant temperature. It is then removed from this source, and the expansion continues without any heat input and the temperature of the working substance decreases to that of the heat sink. Placed in contact with the cooler body, the gas is compressed at constant temperature and then is removed and further compressed until the cycle is closed and the working substance attains the temperature of the heat source. The Carnot cycle is then complete. However, Carnot's analysis of its operation was in terms of caloric theory. In the first operation, the caloric is removed from the heat source to the working substance, and motive power is produced as the caloric attains the temperature of the heat sink. Equilibrium is restored as the caloric again reaches the temperature of the heat source. Using the incomplete image of a waterfall used to drive machinery, the caloric is conserved. It has fallen from one temperature to a lower one, and in this fall we can extract motive power just as we do from the gravitational fall of water.

The caloric theory of heat was accepted across Europe by the latter part of the 18th century. Experimental research into heat occupied many chemists across Europe and had an important place in the development of the reformed chemistry of Antoine Lavoisier. "Airs" (gases) were also the object of avid research throughout Europe by many chemists from before the flight of the Montgolfier brothers in 1784 through the first third of the 19th century. Caloric theory drew these two strands of research together. During the 1780s, Lavoisier had measured the specific heats of a number of substances with Pierre Simon de Laplace. These experiments were aspects of Lavoisier's research on airs and heat. In another series of such experiments, Lavoisier studied digestion and animal heat. Using an ice calorimeter, he demonstrated that respiration was a slow form of combustion and predicted heat outputs that agreed with his experimental results. Both Lavoisier and Laplace believed that heat was a

material substance that manifested itself in two forms: free caloric (detected by thermometers) and latent caloric. Laplace and Simon Poisson developed caloric theory mathematically by assuming that the caloric, Q, in a gas was a function $Q = f(P, \rho, T)$, where P is the pressure, T is the temperature, and ρ is the density of the gas. Manipulating the partial differential equations in a theory of adiabatic compression, this general equation led to relationships in accord with known experimental results such as the ratio of the specific heats of gases. Caloric was connected to the notion of work through Laplace's theory. Free heat was vis viva, combined (latent) heat was the loss of vis viva, and heat released was an increase in vis viva. Caloric theory was sophisticated and was taught at the Ecole Polytechnique when Sadi Carnot was a student, although toward the end of his short life he rejected it to explore the idea that heat was the motion of particles.

Carnot used the latest experimental results on the properties of gases to demonstrate his contentions and then discussed the limitations of heat engine design at the time. He concluded that because of gases' great expansion and contraction with heating and cooling, and because of the need for large differences in temperature to gain the most motive power, gases were the only feasible working substances for heat engines. He used current research on gases to demonstrate that any gas might be used.

Although Carnot's small treatise, *Reflections on the Motive Power of Heat,* received a good review from the *Revue Encyclopedique,* only one engineering professor at the Ecole Polytechnique developed Carnot's ideas. Emile Clapyeron gave geometrical and analytical expression to Carnot's analysis, creating the first pressure–volume (PV) diagram of a Carnot cycle. In his analysis, this cycle was reduced to an infinitesimally small parallelogram for the case pursued by Carnot. Again using the results from work on the properties of gases, Clapyeron obtained an expression for the "maximum quantity of action" developed in a cycle from the area within the parallelogram of the heat cycle for a perfect gas heat engine. Again, the analytical expressions were in terms of general expressions for heat, $Q = Q(P, V, T)$, $P = P(V, T)$, and $V = V(P, T)$. In the concluding paragraphs of his memoir, Clapyeron described the mechanical work available from his heat engine in terms of the vis viva the caloric developed in its passage from the boiler to the condenser of the engine. Many of his expressions were independent of any particular theory of the nature of heat and remained valid even as their physical explanation changed during the 1850s.

The works of both Carnot and Clapyeron sank into oblivion. Carnot died young and tragically in an asylum, but his oblivion was partially political. His father, Lazare, had enthusiastically embraced the French Revolution, was important in its early wars, and rose to prominence in Napoleon's regime. He remained loyal to Napoleon until the end and died in exile. The Carnot name (except that of his brother, a prominent politician), was politically tainted even into the 1830s. And however useful caloric might be for engineers, among scientists the caloric theory held less interest during this decade. They extended the wave theory of light to heat, and through it they explained other scientifically pressing problems such as the conduction and radiation of heat.

So deep was Carnot's oblivion that in 1845 William Thomson, a young ambitious Scottish academic, could find no copies of his treatise in Paris. Thomson had recently graduated Second Wrangler and first Smith's Prizeman from Cambridge University and was actively seeking the empty chair in natural philosophy at Glasgow University. The Glasgow that Thomson grew up in was Presbyterian and was the center of heavy industry (especially shipbuilding) and chemical engineering in Great Britain. Although his father was a distinguished professor of mathematics at Glasgow, the younger Thomson needed the support of the rest of the faculty and of the town representatives to ensure his election. He had to cultivate the image of the practical experimenter rather than flaunt his mathematical abilities, hence Thomson's sojourn at the laboratory of Henri Victor Regnault, well known for his empirical studies of steam engines. These were the circumstances that led Thomson to search out Carnot's work. Undaunted, Thomson made full use of Clapyeron's publications in a series of papers (published in 1847 and 1850) extending Clapyeron's analysis. In his first paper he postulated an absolute zero of temperature implied by Clapyeron's analysis and put forward the notion of an absolute scale of temperature. Both emerged from the mathematical forms independent of their physical interpretation. Thomson's brother, James, noted that one of the results of Clapyeron's work was that the freezing point of water decreased with an increase in pressure. William Thomson confirmed this through experiments done at Glasgow. The position of caloric theory seemed to be becoming firmer once again. However, in the middle of this theoretical and experimental triumph, Thomson went to the 1847

annual meeting of the British Association for the Advancement of Science (BAAS), where he heard a paper given by James Joule on experiments whose results contradicted caloric theory.

3. EQUIVALENCE OF HEAT AND WORK

Joule was the son of a very successful brewer in Manchester, England. He was educated privately and with ambitions to spend his life in science. His first papers dealt with "electromagnetic engines," and in the course of his experiments he focused on what we now call Joulean heating in wires carrying electric currents. This drew attention to the production of heat in both physical and chemical processes. From his experiments and a developing theory, he concluded that an electric generator enables one to convert mechanical power into heat through the electric currents induced into it. Joule also implied that an electromagnetic "engine" producing mechanical power would also convert heat into mechanical power. By 1843, he presented his first paper to the British Association for the Advancement of Science (BAAS) on the "mechanical value of heat." Joule believed that only the Creator could destroy or create and that the great agents of nature, such as electricity, magnetism, and heat, were indestructible yet transformable into each other. As his experiments progressed, he elaborated his mechanical theory of nature. Important to observers, and to Thomson, was Joule's idea that mechanical work and electricity could be converted into heat and that the ratio between mechanical work and heat could be measured. In a series of papers on the mechanical equivalent of heat presented annually to BAAS and politely received, Joule labored alone until the stark reactions of Thomson in 1847. Some scientists could not believe Joule's contentions because they were based on such small temperature changes. Presumably, their thermometers were not that accurate. However, Thomson understood that Joule was challenging caloric theory head on. Until 1850, Thomson continued publishing papers within a caloric framework and an account of Carnot's theory. He had finally located a copy of the treatise. Clearly, he began to doubt its validity, noting that more work on the experimental foundations of the theory were "urgent," especially claims that heat was a substance. It took Joule 3 years to convince Thomson of the correctness of his insight. Only in 1850 did Thomson concede that heat was convertible into work after he read Rudoph Julius Emmanuel Clausius's papers and those of William McQuorne Rankine on heat. He then reinterpreted some parts of his caloric papers that depended on heat as a substance in terms of the interconversion of heat and work. Also, he rederived and reinterpreted his expression for the maximum amount of work that could be extracted from the cycle of a perfect heat engine and other results and then gave his version of the second law of thermodynamics.

4. CONSERVATION OF ENERGY

Others had come to conclusions similar to those of Joule before Thomson. One of the most notable was Hermann von Helmholtz. Trained as a physician while studying as much mathematics and physics as possible, his doctoral dissertation was a thorough microscopic study of the nervous systems of invertebrates. His mentor was Hermann Mueller, who (along with Helmholtz's fellow students) was determined to erase vitalism from physiology. Their goal was to make physiology an exact experimental science and to base its explanatory structures in physics and chemistry. Helmholtz managed to continue his physiological research while a physician in the Prussian army, where he was dependent on his army postings and his early research on fermentation and putrefaction was cut short by the lack of facilities. He turned to a study of animal heat and the heat generated by muscular action and found that during contraction chemical changes occurred in working muscles that also generated heat. In the midst of this research, Helmholtz published "On the Conservation of Force." This paper displayed the extent of his mechanical vision of life and his commitment to a Kantian foundation for his physics. Central forces bound together the mass points of his theory of matter. These commitments resurfaced during the 1880s with his work on the principle of least action.

With considerable pressure brought to bear on the ministry of education and the Prussian military, Helmholtz was released from his army obligations and entered German academic life. He never practiced medicine again, yet his research over the subsequent two decades began with questions in physiology that expanded into studies on sound, hearing, and vision. His work on nerve impulses in 1848 developed into a study of electrical pulses and then into a study of the problem of induction before

growing into a critical overview of the state of the theory of electricity during the 1870s.

The closeness of Helmholtz's work in physics and physiology is evident in his paper on the conservation of force. He had actually stated the principle earlier (in 1845) in a paper reviewing the work of Humphry Davy and Antoine Lavoisier on animal heat. He concluded that if one accepts the "motion theory of heat," mechanical, electrical, and chemical forces not only were equivalent but also could be transformed one into each other. This assumption bound together results already known from chemistry, electricity, and electromagnetism. The structure of this review formed the outline of Helmholtz's 1847 paper. He added a theory of matter that he believed demonstrated the universality of the principle of conservation. Helmholtz began in metaphysical principles from which he extracted specific physical results. In his model, force was either attractive or repulsive. Change in a closed system was measured by changes in vis viva expressed as changes in the "intensity of the force," that is, the potential of the forces that acted between the particles, $\frac{1}{2}mv_1^2 - \frac{1}{2}mv_2^2 = \int_r^R \varphi dr$, where m is the mass of the particle, whose velocity changes from v_1 to v_2, and ϕ is the intensity of the force, whose measure is constructed by looking at changes in v^2. Because both velocities and forces were functions of the coordinates, Helmholtz expressed these last changes as $\frac{1}{2}md(v^2) = Xdx + Ydy + Zdz$, where $X = (x/r)\phi$. He demonstrated that for central forces, vis viva was conserved and, thus, so was the intensity of the force. He then expanded the argument to a closed system of such forces. The unifying concept in the paper was "tensional force" and its intensity ϕ. This allowed Helmholtz to extend his argument to electricity to obtain the "force equivalent" of electrical processes. In this case, the change in vis viva of two charges moving from distance r to R apart was $\int_1^2 \frac{1}{2}md(v^2) = -\int_r^R \varphi dr$, where he identified ϕ with Gauss's potential function, to which Helmholtz gave physical meaning. However, he did not give this entity a new name. He was reinterpreting physically a well-known approach used when mathematizing a physical problem. Unlike mathematicians, Helmholtz drew important physical conclusions from his equations, including the Joulean heating effect. He also included Franz Neumann's work on electrical induction in his conceptual net. Helmholtz used examples from chemistry and electricity to argue that heat was a measure of the vis viva of thermal motions, not a substance. He used a broad range of experimental results across the sciences to illustrate the plausibility of his contentions.

Helmholtz had difficulty in getting these speculations into print. They eventually appeared as a pamphlet. The thrust of the speculations went against prevailing standards in the emerging field of physics. German physicists had rejected grand speculative visions of nature and its operations for precision work in the laboratory, from which they assumed theoretical images would emerge. The model of this new approach was Wilhelm Weber. In Weber's case, precise, painstaking numerical results, analyzed in detail for possible sources of error, led to a center-of-force model that explained the range of electrical phenomena he had researched experimentally. Neumann was known for his careful experiments on crystals and their optical properties. Helmholtz presented his mechanical model before his analysis of the experimental work of others that ranged across disciplinary boundaries. This approach also dissolved the recently hard-won academic and specialist designation of physics in German academia.

Helmholtz was not the only German claimant to the "discovery" of the principle of conservation of energy. After his graduation in 1840, Julius Robert Mayer, a physician, became a doctor on a ship bound for the East Indies. Mayer noted the deeper color of venous blood in the tropics compared with that in Northern Europe. He interpreted this as more oxygen in the blood, with the body using less oxygen in the tropics because it required less heat. He speculated that muscular force, heat, and the chemical oxidation of food in the body were interconvertible forces. They were not permanent but were transformable into each other. On his return, Mayer learned physics to demonstrate the truth of this conviction and then set about to calculate the mechanical value of a unit of heat. To do this, he looked at the works on the thermal properties of gases and took the difference between the two principal specific heats as the amount of heat required to expand the gas against atmospheric pressure. This heat was transformed into work, and from this he computed the conversion factor between heat and work. Further memoirs followed, extending his use of available experimental results from the chemistry and physics of gases. Mayer even included the solar system. He argued that meteorites falling under gravity built up enormous amounts of mechanical force that yielded considerable amounts of heat on impact. This could account for the sun's seemingly inexhaustible supply of heat.

Mayer also encountered difficulties in publishing his work. His first paper of 1842, "Remarks on the

Forces of Inorganic Nature," was published in a chemistry journal rather than a physics journal. As with Helmholtz, the combination of metaphysics and calculations derived from other people's experiments was not acceptable in the physics community in the German states. In addition, Mayer did not have the academic credentials or connections of Helmholtz. In fact, Mayer had to underwrite the publication of his later papers. At the end of the decade, Mayer suffered personal tragedy and in 1850 tried to take his own life. He was confined to a series of mental institutions. However, from 1858 onward, he began to gain recognition for his work from his colleagues, including Helmholtz, and he received the Copley Medal from the Royal Society in 1871.

So far, there were several approaches leading to the statement of the first—and even the second—law of thermodynamics. None of these claimants replaced Carnot's and Clapyeron's work mathematically from a thermodynamic point of view except the work of Clausius. In his 1850 paper, Clausius began from the observation that the conduction of heat is possible only from a hotter to a colder body. This led him logically to replace Carnot's principle, that heat is conserved in the cycle of a perfect heat engine, to that of the equivalence of heat and work. As a consequence, when work is done, heat is consumed in any heat engine. Although he kept many of the mathematical forms developed by Clapyeron, Clausius assumed that the mathematical characteristics of the functions that entered into his equations determined the physical characteristics of his system. In analyzing the Carnot cycle, Clausius chose the perfect gas as his physical system and looked at the state of the gas as it traversed an infinitely small cycle. Following Clapyeron, he constructed the heat added or expended for each leg of the cycle and constructed an expression for the inverse of the mechanical equivalent of heat, A, as the ratio of the heat expended over the work produced. To construct an expression for the work done in completing the cycle, Clausius calculated the expressions for the heat added at each leg in the cycle and constructed

$$\frac{d}{dt}\left(\frac{dQ}{dV}\right) - \frac{d}{dV}\left(\frac{dQ}{dt}\right) = \frac{A \cdot R}{V},$$

where dQ is the heat added, V is the volume of the gas, and R is the gas constant. Arguing mathematically that Q could not be a function of A and t so long as they were independent of each other, the right-hand side of the equation must be zero. This brought the equation into the form of a complete differential,

$$dQ = dU + A \cdot R\left(\frac{a+t}{V}\right)dV,$$

where $a + t =$ the absolute temperature and $U = U(V, t)$, is the heat necessary for internal work, and depends only on the initial and final states of the gas. The second term on the right-hand side, the work done, depended on the path taken between the initial and final states of the gas. Clausius then introduced $1/T$, where T is the temperature that, as a multiplier of dQ, makes a complete differential of the form $Xdx + Ydy$. He applied this combination of physics and mathematics to the behavior of gases and saturated vapors, with the latter being of great interest due to the engineering importance of steam engines during the mid-19th century. In this analysis, he restated the results, some mathematically equivalent to those of Clapyeron and Thomson, but interpreted them in terms of the equivalence of heat and work. At the end of this first paper, he gave an initial statement of the second law of thermodynamics.

Therefore, there are several different paths and claimants to some statements equivalent to the first law of thermodynamics. During the decades immediately following 1850, many others were championed as "discoverers" of the principle of the conservation of energy among the engineers and scientists working on related problems. These claims came in the wake of a change in language from the equivalence of heat and work to the conservation of energy. Some of the assertions gave rise to bitter complaints, especially the contentions of Peter Guthrie Tait's, whose *Thermodynamics* trumpeted Joule as the "discoverer" of the equivalence of heat and work yet readily recognized Helmholtz's priority in the matter of the first law. Mayer's work was recognized belatedly by Helmholtz and Rankine against Joule's claims. Mayer published an estimate of the mechanical equivalent of heat a year before the appearance of Joule's first value. And so it went. A great deal was at stake in these disagreements, which reflected on industrial prowess as well as on intellectual authority in the growing political and industrial rivalry between Germany and Great Britain. Using hindsight, later historians also tried to claim some of the credit for other Europeans.

5. ENERGY PHYSICS

A decade before the development of the mechanical meanings of the conservation of energy, Thomson,

Clausius, and Helmholtz had extended its reach into electrostatics, thermoelectricity, the theory of electromagnetism, and current electricity. During the 1860s, one of the more important applications of this new approach was the establishment of an absolute system of electrical units. Again this involved national prestige given that the Germans already had their own system based on the measurements of Weber. The BAAS committee included Thomson, James Clerk Maxwell, and the young electrical engineer Fleming Jenkin. Maxwell and Jenkin measured resistance in absolute (mechanical) units and deduced the others from there.

So far, what existed were disparate statements and uses for energy, none of which shared a common language or set of meanings. Thomson and Tait created that language for energy in their 1868 textbook on mechanics, *A Treatise on Natural Philosophy*. The text was directed to the needs of their many engineering students. During the 19th century, numerous new technical terms were coined, many of which derived from ancient Greek, and energy was no exception. Derived from the Greek word *ενεργεια*, energy was commonly used for people who demonstrated a high level of mental or physical activity. Also, it expressed activity or working, a vigorous mode of expression in writing or speech, a natural power, or the capacity to display some power. More technically, Thomson and Tait used energy as the foundation of their 1868 text. They claimed that dynamics was contained in the "law of energy" that included a term for the energy lost through friction. They introduced and defined kinetic and potential energy and even claimed that conservation of energy was hidden in the work of Newton. William Whewell grumbled about the book's philosophical foundations, but it was translated into German as well as being widely used in Great Britain.

Conservation of energy was becoming the foundation for teaching physics as well as for new approaches in research in the physical sciences. Thomson considered that conservation of energy justified a purely mechanical vision of nature and tried to reduce all phenomena, including electromagnetic phenomena, to purely mechanical terms. In his early papers in electromagnetism, Maxwell used conservation of energy in constructing and analyzing mechanical models as mechanical analogues to electrical and magnetic phenomena. He then abandoned this mechanical scaffolding and established his results on general energetic grounds, including his electromagnetic theory of light. In defense of his approach to electromagnetism, Maxwell criticized Weber's action-at-a-distance approach as being incompatible with the conservation of energy. In 1870, Helmholtz began a systematic critique of Weber's electrical theory and repeated Maxwell's judgment that he located in Weber's use of velocity-dependent forces. Helmholtz had first stated this criticism in his 1847 paper on the conservation of force, and it became a starting point for alternative approaches to electromagnetism among German physicists. These included Neumann and Bernard Riemann, who developed the notion of a retarded potential. By 1870, this criticism undermined the validity of his work and Weber was forced to answer. He demonstrated that, indeed, his theory was compatible with conservation of energy.

Clausius was the first to try to develop an expression for the laws of thermodynamics in terms of the motions of the molecules of gases. His models were mechanical, whereas the later ones of Maxwell and Ludwig Boltzmann were statistical. Maxwell became more interested in the properties of gases, such as viscosity, heat conduction, and diffusion, relating them to statistical expressions of energy and other molecular properties. Boltzmann developed a statistical expression for the first and second laws of thermodynamics. During the late 19th century, thermodynamics itself was subject to great mathematical manipulation, which in some hands became so remote from physics that it once again became mathematics.

By the time of Maxwell's death in 1879, the conservation of energy had become a principle by which physical theories were judged and developed. This was particularly true in Great Britain, where "energy physics" propelled the theoretical work of most physicists. Much of this work, such as that of Oliver Lodge, was directed toward the development of Maxwell's electromagnetism, although the work was expressed as mechanical models. Lodge also drew up a classification of energy in its varied physical forms, from the level of atoms to that of the ether. He also used an economic image for energy, equating it with "capital." He extolled John Henry Poynting's reworking of Maxwell's *Treatise* in 1884 and saw it as a new form of the law of conservation of energy. In his treatment of the energy of the electromagnetic field, Poynting studied the paths energy took through the ether. He investigated the rate at which energy entered into a region of space, where it was stored in the field or dissipated as heat, and found that they depended only on the value of the electric and magnetic forces at the boundaries of the space. This work was duplicated by Oliver Heaviside,

who then set about reworking Maxwell's equations into their still familiar form. The focus was now firmly on the "energetics" of the electromagnetic field.

The emphasis on energy stored in the ether separated British interpretations of Maxwell's work from those on the continent, including that of Heinrich Hertz. However, energy in the German context became the foundation for a new philosophy of and approach to the physical sciences, energetics. Initiated by Ernst Mach, energetics found its most ardent scientific advocate in Wilhelm Ostwald. The scientific outcome of Ostwald's work was the development of physical chemistry. Yet his dismissal of atoms led to clashes with Boltzmann, and during the 1890s his pronouncements drew the criticism of Max Planck. Planck had, since his dissertation during the 1860s, explored the meaning of entropy, carefully avoiding atoms and molecules. He was forced into Boltzmann's camp in his research on black body radiation, essentially a problem of the thermodynamics of electromagnetic radiation. Planck was not alone in understanding the importance of the problem; however, other theorists were tracking the energy, whereas Planck followed the entropy changes in the radiation.

6. ENERGY IN 20TH CENTURY PHYSICS

As an editor of the prestigious journal *Annalen der Physik,* Planck encouraged the early work of Albert Einstein, not only his papers in statistical mechanics but also those on the theory of special relativity. Planck then published his own papers on relativistic thermodynamics, even before Einstein's 1907 paper on the loss of energy of a radiating body. Two years before this, Einstein had investigated the dependence of the inertia of body on its energy content. In this case, the mass of the body in the moving system decreased by E/c^2, where E is the energy emitted as a plane wave and c is the velocity of light. Einstein concluded that "the mass of a body is a measure of its energy content." Einstein constructed his relativistic thermodynamics of 1907 largely on the foundation of Planck's work and reached the conclusion that $dQ = dQ_0\sqrt{(1 - v^2/c^2)}$, where dQ is the heat supplied to the moving system, dQ_0 is that for the system at rest, and v is the velocity of the moving system to that at rest.

Thus, the law of conservation of energy was integrated into relativity theory. However, its passage through the quantum revolution was not as smooth. In 1923, Niels Bohr expressed doubts about its validity in the cases of the interaction of radiation and matter. This was after Arthur Holly Compton explained the increase in the wavelengths of scattered X rays using the conservation of energy and momentum but treating the X rays as particles. The focus of Compton's research, the interaction of radiation and atoms, was shared with many groups, including that around Bohr in Copenhagen, Denmark. Bohr did not accept Einstein's particle theory of light, even with the evidence of the Compton effect. By using the idea of "virtual oscillators" that carried no momentum or energy, Bohr, H. A. Kramers, and John Slater (BKS) hoped to maintain a wave theory of the interaction of radiation and matter. To do so, they had to abandon the idea of causality, the connection between events in distant atoms, and the conservation of energy and momentum in individual interactions. Struggling with the meanings of the "BKS theory" became crucial in the development of Werner Karl Heisenberg's formulation of quantum mechanics. However, the BKS theory was bitterly opposed by Einstein as well as by Wolfgang Pauli, who, after a brief flirtation with the theory, demonstrated its impossibility using relativistic arguments.

Conservation laws were reinstated at the center of physical theory. Conservation of energy became the criterion for abandoning otherwise promising theoretical approaches during the late 1920s and early 1930s in quantum field and nuclear theory. In the latter case, Pauli posited the existence of a mass-less particle, the neutrino, to explain beta decay. The neutrino remained elusive until the experiments of Ray Cowan and Fred Reines during the 1950s. Conservation of energy also governed the analysis of the collisions of nuclear particles and the identification of new ones, whether from cosmic rays, scattering, or particle decay experiments. Relativistic considerations were important in understanding the mass defects of nuclei throughout the atomic table. The difference between the measured masses and those calculated from the sum of the masses of their constituent particles was explained as the mass equivalence of the energy necessary to bind those components together.

7. ENERGY CONSERVATION BEYOND PHYSICS

Beyond physics, conservation of energy became a foundation for other sciences, including chemistry

and meteorology, although in this case the second law of thermodynamics was more powerful in explaining vertical meteorological phenomena. In the natural sciences, the most dramatic demonstration of the conservation of energy's usefulness was in the study of animal heat and human digestion. The production of heat as synonymous with life can be seen in the aphorisms of Hippocrates during the fifth century BCE. In Galen's physiology, developed during the second century AD, the source of bodily heat was the heart. Heat was carried in the arterial system throughout the body, where it was dissipated through the lungs and the extremities. Aspects of Galen's physiology remained in William Harvey's explanation of the function of the heart. Yet Harvey's experiments, published in 1628 as *Anatomical Experiments on the Motion of the Heart and Blood in Animals,* demonstrated the circulation of the blood and the pumping action of the heart. This mechanical function of the heart became part of Descartes's mechanical philosophy that changed debates over the source of heat in the body. Was it generated in the heart or from friction as the blood flowed through the vessels of the body? Harvey's argument that the lungs dissipated the heat from the body fitted into this mechanical vision of the body.

During the 18th century, Descartes's mechanistic vision of the body was replaced by vitalism based on chemical experiments. This was the vitalism that Helmholtz sought to root out from medicine during the 1840s. The experiments were those performed on "airs" consumed and expended by plants that vitiated or revived the atmosphere. Combined with these experiments were others showing that combustion and respiration made the atmosphere less able to sustain life, usually of birds and small mammals. After the introduction of Lavoisier's reform of chemical nomenclature, which also carried with it a new theory of chemistry and chemical reactions, experiments on animal chemistry became more quantitative and lay within his new theoretical framework. Treadmill studies by Edward Smith in 1857 demonstrated that muscles derived energy, measured by heat output, from foods not containing nitrogen. However, Justus Leibig's idea that bodily energy derived from muscle degradation was not easily deposed. It took the mountain-climbing experiments of Adolph Fick and Johannes Wislicenus in 1866 and the analysis of them by Edward Frankland to finally displace Leibig's theory. Frankland pointed out that Fick and Wislicenus's results could be interpreted by looking at the energy involved in combustion in a unit of muscle tissue and considering the work output the equivalent of this energy. Conservation of energy was necessary in understanding the dynamics of the human body. Frankland also showed that the energy the body derived from protein equaled the difference between the total heat of combustion and the heat of combustion of the urea expelled, with the urea being the pathway for the body to rid itself of the nitrogen in proteins. Thus, conservation of energy entered the realm of nutrition, biophysics, and biochemistry. In later experimental studies, W. O Atwater measured, by a whole body calorimeter, that the heat given off by the human body equaled the heat released by the metabolism of the food within it. Even when the complexities of the pathways of amino acids and other sources of nitrogen were understood, along with their routes through the digestive system, conservation of energy remained the foundation for interpreting increasingly complex results.

Thus, conservation of energy has proved to be crucial for analyzing the results of experiments across—and in some cases beyond—the scientific spectrum. There have even been experiments—so far unsuccessful—aimed at measuring the mass of the soul on its departure from the body.

SEE ALSO THE FOLLOWING ARTICLES

Heat Transfer • Mechanical Energy • Storage of Energy, Overview • Thermodynamic Sciences, History of • Thermodynamics, Laws of • Work, Power, and Energy

Further Reading

Adas, M. (1989). "Machines as the Measure of Man." Cornell University Press, Ithaca, NY.

Caneva, K. L. (1993). "Robert Mayer and the Conservation of Energy." Princeton University Press, Princeton, NJ.

Cardwell, D. (1971). "From Watt to Clausius." Cornell University Press, Ithaca, NY.

Carpenter, K. J. (1994). "Protein and Energy: The Study of Changing Ideas in Nutrition." Cambridge University Press, Cambridge, UK.

Cassidy, D. (1991). "Uncertainty: The Life and Times of Werner Heisenberg." W. H. Freeman, San Francisco.

Cooper, N. G., and Geoffrey, W. B. (eds.). (1988). "Particle Physics: A Los Alamos Primer." Cambridge University Press, Cambridge, UK.

Einstein, A. (1905). Does the inertia of a body depend upon its energy content? *In* "The Collected Papers of Albert Einstein" (A. Beck Trans, Ed.), vol. 2, Princeton University Press, Princeton, NJ.

Einstein, A. (1907). On the relativity principle and the conclusions drawn from it. *In* "The Collected Papers of Albert Einstein" (A. Beck Trans, Ed.), vol. 2, Princeton University Press, Princeton, NJ.

Everitt, C. W. F. (1975). "James Clerk Maxwell: Physicist and Natural Philosopher." Scribner, New York.

Helmholtz, H. (1853). On the conservation of force. *In* "Scientific Memoirs" (J. Tyndall and W. Francis, Eds.). Taylor & Francis, London. (Original work published 1847.)

Smith, C. (1998). "The Science of Energy: A Cultural History of Energy Physics in Victorian Britain." University of Chicago Press, Chicago.

Smith, C., and Norton Wise, M. (1989). "Energy and Empire: A Biographical Study of Lord Kelvin." Cambridge University Press, Cambridge, UK.

Conservation of Energy, Overview

GORDON J. AUBRECHT II
The Ohio State University
Columbus, Ohio, United States

1. Energy Preliminaries
2. The Physical Principle of Energy Conservation
3. Thermal Systems and Energy
4. Energy Conservation as Preservation of Resources

Glossary

conservative force The force that allows a potential energy to be constructed. A conservative force that acts along a closed path does no work. See potential energy.

district heating The use of waste heat from energy generation for space heating within a specified distance of the generating plant.

first law of thermodynamics The statement that the total internal energy of a system is changed both by heat transferred into or out of a system and the work done by or on a system. The principle of conservation of energy was at first applied only to mechanical systems (conservation of mechanical energy). When it was realized that thermal energy was another form of energy, the reason for the "disappearance" of energy from frictional work was realized.

heat Thermal energy that is transferred between reservoirs at different temperatures.

hybrid An automobile that has both a gasoline motor and an electric motor. It is more fuel efficient than a pure gasoline engine because braking recharges the battery and because the gasoline motor can be shut off when it is not accelerating the car.

industrial ecology An approach to design of industrial systems that uses analogies to biological systems to minimize energy and resource demand. The focus of industrial ecological design is thus on the physical parameters such as minimization of resource flows and efficiency of energy budgets.

potential energy A potential energy $U(\mathbf{r})$ can be constructed as

$$U(\mathbf{r}_2) - U(\mathbf{r}_1) = -\int_{\mathbf{r}_1}^{\mathbf{r}_2} \mathbf{F} \bullet d\mathbf{r}.$$

The conservative force F does work that is independent of the path taken. In this case, only the end points of the motion matter, and because only the difference in potential energy is defined physically, the potential energy $U(\mathbf{r})$ may only be defined up to an overall constant. In many cases, this constant is fixed by the choice $U(\mathbf{f}) = 0$, as is usually the case for the electrical potential energy between two charges q_1 and q_2:

$$U(\mathbf{r}) = \frac{1}{4pe0}\frac{q_1 q_2}{\mathbf{r}}.$$

principle of conservation of energy The idea that the total energy of the universe is conserved (it is neither created nor destroyed). The homogeneity of the universe with respect to time translations is a symmetry of nature. As a consequence of this symmetry, the universe's energy is conserved. For example, it is axiomatic for science that an experiment, prepared identically, will return the same result no matter when the experiment is performed.

second-law efficiency The actual work performed divided by the available work. The available work is the maximum amount of work along any path that a system (or a fuel) can attain as it proceeds to a specified final state that is in thermodynamic equilibrium with the atmosphere.

Second Law of Thermodynamics There are many formulations of the second law. A useful one is that there is no way to get work from a single temperature reservoir (one at a fixed temperature). Clausius proved that systems that are isolated will move toward equilibrium with their surroundings.

thermal systems A system that depends on the temperature. Heat transfer is a central feature of thermal systems.

waste heat Thermal energy transferred to the low-temperature reservoir in a thermal engine.

The physical principle of conservation of energy always holds. It reflects the homogeneity of physical laws with respect to time. In the context of thermal systems, this principle is known as the First Law of Thermodynamics. Energy obtained through thermal transformations is subject to a maximal efficiency due to the second law of thermodynamics, which severely constrains the utility of thermal energy as it is transformed to other kinds of energy. Not all sorts

 © 2004 Elsevier Inc. All rights reserved.

of energy have the same ease of use, even though it is always possible to transform one into another. High-quality energy is the most versatile, and thermal energy the least versatile. Energy conservation in ordinary speech refers to preservation of resources the provision of energy resources more efficiently or at lower cost. Many opportunities still exist for reducing costs by raising efficiencies, by encouraging cooperation so that waste is minimized, and by rethinking traditional designs. The transportation and agricultural systems may have the best prospects for increasing energy efficiency.

1. ENERGY PRELIMINARIES

Energy is defined as the capacity to do work, but as a practical matter, energy is difficult to measure with a machine (but it is easy to calculate). Work is defined by the differential relation

$$dW = \mathbf{F} \bullet d\mathbf{r},$$

where $d\mathbf{r}$ is an infinitesimal displacement of the massive object subject to the force **F**. If the force is constant, we find that the work done in displacing the object by $\Delta\mathbf{r}$ is

$$\Delta W = \mathbf{F} \bullet \Delta\mathbf{r} = \mathbf{F}|\Delta\mathbf{r}|\cos\theta,$$

where θ is the angle between **F** and $\Delta\mathbf{r}$. In many texts, this relation is simplified to read

$$\Delta W = \mathbf{F}_{||} d = \mathbf{F} d_{||},$$

where $\mathbf{F}_{||}$ is the component of the force along the displacement direction, d is the total distance traveled, and $d_{||}$ is the displacement component along the direction of the force. A key to understanding energy is to recognize that different forms of energy may be transformed into other forms of energy. In some cases, the work that has to be done against a force to move an object from one particular point A to another particular point B is the same, no matter what way the work is done. The work does, of course, depend on which points are used as end points, but not on the path traversed between them. In such a case, the force is called conservative. The gravitational force is an example of a conservative force. For such a conservative force, work may be done against the force to move a massive object from one position to another. This work represents the change in energy of the massive object. For example, a person may pick up an object that is on the floor (such as a fork) and put it on a table. If the external force on the object is **F** (in this case, **F** would be the gravitational force on the fork, otherwise known as the fork's weight), the work the person does opposing **F** to pick the fork up is

$$-\int_{\mathbf{r}_1}^{\mathbf{r}_2} \mathbf{F} \bullet d\mathbf{r}.$$

If the fork were to fall off the table onto the floor, we would expect that work to reappear somehow (as some other form of energy, perhaps). It has the potential to reappear if the fork is put into a position to fall, and so this energy is known as gravitational potential energy.

In general, there are many types of potential energy; the general potential energy of a particle is defined as

$$U(\mathbf{r}_2) - U(\mathbf{r}_1) = -\int_{\mathbf{r}_1}^{\mathbf{r}_2} \mathbf{F} \bullet d\mathbf{r}.$$

Potential energy may only be defined for a conservative force. For a conservative force **F**, this integral is independent of the path and can be expressed in terms of the end points only; here these are the starting and ending positions $\mathbf{r}_1$ and $\mathbf{r}_2$. It should be clear from the definition of potential energy that only differences in potential energy matter.

Energy in science is measured in the International System of Units as the product of the units of force and distance (newtons times meters); this energy unit is given the special name joule (1 J = 1 N m), but many other energy units exist. The most common are the British thermal unit (1 Btu = 1055 J), the kilowatt-hour (3.6×10^6 J = 3.6 MJ), and the kilocalorie [food calorie] (1 kcal = 4186 J). The quadrillion Btu ("quad") is often used to describe American energy use. In 2001, this was 98.5 quads (1.04×10^{20} J, or 104 EJ).

We may recognize two uses of the phrase energy conservation. One use is scientific and refers to the physical principle taught in school and essential to contemporary understanding of nature; this is discussed in Section 2. A background in thermal physics, which bridges the two uses of the word, is sketched in Section 3. The other use of the word refers to the preservation or wise use of energy resources and is discussed in detail in Section 4.

2. THE PHYSICAL PRINCIPLE OF ENERGY CONSERVATION

The impetus for the principle of energy conservation originally is the observation that in simple systems on Earth, one may define an energy of position (the gravitational potential energy) and connect it to

another form of energy, kinetic energy. For a simple mechanical system with no friction, using Newton's second law ($\mathbf{F} = m\mathbf{a}$), we recast the work as

$$\int_{r_1}^{r_2} \mathbf{F} \bullet d\mathbf{r} = \int_{r_1}^{r_2} m\mathbf{a} \bullet d\mathbf{r} = m \int_{r_1}^{r_2} \mathbf{a} \bullet d\mathbf{r}$$
$$= m \int_{t_1}^{t_2} \mathbf{a} \bullet \frac{d\mathbf{r}}{dt} dt = m \int_{t_1}^{t_2} \mathbf{a} \bullet v dt$$
$$= m \int_{v_1}^{v_2} v \bullet dv = \frac{1}{2} m v_2^2 - \frac{1}{2} m v_2^1.$$

The quantity $T = \frac{1}{2} mv^2$ is called the kinetic energy (it represents energy associated with the motion of a massive object). With this nomenclature, we find from the definition of potential energy that

$$U(\mathbf{r}_2) - U(\mathbf{r}_1) = -[T_2 - T_1],$$

and rewriting slightly, we obtain

$$T_2 + U(\mathbf{r}_2) = T_1 + U(\mathbf{r}_1).$$

Note that the quantity on the left refers only to position 2, while that on the right refers only to position 1. Since 1 and 2 were completely unspecified, this equation states a general result: in such a simple system, the quantity $T + U$ must be a constant (or conserved). This is referred to as conservation of mechanical energy (and of course, "conservative force" is a retrodiction that follows from the observation of conservation of mechanical energy).

More generally, the total potential energy is the sum of all the potential energies possible in a given situation. There is no new physics in the principle of conservation of mechanical energy. Everything follows from Newton's laws (refer back to the introduction of kinetic energy and potential energy).

It was natural to physicists to expand this idea by including every possible form of energy and to believe that the total amount of energy in a closed system is conserved. As a consequence of work by the German mathematical physicist Emmy Noether, it became clear that each conservation law (such as that of energy or momentum) is connected to a symmetry of the universe. The law of conservation of momentum is connected to translation invariance, the law of conservation of energy to temporal invariance. This means, respectively, that moving an experiment in a reference frame and repeating it when prepared identically has no effect of the observed relations among physical quantities, and that doing an experiment prepared identically at different times should produce identical results.

From these observations it is clear that relations among reference frames are essential to consider if discussing conservation laws. In the early 20th century, Albert Einstein worked to understand such relations in the context of electromagnetism and extended it beyond electromagnetism to more general physical situations. Einstein created two theories, the special theory of relativity (1905) regarding transformation among inertial reference frames, and the general theory of relativity (1915), which allowed extension of special relativity to accelerated (non-inertial) reference frames.

In special relativity, in an inertial frame of reference an energy-momentum four-vector p^μ consisting of a time component (E/c) and spatial momentum components p_x, p_y, and p_z, may be defined that satisfies $p^\mu p_\mu = m^2c^2$, which is an expression of energy conservation in special relativity (written in more conventional notation as $E^2 - p^2c^2 = m^2c^4$). That is, the fact that the vector product of two four-vectors is invariant leads to conservation of energy. In special relativity, there is a special frame of reference for massive objects known as the rest frame; in that frame $p = 0$, so the relativistic energy equation is $E = mc^2$, a rather widely known result. All massive objects have energy by virtue of possessing mass, known as mass-energy.

The relativistic energy relation $E^2 = p^2c^2 + m^2c^4$ may be recast as $E = \gamma mc^2$ for massive objects (and, of course, as $E = pc$ for massless objects), where the relativistic momentum is $\mathbf{p} = \gamma m\mathbf{v}$ and $\gamma = \left[1 - \left(\frac{v}{c}\right)^2\right]^{-1/2}$. In this case, the kinetic energy is $T = (\gamma - 1)mc^2$ because T plus mass-energy (mc^2) must be equal to total energy E.

In general relativity, the connection to the Noether theorem is clearer, as one defines a so-called stress-energy tensor, $T^{\mu\nu}$. The stress-energy tensor describes momentum flow across space-time boundaries. The flow (technically, flux) of the stress-energy tensor across space-time boundaries may be calculated in a covariant reference frame and the covariant derivative of the stress-energy tensor turns out to be zero. This means that energy is conserved in general relativity.

There is no proof of the principle of conservation of energy, but there has never so far been disproof, and so it is accepted. It is a bedrock of modern physics. In physics, as in other sciences, there is no possibility of proof, only of disproof. The cycle of model (hypothesis), prediction, experimental investigation, and evaluation leads to no new information if the experimental result is in accord with the prediction. However, if there is disagreement between prediction and experiment, the model must be replaced. Over the historical course of science, faulty

models are thrown away and the surviving models must be regarded as having a special status (until shown to be false as described earlier). Special relativity, general relativity, and conservation of energy are theories that have withstood the test of time. (In fact, by superseding Newton's law of universal gravitation, Einstein's theory of General Relativity shows where Newton's successful theory's boundaries of applicability are.)

3. THERMAL SYSTEMS AND ENERGY

As mentioned in Section 1, energy comes in many forms, all of which may be transformed into other forms of energy. Someone who knew only the physics principle of conservation of energy could be forgiven for wondering why people were speaking of energy conservation and meaning something else entirely. Why was generalization of mechanical energy conservation so difficult? The reason lies with the characteristics of thermal systems.

3.1 Thermal Energy

To help clarify the issues involved, let us consider a simple system: a box resting on flat ground to which a rope is attached. The rope is pulled by a person so that the force on the box is constant. Anyone who has actually done such an experiment (a class that contains virtually the whole human population of Earth) knows that, contrary to the assertion that the kinetic energy increases, it is often the case that the speed of the box is constant (the kinetic energy is constant).

What is missing is consideration of the friction between the box and the ground. Figure 1 shows the situation. In Fig. 1A, the rope's force on the box (caused by the person pulling the rope) is shown and the box is moving along the ground. However, many forces that act on the box are not shown. Figure 1B shows the free body diagram for the box, including all the forces that act. The box has weight, **W**, which must be included. The ground pushes directly upward to prevent the box from falling through the surface with a force $\mathbf{F}_{\text{ground on box}}$ and exerts a frictional force **f** on the box; and, of course, a force is exerted by the person on the rope, and hence the rope exerts a force on the box, **F**.

The presence of the frictional force is what is responsible for the lack of change in the kinetic

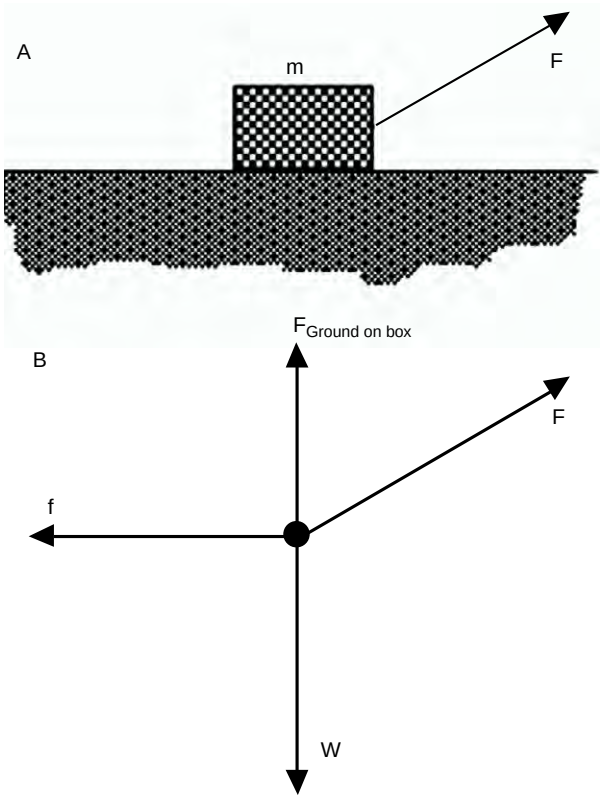


FIGURE 1 (A) A box being pulled along the ground. (B) A free body diagram showing all forces acting.

energy. That is, the speed can be constant because the horizontal components of the two forces add to zero. So where has all the work done gone? It has apparently disappeared from the mechanical energy. In fact, the work is still there, but thermal energy, the random motion of the atoms in the box has increased at the expense of the work done. In Fig. 2, an energy bar chart representation shows what has happened: The thermal energy of the system has increased.

Thermal energy is a new form of energy that must be considered in working with conservation of energy. In fact, the consideration of thermal energy was first codified by the physicist Sadi Carnot in the early part of the 19th century. We might call the 19th century the century of thermal energy because of the work of other giants of thermal physics such as Boltzmann, Clausius, Joule, Kelvin, Maxwell, Planck, and many others who created the science of thermodynamics.

3.2 Thermodynamics: The Zeroth Law

Systems involving measurements of temperature or exchanges of thermal energy are thermal systems.

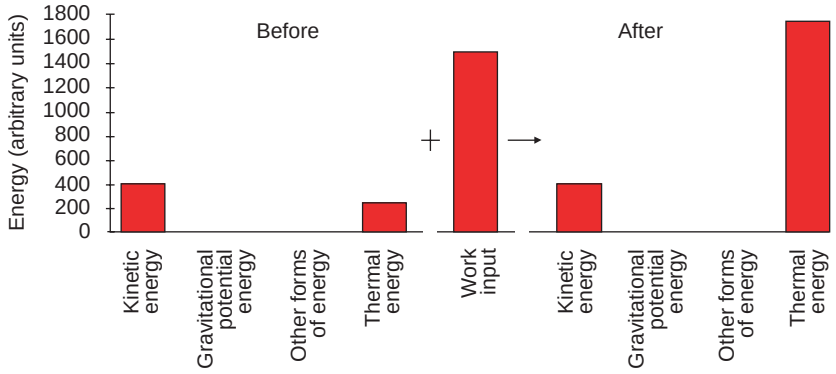


FIGURE 2 An energy bar chart representation of the change in energy of the box because energy is brought into the box system from outside by the rope.

Thermal systems are characterized by their temperatures. Temperature is determined by the mean of the random motions of the atoms making up a material. A device that measures temperature is known as a thermometer. Temperatures are measured by assigning numbers to systems we designate as cold or hot and by using the idea of thermal equilibrium.

If a spoon from the drawer is put into a pot of boiling water, thermal energy will flow from hot water to cooler spoon. Thermal energy is exchanged between the water and the spoon, but eventually the spoon and water do not exchange any more energy. Thermal equilibrium is the state in which there is no net interchange of thermal energy between bodies. The spoon comes to thermal equilibrium with the boiling water.

Temperatures are measured by allowing the thermometer and the system to come to thermal equilibrium. Two bodies are said to be in a state of thermal equilibrium if there is no net interchange of thermal energy between those bodies when they are brought into thermal contact. The measurement of the temperature of the thermometer therefore also gives a measurement of the temperature of the system.

The most common temperature scale used on Earth, and often used in science, is the Celsius scale. It sets the zero temperature at the freezing point of water and sets 100°C at the boiling point of water at sea level under standard conditions. The choice of the boiling end point is arbitrary scientifically. As virtually everyone living in the temperate regions of Earth knows from personal experience, there are temperatures below zero using this scale. A brisk day might have a temperature around 10°C, and a very hot day a temperature around 40°C. Human body temperature is 37°C. Comfortable room temperature is about 20°C.

In science, the arbitrariness of the Celsius scale makes it less useful in communication. Lord Kelvin devised a temperature scale, known as the absolute temperature or the Kelvin temperature scale, that begins with the lowest temperature possible in the universe as 0 K (absolute zero) and has a temperature of 273.15 K at the triple point of water (the temperature at which ice, water, and water vapor are in thermal equilibrium). With this definition, the kelvin interval is the same as the degree Celsius (1 K = 1°C).

3.3 Thermodynamics: The First Law

In the 18th century, steam engines began to contribute mechanical energy to replace human labor. Carnot began to study various engines that operate, as steam engines do, between two different temperatures. In a steam engine, burning fuel boils water to make steam, which pushes a piston, turning the steam's thermal energy to mechanical energy. That mechanical energy can be used in a machine. In early steam engines, the spent steam was vented to the air, the piston retracted, and the cycle began again (modern steam engines work slightly differently). The spent steam's thermal energy was released from the system.

The word "heat" refers to energy in transit between bodies at two different temperatures. It is one of the most misused words in physics. Heat is not a separate form of energy, but an artifact of the transfer of thermal energy to a body. There is no heat content to a body, as was first shown in 1798 in an experiment carried out by the American inventor Benjamin Thompson (1753–1814), also known as Count Rumford. As the person in charge of making cannon for the king of Bavaria, he was able to bore channels for the shot in cannon; there was need for

cooling water as long as the boring proceeded. There was no end to the amount of heat being transferred.

Any thermal engine operates between two temperatures. It is useful to think of the original steam at the higher temperature and the air at the lower temperature as temperature reservoirs, or large "containers" of thermal energy, so large that taking some or adding some leaves the system unchanged. A thermal engine runs through a cycle in which thermal energy is taken in at the high temperature reservoir, work is done, and thermal energy is exhausted at the low temperature reservoir.

Of course, energy is conserved in this cycle. The energy the system takes in from the high temperature is equal to the sum of the work done by the system and the energy the system exhausts to the low temperature reservoir. As applied to thermodynamics, the principle of conservation of energy is known as the first law of thermodynamics. From the first law of thermodynamics, no machine is possible that could create energy from nothing.

3.4 Thermodynamics: The Second Law

We define the efficiency of any energy transformation as

$$\text{Efficiency} = \frac{\text{energy output}}{\text{energy input}};$$

for a machine that produces work from other forms of energy, this relation would be

$$\text{Efficiency} = \frac{\text{work output}}{\text{energy input}}.$$

In the case of a heat (or thermal) engine, the first law of thermodynamics reads

$$\text{Energy input at } T_H = \text{Work output} + \text{Energy output at } T_L.$$

As a result, the efficiency is

$$\text{Efficiency} = \frac{\text{energy input at } T_H - \text{energy output at } T_L}{\text{energy input at } T_H},$$

clearly less than 100%. Note that the necessity of exhausting thermal energy as heat to the low-temperature reservoir necessitates a severe restriction on the efficiency of a heat engine. This also has deleterious consequences, because this so-called waste heat must be absorbed by the local environment (usually local water is used as the low temperature reservoir) and can adversely affect aquatic life.

Carnot studied a special cycle (Fig. 3) in which the thermal energy is absorbed at the high temperature, the system expands, then the system contracts as thermal energy is exhausted at the low temperature. This cycle is known as the Carnot cycle. By studying this special Carnot cycle, which was the most efficient cycle that could possibly exist between the two temperatures, Carnot could find the maximum possible efficiency in terms of the absolute temperatures of the reservoirs for a heat engine operating between these two temperatures.

The second law of thermodynamics, originally conceived by Rudolf Clausius, is the statement that any real thermal engine is less efficient than the Carnot cycle. (There are many other formulations of the second law.) The Carnot efficiency is found to be

$$\text{Maximum possible efficiency of a heat engine} = \frac{T_H - T_L}{T_H}.$$

Again, because of the need to exhaust thermal energy at the low temperature, the efficiency of even the ideal Carnot engine is restricted. For example, suppose that a heat engine (an electric generator) operates between a high temperature of 500°C and a low temperature of 20°C (room temperature). The maximum possible efficiency of operation is

$$\text{Maximum possible efficiency of a heat engine} = \frac{773.15\,\text{K} - 293.15\,\text{K}}{773.15\,\text{K}} = 0.621(62.1\%).$$

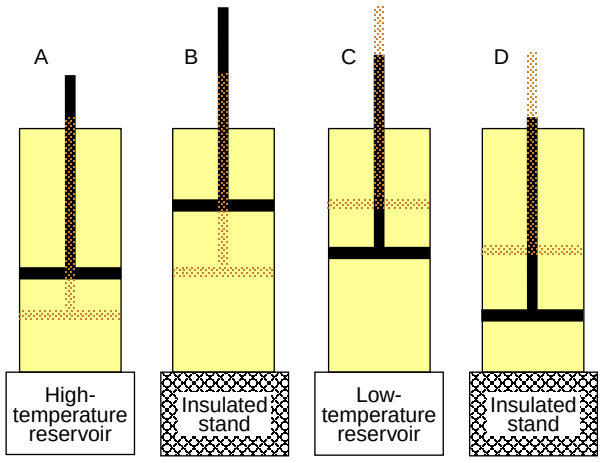


FIGURE 3 The four parts of the Carnot cycle. (A) The system expands, drawing in thermal energy, at absolute temperature T_H while it is attached to a high-temperature reservoir. (B) The system is isolated, so no energy enters or leaves, and the system expands further, lowering the absolute temperature from T_H to T_L. (C) At absolute temperature T_L, the system contracts as the piston pushes inward, transferring thermal energy to the low-temperature reservoir. (D) The system is isolated, so no energy enters or leaves, and the system contracts back to the original volume.

Real electric generators average about 33% efficiency, much lower than the maximum possible (Carnot) efficiency.

3.5 Quality and Second-Law Efficiency

It may be seen from the above that when transforming thermal energy into other forms of energy, the amount available is not the total energy available, but is smaller. However, if work is done by an agent, *all* the work is available; if electrical energy is available, *all* of it may be transformed directly into other, usable, forms of energy. Thermal energy is said to be low-quality energy because of its limitations in the conversion process, while work and electrical energy are high-quality energy.

Thermal energy may be used directly in space heating and in cooking, but if we want to run a motor with it, not all the thermal energy would be available to us. It would make sense to use forms of energy (if available) that could deliver all their energy to the motor. Conversely, if we want to increase the temperature of a room, it is silly to use high-quality energy to do it.

While the equation for efficiency is found using the first law of thermodynamics, the American Physical Society study "Efficient Uses of Energy" points out that the second law of thermodynamics, which shows that the Carnot engine operating between two temperatures is the most efficient, allows us to compare the actual efficiency to the theoretical maximum efficiency for producing a given amount of work; this is called the second-law efficiency for the system. For real thermal systems, the second-law efficiency measures how much room for improvement remains.

4. ENERGY CONSERVATION AS PRESERVATION OF RESOURCES

Often exhortations to conserve energy or other resources such as land or water are heard. Many people view the probable effect of the requested conservation measures in their daily lives as having to "do without" something desirable. Conservation might represent "personal virtue" through sacrifice; it involves paying a personal price. It may represent the religious idea of stewardship. Alternatively, they may see these measures as making their lives more complicated in a world that already seems full of intimidating tasks. They are interested in living their lives, in having life be as comfortable and convenient as possible.

Energy conservation here refers to the provision of the same level of goods or services with a smaller drain than at present on stored energy resources. In their common perceptions about conservation, those people who are wary of conservation efforts are both wrong and right. They are wrong because people who conserve energy are not sacrificing or doing without but are benefiting from increased efficiency or reduced costs. However, they are right because this reorientation may have a cost in personal convenience since it may involve additional time, thought, or change in habits.

In the following, we discuss possibilities for "energy conservation" with the understanding that, physically, energy is always conserved as discussed in Section 2; we are discussing here reduction of waste and increase of efficiency of the economic and physical system. The reliance on thermal systems for transforming energy makes the second-law efficiency a good measure of the efficacy of a proposed conservation measure. The following examples show how it has been or may in the future be possible to provide the same services at reduced monetary and energy cost.

4.1 Space Conditioning

Space heating and cooling are responsible for a considerable amount of energy use. In the United States in 1993, 4.3×10^{17} J of electricity, representing 1.3×10^{18} J of thermal energy, 3.9×10^{18} J of thermal energy from natural gas, and 1.3×10^{18} J of thermal energy from kerosene, fuel oil, and liquefied petroleum gas (a total of 6.5×10^{18} J = 6.5 EJ of thermal energy) was used for residential space heating. Another 1.50 EJ was used for residential air conditioning, of which 18% (0.26 EJ) was used for room air conditioners. This represents over 8% of national energy use. Commercial buildings in the United States in 1995 used 8.3 EJ of thermal energy as electricity, 2.1 EJ of thermal energy from natural gas, and 0.25 EJ of thermal energy from fuel oil, a total of 10.6 EJ of thermal energy, roughly 10% of national energy use. The great majority of energy use is for space conditioning (4.1 EJ) and lighting (3.3 EJ).

In the aftermath of the 1973 energy crisis, many buildings were built with more attention paid to energy-conserving features, resulting in a lowered energy use per area, but this impetus was lost with a return to cheap oil in the 1980s and 1990s, and in early 1990s commercial buildings were built to use

one-third more energy per square meter than average 1980s buildings.

Clearly, with over 10% of United States energy use committed to space conditioning, systematic savings could have a great impact. Savings are available from installation of replacement glazing, increased insulation, and replacement of antiquated heating and cooling systems.

Active and passive solar energy systems have been used for millennia to achieve space conditioning at little or no cost, and these methods continue to hold promise of reduced reliance on fossil fuels and uranium. Passive systems are characterized by construction methods that allow people to cool rooms during the day and heat by night through thermal storage—for instance, in masonry walls—of energy from the sun. Use of surrounding vegetation to help maintain house temperature has been common since humans began building shelters. Natural ventilation systems are also passive in nature. Active solar systems use some fossil energy to pump fluids, but the main input is free energy from the sun.

Heat pumps, especially groundwater-based heat pumps, use a heat engine in a clever way. Groundwater is about the same temperature the year around, and in the winter it is much warmer than the ambient air in colder climates, while in summer it is much cooler than the ambient air. In winter, the heat pump is run by taking in thermal energy at the groundwater low-temperature reservoir, doing work on it, and exhausting the thermal energy plus the work at the high-temperature reservoir (the room). Only the extra energy (the work input) must be supplied by fossil fuels. The advantage of this system is that the user pays only for the work done, and the extra energy used to heat the space is the "free" thermal energy in groundwater or the atmosphere. In summer, the heat pump works backward, taking thermal energy from the high-temperature reservoir (the room), doing work, and exhausting thermal energy to the low-temperature reservoir (the groundwater). Commercial freestanding heat pumps use the atmosphere as the reservoir and can be substantially less efficient than groundwater-based heat pumps.

In cold weather, much ambient thermal energy is transferred out of buildings, where it is desirable to keep it, to the outside through walls and windows. In warm weather, ambient thermal energy is transferred from outside to inside, where it is undesriable, again through the walls and windows. Advanced technology can contribute to reduced fossil energy use through design of various types of windows—low emissivity coatings, double- or triple-glazed windows with evacuated spaces, or smart window coatings that can act as an insulating shade. Aerogel-based insulating materials may also contribute to future energy savings in buildings, as they already have for appliances.

Air conditioners have become more efficient over the years since the 1970s energy crisis. The state of California has led the way by mandating increasingly efficient room and home air conditioning units, and the federal government has followed. The rollback of Clinton administration seasonal energy efficiency ratio (SEER) regulations by the incoming Bush administration (which may have been illegal, since the rules had been published in the *Federal Register*) was a setback to this trend. The payback time for the increased cost of a SEER-13 unit over a SEER-12 unit is a little over 1 year. Fortunately, because it has such a large population, California's energy-efficiency regulations tend to drag the rest of the nation along despite federal foot-dragging, and even the rolled-back SEER targets are far above the average SEER of 5 to 6 in 1994. The efficiency regulation might actually be looked at as an antipoverty regulation, because the least-efficient units (having the lowest purchase price) are bought by landlords, while the typically lower-income tenants must pay the electric bills. Wealthier Americans tend to buy the higher-SEER units because they themselves pay their electricity bills.

4.2 District Heating

In many European cities and some older American cities, electric utilities use the waste heat from the boilers or heat exchangers to keep buildings warm in the cooler months of the year. Hot water is circulated through pipes through nearby buildings, and cooler water returns to the utility. This is known as district heating because an entire region can be served by this system. In the United States, 5.6×10^{17} J (0.56 EJ) of energy was distributed as district heating in 1995.

Savings are possible because of economy of scale (one central heating system instead of many individual systems), the reduced maintenance per unit of output characteristic of a larger system, and because the waste heat would have to be disposed of by other means were it not used for this useful purpose. District heating is only feasible when there is a large but compact region having a relatively large population density through at least part of the day. This is more characteristic of European and Asian conditions than American or African ones.

4.3 Illumination

As noted earlier, commercial buildings use a substantial amount of energy for lighting, roughly 3% of total American energy use. Lighting is also an important part of household and commercial energy consumption. There are many ways of saving fossil energy. For example, commercial buildings may be built with daylighting features, in which the architect designs the space and the windows such as clerestories to provide light to workspaces during the day (clearly lamps are needed when it is not light enough outside, but in most locations days are not severely overcast most of the time).

Incandescent lighting, in which a filament is heated to a white-hot temperature and the resulting light is used, converts only about 5% of the electric energy into radiant energy. Ordinary fluorescent lighting is four times as efficient, converting about 20% of the electric energy into radiant energy. The development of electronic ballasts for fluorescent bulbs in the 1970s after the first energy crisis contributed significantly to fluorescents' energy efficiency advantage.

In the 1990s, compact fluorescent bulbs became available. Now these bulbs, while more expensive than incandescent bulbs to buy, deliver light for a longer time and use much less electricity and so are much cheaper to use. Consider a 100 W incandescent bulb that costs \$0.25 in a package of four bulbs for \$1, and a 23 W compact fluorescent that delivers about the same amount of light but costs \$4.95. The cheap incandescent bulb last on average 500 hours, while the compact fluorescent lasts 10,000 hours. That means it takes 20 incandescent bulbs to replace the compact fluorescent, at a cost of \$5 (the same as the price of the compact fluorescent bulb). However, over 10,000 hours of operation, compact fluorescent use costs (at a rather typical 8 cents per kilowatt-hour)

$$23\,\mathrm{W} \times 10,000\,\mathrm{hours} \times \$0.08/\mathrm{kWh} = 230 \times \$0.08 = \$18.40,$$

while incandescents' use cost

$$100\,\mathrm{W} \times 10,000\,\mathrm{hours} \times \$0.08/\mathrm{kWh} = 1000 \times \$0.08 = \$80.00.$$

Obviously, it is much more rational for a homeowner or businessperson to use the compact fluorescent. However, inertia—the resistance to change—and the high initial cost has limited compact fluorescents' market penetration, only about 10% of lighting sales. The Department of Energy has estimated that a complete switch to compact fluorescents would save the United States about 32 billion kilowatt-hours of electric energy annually, or 0.36 EJ of fossil fuel energy—the equivalent of the output of almost four 1000 MW generating plants.

4.4 Appliances

In the aftermath of the energy crises of the 1970 s and 1980 s, attention turned to energy use in appliances as a means of energy savings. Newer compressors, more effective insulation, and better design of refrigerators has decreased continuing expenses substantially. In 1973, the average refrigerator used over 2000 kWh/year. By 2000, this average was reduced to 1140 kWh/year, and the most energy-efficient units (volume $20\,\mathrm{ft}^3 \sim 0.6\,\mathrm{m}^3$) used just 440 kWh/year. Several factors contributed to increases in efficiency. In the aftermath of the energy crises, legislation mandated the posting of energy guides on appliances (including refrigerators) so consumers could make rational decisions on purchases. In addition, California introduced efficiency regulations for refrigerators (and the California market is so large, it made no sense for manufacturers to make a separate California model). Finally, a \$30 million prize was offered for development of the most energy-efficient refrigerator (the prize was won by Whirlpool).

Europeans have long paid substantially more for energy than their American counterparts, and appliance manufacturers have developed many cost savings that had not been adopted in the United States. With the consolidation of appliance manufacturers worldwide, more European energy-efficient technology has made its way west. Whirlpool, for example, is now manufacturing virtually identical energy-efficient washing machines in Europe and the United States.

A uniquely American idea is the EnergyStar rating system, which is an extension of the energy guides to appliances not covered by the legislation voluntarily entered into by the appliance manufacturers. Appliances meeting EnergyStar guidelines display the EnergyStar sticker and consumers can know that they are among the most efficient available. The criteria change with time; for example, for dishwashers to be eligible for EnergyStar designation prior to January 2001, their efficiency had to be greater than 13% above federal standards, while to earn the EnergyStar after January 2001, the dishwasher had to be over 25% more energy efficient. The efficiencies are achieved by locating heater

boosters inside the dishwashers and by use of less water per cycle.

Computers have been increasingly used in many households and businesses. The use of flat screens for televisions and computer monitors saves considerable energy compared to the CRTs typically used, which is why they were originally introduced for the notebook computers, which must often run from batteries. The sleep mode of computers and computer peripherals uses much less energy than would be used if the computers were fully "awake."

Some years ago, appliances had to "warm up" before they came fully on. Consumers disliked this wait time, and manufacturers of televisions, radios, and other appliances began to manufacture instant response units. Consumers liked the feature, but didn't realize the hidden costs. The units are turned on all the time and collectively use between 10% and 25% of the electricity used in households. Such devices have been called energy vampires. The problem may be limited to developed countries, but the cost is not a necessary one. Electronic transformer use would substantially decrease this drain, and they already are used in countries where mandated by law (they cost only 0 to 20 cents more per unit, but are not installed voluntarily by cost-conscious manufacturers). Computers use much more than the 10 W or so used by VCRs and TVs in their sleep mode, though, of course, the power is much lower than when the computer is awake.

4.5 Transportation

Automobile engines have a Carnot efficiency of around 60% but are about 25% efficient in practice. As long as the internal combustion (Otto) engine continues to be used, large increases in efficiency are unlikely, but sizable incremental savings are still available. Alternatives include electric vehicles and fuel cell vehicles. The transportation sector is important because transportation accounts for 27% of total United States energy use. Most of the energy used for transportation is petroleum based. That means that importing countries depend on the exporting countries' good faith and on their continued willingness to supply oil at a relatively cheap price. While the United States pays some of the highest prices for oil, both because domestic oil is more expensive than imported oil and because the average transportation cost is large, the price of gasoline at the pump is among the lowest in the world. This is mainly due to the difference in tax structures among countries.

In Europe, automobiles are generally smaller than in the United States because of the higher energy prices consumers pay for gasoline and diesel fuel (because of the high local taxes). Smaller cars use less material to manufacture and, because the weight is smaller, cost less to run. Many drivers in the United States are reluctant to purchase small cars because they are concerned about safety (this is apparently partly responsible for the boom in sport utility vehicle, or SUV, sales). Ironically, though people in SUVs are less likely to be killed or injured in a direct front, back, or side collision, the overall death rate is higher in SUVs than in smaller cars because of the large number of fatalities in rollover accidents, which have a relatively high probability of occurrence.

American cars are much lighter than in the 1970s, a direct result of the first energy crisis. A conventional suburban car gets a mileage around 30 mi/gal. It is clear that the weight of most cars could be reduced further. A countervailing trend is for more and more SUVs and pickup trucks to be sold. These vehicles weigh more than normal cars and as a result get worse mileage. By 2000, fewer than half of all vehicles purchased for personal transportation in the United States were ordinary automobiles. A large SUV might get a mileage of just 17 mi/gal, while a small pickup gets 25 mi/gal. The reason for the better car mileage is legislative—Congress adopted Corporate Average Fuel Economy (CAFE) standards after the first energy crisis to force manufacturers to raise mileage, and it worked. The rise in SUVs and pickups have pleased the manufacturers because these vehicles were not covered by the CAFE standards. In addition, the low average energy prices of the late 1990s lulled drivers into expecting cheap imported oil would continue to be available. Political pressure for increasing CAFE standards abated, and car mileage stagnated. As a result, the fleet average for American vehicles experienced a decrease in mileage. The lowest fleet average mileage was achieved about 1989. Even imported cars' mileage has slipped. On the basis of history, only legislative mandate will be successful in increasing mileage in the absence of external threats such as giant leaps in world oil prices.

Before the 1970s, most American cars were rear-wheel drive; now most are front-wheel drive, which is more energy efficient. Most modern automobile engines are more efficient overhead cam engines. However, much more energy could be saved by continuing to pursue ways to decrease drag, and the installation of advanced automatic transmissions that are controlled electronically. Other measures, some partly introduced include improved fuel injection

systems, improved tires, improved lubricants, and reduced engine friction. All these measures are good conservation measures because they come at no or low cost (low cost means short payback times).

California's quest for less-polluting cars led to designation of quotas for lower-emission vehicles. Many ideas for automobiles that run on electricity or energy stored in high-tech epoxy-based flywheels have been proposed. The high hopes California had for so-called zero-emission vehicles were apparently unwarranted, and few total electric vehicles have been designed or produced. There have been difficulties with battery design (conventional lead-acid automobile batteries weigh quite a lot). The nickel-metal hydride battery is the apparent technology of choice for electric vehicles. In addition, battery-powered cars do not perform well at low ambient temperatures, so these cars are not common outside the southwest. Another strike against battery-powered cars is the lack of charging infrastructure. And such vehicles are not really *zero* emission, because gas, oil, or coal must be burned to produce the electricity used to recharge the batteries.

California's quest has had the effect of boosting prospects for low-emission vehicles that are hybrid (combined gasoline engines and electric systems that work together). Hybrid features that save fuel include automatic shutdown of the engine at stops and regenerative braking (converting the work done in braking into energy stored in the battery). In hybrids, the gasoline engine runs only at or near peak efficiency, and excess energy recharges the batteries. Since the fuel is conventional and the electric system recharges in use, there is no need for a special infrastructure as is the case with pure electrics. Honda introduced the first commercial hybrid, the Insight, followed quickly by Toyota with the Prius. Honda has begun to sell Civic hybrid models.

Fuel cells, which are devices that combine fuel (ideally hydrogen) with oxygen from the air without combustion, are expected to supply automobiles with energy in the near future. If the hydrogen were produced by electrolysis using solar energy, such a vehicle really would involve zero emission (outside the water made from combining hydrogen and oxygen). Fuel cells have been used in the past for transportation, but usually liquid metal fuel cells in buses because of the relatively high temperature of operation (300 to 1200°C). The development of low-temperature proton-exchange membrane fuel cells have led the auto manufacturers to engineering development. These are typically 50% efficient at transforming fuel to energy, about twice as efficient as Otto engines. (Very small fuel cells are being used as an alternative to battery packs in some electronic devices.)

The fuel cell option suffers from a lack of an infrastructure to supply the hydrogen to vehicles and the problem of how to carry sufficient hydrogen. Hydrogen is a dilute gas, and it has a very much smaller amount of energy per volume than liquid fuels. Possible solutions to the storage problem range from mixing hydrogen into borides to pressured tanks to storage as hydrides in metal containers. However, many questions still surround the storage question. Alternatively, methanol (a liquid) can be chemically transformed to produce hydrogen, and this may be a more probable choice for the short term. The disadvantage is the presence of carbon dioxide in the exhaust stream.

4.6 Industry

Together, commercial and industrial energy uses total 54.3 EJ, of which 41 EJ is industrial energy use. Industry was hardest hit by rising energy prices in the aftermath of the energy crises and had the largest incentive to reform. Indeed, the industrial sector has saved a considerable amount of energy since then (Fig. 4 shows how the industrial sector energy use breaks down). Work has been done on design of variable-speed motors and control mechanisms for machines used in industrial processes. Motors are much more efficient than those of a generation ago. Compressors are especially important to consider, as many heavy industries use compressors for many different purposes.

Many industries have the same opportunity as commercial business for reduction in energy use in offices—installation of newer, more efficient HVAC systems, replacement of ballasts for fluorescent lighting, or installation of compact fluorescents in place of incandescents, and so on. Industries often buy their electricity on a "interruptible" basis, meaning that the utility can cut them off if demand from other users is high; the industry gets concessional electricity rates in return. The processes can be designed to result in "load leveling" for the utility's demand. For example, the industry can use high demand when the utility's other customers are near the low point of demand; this helps the utility by increasing base load slightly (base load is cheapest). Many energy management control systems have been designed that automatically shuffle load on and off in response to external conditions.

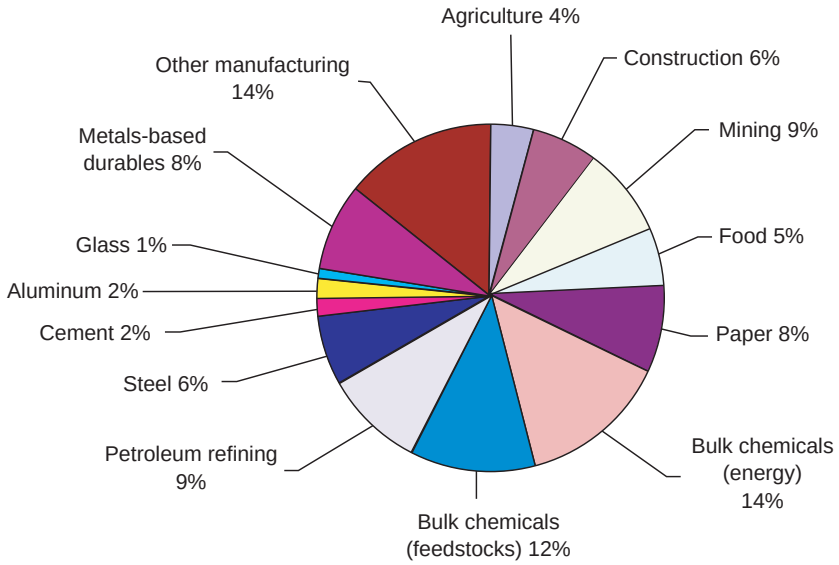


FIGURE 4 Breakdown of the energy use by industrial subsector 1992. From Interlaboratory Working Group, "Scenarios for a Clean Energy Future," Figure 5.1.

While much work has been done, opportunities remain for reducing industrial process energy use. In a 1992 study concerning reduction of carbon dioxide emissions by the United States, a National Academy expert panel identified billions of dollars in savings that would result from changing processes and materials with the aim of lowering carbon emissions (i.e., negative real cost), even in as mature an economy as was current in the United States at that time. Few of its recommendations have been adopted, mostly, it seems because of the attitude that we have not done it that way in the past (in other words, another case of inertia holding back rational change).

Opportunities remain, especially in mining technology, metals refining and working, and chemicals (large subsectors from Fig. 4). In steel processing, hot rolling of metals is being replaced by direct casting and rerolling in small hot strip mills as needed. Scrap to be melted can be preheated using furnace waste heat, reducing overall energy use. More than 50 such energy-saving measures have been identified in the steel industry alone.

Dow Chemical has been a leader in making waste streams into resource streams for other parts of the company or for other companies. This benefits society both by reducing waste emissions and reducing the impact of obtaining the resource. Dow has experienced economic returns over 100% on some programs. If such a well-managed company can find such opportunities, it is clear that practically all other companies can benefit. Experiences such as this and the success of the independent companies colocated in Kalundborg, Denmark, in exchanging materials to mutual benefit have led people to speak of a new field of study they call industrial ecology. There is a hope that bringing attention to these possibilities will result in greater savings and efficiencies in the future and that companies can learn from the best practices of other companies.

The science of chemical catalysis, originally strongest in the petroleum industry, has spread to other industrial sectors. Catalysts are chemicals that increase reaction rates without being themselves consumed in the process. Increases in reaction rates at low temperatures saves the application of thermal energy to raise the temperature (as reaction rates generally increase with temperature).

Glassmaking is an industry that uses a lot of energy per kilogram of output because of the necessity of melting the constituents of glass (or melting glass when it is recycled). New refactory materials have reduced energy use in the glass industry at no cost.

The agricultural industry is one area where penetration of ideas is slow; more energy conservation measures need to be implemented. Individual farmers are rugged individualists and apt to think in many instances that any change is for the worse. One promising avenue is for a combination of chemical and agricultural industry to grow chemical feedstocks

and to achieve ways to grow plastics in the field. For this to be entirely successful, negative attitudes held by the general public toward genetic manipulation may have to be abandoned.

4.7 The Utility Industry

Utilities are large-scale users of fossil fuels. In 2000, utilities used 42.6 EJ of mostly fossil energy, 41% of total U.S. energy use. Because they use so much energy, there is a built-in incentive to adopt the most energy-efficient equipment possible. While the average 33% efficiency in transforming thermal energy into electrical energy seems small, it is far improved from the original generators in the industry a century ago. Progress has generally been incremental, but the advent of integrated gas combined cycle power plants (known as IGCC plants) in the late 1980s and 1990s has boosted output efficiency substantially. Some combined cycle plants are operating near 60% efficiency, not far from the Carnot cycle limit. Gas turbine generators now use aircraft turbines almost unmodified and are much higher in efficiency than previous gas-fired turbines.

One way for the utility to save money is to even out its demand, as referred to briefly earlier. It is to the utility's benefit to assure that as much energy sold as possible is base load, the cheapest energy the utility can generate. The smaller the demand for high-cost peak energy, the better off the utility is. Therefore, utilities offer preferential rates to industrial users who are willing to have their electricity shut off or radically decreased when demand spikes from other users. This is known as load leveling.

Even individual users can sign up for interruptible rates if they are willing to allow the utilities to shut them (or their largest appliances such as air conditioners or water heaters) down as needed. Electronic devices can be installed that give the utility control over whether the individual interruptible appliances are on or off. The homeowner or business owner agrees to allow the utility to shut them off as needed when demand for electricity is high.

4.8 Cogeneration and Net Metering

By cogeneration, we mean generation of electricity as an incidental outcome of industrial production. For example, the paper industry uses a great deal of hot water, and produces the hot water in boilers. It could use the steam produced to run turbines and generate electricity. In the early days of electricity distribution, this occurred regularly, but with the formation of utility monopolies in the late 1800s, industrial producers were not allowed to sell their energy offsite. Given the obvious advantage (two gains for the price of one), restrictions were first loosened after the energy crises of the 1970s and 1980s, and then vanished altogether. Cogeneration has become more widespread since the 1980s. The utility typically signs a purchase agreement with the cogenerator with a contracted rate schedule. The sale of electricity contributes to the seller's bottom line.

The same dynamic is at work on the much smaller scale of the individual electric utility customer in the question of whether states have "net metering" laws. Cogeneration can occur at the scale of an individual home. Customers who install solar energy generators but remain connected to the grid for backup in states without net metering can decrease their electricity costs by using "home-generated" electricity in place of buying, but they cannot sell any excess back (the utility gets it for free). In states with net metering, the excess electricity is bought by the utility (albeit at a lower rate than the customer usually pays for electricity), in some cases leading to reverse bills—utility payments to customers for energy sold back that exceed the value of energy sold to the customer.

SEE ALSO THE FOLLOWING ARTICLES

Cogeneration • Conservation Measures for Energy, History of • District Heating and Cooling • Mechanical Energy • Storage of Energy, Overview • Thermodynamic Sciences, History of • Thermodynamics, Laws of • Work, Power, and Energy

Further Reading

Carnahan, W., Ford, K. W., Prosperetti, A., Rochlin, G. I., Rosenfeld, A., Ross, M., Rothberg, J., Seidel, G., and Socolow, R. H. (eds.). (1975). "Efficient Use of Energy: The APS studies on the Technical Aspects of the More Efficient Use of Energy." American Institute of Physics, New York.

Department of Energy (1995). "Buildings and Energy in the 1980s." U.S. Government Printing Office, Washington, DC.

Department of Energy (1995). "Household Energy Consumption and Expenditures, 1993." U.S. Government Printing Office, Washington, DC.

Department of Energy (1997). "1995 Commercial Buildings Energy Consumption Survey." U.S. Government Printing Office, Washington, DC.

Department of Energy (2001). "Annual Energy Review, 2000." U.S. Government Printing Office, Washington, DC.

Interlaboratory Working Group (2000). "Scenarios for a Clean Energy Future." Oak Ridge National Laboratory, Oak Ridge,

TN, ORNL/CON-476, Lawrence Berkeley National Laboratory, Berkeley, CA, LBNL-44029, and National Renewable Energy Laboratory, Golden, CO, NREL/TP-620-29379.

International Energy Agency (2001). "Things That Go Blip in the Night: Standby Power and How to Limit It." Organization for Economic Cooperation and Development, Paris.

National Energy Policy Development Group (Cheney, R., Powell, C. L., O'Neill, P., Norton, G., Veneman, A. M., Evans, D. L., Mineta, N. Y., Abraham, S., Allbaugh, J. M., Whitman, C. T., Bolten, J. B., Daniels, M. E., Lindsey, L. B., and Barrales, R.) (2001). "National Energy Policy: Report of the National Energy Policy Development Group." U.S. Government Printing Office, Washington, DC.

Panel on Policy Implications of Greenhouse Warming (Evans, D. J., Adams, R. M., Carrer, G. F., Cooper, R. N., Frosch, R. A., Lee, T. H., Mathews, J. T., Nordhaus, W. D., Orians, G. H., Schneider, S. H., Strong, M., Tickell, C., Tschinkel, V. J., and Waggoner, P. E.), National Academy of Sciences (1992). "Policy Implications of Greenhouse Warming: Mitigation, Adaptation, and the Science Base." National Academy Press, Washington, DC.

Van Heuvelen, A., and Zou, X. (2001). Multiple representations of work-energy processes. *Am. J. Phys.* **69,** 184–194.

Consumption, Energy, and the Environment

SIMON GUY
University of Newcastle
Newcastle Upon Tyne, United Kingdom

1. Introduction
2. Theories of Technical Change
3. Barriers to Energy Efficiency
4. Leaping the Barriers
5. Contexts of Consumption
6. Sociotechnical Theories of Change
7. Conclusion: Rethinking Energy Consumption

Glossary

nontechnical barriers The idea that the "attitudes" of decision makers toward technology are the key determinants of innovation and that apathy, ignorance, and lack of financial interest are the major obstacles.
socially relevant actors The broad networks of individuals and organizations that may have an influence on technical change.
sociotechnical The idea that science is a sociocultural phenomenon and that the technical is always in relationship with wider social, economic, and political processes.
technical frames The sets of assumptions and culturally conditioned ways of seeing that underpin individual or organizational perspectives on technical change.
technoeconomic The idea that if technical knowledge is rigorously tested and demonstrably proved, then consumption choices will be made rationally.
universe of choice The multiple rationalities shaping technical innovation and how these relate to the contrasting contexts in which design and development actors operate.

The contribution of building science to our stock of energy efficiency knowledge is substantial. Energy-saving technologies and materials have been successfully developed and manufactured, and energy-efficient building designs have been constructed, tested, and widely publicized. In fact, for some time, designers have possessed the technical knowledge and identified the best practice design techniques necessary to construct zero-energy buildings. Moreover, extensive monitoring of local, national, and international building stocks means that more is known than ever before about the precise potential for improved energy performance. Having perfected energy efficiency technologies, the next task is to explain why apparently proved technical knowledge is often ignored in design practices, or why building occupants consistently fail to adopt energy-saving techniques. Facing the "inadequate diffusion" of apparently cost-effective energy-conserving technologies, the practice of building science is extended to consider the nontechnical processes of technology transfer and arrives at the complex question of consumption.

1. INTRODUCTION

> *Those who call ourselves energy analysts have made a mistake…we have analysed energy. We should have analysed human behaviour.*
>
> *—Lee Schipper, 1987*

> *A renewal of the social theory which informs energy consumption and conservation is called for in the face of environmental challenges.*
>
> *—Hal Wilhite, 2003*

Many researchers, like Schipper and Wilhite, are now questioning the notion that the challenge of improving the energy performance of buildings is simply a task for "building science," a term now widely used to describe the growing body of knowledge about the relevant physical science and its application to buildings. Across international boundaries and research cultures, the common aim of building science is to produce more and better

 © 2004 Elsevier Inc. All rights reserved.

knowledge about buildings and how they perform in order to identify promising solutions to the problem of energy inefficiency. The critical issue here is the definition of "better knowledge" about building performance and the subsequent identification of problems and solutions within a scientific, or techno-economic, paradigm. As with other branches of physical science, the reactions of materials and other inanimate objects are the legitimate interests of the building scientist. Like problems of rain penetration, fire safety, or structural soundness, the issue of energy efficiency in buildings becomes defined as a technical problem amenable to scientific methods and solutions. The belief is that proved technical solutions are transferable and readily applicable to other technically similar situations. Consequently, the research agenda of building science is geared toward the scientific resolution of what are taken to be physical problems. Energy is no different. As Loren Lutzenhiser points out, to date, energy management strategies have "focused almost entirely on the physical characteristics of buildings and appliances." Vast sums of money have been, and continue to be, invested in scientific research to identify and fulfill (at least experimentally) the technical potential for energy savings in buildings.

2. THEORIES OF TECHNICAL CHANGE

Acknowledgment that an actual improvement in buildings has not always matched growth in knowledge has led to the realization that building scientists must appreciate the contribution of the life sciences. Here, the most natural fit has proved to be between building science and economic theory. Economists reinterpret social processes in terms of a market arena that neatly divides the world into separate, but interlinked, domains of knowledge and action. A world of perfect information, on the one hand, and a definable logic of (utility-maximizing) rationality, on the other hand, are posited. This confidence in the capacity of individual decision makers to quantify the benefits of energy efficiency economically is central to the technoeconomic view of technology transfer. The ability of building users literally to count the costs of energy consumption will ultimately lead to more rational energy use. This approach has found eloquent expression in the field of energy economics, and the use of energy audits that enable the new user to cope with the complex conversion equations and calculation of energy costs per standard unit. As Chadderton makes clear, the link to economics is explicit:

> *An energy audit of an existing building or a new development is carried out in a similar manner to a financial audit but it is not money that is accounted. All energy use is monitored and regular statements are prepared showing final uses, costs and energy quantities consumed per unit of production or per square metre of floor area as appropriate. Weather data are used to assess the performance of heating systems. Monthly intervals between audits are most practical for building use, and in addition an annual statement can be incorporated into a company's accounts.*
>
> —*D. V. Chadderton, 1991*

This model of rational action suggests that financial and energy consumption-related decisions are comparable, and that business and domestic users alike are self-interested, knowledgeable, and economically calculative when considering energy measures. Lutzenhiser illustrates this in regard to business users in particular who, skilled at marginal cost and future calculations, would "only seem to need to see the potential competitive economic advantage of innovation to move towards energy efficiency." Following a similar logic, the development of a visible market for energy efficiency would, in turn, encourage domestic consumers to minimize their costs by substituting more efficient ways of satisfying needs for energy services such as heating, cooling, or water heating. Here, the methodological presumption of building science, that with careful monitoring and scientific control it is possible to reproduce the technical achievements of the laboratory universally, is mirrored by an economic logic that promises that technological potential will be fulfilled in any market situation that demonstrates a "static, intertemporal, and intergenerational Pareto optimality." Put simply, the view is that economically rational actors, replete with the necessary technical and economic information, will consistently put science into practice. As Hinchliffe puts it, "a rational, profit maximising man is visualised at work, at home and at play," whereas "the engineering model tends to picture humans as optimally utilising technologies after the fashion of their creators." This epistemological coalition of building science and economics has led, internationally, to a prescriptive view of technological diffusion based on the twin technical and economic logics of proved, replicable, science and idealized consumer behavior. As Lutzenhiser suggests,

> *As a result, a physical–technical–economic model (PTEM) of consumption dominates energy analysis, particularly in energy demand forecasting and policy planning. The behaviour of the human 'occupants' of buildings is seen*

as secondary to building thermodynamics and technology efficiencies in the PTEM, which assumes 'typical' consumer patterns of hardware ownership and use.

—L. Lutzenhiser, 1993

The technoeconomic view of energy efficiency suggests that if technical knowledge is rigorously tested and demonstrably proved, and if "market forces" are not "disturbed," then consumption choices should be taken rationally, with the "right decisions being taken by millions of individual consumers, both at home and in their place of work." The role of policymakers is clear, "to set the background conditions and prices such that consumers will take decisions which are both in their own and the national interest."

3. BARRIERS TO ENERGY EFFICIENCY

This way of seeing technical change is not merely a matter of the attitudes or perceptions of individual policymakers. Rather, it provides an organizational logic or operating principle on which energy research and policy proceeds. This epistemic view represents a specific social organization of technical development and diffusion. The outcome of this view in policy terms is that human and financial resources are, almost exclusively, committed to demonstrating continually the technical efficacy of energy-efficient innovation through regulation, the provision of information about technical means to reduce energy consumption, and the development of best practice demonstration schemes. That policymakers, industry, and nongovernmental organizations all find consensus in this model of technological diffusion is perhaps not surprising, given its strikingly familiarity in the wider world of environmental policy debates.

Stephen Trudgill has, approvingly, formalized the technoeconomic way of seeing innovation in his formulation of the Acknowledgment, Knowledge, Technology, Economic, Social, Political (AKTESP) barriers to the resolution of environmental problems. Trudgill presents an image of technological innovation as a path from ignorance to enlightenment, with the evils of social, political, and economic reality cast as unpredictable obstacles to an otherwise assured technical utopia. Critically, the key to overcoming AKTESP barriers is almost always located within individual motivations. As Trudgill puts it, "motivation for tackling a problem comes from our moral obligation and our self-interest in enhancing the resource base and its life—thus enhancing, rather than destroying, planetary ecosystems and plant and animal species, including ourselves." When such motivation arises, we are in a position to solve the remaining barriers to the solution and to the implementation of environmental problems. The vocabulary of solution and implementation contains similarly individualized terms: "inadequacy of knowledge," "technological complacency," "economic denial or complacency," "social morality/resistance/leadership," and "political cynicism/ideology." Within this model of technical change, individuals are consistently the linchpins of effective energy-saving action. Changes in the level of the energy demand of the built environment is seen as a process involving "thousands" of individual judgments by property owners and other decision makers. The technical, organizational, and commercial complexity of energy-related decision-making is here replaced by an image of autonomous actors, each free to commit themselves to a more sustainable urban future. As Nigel Howard has argued, "there are lots of decision-takers. There are lots of people who have to be influenced, right from government to local authorities, developers, designers, material producers, professional and trade bodies. They all have a role to play. And what we have to do is try and influence all of them." This way of seeing energy views more or less "rational" individuals as both the solution to and the cause of energy-related environmental problems. Bewildered by the apparent irrationality of the social world, government, commercial, and voluntary interests in energy efficiency have all committed to a technoeconomic view of technology transfer, whereby established, proved, scientific knowledge is impeded by lack of information and/or by other nontechnical barriers. As Kersty Hobson puts it, "the drive to provide individuals with information, either to 'create' responsibility or to affect consumption, underpins the choice of national policy mechanisms used to forward sustainable consumption issues."

We are faced here with a self-sustaining, mutually reinforcing package of beliefs spinning between the realms of technology and policy without really belonging to either. Each element of the pervasive bundle, the transferability of technical knowledge, the individualistic theory of technical change, the sequential logic of research and development, and the implicit distinction between the social and the technical, feeds into the next, creating a web of belief strong enough to encapsulate technical researchers and their project officers and elastic enough to span countries and continents.

4. LEAPING THE BARRIERS

> *By defining what is acceptable as evidence, certain privileged methods also act to exclude other sorts of data. It is in this way that certain questions remain unmasked, and certain types of evidence are ignored or dismissed as invalid.*
>
> —*M. Leach and R. Mearns, 1995*

Leach and Mearns' observations highlight how wide acceptance of the technoeconomic view of technology transfer has hitherto marginalized explanations of technical innovation concerned with the social shaping of technical change. As Thomas Hilmo suggests, the notion of barriers suggests that "problems are absolute" and that "some actors [scientists] know the truth about a problem," whereas "other actors [nonscientists] do not and obstruct the solutions in different ways." Hence, the role of social scientist is in turn reduced to that of market researchers, typically undertaking attitudinal surveys designed to identify the human barriers to good energy practice that the promotional campaigns are designed to overcome. This research has typically been described as exploring the behavioral or human dimensions of energy efficiency. While drawing on diverse intellectual sources from economic sociology to environmental psychology, such research shares a view of energy efficiency as the product of a slow cascade of more or less rational individual choices. Although Stern and Aronson usefully identify five archetypes of energy user (the "investor," "consumer," "member of a social group," "expressor of personal values," and "avoider of problems"), their narrow focus on the attitudes and motivations of individual energy consumers tends to isolate and atomize the decision-making process. Little is heard in this research about the consumption practices of organizations, or the role of the organizational actors and groups that actually design, finance, and develop buildings. As Janda points out, although this approach "usefully delineates differences in individuals' attitudes, human dimensions research does not reveal where these differences originate, how they develop, or if they can be changed." Attitudes and decisions are always shaped and framed within wider social processes. Abstraction of the opinions and outlook of energy consumers, and of design and development actors from the contexts of production and consumption, tends to isolate and freeze what are always contingent practices. This narrow, behavioral view of technical change fails to recognize the routine complexities of energy-related decision making. In particular, there is little room in this model to view technical innovations and social processes as interrelated, or to assess how energy-efficient choices may be embedded in the mundane routines of domestic life and commercial practice.

The technoeconomic way of seeing energy efficiency can be contrasted with an alternative view, the sociotechnical, which has questioned what Michael Mulkay has termed the "standard view of science," a view that exists, already formed, outside society. In its place, a broad church of sociologists have presented a revised image of science as a "sociocultural phenomenon" that questions the "authority of science," locates "knowledge-claims in their social context," and identifies the "relationship between such contexts and wider economic and political processes." This approach to understanding science and technical change can be characterized by a series of questions. Volti asks, for example, "What accounts for the emergence of particular technologies? Why do they appear when they do? What sort of forces generate them? How is the choice of technology exercised?" Rather than analyze the process of technical change internally, in terms of developments within the technology, or by reference to the ideas of famous scientists, inventors, and entrepreneurs, such research is interested in discovering to what extent, and how, does the kind of society we live in affect the kind of technology we produce. Or, as McKenzie and Wajcman ask, "What role does society play in how the refrigerator got its hum, in why the light bulb is the way it is, in why nuclear missiles are designed the way they are?"

Questioning the notion that technological change has its own logic, this way of seeing technical change recognizes the wider social contexts within which design solutions emerge and patterns of consumption evolve. Rather than viewing science and technology as asocial, nonpolitical, expert, and progressive, innovation is viewed by Andrew Webster as a "contested terrain, an arena where differences of opinion and division appear." In relation to analysis of energy problems, such research avoids individualist explanations of technological innovation (the rational energy consumer), rejects any form of technological determinism (technical innovation as handmaiden to an energy-efficient economy), and, critically, refuses to distinguish prematurely between technical, social, economic, and political aspects of energy use. As Thomas Hughes has graphically illustrated, in exploring the development of the American electricity system, "sociological, technoscientific and economic analyses are permanently

woven together in seamless web." In this world without seams, social groups and institutions are considered, alongside technological artifacts, as actors who actively fashion their world according to their own particular logic of social action. This sociotechnical analysis of energy use replaces technoeconomic descriptions of universal barriers to energy-efficient innovation (apathy, ignorance, lack of financial interest), with analysis of the ways in which the changing social organization of energy-related choices structures opportunities for more efficient energy use.

There are complementary methodological issues at stake here. Bridging the theoretical gap between the social and technical features of energy use demands a more qualitative research agenda. Rather than solely relying on positivist research tools such as surveys, opinion polls, and statistical analysis, undertaking sociotechnical research means attempting to peer over the shoulder of the actors making energy-related decisions by following actors through their professional and personal routines. In this way, research into energy efficiency finds itself on what Michel Callon has termed a new terrain: that of "society in the making." This refocuses analysis away from pure energy questions to a wider set of debates about design conventions, investment analysis, development costs, space utilization, and market value. Idealized notions of a rational energy user or best technical practice make little sense here. Instead, the social and the technical form a network of associations that serve to frame the meaning of energy efficiency, thereby encouraging or delimiting opportunities for innovation over time and space, and between organizational settings.

5. CONTEXTS OF CONSUMPTION

A significant amount of academic work has been undertaken in an attempt to put energy use into its social context. Social psychologists and sociologists have emphasized the importance of seeing energy problems and solutions in terms of "social systems rather than single causes," the need to design energy systems for "adaptability as an alternative to detailed planning," and, as Stern and Aronson argue, to treat energy policies and programs as forms of "social experiment." Reviewing this work, Lutzenhiser found "a consensus in the literature" that to understand the sociotechnical complexity of energy-saving action, policymakers must be concerned more directly with "the social contexts of individual action." As Lutzenhiser points out,

> *While the physical–technical–economic model assumes consumption to be relatively homogenous and efficiency to be driven by price, the empirical evidence points towards variation, non-economic motives, and the social contexts of consumption. Economics can supply normative guides regarding when investments would be economically desirable, but it tells us little about how persons actually make economic decisions.*
>
> —*L. Lutzenhiser, 1993*

Both Lutzenhiser and Wilhite draw attention to the social nature of energy consumption and the social dynamics of change as a form of "social load." These social loads have peaks that drive the dimensioning of peak energy loads. Wilhite provides examples from his anthropological research into lighting use in Norway. Here, home lighting is designed not simply to provide a sufficient degree of brightness to support basic human activity, but rather to provide a "cozy aesthetic." A dark house was described by Wilhite's respondents as a "sad house." Extra lighting fittings are always fitted and utilized to make sure guests feel cozy, and for social visits on a winter evening, lights are left on in every room to provide a welcoming glow as guests arrive. This, of course, means the use of extra energy. In his study sample, Wilhite discovered an average of 11.5 lights per living room. This social peak "drives the dimensioning of the material and energy system behind lighting, which based on a straightforward provision of lumens would be far smaller." As Wilhite argues elsewhere, "the things we use energy to achieve—a comfortable home, suitable lighting, clean clothes, tasty food—have also been assumed in models of consumption to be generic and physically determined." Wilhite goes on to argue that there is an "urgent need for the development of a more robust theory of consumption, one which incorporates social relations and cultural context, as well as perspectives on individual agency and social change."

6. SOCIOTECHNICAL THEORIES OF CHANGE

In Table I, the two broad positions outlined so far are contrasted. The effect of these analytical assumptions is cumulative, with a technoeconomic or sociotechnical view of buildings as an artifact leading correspondingly to a particular way of viewing energy-efficient design, energy-saving action, technical innovation, or market failure. Commitment to

TABLE I

Ways of Viewing Energy Efficiency in Buildings

	Perspective	
Aspect	**Technoeconomic**	**Sociotechnical**
Buildings	Materially similar, physical structures	Material product of competing social practices
Energy-efficient design	Replicable technical solutions	Outcome of conflicting sociocommercial priorities
Energy-saving action	Individual, rational decision-making in a social vacuum	Socially structured, collective choices
Technological innovation	Series of isolated technical choices by "key" decision makers	Technical change embedded within wider sociotechnical processes
Market failure	Existence of social barriers	Lack of perceived sociocommercial viability
Image of energy consumers	More or less rational	Creative, multirational, and strategic
Role of social science research	Evaluation of technical potential and the detection of environmental attitudes and nontechnical barriers	Identification of context-specific opportunities for technological innovation
Energy policy	Provision of information, granting of subsidies, and setting of regulations	Forging of context-specific communities of interest and promotion of socially viable pathways of innovation

either a predictable linear or socially shaped diffusion pathway for energy-efficient technologies frames the contribution of social science research to understanding innovation and the direction of energy policy-making.

These competing ways of seeing energy consumption raise a number of research dilemmas. Commitment to a technoeconomic or sociotechnical perspective on innovation focuses attention on different actors (key decision makers vs. relevant social groups), different practices (technical design vs. development strategies), and different processes (technological diffusion vs. social and commercial change). We are seemingly faced with a series of choices over how we conceptualize the problem of energy consumption in buildings. As Groak suggests, we can view buildings in terms of their physical attributes, as "essentially static objects" formed in a relatively standardized manner from an assembly of interconnecting construction materials. Seen this way, buildings appear remarkably similar. Irrespective of geographical location, ownership patterns, or operational function, the technical character of building form appears comparatively homogeneous. Alternatively, we could view buildings as material products of competing social practices. This would suggest a different analytical approach. For sociologists such as Bruno Latour, understanding "what machines are" is the same task as understanding "who the people are" that shape their use. Seen this way, technologies and technological practices are as Bijker and Law have illustrated, "built in a process of social construction and negotiation," a process driven by the shifting social, political, and commercial interests of those actors linked to technological artifacts of design and use. Although the complexity of buildings differs from the individual technologies often studied within sociological studies of science and technology, we can nevertheless develop a similar analytical approach. Thus, to understand buildings we must trace the characteristics of the "actor world" that "shapes and supports" their production. Adopting this perspective would mean relating the form, design, and specification of buildings to the social processes that underpin their development. So, although two identical buildings, standing side by side, may well appear physically and materially similar, investigation of their respective modes of production and consumption may reveal profoundly different design rationales, which in turn might help explain variations in energy performance.

This stress on the social organization of design is at odds with the technoeconomic perspective that emphasizes how a repertory of well-tried technical solutions provides reliable precedents for designers. Here, new technical challenges are seen as solvable by shifts in design emphasis, mirroring the march of scientific progress. Although the form and specification of buildings may well vary spatially with climate and culture, the objective is always viewed as the

same, i.e., the provision of the universal needs, shelter and comfort. Technical design is viewed as a process of adaptation and modification to suit changing physical circumstances. Emphasizing the social logic of design raises a different set of questions. Rather than supporting a linear model of innovation, studies in the sociology of technical change have revealed the multidirectionality of the technical design process. Rather than one preordained process of change, we are faced with competing pathways of innovation. Typically, Pinch and Bijker describe the development process of a technological artifact as "an alternation of variation and selection." For example, Wiebe Bijker's study of the development of fluorescent lighting points to a range of innovation pathways that could not be resolved by appeals to technical superiority. Instead, a standoff between lighting manufacturers (who supported a high-efficiency daylight fluorescent lamp) and electric utility companies (who, worried about the effect on electricity sales, supported an energy-intensive tint-lighting fluorescent) led to the introduction of a third design alternative, the high-intensity daylight fluorescent lamp, which combined efficiency with a high light output, thereby maintaining electricity demand. Bijker's study illustrates how a socially optimal design was established from a range of technically feasible possibilities through a process of compromise between competing social interests. In doing so, he highlights the "interpretative flexibility" of design. We might similarly ask how rival energy-efficient designs are valued by different members of the design and development process and how this process of contestation and compromise frames the resulting design strategies.

In approaching these questions, we might begin to draw attention to the idealism surrounding the technoeconomic image of enlightened, rational individuals motivated by a growing stock of technical knowledge. Although these more, or sometimes less, knowledgeable individuals are placed in a hierarchy of influence—from the key decision makers, designers, and top managers, to technicians and lower management, to domestic consumers—their social, spatial, or temporal situation appears of marginal importance. From a sociotechnical perspective, the relationship between individual and context is emphasized and made fluid. Particular technical choices are viewed as expressive of the prevailing social, political, and commercial pressures operating within spatially and temporally contingent contexts. Here, technical choices are not considered to be solely determined by knowledge or motivation, but are shaped by the existence of a more or less socially favorable context. For a variety of reasons, consumers may be unable or prefer not to use particular technologies, or may even use technologies in unpredictable ways not envisaged in the original design. To understand this social structuring of technical choice, Ruth Schwartz Cowan, in her history of home heating and cooking systems in America, treats the consumer "as a person embedded in a network of social relations that limits and controls the technological choices that she or he is capable of making." For Cowan, consumers come in "many different shapes and sizes," and operate in a variety of social contexts. Her analysis of the introduction of cast iron stoves to replace open hearth fires for cooking traces the interconnections between stove producers and merchants, fuels suppliers and merchants, and their networks of influence through production, wholesale, retail, and household domains over both urban and rural consumers. This emphasis on the embedding of the decision maker within wider social networks focuses analytical attention on the "place and time at which the consumer makes choices between competing technologies." This allows Cowan to unpack the "elements" more "determinant of choices" and the technical pathways that "seemed wise to pursue" or that appeared "too dangerous to contemplate." As Cowan points out, "today's 'mistake' may have been yesterday's 'rational' choice." Drawing on this approach, a sociotechnical perspective on energy efficiency might identify the multiple rationalities shaping innovation and how these relate to the contrasting "universe of choice" in which design and development actors operate.

If we begin to accept that the nature and direction of technical change are subject to interpretative flexibility, underpinned by multiple rationalities of context-specific choice, then we would also have to begin to alter our understanding of the process of technological innovation. As we have seen, the technoeconomic perspective views technical change as following an almost preordained pattern of design, development, and diffusion. By contrast, a sociotechnical analysis would explore why particular technical solutions emerged at a certain time and in a particular place. For example, in studying the electrification of cities, David Nye illustrates how the emergence of street lighting was less connected to the rational technical ordering of urban space, and more intimately linked to the need of utility companies to increase load and to the commercial instincts of shopkeepers keen to attract business. Nye

suggests that "shopkeepers understood lighting as a weapon in the struggle to define the business center of the city dramatizing one sector at the expense of others." Framed this way, "electric lighting could easily be sold as a commercial investment to increase the competitiveness of a business." As a result, the subsequent spread of street lighting was accelerated as "electrification of one street quickly forced other commercial areas to follow suit or else lose most of their evening customers." Following this lead, a sociotechnical analysis of energy efficiency would ask what roles were played by architects, developers, governments, investors, manufacturers, retailers, and consumers in fashioning innovation in building design and use. This approach would mean widening the nature of analytical inquiry away from explicit decisions about energy efficiency studied in isolation, to an examination of the embedding of energy-related choices in the manufacture, distribution, and retailing of the technologies, and the commercial processes framing building design and development.

Finally, asking different questions about the process of technical innovation may provide a different understanding of the market success or failure of proved energy-saving technologies. Instead of characterizing nontechnical barriers as both universal and timeless in nature, a sociotechnical approach would explore the degree to which the marketability of technical innovation can be identified as a socially, and temporally, contingent process. For example, Gail Cooper highlights how the seemingly pervasive nature of air-conditioning systems in the United States masks a more contested story about the emergence of heating, cooling, and ventilating technologies. She shows how attempts by engineers to promote the most "technically rational design," which demanded sealed buildings and passive use, was resisted by what engineers saw as "irrational users" who preferred mobile, plug-in air-conditioning systems that, although less efficient, provided greater flexibility and active control. As Cooper suggests, "the engineering culture that characterised the custom-design industry did not produce the best technology, neither did the market forces that dominated the mass production industry." The result is localized compromises that reflect the "seesawing power relations surrounding the development of air-conditioning." In particular, Cooper's study underlines the contrasting image of the energy consumer underpinning the technoeconomic and sociotechnical views. Put simply, the techno-economic "irrational user" is translated in the socio-technical literature into a "guerrilla fighter of those disenfranchised from the design process." Herein lies a key distinction. Rather than assume the intrinsic marketability of technically proved innovations, a sociotechnical approach would assess the socio-commercial viability of the artifact in varying social contexts. Instead of explaining market failure in terms of ubiquitous and timeless barriers, a socio-technical analysis would seek to explain market success or failure in situationally specific situations. Some contexts may favor innovation, others may not. Mapping what Bijker terms the "technical frames" of "socially relevant actors" is vital here, for, as Cowan points out, any one of those "groups or individuals acting within the context of their group identity…may be responsible for the success or failure of a given artefact."

7. CONCLUSION: RETHINKING ENERGY CONSUMPTION

> *The understanding of present and future energy use depends upon understanding how conventions and practices evolve over time and within and between cultures.*
>
> —*Hal Wilhite, 2003*

The aim of this new agenda is not to produce abstract social theory. Instead, a growing number of social scientists are striving to articulate a new role for social science in research and policy debates about energy consumption and environmental change. These researchers are exploring the heterogeneous and contested nature of consumption practices that shape energy use. Rather than simply assuming that people use energy, they are analyzing how energy intersects with everyday life through diverse and cultural inscribed practices such as heating and cooling, cooking, lighting, washing, working, and entertaining. To achieve this, they are drawing on methods and theories beyond science and economics, including sociology, anthropology, geography, psychology, and cultural studies, and are learning lessons from consumption debates beyond energy and the environment, including fashion, food, and shopping.

The scope of this research agenda takes us far beyond the technoeconomic analysis of energy consumption, which tends to be dedicated to identifying the potential scope and scale of energy performance improvements in different building types and in different building sectors, and to the setting of technically feasible CO_2 abatement targets. It also suggests a role for social scientists much

deeper than the forms of market research, evaluating the attitudes of what are taken to be key decision makers toward energy that typify research on human behavior and attitudes. In developing a sociotechnical approach to energy consumption, the concern is more with identifying context-specific opportunities for inserting technically proved technologies into appropriate social practices. But rather than being led by calculations of technical potential, the task is to seek to identify how specific social, spatial, and temporal configurations of energy consumption encourage, or mitigate against, effective energy-saving action. The result is stories about the circumstances in which energy efficiency and conservation practices do or do not flourish. This focus takes us far from the world of building science and the paradigmatic certainties of the technoeconomic perspective, and instead reveals the construction of energy knowledge in varying social worlds and reflects the contested nature of energy consumption practices.

SEE ALSO THE FOLLOWING ARTICLES

Conservation of Energy, Overview • Economics of Energy Efficiency • Energy Efficiency and Climate Change • Energy Efficiency, Taxonomic Overview • Obstacles to Energy Efficiency • Technology Innovation and Energy • Vehicles and Their Powerplants: Energy Use and Efficiency

Further Reading

Bijker, W. E. (1995). Sociohistorical technology studies. *In* "Handbook of Science and Technology Studies" (S. Jasanoff *et al.*, Eds.), pp. 229–256. MIT Press, Cambridge, Massachusetts.

Bijker, W. E., and Law, J. (eds.). (1992). "Shaping Technology/Building Society." MIT Press, Cambridge, Massachusetts.

Bijker, W. E., Hughes, T. P., and Pinch, T. (eds.). (1987). "The Social Construction of Technological Systems." MIT Press, Cambridge, Massachusetts.

Callon, M. (1987). Society in the making: The study of technology as a tool for sociological analysis. *In* "The Social Construction of Technological Systems" (W. E. Bijker, T. P. Hughes, and T. Pinch, Eds.), pp. 83–103. MIT Press, Cambridge, Massachusetts.

Chadderton, D. V. (1991). "Building Services Engineering." E & FN Spon, London.

Cooper, G. (1998). "Air Conditioning America: Engineers and the Controlled Environment 1900–1960." Johns Hopkins Univ. Press, Baltimore.

Cowan, R. S. (1987). How the refrigerator got its hum. *In* "The Social Shaping of Technology" (D. McKenzie and J. Wajcman, Eds.), pp. 202–218. Milton Keynes: Open Univ. Press, Philadelphia.

Groak, S. (1992). "The Idea of Building: Thought and Action in the Design and Production of Buildings." E & FN Spon, London.

Guy, S., Marvin, S., and Moss, T. (2001). "Urban Infrastructure in Transition: Networks, Buildings, Plans." Earthscan, London.

Hilmo, T. (1990). Review of barriers to a better environment. *Geografiska* **72B**, 124.

Hinchliffe, S. (1995). Missing culture: Energy efficiency and lost causes. *Energy Policy* **23**(1), 93–95.

Hobson, K. (2002). Competing discourses of sustainable consumption: Does the 'rationalisation of lifestyles' make sense? *Environ. Politics* **11**(2), 95–120.

Howard, N. (1994). Materials and energy flows. *In* "Cities, Sustainability and the Construction Industry," pp. 11–14. Engineering and Physical Science Research Council, Swindon.

Hughes, T. P. (1983). "Networks of Power: Electrification in Western Society, 1880–1930." Johns Hopkins Univ. Press, Baltimore.

Janda, K. (1998). "Building Change: Effects of Professional Culture and Organisational Context on Adopting Energy Efficiency in Buildings." Unpublished Ph.D. thesis, University of California, Berkeley.

Latour, B. (1987). "Science in Action." Milton Keynes: Open University Press, Philadelphia.

Leach, M., and Mearns, R. (1995). "The Lie of the Land: Challenging Received Wisdom in African Environmental Change and Policy." James Currey, London.

Lutzenhiser, L. (1993). Social and behavioural aspects of energy use. *Annu. Rev. Energy Environ.* **18**, 247–289.

Lutzenhiser, L. (1994). Innovation and organizational networks: Barriers to energy efficiency in the US housing industry. *Energy Policy* **22**(10), 867–876.

McKenzie, D., and Wajcman, J. (1985). "The Social Shaping of Technology." Milton Keynes: Open Univ. Press, Philadelphia.

Mulkay, M. (1979). "Science and the Sociology of Knowledge." George Allen and Unwin, London.

Nye, D. E. (1998). "Consuming Power: A Social History of American Energies." MIT Press, Cambridge, Massachusetts.

Pinch, T., and Bijker, W. T. (1989). The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *In* "The Social Construction of Technological Systems" (W. E. Bijker, T. P. Hughes, and T. Pinch, Eds.), pp. 17–50. MIT Press, Cambridge, Massachusetts.

Schipper, L. (1987). Energy conservation policies in the OECD: Did they make a difference? *Energy Policy*, December, 538–548.

Stern, P. C., and Aronson, E. (eds.). (1984). "Energy Use: The Human Dimension." W. H. Freeman, New York.

Trudgill, S. (1990). "Barriers to a Better Environment." Belhaven Press, London.

Volti, R. (1992). "Society and Technological Change." St. Martins Press, New York.

Webster, A. (1991). "Science, Technology and Society." Macmillan, London.

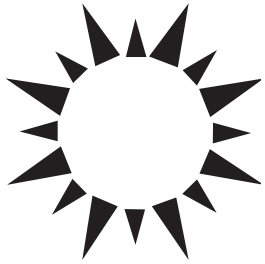
Wilhite, H. (1997). "Cultural Aspects of Consumption." Paper (unpublished manuscript) presented at the ESF-TERM workshop on Consumption, Everyday Life and Sustainability, Lancaster University, Lancaster, U.K.

Wilhite, H. (2003). What can energy efficiency policy learn from thinking about sex? *In* "Proceedings of European Council for an Energy-Efficient Economy (ECEEE)," pp. 331–341. ECEEE, Paris.

Wilhite, H., and Lutzenhiser, L. (1998). Social loading and sustainable consumption. *Adv. Consumer Res.* **26,** 281–287.

Wilhite, H., Shove, E., Lutzenhiser, L., and Kempton, W. (2000). "Twenty Years of Energy Demand Management: We Know More about Individual Behaviour but How Much Do We Really Know about Demand?" Proceedings of the American Council for an Energy-Efficient Economy (ACEEE) 2000 Summer School on Energy Efficiency in Buildings. ACEEE, Washington, D.C.

Conversion of Energy: People and Animals



VACLAV SMIL
University of Manitoba
Winnipeg, Manitoba, Canada

1. **Mammalian Metabolism**
2. **Human Energetics**
3. **Sustained Output and Extreme Exertions**
4. **Working Animals**
5. **Bovines**
6. **Horses**

Glossary

basal metabolic rate A measure of nutritional energy needed by an organism during a fasting and resting state.

bovines Large ungulate mammals belonging to genus *Bos* (cattle) and *Bubalus* (water buffalo) that were domesticated for dairy production and draft labor.

draft animals Domesticated large ungulate mammals of several families (bovines, horses, and camelids) harnessed for fieldwork or transport duties.

horses Large ungulates of the genus *Equus* that were domesticated for heavy agricultural labor and transport draft as well as for riding and warfare.

metabolic scope Multiple of the basal metabolic rate achieved during periods of maximum exertion.

The power of human and animal muscles, limited by metabolic rates and mechanical properties of working bodies, restricted the productive capacities, and hence the typical quality of life, of all preindustrial civilizations. Only the widespread adoption of inanimate energy converters replaced these animate (or somatic) energies as the most important prime movers of human society. This epochal energy transition began in parts of Western Europe and North America during the early 19th century and it was basically completed during the first half of the 20th century. In contrast, in scores of low-income countries both the subsistence agricultural production and many industrial tasks are still energized by human and animal muscles.

Human labor was the only prime mover in all subsistence foraging (gathering and hunting) societies. Simple wooden, stone, and leather tools (including digging sticks, bows and arrows, spears and knives, and scrapers) were gradually invented and adopted in order to increase and extend the inherently limited muscle power. The only extrasomatic energy conversion mastered by foraging societies was the use of fire for warmth and cooking. Even those early agricultural societies that domesticated large animals continued to rely largely on human muscles, but they increased their effectiveness by multiplying individual inputs through concentrated applications of labor and by using simple but ingenious mechanical devices.

The first approach had logistic problems. For example, only a limited number of people can grasp a small but heavy object to carry it, and merely multiplying the number of available hands makes no difference if the object is to be raised far above the people's heads. The second approach relied largely on the three simplest aids, levers, inclined planes, and pulleys, and their common variations and combinations, including wedges, screws, wheels, windlasses, treadwheels, and gearwheels. These tools and machines were used by virtually all old high cultures. They enlarged the scope of human action but their use had obvious physical limits (e.g., the maximum practical length of levers or pulleys). However, some societies that commanded just the muscle power were able to complete, aided by careful organization of complex tasks, such astounding projects as the megalithic structures at Stonehenge or on the Easter Island, the pyramids of Egypt and Mesoamerica, and

 © 2004 Elsevier Inc. All rights reserved.

massive stone temples, walls, and fortresses on every continent except Australia.

Harnessing of large domesticated animals made the greatest difference in agricultural productivity. Depending on the soil type, a peasant working with a hoe would take 100–180 h to prepare 1 ha of land for planting cereals. A plowman guiding a single medium-sized ox pulling a primitive wooden plough could finish this task in just over 30 h. Use of a pair of horses and a steel plough reduced the time to less than 10 h and in light soil to just 3 h. Hoe-dependent farming could have never reached the scale of cultivation made possible by draft animals, and many tasks, including deep plowing of heavy soils, would have been entirely impossible without draft animals.

Draft animals harnessed to wheeled vehicles also made major contributions to land transport, construction, and industry. Wheelbarrows provided the most efficient way for an individual to transport goods, with an adult man covering approximately 10–15 km a day and moving usually less than 50 kg. Oxen were not much faster but could pull wagons up to 10 times that weight. Heavy horse-drawn wagons covered 30–40 km a day, and passenger horse carts on good roads could travel 50–70 km (galloping messengers on Roman roads could ride up to 380 km per day). Oxen and horses were deployed to transport heavy materials to building sites, such as medieval cathedrals or Renaissance villas. Horses were also used for many tedious tasks in the early stages of expanding coal extraction and incipient industrialization of the 17th and 18th centuries. Here, too, there were obvious logistic limits, no matter if the animals worked alone or in groups. Large and heavy animals could not be deployed in low and narrow mine tunnels, and smaller animals were often inadequate to perform tasks requiring hours of nonstop draft.

Harnessing of more than 20 animals to be guided as a single force is difficult, and harnessing more than 100 of them to get 100 kW of sustained power is impossible. Stabling of horses placed enormous demand on urban space, and during the late 19th century dense horse-drawn traffic often created scenes of chaos on the streets of the largest cities (Fig. 1). Even heavy animals could not demonstrate their powerful draft because of the poor preindustrial roads. Late Roman specifications restricted the loads to 326 kg for horse-drawn and up to 490 kg for slower ox-drawn carriages. By 1850, the maximum allowable load on French roads was approximately 1.4 tons, still a small fraction of today's truckloads (commonly more than 15 tons).

FIGURE 1 High density of horse-drawn traffic on a London bridge in 1872.

Maintaining larger numbers of more powerful animals also required more feed (even during periods when the animals were idle), and this was often impossible given the limited amount of land that could be taken away from food cultivation. Combustion of fossil fuels and generation of primary electricity displaced virtually all hard labor in affluent nations, but reliance on animal and human muscles remains high in most low-income countries. Unfortunately, the latter toll still includes such debilitating tasks as producing gravel by shattering rocks with hand hammers or transporting outsize loads on bent backs of porters. Indeed, this dependence on animate prime movers is one of the most obvious physical distinctions between the rich and the poor worlds.

1. MAMMALIAN METABOLISM

Basal metabolic rate (BMR) is the amount of energy needed to maintain critical body functions of any heterotrophic organism. The rate varies with sex, body size, and age, and its measurements in adults show large individual variations from expected means. To determine BMR in humans, the person must be at complete rest in a thermoneutral environment at least 12 h after consuming a meal. Neutral temperature (approximately 25°C) is required to preclude any shivering and sweating, and the wait eliminates the effects of thermogenesis caused by food digestion, which peaks approximately 1 h after the meal and can add up to 10% to the BMR. Resting energy expenditure (REE) is higher than BMR because it includes the thermic effect of meals. Mammals thermoregulate at 34–40°C (humans at approximately 37°C) and heart, kidneys,

liver, and brain are the metabolically most active organs; in humans, these four organs account for slightly more than one-half of adult BMR.

Max Kleiber pioneered the systematic study of mammalian BMR during the 1930s, and in 1961 he concluded that dependence of this rate (expressed in watts) on total body weight (mass in kilograms) is best expressed as $3.4\,m^{0.75}$ (the exponent would be 0.67 if BMR were a direct function of body surface area). Kleiber's famous mouse-to-elephant line became one of the most important and best known generalizations in bioenergetics (Fig. 2). Closer examination of available BMRs shows slightly different slopes for several mammalian orders, and individual outliers illustrate genetic peculiarities and environmental adaptation. For example, cattle have a relatively high BMR, requiring more feed per unit of body weight gain than would be expected based on their mass, whereas the relatively low BMR of pigs makes them efficient converters of feed to meat and fat. Various explanations of the three-fourths slopes have been offered in the biological literature, but the definitive solution remains elusive.

In order to obtain total energy requirements, BMRs must be increased by energy needs for growth and activity. Conversion efficiency of growth depends on the share of newly stored proteins (which are always more costly to synthesize than lipids) and on age (efficiencies range from more than 40% during postnatal growth to well below 10% for mature adults). In humans, the average food energy need is approximately 21 kJ/g of new body mass in infants, equal to a conversion efficiency of approximately 50%; adult growth efficiencies are up to 35% lower. Regarding the conversion of chemical energy of food and feed to kinetic energy of working muscles, it has been known since the late 19th century that its peak efficiency is slightly more than 20%.

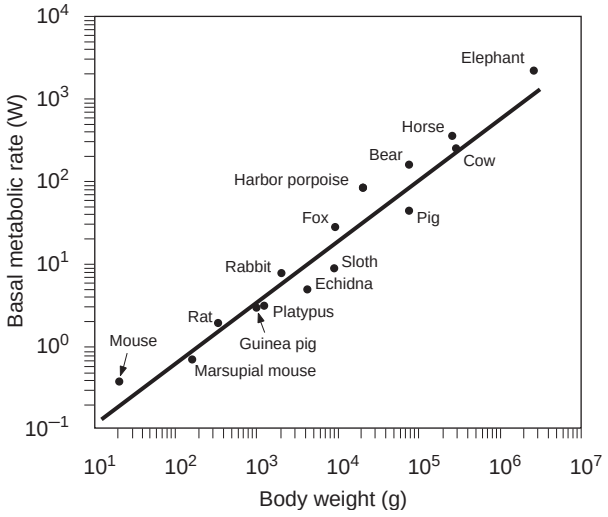


FIGURE 2 Kleiber's line showing the dependence of basal metabolic rate on body mass.

The first fairly accurate estimates of animate power were made in 1699 when Guillaume Amontons studied the effort of French glass polishers. In 1782, James Watt calculated that a mill horse works at a rate of 32,400 foot-pounds per minute, and the next year he rounded this figure to 33,000 foot-pounds, making 1 horsepower equivalent to 745.7 W. By 1800, it was known, correctly, that the power of useful human labor ranges from less than 70 to slightly more than 100 W for most steadily working adults. This means that when working at 75 W, 10 men were needed to equal the power of a good horse for tasks in which either of these animate prime movers could be used.

2. HUMAN ENERGETICS

Healthy people fed balanced diets can convert chemical energy in the three basic classes of nutrients with high efficiency: 99% for carbohydrates, 95% for lipids, and 92% for proteins (however, more than 20% of dietary protein is lost daily through urine). With the exception of people engaged in very heavy work or in strenuous and prolonged exercise, most of the digested energy is used to satisfy the basic metabolic needs. Much BMR and REE data have been acquired mostly by measuring oxygen consumption, which can be easily converted into energy equivalents: One liter of O_2 equals 21.1 kJ/g of starch, 19.6 kJ/g of lipids, and 19.3 kJ/g of protein.

The relationship between BMR and body size can be expressed by simple linear equations (Fig. 3), but these are excellent predictors of individual BMRs only for children and adolescents. More important, using these equations may lead to exaggerated estimates of energy requirements for populations outside the affluent world in general and outside of Europe in particular. For example, BMR equations favored by the Food and Agriculture Organization (FAO) overpredict BMRs of people living in tropical regions by 1.5–22.4%, with the difference being largest for adults older than age 30.

Adult BMRs vary substantially not only among individuals of the same population but also among different population groups. There are no indisputable, comprehensive explanations for these disparities.

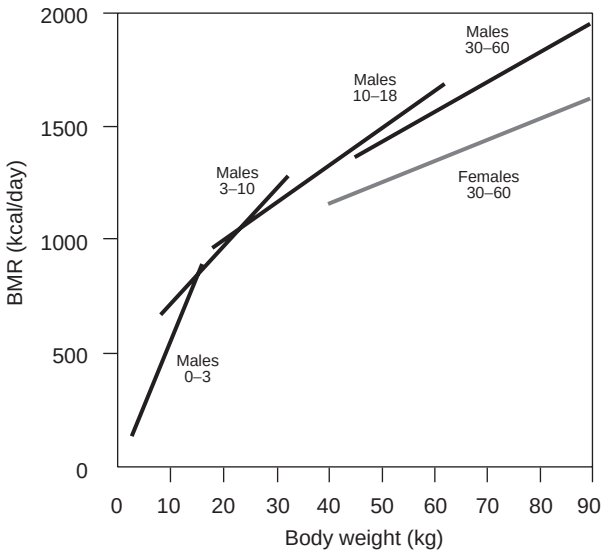


FIGURE 3 Linear predictions of basal metabolic rate (BMR) based on equations recommended by FAO/WHO/UNU.

Disparities among different populations may be due simply to different shares of metabolizing tissues (i.e., muscles and especially internal organs) in the overall body mass. Although individual BMRs vary widely, their lifetime course follows a universal trend. After peaking between 3 and 6 months of age, they decline to approximately half the peak value by late adolescence; this reduction is followed by an extended plateau. Metabolic decline, due to a steady loss of metabolizing lean tissue and increased fat deposition, resumes during the sixth decade and continues for the rest of life. By the seventh decade, fat comprises approximately one-fourth of typical male and slightly more than one-third of female body mass.

As for other mammals, human energy requirements for growth become marginal after reaching adulthood, and food intakes are accounted for almost completely by the combination of BMR and energy needed for various activities. Determination of energy costs of labor and leisure activities has a century-long tradition based on respirometry that requires masks, hoods, or mouthpieces and hoses to measure gas exchange. This is not a problem when testing a cyclist on a stationary bicycle, but it is a challenge when done in restricted workspaces and during aerobic sport activities. Nevertheless, many diverse and interesting values, including those for playing cricket in India and climbing Alpine peaks, have been obtained during decades of such research. Portable systems that can be used for continuous measurements of oxygen uptake for long periods are in widespread use, and two convenient methods allow nonintrusive measurement of energy expenditures.

The first, developed in the mid-1950s, has been used in humans since the mid-1980s: It monitors exponential losses of the doubly labeled water (marked with stable heavy isotopes of ^{2}H and ^{18}O) drunk by subjects during a period of 1 to 2 weeks. When the two heavy isotopes are washed out of the body and replaced with dominant ^{1}H and ^{16}O, loss of deuterium is a measure of water flux and the elimination of ^{18}O indicates not only the water loss but also the flux of CO_2, a key final product of metabolism, in the expired air. The difference between the washout rates of the two isotopes can be used to determine energy expenditure over a period of time by using standard indirect calorimetric calculations. The other technique involves continuous monitoring of heart rate and its conversion to energy expenditure by using calibrations predetermined by a laboratory respirometry.

3. SUSTAINED OUTPUT AND EXTREME EXERTIONS

Energy costs of individual activities, or so-called physical activity levels (PALs), are best expressed as multiples of the BMR or the REE, that is, energy flows equal to approximately 70–90 W for most adult males and 55–75 W for females. Typical averages, corresponding to the reference man of 65 kg and the reference woman of 55 kg, are 78 and 64 W, respectively. Contrary to a common belief, the high basal metabolism of the brain, which always requires approximately one-fifth of the adult food energy intake, increases only marginally when a person is engaged in the most challenging mental tasks: Typical energy increase is only 5% higher than the BMR.

Minimal survival requirements, including metabolic response to food and energy needed to maintain basic personal hygiene, require PALs between 1.15 and 1.2. Standing requires the deployment of large leg muscles, resulting in a PAL 1.3–1.5 times the BMR. Light exertions typical of numerous service and manufacturing jobs that now dominate the modern economies, such as secretarial work, truck driving, assembly line work, or car repair, require no more than 2.5 times the REE. Frequent carrying of loads, lifting of heavy burdens, deep digging, and cleaning of irrigation canals increase the exertion averages into the moderate or heavy expenditure range.

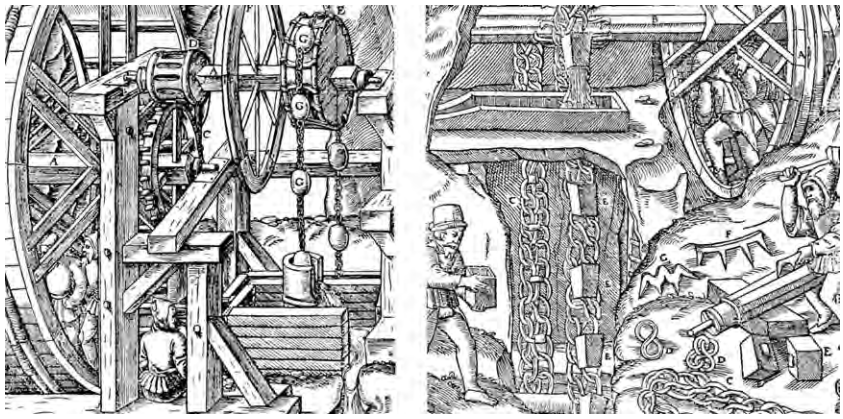


FIGURE 4 Men powering large treadwheels depicted in Agricola's famous "De re Metallica" (1556).

Activities requiring moderate (up to five times the REE), and sometimes even heavy (up to seven times the REE), exertion are now common only in traditional farming (ploughing with animals, hand weeding, and cleaning of irrigation canals are particularly taxing), forestry, and fishing. From the consensus meeting organized by the FAO, the World Health Organization, and the United Nations University PALs of 1.55 for males and 1.56 for females were determined to be the average daily energy requirements of adults whose occupations require light exertions; PALs of 1.64 for females and 1.78 for males engaged in moderately demanding work and 1.82 for females and 2.1 for males with jobs demanding heavy exertions were chosen.

A great deal of experimental work has been performed to determine the limits of human performance. Tiny amount of ATP stored in muscles (5 mmol/kg of wet tissue) can be converted to just 4.2 kJ, enough to energize contractions for less than 1 s of maximal effort. Replenishment through the breakdown of phosphocreatine can add a total of 17–38 kJ, enough to support a few seconds of intensive exertion. Maximum metabolic power achievable in such ephemeral outbursts is large—3.5–8.5 kW for a typical Western man and as much as 12.5 kW for a trained individual. Longer exertions, lasting 30 s to 3 min, can be energized by muscular glycogen. Maximum power of this anaerobic glycolysis is 1.8–3.3 kW for average men. All prolonged efforts are powered primarily by aerobic metabolism, with higher energy demands requiring linear increases in pulmonary ventilation.

Healthy adults tolerate many hours of work at 40–50% of their maximum aerobic capacity. This translates into 400–500 W for most adult men, and with typical kinetic efficiencies of approximately 20% it amounts to 80–100 W of useful work. Untrained individuals have a metabolic scope (the multiple of their BMR) no higher than 10; for active individuals and laborers habituated to hard work, metabolic scope is higher than 15, and for elite endurance athletes (none being better than Nordic skiers and African long-distance runners) the peak aerobic power can be 25 times their BMR or more than 2 kW. Among all mammals, only canids have higher metabolic scopes. Limits of human performance are variable. In total energy terms, peak aerobic capacities are l.5–3.5 MJ for healthy adults, higher than 10 MJ for good athletes, and peak at an astonishing 45 MJ. Both glycogen and fat are the substrates of oxidative metabolism, with lipids contributing up to 70% during prolonged activities.

Overall energy conversion efficiency of aerobic metabolism is 16–20%, whereas anaerobic conversion is less efficient (10–13%). The most efficient, albeit clearly demeaning and stressful, way of using people as prime movers in preindustrial societies was to put them inside (or outside of) treadwheels (Fig. 4). Their operation deployed the body's largest back and leg muscles to do tasks ranging from lifting loads at construction sites and harbors to operating textile machinery. The most efficient use of humans as modern prime movers is in cycling. Bursts of 1 kW are possible for a few seconds while pedaling, and rates of 300–400 W can be sustained for up to 10 min.

4. WORKING ANIMALS

Domesticated animals, mainly bovines (cattle and water buffalo) and equines (horses, ponies, mules, and donkeys), have been the most powerful and highly versatile providers of draft power since the

beginning of Old World sedentary farming. Horses and camels also provided the fastest means of land transport in preindustrial societies, and these and other species (including donkeys, yaks, and llamas) were also used as pack animals, whose caravans were particularly useful in mountainous or desert terrain.

Some Asian societies used working elephants (particularly in forestry operations), several Eurasian circumpolar cultures relied on reindeer as pack animals, and in preindustrial Europe dogs turned spits over kitchen fires or pulled small carts or wheelbarrows. In the Western world, the use of draft animals continued for several decades after the introduction of internal combustion engines and electric motors at the end of the 19th century, and tens of millions of them are still indispensable for fieldwork and transportation in low-income societies of Asia and Africa and in poorer regions of Latin America.

The power of working animals is approximately proportional to their weight (less than 200 kg for small donkeys and approximately 1 ton for heavy horses), but their actual field performance is also determined by a number of external variables. By far the most important are the animal's sex, age, health, and experience; soil and terrain conditions; and the kind of harness used. Only heavier animals could handle such tasks as deep plowing or extracting deep tree roots, but lighter animals were superior in muddy and dandy soils or on steep paths. Moreover, mechanical considerations favored smaller draft animals: Their line of pull is lower, and a more acute angle between this line and the direction of traction will produce a more elongated parallelogram of vectors and hence yield a greater efficiency of work.

Sustainable pulls are equal to approximately 15% of body weight for horses and only 10% for other species. Working speeds range mostly from 0.6–0.8 m/s for oxen to 0.9–1.1 m/s for horses. Thus, depending on their body mass, these draft animals can work at sustained rates of 300–500 and 500–800 W, respectively (Fig. 5). Assuming sustained human exertion at 75 W, the power of a working equine or bovine animal is typically eight times, and usually not less than six times, greater. Actual draft needed for specific fieldwork varies mainly with the soil type.

Deep plowing requires drafts of 120–170 kg, with a range of 80–120 kg for shallow plowing, heavy harrowing, and mowing. Cereal harvesting with a mechanical reaper and binder requires approximately 200 kg. Small horse (350 kg) or a good ox

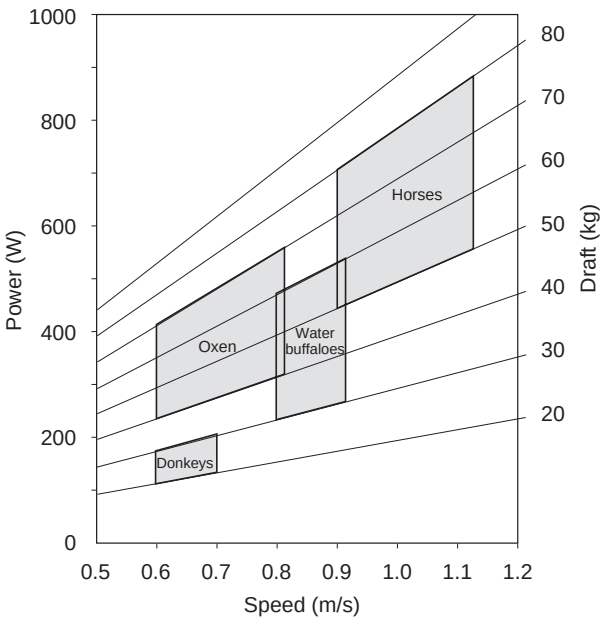


FIGURE 5 Typical performances of draft animals.

(500 kg), capable of sustained draft of approximately 50 kg, will thus not be able to adequately perform most field tasks alone. Teams of two, and later four, draft animals were common in European and American fieldwork, and efficient harnessing eventually resulted in more than 30 horses pulling the world's first combines during the 1880s.

Although the numbers of draft animals continued to increase in every country well into the 20th century, their combined power was rapidly overshadowed by new mechanical prime movers. In the United States, draft animals accounted for more than half of all prime mover power as late as 1870, but by 1950 their share was less than 0.2%. By 2000, the combined power of the world's half a billion draft animals was approximately 200 GW, less than 0.5% of the power installed in internal combustion engines.

5. BOVINES

Currently, bovines account for approximately two-thirds of the remaining draft power. More common reliance on bovines rather than on draft horses everywhere during the earlier stages of agricultural intensification, and in poorer countries and regions today, is easy to explain. Oxen (castrated males) may be difficult to train, less powerful than equally massive horses, and slower paced, but these

drawbacks are compensated by their stolidity, inexpensive harness, the lack of need for shoeing, and especially the fact that they are easy to feed. The efficient digestive system of these ruminants, with microbes (bacteria, fungi, and protists) in their rumen that are able to decompose cellulose of plant roughages that other mammals are unable to digest or can use only very inefficiently, means that during slack periods they can be kept fit only on grasses (fresh or dried) and crop residues (mostly cereal straws) that other animals find difficult or impossible to consume.

Consequently, ruminants require supplementary feeding by highly digestible concentrates (grains or milling residues) only during periods of hard work. Cattle, except for their susceptibility to trypanosomiasis, also perform better in the tropics than do horses. However, water buffaloes are unmatched in wet tropics. Their body mass ranges from 250 to 700 kg, but even the most massive buffaloes are quite nimble, moving easily on narrow dividers between rice fields. Their large hoofs and flexible pastern and fetlock joints made it easier to walk in the slippery, deep mud of rice fields. They eat many grasses, can graze on aquatic plants while completely submerged, require approximately 40% less feed energy per unit of gain than do cattle, can work reliably for at least 10–15 years, and children can take care of these docile beasts.

The energetic advantages of working bovines did not end with their draft power. After the end of their working lives, the animals became a valuable source of meat and leather, and during their lifetime they provided a great deal of organic wastes that were (usually after fermentation) returned to fields in all traditional agricultures that practiced intensive cropping. However, there were some drawbacks when working with bovines. Both the neck and head yokes used to harness draft cattle had a high traction point and required a pair of equally sized animals even when one beast would suffice for lighter work (Fig. 6). The slow pace of bovine work was another major limitation, particularly where multicropping was practiced, in which speed was essential for timely harvesting and planting of a new crop. For example, a pair of oxen working in a dry soil could plow 0.2 or 0.3 ha per day, whereas a good pair of horses could plow 0.5–0.8 ha. European experience during the 19th century was that a pair of good horses easily did 25–30% more fieldwork in a day than a team of four oxen.

In addition, oxen have neither the endurance nor the longevity of horses. The working day for many oxen was 5 h, whereas horses commonly worked twice as long. Also, although both oxen and horses started working at 3 or 4 years of age, oxen lasted usually less than 10 years, whereas horses could work for 15–20 years. Poor peasants were often forced to use their only head of cattle, a family cow, as a draft animal, compromising her milk production and getting only weak traction. More powerful oxen were, and remain, the most commonly used draft bovines: At least 250 million of them were working worldwide in 2000, mostly in Asia's two most populous countries.

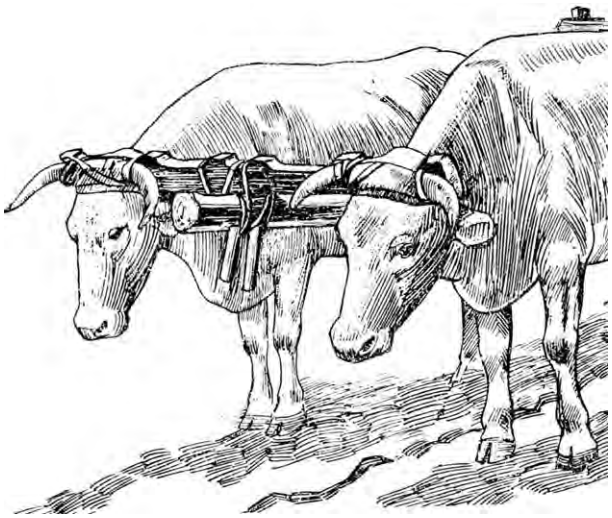


FIGURE 6 Head yoke commonly used for harnessing working oxen.

6. HORSES

There are five major reasons why horses are better draft animals than cattle. Unlike in cattle, frontal parts of horses' bodies are heavier than their rears (the ratio is approximately 3:2) and this gives them an advantage in inertial motion. Their unique arrangement of the suspensory ligament and a pair of tendons makes it possible to "lock" their legs without engaging any muscles and hence without any additional energy cost incurred during standing. Horses also grow generally larger and live longer than cattle, and they have greater work endurance. However, their superior strength was efficiently harnessed only with the widespread adoption of the collar harness and iron horseshoes, and their inherent power and endurance became readily available only with better feeding.

Fitted, and later also padded, collar harnesses provided a desirably low traction angle and allowed for the deployment of powerful breast and shoulder muscles without any restriction on the animal's breathing (Fig. 7). Its precursor was first documented in China in the 5th century of the CE, and an improved version spread to Europe just before the end of the first millennium. Another important innovation in harnessing was swingletrees attached to traces in order to equalize the strain resulting from uneven pulling: Their use made it possible to harness an even or odd number of animals. Iron horseshoes prevented excessive wear of hooves and they also improved traction.

However, collars and horseshoes alone could not guarantee the widespread use of horses: Only larger and better fed horses proved to be superior draft animals. The body mass, and hence power, of European draft horses began to increase only after the working stock benefited from several centuries of breeding heavy war animals needed to carry armored knights. However, these more powerful horses needed at least some cereal or legume grains, not just grasses fed to weaker animals and cattle. Production of concentrate feed required intensification of farming in order to provide yields high enough for both people and animals. Agricultural intensification was a slow process, and it first began in northwestern Europe during the late 18th century.

Even larger horses were taxed when pulling wooden ploughs, whose heavy soles, wheels, and moldboards generated enormous friction, particularly in wet soils. Moreover, the absence of a smooth, curved fitting between the share and the flat moldboard caused constant clogging by compressed soil and weeds. Iron moldboard ploughs were introduced

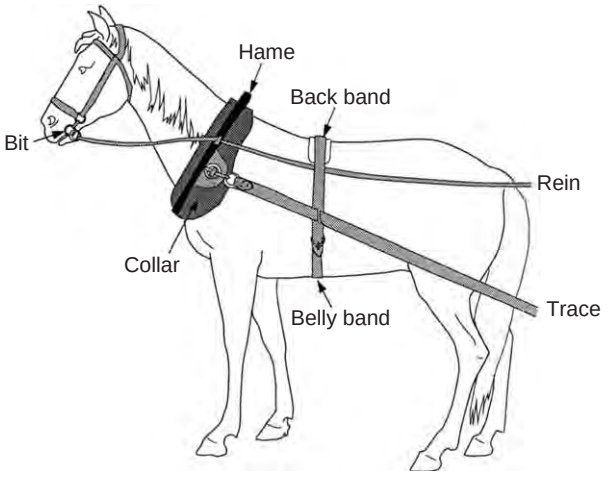


FIGURE 7 Basic parts of the collar harness.

to Europe from China during the 17th century, and cast iron shares were replaced with smooth, curved steel ploughshares by the mid-19th century during the rise of the modern steel industry. Horses then became the principal energizers of the world's largest extension of arable land that took place on the plains and prairies of North America, pampas of Argentina, and grasslands of Australia and southern Russia during the latter half of the 19th and the first half of the 20th century.

The heaviest horse breeds—French Percherons, English Shires, and German Rheinlanders—could work briefly at rates of more than 2 kW (approximately 3 horsepower), and they could steadily deliver 800–1000 W. Feed requirements of these animals were high, but their net energy benefit was indisputable: A horse eating 4 kg of oats per day preempted cultivation of food grain that would have fed approximately 6 adults, but its power could supplant that of at least 10 strong men. Larger and better fed horses also provided essential traction during the initial stages of industrialization, powering road and canal transport, turning whims in mining, and performing tasks in food processing and in numerous manufactures. In most of these tasks, they were displaced by steam engines before the mid-19th century, but even as the railways were taking over long-distance transport, horse-drawn carts, trucks, and streetcars continued to move people and goods in all rapidly growing cities of the late 19th century. Only the diffusion of internal combustion engines and electric motors ended the use of horses as urban prime movers.

However, in fieldwork, horses remained important well into the 20th century. Obviously, their largest numbers were supported where abundant farmland made it easy to produce the requisite feed. The total number of farm horses (and mules) peaked at 21 million animals in 1919 in the United States, when at least 20% of the country's farmland was needed to cultivate their feed. Subsequent mass adoption of tractors and self-propelled field machines was inevitable: Even small tractor engines could replace at least 10 horses and needed no farmland for support. The U.S. Department of Agriculture discontinued its count of draft animals in 1960, when approximately 3 million working horses were still left on American farms. However, in China, the total number of draft horses continued to increase even after 1980, when the country began its economic modernization. Horses, as well as water buffaloes, yaks, camels, and donkeys, remain important draft and pack animals in parts of Asia and Africa.

Virtually complete substitution of animate prime movers by engines and motors is one of the key attributes of modern high-energy society, in which humans act as designers and controllers of increasingly more powerful energy flows rather than as weak prime movers, and in which a rare working animal is an object of curiosity. Nothing illustrates the wide gap between the affluent and the subsistence areas of the world better than the fact that heavy, repetitive, and dangerous human exertions (including those of millions of children) and the draft of hundreds of millions of animals continue to be indispensable prime movers in large areas of the three continents on which modernity is still a promise for tomorrow.

SEE ALSO THE FOLLOWING ARTICLES

Conservation of Energy, Overview • Cultural Evolution and Energy • Human Energetics • Transitions in Energy Use • Work, Power, and Energy • World History and Energy

Further Reading

Aiello, L. C., and Wheeler, P. (1995). The expensive-tissue hypothesis. *Curr. Anthropol.* **36,** 199–221.

Alexander, R. M. (1992). "The Human Machine." Columbia Univ. Press, New York.

Brown, J. H., and West, G. B. (eds.). (2000). "Scaling in Biology." Oxford Univ. Press, Oxford.

Burstall, A. F. (1968). "Simple Working Models of Historic Machines." MIT Press, Cambridge, MA.

Durnin, J. V. G. A., and Passmore, R. (1967). "Energy, Work and Leisure." Heinemann, London.

Food and Agriculture Organization/World Health Organization/United Nations University (1985). "Energy and Protein Requirements." World Health Organization, Geneva.

Hochachka, P. W. (1994). "Muscles as Molecular and Metabolic Machines." CRC Press, Boca Raton, FL.

Kleiber, M. (1961). "The Fire of Life." Wiley, New York.

Lacey, J. M. (1935). "A Comprehensive Treatise on Practical Mechanics." Technical Press, London.

Landels, J. G. (1980). "Engineering in the Ancient World." Chatto & Windus, London.

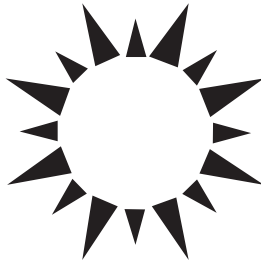
Needham, J., *et al.* (1965). "Science and Civilisation in China. Vol. 4, Part II: Physics and Physical Technology." Cambridge Univ. Press, Cambridge, UK.

Smil, V. (1991). "General Energetics: Energy in the Biosphere and Civilization." Wiley, New York.

Smil, V. (1994). "Energy in World History." Westview, Boulder, CO.

Whitt, F. R., and Wilson, D. G. (1993). "Bicycling Science." MIT Press, Cambridge, MA.

Corporate Environmental Strategy



BRUCE PIASECKI
American Hazard Control Group
Saratoga, New York, United States

1. Corporate Environmental Strategy
2. Beyond Compliance Initiatives
3. Making the Business Case for Sustainable Development
4. Relating Corporate Environmental Strategy to Corporate Change Initiatives
5. How CES Fits into a Normal Corporate Structure
6. Relating CES to Product Innovation and Strategy
7. The New Urgency for CES after Climate Change
8. Social Response Product Development
9. Case Example of CES in Practice
10. Summary

Glossary

corporate environmental strategy (CES) A functional business slogan represented by journals by the same name, owned by Elsevier, and a set of mimic journals and books utilizing the same title and concept. In addition, it is a measured business practice, starting after the tragic accident at Bhopal, India, the Exxon *Valdez* oil spill, and other environmental catastrophes at the end of the 20th century, that allows a firm to lessen its environmental liabilities, compliance costs, and product development costs as well as pursue competitive advantage in an organized and reportable fashion.

governance and management systems The means by which the board and key advisors of a firm manage its direction and performance. After the Bhopal and *Valdez* disasters, these systems increasingly involved environmental and public liability reduction efforts.

leading change, corporate change A studied area of management whereby systems are put in place to manage change in products, leadership, and corporate direction.

market-based environmentalism A term designed to suggest that leading firms based their search for environmental performance improvements on market opportunities, not just on environmental regulation and law.

social response product development Companies restructure their operations to actively shape market desires by creating new products that bridge the gap between traditional expectations of performance, safety, and environmental responsibility.

sustainable value creation A term developed after the demise of Enron and WorldCom in the early parts of the 21st century and designed to capture in a focus phrase the need for reform in how companies position themselves and report to the investment community on the paths they are taking to make their firms more sustainable.

Corporate environmental strategy (CES) involves the tools, management programs, processes, and product development choices that allow a firm to pursue competitive advantage through environmental management strategies.

1. CORPORATE ENVIRONMENTAL STRATEGY

Management scholars such as Deming and Juran spent several decades after World War II making sure that quantity and quality processes entered the plans of corporate strategy, along with the classical concerns of price, technical quality, and distribution matters. In a similar but often more diffuse manner, the proponents of corporate environmental strategy began, in the l970s through the l990s, to alter the standard decision models of corporate strategy to include externalities that challenged the future growth of corporations, such as new environmental regulations or irregularities in energy markets and pricing.

 © 2004 Elsevier Inc. All rights reserved.

This massive set of changes in corporate behavior was first motivated by the need to extinguish both legal and public liabilities of the major multinationals, after the tragic accident at Bhopal and the highly visible oil spill of the Exxon *Valdez*. Later in the 21st century, CES took on a more "proactive" or "sustainable value creation" thrust, as described in this article.

Classic post-Bhopal examples of CES include the following:

1. Honda's and Toyota's development of the hybrid powertrain automobile that uses an electric battery recharged by braking and the conventional combustion engine.
2. British Petroleum's (BP's) and Shell's investments in solar power and their "beyond petroleum" campaigns of the late 20th century.
3. A range of high-efficiency appliances from Whirlpool and Electrolux that go beyond regulatory requirements.

By 2003, CES was a visible and stated corporate strategy and management goal for many multinational corporations. Some suggest that CES became a stated mantra of the Fortune 500; others believe it was spotty in the top companies across the globe. In either case, CES began to have clout in leadership councils and began to tie together energy conservation, materials savings in product design, and the classical environmental regulatory elements.

The core innovations in CES in the last decade of the 20th century occurred in the heavy industries, such as petroleum refining, where firms such as Shell and BP competed on lessening their environmental footprint to attract new partners and new investment capital.

CES also spilled rapidly into the chemical industry, where firms such as DuPont, Celanese AG, and Dow offered new products, such as water-based emulsions.

2. *BEYOND COMPLIANCE INITIATIVES*

Competition in domestic and international markets and legal requirements led to the rapid development of CES in the late l990s. The approaches signified by this concept are not academic in nature and were not, on the whole, developed by academia, which in the l990s spent much of its intellectual energy developing concepts in sustainable development, industrial ecology, and alternative energy.

Instead, CES is an evolving practitioner concept that is defined by corporate practice and function and has been implemented directly by thousands of companies since Bhopal.

There are at least three competing forms of CES. These are as follows: grudging compliance by companies with regulatory requirements while seeking to turn a profit; going beyond regulations and standards while developing products that have market value and demand but are also environmentally sound; a mix of the first two—namely, the combination of willing compliance with innovation in the market.

Whereas most leadership councils felt in practice that the third mix was the best, some firms pursued the purity of the first two strategies into the new century. Then, market pressures, after the Enron and WorldCom scandals, encouraged the attention of corporate leaders on governance reforms and Sarbanes Oxley laws upgrading management systems, corporate reporting, and the overall transparency of modern corporate reporting.

During the first decade of the 21st century, these corporate reforms became attached to CES in complex and still to be studied ways.

Many attempts were made in the l990s and the new century to differentiate the essential or shared ingredients of CES, but most firms found that "one size does not fit all," as CES applications range widely from the mining and extractive industries (where the cost of energy is the focus), to automotive and petrochemical firms (where efficiency is often the key), to consumer-based firms (where firms such as Avon, Coca Cola, and Nike began to compete on elements of corporate social responsibility that are based on environmental strategies). As a result, what is shared is the approach, some tools and metrics, and general direction. Yet there is tremendous variance in results.

For example, CES has become a rallying term that often appears all too inclusive. In fact, sometimes CES seems, in sloppy parlance, a surrogate for classic business strategy and simply another way to make money by serving changing customer expectations. Some view it as a tool to measure emerging issues, where Intel or Hewlett Packard may use CES tools to track how their core customers are looking at new environmental risks in the metals and plastics used in computer and information technology systems.

Some corporations view CES as a tool for new product positioning, so that Rohm and Haas and Celanese AG are building emulsions plants in Asia that bypass the usual environmental legacy questions

from traditional manufacturing of adhesives. Others see CES as a pathway whereby a firm can achieve near-term shareholder value through corporate responsibility.

It is precisely because of this significant variance of results that the history and the readiness of firms to use the term CES have been at best checkered since Bhopal.

3. MAKING THE BUSINESS CASE FOR SUSTAINABLE DEVELOPMENT

By 2003, a number of advantages to CES were being investigated and documented by the management consulting profession and by groups such as the Global Environmental Management Initiative (GEMI) and the Conference Board. These advantages serve to explain why a firm would invest in efforts to move it beyond basic legal compliance with all laws.

Competitive advantages gained by environmental strategies include the following:

1. Margin improvement—seeking cost savings at every stage of the product life cycle through more efficient use of labor, energy, and material resources.
2. Rapid cycle time—reducing time to market by considering environmental issues as part of the concurrent engineering process during the early stages of design.
3. Market access—developing global products that are environmentally "preferable" and meet international ecolabeling standards in Europe, Japan, and other regions.
4. Product differentiation—introducing distinctive environmental benefits, such as energy efficiency or ease of disassembly, that may sway a purchase decision.

CES is the blend of classical environmental regulatory compliance matters with these new strategic dimensions and elements. CES, at its best, is an integral input in the creation of business strategy, product selection, and a corporation's public affairs and reporting under Sarbanes Oxley, and related laws of disclosure both before the U.S. Securities and Exchange Commission and before the investment community.

From l985 to 2003, different firms excelled in these four highlighted attributes: margin improvement, rapid cycle time, product differentiation, and market access. Intel was motivated to pursue CES to enhance its needs for rapid cycle time reductions, and marketing-based firms such as Nike and Proctor & Gamble set up CES goals for reasons of basic product differentiation. In the end, CES began as a means of improving corporate bottom lines through cost containment and new product opportunity programs.

If one were to estimate the degree of penetration of CES into corporate boardrooms, it kept its stigma as being an added cost, or what some have referred to as a "bag on the side of corporate strategy," for many years after Bhopal. Although the diligent and deliberate attempt to lessen a firm's environmental liabilities is a prime force behind CES, the relentless thirst for profit has proven a more reliable and less distracting guide in the new thinking regarding CES. The list of four key elements and goals of CES can be thought of as the turning point from the old century thinking into the new century, where the pressures for sustainable product mixes and policies became far more palpable.

A firm can achieve CES by a sequence of related efforts, including the following:

1. Make environmental management a business issue that complements the overall business strategy.
2. Change environmental communication within the company to use traditional business terms that reflect business logic and priorities.
3. Adopt metrics to measure the real costs and business benefits of the environmental management program.
4. Embed environmental management into operations, similar to the successful design-for-environmental models (here, environmental footprint reduction elements are entered into the equation at the product design stage).
5. Radically change the job descriptions and compensation of environmental managers—and line managers—to reflect the realities of doing business.

Others answered the question of how a firm would begin to move beyond compliance by pointing to the cost and administrative inconvenience of penalties and fines. The classic and more vivid example is the demise of Union Carbide, and its eventual purchase by Dow, after Bhopal, a tragic accident that killed over 2700 people in Bhopal, India. Lessons from Union Carbide for practitioners of improving environmental management systems, under the rubric of CES, include this summary of Union Carbide's major changes after Bhopal:

1. An intricate classification scheme of standards, regulations, and audits, which allows executives more

insight into the firm's liabilities, priorities, and high-risk areas. Technical staff functions of regulatory compliance can then be linked to the legal department's liability containment measures and strategies.

2. A large database of audit findings statistically amenable to analysis. From this corporate base, executives from different divisions compete for capital resources on more exact terms. Its size and consistency make it a management tool of considerable consequence.

3. A systematic monitoring and control program, more reliable than the traditional environmental, health, and safety systems. For instance, Union Carbide's worldwide computer program prioritizes risk at its 1200 facilities worldwide and assists senior management with its growth or phase-down plans.

4. New management tools, based on the audit system, provide quality assurance and executive reassurance that the company's environmental practices do not violate the law or threaten human health or the environment. Nationwide, more than 200 executives have served jail time for environmental crimes. This executive assurance program—developed for Union Carbide by A. D. Little—helps senior management feel "in touch" with the firm's liabilities, compliance issues, and long-term viability.

Other recognized drivers in corporate behavior for corporate environmental strategy included Superfund Authorization and Reform Act (SARA) Title III, which put public pressure on companies by requiring disclosures on chemical stockpiles, warehouses, and such, and the Chemical Manufacturers Association's responsible care initiative, which outlined a code of behavior.

The disruption caused by the Union Carbide accident spurred the above regulations and conduct codes throughout the industry. The public lost faith in the industry and the negative opinion polls increased. Corporate environmental strategy is, in part, a response to the negative public image and public concerns about environmental, health, and safety performance.

4. RELATING CORPORATE ENVIRONMENTAL STRATEGY TO CORPORATE CHANGE INITIATIVES

There are several options that, ultimately, must be integrated when attempting to realize a change in corporate culture: the legal remedy, which is necessary to spur on the slow movers; the money remedy, which does not mean blindly throwing money at an environmental problem, but instead taking a more strategic approach; and finally, the regulatory remedy, which moves beyond simple regulatory compliance to a level of strategy that allows a company to stay ahead.

Some feel that without law there would be no corporate environmentalism, and yet there are those that stood at the forefront of an issue, pushing it to become law. These leaders from industry, academia, government, and the environmental movement are leading CES more than the law is, and for a company to find success through CES, it must stay well ahead with those leaders, not lagging behind with the laws. The law is functional and offers a glimpse at a company's next step, but this is not strategy.

The money remedy does not translate, as some might believe, to throwing money away, the way some companies have done. It is about information management, including audit information, environmental liability and insurance information, product selection, announcement, and positioning information, stockholder and stakeholder information, new and international market information, regulatory and government liaison information, technical information, legal information, and strategic and financial information. Without the full integration of these nine types of information, the amount of money invested in them will not matter.

Finally, there is the regulatory remedy, which can be seen in companies who have begun mass disclosure measures to create corporate transparency, well beyond simple regulatory compliance. Strategy here allows leaders to stay ahead of environmental issues and avoid the old way of constantly cleaning up past mistakes; both company executives and the public are able to see the company's liabilities and assets more easily. This also creates an environment of corporate trust for the public.

These three remedies together can create a strategic advantage for companies willing to make the effort of changing "business as usual." Table I outlines a list of five elements of success in shifting corporate culture.

Table I was based on a survey by the author's firm of over l50 companies of the Fortune 500. Whereas the five corporate change elements noted above are quite inclusive, and extend well beyond what lawyers or engineers in a firm would normally consider environmental management, they do capture the "strategic" or management and leadership council feel of CES in practice.

Without a doubt, the actual benefits of CES are most often kept quite confidential, as they are close enough to the heart of a corporation's business

strategy, when they are functional and not based on public relations, that they are kept close to the chest of the leading executives in a firm. This makes it difficult for scholars in the field to do much in their diagnostics except case studies and retrospective summaries of these elements of success.

5. HOW CES FITS INTO A NORMAL CORPORATE STRUCTURE

After Bhopal and the Exxon *Valdez* spill, it became clear that environmental management was not sufficiently integrated into normal business functions, reviews, and organizations. A l991 survey, seen in Fig. 1, revealed some interested gaps. What the study indicated, as it was replicated for different firms over subsequent years, is that there is no natural or ultimate home for CES in an organization. Depending on the firm, it might be centered in the legal, risk reduction, or regulatory divisions of a firm. But consistently, most firms feel there is a wall between these CES functions and the financial disciplines that run a firm. As a result, CES functions are often perceived as being of secondary or tertiary importance in many firms.

TABLE I

Elements of Success in Shifting Corporate Culture

1. Role for strategic recognition:
Including mission statements and performance reviews
2. Role for changes in staff:
Title and function
3. Adaptations in conventional management tools:
Such as audits or return-on-investment equations
4. Increasing role for external strategies:
Such as community relations and voluntary disclosure
5. The unexpected role for strategic alliances

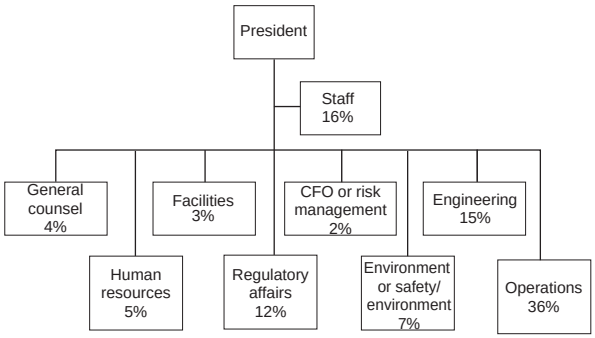


FIGURE 1 Reporting requirements in environmental management. From Coopers and Lybrant (1991), "Environmental Management in the 1990s: A Snapshot of the Profession." Reprinted from Piasecki (1995), with permission from John Wiley & Sons, NY.

6. RELATING CES TO PRODUCT INNOVATION AND STRATEGY

In the pursuit of superior cars, electronic products, and computing, several leading multinational corporations began in the last quarter of the 20th century to tie CES to their product strategy. Leaders such as Toyota and Honda are classic examples in the automotive industry, as are Shell and BP in the petroleum sector.

Those firms that had successfully integrated CES into their normal business functions by the new century shared a common set of attributes. These attributes are described below as "generic" elements of allowing CES to be elevated within a corporate setting, as they often depend on organizational dynamics or a unique set of executive interests and needs. These organizational elements of CES involve "action plans" or goals, such as the following:

1. A role for strategic recognition including mission statements and performance reviews.
2. A role for changes in staff, title, and functions where chief environmental officers and their staff become more visible and respected in corporate organizations.
3. Adaptations in conventional management tools, such as audits or return-on-investment equations.
4. An increasing role for external strategies, such as community relations and voluntary disclosure efforts.
5. In some of the more curious and interesting developments, CES also involved the unexpected role of strategic alliances, where past enemies meet.

Some individuals, such as Steve Percy, the former CEO and Chairman of BP America, have claimed that BP's purchase of Amaco was motivated by CES, since BP saw and appreciated the large natural gas holdings of Amaco and their solar research portfolio as an opportunity to advance BP's need for progress in CES as a reputation enhancer.

Other examples of these strategic alliances were found in the new century in utilities such as Exelon, chemical giants such as Dow, Celanese, and DuPont, and select automakers such as Toyota and Honda.

CES includes knowing when to play by the conventional corporate and financial rules and when

to change the game. In a small set of successful corporations, this involved the integration of "new rules" with the ever-changing game of strategy, product positioning, and product reformulation. Evidence of such "radical" or "incremental" changes became hotly debated with GEMI companies, for example. In the new century, GEMI extended its sunset clause and allowed its continued existence in order to provide a range of companies with these stated requirements in their membership rules.

GEMI companies must reach toward corporate environmental strategy:

- By sharpening existing tools and measurement systems to help managers refine and enhance their environmental performance;
- By popularizing to the press and the public the belief that corporations have a role in the search for environmental excellence, both domestically and internationally;
- By explaining how knowledge from certain business sectors can be shared, coordinated, and then adapted across different industrial sectors and nations;
- By asserting boldly that environmental strategy involves more than regulatory compliance.

7. THE NEW URGENCY FOR CES AFTER CLIMATE CHANGE

Although it is significant that many of the above four imperatives are qualitative, not quantitative, such new efforts have begun to add an elevated level of urgency to corporate environmental strategy.

Whereas the Bhopal and *Valdez* tragedies may have precipitated the birth of CES, global environmental concerns, such as water shortages in China, the energy crisis and electricity grid problems in the northeastern United States, and especially the complexities surrounding the economic impacts and costs of climate change, have begun to add a new turn-of-the-century urgency to the debates over CES.

By 2003, CES began to help firms link price, technical performance, and social response product development. In other words, conventional business tools focusing on competitive price and product strategies, in some instances, began to notice the appeal of shaping or reformulating products relative to external energy and environmental threats, such as limitations in water supply, spiking energy costs, or climate change and other carbon-related challenges. Figure 2 sums up these evolving links between CES and product development.

When multinational firms such as Toyota, Honda, Electrolux, and Whirlpool began making claims before the investment community that their efforts in corporate environmental strategy were paying off, the new emphasis on linking CES to product development and differentiation began to be in vogue.

8. SOCIAL RESPONSE PRODUCT DEVELOPMENT

Historians of business know that vogue is an element of consequence, as evidenced by the passion for Six Sigma tools that General Electric brought to firms as different as 3M, Dow, and Celanese AG. This new product-based vogue for CES enabled a more rapid embrace of CES in select companies, as illustrated in Fig. 3.

"Social response" (i.e., incorporating the sustainability dimension) product development means that instead of manufacturing products solely in response

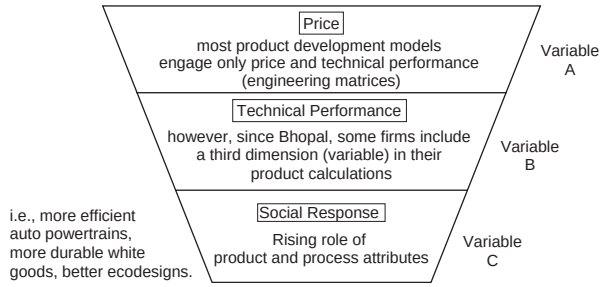


FIGURE 2 Social response product development.

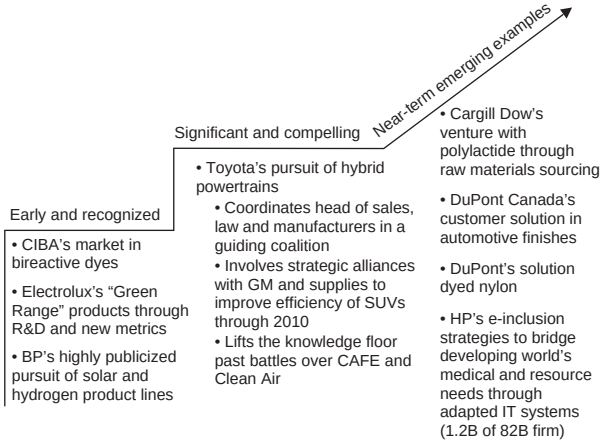


FIGURE 3 Defining social response product development through select examples.

to consumer preferences as discerned and measured by price and technical performance through market research methodologies, companies restructure their operations to actively shape market desires by creating new products such as Toyota's and Honda's hybrid cars.

Although such efforts are logically supported by new manufacturing, quality control, and resource management techniques, what is new is that they bridge the gap between traditional expectations of performance, safety, and environmental responsibility. Such expanded product development models allow a more sustainable value creation at present and through the coming decades. Social response product development also allows product differentiation in mature product lines in a fashion that attracts new capital and new forms of partnering. Examples here include Toyota, BP, Shell, and DuPont. A sustainable value approach, based on this kind of product differentiation and development, supports all other actions in the company. Examples of social response approaches include the following:

- From 1996 to 2001, Anheuser-Busch established a multiyear annual "Pledge and Promise" staff and stakeholder retreat to award eco-innovation teams. The retreat includes presidents and chief executive officers of key Anheuser-Busch stakeholder groups as external judges to select top five sustainable value leaders each year.
- From 1998 to 2002, British Petroleum established a multiyear technical task force to identify, select, and finance sustainable value product lines and plans.

Social response product development shapes how executives approach their task of accomplishing seemingly disparate goals. In chemicals, in high-performance plastics, and in new markets for cars and computers, social response product development delivers both resource efficiency and customer-demanded performance. As an example of social response product development being used as a CES strategy, Toyota's search for a superior automobile is explored.

9. CASE EXAMPLE OF CES IN PRACTICE

In this new century, there is considerable pressure on the top six automakers to reduce their environmental footprint. The automaker that wins the race to build and sell the superior car will shape consumer preferences, thereby boosting sales and profits. The winning firm will fashion a corporate strategy that drives automobile emissions to near zero while simultaneously providing high levels of performance, safety, and comfort.

The model of social response product development as a CES strategy is built on four dimensions, like the structure of a house: knowledge depth—the ever-growing basement; knowledge durability—the expanding sides; knowledge dependence—the ceiling of use; knowledge floors—the ground base. The breakthrough at Toyota was recognizing that new technologies mean that a new knowledge floor can be laid, one that makes it possible to meet all consumer expectations of performance, safety, and environment (fuel efficiency, emissions, resource efficiency, alternative fuels) without the tradeoffs previously required.

Toyota and Honda have used a new form of CES to change the rules of the game by leveraging their knowledge depth to introduce superior new cars, such as the four-door, five-passenger Prius. The Prius achieves fuel efficiency rates of 55 miles per gallon (mpg), twice the regulated corporate average fleet requirement. The move to hybrid cars—those that can run on multiple fuels, most commonly, gasoline and electricity—will be understood as truly exceptional when the realization sinks in how absolutely ordinary it will make the hundreds of millions of cars now littering the current knowledge floor. This floor—containing everything from gas-guzzling sport utility vehicles and commercial trucks to Europe's newest smart cars that are still dependent on the 100-year-old combustion engine—has been permanently superseded.

Some firsts of the Toyota Prius include the following:

1. The Prius has been designed to be recycled. Although this is most relevant to Japanese and European consumers, it does highlight how forward-thinking Toyota has become. The amount of waste avoided through this redesign is staggering, separating Toyota from General Motors, Ford, and DaimlerChrysler.
2. At 2003 gas prices, the average North American would save at least $1/day, or a total exceeding $360, at the pump by driving a Prius rather than a typical compact sedan.
3. Consumers will enjoy the convenience of driving 600 miles between fill-ups.
4. The American version of the Prius has a battery pack with 20% more power than the Japanese equivalent, yet weighs 20% less.

5. The Prius has fundamentally changed the debate over clean cars, almost eliminating the shift to all-electric cars.

6. The Prius has already become legend. It can be found at the New York Museum of Modern Art's exhibit "Different Roads." In June 1999, Toyota became the first automaker in the world to receive the United Nations Global 500 Award, not only because of the Prius, but also because of all of Toyota's environmental commitments to reduce dependence on landfills.

What is noteworthy about this case example is that much of this innovation occurred in anticipation of changes in corporate average fuel economy standards and related air pollution legislation across the globe. This represents a new form of CES as it involves aggressive product change before the regulations take final shape, as a new means of competitive advantage.

10. SUMMARY

Although it would be historically inaccurate to claim that the above-described efforts of BP, Toyota, and DuPont represent a norm, they do signify a way in which corporate environmental strategists are succeeding in the early part of the 21st century.

More and more corporate boards, after the Bhopal and *Valdez* accidents and following a second wave of reforms starting after Sarbanes Oxley required a more accurate disclosure of a firm's lasting liabilities and product assets, are taking this series of steps in linking CES to their larger corporate and social responsibility obligations and plans. Their eventual success will prove to be based on linking their mission and goals, originally compelled by a desire to meet regulations and limit a firm's liability, with products.

As in most corporations, where the product is king, this move may prove the means by which CES reinvigorates itself in the new century.

Acknowledgments

Celeste Richie, of American Hazard Control Group, is gratefully acknowledged for performing the research for this article.

SEE ALSO THE FOLLOWING ARTICLES

Business Cycles and Energy Prices • Clean Air Markets • Consumption, Energy, and the Environment • Ecological Risk Assessment Applied to Energy Development • Energy Services Industry • Risk Analysis Applied to Energy Systems • Service and Commerce Sector, Energy Use in • Subsidies to Energy Industries • Trade in Energy and Energy Services

Further Reading

American Hazard Control Group. (2002–2003). "Corporate Strategy Today," Executive Monograph Series. Lexis Nexis, Saratoga Springs, NY.

Baram, M. (1991). "Managing Chemical Risks." Tufts Center for Environmental Management, Bedford, MA.

Elsevier. (1994–2003). *Corp. Environ. Strategy.*

Epstein, M. (1995). "Measuring Corporate Environmental Performance." The IMA Foundation for Applied Research, New York.

Friedman, F. (1997). "Practical Guide to Environmental Management," 7th ed., pp. 68–70. Environmental Law Institute, Washington, DC.

Hoffman, A. (2000). "Competitive Environmental Strategy." Island Press, Washington, DC.

Kotter, J. (1996). "Leading Change." Harvard Business School Press, Boston, MA.

Laszlo, C. (2003). "The Sustainable Company: How to Create Lasting Value through Social and Environmental Performance." Island Press, Washington, DC.

Piasecki, B. (1995). "Corporate Environmental Strategy: The Avalanche of Change since Bhopal." Wiley, New York.

Piasecki, B., and Asmus, P. (1990). "In Search of Environmental Excellence: Moving beyond Blame." Simon & Schuster, New York.

Renfro, W. (1993). "Issues Management in Strategic Planning." Quorum Books, Westport, CT.

Cost–Benefit Analysis Applied to Energy

J. PETER CLINCH
University College Dublin
Dublin, Ireland

1. **Definition of Cost–Benefit Analysis Applied to Energy**
2. **History of Cost–Benefit Analysis**
3. **Rationale for Cost–Benefit Analysis Applied to Energy**
4. **Cost–Benefit Methodology**
5. **Valuation of Energy Externalities**
6. **Further Issues in the Application of Cost–Benefit Analysis to Energy**

Glossary

benefits transfer A process of taking environmental valuation estimates from existing studies and applying them outside of the site context in which they were originally undertaken.
discounting Adjusting a cost or benefit in the future so that it may be compared to costs and benefits in the present.
energy externality A cost or benefit that arises from an energy project but is not borne or appropriated by those involved in the energy project (e.g., pollution from energy generation).
environmental valuation The process of placing monetary values on environmental effects.
revealed preference methods Environmental valuation methods that use data from actual behavior to derive values for environmental assets.
sensitivity analysis A presentation of various possible outcomes of a future project or policy under conditions of uncertainty.
shadow pricing The process of adjusting a distorted market price to better reflect the true opportunity cost to society of using that resource.
stated preference methods Direct environmental valuation methods that ask people in a survey to place a value on an environmental asset.
total economic value A categorization of the various elements of value of a resource, project, or policy.
value of statistical life The translation of the value of a small reduction in the risk of death into a value for a project that would involve the equivalent of one life saved or lost.
welfare criterion The criterion in a cost–benefit analysis to assess whether a project or policy benefits society.
welfare measure A measurement tool to assess the strength of consumers' preferences for or against a project or policy in order to assess its costs and benefits.

Cost–benefit analysis applied to energy is the appraisal of all the costs and all the benefits of an energy project, policy, or activity, whether marketed or not, to whomsoever accruing, both present and future, insofar as possible in a common unit of account.

1. DEFINITION OF COST–BENEFIT ANALYSIS APPLIED TO ENERGY

Cost–benefit analysis (CBA), also known as benefit–cost analysis, is rooted in applied welfare economics. It is a way of organizing and analyzing data as an aid to thinking. It provides a set of procedures for comparing benefits and costs and is traditionally associated with government intervention and with the evaluation of government action and government projects. The underlying rationale for CBA is rational choice; that is, a rational agent will weigh the costs and benefits of any proposed activity and will only undertake the activity if the benefits exceed the costs.

A confusion that sometimes exists relates to the difference between social cost–benefit analysis and cost–benefit analysis. However, to an economist, social impacts and environmental impacts are economic impacts, so any analysis purporting to be a CBA must address social and environmental impacts.

 © 2004 Elsevier Inc. All rights reserved.

In this way, CBA differs from financial analysis in that it assesses the costs and benefits to the whole of society rather than to a portion of society. For example, a financial analysis of a proposed new power plant by a private energy supplier will not take into account the costs imposed on wider society (such as environmental impacts), whereas a CBA, by definition, will do so. In addition, there are various other subtle but important differences between financial analysis and CBA related to the fact that market prices do not always reflect the true value of goods and services to society. This provides the basis for a technique known as shadow pricing, which adjusts distorted market prices such that they may be used to value costs and benefits appropriately.

As the previous definition implies, a CBA must include all the costs and all the benefits to whomsoever accruing. Thus, the definition instructs that whether the costs or benefits are marketed or not makes no difference. In other words, even if some of the costs or benefits are not exchanged in markets at all and so do not have a price, they may still be of value. This provides the basis for the inclusion of environmental and wider social impacts in CBAs of energy projects and policies. The common unit of account in CBA is money. Therefore, insofar as is possible and reasonable, all impacts, including environmental and social, should be evaluated in monetary terms so that they can be compared with one another. This provides the basis for a technique known as nonmarket or environmental valuation—that is, placing monetary values on goods and services not exchanged in markets, such as environmental improvements and damages.

In many cases, the costs and benefits of a project or policy will arise at different times. This is the case in relation to most energy activities, where the consequences of actions may not be known until far into the future, particularly in the case of environmental impacts such as global warming. This provides the basis for a technique known as discounting, which allows future costs and benefits to be compared with those in the present.

CBA has been used *inter alia* to estimate the environmental costs of power generation, the net social benefit of different power supply options, the net benefit of the use of renewable energy sources (e.g., solar, wind, and hydroelectric power), the net benefits of alterative modes of transport, the benefits of energy efficiency programs in the commercial and domestic sectors, as well as for policy analysis of energy efficiency regulations such as energy rating and labeling.

2. HISTORY OF COST–BENEFIT ANALYSIS

Detailed histories of CBA can be found elsewhere, but the essence is as follows. In 1808, U.S. Secretary of the Treasury Albert Gallatin gave what is believed to be the first recommendation for the employment of CBA in public decision making when he suggested that the costs and benefits of water-related projects should be compared. In 1936, the U.S. Flood Control Act deemed flood control projects desirable if "the benefits to whomsoever they may accrue are in excess of the costs." In 1950, after a debate on the foundations of CBA in welfare economics, the U.S. Federal Inter-Agency River Basin Committee published a guide to CBA known as the "Green Book." From then on, there was considerable interest in CBA and the development of cost–benefit rules, with seminal contributions by a number of authors, particularly during the 1970 s. Although the distortions created by government intervention in the 1970 s provided much room for the application of CBA, the technique went through a lull in the early 1980s when structural reform replaced piecemeal intervention. However, CBA has made a comeback since then, mainly as a result of increased public awareness of environmental degradation and the associated demand that this cost be accounted for in analyses of projects and policies. However, for a number of years, the economist's view that environmental impacts could be assessed in monetary terms had little impact on policy, with many environmentalists having a particular problem with this concept. For example, in the U.S. Clean Air Act (1970), it was forbidden to use CBA, and the U.S. Clean Water Act (1972) set the objective of the elimination of all discharges into navigable waters. However, in the 1980 s, it became a requirement that the U.S. Environmental Protection Agency perform CBAs on all environmental directives to assess their economic efficiency and significant resources were devoted to researching cost–benefit techniques and methods of placing monetary values on the environment. Although it was not until the Single European Act was introduced in 1987 that a legal basis for European Union (EU) environmental policy was provided, the nature of environmental policy reflected the origins of the EU being an economic one. EU policy that "the potential benefits and costs of action or lack of action must be taken into account in the preparation of policy on the environment" provides an obvious platform for the use of CBA.

3. RATIONALE FOR COST–BENEFIT ANALYSIS APPLIED TO ENERGY

The need for CBA applied to energy results from the existence of market failure. If there were no market failure and all energy suppliers were privately owned, there would be no need for CBAs of energy projects and policies. If market signals were appropriate, those activities and projects that are in the best interests of society would be implemented automatically and financial analyses by firms would simply replace any need for CBAs. However, the market does not work perfectly. Market failure necessitates CBA of energy projects, policies, and activities as a result of the existence of externalities, the associated need for environmental policy instruments, state ownership of energy generation and supply, assessment of sustainability, and "green" national accounting.

3.1 Energy Externalities

One of the most significant areas of environmental concern today relates to the environmental impacts of energy generation from fossil fuels, particularly acidification precursors, greenhouse gases, and localized air quality impacts. Much of these emissions result from power generation and the transport sector. These environmental impacts occur at local, national, regional, and global levels.

The environment can be defined as those parts of our physical and psychological endowment that we somehow share, which are "open access." An economic analysis shows that it is precisely this shared nature of many environmental endowments that threatens their quality and character. In a market system, goods are allocated by the price mechanism. In a free market, the price of a good is determined by the demand for, and supply of, the good. Price is therefore a reflection of society's willingness to pay for, or the valuation it places on, the good in question. However, the shared nature of many environmental assets, such as air, the atmosphere, and water resources, means that they are not owned and so do not have a price. When goods are seen as free they are overused. Thus, the market fails to protect environmental assets adequately.

This failure of the market can be best demonstrated with a simple example: If a power plant produces SO_2 emissions resulting in acid rain, which causes forest damage and damage to buildings, in a free market these costs will not be reflected in the cost of power generation. Thus, the price of the energy will not reflect the true cost to society of generating it. In this case, more energy will be generated and more of the environmental asset depleted than is in the best interest of society. When a cost is imposed on those other than the person or business that produces the cost, this is known as an external cost. Of course, many energy projects will reduce such externalities, such as through renewable energy sources and improved heating and insulation systems.

When assessing the social efficiency of energy activities using CBA, it is necessary to include these externalities and other wider societal impacts. However, because by definition such costs and benefits are not traded in markets, there is no price to reflect their value. This provides the basis for the tools of environmental valuation.

3.2 Calibration of Environmental Policy Instruments

Unless some mechanism is found to internalize these externalities, they will cause a misuse of society's scarce resources. This provides the basis for government intervention in the market. The notion of externalities was addressed by economists as far back as the 1930s. Pollution was seen as the consequence of an absence of prices for scarce environmental resources and so the prescription for solving the problem was to introduce a set of surrogate prices using environmental (Pigouvian) taxes and charges. A large literature has developed on these so-called market-based instruments. One of the more controversial concepts in environmental economics is that there is an optimal level of pollution. Pollution is usually the by-product of a useful activity, such as the production of goods and services; in other words, it is an input into the production process. Reducing pollution usually involves a cost in terms of output forgone and/or the cost of the technical process of abatement. Therefore, there is a trade-off between the environmental improvement and the costs of pollution reduction. The optimal level of pollution occurs where the benefit of another unit of abatement (the marginal benefit) equals the (marginal) cost, with the (marginal) cost being higher after that point. In the case of an extremely hazardous pollutant, the optimal level of pollution will be zero or close to it, but in most cases the optimal level of pollution will be positive. CBA provides the tools to assess these trade-offs. In reality, few economic instruments are actually based on such valuations. Research on so-called target setting using

environmental valuation is still in a preliminary stage. However, it could be used to assess the optimal level of an energy tax.

3.3 State Ownership of Energy Generation and Supply/State Support of Energy Efficiency Initiatives

In some countries, energy supply companies are owned by the state. In some cases, the original rationale for such ownership was that no private company would bear the large fixed costs of developing an energy supply system. In addition, it is sometimes suggested that state companies can perform as natural monopolies and thereby exploit economies of scale while producing at minimum average cost and subsidizing energy for certain sections of society. CBA has a role to play in assessing the merits of such state initiatives and is particularly relevant when assessing the merits of government intervention to improve energy efficiency, for example, by subsidizing wind power or a domestic energy efficiency program.

3.4 Sustainability Assessment and Green National Accounting

CBA is useful in assessing sustainability; that is, what is the value of our resource stock, how is it being depleted, and is man-made capital that is replacing it of equal value? Green accounting adjusts traditional measures of economic performance such as gross national product to reflect environmental damage and the depletion of resources. In relation to energy, CBA may be used to assess the social costs and benefits of depleting energy resources such as fossil fuels.

4. COST–BENEFIT METHODOLOGY

4.1 Welfare Criterion

In using CBA to assess the costs and benefits of a project or policy to society, it is first necessary to define the social efficiency criterion. The most frequently used criterion in microeconomic theory is Pareto optimality. Using this criterion, a project would be deemed to pass a CBA if it made at least one person better off without making anybody else worse off. For an investment to pass the Pareto optimality test, compensation must be paid to those bearing the costs so as to leave them indifferent between their welfare *ex post* and *ex ante*. Most policies and investments leave somebody worse off and thus most would fail to meet the Pareto criterion. Frequently in CBA, the Kaldor–Hicks test is used. This test embodies the potential compensation principle. Kaldor categorized an action as efficient if those who gain from the action could compensate the losers and still be better off. Hicks framed the question slightly differently and labeled the action efficient if the potential losers from the action could compensate the potential beneficiaries of the action for not initiating the action. In both cases, the action is an improvement regardless of whether the compensation is actually paid. Thus, if the CBA shows the project to result in a welfare improvement under the Kaldor–Hicks test, this is essentially a potential Pareto improvement since there is the potential, in theory, to pay compensation to the losers such that nobody would be worse off. This will be the case as long as the benefits of the project outweigh the costs (i.e., the benefit:cost ratio is greater than 1).

Whether or not this potential payment is enough to show that a project is good for society has been the matter of much debate. Some authors have suggested that despite the shortcomings of the Kaldor–Hicks criterion, it is often useful to measure the effects of a change in the total value of output independently of the distribution of output; however, in judging the social efficiency of a project, equity aspects should also be considered. A two-part test may be used in which a project or policy is put to the Kaldor–Hicks test and then tested to determine if it improves (or at least does not disimprove) the distribution of income. Approaches that consider questions of equity therefore allow the possibility of acceptance of a project or policy that returns a negative figure for the sum of individual welfare changes if the gains from income distribution outweigh these losses. However, when considering equity the difficulty is in the assignment of weights to individual welfare changes. Consideration of equity issues and distributional consequences are particularly relevant in the case of projects that affect the price of energy and/or health. Energy price increases may be regressive unless poverty-proofing measures are put in place because energy is a necessity and therefore requires a larger proportion of the budget of less well-off people. On the other hand, improvements in domestic energy efficiency may bring significant health and comfort benefits to the elderly, who in some countries have been shown to be more likely to live in energy-inefficient homes and therefore more likely to be subject to winter morbidity and mortality.

4.2 Welfare Measures

In order to assess the costs and benefits of a project or policy, it is necessary to examine the strength of consumers' preferences for or against the project. The standard approach is to measure the benefit an individual obtains from a project or policy by the maximum willingness to pay on his or her part for such a benefit or the maximum willingness to accept compensation for not receiving such a benefit. Similarly, a cost can be measured by the maximum willingness to pay on the part of an individual to avoid such a cost or the maximum willingness to accept compensation for such a cost being imposed on the individual. Such welfare measures may be approximated using shadow pricing and/or direct valuation methods from surveys.

4.3 Temporal Considerations

Account must be taken of the fact that costs and benefits arise at different times. Thus, future costs and benefits must be discounted so that they can be compared with present costs and benefits. The rate of return for an investment can be strongly affected by the discount rate if there is a long time span involved. Therefore, the choice of discount rate is very important. In some estimates of damages from global warming resulting from the burning of fossil fuels, a small change in the assumed discount rate will have significant implications for climate change policy. However, despite the enormous literature on discounting, it is a difficult task to calculate the social rate of discounting. The discount rate is measured by some combination of a social rate of time preference and a measure of the opportunity cost of capital. Economists have been unable to provide a definitive figure for the social rate of discount, and so a rate tends to be chosen rather than measured.

In an application of CBA to any project or policy, a range of discount rates should be used. Often, for a government project, a published government test discount rate is used for making policy recommendations. These vary by jurisdiction. For example, UK treasury guidelines for the opportunity cost of public investment take the rate of return to marginal private sector investment as 6%. For pricing output sold commercially, public sector agencies are expected to achieve a required average rate of return of 8% to be consistent with the private sector. In the United States, recommended rates for investment projects vary. The Office of Management and Budget recommends a rate of 10%, whereas the General Accounting Office prefers a rate of 2.5% based on government bond yields. The World Bank has traditionally used an 8% rate in project analysis.

4.4 Uncertainty and Risk

In *ex ante* CBAs, investments with long time horizons are subject to uncertainty, particularly in relation to future prices. In addition, when evaluating costs and benefits *ex ante*, there may be considerable uncertainty regarding the magnitude of various parameters. This is particularly the case in relation to potential environmental impacts and their valuation. In cases in which there is a high degree of uncertainty in relation to the coefficients being used to give the monetary value of outputs of a project, lower bound estimates are sometimes used because, if the results are to be the basis for the recommendation of a change in policy, it is thought by some authors that it is better for the benefits to be underestimated rather than overestimated. It is important to note that lower bound results do not necessarily imply the lowest possible results. Rather, they denote the lower limits of a range of reasonable estimates.

Sensitivity analysis, whereby results are calculated for various scenarios, should be used in all CBAs. In addition, confidence intervals should be presented where possible. A general rule is that the more results the better while giving guidance and reasons for the particular estimates chosen by the author(s) for the purposes of recommendations based on the CBA results. Risk assessment may be incorporated by including probability estimates for various scenarios.

4.5 Shadow Prices and Wages and Community Stability

Shadow prices must be used when there are distortions within an economy that render certain prices invalid as measures of the social opportunity costs and benefits of using resources in the project in question. In the real world, market prices may not reflect the social value of goods or services because of distortions in the market, such as monopoly power (which inflates price), indirect taxes and subsidies (unless they are correcting some other distortion), and unemployment. Thus, when calculating each value of a project, it is necessary to assess the appropriateness of using the market value and, where necessary, these values should be adjusted for market distortions. This can be a problem in energy projects for which a particular piece of technology may be supplied by one

company and/or estimates must be made regarding how price would decline should such a piece of technology be widely adopted, for example, in a national domestic energy efficiency retrofit.

The approach to valuing labor is worth detailing. In reports purporting to be CBAs, labor or "employment" is sometimes treated as a benefit rather than a cost. Labor is an input into a project and, therefore, a cost even though it may have some additional benefits such as community stability, as discussed later. However, that cost may be less than the market wage would suggest. If labor markets clear, the shadow price of labor will equal the market wage. Any increase in employment in one sector of the economy will merely reduce the availability of labor to another sector. This is likely to be the case when the jobs are highly skilled and/or there is a shortage of such skills. However, for a country with a high unemployment rate, it could be argued that an increase in the demand for labor in one sector of the economy will not necessarily displace a job in another sector; that is, there may be employment additionality whereby if the new job is filled by a person who was previously unemployed, a very low cost (not zero, unless it is considered that the unemployed's spare time is totally unproductive) in terms of output forgone is imposed on society. In this case, the shadow price of labor would be close to zero. It has been the practice of some government CBAs to assume a shadow price of labor of zero. However, this is against the international practice of setting the shadow price of labor at most a fraction below the market wage. For example, in Canada, the shadow wage is usually 95% of the market wage, and in the United Kingdom they are usually considered equal. In many energy projects, specialized skills are required and so a shadow wage close to the market wage will usually be appropriate. However, in some household energy efficiency improvement schemes, long-term unemployed people have been trained how to fit various sorts of insulation, for example. In this example, a case could be made for using a lower shadow wage rate.

A related value that should be considered is whether an energy project may contribute to community stability. For example, would the provision of a new power plant in a rural area provide jobs such that depopulation of that area would halt? This value is not the value of the job per se. Rather, it is the added value of a job being in a particular place. These values are difficult to assess but are real. For example, it is often observed that employees will be willing to accept lower salaries to live in their hometown.

4.6 Valuing Changes in Energy Use

When calculating the benefits of reductions in energy use, the price of energy may be used to value such savings. However, it may be necessary to shadow price such savings. If the state supplies energy, an assessment must be made of whether there is any form of subsidy involved; if so, this should be removed. Where there is monopoly power in the energy market, either locally or, for example, through the monopolistic behavior of the Organization of Petroleum Exporting Countries, the price of energy is unlikely to reflect the true cost to society of its generation.

Energy supply is considered so essential that a premium may be placed on its supply from domestic sources if this reduces the risk of supply interruption. Most countries have assessed the importance of such supply, and this information may be used to assess any premium that should be applied. If the increase in supply that would be generated by the energy project under consideration is marginal to the total supply in a country, it is unlikely that such a premium would be significant.

In an *ex ante* analysis, a sensitivity analysis should be undertaken for alternative scenarios regarding the possible future changes in energy prices. These are difficult to predict because there are a number of offsetting effects. Possible deflationary pressures include the development of more energy-efficient machinery and technology, improved availability (and therefore reduced prices) of non-fossil fuel energy sources and systems, and any change in the regulation of the energy sector. Possible inflationary pressures include rapid economic growth and a consequent increased demand for energy, the exercising of market power in oil production, future scarcity of fossil fuels, and the introduction of carbon taxes and/or tradable permits in light of the Kyoto Protocol. However, if the price of energy is inflated by a carbon tax/emissions-trading system set to reflect the external costs of energy use, it is important to avoid summing external costs estimated by the techniques described later with the energy savings using this price because if the price already reflects the external effects, it would be inappropriate to include them a second time.

4.7 Externalities

Shadow prices adjust for one form of market failure. However, a CBA must also take into account the potential existence of externalities and public goods, as explained previously. The consideration of such

costs and benefits is extremely important in CBAs because all of the costs and all of the benefits to the whole of society are being considered. The difficulty is that these external costs and benefits are not exchanged within markets, so there is no price to reflect their value. For this reason, it is necessary to use nonmarket or environmental valuation approaches to place monetary values on these externalities.

5. VALUATION OF ENERGY EXTERNALITIES

Most energy projects have some impact on the environment that must be factored into a CBA. For example, a new coal-fired power plant will have implications for climate change, localized air pollution, and the production of acidification precursors. A cost–benefit analyst assessing the merits of a renewable energy project or an energy efficiency program will wish to factor in the value of reductions in such environmental effects.

5.1 Economic Value Concepts

A useful framework for assessing the effects of environmental impacts on humans is known as the total economic value approach. This expands the traditional concepts of economic value to include option value and existence value. In essence, an environmental asset is considered to have both use and nonuse value.

Use value can be broken down into actual use value and option value. Suppose we are estimating the value of damage caused to woodland from acid rain resulting from power generation. Actual use value is derived from the actual use of the asset and is composed of direct use value, such as timber from the woodland, and indirect use value, such as the value that would be yielded to inhabitants downstream should a forest reduce flooding.

Option value can be broken down into a number of components: the value individuals place on the preservation of an asset so that they may have the option of using it in the future (value of future options for the individual) for recreational purposes, for example; the value they attach to preserving the asset because it can be used by others (vicarious value); and the value they attach to preserving the asset so that future generations have the option of using it (bequest value). Thus, total use value is defined as

$$\begin{aligned}\text{Total use value} = {} & \text{direct use value} \\ & + \text{indirect use value} \\ & + \text{option value},\end{aligned}$$

where

$$\begin{aligned}\text{Option value} = {} & \text{value of future options for the} \\ & \quad \text{individual} \\ & + \text{vicarious value} \\ & + \text{bequest value}.\end{aligned}$$

Nonuse values are existence or passive-use values. Existence values reflect the benefit to individuals of the existence of an environmental asset, although the individuals do not actually use the asset. These values may exist due to sympathy for animals and plants and/or some human approximation of the intrinsic value of nature (i.e., this value reflects a sense of stewardship on the part of humans toward the environment). Existence value is classed as a nonuse value such that total economic value (TEV) is defined as

$$\begin{aligned}\text{TEV} = {} & \text{actual use value} + \text{option value} \\ & + \text{existence value}.\end{aligned}$$

It is important to note that the TEV approach attempts to encapsulate all aspects of a resource that enter into human's utility functions (i.e., it is anthropocentric). Thus, it does not attempt to find an intrinsic value for the environment, where intrinsic means the value that resides in something that is unrelated to humans. However, it does attempt to include the intrinsic value that humans bestow on nature. There is no agreement on the exact breakdown of the values in the TEV framework; however, the important point in CBA is to account for all these values as best as possible.

5.2 Environmental Valuation Methods

There are two broad categories of nonmarket valuation methods: revealed preference methods and stated preference methods.

5.2.1 Revealed Preference Methods

Revealed preference studies use data from actual behavior to derive values for environmental assets. Suppose the value of a marketed good depends in part on an environmental good. These methods try to isolate the value of the environmental good from the

overall value of the marketed good. Such techniques include the following:

Hedonic pricing method: The value of a house depends, in part, on its surrounding environment. This method examines the determinants of the house's value and then isolates the effect that the environmental feature (e.g., a nearby park) has on the house. This is used as an estimate of the value of the environmental asset in question.

Travel cost method: The value of the park can be estimated by the willingness to pay to attend the park in terms of costs of travel and entrance to the park.

Production function approaches: These value a decrease (increase) in environmental quality as a result of some activity by the loss of (addition to) the value of production due to that activity. For example, the cost of acid rain resulting from pollution from fossil fuel-based power plants might be measured by the lost output of timber from forests (dose–response functions), the cost of restoring the damage to buildings (replacement cost), or the cost of defensive expenditures.

The production function approaches may seem more applicable to estimating externalities of energy generation, but the hedonic pricing method and travel cost method may be used to estimate the value of an amenity such as a forest park. This value could then be used to calculate the cost of damage to the park from air pollution.

5.2.2 Stated Preference Methods

Direct valuation methods ask people in a survey to place a value on the environmental asset in question by rating (contingent valuation), ranking (contingent ranking), or choosing trade-offs between various policy alternatives (choice procedures). These direct methods via questionnaires are the only approaches that can estimate existence value. The best known and most controversial of these methods is contingent valuation.

The contingent valuation method (CVM) collects information on preferences by asking households how much they are willing to pay for some change in the provision of an environmental good or the minimum compensation they would require if the change were not carried out. Since the mid- to late 1980 s, this method has become the most widely used approach for valuing public goods and there has been an explosion in the number of CV studies.

The CVM approach is to choose a representative sample, choose an interview technique (face-to-face, telephone, or mail), and then construct a questionnaire containing:

- A detailed description of the good being valued and the hypothetical circumstance in which it is made available to the respondent.
- Questions that elicit the respondents' willingness to pay for the good being valued: open ended (How much would you pay?), bidding (Will you pay x, if not, will you pay $x-y$), payment card (Which of these amounts would you choose?), and binary choice (If it costs x, would you pay it? Yes/no).
- Questions about the respondents' characteristics (e.g., age, education, and income), their preferences relevant to the good being valued (e.g., how concerned or interested are they in the environment), and their use of the good.

The mean or median willingness to pay from the sample of households surveyed can be multiplied up to the population level to give a total valuation.

5.3 Valuing Environmental Damage and Improvements from Energy Projects

The most significant environmental effects of energy projects that are included in CBAs tend to be the contribution of such projects to global warming from changes in CO_2 emissions and acid rain resulting from the production of NO_x and SO_2. In addition, many energy projects have implications for the production of particulate matter that may have consequences for human health and visibility.

The impacts of climate change are as follows: storm and flood damage (increased in some regions, such as the Caribbean and Pacific, but not others), increased mortality and morbidity, effects on small islands (sea levels rise due to heating water and melting ice caps), increasing salinity of estuaries, damage to freshwater aquifers, and effects on agriculture and forestry. It is estimated that a substantial portion of the existing forested area of the world will undergo major changes.

There are three approaches to placing a monetary value on the benefits of greenhouse gas reductions. The damage-avoided approach values 1 ton of carbon not emitted by the cost of the damage that would have been done by global warming in the event that it had been emitted. The offset approach measures the value of not emitting 1 ton of carbon using one method compared to the next cheapest alternative method. The avoided cost of compliance

approach measures the 1 ton of saved carbon by the avoided cost of compliance with a global/regional CO_2 emissions reduction agreement. A variety of estimates have been produced regarding the benefits of reducing CO_2, including studies by the United Nations Intergovernmental Panel for Climate Change.

Several major research projects have endeavored to place monetary values on all the environmental effects related to energy. Two EU-funded projects, one on the external costs of power generation (ExternE) and the other on European environmental priorities, are very useful sources of information on such values.

The ExternE project used an "impact pathway" methodology. Under this methodology, the fuel cycle, or life cycle of the fossil fuel used for generation, is first divided into stages, such as extraction, transport, combustion, and waste disposal. The environmental burdens imposed at each stage and the likely impact of these burdens on the environment and the public are isolated. The impacts are then valued using a variety of economic valuation methods, such as contingent valuation, direct cost of damage, or cost of mitigation. These costs are then normalized to give a cost per unit energy output and a value per ton of a pollutant reduced.

The "priorities" project endeavors to derive values for various environmental externalities suitable for use in a European context. The impacts of SO_2 on acute mortality, morbidity, materials, and the fertilization effect are all included; however, some impacts are omitted, such as forest damage, adverse effects on ecosystems and cultural heritage, and the dampening effect that SO_2 has on global warming. The main areas of uncertainty relate to the value of mortality. NO_x emissions are dealt with under two headings, acidification and tropospheric ozone. The fertilization effect of NO_x is examined under acidification. For damage to health and crops, the effects of NO_x on tropospheric ozone are used. Again, the bias is toward underestimation because damage caused to forests, biodiversity, non-crop vegetation, and materials is excluded.

Care should be taken when using unit damage values such as those in the studies mentioned previously. First, the confidence intervals are large (and should always be quoted in a CBA). In addition, there is confusion about which emissions cause which adverse effects. For this reason, it is very easy to double count when using estimates derived from different studies. For example, in some studies negative health effects are included in SO_2 damage estimates, whereas in others the most significant health effects result from particulate matter emissions. Finally, the damage caused by 1 ton of most air pollutants (excluding CO_2) will vary depending on the geographical origin of the emission. However, most estimates are derived from an exercise known as benefits transfer.

5.4 Benefits Transfer

To reduce costs and where there is no alternative procedure, it is standard practice in CBAs to apply existing monetary valuation studies outside of the site context in which they were originally undertaken. This is known as benefits transfer. There are a number of rules when using this method, the most obvious being that the contexts of the studies should be very similar. In addition, appropriate statistical techniques should be used, including meta-analyses of best practice studies; the studies should have similar population characteristics; and property rights should be similarly distributed. The method is most accurate when the impacts are not affected by local characteristics (e.g., income and population size). The most comprehensive studies of energy externalities referred to previously involve significant benefits transfer.

5.5 Valuing Mortality and Morbidity

Placing monetary values on life and death is perhaps the most controversial area of economics. There are two major approaches for carrying out this task. The first values life directly through livelihood. This is known as the human capital method and is the oldest and most easily applied. It equates the value of life of an individual as the present value of future lost output of the individual, which is usually estimated using income. However, it is important to note the substantial drawbacks of this approach. Most obvious, it values livelihood rather than life per se. It is quite improbable that there would be a 1:1 relationship between individuals' earnings and their demand for risk reduction. In other words, the human capital approach makes no reference to how humans value risks to their own lives. One of the most glaring flaws of this approach, however, is the fact that it places a zero value on the retired population and those who do not "produce" marketed output (e.g., those who work in the home).

The appropriate way of valuing reductions in mortality is not to value a life per se but rather to value people's willingness to pay for a reduction in

the risk of mortality. These data derive primarily from one of three study formats: the wage risk model (also known as the hedonic wage model), the avertive behavior method, and the contingent valuation method. Although space does not permit a detailed explanation of these methods, the general approach is to value a small reduction in risk and then translate this into the equivalent of a life saved. This is the value of a statistical life (VSL).

There has been controversy regarding this approach in relation to the estimation of the costs of power generation. The VSL estimates would be expected to be considerably lower for a developing country because the willingness of people to pay to reduce risk of death is partially dependent on income. Thus, the VSL used for a developing country (e.g., for deaths from coal mining) would be expected to be lower than the VSL used for a developed area (e.g., for deaths from air emissions in Europe by the burning of such coal). It is notable that the ExternE project steered clear of controversy by using the same VSL for each area.

The value of morbidity may be assessed using a cost of illness approach. In the case of air pollution from energy generation, this would be the sum of the cost, both public and private, of hospital stays and drugs for the treatment of associated illnesses. More positively, the health benefits of improving domestic energy efficiency would be the associated avoided costs of illnesses. Other costs avoided by wider society include the loss of productivity from those who are ill and unable to work. The avoided cost to the individual beyond drugs and hospitalization direct costs (if any) is more difficult to calculate. One approach is to estimate individuals' willingness to pay to avoid "restricted-activity days." The priorities study referred to previously contains such estimates.

6. FURTHER ISSUES IN THE APPLICATION OF COST–BENEFIT ANALYSIS TO ENERGY

When undertaking or evaluating a CBA of an energy project, there are a number of additional issues that must be considered. The tool is not without its controversies. These arise because of its use of money as the common unit of measurement. First, assessing welfare changes using willingness to pay and/or willingness to accept is perfectly valid in theory if one believes that income is distributed optimally. However, care must be taken when the population(s) whose welfare is being assessed has a high income variance. Poor people will be willing to pay less, all else being equal, simply because of their budget constraint. In addition, because the values and market prices used in a CBA of an energy project or policy are in part a product of the existing income distribution, if that distribution changes, prices will change, and so will market and nonmarket values. Second, distributional effects tend to be ignored in CBAs and yet energy policies often have very different consequences for different socioeconomic groups. Because energy is a necessity, energy efficiency schemes may benefit those on lower incomes more than average. However, without poverty-proofing measures, energy taxes will comprise a much higher proportion of their income than average. Third, not all support attempts to place monetary values on environmental assets, and the value of the statistical life concept is particularly controversial. Some find it unpalatable to consider that the environment might be traded with other goods and services let alone "lives." In addition, as in all public policy or human actions, the environmental values used are by definition anthropocentric.

Finally, the environmental valuation methods are limited in their ability to assess values. Many environmental outcomes assessed by natural scientists are subject to large confidence intervals even before economists endeavor to value them. Also, the valuation methods vary in their ability to assess environmental values. Economists have identified a number of criteria by which environmental valuation methods should be evaluated: validity (the degree to which it measures the total economic value of a change in the environment), reliability (the extent to which the variability in nonmarket values is due to random sources), applicability (assesses where the techniques are best applied), and cost (the practicality and viability of carrying out the valuation study). In general, revealed preference methods (e.g., productivity loss and replacement cost) are widely accepted valuation methods that provide valid and often reliable estimates of use values. However, they are limited in that they only provide estimates of use values, and it is often difficult to find suitable and reliable links between market goods and environmental amenities. The estimated values are often sensitive with respect to modeling assumptions, reducing their reliability in such cases. In general, however, revealed preference methods are applicable for valuation of policies that primarily affect use values and can provide very reliable results. Stated

preference methods such as contingent valuation are capable of measuring both use and nonuse (existence/passive-use) values. They are applicable to a larger array of environmental issues than are revealed preference methods, and they are often the only applicable methods for benefit estimation. However, they do have a number of limitations and may be subject to a number of biases.

There are some simple rules for using cost–benefit techniques. Simpler environmental valuation methods should be used where possible (e.g., production function approaches) while recognizing these are primarily of interest when the policy changes will mainly influence use values. A high degree of reliability in stated preference methods may be achieved by adhering to sound principles for survey and questionnaire design and if the method is applied to suitable environmental goods. The methods work best for familiar environmental goods and for salient environmental issues. The methods are likely to be least reliable where the environmental good being valued is complex (e.g., in the case of biodiversity).

Finally, in CBAs of energy initiatives, it is important to note that the estimates of values are made at the margin. The plausibility of the techniques described depends in part on the fact that most things remain the same, with the implicit assumption that there is no paradigm shift in values, power, income, and the like during the time period of assessment.

SEE ALSO THE FOLLOWING ARTICLES

Decomposition Analysis Applied to Energy • *Ecological Risk Assessment Applied to Energy Development* • *Energy Development on Public Land in the United States* • *Environmental Change and Energy* • *Equity and Distribution in Energy Policy* • *External Costs of Energy* • *Green Accounting and Energy* • *Value Theory and Energy*

Further Reading

Clinch, J. P., and Healy, J. D. (2001). Cost–benefit analysis of domestic energy efficiency. *Energy Policy* **29**(2), 113–124.

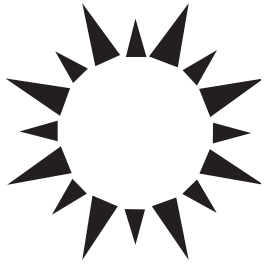
Commission of the European Communities—DG Environment (2003). European environmental priorities Web site: An integrated economic and environmental assessment. Available at europa.eu.int/comm/environment. ExternE Web site: http://externe.jrc.es.

Hanley, N., and Spash, C. (1993). "Cost–Benefit Analysis and the Environment." Elgar, Aldershot, UK.

Layard, R., and Glaister, S. (eds.). "Cost–Benefit Analysis," 2nd ed. Cambridge Univ. Press, Cambridge, UK.

Zerbe, R., and Dively, D. (1994). "Benefit Cost Analysis in Theory and Practice." HarperCollins, New York.

Crude Oil Releases to the Environment: Natural Fate and Remediation Options



ROGER C. PRINCE
ExxonMobil Research and Engineering Company
Annandale, New Jersey, United States

RICHARD R. LESSARD
ExxonMobil Research and Engineering Company
Fairfax, Virginia, United States

1. **Introduction**
2. **Composition of Crude Oils and Refined Products**
3. **Releases to the Environment**
4. **Natural Fate of Oil in the Environment**
5. **Responses to Spills**

Glossary

API (American Petroleum Institute) gravity A scale describing the density of an oil. Oils with gravities $>40^\circ$ are known as light oils, and those with gravities $<17^\circ$ are said to be heavy.

asphaltenes Large complex molecules in crude oils and heavy fuels, often with molecular weights in the many thousands, composed of carbon, hydrogen, oxygen, and other heteroatoms. They are responsible for most of the color of crude and heavy oils and typically are not "oily" to the touch.

bioremediation Stimulating the natural process of oil biodegradation. To date, this has been restricted to the addition of fertilizers to alleviate nutrient limitations for these organisms. Adding bacteria has not been shown to be necessary.

booms Floating devices for corralling, or sometimes absorbing, oil on water. There are many different designs for different situations, including stationary booms in rivers, towed booms for collecting oil, and fireproof booms for holding oil for combustion.

Bunker C Fuel oil used to power ships' engines. It is a very heavy, viscous fuel derived from the heaviest hydrocarbon compounds that remain after processing whole crude oil.

crude oil Natural product that is the remains of biomass that was typically alive 100 million years ago. It is a complex mixture of hydrocarbons, resins, and asphaltenes.

dispersants Concentrated solutions of surfactants that can be sprayed on floating oil to cause it to disperse into the water column, where it is naturally biodegraded.

hydrocarbons Compounds composed of only carbon and hydrocarbon; these are the most abundant molecules in most crude oils and refined fuels. They are a diverse mixture of linear, branched, and cyclic saturated compounds with 1–80 carbon atoms, together with aromatic compounds, most commonly with more than one ring and with multiple alkyl substituents.

natural attenuation The natural processes that eventually remove hydrocarbons from the atmosphere.

resins Moderately large complex molecules in crude oils and heavy fuels, often with molecular weights in the thousands, composed of carbon, hydrogen, oxygen, and other heteroatoms.

skimmers Mechanical devices to remove floating oil from water. There are many designs, including moving belts, weirs, mops, and vacuum pumps.

Crude oils and other hydrocarbons have been a part of the biosphere for millennia, so it is no surprise that a ubiquitous, diverse population of hydrocarbon-consuming microorganisms are able to exploit this resource. These organisms can be recruited to help respond to human-associated oil spills, although sometimes only after physical treatments to remove bulk oil, such as mechanical pick-up or combustion:

- If oil is floating at sea, the addition of dispersants to dramatically increase the surface area available for microbial attack will stimulate biodegradation

 © 2004 Elsevier Inc. All rights reserved.

and also inhibit the formation of troublesome tar balls and minimize the amount of oil that reaches the shore.
- If oil is on a shoreline, the addition of fertilizers to provide necessary ancillary nutrients for the degrading microorganisms will stimulate oil biodegradation.
- If oil is in surface soil, adding fertilizers and tilling the soil will provide both the nutrients and the air needed to stimulate biodegradation.
- If soluble hydrocarbons are in an underground water plume, adding a source of oxygen, whether as a gas or a soluble or slow-release peroxides, will stimulate aerobic biodegradation.

Together, these approaches exemplify modern environmentally responsible remediation activities: working with natural processes to stimulate them without causing any additional adverse environmental impact.

1. INTRODUCTION

Crude oils have been a part of the biosphere for millennia, and human use extends back thousands of years. Early uses included hafting stone axes to handles, use as an embalming agent, and use as a medical nostrum. In the Bible, Genesis implies that bitumen was used as the mortar for the Tower of Babel. It also seems likely that several religions started near natural gas seeps, either as eternal flames or as sources of hallucinogenic vapors. However, these were very minor uses, and it is only in the past century and a half that oil has come to play a truly central role in our lives. Terrestrial seeps were the first locations to be drilled when oil production began in earnest in the 19th century, such as the 1859 Drake well in Pennsylvania. Today, oil is recovered from all over the world, except the Antarctic, and from great depths, including coastal margins such as offshore West Africa and South America. The scale of oil production is staggering; global use is on the order of 10^{12} U.S. gallons per year (3.25×10^9 tonnes or 3.8×10^{12} liters/year), and much of it is transported thousands of miles before it is used. For example, the United States imported 350,000 tonnes of oil per day from the Middle East alone in 1999. Unfortunately, despite the best efforts of the oil industry, some (a relatively small fraction) of this material has been spilled at production and refinery facilities and from pipelines and vessels. Such spills generate enormous public pressure for speedy clean-up, but the obvious approach of simply sucking up the oil from the environment is rarely simple or complete. This article addresses the natural fate of oil in the environment and how understanding the underlying processes allows intervention to safely stimulate them and thereby minimize the environmental impact of spills.

2. COMPOSITION OF CRUDE OILS AND REFINED PRODUCTS

2.1 Origins

Crude oils are natural products, the result of the burial and alteration of biomass from the distant past. The alteration processes are known as diagenesis and catagenesis. The initial process of diagenesis typically occurs at temperatures at which microbes partially degrade the biomass and results in dehydration, condensation, cyclization, and polymerization of the biomass. Subsequent burial under more sediments, and thus at higher temperatures and pressures, allows catagenesis to complete the transformation of the biomass to fossil fuel by thermal cracking and decarboxylation.

It is generally accepted that most petroleum reserves are derived from aquatic algae, albeit sometimes with some terrestrial material, whereas the great coal reserves of the world arose from deposits of higher plants, typically in nonmarine environments. The generation of oil and gas invariably occurs at significant depth and under great pressure, which forcibly expels the oil and gas from the initial source rock. Commercially valuable reservoirs are formed if the migrating oil becomes trapped in a geological structure so that the oil and gas can accumulate. Most petroleum reservoirs are found in sandstones, siltstones, and carbonates with porosities of 5–30%. The traps are never completely full of oil, and there is always some water, usually containing substantial amounts of inorganic salts (brine), underlying the oil. Typical commercial oils originated from biomass that was produced on the order of 100 million years ago, although the oldest commercially valuable oils are from Ordovician biomass (486 million years ago), whereas others are as young as the late Tertiary (a few million years ago).

2.2 Hydrocarbons

Crude oils and refined products are very complex mixtures of molecules, both hydrocarbons and

compounds containing other elements, such as sulfur, oxygen, and nitrogen. The hydrocarbons, containing only hydrogen and carbon, are the best studied and understood, and they can be divided into broad classes based on their overall chemical properties. The hydrogen to carbon ratio of crude oils is between 1.5 and 2.0; the organic molecules are thus principally saturated molecules (i.e., the predominant form of carbon is $-CH_2-$). The convention in the oil industry is to call linear alkanes paraffins and cyclic alkanes naphthenes. There are also significant amounts of aromatic carbon in all crude oils and polar molecules containing the heteroatoms oxygen, nitrogen, and sulfur. These latter molecules can be fractionated by column chromatography, and the fractions have a variety of names, including resins and asphaltenes. In their classic 1984 book "Petroleum Formation and Occurrence," Tissot and Welte state that the average composition of 527 crude oils is 58.2% saturates, 28.6% aromatics, and 14.2% polar compounds, although the absolute values vary widely in different oils. On average, there is rough parity between paraffins, naphthenes, and aromatics.

The paraffins include the linear alkanes from methane to waxes with up to 80 carbons. Linear alkanes typically make up 10–20% of a crude oil, although their content can range from essentially undetectable to as high as 35% depending on source and reservoir conditions. There are also branched alkanes, usually most abundant in the C_6–C_8 range. Two branched alkanes found in almost all crude oils, pristane ($C_{19}H_{40}$) and phytane ($C_{20}H_{42}$), are molecular relics of the phytol chains of chlorophylls and perhaps other biomolecules in the original biomass that gave rise to the oil.

The naphthenes include simple compounds such as cyclopentane, cyclohexane, and decalin, together with their alkylated congeners. Tissot and Welte state that the average composition of the naphthene fraction of 299 crude oils is 54.9% one- and two-ring naphthenes, 20.4% tricyclic naphthenes, and 24.0% tetra- and pentacyclic naphthenes. These latter molecules are molecular relics of membrane components of the original biomass, and they are very useful fingerprints of individual oils.

The aromatic molecules in crude oils range from benzene to three- and four-ring structures, with traces of even larger ring systems. Because of the separation procedures used in the characterization of crude oils, any molecule containing at least one aromatic ring is included in the "aromatic" fraction, regardless of the presence of saturated rings and alkyl substituents. Crude oils typically contain aromatic molecules with up to four aromatic rings; one series contains just six-membered rings and their alkylated derivatives—benzene (one ring), naphthalene (two rings), phenanthrene (three rings), chrysene (four rings), etc. Another series includes one five-membered ring in addition to the six-membered ones—fluorene (three rings), fluoranthene (four rings), etc. Because of the operational definition based on chromatography, sulfur aromatic heterocycles, such as thiophenes, benzothiophenes, and dibenzothiophenes, are included in the aromatic category. Indoles and carbazoles, usually the most abundant nitrogen-containing species, and the less abundant basic nitrogen species such as quinolines are also included in the aromatic category. Alkylated aromatic species are usually more abundant than their parent compounds, with mono-, di-, and trimethyl derivatives usually being most abundant. Nevertheless, the median aromatic structure probably has one or two methyl substituents together with a long-chain alkyl substituent.

2.3 Nonhydrocarbons

The polar molecules of crude oils are the most difficult to characterize because they are often unamenable to gas chromatography, the usual method of choice for the molecular characterization of petroleum. All are thought to contain heteroatoms, such as nitrogen, oxygen, and/or sulfur, and the category includes the porphyrins (usually nickel or vanadium species) and naphthenic acids. Some of these molecules have molecular weights in the thousands or higher, and many are suspended in the oil rather than dissolved in it. The polar fraction of the oil contains the majority of the color centers in crude oil, and in isolation these materials are difficult to distinguish from recent biological residues, such as humic and fulvic acids.

2.4 Oil Classification

Oils are classified by several criteria, but among the most important is the specific gravity. The oil industry uses a unit known as API (American Petroleum Institute) gravity, which is defined as [141.5/(specific gravity)]−131.5 and expressed in degrees. Thus, water has an API gravity of 10°, and denser fluids have lower API gravities. Less dense fluids (e.g., most hydrocarbons) have API gravities $>10°$. For convenience, oils with API gravities $>40°$ are said to be light oils, whereas those with API gravities $<17°$ are said to be heavy (Table I). Light

TABLE I

Conversion Table for a Medium Crude Oil

1 tonne	308 U.S. gallons
	7.33 barrels
	858 liters

oils have higher proportions of small molecules; heavy oils are rich in larger molecules. Viscosity is inversely proportional to API gravity, but it is also dependent on the physical state of the polar compounds and longer alkanes in the oil, and it is highly dependent on temperature.

2.5 Refined Products

Crude oil is transported in the largest volumes, both in pipelines and in tankers, but refined products are also transported long distances, particularly in pipelines. Refining starts with distillation, and the simplest distinction of the various refined products can be related to this process. The most volatile liquid product is aviation gasoline, followed by automobile gasoline, jet fuels, diesel and heating oils, and the heavy oils that are used for fueling ships and some electrical generation. All ships contain large volumes of heavy fuel oil, often called Bunker C, that is barely liquid at ambient temperatures and must be kept warm to be pumped into engines.

Most of the molecules in gasoline have 4–10 carbons (Table II), most in diesel fuel have 9–20, and heavy fuel oils typically have very few molecules with less than 15 carbon atoms except for some added as a diluent to achieve the appropriate viscosity. It is important to recognize that fuels are not sold by their composition but by their properties, such as octane, cetane, or viscosity, and there are many chemical mixtures that meet these requirements. The chemical composition of fuels with the same name can thus be very different, even though all meet their specifications. The lightest products (i.e., those with the lowest boiling points, such as gasolines and diesels) are almost entirely hydrocarbons, whereas the heavy fuel oils are enriched in the polar constituents such as asphaltenes; this is reflected in the color of the products.

TABLE II

Number of Carbon Atoms in the Majority of Compounds in Refined Products

Product	Approximate carbon range
Aviation gasoline	C_4–C_8
Regular gasoline	C_4–C_{10}
JP4 jet fuel	C_5–C_{11}
JP8 jet fuel	C_8–C_{15}
Kerosene	C_9–C_{16}
Diesel No. 1	C_9–C_{17}
Diesel No. 2	C_9–C_{20}
Bunker C	C_8–C_{10}, C_{15}–C_{50}
Hydraulic oil	C_{18}–C_{36}
Automatic transmission fluid	C_{19}–C_{36}

TABLE III

National Research Council's Estimates of Annual Input of Oil to the Sea

Source	North America, tonnes	World, tonnes
Seeps	160,000	600,000
Consumers	90,000	480,000
Spills	12,100	160,000
Extraction	3,000	38,000
Total	260,000	1,300,000

3. RELEASES TO THE ENVIRONMENT

Crude oil has been released to the natural environment from seeps for millennia and, as mentioned previously, these have proved useful sources for humankind. Some seeps are relatively small, and the oil affects only a very small area. This was true for the seeps that gave rise to "perpetual flames" in Iraq and bitumen reserves in eastern North America. However, some are vast and have given rise to huge asphalt lakes and tarpits. The La Brea tarpits in Los Angeles are well-known, not least for the impressive bones from tigers that unwittingly became entrapped, and the large asphalt lakes in Trinidad (140 acres) were once exploited to deliver asphalt for paving in Newark, New Jersey. Clearly, a major determinant of the environmental impact is the scale of the release. Table III shows 2002 estimates of the scale of oil releases into the sea both in the United States and throughout the world.

3.1 Natural Seeps

Natural seeps are found all over the world, and the scale of their releases is quite large. In 2002, the National Research Council estimated that the total annual input of petroleum into the sea from all

sources is approximately 1.3 million tonnes, with almost 50% coming from natural seeps. For example, natural seeps in the Gulf of Mexico release an estimated 140,000 tonnes of oil per year into the sea, those offshore Southern California release 20,000 tonnes per year, and those offshore Alaska release 40,000 tonnes.

3.2 Municipal Runoff

The National Research Council estimates that releases due to the consumption of crude oil, principally municipal runoff but also including spills from nontanker ships, recreational uses, and atmospheric deposition, are the second largest input of oil into the world's oceans, far dwarfing that from catastrophic tanker spills. Again, the sheer scale of oil use is staggering; it is estimated that just the input of lubricating oil from two-stroke engine use in the U.S. coastal seas is 5300 tonnes per year.

3.3 Tanker Spills

Almost half of all oil moves by seagoing tanker at some stage between production and use, and spills from these vessels attract by far the largest amount of attention from the media. As discussed previously, their total input is relatively small compared to other inputs, and a recent estimate is that more than 99.98% of the oil in transit reaches its destination safely. The issue, of course, is that when they do occur, tanker spills are very localized, and they can have substantial localized environmental impacts that call for a cleanup response. The industry has responded to public concerns, and tanker spills are fortunately becoming rarer; for the period 1990–1999, spillage from vessels in U.S. waters was less than one-third that released in 1980–1989. Similar improvements have occurred worldwide, although there is always need for improvement.

3.4 Pipeline Spills

The Minerals Management Service estimates that there are 21,000 miles of undersea oil pipelines installed on the U.S. continental shelf delivering oil from wellheads to collection points and refineries. Spills are rare but are estimated to be approximately 1,700 tonnes per year, mainly in the Gulf of Mexico. To put this in perspective, the largest spill in the Gulf occurred in 1979, when the IXTOC-1 well "blew out" approximately 80 km off Ciudad del Carmen on the Mexican coast. The blowout lasted 9 months and spilled 476,000 tonnes of crude oil before it could be controlled.

Total U.S. liquid petroleum pipelines include a total of 114,000 miles of crude oil pipelines and 86,500 miles of pipeline for refined products, which together transport two-thirds of all U.S. petroleum. Pipeline leaks, often caused by excavation errors, lead to annual spills of approximately 19,500 tonnes, and again there has been a substantial improvement in safety during the past few decades. Because of the physical barriers around most pipeline spills, they are usually contained, and most spilled oil is collected with pumps. Unfortunately, this is not always the case; in 1997, an underwater pipeline released 940 tonnes of south Louisiana crude oil into Lake Barre in Louisiana. The spill affected 1750 ha with a light sheen, but only 0.1 ha of marsh died back. Much of the oil evaporated, some pooled oil at the edges of some marshes was recovered, and the lightly oiled areas appeared to be clean within 2 months.

3.5 Wars, Environmental Terrorism, and Vandalism

The past decade has seen a resurgence of oil spills associated with warfare, terrorism, and vandalism. Many ships were sunk in World War II, and some still contain substantial amounts of oil; an example is the USS Mississinewa, which was sunk in Ulithi Lagoon in Micronesia in 1944 and recently located. It is estimated to contain 3300–10,000 tonnes of heavy fuel oil, and this is beginning to leak from the decaying vessel. By far the largest oils spills in history were those deliberately released by Iraqi forces retreating from Kuwait in 1991. At least 10 million barrels of oil were deliberately poured into the Arabian Gulf, and 700 oil wells were ignited or otherwise destroyed. Vast lakes of oil covering approximately 49 km^2 were formed, and most remain after a decade. The volumes are so huge that the totals remain unknown.

Smaller wars also cause the release of substantial volumes of oil, although again the totals are difficult to estimate. Colombia's 780-km Cano Limon pipeline, which ships 110,000 barrels of oil per day, was bombed at least 170 times in 2001, and estimates of total oil spilled are far higher than those of well-known tanker spills. The French supertanker Limberg was apparently attacked by terrorists offshore Yemen in 2002. Even the Trans Alaska Pipeline is not immune to vandalism; 925 tonnes were spilled when a high-velocity rifle slug penetrated the pipeline in 2001.

4. NATURAL FATE OF OIL IN THE ENVIRONMENT

The description of the various inputs of oil into our environment, mainly from natural seeps and municipal runoff, makes it clear that oil has been part of the biosphere for millennia. However, we are not overwhelmed by it. This is because a variety of natural processes contribute to its removal, the most important of which is biodegradation. Crude oils and refined products provide an excellent source of carbon and energy to organisms, principally microorganisms, able to exploit it. These organisms are responsible for the consumption of most of the hydrocarbons that get into the biosphere and that are not collected or burnt. Nevertheless, this process is relatively slow, and several physical processes typically act on spilled oil before it is eventually biodegraded.

4.1 Flotation and Spreading

Almost all oils in commerce float; thus, they spread rapidly if the spill is on water. This has the effect of providing a dramatically larger surface area for evaporation and dissolution, and eventually the floating layer becomes subject to physical disruption by waves. A few very heavy oils, such as that spilled in the Delaware River from the Presidente Rivera in 1989, are sufficiently dense that they form almost solid lumps that submerge in fresh water. They do not usually sink, however, and can be collected by subsurface nets.

4.2 Evaporation

Small hydrocarbons are very volatile, and molecules smaller than those with approximately 15 carbons typically evaporate quite readily if a spill is on the surface. This can be effectively complete in a matter of days, although underground spills do not typically evaporate very rapidly. Spills are not the only source of volatile hydrocarbons in the atmosphere; trees liberate vast quantities of isoprene and other terpenes. Whatever the source, some interact with reactive nitrogen oxides in the presence of light and in high enough concentrations can contribute to photochemical smogs, The Blue Mountains got their name from this natural phenomenon long before the onset of the petroleum era. Eventually, most volatile compounds, and the reaction products of photochemical oxidations, are "rained out" and deposited on the ground, where they are biodegraded. Some biodegradation may even take place on water droplets in the atmosphere because there are many bacteria in the air.

4.3 Dissolution

A few hydrocarbons are sufficiently soluble in water that they "wash out" of floating surface or underground spills into the water phase. This phenomenon is usually only significant for the BTEX compounds (benzene, toluene, ethylbenzene, and the xylenes) and oxygenated compounds such as methyl-tertiarybutylether. Such compounds have the tendency to wash out from floating spills, whether on surface water or on the water table underground. This phenomenon is of particular concern underground because the soluble compounds have the potential to migrate and contaminate wells and surface water. Fortunately, they are biodegradable under both aerobic and anaerobic conditions.

4.4 Dispersion

Although oils do not dissolve in water, they can disperse. Wave action breaks up floating slicks and may disperse the oil so finely that it is readily biodegraded. This is what happened to most of the 85,000 tonnes of oil lost in the 1993 spill off the Shetland Islands from the Braer. Adding chemical dispersants to encourage this process is an important oil-spill response tool. If the oil is merely broken into large globs, however, these can coalesce and form tar balls that can drift for thousands of miles before eventually beaching. It is not unusual for tar balls to accumulate other debris during their travels or even to be colonized by barnacles.

4.5 Emulsification

Many oils will form very stable water-in-oil emulsions, particularly after they have weathered sufficiently to form a critical concentration of asphaltenes that can stabilize such emulsions. This can multiply the mass of the original oil by up to five times, increasing the recovery and disposal burden. Water-in-oil emulsions can achieve extremely high viscosities, making relatively light oils behave more like heavy fuel oils. This can slow the evaporative and dissolution processes so that if the emulsions are broken, the released oil's properties are more similar to those of the original material than if emulsification had not occurred.

4.6 Physical Interactions with Sediments

The formation of long-lived "pavements" is a poorly understood phenomenon that can have a significant effect on the persistence of oil in the environment. At high enough concentrations, crude oils and heavy refined products such as bunker fuels interact with sand and gravel so that the oil becomes completely saturated with sediment and vice versa. Even quite liquid oils and fine sediments can form pavements that mimic commercial asphalt, and when this occurs the inner oil is essentially entombed and protected from biodegradation. It can then last for years without significant weathering, as happened in Tierra del Fuego following the 1974 spill of 50,000 tonnes of oil and fuel from the Metula. Spill response efforts aim to minimize and prevent the formation of these pavements.

At the microscale, however, the interaction of oil with fine sediments is proving to be an important natural process that removes oil from oiled shorelines. Fine particles become associated with droplets of oil and effectively disperse into the water column, where biodegradation occurs quite rapidly. The scale and importance of this phenomenon have only been recognized during approximately the past decade.

4.7 Photooxidation

Photooxidation by sunlight is very effective at oxidizing three-, four-, and five-ring aromatics in floating slicks and stranded oil. The molecules are not completely destroyed, but they seem to be polymerized and incorporated into the resin and polar fractions of the oil, where they are probably no longer bioavailable. Photooxidation of the surface of pavements and tar balls may contribute to their longevity. Photooxidation seems to have far less of an effect on the saturated components of oil.

Photooxidation is an important phenomenon not because it removes large volumes of oil from the environment but because it removes the molecules of most toxicological concern. Crude oils and heavy fuels contain trace amounts of potentially carcinogenic polycyclic aromatic hydrocarbons such as benzo[*a*]pyrene. Fortunately, these are among the most susceptible to photooxidation, and they are readily destroyed if an oil is exposed to sunlight.

4.8 Combustion

The natural ignition of gaseous seeps is still a source of wonder whenever it occurs. All hydrocarbons are combustible, and sometimes shipwrecks catch on fire (e.g., the 144,000-tonne spill from the Haven near Genoa in 1991) or are deliberately ignited (e.g., the 1360 tonnes of fuel oil on board the New Carissa offshore Oregon in 1999). A major problem is that the oil has to maintain a high temperature to continue burning, and heat transfer to water underneath a slick usually extinguishes any fire as the slick thins. Burning spilled oil thus usually requires containment within fireproof booms, but if successful it can be very effective; combustion of $>90\%$ has been reported in several trials. Burning may be particularly useful in ice-infested waters, in which skimming is very difficult and the ice can act as an effective boom. The residue, although viscous, is not very different from oil that weathers without heat, and there is only a small increase in pyrogenic compounds in the residue. Burning oil, however, can generate a large amount of black smoke. The smoke is very low in residual hydrocarbon, but may, in certain circumstances, present a potential health concern due to fine particulate matter. If burning occurs away from population centers, the potential for environmental and health impacts of the smoke is minimal.

4.9 Biodegradation

4.1.1 Aerobic

Biodegradation is the ultimate fate of hydrocarbons released into the environment, although it is sometimes a slow process. Biodegradation under aerobic conditions has been studied for at least a century, and the variety of organisms able to consume hydrocarbons and the diverse pathways and enzymes used in the process are well understood. More than 60 genera of bacteria and 95 genera of fungi have been described that catalyze this process, along with a few species of archaea and algae. These organisms share the ability to insert one or two of the oxygen atoms of diatomic O_2 into a hydrocarbon molecule, thereby activating it and making it accessible to the central metabolism of the organism. The oxygenated hydrocarbon then serves as a source of reductant for the growth of the organism, and much more oxygen is required as the terminal oxidant.

Almost all hydrocarbons that have been studied have been shown to be biodegraded in the presence of oxygen. The small aromatics (benzene and the substituted benzenes) and alkanes are most readily consumed, followed by the larger alkanes and the two- and three-ring aromatics and their alkylated

derivatives. The branched alkanes are slightly more resistant, but they are still readily degraded. The biodegradation of alkylated three- and four-ring aromatics is slower, and some of the four- and five-ring saturated compounds, the hopanes and steranes, are among the most resistant. Indeed, they are so resistant that they can serve as conserved internal markers in the oil, and their concentration in the residual oil increases as biodegradation proceeds. They remain as "fingerprints" of the initial composition of the oil until the very last stages of biodegradation.

Estimates for the total biodegradability of the hydrocarbon fraction of different crude oils range from 70 to 97%. Less is known about the biodegradability of the resins, asphaltenes, and other polar species in crude oils, and the larger ones are not thought to be very susceptible to the process. Fortunately, these compounds lack "oiliness," are friable in small quantities, and are essentially biologically inert. Indeed, they are essentially indistinguishable from other relatively inert biomass in the environment, such as the humic and fulvic acids that make up the organic fraction of soils and sediments and are so essential for plant growth.

Aerobic oil-degrading microorganisms are ubiquitous, having been found in all natural environments in which they have been diligently pursued. Indeed, aerobic hydrocarbon-degrading organisms are the foundations of the extensive ecosystems that exploit marine oil seeps, which often include large multicellular organisms that feed on the hydrocarbon-oxidizing microbes. Oil seeps may even serve as the primary food sources for some fisheries, such as those of Atlantic Canada.

4.1.2 Anaerobic

For some time, it has been known that hydrocarbons are also biodegraded under anaerobic conditions. The hydrocarbon still acts as a reductant, and in the absence of oxygen something else must serve as a terminal electron acceptor. Sulfate (reduced to sulfide), nitrate (reduced to nitrogen), chlorate (reduced to chloride), ferric and manganic ions (reduced to ferrous and manganous ions), and carbon dioxide (reduced to methane) have all been shown to be effective oxidants for hydrocarbon degradation, at least in mixed cultures of organisms. Only a few genera of anaerobic bacteria that can catalyze these reactions have been identified, but this undoubtedly is just the tip of the iceberg and probably more an indication of the difficulty of isolating these organisms than of their scarcity in anaerobic environments. In the absence of oxygen, something else must be added to the hydrocarbon to initiate its metabolism, and at least two pathways have been identified. One adds a molecule of fumarate to the hydrocarbon, whereas the other seems to add a molecule of carbon dioxide. The anaerobic biodegradation of small hydrocarbons, especially the BTEXs, which are of most concern if they leak from storage tanks, is becoming quite well understood.

An important feature of crude oils and refined products is that although they are a rich source of carbon and energy for microbes able to consume them, they are almost uniquely unbalanced foods. Microbes typically have a carbon:nitrogen:phosphorus ratio of approximately 100:10:1, but oil contains no significant amounts of either nitrogen or phosphorus. Oil-degrading microbes must therefore obtain these elements in useful forms from elsewhere in the environment, and if there is a significant amount of hydrocarbon it is likely that the growth of the degrading organisms will be limited by the supply of these nutrients.

5. RESPONSES TO SPILLS

The first response to a spill is to stop any further release and to contain and pick up as much as possible. On land, this is relatively simple, albeit time-consuming and expensive. Indeed, most of the "oil lakes" left in Kuwait from the Gulf War are still awaiting treatment. At sea and on lakes and rivers, however, any collection requires corralling the oil with booms to stop it from spreading.

5.1 Booms, Skimming, and Combustion

Many different types of boom are available; some are inflatable, and others are rigid. Some are designed to withstand ocean waves, and others are designed to work in sheltered harbors. Often, booms are used to keep oil out of particularly sensitive areas, such as fish hatcheries and marshes. Others are designed to be towed to "sweep up" oil, and still others are fire resistant and designed to allow the oil within the boom to be ignited.

Once the oil has been corralled, it has to be collected. This is done with skimmers, which remove oil while accumulating as little water as possible. There are many different designs; mops, moving belts, and floating suction devices are all used. Each is best suited for a particular type of oil and may be

ineffective against other oils, but all attempt to collect the oil without collecting much water or debris.

Igniting oil in fireproof booms is not a trivial matter, and special igniters are usually required.

5.2 Physical Removal

If oil reaches a sandy shoreline, it can often be collected with scrapers and bulldozers. This was very effective following the 1996 spill of 72,000 tonnes of crude oil from the Sea Empress off the south coast of Wales. Sometimes, manual removal by teams of workers with shovels is appropriate, as in the case of the 2002 spill from the Prestige off the coast of Spain. If the shoreline is inappropriate for heavy equipment—such as the shorelines of Prince William Sound, Alaska, affected by the spill from the Exxon Valdez—the oil can be flushed back into the water, corralled within booms, and collected with skimmers. Recent developments include special surfactant products that help lift oil from the shoreline while allowing its retention within booms for subsequent skimming, thus allowing more oil to be released with cooler flushing. In the United Kingdom, it is permissible to use dispersants for this purpose in certain circumstances, provided that they have passed the required approval protocols. In North America, use of dispersants on shores is not permitted, and so a new generation of products have been developed that loosen the oils for removal by water stream without the unwanted effect of oil being dispersed into the near-shore water column. These products are specially formulated with amounts of surfactants sufficient to wet the surface but insufficient to foster dispersion. Some of these products have been successfully used on spills in Canada and the United States. Their main objective is to allow weathered oils to be removed by washing without having to use heated water, which could have detrimental side effects on biological communities.

5.3 Bioremediation

Although the physical removal of small near-shore spills can usually be carried out quite effectively and completely, this is rarely possible with large spills at sea. Even when physical collection or combustion are fairly effective, there is almost always some oil that escapes capture. As discussed previously, the ultimate fate of this residual oil is biodegradation, and there are several environmentally responsible ways of stimulating this process. Colloquially known as bioremediation, these can be divided into two complementary strategies. While oil is afloat at sea, the surface area available for microbial attack limits its biodegradation. Adding dispersants to disperse the oil into the water, and dramatically increase the surface area for microbial colonization, is an important option. In contrast, the biodegradation of oil remaining on a shoreline is likely limited by the supply of other necessary microbial nutrients, such as biologically available nitrogen and phosphorus, and carefully adding fertilizers is a very useful option.

5.3.1 Dispersants

Dispersants are concentrated solutions of surfactants that can be applied from the air or from ships equipped with spray booms or fire hoses. Typical application rates are approximately 5% of the oil volume, so quite large stockpiles are required, and indeed are available, in central depots. The dispersants need to be added fairly soon after a spill, but there is usually a window of several days during which they are effective. Even mild wave energy disperses the oil very effectively into the water column, and it has been demonstrated that this substantially increases the rate of biodegradation. Some dispersants also include molecules that actively stimulate oil biodegradation either by providing limiting nutrients or by adding preferred substrates. Dispersion also minimizes the beaching of spilled oils. Dispersants were used very effectively following the 1996 spill of 72,000 tonnes of a relatively light crude oil from the Sea Empress off the south coast of Wales.

5.3.2 Fertilizers

Heavily oiled sites that have been washed or otherwise physically treated to remove the bulk of the oil always retain some residual oil, and this is an ideal target for bioremediation through the addition of fertilizers. Lightly oiled beaches may also provide excellent opportunities for use of this technology without pretreatment. Aerobic oil-degrading microbes are ubiquitous, but their abundance is usually limited by the availability of oil. An oil spill removes this limitation, and there is invariably a "bloom" of oil-degrading microbes, both in terrestrial and in aquatic environments. Their growth becomes limited by the availability of something else. In marine environments, this is usually nitrogen, and although a few oil-degrading microorganisms are able to fix atmospheric nitrogen, most have to await the arrival of low levels of nitrogen with every tide. Carefully applying fertilizers to partially

overcome this limitation can substantially increase the rate of biodegradation with no adverse environmental impact. Slow-release and oleophilic fertilizers (designed to adhere to the oil) seem most effective, and they were used with significant success following the 1989 spill from the Exxon Valdez in Alaska. The oleophilic fertilizer was applied with airless sprayers, and the granular slow-release fertilizer was applied with agricultural "whirligigs." Careful monitoring revealed no adverse environmental impacts from this approach and a stimulation of the rate of oil biodegradation of between two- and fivefold. In other words, although the technique did not make the oil disappear overnight, it sufficiently stimulated a process predicted to take a decade so that it occurred in a couple of years.

Since then, bioremediation has been used on a limited site as part of the cleanup of the 1996 Sea Empress spill and has been demonstrated on experimental spills in marine or brackish environments on the Delaware Bay, a Texas wetland, a fine-sand beach in England, mangroves in Australia, and an Arctic shoreline in Spitsbergen. Experience suggests that fertilizer should be added to maintain soluble nitrogen at approximately 100 μM in the interstitial water of the oiled sediment, and this may require reapplication every few weeks. Such a level of fertilizer nutrients does not seem to cause any adverse environmental impact but stimulates the rate of oil biodegradation severalfold.

Fertilizers have also been used with success at many terrestrial spill sites after the bulk of the spilled oil has been collected. Often, biodegradation at terrestrial sites is also limited by the availability of oxygen, and this can be ameliorated by tilling of surface soils or by the addition of oxygen as air or dissolved or slow-release peroxides for subsurface plumes. One approach suspends slow-release peroxides in a series of wells that intercept flowing groundwater plumes; others sparge air into similar wells.

5.3.3 Natural Attenuation

Biodegradation is such an effective environmental process that in some cases it can be relied on to clean an oil spill without further human intervention. Sometimes this is necessary because human intervention may cause more environmental impact than the spill, such as attempts to remove small amounts of oil from marshes. Other situations in which such an approach can be appropriate include underground spills, in which soluble hydrocarbons enter the groundwater. Usually, these are degraded *in situ*, and if the source is plugged the contaminants may be degraded before the plume intercepts any sensitive receptor. In such a case, it may be best to rely on the natural biodegradation process rather than attempt to stimulate it. Clearly, it is important to have a realistic understanding of the natural processes that will remediate the spill before natural attenuation is accepted as the appropriate response.

Together, these approaches exemplify modern environmentally responsible remediation activities—working with natural processes to stimulate them without causing any additional adverse environmental impact.

SEE ALSO THE FOLLOWING ARTICLES

Acid Deposition and Energy Use • Crude Oil Spills, Environmental Impact of • Ecosystems and Energy: History and Overview • Hazardous Waste from Fossil Fuels • Occupational Health Risks in Crude Oil and Natural Gas Extraction • Oil Pipelines • Petroleum System: Natures's Distribution System for Oil and Gas • Public Reaction to Offshore Oil • Tanker Transportation

Further Reading

Canadian Coast Guard (1995). "Oil Spill Response Field Guide." Canadian Coast Guard, Ottawa.

Fingas, M. (2000). "The Basics of Oil Spill Cleanup." 2nd ed. CRC Press, Boca Raton, FL.

National Research Council (1989). "Using Oil Spill Dispersants in the Sea." National Academy Press, Washington, DC.

National Research Council (1993). "In Situ Bioremediation. When Does It Work?" National Academy Press, Washington, DC.

National Research Council (2002). "Oil in the Sea III: Inputs, Fates and Effects." National Academy Press, Washington, DC.

Ornitz, B. E., and Champ, M. A. (2002). "Oil Spills First Principles." Elsevier, New York.

Tissot, B. P., and Welte, D. H. (1984). "Petroleum Formation and Occurrence." Springer-Verlag, Berlin.

Crude Oil Spills, Environmental Impact of

STANISLAV PATIN
Russian Federal Research Institute of Fisheries and Oceanography
Moscow, Russia

1. **Crude Oil Spills: Characteristics, Causes, and Statistics**
2. **Ecotoxicology of Oil in Marine Environments**
3. **Environmental Impacts of Oil Spills**
4. **Impacts on Living Resources and Fisheries**
5. **Oil Spill Response and Prevention Strategies**

Glossary

assessment An orderly process of gathering information about a system and determining the significance and causes of any observed changes following a stress on the system.

benthos The region at the sea bottom and the complex of living organisms and their specific assemblages that inhabit the bottom substrates.

biota The totality of all forms and species of living organisms within a given area or habitat.

disturbance A chemical or physical process (including stresses), often caused by humans, that may or may not lead to a response in a biological system, at the organismic, population, or community level.

ecotoxicology A discipline aimed at studying adverse effects of chemicals on biological and ecological systems.

hazards Combinations of properties and characteristics of a material, process, object, or situation that are able to harm and cause environmental, economic, or any other kind of damage.

littoral A nearshore ecological zone influenced by periodic tidal processes.

pelagial An offshore ecological zone located in open waters and inhabited by water-column organisms.

plankton The microscopic and near-microscopic plants (phytoplankton) and animals (zooplankton) that passively drift in the water column or near the sea surface.

response to stress The physiological, biochemical, reproductive, and other adaptive reactions, induced by stress, of a single organism or a whole assemblage (environmental impact).

risk The probability of realization of a particular hazard under specific conditions and within a certain period of time.

stress A chemical or physical process that leads to a response from a single organism, or at the level of a whole population or community. Acute and chronic stresses are distinguished, depending on duration of impact in situations of short-term and long-term exposure.

The largest percentage of accidental oil input into the sea is associated with oil transportation by tankers and pipelines (about 70%), whereas the contribution of drilling and production activities is minimal (less than 1%). Large and catastrophic spills belong to the category of relatively rare events and their frequency in recent decades has decreased perceptibly. Yet, such episodes can pose serious ecological risk and result in long-term environmental disturbances. Because oil spills are under particularly close scrutiny by environmentalists and strongly shape public opinion toward offshore oil development, they deserve special consideration.

1. CRUDE OIL SPILLS: CHARACTERISTICS, CAUSES, AND STATISTICS

Accidents and associated oil spills have been and continue to be inevitable events at all phases of oil and gas field development. On land, oil spills are usually localized and thus their impact can be eliminated relatively easily. In contrast, marine oil spills may result in oil pollution over large areas and present serious environmental hazards. A wealth of statistical data on accidental oil spills has been accumulated over several decades of hydrocarbon

 © 2004 Elsevier Inc. All rights reserved.

production on the continental shelves of many countries. Some of these data are presented in Fig. 1 and Tables I and II. These data as well as other corresponding materials demonstrate a number of points:

- Overall (with the exception of catastrophic spills that occurred during the Persian Gulf War in 1991), there is a trend toward lower volumes of oil spills during all maritime operations, as compared to the late 1970s.
- Particularly large oil losses result from accidental spills during tanker shipments, whereas the contribution of offshore drilling and production activities is minimal and the losses associated with oil storage and pipeline transportation occupy an intermediate position.
- Small and quickly eliminated oil spills are the most usual types of spills; because they occur frequently, they have the potential to create "stable" (long-term) oil contamination in areas of intense offshore oil production and transportation.
- Catastrophic spills (releasing more than 30,000 tons of oil) occur at a rate of zero to several incidents per year, yet these spills are responsible for the highly uneven distribution of anthropogenic input of crude oil into the sea and for its overall volume.

At least half of the oil produced offshore is transported by more than 6000 tankers. In spite of the clear tendency toward accident-free tanker operations (see Fig. 1), tanker-related accidents remain one of the major sources of ecological risk. The list of very large oil spills (tens of thousands of tons each), long as it is, may become even longer at any time. The probability of an accident involving a crude oil spill from an oil tanker is usually estimated as a function of the covered transport distance. According to one such estimate, based on many years of statistical data, the probability of an accident with more than 160 tons of oil spilled is 5.3×10^{-7} per kilometer of the tanker's route. Similar estimates can also be derived based on the total amount of transported oil.

According to a classification by the International Tanker Owners Pollution Federation (ITOPF), oil spills are categorized into three size groups: small (less than 7 tons), intermediate (7–700 tons), and large (over 700 tons). Information is now available on more than 10,000 spills, the vast majority of which (over 80%) are less than 7 tons. Analysis of causes of intermediate spills from 1974 to 2000 indicates that 35% of such spills occurred during routine operations (mainly oil loading/unloading). Accidents such as grounding and collisions gave rise to 42% of spills in this size category. The major cause of large spills relates to accidents (83%), with grounding and collisions alone making up 64%. The probability of catastrophic blowouts and large spills (more than 1000 tons) during drilling and production operations is rather small (one accident per 10,000 wells).

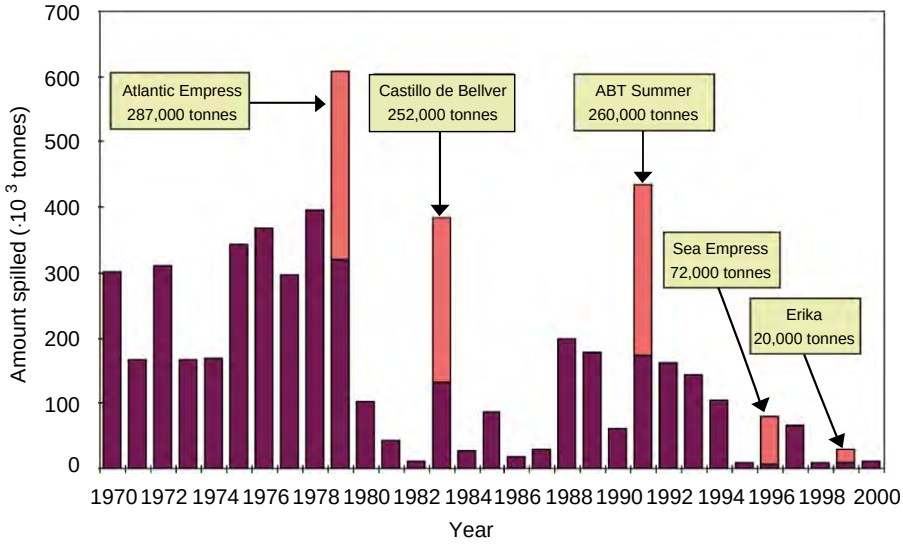


FIGURE 1 Quantities of oil spilled as a result of tanker incidents. Data from the International Tanker Owners Pollution Federation.

TABLE I

Oil Spills Associated with Different Types of Offshore Activities in 1997[a]

Activity	Number of spills	Average spill amount (tons)	Total spill amount	
			Tons	Percentage[b]
Drilling and production	4	191	764	0.5
Oil storage and handling	12	2579	30,948	19.4
Pipeline transport	63	788	49,644	31.2
Tanker transport	23	3391	77,993	48.9
Total	102	1562	159,349	100

[a] Summary from the International Oil Spill Database.
[b] Percentage of total amount spilled in 1997.

TABLE II

Selected Major Oil Spills and Accidental Releases

Source of accident and spill	Year	Location	Oil lost (tons)
Tankers, terminals, installations, Persian Gulf War events	1991	Kuwait; off coast of Persian Gulf and in Saudi Arabia	Up to 1,000,000
Exploratory well *Ixtoc I*	1979	Mexico; Gulf of Mexico, Bahia del Campeche	480,000
Tanker *Atlantic Empress*	1979	Off Tobago, West Indies	287,000
Platform No. 3 well (Nowruz)	1983	Iran; Persian Gulf, Nowruz Field	270,000
Tanker *ABT Summer*	1991	Open water; 1300 km off Angola	260,000
Tanker *Castillo de Bellver*	1983	South Africa; 64 km off Saldanha Bay	252,000
Tanker *Amoco Cadiz*	1978	France; off Brittany	223,000
Tanker *Haven*	1991	Italy; port of Genoa	144,000
Tanker *Odyssey*	1988	Canada; 1175 km off Nova Scotia	132,000
Tanker *Torrey Canyon*	1967	United Kingdom; Scilly Isles	130,000
Pipeline, Kharyaga-Usinsk	1994	Russia; Usinsk	104,000
Tanker *Braer*	1993	United Kingdom; Shetland Islands	85,000
Tanker *Aegean Sea*	1992	Spain; La Coruna harbor port	74,000
Tanker *Sea Empress*	1996	United Kingdom; Mill Bay, Milford Haven harbor port	72,000
Tanker *Prestige*	2002	Spain; over 200 km off Spanish coast	70,000
Tanker *Exxon Valdez*	1989	United States; Alaska, Prince William Sound	37,000

Over 30 years of offshore oil operations, only a few such episodes have occurred, primarily in the early 1970s.

To date, more than 100,000 km of subsea pipelines have been laid worldwide for transporting oil and other hydrocarbons. According to different estimates, the probability of a pipeline accident accompanied by a spill is anywhere between 10^{-4} and 10^{-3} spills/km/year. The data summarized in Table I suggest that the probability of pipeline accidents can be estimated at about 6.3×10^{-4} spills/km/year, with an average spill size of about 800 tons and worldwide oil losses of about 50,000 tons/year. Regional estimates may, of course, vary widely, depending on the local conditions. Accidental spills may be caused by many factors, including pipe corrosion, mechanical damage due to tectonic shifts, and impact by ship anchors. Pipelines become less prone to accidents as their diameter increases, but in all cases the probability of damage and ecological risk increases with aging of seabed pipelines. In general, pipelines are considered to be the safest way to accomplish hydrocarbon transport at sea and on land.

2. ECOTOXICOLOGY OF OIL IN MARINE ENVIRONMENTS

Crude oil is a naturally occurring material; its influx from seabed seepage alone amounts to more than 600,000 tons/year. Additionally, more than 10 million tons/year of aliphatic and other hydrocarbons are produced by marine organisms (mainly bacteria and phytoplankton). Unlike permanent natural sources, accidental oil spills usually occur within a short period of time and produce local zones of abnormally high hydrocarbon concentrations that are hazardous to marine life.

Chemically, crude oil (petroleum) is a complex of natural (mainly organic) substances. The basic component (up to 98%) consists of saturated and unsaturated aliphatic and cyclic hydrocarbons. In seawater, the original petroleum substrate rapidly (within a few hours or days) disappears as it breaks down into fractions (forms) in different states of aggregation. The most common aggregates include surface films (slicks), dissolved and suspended forms, emulsions (oil-in-water, water-in-oil), solid and viscous fractions precipitated to the bottom, and compounds accumulated in marine organisms. Under conditions of long-term (chronic) pollution, the dominant form is usually emulsified and dissolved oil. The fate (behavior, distribution, migration) of oil substances in the marine environment is controlled by a variety of interrelated processes. The major processes, usually known as weathering, include physical transport, evaporation, dissolution, emulsification, sedimentation, biodegradation (decomposition by microorganisms), and oxidation (Fig. 2).

The background contamination levels of petroleum hydrocarbons in the marine environment vary rather widely, from 10^{-5} to 10 mg/liter in seawater and from 10^{-1} to 10^{4} mg/kg in bottom sediments, depending on a spectrum of natural and technogenic factors. Maximum concentrations tend to occur in coastal and littoral waters as a result of many types of offshore, coastal, and land-based activities (developing hydrocarbon resources, transportation, industrial discharges, wastes dumping, etc.). Bioaccumulation (i.e., intake and assimilation by living organisms) of oil substances strongly depends on their hydrophilic and lipophilic properties. The concentration of polycyclic aromatic hydrocarbons (PAHs) in marine organisms is often up to three orders of magnitude above their content in seawater. The most pronounced capacity to accumulate PAHs without noticeable metabolic degradation in tissues is exhibited by filter-feeding bivalves (e.g., mussels).

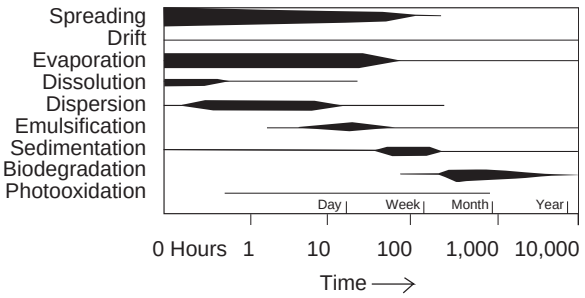


FIGURE 2 Processes acting on an oil slick as a function of time following a spill. The line thickness indicates the relative magnitude of each process. Reproduced from Swan *et al.* (1994) after Wheeler (1978).

From the ecotoxicological standpoint, oil is a complex multicomponent toxicant with nonspecific effects. Its basic component is a group of the most soluble and toxic aromatic hydrocarbons. High-molecular-weight PAHs with five and more benzene rings (e.g., benzo[*a*]pyrene) are able to produce carcinogenic effects, especially in benthic organisms, which inhabit heavily polluted bottom sediments. The primary responses of marine biota exposed to background levels of PAHs can be reliably detected by specific biochemical methods (e.g., measuring the induced activity of enzyme systems in individual marine organism tissues and organs).

For different groups of marine biota, toxic concentrations of oil (primarily aromatic hydrocarbons) in seawater and bottom sediments vary within several orders of magnitude. These differences are most pronounced in benthic communities of coastal zones polluted by oil, including those impacted by spills. Most species of the marine fauna exposed to oil are particularly vulnerable at early life stages (embryos and larvae). A simplified sequence of biological responses in marine biota, depending on oil concentrations in seawater and bottom sediments, is shown in Fig. 3.

A summary of experimental (toxicological) data and actual (measured) levels of oil content in seawater and bottom sediments of different areas of the worldwide oceans is given in Fig. 4. The upper limits of the zone in which there are "no observed effect" concentrations (NOECs), in which biological effects are either absent or manifest as primary (mostly reversible) physiological and biochemical responses, lie within the ranges of 10^{-3} to 10^{-2} mg/liter in seawater and 10–100 mg/kg in

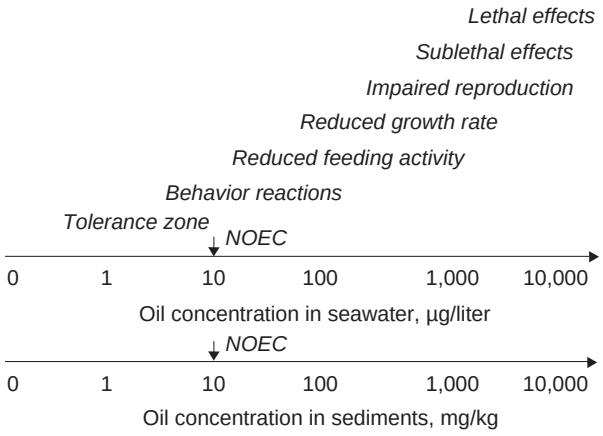


FIGURE 3 Developing characteristic effects and responses in marine biota, depending on the concentrations of oil in seawater and bottom sediments. NOEC, "No observed effect" concentration; see text for definition.

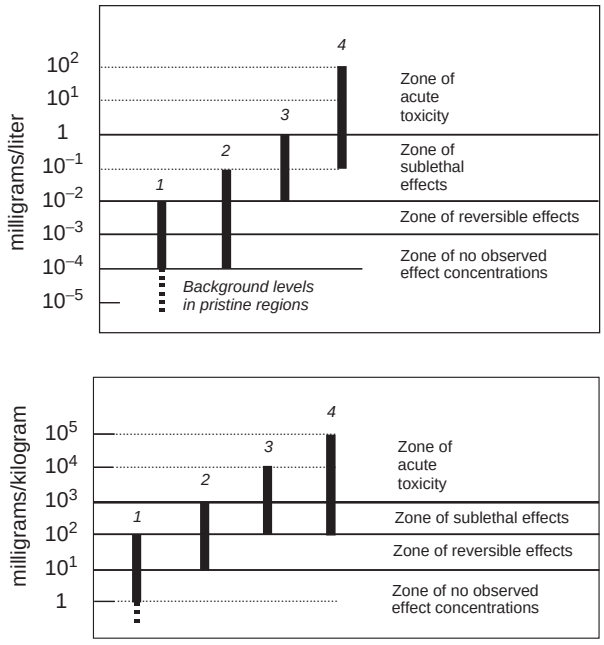


FIGURE 4 Approximate zones of biological effects and ranges of typical concentrations of oil hydrocarbons (mainly aromatic) in seawater (top) and bottom sediments (bottom). Zones: 1, pelagic areas (open waters); 2, coastal and littoral areas; 3, estuaries, bays, and other shallow semiclosed areas; 4, areas of local pollution, including oil spills.

bottom sediments. These ranges can be roughly considered as the limits of maximum permissible (safe) concentrations of oil hydrocarbons dissolved in seawater and accumulated in bottom sediments, respectively.

3. ENVIRONMENTAL IMPACTS OF OIL SPILLS

3.1 Methodology of Environmental Impact Assessment

Oil spills show one of the most complex and dynamic patterns of pollutant distribution and impact in the marine environment. In a sense, each spill is unique and unrepeatable because of the practically unlimited possible combination of natural and anthropogenic factors at a given location and at a given time. This poses a challenge that complicates the modeling of crude oil distribution, not to mention making quantitative predictions of the ecological effects of spills. At the same time, over the past several decades, a great number of special studies and field observations on oil behavior and the environmental hazards of oil spills have been conducted in a variety of ocean regions worldwide. These studies provide a good base for realistic assessment of the negative impact of oil spills on the marine environment and on living resources.

It should be noted that, to this day, no standard and widely recognized methodology of environmental impact assessment (EIA), including EIAs with respect to oil spills, actually exists. The most common methods used at different stages of EIAs include expert estimations, matrix analysis, simulations (modeling), hazards and risks theory, ecotoxicological studies, and field and laboratory experiments. Apparently, such a great diversity of methods and approaches extremely complicates standardization of the EIA procedure. At the same time, the general outlines of this procedure, including the contents and sequence of its stages, have already taken final shape in many countries.

Any system for assessing the state of environmental and human impacts has to be accompanied by a number of selected scales (classifications) for ranking spatial and temporal parameters of impact, as well as for specific evaluation of environmental hazards and risks. One of such classification used here is shown in Table III. Certainly, any scales and relative estimates of this kind will inevitably be circumstantial, but there is no other way to render the results of EIA more specific or objective. Otherwise, we are doomed to ineffective attempts to describe highly complex multifactorial natural systems with an endless number of direct and feedback relations (most often the nonlinear ones). This would leave plenty of room for subjective interpretations and uncertain estimates without any

TABLE III

Impact Scales and Gradation of Oil Spill Ecological Hazards and Consequences in Marine Environments

Impact	Definition
	Spatial scale
Point	Area under impact is less than $100\,m^2$
Local	Area under impact ranges from $100\,m^2$ to $1\,km^2$
Confined	Area under impact ranges within $1–100\,km^2$
Subregional	Area under impact is more than $100\,km^2$
Regional	Area under impact spreads over shelf region
	Temporal scale
Short term	From several minutes to several days
Temporary	From several days to one season
Long term	From one season to 1 year
Chronic	More than 1 year
	Reversibility of changes
Reversible (acute stress)	Disturbances in the state of environment and stresses in biota that can be eliminated naturally or artificially within the time span of several days to one season
Slightly reversible	Disturbances in the state of environment and stresses in biota that can be eliminated naturally or artificially within the time span of one season to 3 years
Irreversible (chronic stress)	Disturbances in the state of environment and stresses in biota that exist longer than 3 years
	General assessments
Insignificant	Changes in the state of environment and biota are absent or not discernible against the background of natural variability
Slight	Observed disturbances in the environment and short-term reversible stresses in biota (below minimum reaction threshold–0.1% of natural population reactions) are possible
Moderate	Disturbances in the environment and stresses in biota are observed without signs of degradation and loss of a system capacity for recovery; changes up to 1% of natural population reactions are possible
Severe	Stable structural and functional disturbances in biota communities (up to 10% of natural population and community parameters) are observed
Catastrophic	Irreversible and stable structural and functional degradation of a system is evident; disturbances in ecosystem parameters are more than 50% of statistical norm

assurance that the ultimate objectives of the EIA will be attained at all.

3.2 Oil Spill Types and Scenarios

From the environmental point of view, distinction should be made between two basic types (scenarios) of oil spills. One scenario includes spills that occur in open waters, without having the oil contact the shoreline and/or bottom sediments (pelagic scenarios). In this case, under the influence of wind, currents, and other hydrodynamic processes, the oil slicks are quickly (within a few hours or days) transformed and dispersed atop and under the sea surface. The dispersed and emulsified oil undergoes active biodegradation, its toxicity significantly decreases, and biological systems display attributes of acute (short-term) stress. The other type of spill, unfortunately more common and potentially more hazardous, involves influx of the oil onto the shore, its accumulation in sediments along the coast, and subsequent long-term ecological disturbances both in the littoral (intertidal) areas and on the shore. Such situations should certainly be treated as chronic stresses. The two scenarios often develop simultaneously, especially when an accidental oil spill occurs within a marine coastal zone not far from the shoreline.

According to available statistics, a majority of accidents and accompanying oil spills take place in coastal waters. In such cases, the likelihood of oil contact with the shoreline depends on many factors, including oil spill characteristics (volume and type of oil spilled, distance from the coast, etc.) and local oceanographic and meteorological conditions, particularly velocity and direction of winds and currents. In most accidents, the probability of an oil slick migration toward the coast and contact with the

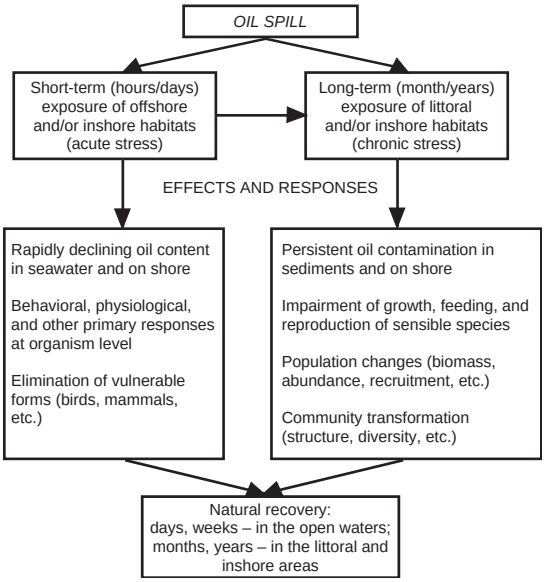


FIGURE 5 Conceptual scheme of characteristic biological effects and environmental hazards of oil spills in the sea.

shoreline is quite high (up to 50%), and the length of the coastal line impacted by oil can reach tens and even hundreds of kilometers. The conceptual scheme of developing biological effects and consequences of an oil spill under acute and chronic stresses is presented in Fig. 5. Depending on the duration, level, and scale of oil contamination, a wide range of effects may occur both in the water column and in littoral sediments—from behavioral responses of organisms at the initial phases of the spill up to long-term populational disturbances under chronic impact in the littoral zone.

3.3 Pelagic Scenarios (Oil Does Not Contact Shoreline and Bottom Sediments)

Ecological effects of oil spills in the open sea are mainly associated with short-term oil pollution of the surface water layers, where organisms such as plankton, nekton (mainly fishes), marine birds, and mammals may be affected when the situation is one of acute oil stress. Pelagic scenarios involve the formation of oil slicks in which total oil concentration ranges from 1 to 10 mg/liter in the surface seawater layers, at a depth of several meters. At horizons deeper than 10 m, oil contamination usually does not exceed background levels. After several days (sometimes weeks), surface slicks on open water are usually dispersed and eliminated under the influence of the processes and factors reflected in Fig. 2.

3.3.1 Impact on Plankton

Among all ecological forms, the hyponeuston, species that live within the most polluted surface microlayer (several centimeters deep at the atmosphere/sea interface), might be assumed to experience the heaviest toxic impact from oil spills. Nevertheless, in open waters, both hyponeuston and planktonic biota as a whole do not demonstrate any persistent structural and functional changes during pelagic oil spills. This is the result of a number of factors and processes, including a rapid decrease in the concentration of spilled oil to harmless levels, due to its dispersion and degradation in seawater; the rapid recovery of phytoplankton and zooplankton populations (within a few hours and days) due to high reproduction rates; and the influx of planktonic organisms from adjoining marine areas. There is not a single known publication in which oil spills have been shown to produce irreversible long-term impact on the planktonic populations of open waters.

3.3.2 Impact on Fish

Conclusions about the absence of any persistent long-term changes at the population level in planktonic biota after pelagic oil spills are fully valid in relation to fish populations. Due to oil localization within the surface layer of seawater, the mortality of water column organisms, including pelagic fish, is practically excluded. Fish mass mortality has never been observed even after catastrophic oil spills in the open sea. Available publications on the subject suggest that adult fish are capable of detecting and avoiding zones of heavy oil pollution.

An adverse impact of oil spills on fish is most likely to be observed in the shallow coastal areas of the sea where the water dynamics are slow. Fish in early life stages are known to be more vulnerable to oil, compared to adults, and, therefore, some younger fish may be killed by exposure to high concentrations of toxic components of crude oil. However, calculations and direct observations indicate that such losses could not be reliably detected against the natural background of high and variable mortality at embryonic and postembryonic stages of fish development.

3.3.3 Impact on Birds and Mammals

Sea birds and mammals are among the most vulnerable components of marine ecosystems in relation to oil pollution. Even short-term contact with spilled oil affects the insulating functions of hairy or feathery coats and results in quick death.

Studies of mass mortality (by tens of thousands of individuals) of birds and mammals after oil spills in coastal waters (even without oil contacting the shoreline) have been described in numerous publications.

The exact consequences of such spills depend primarily on the population traits of different species. Abundant populations with a high reproductive potential are the least susceptible to population stresses. At the same time, adverse impacts for the less abundant species with longer life spans are usually more serious and protracted. Of particular importance are the distribution and congregation patterns of birds and mammals. Even a small oil spill may severely affect a large number of birds or sea mammals that form compact colonies within a small area. In assessing such impacts, one must take into consideration the effect of natural factors, particularly natural mortality caused by extreme weather conditions, lack of food, epizootic diseases, etc. Most data suggest that populations of marine birds and mammals are able to return to their natural level (optimal for given conditions) within several years (seasons) after the oil spill impact.

3.3.4 General Assessment

Using the scales and ratings in Table III, typical oil spill impacts in pelagic scenarios can be assessed as varying from local to confined and from short term to temporary. Potential ecological hazards and biological effects in the form of acute stress for the marine pelagic biota can be assessed as predominantly reversible and varying from insignificant to moderate.

3.4 Littoral Scenarios (Oil Contacts Shoreline and Bottom Sediments)

When spilled oil reaches the shoreline, the major processes of oil accumulation, transport, and transformation occur in the littoral zone exposed to wind-induced waves, storms, surf, currents, and tides. The width of this zone may vary within tens of meters. Its potential self-cleaning capacity depends, first of all, on the geomorphology of the shoreline and characteristics of near-shore sediments and shore deposits (composition, fineness, etc.), as well as on the energy of wave and tidal processes. Numerous observations in different parts of the world indicate that oil persistence and, consequently, its adverse impact sharply increase from open rocky shores to sheltered gravel and pebble coasts (Table IV).

Some examples and characteristics of potential ecological impacts of oil spills in the littoral zone and adjoining shallow (several meters deep) sublittoral habitats are given in Table V. These assessments are based on the analysis of nearshore spill situations and describe biological effects resulting from oil pollution of both littoral sediments and seawater. Typically in such cases, local transformations of species population structures in littoral communities result in the predominance of more resistant species (e.g., some polychaetes) and rapid elimination of sensitive ones (especially crustaceans, such as amphipods, and marine birds and mammals). The time of return to environmental quality and community structure may vary widely, from a month to several years, depending on the exact combination of different natural and anthropogenic factors.

TABLE IV

Classification of Shorelines Based on Their Potential Oil Spill Recovery Capabilities

Self-cleaning capacity	Type of shoreline	Process characteristics
High (type I)	Open rocky shoreline	As a result of wave and surf action, oil is removed within several weeks; there is no need to clean inshore habitats
Medium (type II)	Accumulative coasts with fine-sand beaches	Oil does not usually penetrate deeply into fine, compact sand and can be removed by mechanical means; oil contamination persists for several months
Low (type III)	Abrasive coasts with coarse sand, pebble, and gravel stretches	Oil rapidly penetrates into deeper layers (up to 1 m) and can persist for years
Very low (type IV)	Sheltered coasts with boulder, gravel, and pebble material; wetlands (marshes, mangroves)	Favorable conditions for rapid accumulation and slow degradation of oil; tars and asphalt crusts can be formed

TABLE V

Potential Biological Effects of Oil Spills in the Littoral (Intertidal) Zone

Shoreline type[a]	Recovery capacity	Typical oil content levels in Water[b] (mg/liter)	Typical oil content levels in Bottom (mg/kg)	Characteristic biological effects
I	High	<0.1	$<10^2$	Mortality among the more sensitive species within the first days of contact; sublethal effects; primary structural changes in communities; recovery time, up to a month
II	Medium	0.1–1.0	10^2–10^3	Elimination of crustaceans (especially amphipods), predominance of polychaetes; decreasing biomass and structural changes in communities; recovery time, up to half a year
III	Low	1–10	10^3–10^4	Mass mortality among the more sensitive species (crustaceans and bivalves); steadily decreasing biomass and poorer species diversity; recovery time, up to a year
IV	Very low	>10	$>10^4$	Mass mortality among most species; pronounced declined biomass, abundance, and species diversity; recovery time, more than a year

[a] See classification in Table IV.
[b] Dissolved and emulsified forms of oil.

There are numerous records of oil spills in coastal and littoral areas in many marine regions, as well as hundreds of publications with detailed descriptions of associated environmental impacts. Besides the well-known catastrophic accident involving the tanker *Exxon Valdez* in the nearshore area of Alaska in 1989, a number of similar events have been documented, including oil spills in the coastal waters of the United States, France, United Kingdom, and Spain, among other countries (see Table II). One of the largest oil spills occurred in February 1996 when the tanker *Sea Empress* ran aground within a kilometer of the South Wales shore and 72,000 tons of crude oil were released. Despite favorable weather conditions (only 5–7% of the spilled oil reached the shore) and effective clean-up operations, the accident resulted in heavy pollution of 200 km long the coastline and littoral zone. About 7000 oily birds were collected on the shore alone. At the same time, no mass mortality of commercial fish, crustaceans, or shellfish was recorded. The time of environmental recovery in different locations was quite variable (from several months to a year and more), depending on adopted criteria and specific conditions within the affected littoral zone.

The list of this kind of accident in coastal areas is rather long. All these situations, unlike the purely pelagic spills (oil does not contact shoreline and bottom sediments), have two major phases, interpretable as acute and chronic stresses. Acute stress usually occurs within the first few days (sometimes weeks) and may result in mortality of the oil-impacted organisms (especially birds and mammals). Chronic stress lasts for several months, seasons, or even years and manifests mainly as adaptive modifications of benthic communities in response to the long-term oil contamination of littoral and coastal (inshore) sediments. In a majority of cases (excluding heavy oil pollution during very large catastrophic spills), these modifications do not exceed critical population levels and do not result in irreversible changes in structure and function of benthic populations.

3.4.1 General Assessment

In terms of ratings and definitions in Table III, typical oil spill impacts in littoral scenarios can be assessed as varying from local to subregional and from temporary to chronic. Environmental hazards and biological effects in the form of

chronic stress for the marine biota (mainly for benthic organisms) can be assessed as predominantly slightly reversible and varying from slight to severe.

4. IMPACTS ON LIVING RESOURCES AND FISHERIES

Until recently, there has been no direct evidence of any detectable impact of oil spills on the stock and biomass of commercial species. This conclusion is supported by a number of special studies, including those conducted in the situations of catastrophic accidents. Results of modeling and calculations of potential losses of commercial fish and invertebrates during oil spills lead to similar conclusion. Most estimates indicate that losses of commercial species during spills, even in the most pessimistic scenarios, usually do not exceed hundreds/thousands tons of biomass and cannot be reliably distinguished against the background of extremely high population variability (mainly due to environmental changes, natural mortality, and fishing).

Of particular concern is the possibility of the negative impact of oil spills on the anadromous fish (primarily, salmons) during their spawning migration shoreward. In this regard, it should be remembered that in the open sea, the oil concentration beneath the oil slick decreases sharply with depth, approaching background values at horizons of 5 to 10 m. Detailed studies conducted during a number of years after oil spills near the shores of Alaska and in other salmon breeding regions have shown that all basic parameters of commercial populations (abundance, structure, reproduction, migration, stock, and catches) remain within the natural ranges. Thus, there is a reasonable consensus that negative effects of oil spills on marine living resources in terms of their population parameters are negligible and could not be practically detected due to the enormous natural variability and dynamics of marine populations and communities. This conclusion does not mean, however, that oil spills do not exert any adverse impact on fisheries and the entire fishing industry. Numerous studies have shown that the actual economic losses to fisheries occur as a result of temporary restrictions on fishing activities, interference with aquaculture (i.e., growing and reproduction of marine organisms in marine farms) routines, loss of marketable values of commercial species due to petroleum taste (tainting), and oil contamination of fishing gear. The most serious economic losses for fisheries result from the limitations (bans, closures) imposed on fishing following oil spills in impacted coastal areas. Commercial fishermen are deprived of their livelihood unless they are able to exploit alternative stocks. Naturally, the exact nature and extent of such losses may vary widely, depending on a combination of diverse factors, such as oil spill size, time of year, and weather conditions. In some cases, during the catastrophic accident involving the tanker *Exxon Valdez* near Alaskan shores in the spring of 1989, for example, the losses to local fisheries were estimated at hundreds millions of dollars. This amount was eventually paid by court decisions in the United States as compensation for the restriction on fishing because of heavy oil pollution. Significant losses are known to have been sustained by aquaculturists, as was the case with many farms specializing in growing fish and shellfish during oil spills near British shores in 1993 and 1996. Similar events occurred in the winter of 2002 due to catastrophic oil pollution of the Atlantic coasts of Spain and Portugal after an incident involving the tanker *Prestige* caused release of 70,000 tons of heavy fuel oil.

Another rather common yet poorly controlled type of damage to fisheries inflicted by oil spills is associated with accumulation of petroleum hydrocarbons (especially light aromatic compounds) in marine organisms. This does not usually lead to intoxication of the organisms but results in an oily odor and taste, which inevitably means deterioration of marketable qualities of marine products and corresponding economic losses.

Legal and regulatory mechanisms of compensation for losses sustained by the fishing sector as a result of oil spills are yet to be fully elaborated, although two special international conventions are already in effect: the International Convention on Civil Liability for Oil Pollution Damage (Brussels, 1969) and the International Convention on the Establishment of an International Fund for Compensation for Oil Pollution Damage (London, 1971; amended by the 1992 Protocol). These conventions, administered by member states of the United Nations, set forth a number of measures and provisions for paying compensations to countries in cases of accidental oil spills within the bounds of their continental shelf. In the context of the conventions, fishery closures are expected to minimize or prevent economic damage that might otherwise occur.

5. OIL SPILL RESPONSE AND PREVENTION STRATEGIES

Public concern and anxiety about oil spills in the sea have been clearly augmented since the *Torry Canyon* tanker accident off the southwest coast of the United Kingdom in March 1967, when 130,000 tons of spilled oil caused heavy pollution of the French and British shores, with serious ecological and fisheries consequences. This accident was followed by a number of similar catastrophic episodes (see Table II). These events produced an acute public reaction and clearly elucidated a need for developing effective prevention and response strategies for appropriate reactions in such situations. Since the time of these spills, an impressive technical, political, and legal experience in managing the problem has been acquired in many countries and on the international level, mainly through the International Maritime Organization. In 1990, the International Convention on Oil Pollution Preparedness, Response, and Co-operation was adopted by a number of countries. The convention provided the framework for coordination of the national and international efforts to create special funds, centers, technical means, and methods for dealing with the consequences of oil spills in coastal areas. In practically all nearshore countries, specialized oil spill response centers and services have been established. Additionally, any offshore oil project is required to develop "action emergency plans." In many countries, national plans on preparedness and responses to oil spill accidents have been developed and appropriate legal rules, norms, and regulations are enforced.

An idea about the scope of economical losses and costs of oil spills may be derived from the *Exxon Valdez* incident in 1989; the clean-up operations were conducted during 2 years with about 10,000 people participating, and response costs totaled approximately $2.8 billion (total costs exceeded $9 billion). In the aftermath of this incident, the U.S. Congress adopted the 1990 Oil Pollution Act, which expanded the oil spill prevention, preparedness, and response responsibilities of both the federal government and industry.

The wide spectrum of contemporary techniques for oil spill response and clean-up operations may be divided into four major groups:

1. Mechanical means (booms, skimmers, etc.) for collection and removal of oil from both the sea surface and inshore habitats.
2. Burning of spilled oil at the site.
3. Chemical methods for dispersing surface oil slicks, aimed at accelerating oil degradation by natural factors.
4. Microbiological methods for enhancing oil degradation, usually applied in combination with mechanical and chemical methods.

The great variety of means and methods are quite relevant, taking into account the enormous variation in emergency situations as well as the extremely complex behavior of crude oil in marine environments. In the field, a combination of diverse methods is commonly used, with preference for mechanical means of oil containment and removal. Chemical and microbiological methods are applied predominantly in addition to the mechanical operations.

There are two situations when oil spill response and clean-up operations are not justified. First, oil spills in the open sea in areas with active surface water mass hydrodynamics, at great depths, and far from any shoreline need not be mitigated. As already noted, in such situations, oil slicks are relatively rapidly (usually within several days) dispersed without any significant damage to water column organisms and pelagic ecosystems. Second, shorelines are classified according to their vulnerability to oil spills. As was shown in Table IV, along open rocky shorelines, oil can be quickly dispersed and eliminated as a result of favorable natural processes (i.e., strong wave, tide, and wind activities). It is quite evident that in these situations human intervention may be senseless. Unfortunately, such situations are the exception rather than the rule. The littoral scenarios, with oil entering and accumulating in shallow coastal zones with low-energy hydrodynamic processes, are more usual. In these cases, the necessity of cleaning operations is quite evident.

Clearly, the techniques and means for alleviating the impacts of oil spills describe herein represent only a few of the armamentarium of contemporary oil spill response and prevention strategies. These strategies have been implemented worldwide and represent the development of regional systems of accident monitoring and reporting and coastal sensitivity analysis and mapping, involving adoption of detailed plans for dealing with emergencies and production of databanks and information systems for assisting spill response management. In some countries (e.g., the United Kingdom), integrated booming strategy plans (employing anchored oil-blocking booms) are being developed for protection of environmentally sensitive areas.

In conclusion, it should be emphasized that accidental oil spills are only one of a multitude of various factors of human impact on marine environments. The protection of oceans has always been and shall always be a priority task of international significance.

SEE ALSO THE FOLLOWING ARTICLES

Crude Oil Releases to the Environment: Natural Fate and Remediation Options • Ecological Risk Assessment Applied to Energy Development • Gulf War, Environmental Impact of • Hazardous Waste from Fossil Fuels • Occupational Health Risks in Crude Oil and Natural Gas Extraction • Oil and Natural Gas: Offshore Operations • Oil Pipelines • Tanker Transportation • Wetlands: Impacts of Energy Development in the Mississippi Delta

Further Reading

Cairns, W. J. (ed.). (1992). "North Sea Oil and the Environment. Developing Oil and Gas Resources, Environmental Impacts and Responses." Elsevier Applied Science, London and New York.

Davies, J. M., and Topping, G. (eds.). (1997). "The Impact of an Oil Spill in Turbulent Waters: The Braer." The Stationery Office, Edinburgh.

Etkin, D. S. (1999). Historical Overview of Oil Spills from All Sources (1960–1998). *In* "Proceedings of the 1999 International Oil Spill Conference." American Petroleum Institute, Washington, D.C.

Joint Group of Experts on the Scientific Aspects of Marine Pollution (GESAMP). (1993). "Impact of Oil and Related Chemicals and Wastes on the Marine Environment." GESAMP Reports and Studies, No. 50. International Maritime Organization, London

Patin, S. A. (1999). "Environmental Impact of the Offshore Oil and Gas Industry." EcoMonitor Publ., New York.

Swan, J. M., Neff, J. M., and Young, P. C. (eds.). (1994). "Environmental Implications of Offshore Oil and Gas Development in Australia." Australian Petroleum Exploration Association, Sydney.

Cultural Evolution and Energy

RICHARD N. ADAMS
University of Texas at Austin
Austin, Texas, United States

1. Introduction
2. Energy as a Conceptual Tool for Social Analysis
3. Dynamics of Energy Flow
4. Energy in Evolution of Human Society
5. Concluding Note

Glossary

circumscription theory As formulated by Robert L. Carneiro, the process whereby threats to the survival of a society stemming usually from population pressure on resources, but also from competition from other societies, and environmental constraints, lead to the increasing complexity of the sociopolitical organization.

cultural evolution (1) The process whereby human beings adapt their social and cultural behavior to deal with challenges to their survival; (2) the process whereby human societies become more complex as they adapt to deal with challenges to their survival; such challenges may be internal or external.

dissipative structure An open energy system with a form created by the constant incorporation of energy and matter as inputs and dissipation as output.

energy maximization principle As formulated by Alfred Lotka, the principle that, among living organisms, "natural selection will so operate as to increase the total flux through the system, so long as there is presented an unutilized residue of matter and available energy."

energy triggers The process whereby the dissipation of one energy flow releases dissipation from another energy source.

levels of sociocultural integration See **stages of cultural evolution.**

minimum dissipation principle As formulated by Ilya Prigogine and J. M. Waime, the principle that "when given boundary conditions prevent the system from reaching thermodynamic equilibrium (i.e., zero entropy production) the system settles down in the state of 'least dissipation'."

stages of cultural evolution Successive series of configurations of sociocultural organization that emerge as societies become increasingly complex; these various configurations are also known as levels of sociocultural integration.

Human societies, like all other living phenomena, exist by virtue of extracting energy from the environment and dissipating it in specific ways. The laws, principles, and measures applying to energy dissipation are as applicable to human society and culture as they are to the rest of nature. Theories of social and cultural evolution have developed with limited attention to this fact. The bulk of the present account is divided into three parts. The first two concern using energy dynamics in social analysis and are directed at social scientists interested in the question. The second is concerned with cultural evolution, and are directed more at those unfamiliar with the development of theory on this topic.

1. INTRODUCTION

Sociocultural evolution was first seriously explored in the 19th century by Herbert Spencer, Edward B. Tylor, and Lewis Henry Morgan, but fell under the cloud of cultural relativism in the early 20th century. The concept of sociocultural evolution was subsequently revived by Leslie White, V. Gordon Childe, and Julian Steward, and was further developed by Elman R. Service, Marvin Harris, and Robert L. Carneiro and by many of their students. Sociocultural evolution has proved to be of interest in archeology for modeling sociocultural change over extended eras of time and comparatively between different cultural traditions. It has been taken up in sociology by Gerhard Lenski and Stephen K. Sanderson and in a more Darwinian form by Donald Campbell. Also, under the influence of sociobiology, a distinctive branch of coevolutionary thinking (which will not concern us here) has been explored

 © 2004 Elsevier Inc. All rights reserved.

by Robert Boyd and Peter J. Richarson, Charles J. Lumsden and Edward O. Wilson, and William H. Durham.

The notion that energy was relevant to understanding the social process was early described by Herbert Spencer in his *First Principles* (1863): "Based as the life of a society is on animal and vegetal products, and dependent as these are on the light and heat of the Sun, it follows that the changes wrought by men as socially organized, are effects of forces having a common origin…Whatever takes place in a society results either from the undirected physical energies around, from these energies as directed by men, or from the energies of the men themselves." Following this assessment, the connectedness of energy and cultural evolution was primarily pursued by scholars with interests other than cultural evolution. By the early 20th century, it had been seriously discussed by Wilhelm Ostwald, Patrick Geddes, and Frederick Soddy, to be followed by Alfred Lotka, T. N. Carver, and Lewis Mumford. In 1943, anthropologist Leslie White proposed it as a tool for social analysis. It was then explored in more detail by sociologist Fred Cottrell, ecologist Howard Odum, economists Nicolas Georgescu-Roegen and Herman Daly, anthropologist Richard N. Adams, and, recently, sociologists Eugene A. Rosa and Gary E. Machlis. In the latter part of 20th century, world events produced an oil crisis and with it an explosion of interest in the political economy of energy in society.

2. ENERGY AS A CONCEPTUAL TOOL FOR SOCIAL ANALYSIS

Because energy is dissipated when anything happens, it is a substantive process that provides a common denominator for all activities, human and nonhuman. The second law of thermodynamics defines the irreversibility (and therefore in an important sense, directionality) of the model and thus is a measure for everything that happens. Energy can be thought of as a material cause of everything, but in an instrumental, or agential, sense energy causes nothing; it merely requires that all events evolve so as to reduce the capacity to do work that may be available in the system. This provides a measure for the analysis of human history, one that is culture free and value free and uncolored by political or moral concerns. This permits comparison among all historically known societies of any culture.

Although not always explicit, two perspectives—the agential and holistic—have been used to describe the relationship between energy and human society. The first treats energy as a resource that is controlled by human beings to secure their survival and adaptation, i.e., to sustain their life. Seen in this way, the environment is full of energy resources—biota (including human beings), fossil fuels, water and solar radiation, subnuclear sources, etc.—that human beings harness and control to work for them. Somewhat problematic in this picture is human labor, because individuals control their own labor and may control the labor of others. Thus, when referring to the energy per capita expended by a society in a contemporary industrial nation, we are usually talking about nonhuman expenditure, but when dealing with preindustrial societies and particularly in hunting and gathering societies, human labor is the major energy source. Two phases emerge from this situation: (1) energy (including one's own energy) as controlled by individual human beings and (2) energy as controlled by society, usually by some agent for the collectivity. In complex societies (i.e., with two or more levels of social integration), leaders or authorities decide how energy resources, both nonhuman and human collectivities, are to be used. This perspective readily separates human energy expenditure as an agential process that controls and is impacted by on-going energy-based events.

In the holistic perspective, human beings and the society and the environment relevant to them are a continuing and fluctuating field of multiple energy flows. Within this field, there are islands or sectors of potential energy (materials, fuels, stored goods) held in equilibrium, the energy of which may be released for human use when an appropriate technology is applied. There are also dissipative structures (see later) wherein human labor is seen as merely one part of the flow, as simply one more energy resource. In this perspective, the triggers (see later) that release potential energy are energy flows. This perspective allows us to see social processes in terms of a series of principles, laws, and working hypotheses derived from the physical and biological sciences. If sociocultural evolution is merely one facet of general biological evolution, then these concepts and principles should be equally applicable to social evolution.

3. DYNAMICS OF ENERGY FLOW

There are a number of energetic processes useful for understanding sociocultural evolution.

3.1 Laws of Thermodynamics

The basic dynamics of energy are described in the first and second laws of thermodynamics, which require no elaboration here. The fact that energy is dissipated makes it unidirectional and measurable. What has happened can never be repeated—no matter how much new events may appear to duplicate history. The presence of historical similarities is a challenge to find out why and how they came to be.

3.2 Equilibrium and Dissipative Structures

Although it was recognized that energy was central to the social process, Ilya Prigogine's 1947 formulation of the thermodynamics of nonequilibrium structures provided a key concept—the dissipative structure—for treating human beings and human societies as dissipative energy forms within a field of both equilibrium and dissipative structures. Equilibrium structures are energy forms in relative equilibrium, in which no exchange of matter or energy is, for the moment, taking place with the environment. Dissipative structures are open energy systems that constantly incorporate energy and matter as inputs and dissipate them as output. The form they take is dissipation. They are characterized by fluctuations that are essentially indeterministic when seen in a microscopic field, but, with the operation of natural selection, may contribute to directionality over longer courses of time.

As adapted from Prigogine's model the dissipative structure provides a remarkably accurate representation of how living bodies and societies operate. It combines the inputs of matter and energy but the output also involves the discarding of waste materials that may impact and change other energy flows. Particularly significant are impacts that release potential energy of equilibrium structures and that divert or modify other dissipative processes. To use dissipative structure with sociological models requires only that we find the energy values associated with the reality being represented.

One criticism of dissipative structures as an analytical model for human societies is that it was "a nice analogy, yet while it can explain certain qualitative changes in a system by showing how order emerges from fluctuations, it cannot account for complex cultural systems with their interdependent subsystems, which influence self-organization. And it cannot explain the new structure after a phase of instability." The point is not well taken because such models are not intended to explain the specifics of historical cases. Rather, they provide a model of structural dynamics—in this case, the presence of fluctuations inherent in dissipative systems, which, when energy flow is increased, may cause the structure to become more complex. This happens at critical points where the expansion forces new structures to emerge, or the system will fragment or collapse. Such dynamics characterize Carneiro's 1970 theory of the emergence of the state. The model cannot predict the cultural details of how it occurs—things that depend on specific historical antecedents.

3.3 Energy Maximization

Treating society as a dissipative process allows us to map sociological cases on the energy model and permits the measurement of the energy dynamics at work. With this, there are some energy principles that become useful. One of the earliest of these was Alfred Lotka's 1922 principle of energy maximization: "In every instance considered, natural selection will so operate as to increase the total flux through the system, so long as there is presented an unutilized residue of matter and available energy." It is not difficult to find illustrations of this, but wars are an obvious case. Cavalry dominates over foot soldiers, firearms dominate over the bow, nuclear bombs dominate over gunpowder. In more general terms, one indication that natural selection favors societies with the greatest dissipation of energy may be seen in the reality that societies that have survived and flourished into the present day, dominating competitors, are precisely those that have dissipated the most energy. The society or species that monopolizes resources leaves none for others. The second clause of Lotka's statement, however, states that when energy is not available, maximization cannot take place. This may result from the exhaustion of resources (the path to the "the limits of growth") but may also be due to the presence of boundary conditions or technological inadequacies that prevent access or expansion.

3.4 Minimum Dissipation

Boundaries to energy systems are critical because they may be imposed from the outside, or they may be created by society. This points to a second proposition, the Prigogine–Waime principle of minimum dissipation: "When given boundary conditions prevent the system from reaching thermodynamic

equilibrium (i.e., zero entropy production) the system settles down in the state of 'least dissipation'." This states the conditions under which maximization will not take place. Adult human life is a fluctuating steady state, a homeostatic system, in which internal controls maintain a level of energy use that does not exceed the available and anticipated rate or amount of input. Boundaries may be imposed by disastrous droughts, floods, and plagues, or gradually but consistently, such as by endemic diseases or endemic warfare. Boundaries are often imposed by society through controlling life, such as infanticide, careful birth spacing practices, and putting the elderly to death. Steady-state societies are those that have found ways of controlling the consumption of energy so as to survive on the energy available.

Another way that the principle operates may be seen in the splitting off of part of a society when it is growing too large for comfort. Thus, growing hunting bands and agricultural villages, reaching a certain population size, will hive off, forming new bands or settlements. New boundaries are created because the old boundaries prove inappropriate for survival. In this way, societies follow the minimum dissipation principle because the amount of energy flow is kept to some minimum within the society; but they may succeed in doing this only by creating other societies with boundaries to separate them from the old.

Classic cases have been the peasant philosophy of the "limited good," a pattern of social control whereby those who try to surpass others in some endeavor—and thereby stretch the exploitation of available resources—are condemned and punished by their fellows and restrained from excelling. Some highland New Guinea societies sponsor periodic pig feasts, as described by Roy Rappaport: "The operation of ritual…helps to maintain an undegraded environment, limits fighting to frequencies which do not endanger the existence of the regional population, adjusts man–land ratios, facilitates trade, [and] distributes local surpluses of pig throughout the regional population…." Socially controlling for minimum dissipation is usually achieved in societies only after hard experiences with the consequences of exceeding energy resources.

The advent of capitalism and industrial exploitation of the fossil energy resources, when coupled with an era of rapid population expansion, has generally obscured the implications of the Prigogine–Waime principle. Contemporary industrial societies have sought to expand energy use rather than to limit it. Although societies differ in the degree to which they attempt to shift to more dependable sources—e.g., energy derived directly from solar radiation, river or tidal waters, wind, or hydrogen—none has as yet made a marked shift in that direction. In most industrial countries, such concerns are marginalized by commercial and political interests that promote a policy of continual expansion of dependency on traditional sources. Everywhere, however, the Prigogine–Waime principle operates as available energy resources decline.

3.5 Trigger-Flow Ratio

Energy enters society principally as solar energy but also as moving water or wind and in equilibrium or quasiequilibrium forms, as potential energy—foodstuffs, minerals, forest products, etc. The release of potential energy is literally triggered and then, depending on the technology, controlled as it dissipates into lower grades and entropy, or leaves as a residue of material artifacts and waste. Because triggers and subsequent controls require energy to operate, it is generally expected that the ratio of their energy cost to the amount of energy flow dissipated should be low. Consistent failure to achieve this can lead to the trigger energy exhausting the energy available to the system.

Energy triggers are parts of a larger process, sometimes seen as an investment of energy. It has been proposed that this energy investment is a useful analytical tool, and as such has been referred to variously as embedded energy or, as by Howard Odem, as emergy. The relationship of this investment to the energy it causes to be released is seen as a net energy gain. Odum argued that an increase in the ratio of embedded energy to the energy released constitutes an increase in energy quality. The utility of this concept of energy quality is problematic because, as Joseph A. Tainter and colleagues have noted, "Energy quality in a human system varies with context, as it may also in other living systems." Yasar Demirel prefers to see quality as "the ability of energy to be converted into mechanical work under conditions determined by the natural environment." Emergy and embedded energy are not easy to apply to real situations because they imply an infinite series of antecedent inputs; in fossil fuels alone they include "the energy that previously went into the organism from which these fuels originate as well as the energy from the chemical and geological processes that transformed dead organisms into coal, gas, or petroleum." The importance of net energy lies in economic analyses of the more immediate energy costs.

Because food is everywhere required by human beings, changes in its dynamics reveal much about the course of cultural evolution. History has seen four major kinds of trigger technologies to produce food, all of which require human input: (1) human power in hunting and gathering, (2) cultivation based solely on human energy, (3) cultivation using work animals, and (4) industrialized agriculture. The ratio of trigger energy costs to resulting dissipation changes markedly in these different phases.

Minimally, trigger costs must involve human beings, but other energy sources are also required (tools, fuel, etc.). Table I orders data comparing societies at various levels of energy consumption in terms of the relative amounts of human and nonhuman energy used in food procurement. Early human beings, like contemporary hunters and gatherers, depended almost wholly on their own human energy, as illustrated by the !Kung bushmen (Table I, column C). !Kung gatherers' energy output was almost four times as much as their food energy input (Table I, column F). The development of cultivation increased this gain from 11 to 24 times as much (column F). However, because other energy resources were also used, the output/input gain was somewhat less (column E). With the introduction of work animals, the input of nonhuman energy increases (column C). In general, however, the use of draft animals adds to the trigger costs and provides relatively less yield than can be gained by human beings alone (compare the Mexican production in Table I); the advantage lies in the saving in time it allows. Industrialization brings a major increase in yield over human energy inputs (Table I, column F). With respect to total trigger costs, however, industrialization reduces the net energy gain to a level below that of the !Kung, and much below that of any all the subsistence agriculturalists (column E). In some cases, when the caloric yield is not the object of the cultivation—as with spinach and brussels sprouts—there is a clear net energy loss in the production (column E). Human activity continues to be important in industrial situations. Comparing Japanese and U.S. rice production shows that human activity yields greater efficiency in industrial situations. The Japanese farm requires much more human labor (column C), but

TABLE I

Trigger Energy Cost and Released Flow under Different Food Production Regimes[a]

	A	B	C	D	E	F
Food production effort	Human energy input (kcal)	Total energy input (kcal)	Total energy input/human energy input	Energy output (kcal)	Energy output/ total energy input	Energy output/ human energy input
Human energy gathering						
!Kung bushmen nuts	2680	2680	1.0	10,500	3.92	3.92
Human energy cultivation						
African cassava	821,760	838,260	1.0	19,219,200	22.93	23.39
New Guinea swidden	686,300	739,160	1.1	11,384,462	15.40	16.59
Mexican corn	589,160	642,338	1.1	6,901,338	10.74	11.71
Iban Borneo rice	625,615	1,034,225	1.7	7,318,080	7.08	11.70
Work animal energy						
Thailand peanuts (buffalo)	585,040	1,923,410	3.3	4,992,000	2.60	8.53
Mexican corn (oxen)	197,245	770,253	3.9	3,340,550	4.34	16.94
India wheat (bullocks)	324,413	2,827,813	8.7	2,709,300	0.96	8.35
Industrialized agriculture						
Japanese rice	297,600	8,221,040	27.6	22,977,900	2.80	77.21
U.S. brussels sprouts	27,900	8,060,328	288.9	5,544,000	0.69	198.71
U.S. spinach	26,040	12,759,849	490.0	2,912,000	0.23	111.83
U.S. rice	11,000	11,017,000	1001.5	23,642,190	2.15	2149.29
U.S. corn	4650	10,535,650	2265.7	26,625,000	2.53	5725.81

[a] From Pimentel and Pimentel (1996).

has a higher total net yield compared to the much more energy-expensive U.S. production (column E).

Although high-energy production is costly in kilocalories, it saves time. Pimentel and Pimentel calculate that the energy required to till a hectare of land by human power alone requires 200,000 kcal in 400 hr; with oxen, energy input increases to ≈300,000 kcal in 65 hr; with a 6-hp tractor, it increases to ≈440,000 in 25 hr; with a 50-hp tractor, it increases to ≈550,000 in 4 hr. Time, of course, may be critical to survival.

Comparable changes took place in the other areas in the Industrial Revolution. Whereas human labor in fossil-fueled industry at first expanded, the fossil fuels gradually displaced human labor in most of the production process. Adams showed that in Great Britain, for example, the human population in agriculture and industry grew to over 30% of the total in 1860, then declined by 1990 to well below 20%. In Japan, between 1920 and 1990, it also declined, from 36% to, again, below 20%. All countries that have become significantly industrialized have experienced this decline. However, although human energy used in triggers has declined, total energy costs have increased profoundly, drawing on what was thought to be an unlimited availability of fossil and gravitational energy. In industrial society, the energy cost of triggers grew not only through the accessibility of fossil fuels, but through the emergence of a vast complex system of linked triggers, comprising commercial, industrial, governmental, and societal organizations. Even where consumption is relatively low, the ever-increasing need for energy is a serious challenge to the future of industrial societies.

3.6 Entropy and Energy

Some scholars of energy and society have found entropy to be useful for societal analyses. Particularly influential was Nicholas Georgescu-Roegen's work in economics, directly arguing the implications of the second law of thermodynamics against more classical views of the inevitably progressive dynamics of the market. The focus on entropy in the biological and social sciences, however, took wings when Claude Shannon and Warren Weaver used the term to characterize the statistical patterns in information processes, because they conformed to Ludwig Boltzmann's algebraic formula for thermodynamic entropy.

Scholars differ in their acceptance of the appropriateness of identifying the informational process with thermodynamics. Jeffrey Wicken argues that the former does not refer to changes in physical space but rather in complexity of structure. He argues that it is really an "entropy analogue." It is not a process for which reversibility is highly improbable because complex states may have a low or high probability of becoming more or less complex. The identity has, nevertheless, been found useful by various scholars. Daniel R. Brooks and E. O. Wiley propose a formulation of biological evolution whereby "the strictly thermodynamic entropy of the organism…increases along with the informational entropy, since increasing structural complexity requires increasing energy to maintain the steady state." Sociologist Kenneth D. Bailey tends to put aside thermodynamic questions and elaborates a macrotheory of society that focuses on entropy in terms of the probabilities of order and variation.

Students of cultural evolution have not generally taken up the challenge of the entropy formulations, although Jeremy Rifkin, in a popular account, traced the increasing use of energy as being inextricably interrelated with the creation of disorder, chaos, and waste over the course of human history. The apparent preference of scholars for the analytic utility of energy over entropy may be because the materials of social evolution and history are more easily measurable in terms of the former than the latter.

4. ENERGY IN EVOLUTION OF HUMAN SOCIETY

4.1 Culture and Evolution

Sociocultural evolution is a part of the evolution of life, a process that has been best formulated in Darwinian terms. Although some discussions differentiate biological and social evolution, in fact, the individual and the social setting are necessarily complementary parts of a single process. Human life is generated in the biology of individuals, but it survives and reproduces by virtue of the aggregate activities and interactions that take place as a part of societies. Whereas all living forms are in some way complex, social complexity is best understood as the compounding of successive levels of integration, of which the individual is one. Thus, no matter the complexity of a human cell, an individual, a family, a band, or a nation, this complexity is compounded by requiring additional interrelationships—and hence organization—among the aggregated components, thereby creating new levels of integration.

Human society grows by increasing energy flow. This is done first by biological expansion of the population. Natural selection has favored those that organized in order to reproduce, thus creating societies—what Spencer called the superorganism. It is the evolution of human superorganisms that we call social or cultural evolution. The Darwinian processes of reproduction, variation, and natural selection concern individuals, but societies exist to propagate individuals and continue to survive while individuals die. Human society is a facet of natural selection in that it organizes people to reproduce the human species.

To a degree unmatched by other species, human beings control a vast range of extrasomatic energy forms that are captured, gathered, cultivated, and mined from the environment and other societies. This extraction of energy is a cost to the environment caused by the emergence and evolution of human society. It may also be seen as a coevolution in which the human component differentially promotes and inhibits the evolution of many other species. Culture consists of the mental models people make to deal with the external world, and the behavioral and material forms that correspond to these representations. People use these mental models to manipulate the external world, to control external energy forms. In actual historical fact, the models may be the basis for creating controls, or the model may be created to describe a control that has been discovered to work; in any event, the external controls and their models develop in an interactive process. Mental models trigger energy flows. When we speak of the evolution of culture, we are speaking of both the change in the mental models and the changes in energy forms that correspond to these models. Given that both are involved, the term "cultural evolution" has been used to refer to two phases of changing human behavior: (1) the process whereby human beings adapt their mental models and social and cultural behavior to deal with challenges to their survival, and (2) the process whereby mental models, societies, and cultures become more complex as they adapt to deal with challenges to their survival.

Some treatments of cultural evolution see the first as adaptation and prefer to see the increasing complexity in terms of more complicated or compounded social organizations, or in more complicated technology that allows for the control of greater sequences of energy. The relationship between energy and cultural evolution lies in two phases of the process. One is the construction of ever longer and more complex linked trigger-flow systems. The other is the harnessing of additional energy from the environment, allowing more opportunities to tap the flow through its successive stages of dissipation. The Australian aborigines set savanna fires that release an immense amount of energy but exercise no further control over it. There is no successive system of triggers. In contrast, the fire in a steam locomotive converts water to steam that pushes a piston, that directly turns the wheels, that travels along tracks and pulls a number of railway cars, etc. The detonation of a nuclear bomb requires the careful arrangement of a chain of trigger actions before the explosion emerges.

4.2 Population Growth

The most important dynamic process in sociocultural evolution is population expansion, a thing that, in itself, entails a proportionally greater extraction of nutritional energy from the environment. That energy, however, is then further consumed by human activity that further promotes human life.

Population pressure is central to Robert L. Carneiro's circumscription theory and to Marvin Harris' depletion theory of cultural evolution; it is also at the core of Esther Boserup's concern with technological change. As such, it risks being tautological; but it need not be. Circumscription theory proposes that an expanding population pressing against an unyielding environment produces changes in social organization. Harris's depletion theory holds that societies inevitably deplete their environment and must, therefore, intensify their production to retain their living standard. The problem is that the evidence for the population pressure and for depletion lies in the fact that the social organization changes. To avoid the tautology, the problems entailed by this pressure may be specified. The amount of energy flow is central to both processes. In both hunter-gatherer and agricultural societies, changes in population density or in level of food production threaten the availability of energy inputs to the society. Land:human ratio reduction and/or environmental depletion will reduce food availability to the point that social survival is threatened. The same argument applies in defense of Boserup's thesis that farming intensified because of declining production. Specifying the connectivity avoids the tautological by showing how social organization changes due a demonstrable need to improve the food supply.

A response to such pressure in simple societies is that the society fragments, allowing the separate

segments to hive off in the search for other lands. Another response is that the society reaches out for other resources by warfare or trade, using local resources, such as mineral or aquatic resources, to trade for the inputs needed. Where no other possibility is available—or at least known within the society's cultural repertory—members may turn to warfare. This in itself, especially in complex societies, requires additional expenditures of energy. Wars may fail or may succeed in eliminating the enemy and in taking their land and resources without producing major changes in social complexity beyond those necessary for war. However, if they capture the other society and control its production, then the emergence of political subordination, slavery, etc. requires a new level of integration, i.e., greater complexity. It is only in the last case that new external sources of energy are incorporated, and it is there that the society becomes more complex.

Although there is good reason to believe that population pressure and density led to social expansion and intensification in the preindustrial eras of human history, the emergence of states decisively changed the dynamics. States often want to hold a territory, regardless of whether it has any immediate productive use, and also to hold territories that may be important for production, but may be sparsely inhabited. States therefore may occupy vast territories that have little relevance to the immediate needs of their population. Given high levels of economic control and the relative ease of international migration, relatively small states may feel population pressures that are continually eased by emigration, both temporary and permanent. Urbanization, begun with the appearance of advanced agricultural systems, continued to be a central part of the solution under the state. The consequence of these differences is that even though hunting bands combine low energy consumption with low population density, there are some states that have low energy consumption—such as Haiti, Rwanda, and Bangladesh—that have very high population density; and there are states with high-level energy consumption—Australia, Canada and the Iceland—with very low densities. Such differences would not be possible in early (i.e., low energy consumption) phases of sociocultural evolution.

R. B. Graeber has proposed a mathematical theory of cultural evolution, relating population, territorial area, societal numbers, density, and mean societal size. He concludes that density and mean societal size are "indicators of the cultural-evolutionary state attained by a population." Although suggestive for earlier phases, the argument fails to deal with the contradictory cases among contemporary states.

4.3 Stages of Sociocultural Evolution

Anthropologists have long used stages of sociocultural evolution to characterize classes of societies in evolution. Although various sequences and scales of complexity have been proposed, no systematic effort has been made to relate them to the level of energy use. In the 19th century, Morgan, Tylor, Spencer, and Karl Marx all proposed sequences that were generally progressive and argued to constitute some kind of an "advance." In the past century, Childe, Cottrell, Service, and Morton Fried have made such propositions. The formulation of stages, however, has consistently reflected an author's particular interests, and some of the proposals were not made with energy in mind. Karl Marx proposed stages concerning modes of production (tribal, ancient, feudal, and capitalist) but he evinced no interest in thermodynamic issues. His follower, Friedrich Engels, was more explicit; for him to consider energy in economics was "nothing but nonsense."

The concern for technology early appeared in Morgan's sequence of savagery, barbarism, and civilization, whereby the transition between the first two was specifically based on domestication, and therefore with energy consumption and technology. The criteria for a civilization referred not to energy changes but to the appearance of cities and the population densities that were thereby implied. Energy conversion technology implicitly came to include control of human energy by the state. On this basis, Childe proposed two major revolutions in prehistory. The first was the Neolithic Revolution that specifically increased food production. The second was the Urban Revolution that was based on the invention of the plow, but its major manifestation was the emergence of cities based on the large economic surpluses. The sequence suggested in 1971 by Service—bands, tribes, chiefdoms, and state—shifted the concern from a technology to the relative complexity of social organization.

Evolutionary stages have been useful as sequences of kinds of societies reflecting differences in, among other things, energy consumption. Otherwise, the interest shown by social anthropologists in energy has been related to interests in ecology from an ethnographic perspective. Actual measurements have been forthcoming in a few field studies, such as those by Roy A. Rappaport in New Guinea, W. B. Kemp among Eskimo, Pertti Pelto with the Skolt Lapps,

and Richard Lee and L. J. Marshall with the !Kung Bushmen.

Contemporary treatments of the stages of cultural evolution have dispensed with the notions of progress or advance that were inherent in 19th and early 20th century versions. However, there have been few major formulations that take into account the potential exhaustion of energy resources for the species as a whole and the consequent potential decline that it implies. Tainter has explored the process of decline and collapse of specific societies and civilizations, but, given the dominant role of historical contingencies, there is little basis for predicting just how cultural devolution will emerge. There is reason to suspect, however, that although the societies and cultures became more complex with increases in energy, the internal dynamics of complex organizations will lead to decreasing returns in the future. Moreover, the social life of the species has evolved as innumerable separate societies, and these will each continue to trace their own course.

4.4 Energy and Complexity

No matter how heuristically useful stages of growth have been in ordering materials, they tend to imply an unwarranted punctuated quality to the course of cultural evolution. In fact, the process has been much more gradual, fluctuating over time and space, and the central quality that reflects change in energy flow is the change in complexity.

In 1943, Leslie White formulated the first specific proposition relating energy consumption to cultural evolution:

"Culture develops when the amount of energy harnessed by man per capita per year is increased; or as the efficiency of the technological means of putting this energy to work is increased; or, as both factors are simultaneously increased." Although it has recently been claimed that "the thermodynamic conception of culture has suffused cultural evolutionism," there is little work in the anthropological literature to support this. Whereas the domestication of plants and animals was accompanied by increases in population, fluctuating increases in energy consumption, and imperial expansions over the past 5000 years, the exponential per capita growth of both population and energy flow most markedly began with the Industrial Revolution. Boserup calculates that in 1500 AD the population density in Europe was still under 8 persons/km^2, and by 1750 it was still below 16 persons/km^2; by 1900 it had reached averages as high as 128 persons/km^2.

In defining sociocultural complexity, it helps to differentiate the social from the cultural. It is useful to see social complexity as the creation or differentiation of social relationships in new configurations. Most importantly, sociocultural complexity refers to the creation of new levels of social integration, hence to stratification, centralization, inequality, and emergence of hierarchy. Service's four stages outlined a succession of compounding organizations. Cultural complexity, however, refers to the heterogeneity and multiplicity of culture, i.e., meaning systems and the corresponding material and behavioral forms—as in technology (communication, engineering), social organization (voluntary associations, networking), and mental models (intellectual and symbolic systems) allowing for mathematics and astronomy, to suggest just a few. A number of scales of complexity have been proposed for simpler societies. One by Carneiro proposed the relevance of over 600 traits in 72 societies and another by George Peter Murdock and Catarina Provost was based on 10 traits in 180 societies; the two scales correlate highly. Henri J. M. Claessen and Pieter van de Velde have proposed a specific model of evolution as a complex interaction.

The salience of comparative energy consumption has been clearly related to social complexity in the study of contemporary societies at the national level. Somewhat contrary concerns for energy shortages on the one hand, and environmental damage on the other, have stimulated a concern over the past four decades for the relation between levels of energy consumption and various aspects of living. The interests here differ from those concerning the stages of cultural evolution, not merely because they are contemporary and deal with nation states, but because they are motivated by the concern for the survival of society in the short and long terms. At this level, there is little doubt that there is a relationship between various measures of quality of living and energy consumption. In 1970, Boserup divided 130 nation-states evenly into five groups based on technology and demonstrated the consistent difference in some indices of development and welfare. The first group had an average per capita energy consumption of 50, the third had 445, and the fifth had 4910. The corresponding life expectancies were 41, 55, and 72 years. The percentage of the population that was literate was 12, 46, and 98%, respectively. However, in a much-cited 1965 study, Ali Bulent Cambel reported that in countries with over $1500 per capita, the energy consumption varied widely from below 50 to over 175 Btu per capita.

In relating the amount of energy used by a society to its complexity, there are a number of conditioning factors. First, in both the agential and holistic perspectives, the control of energy must distinguish between controls that merely release energy but do not further control it; and controls that not only release but can limit or amplify the flow, its direction, and its impact, i.e., modulate the flow. In the first case, the energy is totally an output, a degrading that removes it from the society, such as setting an uncontrolled fire, detonating a bomb, or pushing a person off a cliff; the impacts are randomized. In the second, the available energy is kept within control of the society, and the impacts are controlled, such as the adjustment of the level of electric current, the ability of a court to vary a prison sentence, or of a farmer to decide when to plant a crop. The Australian aborigines who set extensive range fires to clear out savanna litter have no control over its dissipation. According to one estimate, such fires may release up to 500 million kcal of biomass energy a year—six times that used by the average American on a per capita basis—but the fires and the dissipation cannot be controlled.

Second, the increase of societal complexity depends not merely on the total energy released, nor even that for which there is control over the dissipation, but whether the controlled energy is effective in the self-reproduction of the society. In terms of comparative societal evaluation, this is technically more difficult to ascertain. One has to judge whether ceremonial structures and performances dedicated to appeasing a rain god are in fact more cost-effective than the construction of an irrigation system, or at what point the energetic costs of imperial expansion cease to be productive for the society's reproduction. Returning to White's proposition, a revised version might be as follows: the evolution of sociocultural complexity varies directly with the amount of energy controlled per capita per year and with the efficiency of the technological means of putting this energy to work in behalf of societal reproduction.

Although there are cases in which culture appears to be devolving and to be returning to some earlier form, in fact such a return simply does not happen. If there were a vast and gradual decline of energy available for consumption, it would require cultural simplification and reduction in complexity, but this would require new adaptations. But there would be no strict reversion to the social and cultural forms that emerged during the expansive phases of the species over the past few million years, or of the human civilizations of the past 50,000 years. The so-called devolution would, among other things, be the evolution of cultures that would be formally different from their predecessors, but necessarily simpler in the diversity of traits and relationships.

4.5 Partitioning of Energy

A facet of energy flow analysis that can help clarify the nature of society is how it is differentially partitioned among different internal functions and external relations. The way that energy is channeled and how this changes throw light on what is taking place in the structure of the society. For example, the proportion of the total energy flow involved in feeding the society seems remarkably similar in both simple and complex industrial societies. John A. Whitehead compared two Australian aboriginal bands with the population in the United States. Whereas the aboriginal bands used 20–30% of their energy resources for food gathering, in the United States, household plus commercial (which included more than food) consumption was estimated at 27%. Joel Darmstadter and colleagues found in eight Euro-American countries that the household–commercial sector constituted between 24 and 32% of the output, and 19% in Japan. This suggests that there is a consistent relationship between the nutrition requirements of the human beings and the economic process that surrounds them, irrespective of the energy level of the society.

In contrast, some differences appear in human energy partitioning as the total energy flow increases. In comparing a series of nations in the past 150 years, Adams found that the proportion of the population engaged in productive activities—i.e., agriculture and industry—was as high as 40% in the 19th century, but then dropped, often below 20% by the end of the 20th century. This was true of both highly industrialized countries as well as much of the Third World that increasingly received the industrial produce of advanced countries. In contrast, the population engaged in administration and commerce was everywhere between 2 and 4% in the early censuses, but then rose to as high as 20%, varying directly with the increase in nonhuman energy. The expansion of this regulatory sector of the population was systematically more marked in the countries with high levels of energy consumption.

Changes in the partitioning of energy consumption were fundamental in the evolution of the capacity for culture. Leslie C. Aiello and Jonathan C. K. Wells write that "the emergence of *Homo* is

characterized by…an absolute increase, to greater body size [and] a shift in relative requirements of different organs, with increased energy diverted to brain metabolism at the expense of gut tissue….” The nutrition patterns changed profoundly as the range of the human diet allowed greater territorial range, and digestive efficiency increased and the relative amount of energy required decreased. This decreasing proportion of energy dedicated to food processing was accompanied by an increase in the size of the brain and the relative amount of energy contributing to it. The increased efficiency of food processing was thus related to an increase in diet quality, which is consistent with the observation that primates with higher diet quality have higher levels of social activity.

These changes occurred relatively rapidly as the hominid brain from *Australopithecus* to *Homo sapiens* grew at a rate almost three times that which characterized ungulates and carnivores in general. Part of this growth was increase in human brain size, and more particularly the appearance of fissures that Harry J. Jerison argued allowed “the evolutionary building of integrated neural circuits through which diverse sensory information could be used to create increasing complex mental images.” This suggests that the ability to create models of reality—an important component in culture—existed 50 millions years ago. The energetic dimension of this change was reflected in the changing partitioning of energy use between the brain and the gut. Today, according to Marcus E. Raichle and Debra A. Gusnard, the brain represents 2% of body weight, but accounts for 20% of calorie consumption by the body.

4.6 The Limits of Growth

Although the conveniences made possible by increasing energy use and intensifying consumption were viewed as human progress by some, the vulnerability it implied was also recognized. Soddy early devoted a number of works to this aspect of contemporary societies, arguing that energy could be a limiting factor in social growth. He recognized that the shift from renewable agricultural resources to an increasing dependence on coal and, later, petroleum posed serious problems for advanced societies. Lotka’s proposition on the maximization of energy in living systems carefully specified that the dynamics of greater energy flows exercising dominance over lesser flow could take place only “so long as there is presented an unutilized residue of matter and available energy.”

For some, the question of energy shortages was less critical than a preoccupation with the damage done to the environment by the impacts of increasing energy expenditure. Concern for this was evinced in the 19th century by George Perkins Marsh and Alexander Ivanovich Voikof; more recently, Rachel Carson’s *Silent Spring* (1962) helped trigger the environmental movement in the United States. The remainder of the 20th century saw the expansion of environmental politics and studies concerned with the destruction of natural resources over much of the world. Besides the exhaustion of nonrenewable energy resources, another major concern has been the effect of excessive combustion of fossil fuels on the atmosphere, global warming, and related deterioration of the ecology of the planet. The chemical and material pollution of oceans, lakes, and the air was coupled with the deposition of millions of tons of waste over the earth’s surface. In all this, the role of increasing energy use was fundamental, both in the despoiling of resources and in polluting the environment. The central question is shifting from simply how to get more energy to how to reduce consumption.

The concern with the availability of energy resources in the nations consuming high levels of energy, especially the United States, replays the problems of circumscription and depletion in earlier human societies. The concern has dominated both economic and political policy and action to a degree that is seldom explored to the fullest. Availability has directly affected the development of environmental concerns in areas of potential oil exploitation, pitting those concerned for the long-term environmental and human sustainability issues against those interested in the monetary yield of exploiting as many resources as fast as possible and controlling resources in other parts of the world. The response of the United States to the terrorist attacks of 2001 was not unlike those postulated for early societies facing threats to their resources created by population growth and competitive societies.

5. CONCLUDING NOTE

Spencer recognized the importance of energetic processes to the evolution of culture and society in his formulation of the modern concept of evolution in the 19th century. It was formally articulated in 1943 by Leslie White, and interest since that time has been limited mainly to archaeologists, who have found in it a useful basis for comparing societies and

for modeling the activities of prehistoric societies, and to a few ethnologists working in technologically limited societies. Independently an understanding of nonequilibrium thermodynamic systems was opened in mid-20th century by the work of Prigogine, yielding the model of the dissipative structure. This has provided the basis for the mapping of social processes onto energetic dynamics, a potential as yet little explored. The importance of energy to the organization of society and to its future has become a matter of major concern to contemporary nations because of their increasing demand for energy, and by environmentalists and globally conscious politicians and economists in their concern for ecological survival and the planet's ability to support human life. What may seem to be an arcane interest in the role of energy in cultural evolution is directly related to the imperative concern for the survival of human society.

SEE ALSO THE FOLLOWING ARTICLES

Development and Energy, Overview • *Early Industrial World, Energy Flow in* • *Earth's Energy Balance* • *Global Energy Use: Status and Trends* • *Oil Crises, Historical Perspective* • *Population Growth and Energy* • *Sociopolitical Collapse, Energy and* • *Sustainable Development: Basic Concepts and Application to Energy* • *War and Energy*

Further Reading

Adams, R. N. (1988). "The Eighth Day: Social Evolution as the Self Organization of Energy." University of Texas Press, Austin.

Boserup, E. (1981). "Population and Technological Change." University of Chicago Press, Chicago.

Carneiro, R. L. (2003). "Evolutionism in Cultural Anthropology." Westview Press, Boulder, Colorado.

Claessen, H. J. M., van de Velde, P., and Smith, M. E. (1985). "Development and Decline, The Evolution of Sociopolitical Organization." Bergin & Garvey Publ., South Hadley, Massachusetts.

Darmstadter, J., Dunkerley, J., and Alterman, J. (1977). "How Industrial Societies Use Energy." Johns Hopkins University Press, Baltimore.

Khalil, E. L., and Boulding, K. E. (eds.). (1996). "Evolution, Order and Complexity." Routledge, London and New York.

Lenski, G., and Lenski, J. (1996). "Human Societies." Fifth Ed. McGraw-Hill, New York.

Pimentel, D., and Pimentel, M. (eds.). (1996). "Food, Energy and Society," Rev Ed. Niwot, University of Colorado Press, Colorado.

Prigogine, I., and Stengers, I. (1984). "Order out of Chaos." Bantam Books, New York.

Rosa, E. A., and Machlis, G. E. (1988). Energetic theories of society: An evaluative review. *Sociol. Inquiry* **53**, 152–178.

Sanderson, S. K. (1990). "Social Evolutionism." Blackwell, Cambridge and Oxford.

White, L. (1959). "The Evolution of Culture." McGraw-Hill, New York.

Decomposition Analysis Applied to Energy

B. W. ANG
National University of Singapore
Singapore

1. Introduction
2. Basics of Decomposition Analysis
3. Methodology and Related Issues
4. Application and Related Issues
5. Conclusion

Glossary

aggregate energy intensity The ratio of total energy use to total output measured at an aggregate level, such as industrywide and economywide.

decomposition technique The technique of decomposing variations in an aggregate over time or between two countries into components associated with some predefined factors.

Divisia index A weighted sum of growth rates, where the weights are the components' shares in total value, given in the form of a line integral.

energy intensity The amount of energy use per unit of output or activity, measured at the sectoral or activity level.

gross domestic product (GDP) A measure of the total flow of goods and services produced by the economy during a particular time period, normally 1 year.

index number A single number that gives the average value of a set of related items, expressed as a percentage of their value at some base period.

Laspeyres index An index number that measures the change in some aspect of a group of items over time, using weights based on values in some base year.

structure change A change in the shares or composition of some attribute, such as sector output in industrial production and fuel share in energy use.

This article discusses the decomposition techniques formulated using concepts similar to index numbers in economics and statistics. They have been called the index decomposition analysis. Approximately 200 energy and energy-related studies have been reported since 1978. Of these, slightly less than 20% deal with methodological or related issues and the remaining are application studies. Several other decomposition analysis methodologies have also been adopted in energy analysis, including the input–output structural decomposition analysis, which uses the input–output approach in economics to carry out decomposition. In terms of the number of studies reported, these techniques are relatively less popular compared to index decomposition analysis. The basics of index energy decomposition analysis are presented here. This is followed by a review of the methodological and application issues.

1. INTRODUCTION

Accounting for the largest share of primary energy demand in most countries, energy use in industry attracted considerable attention among policymakers, researchers, and analysts in the aftermath of the world oil crisis in 1973 and 1974. Effort was made by researchers and analysts to understand the mechanisms of change in industrial energy use through engineering process studies, energy audits, and industrywide analyses. At the industrywide level, a new research area began to emerge in approximately 1980, namely the study of the impact of changes in industry product mix, or structural change, on energy demand. It was realized that structural change could greatly influence energy demand trends, and there was a need to identify and quantify its impact to assist in policymaking. This line of research has since expanded substantially in terms of both methodology and application. It is now called decomposition analysis in the energy literature.

An important feature of decomposition analysis in the 1980s was its close linkages with the study of the

 © 2004 Elsevier Inc. All rights reserved.

aggregate energy intensity of industry. As an aggregate for the entire spectrum of industrial sectors/activities, this indicator is defined as the total industrial energy use in a country divided by its total industrial output. A decrease in the aggregate energy intensity is generally preferred since it means that a lower level of energy requirements is needed to produce a unit of industrial output. At approximately the same time, several cross-country comparative studies of industrialized countries were conducted, and countries with comparatively low aggregate energy intensity were considered more effective in energy utilization.

Researchers and analysts then began to examine factors contributing to changes in the aggregate energy intensity of countries over time (and variations in the aggregate energy intensity between countries). In the simplest form, two contributing factors were identified, namely structural change and energy intensity change. Industry is an aggregate comprising a wide range of sectors (or activities), such as food, textiles, petrochemical, transport equipment, and electronics. Structural change is associated with the varying growth rates among these sectors, which lead to a change in the product mix. Energy intensity change is determined by changes in the energy intensities of these industrial sectors. The sectoral energy intensity, or simply energy intensity, is the ratio of energy use to output for a specific sector defined at the finest level of sector disaggregation in the data set. To determine the impacts of these two factors, the observed change in the aggregate energy intensity is decomposed using an appropriate decomposition technique. Since 1980, numerous decomposition techniques have been proposed.

Information obtained through decomposition analysis has direct policy implications. The composition of industrial activity is effectively beyond the control of the energy policymaker except through indirect measures such as energy pricing. On the other hand, many measures can be taken to influence sectoral energy intensity change, such as taxation, regulatory standards, financial incentives, and information programs. The ability to accurately quantify the relative contributions of structural change and energy intensity change provides insight into trends in industrial use of energy. The success or failure of national energy conservation programs may be evaluated. The decomposition results also provide a basis for forecasting future energy requirements and for evaluating appropriate courses of action.

2. BASICS OF DECOMPOSITION ANALYSIS

Since index decomposition analysis was first applied to analyze energy use in industry, it is appropriate to explain the underlying concept with reference to the decomposition of a change in the aggregate energy intensity of industry. To keep the discussion simple, we consider the two-factor case in which a change in the aggregate intensity is decomposed to give the impacts of structural change and sectoral energy intensity change.

2.1 Notations and Basic Formulae

Assume that total energy consumption is the sum of consumption in n different industrial sectors and define the following variables for time t. Energy consumption is measured in an energy unit and industrial output in a monetary unit.

E_t = Total energy consumption in industry
$E_{i,t}$ = Energy consumption in industrial sector i
Y_t = Total industrial production
$Y_{i,t}$ = Production of industrial sector i
$S_{i,t}$ = Production share of sector i $(= Y_{i,t}/Y_t)$
I_t = Aggregate energy intensity $(= E_t/Y_t)$
$I_{i,t}$ = Energy intensity of sector i $(= E_{i,t}/Y_{i,t})$.

Express the aggregate energy intensity as a summation of the sectoral data:

$$I_t = \sum_i S_{i,t} I_{i,t},$$

where the summation is taken over the n sectors. The aggregate energy intensity is expressed in terms of production share and sectoral energy intensity. Suppose the aggregate energy intensity varies from I_0 in Year 0 to I_T in Year T. Such a change may be expressed in two ways, $D_{\text{tot}} = I_T/I_0$ and $\Delta I_{\text{tot}} = I_T - I_0$, respectively a ratio and a difference. Accordingly, decomposition may be conducted either multiplicatively or additively, respectively expressed in the forms

$$D_{\text{tot}} = I_T/I_0 = D_{\text{str}} D_{\text{int}}$$
$$\Delta I_{\text{tot}} = I_T - I_0 = \Delta I_{\text{str}} + \Delta I_{\text{int}},$$

where D_{str} and ΔI_{str} denote the impact of structural change, and D_{int} and ΔI_{int} denote the impact of sectoral intensity change. In the multiplicative case, all the terms are given in indices, whereas in the additive case all the terms, including the aggregate being decomposed, have the same unit of measurement.

To quantify the impacts, two approaches—the Laspeyres index approach and the Divisia index approach—have been widely adopted by researchers and analysts.

2.2 Laspeyres Index Approach

The Laspeyres index approach follows the Laspeyres price and quantity indices by isolating the impact of a variable by letting that specific variable change while holding the other variables at their respective base year values (in this case, Year 0). The formulae for multiplicative decomposition are

$$D_{\text{str}} = \sum_i S_{i,T} I_{i,0} / \sum_i S_{i,0} I_{i,0}$$
$$D_{\text{int}} = \sum_i S_{i,0} I_{i,T} / \sum_i S_{i,0} I_{i,0}$$
$$D_{\text{rsd}} = D_{\text{tot}} / (D_{\text{str}} D_{\text{int}}).$$

In additive decomposition, the formulae are

$$\Delta I_{\text{str}} = \sum_i S_{i,T} I_{i,0} - \sum_i S_{i,0} I_{i,0}$$
$$\Delta I_{\text{int}} = \sum_i S_{i,0} I_{i,T} - \sum_i S_{i,0} I_{i,0}$$
$$\Delta I_{\text{rsd}} = \Delta I_{\text{tot}} - \Delta I_{\text{str}} - \Delta I_{\text{int}}.$$

The Laspeyres index approach does not give perfect decomposition. The residual terms D_{rsd} and ΔI_{str} denote the part that is left unexplained. Decomposition is considered perfect if $D_{\text{rsd}} = 1$ and $\Delta I_{\text{str}} = 0$. These are also target values of the residual terms for evaluating the performance of a decomposition technique.

2.3 Divisia Index Approach

The Divisia index is an integral index number. In this approach, first the theorem of instantaneous growth rate is applied to $I_t = \sum_i S_{i,t} I_{i,t}$, which gives

$$d\ln(I_T/dt) = \sum_i \omega_i \left[d\ln(S_{i,t})/dt + d\ln(I_{i,t})/dt \right],$$

where $\omega_i = E_{i,t}/E_t$ is the sector share of energy consumption and is known as the weight for sector i in the summation. Integrating over time from 0 to T,

$$\ln(I_T/I_0) = \int_0^T \sum_i \omega_i \left[d\ln(S_{i,t})/dt \right] + \int_0^T \sum_i \omega_i \left[d\ln(I_{i,t})/dt \right],$$

Taking the exponential, the previous equation can be expressed in the multiplicative form $D_{\text{tot}} = D_{\text{str}} D_{\text{int}}$, where

$$D_{\text{str}} = \exp\left\{ \int_0^T \sum_i \omega_i \left[d\ln(S_{i,t})/dt \right] \right\}$$
$$D_{\text{int}} = \exp\left\{ \int_0^T \sum_i \omega_i \left[d\ln(I_{i,t})/dt \right] \right\}.$$

Since in practice only discrete data are available, the weight function has to be explicitly defined in actual application. Two of the weight functions that have been proposed in energy decomposition analysis are described next.

The first is the arithmetic mean of the weights for Year 0 and Year T, where

$$D_{\text{str}} = \exp\left\{ \sum_i (\omega_{i,T} + \omega_{i,0})/2 \ln(S_{i,T}/S_{i,0}) \right\}$$
$$D_{\text{int}} = \exp\left\{ \sum_i (\omega_{i,T} + \omega_{i,0})/2 \ln(I_{i,T}/I_{i,0}) \right\}.$$

Like the Laspeyres index approach, the approach does not give perfect decomposition, and we can write $D_{\text{tot}} = D_{\text{str}} D_{\text{int}} D_{\text{rsd}}$. The additive version can be derived in the same manner, and the relevant formulae are

$$\Delta I_{\text{str}} = \sum_i (E_{i,T}/Y_T + E_{i,0}/Y_0)/2 \ln(S_{i,T}/S_{i,0})$$
$$\Delta I_{\text{int}} = \sum_i (E_{i,T}/Y_T + E_{i,0}/Y_0)/2 \ln(I_{i,T}/I_{i,0}).$$

The decomposition is also not perfect, and we have $\Delta I_{\text{tot}} = \Delta I_{\text{str}} + \Delta I_{\text{int}} + \Delta I_{\text{rsd}}$.

The second and more refined weight function is given by the logarithmic mean of the weights for Year 0 and Year T. The corresponding formulae are

$$D_{\text{str}} = \exp\left\{ \sum_i \frac{L(E_{i,T}/Y_T, E_{i,0}/Y_0)}{L(I_T, I_0)} \ln(S_{i,T}/S_{i,0}) \right\}$$
$$D_{\text{int}} = \exp\left\{ \sum_i \frac{L(E_{i,T}/Y_T, E_{i,0}/Y_0)}{L(I_T, I_0)} \ln(I_{i,T}/I_{i,0}) \right\}$$
$$\Delta I_{\text{str}} = \sum_i L(E_{i,T}/Y_T, E_{i,0}/Y_0) \ln(S_{i,T}/S_{i,0})$$
$$\Delta I_{\text{int}} = \sum_i L(E_{i,T}/Y_T, E_{i,0}/Y_0) \ln(S_{i,T}/S_{i,0}),$$

where the function $L(a, b)$ is the logarithmic mean of two positive numbers a and b given by

$$L(a,b) = (a-b)/(\ln a - \ln b), \text{ for } a \neq b$$
$$= a, \text{ for } a = b$$

The advantage of the logarithmic mean weight function is that perfect decomposition is ensured. It

can be proven that the residual term does not exist (i.e., $D_{rsd} = 1$ and $\Delta I_{str} = 0$).

2.4 An Example

Table I shows a hypothetical case in which industry comprises two sectors. The sectoral energy intensity decreases from Year 0 to Year T for both sectors. In percentage terms, it decreases by 33% in sector 1 and by 20% in sector 2. By examining the sectoral data, one would conclude that at the given sectoral level, there are significant improvements in energy efficiency. At the industrywide level, however, the aggregate energy intensity increases by 20%. This apparent contradiction arises because of structural change. Sector 1, the more energy intensive of the two sectors, expands substantially in production level and its production share increases from 20 to 50%. The achievements in energy conservation, assuming that conservation measures have been successfully implemented within sectors 1 and 2, are therefore not reflected in the industrywide aggregate indicator.

Application of appropriate decomposition techniques to give the impacts of structural change and energy intensity change helps to resolve the previous apparent contradiction. The decomposition results obtained using the Laspeyres index method and the two Divisia index methods are summarized in Table II. Based on the results for the logarithmic mean Divisia method, we conclude that in multiplicative decomposition, the contribution of structural change increases by 70%, whereas that of sectoral energy change decreases by 29%, resulting in a net 20% increase in the aggregate energy intensity as observed. In the case of additive decomposition, structural change accounts for an increase by 0.58 units, whereas sectoral energy intensity change accounts for a decrease by 0.38 units, leading to a net increase in the aggregate energy intensity of 0.20 units.

It may be seen that the impacts of structural change and energy intensity change can be conveniently estimated and the relative contributions identified. In actual application, where very fine data are available, the number of sectors can be as many as a few hundred. The decomposition results obtained are method dependent, as shown in Table II, and the choice of an appropriate technique is often an application issue. The methodology has been applied to study problems with more than two factors and in many energy-related areas.

TABLE I

An Illustrative Example (Arbitrary Units)

	Year 0				Year T			
	E_0	Y_0	S_0	I_0	E_T	Y_T	S_T	I_T
Sector 1	30	10	0.2	3.0	80	40	0.5	2.0
Sector 2	20	40	0.8	0.5	16	40	0.5	0.4
Industry	50	50	1.0	1.0	96	80	1.0	1.2

TABLE II

Decomposition Results Obtained Using the Data in Table I

		Divisia approach	
	Laspeyres approach	Arithmetic mean	Logarithmic mean
Multiplicative			
D_{tot}	1.2000	1.2000	1.2000
D_{str}	1.7500	1.6879	1.6996
D_{int}	0.7200	0.7020	0.7060
D_{rsd}	0.9524	1.0127	1[a]
Additive			
ΔI_{tot}	0.2000	0.2000	0.2000
ΔI_{str}	0.7500	0.5920	0.5819
ΔI_{int}	−0.2800	−0.3913	−0.3819
ΔI_{rsd}	−0.2700	−0.0007	0[a]

[a] Perfect decomposition.

3. METHODOLOGY AND RELATED ISSUES

The methodological development of energy decomposition analysis since the late 1970s may be divided into three phases: the introduction phase (prior to 1985), the consolidation phase (1985–1995), and the further refinement phase (after 1995). It is a multifaceted subject, and the advances have been made due to the collective effort of researchers and analysts with backgrounds varying from science to mathematics, engineering, and economics.

3.1 The Introduction Phase (Prior to 1985)

The techniques proposed prior to the mid-1980s are intuitive and straightforward. Typically, the impact of structural change was singled out by computing the hypothetical aggregate energy intensity that

would have been in a target year had sectoral energy intensities remained unchanged at their respective values in a base year. The difference between this hypothetical target year aggregate energy intensity and the observed base year aggregate energy intensity is taken as the impact of structural change. The difference between the observed aggregate energy intensity and the hypothetical aggregate energy intensity, both of the target year, is taken as the impact of sectoral energy change. These techniques were first proposed by researchers to study trends of industrial use of energy in the United Kingdom and the United States in approximately 1980. No reference was made to index numbers in these studies, which means that the techniques were developed independent of index number theory. However, it was later found that these techniques were very similar to the Laspeyres index in concept. Methodologically, it is appropriate to refer to them as the Laspeyres index approach or the Laspeyres index-related decomposition approach.

3.2 The Consolidation Phase (1985–1995)

In this phase, several new techniques were proposed, with the realization that the Laspeyres index concept need not be the only one that could be adopted. The arithmetic mean Divisia index method was proposed in 1987 as an alternative to the Laspeyres index method. At approximately the same time, as an extension to the Laspeyres index formula that assigns the total weight to the base year, methods with equal weight assigned to Year 0 and Year T were formulated. This modification leads to formulae that possess the symmetry property.

In the early 1990s, attempts were made to consolidate the fast growing number of methods into a unified decomposition framework. A notable development was the introduction in 1992 of two general parametric Divisia index methods that allow an infinite number of decomposition methods to be specified. In these two methods, the weights for Year 0 and Year T are treated as variables whose values are explicitly defined. By choosing the weights appropriately, many earlier methods, including those based on the Laspeyres index approach and the Divisia index approach, can be shown to be special cases of these two general methods. Alternatively, the weights can be determined in an adaptive manner based on energy consumption and output growth patterns in the data set. In chaining decomposition, the weights would be updated automatically over time. By the mid-1990s, this technique, called the adaptive weighting parametric Divisia index method, was the most sophisticated decomposition technique. It was later adopted by some researchers and analysts.

3.3 Further Refinement Phase (after 1995)

Until the mid-1990s, all the proposed energy decomposition techniques left an unexplained residual term. Studies found that in some situations the residual term could be large, especially for the Laspeyres index approach. This issue became a concern because if a large part of the change in the aggregate is left unexplained, the interpretation of decomposition results would become difficult if not impossible, and the objective of decomposition analysis would be defeated.

Researchers began to search for more refined decomposition techniques, and the first perfect decomposition technique for energy decomposition analysis was proposed in 1997. This technique is similar to the logarithmic mean Divisia index method. In addition to having the property of perfect decomposition, the logarithmic mean Divisia index methods are also superior to the arithmetic mean Divisia method in that they can handle data with zero values. After the application of decomposition analysis was extended to include the effect of fuel mix in the early 1990s, the existence of zero values in the data set became a problem for some decomposition techniques. Zero values will appear in the data set when an energy source, such as nuclear energy or natural gas, begins or ceases to be used in a sector in the study period.

In the Laspeyres index approach, the residual terms, such as those shown in Table II, are known to researchers to consist of interactions among the factors considered in decomposition. A scheme to distribute the interaction terms, whereby perfect decomposition is achieved, was proposed in 1998. It has been referred to as the refined Laspeyres index method because it may be considered a refinement to the Laspeyres approach. In 2002, researchers introduced the Shapley decomposition, a perfect decomposition technique that has long been used in cost-allocation problems, to energy decomposition analysis. It has been determined that the refined Laspeyres index technique introduced in 1998 and the Shapley decomposition technique are exactly the same. Several other techniques that give perfect decomposition have been reported, including one that uses the concept of the Fisher ideal index.

3.4 Other Methodological Issues

In addition to method formulation, the following methodological issues are relevant to the application of decomposition analysis: multiplicative decomposition versus additive decomposition, disaggregation level of the data, and decomposition on a chaining basis versus a nonchaining basis. Each of these issues is briefly described.

For a given data set, the analyst may use either the multiplicative scheme or the additive decomposition scheme, and generally only one will be chosen by the analyst. The decomposition formulae for the two schemes are different, as are the numerical results obtained. However, the main conclusions are generally the same. As previously mentioned, the numerical results of the estimated effects are given in indices in multiplicative decomposition, but they are given in a physical unit in additive decomposition. Preference of the analyst and ease of result interpretation are the main considerations in the choice between the two decomposition schemes. Methodologically, the logarithmic mean Divisia index method has an advantage over other methods because the results obtained from the multiplicative and the additive schemes are linked by a very simple and straightforward formula. This unique property allows the estimate of an effect given by one of the two schemes to be readily converted to that of the other, and the choice between the multiplicative and additive schemes made by the analyst before decomposition becomes inconsequential.

The disaggregation level of the data defines the level or degree of structural change to be estimated. In the case of energy use in industry, past decomposition studies used levels varying from 2 sectors, which is the minimum, to approximately 500 sectors. The choice of disaggregation level dictates the results since structural change and energy efficiency change are dependent on the level chosen. Hence, result interpretation must be done with reference to the level defined and generalization should be avoided. A fine level of sector disaggregation is normally preferred. For instance, energy efficiency change is often taken as the inverse of the estimated energy intensity change. Derived in this manner, a good estimate of the real energy efficiency change requires that industry be broken down into a sufficiently large number of sectors. The quality of the estimate improves as the number of sectors increase.

When year 0 and year T are not consecutive years but two points in time separated by 1 year or more, nonchaining decomposition simply uses the data of the two years in the analysis without considering the data of all the intervening years. In contrast, decomposition on a chaining basis involves yearly decomposition using the data of every two consecutive years starting from year 0 and ending at year T, with the decomposition results computed on a cumulative basis over time. Chaining decomposition has also been referred to as time-series decomposition and rolling base year decomposition. In nonchaining decomposition, the year-to-year evolution of factors over time between year 0 and year T is not considered. It has been shown that there is less variation between the results given by different decomposition techniques when decomposition is performed on a chaining basis. This consistency, which arises because decomposition is "path dependent," is a distinct advantage. As such, the chaining procedure should be preferred when time-series data are available. However, nonchaining decomposition is the only choice in some situations, such as in studies that involve a very large number of sectors for which energy and production data are not collected annually.

3.5 Further Remarks

The methodological development of energy decomposition analysis has also been driven by the need of energy researchers to study an increasingly wider range of problems. In the simplest form and with only two factors, many decomposition techniques have a strong affinity with index numbers (i.e., product mix and energy intensity as equivalent to commodity quantity and price on an aggregate commodity value in economics). However, the problems studied after 1990 seldom have two factors. A typical decomposition problem on energy-related national carbon dioxide (CO_2) emissions, for instance, involves five factors—total energy demand per unit of gross domestic product (GDP), per capita GDP, population, energy demand by fuel type, and the CO_2 emission coefficient by fuel type.

There is a good collection of decomposition techniques that can cater to almost every possible decomposition situation or suit the varied needs of the analyst. Each of these techniques has its strengths and weaknesses. Perfect decomposition has been widely accepted as a very desirable property of a decomposition technique in energy analysis. In some situations, especially in cross-country decomposition, some of the techniques that do not give perfect decomposition could have a residual term many times larger than the estimated main effects. In addition, there are techniques that fail when the data set contains zero values or

negative values. Some techniques, such as those based on the Divisia index, have the advantage of ease of formulation and use, whereas others, including those based on the Laspeyres index, have complex formulae when more than three factors are considered.

In general, whether a specific technique is applicable or the "best" among all depends on the following: (i) the properties of the technique, which are related to the index number with which the technique is associated with; (ii) ease of understanding and use; (iii) the problem being analyzed, such as whether it is a temporal or cross-country problem; and (iv) the data structure, such as whether zero or negative values exist in the data set. A good understanding of the strengths and weaknesses of the various decomposition techniques would allow analysts to make sound judgments as to which technique to use. These issues are addressed in some publications.

4. APPLICATION AND RELATED ISSUES

The origin of energy decomposition analysis is closely linked to the study of energy use in industry. Not surprisingly, this application area accounted for most of the publications on the subject until 1990. Since then, the application of decomposition analysis has expanded greatly in scope. The main reasons are the growing concern about the environment and the increased emphasis on sustainable development globally. Moreover, decomposition analysis can be conveniently adopted as long as structural change can be meaningfully defined and data are available at the chosen disaggregation level. Indeed, its simplicity and flexibility opens up many possible application areas. Some of these, such as the development of national energy efficiency indicators to monitor national energy efficiency trends, are unique to index decomposition analysis as a decomposition methodology. In terms of application by country, decomposition analysis has been adopted to study energy and energy-related issues in a large number of countries, including all major industrialized countries, most Eastern European countries including Russia, and many developing countries. The following sections focus on developments in several key application areas.

4.1 Energy

With the successful application to study trends of energy use in industry, extensions to decomposition analysis were made to study energy use in the economy as a whole, in energy-consuming sectors other than industry, and in the energy supply sector. In the case of national energy demand, the aggregate to be decomposed is generally energy use per unit of GDP, commonly known as the energy-output ratio in the energy literature, where energy use covers all sectors of the economy and structural change refers to changes in GDP product mix. However, such studies often have difficulty matching energy consumption with economic output. This is due to the fact that energy uses in some sectors, such as private transportation and residential, have no direct match with output sectors/activities in the GDP, and as such some adjustments to sector definition and data are often needed.

Regarding energy consumption in major sectors other than industry, studies have been performed on energy use in passenger transportation and in freight transportation. For passenger transportation, aggregate energy use per passenger-kilometer has been decomposed to give the impacts of transportation modal mix and modal energy intensity given by modal energy requirement per passenger-kilometer. Similarly, for freight transportation, aggregate energy use per tonne-kilometer of freight moved has been decomposed. For residential energy use, decomposition has been conducted to quantify the impact of changes in the mix of residential energy services consumed. For electricity production, aggregate generation efficiency has been decomposed to give the impacts of generation mix and the generation efficiency of each fuel type.

4.2 Energy-Related Gas Emissions

Since the early 1990s, due to the growing concern about global warming and air pollution, an increasing number of studies on the decomposition of energy-induced emissions of CO_2 have been performed. The extension of the methodology presented in Section 2 to decompose an aggregate CO_2 indicator is fairly straightforward. Assume that the ratio of industrial energy-related CO_2 emissions to industrial output is the aggregate of interest. A declining trend for the indicator is generally preferred, and the motivation is to study the underlying factors contributing to changes in the aggregate. This would normally be formulated as a four-factor problem with the following factors: industrial output structure change, sectoral energy intensity change, sectoral fuel share change, and fuel CO_2 emission coefficient change. Similar decomposition procedures have been applied to study energy-related CO_2

emissions for the economy as a whole, other major energy-consuming sectors, and the energy supply sector. Although the focus among researchers and analysts has been on CO_2 emissions, some studies have dealt with emissions of other gases, such as sulfur dioxide and nitrogen oxides. The number of decomposition studies dealing with energy-related gas emissions equaled that of studies on energy demand by the late 1990s and exceeded the latter in 2000.

4.3 Material Flows and Dematerialization

Another extension of decomposition analysis is the study of the use of materials in an economy. The materials of interest to researchers include a wide range of metals and nonmetallic minerals, timber, paper, water, cotton, etc. The list also includes oil, coal, and natural gas, which are treated as materials rather than energy sources. The aggregate of interest would be the consumption of a material divided by GDP or some other meaningful output measure. In material flows studies, dematerialization refers to the absolute or relative reduction in the quantity of material to serve economic functions, and its concept is similar to reduction of aggregate energy intensity. The technique of energy decomposition analysis can be easily extended to material decomposition analysis to study trends of material use. Indeed, a number of such studies have been performed, and it has been found that decomposition analysis is a useful means of analyzing the development of the materials basis of an economy.

4.4 National Energy Efficiency Trend Monitoring

Increasingly more countries see the need to have appropriate national energy efficiency indicators or indices to measure national energy efficiency trends and progress toward national energy efficiency targets. The energy efficiency indicators may be defined at the national level or at the sectoral level of energy consumption, such as industrial, transportation, commercial, and residential. Here, the primary focus of decomposition analysis is not to study the impact of structural change. Rather, it is the elimination of the impacts of structural change and other factors from changes in an aggregate such as the ratio of national energy consumption to GDP, whereby the impact of energy intensity change computed is sufficiently refined to be taken as inversely proportional to real energy efficiency change. Advances in the decomposition analysis methodology, including the use of physical indicators to measure output or activity level, have allowed the development and use of more meaningful and robust energy efficiency indicators or indices. A related methodological framework that would more accurately track trends in energy efficiency has been developed by many countries and international organizations, including the Energy Efficiency and Conservation Authority of New Zealand, the U.S. Department of Energy, the Office of Energy Efficiency of Canada, the International Energy Agency, and the European Project and Database on Energy Efficiency Indicators.

4.5 Cross-Country Comparisons

Cross-country comparisons involve the quantification of factors contributing to differences in an aggregate, such as aggregate energy intensity or the ratio of energy-related CO_2 emissions to GDP, between two countries. The factors considered are similar to those in the study of changes in the indicator over time in a country. In decomposition analysis, the data for the two countries are taken to be the same as those for two different years in a single-country study. Such studies have been carried out for a number of countries. Studies have also compared CO_2 emissions between world regions, such as between North America and the developing world. For the factors considered, because the variations tend to be larger between two countries/regions than between two years for a country, the decomposition tends to give a large residual term if nonperfect decomposition techniques are applied. In general, the use of a perfect decomposition technique is recommended for cross-country decomposition analysis.

5. CONCLUSION

The methodology development and application of index decomposition analysis have been driven by energy researchers since the 1970s, immediately after the 1973 and 1974 world oil crisis. In the case of the basic two-factor decomposition problems, the tool has a strong affinity with index number problems. Nevertheless, over the years, energy researchers and analysts have continuously refined and extended the methodology for the unique situations and requirements in the energy field. As a result, decomposition analysis has reached a reasonable level of sophistication and its usefulness in energy studies has been firmly established.

A well-defined set of decomposition approaches/methods is now available to accommodate an arbitrary number of factors to be decomposed. Decomposition can be conducted multiplicatively or additively, on a chaining or nonchaining basis, and on a temporal or cross-sectional basis. Due to its simplicity and flexibility, the index decomposition analysis has been adopted in a wide range of application areas. The number of application studies has increased rapidly over time, particularly since the early 1990s, and interest among energy researchers in the methodology has been growing.

The previously mentioned developments were driven by the concern about world energy supply that began in the 1970s and were reinforced by concern about global warming and the need to reduce the growth or cut emissions of CO_2 and other greenhouse gases since the late 1980s. Because energy and the environment will remain important global issues and sustainable development will become increasingly more important in the following years, it is expected that index decomposition analysis will continue to play a useful role in providing information for policymakers to address energy, environmental, and resource depletion problems that mankind will continue to face.

SEE ALSO THE FOLLOWING ARTICLES

Aggregation of Energy • Commercial Sector and Energy Use • Industrial Ecology • Industrial Energy Efficiency • Industrial Energy Use, Status and Trends • International Comparisons of Energy End Use: Benefits and Risks • Modeling Energy Supply and Demand: A Comparison of Approaches • National Energy Modeling Systems

Further Reading

Ang, B. W., and Choi, K. H. (1997). Decomposition of aggregate energy and gas emission intensities for industry: a refined Divisia index method. *Energy J.* **18**(3), 59–73.

Ang, B. W., and Zhang, F. Q. (2000). A survey of index decomposition analysis in energy and environmental studies. *Energy* **25,** 1149–1176.

Ang, B.W., Liu, F.L.,Chew, E. P. (2003). Perfect decomposition techniques in energy and environmental analysis. *Energy Policy* **31,** 1561–1566.

Boyd, G., McDonald, J. F., Ross, M., and Hanson, D. A. (1987). Separating the changing composition of U.S. manufacturing production from energy efficiency improvements: a Divisia index approach. *Energy J.* **8**(2), 77–96.

Farla, J. C. M., and Blok, K. (2000). Energy efficiency and structural change in The Netherlands. 1980–1995. *J. Ind. Ecol.* **4**(1), 93–117.

Jenne, J., and Cattell, R. (1983). Structural change and energy efficiency in industry. *Energy Econ.* **5,** 114–123.

Lermit, J., and Jollands, N. (2001). "Monitoring Energy Efficiency Performance in New Zealand: A Conceptual and Methodological Framework." National Energy Efficiency and Conservation Authority, Wellington, New Zealand.

Liu, X. Q., Ang, B. W., and Ong, H. L. (1992). The application of the Divisia index to the decomposition of changes in industrial energy consumption. *Energy J.* **13**(4), 161–177.

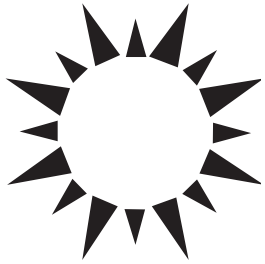
Marlay, R. (1984). Trends in industrial use of energy. *Science* **226,** 1277–1283.

Schipper, L., Unander, F., Murtishaw, S., and Mike, T. (2001). Indicators of energy use and carbon emissions: Explaining the energy economy link. *Annu. Rev. Energy Environ.* **26,** 49–81.

Sun, J. W. (1998). Changes in energy consumption and energy intensity: a complete decomposition model. *Energy Econ.* **20,** 85–100.

Wade, S. H. (2002). "Measuring Changes in Energy Efficiency for the Annual Energy Outlook 2002." U.S. Department of Energy, Energy Information Administration, Washington, DC.

Depletion and Valuation of Energy Resources



JOHN HARTWICK
Queen's University
Kingston, Ontario, Canada

1. **Depletion Pressure**
2. **Depletion Economics**
3. **The Scarcity of Clean Energy Supplies**
4. **Summary**

Glossary

backstop A substitute supply that is expected to become active when current supplies are exhausted. This is taken as a "renewable" supply available indefinitely at a constant unit cost. The "price of the backstop" would be this "high" unit cost. Commercially viable fusion reactors have been taken as the backstop energy source.

hotelling rent The rent per unit extracted that exists because of the finiteness (exhaustibility) of the stock being extracted from.

marginal cost of extraction Given an equilibrium, the cost of extracting one more unit

Nordhaus (Herfindahl) model A linked sequence of "Hotelling models," each with a higher unit cost of extraction. The demand schedule is unchanging in the Herfindahl version.

rent The positive gap between market price per unit and marginal cost for supplying an additional unit.

Ricardian rent The rent per unit that exists because of a quality advantage of the unit in question.

royalty A payment to the owner of a deposit by the extractor, who in effect is renting or borrowing the deposit.

In the early years of the 21st century, industrialized nations exist in a world of cheap energy. Direct outlays by households for electricity, household heating, and fuel for transportation occupy small fractions of almost all family budgets. Canada and the United States, with 5% of the world's population, consume 25% of global energy. In per capita terms, the United States consumes 354 million Btu/year, Western Europe consumes 170 million Btu/year, and India consumes 12 million Btu/year. A Btu is the amount of heat required to raise the temperature of 1 lb of water 1°F. Energy use in North America grew 31% between 1972 and 1997. Is an era of high energy costs just around the corner? Is there a signal of a higher priced era to be read out there today? In the period following oil price hikes in 1973–1974, the U.S. Department of Energy predicted $150 a barrel for oil in the year 2000 in year 2000 prices. Exxon predicted a price of $100. What has abundant energy accomplished for us? Since approximately 1800, the four most revolutionary changes in a person's life in industrialized countries are (1) the substitution of machine power for animal and manpower, (2) the more than doubling of life expectancy, (3) the arrival of relatively low-cost transportation for persons and commodities, and (4) the arrival of global and low-cost communication. Clearly (1) and (3) are outcomes of new forms of energy use. The dramatic increase in longevity is probably mostly a consequence of improvements in the processing of water, food, and wastes—that is, improvements in the public health milieu, broadly conceived rather than a result of low-cost power, directly. But the impact of reduced demands on humans and animals for power on a person's life expectancy should not be dismissed. Improved, low-cost communication is related to the arrival of electricity and electronics, not power per se. The words labor, from Latin, and work, from Greek, both capture the idea of an expenditure of painful physical or mental effort, of energy expenditure. People who do no labor or work then are those expending little energy. Revolutionary in human history is the substitution of machine power for human and animal power. Hoists (pulleys) and levers

 © 2004 Elsevier Inc. All rights reserved.

were early simple machines that magnified human and animal power considerably. Watermills and windmills arrived and showed what could be accomplished with new energy sources captured by machines and new types of power output—namely, steady and continuous rather than episodic. But it was the arrival of steam engines (heat machines) that revolutionized the concept of usable power. The concept of transporting goods and people changed dramatically and powering a factory was freed from ties to watermills and windmills.

1. DEPLETION PRESSURE

It is easy to make the case that depletion pressures induced invention and innovation in energy use. Since at least the time of Great Britain's Elizabeth I, say 1600, conservation of forests for future fuel and timber supply was prominent on the public agenda. In the early 1700s, English iron production was using up approximately 1100 km^2 of forest annually. The notion that forest depletion drove the development of British coal fields is a popular one. But causality is tricky here — did the arrival of cheap coal lead to a liquidation of remaining forests or did the depletion of forests induce innovation in coal supply? This really has no ready answer. Fuel scarcity apparently led to the importation of iron by England in the mid-18th century. Around 1775, coal became viable in smelting iron and England's importing ceased. Coinciding with the coalification of the British economy was the development of the steam engine, first as a water pump and later as a source of power in factories, such as cotton mills. Jevons (1865, p. 113) opined that "there can be no doubt that an urgent need was felt at the beginning of the 17th century for a more powerful means of draining mines." Savery's steam pump was operating in 1698 and Smeaton made Newcomen's engine, with a piston, sufficiently efficient to be commercially viable. In 1775, the new steam engine of James Watt was developed.

Given the documented concern in Great Britain about timber supply, say, before 1700, one should observe a rise in the price of wood for fuel, house-building, and ship-building at some point. The year 1700 seems like a reasonable year to select as having recorded an increase in wood price. Firewood prices grew three times faster than the general price level in England between 1500 and 1630 and the 17th century was one of "energy crisis." Over one-half of the total tonnage entering British ports in the 1750s was timber, and fir imports grew 700% from 1752 to 1792. France also experienced rapid increases in fuel wood prices over this latter period. The prevailing "high" price in 1700 should, of course, spur the search for substitute energy sources. Importation is one form of substitution and coal development would be another. Given flooding problems with new coal deposits, an investment in reasonable-quality pumps was a low-risk strategy. The deposits were known to exist and the demand for coal seemed certain. The Saverys and Newcomens rose to the occasion. This is a possible explanation for the invention and introduction of steam pumps after 1700. It shifts the emphasis from spontaneous invention to profit-driven invention and innovation. A wood supply crisis is the ultimate impetus in this scenario. Current or anticipated fuel scarcity raises fuel costs and induces a search for substitutes. Coal is the obvious substitute but mine-draining looms up as a serious constraint on coal production—hence, the pressure to invent and to make use of steam pumps. Cheap coal transformed the British economy. In 1969, Landes estimated coal consumption in Great Britain at 11 million tons in 1800, 22 million tons in 1830, and 100 million tons in 1870. The British appear to have been sensitive to energy use over the long run. The Domesday Book, published in 1086, records over 5000 water-wheels in use in England—one mill for every 50 households.

Attractive in this scenario is the low-risk nature of investment in pumping activity. The coal was there, the demand was there, and one simply had to keep the water level in the mines low. Recall that a Newcomen engine was part of a multistory structure and thus represented a serious outlay. However, the technology had been proven in small-scale engines and the minerals were there for the taking. The second watershed moment in the history of steam power, associated with James Watt, was its development for powering whole factories. It can be argued that it was the depletion of good watermill sites, well documented for the Birmingham region by Pelham in 1963, that spurred invention. Richard Arkwright's invention of the cotton-spinning machine and mill and James Watt's invention of the improved steam engine compete in the literature on the industrial revolution for the label, "the key events." Spinning mills required a power source and water-wheels were the standard at the time. But convenient sites for water-wheels had mostly been taken up by the 1770s. The price of those sites would have exhibited a significant rise in, say, 1775, the year Watt was

granted his special patent for his engine with a separate condenser, an engine vastly more efficient than the Newcomen type. The climate was right for invention in the power supply field. Depletion of good water-wheel sites drove up their prices and the resulting umbrella of profitability made invention in this area attractive. At approximately this time, serious research of water-wheels was undertaken by Smeaton with a view, of course, to obtain more power from the waterflow. Obviously, a profit-driven view of events is not the only plausible one, but there is no doubt that profits accrued to Watt and his partner, Boulton.

Coal and oil compete as bulk fuels in many parts of the world for powering electricity-generating plants but steam engines for motive power have been displaced by gasoline, diesel, and electric engines in most areas of the world. Steam locomotives were common in industrialized countries as late as the 1950s. This is the classic case of a low-cost, clean energy source (e.g., a diesel engine for a locomotive) being substituted for a dirtier and, presumably, higher cost source. The conspicuous aspect of energy use is the capital goods that come with the energy flow. For example, wind is the energy source and the mill transforms the energy flow into usable power in, for example, grinding wheat. Similarly, the rushing stream is the energy source and the wheel and mount are the capital goods that allow for the production of usable power. What Savery, Newcomen, Watt, and others did was to invent a capital good that turned a heat flow into usable power. The role of the inventor is to envisage a capital good, a machine, that allows for the transformation of an energy flow into a power flow. From this perspective, it is difficult to conceive of an energy supply crisis. One looks forward to larger and better machines harnessing old energy flows better and new flows well. An oil supply crisis in the early years of the 21st century would translate into a transportation cost crisis but not an end-of-civilization crisis. Approximately one-fifth of electricity in the United States was generated with petroleum in 1970. In 2002, that figure is one one-hundredth. Coal burning produces approximately one-third more carbon dioxide than oil burning and approximately twice as much as the burning of natural gas, per unit of heat. Transportation fuel can be synthesized from coal and methane, but only at relatively high costs, given today's technology. The Germans used the Fischer-Tropsch process to produce gasoline from coal during World War II and the search for a catalyst to produce gasoline from methane is under way.

The appropriate aggregate production function to consider for an economy takes the form of output flow, $Q = F[N, K, E(K^E, R)]$, where $E(\cdot)$ is power derived from an energy flow R, say, oil, and capital equipment, K^E, K is other capital, and N is labor services. Before 1800, K^E was windmills and water-wheels for the most part and R was wind and water flow. In the 21st century, food is cooked in and on electric stoves rather than in hearths, light is provided by electric units rather than candles or lanterns, and heat is provided by furnaces and distribution devices rather than by fires, fireplaces, and stoves. $E(K^E, R)$ is dramatically different even though basic houses are similar in concept and shape and eating styles and habits are not hugely different. What would the world be like with steam power but no electricity? There was an interval of approximately 75 years during which this state existed. Steam power was decisive in liberating people from much heavy labor. Steam power could drive hoists, diggers, and pumps. It could also power tractors, locomotives, and even automobiles. Factories could be well mechanized with steam engines. In 1969, Landes suggested that the 4 million hp of installed steam engines in Britain in 1870 was equivalent to 80 million men, allowing for rest, or 12 million horses. A worker with steam power was like a boss with, say, 16 healthy workers at his bidding. In 1994, Smil estimated that a worker can put out 800 kJ net per hour. From this, he inferred that the Roman empire's 85,000 km of trunk roads required 20,000 workers, full time, working 600 years for construction and maintenance. The gain in bulk power production from the use of electric motors rather than steam motors seems marginal when history is viewed on a large time-scale. Electricity rather created new activities such as listening to the radio, watching television, and talking on the telephone. These are large gains in the range of consumer goods, essentially new attractive commodities, but they are on a different order from innovations relieving humans of great and steady physical exertion. Electric stoves and washers have undeniably made housework easier, and air-conditioning (room cooling) has the lessened the discomfort of labor for many as well as altered the geographical distribution of settlement.

With regard to personal travel, steam power might have been revolutionary enough. Steamships were quite different from sailing ships and trains were quite different from stagecoaches. Steam-powered automobiles and trucks are technically feasible. The internal combustion engine, drawing on gasoline and

diesel fuel, is not a crucial technology, though steam power for vehicles was quickly displaced. Steam power revolutionized the concept of daily work and of personal travel. Huge numbers of people ceased to be beasts of burden. Huge numbers of people were able to visit relatively distant places quite rapidly. Diets changed significantly when new foods could be transported from distant growing places. Crucial to modernization, however, was the substitution of machine power for human and animal power, which steam engines effected, and the spread of reasonable living standards to low-income families. Steam power brought about this latter effect only indirectly, by making mass production techniques widespread. Mechanization brought down the prices of many goods (e.g., cotton) but displaced workers in the handicraft, labor-intensive sectors. Wages rose in Great Britain some 50 years after the industrial revolution was well rooted. Machine-intensive production relieves workers of much grinding physical exertion but usually eliminates many traditional jobs as well. A large virtue of market-oriented (capitalistic) development is that entrepreneurs create and often invent new jobs for the workers displaced by mechanization. Wages and living standards have risen, particularly with a decline in family size.

2. DEPLETION ECONOMICS

Though in 1914 Gray first formalized the idea that impending natural stock (say, oil) scarcity would show up in current price as a rent, a "surplus" above extraction cost, Hotelling presented the idea most clearly in 1931. A royalty charged by an owner of a deposit to a user of the deposit is, in fact, "Hotelling rent." Thus, the idea of "depletion" being a component of current mineral price is ancient. Competing, however, with the Gray–Hotelling theory is a theory that has current oil price rising up the market demand schedule because each subsequent ton is becoming more expensive to extract. Deeper is more costly. This approach was adopted by the great W. S. Jevons in his monograph on coal use in Great Britain: "…the growing difficulties of management and extraction of coal in a very deep mine must greatly enhance its price. It is by this rise of price that gradual exhaustion will be manifested …" (Jevons, 1865, p. 8) Jevons' son, also a professor of economics, reflected on his father's work, pointing out that the price of coal had indeed risen over the 50 years since his father's book was written, by, however, just 12% relative to an index of prices. And reflecting on his father's analogy of high energy prices being like a tax, Jevons recommended accumulating alternative forms of capital to compensate future generations for their shrunken coal capital. A standard modern view has the current rent in oil price, a reflection of the profit or rent derivable from currently discovered new deposits. Hotelling theory sees current oil price as a reflection of (1) how long current stocks will carry society until users are forced to switch to substitute sources and (2) what price those substitute sources will "come in at." It was then the value of remaining stocks that held current price up, rather than pressure from currently steadily worsening qualities. As Nordhaus has made clear, in Hotelling's theory of depletion or "resource rent," current price turns crucially on the cost of supply from the future substitute source (the backstop) and on the interval that current stock will continue to provide for demand. See Fig. 1 for an illustration of Hotelling theory.

r is the interest rate in the market and Hotelling's approach is often referred to as the r percent theory of resource rent because profit-maximizing extraction by competitive suppliers implies that rent rises over time at r percent. Hence, the quantity sequence in an equilibrium extraction program is the one displaying rent rising at r percent per period. Current rent (price minus current unit extraction cost) turns out to be simply discounted terminal rent, with this latter term being defined by the unit cost of energy from the backstop or substitute source. To say that the price of oil is high today could mean that remaining stock is not abundant and substitutes will

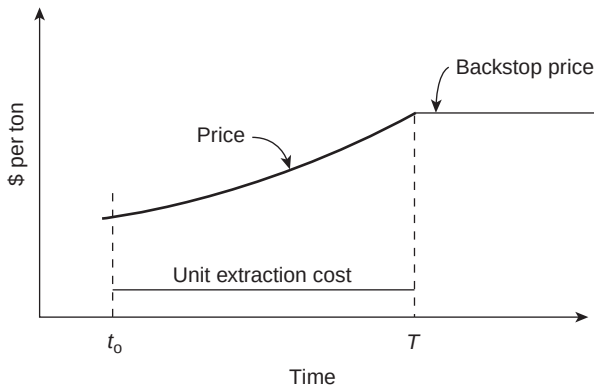


FIGURE 1 Hotelling model. In this competitive case, rent (price less unit extraction cost) rises r% until the stock is exhausted, at date T. Price "fed into" a market demand schedule yields current quantity supplied at each date. Unit cost, backstop price, stock size, and the demand schedule are parameters. Interval (t_0, T) is endogenous.

be available at high prices (Hotelling theory) or that future supplies will come from lower quality, high-cost deposits (Jevons theory). These two theories have been brought together in one model in a simple version by Herfindahl (see Fig. 2) and in a more complicated version in 1977 by Levhari and Leviatan and many others. The Herfindahl version has scarcity or depletion rent rising at r percent and, although unit extraction costs are constant, then jumping down as the next deposit or source is switched to. For the case of many deposits, Fig. 2 depicts a combination of intervals with rising rent linked by jumps down in rent. Quality decline causes rent to jump down. The Hotelling or depletion effect moves counter to the quality-decline effect. Thus, even in this simple world of no uncertainty or oligopoly in supply, one cannot be accurate in saying that exhaustibility is showing up in current price movement as an increasing rent or depletion effect (Fig. 2).

Hotelling provided a systematic theory of how depletion would show up in current price and how finiteness of the stock would show up in the pace of stock drawdown. In the simplest terms, exhaustibility is represented by a wedge between current unit cost and price. Thus, if world oil stocks were dwindling rapidly, one would expect oil prices to be high, well above unit extraction costs. And in fact, episodes of high world oil prices have been observed in recent decades, times when average world oil prices were well above unit extraction costs. Such wedges are labeled "rents" in standard economics and, with something like oil, the wedge might be referred to as scarcity rent or Hotelling rent. The world demand schedule does not shift in these episodes but current extracted output does shift back

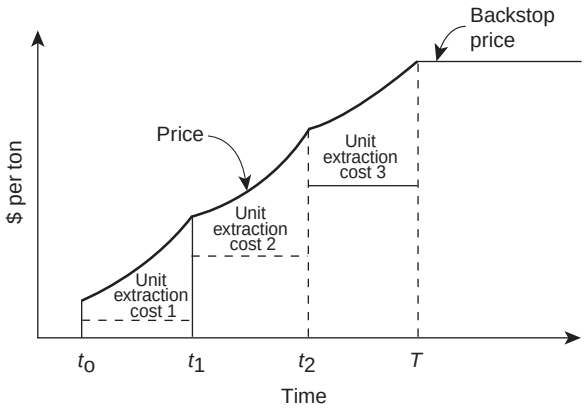


FIGURE 2 Nordhaus (Herfindahl) model. A three-source, linked Hotelling model. Low-cost supplies are exhausted first. Rent rises r% except at source switch. Rent jumps down across source switch. Intervals of source use are endogenous.

for some reason. It is these fairly sudden contractions in current production that cause current prices to jump up and "create" extra rent. The first large jump was in the fall of 1973. Another occurred in 1979, with the collapse of the Palhevi monarchy in Iran, and a large drop in price occurred over mid-1985 to mid-1986. Here the argument is that discipline among a cartel of oil suppliers broke down and world supply increased fairly quickly.

Price changes are largely driven by changes in current production flow or supply and current production is a complicated function of parameters, including estimated remaining stock or cumulative supply. Hotelling theory is an intertemporal theory of discounted profit maximization by competitive producers that links current production to remaining stock, extraction costs, demand conditions, and backstop price. It makes precise the exhaustibility effect or depletion in current resource price. Hotelling theory obtains a "depletion effect" in current price relatively simply in part because it makes use of the traditional assumption of perfect foresight by profit-maximizing agents. Economic theory relies on this assumption because it becomes an essential component of intertemporal profit maximization. The real world is, of course, replete with events that are very difficult to anticipate, such as wars, government regime switches, and technical changes. For example, research suggests that oil may be generated deep in the earth by unexpected geothermal processes. Uncertainty makes simple calculation of a depletion effect in current price a dubious course.

In its original form, Hotelling theory abstracted from uncertainty completely. And Hotelling did not take up quality variation in current deposits or stocks. This has, fact, led many observers to argue that Hotelling theory is fundamentally flawed as an explanation of rent or price paths for, say, oil. Also, Hotelling did not address oligopoly in supply and such market structures have been considered to be crucial in explaining oil price movements in recent decades. In the face of those well-known difficulties in arriving at a theory of oil price determination, in 1985 Miller and Upton tested "Hotelling's valuation principle" for a sample of oil-producing companies whose shares were being traded on the New York Stock Exchange. Employing heroic simplifying assumptions, they derived an equation to estimate in a sequence of straightforward steps and fitted their equation to the data, with positive results. It seems undeniable that extraction to maximize discounted profit is the reasonable maintained hypothesis and in a sense Miller and Upton confirm this position. But

profit maximization by an extractive firm in an oligopoly setting, for instance, yields a quite different extraction program than that for a firm in a competitive setting. Miller and Upton provide weak evidence in favor of competitive extractive behavior by oil extraction companies. Implicit in their work is an argument that Hotelling rent is positive and present in the price of oil. However, the plausible "crowding" of such rent by stock quality decline cannot be evaluated from their analysis. In 1990, Adelman argued that current oil rent reflects the marginal cost of discovering new supply and most rent earned by producers is a Ricardian rent, one attributable to a quality advantage possessed by current suppliers lucky enough to be currently working a low-cost deposit.

Consider the reality of a backstop source of energy. Suppose the world does have a 20-year supply of "cheap" oil available, as Deffeyes suggested in 2001. (The Edison Electric Institute estimates that proven reserves of oil will last 37 years, those of natural gas will last 61 years, and those of coal will last 211 years.) In 2001, Kaufmann and Cleveland presented a careful, empirically based critique of Hubbert's bell-shaped supply schedule. Why is a foreseeable energy supply crisis not showing up in markets today? The answer appears to be that backstop supplies are fairly abundant, in the form of large deposits of coal, natural gas, tar sands oil, shale oil, and uranium. None of these supplies is as attractive and as low-cost as oil but it would not be complicated to power a modern economy with a combination of these energy feedstocks. Directly difficult would be fuel for transportation, given current engine technology and driving habits. Whatever transpires with long-term oil supplies, the era of low-cost individualistic transportation will likely end. The century 1920–2020 may well go down in history as the curious low-cost, automobile period, an era of individuals dashing about the surface of continents in big cocoons of steel on expensive roadways, while seriously polluting the atmosphere.

In 1974, Nordhaus brought his combination of Hotelling and backstop economics together in an ingenious empirical exercise. Given a backstop supply price and a sequence of world demands for energy into the distant future, and given world stocks of oil, coal, and uranium and unit extraction costs, he let his computer solve for the Hotelling transition path from his base period, 1973, to the era of constant cost fusion power. Naturally, stocks with lower unit extraction costs provide energy early on. See Fig. 2. Thus, one observes an oil supply era of endogenous length, a coal era, a uranium era, and then the fusion power era with constant unit cost into the indefinite future. A critic can easily dismiss this investigation as computer-aided, crystal-ball gazing. However, the Nordhaus model was subjected to sensitivity testing at low cost, and the empirical output appeared to be fairly robust. Central to this is, of course, discounting. Events beyond 30 years register little today, given discounting. Hence, fusion power costs largely fade out in the current price of energy, as do many other parameters of the model. It remains, nonetheless, a very good framework for organizing one's thinking about the future of energy. When the oil runs out, modern life does not end, just the curious era of abundant, low-cost energy, especially for personal transportation.

Nordhaus's "simulation" yielded a current price (1973) of energy somewhat below the observed market price. He inferred that this occurred because his model was "competitive," *a priori*, whereas the real world contained much oligopoly in energy production—hence, an obvious need to refine the computer formulation. However, the introduction of oligopoly to extraction models is tricky, if only for reasons of incorporating the dynamic consistency of agents' strategies. There must be no obvious possibilities for profit-taking simply by reneging on one's "original" strategy selected at time zero. To rule out such possibilities for profits, problems must be analyzed and solved backward, from tail to head, and this turns out to be much more complicated than, for example, Nordhaus's essentially competitive formulation. There is an important lesson here. If one believes that current energy prices reflect noncompetitive behavior among suppliers, which seems most reasonable, then quantifying such capitalizations is known to be difficult. One can rule out simple explanations for current energy price levels. The combination of Hotelling theory and backstop theory indicates that traditional supply and demand analysis is an inadequate way to think about current energy price formation. The addition of oligopoly considerations complicates an already complicated process of current price determination.

In the simplest case of current price shift, as for, say, October 1973, suppliers revise estimates of remaining stock downward and current price jumps up. "Mediating" this price shift is a process of agents discounting back from "the terminal date" to the present, discounting along a Hotelling extraction path. Current quantity extracted jumps down and current price jumps up. If demand is inelastic in the region, aggregate revenue will jump up. Thus, the

motivation to raise current price could be for reasons of revenue. With some coordination of quantities in, say, a cartel, the current price jump may be an oligopoly or cartel phenomenon rather than an outcome of stock size revision. Somewhat analogous is a revision by suppliers of the anticipated backstop supply price. A revision of expectations upward would lead to a current price jump upward and vice versa. A price drop occurred briefly when University of Utah researchers announced "cold fusion." And the popular interpretation of the large drop in world oil prices during 1985–1986 is that cartel discipline broke down Organization of Petroleum Exporting Countries (OPEC) and key members raised their outputs above their quota levels.

It is undeniable that the current interaction of the world supply curve for oil and the world demand curve yield the current price. At issue is how depletion rent and oligopoly rent are capitalized in the current costs of production, costs that determine the characteristics of the effective supply schedule. In the mid-1970s, reasonable observers debated whether the 1973 jump in world oil price was the result of production restraint by OPEC, an oligopoly effect, or a reckoning that future oil stocks would be rather smaller than had been anticipated. In any case, current production was cut back and world prices leapt up. A popular view of the poor performance of the U.S. economy in the later 1970s was that it had to accommodate unanticipated supply-side shocks in the energy sector. The notion of cost-push inflation became popular and an extensive literature emerged on this issue.

Since impending exhaustibility is predicted to manifest itself in currently high price, which is reflecting a large scarcity rent, a standard view is that the current and anticipated high prices should induce a search for substitute supply. How tight is this link between, say, rising current oil price and a spurt of invention to develop an alternative supply? It was argued above that this link is relevant for explaining two of the most significant events in human history—the effect of timber scarcity in Great Britain on the timing and intensity of the invention of mine-draining machinery (the steam engine) and the effect of watermill site scarcity on the timing and intensity of invention in steam engine efficiency (Watt's separate condenser). It may be the anticipated high price rather than the current actual high price per se that induces inventors to perform, to come up with substitute supply. Simply anticipating that, say, oil will be exhausted soon while current price is not yet high may be sufficient to get invention going. Inventors can have foresight. They need a process or invention that brings in a competitive supply of energy when the oil runs out. In 2003, Mann argued that recent low oil prices have slowed innovation in the alternative energy sector.

Interesting models in which the substitute supply is, in fact, invented or perfected today in the public sector in order to force current oil suppliers to lower current prices have been developed. The invention is annouced and mothballed until the oil is exhausted. In such a scenario, it is not high prices today that speed up invention, rather it is the anticipation of the high prices in the future that induces inventors to perfect the substitute today.

3. THE SCARCITY OF CLEAN ENERGY SUPPLIES

Attention has shifted from energy sources with low cost, per se, to sources that can provide energy cleanly, at low costs. Smog in Los Angeles seems to have spurred policymakers to move toward designs for a future with cleaner air and with low-emission vehicles, homes, and factories. Eighty-five percent of world energy supply is currently derived from fossil fuels and 1% is renewable (excluding hydropower). The recent DICE model of global warming of Nordhaus and Boyer has unlimited energy produced at constant unit extraction cost from hydrocarbons but energy production contributes directly to an atmospheric temperature increase, via carbon dioxide emissions. The temperature increase shows up in explicit current economic damage. This is a large departure from the Hotelling view because Nordhaus and Boyer have no stock size limitation for, say, oil. Price, inclusive of an atmospheric degradation charge (a Pigovian tax on energy use), rises along the industry demand schedules because, essentially, of the increasing scarcity of moderate atmospheric temperatures. It is "moderate temperature" that is being "depleted" in this view and a "depletion charge" in the form of a marginal damage term (a carbon tax) is required for full cost pricing of energy. A world without carbon taxes is, of course, one of "low" energy prices, one in which the atmosphere is a carbon sink, with carbon dioxide dumped in at zero charge. User fees for dumping? "If U.S. consumers pay $1 billion a year, would that be enough to cover the problem? How about $10 billion? The level of economic damage that might be inflicted by greenhouse gas abatement is so

uncertain that even the Kyoto treaty on global warming … says not a word about what the "right" level of emissions should be." (Mann, 2002, p. 38) In 2002, Sarmiento and Gruber reported that atmospheric carbon dioxide levels have increased by more than 30% since the industrial revolution (the mid-18th century).

A formal statement of the world energy industry along Nordhaus–Boyer lines looks much like the Hotelling variant of Levhari and Leviatan amended to incorporate a continuous quality decline of the oil stock. Aggregate benefits (the area under the industry demand schedule) from current energy use, q, are $B(q)$. Temperature interval, T, above "normal," increases with q in $dT/dt = \alpha q$. Energy users (everyone) suffer current economic damage, $D(T)$. Thus, the only way to stop the full cost price of energy from rising is to switch to a source that does not drive up global temperatures. Here, the current price is $[dB(q)]/(dt)$ and the carbon tax in price is the discounted sum of marginal damages from temperature increments. (A constant unit extraction cost changes the formulation very little.) This price needs the right Pigovian tax on carbon emissions, a dumping fee, to be the full market price for energy.

Will the current, dirty, energy-generating technologies be invented around? Would they be invented around if current energy prices were higher? Mann thinks so. How abundant will fuel cells, solar generation installations, "pebble bed" nuclear reactors, windmill farms, synthetic fuels, and more sophisticated electric power transmission grids be 20 years from now? Are carbon taxes and attendant high current energy prices the best prescription for a cleaner energy future? Many policymakers have not espoused this approach, particularly in the United States and Canada.

4. SUMMARY

Depletion theory is the attempt to explain price formation for commodities such as oil, those with current flows being drawn from finite stocks. Price rises above unit extraction cost because of a depletion effect and one analyzes the nature of this effect in order to try to predict future price paths. Until recently, attention was focused on how the progressive scarcity of an oil stock was translating into the current high price for oil in use. In the past decade, attention has shifted from how the limitation on future supply is showing up in, say, oil price to how the effects of pollution from fossil fuel burning should be incorporated into current price. The depletion effect on the clean environment becomes a component (a pollution tax) in current full price. The hope is that full pricing of energy will signal to users how best to use energy currently and will signal to suppliers the potential profitability of innovating in the energy supply process.

SEE ALSO THE FOLLOWING ARTICLES

Coal Industry, History of • Economics of Energy Supply • Innovation and Energy Prices • Modeling Energy Supply and Demand: A Comparison of Approaches • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Petroleum Property Valuation • Prices of Energy, History of • Transitions in Energy Use • Value Theory and Energy • Wood Energy, History of

Further Reading

Adelman, M. A. (1990). Mineral depletion with special reference to petroleum. *Rev. Econ. Statist* **72,** 1–10.

Deffeyes, K. (2001). "The View from Hubbert's Peak." Princeton University Press, Princeton NJ.

Gray, L. C. (1914). Rent under the assumption of exhaustibility. *Q. J. Econ.* **28,** 466–489.

Hotelling, H. (1931). The economics of exhaustible resources. *J. Polit. Econ.* **39,** 135–179.

Jevons, H. S. (1915, reprint 1969). "The British Coal Trade." A. M. Kelley, New York.

Jevons, W. S. (1865, reprint 1968). "The Coal Question: An Inquiry Concerning the Progress of the Nation and the Probable Exhaustion of Our Coal-mines," 3rd ed. A. M. Kelley, New York.

Kaufmann, R. K., and Cleveland, C. J. (2001). Oil production in the lower 48 states: Economic, geological, and institutional determinants. *Energy J.* **22,** 27–49.

Landes, D. S. (1969). "The Unbound Prometheus." Cambridge University Press, Cambridge, MA.

Levhari, D., and Leviatan, D. (1977). Notes on Hotelling's economics of exhaustible resources. *Can. J. Econ.* **10,** 177–192.

Mann, C. C. (2002). Getting over oil. *Technol. Rev.* **January/February,** 33–38.

Miller, M., and Upton, C. (1985). A test of the Hotelling valuation principle. *J. Polit. Econ* **93,** 1–25.

Nordhaus, W. (1974). The allocation of energy resources. *Brookings Papers Econ. Activ.* **3,** 529–570.

Nordhaus, W. D., and Boyer, J. (2002). "Warming the World: Economic Models of Global Warming." MIT Press, Cambridge MA.

Pelham, R. A. (1963). The water-power crisis in Birmingham in the eighteenth century. *Univ. Birmingham Hist. J.*, 23–57.

Pomeranz, K. (2002). "The Great Divergence: Europe, China and the Making of the Modern World Economy." Princeton University Press, Princeton NJ.

Rae, J., and Volti, R. (2001). "The Engineer in History," Revised ed. Peter Lang, New York.
Sarmiento, J. L., and Gruber, N. (2002). Sinks for anthropogenic carbon. *Phys. Today* **55**, 30–36.
Smil, V. (1994). "Energy in World History." Westview Press, Boulder, CO.
Thomas, B. (1980). Towards an energy interpretation of the industrial revolution. *Atlantic Econ. J.* **3**, 1–15.
United Nations. (2002). "North America's Environment: A Thirty Year State of the Environment and Policy Retrospective." United Nations, New York.

Derivatives, Energy

VINCENT KAMINSKI
Reliant Resources, Inc.
Houston, Texas, United States

1. **Classification of Energy Derivatives**
2. **Forwards and Futures**
3. **Energy Commodity Swaps**
4. **Options**
5. **Main Energy Derivatives Markets**
6. **Options Embedded in Bilateral Contracts and Physical Assets (Real Options)**
7. **Valuation of Energy-Related Options**

Glossary

derivative The term used in finance to describe a financial product with value dependent on the value of another instrument, commodity, or product.

forward Forward contract is a transaction that provides for delivery of the underlying commodity at a future date, at the price determined at the inception of the contract, with cash being paid at or after an agreed number of days following the delivery.

NYMEX Commodity exchange located in New York City, offers a number of listed energy contracts.

option A contract that gives the holder the right, but not the obligation, to buy (sell) a financial instrument or commodity (the underlying) at a predetermined price (a strike or exercise price) by or on certain day.

real options In the energy markets, flexibilities or rigidities embedded in bilateral transactions or physical assets that can be valued using option pricing technology developed for the financial markets instruments.

swap An exchange of streams of cash flows over time between two counterparties. The level of the cash flows depends on prices of financial instruments or commodities referenced in the swap agreement. In a typical commodity swap, one side pays a fixed price; the second side pays a floating (varying) price that corresponds to the market price of the underlying commodity.

volatility The most important input to the option pricing models. Volatility is typically defined as the annualized standard deviation of price returns.

A derivative is a financial instrument with a value based on, or derived from, the value of another stock, bond, or a commodity (the underlying). In the case of energy derivatives, the underlying is an energy commodity, energy-related derivative or a physical asset used to produce or distribute energy.

1. CLASSIFICATION OF ENERGY DERIVATIVES

Energy derivatives can be classified based on (1) the contract definition specifying the rules by which the instrument derives its value from the price of another traded instrument, (2) the market or platform on which they are traded, or (3) the underlying commodity or commodities. Most energy derivatives can be classified as forward (futures) contracts or options. Both types of instruments can be negotiated as physical contracts that require actual delivery of the underlying commodity as financial contracts that are cash settled, based on the prices of the underlying observed in the marketplace or published by an industry newsletter or a specialized business.

Energy derivatives are traded and valued as stand-alone contracts or may embedded in bilateral transactions and physical assets. Such options existed for a long time and only recently they have become a subject of rigorous analysis directed at their valuation and risk management of the exposures they create.

2. FORWARDS AND FUTURES

A forward contract is a bilateral transaction that provides for delivery of the underlying commodity at a future date, at the price determined at the inception of the contract, with cash being paid at or after an agreed number of days following the delivery. In many markets, the fair price of a forward contract

 © 2004 Elsevier Inc. All rights reserved.

can be derived from the spot commodity price, interest rate, and the storage cost, using the so-called arbitrage principle. Specifically, the forward price (F) is equal to the spot price (S) adjusted for the financing and storage cost (r – annualized interest rate, s – unit storage cost), or $F = S^*(1 + r + s)^{(T-t)}$ where $T-t$ represents the life of the contract (t is today's date, T is the delivery date, and both are measured in years). The logic behind this equation is based on the arbitrage argument that defines the upper and lower bounds on the forward price; one can prove that both bounds are equal giving the unique forward price. The buyer of the forward contract can lock in the economic outcome of the transaction by borrowing funds at rate r, buying the commodity at spot (S) and storing the commodity and taking a short position in the forward contract (selling). At contract expiration, the transaction is reversed: the commodity is delivered into the forward contract at price F and the financing and storage costs are recovered. The arbitrageurs will engage in such transactions, as long as $F > S^*(1 + r + s)^{(T-t)}$. Alternatively, the lower bound can be established by recognizing that one can borrow the commodity, sell it in the spot market, and invest the funds at rate r, taking a long position in the forward contract. This argument breaks down often in the case of commodity-related forward contracts for two reasons. Some energy commodities may be unstorable (like electricity) or the supply of storage may be limited at times. In some cases, the existing market framework does not allow for shorting of the underlying physical commodity, eliminating one of the arbitrage bounds.

Given that both the storage cost and interest rate are positive, the arbitrage argument implies that the forward price is always greater than the spot price and that the forward price curve (a collection of forward prices for different contract horizons) is an increasing function of time to expiration (the curve is always in contango). In reality, forward price curves for many commodities, including natural gas and crude, are often backwardated (the forward prices in front of the curve exceed the prices in the back). This arises from the fact that for many market participants holding physical commodity has a higher value than controlling a forward contract. The stream of benefits that accrue to the owner of physical commodity is called a convenience yield. Assuming that the convenience yield can be stated per unit of commodity, δ, the full cost-of-carry formula for an energy forward contract is given by the following equation:

$$F = S^*(1 + r + s - \delta)^{(T-t)}.$$

Futures are forward contracts traded on an organized exchange that provides credit guarantees of counterparty performance, provides trading infrastructure (typically an open outcry system or screen trading), and provides the definition of a standardized contract. Credit protection is provided by margining and by the capital contributed by the exchange members and ultimately their credit. At inception, each counterparty posts the so-called initial margin that is followed by the so-called maintenance margin deposited if the original margin is eroded by losses on the contract. The practice of marking to market the contract and providing (receiving) additional margin is the most important difference between the futures and forward contracts. The most important exchanges offering energy contracts (crude, natural gas, heating oil, gasoline, and electricity) are the New York Mercantile Exchange (NYMEX), International Petroleum Exchange (IPE) in London, Nordpool in Norway. The Intercontinental Exchange offers many energy-related contracts but should be treated more as a trading platform, not as an organized exchange.

Under some fairly restrictive conditions, the futures and forwards have the same value. Potential differences arise from the practice of margining the contract that results in stochastic financing cost (one has to borrow/invest the funds paid/received) and from the differences in default risk. An organized exchange offers more layers of credit protection than bilateral arrangements. In practice, the differences between futures and forwards have been found to be rather small in most empirical studies. The energy industry typically ignores this distinction in routine transactions.

The distinction between the forward and futures contracts has become more ambiguous in many markets. The practice of requiring and posting periodically collateral to provide guarantee of performance under a bilateral forward contract, which has become a standard practice in the United States and many other countries, has obliterated one major difference between forwards and futures. Classical forwards, unlike futures, did not trigger cash flows before contract maturity.

3. ENERGY COMMODITY SWAPS

Swaps can be looked at as a basket of forwards that has to be executed as one contract (no cherry picking is allowed). In the general sense, a swap is the exchange of two streams of cash flows: one of them is based on a fixed (predetermined) price and the

second on a floating (variable) price. At inception, the fixed price is determined by finding the fixed price that makes the fixed and floating legs of the swap equal. The fixed leg is equal to the sum of the periodically exchanged volumes of the underlying commodity multiplied by the fixed price, discounted back to the starting date. The floating leg is equal to the sum of the volumes multiplied by the floating price, discounted back to the starting date. The floating price is obtained from the forward price curve, a collection of forward prices for contracts with different maturities. In other words, the fixed price of the swap is derived from the following equality:

$$\sum_{t=1}^{n} \frac{F_f \times v_t}{(1+r_t)^t} = \sum_{t=1}^{n} \frac{F_t \times v_t}{(1+r_t)^t}, \qquad (1)$$

where F_f denotes the fixed price of the swap, F_t – the floating price, r the applicable discount rate, and v_t – periodic deliverable volumes.

In practice, a swap may be physical or financial. A sale of an energy commodity at a fixed price, with deliveries distributed over time, is referred to and priced as a swap. A swap may be a purely financial transaction, with one counterparty sending periodically to another the check for the fixed-floating price difference (multiplied by the underlying volume), with the direction of the cash flow dependent on the sign of the difference.

4. OPTIONS

An option is a contract that gives the holder the right, but not the obligation, to buy (sell) a financial instrument or commodity (the underlying) at a predetermined price (a strike or exercise price, usually denoted by K), by or on certain day. The option to buy is referred to as a call, the option to sell as a put. An option that can be exercised only at maturity is known as a European option, an option that can be exercised at any point during its entire lifetime is known as an American option. In the case of an energy-related option, the underlying is an energy commodity or another energy derivative (a forward, a swap, a futures contract, or even another option). Some popular energy options are defined in terms of a number of several different commodity prices (have multiple underlying commodities or instruments).

Most energy options traded in bilateral transactions have a forward contract as the underlying and this has important implications for the valuation procedures. Through the rest of this article, we are assuming that the underlying is a forward or a futures contract. Most options transacted in the energy markets are plain vanilla European calls and puts. Certain features of energy markets created the need for more specialized structures, addressing specialized needs of commercial hedgers, with complicated payoff definitions. Most popular options with complex payoff definitions and many practical applications are Asian and spread options.

Asian options are defined in terms of an average price. The payoff of an Asian option at maturity is defined as max (avg (F)-K, 0) for a call option and max (K-avg (F), 0) for a put option, where F stands for the forward price of the underlying and K for the strike price. The term avg (F) corresponds to the average price calculated over a predefined time window. The rationale for Asian options arises in most energy markets from the fear of price manipulation on or near option expiration date, given relatively low trading volumes and the relative ease to affect the price through targeted buying or selling. Averaging mitigates this risk, reducing the impact of any single day price on the option payoff. In many cases, the exposures facing commercial energy hedgers are defined in terms of average prices. For example, an electric utility buying natural gas as a fuel is exposed to an average cost over time, given its ability to pass on its average costs to the end users. Asian options tend to be cheaper than corresponding European options, as price averaging reduces price volatility, an important input to the option pricing models (discussed later).

Spread options are defined in terms of a price difference. The payoff of a call spread option is defined as max (F_1-F_2-K, 0) and the payoff of the put as max (K-F_1-F_2, 0). The price difference may correspond to the location spread (for example, the difference between Brent crude produced in the North Sea and West Texas Intermediate, WTI) or time difference (calendar spread between price of natural gas corresponding to December natural gas and July natural gas contracts with delivery at Henry Hub location in Louisiana). Quality spread corresponds to the price difference of different grades of the same commodity (for example, sweet and sour crude oil). The rationale for spread options arises from the exposures that many commercial operations face in the energy markets: exposures to price differentials as opposed to exposures to a single price level. A refinery is exposed to the spread between the composite price of the basket of refined products and the price of crude. A power plant is exposed to the

difference between price of electricity and the price of fuel. An oil company buying gasoline in Europe and shipping it to the United States faces combined location and calendar spread risks when the cargo is in transit.

The importance of price differential related exposures created specialized terminology describing different types of spreads. The difference between the price of heating oil and the price of crude is referred to as a heat spread; the difference between the price of gasoline and crude in referred to as the gas crack spread. The difference between natural gas prices at one of many trading hubs in the United States and the price at Henry Hub (the delivery point for the NYMEX natural gas futures contracts) is routinely referred to as *basis* in the U.S. natural gas markets (in more general sense, basis denotes price difference). The difference between price of electricity and the price of natural gas is known as the spark spread. In the case of the spark spread, the price of natural gas, denominated typically in $/MMBTU, has to be converted into $/MWh. This is accomplished in practice by multiplying natural gas price by the heat rate (MMBTU/MWh) that corresponds to the thermal efficiency of a specific (or notional) power plant (the energy input required to produce one MWh of electricity).

Spread options are used to value many transactions in the energy markets. One example is the contract for transmission, known as a firm or fixed transmission right (FTR). The FTR protects a provider of electricity against basis risk. This type of exposure happens when a supplier offers delivery at one point but owns generation at another point in a power pool. If no transmission capacity is available between these two points (i.e., congestion develops on the transmission grid), the prices at two locations may decouple and the supplier may be unable to deliver into his contractual obligation. The supplier would have to buy electricity at the point where his or her load is located, running the risk that he will incur a loss. To hedge against this risk he may buy FTRs offered by many power pools through special auctions. An FTR contract may be structured as a swap or an option. A swap pays the difference between the two locational prices and allows lock-in the economic outcome of the original power supply transaction, irrespective of the potential congestion. An FTR contract that is structured like an option may be priced as a spread option.

A special type of energy related option was invented to offer protection against combination of volumetric and price risks. Many energy producers and buyers (especially utilities and local distribution companies) are exposed to both uncertain prices and quantities. This risk is addressed through the so-called swing option that has a number of unique characteristics. It is typically structured as a forward start option: the strike price is set equal to the market price of the underlying on some future date. In the U.S. natural gas markets, this price is typically equal to the first of the month price for base load deliveries over the entire calendar month, published by one of the industry newsletters (like, for example, Inside FERC Natural Gas Market Report). Some swing options use a strike price that is fixed in a bilateral contract. The holder of the option is entitled to buy the commodity at the strike price n times over the defined period (typically one calendar month) within certain volumetric limits, referred to as a minimum and maximum daily quantity. The option definition may be complicated further by ratchets that allow the holder to reduce or increase the purchased volumes in certain increments. The contract may provide also for minimum and maximum takes over certain time periods (for example, months, quarters, and years) and for penalties if the periodic bounds are exceeded. The swing options are popular in the U.S. natural gas and electricity markets as a form of protection against a jump in prices combined with an increase in consumption requirements. Pricing such options is quite complicated due to the need to keep track of both types of optionality (with respect to price and volume).

Basket options are defined as contracts on a composite commodity like, for example, a package of natural gas and crude oil. The price of the underlying in this case is a (weighted) sum of prices of commodities included in the basket. Such options are typically used by producers who have exposure to multiple commodity prices and want to hedge the overall exposure, as the price fluctuations of the basket components may be mutually offsetting, providing a natural hedge. The value of a basket option, like the value of a spread option, is heavily influenced by the correlations between basket components.

5. MAIN ENERGY DERIVATIVES MARKETS

An alternative classification of energy derivatives is based on the market at which they are transacted. The most transparent and liquid markets are those

provided by the organized or electronic exchanges. An organized exchange may be based on the principle of open outcry or may be based on screen-based trading. The most important exchanges in the United States are the New York Mercantile Exchange (NYMEX) and the Intercontinental Exchange. The latter provides an electronic trading platform and other services such as netting and settlements (with London Clearing House and Merchant Exchange). NYMEX offers both futures and options on futures (options that have a futures contract as the underlying and are exercised into a futures contract that has to be taken into delivery or closed in a separate transaction). NYMEX trading is based on the principle of the open outcry but electronic trading is available as well for certain contracts or after regular trading hours. Table I provides the summary of the contract available on NYMEX (only contracts with significant volumes are listed, as opposed to all listed contracts that may have no significant activity).

Other exchanges important in the energy markets are International Petroleum Exchange of London (futures and options on futures for Brent oil and gas oil, futures contracts for natural gas), Powernext in Paris (day-ahead electricity contracts), European Energy Exchange (EEX) in Leipzig, Germany (electricity), and Nordpool in Scandinavia (electricity).

The exchange-based transactions are supplemented by the over-the-counter (OTC) market that provides the benefits of greater flexibility of contract design and allows for customization to fit exactly the needs of the buyer. OTC derivatives lack the credit support provided by an exchange and their pricing is more opaque, given that in many cases they can be quite complicated. The most important forms of the OTC energy contracts are swaps and options.

6. OPTIONS EMBEDDED IN BILATERAL CONTRACTS AND PHYSICAL ASSETS (REAL OPTIONS)

One very important class of energy related derivatives are options embedded in energy contracts. Unlike exchange traded and OTC derivatives, they are not traded as stand alone contracts but are packaged in bilateral contracts. They typically correspond to different flexibilities and rigidities embedded in the transactions that correspond to practical needs of doing business. In principle, every flexibility granted to, or accorded by, a counterparty in terms of quantity, delivery timing, and location or quality can be translated into an option, swaption, or another derivative. The challenge of pricing options embedded in the contracts is that one is dealing with many layers of interacting optionality, and it is often difficult to separate different options (they may not be additive) and price them individually using commercially available software. It is, however, important to price such options to avoid giving them away for free or at reduced prices and to manage the exposures they create over the life of the contract.

One example of a complicated bilateral contract that may extend over many years is so-called full requirements contract. Under this contract, the energy provider takes over the resources of the customer (for example, generation plants, storage, and contractual assets like transportation and transmission contracts of a municipal utility) and commits to satisfy all the energy needs of the buyer at an agreed price. Energy is delivered using the assets of the buyer, controlled and dispatched by the supplier over the duration of the contract, and the surplus (deficit) is sold (covered) in the market. Providers also have the option to shut down the generation assets they control in case

TABLE I

NYMEX Energy Futures

Contract	Size	Delivery point	Options	Settlement type
Natural gas	10,000 MMBTUs	Henry Hub, LA	Yes	Delivery, ADP, EFP
Light sweet crude oil	1000 barrels	Cushing, OK	Yes	Delivery, ADP, EFP
Heating oil	1000 barrels	New York Harbor	Yes	Delivery, ADP, EFP
Unleaded gasoline	1000 barrels	New York Harbor	Yes	Delivery, ADP, EFP
Brent oil	1000 barrels		Yes	Financial
Central Appalachian coal	1550 tons	Ohio River, Big Sandy River	No	Delivery, EFP
PJM electricity	40 MW on peak per day, 19 to 23 days a month		No	Financial

electricity may be acquired at a lower price from the market and sell in the market fuel purchased for the generation plants. In most cases, the provider assumes the volumetric risk associated with demand fluctuations (due to weather changes) and load growth due to economic and demographic changes, within certain predetermined bounds.

Derivatives embedded in physical assets such as power plants or natural gas storage facilities are the subject of the so-called real options, the term that describes the application of the technology developed for financial markets to the analysis of investment decisions. For example, a gas-fired natural gas power plant can be looked at as a strip of spark spread options. The underlying assumption is that the power plant is dispatched if the spark spread, as defined previously, exceeds variable marginal cost. This approach is very effective but should be used with caution. Valuation of a power plant with a life extending over many years requires many inputs for which market information is not directly available. For example, one has to provide information about forward price curves for natural gas and electricity for as many as 20 years, assumptions regarding price volatility and correlations between price returns for both commodities.

In many cases, there are no markets extending that far into the future, and this means that one has to substitute forecasts for what is supposed to be market derived inputs. A more serious shortcoming of this method is that a power plants operations are subject to many physical and regulatory constraints that are difficult to accommodate in the option pricing model. For example, one has to recognize constraints related to planned and forced outages, startup and shutdown costs, emission constraints that limit the total number of hours that a plant may operate in a year, and the restriction on the number of starts (due to provisions in maintenance and service contracts). Incorporation of such restrictions increases the dimensionality of the problem and greatly increases the computational resources required to come up with realistic valuation.

In spite of the shortcomings, the real options approach represents a much better alternative to the standard discounted cash flow approach, traditionally used in making investment decisions under uncertainty. The traditional discounted cash flow approach can only address uncertainty through comparisons of a limited number of scenarios that may not capture the entire range of possible future states of the world and may sometimes represent wishful thinking of the decision makers.

Other frequent uses of the real options technology are natural gas storage facilities and pipelines. A natural gas storage facility can be looked at as a portfolio of calendar spread options. A pipeline transportation contract can be valued as a portfolio of basis (locational spread) options.

The classification of options depending on the underlying commodity is quite straightforward. One possible classification is based on the physical commodity or commodities to which the option is related. One can talk about options on crude, natural gas, and so on. If the option is written on a specific energy-related instrument, the options are referred to as options on forward, futures, other options (compound options), and swaps (swaptions).

7. VALUATION OF ENERGY-RELATED OPTIONS

Valuation of energy-related options offers many challenges due to a number of complications in their definition and market behavior of the underlyings and is an area of very intensive research efforts in a number of academic centers and many energy companies. European options on forwards and futures are valued using the so-called Black formula that represents a version of the well known Black-Scholes formula for pricing options on stocks that don't pay dividends. The call price c is given by

$$c = e^{-r(T-t)}(FN(d_1) - KN(d_2)),$$

and the put p price is given by

$$p = e^{-r(T-t)}(KN(-d_2) - FN(-d_1)),$$

where

$$d_1 = \frac{\ln\left(\frac{F}{K}\right) + \frac{\sigma^2}{2}(T-t)}{\sigma\sqrt{T-t}},\ d_2 = d_1 - \sigma\sqrt{T-t}.$$

In these formulas, F stands for the forward price as of the time of valuation (t), K represents the strike price. The expiration date is given by T. Both t and T are measured in years (1 year is equal to 1, 1 month to 1/12, etc.) and therefore $T-t$ is equal to the remaining life of the option. The risk-free interest rate is given by r and the expression $\exp(-r(T-t))$ is a discount factor. The critical input to the formulas is σ, or the volatility coefficient. Volatility, unlike other inputs, is not directly observable and represents the most critical assumption in the option pricing process. In practice, it is based on the insights

obtained from historical estimates or is inferred from market prices of traded options. In the first case, volatility is estimated as the standard deviation of the logarithmic price returns (natural logs of the price ratios, defined as price P in period t divided by price P in period $t-1$). Volatility is quoted by convention as an annualized number. Annualization is accomplished by multiplying the standard deviation of price returns by the factor that depends on the price data frequency. In the case of monthly data, the factor is the square root of 12. In the case of weekly data, the annualization factor is the square root of 52. In the case of daily prices, one uses by convention the square root of 250 (or 260)—an approximate number of trading days in a year. In some energy markets, trading continues on Saturdays (Western power markets in the United States) or over the weekend. In such cases, one should use the appropriate number of trading days in a year.

For other energy options, there are typically no closed form valuation formulas and one must use different approximation formulas or numerical solutions. The most widely used approaches are Monte Carlo simulations and binomial (or more complicated) trees (see Table II). The Monte Carlo approach is based on repetitive simulations of price trajectories and calculating option payoffs at expiration. The payoffs are then averaged and discounted back to the current date. This approach is very flexible and allows for quick implementation. The shortcoming of this approach is that it requires significant computational resources and may be time intensive. It is more appropriate for path-dependent options (for example, Asian options) where the payoff depends not only on the terminal price but also on the trajectory along which this price was reached. In general, a Monte Carlo approach is not suitable for American options valuation; though some numerical techniques address this problem.

The binomial tree technique is well suited for pricing American options and may be described as a structured simulation. The technique is based on the assumption that a price at any point in time may evolve only into two other states (this explains the use of the term binomial), up or down from the starting level, with probabilities and the magnitudes

TABLE II

Option Valuation Using a Binomial Tree

Time	1	2	3	4	5
					89.06561
					0
				79.35276	
				0	
			70.69912		70.69912
			0		0
		62.98919		62.98919	
		0.786522		0	
	56.12005		56.12005		56.12005
	2.58979		1.499718		0
50		50		50	
5.22672		4.237387		2.859618	
	44.54736		44.54736		44.54736
	7.658791		6.743557		5.452637
		39.68935		39.68935	
		10.82827		10.31065	
			35.36112		35.36112
			14.63888		14.63888
				31.50489	
				18.49511	
					28.0692
					21.9308

of changes can be calibrated from the standard option pricing assumptions (including the volatility assumption). Specifically, probability of the up move is given by

$$p = \frac{1-d}{u-d},$$

and the probability of the down move is given by $q = 1-p$. The price level in the up move is given by uF, where F is the price level and the u multiplier is given by $u = e^{\sigma\sqrt{\Delta t}}$. The down move price level is given by dF, where d is given by $d = 1/u$. The term Δt corresponds to the length of the time step and is chosen by the modeler depending on the computation time and required precision requirements. The time step determines how dense is the tree (how many nodes it has for a given contract duration). Parameters u, d, and p allow for constructing the tree describing possible evolution of the price from the current level till maturity. Options can be priced through the process of backward induction by calculating the payoff of the option at terminal nodes (it is either zero or a difference between the terminal price and the strike price) and moving back in tree. At each node, (t, i), where t corresponds to the time step, i to the node level counted from the bottom of the tree, the value of the option is given by the comparison of the payoff of the option if it is immediately exercised with the expected value one obtains from keeping the option alive. This expected value is equal to the probability weighted values at two nodes connecting to the given node at time $t+1$, discounted back over the length of one time step.

The example that follows shows a tree developed for a put option with a life of 5 months using five time steps. This means that the length of one time step is 1/12 or 0.08333. The starting price is 50, the strike price K is 50, and the volatility is assumed to be 0.4 or 40%. The probability p of the up move at each node is equal to 0.471165. The risk free interest rate is 10% and this translates into a 1-month discount factor of 0.991701. For example, the value of the option at the node (3, 2), where the price level is equal to 44.554736, is equal to 6.743557. This value is calculated as follows: the immediate exercise of the option gives the option holder $50-44.54736 = 5.452$ (it's a put option). Keeping the option alive gives the holder 6.743557 that is calculated as probability weighted average of the option values at the nodes (4,3) and (4,2), discounted back one period. To be exact, $6.743557 = 0.991701 \times (0.471165 \times 2.859618 + (1-0.471165) \times 10.31065)$. Keeping the option alive is better than killing it (exercising it immediately) and therefore the last calculation gives the value of the option at the node (3,2).

The Black formula is based on the assumption that the price of the underlying commodity follows the stochastic process known as Geometric Brownian Motion. Under this process, the instantaneous price return (dF/F) is given by the following formula:

$$dF = \mu F dt + \sigma F dZ, \tag{2}$$

where μ stand for the so-called drift (instantaneous rate of return) and dz is the so-called Wiener variable ($dz = \varepsilon\sqrt{dt}$, where ε is drawn from the standard normal distribution). In case the underlying is a forward contract, the drift is equal to zero. This can be explained by the fact that entering into a forward contract has zero cost. A nonzero drift implies that the expected return is nonzero, and this implies that one would enter into a transaction with an expected loss or gain and no cost. In the first case, the market participants would bid down the price by shorting it; in the second case (gain) they would bid the price up till the potential gain disappeared.

The challenge of pricing energy derivatives is that the prices in many commodity markets do not follow this process. Energy commodity prices display behavior that is characterized by seasonality, the tendency to gap due to shocks that affect either the demand or supply side, and also they cannot be modeled without paying attention to production costs. The research in the area of energy option valuation evolves around development of new models that are based on alternative stochastic assumptions regarding price evolution. Two examples of such processes are a mean reversion process and a jump-diffusion process. Mean reversion process recognizes the fact that prices gravitate toward the levels determined by the cost of production. The simplest form of the mean reversion process is given by the so-called Ornstein-Uhlenbeck process:

$$dF = \alpha(m - F)dt + \sigma dz, \tag{3}$$

where α (>0) denotes the speed of mean reversion and the m denotes the mean price level toward which the price gravitates. If the price drops below the mean level, the difference $m-F$ is positive and given the speed reversion that is positive, the deterministic part of the Eq. (3) will be positive as well. This will help to bring the price back to the mean level. Of course, the stochastic part (σdz) may reinforce or counter the contribution of the deterministic part. If the price F is above the mean level, the mean reversion part of the process will act to bring it down. It is important to

notice that the volatility parameter has a different interpretation than the volatility parameter in the Geometric Brownian Motion process. The price change on the left-hand side of Eq. (3) is measured in dollars per commodity unit. The Wiener variable dz is dimensionless and, therefore, σ is measured in the same monetary units as dP. Volatility σ in the Eq. (2) is unitless.

The propensity of energy pieces to jump reflects one of the structural features of the energy markets. The price at any location is likely to jump (gap) due to sudden changes in supply or demand conditions that cannot be addressed due to the rigidities of the physical infrastructure (inflexible transmission or transportation grid, limited deliverability from and into storage). The modeling approach to this problem is a combination of a diffusion process (Geometric Brownian Motion or Ornstein-Uhlenbeck, for example) and a jump process. One example is the Poisson process combined with the Geometric Brownian motion:

$$dF = \mu F dt + \sigma F dz + (J - 1) F dq, \qquad (4)$$

where the variable dq assumes two possible values, 1 if a jump occurs, 0 if it does not. It is assumed that the jump occurs with the probability λdt, where dt is the time step and λ is known as the intensity of the process. J is the dimension of the jump. If the jump happens and J is equal to 0.8, then the forward price F drops immediately by 20%, in addition to changes resulting from the diffusion part of the Eq. (4). When a jump occurs, it is typically assumed to follow a log-normal distribution ($ln(J) \sim N(\gamma,\delta)$), where γ denotes the mean of the natural logarithm of the jump and δ its standard deviation. The jump diffusion process parameters may be estimated from the historical price return data and used in option valuation models based on numerical techniques. The jump diffusion approach, although very popular, suffers from many shortcomings. The parameter estimates may be unstable and it takes a long time for the results of the jump to taper off, contrary to many empirical observations that show a quick return to more normal price levels following a jump.

Mean reversion and jump diffusion are just two fairly elementary examples of the efforts under way to come up with better valuation techniques for energy options that recognize the unique features of energy markets.

SEE ALSO THE FOLLOWING ARTICLES

Business Cycles and Energy Prices • Energy Futures and Options • Energy Services Industry • Investment in Fossil Fuels Industries • Oil Price Volatility • Stock Markets and Energy Prices • Trade in Energy and Energy Services

Further Reading

Eydeland, A., and Wolyniec, K. (2002). "Energy and Power Risk Management: New Developments in Modeling, Pricing and Hedging." John Wiley, New York.

Fusaro, P. C. (1998). "Energy Risk Management: Hedging Strategies and Instruments for the International Energy Markets." McGraw-Hill, New York.

"Managing Energy Price Risk" (1999). Risk Books, London.

Pilipovic, D. (1997). "Energy Risk: Valuing and Managing Energy Derivatives." McGraw-Hill, New York.

Ronn, E. (ed.). (2002). "Real Options and Energy Management: Using Options Methodology to Enhance Capital Budgeting Decisions." Risk Books, London.

Desalination and Energy Use

JOHN B. TONNER and JODIE TONNER
Water Consultants International
Mequon, Wisconsin, United States

1. Introduction
2. Minimum Energy of Separation
3. Thermal Processes

Glossary

desalination The removal of salt, especially from seawater, to make it drinkable.
membrane desalination The removal of salt from seawater by using membranes to differentiate and selectively separate salts and water.
minimum energy of separation The thermodynamic minimum amount of energy needed to separate salts from water in the process of desalination.
thermal desalination The removal of salt from seawater by condensing purified vapor to yield a product in the form of distilled water.

This article focuses on processes involving the separation of solute and solution with the objective of reducing the solute concentration in the product stream. The focus of the following section is practical economic processes that have demonstrated sustained commercial viability.

1. INTRODUCTION

All desalination processes can be generically represented by the diagram shown in Fig. 1. Desalination processes can be segregated into two broad types, as follows:

1. Thermal processes, the most common being multistage flash (MSF); multiple-effect distillation (MED) and its variants; and vapor compression, which can be subclassified into thermal vapor compression (TVC)—most commonly combined with MED—and mechanical vapor compression;

2. Membrane processes, including reverse osmosis (RO); electrodialysis (ED); and electrodialysis reversal (EDR).

The thermal processes share the common step of condensing purified vapor to yield a product in the form of distilled water. The membrane processes have two further subclassifications with significantly different modes of operation: pressure-driven membranes and charge-selective membranes.

2. MINIMUM ENERGY OF SEPARATION

The minimum energy to separate a solute from solution can be derived from first principles. The most common approach is to consider the thermodynamic properties of pure water and a saline water, such as seawater. Two physically identical vessels, at the same temperature, containing fresh water and seawater, respectively, will have different pressures. This is due to the different vapor pressures of the two chambers, which differ because of the respective purity and salinity of the waters, and is shown diagrammatically in Fig. 2.

If the two vessels were to be connected, then the pressure difference that would exist would drive a flow in the direction of the arrow, from fresh water to saline water. A flow in this direction is not desirable for a desalination process. A day-to-day result of this situation is the reason that reverse osmosis is called reverse osmosis; some energy must be added to the system to overcome the natural direction of flow.

To achieve desalination, the pure water from the vapor space above the saline water (typically seawater) must be transported to the freshwater chamber (see Fig. 3). This can be achieved by a reversible process utilizing a hypothetical vapor compressor that can raise the pressure from the

 © 2004 Elsevier Inc. All rights reserved.

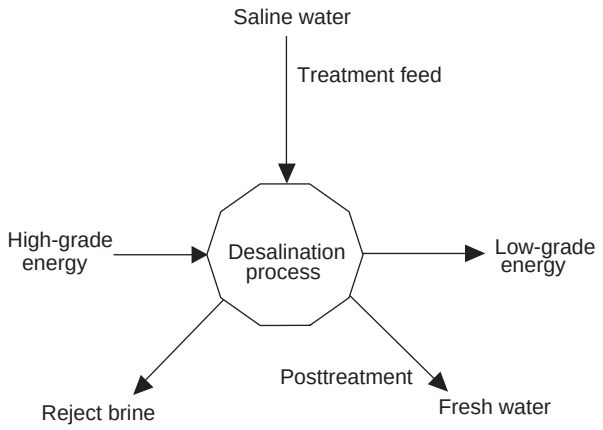


FIGURE 1 The desalination process.

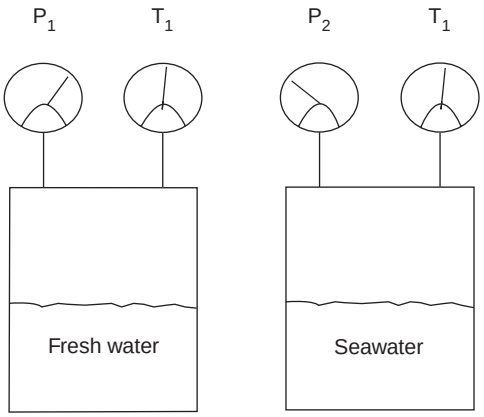


FIGURE 2 Two physically identical vessels, at the same temperature, containing fresh water and seawater, respectively, have different pressures.

freshwater vessel to a value just above that of the saline vessel. If the two vessels are held within a suitably large constant temperature reservoir, then it can be assumed that there is no temperature rise associated with this process.

Using this model, the minimum reversible energy consumption of a desalination process can be found by calculating the compressor power consumption that is a function of the vapor flow rate and partial pressure difference. This shows that the minimum energy of separation is a function of salinity (and also of temperature, since temperature relates directly to partial pressure). For seawater at 25°C, the vapor pressure is 3103 Pa and the specific volume is 44.17 m^3/kg; comparable figures for pure water are 3166 Pa and 43.40 m^3/kg. This yields a minimum energy consumption for the desalination of seawater of approximately 0.7 kWh per ton of water produced

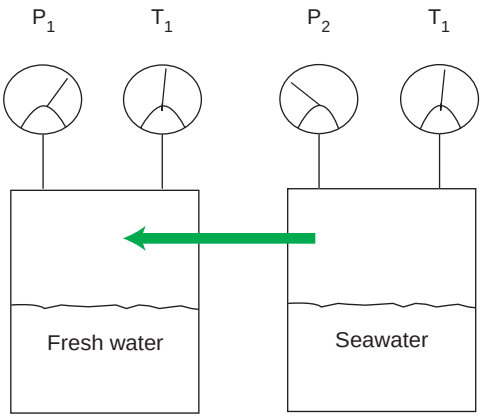


FIGURE 3 To achieve desalination, pure water from the vapor space above the seawater must be transported to the freshwater chamber.

(0.75 kWh/ton, or 2.734 kJ/kg). This minimum energy varies from 0.65 kWh/ton at 0°C to 0.90 kWh/ton at 100°C.

This analysis was considered for vapor compression processes but is valid for all processes. Why? The analysis is based on a theoretical ideal reversible process, which means that the results are valid as minima, which cannot be bested by practical processes. All practical desalination processes have these same values as targets although each of them may be restricted by different irreversibilities.

Each desalination process has a basic thermodynamic parameter that is a primary irreversible and unavoidable loss. For example,

- Distillation must always account for ΔT_b, the boiling point elevation.
- Freezing must similarly allow for ΔT_f, the freeze point suppression.
- Reverse osmosis must overcome ΔP_o, the osmotic pressure.
- Ion exchange must overcome ΔV_p, the polarization voltage, and provide ΔV_d, the diffusion voltage.

A detailed but less intuitively obviously minimum energy analysis could be completed considering each process; however, the net result would be the same. A more practical approach is to consider the state of the art for the most commonly utilized processes.

3. THERMAL PROCESSES

MSF, MED, and TVC processes are all driven by a high-enthalpy energy source, typically steam, but hot

water or oil can also be utilized. In general, all these processes incorporate a maximum operating temperature limited to between 70 and 115°C to avoid scale and corrosion associated with the solutions. This is known as the top brine temperature (TBT), which as a design parameter varies within the specified range due to slight differences in the configuration of the respective processes. This means that these processes can be heated by any energy source that is a few degrees warmer than the TBT; however, some processes utilize significantly higher enthalpy streams.

3.1 MED

MED is one of the oldest desalination methods and the simplest process with which to explain the principle of gained output ratio (GOR). GOR is a measure of how much thermal energy is consumed in a desalination process. The simplest definition is as follows:

How many kilograms of distilled water are produced per kilogram of steam consumed? Obviously, the same value for GOR is obtained when U.S. customary units of pounds are used.

Typically, the value of GOR ranges from 1 to 10 kg (H_2O)/kg (steam). Lower values are typical of applications where there is a high availability of low thermal energy. Higher values, even up to 18, have been associated with situations where local energy values are very high, when the local value or need for water is high or a combination of both.

A single-effect distiller, as shown in Fig. 4, operates with a GOR of 1 if there are no thermodynamic losses. In the absence of losses, each kilogram of steam that condenses in the effect will release sufficient heat of condensation to evaporate an equal quantity of vapor from the seawater, which then condenses in the condenser. This further assumes that the heat transfer area is infinite and there is a negligible temperature difference. Each kilogram of steam is therefore returned to the boiler as condensate and a kilogram of distilled water has been produced for other uses.

In the late 1800s, it was first noticed that this single-step process could be repeated with a gain in production, and therefore efficiency, for each effect added. This became known as Rillieux's First Principle: "In a multi-effect, one pound of steam applied to the apparatus will evaporate as many pounds of water as there are bodies in the set." This can be explained by examining the MED system depicted in Fig. 5. Once again, if losses are neglected, each kilogram of steam will release heat of condensation, which is equal to the heat of vaporization, and thus a four-effect system would produce 1 kg of condensate and 4 kg of distillate for every kilogram of heating steam utilized.

There are, however, many irreversibilities and other losses within these processes but the basic concept of Rillieux's First Principle is substantially correct and requires a 0.8–0.88 correction factor to be applied. This means that a four-effect MED system will typically yield a GOR of between 3.2 and 3.5.

MED systems have been designed and successfully operated with GORs ranging from 0.8 to approximately 10. Higher GOR designs invariably incorporate larger heat transfer areas, which then increases capital costs. The justification of higher capital costs can be made only with energy values that are appropriate or where greater demand for water suitably impacts the economics. The trade-off between capital cost (CAPEX) and operating cost (OPEX) is shown diagrammatically in Fig. 6, where it can be seen that a minimum will exist, which is project-specific, based on prevailing energy values and interest rates.

Typically, low-GOR applications include those where low-grade, low-value energy is abundant; examples include refineries and methanol facilities in arid locations where a traditional source of water is not available and a GOR of 3–4 is common. Higher GOR applications include production of distilled water for municipal supplies in areas with high-energy values such as the Caribbean Islands, where a GOR of closer to 10 is common.

Important additional points to note about the MED process are as follows:

- Each effect has an unavoidable BPE loss. Practical restrictions on TBT mean that utilizing too many effects in order to achieve high GORs is

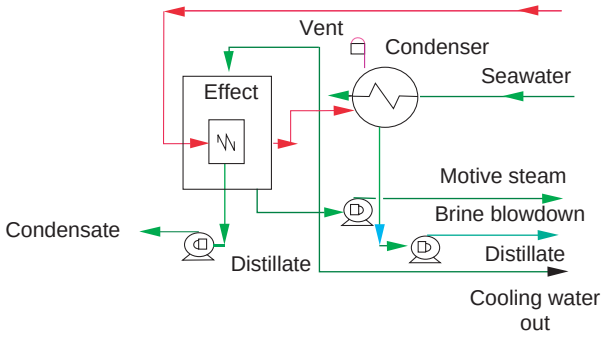


FIGURE 4 A single-effect distiller.

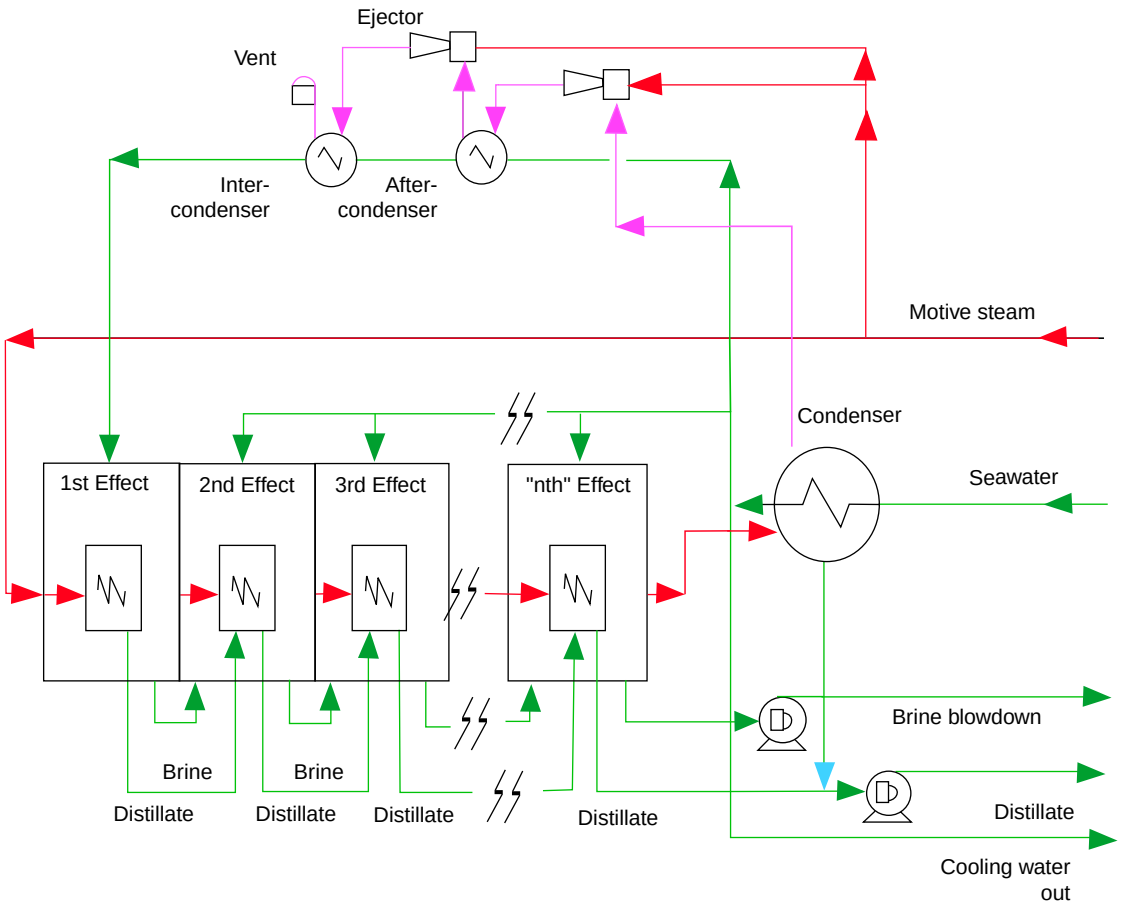


FIGURE 5 Multiple-effect distillation system.

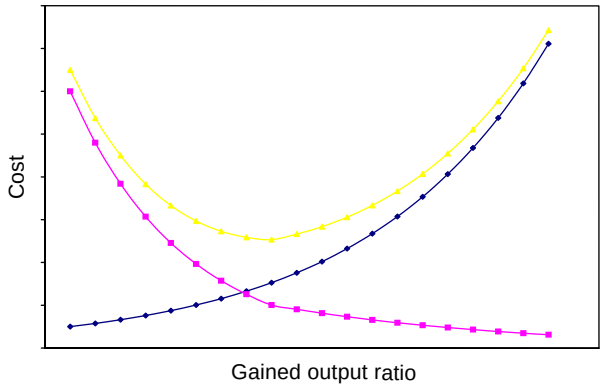


FIGURE 6 Trade-off between capital cost (CAPEX) and operating cost (OPEX). Diamonds, CAPEX; squares, OPEX; triangles, total cost.

impractical as the total BPE loss consumes a larger portion of the available ΔT.

- Pumping power is typically 0.7–1.2 kWh/ton. This includes the process pumps shown in the figures but excludes external pumps such as intake pumps. Excluding these pumps is usually valid since the cooling water that they provide is normally required even if the desalination plant is not operating as part of the facility heat balance (i.e., the steam must still be condensed). The intake pump flow rate is related to the GOR and the total dynamic head is related to site conditions, both of which also justify considering only the pumping energy within the process for comparative purposes.

3.2 MED with Thermal Vapor Compression

MED with thermal vapor compression (MED-TVC), or simply TVC, is the most rapidly growing distillation process. The process combines the basic MED system with a steam jet ejector, which enables the plant to utilize enthalpy and kinetic energy from motive steam and is shown in Fig. 7. Like the MED system, each kilogram of steam entering the first effect can yield a distillate production of $0.85N$ kg, where

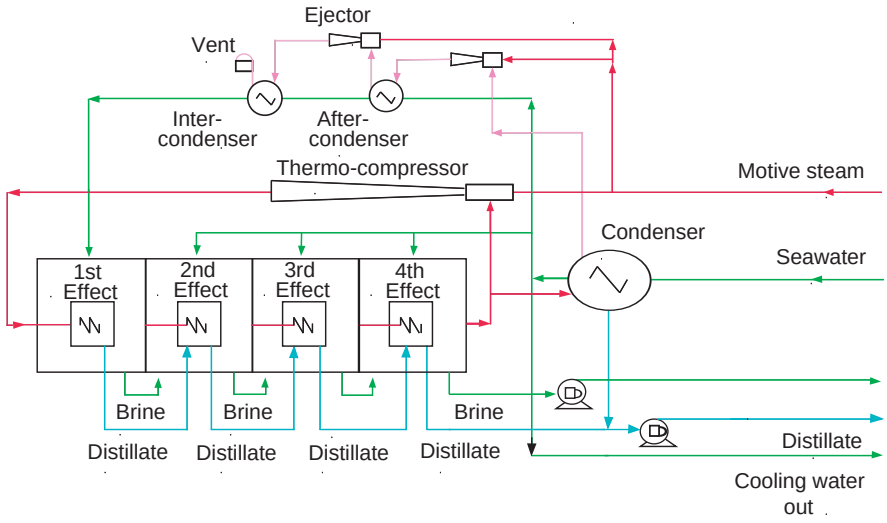


FIGURE 7 Multiple-effect distillation system with thermal vapor compression.

N is the number of effects. The important point to note is that each kilogram of steam entering the first effect is a combination of live steam and recycled low-pressure process vapor. The net result is that when utilizing medium-pressure steam, it is possible to achieve a GOR of over 8 with only four effects.

The addition of a thermocompression device adds approximately 1% to the CAPEX of a basic MED unit, yet the efficiency can be doubled. In cases where the steam is produced from waste gases, or where there is little commercial value difference between medium- and low-pressure steam, the use of TVC has grown tremendously. Using medium-pressure steam, it has been possible to achieve GORs of over 15. Such high GOR values also mean that less heat must be rejected from the system via cooling water, which can be an important consideration in many projects (for environmental or permit reasons).

The most common method for large-scale desalting of seawater had been the MSF process, which typically was combined with power production and utilized turbine extraction steam at 5 bar. Higher heat transfer coefficients, lower pumping pressures (and therefore lighter mechanical design requirements), and greater design flexibility mean that TVC designs are producing the same amount of water from this 5 bar steam but at lower CAPEX and OPEX than MSF. For distillation units in the range of 1000–25,000 ton/day capacity, the TVC process is beginning to dominate.

3.3 MSF

The MSF process is the most prevalent method of desalting seawater based on the volume of freshwater produced daily on a global basis. The process was commercialized during the 1960s based on competing Scottish and U.S. patented technology. MSF is a classic forced circulation process wherein seawater is preheated by the condensation of the vapor, which has been flashed from hot seawater. A once-through MSF process is shown in Fig. 8.

The relationship between the number of stages and the GOR is complex and also depends on the orientation of the tubes. For long tube designs where the tubes and flashing brine are always countercurrent, the number of stages, *N*, is typically 4 times the GOR. For cross-tube designs where there can be multiple tube passes within each stage and the flow through the tubes is perpendicular to the flashing brine, *N* is typically 2.5–3 times the GOR.

The once-through configuration is found mostly on marine distillation applications or similar applications that are relatively low capacity and low GOR. A typical plant of this design will have a GOR of 3 and a capacity of 1000 tons/day. Pumping power is typically 5–6 kWh per ton of distilled water. The U.S. nuclear carrier fleet utilizes this technology primarily to provide high-quality feedwater for the steam catapult launch-assist systems.

The most common MSF plants utilize cross-tube configurations in brine recycle flow schemes, shown in Fig. 9, and have unit production capacities of up to 55,000 tons/day. Many of these large systems are provided in the countries of the Gulf Cooperation Council, in Saudi Arabia, the location of the largest desalination facility at Al Jubail, and in the United Arab Emirates, the location of the largest units at Al Taweelah.

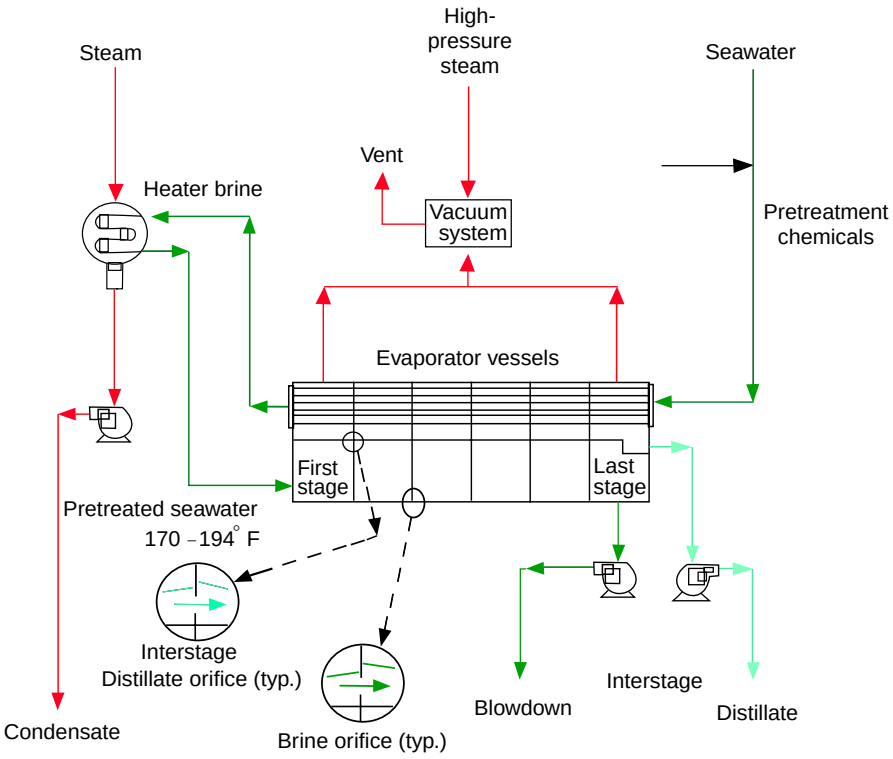


FIGURE 8 A once-through multistage flash process.

The typical GOR for these large units ranges from 7.7 to 9.0. All MSF plants are designed with electrically driven pumps but past practices included steam turbine drivers for the larger pumps. Exhaust steam from the turbines was then condensed in the brine heater, providing much of the MSF process heating. Such steam turbine drives have been replaced by electric motors not for efficiency reasons but for ease of use, simplicity, and operational flexibility. Typical pumping power for MSF units is 3.7–4.5 kWh per ton of distilled water.

3.4 RO

The RO process takes advantage of semipermeable properties of certain materials that allow solvents to pass through the material but impeding the passage of dissolved ions. The most commonly treated solvent is water. The semipermeable material can be configured in many ways but the most common is to roll flat sheets around a central pipe to create what is known as a membrane element. Even though membranes are available for performing very fine filtration, RO is not to be confused with filtration, which removes only material that is suspended within the fluid.

When a membrane is installed between saline water and fresh water, the natural flow will be toward the saline solution. The application of pressure to the saline side can stop this process, with the point of no flow being called the osmotic pressure. If a pressure above the osmotic pressure is applied to the saline side, then the flow reverses and fresh water can be produced. The basic RO flow diagram is shown in Fig. 10.

Osmotic pressure is a colligative property of the solution and generally increases as the concentration of dissolved ions increases. This means that as an RO system produces fresh water, known as the permeate, the salinity of the saline side increases, resulting in an increase in osmotic pressure. For a practical process, the RO system must be configured so that the average pressure along the membrane surface is suitably higher than the osmotic pressure, resulting in a net driving force for the process.

The osmotic pressure of typical municipal water is on the order of 0.33 bar, which is why faucet-mounted RO units can operate using the pressure of the municipal supply (typically 2 bar). The osmotic pressure of standard seawater is approximately 23 bar and seawater reverse osmosis (SWRO) systems typically operate such that the concentrated brine is

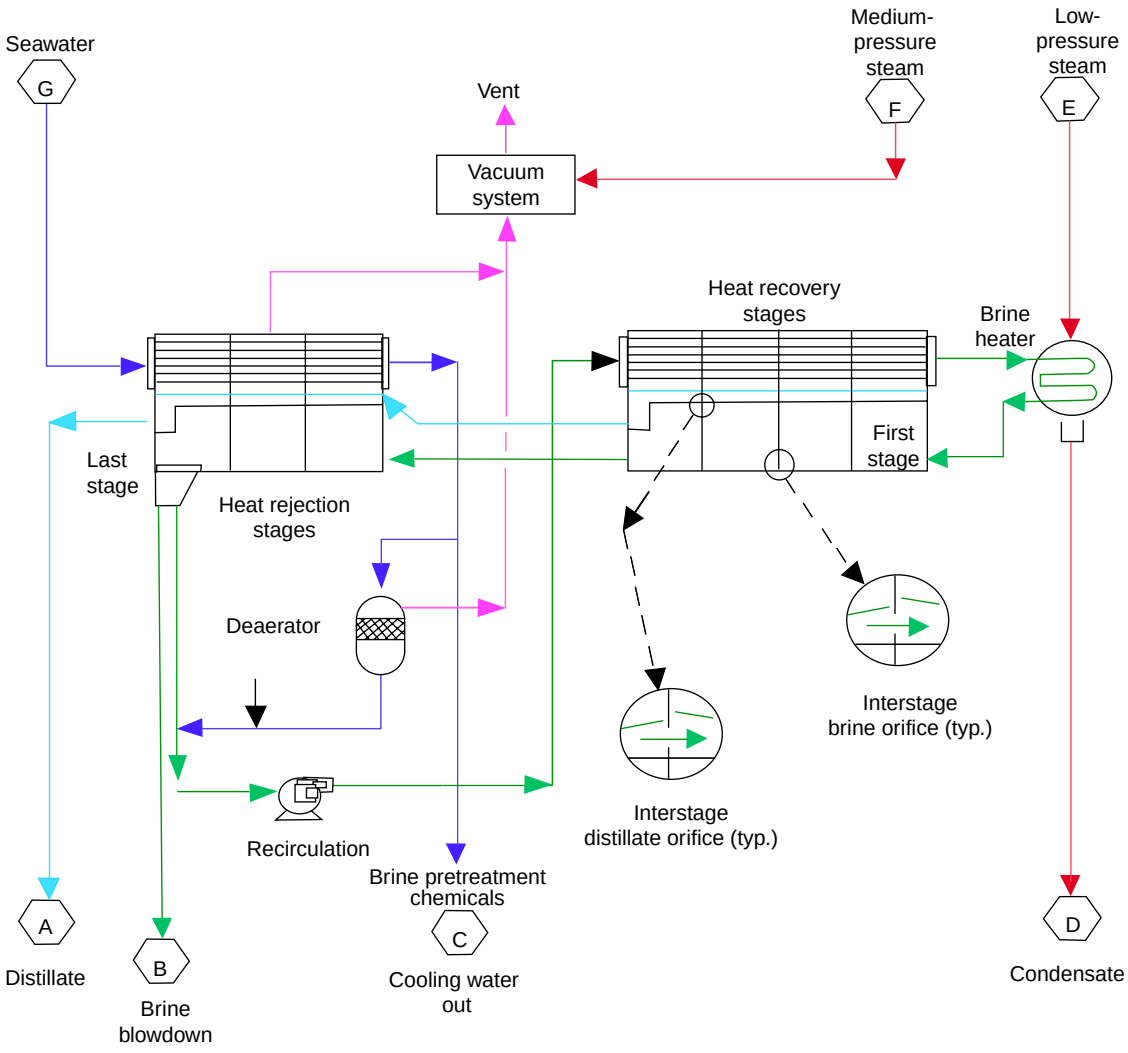


FIGURE 9 Brine recycle flow schemes.

almost two times the inlet concentration (which would be known as 50% recovery). This results in the brine having an osmotic pressure of approximately 45 bar. To achieve a practical economic driving force, the systems operate at 65–70 bar. This is close to the maximum mechanical pressure that the membranes can bear and the concentration is near the limits of solubility for certain salts. There are limited efforts in the marketplace to achieve higher concentrations, which would permit yields of more permeate than current recovery rates, which are approximately 45%.

Energy consumption for the process shown in Fig. 10 is typically 6.5 kWh per ton of permeate, which is unacceptable. The relatively large volume of high-pressure brine contains too much energy to be simply dissipated by the concentrate control valve. Energy recovery devices are installed in the brine stream to recapture the energy, as represented in Fig. 11. The most common devices are reaction turbines; often a reverse running pump; impulse turbines; normally a Pelton wheel; and pressure or work exchangers—these devices rely on proprietary methods of transferring pressure energy directly to the inlet feed stream.

The state of the art for SWRO energy consumption is approximately 2.0 kWh per ton of permeate produced; however, the average is closer to 2.75 kWh per ton of permeate produced.

Less saline water sources have much lower osmotic pressures and generally can be concentrated to 80–90% recovery before solubility limits are reached. The nature of these surface and subsurface natural waters is such that they are highly variable

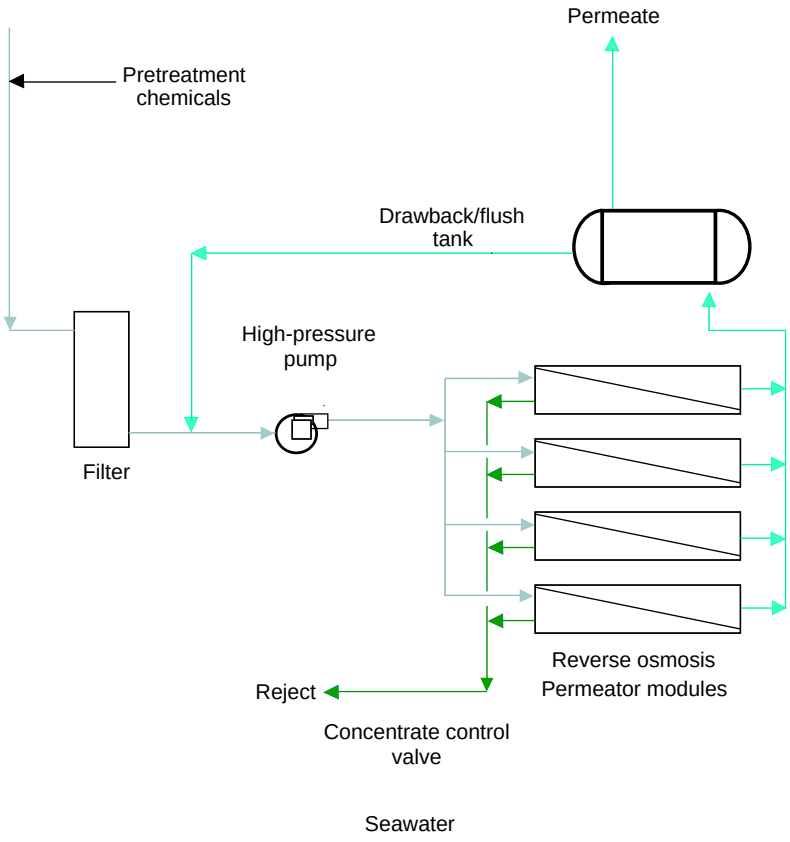


FIGURE 10 Reverse osmosis flow diagram.

from location to location and generalizations are difficult to make (unlike "standard" seawater). It is not uncommon to find waters that are already saturated in one or more ionic species and which cannot be treated (economically by standard desalination processes).

The typical brackish water RO (BWRO) operates at 85% recovery and pressures of approximately 25 bar, yielding energy consumption of 1.0 kWh/m^3. Since the pressures and brine flow volumes for BWRO are much lower than for SWRO, there has not been an economic driving force to develop energy recovery devices for BWRO. New applications of existing technology have been proposed, allowing multiple BWRO brine streams to be joined, yielding flow volumes that merit using energy recovery.

There has been work published that highlights the finding that the mechanical devices involved in RO processes (pumps, energy recovery, etc.) are reaching their practical limits and that future reductions in energy consumption must come from the membrane.

3.5 ED and EDR

Electrodialysis is quite unique among practical desalination processes since it literally desalts the feed source, whereas other processes actually remove pure water. ED is a membrane process utilizing charge-selective membranes and is pressure-based, like RO. It is possible to manufacture membranes that are charge-selective and that allow only ions with the appropriate charge to pass.

The most common configuration of ED utilizes a technique of periodically changing the polarity of the system electrodes returning in a reversed flow through the system, yielding some back-flushing of the membrane surface. This is known as electrodialysis reversal.

There are two main components of ED/EDR energy consumption: The energy of separation can be approximated as requiring 1 kWh/1000 usg per 1000 ppm of salt removed. This energy requirement varies significantly with temperature and is stated reflecting typical ambient conditions of 18–22°C.

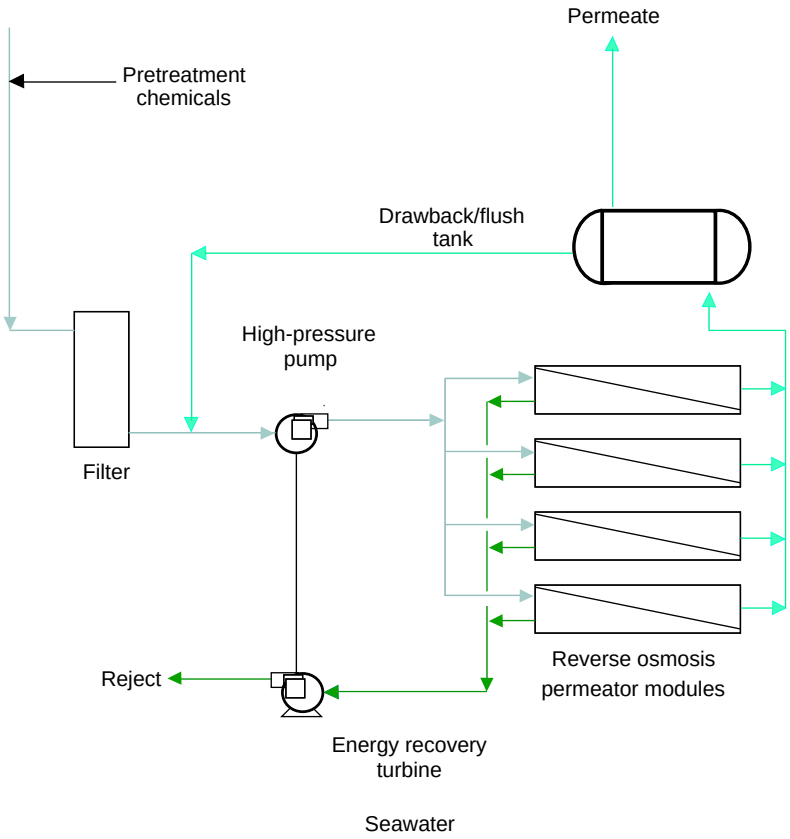


FIGURE 11 Energy recovery devices installed in the brine stream.

The energy of pumping is approximately 2 kWh/1000 usg of water produced.

SEE ALSO THE FOLLOWING ARTICLES

Ocean, Energy Flows in • *Ocean Thermal Energy* • *Solar Water Desalination*

Further Reading

El-Dessouky, H. T., and Ettouney, H. M. (2002). "Fundamentals of Salt Water Desalination." Elsevier, San Diego, CA.

Pankratz, T. M. (2003). "Seawater Desalination Processes: Operations and Costs." Technomic, Lancaster, PA.

Lattemann, S., and Hopner, T. (2003). "Seawater Desalination: Impacts of Brine and Chemical Discharge on the Marine Environment." Desalination Publications, L'Aquila, Italy, and DEStech Publications, Lancaster, PA.

Development and Energy, Overview

JOSÉ GOLDEMBERG
University of São Paulo
São Paulo, Brazil

1. Introduction
2. The Importance of Noncommercial Energy
3. Energy Intensity
4. Human Development Index
5. Generation of Jobs in Energy Production
6. Conclusions

Glossary

energy intensity The ratio of energy demand to gross domestic product.

gross domestic product (GDP) The total output produced within the geographical boundaries of the country, regardless of the nationality of the entities producing the output.

gross national product (GNP) Measure of the country's output of final goods and services for an accounting period (valued at market or purchaser's price).

human development index (HDI) Measures the average achievement of a country in basic human capabilities. The HDI indicates whether people lead a long and healthy life, are educated and knowledgeable, and enjoy a decent standard of living. The HDI examines the average condition of all people in a country: Distributional inequalities for various groups of society have to be calculated separately. The HDI is a composite of three basic components of human development: longevity, knowledge, and standard of living. Longevity is measured by life expectancy. Knowledge is measured by a combination of adult literacy (two-thirds weight) and mean years of schooling (one-third weight). Standard of living is measured by purchasing power, based on real GDP per capita adjusted for the local cost of living (purchasing power parity).

purchasing power parity (PPP) A rate of exchange that accounts for price differences across countries allowing international comparisons of real output and incomes. At the U.S. PPP rate, the PPP of $1 has the same purchasing power in the domestic economy as $1 has in the United States.

tonnes of oil equivalent (toe) Energy unit equivalent to 41.84 GJ or approximately 0.93 tonnes of gasoline, 0.99 tonnes of diesel oil, 0.96 tonnes of kerosene, 1.04 tonnes of fuel oil, 0.93 tonnes of liquified petroleum gas, 1.61 tonnes of coal, 6.25 tonnes of bagasse, 2.63 tonnes of fuel wood, and 1.35 tonnes of charcoal.

terawett-hour (TWh) Energy unit equivalent to 3.6 million GJ.

Although energy is a physical entity well understood and quantitatively defined, the concept of development is less well defined and there are different perceptions about its meaning. The World Bank measures development by the gross national product (GNP) and nations are classified in categories according to their GNP "per capita." This "monetization" of the concept of development is not well understood and is not accepted by many, particularly in developing countries, where income per capita varies dramatically between the poor and the rich. In 1999 U.S. dollars, high income is considered to be $9,266/year or more, whereas middle income is between $756 and $9,265 and low income is $755 or less. This is not the case in the Organization for Economic Cooperation and Development countries, where there is a large middle class and variations in income are not very large.

1. INTRODUCTION

The poor in developing countries, who represent 70% of the world's population, aspire to a "better life," meaning jobs, food, health services, housing (rural or urban), education, transportation, running water, sewage service, communication services, security of supplies of necessities, and a good environment. These things are usually measured in industrialized countries by monetary transactions,

 © 2004 Elsevier Inc. All rights reserved.

but this is not necessarily the case in many other countries. Climate and abundant and easily available natural resources can lead to a better life without great monetary resources. In some countries, cultural values are such that some items are less desirable than they are in other countries. In others, the political system privileges some solutions over others that cost much less. Thus, to compare stages of development only by gross domestic product (GDP) per capita can be quite misleading.

Energy in its various forms (mechanical, chemical, electrical, heat, and electromagnetic radiation) is essential for all the aspirations listed previously, and thus it is closely linked to a range of social issues. Conversely, the quality and quantity of energy services and how they are achieved have an effect on social issues.

The stages of the development of man, from primitive man (1 million years ago) to today's technological man, can be roughly correlated with energy consumption (Fig. 1).

- Primitive man (East Africa approximately 1 million years ago), without the use of fire, had only the energy of the food he ate (2000 kcal/day).
- Hunting man (Europe approximately 100,000 years ago) had more food and also burned wood for heat and cooking.
- Primitive agricultural man (Fertile Crescent in 5000 BC) grew crops and used animal energy.
- Advanced agricultural man (northeast Europe in 1400 AD) had coal for heating, water power, wind power, and animal transport.

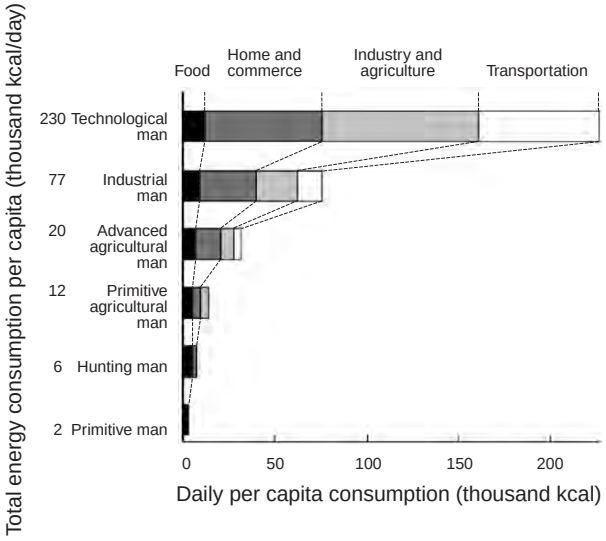


FIGURE 1 Stages of development and energy consumption. Reproduced from Cook, (1976).

- Industrial man (in England in 1875) had the steam engine.
- Technological man (in the United States in 1970) consumed 230,000 kcal/day.

From the very low energy consumption of 2000 kcal per day, which characterized primitive man, energy consumption increased in 1 million years to 230,000 kcal per day. This enormous growth of per capita energy consumption was only made possible by the increased use of coal as a source of heat and power in the 19th century; the use of internal combustion engines, which led to the massive use of petroleum and its derivatives; and Electricity generated initially in hydroelectric sites and later in thermoelectric plants. Income per capita has also increased, and one is thus tempted to determine if there is a clear correlation between energy consumption/capita and GDP/capita. Data are available for a number of countries (Fig. 2).

A linear relationship between energy consumption/capita and GDP/capita could be expected since higher income means more appliances and automobiles, larger homes, more travel, and many other activities, but empirical evidence shows that this is not the case. Clearly, there are many countries that do not fit into a linear relationship.

Moreover, some countries with the same GNP/capita can have very different energy consumption per capita, which is the case for Russia, China, and Morocco. Their GNP/capita is approximately the same, but energy consumption in Russia is four times higher than in China and 15 times higher than in Morocco.

2. THE IMPORTANCE OF NONCOMMERCIAL ENERGY

With regard to developing countries, a problem with energy data (Fig. 2), is that the vertical axis considers only "commercial energy," which is the fraction of total energy consumption that can be quantified since, by definition, it involves monetary transactions. In reality, in many areas noncommercial energy sources, such as dung, agricultural residues, and fuelwood, are used. In some countries, it represents a large percentage of total consumption, particularly for the poor. Table I shows the importance of noncommercial energy in a variety of countries.

The importance of noncommercial energy is also clearly exemplified by data that show the average energy demand by income segment in Brazil (Fig. 3).

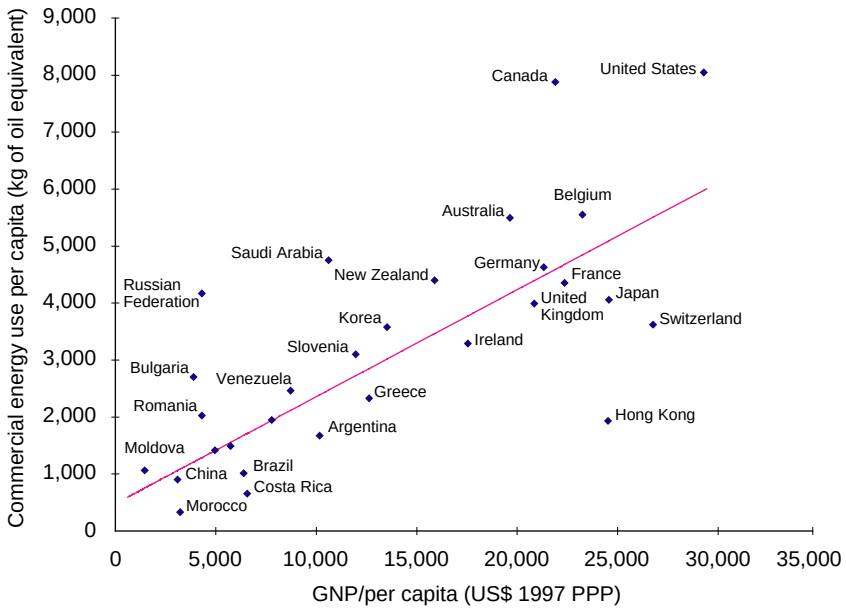


FIGURE 2 Energy use versus GNP. Reproduced from World Bank (1999).

TABLE I

Noncommercial Energy in Africa[a]

Country	1973 (%)	1985 (%)
Uganda	83	92
Malawi	87	94
Guinea-Bissau	72	67
Gambia	89	78
Guinea	69	72
Burundi	97	95
Mali	90	88
Burkina Faso	96	92

[a] Data from the World Resources Institute, United Nations Development Program, United Nations Environmental Program, and World Bank (1998).

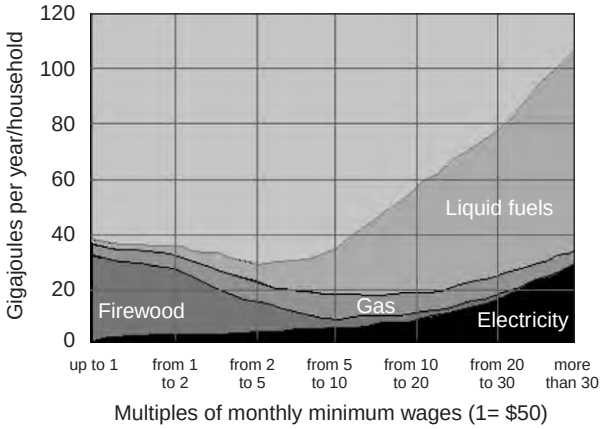


FIGURE 3 Average energy demand by income segment in Brazil, 1998. Reproduced from Almeida and Oliveira (1995).

For low-income households, firewood (usually a noncommercial energy source) is the dominant fuel. At higher income, firewood is replaced by commercial fuels and electricity, which offer much greater convenience, energy efficiency, and cleanliness. In most of Africa and India, dung and agricultural residues are used in lieu of firewood.

3. ENERGY INTENSITY

Another reason for the lack of a linear relationship between energy and GDP per capita is that the amount of additional energy required to provide energy services depends on the efficiencies with which the energy is produced, delivered, and used. Energy intensity often depends on a country's stage of development. In Organization for Economic Cooperation and Development (OECD) countries, which enjoy abundant energy services, growth in energy demand is less tightly linked to economic productivity than it was in the past (Fig. 4).

The evolution of the energy intensity ($I = E/\text{GDP}$) over time reflects the combined effects of structural changes in the economy (built into the GDP) and changes in the mix of energy sources and the

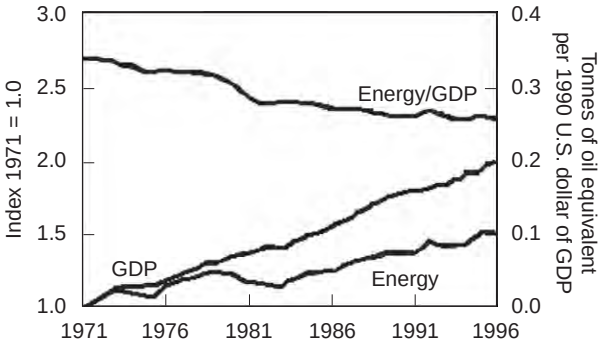


FIGURE 4 GDP and primary energy consumption in OECD countries, 1971–1996. Reproduced from the International Energy Agency (1999).

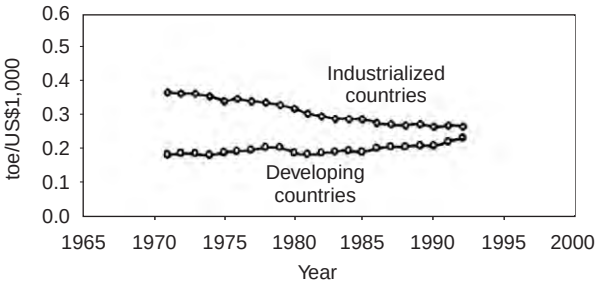


FIGURE 5 Energy intensity. Reproduced from Mielnik and Goldemberg (2000).

efficiency of energy use [built into the primary energy consumed (E)].

Although admittedly a very rough indicator, energy intensity has some attractive features: Whereas E and GDP per capita vary by more than one order of magnitude between developing and developed countries, energy intensity does not change by more than a factor of two. This is due in part to common characteristics of the energy systems of industrialized and developing countries in the "modern" sector of the economy and in part to the fact that in industrialized countries energy-intensive activities, such as jet travel, are increasingly offsetting efficiency gains in basic industries.

Energy intensity (considering only commercial energy sources) declined in OECD countries during the period 1971–1991 at a rate of approximately 1.4% per year. The main reasons for this movement were efficiency improvements, structural change, and fuel substitution. However, in the developing countries the pattern has been more varied. One study indicates that during the period 1971–1992 the energy intensity of developing and industrialized countries was converging to a common pattern of energy use. For each country, energy intensity was obtained as the ratio of commercial energy use to GDP converted in terms of purchasing power parity (PPP). The path of energy intensity of a country was given by the yearly sequence of energy intensity data during the period 1971–1994. The same procedure was followed to derive the energy intensity paths for a set of 18 industrialized countries and for 1 of 23 developing countries. The energy intensity data for each of these subsets were given by the ratio of total commercial energy use to total PPP-converted GDP for each group of countries for each year during the period 1971–1994 (Fig. 5).

4. HUMAN DEVELOPMENT INDEX

Since a clear correlation between energy and income is difficult to establish, one is tempted to search for other correlations between energy consumption and social indicators such as infant mortality, illiteracy, and fertility (Fig. 6). It is clear from Fig. 6 that energy use has many effects on major social issues, resulting in a relationship that is not necessarily casual between the parameters represented but there is strong covariance. Such behavior encouraged analysts to develop an indicator that would not only include GDP per capita but also take into account longevity and literacy. The Human Development Index (HDI), developed by the United Nations Development Program in 2001, is one way of measuring how well countries are meeting not just the economic but also the social needs of their people—that is, their quality of life. The HDI is calculated on the basis of a simple average of longevity, knowledge, and standard of living. Longevity is measured by life expectancy, knowledge is measured by a combination of adult literacy (two-thirds weight) and mean years of schooling (one-third weight), and standard of living is measured by purchasing power, based on real GDP per capita adjusted for the local cost of living (PPP).

The HDI measures performance by expressing a value between 0 (poorest performance) and 1 (ideal performance). HDI can be presented as a function of commercial energy consumption for a large number of countries in tons of oil equivalent (toe) (Fig. 7). Table II lists the characteristics of a few countries, their HDI values, and HDI rank.

An analysis of Table II indicates the importance of factors besides GDP per capita in determining the HDI rank of different countries. For example, Russia and Brazil have approximately the same GDP per capita but adult literacy in Russia (99.5%) is much

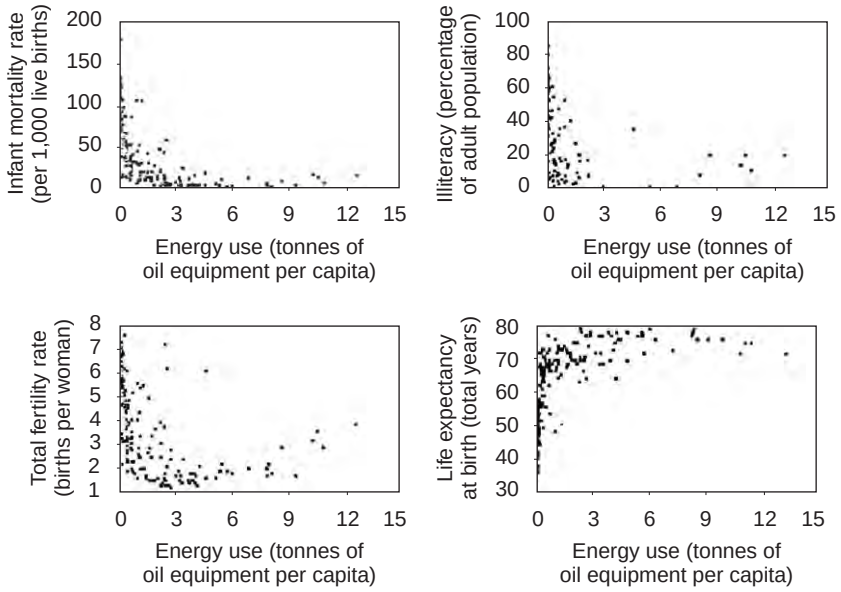


FIGURE 6 Commercial energy use and infant mortality rate, illiteracy, life expectancy at birth, and total fertility rate. Reproduced from United Nations Development Program, United Nations Department of Economic and Social Affairs, and World Energy Council (2000).

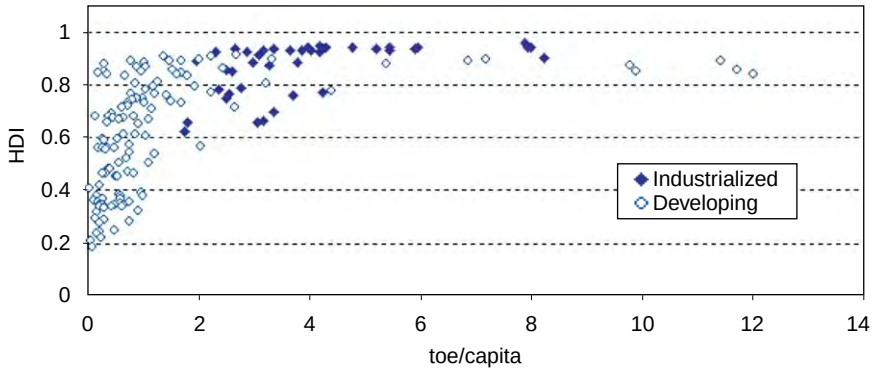


FIGURE 7 Human development index (HDI) and energy use by country. Reproduced from the World Bank (1999) and United Nations Development Program (1998).

TABLE II

Human Development (HDI) Index for Some Countries[a]

Country	Life expectancy at birth (years)	Adult literacy rate (% age 15 or older)	GDP per capita (PPR U.S. $)	HDI	HDI rank
Russia	66.1	99.5	7473	0.775	55
Brazil	67.5	84.9	7037	0.75	69
Sri Lanka	71.9	91.4	3279	0.735	81
China	70.2	83.5	3617	0.718	87
Morocco	67.2	48.0	3419	0.596	112
India	62.9	56.5	2248	0.571	115

[a] Data from the United Nations Development Program (2001).

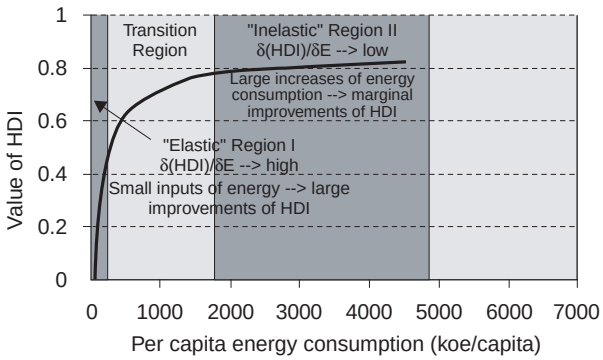


FIGURE 8 HDI versus per capita energy consumption. Reproduced from Reddy (2002).

higher than that in Brazil (84.9%), which ranks the former 55 compared to 69 for Brazil. Sri Lanka and China have a GDP per capita that is approximately the same as that of Morocco but the life expectancy and adult literacy of the former are much higher, which give Sri Lanka a ranking of 81, China 87, and Morocco 112.

It is apparent from Fig. 7 that for an energy consumption more than 1 toe/capita/year, the HDI is higher than 0.8 and essentially constant for all countries. Therefore, 1 toe/capita/year seems to be the minimum energy needed to guarantee an acceptable level of living as measured by the HDI, despite many variations in consumption patterns and lifestyles across countries.

The relationship between HDI and energy has several important implications. The relationship can be considered to consist of two regions (Fig. 8). In region I, the elastic region, – the slope δ (HDI)/δ (E) of the HDI vs E curve is high. In region II, the inelastic region, the slope δ (HDI)/δ (E) of the HDI vs E curve is low. Even large inputs of energy (large improvements of energy services) result in only marginal improvements in HDI; that is, the HDI–energy (benefit–cost) ratio is very low. In region I, enhanced energy services lead directly to improvement. Thus, the implication of the elastic and inelastic regions is that in the elastic region increased energy services guarantee direct improvement of HDI, whereas improvement of HDI via income depends on the uses of the income.

5. GENERATION OF JOBS IN ENERGY PRODUCTION

A variety of sources provide estimates of the number of jobs directly related to the production of energy.

TABLE III

Jobs in Energy Production[a]

Sector	Jobs (person-years), TWh
Petroleum	260
Offshore oil	265
Natural gas	250
Coal	370
Nuclear	75
Wood energy	1000
Hydro	250
Minihydro	120
Wind	918
Photovoltaics	7600
Ethanol (from sugarcane)	4000

[a]Data from Goldemberg (2003).

Some of these estimates are shown in Table III. These data include jobs involved in operating the generating stations, excluding the jobs involved in producing and commissioning the equipment. The very large number of jobs generated by the truly decentralized options, which are photovoltaics and ethanol from sugarcane, is striking.

Photovoltaics energy is usually generated (and used) in small modules of 100 W and the generation of 1 TWh typically requires 10 million modules to be installed and maintained. Ethanol production from sugarcane in Brazil involves agricultural production using approximately 4 million ha and industrial production in sugar–alcohol distilleries, accounting for the large number of jobs directly related to the production of energy.

There are also some rough estimates of the number of jobs generated by energy conservation. A study performed in Sacramento, California, showed that saving enough energy to operate at less than 100-MW power plant capacity requires 39 jobs compared with 15–20 jobs required to operate at the same amount of capacity at a modern coal or gas-fired power plant.

6. CONCLUSIONS

Energy has a determinant influence on the HDI, particularly in the early stages of development, in which the vast majority of the world's people, particularly women and children, are classified. The influence of per capita energy consumption on the HDI begins to decline between 1000 and 3000 kg of

oil equivalent (koe) per inhabitant. Thereafter, even with a tripling of energy consumption, the HDI does not increase. Thus, from approximately 1000 koe per capita, the strong positive covariance of energy consumption with HDI starts to diminish. The efficiency of energy use is also important in influencing the relationship between energy and development.

Another aspect of the problem is the mix of supply-side resources that dominate the world's energy scene today. Fossil fuels have a dominating role (81% of supply in OECD countries and 70% in developing countries), although, as a rule, renewables are more significant for low-income populations. However, there are significant advantages to increasing the role of renewable sources since they enhance diversity in energy supply markets, secure long-term sustainable energy supplies, reduce atmospheric emissions (local, regional, and global), create new employment opportunities in rural communities offering possibilities for local manufacturing, and enhance security of supply since they do not require imports that characterize the supply of fossil fuels.

More generally, development, including the generation of jobs, depends on a number of factors in addition to GNP per capita. Furthermore, although an essential ingredient of development, energy is more important with regard to low rather than high incomes.

SEE ALSO THE FOLLOWING ARTICLES

Cultural Evolution and Energy • Economic Growth and Energy • Energy Ladder in Developing Nations • Global Energy Use: Status and Trends • International Comparisons of Energy End Use: Benefits and Risks • Population Growth and Energy • Sustainable Development: Basic Concepts and Application to Energy • Technology Innovation and Energy • Transitions in Energy Use • Women and Energy: Issues in Developing Nations

Further Reading

Almeida, E., and Oliveira, A. (1995). Brazilian life style and energy consumption. *In* "Energy Demand, Life Style Changes and Technology Development." World Energy Council, London.

Cook, E. (1976). "Man, Energy, Society." Freeman, San Francisco.

Goldemberg, J. (2003). "Jobs Provided in the Energy Sector." [Adapted from Grassi, G. (1996). Potential employment impacts of bioenergy activity on employment. *In* "Proceedings of the 9th European Bioenergy Conference" (P. Chartier *et al.*, Eds.), Vol. 1, pp. 419–423. Elsevier, Oxford, UK; Carvalho, L. C., and Szwarc, A. (2001). Understanding the impact of externalities: case studies. Paper presented at the Brazil International Development Seminar on Fuel Ethanol, December 14, Washington, DC; Perez, E. M. (2001). "Energias Renovables, Sustentabilidad y Creacion de Enpleo: Una Economía Impulsionada por el Sol (Renewable Energy, Sustainability and Job Creation: An Economy Pushed Forward by the Sun), ISBN: 84–8319 115-6; and Renner, M. (2000, September). Working for the environment: A growing source of jobs, Worldwatch Paper No. 152. Worldwatch Institute]

International Energy Agency (1999). "Energy Balances of OECD Countries." International Energy Agency, Paris.

Mielnik, O., and Goldemberg, J. (2000). Converging to a common pattern of energy use in developing and industrialized countries. *Energy Policy,* pp. 503–508.

Reddy, A. K. N. (2002). Energy technologies and policies for rural development. *In* "Energy for Sustainable Development" (T. Johansson and J. Goldemberg, Eds.). International Institute for Industrial Environmental Economics University and United Nations Development Program, New York.

United Nations Development Program (1998). "Human Development Report." United Nations Development Program, New York.

United Nations Development Program (2001). "Human Development Index." United Nations Development Program, New York.

United Nations Development Program, United Nations Department of Economic and Social Affairs, and World Energy Council (2000). "World Energy Assessment" (J. Goldemberg, Ed.). United Nations Development Program, United Nations Department of Economic and Social Affairs, and World Energy Council, New York.

World Bank (1999). "World Development Indicators for Energy Data." World Bank, Geneva.

World Resources Institute, United Nations Development Program, United Nations Environmental Program, and World Bank (1998). "World Resources 1998–1990—Environmental Change and Human Health." World Resources Institute, United Nations Development Program, United Nations Environmental Program, and World Bank, New York.

Diet, Energy, and Greenhouse Gas Emissions

ANNIKA CARLSSON-KANYAMA
Environmental Strategies Research Group at FOI
Stockholm, Sweden

1. **Introduction**
2. **Calculations of Energy-Efficient and Low-Pollution Diets**
3. **Trends in Diets**
4. **Policy Instruments for Dietary Change**
5. **Conclusions**

Glossary

input/output analysis (**IOA**) A method traditionally used to describe the financial linkages and network of input supplies and production that connect industries in a regional economy (however defined) and to predict the changes in regional output, income, and employment; also used for environmental impact analysis today.

Intergovernmental Panel on Climate Change (**IPCC**) - Jointly established by the World Meteorological Organization and the UN Environment Program in 1998 to assess available scientific information on climate change, assess the environmental and socioeconomic impacts on climate change, and formulate response strategies.

International Vegetarian Union (**IVU**) Established in 1908 as a nonprofit-making organization with membership open to any nonprofit organization whose primary purpose is to promote vegetarianism and that is governed exclusively by vegetarians; aims to promote vegetarianism throughout the world.

life cycle assessment (**LCA**) A method to assess the potential environmental impact of a material, product, or service throughout its entire life cycle, from the extraction of raw materials, through the production process, through the user phase, and to the final disposal.

Energy use and greenhouse gas emissions in the current food system are substantial. The food system comprises the whole chain of events resulting in edible food—production of farm inputs, farming, animal husbandry, transportation, processing, storage, retailing, and preparation. Ecological impacts are not evenly distributed among products; therefore, dietary change is a possibility for environmental improvement. From a number of studies, it is possible to detect patterns of dietary recommendations that would lower energy use and greenhouse gas emissions in the food system. They include reducing consumption of meat, exotic food, food with little nutritional value, and overprocessed food. Overall trends in dietary patterns are not conducive to higher environmental efficiency, but the trends toward vegetarianism have an environmental potential. Policy instruments for encouraging more efficient diets include information, supportive structures, and financial incentives. However, they have scarcely been used in mitigation strategies for climate change.

1. INTRODUCTION

For a long time, it has been known that the production of certain food products takes a larger toll on natural resources than that of others, and at times dietary change has been proposed as a way of conserving such resources. During the 1960s, when global concern was about overpopulation and starvation, Georg Borgström proposed a diet lower down on the food chain, implying less use of cereals for animal feed. During the same period, Francois Lappé composed a diet for a small planet based on vegetarian ingredients only to save agricultural land. During the 1970s, when the concern was about energy resources, David and Marcia Pimentel presented calculations of energy inputs over the life cycle for bread and beef. The beef was found to be

 © 2004 Elsevier Inc. All rights reserved.

more energy demanding by a factor of 29 when the two products were compared on a weight basis.

Today, when climate change has emerged as perhaps the most controversial global environmental issue, the question of dietary change as a mitigation strategy has partly resurfaced. This is because energy use and greenhouse gas emissions from the current food system are substantial. Energy use in the food system typically amounts to 12 to 20% of the total energy consumed in developed countries. Current anthropogenic emissions of carbon dioxide (CO_2) are primarily the result of the consumption of energy from fossil fuels, the main energy source in the developed parts of the world. Another 10 to 30% of the current total anthropogenic emissions of CO_2 are estimated to be caused by land-use conversion, partly related to production of cash crops. In addition, the agricultural sector is a large source of emissions of methane and nitrous oxides. It accounts for half of the global anthropogenic emissions of methane through rice farming and rearing of ruminants, whereas 70% of the global anthropogenic emissions of nitrous oxide come from agricultural soils, cattle, and feedlots.

In looking more closely at where energy is used in the food system, one finds that a surprisingly large amount is used by the consumer. For the United States, a review of several studies showed that home preparation contributed between 23 and 30% of the total, figures comparable to processing, which contributed 24 to 37%. For Sweden, a recent survey showed that household energy use for cooking and storing food was 28% of the total, whereas processing contributed 25% (Fig. 1). A different pattern may emerge when looking at emissions of carbon dioxide (Fig. 1). Here, the system for electricity production may alter the energy pattern. For Sweden and some other Nordic countries, extensive use of hydropower results in low emissions, as does nuclear power production.

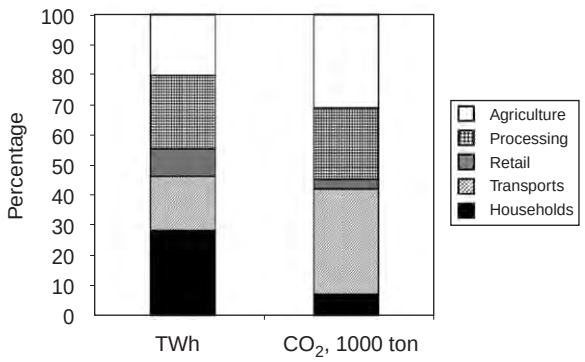


FIGURE 1 Energy use and emissions of carbon dioxide in the Swedish food system in the year 2000.

The purpose of this article is to provide an overview of current research about energy, greenhouse gas emissions, and dietary change. The approach is product oriented with a life cycle perspective, meaning that individual products are considered "from cradle to grave."

The article begins with some recent results from calculations of the environmental impacts over the life cycle for food products and diets. Several studies conducted in Europe and the United States provide sufficient basis for an informed discussion about the potential for change. In the subsequent section, some recent trends in global food consumption are reviewed for their environmental implications, and the trend toward vegetarianism is evaluated. Some lessons from research about policy instruments and dietary change are discussed with a focus on information and its efficiency. Finally, conclusions of relevance for further mitigation efforts are drawn.

2. CALCULATIONS OF ENERGY-EFFICIENT AND LOW-POLLUTION DIETS

Investigating possibilities for low-impact diets involves analyzing the resource use and emissions of the numerous individual food products that constitute a diet throughout the life cycle. Several attempts have been made to this end, and as a result broad patterns of dietary recommendations can be detected. Based on several studies, a diet with less meat and cheese, more in-season vegetables, more locally produced and fresh foods, better use of products, and more use of products from the "wild" could be recommended. The results from some of these studies are discussed in what follows, as are the methods used for computing the environmental impacts.

A team of Dutch researchers used a combination of life cycle assessment (LCA) and input/output analysis (IOA) to estimate greenhouse gas emissions from food products consumed in The Netherlands. The results were expressed as greenhouse gas emissions per Dutch florin (Dfl), and calculations were performed for more than 100 products. This method is useful in that the results can be directly combined with household expenditure profiles to estimate impacts of a whole consumption pattern but lack some detail when it comes to differentiation among various processing methods and so on. Some

results from these calculations were that greenhouse gas emissions during the life cycle of rice are nearly five times the emissions during pastry life cycles and that tomatoes emit four times as much greenhouse gases as do green beans. The reasons for these differences relate to the paddy field cultivation methods for rice and the fact that tomatoes are usually cultivated in fossil fuel-heated greenhouses. Regarding the food consumption emission profile of a Dutch household, methane emissions came mainly from products originating from cattle farming and dairy production. Enteric fermentation by ruminants explains much of this. The dietary recommendations made by these authors included substituting meat and cheese with eggs, nuts, or certain vegetable products and shifting from greenhouse-cultivated vegetables to those grown in the open.

In the United States, two researchers used IOA to calculate greenhouse gas emissions for 44 food products sold in the country. Their results were expressed as amount of emissions per dollar and were combined with household expenditures to calculate total household emissions. Regarding dietary change, the authors' proposal was to lower the expenditure on meat and poultry in favor of expenditure on vegetable products such as pasta. This was because the emissions intensity of beef and pork was 166 g per U.S. dollar, whereas the same intensity for pasta was only 75 g per U.S. dollar. Lowering meat consumption was also considered to be of major importance when all activities of the household were assessed for greenhouse gas emissions and other environmental pollutants.

A Swedish researcher investigated energy use over the life cycle for food supplied in Sweden using a life cycle inventory approach. The study covered 150 food items, and results were expressed as megajoules per kilogram of product. The advantage of this approach is that results can easily be combined with dietary information and that the level of detail in the environmental calculations is high. Among the results from this study are that energy use per kilogram of ready-to-eat food can vary from 2 to 220 MJ/kg, that is, by a factor of more than 100. The highest values were found for certain kinds of fish and seafood. A multitude of factors determine the level of energy inputs, and they relate to animal or vegetable origin, degree of processing, choice of processing and preparation technology, and transportation distance. The study confirmed several earlier findings about environmentally efficient diets while adding some new perspectives. One is the role of food preparation (which can add substantial energy inputs if food is cooked in small portions), another is the remarkably high energy inputs for certain fish-harvesting techniques, and a third is the impacts from food losses on total energy inputs in food life cycles. Examples of daily total life cycle energy inputs for diets with similar dietary energy were calculated, and the result was that energy inputs could vary by a factor of 3. Table I shows meal components for two meals with high and low energy inputs but with the same amount of dietary energy.

In a study of six food products for which LCA was used, detailed calculations of greenhouse gas emissions confirmed the large differences between some products of animal origin and some of vegetable origin. In this study, emissions of greenhouse gases per kilogram of ready-to-eat beef were 135 times higher than the emissions per kilogram of ready-to-eat potato (Table II). There are several explanations for the high emissions for beef: enteric fermentation causing emission of methane, energy use during animal feed production, and transportation of food ingredients. The reasons for lower emissions for pork and chicken are partly related to the digestive systems of pigs and poultry but are also due to the fact that they are much more efficient feed converters than cattle.

TABLE I

Meal Components, Dietary Energy, and Life Cycle Energy Inputs for Two Different Dinners: High and Low

Meal component	Kilograms	Megajoules dietary energy	Megajoules life cycle inputs
Dinner: High			
Beef	0.13	0.80	9.4
Rice	0.15	0.68	1.1
Tomatoes, greenhouse	0.070	0.06	4.6
Wine	0.30	0.98	4.2
Total	0.65	2.51	19
Dinner: Low			
Chicken	0.13	0.81	4.37
Potatoes	0.20	0.61	0.91
Carrot	0.13	0.21	0.50
Water, tap	0.15	0.23	0.0
Oil	0.02	0.74	0.30
Total	0.60	2.61	6.1

Source. Carlsson-Kanyama *et al.* (2003).

TABLE II

Emission of Greenhouse Gases per Kilogram Ready-to-Eat Food for Six Food Products Supplied in Sweden

	Grams CO_2 equivalents per kilogram ready-to-eat product
Beef	13,500
Pork	4,200
Chicken	1,400
Potatoes	100
Milk	870
Bread	500
Lettuce	230

When David and Marcia Pimentel revised their book, *Food, Energy, and Society,* in 1996, they concluded,

> *One practical way to increase food supplies with minimal increase in fossil energy inputs is for the world population as a whole to consume more plant foods. This diet modification would reduce energy expenditures and reduce food supplies, because less food suitable for human consumption would be fed to livestock.*

As described previously, a number of studies have certainly shown this to be true, although detailed calculations of energy use and greenhouse gas emissions over the life cycle of different kinds of meat show that there are large differences between, for example, chicken and beef. This opens the scope for discussing the role of meat in a more environmentally efficient diet as well as for questioning the role of certain kinds of fish and vegetables and the total amount of food consumed.

How can food consumption patterns be influenced, what are the current food trends, and what is the role of information for disseminating environmental knowledge about diets? The next section contains a review of some significant elements in current food trends and their implications for energy use and greenhouse gas emissions.

3. TRENDS IN DIETS

From the statistics databases provided by the Food and Agricultural Organization (FAO), it is possible to provide some global figures on dietary trends of interest for the environmental impacts in the food system. Figure 2 shows some developments with respect to this using starchy roots, wheat, pulses, meat, fish, and sweeteners as examples. The rationale

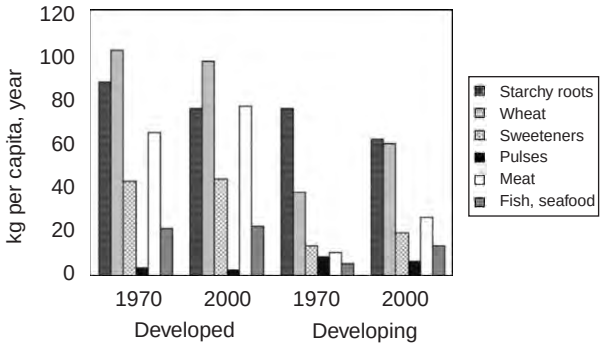


FIGURE 2 Consumption levels of some food products in 1970 and 2000 in the developed and developing worlds.

for choosing these products was that the life cycle-related levels of energy inputs and greenhouse gas emissions are commonly substantially different. Starchy roots and pulses are expected to have much lower impacts than do meat and fish, whereas sweeteners are interesting because they represent food products with considerable production-related energy inputs but with little or disputed nutritional value. Wheat is grown mainly in the Northern Hemisphere but is increasingly replacing traditional starch-rich crops, such as cassava and millet, in the developing parts of the world. Although wheat is not a very energy-demanding food item compared with meat and fish, it requires higher life cycle energy inputs than do some domestic starchy crops due to long transportation distances and intensive cultivation methods.

Figure 2 shows that the consumption of the selected food products in the developing parts of the world is still quite different from that in the developed world. This is particularly evident when it comes to meat consumption and sweeteners, with the 2000 levels per capita in the developed world exceeding those in the developing world by a factor of between 2 and 3. Figure 2 also shows that the most notable changes in the developed world during the past three decades were a 25% decrease in the consumption of pulses and an 18% increase in the consumption of meat. In the developing world, the changes were even more spectacular in that the consumption of meat and fish more than doubled and the consumption of wheat and sweeteners increased by approximately 50%.

The environmental implications of these dietary changes can only be interpreted as negative with respect to the prospects for lowering energy use and greenhouse gas emissions. Whether or not actual emissions have increased depends on the magnitude

of efficiency improvements in the food supply system. More efficient trucks, ships, processing equipment, and refrigerators have to some extent counteracted the negative environmental impacts of dietary change.

Along with the negative dietary changes experienced on a global scale, certain trends in the developing parts of the world (e.g., vegetarianism) hold some promise for more efficient diets in an affluent environment. Vegetarianism is not a new phenomenon in Western societies but has become a mainstream feature in several countries. According to the International Vegetarian Union (IVU), there is a vegetarian subculture in nearly every European country today. It is becoming increasingly common to have vegetarian options in nonvegetarian restaurants as an alternative for those who do not adhere to a vegetarian diet on an everyday basis. No statistics exist on how many vegetarians there are globally, but estimates from several European countries show the highest percentages in the United Kingdom, Germany, and Switzerland (~8–9% of these populations) and much lower percentages in Spain, Norway, and France (~2% of these populations).

The main reasons for choosing a vegetarian diet are commonly better health and nutrition as well as animal welfare considerations. The potential for environmental efficiency has been less investigated, and when it is claimed it is mostly mentioned as a way in which to avoid excessive use of agricultural land. Recent findings about energy inputs and greenhouse gas emissions during the life cycle of food products provide a new tool for evaluating the effects of vegetarianism. A problem with such an evaluation is that the dietary recommendations for vegetarians are only exclusive. They describe what not to eat but do not prescribe any alternatives. So, meat may be replaced by a very energy-efficient alternative, such as legumes, but also by other less efficient alternatives, such as vegetables grown in greenhouses and fruits transported by airplane. However, the trend toward vegetarianism should be considered as an environmentally promising one provided that the alternatives to animal food products are chosen among those vegetable products with moderate life cycle-based resource demands. When discussing potential environmental gains from a vegetarian diet, it is important to know that the environmental impacts from meat in an ordinary diet might not be dominant. One example of this is shown in Fig. 3, where the life cycle-based energy inputs for typical diets in Sweden are displayed. The figure shows that the energy inputs from meat and fish may constitute about a third of the total. The meat and fish part of the diets in other countries may contribute differently, and if the greenhouse gas emissions profile becomes available, it could also alter this conclusion to some extent.

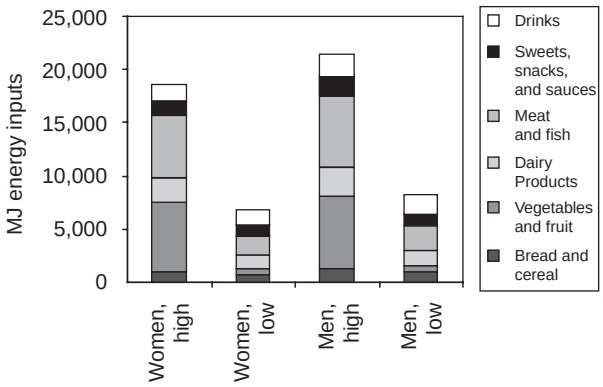


FIGURE 3 Life cycle energy inputs for the annual food consumption of average men and women in Sweden. The ranges indicate possible levels with various food choices.

Having reviewed current trends and commented on those with potential for environmental efficiency, we now turn to the issue of how policy could be used to influence diets so as to decrease energy inputs and greenhouse emissions from the food system. First, it is important to remember that experiences from large-scale campaigns for environmentally friendly diets are lacking and that the few results that can be drawn on come from isolated and relatively small research projects. In this context, it is also interesting to note that the Intergovernmental Panel on Climate Change (IPCC) does not mention dietary change among the mitigation potentials in its recent report about mitigation. Unfortunately, the same "blindness" toward dietary change as a possible environmental measure is not uncommon in policy documents. This article does not speculate on the reasons for this but rather concentrates on the results of some research in which information was used as a policy instrument.

4. POLICY INSTRUMENTS FOR DIETARY CHANGE

How to make everyday behavior more environmentally friendly has been on the agenda for nearly a half-century, and much has been learned. Both external and internal barriers must be overcome. There has been an optimistic belief in the use of information in the hope of fostering environmentally

sound behavior among members of the general public. However, it has been found that there is no automatic link between knowing and doing. Norms, values, feelings, and consumption opportunities have been shown to be a vital part of the picture.

However, there are several policy tools, in addition to information, that can be used for changing everyday behavior, and they differ in influence and effect (Table III).

Four groups of policy instruments can be identified—information, economic instruments, administrative instruments, and physical improvements—and they are often used in various combinations to increase efficiency.

Information can be mediated in several ways. When conveying environmental information about products, written information in pamphlets and advertisements is often used, as is product labeling. The common denominator for all of these ways is that they should attract attention. The receiver of information is supposed to notice and benefit from the new arguments voluntarily.

Economic instruments include taxing, pricing, and refund schemes. Besides increasing the costs for pollution and providing economic benefits for environmentally friendly consumption, these instruments inspire polluters to reflect over changes in behavior or use of technology and to make plans for behavioral change to minimize the costs for polluting the environment. Economic instruments function as catalysts for changes in the future.

Administrative instruments have an immediate effect on all of the actors they are intended to affect from an announced date. The success depends largely on the negative sanctions with which those with deviant behavior are punished. Examples include fines, imprisonment, and prohibition of trade.

Physical improvements, such as construction of waste depositories and bicycle lanes, are intended to facilitate a new pattern of behavior. Such improvements are very often combined with some other instrument pushing the behavior in the same direction. For example, separate lanes for cars and bicycles are followed by traffic rules regulating where the two kinds of vehicles may be driven.

TABLE III

Policy Instruments, Influence on Actors, and Effects

Instrument	Influence	Effect
Information	Voluntary	Slow
Economic instruments	Catalytic	Short range
Administrative instruments	Immediate, forcing	Middle range
Physical improvements	Reminding, repeating	Change habits

From this knowledge, it is obvious that information will be more efficient if it is combined with other policy instruments. As an example, if households are given life cycle-based environmental information about their diets and food choices, the chances that they will lower environmental impacts of their diets increase if (1) there are tax subsidies for low-polluting food products (economic instrument), (2) the sale of food products with very high environmental impacts is prohibited (administrative instrument), and (3) a steady and diverse supply of low-polluting food products is available in nearby shops and restaurants (physical improvement).

Having reviewed the various policy instruments and how they can be integrated, this article now presents some results from research where the potential or actual impacts of environmental information given to households have been tested. In both of these examples, scientific methods were used for evaluation.

4.1 Negative versus Positive Labels

A team of environmental psychologists tested the impact of negative versus positive ecolabels on consumer behavior. The background was that all ecolabels today signal a positive outcome, for example, "Choose this product; it is better for the environment than the average product." Another strategy would be to signal a negative outcome with the purpose of making consumers avoid a product. A laboratory test was carried out on 40 students in Sweden who were asked to choose between products labeled with a three-level ecolabel inspired by traffic lights: red, green, and yellow. The red label was defined as much worse in terms of environmental consequences than the average, the green label was defined as much better than the average, and the yellow label was defined as average. All participants were asked to rate the importance attached to environmental consequences when purchasing everyday commodities. The results showed that those who attached high importance to environmental consequences when purchasing everyday commodities were most strongly affected by ecolabels of any kind. The results also showed that those who attached little importance to environmental consequences when purchasing everyday commodities were unaffected by any kind of label. However,

those who had an intermediately strong interest in environmental issues were more affected by the red label than by the green one. The conclusion was that the ecolabels in use today affect mostly individuals with a strong interest in environmental issues and that a negative label influences a wider range of individuals.

4.2 Households and Researchers in Partnership When Experimenting with Alternative Diets

In a research project titled "Urban Households and Consumption-Related Resource Use," 10 households participated in an experiment where they tried to reduce life cycle energy inputs related to their food consumption. A manual for low-energy foods habits was developed, and based on this information, the households planned and implemented food habits demanding fewer life cycle energy inputs. The effect of participation in the experiment was evaluated.

Among the changes of most importance for lowering energy inputs were a reduction in consumption of meat (mostly pork); an increase in vegetables and fruits, especially vegetables grown in the open; an increase in poultry, eggs, and fish; and an increase in legumes. The households frequently bought organic food, chose products from Sweden instead of imported ones, and used electric kettles for boiling water for coffee and tea as part of their efforts to achieve more energy-efficient diets. Life cycle energy inputs decreased by at most a third. Environmental aspects other than energy inputs (e.g., animal welfare) were also important to the households. Some recommendations, such as supplementing meat with pulses, were hard to follow because the households did not know how to prepare them.

The overall conclusion from this project was that it takes time to educate households to become "green" and that forces from both inside and outside the household must support such an effort. Being involved in an experiment can affect everyday habits, as can feedback on one's own resource use. However, misconceptions and distrust are deeply rooted, and complex information is difficult to convey to consumers. All of this must be considered if we want to count on the collaboration of households for ecologically sustainable development. Table IV shows the changes in food consumption that occurred as a result of the experiment.

TABLE IV

Changes in Quantities of Food Consumed in the 10 Households Participating in the "Urban Households and Consumption-Related Resource Use" (URRU) Project

Food category	Percentage change
Legumes and seeds	132
Egg	97
Poultry	55
Vegetables	86
Fruit, berries, and jam	25
Fish	22
Pork	−30
Lamb	−22
Bread, beef, drinks, fermented milk, cheese, and yogurt	Negligible
Food composed of several food items, porridge and cereals, potatoes, pasta and rice, pastries and sweets, ice cream and desserts	Negligible

Source. Carlsson-Kanyama *et al.* (2002).

5. CONCLUSIONS

Greenhouse gas emissions and energy inputs in the current food system are substantial. Results from case studies of food products followed throughout their life cycles show important differences depending on animal or vegetable origin, degree of processing, transportation distance and mode of transport, preparation methods, and so on. This leaves room for discussing dietary change as a mitigation strategy for curbing climate change. The theoretical potential for such a strategy is large and is open mainly to affluent consumers.

A successful mitigation package will have to recognize that a combination of policy instruments will be most efficient and that any information applied by itself will have a slow change potential. However, there are many different ways in which to produce and disseminate information that will affect the outcome in terms of changes in general behavior. In general, information with feedback is more efficient than one-way communication. Moreover, involving households in experiments with alternative lifestyles has been shown to produce lasting results. Information should also be tailored to the everyday situation of the household for most impact. Although current global trends in food consumption are not promoting higher environmental efficiency, there are some trends that could develop in that direction.

Policymakers should recognize and support such trends for successful mitigation.

SEE ALSO THE FOLLOWING ARTICLES

Development and Energy, Overview • Global Material Cycles and Energy • Greenhouse Gas Abatement: Controversies in Cost Assessment • Greenhouse Gas Emissions, Alternative Scenarios of • Greenhouse Gas Emissions from Energy Systems, Comparison and Overview • Input–Output Analysis • Life Cycle Assessment and Energy Systems • Suburbanization and Energy

Further Reading

Borgström, G. (1972). "The Hungry Planet: The Modern World at the Edge of Famine." Collier Books, New York.

Brower, M., and Leon, W. (1999). "The Consumer's Guide to Effective Environmental Choices: Practical Advice from the Union of Concerned Scientists." Three Rivers Press, New York.

Carlsson-Kanyama, A., Pipping Ekström, M., and Shanahan, H. (2002). Urban Households and Consumption-Related Resource Use (URRU): Case studies of changing food habits in Sweden. *In* "Changes in the Other End of the Chain" (Butjin *et al.*, Eds.). Shaker Publishing, Maastricht, Netherlands.

Carlsson-Kanyama, A., Pipping Ekström, M., and Shanahan, H. (2003). Food and life cycle energy inputs: Consequences of diet and ways to increase efficiency. *Ecol. Econ.* **44,** 293–307.

Grankvist, G., Dahlstrand, U., and Biel, A. (2002). The impact of environmental labelling on consumer preference: Negative versus positive labels. *In* "Determinants of Choice of Eco-Labelled Products" (G. Grankvist, doctoral thesis). Department of Psychology, Göteborg University, Sweden.

Jungbluth, N., Tietje, O., and Scholz, R. W. (2000). Food purchases: Impacts from consumers' point of view investigated with a modular LCA. *Intl. J. Life Cycle Assessment* **5**(3), 134–142.

Kramer, K. J., Moll, H. C., Nonhebel, S., and Wilting, H. C. (1999). Greenhouse gas emissions related to Dutch food consumption. *Energy Policy* **27,** 203–216.

Linden, A-L., and Carlsson-Kanyama, A. (2003). Environmentally friendly behavior and local support systems: Lessons from a metropolitan area. *Local Environ.* **8,** 291–301.

Pimentel, D., and Pimentel, M. (eds.). (1996). "Food, Energy, and Society," rev. ed. University Press of Colorado, Boulder.

Discount Rates and Energy Efficiency Gap

RICHARD B. HOWARTH
Dartmouth College
Hanover, New Hampshire, United States

1. Introduction
2. The Hidden Costs Argument
3. High Discount Rates?
4. Imperfect Information and Bounded Rationality

Glossary

bounded rationality The notion that people behave in a manner that is nearly but not fully optimal with respect to their goals as their available information and cognitive limitations will allow.

discount rate The rate at which someone prefers present over future benefits in a multiperiod model.

efficiency gap The difference between the actual level of investment in energy efficiency and the higher level that would be cost beneficial from the consumer's standpoint.

imperfect information This refers to a situation in which an individual makes decisions without full knowledge of costs and benefits.

Investments in enhanced energy efficiency typically involve a trade-off between short-term costs and long-term benefits. Consider, for example, a home insulation project that would generate energy savings of $130 per year during a 30-year expected life span. If the project required an initial investment of $900, would it be justified from a financial perspective? A naive answer to this question is that the project would provide a long-term net savings of $3000, thereby warranting its implementation. However, although this type of reasoning seems intuitive to many people, it departs from the standard methods of financial analysis, in which costs and benefits that accrue in the future should be discounted relative to the present.

1. INTRODUCTION

The logic behind discounting is linked to people's observed behavior in financial markets. Suppose, for example, that an investor would receive an r percent rate of return on standard financial instruments such as corporate stocks. If the investor was well informed and rational, then he or she would be willing to pay no more than $B/(1+r)^t$ dollars in the present for an asset that he or she could resell to obtain B dollars t years in the future; otherwise, the investor would be better off investing his or her money at the market rate of return. In this example, economists say that r measures the discount rate or time value of money (i.e., the rate of return that people demand on each unit of investment). Based on typical rates of return in the private sector, analysts commonly use "real" (i.e., inflation-adjusted) discount rates on the order of 6% per year in evaluating projects with average degrees of risk.

More generally, consider a project that would yield a stream of costs and benefits C_t and B_t in years $t = 0, 1, \ldots, T$. In this formulation, year 0 represents the present, whereas the project yields impacts that extend a total of T years into the future. Then the project is considered financially attractive if its net present value (NPV), calculated according to the formula

$$NPV = \sum_{t=0}^{T} (B_t - C_t)/(1+r)^t,$$

is positive. In this event, the project's discounted net benefits exceed the initial investment cost. This implies that the project generates a rate of return that exceeds the discount rate, thereby justifying the investment from an economic perspective.

Given a 6% discount rate, the net present value of the home insulation project described previously is

 © 2004 Elsevier Inc. All rights reserved.

$NPV = -900 + 130/1.06 + 130/1.06^2 + \ldots + 130/1.06^{30} = \$889$. Although this example is purely illustrative, it signals an important fact about real-world investments in energy efficiency: In many cases, energy-efficient technologies are cost-effective in the sense that they generate positive net benefits when evaluated using conventional discount rates.

Empirical work on a broad range of energy use technologies, however, points to an important paradox that is known as the energy efficiency gap. Replacing incandescent light bulbs with compact fluorescents, for example, typically reduces electricity use by 75% while generating simultaneous cost savings. In a similar vein, a study by the U.S. National Academy of Sciences found that the fuel economy of passenger vehicles could be increased by as much as 33% for automobiles and 47% for light trucks (pickups, minivans, and sport utility vehicles) by implementing cost-effective design options that leave vehicle size and performance unchanged. On an economywide basis, the Intergovernmental Panel on Climate Change estimates that the energy use of advanced industrialized nations could be reduced by up to 30% through the full adoption of cost-effective technologies. However, households and businesses have shown little enthusiasm for energy-efficient technologies, so market decisions seem insufficient to realize this potential.

Three interpretations of the efficiency gap have been set forth in the literature:

1. Energy-efficient technologies have "hidden costs" that are overlooked by conventional financial analysis.
2. Households and businesses use unusually high discount rates in decisions concerning energy efficiency.
3. The adoption of cost-effective, energy-efficient technologies is impeded by issues of imperfect information, "bounded rationality," or structural anomalies in markets for energy-using equipment.

The remainder of this article discusses these themes.

2. THE HIDDEN COSTS ARGUMENT

Studies of the costs and benefits of energy-efficient technologies typically employ an engineering approach that accounts for purchase price, operations and maintenance costs, and anticipated energy savings. However, energy use technologies can have intangible costs that are not easily measured. The failure to include such hidden costs introduces a potential bias or source of error in financial analysis.

Consider the case of compact fluorescent light bulbs mentioned previously. Although this technology yields well-documented direct cost savings, it involves two types of hidden costs that are well recognized in the literature. First, compact fluorescent bulbs tend to be bigger and bulkier than the conventional incandescents that they replace. Although progress has been made in addressing this problem since this technology was first mass marketed in the 1990s, the fact remains that compact fluorescents do not work well with certain lighting fixtures. Second, compact fluorescents produce a different light spectrum than incandescent light bulbs. Some people find their light to be "cold" and aesthetically inferior. Although these costs are difficult to measure in financial terms, they legitimately affect people's willingness to adopt this technology.

A second example is provided by the case of automobile fuel economy. In the United States, the energy efficiency of new vehicles is regulated under the Corporate Average Fuel Economy (CAFE) standards that were introduced in response to the Arab oil embargo of 1973. Although these standards have been effective in reducing fuel consumption, some analysts argue that car manufacturers achieved increased fuel economy by reducing the size and weight of new vehicles. Since (all else being equal) the crashworthiness of vehicles is inversely proportional to their size and weight, authors such as Crandall and Graham argue a case can be made that the CAFE standards have compromised safety, increasing highway fatalities by up to 3900 persons per year. This impact, of course, is not reflected in standard engineering studies of the economic costs of improved fuel economy, although it clearly is relevant to the purchase decisions of car buyers. To be fair, it is important to note that the energy savings potential identified by the U.S. National Academy of Sciences attempted to control carefully for this effect. However, safety considerations trump fuel economy in the eyes of many consumers.

In general, examples such as compact fluorescent light bulbs and vehicle fuel economy suggest that the hidden costs issue is a real-world phenomenon with significant implications for consumer decisions. Nonetheless, it is important to consider at least two caveats that cast doubt on the generality of this reasoning. First, certain energy efficiency measures yield intangible benefits that are typically overlooked

in conventional cost calculations. Weather stripping a leaky building shell, for example, yields direct cost savings through reduced energy use. In addition, it improves the comfort of building occupants by reducing drafts and hence the subjective sense that a building is cold. In a similar vein, the use of day lighting in passive solar building technologies can substantially enhance a building's aesthetic appeal, although this benefit is difficult to reduce to simple economic terms.

Second, there are many technologies for which changes in equipment characteristics yield energy savings without altering the quality of energy services provided to end users. In refrigerators, for example, the use of energy-efficient motors and compressors leads to pure energy savings through technical measures that have no observable effect on the equipment's size, functionality, or performance. Similarly, programming computers to enter a "sleep state" during periods of nonuse generates substantial energy savings without reducing the benefits of this technology.

The bottom line is that energy analysts should be aware of and sensitive to issues of comfort, convenience, and safety. In some instances, hidden costs can provide good reasons not to adopt a technology that yields apparent cost savings. On the other hand, there is no good evidence that hidden costs explain the overall empirical magnitude of the efficiency gap.

3. HIGH DISCOUNT RATES?

As explained previously, the discount rate reflects the rate at which people are willing to exchange present and future economic benefits. According to standard economic reasoning, discount rates are revealed by people's observed behavior in markets for savings and investment. A homeowner who takes out a loan at a 4% real (inflation-adjusted) interest rate, for example, reveals a willingness to pay back $3.24 30 years from the present to obtain just $1 dollar today. Similarly, a worker who invests $1 in a mutual fund would expect to receive $10.06 upon retiring three decades in the future given an 8% annual return.

Economists generally agree that market rates of return reveal people's discount rates. It is important to note, however, that different types of financial assets yield different rates of return. In particular, safe investments, such as money market funds, certificates of deposit, and short-term government bonds, yield real returns on the order of just 1% per year. Risky assets such as corporate stocks, in contrast, pay average real returns of 6% or more on a long-term basis. The high returns paid by stocks constitute a risk premium that rewards investors for accepting the fluctuations associated with financial markets.

If households and businesses were well informed and rational, then economic theory suggests that they would discount the benefits of energy-efficient technologies at a rate equal to the market return available on investments with similar risk. In general, the volatility of energy prices implies that the benefits of energy efficiency are uncertain. On the other hand, investments in energy efficiency reduce people's vulnerability to large swings in energy prices. In this sense, energy efficiency measures have ambiguous effects on the overall financial risk faced by households and businesses. In practical analysis, it is common to assume that investments in energy efficiency have risk characteristics similar to those associated with typical private sector investments. In theoretical terms, this assumption favors the use of a 6% discount rate.

Despite this body of theory, empirical studies have found that people behave as if they discounted benefits of energy-efficient technologies at an annual rate far in excess of 6% per year. An early study by Hausman, for example, suggested that people discounted the energy savings provided by air conditioners at a 25% annual rate. Later studies summarized by Train and Ruderman *et al*. pointed to the existence of implicit discount rates ranging from 25% to 300% for refrigerators, heating and cooling systems, building shell improvements, and a variety of other technologies. These studies employ sophisticated econometric and/or engineering–economic models to gauge the rate of time preference at which people's actual technology choices could be considered economically rational. As noted by Howarth and Sanstad, the fact that people reject efficiency technologies that yield high rates of return suggests that actual behavior may diverge from the assumption of perfect rationality.

Analysts such as Hassett and Metcalf have attempted to explain this anomaly by arguing that improvements in the state of technology over time can provide a reason to defer investing in technologies that are apparently cost-effective in the short term. The reason is that purchasing an item such as a fuel-efficient car implicitly raises the cost of buying an even more efficient vehicle a few years later. To better understand this point, consider the case of a mid-1990s computer user weighing the merits of purchasing an upgraded machine. For a cost of

$2000, the user could step up from a good system to a very good system with enhanced speed and performance. Alternatively, the user could defer the purchase for 1 or 2 years and expect to buy a far better machine for an equal or lower price. Hassett and Metcalf reasoned that this type of behavior could explain the high discount rates revealed in markets for investments in energy efficiency.

According to this perspective, the benefits of waiting imply that the early adoption of energy-efficient technologies may have a type of hidden cost (or option value) that is sometimes ignored in naive technology cost studies. An empirical analysis by Sanstad *et al.*, however, suggests that the magnitude of this effect is too small to have much impact on decisions about energy efficiency given anticipated developments in energy prices and the pace of technological change. This line of reasoning therefore appears not to explain why technologies that yield positive discounted net benefits or (equivalently) attractive rates of return are often passed up in markets for energy-using equipment.

4. IMPERFECT INFORMATION AND BOUNDED RATIONALITY

The preceding arguments attempt to explain the paradoxes surrounding discount rates and the efficiency gap in terms of the workings of what Sutherland terms normal or economically efficient markets. An alternative explanation offered by Stanstad and Howarth focuses on market barriers or market failures that impeded the adoption of cost-effective, energy-efficient technologies. In broad terms, at least two sources of market failure are emphasized in this context: imperfect information and bounded rationality.

Behavioral studies have established that households and businesses are often poorly informed about the technical characteristics of energy-using equipment. In summarizing this literature, Stern argues that people's knowledge of opportunities to save energy is "not only incomplete, but systematically incorrect. Generally speaking, people tend to overestimate the amounts of energy used by and that may be saved in technologies that are visible and that must be activated each time they are used." At a cognitive level, many people do not focus on energy trade-offs in making decisions about purchasing goods, such as appliances, new cars, or even a home. This holds true because (i) the energy requirements of most goods are not vividly observable prior to a purchase decision, and (ii) energy costs are typically only a small fraction of the total cost of equipment ownership. Hence, buyers emphasize factors such as a refrigerator's features and aesthetics, a vehicle's acceleration, and a home's location and architecture without fine-tuning their decisions to optimize energy efficiency.

The problem of imperfect information is linked to the concept of adverse selection in information economics. As Akerlof showed in his seminal "markets for lemons" article, low-quality products can crowd out high-quality products in markets in which product quality is not easily observed by consumers. In the case of the efficiency gap, the inference is that poorly informed buyers would pass up opportunities to purchase technologies that would simultaneously save energy and produce net economic benefits.

Issues of imperfect information have been partially addressed through measures such as advertising campaigns and energy labeling for appliances, vehicles, and (in recent years) new homes. According to Robinson, these efforts have typically met with limited success. One reason is that from a psychological perspective, human beings do not have the cognitive capabilities required to make perfectly informed, rational decisions in complex, real-world settings. Instead, Kempton and Layne argue that people are boundedly rational, using simple but cognitively efficient rules of thumb to make decisions in environments that are too complex to be fully understood. The literature on this subject suggests that people aim to achieve goals such as cost minimization but in practice make decisions and are flawed and (in many cases) systematically biased.

In cases in which consumer decision making is marked by imperfect information and/or bounded rationality, targeted policy interventions can often facilitate the adoption of cost-effective, energy-efficient technologies. Under the U.S. Appliance Efficiency Standards, for example, equipment manufacturers must meet minimum performance standards that are explicitly based on the criterion of cost-effectiveness. Studies by McMahon *et al.* and Geller suggest that these standards alone will yield 24 exajoules of energy and $46 billion between 1990 and 2015. These standards are set in a process of information exchange and rule making that involves a relatively small number of technology experts in government and industry. Through this process, technologies are optimized in a manner that does not occur through the engagement of buyers and sellers in the unregulated marketplace.

A second policy that has facilitated the adoption of cost-effective, energy-efficient technologies is the Green Lights program of the U.S. Environmental Protection Agency (EPA). Under Green Lights, the EPA provides technology assistance and public recognition for businesses and nonprofit organizations that voluntarily agree to implement energy-efficient lighting technologies that yield net cost savings without negative impacts on energy services. In its first 5 years of operation, EPA estimates that this program achieved $7.4 billion kWh of electricity savings that remained unexploited prior to its implementation in 1991. According to DeCanio, the typical investment carried out under Green Lights yielded a 45% annual rate of return. This example illustrates how market barriers and the efficiency gap can affect the production sector of the economy as well as how voluntary cooperation between government and the private sector can facilitate improved market performance.

Of course, programs and policies can have administrative costs that should be carefully considered in evaluating their net economic benefits. These costs are small for the U.S. Appliance Efficiency Standards and Green Lights but are quite substantial in certain situations. Joskow and Marron, for example, found that administrative costs are a major component of the total cost of reducing residential electricity use through utility-sponsored demand-side management programs. In addition, regulations that depart from the goal of cost-effectiveness can have unintended negative impacts on product characteristics and consumer choice. This issue is illustrated by the possible increase in highway fatalities caused by the CAFE standards imposed on new vehicles in the 1970s.

Nonetheless, the empirical literature on the efficiency gap suggests that there is ample scope to improve energy efficiency through well-designed policy interventions. In this context, using the net present value criterion and conventional discounting procedures provides a mechanism to identify programs and policies that would enhance consumer welfare and economic efficiency, provided that all of the relevant costs and benefits are taken into account.

SEE ALSO THE FOLLOWING ARTICLES

Cost–Benefit Analysis Applied to Energy • Economics of Energy Efficiency • Energy Efficiency, Taxonomic Overview • Industrial Energy Efficiency • Obstacles to Energy Efficiency • Vehicles and Their Powerplants: Energy Use and Efficiency

Further Reading

Akerlof, G. (1970). The market for lemons: Quality uncertainty and the market mechanism. *Q. J. Econ.* **89,** 488–500.

Crandall, R. W., and Graham, J. D. (1989). The effect of fuel economy standards on automobile safety. *J. Law Econ.* **32,** 97–118.

DeCanio, S. J. (1998). The efficiency paradox: Bureaucratic and organizational barriers to profitable energy-saving investments. *Energy Policy* **26,** 441–454.

Geller, H. (1997). National appliance efficiency standards in the USA: Cost-effective federal regulations. *Energy Buildings* **26,** 101–109.

Hassett, K. A., and Metcalf, G. (1993). Energy conservation investment: Do consumers discount the future correctly? *Energy Policy* **21,** 710–716.

Hausman, J. A. (1979). Individual discount rates and the purchase and utilization of energy-using durables. *Bell J. Econ.* **10,** 33–54.

Howarth, R. B., and Sanstad, A. H. (1995). Discount rates and energy efficiency. *Contemp. Econ. Policy* **13,** 101–109.

Intergovernmental Panel on Climate Change (1996). "Climate Change 1995: Economic and Social Dimensions of Climate Change." Cambridge Univ. Press, New York.

Interlaboratory Working Group (2000). "Scenarios for a Clean Energy Future." Oak Ridge National Laboratory, Oak Ridge, TN.

Joskow, P. L., and Marron, D. B. (1992). What does a negawatt really cost? Evidence from utility conservation programs *Energy J.* **13**(4), 41–74.

Kempton, W., and Layne, L. (1994). The consumer's energy analysis environment. *Energy Policy* **22,** 857–866.

McMahon, J. E., Berman, D., Chan, P., Chan, T., Koomey, J., Levine, M. D., and Stoft, S. (1990). Impacts of U.S. appliance energy performance standards on consumers, manufacturers, electric utilities, and the environment. *In* "ACEEE Summer Study on Energy Efficiency in Buildings," pp. 7.107–7.116. American Council for an Energy Efficient Economy, Washington, D.C.

Robinson, J. B. (1991). The proof of the pudding: Making energy efficiency work. *Energy Policy* **7,** 631–645.

Ruderman, H., Levine, M. D., and McMahon, J. E. (1987). The behavior of the market for energy efficiency in residential appliances including heating and cooling equipment. *Energy J.* **8,** 101–124.

Sanstad, A. H., and Howarth, R. B. (1994). Normal markets, market imperfections, and energy efficiency. *Energy Policy* **22,** 811–818.

Sanstad, A. H., Blumstein, C., and Stoft, S. E. (1995). How high are options values in energy-efficiency investments? *Energy Policy* **23,** 739–743.

Stern, P. C. (1986). Blind spots in policy analysis: What economics doesn't say about energy use. *J. Policy Anal. Management* **5,** 200–227.

Sutherland, R. J. (1991). Market barriers to energy-efficiency investments. *Energy J.* **12,** 15–34.

Train, K. (1985). Discount rates in consumers' energy-related decisions: A review of the literature. *Energy* **10,** 1243–1253.

U.S. Environmental Protection Agency (1997). "Building on Our Success: Green Lights and Energy Star Buildings 1996 Year in Review." U.S. Environmental Protection Agency, Office of Air and Radiation, Washington, DC.

U.S. National Academy of Sciences (2002). "Effectiveness and Impact of Corporate Average Fuel Economy (CAFE) Standards." National Academy Press, Washington, DC.

Distributed Energy, Overview

NEIL STRACHAN
Pew Center on Global Climate Change
Arlington, Virginia, United States

1. **Introduction**
2. **Definition of DE**
3. **Historical Development of DE**
4. **DE Market Status**
5. **Overview of DE Technologies**
6. **Economic Characteristics of DE**
7. **Environmental Characteristics of DE**
8. **Other Characteristics of DE**
9. **Future Role of DE**

Glossary

buyback tariff The price paid to the owner of the distributed energy (DE) resource for electricity sold back to the electricity grid. The price paid depends on how much and when power is exported, how regularly it is guaranteed, and where in the distribution system the DE resource is located (and hence the impact on the operation of the network).

capacity factor Also known as load factor. Electricity delivered during a time period divided by the rated electricity capacity.

combined heat and power (CHP) Also known as cogeneration. The use of waste heat to provide steam or hot water, or cooling via an absorption chiller. This can raise the overall efficiency to over 80%

efficiency Electric efficiency is the total electricity output divided by the total fuel input taking into account generation efficiency and transmission losses. Total efficiency includes the use of waste heat.

heat to power ratio (HPR) The ratio between heat use and electricity generation in a combined heat and power application. For example, if electric efficiency and total efficiency are 30 and 80%, respectively, then HPR = (80−30)/30, or 1.67.

higher heating value (HHV) Includes the heat of vaporization of water. For natural gas, HHV efficiency = lower heating value efficiency/1.1.

installation cost The cost necessary to implement a distributed energy technology or resource in an application. Includes technology cost (dollars per kilowatt), engineering studies, permitting, interconnection, and setup expenses.

interconnection The link between a distributed energy generator and the load being served by the utility electricity network

lower heating value (LHV) Assumes the heat of vaporization of water cannot be recovered. For natural gas, LHV efficiency = higher heating value efficiency × 1.1.

modularity Increments of new capacity that can be simply and flexibly added or removed to change the unit's energy output.

operation and maintenance costs (O&M Costs) These are divided into fixed costs and variable costs depending on the relation to hours of operation and capacity factor.

photovoltaic (PV) cells Also known as solar cells. A distributed energy technology that converts sunlight directly into electricity.

Distributed energy (DE) represents an alternative paradigm of generating electricity (and heat) at, or close to, the point of demand. DE includes fossil technologies (fuel cells, microturbines, internal combustion engines and Stirling engines), renewable technologies (photovoltaics and wind turbines), and energy storage options. Potential DE advantages include higher efficiency and lower cost through waste heat recovery and avoidance of transmission and distribution, reduced global air pollutants, enhanced flexibility of electricity grids, reduced investment uncertainty through modular increments of new capacity, and greater system and infrastructure reliability and security. Potential DE disadvantages include higher costs through loss of economies of scale, higher local air pollution near population areas, economic losses from stranded capital investments in generation and distribution, and increased reliance on natural gas.

1. INTRODUCTION

Distributed energy (DE) represents an alternative paradigm of generating electricity (and heat) at

 © 2004 Elsevier Inc. All rights reserved.

or close to the point of demand. DE includes fossil technologies—fuel cells, microturbines, internal combustion engines, and Stirling engines; renewable technologies—photovoltaics and wind turbines; energy storage options; and energy efficiency. Potential DE advantages include higher efficiency and lower cost through waste heat recovery and avoidance of transmission and distribution, reduced global and local air pollutants, enhanced flexibility of electricity grids, reduced investment uncertainty through modular increments of new capacity, and greater system and infrastructure reliability and security. Potential DE disadvantages include higher costs through loss of economies of scale, higher local air pollution near areas of population, economic losses from stranded capital investments in generation and distribution, and increased reliance on natural gas.

DE has generally been considered and studied in the context of niche applications, as emergency back-up power, or as limited to a small portion of grid-connected electricity supply. In many countries, the economies of scale of centralized generation, the low price of coal as a fuel for electricity generation, and regulatory barriers or disincentives to on-site generation have precluded the widespread adoption of DE. These institutional barriers have included lack of interconnection protocols, low electricity buy-back tariffs, and little consideration of the system's benefits of distributed resources. However, changes in the relative economics of centralized versus distributed energy, the increasing use of natural gas, restrictions on new electricity transmission lines, recognition of the environmental benefits of DE, and improved DE control technologies have resulted in the reconsideration of the widespread use of DE. For example, the Netherlands has seen the greatest penetration of DE technologies. By the year 2000, 6% of national electricity capacity was DE units of <1 MWe and 35% of national electricity capacity was DE units of <50 MWe.

This article gives an overview of DE. Sections 2 and 3 define DE and give an historical context. Sections 4 and 5 summarize the market status of DE and the major DE technologies. Sections 6, 7, and 8 summarize the characteristics of DE technologies including their economic, environmental, and reliability performance, as well as compatibility with existing electricity infrastructure and investment requirements. Finally, Section 9 details the opportunities and barriers to future DE deployment.

2. DEFINITION OF DE

DE has been defined by the size of the technology or by specific technologies. It has also been defined on the basis of connection (connected to either the electricity distribution system or, in some cases, the medium-voltage network), dispatch (not centrally dispatched), whether it is located on the customer side of the meter, and noninclusion in system production optimization studies.

However, defining DE by its relationship to the electricity network ensures that DE is a function of the characteristics of that network. For example, a country dominated by very large centralized units, such as France, may define DE as <100 MWe, but another system configuration, such as that in the Netherlands, may define DE as <1 MWe, or two orders of magnitude smaller. The size range of DE technologies thus varies from a few watts to approximately 50 MWe.

A preferable definition is as follows: Distributed energy represents an alternative paradigm of generating electricity (and heat) at, or close to, the point of demand. This definition allows discussion across national electricity infrastructures and across demand sectors, as various applications may require different characteristics of energy supply and different institutional arrangements may require different coupling to the electricity network. Second, it allows for DE to cover a range of sizes of generation for residential through industrial applications. Third, it emphasizes the opportunity for energy efficiency through the use of waste heat, which cannot be transported across distances outside the immediate area. Fourth, it allows for the consideration of a host of current and future technologies and, crucially, also energy storage and energy efficiency in a broad definition of DE. And finally, this definition does not constrain DE to specific functions within current energy delivery networks. These functions include standby power, peak load shaving, remote sites, and grid support. However, an energy infrastructure based solely on DE, bypassing the need for large-scale electricity generation, is feasible and may be desirable.

3. HISTORICAL DEVELOPMENT OF DE

The shift toward DE can in many ways be viewed as power generation coming full circle. In the 1880s, Thomas Edison's Pearl Street electricity system

serving Wall Street and the surrounding city blocks was distributed power. This paradigm was continued over the next 20 years with schemes in the United States and across the world serving limited urban areas with small-scale direct current (DC) systems. Also popular were installations serving individual factories and supplying electricity and heat in combined heat and power (CHP) applications. One large drawback of using DC was the large losses from this low-voltage system when transmitting power over distance. An alternative system based on alternating current (AC) was promoted by a number of competitors, including George Westinghouse. AC systems had the advantage when serving larger and more spread-out service areas, as the voltage could be stepped up using a transformer to minimize losses. The superiority of the AC system was ensured when, in the late 1880s, Nikola Tesla invented the three-phase AC arrangement, which simplified the number and size of wires, and in 1893 when the universal system, a series of technical standards, allowed the interconnection of AC and DC systems and users.

With technology facilitating AC systems, a significant financial impetus for larger, more centralized systems was the requirement for consistently high-load factors on the electricity network. Large amounts of electricity cannot be stored, thus requiring that demand meets supply at all times. Economical use of the transmission and generation capacity thus requires that a diverse customer base maintain electricity demand through the days and seasons. A combination of residential, commercial, and industrial applications enables the demand profile to be smoothed out as much as possible and this generally requires a large service area.

Another stimulus of larger electricity networks was the availability of energy inputs. Hydroelectric power and coal mines for electricity generation are usually located some distance from population and industrial centers. Either the fuel source was transported to distributed power plants (an impossibility in the case of hydroelectricity sites) or electricity was transmitted at high voltages.

An additional driving force of the centralized electricity network was the rise of natural monopolies. Although the institutional structure of emerging systems was very different, ranging from private companies to state-owned enterprises, competition in electricity generation was viewed as infeasible due to the prohibitive costs of laying competing wires and the institutional difficulties (at that time) of being able to verify and charge for use of an electricity network by different firms and organizations. By the 1930s, industrialized countries had set up large electricity utilities, coalescing around the dominant generation design of the steam turbine. Smaller systems were either absorbed or shut down.

This centralized paradigm led to the developments of larger and larger generating plants as economies of scale were pursued to raise total electricity system efficiencies and reduce costs. The size of the largest generating units jumped from 80 MWe in 1920, to 600 MWe in 1960, to 1400 MWe in 1980. In order to finance these enormous capital investments in generation and related transmission, utility monopolies relied on guaranteed revenue streams from predetermined electricity prices to their "locked-in" customer base. Other advantages of this system included reliability due to the integrated nature of the electricity network and reductions of local air pollution near population centers as generation took place in remote areas. In addition, large-scale nuclear power stations were well suited to this paradigm as they required the enormous initial investment and guaranteed investment recovery promised by the centralized electricity network paradigm.

The turning point of the return to DE came with the oil shocks of the 1970s and the drive toward higher efficiencies. Despite economies of scale, overall electricity system efficiencies were restricted to 30–33% of input energy. A major reason for this was the inability to capture the large amounts of waste heat during the electricity production process and hence renewed efforts were made toward technological and regulatory arrangements to allow distributed CHP facilities with overall efficiencies of greater than 80%. The most impressive technological development was in high-efficiency combined-cycle turbines (CCGT) using natural gas. This technology became the mode of generation of choice, capturing most incremental capacity growth in most markets. Progress on smaller DE technologies (see Section 5) included significant cost, reliability, and emissions improvements of engines, microturbines, fuel cells, and photovoltaics (PV). In parallel, and especially in the past 10 years, there have been developments in control and monitoring technologies. The telecommunications revolution has enabled cheap, real-time operation and integration of small-scale power technologies.

Institutional progress has centered around the drive to restructure electricity markets, from monopolies (often publicly owned) to a competitive generation market in which market participants are free to choose technology type and their customer base. The motivations for this shift were to lower power costs and improve customer choice and also to raise revenues in the case of state-owned enterprises.

A major consequence of this market restructuring was the increase in volatility and the resultant uncertainty in recovering investment costs. Firms were less and less willing to commit huge sums in centralized plants with repayment occurring over decades. The average size of a new generating unit in the United States declined from 200 MWe in the mid-1980s, to 100 MW in 1992, and to 21 MWe in 1998. This is approximately equivalent to the average unit sizes in the years 1915–1920.

Additional incentives have stemmed from changing public pressures. It has become increasingly difficult to build additions to electricity networks, especially near population centers. In addition, the demand for very reliable and higher quality power has emerged, spurred by the rapid penetration of electronic appliances and industrial processes that are extremely sensitive to supply disruptions or degradations. Finally, other energy delivery infrastructures, especially the natural gas system, have emerged to compete with electricity transmission. Energy must be moved from source to demand, but where the conversion to electricity is carried out, and at what scale, will be part of the emerging paradigm to come. This is particularly relevant to developing countries where it may or may not be optimal to construct a capital-intensive centralized electricity network.

4. DE MARKET STATUS

It is extremely difficult, or even impossible, to provide values for the total installed capacity of DE technologies and resources. This is due to the range of technologies, their varying levels of use in different countries, and ambiguity over what constitutes a DE application as unit sizes become larger and units are sited further from demand centers. In addition, some of the most promising technologies, including fuel cells, photovoltaics, and microturbines, either are just entering commercialization or are used only in niche applications. The relative rate of technological progress and the regulatory structure of energy markets will determine to what degree and how quickly DE will penetrate the mainstream market. In addition, fuel constraints, stemming from the availability of natural gas, hydrogen, or other fuel networks, and the available renewable energy resource may limit some of the spread of DE technology.

An example of DE installed capacity from the very broadest definition is the DE database of Resources Dynamics Corporation, which as of 2003 lists DE capacity in the United States at 10.7 million DE units with an aggregate capacity of 168 GW. This constitutes 19.7% of the total U.S. installed capacity, although the definitions in this database include back-up and energy units, remote-application small-scale technologies, and a range of larger CHP units.

Better documented market penetration totals and rates can be given for some countries and DE technologies to elucidate the current and potential global DE market size. The Netherlands has the largest market percentage of DE technologies in the world. During the 1980s and 1990s, this was primarily in the form of natural gas-fired engines of less than 2 MWe each in size. By the end of the 1990s, this DE technology was found in over 5000 installations with 1500 MWe of installed capacity. This represents over 6% of the total national capacity.

The Dutch DE units are primarily operated as base-load, CHP applications. If larger gas turbine units are also considered as DE, then the total DE capacity rises to 7500 MWe or 30% of national capacity. Other DE technologies, including microturbines, fuel cells, and Stirling engines, are being introduced as relative costs, emission characteristics, reliability, and power quality considerations are traded off in different applications.

The range of typical DE applications in the Netherlands is given in Fig. 1. The largest sector was commercial horticulture, with greenhouses providing an ideal DE–CHP application with demands for heat as well as electricity for artificial lighting. The Netherlands has many DE units in various commercial building sectors, including hospitals, industry, multiresidential complexes, industrial sites, and sewage operation facilities, where they can also utilize waste gases.

Other countries, including the United Kingdom, Germany, Japan, and the United States, have also seen considerable penetration of internal combustion (IC) engines as a DE resource. These can be either diesel engines, which are often used for back-up and emergency power (as is common in the United States), or gas engines, which are used in CHP applications and can replace the need for an on-site heat boiler plant, and, depending on the regulatory structure and buy-back tariffs, can provide electricity back to the grid. Figure 2 details the numbers and capacity of total worldwide sales of engines that are 1 to 5 MWe each in size, showing both the growth in overall sales and the trend toward gas-fired base-load DE units.

Finally, regarding a nonfossil DE technology, Fig. 3 charts the shipments of PV units in the United States. The largest sectors have traditionally been in applications that are incorporated into buildings to

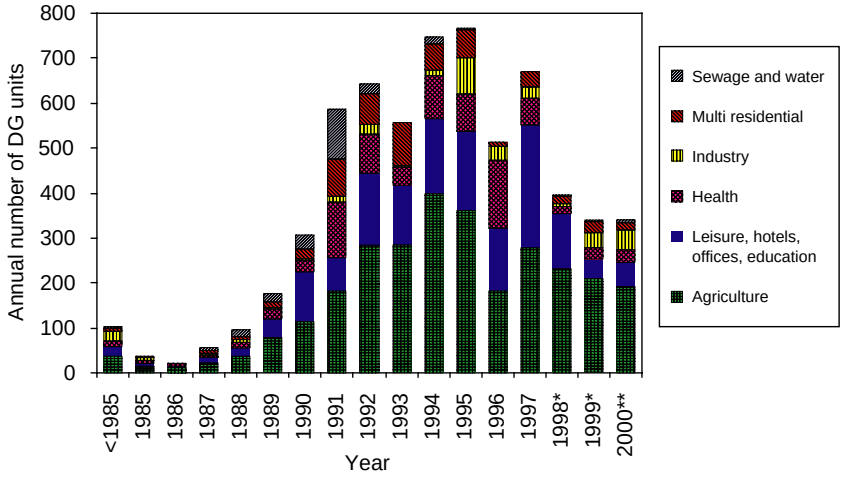


FIGURE 1 Gas IC DG (distributed generation; equivalent to distributed energy) installations by sector in the Netherlands. Years marked with an asterisk are provisional data; those marked with two asterisks are supply firms' market estimates. From Central Bureau of Statistics (1998). "Decentralized Energy Statistics of the Netherlands." Voorborg/Herleen, The Netherlands; and Cogen Nederland (1999). "Nederland Cogeneration Statistical Information." Driebergen, The Netherlands.

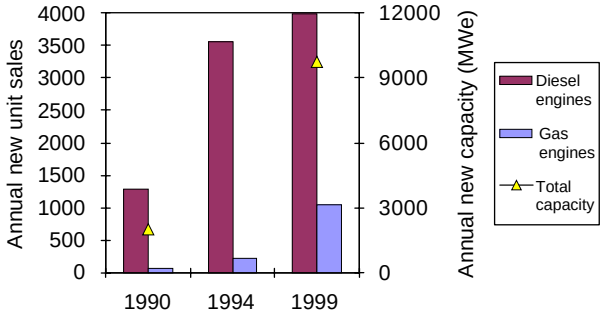


FIGURE 2 Worldwide new unit sales and capacity of DE engines. From Diesel and Gas Turbine Worldwide Power Generation Survey, available at http://www.dieselpub.com.

provide water heating and electrical requirements and in remote applications, such as communication facilities. However, grid-connected PV units have seen the fastest growth.

5. OVERVIEW OF DE TECHNOLOGIES

DE technologies have a range of technical characteristics, making different technologies more suitable in specific applications, depending on whether the determining criterion is electrical output, provision of usable heat, fuel availability, reliability, or emission levels, to name just a few criteria. However, all DE technologies have the capacity to be remotely operated and controlled. This is achieved through monitoring a range of operating parameters via a communication link and allows simplified maintenance, avoids unexpected shutdowns, and eases the grid control issues that arise from locating DE units within the electric distribution network. Another common characteristic of DE technologies is that following factory fabrication they are straightforward to install, thus minimizing implementation costs.

A useful parameter when comparing DE technologies is the heat to power ratio (HPR). This is the ratio between heat use and electricity generation in a CHP application. For example, if electric and total efficiencies are 30 and 80%, respectively, then HPR = (80−30)/30, or 1.67. Depending on the application, electrical or thermal output may be valued or required. Also, note that all electrical efficiencies in this article are quoted as higher heating value (HHV). This includes the heat of vaporization of water. For natural gas, HHV efficiency = Lower Heating Value efficiency/1.1.

This section provides a brief overview of the major DE technologies, discusses their principal strengths and weaknesses, and provides a quantitative comparison of their technical characteristics (see Tables I and II). This section is designed to be illustrative as site-specific variation, differing models, and continuing technical progress will alter these characteristics. Economic and environmental comparisons are detailed in Sections 6 and 7.

5.1 Engines

Reciprocating IC engines are an established and well-known technology. Engines for stationary power are

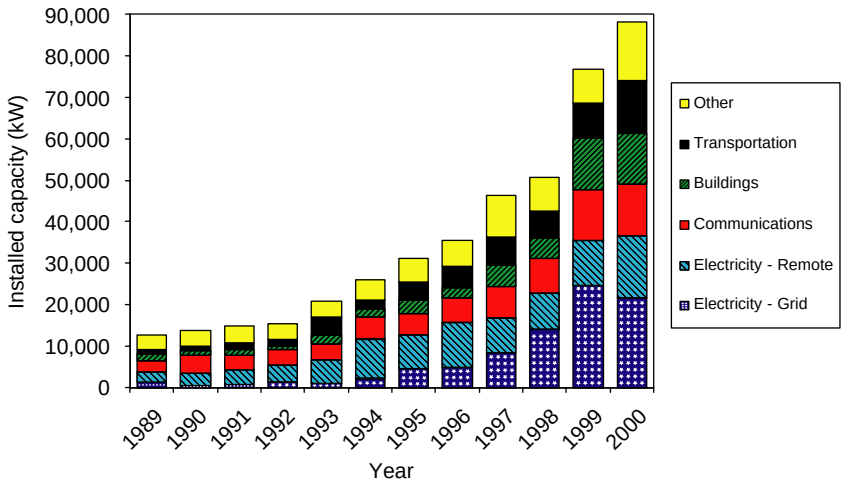


FIGURE 3 U.S. photovoltaic shipments. From Energy Information Administration (2001), Renewable Energy Annual, Table 10, U.S. Department of Energy, Washington, DC.

TABLE I

Qualitative Strengths and Weaknesses of DE Technologies

Technology	Strengths		Weaknesses	
Reciprocating engines	Low capital cost	Fuel flexibility	High air pollutant emissions	Noisy
	High reliability	Quick startup	Frequent maintenance	Thermal output at low temperatures
	Established technology			
Gas turbines	Proven reliability and availability	Low capital cost	Reduced efficiencies at part load	Sensitivity to ambient conditions (temperature, altitude)
	Established technology	High-temperature steam	Required maintenance	
	Larger sizes available			
Microturbines	Compact size	Lightweight	Low electricity efficiencies	Sensitivity to temperature, altitude
	Low emissions	Can utilize waste fuels		
Fuel cells	High electrical efficiencies	Low emissions	High capital costs	Fuel infrastructure constraint
	Quiet	Modular	Reforming natural gas	
	Operates well at part load	High reliability		
	Synergy with vehicle power trains			
Stirling engines	Low noise and vibrationless operation	Low emissions	High capital costs	Low electrical efficiency
	Low maintenance and high reliability	Multifuel capability, including solar power	Durability unresolved	
	Long life			
PV	Low maintenance	No emissions	High capital costs	Intermittency
	Free fuel		By-products of manufacture and disposal	
Wind	Low maintenance	No emissions	High capital costs	Variable power output
	Free fuel		Visual impact	

TABLE II

Quantitative Technical Comparison of DE Technologies

	Reciprocating engines		Turbines		Fuel cells						
	Gas	**Diesel**	**Gas**	**Micro**	**MCFC**	**PAFC**	**PEMFC**	**SOFC**	**Stirling engine**	**PV**	**Wind**
Commercialized	Yes	Yes	Yes	Yes	No	Yes	No	No	No	Yes	Yes
Size (kWe)	30–1000	50–2000	500–50,000	30–500	250–3000	200–500	10–250	200–2000	1–30	10–1000	10–10,000
Electrical efficiency (%)	27–30	28–35	24–38	20–30	48–55%	33–40%	33–40%	42–58%	20–30		
CHP	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No
Thermal temperature (°F)	300–500	190–900	500–1100	400–650	170–710	140–250	140–170	350–420	150	—	—
HPR (85% total efficiency)	2.15–1.83	2.04–1.43	2.54–1.24	3.25–1.83	0.77–0.55	1.58–1.13	1.58–1.13	1.02–0.47	3.25–1.83	—	—
Dispatchable	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No
Fuels	Gas	Diesel	Gas, waste fuels	Gas, waste fuels	Natural gas, H_2	Natural gas, H_2	Natural gas, H_2	Natural gas, H_2	Any	Solar	Wind
Annual maintenance (h/year)	20–250	20–250	50–700	1–20	120	50	10	40	5–10	0	0
Hours between maintenance	500–1000	1500–2000	4000–8000	750–10,000	800	2200–8700	8700	8700	8700	1–40	4000–30,000
Startup time (min)	0.02–0.2	0.2	1–100	0.5–1	1200–2400	180	60	2	0.5–1	0	0.1–10
Physical size/power output (ft^2/kW)	0.22–0.3	0.22	0.2–0.6	0.15–1.5	1–4	4	0.6–3	1.2	1–3	400–500	0.25–110
Noise (dB) (without acoustic enclosure)	60–85	80–90	60–85	50–60	60	62	50–65	60	40–50	0	55–65

Note. MCFC, molten carbonate fuel cell; PAFC, phosphoric acid fuel cell; PEMFC, proton-exchange membrane fuel cell; SDFC, solid oxide fuel cell.

derived from automobile engines. These engines are fired by either natural gas or diesel and use spark ignition or compression ignition depending on the compression ratio used. A typical energy balance is electricity, 26–39% (this range is due to type, size, and operation of engine); useful heat, 46–60%; losses (radiation from engine and exhaust, lubrication, gearbox, generator), 10–20%.

Power production engines are typically less than 1 MWe and have become the most common DE technology for both standby and peaking applications and are increasing in use as base-load combined heat and power units. Engines have low capital costs and high reliability, although they require a regular maintenance and overhaul schedule. Regular maintenance is essential for good performance and ranges from weekly oil changes to major overhauls after 25,000–40,000 h of operation. Heat recovery is from the engine jacket and cylinder head cooling water and from the hot engine exhaust, resulting in hot water at 80–120°C. A combination of combustion modifications and the use of catalytic converters has greatly reduced local air pollutant emissions, although in this regard engines are still inferior to other DE technologies.

5.2 Gas Turbines

Derived from aero-applications, power generation with gas turbines is commonplace, with an extensive installed capacity and sizes ranging from 500 kWe to ≥ 50 MWe. Low maintenance costs and high-grade heat recovery have made gas turbines a favorite in industrial DE applications.

The unit has a turbine and compressor on the same shaft and compressed combustion products at high temperature drive the turbine. Compressor losses and materials restrictions inhibit the upper temperature of the cycle and thus constrain efficiency. Hot exhaust gases are used to preheat incoming air (regeneration). The remaining thermal energy of these exhaust gases can be reclaimed as waste to heat a boiler. A typical energy balance is as follows: electricity, 30%; high-grade heat, 50%; stack losses, 14.5%; other losses, 6.5%. Turbines are classified as temperature-limited cycles and increases in efficiency continue to occur through materials developments and cooling technologies to allow the turbine to run at higher maximum temperatures.

5.3 Microturbines

Microturbines follow the same cycle as conventional gas turbines, although they are at a less developed stage of commercial development. Their development has come from the design of small, very high speed turbines (with compressors on the same shaft) rotating up to 100,000 rpm. Additional scaling down of components, particularly nozzles and burners, has been achieved. The marketed size range is 30–500 kWe.

Microturbines use regeneration, with resultant lowered exhaust gas temperatures (~ 300°C). The unit is air-cooled. Maintenance requirements are reduced due to the elimination of an oil-based lubricating system in favor of air bearings and because the microturbines have no gearbox. Efficiencies are a few percentage points less than those of conventional gas turbines, due primarily to lower operating temperatures.

5.4 Fuel Cells

Fuel cells convert chemical energy directly into electrochemical work without going through an intermediate thermal conversion. Thus, the second law of thermodynamics does not apply, they are not limited by Carnot's theorem, and, therefore, they offer the potential of very high electrical efficiencies. In a fuel cell, the fuel (H_2) and the oxidizer (O_2) are supplied continuously (unlike in a battery). A dynamic equilibrium is maintained, with the hydrogen being oxidized to water with electrons going around an external wire to provide useful electrical work. A major limitation of fuel cells is that direct efficient oxidation of natural gas (CH_4) is not yet possible. Therefore, a reformer is used to convert CH_4 to H_2, with resulting losses in overall efficiency and creation of CO_2 as a by-product.

There are four main types of fuel cells in development. Fuel cells can either use additional heat for the reformer or their own waste heat. They are compared in Table III.

Fuel cells have many admirable qualities for power production including a high potential efficiency, no loss in performance when operating at partial load, ultralow emissions, and quiet operation. Their high capital costs are a major barrier to development.

5.5 Stirling Engines

Stirling engines are an external combustion engine, where the fuel source is burned outside the engine cylinder. This energy source drives a sealed inert working fluid, usually either helium or hydrogen, which moves between a hot chamber and a cold chamber. Stirling engines can be very small in scale (as little as 1 kWe), which in combination with very

TABLE III

Categorization of Major Fuel Cell Types

Type	Operating temperature	Reforming	Efficiency (HHV)	Comments
Phosphoric acid (PAFC)	200°C	External heat required	33–40%	Commercialized
Proton-exchange membrane (PEMFC)	80–100°C	External heat required	33–40%	Poisoned by traces of CO
Molten carbonate (MCFC)	600–650°C	Internal heat	48–55%	Internal reforming can cause variable H_2
Solid oxide (SOFC)	900–1000°C	Internal heat	42–58%	Current R&D favorite

low maintenance, low noise, and low emissions makes them ideal for residential CHP applications. Another advantage of Stirling engines is that they can utilize almost any fuel, including direct solar energy. However, significant challenges remain in reducing the high capital costs and overcoming remaining technical barriers, primarily related to durability.

5.6 Photovoltaics

PV cells, or solar cells, convert sunlight directly into electricity. PV cells are assembled into flat plate systems that can be mounted on rooftops or other sunny areas. They generate electricity with no moving parts, operate quietly with no emissions, and require little maintenance. An individual photovoltaic cell will typically produce between 1 and 2 W. To increase the power output, several cells are interconnected to form a module. Photovoltaic systems are available in the form of small rooftop residential systems (less than 10 kWe), medium-sized systems in the range of 10 to 100 kWe, and larger systems greater than 100 kWe connected to utility distribution feeders.

Two semiconductor layers in the solar cell create the electron current. Materials, such as silicon, are suitable for making these semiconducting layers and each has benefits and drawbacks for different applications. In addition to the semiconducting materials, solar cells consist of two metallic grids or electrical contacts. One is placed above the semiconducting material and the other is placed below it. The top grid or contact collects electrons from the semiconductor and transfers them to the external load. The back contact layer is connected to complete the electrical circuit.

Commercially available PV modules convert sunlight into energy with approximately 5 to 15% efficiency. Efforts are under way to improve photovoltaic cell efficiencies as well as reduce capital costs. Considerable attention is also being given to fully building-integrated PV cells, where the PV cells are an alternative to other construction materials. A principal drawback of PV cells is their reliance on an intermittent power source.

5.7 Wind Turbines

Wind-based power generation may or may not be a DE technology, depending on whether it is located near the demand source or at a remote wind farm. Generally, wind turbines are located in areas with good winds and have annual capacity factors ranging from 20 to over 40%. The typical life span of a wind turbine is 20 years. Maintenance is required at 6-month intervals.

A wind turbine with fan blades is placed at the top of a tall tower. The tower is tall in order to harness the wind at a greater velocity and to be free of turbulence caused by interference from obstacles such as trees, hills, and buildings. As the turbine rotates in the wind, a generator produces electrical power. A single wind turbine can range in size from a few kilowatts for residential applications to greater than 5 MW.

Drawbacks of wind power include high capital costs and reliance on an intermittent resource.

5.8 Energy Storage

Energy storage technologies produce no net energy but can provide electric power over short periods of time. The principal storage options include the following.

5.8.1 Battery Storage

The standard battery used in energy storage applications is the lead–acid battery. A lead–acid battery reaction is reversible, allowing the battery to be reused. There are also some advanced sodium–sulfur, zinc–bromine, and lithium–air batteries that are

nearing commercial readiness. Batteries suffer from limited storage capacity and environmental concerns over their manufacture and disposal processes.

5.8.2 Flywheel

A flywheel is an electromechanical device that couples a motor generator with a rotating mass to store energy for short durations. Conventional flywheels are "charged" and "discharged" via an integral motor/generator. The motor/generator draws power provided by the grid to spin the rotor of the flywheel. Traditional flywheel rotors are usually constructed of steel and are limited to a spin rate of a few thousand revolutions per minute and hence offer limited storage capacity.

5.8.3 Superconducting Magnetic Energy Storage

Superconducting magnetic energy storage (SMES) systems store energy in the field of a large magnetic coil with DC flowing. It can be converted back to AC electric current as needed. Low-temperature SMES cooled by liquid helium is commercially available. High-temperature SMES cooled by liquid nitrogen is still in the development stage and may become a viable commercial energy storage source in the future. SMES systems are large and generally used for short durations, such as utility switching events.

5.8.4 Supercapacitor

Supercapacitors (also known as ultracapacitors) are DC energy sources and must be interfaced to the electric grid with a static power conditioner. Small supercapacitors are commercially available to extend battery life in electronic equipment, but large supercapacitors are still in development.

5.8.5 Compressed Air Energy Storage

Compressed air energy storage uses pressurized air as the energy storage medium. An electric motor-driven compressor is used to pressurize the storage reservoir using off-peak energy and air is released from the reservoir through a turbine during peak hours to produce energy. Ideal locations for large compressed air energy storage reservoirs are aquifers, conventional mines in hard rock, and hydraulically mined salt caverns. Air can be stored in pressurized tanks for small systems.

5.9 Hybrid Systems

A combination of various DE technologies or inclusion of energy storage options (including batteries and flywheels) has been proposed to overcome specific limitations (e.g., intermittency) and to boost primary performance characteristics (e.g., electrical efficiency). Several examples of hybrid systems include the following:

- Solid oxide fuel cell (SOFC) combined with a gas turbine or microturbine;
- Stirling engine combined with a solar dish;
- Wind turbines with battery storage and diesel backup generators;
- Engines combined with energy storage devices such as flywheels.

The SOFC/gas turbine hybrid system can provide electrical conversion efficiencies of 60 to 70%, through a combined cycle where the waste heat from the fuel cell reaction drives a secondary microturbine. Stirling engine/solar dish hybrid systems can also run on other fuel sources during periods without sunlight. Wind turbines can be used in combination with energy storage and some type of backup generation (e.g., an IC engine) to provide a steady power supply to remote locations not connected to the grid. Energy storage devices such as flywheels are being combined with IC engines and microturbines to provide a reliable backup power supply. The energy storage device provides ride-through capability to enable the backup power supply to get started. In this way, electricity users can have an interruption-free backup power supply.

6. ECONOMIC CHARACTERISTICS OF DE

6.1 DE Technology Comparison

Table IV details the major cost parameters of DE technologies. These values are intended only to be illustrative due to the site-specific and application-specific factors that alter individual plant costs. Installed cost covers the capital price of the unit plus implementation costs to connect the unit to its electric load and integrate it with its fuel infrastructure, the electricity network, and, if required, the heating system of the building. Operation and maintenance (O&M) costs cover regular maintenance and scheduled overhauls. Fuel costs are based on typical U.S. prices for industrial customers and do not include any additional mark-up from distribution to commercial or residential sites.

As Table IV shows, the more mature DE technologies, engines, gas turbines, and wind turbines, have

TABLE IV

Economic Comparison of DE Technologies

	Installed cost ($/kWe)	O&M cost (cent/kWh)	Input fuel cost (cent/kWh)
Reciprocating engines			
Gas	600–700	0.8–1.4	0.7–1.7
Diesel	300–400	0.7–1.2	0.8–1.5
Gas turbines			
Simple	500–700	0.2–0.8	0.7–1.7
Micro	800	0.3–1	0.7–1.7
Fuel cells			
MCFC	2500–3000	0.3	0.7–1.7
PAFC	3000	0.3–0.5	0.7–1.7
PEMFC	3000–4000	0.1–0.4	0.7–1.7
SOFC	2500–3000	0.2–0.3	0.7–1.7
Stirling engines	2500–4000	0.1	0–1.5
PV	5000–8000	0	0
Wind	850–2500	0.1	0

Note. MCFC, molten carbonate fuel cell; PAFC, phosphoric acid fuel cell; PEMFC, proton-exchange membrane fuel cell; SOFC, solid oxide fuel cell.

the lowest costs. DE technologies utilizing renewable resources have the advantage of an essentially free fuel. Finally, a number of DE technologies with highly appealing technical and environmental characteristics (fuel cells and PV) have high capital costs that are a focus of considerable research and development efforts.

The more mature DE technologies are competitive with centralized electricity generation, especially if they are operated as CHP units, and the comparison is made with both electricity generation and on-site heat boilers. Any investment in DE technologies will depend on a number of site-specific factors including fuel availability, input fuel costs, average and peak power and heating requirements, and the negotiation of interconnection, emergency backup, safety, maintenance, and electricity export charges and tariffs. In addition, DE is likely to impose charges or produce benefits for the wider electricity network, including reduced use of existing assets, retention of stand-by capacity, or the defraying of costs of distribution upgrades.

DE technologies have additional characteristics that influence their economic attractiveness. Whereas centralized units benefit from economies of scale from larger plant size, DE technologies achieve economies of scale through mass production. This effect is expected to be enhanced by the phenomena of learning by doing and of learning by using in the production and use by early adopters. In addition, DE technologies that require significant maintenance benefit from economics of geographic scale. As a greater density of DE units is installed in a given area, the cost of maintaining each unit falls because a significant proportion of maintenance costs arises from on-site visits to maintain and repair the DE unit. Taken together, as more and more DE units penetrate the market, the economic costs of any single unit will fall, leading to higher levels of dispersal and a cycle of improving economics.

A number of factors will limit this cycle of improving costs, including costs imposed on the electricity network, less suitable sites and applications for DE units, rising costs on input fuels, and stresses on the fuel infrastructure. Unless the DE unit is powered by renewable resources at the point of use, the primary fuel delivery infrastructure will likely be natural gas. A rapid rise in natural gas demand from DE technologies will be tempered by rising natural gas prices and increasing delivery costs as the gas network is extended and reinforced.

A final general driving force of economic costs is path dependency. The lifetime of traditional large-scale energy investments (both plants and networks) is measured in years or decades. This gives rise to path-dependent technological trajectories that can "lock-in" energy supply into nonoptimal portfolios of generation and transmission infrastructure. As large amounts of DE market penetration would amount to a paradigm shift in the generation and supply of energy, closing of an older plant and bypassing of existing transmission capacity may allow a more efficient overall system, especially if DE technologies are operated as CHP plants. However, recompense for long-lived generation and transmission assets remains one of the most contentious issues as energy markets are liberalized.

6.2 Valuing DE within Electricity Networks

The benefits or costs that DE technologies bring to an electricity network are a function of the level of penetration. If DE is limited to a relatively small percentage of overall capacity, it can be incorporated into the existing paradigm of remote generation of electricity with extensive transmission and distribution networks. Within this paradigm, DE can fulfill a number of functions in addition to providing electricity and heat to individual customers. These include

grid reinforcement, standby power, meeting peak demands, remote sites, and hedging against fuel volatility.

The use of DE in areas where it is difficult or expensive to upgrade the distribution network can alleviate bottlenecks in the system and enable greater system capacity and reliability. For grid reinforcement benefits to occur, utilities should know, at a minimum, when, where, and how much electricity will be exported to the grid. The use of DG for network management is much improved if utilities also have some measure of control over DG resources. DE can also be used for peak shaving to meet demands for electricity at times of highest usage. This can be particularly cost-effective using low-capital-cost DE technologies, such as engines. Other DE technologies can be used in remote areas where it is not economically feasible to extend the distribution network. Historically, this has been the application where renewable DE technologies have been employed, despite their high capital costs. Finally, as DE technologies have a range of fuel sources, including natural gas, waste gases, oil, and renewable fuels, they can contribute in a portfolio of supply technologies to hedge against volatility in fuel costs.

As DE capacity increases in the electricity sector, a range of issues related to costs imposed on the wider electricity sector become more prominent. If, as in many developing countries, the national electricity grid is only in its infancy, then a system based on DE and alternate fuel delivery options is perfectly feasible and will offer distinct characteristics in terms of environmental, economic, and reliability performance. However, in all developed nations, a national electricity network that has been designed to move large amounts of power from remote generation sites to meet a diverse set of demand loads is already in place. Incorporating DE into this system necessitates consideration of interconnection, buy-back (or export) tariffs, net-metering, stranded investment costs, and impacts of changing demands.

A general theme in the relationship between DE and the electricity network is that DE users should compensate the utility for any costs imposed on the centralized network, but also be rewarded for any benefits the use of DE gives to the network. Experience has shown that coordination between users, utilities, regulators, and technology vendors is essential for optimal use and recompense of DE to be achieved. Specific regulatory structure will dictate whether utilities can be direct partners in DE projects. Utilities constrained by a cost plus revenue arrangement, where their income is driven by the amount of electricity that flows down their wires, are most likely to view DE as a threat to their core business.

Interconnection charges are imposed on the DE user to connect a unit to the electricity network for both emergency and backup power and to allow export of DE-generated electricity back to the grid. The general small size of DE schemes makes individual assessments impractical and cost-prohibitive and a number of countries have, or are developing, protocols for simple and equitable interconnection of DE. Buy-back tariffs (or prices for electricity export) are the price paid to the DE user for electricity sold back to the grid. The fair price for this electricity will depend on where the DE unit is located in the distribution network, how much electricity is exported, when it is exported (especially in reference to peak demand times), and whether electricity exports are guaranteed. Figure 4 illustrates the range in electricity buy-back tariffs in the Netherlands and United Kingdom. Larger volumes for longer periods of time are valued more highly. The overall difference between the two countries is a result of the level of cooperation with the incumbent utilities and their willingness to use DE as an integral part of their generation and distribution portfolios. An extension of buy-back tariffs is net-metering, where a DE user can reverse the flow of electricity from the grid to offset electricity purchases at other times. A flat rate for net-metering unfairly imposes costs on the electricity utility, whereas flexible net-metering tariffs dependent on time, place, and level of guarantee have much to recommend them.

Wider cost issues for electricity networks include stranded assets, where a capital-intensive generation and distribution plant is not utilized and hence does not recover its investment. This can also occur if

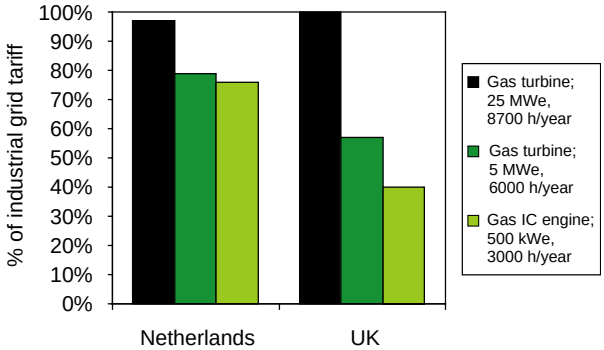


FIGURE 4 Electricity buy-back tariffs for DE technologies. From Cogen Europe (1995), The Barriers to CHP in Europe, SAVE Program of the European Commission, Brussels, Belgium.

capacity must be kept available to meet energy demands from DE unit failure. As electricity markets have been liberalized in many countries, recompense for stranded investments has been a dominant feature of the regulatory debate. Finally, as DE capacity grows, it will affect the demand profile of the area the electricity network serves. If the DE units primarily meet base-load demand, the remaining load will become more variable and hence more expensive to service with a conventional centralized generation portfolio.

7. ENVIRONMENTAL CHARACTERISTICS OF DE

DE technologies exhibit a range of emission characteristics. Carbon dioxide (CO_2) and sulfur dioxide (SO_2) emissions depend only on the carbon or sulfur content of the fuel and the efficiency of generation and supply. Local pollutants including nitrogen oxides (NO_x) also depend on the specifics of the combustion or generation process. The relative environmental performance of DE depends on what centralized electricity technology and fuel they are competing against and whether the DE technology is being used for electricity only or CHP applications. If it is a DE–CHP application, then the emissions from the heat boiler plant must also be factored in. Higher relative DE efficiencies are also aided by the avoidance of electricity distribution losses. Further issues in comparing environmental characteristics of technologies include what control technologies are assumed to be fitted and whether the comparison units are the average of the installed capacity base or only of new units.

Table V gives the latest controlled technology, efficiencies, and controlled emission rates of a range of DE technologies, as well as for a centralized coal steam turbine plant, combined cycle gas turbines (CCGT), and an on-site heat boiler plant. For calibration to alternative units, 1 lb/MWh is approximately equivalent to 0.45 g/kWh and approximately equivalent (for NO_x only) to 22 ppm.

Table V shows that some DE technologies (fuel cells, Stirling engines, PV, and wind turbines) have very low levels of global and local air pollutants. Diesel engines and coal steam turbines are the most polluting distributed and centralized technologies, respectively. Although CCGT is a much cleaner centralized technology, once emissions from heat boilers are factored in, even gas engines are equivalent in emission rates for CO_2, SO_2, NO_x, and particulate matter (PM_{10}). Engine emissions of carbon monoxide (CO) and unburned hydrocarbons (HC) remain a significant problem.

Comparing emission rates is only a first step in evaluating the environmental impact of DE technologies. It is also important to note that emissions will occur closer to population centers, and hence increase the health impacts of local air pollutants, and may also happen at specific times when environmental loading is at its peak. An example of this would be significant levels of diesel engine use in urbanized areas during hot summer afternoons when ozone (O_3) formation is at its peak. Thus, the specific health and environmental impacts of significant DE use require atmospheric

TABLE V

Emission Characteristics of DE and Power/Heat Technologies (lbs/MWh)[a]

	Control technology	CO_2	SO_2	NO_x	CO	PM_{10}	HC	Mercury (Hg)
Reciprocating gas engine	Ignition timing and SCR	1394	0.08	1.11	4.01	0.03	1.20	0
Reciprocating diesel engine	Ignition timing and SCR	1639	2.95	4.75	6.24	0.80	3.68	0
Gas turbine	Staged combustion	1406	0.08	0.67	0.68	0.10	0.98	0
Microturbine	SCR	1617	0.09	0.45	1.05	0.09	0.31	0
Fuel cell (phosphoric acid fuel cell)	None	1064	0.06	0.03	0	0	0	0
PV	None	0	0	0	0	0	0	0
Stirling engine	None	1617	0.09	0.1	0	0	0	0
Wind	None	0	0	0	0	0	0	0
Combined-cycle gas turbine	Staged combustion	980	0.06	0.47	0.17	0.10	0.12	0
Coal steam turbine	Low NO_x burners, FGD and SCR	2345	13.71	4.13	0.17	0.31	0.12	0.0038
Gas heat boiler	Low NO_x burners	449	0.03	0.49	1.05	0.09	0.11	0

[a] Divide by 2.2 to convert to g/kWh.

SCR, selective catalytic reduction; FGD, flue gas desulfurization.

modeling. As many DE technologies have thus far been "under the radar" in regulatory efforts for air pollution, such detailed analysis will be paramount.

Finally, it should be noted that there are non-emission environmental issues that ensure that all DE technologies have some environmental concerns to be addressed. Noise pollution, particularly near workplaces and residences, is a concern (see Table II for uncontrolled noise levels of DE technologies). Even renewable DE technologies with no noise or air pollution consequences need to address concerns over hazardous materials during manufacture and disposal (PV and batteries) or visual impacts (wind turbines).

8. OTHER CHARACTERISTICS OF DE

8.1 Reliability

Reliability considerations for DE can be divided into plant and system impacts. At the plant level, high levels of availability of DE units are achieved through remote monitoring and real-time control. As there can be a large number of DE units, reliability can be ensured for scheduled maintenance and unscheduled outages by transferring the electricity load to the remaining DE units. Building or process management systems can turn off nonessential loads if required. Energy storage devices can also be used to maintain supply for short periods. Finally, the dual-fuel capabilities of many DE technologies or the use of different DE units at the same location can ensure supply in the case of fuel disruptions.

In comparing DE reliability to a centralized electricity network, it should be noted that the existing electricity grid is extremely reliable, operating in excess of 99.9% availability. The majority of outage events in the centralized network are due to accidents and climatic events in the transmission and especially distribution system. DE offers a route to bypass these outages. In addition, in countries such as India, where widespread theft of electric power from the distribution network, and hence reduced reliability, is commonplace, DE can offer an alternative.

The comparison, and possible coordination, between DE and the electricity network depends on which generation source is the back-up. Traditionally, the electricity grid has been employed as emergency or stand-by generation for DE units. This imposes costs on the electricity grid (see Section 6.2) as the utility must maintain spare generation and distribution capacity to meet a potential additional load from a DE outage event. However, this disadvantage is avoided if a cluster of DE units can provide emergency power without the grid. Taking this paradigm one step further, a DE cluster can act as a source of emergency or stand-by generation to the grid, especially in times of peak demand when generation costs or distribution bottleneck issues are of utmost concern.

A final reliability issue is the comparison between the electricity network and the fuel distribution (primarily natural gas) that DE units would depend on. This is discussed below in the context of robustness.

8.2 Robustness

The robustness of an energy generation and distribution network has received considerable attention for its performance under conditions of conflict, whether in terms of war or terrorist events. A system based on DE would do little to alleviate dependence on oil imports, which is primarily related to transportation. On an individual generation unit basis, the small size and low profile of DE units make them a less important target. Although DE can be powered by a range of renewable, waste, and fossil fuels, the reliance on a natural gas infrastructure has been debated and compared to the system robustness of the electricity network. Table VI details some of the advantages and disadvantages of a natural gas DE network.

8.3 Power Quality

Evolving demands for high-quality electricity in a range of sensitive industrial processes and information technologies has refocused attention on the need for extremely high-quality power. DE technologies can help respond to this demand in a number of ways. The use of DE technologies can avoid movement of electricity where the majority of power quality problems, especially voltage sags, occur. Energy storage devices are a DE resource that is often used to provide uninterrupted power supply either as a support to the power grid during periods of very short term instability or as a bridge to allow stand-by DE technologies to begin operation and ramp up to meet the required load.

TABLE VI

Potential Robustness Advantages and Disadvantages of DE Systems

Features of natural gas DE	Conflict context advantages	Conflict context disadvantages
Increased number and smaller size of DE generators	When one generator is damaged, a much smaller proportion of the generating capacity is unavailable	Harder to physically protect individual generation sites
Greater use of natural gas	May replace nuclear facilities	More prone to risks of foreign supply interruptions (pipelines and LNG)
Decreased reliance on electricity transmission and distribution	The electricity T&D system is harder to protect than generators; having generation close to the load reduces the reliance on the vulnerable transmission system	Unless DE units are arranged in clusters to back up one another, may still have to deal with electricity grid outage effects
Increased reliance on natural gas transmission and distribution	Underground natural gas T&D systems (which would be needed anyway for on-site heat-only plants) are generally underground and therefore better protected than electrical T&D lines	Long-distance over-ground natural gas pipelines are very difficult to protect and are vulnerable to attack (with explosives or high-powered rifles)
T&D real-time operational advantages	Gas pipelines do not have the same strict real-time operational problems as electric power grids; in electric power grids, a disturbance to one part of the grid can result in cascading failures that knock out power to a much wider area than the original disturbance	
Fuel substitutability and storage	A natural gas DE system can have greater robustness through dual fuel technologies and local storage facilities for gas and/or other fuel suitable for use in ICE DE units	A centralized electricity network can also build in a diverse supply portfolio

Note: T&D, transmission and distribution; LNG, liquified natural gas.

As there are a range of customer types, each requiring, and willing to pay for, a range of power quality levels, those users requiring extremely high quality power (e.g., information technology firms) have been investigating the concept of microgrids, where a range of DE technologies, along with energy storage options, are geared to provide very high quality power in a designated geographical area. DE microgrids could be based on DC power if required for sensitive electronic equipment.

8.4 Modularity

A final characteristic of DE technologies is that they represent modular or increment investment and capacity additions, especially in comparison to large centralized power plants. This feature has been even more advantageous following the liberalization of many nations' electricity industries and the resultant increase in electricity price uncertainty. The ending of the monopoly systems where investment revenues were guaranteed to be recovered has placed a premium on projects with a shorter, less expensive construction period. Many factory-made and –assembled DE units come packaged with heat recovery, acoustic enclosure, and control interface and can be installed at a site and connected to the electricity and heating systems in 1 or 2 days.

In addition to improving investment under price uncertainty, modular capacity extensions can more easily follow increments in demand growth, alleviating shortages in generation capacity. Adding further value, the geographical flexibility of modular DE capacity allows it to alleviate bottlenecks in the distribution capacity. Finally, mobile DE units can be quickly moved to meet peak demand periods or moved on a seasonal basis to meet climate-driven high-demand periods.

9. FUTURE ROLE OF DE

The traditional policy driving forces for DE, and indeed for all energy supply and distribution options, have been the lowest economic cost and security of supply. Increasingly, motives for evolving policies will also be concerns over global and local environmental issues, reliability, robustness, and quality of electricity supply, and integration with energy infra-

structures, notably electricity, natural gas, and any future hydrogen infrastructure.

Along with continuing technical developments in DE technologies, cost reductions will be largely delivered by economies of scale in production, geographical economies of scale, and the processes of learning by doing and learning by using as DE penetrates the market. Regulatory developments, especially facilitating flexible ownership and operational arrangements, will be necessary if these cost reductions are to be realized.

Any new regulatory design will have to accommodate and value the impacts of a widespread penetration of DE technologies on existing infrastructures and especially electricity generation and distribution utilities. Utilities need to be fairly compensated for costs incurred for interconnection, ensuring emergency supply, altered demand profiles, and possible stranding of investments in generation and distribution equipment. These costs need to be transparent and standardized, as the introduction of DE can be incorporated to reduce peak requirements, defray distributional capacity expansions, and meet the range of daily and seasonal demands at lowest cost. Regulatory decisions over the ownership, technical protocols, market structure, and liability aspects of the electricity industry will all affect whether and how DE is incorporated into the electricity paradigm.

A diverse range of DE technologies, including renewable DE and multifuel DE units such as microturbines and Stirling engines, can help contribute to the diversity and hence the security of energy supplies. However, if the great majority of DE units are fired on natural gas, this could exacerbate price volatility and upward price pressure on natural gas supplies, particularly in North America.

Renewable DE and CHP DE technologies have the potential to significantly reduce emissions of CO_2 and thus may be a major tool in any concerted drive to limit and reduce greenhouse gas emissions. Although most DE technologies emit very low levels of SO_2, only some (fuel cells, microturbines, Stirling engines, PV) have very low levels of a variety of local air pollutants. Given the distributional location of DE near population centers, all local air pollution impacts need to be fully assessed and further controlled if necessary. In addition, any doubts of affected local populations need to be understood and addressed.

Evolving demands for high reliability, high robustness, and high quality in liberalized electricity markets will likely result in customers specifying criteria of supply. For those customers who have very strict requirements for reliable and high-quality power, DE technologies can either supplement the grid or be used in microgrids to meet specialized demands. The use of DE can hedge against both supply and price uncertainties, particularly in view of the modularity of capital investments required for DE.

Significant expansion of DE capacity may impose greater strain, increased price levels, and greater price volatility on the natural gas network. It may also change seasonal demand profiles as more summer electricity is generated and may require rethinking of natural gas storage requirement and locations. Finally, DE may or may not be integral or compatible with the growing use of hydrogen. If hydrogen is created at central facilities (for example, using fossil fuels and carbon sequestration), the energy distribution carrier could be electricity, taking advantage of economies of scale in plants, or hydrogen, taking advantage of economies of scale in DE unit production and higher efficiencies with CHP DE units. If the transportation sector moves to a hydrogen system with a distribution pipeline to fueling stations, this is likely to support hydrogen-fueled stationary DE applications.

Finally, DE technologies and their associated infrastructure impacts and ownership arrangements potentially represent a very different paradigm than the norm of centralized electricity generation and distribution with on-site heat provision. These two paradigms have advantages and disadvantages, but in countries that already have a developed energy infrastructure, any widespread move toward DE will entail significant transitional concerns over a period of decades due to the long life of electricity-related equipment.

SEE ALSO THE FOLLOWING ARTICLES

Electricity, Environmental Impacts of • *Electricity Use, History of* • *Electric Power Reform: Social and Environmental Issues*

Further Reading

Ackermann, T., Andersson, G., and Soder, L. (2001). Distributed generation: A definition. *Electric Power Systems Res.* **57**, 195–204.

Borbely, A., and Kreider, J. F. (eds.). (2001). "Distributed Generation: The Power Paradigm for the New Millennium." CRC Press, Boca Raton, FL.

Cowart, R. (2001). "Distributed Resources and Electric System Reliability." The Regulatory Assistance Project. Gardiner, ME.

Dunn, S. (2000). "Micropower: The Next Electrical Era," WorldWatch Institute Paper 51. WorldWatch Institute, Washington, DC.

Feinstein, C., Orans, R., and Chapel, S. (1997). The distributed utility: A new electric utility planning and pricing paradigm. *Annu. Rev. Energy Environ.* **22,** 155–185.

Gas Research Institute. (1999). "The Role of Distributed Generation in Competitive Energy Markets." Gas Research Institute, Chicago, IL.

Laurie, R. (2001). Distributed generation: Reaching the market just in time. *Electric. J.,* 87–94.

Meyers, E., and Hu, M. G. (2001). Clean distributed generation: Policy options to promote clean air and reliability. *Electric. J.,* 89–98.

Morgan, M. G., and Zerriffi, H. (2002). The regulatory environment for small independent micro-grid companies. *Electric. J.,* 52–57.

Patterson, W. (2000). "Transforming Electricity." Earthscan, London.

Smeers, Y., and Yatchew, A. (eds.). (1998). Distributed resources: Toward a new paradigm of the electricity business. *Energy J.,* Special Issue.

Strachan, N., and Farrell, A. (2002). "Emissions from Distributed Generation," Carnegie Mellon Electricity Industry Center Working Paper 02–04. Carnegie Mellon Electricity Industry Center, Pittsburgh, PA.

Strachan, N., and Dowlatabadi, H. (2002). Distributed generation and distribution utilities. *Energy Policy* **30,** 649–661.

Willis, H. L., and Scott, W. (2000). "Distributed Power Generation: Planning and Evaluation." Dekker, New York.

Zerriffi, H., Dowlatabadi, H., and Strachan, N. (2002). Electricity and conflict: Advantages of a distributed system. *Electric. J.* **15,** 55–65.

District Heating and Cooling

SVEN WERNER
Chalmers University of Technology
Gothenburg, Sweden

1. Five Strategic Resources
2. Market Penetration
3. History of District Heating
4. Technical Design and Construction
5. Customer Relations
6. Energy Supply
7. Environmental Impact
8. District Cooling
9. Conclusions

Glossary

cogeneration Method of simultaneous generation of electricity and heat; also known as combined heat and power (CHP).
community heating Synonym for district heating, common in Great Britain.
customer substation Installation in customer building transferring heat from the heat network to the internal heat distribution systems in the building.
district cooling Method of fulfilling many cooling demands by a seldom citywide cooling distribution network, chilled from one or many large chillers or transferring facilities.
district energy Collective term for district heating and cooling.
district heating Method of fulfilling many city heat demands by an often citywide heat distribution network, which receives heat from one or many large heat-generation or transmission facilities.
trigeneration Method of simultaneous generation of electricity, heat, and cold.

The fundamental idea of district heating is to use local fuel or heat resources that would otherwise be wasted to satisfy local customer heat demands by using a heat distribution network of pipes as a local marketplace. This idea contains the three obligatory elements of a competitive district heating system: the suitable cheap heat source, the market heat demands, and the pipes as a connection between source and demands. These three elements must all be local in order to obtain short pipes for minimizing the capital investment in the distribution network. Suitable heat demands are space heating and preparation of domestic hot water for residential, public, and commercial buildings. Low-temperature industrial heat demands are also suitable.

1. FIVE STRATEGIC RESOURCES

The five suitable strategic local energy resources are useful waste heat from thermal power stations (cogeneration); heat obtained from refuse incineration; useful waste heat from industrial processes; natural geothermal heat sources; and fuels difficult to manage, such as wood waste, peat, straw, or olive stones. These heat sources must be cheap in order to compensate for capital investments in the distribution network and complementing heat-generation plants for peak and backup heat demands. The latter is needed in order to meet customer heat service demands at extremely low outdoor temperatures and when the regular heat source is temporarily unavailable.

District heating is an energy service provided for immediate use directly by the customer and was commercially introduced in the 19th century as a very early example of outsourcing.

2. MARKET PENETRATION

Citywide district heating systems exist in Helsinki, Stockholm, Copenhagen, Berlin, Munich, Hamburg, Paris, Prague, Moscow, Kiev, Warsaw, and other large cities. Many systems supply a downtown district (such as in New York, San Francisco,

 © 2004 Elsevier Inc. All rights reserved.

Minneapolis, St. Paul, Seattle, Philadelphia, and other cities) or a university, military base, hospital complex, or an industrial area.

Total annual heat turnover is approximately 11 EJ in several thousand district heating systems operating throughout the world. The amount of heat delivered corresponds to 3.5% of the total global final energy consumption (1999). The market penetration for district heating in various countries is presented in Table I which is based on international energy statistics. However, many countries have undeveloped or no routines for gathering statistical information from district heating systems. This is valid for the United States, Great Britain, France, China, and some other countries. An independent analysis estimated the total annual heat deliveries from all district heating systems in the United States to be more than 1000 PJ. Hence, the real-world market penetration is probably higher than that presented in Table I.

Fulfilling heat demands for space heating and domestic hot water preparation for residential, public, and commercial sectors constitutes the majority of deliveries. Among the Organization for Economic Cooperation and Development (OECD) countries, district heating has a strong, almost dominating, market position in Denmark, Finland, Sweden, Poland, and the Czech Republic. Among non-OECD countries, district heating has a strong market position in Russia, Belarus, Romania, and the three Baltic states.

3. HISTORY OF DISTRICT HEATING

Birdsill Holly, an inventor and hydraulic engineer, is often credited with being the first to use district heating on a successful commercial basis. As a result of an experiment in 1876 involving a loop of steam pipes buried in his garden, Holly developed a steam supply system in October 1877. Several district heating systems were started in North American cities in the 1880s. The fuel source was steam coal. In New York, the Manhattan steam system went into operation in 1882. It exists today as the steam division of Consolidated Edison, and it delivered 29 PJ of heat during 2000.

The oldest district heating system still in operation is located in Chaudes-Aigues, France. It is based on a geothermal heat source with a temperature of 82°C and was in operation as early as the 14th century. Old municipal documents indicate that two citizens did not pay their heat fees properly in 1332. The hot water was partly distributed using drilled tree trunks as distribution pipes.

In Europe, an early district heating system was built in Dresden in 1900, although it was not a commercial project. The main purpose was to reduce the fire risk in 11 royal and public buildings containing invaluable art treasures. A more commercial project was initiated by Fernheizwerk Hamburg Gmbh for the city of Hamburg in 1921. According to Abraham Margolis, the company chief engineer, the driving force for the project was the high cost of fuel after World War I. The Hamburg system was followed by systems in Kiel in 1922, Leipzig in 1925, and Berlin in 1927. Outside Germany, district heating systems were started in Copenhagen in 1925, Paris in 1930, Utrecht in 1927, Zürich in 1933, and Stockholm and Helsinki in 1953. Reykjavik, Iceland, started a geothermal district heating system in 1930, which today supplies almost all the 160,000 inhabitants with heat for space heating and domestic hot water (10 PJ during 2000).

In Russia, a general utilization of cogeneration and district heating was outlined in the electrification plan in 1920 in order to reduce future fuel demand. The first heat was delivered in St. Petersburg in 1924. Teploset Mosenergo was established in 1931 to manage heat distribution in Moscow, although the heat deliveries had begun in 1928. This department of Mosenergo delivered 268 PJ during 2000 and an additional 70–80 PJ was supplied by Mosteploenergo, another local distributor. Together, these companies constitute the most extensive district heating system in the world. The second largest is the St. Petersburg system.

4. TECHNICAL DESIGN AND CONSTRUCTION

Heat is transferred from the network to the building heating systems by customer substations located in the connected buildings. A building has at least two internal distribution systems that must be heated—one system for supplying heat to the radiators and one system for distribution of domestic hot water. Sometimes, a separate system also provides heat for heating the supply air in the mechanical ventilation system.

Each internal system is heated and regulated separately. The heat is often transferred by the use of heat exchangers, in which case the connection is indirect. There are also direct variants with only valves and elevator pumps. If domestic hot water is prepared by mixing district heating water and

TABLE I

Heat Flows through District Heating Systems for Various Countries and Regions in 1999[a]

Region or country	Total heat supply (PJ)	Total deliveries (PJ)	Industrial customers (PJ)	Other customers (PJ)	% of all final consumption	% of all industrial demands	% of all other demands
World	11,473	10,852	3782	6079	3.5	4.2	5.5
OECD total	2296	2120	535	1452	1.3	1.2	3.0
United States	406	345	226	83	0	2	0
Canada	33	33	33	0	0	1	0
Korea	166	161	0	161	3	0	10
Japan	25	22	0	21	0	0	0
Switzerland	16	14	6	8	2	3	2
Norway	8	6	1	5	1	0	2
Poland	371	371	57	262	12	7	21
Czech Republic	153	146	39	85	12	9	22
Hungary	71	71	19	43	9	11	11
Slovak Republic	31	29	1	23	4	0	11
European Union	1007	913	155	751	2	1	5
Germany	371	343	44	292	3	2	7
Sweden	168	158	13	145	11	2	26
Finland	126	116	27	89	11	5	26
Denmark	122	98	6	91	15	5	30
The Netherlands	105	90	39	51	4	5	6
Austria	48	43	5	38	4	2	9
France	31	31	0	31	0	0	1
Other countries in European Union	36	34	21	13	0	0	0
Other countries in OECD	9	8	0	8	0	0	0
Non-OECD total	9178	8732	3247	4627	5.9	7.2	7.4
Russia	6332	6158	2098	3607	33	36	47
China	1344	1327	719	249	3	5	2
Ukraine	576	404	258	141	11	16	8
Belarus	300	281	118	159	36	40	45
Romania	216	187	20	146	17	5	34
Uzbekistan	105	105	0	105	7	0	11
Former Yugoslavia	67	64	8	53	7	3	13
Bulgaria	55	47	11	29	10	6	22
Lithuania	52	40	5	32	19	11	37
Latvia	36	30	1	27	20	4	36
Estonia	29	24	2	21	22	7	39
Other countries in non-OECD	66	65	8	57	0	0	0

[a] According to the energy balances for OECD and non-OECD countries published by the International Energy Agency. Differences between supply and deliveries constitute distribution losses. Differences between total deliveries and sum of deliveries to customers constitute own use. All other demands do not include the transportation sector.

domestic cold water, the open connection method is used, whereas when a heat exchanger is employed, the closed connection method is used. The open connection method occurs is used in Russia, resulting in high feed water demands in the distribution network.

The most typical part of a district heating system is the citywide distribution network of pipes buried

in the ground under streets, pavement, and park lawns or installed in building basements. Water is normally used as the heat carrier in heat distribution networks. Steam is completely or partly used in systems that were installed before approximately 1940 and for high-temperature industrial heat demands.

The water temperature in the forward pipes varies between 70 and 150°C, with an annual average of 80–90°C. The higher temperature is used at extreme low outdoor temperature, whereas the lower temperature is used during the summer to enable domestic hot water preparation. The temperature of the return water varies between 35 and 70°C, with an annual average of 45–60°C. However, return temperatures of 30–35°C may be obtained in the future. The current high return temperatures depend on malfunctions in substations, short-circuited radiator systems in customer buildings, and short-circuits between forward and return pipes in the distribution networks. The lower the return temperature, the more heat can be transported in the network.

The pipes are constructed so that they accommodate thermal expansion and avoid outer corrosion. Construction methods for pipes buried in the ground have varied throughout the years. In the early days of district heating, double steel pipes were insulated with mineral wool and laid in a common concrete square box duct. This early generation of distribution pipes were expensive to build but were reliable if the ducts were well ventilated and well drained. Today, the most common method is to use prefabricated steel pipes with polyethylene casing and insulated with polyurethane foam. This method has the advantages of low distribution cost, low heat losses, and high reliability. During the past 30 years, plastic pipes have been developed as heat carrier pipes. With respect to limitations in pressure and possible long-term temperatures, the new generation of district heating pipes are mostly used in secondary distribution networks separated from the primary distribution networks by heat exchangers.

The heat loss in the distribution network depends on the thermal resistance in the pipe insulation, the pipe size, the supply and return temperatures, and the linear heat density. The latter is the heat sold annually divided by the route length of double pipes in the network. Typical values for linear heat densities are 15–25 GJ/m for whole networks, more than 40 GJ/m in concentrated downtown and commercial areas, and less than 5 GJ/m in blocks with single-family houses. Typical heat loss in whole networks is 5–10% of heat generated in plants, but up to 20–30% can occur in areas with low linear heat densities. Also, the distribution cost depends on the linear heat density. The cost is low beyond 20–30 GJ/m but increases very quickly below 6–8 GJ/m.

The district heating system is operated using control equipment at four independent levels. Two of these are located in the buildings connected and two are managed by the district heating operator. The first level is the heat demand control by thermostatic valves at the radiators and mixing valves for domestic hot water. The second level is in the customer substations. The heat transfer is controlled by valves that adjust the flow of district heating water. One valve is used for each connected customer system. The third level is the control of the pressure difference between the supply and return pipes in the network, which is accomplished by adjusting the speed of the distribution pumps. This control level allows for all customer substations to receive heat since this pressure difference is the driving force for the flow to circulate through the customer substations. The final level is the control of the supply temperature by adjusting the capacity in the heat-generation facilities.

These levels of control result in an efficient demand-oriented district heating system. The district heating operator can never deliver more heat than the customers request. However, the first three levels are often not available in Russian and Eastern European systems. This situation creates a production-oriented system in which it is difficult to properly allocating the flow in the network, resulting in both low and high indoor temperatures in the buildings connected. When these systems are rehabilitated, the first step is to install at least control at the second and third levels.

5. CUSTOMER RELATIONS

The customers usually pay for the heat received according to a heat rate, which normally consists of one fixed portion related to the capacity needed and one variable portion related to the amount of heat bought. Sometimes, a water volume portion is also used in order to promote low return temperatures from the customer substations. The capacity portion is based on either the maximum heat load or the maximum water flow capacity. The amount of heat sold to the customer is recorded by integrating the product of water flow and the temperature difference between the supply and return pipes. A heat meter in each customer substation measures the district

heating water flow, the supply temperature, and the return temperature. These three measurements are integrated with media constants to derive an accumulating heat volume in gigajoules or megawatt hours by the heat meter.

The point of delivery varies from system to system. If the customer owns the substation, the point of delivery is before the substation; otherwise, it is after the substation. Sometimes the customer pays a connection fee.

The pricing principle for the heat rate varies. Municipal district heating systems generally have lower prices, reflecting lower returns of capital investments. Privately owned systems often price the district heat according to the available alternative for the customer, usually the price of light fuel oil or natural gas with a discount in order to keep the customer. In some countries, the price levels in district heating systems are supervised or approved by national or regional regulatory authorities.

Typical prices for district heat deliveries were $6–13/GJ in OECD countries in 1999 and 2000. An exception was the geothermal system in Reykjavik, which sold heat for only $4/GJ. Typical prices in Eastern Europe were $3–7/GJ, except in Russia, where world market energy prices are not allowed due to a low ability to pay market prices. Heat was sold from the Moscow system for approximately $1/GJ. All these prices are lower than or equal to those from local boilers using fossil fuels with different national energy taxes, such as carbon dioxide taxes.

The lower prices in countries in Eastern Europe and the former USSR just cover the operating costs. The returns of capital investments made in the former planned economies are currently kept low by the national regulatory authorities since the average ability to pay for heating is low in the population.

Requests for deregulation of district heating systems, allowing third-party access, have been put forward from customers and competitors, for example, in New York. The requests are for the deregulation processes of telecommunication, aviation, railway, gas, and electricity markets. These markets are large, international, and mature with many market actors. The goal of deregulation of these markets is an effective price determination with many sellers and buyers in each market. However, district heating systems are normally small local markets, which are often immature, with low market penetration. One operator can easily supply a town of 150,000 inhabitants with heat from one site; thus, more operators are not needed in many district heating systems. Hence, deregulation of district heating systems cannot easily be implemented.

6. ENERGY SUPPLY

The base load heat-generation capacity in a district heating system must have a strong correlation with the five strategic resources discussed previously: cogeneration, refuse incineration, industrial waste heat, geothermal heat, and fuels difficult to manage. Internationally, the strongest driving force for district heating is the simultaneous generation of electricity and heat in cogeneration plants. Base load capacities are associated with low operating costs and high capital investment costs.

Conventional fossil fuels dominate the fuel supply for cogeneration, but the use of biomass is increasing in some countries. One example is the newly built biomass cogeneration plant in Eskilstuna, Sweden, which meets the majority of heat demand of a city with 65,000 inhabitants. Another example is the new biomass cogeneration plant for the St. Paul system. In general, cogeneration plants use more carbon lean fuels than conventional thermal power stations since natural gas, renewables, and waste are more common fuels than coal and oil. It is also possible to recover heat from nuclear power stations, but this is employed to a very limited extent due to long distances between nuclear reactors and heat demands in cities, loss of some electricity generation, fear of strong dependence on one large heat source, and lack of trust in nuclear technology in some countries.

Waste incineration is an acceptable treatment method compared to landfill for the burnable industrial and municipal waste stream. This waste remains after separation of streams for hazardous waste, recycling, and biological waste treatment. Normally, this waste stream from a city comprises only approximately 5–15% of the design capacity in a mature district heating system with a high local market penetration. However, 10% of the design capacity corresponds to approximately 30% of the annual heat generation. Within the European Union, up to 50% of municipal waste is allocated for incineration. Denmark, Switzerland, and Sweden incinerate the highest percentages of waste.

Many industrial processes release heat at a temperature higher than 100°C either in flue gases or in cooling water. This heat can be directed to a district heating system by using heat-recovery boilers or heat exchangers. In Gothenburg, the district heating system receives useful waste heat from two

oil refineries, comprising 30% of the annual heat demand in the system. Revenues from the heat sale cover a major portion of staffing, operating, and maintenance costs for the refineries. This heat recovery gives them a competitive advantage over other oil refineries in Europe.

Geothermal heat is a natural energy source for heating in the same manner that wind, hydro, and solar power are natural renewable energy sources for power generation. However, this resource is not available everywhere since it is restricted to areas with the appropriate geophysical conditions. There are geothermal district heating systems in Iceland, the United States, France, China, Japan, Turkey, Poland, Hungary, and some other countries. However, the total heat delivery worldwide is low, estimated to be approximately 30 PJ/year, where one-third is supplied in the Reykjavik system.

Biomass is an example of a fuel that is normally difficult to manage for conventional heating. Regarding transport, fuel feeding into boilers, and environmental impact, these types of fuel are easier to handle in one central boiler than in several small boilers. Biomass fuels can be obtained from agricultural sources or from the forest industry. Wood waste can be delivered as raw chips or refined as small pellets or powder from forest companies or sawmills. In the early 1990s, the Cree Nation of Ouje-Bougoumou determined that a biomass district heating system would be an enormously positive tool in contributing to the development of the community's future financial base. This aboriginal community, located 1000 km north of Montreal, Canada, uses wood waste from a nearby sawmill for their two biomass boilers, which are connected to the district heating system. Importing fuel oil to this village was considered a less favorable option.

Peak and backup capacity plants complement the base load capacity plant in a district heating system. These plants are needed to meet customer heat service demands at extremely low outdoor temperatures and when the regular heat source is temporarily unavailable. These plants are normally associated with high operating costs and low capital investment costs.

The optimal heat-generation cost is obtained by choosing the base load capacity so that its marginal capacity will have the same annual total cost as peak load capacity for the same operating hours. The typical optimal base load capacity is approximately half the nominal system design capacity based on the local design outdoor temperature. The base half of the capacity represents approximately 90% of the volume of heat generated.

The frequency distribution for the outdoor temperatures determine the limit of capacity utilization in a district heating system. The annual average capacity factor is 25–50% for extreme inland and maritime climate locations. The capacity factor is defined as the actual heat generation divided by the heat generation that is possible. This situation occurs since less heating is required during autumn and spring compared to winter, and no heating is needed during summer.

7. ENVIRONMENTAL IMPACT

High contents of dust, sulfur dioxide (SO_2), and nitrogen oxides (NO_x) in urban areas were among the first pollution problems to be identified. District heating improved the urban air quality in many towns since local heat boilers were replaced and high chimneys for heat-generation plants were required.

Acidification from SO_2 and NO_x is a major problem in many regions. District heating systems provide a centralized and cheap way to eliminate these emissions by separation, use of fuels with low sulfur content, have higher conversion efficiencies, and have higher combustion quality, resulting in lower NO_x emissions.

Global warming due to the greenhouse gases is a problem more difficult to solve than the problems of urban air quality and acidification. In this respect, district heating can be considered the fifth possibility for carbon dioxide emission reduction, after deposition, lowering final energy demands, higher conversion efficiencies, and fuel substitution. District heating uses mainly waste heat from other processes, replacing the use of fossil fuels for heating buildings.

It is estimated that in 1998, all cogeneration plants and district heating systems throughout the world reduced carbon dioxide emissions by 900 Mtons. This corresponds to approximately 4% of all carbon dioxide emitted from fuel combustion. Hence, district heating systems can play an important role in the mitigation of carbon dioxide emissions.

8. DISTRICT COOLING

District cooling systems distribute chilled water to supply air-conditioning or process cooling. Cities with major downtown or commercial districts with district cooling systems include Stockholm, Hamburg, Paris, Tokyo, Chicago, Minneapolis, St. Paul, New Orleans, and Houston. The first major district

cooling system was employed in Hartford, Connecticut, in 1962. Other systems were built in the 1960s and 1970s. District cooling experienced a renaissance during the 1990s, when many countries demanded the termination of the use of CFC due to ozone depletion. District cooling systems are much smaller than district heating systems. Annual cold generation is approximately 1 or 2 PJ in the largest district cooling systems.

The distribution network is similar to that of a district heating system, with one supply pipe and one return pipe. Cold water is circulated instead of warm water. The supply temperature is normally 6 or 7°C, but an ice mixture of 0°C is sometimes used. Customers heat the chilled water in the supply pipe by cooling air supplied to a building or a industrial process. The typical temperature in the return pipe is 12–17°C. The water is again chilled when it circulates through the cold-generation plant. District cooling systems are never used citywide. The low temperature difference makes district cooling distribution expensive, so district cooling networks can only be built in areas with high cooling demand densities, such as downtowns and commercial areas.

Cold generation can be accomplished by central compression or absorption chillers. Free cooling with deep, cold lake or seawater can also be used, as is the case in Toronto and Stockholm. However, this is only possible in areas with cold winters, which annually create cold reservoirs for summer use. Since the diurnal variation of the cooling demand is large at peak load, a significant portion of the peak demand can be met by cold energy storage capacity with a cold water storage tank or an ice storage facility.

Two major associations exist between district heating and district cooling systems. The evaporator of a heat pump can be used for cold generation in a district cooling system and the condenser can simultaneously generate heat in a district heating system. One part of bought electricity will then create two parts of sellable cold and three parts of sellable heat, giving a higher return of the heat pump investment. The other association is the possible use of low-cost district heat for operation of local absorption chillers in connection with customer cooling demands. In this way, citywide deliveries of cooling can be accomplished without installing a citywide district cooling system. This variant is sometimes called warm district cooling, in contrast to cold district cooling, which refers to ordinary district cooling systems. District heating systems using this warm variant are employed in Gothenburg, Mannheim, New York, and Seoul.

These two associations can be integrated into a trigeneration plant, in which power generation is integrated with the generation of both cold and heat. First, a cogeneration unit can generate electricity and heat. These products are then used for feeding a compression chiller or a absorption chiller.

There is an operation optimization benefit for district cooling. Cold generation can be optimized by using different generation technologies to meet various cooling demands. The base load demand can be met by an absorption chiller if low-cost heat is available. Next-generation units in merit order include a compression chiller with heat recovery and a similar chiller without heat recovery. Extreme peak load can be met by a cold storage.

Major advantages of district cooling systems are reduced grid power demand by supplying cooling energy through the district cooling system rather than through the local power grid, shifting power demand to off-peak periods through cold energy storage, and the use of free cooling. Advantages for customers with district cooling are the requirement for less space for cooling equipment inside the building, less operation and maintenance costs, and less electricity cost. The subscribed cooling capacity can also be adjusted to the actual measured customer cooling demand. The customer does not have to buy an expensive oversized chiller.

9. CONCLUSIONS

District heating significantly increases the overall energy system efficiency since the heat supply is coordinated with generation of electricity, refuse incineration, or industrial processes. It is also possible to use geothermal heat sources or biomass fuels such as wood waste. Demand for commercial fossil fuels for heating is reduced, resulting in lower carbon dioxide emissions.

District heating also benefits from economy of scale. This is especially valid with regard to distribution since the cost of pipes is proportional to pipe size, whereas transmission capacity increases with the square of the pipe size. Also, the investment cost for heat-generation capacity decreases with size. The variation in heat demand is also lower, especially with respect to the demand for domestic hot water preparation. However, these advantages are not sufficient to motivate centralized use of conventional fossil fuels in district heating systems. At least one of the five strategic resources must be used to develop a competitive district heating system. When using

more than one of the strategic resources, an internal competition is created within the system, which will increase the competitiveness of the system.

A district heating system becomes more competitive when the advantages of higher overall system efficiency and lower environmental impact are priced by the heat market. Hence, district heating is highly competitive when fuel prices are high and when low environmental impact is appreciated. Many countries use fiscal taxes for fossil fuels and are also introducing carbon dioxide taxes for fuels. International, regional, and national carbon dioxide trading schemes have also been discussed. All these measures increase the long-term competitiveness of district heating systems.

Although district heating has many advantages, the method is still unknown in many communities, whereas it is taken for granted in other communities. Since district heating reduces the operating costs of heating buildings due to capital investment in central heat-generation facilities and distribution pipes, the local view of long-term investments in urban infrastructure is very important. Historically, district heating systems have been built in communities in which district heating has been considered part of the urban infrastructure, along with roads, bridges, water supply, and electricity distribution. The current strong market penetration of district heating in some countries in Europe can be explained by this view. The municipal electrical and technical departments became the district heating entrepreneurs in Finland, Sweden, Denmark, Germany, and Austria since by tradition they were responsible for the urban infrastructure.

SEE ALSO THE FOLLOWING ARTICLES

Biomass for Renewable Energy and Fuels • Cogeneration • Commercial Sector and Energy Use • Geothermal Direct Use • Urbanization and Energy • Waste-to-Energy Technology

Further Reading

Euroheat & Power (1999). "District Cooling Handbook," 2nd ed. Euroheat & Power, Brussels.

Euroheat & Power (2001). "District Heat in Europe, Country by Country—2001 Survey." Euroheat & Power, Brussels.

European Institute of Environmental Energy (1997). "District Heating Handbook." European Institute of Environmental Energy, Herning, Denmark. [Available in Chinese, English, Estonian, German, Latvian, Lithuanian, Polish, and Russian]

Frederiksen, S.,Werner, S. (1993). Fjärrvärme—Teori, Teknik och Funktion" ("District Heating—Theory, Technology, and Function"). Studentlitteratur, Lund. [In Swedish].

Gochenour, C. (2001). "District energy trends, issues, and opportunities, Technical Paper No 493." World Bank, Washington, DC.

Hakansson, K. (1986). "Handbuch der Fernwärmepraxis," 3rd ed. Vulkan Verlag, Essen. [In German].

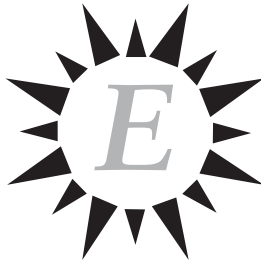
International District Heating Association (currently the International District Energy Association) (1983). "District Heating Handbook," 4th ed. International District Heating Association. Westborough, MA.

Sokolov, E. J. (1982). "Teplofikazija i Teplovye Seti" ("Cogeneration and District Heating Networks"), 5th ed. Energoizdat, Moscow. [In Russian].

Werner, S. (1999). "50 Years of District Heating in Sweden." Swedish District Heating Association, Stockholm.

Westin, P., and Lagergren, F. (2002). Re-regulating district heating in Sweden. *Energy Policy* **30,** 583–596.

Early Industrial World, Energy Flow in



RICHARD D. PERIMAN
Rocky Mountain Research Station, USDA Forest Service
Albuquerque, New Mexico, United States

1. **Agriculture and Urbanization**
2. **Competition for Energy Resources**
3. **Wood Energy**
4. **Wind and Water Energy**
5. **The Transition to Coal Energy**
6. **Steam Power**
7. **Steam Engines and Transportation**
8. **Petroleum**
9. **Electrical Power**
10. **Energy flow and Industrial Momentum**

Glossary

agricultural revolution The period prior to England's industrial revolution, when significant improvements in agricultural production were achieved through changes in free labor, land reform, technological innovation, and proto-industrial output.

capitalism A socioeconomic and political system in which private investors, rather than states, control trade and industry for profit.

industry Economic activity concerned with processing raw materials and producing large volumes of consistently manufactured goods, with maximum fuel exploitation.

industrial revolution The period of technological innovation and mechanized development that gathered momentum in 18th-century Britain; fuel shortages and rising population pressures, with nation-states growing in power and competing for energy, led to dramatic improvements in energy technology, manufacture, and agriculture and to massive increases in the volume of international trade.

plague Referred to as the Black Death, or Bubonic Plague, which killed one-third the population of England and Europe from the mid-14th century until around 1500; millions of people perished, and commerce, government, energy production, and industry waned as farms were abandoned and cities emptied.

proto-industry Supplemented agricultural income via the production of handcrafted work; this piecework was especially common in the creation of textiles by spinning, weaving, and sewing.

steel A strong hard metal that is an alloy of iron with carbon and other elements; it is used as a structural material and in manufacturing.

work energy Defined by natural scientist James Joule as kg m^2/s^2 or a joule (J); it enabled physicists to realize that heat and mechanical energy are convertible.

Population growth and competition for resources have continuously been an impetus for the control and use of energy. In harnessing energy, inventions evolved from previous innovations, ebbing and flowing with changes in population pressures and the availability of resources—necessity shaping history. For millennia, societies depended on resources provided by solar energy (radiation stored as vegetation), wind (a type of solar energy), animal and human energy, water, and fossil fuels. The advent of mechanized labor and its synchronized production demands transformed the lives of workers previously tuned to seasonal energy rhythms. Technological innovations during the early 18th century resulted in enormous growth in the economic wealth and sociopolitical power of nation-states. Sustained growth in Western Europe and Britain spread via exploration and trade. The competition to colonize and control the raw natural resources in the Caribbean and the Americas, in conjunction with imperialism in Indochina, the Mideast, and the Mediterranean, changed the global energy map.

1. AGRICULTURE AND URBANIZATION

Preindustrial economies in Britain and Europe were widespread in scale, dominated by agricultural

 © 2004 Elsevier Inc. All rights reserved.

production, processing, and manufacture, with populations based primarily in rural villages. A range of processed goods were essential to farms and households, although restricted incomes reduced purchasing power. Solar energy, water, regional fuels, and the ebb and flow of the seasons dictated the energy of labor expenditure in workers' lives.

There was little capital accumulation or technological progress during the widespread population destruction of the Black Death in medieval Britain and Europe. In Italy, Spain, France, and Poland, agricultural output per worker diminished, and nonagricultural employment tended to be based in rural areas, corresponding with only minor increases in urban populations. The subsequent recovery of populations after the plague led to the growth of urbanization, and local and national government structures increased their control of energy resources.

The Netherlands, Belgium, and England were distinguished from the rest of Europe by the interrelated factors of rising agricultural productivity, high wages, and urbanization. At the end of the medieval period, rural industry and urban economies expanded. Despite rates of population growth exceeding those elsewhere in Europe, economic development offset population growth, wages were increased or maintained.

By 1750, Britain had the highest portion of population growth in cities, and agricultural output per worker surpassed Belgian and Dutch levels. Urban and rural nonagricultural populations grew more rapidly in England than anywhere else in Europe. The share of the workforce in agriculture dropped by nearly half, while the trend of providing real wages rose.

The wool and cotton textile industry was the first widespread manufacturing enterprise in Britain. When farm laborers began producing piecework for cash, the social and labor organization of agrarian economies changed dramatically. Labor–market interactions and productivity growth in agriculture required a degree of labor mobility that was incompatible with previous agricultural economic models such as serfdom. England's division between landlords and tenant farmers caused reorganization to occur through reductions in farm employment, enclosure, and farm amalgamation. In Europe, where landowners/occupiers were more prevalent, farm amalgamation and enclosure were disruptive to productivity and institutional change was less profitable.

Rural industrialization was a response to population growth combined with increased urban demands for farm products, and sustained urbanization created dependency on labor based in the countryside. Substantial migrations of people to cities, and the spread of industry to the countryside, provided opportunities to escape the restrictions of guilds.

Providing enough farm produce for cities required increasing agricultural output and value per acre and introducing new crops. By the dawn of the 18th century, the industrial and commercial sectors of the British economy contributed at least one-third of the gross national product.

2. COMPETITION FOR ENERGY RESOURCES

For millennia, the Mediterranean had been the center of international trade, with trade routes among Europe, Africa, and Asia. After the plague during the 15th and 16th centuries, Mediterranean networks of international trade with Europe declined. European nation-states were not powerful enough to govern beyond their transient borders, and inadequate local resources could not support population growth. This led to each state employing protective measures to stimulate domestic industrial production, self-sufficiency, and competition for resources. During the 16th and 17th centuries, these policies were formalized, extended, and consolidated, severely limiting the growth of trade.

The high costs of inland transportation and restrictive national practices inhibited trade opportunities and led to an economic crisis. As trading opportunities ceased to expand after the plague's decimations, the need for new sources of revenue provided the impetus to search far afield for resources. With the diminishing extent of accessible markets, Europeans explored the Atlantic, Africa, the Caribbean, and the Americas, where raw resources and cheap commodity production offered immense opportunities for trade expansion. During the 15th century, the Portuguese established trading posts along the west coast region of Africa, produced sugar on slave-worked plantations, and traded gold.

The advent of the Atlantic world trading system and its abundant resources extended the production and consumption frontier of Western Europe. Because available transportation of goods was slow and expensive, unit cost of production in the Americas had to be sufficiently low for commodities to bear the cost of trans-Atlantic transportation. Large-scale production in the Americas increasingly depended on coerced labor. Indigenous peoples of the Americas were forced into slavery to mine silver in the Spanish

colonies and provision Europeans. Native American land was appropriated, and populations were decimated by an unwelcome import, Old World diseases. Colonizers obtained vast quantities of acreage, becoming land rich beyond the limits of European caste and class systems. The average population density in the Americas was less than one person per square mile during the 17th century.

As indigenous peoples in the colonies suffered disease, displacement, and slavery, the production of commodities for Atlantic commerce became more dependent on slaves from Africa. Labor costs on plantations throughout the Caribbean and the Americas were below subsistence costs, enabling large levels of production and maximum exploitation of human energy resources. From Barbados to the Carolinas and the Mississippi delta, a major source of power were the thousands of enslaved Africans and Native Americans as well as indentured servants from England, Scotland, and Ireland. Plantations prospered due to the high value of cotton in the world market and by producing coffee, sugar, cocoa, tobacco, and rum. As these export commodities increased in volume, prices fell in Europe, with the commodities becoming commonly consumed goods rather than rarified luxuries. The production of raw materials for mass markets contributed greatly to the development and economic homogenization of industry and energy exploitation.

England combined naval power and commercial development to secure choice Atlantic territories and create advantageous treaties to control the colonies' resources. The majority of commodity shares, production, and trade of the slave-based economy of the Atlantic world trading system were primarily British controlled.

The estimated percentage share of these commodities produced by enslaved Africans and indentured servants in the Americas was 50 to 80%. Between 1650 and 1850, the slave-based economy of the Atlantic system transformed international trade, transformed European and British economics, and spurred technological innovation. Slavery failed as an energy source due to the socioeconomic upheavals that eventually prevented and censured its expansion as well as the fact that it could not compete with developing sources of industrial energy such as steam power. Slave labor, rather than free labor, failed as a socioeconomic source of power.

In per capita terms, the exposure of England's economy and society to the development of the Atlantic world market was greater than that of any other country, although Europe profited immensely from participation in the slave-based Atlantic world economy. Newly acquired gold and silver from the Americas bolstered weak European economies. Imported textiles from England were welcomed in the Americas, with these sustained sales creating employment in England's manufacturing regions and stimulating population growth and changes in agrarian social structures. Combined with export demands, this created a ripe environment for transforming the organization and technology of manufacturing in domestic and export industries. This international economic energy propelled the technological industrial revolution.

The transition of intensified industrialization depended on improvements in transportation systems to provide dispersion networks of fuels and raw materials. With the advent of the Victorian empire, improvements in the infrastructure of transportation (e.g., roads, canals, shipping) spread. The construction of railroad systems in Britain, Europe, Asia, Russia, Africa, and the Americas connected, amalgamated, and exploited energy resources in most regions of the world, creating an engine of economic dynamism that itself was a product of mechanized industry.

3. WOOD ENERGY

Woodlands have been an essential resource for constructing homes, vehicles, ships, furniture, tools, and machines as well as for providing heat for households and cooking. Poor management of woodlands caused low productivity, and as timber stands became more remote, logging and transporting timber became less economical. During the 13th century in England, lime kilns devoured hundreds of ancient oaks in Wellington Forest. By the 1540s, the salt industry was forced to search far afield for the wood used in its processing furnaces. British navel power judged this shortage to be a national security threat. Authorities attempted to impose stiff fines for timber poaching and forbade bakers, brewers, and tile makers from purchasing boatloads of wood.

Severe timber shortages reached national crisis proportions in Britain, impelling the import of expensive timber from the Baltic and North America. Native turf, peat, reeds, gorse, and other plant fuels simply could not provide adequate supplementation. The climate did not cooperate. Between around 1450 and 1850, winters became unusually long and harsh, the Thames river froze, and diseases and cold increased in their killing power. Populations suffered

as low birth rates could not keep up with expanding death rates.

In Scotland, woodlands were carefully managed during the early medieval period, and native oak supplied the needs of developing burghs. Norse incursions pushed populations inland, allowing woodland regeneration in coastal areas. During the 15th century, the quality and quantity of native oak dwindled, and alternative sources (e.g., imported oak, conifers) were increasingly exploited. France had abundant internal supplies of timber, so there was no need to invest time, money, and politico–military resources in obtaining timber from elsewhere.

By the close of the 17th century, industry and manufacturing devoured more than a third of all the fuel burned in Britain. The textile industry's immense scale made it a leading consumer of fuels, although fuel cost was not a major constituent in manufacturing textiles, dyeing, calendaring, and bleaching. Energy consumption by industries increased in local and national terms because industry was not uniformly distributed, and this exacerbated and caused severe local and regional scarcity. Local needs concerning brewing, baking, smithing, potting, and lime processing had survived on whatever fuel was available. High temperatures were required to brew ale and beer and to process salt and sugar. The construction industry depended on kilns and furnaces for the production of bricks, tiles, glass, firing pottery, and burning lime. Smelting and working metals, and refining the base compounds of alum, copperas, saltpeter, and starch, were expensive processes.

Even when located within extensive woodlands, levels of consumption by iron forges, glass factories, and lime kilns led to shortages. Fuel needs grew voraciously as industries expanded in scale and became more centralized and urbanized. There was a fine balance between fuel needs in communities and industrial demands. Urban centers consisting of large concentrations of consumers inherently tended to outgrow the capacity of adjacent woodland fuel supplies. Transport costs commensurately inflated as supply lines were extended.

4. WIND AND WATER ENERGY

Windmills were developed in Persia by the 9th century BCE. This technological innovation was spurred by the need to mill corn in areas lacking consistent water supplies. Early windmills used an upright shaft, rather than a horizontal one, to hold the blades. This system was housed in a vertical adobe tunnel with flues to catch the wind (similar to a revolving door). The concept of harnessing wind energy via windmills reached Britain and Europe by 1137, where it underwent significant changes, with horizontal (rather than vertical) shafts and horizontal rotation. A form of solar energy, climate-dependent wind power has been unreliable and difficult to accumulate and store.

Roman aqueducts provided a system of fast-flowing water to watermills. By the 1st century BCE, Romans used an efficient horizontal axis where the waterwheel converted rotary motion around a horizontal axis into motion around a vertical axis; water passed under the wheel, and kinetic water energy turned the wheel. Draft animals and waterpower were harnessed to turn millstones. Rudimentary primitive watermills used vertical axes, with millstones mounted directly to waterwheel shafts. In mountainous regions, water was channeled through a chute into the wheel. Although power output was minimal ($\sim$300 W), this type of watermill was used in Europe until the late medieval period.

During the late 18th and early 19th centuries, waterwheel efficiency was improved. English engineer John Smeaton, expanding on the systems of early Roman waterwheels, found that more energy was generated when water was delivered to the top of the wheel via a chute. An overshot waterwheel used both kinetic and potential water energy and was capable of harnessing 63% more potential energy than was the undershot. The growing demand for energy intensified use of the waterwheel. While toiling in the gold mines of California during the 1870s, a British mining engineer, Lester Pelton, discovered that a waterwheel could be made more powerful if high-pressure jets of water were directed into hemispherical cups placed around its circumference. This waterwheel design, generating 500 horsepower (hp), was used 20 years later in the sodden mines of Alaska.

5. THE TRANSITION TO COAL ENERGY

Coal had been widely harvested and exploited as an energy source for thousands of years. In Asia, coal provided little illumination but plenty of heat. Romans in Britain found that stones from black outcrops were flammable, and Roman soldiers and blacksmiths used this coal to provide heat. During the 13th and 14th centuries, a multitude of industries

became dependent on a regular supply of coal. In salt production, coal was used to heat enormous pans of sea water. If woodlands had not been devoted nearly exclusively to the production of timber, their yield could not have met the fuel value generated by that produced by coal mines. It took an annual yield from 2 acres of well-managed woodlands to equal the heat energy of a ton of coal. Timber became more scarce and expensive to procure during the Tudor and Stuart periods. Consequently, the processes of salt boiling, ironworking, brewing, baking, textile manufacturing, and lime burning adopted coal, rather than wood, as a readily available and economical fuel source.

Confronted by the unrelenting scarcity and high price of timber, manufacturers in fuel-intensive industries were further prompted to restrain costs and enhance profits by switching to coal. The problems caused by burning coal in processes previously dependent on wood (e.g., producing glass, bricks, tiles, pottery, alum, copperas, saltpeter, and paper; smelting lead, tin, and copper; refining soap and sugar) swelled the demand for progressive technology. Increased energy efficiency in industrial processes stimulated competition, innovation, and the emergence of fuel economies. Urbanization expanded, and the availability of inexpensive coal enabled agriculture to supplant arboriculture.

During the late 17th century, British coal mining employed 12,000 to 15,000 people, from miners to those involved in the transportation and distribution of coal to the consumer. During this period, coal prices remained virtually constant and were not dependent on revolutionary changes in structure or technology, and mine shaft depth did not increase dramatically (safe ventilation of deep pits did not occur until the 19th century). Conservatively estimated, coal production increased 370% between 1650 and 1680, creating an environment in which profound industrial advances in mining and trade occurred.

The substantial expansion of the textile industry was accommodated within preexisting systems of production, relying on traditional procedures, artisan workshops, and piecework arrangements. Advances were built on the potential that had existed in preindustrial economies, while labor and finance, marketing, transportation, and distribution evolved. Industrial enterprises incorporated factories, warehouses, forges, mills, and furnaces, and small collieries produced thousands of tons of coal per year. By 1750, the British population had nearly doubled, and urbanization and industrialization continued to amalgamate.

During the 1790s, French engineer Philippe Lebon distilled gas from heated wood and concluded that it could provide warmth, conveniently light interiors with gas in glass globes (fuel distributed via small pipes within walls), and inflate balloons. Gregory Watt, son of inventor James Watt, journeyed to France to investigate these experiments. In Cornwall, England, William Murdock lit a house with gas generated from coal. Concerned with commercial possibilities, he experimented with coal of varying qualities, different types of burners, and devices that could store enough gas to make portable light. In 1798, he returned to the foundry in Soho, England, where he illuminated the main building with coal gas using cast-iron retorts.

By 1816, London had 26 miles of gas mains, and soon there were three rival gas companies. With the availability of kerosene and gas light, the whale oil industry futilely attempted to compete. The advent of gas light enabled Victorian readers to become increasingly literate. During the 1860s, gas utilities spread in England and Europe, while the eastern United States and Canada adopted gas lighting for streets. At the turn of the 20th century, coal gas cooking and radiant heat to warm homes from gas fires became more common. The gas industry was determined to increase efficiency in all of these domestic and industrial heating functions.

Massive petroleum and gas industries resulted from the need for improved lighting during the industrial revolution. Although oil lamps had improved, the illuminating capabilities of vegetable and animal oils were limited. Gas light was superior but was restricted to urban areas. Gas oil, consisting of the fraction intermediate between kerosene and lubricants, was inexpensive and less dangerous than petroleum and also provided a convenient means of enriching the coal gas.

At the turn of the 20th century, when the automobile industry was still rudimentary, crude petroleum and rubber remained insignificant, while coal gas reached its zenith of full technological development as an energy source. Versatile and inexpensive kerosene continued to provide illuminating energy in much of the world, while gas and electricity competed in urban centers. As an illuminant, gas light began to lose momentum, while electrical power grew enormously.

The turning point for the British coal industry, and a defining moment of the industrial revolution, came with the innovation of the steam engine. It was used in processing both coal and metals, and by the close

of the 18th century more than 15 million tons of coal were produced annually.

6. STEAM POWER

In the 6th century BCE, Greek philosophers described the universal elements as consisting of air, earth, fire, and water. It was established that water displaced air and that steam condensed back to liquid and could rotate wheels. During the 17th century, the physical nature of the atmosphere was explored in qualitative rather than quantitative terms; water and other liquids entering vacuous space was attributed to nature abhorring a vacuum. The persistent problem of mine drainage inspired understanding of atmospheric pressure, leading to the earliest steam engines.

When engineers failed to drain excess water from Italian mines during the 1600s, Galileo was enlisted. In 1644, his pupil, Evangelista, found that atmospheric pressure was equal to that of a column of mercury 30 inches high, corresponding to a column of water 30 feet high, and that atmospheric pressure falls with increasing altitude. Torricelli postulated that a method of repeatedly creating a vacuum would enable atmospheric pressure to be a source of power and could be used in mine drainage. Steam at l00°C equals sea-level atmospheric pressure and so can raise water 30 feet. Tremendous pressure can be generated by superheating water to create high-pressure steam; at 200°C, pressure is 15 times greater than at l00°C, lifting water 450 feet.

The first steam engines were termed atmospheric engines. In 1680, Christian Huygens, a Dutch scientist, designed a system where gunpowder exploding in a cylinder created hot gases that were expelled through relief valves. On cooling, the valves closed and a partial vacuum existed within the cylinder. The gunpowder gas occupied a smaller volume than when heated; thus, atmosphere pressure drove a piston in the cylinder. In 1690, Huygens's assistant, Denis Papin, found that 1 volume of water yields 1300 volumes of steam. A vacuum could be achieved by converting steam back into water via condensation, propelling a piston up and down in a cylinder.

Prolific inventor Thomas Savery improved on this in 1698, designing a steam pump using heat to raise water, generating approximately 1 hp. Steam from a boiler passed through a valve-fitted pipe into a vessel full of water, expelling the water up through a second pipe. When steam filled the vessel, it was condensed by pouring cold water onto the outer surface. This created a partial vacuum. When the vessel was connected with another pipe with water at a lower level, atmospheric pressure forced the water up. Savery's engine pumped water for large buildings and waterwheels, but its maximum lift was inadequate for mine drainage.

Thomas Newcomen, an inventor and blacksmith, adopted Papin's cylinder and piston and built his own version of the steam pump, replacing costly horse-drawn pump engines. Newcomen's direct experience with drainage problems in Cornish tin mines led to his construction of an efficient engine using atmospheric pressure rather than high-pressure steam. The Newcomen system had a boiler produce steam at atmospheric pressure within a cylinder, where a pump rod hung from a beam so as to apply weight to drive a piston. When the cylinder was full of steam and closed by a valve, a jet of cold water condensed the steam and atmospheric pressure forced the pump rod to push the piston down. Generating approximately 5 hp, this engine was a success in draining coal mines, lifting 10 gallons of water 153 feet through a series of pumps. Its primary users were coal mine proprietors who ran their engines on low-grade coal.

With improved control of mine flooding, the depth of mines and the coal industry grew exponentially. Use of the steam engine spread throughout Europe and to the American colonies. Newcomen's design was so reliable that even as other fuel technologies were developed, the last engine of this type worked for more than a century in Yorkshire, England, without serious breakdown (it was finally dismantled in 1934). Steam power was also used to pump water for waterwheels in cotton mills, accommodating the growing energy needs of textile manufacture as well as a multitude of other industries.

Between 1769 and 1776, James Watt discovered that energy was lost in the process of cooling and reheating the Newcomen engine. The addition of a separate condenser cylinder permitted a steady injection of steam into the first cylinder without having to continuously reheat it. In 1774, Watt and Matthew Boulton, an English manufacturer in Soho, became partners, providing Watt with skilled engineers and workers.

An impediment to steam engine design was the shortage of precise cylinders; imprecise boring allowed steam to escape between the cylinder walls and piston. John Wilkinson designed a mill for boring iron cannons that was capable of making

precise cylinders for any purpose. His blast furnaces, with air from Watt's engine, produced consistent cylinders. Using this newfound precision in cylinder construction, the rotative Watt engine extracted four times as much energy, a minimum of 10 hp, from coal as the Newcomen model. By the turn of the 19th century, Watt's and Boulton's Soho factory had produced hundreds of engines powering industrial machinery.

Improvements in steam technology continued to be made during the 19th century. In 1802, a Cornish mining engineer, Richard Trevithick, built a small pumping engine with a cylinder 7 inches in diameter, a cast-iron boiler 1.5 inches thick, and steam pressure of 145 pounds per square inch—10 times atmospheric pressure. Two years later, Trevithick built 50 stationary engines and the first successful railway locomotive. By 1844, high-pressure steam engines averaged 68 million foot pounds per bushel of coal, compared with 6 million foot pounds capacity for the Newcomen engine. When Watt's engine was combined with Trevithick's design, a pressure of up to 150 pounds per square inch was achieved, referred to as "compounding." Trevithick's steam engine eliminated the huge rocking beam of Watt's design and became the new standard, suited to factories with limited space. His engines were unrivaled and remained in a variety of industrial purposes, including pumping, iron works, corn grinding, and sugar milling, until the end of the 19th century.

7. STEAM ENGINES AND TRANSPORTATION

Steam-powered engines were crucial in the development of transportation. Roads often were little more than deeply gouged tracks, slowing people, carriages, and carts in an energy-hindering quagmire. The growing demands for bulk fuel beyond what was locally available, with the high cost of transport added to the expense of energy resources, expedited innovation.

During the 1700s, British canals were improved, and tow paths allowed heavily burdened vessels to be pulled by teams of horses. Combining water and steam power, a paddlewheel steamboat ascended a river in 1783 near Lyons, France. William Symington, an English engineer, constructed an atmospheric steam engine to propel a small boat in 1788 using a horizontal double-acting cylinder with a connecting rod driving the paddlewheel's crankshaft. In the United States, Robert Fulton used a Boulton and Watt engine to operate a commercial steamer, and during the 1840s the U.S. Navy adapted this propulsion system in military craft, locating the engine below the waterline for protection.

In 1801, Trevithick built his first steam carriage (later driving one wildly through the streets of London). He combined the steam locomotive with railways, demonstrating its usefulness in conveying heavy goods and passengers. Rails had already been constructed for horses to tow carriages and coal carts more efficiently than on unreliable roads. The development of railway passenger services was a crucial formative factor in the phenomenon of the industrial revolution. Horsepower was replaced by the more powerful steam engine. By 1830, George Stephenson developed a locomotive railway line between Liverpool and Manchester without the need for stationary steam engines or horses in steep areas to supplement power. Railroad carriages were finally capable of hauling tons of coal and hundreds of passengers. Britain exported locomotives to Europe and America, where railway systems were rapidly established. Rail speeds exceeding 60 miles per hour were common, and engine efficiency increased, requiring only 23 pounds of coal per mile, although countries with abundant forests, such as the United States and Canada, still used wood fuel. Improved materials and methods of manufacture, as well as the steam turbine, resulted in higher power/weight ratios and increased fuel economy. During the 1880s, industrial steam power transformed networks of world communication, extending and unifying international economic systems.

8. PETROLEUM

Coal gas and petroleum energy share the property of having hydrocarbons as their principal constituents; coal gas is rich in the hydrocarbon methane, whereas petroleum consists of a complex mixture of liquid hydrocarbons. The use of petroleum products long antedates the use of coal gas. In Mesopotamia, concentrations of petroleum deposits provided commerce and production. Babylonians called inflammable oil "naphtha." For thousands of years, mixtures containing bitumen were used in roads, ship caulking, floor waterproofing, hard mortar, and medicines. Naphtha was militarily important as a weapon, although Romans derived their bitumen from wood pitch. Aztecs made chewing gum made with bitumen, and during the 16th century oil from seepages in

Havana was used to caulk ships. By the 15th and 16th centuries, there was an international trade in petroleum.

Distillation of crude oil revealed that its derivatives were suitable in axle grease, paints and varnishes, dressing leather, lamp fuel, and mineral turpentine. A pliable caulking for ships was made by thinning pitch thinned with turpentine, and hot asphalt pitch and powdered rock were used in flooring and steps. During the 19th century, asphalt mixed with mineral oil was used in pavements and roads, and oil cloth and linoleum became popular and inexpensive household items.

During the early 19th century, deep exploratory drilling for water and salt required sufficiently hard drills and mechanical power. Steam engines provided the energy, and the development of derricks aided in the manipulation of rapid-drilling machines. Prospecting for oil ceased to be dependent on surface seepages when hollow drills enabled drill samples to be removed, revealing the structure of underground formations. The compositions of crude oil varied; thus, methods of oil refinery were important. During the 1860s, the United States had a steady supply of 10 million barrels of petroleum annually. By 1901, Russian crude oil output reached 11 million tons, and petroleum-fueled railways connected world markets.

Kerosene was distilled from petroleum with sulfuric acid and lime, providing inexpensive and plentiful fuel for lamplight. Although gasoline remained a worthless and inflammable by-product of the industry, oil derivatives provided lubricants that became increasingly necessary with innovations in vehicles and machinery, and heavy machine oils were used on railways.

9. ELECTRICAL POWER

In early Greece, it was found that the *elektron* (amber), when rubbed, acquired the power of attracting low-density materials. The widespread use of electricity for heat, light, and power depended on the development of mechanical methods of generation. During the 17th and 18th centuries, static electricity was found to be distinct from electric currents. Electricity could be positive or negative as charged bodies repelled or attracted each other, distinguishing conductors from nonconductors.

In 1754, John Canton devised an instrument to measure electricity based on the repulsion of like-charged pith suspended by threads, and this was later standardized and redesigned as the gold leaf electroscope. Benjamin Franklin identified the electrical discharge of lightning, leading to the invention of the lightning conductor. In Italy, Allesandro Volta, a natural scientist, found that electricity was derived when alternate metal plates of silver or copper and zinc contacted each other within a solution. The importance of this discovery was that it provided a source of continuous electric current. The original voltaic energy battery gave stimulus to the experimental study of electricity. In a matter of months, laboratories produced electric batteries, converting the energy released in chemical reactions.

In 1820, Danish physicist H. C. Oersted described the magnetic field surrounding a conductor carrying an electric current. The relationship between the strength of a magnetic field and its electric current established that a continuous conductor in a magnetic field causes electric currents to flow. From 1834, rotating coil generators were made commercially in London. The earliest generators produced alternating currents. Converting alternating energy into direct electric currents was resolved via a mechanical commutator and rectangular coils that rotated in a magnetic field. Maximum voltage in each coil was generated in succession, alleviating the irregularities at a given speed of rotation providing constant voltage.

By 1825, electromagnets were used as an alternative to permanent magnets by the founder of the first English electrical journal, William Sturgeon. Electromagnets possessed enough residual magnetism in their iron cores to provide the necessary magnetic field to start output from an electric generator. The electric generator became a self-contained machine that needed only to be rotated to produce electrical energy. The application of a steam engine to rotate the armature generated enough electricity to supply arc lamps for lighthouses, bringing large-scale use of electricity a step closer.

An armature using a continuous winding of copper wire increased the possibilities of using arc lamps for lighting streets. Arc lamps developed after the discovery that a spark struck between two pieces of carbon creates brilliant light. Inventors Thomas A. Edison and Sir Joseph Swan developed filament lamps for domestic use during the 1870s, and an incandescent filament lamp used an electric current to heat its conductor.

The growing needs of industrial light required electric power generators to be substantial in size. In New York, Edison's electric-generating station was in

operation in 1882. Electric storage batteries were used for lighting railway carriages and propelling vehicles.

At Niagara Falls during the 1890s, George Westinghouse developed the first large-scale hydroelectric installation, with a capacity of 200,000 hp. Hydroelectric generators demanded a considerable capital outlay for successful operation, so coal-fired steam engines lingered as a favored energy source at the end of the 19th century. During the late 1800s, the electricity industry expanded, output per electric power stations tripled, and coal consumption per unit of electricity fell by nearly half.

Conductors consisting of concentric copper tubes were used with alternating currents because copper had a lower electrostatic capacity and the concentric cable had no inductive effect on neighboring telegraph, telephone, or other electrical installations. The electric telegraph required a reliable battery cell to provide constant voltage that was capable of prolonged output. Attendant construction needs for telegraph poles required the production of enormous quantities of timber, porcelain insulators, and rubber insulation for cables.

Local electrical generation evolved into large centralized distribution utilities, and surplus electricity was sold to local consumers. In 1889, the London Electricity Supply Corporation's power station provided high voltage by operating four 10,000-hp steam engines to fuel 10,000-volt alternators. The economic advantages of central power stations to generate electricity at high voltages for serving large areas brought new practical and economic problems of distribution and, thus, further innovations. By the end of the 19th century, underground distribution systems of electricity had color-coded cables. Each strand was identified in this manner, a major consideration given that the electrical network beneath city streets increased in complexity.

10. ENERGY FLOW AND INDUSTRIAL MOMENTUM

During the early industrial period, increases in scale and diversity of manufacturing and trade corresponded with a shortage of readily available consistent energy, and this pressure induced developments in fuel technology industries. Competition for dwindling resources spurred the expansion of international trade and exploration, and fuel efficiency became more crucial. The fusion between mechanized growth and economic fuel use created the momentum of energy exploitation and technological innovation, and socioeconomic systems became known as the industrial revolution. Transportation systems connected a myriad of worldwide energy sources, and improved communication expedited the process of efficient global fuel exploitation and industry. Global commerce and fuel exploitation decisively affected the flow of materials via worldwide transportation and trade systems. As technological innovations were applied to harvesting energy and the subsequent momentum of labor mechanization, production uniformity and the quantity of fuel used in mass processing vast quantities of goods coalesced into an enormous worldwide energy flow.

SEE ALSO THE FOLLOWING ARTICLES

Coal Industry, History of • Electricity Use, History of • Energy in the History and Philosophy of Science • Technology Innovation and Energy • Thermodynamic Sciences, History of • Transitions in Energy Use • Wind Energy, History of • Wood Energy, History of • World History and Energy

Further Reading

Cotterell, B., and Kamminga, J. (1990). "Mechanics of Pre-Industrial Technology." Cambridge University Press, Cambridge, UK.

Crone, A., and Mills, C. M. (2002). Seeing the wood and the trees: Dendrochronological studies in Scotland. *Antiquity* **76,** 788–794.

Derry, T. K., and Williams, T. I. (1960). "A Short History of Technology from the Earliest Times to A.D. 1900." Oxford University Press, Oxford, UK.

Fields, G. (1999). City systems, urban history, and economic modernity: Urbanization and the Transition from agrarian to industrial society. *Berkeley Planning J.* **13,** 102–128.

Freese, B. (2003). "Coal: A Human History." Perseus Publishing, Cambridge, MA.

Harman, P., and Mitton, S. (eds.). (2002). "Cambridge Scientific Minds." Cambridge University Press, Cambridge, UK.

Harrison, R., Gillespie, M., and Peuramaki-Brown, M. (eds.). (2002). "Eureka: The Archaeology of Innovation and Science—Proceedings of the Twenty-Ninth Annual Conference of the Archaeological Association of the University of Calgary." Archaeological Association of the University of Calgary, Alberta, Canada.

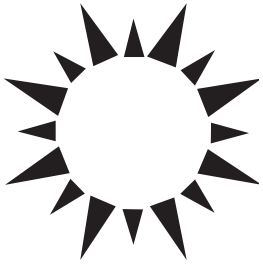
Hatcher, J. (1993). "The History of the British Coal Industry," vol. 1: "Before 1700: Towards the Age of Coal." Clarendon, Oxford, UK.

Licht, W. (1995). "Industrializing America: The Nineteenth Century." Johns Hopkins University Press, Baltimore, MD.

Nef, J. (1964). "The Conquest of the Material World: Collected Essays." Chicago University Press, Chicago.
Theibault, J. (1998). Town and countryside, and the proto-industrialization in early modern Europe. *J. Interdisc. Hist.* **29,** 263–272.
Toynbee, A. (1956). "The Industrial Revolution." Beacon, Boston.
Wallerstein, I. (1974). "The Modern World-System I: Capitalist Agriculture and the Origins of the European World-Economy in the Sixteenth Century." Academic Press, San Diego.
Wallerstein, I. (1980). "The Modern World-System II: Mercantilism an the Consolidation of the European World-Economy 1600–1750." Academic Press, San Diego.
Wallerstein, I. (1989). "The Modern World-System III: The Second Era of Great Expansion of the Capitalist World-Economy 1730–1840s." Academic Press, San Diego.
Zell, M. (1994). "Industry in the Countryside: Wealden Society in the Sixteenth Century." Cambridge University Press, Cambridge, UK.

Earth's Energy Balance

KEVIN E. TRENBERTH
National Center for Atmospheric Research
Boulder, Colorado, United States



1. **The Earth and Climate System**
2. **The Global Energy Balance**
3. **Regional Patterns**
4. **The Atmosphere**
5. **The Hydrological Cycle**
6. **The Oceans**
7. **The Land**
8. **Ice**
9. **The Role of Heat Storage**
10. **Atmosphere–Ocean Interaction: El Niño**
11. **Anthropogenic Climate Change**
12. **Observed and Projected Temperatures**

Glossary

aerosol Microscopic particles suspended in the atmosphere, originating from either a natural source (e.g., volcanoes) or human activity (e.g., coal burning).

albedo The reflectivity of the earth.

anthropogenic climate change Climate change due to human influences.

anticyclone A high-pressure weather system. The wind rotates around anticyclones in a clockwise direction in the Northern Hemisphere and counterclockwise in the Southern Hemisphere. They usually give rise to fine, settled weather.

convection In weather, the process of warm air rising rapidly while cooler air usually subsides more gradually over broader regions elsewhere to take its place. This process often produces cumulus clouds and may result in rain.

cyclone A low-pressure weather system. The wind rotates around cyclones in a counterclockwise direction in the Northern Hemisphere and clockwise in the Southern Hemisphere. Cyclones are usually associated with rainy, unsettled weather and may include warm and cold fronts.

dry static energy The sum of the atmospheric sensible heat and potential energy.

El Niño The occasional warming of the tropical Pacific Ocean from the west coast of South America to the central Pacific that typically lasts approximately 1 year and alters weather patterns throughout the world.

El Niño–Southern Oscillation (ENSO) El Niño and the Southern Oscillation together; the warm phase is El Niño and the cold phase is La Niña.

enthalpy The heat content of a substance per unit mass. Used to refer to sensible heat in atmospheric science (as opposed to latent heat).

greenhouse effect The effect produced as certain atmospheric gases allow incoming solar radiation to pass through to the earth's surface but reduce the outgoing (infrared) radiation, which is reradiated from Earth, escaping into outer space. The effect is responsible for warming the planet.

greenhouse gas Any gas that absorbs infrared radiation in the atmosphere.

Hadley circulation The large-scale meridional overturning in the atmosphere in the tropics.

La Niña A substantial cooling of the central and eastern tropical Pacific Ocean lasting 5 months or longer.

long-wave radiation Infrared radiation in the long wave part of the electromagnetic spectrum, corresponding to wavelengths of 0.8–1000 μm. For the earth, it also corresponds to the wavelengths of thermal-emitted radiation.

short-wave radiation Radiation from the sun, most of which occurs at wavelengths shorter than the infrared.

Southern Oscillation A global-scale variation in the atmosphere associated with El Niño and La Niña events.

thermocline The region of vertical temperature gradient in the oceans lying between the deep abyssal waters and the surface mixed layer.

total solar irradiance The solar radiation received at the top of the earth's atmosphere on a surface oriented perpendicular to the incoming radiation and at the mean distance of the earth from the sun.

troposphere The part of the atmosphere in which we live, ascending to approximately 15 km above the earth's surface, and in which temperatures generally decrease with height. The atmospheric dynamics we know as "weather" take place within the troposphere.

 © 2004 Elsevier Inc. All rights reserved.

urban heat island The region of warm air over built-up cities associated with the presence of city structures, roads, etc.

The distribution of solar radiation absorbed on Earth is very uneven and largely determined by the geometry of the sun–earth orbit and its variations. This incoming radiant energy is transformed into various forms (internal heat, potential energy, latent energy, and kinetic energy) moved around in various ways primarily by the atmosphere and oceans; stored and sequestered in the ocean, land, and ice components of the climate system; and ultimately radiated back to space as infrared radiation. The requirement for an equilibrium climate mandates a balance between the incoming and outgoing radiation and further mandates that the flows of energy are systematic. These drive the weather systems in the atmosphere and the currents in the ocean, and they fundamentally determine the climate. They can be perturbed, causing climate change. This article examines the processes involved and follows the flows, storage, and release of energy.

1. THE EARTH AND CLIMATE SYSTEM

Our planet orbits the sun at an average distance of 1.50×10^{11} m once per year. It receives from the sun an average radiation of 1368 W m^{-2} at this distance, and this value is referred to as the total solar irradiance. It used to be called the "solar constant" even though it does vary by small amounts with the sunspot cycle and related changes on the sun. The earth's shape is similar to that of an oblate spheroid, with an average radius of 6371 km. It rotates on an axis with a tilt relative to the ecliptic plane of 23.5° around the sun once per year in a slightly elliptical orbit that brings the earth closest to the sun on January 3 (called perihelion). Due to the shape of the earth, incoming solar radiation varies enormously with latitude. Moreover, tilt of the axis and the rotation of the earth around the sun give rise to the seasons, as the Northern Hemisphere points more toward the sun in June and the Southern Hemisphere points toward the sun in late December. The earth also turns on its axis once per day, resulting in the day–night cycle (Fig. 1). A consequence of the earth's roughly spherical shape and the rotation is that the average energy in the form of solar radiation received at the top of the earth's atmosphere is the total solar irradiance divided by 4, which is the ratio of the earth's surface area ($4\pi a^2$, where a is the mean radius) to that of the cross section (πa^2).

The earth system can be altered by effects or influences from outside the planet usually regarded as "externally" imposed. Most important are the sun and its output, the earth's rotation rate, the sun–earth geometry and the slowly changing orbit, the physical makeup of the earth system such as the distribution of land and ocean, the geographic features on the land, the ocean bottom topography and basin configurations, and the mass and basic composition of the atmosphere and ocean. These components affect the absorption and reflection of radiation, the storage of energy, and the movement of energy, all of which determine the mean climate, which may vary due to natural causes.

On timescales of tens of thousands of years, the earth's orbit slowly changes, the shape of the orbit is altered, the tilt changes, and the earth precesses on its axis like a rotating top, all of which combine to alter the annual distribution of solar radiation received on the earth. Similarly, a change in the average net radiation at the top of the atmosphere due to perturbations in the incident solar radiation from the changes internal to the sun or the emergent infrared radiation from changes in atmospheric composition leads to a change in heating. Changes in atmospheric composition arise from natural events such as erupting volcanoes, which can create a cloud of debris that blocks the sun. They may also arise from human activities such as the burning of fossil fuels, which creates visible particulate pollution and carbon dioxide, a greenhouse gas. The greatest variations in the composition of the atmosphere involve water in various phases, such as water vapor, clouds of liquid water, ice-crystal clouds, and hail, and these affect the radiative balance of the earth.

The climate system has several internal interactive components. The atmosphere does not have very much heat capacity but is very important—the most volatile component of the climate system with wind speeds in the jet stream often exceeding 50 m s^{-1}—in moving heat and energy around. The oceans have enormous heat capacity and, being fluid, can also move heat and energy around in important ways. Ocean currents may be >1 m s^{-1} in strong currents such as the Gulf Stream, but are more typically a few centimeters per second at the surface. Other major components of the climate system include sea ice, the land and its features [including the vegetation, *albedo* (reflective character), biomass, and ecosystems], snow cover, land ice (including the semipermanent ice sheets of Antarctica and Greenland, and glaciers),

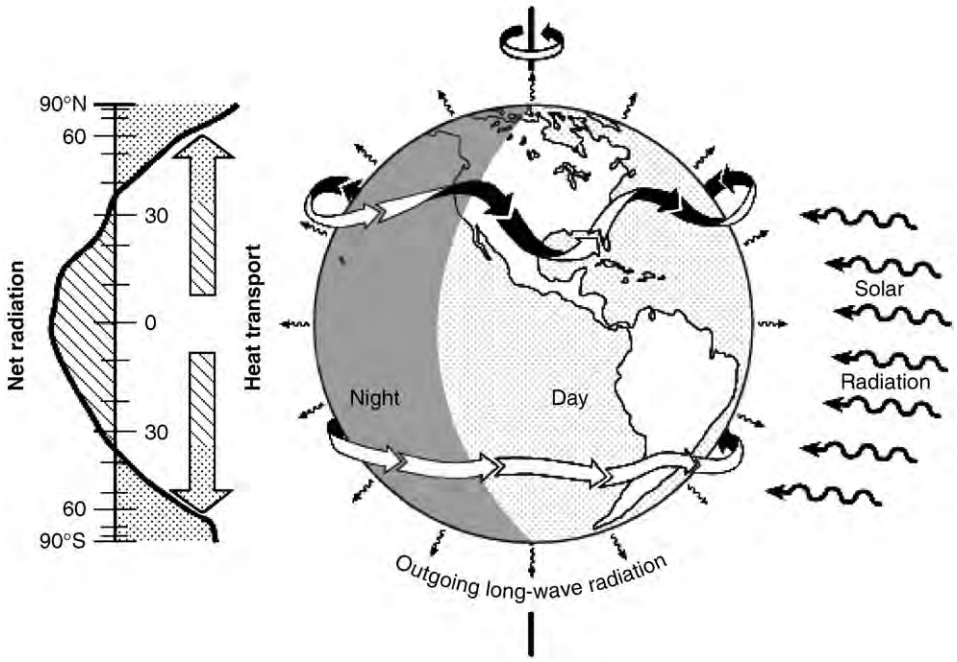


FIGURE 1 The incoming solar radiation (right) illuminates only part of the earth, whereas the outgoing long-wave radiation is distributed more evenly. On an annual mean basis, the result is an excess of absorbed solar radiation over the outgoing long-wave radiation in the tropics, whereas there is a deficit at middle to high latitudes (left) so that there is a requirement for a poleward heat transport in each hemisphere (arrows) by the atmosphere and the oceans. The radiation distribution results in warm conditions in the tropics but cold at high latitudes, and the temperature contrast results in a broad band of westerlies in the extratropics of each hemisphere in which there is an embedded jet stream (ribbon arrows) at approximately 10 km above the earth's surface. The flow of the jet stream over the different underlying surface (ocean, land, and mountains) produces waves in the atmosphere and geographic spatial structure to climate. [The excess of net radiation at the equator is 68 $W m^{-2}$ and the deficit peaks at -100 $W m^{-2}$ at the South Pole and -125 $W m^{-2}$ at the North Pole. From Trenberth *et al.* (1996).

and rivers, lakes, and surface and subsurface water. Their role in energy storage and the energy balance of the earth is discussed later.

Changes in any of the climate system components, whether internal and thus a part of the system or from external forcings, cause the climate to vary or to change. Thus, climate can vary because of alterations in the internal exchanges of energy or in the internal dynamics of the climate system. An example is El Niño–Southern Oscillation (ENSO) events, which arise from natural coupled interactions between the atmosphere and the ocean centered in the tropical Pacific. Such interactions are also discussed later from the standpoint of energy.

2. THE GLOBAL ENERGY BALANCE

The incoming energy to the earth system is in the form of solar radiation and roughly corresponds to that of a black body at the temperature of the sun of approximately 6000 K. The sun's emissions peak at a wavelength of approximately 0.6 μm and much of this energy is in the visible part of the electromagnetic spectrum, although some extends beyond the red into the infrared and some extends beyond the violet into the ultraviolet. As noted earlier, because of the roughly spherical shape of the earth, at any one time half the earth is in night (Fig. 1) and the average amount of energy incident on a level surface outside the atmosphere is one-fourth of the total solar irradiance or 342 $W m^{-2}$. Approximately 31% of this energy is scattered or reflected back to space by molecules, tiny airborne particles (known as aerosols), and clouds in the atmosphere or by the earth's surface, which leaves approximately 235 $W m^{-2}$ on average to warm the earth's surface and atmosphere (Fig. 2).

To balance the incoming energy, the earth must radiate on average the same amount of energy back to space (Fig. 2). The amount of thermal radiation

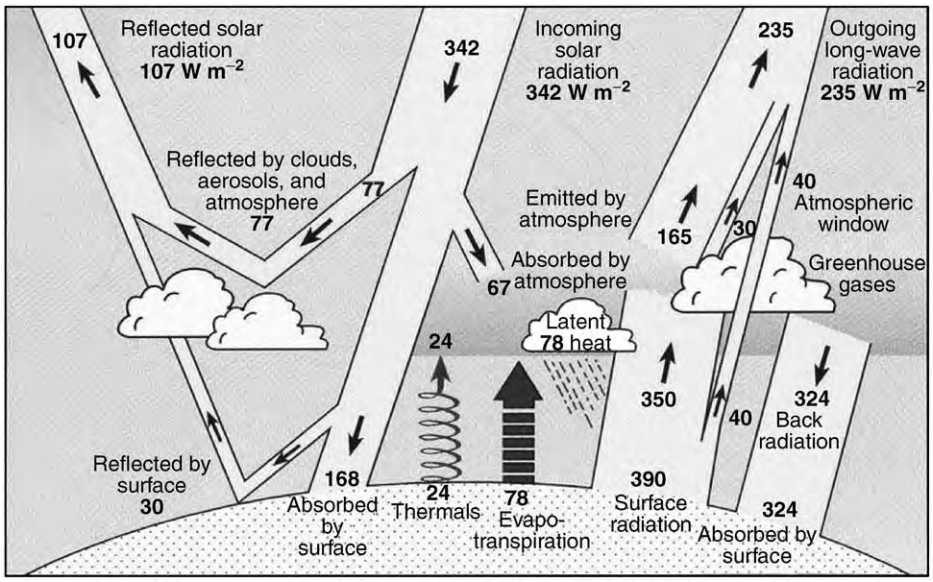


FIGURE 2 The earth's radiation balance. The net incoming solar radiation of 342 W m^{-2} is partially reflected by clouds and the atmosphere or at the surface, but 49% is absorbed by the surface. Some of this heat is returned to the atmosphere as sensible heating and most as evapotranspiration that is realized as latent heat in precipitation. The rest is radiated as thermal infrared radiation and most of this is absorbed by the atmosphere and reemitted both up and down, producing a greenhouse effect, since the radiation lost to space comes from cloud tops and parts of the atmosphere much colder than the surface. From Kiehl and Trenberth (1997).

emitted by a warm surface depends on its temperature and on how absorbing it is. For a completely absorbing surface to emit 235 W m^{-2} of thermal radiation, it would have a temperature of approximately −19°C (254 K). Therefore, the emitted thermal radiation occurs at approximately 10 μm, which is in the infrared part of the electromagnetic radiation spectrum. Near 4 μm, radiation from both the sun and the earth is very small, and hence there is a separation of wavelengths that has led to the solar radiation being referred to as short-wave radiation and the outgoing terrestrial radiation as long-wave radiation. Note that −19°C is much colder than the conditions that actually exist near the earth's surface, where the annual average global mean temperature is approximately 14°C. However, because the temperature in the lower atmosphere (troposphere) decreases quite rapidly with height, a temperature of −19°C is reached typically at an altitude of 5 km above the surface in midlatitudes. This provides a clue about the role of the atmosphere in making the surface climate hospitable.

2.1 The Greenhouse Effect

Some of the infrared radiation leaving the atmosphere originates near the earth's surface and is transmitted relatively unimpeded through the atmosphere; this is the radiation from areas where there are no clouds and which is present in the part of the spectrum known as the atmospheric "window" (Fig. 2). The bulk of the radiation, however, is intercepted and reemitted both up and down. The emissions to space occur either from the tops of clouds at different atmospheric levels (which are almost always colder than the surface) or by gases present in the atmosphere that absorb and emit infrared radiation. Most of the atmosphere consists of nitrogen and oxygen (99% of dry air), which are transparent to infrared radiation. It is the water vapor, which varies in amount from 0 to approximately 3%, carbon dioxide, and some other minor gases present in the atmosphere in much smaller quantities that absorb some of the thermal radiation leaving the surface and reemit radiation from much higher and colder levels out to space. These radiatively active gases are known as greenhouse gases because they act as a partial blanket for the thermal radiation from the surface and enable it to be substantially warmer than it would otherwise be, analogous to the effects of a greenhouse. Note that while a real greenhouse does work this way, the main heat retention in a greenhouse actually comes through protection from the wind. In the current climate, water vapor is estimated to account for approximately 60% of the greenhouse effect, carbon

dioxide 26%, ozone 8%, and other gases 6% for clear skies.

2.2 Effects of Clouds

Clouds also absorb and emit thermal radiation and have a blanketing effect similar to that of greenhouse gases. However, clouds are also bright reflectors of solar radiation and thus also act to cool the surface. Although on average there is strong cancellation between the two opposing effects of short-wave and long-wave cloud heating, the net global effect of clouds in our current climate, as determined by space-based measurements, is a small cooling of the surface. A key issue is how clouds will change as climate changes. This issue is complicated by the fact that clouds are also strongly influenced by particulate pollution, which tends to make more smaller cloud droplets, and thus makes clouds brighter and more reflective of solar radiation. These effects may also influence precipitation. If cloud tops become higher, the radiation to space from clouds is at a colder temperature and this produces a warming. However, more extensive low clouds would be likely to produce cooling because of the greater influence on solar radiation.

3. REGIONAL PATTERNS

The annual mean absorbed solar radiation (ASR) and outgoing long-wave radiation (OLR) are shown in Fig. 3. Most of the atmosphere is relatively transparent to solar radiation, with the most notable exception being clouds. At the surface, snow and ice have a high albedo and consequently absorb little incoming radiation. Therefore, the main departures in the ASR from what would be expected simply from the sun–earth geometry are the signatures of persistent clouds. Bright clouds occur over Indonesia and Malaysia, across the Pacific near 10°N, and over the Amazon in the southern summer, contributing to the relatively low values in these locations, whereas dark oceanic cloud-free regions along and south of the equator in the Pacific and Atlantic and in the subtropical anticyclones absorb most solar radiation.

The OLR, as noted previously, is greatly influenced by water vapor and clouds but is generally more uniform with latitude than ASR in Fig. 3. Nevertheless, the signature of high-top and therefore cold clouds is strongly evident in the OLR. Similarly, the dry, cloud-free regions are where the most surface radiation escapes to space. There is a remarkable cancellation between much of the effects of clouds on

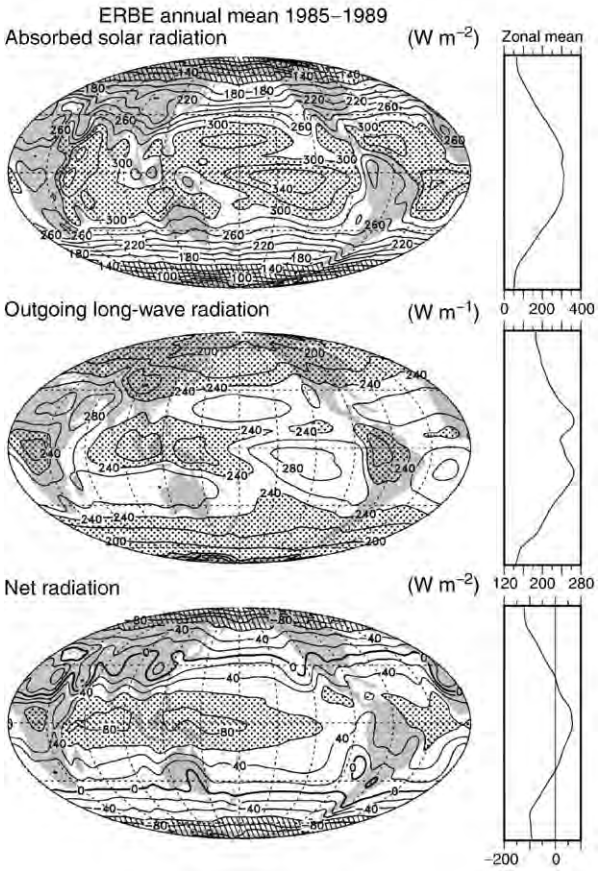


FIGURE 3 Maps of the annual mean absorbed solar radiation, outgoing long-wave radiation, and net radiation at the top of the atmosphere during the period between April 1985 and February 1989, when the Earth Radiation Budget Experiment was under way. From Trenberth and Stepaniak (2003).

the net radiation (Fig. 3). In particular, the high convective clouds are bright and reflect solar radiation but are also cold and hence reduce OLR. The main remaining signature of clouds in the net radiation from Earth is seen from the low stratocumulus cloud decks that persist above cold ocean waters, most notably off the west coasts of California and Peru. Such clouds are also bright, but because they have low tops they radiate at temperatures close to those at the surface, resulting in a cooling of the planet. Note that the Sahara desert has a high OLR, consistent with dry, cloud-free, and warm conditions, but it is also bright and reflects solar radiation, and it stands out as a region of net radiation deficit.

For the earth, on an annual mean basis there is an excess of solar over outgoing long-wave radiation in the tropics and a deficit at mid- to high latitudes (Fig. 1) that set up an equator-to-pole temperature gradient. These result, along with the earth's rotation,

in a broad band of westerlies and a jet stream in each hemisphere in the troposphere. Embedded within the midlatitude westerlies are large-scale weather systems that, along with the ocean, act to transport heat polewards to achieve an overall energy balance.

4. THE ATMOSPHERE

In the atmosphere, phenomena and events are loosely classified into the realms of "weather" and "climate." Climate is usually defined as average weather and thus is thought of as the prevailing weather, which includes not just average conditions but also the range of variations. Climate involves variations in which the atmosphere is influenced by and interacts with other parts of the climate system and the external forcings. The large fluctuations in the atmosphere from hour to hour or day to day constitute the weather but occur as part of much larger scale organized weather systems that arise mainly from atmospheric instabilities driven by heating patterns from the sun.

Much of the incoming solar radiation penetrates through the relatively transparent atmosphere to reach the surface, and so the atmosphere is mostly heated from below. The decreasing density of air with altitude also means that air expands as it rises, and consequently it cools. In the troposphere, this means that temperatures generally decrease with height, and warming from below causes convection that transports heat upwards. Convection gives rise to the clouds and thunderstorms, driven by solar heating at the earth's surface that produces buoyant thermals that rise, expand and cool, and produce rain and cloud.

Another example is the cyclones (low-pressure areas or systems) and anticyclones (high-pressure systems) and their associated cold and warm fronts, which arise from the equator-to-pole temperature differences and distribution of heating (Fig. 1). The atmosphere attempts to reduce these temperature gradients by producing these weather systems, which have, in the Northern Hemisphere, southerly winds to carry warm air polewards and cold northerly winds to reduce temperatures in lower latitudes. In the Southern Hemisphere, it is the southerlies that are cold and the northerly winds are warm. These weather systems migrate, develop, evolve, mature, and decay over periods of days to weeks and constitute a form of atmospheric turbulence.

Energy in the atmosphere comes in several forms. The incoming radiant energy is transformed when it is absorbed into sensible energy. On the surface of land, the heat may be manifested as increases in temperature or as increases in evaporation, and the partitioning depends on the available moisture and nature of vegetative ground cover. Increased evaporation adds moisture to the atmosphere, and this is often referred to as latent energy because it is realized as latent heating when the moisture is condensed in subsequent precipitation. Increases in temperature increase the internal energy of the atmosphere, which also causes it to expand. Consequently, it changes the altitude of the air and increases the potential energy. Therefore, there is a close relationship between internal and potential energy in the atmosphere, which are combined into the concept of enthalpy or sensible heat. Once air starts to move, usually because temperature gradients give rise to pressure gradients that in turn cause wind, some energy is converted into kinetic energy. The sum of the internal, potential, kinetic, and latent energies is a constant in the absence of a transfer of heat. However, the possibilities for conversions among all these forms of atmospheric energy are a key part of what provides the richness of atmospheric phenomena.

In terms of poleward transport of energy, the main transports at middle and higher latitudes occur in the form of sensible heat and potential energy, which are often combined as dry static energy, whereas at low to middle latitudes, latent energy also plays a major role. In the tropics, much of the movement of energy occurs through large-scale overturning of the atmosphere. The classic examples are the monsoon circulations and the Hadley circulation. At low levels, the moisture in the atmosphere is transported toward areas where it is forced to rise and hence cool, resulting in strong latent heating that drives the upward branch of the overturning cells and constitutes the monsoon rains. In the subtropics, the downward branch of the circulation suppresses clouds and is very dry as the "freeze-dried air" from aloft slowly descends, and the air warms as it descends and expands. Hence, in these regions the surface can radiate to space without much interference from clouds and water vapor greenhouse effects. This is what happens over the Sahara. However, another key part of the cooling in the subtropics that drives the monsoons is the link to midlatitude weather systems. The latter transport both sensible and latent heat polewards in the cyclones and anticyclones and hence cool the subtropics while warming the higher latitudes.

Thus, the atmosphere is primarily responsible for the energy transports (Fig. 1) that compensate for the

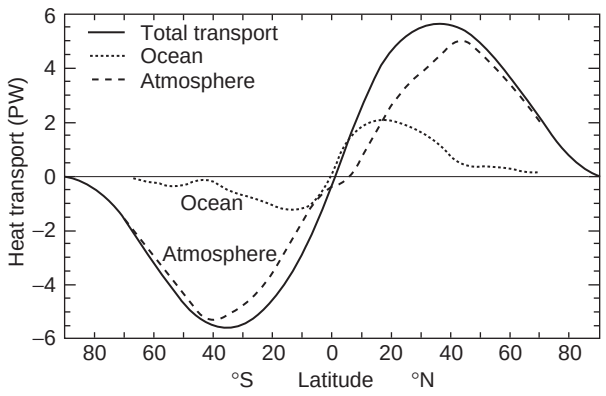


FIGURE 4 Poleward heat transports by the atmosphere and ocean and the total transport. From Trenberth and Caron (2001).

net radiation imbalances. Annual mean poleward transports of atmospheric energy of 5.0 PW peak at 43°N and with similar values near 40°S. Poleward ocean heat transports are dominant only between 0 and 17°N. At 35° latitude, where the peak total poleward transport in each hemisphere occurs, the atmospheric transport accounts for 78% of the total in the Northern Hemisphere and 92% in the Southern Hemisphere (Fig. 4).

5. THE HYDROLOGICAL CYCLE

Driven by exchanges of energy, the hydrological cycle involves the transfer of water from the oceans to the atmosphere, to the land, and back to the oceans both on top of and beneath the land surface. In the tropics in summer, land warms relative to the ocean and sets the stage for monsoon development. Water is evaporated from the ocean surface, cooling the ocean. As water vapor, it is transported perhaps thousands of kilometers before it is involved in clouds and weather systems and precipitated out as rain, snow, hail, or some other frozen pellet back to the earth's surface. During this process, it heats the atmosphere. Over land, soil moisture and surface waters can act through evaporative cooling to moderate temperatures. Some precipitation infiltrates or percolates into soils and some runs off into streams and rivers. Ponds and lakes or other surface water may evaporate moisture into the atmosphere and can also freeze so that water can become locked up for awhile. The surface water weathers rocks, erodes the landscape, and replenishes subterranean aquifers. Over land, plants transpire moisture into the atmosphere.

An estimate of the global hydrological cycle in units of $10^3\,km^3$ per year is that evaporation over the oceans (436) exceeds precipitation (399), leaving a net of 37 units of moisture transported onto land as water vapor. On average, this flow must be balanced by a return flow over and beneath the ground through river and stream flows and subsurface ground water flow. Consequently, precipitation over land exceeds evapotranspiration by this same amount (37). The average precipitation rate over the oceans exceeds that over land by 72% (allowing also for the differences in areas). It has been estimated that, on average, more than 80% of the moisture precipitated out comes from locations more than 1000 km distant, highlighting the important role of the winds in moving moisture.

6. THE OCEANS

The oceans cover 70.8% of the surface of the earth, although there is a much greater fraction in the Southern Hemisphere (80.9% of the area) than the Northern Hemisphere (60.7%), and through their fluid motions and high heat capacity they have a central role in shaping the earth's climate and its variability. The average depth of the ocean is 3795 m. The oceans are stratified opposite to the atmosphere, with warmest waters near the surface. Consequently, in the ocean, convection arises from cooling at the surface and transport of heat upwards occurs through colder and denser waters sinking and being replaced by lighter and more buoyant waters. Another vital factor in convection is the salinity of the water because this also affects density. Consequently, the densest waters are those that are cold and salty, and these are found at high latitudes where there is a deficit in radiation and atmospheric temperatures are low. The formation of sea ice also leads to a rejection of brine and increases the salinity of the surrounding waters. The cold, deep, abyssal ocean turns over only very slowly on timescales of hundreds to thousands of years.

The most important characteristic of the oceans is that they are wet. Water vapor, evaporated from the ocean surface, provides latent heat energy to the atmosphere when precipitated out. Wind blowing on the sea surface drives the large-scale ocean circulation in its upper layers. The ocean currents carry heat and salt along with fresh water around the globe (Fig. 4). The oceans therefore store heat, absorbed at the surface, for varying durations and release it in different places, thereby ameliorating temperature changes over nearby land and contributing substantially to variability of climate on many timescales.

Additionally, the ocean thermohaline circulation, which is the circulation driven by changes in seawater density arising from temperature (thermal) or salt (haline) effects, allows water from the surface to be carried into the deep ocean, where it is isolated from atmospheric influence and hence it may sequester heat for periods of 1000 years or more.

The main energy transports in the ocean are those of heat associated with overturning circulations as cold waters flow equatorwards at some depth while the return flow is warmer near the surface. There is a small poleward heat transport by gyre circulations whereby western boundary currents, such as the Gulf Stream in the north Atlantic or the Kuroshio in the north Pacific, move warm waters polewards while part of the return flow is of colder currents in the eastern oceans. However, a major part of the Gulf Stream return flow is at depth, and the most pronounced thermohaline circulation is in the Atlantic Ocean. In contrast, the Pacific Ocean is fresher and features shallower circulations. A key reason for the differences lies in salinity. Because there is a net atmospheric water vapor transport in the tropical easterly trade winds from the Atlantic to the Pacific across the central American isthmus, the north Atlantic is much saltier than the north Pacific.

7. THE LAND

The heat penetration into land is limited and slow, as it occurs mainly through conduction, except where water plays a role. Temperature profiles taken from bore holes into land or ice caps provide a coarse estimate of temperatures in years long past. Consequently, surface air temperature changes over land occur much faster and are much larger than those over the oceans for the same heating, and because we live on land, this directly affects human activities. The land surface encompasses an enormous variety of topographical features and soils, differing slopes (which influence runoff and radiation received), and water capacity. The highly heterogeneous vegetative cover is a mixture of natural and managed ecosystems that vary on very small spatial scales. Changes in soil moisture affect the disposition of heat at the surface and whether it results in increases in air temperature or increased evaporation of moisture. The latter is complicated by the presence of plants, which can act to pump moisture out of the root zone into the leaves, where it can be released into the atmosphere as the plants participate in photosynthesis; this process is called transpiration. The behavior of land ecosystems can be greatly influenced by changes in atmospheric composition and climate. The availability of surface water and the use of the sun's energy in photosynthesis and transpiration in plants influence the uptake of carbon dioxide from the atmosphere as plants transform the carbon and water into usable food. Changes in vegetation alter how much sunlight is reflected and how rough the surface is in creating drag on the winds, and the land surface and its ecosystems play an important role in the carbon cycle and fluxes of water vapor and other trace gases.

8. ICE

Major ice sheets, such as those over Antarctica and Greenland, have a large heat capacity but, like land, the penetration of heat occurs primarily through conduction so that the mass involved in changes from year to year is small. Temperature profiles can be taken directly from bore holes into ice and it is estimated that terrestrial heat flow is 51 mW/m^2. On century timescales, however, the ice sheet heat capacity becomes important. Unlike land, the ice can melt, which has major consequences through changes in sea level on longer timescales.

Sea ice is an active component of the climate system that is important because it has a high albedo. A warming that reduces sea ice, reduces the albedo and hence enhances the absorption of solar radiation, amplifying the original warming. This is known as the ice albedo feedback effect. Sea ice varies greatly in areal extent with the seasons, but only at higher latitudes. In the Arctic, where sea ice is confined by the surrounding continents, mean sea ice thickness is 3 or 4 m deep and multiyear ice can be present. Around Antarctica, the sea ice is unimpeded and spreads out extensively but, as a result, the mean thickness is typically 1 or 2 m.

9. THE ROLE OF HEAT STORAGE

The different components of the climate system contribute on different timescales to climate variations and change. The atmosphere and oceans are fluid systems and can move heat through convection and advection in which the heat is carried by the currents, whether small-scale short-lived eddies or large-scale atmospheric jet streams or ocean currents. Changes in phase of water, from ice to liquid to

water vapor, affect the storage of heat. However, even ignoring these complexities, many facets of the climate are determined simply by the heat capacity of the different components of the climate system. The total heat capacity considers the mass involved as well as its capacity for holding heat, as measured by the specific heat of each substance.

The atmosphere does not have much capability to store heat. The heat capacity of the global atmosphere corresponds to that of only 3.2 m of the ocean. However, the depth of ocean actively involved in climate is much greater. The specific heat of dry land is approximately a factor of $4\frac{1}{2}$ less than that of seawater (for moist land the factor is probably closer to 2). Moreover, heat penetration into land is limited and only approximately the top 2 m typically plays an active role (as a depth for most of the variations on annual timescales). Accordingly, land plays a much smaller role in the storage of heat and in providing a memory for the climate system. Similarly, the ice sheets and glaciers do not play a strong role, whereas sea ice is important where it forms.

The seasonal variations in heating penetrate into the ocean through a combination of radiation, convective overturning (in which cooled surface waters sink while warmer, more buoyant waters below rise), and mechanical stirring by winds through what is called the "mixed layer" and on average involve approximately 90 m of ocean. The thermal inertia of the 90-m layer would add a delay of approximately 6 years to the temperature response to an instantaneous change. This value corresponds to an exponential time constant in which there is a 63% response toward a new equilibrium value following an abrupt change. Thus, the actual change is gradual. The total ocean, however, with a mean depth of approximately 3800 m, if rapidly mixed would add a delay of 230 years to the response. Mixing is not a rapid process for most of the ocean, so in reality the response depends on the rate of ventilation of water between the mixed upper layers of the ocean and the deeper, more isolated layers through the thermocline (region of temperature gradient). Such mixing is not well-known and varies greatly geographically. An overall estimate of the delay in surface temperature response caused by the oceans is 10–100 years. The slowest response should be in high latitudes, where deep mixing and convection occur, and the fastest response is expected in the tropics. Consequently, the oceans are a great moderating effect on climate variations, especially changes such as those involved with the annual cycle of the seasons.

Generally, the observed variability of temperatures over land is a factor of two to six greater than that over the oceans. At high latitudes over land in winter, there is often a strong surface temperature inversion. In this situation, the temperature increases with altitude because of the cold land surface, and it makes for a very stable layer of air that can trap pollutants. The strength of an inversion is very sensitive to the amount of stirring in the atmosphere. Such wintertime inversions are greatly affected by human activities; for instance, an urban heat island effect exceeding 10°C has been observed during strong surface inversion conditions in Fairbanks, Alaska. Strong surface temperature inversions over midlatitude continents also occur in winter. In contrast, over the oceans, surface fluxes of heat into the atmosphere keep the air temperature within a narrow range. Thus, it is not surprising that over land, month-to-month persistence in surface temperature anomalies is greatest near bodies of water. Consequently, for a given heating perturbation, the response over land should be much greater than that over the oceans; the atmospheric winds are the reason why the observed factor is only in the two to six range.

Another example is the contrast between the Northern Hemisphere (NH) (60.7% water) and Southern Hemisphere (SH) (80.9% water) mean annual cycle of surface temperature. The amplitude of the 12-month cycle between 40 and 60° latitude ranges from <3°C in the SH to ~ 12°C in the NH. Similarly, in midlatitudes, the average lag in temperature response relative to the sun for the annual cycle is 33 days in the NH versus 44 days in the SH, again reflecting the difference in thermal inertia.

10. ATMOSPHERE–OCEAN INTERACTION: EL NIÑO

The climate system becomes more involved as the components interact. A striking example is a phenomenon that would not occur without interactions between the atmosphere and ocean, El Niño, which consists of a warming of the surface waters of the tropical Pacific Ocean. It takes place from the International Dateline to the west coast of South America and results in changes in the local and regional ecology. Historically, El Niños have occurred approximately every 3–7 years and alternated with the opposite phases of below average temperatures in the tropical Pacific, called La Niña. In the atmosphere, a pattern of change called the Southern

Oscillation is closely linked with these ocean changes so that scientists refer to the total phenomenon as ENSO. El Niño is the warm phase of ENSO and La Niña is the cold phase.

El Niño develops as a coupled ocean–atmosphere phenomenon, and the amount of warm water in the tropics is redistributed and depleted during El Niño and restored during La Niña. There is often a mini global warming following an El Niño as a consequence of heat from the ocean affecting the atmospheric circulation and changing temperatures throughout the world. Consequently, interannual variations occur in the energy balance of the combined atmosphere–ocean system and are manifested as important changes in weather regimes and climate throughout the world.

11. ANTHROPOGENIC CLIMATE CHANGE

11.1 Human Influences

Climate can vary for many reasons, and in particular, human activities can lead to changes in several ways. However, to place human influences in perspective, it is worthwhile to compare the output from a power plant with that from the sun. The largest power plants that exist are on the order of 1000 MW, and these service human needs for electricity in appliances and heating that use power in units of kilowatts. Figure 2 shows that the energy received at the top of the atmosphere is $342\,\mathrm{W\,m^{-2}}$ on an annual mean basis averaged over the globe. This is equivalent to 175 PW, although only approximately 120 PW is absorbed. One petawatt is 1 billion MW or 1 million huge power plants. Consequently, it is easily recognized that human generation of heat is only a tiny fraction of the energy available from the sun. This also highlights the fact that the main way in which human activities can compete with nature is if they somehow interfere with the natural flow of energy from the sun, through the climate system, and back out to space.

There have been major changes in land use during the past two centuries. Conversion of forest to cropland, in particular, has led to a higher albedo in places such as the eastern and central United States and changes in evapotranspiration, both of which have probably cooled the region in summer by perhaps 1°C and in autumn by more than 2°C, although global effects are less clear. In cities, the building of "concrete jungles" allows heat to be soaked up and stored during the day and released at night, moderating nighttime temperatures and contributing to an urban heat island. Space heating also contributes to this effect. Urbanization changes also affect runoff of water, leading to drier conditions unless compensated by water usage and irrigation, which can cool the area. However, these influences, although real changes in climate, are quite local. Widespread irrigation on farms can have more regional effects.

Combustion of fossil fuels not only generates heat but also generates particulate pollution (e.g., soot and smoke) as well as gaseous pollution that can become particulates (e.g., sulfur dioxide and nitrogen dioxide, which get oxidized to form tiny sulfate and nitrate particles). Other gases are also formed in burning, such as carbon dioxide; thus, the composition of the atmosphere is changing. Several other gases, notably methane, nitrous oxide, the chlorofluorocarbons (CFCs), and tropospheric ozone, are also observed to have increased due to human activities (especially from biomass burning, landfills, rice paddies, agriculture, animal husbandry, fossil fuel use, leaky fuel lines, and industry), and these are all greenhouse gases. However, the observed decreases in lower stratospheric ozone since the 1970s, caused principally by human-introduced CFCs and halons, contribute to a small cooling.

11.2 The Enhanced Greenhouse Effect

The amount of carbon dioxide in the atmosphere has increased by more than 30% in the past two centuries since the beginning of the industrial revolution, an increase that is known to be in part due to combustion of fossil fuels and the removal of forests. Most of this increase has occurred since World War II. Because carbon dioxide is recycled through the atmosphere many times before it is finally removed, it has a long lifetime exceeding 100 years. Thus, emissions lead to a buildup in concentrations in the atmosphere. In the absence of controls, it is projected that the rate of increase in carbon dioxide may accelerate and concentrations could double from pre-industrial values within approximately the next 60 years.

If the amount of carbon dioxide in the atmosphere were suddenly doubled, with other things remaining the same, the outgoing long-wave radiation would be reduced by approximately $4\,\mathrm{W\,m^{-2}}$ and instead trapped in the atmosphere, resulting in an enhanced greenhouse effect. To restore the radiative balance, the atmosphere must warm up and, in the absence of

other changes, the warming at the surface and throughout the troposphere would be approximately 1.2°C. In reality, many other factors will change, and various feedbacks come into play, so that the best estimate of the average global warming for doubled carbon dioxide is 2.5°C. In other words, the net effect of the feedbacks is positive and approximately doubles the response otherwise expected.

Increased heating therefore increases global mean temperatures and also enhances evaporation of surface moisture. It follows that naturally occurring droughts are likely to be exacerbated by enhanced drying. After the land is dry, all the solar radiation goes into increasing temperature, causing heat waves. Temperature increases signify that the water-holding capacity of the atmosphere increases and, with enhanced evaporation, the atmospheric moisture increases, as is currently observed in many areas. The presence of increased moisture in the atmosphere implies stronger moisture flow converging into all precipitating weather systems, whether they are thunderstorms, extratropical rain, or snow storms. This leads to the expectation of enhanced rainfall or snowfall events, which are also observed to be happening. Globally, there must be an increase in precipitation to balance the enhanced evaporation and hence there is an enhanced hydrological cycle.

The main positive feedback comes from water vapor. As the amount of water vapor in the atmosphere increases as the earth warms, and because water vapor is an important greenhouse gas, it amplifies the warming. However, increases in clouds may act either to amplify the warming through the greenhouse effect of clouds or reduce it by the increase in albedo; which effect dominates depends on the height and type of clouds and varies greatly with geographic location and time of year. Decreases in sea ice and snow cover, which have high albedo, decrease the radiation reflected back to space and thus produce warming, which may further decrease the sea ice and snow cover, known as ice albedo feedback. However, increased open water may lead to more evaporation and atmospheric water vapor, thereby increasing fog and low clouds, offsetting the change in surface albedo.

Other, more complicated feedbacks may involve the atmosphere and ocean. For example, cold waters off the western coasts of continents (such as California or Peru) encourage development of extensive low stratocumulus cloud decks that block the sun and this helps keep the ocean cold. A warming of the waters, such as during El Niño, eliminates the cloud deck and leads to further sea surface warming through solar radiation.

11.3 Effects of Aerosols

Aerosols occur in the atmosphere from natural causes; for instance, they are blown off the surface of deserts or dry regions. As a result of the eruption of Mt. Pinatubo in the Philippines in June 1991, considerable amounts of aerosol were added to the stratosphere that, for approximately 2 years, scattered solar radiation leading to a loss of radiation and a cooling at the surface.

As noted previously, human activities also affect the amount of aerosol in the atmosphere. The main direct effect of aerosols, such as sulfate particles, is the scattering of some solar radiation back to space, which tends to cool the earth's surface. Aerosols can also influence the radiation budget by directly absorbing solar radiation, leading to local heating of the atmosphere, and, to a lesser extent, by absorbing and emitting thermal radiation. A further influence of aerosols is that many of them act as nuclei on which cloud droplets condense. A changed concentration therefore affects the number and size of droplets in a cloud and hence alters the reflection and the absorption of solar radiation by the cloud.

Because man-made aerosols typically remain in the atmosphere for only a few days before they are washed out by precipitation, they tend to be concentrated near their sources, such as industrial regions, adding complexity to climate change because they can help mask, at least temporarily, any global warming arising from increased greenhouse gases.

12. OBSERVED AND PROJECTED TEMPERATURES

Changes in climate have occurred in the distant past as the distribution of continents and their landscapes have changed, as the so-called Milankovitch changes in the orbit of the earth and the earth's tilt relative to the ecliptic plane have varied the insolation received on earth, and as the composition of the atmosphere has changed, all through natural processes. Recent evidence obtained from ice cores drilled through the Greenland ice sheet indicates that changes in climate may often have been quite rapid and major and not associated with any known external forces.

Observations of surface temperature show a global mean warming of approximately 0.7°C during the past 100 years. The warming became noticeable from the 1920s to the 1940s, leveled off from the 1950s to the 1970s, and increased again in the late 1970s. The calendar year 1998 is the warmest on record. The 1990s are the warmest decade on record. Information from paleodata further indicates that these years are the warmest in at least the past 1000 years, which is as far back as a hemispheric estimate of temperatures can be made. The melting of glaciers throughout most of the world and rising sea levels confirm the reality of the global temperature increases.

Climate modeling suggests that solar variability has contributed to some of the warming of the 20th century, perhaps 0.2°C, up to approximately 1950. Changes in aerosols in the atmosphere, both from volcanic eruptions and from visible pollutants and their effects on clouds, have also contributed to reduced warming, perhaps by a couple of tenths of a degree. A temporary cooling in 1991 and 1992 followed the eruption of Mount Pinatubo in June 1991. Heavy industrialization and associated pollution following World War II may have contributed to the plateau in global temperatures from approximately 1950 to 1970. Interactions between the atmosphere and the oceans, including El Niño, have contributed to natural fluctuations of perhaps two-tenths of a degree. It is only after the late 1970s that global warming from increases in greenhouse gases has emerged as a clear signal in global temperatures.

Projections have been made of future global warming effects based on model results. Because the actions of humans are not predictable in any deterministic sense, future projections necessarily contain a "what if" emissions scenario. In addition, for a given scenario, the rate of temperature increase depends on the model and features such as how clouds are depicted so that there is a range of possible outcomes. Various projections in which the concentrations of carbon dioxide double 1990 values by the year 2100 indicate global mean temperature increases ranging from 2 to 4°C higher than 1990 values. Uncertainties in the projections of carbon dioxide and aerosols in the atmosphere add to this range. However, a major concern is that the rates of climate change as projected exceed anything seen in nature in the past 10,000 years.

Acknowledgments

The National Center for Atmospheric Research is sponsored by the National Science Foundation.

SEE ALSO THE FOLLOWING ARTICLES

Climate Change and Energy, Overview • Climate Change: Impact on the Demand for Energy • Climate Protection and Energy Policy • Ecosystems and Energy: History and Overview • Environmental Gradients and Energy • Greenhouse Gas Emissions from Energy Systems, Comparison and Overview • Lithosphere, Energy Flows in • Ocean, Energy Flows in

Further Reading

Bond, G., *et al.* (1997). A pervasive millennial-scale cycle in North Atlantic Holocene and glacial climates. *Science* **278,** 1257–1266.

Dahl-Jensen, D., Mosegaard, K., Gundestrup, N., Clow, G. D., Johnsen, S. J., Hansen, A. W., and Balling, N. (1998). Past temperatures directly from the Greenland Ice Sheet. *Science* **282,** 268–271.

Hansen, J., *et al.* (1996). A Pinatubo climate modeling investigation. *In* "The Mount Pinatubo Eruption: Effects on the Atmosphere and Climate" (G. Fiocco, D. Fua, and G. Visconti, Eds.), NATO ASI Series I, No. 42, pp. 233–272. Springer-Verlag, Heidelberg.

Intergovernmental Panel on Climate Change (2001). *In* "Climate Change 2001. The Scientific Basis" (J. T. Houghton *et al.*, Eds.). Cambridge Univ. Press, Cambridge, UK.

Kiehl, J. T., and Trenberth, K. E. (1997). Earth's annual global mean energy budget. *Bull. Am. Meteorol. Soc.* **78,** 197–208.

Levitus, S., Antonov, J. I., Wang, J., Delworth, T. L., Dixon, K. W., and Broccoli, A. J. (2001). Anthropogenic warming of the earth's climate system. *Science* **292,** 267–270.

Pollack, H. N., Huang, S., and Shen, P. -Y. (1998). Climate change record in subsurface temperatures: A global perspective. *Science* **282,** 279–281.

Trenberth, K. E. (1983). What are the seasons? *Bull. Am. Meteor. Soc.* **64,** 1276–1282.

Trenberth, K. E. (1998). Atmospheric moisture residence times and cycling: Implications for rainfall rates with climate change. *Climate Change* **39,** 667–694.

Trenberth, K. E. (2001). Stronger evidence for human influences on climate: The 2001 IPCC Assessment. *Environment* **43**(4), 8–19.

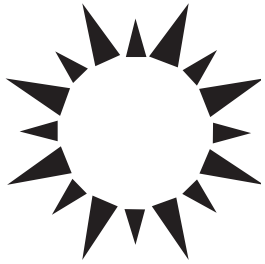
Trenberth, K. E., and Caron, J. M. (2001). Estimates of meridional atmosphere and ocean heat transports. *J. Climate* **14,** 3433–3443.

Trenberth, K. E., and Hoar, T. J. (1997). El Niño and climate change. *Geophys. Res. Lett.* **24,** 3057–3060.

Trenberth, K. E., and Stepaniak, D. P. (2003). Seamless poleward atmospheric energy transports and implications for the Hadley circulation. *J. Climate* **16,** 3705–3721.

Trenberth, K. E., Houghton, J. T. and Meira Filho, L. G. (1996). The climate system: An overview. In "Climate Change 1995. The Science of Climate Change" (J. T. Houghton, L. G. Meira Filho, B. Callander, N. Harris, A. Kattenberg, and K. Maskell, Eds.), Contribution of WG 1 to the Second Assessment Report of the Intergovernmental Panel on Climate Change, pp. 51–64. Cambridge Univ. Press, Cambridge, UK.

Easter Island: Resource Depletion and Collapse



JAMES A. BRANDER
University of British Columbia
Vancouver, British Columbia, Canada

1. **The Mystery of Easter Island**
2. **First Discovery and Early History**
3. **Resource Depletion and Collapse**
4. **Malthusian Population Dynamics**
5. **A Predator–Prey System with Unfortunate Parameter Values**
6. **The Role of Property Rights and Open Access**
7. **Aftermath**
8. **Lessons of Easter Island**

Glossary

core sample A specimen obtained in a long cylinder, taken from a physical site for subsequent study; used in many fields, including oil exploration and archaeology. In archaeology, elements in the core sample can be carbon dated, allowing a reconstruction of the implicit historical record.

Malthusian population dynamics A demographic theory originated by Thomas Malthus, suggesting that the human population has a natural tendency to grow sufficiently rapidly to offset and perhaps even overshoot productivity gains in food production. Malthus also made the related assertion that population growth will therefore be sensitive to per capita consumption—tending to increase when consumption rises and to decrease when consumption falls. Hence, the term refers to a situation in which population growth is positively related to per capita consumption. It is noteworthy that the term does not describe societies that have gone through the so-called demographic transition in which fertility tends to fall at sufficiently high income levels, it does seem to describe societies at the level of development of Easter Island prior to European contact.

market failure A situation in which markets fail to achieve economic efficiency. Open-access resources and, more broadly, difficulty in establishing property rights constitute an important source of market failure.

open-access resource A resource to which any potential user can gain access. Various communities have sometimes had land to which everyone had access, e.g., a "common," consisting of grazing land. It was commonly observed that such common land was often overgrazed. The tendency to overuse open-access resources is sometimes referred to as the "open-access" problem, which is an extreme version of "incomplete property rights," arising when the owner or controller of an economic resource can exclude other users only at significant cost.

predator–prey system A system in which a "prey" forms an important part of the consumption for a "predator." It has often been observed that predators and prey (such as lynx and rabbits) sometimes go through dramatic population cycles. In the early 20th century, Lotka and Volterra independently studied the formal properties of simple linear two-equation dynamic systems in an effort to shed light on predator–prey systems. They discovered that population cycles are a natural outcome of such systems for particular parameter values. Extinction and monotonic convergence to a steady state are also possible dynamic trajectories in predator–prey models.

Easter island (or rapa nui) is remarkable for being a small geographic entity that has generated enormous academic and popular interest. This interest arose primarily from the enormous stone statues, or moai, positioned at various locations around the island. These moai are sometimes referred to as "the mystery of Easter Island." The mystery of the moai remained unsolved until the last decade of the 20th century, when it became clear that the statues were intimately connected to a story of resource depletion and collapse.

 © 2004 Elsevier Inc. All rights reserved.

1. THE MYSTERY OF EASTER ISLAND

1.1 Early European Contact

First European contact with Easter Island occurred on Easter Day of 1722, when three Dutch ships under the command of Jacob Rogaveen stopped for a 1-day visit at the very isolated South Pacific island, just over 3000 km west of Chile. The visitors observed a small, treeless island populated by what they thought were about 3000 islanders and, to their surprise, by a large number of giant statues. The outside world had no previous knowledge of Easter Island and the islanders apparently had no awareness of the outside world. The visit must have been a shock to the Easter Islanders, but no systematic account of their reaction survives and no further contact with Europeans occurred until 1770, when a Spanish vessel made a brief stop.

The first systematic study of Easter Island dates from the 1774 visit of James Cook. At least one member of Cook's crew could understand the Polynesian dialect spoken by the Easter Islanders. This allowed Cook to infer that the Easter Island population was in fact Polynesian. In addition, Cook learned that the islanders believed that the statues had been built and moved to their then-current locations by some supernatural force. Certainly the islanders of 1774 had no concept of how to carve or move the statues. Cook estimated that the island's population was about 2000, which would have been insufficient to move the statues using any technology available on the island.

1.2 Explanations of the Moai

The mystery of how the statues came into existence remained a topic of increasingly fanciful speculation from Cook's day forward. Among the more striking theories advanced to explain the mystery include Thor Heyerdahl's view that the statues had been built by an advanced South American civilization that had discovered and colonized Easter Island before somehow being displaced by a less advanced Polynesian society. In an effort to support his theory, Heyerdahl undertook a famous voyage to Easter Island, in a reed boat named the Kon Tiki, starting from off the coast of Chile. An alternative and even more exotic theory was that Easter Island was the remnant of the hypothesized archipelago civilization of "Mu" that, similar in concept to Atlantis, had sunk beneath the ocean due to geological events. Perhaps the most remarkable theory was that of Eric von Daniken, who wrote a best-selling book (*Chariots of the Gods*) and produced a popular 1970s television series, which proposed that the statues had been built by extraterrestrials.

These imaginative explanations have been displaced by scientific evidence, including carbon dating of core samples taken from Easter Island and a genetic study of skeletal remains on Easter Island. This evidence has allowed the development of a rather different understanding of Easter Island's past. This understanding, based on resource depletion and collapse, explains both the statues and the decline of the Easter Island civilization.

2. FIRST DISCOVERY AND EARLY HISTORY

2.1 The Virgin Forest

Although there is some uncertainty about dates, it now seems that Easter Island was first discovered by a small group of Polynesians sometime between 400 and 700 AD. The striking surprise that emerged from analysis of core samples is that Easter Island was at this time covered by a dense forest of *Jubaea chilensis* (the Chilean wine palm). This particular palm is a large, slow-growing tree that grows in temperate climates.

2.2 The Easter Island Economy

Following first discovery of Easter Island, the new inhabitants developed an economy based on the wine palm. Harvested trees were the raw material for canoes that could be used for fishing, leading to an early diet that depended very heavily on fish, gradually supplemented from other sources, including yams grown on deforested sections of the island, and domestic fowl. The population grew steadily and the civilization grew more sophisticated, allowing a substantial diversion of labor into ceremonial and other nonsubsistence activity. However, as time progressed, the forest stock gradually diminished. At first, liberation of forested land for agricultural uses, for dwelling space, and for other purposes would have been seen as a benefit. The rate of forest decline was slow. In a typical lifetime (somewhere between 30 and 40 years for those who survived early childhood), it would have been hard to notice a dramatic change in the forest cover because the stock would not have declined by more than about 5 or 6% over a 30- to 40-year period, until relatively near the end of viable forest.

3. RESOURCE DEPLETION AND COLLAPSE

3.1 An End to the Golden Age

By the turn of the millennium in 1000 AD, the population was still rising and a statue-carving culture had developed. The statues were carved in Easter Island's lone quarry and then transported to desired locations using logs as rollers. The forest stock would have been down to perhaps two-thirds of its original level and the islanders probably would have regarded as apocryphal any claims that the island was once fully forested. However, this much loss of forest cover would have begun to reduce rainfall (because low clouds could pass overhead more easily) and would also have reduced the capacity of the soil to retain water. These trends continued for the next several hundred years. The peak population (of probably between 10,000 and 14,000), the greatest intensity of statue carving, and the virtual elimination of the forest stock all occurred in the neighborhood of 1400 AD.

3.2 Collapse of a Civilization

The subsequent 100 years on Easter Island must have been very difficult. Statue carving declined and soon ceased entirely. A new weapon, almost certainly used for hand-to-hand combat, enters the archaeological record at this time. Chickens were moved into fortified enclosures, much of the population moved into caves, and there is evidence of cannibalism from this period. The population began to decline steadily. The population was apparently still decreasing in the 18th century, because the 1774 population estimate by Cook was significantly less than Rogaveen's estimate of 50 years earlier. In addition, Cook noted that many of the statues had been pushed over on their faces, but no such phenomenon was observed by Rogaveen. It seems likely that the middle of the 18th century was a turbulent time on Easter Island. As has been frequently observed, the Easter Island civilization temporarily surpassed its limits and crashed devastatingly.

4. MALTHUSIAN POPULATION DYNAMICS

4.1 A Cautionary Tale?

The basic facts regarding Easter Island are no longer in serious dispute. It is, however, important to understand why Easter Island suffered an internally generated collapse based on resource overuse. The real question is whether Easter Island is an unusual and isolated case or whether it is a cautionary tale for the modern world. It is therefore important to uncover any general principles that might be inferred from the Easter Island experience.

4.2 Malthusian Population Dynamics and Overshooting

One important idea that might shed light on Easter Island derives from Malthusian population dynamics. Polynesian societies (along with virtually all other societies at comparable stages of development) were Malthusian in the sense that fertility was positively related to consumption (or real income) and mortality was negatively related to consumption. Correspondingly, in times of increasing hardship, fertility would decrease and mortality would increase. Depending on the precise responsiveness of fertility and mortality to real income (or food supply), and on other aspects of the environment, it is possible for Malthusian populations to "overshoot." Such populations would tend to increase too rapidly to be sustained, leading to an ultimate population "crash" enforced by some of the more unpleasant instruments of mortality, such as disease, starvation, and violent conflict. This possible feature of Malthusian models led subsequent readers of economist Thomas Malthus to refer to economics as the "dismal science."

4.3 Mitigating Factors and Reinforcing Factors

Various factors can mitigate Malthusian population dynamics (as Malthus recognized), including technological progress and the possibility of a "demographic transition" under which fertility declines at sufficiently high incomes. (The combination of technological progress and a demographic transition has, of course, allowed most of the modern world to avoid the Malthusian trap over the two centuries that have passed since the time of Malthus.) On the other hand, as Malthus also realized, there are factors that tend to exacerbate Malthusian dynamics, including reliance on open-access resources and lack of property rights more generally.

4.4 Easter Island and Malthusian Overshooting

At a superficial level, Easter Island might look like a simple example of Malthusian population

overshooting. The mitigating factors were not present, in that little technological progress occurred and the level of income was never high enough for a demographic transition to be relevant. In addition, the Easter Island economy did suffer from heavy reliance on a renewable resource that very likely was characterized by open-access problems or at least by incomplete property rights. However, Malthusian overshooting cannot be the whole story, for most Polynesian societies did not experience the boom-and-bust pattern exhibited by Easter Island, even though these societies were similar to the Easter Island culture in most important respects. Thus, for example, islands such as Tahiti and New Zealand's North Island appeared to be in something approaching a long-run steady state at the time of first European contact. Therefore, there must be more to the story than simple Malthusian overshooting.

5. A PREDATOR–PREY SYSTEM WITH UNFORTUNATE PARAMETER VALUES

5.1 Easter Island as a Predator–Prey System

As reported in 1998 by Brander and Taylor, the understanding of Easter Island can be greatly advanced by using formal mathematical analysis. The formal description provided here relies on that analysis. One key point is that the Easter Island economy can be viewed as a "predator–prey" system. The forest resource can be considered as the prey and the Easter Islanders as the predators. One important aspect of even relatively simple predator–prey systems is that small to moderate changes in one or more key parameters can give rise to major qualitative changes in the dynamic pattern of population growth and decline. In particular, depending on parameter values, it is possible for a given model structure to give rise to smooth, or "monotonic," movement toward a steady state or, alternatively, to a boom-and-bust cycle. One possible behavior is long-run stability of the type apparently observed in Tahiti. Another possible behavior is the rise and fall that characterizes Easter Island. More broadly, it is quite possible that the same general model might describe a variety of Polynesian societies, with plausible variations in parameter values giving rise to the observed differences in demographic and economic experience.

5.2 Modeling the Renewable Resource Stock

The economy and ecology of Easter Island were relatively simple, but obtaining a tractable mathematical representation still requires significant abstraction and approximation. At a minimum, it is necessary to model the behavior of two "stocks" or "state variables": the forest stock and the population. In fact, it is an oversimplification to focus just on the forest stock. Even in the absence of the forest stock, grass, shrubs, and certain vegetables could be grown and this biomass was sufficient to support a diminished but significant population. Rather than introduce a third "stock" representing other biomass, it is simpler to aggregate the renewable resource base growing on the land into a single resource stock, denoted as S. For simplicity, sometimes just the term "forest" is used, but it should be kept in mind that other components of the flora biomass are also relevant.

5.3 Logistic Growth

The starting point for describing the evolution of a renewable resource stock is the logistic growth function. Using t to denote time, a simple logistic growth function has the form $G(t) = rS(1-S/K)$. The variable r is the intrinsic growth rate and K is the environmental carrying capacity, or maximum possible size of the resource stock. $G(t)$ is the growth rate defined in biomass units and G/S is the proportional growth rate (i.e., a number such as 0.1 or 10%). If the forest stock S reaches a level equal to carrying capacity K, then $G = 0$ and no further growth occurs. Similarly, if there is no forest stock, then $S = 0$ and no growth occurs. If the forest stock is small relative to carrying capacity (but positive), then S/K is negligible and $G = rS$. In this case, proportional growth G/S approaches r. Thus, we can think of r as the proportional growth rate in the absence of congestion effects. As the forest stock increases, there is congestion (competition for space, rainfall, sunlight, etc.), with the result that the proportional growth rate tends to decline. However, the base to which that growth rate is applied increases as S increases. Therefore, the absolute growth rate in biomass units increases with S up to a critical point beyond which the congestion effect dominates the increasing base effect and the absolute growth rate declines. A logistic growth function is illustrated in Fig. 1. It is clear from Fig. 1 that the maximum sustainable yield occurs at the maximum

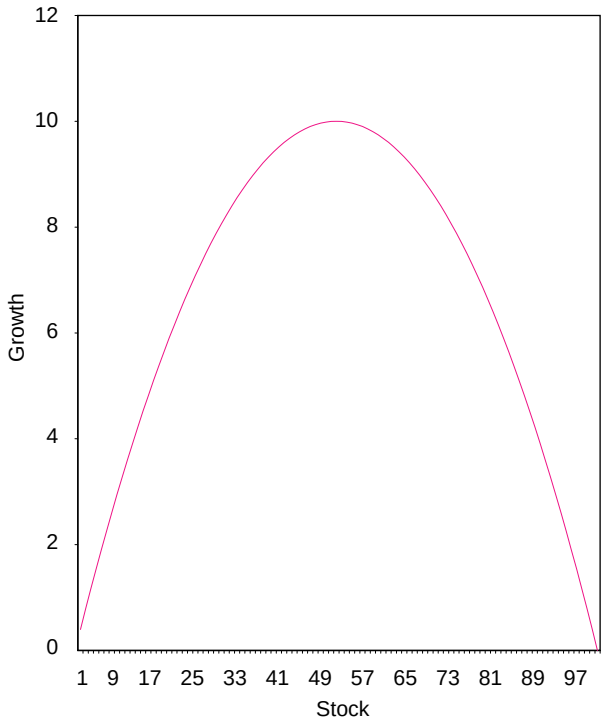


FIGURE 1 Logistic growth.

growth rate. In other words, it would be possible to maintain a stock of $K/2$ and harvest at this maximum growth rate in steady state.

5.4 Other Growth Functions

The logistic growth function is no doubt far from being a perfect approximation to the growth of the *Jubaea chilensis* forest on Easter Island. Various extensions to the logistic growth function are possible. For example, it is quite possible that at very low values of the stock, the proportional growth rate actually tends to decline. This is referred to as "depensation" or the "Allee effect." The logistic growth function can be augmented to allow for depensation, but that will not be pursued here. The basic logistic growth function does allow for the critical elements of biological renewable resource growth that are needed to capture at least the qualitative experience of Easter Island.

5.5 Open Access, Property Rights, and Harvesting Behavior

The natural growth of the forest stock provides part of the story in characterizing the evolution of the forest stock. The other part of the story depends on the harvesting behavior of the Easter Islanders. The key economic point is that under conditions of incomplete property rights, it is natural to expect that, for any given stock, the amount harvested at any given time period, $H(t)$, is increasing with population size. In other words, for any given set of external conditions and any given forest stock, we expect more harvesting to occur if there is a larger population seeking to use the forest stock. Similarly, for a given population, harvesting is easier if the stock is larger, and therefore more harvesting would be likely to occur. In the case of pure open access (no enforceable property rights) and standard demand and production functions, the harvesting function can be shown to take on the form $H = aSP$, where a is a parameter reflecting demand conditions and the harvesting technology and P is the population. This functional form is used here for concreteness, although other related functional forms would have similar properties. The differential equation that governs the evolution of the forest stock can therefore be written as

$$dS/dt = G(t) - H(t) = rS(1 - S/K) - aSP, \quad (1)$$

where state variables S and P are functions to time t.

5.6 Population Growth

To describe population growth, we adopt the linear form used by Lotka and Volterra in their original predator–prey formulations. The first component is a base net fertility rate, n, that would apply in the absence of a resource stock (and would presumably be negative in our environment). In addition, there is a second term that, from our point of view, captures the Malthusian idea that net fertility rises when per capita consumption rises. This implies that proportional net fertility should rise when, other things equal, the resource stock is larger. The effect follows because a higher stock (for a given population) would normally give rise to higher per capita consumption. This reasoning leads to the following population growth function:

$$dP/dt = nP + bPS, \quad (2)$$

where b is a parameter that reflects economic conditions. It can be seen from Eq. (2) that proportional population growth rate $(dP/dt)/P$ equals the base rate plus some increment related to the per capita resource stock. The functional form could be changed to include a "congestion effect," similar to the congestion effect that arises in the logistic growth function that is used for the resource stock, but

Eq. (2) should allow a reasonable approximation to actual events.

5.7 A Dynamic System

Equations (1) and (2) form an interdependent dynamic system similar to the Lotka–Volterra predator–prey model. Equation (1) shows the evolution of prey (the forest stock) and has the key property that the prey tends to diminish more quickly when there are more predators (humans). Equation (2) shows the evolution of the predator and has the property that the predator increases more rapidly when prey are more numerous. Figure 2 illustrates the relationship between the principal "stocks" and "flows" in the model using a formal flow diagram.

It is possible to examine the properties of this model analytically. Solutions for steady states can be determined and the dynamic path leading from any initial conditions can be characterized. Depending on parameter values, three steady states are possible, one with no humans and the forest at environmental carrying capacity, one with no humans and no resource stock, and one "interior" steady state with positive stocks of humans and forest biomass. The dynamic evolution of the model can be either monotonic or cyclical, once again depending on parameter values.

5.8 Simulation of Easter Island

Perhaps the most revealing method of analysis of this dynamic system is through simulation. It is possible to estimate plausible values for the parameters in Eqs. (1) and (2) and simulate the evolution of Easter Island. The necessary parameters are a, b, n, r, and environmental carrying capacity K. It is also necessary to specify the size $P(0)$ of the initial founding population.

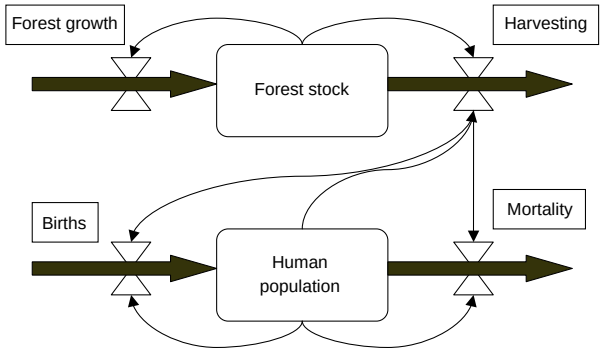


FIGURE 2 A flow diagram of stocks and flows.

The environmental carrying capacity and the initial forest stock are taken to be the same. Of course, the size of this stock can be normalized to any value. For scaling purposes, it is convenient if the resource stock is normalized to be of approximately the same magnitude as the maximum population, so we assume the initial resource stock (and environmental carrying capacity) is 12,000 biomass units. The initial population probably arrived in one or a few large ocean-going canoes that had traveled a considerable distance in search of new land. Take the founding population to be 40, although it makes very little difference if the founding population were taken to be twice or three times as large or, for that matter, if it were smaller. The net fertility rate, n, is the proportional (or percentage) rate of growth in the absence of the resource stock. In the simulation, discrete periods of 10 years will be used. Using a value of -0.1 for n indicates that, without the biomass in place, the population would tend to decline at a rate of 10% per decade. Parameters a and b reflect particular underlying economic conditions related to harvesting technology and to the preferences of the Easter Islanders. Because Eqs. (1) and (2) are being used as the starting points for the formal analysis here, parameters a and b could be viewed as essentially arbitrary parameters that we choose appropriately to "fit" the model to the experience of Easter Island. However, Eqs. (1) and (2) can be derived from more fundamental economic structures that suggest plausible magnitudes for a and b. A plausible magnitude for a is 4×10^{-5} and a plausible magnitude for b is 1.6×10^{-4}.

The key parameter of interest is the intrinsic growth rate, r. As it happens, *Jubaea chilensis* is a well-known species. It prefers temperate climate and grows slowly, normally requiring 40 to 60 years from first planting to the time it bears its first fruit. An intrinsic growth rate on the order of 4% per decade would be plausible. The Easter Island flora biomass consisted of more than just *J. chilensis*, but the forest was by far the most important part of the resource stock, so 4% (or 0.04) is used in the base case simulation for Easter Island.

To obtain simulations for Easter Island, Eqs. (1) and (2) are converted to discrete format. Thus, the two simulated equations are as follows:

$$S_{t+1} = S_t + rS_t(1 - St/K) - aS_tP_t, \qquad (1A)$$

$$P_{t+1} = nP_t + bP_tS_t. \qquad (2A)$$

The units of time are decades and the model is run for about 1200 years. The results are shown in Fig. 3.

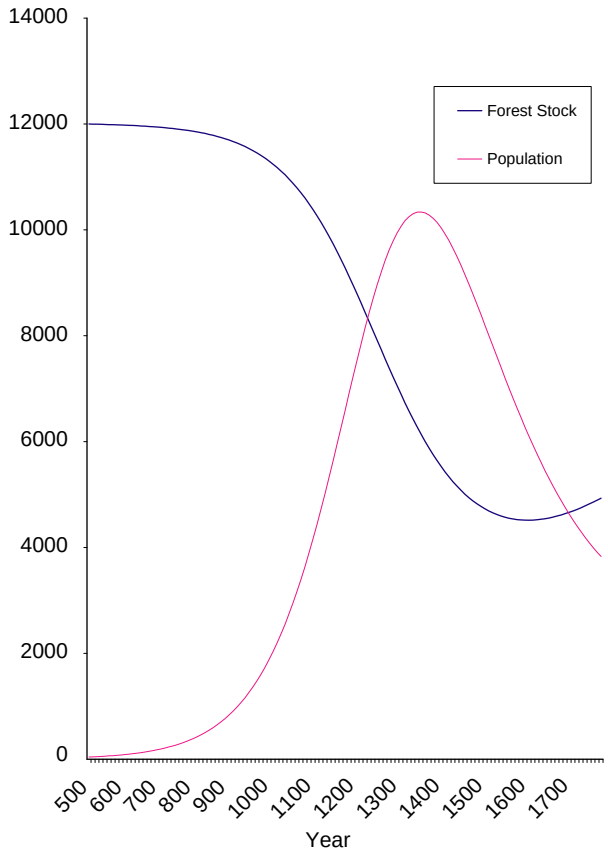


FIGURE 3 Easter Island simulation.

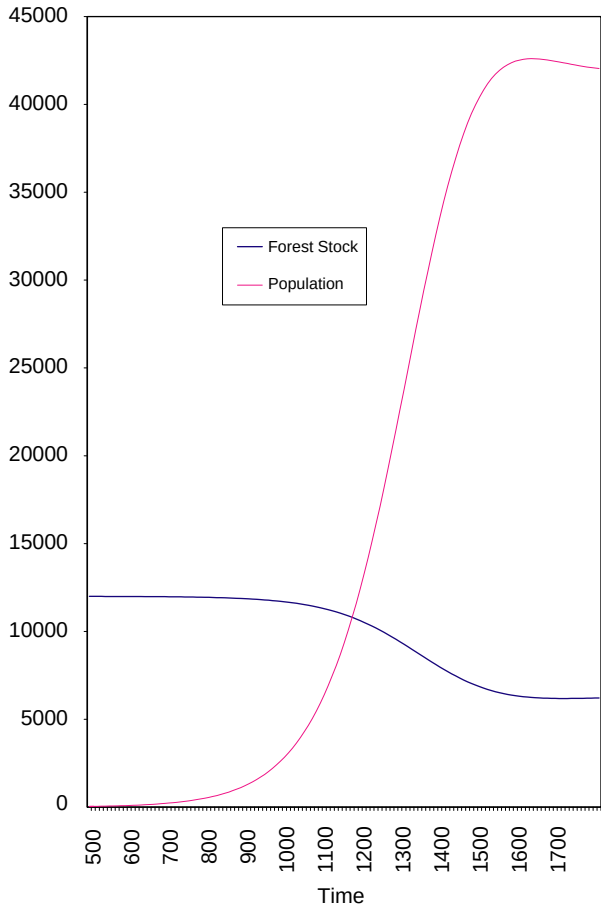


FIGURE 4 A fast-growing resource.

As can be seen in Fig. 3, this relatively simply model tracks the known facts about Easter Island quite well. If it is assumed that first colonization occurred in 500 AD, then peak population occurs in the 14th century at just over 10,000. From there, the population decreases and is down to a little over 4000 by the time of first European contact in 1722—a little high relative to actual estimates, but quite close. The resource stock is down to less than half its initial level by 1400 AD. On the actual Easter Island, this was about the time the wine palm forest completely disappeared, so for this to be accurate it would need to be the case that the flora biomass in the absence of the palm forest was about half of the biomass with the forest. This is certainly high, because the original palm forest had far more than twice the biomass of whatever shrubs, grass, and other vegetation was left after the forest was gone. Still, the general qualitative pattern shown in Fig. 3 is a reasonable representation of the Easter Island experience.

A particularly interesting variation of this simulation arises from making one change in a model parameter. Most of Polynesia did have palm forests. However, the palms growing on other islands grew much faster than the Chilean wine palm. The two most common large palms are the *Cocos* (coconut palm) and the *Pritchardia* (Fijian fan palm). These palms reach the fruit-growing stage after 7 to 10 years (much less than the 40- to 60-year period for the wine palm) and would have an intrinsic growth rate on the order of 35% per decade (rather than 4%). The resulting simulation is shown in Fig. 4, which illustrates a dynamic pattern strikingly different from that of Fig. 3. No boom-and-bust pattern is discernible. In fact, there is a very slight cycle that would not disappear entirely unless the intrinsic growth rate exceeded 71% per decade, but, even at 35%, the very gentle cycle would be much too damped to be evident to archaeologists and the approach to steady state is virtually monotonic. The forest stock declines smoothly to a new steady state that would correspond to having some standing forest in steady state. Low intrinsic growth rates, on the other hand, produce sharp cyclical fluctuations.

Therefore, in the predator–prey model described here, an island with a slow-growing resource base will exhibit overshooting and collapse. An otherwise identical island with a more rapidly growing resource base will exhibit a near-monotonic adjustment of population and resource stocks toward steady-state values. This fact alone can explain the sharp difference in experience between Easter Island and other Polynesian islands.

There are 12 so-called mystery islands in Polynesia that had been settled by Polynesians but were unoccupied at the time of first European contact. All these islands but one had relatively small carrying capacities. Applying our model, we see that if K (carrying capacity) is too small, then there is no interior steady state, implying that a colonizing population might arrive and expand but would eventually be driven to extinction.

6. THE ROLE OF PROPERTY RIGHTS AND OPEN ACCESS

A presumption underlying the analysis here is that property rights were incomplete, perhaps not to the point of full open access, but at least to the point where it was hard to prevent increased harvesting as the population increased. Specifically, in Eq. (1) harvesting is assumed to be increasing with population size for any given stock. This need not be the case if property rights are complete, because an efficient owner or manager of the resource would restrict access beyond some optimal point.

It is unlikely that Easter Island was characterized by complete open access, but it is very likely that property rights were incomplete in the sense that different competing harvesters probably had access to some common part of the forest. Incomplete property rights give rise to market failure, with the result that overharvesting occurs. In other words, the individual incentives faced by individual harvesters (or by small kinship groups) would lead to overharvesting from the collective point of view. As shown in Fig. 4, even with incomplete property rights, it is still possible to have monotonic adjustment to a steady state if parameter values are suitable. However, it would be even better if the society could enforce strict property rights. If a dictator or "social planner" could impose strict property rights, that person could, for example, set harvesting at the level of maximum sustainable yield and keep harvesting at that level independent of population size. However, by all accounts, that level of control was very difficult to impose and maintain in Polynesian and in most other societies.

In addition to the property rights problem, there was also a significant intertemporal market failure problem. Specifically, no one who planted or nurtured a young wine palm could expect to make personal use of that tree in his or lifetime, because life expectancy was short relative to the time taken for trees to reach maturity. Even the planter's children would be unlikely to benefit from the such activity. It would be the generation of one's grandchildren who would benefit from current planting activity. Thus, the incentives to undertake investments in planting and growing were weak. In modern societies, someone who plants a stand of trees can sell the stand at any time to someone else willing to hold the tree lot for some period of time—secure in the knowledge that it can be sold to still another person in the future. This mechanism of overlapping ownership connected through markets allows, in effect, future generations to "pay" for current investments. This is necessary for economic efficiency. However, in the Easter Island context, there would be no such markets. Thus, there would be a "market failure" leading to under investment in planting and growing trees. In primitive societies, social norms often take the place of market transactions in enforcing particular economic investments, but social norms are highly unreliable. In the Easter Island case, it is quite possible that the statue-carving norm dominated any conservation norms and worsened rather than mitigated the overharvesting problem.

7. AFTERMATH

It is useful to consider the history of Easter Island after the 18th century. The story is a very sad one—much worse than the endogenous boom-and-bust cycle that had already occurred on Easter Island. From the time of James Cook in the late 18th century through the mid-19th century, conditions on Easter Island gradually improved and the population increased gradually. Population was reliably estimated at something exceeding 3000 in 1862. In 1862 and 1863, slave traders from the Spanish community in Peru invaded the island and took about one-third of the islanders as slaves. Most of these slaves died of smallpox and other infectious diseases within a few years. A few returned to Easter Island, inadvertently causing a smallpox epidemic that killed most of the remaining Islanders.

By 1877, the population reached its low point of 111. Some of the deaths would have been due to causes other than smallpox, but, even allowing for these other causes, these numbers imply that the death rate from smallpox exceeded 90% of the base population—probably the highest death rate ever observed from an infectious disease invading a population of significant size. Although this has nothing to do with the pattern of resource overuse and collapse, it illustrates an important force in human history: the devastating effect that a new infectious disease can have on a long-isolated population.

From 1877, the population gradually increased through natural increase, immigration from other Polynesian islands (especially Tahiti), and immigration from the South American mainland. In 1888, the island was annexed by Chile, which maintains possession of the island at present. Population on the island has risen to something approaching 3000 and is devoted primarily to agriculture and to tourism. There is some mild tension between the predominantly Polynesian traditional or indigenous peoples and the South American population in Easter Island, but, on the whole, Easter Island is a peaceful and pleasant island that has become a major tourist destination of particular interest to "cultural tourists" who wish to see the now carefully maintained and protected moai.

8. LESSONS OF EASTER ISLAND

Some observers like to see Easter Island as a metaphor for the modern world. However, it is important to be aware of the differences. The first point, as has been emphasized already, is that societies of the Easter Island type do not necessarily have a dramatic decline. In the case of Easter Island, the civilization was "unlucky" to be on an island where the main resource was very slow growing. If the resource had been faster growing, then the boom-and-bust cycle probably would have been avoided, as it was in most of Polynesia. Therefore, even if Earth as a whole could be viewed as Easter Island writ large, there would no presumption about any inevitability of the imminent decline and fall of modern civilization.

On the other hand, Easter Island should not be viewed as an isolated case. As modern archaeology has made increasing use of sophisticated scientific methods, it has become increasingly clear that resource degradation has played an important role in the rise and fall of many civilizations. The truly unique feature of Easter Island was its isolation, which prevented migration as a response to resource degradation. In other parts of the world, overshooting population and resource degradation have led to substantial out-migration that mitigated, but did not eliminate, the population losses and the cultural losses associated with the boom-and-bust cycle. The lesson here is that resource management is very difficult, especially in conditions approaching open access, and overshooting is a genuine concern. This applies to the modern world just as it did to Easter Island. Thus, for example, major fisheries, major wildlife resources, and major forest resources around the world have been significantly compromised already. It is quite possible that various populations around the world, especially in areas with high population growth, such as Africa and parts of the Middle East, are on an overshooting trajectory.

Another lesson to be drawn from Easter Island concerns the danger of making simple linear projections based on a short time series. If an Easter Islander had, in the year 1300, extrapolated trends in population and real income based on the previous few centuries, that person would have failed to anticipate the "turning point" in population and real income that was coming. It is a characteristic of resource systems and predator–prey systems more broadly that cyclical patterns are common.

Even the two centuries of remarkable economic performance that have occurred in the time since Malthus should not encourage complacency about avoiding future Malthusian adjustment, especially if major renewable resource stocks (fish, forests, and soil) continue to decline. The world's current population growth rate is about 1.3% per year, significantly lower than its peak of 2.2% (reached in 1963), but still high by historical standards. Current population growth rates imply a population doubling in just over 50 years. Current resource stocks, even if degradation could be slowed or halted, would have a hard time supporting such a population at anything like current levels of real income, even adjusting for likely technological improvements over the next 50 years. A subsequent doubling in the latter half of the 21st century would seem completely infeasible. Therefore, it seems that population growth will have to fall dramatically from current levels over the next 100 years. Whether this is achieved through a benign demographic transition or through the more unpleasant Malthusian mechanisms of disease, famine, and violent conflict is still very much an open question at this stage.

SEE ALSO THE FOLLOWING ARTICLES

Biomass Resource Assessment • Depletion and Valuation of Energy Resources • Ecological Footprints and Energy • Ecological Risk Assessment Applied to Energy Development • Ecosystem Health: Energy Indicators • Population Growth and Energy • Sociopolitical Collapse, Energy and

Further Reading

Bahn, P., and Flenley, J. (1992). "Easter Island, Earth Island." Thames and Hudson, London.

Brander, J. A., and Taylor, M. S. (1998). The simple economics of Easter Island: A Ricardo–Malthus model of renewable resource use. *Am. Econ. Rev.* **88**(1), 119–138.

Clark, C. W. (1990). "Mathematical Bioeconomics, The Optimal Management of Renewable Resources." 2nd ed. Wiley, New York.

Dark, K. R. (1995). "Theoretical Archaeology." Cornell University Press, Ithaca, N.Y.

Lotka, A. J. (1925). "Elements of Physical Biology." Williams & Wilkins, Baltimore.

Malthus, T. R. (1798). "An Essay on the Theory of Population." Oxford University Press, Oxford.

Ostrum, E. (1990). "Governing the Commons, The Evolution of Institutions for Collective Action." Cambridge University Press, Cambridge.

von Daniken, E. (1970). "Chariots of the Gods? Unsolved Mysteries of the Past." Putnam, New York.

Ecological Footprints and Energy

MATHIS WACKERNAGEL and CHAD MONFREDA
Redefining Progress
Oakland, California, United States

1. **Ecological Overshoot**
2. **Conceptual Predecessors of the Ecological Footprint**
3. **Component- and Compound-Based Approaches**
4. **Method of National Ecological Footprint Accounting**
5. **Bioproductive Areas of the Ecological Footprint**
6. **Fossil Energy**
7. **Nuclear Energy**
8. **Renewable Energy**
9. **Energy Footprint of Preindustrial Society**
10. **Energy Footprint of Industrial Society**

Glossary

biocapacity The potential productivity of the biologically productive space within a specified country, region, or territory.

biologically productive space Areas of land and water capable of supporting photosynthesis at sufficient rates to provide economically useful concentrations of biomass. Marginal and unproductive regions, such as deserts, tundra, and the deep oceans, are excluded. The global biologically productive space totals 11.4 billion ha.

ecological deficit The amount by which the ecological footprint of a population (e.g., a country or region) exceeds the biocapacity of the space available to that population.

ecological footprint A measure of how much biocapacity a population, organization, or process requires to produce its resources and absorb its waste using prevailing technology.

ecological overshoot Ecological deficit met through the overexploitation of resources or accumulation of waste. William Catton defines the term as "growth beyond an area's carrying capacity, leading to crash."

embodied energy The energy used during a product's entire life cycle for manufacture, transport, operation, and/or disposal.

global hectare One hectare of biologically productive space with world-average productivity.

productivity A measurement of the potential rate of biological production of a given area. Various indicators can be used to weight productivity, including agricultural suitability, potential net primary production, and useful biomass extraction.

The ecological footprint is a measurement that compares rates of human resource consumption and waste generation with the biosphere's rates of resource regeneration and waste assimilation, expressed in terms of the area necessary to maintain these flows. Ecological footprints represent the biologically productive space required to produce the resources and absorb the wastes of a given population, organization, or product using prevailing management and technology.

1. ECOLOGICAL OVERSHOOT

The nature of ecological limits may be one of the most influential misconceptions of the sustainability debate. A former executive director of the United Nations (UN) Population Fund stated that "many environmentalists think [that the carrying capacity of Earth] is 4 billion, maximum. But now we have 6 billion people." The continued growth of the human population and resource consumption, however, is only an apparent contradiction. In reality, ecological limits can be exceeded for a period of time because nature reacts with inertia. More precisely, natural capital can be harvested faster than it regenerates, thereby depleting the natural capital stock. Ecologists call this state overshoot.

The concept of overshoot reconciles the apparently contradictory assertions that standards of living are waxing while ecological capacity is waning. Nevertheless, uncertainties remain regarding the possible impacts of prolonged overshoot. Some fear that unsustainable harvests will trigger the sudden collapse of ecosystems, leaving fisheries and other dynamic resource stocks irreversibly damaged.

 © 2004 Elsevier Inc. All rights reserved.

Others hope that the contraction of ecosystem services will be slower and more forgiving.

The delayed effects of overshoot and difficulty of valuing ecosystem services prevent markets and uninformed policy from forming corrective feedback mechanisms. Without adequate feedback mechanisms, such as systematic accounting that compares human demands on nature with nature's capacity to regenerate, overshoot can go unnoticed, thereby undermining humanity's future. Industrial economies in particular can exceed "sustainable limits" by masking resource depletion and decreasing energy returns on investment with technological advances, inexpensive energy, and easy access to distant resources. Particularly on the waste side, feedback is weak: CO_2 from fossil fuel burning leaves tailpipes with ease, independent of the increasing CO_2 concentration in the atmosphere.

No national government or UN agency operates a comprehensive accounting system to document the extent to which human demands fit within the capacity of existing ecosystems. The ecological footprint is one of the few tools that attempts integrated resource accounting.

The purpose of ecological footprints is not only to illustrate the possibility of overshoot but also to offer a robust tool to demonstrate its occurrence. As a resource accounting tool, it can help determine who is contributing, and how much, to the overall impact of humanity. Most important, it can evaluate potential strategies for avoiding overshoot.

2. CONCEPTUAL PREDECESSORS OF THE ECOLOGICAL FOOTPRINT

Biophysical assessments of human dependence on nature date back thousands of years to tales about the relationship between people and land. Records on the concept of carrying capacity trace at least as far back as Plato's "Laws," Book V, which states that a "suitable total for the number of citizens cannot be fixed without considering the land and the neighboring states. The land must be extensive enough to support a given number of people in modest comfort, and not a foot more is needed."

The first scholarly book in English on sustainable resource use may have been John Evelyn's 1664 "Sylva: A Discourse of Forest Trees and the Propagation of Timber." In North America, George Perkins Marsh's widely read 1864 study, "Man and Nature," was influential in increasing the awareness of nature's limited capacity to provide for human demands. Formal ecological accounting dates to 1758, when François Quesnay published "Tableau Economique," which captures the relationship between land productivity and wealth creation. Intellectual groundwork for more recent studies was laid in the 1960s and 1970s, particularly through George Borgstrom's concept of "ghost acreage," Howard Odum's energy analyses, and Jay Forrester's model on world resource dynamics presented in Donella Meadows *et al*.'s "Limits to Growth" report to the Club of Rome in 1972.

Natural capital accounts focused on regional and global limits are increasingly prominent in the political debate. Peter Vitousek's 1986 landmark study, "Human Appropriation of the Products of Photosynthesis," compiled consumption data to determine the human economy's draw on terrestrial net primary productivity. Based on 1980s data, the study estimates that humanity co-opts 40% of the photosynthetic materials produced by terrestrial ecosystems, excluding the photosynthesis that assimilates fossil fuel emissions and other anthropogenic waste. Fischer-Kowalski and Hüttler used material flow analysis to advance measurements of societal metabolism, a macroindicator of the environmental performance of societies. The World Resources Institute's biennial "World Resources" report and the United Nations Environment Programme's (UNEP) "Global Environmental Outlook" provide detailed documentation of pressures facing global ecosystems but do not contain aggregated impact assessments.

3. COMPONENT- AND COMPOUND-BASED APPROACHES

There are two basic ecological footprinting methods: component- and compound-based. The first calculates the footprint of individual goods or organizations. This bottom-up, component-based method first identifies all relevant items and then assesses the ecological footprint of each item using life-cycle data. The overall accuracy of the final result depends on the completeness of the component list as well as the reliability of the life-cycle assessment (LCA) of each identified component. This approach provides a high level of detail but has limitations due to LCAs' frequent inaccuracies, incompleteness, and poorly defined boundaries. The component-based method also has difficulty with data gaps and indirect expenditures outside of an organization's inventories, such as public infrastructure and military.

The compound-based method uses a top-down approach to aggregate economic and environmental statistics into national ecological footprints. Compound footprinting, which underlies the national results presented in this article, calculates the ecological footprint using aggregate national data. This information is more complete than subnational sources and LCAs and captures both direct and indirect consumption. For instance, to calculate the paper footprint of a country, information about the total amount consumed is typically available and sufficient for the task. In contrast to the component method, there is no need to know which portions of the overall paper consumption were used for which purposes, aspects that are poorly documented in statistical data collections.

4. METHOD OF NATIONAL ECOLOGICAL FOOTPRINT ACCOUNTING

National ecological footprint accounts consist of two measurements. The footprint aggregates the total area a given population, organization, or process requires to produce food, fiber, and timber; provide space for infrastructure; and sustain energy consumption. The biocapacity aggregates the total area available to supply these demands.

The ecological footprint focuses on six potentially renewable demands: cropland, pasture, forests, built-up land, fisheries, and energy. Activities in these components are deemed sustainable if rates of resource use and waste generation do not exceed a defined limit that the biosphere can support without degrading the resource stock. By including a core set of potentially sustainable activities, the ecological footprint defines minimum conditions for sustainability.

The results underestimate human impact and overestimate the available biocapacity by counting each footprint component only once if multiple demands come from the same area; choosing the more conservative footprint estimates and more optimistic biocapacity estimates when in doubt; and including agricultural practices as if they would not cause long-term damage to soil productivity. The accounts also exclude activities for which there is insufficient data or understanding, such as acid deposition. Activities that systematically erode nature's capacity to regenerate are also excluded, including aquifer depletion, soil erosion, and the release of chlorofluorocarbons.

In order to provide a quantitative answer to the research question of how much regenerative capacity is required to maintain a given resource flow, ecological footprint accounts use a methodology grounded on six assumptions:

1. It is possible to use annual national statistics to track resource consumption and waste generation for most countries.
2. Resource flows can be measured in terms of the bioproductive area necessary for their regeneration and the assimilation of their waste. (Resource and waste flows that cannot be measured are excluded from the assessment.)
3. Bioproductive areas of different productivity can be expressed in a common unit of standardized usable biological productivity. Usable refers to the portion of biomass used by humans, reflecting the anthropocentric assumptions of the footprint measurement.
4. The sum of mutually exclusive areas needed to maintain resource flows expressed in a common unit represents aggregate demand; the sum of mutually exclusive bioproductive areas expressed in a common unit represents aggregate supply.
5. Human demand (footprint) and nature's supply (biocapacity) are directly comparable.
6. Area demand can exceed area supply, meaning that activities can stress natural capital beyond its regenerative capacity. For example, the products from a forest harvested at twice its regeneration rate have a footprint twice the size of the forest. A footprint greater than biocapacity indicates ecological deficit. Ecological deficits are compensated in two ways: Either the deficit is balanced through imports (ecological trade deficit) or the deficit is met through overuse of domestic resources, leading to natural capital depletion (ecological overshoot).

Cropland, forests, pastures, and fisheries vary in biological productivity, or their capacity to provide ecological goods and services through photosynthesis. One hectare of arid rangeland, for example, has less capacity to recycle nutrients, produce food, and support diverse land-use patterns than one hectare of temperate forest. The ecological footprint therefore normalizes each bioproductive area—cropland, pasture, forests, built-up land, and fisheries—into common units of "global hectares" (gha).

Current accounts weight productivity according to agricultural suitability—a function of numerous factors, including temperature, precipitation, soils, and slope. The Food and Agriculture Organization and International Institute for Applied Systems

TABLE I

Global Biocapacity[a]

Area	Equivalence factor (gha/ha)	Global area (billion ha)	Biocapacity (billion gha)
Cropland	2.1	1.5	3.2
Pasture	0.5	3.5	1.6
Forest	1.3	3.8	5.2
Built-up land	2.2	0.3	0.6
Fisheries	0.4	2.3	0.8
Total	1.0	11.4	11.4

[a] The relative productivities of major global bioproductive areas, expressed in actual area (hectares) and productivity-weighted area (global-hectares). Equivalence factors are based on the suitability index of Global Agro-Ecological Zones 2000 (FAO/IIASA). Data are rounded.

Analysis have created a spatially explicit distribution of these variables in the "suitability index" of Global Agro-Ecological Zones 2000. Recent ecological footprint accounts use this index to translate unweighted hectares into global hectares (Table I). Other possibilities for weighting productivity include potential primary production and the rate of economically useful biomass production.

5. BIOPRODUCTIVE AREAS OF THE ECOLOGICAL FOOTPRINT

Earth has a surface area of 51.0 billion ha, of which 36.6 billion ha is ocean and inland waters and 14.4 billion is land. Ecological footprint accounts estimate that less than one-fourth of Earth's surface—9.1 billion ha of land and 2.3 billion ha of water—provides sufficient concentrations of economically useful biomass to be considered biologically productive. The 11.4 billion ha of bioproductive area is differentiated into five components: cropland, pastures, forests, built-up land, and fisheries. The remaining 39.6 billion ha is marginally productive or unproductive for human use because it consists of the deep oceans, is covered by ice, or lacks fertile soils and fresh water.

Cropland occupies approximately 1.5 billion ha of the most productive land worldwide. Current accounts include 74 crops and 15 secondary products and report details on how food crops are used (waste, seeds, feed, food, and processing), traded, and stored. These numbers document the area dedicated to cultivation and omit impacts from unsustainable agricultural practices, such as long-term damage from topsoil erosion, salinization, and contamination of aquifers with agrochemicals.

Pastures, which cover 3.5 billion ha, supply a substantial part, perhaps half, of the metabolic needs of livestock worldwide. Differing definitions of what constitutes pasture, poorly documented temporal and spatial variations in forage production and quality, and difficulties in separating metabolic needs supplied through concentrate feeds, crop residues, and other nonpasture sources make the pasture footprint difficult to determine. Healthy pastures are a critical component of food security, and even a preliminary assessment of their global capacity to support livestock would improve the current dearth of data.

Forests provide lumber, pulp, and fuelwood and include both natural forests and plantations. The Food and Agriculture Organization's "Forest Resource Assessment 2000" places the total area at 3.8 billion ha. Estimates of timber productivity are derived from the UN's Economic Commission for Europe's "Temperate and Boreal Forest Resource Assessment 2000" and data from the Intergovernmental Panel on Climate Change. These sources also provide information on plantation type, coverage, timber yield, and areas of protected and economically inaccessible forest.

Fisheries are concentrated on a relatively narrow portion of the ocean located on continental shelves. Excluding inaccessible or unproductive waters, these include an estimated 2.0 billion ha. Although a mere fraction of the ocean's 36.3 billion ha, this 2.0 billion ha provides 95% of the marine fish catch. Inland waters comprise an additional 0.3 billion ha.

Higher trophic-level fish—species high on the food chain—consume a proportionally greater portion of the ocean's primary productivity than lower trophic-level fish. It is estimated that each level up in the food chain translates into a 10-fold increase in demand for photosynthetic production per unit of biomass. Thus, 1 ton of cod at trophic level 4 has a footprint 10 times larger than 1 ton of sardines at trophic level 3. (Trophic level 1 corresponds to photosynthetic species at the base of the food chain—the primary producers.)

Built-up land for housing, transportation, industrial production, and capturing hydroelectric power occupies approximately 0.3 billion ha. Built-up land is poorly documented at a global level, partially due to different definitions used in various studies and the difficulty of satellites to view smaller objects and linear developments such as roads. However, some

countries, such as those in the European Union, have more reliable data sets. Current accounts assume that built-up land replaces cropland because most human settlements favor fertile regions well suited for agriculture. Hence, urban expansion encroaches typically on cropland.

6. FOSSIL ENERGY

The footprint of fossil fuels can be assessed by a number of approaches. The question of how much regenerative capacity is required to maintain the throughput of fossil fuel through the human economy can be addressed from three perspectives: absorbing the waste, maintaining the energy supply, and maintaining the resource. While locally significant, the area required for extraction, solid waste disposal, and power production is comparatively small at 3.2 ha/MW. In contrast, the waste assimilation method places the footprint for carbon emissions from the operation, construction, and maintenance of a typical coal-fired facility at 1900 gha/MW and the footprint of a similar-sized natural gas plant at 1050 gha/MW.

6.1 Waste Assimilation Method

The most common method for determining the ecological footprint of fossil fuels is based on the assumption that the primary limitation for fossil fuel use is the biosphere's ability to absorb CO_2 waste emissions. Following the dictum of the 1992 UN Framework Convention on Climate Change that calls for CO_2 stabilization in the atmosphere, the waste assimilation approach estimates how much larger the biosphere would need to be to provide the absorptive services to remove excess carbon emissions. Most accounts calculate this area by estimating the additional forest area required to fully sequester CO_2 emitted from the burning of fossil fuels.

Of the average annual 6.3 Gt of anthropogenic carbon emitted from fossil fuel combustion and cement production during the 1990s, the ocean absorbed an estimated 1.7 Gt and the land another 1.4 Gt. This imbalance has led to an increase in the atmospheric concentration of CO_2 from a preindustrial level of 280 parts per million (ppm) in 1750 to 370 ppm at the end of the 20th century. Given that carbon has a much shorter residence time in the terrestrial sink than in the oceanic sink, reflects prior biomass removal, and is susceptible to re-release from land-use and climate changes, current accounts only deduct the emissions sequestered in the ocean from the footprint.

Although possibilities for carbon sequestration include the fertilization of oceans and the pumping of carbon emissions underground, the reversion to forests from human-altered landscapes—cropland, pastures, and settlements—is used to calculate the footprint because it best reflects the biosphere's sequestration potential without human management. Reforestation measures have been implemented on a small scale and actively debated as a legitimate source of carbon credits in international climate negotiations. However, this approach does not suggest that the problem of increasing CO_2 concentration in the atmosphere can be solved through sequestration alone. Rather, it shows that the biosphere does not provide sufficient biological capacity for sequestration, and that it is highly unlikely that the CO_2 concentration in the atmosphere can be stabilized without reducing carbon emissions.

According to this method, the ecological footprint of fossil fuels equals the area that would be needed to assimilate carbon emissions through reforestation. Although sequestration rates vary with the temporal and spatial scales involved, and global average rates cannot be determined with great precision, sequestration would decrease significantly after 30–50 years as replanted forests mature. Although some studies have found that mature forests have retained their sink capacity, others have found that they loose their capacity or even become net carbon sources. Using the waste assimilation approach, the ecological footprint of fossil fuels provides a conservative estimate of the bioproductive area needed to assimilate carbon emissions. Assuming a generous sequestration rate of 0.95 tons C/ha/year and deducting the oceanic sink, a reforested area of more than 6 billion gha would be needed to balance current emissions. The total bioproductive land mass amounts to only 9.1 billion ha, equivalent to 10.6 billion gha.

6.2 Renewable Substitution Method

This approach builds on the idea of maintaining the energy stock. This means that depleted fossil fuel would be compensated with equivalent amounts of energy stocks. The most common long-term energy stock the biosphere accumulates is in wood. Hence, this method for calculating the ecological footprint of fossil fuels, and nonrenewable fuels in general, estimates the area required to supply their energy equivalent in fuelwood.

The ecological footprint of fossil fuels thus becomes the area needed to renew the energy equivalent of fossil fuel combustion. Unlike large-scale hydropower or geothermal energy, the energy stored in fuelwood is available in unmanaged systems and has been the mainstay of most pre-industrial societies (although many have managed fuelwood stocks). The ecological footprint expressed in fuelwood compares societies' nonrenewable energy demands with the biosphere's unmanaged renewable supply. On average, forests produce approximately 40 GJ of fuelwood/gha/year. Harvesting young, quickly growing trees may increase this capacity to 60 GJ/gha/year. Much of the energy captured by forests through net primary production is lost to decomposition and herbivory, and a portion of the remaining biomass energy, or net ecosystem productivity, is in foliage, roots, or other sources less suitable as fuel.

Variations of this technique describe the area needed to derive the energy equivalent of fossil carriers not in fuelwood but in cultivated biomass such as switchgrass, ethanol, or methanol derived from agriculture or forestry. Converting biomass fuels into liquid and gaseous mediums for use in prevailing infrastructure would net less energy than the total captured in unconverted biomass.

Andrew Ferguson of Optimum Population Trust promotes a slightly different approach by estimating an appropriate mix of renewable energy sources—photovoltaic arrays, offshore and onshore wind turbines, biomass-based ethanol, etc. The composition of the mix reflects the limitations of wind power due to its irregular availability and the cost barriers of photovoltaic arrays. He estimates that this mix could generate the current energy demand at approximately 95 GJ/ha/year. This would most likely occupy above-average-quality land, indicating an energy yield less than 95 GJ/gha/year.

6.3 Fossil Carrier Regeneration

This method treats fossil fuels like a renewable resource. The waste assimilation and renewable substitution approaches give the areas needed to balance emissions and energy stocks, not the area required to renew fossil fuels per se. This approach considers fossil fuels as a very slowly renewable resource limited by sedimentation and accumulation rates, and it defines their use as sustainable only when the fossil fuel stocks in the lithosphere are neither diminished nor transferred to the atmosphere (Table II). This method provides an assessment of the area needed to continue fossil fuel consumption indefinitely. With a global footprint of 100 gha/capita—50 times greater than global biocapacity and 100 times greater than the footprint from carbon sequestration and fuelwood substitution—it also demonstrates that the continued disparity between rates of fossil fuel consumption and creation will lead to the stock's depletion.

TABLE II

Regeneration Rates of Fossil Energy[a]

Fossil energy carriers	Yield (MJ/m^2/year)
Coal	0.05
Fuels	0.08
Natural gas	0.01

[a] Annual accumulation rates of fossil energy carriers based on estimated rates of carbon sedimentation. The data are presented in unweighted hectares. The ecological footprint typically weighs hectares by their productivity. From Ströglehner (2003).

7. NUCLEAR ENERGY

As with fossil fuels, the area occupied through extraction, production, and waste disposal of nuclear energy is relatively minor, amounting to 0.4 ha/MW. The embodied energy of nuclear infrastructure is small relative to the energy output because of the high energy density of nuclear fuel. By itself, nuclear fuel makes relatively few demands on biological productivity when contained, but intentional and accidental releases of radioactive materials have compromised human and environmental health. Failures of nuclear power plants can appropriate large bioproductive areas by making them unsuitable for human use for extended periods of time. The meltdown of Chernobyl has completely removed a 170,000-ha "zone of alienation" from economic turnover and restricted activities on hundreds of thousands of additional hectares since the 1986 accident and possibly for thousands of years into the future. Fallout contaminated larger areas extending from nearby Ukraine, Belarus, and southwestern Russia to distant Poland, Norway, and Sweden. The actual biocapacity occupied by nuclear energy is uncertain and speculative. Some studies treat the nuclear footprint as negligible or subsume these demands within national built-up areas and fossil fuel expenditures. Most nuclear footprints use the footprint of the fossil fuel energy equivalent.

8. RENEWABLE ENERGY

Some studies have examined the ecological footprint of renewable energy sources but more work is needed (Table III). This inattention is partly due to the fact that renewable sources provide an insignificant fraction of all commercial energy. Moreover, rapidly developing technologies make generalizations about conversion efficiencies, embodied energies, and life spans difficult. Although the energy embodied within the fossil fuel infrastructure is miniscule when compared with the energy expended in the fuels' extraction, transportation, and combustion, the energy embodied in most renewables' infrastructure dominates an essentially free energy source.

8.1 Hydropower

Hydropower is the largest renewable source of commercial energy. In 1995, it provided 7% of all commercial energy and 20% of electricity globally. The ecological footprint of hydropower is based on the area inundated by the reservoir created by the dam, the energy embodied in the dam, and the biological productivity lost because of the dam, primarily through depleted fish stocks. Sergio Pacca and Arpad Horvath assessed the carbon emissions released in the construction, maintenance, and operation of various electricity facilities. Basing the hydropower analysis on the Glen Canyon Dam, which generated 5.55 TWh of electricity in 1999, they determined these processes would release 25,000 tons of CO_2 per year averaged over a 20-year life cycle. The associated carbon footprint is 10.5 gha/MW.

The ratio of energy output to inundated area and the productivity of inundated land are highly variable. Energy output fluctuates with power demand and reservoir levels, although typical outputs for hydropower facilities in the United States are 40% of maximum capacity. A random survey of 24 hydropower dams in the Pacific Northwest of the United States found that the inundated areas produce an average of 3200 GJ/ha/year (10 ha/MW), varying widely between 15 GJ/ha/year (lower-course dams) and 11,000 GJ/ha/year (upper-course dams in mountain areas). A similar survey of 28 dams in 17 countries suggests an annual production of 900 GJ/ha/year (35 ha/MW), ranging from a minimum of 12 GJ/ha/year for Ghana's 833-MW Akosombo Dam to a maximum of 15,300 GJ/ha/year for Chile's 500-MW Pehuenche Dam. These values are highly variable and underestimate energy output per hectare by including the original river area as part of the inundated area. Hydroelectric plants located far from populated centers have long-distance transmission inefficiencies that increase the footprint of useful electricity to reach homes and commercial facilities.

TABLE III

Energy Yield of Renewable Sources[a]

Renewable sources	Yield (MJ/m^2/year)
Electricity, photovoltaic	450
Electricity, wind	900
Electricity, water power	100
Solar heating	1200
Wood extensive	7
Wood intensive	190
Straw as by-product	4
Miscanthus	25
Rape oil	6
Methyl ester of rapeseed	5
Organic ethanol from sugarcane	9
Organic ethanol from sugar beet	8
Organic ethanol from wood	5
Organic ethanol from wheat	5
Organic ethanol from maize	2
Organic methanol from wood	12
Organic biogas, digestor gas, and landfill gas	5

[a] The data are presented in unadjusted hectares that do not reflect differences in biological productivity. The ecological footprint weighs hectares by their productivity. These data may differ from other estimates depending on technology or other factors. For instance, sources indicate slightly higher ethanol yields in Brazil of 12 MJ/m^2/year. From Stöglehner (2003).

8.2 Wind Power

The most attractive sites for wind power are often offshore or in remote regions, potentially limiting demands for biocapacity. Moreover, the actual area occupied by turbines and other infrastructure is only 5% of the total area of wind farms, leaving the remainder available for pasture, fisheries, or other dual uses. At an average output of 8200 GJ/ha/year, the built-up footprint of a wind power plant on average-quality pasture is 1.8 gha/MW.

The energy embodied in the extensive infrastructure needed to capture a diffuse and intermittent energy source forms a large part of the footprint of wind power. Pacca and Horvath estimate that the CO_2 embodied in the construction and materials of a wind farm capable of generating 5.55 TWh/year

would total 800,000 tons over 20 years—approximately 10 times greater than the embodied emissions of a similarly sized coal-powered facility (excluding operational emissions). Assuming a 20-year life span, the carbon sequestration footprint is 16.9 gha/MW, one order of magnitude greater than the built-up area of the wind farm. If wind or another renewable carrier supplied the construction energy, the footprint of embodied energy would be greatly reduced.

8.3 Photovoltaic

Like other renewable energy technologies, photovoltaic (PV) arrays contend with low power densities. Nevertheless, annual energy output can be impressive. Assuming an annual power density of 15 W/m^2 typical in commercial systems and allowing additional space for shading and access within large-scale installations, PV systems occupy 10 ha/MW. One hectare of PV panels can deliver the thermal energy contained in 100 tons of oil annually. Without including the excess capacity required in a working infrastructure, PV arrays covering a square 160 km on each side could supply the electricity demand of the United States. With a population of 280 million people, this amounts to 0.01 ha/capita. The global area amounts to .002 ha/capita, covering a square 310 km on each side. The complete replacement of fossil fuels with a hydrogen fuel cell-based economy may require an additional area manyfold larger. PV installations atop buildings or in arid regions can mitigate competition for bioproductive areas, an advantage shared with wind power. Also, although potential PV sites may be more widespread than those for wind power, the embodied energy is also greater. Pacca and Horvath estimate the CO_2 emissions embodied in a PV system producing 5.55 TWh/year to be 10 million tons over 20 years, or 211 gha/MW.

9. ENERGY FOOTPRINT OF PREINDUSTRIAL SOCIETY

Historically, renewable energy captured chemically through photosynthesis drove nearly every aspect of society. Renewable kinetic sources—wind and water—held much less global prominence, although they expanded rapidly in the 18th and 19th centuries, when wind mills spurred the colonization of the American frontier and waterwheels became increasingly important in industrialization. In most areas, however, fuelwood supplied the majority of heat energy for small-scale activities, such as domestic heating and cooking, and larger scale enterprises such as metal smelting.

In many areas, fuelwood limitations were one factor restricting the size of preindustrial cities. According to Vaclav Smil, 12th-century Paris, then a large city of 100,000 people, may have required 1.5 tons of fuelwood per person annually, amounting to 3 PJ/year. Even if half of this demand was met through wood grown on cropland or other nonforest sources, as in some low-income countries today, the fuelwood footprint would have approached 0.4 gha/capita. The forest area required for energy demands alone equaled 50–200 times the urban area. Overland transport of fuelwood was prohibitively expensive at distances of more than 30 km, and cities with river access drew resources from great distances. Preindustrial Beijing, for example, imported huge rafts of wood from as far as Szechwan province more than 1000 km away. Fuelwood limits halted economic activities in less fortunate cities, where villages and industry competed for energy sources. According to one estimate, in 17th-century Europe fuelwood shortages halted production every other year in a typical ironworks furnace, which in 1 year of operation consumed 4000 ha of forest. Assuming a typical European growth rate of 5 tons/ha/year, this would equal a footprint of approximately 10,000 gha. Fuelwood demands increased severalfold during industrialization. By the mid-19th century, the United States fuelwood demand was 5 tons/capita, with a possible fuelwood footprint of 2.5 gha/capita—a national total of 60 million gha.

Crop residues and other cultivated biomass have played a significant but limited role in providing heat since their value as food and fertilizer introduced competing demands. David Pimentel estimates that currently fuelwood supplies half of all the world's biomass fuel by weight, whereas crop residues supply one-third. The higher energy density of fuelwood, however, means that the actual contribution of crop residues to biomass energy is smaller. More important, the chemical energy of cultivated biomass fed people and livestock, the predominant source of mechanical energy for transportation and labor. Powerful draft horses, for example, provided the labor of 10 men but could reach peak performance only when well nourished. Such an animal would eat the daily equivalent of 5 kg of oats, matching the food needs of 6 people. Nevertheless, the mechanical energy available in draft animals retained its advantage well into industrialization, and in the

United States, the demand for horses reached its height in the second decade of the 20th century. At an average yield of just over 1100 kg of oats/ha/year, a working horse may have had a cropland footprint of 3.5 gha. Some estimates place horse feed as constituting 20% of cultivated cropland in the United States at that time—amounting to approximately 60 million gha or 0.55 gha/capita.

The ecological footprint of energy obtained through fuelwood and agriculture faced local limits. Transgression of these constraints leads to an ecological deficit, a state in which ecological demands exceed local capacity. An ecological deficit signifies either imports of ecological capacity from abroad or ecological overshoot of local resources. Although ecological imports may contribute to overshoot abroad, local overshoot has immediate domestic implications. In the case of renaissance England, an escalation in charcoal demands for metal smelting led to widespread deforestation. Abundant coal supplies yielded a ready charcoal replacement and weaned England from fuelwood by the late 17th century, ushering in an early industrialization and the dawn of the first fossil fuel-dependent society.

10. ENERGY FOOTPRINT OF INDUSTRIAL SOCIETY

Tapping fossil fuel reservoirs accumulated over hundreds of millions of years exceeds local rates of fossil fuel formation, leading to local overshoot. The widespread transition to fossil fuels, however, has resulted in global overshoot and a qualitative shift in humanity's relation to the biosphere (Table IV; Fig. 1). Despite radical local and regional ecological changes from preindustrial societies, the human economy only surpassed global limits when fossil fuel use exceeded ecologically sustainable rates and induced the expansion and intensification of extractive industries such as agriculture and forestry. Although local overshoot portends immediate domestic resource constraints, global overshoot from fossil fuels externalizes costs, diffusing the feedback needed to adjust behavior.

Renewable energy sources bounded the growth of preindustrial economies, but the switch to fossil fuels curtails photosynthetic limitations. Energy production from biomass equaled that from coal and oil at the end of the 19th century. Fifty years later, however, biomass energy amounted to only one-fourth of global energy production, and its contribution to useful work was far smaller. Meanwhile, oil has increased in global prominence to become the largest primary energy source (Table V). The magnitude of this transformation has been immense. At more than 320 EJ/year, the global production of fossil fuels, for example, dwarfs the energy captured in food production by an order of magnitude.

With the advent of industrial civilization, nonrenewable sources became the prime economic drivers, supplanting not only fuelwood for heat but also livestock for transportation and labor. Fertile areas freed by the shift from biomass energy to fossil energy, decreasing transportation costs, global markets, and the diffusion of burdens from fossil fuel combustion helped spark an enormous expansion,

TABLE IV

Global Ecological Footprint and Biocapacity, 1999

Area	Footprint (gha/cap)	Biocapacity (gha/cap)
Cropland	0.5	0.5
Pasture	0.1	0.3
Forest	0.3	0.9
Fisheries	0.1	0.1
Built-up land	0.1	0.1
Fossil fuels and nuclear energy	1.2	0.0
Total	2.3	1.9

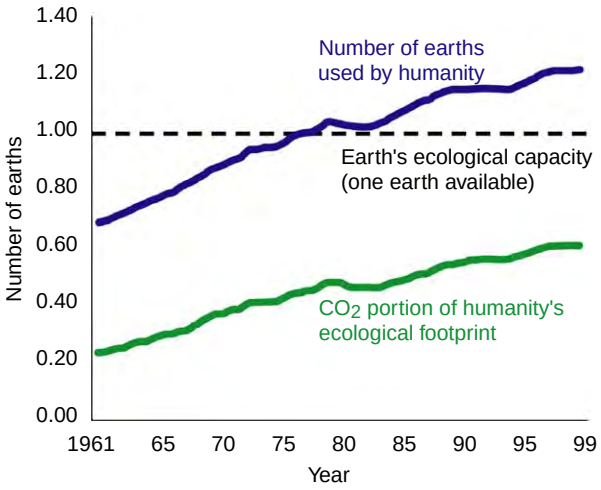


FIGURE 1 Global ecological footprint versus biocapacity, 1961–1999. Human demand exceeded the biosphere's supply during this time period, most clearly manifest in increasing CO_2 concentrations and overexploited fisheries.

TABLE V

Global Ecological Footprint of Energy, 1999

Energy source	Footprint (gha/cap)
Oil	0.5
Coal	0.4
Natural gas	0.2
Nuclear	0.1
Fuelwood	0.05
Hydro	0.002
Total	1.2

Data are rounded.

beginning with the human population in the mid-19th century and soon followed by the global economy. Today, the footprint of timber products in high-income countries is 0.7 gha/capita. Mechanized forestry boosted labor productivity with modern mills, harvesters, and chainsaws—Canadian timber extraction, for example, was 3.1 gha per Canadian in 1999. In contrast, the biocapacity of global forests approaches 0.9 gha/capita, of which only 0.4 gha/capita is exploitable. The fossil fuel transition also enables industrial fishing fleets to harvest previously inaccessible stocks, raising the footprint of the global marine catch to its estimated biocapacity of 0.1 gha/capita. The footprint of marine fish consumption averages at least triple this rate in high-income countries.

The mechanization of agriculture relieved the 20% of cropland used to feed work horses in the early 20th-century United States. The freed area became a source of meat and milk production, and in 1999 meat- and milk-producing livestock consumed 60% of U.S. maize. At the same time, the intensification of agriculture, along with new breeding techniques, initiated dramatic increases in crop yields and the human population. A substantial portion of the food energy consumed by people relies on fertilizers and agricultural chemicals synthesized through fossil fuels.

Fossil fuels have created a transitory carrying capacity of significant proportions. The ecological footprint accounts make this "phantom planet" visible. As the coiner of this term, William Catton, points out, this extra capacity gives humanity the illusion of living on a larger planet and encourages human expansion beyond the regenerative capacity of the biosphere. Consider the following comparison: 0.7 TW of human food produced today occupies more than 40% of the biosphere's regenerative capacity; in comparison, fossil fuel consumption exceeds 10 TW. From this perspective, the fossil fuel provides humanity with a temporary income that corresponds to the capacity of more than five biospheres (0.4% biospheres $\times$ 10 TW/0.7 TW = 5.7 biospheres). Therefore, some biologists liken humanity's aggregate behavior to that of species who come in contact with a large but finite food stock. Examples are deer introduced to an island or yeast cells placed in a cup full of sugar. Such an abundance of food typically leads to population growth beyond the habitat's carrying capacity, resulting in a population crash. So far, the growth curves of the fossil fuel economy have resembled the yeast cells in a finite cup of sugar. The difference is that people have choice. Like financial budgeting, ecological footprint accounts help society choose the most desirable allocation of the limited budget nature provides. The alternative is to continue liquidating natural capital.

SEE ALSO THE FOLLOWING ARTICLES

Aquaculture and Energy Use • Complex Systems and Energy • Earth's Energy Balance • Easter Island: Resource Depletion and Collapse • Ecological Risk Assessment Applied to Energy Development • Ecosystem Health: Energy Indicators • Ecosystems and Energy: History and Overview • Modeling Energy Supply and Demand: A Comparison of Approaches • Sustainable Development: Basic Concepts and Application to Energy

Further Reading

Costanza, R. (2000). The dynamics of the ecological footprint concept. *Ecol. Econ.* **32**(3), 341–346.

Ferguson, A. (1999). The logical foundations of ecological footprints. *Environ. Resource Econ.* **1**, 149–156.

Ferguson, A. (2002). The assumptions underlying eco-footprinting. *Population Environ.* **23**, 303–313.

Hoffert, M., Caldeira, K., Benford, G., Criswell, D., Green, C., Herzog, H., Jain, A., Kheshgi, H., Lackner, K., Lewis, J., Lightfoot, H., Manheimer, W., Mankins, J., Mauel, M., Perkins, L., Schlesinger, M., Volk, T., and Wigley, T. (2002). Advanced technology paths to global climate stability: Energy for a greenhouse planet. *Science* **298**, 981–987.

Loh, J. (ed.) (2002). "Living Planet Report 2002." Worldwide Fund for Nature. Gland, Switzerland.

Moffatt, I. (2000). Ecological footprints and sustainable development. *Ecol. Econ.* **32**(3), 359–362.

Pacca, S., and Horvath, A. (2002). Greenhouse gas emissions from building and operating electric power plants in the

Upper Colorado River Basin. *Environ. Sci. Technol.* **36,** 3194–3200.

Rees, W. (2000). Eco-footprint analysis: Merits and brickbats. *Ecol. Econ.* **32**, 371–374.

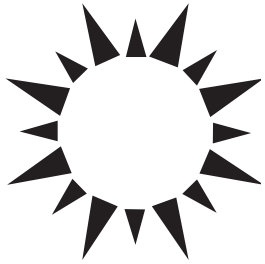
Simmons, C., Lewis, K., and Moore, J. (2000). Two feet–two approaches: A component-based model of ecological footprinting. *Ecol. Econ.* **32**(3), 375–380.

Stöglehner, G. (2003). Ecological footprint—A tool for assessing sustainable energy supplies. *J. Cleaner Production* **11,** 267–277.

Wackernagel, M., Monfreda, C., and Deumling, D. (2002a). The ecological footprints of nations: How much nature do they use? How much nature do they have, Sustainability Issue Brief. Redefining Progress, Oakland, CA.

Wackernagel, M., Schulz, N., Deumling, D., Callejas Linares, A., Jenkins, M., Kapos, V., Monfreda, C., Loh, J., Myers, N., Norgaard, R., and Randers, J. (2002b). Tracking the ecological overshoot of the human economy. *Proc. Natl. Acad. Sci.* **99**(14), 9266–9271.

Ecological Risk Assessment Applied to Energy Development



CHARLES R. HARMAN
AMEC Earth & Environmental
Somerset, New Jersey, United States

WILLIAM R. ALSOP and PAUL D. ANDERSON
AMEC Earth & Environmental
Westford, Massachusetts, United States

1. **Components of an Ecological Risk Assessment**
2. **Application to Specific Energy-Development Activities**
3. **Conclusions**

Glossary

analysis The second phase of the ecological risk assessment. It consists of the characterization of exposure, which is the measurement of the temporal and spatial distribution of a stressor, and characterization of effects, which is the identification and quantification of adverse effects elicited by the stressor.

assessment endpoint An explicit written expression of an environmental value that is to be protected in an ecological risk assessment. The environmental value consists of an entity (e.g., fish-eating birds) and a property of that entity (e.g., abundance). For example, sustaining the population of fish-eating birds in the vicinity of a water body.

conceptual site model A written description or visual representation of predicted relationships regarding how a stressor may affect the ecological components (ranging from an individual organism to a specific population, communities of populations, etc.) of the natural environment.

ecological risk assessment A process that evaluates the likelihood that adverse ecological effects may occur or are occurring as a result of the exposure of terrestrial, avian, and aquatic receptors to one or more stressors. As currently practiced in the United States, the process consists of three steps: problem formulation, analysis, and risk characterization.

measurement endpoint A concentration-dependent measurable response (typically an adverse effect such as lowered reproduction) to a stressor that is used to evaluate potential adverse effects to the environmental value specified in the assessment endpoint. For example, a comparison of the number of kingfisher (a fish-eating bird) young produced per nest at an impacted water body with the number of kingfisher young produced per nest at a nonimpacted water body.

problem formulation The first step in the ecological risk assessment, which includes a preliminary characterization of exposure; examination of available data and data needs, policy, and regulatory issues; and an evaluation of site-specific factors to define the feasibility, scope, and objectives of the ecological risk assessment. The outcome of the problem-formulation step is the selection of assessment and measurement endpoints and the development of a conceptual site model.

receptor A biotic component (from an individual organism to a population, a species to a community, etc.) of the ecosystem that is, or may be, adversely affected by a stressor.

risk characterization The third and final phase of the ecological risk assessment. It uses the results of the exposure and effects analyses to evaluate the likelihood of adverse ecological effects associated with exposure to a stressor.

stressor Any physical, chemical, or biological entity that can induce an adverse effect on a receptor.

Since the mid-1970s, public concern regarding the effects of chemicals released into the environment has increased, leading state and federal regulatory agencies to develop and adopt procedures for assessing both potential and actual ecological impacts that occur as the result of these releases. These effects are evaluated through a process called an ecological risk assessment (ERA). The requirement for ERAs was brought about by the Comprehensive Environmental Response Compensation and Liability Act (also

 © 2004 Elsevier Inc. All rights reserved.

known as Superfund), Resource Conservation and Recovery Act, and other environmental statutes. However, it was not until the late 1980s when the Science Advisory Board, the research arm of the U.S. Environmental Protection Agency, recommended that the agency place as much emphasis on evaluating ecological risks as on human health risks. Efforts to then standardize the ERA process began.

The existence of a well-established process to evaluate potential human health risks has had an important effect on the ERA process. In human health risk assessment, humans are the receptor of concern (unlike in ecological risk assessment, in which many receptors can be exposed and can potentially be at risk). Society also places a great deal of value on each individual human and on the prevention of very subtle adverse effects. In human health risk assessment, for example, an increase in mortality of 1% caused by a chemical release would be unacceptable. In ecological risk assessment, an increase in mortality of more than 1% may be acceptable if it occurs in a common and abundant species. In order to ensure that unacceptable levels of risk do not occur in humans, human health risk assessment uses numerous conservative assumptions. A conservative assumption is one that leads to overprotection rather that underprotection. Because society generally places more value on humans than on ecological receptors, and also places a unique value on different groups of ecological receptors, practitioners and users of ecological risk assessment need to be cognizant of the degree of conservatism contained in an ERA, that it is appropriate, and that it is consistent with societal norms. There is a tendency to allow the conservatism found in human health risk assessments to creep into ecological risk assessments and for ecological relevance to be lost.

With regard to the development of energy, ERAs are typically used to evaluate the effects of chemicals released into the environment, although the process has been developed to deal with "stressors" in general. The development of energy presents special challenges to evaluating impacts to environmental quality because exploration, development, refining, distribution, and use of energy from different types of energy systems have the potential for releasing a variety of chemicals into the environment over scales ranging from the small and local to regional and global levels. This article outlines concerns and issues to be considered during the preparation of ERAs for energy systems with examples from associated activities. This articles focuses on the potential effects of chemicals released during energy development and use and not on other stressors, such as nuclear radiation, heat, or the habitat disturbance that may occur during mining or oil drilling. A brief discussion of other stressors (e.g., physical effects to habitats from hydropower) associated with energy development is presented.

1. COMPONENTS OF AN ECOLOGICAL RISK ASSESSMENT

Ecological risk assessments (ERAs) typically consist of three steps: problem formulation, analysis, and risk characterization. Problem formulation provides the overall rationale and justification for the subsequent steps in the ERA. Problem formulation summarizes available information and uses it to develop clearly defined goals for the ERA as well as unambiguous criteria for evaluating potential risks. Endpoints that are ambiguous or difficult to define and measure (such as "ecosystem health") will only result in disagreement about what was intended (e.g., protection of individual receptors versus protection of a population of receptors versus protection of certain ecosystem processes such as providing a food source). The results of the problem-formulation step are a conceptual site model (CSM) and the selection of assessment and measurement endpoints. The CSM considers the dominant physical, chemical, and biological characteristics of a system; identifies relevant receptors; and identifies potentially complete pathways by which relevant receptors can be exposed to stressors. It is common for a CSM to identify 5–10 receptors to evaluate in an ERA. Typically, 1 receptor is evaluated in each assessment endpoint. For example, for a release of chemicals to a water body, one assessment endpoint may be the protection of fish-eating birds relying on that water body as a food source. The selected receptor for this endpoint might be a kingfisher (a fish-eating bird species) that has been observed feeding at the water body.

The CSM provides the basis for the selection of assessment endpoints and the measurement endpoints used to evaluate them. Site-specific assessment endpoints should be identified that address stressor-specific potential adverse effects to individual organisms (in the case of threatened or endangered species) or to local populations or communities of plants and animals. Typically, several assessment endpoints are selected to evaluate potential ecological effects based on the potentially complete pathways between the source of the chemicals and the receptors that might

be affected. For the example of a chemical release to a water body (discussed previously), an additional assessment endpoint might be the continued survival of the largemouth bass population. A third assessment endpoint would be the evaluation of the potential adverse effects to a terrestrial mammal such as a mink or raccoon that uses the water body for food and drinking water.

Measurement endpoints represent observations, measurements, or analyses that evaluate an assessment endpoint. One to several measurement endpoints are used to evaluate each assessment endpoint. For the kingfisher described previously, one set of measurement endpoints might focus on whether the aquatic community in the pond can still provide enough food to sustain the kingfisher population living near the pond. Measurement endpoints may include direct comparison of water body chemical concentrations to ambient water quality criteria; laboratory bioassays of site water with standard aquatic organisms; and field surveys of aquatic communities comparing the site to a reference area that has the same features as the site water body, except for the presence of site-related chemicals.

For measurement (and assessment) endpoints evaluating ecological communities, the distinction between structure and function of the endpoints is important. For the example of a chemical release to surface water, historically many ecological risk assessments have included the structure of the benthic community (e.g., diversity) as an assessment endpoint. It may be more appropriate to focus on the functional aspects of the benthic community (e.g., its function as a prey base for higher trophic-level predators). A change in the structure (i.e., the diversity of organisms present) of a lower trophic-level community that serves as a prey base for higher trophic-level predators may be less important than a change in the biomass of prey produced (i.e., the weight of prey available to be eaten by predators) by the community.

A second important distinction is between chemicals that can be directly toxic but do not bioaccumulate and chemicals that are directly toxic and also bioaccumulate. For chemicals that can bioaccumulate (e.g., PCBs and mercury), assessment endpoints generally include higher trophic-level predators that may accumulate these chemicals from the food chain. For these compounds and receptors, endpoints evaluating indirect toxicity through food chain exposures may be more relevant than endpoints evaluating direct toxicity. Conversely, for chemicals with low potential for bioaccumulation, selection of assessment endpoints may concentrate on lower trophic-level receptors that may have higher direct exposures to the chemicals. For these compounds and receptors, assessment endpoints evaluating direct toxicity to fish or benthic invertebrates (e.g., reduced growth, reduced reproduction, or even mortality) may be more relevant than those evaluating indirect toxicity from food chain exposures.

Analysis is the second step in the ERA process. In this step, exposures of receptors to stressors are evaluated and the relationship between stressor levels and ecological effects is documented. Exposure characterization describes potential or actual contact or cooccurrence of stressors with receptors.

In large and complex ERAs, exposures of ecological receptors to chemicals may vary over time and space depending on the characteristics of the selected receptors. Migratory species will spend less time being exposed to site-specific chemical concentrations than resident species. For species with large home ranges that forage over large areas that do not contain site-related chemicals, exposures will be lower than those for species with smaller home ranges that may be exposed to site-related chemicals for their entire lifetimes. Exposures may be measured directly by sampling receptors or the organisms that comprise their diet, but more commonly exposures are estimated using a series of standard equations. The equations require information about the concentration of chemicals in food, foraging areas and habits, food consumption rates, body weight, water consumption rates, and incidental media-specific ingestion rates (e.g., the amount of soil ingested by robins feeding on worms). This information is obtained from sources ranging from literature summary values to species-specific literature values and results from experimental studies conducted in the field.

Ecological effects characterization identifies the effects (e.g., mortality and reduced reproductive success) elicited by a stressor, the relationship between these effects and the assessment endpoints (e.g., sustaining the population of fish-eating birds in the vicinity of a water body), and provides an evaluation of how these effects change with varying stressor levels. The characterization begins by evaluating available effects data to specify (i) the effects that are elicited by the stressor, (ii) the plausibility that effects may occur or are occurring as a result of exposure to the stressor consistent with the conceptual site model, and (iii) the links between measurable ecological effects and assessment endpoints when the latter cannot be directly measured.

Effects data may consist of chemical-specific literature values from laboratory studies on the selected receptor species, results of chemical-specific studies on taxonomically related species, and results of studies on structurally similar chemicals. Site-specific field studies may also be evaluated as part of the effects characterization, depending on the selected assessment endpoints (e.g., sustaining the population of fish-eating birds in the vicinity of a water body) and measurement endpoints (e.g., field studies of reproductive success). Data typically collected from laboratory studies include effects ranging from the molecular level (e.g., enzyme induction) to the organ level (e.g., liver or kidney toxicity) and the organism level (e.g., mortality). Data typically collected from field studies include population-level effects (e.g., reduced numbers or reduced reproduction).

The selection of the appropriate study to use in the ecological effects characterization requires linking the study to the selected assessment endpoint. If the assessment endpoint includes a threatened or endangered species, then the effects characterization may focus on effects to an individual. However, if the assessment endpoint is sustaining the population of fish-eating birds in the vicinity of a water body and one measurement endpoint is sustaining the food supply for the fish-eating birds, then the effects characterization may focus on population-level effects to prey fish species.

The last step of the ERA process is risk characterization. This step combines the results of the analysis phase to develop estimates of the potential risks posed to the ecological receptors included in the assessment endpoints identified in the problem-formulation step. Risk characterization evaluates whether the doses measured or estimated in the exposure assessment are higher than the doses (identified in the effects assessment) that cause adverse effects and whether the estimated exposures are likely to cause adverse effects to the selected assessment endpoint. Because ERAs generally contain conservative assumptions, it is possible that an estimated exposure exceeds the allowable exposure but that an adverse effect is not expected to the assessment endpoint. Risk estimates developed from measurement endpoints are evaluated in the context of the significance of any adverse effects and lines of evidence supporting their likelihood of occurrence. For the assessment endpoint of a threatened or endangered species, this may include an evaluation of whether individual-level effects are likely to occur.

For assessment endpoints that are being evaluated using multiple measurement endpoints, a weight of evidence approach is typically used. In this approach, the weight conferred on each measurement endpoint, the magnitude of the response to each endpoint, and the concurrence among measurement endpoints are used to evaluate whether significant ecological effects are likely to occur. The weight of each measurement endpoint is a function of the degree of association between the measurement and assessment endpoints, the study design, and the quality of the data supporting the measurement endpoint. In the second component, strong unambiguous responses provide greater weight than weak responses. Finally, when a number of different measurement endpoints indicate similar responses, this provides additional weight. For example, one measurement endpoint [e.g., comparison of the number of kingfisher (a fish-eating bird) young produced per nest at an impacted water body with the number of kingfisher young produced per nest at a nonimpacted water body] may indicate no adverse effects to the assessment endpoint (e.g., sustaining the population of fish-eating birds), but an additional assessment endpoint including an evaluation of the potential for adverse effects to the aquatic community that serves as the food source for the kingfisher may indicate the potential for an adverse effect. A weight of evidence approach is used to evaluate whether the estimated effects to the food supply are sufficient to cause adverse effects to the selected assessment endpoint. Considerations in this evaluation also include the presence of alternative sources of food and the likelihood of their use by the kingfishers.

2. APPLICATION TO SPECIFIC ENERGY-DEVELOPMENT ACTIVITIES

The development of energy, although providing a benefit to society, can adversely affect the environment. ERAs can be used to assess the magnitude or severity of the perturbation and in determining cleanup levels in circumstances in which a cleanup is required. As previously mentioned, the focus of this article is on the evaluation of potential ecological effects associated with chemical releases from energy development.

The ERA process can be used to evaluate other nonchemical stressors, such as physical disturbance, radiation, and thermal changes in an environment. For example, nuclear power-generating stations have

the potential to release radioactive materials into the environment. As a stressor, radionuclides can be examined via the same ERA process as that used for chemical stressors. This process includes identification of appropriate potential receptors and endpoints, determination of levels of radionuclides in the environment, analysis of the levels of existing radiation versus benchmarks levels for organisms of concern or the endpoints that are being assessed, followed by the characterization of the risks to populations of various organisms in the release area. As with other energy-development activities, the critical step in the process is a thorough and appropriate problem-formulation step, including the development of a conceptual site model that emphasizes the circumstances of the release area and takes into consideration the movement and ultimate effects of radiation on natural ecosystems.

Although the ERA process evaluates the potential for chemicals to affect biological organisms, it is important to understand the impact that physical disturbance and habitat alteration may have on the ecosystem versus the release of development-related chemicals. Hydroelectric power, which releases few chemicals to the environment, may block the movement of fish up and downstream; may reduce large fluctuations in hydrologic regimes that may hinder plant and animal species that are competitively more adapted to living in flood-prone areas; or may change nesting areas, resulting in the modification of faunal communities in an area. Nuclear and fossil fuel power plants may impinge or entrain large quantities of fish and aquatic life through water intake for cooling purposes. These facilities may discharge heated water into the environment that may kill fish, block fish migrations, or result in the encroachment of nuisance species. Mining for coal can result in the stripping of mountaintops and the resulting loss of habitat used by terrestrial wildlife. All these are "nonchemical" stressors with the potential to pose ecological risk that can be evaluated using ERAs.

The point is that an ERA is a tool used to evaluate the effects of stressors on the environment. Its results must be placed in a realistic context of the breadth of stressors that result from a given energy-development activity. As discussed previously, and discussed throughout much of this section, the problem formulation stage is the critical first step in identifying what potential issues must be evaluated in the ERA and how they relate to other stressors in the environment. The following sections discuss issues related to the use of ERAs for specific energy-development activities.

2.1 Ecological Risk Assessment Applied to Oil and Gas Activities

Because energy development is dominated by the use of oil and gas as fuel sources (approximately 40% of the world's energy supply is from crude oil), this article focuses on the ecological effects associated with the various stages of energy development from oil and gas. Oil- and gas-related activities have the potential to adversely affect ecological receptors through the release of a variety of chemicals to the environment. Glen W. Suter outlines a general framework for conducting ecological risk assessments at sites at which petroleum hydrocarbons are the dominant chemicals that may potentially pose an ecological concern.

2.1.1 Petroleum Hydrocarbons in the Environment

In 1989, the U.S. Environmental Protection Agency (EPA) reported that 20.9 million barrels of production waste and 361 million barrels of drilling waste were generated from approximately 835,000 oil and gas wells nationwide. Additionally, they reported that on average approximately 9000 oil spills occurred each year, releasing approximately 11 million barrels of oil into the environment. Ecological impacts associated with these releases ranged from short-term effects such as fish kills to long-term impacts such as chronic changes in benthic macroinvertebrate community structure (e.g., the number and diversity of invertebrates living in sediments).

The ecological impacts produced by these releases occurred as a result of physical actions, such as smothering and fouling, and the direct toxicity of the various chemicals in the petroleum mixtures. Key organic compounds in such releases include polycyclic aromatic hydrocarbons [e.g., phenanthrene and benzo(*a*)pyrene] and volatile organic compounds (e.g., benzene). Although the potential effects from inorganics (e.g., metals) are generally not of concern where petroleum mixtures are spilled in the environment, metals such as arsenic, barium, cadmium, chromium, copper, lead, mercury, and selenium can be released at levels that are toxic to terrestrial and aquatic receptors during oil and gas production activities.

Petroleum is a mixture of hydrocarbons that can be classified into two broad categories: aliphatic hydrocarbons, which can be further divided into alkanes, alkenes, and cycloalkanes, and the aromatic hydrocarbons. The percentage of each group present varies with both the source of crude oil and the

refined product. The compounds within these two chemical groups exhibit a range of toxicity, with some being extremely toxic, whereas others are relatively benign. In general, the toxicity of hydrocarbons is related to their chemical structure and their affinity for water, with more soluble compounds being relatively less toxic. Although this is a general relationship and can be affected by chemical structures, compounds that are more soluble in water are less soluble in fat. As the solubility in fat or lipids increases, solubility in water decreases and the ability of the hydrocarbon to cross the cell membrane and produce biological damage increases. This relationship also suggests that as lipid solubility increases, the potential for the chemical to be passed up through the food chain (from an original food source to a lower level predator and then to ever higher levels of predators) increases.

Understanding these toxicological properties is important because of the general manner in which many regulatory agencies evaluate the presence of petroleum at sites. As opposed to measuring the concentration of each chemical in a hydrocarbon mixture, many states regulate and evaluate petroleum sites based on the quantification of total petroleum hydrocarbons (TPHs) in surface water, sediments, and soil. The EPA defines TPH as "a measure of the concentration or mass of petroleum hydrocarbon constituents present in a given amount of air, soil, or water" and further notes that

> *the term total is a misnomer, in that few, if any, of the procedures for quantifying hydrocarbons are capable of measuring all fractions of petroleum hydrocarbons present in the sample. Volatile hydrocarbons are usually lost in the process and not quantified. Additionally, some nonpetroleum hydrocarbons may be included in the analysis.*

To the extent that the hydrocarbon extraction efficiency is not identical for each method, the same sample analyzed by different TPH methods will produce different TPH concentrations. TPH analytical procedures use a variety of techniques, such as gas chromatography (GC), infrared spectroscopy (IR), gravimetric techniques, and immunoassays. Because of different technical processes, each technique detects a different range of hydrocarbons. GC techniques normally measure the broadest range of hydrocarbons, whereas IR techniques measure most hydrocarbons with the exception of volatiles and compounds with high numbers of carbon atoms. Gravimetric methods measure anything that is extractable except volatiles, and immunoassays measure only aromatic hydrocarbons, such as polycyclic aromatic hydrocarbons and benzene, toluene, ethylbenzene, and xylenes compounds. Interference with the analytical procedure of some of the methods can result from the presence of natural oils, humic substances, and peat moss. TPH measurements have been made of grass (14,000 mg/kg of TPH), pine needles (16,000 mg/kg of TPH), and oak leaves (18,000 mg/kg of TPH). Assessing risk based only on TPH is difficult because the analytical procedures identify all hydrocarbons and not specific ones, regardless of the source. As such, TPH analysis does not adequately characterize a petroleum mixture and provides little basis for characterizing the fate, transport, and toxicological properties of the mixture.

Under these conditions, using TPH leads to overestimates of ecological risk if, for example, the allowable TPH concentration was based on a "toxic" petroleum mixture and the petroleum being assessed was from a relatively "inert" source. Conversely, ecological risk may be underestimated if the reverse were true—the TPH in the field was from a petroleum source that was more toxic than that used to establish the allowable TPH level. The more appropriate manner for characterizing site conditions related to petroleum releases is the use of analytical procedures that evaluate the semivolatile and volatile organic compounds using standard EPA protocols. This is supported by guidance documents from Oak Ridge National Laboratory that state that TPH analysis provides little basis for performing an ERA because it supplies limited information about the composition and, therefore, the properties that determine the potential fate and toxicity of the material. However, TPH may be useful for determining the extent of contamination or the locations of the greatest contamination.

In addition to the shortcomings in adequately characterizing differences in toxicity in different petroleum hydrocarbons, TPH analyses also fail to consider the processes that cause petroleum hydrocarbons to partition into the environment. Mixtures of petroleum hydrocarbons change once released into the environment because weathering, the process by which natural chemical and physical processes alter the composition of the mixture, changes the bioavailability of the remaining hydrocarbons and hence their toxicity. Weathering is primarily driven by dissolution in water, biodegradation, and volatilization.

Because weathering affects each class of hydrocarbons differently, the toxicity of a particular mixture changes as it weathers. The solubility and volatility of all compounds generally decrease with

an increase in molecular weight. Aromatic hydrocarbons, those hydrocarbons with one or more benzene rings, are more soluble in water, whereas low-molecular-weight aliphatic hydrocarbons, characterized by a straight chain of carbon atoms, are more volatile. Low-molecular-weight compounds, which are more soluble and volatile, tend to migrate from contaminated soil at a much faster rate than high-molecular-weight compounds. Residual compounds left in highly weathered hydrocarbons tend to be the higher molecular-weight compounds that are less mobile and generally less toxic.

2.1.2 Oil and Gas Production

Potential ecological impacts from releases of chemicals during oil production activities are generally small in scale, although they can produce severe effects to locally affected communities. Compounds of ecological concern can occur in drilling mud and fluids, which when released into the environment can produce fish kills, smother benthic invertebrate communities, and destroy vegetation. Such compounds include arsenic, chromium, copper, iron, and other metals as well as oil, grease, chlorides, nitrates, and other substances. These substances tend to be directly toxic to fish, wildlife, and other biological receptors but generally do not bioaccumulate at successively higher trophic levels of the food chain.

As a result of the small scale of individual production activities at a single well, ERAs may overestimate risks to populations if they used conservative assumptions and focused on individuals rather than having the problem-formulation phase consider the issue of scale in the context of appropriate receptors and endpoints. In examining a single drilling site or release via the standard three-step ERA process, consideration must be given to applying realistic exposure parameters to the typical receptors found in a given area. Because the focus of the ERA must be on evaluating the potential for ecological impacts to local populations, care must be taken to ensure that the inference of impacts to local organisms is not considered as a significant impact to the population. A drill rig located on a $\frac{1}{2}$-acre site in a 1000-acre tract that supports a population of the meadow vole *Microtus pennsylvanicus* might only affect 4–6 animals based solely on the territory that the animals occupy. Out of the greater population of animals, in which an individual female may bear as many as 150 young per year, this would place into question the ecological relevance of a risk assessment that is not focused in the problem formulation on population issues. However, if the 1000-acre tract were covered with a large number of drilling sites, the cumulative effect of many sites, all releasing approximately the same quantity of compounds, could result in ecological impacts to populations of animals and plants.

2.1.3 Petroleum Refining, Storage, and Processing

From an ecological standpoint, evaluating the ecological impacts associated with residual petroleum releases at major oil-production facilities can be difficult because of the complexity of the sites, the large number of chemicals involved, and the ongoing production practices. An examination of many of the large refineries or terminals shows that these facilities are often located adjacent to sensitive environments, such as coastal waterways, rivers, and harbors. In New Jersey, the largest refineries can be found at locations along the Delaware River in the south and adjacent to Raritan Bay, Arthur Kill, and Newark Bay in the New York/New Jersey Harbor area. These locations present ample opportunity for the discharge of petroleum and petroleum-related products into productive ecosystems containing receptors that will be in close contact with the released materials.

Part of the complexity in dealing with such facilities is their size, the diversity of habitats and the variety of ecological receptors that can be exposed, and the age of the facilities. Refineries can be very large, in excess of 150 acres, and contain freshwater and coastal wetlands, ponds, and forests. Although production activities within refineries and terminals dominate land use, most refineries also contain land that is undisturbed, fallow, or covered with storage tank fields, landfills, and waste management units. Many of the areas supporting seemingly natural habitats may have historically been used for storage or production. At several refineries in New Jersey, forests and shrublands have grown on parts of the property that previously contained storage tanks that were removed from service and dismantled. These areas of terrestrial vegetation support a number of wildlife species, including white-tailed deer (*Odocoileus virginianus*) and avian predators such as red-tailed hawks (*Buteo jamaicensis*). However, these habitats have developed in areas created by fill and often contain residual levels of petroleum hydrocarbons. From a gross standpoint, the fact that these areas have revegetated substantiates the limited effects of residual hydrocarbons in the environment. Nevertheless, ERAs are generally required in such circumstances to evaluate any potential risks associated with residual levels.

In preparing an ERA for a refinery or large production facility, great care must be given to the problem-formulation step. As opposed to ERAs conducted under production or drilling scenarios, refinery ERAs must consider three complex issues. First, refineries often contain hundreds of existing or former discharge sources. These sources can include crude oil or basic petroleum products as well as more refined petrochemical products that may potentially be more toxic than crude oil. Second, the pathways associated with a refinery ERA often include both aquatic and terrestrial environments, with numerous assessment endpoints in each of these environments that may need to be evaluated. Typical assessment endpoints include the maintenance of viable terrestrial plant communities; maintenance of viable aquatic and wetland plant communities; maintenance of aquatic wildlife, fish, and benthic invertebrate communities; and maintenance of both resident and migratory terrestrial wildlife populations. Third, developing a conceptual site model for a refinery or terminal is often complicated by the age and location of these facilities. In industrial sections of the northeastern United States, some refineries began operations in the late 1800s or early 1900s, whereas general industrialization extends back to the late 1700s. In these settings, it is often difficult to distinguish impacts associated with refinery operations from those that occurred as a result of general urbanization and to isolate site-related ecological impacts from impacts that may have occurred as a result of neighboring facilities that may also be releasing constituents of concern to the environment.

In settings in which refineries have residual concentrations of chemicals but still support viable populations of terrestrial and aquatic flora and fauna, comparison to background chemical data can be an important indicator of site-related effects. In this case, it is essential to collect background chemical data from locations similar in habitat but still subject to the same level of historic disturbance as the refinery. Refineries are typically located adjacent to, or close to, major harbors, with more than 60 petroleum facilities located in the Newark Bay estuary alone. Between 1982 and 1991, a total of 1453 accidental incidents released more than 18 million gallons of hazardous materials and petroleum products into Newark Bay. Thus, identification of background locations can be a contentious issue in industrialized settings because chemical concentrations can be quite high—high enough that they would require cleanup if the site were in a pristine setting.

The EPA has issued guidance on use of background data in ERAs. Although the agency recognizes that sources other than the facility being investigated may release chemicals into the environment, the expectation is that the effect of background concentrations will be included in risk estimates, even if background is the primary source of risk, and be qualitatively discussed in the risk characterization section of the ERA. The importance of the background concentrations of chemicals of concern is applicable in the risk management phase following the completion of the risk assessment. The assumption is that even if the background chemical concentrations were anthropogenically produced, taking action to address their presence at a particular site would be neither practical nor cost-effective if the background levels are widespread. Furthermore, remediated sites, which receive a constant stream of off-site contamination, will only be recontaminated from upstream sources. Particularly in aquatic sediments, a remediated area can be recontaminated by the movement of contaminated sediments from upstream areas through the action of currents or by the release of pollutants from upstream facilities.

2.1.4 Distribution

Distribution of oil and gas through pipelines can cause large-scale effects that are not related to chemical releases (e.g., alteration of many miles of habitat during construction) and relatively local effects from releases of oil during a pipeline failure or releases of other chemicals associated with transport of oil and gas. The most well-known of these was the release of PCBs at gas pipeline compressor stations. Unlike most petroleum constituents, PCBs have the potential to bioaccumulate. Consequently, assessments of ecological risks associated with PCB releases required food web modeling to predict ecological risks to upper trophic-level predators.

The other common mode of oil distribution is via ships. The release of oil in aquatic ecosystems has two primary modes of impact. One is physical fouling or impairment of biological organisms, such as birds, fish, and plants. The other is the toxic effect of the chemicals in the oil or refined product or any residual compounds. Physical fouling can be the most destructive action, but its effects are short term and mitigated by the dilution and degradation of the released oil. The toxic effects are often longer term and are assessed through the ERA process.

2.1.5 Retail Service Stations

Typically, individual gasoline service stations do not pose a significant threat to ecological resources. Service stations are small in relation to the surrounding environment and most are located in developed, urban areas or along paved roadways. These areas generally do not support large communities of native plants or animals that can be exposed to petroleum-related products nor are they generally located near sensitive habitats. Additionally, the major contamination sources at service stations are leaking underground storage tanks used to store gasoline and other products. Because the tanks are located in the ground, exposure pathways to ecological receptors in the surface environment are generally not complete. Without a complete exposure pathway, there can be no risk. However, exposure can occur if a leaking tank contaminates groundwater, which in turn is discharged to a surface water body or a wetland. In such instances, chemicals to which ecological receptors may be exposed include volatile organic compounds such as benzene, toluene, xylene, ethylbenzene, and methyl *tert*-butyl ether (MTBE) and, depending on the age of the release, inorganic constituents such as lead.

In evaluating the ecological risks associated with groundwater contamination originating from a service station located in the Pine Barrens Management Area of southern New Jersey, environmental sampling indicated that ethylbenzene, xylenes, MTBE, and lead were being released over a 50-ft stretch of a small steam flowing through a mixed hardwood forested wetland. The problem formulation indicated that the release had the potential to affect invertebrates within the stream sediments, fish living in the stream, and wildlife that either used the stream as a water source or fed on plants growing adjacent to the stream. Investigation of the release found maximum concentrations in sediments of 34 μg/kg of xylenes, 50 μg/kg of MTBE, and 24 mg/kg of lead. The only chemicals identified in surface water were MTBE (0.5 μg/liter) and lead (26.5 μg/liter). To characterize potential risk, chemical concentrations in surface water and sediments were compared to values for specific chemicals that are meant to be protective of the environment under most conditions. Slight exceedances existed, so a site-specific biological survey of benthic invertebrates in sediments was conducted. Benthic invertebrate surveys conducted at reference locations and site-sampling locations found no difference between the two areas in the kinds of species present, indicating that the aquatic community in the affected stream was similar in all respects to the aquatic community in unaffected streams. The ERA concluded that no adverse ecological impacts were associated with the discharge of contaminated groundwater. Although remedial actions to address groundwater contamination and potential human health risks were implemented, no actions were necessary within the stream or wetlands to protect ecological resources.

2.2 Ecological Risk Assessments Associated with Coal Mining

Whereas ecological risk issues complicated by industrial development and background contamination dominate some aspects of the petroleum industry, coal mining is generally an activity that occurs in more natural settings. The obvious ecological impact associated with coal mining is the destruction of habitat. However, chemicals can be released into the environment from the development of ore piles, mine tailings, and spoil heaps.

The chemical release of greatest ecological concern is acid mine drainage (AMD). In certain circumstances, surface water and shallow groundwater will react with rocks containing sulfur-bearing minerals to create sulfuric acid. This acidified water then leaches metals such as aluminum, iron, and manganese from rocks that come in contact with the water. Aluminum concentrations of 57 mg/liter have been measured in AMD from mining in Appalachia. The EPA chronic water quality criterion for aluminum is 87 μg/liter under alkaline conditions (the potential for aluminum toxicity increases with increasing acidity), which is approximately 1000 times lower than the aluminum concentration in AMD. The EPA chronic water quality criteria for acidity (pH) ranges between 6.5 and 9 standard units, whereas pH in AMD from Appalachia has been reported to be as low as 2.7. AMD has caused the complete elimination of fish from some lakes and streams in the Appalachians or the presence of only a few acid-tolerant fish species in other waters.

2.3 Ecological Risk Assessment Procedures Applied to Power Generation

Power generation through the burning of fossil fuels can lead to a variety of chemical-related impacts to the environment. Also, direct habitat loss can occur as a result of facility construction, tank storage, or coal pile maintenance. Chemical impacts can occur as the result of catastrophic releases of oil or waste materials

such as caustic soda, the generation of AMD (from coal piles), or the improper storage or mishandling of waste sludges and fly ash wastes. Fly ash wastes from coal-fired power plants contain metals, such as aluminum, arsenic, cadmium, chromium, copper, mercury, and selenium, that can be discharged at concentrations sufficient to significantly impact benthic invertebrate populations and fish communities as well as terrestrial plant and animal communities.

The most wide-ranging ecological issue associated with power plants is the release of chemicals through stacks into the atmosphere. Conceptual site models for air-dispersion ERAs must take into account the numerous potentially complete exposure pathways that can spread hundreds of miles from the site and affect many receptors. In such instances, ERAs must focus on select indicator organisms that are used as the most likely impacted organisms as well as sensitive receptors that could be most adversely impacted by the stack emissions. Particularly important issues related to air emissions, mercury, and acid rain are discussed in detail in the following sections.

2.3.1 Potential Ecological Impacts from Mercury Emissions

The EPA completed an assessment of the potential impacts of mercury emissions, part of which consisted of an ERA. It estimated that approximately 87% of the estimated 144 Mg (158 tons) of mercury emitted annually into the atmosphere by anthropogenic sources in the United States is from combustion point sources, 10% is from manufacturing point sources, 2% is from area sources, and 1% is from miscellaneous sources. Four specific source categories account for approximately 80% of the total anthropogenic emissions: coal-fired utility boilers (33%), municipal waste combustion (19%), commercial/industrial boilers (18%), and medical waste incinerators (10%). These sources represent high-temperature waste combustion or fossil fuel processes. For each of these operations, mercury is present as a trace chemical in the fuel or feedstock. Because of its relatively low boiling point, mercury is volatilized during high-temperature operations and discharged to the atmosphere with the exhaust gas.

Mercury is emitted from stacks primarily in its elemental or inorganic (combined with other inorganics, such as sulfide or chloride) form. Mercury is transported and deposited onto soils and surface waters, where some is converted to organic mercury (e.g., methylmercury). Inorganic and elemental mercury are directly toxic through ingestion and inhalation, but organic mercury tends to bioaccumulate and toxicity is expected through the food chain in higher trophic-level organisms, not in directly exposed lower trophic-level organisms. That is, because mercury accumulates through the food chain, exposures experienced by higher trophic-level organisms to bioaccumulated levels of mercury can cause direct toxicity, whereas exposures experienced by lower trophic-level organisms to nonbioaccumulated levels may not be sufficient to cause toxicity.

The EPA's risk assessment evaluated the fate and transport of mercury emissions and used estimated atmospheric mercury concentrations and deposition rates as inputs in a model to derive concentrations of mercury in soils and surface waters over the drainage basis for an entire water body. These soil and water concentrations were used to estimate mercury (as methylmercury) concentrations in the associated biota and fish, which humans and other animals were assumed to consume. Methylmercury intake (on a per body weight basis) was estimated for six wildlife species: kingfisher, river otter, loon, mink, osprey, and bald eagle. Methylmercury exposures for the most exposed wildlife species (the kingfisher) were estimated to be up to two orders of magnitude higher than human exposures from contaminated freshwater fish. For wildlife, the importance of site-specific effects of mercury exposure is anticipated to be greater than that for humans in the general population because fish-eating wildlife obtain their fish from a much more limited geographic area than do people, and the diet of some wildlife species is composed almost entirely of fish.

The broad ecosystem effects of mercury are not completely understood. No applicable studies of the effects of mercury on intact ecosystems were found. Consequently, characterization of risk for nonhuman species did not attempt to quantify effects of mercury on ecosystems, communities, or species diversity. The characterization focused on quantities of mercury that may adversely affect the health of sensitive subpopulations of wildlife species and on the collocation of these populations with areas of elevated mercury exposures from anthropogenic emissions of mercury.

The results of the EPA's assessment suggest that ecosystems most at risk from airborne releases of mercury exhibit one or more of the following characteristics:

1. They are located in areas where atmospheric deposition of mercury is high.
2. They include surface waters already impacted by acid deposition.

3. They possess characteristics other than low pH that result in high levels of bioaccumulation.
4. They include sensitive species.

2.3.2 Ecological Impacts Associated with Acid Rain

Acid rain is the generic term for the wet deposition of substances that make the rain more acidic than the normal pH 5.6 of carbon dioxide in equilibrium with distilled water. Acidic deposition includes both wet and dry deposition of material more acidic than normal based on the concentrations of sulfates and oxides of nitrogen. In 1997, the major sources of sulfur dioxide emissions were electric utilities (60%), industrial combustion (17%), and industrial processes (8%). Transportation sources (including cars, trucks, and nonroad vehicles such as construction equipment) accounted for more than 50% of all nitrogen oxide emissions. Other major sources of nitrogen oxides include electric utilities (26%) and industrial combustion (14%). According to regional studies conducted by the EPA since 1990, agricultural activities, particularly manure handling, constitute the largest source of ammonia emissions. Motor vehicles and industrial processes also contribute to ammonia emissions.

Although regulatory controls in the 1970s and 1990s resulted in decreased sulfur dioxide emissions, levels remained high compared to background conditions. From 1995 to 1997, wet deposition of sulfate in the northeast was approximately 20% lower than levels in the preceding 3 years with implementation of the 1990 Clean Air Act Amendments. Controls on nitrogen oxides and ammonia were not fully addressed during the same period; consequently, emissions of these compounds have remained largely unchanged.

The 1990 National Acid Precipitation Assessment Program report to Congress concluded that there was insubstantial evidence that acid deposition had caused the decline of trees, other than red spruce, growing at high elevations. Research has shown that acid deposition contributes to the decline of red spruce trees throughout the eastern United States and sugar maple trees in central and western Pennsylvania. Reported symptoms of tree decline include poor crown condition, reduced tree growth, and unusually high levels of tree mortality.

Acid deposition degrades water quality by lowering pH levels (i.e., increasing acidity), decreasing acid-neutralizing capacity, and increasing aluminum concentrations. Although sulfate concentrations in lakes and streams have decreased during the past 20 years, they remain high compared to background conditions. Moreover, improvement in the acidity of lakes and streams in the northeast has been limited. For example, many Adirondack region lakes are still acidic or subject to short-term pulses in acidity associated with precipitation events.

3. CONCLUSIONS

The ERA process consists of three steps: problem formulation (including developing a CSM and selecting assessment and measurement endpoints), analysis (exposure and effects), and risk characterization. Problem formulation (including selection of the ecological values to be protected) is the most important step because it identifies the ecological values that need to be protected and evaluated. Although scientific considerations play a role in identifying these values, value judgments (subjectivity or professional judgment) have an equal if not more important role, for example, in selecting which species to use as an indicator, whether to examine individual or population effects, or whether to consider structure or function of a plant or benthic invertebrate community. Problem formulation also involves the evaluation of complete versus incomplete exposure pathways. The analysis phase includes the evaluation of potential exposures of receptors to chemicals and the evaluation of the potential toxicity of these chemicals to the receptors.

The last step of the ERA process is risk characterization, which evaluates whether the doses measured or estimated in the exposure assessment are higher than the doses (identified in the effects assessment) that cause adverse effects and whether the estimated exposures are likely to cause adverse effects to the selected assessment endpoint. Because ERAs generally contain conservative assumptions, it is important to evaluate the contribution of background concentrations.

In the development of energy and energy sources, wherever release of chemicals to the environment occurs, an ERA may be needed to evaluate the potential for ecological impacts associated with that release. ERAs must be tailored to meet the technical requirements and challenges of the energy-development industry.

SEE ALSO THE FOLLOWING ARTICLES

Acid Deposition and Energy Use • Climate Change and Public Health: Emerging Infectious Diseases •

Crude Oil Spills, Environmental Impact of • Ecological Footprints and Energy • Ecosystem Health: Energy Indicators • Ecosystems and Energy: History and Overview • Hazardous Waste from Fossil Fuels • Nuclear Power: Risk Analysis • Risk Analysis Applied to Energy Systems

Further Reading

Suter, G. W. II. (1993). "Ecological Risk Assessment." Lewis, Boca Raton, FL.

Suter, G. W. II. (1997). "A framework for assessing ecological risks of petroleum-derived materials in soils, Oak Ridge National Laboratory Environmental Sciences Division Report ORNL/TM-13408." Oak Ridge National Laboratory, Oak Ridge, TN.

U.S. Environmental Protection Agency (1989). "The nature and extent of ecological risks at Superfund sites and RCRA facilities, Office of Policy Planning and Evaluation EPA-230-03-89-043." U.S. Environmental Protection Agency, Washington, DC.

U.S. Environmental Protection Agency (1989). "Summary of ecological risks, assessment methods, and risk management decisions in Superfund and RCRA facilities, Office of Policy Planning and Evaluation EPA-230-03-89-046." U.S. Environmental Protection Agency, Washington, DC.

U.S. Environmental Protection Agency (1992). "Framework for ecological risk assessment, Office of Research and Development/Risk Assessment Forum EPA/630/R-92/001." U.S. Environmental Protection Agency, Washington, DC.

U.S. Environmental Protection Agency (1993). "Wildlife Exposure Factors Handbook, EPA/600/R-31/187a." U.S. Environmental Protection Agency, Office of Research and Development, Washington, DC.

U.S. Environmental Protection Agency (1997). "Ecological risk assessment guidance for Superfund: Process for designing and conducting ecological risk assessments, Interim final, EPA 540/R-97/006." U.S. Environmental Protection Agency, Washington, DC.

U.S. Environmental Protection Agency (1998). "Guidelines for ecological risk assessment, 63FR26845, EPA/630/R-95/002F." U.S. Environmental Protection Agency, Washington, DC.

U.S. Environmental Protection Agency (1999). "Issuance of final guidance: Ecological risk assessment and risk management principles for Superfund sites, Office of Solid Waste and Emergency Response Directive 9285.7-28 P." U.S. Environmental Protection Agency, Washington, DC.

Economic Geography of Energy

BARRY D. SOLOMON
Michigan Technological University
Houghton, Michigan, United States

1. Eight Basic Problems
2. Fossil Fuels
3. Nuclear Power
4. Renewable Energy
5. International Comparisons of Energy Consumption

Glossary

energy reserves The portion of the total resource base that has been located with certainty that can be economically recovered and used with present technology. This quantity is not fixed.

energy resources The portion of the total resource base that has been estimated with at least a small degree of certainty, and which might be recovered in the future.

location theory Spatial theory developed to solve optimization problems for selecting a preferred zone or location for a facility, which involves trading off transportation costs with production and processing costs, production balances, and market delivery levels.

Organization of Petroleum Exporting Countries (OPEC) Members include Algeria, Indonesia, Iran, Iraq, Kuwait, Libya, Nigeria, Qatar, Saudi Arabia, United Arab Emirates, and Venezuela.

resource/production ratio The ratio of total oil or gas resources to the current production level; provides a rough estimate of the scarcity of fossil fuels in a country, region, or the world.

The economic geography of energy involves quantitative and qualitative study of the spatial patterns of energy production, transportation, and use at a variety of geographic scales, and analysis of the determinants of those patterns. It is a surprisingly diverse subject area, encompassing everything from locational and transportation models, to economic impact studies of energy development and use, to international, national, regional, and site-specific problems associated with particular energy sectors. Site-specific sector problems include transportation constraints associated with natural gas in various countries, the growing scarcity of oil and gas, emergency planning at nuclear power plants and radioactive waste disposal, environmental impacts, and factors influencing the international rate of diffusion of wind power technology. Given this diversity of subject matter, the discussion here provides a broad overview of the economic geography of energy.

1. EIGHT BASIC PROBLEMS

Eight broad research problems are representative of the rich range of energy subjects that are investigated in the field of economic geography: supply and demand location, fossil fuel scarcity, environmental impacts, nuclear proliferation, renewable energy development, electric utility deregulation, geopolitics, and technological changes and energy efficiency.

1.1 Separation between Supply and Demand Centers

From prehistory until the Industrial Revolution, most energy sources used by humans were localized (i.e., available within 5–10 miles of end users). Energy sources included draft animals, human slaves, and renewable sources such as biomass (wood and wastes), water mills, and wind power. Following the onset of the Industrial Revolution, with advancements in transportation technology and increased rates of deforestation in many regions, societies increasingly relied on long-distance shipments of coal, wood, and eventually oil, natural gas, and electric power. With increased separation between the location of energy resources and demand centers came a need to investigate the optimal transportation modes, production rates, inventories, and scheduling of energy production and delivery systems. Industrial location theory and mathematical programming techniques

 © 2004 Elsevier Inc. All rights reserved.

have been used by geographers to analyze the trade-off between energy development and production costs, and shipment costs to market centers to explain the siting pattern of energy facilities. A major factor determining the economic value of an energy source or carrier to its consumers is whether it can be economically stored or controlled for later use, which in turn is a function of the demand at a particular location. This interplays with other determinants of the desirability of a specific energy source or carrier, such as pecuniary cost, convenience of access and use, energy efficiency, steadiness of flow and renewability, safety, quality and other environmental properties, and end-use technology.

1.2 The Growing Scarcity of Fossil Fuels

Policy concerns have been raised about the scarcity of fossil fuels, especially oil and natural gas, since the 1970s. The important distinction between oil and gas reserves and resources leads to different prognoses in different parts of the world (see the Glossary). A useful indicator of scarcity is the resource/production ratio, which can be instructive to calculate for various nations. For instance, Indonesia had 7.4 billion barrels of oil resources in the ground in 2000 and produced 515 million barrels in 2001, leading to a ratio of 14.4, an indicator of the number of years of production left before its supply is exhausted, *ceteris paribus*. The same ratio for the United States in 2001 was 62.1. These ratios will not change with shifts in production technology or in the availability and cost of alternative energy sources.

1.3 Environmental Impact

The production and use of all energy sources results in undesirable environmental effects, which vary based on the health of the existing ecosystem, the size and health of the human population, energy production and consumption technology, and chemical properties of the energy source or conversion device. A shorthand equation for the environmental impact of energy production and use is $I = P \cdot A \cdot T$, where I is environmental impact, P is the size of the human population, A is the affluence of the population (e.g., per capita income), and T is technology (e.g., energy efficiency, emission rate of air and water pollution). Such impacts have been operationalized by geographers in case studies around the world. In some cases, regional econometric and input–output models have been used in these studies.

1.4 Nuclear Proliferation

Geographers researching the development of nuclear power have shifted emphasis from commercialization, cost, risk, public acceptance, and power plant siting in the 1950s through the early 1980s to reactor decommissioning and radioactive waste disposal since then. With nuclear power development on hold in most countries, attention has also been given to nuclear weapons facilities and weapons proliferation in an increasingly dangerous world. With the distinction, shaky at best, between civil and military nuclear programs made by the United States and United Kingdom under the Nonproliferation Treaty of 1970, nuclear programs or policies of nations with unofficial nuclear weapons programs (i.e., South Africa, Israel, Argentina, Brazil, Taiwan, and perhaps Iran, among others) will come under increased scrutiny. This is in addition to the eight or more nations with official nuclear weapons (and power) programs by 2003: the United States, Russia (and perhaps other former Soviet republics), United Kingdom, France, China, India, Pakistan, and North Korea.

1.5 Development of Renewable Energy Sources

Amid growing concern in the international scientific community concerning the scarcity of fossil fuels, greenhouse gas emissions, and global climate change, analytical interest in renewable energy development has grown. Because renewable energy sources are widely yet unevenly available, geographic analysis of development potential and policies is important. Biomass energies and hydroelectric power are the traditional sources of renewable energy; development policy has focused on wind, solar, geothermal, and hydrogen power technologies. Wind energy development in the United States has been slow but steady since the 1990s, whereas Germany became the world leader in installed capacity in 1997. Other nations with significant programs in renewable energy development include Spain (wind), Italy (geothermal), Denmark (wind), Iceland (hydrogen and geothermal), Israel (solar), Japan (solar and geothermal), Philippines (geothermal), India (wind and solar), Indonesia (geothermal), Brazil (biomass and hydrogen), Mexico (geothermal), and South Africa (solar).

1.6 Electric Utility Deregulation

Electric utility deregulation offers the great promise of market forces leading to lower electric rates, lower

air pollution, greater energy (and economic) efficiency, and perhaps greater use of renewable energy sources. Ideally, deregulation involves the restructuring of a previously monopolized or nationalized electric utility into separate generation, transmission, distribution, and marketing companies, and allowing wholesale and retail choice of generation company or power marketer. Deregulation has occurred to varying degrees since 1989 in the United Kingdom, Norway, Australia, New Zealand, Chile, Argentina, and about 20 states in the United States. There have been promising results in a few countries and in some U.S. states in some respects, especially lower rates and lower air pollution. In most cases, competitive markets have yet to be realized and lower rates can be attributed to other causes, such as previously planned amortization or retirement of expensive power plants, unexpected surplus in natural gas, rate caps, etc. In addition, deregulation has had only a slight beneficial effect on the use of renewable electricity sources. The promise of electric utility deregulation is thus unfulfilled and deserves further study.

1.7 Role of Geopolitics

Geopolitical considerations have played a major role in many energy policy decisions, e.g., in domestic debates over gasoline taxes, pipeline construction, radioactive waste disposal, and acid rain control legislation in the United States, and in petroleum-related violence in Nigeria. The most prominent role for geopolitics in energy policy has probably involved international discussions on controlling greenhouse gas emissions, and in oil markets. In the cases of the Kyoto Protocol of 1997 and the 1992 Framework Convention on Climate Change, nations carefully considered their national economic interests, domestic politics, and international trade during the negotiations. European countries, with the lowest rates of population and economic growth along with strong domestic environmental lobbies, have pursued a greater rate of greenhouse gas reduction. The United States, in contrast, has been stubbornly cautious and backed out of the treaty in 2001 (arguing it is not in its economic best interests), and the oil-rich nations of the Middle East have been least supportive of any emissions controls. In the case of oil markets, with the United States now dependent on imports for over half its supply, energy policy and trade strategy have played major roles in the pursuit of new oil discoveries in Alaska and in warfare in Kuwait, Iraq, and perhaps Afghanistan.

1.8 Technological Change and Energy Efficiency

Although technological change (usually involving an improvement in energy efficiency) is not inherently a geographic process, it does not occur uniformly over geographic space. Consequently, mathematical modeling and behavioral and innovation diffusion studies by geographers have improved our understanding of the pervasive phenomenon of technological change in energy markets. Similarly, creative government policies in Germany since 1990 have led to a faster rate of adoption of renewable energy sources, compared to the United States. Other studies have helped to explain why Japan, Italy, Germany, and Scandinavian countries have lower levels of energy consumption per capita or per unit of gross domestic product (GDP) than is the case in the United States and Canada. The answer to this question is a complex mix of price, government incentives and standards, economic structure, education, and lifestyle factors.

2. FOSSIL FUELS

2.1 Oil

The economic geography of the world oil industry has been studied in detail since the 1963 seminal study by Peter Odell. This study provided an analysis of the factors that contributed to the spatial distribution of the production, refining, transport, distribution, and consumption of oil. One interesting comparison is that although the share of proved crude oil reserves located in the Middle East at the time of the study was 60%, the share has risen only slightly since then, to about 65% by 2002. The top five nations of the world in terms of crude oil reserves are all in the Middle East, with the United States a distant eighth (Table I). Given these patterns, it is not surprising that the Organization of Petroleum Exporting Countries (OPEC) accounted for 39% of the world's crude oil production in 2002, with 26% coming out of the Persian Gulf. Saudi Arabia was the leading producer at 8.4 million barrels/day, followed by Russia and the United States. Before the rise of OPEC in the 1960s, the majority of global production was controlled by multinational oil companies. Although OPEC's production share peaked at 55% in 1973 and steadily declined following the OPEC oil embargo of that year to a low of 29% in 1985, it has gradually risen again since then to the current 39%

TABLE I

Proved Reserves of Fossil Fuels in Selected Countries in 2002[a]

Nation	Oil (10^9 barrels)	Gas (10^{12} m^3)	Coal (10^6 short tons)
Saudi Arabia	261.8	6.36	0
Iraq	112.5	3.11	0
United Arab Emirates	97.8	6.01	0
Kuwait	96.5	1.49	0
Iran	89.7	23.00	1885
Venezuela	77.8	4.19	479
Russia	60.0	47.57	157,010
United States	30.4	5.19	249,994
Libya	29.5	1.31	0
Mexico	12.6	0.25	1211
China	18.3	1.51	114,500
Qatar	15.2	14.40	0
India	5.4	0.76	84,396
Australia	3.5	2.55	82,090
Germany	0.3	0.34	66,000

[a]Data from the British Petroleum Company (2003) and the Energy Information Administration (2003).

level. Oil prices have been remarkably stable since 1974, except for the period from 1980 to 1986 following the Iranian Revolution and during the Iran–Iraq War.

The global patterns in crude oil refining, consumption, and trade illustrate how integrated the oil markets have become in recent decades. In the case of refining, the United States is by far the global leader. It refined more than 15 million barrels/day in 2001, accounting for 21.6% of the world total. Europe as a whole (not counting the European portion of the former Soviet Union) was not far behind at 21.2%, though capacity is split widely among Germany, Italy, France, the United Kingdom, Netherlands, Spain, and others. The Middle East accounted for 8% of the global refining throughput in 2001, led by Saudi Arabia and Iran. South America and Central America were next at 7.9%, led by Brazil and Venezuela. The next largest refiner was the former Soviet Union at 7% (mostly Russia), though its capacity has fallen by about 30% since 1991. Other leading oil refiners include China and Japan. Global patterns of oil consumption are more geographically concentrated than is refining, with the United States, Europe (except for the former Soviet Union), Russia, Japan, and China accounting for 64% of the total.

The global trade pattern in crude oil and refined petroleum products is show in Fig. 1. The large production coming out of the Middle East, which accounts for more than 40% of the total trade flow, is marketed to Japan, the Asia Pacific region, Europe, and the United States. The United States made a concerted effort after the 1973 OPEC embargo to be less dependent on the OPEC supply, but it actually increased its foreign crude oil and refined product import dependence over this period, going from 36% in 1973 to over 60% in 2002. The former Soviet Union is also a major oil trader, with most of its exports destined for refineries in western Europe.

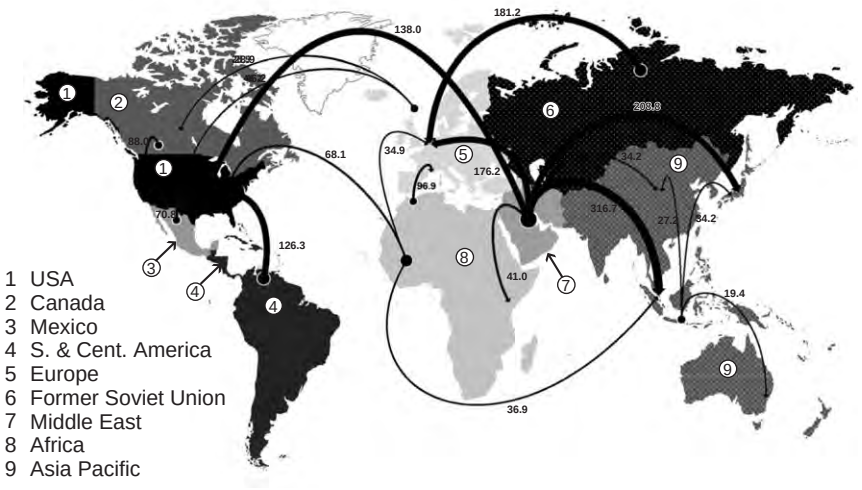


FIGURE 1 Major trade movements of oil (million tonnes). From the British Petroleum Company (2002), with permission.

Other significant oil shipment routes include Venezuela to the United States, Algeria and Libya to southern Europe, and Nigeria to the United States and the Asia Pacific region (Fig. 1).

Synthetic forms of oil supplies also can be developed. Although the vast technical potentials of tar sands and shale oil have been known for decades, these resources are beset by economic, environmental, and political problems and the industries are only in their infancy. Tar sands reserves, which are more properly known as bitumen or petroleum with very high viscosity, low resource quality, and low energy content, are concentrated in northern Alberta, Canada, where total production through surface and *in situ* mining amounted to just 30 million tonnes in 1999; an additional 3.3 million tonnes was produced in Venezuela. Canada has 979 million tonnes of proved recoverable reserves of synthetic crude oil or raw bitumen; Venezuela has 373 million tonnes. This compares with an additional 4231 million tonnes of additional reserves of foreseeable economic interest in the United States (half of the total is located in Utah and Alaska). The tar sands reserves of the United States and Canada are slightly larger than their oil reserves, whereas the tar sands reserves are much smaller than oil reserves in Venezuela.

The ultimate development potential of oil shales, if development proves to be economically and environmentally sound, is much greater than that of tar sands. The term "shale oil" is commonly used, although it is a misnomer because the organic material is chiefly kerogen and the hard rock it is usually found in is marl. This alternative oil resource can also be mined by conventional or *in situ* methods. The United States has by far the largest quantity of proved recoverable reserves of shale oil, at 60,000–80,000 million tonnes of oil equivalent. A small commercial facility was founded in Autun, France in 1839. Other nations that have produced shale oil include Canada, Scotland, Russia, China, Brazil (195 thousand tonnes in 1999), Estonia (151 thousand tonnes in 1999), and Australia (5 thousand tonnes in 1999, though production has grown since then).

2.2 Natural Gas

The spatial patterns of the world natural gas industry are somewhat different than those for oil, because gas has not been explored for and developed with the same vigor, is much less convenient to store, and is much more expensive to ship because of its low energy density (especially internationally). Geographic studies of this industry have concentrated on its production potential and have modeled its expanded development, transport via pipeline, and trade, especially in North America, Russia, and the rest of Europe.

The gas reserve base is geographically concentrated, with Russia, Iran, and Qatar accounting for 55% of the world total in 2002 (Table I). Other leading holders of gas reserves are Saudi Arabia, United Arab Emirates, the United States, Algeria, and Venezuela. The gas reserve base does not include coal bed methane, the reserve of which may be equal to or greater than that of conventional world gas in Canada and Russia alone. The United States and Russia are the only countries that account for more than 10% of the share of global gas production, each at about 22%. United States production totaled 555 billion m^3 in 2001; Russia produced 542 billion m^3. Canada was a distant third, at 172 billion m^3/yr. Although Russia used to dominate global output, it has alternated places with the United States as global leader since 1997.

Gas production figures, of course, reflect a nation's consumption level and involvement in international trade. The United States, the world's leading gas consumer, accounts for more than one-fourth of global consumption and is thus a slight importer, mostly from Canada (Fig. 2). Russia, conversely, only uses 15.5% of the world's gas output, and is a net exporter, to Europe and Turkey. European gas consumption, led by the United Kingdom, accounts for nearly 20% of global output. The total volume of liquefied natural gas (LNG) in world trade was 143 billion m^3 in 2001, or one-third the level of movement by pipeline in gaseous form. The leading shippers of LNG in recent years have been Indonesia, Algeria, Malaysia, and Qatar. About half of this LNG is sold to Japan, with another seventh of the total going to South Korea.

2.3 Coal

Economic geography studies of the coal industry began in the 1930s. The early work focused on the location of coal seams, local economic development, and transportation to markets in the United States and Europe. Geographic research since the 1970s has been more analytical and has often used mathematical programming models, because such studies have emphasized optimal coal development and shipment patterns, power plant siting, the response of the industry to technological change, growing environmental constraints, and electric utility deregulation. In the United States, e.g., coal mining has gradually

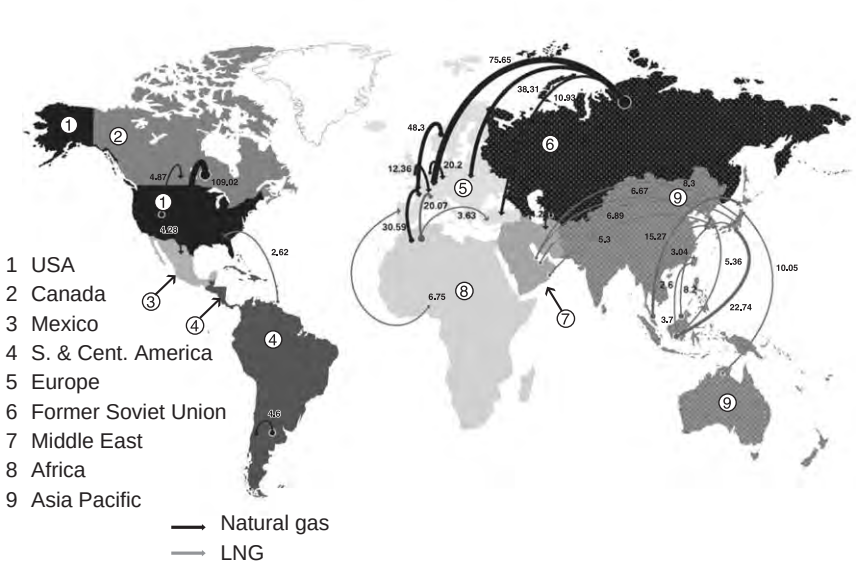


FIGURE 2 Major trade movements of natural gas (billion cubic meters). From the British Petroleum Company (2002), with permission.

shifted westward, because the economic advantages of that region's extensive near-surface deposits and lower sulfur content have increasingly outweighed the greater distances from electric power plants and other markets. European coal mining in the United Kingdom, Germany, Poland, and Russia, in contrast, has felt the greater brunt of acid rain and global warming restrictions, along with declining economics.

The coal reserve base is even more geographically concentrated than that of natural gas, with the United States, Russia, China, India, Australia, and Germany accounting for 77% of the world total in 2002 (Table I). Currently, only the United States, India, and Australia seem intent on using most of this resource base. Although Russian coal production has fallen by about half since the collapse of the former Soviet Union in 1991, the status of China is more uncertain. China led the world in coal production at 1545 million short tons in 1996, which fell to 1314 million short tons in 2000 as part of national policy to control air pollution and improve production efficiency (Germany implemented similar policy), but grew again in 2001 by more than 9%. Regarding global climate change, given the conflicting and unsettled policies of the United States, China, India, and Australia, the disposition of the huge coal resource base remains highly uncertain. The United States and China alone account for almost half of the world's coal consumption.

3. NUCLEAR POWER

The development of civil nuclear power was a stepchild of the military program of the United States to end World War II and of U.S. President Dwight D. Eisenhower's Atoms for Peace Speech to the United Nations General Assembly in 1953. The first commercial plants were commissioned in the United Kingdom, United States, and Russia in the late 1950s, and these nations actively developed this industry in the 1960s. Following the OPEC oil embargo in 1973, France and Japan developed major nuclear programs as well as strategies for energy autarky.

Significant political opposition to nuclear power plants and technology began to surface in the middle to late 1970s, first in the United States and later in several western European nations. Public opposition to nuclear power even spread to Taiwan in the 1990s. The opposition focused on the risk of accidents and radiological contamination, a prospect made worse by reactor sitings close to human population centers in New York, Boston, Philadelphia, Chicago, and coastal United Kingdom. In the aftermath of the accident at the Three Mile Island Unit 2 reactor near Harrisburg, Pennsylvania on March 28, 1979, it was clear that evacuation plans and procedures were inadequate in most nations (though they were improved thereafter).

The United States continued its global lead in nuclear power generation, with 104 nuclear reactor

units and 98.5 gigawatts (electrical; GWe) of installed capacity in 2003 (Table II). France leads large countries in reliance on nuclear power, which accounted for three-fourths of its electricity generation in 2002. Although the commitment to nuclear power and the location of the world's 441 nuclear reactors as of September 2003 (with 362 GWe of generating capacity) was widely distributed, most facilities outside of North America were located in Europe. Of the top 12 nuclear power users, 8 are European nations, with Japan and South Korea also making this list (Table II). The commitment to new reactor sitings and development, however, has stalled. Since 1995, opening of new nuclear power plants occurred only in France, the Czech Republic, Brazil, and especially in Asia (though a few others countries, such as the United States, have continued to expand their output as the capacity factor of many nuclear reactors has increased). There has been a *de facto* moratorium on new nuclear power in most of the world since 1980, with phaseout decisions pronounced in Sweden, Germany, Switzerland, Spain, Belgium, and the Netherlands.

In addition to heightened concern about nuclear proliferation and reactor safety in the 21st century, the nuclear power industry has been stifled by the twin Achilles' heels of reactor decommissioning and radioactive waste disposal. Although disposal sites have been in operation for shorter-lived, low- and medium-level wastes in several countries for decades, most attention has been given to the development of long-term, deep-geologic-disposal sites for high-level wastes (spent fuel rods and vitrified reprocessing wastes). Most progress on this front has been made in the United States and Scandinavia. The Finnish Parliament in May 2001 approved a plan in principle for deep-geologic disposal, though the repository would not be built until the 2010s and would not open until 2020. The disposal site is expected to be near the Olkiluoto nuclear plant in the town of Eurajoki on the west coast (a low-level waste site already exists there). In Sweden, repository siting is at an advanced stage and has been narrowed down to the Oskarshamn and Forsmark nuclear plant sites. A low- and medium-level waste site has existed near Forsmark since 1988 in offshore granite bedrock, 60 m below the Baltic Sea. In the United States, the Waste Isolation Pilot Plant in southeastern New Mexico opened on March 26, 1999 after a delay of 25 years; it accepts military transuranic and mixed hazardous wastes in bedded salt caverns 650 m underground. A high-level waste site in volcanic tuffs at Yucca Mountain, Nevada, 100 miles northwest of Las Vegas (also opposed and delayed), was approved by the U.S. Senate on July 9, 2002 and could open by 2010. Even so, scientific problems continue to be raised at the U.S. waste sites.

TABLE II

Top 12 Nuclear Power Producers in 2003[a]

Nation	Nuclear reactors	Capacity (GWe)[b]	Share of electricity (%)
United States	104	98.5	20
France	59	63.3	78
Japan	53	44.1	39
Germany	19	21.1	30
Russia	30	20.8	16
South Korea	18	14.9	39
United Kingdom	27	12.1	22
Canada	16	11.3	13
Ukraine	13	11.2	46
Sweden	11	9.5	46
Spain	9	7.4	26
Belgium	7	5.7	58

[a] Data from the World Nuclear Energy Association (2003); available on the Internet at http://www.uic.com.au/reactors.htm.
[b] GWe, gigawatts (electrical).

4. RENEWABLE ENERGY

The major renewable sources of energy include hydroelectricity, biomass, wind, solar, and geothermal. Renewable energy sources account for over one-fifth of total world energy use. Geothermal energy is categorized by federal statute in the United States as renewable, but strictly speaking, it is not, because it can be depleted at specific sites. All renewable energy resources can be used to generate electricity. In addition, biomass, solar, and geothermal energies are used for heating and cooking, and biomass is also used for transportation (as ethanol or methanol).

4.1 Hydroelectricity

Hydroelectric power is the largest and cheapest renewable electricity source in the world, although biomass easily surpasses it when other energy applications are considered. This is not to say that hydro power is without adverse environmental impacts, however. Large-scale dams are well known to lead to significant ecological damages, including loss of biodiversity, localized deforestation, soil

erosion, and disruption of free-flowing rivers, as well as displacement of large numbers of humans in many cases. Even so, hydroelectric power releases no direct greenhouse gas emissions and is a significant source of power in many parts of the world.

The United States, China, Canada, and Brazil are the leading producers of hydroelectricity (Table III). Brazil and Canada are the most dependent on hydro power, which accounted for 89 and 61%, of the electricity used in these two countries, respectively, in 2001, although persistent drought and rationing led to power cutbacks in Brazil in that year. Overall, Brazil relies on renewable energy for more than half of its total energy supply. Yet some small nations are even more dependent on renewable sources of electricity. Norway and Iceland, for instance, rely on renewable energy for 99% of their electricity consumption. In recent years, there have been few new large-scale hydroelectric power projects in developed nations because the best dam sites have generally already been tapped. A major exception to this pattern is the 2000-megawatt (MW) Lower Churchill Project on Gull Island in Newfoundland, Canada. In contrast, there have been several highly publicized dam removals and downstream ecological restoration projects in the northern United States. These trends, plus occasional drought, reduced the reliance of the United States on hydro power for electricity to 6% in 2001, down from a peak of 356 million MW-hr in 1997.

Large hydro projects can be found in many developing countries, such as China, India, Malaysia, Vietnam, Brazil, and Azerbaijan. China is building the controversial 18,200-MW Three Gorges Dam Project (the largest in the world) on the Yangtze River; the project is expected to reach full capacity by 2009. More than 1 million people will have to be relocated before the dam's reservoirs can be fully flooded. India, with 150-GWe of hydroelectric potential, has begun to focus on expanding this renewable source of energy, but not without difficulty in obtaining financing and in addressing protests from environmental and human rights groups. Most attention has focused on the Narmada Valley Development Project, which is planned to include up to 30 large dams in addition to numerous smaller ones.

TABLE III

Power Capacity From Hydro, Wind and Geothermal in Selected Countries, 2003[a]

	Gigawatts (electrical)		
Nation	Hydroelectric	Wind	Geothermal
United States	98.9	5.3	2.8
China	70.0	0.4	—[b]
Canada	67.2	0.2	0
Brazil	59.0	—[b]	0
Russia	43.4	—[b]	—[b]
Norway	27.4	—[b]	0
India	24.5	1.7	0
Japan	21.5	0.3	0.5
Italy	13.1	0.7	0.8
Spain	11.5	5.0	0
Mexico	9.6	0.3	0.9
Germany	3.0	12.8	0
Philippines	2.3	—[b]	1.9
Denmark	—[b]	2.8	0

[a] Data from the Energy Information Administration (EIA) (2003); the EIA, Renewable Energy Annual 2002, DOE/EIA-0603(2002), Washington, D.C., November 2002; the International Energy Agency, Wind Energy Annual Report 2003, Paris; and the International Geothermal Association, 2002, available on the Internet at http://iga.igg.cnr.it/electricitygeneration.php.

[b] Less than 100 MW.

4.2 Wind and Geothermal

Wind power has been the fastest growing energy source in the world since it became more cost-effective in the 1990s, though it has a long way to go to catch up with hydro power. Leading the way is Germany, supported by its 1990 Electricity Feed Law that requires electric utilities to pay 80–90% of the retail rate for wind power purchases. Germany passed the United States as world leader in wind power development in 1997, and now has almost 13 GWe in capacity (Table III). Other leading wind power producers include Spain, Denmark, and India, though Italy, the United Kingdom, Ireland, and several other European nations are rapidly expanding their wind power programs, including at many offshore sites. Denmark stands out as being most dependent on wind power, which accounts for 20% of that nation's electricity supply.

The two leading developers of geothermal energy are the United States and the Philippines, with most of their facilities being committed to electric power generation and much less to direct use (Table III). In the case of the Philippines, geothermal energy accounted for 21.5% of the nation's electricity supply in 2000. Other leading producers include Mexico, Italy, Indonesia, Japan, and New Zealand. Among these nations, the United States, Mexico,

Indonesia, and the Philippines have the largest expansion plans for geothermal energy facilities.

4.3 Biomass and Solar

Biomass energy in its various forms and uses (firewood, charcoal, crop residues, alcohol fuels, municipal and animal wastes, electricity) accounts for one-seventh of the world's primary energy supplies and over 90% in some developing nations. The United States is the leading electric power producer, with 10.1 GWe of capacity from biomass fuels (wood and wastes). Over half of this is cogenerated, mostly from the paper industry. This makes biomass the next largest source of renewable electricity in the United States after hydro power. The United States also has a significant (though subsidized) ethanol industry, which produced over 2 billion gallons in 2002, mostly from corn, with a refining capacity of over 3 billion gallons. Brazil has an even larger ethanol fuel program from sugarcane; production, which was subsidized, was between 3 and 3.5 billion gallons in recent years. Wood is a common source of biomass energy in Scandinavia, China, Pakistan, India, and sub-Saharan Africa. Unfortunately, much of the biomass energy data are incomplete or unreliable, rendering international comparison difficult. Consequently, most geographers working on this subject conduct case studies, especially of wood fuel use and sustainable development.

Solar energy has many applications. Most attention has been given to low-cost solar cookers, water and space heaters, and photovoltaic cells, though California also 384 MW of solar power plant capacity. The price of photovoltaic modules declined a hundred old from 1972 to 1992. The module price had fallen further to $3.40/W and $2.46/W by 2001 for low- and high-volume purchases, which makes them cost-effective in limited applications. Japan and Germany lead the market for photovoltaics, with over 400 and 250 MW, respectively. Japan is also the world's leading user of household solar thermal collectors. Israel is even more dependent on solar energy. Over 80% of Israeli households have solar water heaters (more than 1.3 million installations), and solar energy accounted for 3% of its total energy use in 2001.

5. INTERNATIONAL COMPARISONS OF ENERGY CONSUMPTION

International comparisons of energy consumption (per capita, per unit of GDP, etc.) are often used as indicators of energy efficiency and gluttony. On this basis, the United States, Canada, Russia, Japan, Australia, Saudi Arabia, and much of western Europe are the "haves," whereas China, India, Pakistan, Bangladesh, Indonesia, the Philippines, Vietnam, Brazil, and most of Africa are the "have nots" (Fig. 3). In some cases these differences in energy use per capita are more than 5:1. Such comparisons, although interesting, are flawed. Aggregate indicators of energy use per capita mask the real determinants of energy use. Decomposition of these patterns has

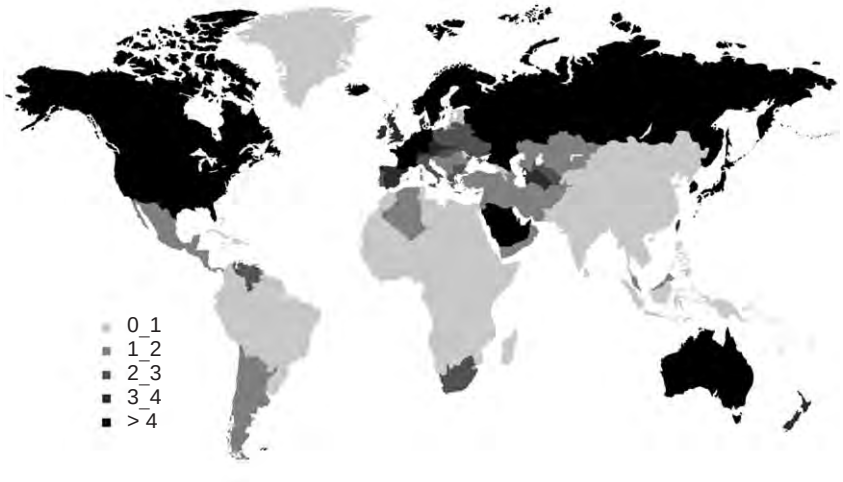


FIGURE 3 Energy consumption per capita by nation (tonnes of oil equivalent). From the British Petroleum Company (2002), with permission.

revealed the varying roles played by technological efficiency, structural economic change, and economic growth in determining energy consumption. Nonetheless, whatever the future energy trends and geographic patterns are in the world, such large inequities in energy use will not be ignored.

SEE ALSO THE FOLLOWING ARTICLES

Economics of Energy Demand • Economics of Energy Supply • Geographic Thought, History of Energy in • Global Energy Use: Status and Trends • Global Material Cycles and Energy • Natural Gas Resources, Global Distribution of • Oil and Natural Gas Liquids: Global Magnitude and Distribution

Further Reading

British Petroleum Company. (2003). "BP Statistical Review of World Energy." British Petroleum Corporate Communication Services, London.

Cleveland, C. J., and Kaufmann, R. K. (2003). Oil supply and oil politics: Deja Vu all over again. *Energy Policy* **31,** 485–489.

DiFrancesco, R. J., and Anderson, W. P. (1999). Developing Canada's Arctic oil resources. *Environ. Planning A* **31,** 459–476.

Elmes, G. A., and Harris, T. M. (1996). Industrial restructuring and the changing geography of the United States coal-energy system, 1972–1990. *Ann. Assoc. Am. Geogr.* **86**(No. 3), 507–529.

Energy Information Administration. (2003). International Energy Annual 2000. DOE/EIA-0219 (2001). U.S. Department of Energy, Washington, D.C.

Kuby, M. (1996). "Population growth, energy use, and pollution: Understanding the driving forces of global change." Commission on College Geography 2. Association of American Geographers, Washington, D.C.

Lakshmanan, T. R. (1983). A multiregional model of the economy, environment, and energy demand in the United States. *Econ. Geogr.* **59,** 296–320.

Luten, D. B. (1971). The economic geography of energy. *Sci. Am.* **225**(No. 3), 109–117.

Miller, E. W. (2002). Renewable energy's place in the global energy market. *In* "Renewable Energy: Trends and Prospects" (S .K. Majumdar, E. W. Miller, and A. I. Panah, eds.), pp. 1–18. Pennsylvania Academy of Science, Easton, Pennsylvania.

Odell, P. R. (1963). "An Economic Geography of Oil." Praeger, New York.

Schipper, L., Unander, F., Murtishaw, S., and Ting, M. (2001). Indicators of energy use and carbon emissions: explaining the energy economy link. *Annu. Rev. Energy Environ.* **26,** 49–81.

Smil, V. (1994). "Energy in World History." Westview Press, Boulder, Colorado.

Smil, V. (2001). Geography of energy use. *In* "Macmillan Encyclopedia of Energy" (J. Zumerchik, ed.), Vol. 2, pp. 567–572. Macmillan Reference, New York.

World Energy Council. (2003). WEC Survey of Energy Resources 2001. London. Available on the Internet at http://www.worldenergy.org/wec-geis/publications/reports/ser/. Last accessed January 2, 2003.

Economic Growth and Energy

DAVID I. STERN
Rensselaer Polytechnic Institute
Troy, New York, United States

1. **Introduction**
2. **Theory of Production and Growth**
3. **Factors Affecting Linkage between Energy and Growth**
4. **Empirical Testing**
5. **Environmental Implications**

Glossary

capital There are two meanings in economics. One is the monetary value of assets owned by an individual firm, etc. or the amount that is available to invest. The second meaning, used herein, refers to a human-made input into the production process, including things such as buildings and machinery, and infrastructure (telecommunications systems and the like). Capital in these forms is also described as reproducible capital, manufactured capital, or human-made capital. Some writers use the term "capital" more widely, to include human capital and/or environmental or natural capital. Human capital refers to expertise or know-how embodied in people through processes of education and training. The term "natural capital" has been used to refer to the stock of environmental resources.

demand elasticity The percentage response in the quantity of a good demanded for a 1% increase in its price.

diminishing returns If use of an input to production is increased, holding the amounts of the other inputs constant, the law of diminishing returns states that eventually the marginal product of an input will decline, though not necessarily becoming negative.

ecological economics A transdisciplinary field including ideas from economics, natural sciences, and other social sciences, in which scholars and practitioners are interested in the broad question of the use of resources by society and in understanding the nature, behavior, and dynamics of the economy–environment system. Many, but not all, ecological economists are highly critical of the mainstream (neoclassical) school of economics.

econometrics The specialized branch of mathematical statistics applied to the analysis of economic data.

economic growth An increase in economy-wide economic production, usually measured by an increase in gross domestic product (GDP); also, the process of the economy growing over time.

elasticity of substitution A measure of the responsiveness of the ratio of the quantities of two productive inputs to a change in the relative prices of the quantities. For example, if the price of energy increases 1% relative to the price of capital and the amount of energy used per dollar of capital decreases by 1%, then the elasticity of substitution is 1. If the decrease in energy use is less, then the elasticity of substitution is less than 1, and vice versa.

embodied energy Used directly in the production of goods and services, energy is also used indirectly through its involvement in producing the other inputs to production. The sum of the two uses is called the embodied energy of the good or service.

externality A negative externality, or external effect, occurs when a production or consumption activity has unintended damaging effects, such as pollution, on other firms or individuals, and no compensation is paid by the producer or consumer to the affected parties. A positive externality occurs when activities have beneficial effects for others who do not pay the generator of the effect, such as freely available research results.

institutions Commonly accepted rules, practices, etc. that govern human social and economic behavior, such as laws, customs, markets, etc.

marginal product If use of an input to production is increased, holding the amounts of the other inputs constant, the increase in output produced is called the marginal product of that input.

neoclassical economics The mainstream school or paradigm of economics characterized by mathematical models based on optimization by individual consumers and producers. A paradigm that arose in the late 19th century and emerged to its dominant position after the World War II; other paradigms include Keynesian, Marxist, Institutional, and neo-Ricardian, and post-Keynesian models.

nonrenewable resource A natural resource such as petroleum, which has zero natural growth over ordinary human timescales; though it may be replenished over geological time. Renewable resources such as forests

 © 2004 Elsevier Inc. All rights reserved.

grow or regenerate at rates that are relatively fast on human timescales.

production function A mathematical representation of the quantitative relationship between productive inputs and the level of output of some good or service, expressing the maximum quantity of output that can be obtained from any particular amounts of inputs.

productivity An index of economic output relative to input. Increases in productivity are often seen as being due to improvements in technology but may also be due to other factors. The most general productivity index is known as total factor productivity. Labor productivity is just output per worker hour and is the measure usually discussed in the financial media.

public goods The consumption of a public good by one person, which does not reduce the amount available for others and once provided to some people, other people cannot be excluded from consuming them.

substitution Shifting the mix of inputs used to produce a given quantity of output. Substitution occurs among already known production methods and does not imply that innovations are introduced.

sustainability As used herein, nondeclining individual consumption or utility over time; many other definitions exist.

technological change The invention and introduction of new methods of production as well as new products; the main focus in economics and herein is on the former.

utility The level of satisfaction (or happiness) of an individual. In a more aggregate context, we refer to the social welfare of a group of people, which is an indicator related to the utilities of the individuals in society.

Resource and ecological economists have criticized the theory that energy plays a minor role in economic growth on a number of grounds, especially the implications of thermodynamics for economic production and the long-term prospects of the economy. Although a fully worked out alternative model of the growth process does not seem to exist, extensive empirical work has examined the role of energy in the growth process. The principal finding is that energy used per unit of economic output has declined, but that this is to a large extent due to a shift in energy use from direct use of fossil fuels such as coal to the use of higher quality fuels, and especially electricity. When this shift in the composition of final energy use is accounted for, energy use and the level of economic activity are found to be tightly coupled. When these and other trends are taken into account, the prospects for further large reductions in the energy intensity of economic activity seem limited. These findings have important implications for environmental quality and economic and environmental policy.

1. INTRODUCTION

This article surveys the relation between energy and economic growth, and more generally the role of energy in economic production. Business and financial economists pay significant attention to the impact of oil and other energy prices on economic activity, but the mainstream theory of economic growth pays little or no attention to the role of energy or other natural resources in promoting or enabling economic growth. The extensive discussions concerning the "productivity slowdown" following the 1970s oil crises were an exception.

This article is structured to cover these key points in a systematic fashion. In the first section there is a review of the background theory of production and growth from different points of view—those based in economics and those based in the natural sciences. The starting premise is that gaining an understanding of the role of energy in economic growth cannot be achieved without first understanding the role of energy in production. The scientific basis of the role of energy in production, and hence also in the increasing scale of production involved in economic growth, is considered first. However, institutional phenomena also affect how this role plays out and therefore the economic view of growth and production and of the potential role of energy is necessarily more complex than just this scientific understanding. The mainstream theory of economic growth is, therefore, reviewed next. The limitations of its consideration of energy and other resource issues have been the subject of strong criticism grounded in the biophysical theory of the role of energy. A review of these alternative viewpoints completes this first section of the article.

The mainstream economics concept of the production function is next used to examine the key factors that could reduce or strengthen the linkage between energy use and economic activity over time. This production theory is very general and is less subject to criticism than are the specific models of economic growth. These key factors are (1) substitution between energy and other inputs within an existing technology, (2) technological change, (3) shifts in the composition of the energy input, and (4) shifts in the composition of economic output. Each of these themes will be discussed in detail.

Numerous ideas and views exist about the potential linkages between energy and economic growth. Choice between these theories has to be on the basis of both inherent plausibility and consistency and, perhaps more crucially, empirical

evidence. Therefore, studies that investigate the strength of the linkage between energy and growth are reviewed. To be useful, such studies must not be grounded in a single theory, potential mechanism, or school of thought. Therefore, the studies reviewed here do not specify structural linkages between energy and output. Because correlation and regression analysis does not imply causality from one variable to another, most of these studies employ the econometric notions of Granger causality and cointegration to test the presence of and direction of causality between the variables.

Finally, the implications of the theory for environmental quality are examined and the recent empirical literature on that topic is discussed.

2. *THEORY OF PRODUCTION AND GROWTH*

2.1 Energy in Production: Physical Theory and Economic Models

Reproducibility is a key concept in the economics of production. Some inputs to production are nonreproducible, whereas others can be manufactured at a cost within the economic production system. Primary factors of production are inputs that exist at the beginning of the period under consideration and are not directly used up in production (though they can be degraded and can be added to), whereas intermediate inputs are created during the production period under consideration and are used up entirely in production. Mainstream economists usually think of capital, labor, and land as the primary factors of production, and goods such fuels and materials as the intermediate inputs. The prices paid for all the different inputs are seen as eventually being payments to the owners of the primary inputs for the services provided directly or embodied in the produced intermediate inputs. In the theory of growth, this approach has led to a focus on the primary inputs, in particular on capital and land, and a much lesser and somewhat indirect treatment of the role of energy in the growth process. The primary energy inputs are stock resources such as oil deposits. But these are not given an explicit role in the standard growth theories, which focus on labor and capital. However, capital, labor, and, in the longer term, even natural resources are reproducible factors of production, whereas energy is a nonreproducible factor of production, although, of course, energy vectors (fuels) are reproducible factors. Therefore, natural scientists and some ecological economists have placed a very heavy emphasis on the role of energy and its availability in the economic production and growth processes.

The first law of thermodynamics (the conservation law) implies the mass-balance principle. In order to obtain a given material output, greater or equal quantities of matter must enter the production process as inputs, with the residual as a pollutant or waste product. Therefore, there are minimum material input requirements for any production process producing material outputs. The second law of thermodynamics (the efficiency law) implies that a minimum quantity of energy is required to carry out the transformation of matter. All production involves the transformation or movement of matter in some way, and all such transformations require energy. Therefore, there must be limits to the substitution of other factors of production for energy. Energy is also an essential factor of production. Though all economic processes require energy, some service activities may not require the direct processing of materials. However, this is only true at the micro level. At the macro level (economy-wide level), all economic processes require the indirect use of materials, either in the maintenance of labor or in the production of capital goods.

Some aspects of organized matter—that is, information—might also be considered to be nonreproducible inputs. Several analysts argue that information is a fundamentally nonreproducible factor of production in the same way as energy. Energy is necessary to extract information from the environment, but energy cannot be made active use of without information and possibly accumulated knowledge. Unlike energy, information and knowledge cannot be easily quantified. But these latter factors of production must be incorporated into machines, workers, and materials in order to be made useful. This provides a biophysical justification for treating capital, labor, etc. as factors of production. Though capital and labor are easier to measure than information and knowledge, their measurement is, still, very imperfect compared to that of energy.

In the mainstream neoclassical economics approach (discussed later), the quantity of energy available to the economy in any period is endogenous, though restricted by biophysical constraints (such as the pressure in oil reservoirs) and economic constraints (such as the amount of installed extraction, refining, and generating capacity), and the possible speeds and efficiencies with which these processes can proceed. Nevertheless, this analytical approach leads

to a downplaying of the role of energy as a driver of economic growth and production. Some alternative, biophysical models of the economy propose that energy is the only primary factor of production. This could be understood as there being a given stock of energy that is degraded (but, due to the law of the conservation of energy, not used up) in the process of providing services to the economy. But this means that the available energy in each period needs to be exogenously determined. In some biophysical models, geological constraints fix the rate of energy extraction. Capital and labor are treated as flows of capital consumption and labor services rather than as stocks. These flows are computed in terms of the embodied energy use associated with them and the entire value added in the economy is regarded as the rent accruing to the energy used in the economy. The actual distribution of the surplus depends on the relative bargaining power of the different social classes and foreign suppliers of fuel. Therefore, the owners of labor, capital, and land appropriate the energy surplus. The production process in the economy as a whole can be represented as an input–output model as originally developed by Nobel Laureate Wassily Leontief and adapted in the energy context by Bruce Hannon and others. The input–output model specifies the required quantities of each input needed to produce each output, with each output potentially an input to another process. It is also possible to trace all final production back to its use of a single primary factor of production, in this case energy.

These ecological economists argue that the energy used to produce intermediate resource inputs such as fuels increases as the quality of resources (such as oil reservoirs) declines. This is like negative productivity growth or technological change. Thus, changing resource quality is represented by changes in the embodied energy of the resources rather than by changes in the input–output coefficients. If resource stocks were explicitly represented, energy would no longer be the only primary factor of production. The neo-Ricardian models developed by Charles Perrings and Martin O'Connor, like all other neo-Ricardian models, have a fixed-proportions technology in terms of capital stocks instead of the flows in the Leontief model. They do not distinguish between primary and intermediate factors of production. Yet that approach can still take biophysical constraints such as mass balance and energy conservation into account.

If the economy can be represented as an input–output model where there is no substitution between factors of production, the embodied knowledge in factors of production can be ignored. An accurate accounting for all energy used to support final production is important. But in the real world the contribution of knowledge to production cannot be assumed to be proportional to its embodied energy. Though thermodynamics places constraints on substitution, the actual degree of substitutability among capital stocks embodying knowledge and energy is an empirical question. Neither the Leontief nor the neo-Ricardian model allows substitution between inputs. The neoclassical production models that are considered next, however, do.

2.2 The Mainstream Theory of Growth

As already explained, there is an inbuilt bias in mainstream production and growth theory to downplay the role of resources in the economy, though there is nothing inherent in economics that restricts the potential role of resources in the economy. The basic model of economic growth is the Nobel prize-winning work by Robert Solow that does not include resources at all. This model subsequently was extended with nonrenewable resources, renewable resources, and some waste assimilation services. These extended models, however, have only been applied in the context of debates about environmental sustainability, not in standard macroeconomic applications.

2.2.1 The Basic Growth Model

Economic growth models examine the evolution of a hypothetical economy over time as the quantities and/or the qualities of various inputs into the production process and the methods of using those inputs change. Here, the simplest model based on the work of Solow is described. In this model, a constant-sized labor force using manufactured capital produces output in terms of gross domestic product (GDP). The model assumes that output increases at a decreasing rate as the amount of capital employed rises. The uppermost curve in Fig. 1 shows this relationship between output (Y) and capital (K). This is termed diminishing returns to capital. It is assumed that a constant proportion, s, of the output is saved and invested in the capital stock. A constant proportion of the existing capital stock depreciates (becomes productively useless) in each period of time. The capital stock is in equilibrium (and so unchanging in size) when saving equals depreciation. This is also shown in Fig. 1. The dynamics implied by Fig. 1 are very simple. To the left of K^*, where capital per worker is scarce, capital investment generates a relatively large increase in future income, and so will

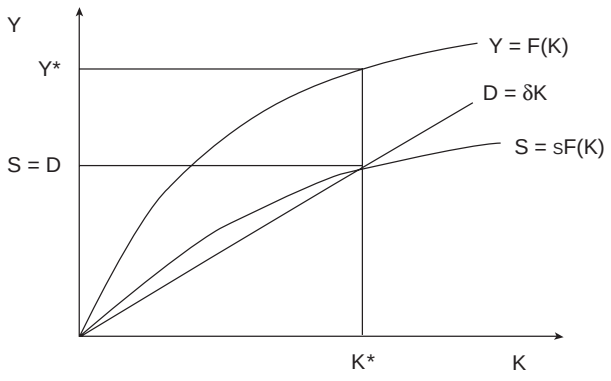


FIGURE 1 The neoclassical growth model.

offer a high rate of return. Moreover, it is clear from the relative positions of the *S* and *D* curves to the left of K^* that the addition to the capital stock (*S*) is greater than depreciation (*D*) and so capital rises. When the capital stock has reached K^*, it will be at a stationary, or equilibrium, state. Additions to capital due to saving are exactly offset by reductions in capital from depreciation.

This simple economy must sooner or later reach a stationary state in which there is no net (additional) investment and economic growth must eventually halt. In a transition process, while a country is moving toward this stationary state, growth can and will occur. An underdeveloped economy, with a small capital stock per worker, can achieve fast growth while it is building up its capital stock. But all economies will eventually settle into a zero growth equilibrium if the savings rate remains constant. No country can grow in perpetuity merely by accumulating capital, due to the diminishing returns to capital accumulation. If the savings rate is increased, growth will occur for a while until a new equilibrium is reached. Of course, the higher the savings the rate, the lower the current standard of living of the population.

If the labor force grows at a fixed rate over time, the total capital stock and the total quantity of output will rise, but capital per worker and output per worker will remain constant once an economy has developed to its equilibrium. The only necessary adjustment to Fig. 1 is that all units are now measured in per capita terms.

According to neoclassical growth theory, the only cause of continuing economic growth is technological progress. As the level of technological knowledge rises, the functional relationship between productive inputs and output changes. Greater quantities or better qualities of output can be produced from the same quantity of inputs. In the simple model being examined, technological progress continually shifts the output function upward, and so raises the equilibrium per capita capital stock and output levels. Intuitively, increases in the state of technological knowledge raise the rate of return to capital, thereby offsetting the diminishing returns to capital that would otherwise apply a brake to growth.

2.2.2 Endogenous Technological Change

The simple model just described does not explain how improvements in technology come about. They are just assumed to happen exogenously, so that this model is said to have exogenous technological change. More recent models attempt to endogenize technological change, explaining technological progress within the growth model as the outcome of decisions taken by firms and individuals. In endogenous growth models, the relationship between capital and output can be written in the form $Y = AK$. Capital, *K*, is defined more broadly than in the neoclassical model. It is a composite of manufactured and knowledge-based capital. Endogenous growth theorists have been able to show that, under reasonable assumptions, the *A* term in the preceding expression is a constant, and so growth can continue indefinitely as capital is accumulated.

The key point is that technological knowledge can be thought of as a form of capital. It is accumulated through research and development (R&D) and other knowledge-creating processes. Technological knowledge has two special properties. First, it is a public good: the stock of this form of capital is not depleted with use. This is important because it implies that the knowledge stock can be stored over time, even when it is being used. Second, it generates positive externalities in production: although the firm doing R&D obtains benefits from the knowledge acquired, others benefit too—the benefits that the firm accrues when it learns and innovates are only partly appropriated by itself. There are beneficial spillovers to the economy from the R&D process so that the social benefits of innovation exceed the private benefits to the original innovator. These externalities create momentum in the growth process. As firms install new capital, this tends to be associated with process and product innovations. The incentive to devote resources to innovation comes from the prospect of temporary monopoly profits for successful innovations. The growth of *K* thus means the growth of a composite stock of capital and disembodied technological knowledge. Therefore, output is able to rise as a constant proportion (*A*) of the

composite capital stock, and is not subject to the diminishing returns shown in Fig. 1.

So, in an endogenous growth model, the economy can sustain a constant growth rate in which the diminishing returns to manufactured capital are exactly offset by the technological growth external effect just described. The growth rate is permanently influenced by the savings rate; a higher savings rate increases the economy's growth rate, not merely its equilibrium level of income.

2.2.3 Growth Models with Natural Resources

The growth models examined so far do not include any natural resources. All natural resources exist in finite stocks or flows, though some, such as sunlight or deuterium, are available in very large quantities. Some environmental resources are nonrenewable, and many renewable resources are potentially exhaustible. Finiteness and exhaustibility of resources make the notion of indefinite economic growth problematic. Even maintaining the current level of output indefinitely—so-called sustainable development—may be difficult.

When there is more than one input (both capital and natural resources), there are many alternative paths that economic growth can take. The path taken is determined by the institutional arrangements that are assumed to exist. Analysts have looked at both optimal growth models that attempt to either maximize the sum of discounted social welfare over some relevant time horizon (often an infinite horizon) or achieve sustainability (nondeclining social welfare) and models intended to represent real economies with perfectly competitive markets or other arrangements.

The neoclassical literature on growth and resources centers on what conditions permit continuing growth, or at least nondeclining consumption or utility. Technical and institutional conditions determine whether such sustainability is possible. Technical conditions refer to things such as the mix of renewable and nonrenewable resources, the initial endowments of capital and natural resources, and the ease of substitution among inputs. The institutional setting includes things such as market structure (competition versus central planning), the system of property rights (private versus common property), and the system of values regarding the welfare of future generations.

In a 1974 paper published in a special issue of *Review of Economic Studies*, Solow showed that sustainability was achievable in a model with a finite and nonrenewable natural resource with no extraction costs and nondepreciating capital, which was produced using capital and the natural resource. However, the same model economy under competition results in exhaustion of the resource and consumption and social welfare eventually falling to zero. In another paper in the same special journal issue, Dasgupta and Heal showed that with any constant discount rate the so-called optimal growth path also leads to eventual depletion of the natural resource and the collapse of the economy. A common interpretation of this theory is that substitution and technical change can effectively decouple economic growth from resources and environmental services. Depleted resources or degraded environmental services can be replaced by more abundant substitutes, or by "equivalent" forms of human-made capital (people, machines, factories, etc.). But this is a misinterpretation. Neoclassical economists are primarily interested in what institutional arrangements, and not what technical arrangements, will lead to sustainability, so that they typically assume a *priori* that sustainability is technically feasible, and then investigate what institutional arrangements might lead to sustainability if it is technically feasible. However, there is a tendency among mainstream economists to assume that sustainability is technically feasible unless proved otherwise.

The elasticity of substitution (σ) between capital (factories, machines, etc.) and inputs from the environment (natural resources, waste assimilation, ecosystem services) is a critical technical term that indicates by how much one of the inputs must be increased to maintain the same level of production, when the use of the other input is reduced. A large value of σ implies that this substitution is easy, and vice versa. Figure 2 shows the different combinations of two inputs that can produce a given level of output for different values of σ. Different levels of output were chosen for the three values of σ to make the figure clearer.

The marginal product is the additional contribution to production of using one more unit of an input, holding the levels of the other inputs constant. A unitary elasticity of substitution ($\sigma = 1$), referred to as "perfect substitutability," means that as the ratio of the two inputs is changed by a given percentage, holding output constant; the ratio of their marginal products changes by the same percentage (in the opposite direction). This relation is shown by the curve (known as an isoquant) in Fig. 2, which is asymptotic to both axes. As resource use decreases toward zero, production can be maintained by increasing capital use toward infinity. Additionally,

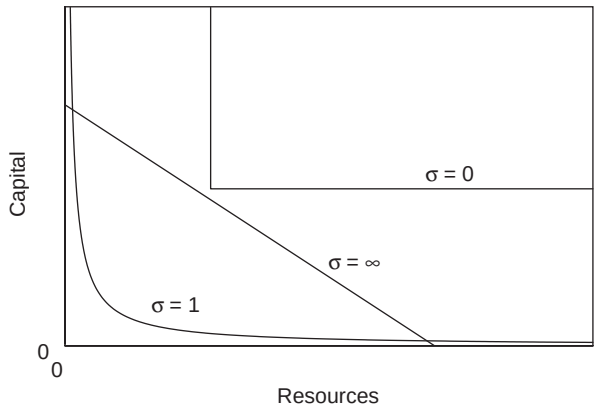


FIGURE 2 The elasticity of substitution.

the total cost of production is constant along the isoquant. Perfect substitutability does not mean that resources and capital are equivalently useful—that case is termed "infinite substitutability" ($\sigma = \infty$). In the latter case, producers see no difference between the two inputs and use the cheapest one. Whenever the elasticity of substitution is greater than 1, the isoquants cross the axes and inputs are nonessential for production. The figure also illustrates the case where no substitution is possible ($\sigma = 0$). In this case, the two inputs must be used in a fixed ratio. As is discussed later, perfect substitutability is an unrealistic assumption from a biophysical perspective, at least if it is assumed to apply at all possible ratios of capital and resources. Demand elasticities for energy, which in theory are related to the elasticity of substitution, also indicate that the elasticities of substitution between energy and other inputs and among different fuels may be between 0 and 1.

Economists such as Solow explicitly dispose of cases where σ for nonrenewable resources and capital is greater or less than unity. In the former case, substitution possibilities are large and therefore the possibility of nonsustainability is not an issue. In the latter case, sustainability is not feasible if an economy uses only nonrenewable resources. Of course, where there are renewable resources, sustainability is technically feasible, at least in the absence of population growth. Neoclassical economists argue that the class of growth models that includes resources can account for mass balance and thermodynamic constraints with the "essentiality condition." If σ is greater than 1, then resources are "nonessential." If σ is less than or equal to 1, than resources are "essential." Essential in this case means that, given positive nonresource inputs, output is only zero when the resource input is zero, and strictly positive otherwise. The Cobb–Douglas production function, a form frequently used in growth models, has the essentiality condition. Economists argue that this at least accounts for the fact that some amount of energy and materials is required to produce goods and services. But when the elasticity of substitution is unity, this "essential" amount can be infinitesimal if sufficient manufactured capital is applied. Economists also note that resources and capital are interdependent in the neoclassical models in that some positive quantity of resources is required to produce capital assets. Thus, the capital stock cannot be increased without depleting the resource stock.

Substitution that is technically possible will not occur unless society invests in sufficient capital over time to replace the depleted natural resources and ecosystem services. How much investment does take place depends on the institutional setting of the economy. For example, in an economy in which sustainability is just technically feasible ($\sigma = 1$) and there are only nonrenewable resources, sustainability will not occur in either a competitive or centrally planned economy where the decision rule is the maximization of the discounted flow of utility of future generations using a constant and positive discount rate. Consumption per capita will eventually decline to zero after an initial period of economic growth because resources and ecosystem services are depleted faster than capital can be accumulated to replace them. However, if the utility of individuals is given equal weight without regard to when they happen to live, and the aim is to maximize the sum of utilities over time, then growth in consumption can occur indefinitely. This is equivalent to maximizing net present value with a zero discount rate. The Hartwick rule is an investment rule that can, in theory, achieve sustainability if sustainability is technically feasible. A constant level of consumption can be achieved by reinvesting resource rents in other forms of capital, which, in turn, can substitute for resources.

The other key factor permitting growth in the face of a limited resource base is technological change. A technological improvement is defined as a gain in total factor productivity, which implies that output increases while a weighted sum of the inputs to production is held constant. Therefore, growth would be possible in the face of finite resources. Studies that examine the roles of resources in growth models with endogenous technological change have so far been less general in their assumptions, compared to research using the exogenous technological change

assumption, and do not yet provide necessary or sufficient conditions for the achievement of sustainability. For example, Sjak Smulders and Michiel de Nooij assume in their model that energy use per capita continues to grow over time and that conservation policies consist of either a reduction in the growth rate or a one-time cut in the level of energy use. This set of assumptions is very different from those in the models just discussed. The results depend on whether the growth rate of energy use is reduced or the level of energy use is cut and on the value of the elasticity of substitution. In these models, energy scarcity or conservation policies can induce technological change. Though this can mitigate the effects of scarcity, non-energy-related research and development may be crowded out. This means that attempts to conserve energy can depress the long-run economic growth rate, even when they would not in a model with exogenous technological change. Integrating these ideas into research on sustainability policy will clearly be a future area of active research.

2.3 Critique and Alternative Views

Many ecological economists have a "preanalytic vision" of the economic process that is fundamentally different from that presented in neoclassical economics. Mainstream growth theory focuses on institutional limits to growth. Ecological economists tend instead to focus on the material basis of the economy. The criticism of growth theory focuses on the limits to substitution and the limits to technological progress as ways of mitigating the scarcity of resources. Substitution of manufactured capital for resources and technological change could potentially get more production out of a limited resource input and circumvent the limited capacity of natural environments to absorb the impacts of energy and resource use by reducing those impacts. But if these two processes are limited, then limited resources or excessive environmental impacts may restrict growth (empirical evidence on these questions is reviewed in Sections 4 and 5).

2.3.1 Limits to Substitution

There is more than one type of substitution between inputs and, therefore, there is more than one reason why substitution may be limited. There can be substitution within a category of similar production inputs (for example, between different fuels) and between different categories of inputs (for example, between energy and machines). There is also a distinction to be made between substitution at the micro level (for example, in a single engineering process or in a single firm) and at the macro level (in the economy as a whole). Additionally, some types of substitution that are possible in a single country are not possible globally.

The first type of substitution (within category), and, in particular, the substitution of renewable for nonrenewable resources, is undoubtedly important. The long-run pattern of energy use in industrial economies has been dominated by substitutions: from wood to coal, to oil, to natural gas, and to primary electricity. It is possible that the elasticity of substitution for within-category types of substitution exceeds unity. This would imply that some particular inputs are nonessential.

Ecological economists emphasize the importance of limits to the other type of substitution, and, in particular, the substitution of manufactured capital for natural capital. Natural capital is needed both for energy capture and for absorbing the impacts of energy and resource use—the sink function. Even if smaller amounts of energy were needed for production, all productive activity, being a work process transforming the materials using energy, will disrupt the natural environment. Often one form of environmental disruption (for example, pollution) is replaced by another form of environmental disruption (for example, hydroelectric dams). Potential reasons for limited substitutability are discussed in the following section.

2.3.1.1 Thermodynamic Limits to Substitution

Thermodynamic limits to substitution are easily identified for individual processes by an energy–materials analysis that defines both the fundamental limitations of transforming materials into different thermodynamic states and the use of energy to achieve that transformation. These types of analyses have shown where technological improvements exhibit strong diminishing returns due to thermodynamic limits, and where there is substantial room for improvements in the efficiency of energy and material use. For example, the thermal efficiency of power plants has been relatively constant for many years, reflecting the fact that power plant design is approaching the thermodynamic limit.

2.3.1.2 Complementarity Limits Substitution

Production is a work process that uses energy to transform materials into goods and services. Nicholas Georgescu-Roegen's fund-flow model describes production as a transformation process in which a flow of materials, energy, and information is transformed

by two agents of transformation: human labor and manufactured capital. The flow of energy, materials, and services from natural capital is what is being transformed, while manufactured capital effects the transformation. Thus, some ecological economists argue that, for example, adding to the stock of pulp mills does not produce an increase in pulp unless there also is the wood fiber to feed them. The latter is essential an argument about material balance.

Mainstream economists think about this question differently. First, they argue that though additional capital cannot conjure wood fibers out of a vacuum, more capital can be used with each amount of wood fibers to produce more sophisticated and valuable products from them and that this is the relevant substitution between capital and resources. In the energy industries, more capital can extract more oil from a petroleum reservoir and extract more useful work downstream in cleaner ways, only subject to thermodynamic limits. Even thermodynamic limits apply only to production of physical product. There is no limit in this view to the potential value of product created through sophisticated manipulation using larger amounts of capital.

2.3.1.3 Physical Interdependence and Macroeconomic Limits to Substitution The construction, operation, and maintenance of tools, machines, and factories require a flow of materials, i.e., energy from natural capital. Similarly, the humans that direct manufactured capital consume energy and materials (i.e., food and water). Thus, producing more of the "substitute," i.e., manufactured capital, requires more of the thing for which it is supposed to substitute. Ecological economists argue that production functions used in growth models do not account for this interdependence, and thus assume a degree of substitutability that does not exist. But both environmental and ecological economics have erred by not distinguishing among the micro- and macro-applications of production functions. Substitution is fundamentally more constrained at the macro level of analysis than at the micro level. For example, home insulation directly substitutes for heating fuel within the household sector. But interdependence means that insulation requires fuel to manufacture, so, for the economy as a whole, the net substitution of insulation for fuel is less than that indicated by an analysis of the household sector, in isolation from the rest of the economy.

In Fig. 3, the curve $E = f(M)$ is a neoclassical isoquant for a constant level of output, where E is energy and M is materials. The indirect energy costs of materials are represented by $g(M)$. For simplicity, the diagram unrealistically assumes that no materials are required in the extraction or capture of energy. Addition of direct and indirect energy costs results in the "net" isoquant $E = h(M)$. Generalizing for material costs to energy extraction appears to indicate that there are eventually decreasing returns to all factors at the macro level, and therefore the socially efficient region of the aggregate production function does not include areas with extreme factor ratios.

At a global level, a country such as Kuwait or Nauru can deplete its natural resources and invest in manufactured capital offshore through the financial markets. But this route to substituting manufactured capital for natural capital is clearly not possible for the world as a whole.

2.3.1.4 Critical Natural Capital Limits Substitution Ecological economists have also argued that at the macro level some forms of "natural capital" are not replaceable by produced capital, at least beyond certain minimum stock sizes. These stocks may provide life-support services to the economy or represent pools of irreplaceable genetic information or "biodiversity." The limited substitutability argument has also been extended to incorporate nonlinear dynamics and irreversible changes. The fear is that excessive substitution of human-made capitals for natural capital will cause the system to approach a threshold beyond which natural systems will lose resilience and suffer catastrophic collapse. Though we cannot demonstrate these forms of nonsubstitutability from basic physical laws, they may be just as important as thermodynamics in constraining actual production functions. In the energy context, this argument is most relevant regarding the sink func-

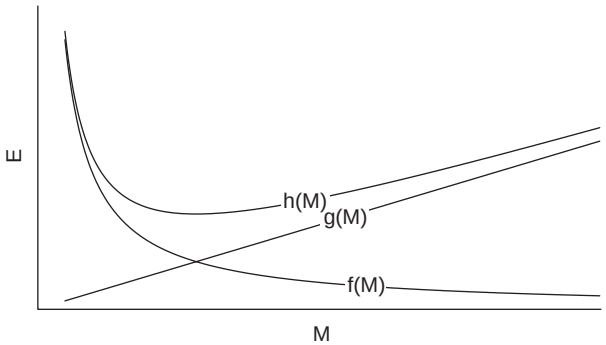


FIGURE 3 Macro-level limits to substitution.

tion of the environment. Using more and more of the environment as a sink for pollution means less and less of it is available for other life-support uses. Compared to current fossil fuel technologies, alternative energy sources may require larger areas of the environment for energy capture and may also generate wastes, etc. in the production of the energy capture and transmission capitals.

2.3.2 Limits to Technological Change

Even if substitution possibilities are limited, sustainability is possible if technological change is natural capital augmenting and unlimited in scope. The arguments for technological change as a solution would be more convincing if technological change were really something different from substitution. The neoclassical approach assumes that an infinite number of efficient techniques coexist at any one point in time. Substitution occurs among these techniques. Changes in technology occur when new, more efficient techniques are developed. However, in a sense, these new techniques represent the substitution of knowledge for the other factors of production. The knowledge is embodied in improved capital goods and more skilled workers and managers, all of which require energy, materials, and ecosystem services to produce and maintain. Thus, however sophisticated the workers and machinery become, there are still thermodynamic restrictions on the extent to which energy and material flows can be reduced.

Another question is whether technology will follow the "right" direction. If natural resources are not priced correctly due to market failure—a common and pervasive phenomenon that is the main topic of study of mainstream environmental economics—then there will be insufficient incentives to develop technologies that reduce resource and energy use. Instead, technological change would result in more resource use, not less.

3. FACTORS AFFECTING LINKAGE BETWEEN ENERGY AND GROWTH

There has been extensive debate concerning the trend in energy efficiency in the developed economies, especially since the two oil price shocks of the 1970s. Taking the example of the U.S. economy, energy consumption hardly changed in the period 1973 to 1991 (Fig. 4). This was despite a significant increase in GDP. These facts are indisputable. The reasons for the break in the trend have been the subject of argument. It is commonly asserted that there has been a decoupling of economic output and resources, which implies that the limits to growth are no longer as restricting as in the past.

The discussion here starts from the neoclassical perspective of the production function to examine the factors that could reduce or strengthen the linkage between energy use and economic activity over time. A general production function can be represented as follows:

$$(Q_1, \ldots, Q_m)' = f(A, X_1, \ldots, X_n, E_1, \ldots, E_p), \quad (1)$$

where the Q_i are various outputs (such as manufactured goods and services), the X_j are various inputs (such as capital, labor, etc.), the E_k are different energy inputs (such as coal, oil, etc.), and A is the state of technology as defined by the total factor productivity indicator. The relationship between energy and an aggregate of output such as gross domestic product can then be affected by substitution between energy and other inputs, technological change (a change in A), shifts in the composition of the energy input, and shifts in the composition of output. Also, shifts in the mix of the other inputs—for example, to a more capital-intensive economy from a more labor-intensive economy—can affect the relationship between energy and output, but this issue has not been extensively discussed in the literature and so will not be pursued further here. It is also possible for the input variables to affect total factor productivity, though in models that invoke exogenous technological change, this is assumed not to occur.

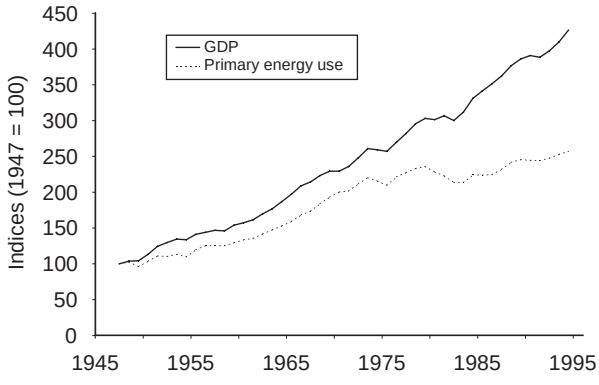


FIGURE 4 United States gross domestic product (GDP) and total primary energy use. GDP is in constant dollars, i.e., adjusted for inflation. Both variables are indexed to 100 in 1947.

3.1 Energy and Capital: Substitution and Complementarity

Econometric studies employing the translog and other functional forms have come to varying conclusions regarding whether capital and energy are complements or substitutes. These studies all estimate elasticities at the industry level. It seems that capital and energy are at best weak substitutes and possibly are complements. The degree of complementarity likely varies across industries, the level of aggregation considered, and the time frame considered.

There are few studies that look at macroeconomic substitution possibilities. In a 1991 paper, Robert Kaufmann and Irisita Azary-Lee demonstrated the importance of accounting for the physical interdependency between manufactured and natural capital. They used a standard production function to account for the indirect energy used elsewhere in the economy to produce the capital substituted for fuel in the U.S. forest products sector. They found that from 1958 to 1984, the indirect energy costs of capital offset a significant fraction of the direct fuel savings. In some years, the indirect energy costs of capital are greater than the direct fuel savings. The results of Kaufmann and Azary-Lee's analysis are consistent with the arguments made previously that scale is critical in assessing substitution possibilities. In this case, the assessment of substitution at one scale (the individual sector) overestimates the energy savings at a larger scale (the entire economy).

3.2 Innovation and Energy Efficiency

Substitution between different techniques of production using different ratios of inputs occurs due to changes in the relative prices of the inputs. Changes in the energy/GDP ratio that are not related to changes in the relative price of energy are called changes in the autonomous energy efficiency index. These could be due to any of the determinants of the relationship between energy and output listed at the beginning of this section, not just technological change. Actually, when there is endogenous technological change, changes in prices induce technological changes. In reality, technological change is at least in part endogenous. As a result, an increase in energy prices does tend to accelerate the development of energy-saving technologies. Periods of decreasing energy prices may result in technological development tending toward more intense use of energy.

Estimates of the trend in autonomous energy efficiency or the related energy augmentation index are mixed. This is likely because the direction of change has not been constant and varies across different sectors of the economy. Jorgenson and Wilcoxen estimated that autonomous energy efficiency in the United States is declining. Jorgenson also found that technological change tended to result in an increase in the share of production costs allocated to energy use—so-called biased technological change. Berndt *et al.* used a model in which the energy augmentation index is assumed to change at a constant rate. They estimate that in the U.S. manufacturing industry between 1965 and 1987, the energy augmentation index was increasing between 1.75 and 13.09% per annum, depending on the assumptions made. Judson *et al.* estimate separate environmental Kuznets curve relations for energy consumption in each of a number of energy-consuming sectors for a large panel of data using spline regression. The sectors are industry and construction, transportation, households and others, energy sector, nonenergy uses, and total apparent consumption, as well as households and agriculture, which are subsets of households and others. Judson *et al.* estimate time effects that show rising energy consumption over time in the household and other sectors but flat to declining time effects in industry and construction. Technical innovations tend to introduce more energy-using appliances to households and energy-saving techniques to industry.

Figure 5 presents the author's estimate of the autonomous energy efficiency index using the Kalman filter, an advanced time series technique that can extract unobserved stochastic series from time series data given a theoretical model that defines the trend of interest. This approach is more sophisticated than that of the studies cited previously, which assume that the index follows a simple linear or quadratic time path. Relative to an overall upward trend, the index shows large fluctuations. Until the mid-1960s, autonomous energy efficiency is increasing and then it starts a sharp decline. However, the results show that the first oil shock in 1973 does not disrupt the overall downward trend in energy efficiency. Only after the second oil shock does the trend reverse and energy efficiency increase.

The Khazzoom–Brookes postulate argues that energy-saving innovations can end up causing even more energy to be used, because the money saved is spent on other goods and services that require energy in their production. Energy services are demanded by the producer or consumer and are produced using

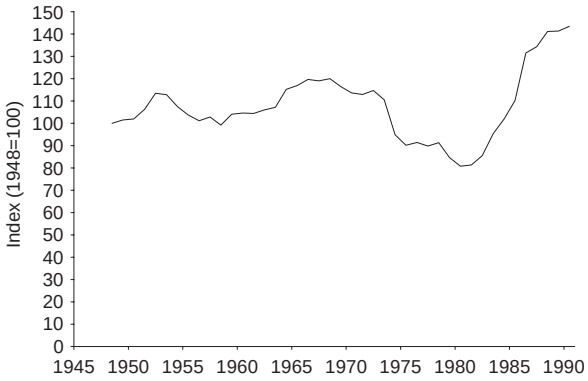


FIGURE 5 United States autonomous energy efficiency index.

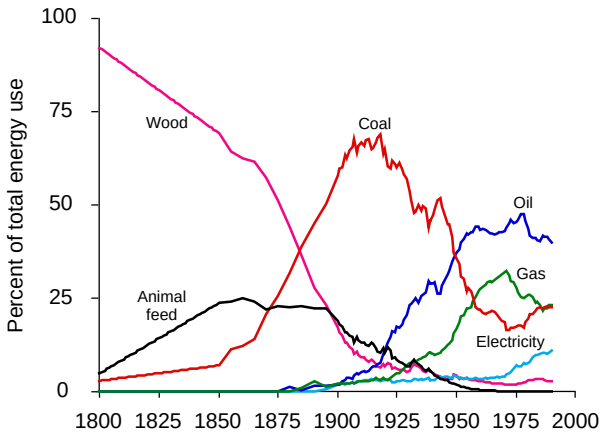


FIGURE 6 Composition of U.S. energy input.

energy. An innovation that reduces the amount of energy required to produce a unit of energy services lowers the effective price of energy services. This results in an increase in demand for energy services and therefore for energy. The lower price of energy also results in an income effect that increases demand for all goods in the economy and therefore for the energy required to produce them. There may also be adjustments in capital stocks that result in an even further increased long-run demand response for energy. This adjustment in capital stocks is termed a "macroeconomic feedback." In a 1997 paper, Richard Howarth argued persuasively that the rebound effect is less than the initial innovation-induced reduction in energy use, so improvements in energy efficiency do, in fact, reduce total energy demand.

3.3 Energy Quality and Shifts in Composition of Energy Input

Energy quality is the relative economic usefulness per heat equivalent unit of different fuels and electricity. One way of measuring energy quality is the marginal product of the fuel, which is the marginal increase in the quantity of a good or service produced by the use of one additional heat unit of fuel. These services also include services received directly from energy by consumers. Some fuels can be used for a larger number of activities and/or for more valuable activities. For example, coal cannot be used directly to power a computer whereas electricity can. The marginal product of a fuel is determined in part by a complex set of attributes unique to each fuel: physical scarcity, capacity to do useful work, energy density, cleanliness, amenability to storage, safety, flexibility of use, cost of conversion, and so on. But also the marginal product is not uniquely fixed by these attributes. Rather the energy vector's marginal product varies according to what activities it is used in, how much and what form of capital, labor, and materials it is used in conjunction with, and how much energy is used in each application. Therefore, energy qualities are not fixed over time. However, it is generally believed that electricity is the highest quality type of energy, followed by natural gas, oil, coal, and wood and other biofuels in descending order of quality. This is supported by the typical prices of these fuels per unit of energy. According to economic theory, the price paid for a fuel should be proportional to its marginal product.

Samuel Schurr was among the first to recognize the economic importance of energy quality. Noting that the composition of energy use has changed significantly over time (Fig. 6), Schurr argued that the general shift to higher quality fuels reduces the amount of energy required to produce a dollar's worth of GDP. If this is ignored, apparent total factor productivity (TFP) growth is greater than is really the case. Researchers such as Cutler Cleveland, Robert Kaufmann, and the Office of Technology Assessment, have presented analyses that explain much of the decline in U.S. energy intensity in terms of structural shifts in the economy and shifts from lower quality fuels to higher quality fuels. In a recent paper, Kaufmann estimated a vector autoregressive model of the energy/GDP ratio, household energy expenditures, fuel mix variables, and energy price variables for the United States. He found that shifting away from coal use and in particular shifting toward the use of oil reduced energy intensity. This shift away from coal contributed to declining energy intensity over the entire 1929–1999 time period.

Figure 7 illustrates the increase in the second half of the 20th century in U.S. GDP and a quality-adjusted index of energy use computed by the author. The index accounts for differences in the productivity of different fuels by weighting them by their prices. There is clearly less evidence of decoupling of energy use and GDP in these data than in those in Fig. 4. If decoupling is mainly due to the shift to higher quality fuels, then there appear to be limits to that substitution. In particular, exhaustion of low-cost oil supplies could mean that economies have to revert to lower quality fuels such as coal.

3.4 Shifts in the Composition of Output

Typically, over the course of economic development, the output mix changes. In the earlier phases of development there is a shift away from agriculture towards heavy industry, and in the later stages of development there is a shift from the more resource-intensive extractive and heavy industrial sectors toward services and lighter manufacturing. Different industries have different energy intensities. It is often argued that this will result in an increase in energy used per unit of output in the early stages of economic development and a reduction in energy used per unit output in the later stages of economic development. Pollution and environmental disruption would be expected to follow a similar path.

Service industries still need large energy and resource inputs. The product being sold may be immaterial but the office towers, shopping malls, warehouses, rental apartment complexes, etc. where the activity is conducted are very material intensive and energy is used in their functioning as well as in their construction and maintenance. Other service industries such as transport clearly heavily draw on resources and energy. Furthermore, consumers use large amounts of energy and resources in commuting to work, to shop, etc. Therefore, a complete decoupling of energy and growth as a result of shifting to the service sector seems unlikely. When the indirect energy use embodied in manufactured products and services is taken into account, the U.S. service and household sectors are not much less energy intensive than are the other sectors of the economy, and there is little evidence that the shift in output mix that has occurred in the past few decades has significantly lowered the energy/GDP ratio. Rather, changes in the mix of energy used are primarily responsible. Furthermore, on a global scale, there may be limits to the extent to which developing countries can replicate the structural shift that has occurred in the developed economies and, additionally, it is not clear that the developed world can continue to shift in that direction indefinitely.

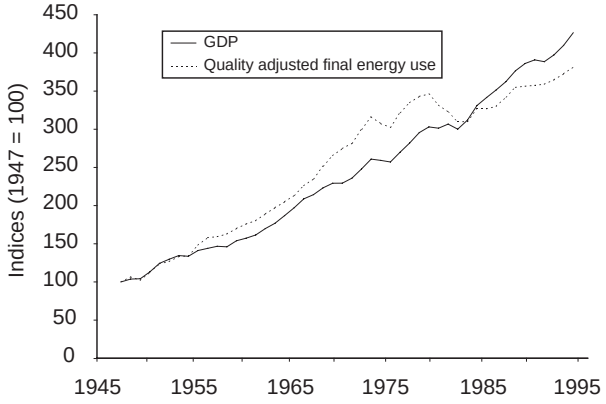


FIGURE 7 United States gross domestic product (GDP) and quality-adjusted final energy use GDP is in constant dollars, i.e., adjusted for inflation. Energy use is a Divisia index (the accumulation of weighted sums of growth rates) of the principal final energy use categories (oil, natural gas, coal, electricity, biofuels, etc.). The different fuels are weighted according to their average prices.

4. EMPIRICAL TESTING

Here the focus is on the empirical evidence for the tightness of coupling between energy use and economic output. Ordinary linear regression or correlation methods cannot be used to establish a casual relation among variables. In particular, it is well known that when two or more totally unrelated variables are trending over time they will appear to be correlated simply because of the shared directionality. Even after removing any trends by appropriate means, the correlations among variables could be due to causality between them or due to their relations with other variables not included in the analysis. Two methods for testing for causality among time series variables are Granger causality tests and cointegration analysis.

Granger causality tests whether one variable in a linear relation can be meaningfully described as a dependent variable and the other variable as an independent variable, whether the relation is bidirectional, or whether no functional relation exists at all. This is usually done by testing whether lagged values of one of the variables significantly add to the explanatory power of a model that already includes

lagged values of the dependent variable, and perhaps also lagged values of other variables.

Whereas Granger causality can be applied to both stationary and integrated time series (time series that follow a random walk), cointegration applies only to linear models of integrated time series. The irregular trend in integrated series is known as a stochastic trend, as opposed to a simple linear deterministic time trend. Time series of GDP and energy use are integrated. Cointegration analysis aims to uncover causal relations among variables by determining if the stochastic trends in a group of variables are shared by the series. If these trends are shared, either one variable causes the other or they are both driven by a third variable. It can also be used to test if there are residual stochastic trends that are not shared by any other variables. This may be an indication that important variables have been omitted from the regression model or that the variable with the residual trend does not have long-run interactions with the other variables. The presence of cointegration can also be interpreted as the presence of a long-run equilibrium relationship between the variables in question.

Both cointegration and Granger causality tests are usually carried out within the context of vector autoregression models. These models consist of group of regression equations in which each dependent variable is regressed on lagged values of itself and of all the other variables in the system.

Many analysts have used Granger causality tests or the related test developed by Christopher Sims to test whether energy use causes economic growth or whether energy use is determined by the level of output in the context of a bivariate vector autoregression. The results have been generally inconclusive. When nominally significant results were obtained, the majority indicated that causality runs from output to energy use. In a 1993 study, the author tested for Granger causality in a multivariate setting using a vector autoregression (VAR) model of GDP, energy use, capital, and labor inputs in the United States over a the post-World War II period. He also used a quality-adjusted index of energy input in place of gross energy use. The multivariate methodology is important because changes in energy use are frequently countered by the substitution of other factors of production, resulting in an insignificant overall impact on output. When both of these innovations are employed, energy is found to "Granger-cause" GDP. These results are supported by other studies that found that changes in oil prices Granger-cause changes in GDP and unemployment in VAR models, whereas oil prices are exogenous to the system.

Yu and Jin were the first to test whether energy and output cointegrate. They found that no such relationship exists between energy use and either employment or an index of industrial production in the United States. However, the lack of a long-run equilibrium relationship between gross energy use and output alone does not necessarily imply that there is no relation between the variables. Few analysts believe that capital, labor, and technical change play no significant role in determining output. If these variables are integrated, then there will be no cointegration between energy and output, no matter whether there is a relationship between the latter two variables. Also, decreasing energy intensity (due to increased energy efficiency), shifts in the composition of the energy input, and structural change in the economy mean that energy and output will drift apart. Similar comments apply to the bivariate energy-employment relationship. Further, using total energy use in the economy as a whole but measuring output as industrial output alone may compound the insensitivity of the test.

It would seem that if a multivariate approach helps in uncovering the Granger causality relations between energy and GDP, a multivariate approach should be used to investigate the cointegration relations among the variables. The author investigated the time series properties of GDP, quality weighted energy, labor, and capital series, estimating a dynamic cointegration model using the Johansen methodology. The cointegration analysis showed that energy is significant in explaining GDP. It also showed that there is cointegration in a relationship including GDP, capital, labor, and energy. Other analysts have found that energy, GDP, and energy prices cointegrate and that when all three variables are included there is mutual causation between energy and GDP. The inconclusive results of the earlier tests of Granger causality are probably due to the omission of necessary variables—either the quantities of other inputs (and quality adjustment of the energy input) or energy prices.

5. ENVIRONMENTAL IMPLICATIONS

5.1 Theory

The argument has been made here that there is a very strong link between energy use and both the levels of

economic activity and economic growth. What are the implications of this linkage for the quality of the environment? The focus here is now on immediate environmental impacts (issues of sustainability were discussed in Section 2).

Energy use has a variety of impacts. Energy extraction and processing always involve some forms of environmental disruption, including both geomorphological and ecological disruption as well as pollution. Energy use involves both pollution and other impacts, such as noise from transport, and land-use impacts, such as the construction of roads, etc. As all human activities require energy use; in fact, all human impacts on the environment can be seen as the consequences of energy use. Energy use is sometimes seen as a proxy for the environmental impact of human activity in general. Creating order in the economic system always implies creating disorder in nature, though this disorder could be in the sun or space rather than on Earth. The factors that reduce the total amount of energy needed to produce a dollar's worth of GDP, discussed in Section 3, therefore, also act to reduce the environmental impact of economic growth in exactly the same way as they reduce energy consumption. However, not all impacts of energy use are equally harmful to the environment and human health.

A shift from lower to higher quality energy sources not only reduces the total energy required to produce a unit of GDP, but also may reduce the environmental impact of the remaining energy use. An obvious example would be a shift from coal use to natural gas use. Natural gas is cleaner burning and produces less carbon dioxide per unit of energy derived. However, we need to be careful here. Nuclear-generated electricity is a higher quality fuel compared to coal, at least as measured by current prices, but its long-term environmental impacts are not necessarily lower. Incorporating the cost of currently unpriced environmental externalities into the prices of fuels would, though, raise the apparent quality of those fuels. This is not a contradiction, because a higher market price would mean that those fuels would only be used in more productive activities.

The environmental impact of energy use may also change over time due to technological innovation that reduces the emissions of various pollutants or other environmental impacts associated with each energy source. This is in addition to general energy-conserving technological changes that reduce the energy requirements of production, already discussed. Therefore, despite the strong connections between energy use and economic growth, there are several pathways through which the environmental impact of growth can be reduced. Again, however, if there are limits to substitution and technological change, then the potential reduction in the environmental intensity of economic production is eventually limited. Innovations that reduce one type of emission (for example, flue gas desulfurization) often produce a different type of waste that must be disposed of (in this case, gypsum that is possibly contaminated with heavy metals) as well as other disruptions required to implement the technology (in this example, limestone mining).

5.2 Empirical Evidence: The Environmental Kuznets Curve

In the 1970s, the debate on the relation between growth and the environment focused on the report, *The Limits to Growth*, by the Club of Rome. The common wisdom was that economic growth meant greater environmental impacts and the main way to environmental gains was reduced population and consumption. Economists and others argued that substitution and innovation could reduce environmental impacts (as was described in Section 2). But these were minority views. The mainstream view was that environment and economy were conflicting goals. By the late 1980s, however, with the emergence of the idea of sustainable development, the conventional wisdom shifted to one of "too poor to be green." Less developed countries lacked the resources for environmental protection and growth, and development was needed to provide for environmental protection. This idea was embodied in the empirical models that became known as environmental Kuznets curves (EKCs). The hypothesis states that there is an inverted U-shape relation between various indicators of environmental degradation and income per capita, with pollution or other forms of degradation rising in the early stages of economic development and falling in the later stages. The EKC is named for the economist Kuznets, who hypothesized that the relationship between a measure of inequality in the distribution of income and the level of income takes the form of an inverted U-shape curve. However, Kuznets had no part in developing the EKC concept.

The idea has become one of the "stylized facts" of environmental and resource economics. This is despite considerable criticism on both theoretical and empirical grounds. The EKC has been inter-

preted by many as indicating that no effort should be made to adopt environmental policies in developing countries, i.e., when those countries become rich, the current environmental problems will be addressed by policy changes adopted at that later time. As a corollary, it is implied that little in the way of environmental clean-up activity is being conducted in developing countries. These views are challenged by recent evidence that, in fact, pollution problems are being addressed and remedied in developing economies. In addition to the data and case studies provided by Susmita Dasgupta *et al.*, the author shows that for sulfur (widely believed to show the inverted U-shape relation between emission and income per capita), emissions in fact rise with increasing income at all levels of income, but that there are strong time effects reducing emissions in all countries across all income levels. In the author's opinion, this new evidence supersedes the debate about whether some pollutants show an inverted U-shaped curve and others—for example, carbon dioxide and "new toxics"—show a monotonic relationship. All pollutants show a monotonic relation with income, but over time pollution has been reduced at all income levels, *ceteris paribus*. Similarly, the debate about whether the downward sloping portion of the EKC is an illusion resulting from the movement of polluting industries to offshore locations is also now moot. This phenomenon might lower the income elasticity of pollution in developed economies relative to developing economies, but it does not seem sufficient to make it negative. The true form of the emissions–income relationship is a mix of two of the scenarios proposed by Dasgupta *et al.*, illustrated in Fig. 8. The overall shape is that of their "new toxics" EKC—a monotonic increase of emissions in income. But over time this curve shifts down. This is analogous to their "revised EKC" scenario, which is intended to indicate that over time the conventional EKC curve shifts down.

Although the environmental Kuznets curve is clearly not a "stylized fact" and is unlikely to be a useful model, this does not mean that it is not possible to reduce emissions of sulfur and other pollutants. The time effects from an EKC estimated in first differences and from an econometric emissions decomposition model both show that considerable time-related reductions in sulfur emissions have been achieved in countries at many different levels of income. Dasgupta *et al.* provide data and case studies that illustrate the progress already made in developing countries to reduce pollution. In addition, the income elasticity of emissions is likely to be less than 1—but not negative in wealthy countries, as suggested by the EKC hypothesis. In slower growing economies, emissions-reducing technological change can overcome the scale effect of rising income per capita on emissions. As a result, substantial reductions in sulfur emissions per capita have been observed in many Organization for Economic Cooperation and Development (OECD) countries in the past few decades. In faster growing middle income economies, the effects of rising income overwhelm the contribution of technological in reducing emissions.

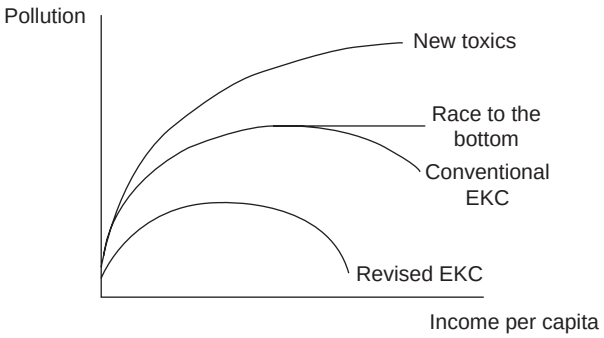


FIGURE 8 Environmental Kuznets curve (EKC): different scenarios. From Dasgupta *et al.* (2002).

This picture relates very closely to the theoretical model described in the previous Section 5.1. Overall, there is a strong link between rising energy use, economic growth, and pollution. However, the linkages between these three can be mitigated by a number of factors, including shifting to higher quality fuels and technological change aimed both at general increases in economic productivity and, specifically, at reducing pollution. However, given the skepticism regarding the potential for unlimited substitution or technological progress, there may be limits to the extent that these linkages can continue to be loosened in the future.

SEE ALSO THE FOLLOWING ARTICLES

Development and Energy, Overview • Economic Growth, Liberalization, and the Environment • Economics of Energy Demand • Economics of Energy Supply • Energy Ladder in Developing Nations • Environmental Kuznets Curve • Population Growth and Energy • Sustainable Development:

Basic Concepts and Application to Energy • Thermodynamics and Economics, Overview

Further Reading

Berndt, E. R., Kolstad, C., and Lee, J-K. (1993). Measuring the energy efficiency and productivity impacts of embodied technical change. *Energy J.* **14,** 33–55.

Brookes, L. (1990). The greenhouse effect: The fallacies in the energy efficiency solution. *Energy Policy* **18,** 199–201.

Cleveland, C. J., Costanza, R., Hall, C. A. S., and Kaufmann, R. K. (1984). Energy and the U.S. economy: A biophysical perspective. *Science* **225,** 890–897.

Dasgupta, P., Heal, G. (1974). The optimal depletion of exhaustible resources. Review of economic studies. Symp. Econ. Exhaustible Resources, pp. 3–28.

Dasgupta, S., Laplante, B., Wang, H., and Wheeler, D. (2002). Confronting the environmental Kuznets curve. *J. Econ. Perspect.* **16,** 147–168.

Georgescu-Roegen, N. (1976). "Energy and Economic Myths." Pergamon, New York.

Hannon, B. (1973). The structure of the ecosystem. *J. Theoret. Biol.* **41,** 535–546.

Hartwick, J. M. (1977). Intergenerational equity and the investing of rents from exhaustible resources. *Am. Econ. Rev.* **66,** 972–974.

Howarth, R. B. (1997). Energy efficiency and economic growth. *Contemp. Econ. Policy* **25,** 1–9.

Jorgenson, D. W. (1984). The role of energy in productivity growth. *Energy J.* **5**(3), 11–26.

Jorgenson, D. W., and Wilcoxen, P. J. (1993). Reducing U.S. carbon emissions: An econometric general equilibrium assessment. *Resour. Energy Econ.* **15,** 7–25.

Judson, R. A., Schmalensee, R., and Stoker, T. M. (1999). Economic development and the structure of demand for commercial energy. *Energy J.* **20**(2), 29–57.

Kaufmann, R. K. (2003). The mechanisms for autonomous energy efficiency increases: A cointegration analysis of the US energy/GDP ratio. *Energy J.* (in press).

Kaufmann, R.K., Azary-Lee, I.G. (1991). A biophysical analysis of substitution: Does substitution save energy in the U.S. forest products industry? Proc. Ecol. Econ.: Implications Forest Manage. Practice, 1991.

Khazzoom, D. J. (1980). Economic implications of mandated efficiency standards for household appliances. *Energy J.* **1**(4), 21–39.

O'Connor, M. P. (1993). Entropic irreversibility and uncontrolled technological change in the economy and environment. *J. Evolut. Econ.* **34,** 285–315.

Perrings, C. A. (1987). "Economy and Environment: A Theoretical Essay on the Interdependence of Economic and Environmental Systems." Cambridge University Press, Cambridge.

Schurr, S., and Netschert, B. (1960). "Energy and the American Economy, 1850–1975." Johns Hopkins University Press, Baltimore.

Sims, C. A. (1972). Money, income and causality. *Am. Econ. Rev.* **62,** 540–552.

Smulders, S., and de Nooij, M. (2003). The impact of energy conservation on technology and economic growth. *Resour. Energy Econ.* **25,** 59–79.

Solow, R. M. (1956). A contribution to the theory of economic growth. *Q. J. Econ.* **70,** 65–94.

Solow, R.M. (1974). Intergenerational equity and exhaustible resources. Review of economic studies. Symp. Econ. Exhaustible Resources, pp. 29–46.

Solow, R. M. (1997). Reply: Georgescu-Roegen versus Solow/Stiglitz. *Ecol. Econ.* **22,** 267–268.

Stern, D. I. (1993). Energy use and economic growth in the USA. A multivariate approach. *Energy Econ.* **15,** 137–150.

Stern, D. I. (2000). A multivariate cointegration analysis of the role of energy in the U.S. macroeconomy. *Energy Econ.* **22,** 267–283.

Stern, D. I. (2002). Explaining changes in global sulfur emissions: An econometric decomposition approach. *Ecol. Econ.* **42,** 201–220.

Stern, D. I., and Common, M. S. (2001). Is there an environmental Kuznets curve for sulfur? *J. Environ. Econ. Manage.* **41,** 162–178.

U.S. Congress, Office of Technology Assessment (1990). "Energy Use and the U.S. Economy." OTA-BP-E-57. U.S. Government Printing Office, Washington, D.C.

Yu, E. S. H., and Jin, J. C. (1992). Cointegration tests of energy consumption, income, and employment. *Resour. Energy* **14,** 259–266.

Economic Growth, Liberalization, and the Environment

SJAK SMULDERS
Tilburg University
Tilburg, The Netherlands

1. **Interactions between the Economy and the Environment**
2. **Trends in Economic Growth and Resource Use**
3. **Explaning the Link between Growth and Resource Use**
4. **International Trade and the Environment**
5. **Environmental Regulation and Productivity**

Glossary

comparative advantage A country has a comparative advantage in a particular good if the ratio of costs to produce the good, relative to the cost to produce another good, is lower than the same cost ratio in other countries. Comparative advantage is the basis for trade: Countries gain if they export the good they produce relatively most efficient.

economic growth Steady increases in gross domestic product.

energy intensity Energy supply divided by gross domestic product.

externality A situation in which market transactions affect persons who are not involved in the transaction directly (i.e., not through market prices).

gross domestic product (GDP) The sum of all output produced by economic activity within a country. Gross national product includes net income from abroad (e.g., rent and profits).

general equilibrium A situation in which all markets in the economy are in equilibrium.

market equilibrium A situation characterized by a certain volume of trade and a price of a certain good such that no market party wants to change behavior: No seller of the good wants to increase or decrease supply, and no consumer wants to reduce or increase demand.

Organization for Economic Cooperation and Development (OECD) The North American and Western European countries, Greece, Turkey, and Mexico.

real price Relative price of a commodity in terms of another basket of commodities. GDP is often measured in real prices (also labeled "at constant prices") to eliminate the effects of inflation (where inflation is the upward trend common in all prices, leading to a downward trend in the purchasing power of money).

total factor productivity (also multifactor productivity) Outputs per unit of input. Changes in total factor productivity measure technological change by accounting for how much of the increase in production is due to other factors besides changes in inputs.

Economic growth is the steady increase in output produced by economic activity within a country (gross domestic product). A central question is whether economic growth is compatible with a nondeteriorating environment and whether the environmental consequences of economic growth have a major impact on welfare levels derived from growing output of marketed goods. International trade has grown even faster than national income. It has been questioned whether increased trade has been bad for the environment and whether it has caused rich countries to relocate their polluting industries to developing countries. The following is another important question: How large are the costs associated with environmental regulation in terms of competitiveness, productivity, and employment? This article deals with these questions by reviewing empirical evidence and economic theory on resource use and economic growth.

1. INTERACTIONS BETWEEN THE ECONOMY AND THE ENVIRONMENT

1.1 Economic Growth and Resource Use

Statistics on national production levels and indicators of environmental pressure have been collected during

 © 2004 Elsevier Inc. All rights reserved.

the past few decades (and reconstructed for the more distant past) to document the link between economic growth and the environment. The theory of economic growth and economic theories of natural resources provide explanations and interpretations of the trends as well as methods to assess future developments. The basic approach in most statistical and theoretical analyses is the decomposition of aggregate production of a nation into, on the one hand, the current state of technology and, on the other hand, the inputs into the national production process. The main inputs are labor (hours worked), capital (produced means of production that can be used over a certain time period), and natural resources.

Natural resources comprise energy inputs and material inputs. An important distinction is between renewable and nonrenewable resources. Renewable resources are biotic populations, flora, and fauna, which have the potential to grow by means of natural reproduction. For example, the theory of renewable resources studies fishery and forestry. Nonrenewable resources are minerals of which physical supplies do not grow on a relevant timescale for humans. Examples are metals (e.g., aluminum, copper, and gold) and fossil fuels (e.g., oil, coal, and natural gas).

The basic characteristic of natural resources is that they cannot be produced, but they serve as an essential input in production. A higher rate of resource usage implies faster depletion of the available stock of the resource. In this sense, resource inputs differ in a fundamental way from other inputs used in production, such as labor and capital: Labor is not used up and man-made capital can be accumulated. Another important difference is that the available resource stock may directly affect welfare as an amenity (nonuse value). Resource economics studies the trade-off between extraction of the resource for current consumption and conservation for later dates. Environmental quality can also be considered as an input in the aggregate production process. Pollution is an inevitable by-product of production. Pollution can be seen as a degradation of environmental resources, such as clean air and soil.

1.2 Externalities, Markets, and Incentives

Economic growth is the result of growth in inputs and increases in the productivity of the inputs—that is, changes in technology. Economics further explains why inputs and technology change over time. Inputs such as materials, labor, and capital are traded in factor markets. Prices match demand and supply and provide consumers and suppliers with information and incentives. Population growth, for example, increases the potential working force and the supply of labor inputs in the economy increases. This fuels growth if wages adjust so that the growing working force is employed in production. With more labor in production, capital can be operated at a higher productivity level. This increase in the return to capital fuels investment in new capital so that growth in capital inputs also fuels output growth. Similarly, the demand for materials and energy increases. If these are available in limited supply, their prices increase, which provides incentives to develop technologies that save on their use. Thus, although growing scarcity of inputs may impose a drag on growth, it may also induce technological change that fuels economic growth.

Markets provide a powerful coordination mechanism. In sectors of the economy in which the demand for resources such as materials or labor is high, prices increase and this attracts resources to these sectors, away from other sectors. Thus, resources tend to be allocated to those sectors in which they are in highest demand. The high demand reflects a high willingness to pay for the resources and implies that their value is considered to be high. This explains why the market mechanism in principle can enhance efficiency and welfare in the economy.

The market can only coordinate social preferences (demand for certain goods or services) and (technical) possibilities and resource endowments in a welfare-maximizing way if prices reflect these preferences and possibilities. This requires (i) that the owners of production factors (the inputs such as labor, capital, and materials) can charge a price that reflects their production cost and (ii) that consumers can express their willingness to pay. For many environmental goods, these preconditions are not met. In the absence of environmental regulation, there are no markets for clean air or climate. Pollution entails a cost to society but not a cost to the polluting producers. Waste can be disposed freely but imposes a cost on society. When economic activity (i.e., market transactions) imposes costs or benefits on persons who are not involved in the transaction directly, externalities are said to be present. Externalities cause markets to function imperfectly and welfare to be impaired. This is why environmental policies may improve welfare. These policies have to correct environmental externalities, for example, by making firms pay for the pollution costs they impose on society.

We can now more precisely define the cost of pollution for society. From an economic perspective,

pollution (and, similarly, environmental degradation or resource depletion) is costly to the degree that it causes externalities. Because the value of environmental quality is not reflected in the prices in the economy, the economic decisions that are based on these prices give rise to more pollution than society finds optimal.

The ideal (optimal) environmental regulation imposes measures to eliminate all external costs of pollution; that is, it makes all parties in society internalize all social costs of pollution. In such a situation, welfare is maximal, but pollution is not necessarily (and very unlikely) at the technical minimum. Although depletion of natural resources and polluting consequences of production are by definition bad for the environment, they are not necessarily bad for welfare. They provide the inputs for production of valuable things. As long as the benefits from increases in production outweigh the costs of environmental degradation, economic growth enhances welfare.

The following sections examine how economic growth, international trade, and environmental regulation affect the environment and welfare. To examine the impact on aggregate welfare, changes in the physical state of the environment have to be made comparable to changes in economic production, which requires a valuation method to place values on environmental improvements. Since this is a difficult and subjective task, I also discuss the physical effects, by which I mean the impact on pollution, pollution concentration, and resource stocks measured in physical terms. This is the conservationist perspective, to be contrasted with the economic approach that deals with welfare effects.

2. TRENDS IN ECONOMIC GROWTH AND RESOURCE USE

2.1 Trends in Economic Growth

From 1820 to 1995, total world production [gross domestic product (GDP)] grew on average at an annual rate of 2.2%, which amounts to a more than 40-fold increase over a mere 175 years. This rate of change is unprecedented: In the 1800 years before this period, world GDP increased less than 7-fold. Per capita economic growth has also been enormous during the past century. With the current world population of approximately 5.5 billion, five times as large as in 1820, per capita GDP is now almost eight times higher than it was 175 years ago.

Differences among countries are major. In 1988, output per worker in the United States was more than 30 times higher than that in sub-Sahara Africa. Developing countries often have per capita income levels that are less than 10% of U.S. levels (e.g., in 1988, India, 9%; China, 6%). Per capita income in Western Europe is on average 30% lower than that in North America, but income gaps are smaller than they were four decades ago. Typically, among the richer countries, there is convergence in income levels, whereas the gap between the poorest and richest countries grows.

2.2 Trends in Energy and Materials

Economic growth directly translates into growing demand for energy resources and materials since they are necessary inputs in production. However, in the process of growth, some energy resources or materials may be substituted for others so that not all of them are used at an increasing rate. Technological change has allowed many sectors of the economy or even entire economies to produce at a lower overall energy intensity.

Primary energy supply in the Organization for Economic Cooperation and Development (OECD) countries was approximately 50% higher in 1999 than in 1971, but GDP increased faster. Indeed, from 1960 to 2000, energy use was decoupled from economic growth, at least in the European Union, Japan, and the United States (at annual rates of 0.4, 0.2, and 1.4%, respectively). The oil shortages in the 1970s induced energy conservation programs. Also, the growing relative importance of services relative to manufacturing caused energy efficiency to increase. In recent years, growth in energy use has accelerated due to increased transport and electricity use in service sectors. Trends in materials use vary considerably for different materials. Iron and steel have grown slower than GDP, paper use has grown in line with GDP, but plastics and aluminum have grown faster than GDP.

The concern is that growing energy and materials use results in scarcity of resources and imposes a threat to future growth. Stocks of virgin ores decline. The stock of materials embedded in products grows, leaving less available for recycling. If growth fuels demand for resources and supply becomes smaller, prices can be expected to rise. However, most price indicators do not reflect growing scarcity. The prices of most commodities have declined since 1960. Here, prices are measured in real terms (i.e., relative to the prices of a representative basket of produced goods). Most economists have concluded that resource scarcity will not limit growth, at least not during

the coming decades and not if we consider scarcity of those resources that are traded in well-developed global markets, such as most minerals and energy resources.

Declining prices may reflect technological progress, which allows mining at lower cost or exploration of new resource deposits. According to most economic theories, this is likely to be a temporary phenomenon, and in the long term prices of nonrenewable resources are expected to increase. This does not necessarily imply that future growth will no longer be fuelled by lower costs of energy and materials, since technological change, as well as exploration and development of new economically useful resources, may overcome scarcity limits.

The growing use of certain energy resources and materials has severe impacts on the natural environment. In general, economists agree that free market forces do not serve very well society's demand for environmental amenities or conservation of natural resources for which markets are illegal, incomplete, or do not exist (e.g., water and air quality, rare species, biodiversity, and ecosystem services). Emissions of CO_2 from fossil fuel combustion pose major threats to air pollution and climate. Despite the decoupling of energy use from economic growth in the developed economies and despite the decline in the carbon content per unit of energy used, total CO_2 emissions have risen in almost all countries in the past few decades. In large developing countries such as India, China, and Brazil, absolute CO_2 emissions have increased very rapidly. In contrast to the OECD experience, their CO_2 emissions per unit of GDP have risen at fast rates (up to 2.6% percent per year in India from 1971 to 1998).

Other major threats to the environment come from the use of hazardous industrial waste, emissions, and toxic chemicals that are used in the production process and discarded as some form of industrial waste. As a result of environmental regulation, emissions from production activities (industry) have actually declined. However, waste and emissions from consumption activities are growing. Municipal waste and hazardous waste have increased in recent decades in developed countries at a rate faster than those of the population and GDP.

2.3 Cross-Country Differences in Resource Intensities

Differences in energy used per unit of GDP are large. In North America, per capita energy use is more than twice that in Japan, France, or the United Kingdom and approximately 1.5 times that in Germany and Australia, mainly due to differences in per capita income and in energy used per unit of output. Differences in the structure of production (in terms of the relative importance of industrial production sectors) are hardly responsible for the differences in per capita energy use.

3. EXPLANING THE LINK BETWEEN GROWTH AND RESOURCE USE

3.1 The Environment–Income Relationship

Economists have tried to identify underlying mechanisms in resource use and growth trends. A central question is how environmental pressure changes with economic growth. One hypothesis is the so-called environmental Kuznets curve (EKC) hypothesis, according to which environmental pressure or resource use first increases with GDP but, after a certain threshold income level, declines with per capita GDP. The hypothesis has been tested for several resources, energy sources, and emissions. The evidence is mixed: The EKC shows up, but only for some pollutants and not in all countries.

In the typical approach, pollution in the major cities (or other measurement stations) of a set of countries is compared and related to the level of income in these cities. If pollution rises with the level of income but also declines with the square of the level of income, the EKC holds.

The EKC is found for water pollutants and for air pollutants, such as SO_2, suspended particulate matter (dark matter as well as heavy particles), NO_x, and CO. These pollutants cause health problems close to the place where they are emitted (i.e., these are local pollutants), whereas their impact occurs simultaneous with emissions (i.e., these are short-lived or flow pollutants). The EKC is generally rejected for municipal waste, CO_2, and aggregate energy consumption: These variables monotonically increase with the level of income in a cross section of countries. Since it is the accumulation of municipal waste in landfills, rather than the annual flow, that imposes environmental pressure, it is a stock pollutant. CO_2 is partly a local flow pollutant, but its main impact is through concentration of CO_2 in the atmosphere on the global climate (i.e., it is a global pollutant as well as a stock pollutant). Other evidence on the main global stock pollutants and environmental indicators is only

indirect. For example, wilderness land is monotonically declining with the level of income, which is a crude measure of the negative relation between income and biodiversity.

3.2 The Role of Environmental Regulation

The common pattern is that local pollutants and flow pollutants are more likely to first rise and then fall with income than global pollutants or stock pollutants. The main driving force behind the lower pollution levels in richer countries is policies to reduce pollution. In richer countries, citizens are willing to pay more for improvements in environmental quality and succeed in convincing their local and national governments to impose the necessary regulation to reduce urgent pollution problems. Since reducing local flow pollutants has the most visible effects for the constituencies of the local authorities, regulation is mainly directed at these rather than global or stock pollutants. The latter affect other regions or future generations, respectively, but not so much the authorities' electorate. In this interpretation, environmental improvement only occurs when environmental regulation becomes more stringent in richer economies. That is, environmental problems are not solved automatically by economic growth. Growth will be accompanied by reductions in pollution only if a number of conditions are fulfilled.

First, when becoming richer, people must care more about environmental quality. There is no strong evidence in favor of this claim. For example, in developing countries, environmental quality and availability of natural resources are often much more crucial for households than in developed countries. However, within developed countries, willingness to pay for improvements in health increases with income. As a result, richer citizens may be more in favor of costly measures to reduce pollutants with direct health impacts.

Second, only if the costs of reducing pollution do not increase too much with national income will citizens vote for pollution reduction. Technological change is important in this respect. On the one hand, cleaner production processes have been developed, mitigating the cost of pollution reduction. On the other hand, some new production processes and the production of new goods require more energy and resource use. For example, the rapid spread of computers and the growing demand for transport make the economies heavily dependent on electricity and fossil fuels so that reductions in energy use have a major cost in terms of loss of production. This may explain why electricity and fossil fuel use do not decline with income, despite concern about the environment.

Third, actual environmental policies need to reflect societies' preferences for environmental quality. Indeed, in countries with a relatively high degree of political and civil freedom, environmental problems decline more rapidly with income. In countries with high indicators of corruption, natural resource depletion is faster.

Pollutants differ in terms of their (health) impact, the degree to which they are local and short-lived, and the cost of reduction. As a result, when comparing emissions of different pollutants in a particular country, some may increase and others may decrease with income. That is, the Kuznets curves, if they exist at all for the pollutants at hand, overlap. It implies that although environmental quality improves due to the reduction in one pollutant, aggregate measures of environmental quality may deteriorate because this pollutant is replaced by other pollutants. This may be a problem particularly since stock pollutants replace short-lived pollutants, which amounts to a shift of the environmental burden to future generations. Similarly, Kuznets curves of different countries overlap. While rich countries reduce emissions of a certain pollutant, poor countries may simultaneously increase emissions in their growth process. Thus, pollution is shifted from rich to poor countries. As a result, the existence of an EKC for a certain pollutant does not guarantee that worldwide environmental quality improves over time. This phenomenon is particularly important because the poor countries, which tend to increase emissions with production, account for a much larger share of the world economy than rich ones and some of them are growing at above-average rates (notably China).

4. INTERNATIONAL TRADE AND THE ENVIRONMENT

4.1 The Pollution Haven Hypothesis

The evidence on the EKC indicates that national environmental policies differ between high-income and low-income countries, and that as a result of income growth pollution in the Northern hemisphere may be reduced over time at the cost of increasing pollution in the Southern hemisphere. There have

been concerns that this process is reinforced by international trade of goods and by foreign direct investment between North and South. More stringent environmental policies in the North force pollution-intensive industries located in the North to raise their prices relative to the prices charged by firms in the South. Consumers in the North may then profitably import more goods from the South and firms in the North may profitable relocate their plants to the South. This claim on the effects of differences in stringency of environmental regulation is called the pollution haven hypothesis.

Empirical evidence for the pollution haven hypothesis is not very strong. The hypothesis has been tested in various (mainly indirect) ways. Typically, in a first step polluting industries are identified based on their emissions per unit (dollar) of output or pollution control costs as a fraction of value added. At the top of these rankings are industries such as iron and steel, nonferrous metals, mining, pulp and paper, industrial chemicals, petroleum refineries, rubber, and leather products. In a second step, the performance of these sectors in different countries is related to the country's environmental stringency, controlling as much as possible for other country differences. Stringency is either measured as an index based on law and regulation or a measure of pollution control expenditures.

Within the United States, polluting industries tend to export smaller volumes and build fewer new plants, and there is less investment in states with more stringent environmental regulation than in states with more lenient environmental regulation. Among different countries, the effects of differences in environmental regulations are more difficult to trace. Overall, industrial countries have reduced their share in world trade of goods from polluting industries relative to developing countries during the past few decades. However, there is no evidence that this trend is mainly due to differences in environmental regulation or due to international trade. The pattern of growth in these countries may be caused by a shift to pollution-intensive industries (as a movement along the upward sloping part of the EKC). Moreover, industrial countries remain the largest exporters in the most polluting industries. Hence, we must reject the claim that internationally most pollution is created in countries with lax environmental standards. The main explanation for this finding is that other factors besides environmental regulation are much more decisive in determining production cost or location advantages. The share of compliance costs for environmental regulation is small relative to the cost savings that can be gained from the availability of skilled workers able to operate advanced technologies, access to large consumer markets, and political stability. Polluting industries sometimes have high transport costs, so proximity to consumer markets is important. Polluting industries are also typically capital-intensive and operated by skilled workers. These factors make developed countries a more attractive location for polluting industries. Hence, differences between North and South in availability (i.e., endowments) of capital and skilled workers offset the differences in environmental regulation.

4.2 Is Trade Beneficial?

For given differences in endowments of capital and labor, relatively lax environmental regulation in some countries, particularly poorer ones, is likely to shift some of the dirty industries to the South. The question is whether this is always undesirable. According to a standard economic argument, there may be mutual benefits from relocation of pollution-intensive industries between different countries. For example, countries with low income but enough space to keep pollution away from their populations may be willing to accept pollution if this can greatly boost income. High-income countries with high population density care less about increases in income if they are obtained at the cost of pollution problems. In this case, both types of countries benefit if pollution is relocated from high-income to low-income countries.

Although lax environmental policies in the South seem to burden citizens in the South with pollution havens, workers in the North may be concerned about the stringent environmental policies in their countries. The North loses jobs in pollution-intensive industries and export performance in these sectors declines when these industries move to the South. However, at the same time, trade with the South allows the North to improve performance in clean industries so that again trade creates mutual benefits for North and South (although these benefits may be obtained only after some time-consuming, and therefore costly, adjustments in the economy). Indeed, also if the North did not engage in trade with the South, environmental regulation would result in job losses in polluting industries and shifts to clean industries, but it would forego the benefits from importing and consuming goods that can be produced at lower costs in other countries and the opportunity to export goods that can be produced at a comparative advantage in the North.

In general, differences in environmental regulation may reflect differences in social costs to deal with pollution and form a true basis for mutually beneficial trade. However, if national environmental regulation ignores the average citizen's preference for pollution and fails to reflect the social costs of pollution, trade is not always mutually beneficial. This is why large multinational corporations have been criticized when they negotiate concessions with elite governments and exploit their monopoly power. However, evidence on unbalanced benefits and costs from the activities of such firms is not available.

Other situations in which trade does not necessarily lead to mutual gains occur when pollution is transboundary—that is, when pollution from one country harms citizens of other countries. Since national governments are not likely to take into account these pollution spillover effects, they tend to set pollution regulation too lax. For all countries to gain from trade, environmental regulation in each country should reflect the cost of pollution not just to its own citizens but also to the world community as a whole. Thus, if SO_2 emissions in the United Kingdom hardly affect British citizens but cause forests in Scandinavia to die, the United Kingdom may be tempted to impose lax regulation of SO_2 emissions, thus creating a comparative advantage in SO_2-intensive goods and boosting export of these goods. This trade would harm Scandinavia.

A final situation of harmful international trade may arise when countries have strategic reasons to set overly lax or overly stringent environmental policies. Some environmentalists fear that governments set environmental policies overly lax. Governments may have an incentive to accept environmental standards that are lower than is socially desirable in order to attract foreign direct investment or give domestic industry a cost advantage over foreign firms, thus enabling them to make higher profits (which has been called ecological dumping). However, if many governments try to protect their polluting industries in this way, they may all fail in a "race to the bottom": The lax standard in one country is offset by the response in the other country to also lower environmental regulation. Only if firms' location choices and cost levels are sufficiently responsive to differences in environmental standards can national governments use environmental policies as a strategic instrument against other countries. In the past, the lack of strong evidence for pollution havens has been interpreted as evidence against significant firm responses to environmental stringency and the fear of strategic use of environmental policy seemed ungrounded. Recently, however, evidence of firm responses has become stronger so that strategic environmental policy may still be reason for concern.

4.3 Is Trade Good for the Environment?

Although there is no hard evidence on the magnitude of the net benefits (or costs) of trade in polluting goods, and there is no strong evidence regarding major effects of environmental regulation on trade and foreign direct investment, there are detailed analyses of the effects that an increase in world trade has on pollution in different countries. Trade has significant effects on pollution, but often more trade implies a cleaner environment.

World trade has increased enormously during the past few decades. Between 1973 and 1998, world merchandise exports and global foreign direct investment increased at annual rates of 9 and 14%, respectively. This expansion of trade is not responsible for increases in world pollution. The share of dirty products in world trade has declined. Comparing open economies (countries with high shares of exports and imports in GDP) to closed economies (countries with little foreign trade), statistical analysis shows that open economies have less pollution per unit of GDP (other things being equal). Open economies tend to adopt clean technologies more rapidly.

Trade often has a beneficial impact on the environment because of the interaction between comparative advantages and adjustment of environmental regulation. High-income countries tend to have higher environmental standards. However, high-income countries can also produce pollution-intensive (which are often skilled-labor or capital-intensive) products at a relatively low cost due to access to advanced technology and large endowments of skilled labor and capital. This implies that high-income countries have a comparative advantage in producing pollution-intensive goods. When they liberalize trade, they tend to export more of these goods. Thus, world trade becomes greener since trade shifts production of polluting goods to countries in which environmental standards are higher. Moreover, international trade increases income, which leads to a policy response since richer countries tend to impose more stringent environmental regulation. Thus, countries with lower incomes tighten their environmental regulation. Multinational firms adopt environmental practices overseas that are (partly) determined by environmental regulations in their home countries under pressure of consumer actions and for fear of loss of reputation.

Although evidence on an aggregate level suggests that international trade is benign to the environment, in certain areas trade has a serious adverse impact on the environment. Certain types of pollution or natural resource use remain unregulated, and it is in these areas that there are concerns. One example is transport. The growth in world trade has been accompanied by increases in transport movements, which are heavily energy intensive. The pollution costs (externalities) associated with transport are not fully reflected in regulation so that welfare losses and excessive pollution may stem from the increase in transport. Since the costs of transport cross borders, appropriate policies require international coordination, which is often difficult to establish. Second, transport involves the problem of invasive species: The accidental or intentional introduction of harmful nonindigenous species of plants and animals can damage environmental resources. The cost of screening and quarantines is often prohibitive. Third, the concentration of agriculture in certain locations without appropriate regulation puts stress on biological diversity with potential loss of soil quality in larger regions in the long term. It destroys the habitat of species and may lead to extinction of species and loss of biodiversity.

5. ENVIRONMENTAL REGULATION AND PRODUCTIVITY

5.1 The Cost of Environmental Regulation

Environmental regulation aims at improving environmental quality but may come at the cost of jobs, productivity, or other undesirable economic effects. To justify environmental regulation on economic grounds, it should be shown in a cost–benefit analysis whether the benefits are large enough to incur the costs. In some cases, the costs are found to be small or even negative; that is, there are benefits to be reaped not only in terms of environmental improvement but also in other fields. Estimates of the total costs of environmental regulation vary enormously between studies.

The traditional argument is that environmental regulation imposes costs on firms as well as governments and consumers. Governments incur costs of administration of the regulation. Firms are burdened with costs because they have to install and operate pollution control equipment and intensify monitoring, and they face higher administrative burdens. Second, under certain types of regulation, they may face pollution charges, expenditures on buying pollution permits or quotas, or fines insofar as they emit more than the quotas that are allocated to them. Third, depending on the degree that they pass on costs through higher prices, demand for their output will decline and they will suffer from lower profits.

The total burden on polluting firms depends heavily on firm characteristics and the employed technology. Estimates of direct costs of complying with environmental regulation vary from less than 1% of GDP in 1972 to more than 2.5% at the end of the 1990s in the United States and slightly lower in Western Europe. Petroleum and coal industries faced pollution control capital expenditures up to one-fourth of total capital expenditures, whereas for rubber and plastics production industries the number was only 2%.

More crucial from an economic perspective is that the burden of regulation depends on the design of the regulation, particularly the flexibility firms are given to choose their own way to comply with the regulation. At the one extreme, the regulator prescribes the technology the firms have to use or the pollution control actions they have to take, leaving little flexibility for the firms. At the other extreme, the regulator allows firms a specified emissions allowance and leaves it to the firms' discretion to determine by what means to reduce emissions as well as whether to trade pollution allowances with other firms (system of marketable pollution permits). Compared to the former type of regulation (the command-and-control approach), the latter type (the market-based approach) minimizes the costs of regulation. First, it allows firms to choose the cheapest way to reduce emissions. Profit-maximizing firms will have an incentive to do so. Second, it gives incentives to firms that can reduce emissions at lowest cost to do most of the reduction and sell their pollution allowances to other firms. Thus, total pollution control costs are minimized. Alternative market-based instruments are charges (taxes) proportional to emissions.

Market-based instruments have been applied less frequently than command-and-control instruments, but their importance has increased. Marketable permits have been successful with respect to reductions of lead content in fuel, air pollutants, and SO_2 emissions in particular. Estimates of the cost savings potential of changing from command-and-control to market-based instruments range from 50 to 90%.

Sometimes, regulated firms have zero cost options to comply with more stringent environmental

regulation, or even make profits. Case studies have documented situations in which the reductions of materials or toxic chemical use for environmental reasons saved enough of the firm's expenditures on inputs to easily repay the investment cost of adjusting the production process. Even without environmental regulation, the investment would have been justified since it allowed for a real efficiency improvement.

It is unlikely that these situations arise systematically since managers of firms are supposed to search for opportunities to improve efficiency and exploit them even without intervention by environmental regulators. Thus, when firms maximize profits and regulation constrains their choice of technologies, the smaller choice menu must lead to fewer opportunities to operate at certain profit levels and regulation is costly. This conclusion is unlikely to change if we acknowledge the limits to profit maximization because of informational and managerial problems. No doubt, inefficiencies in firms' production processes and organization remain pervasive since managers' span of control is insufficient to assess and eliminate all of them. Because of the small share of environment-related costs, managers may systematically ignore environment-related inefficiencies. Then, environmental regulation may make managers more focused on cost savings potentials related to pollutants and materials use. However, it may come at a cost by turning their attention from other inefficiencies so that on balance it is unlikely that environmental regulation can systematically (i.e., economywide) improve overall efficiency.

Opportunities for cost savings have been explored by engineering or "bottom-up" studies. These studies start from an emissions reduction target for the economy as a whole and then identify all the technological measures by which this target can be achieved. For each of them, the researcher estimates the cost incurred by firms and adds these up to an economywide total. Often, the numbers are surprisingly low, suggesting that technological opportunities abound to improve the environment at low cost. The problem with these studies is that many hidden costs are ignored. Adoption of new technologies will not only change firms' emissions but also change product characteristics and quality, employment structure, and input demand. It requires adjustments beyond the simple installation of pollution control equipment. The entire production and marketing process may be affected.

The alternative method for assessing pollution control costs, the "top-down" approach, involves a statistical (and econometric) analysis of the relationship between production costs and environmental regulation and between emissions and regulation across industries. The statistical results are used to make a projection of how costs are affected by a hypothetical change in emissions, which corresponds to the policy target. This requires building a stylized model of the entire economy, which involves the incorporation of many assumptions and additional statistical results regarding how markets work and react to policy changes. Top-down studies suggest much higher costs of environmental policy than do bottom-up studies. Hence, adjustment costs are substantial. Drawbacks of this method are that it necessarily relies on historical data on firms' reactions to regulation and it cannot incorporate the details of available new technologies to meet the environmental target. The results are also quite sensitive to the modeling structure and assumptions made.

Econometric techniques have also been used to estimate the effects of environmental regulation on employment. One would expect to find a negative correlation between stringency of environmental regulation and employment changes in pollution-intensive industries. Most studies find minor effects when relating environmental regulation across U.S. states to employment. Often, the correlation between employment in dirty industries and environmental regulation is positive. The main explanation is that we should not expect a causal relation from environmental stringency to employment but rather view both employment and regulation as being determined simultaneously by economic and political factors. States with a strong economy, industries with high productivity, and high income can afford—and may actually prefer—more stringent regulations than poorer states; at the same time, the former may experience strong employment growth as a result of the favorable economic conditions.

5.2 General Equilibrium Effects

To assess the overall costs of environmental regulation, one must measure not only how it directly imposes costs on a regulator and regulated firms but also how it affects other areas of the government's budget and how it affects customers and suppliers of regulated firms. Almost all activities in the economy are linked to each other: A consumer buys goods from firms and receives his or her income from (other) firms and pays taxes to governments, which subsidize firms and provide infrastructural goods for firms. Hence, regulating certain firms may affect the entire economy through these linkages. Such indirect

effects are labeled general equilibrium effects because they reflect the effects after all markets in the economy have reacted and adjusted such that a new equilibrium in the economy as a whole results.

The linkages between economic activities may sometimes mitigate the costs of environmental regulation and sometimes reinforce them. When regulated firms reduce their polluting activities, they buy less from their suppliers, which also experience a loss in sales. However, other firms may benefit. Regulated firms may need to buy specialized equipment to control pollution, so firms selling this equipment benefit. When regulated firms pass on their higher costs through higher prices, their customers switch to other producers. Hence, firms likely to benefit from regulation are those that supply goods that customers view as a suitable substitute for the regulated firms' goods.

The overall cost of environmental regulation is the sum of all gains and losses. The indirect gains are not likely to fully offset the direct costs for one obvious reason: Regulation aims at reducing pollution, and this requires reductions in polluting inputs. If the economy as a whole has to cut back on inputs, total output is likely to decline, too. This effect may be reinforced over time through effects on investment. Investment in new capital goods fuels growth in the economy. The return to investment is likely to be smaller if new capital goods have to be run with less energy or polluting inputs so that the incentive to invest is crowded out through more stringent environmental regulation. Thus, growth may be impaired by environmental regulation.

The economywide impact of environmental regulations can only be estimated, rather than directly measured, since we cannot observe how the economy would have evolved in the absence of the regulation. Estimates therefore heavily depend on the model that is used for the estimation. Based on standard macroeconomic or growth-theoretic models, estimates of the total cost are on the same order of magnitude as the direct pollution control costs, approximately 2% of GDP. Some researchers have noted that the surge in regulatory costs coincides with the growth slowdown in the 1970s. However, all calculations show that only a very small portion of the slowdown can be attributed to environmental regulation.

5.3 In Search of "Double Dividends"

Despite the fact that environmental regulation amounts to restricting inputs in the economy and is therefore likely to reduce output, counterarguments have been advanced that environmental regulation may bring about large gains. The key argument is that environmental policy shifts economic activity away from polluting sectors to other sectors in which the contribution of economic activity to welfare is larger. Hence, although regulation forces the economy to reduce polluting inputs, it may increase total output by improving the average efficiency at which inputs are employed.

5.3.1 Elimination of Inefficient Subsidy and Tax Programs

Efficiency gains can arise from environmental regulation if efficiency differs between sectors, particularly if efficiency in polluting sectors is lower than in clean sectors. Normally, differences in efficiency are eliminated by market forces, but (nonenvironmental) regulation may distort these market forces and create sectors in which resources are inefficiently used. Important examples are agriculture and energy. In the OECD, agricultural output is heavily subsidized so that high-cost/low-productivity agricultural activities in these countries remain viable. This support creates inefficiencies, and the tax revenue could be better spent from an efficiency standpoint. The widespread use of energy subsidies to firms also creates support for low-productivity, energy-intensive production. If these subsidies are abolished and replaced by environmental regulation that reflects, for example, the adverse impacts of pesticide on soil and water quality or those of energy use on world climate, efficiency gains can be large by shifting economic activity away from inefficient farming and energy-intensive industries.

5.3.2 Environmental Tax Reform

Gains from environmental regulations have also been claimed to be feasible when they can be combined with reductions in certain taxes. On the one hand, governments have relied on taxes on labor and capital to fund their budgets. These taxes may reduce the incentive to participate in the labor market or to invest in capital so that the taxes distort markets and create inefficiently low levels of employment and investment. On the other hand, environmental taxes (charges on emissions and auctioning off of pollution permits) not only give incentive to address environmental problems but also raise revenue.

Indeed, the first important insight from the theory of public finance theory is that it is better to "recycle" the revenues from environmental taxes through reductions in distortionary taxes than to use them

to fund the general government budget. However, the second insight is that replacing distortionary taxes on labor and capital by revenue-raising environmental taxes will reduce the efficiency of the tax system and increase the burden of taxation. That is, the improvement in the environment (a first dividend) is not likely to be offset by a cheaper way for the government to raise taxes (which would give a second, nonenvironmental, dividend). The reason is that a change from taxing labor and capital to taxing pollution narrows the base on which the taxes can be levied. Capital and labor income each constitute a much larger fraction of total income than the share of pollution in total cost. As noted previously, environmental taxation reduces inputs in production and forces firms to spend more on pollution control. This reduces output and income in the economy, the tax base becomes smaller, and the government raises less tax revenue from nonenvironmental taxes. The revenues from environmental taxes are insufficient to make up for the loss of revenues from nonenvironmental taxes.

In practice, the replacement of nonenvironmental taxes by revenue-raising environmental taxes may yield a double dividend if the existing tax system is highly inefficient in other, nonenvironmental dimensions (e.g., if capital is excessively taxed relative to labor). In this case, the tax system can be improved even without the introduction of environmental taxes, but vested interests or distributional concerns may prevent these tax reforms. Results from economic simulation models suggest that gains from the combination of environmental taxation and tax reform may yield substantial benefits in practice but not for all groups in society. The major challenge is to find ways to compensate losers from tax reform without providing the wrong incentives. For example, workers and capital owners in polluting industry are likely to be negatively affected by environmental tax reform, but compensating them for losses should not reduce their incentive to invest funds and to find alternative employment outside the polluting industry.

5.3.3 Environmental Quality as a Source of Productivity

Environmental regulation has short-term costs but long-term gains. These gains are primarily in terms of a cleaner environment; they take time to materialize since the recovery of environmental quality is time-consuming. However, as a feedback effect, a cleaner environment may improve productivity in the economy, which may offset the economic costs. The most obvious example is the effect on productivity in sectors that heavily rely on environmental resources. Policies to improve water and air quality of certain regions may boost the attractiveness of these regions for tourism or residential development. Hence, profits and productivity of tourism and construction may increase. Policies aimed at marine resources may also boost fish populations, thus improving productivity of the fishery. Forest management may give rise to both environmental benefits and improved yields for forestry.

In general, productivity of sectors or the economy as a whole is measured by multifactor productivity, which is a measure of the amount of outputs per unit of inputs. The factor productivity of conventional inputs, such as labor and capital, may increase if environmental quality improves. Capital equipment and residential buildings may depreciate more slowly if there is less air pollution. The productivity of labor may benefit from health improvements that are induced by better environmental quality. Less working days are lost if health improves, which increases output per unit of input. Health care costs decrease, alleviating private and government budgets. Substantial costs can be avoided if climate change, such as an increase in sea level (which directly affects agriculture, urban areas in coastal regions, and dike maintenance costs) and the loss of harvest in agriculture due to draughts or changes in precipitation, is mitigated.

5.4 Confronting Costs and Benefits

Most empirical research suggests that improvements in environmental quality require costly regulation, despite the various general equilibrium effects on efficiency and productivity that may mitigate the direct costs of regulation. The focus on costs of environmental regulation may distract attention from its benefits: These costs are incurred to achieve environmental benefits. It is worthwhile to incur the costs if the benefits outweigh the costs in a social cost–benefit analysis.

The social benefits from environmental improvements are often major, but they are less visible than the costs of environmental regulation. Measuring the benefits requires putting a value on improvements in environmental quality, health, and environmental amenities. When questioned in surveys, individuals indicate that they are willing to give up large amounts of money and income to pay for these improvements. We can also infer individuals' valuation of environmental amenities from the amount of money individuals spend on traveling to locations with

environmental amenities or the premium people are willing to pay on houses located near them. The data from these studies can be used to construct an aggregate indicator of the value of environmental improvements, which can then be compared to the cost estimates to undertake a social cost–benefit analysis.

The problem that the value of environmental amenities is unobservable (that there are no market prices) can thus be solved. However, perceptions may still be biased. Since neither market transactions nor market prices are associated with environmental goods, the current system of national accounts does not include the value of environmental goods; it only measures the value of marketed goods. As a result, the environmental benefits that environmental policies create are not part of national income, and environmental improvements do not count as economic progress. Attempts have been made to adjust national income data so as to include imputed values of environmental improvements and degradation. These "green GDP" data would more appropriately measure well-being and progress than do current GDP data. As long as these adjustments are not standard, there seems to be a trade-off between environment and growth. Once we take into account the value of environment for well-being, however, there is a close positive correlation between environment and economic progress.

SEE ALSO THE FOLLOWING ARTICLES

Cultural Evolution and Energy • Development and Energy, Overview • Economic Growth and Energy • Energy Ladder in Developing Nations • Equity and Distribution in Energy Policy • Global Energy Use: Status and Trends • Market-Based Instruments, Overview • Market Failures in Energy Markets • Population Growth and Energy • Sustainable Development: Basic Concepts and Application to Energy • Technology Innovation and Energy • Transitions in Energy Use

Further Reading

Bovenberg, A. L., and Goulder, L. H. (2002). Environmental taxation and regulation. *In* "Handbook of Public Economics" (A. Auerbach and M. S. Feldstein, Eds.), Vol. 3. North-Holland, Amsterdam.

Jaffe, A. B., Peterson, S. R., Portney, P. R., and Stavins, R. N. (1995). Environmental regulation and the competitiveness of U.S. manufacturing: What does the evidence tell us? *J. Econ. Literature* **33,** 132–163.

Krautkraemer, J. (1998). Nonrenewable resource scarcity. *J. Econ. Literature* **36,** 2065–2107.

Lieb, Chr. M. (2003). "The environmental Kuznets curve—A survey of the empirical evidence and of possible causes, Study No. 391." Interdisciplinary Institute of Environmental Economics, University of Heidelberg, Heidelberg, Germany. [Available at http://www.uni-heidelberg.de].

Schulze, G., and Ursprung, H. (eds.). (2001). "International Environmental Economics; A Survey of the Issues." Oxford Univ. Press, Oxford, UK.

Simpson, D., and Toman, M. (eds.). (2004). "Scarcity and Growth in the New Millennium." Resources for the Future, Washington, DC.

Smulders, S. (2000). Economic growth and environmental quality. *In* "Principles of Environmental Economics" (H. Folmer and L. Gabel, Eds.), pp. 602–664. Elgar, Cheltenham, UK.

Economics of Energy Demand

KENNETH B. MEDLOCK III
Rice University
Houston, Texas, United States

1. **The Demand for Energy: An Introduction**
2. **Modeling Energy Demand: An Analytical Framework**
3. **Modeling Energy Demand: Empirical Analysis**

Glossary

asymmetry In the present context, this refers to the asymmetric effect of changes in income or price on energy demand. Specifically, changes in energy demand caused by a Δ increase in income or price will not necessarily be of equal magnitude to changes caused by a Δ decrease in income or price (see also imperfect price-reversibility).

dematerialization The reduction of raw material (energy and material) intensity of economic activities, measured as the ratio of material (or energy) consumption in physical terms to gross domestic product in deflated constant terms.

economic structure The structure of production and consumption in an economy. Specifically, the share of goods and services produced and consumed by various economic sectors.

energy security The concept of maintaining a stable supply of energy at a "reasonable" price to avoid the macroeconomic costs associated with interruption of energy supply or increases in energy price.

imperfect price-reversibility In the present context, this refers to the asymmetric effect of changes in price on energy demand. Specifically, changes in energy demand caused by a Δ increase in price, though opposite in impact, will not necessarily be of equal magnitude to changes caused by a Δ decrease in price.

income elasticity of energy demand The percentage change in energy demand caused by a 1% change in income, holding all other factors constant.

price elasticity of energy demand The percentage change in energy demand caused by a 1% change in price, holding all other factors constant.

primary energy requirements The quantity of energy consumed by the end user (total final consumption) plus the losses incurred during the conversion, transmission, and distribution.

rental price of capital The opportunity cost of purchasing/owning capital.

total final consumption The total quantity of energy consumed at end use.

user cost of capital The rental price plus the cost of operating and maintaining capital.

The demand for energy is a *derived* demand. Energy's value is assessed not by its intrinsic appeal, but rather by its ability to provide a set of desired services. When combined with energy-using capital equipment, energy provides the work necessary to produce goods and services in both industry and in the household. As such, energy consumption, whether in the firm or household, is the result of a set of simultaneous decisions. Energy consumers must determine the quantity and type of capital equipment to purchase, where type is differentiated by technological characteristics such as efficiency and fuel input, and they must also determine a rate of capital utilization. Thus, the quantity of energy consumed is a function of capital utilization, as well as the engineering characteristics of capital equipment.

1. THE DEMAND FOR ENERGY: AN INTRODUCTION

To the extent that technologies make it possible, consumers of energy can choose between different sources of energy, and, in fact, energy use by source varies greatly across countries. In industrialized countries, the majority of energy consumed is referred to as commercial energy, which is differentiated from traditional forms of energy. The distinction between commercial and traditional energy can be made as an economic one. Specifically, commercial energy (for example, oil, natural gas, coal, and electricity) is traded in broader, larger-scale markets, whereas, traditional energy (for example,

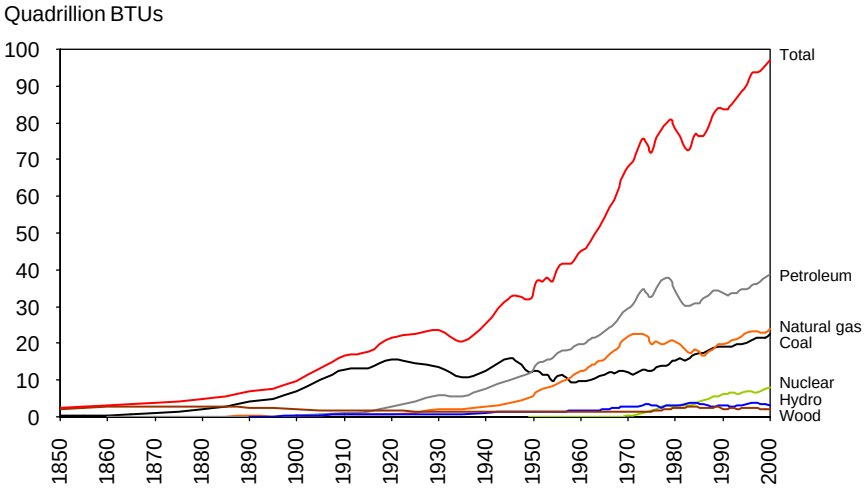
 © 2004 Elsevier Inc. All rights reserved.

waste and other biomass) is usually only traded, if at all, in smaller local markets. Typically, only in less developed countries will there be large shares of total energy consumed in traditional forms, since the proliferation of stocks of energy-using capital equipment, an occurrence associated with rising incomes, begets a substitution to commercial energy.

On a global scale, the 20th century was dominated by the use of fossil fuels. According to the U.S. Department of Energy, in the year 2000 global commercial energy consumption consisted largely of petroleum (39%), followed by coal (24%), and natural gas (23%), with hydro (6%) and nuclear (7%) splitting the majority of the remainder. However, this breakdown has not been constant over time either globally or within nations. In the United States, for example, wood was the primary source of energy consumption prior to industrialization (see Fig. 1). At the onset of industrialization, however, fossil fuels became the dominant energy source, and a number of factors, such as relative prices and environmental considerations, have dictated the composition of energy use since.

1.1 Energy Demand: A Context for Discussion

Energy is crucial to the economic progress and social development of nations. The development of new technologies and emergence of new production processes during the 19th and 20th centuries enabled an increase in productivity through the substitution of mechanical power for human and animal power. This, in turn, led to unprecedented economic growth and improved living standards, as well as, of course, large increases in energy consumption. For example, the consumption of coal and crude oil products for the locomotive, and later the automobile, significantly reduced the cost of transportation of people and goods, thereby creating growth opportunities for businesses and entrepreneurs. In addition, electrification has played a pivotal role in social development and welfare by making possible greater access to information via radio and television, providing a cleaner, more efficient means of storing and preparing food, and controlling enclosed environments to provide warmth and cooling. In the industrialized



Shares of Total Energy by Select Year

1850	1900	1950	2000
Coal: 9.3%	Coal: 71.4%	Coal: 35.6%	Coal: 23.0%
Petroleum: 0.0%	Petroleum: 2.4%	Petroleum: 38.5%	Petroleum: 39.1%
Natural gas: 0.0%	Natural gas: 2.6%	Natural gas: 17.2%	Natural gas: 24.5%
Wood: 90.71%	Wood: 21.0%	Wood: 4.5%	Wood: 2.3%
Hydro: 0.0%	Hydro: 2.6%	Hydro: 4.2%	Hydro: 3.1%
Nuclear: 0.0%	Nuclear: 0.0%	Nuclear: 0.0%	Nuclear: 8.0%

FIGURE 1 Energy demand in the United States increased fairly consistently from 1850–2000, but the source of that energy continually changed. Initially, wood was the primary form of energy use, but, driven by structural and technological change, coal, and later petroleum and natural gas, took the majority share of energy demand as the U.S. economy expanded. From the U.S. Energy Information Administration.

society of the 21st century, energy is necessary lest modern economic activity would cease and modern standards of living would diminish. Moreover, in developing nations, a lack of availability of energy supplies will prohibit further economic and social development.

Due to the importance of energy to economic and social progress, it is difficult to separate a discussion of energy demand from energy policy. In fact, a reoccurring theme in energy economics in general has been that of energy security. According to Bohi, Toman, and Walls, "Energy security refers to loss of economic welfare that may occur as a result of a change in the price or availability of energy" (p. 1), thus suggesting a link between energy and economy. The energy-economy link is supported by an abundance of empirical research demonstrating a strong correlation between increases in oil prices and decreases in macroeconomic performance, for both the United States and many other oil-importing industrialized countries. In fact, almost every recession in the post–World War II era in the United States, as well as many other energy-importing nations, has been preceded by a spike to the price of energy (see, for example, Hamilton; Ferderer; and Mork, Mysen, and Olsen). Because of this, governments, especially in those countries that are dependent on imported energy resources, have been concerned with maintaining security of supply at a reasonable price.

Projections of energy demand are necessary to develop national energy policies aimed at minimizing exposure to international price fluctuations and maximizing energy availability. On a less global scale, projections of energy demand are also important to the development of corporate business plans and the establishment of operational parameters for public utilities that must develop strategies to fulfill public service obligations. Therefore, the modeling and projection of energy demand is very important in the formation of policy in general, at both the national and local levels as well as in the private sector. Sound energy policy requires an adequate understanding of the economics of energy.

The economics of energy are generally concerned with the manner in which both nonrenewable and renewable energy resources are allocated over time and across space. A resource is considered nonrenewable when the sum, over time, of the services it provides is finite, or at least not replaceable in a reasonable time frame. Energy economists have been heretofore concerned primarily with the allocation of nonrenewable resources, largely because global energy consumption has consisted primarily of the consumption of fossil fuels. Dating back to Hotelling's classic paper, there have been numerous theoretical studies of optimal depletion rates and associated pricing rules given different sets of assumptions, as well as empirical studies designed to test theory. The well-known r-percent rule, as well as an apparent disregard for the impact of still underappreciated factors such as technological change in the extractive industries, the development of lower cost alternatives, and expansion of the resource base through exploration activities, influenced predictions by many scholars and industry analysts in the late 1970s and early 1980s that oil prices would top $100 per barrel by 2000. This projection never occurred.

Also of considerable influence to these erroneous projections was an apparent extrapolation of demand trends without properly considering the impact of allowing for, among other things, structural change in economies as they develop and the effect of increased energy prices on the composition and characteristics of capital stocks. In general, as economies develop they move from agricultural to industrial to service economies. This structural transformation has important implications for the manner in which energy demand grows relative to income. Specifically, as less energy intensive activities make up a larger proportion of total economic activity the income elasticity of demand will decline. In short, the income elasticity of nations is not constant.

The characteristics of energy-using capital stocks are also an important determinant of energy demand that cannot be ignored when forecasting. When considering an entire economy, the type of capital employed by firms and in households is largely dependent on economic structure, but certain characteristics of capital, such as embodied technology, are sensitive to price. Therefore, price-induced technological change will alter the energy requirements for a given level of output, allowing for an expansion of economic activity without an appreciable increase in energy demand. To illustrate, consider the United States transportation sector. According to the Energy Information Administration, the average fuel rate of a passenger automobile increased from about 14.3 miles per gallon in 1978 to about 20.2 miles per gallon in 1990, an average annual increase of 2.9%. During the same time period, transportation fuel consumption increased at only 0.7% per year, despite increases in both motor vehicle stocks and motor vehicle utilization (miles driven per vehicle). Clearly, efficiency gains offset increased utilization and growing vehicle stocks thus

limiting the increase in fuel consumption. Apart from allowing output to increase for a given level of fuel consumption, an increase in the energy efficiency of capital stocks also leads to asymmetry in the response of energy demand to changes in price, as noted in Gately and Huntington. Specifically, if certain capital investments aimed at increasing energy efficiency are made in response to a price increase, they will not be undone when price falls. Thus, energy demand tends to respond differently to price increases than to price decreases.

1.2 Elasticity of Energy Demand

The experiences of the 1970s and 1980s led to an explosion in the number of energy demand studies, a trend that has been somewhat revitalized by the emergence of concerns about the emissions of greenhouse gases from the combustion of fossil fuels. A large majority of these studies have resulted in the estimation of, or made use of previously estimated, income and price elasticities of energy demand. Income and price elasticities of energy demand are of the utmost importance to forecasting energy demand. Moreover, the manner in which these relationships may or may not evolve in response to future economic growth, changes in economic structure, and various policy initiatives is also important. Income and price elasticities are, in general, the rules of thumb by which broader energy policy is directed. There is, however, no real consensus as to an appropriate value of income or price elasticity. Much of the disagreement in empirical research is certainly due to different model specifications, but it is also due to some confounding issues that are difficult to isolate, such as the effects of changes in technology, changes in capital stock utilization, capital stock turnover, shifts in economic structure, and changes in energy policy.

The income elasticity of energy demand is defined as the percentage change in energy demand given a 1% change in income holding all else constant, or

$$\varepsilon_y = \frac{\%\Delta e}{\%\Delta y} = \frac{de}{dy} \cdot \frac{y}{e},$$

where e denotes energy demand and y denotes income. This measure provides an indication of how demand will change as income changes. The income elasticity of energy demand had long been reported as being greater than or near to 1. Research, however, has indicated that this may be greatly overstated for industrialized countries, as there is increasing evidence that energy intensity is inversely related to economic development, being greatly influenced by changes in economic structure and the state of technology.

The price elasticity of energy demand is defined as the percentage change in energy demand given a 1% change in price holding all else constant, or

$$\varepsilon_p = \frac{\%\Delta e}{\%\Delta p} = \frac{de}{dp} \cdot \frac{p}{e},$$

where p denotes the price of energy. This measure gauges the influence of energy price on energy demand and is often used as a barometer to assess the impact of various conservationist policies, such as emissions abatement programs or energy tax/subsidy schemes. For example, one can approximate the reductions in emissions achieved by taxing energy if the impact of that tax on the quantity and composition of energy consumption is known. The price elasticity can also be used to determine the impact of energy price disruptions on energy demand and, hence, any subsequent loss in macroeconomic output. However, the long- and short-run effects of energy prices can be difficult to disentangle. Energy demand is predicated on ownership of an energy-using capital good, the owner's decision as to the energy efficiency of the capital good, which is made when the capital is purchased, and utilization of the capital good. Once a capital good has been purchased, the capital utilization decision will dictate the responsiveness of energy demand to price in the short run, assuming the short run implies that capital and technology are fixed so that the consumer cannot "trade up" to higher energy efficiency. In the long run, however, the consumer is free to turn over capital equipment in favor of higher energy efficiency. Thus, the long-run response to price involves a decision on both capital utilization rates and energy efficiency.

There is an important consideration that bears mention here with regard to calculating income and price elasticity of energy demand. When estimating an income elasticity of energy demand, for example, it is not sufficient to simply divide the percentage change in energy demand by the percentage change in income. Nor it is not sufficient to simply use linear regression analysis to estimate an income elasticity of energy demand if income is the only explanatory variable. Both of these approaches completely ignore the influence that other variables may have on energy demand, which can lead to erroneous, statistically biased results. Nonetheless these are common mistakes often made by well-meaning energy analysts.

To illustrate, recall that income elasticity of energy demand is the percentage change in energy demand given a 1% change in income, *holding all else constant*. If we take energy demand to be a function of income, price, and possibly some other set of variables, then, in general, the total derivative of energy demand is given as

$$de = \frac{\partial e}{\partial y} dy + \frac{\partial e}{\partial p} dp + \cdots .$$

If we hold all factors but income constant, the preceding equation reduces to

$$de = \frac{\partial e}{\partial y} dy.$$

We can then multiply both sides by

$$\frac{1}{dy} \cdot \frac{y}{e}$$

to yield

$$\frac{de}{dy} \cdot \frac{y}{e} = \frac{\partial e}{\partial y} \cdot \frac{y}{e}.$$

The left-hand side of this equation is the income elasticity of energy demand, as defined earlier, but we only arrive at this result when all variables except income are held constant. It is a simple matter to show that not holding all other variables constant will result in a remainder term on the right-hand side of the preceding equation (i.e., the term in brackets):

$$\frac{de}{dy} \cdot \frac{y}{e} = \frac{\partial e}{\partial y} \cdot \frac{y}{e} + \left\{ \frac{\partial e}{\partial p} \cdot dp \cdot \frac{1}{dy} \frac{y}{e} + \cdots \right\}.$$

If, therefore, we seek an estimate of income elasticity, we cannot ignore the other relevant variables as it introduces a bias similar to the term in brackets. For example, if energy consumption has been increasing at 2% per year for 10 years, and income has been increasing at 3% per year for the same time period, a naïve approximation of income elasticity of energy demand would be 0.67 (= 2%/3%). However, if we consider that price may have been changing during the past 10 years, then our naïve estimate is incorrect. Specifically, if price had been falling, then, given a downward sloping demand curve, an income elasticity of 0.67 is an overestimate and can lead to serious problems when forecasting future demand. Therefore, when modeling energy demand, it is important to recognize that many variables influence energy use, and those variables must be carefully accounted for. This is precisely why a sound theoretical foundation is necessary when analyzing energy demand.

1.3 A Note on Energy Accounting

Before discussing the aspects of energy demand modeling about which the researcher must be conscious, it is important to highlight exactly what is referred to when discussing energy demand. When modeling energy demand, the researcher is usually interested in the quantity of energy required to perform a specified activity. For example, operating a drill-press requires electricity, driving an automobile requires gasoline or diesel fuel, and heating a household requires natural gas, heating oil, or electricity. Simply accounting for the energy consumed at end use, however, does not fully account for all the energy consumed to make the activities listed above possible. Specifically, raw energy supplies (such as coal, crude oil, and wet natural gas), while they can be directly consumed, are generally converted into products (such as gasoline, heating oil, propane, ethane, dry natural gas, and electricity) that are then consumed in various end uses. The energy conversion industries, such as refining, processing, and electricity production, use primary energy inputs to produce usable products. Energy balance tables are a means of accounting for the process of energy conversion and eventual consumption. Figure 2 illustrates the general flow that energy balance tables capture. Total primary energy requirements (TPER) are the energy needed to produce the energy necessary for total final consumption (TFC). TPER is greater than TFC due to the energy expended in the conversion, transmission, and distribution activities.

2. *MODELING ENERGY DEMAND: AN ANALYTICAL FRAMEWORK*

As stated earlier, the demand for energy is derived from the demand for goods and services provided by energy-using capital equipment. In residences and commercial establishments, energy is consumed, in conjunction with some energy-using durable good, to provide services such as space heating and cooling, lighting, cooking, and cleaning. Similarly, in industry, energy is used to produce steam and power automated processes, and in transportation, energy is required to make use of the services provided by automobiles, airplanes, trains, and so on. Therefore,

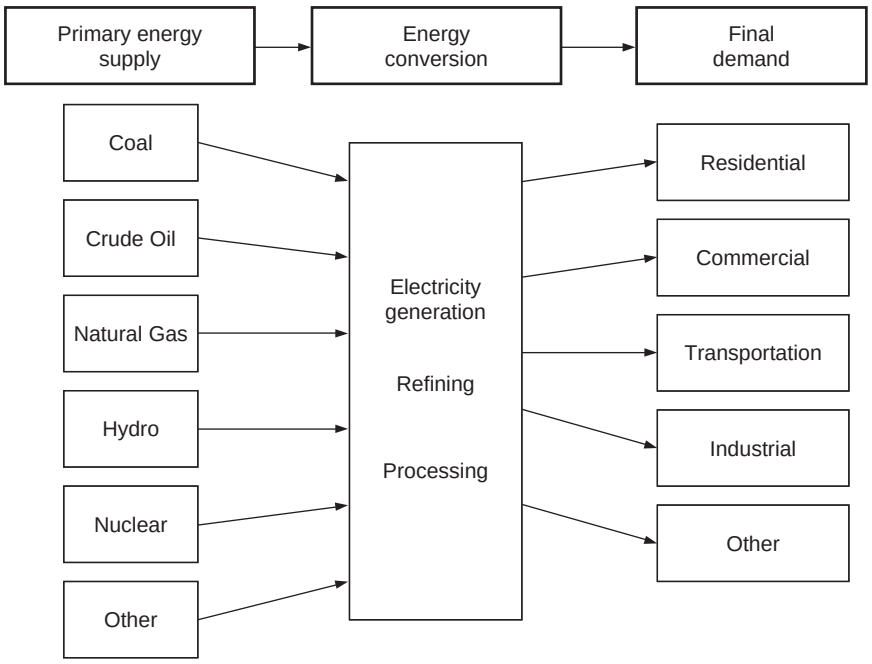


FIGURE 2 Primary energy supply is the energy required to meet final demand. Some energy is expended in the energy conversion industries in order to make usable energy products, so primary energy supply is greater than final demand.

implicit in the decision to consume energy is an underlying set of decisions that each energy consumer faces. Consumers of energy must decide to purchase the capital equipment that provides the desired service (heating, cooling, transportation, and manufacturing), they must choose the characteristics of the capital equipment (energy efficiency and type of energy input), and they must choose the rate at which the capital will be utilized. These decisions are conditional on, among other things, the energy cost of capital utilization. It is this particular aspect of demand modeling that is sometimes difficult to capture. In the short run, capital stocks are assumed fixed, so energy price fluctuations will influence the utilization rate of capital. In the long run, the energy consumer can replace capital equipment thereby improving energy efficiency. In empirical studies, this can pose difficulties due to limited time series data (both in length and breadth), thereby confounding the differentiation between the long- and short-run impacts of energy price movements. It is therefore important to develop a theoretical framework that is capable of addressing these concerns in order to lend guidance to assessing and conducting empirical analysis.

The quantity of energy consumed in the production of goods and services is dependent on the technical nature of energy-using capital equipment. The relationship between energy consumption and capital equipment can be given as

$$\text{energy} \equiv \left(\frac{\text{capital utilization rate}}{\text{capital equipment efficiency}}\right) \times \text{capital stock}, \quad (1)$$

where efficiency is an equipment-specific technological parameter, defined to be service output per unit energy. It is important to note that this relationship is, in general, equipment specific, and there is one such relationship for each type of capital equipment. This is important because it is an oversimplification to aggregate capital stocks and energy if one wishes to analyze the relationship between capital and energy. In the case of motor vehicles, efficiency is miles per gallon, and in the case of central air conditioners, efficiency is British thermal units (BTUs) per watt-hour. As efficiency increases, the energy required per unit of service declines. The capital utilization rate denotes the service rendered by capital equipment, such as miles driven per vehicle, thus energy use is positively related to capital utilization. Equation (1) is an identity because its units simplify to energy≡energy.

Notice that the right-hand-side of Eq. (1) contains the three choice variables involved in any decision to consume energy. Because capital stocks and technology are key determinants of energy use, static models of energy demand are limited in their ability to

capture the full response of energy use to changes in economic variables, such as income and price. Nevertheless, the static model built on sound microeconomic principles provides a reasonable starting point. As such, and because static models are widely used, we will begin with a brief description of them before moving to a dynamic framework.

2.1 Static Models

Following the energy balance structure indicated in Fig. 2, the demand for energy originates in different end-use sectors: industrial, residential, commercial, transport, and other. More generally, energy consumption can be modeled as either providing utility to households or as an input in the production process for firms. For treatment of the former problem, it is generally assumed that a consumer maximizes his or her welfare by consuming some bundle of goods and services subject to a budget constraint. Energy is assumed to be proportional to the service it provides, and therefore consumer welfare is influenced by energy consumption. (As demonstrated later, if one wishes to allow for substitution possibilities across types of energy commodities, it is a simple matter to incorporate various types of energy along with their specific costs.) The resulting demand for energy by the representative individual in this problem will be influenced by a number of factors, including but not limited to energy price, the price of other goods, and income (other variables, such as weather and durable goods stocks, can also have an influence, depending on the model specifications).

Expressing the preceding argument mathematically, we have that a representative consumer maximizes utility, $U(z, e)$, which is function of energy consumption, e, and all other consumption, z, subject to the constraint that expenditures cannot exceed income, y. As usual, we assume utility is a concave twice-differentiable function of its arguments. If the modeler wishes to examine the substitution possibilities across energy products, the energy variable can be a vector of n energy products, $e = [e_1, e_2, \ldots, e_n]$. Allowing the price of good j to be represented as p_j, the consumer is assumed to

$$\max_{z,\, e_1, \ldots, e_n} U(z, e_1, \ldots, e_n)$$

subject to $y \geq p_z z + p_{e_1} e_1 + \cdots + p_{e_n} e_n$.

The first-order necessary conditions for a maximum for this problem can be solved to yield demand equations for each of the energy products and for all other consumption. The resulting function for energy demand is given as

$$e_j^* = e_j(y, p_z, p_{e_1}, \ldots, p_{e_n}), \tag{2}$$

where the star denotes an optimal choice. Thus, the demand for energy commodity j is a function of income, own price, the price of all other energy commodities, and the price of all other consumption goods.

A firm also demands energy to produce goods and services. The typical firm is interested in maximizing profits, or minimizing the cost of producing a given level of output, and it must purchase energy as well as all other inputs to production. As it turns out, the firm's problem is not unlike the household's problem. The firm's demand for energy in this problem can be shown to be a function of its output, the price of energy, and the price of all other inputs.

Formalization of this problem is as follows. Consider a representative firm that minimizes costs, c, for a given level of output, $\bar{q}$. Output is a function of capital (k), materials (m), labor (l), and energy, $q = f(k, l, e_1, \ldots, e_n, m)$, and costs are the sum of payments to the factors of production, $c = rk + wl + p_{e_1} e_1 + \cdots + p_{e_n} e_n + p_m m$, where r are rent payments to capital, w are wages paid to labor, and p_j is the price of factor j. The firm's problem is then stated as

$$\min_{k, l, e_1, \ldots, e_n, m} c$$

subject to

$$\bar{q} = f(k, l, e_1, \ldots, e_n, m)$$

$$c = rk + wl + p_{e_1} e_1 + \cdots + p_{e_n} e_n + p_m m.$$

The reader can verify that the solution to this problem is not very different than the solution to the utility maximization problem, other than a change of terminology. (In fact, the dual to the household's [firm's] problem is one of cost minimization [profit maximization].) The firm will choose inputs such that their marginal values in production will be equal. Solution of the first-order conditions for a maximum for this problem yields a general function for the demand for energy

$$e_j^* = e(\bar{q}, r, w, p_{e_1}, \ldots, p_{e_n}, p_m).$$

Thus, by Eq. (3), the firm weighs the price of all possible factor inputs when choosing its energy consumption for a given level of output.

These problems are static representations of energy demand. As such, they are most appropriate in deriving short-run demand functions, when capital and technology are fixed. Comparative statics can be a useful tool in this setting for analyzing the impact of a rise in capital stocks, say, but they disregard the

intertemporal choices an energy consumer faces with regard to type of capital, utilization of capital, and efficiency of capital. A more general approach would allow for a dynamic specification of demand. A dynamic model accounts for intertemporal choice of capital stock characteristics and utilization, as well as the effects of each on energy demand. As such, a dynamic framework is capable of addressing the asymmetries that exist in the energy-income and energy-price relationships. In the static setting such explorations are left to pontification and conjecture.

2.2 Dynamic Models

Dynamic models of capital ownership and energy use allow for a more complete analysis of energy demand because, for one, they are capable of capturing factors that generate the asymmetries mentioned previously. In addition, dynamic models incorporate the intertemporal choices that a consumer (firm) must make when maximizing welfare (profits) over some time horizon. Medlock and Soligo (2002) developed a dynamic model of motor vehicle ownership and fuel use that generalizes to many other types of energy-using capital equipment. They showed that the demand for energy is determined by the consumer's wealth, the price of energy, the price of other goods, the user cost (or cost of owning and maintaining) capital, and the energy efficiency of the energy-using capital stock. Because the model they developed is dynamic, it captures the notion that the decision to consume energy is made jointly with the decision to purchase and maintain a stock of energy-using capital equipment, the latter of which incorporates the elimination of potential intertemporal arbitrage opportunities.

It is useful to consider the framework developed by Medlock and Soligo (2002). To begin, consider a modification of the household's problem. For ease of exposition, energy consumption has been denoted as a single commodity. Relaxation of this constraint allows for multiple types of capital and hence multiple types of energy consumption. Denoting time using the subscript t the consumer will maximize the discounted sum of lifetime utility,

$$\sum_{t=0}^{T} \beta^t U(z_t, e_t),$$

subject to the constraint that capital goods purchases (i_t), purchases of other goods (z_t), purchases of energy (e_t), and savings (s_t) in each period cannot exceed this period's income (y_t) plus the return on last period's savings ($(1+r)s_{t-1}$). It is assumed that capital goods depreciate at a rate δ, savings earn a rate return r, the discount rate is such that $0<\beta<1$, and all initial conditions are given. (Note we can complicate the problem somewhat be allowing access to credit markets, but this adds little to the present context.)

The problem is thusly formulated. Consumers will

$$\max_{z,e,s} \ \sum_{t=0}^{T} \beta^t U(z_t, e_t)$$

subject to

$$\begin{aligned} & p_{z,t}z_t + p_{e,t}e_t + p_{k,t}i_t + s_t \le y_t + (1+r)s_{t-1} \\ & e_t \equiv \frac{u_t}{\varepsilon_t} k_t \\ & i_t = k_t - (1-\delta)k_{t-1} \\ & z_t, u_t, k_t \ge 0 \quad t = 1, \ldots, T \end{aligned}$$

where $p_{j,t}$ is the price of good j and all initial values are given. Note the second constraint is a restatement of Eq. (1). It is through this constraint that the relationship between energy and capital is made explicit.

The marginal conditions for this consumer's problem lend considerable insight into the manner in which consumers will respond to changes in income and energy prices. Specifically, the marginal conditions yield

$$U_k \frac{u_t^*}{\varepsilon_t} = U_z \left(p_{e,t} \frac{u_t^*}{\varepsilon_t} + p_{k,t} - p_{k,t+1} \left(\frac{1-\delta}{1+r} \right) \right), \quad (4)$$

where the star denotes the optimality. Equation (4) indicates that the consumer will allocate income such that the marginal value of the energy services attained from the capital stock is equal to the marginal value of consumption of all other goods. The term in brackets in Eq. (4),

$$p_{e,t} \frac{u_t^*}{\varepsilon_t} + p_{k,t} - p_{k,t+1} \frac{(1-\delta)}{(1+r)},$$

is defined as the "user cost," μ_k, of capital stock, and it is this term that bears significant mention. The first term, $p_{e,t}\frac{u_t^*}{\varepsilon_t}$, indicates that the consumer chooses user cost to the extent that capital utilization is a choice variables. If the optimal choice of utilization were such that $u_t^* = 0$ user cost reduces to

$$p_{a,t} - p_{a,t+1} \frac{(1-\delta)}{(1+r)},$$

which, following Diewert, is the rental price $(p_{R,t})$ of owning capital equipment.

Simultaneous solution of the first-order conditions will yield a set of demand equations, one for each of the arguments in the utility function at each of the

$t+1$ time periods. By Eq. (1), once a solution is obtained for k_t^* and u_t^*, we can solve for the optimal consumption of energy. Thus, the optimal choice of energy demand is determined by the optimal choice of capital stocks and capital utilization, with efficiency given. The demand for energy is found to be a function of income, the user cost of capital, and the price of all other goods. Since user cost is a function of energy price, energy efficiency, and the rental price of capital, we can generically state the demand for energy as

$$e_t^* = e\left(y_t, p_{z,t}, p_{e,t}, p_{R,t}, \varepsilon_t\right). \tag{5}$$

Note that Eq. (5) is altered if efficiency is taken to be choice variable, although a distinction must be made between the long and short run. This is not done here for ease of exposition, as it would require the differentiation of types of possible capital purchases by the use of additional constraints. (One such example is Tishler.) The added insight gained from allowing a choice across multiple energy efficiencies is that consumers may choose to increase efficiency if energy prices rise in order to prevent the user cost of capital from rising. This type of response allows for the previously mentioned asymmetries.

We could also increase the complexity of the problem by allowing for multiple types of capital equipment (a household typically has refrigerators, air conditioners, ovens, etc.) and multiple fuel sources, which would result in, for example, natural gas prices appearing in, say, an electricity demand function. Again, however, this is not done here for ease of exposition. Such analysis is left to the reader for further exploration.

3. MODELING ENERGY DEMAND: EMPIRICAL ANALYSIS

Empirical studies often use the general form of Eq. (2) or Eq. (5) to specify a function to be estimated as an approximation of the "true" demand function. This can lead to substantial variation in published results since parameter estimates are sensitive to the model specification. This is not an unimportant fact since policy recommendations are often made on the basis of empirical research. Moreover, it highlights the importance of sound theoretical justification for model specification.

A simple and popular demand specification is the log-linear demand equation given as

$$\ln e_t^* = a + \alpha \ln y_t + \eta \ln p_{e,t}, \tag{6}$$

where e^* is the "optimal" quantity of energy demanded, y is income, p_e is energy price, and a is a constant. The subscript t denotes the time period. (Such a demand specification implies a demand function of the form $e_t^* = A y_t^{\alpha} p_{e,t}^{\eta}$). Equation (6) is easily estimated to directly yield an income elasticity, α and a price elasticity, η. Because energy demand is not necessarily a function of only income and price, additional parameters are also sometimes included in the specified demand equation. For example, when estimating energy demand across countries, a *per capita* income measure is generally preferred because it normalizes income with respect to population. Population may also be added explicitly as an independent variable, as well as other parameters of importance such as capital stocks, weather/climate, and the prices of alternative fuels, the latter two of which are particularly important when modeling the demand for individual fuel types whose consumption varies seasonally.

An alternative specification to Eq. (6) allows energy demand to be a function of past realizations of income and price. It can be shown that this is equivalent to assuming the existence of a lag adjustment mechanism of the form

$$\left(\ln e_t - \ln e_{t-1}\right) = \gamma\left(\ln e_t^* - \ln e_{t-1}\right),$$

where $\gamma \in [0, 1]$ is the speed of adjustment. Eliminating the term $\ln e^*$ in Eq. (6) by direct substitution results in the equation

$$\ln e_t = \gamma a + \gamma\alpha \ln y_t + \gamma\eta \ln p_{e,t} + (1-\gamma)\ln e_{t-1}, \tag{7}$$

where α is the long-run elasticity of demand with respect to income, η is the long-run elasticity of demand with respect to price, and $\gamma\alpha$ and $\gamma\eta$ are the respective short-run elasticities.

While econometric specifications such as Eqs. (6) and (7) are attractive because they are easily estimated and yield elasticity terms directly, they implicitly assume that elasticity is constant across all ranges of income and price, a point that has been challenged in the literature. Moreover, it does not adequately capture the influence of changes in the composition of capital stocks over time, a phenomenon that could alter energy required per unit output and result in asymmetric demand responses to changes in income and price. One means of controlling for these factors directly is to estimate the components of Eq. (1) in order to derive an estimate of energy demand for a specific end use. Olof and Schipper take such an approach with regard to forecasting long-term motor fuel demand in several

countries. The advantage of such an approach is that it is robust to the simultaneous set of decisions the consumer makes when determining energy consumption.

Other adjustments can and have been made to Eqs. (6) and (7). For example, one can employ dummy variables on the price and income coefficients, thereby separating price and income effects into increases and decreases, in order to capture any asymmetric response of energy demand. One can also specify a demand function that is nonlinear in all or some of the included variables. Such a specification could allow, for example, a declining income elasticity of energy demand. Still other approaches involve an entirely different approach to deriving an equation for energy demand. This class of models involves estimating the first-order conditions that would arise from a theoretical model similar to that presented earlier. Thus, rather than using general solutions such as those given in Eqs. (2) or (5), consumer utility functions or production functions are to be specified, and analytical solutions are obtained. While potentially cumbersome mathematically, this type of approach has the advantage of including all relevant variables, and the interactions therein, from theoretical framework in a simultaneous equations setting.

In sum, there are numerous specifications that can be potentially explored when analyzing energy demand. But it must be noted that the different types of alternative specifications are tailored to identify specific characteristics of the data being analyzed and are usually specific to the question posed by the researcher. While no one specification is necessarily more correct than another, some are certainly more adequate than others in capturing certain aspects of energy demand, and thus they highlight what should, and sometimes should not, be considered in empirical modeling.

Regardless of the specification chosen, a strong understanding of the economic forces at work in individual markets is crucial to empirical analysis. For instance, when modeling the near-term demand for natural gas, it is important to account for both weather and alternative fuel prices (heating oil for example), as well as income and own price. This is due to the fact that natural gas demand increases considerably in the winter months as heating demand increases. Therefore, for reasons discussed earlier, not accounting for weather and alternative fuel prices will lead to biased estimates of income and price elasticity.

3.1 Empirical Considerations: The Energy-Income Relationship

3.1.1 Economic Development and Dematerialization

The assumption of constant income elasticity of energy demand is inconsistent with empirical evidence that energy intensity increases then decreases during the course of economic development. In general, econometric analysis of countries across time and space yield the result that the income elasticity of energy demand is higher in lower income countries. The energy intensity of an economy resembles an inverted U-shape across increasing levels of per capita income, as discussed in 2001 by Medlock and Soligo. This arises because various sectors grow at different rates and at different times during the course of economic development. Industrialization results in large increases in commercial energy use. This is followed by smaller increases in energy demand during the post-industrial phase of development, in which the service sector grows relative to all other sectors. Figure 3 illustrates the general pattern of energy intensity that arises during the course of economic development, as well as the general relationship between energy intensity and income elasticity of energy demand. The reader can readily verify that if the income elasticity of energy demand exceeds unity, then, holding all else constant, energy intensity is rising. Similarly, if the

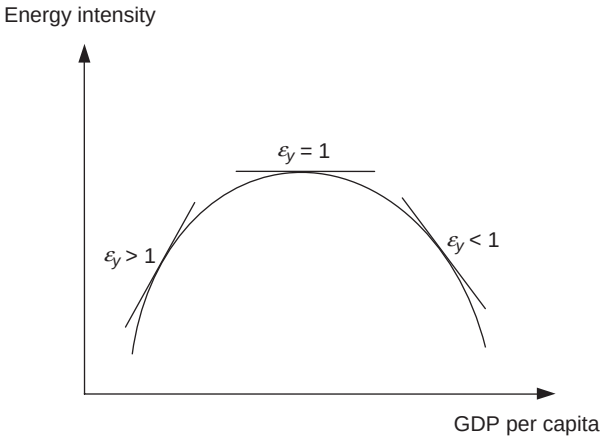


FIGURE 3 Generally, due to structural transformation and technological change, energy intensity follows an inverted U-shape during the course of economic development. This has implications for income elasticity. When income elasticity of energy demand exceeds unity, then, holding all else constant, energy intensity is rising. Similarly, if the income elasticity of energy demand is less than unity, then, holding all else constant, energy intensity is falling.

income elasticity of energy demand is less than unity, then, holding all else constant, energy intensity is falling. Note that declining energy intensity does not imply that energy demand declines, only that energy demand grows more slowly than output.

A rising then declining energy intensity is generally consistent with the theory of dematerialization, which was defined by Bernardini and Galli in 1993 as "the reduction of raw material (energy and material) intensity of economic activities, measured as the ratio of material (or energy) consumption in physical terms to gross domestic product in deflated constant terms." There are two basic ideas that constitute the theory of dematerialization. One, intensity-of-use of energy and materials initially increases then decreases with increasing GDP in all countries. Two, the later in time a country develops the lower its maximum intensity-of-use in energy and materials will be. These concepts, as they pertain to energy, are illustrated in Fig. 4.

The idea that the intensity-of-use of energy first increases then decreases is related to the impact of changing economic structure on energy use. In general, the process of industrialization results in large increases in commercial energy consumption, which causes energy intensity to initially increase. As consumer wealth rises, however, the structure of consumption changes increasing the demand for light manufactures and services. As these new demands are met, the structure of a nation's capital stock will change to reflect its output. Growth in the service sector relative to other sectors serves to reduce energy intensity as the energy required per unit of output from capital employed in the service sector is generally lower than that in the industrial sector.

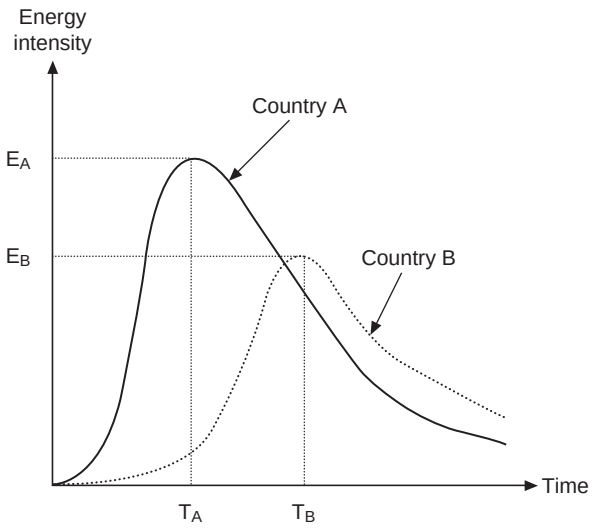


FIGURE 4 Consider two countries, country A and country B. Assume that gross domestic product per capita is positively correlated with time, and allow country A to develop more rapidly than country B. *Ceteris paribus*, the energy intensity of country A, reaches its peak (T_A, E_A) before country B (T_B, E_B). Moreover, country B benefits from the production processes and technologies developed by country A. Therefore, the peak energy intensity in country B is lower than the peak energy intensity in country B. Or, otherwise stated, $E_A > E_B$ because $T_A < T_B$.

The decline in intensity-of-use of energy and materials can also be hastened by the introduction of newer energy resources and materials with price and/or quality advantages, and also by the introduction of technologies that increase the energy efficiency of capital equipment. Industrialization in the West saw massive increases in the use of coal, as it was the primary fuel of choice for meeting energy requirements. However, increasing wealth generated new demands, and better, more suitable alternatives, namely oil and natural gas, eventually took the place of coal. For example, growth in private and small-scale transportation could be more easily facilitated with petroleum products than with coal, since the latter is more cumbersome and less efficient to meet that end. As environmental pressures mount and the associated costs are internalized, it is probable that ever cleaner alternative energy sources will become economically competitive, and therefore it is likely the intensity-of-use of other fuel sources will continue to rise over time as the intensity-of-use of fossil fuels continues to decline.

The second of the ideas central to the theory of dematerialization, which comes to bear because the diffusion of newer technologies and the economic/environmental lessons learned in the industrialized world could contribute to lowering the maximum energy intensity of developing nations, is difficult to test in a casual manner. While one can examine cross-country data to make comparisons, it is necessary to control for factors specific to individual countries that could influence energy intensity. Such research has not to date been completed. Nonetheless, the implication of the time dependency of maximum energy intensity is substantial. If developing countries, in general, will never reach the maximum intensity of energy use that was reached in industrialized countries in the 20th century, then expectations should be gauged accordingly. Long-term forecasting, therefore, should anticipate the possibility of these developments, particularly when assessing the impact of future economic development on the demand for energy.

3.1.2 Asymmetry in the Energy-Income Relationship

When considering the energy-income relationship, we must be mindful of the fact that the manner in which income influences energy demand is not symmetric. In other words, an increase in macroeconomic output will increase energy demand, but a decline in output will not necessarily have an equal and opposite effect. The reason for this is related to the arguments underpinning dematerialization, and possible explanations can be tied to Eq. (1). Specifically, different sectors grow at different rates and at different times during the course of economic development. As mentioned earlier, it is an oversimplification to assume Eq. (1) holds at the aggregate level. Doing so ignores the fact that capital stock characteristics are tied to specific capital stocks, and the proliferation of types of capital is very much dependent on economic structure. For example, the capital employed in industrial and manufacturing facilities is differentiated from the capital employed in households by the service they provide and the engineering characteristics they possess. Moreover, these items are more or less prevalent given the level of economic and social development, as well as the structure of production and consumption, within a nation.

Asymmetry in the energy-income relationship arises largely because not all expansions (recessions) are created equal. Some sectors of an economy may expand (contract) more rapidly than others during certain periods of time. Since different economic sectors are generally characterized by different types of capital, it is possible that some periods of macroeconomic expansion (contraction) will result in the expansion (contraction) of a sector in which energy intensity is very high relative to other sectors. Therefore, growth (reduction) in energy use could be larger in those time periods.

To illustrate, consider a two-sector economy (denoted as sectors A and B). Total energy consumption is given as the sum of energy use in each of the two sectors, $e = e_A + e_B$. Sector A has an energy intensity measured as

$$\frac{e_A}{y_A},$$

and sector B has energy intensity measured as

$$\frac{e_B}{y_B},$$

where y_j is the output of sector j and total output is given as $y = y_A + y_B$. Assume that sector A is more energy intensive than sector B, or $\frac{e_A}{y_A} > \frac{e_B}{y_B}$. Note that total energy intensity can be written as

$$\frac{e}{y} = \frac{e_A + e_B}{y} = \frac{e_A}{y_A} \cdot \frac{y_A}{y} + \frac{e_B}{y_B} \cdot \frac{y_B}{y}$$
$$= \frac{e_A}{y_A} \cdot \theta_A + \frac{e_B}{y_B} \cdot \theta_B,$$

where θ_j is the sector j share of total output.

It follows for a given rate of total economic growth, if sector A grows faster than sector B, the effect on total energy demand will be larger than if sector B grows faster than sector A. To see this, hold energy intensity within each sector constant. Taking the partial derivative of the preceding equation with respect to the output shares yields

$$\frac{\partial\left(\frac{e}{y}\right)}{\partial\theta_A} = \frac{e_A}{y_A} \text{ and } \frac{\partial\left(\frac{e}{y}\right)}{\partial\theta_B} = \frac{e_B}{y_B}.$$

But we know, by assumption, that

$$\frac{e_A}{y_A} > \frac{e_B}{y_B},$$

so growth in sector A has a larger impact on total energy intensity than growth in sector B.

It should be noted that reality is not as simple as this example. Specifically, technological innovations do cause sector-specific energy intensities to change over time (although, as the example illustrates, they need not do so for total energy intensity to change over time). However, if we assume that the time period is reasonably short, our assumption of constant sector-specific energy intensity is not a gross approximation. Allowing for a dynamic sector-specific energy intensity would, however, indicate just how complicated the long-term energy-income relationship can become.

3.2 Energy Demand and Imperfect Price Reversibility

It is well documented that energy prices have a significant negative impact on both energy demand and macroeconomic performance. However, this relationship is not symmetric. In other words, much as is the case with the energy-income relationship, rising energy prices will reduce growth in macroeconomic output and energy demand, but falling prices will not have an equal and opposite effect. This imperfect price reversibility arises due to the characteristics of the energy-consuming capital stock. When energy prices rise, there is incentive for the holders of energy-using capital stocks to take steps to increase efficiency. This could come in the

form of either retrofitting existing capital or replacing it with newer capital. In either case, these steps require capital investment that will not be reversed when energy prices fall. Likewise, improvements in technologies (most of which are embodied in newer capital) induced by higher energy prices will not be reversed when prices fall.

Imperfect price reversibility follows from examination of Eq. (1). In the short run, when energy prices increase, the user cost of capital also increases (see Eq. (4)) since capital and technology are fixed. Thus, the energy consumer will re-optimize his or her portfolio by reducing energy use where possible. Since capital and technology are fixed in the short run, this can only be done by reducing capital utilization rates (u_t). Not only will the decline in capital utilization result in a decline in energy demand, it will also result in a reduction in economic activity.

In the long run, allowing capital and technology to vary will, in general, still result in a decrease in energy demand, but it can complicate the macroeconomic response to energy price increases. As technologies change and capital stocks are turned over, older stocks are replaced with stocks in which newer, more energy-efficient technologies are embodied. The improved efficiency can generate asymmetry in the response of demand to price by preventing energy use from rising appreciably when energy prices decline. In addition, better efficiency can reduce the impact on energy demand of future price increases because, by Eq. (4), increasing efficiency lowers the user cost of energy-using capital. Thus, a Δ price decrease will not have an equal and opposite effect as a Δ price increase. Similarly, a Δ price increase in the future will not necessarily have the same impact as a Δ price increase in the present. To emphasize the point, we can again consider the U.S. experience in the consumption of motor fuels to illustrate the dynamics at work in the long and short run. As indicated in Fig. 5, average fuel efficiency of passenger vehicles in the United States was virtually unchanged from 1960 to 1978. For the majority of this period, oil prices were relatively stable, and vehicle utilization (measured as miles per vehicle) grew steadily. However, the oil price shocks of the 1970s and early 1980s were followed by an immediate decrease in vehicle utilization. During the same period, average vehicle efficiency was relatively stable. From 1979 to 1992, however, average vehicle efficiency increased at close to 5% per year, prompted by the energy efficiency policies adopted following the oil price shocks of the 1970s. The increases in efficiency in the early 1980s offset increasing prices, as 1981 saw vehicle utilization, fuel efficiency, and the real price of gasoline all increase.

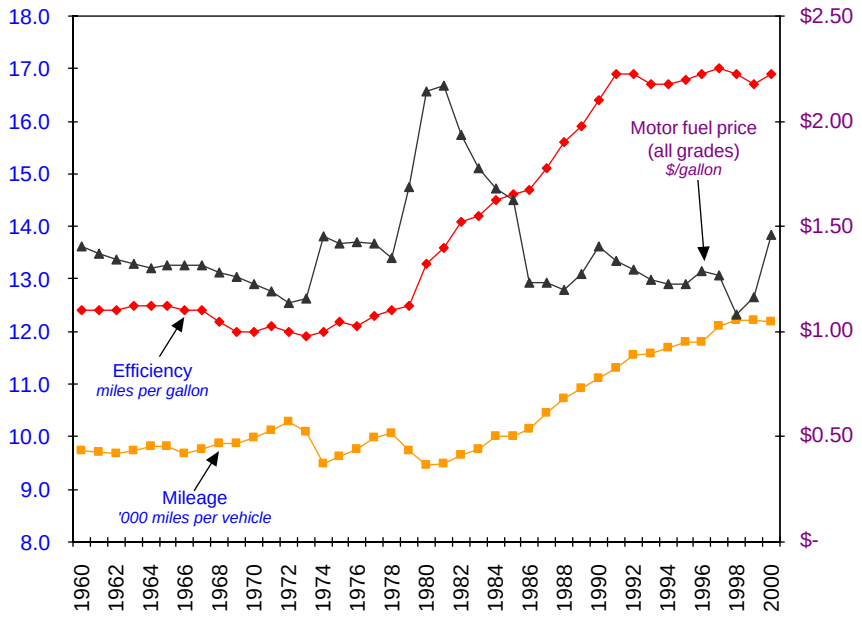


FIGURE 5 Energy prices for transportation increased dramatically from 1974 through 1982. In response, motor vehicle utilization declined initially, but a subsequent, albeit later, increase in efficiency of the motor vehicle fleet served to offset higher fuel prices, thereby reducing user costs.

SEE ALSO THE FOLLOWING ARTICLES

Business Cycles and Energy Prices • Climate Change: Impact on the Demand for Energy • Economic Geography of Energy • Economic Growth and Energy • Economics of Energy Supply • Energy Futures and Options • Market Failures in Energy Markets • Modeling Energy Supply and Demand: A Comparison of Approaches

Further Reading

Balke, N. S., Brown, S. P. A., and Yücel, M. K. (2002). Oil price shocks and the U.S. economy: Where does the asymmetry originate? *The Energy J.* **23**(3), 27–52.

Bernardini, O., and Galli, R. (1993). Dematerialization: Long-term trends in the intensity of use of materials and energy. *Futures*, May, 431–448.

Bohi, D., Toman, M., and Walls, M. (1996). "The Economics of Energy Security." Kluwer Academic Publishers, Boston, MA.

Diewert, W. E. (1974). Intertemporal consumer theory and the demand for durables. *Econometrica* **42,** 497–516.

Ferderer, J. P. (1996). Oil price volatility and the macroeconomy. *J. Macroeconomics*, Winter, 1–26.

Galli, R. (1998). The relationship between energy intensity and income levels: Forecasting long term energy demand in Asian emerging countries. *Energy J.* **19**(4), 85–105.

Gately, D., and Huntington, H. (2002). The asymmetric effects of changes in price and income on energy and oil demand. *Energy J.* **23**(1), 19–55.

Hamilton, J. D. (1983). Oil and the macroeconomy since World War II. *J. Political Economy* **91**(2), 228–248.

Huntington, H. (1998). Crude oil prices and U.S. economic performance: Where does the asymmetry reside? *Energy J.* **19**(4), 107–132.

Johansson, O., and Schipper, L. (1997). Measuring the long-run fuel demand of cars: Separate estimations of vehicle stock, mean fuel intensity, and mean annual driving distance. *J. Transport Economics and Policy* **31**(3), 277–292.

Judson, R., Schmalensee, R., and Stoker, T. (1999). Economic development and the structure of the demand for commercial energy. *Energy J.* **20**(2), 29–58.

Medlock, K. B., and Soligo, R. (2001). Economic development and end-use energy demand. *Energy J.* **22**(2), 77–105.

Medlock, K. B., and Soligo, R. (2002). Car ownership and economic development with forecasts to 2015. *J. Transport Economics and Policy* **36**(2), 163–188.

Mork, K., Mysen, H. T., and Olsen, O. (1994). Macroeconomic responses to oil price increases and decreases in seven OECD countries. *Energy J.* **15**, 15–38.

Tishler, A. (1983). The demand for cars and gasoline: A simultaneous approach. *European Economic Rev.* **10,** 271–287.

Economics of Energy Efficiency

ADAM B. JAFFE
Brandeis University and National Bureau of Economic Research
Waltham, Massachusetts, United States

RICHARD G. NEWELL
Resources for the Future
Washington, D.C., United States

ROBERT N. STAVINS
Harvard University
Cambridge, Massachusetts, United States

1. **Overview**
2. **The Process of Technological Change**
3. **Understanding the Energy Efficiency Gap**
4. **Market Failures**
5. **Non-Market-Failure Explanations of the Energy Efficiency Gap**
6. **Discounting and Implicit Discount Rates**
7. **Summary**

Glossary

externality An economically significant effect of an activity, the consequences of which are borne (at least in part) by parties other than the party who engages in the activity, and which are not accounted for through trade.

diffusion The gradual adoption of new process or product innovations by firms and individuals.

innovation The initial market introduction or commercialization of new process or product inventions.

invention The development and creation of a prototype new idea, process, or piece of equipment.

market barriers Disincentives to the diffusion and/or use of a good, such as high costs or prices, which may or may not represent market failures.

market failures The failure of private markets to provide certain goods at all or at the most desirable level, typically arising from a situation in which multiple parties benefit from a good without decreasing one another's benefits, and in which those who have paid for the good cannot prevent others from benefiting from it.

technological change The process of invention, innovation, and diffusion whereby greater and/or higher quality outputs can be produced using fewer inputs.

This article reviews economic concepts relevant to decision making about energy efficiency investments and related policy choices. We describe economic perspectives on the process of energy-saving technological change, the distinction between market failures and market barriers in energy-using product markets, and the important role that discounting plays in this area.

1. OVERVIEW

Energy efficiency is defined here to mean energy services provided per unit of energy input (for example, gallons of water heated to a specified temperature per British thermal units of natural gas input). Within this framework, energy efficiency is conceived primarily at the disaggregated, product level, rather than at a more aggregated sectoral level. As with virtually all economic problems, the economics of energy efficiency is at its heart a question of balancing of costs and benefits. For the individual energy user, this involves weighing the higher initial cost of purchasing energy-efficient products against the expected benefits of future cost savings when operating the products, among other considerations. Due to this difference in the timing of energy efficiency costs and benefits, issues related to discounting feature prominently in analyses of energy-efficient technology adoption. For suppliers of energy-using products, decisions regarding energy-efficient innovations likewise depend on the expected profits from such technology development. Profits from innovation depend in turn on the expected

 © 2004 Elsevier Inc. All rights reserved.

demand for energy-efficient technologies and the degree to which firms can appropriate the value created by their innovations.

On both the demand and supply side of this market for energy-efficient technology, potential market imperfections can lead to underinvestment in energy efficiency. This raises the possibility that corrective government policies could provide economic gains if they are practicable and provide net benefits after inclusion of all public and private implementation costs. The degree to which such opportunities exist in practice is the subject of significant debate. Finally, environmental pollution associated with energy production—particularly carbon dioxide emissions from fossil fuels—represents an additional reason why private markets might underprovide energy efficiency if energy users do not face the cost of any resultant environmental harm.

2. *THE PROCESS OF TECHNOLOGICAL CHANGE*

Many readers may be unfamiliar with the way economists typically view the process of technological change, thus it is useful to first establish this common understanding. Furthermore, to understand the potential for public policy to affect energy efficiency, it is also necessary to understand the process through which technology evolves: invention, innovation, diffusion, and product use. Invention involves the development of a new idea, process, or piece of equipment. This activity typically takes place inside the laboratory. The second step is the process of technology innovation, in which new processes or products are brought to market. Another way of describing this stage is commercialization. The third step is diffusion, the gradual adoption of new processes or products by firms and individuals, which then also decide how intensively to use new products or processes.

Tying this all together, it would be possible, for example, to think of a fundamentally new kind of automobile engine being invented. This might be an alternative to the internal combustion engine, such as a system dependent on energy-efficient fuel cells. The innovation step would be the work carried out by automobile manufacturers or others to commercialize this new engine—that is, bring it to market, then offer it for sale. The diffusion process, then, would reflect the purchase by firms and individuals of automobiles with this new engine. Finally, the degree of use of these new automobiles will be of great significance to demand for particular types of energy. The reason it is important to distinguish carefully among these different conceptual steps—invention, innovation, diffusion, and use—is that public policies can be designed to affect various stages and will have specific and differential effects.

2.1 Technology Adoption and Diffusion

Beginning at the end of the technological change process, research has consistently shown that diffusion of new, economically superior technologies is never instantaneous. The S-shaped diffusion path shown in Fig. 1 has typically been used to describe the progress of new technologies making their way into the marketplace. The figure portrays how a new technology is adopted at first gradually and then with increasing rapidity, until at some point its saturation in the economy is reached. The explanation for this typical path of diffusion that has most relevance for energy conservation investments is related to differences in the characteristics of adopters and potential adopters. This includes differences in the type and vintage of their existing equipment, other elements of the cost structure (such as access to and cost of labor, material, and energy), and their access to technical information. Such heterogeneity leads to differences in the expected returns to adoption and, as a result, only potential adopters for whom it is especially profitable will adopt at first. Over time, however, more and more will find it profitable as the cost of the technology decreases, quality improves, information about the technology becomes more widely available, and existing equipment stocks depreciate. The longevity of much energy-using equipment reinforces the importance of taking a longer term view toward energy efficiency improvements—on the order of decades (see Table I).

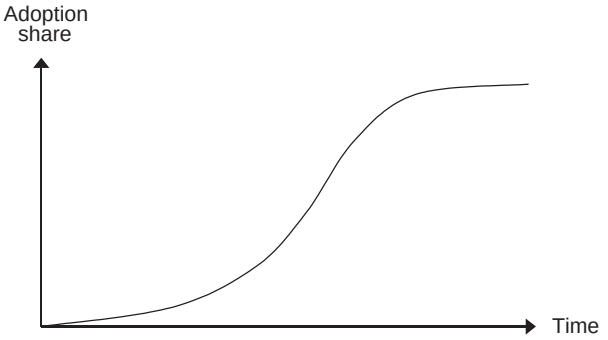


FIGURE 1 The gradual S-shaped path of technology diffusion.

Several studies have explored the effect of energy prices and technology adoption costs on energy efficiency investments. From a policy perspective, the effect of higher energy prices can be interpreted as suggesting what the likely effects of taxes on energy use (or carbon dioxide) would be and the effects of changes in adoption costs can be interpreted as indicating what the effects of technology adoption subsidies would be. As suggested by the economic cost–benefit paradigm, these studies have found that higher energy prices increase and adoption costs decrease the extent of adoption of energy-efficient technology (for example, building insulation, more efficient home appliances, or more efficient industrial motors). An additional interesting finding in this line of research is that the adoption of these technologies is more sensitive to the cost of the equipment than it is to the expected cost of energy. This implies that a policy of subsidizing the purchase of new efficient equipment may be more effective than a policy of taxing resource use, for policies that should in theory create the same magnitude of economic incentive.

There are at least three possible explanations for this divergence. One possibility is a behavioral bias that causes purchasers to focus more on up-front cost than they do on the lifetime operating costs of an investment. An alternative view is that purchasers focus equally on both, but uncertainty about future energy prices or whether they will face these costs (because they could move, for example) makes them give less weight to energy prices than they do to capital cost, which is known. A final interpretation might be that consumers have reasonably accurate expectations about future energy prices, and their decisions reflect those expectations, but the proxies for these expectations that are used by researchers are flawed, causing their measured effect to be smaller than their true effect. For example, studies often use current realized energy prices as a proxy for expected future energy prices. Current prices fluctuate more than expected future prices, however, leading to a downward bias in the coefficient on the energy price proxy relative to the true relationship with expected prices.

Although empirical evidence indicates that subsidies may be more effective than comparable taxes in encouraging technology diffusion, it is important to recognize some disadvantages of such subsidy approaches. First, unlike energy prices, adoption subsidies do not provide incentives to reduce utilization. Second, technology subsidies and tax credits can require large public expenditures per unit of effect, because consumers who would have purchased the product even in the absence of the subsidy still receive it.

TABLE I

Technology Diffusion and the Rate of Capital Stock Turnover[a]

Type of asset	Typical service life (years)
Household appliances	8–12
Automobiles	10–20
Industrial equipment/machinery	10–70
Aircraft	30–40
Electricity generators	50–70
Commercial/industrial buildings	40–80
Residential buildings	60–100

[a] Technology diffusion is closely related to the concept of "capital stock turnover," which describes the rate at which old equipment is replaced and augmented by new. New equipment can be purchased either to replace worn out and obsolete units or as a first-time purchase. A primary driver of replacement purchases for durable energy-using products is the useful lifetime of the product. The rate of economic growth is also important, especially for first-time durable-goods purchases; the rate of home construction is particularly relevant for residential equipment. The typical lifetimes for a range of energy-using assets illustrate that the appropriate time frame for thinking about the diffusion of many energy-intensive goods is on the order of decades.

2.2 Technology Innovation and Invention

Now it is possible to move back in the process of technological change from diffusion to innovation. In the energy efficiency area, it is helpful to think of the innovation process as affecting improvements in the attributes or characteristics of products. In Fig. 2, this process is represented as the shifting inward over time of a curve representing the trade-offs between different product characteristics for the range of products available on the market. On one axis is the cost of the product, and on the other axis is the energy flow (use; that is, the energy intensity) associated with a product. The downward slope of the curves indicates the trade-off between equipment cost of energy efficiency. Innovation means an inward shift of the curve—greater energy efficiency at the same cost, or lower cost for a given energy efficiency.

As with technology diffusion, studies have shown that increases in the price of energy have induced technological innovations in energy efficiency of commercialized products, such as household appliances (e.g., air conditioners, water heaters), automobiles, tractors, and jet aircraft. Moving back even further in the process of technological change to

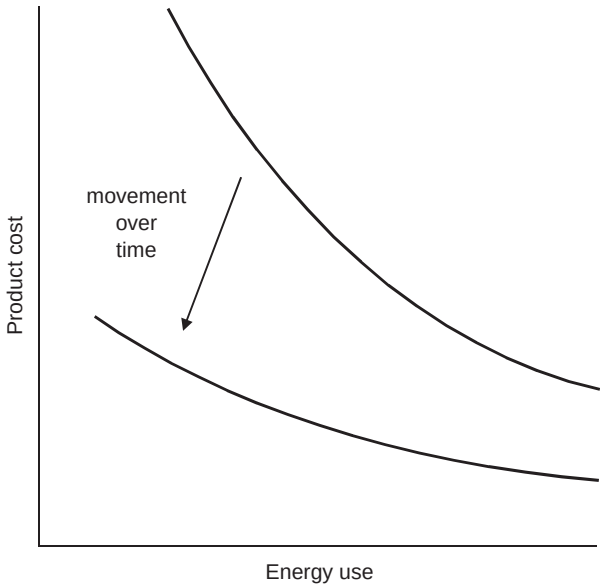


FIGURE 2 Innovation in product characteristics.

examine invention, other studies have analyzed U.S. patent data, finding that the rate of energy-conserving device patent applications (e.g., for waste heat, heat pumps, solar energy, and fuel cells) is significantly and positively associated with the price of energy.

3. UNDERSTANDING THE ENERGY EFFICIENCY GAP

Renewed attention is now being given by policymakers to energy efficiency, primarily due to concerns about global climate change associated with carbon dioxide emissions from fossil fuels. Much attention has been placed on the role that technological improvements in energy efficiency can play in reducing carbon emissions and in lowering the cost of those reductions. In such policy contexts, it is frequently asserted that there exists an "energy efficiency gap" between current energy use, on the one hand, and optimal energy use, on the other hand. From this perspective, the energy efficiency gap discussed earlier can now be thought of as a debate mainly about the gradual diffusion of energy-saving technologies that seem to be cost-effective.

Clearly, an estimate of the magnitude of this gap will be determined mainly by how optimal behavior is defined. Several major conceptual issues surround different approaches to defining an optimal energy-use scenario, with implications for guidance regarding public policy for energy and energy technologies. The standard economic approach defines "good" public policy to be that which maximizes the appropriately weighted sum of the values of goods and services enjoyed by society throughout time (including intangibles such as the environment). From this perspective, energy efficiency is not a goal in itself, but only a means to the end of overall economically efficient (and equitable) resource allocation. To the extent that energy generation and/or use creates environmental problems—one of the primary reasons why energy use is a public policy concern in the first place—such effects can in principle be incorporated in analyses by placing appropriate values on the environmental and nonenvironmental benefits and costs associated with energy generation and use.

The crux of the debate surrounding the energy efficiency gap lies in differing interpretations of what has been called the paradox of gradual diffusion of apparently cost-effective energy-efficient technologies. Why are compact fluorescent light bulbs, improved thermal insulation materials, and energy-efficient appliances not more widely used? Many studies have demonstrated that there exist such technologies (and processes)—ones that simple net-present-value calculations show to be cost-effective at current prices and market interest rates, but which enjoy only limited market success. The phrase market barriers has been used to refer to any factors that may account for this apparent anomaly. Others have used the phrase market barriers even more broadly to include, for example, low energy prices that are a disincentive to the adoption of more energy-efficient technologies. Differing views about the nature of these barriers lead to fundamentally different views about optimal energy use and the role of government policy.

As already noted, the diffusion of economically superior technologies is typically gradual. Awareness of this empirical reality should make the existence of the energy efficiency gap much less perplexing, but it does not answer the question of whether the optimal rate of diffusion is greater than the observed rate. To some degree, responses to the observation that new technologies always diffuse slowly depend on subjective perspectives. Some are inclined to note that if technology diffusion is typically not optimal, then there is no basis for presuming that the market is working efficiently, and hence no reason to eschew government intervention designed to improve it. On the other hand, others are more inclined to maintain that the government cannot possibly worry about trying to optimize the diffusion rate for all technologies, and that the government should hence stay out altogether, even if the no-intervention result is not optimal.

As will be described further here, at a theoretical level there are several reasons why technology diffusion will not, in general, occur at an economically efficient rate. But, if the purpose of assessing the efficiency gap is to identify desirable government policy interventions, then it is necessary to know whether the market barriers that cause slow diffusion can be mitigated by government intervention such that overall resource allocation is improved. In this context, the appropriate definition of the optimal level of energy efficiency is that which is consistent with efficient overall resource use, including efficient use of government resources.

Those market barriers that might justify a public policy intervention to overcome them are referred to in economics parlance as market failures. Normative economics teaches that in the presence of market failures, unfettered markets may not operate to produce outcomes that are socially optimal. When analysts speak of no-cost climate policies based on energy efficiency enhancement, they are often implicitly or explicitly assuming the presence of market failures in energy efficiency. By implication, there are some market barriers—such as "low" energy prices, "high" technology costs, and "high" discount rates—that are not necessarily market failures. In such cases, the existence of an energy efficiency gap does not, in and of itself, call for policy responses. On the other hand, there are also some market failures that do not relate to problems in individual decision-making per se, but that might still provide justification for policy interventions. In other words, even if the energy efficiency paradox is resolved in the sense that people's decisions are found to be consistent with the costs they face, there could be other reasons—particularly environmental pollution from energy use—why the resulting behavior would deviate from the social optimum. These classes of market failures and non-market-failure explanations of the energy efficiency gap are considered next.

4. MARKET FAILURES

In this section, the focus is on several sources of potential market failure that may affect energy-conserving technology adoption rates. Three of these relate to the availability of information. There are also market failure issues that do not help explain nonadoption at current technology costs and energy prices, but which are still relevant to policy debates about the energy efficiency gap. These include environmental externalities, broader innovation and adoption externalities, and issues related to energy supply pricing and national security.

4.1 Inadequate Information

First, information has important public good attributes: once created it can be used by many people at little or no additional cost. It may be difficult or impossible for an individual or firm that invests in information creation to prevent others who do not pay for the information from using it. It is well known that such public goods will tend to be underprovided by ordinary market activity. Second, if the act of adopting a new technology is inherently a source of useful information for others, then the act of adoption creates a positive externality by providing information to others for which the adopter is unlikely to be compensated. It takes time for the many potential users to learn of the new technology, try it, adapt it to their circumstances, and become convinced of its superiority. An important mechanism in this learning process is the observation of the adoption of the new technology by others. If a neighbor or competitor tries something new, and others see that it works, it becomes much safer and easier for the observers to try it. Hence the adopter of a new technology creates a positive externality for others, in the form of the generation of information about the existence, characteristics, and successfulness of the new technology. This phenomenon is often called learning by using. This (positive) externality is another form of market failure.

Incomplete information can also foster principal-agent problems, as when a builder or landlord chooses the level of investment in energy efficiency in a building, but the energy bills are paid by a later purchaser or a tenant. If the purchaser has incomplete information about the magnitude of the resulting energy savings, the builder or landlord may not be able to recover the cost of such investments, and hence might not undertake them. This is another potential form of market failure. A home-builder, acting as agent for the ultimate home-buyer, may have incentives to take actions that are different from what the principal would prefer. The extent to which this market failure is significant in practice is not well established.

4.2 Environmental Externalities

Economic analysis of environmental policy is based on the idea that the potentially harmful consequences of economic activities on the environment constitute

an externality. An externality is an economically significant effect of an activity, the consequences of which are borne (at least in part) by a party or parties other than the party who controls the externality-producing activity. An electricity generator that pollutes the air imposes a cost on society. The firm that owns the generator has an economic incentive to use only as much fuel as it can productively employ, because this input is costly to the firm. In jargon, the cost to society of having some of its fuel used up is internalized by the firm, because it has to pay for that input. But the firm does not (in the absence of appropriate environmental policy intervention) have an economic incentive to minimize the external costs of pollution associated with that fuel.

Although the details and refinements are important, all environmental policies, at their core, are designed to deal with this externality problem, either by internalizing environmental costs so that polluters will make efficient decisions regarding their consumption of environmental inputs, or else by imposing from the outside a level of environmental pollution that policymakers believe to be more efficient than that otherwise chosen by firms. If environmental externalities are not fully addressed by environmental policy—for political or other reasons—the resulting level of energy efficiency will likely be too low. But this may not be true if the government is already intervening to change the energy efficiency of products available in the market place (or is otherwise intervening to control the pollution associated with energy use). For example, Corporate Average Fuel Economy (CAFE) standards result in the production and sale of cars that are more efficient than consumers would otherwise demand. Depending on the actual magnitude of environmental externalities, it is possible that the efficiency of automobiles is greater than would be socially optimal.

4.3 Broader Innovation and Adoption Externalities

In addition to the externality associated with pollution, innovation and diffusion are both characterized by externalities, as well as other market failures. In the case of pollution, the problem is that a polluter imposes costs on others, and hence has an inadequate incentive, from a social perspective, to reduce those costs. With respect to technology, the problem is the reverse. A firm that develops or implements a new technology typically creates benefits for others, and hence has an inadequate incentive to increase those benefits by investing in technology. Pollution is a negative externality, and so the invisible hand allows too much of it. Technology creates positive externalities, and so the invisible hand produces too little of it.

With respect to innovation, the positive externality derives from the public-good nature of new knowledge. Whereas patents and other institutions try to protect firms' investments in innovation, such protection is inherently imperfect. A successful innovator will capture some rewards, but those rewards will always be only a fraction—and sometimes a very small fraction—of the overall benefits to society of the innovation. Hence innovation creates positive externalities in the form of knowledge spillovers for other firms, and spillovers of value for the users of the new technology.

In addition, production costs tend to fall as manufacturers gain production experience in a process commonly called learning by doing. If this learning spills over to benefit other manufacturers, it can represent an additional externality. In any event, the existence of these innovation and adoption externalities suggests a rationale for public support of research and development in general, not for energy-efficient technology in particular. An argument particularly supporting energy-efficient technologies in this regard would require stating that spillovers related to these technologies were somehow greater than for technologies more generally in the economy.

4.4 Market Failures in Energy Supply

4.4.1 Average-Cost Pricing

Actual energy prices, particularly for electricity, may differ from marginal social cost because of subsidies and because of pricing based on average rather than marginal cost. During the 1980s, it was widely perceived that the incremental costs of increasing electricity supplies were significantly greater than the average costs of existing electrical capacity. Because the prices consumers pay are typically based on historical average costs, it was frequently suggested that consumers faced inadequate incentives to conserve electricity. Each kilowatt of capacity that did not need to be built saved society much more than it saved the persons whose conservation decisions reduced the need for such new capacity. If true, this would be one reason why public policy should seek to promote greater energy efficiency than private individuals choose on their own.

Although this argument remains conceptually valid in part, several things have changed since the 1980s that weaken the argument considerably. First, the deregulation of electricity production for about 40% of the U.S. population has meant a movement toward competitive electricity prices; this trend will continue. Second, there is widespread excess capacity in the electricity industry in most of the United States, although there is substantial regional variation. Thus, it is simply no longer true that the incremental cost of capacity is well above the price paid by most consumers. Indeed, given the availability of wholesale bulk power at prices closer to variable cost than to total cost, it could be argued that the social value of reducing electrical utility loads is actually less than the prices paid by most consumers. Furthermore, this situation is likely to prevail for many years, depending on the growth of electricity demand. Thus, regulatory distortions in the utility industry no longer provide an argument for policy interventions to foster energy conservation. One caveat, however, is that at peak demand periods, the marginal cost of supply often exceeds the price. This provides an argument not for energy efficiency in general, but rather for load-shifting to off-peak periods, energy conservation during peak periods, or time-of-day pricing.

With respect to technology incorporated in new buildings, it can be argued that these buildings are likely to last long into the future, at which time new electrical capacity may once again be needed, so that consumers choosing energy-inefficient technology are imposing a social cost above their own energy costs, albeit one that is very uncertain and remote in time. With respect to retrofit investments, however, even this argument is unavailable, because these investments can simply be postponed until such time as they are needed to forestall the need to expand capacity.

4.4.2 Security Externalities

It has also been suggested that there are externalities associated with the economic and military security costs resulting from domestic U.S. dependence on oil imported from politically unstable regions. For this argument to be valid, it would have to be the case that, at the margin, these national security costs are reduced if oil consumption is marginally reduced. This seems unlikely to be the case.

5. NON-MARKET-FAILURE EXPLANATIONS OF THE ENERGY EFFICIENCY GAP

Non-market-failure explanations of the energy efficiency gap consist essentially of explaining why observed behavior is indeed optimal from the point of view of individual energy users. To be useful, such explanations must advance beyond the tautological assertion that if the observed rate of diffusion is less than the calculated optimal rate, there must be some unobserved adoption costs that would modify calculations of what is optimal.

A slightly less tautological approach would be to maintain the assumption—based on general experience—that private agents act in their own interests, unless it can be shown that specific market failures exist. Under this approach, it can quickly be concluded that no paradox exists, unless and until it can be demonstrated that the theoretical market failures already discussed are important in explaining observed behavior. This amounts to a policy prescription that markets should be "considered innocent until proved guilty." There is no scientific basis for such a prescription. On the other hand, it is important to keep in mind that although market failure may be considered a necessary condition for government policy intervention, it is not a sufficient condition. For government policy to be desirable in economic terms, it must be the case that a market failure exists and that there exists a government policy that, in overcoming the market failure, generates benefits in excess of the costs of implementation. Analyses of past government and utility energy conservation programs demonstrate that although some such programs certainly have had positive net benefits, the benefits of others have been lower than predicted.

5.1 Irreversibility and the Option to Wait

One such non-market-failure explanation is that uncertainty about the future benefits of energy-efficient technologies, combined with the irreversible nature of the efficiency investment, makes the effective discount rate for analyzing the net present value of energy savings significantly greater than is typically used in the calculations that suggest the existence of a paradox. Simple net-present-value analysis typically does not account for changes over time in the savings that purchasers might enjoy from an extra investment in energy efficiency, which

depends on trends and uncertainties in the prices of energy and conservation technologies. When making irreversible investments that can be delayed, the presence of this uncertainty can lead to an investment hurdle rate that is larger than the discount rate used by an analyst who ignores this uncertainty.

The magnitude of this option-to-wait effect or option value depends on project-specific factors, such as the degree of energy price volatility, the degree of uncertainty in the cost of the investment, and how fast the prices of energy and conservation technologies are changing over time. The effect is magnified when energy and technology price uncertainty is increased and when energy prices are rising and technology costs are falling more quickly. Under conditions characterizing most energy conservation investments, this effect could raise the hurdle rate by up to about 10 percentage points. On the other hand, if there is no opportunity to wait, this effect can be ignored.

Note that uncertainty, in contrast to imperfect information, is not a source of market failure in and of itself. It is reasonable and appropriate for individuals to take uncertainty into account in making investment decisions, and to apply relatively high hurdle rates to irreversible investments when returns are uncertain. To the extent that consumers' true effective discount rates are high for this reason, this would not represent a market failure or a reason for policy intervention from an economic perspective.

5.2 Unaccounted for Costs and Overestimation of Energy Savings

There are typically costs of adoption that are not included in simple cost-effectiveness calculations. It is by no means costless to learn how a technological improvement fits into a home or firm or to learn about reliable suppliers. Even after basic information about a technology has been disseminated, the purchase price of a new product is only a lower bound on its adoption cost. Another type of hidden cost is the possibility that qualitative attributes of new technologies may make them less desirable than existing, less efficient technologies. An obvious example is the difference in hue between florescent and incandescent lighting and the delay time in achieving illumination with many florescent lights.

Some have argued that not only costly information acquisition but also biased estimates of likely energy savings play a role. There is evidence that analysts have substantially overestimated the energy savings that higher efficiency levels will bring, partly because projections often are based on highly controlled studies that do not necessarily apply to actual realized savings in a particular situation. For example, studies have found that actual savings from utility-sponsored programs typically achieve 50 to 80% of predicted savings. Other studies have drawn similar conclusions based on analysis of residential energy consumption data, finding that the actual internal rate of return to energy conservation investments in insulation was substantially below typical engineering estimates for the returns from such investments. On the other hand, the bias could be offset in the opposite direction, because some studies indicate that consumers systematically overestimate energy savings associated with some types of new technologies.

5.3 Heterogeneity in Energy Users

Another possible non-market-failure explanation for the energy efficiency gap is associated with the fact that even if a given technology is cost-effective on average, it will mostly likely not be for some individuals or firms. If the relevant population is heterogeneous—with respect to variables such as the purchaser's discount rate, the investment lifetime, the price of energy, the purchase price, and other costs—even a technology that looks very good for the average user will be unattractive for a portion of the population. Heterogeneity in these and other factors leads to differences in the expected value that individual purchasers will attach to more energy-efficient products. As a result, only purchasers for whom it is especially valuable may purchase a product. Depending on how the efficiency gap is measured, such underlying heterogeneity can provide yet another non-market-failure explanation. For example, it may not make sense for someone who will only rarely use an air conditioner to spend significantly more purchasing an energy-efficient model—they simply may not have adequate opportunity to recoup their investment through energy savings. Analysis based on single estimates for the important factors previously discussed—unless they are all very conservative—will inevitably lead to an optimal level of energy efficiency that is too high for some portion of purchasers. The size of this group, and the magnitude of the resulting inefficiency should they be constrained to choose products that are not right for them, will of course depend on the extent of heterogeneity in the population and the assumptions made by the analyst.

6. DISCOUNTING AND IMPLICIT DISCOUNT RATES

Much discussion surrounding the energy efficiency gap is couched in terms of arguments about the appropriate discount rate to use in evaluating energy savings. It is useful to distinguish among a number of separate and distinct questions regarding the appropriate discount rate. First, an attempt can be made to estimate the implicit discount rate that consumers appear to be using when they make energy efficiency decisions. Second, there can be speculation about the correct discount rate for consumers to use in making such decisions. Third, the implicit discount rate (explained below) or the correct rate can be compared with other private discount rates in the economy. Fourth and finally, it can be asked whether any of these rates are equal to the social discount rate that ought to be applied in evaluating future energy savings when making public policy decisions.

The observation that consumers have high implicit discount rates when they make energy efficiency decisions is actually neither more nor less than a restatement of the existence of the energy efficiency gap. To estimate implicit discount rates, studies examine decisions actually made and calculate the discount rate that makes those decisions privately optimal, given estimates of the costs and future energy savings of the investments and given the assumption that there are no important market failures impeding the adoption of efficient technologies. There is substantial empirical evidence of implicit discount rates for energy-conservation investment decisions in the range of 20 to over 100% for products such as air conditioners, space heating, and refrigerators. The energy efficiency gap can thus be restated as the observation that these implicit discount rates appear to be much higher than other interest rates in the economy. To observe that implicit discount rates are high, however, says nothing about the reason people make the decisions they make. One possibility is that people are applying normal discount rates in the context of significant market failures; another possibility is that people actually utilize high discount rates in evaluating future energy savings. Note that "actually" does not mean "literally." It is not necessary that people know how to perform net-present-value calculations; rather, they make decisions balancing costs today and costs tomorrow, with those trade-offs incorporating their true discount rate. The truth is likely some combination of the two. Thus, high implicit discount rates, on their own, are neither a market failure nor an explanation of observed behavior.

Further, if only the relevant investment decisions are observed, it is fundamentally impossible to determine whether the observed behavior results from market failures, from truly high discount rates, or perhaps from misestimation of the true costs and savings of adoption. To make that distinction would require either observing something that distinguishes the market-failure explanations from the non-market-failure ones, such as whether people with better information are more likely to purchase more efficient models, or else calculating from some basis other than the investment decisions what an appropriate discount rate for these investments would be. Thus, to investigate the energy efficiency gap by changing the discount rates people use in making investment decisions amounts to assuming the answer. If the outcome with lower discount rates is considered the optimal result, then it is implicitly assumed that all market barriers are indeed market failures. Conversely, postulating that the optimal result is the one in which consumers are assumed to discount at observed high implicit rates, then it is implicitly assumed that there are no market failures.

To make this a bit more concrete, suppose that through calculations it is found that consumers' true discount rates for some set of energy efficiency investments are approximately 20%—higher than mortgage interest rates, but in the range of rates on personal credit cards. What would this suggest for public policy? If mandatory efficiency standards were being considered, for example, then requiring efficiency any greater than what someone with a 20% discount rate would choose would have the effect of making that consumer worse off. Thus, to the extent that high implicit discount rates correspond to truly high discount rates, rather than to market failures, there is nothing particularly wrong with those high rates, and they do not correspond to any efficiency gap that ought to be addressed by public policy. Instead, the question of whether relevant market failures exist is again the issue.

Finally, there is the issue of the social discount rate. Several decades of academic and policy debates have failed to resolve fully the question of what the social rate of discount should be, even as a conceptual matter. However, there is a strong argument that policies to increase energy efficiency can be justified only on the basis of the appropriate true private rate of discount, not some (lower) social rate. If the only reason why investment in energy-efficient technology is suboptimal is the divergence

between the private and social discount rates, then it is certainly true that a theoretical argument could be made to support public policies to increase that investment. The problem is that this same argument would also apply with equal force to all other forms of investment—plant and equipment, research, education, and so on. It may be that the government should be doing more to encourage all of these forms of investment, but that is surely an issue beyond the scope of energy policy.

Further, if the social rate of discount is really very low, then the set of available investment opportunities that should be undertaken probably exceeds the nation's annual gross national product. Obviously, the government can not undertake all of these projects. If the notion that public policy should have as its objective undertaking all investments for which rates of return exceed the social discount rate is taken seriously, then the reasonable prescription is to increase investment, beginning with the investments with the highest rates of return. But increasing investment reduces consumption, and long before investment opportunities were exhausted, consumption would be so significantly reduced that the social discount rate would rise. Increasing investment up to the point where the private and social rates of return coincide can be imagined, but it is impossible to know how much greater investment would be at that point or what the discount rate would be. Given this, a policy prescription to increase investment in energy efficiency should be based on a conclusion that the (social) rate of return to this form of investment is higher than the rate of return available on other forms of investment in the economy. Otherwise, this form of investment could well be increasing at the expense of other forms that are even more beneficial. Of course, saying that the social rate of return on energy efficiency investments exceeds the social rate of return on other investments means precisely that current investment in energy efficiency is inadequate when evaluated at the appropriate private rate of discount.

7. SUMMARY

It is possible to synthesize much of the above discussion by examining graphically the concepts behind the major, alternative notions of the energy efficiency gap. To understand the basic elements of the debate, it is helpful to distinguish first between energy efficiency and economic efficiency, as in Fig. 3. The vertical axis measures increased energy efficiency (decreased energy use per unit of economic activity). The horizontal axis measures increased economic efficiency (decreased overall economic cost per unit of economic activity, taking into account energy and other opportunity costs of economic goods and services). Different points in the diagram represent the possible energy-using technologies available to the economy as indicated by their energy and economic efficiency. As a concrete illustration of this distinction, consider two air conditioners that are identical except that one has higher energy efficiency and, as a result, is more costly to manufacture because high-efficiency units require more cooling coils, a larger evaporator, and a larger condenser, as well as a research and development effort. Whether it makes sense for an individual consumer to invest in more energy efficiency depends on balancing the value of energy that will be saved against the increased purchase price, which depends on the value of the additional materials and labor that were spent to manufacture the high-efficiency unit. The value to society of saving energy should also include the value of reducing any associated environmental externalities; but again this must be weighed against the costs. Starting from the baseline situation, as represented by the lower left corner of Fig. 3, the adoption of more energy-efficient technology is represented as an upward movement. But not all such movements will also enhance economic efficiency. In some cases, it is possible simultaneously to increase energy efficiency and economic efficiency. This will be the case if there are market failures that impede the most efficient allocation of society's

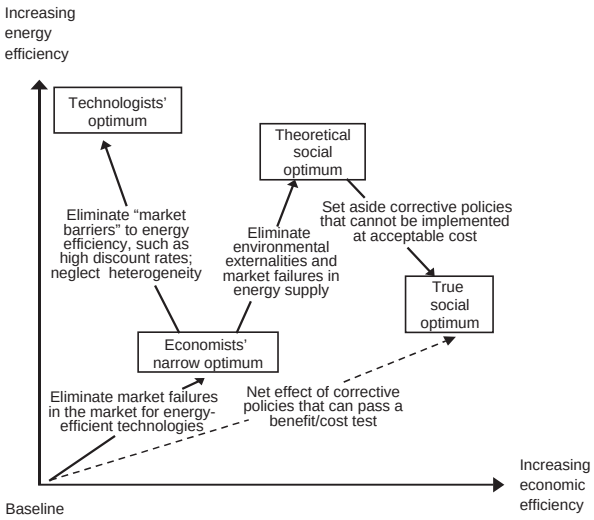


FIGURE 3 Alternative notions of the energy efficiency gap.

energy, capital, and knowledge resources in ways that also reduce energy efficiency. These are examples of what economists and others refer to as win-win or no regrets measures.

At the risk of excessive simplification, technologists can be characterized as believing that there are plentiful opportunities for low-cost or even negative-cost improvements in energy efficiency, and that realizing these opportunities will require active intervention in markets for energy-using equipment to help overcome barriers to the use of more efficient technologies. In terms of Fig. 3, the economist's notion of a narrow optimum is where market failures in the market for energy-efficient technologies have been corrected, the result being greater economic efficiency and energy efficiency. This optimum is narrow in the sense that it focuses solely on energy technology markets and does not consider possible failures in energy supply markets (such as underpriced energy due to subsidies or regulated markets) or, more important, environmental externalities associated with energy use (such as global climate change). Market failures in the choice of energy-efficient technologies could arise from a variety of sources, as discussed previously.

Eliminating broader market failures takes us to what is called the theoretical social optimum in Fig. 3. This represents both increased economic and increased energy efficiencies, compared with the economists' narrow optimum. This ignores the reality that the baseline efficiency already reflects policy measures motivated by environmental concerns. As emphasized previously, it is possible that these measures already achieve as much or more additional efficiency as would be justified by the environmental externalities of energy use. But not all market failures can be eliminated at acceptable costs. In cases in which implementation costs outweigh the gains from corrective government intervention, it will be more efficient not to attempt to overcome particular market failures. This takes us from a theoretical social optimum to what is referred to as the true social optimum in Fig. 3. Market failures have been eliminated, but only those for which elimination can pass a reasonable benefit–cost test. The result is the highest possible level of economic efficiency, but a level of energy efficiency that is intermediate compared with what would be technologically possible.

In contrast to the economist's perspective, if non-market-failure market barriers (such as high discount rates caused by uncertainty about payback) were also removed, the technologists' optimum would be achieved. This alternative notion of an optimum is found by minimizing the total purchase and operating costs of an investment, with energy operating costs being discounted at a rate the analyst (not necessarily the purchaser) feels is appropriate. Clearly, if no distinction is made between barriers to adoption that are market failures and those that are not, and no choice is made to eliminate all barriers, a higher estimate of energy efficiency will be achieved, but not economic efficiency. An important implication of this perspective is that comparisons of an engineering ideal for a particular energy use with the average practice for existing technology are inherently misleading, because the engineering ideal does not incorporate all the real-world factors influencing energy technology decision-making.

How these alternative definitions of optimal energy efficiency and the energy efficiency gap relate to specific empirical estimates depends, of course, on the assumptions that underlie those estimates. Many published studies of the potential for energy efficiency correspond to what is labeled technologists' optimum. That is, they assume that the resolution of the energy paradox must be that the simplest calculations are correct and that a host of market failures explain observed behavior. Certainly, estimates of efficiency potential derived from forcing low discount rates into analyses of consumers' and firms' decisions do not correspond to any policy-relevant notion of optimal. Unfortunately, it is easier to explain what is wrong with existing approaches than it is to specify what is the right approach. It is clear, however, that in order to understand what is wrong with some of the existing approaches to estimating the energy efficiency gap and to begin the process of searching for the right measure, the first requirement is to disentangle market-failure and non-market-failure explanations for observed decisions regarding energy-efficiency investments.

SEE ALSO THE FOLLOWING ARTICLES

Conservation Measures for Energy, History of • *Consumption, Energy, and the Environment* • *Discount Rates and Energy Efficiency Gap* • *Economics of Energy Demand* • *Economics of Energy Supply* • *Energy Efficiency and Climate Change* • *Energy Efficiency, Taxonomic Overview* • *Industrial Energy Efficiency* • *Obstacles to Energy Efficiency*

Further Reading

Anderson, S. T., and Newell, R. G. (2003). Information programs for technology adoption: The case of energy-efficiency audits. *Resource Energy Econ.* (in press).

DeCanio, S. J., and Watkins, W. E. (1998). Investment in energy efficiency: Do the characteristics of firms matter? *Rev. Econ. Statist.* **80**(1), 95–107.

Hassett, K. A., and Metcalf, G. E. (1993). Energy conservation investment: Do consumers discount the future correctly? *Energy Policy* **21**(6), 710–716.

Hassett, K. A., and Metcalf, G. E. (1995). Energy tax credits and residential conservation investment: Evidence from panel data. *J. Public Econ.* **57,** 201–217.

Hausman, J. A. (1979). Individual discount rates and the purchase and utilization of energy-using durables. *Bell J. Econ.* **10,** 33–54.

Hirst, E. (1986). Actual energy savings after retrofit: Electrically heated homes in the Pacific Northwest. *Energy* **11,** 299–308.

Jaffe, A. B., and Stavins, R. N. (1994). The energy paradox and the diffusion of conservation technology. *Resource Energy Econ.* **16**(2), 91–122.

Jaffe, A. B., and Stavins, R. N. (1994). The energy-efficiency gap: What does it mean? *Energy Policy* **22**(1), 804–810.

Jaffe, A. B., Newell, R. G., and Stavins, R. N. (2003). Technological change and the environment. *In* "Handbook of Environmental Economics," Vol. 1 (K.-G. Mäler and J. R. Vincent, Eds.), pp. 461–516. Elsevier, Amsterdam.

Joskow, P. L., and Marron, D. B. (1992). What does a negawatt really cost? Evidence from utility conservation programs. *Energy J.* **13**(4), 41–74.

Levine, M. D., Koomey, J. G., McMahon, J. E., and Sanstad, A. H. (1995). Energy efficiency policy and market failures. *In* "Annual Review of Energy and the Environment" (J. M. Hollander, Ed.), pp. 535–555. Annual Reviews Inc., Palo Alto.

Lind, R. C. (ed.). (1982). "Discounting for Time and Risk in Energy Policy." Resources for the Future, Washington, D.C.

Metcalf, G. E., and Hassett, K. A. (1997). Measuring the energy savings from home improvement investments: Evidence from monthly billing data. *Rev. Econ. Statist.* **81**(3), 516–528.

Newell, R. G., Jaffe, A. B., and Stavins, R. N. (1999). The induced innovation hypothesis and energy-saving technological change. *Q. J. Econ.*, August, 941–975.

Sanstad, A. H., Blumstein, C., and Stoft, S. E. (1995). How high are option values in energy-efficiency investments? *Energy Policy* **23**(9), 739–743.

Sebold, F. D., and Fox, E. W. (1985). Realized savings from residential conservation activity. *Energy J.* **6**(2), 73–88.

Sutherland, R. J. (1991). Market barriers to energy-efficiency investments. *Energy J.* **12**(3), 15–34.

Economics of Energy Supply

JEFFREY A. KRAUTKRAEMER
Washington State University
Pullman, Washington, United States

MICHAEL A. TOMAN
Johns Hopkins University
Washington, D.C., United States

1. **Introduction**
2. **Basic Hypotheses in the Economics of Depletable Energy Supply**
3. **The Simple "Cake-Eating" Model of Resource Extraction**
4. **A More General Theory of Economically Efficient Depletable Resource Supply**
5. **Effects of uncertainty**
6. **Market Failures and Distortions**
7. **Empirical Analysis of Depletable Energy Supply**
8. **Concluding Remarks**

Glossary

depletable resources Those natural resources for which total physical availability is fixed over any meaningful time horizon for economic decision-making; oil and coal are classic examples. The quality and other attributes of different deposits may vary.

economic scarcity The result of not having enough resources to meet all of one's needs. Consequently, choices must be made and something of value must be foregone or given up in order to gain something else of value. In the case of depletable resources, if there is not enough of the resource to meet all needs in each time period, then there is a scarcity value to the resource stock.

Hotelling rule Named for Harold Hotelling, the intertemporal efficiency condition in a simple depletable resource model that the *in situ* value of the resource stock increases at the rate of interest. The rule does not hold in such a simple form in more complex models, such as when there are stock effects.

Hotelling valuation principle The valuation of a resource deposit on the basis of current price, marginal extraction cost, and the amount of the resource in the deposit.

***in situ* value** The incremental increase in present value that would occur with a one-unit increase in the initial resource stock. It also is referred to in the literature as the "user cost" or "scarcity rent" associated with depletion of the resource stock.

marginal cost The added economic cost of increasing an activity such as resource extraction or reserve discovery.

stock effect Usually, the increase in marginal extraction cost that occurs when a resource stock is depleted, either because it is physically more difficult to extract the resource (e.g., well pressure declines) or because lower quality or higher cost deposits are exploited.

The basic logic underlying all neoclassical economic theories of "optimal" energy supply is maximization of the present value of some stream of economic returns. The economic theory of optimal resource depletion has evolved since Hotelling's classic 1931 article. The power of the theory to support improved empirical understanding of actual behavior is considered herein. The empirical literature indicates that this theory has so far provided only limited empirical understanding.

1. INTRODUCTION

The economic principles governing the efficient production of depletable energy resources (oil, natural gas, and coal) have been a subject of interest and controversy for over 70 years. The pathbreaking article on the subject by Hotelling in 1931 was motivated by then-current concerns about scarcity of energy and other natural resources that have a remarkably familiar ring to contemporary ears. More recently, the energy crisis of the 1970s motivated new concerns about the availability of sufficient energy resources

 © 2004 Elsevier Inc. All rights reserved.

and the capacity of economic markets to allocate those resources effectively over space and time.

In this article, the basic logic underlying all neoclassical economic theories of optimal energy resource provision, maximization of the present value of some stream of economic returns, is first presented and discussed. Then, how the economic theory of optimal resource depletion has evolved since Hotelling's article is assessed. Considered in turn are theories of economically efficient energy resource extraction and new resource development. After presenting the basic theoretical framework, the discussion centers on extensions incorporating investment in fixed production capacity, uncertainty, and various market distortions, as well as extensions incorporating nondepletable "backstop" energy resources.

The utility of an economic theory is to be found not merely in its conceptual elegance, but also in its power to support improved empirical understanding of actual behavior. The empirical literature indicates that this work has so far provided only limited empirical understanding of observed energy supply behavior. Some key examples in which the theory and application of depletable energy supply have been developed in the literature. Some relatively recent surveys provide more detailed references (those by Cairns, Walls, Epple and Londregan, Toman and Walls, and Krautkraemer).

2. BASIC HYPOTHESES IN THE ECONOMICS OF DEPLETABLE ENERGY SUPPLY

The fundamental assumption underlying all of mainstream economic theory is that economic actors are motivated to make choices by the goal of maximizing some measure of economic returns. The textbook model of commodity supply is a simple case in point. In that model, competitive producers (producers lacking any market power, taking prices of inputs and outputs as given) seek to maximize profit, the excess of gross revenue over their opportunity cost of production, including returns to capital and entrepreneurial effort. Profit is maximized when the market price of output, a measure of incremental gain from expanding production, equals the incremental cost of output expansion taking into consideration all possible technological possibilities for cost-effectively combining various inputs.

The same assumption underlies the economics of depletable energy supply. However, in that theory, the measure of returns assumed to motivate behavior is a present value of current and future net revenues that accrue over time from a sequence of production decisions. A present-value metric is used in depletable energy economics because (for reasons explained further later) the act of producing or developing more or less resource today has unavoidable consequences for future potential returns, owing to the interplay of economics, geology, and technology. If these future consequences were ignored by totally myopic producers, their cumulative effects would come as an unanticipated and unwelcome surprise later. Because there is every reason to assume that producers are aware of these intertemporal connections and, being aware, are motivated to account for them, it makes more sense from the standpoint of theoretical logic to suppose that producers systematically account for them in evaluating the present value of a sequence of decisions over time.

The foregoing does not imply that energy producers are omniscient or free of forecast error. That is prima facie not the case. Later in this article there is a discussion of how the theory of depletable energy supply has been extended to incorporate various kinds of uncertainties. Those extensions emphasize the key role that expectations play in influencing supply behavior. What is assumed in the economics of depletable energy supply, however, is that producers are capable of avoiding systematic and persistent errors over time. This assumption can and sometimes is questioned, but no known persuasive economic theory based on alternative assumptions has been formulated.

It is worth noting in this regard that the calculation of a present value involves the choice of discount rate(s) to deflate future returns relative to current returns. In characterizing producer behavior in depletable energy supply theory, the appropriate discount rates would reflect the competitive rate of return capital invested in the energy industry could earn in other uses, including adjustments for various kinds of financial risks. In practice, market-based rates of return put disproportionate weight on returns in the first few years of an activity, thus the assumption that producers are not systematically biased in their expectations of future returns is not as difficult to swallow as it might first seem.

The standard theory of depletable energy supply can be extended well beyond a simple description of a single competitive firm. In addition to uncertainty, already noted, the theory can be extended to incorporate a number of market distortions, including property rights problems, environmental spillovers, taxes, and market power. However, efforts to

describe the exercise of market power by energy producers such as the Organization of Petroleum Exporting Countries (OPEC) and its members have been especially unsuccessful. A number of political and other factors must be incorporated to treat market power properly.

3. THE SIMPLE "CAKE-EATING" MODEL OF RESOURCE EXTRACTION

In the simplest case, it is assumed that a known, finite stock of the energy resource is being allocated over some number of time periods to maximize the present value of net returns. Producers calculate returns to extraction in different periods based on price expectations (in the simple model, expectations are assumed always to be realized), and knowledge of costs based on technological and geological information. Returns across periods are linked in this model by the simple fact that any energy resource produced in some period is by definition not available for production in future periods (energy resources obviously cannot be recycled). Only a finite amount of energy resource is available in this framework, and all of it will eventually be extracted. But the finite availability of fossil fuels is not the only key feature of energy supply, even in the simplest framework. In this framework, the present value of net returns is maximized when the "equimarginal principle" is satisfied. This principle requires that the marginal net return from the extraction and sale of the energy resource be the same in every time period with positive extraction. This, in turn, requires the current value of marginal net profit to be increasing at the rate of discount applied to future returns.

This principle can be stated in simple mathematical terms. Let P_t denote the price of extracted energy resource at time t, and q_t is the quantity extracted at time t, $C(q_t)$ is the cost of extracting P_t, and δ is the discount rate. Then, under the equimarginal principle, it must be that

$$P_0 - C'(q_0) = \frac{P_1 - C'(q_1)}{(1+\partial)^1} = \frac{P_2 - C'(q_2)}{(1+\partial)^2} = \frac{P_t - C'(q_t)}{(1+\partial)^t} = \lambda \quad (1)$$

where λ denotes the common present value of marginal net profit. This number can be interpreted as the present-value shadow price or *in situ* value of the resource stock, because it reflects the incremental increase in present value that would be enjoyed if a resource owner could experience an augmentation of resource stock. It also is referred to in the literature as the "user cost" or "scarcity rent" associated with depletion of the resource stock.

At any point in time, then,

$$P_t = C'(q_t) + (1+\partial)^t \lambda. \quad (2)$$

In any successive time periods,

$$\frac{p_{t-1} - C'(q_{t-1})}{(1+\delta)^{t-1}} = \frac{P_t - C'(q_t)}{(1+\partial)^t} \quad \text{or}$$
$$P_t - C'(q_t) = (1+\delta)[P_{t-1} - C'(q_{t-1})]. \quad (3)$$

These conditions are known as the Hotelling rule (see below) in honor of Hotelling's contribution to the literature. The first efficiency condition requires that at any point in time, the marginal gain from additional extractions equals the full opportunity cost of extraction, including the present value of future losses from a diminished future resource stock as well as conventionally defined incremental production costs. The second efficiency condition requires that the rate of return to holding the asset be equal to the rate of discount. This condition can be thought of in terms of asset market equilibrium, because presumably the discount rate is the rate of return that could be earned by investing in other assets. The markets for assets will not be in equilibrium if the rate of return to holding one asset is greater than the rate of return to holding some other asset. If that were the case, investment in the higher (lower) return asset would increase (decrease). Thus, asset market equilibrium requires that the rate of return to holding a nonrenewable resource asset be the same as the rate of return to other assets in the economy as represented by the discount rate.

The user cost, or shadow price of the remaining stock, depends on future extraction costs and the future price path. These depend, in turn, on developments in resource production technology, the size of the remaining stock, and future availability of substitutes. The future price of the resource can depend on the general level of economic activity, so expectations about population and economic growth are important determinants of the scarcity rent or user cost. Technological progress can affect extraction cost, the availability of substitutes, and the remaining resource stock. For example, the development of directional drilling techniques has lowered the cost of extraction and increased the amount of recoverable oil from some deposits. Increased computing capability has enhanced the ability to use information from exploration activities

and lowered the cost of discovering and developing new oil reserves.

The model just described can be extended in several directions. Switching from considering a single competitive producer to a price-setting monopolist, it is possible to infer that the monopolist will reallocate output from earlier to later in the life of the resource base, because this generates higher prices early on, and near-term returns are worth more than discounted future returns. But because the stock of resource is fixed and is necessarily exhausted under either market structure, resources that are not produced earlier will be produced later. If, to take another scenario, the stock of the resource is fixed but uncertain, then, again, extraction early on would be less than in a situation with certainty about the resource stock, because producers would put weight on the possibility of more scarce and therefore more valuable carry-over stocks. If, on the other hand, there is a renewable backstop energy form that can be supplied if the price of depletable energy climbs high enough, the Hotelling rule will still hold, but the net price path will rise at the rate of discount until it meets the unit cost of the backstop just at the point of full depletion.

Despite its seeming intuitive appeal, however, the basic Hotelling model is built on very shaky assumptions. It assumes a known stock of a resource of homogeneous quality and that the extraction technology does not change over time. In fact, nonrenewable resource stocks are not a given: exploration for new deposits as well as the further development of existing deposits are important features of fossil fuel supply. For many nonrenewable resources, including fossil fuels, the discovery of additional deposits has exceeded consumption, so that known, or proved, reserves has actually increased over many time periods. For example, U.S. oil reserves increased from 13 billion barrels in 1930 to 33.8 billion barrels at the end of 1990, and production in that period was 124 billion barrels. More than 30 billion barrels were produced from 1991 to 2000, but reserves at the end of 2000 were 29.7 billion barrels. The increase in world oil reserves is even more dramatic—proved reserves increased from 660 billion barrels at the end of 1980 to 1009 billion barrels at the end of 1990. This increased slightly to 1046 billion barrels at the end of 2000, even though 250 billion barrels were produced from 1991 to 2000.

Nonrenewable resources, including energy resources, also vary in quality and in their cost of extraction. Moreover, cost conditions for extraction in a specific petroleum reservoir change over the

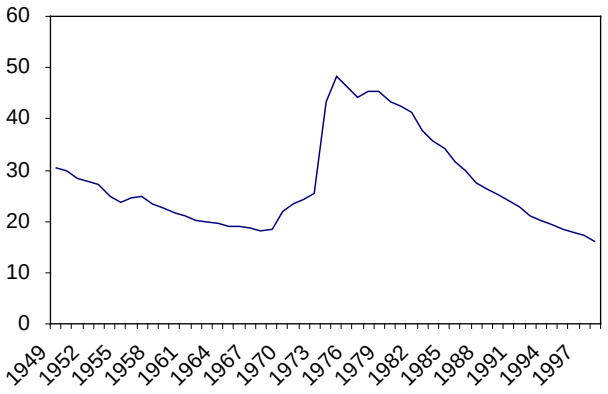


FIGURE 1 Real price of coal, 1949–1999, in chained 1996 dollars, calculated by using gross domestic product implicit price deflators. From the Energy Information Administration, United States Department of Energy, *Annual Energy Review.* Available on the Internet at http://www.eia.doe.gov/emeu/aer/txt/tab0708.htm.

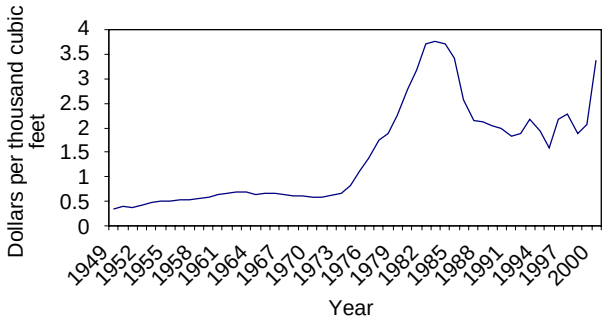


FIGURE 2 Real price of natural gas, 1949–2000, in chained 1996 dollars, calculated by using gross domestic product implicit price deflators. From the Energy Information Administration, United States Department of Energy, *Annual Energy Review.* Available on the Internet at http://www.eia.doe.gov/emeu/aer/pdf/pages/sec6_19.pdf.

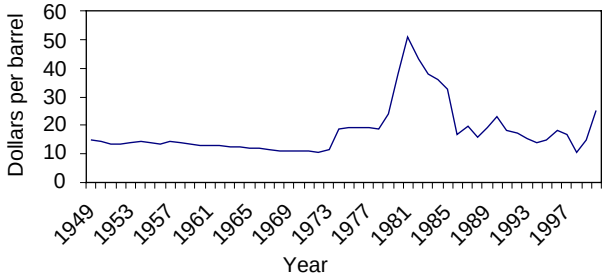


FIGURE 3 First purchase price, 1949–2000, of crude oil (domestic), in chained 1996 dollars, calculated by using gross domestic product implicit price deflators. From the Energy Information Administration, United States Department of Energy, *Annual Energy Review.* Available on the Internet at http://www.eia.doe.gov/emeu/aer/pdf/pages/sec5_39.pdf.

economic life of the reservoir: as natural drive is depleted, cost rises and yield per unit extraction effort falls, unless costly secondary and other recovery enhancement measures (e.g., injection of water or carbon dioxide) are undertaken to stimulate reservoir pressure. In short, resource stocks are neither fixed nor homogeneous, contrary to the assumptions underlying the simple Hotelling rule. It is not surprising, therefore, that the simple rule has done poorly in describing actual supply behavior (see later). We must look to a richer theory if we want to better understand observed behavior.

Empirically, the resource price should reflect both the marginal extraction cost and the scarcity rent, therefore generally it would be expected from Eq. (3) that the price of a nonrenewable resource would be increasing over time. However, there has not been a persistent increase in prices of fossil fuels (or other nonrenewable resources), but rather fluctuations over various time periods. Fossil fuel prices fell over much of the 20th century, peaked dramatically in the 1970s and early 1980s, and have generally fallen since then (see Figs. 1–3). This further illustrates the need for a richer theory to capture observed behavior.

4. A MORE GENERAL THEORY OF ECONOMICALLY EFFICIENT DEPLETABLE RESOURCE SUPPLY

A more useful general theory must take into account the stylized facts already noted—that resources are heterogeneous, with extraction costs that vary systematically over time with depletion of resource deposits—and that resources can be augmented through exploration and development. Moreover, a complete theory must account for the high degree of capital intensity of the extractive sector and the implications of this in terms of potential constraints on productive capacity.

4.1 Resource Extraction

Let us first address generalization of the Hotelling extraction model to deal with resource heterogeneity. Among the many good papers to deal with this case, those of Pindyck (in 1978) and Heal (in 1976) are particularly noteworthy. To examine this part of the problem, continue to treat reserves as fixed for the moment, and ignore capacity constraints.

The cost of extraction is assumed to be given by some function, $C(q_t, R_t)$, where R_t is the stock of resource that has not yet been depleted. Incorporating this "stock effect" into extraction cost is a simple if somewhat imprecise way to introduce resource heterogeneity into the extraction cost relationship. The relationship is relatively straightforward if we focus on a single energy resource development: as the resource in this development declines, cost is assumed to rise (deeper pits to be dug, more waste material to be removed, weaker reservoir pressure). It is less straightforward if we are attempting to describe heterogeneity among resource deposits (as is discussed later). With cost rising as the resource is depleted, the assumption of complete physical exhaustion from extraction—which is not empirically realistic anyway—no longer makes sense in the theoretical model either. Instead, what drives extraction behavior is economic exhaustion of the resource—at some point, cost becomes so high relative to the price what buyers are willing to pay that a particular mine or field is abandoned.

With resource heterogeneity and incomplete physical exhaustion, the simple Hotelling model summarized by the efficiency conditions [Eqs. (1)–(3)] no longer applies. Instead, the equimarginal principle must be modified as in the following equation:

$$\begin{aligned} P_t &= C_{qt} + (1+\partial)^t \lambda_t \\ &= C_{qt} + \sum_{s=t+1}^{T} (-C_{Rs})(1+\partial)^{t-s}. \end{aligned} \tag{4}$$

In Eq. (4), subscripts q and R on cost C signify rates of change of cost with respect to the extraction rate and the depletion of reserves, respectively. The term λ_t is a generalized measure of user cost that reflects the rate at which future extraction cost is increased as a consequence of future depletion. This stock effect is part of the full opportunity cost of current extraction. Manipulating Eq. (4) a bit algebraically yields

$$P_t - C_{qt} = (1+\partial)\left[P_{t-1} - C_{q,t-1}\right] + C_{Rt}. \tag{5}$$

Equation (5) shows how with increased stock providing a kind of "dividend" in terms of holding down extraction cost over time ($C_{Rt} < 0$), the net return from depletion no longer has to rise at the rate of the discount rate to induce resource owners to retain their holdings. Application of this framework to describe the depletion over time of different energy deposits is frequently done, but is somewhat more problematic. In the simplest case, where resource quality varies across deposits but is homogeneous within a deposit, the optimal extraction pattern generally requires exploiting those deposits in a

sequence from low cost to high cost. The user cost for a lower cost deposit is greater than the user cost for a higher cost deposit. At the time of transition from one deposit to the next most costly deposit, the marginal extraction cost plus user cost is the same at each deposit. This implies that the resource price rises at a rate slower than the rate of interest during the transition. Simultaneous extraction from different deposits can be optimal when marginal extraction cost at a deposit increases with the extraction rate, or if extractive capacity is fixed.

The current pattern of petroleum extraction obviously violates the rule of exploiting lower cost deposits before higher cost deposits. The cost of extraction from petroleum reserves in the Middle East is much lower than the cost of extraction from reserves in the United States. This suggests that the simple modeling device of incorporating resource inventory in the extraction cost relationship probably aggregates over and therefore blurs some important distinctions. The case of petroleum is further complicated by the actions of large oil producers in OPEC using market power to restrict their output to levels below what would occur under perfect competition. The higher-than-competitive price that results provides the incentive for higher cost producers to extract from their reserves. The higher noncompetitive price also distorts the capital investment decision, because the marginal cost of investment in capacity varies widely, from $343 in Saudi Arabia to over $10,000 in the United States. Efficient allocation would result in the equalization of marginal investment cost across geographical regions.

4.2 Exploration and New Resource Development

Exploration is an economic activity and, like other economic activities, marginal benefit and marginal cost play a key role in determining how much exploration occurs. Again, Pindyck published a key paper in the development of this analysis. Expenditures on exploration and development reduce current profit with an expected return of adding valuable reserves for future exploitation. The efficient level of exploration activity balances the expected marginal cost of exploration with the expected benefit of exploration.

The expected marginal benefit of exploration is not just the discovery and development of new reserves per se; in a model of incomplete resource exhaustion, the quantity of the resource stock is not determinative. The marginal benefit of exploration comes from the discovery of reserves for which the cost of development and extraction is less than or equal to the cost of the reserves currently being depleted. In other words, the marginal value of new reserves equals the user cost of current resource depletion along an efficient time path of extraction and new reserve development.

One feature of the full opportunity cost of exploration activity is that it reduces the stock of exploration opportunities, so that future additions to reserves tend to become more costly. In other words, there is a user cost similar to that in Eq. (4) associated with depleting the stock of exploration opportunities, in addition to direct exploration costs. If the conditions for efficient exploration, development, and extraction are then combined, we get

$$\begin{aligned}\text{Price} &= (\text{marginal extraction cost}) \\ &\quad + (\text{extraction user cost}) \\ &= (\text{marginal extraction cost}) \\ &\quad + (\text{marginal direct cost of finding new reserves}) \\ &\quad + (\text{user cost of reduced long-term development prospects}). \end{aligned} \tag{6}$$

The heuristic equality in Eq. (6) can be related to intuition about opportunity cost in economics more generally. The extraction user cost is the cost of depleting an existing inventory, which can be replenished, but only (in the model) at an ever-rising cost. The sum of the marginal extraction cost and the marginal direct cost of finding new reserves gives a measure of the cost of using and replacing the inventory, but it ignores the ultimate cost of natural capital depreciation reflected in the user cost of reduced long-term development prospects.

Once we allow reserves to vary endogenously with economic (and technological) conditions, the inexorable upward movement of prices implied by Eq. (3) or even Eq. (5) no longer need hold. Depending on the degree to which current extraction costs are rising as a consequence of depletion and the cost of new reserve additions, including their user cost, it is possible in the model for new discoveries to outstrip depletion, at least for awhile, causing a U-shaped energy price path. This is a reassuring finding, because such behavior has been observed in energy and other resource markets, and a model incapable of replicating such behavior would have little use. The ability to de facto enlarge reserves through technological breakthroughs that increase recoverability and lower the costs of extraction from existing deposits, or make new deposits easier to find and access (lowering reserve

replacement costs), likewise could exert downward pressure on prices, possibly for a long period of time. The simple model of endogenous exploration and reserves makes the simplifying assumptions that reserves are found on average in order of increasing cost and can be found in infinitely divisible amounts. Given the uncertainties surrounding the exploration process, the former assumption obviously is an oversimplification. When the discovery of deposits is more random, the aggregate extraction cost function cannot be described by a reserve stock indicator alone.

If the discovery of deposits is "lumpy," exploration and discovery can occur in "episodes," with intervals of no exploration between those episodes. Exploration does not generally occur in strictly discrete episodes, but it does vary with the resource price, and large discoveries do put significant but temporary downward pressures on price. The theory would then predict that the *in situ* value of the resource increases between major discoveries and jumps up or down when exploration occurs, depending on whether the exploration outcome is less or more favorable than expected. The resource price path can follow a "sawtooth" pattern while still having a downward trend over some time periods, depending on the actual exploration outcomes. The likelihood that the resource price will be increasing over time increases as unexplored prospects become scarcer.

Exploration also provides new information about future exploration prospects that can lead to revised expectations about the future value of the resource stock and generate new expected time paths for the resource price and quantity extracted. For example, a firm may revise its expectations about the probability of successful exploration either upward or downward in response to information obtained from its current exploration. In this case, a variety of patterns are possible for the realized price path, including a generally downward trend. The arrival of previously unanticipated information can alter the resource price, extraction, and exploration paths so that the observed price path deviates systematically from a deterministic calculation. The observed time paths for the resource price and *in situ* resource value may represent a combination of the initial portion of many different expected price paths rather than outcomes along one fully anticipated price path. Moreover, because individual energy producers can learn something about the information possessed by others via observing their behavior (e.g., by another producer not drilling in a certain area), information has certain public good aspects. To ensure that enough information about energy resources is collected and disseminated to society at large, it may be necessary for the government to undertake or underwrite investments in resource delineation.

4.3 Capacity Investment and Capacity Constraints

The basic nonrenewable resource model does not capture the capital intensity of extractive industries. The development and extraction of fossil fuels are capital intensive, and the timing and size of investment in extractive capacity are functions of the cost of capital and the expected price path. Once the existence of capacity constraints is incorporated, the behavior of the energy resource price along an efficient path generally will depart from the kind of behavior implied by Eq. (5) (or its extension to incorporate new reserve additions). This problem, too, has been studied by a number of authors (Lasserre, Powell and Oren, and Cairns and Lasserre).

If extraction is expected to decrease over time, either because the present value price of the resource is declining or the extraction cost increases with depletion, the extractor would seek to reduce its capital input over time if it were able to do so. However, this is not possible if capital is nonmalleable (i.e., once invested in oil fields, it cannot easily be converted into rail car investments). The energy-producing firm takes this into account when the initial investment is made, and the initial capital investment is lower than it would have been if capital were malleable. Once in place, this lower initial extractive capacity is likely to constrain the rate of extraction from the deposit, at least early in the extraction horizon. In addition, it may be relatively costly to increase extractive capacity, at least quickly. Firms may invest in advance of the need for some capacity in order to smooth out their investment costs over time, but they will also invest slowly in response to previously unanticipated price increases. Consequently, the short-run supply curve for an energy resource may be very inelastic and so variations in market demand (especially increases) may be mediated more by price changes than by quantity changes. The result is that short-term energy prices can be relatively volatile.

In the standard nonrenewable resource model, an increase (decrease) in the interest rate decreases (increases) the relative value of extraction in the future relative to current extraction, and so tilts the depletion path toward the present and away from the future. However, if a large capital investment is necessary before extraction begins, an increase in the

interest rate increases the cost of the capital investment. Thus, a higher interest rate reduces the incentive to use capital, and can lead to lower rather than greater initial extraction whether or not capital is malleable. The effect of the higher interest rate on the capital investment outweighs the effect of the higher interest rate on user cost if the initial resource stock is relatively large. A lower interest rate increases conservation if the interest rate is low (when user cost is high and capital cost is low), but decreases conservation when the interest rate is high (user cost is low and capital cost is high). A higher (lower) interest rate also can increase the capital cost of a backstop technology, increasing the maximum, or choke, price for the depletable energy resource and resulting in less (more) rapid extraction.

5. EFFECTS OF UNCERTAINTY

There is uncertainty regarding the future values of many important influences on the intertemporal supply of depletable energy resources: future price, future extraction cost, the costs and outcome of exploration and development activities, and the cost and timing of the availability of a backstop technology. The intertemporal extraction and resource development patterns and price path are affected by uncertainty in a variety of ways, depending on the source and nature of the uncertainty. From a financial perspective, the risk associated with holding reserves of a depletable energy resource can be mitigated by also owning other assets. The necessary return to holding the energy resource asset—including the rate of *in situ* appreciation plus the benefits of increased reserves in moderating costs—can be less if the return is negatively correlated with the return to other assets in a portfolio, thereby reducing overall risk.

Even with modern developments in financial markets, such as energy commodities futures that more tightly link energy and financial markets and provide added options for risk management, in practice it is not possible for energy producers to completely "insure" away fluctuations in net income. Moreover, firms will not be indifferent to risks even if their basic management posture is what economists call "risk neutral." Work by Pindyck and others shows that these impacts depend on the curvature of incremental cost relationships, among other factors. For example, uncertainty about the growth of demand would shift extraction from the future to the present; if the degree of uncertainty increases with time, a limiting case would be the risk of expropriation, meaning no return from future production. Uncertainty about reserves could increase or decrease the marginal value of new discoveries compared to a situation of the same average conditions with no uncertainty.

6. MARKET FAILURES AND DISTORTIONS

Assessing the effects of various market failures and distortions on depletable energy supply is complicated by the inherently intertemporal nature of these resources. Whether a market imperfection results in depletion that is more or less rapid or extensive than the socially efficient depletion rate or volume depends on how the distortions change over time and interact with dynamic changes in resource scarcity signaled by the extraction and exploration user costs. For example, a monopolist would want to restrict output to raise price and increase returns. This would mean a lower cost of extraction in the future compared to a competitive industry facing the same demand conditions and extracting more quickly. (In actuality, demand conditions would change in response to the higher monopoly prices, as energy users adapted to more energy-efficient and alternative-energy technologies.) This, in turn, would encourage more output. Nevertheless, the monopolist could restrict output over time in such a way that the cumulative depletion of the resource (and cumulative development of new resources over time) was less than under competition. This is a different prediction than the one obtained from a simple model with fixed homogeneous reserves.

Government interventions also can tilt the market's depletion path toward the present or the future. Energy resource industries are subject to a variety of taxes in addition to the taxes paid by firms generally. One motivation for depletable resource taxation is that depletable natural resources such as fossil energy are part of the national heritage, and resource rents should accrue to the general welfare. Depletable energy resources can be subject to severance taxes per unit extracted or royalty payments as a percentage of the resource price. In general, the effect of such a tax on the intertemporal extraction pattern depends on how the present value of the tax changes over time as well as on the magnitudes of depletion effects in the cost structure. For example, the present value of a constant severance tax decreases over time and so shifts extraction from the present to the

future. However, the tax still reduces net returns in any period compared to a no-tax alternative, thereby reducing the contemporaneous incentive to extract. Although less extraction would moderate the depletion effect on cost, it is still likely that total recovery of the energy resource will fall over time as a result of the tax. In particular, reserves that were formerly marginally economic to extract become subeconomic. Moreover, the tax would reduce the return to investment in new reserves.

The extraction and consumption of energy resources are associated with a wide variety of environmental externalities. The effect of externalities on the market's depletion path can depend on the exact nature of the externality. There are two broad categories of externalities: flow externalities and stock externalities. Flow externalities represent a situation in which the environmental damage is a function of the current rate of energy production or consumption. An example relevant to energy supply would be the air pollution by-products of energy use in the process of mining. More persistent stock externalities arise when the environmental damage is a function of cumulative emissions. Examples of more persistent externalities would include atmospheric accumulation of carbon dioxide and its effect on the global climate, contamination of ground water from oil or coal extraction that is only slowly reversed by natural processes, and unremediated damage to natural landscapes through strip mining.

Flow externalities can result in a depletion rate that is too rapid or too slow relative to the socially efficient supply rate, depending on how the present value of the marginal external damage changes over time. The effect of stock externalities is similar to the impact of resource depletion on the cost of extraction: these stock effects imply a slower rate of depletion and less total depletion than the market would choose on its own.

7. EMPIRICAL ANALYSIS OF DEPLETABLE ENERGY SUPPLY

Efforts to test empirically the theoretical implications of resource depletion models as descriptions of observed energy supply behavior are hampered by a lack of data. Data for *in situ* reserve values are not readily available. Market data on the value of firms are generally available, but exclude depletable resource values, and data on extraction and development costs are usually proprietary information. Consequently, the empirical economic analysis of resource supply and scarcity has taken a variety of less direct paths. Analysts have attempted in a number of ways to use available data to construct dynamic empirical models based on the basic theoretical tenets sketched in the preceding sections.

7.1 The Hotelling Valuation Principle for Assessing Resource Scarcity

An illustration of the methods for testing the resource scarcity implications of the depletable energy supply model involves the relationship between average reserve value and current net price using cross-section data. Miller and Upton argued that this relationship can be written as follows:

$$\frac{V_0}{S_0} = \alpha + (P_0 - C_0). \tag{7}$$

Here V_0 denotes reserve value, S_0 is the resource endowment, and $P_0 - C_0$ is the price net of extraction cost; the term α, which can be positive or negative, reflects influences such as changing cost and price trends. The model was found to be consistent with pooled, cross-section data from December 1979 to August 1981 for 39 oil- and gas-producing firms in the United States. A subsequent test of the Hotelling valuation principle using data from August 1981 to December 1983 produced a quite different result: in that analysis, an increase in the net price translated into an increase in average reserve value less than half as large.

An explanation for why the Hotelling Valuation Principle might overvalue reserves, at least in the case of oil and natural gas production, is that it affords producers greater flexibility for choosing output than they actually have. The extraction of petroleum over time is restricted by declining well pressure as the reservoir is depleted. If the rate of extraction declines at the rate a because of declining well pressure, the average reserve value is

$$\frac{V_0}{S_0} = \frac{a}{a + r - g}(P_0 - C_0), \tag{8}$$

where g is the expected rate of change in net price. This example and others in the literature provide strong empirical evidence that a simple model of finite resource depletion does not adequately explain the observed behavior of depletable energy resource prices and in situ reserve values. This is not terribly surprising, given the many other features of depletable resource supply already discussed—such as exploration for and discovery of new deposits, technological

change, and capital investment—that alter the implications of finite availability. It seems clear that these other factors have overshadowed finite availability of the resource as determinants of the observed dynamic behavior of nonrenewable resource prices and *in situ* values.

7.2 Data on Reserves

In the theoretical models of depletable energy supply we have discussed, measures of reserves or resources play a significant role as indicators of cost drivers, not just in terms of physical availability. Before looking at efforts to model empirically depletable energy supply, therefore, it is worth making some brief comments on the nature of available data for reserves.

The classification scheme of the United States Geological Survey defines total resources, including energy resources, as materials that have been discovered or might be discovered and used. Reserves are that portion of total resources that can be economically extracted. Undiscovered resources are classified as hypothetical, if in a known mining district, or speculative. Identified but currently noneconomic resources are categorized as paramarginal or submarginal. The physical measures of reserves often are compared to measures of the rate of use in order to determine the remaining life of the reserves.

Because reserves are defined in terms of economic recovery, whether a deposit is classified as a reserve changes with the resource price and extraction cost. In addition, because costly investment is required to "prove" reserves, there is a limited incentive to prove reserves beyond a certain point. The ratio of reserve to consumption for petroleum increased from 35 in 1972 to 45 in 1990, even though commercial energy consumption increased by more than 50% between 1971 and 1991. Physical measures of reserves thus probably have more meaning as an indicator of inventory on hand than as a more fundamental indicator of resource cost.

7.3 Models of Depletable Energy Supply Behavior

Behavioral models of depletable energy or other resource supply seek to develop empirical counterparts to theoretical efficiency conditions [such as Eqs. (4)–(6)] in order to describe how supply decisions (reserve development and extraction) respond to economic incentives (prices and cost drivers). Much of this work has focused on oil and gas supply, though some applications to mineral resources (e.g., nickel) also can be found in the literature. Relatively little of this kind of modeling work has been done for coal, because the depletion effects central to the development of the dynamic efficiency conditions (discussed previously) seem to be of very limited relevance to coal supply (with large reserves and continued technical progress in resource exploitation).

The problem of developing empirical behavioral models based on the theory of efficient depletable energy supply is quite difficult for several reasons. One problem, noted previously, is the poor data on reserves and costs. But even with better data, it is inherently complicated to develop reliable empirical models for complex dynamic process, and resource depletion is inherently dynamic. Moreover, because behavior depends so much on expectations of unknown future influences in this dynamic framework, the empirical analyst must formulate a hypothesis about the nature of the expectations driving behavior.

Finally, the theory discussed previously applies most directly to an individual decision-making unit and perhaps even to decisions within a specific resource-bearing location. Once the analysis is extended to cover multiple extraction and investment decisions from heterogeneous deposits, some difficult aggregation problems arise. Care is needed not to omit or mix together key influences. For example, the yield from exploratory drilling likely depends on both how much the best resource prospects have been depleted by past activity and what advances in drilling technique can do to raise discoveries per unit of effort.

Basically, two types of modeling approaches are found in the literature: process optimization simulation models and econometrically estimated models. Process optimization simulation models specify and numerically solve sets of equations based on first-order efficiency conditions such as Eqs. (4)–(6). For example, the model could incorporate a specification in which the reserves discovered per foot of new drilling depend on cumulative past drilling in that province. This specification could be used along with other information to determine the efficient level of reserve development as a function of the direct cost and user cost of drilling, plus the marginal (shadow) value of additional reserves. One prominent example of such a model in the United States is the oil and gas supply component of the National Energy Modeling System (NEMS) operated by the Energy Information Administration (EIA); Deacon has provided a smaller-scale example.

Numerical parameters in such models come from a variety of sources: statistical curve fitting, engineering experience, and informed judgment, among others. Because these models are not derived from statistically fitting equations to data, it is not possible to construct classical confidence intervals and statistics for assessing the reliability of the model outputs. They are useful principally for "what if" exercises—by supplying different sets of input variables (e.g., oil prices and technical recovery parameters for drilling), the consequences of changes in these assumptions can be seen. Whether the model structure is properly specified empirically cannot be independently assessed. The simulation models also do not incorporate uncertainty and expectations formation. Implicitly in dynamic optimization models, individuals are assumed to judge the future accurately.

Econometrically estimated models also are rooted in the theoretical efficiency conditions. However, these models are parameterized using statistical methods that seek to define supply decisions as functions of causal variables. The most sophisticated of these frameworks explicitly model expectations formation using the rational expectations time series approaches pioneered in macroeconomics by Epple, Hendricks and Novales, and Walls. In brief, these methods explicitly include representations of expectations formation based on past observations of price and cost influences; assuming that producers are not systematically biased in their expectations and fairly rapidly learn about a systemic change in market or technological conditions, statistical techniques can be used to estimate the parameters of the expectations process and the underlying parameters linking supply choices to the relevant economic influences. The point is to provide consistent estimates of the basic technological and geological parameters undergirding the supply process, so that the effects of policy intervention or shifts in market conditions can be separated out and understood. However, success with this class of models has been limited, notwithstanding their conceptual appeal. The models are difficult and data intensive to construct and estimate; they are not well suited to geographical or other finer grained breakdowns of activity into constituent parts; and the results have not always been very powerful. In particular, some approaches have not coped well with what appear to be basic nonlinearities in the relationship between unit supply cost and reserves. A prospect for future work is a better blending of the engineering expertise found in energy systems modelers with the dynamic economic models discussed here.

8. CONCLUDING REMARKS

Finite availability is one defining characteristic of a depletable energy resource and generates the "Hotelling rule" that the marginal value of a nonrenewable resource stock increases at the rate of interest. However, many other factors, including exploration, capital investment, and heterogeneous reserve quality, also are important to the economics of energy depletion and development. The investigation of how these other factors affect the empirical implications of the Hotelling model has been spurred by the frequent failure of the basic Hotelling model to explain the observed dynamic behavior of nonrenewable resource prices and *in situ* values.

These other factors can affect price and depletion paths in a number of ways, particularly when considered in combination with each other. The variety of possible outcomes makes it difficult, if not impossible, to make any general predictions about the overall impact on price and extraction paths. These other factors, particularly the discovery of new deposits and technological progress that lowers the cost of extracting and processing nonrenewable resources, appear to play a relatively greater role than does finite availability in determining observed empirical outcomes. Although models that include these other factors have enriched the Hotelling model, they have not completely reconciled the economic theory of nonrenewable resources with the observed data. The distinction between the responses of supply to anticipated versus unanticipated changes in extraction cost, interest rate, reserve discoveries, availability of backstop substitutes, and other factors figures prominently as a topic for future empirical research. The empirical evidence also indicates that the discovery of new deposits and technological progress in resource extraction and discovery have significantly mitigated the impacts of finite geological availability on the relative cost and supply of depletable energy resources. Although the future holds promise with respect to the development of new substitutes for fossil energy and greater energy efficiency, improvements in extraction and exploration technologies that reduce costs and allow the economical use of lower grade reserves also are likely to continue.

SEE ALSO THE FOLLOWING ARTICLES

Depletion and Valuation of Energy Resources • *Economic Geography of Energy* • *Economic Growth*

and Energy • Economics of Energy Demand • Energy Futures and Options • Market Failures in Energy Markets • Modeling Energy Supply and Demand: A Comparison of Approaches

Further Reading

Adelman, M. A. (1990). Mineral depletion with special reference to petroleum. *Rev. Econ. Statist.* **72**(1), 1–10.

Adelman, M. A. (1993). Modelling world oil supply. *Energy J.* **14**(1), 1–32.

Bohi, D. (1999). Technological improvement in petroleum exploration and development. *In* "Productivity in Natural Resource Industries" (R. D. Simpson, Ed.). Resources for the Future, Washington, D.C.

British Petroleum (2001). BP Statistical Review of World Energy, June. Available on the Internet at http://www.bp.com/(accessed November 12, 2002).

Brobst, D. A. (1979). Fundamental concepts for the analysis of resource availability. *In* "Scarcity and Growth Reconsidered" (V. K. Smith, Ed.), pp. 106–159. Johns Hopkins University Press (for Resources for the Future), Baltimore.

Cairns, R. D. (1990). The economics of exploration for non-renewable resources. *J. Econ. Surv.* **4**(4), 361–395.

Cairns, R.D., Lasserre, P. (1990). Investment in a multiple-deposit non-renewable resource industry: Some results and remaining puzzles. *J. Environ. Econ. Manage.* 21(1), 52–66

Darmstadter, J. (1999). Innovation and productivity in U.S. coal mining. *In* "Productivity in Natural Resource Industries" (R. D. Simpson, Ed.). Resources, Washington, D.C.

Deacon, R T. (1993). Taxation, depletion, and welfare: a simulation study of the U.S. petroleum resource. *J. Environ. Econ. Manage.* **24**(2), 159–187.

Energy Information Administration. (2000). "The National Energy Modeling System: An Overview 2000." U.S. Department of Energy, Washington, D.C.

Epple, D. (1985). The econometrics of exhaustible resource supply a theory and an application. *In* "Energy Foresight and Strategy" (T. J. Sargeant, Ed.). Resources for the Future, Washington, D.C.

Epple, D.,Londregan, J. (1993). Strategies for modeling exhaustible resources supply. *In* "Handbook of Natural Resource and Energy Economics," Vol. 3 (A. V. Kneese and J. L. Sweeney, Eds.). Elsevier, Amsterdam.

Farzin, Y. H. (1984). The effect of the discount rate on depletion of exhaustible resources. *J. Politic. Econ.* **92**(5), 841–851.

Heal, G. M. (1976). The relationship between price and extraction cost for a resource with a backstop technology. *Bell J. Econ.* **7**(2), 371–378.

Hendricks, K., and Novales, A. (1987). Estimation of dynamic investment functions in oil exploration. Stanford Hoover Institute Working Papers in Economics, pp. E87–E52. Hoover Institution, Stanford, CA

Hotelling, H. (1931). The economics of exhaustible resources. *J. Politic. Econ.* **39**(2), 137–175.

Krautkraemer, J. A. (1998). Nonrenewable resource scarcity. *J. Econ. Lit.* **36**(4), 2065–2107.

Lasserre, P. (1985). Capacity choice by mines. *Can. J. Econ.* **18**(4), 831–842.

Livernois, J. R., and Uhler, R. S. (1987). Extraction costs and the economics of nonrenewable resources. *J. Politic. Econ.* **95**(1), 195–203.

Miller, M. H., and Upton, C. W. (1985). A test of the hotelling valuation principle. *J. Politic. Econ.* **93**(1), 1–25.

Miller, M. H., and Upton, C. W. (1985). The pricing of oil and gas: some further results. *J. Finance* **40**(3), 1009–1018.

Pindyck, R. S. (1978). The optimal exploration and production of nonrenewable resources. *J. Politic. Econ.* **86**(5), 841–861.

Pindyck, R. S. (1980). Uncertainty and exhaustible resource markets. *J. Politic. Econ.* **88**(6), 1203–1225.

Powell, S. G., and Oren, S. S. (1989). The transition to non-depletable energy social planning and market models of capacity expansion. *Oper. Res.* **37**(3), 373–383.

Slade, M. E. (1982). Trends in natural resource commodity prices: an analysis of the time domain. *J. Environ. Econ. Manage.* **9**(2), 122–137.

Slade, M. E. (1987). Natural resources, population growth and economic well-being: Issues and evidence. *In* "Population Growth and Economic Development" (D. G. Johnsson and R. D. Lee, Eds.), pp. 331–372. University of Wisconsin Press, Madison.

Swierzbinski, J. E., and Mendelsohn, R. (1989). Exploration and exhaustible resources: the microfoundations of aggregate models. *Int. Econ. Rev.* **30**(1), 175–186.

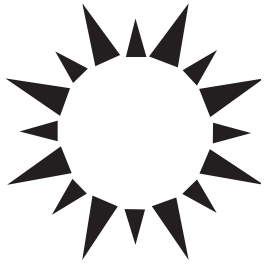
Toman, M. A., and Walls, M. (1995). Nonrenewable resource supply: theory and practice. *In* "Handbook of Environmental Economics" (D. W. Bromley, Ed.). Blackwell, Oxford and Cambridge.

Walls, M. (1992). Modeling and forecasting the supply of oil and gas: a survey of existing approaches. *Resour. Energy* **14**(3), 287–309.

Walls, M. (1994). Using a "hybrid" approach to model oil and gas supply: a case study of the gulf of mexico outer continental shelf. *Land Econ.* **70**(1), 1–19.

World Resources Institute (1994). "World Resources 1994–95: People and the Environment." Oxford University Press, Oxford.

Economic Systems and Energy, Conceptual Overview



JOHN PEET
University of Canterbury
Christchurch, Canterbury, New Zealand

1. **Physical Foundations of Economic Activity**
2. **Energy in the Economic System**
3. **The Economics and Physics of Production**
4. **Energy Analysis of Economic Processes**
5. **Physical Accessibility and the Concept of Net Energy**
6. **Energy and its Importance to Policies for Sustainability**
7. **Aims of the Biophysical Systems Worldview**

Glossary

direct energy The energy used in the specific process of production of a good or service.

embodied energy The energy associated, directly or indirectly, with the production of a good or service.

energy analysis A family of tools that follow economic production and consumption processes through their (physical) energy consequences.

energy carrier (also energy vector) A portable substance that can release energy when desired (e.g., liquid petroleum fuel).

indirect energy The energy used in preparation of intermediate inputs to a process.

net energy The amount of energy available from an energy transformation system after deduction of the energy expended in creating the capital equipment for mining and refining and the energy used in operating the system.

Human society is a complex system, in turn embedded in another—the natural environment. Since humans are totally dependent on the natural environment for the necessities of life, and since human economic activities strongly influence both human and ecosystem health, we need to be well informed about key aspects of the state of the human system and the natural "supersystem" within which humans exist. One such key aspect is associated with the flows of energy, both natural and human influenced, throughout the biosphere.

Economic systems depend on the availability and use of energy resources for their every function. In this context, the ability of a form of energy to "do work" in a thermodynamic sense (e.g., run a machine) is often referred to as its availability (or exergy). Highly "available" forms of energy such as electricity are better able to perform a range of important functions in the economy than, for example, fuelwood, which has a significantly lower availability. The energy in seawater is much less available, and although the overall quantity is huge, its ability to do economically useful tasks is small. The term quality is often used to describe a fuel's economic usefulness (e.g., as gross domestic product produced per heat unit of fuel used), and although this is related to its ability to do work in a strict thermodynamic sense, it is not the same thing.

From the mainstream economic standpoint, the economy is seen as a system of production, organized according to a system of signals in the form of prices paid by consumers to producers in order to obtain a good or service. These prices are determined in a marketplace, where willing buyers and willing sellers reach an understanding of the price to be paid for the product. Anything that is scarce has a price, provided ownership can be established.

On the other hand, from the biophysical systems standpoint, an economy is seen as a social system characterized by the physical activities that take place in it. That social processes are involved in the myriad of decisions made in that society is to some extent secondary to the physical processes that are occurring. From this perspective, physics, particularly thermodynamics, provides the tools for understanding both what is happening now and what is either possible or impossible, in physical terms, in the future. Clearly, the two perspectives are very different; part of the purpose of this article is to show

 © 2004 Elsevier Inc. All rights reserved.

how at least a partial integration of perspectives is possible.

1. PHYSICAL FOUNDATIONS OF ECONOMIC ACTIVITY

The term energy not only applies to wood, coal, oil, gas, and electricity but also has a fuller scientific meaning. This is because the laws of thermodynamics apply, without exception, to all physical processes on Earth. They are true whether applied to the actions of people and machines in an economy, all forms of life, circulation in the atmosphere and the oceans, or movements of the planets. Inputs of available energy are essential to the functioning of any economy, not only directly as fuels but also indirectly through the natural solar cycles that provide the "ecosystem services" (air, water, soil, etc.) essential for all forms of life, including the economic.

The associated transformation of available into unavailable energy resources (in accordance with the second law of thermodynamics) is an unavoidable part of the functioning of all economies. There is no process known that did not begin by using some form of high-quality raw material (available energy) and that will not end up somewhere as waste (unavailable energy). Directly or indirectly, most source energy comes from or came from the sun, and all sink energy is eventually dissipated to outer space. Heat from the earth's interior is also an important source of energy for some economies, and it is obviously relevant in many geological processes important to human life.

Early human societies evolved in a context in which they obtained the essentials of life from their surroundings via natural mechanisms by which most of the products they used were obtained, directly or indirectly, from the sun. Needs for food, clothing, shelter, and so on were met from plants and animals by hunting and gathering, and humans traveled or migrated as necessary to ensure continued supply. Early industrial societies also used solar-derived energy resources, especially wood as fuel, often converted to charcoal for more specialized purposes (e.g., high-temperature metal smelting). By the 18th century, large-scale mining of coal, together with associated technologies for the production of iron and steel, was fueling the industrial revolution and the rapid economic and population growth that eventuated. Later technologies, similarly based on prior use of high-quality energy resources, enabled exploitation of oil and gas, in turn powering further economic growth. The ability of these fuels, together with their associated technologies, to do "work" in economic production processes vastly increased the ability of humans to control and transform their physical environment.

Each stage of industrialization was built on a foundation of previous industrialization. Without exception, all required prior consumption of available energy from the environment and, implicitly, the "investment" of that energy in the creation of the artifacts (machines, processes, and buildings) needed for physical production of goods and services.

From the physics standpoint, it is also important to note that what the marketplace calls energy is actually consumer fuels (coal, oil, gas, and electricity) since they are the form in which energy is traded commercially. In the generalized scientific sense, however, energy has no substitutes. Only fuels, which can be seen as "energy carriers," can often be substituted, one for the other (e.g., oil for coal). An energy carrier is a portable substance that can release energy when desired (e.g., petroleum fuel such as gasoline).

The interaction between energy and matter is the basis of life; therefore, it is also the basis for human life and culture. However, this is not to suggest that life is only energy. Far from it. It is nevertheless vital to recognize the key role of energy in the physical world because otherwise it is fatally easy to fall into the trap of thinking that all processes of importance in human decision making are internal to the economy. Unfortunately, this view is typical of all major economic theories. This does not leave much room in the political spectrum for other views, but it is essential to do so and thereby add important information to both.

Probably the most notable exponents of the physical basis of economics are Nicholas Georgescu-Roegen, Kenneth Boulding, and Herman Daly, economists who have made extensive contributions to the literature on physical limits to the use of resources in economic production.

The inclusion of energy in a worldview highlights the help that the biophysical systems perspective can provide in clarifying questions about the relationship between a social system and its physical environment. It also helps us see where the biophysical viewpoint contrasts with the political–economic and thus indicates areas where conventional assumptions may need deeper examination.

2. ENERGY IN THE ECONOMIC SYSTEM

Starting with primary resources such as coal, oil, and gas in the ground, raw energy is processed into "consumer" form, suitable for use in society, via power stations, cables, oil refineries, gas pipelines, and so on. The resultant fuels, such as gasoline, piped natural gas, and electricity, are then used to produce energy services—transport, warmth, cooking, lighting, and so on.

Like energy, matter also enters an economic system as raw resources from the environment. Even if temporarily accumulated within the economy (e.g., embodied in capital goods) or recycled for awhile within the economic process, it must be disposed of. This gives rise to pollution, acknowledged by the concept of externalities in economic terminology. These are also seen as feedbacks in systems terminology.

The flow of matter into and through an economic system can be characterized in terms of the high-quality available energy required to carry out the transformation processes in which the matter is involved. Thus, energy consumption can be used as a "numeraire" enabling a wide variety of goods and services to be accounted for in terms of the quantity of one important characteristic, the energy expended in the processes of transforming them into marketable products. From this viewpoint, capital accumulation and stocks of durables are seen as a temporary store for energy resources. In time, these depreciate and wear out, and they eventually end up as waste.

In physical terms, as in economic terms, energy is only of value by virtue of its ability to cause physical or chemical change or to do "work." From the biophysical systems perspective, what is important is the amount (i.e., the physical cost) of energy used in a process and thereby embodied in a product, not its price in the marketplace. In contrast to the political–economic perception, however, energy analysis follows only the physical processes of production and not the social valuation of the product in the marketplace.

Energy is a means of following the way in which an economy works, usually via techniques of energy analysis. Embodied energy is the energy associated, directly or indirectly, with the production of a good or service. In this context, direct energy is the energy used in the specific process of production; indirect energy is used in the preparation of intermediate inputs to that process. Although the term embodied energy is widely used, it has many different interpretations, depending on the context.

The use of energy as a complementary means of describing the workings of a social system can also be seen as an attempt to address the apparent inability of standard economics to take full account of the central position of physics in economic production.

3. THE ECONOMICS AND PHYSICS OF PRODUCTION

From a conventional economic perspective, the marketplace is where decisions are made about the uses of energy. Energy is simply a tradable commodity, like all scarce resources, the demand for which is affected by its price. Everything the economy needs to know about energy is incorporated in its price, and the freer and more competitive the market, the more accurate is the price evaluated by it.

Goods are said to be substitutes if (other things being equal) an increase in the price of one leads to a shift toward consumption of another, whose price is lower. Thus, if an increase in the price of coal causes some consumers to switch to natural gas as a heating fuel, these are substitutes. Note, however, that in the generalized scientific sense, energy has no substitutes. Only fuels, which are energy carriers, can often be substituted, one for the other.

Energy consumption has been used for decades as an indicator of the level of wealth of an economy. As a broad generalization, it appears that there has often been a strong relationship between energy consumption and gross domestic product (GDP), with the ratio E:GDP commonly known as the energy intensity of GDP. There are problems with this criterion, however, because some very different countries have approximately similar energy demands, whereas other similar countries have different demands. Nevertheless, greater energy consumption was long viewed as a necessary prerequisite for economic growth.

The relationship between energy consumption and economic activity is significantly less direct than was the case in the past for a number of reasons, including substantial improvements in the technical efficiency of use of energy, especially in technologically advanced countries that previously consumed large quantities of fuel. In some cases, this increase in efficiency was the result of switching to other, higher quality fuels, which enabled a greater amount of GDP to be produced per unit of fuel used.

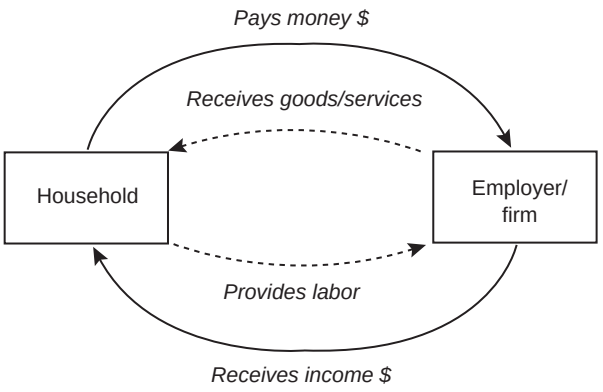


FIGURE 1 The basic macroeconomic model.

3.1 Economic Models of Production

The circular flow diagram of Fig. 1 characterizes the conventional model of the economic process. In this model, goods and services made by employers/firms (producers) are sold to households (consumers), who in turn obtain money either by selling their labor to or investing their savings in the producing firms (often via financial institutions).

The model is developed further in Fig. 2, which shows the distinction between different types of households and introduces government and financial institutions. The model is nevertheless conceptually the same as that in Fig. 1. In this model, money flows, which represent the tokens that are exchanged for goods and services, labor, and capital, are one representation of reality. It is easy to forget that the underlying reality is of goods and services and people, which is apparently forgotten in some economics textbooks and is implicit in Fig. 2. The whole economy then becomes something that is described and quantified in monetary terms, with the total of "net income" transactions being equal to GDP.

The model of Fig. 2 is actually a map of the flows of exchange value in an economy—no more and no less. Exchange relationships between capital (employers/firms) and labor (households) are the primary determinants of behavior of the system.

This model of production and consumption linked in a circle of flows of exchange value can be modified by acknowledging the essential role played by the flows of raw resources from the environment through the economy and out as pollution, as shown in Fig. 3. Although the monetary exchange system is circular on one level (money being exchanged for goods), the physical products start at one place and end up somewhere else. In the process, much energy is consumed, and major disposal problems are generated. This model therefore illustrates the "metabolic" aspect of economic activity, in which the economy is "fed" by raw resources from the environment and "excretes" wastes back to that environment.

The model of Fig. 3 is also the basis for the application of conventional welfare economic concepts via resource and environmental economics. In both, the key issue is that of clarifying property rights as an essential step toward the evaluation of

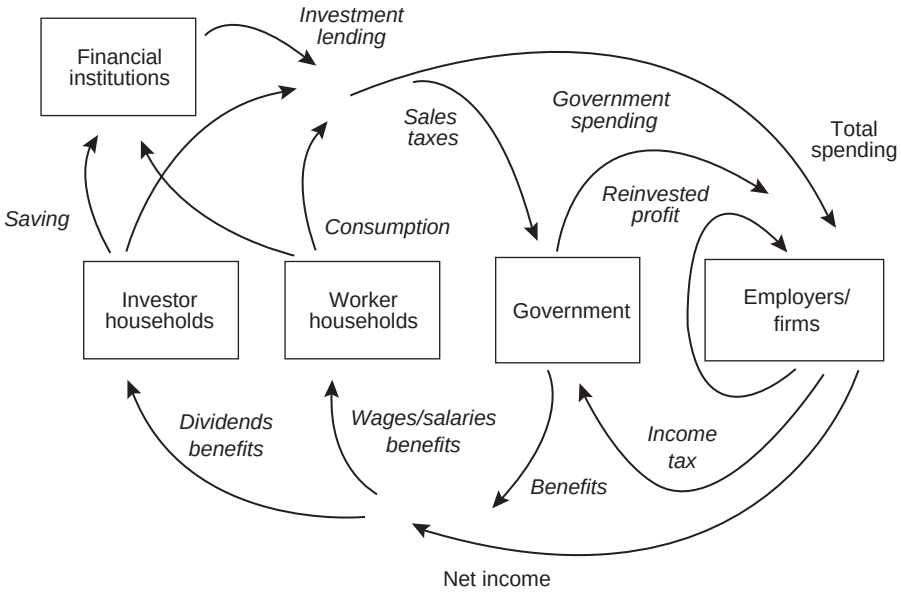


FIGURE 2 The extended macroeconomic model.

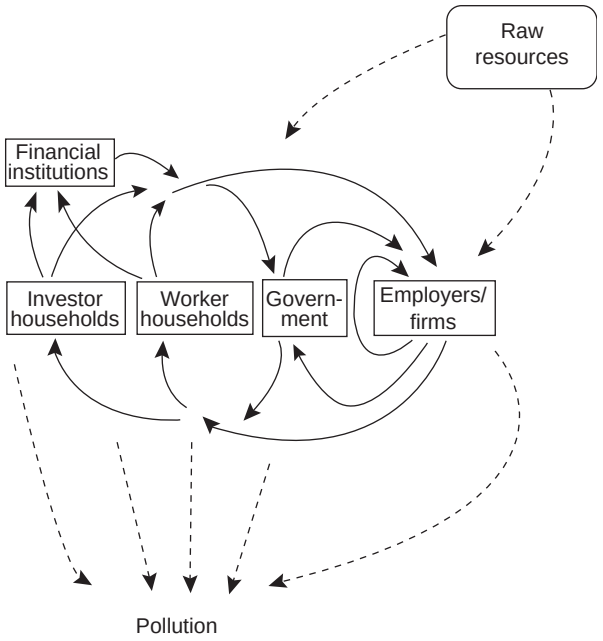


FIGURE 3 Resource requirements of economic activity.

"externalities." In environmental economics, the solution to environmental problems becomes one of creation of markets in pollution rights and ecosystem services on the principle that damage must be paid for, which acts as a disincentive to pollute. Related tools in resource economics involve application of discounting of the value of future events such as resource depletion in order to determine the most economically efficient use of resources over time.

Therefore, there is a sense in which the mainstream economic model regards the environment as simply a subsystem of the economy—a source of resources and sink for wastes, with the primary purpose being to enable the economy to grow without limit. Figure 4 illustrates Daly's critique of this perception. The special place of energy in the biophysical worldview is completely absent from this model, other than as just another resource.

All these attempts to incorporate the reality of environmental pollution into social policy development have involved biophysical issues interpreted via economic theories. From a scientific viewpoint, one may observe that in this approach the needs of the total system are interpreted in terms of those of one of its (dependent) subsystems (the economy) for reasons that appear to owe little to science. Nevertheless, there is little doubt that if properly applied, tools such as environmental and resource economics could give rise to more environmentally and socially benign outcomes than current outcomes via the use

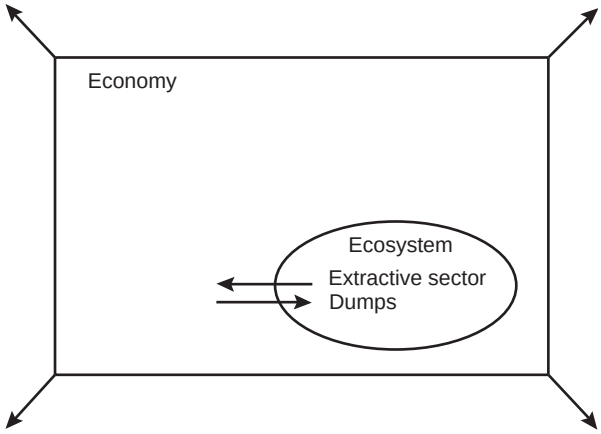


FIGURE 4 Ecosystem as subsystem of the economy. From Daly (1999), with permission.

of improved technologies for converting resources into end-use human services.

3.2 Physically Modified Economic Models of Production

The model in Fig. 3 has limitations due to the particular place energy has in the processes of production and consumption. If we separate out raw energy resources from other material resources, we obtain a different picture, despite the fact that some raw energy resources, such as coal, oil, and gas, may also be raw resources in their own right (e.g., when used for petrochemicals production). This is not a problem as long as we remember the distinction between the two functions.

Figure 5 indicates the special function of the energy sector as an energy transformation system. A flow G of raw energy resources (coal, oil or gas in the ground, water in mountain lakes, geothermal heat, solar input, and so on—often referred to as primary energy) is transformed into a flow E of consumer energy (often referred to as secondary energy)—such as delivered, refined fuels, electricity, process heat, etc.—transformed in coal mines, oil/gas wells, refineries, thermal power stations, hydro dams, boilers, and so on.

These processes always involve loss of some of the input of raw energy resources, as a direct reflection of the processes involved, such as oil drilling, pumping and refining, thermal or other electricity generation, and so on. Some of the consumer energy output of these plants is often needed to run the plants (e.g., electricity to run auxiliary machinery in a thermal power station). Note that flows of waste energy and material pollution are not included in this simplified model but are also important.

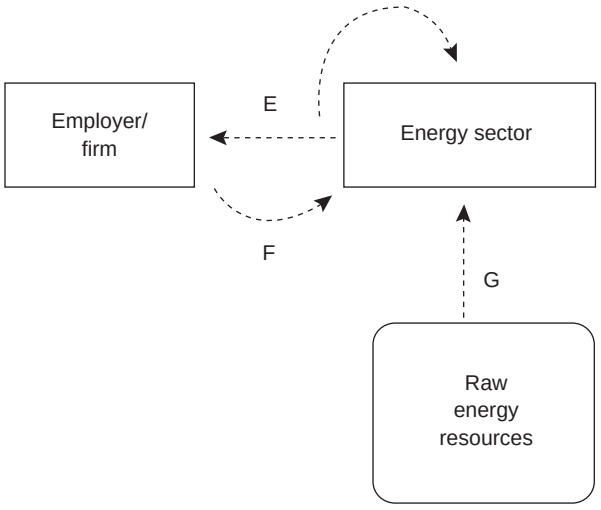


FIGURE 5 Energy transformation system.

The energy transformation system comprises large capital works constructed from steel, concrete, and other materials, in turn created using energy and material resources in the past. Figure 5 includes a feedback arrow showing the energy flow *F* embodied in the goods and services (as manufactured capital, etc.) needed from the production sector to build and maintain the energy sector.

Even consumer, or secondary, energy is not the real end product. What economic production or consumption processes actually need from energy is the services that energy provides, such as heat, light, motive power, chemical conversion, and operation of electronic equipment. This is illustrated in Nørgård's model in Fig. 6, which shows the chain of conversion of raw, primary energy through secondary energy to energy services and to the ultimate end, the energy services and other inputs to lifestyle and satisfaction of needs. In this model, the fact that energy is a means to an end is very clear. Also very clear is the importance of efficiency of conversion at each stage and the fact that each increment of satisfaction of needs requires a substantial infrastructure of capital equipment and a steady consumption of raw resources. Sooner or later, there is an associated and inevitable conversion of all the resources into waste.

Gilliland's model (Fig. 7) integrates the perspective of Fig. 5 into the economic model of Fig. 2 and shows how a significantly richer picture of economic activity is obtained if resources are separated into their material and energy components. This thermophysical model shows the economy as a pathway by which energy resources are used to run the processes

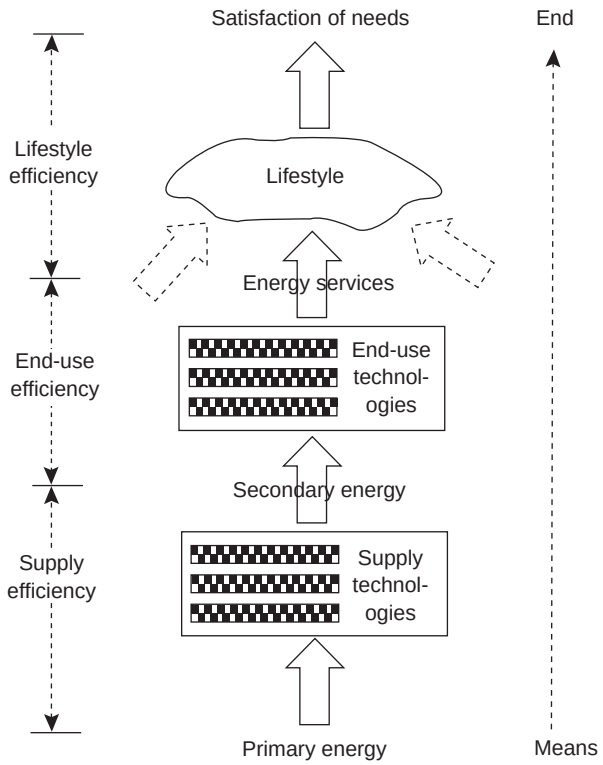


FIGURE 6 Conversion of primary energy to energy services. Courtesy of Nørgård, J. S.

that convert raw materials into the goods and services needed for everyday functioning of the economic system.

In Fig. 7, it is clear that money flows in the opposite direction as energy. In addition, although money flows are circular, energy flows are essentially linear (in one end and out the other), as is typical of all metabolic processes.

In the thermophysical context, energy use—or, more correctly, the degradation of available input energy into unavailable output energy—can be seen as representing physical "cost." However, this is a different concept from money flow in the economic perspective, which normally represents market price. Therefore, it is not suggested that energy consumption is an adequate measure of social value in an economic context, although undoubtedly it can have significant long-term implications for prices.

Again, however, to make our approach operational, it is desirable to view the place of important component subsystems in our conceptual model of the system of society within the supersystem environment. Figure 8 shows the thermophysical economy of Fig. 7 within its total ecological environment of planet Earth. The energy concept is central to this description of the thermobiophysical "engine" that enables

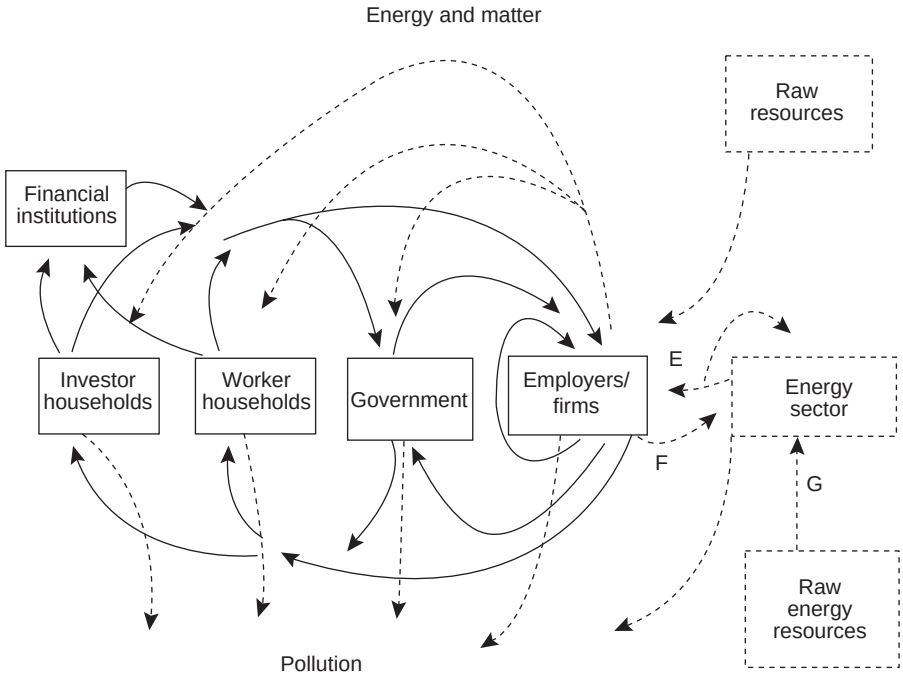


FIGURE 7 The thermophysical macroeconomic model of Gilliland.

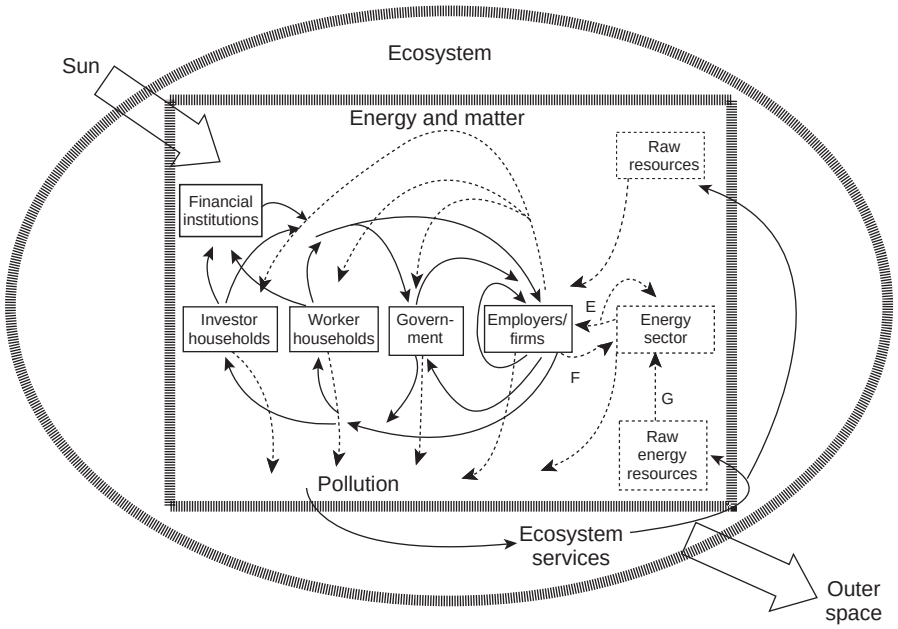


FIGURE 8 Biophysical economic model as part of planet Earth.

the global ecosystem to function, with high-quality incident solar energy driving all the natural processes on which the activities of living things depend, before eventually being rejected as waste heat into the vast sink of outer space. Within the earth's system, that energy is stored for a short time in processes such as circulation of the atmosphere, for longer times in plants that absorb energy for photosynthesis, and for even longer times in geological deposits.

The energy in fossil fuels was originally derived from solar energy many millions of years ago, captured mainly through photosynthesis in plants

and then via the plants (and the animals that fed on the plants) whose remains were transformed over very long periods of time by geological processes into the hydrocarbon fuels coal, oil, and gas. The fact that humans currently have access to these resources (for a period that will be very short in the context of geological or even human time) does not affect the validity of Fig. 8 as a general model.

4. ENERGY ANALYSIS OF ECONOMIC PROCESSES

An important characteristic of an energy transformation sector is that it functions by absorbing capital, operating, maintenance, and other inputs from the economy, represented in Fig. 5–7 by the feedback flows F of energy embodied in those goods and services. It is only when the flow of useful, consumer energy E is greater than the sum of embodied feedback flows that there is a supply of net energy to the economy. This concept is analogous to one common in economics in that revenue or turnover alone are not enough to indicate financial success of a project; net cash flows or profit are important.

Energy analysis comprises a family of tools that follow economic production and consumption processes through their energy consequences, usually expressed in terms of the first law of thermodynamics (i.e., the heating abilities of energy). More advanced forms of energy analysis, such as exergy analysis, address the ability of energy to do work in terms of the second law of thermodynamics.

In its basic form, energy analysis involves the determination of the energy embodied in the process of delivering consumer energy or end-use energy or in making a good or service. As Fig. 7 indicates, both the losses of raw energy and the energy required to run the energy sector may be substantial. In the case of coal-fired thermal electricity generation, approximately two-thirds of the raw (primary) energy is dissipated in producing the consumer (secondary) electricity output, even before the efficiency of conversion into end-use forms is taken into account.

There are two common techniques for evaluating the embodied energy of goods and services—process analysis and input–output analysis. Process analysis involves the systematic study of the inputs to and outputs from a process, with determination of energy requirements for each input. These are then summed to give the total energy requirement of the product.

For example, a process energy analysis of running a home would first concentrate on electricity and heating fuel consumption since these are the most obvious large energy inputs. Less important direct inputs would then be assessed. Energy for agriculture, harvesting, processing, transport, merchandising, and cooking of food also needs to be evaluated, together with the energy involved in the manufacture of clothing, furniture, cleaning materials, paint, carpet, and so on from their raw materials and the energy to transport and deliver them to shops and the home. If one goes back further, one could consider the energy required to make or process the house construction materials (steel and concrete for foundations, lumber and other materials for the structure, and the energy needed to build the machines and factories where these parts were made) and, before that, the energy required to build and run the mines, forests, and so on from whence came the raw materials, and so on.

Clearly, the task is daunting, and few process analyses have been carried out beyond the first or second stage. This technique is good at giving an accurate estimate of direct energy inputs but not for evaluating the indirect inputs, such as the energy required to process the timber for the building, the lumber mill, or the steel mill or the mines from which the iron ore came to make the steel. These indirect inputs can often be very large, especially in the service-sector industries, which on the surface often appear to have relatively small energy demands but whose indirect demands may be many times higher.

The second common technique is input–output energy analysis, a modification of a standard economic tool. This employs a detailed survey of transactions in an economy. In effect, it addresses the production part of the economy, with explicit accounting for the energy inputs required to keep the economy going. By a straightforward computer-based method, all the energy inputs to the processes that preceded the output of a given good or service in the economy are summed, back to the coal mine, oil well, power station, or gas field.

In general, process analysis is appropriate for study of specific production technologies and is a technique widely used in engineering. Input–output analysis is more appropriate for examining the effects of changes in technology, resource use, etc. for a whole economy. The difference in quality, or availability to do work, of the various forms of energy is not specifically determined by either approach, although it is possible to evaluate it.

Both techniques are involved, as appropriate, in applications such as life-cycle analysis.

Incorporation of the extra information gained from a material/energy analysis inevitably raises problems of commensurability of units because activities within the economy are conventionally represented in monetary units, whereas those in ecological and industrial systems are better represented by measures of material and energy/entropy flows. The emerging field of ecological economics addresses these and related issues.

5. PHYSICAL ACCESSIBILITY AND THE CONCEPT OF NET ENERGY

There are significant differences in the ways different researchers measure the net energy performances of energy transformation systems that deliver a flow of consumer (secondary) energy E to an economy (Fig. 5) from a raw (primary) energy resource flow G using a feedback flow F. However, the net energy criterion (in whatever form) is the means of indicating the physical accessibility of a resource. It is the amount of effort (measured in energy terms) required to extract the resource, process it as required, and deliver it to the consumer. The greater the effort, the lower the accessibility of the resource. As indicated previously, production of electricity from coal via conventional thermal power stations results (in heat equivalent terms) in only approximately one-third of the energy in the mined coal being delivered to consumers in the form of electricity. However, the ability of the electricity to do work, especially running machines and electronics, makes it a more valuable energy source to an economy than the raw coal.

During the past three centuries, economies in industrialized countries have changed from an almost complete dependence on biomass-based fuels (mainly wood) to coal and then to oil and natural gas. Each transition was characterized by change to a more accessible energy source, but that change was not possible until technology developed to the point at which the improved accessibility was achievable. Thus, coal could not be supplied in large quantities until underground mining technology was developed, and open-cut coal required heavy earth-moving machines to be developed. Oil and gas were not accessible until drilling, transportation, and refining technologies were applied. All of these depended heavily on the development of systems for long-distance transportation of both raw materials and products.

In each case, the accessibility of the new resource (measured by both net energy and economic methods) was better than that of the previous one. This meant costs were lower, so substitution could proceed rapidly after the technology developed. This outcome is believed by some to be the result of factors internal to economic systems. To the extent that technological developments were required, this is partially true. The underlying physical accessibility mechanism is a separate issue, however, subject to the laws of physics and should not be excluded from the argument.

There is increasing evidence that energy is becoming more difficult to obtain and process into readily usable forms. Geological resources are not homogeneous, and humankind has exploited the easily accessible, high-quality resources first. Lower quality, less accessible resources are left for future generations in the expectation that the price mechanism will offset any shortages.

As an economy changes its sources over time from high-quality (high-availability) energy sources to low-quality (low-availability) sources, the second law of thermodynamics indicates that the efficiency with which work may be done by these resources will decline. In addition, the change from resources that are easy to mine to those that are deeper and/or more difficult to mine means that a greater amount of economic (e.g., engineering) effort will have to be expended in order to tap these resources and to supply useful net energy to the economy. Thus, the resources become less accessible. In both cases (Figs. 5–7), a greater flow of either G or F (or both) is needed to maintain a constant flow E. Over time, the energy transformation industry will have to grow and absorb a steadily increasing quantity of economic output just to maintain the system.

If there is a steady increase in the resources that need to be expended from the economy (flow F) in order to obtain a given quantity of energy E, then in the long term one can envisage a situation in which no net energy is supplied. In other words, the energy content (embodied in buildings, machinery, maintenance, and operation) of the economic effort needed to extract energy and deliver it could increase to such a level that as much embodied energy flows back into the energy supply industry as is supplied to the economy. If $F = E$, a point of futility has been reached, where much socioeconomic effort is going into a process equivalent (in energy terms) to going around in circles.

When resources are relatively inaccessible, the primary resource flow will be greater than that for

the earlier, high-accessibility case. Thus, the rate of depletion of primary resources will also increase for a given (constant) flow of net energy to the economy. Pollution and other externalities will also be greater due to the increased rejection of unavailable energy.

Of course, technical change can often act to lower energy and other costs, thereby offsetting to a greater or lesser extent the depletion effect. Improved technologies may also have an energy cost, however, especially if they involve heavy capital investment with the associated indirect energy costs of manufacture. Nevertheless, many such technologies show substantial benefits. Although very valuable in the short term, in the longer term the constraints set by the laws of physics must prevail.

Thermodynamics is concerned with the properties of systems, not of things on their own. In energy analysis, the expenditure of energy and time to do work is a linking mechanism between the natural environment system and the human socioeconomic system. This link may enable a unified, more complete analysis of the combined environment–economy system to be carried out.

5.1 Net Energy of Resources of Declining Accessibility

As an example of the type of a situation that could happen, consider a simple coal mine. Assume that the coal seam is vertical, starting at the surface. (This example is exaggerated to make a point, but it provides a good indication of the type of situation that will happen in the long term because coal that is closest to the surface will always be extracted first, leaving deeper seams until later.) Initially, only simple materials and resources are needed to mine the coal and deliver it to trucks or railway cars. As the coal face goes lower, the total work that has to be done to get the coal to the surface increases because of the greater height which it has to be lifted. The greater the height, the heavier and more complex the lifting mechanism has to be and the faster it has to run just to maintain a constant rate of delivery of coal at the surface. At this stage, a substantial proportion of the energy in the coal delivered to the surface will be needed just to manufacture, maintain, and run the machinery needed to get it out. This is quite separate from the effort expended in other activities, such as pumping water to keep the mine dry and the precautions necessary to ensure safety at the coal face, where the working conditions are becoming ever more inhospitable.

As the depth increases, the amount of net energy delivered to the surface declines because increasingly more of the mine's output is needed just to drive the mining machinery. To ensure even a constant supply of net energy E delivered to the economy at the surface, the rate of extraction of coal will have to increase. It is not difficult to appreciate that the mining rate G will accelerate in such circumstances. The faster the miners try to get the coal out and the greater the demands placed on the mine by the economy, the faster the process approaches the stage at which the whole operation is energetically pointless. Clearly, the end result is stalemate in that the mining operation does not actually do anything for the economy other than supply employment. Beyond the point at which the net energy yield is zero, the operation is in fact taking more than it supplies. This will happen irrespective of the amount of coal left in the mine. After a certain stage, therefore, further mining is pointless and the resource is economically and energetically useless.

It is possible to predict from dynamic energy analysis the maximum depth at which the work required to lift the coal exactly equals that available from the energy in the coal lifted. The lower the quality of the energy resource, the sooner this will happen. Peat or lignite will be accessible to shallower depths than bituminous coal or anthracite. The direct and indirect energy requirements of the other parts of the mining activity will ensure that in practice, the depth that is mineable will be very much shallower than that which is calculated from the energy requirements of lifting alone.

From the economic viewpoint, many people support the notion that improvements in technology will ensure that the economy will always get what it wants. Such a view is not supported by the physics. It is quite possible that lifting machinery and water pumps will be improved, but there is an absolute minimum energy requirement for lifting a tonne of coal to the surface and pumping a liter of water that is absolutely unalterable by technology, being a consequence of the second law of thermodynamics. The most that can be hoped for from technology is that it will enable mining to go a little deeper; the constraint remains. It should be mentioned, however, that few if any mines have reached the depth at which the net energy yield is zero; the point of this discussion is that the raw energy requirement of delivered consumer energy will increase, and at an accelerating rate, in the future, not that we will run out of the resource.

5.2 Dynamic Energy Systems Analysis

The example of the coal mine illustrates stages in the evolution of energy supply to an economy. To generalize the analysis, we must acknowledge that in reality the energy system and the economy are changing all the time. As described previously, the inevitability of declining resource quality means that accessibility will steadily decrease. For awhile, improvements in technology can be expected to compensate for this problem or even improve the overall situation. However, the second law of thermodynamics ensures that there are absolute limits to the improvements that are possible in the efficiency of conversion of primary energy into consumer energy and then into end-use energy. It is therefore necessary to examine the way in which the energy supply system and the economy interact over a long period of time in order to understand the nature of net energy effects.

Empirical data confirm that as an energy system expands into resources of decreasing accessibility, the primary energy intensity of economic activity does indeed increase, with the result that in some circumstances the total primary energy input to the economy can be expected to increase very rapidly for only modest increases in production of goods and services. In addition to the increased energy intensity of economic activity that results from such system growth are increases in the cost of substitute forms of energy. In an economy in which energy costs are increasing rapidly in real terms (as is likely in the later stages of life of many major natural gas fields, which is already occurring), it can be expected that the cost of plant to manufacture substitutes (e.g., synthetic natural gas from coal) will also increase.

From our understanding of the behavior of complex systems, we also must acknowledge that the future is indeterminate. Thus, we cannot make a realistic forecast of when a gas field will run out, let alone an estimate of the costs of substitute gas. The only reliable conclusion we can make is that current predictions appear to seriously underestimate the current value of gas and other high-quality scarce fuels and encourage wasteful use.

5.3 The Importance of Growth Rate

A great deal of energy-intensive capital investment goes into the development and expansion of energy sources, and under high-growth rates it is quite conceivable for energy-supply programs to be net consumers of energy. A very important factor here is the rate of expansion; the lower the rate, the better the net energy performance.

It appears that the rate of growth of consumer energy demand E may be a critical factor in evaluating the net energy obtainable from a system because of the strong positive-feedback energy flows F required for rapid system growth. Although net energy is not the only resource that an energy technology must supply, one can conclude that it is the most important.

It is a basic tenet of neoclassical economics that substitution is always possible, which effectively makes scarcity only relative. From biophysical systems reasoning, however, substitution of a resource by another of lower accessibility is not a situation in which the two resources can be considered separately: They exist in a systemic relationship, in which one affects the other. The increased energy requirements of access to new primary energy resources suggest that the cost of delivered energy could increase at an accelerating rate in the future and that substitution is by no means automatic. The extent to which this argument applies to the transition from fossil fuels to renewable energy sources is still an open question.

5.4 Renewable Energy

Solar inputs are essential to the functioning of any economy, not only directly but also indirectly through their part in "driving" the natural cycles that provide the ecosystem services that are essential for all forms of life, including economic. Solar sources also provide small but rapidly increasing supplies of electricity (photovoltaic and wind) and refined biomass fuels (ethanol and biodiesel) to many economies. Although environmentally more benign than the use of fossil fuels, these still have to obey the thermodynamic imperatives summarized in Figs. 5–7 via energy transformation sectors. As with fossil resources, the accessibility of solar energy is not uniform, either with latitude or with the seasons. However, the fact that annual solar inputs to Earth are many orders of magnitude greater than fossil fuel inputs means that the potential for contributions to economic activity can be expected to increase enormously in the future.

A caveat is important at this point. Investment in solar energy capture for electricity, for example, requires preinvestment from fossil-fueled economic processes. In other words, solar capture technologies (wind machines, photovoltaic cells, etc.) must be manufactured from metals, glass, plastics, ceramics,

and other products of an economy in which these materials are produced predominantly by fossil-fueled industries. The transition from present-day economies to future economies that are entirely solar powered will not be easy and can be expected to take a long time. It is not clear whether present-day patterns of consumption of goods and services can be maintained in an economy that is entirely solar based.

6. ENERGY AND ITS IMPORTANCE TO POLICIES FOR SUSTAINABILITY

In the early 21st century, the issue of sustainability of economic activity has moved to center stage of policy development in many countries. For many, the policy issues involve integration of the environmental, the social, and the economic into a policy-formation nexus. Figure 9 inverts the mainstream economic model of Fig. 4, showing from the biophysical perspective Mitchell's concentric model, in which the economy's sustainability is subordinate to that of the society, whose sustainability is in turn subordinate to that of its surrounding and sustaining environment. Fig. 9 also indicates the place of branches of thermodynamics and economics in addressing the sustainability of these parts.

The fundamental question of purpose arises. In one respect, there is no doubt that by replacing fossil fuel inputs with solar inputs to the economy there would be very substantial environmental advantages. There would also likely be substantial economic advantages (in terms of increased employment and improved health due to reduced localized pollution).

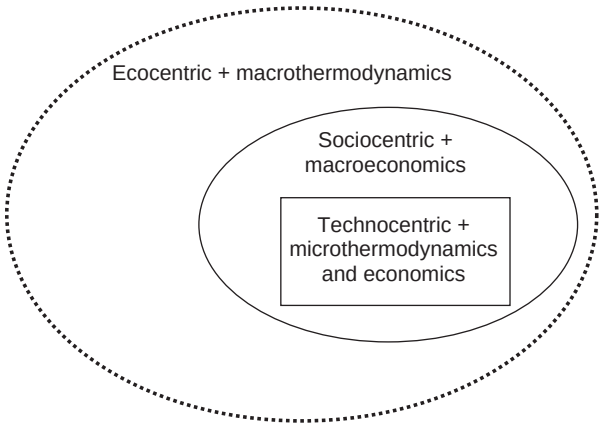


FIGURE 9 Mitchell's concentric model. From Mitchell, C. A. (2000). *Trans. Instit. Chem. Eng.* **78**, B237–242, with permission.

However, there remains the question: What is it all for? If one only envisages the replacement of fossil fuels by clean solar fuels, does that mean continuing increases in the use of private cars? Does that mean 3-ton "green" SUVs spending hours in traffic jams? To take the question further, could it mean solar photovoltaic-powered or wind-powered aluminum smelters to produce green aluminum for throwaway beer or soft drink cans?

These examples raise important questions about the relationship between energy and economics. As noted previously and demonstrated in Fig. 6, the ultimate purpose of use of energy resources is to satisfy the needs of people. In other words, the key requirement is to match means with ends.

Substituting solar for fossil fuels simply addresses the requirement of minimizing adverse impacts on the natural system while doing nothing directly to better address people's basic needs. The overall issue is much more complex than the simple substitution of one primary source of energy by another, no matter how apparently benign. The ethical positions of society are inherently relevant to our discussion and need to be applied to decisions about future economic processes.

7. AIMS OF THE BIOPHYSICAL SYSTEMS WORLDVIEW

From the previous discussion, it is clear that the biophysical systems perception has little to contribute to the determination of how people are likely to react in social situations because it is based on and applicable to issues involving the physical world, not the social. With few exceptions, most physical scientists would agree that the perception is applicable to the physical constraints and long-term limits to social activities and not to the day-to-day activities of people within those limits.

The universality of the process of consumption of available energy and its rejection as unavailable energy by physical systems makes energy one logical means of studying the structure and dynamic behavior of real physical systems. It also provides one means of assisting development of a more unified understanding of the behavior of complex physical systems.

The most important attribute of the biophysical systems approach, however, is that it is an entirely different way of thinking—a different worldview. In any real system, the ruling state is metastable, reflecting physical disequilibrium. This understanding is essential to a proper understanding of economic

activity within its physical environment. This concept gives an entirely different viewpoint on development in an economy, and of its impacts on options for future development, than is gained from the standard economic perspective. Together, they help us gain a perspective that is clearer and more powerful than those obtained from either used alone.

If the insights obtained from this perspective alone were to be incorporated into resource economics and the treatment of subjects such as depletion, then the economics of sustainable alternatives such as energy demand management (energy conservation) would be seen to be much more favorable. The approach makes it clear that there are strong indications of the existence of limits to growth of economies, in any sense involving the production and consumption of material goods and services. Although the immediate aims of the biophysical systems perception are modest, the effects of acceptance of the findings would be far-reaching, not least in the field of the use of energy resources in society.

SEE ALSO THE FOLLOWING ARTICLES

Economic Growth and Energy • Economics of Energy Demand • Economics of Energy Supply • Entropy and the Economic Process • Sustainable Development: Basic Concepts and Application to Energy • Thermodynamics and Economics, Overview

Further Reading

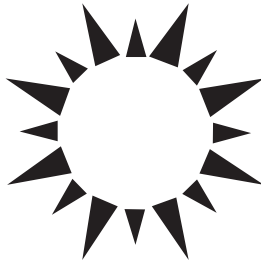
Daly, H. E. (1999). "Ecological Economics and the Ecology of Economics: Essays in Criticism." Elgar, Cheltenham, UK.

Peet, J. (1992). "Energy and the Ecological Economics of Sustainability." Island Press, Washington, DC.

Slesser, M. (1982). "Macmillan Dictionary of Energy." Macmillan, London.

Ulgiati, S. (ed.). (1998). "Proceedings of the International Workshop: Advances in Energy Studies Energy Flows in Ecology and Economy, Porto Venere, Italy, 26/30 May 1998."Museum of Science and Scientific Information, Rome.

Economic Thought, History of Energy in



PAUL CHRISTENSEN
Hofstra University
Hempstead, New York, United States

1. Introduction
2. Early Energy Ideas
3. Early Physiological Economics: Hobbes, Petty, Boisguilbert
4. François Quesnay: The Energetic Premises of Physiocracy
5. Adam Smith: The Work and Powers of Nature and Labor
6. J.-B. Say: General Theory of the Work of Nature, Labor, and Machines
7. Malthus and Ricardo: Limits on Powers of Land
8. Late Classical Theory: Babbage, Senior, and J. S. Mill
9. How Energy Came Up Missing in Neoclassical Economics

Glossary

ancient atomism Natural bodies composed of indivisible units of matter possessing size, shape, weight, and an intrinsic source of motion.
classical economics Early modern theories based on the primacy of production processes.
early idea of energy The principle of activity, force in action, the "go of things."
fire A very active and subtle substance or element used to explain motion and activity in nature and in organisms, including sensitivity in antiquity and in early modern chemistry and physiology.
neoclassical theory The modern exchange approach to economics based on scarcity and marginal utility.
phlogiston A volatile component of combustible substances in 18th century chemistry captured from sunlight by plants and released in vital processes, combustion, and reduction.
work A late-18th-century measure of the effect produced by machines and labor employing natural powers.

Economic theory and its history conveniently divide in two parts: an "early modern" production theory focused on how a society uses resources provided by nature to produce physical goods and services and a "modern" exchange theory centered on the study of the efficient allocation of scarce resources to satisfy the unlimited demands of consumers. The first, pre-classical and classical economics (1650–1850), evolved in a close and ongoing relationship with early modern scientific theories of nature's physical fecundity and organization including mechanics, matter theory, chemistry, physiology, and psychology. Mainstream neoclassical economics (1870 to the present) took its direction from the conceptual and mathematical framework of 19th-century analytical mechanics. Production is treated as an analogue of a spatial (and nondissipative) energy/output gradient, which effectively eliminated any substantive treatment of material and energy resources and physical "converters" from production theory. Neoclassical economics neglects material and energy resources and technologies and is inconsistent with their inclusion. The theory is inherently static and cannot explain physical processes, the dynamics of change in economics, or even how an economic system is coordinated.

1. INTRODUCTION

The production approach to economics that emerges in the mid-17th century took its inspiration form a broad range of sciences: Galilean mechanics, ancient ideas of active matter, Harvey's new circulation physiology, the proto-ecological natural histories developed by the scientific academies, and psychological and political theories of sentiment and organization with roots in antiquity. For nearly two centuries, economic ideas of physical production and economic behavior drew on the natural sciences,

 © 2004 Elsevier Inc. All rights reserved.

including physiology and medicine. This early connection between economics and natural philosophy has been little studied by economists but provides an essential key for understanding the evolution and structure of the classical approach.

In the second half of the 19th century, the focus of theoretical inquiry shifted from a production economics to a theory of market exchange. Neoclassical economics was constructed on two related ideas: the universal scarcity of resources and diminishing marginal productivity applied to resources and consumption goods alike. This basic idea, formulated in terms of simple partial derivatives, is that additions of any single input (or good), everything else held constant, yield a stream of positive albeit diminishing additions to output (or psychic benefit). But according to the conservation matter and energy, physical output cannot be increased if material and energy inputs are held constant. The central idea of neoclassical theory, the doctrine of marginal productivity, violates the laws of thermodynamics. Basing economic theory and its regulatory mechanisms on the conceptual structures and mathematics of analytical mechanics and field theory eliminated material and energetic processes (and much else including technology, organization, and psychology) from economic theory. Economics became an autonomous science isolated any real connection to nature or to society.

Our concern is to show how the material and energy ideas of early modern science were used to shape early economic theory and how these ideas were gradually lost. These early connections between science and economics have been hidden for the reason that they were considered heretical and were not generally openly avowed by the scientists and the economic thinkers who took them up to avoid censorship or religious condemnation. Contemporary readers could be assumed to know the coded language of their presentation. Modern investigators face the double difficulty of not knowing the language of the early sciences or the coded messages of radical materialism. They are also trained in economic theories that have severed any connection to nature and therefore lack any conception of the necessary interdependence that exists between material and energy processes in nature and in human economies. The discovery of these connections requires a sustained investigation of the shifting and forgotten labyrinths of early science. Thanks to the breadth of work in the history of science, an archaeology of the material and energetic foundations of early economics is now under way.

The next section surveys the main energy ideas employed in the sciences and economic thought during the first two centuries of early economic theory. Subsequent sections focus on the main economic writers and their very individual connections to the scientific ideas of their time starting with Thomas Hobbes and William Petty in the 17th century. The main developments move back and forth between Britain and France. Classical theory ends in Britain with J. S. Mill in the mid-19th century who moves away from grappling with the fundamental question of the meaning of energy for economic production. A final section briefly sketches the analytical focus that has governed neoclassical theory.

2. EARLY ENERGY IDEAS

The idea of energy we use is Wicken's suggestion that energy "is the go of things." We are interested in concepts that provided explanations of physical and vital motions in early modern natural philosophy and the influence of these ideas on the formation of early economic theory. Since energy comes in a material form and is not used for itself but to make or do something, we are also concerned with material transformations and physical technologies that convert energy and materials. These components of production are inextricably linked in the economic process.

Material and energy conversions are subject to physical principles: (1) matter and energy are neither created nor destroyed in ordinary processes; (2) energy is degraded in quality as it is used and becomes unavailable for use (useless energy); and (3) material processes also follow physical and chemical rules. Theories of economic production must be consistent with scientific principles including the basic complementarity that exists between materials, energy, and physical converters. If economic principles are not consistent with the sciences of nature, then economic theory must be reformulated. Economics is not an autonomous science.

The first energetic concept we isolate is the material fire, which provided the active principle of matter theory in the natural philosophies of antiquity and early modern era. Fire was a very subtle and active matter used to explain the visible fire of combustion (by a disassociation of the matter of combustible bodies by the penetration of fire atoms), the invisible flame of the vital heat in animal bodies, and sensitivity in plants and animals.

The mechanist philosophies that emerged in the wake of the victory of the new science of mechanics of Kepler and Galileo are generally assumed to have regarded matter as entirely passive. But nature required sources of motion and ideas of active matter remained central in many 17th-century philosophies. The revival of Stoic, Epicurean, and Democritean theories of matter, the close connection between a mechanical explanation of nature and ideas of active matter in antiquity, the Renaissance Aristotelianism taught in the Italian medical schools, and the growing influence of an alchemy joined to atomism all promoted ideas of active matter. Fire, the active principle of matter theory, was a basic doctrine in these theories.

Attempts to construct a unified mechanical philosophy of nature reopened the conflict between doctrines of active matter and religion. Nature, in church doctrine, was passive and God the sole source of motion in the universe. After Galileo, matter theory proved more problematic for natural philosophers than Copernicanism. Descartes canceled plans to publish *Le Monde* with its naturalistic account of the universe and moved to ground his natural philosophy on a metaphysical argument that all motion was due to God. Despite a prohibition against teaching atomism, Gassendi, a Catholic priest and astronomer protected by powerful friends, proposed a version of atomism consistent with the creation. God, according to Gassendi, could have made the world in any number of ways but chose to make it from atoms endowed with an enduring source of motion. His union of atomism with 17th-century atomistic chemistry and with Hippocratic biology provided an attractive package for the development of chemical and physiological ideas and was quickly adopted by physicians in France and by Harvey's disciples in England. His followers had to express their ideas clandestinely. Hobbes developed a entirely naturalistic account of motion but had to parry accusations of antagonists that he held a theory of self-moving matter. Newton fought materialism all his life and denied material powers to "brute matter." But by treating gravity and other active principles as powers of bodies and speculating on material ethers and vegetative powers, he fostered the materialism of the Enlightenment.

Thanks perhaps to the flux in religious regimes, England was fertile soil for energetic and atomistic ideas of matter as is evident in the work of Thomas Harriot and Walter Warner in the late 1500s and after, Bacon's atomistic speculations in the 1620s, and the Cavendish circle at Welbeck in the 1630s (which included Hobbes) who knew Warner's work. Harvey's *De generatione*, which circulated for two decades before publication in 1651, was another source of energetic ideas. His physician disciples in Oxford and London were quick to take up the "Democritean and Epicurean system" associated with Hobbes and Gassendi. The latter had a considerable influence thanks to the flexibility of his system and close connection to chemistry.

The introduction of Gassendi's atomism in England has been credited to Walter Charleton's translations of Gassendi's early works on Epicurus in the 1650s. But the young physician and future economist William Petty provides an earlier source. Petty met and became friends with Gassendi during the mid-1640s in Paris when he served as Hobbes's scribe and tutor in anatomy. The two also took a course in chemistry. We can assume he introduced the ideas of Hobbes and Gassendi to his colleagues, including Robert Boyle, in Hartlib's circle in London, where he learned the chemistry of Cornelius Drebbel; in Oxford in 1649 where he led research meetings in his lodgings, and to his cousin, Nathanial Highmore, who became a Gassendian.

Gassendi's synthesis of atomistic chemistry and Hippocratic physiology provided the basic framework that Harvey's physician disciples employed for four decades to develop the new circulation physiology. They conducted a remarkable series of experiments investigating combustion and the nutrition and respiration of plants and animals using nitre, sulfur, and other chemically active substances that brought them close to modern ideas about combustion and respiration. They also used ideas of fire as the active element of the nutritive, animate, and intellectual souls of living things, concepts Descartes had supposedly banned from biology.

The union of Harvey's circulation physiology and atomistic chemistry was quickly embraced by the scientists of the newly organized Academy of Sciences in Paris who initiated anatomical and chemical studies of plant and animals that were sustained, thanks to Royal patronage, over many decades. Early papers extended the circulation model to the preparation and movement of sap in plants. Long the neglected orphans of science, plants were now recognized as occupying the critical position in the physical and chemical circuits linking living organisms to the soil, air, and sunlight. They were the assemblers of the materials and energetic substances feeding and maintaining the other organisms. Cycles of production and consumption bound

all the parts of nature together in conditions of "mutual need." Harvey's circulation physiology extended to larger systems of circulating nutrients and materials provided a new vision of the unity and interdependence of nature and an obvious model for the human economy.

A second and closely related locus of energy ideas in 18th-century scientific and economic thought was the phlogiston chemistry formulated by the German chemist and physician, Georg Stahl, at the beginning of the century. Phlogiston, or "matter of fire," was an imponderable (unweighable) substance that joined other substances in chemical reactions (synthesis) and was released in chemical decompositions (analysis). Combustible substances such as plants, alcohol, and coal were assumed to have high concentrations of phlogiston. One of Stahl's key contributions was to emphasize the importance of the principle of conservation of matter in the methods of analysis and synthesis of compounds that became a basic tool of chemical research thereafter. Together with the "elective affinity" that different elements of matter revealed for undergoing reactions with other elements, phlogiston chemistry provided the main framework for chemical theorizing until it was supplanted by Lavoisier's oxygen theory. As an early version of potential energy, it continued to provide an energetic explanation of nature's operations into the next century.

Phlogiston chemistry was taught in the medical schools in Montpelier and Edinburgh and employed in conjunction with physiological ideas of sensitive matter. G.-F. Rouelle taught a version of phlogiston chemistry at the Royal Garden under Buffon. His students, some of whom wrote the chemistry articles in Diderot's *Encyclopédie,* rejected mechanistic chemistry in favor of ideas of energetic and sensitive matter. This vision of sensitive molecules informed the theories of psychological sensibility (sentiment) advanced by the scientists of the radical Enlightenment. These ideas informed François Quesnay's *Essay on the Animal Oeconomy* (1847) and were implicit in his doctrine of economic productivity. Adam Smith in turn praised the "philosophical chemistry" of his friends for the theoretical connections it made between the processes of the art of the furnace and the "vital flame" in animals. His associates justified their physiological and medical theories of sensitive substances by referencing the psychological theories of Hume and Smith. Phlogiston chemistry and its ideas of energetic substance were also central to the 18th-century scientists who influenced Thomas Malthus and David Ricardo.

The third basic energy idea to shape economic thinking in the classical period was the new understanding and definition of work done in harnessing and using natural powers. The British engineer, William Smeaton, made the critical contribution in a study of the efficiency of different types of water wheels in converting a given head of waterpower into work. His study, published in 1759, established the superior efficiency of overshot wheels and enabled an increase in effective water power capacity in Britain by a third in a period when water power was critical to economic growth. He defined work as the capacity to lift a given weight a given distance in a given amount of time and made comparisons of the "work done" by horses, humans, and wind, water, and steam engines. Adam Smith adopted his definition of work but unfortunately confined its consideration to human labor, which he made the center of this theory.

The new mechanical theory of work was taken up by the French engineers Charles Coloumb and Lazare Carnot. Coulomb's thrust was to assimilate machines to the work done by humans and to make machines "an economic object, a producer of work." Carnot's great study, published in 1783, was concerned with a theoretical analysis of machines as devices for transmitting motion form one body to another. Carnot was concerned with general principles irrespective of the source of power. He emphasized the notion of power over that of force and machines as agents for the transmission of power. His work received little notice until he was promoted to the top leadership in the new Republic and celebrated for his reorganization of the military as "the organizer of the victory." His contributions to the theory of machines were taken up by J.-B. Say who recognized Smith's neglect of the central role of machines employing natural powers in industrial productivity. Say's construction of a general theory of production based on materials, forces of nature, and machines to multiply labor will receive important elaborations by the mathematician and engineer Charles Babbage, the economist Nassau Senior, and the physicist William Whewell. Their contributions do not, however, receive any sustained attention by subsequent economists. As happens to many good ideas, they were passed over in silence by the leading theorists of the time who followed J. S. Mill's shift of focus from a physical theory of production to economizing, which brought the classical epoch to an end.

3. EARLY PHYSIOLOGICAL ECONOMICS: HOBBES, PETTY, BOISGUILBERT

Despite his considerable contributions to the psychological and political framework of early economics, Thomas Hobbes is considered to have made little or no contribution to the core ideas of economic theory. He is also widely seen as a mechanistic philosopher who reduced everything to Galileo's new mechanics. Most historians of philosophy and economics have not understood the central importance physiological ideas have in his system, his hierarchical conception of the sciences, and the use he makes of Harveyian physiology to extend the materials provisioning approach Aristotle initiated in his *Politics* but failed to complete. Hobbes in fact makes the first sketch the core ideas that will characterize preclassical and classical theory thereafter.

Like Gassendi, Hobbes constructs a conception of nature and society that proceeds from a basic energetic physics of matter in motion and avoids reductionism by treating each level of activity according to what he took to be the organizing principles and evidence for the topic at hand. In addition to the atomism of Democritus and Galileo, he knew and used Aristotle's biological and psychological ideas; Harvey's theories of generation, organic sensibility, and circulation; and the economic ideas of Aristotle's *Politics* and the pseudo-Aristotelian *Oeconomia*. This physics and physiology informs the production approach to economics he sets out under the heading of "The nutrition and procreation of commonwealths" in *Leviathan* in 1651. An economy, like living bodies, obtains the materials and "nutriments" it needs to produce a commodious existence from the land and sea by human industry. These material resources are passed through various stages of concoction (heating) and preparation and circulated via the conduits of trade to all parts of the body politic. His chapter also covers the institutions of property, contracts, and so on that were part of the evolutionary sociopolitical framework initiated by Democritus and his successors.

The energetic thrust of Hobbes's ideas is subtle but clear enough. He knew the atomism and active matter theories of Harriot, Warner, Bacon, and Gassendi but adopts the more dynamical approach of Galileo. He had met Galileo in 1635 and reports carrying away a vision of a world filled with motion. Every seemingly solid body was alive with the internal motions (agitations) of its smallest parts. He defines each specie of body by the imperceptible motions of its smallest particles. These motions constituted the specific capacities of each body for the initiating motions (endeavors) triggered by interactions with other bodies (the explosion of gunpowder by a spark, etc.). Hobbes's idea of endeavor as an imperceptible but actual motion is recognized as an influence on the very different ideas of force developed by Leibniz and Newton.

The basic source of motion in nature was the undulating wave-like motions of sunlight. Hobbes did not attempt any explanation of the connection between sunlight, plants, and the food consumed by animals. In his *De homine* (of man) of 1658, he attributes the motions of the heart and muscles to motions of "the invisible atoms … of a salt or niter" taken into blood via respiration. This idea, long popular among Harvey's and Hobbes's alchemically minded friends, was restated in 1654 by Ralph Bathurst, a supporter of Hobbes and associate of Petty at Oxford.

Hobbes's sketch was fleshed out by the brilliant William Petty (who may have contributed to their formulation in Paris). Petty had studied navigation, geography, physics, mathematics, and medicine in France and Holland, and he was familiar with the work of Hobbes and Gassendi and the scientific ideas of the Puritan reformers in the Hartlib's circle in London. He returned to take his medical degree at Oxford in 1650 where he organized research projects and quickly attained high positions in the university. The siren of political preferment and reform called him to Ireland. Briefly chief physician for the army, he reorganized and completed the Down Survey of Irish lands in breathtaking speed and became one of Ireland's largest landholders. Spending most of his time defending and managing his estates and seeking support for his political and economic projects, his scientific interests shifted towards the socioeconomic topics he called "political arithmetic." His Baconian-inspired works include a succession of acute but scattered theoretical insights central to classical economics, which were never drawn together in a single work.

His first contribution was the elaboration, in a *Treatise on Taxes* in 1662, of the idea of a surplus that the ancient atomists made the basis of social and economic evolution. Petty takes up the surplus in order to explain the source of taxable revenue. The surplus is the difference between the total food produced on the land and the portion that must be reserved to replace the inputs (food, fodder, seed, and equipment) used in production. It is effectively the energy available to support and feed the rest of

society, including the artisan and commercial sectors. His proposal was to tax the revenue going to the wealthiest land owners and direct it to projects increasing access to markets, land productivity, and the skills and habits of unemployed workers.

He sketches the implications of the surplus for a theory of competitive profits across sectors, which he terms "neat rents" (not the differential rents earned on more productive or favorably located soils). A "neat" or net rent is the ratio of net earnings from an activity to the cost of carrying it out. He applies this idea to agriculture and extends it to mining and casting coin. Higher returns in one sector, he suggests, lead to a flow of labor and capital goods to that sector, equalizing the rate of return between activities. In addition to this idea of an equal rate of profit (our term), he also developed a concept of differential rent for favorably located lands.

Energy ideas inform a production approach to prices in his *Political Arithmetic*. These are cost-based prices set by the food, materials, and tools needed plus profit. He reduces these in turn to labor, land, and food, given that humans and beasts must "eat, so as to live, labor and generate." The ultimate determinate of prices is thus the "energy" required, directly and indirectly in production. He distinguishes between short-run prices (reflecting day to day movements of market forces) and the "intrinsic" prices determined by production costs. This distinction becomes a basic feature in classical economics via Richard Cantillon and Adam Smith.

The next significant development of a physiological economics was made by Pierre Boisguilbert, the Lieutenant-General of Normandy (active 1695–1714). Boisguilbert was obviously aware of Hobbes's *Leviathan* and Petty's theory of the surplus, which he made the basis of the emergence of the professions in modern society. A deep study of agriculture and its commerce inform his writings. He was also a close student of the new sciences of nature of the Academy of Sciences in Paris and the work of his younger cousin, Bernard Fontenelle, who was establishing himself as a leading interpreter of the new science. Neither he nor his more radical cousin were Cartesians. Boisguilbert's commitments were to the Stoicism of his uncle (Pierre Corneille) and the Epicureanism of Gassendi and François Bernier. The latter was well known for his travels in the East and critique of Oriental Despotism—the aggrandizement of resources by the few and the absence of property rights and basic civil protections, which the Epicureans regarded as critical incentives for social intercourse, production, and prosperity. Bernier implies a critique of France's aristocratic economy, which Boisguilbert makes explicit.

The work of the Royal Academy of Science provided a new conception of nature linked together in a circuit of chemical exchanges. Plants were the prime producers of the energetic substances maintaining all organisms in "a state of reciprocal and mutual need." Boisguilbert develops similar ideas in economics. Agriculture produces the "basic" goods (grain, wine, and plant and animal tissues) that, like the chemical elements, produce the rest. The fecundity of nature supplies the means that give birth to the 200 or so professions of the state.

The academy's model of the circulation of sap in plants provided Boisguilbert (as Harvey's theory did for Hobbes) a model for economic production and circulation. The plant model makes a clear separation of production sites. Primary materials assembled and processed at one site are perfected at others and move in distinct "circuits." Boisguilbert likewise sees goods produced on the land and in artisan shops moving along distinct circuits between farms and towns. He extends a distinction between the sites of agricultural and manufacturing production to recognize and explain the very different price and quantity behavior between the two sectors.

He recognizes that the strict requirements that nutritional needs impose for maintaining the production and flow of food to consumers requires in turn the reciprocal needs of artisans for markets for their goods and services. Any break in this mutual circulation threatens a widening disequilibrium endangering the health of the whole society. Given the ferocious pursuit of self-interest by traders, equilibrium is maintained on the edge of a sword where a strong movement of supplies or fall in demand can become self-feeding. Such instability is magnified by the sectoral differences in conditions of production and market behavior. The price instability of agricultural markets has disastrous effects for maintaining productive investments and fertility of the soil. In urban industries, workers resist cuts in pay and employers respond to falling demand by laying off workers. The result exacerbates the crisis. His acute analysis captures an asymmetry in market behavior that continues to perplex modern equilibrium theorists.

Prices obviously play an important role in maintaining economic equilibrium but the analysis goes beyond the sphere of exchange to assume that nature's conserving operations extend to the economy. Nature has the capacity to sense injury to the economy and the power to regulate and to heal

itself. His ideas reflect the Stoic vision of natural order and harmony that Gassendi and Bernier extend from animate nature to the universe and the Hippocratic medicine advocated by Gassendian physicians. Good policy can maintain health but when illnesses arrive they must follow a natural course. Nature not the physician heals, thus, "let nature alone." These properties of self-regulation are the laws, God's invisible hand, that maintain order in the world.

Boisguilbert's psychological and social ideas also reflect Gassendi's reinterpretation of Epicurean psychological and ethics. Living organisms are designed with natural capacities for sensing what is good for them and what to avoid. Pleasure and pain are part of God's design to direct the activity of nature's creatures to life, health, and reproduction. This naturalist psychology extends to the direction and regulation of human society where the desire for the goods of the world, property, and security induce individuals to establish collective agreements to respect and protect the property and rights of others. The ruler is bound in turn to protect property and individual rights and to advance the common good, which includes protecting the weak from the depredations of the strong and wicked. Boisguilbert's use of this clandestine current of ideas links him to a lineage of economic thinkers going back to Hobbes and forward to Smith.

4. FRANÇOIS QUESNAY: THE ENERGETIC PREMISES OF PHYSIOCRACY

Before Quesnay became an economist, he was a surgeon and physician whose *Physical Essay on the Animal Economy* of 1747 was dedicated to establishing the physical and chemical foundation of the nutrition, generation, and sensitivity of living organisms. For Quesnay, this was the ancient physics revived by Gassendi, the energetic chemistry taught by Rouelle, and an Epicurean psychology extended by the physiological and psychological theories of sentiment of the Montpelier physicians—the chemical and psychological doctrines of Diderot's *Encyclopedia*. Quesnay presents this radical energetic materialism in the form of a critique of ancient ideas, a skeptical strategy designed to avoid censorship (which has also misled historians of economics).

Thanks to his skills as a surgeon and systematic thinker, Quesnay rose from lowly origins to become physician to Madam de Pompadour at the Court of Versailles. The grant of a rural estate by the king turned his attention to the problems of French agriculture. His articles in Diderot's *Encylopedia* on farming in the 1650s and his famous *Economic Table* at the end of the decade include his infamous economic doctrine of the exclusive productivity of agriculture and corresponding sterility of commerce and manufacturing. Developed within the context of an agrarian economy, his argument was that only agriculture is regenerative in the sense of replacing what is consumed. Artisan industry (assuming all inputs come from the land) only transforms and consumes things that have been produced by the farmers. These ideas can be seen to depend on the energetic theory of his physiology that plants are the only producers in nature and the rest of nature (and humans) only exist by consuming what plants have made.

According to the 1747 *Essay*, the cycle of nature's production begins "in the womb of the earth" where elements from the atmosphere and soil are chemically combined in simple compounds and taken up by the plants where they are joined to the "very subtle and active matter" of fire provided by the sun. These organic compounds provide the active energetic substance of fire circulated throughout nature by the animals that eat the plants, the animals that eat animals, and so on. The materials are returned to the earth (recycled), but the subtle energetic matter is lost in the decaying compounds and has to be replaced by a new cycle of primary production.

Two basic criticisms have been leveled at Quesnay. The first is that he mistakenly assumed that only agriculture was capable of producing a surplus (and profit) and that he neglected the obvious importance of manufacturing and growing industrial productivity to human welfare. The criticism is valid. The second criticism is that he failed to apply the conservation of matter to agriculture in the same way he does in artisan manufacturing. Quesnay, however, would agree that production in both sectors obeys the conservation of matter. No output, let alone a surplus of grain, can be produced without the appropriate flow of material inputs from the soil and air. His point is rather that only plants can produce the vital energetic matter. His argument neglects the energy resources and inorganic materials used in industry that do not come from agricultural land. But within the assumption of a renewable resource economy limited to the use of plant biomass for materials and energy, his theory of the exclusive productivity of agriculture holds.

5. ADAM SMITH: THE WORK AND POWERS OF NATURE AND LABOR

Adam Smith is popularly regarded as the founder of classical economics, a judgment that owes more to the accessibility of his vision of a self-regulating market economy operating within an institutional framework offering something for everyone than to any fundamental contribution to theory. His laissez-faire vision excites some, his production orientation others. The heart of his production theory is the proposition that increases in the division of labor are the prime driver of productivity growth and the wealth of nations. He employs the new engineering treatment of work and natural powers to set out his theory. What is astounding, however, is the absence of any mention of the work done by natural powers other than labor (and land). This lacunae will have far reaching consequences for the treatment of energy and technology by later writers.

Smith wrote impressive histories of astronomy and ancient physics and reviews of Buffon, Linnaeus, Diderot, Rousseau, and others designed to introduce the ideas of the Enlightenment to his countrymen. He knew Gassendi's work and was good friends with Quesnay. His intimate circle included the philosopher David Hume; the chemist-physicians William Cullen, Joseph Black, and James Hutton (famous as a geologist); and the physicist John Playfair. Smith's early writings and *Theory of Moral Sentiments*, published in 1756, suggest theories of nature's design and operations, organic sensibility, and social evolution at the center of Enlightenment thought.

Smith's references to the system of the world as a great self-regulating machine have been taken to indicate an adherence to Newtonian orthodoxy. This inference is problematic. Smith, as did Hume, placed himself under Newton's mantle, but his ideas of nature's design and principles of operation are at variance with the theistic, antimaterialistic, and neoPlatonic metaphysics of immaterial principles connected with Newtonian "orthodoxy." In his "History of Astronomy," Smith interpreted gravity and inertia as balancing "material principles" that ensure a self-regulating universe. In his "Essay on Ancient Physics," he favorably compares the Stoic system of the world which used the idea of fire as the vital principle forming the world and living things to unify the design and the operation of the world, to the theism of Plato (and Newton), which artificially separated the principles of intelligence that formed and animated the world from those carrying out the ongoing operations of nature.

He also praises the "philosophical chemistry" of his associates for the "connecting principles" it provided for explaining the operation of fire in furnaces and the "invisible fire" or "vital principle" of plants and animals. He noted with regret how little this science was known outside the chemists who practice the "art of the furnace." Indeed, his idea of fire and what he calls the Stoical system was straight out of Bernier, which makes Gassendi a central source for the vision of nature's capacities of operation and self-regulation he both lauds and applies in his own work. Like Hutton and Playfair, who saw an eternal and self-regulating universe operating by natural laws, Smith appears closer to the deism of the Enlightenment than to Newton's theism.

The chemical ideas of fire and phlogiston continue to inform Smith's idea of the productivity of land in the *Wealth of Nations*. In this regard, he remains close to Quesnay. Agriculture is the source of the energetic matter feeding the human and animal populations. But he employs the new terminology of work and natural powers to displace Quesnay's view that "land is the unique source of wealth" and shift the emphasis of production theory to labor and the factors contributing to labor productivity as the source of wealth. The productivity differences characterizing nations and regions are not related to differences in nature but to the effects of the division of labor, which augments the efficiency of workers. Labor (work) is the primary source of growing wealth, which he contrasts to the "niggardliness of nature."

Smeaton's reports on the conversion of mechanical power by hydraulic and animate engines and definition of work as a universal measure of mechanical effect and value are the obvious source for Smith's ideas. He likely knew these ideas directly (the volumes are in his library) and discussed them with his friend James Watt in the 1760s when they were colleagues at the University of Glasgow. We know that Watt discussed his innovation of a separate condenser for the steam engine with Smith. Smeaton provides the engineer's concern with efficiency in the use of a given quantity of work capacity, the idea of work as a universal measure of effect and value in all times and places, and the notion of work as equivalent to effort or pain exerted over a defined period. All these ideas are prominent in Smith.

While this framework is pregnant with possibilities for a new foundation for production, Smith makes no mention of the work of machines in harnessing nature's resources and powers in his

economics. The new ideas of work are confined to labor and the related ideas of natural powers are confined to land and human labor (and working animals). The efficiency concerns of Smeaton's analysis picked up, but there is no discussion of the work done in the economy by natural powers of wind, falling water, or the immense sources of heat from coal, which had replaced wood in the industrial processes powering Britain's industrial growth.

Smith notes in passing the great investments made in "furnaces, forges, and slit mills." He would have known the great Carron ironworks near Glasgow—the first integrated iron works in Britain to employ coke in all its operations, whose blowing engines for the blast furnaces and water wheels and steam pumping engines supplying the mills with a continuous source of motive power were designed by Smeaton. He should also have known that nails and pins were made by power-driven machinery in rolling, slitting, and wire-drawing mills. Yet he illustrates the principle of the division of labor using the example of an artisan pin factory unassisted by machinery.

Modern economists tend to see Smith's division of labor and increases in the efficiency of labor in combination with increasing market size as an ever enlarging spiral of growth. It could not have been in Smith's theory since the original source of natural "powers" converted by labor in his theory continued to be, as it was in Quesnay, the energy fixed by plants. In Book II of the *Wealth of Nations*, he declares agriculture to be more productive than manufacturing because it employs two powers (nature and labor) in contrast to manufacturing which only employs one (labor). But the energy supplied by plants must be divided between workers in agriculture, manufacturing, transport, and a multitude of other uses and remains the fundamental limit on the economy. Hence his emphasis on the division of labor and productive employment as opposed to unproductive servants. His lack of clarity about the role of energy in agriculture and his neglect of industrial sources of power are major reasons for the neglect of energy in economics.

6. J.-B. SAY: GENERAL THEORY OF THE WORK OF NATURE, LABOR, AND MACHINES

The first general application of the new ideas of work in production theory was made by the French economist Jean-Baptiste Say in 1803. Despite strong literary interests, his father sent him to a commercial school in England and then apprenticed him to a powerful French banker. Say left business to work for a newspaper run by Mirabeau (Quesnay's close associate), which led to his appointment as an editor for an important revolutionary journal that included a coverage of science. Aided by his brother, an engineer, he attempted to acquaint himself with "all of science." He knew Lavoisier's chemistry and the contributions of Lazare Carnot and Charles Coloumb to the new science of work. Reading Smith he was immediately critical of Smith's emphasis on labor and neglect of nature and machines. The greatest development of productive forces, he argued, came not from the division of labor but from machinery using the forces of nature. Smith did not understand the true theory of machines in producing wealth.

Say was even more critical of the physiocrats who failed to see that all production, whether it was the production of wheat in a field or the making of a watch in a workshop, did not create new matter but only rearranged or assembled existing matter. All production was subject to the conservation of matter and required the forces of nature. Quesnay, we have seen, would not have disagreed but was making an argument that only agriculture renewed the energetic properties of materials required for life. Say's contribution was to see that human production was exploiting other sources of power by inventing new machines and that production theory must be concerned with the fundamental contributions of materials and forces of nature in human production. His ideas were clearly shaped by Lavoisier's emphasis on the importance of the conservation of matter in chemical analysis and the energetic ideas employed by French engineers and scientists. He knew the energy mechanics of Lazare Carnot (a revolutionary hero) whose work was reviewed in Say's journal, Charles Coulomb's comparisons of the work of machines, animals, and labor, and most likely Lavoisier's ideas of energy conversions by animals and machines.

Say defined the work of humans and the work of nature as "the continuous action which is exerted for executing an operation of industry." He argues that all material production whether in the field or the factory requires the same basic categories of elements (i.e., materials, forces of nature, and actions of labor, and machines). He compares land to a machine but knows that labor and machines are powerless to produce on their own without materials and a source

of power. He understands the importance of production in economics and insists that economic theory must begin with production before taking up the topics of value, distribution, and consumption.

Say provided the first general physical theory of production based on the new theory of work and mechanical energy. Unfortunately, his achievement was obscured by a diffuse presentation (especially in the first edition of the *Treatise*) and his failure to develop the implications of his new approach for a production-based theory of prices. He incorrectly argues that production creates nothing new, only utilities. While production does not create new matter, a watch is more than the gross weight of its parts. It is a new physical entity, and its conditions of production as well as its utility are important for how it is priced or valued. Say's attempt to assimilate his theory to the French scarcity and utility approach to value ignored the interdependence between production and value and contributed to the misunderstanding of his important contribution to classical production theory by the English classicals who failed to take up his production approach and its specific concerns with materials, energy, and machines.

7. MALTHUS AND RICARDO: LIMITS ON POWERS OF LAND

Like Smith, Robert Malthus has been seen as a disciple of Newton in his scientific, economic, and theological ideas. Malthus took his degree in mathematics at Cambridge, which was basically the rational mechanics of Newton and he made good use of his training in formulating regulatory mechanisms for population and price theory. But the natural philosophy and theology he outlined in the first *Essay on Population* in 1798 is neither orthodox nor Newtonian. His ideas of an ongoing creation, the evolution of mind from "a speck of matter," and the forging of mental and moral qualities by the struggle to overcome evil have their sources in the associationist psychology and theology of David Hartley; the scientific and theological ideas of Joseph Priestley, Hartley's leading disciple; and the evolutionary physiology of the physician and biologist, Erasmus Darwin, grandfather of Charles Darwin.

Out of this heady brew, Malthus constructs a decidedly heterodox philosophy and theology which sees nature, mind, and human progress as products of a long struggle against adversity. He makes an important break with ideas of natural harmony by drawing out the implications of Priestley's theological vision of God's imperfect but evolving creation and recognition of limited resources (even God faced limits in what he could do). Given the inherent drive to reproduce and the scarcity of resources, the capacities of the population of each species to outrun its food supply must eventually produce a struggle for existence in nature. Malthus recasts Priestley's progressive theology of struggle and growth into a theodicy of scarcity refracted through the mathematics of exponential growth and the focusing lens of regulatory mechanics. His ideas also reflect the absolute limits of an energetic perspective. Priestley was a phlogiston chemist and co-discoverer of photosynthesis. In a period when the nutrient limits of soils were not understood, the primary limit on agricultural production appeared to be the ability of plants to fix energetic matter as opposed to a shortage of particular (and augmentable) nutrients.

Malthus's preoccupation with the availability of food was reinforced by a theological vision of nature's design. Struggle against limits was necessary for progress. A preoccupation with food and limits, reflecting energy theory and theology, helps explain why Malthus's contributions to areas of production theory outside agriculture were so limited. Despite his acute observations on the central differences between returns and costs in industry (where techniques are replicated and constantly improved) compared to agriculture and an understanding of the energetic basis of production, he does not develop a general theory of the material and energetic transformations specific to industrial production. He certainly knows Say's production theory. It would appear that the idea of constraints and limits, which were rooted in theology and energy theory, carry implications for limits on human progress that explain his reluctance to engage Say's more general theoretical framework with its quite different vision of human progress.

The scientific underpinning of David Ricardo's powerful theorizing has received little attention from historians. What is known of his scientific interests and cast of his thinking suggests that natural philosophy played an important methodological and conceptual role in his economics. He had early interests in chemistry and geology. A convert to Unitarianism, we can expect he was familiar with the philosophical and chemical ideas of Priestley, the intellectual leader of the movement. The fact that Ricardo was a trustee of the London Geological Society, whose directors sought to advance James Hutton's uniformitarian geology, suggests he knew Hutton's powerful energetic arguments about the

geological and biological processes forming the earth and his method of dynamic (energy) equilibrium (and his tendency to ignore historical processes). Hutton was one of the leading energy thinkers of his generation and the likely source for Ricardo's similarly powerful focus on dynamic processes.

Energy ideas play a central role in Ricardo's early corn theory of profit, which is set out as a ratio of the surplus of corn to its costs of production (measured in corn). If land (and energy) increase relative to labor, profits rise. Corn continues to play a central role in the dynamics of his labor theory of production and value after Malthus convinced him that capital could not be limited to corn. The relative availability and cost of corn (the energy source running the organic machines of the agricultural and artisan economy) continued to provide the foundation of his theory of relative costs, profits, and rents, and thus the long-run dynamics of the economy.

Although Ricardo noted the merit of Say's division of economic theory into production, distribution, and value, he did not make the explicit treatment of the production principles that Say's plan suggested. Instead, he begins his *Principles* with a chapter on how the values of commodities are determined. His analysis is clearly informed by a vision of production but the physical perspective of contemporary scientific ideas that informs it must be reconstructed from the wording Ricardo employs, a footnote or two, and the tenor of his arguments. We can catch a glimpse of his acute understanding of the contribution of the natural powers employed in industry when he refutes Smith's notion that nature does nothing in manufactures. The passage reflects Say's language but is pure Hutton in its energetic focus:

> *Are the powers of wind and water which move machinery, and assist navigation, nothing? The pressure of the atmosphere and elasticity of steam which move the most stupendous engines—are they not the gifts of nature? to say nothing of the effects of the matter of heat in softening and melting metals, of the decomposition of the atmosphere in the process of dyeing and fermentation. There is not a manufacture … in which nature does not give her assistance to man, and give it too, generously and gratuitously.*
>
> *—Ricardo, Principles of Political Economy, [1817], chapter 2, n10*

8. LATE CLASSICAL THEORY: BABBAGE, SENIOR, AND J. S. MILL

Until the mid-1830s, the energy ideas entertained by classical economists, apart from Say and Ricardo, were effectively confined to plants and biological engines. Thanks to Babbage's pioneering study of the new industrial "Economy of Machinery and Manufactures" in 1832 and Andrew Ure's study of the factory system in 1835, the importance of coal and steam engines for industrial production and transportation and for Britain's industrial ascendancy were now in the public eye. In reviews of Babbage and Ure, the economist John McCulloch noted the link between Britain's industrial prosperity and the exploitation of coal technologies. He argued that the increase in the quantity of iron made with coal lowered the costs of machinery (and steam engines) leading to more output at lower unit costs and more demand for coal, iron, and machines. He did not, however, attempt to incorporate these ideas into a theory of production.

Babbage's great study of principles of a machine economy distinguished two types of machines, those which produced power and those designed to transmit force and execute work. The first class included water and wind mills. But it was the adoption and continued innovation in steam engines employing coal for fuel that captured Babbage's imagination (the heat-based industries need equal attention as well). Economies of scale and high fixed costs pushed mill owners to continuous operation and opened opportunities for new professions (accountants, etc.) and industries, such as gas lighting. New power sources and technologies created new products and materials and new uses for materials that formerly had little value. The principles of the machine (and coal and iron) economy extended beyond the factory to the larger political economy and society generating new knowledge and new models for administration.

The only economist to apply Babbage's theory of machines to a theory of production inputs and returns was Nassau Senior, whose contributions in this regard, however limited, were completely ignored by his contemporaries. Babbage's suggestion that "practical knowledge" would continue to grow without limit appeared in the third of Senior's "elementary propositions of economic science," which stated that "the powers of labor and the other instruments that produce wealth can be increased indefinitely." Senior's fourth proposition posited the existence of diminishing returns to additional labor employed on the land and increasing returns in manufacturing. While he understands the conservation of matter in manufacturing industry when he states that "no additional labour or machinery can work up a pound of raw cotton into more than a pound of manufac-

tured cotton," his attribution of diminishing returns in agricultural production to the employment of additional labor "on the *same* material" does not make it clear that additional output here requires additional materials from the soil and atmosphere and additional solar energy to produce another bushel of corn or pound of cotton seed.

Senior used Babbage to advance Adam Smith's distinction between fixed and circulating capital (structures and agents versus material flows). He distinguished tools and instruments from materials embodied in production and subdivided machines into those producing and transferring power. He separated natural powers from ordinary materials on the grounds the former are not embodied in output but did not know where to put food, coal, and other natural powers. Instead of distinguishing a subcategory of natural powers or energetic flows along side materials, he lumped food and coal with "fixed capital." Clearly, more insight about the nature of power and the agents and machines applying and using power was needed. His laudable efforts to grapple with a physical classification of production inputs and their implications for production returns were not, unfortunately, taken up by subsequent economic writers. Indeed, J. S. Mill attempts to undermine both Senior's classification efforts and treatment of production returns.

Another extremely important treatment of the ideas of the sources and application of power in human economy was that of the scientist and mathematician William Whewell in 1841, who extended Babbage's classification of machinery to include moving power, trains of mechanism connecting power to work and work done. Whewell also recognized that the engine economy and its sources of power were dissipative and would eventually be exhausted (including the engine of the sun). He concluded, Norton Wise points out, that no balance existed in nature (or economy). In Whewell, the steam economy replaced the equilibrium of the balance with disequilibrium and decay, foreshadowing the second law of thermodynamics. Whewell embraced decay as part of a Christian escatology but would have not been happy that evolutionary processes are the other side of a dissipative and temporal universe.

Despite his friendship with Babbage and adoption of J.-B. Say's ordering of economic theory (Book I of his *Principles* is devoted to production), the socially progressive J. S. Mill made no attempt to apply physical or engineering insights to production theory. He follows Say (and Senior) in classifying production inputs under the headings of labor, natural agents, and capital, but his concern throughout is to emphasize economic aspects of production. He shows no interest in the physical foundations of economic theory. Book I has three chapters on capital and at least two on labor but none on the characteristics of land or natural agents. He isolated "labor and natural objects" as the two "requisites of production." But in contrast to his elaborate classification of indirect labor, and the separation he makes between materials and powers in nature, he refuses to make any distinction between materials, food, and fuels as circulating capital.

Ignoring Babbage, Senior, and Whewell, Mill argued that "political economists generally include all things used as *immediate* means of production either in the class of implements or materials to avoid a multiplication of classes [and] distinctions of no scientific importance." He made the distinction between materials and machines turn on whether an "instrument of production can only be used once" and declared that a "fleece is destroyed as a fleece" in making a garment just as a fuel is consumed to make heat (which indicates he did not understand the distinction between energy and materials). He responded to the "able reviewer" of the first edition of his work (Senior) who objected to subsuming coals for engines and food for workers under the class of materials by admitting that while his terminology was not in accord with the physical or scientific meaning of material, the distinction between materials and fuels is "almost irrelevant to political economy."

Mill correctly recognized that tools and machinery lack productive powers of their own. But after declaring that the only productive powers are labor and natural agents he denied this status to "food and other materials" (and thus fuels). The fact that steam engines burning fuels are productive and a worker lacking food is not exposes the weakness of his position. By undermining Babbage and Senior's extension of Smith's distinction between physical structures that process and store materials and flows of materials and energetic powers, Mill brought the development of engineering ideas about the machines and powers in production theory to a halt.

Mill recognized the existence of increasing returns in manufacturing and argued that the causes increasing "the productiveness of [manufacturing] industry preponderate greatly over the one cause [the increasing expense of materials] which tends to diminish that productivity." Yet he treated increasing returns as a subsidiary proposition to "the general

law of agricultural industry," that is, diminishing returns, which he calls "the most important proposition in political economy." Increasing returns is not given an independent development. The statement of neither proposition includes a consideration of the essential inputs required to increase output according to the conservation laws which tie output to material and energy input. Mill was aware of the complementarity between capital and natural powers and materials and output but his lack of concern with accurate scientific terminology and any systematic consideration of the physical principles of production fail to preserve or advance the physical production framework advanced by Say, Ricardo, and Senior let alone Babbage or Whewell. Given the that Mill's work was a common point of departure for neoclassical theorists, his failure was a critical factor in the neoclassical failure to seriously grapple with energy and scientific approach to production theory.

9. HOW ENERGY CAME UP MISSING IN NEOCLASSICAL ECONOMICS

The main early neoclassical economists, including William Stanley Jevons, Leon Walras, Alfred Marshall, and Wilfredo Pareto, took their inspiration from the rational and analytical mechanics of the 18th century and 19th century. Thanks to the construction of mechanics in terms of the equilibrium between forces, mechanics provided economics with a powerful mathematical language of analysis that included a model of self-regulating activity produced by the inexorable operation of opposing forces. Neoclassical economics abandoned the production approach of the classicals to reconstruct economics as a mathematical science of behavior and markets expressing the interplay of demand and supply based on theories of marginal utility, marginal productivity, and universal scarcity, the latter being the consequence of the substitution assumption or "law of variable proportions," which assumes that any nonscarce good of interest to economics will become scarce by being substituted for other goods.

We have already noted how the standard formulation of diminishing marginal productivity violates the basic laws of thermodynamics. The adoption of the idea of marginalism necessarily pushed materials and energy and any scientific formulation of production from the scene. William S. Jevons eliminated fixed capital from his theory and provided no treatment of the role of energy in his treatment of production despite having written and entire book on the critical importance of coal to British industrialization just a few years before. Leon Walras had initially specified raw materials as an intermediate input but then proceeded to vertically aggregate production so that output was a direct function of the "original factors of production," land, labor, and capital. Capital was assumed to be like land so that adding more labor to a given amount of capital could be assumed to add more output albeit subject to diminishing returns.

But the theoretical problems of marginal productivity are not confined to a special assumption (individual inputs are productive everything else constant) that can be relaxed by adding more inputs. The assumption reflects a structural feature of the theory, which effectively prevents any realistic physical reformulation of the theory. Thanks to the adoption of the mathematics of field-theory as the structural framework informing utility and production theory, every input in the production function (which is based on the potential function of analytical mechanics) is effectively assumed to have productive capacities because of the existence of an energy gradient in the space over which the production function is defined. Each production input is effectively considered to be an axis in n-dimensional space. Thus, an increase in any input (an increase along one axis) automatically involves a movement to a higher energy gradient. Output, defined as the potential function of the field, increases. These movements, moreover, are always reversible since the field theoretical framework is nondissipative and conservative.

In the 18th and early 19th centuries, nature (economics included) had been assumed to be self-regulating as a metaphysical principle grounded in nature's design. In the late 19th century, despite the development of a new energy physics based on the laws of thermodynamics and an understanding of the inexorable processes of entropic dissipation, economics had found a new and powerful mathematical and conceptual framework in field theory for expressing and developing its vision of order, stability, harmony, and rational allocation. For economics, the virtue of a theory of electromagnetic or gravitational fields is that there is no dissipation. Unlike other energy conversions, which are dissipative, the energy principles of field theory are conservative (no dissipation takes place). This provides a framework taken from physics that

isolates economics from the troubling questions about natural resources or sustainability. These issues have little meaning once the assumptions of the theory have been accepted.

SEE ALSO THE FOLLOWING ARTICLES

Conservation Measures for Energy, History of • Conservation of Energy Concept, History of • Early Industrial World, Energy Flow in • Economics of Energy Demand • Economics of Energy Supply • Energy in the History and Philosophy of Science • Geographic Thought, History of Energy in • Prices of Energy, History of • Sociopolitical Collapse, Energy and • Thermodynamic Sciences, History of

Further Reading

Aspromourgos, T. (1986). Political economy and the social division of labour: The economics of Sir William Petty. *Scottish J. Pol. Econ.* **33,** 28–45.

Ayres, R. U. (1978). "Resources, Environment, and Economics: Applications of the Materials/Energy Balance Principle." Wiley, New York.

Christensen, P. (1989). Hobbes and the Physiological Origins of Economic Science. *History Pol. Econ.* **21,** 689–709.

Christensen, P. (1994). Fire, motion, and productivity: The proto-energetics of nature and economy in François Quesnay. *In* "Natural Images in Economic Thought, 'Markets Read in Tooth and Claw'" (P. Mirowski, Ed.), pp. 249–288. Cambridge University Press, Cambridge, UK.

Christensen, P. (2003). Epicurean and stoic sources for Boisguilbert's physiological and Hippocratic Vision of nature and economics. *In* "Oeconomies in the Age of Newton" (M. Schabas and N. De Marchi, Eds.), pp. 102–129. Duke University Press, Durham, NC.

Debeir, J.-C., Deleage, J.-P., and Hemery, D. (1991). "In the Servitude of Power: Energy and Civilization through the Ages" (John Barzman, trans.). Zed Books, Atlantic Highlands, NJ; London, UK.

Dixon, B. L. (1988). "Diderot, Philosopher of Energy: The Development of his Concept of Physical Energy, 1745–1769." Voltaire Foundation, Oxford, UK.

Frank, R. G. (1980). "Harvey and the Oxford Physiologists." University of California, Berkeley, CA.

Garber, D., Henry, J., Joy, L., and Gabbey, A. (1998). New Doctrines of Body and its Powers. *In* "The Cambridge History of Seventeenth Century Philosophy, I," pp. 553–623. Cambridge University Press, Cambridge, UK.

Golinski, J. (2003). Chemistry. *In* "The Cambridge History of Science" vol. 4, "Science in the Eighteenth Century" (Roy Porter, Ed.), pp. 375–396. Cambridge University Press, Cambridge, UK.

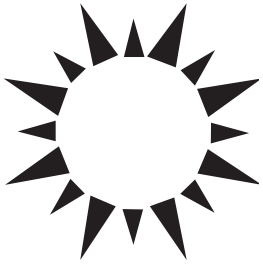
Vaggi, G., and Groenewegen, P. (2003). "A Concise History of Economic thought: From Mercantilism to Monetarism." Palgrave, New York.

Vatin, F. (1993). "Le Travail: Economie et Physique: 1780–1830." Presses Universitaires de France, Paris.

Wicken, J. S. (1987). "Evolution, Thermodynamics, and Information: Extending the Darwinian Program." Oxford University Press, New York and Oxford, UK.

Wise, N., and Smith, A. C. (1989). Work and waste: Political economy and natural philosophy in nineteenth century Britain. *History Sci.* **27,** 263–301, 391–449.

Ecosystem Health: Energy Indicators



DANIEL E. CAMPBELL and TINGTING CAI
United States Environmental Protection Agency
Narragansett, Rhode Island, United States
THOMAS W. OLSEN
Consultant
Kingston, Rhode Island, United States

1. **Ecosystem Health and Ecological Integrity**
2. **Historical Background on Ecosystem Health**
3. **Energy Systems Analysis, Health and Emergy**
4. **Energy and Ecosystems**
5. **Direct Measures of Ecosystem Health**
6. **Indirect Measures of Ecosystem Health**

Glossary

ascendency A summary measure of the magnitude of the fluxes within a network and of the degree to which the paths of flow within the network are organized or preferentially directed.

available energy The potential to do work.

ecological integrity The state of structural completeness of components and maximum function of processes in an ecosystem operating under a given set of external forcing energies. In this state, the empower of the ecosystem network is near the maximum possible.

ecosystem An organized system of biotic and abiotic components including land, water, mineral cycles, living organisms, and their behavioral control mechanisms.

ecosystem health The functional integrity of an ecosystem as characterized by a state of vigorous self-renewal in the ecosystem as a whole and in its components (soil, water, animals, plants, etc.). One measure is the empower gained by the ecosystem from a given set of external energy sources.

emergy The available energy of one kind previously used up both directly and indirectly in making a product or service. Emergy has units of emjoules to connote energy used in the past.

empower Emergy per unit time or emergy flux.

energy flow network The interconnected energy flows in an ecosystem.

energy (emergy) signature The suite of external forcing energies from which an ecosystem can capture available energy. When the external energies are converted to emergy, the entire suite is called the emergy signature of the ecosystem.

enthalpy (heat content) The internal energy plus the product of the pressure times the volume. Relative changes in enthalpy are easily measured by calorimetric methods, which have been widely used to determine the heat content of biomass.

exergy The potential that energy has to do work relative to a reference state. As used by some ecologists, exergy includes the higher potential that energy associated with lower-probability structures has for the capture of additional available energy.

homeorhetic The tendency for a system to maintain constancy in its dynamic properties (e.g., energy flow).

maximum power (empower) principle The principle originally formulated by Alfred J. Lotka as "natural selection tends to make the [available] energy flux through the system a maximum, so far as compatible with the constraints to which the system is subject." H. T. Odum realized that available energies of different kinds have different abilities to do work when used within an ecosystem network. Conversion to emergy flux (empower) expresses the energy fluxes of many kinds in units with the same quality (e.g., solar joules), allowing the maximum to be accurately determined.

pulsing paradigm The theory that alternating production/storage with consumption/release on all hierarchical scales maximizes power.

self-organization The use of net energy to build increasingly complex patterns of feedback and design to capture more available energy for the ecosystem; the properties of this process as they govern the creation of order in far-from-equilibrium systems.

solar transformity The solar emergy required to make a unit of energy of a product or service. Its standard units are solar emjoules per joule.

specific emergy The emergy per unit mass of a flow, product, or service, its standard units being solar emjoules per gram.

Just as for human beings health is a concept that applies to the condition of the whole organism, the health of an ecosystem refers to the condition of the ecosystem as a whole. For this reason the study and characterization of ecosystems is fundamental to establishing accurate measures of ecosystem health. By definition, the idea of health is closely related to the optimal functioning of an organism or other entity. This idea can be applied to ecosystems, economies, and other organized systems in general and not simply to living organisms. Health is more than the simple absence of disease; rather it is a state that reflects the overall well-being of the system, whether it is a human organism, an ecosystem, or the economy of a nation.

1. ECOSYSTEM HEALTH AND ECOLOGICAL INTEGRITY

Integrity is a complex idea closely related to health that has two meanings relevant to ecosystems. The first important meaning of integrity is that it requires an entity to be unimpaired when compared to an original condition. This implies that a reference condition must be defined and that the presence or absence of integrity can be measured relative to this standard state. Health is also often measured with respect to a reference condition. A second relevant meaning of integrity is that it defines a state of organic unity, or wholeness within a system. Health and integrity are complementary ideas that together are capable of describing the well-being of any system. Because these ideas are so closely related (both share the meaning of soundness), they have not always been clearly distinguished in scientific and other writings.

2. HISTORICAL BACKGROUND ON ECOSYSTEM HEALTH

Early depictions of ecosystems as organized networks by A. J. Lotka, A. Leopold, G. E. Hutchinson, and R. L. Lindemann and their insights into the role of energy flow, storage, and conversion efficiency in the healthy organization and functioning of ecosystems provided the basis for later assessment of ecosystem health in terms of the status and dynamics of various energy measures. However, the direct application of health, and its sister concept integrity, to ecosystems can be traced to the writings of Aldo Leopold during the late 1930s and 1940s. Leopold developed the concepts of a land ethic and land health that presaged the scientific interest in ecological integrity and ecosystem health, respectively. Leopold equates health of the land to its functional integrity, which he defines as a state of vigorous self-renewal in the land as a whole and in its components (soil, water, animals, plants, etc.). Leopold's concept of land is similar to our present idea of an ecosystem. Leopold's poetic description of land as a hierarchy of energy circuits implicitly called for the development of a quantitative tool like the energy circuit language, which H. T. Odum developed 30 years later.

The scientific interest in ecosystem health had its origin in the field of stress ecology that developed during the late 1970s and 1980s. Some research in stress ecology built on the earlier work of H. T. Odum, who showed that stress in ecosystems had the common property of acting as an energy drain, preventing available energy from flowing into useful structures and functions. Hans Selye discussed evidence for a nonspecific, generalized stress syndrome in humans. This work served as a basis for moving from the idea of stress in an organism to that of stress in an ecosystem, which led David Rapport and others logically to a consideration of ecosystem health. In the late 1980s, more research on ecosystem health, based on an analogy with medicine, began to appear. The First International Symposium on Aquatic Ecosystem Health was held in July 1990 and set the stage for a new publication, *Journal of Aquatic Ecosystem Health*. David Rapport characterized ecosystem health as a transdisciplinary science, which includes the study of ecosystem structure and function, along with socioeconomic factors and human health as they relate to the overall well-being of an environmental system.

3. ENERGY SYSTEMS ANALYSIS, HEALTH AND EMERGY

Robert Costanza proposed that the health of an ecosystem could be characterized by the vigor, network organization, and resiliency of the system.

These three general properties of ecosystems are essentially the total energy flow through the network, the arrangement of those energy flows into an organized network, and the excess energy flow or scope for growth that an ecosystem can use to recover from disturbance. These three characteristics relate to ecosystems and their functioning over time, and therefore systems analysis and simulation are fundamental tools needed to characterize ecosystem health. In this article the concepts of ecosystem health and ecological integrity are defined, measured, and interpreted from the perspective of energy systems theory and other systems approaches, such as information theory (ascendency), that are capable of analyzing networks of energy flow. Since the focus is on energy indicators of ecosystem health, the authors judge this perspective to be the most informative. Other perspectives on ecosystem health exist and are valid for certain purposes, but they are not considered here.

Energy systems theory provides tools and methods for describing and analyzing the hierarchical organization of ecosystem networks. Analyzing the energy hierarchy led to the insight that all energies are not equal in their ability to do work in a network. Emergy measures the relative ability to do work of the energy flowing through different components in an ecosystem network. Emergy values are always based on an underlying energy flow or storage times an appropriate emergy per unit coefficient. Emergy places energies of different kinds on the same scale so that they are directly comparable. The base unit for expressing emergy in ecosystems is the solar emjoule, sej. An emjoule is a joule of a particular kind of available energy that has been used in the past to make a product or service. Where the coefficient converts energy to solar emergy it is called the solar transformity and has units of sej/J. Also, mass in grams can be multiplied by the specific emergy of the item in sej/g to obtain emergy. Power must be converted to empower to gain a true picture of the relative importance of changes in different energy flows. The maximum power principle must be evaluated in terms of empower to obtain an accurate criterion for judging the relative health of ecosystems.

4. ENERGY AND ECOSYSTEMS

Available energy drives every known process at every level of organization in the universal hierarchy of natural phenomena. The existence of available energy potentials of sufficient magnitude is the fundamental basis for all self-organizing structures that exist far from thermodynamic equilibrium (e.g., ecosystems, organisms, and societies). The use of available energy potentials enables ecosystems and their components (e.g., individuals, populations, and communities) to perform their natural functions. Ecosystems self-organize into hierarchical networks of components connected by energy flow. These flows interact to link ecosystem components into energy flow networks through pathways of power and control (Fig. 1). Some controlling feedbacks interact with external energy sources to capture more available energy for use in building the network. Hierarchical levels of organization are also linked by energy flow (Fig. 1). Energy is a common denominator among ecosystems that makes it possible to identify and characterize similar processes and functions in ecosystems of many kinds. Because of the fundamental role of power (energy flow) in ecosystem self-organization, it is logical to expect that measures of power will be useful in characterizing the functional integrity (health) of ecosystems and that changes in energy flow will measure possible disturbances in healthy ecosystem processes.

4.1 Energy Intake and Ecosystem Health

Ecosystems consist of biotic and abiotic elements organized as structures, gradients, and flows that are maintained far from thermodynamic equilibrium through the capture and transformation of available energy. The persistence of such organization is continuously threatened and disrupted by the dissipative tendencies of the established structures and gradients within the system and its surroundings. The second law of thermodynamics governs these tendencies toward disorder, and ecosystems as well as organisms must continuously draw in available energy to repair and counteract them. For ecosystems, many external forces such as solar radiation, heat, wind, nutrients, and invasive organisms can be either sources of available energy or disordering forces depending on their form, magnitude, or intensity. The similarity between the condition of ecosystems and organisms with regard to the maintenance of necessary structure has motivated the development of the various energy indicators of ecosystem health—for example, the ratio of the energy used in respiration to the structure that it supports is called the Schrödinger ratio, a measure of the energy cost to maintain a unit of structure.

Both the persistence and growth of ecosystems depend on the continuous influx of available energy

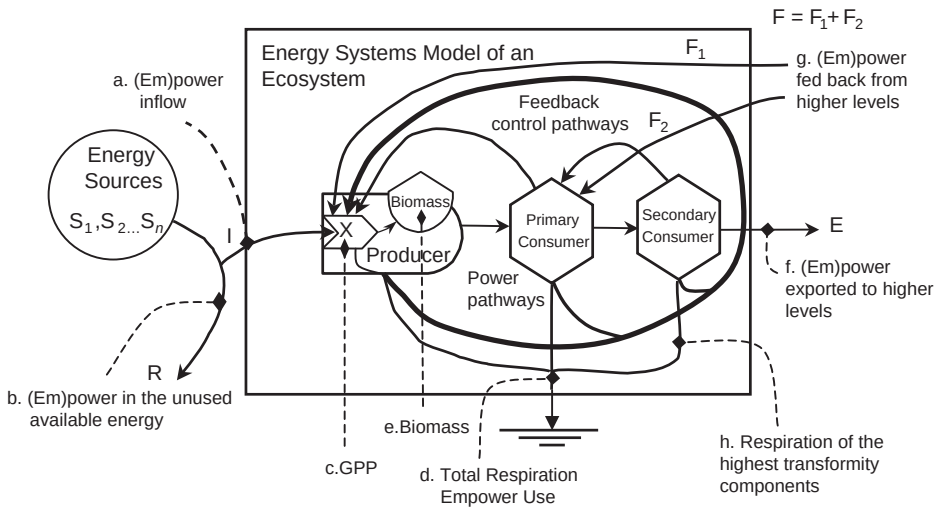


FIGURE 1 Energy systems model showing some of the general characteristics of ecosystems and the points where power and empower might be measured (labeled daimonds) to indicate ecosystem health. Symbols are briefly defined as follows: circles are energy sources, tanks are stored quantities, bullet symbols are producers, hexagons are consumers, and the block arrow is a work gate or interaction. Energy flows on the lines in the direction indicated by the arrowheads; light gray lines show flows of used energy to the heat sink (ground symbol), and the thicker, darker gray lines show the recycle of materials. S_i's are sources, F_i's are feedbacks from the larger systems, and E is the export.

and the transformation of this energy within the system to provide for its own self-renewal (i.e., structural maintenance and reinforcement of energy inflows) under the conditions imposed on it by its surroundings. When the structures, flows, and gradients within an ecosystem are organized to provide a capacity for maintenance and self-renewal that results in a sustained flourishing of the ecosystem, the system is healthy. For an organism or cell to remain healthy, both a sufficient metabolic rate *in toto* and an appropriate balance among the rates of specific metabolic processes are required to maintain the structural integrity and the functions on which its life depends. This same condition applies to ecosystems. Energy systems theorists have hypothesized that the maximum power principle governs the acquisition of energy by ecosystems. According to this principle, ecosystems prevail when the total rate of energy acquisition and conversion is sufficient to be competitive compared with the power acquisition rates of alternative ecosystem designs. If this principle is correct, ecosystem health depends on meeting this criterion for survival in evolutionary competition, and healthy systems will have total power flows that approach the maximum possible for a given set of inputs. To capture more of the available energy in external sources, ecosystems must develop an appropriate balance among the specific applications of this energy to maintain the internal structures and functions on which the viability of the ecosystem depends. Developing and maintaining the capacity to respond to pulsing on several hierarchical scales (the pulsing paradigm) is a condition for maximizing empower when energy is stored. Using the energy systems perspective and substituting emergy for energy and empower for power, the preceding concepts of ecological integrity and ecosystem health are defined and related in the following statement: ecological integrity is an emergent property of ecosystems that is greatest when all the structural components of a system that should be there based on the emergy sources received by the ecosystem are there (i.e., structural integrity is complete) and all the processes operating within this system of components are functioning at the optimum efficiency for maximum empower (i.e., the ecosystem is healthy). Therefore, ecosystem health (maximum network empower) is one component of the overall ecological integrity of an ecosystem, and structural integrity (completeness) is the other.

4.2 Properties of Healthy Ecosystems

Healthy ecosystems can be described by a relatively few properties that accompany maximal energy but more specifically emergy intake. Since ecosystems undergo a regular cycle of change, the expected values for these properties will be different depending on the stage of ecosystem development. One sign of stress for the ecosystem is a change in the expected course of succession and regression that characterize this cycle. The maximum power principle indicates

that healthy ecosystems will maintain optimally efficient energy transfers and nutrient cycling to maximize energy intake from available energy sources. This optimum ecological organization includes characteristic sets of species. These species sets vary with the stage of ecosystem development and with the particular ecosystem type, but in general a diverse species assemblage dominated by long-lived larger organisms characterizes the mature (climax) stage of ecosystem development. Healthy ecosystems also have the energy and information reserves to resist or adapt to change, and therefore they show resiliency in the face of perturbations or debilitating factors. The state of health involves optimization in terms of the allocation of resources because adaptation to any disordering factor entails tradeoffs with other factors including the maximum growth potential of the system.

4.3 Energy Indicators of Ecosystem Health

The characterization of the state of health of an ecosystem as defined earlier requires measures that express the condition of the ecosystem as a whole or some aspect of the system that can be quantitatively related to overall condition. Self-organization of ecosystems for maximum power has resulted in general design properties that are common to all ecosystems. The energy systems language model in Fig. 1 illustrates the following common design properties: (1) a set of external energy sources, (2) a web of interacting components, (3) convergence of successive energy transformations in a chain of increasing energy quality (emergy per unit energy), (4) storage of mass and information, (5) recycle of materials, (6) feedback control systems, (7) interactions of energy flows of different quality to maximize power by rewarding components that contribute, and (8) connection to other hierarchical levels of organization through imports and exports of energy. Important points where energy indicators of ecosystem health can be measured are marked on the diagram. The flows and storages of available energy that provide information about the health of ecosystems are (a) the energy inflow, (b) the albedo or unused energy, (c) gross primary production (GPP), (d) respiration of the whole system, populations, or groups of populations, (e) total biomass, the biomass of system components, and changes in biomass, (f) energy export from the system, (g) energy import into the system, and (h) the energy used by the highest transformity elements of the ecosystem. In the system diagram in Fig. 1, the total inflow of empower in the available energy captured (F + I) will be close to a maximum in healthy systems operating under a given set of external energies. When the external energy sources are constant or remain within an expected range, ecosystem health might be measured by comparison to reference levels of (F + I) or reference values of any other flow that has a known relationship to (F + I). When the spectrum of external emergies (the emergy signature) is changing, health can only be judged by comparison of the network empower of alternative system designs.

5. DIRECT MEASURES OF ECOSYSTEM HEALTH

Figure 1 shows many points at which the energy flow in an ecosystem network can be measured. Other measurement points are also possible. Each point of measurement tells us something different about ecosystem function and all will have their usefulness. The indices that will be the most useful in characterizing ecosystem health remain to be proven by empirical testing.

5.1 Biomass: Energy Storage as an Indicator

The most direct and earliest measures of energy status and dynamics include the enthalpy stored by the biotic community and the rates of enthalpy capture by and transfer within this community. Measures of community and population biomass provide an estimate of energy storage and its organization when standard factors are used to convert biomass dry-wt. values to enthalpies. Based on representative carbohydrate, lipid, and protein fractions in the major biological taxa, typical conversion factors range from 19 kJ/g for terrestrial plants to 26 kJ/g for vertebrates on an ash-free dry-wt. basis. Because total community biomass increases during ecosystem succession and generally decreases in response to stress, this measure can serve as an indicator of ecosystem health, particularly if a reference for comparison can be established. Relatively little effort has been made to identify such reference values due to practical difficulties arising from the great variability among ecosystems in community biomass and in the factors affecting it. If these factors could be adequately accounted for, total community biomass might provide a reasonable

initial indication of an ecosystem's health in terms of its energy flow.

Biomass alone could be a very misleading indicator, if diversity and balance among ecosystem elements are important components of ecosystem health. The productivity that produces a high biomass could result from the predominance of a minimal number of energy flows and conversions that are highly efficient under a narrow range of conditions. When such efficiency is attained at the expense of the redundancy and diversity of flows and conversions that are critical to the long-term persistence of the ecosystem, high biomass does not accurately indicate a robustly healthy ecosystem. Diversity measures can also be derived from the stored biomass in an ecosystem's respective populations to provide an index of organization that has been proposed as a predictor of ecosystem stability and resilience. However, no consistent relation between these health-related properties and such diversity has been established.

5.2 Productivity: Energy Flow Rate as an Indicator

Gross primary productivity is the rate of conversion of the energy captured by the biotic community or its respective autotrophic populations to chemical energy in biomass. It is an indicator that is closely related to energy flow within the system and thus should be more sensitive than biomass is to changes in health. However, if gross primary productivity is measured only at the community level, the homeorhetic nature of biological production and respiration severely diminishes the sensitivity of this indicator. A decrease in the productivity in one population is often compensated for by an increase in the productivity of another. Productivity is also quite variable and measurements must be taken over a long enough period of time to obtain a representative average. The diversity (Shannon-Weiner index) of productivity (energy flows among an ecosystem's populations) has also been considered as a measure of ecosystem health based on a hypothesized dependence of stability (a characteristic of healthy systems) on diversity. The extent of this dependence has yet to be established. Some increase in homeorhetic potential with increasing diversity is expected at low diversities. In general, diversity appears to correlate with ecosystem but not population stability. Even though methods exist to estimate gross primary production from measurements of total system metabolism, the difficulty of measuring gross productivity directly (because respiration simultaneously consumes the organic matter produced) has been an impediment to the successful development of productivity-related indicators of ecosystem health.

5.3 Respiration: Energy Dissipation as an Indicator

The community respiration rate has also been proposed as an indicator of ecosystem health because of its close association with processes of maintenance and repair. As one of the first effects of stress, the demand for these processes is expected to increase, and thus an increase in respiration could be an initial indication of the exposure of an ecosystem to an environmental change that adversely affects its health. In the longer term, community respiration typically declines as the productivity of an ecosystem is diminished by an adverse condition that reduces energy intake. Because processes of biotic self-organization and renewal, including growth and reproduction as well as maintenance and repair, are driven by the energy used in respiration, ecosystem health as described here is clearly related to respiration.

A healthy ecosystem might be characterized as one that maintains its respiration rate through periods of relative stability and in response to slow changes in its energy sources. Since stress acts as an energy drain, an increase in respiration rate over that expected for a particular ecosystem might indicate a decline in health. Optimal (or healthy) values for an ecosystem's respiration rate and for the time-dependent variability in this rate will differ with prevailing environmental conditions, and standards appropriate to any given set of conditions are not yet available. If such standards could be developed, deviations from the standard could provide a useful indication of a lack of healthy adaptation to extant stress. Where opposing effects of multiple stressors affect metabolic rates, simulation models would be needed to predict impairment. The difficulty and infrequency with which community respiration is measured and the complications involved in developing and validating diagnostic standards present obstacles to the use of community respiration rate as an indicator of ecosystem health.

5.4 Energy Absorbed: Energy Intake as an Indicator

The solar energy absorbed by terrestrial ecosystems and its complement the albedo are arguably the most direct measures of ecosystem health for these systems. This measure varies greatly among system

types, but it also varies within a system according to the overall condition of the system. Remotely sensed measures of ecosystem condition include the Normalized Difference Vegetation Index (NDVI), or greenness index, which shows clearly the effects of wet and dry years on the condition of forests and other terrestrial vegetation. Methods to determine the long-term changes in energy capture by an ecosystem show promise in documenting the regional effects of global climate change on ecosystem health. Because of the efficiency and power of remote sensing in characterizing large areas with high-resolution data, these methods may hold the most promise for developing future energy indicators of ecosystem functioning.

5.5 Energy Ratios as Indicators

Ratios among the preceding measures (i.e., *P*/*R*, *P*/*B*, and *R*/*B*, where *P* is gross community productivity, *R* is community respiration, and *B* is community biomass) have also been proposed as indicators of ecosystem health or disease (i.e., lack of healthy adaptation to an extant stress). Production and respiration are the two essential processes within ecosystems, and the ratio of these two energy flows provides considerable information on the state of the ecosystem (Fig. 2). In ecosystems that are not subject to large net imports or exports of organic matter, the *P*/*R* ratio will eventually become balanced (at close to 1). Imbalances in the forcing energy signature as well as stressful factors that alter community production or respiration or that increase respiration in response to biotic damage can shift this ratio away from 1, indicating impaired ecosystem health for these systems. The *P*/*R* ratio is less dependent on a comparison to previous measures of community metabolism or established standards for ecosystem types than is an indicator based on *P*, *R*, or *B* considered independently. Its interpretation is dependent on knowledge of external factors such as imports and exports of organic matter and of any

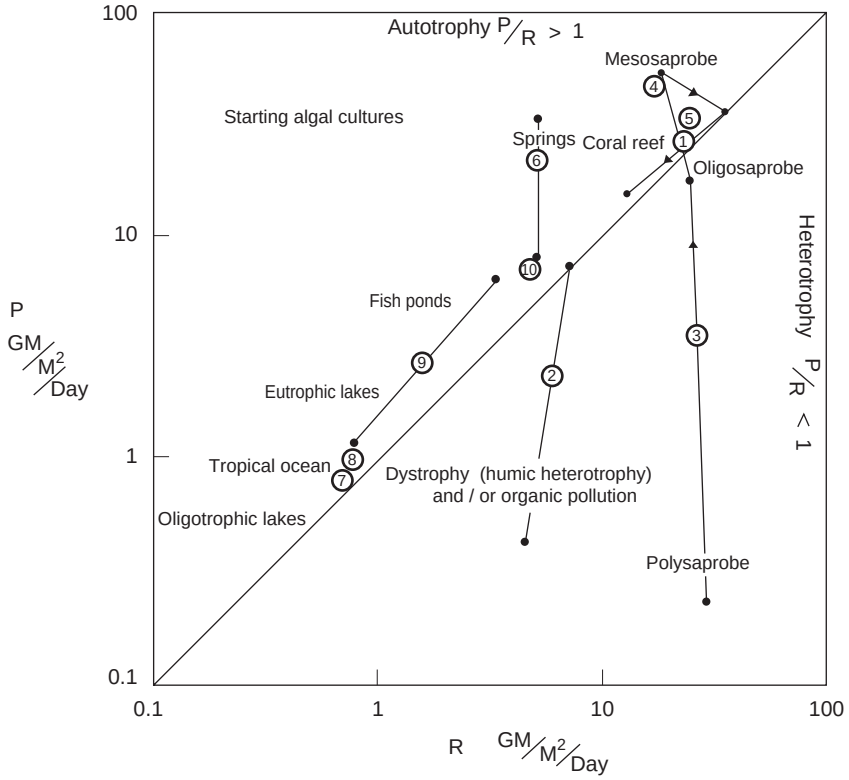


FIGURE 2 A diagram of the classification of ecosystems based on autotrophic (P) and heterotrophic (R) metabolism. Reprinted with permission from *Limnology and Oceanography* (1956), **1**(2),115. The original caption read as follows: A diagram for the functional classification of communities according to the total and relative dominance of autotrophic and heterotrophic metabolism. Specific examples are indicated as follows: (1) Coral reefs; (2) Itchen River, England; (3) recovery sequence on the White River, Indiana; (4) Birs, Switzerland; (5) Turtle grass; (6) Silver Springs; (7) Tropical Ocean; (8) Linsley Pond, Conn. (Riley, 1940); (9) Little John Pond, Wisconsin (Juday, Blair, Wilda, 1943); (10) fish ponds (from Winberg, 1937, as quoted by Juday et al., 1943); Ryther (1954) described the behavior of the P/R ratio in algal cultures.

large disturbances to which the ecosystem may have been subjected.

The ratios *P*/*B* and *R*/*B* have been advanced in association with the hypothesis that the self-organization of healthy ecosystems results in a maximization of biomass subject to available energy constraints. Thus, decreases in these ratios under a stable energy regime would indicate a stress that had undermined in some manner the healthy self-organization of the ecosystem. From an energy systems perspective trends in biomass are viewed as secondary to the priority of maximizing empower in the network. Thus, ecosystems build structure and organization as the means of increasing the use of the available energies.

The *R*/*B* ratio (sometimes also called the metabolic quotient, especially when measured in units of carbon stored and respired, or the Schrödinger ratio) has been the more thoroughly investigated of the two, especially with respect to soil ecosystems. The Schrödinger ratio is considered by some to be a thermodynamic order function in an ecosystem. E. Schrödinger, a physicist, showed that biological systems with molecular order affected by thermal vibrations must continuously pump out disorder if they are to maintain their internal order. Thus, the larger the biomass, the greater the maintenance cost. Although the energy required for maintenance of biological structure relative to the energy stored in that structure might be expected to increase with the application of stress, the observed direction of change in this ratio in the presence of common stress factors has varied with the ecosystem and the stress.

6. INDIRECT MEASURES OF ECOSYSTEM HEALTH

There are many possible methods for analyzing the properties of networks to indirectly measure ecosystem functional integrity or health. The three methods discussed in this section have been used in the literature to characterize ecosystem health. In addition, the reader should examine the references given in the bibliography.

6.1 Emergy Measures and Models

Emergy addresses the differences in the effects of different forms of available energy when used in a network. When the direct energy measures presented earlier are converted to emergy before calculating indices, a more accurate picture of the balance within the ecosystem should be gained. For example, the respiration of an ecosystem might decrease but the empower of respiration could increase if more of the energy in metabolism was due to higher transformity components. We expect that indices constructed using empower will be able to explain observations where power alone has failed. The fundamental theoretical connection between emergy measures and the response of systems to selective pressures might make these measures particularly suitable for indication of the long-term health status of ecosystems. Three emergy indicators of ecosystem health are (1) the empower used to maintain a given quantity of stored emergy in structure (similar to the respiration per unit biomass for an organism), (2) the dynamic transformity for the whole ecosystem or the ratio of the emergy required to produce a unit of energy flowing along all the pathways of an ecosystem to that unit of energy, which is an intensity factor for emergy in the ecosystem just as the temperature of an organism is an intensity factor for heat in the body, and (3) the ratio of the emergy stored to the emergy consumed in an ecosystem or the emergy *P*/*R* ratio, which may have characteristic values for systems with different characteristic emergy signatures. Figure 3 shows an example of two of these indicators calculated for control ecosystems and ecosystems impacted by hot water at Crystal River, Florida. Although emergy indicators are theoretically promising, they have not been tested and evaluated in large-scale studies of ecosystem health.

Simulation modeling is a viable means to develop indicators of ecological integrity and to identify criteria for ecosystem health. Ecological integrity and ecosystem health can be characterized through an analysis of validated simulation models of energy flow, when it has been demonstrated that the network of interactions in the model captures those factors that are primarily responsible for the behavior of the real system. In model simulations, ecosystem health is determined by calculating the empower at various points in the network of energy flows being simulated. The empower at any point in the network is calculated by multiplying the energy flow along the pathway by a dynamic transformity for that pathway. The value of the transformity depends on the external sources delivering emergy to the system as well as on the pathways of energy transformation within the system. Relative ecosystem health has been hypothesized to increase with empower on the pathways. In some studies of simulated ecosystems, this value has been found to

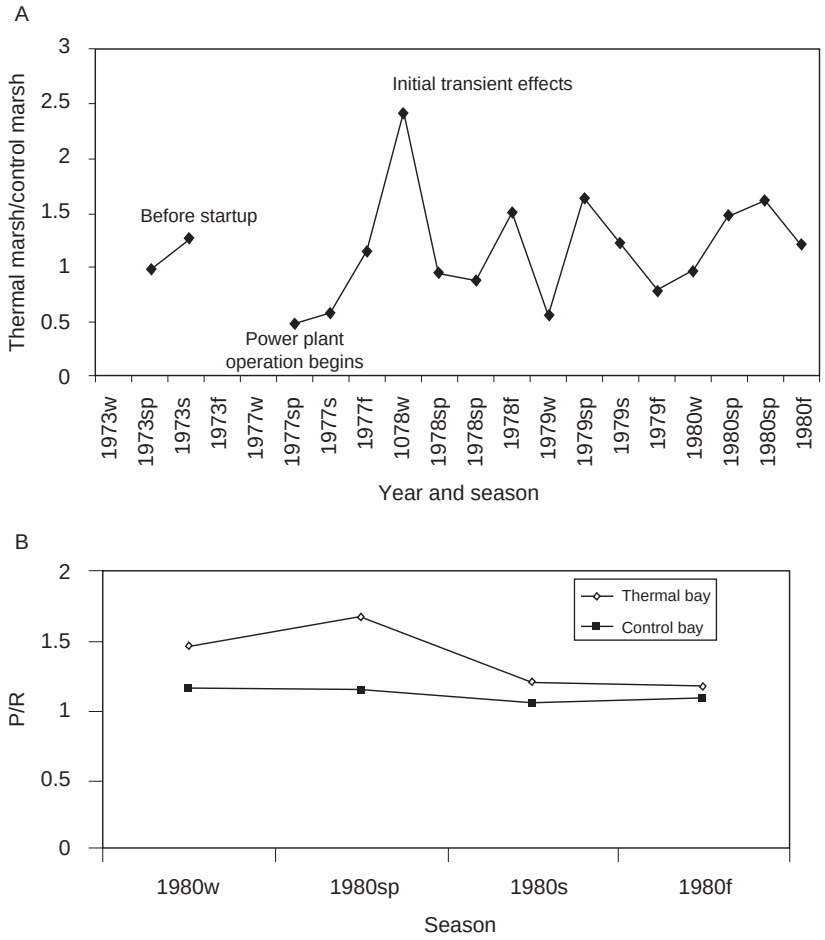


FIGURE 3 Emergy ratios illustrating patterns of health and disturbance. (A) Comparison of Schrödinger ratios for thermally impacted and control marshes. (B) The emergy P/R ratio of a thermally impacted versus a control bay. Reprinted with permission from Campbell, D. E. (2000), Using energy systems theory to define, measure, and interpret ecological integrity and ecosystem health. *Ecosystem Health*, **6**(3), 195. The original caption read as follows: Emergy indicators of ecological integrity calculated using data from ecosystems affected by hot water effluent and control ecosystems at Crystal River, Florida. (A) The ratio of the fraction of emergy in live biomass dissipated in daily maintenance for *Spartina alterniflora* from thermally affected and control marshes. (B) The emergy P/R ratio for the inner thermally impacted bay and its control bay.

approach a maximum for a given set of external energy sources that remain constant. There are special rules for determining the emergy of storages. The empower of the network includes information on the energy flow on each pathway as well as information on the position and connectivity of the pathways, which is supplied by the dynamic transformity. Empower can be determined at any point in the system where there is an energy flow.

The capacity to measure anthropogenic stress factors as emergy flows and to integrate these flows within a single dynamic systems model of the human-ecosystem interaction provides a basis for both ex ante and ex post assessments of the effects of human decisions and activities on ecosystem health. This capacity also permits derivation of relative measures of ecosystem health that reflect dynamically changing human preferences for ecosystem services and functions. Within such an integrated dynamic systems analysis, not only can the subjective differences in human preferences be incorporated, but also the contributions of these preferences to the long-term likelihood of persistence and prevalence of both ecosystem and human economic subsystems can be estimated in terms of emergy acquisition and use. Furthermore, such dynamic analysis permits assessment of adaptation to fluctuating emergy inputs and the consequences of external or internal patterns of pulsing (i.e., emergy storage and release) that would not be possible through a static analysis.

6.2 Exergy

Exergy (in the sense in which this term has typically been used by systems ecologists) is a measure of the thermodynamic potential of an ecosystem with respect to a system at the same temperature and pressure and with the same elemental composition but in a state of thermodynamic equilibrium. This potential includes both the maximum work that could be acquired (through an entropy-free transfer of energy) from the ecosystem as it attained this equilibrium state and an estimate of the enhanced capacity for capturing or degrading a flow of available energy provided by low-probability structures maintained within the ecosystem. The first component of this potential (the maximum attainable work) has been estimated as the chemical potential of the organic matter in an ecosystem relative to that in the equilibrated system (i.e., using $-\Delta G_{org}$, the Gibbs free energy released by a unit of this organic matter—or of a representative combination of carbohydrates, fats, and proteins—upon complete oxidation).

There is no consensus on how (or in reference to what flow or process) the second component or capacity might be calculated, but the inclusion of this component reflects the aforementioned understanding that different forms of (available) energy can have very different capacities to affect ecosystem organization and function. One approach, which was proposed by S. E. Jorgensen, hypothesizes that the capacity of living organisms to increase the available energy capture by an ecosystem is positively correlated with the number of genes that they possess. The corresponding exergy contribution of an ecosystem compartment is then approximated based on the probability of producing the amino acid sequences coded by the DNA of the organisms in that compartment. The relevant compartments are selected by the ecosystem analyst based on their importance in maintaining or altering the thermodynamic gradients associated with the ecosystem. These probabilities are computed under two simplifying conditions: (1) the 20 amino acids used by organisms are all equally likely to appear at any point in the sequence and (2) each gene codes for a sequence of 700 amino acids. Given a typical number of genes possessed by the organisms within a compartment equal to n, the approximation for this probability is thus $20^{-700 \cdot n}$, and the corresponding exergy contribution is defined as $-R \cdot T \cdot c_i \ln 20^{-700 \cdot n}$, where R is the universal gas constant, T is the absolute temperature, and c_i is the concentration (e.g., in g/L of an aquatic system), of carbon, for instance, in compartment i. The total exergy of an ecosystem with M organic compartments, as estimated using Jorgensen's formulation, can thus be expressed as

$$Ex = -\Delta G_{org} \cdot c_{org} + R \cdot T \cdot \sum_{i=1}^{M} c_i \cdot \ln 20^{700 \cdot n_i},$$

where c_{org} is the total concentration (of the chemical element or species of interest) in all organic compartments of the ecosystem (and ΔG_{org} is in J/unit of this element or species in the organic matter).

Due to the first term in this formula, exergy is highly sensitive to changes in available energy capture. Changes in structure related to shifts in species composition affect the second term, although a substantial shift in composition will not affect exergy if the number of genes and concentrations in the new taxa are similar to those in the former taxa. Furthermore, the sensitivity of an exergy index to important changes in ecosystem structure depends on the success of the analyst in selecting those compartments that are most closely related to these changes. The preceding formula is offered only as an approximation that provides one index for the assessment of the relative health of similarly analyzed ecosystems or of the relative effects of changes in an ecosystem or its environment on the health of that ecosystem.

Ecosystem resilience is not captured in any direct manner by the formula, but modeling studies have indicated a relation between exergy and the resistance of ecosystem state variables to perturbations. On average, resistance increases with increasing exergy, but exergy does not provide specific information on the resistance and resilience of critical ecosystem functions with respect to potential changes in particular forcing functions. The structural exergy, or that component of total exergy contributed by the internal structure of the ecosystem, rather than by its size, is more specifically sensitive to changes in diversity that are likely to affect ecosystem health. This index can be calculated as $Ex/(R \cdot T \cdot c_{tot})$, where c_{tot} is the total concentration of the element or species of interest in all compartments of the ecosystem. Increases in diversity will tend to increase the structural component of exergy if these diversity increases reflect a fuller utilization of resources within the ecosystem through the realization of additional potential niches.

6.3 Ascendency

Ascendency (an index proposed by R. E. Ulanowicz) is a summary measure of system activity derived from the network of flows associated with that activity, in which each flow is associated with a source compartment and a destination compartment within the system or its surroundings. As such it integrates a measure of the total flux through the system with an estimate of the extent to which the diversity of the fluxes among the respective compartments is organized by preferential paths of flow. (While the relevant compartments within the system must be determined by the ecosystem analyst, all flows from the system are assigned to one of two external compartments [i.e., either export or dissipation], and all flows into the system are assigned to a single import compartment.) When the flows are measured in units of available energy, the total flux through the system (T) is the sum of all available energy fluxes among all the compartments. As such T reflects the vigor of an ecosystem's processing of available energy and thus its capacity for growth and self-renewal under the prevailing conditions at the time of its measurement.

The (Shannon-Weiner) diversity of these fluxes, when multiplied by T, is suggested as an indicator of the total capacity (C) for organizational development. In other words, the number of compartments, evenness of flux among compartments (at any given scale of compartment aggregation), and T all correlate positively with the capacity for organization of flows within a system. In this context, organization refers to the statistical dependence of the destinations and sources of the flows in a network. Thus, in a system with no network organization, the likelihood of a flow of available energy being directed to any given compartment would be independent of which compartment supplied the flow. Because optimal ecosystem health requires that the flow of available energy be selectively directed to its most efficient and productive processes of maintenance, repair, and increase, optimal health will coincide with an optimal organizational development in the sense indicated here, particularly if the selection of compartments by the analyst reflects the intrinsic division of viability-sustaining functions among the ecosystem's critical processes.

The extent to which flow destination is independent of (or unconditioned by) flow source can be expressed as a function of the conditional influx probabilities of the compartments. The conditional probability of flow to compartment j given that the source of this flow is compartment i is estimated as the flux from i to j divided by the total flux from i ($f_{ij}/f_{i\cdot}$). The conditional probability of flow from compartment i given that the destination of this flow is compartment j is similarly estimated as the flux from i to j divided by the total flux to j ($f_{ij}/f_{\cdot j}$). These terms are then used in the formula

$$O = -\sum_{i,j} f_{ij} \cdot \log_2\left(\frac{f_{ij}}{f_{i\cdot}} \cdot \frac{f_{ij}}{f_{\cdot j}}\right) \quad \text{for } i,\ j = 1 \text{ to } N,$$

where O (the system overhead) is the product of T and a measure of the independence of destinations and sources that has a maximum value equal to the $\log_2$ Shannon diversity of fluxes within an unorganized network, and N is the total number of compartments. Thus, O attains its (theoretical) maximum value of C in a network in which destinations are completely independent of sources. It attains its minimum value of 0 in a network (with all flows supported by a single import compartment) in which flow destination is completely determined by flow source (i.e., in which knowing the source of a flow removes all uncertainty as to its destination). Accordingly, ascendency ($C - O$) is an indicator of the organization of flows within the network scaled by the total flux through the network. Increases in vigor, in evenness among the total fluxes to (or from) a network's compartments, and in the selective direction of available energy in the network all contribute to a higher ascendency. Increases in the number and evenness of the influxes to each compartment, on the other hand, decrease ascendency while increasing O.

Maximum ascendency requires complete determination of flow destination by flow source, which occurs when flow from each compartment is directed exclusively to a single destination compartment. All ecosystem compartments and processes (maintained as they are away from thermodynamic equilibrium) must contribute some flow to the dissipation compartment, however, in addition to whatever flow they contribute to any other compartments that they support. Thus, this theoretical maximum is not attained. Furthermore, although the direction of all flow from each compartment to a single process might be compatible with maximum long-term efficiency and productivity—and thus with peak ecosystem health—in the absence of any environmental variability, optimal health must be achieved under conditions of varying environmental provision, stress, subsidy, and

disturbance. Thus, the ascendency attained at optimal health reflects the maintenance of the diversity of flows required to persist and thrive under this environmental variability.

Due to uncertainty with respect to the optimal balance between immediate gains in vigor through selective reinforcement of the most productive flows and the maintenance of other flows that might be critical to vigor or viability under possible or likely future conditions, no precise value for optimal ascendency is suggested. It might be possible instead to establish ranges in ascendency typically maintained by persistently thriving ecosystems within given environments and energy regimes. Alternatively, it might be necessary for ascendency to be used in conjunction with *O* (or with *O* and *T*) in the assessment of ecosystem health. In a plot of ascendency versus *O* for a given class of ecosystems, for instance, a region typical of persistently thriving ecosystems might be identifiable. Although ascendency is sensitive to all of the components of ecosystem health discussed previously, further empirical investigation is required to determine whether sufficient information is retained by this summary measure (if used alone) to provide an accurate or useful indication of ecosystem health. Of the three related measures, ascendency is most specifically associated with organization selectively reinforced for maximum efficiency and productivity, *T* is most specifically indicative of ecosystem vigor, and *O* of adaptive resiliency. A general use of any of these three measures to indicate health status will require comparison among ecosystems across similar scales and standards of aggregation. A more precise characterization of the optimal balances in the allocation of available energy between the most productive flows and potentially critical alternatives under a relevant range of energy and disturbance (or stress) regimes will also be needed.

Acknowledgments

The authors thank Environmental Protection Agency internal reviewers, H. Walker, S. Brandt-Williams, E. Dettmann, C. Strobel, and W. Munns for their helpful comments. This article is Contribution No. AED-03-096 of the U.S. Environmental Protection Agency's Office of Research and Development, National Health and Environmental Effects Research Laboratory, Atlantic Ecology Division. Although the research for this article has been funded in part by the U.S. Environmental Protection Agency, it has not been subjected to agency review. Therefore, it does not necessarily reflect the views of the agency.

SEE ALSO THE FOLLOWING ARTICLES

Acid Deposition and Energy Use • Aquaculture and Energy Use • Easter Island: Resource Depletion and Collapse • Ecological Footprints and Energy • Ecological Risk Assessment Applied to Energy Development • Ecosystems and Energy: History and Overview • Greenhouse Gas Emissions from Energy Systems, Comparison and Overview • Hazardous Waste from Fossil Fuels

Further Reading

Campbell, D. E. (2000). Using energy systems theory to define, measure, and interpret ecological integrity and ecosystem health. *Ecosystem Health* **6,** 181–204.

Costanza, R., Norton, B. G., and Haskell, B. D. (eds.). (1992). "Ecosystem Health". Island Press, Washington, DC.

Jorgensen, S. E. (1997). "Integration of Ecosystem Theories: a pattern." Kluwer Academic, Dordrecht.

Leopold, A. (1991). A biotic view of land [1939]. *In* "The River of the Mother of God and Other Essays by Aldo Leopold" (S. L. Flader and J. B. Callicott, eds.), pp. 212–217. University of Wisconsin Press, Madison. WI.

Lotka, A. J. (1925). "Elements of Physical Biology." Williams and Wilkins, Baltimore. MD.

Odum, E. P. (1985). Trends expected in stressed ecosystems. *Bioscience* **35,** 419–422.

Odum, H. T. (1994). "Ecological and General Systems." University Press of Colorado, Niwot, CO.

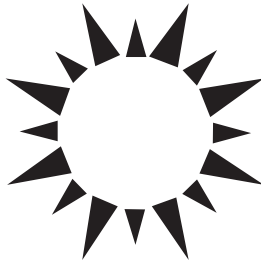
Odum, H. T. (1996). "Environmental Accounting: EMERGY and Environmental Decision Making." John Wiley & Sons, New York.

Odum, W. E., Odum, E. P., and Odum, H. T. (1995). Nature's pulsing paradigm. *Estuaries* **18,** 54–555.

Rapport, D., Costanza, R., Epstein, P. R., Gaudet, C., and Levins, R. (eds.). (1998). "Ecosystem Health." Blackwell Science, Oxford.

Ulanowicz, R. E. (1997). "Ecology, the Ascendent Perspective." Columbia University Press, New York.

Ecosystems and Energy: History and Overview



CHARLES A. S. HALL
State University of New York, College of Environmental Science and Forestry
Syracuse, New York, United States

1. Introduction
2. Energy and the Structure of Ecosystems
3. Origin and Approach of Ecosystem Studies
4. Biogeochemical Cycles
5. Stability and Disturbance
6. The Future of Ecosystems
7. Conclusion

Glossary

biogeochemical Pertaining to the abundance and movements of chemical elements or compounds in and between various components of an ecosystem or between ecosystems.

density dependence The situation in which the number of animals or plants at a future time is dependant on the number at an earlier time.

ecosystem A unit of nature, generally from 1 hectare to some thousands of square kilometers, with boundaries defined by the investigator, that includes plants, animals, microbes, and the abiotic environment.

endogenous A factor from inside the system that influences the system.

exogenous A factor from outside the system that influences the system.

trophic Pertaining to food.

We may consider ecosystems as machines that have been selected to use solar and occasionally other energies to reorganize the raw materials of the earth into systems that support life. Ecosystems, and their component organisms, use this energy to rearrange the chemical molecules of the earth's surface and of the atmosphere into living tissue according to those patterns that have high survival potential. They are, in this context, antientropic—that is ecosystems (and life in general) work against the tendency for the molecules of the world to degrade into relatively random arrangements. As such they are self-designed, in that their organization has emerged through natural selection for structures that capture, hold on to, and invest energy into further structure in a way that appears to be self-designed) the structure is generated to continue life and enhance over time its ability to capture and use energy). More generally, life is most explicitly nonrandom arrangements of molecules, and all organisms use energy to create and maintain these specific molecular arrangements. But individual organisms cannot survive in isolation because they need the supporting context of other species, the soil or other substrate, the atmosphere or water around them, proper hydrological cycling, and so on. Thus, to our knowledge, ecosystems are the minimum units of sustainable life outside of the laboratory. The proper functioning of the earth's ecosystems is also essential for human beings—for atmospheric and climatic stability; for the production of food, fiber, and clean water; and for physical and psychological well-being.

1. INTRODUCTION

1.1 Early Studies at the Ecosystem Level

Although the English term "ecosystem" was coined in 1935 by Tansley, ecosystems have been the subject of human inquiry for millennia. The first written records that survive include the work of Aristotle, an insightful and knowledgeable natural historian, and especially his student Theophrastus, who was interested in how different species of trees grew in different locations. Of course, practical people living on (or off) the land (fishermen, ranchers, hunters, trappers, many farmers) also knew and know a great deal of a practical nature about how ecosystems operate.

 © 2004 Elsevier Inc. All rights reserved.

Well before Tansley, the formal scientific study of ecosystems was well developed by Russian and Ukraine scientists in the early part of this century. The serious student of ecosystems should return to the seminal, often brilliant earlier writings of Stanchinskii, Vernadaskii, Morozov, and their colleagues. Vernadsky in particular was interested in how ecosystems, and even the living world as a whole (the Biosphere), functioned. Unfortunately the work, and sometimes even the lives, of these early ecologists was terminated by Stalin and his henchman Lysenko because their view of nature's limits to human endeavors was in opposition to official communist social engineering.

The practical implications of understanding energy flow in ecosystems became obvious to a number of early investigators who often attempted to determine how much food could be harvested from a unit area of land or sea. Stanchinskii, mentioned earlier, attempted to examine the quantitative importance of the flows at each trophic (food) level in an Ukranian grassland. Wells *et al*. examined qualitatively the flow of energy from the sun through the food chain to the herring fisheries of the North Sea as early as 1939. The first study in North America is usually attributed to Raymond Lindeman, who quantified explicitly the flow of energy from the sun through primary producers to higher trophic levels in a bog lake in Minnesota. Lindeman generated most of his numbers from preexisting information but was extremely creative in organizing the organisms of an ecosystem as trophic (or food) levels, rather than simply by species. Although Lindeman made some errors (for example, by subtracting predation after estimating production from turnover rate), his work remains extremely important for our conceptualization of how energy flows through nature.

An especially important group of studies on the eneregetics of ecosystems were done by Eugene and especially Howard Odum. The brothers worked together measuring photosynthesis in a coral reef at Eniwetok. The most important of these studies was done by Howard on the ecosystem at Silver Springs, Florida. The particular advantage of this site was that the water flowing out of the main spring boil was always at a constant oxygen concentration, so that the deviation of the oxygen content of water downstream (higher during the day, lower at night) could be used to estimate photosynthesis and respiration of the entire ecosystem. Odum also measured explicitly the energy fixed or used by each trophic level by, for example, putting fish into cages and measuring their growth. Odum found that by far the largest proportion of the energy captured at a given trophic level was utilized by that trophic level for its own maintenance respiration, which was consequently unavailable to higher trophic levels. Lindeman had introduced the concept of trophic efficiency, defined as the ratio of production at one trophic level to production at the next. Trophic efficiency is commonly from 10 to 20% but occasionally may be very different. The concept is important and familiar in agriculture where beef or fish production per hectare is much less than the production of plants of that same area, due principally to the large maintenance respiration of both the plants and animals. Both Howard and Eugene Odum continued basic research on the relation of energy to ecosystem structure and function into the new millennia in astonishingly long and illustrative career in research and graduate training.

Research has emphasized that most trophic relations occur not as simple straight-line chains but as more complicated food webs, in which a given species and even different life stages of that species eat from different trophic levels. For example, a herring whose diet contained 50% algae and 50% herbivorous crustaceans would be assigned to trophic level 2.5. Many, perhaps most, organisms are omnivores rather than strictly herbivores or carnivores, and probably the majority eat dead rather than live food. The single most important attributes of food quality, other than its energy content, is its protein content, which is approximately proportional to the ratio of nitrogen to carbon in the food (Table I).

1.2 Ecosystems versus Species-Oriented Ecology

During this time something of a gulf developed between ecosystem-level ecologists and many other ecologists. This was due in part to the perception by

TABLE I

Properties of Food

Food type	Example	Kcal/gm	Percentage protein
Carbohydrate	Grains	4.25	2–14
	Beans	4.25	5–30
Fat	Butter	9.3	0
	Olive oil	9.3	0
Protein	Animal flesh	5.0	20–30

From Whittaker (1975) and USDA nutritional Web site.

some that ecosystem ecologists were paying insufficient attention to the basic laws and mechanisms of nature, including natural selection, that had been especially well developed by biologists at the individual and population level. Some ecosystem-oriented papers appeared to these ecologists to be almost mystical in their message that the system knows best. On the other hand, ecosystem-level ecologists could point to the failure of population dynamics to do a good job in prediction and management of fisheries, in large part because the population-level ecologists tended to ignore spatial and interannual variations in the ecosystem productivity that often determined year class strength. As part of this debate throughout the literature of ecology, there has been considerable discussion and even tension about the relative importance of exogenous (meaning external to the system, such as climate variability, another phrase meaning the same thing is forcing function) controls versus endogenous (meaning internal to the system, such as density dependent survival) controls for determining the structure and function of ecosystems. To some degree the argument is specious, for obviously both are important, although the endogenous aspects have been overly emphasized. Clearly, as described here, climate (an exogenous factor) is critical in determining the basic possibilities for life, and evolution (basically an endogenous factor) is important in determining the response of living creatures in a region to the climate.

Part of the problem leading to this dichotomy was the seemingly completely disparate perceptions of the different groups on what was the important currency of ecology. Energy was perceived as the currency of the ecosystem-level ecologists, and fitness (reproductive potential over time that includes the abilities to both survive and reproduce) was seen as the currency of the population-level ecologists, and each group tended to dismiss, argue against, or ignore the perspectives of the other. These were spurious arguments, and although many previous papers hinted at the importance of energy for fitness, this has been especially well done in papers such as that by Thomas *et al*. titled "Energetic and Fitness Costs of Mismatching Resource Supply and Demand in Seasonally Breeding Birds." This is one of the best papers in ecology, and there is more good science packed into the two-and-one-half pages of this paper than in many books. Simple observations and measurements, and double labeled isotopes, were used to examine energy costs and energy gains of various populations of European tits (chickadees). Those birds that timed their migrations and reproductions appropriately to the availability of their preferred high-energy food (caterpillars, in turn dependent on the phenology of their oak-leaf food source) were much more fit in the evolutionary sense. That is, birds with optimal timing were much more successful at raising more and larger (and hence more likely to survive) offspring and were themselves more likely to survive and reproduce again the next year. Birds that timed their reproduction inappropriately relative to the caterpillar dynamics worked themselves into a frazzle trying to feed their less successful offspring. Finally Thomas *et al*. found that climate change was interfering with past successful timings because the birds were tending to arrive too late, since the phenology of the trees was responding more quickly to temperature increases than was the impetus for the birds to migrate. Presumably birds that had genetically determined behavioral patterns that were getting them on the breeding grounds too early in the past would be selected for if and as the climate continued to become warmer. This study also indicates the importance of the quality of energy. In this case, it was not simply total trophic energy available to the birds that was critical but also their quality as represented in their concentration as large food packages, showing again the importance of energy quality as well as quantity.

The paper by Thomas *et al*. shows elegantly and more explicitly than earlier studies what many suspected for a long time: probably the principal contributor to fitness is the net energy balance of an organism. The idea that energy costs and gains are related directly to fitness is not a new one, found, for example, as Cushing's match/mismatch hypothesis in fisheries, which in turn is a restatement of Hjort's 1914 paper about year class abundance in Scandinavian cod. Energetics, combined with climatic and other external forcing, can explain population dynamics much more powerfully than the density dependant population equations that still seem to dominate much of population ecology. Clearly density dependence exists. But how often is it the principal determinant of population year class strength? Perhaps the best way to consider the importance of density dependence is what Donald Strong calls density-vague relations—that is the empirical information shows that density dependence works sometimes but does not operate importantly every year. The way that we have set up many of our experiments in ecology virtually guarantees that we will find some density dependence even when it is not dominant.

1.3 General Principles of Energy for Ecosystems

Some general principles about energetics that have emerged from all of these studies are that (1) the world of nature is (of course) as beholden to the laws of thermodynamics as any other system; (2) ecosystems can be considered the products of self design—that is, given the environmental circumstances where they are found they "create" through natural selection (operating at various levels, from genes to organisms to species to coevolved groups of organisms) the systems that we see today; (3) virtually every ecological or even biological action or interaction has an important or even dominant energy component; (4) ecological efficiency, or transfer from one place to another, is low, from one or two to a few tens of a percent, principally because of the need for each level of a food chain to support its own maintenance metabolism; (5) each organism is an energy integrator that integrates through growth and reproductive output (i.e., fitness) all energy costs and gains; and (6) each specific organism is adapted to its own biophysical environment in a way that maximizes this difference between energy costs and gains. As a consequence, the collective energy captured and used by that ecosytem is thought to be maximized; (7) species are ways of packaging morphologies, physiologies, and behaviors that allow sufficient, or even maximum, reproductive output as a consequence of having more or less optimal patterns for that given environment. The ecosystems that have survived over time are thought by many to be maximizing the power (by operating at intermediate, not maximum, efficiency) obtainable under the existing physical circumstances.

More recent research on ecosystems and energetics (i.e., since about 1980) has focused on increasingly detailed examinations of the role and activity of particular species, a greater understanding of the factors that control ecosystems, the importance of disturbance, and recovery from disturbance and the role of microbial components in decomposition and regulation. An important component of this has been the 20 or so LTER (long-term ecological research) centers where basic studies are carried out for long time periods to see how, for example, ecosystem components respond to climate variability, long-term exclusion of herbivores, and so on. This approach has allowed the comparison of the basic patterns of ecosystem.

2. ENERGY AND THE STRUCTURE OF ECOSYSTEMS

Neither ecosystems nor their component species can exist without a constant supply of energy to maintain the biotic structures and their functions. The source of this energy is in almost all cases the sun. Clearly the sun runs the energy, carbon, and nitrogen fixation of green plants. Less obviously the sun does many other things for ecosystems: most important it evaporates and lifts water from the ocean and delivers it to continental ecosystems, replenishes carbon dioxide and other gases through winds, pumps water and associated minerals from the roots to the leaves via transpiration, provides nutrients through weathering, and so on.

Ecosystem studies tend to be different from much of the rest of ecology, and indeed science in general, in their emphasis on interdisciplinary and interactive studies. In most other subareas of science, including ecology, the emphasis is on understanding a specific phenomenon and on carefully controlled studies under exacting laboratory conditions. In contrast, when studying ecosystems it is generally necessary to consider aspects of physics, chemistry, geology, meteorology, hydrology, and biology, and in many cases some of the social sciences as well. Since it is impossible to replicate and control all of these factors carefully, as would be the case in the laboratory, ecosystem studies are more likely to be based on observation or correlation than experimental approaches. In these senses and in others, ecosystem science tends to be holistic—that is, it emphasizes the entirety of both a system and a problem rather than just some aspect, and it frequently uses the conceptual tools of systems analysis. For example, if we are to understand the relation between a particular ecosystem and the atmosphere, we obviously must consider the physics of the response of organisms to temperature, the chemistry of the relation of carbon exchange between plants and the atmosphere, the possible effects of changing climate on hydrology and hence the availability of the water needed for the plant to grow, the movement of gases through the atmosphere, and the social factors that are causing the carbon in the atmosphere to increase. In addition, we often need special integrative tools that are outside the traditional scientific disciplines that allow us to examine all or many of these factors in interaction. These tools include quantitative flow diagramming, computer simulation, and theories that are transcomponent.

Such a holistic, or, more accurately, systems perspective is increasingly important as human

impacts are changing many aspects of the earth rapidly. Such a perspective does not obviate other approaches to science; in fact, it tries to incorporate and integrate as possible all relevant levels and techniques of analysis. The problems that the increasing human population and its growing affluence are putting on the world require a whole different approach to science—one that is difficult if not impossible to do in the laboratory. Nevertheless it must adhere to the same basic standards as the rest of science in that the results of the science must make predictions that are consistent with the real world. This can lead to certain problems when scientists concerned with the destruction of ecosystems and their species are faced with attempting to apply their results to saving nature. The science of ecology is full of generalizations about the supposed virtues of undisturbed nature that have led to theories such as diversity and stability, diversity and productivity, and the intermediate disturbance hypothesis whose validity seems to wax and wane with each new study. Separating the robust ecological theory that will help with the genuine environmental crises we face from the good sounding but poorly extrapolateable theories has often been a difficult feat.

3. ORIGIN AND APPROACH OF ECOSYSTEM STUDIES

The word ecosystem is a combination of the words "ecology" and "system." The word ecology is derived from the Greek word *Oikos,* which means pertaining to the household, especially to the management of the household. The academic discipline of ecology refers to the study, interpretation, and management of our larger household, including that of the planet that supports our species and that provides the basis for our economy. System refers to all of the (or all of the important) components of whatever is being considered and the pathways and rules of interaction among the components. Thus, ecological system, or ecosystem, refers to the structure and function of the components that make up household earth and the interactions among those components.

Ecosystems generally are considered from the perspective of their structure, that is their composition and changes in structure (in terms of geological land forms, number, diversity, and abundances of species, biomass, height of vegetation, abundance of critical nutrients, etc.), their function, that is the pathways and rates of energy and nutrient flows, and the regulatory or control processes that govern those changes and flows. We might also want to ask many other kinds of questions about, for example, the diversity of the species that live there or how that diversity effects the function of the ecosystem. Ecosystem structure and function are related over evolutionary time. For example, tall trees functionally need to supply water and nutrients to growing leaf tips, which in the case of the rain forest might be 60 m above the ground. Consequently they have evolved elaborate structures, including a system of pipes, to pump the water to those growing tips. This in turn is possible only where the soil is generally wet and the air is dry, at least periodically.

An extremely useful basic assessment of the most important attributes of ecosystems is found in the classic paper "Relationships between Structure and Function in the Ecosystem" by E. P. Odum. Odum compared the basic attributes of an evergreen forest (such as might be found in the Southeastern United States) and a deep ocean ecosystem (such as might be found a few hundred kilometers to the west of that peninsula). Odum found that these ecosystems are very different in their physical structure—that is the physical characteristics of the ecosystem itself. The organisms themselves constitute the major physical structure in the rain forest. It is possible for the plants to be large here because the soil allows for an anchoring place for the tree, the climate is near optimal for plant growth, and there are no large grazing animals that can eat the tree. Therefore the trees can become very large and become themselves the principal determinant of the physical structure of the ecosystem.

In the ocean, by comparison, the lack of a firm substrate and the constantly changing position of a parcel of water makes it impossible for the organisms to change the nature of the physical medium (the water) significantly, and it is the water itself that forms the basic structure of the ecosystem. Since the organisms are constantly sinking or being swept away by turbulence, they must be small and be able to reproduce rapidly. The total biomass, or living weight, of the aquatic organisms is very small compared to that of the forest. Also, interestingly, the grazers (or animals that eat the plant) have nearly the same biomass as the algae in the water but are much less, as a proportion, in the forest.

One of the most interesting observations when comparing these two ecosystems is that although the biomass of the plants is thousands of times greater in the forest, the rates at which the plants capture the sun's energy (their primary productivity) is much more similar, especially in coastal areas (Table II). In

TABLE II

Ecosystem Structure and Function of Some Typical Ecosystems as Measured by Approximate Biomass and Production Levels

Ecosystem type	Biomass (dry tons/Ha)	Net productivity (dry tons/Ha/yr)
Tropical rain forest	225 (60–800)	10–35
Tropical seasonal forest	175 (60–600)	10–25
Savanna	40 (2–150)	2–20
Temperate deciduous forest	200 (60–600)	6–25
Temperate evergreen forest	200 (60–2000)	6–25
Boreal forest	200 (60–400)	4–20
Temperate grassland	16 (2–50)	2–15
Desert	7 (1–40)	0.1–2.5
Ocean	0.03 (0–0.05)	0.02–4
Estuaries	10 (0.1–60)	2–35

Based on summary by R. L. Whittaker and G. Likens, in Woodwell and Pecan (1973), modified by Sandra Brown, personal communication.

fact, given the greatly varying quantity of plant biomass around the world, most ecosystems have broadly similar rates of production, at least if there is enough water. Thus, we can say that the structure appears much more variable than the function, at least with respect to these particular attributes.

From an energy perspective, we know that the structure of an ecosystem is costly to build and costly to maintain. "Good" structure—that is structure that has been selected for—will be well worth it energetically for it will draw maximum power (i.e., useful energy sequestered per time) given the environmental possibilities, and it will become more abundant than those components that draw less power. Hence, structure and function, and the controls that link them, are intimately and functionally linked through natural selection.

One of the fundamental ideas used to analyze ecosystems is the idea of food chains, which is based on the transfer of energy originally derived from the sun through the biological system. The word "trophic" comes from the Greek word meaning food, and the study of feeding relations among organisms is called trophic analysis. More specifically, the study of trophic dynamics emphasizes the rates and quantities of transfer over time. Trophic processes are the essence of ecosystem processes, for they are the ways that solar energy is passed from the abiotic world to plants and to animals and decomposers, and they allow nonphotosynthetic organisms, including ourselves, to exist. These pathways of food are also a principal way that materials such as nutrients are passed from one location to another within an ecosystem. Trophic studies are an important component of ecology because trophic processes determine energy availability and hence what is and what is not possible for a given organism, and for all organisms collectively, within an ecosystem. In addition, many management concepts and objectives important to people, such as those relating to harvesting fish, timber, or crops, are oriented toward understanding and directing the trophic relations of nature. Energy is stored for a short time in the ecosystem in the biomass of the various organisms, soil, and so on, and some of this energy is transferred from one trophic level to the next (Fig. 1). For example, some of the free energy captured by the plants is eaten by heterotrophs, although a larger part is used by the plant itself. Thus, most of the energy once captured is lost from each trophic level as waste heat as organisms use that energy to fight entropy. Eventually, all the energy originally captured by autotrophs is degraded and returned to the surrounding environment as waste heat. Ecosystems do not recycle energy; instead, they capture a small percentage of the solar energy that falls on them, concentrate it, use it to do work, and then return it to the environment in the form of low-grade heat that is no longer available to do work. The principal work that has been done is the maintenance of life, and this has required energy to sequester and rearrange chemicals, overcome gravity, move either themselves (animals) or their chemical constitutes (plants and animals), generate microclimates, maintain general homeostasis, and then replicate the whole structure and functions through reproduction.

There is constant selective pressure operating on organisms to maximize the useful energy that they capture and to use it in ways that contribute to their survival and to propel their genes into the future. Many think that this process leads to the collective maximization of power at the level of the ecosystem—that is, those organisms and arrangements of nature will be selected for that most fully exploit available energy and use it most effectively compared to alternative organisms or patterns. Obviously this is a difficult idea to test at the level of entire ecosystems and on a scale of human generations, but it does offer intriguing hints that natural selection operates in ways that generate and maintain (or self-organize) entire systems. For example, although it is unlikely that many of the same species were present in the wet montane tropics 300 million years ago as are there now, it is likely that a recognizable

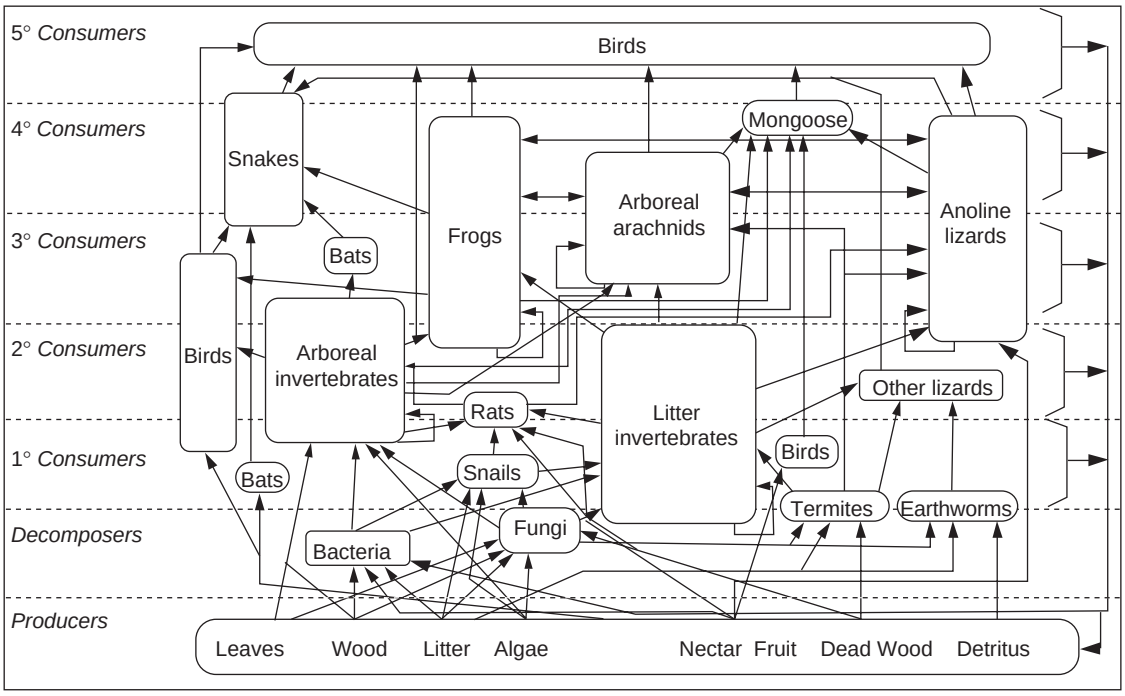


FIGURE 1 Energy flow diagram for the food web of the Luquillo Forest, Puerto Rica. Pictorial representation. Modified from Hall, C. A. S. (1999), Ecosystems, pp. 160–168, in D. E. Alexander and R. W. Fairbridge (Eds.), "Encyclopedia of Environmental Science." Kluwer Academic, Norwell, MA.

tropical wet forest was there that had similar functions and rates of processes as modern tropical wet forests.

3.1 Trophic Processes in More Detail

The pathway of energy conversion and transfer (the eating of one organism by another) goes from the initial capture of solar energy by autotrophs, to the herbivores that eat plants, to the first-level carnivores (that eat herbivores), on to the top carnivores. The principal pathways within a food chain can be represented as energy transfer through a series of steps. The power flow of an organism per unit area, or of a trophic level, is called productivity and normally is expressed in units of kilocalorie per square meter per time.

3.1.1 Autotrophs

In the first step autotrophs, or green plants, use chlorophyll to capture energy from solar-derived photons and store this energy by restructuring the carbon atoms of carbon dioxide derived from the surrounding atmosphere or water into complex organic compounds. Primary production is the fixation of solar energy by green plants. Gross productivity is total energy captured, whereas net production is that minus the energy required for respiration. Net energy is then allocated into tissue increments, leaf turnover, reproductive products, and, in time and through these processes, the replacement of the first plant with a second. In the second step herbivores or primary consumers obtain energy by eating autotrophs. Secondary production is the accumulation of animal or decomposer living tissue. Heterotrophs obtain metabolically useful energy from the consumption of the organic molecules in the food they obtain from other organisms.

3.1.2 Grazers

Since a substantial amount (perhaps 80 to 90%) of the energy captured by an ecosystem is used by the first trophic level for its own necessary maintenance metabolism, food chains are inherently and necessarily "inefficient." Thus, there is and must be much more production of biomass of deer than wolves or cougar in a forest, for the deer must use a large

portion of the energy in the leaves it eats for its own metabolism rather than for growth; much less is left for their own growth and the potential exploitation by the predators.

3.1.3 Decomposers

The world is covered with living creatures, and these creatures are continuously dying. Yet the world is not littered with carcasses, so obviously they disappear in some way. We call this process decomposition, and it is mediated principally by single-celled organisms, especially bacteria and fungi. In addition, the decomposition process gets a start initially by larger animals, from buzzards to maggots to protozoans, and in earlier days, even our own ancestors. Although humans are often disgusted by decomposition (and, in fact, have probably been selected for this quality), without decomposition the biological world would come to a stop, as all of the earth's available materials would be tied up in various carcasses!

Some food chains are complicated indeed. Think of a salmon feeding on the plankton of the mid-Pacific ocean and then swimming far up an Alaskan river, fueled by energy reserves built on this plankton, to spawn and die—or to be eaten by a bear. The offspring of the salmon that did spawn spend their first year of life in the freshwater environment, eating zooplankton that ate phytoplankton whose growth was fertilized by phosphorus leaching from the bones of the dead parents. Another way that food chains are complicated is that much of the energy flows not through live but through dead organisms. For example, in the rain forest most of the plant material is not eaten directly by live leaf munchers but instead by various insects and other small organisms that consume the dead leaf material called detritus.

3.2 Trophic Dynamics and Biomass Pyramids

Thus, of all the energy captured by green plants, progressively less and less flows to the next consumer, or, as it is often called, trophic level. When the rate of flow of energy is graphed, it nearly always will look like a pyramid, with the large base representing the energy captured by the autotrophs and each successive layer representing higher trophic levels, each further removed from the base (Fig. 2). This is called the pyramid of energy. Biomass plotted by trophic level may look like this but often looks more like a square than a pyramid. How can this be? The reason is that larger organisms often found in higher trophic levels generally respire (use up) energy more slowly; in a sense they get less energy, but they hold on to what they do get for a longer time. So in the ocean there is roughly the same biomass of algae, zooplankton, small fish, and large fish. The reason is that the small algae "turn over" much faster, meaning that they have a much higher rate of metabolism and loss to other organisms that eat them.

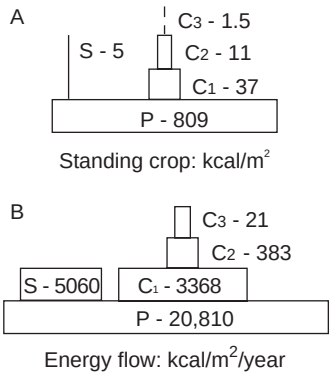


FIGURE 2 The trophic categories given in Fig. 1 represented as (A) an energy pyramid and (B) a biomass pyramid. Data from Silver Springs, FL. Modified from Hall, C. A. S. (1999), Ecosystems, pp. 160–168, in D. E. Alexander and R. W. Fairbridge (Eds.), "Encyclopedia of Environmental Science." Kluwer Academic, Norwell, MA.

4. BIOGEOCHEMICAL CYCLES

A second critical area for the study and understanding of ecosystems is the structure and function of nutrient cycles. Nutrients can mean all of the chemical elements that an organism or an ecosystem needs to grow and function, but most commonly the term is restricted to a consideration of nitrogen, phosphorus, potassium, and less commonly calcium, sulfur, iron, and micronutrients such as copper, cobalt, and molybdenum. Sometimes, but not generally, CO_2 is considered a nutrient, for it too can limit growth of plants. Finally, it also (rarely) can mean complex organic materials necessary for growth, such as vitamins. We refer to the movement of a particular nutrient through an ecosystem, often changing from one form to another, as a nutrient cycle. Nutrient cycling, like all other processes, occurs because of energy forces including evaporation, precipitation, erosion, photosynthesis, herbivory decomposition, and so on, all run by the sun.

4.1 Limiting Nutrients

The study of nutrients in ecosystems was first undertaken systematically by the German chemist Liebig, who found that the growth of plants tended to be limited by only one element at a time. For example, if a plant had insufficient phosphorus relative to its needs, its growth would stop until phosphorus was added. At some point adding more phosphorus had no further effects, but at that time some other element, perhaps potassium, might be limiting the plant's growth, as could be determined by noting the growth response, or lack thereof, of adding potassium. This led to the formulation of Liebig's law of nutrient limitation, which states that the growth of plants tends to be limited by a single nutrient at any one time, that which is least abundant or available relative to the needs of the plant. Nutrient cycles at the global level were initiated by Vernadsky and greatly encouraged by the growth of atomic energy in the 1950s, which provided radioactive tracers that helped follow nutrient pathways, funding to do that research, and also a reason as to why we might be concerned about them—that is, if there were releases of radioactive elements into the environment.

Nitrogen is especially important because plants need it in large quantities and it is often restricted in its availability and energetically expensive to get into a useful form. Plants and animals use nitrogen extensively because it is, after carbon, the most important component of proteins. Plants extract it from the soil, and animals get it by eating plants or other animals. All organisms must concentrate the normally relatively rare nitrogen to get enough to build their proteins. Waste nitrogen is released from animals as ammonia or other compounds in their urine. Since N_2, ammonia, and nitrous oxide are volatile, nitrogen is different from most other nutrients in having a gaseous phase.

The cycle of nitrogen in ecosystems is also complicated because nitrogen is a useful and versatile atom with a valence state that allows for many complex interactions with other atoms. All organisms use nitrogen to make proteins. Some use ammonia as a fuel to gain energy. Others use nitrate for an electron receptor in the same way that humans use oxygen. Still others invest energy to take it out of the atmosphere for their own use. Even though nitrogen is presently about 80% of the atmosphere, plants cannot use it in this form, only after it is fixed (i.e., incorporated as nitrate or ammonia). In nature, this can be done only by the bacteria and blue-green algae. Thus, all members of all ecosystems are dependant on these two types of life for their nitrogen, or at least they were until about 1915, when humans learned how to fix nitrogen industrially through the Haber-Bosch process.

Evolution has found phosphorus to be an especially useful element, and hence it has a particular importance in ecosystems. Phosphorus is used by all organisms for their genetic material (DNA and RNA) and for energy-storage compounds (ATP), and in vertebrates for teeth and bones. It seems to be in a rather peculiar short supply in the world's biogeochemical cycles relative to the importance given it by evolution. Phosphorus tends to be especially limiting for the growth of plants in freshwater, and even where abundant it may be held very tightly by soils so as to be virtually unobtainable by plants. Phosphorus cycles rapidly between plants and the environment and is passed along food chains as one organism eats another (Fig. 3). When organisms die, they decompose, and we say that the nutrients such as phosphorus are mineralized (returned to their nonliving state). Phosphorus is easily dissolved in water, and in many respects the cycle of phosphorus at the ecosystem level is the cycle of water.

4.2 Redfield Ratios

It is not only the abundance of each nutrient that is important, but also their concentration relative to each other. The importance of this idea was first worked out by the oceanographer Alfred Redfield, who sailed around the world's seas on the first oceanographic ship, the *Challenger*, meanwhile taking samples of the oceans and its various life forms. Redfield's chemical analysis was a rather difficult undertaking, since the surface of most of the earth's seas are low in nutrients because the constant sinking of organisms and their feces tends to deplete nutrients in the surface of the sea. Nevertheless, Redfield was rather astonished to find that both the sea and the small plants of the sea had nearly constant ratios of nitrogen to phosphorus, about 15 atoms of nitrogen for each one of phosphorus, in both the living and the nonliving environments. Since we know that aquatic plants need roughly 15 atoms of N for each one of P, then we can determine whether the environment is N or P limited by whether it has more or less than this ratio. If the water is 20 to 1 it is likely to be P limited, and if it is 10 to 1 it is likely to be N limited. Likewise agricultural extension agents can analyze different soils and determine which nutrient is limiting, saving

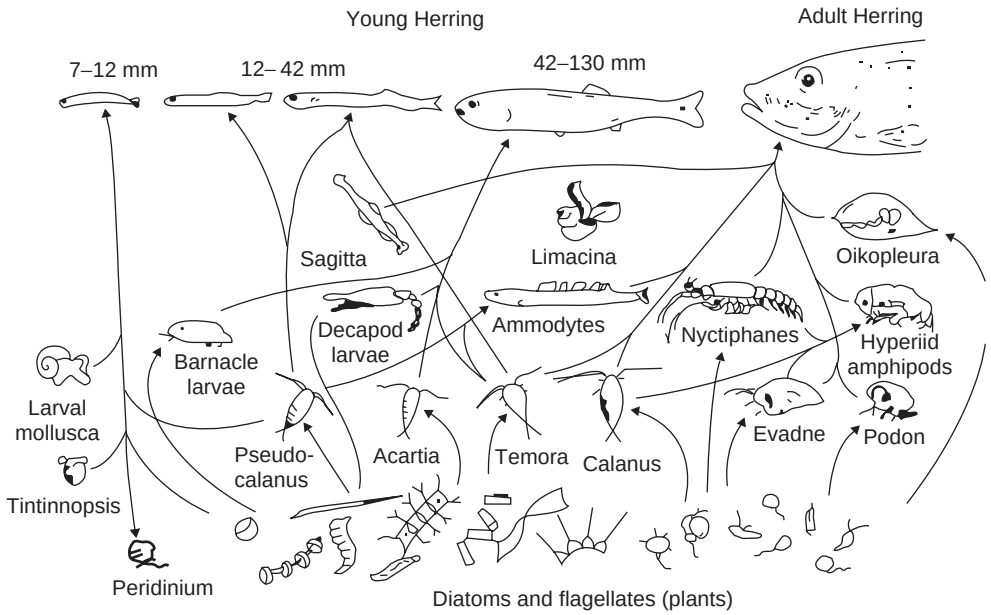


FIGURE 3 An early example of a food chain (from Wells *et al.*, 1939). In this case, the energy is represented as flowing from right to left, the opposite of later convention.

the expense and environmental problems of pouring on unneeded nutrients.

5. STABILITY AND DISTURBANCE

A question that has intrigued many ecologists is the degree to which ecosystems are stable—that is relatively homeostatic or self-correcting. It seems to many early investigators that natural ecosystems tended to have self-correcting properties, so if they were disturbed they would bounce back toward their original conditions.

5.1 Succession

Hence earlier ecological studies tended to emphasized the equilibrium nature of ecosystems. The idea was that there was a strong tendency for ecosystems to "evolve" toward a "climax" state determined by the climatic conditions of the site. In this view, succession was an orderly process that would bring the system back to climax, a condition of stability with no more species change. Climax was considered the normal state of ecosystems, and considerable virtue was associated with this process, a sort of "mother nature knows best." The view is evolving that ecosystems are continuously subjected to disturbance, and are continuously responding in various ways to the various disturbances they are subject to. For example, tropical ecosystems were once thought particularly unchanging. Research has shown that many tropical forested ecosystems are continuously in disequilibrium—that is, recovering in some way from some externally imposed stress or pulse, such as a hurricane, a landslide, a drought or, locally, a tree fall.

5.2 The Role of Species Diversity

An unresolved issue in ecology is to what degree all of the different species found in an ecosystem are required for it's proper functioning. Twenty-five years ago, the idea that diversity (of species) and stability (of populations and ecosystems) were positively connected was very popular. At this time, the best that we can say is that certainly an ecosystem needs many species to function properly and that through coevolution many species are tightly linked to each other. But we do no know in a general sense whether ecosystems *need* all, or even most, of the plethora of species that are found there in order to continue their basic structure and function. At the same time, we are finding an increasing number of examples where animals once thought rather superfluous with respect to ecosystem function are in fact critical for some important function such as distributing or germinating the seeds of a dominant

plant. Species that have a disproportionate effect on the entire ecosystem are called Keystone species. An example is the sea otter of California-to-Alaska coastal waters, which maintains kelp beds by eating sea urchins that would other wise overgraze the kelp. Additionally we are understanding better that ecosystems themselves can be much more resilient than many of their component species, since if one species is knocked out, another (although perhaps one less desirable from a human perspective) can sometimes take over that function. This is an important question as all over the world humans are reducing the diversity of ecosystems, and they are especially impacting the large animals.

5.3 Disturbance

An important aspect of ecosystems that we have learned is that although we once thought of ecosystems as, generally, stable and focused on the tendency toward equilibrium that occurred through the process of succession, we now view ecosystems as much more transient phenomena, generally subject to, and responding to, external forces, which we normally term disturbances. For example, graduate students are often taught that tropical forests were "stable," due principally to a lack of winter and glacial-cycle impacts. In reality, many tropical forests, for example the Luquillo experimental forest in Puerto Rico, are routinely subject to three type of natural disturbances: tree falls, landslides, and hurricanes, so that the forest is almost always responding to, or recovering from, disturbances.

Some important nutrient cycling studies were done at the U.S. Forest Service experimental watersheds at Hubbard Brook, New Hampshire. Here entire watersheds were cut, and vegetation regrowth suppressed, to examine the role of the vegetation in determining nutrient retention by ecosystems (it was substantial). Even so, the scientists there also found that long-term trends in rainfall intensity were equally important in determining the nature of the ecosystem. In other words, even in a relatively undisturbed forest there was a large amount of what we might call disturbance introduced simply by the normal decade-scale cycles of drought and flood, and this greatly interfered with the evolution of the ecosystem toward what we might have once called mature. Thus, at the level of the ecosystem it is clear that both endogenous and exogenous factors are important in determining the structure of the ecosystems and the nature of nutrient movements.

5.4 The Role of Human Disturbance

We previously discussed the importance of natural disturbance in determining the structure and function of existing ecosystems. Unfortunately, often it has been frustratingly difficult to unravel the impact of natural factors on ecosystem structure and function because almost all ecosystems of the world have been heavily impacted by prior human activity. For example, it was widely believed that New England forests were in some kind of pristine condition prior to the American Revolution, or even prior to European colonization. Instead we now know that the New England landscape had been modified severely by the activity of both Native Americans and very early colonists.

Likewise Perlin has documented the intensive deforestation that has taken place in virtually all portions of the world over the past thousands of years due to human activities. The basic cycle, repeated over and over again, has been that initially human settlers had relatively little impact on the forested lands except to clear small areas for agriculture. Often this land was abandoned after fertility declined and a new patch cut while earlier patches reverted to forests. This was not too different from the patterns of natural disturbance. But then, characteristically, some metal was discovered. Huge areas of forests were cut to provide the charcoal to smelt the metal, and the net effect was extensive deforestation. Subsequently there tended to be an increase of agricultural production on these newly cleared and fertile lands, followed by large human population increases, and then substantial soil erosion, river and harbor sedimentation, and eventual agricultural and population collapse. For example, ancient Greece was deforested 4000 years ago and then recovered, only to be again deforested, recover, and be deforested again in relatively modern times. The use of chemical fertilizers has allowed agriculture to be maintained over a longer time period, so that the land is not abandoned to reforestation. Curiously, some few parts of the world, notably the Eastern United States and Western Europe, are now undergoing reforestation because the exploitation of fossil fuels has allowed the forests a respite, albeit perhaps only temporarily.

There are other ways in which people have and continue to impact natural ecosystems. Humans have changed the relative abundance of species through the introduction of new species and through the exploitation of natural populations. For example, the introduction of goats into the Canary Islands 800 years ago

resulted in wholesale destruction of the natural vegetation, and the unique flora and fauna of islands such as the Galapagos and Hawaii are being changed dramatically by human-caused introductions, many of them centuries old. Curiously the flora and fauna of the island of Puerto Rico seem to be much less impacted by invasive species, and when the vegetation is allowed to recover from human disturbance, native species predominate.

One consequence of all of this is that it is difficult for ecologists to measure what is "natural" in nature, for almost all nature that we have had an opportunity to observe is already quite different from what might have existed some tens of thousands of years ago. Meanwhile, climate change, at both a local and global level, are impacting all organisms in every ecosystem, subjecting them to different temperature and rainfall regimes and to different carbon and nitrogen levels than those within which the organisms evolved. We are learning that ecosystems can adapt surprisingly well to at least some of these changes, but we do not know how long this will continue into the future.

6. THE FUTURE OF ECOSYSTEMS

The natural ecosystems of the earth are tremendously and increasingly impacted by human activities. For example, Vitousek and his colleagues have estimated that fully 40% of the earth's primary production is directly or indirectly exploited by our own species, and others believe this proportion is higher. This is remarkable considering that we are but one of millions of species on this planet, and that about two-thirds of the planet is water and as such is difficult for humans to access. People are probably not doing anything conceptually different from other organisms in nature, which also tend to eat or otherwise exploit whatever resources are available as much as possible. Natural selection almost certainly has hardwired us for lust and something resembling greed. The difference is that agricultural, medical, and especially fossil fuel technology have allowed people to be enormously more effective at exploiting resources compared to other species, or indeed compared to our earlier selves. At the same time, the earth's ability to check our own population levels through predation and disease has been greatly reduced. The net result is that the human impact has grown enormously.

The consequences of this tremendous rate of exploitation are only beginning to be understood. They are not necessarily bad from the perspective of ecosystems, at least if we measure good and bad in terms of their biomass and productivity. For example, Nemani and his colleagues assessed the impact of changing climates and nutrient cycles on global vegetation and found that there had been a probable increase in productivity of ecosystems in the temperate regions and a decrease in the tropics. But in a conference on the sources and sinks of carbon, Sampson and his colleagues concluded that human activity was turning what had been natural sinks of carbon for most of the world's ecosystems into carbon sources, probably exacerbating the greenhouse effect. Although various scientists argue whether or not human activity has impacted the earth's climate to date, that is not really the issue. We have changed the concentration of carbon dioxide in the atmosphere by only about 40% so far. If civilization continues along its present path, the change could be 500 to 1000%. Thus, natural ecosystems, and the services that they provide, are likely to continue to decrease in area and naturalness, since the human population continues to grow—in fact, to increase more rapidly than ever before—and most of the earth's human citizens have increasing desires for material goods, which can come only from the earth itself. Young people reading this article will almost undoubtedly see a disappearance of most of the world's natural ecosystems during their lifetimes unless there is an extraordinary effort to reduce human impact or unless that occurs involuntarily.

7. CONCLUSION

Ecosystem science got a relatively slow start in America because of an early species-oriented, stamp collecting attitude toward the large, fascinating, and largely unknown diversity of nature in the new world. In addition, there was competition for funding, attention, and prestige from other levels of ecological and biological inquiry, a largely false sense of the power of formal mathematical analysis that came into vogue in ecology, and the fact that ecosystem studies often were expensive. But in the 1950s, ecosystem studies began to prosper, propelled in part by new techniques for studying the energetics of whole aquatic systems and also by curiosity and funding related to how radioactive materials might move through ecosystems if released by accident or war. Important studies of grasslands, forests, and lakes in the 1960s quantified many basic aspects of

ecosystems and began the process of constructing ecosystem models. In the 1970s, an ecosystems approach to watersheds in New Hampshire (Hubbard Brook) and North Carolina (Coweeta) showed how an experimental approach could be used to address many basic aspects of understanding ecosystems, and Howard Odum began his analyses of humans as components of ecosystems. The importance of ecosystems-level studies is becoming more and more obvious, and ecosystem studies have changed from a relatively obscure part of biology to a healthy and vibrant science all of its own. For example, the U.S. Forest Service in the early 1990s changed the entire focus of its management for our national forests from one based on timber yield to one focused on managing the forests as ecosystems. Another important trend is that ecosystem approaches are moving from the temperate latitudes increasingly to other regions, such as the tropics and tundra. One next phase, barely started, is to use ecosystem approaches to study cities and the relation of cities to natural or seminatural systems. Perhaps our understanding of how natural systems have evolved to prosper in the circumstances of relatively limited and stable energy supplies can give us some useful information as humans, presumable, will have to adapt to a more energy-restricted future.

Acknowledgments

This article is based in part on Hall, C. A. S. (1999), Ecosystems, pp. 160–168, in D. E. Alexander and R. W. Fairbridge (Eds.), "Encyclopedia of Environmental Science." Kluwer Academic, Norwell, Massachusetts.

SEE ALSO THE FOLLOWING ARTICLES

Easter Island: Resource Depletion and Collapse • *Ecological Footprints and Energy* • *Ecological Risk Assessment Applied to Energy Development* • *Ecosystem Health: Energy Indicators* • *Environmental Gradients and Energy* • *Heterotrophic Energy Flows*

Further Reading

Brown, M. T., and Hall, C. A. S. (eds.). (in press). An H. T. Odum primer. "Ecological Modeling."

Craig, B. J. (2001). Eugene Odum. "Ecosystem Ecologist and Environmentalist." University of Georgia Press.

Golley, F. (1993). "A History of the Ecosystem Concept in Ecology: More Than the Sum of the Parts." Yale University Press, New Haven, CT.

Hagen, J. P. (1992). "An Entangled Bank: The Origins of Ecosystem ecology." Rutgers University Press, New Brunswick, NJ.

Hall, C. A. S. (ed.). (1995). "Maximum Power: The Ideas and Applications of H. T. Odum." University Press of Colorado, Niwot, CO.

Hall, C. A. S. (1988). An assessment of several of the historically most influential models used in ecology and the data used in their support. *Ecol. Modelling* **43**, 5–31.

Likens, G., and Borman, H. (1995). "Biogeochemistry of a Forested Ecosystem." 2nd ed. Springer Verlag, New York.

Lindeman, R. (1942). The trophic-dynamic aspect of ecology. *Ecology* **23**, 399–418.

Odum, E. P. (1962). Relationships between structure and function in the ecosystem. *Japanese J. Ecology* **12**(3), 108–118. Reprinted in Bobs Merrill College Division reprint series in the Life Sciences.

Odum, H. T. (1957). Trophic structure and productivity of Silver Springs, FL. *Ecological Monographs* **27**, 55–112.

Odum, H. T. (1988). Self-organization, transformity and information. *Science* **242**, 1132–1139.

Odum, H. T. (199X). Systems Ecology. Wiley Interscience. Reprinted as "General and Ecological Systems." University Press of Colorado, Niwot, CO.

Thomas, D. W., Blondel, J., Perret, P., Lambrechts, M. M., and Speakman, J. R. (2001). Energetic and fitness costs of mismatching resource supply and demand in seasonally breeding birds. *Science* **291**, 2598–2600.

Wiener, D. (1988). "Models of Nature: Ecology, Conservation and Cultural Revolution in Soviet Russia." Indiana University Press, Bloomington, IN.

Woodwell, G. M., and Pecan, E. V. (eds.). (1973). Carbon and the biosphere. Brookhaven Symposium in Biology, 24. National Technical Information Service. Springfield, VA.

Electrical Energy and Power

JOSEPH PRIEST
Miami University
Oxford, Ohio, United States

1. **Atomic Structure**
2. **Electricity**
3. **Electrical Resistance**
4. **Electric Power and Electric Energy**
5. **Magnetic Force and Magnetism**
6. **Electric Generators**
7. **Transformers**

Glossary

alternating current (AC) An electric current that alternates direction at regular intervals.
ampere (A) Unit of electric current. One ampere is a rate of flow of charge equal to 1 C/s ($I = Q/t$).
atom An atom consists of a dense, positively charged nucleus surrounded by a system of electrons equal in number to the number of protons in the nucleus. The atom is bound together by electric forces between the electrons and the nucleus.
atomic number The number of electrons (or protons) in a neutral atom.
coulomb (C) The metric unit of electric charge. In terms of the charge of the electron, the sum of the charges of approximately 6 1/4 billion-billion electrons equals 1 C.
direct current (DC) An electric current for which charges flow in one direction.
electric current Electric charges in motion. The ampere is the measuring unit.
electric field Qualitatively, a region of space in which an electric charge feels a force.
electric force A force between two objects that each have the physical property of charge.
electrical resistance The resistance offered by the structure of a conductor to the flow of electric charges.
hertz (Hz) Unit of frequency equal to one cycle per second. Named for Heinrich Hertz (1857–1894), a German physicist.
kilowatt-hour (kWh) A unit of energy equal to a power of 1 kilowatt (1000 W) acting for 1 h.
magnetic field Qualitatively, a region of space in which the pole of a magnet or a moving charge feels a force.
magnetic field line An imaginary line in a magnetic field along which a tiny compass needle aligns.
nucleus (of an atom) The positively charged central region of an atom. It is composed of neutrons and protons and contains nearly all the mass of the atom.
ohm Unit of electrical resistance. If a potential difference of 1 V across some electrical element causes a current of 1 A, the electrical resistance is 1 ohm ($R = V/I$).
Ohm's law If the electrical resistance of a device (e.g., a toaster element) does not depend on the current in the device, then it is said to obey Ohm's law.
potential difference The work required to move an amount of charge between two positions divided by the strength of the charge ($V = W/Q$). The volt is the measuring unit.
power Rate of doing work. The watt is the metric measuring unit.
transformer An electromagnetic device used to change the strength of an AC voltage.
volt (V) Unit of potential difference. If 1 J of work is required to move 1 C of charge between two positions, the potential difference between the positions is 1 V ($V = W/Q$).
voltage A common expression for potential difference.
watt (W) The metric unit of power. A rate of doing work of 1 J/s is a watt ($P = W/t$).

Electric charge is a property of matter in the same sense that mass is a property of matter. These two properties produce two distinct types of force. Two objects experience a mutual gravitational force because of the property of mass. We see the result of a gravitational force when a pencil is dropped and pulled toward Earth. An electric force is felt by each of two objects possessing the property of charge. Unlike the gravitational force, which always tends to pull masses toward each other, electric forces can tend to pull objects together or push objects apart. If the electric force is attractive, we say the charges acquired by the two objects are unlike. The charges are like when the two objects repel. Two unlike charges are distinguished by calling one positive and the other negative. A toy balloon acquires an electric charge when rubbed on one's hair. The charge on the

 © 2004 Elsevier Inc. All rights reserved.

balloon is negative. Any charge repelled by the balloon is negative; any charge attracted is positive.

It is easy to relate to the property of mass because mass is a measure of the quantity of matter. Charge is a much more subtle property because we cannot see charge; we see only the results of forces produced by charge.

In 1784, Augustin de Coulomb discovered that the force between two charges labeled q and Q and separated a distance r obeys a law of the form

$$F = k\,\frac{qQ}{r^2}. \tag{1}$$

In the metric system of units charges q and Q are measured in Coulombs (C), r is measured in meters, and the constant k has the value

$$k = 8.99 \times 10^9\,\frac{\text{newton meter}^2}{\text{coulumb}^2}.$$

Electric charge, electric force, and energy associated with the electric force play roles ranging from the structure of atoms to the liberation of energy in either a nuclear-fueled or coal-fired electric power plant.

1. ATOMIC STRUCTURE

The charge acquired by rubbing a balloon as well as electric current in a wire has its roots in atoms, which are the basic structural components of matter. An elementary model of an atom incorporates features of an Earth–satellite system. Electrons are satellites and a nucleus acts like Earth. An electron has a mass of 9.11×10^{-31} kg and a negative charge of -1.60×10^{-19} C. A nucleus is positively charged and is much more massive than an electron. The mass of the nucleus is confined to a volume approximating a sphere of radius of approximately 1×10^{-15} m. The closest electron is approximately 1×10^{-10} m. In perspective, if the radius of the nucleus were approximately the thickness of a dime, the nearest electron would be approximately the length of a football field away. Having both mass and electric charge, electrons and a nucleus experience gravitational and electrical forces that tend to bind the electrons to the nucleus. However, the overwhelming binding is due to the electric force.

In a normal atom, the total negative charge of the electrons balances the positive charge of the nucleus and the atom is said to be electrically neutral. The number of electrons in a neutral atom is called the atomic number, which distinguishes different atomic species. For example, a carbon atom has 6 electrons and an oxygen atom has 8 electrons. Most electrical wires in homes are made of copper. Each atom of copper has 29 electrons. As the atomic number increases, the mass of the atom also increases. This results from the additional electrons and by an increase in the mass of the nucleus.

2. ELECTRICITY

A current is defined as "a steady and smooth onward movement, as of water." A steady movement of cars on a highway would be a car current. The rate at which cars pass a point on the highway is a measure of the car current. To determine the car current, one would count the cars, measure the time for the cars to pass, and divide the number by the time to obtain the rate in units such as cars per minute. If each car had two passengers, then multiplying the car rate by two yields a passenger current measured in passengers per minute.

An electric current is a flow of electric charges. In wires and appliances the flow is due to electrons because they are much freer than the positive charges in the nucleus. Electrons moving through a wire are analogous to cars moving on a highway. One cannot see the electrons but in principle one can stand at the edge of the wire and count the electrons passing by in a measured time interval and express the rate in units of electrons per second. Multiplying the electron rate by the charge per electron yields the rate at which charge flows in units of coulombs per second. As an equation,

$$\text{Electric current} = \frac{\text{coulumbs of charge flowing through a wire}}{\text{time required for the flow}}$$

$$I = \frac{Q}{t}. \tag{2}$$

Recording charge (Q) in coulombs and the time interval (t) in seconds yields coulombs per second as the units of electric current (I). A coulomb per second is called an ampere (A). If 100 C of charge flow by some position in a wire in 10 s, the current is 10 A. The electric current in a lit flashlight bulb is approximately 1A. If the bulb were on for 10 min, the total charge passing through the bulb in 10 min is

$$1\ \text{C/s} \times 10\ \text{min} \times 60\ \text{s/min} = 600\ \text{C}.$$

A force must be exerted on the electrons to cause them to flow through a wire just as a force must be applied to water to keep it moving through a pipe. The electric force has its origin in the force between

electric charges. Electrons pulled toward a positive charge create an electric current. Electric forces do work on the electrons just as work is done on an automobile pushed some distance by a person. It is the role of a battery in a flashlight or an electric generator in an electric power plant to supply electrical "pressure" for moving electrons. This electrical pressure is called a potential difference (voltage is a less formal word describing the same idea). A potential difference is analogous to a water pressure difference for moving water through a pipe. Potential difference is defined in terms of the amount of work done by electric forces on a charge as it moves between two positions:

$$\text{Potential difference} = \frac{\text{electrical work on the amount of charge moved}}{\text{amount of charge}}$$

$$V = \frac{W}{Q}. \tag{3}$$

Work (W) is measured in joules (J) and charge (Q) is measured in coulombs, making potential difference (V) joules per coulomb. A joule per coulomb is called a volt (V). If 100 J of work is done by electric forces on 1 C of charge when it moves between two positions, the potential difference between the two positions is 100 V. Batteries for most automobiles provide a 12-V potential difference between two terminals. If 2 C of charge flow between the terminals, electrical forces do 24 J of work on the 2 C of charge.

The notion of a difference in potential is very important. If a building is said to be 500 ft high, this probably means the difference in "height" between the top and bottom is 500 ft. Clearly, however, if the building were on top of a mountain, its height would be more than 500 ft relative to the base of the mountain. Potential differences, like heights, are also relative. The potential difference between the terminals in a household electrical outlet is 115 V. This means that one of the contacts is 115 V relative to the other contact. This other contact in a household outlet is at the same potential as any other part of the room, including a person in the room. You can touch this contact and not get an electrical shock because there is no potential difference between you and the contact. This contact is called ground because a wire connects it to a metal stake driven into the ground. Generally, this ground wire can be seen alongside an electric utility pole holding wires that bring electricity to a house. The other contact in an outlet is referred to as "hot." Electrical shocks occur when you touch both ground and the hot contact.

All electrical appliances have two electrical connections. To operate an appliance, an appropriate potential difference must be connected to these two connections. If two wires were connected to a flashlight bulb and the other ends of the wires touched to the terminals of a flashlight battery, the bulb would light. The lit condition signals a complete electrical circuit. It is complete in the sense that a complete (or closed) path is provided for the flow of electrons. Cutting one wire with a pair of scissors breaks the circuit and the light no longer glows. A switch is a device that allows opening and closing electrical circuits. A switch is analogous to a faucet that can be opened or closed to control the flow of water. An example of a water circuit is shown in Fig. 1. A pump pulls water from the container, forces it through a valve, past a water wheel, and back into the container. In a household electrical circuit, an

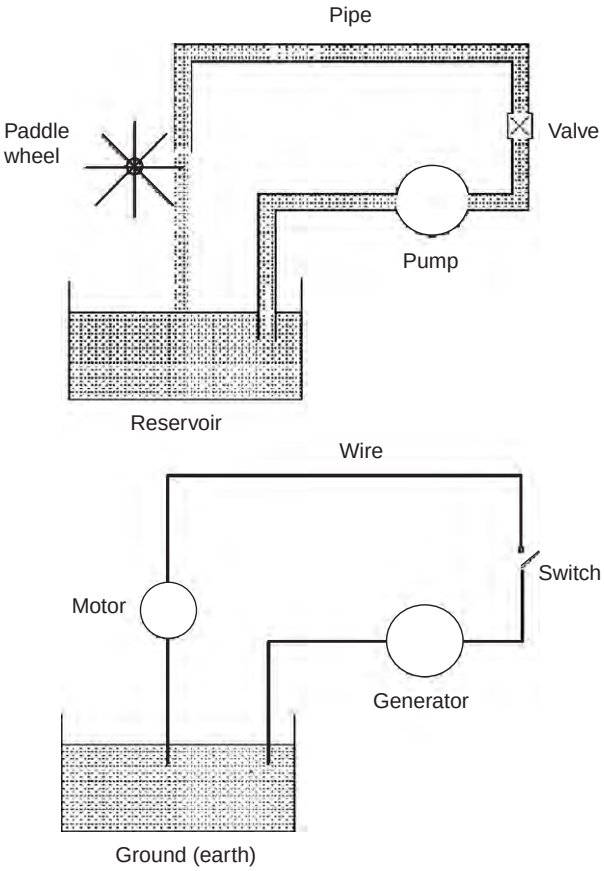


FIGURE 1 Schematic comparison of a fluid circuit and an electrical circuit. The water pump forces water to circulate through the water circuit. The electrical pump, a battery or generator, forces electrons to circulate through the electrical circuit. The sum total of electric charges never changes in the process, just as in the absence of leaks the sum total of water never changes in the water circuit.

electric generator provided by an electric utility pulls electrons from the ground, forces them through a switch, through a motor, for example, and back into the ground. The wire plays the role of a water pipe.

3. ELECTRICAL RESISTANCE

Electrons moving through a wire migrate through a network of atoms. This network impedes the flow, and we say the wire has electrical resistance. The resistance depends on the composition of the wire. For example, an aluminum wire having the same length and diameter as a copper wire offers more resistance than the copper wire. Just as water flows easier through short pipes than through long pipes, electrons flow easier through short wires than through long wires. Also, just as water flows easier through large-diameter pipes than through small-diameter pipes, electrons flow easier through large-diameter wires than through small-diameter wires. Accordingly, the resistance of a wire increases as the length increases but decreases as the diameter increases.

A current is produced in a wire when the ends are connected to a potential difference. The current increases if the potential difference (the electrical pressure) increases. If the resistance increases, the current decreases. This reasoning is modeled as an equation by writing

$$\text{Current} = \frac{\text{potential difference}}{\text{resistance}}$$

$$I = \frac{V}{R}. \tag{4}$$

Resistance has units of volts divided by amperes. This unit is called an ohm. If 1.5 A is produced in a bulb connected to a 12-V battery, the resistance of the bulb is

$$R = V/I = 12\,\text{V}/1.5\,\text{A} = 8\ \text{ohms}.$$

The resistance need not be constant. Doubling the potential difference across the bulb may produce a current of 2 A and a resistance of 6 ohms. If the resistance does not depend on the potential difference, the device obeys Ohm's law. Wires used for electrical conductors and heating elements in toasters and irons obey Ohm's law. Transistors do not obey Ohm's law.

In a household in which the potential difference is fixed at 115 V, the current in an appliance depends on its resistance. As the resistance decreases, the current increases. The energy used to overcome this electrical resistance appears as heat. For example, the heat liberated in a toaster is produced by electrons overcoming the electrical resistance provided by the metallic elements of the toaster.

4. ELECTRIC POWER AND ELECTRIC ENERGY

Like water pulled over a dam by a gravitational force, electrons moved by an electric force acquire kinetic energy as a result of work done on them. Electrons convert this energy in a variety of ways. In a lamp, the energy is converted to heat and light. In an electric motor, the energy is converted to mechanical rotational energy. The rate at which moving charges convert energy is called electric power:

$$\text{Power} = \frac{\text{energy converted in joules}}{\text{time required in seconds}}$$

$$P = \frac{E}{t}. \tag{5}$$

Electric power is expressed in joules per second, or watts. Electric clothes dryers use electric power in the range of a few thousands of watts; electric power plants produce power in the range of hundreds of millions of watts. Thousands of watts and millions of watts are usually expressed as kilowatts (kW) and megawatts (MW), respectively.

A useful relation for electric power follows from combining the equation for the definition of power, $P = W/t$, and the definition of potential difference, $V = W/Q$. Substituting for W in the equation $P = W/t$, we have $P = V(Q/t)$. Recognizing that Q/t is current I, we find $P = VI$. Power equals voltage times current is a general relationship for any electrical device. For a resistance, we can use $V = IR$ to arrive at $P = I^2R$ and $P = V^2/R$.

Companies that provide the voltage for forcing electrons through our appliances are often referred to as power companies. It is reasonable to expect that the companies are paid for power. However, they are actually paid for energy, not power.

Power is the rate of converting energy. A 100-W light bulb requires energy at a rate of 100 J/s regardless of how long it is on. However, the cost of operation depends on how long the bulb is lit. Using Eq. (5), it follows that energy, power, and time are related by

$$\text{Energy} = \text{power} \times \text{time}$$

$$E = Pt.$$

It follows that any unit of power multiplied by a unit of time yields a unit of energy. Watts (or kilowatts)

TABLE I

Power Requirements of Common Household Appliances

Appliance	Power (W)
Cooking range (full operation)	12,000
Heat pump	12,000
Clothes dryer	5,000
Oven	3,200
Water heater	2,500
Air conditioner (window)	1,600
Microwave oven	1,500
Broiler	1,400
Hot plate	1,250
Frying pan	1,200
Toaster	1,100
Hand iron	1,000
Electric space heater	1,000
Hair dryer	1,000
Clothes washer	500
Television (color)	330
Food mixer	130
Hi-fi stereo	100
Radio	70
Razor	14
Toothbrush	7
Clock	2

and hours are units of power and time. The watt-hour (or kilowatt-hour) is a unit of energy. It is peculiar because the units of time (hours) do not cancel the time units of seconds in power.

Table I lists the power requirements for several household appliances. Note that those involving the generation or removal of heat require the most power. These devices tend to have low resistance.

5. MAGNETIC FORCE AND MAGNETISM

We have discussed electric current and the idea that an electrical pressure is required to establish electric current. Nothing has been said about how the power company generates the pressure or how it is transmitted to homes and factories. To understand this important aspect of producing electricity, we must discuss some fundamentals of magnetic force and magnetism.

Images on a TV are produced by electrons impacting with the screen. The picture distorts if a magnet is placed near the face of a black-and-white TV. The distortion results from an interaction of the moving electrons with a magnetic field produced by the magnet. As a general principle, any moving charge "cutting" magnetic field lines experiences a magnetic force. In Fig. 2, magnetic field lines are directed from the N pole to the S pole. A wire moved vertically in the magnetic field "cuts" the magnetic field lines and all charges in the wire experience a force by virtue of moving in a magnetic field. However, only the free electrons in the wire acquire motion. If the ends of the wire are connected to a light bulb, for example, an electric current is produced in the circuit. Had the wire been moved horizontally parallel to the magnetic field lines, there would have been no current in the circuit. Work done by the agent moving the wire is responsible for the energy acquired by the electrons. This is the principle of an electric generator such as one might find on a bicycle or in an electric power plant. On a bicycle generator, the agent moving the wire is a wheel rotating wires mounted on the shaft of the generator. A large steam turbine is the agent in an electric power plant.

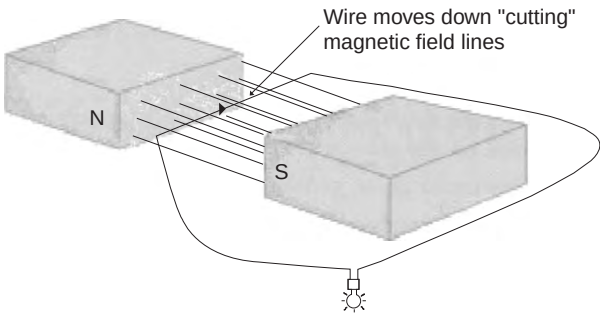


FIGURE 2 A current is established in a wire when it moves in a magnetic field so as to "cut" the magnetic field lines. If the wire is moved directly from the N to the S pole, no magnetic field lines are cut and no current results. The arrow indicates the direction in which electrons move in the wire.

The generator principle was illustrated by holding a magnet fixed and moving the wire in the magnetic field. A current would also be produced if the wire were held fixed and the magnet moved. As long as the wire cuts the magnetic field lines, a current results. If the wire were moved in a direction opposite to that shown in Fig. 2, the electric current changes direction. Cutting magnetic field lines with an oscillatory motion produces a current whose direction alternates. Such a current is called an alternating current (AC). The current produced by a battery is called direct current because the charge flows in one direction only.

6. ELECTRIC GENERATORS

A practical electric generator contains a coil of wire that is rotated in a magnetic field. To understand the operation, we examine a single loop as shown in Fig. 3. Metallic rings, called slip rings, are mounted on the shaft of the rotating loop and the ends of the loop are bonded to the slip rings. Metal contacts, called brushes, rub against the slip rings. The loop, slip rings, brushes, and connected electrical device form an electrical circuit for a current. The magnet produces a magnetic field whose field lines are directed from the N pole to the S pole. Only the sides of the loop labeled A and B are able to cut magnetic field lines. The two segments of the coil labeled A and B always move in opposite directions. Accordingly, the current is always in opposite directions in these two segments. The maximum current results when the wires move perpendicular to the magnetic field lines. There is no current in the circuit when the wire segments move parallel to the magnetic field lines. When the segment designated A in Fig. 3 is in the upmost position of its rotational path, it is moving horizontally to the right. Because the magnetic field lines are directed to the right, the wire is instantaneously moving parallel to the magnetic field lines. At this instant, there is no current in the circuit. As the loop rotates, the segments begin cutting the magnetic field lines and a current develops in the circuit. When the portion of the loop designated A has moved one-fourth of a turn down, instantaneously it moves perpendicular to the magnetic field lines. In this position, the electron current is a maximum. The electron current decreases to zero when the coil makes another one-fourth turn. As the coil rotates further, the electron current increases and changes direction. It reaches a maximum when the coil makes a third one-fourth turn. Finally, the electron current decreases to zero when the coil returns to the initial starting position. A plot of electron current versus position of the coil, or equivalently time since the position depends on time, is shown in Fig. 4. The cycle of current and time shown in Fig. 4 repeats with each revolution of the coil. The number of cycles produced each second is called the frequency. A cycle per second is called a hertz (Hz). In the United States, the frequency of the AC voltage produced by commercial electric power plants is 60 Hz.

When a toaster is plugged into an AC household outlet, the current produced in the wires of the toaster continually changes in magnitude and direction. Nevertheless, the current causes the toaster wires to warm. Otherwise, one would not be able to toast bread. Household voltage is usually labeled 120 V, even though the voltage is continually changing. The 120-V rating means that the voltage is as effective at producing heat as a 120-V battery connected to the toaster. The 120-V AC label on a light bulb is the effective value of an AC voltage needed to power the bulb.

A generator that provides electricity for a headlamp on a bicycle functions very much like the one illustrated in Fig. 3. Mechanical energy is provided by the bicycle wheel rubbing against and rotating a shaft connected to the coil of the generator. The magnetic field is provided by a magnet. This type of generator produces approximately 5 W of electric power. Although the generator in a commercial electric power plant is much larger in size and often produces

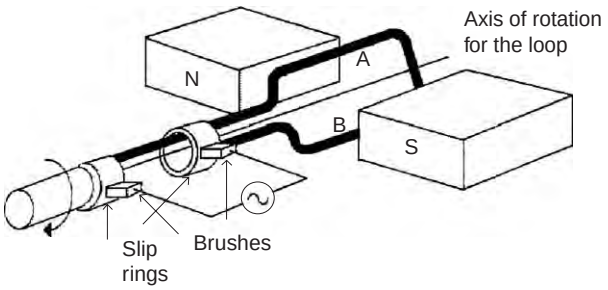


FIGURE 3 Principle of operation of an AC generator. A current develops in the rotating coil when it moves through the magnetic field created by the magnet. The slip rings rotate with the coil and make electrical contact with the brushes that are connected to an electrical device, such as a light bulb or a motor.

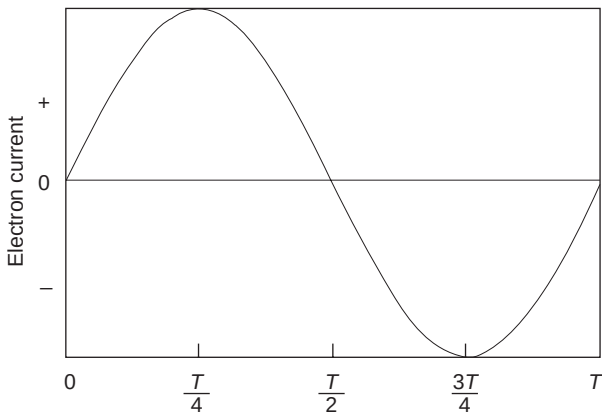


FIGURE 4 Plot of induced current in the rotating coil of an AC generator. Positive and negative currents are used to distinguish the two directions of charge flow. $T/4$, $T/2$, $3T/4$, and T denote times for one-fourth of a rotation, one-half of a rotation, three-fourths of a rotation, and a full rotation of the coil, respectively.

1 billion watts of electric power, the basic physical principle is the same. The commercial generator uses a steam turbine for the input mechanical energy and a massive stationary coil with a circulating electric current to provide the magnetic field.

7. TRANSFORMERS

An AC generator and a battery both deliver energy to such things as light bulbs connected to them. If a light bulb requires 12 V and one has a single battery that can provide 2 V, it is difficult to use the battery to power a device that can change the voltage from 2 to 12 V. If AC voltages are involved, it is reasonably easy to tailor voltages using an apparatus called a transformer. Transformers play very important roles in the transmission of electric power and in adapting voltages for industrial and household uses. The transformer principle is straightforward.

An electric current produces a magnetic field proportional to the current. If the current is constant, as would be the case if a battery were connected to the ends of a wire, the magnetic field is constant. The direction of the magnetic field lines depends on the direction of the current. If the battery is replaced by an AC source, the magnitude of the magnetic field changes as the current changes and the direction of the magnetic field lines changes when the current changes direction. There exists in the vicinity of the wire a changing magnetic field. If a coil of wire is placed in this changing magnetic field, an alternating current develops in it because the magnetic field lines as they change cut the wires of the coil. It is much like changing the magnetic field lines by moving a magnet. This principle is used in transformers to change the size of an AC voltage. The small black box often connected to an electrical outlet and to a portable computer to recharge the computer's battery is a transformer that typically reduces the outlet voltage from 115 to 20 V.

A transformer has two distinct coils of wire wound onto an iron core (Fig. 5). One of the coils, the primary, is connected to an AC voltage such as at the outlet in a house. The other coil, the secondary, is connected to an electrical device such as a light bulb. The AC current in the primary coil produces magnetic field lines that are guided into the secondary coil by the iron core. The interception of the magnetic field lines by the secondary coil produces a current in the secondary coil. The ratio of the primary and secondary voltages is equal to the ratio of the number of turns of wire in the primary and secondary coils of the transformer:

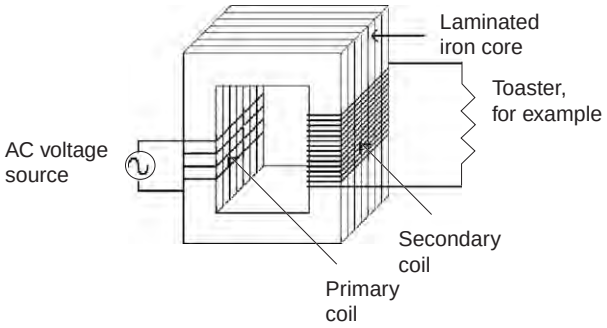


FIGURE 5 Essentially, a transformer consists of two coils of wire (primary and secondary) and an iron core to guide the magnetic field lines from the primary coil to the secondary coil.

$$V_s/V_p = N_s/N_p. \tag{6}$$

Door chimes in houses often require 6 V for operation. Suppose that the voltage available at an outlet is 120 V. If a transformer has 20 times more turns on the primary than on the secondary so that $N_s/N_p = 1/20$, then the secondary voltage will be 6 V when the primary voltage is 120 V. The primary is connected to the available 120 V and the secondary is connected to the chimes.

The secondary of a transformer delivers electric power to whatever is attached to it. This electric power has its origin in the electrical source connected to the primary. The principle of conservation of energy prohibits obtaining more power in the secondary than was supplied in the primary. Transformers tend to warm, indicating that some of the power supplied in the primary appears as heat so that the secondary power is always less than that in the primary. In a well-designed transformer, the efficiency for energy transfer is approximately 99%. The ability of a transformer to tailor voltages and its very high efficiency makes it an extremely useful device.

SEE ALSO THE FOLLOWING ARTICLES

Batteries, Overview • Electricity Use, History of • Electric Motors • Electric Power Measurements and Variables • Electric Power Systems Engineering • Electromagnetism • Mechanical Energy

Further Reading

Hobson, A. (2002). "Physics: Concepts and Connections." 3rd ed. Prentice Hall, Englewood Cliffs, NJ.
Priest, J. (2000). "Energy: Principles, Problems, Alternatives." 5th ed. Kendall/Hunt, Dubuque, IA.
Serway, R. A., and Beichner, R. J. (2000). "Physics for Scientists and Engineers." Brooks/Cole, Pacific Grove, CA.
Serway, R. A., and Faughn, J. S. (1999). "College Physics." Brooks/Cole, Pacific Grove, CA.

Electricity, Environmental Impacts of

ALEXANDER E. FARRELL
University of California, Berkeley
Berkeley, California, United States

1. Activities
2. Impacts

Glossary

coal combustion products (**CCP**) Solid waste left after burning the carbon and hydrocarbons in coal; made up of flyash, bottom ash, slag, and sulfur dioxide scrubber sludge. CCP are often used as construction materials.

direct cooling The transfer of waste heat from power plants to water bodies using water-to-water heat exchangers. This approach does not consume significant amounts of water (most is returned) but can cause significant impacts on aquatic species due to contact with screens or pumps.

electromagnetic field The combination of electric and magnetic fields that surround moving electrical charges (e.g., electrons), such as those in electric currents. Electromagnetic fields apply a force on other charges and can induce current flows in nearby conductors.

indirect cooling The use of a closed loop water system to reject waste heat to the atmosphere. A significant amount of water is evaporated in this process, which must be replaced, but total flows are less than with direct cooling, and water bodies are not heated.

radionuclide An isotope of an element that is radioactive, that is, a radionuclide undergoes spontaneous fission (splitting), releasing energy when it does so. The energy released from radionuclides can cause various health effects, including increased risk of cancer.

submergence Permanent flooding of land behind (upstream from) a dam.

Electric power is simultaneously one of the cleanest energy technologies and one of the dirtiest; it has virtually no environmental impacts at the point of use, but generation and transmission create enormous impacts. This article provides an overview of the environmental impacts of electricity, the most important of which are air pollution, water use, solid waste, and land use.

1. ACTIVITIES

Most of the environmental impacts of electricity are associated with generation and transmission, including equipment manufacture and construction, fuel supply, fuel combustion, and waste disposal. Currently, electric power systems rely on fossil (coal, oil, and gas) powered generation for to ensure adequate electricity is always available and to follow load; but if fossil generation is limited in the future, other technologies will be needed. If so, the environmental impact of the entire system, including the manufacture and use of new technologies such as energy storage, must be considered. Here, only direct effects are discussed.

1.1 Generation

Enormous quantities of fossil fuels are used to generate electricity, so even relatively minor contaminants found in them (especially in coal) lead to very significant emissions of pollutants such as sulfur dioxide (SO_2), nitrogen oxides (NO_x), carbon dioxide (CO_2), toxics such as mercury (Hg) and hydrogen chloride (HCl), and radionuclides such as uranium (U) and thorium (Th). If electricity generation occurs near the point of use, emissions of carbon monoxide (CO) may be an issue.

Another major environmental impact of fossil and nuclear fuel electricity generation is water use, which includes consumption, impacts on aquatic species, and discharge of heat into nearby water bodies. Electricity generation by coal and nuclear plants also creates significant solid wastes that must be dealt

 © 2004 Elsevier Inc. All rights reserved.

with. The majority of the environmental impacts of nuclear technologies occurs before and after generation, especially in fuel production and spent fuel disposal. Nuclear power plants are also vulnerable to accidental release of radionuclides.

Aesthetic concerns are also important; many people find electricity generation and transmission equipment unpleasant to be near or even to look at. While aesthetic concerns may seem relatively unimportant compared to health and environmental quality issues, they can be more difficult to resolve than pollution (which generally has a technical solution).

Renewable electricity generation technologies usually have quite different impacts; they typically create little or no emissions or waste during generation, but their manufacture may have important environmental impacts and they often require the use of large land areas to collect relatively diffuse renewable energy resources. This is especially true of hydroelectric power, which can require flooding vast areas. Fuel cells may also fall into this category because while they emit only water, producing the fuel cells and manufacturing and delivering hydrogen fuel may have significant environmental impacts. In this entry, however, only the direct impacts of operation are considered.

1.2 Transmission, Distribution, and Storage

The primary environmental impacts of transmission and distribution systems are associated with land use and aesthetics, but their scales are quite different. Distribution systems are ubiquitous in developed areas where they affect land use little (if at all), and are largely ignored by the public. In highly urbanized areas, electricity distribution infrastructure may be largely underground. The public often strongly opposes proposals to build transmission lines, especially when they will be close to population centers or highly visible. One reason for this opposition is concern about possible health effects of electromagnetic fields (EMFs), although it is very uncertain what (if any) risk EMFs create. On the other hand, transmission systems reduce the amount of generation capacity that's needed for a reliable system and allows for power plants and their emissions to be located far from populated areas, which tends to decrease their health and aesthetic effects.

Renewable energy sources are often located far from demand centers (cities and factories), extensive transmission systems may be needed to deliver renewable power from where it's available to where it is needed. In addition, wind and solar energy are intermittent (they are not always available—such as on a calm night), so if they are to supply electricity whenever consumers need it, these technologies need some sort of energy storage. Nuclear power plants may rely on storage in the future as well; it can be hard to adjust reactor power up and down rapidly (over the course of a few hours) to follow demand.

Energy storage is extremely challenging, only two practical methods now exist—pumped storage and compressed air. Pumped storage facilities are like hydroelectric dams; energy is stored by pumping water from a lower reservoir to a higher one, and withdrawn by letting it flow back down through turbine generators. At compressed air energy storage facilities, electrically driven compressors fill caves or other underground formations with pressurized air to store energy, which is released later through a turbine. Neither of these technologies is widely used, and the environmental impacts of significant deployments of them are not well known.

1.3 End Use

There are essentially no environmental impacts of electricity in its end-use phase. Thus, as electric technologies replaced end-use fuel consumption for heating, cooking, lighting, and mechanical work, many environmental problems disappeared at the cost of introducing environmental impacts from electricity generation and transmission. This has helped homes and workplaces in many countries around the world become much cleaner and safer since the start of the 20th century, vastly improving public health and environmental quality. The only exception is the potential risk of EMF effects from household and workplace exposures, especially appliances.

2. IMPACTS

2.1 Air Pollution

2.1.1 Gases: SO_2, NO_x, and CO_2

Electricity generation causes a significant fraction of anthropogenic emissions of three major air pollutants, CO_2, NO_x, and SO_2, although the fraction varies significantly around the world, as shown in Table I. Anthropogenic emissions of SO_2 are relatively large compared to natural emissions of sulfur into the atmosphere, and are highly concentrated in the northern hemisphere. Electricity sector

TABLE I

Annual Emissions of Major Gaseous Pollutants from Global Electricity Generation

Pollutant	CO_2		NO_x		SO_2	
Region	Mass (Tg-CO_2)	Percentage of anthropogenic	Mass (Tg-NO_2)	Percentage of anthropogenic	Mass (Tg-SO_2)	Percentage of anthropogenic
Canada	103	20	0.3	8	1.5	53
United States	2045	37	6.2	29	11.0	61
OECD Europe	966	26	2.1	15	5.0	30
Oceania	138	37	0.5	21	0.7	38
Japan	377	29	0.3	9	0.2	8
Eastern Europe	448	45	1.0	37	5.8	55
Former USSR	1278	41	4.4	45	6.9	43
Latin America	185	9	0.6	6	1.8	18
Africa	246	20	1.0	9	1.7	25
Mid-East	258	22	0.7	14	1.8	31
South Asia	410	34	1.7	25	2.5	33
East. Asia	1139	25	4.6	29	12.0	30
Southeast Asia	150	12	0.5	10	1.3	34
Total	7744	29%	23.8	21%	52.1	37%
Percentage of natural flows to the atmosphere	2		6		80	

Source. Oliver and Berdowski (2001).

SO_2 emissions are expected to grow in the future and become a larger fraction of anthropogenic emissions due to more use of coal for power production, compared to flat or declining uses in other sectors. In contrast, electricity sector NO_x emissions are relatively small compared to anthropogenic and natural NO_x emissions. Power sector NO_x emissions can have important effects because they are often concentrated near populated areas and because the nitrogen cycle can be significantly altered by relatively small perturbations. In the future, NO_x emissions from electricity generation are expected to grow, but not as fast as other sources, especially transportation.

The environmental effects of these emissions from power plants are diverse, including acidification, reductions in visibility (also called haze), and eutrophication. Acidification is a change in an ecosystem's natural chemical balance caused by an increase in the concentration of acidic elements and can lead to severe damage to trees and the elimination of fish species from lakes and streams. Eutrophication is excessive plant growth due to the addition of compounds like NO_x, which leads to noxious algae and loss of fish and other aquatic wildlife. Visible haze is increasingly viewed as an important issue in parks and other natural areas, and it is driving some regulatory processes to control power plant emissions.

The most important human health impacts of SO_2 and NO_x emissions are their contributions to urban air pollution. In photochemical smogs, increased concentrations of ozone, polycyclic aromatic hydrocarbons, and fine particles (less than 2.5 micron diameter) can lead to increased morbidity and mortality. These compounds have many sources, but anthropogenic NO_x is among the most important.

Emissions of CO_2 from the electricity sector are not due to fuel contamination, CO_2 is the intended result of the combustion of fossil fuels. Unfortunately, it is also the primary cause of climate change.

2.1.2 Smoke and Particulate Matter

One of the main environmental impacts that electricity has helped lessen over the past century is smoke in urbanized and industrial areas. Smoke is a combination of unburned carbonaceous material (both solid and gaseous), nonburnable ash, sulfur

dioxide, nitrogen oxides, and trace amounts of metals. In the early 20th century, cities like Pittsburgh and London could experience dustfall of 10 to 30 g/m^2 each month. This material was certainly a nuisance and possibly a health hazard as well. Electricity replaced industrial boilers for mechanical work and in some cases as a heat source, reducing some of the smoke burden.

At first, however, electric power plants simply changed the location and character of smoke, or as it has come to be known particulate matter (PM) emissions. Early power plants had little or no PM controls and some power plants in developing countries lack them. In the 1960s, power plants in North America and Western Europe began to be fitted with PM control equipment and now virtually all coal-fired power plants in industrialized countries capture 99% or more of fly ash. In the United States, electricity generation accounts for about 1% of anthropogenic PM emissions of 10 microns diameter or greater (PM-10). In contrast, electricity generation contributes between 15 and 45% of PM emissions in some cities in India.

The greatest health risks of PM are due to the smallest particles, those 2.5 microns and smaller (PM-2.5). These fine particles also scatter visible light, causing haze. PM-2.5 can be emitted directly but are also formed in the atmosphere from gaseous emissions, especially SO_2 and NO_x. Electricity generation contributes similar fractions of direct PM-2.5 as it does PM-10, but it is a much larger contributor to global SO_2 and NO_x emissions and therefore to secondary PM-2.5 in the atmosphere. However, fine particles may have different health risks depending on their chemical composition, so the importance of electricity sector emissions to the health impacts of PM-2.5 pollution is not completely clear.

2.1.3 Toxics

Because many coals contain toxic contaminants and almost two-thirds of global coal production goes to power electricity generation, the sector has substantial toxic emissions. Petroleum products and to a much lesser extent natural gas, contain some toxic contaminants as well. These pollutants are persistent (i.e., they do not degrade) and some of them, notably mercury (Hg), bio-accumulate. The most important compounds that coal-fired power plants may emit include Hg, hydrogen chloride (HCl), sulfuric acid (H_2SO_4), and hydrogen fluoride (HF). These materials are can cause health and environmental effects at lower concentrations than more conventional pollutants like NO_x, and can be a significant local problem. Table II contains approximate annual global emissions of metals from worldwide electricity production and shows how these emissions compare to global anthropogenic emissions and of global natural emissions. Natural emissions of these metals are very small, so electricity generation creates very large perturbations to natural flows.

Geothermal resources (hot water and steam) can contain significant amounts of some of these toxics, as well as the toxic gas hydrogen sulfide (H_2S). In installations with very small emissions these may be vented to the air, in larger facilities they are reinjected into the ground along with most of the water or steam that has been withdrawn, and so they are. However, the contribution of geothermal electricity to all toxics emissions is trivial compared to that of fossil fuels.

Except for Hg and Se, most metals in coal tend to be captured by the fly ash, and their ultimate fate corresponds to that of the ash from the power plant. Many of these metals can also be captured in SO_2 scrubbers, and if so, their environmental impact is then largely determined by the final disposition of scrubber sludge. However, up to two-thirds of the Hg that enters a power plant in the coal remains a gas and escapes to the atmosphere. Controlling Hg emissions may include the conversion of gaseous Hg in power plant exhaust into compounds such as HgCl that can be more easily captured. Another

TABLE II

Approximate Annual Emissions of Metals from Global Electricity Generation

	Cr	Cu	Hg	Ni	Pb	Sb	Se
Emissions (Tg/year)	0.013	0.008	0.0011	0.042	0.013	0.0013	0.0038
Percentage of anthropogenic	40	20	50	80	15[a]	40	60
Percentage of natural flows to the atmosphere	30	30	40	150	110	50	40

Sources. Nriagu (1994) and Pacyna and Pacyna (2002).
[a]Not including leaded gasoline.

strategy may be to inject a sorbent (such as activated carbon) into the exhaust gas to capture the Hg and allow the particle control device to remove it.

Probably the most important public health impact of toxic emissions from electricity generation is neurotoxicity risk to unborn children of Hg. The primary exposure route is through consumption of fish by pregnant women, since aquatic bacteria create the most harmful form (methlymercury), which then accumulates in edible fish such as bass and perch. By 2002, Hg contamination had caused almost all U.S. states to warn women of child-bearing age to limit consumption of some fish from some water bodies. Wildlife may also be at risk due to Hg exposure, especially fish-eating birds and mammals.

As a liquid at typical temperatures, Hg can be transported very long distances through deposition and revolatization. Thus, Hg is a global pollutant, and for many countries more than one-third of Hg deposition originates outside of their borders. The hemispheric temperature gradient drives deposition toward the poles, although most Hg emissions occur in northern midlatitudes, health and ecological effects in the Artic may be most severe.

2.1.4 Radionuclides

Radionuclides present a human health risk by increasing the risk of cancer, genetic changes, and birth defects, and present a similar risk to wildlife. Globally, the background radiation dose is typically about 2.5 milli-Sievert (mSv) per year, of which about one-third is from cosmic radiation and two-thirds from naturally occurring radionuclides, mostly uranium (U), thorium (Th), and especially potassium (K). Medical equipment can add measurably to this amount (for instance, 0.5 mSv/yr. in the United States, on average), as can lifestyle choices (e.g., flying in airplanes frequently). Nuclear power plants are often designed to give a dose of no more than about 0.05 mSv/year (about 2% of background) to the most exposed nonworker group.

Radionuclide contaminants in coal used to generate electricity total about 1.5×10^{14} disintegrations per second (or Becquerels, Bq) each year. The isotopes most important isotopes (of U, Th, and K) are solids by the time they leave the combustion chamber, so radionuclide emissions and exposure from coal generation are dependent on PM controls and solid waste disposal. Significant direct contact with ash or poor disposal techniques may create a slight hazard, but radionuclide concentrations in most coal ash are lower than in many soils and building materials (e.g., limestone). Interestingly, the total radiation doses due to coal-fired electricity and due to nuclear electricity are virtually identical under normal operating conditions (unless the fuel is reprocessed, in which chase the nuclear electricity dose more than doubles).

Nuclear power presents the unique problem of potential accidental release of radionuclides. Operating reactors contain significant amounts of radioactive Iodine (I-131), which tends to be absorbed readily by the body and concentrated in the thyroid, increasing the effective dose to that gland. This radionuclide has a short half-life (about 8 days) and so is essentially eliminated as a risk 2 months after a release. Nonetheless, I-131 can cause sufficient genetic damage during this short period to induce thyroid cancers later. Iodine tablets are an inexpensive way to minimize exposure in case of accidental release.

Almost a dozen nuclear reactors have suffered meltdowns since the mid-1940s, in addition to fires and other accidents. Most of these accidents were associated with experimental or weapons reactors, but two large electric power reactors have had catastrophic meltdowns, Three Mile Island Unit 2 (TMI-2) in 1979 in the United States and Chernobyl RBMK Unit 4 in 1986 in the Ukraine (which was part of the Soviet Union at the time of the accident).

The TMI-2 accident resulted in a release of 6×10^{11} Bq I-131 to the environment and resulted in a dose of less than 1 mSv to the closest neighboring civilians. The containment system prevented further releases despite most of the core melting to the bottom of the steel reactor vessel. The TMI-2 accident has had no observable effect on the health of either workers at the plant or the public. However, the total does for the United States was 20 to 40 person-sieverts, which will cause one or two excess cancer deaths in the United States. Cleanup of TMI-2, which will be completed some time after 2010, will eventually cost about $1 billion, more than the plant's construction cost.

The Chernobyl accident was far worse; inadequate design, incompetent operation, and negligent disaster management combined to cause a large steam (thermal) explosion that exposed the core directly to the environment and caused great human suffering. Contemporary reports talk of a scarlet glow in the sky that could be seen 10 km away due to the light of the red-hot core being reflected in the clouds. The accident released large quantities of radioactive gases and particles, including 3% to 4% of the fuel. The greatest risk was from the release of isotopes that are readily taken up by the body,

including 1.7×10^{18} Bq of Iodine (I-131) and 8.5×10^{16} Bq Cesium (Cs-137).

About 200 tons of fuel escaped as liquid metal from the concrete cavity in which the core sat and mixed with other building material to form several piles that are now crumbling into dust. This dust presents a significant potential radiation risk, as does the 20,000 liters of contaminated water under the remains of the reactor. The reactor is contained within a massive but unstable "sarcophagus" that was constructed quickly and under extremely difficult conditions. To date, cleanup of Chernobyl has cost more than $3 billion; more will likely be needed to replace the sarcophagus and for continued monitoring and cleanup.

More than 30 employees and emergency response personnel working onsite were killed due to radiation exposure during the first several days of emergency response to the accident, and more than 100 other workers were hospitalized with acute radiation sickness. Evidence has been found that considerable numbers of neighboring civilians probably also had some acute radiation sickness, but this was not acknowledged by Soviet officials at the time of the accident.

In the subsequent decade, approximately 800,000 people were involved in cleanup of Chernobyl, most of whom received <25 mSv, but one quarter of whom received >200 mSv. Among these workers, increased (approximately double) rates of various cancers have been observed, as well as genetic damage (although it is not clear if these have hereditary implications). Evidence of genetic mutations in the DNA of children of cleanup workers is mixed. One of the more common problems that Chernobyl cleanup workers have faced is psychological distress, which has caused various neurological symptoms.

Average exposure to the public was lower but still significant. Most of these doses were from I-131 and so were delivered primarily to the thyroid. Most people received doses below 20 mSv, but some went as high as 1000 mSv, with small children and infants being most affected. People in this age cohort who lived within a thousand kilometers of the plant at the time of the accident have experienced about a five-fold increase in thyroid cancers (typically nonfatal), probably totaling a few thousand through their lifetimes. Some researchers have found evidence of additional leukemia among this group as well, while some have found no relationship. There is some evidence of increases in birth defects in places as far away as Berlin and Greece. Estimates of the overall effects are difficult to calculate, but it appears that the Chernobyl accident will cause between 10,000 and 30,000 fatal cancers, most in the Ukraine and Belarus.

Approximately 325,000 people were evacuated or relocated from the vicinity and the area within 30 km of the site is now a permanent exclusion zone. Two of the other reactors at the Chernobyl site operated until 2000 but are now shut down. The area around the site has become economically and socially depressed due to migration and other reasons. Local ecosystems were severely affected immediately after the accident, but are recovering without observable damage (although the extent of any genetic damage has not been determined).

About a dozen reactors with the same design as the reactor in Chernobyl (RBMK) exist elsewhere in the former Soviet Union, but all have received operational and engineering changes that make a repeat extremely unlikely. All other nuclear plants in the world are more similar to the TMI-2 design with massive steel containment vessels and other features (especially negative coefficients of reactivity for temperature and voids) that make an accident exactly like the one at Chernobyl impossible. However, the root causes of both the TMI-2 and Chernobyl accidents were a combination of design flaws, operating errors, and equipment failures, which can afflict any complex system. For instance, in 2002 a massive hole was found in pressure vessel of the Davis-Bessie reactor. Several inches of steel had corroded away without being noticed, leaving only a thin layer of stainless steel in place. No accident occurred, but this new type of situation demonstrates that the complexity of nuclear power plants continues to be a safety challenge.

2.2 Water Use

Water is consumed during virtually all forms of electricity generation, although fossil and nuclear plants that use steam turbines generally use far more than other technologies. Water is used to cool the plant, to make up feedwater, in emission control equipment, for ash handling, and for cleaning. By far the greatest volumes are used in cooling, although most of these withdrawals are returned to the water body (but at a higher temperature). In industrialized countries, water use by electricity generation be significant. In the United States, for example, 708 electricity generating plants (half of the installed capacity in the country) use cooling water, making up about half of all large industrial uses of cooling water in the country. These power plants have a combined average flow of 10,000 m^3/s. About half are located

TABLE III
Water Use by Power Plants

	Plant type and size							
	350 MW CCGT				350 MW coal			
Cooling technology	Direct	Indirect	Hybrid	Air	Direct	Indirect	Hybrid	Air
Withdrawal (m^3/s)	5.26	0.122	0.102	0	11.7	0.388	0.324	0
Return (m^3/s)	5.26	0.061	0.051	0	11.7	0.194	0.162	0
Net withdrawal (m^3/s)	0	0.061	0.051	0	0	0.194	0.162	0
Withdrawal per MW (m^3/s-MW)	0.0150	0.0004	0.0003		0.0334	0.0008	0.0006	

Source. Martin and Todd (1999).

on lakes and reservoirs, a quarter on streams and rivers, and a quarter on oceans and estuaries. Discharge temperatures can be as high as 40°C.

Cooling can be either direct or indirect. In the direct approach, cooling water enters a once-through system with heat exchangers that condense turbine exhaust steam to make return feedwater. All the waste heat goes into the water body, but little water is actually consumed. In addition to heating water bodies, electric power plants with direct cooling kill many marine organisms that come in contact with cooling water systems. Organisms may enter the cooling water system (entrainment) or be trapped on screens (impingement). For example, estimates for individual nuclear power plants in the 1600 to 2200 MW size indicate annual losses of 3 billion to 4 billion larvae, equivalent to 23 million adult fish and shellfish, or 17 to 45 tons of adult fish (creating observable changes in populations).

In the indirect approach, a closed loop of water absorbs heat from the steam heat exchanger to a cooling tower and gives up heat to the atmosphere, partially evaporating. This water must be made up with input from a nearby source. Tall (100 m) natural-draft cooling towers may be used (wet cooling); however, they are unsightly and may cause a visible plume or fog. Dry cooling replaces the cooling tower with finned radiators over which air is blown with fans. Hybrid systems combine wet and dry cooling. The water use and efficiency implications are shown in Tables III and IV. Hybrid cooling also tends to cost less than indirect wet cooling or air cooling and have less visual impact than wet cooling towers.

Overall, direct cooling involves larger flows of water, resulting in larger numbers of marine organisms killed and significant heating, while indirect cooling consumes larger amounts of water.

TABLE IV
Effect of Cooling Technology on Plant Efficiency

Cooling technology	CCGT efficiency (%)
Direct	56.4
Indirect—Dry	55.5
Indirect—Hybrid	56.1

Source. Martin and Todd (1999).

Technologies that do not use a steam cycle (and thus require little cooling), such as gas turbines, need little water. However, if water or steam injection is used for NO_x control, gas turbines can consume considerable amounts of water. Flue gas desulfurization to control SO_2 emissions also requires a considerable amount of water; a 1000 MW unit burning 1.5% sulfur coal would need approximately 0.1 m^3/s for this purpose alone. All solar electricity technologies need some water for cleaning, and those that use sunlight to drive a steam cycle may need cooling water as well.

Large dams can also impact water quality by vastly changing natural flow regimes in terms of flow timing, volume, turbidity, and temperature. Such changes can have very large impacts on riverine species as well as on terrestrial species that depend on aquatic sources of food. Of course, large dams pose insurmountable barriers for species that spawn in rivers and streams but live most of their lives in the ocean, such as salmon.

2.3 Solid Waste

Besides nuclear waste, the major solid wastes created by electricity production are coal combustion pro-

ducts (CCP), which consist of flyash (ash that leaves with the flue gas and can leave the stack), bottom ash (ash that falls to the bottom of the combustion chamber), slag (ash that has melted and resolidified), and sludge from flue gas desulfurization scrubbers. The ash and slag consist of impurities of silica, iron, alumina, and other incombustible matter that are contained in the coal, while the sludge is largely gypsum ($CaSO_4$). Some ash is also produced by petroleum coke or fuel oils. Many toxics are solids by the time they leave the combustion chamber (exceptions include Hg, Cl, and F) and end up wherever the ash goes.

Assuming 10% ash content on average, about 500 million metric tons of CCP are created each year. Most of this ash is dumped in large landfills; sometimes a small valley will be dammed and ash slurry pumped behind the dam until filled. Increasing pressure to protect the environment in addition to increasing interest in controlling costs have led the electricity and coal industries to find alternative uses for CCPs. The largest use is in making cement and concrete; followed by wallboard (gypsum), road construction materials, mine fill, blasting grit, soil additive, and structural fill. Importantly, the use of CCP avoids the need to mine or otherwise produce raw materials for these products. In Europe, almost 90% of CCP finds a use, while only about one-third is in the United States.

Environmental regulations have made the reuse of CCP more difficult. First, low NO_x burners increase the unburned carbon in the ash, which concrete manufacturers limit to 6%. Second, NO_x reduction systems that inject urea or ammonia into the flue gases can contaminate the ash, creating handling difficulties. Looking ahead, requirements to produce renewable fuels may lead to cofiring of biomass and coal, which will add simple organic salts to the mostly alumino-silcate CCPs. The importance of this change has yet to be determined.

Power plant operators have learned how to optimize their scrubbers to produce high-quality production by forcing the oxidation of calcium sulfite and adding chemicals that help stabilize the dry mass. Gypsum can be used to make wallboard or cement. It can also be used to mitigate acidification and acid mine drainage due to its alkalinity, or as a fertilizer and soil enhancer in some locations.

Petroleum fuels do create small amounts of ash, which will have more vanadium and nickel. Ash and slag from the combustion of municipal solid waste (MSW) contains much higher quantities of toxics, as well as potentially valuable metals (like copper). The solid waste from fluidized bed combustors will contain unreacted limestone and products of SO_2 and Cl capture, mostly as calcium compounds.

2.4 Land Use

All forms of electricity production as well as electricity transmission require large equipment and considerable amounts of land. High-voltage transmission lines can have a number of important land use implications and the siting of transmission lines in industrialized countries has become extremely difficult. However, transmission lines are necessary to use some resources effectively, and they can reduce the cost of system operation by allowing distant systems to share resources. Land use is the major environmental impact of renewable technologies. Good locations for hydropower and wind turbines are rarely located at demand centers (e.g., major cities), necessitating long-distance transmission and the associated environmental effects.

The amount of land different electricity technologies require depends on how the calculation is done. Key questions include these: Are the impacts of fuel production and transportation included? (Natural gas production can be very disruptive to wilderness areas and biomass may require vast areas for planting.) Are the impacts of waste disposal included? (Nuclear waste storage removes some areas from other uses.) Are the impacts of pollution measured in area? (Coal fired power plants can case acidification across large regions.) Are the impacts of associated transmission included? (Wind power will often have to be transmitted long distances.)

2.4.1 Aesthetic, Cultural, and Equity Concerns

Many people find electric power plants and transmission towers ugly and undesirable to live or work near. For this reason, property near power plants or power lines have a lower value than comparable properties not near power lines. Electricity infrastructure is typically prohibited in parks, wilderness areas, scenic areas, and so forth as well as historic or culturally important locations. An important and related aspect is equity: Why should one community accept a power line (or power plant) when that facility will serve another community? This problem can create tensions between urban dwellers and rural/suburban dwellers.

Concerns about aesthetics and equity are among the largest challenges facing developers of wind power farms and transmission systems. This is exacerbated by the fact that many of the best

potential wind energy sites are highly visible and prized mountain ridges. This problem has forced many new installations offshore.

As income levels rise, the desire for political freedoms and preferences for environmental amenities tend to rise as well, suggesting that aesthetics and equity will play larger roles in electricity projects in the future. This trend may boost technologies that can reduce the visual impact of transmission systems, such as more compact high-voltage direct current transmission systems and underground cables. It may also boost small, distributed generation technologies, which typically are connected to an underground gas distribution system and may require fewer transmission lines.

2.4.2 Land Submergence and Relocation

The environmental impact of hydroelectric facilities differs greatly from other types of electricity generation in that it often requires the submergence of large land areas and relocation of residents and other ecosystem inhabitants. This may be true of both facilities that use impoundments (dams) and run-of-river facilities, since the latter are often built downstream of a dam and rely on moderated river flows of the dam. Unlike other renewable energy sources that utilize a lot of land for energy collection, submerged land is lost completely, along with rivers and creeks. However, a new, lake-like reservoir is created, with its own amenities. Dams can also change the aquatic ecosystems dramatically. Importantly, dams can reduce the hazard of flooding, which can be considerable for settlements in floodplains.

Globally, about 45,000 large dams (>15 m high or 5 to 15 m high with a reservoir of 3 million m^3) exist, along with 800,000 smaller ones. About 100,000 dams worldwide generate electricity, mostly the larger ones. (The remainder are used for irrigation or water supply only.) The total area covered by reservoirs is 0.5 Million km^2, less than 1% of the land used for human habitation and agriculture. Although dam construction in North America and Europe has largely halted (or even reversed), several nations in Africa, Asia, and Latin America have considerable potential and are involved in large-scale dam development.

Reservoirs can submerge homes, farmland, wilderness, polluted sites (possibly liberating the pollutants), cultural features, and appealing landscapes, especially dramatic river valleys. Larger hydroelectric projects tend to have larger reservoirs, but submerge less land per unit of capacity. Dams under 100 MW generally require over 200 hectares per MW of capacity (ha/MW), those from 100 MW to 500 MW less than 100 ha/MW, and those above ,000 MW less than 50 ha/MW.

Relocation of large numbers of people can be very difficult and expensive, and finding new homes (and possibly farms) for them may cause environmental stress. There are often tensions between different groups, such as between urban dwellers (including elites) who will typically get most of the power from a hydro project and rural dwellers who will be displaced. Tensions between different ethnic and political groups have arisen as well. In addition, dam construction often reduces flood risk, inducing settlement on floodplains for the first time. Dams also reduce sediment loads downstream, which can weaken dikes and levies. These conditions can worsen outcomes when flood control measures of dams fail, which does occasionally happen.

2.4.3 Wilderness and Wildlife

Electric power systems can have major impacts on wilderness and wildlife, mostly because fuels or hydroelectric facilities can be located in wilderness areas or because electric power systems in different areas may find it beneficial to interconnect themselves to share resources. In particular, dams and wind turbines can affect wildlife. Submergence of habitat is the most obvious impact. In addition, the roads and power lines needed to construct, maintain, and connect hydroelectric facilities sometimes run through wilderness areas, fragmenting them. The vegetation below power lines must be trimmed, burned, or controlled with herbicides, and there must be access roads for maintenance, all of which have environmental impacts. These effects may reduce the habitat for threatened species, or cause new species (including threatened species) to move into an area. Birds are known to strike transmission towers and lines, but the numbers doing so are usually not very large, and proper routing and design can reduce bird deaths.

Birds have also been known to strike wind turbines, sometimes in large numbers. This problem can be greatly reduced by modern designs that spin slower and by not placing wind turbines in areas that birds use heavily as flight corridors.

Underground and underwater electric cables are becoming more common, most often due to space restrictions, such as in major cities, and partly because of aesthetic concerns. These cables are filled with dielectric oil and are armored. Breaks that leak the dielectric oil are rare and are usually identified quickly. Underwater cables are typically laid 1 or 2 m

below the seafloor using a mechanical plow or hydraulic jet to create a trench, which is then filled in. This alters the seabed, and there are concerns that it might have permanent effects on fisheries and marine wildlife. Trawling and other fishing techniques, repair of the cable, and retrieval at end of life must all be considered.

2.5 Electromagnetic Fields

When electricity is generated, transmitted, or used, electromagnetic fields (EMFs) are created as a direct consequence. These fields can induce small electrical currents in the body, which have been suspected of causing cancer and other health effects. This idea remains controversial, however, since the currents induced by EMFs seem far to small to cause such effects, leading to a suspicion that contact currents and not EMFs may account for the observed effects.

The main sources of exposure to EMFs are wiring and appliances in the home and workplace, and electric power lines (both transmission and distribution). The use of electricity increased for 100 years without any observed environmental effects from EMFs until a 1979 study reported that children living near power lines might have an increased risk for developing cancer. Since then, studies of human populations, animals, and isolated cells aimed at understanding the potential health effects of EMFs have been conducted, but experts are divided about their interpretation.

Three types of scientific evidence have been examined in the effort to understand the possible health effects of EMFs: human epidemiological studies, mechanistic (test-tube based) research, and animal toxicology. The epidemiological studies have shown a small risk of increased childhood leukemia, and possibly adult brain cancer, amyotrophic lateral sclerosis, and miscarriage. No evidence for other health risks has been found. However, epidemiological research can only show associations, it cannot determine cause and effect. The mechanistic and toxicological research has never found any consistent effect. The failure of these two approaches means scientists cannot explain how EMFs might cause health effects.

The degree to which experts believe that EMFs present a health risk depends on how they weigh these three lines of inquiry. Generally, however, scientists who study EMFs feel they is cannot be recognized as entirely safe, sometimes listing EMFs as possible carcinogens. If there is an effect, the risk probably falls just within the range that is considered sufficient to warrant minor regulatory action in industrialized countries. Various bodies have set voluntary guidelines that are designed to limit the induced current density in the body to $10\,mA/m^2$, sometimes reduced to $2\,mA/m^2$ for exposure to the general public. Exposure in the home is typically less than $0.01\,mA/m^2$, while contact with electrical appliances (e.g., razors, electric blankets) can create currents up to $30\,mA/m^2$.

2.6 Other Impacts

Electric power systems can affect nearby people and the environment in a number of less obvious ways. Some equipment creates noise, including wind turbines, transformers, and power lines. There is some concern that this noise could limit the suitability of land under wind turbines for wildlife or livestock. Wind farms can occupy from 4 to 32 ha/MW of capacity, although they typically only use a small percentage of that land for towers and other equipment, leaving the rest free for other uses. Wind turbines can interfere with television and radio signals. Corona effects on transmission system components can create electrical interference, but this is a highly localized effect.

Transmission lines can interfere with other infrastructures and with aircraft, possibly a significant problem in areas where aerial application of herbicides is used. Wind turbines and tall exhaust stacks can also affect aircraft. Dams will change a regions hydrology and may induce excess seismicity. Geothermal electricity production will generally cause subsidence, although this is usually controlled by reinjection of the highly saline water and steam that is brought to the surface for electricity production.

SEE ALSO THE FOLLOWING ARTICLES

Air Pollution from Energy Production and Use • Air Pollution, Health Effects of • Clean Air Markets • Clean Coal Technology • Electric Power Generation: Valuation of Environmental Costs • Electromagnetic Fields, Health Impacts of • Entrainment and Impingement of Organisms in Power Plant Cooling • Hazardous Waste from Fossil Fuels • Nuclear Waste

Further Reading

Boorman, G. A., Bernheim, N. J., Galvin, M. J., Newton, S. A., Parham, F. M., Portier, C. J., and Wolfe, M. S. (1999). "NIEHS Report on Health Effects from Exposure To Power-Line Frequency Electric and Magnetic Fields." National Institute of Environmental Health Sciences, Research Triangle Park. www.niehs.nih.gov/emfrapid/home.htm.

Garwin, R. L., and Charpak, G. (2001). "Megawatts and Megatons," p. 412. University of Chicago Press, Chicago.

Hall, M. L., and Livingston, W. R. (2002). Fly ash quality, past, present and future, and the effect of ash on the development of novel products. *J. Chem. Technol. Biotechnol.* **77**(3), 234–239.

Hammons, T. J. (2001). Environmental implications and potential of international high-voltage connections. *Electric Power Compon. Sys.* **29**(11), 1035–1059.

Martin, P. C., and Todd, I. J. (1999). The changing picture of water consumption by power stations. *J. Chartered Inst. Water Environ. Management* **13**(5), 369–374.

Neutra, R. R., DelPizzo, V., and Lee, G. M. (2002). "An Evaluation of the Possible Risks From Electric and Magnetic Fields (EMFs) from Power Lines, Internal Wiring, Electrical Occupations and Appliances." Oakland, CA, California EMF Program. www.dhs.cahwnet.gov/ehib/emf/RiskEvaluation/riskeval.html.

Nriagu, J. (1994). Industrial activity and metal emissions. *In* "Industrial Ecology and Global Change" (R. Socolow, C. Andrews, F. Berkhout, and V. Thomas, Eds.), pp. 277–285. Cambridge University Press, New York.

Olivier, J. G., and Berdowski, J. J. (2001). Global emissions sources and sinks. *In* "The Climate System" (J. J. Berdowski, J. R. Guicheirt, and B. J. Heij, Eds.), pp. 33–78. A. A Balkema, Lisse, The Netherlands.

Pacyna, E. G., and Pacyna, J. M. (2002). Global emission of mercury from anthropogenic sources in 1995. *Water Air Soil Pollution* **137**(1–4), 149–165.

Sinha, S. (2001). Environmental guidelines for power plants in India and other nations. *Environ. Qual. Management* **11**(1), 57–69.

Electricity Use, History of

DAVID E. NYE
University of Southern Denmark
Odense, Denmark

1. Inventing and Promoting Electrification
2. The Telegraph, Alarm Systems, and the Telephone
3. Public Lighting
4. Industrial Applications
5. Electrified Transportation and the City
6. Personal Uses of Electricity
7. Domestic Electrification
8. Rural Electrification
9. Air-Conditioning
10. Electricity Use and the Environment

Glossary

alternating current This current is produced when flows of electrons are made to shift directions in a constant pattern. It is easier to transmit long distances than direct current, and it is the form of electricity in most homes and businesses.

arc light An early, brilliant form of illumination produced by passing an electric arc between two carbon rods.

direct current This current is created when electricity moves in one direction along a wire. It was most common in the late 19th century; for example, it was used in Thomas Edison's first light and for early streetcars. It still has many applications, for example, in batteries.

incandescent light Light produced from an enclosed bulb by heating a filament until it glows.

In the 19th century, electricity was at first exotic, and then a luxury. Since then it has become a necessity, but paradoxically both the level of demand and the level of use vary considerably from one nation to another. The telegraph (1838), telephone (1876), and broadcast radio (1919) at first spread slowly; as with many later applications, consumers needed time to discover their uses. An interconnected electrical system grew rapidly from c. 1880, with lighting followed in rough sequence by commercial establishments (such as department stores), public buildings, monuments, streetcars, railways, subways, factories, homes, airports, farms, and stadiums. The precise sequence of adoption and degree of penetration varied from one nation to another, particularly in rural areas. Likewise, the intensity of use varied. In c. 1960, Norway, Canada, and the United States consumed more than twice as much per capita as England and France. Continually improved generating efficiencies did much to meet increasing demand while lowering the cost per kWh, until reaching engineering limits in the 1970s. Since then, new efficiencies have been in the appliances and light bulbs consumers use. Nevertheless, rising demand for electric power creates environmental problems that alternative energy sources such as wind mills can only partially solve.

1. INVENTING AND PROMOTING ELECTRIFICATION

Until the many practical applications of electricity that became available in the late 19th century, no single energy source could provide light, heat, and power. Electricity's versatility made it an enabling technology that allowed people in all walks of life—including architects, designers, homemakers, industrialists, doctors, and farmers—to do new things. These new and often unanticipated uses found for electricity mean that its history must include far more than such matters as dam construction, the extraction of fossil fuels, or improvements in dynamo efficiency. One must also consider energy history in terms of demand. Consumers have been an active force behind the expansion of electricity use. If one considers the rapid increase in electricity demand from the middle of the 19th century to the present, one is struck by how this increase parallels the growth of the middle class, the expansion of cities,

 © 2004 Elsevier Inc. All rights reserved.

the development of mass consumption, and increased demands on the environment. The uses of electricity are more than mere applications; they are inseparable from rising demand, which at times has outstripped the capacity of utilities and led to major programs in demand side management. Indeed, the same pattern repeated itself with the electric light, the electric iron, or the electric vacuum cleaner and a host of appliances and later applications. A new device first became a luxury for the rich, then a desirable item of conspicuous consumption for the middle classes, and, within little more than a generation, a necessity for all.

Practical applications were only possible in the 19th century and after. The understanding of electricity was so rudimentary in the late 18th century that Benjamin Franklin, a clever self-taught amateur in the American colonies, made an important contribution to knowledge by discovering that lightning and electricity were the same thing. The phenomenon of electricity had remained a curiosity since classical times, when static electricity was known but scarcely investigated. Few scientists focused attention on it until after the invention of the Leyden jar (1746), which facilitated a century of experiments that prepared the way for practical applications. Charles Coulomb, Michael Cavendish, Benjamin Franklin, Luigi Galvani, Alessandro Volta, Humphrey Davy, and Andre Ampere developed a rich store of observation and theory that made possible Faraday's studies of electromagnetic induction, which in turn made possible the dynamo (1831). When Faraday was asked whether his experiments would ever result in tax revenue, he boldly declared that they would.

Indeed, during the next half century a series of industries and businesses emerged based on the special properties of electricity. Its movement was so rapid as to be instantaneous, a property exploited by the telegraph, various alarm systems, and the telephone. It could produce electrolysis, making possible metal plating. Electricity could also produce light when passed through certain conductors, eventually making possible arc and incandescent lighting. Electricity provided a way to transmit power over considerable distances, allowing falling water or burning coal to be translated into motive force by distant electric motors. Electricity could also be used to produce precisely controlled heat without fire or the consumption of oxygen, leading to applications in baking, enameling, and industrial heating and burning. Each of these innovations was eventually incorporated into devices used in the home and on the farm.

Given the many potential uses of electrification, it might seem that its adoption was virtually automatic. Such was not the case. As late as 1912, less than 3% of electrical power was residential, whereas almost 80% was used in manufacturing, with the remainder used commercially. Although it might be tempting to treat the history of electrification as the story of inventors producing a stream of practical devices, it was a far more complex developmental process that included government, manufacturing, marketing, and several different groups of consumers. Utilities quickly discovered that large, urban customers were far more profitable than smaller customers in suburbs and the countryside. The spread of electrical service therefore moved from larger to smaller customers, and from urban to rural, in a series of distinct stages. This process began in the 1870s in Europe and the United States and has since spread throughout the world. However, even in the early 21st century electricity has by no means become the dominant source of lighting or power in every location, and in most cases other sources of heat are more economical. If the world's capital cities are electrified, some of the countryside remains in darkness, as revealed by satellite photographs. Indeed, even in the nations in which the process began—England, France, Germany, and the United States—in 1900 almost all factory power still came from falling water or steam, and less than 5% of all homes had electricity. Although in retrospect electrification might appear to have been a steady and inevitable process, closer inspection reveals uneven growth, at times thwarted by competition from gas, coal, and other energy forms. In general, electricity replaced gas for lighting because the incandescent bulb was safer and cleaner than an open flame and because it was more like daylight. In contrast, gas remains a strong competitor to electricity for heating and can also be used for air-conditioning. Yet such generalizations do not apply universally because energy use was inflected by local conditions and by social class. For decades, English workers clung tenaciously to gas lighting, leaving electricity to the middle and upper classes. In contrast, wealthy Swedes were reluctant to give up traditional forms of lighting that they considered genteel, whereas Swedish workers more rapidly embraced the electric light. In 1931, less than 30% of British homes were wired, and few were all electric, at a time when 90% of Swedish homes had been wired. In the United States, most urban but few rural dwellings were electrified by that time, and wiring was less a marker of social class than of geographical location.

From a technical standpoint, electrification was not a unified process that spread like ripples in a pond from certain nodal points. Rather, there were several interrelated processes, which began at many different locations, driven not only by utilities but also by the demands and interests of consumers. Particularly in the early decades, separate systems developed to serve different needs, including small stand-alone generating plants for some hotels, department stores, theaters, and wealthy homes and large, direct current power houses owned by street railways. Until the 1930s, many factories also had their own power stations.

The degree of such fragmentation varied, being considerably greater in Britain than in the United States, for example. In the early 20th century, London had 65 electrical utilities and 49 different types of supply systems. Anyone moving within London risked entering a new grid because there were 10 different frequencies and 32 different voltage levels for transmission and 24 for distribution. Once consumers had converted all the lamp and appliance plugs and made other appropriate changes, they discovered that each utility also had its own pricing system. In the United States, where utilities worked harder to coordinate their systems and interconnect their lines in order to achieve economies of scale, consumers still faced considerable inconsistencies in plug and socket sizes. These were only ironed out in the 1920s with the help of the federal government. Even today, as every tourist knows, the electrical current and the plugs needed to access the system vary among the United States, Britain, and continental Europe, and these three different standards have been reproduced throughout much of the rest of the world.

The large electrical manufacturers, which sold generating equipment and provided technical expertise to new markets, exported these standards. With so many potential customers, the manufacturers did not attempt to enter all markets. Rather, each focused on certain regions, to a considerable degree based on political spheres of influence and colonial empires. For example, General Electric (United States) sold lighting and traction systems to Brazil and much of Latin America, whereas in most cases General Electric (United Kingdom) transferred the same technologies to the British Empire. German and Austrian electrical expertise spread throughout central Europe and the Nordic countries. To some extent, France developed its own systems but relied heavily on investment and partial ownership from the United States.

On the demand side, in all the industrial countries the middle classes developed a demand for electrification in many guises. On the supply side, the electrical industry was an expansive investment opportunity that was more important and longer lasting than the Internet bubble of the 1990s. Between 1875 and 1900, electrical equipment makers and utilities formed the most dynamic sector of the American economy, growing from a few small adjuncts to the telegraph companies into a $200 million industry. After 1900, it grew into a colossus that was inseparable from the health of the national economy. The electrical industry's potential was early recognized by J. P. Morgan and other leading investors who financed Edison's work. Once commercial development began, a flurry of mergers reduced the field from 15 competitors in 1885 to only General Electric and Westinghouse in 1892. Before this time, railroads had been America's largest corporations, demanding the greatest capital investment, but they were a mature industry compared to the emerging electric traction companies, local utilities, and equipment manufacturers.

2. THE TELEGRAPH, ALARM SYSTEMS, AND THE TELEPHONE

The invention of the telegraph in 1838 heralded the first widespread practical application of electricity. However, potential consumers were not certain what to use the telegraph for. Conceived and patented by Samuel F. B. Morse, the telegraph astounded incredulous crowds, and the first telegraph offices often provided seating for the public, who could scarcely believe that it was possible to sever language from human presence. Twenty years after its invention, when thousands of miles of lines linked the United States, the *New York Times* declared that "the telegraph undoubtedly ranks foremost among that series of mighty discoveries that have gone to subjugate matter under the domain of mind" (August 9, 1858).

However, even the inventor of the telegraph did not understand its enormous commercial potential, and he tried to sell it to the federal government. Many European countries chose to make the telegraph part of the postal service, but the U.S. Congress refused, and only after 5 years granted Morse $30,000 to build a line from Washington to Baltimore. People seemed to have little use for it, however, and even when the first line connected New York, Philadelphia, and Washington business still developed slowly. Six months after it started operation, the line averaged less than 2000 words a day, and it was generating

only approximately $600 a week. By 1848, the telegraph linked Chicago to New York, however, and grain traders with immediate access to information soon made fortunes trading in midwestern wheat. In less than a generation, Western Union became one of the largest corporations in the United States. During the same years, newspapers realized the commercial value of Mexican war news and formed an association that inaugurated the wire services. As these examples suggest, the potential of the various forms of electrification was only gradually understood.

During the second half of the 19th century a series of technologies based on the telegraph, such as burglar alarms, stable calls, and door bells, familiarized people with the idea that a current carried over wires could be instantaneously transmitted to virtually any point. However, other practical applications were few because most electricity still came from inefficient batteries. Further use of electricity was limited until large and efficient generators became available in the 1870s. Until then, "impractical" electric lighting was confined to special effects in theaters.

Like the telegraph, the invention of the telephone (1876) at first seemed a curiosity, and Alexander Graham Bell was unable to sell his invention even to Western Union. Setting up on his own, Bell concentrated on businessmen and wealthy households and did not expand service widely until his patent ran out and competition emerged in the 1890s. It took more than a generation before phone companies realized that people wanted phones not just for business and emergencies but also for gossipy conversations. Even so, the telephone took a long time to become a necessity to consumers compared to the electric light. Only after World War II did most American homes have telephones. Even then, they were all black, heavy apparatuses that the subscriber did not own but leased. They were not stylish elements of décor that varied from one home to another but utilitarian objects that were engineered to last decades. In contrast, a greater variety of models were marketed in Europe, despite the fact that universal service only came to some countries during the 1960s, and Spain and Portugal only achieved it at the end of the 1980s. On the other hand, during the 1990s Europeans adopted mobile phones far more quickly than did Americans.

Again, as late as 1919 the U.S. government failed to see the broadcasting potential of radio, understanding it only as a superior form of the telegraph, for wireless point-to-point communication. In the early 1920s, a few amateurs initiated the first public radio broadcasts. Yet even when the potential of a device is grasped, not all societies have made the same choices about how to embed electrical technologies within the social structure. Should suppliers be public or private or a mix of the two? Should pricing encourage extensive or intensive use of a system? What is the best balance between regulation and open competition? Between 1880 and 1900, such questions had to be answered first with regard to public lighting and then street traction systems.

3. PUBLIC LIGHTING

In both Europe and the United States, regular public electric lighting appeared first in the late 1870s as a spectacular form of display in city centers and at expositions and fairs. The first installations were arc lights, which shed a powerful glare over streets and the insides of some stores. Department stores and clothiers quickly adopted such lights because they were cleaner, brighter, and safer than gas and because they proved popular with customers. By 1886, 30 miles of New York's streets had arc lights, installed approximately 250 ft apart. In 1879, Thomas Edison demonstrated his enclosed incandescent bulb, which had a softer glow that was well suited to most indoor and domestic settings. Both arc and incandescent lighting competed with already existing gas systems until the first decades of the 20th century. Edison recognized that his light, its distribution system, and its cost had to be designed to complete with gas. For as much as a generation, some new buildings had both systems. The Edison system was rapidly adopted wherever convenience and fissionability were highly valued, and it spread rapidly into the wealthiest homes, the best theaters, and the most exclusive clubs in New York, London, Vienna, and Paris. However, 20 years after its introduction, the electric light was almost entirely restricted to public places. Less than 1 home in 10 had been wired, and the average person encountered electricity at work, in the street, and in commercial establishments.

From its inception, the public understood lighting as a powerful symbolic medium. The history of lighting was far more than the triumph of the useful. Before gas and electric streetlights people had to find their way with lanterns, and the city at night seemed fraught with danger. Public lighting made the city safer, more recognizable, and easier to negotiate. However, mere functionalist explanations cannot begin to explain why electric lighting had its origins in the theater or why spectacular lighting emerged as a central cultural practice in both Europe and the

United States between 1885 and 1915. As early as 1849, the Paris opera used an electric arc lamp to create a sunrise, and showmen of every description seized on the new medium to attract crowds. During the 1870s, thousands of people turned out to see the first arc lights in Cleveland, Boston, London, and Paris. The new light fascinated them not only because it was so much brighter than existing gas but also because it seemed to violate the natural order. For the first time in history, light was separated from fire. It needed no oxygen. It was unaffected by the wind. It could be turned on in many places simultaneously at the turn of a switch. In 1883, Niblo's Garden featured "novel lighting effects by the Edison Electric Light Company," which included an illuminated model of the Brooklyn Bridge and chorus girls with lights flashing from their foreheads as they flourished electrically lighted wands. Promoters and public alike demanded ever-greater public displays as use far exceeded necessity.

Intensive lighting became a form of symbolic expression at not only theaters but also world's fairs and other public events. Every exposition attempted to outdo all previous efforts by escalating both the scale and the completeness of displays. Intensive arc lighting at Chicago in 1894 gave way to thousands of incandescent lights at Omaha (1898), arrays of colored bulbs in Buffalo (1901), and artificial sunrises and electrical fireworks in San Francisco (1915). Many fairs made illuminated towers their central symbols, notably Buffalo's Electric Tower, San Francisco's Tower of Jewels, and New York's Trylon and Perisphere in 1939. Electricity became more than a theme; it provided a visible correlative for the ideology of progress, and it was placed at the apex of an evolutionary framework, as opposed to "ethnological" villages—compounds representing the life of Mexicans, Cubans, blacks in the Old South, Filipinos, or Africans in the jungle—where the visitor could see staged versions of earlier forms of social evolution. Electrification thus became embedded in a social Darwinist ideology of racial superiority, as darkness represented the "primitive," whereas artificial light exemplified Christianity, science, and progress. At the Buffalo Pan-American Exposition, "the strongest, crudest colors" were nearest the entrances, and there was "a progression from warm buff and ochre walls … through more refined and brilliant hues to … ivory white, delicate blue, greens, and gold" on the central tower.

During the same years, electrical advertising signs spread from a few isolated installations in approximately 1890 to nearly ubiquitous evening displays. Particularly in the United States, main streets became "great white ways" that were more alluring at night than during the day. In each nation, particular locations became famous tourist attractions because of their advertising signs, notably New York's Times Square and London's Piccadilly Circus. In these public squares, the electric sign, the spotlight, and even the streetlight became economic weapons. The competition for attention did not create the majestic serenity of an exposition, dominated by beaux-arts architecture bathed in white light. Commercial lighting produced a potpourri of styles and colors flashing against the night sky. Thousands of individual decisions produced this electric landscape, but it nevertheless had an unanticipated collective impact. Literally millions of lights seen together, particularly from a distance or from the top of a skyscraper, had a strong popular appeal. Electric lighting was also used to dramatize important monuments. One of the most effective early displays was the lighting of the Statue of Liberty in 1886, which made the statue stand out against the darkness of the New York harbor. In the following decades, search lights would be permanently trained on many of America's most important symbols, including many skyscrapers, the White House, Niagara Falls, Old Faithful, Virginia's Natural Bridge, and Mt. Rushmore.

Between 1880 and approximately1920, dramatic lighting transformed the night city into a scintillating landscape that was a startling contrast to the world seen by daylight. Street lighting, illuminated buildings, and electrical advertising simplified the city into a glamorous pattern that signaled more than the triumph of a new technology. By 1920, spectacular lighting had become a sophisticated medium of expression that governments and corporations used to sell products, highlight public monuments, commemorate history, encourage civic pride, and even simulate natural effects.

4. INDUSTRIAL APPLICATIONS

Industry found many uses for electricity, and from 1880 to 2000 it used the preponderance of power in most nations during any given year. Factory electrification began with lighting. Munitions plants and flour and cotton mills quickly adopted electric light because it was far safer than any other form of illumination. Printers, artists, and others whose work required accuracy and true color relations found gas inferior to incandescent light. Factory owners soon realized that light made possible the construction of

large interior spaces without regard for windows. However, if lighting came quickly to the factory, adoption of electric drive required a generation. Existing steam and water power functioned well and cost a great deal to replace. These power sources could not easily be transmitted over more than a few hundred feet, but there were other alternatives. In 1891, to transmit water power 2 km or less, wire rope cost half the price of electricity. At that time, electricity was only cost competitive for networked power systems more than 5 km long. For example, high-pressure hydraulic transmission was a well-established alternative. Introduced by Joseph Bramah in Dublin in 1795, it was applied to cargo handling in many British ports and to centralize urban hydraulic power supply. Hydraulic power was also applied to heavy industry, such as steel production, and water motors using a city water supply could drive small appliances requiring less than 2 horsepower as well as church organs and many elevators in commercial buildings. Milwaukee alone had 600 hydraulic elevators connected to the city water system. Power from illuminating gas also briefly flourished in the 1890s. Clean, safe, compressed air, however, was the most successful alternative to electric power and was well developed in Paris, where in 1899 more than 100 miles of pipe supplied 24,000 horsepower to a wide variety of customers.

Nevertheless, electricity proved so versatile and generation costs declined so quickly that after 1900 it was in the ascendant. An investigation by the Franklin Institute showed why it was preferred. Interestingly, savings in energy costs were considered to be of no great importance since power usually comprised only 1–3% of a workshop's budget. Furthermore, electrical equipment was somewhat more expensive than the familiar mechanical drive, but savings in maintenance and depreciation offset this expense. Electricity's advantages were primarily in creating freedom in the location of machinery, improvements in lighting and ventilation, cleanliness (there were no overhead oil drippings from shafting), reductions in shutdowns, greater precision of electrical equipment, and increases of 20–30% in output as a result of all these factors.

Adoption of electricity in industry varied considerably by nation, however, with the United States taking the lead, closely followed by Germany. Although accurate statistics are difficult to obtain, in 1928 one British expert estimated that 73% of manufacturing power was electrical in the United States compared to 67% in Germany, 48% in Britain, and less for other European countries. Britain particularly lagged behind in iron, steel, and chemicals, but this gap was partly closed during the 1930s, when its economy performed better than that of the United States.

Electricity's versatility extended into virtually all areas of manufacturing. For example, electrolysis enabled silverware manufacturers to coat materials with silver. In coal mines and rock quarries, electric drills were far less cumbersome to move around than drills driven by compressed air or steam. By 1920, many steel rolling mills had installed electric drive and controls that made it possible for one man to guide steel from the ingot furnace to the finished rail. An observer noted, "Every motor replaced a man, but the work is done better and more quickly." In food processing plants, aluminum factories, brass foundries, and potteries, electric furnaces and ovens produced no smoke, consumed no oxygen, and maintained high temperatures precisely. Such factories reduced spoilage and increased throughput.

Electric motors proved especially useful in moving materials, particularly in mines and other closed environments. Electric motors were also adapted to overhead hoists and cranes, and these began to move their loads more freely once overhead drive shafts were removed in favor of electric drive. Indeed, by 1920 electricity had become the basis for a revolution in factory design, brightening and opening up the work space and making possible new work flows, most notably on the assembly line, where electrically powered machines were arranged in combinations that were impossible using older power-delivery systems. This is not to say that electrification necessarily led to a particular form of mass production. Rather, it facilitated experiments and new designs. At the same time, electrification altered the siting requirements of factories, which no longer had to be immediately accessible to rushing streams or coal supplies. Indeed, after the successful demonstrations of alternating current in the 1890s, the short practical transmission range of direct current no longer placed a limit on where power could be delivered. Whereas early generating plants such as Edison's Pearl Street Station in New York had to be near the city or town to reach customers, during the 20th century power stations increasingly were located on the edge of the city or in the countryside. Because transmission lines could reach virtually anywhere, factories and businesses could also move away from railroads and canals to the edge of town or along main highways. Thus, electrification was used to transform both factory layouts and the location of industry.

From a worker's point of view, incandescent lighting reduced the danger of fire, reduced pollution, and made it easier to see, but it also made possible round-the-clock shifts. Furthermore, as electric motors and cranes provided more horsepower to production, they made possible radical changes in the layout of production, most strikingly in the development of Henry Ford's assembly line in 1912, an innovation partly anticipated by Edison's experiments with automating iron mining in the 1890s. The assembly line was literally impossible in any complex industry before machines were freed from the rigid organization required by overhead drive shafts and given the flexibility of electrical devices.

The connection between electrification and mass production at times has been overlooked by historians, who have emphasized scientific management, particularly the work of Frederick Winslow Taylor. It is true that Taylor systematized work and made tasks such as shoveling more efficient. In contrast, Ford's assembly line used electrified machines to eliminate shoveling altogether. Taylor retained piecework, a concept that makes no sense on the assembly line, where all share the same pace. Taylor maximized efficiency in existing production technologies; Ford transformed the means of production. Taylor saved time; Ford sped up time. In 1914, after Taylor imposed his system on the Packard Motor Company, its 4525 workers produced 2984 cars a year, less than 1 per man. This was essentially artisanal production of a luxury vehicle. Ten years earlier, well before the assembly line, Ford produced 12 cars for every worker. By 1914, 13,000 Ford employees, using conveyor belts, electric cranes, and 15,000 specialized machine tools, manufactured 260,000 cars or 20 per worker. Ford's system outperformed Taylor's by more than 2000%.

5. ELECTRIFIED TRANSPORTATION AND THE CITY

As industrial plants grew larger and became concentrated in cities, the urban scale expanded and electricity supplied many forms of transportation. Indeed, it made possible faster and easier movement not only at the street level but also into the sky and down into the earth. The Columbian Exposition of 1894 introduced visitors to electric boats and electrified trains and even installed some electrified moving sidewalks. Although similar sidewalks are found in few places outside airports today, a closely related technology, the escalator, has become almost ubiquitous in department stores and large public buildings.

The clean and quiet electric car initially seemed to be a promising form of urban transportation. A French coach builder, Charles Jeantaud, introduced an electric car in 1894, but it never sold well. In contrast, many American inventors enjoyed initial success, including an "Electrobat" introduced in Philadelphia in the same year. The early leaders were two Hartford-based firms, Pope Manufacturing, which made 500 vehicles in 1897, and the Electric Vehicle Company, which made 2000 electric taxis and several electric trucks the following year. For approximately a decade such vehicles were almost as numerous as fledgling gasoline cars. Even Henry Ford bought an electric car for his wife Clara in 1908. Such vehicles were easier to control than gasoline cars, and with their slower speeds (25–30 mph) and shorter range (60–80 miles) they seemed ideal for women making social calls or shopping but not for men, who demanded faster, long-distance transport. The electric car could easily strand drivers in the countryside, far from any recharging station, and even if one were found recharging took hours. If electric trucks and cars largely disappeared after the early 1920s, however, smaller electric gaffle trucks, golf carts, and indoor delivery vehicles held a niche market. Eight decades later, electric vehicles are still being designed and tested, and hybrid cars are now sold that switch automatically from gasoline motors to electric drive, depending on driving conditions.

Other forms of electric transportation were more widely adopted. Vertically, the elevator made the skyscraper practical, as it carried executives to their preferred offices high above the streets. These offices had electric lights and telephones, their plumbing was supplied with water by electric pumps, and their air could be heated or cooled electrically. The electrified office literally detached executives from the noises, smells, and temperatures of the street.

Horizontally, electric trolleys made a much larger city feasible. In the 1890s, streetcars were faster and cleaner than horsecars, and they could be stopped more precisely and climb steeper grades than cars pulled by steam. Electric cars could easily be lighted at night and heated in winter, and overall their operation cost less than any alternatives. Real estate promoters often invested in or owned traction companies, whose spreading lines defined much suburban and regional development, especially in new cities such as Minneapolis and Los Angeles but also in older metropolises, such as Boston, Copenhagen, or Amsterdam. By

1902, large cities averaged more than 200 trips per citizen per year, and Americans alone took 4.8 billion journeys by streetcar.

In the larger cities, notably Paris, New York, Budapest, and London, the technology of the electric tram was adapted to underground use. Construction began in the 1890s, and millions of commuters and shoppers were literally hidden from view. The subway showed that electric light, power, and ventilation made possible an entire new urban setting beneath the pavement. New department stores added several brightly lighted basement stories, and underground arcades connected some buildings. In Toronto and Montreal, where winter temperatures prevail, such urban designs became particularly popular.

Electrified transport transformed the city, bringing an unprecedented number of rural and urban consumers to the center, which stimulated new forms of entertainment. Vaudeville theaters and movie houses were able to fill thousands of seats several times a day, and new sports arenas (electrically illuminated) drew even larger crowds to see professional baseball, football, and basketball. Traction companies drew some of the crowd to their own amusement parks, which consumed some of their excess electrical output on evenings and weekends. The technology of electrical traction was the basis of the roller coaster, the subterranean tour, and many other rides. Trolley parks spread rapidly after 1888, and by 1901 more than half of all street railway companies operated at least one (often several). These typically generated as much as 30% of the line's overall traffic as well as provided a market for electricity. The trolley made the city the center of an easily accessible network, from the downtown shopping district to streetcar suburbs, the amusement park, and the fringe of rural hamlets on interurban lines.

Urban electrification was not merely a matter of improving existing structures. The electrified city had a fundamentally different built environment than the city before it. At the center were large office buildings, department stores, theaters, and brilliantly lighted public buildings accessible on an expanded scale reaching miles further into surrounding countryside. Practical electrical lighting, elevators, and air-conditioning likewise made it possible to create self-enclosed environments, including the later developments of indoor sports stadiums and shopping malls. The modern city is unimaginable without electricity, which enabled enormous changes its in design.

Yet these changes were by no means the same in all cultures. In the United States, where the automobile became a central form of transportation in the 1920s, the electrified city was quickly modified, sprawling out into the suburbs. In contrast, European cities retained mass transit and bicycles for a generation longer, and as a result the electrified cityscape developed and became naturalized. The urban landscape made possible primarily by electrification (i.e., without intensive use of the automobile) is therefore far more visible in a city such as Oslo than in Los Angeles.

6. PERSONAL USES OF ELECTRICITY

The public became excited by electrical medicine well before Edison had perfected his incandescent light. Electricity seemed to be a mysterious fluid that permeated living organisms, leading Walt Whitman to write of "the body electric." Electricity seemed capable of recharging the body, helping it to overcome a host of diseases and conditions. Neurasthenia and neuralgia became fashionable complaints apparently caused by lack of nerve force, and millions bought patent medicine cures for electrical imbalances. Respected doctors hooked their patients up to generating machines and passed currents through diseased limbs or sought to (re)animate sexual organs. Patent medicines such as Dr. Cram's Fluid Lightning claimed to contain electricity in liquid form so that one could ingest it and revitalize the whole body. By the mid-1890s, magazines such as *Popular Science Monthly* asked, "Is the Body a Human Storage Battery?" So many believed the answer was yes that electrical devices were mass marketed. The *Sears and Roebuck Catalog* for 1901 sold electric belts designed to restore men's sexual powers and to help either sex restore lost energies. Some electrotherapeutic electric belts retailed for as much as $40 and promised a "triple silver plated battery of great power, silver plated nonirritating electrodes, and a very powerful electric sack suspensory." Women bought electric corsets fitted with "magnetic steel" with the expectation that they could mould their figure to any desired form. Even the U.S. Congress had a cellar room in the Capitol building fitted with an electrical medical apparatus. The *Electric Review* noted, "The members say it is splendid after they have exhausted their brain power by speechmaking or listening. A great many members take electricity, and some go to the basement of the Capitol for it every day during the season."

Although these "uses" of electricity are clearly bogus, other applications just as clearly were not. By

the 1890s, it was firmly established that ultraviolet light, either from the sun or from arc lights with special filters, could kill bacteria, and in 1896 Niels Finsen discovered that lupus could be cured by repeated exposure to ultraviolet rays. He won the Nobel Prize for this work in 1903, and such successes seemed to be only the beginning. The Belgian chemist Ernest Solvay proposed that electrical stimulation might accelerate plant growth and increase the yield of crops. Similar ideas inspired experiments by the Nobel Prize-winning chemist Svante Arrhenius. He "tested the effect of electrical stimulation on the growth of children," who, he reported, grew 20 mm more than those in the control group.

When prestigious scientists made such claims, the gullibility of laymen is not surprising, and the enthusiasm for electrical cures lasted well into the 20th century. Despite the efforts of the American Medical Association to discredit them, people bought electrical cures in the 1920s and 1930s. For example, thousands of people paid $60 for an I-ON-A-CO electric collar in hopes of rejuvenation. As this popular faith in electrical medicine suggested, people regarded electricity to be a magical fluid, a nerve-tingling "juice," or an invisible tonic that could restore normal health. From the 1880s until 1920, when few people had wired their homes, electrical medicine offered direct contact with the mysterious force that was transforming work and public life. For approximately half a century, the body literally became the contact zone where men and women explored electricity's possible meanings.

By the time the electrical medicine craze was over, electricity had begun to be used to enable the body in many ways as battery technology improved. Small, lightweight batteries could deliver small but constant amounts of power for a year or more, and starting in the 1950s they first made possible electric watches and hearing aids, and later pacemakers and a host of ingenious medical devices. Slightly larger batteries combined with the new transistors made possible radios that could be carried in a shirt pocket. Consumers have not given up on using electrical devices to reshape their bodies. Some still buy electric belts, advertised on TV in the wee hours of the morning, and attach electrodes to their stomachs with the expectation that they can effortlessly tone up their muscles and look thin. Millions of people pay to use private gyms with Stairmasters and other electrical devices that tell them how long the workout has lasted, monitor heartbeat, and count calories lost.

By the 1990s, improved batteries, weighing less and lasting longer, had become an assumed part of most people's everyday life. They drove electric toothbrushes, cordless drills, and other tools. They powered tiny portable radios, disc players, mobile telephones, and palm desktop computers. They had become essential to thousands of children's toys, even if they were not usually included with those toys when purchased. In outer space, satellites and exploring machines relied on solar-powered batteries to keep them operational for years. In little more than a century, electrical devices had penetrated into the most intimate parts of the body and reached millions of miles into space and, in the process, extended human sight and hearing.

7. DOMESTIC ELECTRIFICATION

In the 1880s, the few domestic consumers of electricity were almost exclusively wealthy, and they chose the new form of illumination as much for its social prestige as for convenience. Some even bought personal generating plants, especially in the countryside. Utilities did not focus on domestic customers until after a generation had passed, during which time a host of appliances appeared that made higher levels of consumption possible. Electric lights alone did not create much demand, but with the marketing of electric irons, fans, toasters, coffee makers, and other devices after 1900, the domestic consumer became more interesting. In the United States, gas continued to supply the vast majority of light and heat to private residences until after 1910, with the major shift to electricity occurring between 1918 and 1929. In Britain, the change occurred later, and as late as 1919 less than 6% of all homes were wired.

In the decade before World War I, the new profession of home economists championed the electrified house, but the domestic market only became particularly attractive to utilities once their overall load curve peaked in the day, creating a need for compensating night demand from homeowners. However, the conversion to electricity was neither automatic nor easy, and in many areas gas remained the stronger competitor, for example, in refrigerators until at least the end of the 1920s and in stoves and home heating until the present. Furthermore, it was not always easy to determine what the consumer would buy. Electric Christmas tree ornaments were a success in the United States but less so in some European countries. Even today, most Scandinavians prefer to put lighted candles on their trees. Electric cigarette lighters caught on in automobiles but not in the home. The air-conditioned bed never sold well,

although there was limited success with electric milk warmers for babies.

Given the uncertainties of public demand, could the electrical utilities create this demand in practice? In Chicago, Samuel Insull showed other utilities that it was possible to build up the load through innovative pricing policies designed to attract a wide variety of customers, whose variable energy use evened out the demand curve and permitted production efficiencies. His variable pricing persuaded large customers, such as factories and traction companies, to abandon their own power systems, creating daytime demand that decreased at night. This created the generating capacity for a large domestic market, where use peaked after working hours.

In addition to Insull, Henry L. Doherty in Denver was another innovator in promotion. Their salesmen marketed electrical devices systematically, selling door to door. They also employed extensive newspaper advertising, built model electric homes that were extremely popular in approximately 1920, and paid for follow-up demonstrations by home economists in schools and at public events. Many people were afraid to use an electric iron, for example, and salesmen were trained to educate consumers and to dispel their fears of the new technologies. They emphasized cleanliness, comfort, convenience, and economy of operation—themes particularly appealing to those responsible for public buildings, especially schools. By approximately 1915, educators were convinced that the air temperature should be 68–70° and that sickness could be controlled through good ventilation. Schools also prepared a new generation of consumers. Students experienced the superior light, heat, and ventilation that electricity made possible, and in class they learned about the scientific background of the new domestic technologies. There were also gender-specific classes: newly invented home economics for girls and shop for boys. Additional domestic appliances made the home an even more attractive market as vacuum cleaners and washing machines became common during the 1920s, followed by radios, stoves, electric phonographs, and refrigerators in the 1930s.

The widespread adoption of labor-saving machines soon spread in Europe. In Britain, by the late 1920s domestic power consumption had outpaced commercial use. France lagged slightly behind and experienced the appearance of electric kitchen appliances as Americanization, although its level of consumption nevertheless was only 40% that of the United States as late as 1937. In that year, Norwegians and Canadians were by far the most avid domestic consumers. They faced long, cold winters but possessed ample hydroelectric resources and as a result used more than twice as much electricity per inhabitant as the United States and four times as much as Germany. Although electrification spread more slowly in some areas than others, once it entered the home it was seldom removed, and by the 1930s it had begun to be seen as a necessity.

U.S. consumption continued to double every decade. *Fortune* noted in a 1969 article on the dangers of brown-outs and utility overloads that "the American likes his home brilliantly lit, of course, and he has a passion for gadgets that freeze, defrost, mix, blend, toast, roast, iron, sew, wash, dry, open his garage door, trim his hedge, entertain him with sounds and pictures, heat his house in winter, and—above all—cool it in summer."

The European pattern of development was much different than the American pattern. In part because of the denser population and in part because electricity was more often considered to be a state service rather than a private business, lighting was widely dispersed because it seemed fair to do so, although the cost was generally higher than in the United States. As with gasoline, the expense of the service discouraged Europeans from adopting a wide range of appliances during the first half of the 20th century. Since approximately 1960, however, the patterns of consumption have converged, and by 2000 a French or German home had much the same repertoire of electrical appliances as one would find in Canada or the United States. Nevertheless, differences remain. What are considered normal and acceptable levels of illumination in many European homes may seem dim to American visitors. Many a temporary resident complains that the owners must have removed most of the floor lamps. In fact, Scandinavians prefer a living room that is dim by American standards. They generally find it cozy to have only small pools of light and a few candles.

8. RURAL ELECTRIFICATION

One of the most striking changes in the past two centuries is the shift away from agricultural life, which rural electrification abetted, although it had begun long before. In 1800, approximately 90% of all Europeans and Americans were farmers; today, farmers comprise less than 4%. Half the United States was still rural until the 1920 census, so much of the decline had little to do with electricity. However, the disappearance of millions of additional farmers was due in large part to the increased

efficiency brought by electrification. Although the scale and organization of individual farms changed continually for 200 years, rural electrification exemplifies the substitution of capital equipment for labor, bringing both reductions in waste (e.g., through refrigeration) and improvements in productivity using such devices as milking machines, incubators, and electric pumps.

Not all applauded these developments. British thinkers, such as G. M. Trevelyan, Vaughan Cornish, and Patrick Abercrombie, developed environmental ideas about the relationship between man and nature. Such thinkers believed that periodic experiences of wilderness could assuage the wounds of urban civilization and therefore attacked the march of electrical pylons across the countryside. In contrast, most agrarians in the United States embraced electrification as a way to reverse flight to the city by bringing new amenities to farmers and small towns. John Crowe Ransom, for example, realized that electrification could make farming more attractive, and Herbert Agar argued that it could bring some urban amenities to the countryside and help preserve the Jeffersonian ideal of equality and independence.

Although the long view makes the process of electrification seem rational and even inevitable, in fact rural electrification was long ignored by utilities, unless prodded by government. Some of Edison's early utilities were in smaller cities, but he quickly realized that the distribution costs for scattered customers were too high. Indeed, the costs were much higher for direct current than for alternating current, which only became commercially available in the 1890s. After that time, in places where rural services were nationalized or subsidized, such as in New Zealand, the countryside was rapidly brought into the grid. In the United States, however, where power was largely privately controlled, more than 90% of all farmers still did not have electricity as late as 1935. Although equipment manufacturers and hardware stores advertised electrical wares to farmers in magazines and displayed appliances at county fairs, few people could buy them because before the 1930s private utilities used most of their resources to electrify urban homes and sponsored few experimental rural projects. The long distances between farms, with only three or four customers per mile, made transmission costs high, and farmers had little money due to the long agricultural depression that persisted throughout the 1920s.

For utilities, this rural poverty was not an incitement to action but a caution against expansion. As a result, by the 1930s electricity had become naturalized as a normal part of daily life in cities, creating a vivid contrast with a countryside still plunged in darkness. New Deal politicians, encouraged by Franklin D. Roosevelt, made rural power a national issue and created a Rural Electrification Administration (REA) that established rural cooperatives throughout the United States. Indeed, one of the selling points for this program was that the nation had fallen behind the successful efforts in Europe to provide electricity to the entire population.

Organizing farmers into REA cooperatives and getting electric lines to their homes were only first steps. A self-sustaining cooperative required higher electrical consumption, and REA coops soon were being visited by "utilization experts," Washington's bureaucratic term for salesmen. Their mission was to convince farmers to buy motor-driven equipment and larger appliances. Customers who never installed more than a few electric lights did not create enough business to enable coops to repay government loans. Washington agencies at first misjudged the farmers, however, and the utilization experts initially pushed heavy farm equipment. In practice, however, only dairies and chicken ranches discovered immediate advantages of electrical equipment. Raising wheat, tobacco, cotton, and many other crops could be done profitably much as before. However, the farmhouse was quite a different story. Farmers adopted washing machines and refrigerators more quickly than urban consumers, and they did so for sound economic and practical reasons. Cities had laundries, but in the country all washing had to be done by hand after water was pumped from a well and heated on a stove. Little wonder half of the REA's customers in the first 3 years bought washing machines, which had a much higher priority than indoor toilets. A 1939 study of the first cooperative in Indiana showed that after 2 years of service 68% had washing machines, whereas only 10% had utility motors; 45% had vacuum machines but less than 1% had bought any machines for agricultural production. Only during World War II did farmers purchase large amounts of such equipment due to manpower shortages and high crop prices. Productivity and profits rose.

Rural electrification was not implemented to drive people off the land but to keep them there by bringing to the country the amenities enjoyed by the city. Paradoxically, however, rural electrification meant fewer farmers were needed. In the United States, people as diverse as Henry Ford, Frank Lloyd Wright, Lewis Mumford, southern Agrarians, and New Dealers all agreed on the desirability of rural electrification

as a means to regenerating life in the countryside. However, in practice relocation programs and rural electrification instead helped make farmers so productive that many became redundant. In the United States, 40% of the population was on the farm in 1930, 15% remained in 1950, and there was only 8.7% by 1960. The percentage kept declining for the rest of the century and fell below 2% in 2000. This transformation cannot be attributed to electrification alone, of course, because mechanization of farm work and transport also played a role. As farmers moved out, however, others began to move beyond the suburbs into the countryside. Although some were hobby farmers, most wanted to escape the city, but they demanded the urban amenities that only electricity made possible, such as indoor running water, washing machines, vacuum cleaners, radio, and television. After 1975, migration into rural America was greater than out-migration, and the trend shows no sign of abating. Thus, rural electrification underlay not only the decline of the farm but also the deconcentration of cities and suburbs.

Most European countries did not repeat this development to nearly the same degree. During most of the 20th century, cities remained more attractive to most citizens, and until at least 1965 those who commuted were more likely to use bicycles, trains, and trolleys than automobiles. At the same time, European Union subsidies made it possible for a higher percentage of farmers to remain on the land, and national land legislation, notably in The Netherlands, often discouraged sprawl into the countryside. In Europe, where rural electrification was first realized, the electrified city remains the focus of life far more than in the United States, suggesting how much cultural values shape the consumption of electricity. Indeed, overall in the 1990s, a North American family was using twice as much energy as its Western European counterpart.

9. AIR-CONDITIONING

The skyscrapers of Dubai or Mexico City would not be habitable without air-conditioning, which has transformed possibilities of life in large areas of the world that are hot during much of the year. Stuart Cramer coined the term air-conditioning in 1904. He developed humidity controls for Southern textile factories, where excessive dryness made threads break easily, creating havoc in production. Cramer focused on ventilation and humidification more than cooling. Willis Carrier, a Cornell engineering graduate, did the most to promote the acceptance of a precisely controlled indoor environment. His Carrier Air Conditioning Company (1908) sold devices that could regulate both a room's humidity and its temperature, and these spread rapidly from textile mills to food processors and breweries.

Engineers used air-conditioning to increase their authority and control in factories. They first insisted on control of the building and later over the activities within. At times, the new technology displaced skilled workers whose tacit knowledge of materials and their response to different atmospheric conditions previously had been crucial to production of such varied things as textiles, macaroni, candy, beer, and cigarettes.

In the 1920s, air-conditioning spread into stores, skyscrapers, theaters, restaurants, and trains, familiarizing the consumer with its benefits and making people more accustomed to an enclosed, artificial environment. By the 1930s, many firms saw enormous potential for domestic sales and began a fierce competition for a market that only developed after World War II. Many early domestic machines did not regulate humidity but focused on cooling only. As a result, the public perception of air-conditioning remained confused, and in 1938 less than 1 home in 400 used it in even one room. After World War II, however, the industry overcame consumer resistance by working with builders and architects and by achieving economies of scale in producing the equipment.

Air-conditioning was not merely something added to existing structures; its use transformed housing design. Builders partially offset the cost of air-conditioning by eliminating attic fans, window screens, sashes, and high ceilings. New homes were often designed to minimize the purchase price, not the operating costs. Sealed windows combined with the heat produced by appliances, television, and the modern kitchen made air-conditioning a necessity, and the resulting electricity demand eventually helped create the national energy shortages of the 1970s.

In 1955, the federal government adopted a temperature policy that in effect required air-conditioning in most of its new buildings. The Federal Housing Association also permitted new mortgages to include it. Two forms of air-conditioning competed in the marketplace. Selling to both home and workplace, the Carrier Corporation led old engineering firms committed to central systems, whereas General Electric and other appliance manufacturers sold smaller units designed for individual rooms. Central systems began to predominate in hot

cities such as Las Vegas, Dallas, and Bakersfield by the 1960s. Air-conditioning reached 24 million homes by 1970. Thereafter, it rapidly changed from being a luxury to a necessity, creating enormous summer demand for electricity. At office buildings such as the United Nations headquarters, floor-to-ceiling windows increased air-conditioning costs by 50%. By 1982, just after the end of the energy crisis, most office towers and more than 60% of all homes had some form of air-conditioning, half of which were central systems for cooling the entire structure. This was high market penetration considering that this technology was unnecessary in areas such as the Pacific Northwest, the northern tier of states, higher elevations generally, and many coastal communities.

10. ELECTRICITY USE AND THE ENVIRONMENT

Compared to burning wood or coal, electricity seems to be a clean and odorless form of energy. Indeed, for generations electricity provided a better alternative to older, more polluting technologies. Unlike gas and kerosene lamps, electric lights left no soot. Clean electric trains eliminated the cascades of soot from steam locomotives. In factories, electric motors emitted far less oil and raised far less dust than overhead drive systems. Hydroelectric plants create energy with no smoke, but unfortunately falling water provides less than 5% of U.S. energy needs and even less in some other nations. Generating electricity in coal-fired power plants causes pollution, even if it is often invisible to consumers. Such plants are usually located far from the sites of consumption, but their smoke, and the acid rain that eventually results, falls on distant consumers.

Fortunately, during the 20th century utilities developed economies of scale, transforming a larger percentage of coal into energy. Gas-fired plants are even more efficient. During the 20th century, the production of electricity became increasingly efficient, which lowered prices and reduced pollution. In New York City, for example, the average residential customer in 1930 paid $29.23 for 404 kilowatt-hours (kWh) of electricity per year, slightly more than 7¢ per kilowatt-hour. By 1956, the average New Yorker's consumption had almost quadrupled to 1536 kWh, but the cost had only doubled to $61.09. For the nation as a whole by that date, the average consumer used much more, 2969 kWh, but paid only $77.19 for it. Electricity became cheaper because engineers found ways to extract more from a ton of coal or barrel of oil. By the 1970s, however, it was difficult to achieve further production efficiencies. Thereafter, the most cost-effective savings lay not in production but rather in more efficient consumption.

Industrial research therefore focused on developing highly efficient products, notably light bulbs, motors, washing machines, stoves, and refrigerators. As a result, during the two decades after 1973 total U.S. energy use increased only 10%, an achievement based more on technological ingenuity than on changes in lifestyle. People did not use fewer appliances, but their new devices were more efficient. Home refrigerator manufacturers tripled the energy efficiency of refrigerators during these two decades, and by the 1990s washing machine electricity use had been halved. However, efforts to become more efficient resembled the mad queen in Alice in Wonderland, who had to run as fast as she could just to stay where she already was. Even greater efficiencies or lower per capita consumption will be necessary in the 21st century as electricity reaches all the world's population. Recognizing this, in 1994 the World Energy Council proposed the idea of national limits in electricity use. It noted that annual electricity use per capita higher than 6 MWh did not seem to improve life expectancy. It therefore suggested that the 30 highest consuming nations should reduce electrical consumption to this level, whereas the rest of the world should not allow consumption to rise higher than this level. Even if these targets are achieved, however, the increased use of batteries in toys and other consumer products poses another environmental problem because the discarded batteries contain highly poisonous materials. The public tends to discard them in ordinary dumps, allowing lead, cadmium, and other toxic wastes to filter into the groundwater.

Overall, demand for electricity poses environmental problems that lay at the heart of the quest for sustainable development. However, whereas coal and oil-fired power plants contribute to global warming and acid rain, new forms of wind mills, solar power, and passive solar heating promise not only to reduce pollution but also to decentralize power production, moving it closer to the consumer. In the future, as in the past, production of electricity will prove inseparable from its use.

SEE ALSO THE FOLLOWING ARTICLES

Batteries, Overview • *Coal Industry, History of* • *Electrical Energy and Power* • *Electromagnetism* • *Hydropower, History and Technology of* • *Manufac-*

tured Gas, History of • Natural Gas, History of • Nuclear Power, History of • Oil Industry, History of • Refrigeration and Air-Conditioning • Solar Energy, History of • Wind Energy, History of • Wood Energy, History of

Further Reading

Beltran, A., and Carré, P. A. (1991). "La fée et la servante: La sociéte fançaise face á l'electrcité xix–xx siééle." Éditions Belin, Paris.

Hannah, L. (1979). "Electricity before Nationalisation: A Study of the Development of the Electricity Supply Industry in Britain to 1948." Macmillan, London.

Hughes, T. P. (1983). "Networks of Power Electrification in Western Society, 1880–1930." Johns Hopkins Univ. Press, Baltimore.

Marvin, C. (1988). "When Old Technologies Were New." Oxford Univ. Press, New York.

Nye, D. E. (1990). "Electrifying America: Social Meanings of a New Technology." MIT Press, Cambridge, MA.

Nye, D. E. (1998). "Consuming Power: A Social History of American Energies." MIT Press, Cambridge, MA.

Platt, H. L. (1991). "The Electric City." Chicago Univ. Press, Chicago.

Thomas de la Peña, C. (2003). "The Body Electric: How Strange Machines Built the Modern American." New York Univ. Press, New York.

Electric Motors

ANIBAL DE ALMEIDA
University of Coimbra
Coimbra, Portugal

STEVE GREENBERG
Lawrence Berkeley National Laboratory
Berkeley, California, United States

1. How Electric Motors Work
2. Development of Electric Motors and Their Role in Industrialization
3. Current Status of End-Use Applications of Electric Motors
4. Energy Efficiency Trends

Glossary

centrifugal The motion away from the center of rotation. Centrifugal fans, pumps, and compressors use centrifugal action to create pressure differences that result in the flow of fluids.

harmonics Electrical voltages or currents with frequencies that are integral multiples of the fundamental frequency, e.g., 150 cycles/second is the third harmonic in a 50-cycle system.

power factor The ratio between real power (measured in watts or kilowatts) and apparent power (the product of the voltage times the current measured in volt-amperes or kilovolt-amperes), expressed either as a decimal fraction or a percentage (0–100%). In the case of pure sinusoidal waveforms (i.e., not distorted by harmonics), the power factor is equal to the cosine of the phase angle between the voltage and current waves in an AC circuit. This value is known as the "displacement" power factor (because it deals with the displacement in time between the voltage and current). If the power factor is less than 1, more current is required to deliver a unit of real power at a certain voltage than if the power factor were 1. In the case of waveforms that include harmonics, the harmonic current adds to the total current without contributing to the real power, so the power factor is reduced. This higher current is undesirable because the energy lost to heat in the wires supplying power is proportional to the square of the current. In motors and other inductive loads operating in AC circuits, the current wave lags behind the voltage wave.

three-phase alternating current (AC) The flow of electrons back and forth along the conductors making up a circuit; the frequency of the alternation is 50 cycles/second in most of the world and 60 cycles/second in North America. Three-phase AC circuits have three conductors, each having an alternation at the nominal frequency, but the voltage waveforms in the three phases are offset from each other by 120° (equally out of phase around the 360° cycle).

Electric motors, devices that convert electrical energy to mechanical energy, power a wide range of applications in all sectors of the economy. Electric motors take over half of the world's electrical energy as their input. In this article, how the most common types of electric motors work, their development and application, and their trends in energy efficiency are described.

1. HOW ELECTRIC MOTORS WORK

All electric motors convert electrical energy into mechanical energy, typically in the form of a rotating shaft; they operate through the interaction of magnetic fields. The operation of the most important motor types is described in this section.

1.1 Induction Motors

1.1.1 Squirrel-Cage Induction Motors

The vast majority of the motors used in industry are squirrel-cage induction motors (Figs. 1 and 2) due to their low cost, high reliability, and fairly high efficiency. These motors also use more of the total electrical energy than any other motor

 © 2004 Elsevier Inc. All rights reserved.

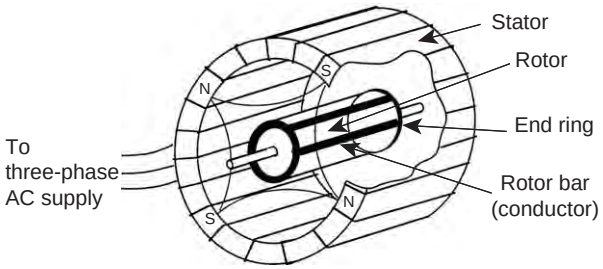


FIGURE 1 Diagram of squirrel-cage induction motor.

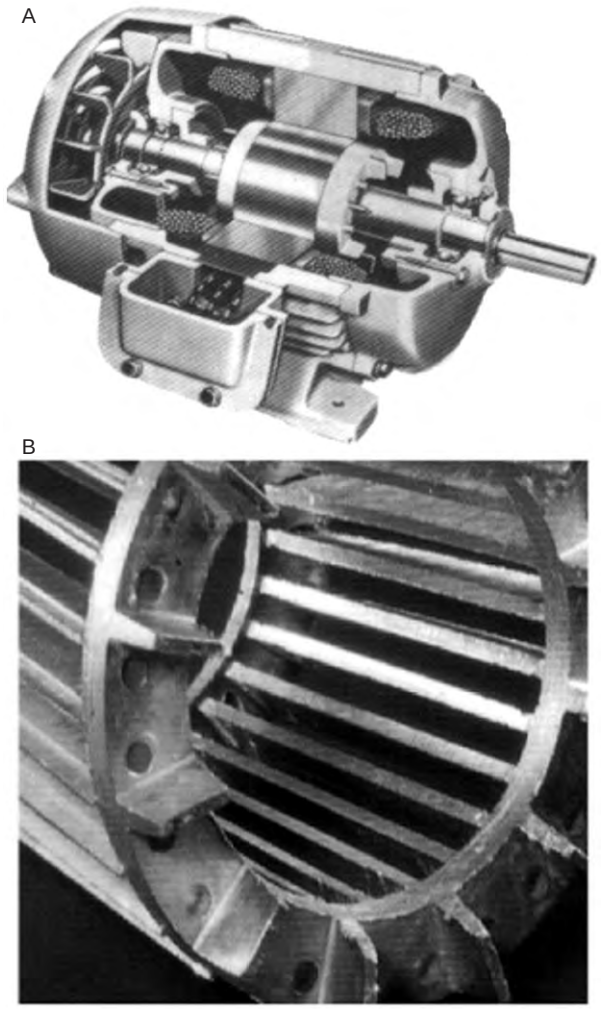


FIGURE 2 Squirrel-cage induction motor. (A) General structure. (B) Rotor squirrel cage (bars and end rings).

type in homes and buildings. The three-phase induction motor, invented in the 1890s, is composed of three sets of stator (the stationary part of the motor) windings arranged around the stator core (a stack of steel laminations with slots for the windings).

There are no electrical connections to the rotor (the rotating part of the motor), which means that there are fewer wearing parts to maintain and replace. When the three stator windings are fed by a three-phase alternating current (AC) supply, a rotating magnetic field is generated, the rotating speed of which is determined by the frequency of the power supply and by the number of magnetic poles of the motor. If the motor is fed by 60 Hz and has two pairs of poles as in Fig. 1, the speed of the rotating field will be 1800 revolutions per minute (rpm). When the rotating field crosses the squirrel-cage rotor (so named because the cylindrical shape of the conducting rotor bars resembles a squirrel cage), the variable magnetic flux induces currents in the rotor bars (hence the name induction motor), which leads to a pulling torque in the direction of the rotating field. The rotor accelerates to a speed close to the speed of the rotating field speed (synchronous speed), but never is able to reach it. If the two speeds were equal, there would not be a change in the magnetic flux (of the stator relative to the rotor), and thus no induced currents, and therefore no torque. The induction motor speed decreases by a few percent when the motor goes from no load to full-load operation. The full-load speed is a function of the motor design and power and is usually in the range of 94–99% of the synchronous speed; the difference is the "slip." Figure 2 shows the typical construction of a squirrel-cage induction motor. The rotor squirrel cage, detached from the steel laminations can also be seen.

1.1.2 Wound-Rotor Induction Motor

Large induction motors can also have a wound rotor. As the name suggests, these motors feature insulated copper windings in the rotor, similar to those in the stator. This type of induction motor was once used in industrial applications in which the starting current, torque, and speed needed to be precisely controlled. In such applications today, squirrel-cage motors with variable-speed drives are much more commonly used.

1.2 Synchronous Motors

Synchronous motors, as the name implies, rotate at a constant (synchronous) speed. The rotor of this type of motor is a wound rotor, which receives the excitation (magnetizing) current from its excitation system (a separate direct current source with controller). Synchronous motors, although they are more costly and require more maintenance compared

to induction motors, are used in applications requiring constant speed (such as in the textile fiber and paper industries), a high degree of operating efficiency, and a controllable power factor. The last two factors are particularly important above 1000 kW. Synchronous motors are often used in large plants to drive the central air compressor. A large synchronous motor can be used to control the power factor of the whole plant, compensating the lagging power factor of a large number of medium and small induction motors.

1.3 Motor Efficiency

It must be emphasized when considering the energy efficiency of electric motors that the motors are part of a system. The loads (cooling, movement of fluids or materials, etc.) imposed on the system are worthy of attention, as well as the motor-driven process or equipment designed to meet those loads: both influence to a large extent the overall efficiency (units produced or service provided/kilowatt-hour) of the system. Automatic controls in industry and other sectors can reduce labor intensity and thus increase productivity, as well as increase energy efficiency.

1.3.1 Motor Losses

Part of the electrical energy supplied to the motor is not converted into useful mechanical energy, ending up instead in losses released to the outside of the motor in the form of thermal energy (Fig. 3). There are four different kinds of losses occurring in a motor: electrical, magnetic, mechanical, and stray. These losses can be reduced by using quality materials as well as by optimizing the design. The electrical losses are of the type I^2R (the product of the square of the electrical current I and the resistance R of the stator and rotor), and they increase rapidly with the motor load (because current and load are roughly proportional). These electrical losses can be decreased by increasing the cross section of the stator and rotor conductors. Magnetic losses occur in the steel laminations ("iron") of the stator and rotor. They can be reduced by increasing the cross section of the iron in the stator and rotor and by using thinner laminations and improved magnetic materials. Mechanical losses are due to friction in the bearings, ventilation, and windage (air-drag) losses. They can be decreased using low-friction bearings and improved fan design. Stray losses are due to leakage flux, nonuniform current distribution, mechanical imperfections in the air gap, and irregularities in the air gap flux density. They can be reduced by optimal design and careful manufacturing.

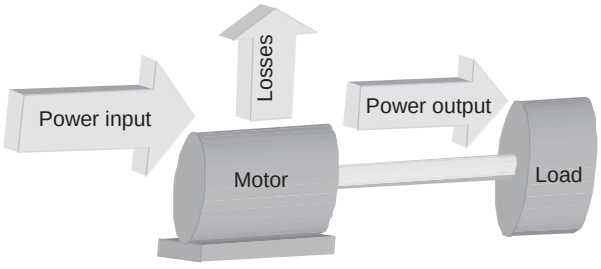


FIGURE 3 Power flow in a motor system (efficiency = output power/input power).

1.3.2 Energy-Efficient Motors

After World War II and until the early 1970s, there was a trend to design motors that minimized the use of raw materials (namely copper, aluminium, and silicon steel). These motors, although inefficient, had lower initial costs and were more compact than previous generations of motors, but due to their low efficiency, their running costs were higher. By the mid-1970s, electricity prices started escalating rapidly. Most of the large motor manufacturers started offering a line of energy-efficient motors in addition to standard-efficiency motors. Efficiency improvements have now been achieved in the standard motor range. There is competition in the market based on motor efficiency. Energy-efficient motors (EEMs) feature optimized design, more generous electrical and magnetic circuits, and higher quality materials. Incremental efficiency improvements are still possible with the use of superior materials (e.g., amorphous silicon steel) and optimized computer-aided design techniques. EEMs typically have 30–50% lower losses (this translates into 2–10% better efficiency, depending on motor horsepower) compared to the equivalent standard motors. EEMs normally carry a price premium of around 20–25% in relation to standard motors. This translates into a price premium of $5–22/kW (U.S. dollars). In a new application, and for a motor with a large number of operating hours, the payback time is normally under 2 years.

Because the reduced losses result in lower temperature rise in the motor, the lifetime of the motor winding insulation, of the bearings, and of the bearing grease increases. Reliability therefore increases, leading to less down time and reduced maintenance costs. Lower losses also provide the following advantages:

- Better tolerance to thermal stresses that typically result from stalls or frequent starting.

- Increased ability to handle overload conditions.
- Better resistance to abnormal operating conditions, such as under- and overvoltage or phase unbalance.
- Higher tolerance to poorer voltage and current wave shapes (e.g., harmonics).
- The ability to use smaller cooling fans, allowing quieter motor operation.

1.3.3 Proper Motor Sizing and Selection

Motors rarely operate at their full rated load. Field tests in the United States and Europe indicate that, on average, motors operate at around 60% of their rated load. Induction motor efficiency typically peaks near 75% of full load and is relatively flat down to the 50% load point. Motors in the larger size ranges can operate with reasonably high efficiency at loads down to 30% of rated load. In Fig. 4, the efficiency and power factor values at partial load points are shown. The adverse effect of the reduction of power factor due to oversizing is often neglected. Motor oversizing is a major reason for the low power factor of some industrial operations. Unless the reactive power is compensated for each motor, the additional line losses from higher currents due to oversizing are an additional incentive for proper motor selection. The original equipment manufacturers (OEMs), who are responsible for choosing the motor in the first place, as well as the users, who influence the buying phase or the replacement of failed motors, should consider that the design criteria leading to oversizing can result in higher first and operating costs, especially when a motor has a working point far below 100% of rated power. Oversizing can also lead to an increase in the mechanical power used by the load. The operating speed is dependent on the applied load, and the loading imposed on a motor is in turn dependent on its size. For example, a 25% loaded 100-kW motor could be replaced by a 50-kW motor with a 50% load factor or a 25-kW motor with a 100% load factor. The oversized motor operates at higher speed, due to having a smaller slip. For centrifugal fans and pumps, even a minor change in the motor's operating speed translates into a significant change in imposed load and annual energy consumption. For most fans, the belt drive between the motor and the fan can be changed to obtain the correct fan speed; most pumps are direct drive, so the motor size, slip, and efficiency must all be carefully considered to make the best selection.

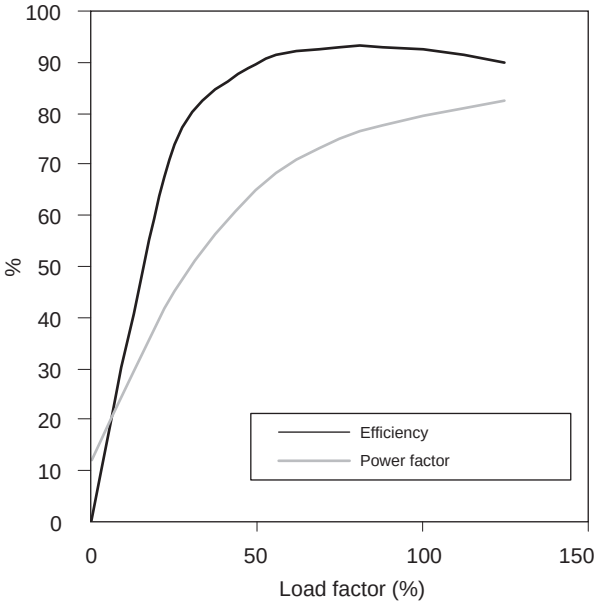


FIGURE 4 Typical efficiency and power factor versus motor load.

2. DEVELOPMENT OF ELECTRIC MOTORS AND THEIR ROLE IN INDUSTRIALIZATION

2.1 Early Motors

The first practical electric motors date back to the late 1800s. They were driven by direct current (DC) supplied by nearby generating plants. The early DC systems were superseded by alternating current, due to the ease of transmitting AC power generated centrally and to the simplicity of the AC induction motor, which was invented in the 1890s. Soon, DC motors were relegated to niche roles, especially when speed and motion control were needed (e.g., elevators and electric transportation).

2.2 Motor Evolution

2.2.1 Effect of Materials and Technology

The ability of the original motor winding insulation systems to withstand high temperatures was limited, so motors needed to be large, efficient, or both in order to keep the insulation from being damaged. As synthetic insulation materials were developed in the mid-20th century, higher operating temperatures were acceptable, allowing motors to be designed with lower operating efficiency, as noted in Section 1. In the 1970s and later, with renewed interest in efficiency, special steels were used in the magnetic

core to reduce energy losses there; computer-aided design and other techniques were used to design and build more efficient motors. The advent of inexpensive power electronics in the late 20th century made controlling the speed of AC motors easy and made DC motors with permanent magnets practical and efficient in applications ranging from residential furnace fans to drive motors for electric cars (see Section 4). Ongoing work includes motors that use high-temperature superconductors in the windings to further increase efficiency and motors that are extremely small (so-called nanomotors, because their size is on the scale of a nanometer).

2.2.2 Energy and Labor Considerations

For most of the 20th century, decreasing electricity costs and the evolution in insulation materials together put downward pressure on motor first cost and efficiency. Decreasing motor first cost and increasing labor rates meant that repairing motors in the United States and Europe gave way to simple replacement, especially for small motors. The electric motors that drive factory and construction equipment have been fundamental contributors to increases in worker productivity.

2.3 Role in Industrialization

2.3.1 Fluid Motion

The movement of fluids (collectively, liquids and gases) is the largest application for electric motors throughout the world.

2.3.1.1 Fans Fans are used for moving air at relatively low pressures; they are sized from a few watts for cooling small computers to megawatts for forcing combustion air through boilers at electric power plants. Fans are commonly used in systems that heat, cool, and ventilate buildings in which people live, learn, and work.

2.3.1.2 Pumps Pumps are used for moving liquids, most commonly water; they are sized from a few tens of watts for circulating water for heating homes to tens of megawatts for supplying drinking water to large cities. Industry uses large numbers of pumps for moving cooling water and a wide variety of liquid chemicals.

2.3.1.3 Compressors The most common use of compressors in industry is for making compressed air for operating air-powered (pneumatic) tools and processes. A very common type of compressor through all sectors of the economy is the refrigeration compressor, used for ice making, air-conditioning, and food refrigeration. The scale of these compressors ranges from below 100 W for small household refrigerators to over 1 MW for large building air-conditioning systems.

2.3.2 Materials Transportation

Electric motors power conveyor belts, slurry pipelines, and pneumatic and air-blown systems used in the transport of all types of materials.

2.3.3 Materials Processing

In industry, materials are processed by a wide variety of electric motor-driven equipment such as grinders, saws, drills, crushers, stampers, shears, and lathes.

2.3.4 Transportation

Electric motors are widely used throughout the transportation sector. Electric railroad locomotives and diesel–electric locomotives both use electric motors in the low-megawatt size range. Likewise, streetcars and subways typically use DC motors, though there is a trend toward the use of AC induction motors with inverter drives. Electric and hybrid electric automobiles use AC induction motor/inverter drives or permanent magnet motors in the range of tens of kilowatts. Elevators, escalators, and people movers provide vertical and horizontal transportation in buildings, all powered by electric motors. Many ships are propelled by electric motors, supplied by onboard generating plants.

3. CURRENT STATUS OF END-USE APPLICATIONS OF ELECTRIC MOTORS

3.1 Common End Uses

3.1.1 Industrial Sector

In the industrial sector, motor-driven loads use about three-quarters of the total sector electricity use (74% of 2007 TWh/year for the United States and European Union combined). Common end uses are dominated by pumps and material processing, with air compressors and fans making significant contributions (see Fig. 5).

3.1.2 Commercial Sector

In the commercial sector (known as the tertiary sector in Europe), 37% of the electricity is used in

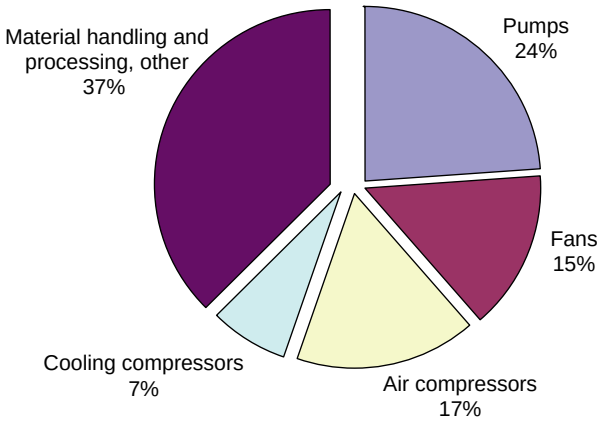


FIGURE 5 Industrial sector motor electricity use in the United States and European Union. Data from Nadel *et al.* (2002) and the University of Coimbra.

motor systems (out of a sector total of 1540 TWh/year for the United States and European Union combined); the dominant end uses are fans, air-conditioning, and refrigeration (see Fig. 6).

3.1.3 Residential Sector

In the residential sector, the dominant end uses in the United States are space conditioning (heating, fans, and air-conditioning) and refrigeration; motor-driven end uses require 38% of the residential sector's total of 1192 TWh/year (see Fig. 7). In Europe, a slightly higher fraction of the residential electricity is estimated to flow into motors.

3.2 Evolving End Uses

3.2.1 High-Performance Motion Control

In many industrial applications (e.g., robotics), there is increasing use of high-performance permanent-magnet or switched-reluctance drives (see Section 4 for more on these drives). With the advent of vector drives (inverters that separately control the speed and torque of AC induction motors), many applications (e.g., elevators) that were the stronghold of DC motors are giving way to simpler and more efficient AC motors.

3.2.2 Increasing Use of Speed Control

The cost of inverters for speed control (see Section 4) is continuing to decrease, and the reliability is increasing as the circuitry becomes more integrated (more functions on fewer components). As a result, the controls are becoming standard in many applications in homes, commercial buildings, and industry

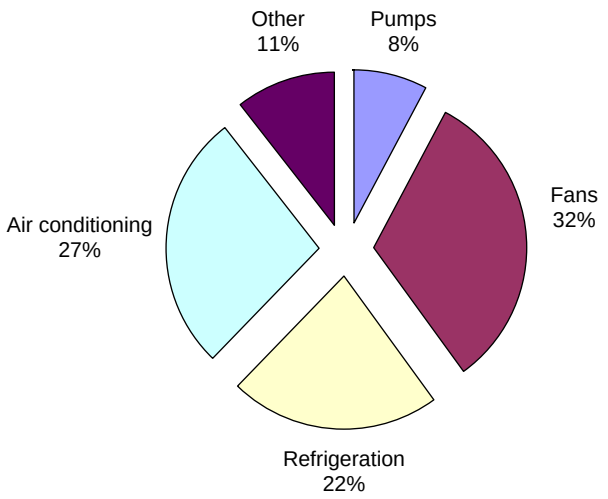


FIGURE 6 Commercial (tertiary) sector motor electricity use in the United States and European Union. Data from Nadel *et al.* (2002) and the University of Coimbra.

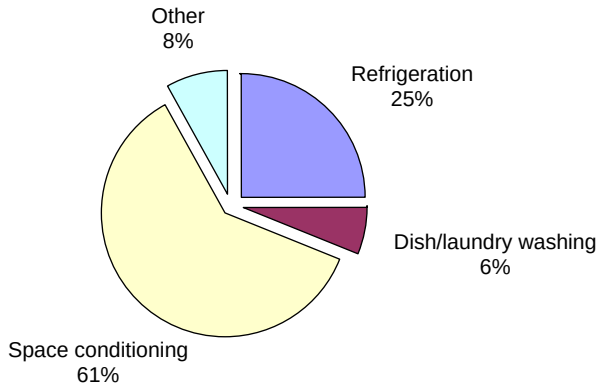


FIGURE 7 Residential sector motor electricity use in the United States. Data from Nadel *et al.* (2002).

that formerly used on/off cycling or throttling to match output to load. Examples include residential furnace fans, ventilation fans, and air-conditioning compressors in office buildings, and many industrial applications. A continuing challenge is the control hardware and especially software to take full advantage of the improved process control and energy-savings opportunities offered by speed control.

3.2.3 Electric Transportation

3.2.3.1 Rail Due to the availability of powerful and compact inverters, there is an increasing tendency to use AC induction motors with inverter drives rather than DC drives. In the future, high-speed rail will include magnetic levitation trains (using magnetic repulsion to float the cars above the

guideway) that use a linear motor to propel the cars along the guideway.

3.2.3.2 Road Electric and hybrid electric vehicles use sophisticated water-cooled drives that include regenerative braking to increase energy efficiency. Some designs use AC induction motors and some use permanent-magnet motors. These will continue to evolve as hybrid vehicles move from internal combustion engines to fuel cells as the primary power source.

3.2.3.3 Air The National Aeronautics and Space Administration in the United States is flying unmanned airplanes powered by numerous electric motors that gather their electricity from onboard solar–electric panels and fuel cells. These can fly at very high altitudes for long periods and could eventually perform many of the functions of satellites at much lower cost.

4. ENERGY EFFICIENCY TRENDS

4.1 Technology

4.1.1 Permanent-Magnet Motors

In the last decades of the 20th century, new technologies led to the introduction of new motor types that have some performance and efficiency advantages over induction motors. Permanent-magnet (PM) motors have a stator winding configuration similar to the three-phase induction motors, but they use permanent magnets in the rotor instead of a squirrel cage or wound rotor. The permanent-magnet rotor tracks with synchronism the stator rotating field, and therefore the rotor speed is equal to the rotating magnetic field speed. A wide variety of configurations is possible, and the two-pole version can be seen in Fig. 8. Motors of this sort have output ranging from about 100 W up to 100 kW. The key advantage of the permanent-magnet motor is that no excitation power is required for the rotor and therefore its efficiency is higher, compared to the induction motor. Early permanent-magnet motors, besides being very expensive, suffered from the tendency of the magnets to be demagnetized by the high stator currents during starting, and from a low maximum allowable temperature. Much improved versions using rare-earth magnets have been developed since the 1970s to overcome these problems. Rare-earth alloys (neodymium, iron, and boron),

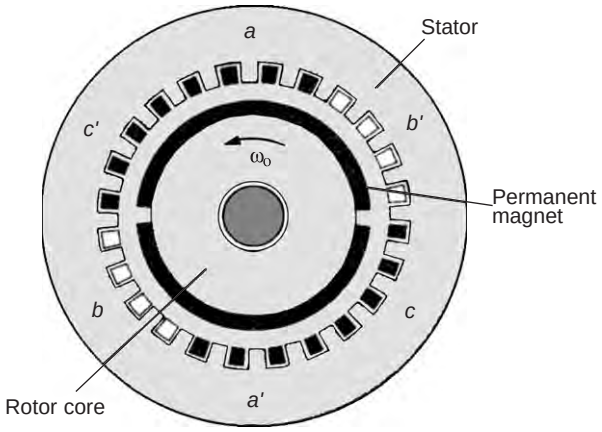


FIGURE 8 Permanent-magnet motor: two-pole version. See text for discussion.

developed in the mid-1980s, have allowed an increase in performance with a decrease in costs. For starting from a fixed-frequency supply, a rotor cage is required; such motors are referred to as "line-start" motors. Because there are no electric and magnetic losses in the rotor, there is less heating than in a conventional motor, so higher specific outputs can be achieved. The rotor inertia can also be less than that of an induction motor rotor, which means that the torque/inertia ratio is better, giving a higher acceleration. The torque-to-weight ratio, the steady-state efficiency, and the power factor at full load of PM motors are in most cases better than those of the equivalent induction motor, and they can pull in to synchronism with inertia loads of many times rotor inertia.

Although better efficiency and low-speed performance can be obtained with permanent-magnet synchronous motors, there would need to be a significant reduction in the cost of rare-earth permanent-magnet material for these motors to the replace induction motors in most applications. Permanent-magnet motors are in most cases used with electronic speed controls, normally called brushless DC (BLDC) motors. In the BLDC motor, the stator winding currents are electronically commutated (switched) by digital signals from simple rotor position sensors; these motors are also known as electronically commutated motors (ECMs) The stator winding also creates a rotating field that creates a rotating torque by pulling the permanent-magnet rotor. This combination permits the motor to develop a smooth torque, regardless of speed. Very large numbers of brushless DC motors are now used, particularly in sizes below 10 kW. The small versions (less than 0.75 kW) are

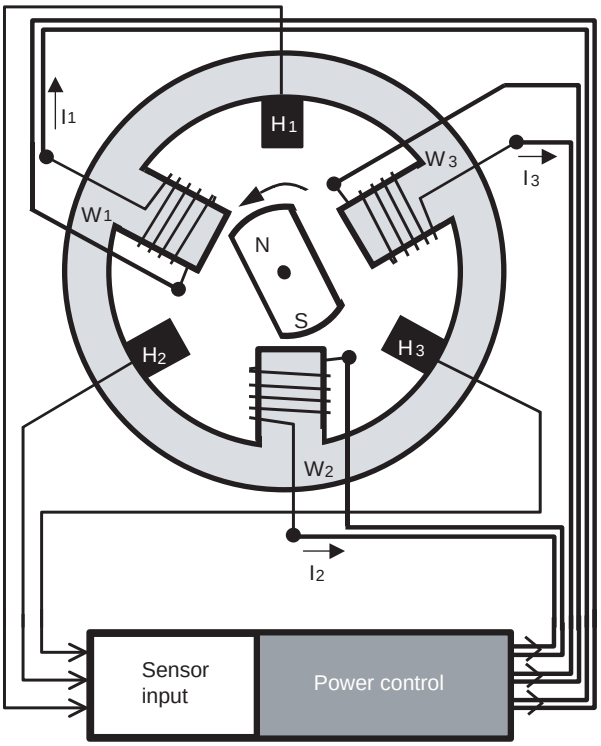


FIGURE 9 Control system scheme of a brushless DC motor. H_1, H_2, and H_3 are magnetic position sensors.

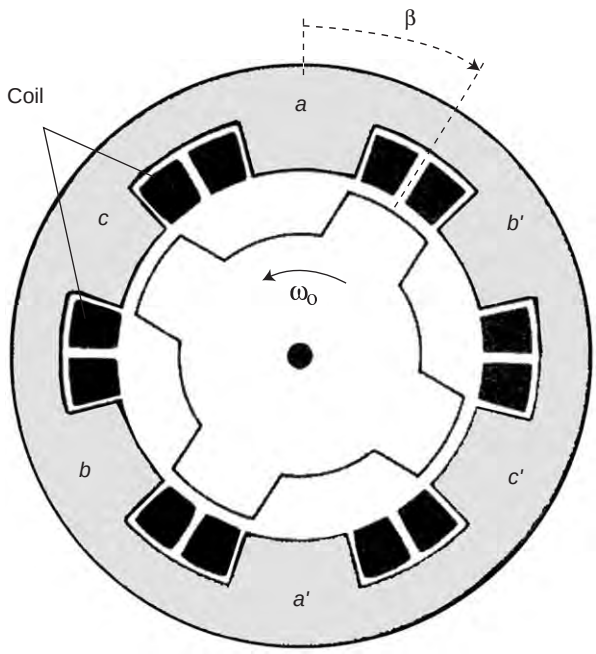


FIGURE 10 Switched-reluctance motor configuration.

increasingly made with all the control and power electronic circuits integrated at one end of the motor. The typical stator winding arrangement and the control scheme are shown in Fig. 9.

Many brushless motors are used in demanding servo-type applications (e.g., robotics and high-performance systems), in which they need to be integrated with digitally controlled systems. For this sort of application, complete digital control systems that provide torque, speed, and position control are available. New applications such as energy-efficient elevators with direct-drive (gearless) permanent-magnet motors are also entering the market.

4.1.2 Switched-Reluctance Motor

The switched-reluctance (SR) drive is a recent arrival (ca. 1990s) on the drives scene, and can offer advantages in terms of efficiency, power density, robustness, and operational flexibility. The drawbacks are that it is relatively unproved, can be noisy, and is inherently not well suited to smooth torque production. Despite being recent, SR technology has been successfully applied to a wide range of applications, including general-purpose industrial drives, traction, domestic appliances, and office and business equipment.

In the switched-reluctance motor, both the rotor and the stator have salient (protruding) poles. This doubly salient arrangement proves to be very effective as far as electromagnetic energy conversion is concerned. The stator has coils on each pole, the coils on opposite poles being connected in series. The rotor, which is made from machined steel, has no windings or magnets and is therefore inexpensive to manufacture and extremely robust. The example shown in Fig. 10 has six stator poles and four rotor poles, and represents a widely used arrangement, but other pole combinations are used to suit different applications. The motor rotates by energizing the phases sequentially in the sequence a–a', b–b', c–c' for clockwise rotation or a–a', c–c', b–b' for counter-clockwise rotation, the "nearest" pair of rotor poles being pulled into alignment with the appropriate stator poles by reluctance (the alignment of a metal object with a magnetic field) torque action. In this way, similar to PM motors, the rotor synchronously tracks the stator rotating magnetic field.

As with the PM motor, the SR motor has no electric or magnetic losses in the rotor. Therefore, the overall efficiency is generally higher compared to induction motor efficiency. The SR motor is designed for synchronous operation, and the phases are switched by signals derived from a shaft-mounted rotor position detector. This causes the behavior of the SR motor to resemble that of a BLDC motor.

Because the direction of torque in a reluctance motor is independent of the direction of the current, it means that the power converter can have fewer switching devices than the six required for three-phase bipolar inverters used in BLDC motor. Some of the early SR motors were deemed to be very noisy, but improved mechanical design has significantly reduced the motor noise.

4.1.3 Speed Controls

Most induction AC motors operate at fixed speed. However, most motor applications will benefit if the motor speed is adjusted to match the process requirements. Motor speed controls are the devices that, when properly applied, can tap most of the potential energy savings in motor systems. Motor speed controls are particularly attractive in applications in which there is a variable fluid flow. In many centrifugal pump, fan, and compressor applications, the mechanical power grows roughly with the cube of the fluid flow; e.g., to move 80% of the nominal flow, only half of the power is required. Centrifugal loads are therefore excellent candidates for motor speed control. Other loads that may benefit from the use of motor speed controls include conveyers, traction drives, winders, machine tools, and robotics.

Conventional methods of flow control used inefficient throttling devices such as valves, dampers, and vanes. These devices, although they have a low initial cost, introduce high running costs due to their inefficiency. Several speed-control technologies can be used to improve motor system operation. Electronic variable-speed drives (VSDs) are discussed in detail here because they are the dominant motor speed-control technology.

As was mentioned in Section 1.1, the speed of the rotating field created by the induction motor stator windings is directly linked with the frequency of the voltage applied to the windings. Electronic variable-speed drives can produce variable-frequency, variable-voltage waveforms. If these waveforms are applied to the stator windings, there will be a shift of the torque–speed curve, and the same slope of the linear operation region of the curve. In this way, the motor speed is proportional to the applied frequency generated by the VSD (Fig. 11). Figure 12 shows the general configuration of most VSDs. The three-phase alternating current supply is initially converted to direct current, then filtered, and, finally, the DC/AC inverter converts the DC voltage to the variable voltage and variable-frequency output applied to the motor. The adjustment of the motor speed through the use of VSDs can lead to better process control,

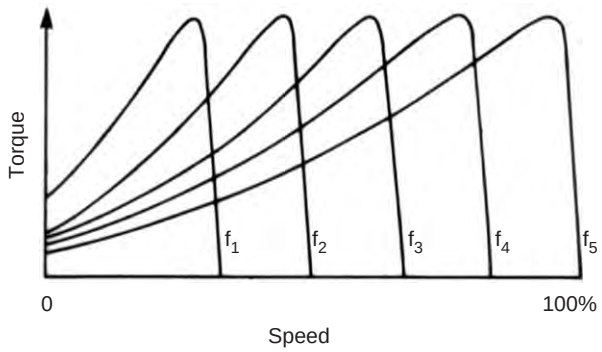


FIGURE 11 Speed–torque curves for an induction motor ($f_1 < f_2 < f_3 < f_4 < f_5$ and $f_5 =$ normal line frequency).

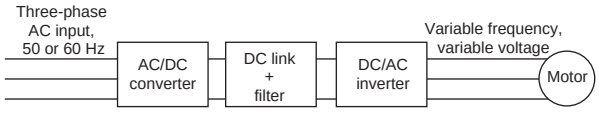


FIGURE 12 General configuration of inverter-based variable-speed drives.

less wear in the mechanical equipment, less acoustical noise, and significant energy savings. However, VSDs can have some disadvantages, such as electromagnetic interference (EMI) generation, current harmonics introduction into the supply, and the possible reduction of efficiency and lifetime of old motors.

There are five basic types of power electronic VSDs: two voltage-source (pulse-width modulation and six-step) types, two current-source (force-commutated and load-commutated) types (typically for large motors), and cycloconverters (typically for large-motor, low-speed applications). Of these, the pulse-width-modulated VSD has come to dominate the market.

4.1.3.1 Pulse-Width-Modulated VSDs The pulse-width-modulated (PWM) voltage source inverter is widely accepted as giving the best overall performance below power levels of 1 MW. The PWM inverter maintains a nearly constant DC link voltage, combining both voltage control and frequency control within the inverter. The objective of the sinusoidal PWM is to synthesize the motor currents as near to a sinusoid as economically possible. The lower voltage harmonics can be greatly attenuated, and therefore the motor tends to rotate more smoothly at low speed. Higher order harmonic motor currents are limited by the motor inductance (the tendency to resist changes in current). Torque

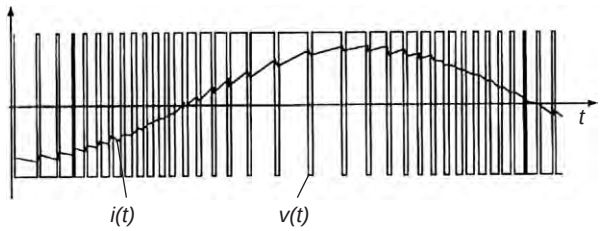


FIGURE 13 Illustration of a sinusoidal pulse-width modulation technique.

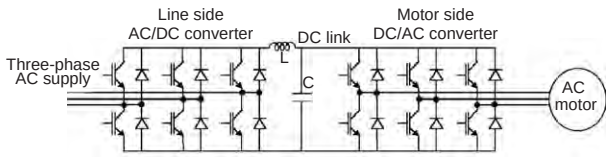


FIGURE 14 Power circuitry of a voltage source inverter pulse-width modulator with regenerative capacity and power factor control. Whereas conventional VSDs use diode rectifiers in the input stage, regenerative units use insulated gate bipolar transistors (IGBTS) at both the input and the output stages to allow bidirectional power flow.

pulsations are virtually eliminated and the extra motor losses caused by the inverter are substantially reduced (see Fig. 13).

The advantages of the PWM VSD are the good displacement power factor throughout the speed range, low distortion of the motor current, wide speed range, and, with proper circuitry, regeneration capability, which allows the motor to absorb energy from the load, generate electrical power, and feed it back into the electrical supply. When the power rating is large and/or when the motor may have to operate for long periods of time in the regenerating region, the line-side converter can be modified to allow regeneration, feeding power back to the line by reversing the direct current in the DC link. New (ca. 2000) PWM circuitry (Fig. 14) allows for the braking energy to be injected back to the source. This feature can be a way of saving a significant amount of energy in applications with high speed/high inertia (e.g., centrifugal separators) and/or long time operations in the braking mode (e.g., elevators, downhill conveyor belts).

4.2 Policies to Promote Energy-Efficient Motor Technologies

Experience from many energy-saving initiatives around the world shows that the most successful programs to achieve the potential savings are based on a combination of technical information, financial incentives, voluntary agreements, and, in some cases, the use of regulation. Strategies involving education, technical support, financial aid, and legislative standards can be pursued.

4.2.1 Education and Training

More education on energy efficiency is needed on many fronts. The information gap should be filled through several means:

- Media, including advertising, trade press releases, direct mail, electronic mail, Web site, and exhibitions; the mix of these depends mainly on the available budgets, with care taken to examine the possible impacts.
- Publications, including a selection of materials from short and simple introductory leaflets and brochures to guides, handbooks, journals, and software; detailed technical information may be needed to satisfy more experienced personnel. Literature should be issued and distributed by manufacturers, trade associations, distributors, energy agencies, etc.
- Training and workshops.
- Dissemination of educational materials, not only through existing energy agency databases and via the media (paper-based products, software programs, and Internet information), but also third parties such as trade associations and equipment suppliers, groups with tremendous potential leverage.

4.2.2 Provision of Technical Assistance

Decision support tools, such as user-friendly software tools, are required for more standardized comparison of efficiency and economics of system configurations. These include support tools for product selection, such as the U.S. Department of Energy's Motor Master and (for Europe) the European Database of Energy-Efficient Motors (EURODEEM) for motors, guides that help users to select particular equipment for an application.

On-site assessment and assistance can provide motor users with information describing their savings potential. Energy-saving publications frequently use comprehensive measurements to ensure the reliability of the figures presented, but in practice most site personnel will have the time and resources only for much simpler measurements. Demonstration projects and pilot actions can be very helpful. One key factor that limits the adoption of energy-efficient technologies is the scepticism about their viability. People doubt equipment durability, performance, and real savings. There already are many demonstration projects and case studies on the

application of EEMs and VSDs; additional demonstration projects aimed at new technologies as well as easier access to the results of existing actions are needed.

4.2.3 Financial Tools

Tax incentives, including tax reduction and accelerated depreciation on efficient equipment, e.g., with VSDs, would be very effective measures to promote efficiency. Another market mechanism for improving energy efficiency involves the development of the energy service company (ESCO) industry. Energy service companies are often privately owned, nonregulated organizations that contract with a utility or an end user to assess, finance, and install energy conservation measures. The financial mechanism used by ESCOs for encouraging the use of their services is performance contracting. The ESCO takes as payment a share of the energy cost reduction, based on either engineering estimates or measurements. Yet another mechanism, rebates, is now seen primarily as a short-term measure to help stimulate the market for energy-saving products; this served in North America as a very useful precursor to legislation on minimum motor efficiency standards.

4.2.4 Efficiency Standards

The establishment of minimum efficiency standards for motors, for end-use devices, and for motor systems is an important market transformation tool. These standards are particularly effective for OEMs, obliging them to incorporate efficient devices in their equipment. A case in point in the United States is the mandatory motor minimum efficiency standards (Energy Policy Act), enforced in October 1997, as well as minimum efficiency standards for appliances dominated by motor use (e.g., refrigerators and air conditioners).

SEE ALSO THE FOLLOWING ARTICLES

Batteries, Overview • Batteries, Transportation Applications • Electrical Energy and Power • Electricity Use, History of • Electromagnetism

Further Reading

De Almeida, A., Fonseca, P., Falkner, H., and Bertoldi, P. (2003). Market transformation of energy-efficient motor technologies in the EU. *Energy Policy* **31**(6), 563–575.

De Almeida, A., Fonseca, P., and Bertoldi, P. (2003). Energy efficient motor systems in the industrial and in the services sectors in the European Union: Characterization, potential barriers and policies. *Energy Int. J.* **28**(7), 673–690.

European Commission (2000). "Improving the Penetration of Energy-Efficient Motors and Drives." Project report, Institute of Systems and Robotics—University of Coimbra. European Commission, Brussels.

Institute of Systems and Robotics (ISR)—University of Coimbra. (2001). "VSDs for Electric Motor Systems." Project report, ISR—University of Coimbra, May 2001. Coimbra, Portugal.

Joint Research Center (JRC), European Commission (2004). European Database of Energy-Efficient Motors (EURODEEM) and Motor Challenge Programme. Information available on the JRC Web site at http://www.energyefficiency.jrc.cec.eu.int/eurodeem/.

Nadel, S., Elliot, N., Shepard, M., Greenberg, S., Katz, G., and de Almeida, A. (2002). "Energy-Efficient Motor Systems." 2nd ed. American Council for an Energy-Efficient Economy, Washington, D.C.

U.S. Department of Energy, Energy Information Administration (EIA). (2004). "Annual Electricity Sales by Sector." Available on the DOE Web site at http://www.eia.doe.gov/.

U.S. Department of Energy, Office of Industrial Technologies (OIT). (2004). Motor programs available on the OIT Web site at http://www.oit.doe.gov.

U.S. Environmental Protection Agency (EPA) and U.S. Department of Energy (DOE). (2004). Energy Star award rules and instructions; continuously updated information available on the EPA/DOE Web site at http://www.energystar.gov.

Electric Power: Critical Infrastructure Protection

ALEXANDER E. FARRELL
University of California, Berkeley
Berkeley, California, United States

HISHAM ZERRIFFI
Carnegie Mellon University
Pittsburgh, Pennsylvania, United States

1. **Defining the Problem**
2. **Turning Out the Lights**
3. **Other Attack Modes**
4. **Mitigation**

Glossary

cascading failure A disruption at one point in a network that causes a disruption other points, leading to a catastrophic system failure.

critical infrastructure protection (CIP) Activities and procedures designed to avoid, prevent, detect, resist, and recover from deliberate attacks on electric power systems.

distributed generation (DG) Electricity production near point of use, typically with small (<10 MW) generators using natural gas. One form of embedded generation.

energy management system (EMS) System of sensors, actuators, communications, and computation hardware and software used to operate an electric power system. Key functions include generation scheduling and control, transmission system analysis and control, and operator training.

outage A temporary suspension of electric service; a loss of electric power.

security The ability of electric power systems to withstand accidental, sudden disturbances such as electric short circuits or unanticipated loss of system elements.

stress Electric power system conditions more severe than those considered in reliability planning analysis, that is, outside of typical operating conditions in industrialized countries. These may include deliberate attacks, multiple failures, persistent failures (or attacks), and conditions that make system restoration difficult (i.e., a lack of spare parts).

survivability The ability of a system to fulfill its mission in a timely manner, despite attacks, failures, or accidents by detecting, resisting, and recovering from attacks. Survivable systems are likely less vulnerable to stress.

The rise of terrorism around the world and the persistence of inadequate investment or maintenance in some areas have given rise to concerns about how to safeguard key infrastructures (energy, communications, banking, roads, etc.). Electric power systems are among the most difficult to protect because of their complexity. The concept of critical infrastructure protection is still relatively new and is likely to evolve over time, possibly away from a "guards, gates, and guns" defensive approach and toward a design approach that yields systems that are inherently harder to successfully attack. Such survivable systems may feature distributed intelligence, control, and operations.

1. DEFINING THE PROBLEM

Deliberate attacks on electric power systems occur around the world, although their character and magnitude vary greatly. In the industrialized world, such events are small and infrequent; they pale in comparison to the largest weather-related events. Modern war-fighting includes tactics and weapons designed to damage electric power systems, but preparation for war is outside of the power system planning processes. Thus, attention paid to deliberate attacks on electric power systems is relatively scarce. However, the growth of terrorist acts within

 © 2004 Elsevier Inc. All rights reserved.

developed countries since the late 20th century has raised concerns about critical infrastructure protection (CIP) for power, fuel, telecommunications, transportation, and other systems.

Even less attention is paid to the reconstruction of power systems damaged by conflict and in places where political tensions may result in frequent deliberate attacks in the future. The most obvious terms for these issues, reliability and security, have different, well-established meanings in electricity system planning. Thus, we will use CIP to refer to activities and procedures designed to avoid, prevent, detect, resist, and recover from deliberate attacks on electric power systems. The understanding of the CIP implications for electric power systems is rather limited, and the ability to manage the threat of deliberate attacks even more so. Significant research and development efforts are expected over the coming decade, with subsequent deployment of new approaches and technologies.

Electricity cannot be stored economically, so generation and consumption must be matched at every instant. In addition, electric power systems must be managed so that several other parameters (including voltage, reactive power, and frequency) remain within specifications at all times. Not only can these parameters change extremely quickly, electric power systems are typically large interconnected grids that are vulnerable as a system due to the possibility of cascading failures in which a small number of failures in a network cause other failures, leading to widespread disruption and outages. Another key feature of electric power systems is that they are inherently vulnerable to attack since they extend over great distances, often through relatively isolated areas. Further, transmission systems are extremely complex with many custom-designed and hard-to-replace components, including transformers, switches, and relays. These components are monitored and operated by technicians both in the field and at central control centers using electronic systems while other groups form electricity markets that use different computer networks. These applications open up the possibility of cyber attack on electric power systems.

To better understand CIP, it is necessary to introduce three key concepts. The first is reliability and how electric power system planners have dealt with normal failures in the industrialized countries. The second concept is stress, and why planning for reliability must change when the system is under stress. The third concept is cascading failures of the electric power system and how they relate to reliability planning and stress.

1.1 Reliability

These constraints make CIP a greater challenge in electric power systems than in any other part of the energy sector. However, the necessity of dealing with these constraints during normal operations has result in a long history of reliability planning, which includes consideration of both normal operation under reasonable expectations of intended operating conditions (adequacy) and of limited disturbances due to weather and occasional equipment failure that are unplanned but not unexpected (security). These conditions fairly well describe most conditions in high-income countries, but they do not describe a variety of situations now considered important in both high- and low-income countries (see Table I). Frequent deviations from intended operating conditions in industrialized countries is defined as a condition of stress, which includes deliberate attacks, multiple failures, persistent failures or attacks, and conditions that make system restoration difficult (e.g., a lack of spare parts). Cascading failures are still largely excluded since stress is a planning concept and concerned with improving reliability under stress, not an operating concept, unless the stress is directly linked to a change in the probability and magnitude of cascading failures.

Generally, reliability concepts are thought of first when CIP is considered by the electric power industry. However, traditional concepts related to electricity reliability may not be appropriate for achieving protection of critical infrastructures from deliberate attack. There are two main reasons for this; first, the assumptions made in reliability analysis about single point failures may not be realistic; second, some attack modes do not necessarily result in service outages, which is the focus of reliability planning.

In the past, reliability concerns have been handled from the supply side under the fairly restrictive assumption of a single mode failure. The goal that planners adopted was to protect generators and transmission and distribution networks from the ill effects of an equipment failure or natural disturbance (e.g., a tornado). Thus, practice is to have available at all times more capacity than actually needed in order to deliver power to the end users under both normal and a limited set of abnormal conditions. The reserve capacity typically required under the single failure mode assumption is sufficient for the largest single generator or transmission system component to fail without causing either damage to other equipment or a loss of power to any customers. This rather brute force, reliability-through-redundancy approach has

TABLE I

Modes of Disturbance in Electric Power Systems

Mode of disturbance	Possible causes	Likely characteristics	Likely impacts
Weather related damage[a]	Hurricanes, tornadoes, floods, ice storms	Random, not repeated, not targeted, regional	Impacts T & D primarily. No long term impacts on failure probabilities, magnitudes, or durations; recovery only hampered by environmental conditions
Equipment failure[a]	Trees, animals, or vehicles contacting power lines, unanticipated wear	Single events occur randomly, but at a known, acceptable rate	Reduces the equipment available to the system by one component; managed to obtain desired balance of cost and reliability
System-wide direct conflict damage	Civil war (e.g., Bosnia), guerilla movement	Persistent, system-wide, impacts all levels of system	Both failure probabilities and magnitude of damage high, recovery difficult and expensive due to continuing conflict
Regional direct conflict damage	Regional insurgency	Persistent but localized, impacts all levels of system	Failure probabilities and magnitudes increase in affected region, recovery difficult
Localized direct conflict damage	Terrorism or sabotage	Targeted, repeated (lower frequency), less damage per attack on average, less damage to large generators	Failure probabilities increase, magnitudes do not increase greatly except for the most extreme acts, recovery relatively unhampered
System-wide indirect conflict damage	Civil war (e.g., Bosnia), guerilla movement	Mobility hampered, increased nontechincal losses creating financial problems	Failure probabilities increase, magnitudes of failures do not increase, recovery more difficult
Regional indirect conflict damage	Regional insurgency	Regional mobility hampered, increased nontechnical losses, financial problems	Failure probabilities increase, magnitude of failures do not increase, recovery more difficult
Lack of investment in new capacity	Capital access, investment uncertainty	Units need to be run more often and for longer as reserve margins decline, transmission system strained	Possible increase in failure rates over time
Poor maintenance	Capital and spare parts access		Failure rates increase over time, repair times increase

[a] These disturbances are part of reliability planning. All others are stress conditions.

proven effective in the past, but its efficacy is challenged both by the advent of restructuring and the potential for terrorist or other military attacks.

The rationale for a supply-side approach to reliability was that by protecting the electricity supply system; the concerns of most end users would automatically be accommodated. These concerns were quantified with reliability indices. However, these indices do not differentiate among the very different needs of the end users, nor are supply-side measures alone likely to meet the needs of all customers. For instance, a sag in voltage that lasts no more than a fraction of a second would go unnoticed by most end users but could be a serious event for some. Eliminating such sags from the supply side is difficult, if not impossible. Similarly, procedures for measuring indices ignore outages less than a minute long, even though a few milliseconds of lost power may be sufficient to interrupt or damage electronic equipment.

No matter what the cause, power outages impose significant costs on electricity users, while there are

significant costs to electricity companies for avoiding outages. Lowering outage rates raises electricity costs, especially for high levels of reliability. Choosing an acceptable level of outages based on supply costs as well as the costs of outages is a key decision for power system planers and regulators. For instance, typical average outage rates in the Untied States are 1.4 outages per year, and a total time without power of about 3.5 hours per year. About half of the time without power is due to problems with the local distribution system and about half are associated with the transmission system.

In the past, a balance between outage costs and reliability costs has been struck through a political/legal process of some sort, whether by a Ministry deciding on power system design and prices or by a public service commission setting rates for private power companies who operate within a set of (usually voluntary) reliability guidelines. While this approach has been effective, it has also served the interests of utility managers in both public and private systems. In public systems, reliability through redundancy meant higher budgets and more employees, generally seen as markers of status in government bureaucracies. In private systems, reliability through redundancy served shareholders by tending to increase the size of the ratebase, and thus profits.

However, it is not clear how this balance will be struck if and when privatization and competition are used to guide most investment and operating decisions in the power sector. Voluntary reliability requirements could be harder to implement, although competition may lead to flexible, lower cost alternatives. CIP adds a new concern to this balancing act; actions to prevent or mitigate deliberate attacks may raise the cost of electricity service directly through new expenses or indirectly by limiting the flexibility of firms to innovate or economize. Minimizing cost increases due to CIP is an important goal—protecting a country's homeland security should not slow its economic growth unduly. A further complication will be deciding who pays for what fractions of the cost of CIP. This will be an area of research, development, politics, and practice for some time.

1.2 Stress

Stress differs from the modes of disturbance analyzed in typical reliability planning in several important ways, as seen in Table I. Since electrification occurred first in industrialized countries and has been studied most extensively there, analysis of the electricity sector reflects conditions in industrialized countries. Typical conditions in low-income countries have not been well studied. In addition, deliberate attacks on infrastructure have not been a significant concern in industrialized countries until relatively recently, so these conditions have not been well studied either. This article focuses on deliberate attacks, which will include one or more of the following general characteristics (to different levels of intensity).

1.2.1 Coordination of Attack

Unlike equipment failures or even extreme weather events, deliberate attacks are not short-term, random events. Both militaries and terrorists alike have the capabilities to coordinate attacks and intentionally maximize damage to the system. For example, the Farabundo Marti National Liberation Front (FMLN) was able to interrupt service to up to 90% of El Salvador at times and even produced manuals on disrupting electricity systems. Attacks can also be repeated as system components are repaired.

1.2.2 Scope of Impacts

The failure of distribution equipment within a half mile of the customer is the cause about half of the outages in the United States. However, in conflict situations, remote transmission lines, generating stations, and transformer substations can also become primary targets or be impacted by indirect damage stresses. The nonconflict stresses discussed earlier (e.g., poor access to spare parts) would also have a wider scope of impacts.

1.2.3 Persistence of Outage

Deliberate attacks may not be single events, and they may occur under conditions (or create conditions) that make restoration of service and repair more difficult. Alternatively, conditions could be created that make even necessary routine maintenance more difficult. When outages do occur, a number of factors can cause them to persist for days or even longer. Such factors would include risks to personnel, impeded transportation of personnel and parts, lack of funds for replacement parts, absence of technical expertise, and the possibility of subsequent sabotage. Long outages are not considered normal by electricity planners in industrialized countries (a sustained interruption is generally classified as one that lasts more than 1 hour, there is no classification level for longer outages).

1.3 Cascading Failures

Large-scale blackouts result from cascading failures, which occur when equipment failure or other disruption cause further failures or disruptions in large portions of the power grid. These events are relatively rare but potentially very costly. Reliability planning is meant to avoid creating situations in which a cascading failure is even possible while accepting that some localized (normal) outages will occur. However, such cascading failures do occur in industrialized countries and are not unique to stressed systems or to concerns about CIP. Reliability improvements may reduce the probability of cascading failures but will not eliminate them, and it may be difficult to determine the precise impact of reliability improvements on the frequency and size distribution of cascading failures.

Changes in electricity system planning to account for stress should reduce the higher incidence of normal outages that would otherwise be expected under these conditions. Avoiding cascading failures is still a goal, but planning for stressed systems is considered with wider impacts of stress and over much longer timescale than any individual cascading failure. In some places, rolling blackouts may already be the norm rather than the exception. These are not the result of cascading failures but are imposed in a controlled fashion on a regular basis due to inadequate generation. In fact, cascading failures may not even be a major concern for many stressed systems. In a tightly coupled, interconnected, and generally well-functioning system like in the United States, cascading failures are a real possibility and concern. On the other hand, the infrastructure in many places may be only loosely interconnected, which may make widespread cascading failures less likely. As a result, electric power systems under stress may or may not suffer cascading failures at a higher rate depending on the context, but they will suffer more and possibly longer localized outages.

2. TURNING OUT THE LIGHTS

One of the primary concerns for CIP in electric power systems is preventing outages, particularly large blackouts. There is much to be learned from the study of past outages. Contingency planning, spare equipment preplacement, emergency preparedness training, and temporary personnel transfers tend to be the keys to recovery from large outages. Power companies often loan trained workers in emergencies under fairly informal arrangements, knowing that they are likely to need to make such requests themselves someday. Several aspects of successful management of electric system outages have been identified: planning, training, analysis, and communications. However, communication during and after an event might be difficult, especially if, as discussed later, the attack is not on the physical part of the electric power system but on the information technology (IT) components.

The experience of several high-profile outages reinforces the need for planning and leadership. One of the largest such event in U.S. history was the Great Northeastern Blackout of 1965, which shut down New York, as well as much of Ontario and New England after an equipment failure. New York City was blacked out for an entire night, and about 30 million people were directly affected. The National Opinion Research Center studied this event and discovered the following:

> *An outstanding aspect of public response to the blackout was the absence of widespread fear, panic, or disorder. There is probably little question that this absence is largely due to the ability of individuals to interpret their initial encounter with power loss as an ordinary event…. Of equal importance in maintaining order was the rapid dissemination of information about the blackout.*

A smaller outage in New York City in July 1977 was far more troubling. Although outright panic did not emerge, there was considerable looting in some parts of the city. At the time, crime had been increasing and there was a general sense of insecurity in New York City. Some of the looting occurred due to poor response by the police, leaving some areas of the city more vulnerable. The 1977 blackout was frightening in ways that ordinary electricity outages are not, suggesting that there *can* be troubling social responses to turning out the lights. In contrast, the August 2003 blackout in the northeastern United States, the largest outage to date, did not generate terror. Commentators who had lived through both the 1977 and 2003 events noted that a more positive attitude in New York pervaded the second time.

There are three primary concerns in protecting the electric power system from an intended failure: physical security, cyber security, and fuel security. Each will be discussed in turn.

2.1 Physical Security

Many CIP concerns are about possible disruption of electric service, or turning out the lights. In places where deliberate attacks on the power system and

interruption of service are frequent, both everyday life and economic development can be damaged. However, it is not clear if an attempt to turn out the lights would permit terrorists to create widespread panic or the shock associated with violent attacks on symbolic objects that cause loss of human life. Electricity customers in most industrialized countries lose power for an average of couple of hours per year. For the most part, individuals and society cope with these outages well. Power companies respond rapidly to restore service, and facilities that have special needs for reliability (e.g., hospitals and airports) typically have backup generators.

There is some experience with large physical damage from natural forces. For example, in January 1998 an ice storm struck southeastern Canada, New York, and New England, felling 1000 transmission towers and 30,000 distribution poles while sending thousands of tree branches into power lines. This event left 1.6 million people without power, some for more than a month. Almost a quarter-million people had to leave their homes to take shelter and at least a dozen people died (mostly due to loss of heat or suffocation from faulty temporary heating setups). Insurance claims reached approximately $1 billion Canadian. This event, although massively destructive, did not cause panic or shock.

It is difficult to imagine a terrorist organization amassing the capability to accomplish damage on this scale, but a terrorist attack could nonetheless be larger and more long lasting than typical equipment- or weather-related problem, and multiple attack modes could be coordinated for maximum effect. Critical points exist in the electricity infrastructure (especially among transmission and substation assets) where attacks could cause more damage. A U.S. National Research Council panel found that "a coordinated attack on a selected set of key points in the system could result in a long-term, multistate blackout." Furthermore, the ability of the electric power sector to deal with coordinated attacks on the system is lacking. As noted by the U.S. National Electricity Reliability Council, "most (electricity sector) members do not have in place business recovery plans that include procedures for dealing with widespread, well-planned attacks on physical facilities. Likewise, they do not include procedures for dealing with deliberate attempts to hamper repair and restoration activities."

High-voltage transformers have been long identified as a key vulnerability in electric power systems, but due to the cost of stockpiling, few spares are available. It might be possible to develop modular, flexible transformers, but this goal is frustrated by the tendency of different power companies to insist on unique standards and often on customized equipment for every situation. The trend toward restructured markets may exacerbate this problem, as noted by a U.S. National Research Council panel, "Deregulation has encouraged efficiency, invested-cost utilization and return on investment rather than redundancy, reliability and security. For the same reasons, power companies keep fewer spares on hand." Other problems have also been identified. Spares are often kept at the substation site, increasing the possibility that an attack will also result in damage to the spares. If spares have to be ordered from the manufacturer, the customized nature of the transformers may mean the units could take a year or more to manufacture. Delivery could also be lengthy given that transformers are increasingly being manufactured in a small number of countries (e.g., China).

Even so, it is not clear if a highly successful physical attack on an electric power system would have the same social effect as terrorist attacks that claim lives or focus on symbolic targets. Responses would likely include outrage and a sense of vulnerability, but evidence from large-scale blackouts indicates that panic and shock do not generally ensue.

2.2 Cyber Security

Traditionally, information technology (IT) in the electric power sector has been considered less vulnerable to attack (at least by outsiders) than the transmission grid. This is partly because power sector IT systems often consisted of dedicated systems without interconnection to other computer networks and with built-in redundancies. However, systems may be more vulnerable to attack due to their use of dialup modems and the Internet. For instance, in January 2003 an Internet-transmitted computer virus infected and disabled several computers at the Davis-Bessie nuclear power plant, although this problem apparently presented no safety concerns.

In the electricity industry, energy management systems (EMS) monitor and control the production, transmission, and distribution of electrical energy. This includes scheduling and controlling generators and monitoring and analyzing the transmission network in order to provide power while maintaining reliability of the system. Competition is leading to more and more use of electricity markets to make decisions about what power plants will be operated when. Since markets are typically open, it may be

impractical (or simply expensive) to verify the identity and reliability of all participants, exacerbating the problem.

Concern about EMS arises because of their use of open protocols, increased use of intelligent devices, and lack of encryption. Remote systems receiving control signals from the control center often do not verify such signals for authenticity. Instead of relying on traditional means to protect these systems, information security typically has relied on uncommon, often proprietary protocols and the need for considerable hard-to-obtain knowledge about operating systems and so forth. However, such measures are insufficient, particularly when one considers the insider threat problem. An employee with the proper knowledge could create serious problems and face minimal resistance from the system due to lack of encryption and authorization requirements.

2.3 Fuel Security

Electric power systems vary in how vulnerable they are to fuel supply disruption. Four major fuels are used: coal, natural gas, uranium, and renewables (wind and water). Each has a different vulnerability to deliberate attack. Several important distinctions between fuel distribution and electricity transmission systems are relevant. First, it is possible to store fuel but not electricity. Second, system stability concerns in electric power systems have no real analog in fuel supply systems. Third, fuels usually have a number of alternate delivery paths, while electricity transmission systems often have few, and some electricity generators can switch fuels. Fourth, almost all of the components of electricity transmission and distribution systems are aboveground, while much of the natural gas delivery system is buried.

2.3.1 Coal

Coal is the fuel least vulnerable to deliberate attacks because it is produced in many locations, can be shipped by multiple means, and can be easily stored. In addition, coal supply systems are made up of large and relatively robust capital equipment. Coal is transported by a variety of means. Rail dominates, and other major modes include river barge and truck. Coal-fired power plants may often have multiple supply routes available to them, although one mode will tend to dominate and generally only that mode will have coal-handling capacity sufficient for long-term supply. Coal can be readily stored on site and coal-fired power plants typically have three to eight weeks of fuel supply on hand. For power plants built at coal mines, fuel security is obviously less of an issue, although these plants may require more transmission assets.

2.3.2 Petroleum

Overall, petroleum supplies are considered more vulnerable to deliberate attack than coal supplies, but less than natural gas. The most important vulnerabilities are in refineries and pipeline pumping stations, only a limited number of which typically supply any particular region. Although petroleum production (wells) and shipping (tankers) can and have been attacked (for instance, the destruction and ignition of many of Kuwait's oilfields by retreating Iraqi forces during the 1992 Gulf War), they are numerous and operate through a global market, so substitutes may be relatively readily found. An additional problem is that several chokepoints vulnerable to mines or other types of attack exist on ocean petroleum shipment routes, including the Straits of Malacca, Bosporus Straits, and Strait of Hormuz.

Table II shows the percentage of electricity generation that is fueled by imported oil for the 10 nations with the greatest fraction. (This table and the next two only consider countries with large imports.) The United States accounts for over 24% of world oil imports; however, those oil imports contribute less than 2% to the total electricity generation and so is not included. In contrast, for the next largest oil importer, Japan, electricity generation accounts for

TABLE II

Percentage of Electricity Generation from Imported Oil, Select Countries 1998–1999

Importer	Percentage of world oil imports	Percentage of generation from imported oil
Singapore	3	65
Italy and San Marino	5	32
Philippines	0.9	28
Portugal	0.7	25
Greece	1	16
Japan	11	16
Thailand	2	16
Belarus	0.5	8
Korea, Republic of	6	7
Spain	3	7
Netherlands	3	7

Source. United Nations.

16% of oil imports. Seven countries rely on oil imports for over 10% of their electricity generation.

2.3.3 Natural Gas

The gas supply system is more vulnerable to deliberate attack than coal or petroleum supplies, but it is more robust than the electric system. It is less vulnerable than the electricity system because it is mostly buried underground, it can be used to store gas (line pack), and it can use relatively secure depleted oil and gas wells for storage. Like the electricity industry, companies in the gas business have long recognized and effectively planned for contingencies that mitigate terrorism. Spare parts are generally kept on hand to effect quick repairs. However, the potential vulnerabilities in gas system supply are illustrated by the August 2000 rupture and explosion caused by corrosion in a 30-inch gas pipe near Carlsbad, New Mexico, killing 12 people. This explosion led to significant increases in gas prices in California, exacerbating the electricity crisis there. The U.S. National Transportation Safety Board subsequently determined that decades of inadequate testing and maintenance by the pipeline company caused the mishap.

Table III shows the percentage of electricity generation that is fueled by imported gas for 11 countries with the highest percentages. Due to the regional nature of gas supply, generation is shown for each major gas supply region. Eleven countries rely significantly on imported natural gas for electricity generation. Russia is a key supplier to many of the European countries that rely on natural gas imports for electricity generation. Algeria is a significant supplier for Italy, Belgium, Spain, and Turkey; while the Mideast is the key supply region for Japan and South Korea.

Considering Tables II and III together, two countries rely on imported oil and gas for over half of their electricity generation (Belarus and Italy). For comparison, the United States uses imported oil and gas for about 4.3% of its electricity generation, over half of which is from Canada.

TABLE III

Percentage of Electricity Generation Resulting from Imports of Natural Gas, Select Countries, 1998–1999

		Percentage of electricity generation from imported natural gas (by supply region)				
Importers	Percentage of world gas imports	Russia	Canada	Algeria	Europe	Asia (incl. Middle East)
Belarus	3	87				
Turkey	2	19		8		2
Italy and San Marino	8	9		12	2	
Belgium	3			7	15	
Japan	13					16
Kazakhstan	0.6	14				2
Finland	0.7	14				
Ukraine	10	13				
Hungary	2	13				
Korea, Republic of	3					11
Slovakia	1	11				

Source. United Nations.

2.3.4 Nuclear

Although nuclear power plants are operating in about 30 countries, uranium fuel is not available in all of them. Although refueling is infrequent (every 18 months or so for most reactors) it does involve risks of deliberate attack associated with production and shipping. The Table IV shows the countries most reliant on nuclear power for their electricity generation and the countries with the greatest known reserves of uranium. The top seven countries, ranked in order of nuclear contribution to electricity generation, have no uranium resources whatsoever (e.g., France, Ukraine, Korea, Germany, Japan, Spain, United Kingdom). However, uranium deposits exist in many countries and are available in significant quantities in several stable democracies.

Considering oil, gas and uranium, six countries import more than half of the energy they use for electricity generation. If coal imports were considered, this number could be larger. Further, given likely policies to limit greenhouse gas emission, gas imports are likely to rise substantially in many nations.

2.3.5 Renewables

Renewable energy sources have little, if any, fuel supply vulnerabilities. However, water impounds behind dams may be a source of concern due to the potential for damage downstream if a large dam is breached. The typically large distances between major renewable energy sources (e.g., rural or offshore locations where wind turbines are both economic and politically acceptable) and major electricity demand centers (i.e., cities) produces an

TABLE IV

Nations with Large Nuclear Energy Sectors and with Recoverable Uranium Resources

Percentage of generation provided by nuclear power		Percentage of global known recoverable uranium resources	
France	77	Australia	28
Ukraine	45	Kazakhstan	15
Korea	37	Canada	14
Germany	30	South Africa	10
Japan	30	Namibia	8
Spain	28	Brazil	6
United Kingdom	23	Russia	4
United States	20	United States	3
Russia	15	Uzbekistan	3
Canada	12		

Columns 1 and 2 are from IEA key world indicators report; columns 3 and 4 are from the World Nuclear Association.

inherent risk due to the likelihood of producing energy in remote, uninhabited areas and the necessity of long-distance transmission.

3. OTHER ATTACK MODES

Deliberate attacks on some parts of the electric power system could cause damage or terror without necessarily creating a power outage. Nuclear power plants, hydroelectric dams, and cooling towers are the major concerns.

The most obvious threat is a direct attack on nuclear reactors themselves. One possible attack mode is by airplane. Most nuclear reactor containment buildings are massive structures designed to withstand significant impacts (exceptions include plants like the Chernobyl facility in the Ukraine). However, no conclusive analysis shows that all containment buildings and the nuclear systems they house can safely withstand a direct hit by a jumbo jet or by a smaller plane loaded with high explosives. In addition, the spent reactor fuel stored in water-filled pools or aboveground containers at all operating reactor sites may be far more vulnerable, especially if the water that cools and shields hot core elements is lost. These temporary facilities have grown around the world due to the failure of any country to develop a permanent reprocessing or disposal plan for nuclear waste. Attacks on nuclear power plants, even if they have little direct impact, could raise public alarm and lead to the closure of facilities out of fear.

Due to the dread the public has of nuclear radiation (whether scientifically supported or not), it might be possible to create significant public distress (or even terror) even with an unsuccessful attack on a nuclear power plant. The core meltdown of the Three Mile Island plant in Pennsylvania in 1976 had a devastating effect on the nuclear power industry even though no one was hurt and only insignificant amounts of radioactivity were released. Even an unsuccessful attack could lead to nuclear plant closures.

Concerns about direct attack by truck bombs or commando squads have led to physical barriers and armed guards that protect all nuclear power plants. However, it is not clear how effective these steps are; a 1999 review by the U.S. Nuclear Regulatory Commission found "significant weaknesses" in 27 of the 57 plants evaluated.

Nuclear power plants also present a unique threat labeled "proliferation," which is the potential for nuclear power plant fuel (or waste) to be used in creating nuclear weapons. These could be either fission bombs, which would be relatively hard to produce, probably requiring the resources of a nation, or much cruder "dirty bombs," which even small groups could construct from conventional explosives and almost any highly radioactive material. One concern is that reprocessing fuels so that fissile material (plutonium) is separated from highly radioactive fission products could increase the proliferation potential of nuclear waste. New reactor designs that are more resistant to physical assault or require no refueling might be developed, but the cost of doing so is unclear.

Typically, the costs of mitigating the CIP risks of nuclear power plants, and other security risks such as proliferation, are not fully included in the cost of electric power these plants produce. For instance, after the terrorist attacks of September 11, 2001, many state governors put National Guard troops at nuclear plants, so the taxpayer in effect subsidized electricity consumers. As the owners of nuclear power electricity facilities have begun to face competition and the bottom line becomes more important, the costs of security are likely to receive greater scrutiny. This subsidy is similar to the support for private nuclear power plant owners in the United States from the Price-Anderson Act, which limits their liability due to accidents.

The second major threat is to hydroelectric dams, which can store significant amounts of energy in the form of water retained behind them. If they were suddenly breached, both property and lives could be lost. It is not very easy to successfully attack most large dams, however, and physical access to them can

be controlled relatively easily. One complicating factor is that smaller upstream facilities such as locks or small dams may be more vulnerable to attack. Breeching one or more of these could send a pulse of water downstream, possibly causing a larger dam downriver to fail catastrophically.

Third, cooling towers, used in many large power plants, could be used in a deliberate attack as a means of dispersing biological or chemical agents. The magnitude of this threat is not well understood since these facilities have some physical security and the effectiveness of cooling towers to disperse various agents is not well studied.

Finally, large electromagnetic pulses (EMP) can induce instantaneous voltages of hundreds of thousands of volts in conductors, creating large disruptions in electric power systems and destroying electrical equipment components such as motors, backup generators, and microprocessors. Concerns about EMP have traditionally been focused on the impact of high-altitude nuclear detonations, but relatively simple, inexpensive devices that can deliver an EMP have been designed and tested by the United States and other nations. Concerns have emerged that even crude devices could be deployed with significant localized impacts. Such weapons could damage the electric power systems and also use the electric power system as a conduit to attack other infrastructures (e.g., telecommunications) without physically penetrating facility perimeters. It is not clear how vulnerable the system is to such assaults, nor what could be done to defend against them. There is very little information from credible sources on the EMP threat from nonstate actors, so this concern is somewhat speculative.

4. MITIGATION

Mitigation efforts are designed to reduce the probability or magnitude of disruptions to electric power systems or the use of the power system for other attack modes. Mitigation measures can take a variety of forms ranging from improved security procedures to changes in the physical and institutional structure of the system. Before describing specific mitigation strategies, it is necessary to introduce the concept of survivability, which should inform any mitigation strategy.

4.1 Survivability

The traditional reliability planning approach does not lend itself well to responding to unexpected, deliberate, and potentially very high impact attacks. For one thing, it is difficult to imagine all possible attack modes or strategies. Even if this problem could be overcome, practices in reliability analysis allow for essentially only one strategy—provide more reserve capacity—a costly solution that competitive firms are not likely to implement, and one that may divert resources from other important uses in CIP or economic growth. Thus, the advent of restructuring and the possibility of terrorism require the reformulation of the reliability problem. Supply-side solutions are not likely to be sufficient any longer; more attention will need to be paid to the economic and social implications of end-use disruptions. One concept that may serve to link reliability and CIP and also be compatible with a competitive electricity industry is survivability, also known as the self-healing grid or the resilient network.

Survivability is the ability of a system to fulfill its mission in a timely manner, despite attacks, failures, or accidents; survivability can be contrasted with the fortress model of security that tries to prevent or counter all attacks but has disastrous outcomes (e.g., cascading failures) when it does inevitably fail. A fundamental assumption of survivability analysis and design is that no individual component of a system is immune from attacks, accidents, or design errors. Thus, a survivable system must be created out of inherently vulnerable subunits, making survivability an emergent property of the system rather than a design feature for individual components.

Due to the complexity and speed of electricity system operations, it may be difficult to recognize attacks until there is extensive damage. Thus, ways must be found to recognize attack early and to protect the system without taking the time to discover the cause of the attack. Survivable systems must be able to maintain or restore essential services during an attack and to recover full service after the attack. Thus, the system must fail gracefully, shedding low priority tasks and then adding tasks in a priority ordering during recovery.

For example, in most cities, traffic signals are powered by the same circuits that provide service to much less critical loads like billboards. During blackouts, injury and property loss may occur due to blank traffic signals. Worsening the problem, blackouts cause gridlock that prevents police and fire vehicles, as well as emergency response crews from reaching their destinations. This demonstrates the fortress aspect of traditional reliability planning – it creates a system in which frequent or large-scale blackouts are not supposed to occur, but when they

do, the consequences are severe. In contrast, a system designed around survivability concepts might use low-power light-emitting diode traffic lights with battery backup to ensure that a blackout did not interrupt traffic flow.

4.2 Guards, Gates, and Guns

One approach to dealing with these issues is to assess, and if necessary and justified, improve the physical security of installations in the electric power system. This approach also probably requires relying on surveillance technologies to detect attack, ranging from automated systems such as sensors and alarms to detect intrusion to manned surveillance. The primary purpose of such technologies is to detect an attack and mobilize the resources (i.e., private security force, local police, military) to repel the attack and prevent damage. A secondary purpose would be to collect evidence to be used in the investigation of any attack.

While there exist a wide variety of monitoring and surveillance technologies, they are not considered adequate for protecting the electricity infrastructure. In the case of remote facilities, monitoring technologies may not be of use in preventing an attack if the response time of an intervention force is too long.

Unfortunately, any technology used for monitoring or surveillance has the risk of failure. Another problem has to do with nuisance alarms. Surveillance and monitoring systems are not always able to distinguish between an intruder and an animal or between a line being deliberately cut and one that has been struck by a falling tree. Guarding against failures and false positives may require backup systems or complementary technologies that can confirm or deny events, increasing costs.

4.3 Increased Backup

The traditional reliability approach has also led to a localized version of reliability through redundancy at facilities that are particularly concerned about electric power reliability: the use of backup generators. Especially since the recognition of a potential Y2K problem (the probability that older computer programs, especially those embedded in operating systems, would have failed on midnight of January 1, 2000), hospitals, airports, water supply systems, and sometimes office buildings have invested in onsite diesel generators. In some cases, building codes or operating licenses require backup generators. These are typically designed to provide a minimal level of electricity supply for up to a few days before needing refueling. Such approaches may not be satisfactory under attack situations, depending on the level of functionality supported by the backup generator, its own reliability, and the feasibility of long-term refueling. Furthermore, such approaches are typically too expensive for residences and small businesses.

4.4 System Architecture and Control

4.4.1 System Architecture

There may be fundamental limitations to the improvements that can be made in the vulnerability of the electric power system solely through improved physical or cyber security of current systems. One possible method of addressing this limit would be a change in the overall physical architecture of the system. Such changes in the system architecture would be designed to provide the electric power system with more of the properties of survivable systems.

One such approach would be to use significant decentralized (or distributed) generation, often called DG, which places energy supply at the location of its use rather than relying on large, central station plants and long-distance transmission. DG facilities are smaller (though they can range widely in size from the few tens of kW to tens of MW) and more dispersed. Until recently, DG technologies have generally found use in either niche applications (e.g., in certain industrial contexts) or as backup power. However, there are areas (e.g., the Netherlands) in which significant portions of generation is supplied by grid-connected distributed units that both supply local power and feed power into the grid.

However, there has been little analysis presented to indicate either the gain in robustness expected by using DG, nor the impact on system cost. Work has been conducted to quantify some of the costs and benefits of a more distributed system by modeling hypothetical centralized and distributed systems under stress. Generally these efforts find that that distributed systems could prove to be significantly more robust under stress conditions and results in lower reserve requirements, both of which can reduce costs. In addition, in many DG applications it is possible to use waste heat, resulting in higher overall efficiencies and lower costs. The potential effect of DG on cascading failures is unknown.

Advanced renewable energy sources (solar, wind, hydro, and biomass) have the potential for significantly increasing the diversity and flexibility of

electricity generation, but they face limitations due to their cost and their dependency on transmission. (Many renewable resources are located far from demand centers.)

Another approach to improving survivability might be demand response technologies that can reduce and make more flexible demand for electricity. Demand reduction could conceivably come either through energy conservation or through demand response technologies that could adjust demand if necessary.

At its most extreme, demand response could lead to shedding of load, otherwise know as rolling blackouts. Intelligent load shedding under emergency situations could help in preventing cascading grid failures while maintaining service to the most critical loads. Such an approach would provide for inherent deterrence from attack by offering fewer large, obvious targets. It will also emphasize differences in importance among end uses during emergencies (subways and traffic lights might be considered more important than billboards and air conditioners, for instance). Such a system might be able to reconfigure itself to isolate faults (adaptive islanding).

4.4.2 Control

The traditional approach to reliability has led to the automatic control of power system components in response to very small, short-term (seconds to minutes) deviations in demand by adjusting power generated to maintain locally measured frequency and voltage within the acceptable limits. Over longer terms (minutes to hours), human operators start and stop generators to meet forecast demand changes. When major equipment outages occur, decisions are made by human operators intimately familiar with the response of each specific system to the given (previously studied) outage. In spite of these efforts, cascading failures have occurred as a result of events that were not easy to detect by the system operator. An important issue is how system control strategies would perform under attack and how they could be improved.

The primary controllers on power plants are tuned to respond to small deviations in system conditions. It is, however, possible, to make these much "smarter" by equipping them with additional sensing and communications effectively resulting in a more observable system whose dynamics are controlled. In addition, it is possible to design these controllers with adaptive, nonlinear logic. Equipping the existing power plants with such power system stabilizers has been known to greatly enhance transient stability of a power grid under stress.

Problems in protective relaying are frequently at the heart of major blackouts. A typical problem is that a protective relay is not set correctly or it fails and incorrectly senses a fault on line it is protecting. If that happens, the breaker disconnects the line, which, in turn, causes line overflows in other transmission lines creating a sequential blackout situation. The fundamental problem with protective relaying is that it only uses local information, and the settings are tuned for a certain range of operating conditions. If these are exceeded, relay functions may fail.

4.4.3 Institutional Design

The robustness of electric power systems, and the ability to restore power quickly after accidental outages, is no accident. Government regulation and private institutions have been established to promote reliability in the electric power sector. In the United States, voluntary coordination councils were set up by the electric utility industry following the 1965 blackout, partly to avoid government regulation of reliability. State public utility commissions play a role as well. All countries have some form of electricity reliability institution(s). However, industry restructuring is drastically changing institutional arrangements, leading to serious questions about whether such private approaches will be adequate to ensure reliability, let alone provide CIP.

Both reliability and CIP share an important characteristic—they are what economists call public goods, services that if supplied to one person is available to others at no cost (i.e., nonrival consumption and nonexcludable access). Public goods are not provided in socially desired quantities by market forces alone, government action of one sort or another is always needed. Thus, government will probably have to act in some way to ensure sufficient reliability and CIP. Following the 2003 blackout in the United States, pressures to move toward regulation and mandatory reliability rules mounted, although the industry still largely preferred self-regulation.

Institutional issues also play out in apparently purely technical domains. For instance, several studies have shown that barriers are often raised to the interconnection of a distributed source to the grid. These barriers could come from market structure issues, the relative roles of DG owners and utilities, and from standards that are developed

for the interconnection of distributed resources. Similar concerns exist about the incentives for cooperation and investment in new control strategies and technologies.

One of the key challenges for the future is ensuring that institutional solutions to emerging reliability and CIP problems are created for whatever new structure the electricity industry takes. This includes looking at threats that go beyond those considered in traditional contingency planning. However, the potential for creating terror through attacks on the electric power system seem limited (with the exception of an attack on a nuclear facility). Loss of power, even if large-scale and frequent, is costly and annoying, and if caused by malicious attacks can be disconcerting, but it pales in comparison to the effect of the large-scale loss of life that is often the purpose of terrorist acts.

SEE ALSO THE FOLLOWING ARTICLES

Electrical Energy and Power • Electric Power Generation: Fossil Fuel • Electric Power Generation: Valuation of Environmental Costs • Electric Power Measurements and Variables • Electric Power Reform: Social and Environmental Issues • Electric Power Systems Engineering • Electric Power: Traditional Monopoly Franchise Regulation and Rate Making • Electric Power: Transmission and Generation Reliability and Adequacy

Further Reading

Amin, M. (2001). "Towards Self-Healing Infrastructures." IEEE Computer Applications in Power, p. 20–28.

Billington, R., and Li, W. (1994). "Reliability Assessment of Electric Power Systems Using Monte Carlo Methods." Plenum Press, New York.

Council on Foreign Relations. (2002). "America Still Unprepared–America Still in Danger: Report of an Independent Task Force Sponsored by the Council on Foreign Relations." Council on Foreign Relations, New York.

Cowart, R (2002). "Electrical Energy Security: Policies for a Resilient Network." Regulatory Assistance Project, Montpelier, VT, www.raponline.org.

Farrell, A. E., Lave, L. B., and Morgan, M. G. (2002). Bolstering the security of the electric power system. *Issues Sci. Technol.* **XVIII**(3), 49–56.

Lipson, H. F., and Fisher, D. A. (1999). Survivability–A new technical and business perspective on security. *In* "New Security Paradigms Workshop." ACM, Caledon Hills, OR.

National Research Council (2002). "Making the Nation Safer: The Role of Science and Technology in Countering Terrorism." National Academies Press, Washington, DC.

Seger, K. (2003). "Utility Security: A New Paradigm." PennWell, Tulsa, OK.

Thorp, J. S., and Phadke, A. G. (1999). Protecting power systems in the post-restructuring era. *IEEE Comp. Appl. Power* **12**(1), 33–37.

Zerriffi, H., Dowlatabadi, H., and Strachan, N. D. (2002). Electricity and conflict: Advantages of a distributed system. *Electricity J.* **15**(1), 55–65.

Electric Power Generation: Fossil Fuel

JÁNOS M. BEÉR
Massachusetts Institute of Technology
Cambridge, Massachusetts, United States

1. **Background**
2. **Fossil Fuel Utilization Systems**
3. **Coal-Fired Power Generation**
4. **Pollutant Emission Control in Coal-Fired Plants**
5. **Natural Gas-Fired Plants**
6. **Power Cycles of Increased Thermodynamic Efficiency**
7. **Conclusions**

Glossary

coal pyrolysis The process of heating coal in an oxygen-deficient atmosphere.

combined cycle A combination of Brayton (gas turbine) and Rankine(condensing steam turbine) cycles; exhaust heat from gas turbines generates steam for steam turbines, and electric power is produced in both cycles. The efficiency of the combined cycle is high because of the complementary temperature ranges of the Brayton and Rankine cycles.

cycle efficiency The energy output of an energy conversion cycle expressed as the percentage of the energy input.

diffusion flame The flame produced when fuel and oxidant are separately injected through a burner, are only partially mixed prior to ignition, and their mixing continues during combustion. The overall rate of combustion of the fuel is determined by the rate of the fuel/oxidant mixing in the flame.

excess air The air required in addition to the theoretical combustion air to burn the fuel completely (usually expressed as a percentage of the theoretical combustion air).

flue gas desulfurization The postcombustion reduction of SO_x emissions. Most systems involve wet scrubbing of the flue gas with lime to produce calcium sulfate as the waste. Capable of 95% or higher sulfur removal.

fluidized bed combustion The combustion of millimeter-sized size coal particles in a dense bed of limestone (SO_2 sorbent) particles. The particles are suspended in a column of air supplied through a distributor plate at the bottom of the bed. Steam-generating tubes immersed in the fluidized bed extract heat and maintain the bed at a low (<1100 K) combustion temperature. This favors the reduction of SO_2 and NO_x emissions.

premixed flame A combustion system in which fuel and oxidant are mixed prior to ignition. The rate of combustion is determined by the chemical rates of combustion reactions, and hence by the reactant concentrations and the temperature.

Power generation can be fueled by coal or natural gas. Pulverized coal and fluidized bed coal combustion are used for steam generation in the Rankine cycle; natural gas is used in gas turbines for power generation in the Brayton cycle. These techniques, including the problems of combustion-generated pollution and the progress in pollution control technologies, are discussed in this article. The current status of advanced, high-efficiency power-generating cycles, capable of lowering significantly emissions of sulfur oxides, nitrogen oxides, mercury, and CO_2, is highlighted and their potential for deployment and the critical needs of research and development are examined.

1. BACKGROUND

The provision of electric power is one of the prerequisites of prosperity; there is strong correlation between electric power-generating capacity and per capita gross domestic product (GDP) (Fig. 1). Across the world, economic indicators signal that there will be continued growth and increased electricity demand; during the next 15 years, 1200 GW of new capacity is projected to be added to the world's present electric generation capacity of about 3500 GW.

The U.S. Energy Information Administration (EIA) data for 2001 show that 54% of the electricity supply in the United States was generated by coal,

 © 2004 Elsevier Inc. All rights reserved.

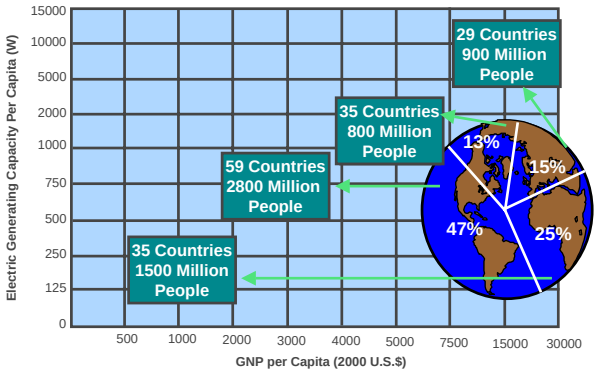


FIGURE 1 Prosperity and electricity in 2003. Generating capacity (watts) per capita as a function of gross national product (GNP; U.S. dollars) per capita. From EPRI (2003).

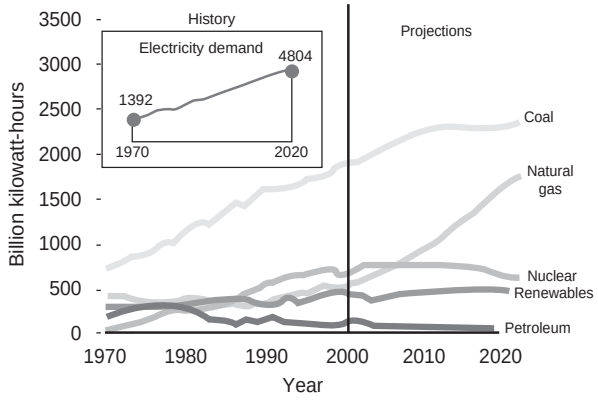


FIGURE 2 United States electricity supply by fuel type from 1970 through 2020. From Annual Energy Outlook (2000).

12% by natural gas (increasing to an estimated 32% by 2020), and about 2% by oil, with the rest produced by nuclear power (21%), hydro power (9%), and by renewable solar or wind (2%). Electricity generation (billion kilowatt-hours) by fuel type from 1970 to 2020 in the United States is illustrated in Fig. 2. The increase in the use of coal and, especially, natural gas for power generation projected for the period from 2000 to 2020 is to meet growing demand for electricity and to offset the projected retirement of nuclear power plants.

Because of the large coal reserves in major developing countries such as China, India, and Indonesia, where most of the new power-generating plants are being installed, it can be expected that coal will remain the dominant source of power generation worldwide during the first half of this century. There is also a strongly growing demand for natural gas, a clean fuel capable of being used in power generation with high efficiency.

Coal as an energy source has the attraction of broad availability (there are large reserves in several countries around the world), low cost, and economic utilization by mature technologies. The disadvantages are due mainly to environmental and health impacts at the mining phase and in the course of coal utilization. Both the mining and the utilization of coal are subject to increasingly tightening standards by governmental regulations. Because there is no presently acceptable practical alternative to coal and natural gas, industrial research and development programs are concentrated on methods of reducing pollutant emissions, especially SO_2, NO_x, and fine particulates, from coal-fired plants, and NO_x from natural gas-fired combustion turbines. New environmental technologies developed and demonstrated under the Clean Coal Technology Program of the U.S. Department of Energy (DOE) during the past 15 years have become commercial and are being applied worldwide. Major advances have also been made in stationary gas turbine technology by increasing the inlet gas temperature to the turbine and by significantly reducing nitrogen oxide emissions.

The reduction of CO_2 emission remains a challenge. Research aimed at CO_2 capture and sequestration is receiving governmental and industrial support, but it may take a decade before an economically feasible solution is found. In the mean time, the most practical approach to this problem is the improvement of electric power-generating efficiency. Improved efficiency will reduce the amount of fuel used for the generation of a given electric energy output and hence will reduce the emissions of all of the pollutants.

2. *FOSSIL FUEL UTILIZATION SYSTEMS*

The steam (Rankine) cycle and the gas turbine (Brayton) cycle are the two main thermodynamic cycles in utility power generation. In the Rankine cycle, steam is created in a boiler, and part of the enthalpy of the high-pressure superheated steam is converted to shaft work in a steam turbine that drives an electric generator. In the Brayton cycle, high-pressure and high-temperature gaseous combustion products enter directly a gas turbine, where part of their enthalpy is converted to shaft work. The gas turbine drives a compressor and the electric power generator. Because of the direct entry of the combustion products into the turbine, the gas turbine

cycle is highly sensitive to the quality of fuel and is restricted to the use of natural gas or distillate fuel oil. Boilers in the steam cycle are, by contrast, fairly tolerant to fuel quality and are capable of being fueled even by low-quality coal or waste fuels. Combustion is the prevailing fuel utilization technology in both of these cycles. In future central power plants, there are good prospects for application of gasification, which enables coal use in gas turbines, because of the environmental advantages and the promise of higher cycle efficiency.

During the past decade, there has been a growing trend toward distributed generation. Compared to large, centralized utility plants, power plants of smaller capacity, can be sited close to consumers, thereby reducing electricity transmission losses and improving operational flexibility. Being situated near population centers, these smaller plants have to be highly efficient and use a clean fuel, usually natural gas. Natural gas is free of sulfur and ash, and because of a carbon content lower than that in coal or oil, the CO_2 emission from natural gas combustion is also lower.

High-efficiency utilization is achieved by combined gas turbine–steam cycles. Also, being close to industrial and population centers opens the opportunity for combined heat and power (CHP) generation, which further enhances the plant's energy efficiency. An example of distributed generation and CHP is the Massachusetts Institute of Technology's natural gas-fired gas turbine/steam Cogen plant, which provides the campus with 22 MW of electric power and 80 t/hr of steam for heating or cooling at an energy efficiency above 80% and with very low emissions. This plant replaced a low-efficiency, more polluting, and more costly energy system in which steam was produced on campus in oil-fired boilers and electric power was purchased from a utility.

3. COAL-FIRED POWER GENERATION

3.1 Pulverized Coal Combustion

Pulverized coal (PC) combustion is presently the system of choice for coal-fired power-generating plants. In PC combustion, the coal is dried and is ground to a specified fineness, with a maximum particle size of 250–300 μm, depending on the reactivity of the coal. Coal preparation, which involves feeding, drying, and grinding of the coal, and the pneumatic transport of the pulverized coal to the burners, is fully integrated with the boiler. For lower reactivity coals, the fineness of grind is increased to create a larger specific surface area so as to improve conditions for ignition and combustion. The powdered coal is pneumatically transported to burners and injected via particle-laden jets into the combustion chamber. The transport air that carries the coal from the mill to the burners is a small fraction of the total combustion air. It is kept at low temperature, limited to about 373 K, for reasons of safety against ignition and explosion in the mill and in the powdered coal transport pipeline between the mill and the burners. The rest of the combustion air, which can be preheated to higher temperatures, is injected separately and admixed with the already ignited coal particle-laden jet in the combustion chamber. A schematic illustration of a PC combustion boiler is shown in Fig. 3. The combustion chamber is typically of parallelepiped shape; the dimensions of a 300-MW coal-fired boiler would be approximately be $15 \times 15\,m^2$ of cross-sectional area and 45–50 m in height. The combustion chamber walls are completely

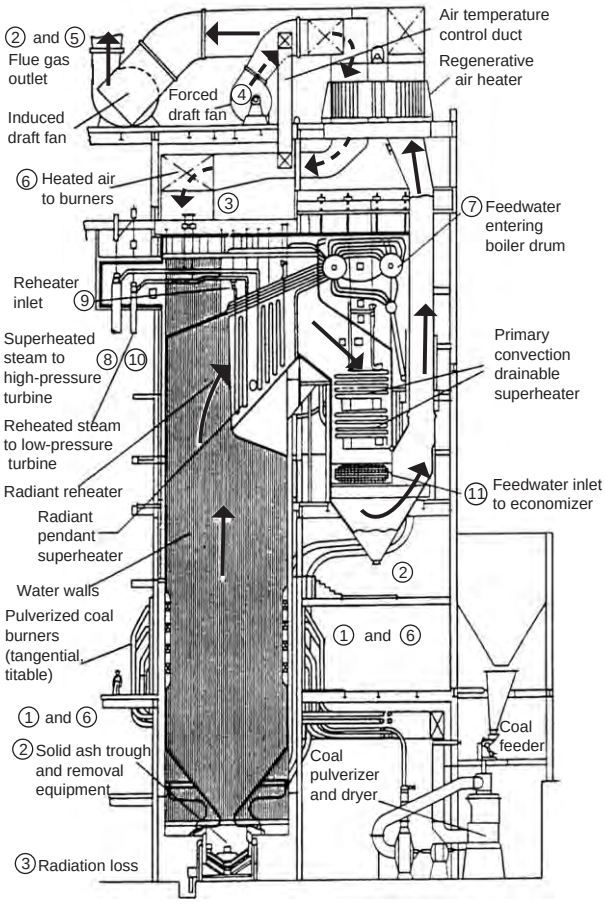


FIGURE 3 Utility boiler fired with pulverized coal.

cooled by steam-generating tubes. As the pulverized coal particles burn, the flame transfers heat, mainly by thermal radiation, to the steam-cooled tube walls of the boiler. The design of the combustion chamber has to provide for sufficient residence time of the burning particle to complete combustion, and for the cooling of the residual fly ash particles to below their "softening temperature" before entry to narrow convective heat exchanger passages, to prevent the build-up of sticky ash deposits. At regular intervals, friable deposits are removed by soot blowing using steam.

Although there are a great variety of burners in PC combustion boilers, the most widespread types are the circular burners and the vertical nozzle arrays. Circular burners are usually positioned perpendicularly to the combustion chamber walls, whereas vertical nozzle arrays are in the corners, firing tangentially onto the circumference of an imaginary cylinder in the middle of the combustion chamber. Designs utilizing circular burners are more concerned with the tailoring of the individual burner flame, whereas vertical nozzle arrays in tangentially fired furnaces are designed to rely more on the furnace volume for the mixing of the fuel and air streams. Complete combustion with a minimum (typically less than 15%) of excess air is a challenging task for a coal combustion system. The excess air, usually expressed as the percentage of the theoretically required combustion air, is the additional air needed for complete burnout of the fuel. The significance of the excess air is that the heat carried by the stack gases into the atmosphere, the largest heat loss (about 10%) of boiler operation, is directly proportional to the volume of the flue gas and, hence, to the excess air.

The fly ash formed in pulverized coal combustion is removed from the flue gas in the form of dry particulate matter; a small proportion ($\sim$10%) of the coal ash that falls off the tube walls as semimolten agglomerated ash is collected from the bottom hopper of the combustion chamber ("bottom ash"). Ash particles are molten in the high-temperature regions of the flame, but cool and solidify as they reach heat exchange surfaces of the boiler, forming friable deposits that are easily removed by soot blowing.

3.2 Fluidized Bed Combustion

In fluidized bed combustion (FBC) (Fig. 4), crushed coal, 5–10 mm in particle size, is burned in a 1000-mm-deep bed of 0.5- to 3.0-mm particles consisting mainly of additive sorbents (limestone or dolomite) and coal ash. The bed is fluidized by a blast of air through a distributor plate at the bottom of the bed.

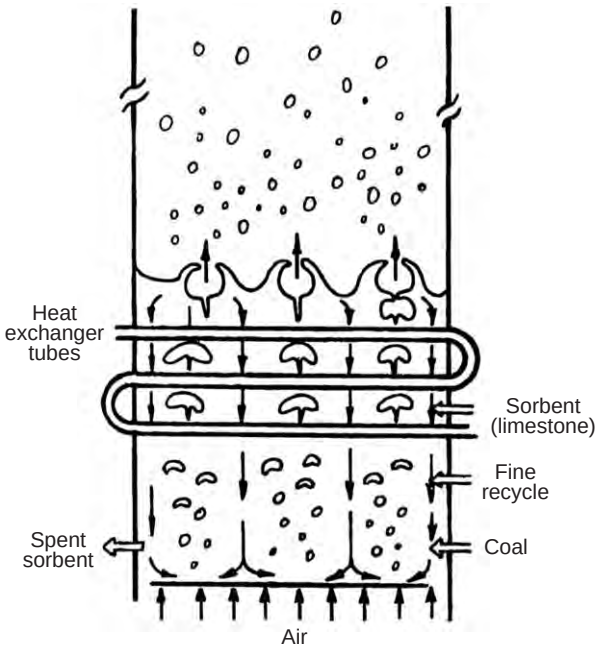


FIGURE 4 Fluidized bed coal combustion with sorbent injection for sulfur capture, and heat extraction by means of steam-generating tubes immersed in the bed. From Sarofim and Beér (1978).

As the airflow through the bed increases, the drag lifts and separates the particles from one another, and the pressure drop across the bed becomes sufficient to support the weight of the bed; the bed is incipiently fluidized. Further increase of the velocity causes the air to form bubbles that rise in the bed like gas in liquid. The bubbles stir the bed by carrying pockets of particles on their tops and in their wake. The particles are released and are returned to the bed as the bubbles burst on the top of the fluidized bed. Less than 2% of the bed material is coal; the rest is coal ash and additive limestone, capable of capturing sulfur after the calcination of the additive limestone and the formation of lime (CaO). Due to the excellent gas–solid contact, the fluidized bed is a favorable reactor for the combustion of millimeter-size coal particles and for sulfur capture in the combustion process. The bed is cooled to a temperature of 1050–1170 K by immersed steam-generating tubes. This relatively low combustion temperature prevents the softening of the coal ash and the chemical decomposition of $CaSO_4$, the product of sulfur capture.

3.2.1 Circulating Fluidized Bed Combustion

In circulating fluidized bed combustion (CFBC) (Fig. 5), the gas velocity is higher than in FBC, and a large part of the solid particles is carried over from

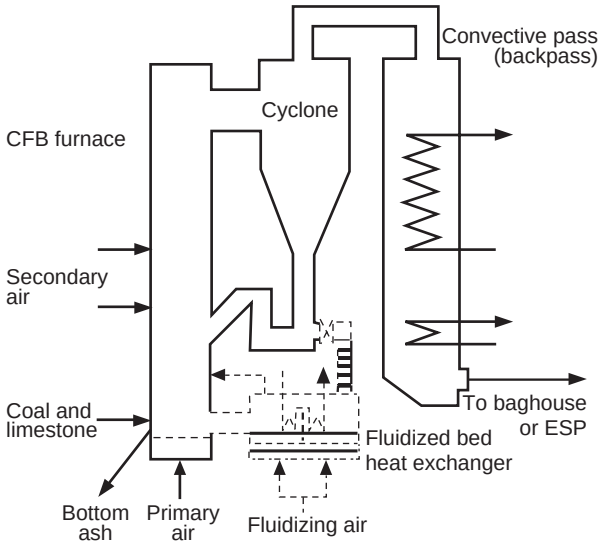


FIGURE 5 Circulating fluidized bed (CFB) furnace. ESP, electrostatic precipitator. From Scott and Nilsson (1999).

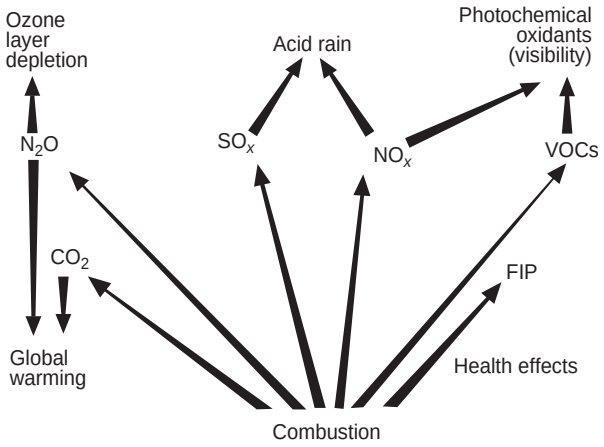


FIGURE 6 Combustion-generated pollutants; environmental and health effects. VOCs, volatile organic compounds; FIP, fine inorganic particles.

the bed into a tall riser, where the particles aggregate into clusters that reflux at the wall. Particles that are carried over from the riser are precipitated by a high-temperature process cyclone and are recirculated into the riser. With the higher gas velocity, the bed cross-sectional area becomes smaller than in FBC, for a given performance. This, in turn, leads to a smaller number of coal feed points required for maintaining uniform coal concentration over the bed cross-sectional area, a significant operational convenience. Also, because of the higher gas velocity and the somewhat smaller sorbent particle size, less additive sorbent is required in CFBC for a given percentage sulfur capture.

FBC and CFBC are commercial, mature technologies; primarily applied to the use of low-grade fuels in smaller unit sizes. The main advantages of CFBC are the capability of capturing sulfur in the combustion process by additive sorbents and the lack of sensitivity to fuel type or fuel quality variation. Due to the refractory lined process cyclone circulating fluidized bed combustors do not lend themselves easily to scale-up to the larger utility boiler sizes ($\geq$660 MW).

4. POLLUTANT EMISSION CONTROL IN COAL-FIRED PLANTS

The combustion-generated pollutants of concern are mercury, fine inorganic particulates, and oxides of sulfur and nitrogen. Figure 6 is an illustration of the pollutants and their effects.

4.1 Sulfur

Sulfur in coal will oxidize to SO_2, a precursor of acid rain. Methods of sulfur capture in the combustion process involve the reactions of sorbents such as calcined limestone (CaO) with SO_2 to produce $CaSO_4$, a stable, disposable solid waste. In the high-temperature, fuel-lean environment of pulverized coal flames ($T_{peak} \sim 2000$ K), however, $CaSO_4$ is unstable; it decomposes, which leaves flue gas desulfurization (FGD) as the viable option for sulfur capture from PC combustion. The development of fluidized bed combustion provided the opportunity to effectively retain sulfur in the combustion process, because $CaSO_4$, is stable at the lower FBC operating temperature of 1050–1170 K. Sulfur capture in fluidized combustion ranges between 70 and 90%, with even higher capture efficiencies obtained for subbituminous coals of high calcium content. FGD technologies are mature and commercially proven; 99% SO_2 reduction has been achieved at some plants, with 90–95% reduction as routine. Sulfur dioxide emissions from utility plants in the United States have been reduced 40% since 1980, even though electricity production increased by 35% during the same period. This was achieved by switching to coals of lower sulfur content and by retrofitting some of the existing coal-fired power plants with scrubbers. This process will have to continue to satisfy the constraints of the Clean Air Act Amendment of 1990, which requires further reductions in SO_2 emissions in the Phase 2 implementation that began in the year 2000.

4.2 Nitrogen Oxides

Nitrogen oxides as pollutants deserve special attention, because of their wide-ranging effect on the environment, including contribution to acid rain, reduction of atmospheric visibility, production of tropospheric (ground level) ozone, and, in the case of N_2O, depletion of stratospheric ozone. It is also noteworthy that NO_x emissions are amenable to reduction by combustion process modifications. The main routes of NO formation in coal flames are the attack of atomic oxygen on molecular nitrogen in high-temperature, fuel-lean flames ("thermal NO"), and the oxidation of the organically bound nitrogen in the coal ("fuel NO"). It is noteworthy that fuel-bound nitrogen and NO can be converted to molecular nitrogen and, hence, rendered innocuous for NO_x emission by coal pyrolysis, a process in which coal is maintained at high temperature in an oxygen-deficient atmosphere. To complete combustion, the rest of the combustion air, called "overfire air" (OFA), ought then to be admixed to the products of the pyrolysis reactions. In this burnout stage of the flame, the temperature needs to be controlled to levels sufficiently high to complete combustion but insufficient for significant thermal NO formation.

NO_x emission control by combustion process modification is accomplished by OFA, low-NO_x burner, and NO reburn technologies. In low-NO_x burners (LNBs) (Fig. 7), air staging is achieved by aerodynamically tailored mixing of the fuel jet with airstreams supplied through each individual burner, rather than through the splitting of the combustion chamber into fuel-rich and fuel-lean combustion zones by the use of OFA. LNBs and OFA represent the most cost-effective methods of achieving reduced NO_x emissions from new plants and also from existing boilers by retrofit.

NO reburning involves a secondary fuel, usually natural gas; the gas, amounting to 10–20% of the total heat input to the boiler, is injected into the combustion chamber and mixed with the NO-bearing coal combustion products. The NO is reduced by reactions with hydrocarbon fragments, CH and CH_2, that are formed in the course of the pyrolysis of the natural gas. Control of NO_x emissions by combustion process modifications and their combinations can yield up to 70% reductions. These technologies are successfully applied in the utility industry; more than 188 GW of electric power-generating capacity currently in operation internationally has been fitted by these combustion measures.

Further reductions in NO_x emissions can be obtained by the injection of additive ammonia into the combustion products. At high temperature (1060–1350 K), no catalyst is needed in the selective noncatalytic reduction (SNCR) process for the removal of up to 40% of the NO formed during coal combustion. In selective catalytic reduction (SCR), the ammonia vapor is injected over a catalyst bed situated at the boiler economizer outlet, where the flue gas temperature ranges from 600 to 650 K. In the presence of the catalyst, the ammonia chemically reacts with NO_x to form water vapor and N_2. The cost of SRC installation is higher than that of SNCR, but much less ammonia has to be injected, and NO_x reductions in excess of 90% can be obtained, depending on the type of coal burned.

4.3 Fine Particulates

Fine particulates are removed by electrostatic precipitators (ESPs) or by baghouse (BH) filters. In the ESP, particles with an electric charge are attracted to collecting electrodes in a high-voltage field. The precipitation efficiency is high, in the 10- to 300-μm size range, but it falls off for smaller particles, especially those in the submicrometer range. These ultrafine particulates have important health effects and their emissions will likely be regulated. Baghouses are fabric collectors in which the dust-laden gas flows

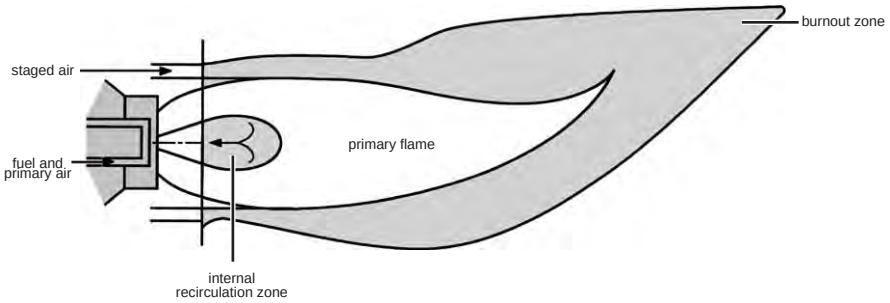


FIGURE 7 Schematic of a low-NO_x burner. From Soud and Fukusawa (1996).

through a cloth tube, where particles are captured by sieving action. As a "filter cake" forms by deposition, particles much smaller than the interstices of the cloth are captured. The particles are then removed by reversed air flow or shaking, and are collected in a bin below the cloth filter tubes. Baghouses are capable of precipitating submicrometer particles and fumes with high efficiency. Their flow resistance is higher than that of the ESP and their operation is, therefore, somewhat more costly.

4.4 Mercury

Mercury has been identified by the U.S. Environmental Protection Agency (EPA) as a toxic substance of great concern, because of its adverse health effects. Regulations for mercury removal from coal-fired utility plants are expected to be issued by the end of 2004. Significant research and development efforts by the U.S. DOE and industry are in progress to find cost-effective ways for the removal of mercury from stack gases. Potential solutions include SO_2 scrubbers that will also capture a high percentage of soluble (oxidized) mercury, SCR catalysts for NO_x removal to oxidize elemental mercury for subsequent capture by electrostatic precipitators or fabric filters, and the absorption of vapor-phase mercury by beds of activated carbon or particles injected into the flue gas and subsequently removed by gas–solid separation devices. The cost of mercury removal is significantly lower in integrated coal gasification combined cycle (IGCC) plants than it is in PC plants because the treated volume of syngas in IGCC plants is much smaller, compared to the flue gas in PC plants. However, this affects mainly the choice of new power plants, because of the small number of IGCC plants presently in operation.

5. NATURAL GAS-FIRED PLANTS

Natural gas is a premium fuel; because of its relatively high price, in the past it was used primarily for "peak-shaving" in gas turbine plants. Reassessment of the long-term availability of natural gas, the development of gas turbine–steam combined cycle (GTCC) plants with high cycle efficiency, environmental pressures for "decarbonization" of the fuel supply, and the rapid growth of distributed power generation have made natural gas the choice fuel for base load power generation. Natural gas has no sulfur, and the carbon:hydrogen ratio of its composition is much lower than that in petroleum fuels or in coals. Because of the lower C:H ratio, CO_2 emissions from the combustion of natural gas are significantly lower compared to the heat equivalent of other fossil fuels.

In the gas turbine cycle, the gaseous combustion product is the direct working medium of power generation. The gas turbine (GT) consists of a compressor, combustor, turbine, and heat exchanger (Fig. 8). Combustion takes place at elevated pressure in the combustor. Large amounts of excess air are used as diluent, to ensure that the combustion product temperature at entry to the gas turbine does not exceed the limit set to protect the structural integrity of the turbine. However, because the cycle efficiency is proportional to GT entry temperature, concerted research and development efforts are in progress to permit higher inlet temperatures by means of improved construction materials and more effective turbine blade cooling technologies. As a result, gas turbine entry temperatures have risen almost steadily over the past two decades to a state-of-the-art value of 1570 K. Further increases can be expected by the introduction of advanced ceramic materials to GT technology.

5.1 Nitrogen Oxides Emission Control in Natural Gas-Fired Gas Turbine Plants

There is no nitrogen organically bound to fuel molecules in natural gas; NO forms primarily by the "thermal NO" process. In the conventional method of combustion, fuel and air are separately injected into the combustor and mix in the course of combustion (diffusion flame). This process is prone to high rates of thermal NO formation because near-stoichiometric (theoretical air:fuel ratio) mixtures prevail on the boundaries of fuel-rich and fuel-lean turbulent eddies in such flames. In contemporary

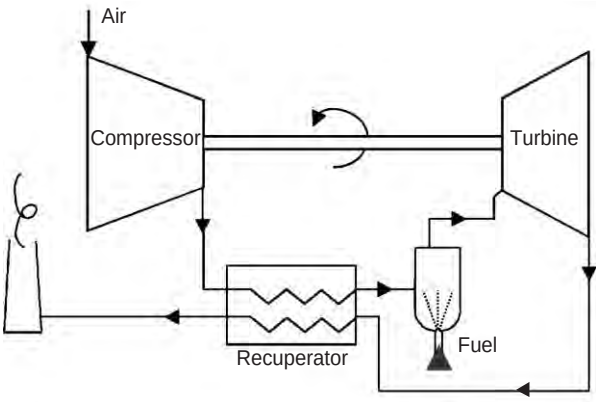


FIGURE 8 Gas turbine cycle with regenerator.

combustion technology's response to this problem, the fuel gas and air are premixed prior to their entry into the combustor, creating a strongly fuel-lean mixture that corresponds to the combustor exit gas temperature. This, so-called ultralean premixed combustion results in very low NO_x emissions, but because the combustor is operated at mixture ratios close to the lean blow-off limit of the fuel gas/air mixture, the flame stability may be impaired. For restoring flame stability, a small amount of non-premixed fuel, say 10%, is injected into the combustor to form a pilot flame and act as a stable source of ignition. Further reduction of NO can be obtained by steam injection into the combustor. Although steam injection will depress the cycle efficiency on account of loss of the latent heat of steam, this effect is somewhat compensated for by the fact that the steam represents additional mass flow through the turbine without the need for expending additional compression work.

6. POWER CYCLES OF INCREASED THERMODYNAMIC EFFICIENCY

Pollutant emissions from electric power-generating plants are inversely proportional to the thermodynamic cycle efficiency of power generation. Some power plants with advanced thermodynamic cycles are in commercial use, others have the potential to be applied in the near and medium term, and there are also long-term research and development programs for very high-efficiency and "zero-emission" cycles in power generation. It has to be borne in mind that in addition to technical feasibility, a novel electric power-generating system needs to be successfully demonstrated at a commercial scale for years, and must be economically viable, before it is accepted for application by the utility industry.

6.1 Pulverized Coal Combustion in Supercritical Steam Cycle

Pulverized coal-fired supercritical steam boilers (e.g. 250 atm, 2×853 K), have been in use since the 1930s, but improvements in materials and increasing demand for higher efficiency are making this system the choice of new coal-fired utility plant worldwide. Their increased efficiency is due to the higher mean temperature of heat addition in the supercritical steam cycle. Because of a strong increase in the moisture content of high-pressure steam as it expands through the last stages of the steam turbine, the steam has to be reheated by taking it back from the turbine to the boiler. Reheating, once or multiple times, will raise also the thermodynamic efficiency of the power cycle because it further increases the mean temperature of heat addition. Advanced supercritical steam power plants with steam parameters of 300 atm pressure, temperatures of 866 K, and cycle efficiencies up to 43% are expected to play a major role in new power generation in the near future. As the steam temperature approaches 973 K, the efficiency may reach 47%, but new advanced designs and materials for the boiler, the steam turbine, and the associated piping will be required; development will likely yield commercial applications after 2010.

6.2 Gas Turbine–Steam Combined Cycles

In gas turbine–steam combined cycle plant, the gas turbine exhaust, typically 750 K and containing 13–15% O_2, raises steam in a heat recovery steam generator for producing power in a steam turbine and/or heat in cogeneration. Because of the complementary temperature ranges of the Brayton (1600–900 K) and Rankine (850–288 K) cycles, their combination can produce significantly improved thermodynamic cycle efficiency, approaching 58% in modern commercial plants. When coal or residual fuel oil is added as a supplementary fuel in the heat recovery steam generator of a GTCC plant (Fig. 9), there is considerable gain in operational flexibility, albeit at the expense of a small reduction in the cycle efficiency. Due to a low first cost ($400–500/kW, less than one-half of the approximate first cost of a PC plant), high efficiency, and low emissions, GTCC plants are the favored choice for new power plants, despite the gas/coal price differential and the uncertainty of future gas prices.

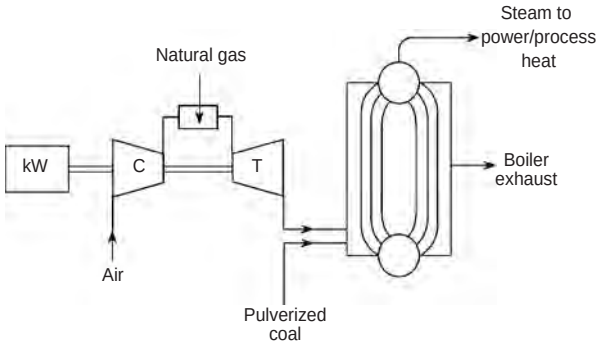


FIGURE 9 Gas turbine–steam combined cycle with fired heat recovery steam generator. C, compressor; T, turbine.

Gas turbines can be used also for retrofitting existing steam plants by injecting the gas turbine (GT) exhaust through the boilers' coal burners ("hot wind box repowering"), or applying it to partial boiler-feed water heating, thereby allowing more of the steam flow in the steam turbine to expand to condenser pressure instead of being bled for regenerative-feed water heating. This will then result in both higher efficiency and increased electric power-generating capacity.

6.3 Pressurized Fluidized Bed Combustion

In pressurized fluidized bed combustion (PFBC), the advantages of sulfur capture by additive sorbents in the bed and an operation insensitive to fuel quality are combined with increased efficiency of the combined steam and gas turbine cycle. Compared to FBC, the heat release rate per unit bed area in PFBC is several times higher, making the "footprint" of the pressurized power plant smaller. The bed height in the PFBC is 3–4 m instead of the typical bed height of 1 m in FBC. Under atmospheric pressure, the acceptable pressure drop of about 100 mbar limits height across the bed. In the PFBC–GT cycle, a 300 mbar pressure drop represents less than 3% of the total pressure ratio. A consequence of the increased bed height is a larger carbon inventory in the bed and lower NO emission due to the reduction of the NO formed by solid carbon. The high carbon load, however, does not reduce the emission of nitrous oxide, N_2O, which is still stable at the relatively low temperature ($\sim$1050 K) in the PFBC. A further disadvantage of this temperature is that it is also low for efficient gas turbine applications. The efficiency of the PFBC combined cycle units in operation is about 41%. Further increases in efficiency can be obtained by raising the turbine inlet temperature using a topping combustor.

6.3.1 The Second Generation of PFBC

The second generation of PFBC is a response to the need for increasing the gas temperature at the inlet to the gas turbine. In this cycle (Fig. 10), coal, usually in the form of a coal–water slurry, is injected into a pressurized fluidized bed carbonizer, where it undergoes mild gasification to produce char and a low-calorific-value gas. The char is burned in a pressurized circulating fluidized bed combustor (PCFBC) with high levels of excess air. Sulfur is captured in the PCFBC by additive dolomite and the flue gas is cleaned of particulates and alkali at high temperature before entry to the GT. Sulfur is also captured in the

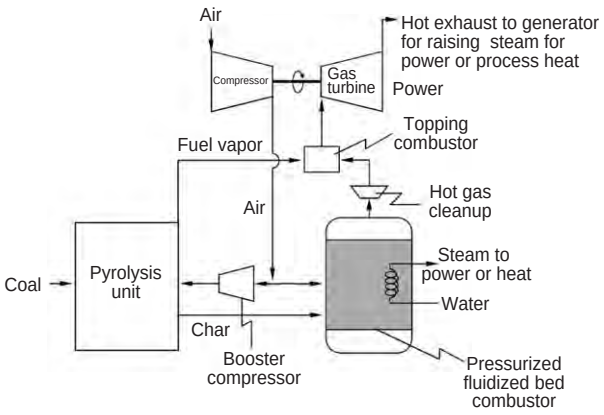


FIGURE 10 Pressurized fluidized bed combined cycle with topping combustion. From Robertson *et al.* (1989).

fluidized bed carbonizer, and the fuel gas, after being cleaned of particulates and alkali, is burned in the topping combustor to increase the temperature at the inlet to the gas turbine to 1470 K. This temperature rise increases the cycle efficiency up to about 45%. Further improvements in efficiency can be obtained by the application of advanced gas turbine technology and, on the steam side, by advanced steam parameters. An additional advantage of this cycle is that the N_2O emission is eliminated because the N_2O formed in the pressurized fluidized combustor decomposes at the elevated temperature in the topping combustor. The successful application of this cycle requires further demonstrations of the hot gas clean up and the topping combustor. The PCFBC with topping combustor is expected to be ready for commercial application after 2010.

6.4 Integrated Gasification Combined Cycle

IGCC involves the total gasification of coal, mostly with oxygen and steam, to produce a medium-calorific-value fuel gas for use in a gas turbine cycle. The fuel gas, after being cleaned of particulates, alkali, and sulfur compounds, is burned in a low-NO_x combustor. The gasifier also produces steam for a steam power cycle. A schematic arrangement of IGCC equipment (gasifier, gas-cleaning tanks, steam-raising tanks, and power-generating apparatus) is illustrated in Fig. 11.

IGCC is the cleanest presently available advanced coal technology. It is working with no major operational problems. The future of IGCC depends on whether it will be possible to reduce its first cost and to increase the cycle efficiency. The cost is

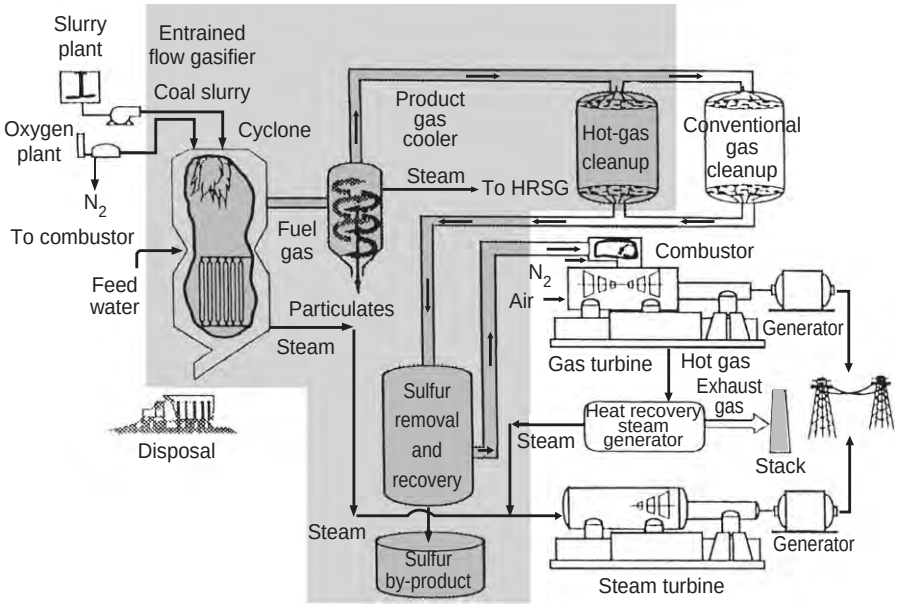


FIGURE 11 Integrated gasification combined cycle. HRSG, heat recovery steam generator. From Ruth (1998).

presently high, mainly because of the oxygen plant needed for the oxygen-blown gasifier and also because of the less than complete integration of the various subsystems, such as the gasifier air separation system, fuel gas cooler and cleanup, and gas turbine and steam plants. Existing IGCC demonstration plants in the United States have efficiencies under 40%, but two newly commissioned European IGCC demonstration plants have higher design efficiencies of 43 and 45%. These higher cycle efficiencies are mainly due to improved gas turbine and steam plant efficiencies and also to better subsystems integration. Further improvements yielding cycle efficiencies up to 50% are possible.

Even with improved cycle efficiencies, however, IGCC plants are presently not competitive with other advanced coal-burning systems, such as PC-fired supercritical steam plants, because of the higher installation costs of IGCC plants. Nevertheless, the following considerations may eventually tilt the balance in favor of IGCC applications:

- The cost of mercury removal in IGCC plants is lower because the treated volume of syngas is much smaller than that of the flue gas in PC-fired plants.
- IGCC lends itself to more efficient removal of CO_2 from high-pressure fuel gas.
- By broadening the fuel supply to the ever-increasing volume of refinery wastes (heavy residual oils and petroleum coke), IGCC could become attractive for clean and efficient central power generation by using fuels of very low or even negative cost (the waste fuel cost is negative if it stands against the cost of disposal).

6.5 Hybrid Gasification/Fuel Cell/Gas Turbine/Steam Combined Cycles

6.5.1 The DOE's Vision 21 Cycle

One of the promising coal-fired advanced cycles expected to be ready for demonstration in the period between 2010 and 2015 is the Vision 21 Cycle of the U.S. DOE. (Fig. 12). The fuel gas produced in an oxygen-blown gasifier at elevated pressure is cleaned at a temperature of 1273 K. The gas, composed mainly of H_2 and CO, enters a solid oxide fuel cell (SOFC) on the anode side, while air from a compressor exhaust, preheated in a recuperator, enters on the side of the cathode. The hydrogen is used to generate electricity in the SOFC, and the CO burns in a combustion turbine that drives the compressor. Electric power is produced in another SOFC and gas turbine, at a lower pressure downstream of the high-pressure turbine, with more power added by a bottoming steam cycle. The efficiency could reach 60% in this "zero-emission" scheme. Concerted research and development efforts by a DOE/industry/university program aim at the development of individual modules of this cycle and the complex control system required for flexible load following.

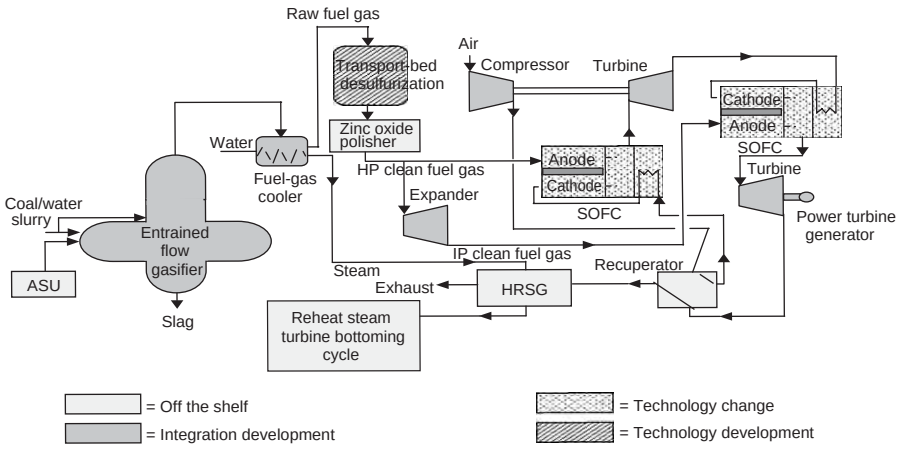


FIGURE 12 Vision 21. Gasification/gas turbine/fuel cell/steam combined cycle plant. SOFC, solid oxide fuel cell; HRSG, heat recovery steam generator; HP, high pressure; IP, intermediate pressure; ASU, air separation unit. From Ruth (1998).

7. CONCLUSIONS

Coal is the prevailing fuel of power generation worldwide, and it is likely to remain important well into the next century. Natural gas in gas turbine combined cycle plants will, in the short and medium terms, replace some coal-fired capacity, especially in distributed power generation and in heat and power cogeneration. Because of coal's role in pollution, and especially due to the expected future limitations on CO_2 emissions, clean coal utilization technology with high thermodynamic efficiency will have to be applied for new coal-fired central power stations. Pulverized coal combustion in supercritical steam boilers (300 atm, 2×866 K) can play a major role in new power generation because of reduced emissions, improved efficiency (43%), and the long experience with pulverized coal combustion technology. As steam temperatures approach 973 K, efficiency may reach 47%, but new advanced designs and materials for the boiler, the steam turbine, and the associated piping will be required; development will likely yield commercial applications after 2010.

Several small, 70-MWe pressurized fluidized combustion combined cycle plants have been operating satisfactorily since 1991. Efficiencies are around 41% and emissions are low, except for N_2O (50–100 ppm). A larger plant of 360 MWe capacity began operation in 1999 in Japan. Combustion of fuel gas produced by pregasification of coal increases turbine inlet temperature to 1470 K, with efficiency increasing to 45% in the topping cycle (second-generation PFBC). N_2O is eliminated at the elevated temperature in the topping combustor. Enabling technologies, e.g., hot gas cleanup and topping combustors, need to be finalized and are expected to be ready for demonstration by 2010.

IGCC is the cleanest of advanced coal-fired technologies. The demonstration plants in the United States were designed with relatively low cycle efficiency ($\sim$40%), but IGCC is capable of higher efficiencies through the application of advanced gas turbine technology and better subsystem integration. The disadvantage of presently higher installation cost compared to PC-fired plants could be compensated for in the future by lower cost of mercury removal and more favorable conditions for CO_2 capture. Interest in gasification has also continued to increase because of favorable prospects for achieving a zero-emission and near-60% cycle efficiency system by combinations of coal gasification, fuel cell, and gas turbine technology (e.g., Vision 21).

Acknowledgments

Parts of this article have been published in a paper entitled *Developments in Power Generation in Response to Environmental Challenges,* in the journal *Progress in Combustion and Energy Science*. Permission of Elsevier to use this material is acknowledged with thanks.

SEE ALSO THE FOLLOWING ARTICLES

Electrical Energy and Power • Electric Power: Critical Infrastructure Protection • Electric Power Generation: Valuation of Environmental Costs • Electric Power Measurements and Variables • Electric Power Reform: Social and Environmental Issues • Electric Power Systems Engineering • Electric Power: Traditional

Monopoly Franchise Regulation and Rate Making • Electric Power: Transmission and Generation Reliability and Adequacy • Hazardous Waste from Fossil Fuels • Investment in Fossil Fuels Industries

Further Reading

Beér, J. M. (1988). "H. C. Hottel Plenary Lecture," *Symp. (Int.) Combust., 22nd,* pp. 1–16. The Combustion Institute, Pittsburgh, Pennsylvania.

Beér, J. M. (1996). Low NO_x burners for boilers, furnaces and gas turbines. *Combust. Sci. Technol.* **121,** 169–191.

Blum, R., and Hald, J. (2000). "ELSAM." Skaerbaek, Denmark.

Boyce, M P. (2002). "Handbook for Cogeneration and Combined Cycle Power Plants." ASME Press, New York.

Electric Power Research Institute (EPRI). (2003). "Electricity Technology Roadmap." [www.epri.com]. EPRI, Palo Alto, California.

Energy Information Administration. (2000). "Annual Energy Outlook 2001 with Projections to 2020." U.S. Department of Energy, Washington, D.C.

Folsom, B., Hong, C., Sommer, T., and Pratapas, J. M. (1993). Report TR103265-V2. *Proc. Joint Symp. Stationary Combust. NO_x Control, 1993,* pp. 7A; 15–34. Electric Power Research Institute (EPRI), Palo Alto, California.

"Fuel Cell Handbook," 5th ed. (2000). DOE National Energy Technology Laboratory (NETL), Morgantown, WV.

Haupt, G., and Karg, J. (1998). *Proc. Conf. Electric Power Supply and Industry (CEPSI)*, 12th. Pattaya/Thailand.

Herzog, H. J., and Drake, E. M. (1993). "IEA Greenhouse Gas R&D Program, IEA/93/0E6." International Energy Agency, Cheltenham, U.K.

Robertson, A., Garland, R., Newby, R., Rehmat, A., and Rebow, L. (1989). "Second Generation Pressurized Fluidized Bed Combustion Plant". Foster Wheeler Dev. Corp. Rep. FWC/FWDC-TR 89/11 to the U.S. DOE, DE-AC21-86MC21023.

Ruth, L. A. (1998). "U.S. Department of Energy Vision 21 Workshop." DOE Federal Energy Technology Center, Pittsburgh, Pennsylvania.

Sarofim, A. F., and Beer, J. M. (1978). *Symp. (Int.) Combust., 17th*, pp. 189–204. The Combustion Institute, Pittsburgh, Pennsylvania.

Scott, D. H., and Nilsson, P. A (1999). "Competitiveness of Future Coal-Fired Units in Different Countries." IEA Coal Research, CCC/14, London.

Soud, H. N., and Fukusawa, K. (1996). "Developments in NO_x Abatement and Control." IEA Coal Research, London

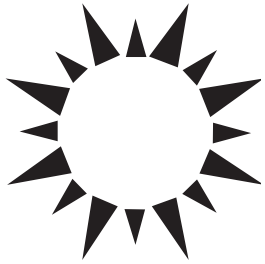
Steinbach, C., Goudeaux, D., Troger, C., and Moore, R. (2000). "Operation Experience with Alstom Power's GT10 Gas Turbine and Ultra Low NOx Combustion System." PowerGen Conference, Orlando, FL.

U.S. National Coal Council Report to the Secretary of Energy. (2003). "Coal Related Greenhouse Gas Management Issues." Washington, DC.

Wendt, J. O., Sternling, C. V., and Mafovich, M. A. (1973). *Symp. (Int.) Combust., 14th*, pp. 897–904. The Combustion Institute, Pittsburgh, Pennsylvania.

Wiswanathan, R., and Bakker, W. T. (2000). "Materials for Ultra Supercritical Fossil Fuel Plants." EPRI Report TR-114750.

Electric Power Generation: Valuation of Environmental Costs



THOMAS SUNDQVIST
Swedish Energy Agency
Eskilstuna, Sweden

PATRIK SÖDERHOLM
Luleå University of Technology
Luleå, Sweden

ANDREW STIRLING
University of Sussex
Brighton, United Kingdom

1. **Introduction**
2. **The Valuation of Electricity Externalities**
3. **Electricity Externality Studies**
4. **Critical Issues in Externality Assessment**
5. **Conclusions**

Glossary

abatement cost The implicit cost associated with controlling or mitigating damage.
damage cost The direct cost associated with damage.
economic valuation Measuring individual preferences for or against change, as expressed by changes in welfare.
environmental cost Externality arising from environmental impacts.
externality An unpriced and uncompensated side effect of one agent's actions that directly affects the welfare of another agent.
social cost The total costs to society of a production or consumption activity; the sum of private and external costs.
neoclassical welfare economics A branch of economics that studies efficiency and the overall well-being of society based on alternative allocations of scarce resources.
utility The welfare or satisfaction obtained from the use or consumption of goods and services.
willingness to accept Monetary measure for a change in welfare; the monetary amount that someone is willing to receive or accept to give up a good or service.
willingness to pay Monetary measure for a change in welfare; the monetary amount that someone is willing to give up or pay to acquire a good or service.

The availability of reliable electric power at low cost is a necessary condition for sustained global economic welfare. However, at the same time, electric power generation is the source of a number of social and environmental problems. Therefore, a wide variety of regulations and economic incentives exist worldwide to promote the introduction of more environmentally benign power generation technologies. When implementing such policy measures, two questions typically arise. First, which technologies are the more relatively environmentally benign? Second, how does one find a proper overall balance between the benefits of electricity production and the costs of environmental degradation?

1. INTRODUCTION

Over recent decades policymakers have shown an increased interest in the market-oriented measures recommended by neoclassical economics. According to this body of analysis, normal market processes are assumed to approximate optimal balances between those private costs and benefits that are reflected in market prices. For those costs and benefits that are "external" to market prices, the neoclassical prescription is to apply nonmarket valuation techniques

 © 2004 Elsevier Inc. All rights reserved.

to determine the magnitudes of these externalities and, thereby, to enable regulatory or fiscal measures to explicitly include these values in market prices. Under this view, normal processes of market optimization will then be based on total social costs (private costs plus external costs) and, thereby, will reflect a proper balance between the benefits and impacts in their widest sense. Formally, then, an externality (i.e., an external cost or benefit) may be defined as an unpriced and uncompensated side effect of one agent's actions (in our case electric utilities) that directly affects the welfare of another agent. Therefore, the challenge lies in finding techniques by which to derive complete and reliable values for the many different types of external cost associated with electricity production.

In response to increasing demands from policymakers during the 1980s and 1990s, this challenge received intense attention from economists and environmental scientists worldwide. A number of large-scale international studies embarked on the task of quantifying and valuing the externalities associated with electric power production. Examples include the ExternE project in the European Community and the External Costs and Benefits of Fuel Cycles study in the United States. The ultimate goal of these studies has been to guide policymaking by providing a robust basis for new regulations and taxes and, thereby, to aid decision makers in their choices among fuels and technologies.

The results obtained by several of these externality studies have already begun to exert some influence on regulatory decision making and wider policy debates. For instance, certain external cost values have been featured prominently in legislative initiatives for "environmental cost adders" in a few U.S. states, and some other results (e.g., those associated with global warming) have from time to time become the subject of high-profile public policy debates. However, the results obtained in the accumulated extensive body of literature on the external environmental costs of electricity production so far have generally exerted only a limited impact on real policy decisions. Some authors argue that this is because electricity environmental impact studies may have raised more questions than they have answered and because there are important limits to their usefulness in deriving robust recommendations for policymakers.

This article reviews the economic theory behind externality valuation as well as the various methods that can be used to assess external costs in practice. It also surveys some results from the vast externality valuation literature and discusses some important practical issues that bear on efforts to improve the environmental performance of electricity systems by adopting these particular values as a basis for regulatory and policy measures.

2. THE VALUATION OF ELECTRICITY EXTERNALITIES

2.1 The Background to Externality Valuation

The presence of an externality indicates a situation in which the first theorem of welfare economics fails to apply. In other words, externalities cause market failures, which then lead to a failure to accomplish Pareto efficiency. The market failures caused by externalities can arise in both production and consumption activities. Either way, they give rise to resource allocations that are nonoptimal from the point of view of society as a whole. Fig. 1 shows the relationship among a number of key concepts associated with market failure through the existence of externalities. The private costs facing a producer represent the best alternative uses of resources available as reflected by the market prices of the specific resources used by the producer. However, the social costs of production equal private costs plus external costs and represent the best alternative use of resources available to society as a whole. Therefore, a negative externality is equivalent to the difference between the private and social costs of an activity such as electricity production. Because by definition there exists no market for an external impact, a profit-maximizing utility has no incentive to integrate this effect into the decision-making process. Thus, private costs are lower than social costs. However, the existence of a difference between private and external costs is not inevitable. If the external costs can—by means of various fiscal or regulatory instruments—be "internalized" in electricity prices (i.e., be made private), producers and consumers in electricity markets will have an

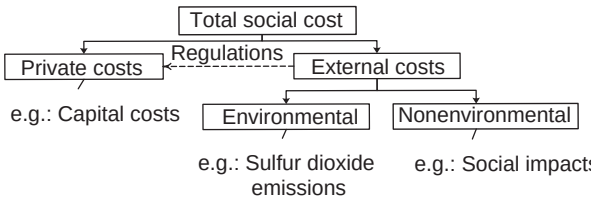


FIGURE 1 Total costs to society of a productive activity.

incentive to undertake actions that help to mitigate the negative environmental impacts of electricity use.

In his seminal work, Coase argued that bargaining between the polluter and the affected agents can, under certain circumstances (e.g., low transaction costs and full information), internalize externalities and achieve an efficient outcome. However, in most cases, because of the large number of parties involved, this calls for some kind of government intervention. One way of correcting the inefficiency of an external cost is the use of so-called Pigovian taxes as originally suggested by Pigou. This implies setting a tax equal to the value of the marginal external cost (at the optimal level of the activity in question) so that the private decision maker is provided with an incentive to economize not only on the traditional input factors but also on unpriced goods and services such as those provided by the natural environment. However, this solution to the externality problem requires that the tax authority is able to identify the external cost function. This returns us to the challenge of how to go about assessing the size of this function and, hence, the value of the damage caused by a negative externality (or the benefits incurred by a positive one).

2.2 Externality Valuation in Theory

The theoretical basis for the economic valuation of externalities (and thus environmental impacts) is outlined in the welfare economics literature. This strand of research recognizes that the economic value of a resource or service is ultimately a function of individual preferences and, therefore, that the tool for analyzing welfare changes is utility theory. For example, consider the welfare of individual i and assume that her utility (u_i) depends on a vector of private consumption possibilities ($\mathbf{x}$) and on a vector of environmental and resource flows ($\mathbf{q}$) given to the individual (i.e., environmental quality). Hence, individual i's utility function,

$$u_i = u_i(\mathbf{x}, \mathbf{q}), \tag{1}$$

shows the various combinations of $\mathbf{x}$ and $\mathbf{q}$ from which the individual derives utility. Furthermore, assume that some of the elements in $\mathbf{q}$ have positive prices, represented by the vector $\mathbf{r}$. The individual maximizes her utility subject to the budget constraint,

$$M_i = \mathbf{px} + \mathbf{rq}, \tag{2}$$

where M_i is individual i's income and $\mathbf{p}$ is a vector of prices for $\mathbf{x}$. The first-order condition for this maximization problem implies that individual i's conditional demand function (i.e., conditioned on the imposed $\mathbf{q}$), x_i, for marketed goods can be written as

$$x_i = x_i(\mathbf{p}, M_i - \mathbf{rq}, \mathbf{q}). \tag{3}$$

Inserting Eq. (3) into Eq. (1) gives the conditional indirect utility function, v, for individual i:

$$v_i = v_i(\mathbf{p}, M_i - \mathbf{rq}, \mathbf{q}). \tag{4}$$

The use of indirect utility functions enables us to express welfare measures in money metric terms. Assuming that $\mathbf{q}$ consists of only one element q, that $\mathbf{x}$ is the numeraire, and that q is an environmental "good" (i.e., *ceteris paribus*, individual i would choose more of q if given the option), consider a project that causes the environmental quality to improve from q^0 to q^1 (i.e., $q^1 > q^0$). The project causes a positive change in individual i's indirect utility (or welfare) so that

$$\Delta v_i = v_i(\mathbf{p}, M_i - rq^1, q^1) - v_i(\mathbf{p}, M_i - rq^0, q^0) > 0. \tag{5}$$

Thus, given the presence of the project, individual i is made better off. However, because utility is not directly observable and because environmental "goods" are not part of market transactions, we need to find alternative ways of assessing the value of this welfare change. Theoretically, two standard monetary measures of quality-based welfare changes are the compensating surplus and the equivalent surplus. The first measure, compensating surplus (CS), is given by the solution to

$$v_i(\mathbf{p}, M_i - rq^0, q^0) = v_i(\mathbf{p}, M_i - rq^1 - CS, q^1). \tag{6}$$

In this case, CS is equal to the maximum willingness to pay (WTP) for the environmental improvement. Second, consider the equivalent surplus (ES) measure, which is given by the solution to

$$v_i(\mathbf{p}, M_i - rq^0 + ES, q^0) = v_i(\mathbf{p}, M_i - rq^1, q^1) \tag{7}$$

Here, ES equals the minimum willingness to accept (WTA) compensation for the environmental improvement not occurring. The interpretation of the CS and ES measures is reversed in the case of an environmental quality deterioration, where CS is equal to the minimum WTA and ES is equal to the maximum WTP.

In empirical studies, it is generally the case that WTA measures tend to be substantially higher than WTP measures for the same change. Thus, the choice of WTP or WTA as a measure of economic value may significantly affect the size of the resulting externality estimate. Even though there is no firm theoretical basis for choosing one approach rather

than another, *WTP* is generally advocated as the most appropriate measure of changes in welfare. This is partly because *WTA* is not constrained by income and, therefore, is seen to introduce an incentive problem. On the other hand, the sensitivity of *WTP* to income introduces certain equity issues.

In sum, the economic valuation of externalities, and thus of many environmental impacts, builds on the assumption that people seek to satisfy their preferences, which are exogenously determined, complete, continuous, and ethically unchallengeable (subjective). The environment is essentially treated as any other private commodity, and people are willing to consider trade-offs in relation to the quantity or quality of environmental "goods." According to the welfare economics literature, the appropriate role of policy in the field of energy externalities is to aggregate the monetary estimates of individual preferences and to weigh them against other (more tangible) economic benefits and costs. Thus, the economics of nonmarket valuation builds on (1) clear but relatively restrictive behavioral assumptions (i.e., utility maximization), (2) a sense of society as the sum of the preferences of its individual members, and (3) a view of the task of public policy involving the internalization of external impacts and with utilitarianism as the ethical principle guiding social choice.

2.3 Externality Valuation in Practice

In practice, there are two basic methodological approaches used for the valuation of external costs in the energy sector: the abatement cost approach and the damage cost approach.

The abatement cost approach uses the costs of controlling or mitigating damage or the costs of meeting legislated regulations as an implicit value of the damage avoided. The rationale behind this approach is that legislatures are assumed to have considered the willingness of the public to pay for alleviation of the damage in setting the standard, thereby providing a revealed preference damage estimate no less reliable than the more direct valuation methods. One of the serious caveats with this approach is that it relies on the rather strong assumption that decision makers always make optimal decisions, that is, that they know the true abatement and damage costs. Also, a necessary condition for social optimality is that the abatement costs used are derived from the pollution control strategy that provides the least cost of control. If not, the estimates cannot reflect damage costs adequately.

Another limitation of the abatement cost approach is that society's preferences change over time as information, analysis, values, and policies change. Hence, past revealed preferences might bear little relation to actual impacts today and their current value to society. For instance, the implicit value of carbon dioxide (CO_2) emissions indicated by a revealed preference analysis would, in many cases, be very low because regulations targeted toward this problem are relatively few in number and low in scale. This built-in "tautology" of the approach means that estimates need to be constantly revised as regulations and policies change. Perhaps more important, given that policy is by definition optimal, the analysis provides no room for relevant policy implications; therefore, one can question why the analysis is needed in the first place.

The damage cost approach is aimed at measuring the net economic damage arising from negative externalities by focusing more or less directly on explicitly expressed preferences. This approach can be subdivided into two main categories: top-down and bottom-up. Top-down approaches make use of highly aggregated data to estimate the external costs of, say, particular pollutants. Top-down studies are typically carried out at the national or regional level, using estimates of total quantities of pollutants and estimates of total damage caused by the pollutants. Specifically some estimate of national damage is divided by total pollutant depositions to obtain a measure of physical damage per unit of pollutant (Fig. 2). These physical damages are then attributed to power plants and converted to damage costs using available monetary estimates of the damages arising from the pollutants under study. The main critique against the top-down approach is that it "generically" cannot take into account the site specificity of many types of impacts or the various stages of the fuel cycle. Another argument that has been raised

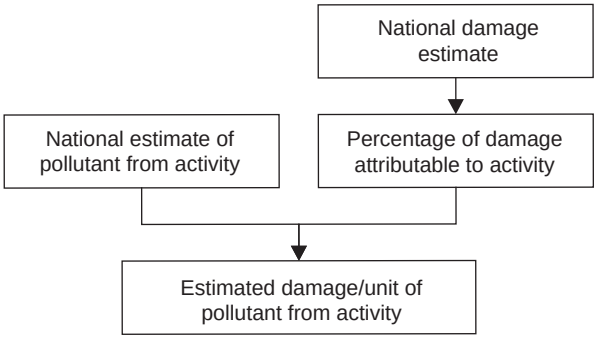


FIGURE 2 The top-down approach.

against the approach is that it is derivative because it depends mostly on previous estimates and approximations.

In the bottom-up approach, damages from a single source are typically traced, quantified, and monetized through damage functions/impact pathways (Fig. 3). This method makes use of technology-specific data combined with dispersion models, information on receptors, and dose–response functions to calculate the impacts of specific externalities. The bottom-up approach has been criticized because applications of the method have unveiled a tendency for only a subset of impacts to be included in assessments, focusing on areas where data are readily available and rendering impact pathways easier to establish. That is, in bottom-up applications important impacts may be left out, especially in cases where data are not readily available. Furthermore, the bottom-up approach relies on models that might not adequately account for complexities in "the real world" given that there may be synergistic effects between pollutants and environmental stresses and

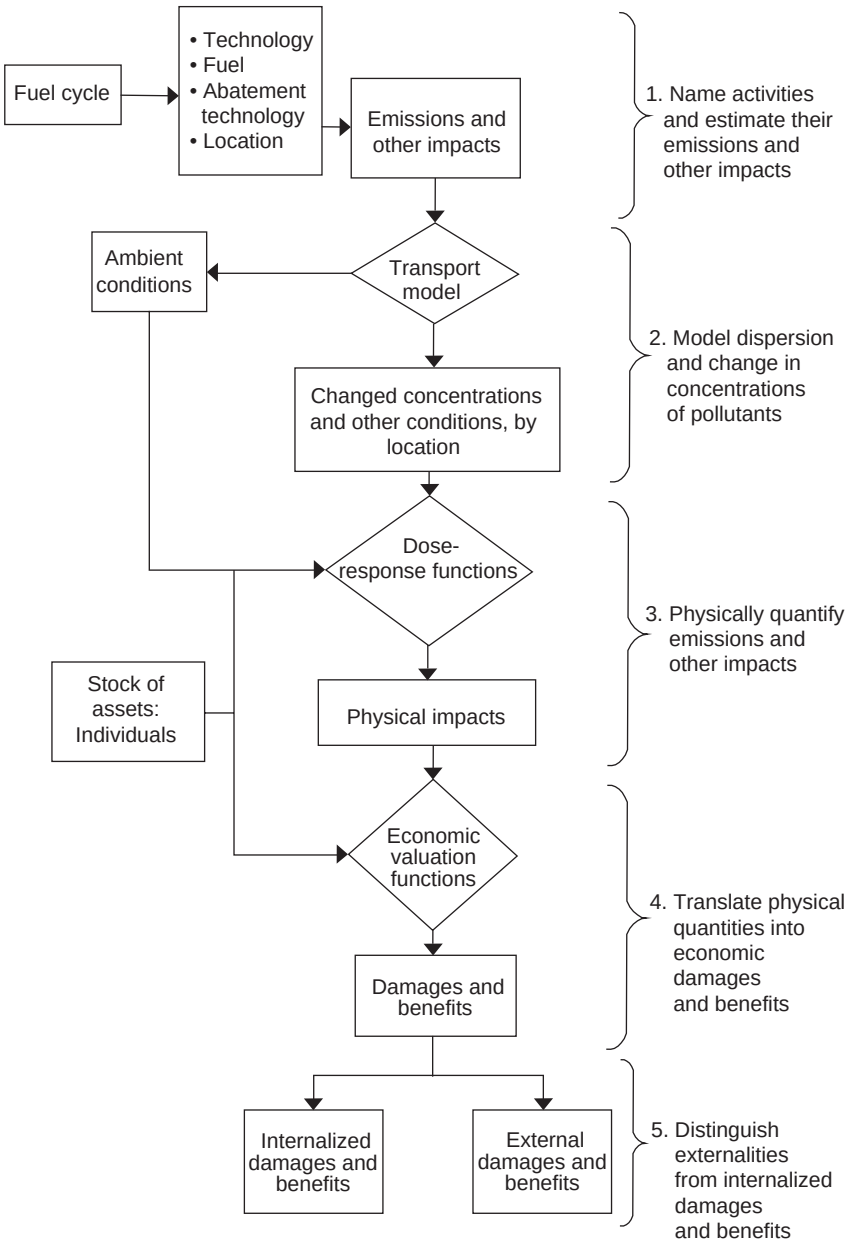


FIGURE 3 The impact pathway (bottom-up approach).

that there may be problems in establishing the timing of effects (e.g., between exposure and impact). Therefore, the argument is that bottom-up approaches might not be sufficiently transparent. Still, this is the approach that, due to its focus on explicit estimates of economic welfare (rather than on implicit ones such as in the abatement cost approach), appears to be most in line with economic theory. As is evident by the methodological choices of recent externality studies, it is also the most preferred approach to the empirical assessment of externalities in the electricity sector (see Section 3.1).

There are several ways of monetizing the identified externalities. The first two approaches discussed previously, abatement cost and top-down damage cost, directly provide a monetary estimate of the damages associated with the externalities. However, in the third approach, bottom-up damage cost, one needs to translate the physically quantified impacts into monetary terms. By definition, externalities are external to the market in question. However, certain costs (e.g., pollution damage to property or agriculture) may be external to electricity markets but subject to pricing in other markets. In such cases, the relevant market prices are used as a direct basis for valuation. Impacts from externalities that are not reflected in prices in any market must then rely on nonmarket valuation methods, which can be subdivided into direct methods and indirect methods.

The direct methods attempt to create a hypothetical market for the environmental good. These methods are direct in the sense that they are based on direct questions to households about willingness to pay. The direct methods possess the advantage that they can assess total economic values, that is, both the use and nonuse values (i.e., existence values) associated with the good. Well-known examples of such techniques include contingent valuation and choice experiments. The indirect methods take their basis in the actual (rather than the hypothetical) behavior of individuals. In terms of willingness to pay, the welfare effects show up as changes in costs or revenues either in observable markets or in markets closely related to the resource that is affected by the externality. Thus, the damage is indirectly valued using an existing relation between the externality and some good that is traded in a market. Examples of indirect methods are hedonic pricing (e.g., looking at property values) and travel costs (e.g., looking at amenity values).

3. ELECTRICITY EXTERNALITY STUDIES

This section gives an overview of the results from the multitude of externality studies carried out during the 1980s and 1990s. This is followed by a brief look at the results from one particularly detailed study, the European ExternE project.

3.1 Overview of Results

A considerable number of externality studies were carried out during the 1980s and 1990s. Figure 4 presents the results from 63 externality studies, with the numbers in square brackets showing the total number of observations for each fuel source (note the use of logarithmic scale). The ranges overlap, rendering ambiguous the ranking of various fuels with respect to externality impacts. Indeed, it is possible, using these overlapping ranges, to argue for virtually any ranking order one likes across the options in question. However, when the preponderance of findings are taken into account, the results do suggest that fossil fuel-fired power (in particular coal and oil) gives rise to the highest external costs, whereas some of the renewable energy sources (solar, wind, and hydropower) tend to have the lowest.

For a specific fuel source, the difference between low and high values is substantial, and this is also true if one looks at single studies. The ranges reported can often vary from a tiny fraction of electricity market prices and the private costs of producing power to a number that is significantly above private cost levels. However, the wider sensitivities are not always fully conveyed or discussed in individual studies, where results are often expressed with a fine degree of precision and are routinely referred to as "real," "full," or "true" values. When attention extends to the literature as a whole, the scale of the discrepancies between the results obtained by various studies for similar fuels does raise concerns about the validity and reliability of an approach to regulatory appraisal based purely on environmental valuation.

This said, it is of course entirely reasonable that there should be some degree of variability among studies. This may be due, for instance, to factors such as (1) a focus on different technologies (e.g., implying separate emission factors); (2) different site-specific characteristics (e.g., population density, income, transport distances); (3) differences in scope (e.g., different selections of externalities may be included, different taxonomies may be used to organize effects,

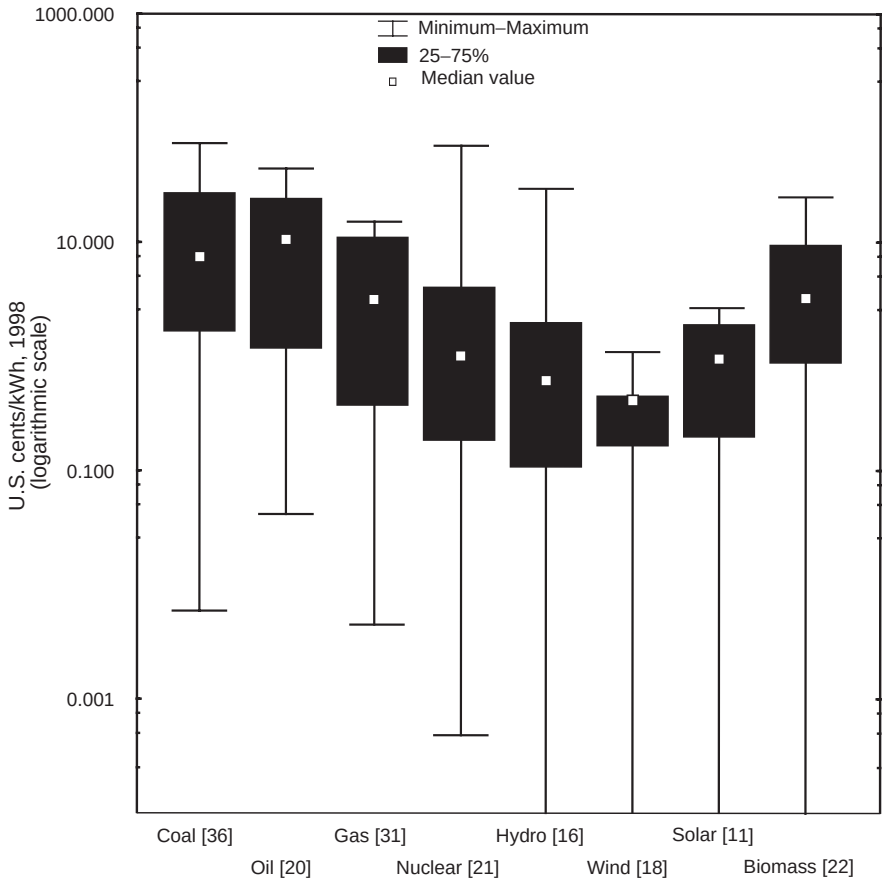


FIGURE 4 Range of external cost estimates for various fuel sources. From Sundqvist, T., and Söderholm, P. (2002). Valuing the environmental impacts of electricity generation: A critical survey. *J. Energy Lit.* **2,** 3–41. Reprinted with the permission of the Oxford Institute of Energy Studies.

different stages in fuel cycles or resource chains may be evaluated); (4) different trade-offs may be implicit among the various dimensions of environmental effects (e.g., injuries and disease, immediate or latent effects, concentrated or dispersed effects); (5) different treatments of complexity, uncertainty, or gaps in knowledge (e.g., *ex ante* rather than *ex post* data; likely technological trajectories; indirect, contingent, or synergistic effects). Some of these issues are further discussed in the following section. For the moment, note that it is difficult to determine whether such factors justify such large variability in results, especially given that there is no objective metric against which to compare the "truth" of the contending empirical estimates.

Sundqvist and Söderholm provided an overview of roughly 40 externality studies and showed that, in addition to the apparent disparity of results, several other conceptual issues are of importance. First, most of the fuel sources available for power generation, including coal, oil, natural gas, nuclear, hydro, wind, solar, and biomass as well as (in a few cases) lignite, waste incineration, geothermal, peat, and orimulsion, have been addressed in previous valuation efforts. However, most studies focus on the traditional fuels such as coal and nuclear. Thus, there is a tendency for many studies to focus on existing technologies rather than on those technologies that are generally accepted as being likely to play a significant role in the future (e.g., wind, biomass). In many cases, this is understandable given that empirical data clearly are more available for existing technologies than for emerging ones. Nevertheless, an important goal of externality valuation in the power sector has been to "level the playing field" in the selection between traditional and new generating technologies. This would probably require a stronger focus on promising but as yet uncommercialized technologies.

Second, a majority of the studies have been carried out for the developed world (mostly for Western

Europe and the United States). Thus, only in some rare cases has the focus been on developing countries, where the pressure for additional power capacity is by far the greatest. There are also reasons to believe that externality estimates should differ substantially between developing and developed countries. In developing countries, incomes are lower, sociocultural priorities are likely to differ, and the environmental effects of power production may be fundamentally different. An important example of the latter is the environmental externalities stemming from hydropower. Hydroelectric development in tropical climates may give rise to global warming impacts due to decay of vegetation left in the reservoir, whereas hydroelectric development in cooler climates will not. This raises serious concerns about transferring values from studies conducted in, say, Western Europe for use in a developing country context.

Third, examining the methodologies used over time reveals that the bottom-up damage cost approach has become the dominant paradigm, whereas the abatement cost and top-down approaches were used predominantly during the 1980s and early 1990s (Fig. 5). An important reason for this development is that the national implementation phase of the ExternE project relies solely on damage cost bottom-up models, and these studies together represent a large share of the total number of projects conducted during the latter part of the 1990s. However, this also indicates that the bottom-up model has been accepted as the most appropriate method with which to assess power generation externalities. The ExternE project has largely served as a vehicle in the methodological development of externality valuation. The sophistication and scope of the ExternE work and the detailed attention to methodology have been widely respected at the international level, and many followers rely heavily on the numbers and methods presented. However, this development raises the question of whether the choice of methodological approach (between abatement costs and damage costs) matters for the results. We return to this question in more detail later and suggest that this choice might very well matter. This, in turn, raises important concerns about the reliability of external cost valuation exercises in the power-generating sector.

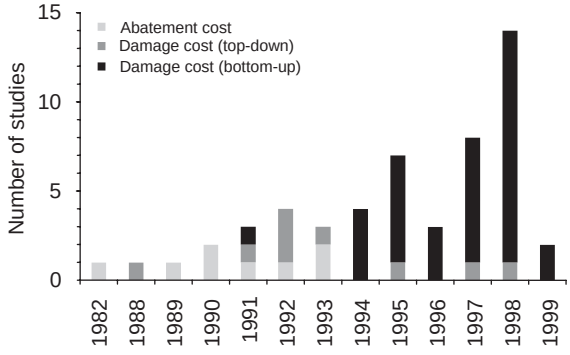


FIGURE 5 Methodological choice over time. From Sundqvist, T., and Söderholm, P. (2002). Valuing the environmental impacts of electricity generation: A critical survey. *J. Energy Lit.* **2**, 3–41. Reprinted with the permission of the Oxford Institute of Energy Studies.

3.2 An Example: The ExternE Core Project

As noted earlier, one of the most influential externality studies is the European ExternE project. This project aimed at evaluating the external costs of the various power generation fuel cycles in the European Union using the bottom-up approach. Because the ExternE project represents one of the most ambitious and internationally respected attempts to yield a "true" external cost estimate for the various power technologies, it serves well as an example. Tables I and II present the various power generation externalities quantified and priced within the ExternE core project. The ExternE core project was followed up by the so-called national implementation project, whose aim has been to develop a European Union-wide set of external cost data for the various fuel cycles and countries as well as methods for aggregation of results to make them useful for policymakers.

When inspecting Tables I and II, we first observe that most of the fuel cycles involve significant health and mortality impacts on humans (public and occupational). In this project, considerable attention was devoted to the evaluation of public health impacts and much was learned, especially about the importance of fine particle emissions. Overall, the results support the rather unexceptional conclusion that fossil fuels, especially coal, present more serious environmental impacts than do renewables such as hydro and wind. It is notable that the most favorable results are attributed to nuclear power. However, most of the externality damages priced in the ExternE core project are reported to reflect an uncertainty range characterized as "medium to high." This raises a number of salient critical issues bearing on the robustness and meaning of these results and their associated prescriptive rankings.

TABLE I

Externalities Priced within the ExternE Core Project: Coal, Oil, and Gas

Externality	Coal	Oil	Gas
Public health	Particulate matter, ozone, and accidents: mortality, morbidity, and transport impacts	Particulate matter and ozone: mortality, morbidity, and transport impacts	Particulate matter: mortality, morbidity, and transport impacts
Occupational health	Diseases from mining and accidents during mining, transport, construction, and dismantling	Accidents: death and injury impacts	Accidents: death and injury impacts
Agriculture	Sulfur, acidification, and ozone: crop and soil impacts	Sulfur, acidification, and ozone: crop and soil impacts	
Forests	Sulfur, acidification, and ozone damages	Sulfur, acidification, and ozone damages	
Marine	Acidification impacts	Accidents with oil tankers	Fishery: extraction impacts
Materials	Sulfur and acidification damages on surfaces	Sulfur and acidification damages on surfaces	Sulfur and acidification damages on surfaces
Amenity	Noise: operational road and rail traffic impacts		Noise: operational impacts
Global warming	CO_2, CH_4, and N_2O damages	CO_2, CH_4, and N_2O damages	CO_2, CH_4, and N_2O damages
Total estimate[a] (U.S. cents/kWh)	2.8–4.1	2.7–2.9	1.7

[a] The global warming impacts constitute roughly half of the reported external cost estimates for coal-, oil-, and gas-fired power.

TABLE II

Externalities Priced within the ExternE Core Project: Nuclear, Hydro, and Wind

Externality	Nuclear	Hydro	Wind
Public health	Radiation and nonradiation: mortality and transport impacts from operations and accidents		Accidents: travel to and from work
Occupational health	Radiation and nonradiation: mortality and transport impacts from operations and accidents	Accidents during construction and operation	Accidents during manufacturing, construction, and operation of turbine
Agriculture		Loss of grazing land	Acidification: damage on crops
Forests		Forest production loss due to flooding and land use	Acidification damages
Marine		Water supply and ferry traffic	Acidification damages
Materials			Acidification damages
Amenity		Visual amenity loss	Noise and visual amenity loss: Operational impacts
Global warming			CO_2, CH_4, and N_2O damages
Recreation		Fishing and hunting	
Cultural objects		Objects of cultural and archaeological interest	
Biodiversity		Terrestrial and aquatic ecosystems	
Total estimate (U.S. cents/kWh)	0.0003–0.01	0.3	0.1–0.3

4. CRITICAL ISSUES IN EXTERNALITY ASSESSMENT

This section discusses a number of critical issues concerning nonmarket valuation of energy externalities and focuses in particular on the role of externality valuation in policymaking.

4.1 Methodological Issues

As already noted, we need not necessarily be concerned about the variability in external cost estimates if we can identify rational reasons for the discrepancies. However, one possible reason for such discrepancies that would itself raise concerns is that the choice of methodology may have a significant impact on the results. By using the results from more than 40 electricity externality studies across eight fuel sources, Sundqvist provided some econometric evidence in support of the conclusion that methodological choice does matter for the results. He reported that the probability of obtaining a "low" externality value is, *ceteris paribus*, generally lower when the abatement cost approach or the top-down damage cost approach is used, whereas the opposite is true when the bottom-up damage cost approach is used. Abatement costs will (theoretically) be representative of damage costs if and only if they are derived from the least cost strategy, but normally most studies employ the most commonly used (or mandated) abatement technology when assessing pollution abatement costs. As a consequence, it can be expected that the results for the abatement cost applications would deviate. The reason why the top-down approach also tends to yield relatively high results is that practical problems may arise in attributing the "exact" damage to each individual source, and this may force researchers to rationalize and use standardized rules for the attribution process. These rules may fail to ascribe the aggregated damage to each and every individual source, especially smaller sources, thereby producing estimates for specific power plants that are positively biased, given that these plants normally are easily identifiable as well as significant sources of pollution.

A fundamental question raised by these results concerns the extent to which the three broad approaches to externality valuation are comparable. So far, we have suggested that the differences reported may be due primarily to a number of technical factors in analysis such as identifying the least cost control strategy and addressing all relevant end points. If these issues can be resolved, it might be hoped (at least in principle) that the methods could begin to generate more similar results. In what follows, we suggest a number of reasons why this is not necessarily the case. However, before turning to this, it may be worth reviewing one final observation that has been made concerning the pattern evident in the picture of environmental valuation results provided in Fig. 4.

Stirling observed that the scope for multiple divergent, but equally legitimate, "framing assumptions" is a feature held in common between environmental valuation and many other forms of regulatory appraisal, including risk assessment and decision analysis. However, one feature that is unique to neoclassical environmental valuation is that what is conveyed to policymaking is not simply the relative performance orderings for different options but also the particular numerical values themselves. These are expressed as particular values for monetary externalities, suitable for adoption as a basis for Pigovian taxes. Stirling argued that this may make environmental valuation peculiarly vulnerable to a phenomenon he referred to as a "price imperative." This describes a situation where, when faced with widely differing but equally valid possible assumptions, the choices made in analysis—implicitly or explicitly—lead results to fall into a domain that is most likely to be practically usable in policymaking. A Pigovian tax set at much less than one-tenth of market prices would have minimal impact on producer or consumer behavior. A Pigovian tax set at 10 times market prices would probably be politically unfeasible. This may be a factor in explaining why, despite ranges of variation spanning more than six orders of magnitude in Fig. 4, the bulk of valuation results lie well within one order of magnitude of existing wholesale electricity prices.

4.2 Externality Classification

The neoclassical welfare economics literature provides a relatively straightforward definition of the concept of externality. However, in practice, the choice of "relevant" externalities tends to differ among various valuation studies. In other words, some studies seem to differ considerably in their definition of what constitutes an externality. As an example, the inclusion of employment benefits (following a power generation investment) as an external benefit strains the definition of what an externality is. For employment effects to be considered external, the local and regional labor markets

must function poorly (i.e., market failures must be present) and workers must be immobile. However, some authors make a strong case for treating employment impacts as external benefits. For instance, to the extent that people obtain disutility (e.g., less self-confidence) from unemployment as such, this notion is valid. However, in such a case, the employment benefits should be compared across all fuel sources and not only for renewable energy projects (as is often the case).

For economists, the existence of externalities motivates regulatory action and the use of taxes and subsidies to improve efficiency. However, lay people and politicians often hold the view that regulations and other policy measures should be used to promote the "good" and discourage the "bad" in a much broader sense. This suggests that economists have to be careful in explaining what the results stemming from externality appraisals show and what they do not show. The preceding also raises the more philosophical question of what should be the ultimate end of policy: economic efficiency based on individual preferences or other value judgments formed in public deliberations (or any mix of these policy principles)? If the latter path is chosen, externality valuation may still provide an important input into the policy decision process, but the notion of some "total" (or "true") cost of electricity production appears less valid. We return to this question later.

4.3 Scope

Even when all of the external costs and benefits of a given power generation source have been identified, there remains the issue of choosing the appropriate level of scope for the valuation effort. At least two choices have to be made. First, which of the externalities are important enough to include in the valuation? Second, should the study address externalities across the entire fuel cycle (including downstream and upstream impacts) or focus only on the generation stage? Studies conducted in the past have made different choices in these respects, and this is likely to explain some of the disparity in the estimates presented. Let us first look at how this question has been dealt with in previous studies.

As a part of the process of deciding what externalities to include, many studies begin with a comprehensive screening of the relevant economic and scientific literature to determine which pathways and end points are likely to be important in the analysis. Thus, some of the externalities are *de facto* assigned a zero value on the basis of their presumed insignificance, whereas others are included in the analysis because the initial screening could not rule out the possibility of significant effects. It should be clear that this approach may lead to total externality values that are significantly downward biased. However, what is perhaps of more importance is those cases where externalities are left out of the analysis because there is insufficient scientific information to establish defensible monetary damage values. The most important example is global warming and CO_2 emissions. During the first phase of the ExternE project, it was noted that the environmental damage cost estimates for greenhouse gas emissions presented in the literature spanned a range of several orders of magnitude, and the main report concluded that "all attempts to value these impacts require important normative judgements, and therefore the potential for synthesis or consensus is remote."

After additional research efforts within the ExternE project, these general conclusions appear to remain valid. However, it is generally agreed that the external costs of CO_2 emissions are substantial and, therefore, may constitute a large share of the total value. Thus, the environmental economist faces a dilemma: is it better to leave out potentially important external damages from the valuation and present biased estimates, or should one make use of rough proxy estimates (e.g., mitigation costs) to provide (or at least approach) some kind of "full-cost" estimate? In the end, the ExternE project chose the latter path and recommended the use of "minimum," "central," and "maximum" estimates.

This raises a number of important issues. The choice of which externalities to include in the assessment cannot be entirely objective but is rather a matter of judgment. The judgment that has to be made is essentially whether the externality under consideration is "mature" enough to "undergo" economic valuation. However, this is not only a question of whether the scientific knowledge is more or less established; it also involves the issue of whether the public is sufficiently informed and, hence, able to form a valid opinion about the issue at hand. Again, economic valuation is ultimately about measuring people's given preferences toward goods, but if no relevant preference structure for a particular good exists (as with global warming), the valuation effort may be seen as arbitrary. We also noted previously that the use of abatement cost estimates (i.e., regulatory-revealed preferences) provides poor substitutes given that such estimates rely on the notion that all relevant preferences

have already been perfectly integrated into policy decisions.

This raises questions about the decision in the ExternE project to present rough estimates of global warming impacts. Moreover, global warming also entails important ethical and justice-related issues such as the question of whether human lives lost in developing countries should count less than corresponding lives in the developed world and to what extent impacts affecting future generations should be discounted (if at all). In such a case, the initial challenge of policy might not lie in "measuring" and aggregating individual preferences but rather might lie in specifying the conditions for public discourse over common ways of understanding the meanings and relative priorities attached to the various pertinent issues.

The preceding also suggests that any notion of "full," "total," or "true" cost of electricity has to be understood as hypothetical at best. The fact that externality studies leave out potentially important externalities might not be a problem per se. On the contrary, nonmarket valuation builds on specific basic assumptions, primarily about the behavior and preferences of the public, and if these are not well articulated, it may be better to refrain from monetary valuation. These concerns are likely to be particularly valid for "goods" with which people have relatively little past experience and that are "complex" and involve far-reaching and unknown consequences. Global warming, like the risk profile for nuclear accidents and radioactive waste, provides examples.

Finally, the choice of what parts of the fuel cycle to focus on also complicates the assessment of the full cost of electricity. Where should one draw the appropriate analytical boundaries and, thus, the limits of the external cost assessment? This choice is perhaps somewhat more technical than that of what externalities to include, but it is no less important. It is also complicated by the fact that the values of many upstream and downstream impacts tend to be highly site specific and depend, for instance, on the mode of transportation and the location of the place where the fuel is extracted. One problem is how one is to define the "entire" fuel supply chain. For instance, if we include in the analysis the externalities incurred by transporting fuels from the mine or well to the power station, should we then not also include the corresponding transport-related externalities involved in bringing building materials and workers to the power plant construction site? The choice of a relevant level of analysis is not easy, and this also suggests that the notion of any "total" or "true" cost of power generation is flawed. Ideally, this choice should be guided by the need for relevant policy information.

4.4 Parameter Input Assumptions

There are essentially two different categories of parameter input assumptions made in electricity externality studies: technical assumptions (e.g., energy efficiency, dose–response functions, emission factors) and economic assumptions (e.g., monetary values, discount rates). Arguably, the assumptions regarding, for example, CO_2 emissions are major drivers of the results. Another striking example concerns the assumptions made about the monetary values used to value mortality impacts due to air pollution. Many studies use the value of a statistical life (VOSL) to assess these types of impacts. For instance, the ExternE core study relies on a VOSL value of U.S. $2.6 million for Europe. In the national implementation part of the ExternE project, the decision was made to introduce an alternative measure on which to base the valuation of mortality impacts due to air pollution. This is the so-called years of life lost (YOLL) approach, which essentially assigns a WTP value to the risk of reducing life expectancy rather than to the risk of death. The YOLL values attributed to the mortality impacts are, as is evident from Fig. 6, reduced by up to two orders of magnitude as compared with the values based on the VOSL method. In addition, the core project that relied on the VOSL approach did not include values for chronic mortality impacts due to air pollution, something that the national implementation studies do.

Thus, it is clear that the assumptions underlying the valuation of human health and mortality impacts, global warming impacts, and dose–response functions are major drivers of external cost estimates. Clearly, this also provides a major explanation as to why the total externality estimates from various studies often differ considerably (even for the same fuel source).

So far in this subsection, we have noted that the input assumptions made in externality assessments play a significant role in affecting the overall external damage estimates and, thus, in guiding policy. To some extent, this is how it should be, of course; different sites and different technologies (depending, e.g., on vintage) incur different emission impacts and, hence, different damages. Still, the large sensitivity in results due to parameter input assumptions also creates problems for policymakers. First, policymakers often wish to gain some notion as to which

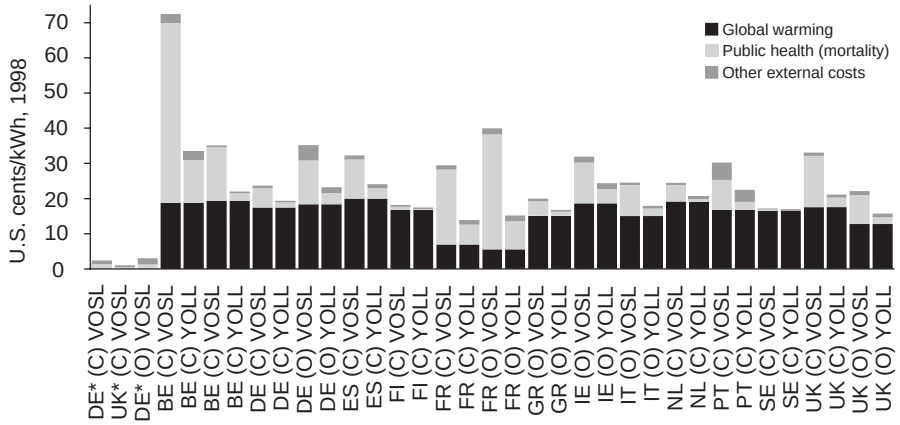


FIGURE 6 External cost estimates in the ExternE core (*) and national implementation projects: coal (C) and oil (O) fuel cycle. From Sundqvist, T., and Söderholm, P. (2002). Valuing the environmental impacts of electricity generation: A critical survey. *J. Energy Lit.* **2**, 3–41. Reprinted with the permission of the Oxford Institute of Energy Studies.

are the most important external costs (i.e., environmental damages) for each power generation source. This enables them to target policy measures, such as subsidies, research and development (R&D) support, and regulations, toward the most important impacts. However, previous research efforts have provided only limited guidance on this particular point. Also, in many studies, a single category of external damages seems to dominate the total external cost estimates. For example, in the initial ExternE coal study, human health-related impacts dominate the aggregate damage (some 76–95% of the total), whereas global warming impacts constitute an important part of the estimated damages from the coal fuel cycle in the national implementation part of ExternE (Fig. 6). Second, in deregulated electricity markets, one of the most important uses of external cost estimates are as inputs in the development of environmental taxes and tradable permit schemes. However, for this reason, policymakers would need relatively "safe bets" about the general impacts involved, but so far previous studies have provided only wide ranges of estimates. Again, the impact of global warming is a good illustration of this.

This discussion raises a fundamental issue in nonmarket valuation that is rarely touched on in the electricity externality debate. Most environmental economists would agree that environmental valuation requires that a relevant "project" has been defined and that it involves the choice between two or more relevant alternatives. In the case of electricity externalities, these "projects" are normally the investments in various power plants (few companies or governments undertake investments in entire fuel cycles). However, in externality studies, these investment projects are often hypothetical; that is, they do not represent an existing (real) situation, and valuation estimates are transferred from other studies. The problem with this is that according to the literature on nonmarket valuation (and indeed that on market valuation as well), economic values are context dependent and project specific. In other words, it might not make much sense to talk about a universal *WTP* for avoiding 1 ton of SO_2 being emitted, even though that is just what many policymakers would like to know about because they often prefer more or less harmonized standards and taxes across different regions.

4.5 Public versus Private Preferences

Critics of environmental valuation often stress that the methods used rely on overly restrictive assumptions. This implies that such studies often produce poor descriptions of the environmental values that people actually hold as well as underlying processes of preference formation. Specifically, externality valuations rest on an assumption that individuals aim at maximizing personal utility and that they possess well-articulated, ethically unchallengeable, and exogenous preferences for any environmental good. They also effectively assume that these individual preferences may be combined to yield an aggregate preference function for society as a whole. Both types of assumption are quite seriously problematic.

This second assumption is problematic even when approached within the confines of neoclassical welfare economics on which environmental valuation is

built. It glosses over the profound intractabilities involved in seeking definitively to compare the intensities of subjective preference displayed by various social agents. Indeed, even where social choices are addressed simply in ordinal terms as a matter of relative option rankings, economist Kenneth Arrow went a long way toward earning his Nobel Prize by demonstrating formally that it is impossible, under basic axioms of rational choice in neoclassical welfare economics, to guarantee any definitive aggregation of preference orderings in a plural society. In other words, the business of aggregating individual preferences, lying at the heart of neoclassical approaches to environmental valuation, violates its own well-established theoretical principles.

Beyond this, the first assumption raises a number of further concerns arising outside the narrow field of rational choice in neoclassical welfare economics. For instance, rather than being taken simply as an expression of personal utility maximization, environmental values may instead be seen to hold a broad ethical content. Because ethics is a matter for cultural and political discourse, environmental valuation should, it is argued, be endogenous to the political process and rely on social agreements. In other words, the initial challenge of environmental policy might lie not in "discovering" private preferences but rather in specifying the conditions for public discourse over what is worth valuing and for what reason.

Sagoff claimed that individuals essentially have at least two different preference orderings. Their private preferences reflect only their own well-being as consumers of private goods, whereas the public preferences reflect moral values about what persons, as citizens, believe is right and just for society as a whole. In their roles as citizens, people may express a rights-based (or deontological) belief system, which denies the principle of utilitarianism (and trade-offs) and instead recognizes the priority of the right over the good. Such a moral position is likely to be particularly prevalent in the case of environmental goods. The environment is often the subject of ethical concern and involves many cross-cutting dimensions that cannot be causally separated. In this way, the "market analogy" and "commodity fiction" of environmental valuation may break down.

There is empirical evidence that people express public, rather than private, preferences when considering complex environmental issues and when confronted with *WTP* questions in contingent valuation surveys. It is also the case that the consumer–citizen distinction provides one explanation of some puzzling phenomena frequently observed in contingent valuation studies. For example, a large number of studies have found very large discrepancies between *WTP* and *WTA* measures for environmental goods, with the latter often being several times larger than the former. This may be explained by the notion that in the *WTA* case, if the environmental good is morally considerable, the acceptable level of compensation will be extremely high or even undefined.

The distinction between private and public preferences is highly relevant when considering the external costs of power generation. Most important, many power generation externalities are either relatively novel (e.g., the risk profiles of nuclear power) or complex (e.g., ecosystem changes due to hydropower developments). For these types of externalities, most people have yet to find an overall ethical position; consequently, there are no "given" preferences on which economic valuation may rely. In addition, most power generation fuel cycles involve significant impacts on the health and deaths of humans. These impacts raise a moral dilemma: to what extent should we treat humans as means to an end (utility) or as ends in themselves? Such a question should ideally be resolved within the realms of public discourse. Thus, the social choice problem with respect to many energy and environmental issues is first, about advancing common ways of understanding the pertinent issues. This implies that environmental research in the social science field must increasingly address the instruments and content of political and moral debate and not simply the technicalities of established valuation methods. In other words, the process may count every bit as much as the outcome.

The preceding discussion also adds a new perspective to the observed differences in reported externality estimates between the abatement cost approach and the damage cost approach. The two methods involve different ethical bases. For example, the damage estimates developed within the ExternE project are considered *ex ante* (i.e., the damages themselves determine whether one power source is "better" than another), whereas the estimates derived using the abatement cost approach are *ex post* (i.e., the price is an outcome of a political process and does not play a direct role in the decision). Therefore, the damages developed in "advance" (as in the ExternE project) might not be directly comparable to "implicit" estimates that are based on the cost of abating as revealed by decision makers given that

they reflect different reasoning processes. Policymakers, in their formulation of regulations, may very well rely on ethical foundations other than economic welfare theory.

5. CONCLUSIONS

This discussion has considered the theoretical framework, reviewed the principal methodologies, and surveyed the detailed results associated with the valuation of external environmental costs of electricity production. Given the magnitude and pressing nature of the environmental imperatives in this area, it is clear that serious careful study of the kind that is well represented in this field has a crucial role to play in informing policy. However, it supports neither the cause of rigorous analysis nor that of environmental protection to neglect the series of profound difficulties and intractable challenges that attend neoclassical environmental valuation. In particular, it is important to note that the theoretical foundations are contestable, the methods are problematic, and the practical results are highly variable and sensitive to divergent "framing assumptions".

Overall, the externality studies carried out have provided a welcome contribution to increasing the knowledge about the economic effects associated with the environmental problems in the electricity sector. However, more research is likely needed before these and future studies will be of practical use in shaping policymaking. Thus, perhaps the main positive conclusion that may be drawn from this critical survey is that the real value of environmental externality studies lies not so much in the particular numerical results they yield but more in the opportunity they provide to focus on the key sources of variability and sensitivity. In this regard, there may be considerable potential to combine these methods with techniques such as sensitivity and scenario analysis to systematically explore the implications of various possible frameworks and assumptions. By embedding such broader based approaches in inclusive deliberative procedures involving different industry and public stakeholders, the emphasis might usefully shift away from the numbers themselves and toward the contending perspectives and assumptions that "frame" these numbers. In this way, environmental valuation might contribute toward establishing a truly well-informed and robust basis for future environmental decision making in the electricity sector.

SEE ALSO THE FOLLOWING ARTICLES

Electrical Energy and Power • Electricity, Environmental Impacts of • Electric Power: Critical Infrastructure Protection • Electric Power Generation: Fossil Fuel • Electric Power Measurements and Variables • Electric Power Reform: Social and Environmental Issues • Electric Power Systems Engineering • Electric Power: Traditional Monopoly Franchise Regulation and Rate Making • Electric Power: Transmission and Generation Reliability and Adequacy

Further Reading

Arrow, K. (1963). "Social Choice and Individual Values." Yale University Press, New Haven, CT.

Arrow, K., Solow, R., Portney, P. R., Leamer, E. E., Radner, R., and Schuman, H. (1993). Report of the NOAA Panel on Contingent Valuation. *Federal Register* **58,** 4601–4614.

Baumol, W. J., and Oates, W. E. (1988). "The Theory of Environmental Policy." 2nd ed. Cambridge University Press, Cambridge, UK.

Coase, R. H. (1960). The problem of social cost. *J. Law Econ.* **3,** 1–44.

European Commission. (1995). "ExternE: Externalities of Energy," vols. 1–6. Office for Official Publications of the European Communities, Luxembourg.

European Commission. (1999). "ExternE: Externalities of Energy," vols. 7–10. Office for Official Publications of the European Communities, Luxembourg.

Garrod, G. D., and Willis, K. G. (1999). "Economic Valuation and the Environment, Methods and Case Studies." Edward Elgar, Cheltenham, UK.

Kelly, J. (1978). "Arrow Impossibility Theorems." Academic Press, New York.

Krewitt, W. (2002). External costs of energy: Do the answers match the questions? Looking back at 10 years of ExternE. *Energy Policy* **30,** 839–848.

Oak Ridge National Laboratory and Resources for the Future. (1994–1998). "External Costs and Benefits of Fuel Cycles," report 2-8. McGraw–Hill Utility Data Institute, Washington, DC.

Pigou, A. C. (1924). "The Economics of Welfare." Macmillan, London.

Sagoff, M. (1988). "The Economy of the Earth." Cambridge University Press, Cambridge, UK.

Sen, A. K. (1970). "Collective Choice and Social Welfare." Elsevier North-Holland, New York.

Spash, C. L. (1997). Ethics and environmental attidtudes with implications for economic valuation. *J. Environ. Mgmt.* **50,** 403–416.

Stirling, A. (1997). Limits to the value of external costs. *Energy Policy* **25,** 517–540.

Sundqvist, T. (2003). What causes the disparity of electricity externality estimates? *Energy Policy*, in press.

Sundqvist, T., and Söderholm, P. (2002). Valuing the environmental impacts of electricity generation: A critical survey. *J. Energy Lit.* **2,** 3–41.

Electric Power Measurements and Variables

MARIESA L. CROW and NIRUP SHETTY
University of Missouri, Rolla
Rolla, Missouri, United States

1. **A Brief History of Electrification**
2. **Electric Power**
3. **Power Transformers**
4. **Transmission System Components**
5. **Substations**
6. **Distribution Systems**
7. **Power System Protection**

Glossary

active power The part of the instantaneous power that can do real work. Real work is done when the resistive component of a load consumes electrical energy. Active power is measured in watts (W).

apparent power The total power (the complex vector sum of active and reactive power), measured in volt-amperes (VA).

bundling Transmission lines are bundled when there is more than one conductor per phase in extra high-voltage lines.

distribution The process of delivering electricity from substations to customers. The distribution system "steps down" power from higher voltage transmission lines to a level that can be used in homes and businesses.

grid The transmission network through which electricity moves from suppliers to customers.

load A device that uses the power supplied from the source for doing work.

reactive power The part of the instantaneous power that does no real work. It is absorbed by the inductive and/or capacitive components of the power system. It is measured in volt-amperes reactive (VAR).

short circuit A fault that occurs in an electric circuit or apparatus usually due to imperfect insulation causing a large current to flow through the system to ground. A short circuit has the potential of inflicting significant damage to the system.

substation The set of apparatus in a power system network that is used to change some characteristic, such as frequency or voltage, of power supplies.

transformer A device that is used to convert high alternating voltage to low alternating voltage and vice versa.

transmission The process that transports electric energy in bulk form, usually at high voltages, from a source of supply to a distribution system.

The modern electric power system is comprised of three primary entities: the generating system, which produces the electricity; the transmission system, which transfers the electricity over high-voltage transmission lines to regional substations; and the distribution system, which distributes the electricity to local customers. The transmission system consists of transmission lines and structures, substations, and transformers to convert the electricity and one voltage level to another. The substations also house a variety of equipment to protect against lightening, short circuits, and other faults that may occur in the system that adversely affect the distribution of the electricity to the customers. Electric power system customers range from residential home and small business owners to commercial customers and large industrial customers. Each type of customer impacts the transmission and distribution of electricity in different ways during daily and seasonal electric power variations.

1. A BRIEF HISTORY OF ELECTRIFICATION

The electrical utility industry is arguably the largest and most complex man-made system in the world and has been hailed as one of the greatest engineering achievements of the 20th century. The invention of the electric incandescent lamp by Thomas A. Edison

 © 2004 Elsevier Inc. All rights reserved.

in 1878 marked the first step toward electrification. Edison is also credited with first proposing the idea of a centrally located power station for distributed lighting and developed the first electric network in the United States providing direct current (DC) power for lighting the lower Manhattan area of New York City. After the invention of the transformer, alternating current (AC) power became more popular than DC, with the first single-phase AC line being constructed in Oregon in 1889. The first three-phase system was installed by Southern California Edison at 2300 V in 1893. The development of electric power generation, transmission, and distribution has been increasing steadily ever since. From the construction of the first 220-V DC grids by Thomas Edison to the recent construction of AC grids with transmission voltages of 765,000 V, the drive to provide more reliable and efficient power has led to many significant developments in the field.

2. ELECTRIC POWER

The sinusoidal waveforms that comprise the voltages in most three-phase power transmission system typically have the form $v(t) = V\cos(\omega t)$ and the current has the form $i(t) = I\cos(\omega t + \phi)$, where ω is the system frequency (typically either 50 or 60 Hertz) and ϕ is the power factor angle. The angle ϕ is introduced due to the inherent inductive or capacitive nature of the transmission system and the system load, and it is the angle difference between the current and voltage waveforms. In an inductive circuit, the current lags the voltage, whereas in a capacitive circuit the current leads the voltage. In a purely inductive system, the current will lag the voltage by exactly 90°, whereas in a purely capacitive system the current will lead the voltage by exactly 90°. A purely resistive system will have the current and voltage in phase with each other. Most power systems are inductive due to the large reactance of the transmission lines. Figure 1 shows the phasor diagram and voltage and current waveforms of an inductive circuit in which the current lags the voltage by 30°.

For a circuit that has a combination of resistances and inductances, the product of the voltage (V) and the component of the current I that is in phase with the voltage ($I\cos\phi$) is the actual power that is used by the load. This quantity is known as the active power, P. The out-of-phase component ($VI\sin\phi$) is not used by the load directly but rather is responsible for the energization and deenergization of the reactive components in the circuit. This component is called

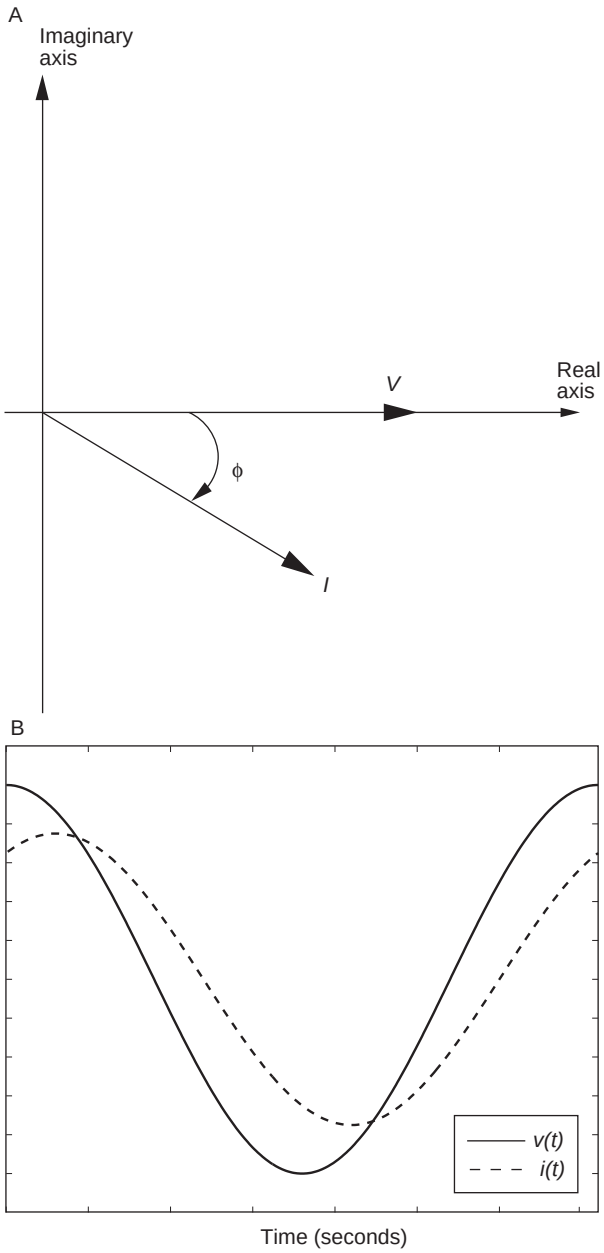


FIGURE 1 Inductive circuit (A) phasor diagram and (B) waveforms.

the reactive power, Q. Given the unique relationship between P and Q, the power in the AC circuit can be represented by $S = P + jQ$, where this complex sum S is called the apparent power with units of volt-amperes (VA). These relationships exist only in AC circuits. In a DC circuit there is no phase difference between the voltage and current; therefore, there is no reactive power and all the power generated is useful active power P. An inductive circuit will absorb both active and reactive power, whereas a

capacitive circuit will absorb active power but will inject reactive power.

2.1 Three-Phase Power

With the invention of the power transformer, AC systems became more popular than DC systems since the electric power could be transmitted across the transmission lines at higher voltage levels. Since power is the product of voltage and current, higher voltage levels require less current flow to transmit power. The reduced current subsequently reduces the active power losses in the line since the losses are given as I^2R, where R is the transmission line resistance.

Most electric power is transmitted in three phases, a–c. The three sets of sinusoidal voltages (and currents) are equal in magnitude but differ in phase angle of $\pm 120^\circ$ from each other. The voltages shown in Fig. 2 have three phases. By convention, the a phase is usually taken as the reference voltage. From these waveforms, it can be seen that the combination of three waveforms at any instant in time is identically zero. The same is true for the a–c phase currents. However, the product of these voltage and current waveforms yields the instantaneous electric power. If the three phases of electric power are combined, the resulting power is always constant at every instant under steady-state conditions even though the voltages and current are varying sinusoidally. This property of constant power is the primary reason that three-phase power is used. The same constant power result cannot be achieved by two-phase or four-phase power; thus, three-phase power is unique.

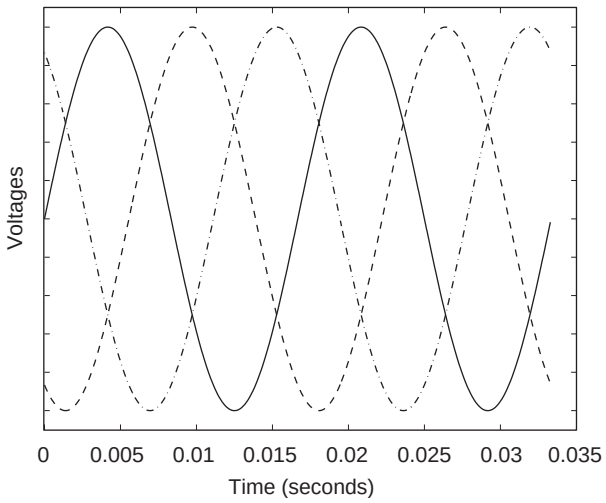


FIGURE 2 Power system voltages. Solid line, a phase; dashed line, b phase; and dashed–dotted line, c phase.

The second reason for using three-phase power is the ease of generation. A single-phase generator always experiences pulsating torque because the generated magnetic field does not rotate. The magnetic field simply changes direction at synchronous speed (i.e., it changes direction every 1/60 of a second). Conversely, the magnetic field generated by three-phase windings smoothly rotates at synchronous speed and does not produce pulsating torque. Similarly, motor loads require three-phase power for smooth and efficient operation. The three-phase system is also capable of eliminating the third harmonic currents that are often present in the distribution system by the use of delta-wye transformers, which are unique to the three-phase system. This idea can be extended to a six-phase system; however, such a system would not be economical because of the cost of additional cables and insulation. Hence, the three-phase system is considered to be optimal for electric power transmission.

2.2 Power System Voltage Levels

The three-phase power is typically generated at a particular voltage and is then "stepped up" to a higher voltage using a three-phase transformer. The higher voltage is used for transmission (to reduce currents and thereby losses) and is then "stepped down" to lower values at the distribution substation. The voltage ratings of large generators employed at generating stations range from 13.8 to 24 kV. This is stepped up to transmission-level voltages of 69–765 kV using transformers. The evolution of voltages from 138 to 765 kV in the United States was a result of the developing transmission technology. These voltage increases met the needs of the larger and more economical generators begin developed. This evolution is shown in Fig. 3. Electricity is delivered from the transmission system to the customer's home through a primary distribution system. This system is composed of a network of power lines delivering electricity at lower voltage. The voltage is stepped down again to the distribution system by a transformer. The voltages on a distribution level range from 220 V to 5 kV depending on the load connected. A typical power system configuration is shown in Fig. 4.

3. POWER TRANSFORMERS

A power transformer is used in a substation to step up or step down the incoming and outgoing voltages.

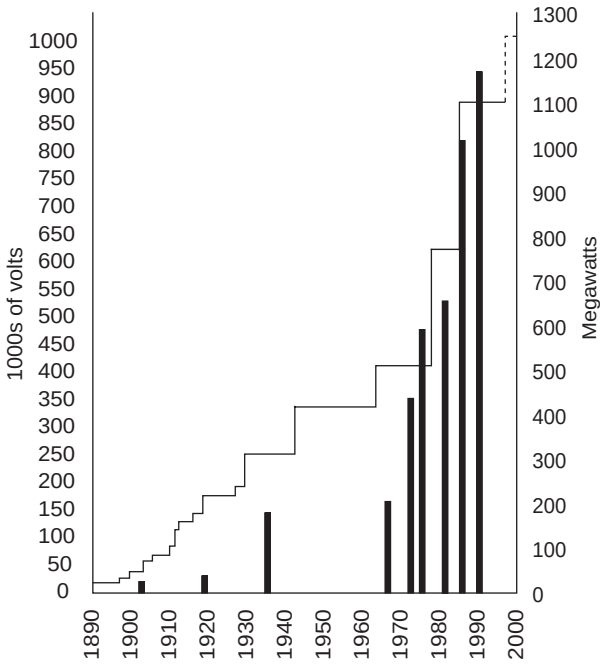


FIGURE 3 Evolution of maximum generating units and maximum transmission voltages.

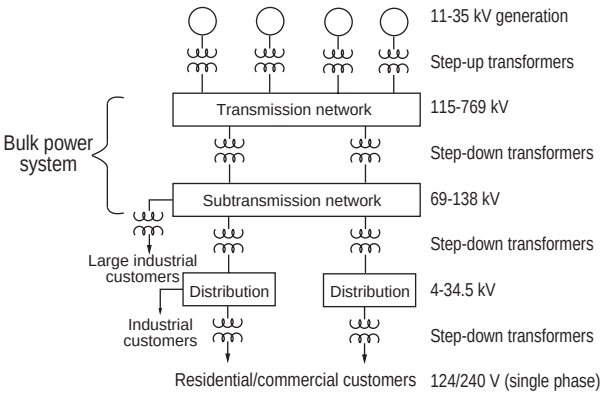


FIGURE 4 Power system configuration.

The transformers at distribution substations always step down the incoming transmission voltages to lower voltages used at the distribution level. Distribution lines emanate from substations at voltages of 4–25 kV. The power is then transmitted to a distribution transformer, where the voltage is further stepped down to the utilization voltage that ranges from 110 to 480 V for direct consumer usage.

When the transformer was invented by William Stanley in 1886, the early restrictions on the range and power levels of power systems were eliminated. In the DC system designed by Thomas Edison, the transmitted power required exceptionally high current levels because the voltage was low. These large current levels caused large voltage drops and power losses in the transmission lines. The use of transformers greatly reduced voltage drops and power losses. The basic premise of an ideal transformer is that if the system voltage is increased, the current levels decrease proportionally to maintain a constant power output.

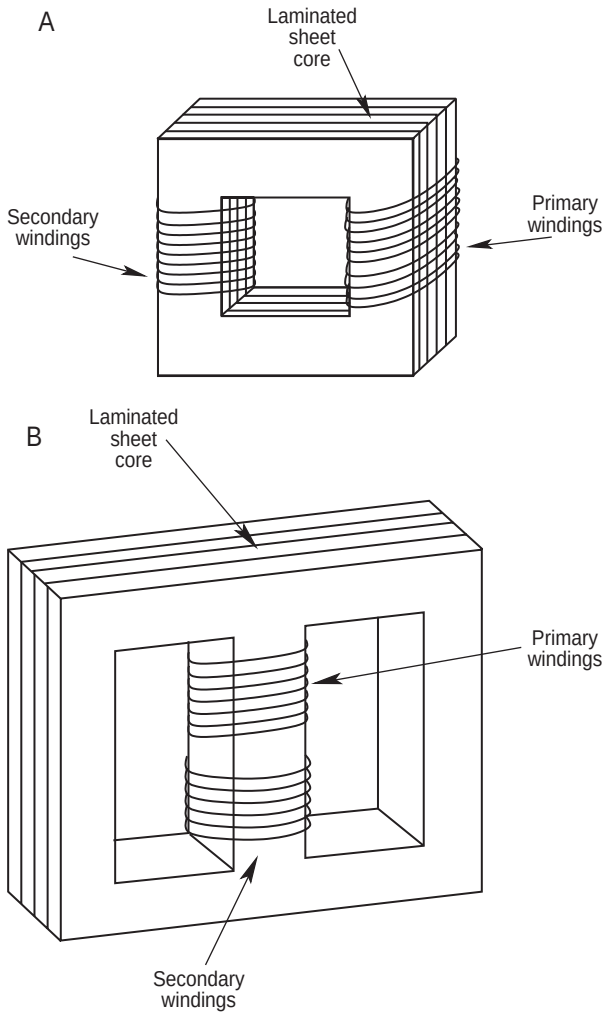


FIGURE 5 Single phase transformers: (A) core type and (B) shell type.

A transformer is typically constructed in one of two ways. The following basic configurations are shown in Fig. 5:

- Core type: This consists of a simple rectangular laminated piece of steel with the transformer winding wrapped on both sides. The winding connected to the source is called the primary winding, and that which is connected to the impedance or load is known as secondary winding.

• Shell type: This type of construction consists of a three-legged core with the windings wrapped around the center leg.

In both types of transformers, the core is constructed of thin laminations electrically isolated from each other in order to minimize eddy currents or those currents that circulate within the core material and cause heating losses.

The transformer works on the principle of magnetic induction. The time-varying voltage, V_p, that is applied to the input (primary) side produces a time-varying flux in the core. This changing flux couples with the secondary winding and induces a secondary voltage, V_s, in proportion to the ratio of the windings. The relationship between the voltages at both the sides of an ideal transformer is given by

$$V_p/V_s = N_p/N_s,$$

where N_p and N_s are the number of turns of the primary and secondary windings, respectively. The relationship between the current flowing in both sides of the ideal transformer is given by

$$N_p I_p = N_s I_s,$$

where I_p and I_s are the primary and secondary currents, respectively. If the transformer is lossless, then the input power is equal to the output power and the voltages and currents are related by $V_p/V_s = I_s/I_p$. Thus, as the voltage increases from secondary to primary, the current decreases; therefore, the current is lower at higher voltages.

In modern power systems, transformers can be found everywhere. Transformers have a variety of names, depending on the purpose for which they are used in a power system. A transformer connected to the output of the generator and used to step up its voltage to transmission levels is called a unit transformer. The transformers that bring down the transmission voltage levels to distribution levels are known as substation transformers. Lastly, the transformers that bring down the distribution-level voltage to the voltage level made available to the customers are known as distribution transformers.

Transformers are typically placed in a variety of locations. Many transformers are overhead transformers hung on poles with direct access to the wires. Underground transformers are typically placed in underground vaults to protect them from public access and environmental disturbances. The most common type of transformer is the pad-mounted transformer, which typically sits on a concrete pad at ground level. Pad mounts are designed for contact with the general public and therefore must adhere to strict American National Standards Institute (ANSI) standards. The last type of transformer is the interior transformer, which is designed to be placed in a building, usually at ground level. Due to concerns about the oil used for coolant, most interior transformers are cooled with a less flammable coolant such as gas and are frequently referred to as dry-type transformers.

A unique set of transformers are the under-load-tap-changing (ULTC) transformers. Since the voltage along a distribution line typically decreases with an increase in load being drawn by customers, it is desirable to regulate the voltage magnitude at the substation. The most economical method of voltage regulation is with an ULTC transformer. The ULTC transformer is built with several tap settings that allow the number of secondary windings to be adjusted based on the desired voltage level. As the ratio of primary to secondary windings varies, the voltage ratio on the primary and secondary side also changes. This changing of the tap setting enables the secondary voltage to be changed in response to changes in load. These taps are fully automatic and operate in response to voltage on the secondary side of the transformer.

4. TRANSMISSION SYSTEM COMPONENTS

An overhead transmission line consists of conductors, insulators, support structures, and, in most cases, shielding wires.

4.1 Power Cables

Copper has been replaced by aluminum for construction of overhead lines due to the availability and cost factor. The most commonly used conductor is aluminum conductor-steel reinforced (ACSR), which has a central core of galvanized steel wire. Surrounding this central core are strands of aluminum, which have good conductivity. The steel strands provide strength and the aluminum conducts the current. This type of conductor is shown in Fig. 6. The other conductors in common use are all-aluminum conductor, all-aluminum alloy, aluminum-clad steel conductor, aluminum-clad steel conductor (Alumoweld), and extra high-strength steel conductor. Fillers such as fiber or paper can be added between the aluminum and steel strands to increase the conductor spacing to reduce the electric field at the conductor surface.

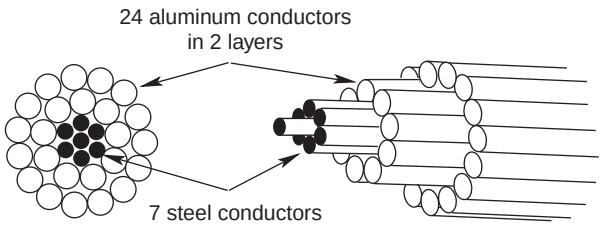


FIGURE 6 ACSR 24/7 conductor.

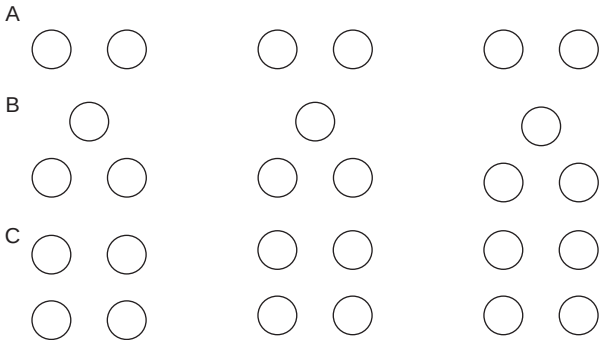


FIGURE 7 Typical bundle configurations.

Extra high-voltage lines often have more than one conductor per phase. This configuration is commonly referred as conductor bundling. Line bundling allows transmission of large amounts of power over long distances. A bundle consists of two–four conductors. Bundling increases the effective radius of the conductor and reduces the electric field strength near the conductors, which reduces corona power loss, audible noise, and radio interference. Line bundling also reduces line reactance, allowing greater current carrying capacity. Figure 7 shows typical bundle configurations.

4.2 Shielding Wires

Shielding wires are located over the load-carrying conductors on transmission lines to prevent a direct lightning strike. Almost all lines above 34.5 kV employ shielding wires. The shield wire provides a 30° protection on either side of the wires. The number and location of the shielding wires are selected such that a lightning strike will terminate on the shielding wires and not on the phase wires. Shielding wires are essentially grounded neutral wires located above the phase wires. They bypass a lightning strike harmlessly to the ground, provided that the tower impedance and tower footing resistance are small.

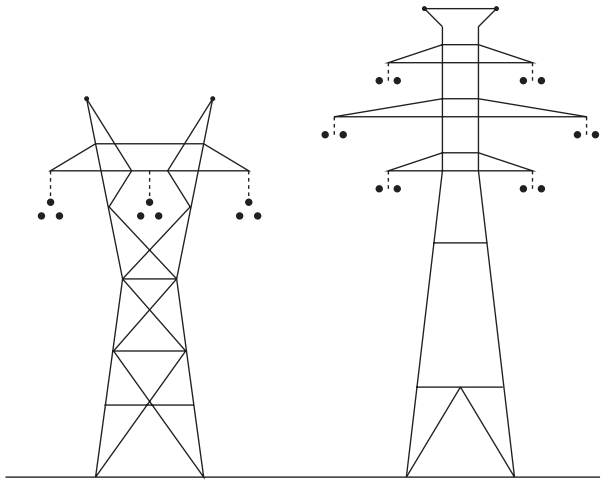


FIGURE 8 Typical lattice structures for 345-kV transmission lines.

4.3 Transmission Line Structures

The primary function of a transmission line structure is to support the transmission lines. The design of the transmission line structure must consider environmental conditions, area requirements, voltage levels, and right-of-way (ROW) requirements. The ROW is the amount of clearance required on either side of the transmission line for safety purposes. Double-circuit 345-kV lines usually have self-supporting steel towers with the phases arranged in a triangular configuration to reduce tower height or in a vertical configuration to reduce tower width. For transmission of voltages of 345 kV and less, wood-frame configurations are commonly used. Typical lattice-type structures for a 345-kV transmission line are shown in Fig. 8. Single-pole, tubular-steel structures are also popular for 345-kV transmission lines. These structures are preferred over the less expensive lattice-type structures when the ROW is restricted. Tubular-steel structures are typically more expensive than lattice-type structures, but lower maintenance requirements may make them competitive in situations in which access is difficult.

4.4 Insulators

Overhead line conductors must be supported on poles or towers such that currents from the conductors do not flow to the ground through the supports. This requires proper insulation of the lines from the supports, which is achieved by securing the line conductors to the supports by insulators. Insulators for transmission lines higher than 69 kV are suspension-type insulators that consist of a string

of discs, typically constructed of porcelain. All the discs are connected in series by metal links in the form of a string. The conductor is suspended to the bottom end of this string. The other end of the string is secured to the cross-arm of the tower. Each disc can only withstand a voltage of approximately 16 kV; therefore, the number of discs per insulator is proportional to the voltage rating of the line. A 120-kV insulator has 6–8 units, whereas a 500-kV insulator is constructed of 26–32 units.

Another type of insulator in operation is the post-type insulator, which is housed in substations to support conductors, busbars, and other equipment. This type of insulator is used for medium- and high-voltage transmission lines in which the insulator replaces the cross-arms of the transmission tower.

Composite insulators are becoming more common as an alternative to porcelain. Composite insulators are constructed around a fiberglass rod, which gives it high mechanical strength. A forged-steel or iron fitting is then compressed to the rod to provide a slippage-free end fitting.

5. SUBSTATIONS

Electrical power is produced at power-generating stations, which are located at favorable but typically geographically remote locations. The power is delivered to the customers at homes or offices through a large network of interconnecting transmission and distribution lines. At many places in the power system, it is necessary to alter some of the electric power characteristics, such as voltage or current. For example, the transmission-level voltages are reduced to distribution-level voltages for the purpose of supplying it to the customers. The place where this conversion of transmission to distribution takes place is known as a substation. Similarly, the generating voltages are also stepped up to transmission levels at substations near the generating unit. A typical distribution substation contains equipment discussed in the following sections.

5.1 Busbar

The busbar, or bus, is used to distribute the power in different directions. Many feeders come out of the substation because they have to cater to different customers in various geographical locations. The bus is a node in the power system circuit. In a power system, when two or more lines are attached to a common point, that point is called a bus. Physically, busbars are copper rods or thin-walled tubes at a common voltage in a substation to which two or more lines (feeders) are connected. If a number of generators or feeders at the same voltage are directly connected at a point, busbars are used as the common electrical component.

5.2 Isolating Switches and Circuit Breakers

In substations, it is often desired to disconnect a part of the system for general maintenance and repairs. This is accomplished with the help of isolators or isolating switches. A circuit breaker is a device that can open or close a circuit under normal as well as fault conditions. It is designed so that it can be operated manually under normal conditions and automatically under fault conditions.

5.3 Converter Stations

Converter stations are present only in high-voltage DC substations. The components are power electronic-based inverters or rectifiers, which are needed to convert the incoming AC components to DC and vice versa. If a DC link exists between two substations, then the voltage has to be converted at the first substation and back to DC by inverters at the second substation.

6. DISTRIBUTION SYSTEMS

The distribution system is the primary means of distributing the electric power from the substation to the system customers. A typical distribution substation serves 1–10 feeder circuits. Each feeder circuit may serve a variety of load types.

6.1 Power System Loads

A load is any device that draws electrical energy from a power system. A load can be resistive, inductive, capacitive, or a combination of these characteristics. Loads in a typical power system are classified as residential, commercial, and industrial. Of these, residential loads are mostly single-phase loads, whereas the other two types are generally three-phase loads. Residential loads have lower voltage levels and include lighting, air-conditioning, heating, and other household appliances. They are usually rated below 2 kW. On the other hand, industrial

loads have the highest voltage levels. Industrial loads include high-power motors used for drilling, pumping, grinding, or other manufacturing purposes and furnaces. Industrial loads are generally rated from 25 kW for small-scale industry to 500 kW or higher for large-scale industry. Lighting, heating, or cooling of large commercial areas such as shopping malls are commercial loads. Figure 9 shows the utility sales by sector for the United States from 1990 to 2001.

The magnitude of a load may vary continuously throughout the day, week, season, and year. Therefore, loads may also be differentiated in terms of time as opposed to end-use sector:

- Base load: The minimum load that has to be served throughout the day. This load is typically served by the largest and/or most inexpensive generators.
- Peak load: The maximum load that has to be served by the utility during the time period of interest. For example, industrial loads usually peak once during each day, but the summer peak load occurs just one day per year.
- Off-peak load: Any load other than peak or base loads is referred to as an off-peak load.

7. POWER SYSTEM PROTECTION

A fail-free power system is not technically feasible since faults occur in a power system at regular intervals for various reasons. Short circuits occur in power systems when equipment insulation fails due to system overvoltages caused by lightning or switching surges. Faults can occur in any power system component, including generators, transformers, buses, or lines. The most prone to faults are the transmission lines since they are the component most exposed to the environment. Typical transmission line faults that can occur are insulation degradation and short circuits caused by contact with tree branches, lightning, and grounding failures. These faults can cause sustained or momentary interruption of power to the customer and may cause considerable economic losses. Approximately 80% of the faults that occur in the power system are temporary in nature. Proper protection schemes can ensure that reliable power can still be supplied to the customer.

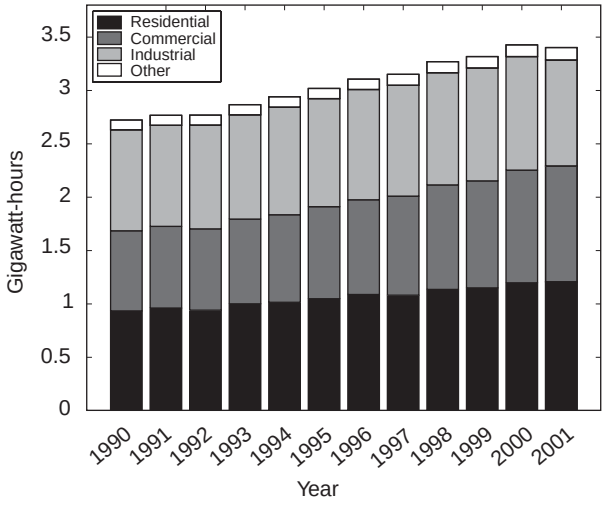


FIGURE 9 U.S. electric utility retail sales of electricity by sector, 1990–2001.

ANSI standards require that faults should be cleared within 3 cycles in extra high-voltage lines and in 5–20 cycles in lower voltage equipment. Faults that occur on a power system are cleared by placing protection equipment throughout the line.

7.1 Circuit Breakers

Circuit breakers are a self-controlled device for automatically interrupting and reclosing an AC circuit with a preset sequence of opening and reclosures. Circuit breakers do not have built-in controls, and they open or reclose based on relay operation. Relays sense the current or voltage and energize the trip coil of the circuit breaker, which in turn causes the poles to separate. As the poles separate, an electric arc is induced across the intervening gap. Different quenching mediums are used to interrupt the arc, thereby opening the circuit:

- Air-break circuit breaker: This type of circuit breaker is generally the indoor type and is installed on vertical panels. They are used for DC circuits up to 12 kV and for medium- and low-voltage AC circuits, usually up to 6.6 kV. In this case, the arc is lengthened in the breaker by a magnetic field and is extinguished by arc splitters.
- Oil circuit breaker: Oil is employed as the quenching medium in this family of breakers. Early bulk-oil breakers used a large oil tank for quenching the arc. However, modern practice is to use minimum-oil circuit breakers in which the arc is extinguished in a small arc control device filled with oil. The voltage rating for this type of breaker is 3.6–245 kV. The disadvantage of this type of breaker is that since oil is hygroscopic, it must be sealed in an air-tight chamber.
- Vacuum circuit breaker: The principle for this type of circuit breaker is that a vacuum is a dielectric

medium and an arc cannot persist in an ideal vacuum. The arc is generally extinguished in half a cycle. Vacuum circuit breakers are generally employed up to 72.5 kV and are smaller than oil circuit breakers.

- Air-blast circuit breaker: To extinguish the arc in an air-blast circuit breaker, a blast of air is blown at the current zero-crossing. These circuit breakers can be employed up to a voltage of 275 kV.
- SF-6 circuit breaker: SF-6 is a heavy, chemically inert, nontoxic, nonflammable gas. Breakers using SF-6 are popular and are used for voltages of 6.6–765 kV.

7.2 Relays

Relays are devices whose primary function is to detect faults and take proper control action to clear the faults. Whenever the relay senses a faulty condition in the power system, it energizes the trip coil of the circuit breaker to clear the fault. Relays and circuit breakers comprise the protection system of the power system. The following are the different types of relays:

- Overcurrent relays: This type of relay operates when the current through it exceeds the pickup value of the relay. For currents lower than this pickup value, the circuit breaker is not tripped. For system-wide coordination of overcurrent relays, it is essential that the relays conform to an inverse-time characteristic in which the operating time (time to trip for a given current) decreases as the relay current increases. Such relays are known as inverse definite minimum time lag relays.
- Directional relays: Directional relays are designed to operate for fault currents flowing in only one direction. When two or more sources are present at different locations, it is frequently difficult to coordinate overcurrent relays. In such cases, directional relays can be used.
- Impedance relays: Impedance relays operate on the basis of the voltage-to-current, or impedance, ratio. The impedance relay is also called a distance relay since the line impedance is proportional to distance. The relay trips when the measured impedance $|Z|$ falls below $|Z_r|$, where $|Z_r|$ is an adjustable relay setting and $|Z|$ is the voltage-to-current ratio at the relay location.
- Differential relays: A differential relay is a suitably connected overcurrent relay operating on the difference between the currents at the two ends of the protected element. It does not operate for faults outside of the equipment. Only for an internal fault will a difference of currents occur, causing the trip coil to be energized. Differential relays are commonly used to protect generators and transformers.

Relays were introduced in the early 1900s and were electromechanical in nature. They are relatively maintenance free and are still found in operation today. Solid-state relays using analog circuits and logic gates provided the same reliability as electromechanical relays but at a more competitive price. The latest technological devices are digital relays that allow the relay characteristics to be changed from a remote location in real time. The principles of digital relays are the same as those of previous electromechanical relay, but these relays have the advantage that a central computer can coordinate between the different local digital relays.

7.3 Fuses

The simplest overcurrent protective element in a power system network is a fuse. Fuses are generally used in large numbers by utility companies because they are inexpensive and maintenance free. The fuse consists of a metal "fusible" link or links encapsulated in a tube, packed in a filler material such as sand. The metal link is generally silver, and it is connected to the contact terminals. If the current flowing through the fuse is normal, the fuse acts as a conductor. However, when abnormal current levels are reached, the link melts. Most of the fuses are current limiting in nature (i.e., the fuse will melt in less than half a cycle for high fault current). In the event of a fault, only the fuse closest to the fault should interrupt and open the circuit immediately downstream. The fuse with the highest continuous current rating is placed closest to the source in the case of a radial feeder. The primary disadvantage of fuses is that they operate only once and cannot operate again without manually replacing the fuse link.

7.4 Circuit Recloser

An automatic circuit recloser is an overcurrent protective device that automatically trips and recloses a preset number of times to clear temporary faults or to isolate permanent faults. It also has provisions for manual opening and reclosing. A recloser unit has a built-in relay and a circuit breaker. Most temporary faults are cleared in one or two recloser operations. Since 80% of the faults that

occur on the power system are temporary in nature, the presence of a recloser helps eliminate the unnecessary tripping of loads for temporary faults. The recloser and the fuses must be coordinated so that only the faulted section is affected.

7.5 Sectionalizer

A sectionalizer is a self-opening device that is used in conjunction with the source-side recloser or circuit breaker to automatically isolate faults. When the source-side protective device operates to deenergize the circuit, the sectionalizer will count the number of operations. After a set number of operations, the sectionalizer permanently opens, thus limiting the outage to the faulted section only. Sectionalizers isolate permanent faults to smaller sections of the line, which in turn improves the reliability of the system. Another advantage is that the sectionalizer does not have a time–current curve and therefore does not have to be coordinated. It can be placed between existing reclosers and fuses without changing its setting.

7.6 Lightning Arrestors

A lightning arrestor, or surge arrestor, is a protective device that conducts the high-voltage surges on the power system to the ground. One end of the arrestor is attached to the terminal of the equipment and the other end is effectively grounded. It consists of a spark gap in series with a nonlinear resistor. The length of the gap is set such that normal line voltage is not enough to cause an arc across the gap, but a dangerously high voltage will break down the air insulation and form an arc. The property of the nonlinear resistance is such that its resistance decreases as the voltage or current increases and vice versa.

SEE ALSO THE FOLLOWING ARTICLES

Electrical Energy and Power • Electric Power: Critical Infrastructure Protection • Electric Power Generation: Fossil Fuel • Electric Power Generation: Valuation of Environmental Costs • Electric Power Reform: Social and Environmental Issues • Electric Power Systems Engineering • Electric Power: Traditional Monopoly Franchise Regulation and Rate Making • Electric Power: Transmission and Generation Reliability and Adequacy

Further Reading

Blackburn, J. (1997). "Protective Relaying: Principles and Applications." Dekker, New York.
Grainger, J., and Stevenson, W. (1994). "Power System Analysis." McGraw-Hill, New York.
Grigsby, L. (2000). "The Electric Power Engineering Handbook." CRC Press, Boca Raton, FL.
Kersting, W. (2001). "Distribution System Modeling and Analysis." CRC Press, Boca Raton, FL.
Winders, J. (2002). "Power Transformers: Principles and Applications." Dekker, New York.

Electric Power Reform: Social and Environmental Issues

NAVROZ K. DUBASH
World Resources Institute
Washington, D.C., United States

1. Introduction
2. Reasons for Reform of the Electric Power Sector
3. Social Concerns in Electric Power Reform
4. Environmental Concerns in Electric Power Reform
5. Governance of Electric Power Reform

Glossary

cross-subsidy A subsidy to one set of consumers of a service to reduce the cost of providing access or the cost of consumption, paid for by charging other consumers more.

electricity service concession A contract to a private sector firm to provide electricity in a defined area under a set of rules and incentives established by the government.

end-use energy efficiency programs Efforts at increasing the efficiency of the energy system by focusing on changing use patterns of and technologies deployed by the final consumer.

environmental dispatch The dispatch of electricity from electricity generators based on a pollution index that allows adjustment of both the order of dispatch and the price to internalize pollution costs.

integrated resource planning A planning framework that allows for comparison of the costs and benefits of alternative electricity supply mechanisms—including demand-side savings and nontraditional electricity sources such as renewable sources—across an electricity system. This framework should include consideration of environmental and social externalities.

lifeline rates Reduced rates, often financed by cross-subsidies, which allow poor households that use minimal amounts of a service to pay less than wealthier households that use more of those services.

universal service obligation A commitment by service providers to provide electricity services to those who request it within their service area, often accompanied by a commitment to provide this service at an affordable price.

The issue of equity and environment in electric power reform is rooted in a larger debate over the nature of the electricity sector. Is electricity a commodity like any other? Should the production, transmission, and distribution of electricity be managed largely by the private sector? Or is electricity an instrument of policy aimed at sustainable development, over which governments should retain policy control, if not ownership? In furthering the debate, a position that has found resonance is that although provision of electricity services may indeed be benefited by some commercial role, electricity provision also has strong positive and negative externalities that call for proactive government involvement.

1. INTRODUCTION

Electric power reform commonly refers to changes in ownership and structure of an electricity system, with the purpose of introducing private ownership and competition. Discussions over how to implement these reforms have focused on improving the technological and financial performance of the sector. Scholars and policy makers have also recognized that power sector reform can have implications for social and environmental outcomes, and that these concerns are not captured by a focus on technical and economic concerns alone. Prominent among relevant social outcomes are the extent to which members of a society have access to electricity, the affordability of electricity, the impact on labor employed in the sector, and the reliability of electricity supply. Relevant among environmental outcomes are the impacts of sector reforms on global, regional, and local pollution caused by production of electricity. Important factors shaping environmental outcomes include the impact of

 © 2004 Elsevier Inc. All rights reserved.

reforms on consumption levels, technology change, efficiency of supply and demand, and regulatory practice. Finally, an important factor in determining equity and environmental outcomes of electric power reform is the quality of governance in the sector.

2. REASONS FOR REFORM OF THE ELECTRIC POWER SECTOR

2.1 Electric Power Reform Defined

As described in a widely used formulation by authors Sally Hunt and Graham Shuttleworth, electric power reform captures changes in the electricity sector along two dimensions: ownership/control and market structure. Changes in control and ownership can be further broken into three categories:

1. Commercialization. A government surrenders control over a state-owned enterprise in order to promote operation along commercial lines. However, ownership remains in government hands, with no change in organizational form.
2. Corporatization. A government formally and legally relinquishes control and management over a public enterprise to create a corporation. Management is independent of the government, although ownership remains in government hands, and the corporation may be subject to regulatory oversight.
3. Privatization. A government sells a corporation to private owners. The new owners' primary accountability is to investors, although the corporation may be subject to regulatory oversight.

Changes in market structure to promote competition can be subdivided into four steps:

1. Monopoly. No competition at any level. A single entity, usually government owned and operated, handles generation and transmission of electricity. This company provides electricity to distribution companies, which have a monopoly relationship with the final consumer.
3. Single buyer. Competition in generation. Public or private power generators sell power to a single purchasing agency. The purchasing agency transmits power to distribution companies, which have a monopoly relationship with the final consumer.
3. Wholesale competition. Competing generating companies sell directly to distribution companies, and also to very large consumers. Generators have open access to the transmission grid in order to deliver power to their consumers. With the exception of large consumers, distribution companies have a monopoly over final consumers.
4. Retail competition. Generating companies sell directly to distributing companies and final consumers. Generators have open access to both transmission and distribution grids. Final consumers can choose their supplier.

This simple framework conceals considerable discussion and debate over the details of creating markets, and the appropriate sequencing of these various stages.

2.2 Drivers of Electric Power Reform

Historically, electric power systems have been publicly owned and operated because of their natural monopoly characteristics. Governments were better able to exploit the economies of scale in power generation, raise the large amounts of capital required, bear the long time periods for returns on investment, and address the networked nature of the system. Over time, electric power became part of a set of essential services, provision of which was associated with governments. In much of the industrialized world, this approach succeeded in providing reliable and affordable power to citizens, although in the developing world success was far more modest. However, it was in industrialized countries that this approach to power provision was first questioned. At least three factors spurred the questioning, albeit with different emphases in different countries. First, demand for electricity began to slow, undermining the predictable sources of growing income on which utility financing had been based. Second, new generation technologies, particularly gas turbines, reversed the scale economies in the sector. Third, costs and risks increased in the sector due to a rising environmental consciousness, a corresponding increase in regulations, and burdensome investments in high-capital-cost nuclear power plants.

Countries began to act on these changed realities. In the 1970s, the United States allowed independent power producers to sell electricity to utilities, undermining the idea that electricity was best run as a natural monopoly. In the 1980s, the United Kingdom and Chile remade their electricity sectors around the objective of promoting competition. In particular, they unbundled the component parts of the electricity system—generation, transmission, and distribution—and offered components for sale to the private sector. Around the same time, Norway also introduced competition, but without an

accompanying process of privatization. The rapid growth and declining costs of information technology facilitated these changes by enabling the development of new control techniques consistent with greater decentralization of the grid.

Advocates of market-based reforms suggest that reforms bring greater incentives for cost-saving and productivity enhancement. Specifically, it is argued that market reforms provide appropriate price signals to consumers (which spurs more efficient economy-wide resources allocation), reward the efficient use of inputs with profit, and force cost cutting through competitive pressures. Critics point to several difficulties with creating a well-functioning market in electricity. These include the difficulties of defending against market power, the higher cost of capital faced by the private sector, which can translate to higher average costs in a capital-intensive sector such as electricity, and the difficulty of planning, particularly for transmission, in a decentralized system when information is hidden and incomplete. In addition, critics challenge the emphasis on reforms aimed at facilitating private participation, and question whether a change in ownership is a critical element in solving the sector's problems.

It is beyond the scope here to settle these debates conclusively, but the empirical record provides a few lessons of relevance to this discussion. Designing and implementing a market for electricity is a complicated affair, and it is certainly possible to make an electricity system worse off through reforms, as the disastrous California experience showed only too well. Where electricity reforms have resulted in some efficiency gains, competition is at least as important as privatization in achieving these gains. Where efficiency gains have led to lower electricity prices, such as in the United Kingdom and Europe, benefits have gone disproportionately to industrial users. Finally, controlling price volatility has been a considerable challenge, in some cases even threatening the reliability of the system. Despite these continued uncertainties, electricity reform has proceeded apace worldwide, including in developing countries. Reforms in developing countries have been driven by financial and economic concerns such as the poor performance of public utilities, a shortage of capital for investment, a broader fiscal crisis due in part to electricity subsidies, and a short-term need to raise capital through the sale of electricity assets. In many cases, countries have embarked on a reform trajectory at the urging of international financial institutions, and as part of a larger package of economic reforms.

The application of a uniform reform approach as just described to countries in the developing world has led to criticism. One prominent critic, A. K. N. Reddy, points out that the approach followed in the developing world was developed for quite a different purpose—to exploit opportunities for efficiency gains in mature electricity systems in industrialized countries. Existing problems—technological and financial performance crises, the failure to provide access to electricity to large portions of the population, and the lack of attention to the environmental trajectory of the sector—find little explicit attention in the reform approach that has been carried out to date. This observation is an important reason for seeking a broader and more context-specific set of design parameters for electric power reform in the developing world.

3. SOCIAL CONCERNS IN ELECTRIC POWER REFORM

Electricity services enable many positive economic and social benefits. Electricity is a key infrastructural element for industrialized growth and for irrigation. At a household level, electricity can have positive impacts on education, due to better lighting; on time and effort savings, by substituting for traditional fuels used for cooking and heating; on health, by reducing indoor air pollution and through use of refrigeration for foods; and on entrepreneurial activity, through better information and digital connectivity. To achieve these gains, citizens must have access to electricity, and this access must be affordable and reliable. In addition, the electricity sector is also a large employer in many countries, which has provoked concerns among labor groups about the impacts of reforms. The discussion here focuses on access, affordability, and labor concerns. Although reliability is also an important concern, it is already an important and explicit goal of the reform agenda, and so is not treated further here.

3.1 Access to Electricity

Worldwide, rates of access to electricity remain low. Two billion people around the world are commonly estimated to lack electricity. The World Energy Assessment states that in rural areas the situation is particularly bad; 56% of the world's rural population has to manage without electricity. The situation is particularly dire in Sub-Saharan Africa, where the World Bank found that in 1990, 62% of the urban

population and 92% of the rural population rely entirely on energy sources other than electricity.

Historically, many public utilities have undertaken grid extension by setting targets, and financing investment through levies or cross-subsidies. For example, South Africa's public utility, Eskom, exceeded its target of 350,000 new connections a year through much of the 1990s, financed by a cross-subsidy from existing users. South Africa, however, is an exception, and the experience with target setting for public utilities has been disappointing in much of the world. There are several reasons for this failure to provide access to electricity. The rural poor are particularly costly to serve because of remote locations, low population density, high transmission line losses, low consumption levels (which raise average costs), poor credit and minimal collateral, and a lack of purchasing and political power. Many of these factors, other than geographically determined factors, also apply to poor urban and periurban populations. These problematic conditions have been compounded by institutional failure at public utilities. Uncollected levies, the failure to apply least-cost electrification approaches, and the use of electricity as an instrument of political patronage have further stymied efforts at electrification.

The proposed new model of private participation in a market-oriented sector may be no better placed to solve this problem than the public utility approach. In a market model driven by profitability, interest in serving the rural or urban poor is likely to be muted in the absence of specific incentives provided by public policy. Research conducted at the World Resources Institute suggests that through the 1990s, access to electricity was rarely a priority in developing country reform efforts. Early indications suggest that the outcome is indeed minimal progress toward increased access to electricity. For example, a review of reforms in Bolivia concluded that expansion of the grid will not take place as a consequence of privatization and restructuring. In addition, the private sector may choose to disconnect nonpaying customers that the public utility had continued to supply as an implicit form of social welfare. For example, in Argentina, reforms led to a measure of deelectrification of the poor as the privatized utility clamped down on nonpayment and informal connections.

As it has become clear that additional policy incentives are necessary if increasing access to electricity is to remain a priority, more attention has been paid to ways of integrating access into reform approaches. In a market-based electricity sector, two possible entry points for access concerns are (1) the conditions of privatization and (2) concession agreements with the private sector for provision of electricity.

3.1.1 Institutional Mechanisms for Promoting Enhanced Electricity Access

Governments may place obligations on private providers of electricity, as a condition of participation in the market. Such an obligation can take various forms. One approach is to specify an "obligatory service" commitment, which requires a company to provide access to its services to all who wish to avail themselves of services at the prevailing tariff. However, many consumers may not be able to afford consumption at the prevailing tariff. A "universal service obligation" addresses this issue by requiring a company to provide a specified level of service to all customers at an affordable tariff. Because this tariff is necessarily below the price the company would otherwise seek to charge, establishment of a universal service obligation also entails finding financing to support this obligation. This can be arranged through three approaches: a fund collected through a special charge on consumers, which is then paid to the service-providing company; by allowing cross-subsidies; or by defining the service area in such a way that profitable and unprofitable areas are bundled together.

For populations beyond the scope of the existing grid, however, more explicit and targeted programs of access are necessary. Electricity concessions provide one such approach that is potentially compatible with a postreform electricity sector. The central idea is that the private sector is awarded a contract to provide electricity in a defined area under a set of rules and incentives, including financial incentives, established by the government. The desired end is access to electricity, and not necessarily grid extension. Because generation closer to the point of consumption will obviate the need for investment in transmission lines, the concession approach provides an incentive to seek out the lowest cost technology for providing access. Consequently, distributed electricity using technologies such as small hydropower, wind, solar, and biomass may prove to be cheaper and more practical options for providing electricity, and may be adopted by concessionaires.

There are several variants on this approach, which vary in their method of subsidy allocation and degree of private participation. For example, in Argentina, the government has sought to maximize cost-effectiveness by awarding concessions to the company

that requires the lowest subsidy to provide service to a concession area. In Chile, private distribution companies, acting in partnership with municipalities or rural communities, submit project proposals to receive subsidies to cover investment, but not operating, costs. Only projects that have a negative private return but a positive social return are eligible. The state agency evaluates these projects on the basis of cost–benefit analysis, financial contribution by communities and the company, and the degree of social impact. In another example, electrification in Morocco is promoted by a combination of a public and private approach. In some cases, the national electric utility maintains full control over installation, maintenance, and collection of fees; in other cases, it contracts private retailers to install distributed power systems in return for a share of fees. In addition, the utility requires municipalities and households to cofinance 45% of the investment costs of electrification.

The discussion here so far has focused on institutional mechanisms for involving the private sector in electricity provision. However, the same mechanisms can work effectively for nonprofit entities, such as community groups or cooperatives. There are precedents for electrification organized around cooperatives, most notably in the United States. Starting in 1935, electrification in the United States was built around rural electricity cooperatives financed by low-interest public funding. These cooperatives applied a principle of universal coverage and drew on standardized engineering across the country to reduce costs. Cooperatives and smaller scale entities also benefit from greater knowledge of local circumstances, which can enable them to better adapt technologies and more effectively fulfill difficult tasks such as bill collection.

3.1.2 Financing Access

It is significant that the approaches described in the preceding sections do not lift from the state the burden of raising funds, but rather provide a new way of channeling those funds, and potentially leveraging additional funds. Historically, electrification has been financed through cross-subsidies—charging some consumers more to subsidize investment in access for others. In the drive to introduce market reforms, cross-subsidies are increasingly frowned on as introducing distorted price signals leading to suboptimal patterns of use. More transparent, but also distorting, are explicit levies on some consumers to be used for electrification. The use of levies requires sufficient state capacity to effectively administer such a fund. A third approach, and one most favored by market-oriented reform advocates, is an explicit allocation from the central budget for electricity access. However, aspirations to transparent and explicit subsidy allocation ignore the political reality in many countries where access is a concern. Where budget allocation is determined by a complex of national and subnational political interests, it is far from clear that forcing subsidy allocation to pass through a national governmental process will yield a satisfactory allocation of funds for promoting electricity access. In general, the choice of financing mechanism should be determined not only by considerations of economic efficiency, but also by considerations of government fiscal and administrative capacity and the likely impact of local politics.

3.1.3 Local-Level Competition for Electricity Provision

The approaches described so far are all built around monopoly provision of electricity at some local scale, even if the selection of the monopoly provider is on a competitive basis. An alternative approach would allow free entry to the market for electricity services. Advocates of competition at the local level suggest that the result will be increased efficiency of service provision and incentives for innovation, with the result that the need for subsidies may be reduced. To realize this vision, arrangement for exclusive provision will have to be opened to all providers, and regulations will have to be designed minimally in order to reduce entry costs. However, open entry and competition for provision of services bring potential problems. First, there is a strong link between market structure and financing of electrification. Where there is an existing service provider, competition could allow an entrant to "cherry pick" rich customers or rich areas, leaving only the poorer customers. This problem is particularly pernicious where services to poor customers are financed through cross-subsidies, or where the incumbent provider has a universal service obligation. In areas where there are no existing providers and where subsidies are necessary to lure entrants, the burden of administering subsidies increases considerably with multiple players. This is particularly true if provision of subsidies is dependent on performance criteria, such as number of customers served. Second, in the absence of adequate regulation, there is scope for formation of cartels and price gouging of customers. Finally, there are basic issues of safety in electricity provision that are left unaddressed in an unregulated or lightly regulated market for electricity services.

3.2 Price

In both industrialized and developing countries, electricity reform leads to price changes, although for different reasons and sometimes in opposite directions. In industrialized countries, reforms are expected to lead to efficiency gains and lower prices for consumers. In practice, where these gains have been realized, they have benefited shareholders and rich consumers more than they have benefited poor consumers. The results can be highly regressive. In Argentina, for example, prices went down postreform for all consumer classes, but the price for high-consumption (and likely wealthy) users dropped by 70% whereas that for low-consumption (and poor) users fell only marginally. The net effect was that, in absolute terms, the rich paid 40% less than the poor paid per unit of electricity. In the United Kingdom, electricity companies reaped the greatest share of benefits, government gained from increased sales revenue and higher taxes on profits, but power purchasers paid higher prices than they would have under public ownership.

In many developing countries, prices for consumers are likely to rise, at least in the short to medium term, as cross-subsidies and price restrictions are removed. In some instances, these price increases have been considerable. In Ghana, the tariff increased by about 300% in the course of implementing reforms. Higher prices disproportionately affect the poor, who tend to spend a larger proportion of their income on energy services than do wealthy consumers. Although the higher prices may be necessary to put the sector on a sound commercial footing, they also bring social costs. Consequently, to ensure that social policy goals of providing effective access to electricity are met may require a measure of public support in the form of subsidies.

The question of subsidizing the price of electricity is closely tied to the discussion on access to electricity (Section 3.1), because a given subsidy allocation can be used either for the cost of facilitating access to electricity or for reducing the per-unit cost for the consumer. Both benefit the consumer, but in different ways. There are good arguments for subsidizing access rather than consumption. Populations without access to electricity are likely to be among the poorest in a country. By focusing on access, subsidies are automatically targeted toward the neediest. In addition, subsidizing use can skew consumption patterns in detrimental ways. For example, electricity subsidies for farmers in India have encouraged heavy use of groundwater, including the introduction of water-intensive crops in areas of water shortage. One means of avoiding this problem is to subsidize the fixed costs associated with consumption, such as connection fees. The result can be substantial savings for low-income groups, even while eliminating the potential for price distortions.

Although subsidizing access is preferable to subsidizing consumption, subsidizing a minimal level of consumption can afford low-income groups access to basic electricity services, with considerable social gains. Such subsidies are normally structured as "lifeline rates," which are cross-subsidies that allow poor households that use basic services to pay less, compared to wealthier households that use more electricity services. Although challenged under a reform approach that seeks to remove market distortions, these lifeline rates continue to be an effective way to balance social objectives with efficiency gains.

Several factors bear consideration in the design of price subsidies. First, subsidies should be well targeted to a population. In many cases, subsidies have become mechanisms for appeasing politically important constituencies rather than an effective way of supporting a social objective. Second, the amount of subsidy should be consistent with the goal of the subsidy program. For example, a lifeline rate should be limited to basic electricity services, and should not be set at a consumption level that encompasses groups beyond the target group. Third, the subsidy should be designed so as to minimize distorting consumer decisions.

Another means of minimizing social pain due to price increases is to increase the efficiency of electricity use. Doing so would enable consumers to extract greater services out of each unit of electricity consumed. In this sense, programs to promote end-use energy efficiency, typically conceived of as environmental programs, also have social benefits.

3.3 Employment

In many countries, the electricity sector is a significant employer. Reform of the sector is often accompanied by job cuts, as privatized utilities seek to cut costs and improve efficiency. In the European Union, for example, the research unit of Public Services International, a labor union, found that between 1990 and 1995, 212,000 jobs were lost in the energy sector, equivalent to a 17% reduction. The United Kingdom, where 42% of jobs in the sector were cut, accounted for more than half of these losses.

In some countries, reform of the sector to restore efficient functioning and financial viability may indeed require some measure of employee reduction. For example, in many African countries, the number of customers per employee hovers in the range of 20–80, as against a nominal international target of around 125. However, these numbers must be read with caution, because they do not account for differences across companies that only generate power and those that also distribute power. Moreover, the relative costs of labor and capital vary across countries, which suggests that the appropriate ratio of customers to employees need not be the same across the world. Nonetheless, it is safe to assume that there are many cases in which effective reform requires a measure of labor retrenchment. In these cases, a socially progressive approach to reform would ensure safeguards for laid-off workers, such as retraining programs and assistance with reemployment. Even more wide-reaching are plans to include workers at large in the capitalization of electric utilities. In Bolivia, for example, labor unions, political parties, business groups, and others forged a "capitalization plan" whereby a portion of the shares of state-owned enterprises, including electricity, were distributed among the broader working-age population. Specifically, half the shares were sold to strategic investors and the other half were credited to workers' individual retirement accounts, with a portion distributed as annuities. The general point is that although employee reductions may be necessary in some cases, the associated social pain can be reduced by appropriate mitigation, and by undertaking privatization in a manner that distributes benefits across the population.

4. ENVIRONMENTAL CONCERNS IN ELECTRIC POWER REFORM

The electricity sector has a considerable environmental footprint. Many of the environmental impacts associated with electricity are related to the heavy reliance on fossil fuels for electricity generation in much of the world. As the World Energy Assessment summarizes, combustion of fossil fuels releases local pollutants such as particulates and lead, regional pollutants such as sulfur dioxide, and global pollutants such as carbon dioxide. These pollutants have consequences for human health and the well-being of ecosystems. In addition, there are upstream environmental impacts such as those caused by mining and fuel transport, storage, and processing, and downstream impacts such as ash disposal. Nonfossil sources of electricity also cause environmental disturbance. For example, hydropower (in particular, large hydropower) causes several land-use-related impacts, such as submergence, displacement of populations, greenhouse gas emissions caused by decaying biomass, and disruption of fisheries. The discussion in this section, however, focuses largely on air pollution impacts.

What will be the net effect of electric power reform on the environment? There is no single answer to this question. Instead, the net effect of the several changes that collectively constitute reform will depend at minimum on the time frame (short versus long term), the specifics of how policies are implemented, and the initial conditions in the reforming country. However, systematic tendencies can be associated with particular reform steps. And, in addition, countries may choose to take proactive measures to address some of these tendencies. The discussions that follow outline these environmental tendencies and spell out possible proactive steps to increase the likelihood of environmentally positive outcomes.

4.1 Environmental Implications of Institutional Change for Electricity Reform

There are several overlapping ways to analyze the environmental implications of institutional change in the electric power sector. One approach is to examine the various components of electricity reform (as described in Section 2.1), with an eye to assessing the environmental implications of each component. This is the approach followed in this discussion. Another would focus on pathways through which reforms would have an environmental impact, such as on choice of generation technologies. These are not exclusive approaches, and there is necessarily some overlap in the categories under each approach. To complement an analysis based on the reform component, Table I, which is intended to be illustrative rather than comprehensive, summarizes a pathway approach.

The remainder of this section follows a framework laid out by Keith Kozloff in studies for the U.S. Agency for International Development and for the Renewable Energy Policy Project. The discussion is organized around five dimensions of reform: commercialization, privatization, unbundling, wholesale competition, and retail competition. The first two dimensions are related to ownership, and

TABLE I

Pathways through Which Electricity Reform Affects Environmental Outcomes

Pathway	Relevant questions	Intermediate variables
Consumption: price response and output substitution	How will electricity price change?	Price elasticity of demand
	Will output substitution occur toward or away from electricity?	Existing generation mix
	Will a market for electricity services be created?	Existing nonelectricity fuel choices
Generation technology: scope for input substitution	Which technologies will be favored under reform, and which will be disadvantaged?	Fixed versus operating cost of technology
		Transmission investment or cross-border trade as a substitute for generation investment
		Extent to which generation facilities using older technologies have been amortized
Efficiency of supply and demand	Will incentives for more efficient generation, transmission, and distribution be introduced?	Commercial pressure to cut costs
	Will end-use energy efficiency become easier or harder to promote?	Transactions costs for end-use efficiency
Regulation: environmental and economic	To what extent will system-wide planning be carried out in a competitive market?	Legal framework for regulation
	How will economic and environmental regulations be coordinated?	Capacity of economic regulators to address environmental incentives and vice versa
		Degree of enforcement capability

the last three, to structure. These categories often blur at the edges, and links listed under one could often feasibly be listed under several other categories.

4.1.1 Commercialization

Commercialization in the power sector is likely to have a range of positive impacts on the environment. First, the principle of cost recovery introduced through commercialization could increase incentives for better management and operation of existing capacity and more efficient transmission and distribution. Cost recovery will also provide an incentive to utilities to implement end-use efficiency programs, especially to reduce losses in rural areas, where the revenue generated often does not meet the marginal cost of supplying electricity.

Second, commercialization is often accompanied by tariff reform that releases controls on electricity rates. Although in industrialized countries reform has led to greater efficiency and often to lower prices, in developing countries reforms are likely to lead to higher prices as government-mandated price caps are removed. Higher prices will provide consumers greater incentive to implement end-use efficiency programs, and may provide market opportunities for firms that provide energy efficiency products and services. However, price effects may be muted because the price elasticity of electricity is considerably lower—perhaps as much as 50% lower—in developing countries as compared to industrialized countries. On the negative side, they will also partially undermine the benefits utilities gain from loss-reducing efficiency programs in rural areas.

Third, the combination of cost recovery and more accurate price signals suggests that a commercialized utility has a strong incentive to identify least-cost ways of extending services to rural areas. This will likely provide an incentive to distributed renewable energy technologies.

4.1.2 Privatization

Privatized utilities have little or no commitment to achieving social or environmental goals, unless they are explicitly required to or are given incentives to do so through regulatory structures. However, on the positive side of the environmental ledger, because a private utility not only seeks to recover costs but also to turn a profit, the incentives for greater efficiency in supply-side efficiency improvements in operation and maintenance also introduced through

commercialization are strengthened. Similarly, a private utility may have an incentive to invest in low-cost technologies, which will promote a switch to cleaner burning natural gas.

On the negative side, the private sector typically has a higher cost of capital and a higher discount rate than does the public sector, both of which provide disincentives to environmentally sensible investments that yield benefits over time. For example, because demand-side management (DSM) measures yield benefits over a number of years, they will yield a lower return under private ownership and will be less attractive. Similarly, higher capital costs and discount rates are unfavorable to renewable energy sources that require large up-front capital investments, and are more friendly to technologies such as gas turbines, which have low capital costs but high fuel costs. In addition, privatization could attract greater investment to the sector, which would increase generation capacity and result in a larger environmental footprint. However, to the extent that new capacity replaces small-scale diesel self-generation by industry, this effect would be mitigated.

If requirements are placed on the private sector to reverse some of the negative incentives just discussed, then it is important that these be put in place prior to privatization. Once existing assets are sold off, it will be politically difficult to change the rules, and sends future investors a signal of inconsistent policy.

4.1.3 Unbundling

As a result of unbundling, the costs of providing generation, transmission, distribution, and retailing services are accounted for separately. As a result, unbundling can provide disincentives to activities that promote energy efficiency. In an unbundled sector, generation and transmission companies do not have direct contact with the consumer in order to undertake end-use efficiency programs. Distribution companies, which are in a position to undertake these programs, are unable to capture the full benefits of greater efficiency because the benefits of reduced generation and transmission requirements are shared with other distribution companies. This separation also limits access to useful information—those on the supply side are cut off from information on consumption patterns, and distribution companies may not have access to billing data from retailing companies. In addition, unbundling makes it more difficult for a single entity to identify and capture the full upstream savings in generation and transmission costs provided by distributed renewable sources. Consequently, there is diminished incentive to invest in distributed renewable power.

Beyond these general observations, the environmental impact of the unbundling process depends on the details of implementation. First, in an unbundled structure, the terms and conditions of access to transmission have environmental implications. Open transmission access could stimulate investments in grid-connected renewable sources. However, this advantage will be negated if renewable power developers are charged higher prices, as is possible under some pricing schemes. In addition, transmission access opens opportunities for interregional electricity exchange that may affect the overall fuel mix. If access to broader markets increases use of low-cost coal plants, which operate under their operating potential, then the result will be negative environmental outcomes. This may be less of a problem in developing countries, where such excess idle capacity typically does not exist. Finally, transmission access could also promote cogeneration. However, in order for this to work, access prices must be low enough for cogeneration to compete with other supply sources.

Second, in the single-buyer model (many generators but one monopoly buyer), the environmental consequences are influenced by the details of power purchase agreements (PPAs). Existing PPAs are structured around low-capital-cost technologies such as fossil fuel plants, which may be unfavorable to renewable energy technologies with equivalent life-cycle costs but higher up-front costs. Important considerations are payment schedules and the treatment of risk in fuel price. In addition, because renewable power projects tend to be smaller but transactions costs tend to be fixed, the transactions costs per unit tend to be higher for renewables. In order to overcome these problems, PPAs need to be tailored to renewable sources and standardized. Moreover, the long-term nature of PPAs is potentially damaging because it locks in technologies that might rapidly become obsolete. A spot market is more likely to take advantage of new technological options as they develop.

4.1.4 Wholesale Competition

Wholesale competition may promote more efficient operation of thermal power plants, but has disincentive effects on energy efficiency. First, when wholesale competition is organized around spot markets in electricity, plant operators have an incentive to run their plants as much as possible in order to maximize returns. Second, price signals in

spot markets are based on short-term costs rather than on life-cycle costs. Because the benefits of energy efficiency are spread over a long period, these benefits tend not to be factored into spot markets. Third, spot markets potentially increase the expected variability in prices, which leads to more uncertain savings from energy efficiency, which are then discounted compared to more certain savings derived from life-cycle costs.

Wholesale competition will likely also negatively affect adoption of renewable energy technologies, because these technologies are more difficult to finance in short-term electricity markets. Spot markets imply a more uncertain revenue stream than do long-term PPAs, thus lenders are less willing to finance the high capital costs of renewable power projects in a short-term market. Moreover, the high debt burden of renewable power projects leaves the projects susceptible to a slump in spot market prices. In addition, some "intermittent" renewable energy sources, such as wind and solar, are at a disadvantage in spot markets because they are unable to provide power on demand in peak periods. This disadvantage needs to be addressed explicitly in the formulation of rules governing the spot market.

4.1.5 Retail Competition

Retail competition, if appropriately structured, may introduce some features that mitigate the negative environmental effects of wholesale competition, whereas others worsen the problem. There are three possible positive effects. First, retail competition requires access to distribution lines. The costs of access require that area-specific marginal costs are assessed. If these costs are charged to customers, their incentive to manage their load is enhanced. Second, a generator who is also a retailer could invest in conservation measures within its customer base, and use the saved energy to compete in other markets. This mechanism is enhanced where generators face an "obligation to serve" in rural areas at an unprofitable tariff. Efficiency measures allow the utility to reduce losses in these unprofitable areas and sell the saved power in profit-making areas. Third, retail competition allows renewable energy suppliers to distinguish themselves by the source of their energy, by, for example, promoting themselves as "green" power suppliers. However, this strategy may be less successful in developing countries, in which electricity prices account for a larger share of budgets, as compared to industrialized countries.

On the other side of the ledger, retail competition will likely lead to an electricity price decline, further diminishing the incentives to undertake efficiency-improving measures. Second, because retail competition allows a customer to switch between suppliers at any time, a retailer is not guaranteed of capturing the full stream of benefits of end-use energy efficiency investments, providing a strong disincentive to make those investments.

4.2 Measures to Improve Environmental Outcomes Following Reform

As the preceding discussion suggests, the environmental impact of electricity reform in a particular place will depend on the details of implementation. In this section, three important elements of implementation are summarized, each with examples of environmentally positive policy steps.

4.2.1 System Planning and Portfolio Management

Effective system planning and portfolio management require that the costs and benefits of alternative electricity supply mechanisms are comparable across a system. This comparison should include consideration of demand-side savings and non-traditional sources such as renewable energy, should internalize environmental and social considerations, and should provide a mechanism for public input. Integrated resource planning (IRP) was the tool devised to meet these requirements under an integrated system. IRP provides a means to minimize the "total resources cost" involved in meeting electricity service needs, which includes not only the costs to the utility and its customers, but also the indirect social and environmental costs.

Centralized IRP at the system level is incompatible with a market structure. However, the essential planning and portfolio management functions remain important, and should be performed through other institutional mechanisms. For example, under wholesale competition, distribution utilities can still conduct, or can be required by law to conduct, an integrated analysis of alternative portfolios, and can draw up a plan based on that analysis. Instead of directly acting on the plan, however, the distribution utility must then solicit bids from the marketplace to meet the plan. Over time, the demand pattern would force a more balanced portfolio of resource investment.

Another approach to internalizing environmental externalities would shift the focus from procurement to dispatch. An "environmental dispatch" approach would use a pollution index to adjust both the order in which electricity is dispatched from generators and the price the generators are paid. Environmentally

clean generators are dispatched preferentially and receive a higher price.

4.2.2 *Environmental Implications of Regulatory Decisions: Transactions Costs and Prices*

Many decisions routinely made by electricity regulators for economic reasons often have environmental implications. Regulatory decisions can unintentionally increase the transactions costs for more environmentally sound forms of electricity. In other cases, opportunities to reduce these transactions costs may be missed. In addition, rules on pricing decisions may systematically underplay the economic benefits that energy efficiency and renewable energy can bring. For example, a regulatory decision that links utility profits to the amount of electricity sold works as an obstacle to energy efficiency. Specifically, a price cap does just this, whereas a revenue cap decouples sales from revenues. In addition, unbundling raises the transactions costs to investing in efficiency—price signals are less clear, distribution utilities have an incentive to "free ride," and generators have inadequate access to consumers. Regulators can play a role in fixing these problems. If retail bills reflect actual service costs, customers receive signals that allow them to tailor demands to reduce costs. On the other hand, if these costs are represented as fixed charges, then effective demand management becomes harder.

With respect to renewable energy, a simple approach to minimizing transaction cost is to offer a standardized contract for power purchase to small-scale renewable energy suppliers. Without such a standard contract, the fixed cost of negotiating a power purchase agreement may be prohibitive for small suppliers. This approach has worked with some success for small hydropower developers in Sri Lanka. In addition, open access to the transmission grid is a prerequisite for renewable energy suppliers to compete in a market. These policies are credited with stimulating a renewable energy industry in India.

The absence of facilities for net metering (a bidirectional meter that registers flow in both directions) introduces a transactions cost that discourages investment in very small-scale renewable energy. The existence of net metering would allow embedded small-scale generation to realize its full potential by allowing the sale of extra power generated on site to the grid. Finally, the bidding rules that govern power pools may be written based on the characteristics of large generation plants, and may systematically disadvantage renewable energy, small generating units, and energy efficiency.

Similarly, there is a set of regulatory decisions that operate through the mechanism by which prices are set. For example, because distribution costs can vary widely across a service area, an average cost-based price conceals the true gains from distributed energy generation and from energy efficiency. By contrast, a deaveraged distribution price would send consumers the appropriate price signals. In another example, subsidies for line extensions also artificially lower the cost of on-grid generation technologies, undermining the competitiveness of off-grid technologies. Finally, transmission prices and terms of access will need to reflect the special circumstances of renewable energy and small generating units if they are to be competitive. For example, distance-based pricing, pricing based on maximum rated capacity rather than actual generation, and a penalty for intermittent transmission could all work against renewable power developers.

4.2.3 *Proactive Policy Measures to Promote Energy Efficiency and Renewable Energy*

In addition to ensuring that economic regulation does not disadvantage energy efficiency and renewable energy, there are a number of proactive measures consistent with competition available to promote these approaches. A number of countries have adopted price-driven, or supply-push, renewable energy incentives programs. In the United States, utilities were required to purchase all the renewable energy offered to them at a price equivalent to the "avoided cost" of generating that power. The United Kingdom has imposed a tax on fossil fuel generation; under the Non Fossil Fuel Obligation program, the tax is used to support nonfossil energy sources. Germany and Denmark have implemented a "feed-in" law, which sets a guaranteed price for the purchase of renewable energy. India has allowed generous depreciation allowances to promote wind energy.

Quantity-driven incentives such as a renewable portfolio standard (RPS) offer another approach. An RPS requires all retail electricity service providers to purchase a specific proportion of their electricity from renewable energy sources. This approach has been adopted in several European countries and in several states in the United States.

A systems benefits charge provides a more general way to support environmentally (and socially) desirable investment requirements by means of a general surcharge across the system. For example, Brazil has adopted a modified version of a system benefits charge, requiring distribution utilities to spend 1% of revenues on energy conservation, of which 0.25%

must be spent on end-use efficiency. More generally, the funds gathered through a systems benefits charge may be spent on increasing access to electricity for disadvantaged populations, on energy efficiency, or on research and development of sustainable energy technologies.

5. GOVERNANCE OF ELECTRIC POWER REFORM

It is increasingly clear that although provision of electricity services would benefit from the application of commercial principles, application of these principles should not exclude a strong government role to maintain essential public benefits (social and environmental) in the electricity sector. As the preceding discussion suggests, there are various mechanisms through which to promote social and environmental interests.

A technical understanding of the links between social and environmental concerns and electric power reform is a necessary, but not sufficient, condition for these interests to be addressed. If decisions about electricity reform are made without advocates for concerns such as access to electricity, price, and environmental outcomes, present and past experience suggests that these issues are likely to be crowded out by more narrow economic considerations. Equitable and environmentally sound reform will require that decision-making processes be open to concerns from rural populations, civil society, environmentalists, and others who typically articulate a public benefits agenda. Hence, equity and environment in electricity reform are equally about an effective system of sectoral governance and a technical or policy understanding of social and environmental concerns in the sector.

SEE ALSO THE FOLLOWING ARTICLES

Electrical Energy and Power • Electric Power: Critical Infrastructure Protection • Electric Power Generation: Fossil Fuel • Electric Power Generation: Valuation of Environmental Costs • Electric Power Measurements and Variables • Electric Power Systems Engineering • Electric Power: Traditional Monopoly Franchise Regulation and Rate Making • Electric Power: Transmission and Generation Reliability and Adequacy • Environmental Injustices of Energy Facilities • Equity and Distribution in Energy Policy

Further Reading

Brook, P. J., Besant-Jones, J. (2000). Reaching the poor in an age of energy reform. *In* "Energy Services for the World's Poor." Energy Sector Management Assistance Program, Washington, D.C.

Burtraw, D., Palmer, K., and Heintzelman, M. (2000). "Electricity Restructuring: Consequences and Opportunities for the Environment." Resources for the Future, Washington, D.C.

Clark, A. (2001). Making provision for energy efficiency investment in changing electricity markets: An international review. *Energy Sustain. Dev.* **V**(2), 26–39.

Dubash, N. K. (ed.). (2002). "Power Politics: Equity and Environment in Electricity Reform." World Resources Institute, Washington, D.C.

Goldemberg, J., and Mielnik, O. (1997). Electricity system reform and sustainable development in Latin America. *Energy Sustain. Dev.* **III**(6), 17–26.

Hamrin, J. (2002). Policy options for building environmental protection into the restructuring process. *In* "Energy Market Restructuring and the Environment" (M. Harriss, Ed.), pp. 197–220. University Press of America, Lanham, Maryland.

Hunt, S., and Shuttleworth, G. (1997). "Competition and Choice in Electricity." John Wiley & Sons, Chichester. .

Karakezi, S., and Kimani, J. (2002). Status of power sector reform in Africa: Impact on the poor. *Energy Policy* **30**(11–12), 923–946.

Kozloff, K. (1998). "Electricity Sector Reform in Developing Countries: Implications for Renewable Energy." 2. Renewable Energy Power Project, Washington, D.C.

Martinot, E. (2000). "Power Sector Reform and Environment: A Role for the GEF?" Global Environment Facility, Washington, D.C.

Reddy, A. (2001). Indian power sector reform for sustainable development: The public benefits imperative. *Energy Sustain. Dev.* **V**(2), 74–81.

U.S. Agency for International Development. (1998). "The Environmental Implications of Power Sector Reform in Developing Countries." 98-05 USAID, Washington, D.C.

Wamukonya, N. (ed.). (2003). "Electricity Reform: Social and Environmental Challenges." United Nations Environment Programme, Nairobi, Kenya.

Electric Power Systems Engineering

MARIJA D. ILIC
Carnegie Mellon University
Pittsburgh, Pennsylvania, United States
JOHN ZABORSZKY
Washington University
St. Louis, Missouri, United States

1. **Introduction**
2. **Basic Objectives of EPSE**
3. **EPSE Tools for System Operation under Normal Conditions**
4. **EPSE for Emergency Operation**
5. **Challenges to EPSE in the Changing Industry**
6. **Objectives of EPSE as a Function of Changing Industry Structures**
7. **A Glimpse into the Future: More Smooth and Secure on a Building Block Structure**

Glossary

forced outage or blackout Loss of electrical power due to the shutdown of a major component, such as a power plant, transformer, or transmission line.
gains Increase in signal power, voltage, or current by amplification.
governor A control device on a machine or engine that is used to maintain a steady speed, pressure, or temperature automatically (as by controlling the fuel supply).
load A device or destination to which electrical power is delivered; also, output of a generator or power plant.
primary controller An electrical instrument or set of instruments that serves as the main regulating or guiding device for a mechanical or electrical system.
transformer An electrical device used to transfer electric energy from one circuit to another, especially equipment that produces such a transfer with a change in voltage.
utility A company that performs a public service (such as sewage disposal or supply of gas for heating) subject to government regulation; in this context, an electricity generation/supply company.

This article is devoted to the electric power systems engineering field, focusing on systems aspects of modeling, analysis, and decision-making tools for operating and planning complex electric power grids. It first describes basic objectives of electric power systems engineering. The primary emphasis of the article is on IT tools and the key challenges that must be overcome to further benefit from using electric power systems engineering (EPSE) tools in support of the electric power industry under restructuring. The material in this article is of direct relevance for managing the electric power systems in a reliable and efficient way without major widespread loss of service to customers.

1. INTRODUCTION

Electric power systems engineering (EPSE) is a subfield of electric power engineering, with the prime concern being performance of an interconnected electric power system. The EPSE subfield has evolved over the past 40 years in response to the industry needs to manage the system (1) securely under unexpected equipment outages and unusually wide variations in load demand patterns and (2) as economically as possible in response to normal load-demand variations. Its growth has been made possible through tremendous progress in the general fields of computers, communications, and control engineering for complex, large-scale systems. Nevertheless, the general tools had to be formulated and made useful in the context of the specific applications to the electric power systems.

This article is being completed the eve after the major blackout in North America that took place on August 14, 2003. Society has once more been caught

 © 2004 Elsevier Inc. All rights reserved.

by surprise in a widespread industry failure to provide electricity to over 60 million customers for over 24 h. Questions about why such an event spread over very large geographical areas are being asked. A typical belief that an electric power grid has automated protection against cascading failures despite an unexpected failure of a major piece of equipment, such as a transmission line or a power plant, has once more been proven wrong. A related category of questions concerns prevention of massive power failures by isolating only parts of the grid that had to be disconnected. Still other questions focus on what should be done to modernize the electric power grid in order to prevent future cascading failures.

This unplanned coincidence between the blackout and the completion of this article has forced the authors to revisit its earlier version. To the authors, the 2003 blackout represents yet another example of challenges to the operating practices and the information technology (IT)-based methods for facilitating reliable and efficient operation of complex electric power grids. It is, unfortunately, one of many in the long list of hard-to-predict, high-impact, cascading events of this type that have taken place since the early 1960s worldwide. It is hoped that the reader will see several connections between the material in this article and the recent events.

The following quotation from *Dynamics and Control of Large Electric Power Systems*, by Ilic and Zaborszky, briefly summarizes the gradual evolution of the U.S. interconnection and points out the rationale for its present complexity:

> *The U.S. structure of power-system operation was developing generally during the twentieth century without much awareness or attention to the implications of the evolution of power-system structure. With the gradual infiltration of the computer into on-line operation in the 1950s, a surge of interconnections developed for utilizing the potential for improved efficiency and economy. As is typically the case in human activities, the development took the form of designs for individual cases of construction based only on the local situation of the case and only during normal, stationary system operation. The vital questions of the effects of off-nominal conditions were essentially ignored. A striking example is the omission in new construction of self-starting house units and the total dependence for starting large power plants on power drawn from transmission lines to neighboring stations. This general attitude was the root cause of the famous Northeast Blackout, a cascading system collapse that disabled ten states for approximately 1 week in 1964 followed by a few other large events, such as that occurring in Manhattan in 1967.*
>
> *After this rude awakening, consciousness of the importance of the structure of the complete system (the interconnected U.S. electric power system) started to gradually creep into the awareness within the field.*
>
> —*Ilic and Zaborszky, 2000, p. 642*

Also, recent electric power restructuring has created very new challenges to the existing EPSE methods. One of the qualitative changes related to the organizational changes in this industry involves the shift from a top-down coordinated monitoring and decision-making to more distributed information managing and decision-making. This change is opening doors to new technical and pricing paradigms for operating complex power grids in a self-healing manner by the system end users adjusting to the changing conditions in a bottom-up manner.

The most recent blackout, the above quote summarizing the beginnings of EPSE, and the recent industry reorganization all concern on-line information essential for determining the system-wide status beyond single components and single utilities. Since the specifics of particular physical and policy/regulatory phenomena vary with the nature of triggering events and system designs, the complexity of data and difficulty of extracting critical information quickly are issues at the heart of any blackout. The critical physical information may concern frequency-, voltage-, or transmission-related physical phenomena; the importance of these factors is well documented in several post mortem analyses of major blackouts. The critical policy/regulatory information may affect capacity available for managing the system as a whole under forced outage and its valuation or it may concern state versus federal responsibility for building stronger transmission lines for exchanging power under outages.

This article is devoted to the electric power systems engineering subfield of electric power engineering, focusing on systems aspects of modeling, analysis, and decision-making tools for operating and planning complex electric power grids. It first describes basic objectives of electric power systems engineering. The primary emphasis of the article is on IT tools and the key challenges that must be overcome to further benefit from using EPSE for operating and planning electric power systems of the future. The final section describes fundamental challenges to developing EPSE tools in support of the electric power industry under restructuring.

Electric power systems engineering must evolve as the industry needs evolve, both as the industry reorganization takes place and as the new hardware technologies become more available and cost-effective. Much could be improved by supporting the

power grid's on-line operations via carefully designed minimal coordination at its various layers. Doing this correctly remains a tremendous challenge in pursuit of increased system-wide efficiency and ability to prevent huge failures.

2. BASIC OBJECTIVES OF EPSE

The main objectives of EPSE are monitoring, analysis, and decision-making tools for facilitating operations and planning of an electric power system as a whole. The EPSE methods are available for monitoring, analyzing, and controlling electric power grids in a semiautomated manner under normal operating conditions, i.e., when the status of the equipment is as planned. The EPSE approaches to managing the system under forced outages are not as mature in comparison with the tools available for normal operation; instead, these methods are generally utility-specific and based on human expert knowledge about the system. The events on August 14, 2003 confirmed once more that industry practices are not fully capable of preventing widespread power failures under forced equipment outages. Moreover, as the electric power industry restructures and new technologies become more cost-effective, the objectives of electric power systems engineering evolve, requiring novel problem formulations and computer-aided tools for relating often conflicting engineering and economic objectives.

The early stages of electric power engineering were primarily concerned with the design and operation of individual pieces of equipment, power plants, transformers, transmission lines, and the like. However, as the electric power grid has expanded over time, it has become essential to understand the interactions among various system components and the effect of these interactions on the overall quality of electricity service. The quality and cost of power service to customers have greatly improved as a result of optimizing utilization of system-wide resources under normal conditions. IT-based power systems engineering tools for scheduling the electric power generation required to meet the time-varying load demand have been basic to efficient use of the least operating cost power generation. In particular, under normal conditions, the cheapest generation is scheduled and delivered across the power system to supply both forecasted demand and demand fluctuations that are hard to predict. Ideally, under unexpected forced outages of major components (power plants, transformers, or transmission lines), power should be delivered from other parts of the system so that the load in the affected area remains served. To achieve this, power systems engineering tools have been developed for scheduling short-term generation reserve to supply customers in case major components of power equipment fail as a result of hard-to-predict forced outages.

The software tools for balancing supply and demand under normal conditions are based on implicit assumptions about the electric power system structure. Naturally this structure allows simplification of very complex electric power system dynamics by the separation of operations into subprocesses that evolve at different temporal rates and at various system levels. It is important to use the IT-based tools to assess the validity of the assumptions and underlying algorithms used, as system conditions vary over time. In particular, the useful structure and unique features of complex power system dynamics are typically violated under forced outages. Moreover, industry restructuring may inadvertently lead to qualitatively new power system architectures, which, in turn, will require careful assessment of the software tools used for operating them.

Some of the fundamental policy- and technology-related changes are discussed in Sections 5 and 6 below. A typical power systems architecture characteristic of normal operation and regulated industry is described in Section 2 in preparation for reviewing the IT tools for operating regulated utilities under normal conditions. The most difficult challenges concern trade-offs between efficient operation under normal conditions and the stand-by reserve to be used in case of forced outages. Since the critical events are unknown and are very hard to predict in a timely manner, conventional probabilistic tools are difficult to use for efficient on-line decision-making.

2.1 Basic Structure of Electric Power Systems

A typical organization of today's electric power systems is shown in Fig. 1. The generation and distribution of electric power geographically are accomplished over one single electric network in the United States. Traditionally in the regulated industry, generation is provided in very large quantities at power plants. Under restructuring, small-quantity distributed generation also enters with a naturally limited scope of providing power locally. Electricity consumption ranges from very large levels (factories) to very small levels (residential).

This immense interconnected network inescapably has an internal structure. Generation in very large

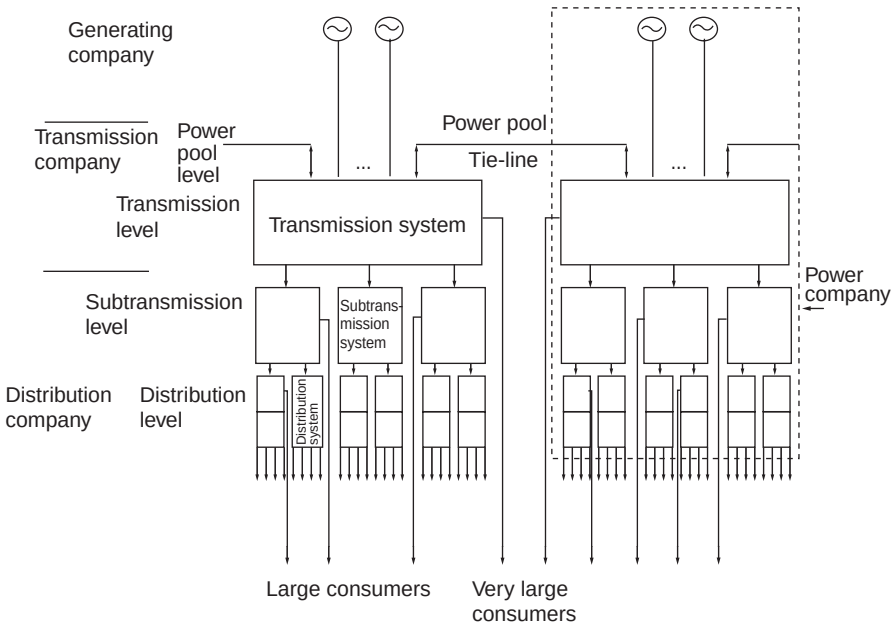


FIGURE 1 Typical organization of existing electric power systems.

quantities in power plants demands high- to very-high-voltage (ranging from 20,000 to 1 million V) transmission line networks. Such networks make up the pool or bulk transmission level network that covers the United States but is also geographically subdivided into large blocks according to ownership, as indicated in Fig. 1. The power distribution process within each geographic bulk transmission block is typically subdivided into geographic subtransmission areas, which historically do not connect to individual generating facilities; instead, only large loads are supplied from the subtransmission networks along with the various distribution sections within each subtransmission area. Finally, the lowest voltage distribution level also splits into two voltage sublevels, ending at the residential households.

The network consists of transmission lines that interconnect nodes, known as buses, at high voltage levels. Different levels of voltage within this structure are interconnected via transformers. Buses are the sole connection points to the transmission system of all equipment:

(1) the sources, that is, generators;
(2) the loads, either large and concentrated like an aluminum plant or distributed like the feeding point of subtransmission networks supplying a composition of smaller loads and eventually leading to distribution lines on residential streets;
(3) reactive-power storage devices, such as capacitor banks or reactors;
(4) increasingly numerous electronically switched power devices (thyristors), such as static Var compensators, the coupling equipment between the high-voltage (HV) alternating current (AC) bus and the high-voltage direct current (HV-DC) transmission lines, and flexible AC transmission systems (FACTS) devices that are in development;
(5) series capacitor compensation; and
(6) lightning arrestors and other equipment connected to a bus by links other than transmission lines.

Buses form large islands of complex equipment connected to the outside world exclusively through the high-voltage lines (AC or DC) of the transmission network.

The business organization of this gigantic system in the traditional regulated sense was single ownership of each vertical column; see Fig. 1. This division preserves the coast-to-coast nature of the total network, but the sale and purchase of power are carried out between companies owning the individual vertical blocks within Fig. 1 under very strict government regulations. In the new industry environment, individual blocks in Fig. 1 are separately owned, and free market trading proceeds for a fee between individually owned transmission blocks under relatively mild government regulation. The above-described strict separation by size is also fading; any amount of power can be bought and sold at any power flow level.

The above-described organizational structure in the regulated industry maps into various power flow levels. The bulk power is transmitted from a very few large power plants to the load centers within a region or from a power pool composed of several subsystems; these subsystems most often are defined as the boundaries of the individual electric power utilities, loosely coupled with the neighboring utilities. The delivery of power between utilities within an electric power pool or within similar horizontal structures occurs by means of an extra-high-voltage (EHV) sparse electric transmission grid, at typical voltages of 500 kV and higher; the pool (or region) level is referred to as a tertiary level. Within each subsystem, the transmission grid is less sparse and at high voltage, typically above 200 kV, known as the secondary level. Finally, major pieces of equipment, such as power plants, transmission lines, transformers, and substations (load centers), are connected to an even lower level voltage network, known as subtransmission; they are referred to as the primary (component) level. The substations are further connected through a low-voltage distribution system to different loads; larger industrial loads are connected at voltage levels similar to the substation voltage level and the smaller loads are progressively connected to lower and lower voltage levels within a distribution system, terminating at the lowest level residential end users.

2.2 Primary Control of an Electric Power System

An electric power system characterized by the above structure serves the basic purpose of producing and distributing electric power to the customers. The system as a whole is operated in anticipation of and in response to the temporal variations in load demand. Electric power is produced by the turbine–generator–governor sets at the power plant locations. The mechanical power is produced by the prime movers. The key control of how much power goes to the network is the governor, which is a local (primary) controller of each modern power plant. Like any other automated controller, a governor has a setpoint and reacts to the deviations from this setpoint automatically. This controller reacts to the frequency deviations from the frequency setpoint value. The faster the turbine turns, the more power is decreased by controlling fuel income to the prime mover by means of a governor (Watt regulator). Standard operating practice has been to adjust the setpoints on governors of generators in each area so that the anticipated area load is fully supplied and the system frequency is nominal. The mechanical power is then converted to the electric power in generators and distributed via the electric power transmission network to the load centers and further distributed to the individual end users via the local electric distribution networks. It is generally hard to predict load exactly. This results in an imbalance between mechanical power produced and power consumed. In this case, according to Kirchhoff laws, the electric power into each node of the electric power network still equals electric power out of the node instantaneously. If there is not enough mechanical power to supply electric power consumed, the generators slow down, resulting in reduced system frequency.

Most of the electric power transmission today is AC, requiring all power generators to operate at the same nominal frequency. In order to maintain frequency very close to its nominal value, each system has several power plants that participate in so-called automatic generation control (AGC). These plants automatically adjust their electric power output by resetting the setpoints on their governors in response to frequency deviations in each control area. Similarly, power plants have so-called automatic voltage regulators (AVRs), which automatically control reactive power produced by the power plant in order to maintain voltage magnitudes at the location of power plants within acceptable voltage deviations. The voltage regulators are also local (primary) controllers. The setpoints of these controllers are adjusted in anticipation of reactive power consumption on the system to ensure that voltage magnitudes at the consumers' side stay within the acceptable limits. Given a setpoint, the primary AVR regulates the reactive power output of a generator so that the voltage magnitude remains within the threshold of the setpoint. This is very similar to the way the governor responds to frequency deviations by producing real power in order to maintain frequency close to the nominal frequency. Some power plants also have power system stabilizers, which represent more advanced local controllers that respond simultaneously to the deviations in frequency and voltage and/or acceleration and have a faster response to the error than the governors and AVRs. If the power system stabilizers are tuned correctly and are implemented on the key power plants, the electric power system as a whole responds in a much more robust way to major disturbances than if the power plants do not have these controllers.

The response of all primary controllers is tuned for the assumed range of operating conditions. The

tuning is achieved by modeling the rest of the complex system as a simple equivalent circuit. However, it is critical to observe that system operating conditions could deviate significantly from the assumed conditions if either (1) load deviates significantly from the assumed or (2) a piece of major equipment is out of operation. In either case, the primary controllers may not perform their function most effectively because their gains are not adjusted to respond effectively to the new conditions. Readjusting these gains in anticipation of or in response to the unusual conditions forms the basis for some form of adaptive control to enable the most effective use of system resources over the wide range of changes in its conditions.

Advanced logic using local measurements in typical primary controllers (e.g., using acceleration measurement in addition to frequency) could contribute significantly to the power plant's ability to adjust better and faster to the changing conditions—such measurements might, for example, result in the development of a controller that would return the system back to its normal frequency and voltage by adapting its power output faster and more effectively. Also, much could be gained from at least partial communication among the primary controllers. This is particularly critical for detecting qualitative changes in system conditions, in order to trigger the change in the logic of the primary controllers determining more adequate settings for the new system conditions. There exist representative examples of the potential gains from wide area measurement systems for stabilizing system conditions under forced outages that would be otherwise uncontrollable without coordination among several controllers; on-load tap changing transformers for regulating voltage at the receiving end of the line have been known to be critical in several voltage problem-related blackouts in France and Belgium; however, the change in their logic when the operating conditions are qualitatively different could potentially prevent these blackouts from occurring.

These seemingly small improvements in primary control of power system equipment could be shown to enhance the ability of the system as a whole to withstand unpredictable difficult conditions without collapsing. The major challenge here is the design of control logic and supporting communications to implement advanced primary control in a robust way—that is, if the controller and/or its communication fail, the system as a whole will still not be affected in a detrimental way. Since this is hard to do with provable robustness, the more common way is to enhance system performance by adding major redundant hardware for stand-by reserves. Establishing a balance between flexibility and over-design has always been the challenge in EPSE.

2.3 System Protection

Under some major forced outages, the system may not be able to return to the nominal frequency after a fault is cleared. In particular, if a fault is not cleared within the so-called critical clearing time, the basic integrity of the operation is endangered. Computing critical clearing time for complex electric power systems still presents an enormous technical challenge.

At each time-step, the proper commands are generated for correcting the abnormality or protecting the system from its consequences. If no abnormality is observed, then the normal operation proceeds for the next sample interval. If the frequency deviations are large and sustained, the under- and over-frequency protective relays automatically disconnect power plants at which the unacceptable deviations occur. Similarly, if a large transmission line goes out of operation, this could lead to current overload in the neighboring transmission lines. A typical scenario here is that a line becomes shorted as a tree touches it, resulting in an unacceptably large current passing through it, and the over-current protective relay of that line disconnects the line to prevent it from further damage. This further activates the over-current protective relays of other lines and they in turn become disconnected.

Protective relaying plays a tremendous role in preventing particular pieces of equipment from becoming damaged under unusual operating conditions. Recall that protective relaying has evolved around the objective of ensuring the safety of individual pieces of equipment under abnormal conditions. For example, transmission lines are allowed to carry only a certain maximum current, blades on turbines can turn only at the speeds specified by the design specifications, and appliances in households operate without being damaged if the voltage applied to them is within the prespecified thresholds of deviations from the nominal voltage at which they are intended to operate. The principles of setting thresholds on relays have evolved over many years and are often system-specific. The relay settings are not changed dynamically to adapt to the changing operating conditions.

For these reasons, relays are key to the cascading events causing major blackouts. The relay settings and their coordination are critical to the efficient use

of the system as a whole. While it is often difficult to differentiate the role of relays in protecting individual pieces of equipment from their role in the system as a whole, it is known that the malfunctioning of protective relaying has played a critical role in several major blackouts. A typical scenario of a blackout of this type starts by a major piece of equipment undergoing forced outage, which, in turn, causes unacceptable over-currents, over- or under-voltages, or unacceptable frequency deviations somewhere else in the system. This situation then triggers the disconnection of the protective relays in the pieces of the affected equipment for safety reasons, which creates further imbalances in the remaining portion of the system, where the protection disconnects the remaining pieces of the equipment. This ultimately results in the disconnection of all pieces of equipment, known as the system blackout or collapse. An even more difficult scenario is related to the false tripping of a relay, which disconnects a piece of equipment without a forced outage actually causing it. Once this piece of equipment is disconnected, the same cascading process of disconnecting the affected pieces of equipment throughout the systems begins. This scenario is known as a "hidden failure." The need to rethink the overall concept of protective relaying in light of system needs and progress in fast computer/communications technologies was recognized some time ago; it has been only recently that this direction is being revived.

Because of this critical role that protective relays play in affecting the interactions among various pieces of the system, it is important to improve their logic and make them both error-free and more adaptive to the changing conditions. Similarly as for primary controllers, this is possible either through a redundant design or through more communications among relays associated with the individual pieces of equipment as these interact to perform system-wide objectives. With the major breakthroughs in the communications technologies, significant progress is being made toward adaptive relaying. Whereas protective relaying could be viewed as the extreme case of primary control, the two problems have not been viewed this way in the past.

2.4 Structure-Based Hierarchical Operation under Normal Conditions

A typical electric power system of the structure described above has millions of nonuniform components and local primary controllers and relays. One of the major goals of EPSE is to design power system control for achieving a desired prespecified performance. Despite many years of research and development, it has remained difficult to provide a reliable, flexible methodology for automated tuning and switching of a variety of control equipment spread throughout the system. In addition to the problem of high dimensionality, the development of such general tools is particularly challenging for this class of systems because of the nonuniformity of control components, reflected in their uneven spatial distribution throughout the system as well as in a large spread of response times.

Instead of developing the most general computer tools for any possible condition, current operating practice is much simpler and different under normal conditions than under the forced outages. Under normal conditions, the power systems are operated hierarchically by various control centers and their on-line interactions. Typically, decisions are made by individual control centers about the status of the area (utility, control area) under often strongly implied or explicitly understood assumptions about the status of the rest of the interconnection and their interactions with the neighboring areas. As explained later in more detail, the entire hierarchical control of today's electric power systems is based on the assumption that any system change would primarily affect the system only locally and that the effects on the rest of the system are secondary. This underlies the decentralized logic of the primary controllers and relays; it further underlies decentralized decision-making by individual control areas within the overall system interconnection. These simplifying assumptions, generally valid under normal operation, lead to the hierarchical control with spatially and temporally separable sub-objectives within an otherwise very complex system.

The critical assumptions under which the interconnection is planned and operated are generally violated under some abnormal conditions associated with forced outages and/or large deviations in load demand from its forecast. When the system enters such conditions, many of the software tools routinely used under normal operation are disabled by the system operators and the main decision maker becomes a system operator. In addition to adequacy problems, i.e., the failure of power plants to supply power, the other major aspect of abnormalities concerns unusual power flow patterns through the delivery system that are very different from the conditions assumed when primary controllers and protective relays are set. It is, therefore, at least in concept, clear why these "malfunction" under

emergency conditions. Under forced outages, system survivability is the prime objective and the cost is secondary. This means that in regulated industry different control areas cooperate by sharing stand-by reserves needed for emergency conditions. The system interconnection as a whole is considered to be viable under contingencies when there are no widespread failures affecting service to a large number of customers and the system grid remains intact; i.e., conditions that may trigger disconnecting pieces of equipment automatically for safety are avoided. The entire concept assumes no active participation by the end users in emergencies.

Because an electric power system is operated very differently in these two modes, one of the largest challenges to effective power systems engineering tools has been to classify the system status according to how close it may be to either its normal condition or the abnormal states. This challenge was conceptualized for the first time by DyLiacco, who introduced the classification of power system operation shown in Fig. 2. There are many degrees and many timescales of abnormal operation that are shown in Fig. 2. Conditions are generally assessed as normal when they are stationary and as expected. All equipment is working that is supposed to work; loads, fuel supplies, water, and weather conditions are within the expected ranges. This classification was later enhanced, but its overall philosophy has remained as initially proposed. Using this classification of system status, a system operator is in a position to schedule power for the forecasted demand and leave the rest to automation at the secondary and primary levels or override the automation and rely on an expert system knowledge about the specifics of the area of interest and act in real time to prevent major system-wide failures. It is worth observing, however, that the boundaries between normal and abnormal conditions are not straightforward to establish, making it necessary to use this classification always with caution. Requiring much margin between normal and abnormal conditions generally results in subefficient use of the existing resources; also, pushing the system to the boundaries of its secure operation for efficiency reasons could result in less robust and less reliable operation. Balancing these two conflicting objectives is generally very difficult.

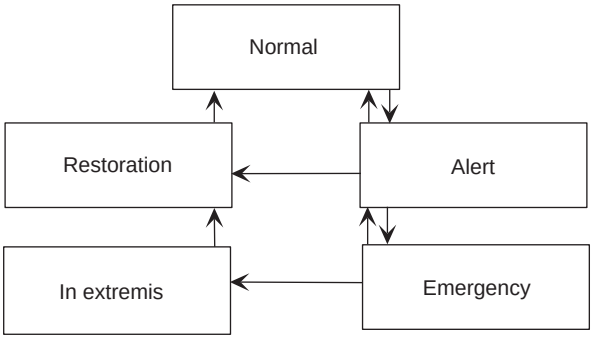


FIGURE 2 Classification of power system operation.

2.5 Operating the System by Decision and Control

To actually operate the system within any of the basic conditions defined in Fig. 2 and during intentional or spontaneous transition between them inherently needs to be done in time-steps where the system state is identified at each time and a decision is made regarding control action to be made at each time-step. After these control decisions are set up at each bus system-wide, then the system responds during the interstate intervals, monitored appropriately to provide the basis for decision at the next sampling time.

The system-wide flow of information and action for this process is of immense size and complexity. Its basic structure is summed up in Fig. 3. Note the two columns for decision and for control at each sampling time. The information base—the system-wide measurement data—is indicated as entering at the bottom of the decision column, where the first task is the classification of the system state into basic operating modes of Fig. 2 or the transition between them. Identifying the position of the current system condition within this column represents the decision regarding the phase of the operation. This information is then passed over to the appropriate block into the control column of Fig. 3 for appropriate control action, which then is passed to the appropriate equipment within the system as indicated in Fig. 3. The equipment then carries out the action called for in system-wide control.

Figures 1 through 3 give a summary of the structure and functioning of a nationwide electric power system. The details and their complexity and volume of the contents of this gross frame of the national power system and economics are immense.

3. EPSE TOOLS FOR SYSTEM OPERATION UNDER NORMAL CONDITIONS

EPSE tools have been developed to accommodate the hierarchical structure of typical electric power

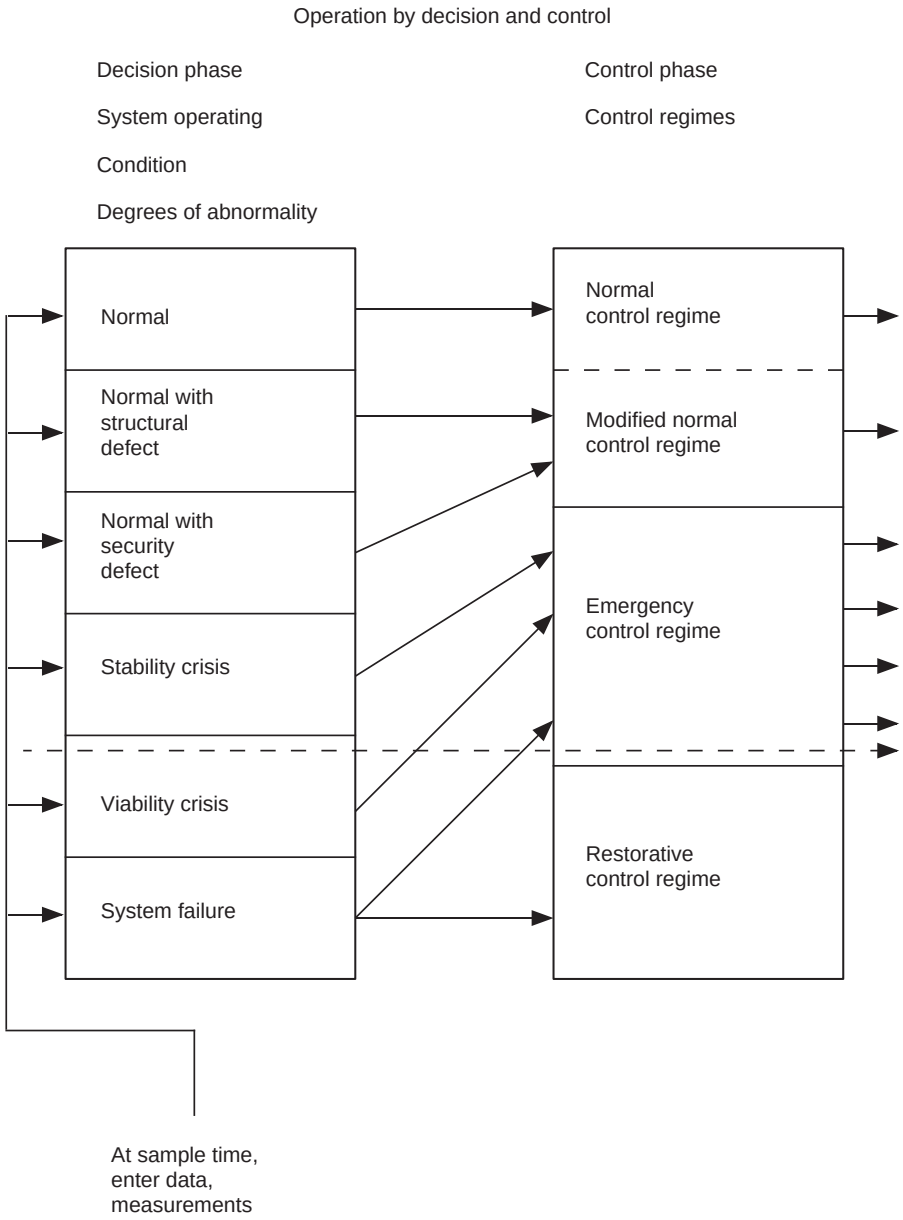


FIGURE 3 Basic structure of system-wide flow of information and action. The dashed line indicates an example of a signal path.

interconnection shown in Fig. 1, horizontally structured into utilities with their own sub-objectives and loosely coupled with the neighboring utilities for sharing generation reserves in case of major equipment outages within the entire interconnection. System monitoring, control, and protection methods are based on these hierarchical structures.

Principal computer-based algorithms for this condition in a decision phase are as follows:

- monitoring and estimation of load and generation;
- static state estimation;
- monitoring of the system loading conditions and security.

Some principal algorithms used for controlling systems under normal operation include the following:

- unit commitment (UC);
- economic dispatch (ED) or optimal power flow (OPF);
- automatic generation control (AGC) to match area real power supply and demand automatically as real power load varies slowly;

- automatic voltage control (AVC) to match area reactive power supply and load and voltage profiles (in some parts of the world only);
- load management.

3.1 System Monitoring

Modern electric power systems are monitored in normal operation to estimate load variations, the status of generators (functional, out for maintenance, or under forced outage), and the status of transmission lines (connected to the rest of the system or shorted out). This information is used as an input to the decision and control phase to balance supply and demand as the system status and levels of load demand vary. In a specific utility area, load variations have typical periodic patterns: annual, seasonal, monthly, weekly, daily, and hourly. Different computer-based methods are used to forecast these system load variations; the long-term forecast (annual and longer) is the hardest and least accurate and the shorter term forecasts are more accurate.

Much progress has been made since the time static state estimation was first introduced to the area of electric power systems. However, this type of estimation is still in use. The basic principle is to use on-line redundant transmission line flow measurements, which estimate false data on the system. The most challenging false data are caused by the incorrectly assumed status of equipment. After many years of researching and developing state estimators, many control centers have them in on-line operations.

The system is also monitored for its proximity to its safe operating limits. The system is generally considered to be in a secure state if none of the operating conditions are violated and the status of the equipment is as expected. The System Control and Data Acquisition System (SCADA) is the basic system in support of area monitoring and decision-making in each control center.

3.2 System Decision and Control

As the loading conditions, power plants' availability status, and the status of transmission lines vary over time, system operators rely heavily on computer-aided software for adjusting overall system resources as needed to supply customers, while maintaining currents/voltage/frequency within the prespecified acceptable threshold of variations. The basic objective of system decision and control in normal operation is to cover the consumers' demand for power using generating sources available throughout the system.

3.2.1 Objectives under Regulated Industry

- The total operating cost of the full system including transmission losses is minimized.
- The price the consumers pay for power is subject to government regulation and control in such a way that the total cost of estimated power production and delivery is covered with adequate profit margins.
- There is enough power available in the system to cover current load estimated statistically using the past year's records plus adequate reserves to cover deviations of the current load from the estimated values and for emergencies of a first-contingency class.
- The power quality is adequate for consumer satisfaction; that is, voltage and frequency are kept within an adequate viability band and kept sufficiently steady. Note that synchronous time frequency is a courtesy to customers. Much less precision would be enough for engineering purposes.
- The system has security that minimizes outages.

3.2.2 Hierarchies in Regulated Generation Control

The only major changes considered in normal operation are the result of load dynamics. Therefore, power systems dynamics under normal conditions are driven primarily by changes in demand. Typical demand curves exhibit several rates of response, ranging from very fast random variations (order of seconds), through hourly, daily, weekly, and seasonal patterns of larger deviations. In today's hierarchical organization, generation-based scheduling is open-loop for balancing anticipated load on weekly, daily, and hourly bases; the secondary level of control (AGC and/or AVC) is closed-loop, automated, and responsive to the voltage and frequency deviations on a minute-by-minute basis, as the unexpected slow and small load deviations in real and reactive power occur. This power balancing on a minute-by-minute basis is performed by each control area within an electric interconnection for the assumed transmission line flows to the other neighboring areas.

There is very little or no coordination of control areas to adjust their line flow exchanges in near real time; instead, control areas agree how much to exchange for the following day and jointly adjust

their internal resources to implement these agreements. However, since the transmission line flows are not directly controllable in typical AC systems, variations from pre-agreed upon schedules take place in real time. This has not been considered to be a big problem, and utilities have established ways to pay back this inadvertent energy exchange approximately once a month, without any major disturbances to the overall system operations, since the system is operated according to the preventive security criteria for the given regional design. As the line flow patterns between control areas deviate for economic transfers—creating considerably different patterns than originally anticipated when the regional system was planned—mainly for implementing energy trades across control areas, the need for on-line minimal coordination of exchanges between control areas at the regional (tertiary) level and even among the regions may become necessary. Several blackouts, including the August 2003 blackout in the Northeast, clearly indicate the need for such monitoring.

Variations of the hierarchies and their automation can be found throughout the world; for example, the system in the United Kingdom does not have automated frequency control, and, moreover, since the entire system is effectively a single control area, tertiary and secondary levels merge into a single level. On the other hand, the European Interconnection is similar in structure to the U.S. horizontal structure, with the control areas corresponding to the countries, instead of particular utilities, respectively. As the electrically interconnected systems open increasingly to the free energy trades across the traditional control area boundaries, the need for advanced EPSE coordinating tools is becoming more pronounced. The blackout of the Italian electric power grid in September 2003, much in the same way as the U.S. August 2003 blackout, was also related to the failure of pieces of equipment outside the affected network; in this case, it was triggered by the failure of an EHV transmission line interconnecting Italian, French, and Swiss subsystems.

3.2.3 Temporal Separation of Basic Control Functions under Normal Operation

The normal-state standard system control functions are summarized as follows:

- Time frame function;
- 2–3 s Inertia, loads, excitation systems;
- 7–10 s Governors;
- 5–10 min AGC;
- 5–10 min AVC;
- 5-30 min ED/OPF;
- 1–10 h UC, restarting and shutting off units.

3.3 System Control

During normal operation, real power and voltage control functions are separable within a certain decoupling error. The operating practices described here are developed assuming this separation.

3.3.1 Real Power Unit Commitment and Economic Dispatch

The unit commitment and economic dispatch are basic computer-aided tools for optimizing total variable operating costs in industry. This is accomplished by using the least expensive units to supply base load and more expensive units for load following and (load) frequency control. At a control center, the most cost-effective units are scheduled on and off to optimize generation use over daily or weekly time intervals in accordance with coordinated unit commitment. The function is open-loop, performed only for anticipated, not actual, system conditions. The longer term economic efficiency is determined by how power plants are turned on and off and cost optimization is achieved by performing economic dispatch as the information about system loads becomes available through a SCADA system.

3.3.2 Automatic Real Power Generation Control

Of course, the actual current load will virtually never match its statistically estimated value for which generation dispatch is done, and the system frequency is also subject to small fluctuations. Consequently, some form of on-line control will be needed to keep the system frequency close to nominal. The specific control generally depends on the horizontal structure of the given electric power system and it could range from a manual frequency control of an isolated system, as in the United Kingdom, through a hierarchical generation control in a multi-control-area setup, as in the United States.

An important observation concerning frequency control is that even the most advanced hierarchical frequency control is strikingly simple with respect to its on-line information and coordination requirements. The area control error (ACE) is the basic measurement to which AGC reacts to balance supply and demand in each control area in near real time. It is defined as the sum of the weighted deviation of the

control area frequency at each particular sampling instant from the nominal system frequency and the deviation of the net power flow exchange with the neighboring areas measured at the same time from the pre-agreed upon net tie-line flow exchange. This entirely decentralized balancing of supply and demand in the entire electric interconnection comprising several control areas is based on the fact that when the ACE is zero for each area, the system is running exactly on schedule. Therefore, if each area independently controls its generation economically to drive its own ACE to zero, then:

- The actual load for the system is covered;
- The system frequency is as scheduled;
- All net tie-line exchanges are at their scheduled values area by area, thus satisfying the scheduled trades;
- If an area is unable to supply its scheduled net tie-line exchange and drive its own ACE to zero, then system frequency will drop and net tie-line exchanges at other areas will rise to cover the deficiency in the load of the deficient area.

3.3.3 Basic Hierarchical System for Real Power Control

• The top (tertiary, regional, interconnection) level simply adjusts the setpoint frequency to drive the actual system frequency and synchronous time back to normal. This one piece of information, the setpoint frequency, needs to be communicated to all areas. This minimal central action will rearrange the area net tie-line exchanges to balance the entire load. This is known as time–error correction in the regulated industry structure.

• Each area individually acts on the change of frequency setpoint by keeping its ACE to zero. No communication is needed between the areas.

• The system is fully decoupled; the areas do not need to communicate with one another or the center. The center needs only to watch the synchronous time to generate its single control command.

• Primary controllers (governors and power system stabilizers) stabilize very fast transients as the power supply/demand balancing at each control area level is performed through AGC.

It is important to recognize the role of accurate short-term load forecasting: the more accurately and more frequently the load is projected, the more efficient the use of available generation will be. The ED and UC are an integral part of a single complex function whose economic performance and technical performance are the result of the actual interdependencies of UC and ED, on one side, and the AGC, on the other. For example, it was estimated that the main inefficiency when performing UC comes from leaving unused generation on some regulating units for frequency control. One could envision, with sufficient computing power, that generation control based solely on dynamic economic dispatch and stochastic unit commitment would be sufficient to balance supply and demand in real time without violating frequency quality specifications.

3.3.4 Reactive Power Economic Dispatch

For an interconnected power system to remain viable, the reactive power must balance at each bus. The problem of reactive power dispatch in the regulated industry is generally bundled with the problem of maintaining (load) voltages within the prespecified limits. The (generator) voltage setpoint values are optimized with respect to certain performance criteria subject to the reactive power balance constraints, the load-voltage acceptable limits, and the limits on generator voltages. Although the voltage control of an interconnected large-scale power system is widely recognized as a very important problem, its basic formulations and solutions are often utility-specific. In the U.S. interconnection, voltage control at the secondary (control area) and tertiary (regional) levels is viewed strictly as a static optimization problem. The most common tool for solving this optimization problem is an algorithm based on OPF.

Compared with generation-based frequency control, generation-based reactive power or voltage control is less standardized and less automated. Only a handful of countries in the world have semiautomated reactive power dispatch of generators.

The generation-based reactive power dispatch falls under the category of the OPF. Depending on the actual choice of the optimization criterion, two formulations are of interest:

1. The cost function is the total fuel cost required to supply a given demand.
2. The cost function is the total transmission loss that occurs as power is delivered from the power plants to customers.

An increased severity of voltage-related operating problems has resulted in an effort to make OPF-based formulations more directly related to these problems than the general OPF methods. Among many developments, the most notable are as follows: (1) an improved formulation of the decoupled OPF;

(2) OPF-based optimization of reactive power margins; and (3) an OPF targeted on improving energy transfer over far electrical distances.

3.3.5 A Hierarchical System for Reactive Power/Voltage Control

A qualitatively different approach to the voltage-control design is hierarchical and is based on temporal separation of reactive power fluctuations and corresponding control levels. Similar to the hierarchical real power-frequency control schemes, the approach is based on decomposing a large system into control areas equipped with decentralized closed-loop secondary-level voltage controllers. A close look into this approach indicates that the objectives of a hierarchical voltage control scheme are analogous to the objectives of frequency control; the main objective at the secondary (control area) level of voltage control is, for example, to update setpoints of generator exciters to maintain voltages at critical ("pilot") points within the prespecified deviations as the reactive power demand and generation deviate from their scheduled (anticipated) values. The hierarchical control for the regulated industry, based on the systematic implementations in several European countries, is often referred to as AVC.

3.3.6 Load Management

It has been well understood for quite some time that having an active response to the changing system conditions seen in frequency and voltage deviations from their nominal values by the end users could significantly contribute to a more robust performance of the system as a whole. In basically the same way that the control area supply/demand imbalance (seen in its ACE) becomes reduced by increasing generation, one could decrease control area load and drive ACE back to zero. Responsive load management is particularly important at times of unexpected shortage in generation and transmission capacity, as well as for using less costly generation reserves instead of peaking units under normal operation.

Except for some early experiments known as demand-side management programs, customers have largely remained passive power consumers. As the industry restructures, the need for more system conditions and price-responsive load will increase. The early vision of homeostatic control, in which all end users react to local deviations in frequency and demand in an automated way, could go a very long way toward more efficient use of resources under normal conditions and reduced reserves for emergency conditions. Sections 5 and 6 state that price-responsive load is the key to the well-functioning electricity markets; the difficult question is providing on-line incentives to the end users to respond at the value as seen by themselves.

4. EPSE FOR EMERGENCY OPERATION

The ultimate challenge to electric power systems engineering at its present stage is reflected in the lack of model-based analytic tools and algorithms for detecting abnormalities in system operation and for deciding what actions to take as the abnormal situation evolves. The classification of potential abnormal modes of operation according to the famous DyLiacco diagram shown in Fig. 2 and its later enhancements suggests that the following operating modes could occur: alert mode and emergency mode.

This section briefly summarizes the tools for monitoring and responding under these abnormal conditions.

4.1 System Monitoring

4.1.1 Alert Operating Mode

SCADA-based monitoring could detect a situation in which conditions are normal; the system is viable and within the acceptable operating constraints. However, the structure has been altered from what was expected by an earlier event—a transmission line is missing, for example. If this is detected, the operating mode is recognized to be under an alert condition.

4.1.2 Emergency Operating Mode

The emergency conditions could be reflected in several different system-wide problems, ranging from stability and viability crisis through system failure. This division, however, is not clear-cut. The conditions overlap, but normally one dominates. Take in sequence:

4.1.2.1 Stability Crisis The system is in a momentary dynamic state, typically caused by a fault, which is sufficiently violent to endanger the integrity of the system.

4.1.2.2 Viability Crisis The system in its present condition is incapable of operating within voltage,

current, frequency, etc., operating constraints, given the available generation and transmission capability for the existing load demand. This condition could result from a fault and is then often preceded by a stability crisis. Nonviable conditions can, however, come on quietly, for instance, if major generation or other equipment does not become available when it is scheduled to come on-line for peak demand.

4.1.2.3 Integrity Crisis The integrity of the system is violated—for instance, load was dropped or the portion of the system is dropped off. Principally, this is a severe and extensive accumulation of stability and viability crises or the damage left by them with islanding and blackout patches in the system. During the crisis phase, events are still occurring and a struggle is still going on to avoid as much disintegration as possible.

4.1.2.4 Restoration This is an extensive and collective control (generally interpreted) effort to remedy the damage left by any combination of the three crises.

4.2 System Decision and Control

4.2.1 Alert Operating Mode

The principal operating algorithms are as for the normal operating mode, but, in addition, future consequences of the structural change must be evaluated and, if necessary, remedied. For example, the structural change may violate security constraints while approaching peak load. Industry practice is preventive in this operating mode, in the sense that the system is operated so that it meets normal operating constraints when any one of the most likely and most severe unexpected structural changes (contingencies) takes place. It is often cost-inefficient because it does not allow for least-cost use of resources without considering future events. To circumvent this problem, the operating limits under structural changes are allowed to be less restrictive in the alert mode than in the normal mode. For example, typically accepted voltage variations under normal operation are within 2% or 0.002 per unit, and in the alert mode they are relaxed to 5%. When a more severe structural change occurs, it may bring the system very close to the limits of acceptable operating constraints. For such conditions, economic criteria are overruled by the security criteria. Different ancillary algorithms for readjusting network flows, bringing on new, more expensive generation, modifying load dispatch, etc., are used to assist the operator under these conditions.

Moreover, since the system operation is still normal under the alert state, the control tools and actions of the normal state are used. The special alert-state control actions consist primarily of equipment switching and changes in control setpoints and schedules to improve security.

4.2.2 Emergency Operating Mode

The decision and control actions in emergency mode are often left to the human operator and are often dependent on the specifics of a particular control area or region. At best, off-line computer-based analysis is performed for scenarios of interest and particular procedures are in place in case the particular scenario occurs. Model-based simulations of such scenarios are generally very time-consuming and are therefore not feasible as on-line means of facilitating decision-making and control in emergency situations, when time is critical. One of the typical decisions concerns the so-called critical clearing time, i.e., the longest time of fault duration, so that when the fault is removed, the system can still return to its normal operation. The following are representative decision and control actions characteristic of an emergency operating mode.

4.2.2.1 Stability Crisis Some of the principal control tools can be:

- Local structural control or protective relaying to eliminate faulty components and local control actions to preserve system stability.
- Stability augmentation utilizing FACTS, technologies, breaking resistors, load skipping, etc.
- Load dropping or major structural changes to separate the system in the least objectionable manner when integrity cannot be maintained. Note that this action pushes the system into the integrity crisis or an *in extremis* condition.

4.2.2.2 Viability Crisis Principal algorithms in an approximate order of increasing undesirability include:

- Frequency reduction of the AGC type;
- Use of spinning reserves or cold reserves;
- Special measures, such as fast turbine run-back;
- Help from neighboring areas;
- Exploiting time-limited overload capability of the equipment;
- Special measures, such as voltage reduction;

- First-stage structural control, such as starting a new generation, returning equipment from maintenance, or load dropping (for minutes or hours);
- Second-stage structural control-islanding involving separation of a power network into parts, some of which are not supplied with power.

4.2.2.3 Integrity Crisis The tools are mostly the same as for the stability and viability crisis.

4.2.2.4 Restoration The main processes taking place during restoration are reconnection of the islands and restoration of loads. This is, of course, an operation by itself and it has been researched extensively over time.

5. CHALLENGES TO EPSE IN THE CHANGING INDUSTRY

Over the past decades, the electric power industry has been undergoing major organizational changes. It is critical to recognize that an electric power structure, along with the objectives and performance metrics used in various EPSE methods for monitoring, decision-making, and on-line operations of a complex electric power system, must be posed in the context of each specific industry structure. The interplay between technical, economic, and policy signals over various time horizons and among various industry layers is fundamental to the overall system performance as a whole. One of the more difficult tasks of EPSE even for fully regulated industry has been how to establish the right balance between system robustness to unexpected events, such as forced outages, and the efficient use of the existing equipment under normal conditions. In other words, the tradeoffs between system security/reliability and the economics of providing power to the end users have been generally difficult to mediate. The performance metrics of the system as a whole have been affected by the utility-dependent performance metrics within the system. Basic security criteria adopted by the regulated industry are discussed above; all software tools for monitoring and decision-making in control centers are based on these well-established objectives.

Electric power restructuring has brought about many fundamental changes to the way EPSE might be performed in the future. To start with, a once vertically integrated utility with a single objective of supplying its native customers has become unbundled into power supply, transmission, and distribution businesses and/or functions. Delivery and consumption are becoming businesses with their own objectives, typically not aligned with the objectives of the power suppliers. In addition, the service is provided through some forms of markets, with marketers managing both the physical, forced outage-related risks and the financial risks with the objectives of making their own profits. Design of markets for various services (delivery, voltage, frequency control) in addition to the basic power market is at rudimentary stages. Providing electric power service competitively asks for complex new economic and policy designs, with a full understanding of their overall effects on the system and the unbundled market participants themselves.

The EPSE tools in support of electric power provision in these evolving industry structures are still rudimentary. One thing is certain, although not widely recognized: it is going to be impossible to continue using the same EPSE models, analysis, and decision-making tools developed for the regulated industry to manage effectively the power system under restructuring. This lack of understanding of the gaps between the engineering tools and market objectives has already caused unexpected outcomes. The remainder of this section briefly summarizes evolving EPSE concepts. The last part of this section provides a possible systems view for bridging this gap between regulated and restructured industry models. The section explains how potentially new paradigms for the electric power system may be based on distributed decision-making by all market participants actively pursuing their subobjectives under incomplete information. Minimal coordination at various industry layers in order to ensure well-understood performance metrics is the ultimate goal. This challenge is briefly elaborated at the end.

6. OBJECTIVES OF EPSE AS A FUNCTION OF CHANGING INDUSTRY STRUCTURES

6.1 EPSE Tools for Vertically Unbundling Industry

6.1.1 Objectives in Deregulated Industry

- Suppliers can obtain the maximum price for their product competitively established in free

commodities market-type economy restricted only in the sense of the rules of exchange.

- Consumers of electric power can pay the lowest price for their purchase competitively established in a free commodities market-type economy restricted only in the sense of the rules of exchange. Note that this may work fine for large customers who can effectively trade at the commodities market to their advantage. Small customers cannot effectively do that, which raises the difficult question of how to establish the price of the power at individual homes, with the potential outcome to retain some kind of government regulation in this area.
- There is enough power available in the system to cover the current load of the system for emergencies of a first contingency class.
- The power supply quality is adequate for consumer satisfaction; that is, voltage and frequency are kept within an adequate viability band and sufficiently steady.
- The system has security to minimize outages.

6.1.2 Evolution from the Hierarchical Operation into Open Access Operation

In the changing power industry, balancing supply and demand in real time is generally market-based. Three qualitatively different market mechanisms are of interest: (1) the primary electricity market(s) for supplying anticipated demand; (2) various "ancillary" markets for ensuring that system conditions remain within prespecified limits as demand deviates in real time from its anticipated pattern; and (3) transmission markets for delivering power requested by (1) and (2).

These markets are necessary for ensuring that the system conditions remain within their technically acceptable limits as power is provided competitively. Here, only some fundamental changes needed in EPSE tools for serving the new industry are summarized.

In addition to the temporal decomposition-based hierarchies present in load dynamics, further temporal decompositions are emerging via the establishment of electricity markets whose on-line decision-making evolves at daily, hourly, and even near-real-time rates. Two observations are relevant here for effective on-line use of generation in deregulated industry: (1) the distinction between predicted and real time is no longer as pronounced as when unit commitment is done assuming demand to be known for the entire next day and (2) it is, therefore, conceptually difficult to "unbundle" the economic generation scheduling process, assuming the predicted load from the on-line control function for responding to unexpected load fluctuations.

Furthermore, some customers are willing to accept power interruptions in exchange for a lower price of electricity—that is, generation is beginning to be provided by some customers on a nonfirm basis. This is a qualitative change from the utility's unconditional obligation to serve and charge its connected customer the same price of electricity over time. This raises a fundamental question about the inherent temporal decomposition of load dynamics and the conceptual applicability of hierarchical power systems operation under competition.

Finally, a transmission grid is managed as a regulated monopoly in a multiownership environment, but is expected to serve markets (1) and (2), which are competitive. This asymmetry of regulatory structure within the single electric interconnection intended to be used in an open access manner, without recognizing ownership boundaries, is creating tremendous challenges for operating the transmission grid consistently. Because of this, systematic policy/regulatory frameworks for moving toward value-based transmission service need to be introduced.

6.2 EPSE for Operating the System under Open Access

Operating under open access implies, at least in principle, the ability for power suppliers to competitively serve customers across utility boundaries within an electric power interconnection. The hierarchical operation in regulated industry must be modified to facilitate delivering power across the boundaries of former control areas established in the old industry without endangering the reliability of the interconnection as a whole. As power is traded across the control areas, each control area must monitor these exchanges in order to balance the remaining needs for power to the customers that it serves as a provider of last resort. The interconnection is being used for power flow patterns different than the flow patterns expected under the originally intended system control and protection design. These market-driven exchanges could result in much stronger couplings among the control areas and, therefore, could require some degree of on-line coordination of flow exchanges between the control areas in order to operate the interconnection as a whole reliably. EPSE tools for coordinating control

areas while enabling power trading across control areas do not yet exist.

One possible approach to operating the interconnection under open access would be a market-based approach to (1) ensuring enough capacity, both generation and transmission, for reliable service and (2) giving incentives to the end users to adjust in an on-line way to the changing technical and/or economic conditions. An early vision of the 20th century electric power operation was introduced by Fred Schweppe: all end users react to variations in local frequency and voltage and the system remains in homeostasis, or balance. The success of this concept, borrowed from biology, must obey two key rules: (1) all end users must respond and (2) the right automation must be implemented. The ongoing industry restructuring has not yet provided a regulatory/policy framework in support of this technological vision.

Furthermore, although the technology for automated participation by the end users is available, it is not cost-effective when implemented by each end user, independent of its size and characteristics. Instead, what are available are cost-effective technologies for managing very large consumers or groups of smaller end-users. Although in concept there should not be any difference between one large end user of X megawatts and several smaller users whose total use is X megawatts, there is a very big difference between two end users of the same average X-megawatt energy consumption, where one end user is adaptive to the system conditions (by "shaving" its peak consumption by redistributing the needs to a different time of the day and/or by reducing its nonessential consumption during extreme system conditions) and the second end user is completely nonresponsive to the system conditions and requires an unconditional supply without considerations for the overall system conditions.

To implement price-responsive energy consumption on-line, communications begin to play a very different new role than in the hierarchical operation under regulation. For an open access system to work well, it becomes necessary to have on-line communication of important technical and/or economic information (1) between the aggregator of various end users and (2) between the aggregators and energy providers, transmission providers, and system operators as well as among several system operators within the interconnection. Note that, at least in principle, a bundled production could be provided by the sellers of power with the delivery service included or the aggregator at the end user's side would look for the best combination of subservices as needed by the end users represented by this aggregator.

A highly effective paradigm of this type has not been actively pursued due to the lack of regulatory policy incentives to differentiate quality of service and value of these differences. In today's industry under restructuring, there is hardly any incentive to the end users to respond to the system conditions, since the service is cost-based with the costs more or less equally distributed among all end users according to their average X-megawatt consumption. The final part of this section provides a possible vision of futuristic EPSE tools that, given the right policy incentives, could lead to the value-based provision of quality of service under open access.

6.3 EPSE for Risk Management under Open Access

As the industry reorganizes, it has become necessary to differentiate between the physical and the financial risks. Major physical risks concern system-wide failures like that experienced during the 2003 blackout, reflected in a fundamental inability of the system to operate as intended and to deliver according to the prespecified performance metrics. This could take place for a variety of reasons, such as (1) underdesign of system hardware for the perceived needs; (2) poor on-line IT-based monitoring, control, and protection coordination of various individual pieces of equipment; (3) human error; and (4) nonresponsive power consumption. The society must take it as given that some very-low-probability, potentially high-impact triggering events will happen. It is, however, not obvious at all how eliminating any of the reasons (1)–(5) would fully guard against widespread system failures. The individual relays and breakers very rarely fail to disconnect pieces of equipment that they are responsible for. However, in order to minimize the effects of triggering events on the system as a whole, the tradeoffs between designing an "optimal grid" for the anticipated use under normal and abnormal conditions and developing novel EPSE tools for managing these physical risks under open access must be understood and managed. No matter how robust the hardware design in place, its system-wide use is not going to be as effective as possible without good software tools.

The role of EPSE software tools in managing physical risks in complex electric power grids has been grossly underestimated. Consequently, the tools are

very rudimentary and very little technology transfer has been attempted. In particular, feed-forward software tools for scheduling available resources for the predicted load are well established for the regulated industry, as reviewed in Section 3. The software and hardware options for managing resources under open access are much more complex than in the regulated industry. EPSE tools for distributed decisions by market participants and by the system operators, in particular with the load being an active decision maker, must be developed. The old industry has been characterized by being truly risk-averse. As the levels of risk-taking become more diverse under open access, EPSE tools must be developed as a means of decision-making by those taking risks.

6.4 Toward EPSE-Supported Protocols under Open Access

Given a strong movement toward the industry restructuring along with an emergence of cost-effective technologies for producing, delivering, and consuming power in qualitatively different ways than in the past, it is essential to develop EPSE tools for flexible provision of service to the end users according to their choice. Instead of pursuing the rigid hierarchical control of the past and/or over-designing the system, an IT framework would provide flexibility in delivery and consumption.

The underlying technological weaknesses that led to the major blackouts of 1966 (Northeast), 1975 (Northeast), 1996 (California), and 2003 (Northeast) are best corrected with information and control technologies. The first weakness is simplistic equipment protection and control. When system conditions evolve to threaten individual pieces of equipment, such as a line, transformer, or generator, protection devices respond by removing that equipment from service. This increases the stress on other parts of the system, possibly causing their protection to act, and so on. The result can be a large-scale cascading blackout. A better alternative would be to shift resources to aid the equipment first threatened, avoiding a blackout altogether. This requires that local controllers of power plants, transmission lines, and consumer consumption act quickly in a coordinated manner to relieve system stress before protective action is required.

Unfortunately, controllers cannot identify emergency conditions and are not flexible enough to act in a coordinated manner, and the protection devices lack the information and decision-making abilities required to postpone their actions. The second weakness is the inability to quickly detect that the system is not operating as desired, because of equipment failure or unusual power demand somewhere within the very large and complex electric power grid. In particular, if individual operators could quickly recognize that their region is under stress, then they could adjust their own resources fast enough to keep operation under control or at least minimize the extent of a blackout. Today's operators do not have the information to do so because there is no on-line monitoring at the regional level or higher. Taken together, these two weaknesses result in a lack of information, flexibility, and coordination that will inevitably lead to future large-scale blackouts as demand increases. With economically minimal investment in information and control technology, this need not be the case.

To improve system operation, the protection and control of equipment, such as power plants, transformers, transmission lines, and loads, must first be improved. This is a challenge because the system response to protection and control actions under stress may be different than under the conditions for which the protection and controls were initially built and tuned. Typical assumptions under normal operation are that (1) the local effects of protection and control are much stronger than their effects elsewhere on the system and (2) producing more power when frequency and voltage drop always helps. These two assumptions are frequently violated when the system is under stress, requiring more on-line information about the rest of the system and more advanced logic to decide how to act.

Next, an on-line coordinated exchange of power between utilities and larger regions of the electric power grid in the United States and Canada must be implemented. This will allow the system operators to better manage large-scale system operation. Managing the entire interconnected grid on-line as a single system is not possible for a variety of technical and business reasons. Instead, a multilayer IT-based framework is required for the system to support the delivery of power between arbitrary generators and consumers. This framework must (1) allow all system users to adjust to changing conditions; (2) allow all consumers to communicate their short- and long-term demands for power and its delivery; (3) have clearly defined technical and contractual responsibilities for the performance of all users; and (4) monitor and control power exchanges between all utilities and regions. Such a framework would facilitate the evolution of the rigid grid of today to an

Internet-like grid in which the users decide when and how to use the system. Not only will this improve reliability, but it will permit value-based reliability in which users pay more (or less) for more (or less) reliable power.

The same IT-based framework for power system management can also support long-term forward markets for reliable service by providing current and future system needs for that service. The money paid in advance for reliable service could be invested in the information and control technology required to implement it. Having forward reliability markets would avoid unpopular regulatory decisions to increase the price of electricity without a good understanding of customer needs for reliable service and their willingness to pay for it. It would also promote the value-based evolution of the grid according to customer needs.

An adequately reliable power system will most economically come from investment in information and control technology. Improvements in system protection and control, built on an IT framework for on-line reliable service, and the forward markets for reliable service complement each other. One without the other does not work and both are long overdue.

7. A GLIMPSE INTO THE FUTURE: MORE SMOOTH AND SECURE ON A BUILDING BLOCK STRUCTURE

7.1 The Current Bulk Power Transmission Network

The electric power system in its present state is a dynamic system of enormous size coast to coast, operating at an equilibrium state that is stable under the slight perturbations imposed by normal load variations. However, faults or mistakes must always be expected even in the best of equipment—and even in humans. An abnormal sudden change like the loss of a large power-generating plant or a large, vital transmission line abruptly changes the system equilibria, including the current operating state, potentially drastically. While the system is trying to move automatically (too fast for human interference) to a new equilibrium, the huge difference in energy present in the network before and after the change between the two operating equilibria is left to swing freely over the system in giant energy waves, the postfault transient oscillation, which challenge the current, voltage, or power-carrying capacity, i.e., the backbone strength of equipment such as transmission lines, generators, and transformers. It is the system operators' duty to maintain a loading condition by their dispatch activity at all times for the existing loading of the system without system breakup for all fault contingencies. Through existing advanced technology including power electronics, computer, and communication, this effort can be successful most of the time, but not always, and one should not forget the new emergence of terrorist possibilities.

7.2 Reorganizing Bulk Power Transmission Network to Building Block Structures Individually Controlled by a Computer Control Center

Drastic dangers call for drastic remedies. The authors' group, led by John Zaborszky, has been developing such a drastic remedy by mostly retaining its currently used power equipment, but rearranging the structure of the interplay of the power system and its controlling computer and communication system for maximum effectiveness.

Recent events served as a reminder that the current power system is vulnerable to widespread failure caused by the failure of even one major component. This vulnerability derives from the fact that after such a failure the system becomes a substantially different one and the operating equilibria cease to be equilibria in the new system. However, the states in the new state space will generally lie in the region of attraction of one of the new equilibria, which in turn will lie quite close to their existing ones except in the disturbed area, so they should move to their new position through a dynamic transient unless the system is becoming unstable. The latter results from the fact that each of the components (building blocks) carries a large amount of energy at an equilibrium and in AC this energy is swinging across any connection at 120 cycles.

So, at either of the equilibria (before or after the event), each building block will carry two components of energy:

- Type 1: the steady state energy content of the building block in the new system after the event.
- Type 2: the difference between the energy content of the building block at the time of the event and the type 1 energy content.

Each of these should be available in the computer control center as it is planning the postevent operation of the system after the event.

7.3 The Outline of the Management of the Envisioned Approach

The proposed new, building block (for example, power plants or generators, or major loads, such as aluminum plant, capacitor banks) structure eliminates any direct control interaction between individual building blocks even if they are neighbors but not their physical dynamic interconnection through the bulk power network system. The only information and control command exchange exists between the control centers and individual building blocks before and after the disturbance.

Each of the individual building blocks in their current physical state contains an amount of energy of type 1. This energy will differ from the energy at the state assigned by the control center to the blocks. In effect, the distribution of the type 1 energy defines the operating state of the system.

The excess energy (type 2) becomes the free swinging power in the system postevent transient oscillation, unless it is eliminated by draining it out or in locally at each of the bulk transmission buses into storage capacitors and then fed back smoothly into the system by routine dispatch (such storage already exists sporadically in the power system). While the excess energy is thus drained out or in (if available in storage) at individual buses, the system voltages and current are sliding on smoothly under dispatch control without wild swinging with only a perturbation ripple, provided enough generation is available. If not, i.e., in an emergency, the central computer control will order dropping of loads in priority order and then they will be reconnected when possible.

In fact, the system may already be broken up into disjoint pieces (e.g., in a disaster situation). If so, the building block structure allows the control centers to make the separate pieces by themselves viable by load dropping and set them up like "individual power companies" to run independently of one another while still connected to the control center until their regular connection can be restored. This approach outlined above can become an effective defense in emergencies of the 9/11 or the New York 2003 type.

7.4 Summary of the Process for Building Block Use Operated and Controlled by the Computer Control Center

In summary, the envisioned building block structure of the power system has the following distinguishing features:

- The system is restructured as a mosaic of building blocks, such as generators, power plant buses, and transmission lines.
- Building blocks are connected only to their neighbors.
- Each block has only one single control command connection going to the system's computer control centers. There is no control command connection, i.e., input, directly between two building blocks, even if they are direct neighbors physically.
- The one single communication between each individual building block and the control center carries full report of the dynamic state of that particular building block to the control center and control command instruction to the block.
- The control center sends to each building block a command to adjust one particular system state variable (located inside in that block) to an assigned value, using a single-variable feedback loop control producing a single control signal input generated independently within the building block for setting the state variable of the block.
- On the basis of block-by-block information received by the control center about components of the system state in each block and the information the center has about the demands on the system, the control center computes the type 1 energy content required for stable operation of the system in each of its blocks.
- Note that this computation is simply the steady-state stable "load flow" for the system.
- Block by block, this information is deducted from the present energy contents of the block.
- This difference is drained in or out (depending on whether it carries a plus or minus sign) of the storage equipment at each particular block.
- The result is that the system has slid over into its desired steady state value.
- No violent transients exist to cause new failures.
- Note that on an all-DC bulk power system, this result derives automatically but it applies as described on a conventional mixed AC and DC bulk power system—with the addition of some more power electronic equipment.

In either case (all HV-DC or conventional bulk transmission), the dynamics of the system trajectories in the state space will change instantly and smoothly to that of the postevent system composition without transient waves but the energy controls of the energy storage devices (typically capacitor banks) will shift virtually unilaterally, to its post event form.

Note the following:

- The above assumes the system to be linear around its operating equilibria and that its response is instantaneous, assumptions that are true only approximately. It needs to be proven that on the actual systems this would cause only a tolerable perturbation around the operating point in case of faults.
- This approach exploits the full potential of power electronics to achieve maximum possible smoothness and security of the system operation without violent transient swings that can cause the spreading of major faults (e.g., East Coast 1965, 2003).
- This, however, requires installation of some power electronics, computer, and storage equipment, which need to be properly analyzed.
- The economy of such full restructuring of the current bulk power system needs to be established to fully utilize the current technology of power system electronics, computers, and communications, even if it requires only the addition of a moderate amount of equipment but major restructuring of their operation into a building block base.

In conclusion, as briefly summarized here, this approach assumes linearity and instant response (both of which can only be approximated, leaving the question: how well?) and assumes that it would be possible to drain out the type 2 energy at each building block so that the system would find itself in the new equilibrium position with no transient. Without any such draining process, as is happening today, the type 2 energy differences create a violent transient, which is proven to cause new faults since it creates highly abnormal voltages, currents, etc., on the equipment until it is dissipated through the line losses and radiation.

With what is proposed here, the system—that is, the bulk power transmission network—would possess a smoothly varying equilibrium state without any over-voltages or currents and hence without the spreading of the disturbances and system breakup following a major fault as in New York 2003 or a 9/11-type attack.

SEE ALSO THE FOLLOWING ARTICLES

Electrical Energy and Power • Electric Power: Critical Infrastructure Protection • Electric Power Generation: Fossil Fuel • Electric Power Generation: Valuation of Environmental Costs • Electric Power Measurements and Variables • Electric Power Reform: Social and Environmental Issues • Electric Power: Traditional Monopoly Franchise Regulation and Rate Making • Electric Power: Transmission and Generation Reliability and Adequacy

Further Reading

Amin, M. (Ed.). (2002). Complex networks. IEEE Control Systems Magazine, February 2002.

Chen, J., Thorp, J. S., and Parashar, M. (2001). "Analysis of Electric Power System Disturbance Data," Hawaii International Conference on System Sciences (HICSS-34), Maui, Hawaii, January 2001.

Cohn, N. (1977). "Research Opportunities in the Control of Bulk Power and Energy Transfers on Interconnected Systems," Special Report on Power System Planning and Operations: Future Problems and Research Needs, EPRI E-377-SR.

Fink, L., and Carlsen, K. (1978). Operating under stress and strain. *IEEE Spectrum* March, 48–53.

Ilic, M., Galiana, F. D., and Fink, L. (Eds.). (1998). "Electric Power Systems Restructuring: Engineering and Economics." Kluwer Academic, Dordrecht, the Netherlands/Norwell, MA.

Ilic, M. D., and Liu, S. X. (1996). "Hierarchical Power Systems Control: Its Value in a Changing Electric Power Industry, Advances in Industrial Control." Springer-Verlag, London.

Ilic, M. D., and Zaborszky, J. (2000). "Dynamics and Control of Large Electric Power Systems." Wiley-Interscience, New York.

Interconnected Operations Services. (1996). NERC WG Report.

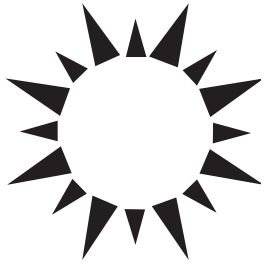
Joskow, P., and Schmalansee, R. (1983). "Markets of Power: An Analysis of Electric Utility Deregulation." MIT Press, Cambridge, MA.

Schweppe, M., Caramanis, F., Tabors, R., and Bohn, R. (1988). "Spot Pricing of Electricity." Kluwer Academic, Dordrecht, the Netherlands/Norwell, MA.

Zaborszky, J., and Ilic, M. (2001). Exploring the potential of the all DC (ADC) bulk power systems. In "Proceedings of the Bulk Power System Dynamics and Control V," August 26–31, 2001, Onomichi, Japan." [Plenary paper]

Zaborszky, J., and Ilic, M. (1999). "General Structure of the Regulated, Deregulated and Other Generation Dispatch." IEEE Power Tech '99 Conference, Budapest, Hungary, Paper No. BPT-440-21.

Electric Power: Traditional Monopoly Franchise Regulation and Rate Making



KENNETH ROSE
Institute of Public Utilities
East Lansing, Michigan, United States

1. **History of Traditional Regulation**
2. **The Utility Franchise**
3. **The Regulatory Process**
4. **The Revenue Requirement Formula**
5. **Rate Base**
6. **Rate of Return**
7. **Expenses**
8. **Cost Allocation and Rate Design**
9. **Limitations of the Traditional Approach**
10. **Summary and Conclusions**

Glossary

cost allocation The process of allocating commission-approved utility cost of service among customer classes based on demand characteristics or other similarities.

cost of capital Percentage return that investors require for the use of their money.

cost of service Total amount of dollars needed to supply the utility's customers, including operating and maintenance expenses, taxes, depreciation expenses, and a fair return on investment.

franchise Exclusive territorial rights granted by state or local authority.

prudent investment test A regulatory standard that requires that costs recovered from customers by a regulated public utility must have been reasonable given what was known or knowable at the time the costs were incurred.

rate base A measure of a utility's total property value, including plant, equipment, materials, supplies, and cash; excluded are accumulated depreciation, amortization, deferred taxes, and customer-provided capital.

rate case Formal process used by state and federal commissions to determine and set rates.

rate design The rate structure or tariff schedule devised for each customer class or subclass; it is intended to collect the cost of service allocated to the customer class or subclass.

rate of return The return on investment received by the utility or its investors; the allowed rate of return is the commission-approved percentage return on the value of the rate base the utility is given an opportunity to earn on its investments that are used and useful for public service.

revenue requirement The total expenses, or the cost of service, determined by a commission to be needed to provide service to the customers.

traditional regulation A form of economic or price regulation conducted by a governmental agency or commission that estimates the regulated firm's cost of doing business and determines a "just and reasonable" price for consumers based on approved costs, including a fair return.

used and useful test Regulatory standard for utility investments and expenditures that must be met to be included in the rate base or recovered as expenses from customers that specifies that investments and expenditures must be used and useful for public service.

Traditional regulation as applied to electric utilities is a form of economic or price regulation that evolved and often changed form over a period of more than 100 years. Also called rate base, cost of service, cost-based regulation, or rate-of-return regulation, this form of regulation is a method that attempts to estimate the regulated firm's cost of doing business and derive a "fair" price for consumers. The concept of public utility regulation was developed from the common law prohibition of unfair competition. It was applied to many industries, including the railroad industry beginning in the mid-19th century and later urban transit (streetcars and buses), telephone, water transportation, manufactured and

 © 2004 Elsevier Inc. All rights reserved.

natural gas, and water utilities. An important goal of this type of regulation is to balance the interests of the firm while protecting customers from monopoly prices and other abuses. The term "traditional regulation" is used here with the caveat that many industries and state and federal regulators have abandoned or significantly modified this basic form of regulation during recent years for one or more sectors. Charles Phillips, Jr., noted that utility regulation is an economic, legislative, and legal concept. These themes are repeated often when regulatory questions arise. Economic concepts are used for rate design, determination of rate of return, and analysis of company incentives for cost control and behavior. Legislative involvement from the enabling statutes, statutory changes made in response to economic conditions and technology changes as well as through the political process, and legal concepts form the basis of regulation and are grounded in constitutional law and set the boundaries for commission authority and utility actions. Other disciplines are also involved in influencing regulation, especially accounting through accounting standards and reporting requirements and engineering through the prominent technical complexities of power generation and delivery.

1. HISTORY OF TRADITIONAL REGULATION

There is some difference of opinion about the origin of traditional utility regulation in the electric industry. Some have argued that it was due to the efforts of Progressive Era reformers who were crusading against monopoly abuses in many industries during the late 19th century. Others maintain that it was the industry itself that decided that regulation was better than either "ruinous" competition or the possibility of government acquisition. Because there is strong evidence to support both views, it is likely that its origin is one of those points in history where opposing interests converge to support a similar solution. However, once a common frame of reference was reached, the actual implementation of the regulatory model was often less congenial.

The regulation of the electric industry generally followed the changing structure of the industry. At the beginning of the industry's history, there was little regulation and it was often thought to be unnecessary. Electric power was seen as a means to further economic development, much as railroads were seen a few decades earlier. The main use for electric power during the late 19th century was for industrial purposes and a few wealthy households. Franchises were granted by municipalities, and rate limits were sometimes set by local governments. Early in the 20th century, most of the power was still generated for industrial use and was produced at the industrial site that used the power. However, central station power, or utility production, became more prevalent and eventually became the leading form of production by the 1920s.

Along with the rise of central station power production and the consolidation of the industry into larger regional companies due to economies of scale in the production and delivery of electricity, utility borders expanded beyond municipalities. As a result, regulation shifted to the state level, beginning in 1907 in New York and Wisconsin and spreading to nearly every state by the 1920s. Large regional and multistate holding companies formed, and investor and consumer abuses were rampant, peaking in 1929 with the financial collapse of the Insull companies. The congressional and regulatory response was to force the breakup of the large holding companies and to place restrictions on the operations and management structures of the remaining regional companies (through the Public Utility Holding Company Act of 1935). This marked the beginning of federal involvement in the industry that continues to the current day. However, state commissions continue to be the primary regulator of retail rates for investor-owned and regulated utilities in the United States.

2. THE UTILITY FRANCHISE

During the late 19th century, it was generally decided that efficient competition was not feasible or practical for the developing electric industry. Rather than allow redundant systems and parallel distribution lines (which occurred in some locations where competing electric companies had their own lines), local municipalities and (later) state governments granted exclusive territorial rights to utilities. This was and is done through an agreement with state or local authority by franchise territorial exclusivity laws, a certificate of public convenience and necessity, or both. All 50 states had laws setting up exclusive retail marketing areas for investor-owned utilities, and many states provided specific service area assignments under territorial exclusivity statutes. These territorial exclusivity statutes provide a

utility with the exclusive right and obligation to serve in an identifiable service area. Many states make service area assignments through the commission granting the utility a certificate of public convenience and necessity.

Because this sanctioned monopoly was by design the only supplier in the area, the utility was required to serve all customers in its service area. In general, utilities have historically had an obligation to serve all retail customers requesting service and to make the investments necessary to provide this service on demand (with limits on this obligation for new customers requiring new lines). To meet this obligation to serve, utilities made investments and entered into legal obligations, such as construction and purchased power contracts, and usually have rights of eminent domain for the construction of new power lines and other facilities. The obligation to serve derives from the utility being granted an exclusive public utility franchise. The requirement to serve is in exchange for an exclusive but revocable right or privilege to serve an area.

In addition to the obligation to serve all who apply for service within their franchise or service areas, public utilities have three other main responsibilities. First, utilities must provide safe and reliable service. Second, they may not engage in undue price discrimination. The prohibition against undue price discrimination requires that all similarly situated customers receiving identical service must be served on the same terms and conditions and for the same price (usually enforced through a required standard tariff filing). Third, utilities may charge only just and reasonable rates and may not earn monopoly profits with rates determined by a commission. What constitutes just and reasonable rates is determined by the regulator, whose decisions are subject to judicial review. These requirements are designed for the express purpose of limiting the potential for monopoly abuse. In return for undertaking these obligations, utilities are granted an opportunity (but not a guarantee) to earn a reasonable return on their prudent investment and to recover their prudently incurred expenses.

The granting of a certificate to serve, along with the accompanying obligations to serve and other legal interpretations and evolving reinterpretations embodied in state and local statutes and case law, constitutes what has become known as the regulatory compact. The regulatory compact does not and never did exist in a written form; rather, it was formed through a series of court cases and commission actions and practices. The regulatory compact is, as Justice Oliver W. Holmes described it in 1912, a balancing of a utility's constitutional rights and its responsibilities.

3. THE REGULATORY PROCESS

The general procedure used to set rates under traditional regulation is through a formal rate case. The full rate case is a quasi-judicial procedure similar to a court proceeding and follows procedures proscribed in the public utility commission's state or federal enabling statutes and the commission's own administrative procedures. Rate cases can last up to a year or longer but are sometimes time limited by statute. Typically, testimony is presented by the participants' witnesses, and cross-examination may be done. A full rate case for a large electric utility may involve thousands of pages of testimony from expert witnesses, hearing transcripts, cross-examination, reply and rebuttal briefs, re-cross examination, and commission statements and decisions. A decision of the commission can be appealed to state and federal courts, which can take several years to resolve the issue. Although deference is often given to the regulatory authority on findings of fact, decisions are sometimes overturned on procedural grounds or due to questions of statutory or constitutional interpretation.

To avoid the time and expense of the full proceedings, settlements may be reached among the parties to the case. Once rates are established, automatic adjustments or rate modification may also occur without the formal rate case procedure. During periods of escalating costs and plant and equipment construction, rate cases may be common and may even overlap or run end to end. However, during periods of stable input costs, low inflation, and limited building activity, regulated firms might not request a rate adjustment for many years.

Rate cases can be initiated by the regulated firm or the commission, and in some states and under certain circumstances, others can initiate rate cases as well. Besides the commissioners, commission staff, and the regulated company, other participants in a rate case may include advocates for large and small consumers, environmental groups, and other "intervenors." Commissions conduct other types of hearings and investigations besides rate cases to consider commission procedures or possible rule changes, certificate applications, utility asset transfers, and other matters.

4. THE REVENUE REQUIREMENT FORMULA

From a purely functional and operational viewpoint, at the heart of traditional regulation is a deceptively simple formula—the revenue requirement formula. The revenue requirement is an approximation of the regulated company's total expenses to provide service to its customers and can be written as follows:

Revenue Requirement
= Operating and Maintenance Expenses
+ Depreciation and Amortization Expenses
+ Taxes
+ [Gross Property Value
− Accrued Depreciation]
× Allowed Rate of Return.

The part of the formula in brackets, gross property value minus accrued depreciation, is referred to as the rate base. Both the rate base and the rate of return are discussed in more detail in the next two sections. Operating expenses include wages, salaries, fuel, advertising, and regulatory costs. Also included are maintenance expenses, depreciation and amortization expenses, and federal, state, and local taxes. Expenses are also discussed later. The revenue requirement formula is designed to provide the regulated firm with an opportunity to recover its investments and expenses to serve its customers. Unfortunately, applying the formula is not just a matter of plugging in the numbers and getting the "correct" answer. Although the formula is straightforward, determining the value for each of the components has been shown, through decades of experience, to be a complex, contentious, and time-consuming task.

Before determining the value of each component, a "test year" or base period must be established. If a historic test year is used, as is most often the case, a 12-month period prior to the filing of the rate case is chosen. Alternatively, a future or forecasted test year, or a "rolling" test year, is used. In any case, the aim is to establish a representative sample of the utility's sales revenues as well as its expenses, taxes, and other costs to serve its customers.

5. RATE BASE

The rate base measures the total value of the firm's property, including plant, equipment, material, supplies, and cash, and may include fuel inventories. Excluded are accumulated depreciation, amortization, deferred taxes, and customer-provided capital. An average rate base for the test year is typically used rather than a value at the end of the test year.

Although all the components of the revenue requirement formula are often contested and argued during rate-making proceedings, the valuation method and what investments are to be allowed in the rate base have been the most contentious and litigious part of the formula over the years. This is because the determination of the size of the rate base often involves large investments that have a significant impact on rates and because a certain amount of subjective opinion and discretion is used to determine it.

The methods and procedures developed over the years to determine rate base value exemplify the fact that traditional regulation is based largely on case law and legal precedent. Before 1900, regulation was price based and set by legislatures or by franchise bargaining by municipalities with companies more often than it was based on the value of companies' properties. In an early and important case in regulatory law, *Munn v. Illinois* (1877), the Supreme Court suggested that although private property cannot be taken away without due process, legislatures can change the law. If there are "abuses by legislatures," the solution was for "the people [to] resort to the polls, not to the courts." This left the matter of regulation to be decided by legislatures and not the courts. The *Munn* case marked the beginning of the regulation of private property held by businesses that are "affected with a public interest," a phrase cited by the Supreme Court with an origin back in 17th-century British common law. Although many of the businesses that were regulated during the 19th and early 20th centuries are no longer regulated or have since faded away (e.g., bakers, innkeepers, milk and ice delivery services), the basis of electric utility regulation still used today by most states has its origins back in the case law that formed the public utility concept.

The judicial noninvolvement of *Munn* was clearly set aside in 1898 by the Supreme Court's decision in *Smyth v. Ames*. In the *Smyth* case, the court laid out its idea as to how utility property should be valued as follows: "The basis of all calculations as to the reasonableness of rates to be charged by a corporation … must be the fair value of the property being used by it for the convenience of the public." Originally, the court had in mind six specific measures of value and allowed for others as well.

This marked the beginning of the "fair value" doctrine. This case was used as a means to determine the rate base for nearly a half-century. Subsequent cases clarified what was meant by fair value. What was upheld by the courts was that fair value was to be defined as either original cost (i.e., the amount actually paid for installing the plant and equipment less accumulated depreciation plus improvements) or reproduction cost (i.e., the cost of plant and equipment estimated at market price levels prevailing at the date of valuation less depreciation).

For the railroad company at the time of *Smyth,* the original cost was higher than the reproduction cost because the investments were made during and just after the Civil War when costs were higher than during the 1890s. The state of Nebraska, represented by William Jennings Bryan, advocated the lower reproduction value. The Supreme Court generally also favored reproduction cost valuation around the turn of the 20th century because of the difficulty in determining original cost (due to different accounting practices and possible fraud). Later, during periods of inflation, positions changed whereby companies supported a reproduction cost standard while regulators and consumer representatives advocated original cost. The principle of reproduction cost remained firmly in place with a majority in the Supreme Court through the late 1920s.

According to Emery Troxel, "the inexact, variable, and sometime inexplicable meaning of fair property value" made determining rate base value extremely difficult. The subjective nature if the calculations inevitably led to the various interested parties coming up with values that differed widely. In an important judicial turning point, Justice Louis D. Brandeis wrote a separate concurring opinion (together with Justice Holmes) in *Missouri ex rel. Southwestern Bell Telephone Co. v. Missouri Public Service Commission* in 1923 that began the move away from reproduction cost. Brandeis stated that he believed that *Smyth v. Ames* was "legally and economically unsound" and that using this method was a "laborious and baffling task."

But it was not until 1944 that the Supreme Court deemphasized the importance of finding a rate base value and the valuing methodology and placed more emphasis on the "end result" of a commission's decision. In *Federal Power Commission v. Hope Natural Gas Co.,* the Supreme Court stated,

> *It is not theory but the impact of the rate order which counts. If the total effect of the rate order cannot be said to be unjust and unreasonable, judicial inquiry under the [Natural Gas] Act is at an end. The fact that the method employed to reach that result may contain infirmities is not then important. … Rates which enable the company to operate successfully, to maintain its financial integrity, to attract capital, and to compensate its investors for the risks assumed certainly cannot be condemned as invalid, even though they might produce only a meager return on the so-called "fair value" rate base.*

In other words, the focus should be on the "end result," not on the specific property of the utility or on a rate base valuation formula. This allowed state and federal regulatory authorities the discretion to also use an original cost standard based on "prudent investment," that is, original cost minus fraudulent, unwise, or extravagant expenditures. The *Hope Natural Gas* case had the effect of changing the emphasis from rate base valuation methodology to return on investment.

To be included in the rate base, the investment must be both "prudent" and "used and useful" to the customers being served. The prudent investment test or standard establishes that costs that were reasonable at the time they were incurred, and given the circumstances and what was known or knowable at the time, are to be included in the firm's rates. This provides the utility with an opportunity to recover these costs. Capital costs are recoverable along with a reasonable return on equity. Reasonably incurred operating costs are also included in the revenue requirement. In general, investments and expenditures have a "rebuttable presumption" that they are prudent when incurred; that is, they are presumed to be prudent unless it can be demonstrated otherwise. Regulators can consider cases of "hidden" imprudence, even if the investments or expenditures are already in the rate base, but are required to consider what was known or knowable at the time the decision was made; in other words, there should be no "Monday morning quarterbacking."

The used and useful test is a much older test established in law. This test holds that utility investments and expenditures must be used and useful to the customers the company serves to be included in the firm's rate base. Also, investments that are no longer used and useful to the public because they are either technologically or economically obsolete can be excluded from the rate base. The used and useful test is intended to prevent a regulated company from deliberately overinvesting in plant and equipment to inflate the rate base and to discourage a utility's investment in assets that will not provide a tangible useful service to its customers. Investments that were once determined to be not

used or useful and were excluded from the rate base may later be allowed in the rate base if conditions change. An example would be if a utility's sales have increased since the last rate case such that additional capacity is needed to serve the customers. In such a case, it may be determined that a previous investment that was excluded from the rate base is now used and useful and should be included in the recalculated rate base.

An additional controversial issue in the determination of the company's rate base is the treatment of plant under construction or "construction work in progress" (CWIP). An issue is how much, if any, of CWIP should be included in the rate base calculation. This treatment of plant under construction, which for electric utilities can be considerable relative to the size of the firm, varies by jurisdiction. Factors that regulatory authorities use to decide whether to add CWIP to the rate base include whether the completion date is near for the work under construction, the project is of a short duration, the company's investment is relatively significant so that excluding it could impair the financial health of the company and its ability to raise capital, the construction project is for a certain type of investment (e.g., pollution control equipment), and the interest charges would be a large portion of the company's earnings. Some jurisdictions will not allow CWIP in the company's rate base at all because, consistent with the used and useful standard, they believe that current rate payers should pay only for facilities that are currently in use for public service. A related and historically older mechanism that state commissions have often used is the "allowance for funds used during construction" (AFUDC), which provides for recovery of the cost of capital for plant under construction once that plant is in service. Because AFUDC was not recovered from customers until the plant was in service, it has been generally regarded as less controversial than CWIP, particularly when the construction costs were viewed as not being excessively burdensome.

In summary, *Munn v. Illinois* laid the constitutional foundation for the regulation of businesses that are "affected with a public interest" such as electric utilities. *Smyth v. Ames* began more judicial involvement and efforts to establish reasonable rates based on fair value. Finally, *Federal Power Commission v. Hope Natural Gas Co.* switched the emphasis from finding a rate base value and the valuing methodology to the importance of obtaining a fair and reasonable end result of the regulatory process.

6. RATE OF RETURN

Alfred Kahn points out that every part of the regulatory process determines the level of allowed earnings for the regulated company. The regulated company's earnings are affected by what investments are included in the rate base, the allowed depreciation methodology, what expenses are allowed to be recovered, and the timing of the recovery, among other factors. But when the commission determines the allowed rate of return, it is explicitly determining the return on the company's property used to serve the public. Kahn noted, as observed by the Supreme Court in the *Hope Natural Gas* case, that there is no scientifically correct allowed rate of return. Commissions must exercise their judgment to determine a "zone of reasonableness," again applying the just and reasonable standard and balancing utility shareholder rights with protecting the public from monopoly abuses.

When a regulatory authority sets the company's allowed rate of return, the authority is determining the return that the utility will be allowed to earn on the rate base, beyond the operating costs and expenses that the company is also allowed to recover. This provides the utility with an opportunity to earn the allowed return, but it does not guarantee such a return. Because conditions (e.g., utility sales) may be different from what was assumed for the test year when the rate of return was set by regulators, the actual return earned by the utility may be different. The utility is not protected from losses due to mismanagement or adverse business conditions. The allowed rate of return refers to the percentage return on the rate base the regulator approves, which may be different from the rate of return received by investors or the utility from the total investment of the utility. Return refers to the total dollars that the utility receives or that investors receive in interest and earnings. In general, there is no provision for retroactive rate making, that is, making up for returns that were allowed but were not received by the utility.

The case law history again played a significant role in determining the method of setting the rate of return. An early and influential case was *Bluefield Water Works & Improvement Co. v. Public Service Commission of West Virginia* from 1923, when the Supreme Court outlined the major elements that regulators have considered over the years when determining rate of return. These elements include that (1) the return should be "reasonably sufficient" to maintain the financial soundness of the utility; (2)

the return should be adequate to maintain the company's credit sufficiently to allow it to raise money necessary for public service, assuming the company is operated efficiently; and (3) the utility should be allowed to earn a return on the value of the property dedicated to public service comparable to that being made on similar investments in unregulated businesses with similar risks, but it is not entitled to earn profits in excess of that. On this last point, the court stated that a public utility "has no constitutional right to profits such as are realized or anticipated in highly profitable enterprises or speculative ventures." The court also noted that a "rate of return may be reasonable at one time and become too high or too low" due to changing financial and business conditions.

In summary, the legal standard is that rate of return should not be confiscatory; rather; it should justly, but not excessively, compensate the utility for use of the shareholders' property. The company is obligated to operate efficiently, invest prudently, and control expenses reasonably. Again, when considering these factors, the regulatory authority attempts to balance the interests of the company's investors with those of the rate payers.

In terms of economic theory, the company should be allowed to earn a "normal" profit, that is, what it could earn on its investments if it were a business with similar risk and capital structure in a competitive industry. This should allow sufficient return on investment so that the utility can compensate its investors for the use of their money, that is, the investors' "opportunity cost."

A utility's actual return to its investors and its cost of capital are determined by financial markets. An investor's return is the sum of the yield of a security (stocks or bonds) and the growth (or loss) of that security. Both yield and growth are dependent on the security's price, which may change each regular business day in the financial markets. If a utility is perceived to be more risky because, for example, investors believe that the firm is less likely to maintain its dividend, the price of its stock will decline. In such a case, investors who hold the utility's stocks or bonds lose part of their investment (on paper at least). The expected return on such an investment increases, and new investors now require an additional risk premium to become or remain investors in the firm. Overall, investments of similar risk will be priced in financial markets to have equivalent expected returns to investors. For example, if something occurs that lowers risk, a security's price will rise, thereby decreasing its yield. Conversely, if risk increases, the price will fall and the yield will increase.

Traditionally, from an investor's standpoint, electric utility stocks and bonds have been considered "quasi-fixed income" securities. Their traditional monopoly status and relatively inelastic demand meant that they were a relatively safe and a steady income source. Growth potential was not the main reason why investors purchased these investments. In general, it was viewed that there was limited opportunity for growth either within or outside the utility's service territory but that the risk of a loss was low as well. This view has changed as the industry is being transformed into a more competitive structure. However, when and where traditional regulation is applied, investor service and rating agencies observe and analyze carefully the results of the regulatory process and how the regulated companies are treated. A key factor they watch is where the allowed rate of return is set.

Leonard Hyman traced historical rates of return and found that between 1915 and 1929 they were in the range of 7 to 8%. During the Depression years of the 1930s, they dropped to the range of 5.5 to 6.0% and remained there until the mid-1960s. After the mid-1960s, because of rising interest rates and inflation, rates of return rose and peaked during the late 1970s and early 1980s. It is not always clear how commissions determined rates of return early on; perhaps they were using prevailing business and interest rates of the time. During the latter half of the 20th century, it became more common for commissions to base the rate of return on the utility's cost of capital.

Table I illustrates how the cost of capital can be determined. Assuming a capital structure of 50% debt and 50% equity (weighted for preferred and common equity), where the utility raises $500,000 from borrowing at 4.0% and another $500,000 at 6.0% from equity, the weighted cost of capital would be 5.0%. In practice, determining the cost of debt and preferred stock is relatively easy because the interest rate on debt and preferred dividends are fixed. However, determining the return on common equity is not so straightforward and has had the affect of increasing the complexity of determining an allowed rate of return considerably. Because common equity is usually a significant part of an electric utility's capital structure, it is important to gauge the return on common equity as accurately as possible.

Hyman described four methods used to estimate return on the common stock equity. First is the comparable earnings method, which is based on

TABLE I

Cost of Capital Example

Capital component	Amount (U.S. dollars)	Cost (percentage)	Weighted cost (U.S. dollars)
Debt	500,000	4.0	20,000
Equity	500,000	6.0	30,000
Total	1,000,000		50,000
Cost of capital = 5.0%			

Source. Based on example from Hyman (1988).

returns on equity of comparable businesses in terms of risk and capital structure. Among the limitations to this approach are that it is difficult to reach an agreement on a sample distribution of comparable companies and that the results are dependent on the selection of companies picked by the interested party in the case. The second approach is the discounted cash flow technique, which is an attempt to objectively estimate a market cost of capital. This is based on the premise that the price of a common share equals the present value of the sum of all future income to be received from the share. This method also has limitations that stem from assumptions used in the estimation. A third approach is the capital asset pricing model. This method divides the return on investment into a risk-free return and an added return for risk and attempts to estimate investors' expectations of a future return from the past performance of stocks with similar characteristics. The fourth method is a group of three financial integrity approaches: (1) set a return to produce a necessary interest coverage ratio, (2) provide equity with a return set at a fixed amount above current debt costs, and (3) provide a return on equity designed to keep the common stock price at or slightly above book value.

These approaches used to determine the rate of return are complex and vary widely by jurisdiction. In actual practice, there are wide gaps among the theory of how rate of return should be determined, opinions on how it should be done, and what in practice is actually done. It is widely recognized that although regulatory authorities may use information from the four methods or other approaches to help inform them of the cost of capital, they still must rely on discretion and judgment, guided by the extensive legal history, to decide on a just and reasonable rate of return within a zone of reasonableness. This explains why the process of determining a rate of return has been described as more like informed conjecturing than like exact science.

7. EXPENSES

As described previously, the total revenue requirements calculation also includes operating and maintenance expenses, annual depreciation expense accrual, income taxes, and other taxes. For electric utilities, fuel and purchased power are typically the largest expenses. A problem with establishing a representative or historic test year is that it is likely that conditions will change, particularly during periods of inflation or fuel cost changes. Of course, changes can either increase or decrease the utility's cost of service. When the utility's earnings decline and it is unable to earn the allowed return on investment, this is referred to as attrition. To remedy this, states may use attrition allowances or use a partially forecasted test year to adjust the rate base or expenses.

Some expenses, such as fuel costs, are nearly certain to change and are very difficult to forecast with any accuracy. This became especially apparent during the rapidly increasing fuel price increases of the 1970s. As a result, most states use automatic adjustment mechanisms that periodically adjust companies' fuel expenses and rates. This allows a utility to recover increased fuel costs without having to go through the time-consuming process of a full rate case. Other expenses have been added to the adjustment mechanism over time, for example, for recovery of purchased power costs.

8. COST ALLOCATION AND RATE DESIGN

Once the total revenue requirement has been determined, the rates charged for each customer class are derived through a two-part process of cost allocation and rate design. The first part allocates the cost of service for each customer class, usually through a cost-of-service study. Customers with

similar patterns of use are grouped into customer classes. For electric service, the general customer classes are residential, business (commercial and industrial), lighting, and miscellaneous. Each general class is often also divided into subclasses. Examples include standard residential service, residential with electric space heating, and small and large industrial and commercial classes. Electric utilities in the United States typically have one to two dozen or more customer tariffs for each subclass with optional service categories such as interruptible, time of use, conservation, or other special rates. A guiding principle of cost allocation is that the rates customers pay should correspond with the costs of serving those customers.

The next step is to design the individual rates for each customer class to provide the utility with an opportunity to collect the authorized revenue requirement. Electric utilities typically have three main components of the customer's rate (often called a rate schedule) for each customer category. First are customer charges that are for costs incurred to serve the customer independent of the amount used such as metering equipment, meter reading, billing, and distribution line. Customer costs vary not by the amount for power used but by the type of customer; these charges are usually expressed in dollars per meter per month. Second are demand charges that are for the fixed costs of plant and equipment capacity needed to meet the customers' maximum or peak demand, usually calculated in terms of kilowatts per month. Third are energy charges that are intended to recover the cost of operation and that vary with the amount of power used such as fuel, labor, and materials. Energy charges are usually in cents per kilowatt-hour and are often seasonal, that is, higher during the summer when demand is higher and power plants with higher operating costs are used to meet demand. Additional charges may be added for an automatic fuel adjustment, conservation or renewable program cost, federal- or state-mandated environmental equipment cost, and/or commission-approved or -mandated programs for low-income household assistance.

The rate schedule will vary by customer class and specifies rates by season, volume use, any discounts, and any additional charges. Seasonal rates are usually used to allow higher prices during the summer (for summer-peaking utilities, e.g., June through September). A uniform rate applies a constant energy charge no matter how much is consumed. If the per unit energy charge falls as quantity consumed increases, this is a declining- or decreasing-block rate (e.g., the first 600 kilowatt-hours per month are 6 cents per hour and each additional kilowatt-hour is 5 cents per hour). Finally, with an increasing-block rate structure, the per unit rate increases as quantity consumed increases.

James C. Bonbright outlined eight "criteria of a sound rate structure" that identify desirable characteristics that regulators have used, or at least should use, to develop rate structures. These criteria were described by Charles Phillips as follows:

1. The related "practical" attributes of simplicity, understandability, public acceptability, and feasibility of application
2. Freedom from controversies as to proper interpretation
3. Effectiveness in yielding total revenue requirements under the fair return standard
4. Revenue stability from year to year
5. Stability of the rates themselves, with a minimum of unexpected changes seriously adverse to existing customers
6. Fairness of the specific rates in the apportionment of total costs of service among the various consumers
7. Avoidance of "undue discrimination" in rate relationships
8. Efficiency of the rate classes and rate blocks in discouraging wasteful use of service while promoting all justified types and amounts of use (a) in the control of the total amounts of service supplied by the company and (b) in the control of the relative uses of alternative types of service (e.g., on-peak vs off-peak electricity)

These principles are widely accepted and used by regulators as guidelines for developing a rate structure. Phillips noted that these principles are broad and ambiguous but are of value because, as Bonbright noted, they are useful for reminding regulators of considerations that might be neglected and for suggesting reasons why practical rate design does not "readily yield to 'scientific' principles of optimum pricing." As with determining the allowed rate of return and other decisions, regulators must again rely on their informed judgment when designing the rate structure.

9. LIMITATIONS OF THE TRADITIONAL APPROACH

Many observers would argue that traditional or cost-based regulation has performed satisfactorily under

many different conditions in terms of cost to consumers, reliability of service, and determining a "fair" outcome. This form of regulation was shown to be adaptable and able to deal with a wide range of crises and challenges placed before it over nearly a century of experience. However, it has come under intense criticism from many observers. One of the more common and persistent limitations pointed out by critics is the lack of incentive a utility has to minimize and control its cost under a cost-based system of regulation. Regulators can attempt to oversee and scrutinize the utility's management and operation of its system, but they cannot possibly impose the same level of cost control that competitive pressures exert on a competitive firm. Other criticisms include the relatively high administrative cost (from being overly cumbersome, legalistic, and bureaucratic), the lack of incentive for a regulated utility to use innovative technologies or rate structures, and the tendency for the utility to attempt to overcapitalize in rate base investments under certain conditions.

In response to these criticisms, regulators have been introducing more incentive- or performance-based mechanisms to ameliorate the reduced incentive to control costs under traditional regulation. The idea of introducing stronger incentives into the regulatory process is not new. For example, the used and useful and prudent investment standards were designed to induce more cost control in utility investments. During more recent years, regulators have used commission-ordered company audits, performance incentives to induce increased plant efficiency, industry benchmarking standards to compare a utility's performance with that of other utilities, price cap regulation that sets rates for several years (with annual productivity and inflation adjustments) and allows the utility to retain any gain (or loss) from improved cost efficiency, and profit-sharing arrangements based on performance.

Much of the electric supply industry's attention recently has been placed on the restructuring of the industry to increase competition in wholesale and (in some states) retail markets. However, traditional regulation continues to be an important regulatory tool. Currently, most states continue to use traditional or cost-based regulation for all electric utility services and will likely continue to do so for the foreseeable future. States that have changed to retail access, allowing retail customers to choose their suppliers, have abandoned cost-based regulation for generation service but will likely continue to use cost-based methods for the distribution services offered by their local distribution utilities.

10. SUMMARY AND CONCLUSIONS

It was recognized that under traditional monopoly franchise regulation and rate making, electric utilities and other regulated companies were subject to a regulatory compact or, as Justice Holmes referred to it in 1912, a "bargain." Based on historical case law and more than a century of development and change, this compact can be summarized as the careful balance between compensatory rates and confiscation of utility property that provides a utility with an opportunity to earn a reasonable return on its investment in exchange for providing safe and reliable power at reasonable cost to all customers who request service. This opportunity is held in check by the used and useful and prudent investment tests as well as by competition from government ownership, fuel substitutes, self-generation, and (more recently) incentive- or performance-based rates and competition from other suppliers. This compact has been rebalanced occasionally to accommodate changing conditions in the industry from economic impacts and technological change. Over decades, state and federal commissions have developed procedures to implement the laws passed and have refined the regulatory mechanisms to determine what investment and operating costs should be recovered, set a return that the regulated company should be allowed to earn on its property, and determine rates for consumers. State and federal commissions have used incentive- and performance-based mechanisms to improve operating efficiency and investment decisions. The recent move to competition for generation nationally and by some states notwithstanding, many aspects of traditional regulation are still in use by most states and will continue to be used for the noncompetitive segments of the industry, such as transmission and distribution services, for the foreseeable future.

SEE ALSO THE FOLLOWING ARTICLES

Electrical Energy and Power • Electric Power: Critical Infrastructure Protection • Electric Power Generation: Fossil Fuel • Electric Power Generation: Valuation of Environmental Costs • Electric Power Measurements and Variables • Electric Power Reform: Social and Environmental Issues • Electric Power Systems Engineering • Electric Power: Transmission and Generation Reliability and Adequacy

Further Reading

Anderson, D. D. (1981). "Regulatory Politics and Electric Utilities: A Case Study in Political Economy." Auburn House, Boston.

Beecher, J. A. (2003). "Economic Regulation of Public Utilities: An Introduction and Overview." Institute of Public Utilities, East Lansing, MI.

Bonbright, J. C. (1961). "Principles of Public Utility Rates." Columbia University Press, New York.

Bonbright, J. C., Danielsen, A. L., and Kamerschen, D. R. (1988). "Principles of Public Utility Rates." Public Utilities Reports, Arlington, VA.

Bryant, J. M., and Herrmann, R. R. (1940). "Elements of Utility Rate Determination." McGraw–Hill, New York.

Burns, R. E., *et al.* (1984). "The Prudent Investment Test in the 1980s." National Regulatory Research Institute, Columbus, OH.

Clemens, E. W. (1950). "Economics and Public Utilities." Appleton–Century–Crofts, New York.

Edison Electric Institute and American Gas Association (2001). "Introduction to Public Utility Accounting." Edison Electric Institute and American Gas Association, Washington, DC.

Fainsod, M., Gordon, L., and Palamountain, J. C. (1959). "Government and the American Economy." Norton, New York.

Hyman, L. S. (1988). "America's Electric Utilities: Past, Present, and Future." Public Utilities Reports, Arlington, VA.

Kahn, A. E. (1988). "The Economics of Regulation: Principles and Institutions." MIT Press, Cambridge, MA.

Phillips, C. F. (1984). "The Regulation of Public Utilities: Theory and Practice." Public Utilities Reports, Arlington, VA.

Posner, R. A. (1999). "Natural Monopoly and Its Regulation." Cato Institute, Washington, DC.

Troxel, E. (1947). "Economics of Public Utilities." Rinehart, New York.

Electric Power: Transmission and Generation Reliability and Adequacy

FRANCISCO F. GARCÉS
Universidad Nacional de San Juan
San Juan, Argentina

1. Introduction
2. Generation System Reliability Evaluation
3. Composite Generation and Transmission Reliability
4. Electric Power Industry Restructuring and Reliability

Glossary

adequacy The ability of the electric systems to supply the aggregate electrical demand and energy requirements of their customers at all times, taking into account scheduled and reasonably expected unscheduled outages of system elements.

bulk electric system The part of the system that encompasses the electrical generation resources, transmission lines, interconnections with neighboring systems, and associated equipment, generally operated at voltages of 110 kV or higher.

distribution system The part of the power system that functions to move electric energy from the bulk supply points to individual consumers.

embedded generation The generation sources that feed into middle- or low-voltage distribution networks.

generation system The part of the power system that functions to convert different forms of primary energy in electric energy.

reliability The ability of a technical system to perform its function properly, for the period of time intended, under the operating conditions intended.

restructuring The transformation process of the electric power industry, from vertically integrated monopoly utilities to market-oriented structures.

security The ability of electric systems to withstand sudden disturbances, such as electric short-circuits or unanticipated loss of system elements.

transmission system The part of the power system that functions to transport electric energy from generation points to bulk supply points.

Power system reliability is the field of engineering devoted to assessing the adequacy and security of power systems. Adequacy is defined as the ability to meet the aggregate electrical demand and energy requirements of customers at all times, taking into account scheduled and reasonably expected unscheduled outages of system elements. Security is the ability of electric systems to withstand sudden disturbances. It is generally understood that adequacy assessment involves steady-state consideration of the systems, whereas security assessment must include analysis of the dynamic system behavior following disturbances. This article deals with the problem of bulk power system adequacy assessment, involving both transmission and generation systems.

1. INTRODUCTION

The behavior of technical systems is conditioned by internal and external factors, many of them of stochastic nature, that can lead to system failure. In order to evaluate the reliability of a system, possible system states should be identified, the ability of the system to perform its function in each of these states should be analyzed, and probabilistic measures about the occurrence of system states should be calculated. The function of a power system is to supply consumers with electric energy economically, at any time and location, observing reliability and quality standards. The practical economic impossibility to store significant amounts of electric energy implies that a balance between generation and load, including network losses, must be maintained at all points in time.

According to the North American Electric Reliability Council, the reliability of the interconnected bulk

 © 2004 Elsevier Inc. All rights reserved.

electric systems is defined through adequacy and security. Both adequacy and security are affected by stochastic events. It is generally understood that adequacy assessment involves steady-state consideration of systems, and that security assessment must include the treatment of the dynamic system behavior following disturbances. Tools for adequacy assessment, based on probabilistic procedures combined with methods for steady-state power system analysis, involve the solution of large nonlinear equation systems. The analysis of power system security requires that sets of differential equations describing the dynamic system behavior be solved. Even considering the powerful computers available today, the combination of dynamic analysis with probabilistic procedures leads to very time-consuming calculations. Generally, security analysis of power systems is carried out within a very short time frame after a disturbance, disregarding the possibility of multiple independent events.

The typical structure of a power system includes the subsystem generation, transmission, and distribution components. Generation and transmission affect bulk power system reliability and are the focus here. The most important types of studies of bulk power system reliability involve generation system reliability and composite-generation and transmission-power system reliability.

During the last 20 years of the 20th century, the industry of electric power supply underwent a great structural transformation, going from a regulated, vertically integrated monopoly to a more competitive industry that is thought to be better suited to provide service more efficiently. The ongoing restructuring process, however, adopting different structures in different regions, generally involves privatization and unbundling of generation, transmission, distribution, and retail activities. The change of paradigm in the electric industry, from minimizing total system costs to maximizing system-participant profits, can strongly affect system reliability. Because most of the tools for reliability assessment currently in use have been adapted from those developed for vertically integrated systems, it is relevant to question the impact of change on reliability.

2. GENERATION SYSTEM RELIABILITY EVALUATION

The aim of generation system reliability calculations is to find the generation capacity needed to achieve a desired (often called optimal) reliability level. Because electricity cannot be stored economically in large quantities, uncertainties affect the amount of load to be supplied, and components are subject to unscheduled outages, it is necessary to have generation reserve for a reliable service.

A first approach in generation reliability evaluation assumes that network components are substantially more reliable, compared to generation units, and that those networks have enough redundancy. In this context, a fundamental simplification can be made, assuming that power can be moved unconstrained from suppliers to consumers.

2.1 Reliability Criteria and Indexes

The optimal reliability level can be obtained, minimizing the total cost, including reserve and outage costs. With increasing reserve, the utility costs will increase due to the investments needed and the outage costs will decrease, and vice versa. The actual determination of the optimal reliability level requires knowledge of outage costs. This is problematic because it involves both objective measurable quantities and subjective perceptions of customers. Generally, the reliability level to be maintained is determined on the basis of estimation of outage costs by regulatory agencies, which set planning and operation standards to guarantee quality, reliability, and security of the service.

Historically, deterministic criteria, such as a percentage of the peak load or the capacity of the largest unit dispatched, were used to define generation reserve requirements. Then, probabilistic criteria were developed and implemented to take into account the stochastic nature of factors affecting reliability. One of the most widely used reliability indexes for generation systems is loss-of-load probability (LOLP), which is the probability of the customer load being greater than the available generation capacity. Other well-known reliability indexes are loss-of-load expectation (LOLE), generally given in number of days per year on which the peak load is expected to exceed the available generating capacity, and expected energy not served (EENS), which gives the expected megawatt-hours not supplied due to capacity shortages in a given time unit (a year, for example).

2.2 Reliability Calculation

The first step in reliability evaluation of generation systems is the modeling of the stochastic behavior of generation units. Many different generator models for reliability studies have been developed, varying in

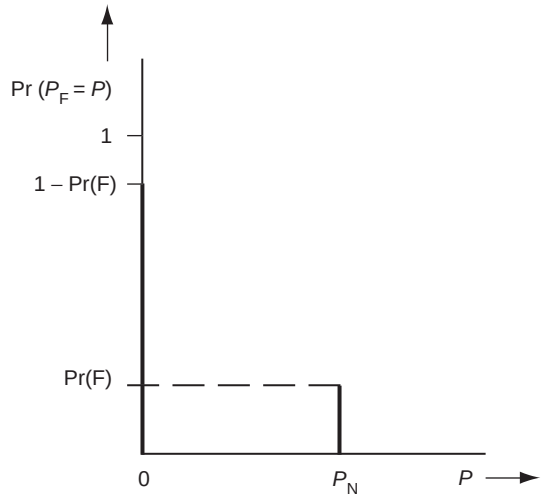


FIGURE 1 Simplified two-state generator model.

the degree of sophistication to describe the actual behavior of generating units. From the system viewpoint, it is important to identify those generator states in which reserve is needed because the generator is not able to supply load. These states can be combined to form the failure state, obtaining a two-state generator model such as the one depicted in Fig. 1. Given the failure probability function $\mathrm{Pr}_i(\mathrm{F})$ of each generator i in the system and assuming independent behavior of generators, the cumulative probability function of unavailable capacity, $\mathrm{F_F}(P) = \mathrm{Pr}(P_\mathrm{F} \geq P)$, can be calculated through simple convolution of the probability functions of all generators in the system. Figure 2 shows the qualitative form of the function $\mathrm{F_F}(P)$.

The reserve capacity P_R of a generation system is defined as the difference between installed generation capacity and maximum load to be supplied. For a given value of P_R, LOLP can be calculated as the value of $\mathrm{F_F}(P)$ for $P = P_\mathrm{R}$ (Fig. 2). According to its definition, LOLP is the probability that the system cannot fully supply the customer load. This happens when the outage capacity exceeds the available reserve. Because the load is known only as a forecast value, it should be modeled through a probability distribution. Usually a normal distribution with mean equal to the forecast value is used. Assuming independence of generation and load, the probability function describing the stochastic load behavior can be included in the convolution calculation to evaluate the function $\mathrm{F_F}(P)$. On the other hand, if a maximum acceptable value of LOLP, called here $\mathrm{LOLP_{max}}$, is given, the generation reserve, P_Rn, needed to achieve the reliability level defined through $\mathrm{LOLP_{max}}$ can be obtained as indicated in Fig. 2. With the exception of the simpler, deterministic approaches, this is probably the most common methodology used by system operators to calculate generation reserve requirements.

LOLE is determined as the annual average value of LOLP calculated for each daily peak load in the year, multiplied by the number of days in the year. The weakness of the indexes LOLP and LOLE is that they do not give any information about the severity of a failure. The index EENS overcomes this problem, encompassing probability and severity of the failure, and can be calculated using the yearly load duration curve shown in Fig. 3. For each value of available generation capacity $P_\mathrm{A} = P_i$, with probability $\mathrm{Pr}(P_\mathrm{A} = P_i)$, there is an energy deficit value, $E_{\mathrm{d}i}$, such that

$$\mathrm{EENS} = \sum_i E_{\mathrm{d}i} \cdot \mathrm{r}(P_A = P_i).$$

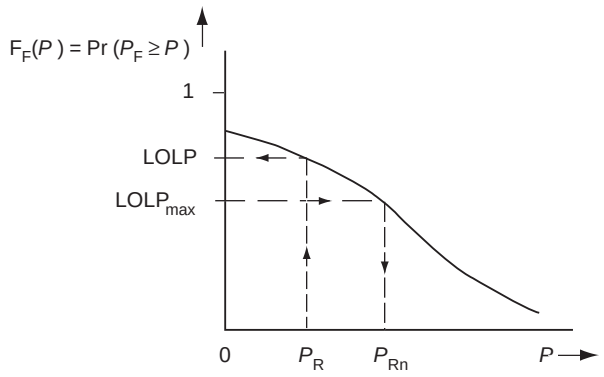


FIGURE 2 Loss-of-load probability and reserve requirement calculation.

2.3 Embedded Generation

An increasing concern about environmental issues has led many governments to establish incentives for

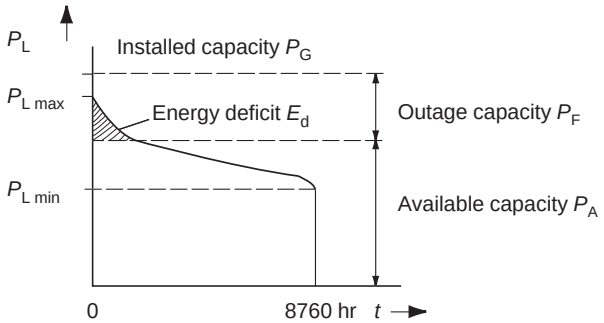


FIGURE 3 Calculation of expected energy not supplied.

clean generation using renewable energy sources such as sun or wind. Unlike large conventional generation units, small solar or wind generators often feed into the distribution networks, creating a completely new set of problems. However, reliability studies of systems with sizable participation of embedded generation should consider this type of generation. Even when network constraints are neglected, typical in reliability evaluation of generation systems, some new problems arise: (1) prediction of embedded generation output is no easy task because often embedded generation is not dispatched by the system operator; (2) primary sources, especially wind, are highly variable and difficult to forecast; (3) there are dependencies in the behavior of generators, and loads are correlated with the availability of primary energy sources, notably the sun; and (4) frequently, distributed generators are not permitted to generate if the distribution network they are feeding into is disconnected from the bulk power system. Although some efforts have been made in order to overcome these problems, further work is needed to develop models that properly consider the contribution of embedded generation in reliability studies of generation systems.

3. COMPOSITE GENERATION AND TRANSMISSION RELIABILITY

Generation systems without the constraints that originate in transmission networks connecting generation with load points were considered in the previous section. Because actual transmission systems are not 100% reliable and have limited transport capacity, reliability evaluation of bulk power supply requires consideration of both generation systems and transmission networks. The distribution of power flows in the transport network follows Kirchoff's laws, which, together with network impedances, generator outputs, and bulk supply point loads, determine the loading of each transmission line in the system and the nodal voltages. A load flow model, one of the most powerful tools for power systems analysis, allows the calculation of voltages and load flows in transmission networks.

The function of the composite system is to generate enough energy to satisfy the load and to transport it to the supply points. As already stated, reliability analysis includes adequacy and security evaluation. Security evaluation requires analysis of the dynamic behavior of a system when disturbances lead to changes of the system state. Although the subject of intense research, probabilistic assessment of the dynamic behavior of power systems has not yet reached the maturity needed for successful implementation in real power systems. Adequacy studies analyze power system states from a steady-state viewpoint using a load flow model. A system state will be termed adequate whenever it has enough generation available to supply the customer load fully and has the ability to transport the energy from the generating plants to the supply points, observing voltage and thermal limits. Dynamic constraints are often taken into account indirectly through reduction of transmission line transport capacity limits on the basis of stability considerations.

3.1 The $(n-1)$ Criterion

The most frequently used approach to assess adequacy of transmission and composite systems is the so-called $(n-1)$ criterion, which states that a system must be able to withstand any single contingency without affecting functionality. Using load flow calculations, all system states, including one major component (generator, line, or transformer) forced out of service, are examined. If the system load is fully supplied, satisfying all constraints (including active and reactive generation limits, nodal voltages, and thermal or stability capacity limits), the considered state is deemed adequate. Otherwise, the state is considered inadequate. Then, reinforcements are proposed until all $(n-1)$ states fulfill the adequacy criterion.

The appeal of the $(n-1)$ criterion derives from its simplicity. The $(n-1)$ criterion is also the most frequently used approach to composite systems security analysis. Replacing the steady-state load flow analysis of each $(n-1)$ state by a dynamic behavior analysis following each possible single failure of the system, the system states can be classified as secure, when all constraints are satisfied, or insecure otherwise.

The weakness of the $(n-1)$ approach to composite system adequacy assessment is that it considers neither the probability of the analyzed states nor the magnitude of the problem in case of inadequate states. It must be pointed out that the $(n-1)$ criterion neglects all multiple contingencies regardless of their probabilities. An extension of the approach to consider multiple contingency states and the introduction of load point-related, individual component-related, and system wide reliability indexes help to overcome these problems. Regarding the

severity of the problems, the $(n-1)$ criterion can be extended to consider a more detailed classification of the system states: (1) a system state with fully supplied load, no constraint violations, and fulfilling the $(n-1)$ criterion is termed the normal state; (2) when the load is fully supplied, no constraint is violated but the $(n-1)$ criterion is not fulfilled, the state is called the alert state; (3) if the load is fully supplied but there are constraint violations, the system state is designated as an emergency state; and (4) any state with load curtailment is called an extreme emergency state. Probabilistic indexes can be calculated for each of these states.

3.2 Reliability Indexes

System-wide reliability indexes reflect the average adequacy of the composite system but do not give any information about reliability as observed at individual bulk supply points. Therefore, it is important to use a combination of both system-wide and supply point-related indexes to describe the reliability of composite power systems. A variety of different indexes have been proposed, including probability, frequency, and duration of failures, load not supplied, energy curtailed, etc. Only a few of them are presented in Table I as examples of often used reliability indexes. For each specific reliability study, the more appropriate reliability indexes should be selected.

Probability indexes can be calculated for a given load state, as in the case of the index LOLP in the previous section, or as average values for a given time period, as in the case of LOLE. It is obvious that energy-not-supplied indexes require the consideration of a certain time period. Because the load state of a power system changes with time, leading to changes in the generation units scheduled to supply load and in the network state, any calculation of average reliability indexes requires definition and analysis of representative load states for the time period of interest.

TABLE I

Reliability Indexes

System-wide indexes	Supply-point or component-related indexes
Probability of failing to fully supply the system load satisfying all technical constraints	Probability of failing to fully supply the load at a given bulk supply point
Probability of any capacity limit violation	Probability of capacity limit violation for a given component
Probability of any voltage limit violation	Probability of voltage limits violation at a given network node
Expected energy not supplied for the whole system	Expected energy not supplied for a given load point

3.3 Reliability Calculation

There are six essential steps in the procedure for reliability evaluation of composite power systems.

3.3.1 Definition of Representative Load States To Be Considered

The question about the load states to be considered depends on the aim and scope of the study. If, for example, the probability of the system not being able to supply the load fully at a given point in time (usually the peak load) is desired, the load at that time point is to be considered. When the expected value of energy not supplied in a given time period is required, many load states representing the possible load conditions in that period must be studied.

3.3.2 Active and Reactive Power Dispatch

Once the load states are defined, it is necessary to schedule the operation of the generation system to supply the load. This has been done in the vertically integrated industry using well-known and generally accepted optimization methods. In the new competitive environment, generators submit to the Independent System Operator (ISO) bid prices at which they are willing to sell energy, and the ISO allocates the resources economically.

3.3.3 Load Flow Analysis

Given the loads and the generation scheduling, a load flow analysis is performed in order to verify compliance of technical constraints, such as transport capacity limits or supply point voltages.

3.3.4 Contingency Simulation

Contingency simulation is the heart of composite systems reliability evaluation and is based on a load flow model. It aims to find out the system states in which there are load curtailments, constraint violations, or both. The selection of the system states to be analyzed can be made using simulative or analytical techniques. Simulative techniques, usually called Monte Carlo methods, are based on the use of computer-generated random numbers to determine the state of each system component. Outages of

components with high failure rates are likely to be simulated more frequently, compared to outages of more reliable components. The simulation process produces a sample of the possible behavior of the system. Reliability indexes are calculated through statistical evaluation of this sample. The number of simulations required to obtain a representative sample can be calculated as a function of the desired accuracy of results.

In the analytical approach, the system states are enumerated, generally in an increasing order of contingency level, stopping either after a given contingency order or when the state probability falls below a given value. Power system components are very reliable, implying that for increasing contingency order, state probabilities rapidly become very small. For each system state, selected either through simulation or enumeration, probabilistic measures regarding its occurrence are calculated and a load flow analysis is performed to verify if any constraint is violated. In case of constraint violations, the procedure must mimic the operator behavior, determining and simulating the introduction of remedial actions. The impact of these corrective actions is then studied through a new load flow calculation. After all corrective possibilities for a given contingency state have been used, the remaining constraint violations determine that the analyzed state is a failure state.

3.3.5 Corrective Actions

The simulated corrective actions must reflect with reasonable accuracy the actual possibilities of the operator in a real system. The possibilities of the system operator include modification of active and reactive power dispatch and change of the network topology through connection or disconnection of transmission facilities. Optimization tools with the objective of operating cost minimization can easily include the possibility of changes in generation dispatch to satisfy technical constraints. It is not so easy to consider topology modifications because of the discrete nature of this problem. The most popular tool used by power engineers to obtain an economic dispatch considering network constraints is optimal power flow (OPF). OPF is an extension of the well-known simple economic dispatch (ED). The objective function of ED is minimization of generator operating costs subject to the constraints that generator output must be within technical limits for each generator and the total output of the dispatched generators must equal the system load, including transmission losses. OPF adds to the economic dispatch a set of alternating current (AC) power flow equations as equality constraints. The solution of the OPF problem is an operating plan with minimal costs and a complete load flow solution. A further development of OPF is the so-called security-constrained OPF, which introduces inequality constraints on power flows and voltages.

Security-constrained OPF allows the calculation of optimal corrections of active and reactive generation aimed to eliminate or mitigate violations of power flow or voltage constraints. Attempts are being made by researchers to develop accurate and robust procedures to determine corrective actions, including all possibilities available to system operators using different modeling approaches, such as expert systems or other artificial intelligence tools. The reliability indexes must reflect the constraint violations remaining after exhausting the possibilities regarding corrective actions.

3.3.6 Calculation of Reliability Indexes

Once the preceding five steps have been carried out, the system states selected through enumeration or simulation can be classified either in normal operative states or in problem states because the load is not fully satisfied or constraints are violated. For any specific class of adequacy deficiency, insufficient generation, or network violations, system-wide, component-, or node-related probabilistic measures such as probability, frequency, or duration can be calculated. The load flow analysis provides the results necessary to link these reliability parameters to the severity of the deficiency. For example, for a bulk supply point, probabilities can be calculated for states with load curtailment exceeding prespecified values.

4. ELECTRIC POWER INDUSTRY RESTRUCTURING AND RELIABILITY

Before successful restructuring in other industry branches led to significant changes in the electric power industry organization and structure, beginning in Chile in the early 1980s and then in Britain in 1988, the most common organizational forms of electricity supply were the state-owned monopoly and the monopoly franchise system with vertically integrated utilities. Reliability planning in these systems was done in similar ways. In state-owned monopoly systems, the supply duty is generally established by law, requiring that all customers are reliably supplied, with costs as low as possible. Because the whole system is owned by the state,

central planning, including generation and transmission, is the natural way to achieve high reliability while minimizing total costs. All necessary data are available to the power company and the problem is easily formulated as an optimization problem with total cost minimization as the objective function, subject to technical and reliability constraints. The reliability constraint is formulated as an inequality, requiring that a reliability index, typically loss-of-load probability or expected energy not supplied, does not exceed a prespecified limit.

A vertically integrated utility does its planning much in the same way as the state monopoly. In its geographically delimited concession region, the vertically integrated utility has to economically supply the customer demand subject to the same constraints as the state-owned monopoly. Reliability standards are set by the state through regulatory agencies and electricity prices to end consumers are regulated on the basis of audited costs and a guaranteed rate of return.

New generation technologies, notably high-efficiency gas-fueled combined cycle thermal plants, broke the notion that scale economies made electricity production a natural monopoly. This, combined with the fact that a regulated monopoly has little incentive to reduce its costs because any reduction is passed on to the consumers through adjustment of the regulated prices, and the success of deregulation in other industry branches, such as telecommunications and airlines, led to the movement toward restructuring in the power industry. On the other hand, there is a feeling that the high reliability standards used in the vertically integrated industry, state or private owned, have sometimes led to overcapacities that probably increase costs beyond the values consumers would be willing to pay for reliability.

In the new deregulated environment, there is no more central planning. Regarding generation investment, instead of central planning it is expected that economic signals will play a fundamental role in defining location, capacity, and time of investments. Concerning short-term operation, it is necessary to have a set of market rules such that reliability-related products can be commercialized in a more flexible and competitive way. It is expected that the adequacy of generation systems is a problem that can be properly addressed in a framework of competitive market rules, in both long-term planning and short-term operation.

Transmission systems are essentially different from generation systems due to the factual impossibility of direct competition. Because of the large investments required and the problems in finding new sites for high-voltage lines, transmission systems will continue to be seen as natural monopolies. How to assure reliability in the long term under the new industry environment is yet an open problem. Even though there is no consensus about the solution to this problem, currently there is an ongoing discussion about considering transmission as a sector that can be contemplated as a performance-based industry rather than a guaranteed rate-of-return industry. With this change in regulation plus the implementation of new transmission technologies such as flexible AC transmission systems (FACTS) and the acceptance of distributed generation, competition for some of the transmission reliability-associated products can be expected.

On the other hand, in the short term, for a given transmission network, the real questions are reserve requirement calculation and reserve allocation procedures. Adequacy problems derived from transmission bottlenecks or transmission failures lead to undesired supply interruptions and highly volatile prices. This problem is much more severe in cases in which demand plays a passive role. To be consistent with the aim of competition in power systems, procedures for reserve allocation should be able to consider explicitly reliability requirements submitted by the demand. Composite systems reliability analysis should be further developed on the basis of the fundamentals discussed here, with the aim to provide tools for proper reliability-related risk management for both system operator and market participants. The ISO needs reliability analysis tools to define the reserve requirements on the basis of the desired reliability level, and market participants need tools to be able to manage their exposure to dispatch risks for any kind of generation capacity. On the basis of this analysis, a market participant can optimize bids for energy and reserve on a portfolio basis.

SEE ALSO THE FOLLOWING ARTICLES

Electrical Energy and Power • Electric Power: Critical Infrastructure Protection • Electric Power Generation: Fossil Fuel • Electric Power Generation: Valuation of Environmental Costs • Electric Power Measurements and Variables • Electric Power Reform: Social and Environmental Issues • Electric Power Systems Engineering • Electric Power: Traditional Monopoly Franchise Regulation and Rate Making

Further Reading

Allan, R., and Billinton, R. (2000). Probabilistic assessment of power systems. *Proc. IEEE,* **88,** 140–162.

Alvarado, F. (2000). The relationship between reliability and market design: Where are we? "NSF/DOE/EPRI Workshop: Future Research Directions for Complex Interactive Electric Networks," Nov. 16–17. Dept. of Energy, Washington, D.C.

Billinton, R., and Allan, R. N. (1996). "Reliability Evaluation of Power Systems." 2nd Ed. Plenum, New York.

CIGRE Task Force 38.03.12. (1997). Power system security assessment: A position paper. *Electra* **175,** 49–78.

North American Electric Reliability Council. (1997). "Operating Policies and Planning Standards." Available on the Internet at http://www.nerc.com.

Electromagnetic Fields, Health Impacts of

DAVID L. MCCORMICK
IIT Research Institute
Chicago, Illinois, United States

1. **Introduction**
2. **General Toxicity Studies of EMF**
3. **Developmental and Reproductive Toxicity Studies of EMF**
4. **Immunotoxicity Studies of EMF**
5. **Genetic Toxicity Studies of EMF**
6. **Oncogenicity Studies of EMF**
7. **Conclusion**

Glossary

clastogenic Capable of inducing DNA strand breaks.
mutagenic Capable of inducing mutations in DNA.
oncogenic Capable of inducing a tumor or other neoplasm.
terata Birth defects.
teratogenesis Abnormal processes of fetal development leading to birth defects.

Efforts to identify and quantitate possible health hazards of exposure to power frequency (50/60 Hz) magnetic fields include both epidemiology studies of human populations and hazard identification and mechanistic studies performed in experimental model systems. The results of these studies demonstrate that magnetic field exposure induces no systemic toxicity in humans or in experimental animals. Similarly, magnetic field exposure has no adverse effects on reproduction or fetal development in humans or in animal models that are commonly used for hazard identification. Magnetic fields do not induce mutations or chromosome damage in standard genetic toxicology test systems, and extensive studies in a wide range of animal models provide no clear evidence that magnetic field exposure is oncogenic in any test system. Epidemiology studies suggest a small increase (relative risk ≤ 2.0) in the risk of leukemia in children exposed to magnetic fields; this increase is not seen in other sites in children, or in any organ site in adults. Because no compelling biological mechanism has been identified through which magnetic fields may induce neoplasia or other adverse health effects, further studies are required to determine whether magnetic fields are causally linked to childhood leukemia.

1. INTRODUCTION

Over the past two decades, the possible relationship between human exposure to power frequency (50 and 60 Hz) electromagnetic fields (EMF) and adverse health outcomes has generated significant interest in both the scientific community and the general population. In the late 1970s and early 1980s, several epidemiology studies were published whose data suggested that adverse health effects may be seen in human populations receiving either occupational or environmental (i.e., residential) EMF exposures. Health effects associated with EMF in these studies included reproductive dysfunction, increased risk of birth defects and other developmental anomalies, and increased risk of cancer in several sites. These suggestive epidemiology data stimulated a large number of additional epidemiologic and experimental investigations, whose goals are to (1) confirm or refute the existence of specific health hazards of EMF exposure and (2) where such hazards are demonstrated to exist, identify EMF flux densities at which adverse health effects are observed, in order to quantify the risks of human exposure to EMF.

On the basis of the type of system studied, scientific investigations into the possible health

 © 2004 Elsevier Inc. All rights reserved.

effects of EMF can be divided into three general categories:

1. Epidemiology studies, in which the risk of a disease is compared in human populations with different levels of EMF exposure
2. *In vivo* experimental studies (animal bioassays), in which the risk of disease is compared in groups of experimental animals with different levels of EMF exposure
3. Mechanistic studies, in which the effects of EMF exposure on biochemical, molecular, or other disease-associated end points are evaluated using experimental animals or *in vitro* model systems

Each category has important strengths and limitations that should be considered in the overall interpretation of EMF bioeffects data.

1.1 Strengths and Limitations of Epidemiology Studies of EMF

Clearly, the most important strength of epidemiology studies is that they are investigations of human responses to EMF exposure. As such, no interspecies extrapolations are required, and the EMF flux densities being investigated in epidemiology studies are, by definition, "real world." For these reasons, any observation of a strong positive association between EMF exposure and risk of a specific disease would provide compelling evidence of a relationship between the exposure and that adverse health effect. Conversely, however, epidemiology studies are limited in that they can only demonstrate associations between an exposure and disease risk. Epidemiology cannot provide conclusive demonstrations of causality: spurious associations may be identified in epidemiology studies, and weak causal associations may be missed due to limited statistical power.

A critical factor that must be assessed in the interpretation of EMF epidemiology studies is the accuracy and relevance of the exposure metric on which comparisons are based: Is the estimate of EMF exposure developed using (in order of decreasing desirability)

- Measurements of EMF flux densities to which an individual is actually exposed, as obtained by a personal dosimeter?
- Measurements of EMF flux densities present at a relevant location in the home or work site?
- Measurements of EMF flux densities made at a constant point outside of the home ("front door" measurements)?
- An imputed or calculated level of EMF exposure that is based on job classification, residential wire code, residential distance from a power line or substation, or other metric?

It has been demonstrated that EMF exposures that are measured or imputed using different strategies can vary substantially; more specifically, important criticisms have been raised concerning the accuracy of residential wire codes that have been used widely as a surrogate for residential EMF exposure levels and job title classifications that have been used as an estimator of occupational EMF exposures.

A parallel consideration is the temporal variability of EMF measurements: Does today's measurement of an EMF flux density in a house or other location provide an accurate estimate of the magnetic field environment that was present in the same location 5 years ago, 10 years ago, or at some other earlier time that is relevant to the etiology of the disease in question? Because EMF flux densities in any specific location can demonstrate considerable temporal variability, the ability of such measurements to support retrospective epidemiology studies remains an important issue.

1.2 Strengths and Limitations of Experimental Studies of EMF

The strengths and weaknesses of studies of the effects of EMF on disease induction in experimental animals are essentially opposite of those seen in epidemiology studies. Because animal bioassays are performed in a nonhuman test system, interspecies extrapolation of study data is required. In addition to possible species differences in response, field strengths studied in animal bioassays often greatly exceed those seen in human exposures and may thereby induce artifactual responses by overwhelming a physiologic system. These two factors provide the most important limitations to the use of experimental animal data to assess the potential risk of human exposure to EMF or any other exogenous agent.

Conversely, however, several factors provide important support for the use of animal models in EMF hazard assessment. A number of animal model systems that have been used in studies of EMF toxicology are also used widely in the safety assessment of pharmaceuticals, environmental chemicals, and other agents to which humans receive broad population exposure. These model systems are well-characterized, have extensive historical databases, and have been shown to be generally

predictive of human responses. Animal studies also provide the opportunity to conduct bioassays under precisely controlled and monitored exposure conditions and to incorporate statistical considerations into study designs. In consideration of the difficulties associated with EMF exposure assessment in epidemiology studies, the ability to conduct animal studies under rigorously controlled exposure conditions is a major strength of this approach to hazard identification. Finally, through the use of genetically engineered animal models, possible effects of EMF on susceptible subpopulations can be investigated. Studies in such susceptible subpopulations may identify hazards that cannot be identified in normal animals or humans, but that may substantially increase the risk of disease in individuals with a genetic predisposition, preexisting condition, or other factor that increases disease risk.

1.3 Role of Mechanistic Studies in EMF Hazard Assessment

The results of mechanistic investigations performed in experimental animals or *in vitro* model systems provide critical data that can establish the biological framework in which observed health effects of EMF can be interpreted. Although identification of a mechanism of action for a toxic agent is not essential for hazard identification, elucidation of one or more biochemical, molecular, or cellular pathways through which an agent exerts biological effects will clearly strengthen hazard assessments developed on the basis of epidemiology and animal bioassays.

Conversely, should the results of epidemiology studies and animal bioassays be negative or conflicting, the identification of a clearly definable mechanism of action can both (1) provide a rational biological context for hazard identification and (2) support the design of optimized hazard identification studies. However, should the results of epidemiology studies and animal bioassays be negative, the lack of an identified biological mechanism of action further weakens the assertion that a substantive human health hazard exists.

2. GENERAL TOXICITY STUDIES OF EMF

In consideration of the absence of documented case reports of systemic EMF toxicity, it is generally agreed that acute, subchronic, or chronic exposure to EMF causes little or no systemic or generalized toxicity in humans. Because certain occupational environments contain "high field" areas in which EMF flux densities exceed those of residential environmental by tens to hundreds of fold, it is logical to expect that such high field environments would be the primary sites of any observed EMF toxicities. However, no case reports of EMF-associated generalized toxicity or malaise in any exposed worker have been published.

The lack of systemic toxicity of EMF exposure is further supported by the results of acute and subchronic animal bioassays conducted in rodent models. Although occasional instances of statistically significant differences from sham controls have been reported, no clear pattern of EMF toxicity identifiable by effects on animal survival, clinical observation, body weight, food consumption, hematology, clinical chemistry, or histopathology has been reported.

As such, no evidence of generalized toxicity associated with EMF exposure has been identified in studies of either humans or experimental animals. On this basis, it can be concluded that EMF exposure presents little or no hazard of systemic acute, subchronic, or chronic toxicity.

3. DEVELOPMENTAL AND REPRODUCTIVE TOXICITY STUDIES OF EMF

Investigations into the possible developmental and reproductive toxicity of EMF were stimulated by early reports of teratogenesis in chick embryos exposed to high EMF flux densities. Although the processes of prenatal development differ substantially in chickens and in humans, the induction of terata in early chick embryos has been interpreted by some investigators as supporting the hypothesis that EMF is a teratogen. By contrast, however, teratologists who use animal models to develop human hazard assessments for chemical and physical agents have questioned the value of the chick embryo model for human hazard identification.

The early teratology data generated in chicken models were extended by reports of small increases in the incidence of birth defects and spontaneous abortions in women who used electric blankets and electric waterbed heaters during pregnancy. These early findings were hypothesis generating, but were limited in that they were retrospective studies that

included no measurements of EMF exposure. Furthermore, the design of these studies essentially precluded the separation of possible EMF effects from possible thermal effects associated with the use of an electric blanket or bed heater.

Since the publication of these early findings, a large number of epidemiology studies have been performed to investigate the hypothesis that EMF exposure is a risk factor for teratogenesis. These studies have examined the incidence of birth defects, growth retardation, and premature delivery, among other end points, in the offspring of women who used electric blankets or electric bed heaters during pregnancy. Other studies have examined rates of these abnormalities in pregnant women with different levels of residential EMF exposure and in pregnant women with different measured or imputed levels of occupational EMF exposure. In general, although occasional EMF-associated differences have been reported, the results of these studies failed to confirm the hypothesis that EMF exposure during pregnancy is a significant risk to the developing fetus. In most studies, no statistically significant relationships were found between EMF exposure and the risk of birth defects or other developmental abnormalities, regardless of whether the increased EMF exposure resulted from the use of an electric blanket or bed heater, from a higher level of EMF in the residential environment, or from EMF exposures that occurred as a result of employment.

Supporting these epidemiologic data are the results of more than a dozen teratology studies that have been conducted in animal model systems. In this regard, it is essential to note that the rodent models and study designs used in the EMF teratology studies are used routinely to evaluate the possible teratogenicity of pharmaceuticals and environmental chemicals; these models have been validated extensively and have been shown to be generally predictive of human responses. The results of these studies, which have been conducted by several different teams of investigators in several countries, have been uniformly negative. When the negative results generated in these experimental evaluations of EMF teratogenesis are considered together with the negative results of most epidemiology studies, the overwhelming majority of available teratology data do not support the hypothesis that EMF exposure during pregnancy constitutes an important hazard to the developing fetus.

As was the case with teratology studies, no body of evidence exists to implicate EMF as a significant risk for fertility. Although the total number of available studies is small, no evidence of reduced fertility has been reported in either men or women receiving occupational exposure to high levels of EMF. Although the number of experimental studies of EMF and reproductive function is similarly small, animal studies of male and female fertility are similarly negative. These studies, which commonly involve exposure of several generations of rodents to the test agent, have failed to identify any decreases in fertility in either male or female animals exposed to EMF.

4. IMMUNOTOXICITY STUDIES OF EMF

Evaluation of the possible effects of EMF exposure on immune function is important to support our understanding of the possible role of EMF in carcinogenesis and as a modifier of host resistance to infection. Any reproducible down-regulation of immune function in EMF-exposed animals could have important consequences in terms of the host's ability to resist infectious disease challenge or to provide effective immune surveillance against neoplastic cells.

The body of studies in which the influence of EMF on immune function parameters has been evaluated is small and is almost entirely limited to studies conducted using rodent model systems. In these studies, a battery of immune function assays (lymphoid organ weight, lymphoid organ cellularity, lymphoid organ histology, B- and T-cell mediated immune responses, NK cell function, lymphocyte subset analysis, host resistance assays) was performed on mice and/or rats exposed to magnetic fields for period ranging from 4 weeks to 3 months. In these studies, no alterations were found in lymphoid organ weights, lymphoid organ cellularity, or lymphoid organ histology; similarly, no deficits in T-cell function or B-cell function were identified, and lymphocyte subset analysis did not reveal any effects of EMF exposure. Interestingly, however, a statistically significant suppression of NK cell function was seen in two studies performed in female mice; this alteration was not seen in male mice of the same strain, nor in either sex of rat evaluated in a parallel study.

Although the authors conclude that this suppression of NK cell function is real, it appears to have little or no biological significance. The primary role of NK cell appears to be in immune surveillance against cancer cells. However, completely negative results were obtained in a 2-year carcinogenesis bioassay of EMF performed in the same strain of mice that demonstrated the reduction in NK cell

function. Because the deficit in NK cell function was not associated with an increase in cancer risk, it is concluded that either (1) the quantitative reduction (~30%) in NK cell function induced by EMF is too small to be biologically significant or (2) other arms of the host immune system were able to perform required immune surveillance in these animals.

5. GENETIC TOXICITY STUDIES OF EMF

The possible activity of EMF exposure in the induction of genetic mutations, DNA breaks, and other types of DNA damage is a key mechanistic consideration that underlies the interpretation of data relevant to the possible activity of EMF as a carcinogen. Although nonmutational mechanisms may also be responsible for carcinogenic activity, mutation and DNA damage are a common mechanism of action for carcinogens. As such, any evidence of genetic toxicity for EMF would have substantial implications for its potential activity as a risk factor for human cancer.

Numerous studies have been performed over the past 20 years to investigate the hypothesis that EMF exposure induces mutations or other genetic damage. Studies to evaluate the possible genetic toxicity of EMF have been performed using a broad range of experimental model systems, including *in vitro* bacterial mutagenesis assay systems (such as the Ames test), *in vitro* mutagenesis and DNA damage test systems using mammalian cells, and *in vivo* assays of DNA damage. A substantial number of the *in vitro* genetic toxicology studies of EMF have been performed using test systems that are used widely in the safety assessment of pharmaceuticals and chemical agents. As such, a large database exists to support the utility of these assays to predict *in vivo* responses.

The results of the overwhelming majority of genetic toxicology studies published in the peer-reviewed literature have been negative; as such, a broad assessment of the genetic toxicology database does not support the hypothesis that EMF is genotoxic in either bacteria or in mammalian cells. Virtually all genetic toxicology studies that have been performed using what are considered to be "standard" genetic toxicology test systems (bacterial cell mutagenesis [Ames test], mammalian cell mutagenesis, micronucleus assays) have generated negative data; the quantity of negative genetic toxicology data for EMF in these test systems would ordinarily be sufficient for a federal regulatory agency to consider EMF to be nonmutagenic.

It should be noted that not all genetic toxicology studies of EMF have been negative. Positive evidence of genetic toxicity has been obtained under certain very limited experimental conditions (i.e., electric shocks and spark discharges) associated with power frequency electric fields; however, these specialized situations are not encountered in most residential or occupational environments and therefore appear to be a little relevance to overall EMF hazard assessment. Most other positive reports have involved studies in nonstandard experimental systems whose relevance to human hazard identification is unclear. A few positive results have been obtained in studies using test systems in relatively wide use in the genetic toxicology community; however, efforts to replicate the positive data have failed.

Aside from specialized situations that are not encountered in routine residential or occupational environments, there is no consistent body of experimental evidence that demonstrates that either power frequency magnetic fields or a combination of power frequency magnetic + electric fields (nonshocking and nonsparking) induce mutations, DNA breaks, or other DNA damage in any standard genetic toxicology test system. On this basis, it is generally accepted within the scientific community that power frequency EMF is not genotoxic.

6. ONCOGENICITY STUDIES OF EMF

Although other possible health effects of residential or occupational EMF exposure are clearly of concern, the primary focus of most EMF hazard assessment studies has been to determine whether exposure to EMF increases the risk of cancer or other neoplasms. The majority of epidemiologic studies of EMF published by the end of the 20th century have been focused on cancer end points. Similarly, a large fraction of experimental resources dedicated to studies of EMF action have been directed toward the question of cancer hazard assessment.

6.1 Epidemiology Studies of EMF and Cancer

Although the results of a number of individual epidemiology studies have provided suggestive evidence of the potential oncogenicity of EMF, the

total body of epidemiologic data linking EMF exposure and cancer risk is by no means conclusive. The results of several EMF epidemiology studies published in the late 1970s and early 1980s suggested that EMF exposure may be a risk factor for cancer of the brain, breast, and hematopoietic system. However, other epidemiologic studies of similar populations failed to confirm these findings, and no clear picture has emerged from this body of literature.

Since 1990, nearly 100 epidemiology studies and meta-analyses designed to investigate the possible association between occupational or residential exposure to magnetic fields and cancer risk have been published. The experimental methods used in these studies are often substantially improved in comparison to methods used in earlier investigations; specific areas of improvement include the use of larger sample sizes and substantial improvements in EMF exposure assessment. However, despite these improvements in epidemiologic methods, EMF cancer epidemiology studies continue to generate both positive and negative results. When considered together, the results of these studies are insufficient to either clearly support or refute the hypothesis that exposure to EMF is a significant risk factor for human cancer in any organ system.

The most comprehensive evaluation of the possible carcinogenicity of power frequency magnetic fields was that published by a working group assembled by the International Agency for Research on Cancer (IARC). The conclusion of this group, which reviewed essentially every epidemiology study of EMF and cancer that has been published in a peer-reviewed journal, was that limited evidence exists to support the hypothesis that EMF exposure is a risk factor for childhood leukemia. The increased risk of childhood leukemia associated with EMF exposure is small (relative risk of <2.0), but is a generally consistent finding across a relatively large number of epidemiology studies. By contrast, however, the conclusion of this group was that the sum of the epidemiology data is inadequate to support the hypothesis that EMF exposure is a significant risk factor for leukemia or any other type of neoplasm in adults and is inadequate to support the hypothesis that EMF exposure is a significant risk factor for any type of malignancy other than leukemia in children. As such, the IARC panel concluded that suggestive evidence of EMF oncogenicity exists only for childhood leukemia and not for any other neoplasm in either children or adults.

Because the increased risk of childhood leukemia in EMF-exposed groups is small, epidemiology studies are continuing in order to quantify more precisely the risks involved, and to confirm that this observed elevation in leukemia risk does, in fact, result from EMF exposure and is not the result of some confounding factor.

6.2 Experimental Studies of EMF and Cancer

In situations where epidemiology does not support the conclusive identification and quantitation of the potential risks associated with exposure to an environmental agent, laboratory studies conducted in appropriate experimental model systems increase in importance. Well-designed and controlled animal studies permit evaluation of biological effects *in vivo* under tightly controlled exposure and environmental conditions, as well as in the absence of potential confounding variables. In consideration of the conflicting results of EMF epidemiology studies and difficulties associated with exposure assessment in such studies, animal studies may provide the best opportunity to identify effects of EMF exposure that could translate into human health hazards.

6.2.1 Assessment of the Possible Oncogenic Activity of EMF as a Single Agent

Several teams of investigators have published the results of studies that are designed to investigate the possible oncogenic activity of chronic exposure to EMF alone (as opposed to EMF exposure in combination with simultaneous or sequential exposure to other agents). These studies have generally been conducted using the chronic rodent bioassay design, a standardized study design in which animals are exposed to the test agent for the majority of their natural life span (2 years), followed by a complete microscopic evaluation of approximately 50 tissues from every animal to identify induced cancers and other lesions. This study design has been used widely to evaluate the potential carcinogenicity of new drugs, chemicals, and environmental agents; the IARC and the United States National Toxicology Program (NTP) have both performed statistical analyses demonstrating the utility of the chronic rodent bioassay as a predictor of human carcinogenicity. As a result, it is generally accepted among toxicologists that the chronic bioassay in rodents currently provides the best available experimental

approach to identify agents that may be carcinogenic in humans.

In all studies conducted using this study design, chronic exposure to power frequency EMF had no effect on the incidence of neoplasia in any of the three putative targets for EMF action (brain, breast, and hematopoietic system) that had been identified by epidemiology. Furthermore, chronic exposure to EMF did not induce any pattern of biologically significant increases in cancer incidence in other tissues. The very close comparability of the results of chronic rodent oncogenicity bioassays conducted independently in three laboratories strengthens the conclusion that exposure to magnetic fields does not induce cancer in rodents. On the basis of the demonstrated predictive nature of the rodent 2-year bioassay, these findings do not support the hypothesis that EMF is a significant risk factor in the etiology of human neoplasia.

6.2.2 Site-Specific Rodent Oncogenicity Studies

It is generally accepted that exposure to power frequency EMF does not induce mutations or other genetic damage that is identifiable using standard genetic toxicology model systems. Because EMF is apparently not genotoxic, it is logical to conclude that any oncogenic effects of exposure are mediated through co-carcinogenic or tumor promotion mechanisms. To address this possibility, a relatively large number of *in vivo* studies have been conducted to identify possible co-carcinogenic or tumor promoting effects of EMF in various target tissues. The design of these studies involves simultaneous or sequential exposure to EMF in combination with another agent (generally a chemical carcinogen or ionizing radiation) and is focused on neoplastic development in a specific organ. Most multistage tumorigenesis studies of EMF have been conducted using animal models that are designed as research tools to study site-specific carcinogenesis. As a result, the utility of these assays to predict human oncogenicity is less completely understood than are the results of the chronic oncogenicity bioassay.

Studies of the effects of EMF on cancer induction in multistage models for cancer of the brain, hematopoietic system, and liver have been uniformly negative: in all cases, exposure to an inducing agent + EMF induced no more neoplasms than were induced by exposure to the inducing agent alone. Conflicting data have been presented for studies of the effects of EMF exposure on the induction of skin and breast cancers in multistage animal models: in each case, positive data have been presented by one laboratory, but independent studies conducted in one or more additional laboratories have failed to confirm the original finding.

Overall, the peer-reviewed literature provides little compelling evidence that exposure to EMF has co-carcinogenic or tumor promoting activity in any organ site in experimental animals. Completely negative results were found in studies conducted using models for cancer in two of the three target organs (brain and hematopoietic system) identified in epidemiology studies as possible targets for EMF action; data in the third putative target tissue (breast) are conflicting. These data generally support the conclusions from the 2-year oncogenicity bioassays and do not support the hypothesis that EMF exposure is a key risk factor for cancer induction in these target tissues.

6.2.3 Oncogenicity Studies in Genetically Engineered Animals

Insights into the molecular mechanisms of disease have provided the scientific basis for the development and use of genetically modified animals to study disease processes and the application of these models in hazard identification. The science supporting the use of genetically engineered animals in safety assessment stems from initial observations that the incidence of malignancy is increased or its latency is decreased in transgenic animal strains into whose germ line a tumor-associated gene (oncogene) has been inserted and in knockout animal strains from whose germ line a tumor suppressor gene has been deleted. Genetically engineered animals demonstrate a predisposition to specific diseases and thereby provide an approach to hazard identification that may identify effects that occur in sensitive subpopulations but rarely or not at all in the general population. As a result, genetically modified models may demonstrate sensitivity to EMF bioeffects that do not occur in standard animal strains.

Although the total number of completed oncogenicity evaluations of EMF in genetically engineered animals is small, the results of published studies are consistent. In three studies conducted in the United States and in Australia, EMF exposure had no effect on the induction of hematopoietic neoplasia in one transgenic mouse strain (two independent studies) and in one knockout mouse strain. Because EMF exposure did not increase the incidence of hematopoietic neoplasia in any experimental group, the data from these genetically engineered mouse models can be considered to confirm and extend the negative

oncogenicity data generated in the 2-year rodent oncogenicity bioasssays of EMF.

6.3 Possible Mechanisms of EMF Action in Tumorigenesis

Over the past two decades, researchers have proposed a number of possible mechanisms though which exposure to EMF may act to induce or stimulate carcinogenesis; these proposed mechanisms include a wide range of potential effects at the molecular, cellular, and whole organism levels. Review of the published scientific literature provides no clear support for any specific mechanism of EMF action in carcinogenesis. Furthermore, no reproducible body of scientific evidence exists to link EMF exposure to alterations in any biological end point that is mechanistically related to neoplastic development in any tissue.

The majority of hypothesized mechanisms of EMF action in carcinogenesis can be grouped into five general types of effects:

1. Genetic toxicity (DNA damage, clastogenesis or mutagenesis)
2. Alterations in gene expression
3. Biochemical changes associated with regulation of cell proliferation
4. Alterations in immune function
5. Alterations in melatonin levels or action

These general classifications are necessarily arbitrary and were developed to simplify the purposes of this review. It is clear that several of these mechanisms overlap and that not all possible mechanisms of EMF action are included in these five groups. However, these groups do encompass the vast majority of mechanisms proposed for EMF bioeffects and do include mechanisms that have received the most extensive study in order to identify possible linkages between EMF exposure and carcinogenesis.

6.3.1 Genetic Toxicity

The possible genotoxicity of EMF exposure was reviewed in Section 5. As indicated in that section, numerous well-designed, complete, reproducible, and apparently well-conducted studies have failed to demonstrate any genotoxic effects of EMF.

It is generally agreed that ELF magnetic fields do not possess sufficient energy to damage DNA via direct interaction. Although it is possible that power frequency EMF could damage DNA through an indirect mechanism (i.e., via generation of free radicals), an overwhelming majority of genetic toxicology studies conducted in standard safety assessment models have failed to identify any significant mutagenic or clastogenic effects of EMF exposure. As such, a consensus has emerged in the scientific community that exposure to power frequency EMF presents little to no risk of genetic toxicity.

6.3.2 Alterations in Gene Expression

The possible influence of EMF exposure on the expression of cancer-related genes is one of the most controversial areas of research in bioelectromagnetics. Over the past decade, several investigators have reported increases in the expression of one or more protooncogenes (generally c-*myc* or c-*fos*) in cell cultures exposed to 60 Hz EMF. However, replication studies conducted by numerous other investigators have not confirmed these findings, and other, more broadly based screening efforts to identify genes that are up- or down-regulated by EMF exposure have generally yielded negative results. On this basis, the robustness of the reports of altered gene expression, as well as the generality of those findings, must be questioned.

Given the huge universe of genetic targets in the mammalian cell, it is considered likely that exposure to 60 Hz EMF may alter the expression of one or more genes. It is important to note, however, that a finding of enhanced expression of an oncogene or any other gene is not, by itself, sufficient evidence to suggest physiologic or pathophysiologic activity. Because many cellular pathways exhibit biological redundancy, alterations in one pathway may be compensated for by an antagonistic change in another pathway. For this reason, although findings of altered gene expression may be considered to suggest a biologically important effect, demonstration of that effect is necessary to establish the biological relevance of the observed alteration in gene expression. Lacking the demonstration of any physiological/pathophysiological effect, an isolated finding of altered gene expression cannot be construed to be biologically significant. Restated with emphasis on the process of carcinogenesis, a finding of altered expression of an oncogene that is associated with EMF exposure can by no means be equated with any activity of EMF as a carcinogen or tumor promoter.

6.3.3 Biochemical Changes Associated with Regulation of Cell Proliferation

Changes in the expression or activity of enzymes and protein products involved in cell proliferation, cell

cycle regulation, and apoptosis would be expected if EMF exposure was significantly associated with either complete carcinogenesis or tumor promotion. Several reports of this type of EMF bioeffect (e.g., alterations in the activity of ornithine decarboxylase) have been presented in the literature. However, these reports generally demonstrate little consistency across investigators or experimental model systems; as such, the results of individual studies do not appear to be supportable as a general mechanism for EMF activity.

As discussed earlier with regard to possible alterations in gene expression, an isolated finding of changes in enzyme activity or other pathway in cells that are exposed to EMF can be considered to be no more than suggestive of a possible adverse outcome. Linkage of these biochemical or cell-based effects to the physiological or pathophysiological effect is essential to demonstrate their biological significance.

6.3.4 Alterations in Immune Function

Down-regulation of the host immune response by EMF is a tempting hypothesis through which EMF exposure could be associated with risk of carcinogenesis. However, as reviewed in Section 4, this hypothesis is supported by little experimental evidence: with the exception of an apparently biologically insignificant effect on NK cell function, the results of all studies of EMF effects using standardized immunotoxicology test systems and end points have failed to demonstrate effects of EMF on any arm of the host immune response. Although the hypothesis that EMF exposure stimulates tumorigenesis through an indirect mechanism involving suppression of host immunity is intellectually appealing, it is unsupported by any substantial body of experimental evidence.

6.3.5 Alterations in Melatonin Levels or Action

The "melatonin hypothesis" is perhaps the most widely accepted potential mechanism for the biological effects of EMF exposure. A simplified version of this hypothesis states that EMF bioeffects are mediated through alterations in the function of the pineal gland; these changes in pineal function are then responsible for alterations in patterns of melatonin synthesis, kinetics, or action. Ultimately, reductions in melatonin synthesis may have a number of potential adverse effects on the host, including reduced activity as an endogenous free radical scavenger, and influences on the hypothalamic-pituitary-gonadal axis.

A considerable body of evidence to support the melatonin hypothesis was generated in bioelectromagnetics studies performed during the 1980s and early 1990s; in these studies, reductions in circulating melatonin levels or time shifts in the circadian rhythm of melatonin biosynthesis were reported in both experimental animals and in humans. However, several leading investigators who previously reported significant effects of EMF exposure on pineal function have been unable to replicate their own data. This failure of replication is often associated with the acquisition of new, more sophisticated EMF exposure and monitoring equipment. Furthermore, the majority of papers from other laboratories have failed to demonstrate any consistent pattern of effects of EMF exposure on melatonin synthesis or kinetics in humans or in rodent model systems.

One notable exception to this trend is a series of studies that have been conducted using the Siberian hamster as a model system; EMF effects on melatonin secretion in this species appear to be a reproducible finding. However, this animal demonstrates a seasonal circadian rhythm that is considerably different from that seen in either humans or rodent species that are commonly used for safety assessments; as such, a relationship between this finding and possible human health effects is difficult to establish.

Concomitant with the weakening of the data set relating EMF exposure to alterations in pineal function, the melatonin hypothesis has been revised to suggest that EMF influences on melatonin may be effected through alteration in melatonin effects at the level of the target cell, rather than through effects on its synthesis or secretion. The database supporting this hypothesis is very limited and is considered insufficient to provide strong support for this potential mechanism of EMF action.

7. CONCLUSION

With one possibly notable exception, the large body of epidemiologic and experimental data generated over the past 20-plus years generally fails to support the hypothesis that exposure to power frequency EMF is associated with adverse health effects in either humans or experimental animals. No substantive evidence of generalized toxicity resulting from acute, subchronic, or chronic exposure to EMF has been reported in either humans or animal model systems. EMF exposure appears to have no biological significant effects on host immunity and does not impair fertility in either sex. Although some

conflicting data have been generated, EMF exposure does not appear to be a significant risk factor for the developing fetus.

Those data notwithstanding, however, the primary consideration in most hazard assessments is whether the agent being evaluated causes cancer. EMF exposure is not genotoxic, and the results of oncogenicity evaluations conducted using standard animal models, multistage animal models, and genetically engineered animal models are overwhelmingly negative. Similarly, the results of epidemiology studies provide no clear evidence of EMF oncogenicity in any organ site in adults. However, a small, but consistent increase in the risk of childhood leukemia has been observed in a substantial number of EMF epidemiology studies; the consistent nature of this finding was considered to be sufficient for the IARC Working Group to conclude that "limited evidence" of activity exists for EMF in the induction of childhood leukemia.

In summary, epidemiologic evidence provides limited evidence that EMF exposure is associated with an increased risk of childhood leukemia; by contrast, epidemiology studies have not identified increases in cancer risk for any other site in children or for any tissue in adults. No compelling body of epidemiologic or experimental evidence exists to demonstrate any other toxicologic effect of exposure to power frequency EMF, and no substantive biological mechanism has been identified for EMF action. Further epidemiology studies will be required to characterize the relationship between EMF and childhood leukemia more completely.

SEE ALSO THE FOLLOWING ARTICLES

Air Pollution, Health Effects of • *Climate Change and Public Health: Emerging Infectious Diseases* • *Electricity, Environmental Impacts of* • *Electric Power Generation: Valuation of Environmental Costs* • *Electromagnetism* • *Radiation, Risks and Health Impacts of*

Further Reading

International Agency for Research on Cancer (2003). IARC Monographs on the Evaluation of Carcinogenic Risk to Humans, Volume 80. Non-Ionizing Radiation, Part 1: Static and Extremely Low-Frequency (ELF) Electric and Magnetic Fields. Lyon, France.

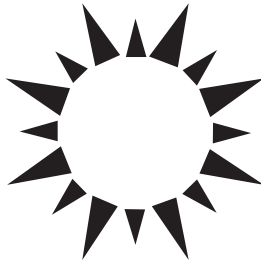
McCann, J., Dietrich, F., and Rafferty, C. (1998). The genotoxic potential of electric and magnetic fields: An update. *Mutation Research* **411,** 45–86.

McCann, J., Dietrich, F., Rafferty, C., and Martin, A. O. (1993). A critical review of the genotoxic potential of electric and magnetic fields. *Mutation Research* **297,** 61–95.

Portier, C. J., and Wolfe, M. S. (eds.). (1998). "Assessment of Health Effects from Exposure to Power-Line Frequency Electric and Magnetic Fields." National Institute of Environmental Health Sciences, Research Triangle Park, NC.

Electromagnetism

DANIEL R. STUMP
Michigan State University
East Lansing, Michigan, United States



1. Introduction
2. Electric Charge and Current
3. The Field Concept
4. Electromagnetic Induction
5. Applications of Electromagnetic Induction
6. The Maxwell Equations of the Electromagnetic Field
7. Electromagnetic Waves
8. Summary

Glossary

electric charge A property of the elementary particles of matter that gives rise to electric and magnetic interactions; electrons have negative charge and protons have positive charge.
electric current Net flow of electric charge due to the motion of charged particles in a conducting material.
electric field Lines of force produced by an electric charge and exerting a force on other electric charges.
electromagnetic induction Phenomena in which a change of magnetic field induces an electric field or electromotive force (EMF); the basis of the electric generator.
magnetic field Lines of force produced by an electric current or ferromagnet and exerting a force on other electric currents or on the poles of a magnet.
Maxwell's equations A set of four equations that describe the electric and magnetic fields as functions of position and time.

Electromagnetism is the science of electric and magnetic fields and of the interactions of these fields with electric charges and currents, and as such it is one of the fundamental interactions of nature. Its physical origin lies in a property possessed by elementary particles of matter—electrons and protons—called electric charge. The electromagnetic interaction also governs light and other forms of radiation. Electricity and magnetism appear in many natural phenomena and are applied in many inventions of technology and everyday life. This article describes the basic science and applications of electromagnetism.

1. INTRODUCTION

The earliest discoveries of electric and magnetic forces were made by philosophers of ancient Greece. They observed that when amber is rubbed with animal fur, it acquires the ability to attract small bits of reed or feathers. This small effect was the first observation of static electricity. (The Greek word for amber, $\varepsilon\lambda\varepsilon\kappa\tau\rho o\nu$, is the origin of our word "electric.") They also observed that a lodestone exerts a force on iron—the first example of magnetism. (The Greek province of Magnesia, where the iron ore magnetite occurs naturally, is the origin of our word "magnetic.") These discoveries of weak and mysterious forces were the first steps toward our scientific understanding of electromagnetism. Today science has dispelled much of the mystery of electricity and magnetism, and created technological power beyond the dreams of the ancient philosophers.

2. ELECTRIC CHARGE AND CURRENT

2.1 Electrostatics and Electric Charge

Electric charge is a property of matter. At the most basic level, the constituents of atoms are charged particles—electrons with negative charge ($-e$) and protons with positive charge ($+e$). An atom has equal numbers of electrons and protons so it is electrically neutral. However, a sample of matter becomes electrically charged if the balance between electrons and protons is broken. For example, when amber is rubbed with fur, electrons are transferred from the fur to the amber; the amber then has net negative charge.

 © 2004 Elsevier Inc. All rights reserved.

Like charges repel and unlike charges attract. That is, if two samples of matter have net positive charge (or both have net negative charge), they exert equal but opposite repulsive forces on one another. If the samples have unlike charges, one positive and the other negative, then each exerts an attractive force on the other. The strength F of the electric force **F** was measured accurately by Charles Augustin de Coulomb (Vectors are indicated in boldface, scalars in plain face). The force is proportional to the product of the charges, q_1 and q_2 and inversely proportional to the square of the distance of separation r,

$$F = \frac{kq_1q_2}{r^2} \quad \text{where} \quad k = 8.99 \times 10^9\ \mathrm{Nm^2/C^2}.$$

This simple mathematical formula has been tested to extreme precision. The dimensional units for the constant k are $\mathrm{Nm^2/C^2}$, where N = newton = unit of force, m = meter = unit of length, and C = coulomb = unit of charge. The proton charge is $e = 1.602 \times 10^{-19}$ C. It forms part of the foundation for the theory of electromagnetism.

2.2 Magnetostatics and Electric Current

Everyone has observed magnets and their forces, which can be quite strong even in small magnets. Every magnet has polarity—north and south poles. Two magnets repel if their north poles approach one another and repel if the south poles approach; but they attract if north approaches south. Magnets have a special attraction to iron; if either pole is placed near an iron object, a magnetic force pulls the magnet and the object toward one another. Science has identified the origins of magnetic forces. Technologies based on magnetic forces are used every day throughout the world.

Magnetism is very closely connected to the electric charge of subatomic particles. The most familiar example of a magnet is a ferromagnet—a piece of magnetized iron. However, a ferromagnet is not the only source of magnetism nor even the most basic. Electric currents also produce magnetic forces, and in some ways the magnetic field associated with electric current is the most basic form of magnetism. The term magnetic field used in this section refers to any magnetic effect. The more technical meaning of the term is explained in Section 3.

2.2.1 Electric Current as a Source of Magnetic Field

An electric current is a stream of electrically charged particles (of one sign) moving in the same direction. The current may be constant in time (DC, or direct current), oscillating in time with a constant frequency (AC, or alternating current), or varying in time. Currents can exist in metals and in several other forms of conducting matter. In a metal, some electrons occupy states that extend over large distances (i.e., not bound to a single atomic core). If an electric force is applied, then these conduction electrons will move in response, creating an electric current. (Ohm's law, $V = IR$ where V = potential difference in volts, I = current in amps, and R = resistance in ohms, expresses quantitatively the relation between the electric force and the current.) In metals, the positively charged atomic nuclei are fixed in a crystalline lattice, so the electric current is due to motion of electrons.

The first observation of the magnetic field associated with an electric current was an accidental discovery by Hans Christian Oersted during a public lecture in 1819. The current existed in a metal wire connected across a battery. Oersted noticed that a nearby compass needle deflected while the current was flowing. The strength of the magnetic field at a distance of 1 centimeter from a 1 ampere current is 2×10^{-5} tesla, comparable to the earth's magnetic field of approximately 5×10^{-5} tesla.

Oersted's discovery was studied in detail by Jean Marie Ampère. The magnetic field can be determined quantitatively by measuring the force on a pole of a magnet or the torque on a compass needle. A compass needle in equilibrium points in the direction of the field vector at the position of the compass. Therefore, a compass can be used to map the field directions. Ampère found that the field direction "curls around" the current. Figure 1 shows a segment of current-carrying wire and the associated magnetic field encircling the wire. The field directions follow the right-hand rule: with the thumb of your right hand pointing along the wire in the direction of the current, the fingers naturally curl in the direction of the magnetic field around the wire. The theory of the magnetic field created by an electric current is called Ampère's law.

Electric currents can also exist in materials other than metals, such as plasmas and ionic solutions. The associated magnetic fields may be measured for scientific purposes. For example, the current in a neuron in the brain creates a magnetic field that is measured in magnetoencephalography. Or a lightning strike (current in an ionized path through the atmosphere) creates a magnetic field that may be measured to study the properties of lightning. The magnetic field of the earth is another natural example

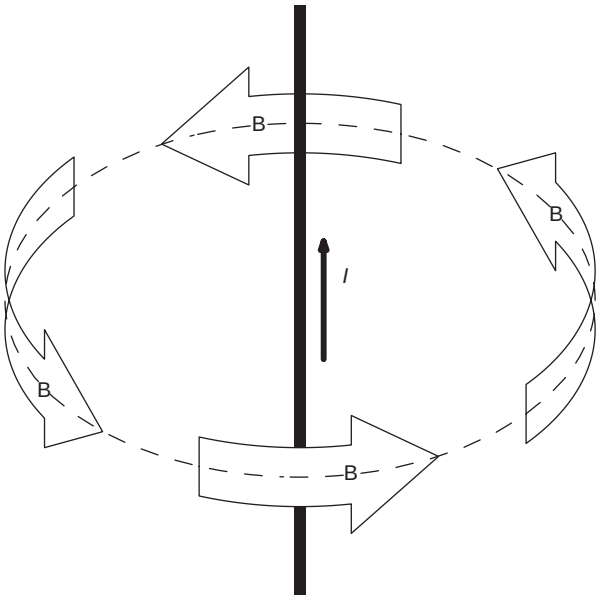


FIGURE 1 The magnetic field **B** curls around the current *I*. The dashed curve indicates an imaginary circle around the wire segment. A compass needle placed near the wire will point in the direction of the field.

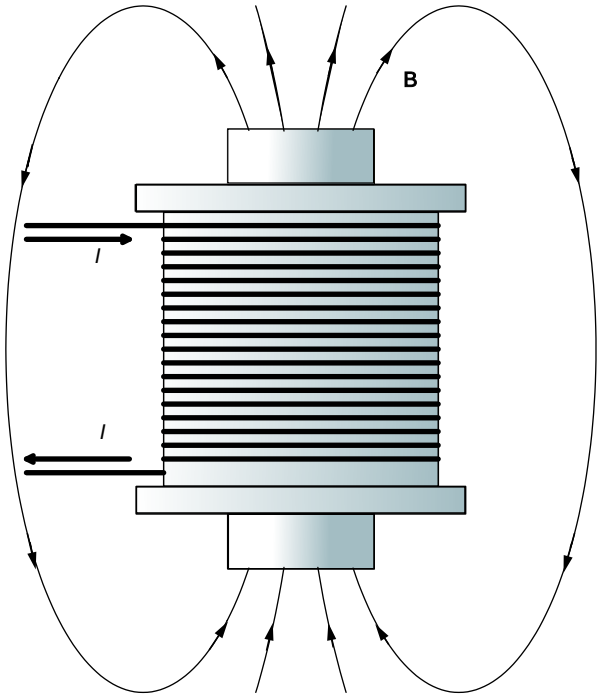


FIGURE 2 An electromagnet.

of a field produced by an electric current. The core of the earth is highly metallic and at high temperature and pressure. Electric currents in this metal core create the earth's magnetism, which we observe at the surface of the earth.

2.2.1.1 Electromagnets Ampère's law is applied in electromagnets. The magnetic field due to a straight length of current-carrying wire is weak. However, if the wire is wound around a cylinder, making a coil with many turns as illustrated in Fig. 2, then the field inside and near the ends of the cylinder can be strong for a practical current. The cylindrical coil is called a solenoid. The field strength can be increased by putting an iron core in the cylinder. Unlike a permanent ferromagnet, the field of an electromagnet can be turned on and off by an electric switch. Electromagnets are commonly used in electric motors and relay switches.

2.2.2 The Magnetic Force on an Electric Current A magnetic field, whether produced by a ferromagnet or by an electric current, exerts a force on any electric current placed in the field. Indeed, there exists a magnetic force on any moving charged particle that moves across the magnetic field vectors. So the interaction between electric charges in motion and a magnetic field has two aspects: (1) an electric current creates a magnetic field (Ampère's law); (2) a magnetic field exerts a force on an electric current.

2.2.2.1 The Electric Motor The magnetic force on an electric current is the basis for all electric motors. The magnetic field may be produced by a permanent magnet (in small DC motors) or by an electromagnet (in large DC and AC motors). The current flows in a rotating coil of wire and may be produced by a battery or some other source of electromotive force. Many practical designs have been invented, with the common feature that a magnetic force acts on the current-carrying coil, in opposite directions on opposite sides of the coil, creating a torque that drives the rotation of the coil. A co-rotating shaft attached to the coil is then connected to a mechanical system to do useful work.

3. THE FIELD CONCEPT

In field theory, electric and magnetic forces are described as the effects of electric and magnetic fields. The theory of electromagnetic phenomena is entirely based on the field concept.

Any electrically charged particle q creates an associated electric field **E**. The field extends throughout the space around q and varies with the position

x in space. If the particle is at rest (in some frame of reference), then its field is independent of time. In a system with many charges, the full electric field is the sum of the electric fields of the individual charges. Thus, the electric field in an electrostatic system, denoted mathematically by $\mathbf{E}(\mathbf{x})$, is a function of position **x** and depends on the locations and charge strengths of all the charged particles.

The electric field extends throughout a volume of space. It exerts a force on any charge in the space. To consider the field concept, suppose two charged particles, q_1 and q_2, are located at positions $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively (see Fig. 3). The electric field at an arbitrary point **x** is

$$\mathbf{E}(\mathbf{x}) = \frac{kq_1}{r_1^2}\,\mathbf{e}_1 + \frac{kq_2}{r_2^2}\,\mathbf{e}_2,$$

where r_1 is the distance from q_1 to **x**, $\mathbf{e}_1$ is the unit vector in the direction from q_1 to **x**, and q_2 and $\mathbf{e}_2$ are the analogous quantities for q_2. A small test charge q placed at **x** will experience a force $\mathbf{F} = q\mathbf{E}(\mathbf{x})$. Since $\mathbf{E}(\mathbf{x})$ is the sum of the fields due to q_1 and q_2, the force **F** on the test charge is the sum of the two forces exerted on q. The charges q_1 and q_2 also experience forces due to the presence of each other. For example, the force on q_2 due to q_1 is $q_2\mathbf{E}_1(\mathbf{x}_2)$, where $\mathbf{E}_1$ is the field due to q_1 alone. The field due to a charged particle is inversely proportional to the square of the distance from the particle, so the force between charges obeys the inverse-square law observed by Coulomb.

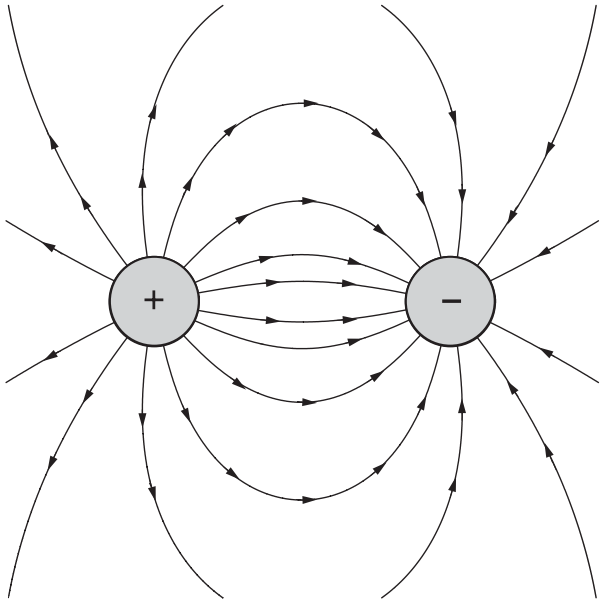


FIGURE 3 The electric field lines of a system of two charges with equal but opposite charge.

In field theory, the force on a charged particle q is attributed to the field $\mathbf{E}(\mathbf{x})$ created by the other charges. Thus, the field concept is significantly different from "action at a distance." The force on the particle, equal to $q\mathbf{E}(\mathbf{x})$, is exerted by the field at the position of q, rather than by direct actions of the distant charges. In other words, an electrostatic system consists of two physical entities: a set of charged particles and an electric field $\mathbf{E}(\mathbf{x})$. The field is just as real as the particles.

Figure 3 illustrates the electric field for a system of two charged particles with equal but opposite charges. The curves in Fig. 3, called the electric field lines, represent the field. The electric field is a vector at each point throughout the space around the charges. A positive test charge q in the field would experience a force in the direction of the field vector at its position. We visualize the field by drawing the field lines, which are curves that are everywhere tangent to the field vector directions.

Electric charges at rest create an electric field $\mathbf{E}(\mathbf{x})$. Ferromagnets and electric currents create another field—the magnetic field $\mathbf{B}(\mathbf{x})$. Figure 4 illustrates the magnetic fields for two elementary current sources: (1) a small current loop and (2) a long current-carrying wire. In a ferromagnet the atoms behave as small current loops with a uniform orientation. The combined atomic fields make up the field of the magnet.

Both **E** and **B** are force fields, which extend throughout a volume of space. But they exert distinct and different forces. The electric field **E** exerts a force on a charge q in the direction of the field vector. The magnetic field **B** exerts a force on moving charges or current-carrying wires. The direction of the magnetic force is perpendicular to both the current and the field, as illustrated in Fig. 5. The magnetic field also exerts forces on the poles of a ferromagnet. The direction of the force is parallel (for a north pole) or antiparallel (for a south pole) to the vector **B**. A compass needle aligns with the magnetic field because the equal but opposite forces on the two poles of the needle compose a torque that twists the needle toward alignment with the field. The magnetic forces on the current elements around a small current loop also compose a torque; in equilibrium, the plane of the loop is oriented perpendicular to the **B** vector. The combined torques on the atoms in a ferromagnet make up the net torque on the magnet.

The field concept was first stated by Michael Faraday. From years of experimental studies on electricity and magnetism, Faraday had formed the idea that a physical entity extends throughout

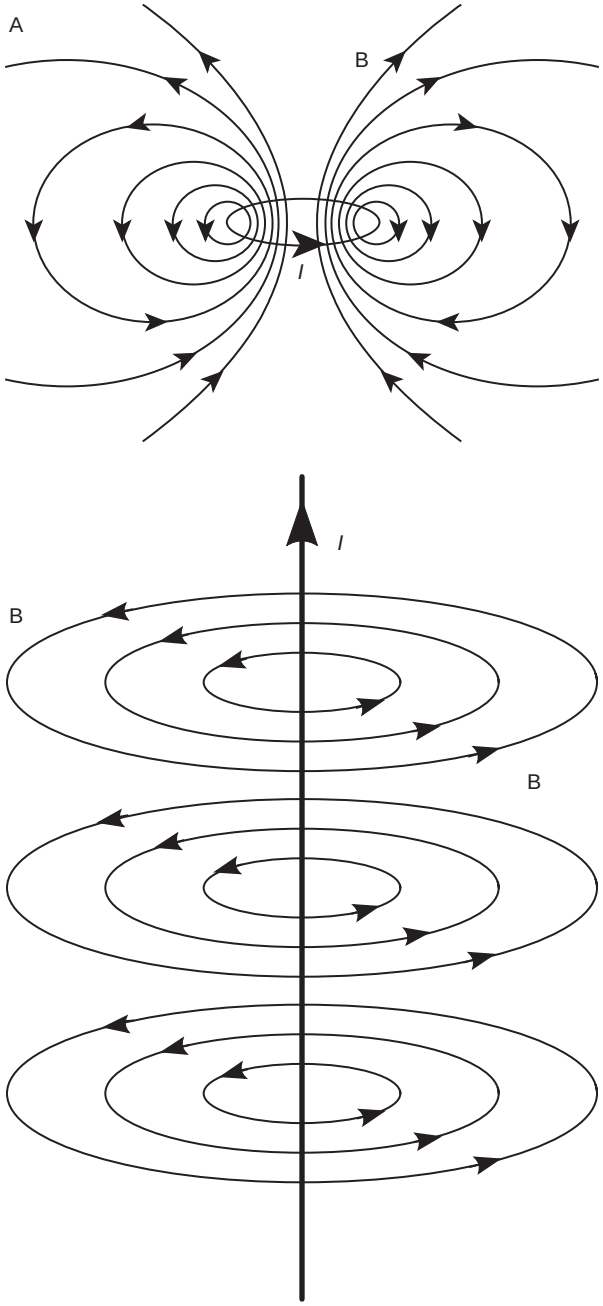


FIGURE 4 Magnetic field of (A) a small current loop and (B) a long straight wire segment.

the space outside charges or magnets and exerts forces on other charges or magnets in the space. He referred to this extended entity as the "lines of force." The term "electromagnetic field" was coined by James Clerk Maxwell, the renowned theoretical physicist who developed a mathematical theory of electromagnetism based on the field concept.

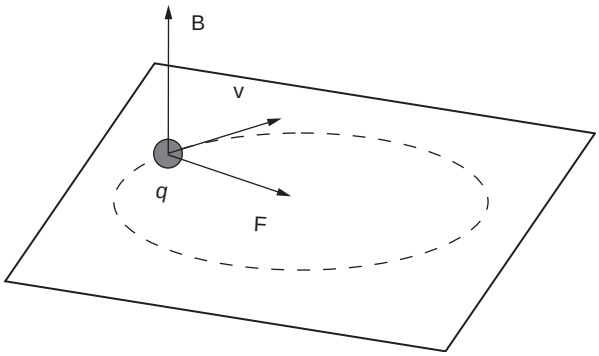


FIGURE 5 The magnetic force on a moving charged particle (or current segment). The direction of the force is perpendicular to both the particle velocity **v** (or current) and the magnetic field: $\mathbf{F} = q\mathbf{v} \times \mathbf{B}$. A charged particle moving on a plane perpendicular to a uniform magnetic field moves along a circle.

4. ELECTROMAGNETIC INDUCTION

The previous sections were concerned with static systems of electric charge or current—that is, in which the fields $\mathbf{E}(\mathbf{x})$ and $\mathbf{B}(\mathbf{x})$ do not change in time. Some relationships between electricity and magnetism were discussed: a steady electric current produces a magnetic field (Ampère's law); a magnetic field exerts a force on any electric charge moving across the field lines. However, for static fields electric and magnetic phenomena appear to be rather distinct.

Time-dependent fields will be described next. The electric field $\mathbf{E}(\mathbf{x}, t)$ and magnetic field $\mathbf{B}(\mathbf{x}, t)$ are functions of time t as well as position $\mathbf{x}$. In dynamic systems, the two fields affect each other significantly. Therefore electric and magnetic phenomena are connected and must be described by a unified theory. Electricity and magnetism are then combined into electromagnetism.

The phenomenon of electromagnetic induction was discovered in 1831 by Michael Faraday in England and independently by Joseph Henry in the United States. The effect is that when a magnetic field changes in time, an electric field is induced in directions that curl around the change of the magnetic field. This phenomenon has important technological applications.

Electromagnetic induction may be observed directly in simple physical demonstrations. Figure 6 shows schematically a coil of conducting wire C connected to a galvanometer G. The galvanometer acts as a current indicator: When no current flows around C the galvanometer needle points toward the

central (zero current) position; when current flows, the needle is deflected. No current source (such as a battery) is connected to the wire coil. In Fig. 6, M is a magnet that can be moved toward or away from the coil C. When the magnet is at rest, no current flows in C and the galvanometer needle points in the central direction.

If the magnet in Fig. 6 moves toward the coil, the galvanometer needle will be deflected in one direction, indicating that a current is flowing in C. The current exists while M is moving. When the motion of M ceases, the needle will return to the central position indicating that the current has stopped. If the demonstration is repeated with the magnet moving away from the coil, the galvanometer needle will be deflected in the opposite direction while M moves. These demonstrations show directly that a change in the magnetic field through the coil induces a current around the coil.

The magnetic field in the demonstration might be varied in a different way. Suppose the bar magnet M in Fig. 6 is replaced by an electromagnet that does not move. Electric current $I(t)$ in the electromagnet solenoid produces a magnetic field $\mathbf{B}(\mathbf{x}, t)$ according to Ampère's law. If the solenoid current is constant, then $\mathbf{B}$ is constant in time and there is no current in the coil C. But if the solenoid current changes, the magnetic field changes. A deflection of the galvanometer needle will be observed while the magnetic field is varying in time. A current around the sensing coil C is again induced when the magnetic field through the coil is changing.

These demonstrations show that when a magnetic field $\mathbf{B}(\mathbf{x}, t)$ changes in time, a current is induced in a conductor that is present in the field. However, the induced current is actually a secondary effect. The current in C is created by an induced electric field $\mathbf{E}(\mathbf{x}, t)$, and $\mathbf{E}$ is the primary effect. Electromagnetic induction is fundamentally a phenomenon of the fields, $\mathbf{B}(\mathbf{x}, t)$ and $\mathbf{E}(\mathbf{x}, t)$. A magnetic field that changes in time induces an electric field in directions curling around the change of the magnetic field. If there happens to be a conducting coil present—as, for example, C in Fig. 6—then the induced electric field drives an electric current around C. But the primary effect is induction of the electric field. The induced electric field $\mathbf{E}(\mathbf{x}, t)$ exists while the magnetic field $\mathbf{B}(\mathbf{x}, t)$ is varying in time.

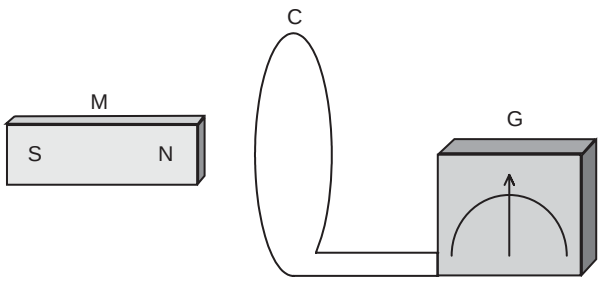


FIGURE 6 Schematic apparatus for demonstrations of electromagnetic induction.

The apparatus shown in Fig. 6 is only schematic. The induced current in C would be very small for an ordinary bar magnet M moving at reasonable velocities. A practical demonstration would require a sensing coil with many turns of wire and a sensitive galvanometer. The effect might be increased by putting an iron core inside the coil to enhance the magnetic field.

Faraday and Henry performed laboratory experiments similar to the demonstrations illustrated in Fig. 6 in their discoveries of electromagnetic induction. Faraday described the results of his detailed studies in terms of the lines of force—his concept of a magnetic field filling space around a magnet. In modern language, a statement summarizing his observations is

> **Faraday's law:** *When the flux of magnetic field through a loop changes in time, an electromotive force (EMF) is induced around the loop.*

This statement is Faraday's law of electromagnetic induction. In equation form, $d\Phi/dt = -\varepsilon$, where Φ is the magnetic flux through the loop, $d\Phi/dt$ is the rate of change of the flux, and ε is the electromotive force around the loop. The "flux" is a quantitative measure of the field lines passing through the loop, defined by the product of the component of the magnetic field vector perpendicular to a surface bounded by the loop and the area of the surface.

The "loop" in Faraday's law is a closed curve (e.g., a circle). The loop may be a conducting loop, in which case the induced EMF drives a current, or it may just be an imaginary curve in space. In either case, the EMF is an induced electric field curling around the change of the magnetic field.

Another, related demonstration may be carried out with the simple apparatus of Fig. 6. Instead of moving the magnet M and holding the coil C fixed, suppose the coil is moved while the magnet is held fixed. Again, a current will be observed around C. The phenomenon in this case is called "motional EMF"; an electromotive force is induced in a conductor that moves relative to a nonuniform magnetic field. In the language of Faraday, when a conducting wire moves through magnetic lines of

force, an induced current flows in the wire. Evidently, any change of the magnetic flux through a conducting loop will induce a current in the loop.

Yet another way to change the magnetic flux through the coil C in Fig. 6 is to change the orientation of the coil. In Fig. 6 the plane of the coil is shown perpendicular to the bar magnet. In this orientation the magnetic field lines pass straight through C; the flux is the product of the magnetic field strength B and the area A of the loop, $\Phi = BA$. Now suppose the coil rotates about a vertical axis, with M fixed. Then the flux of magnetic field through the coil changes as the plane of the loop is at a varying angle to the field vector. When the plane of the loop is parallel to the field lines, the flux is zero because no field lines pass through the coil. While the coil is rotating, a deflection of the galvanometer needle will be observed, consistent with Faraday's law, because of the changing flux. This is another example of motional EMF: The conducting wire moves through the magnetic field lines, and there is an induced current in the wire.

5. APPLICATIONS OF ELECTROMAGNETIC INDUCTION

A number of important inventions are based on the phenomenon of electromagnetic induction. Two that have great technological significance will be described here.

5.1 The Electric Generator

An electric generator (or dynamo) is a machine that generates an electromotive force (EMF) by electromagnetic induction. One or more coils of wire rotate in a magnetic field. The magnetic flux through a coil changes as the coil rotates, inducing an EMF around the coil by Faraday's law, $\varepsilon = -d\Phi/dt$.

Figure 7 shows schematically the design principle of an electric generator. In the model, the square coil rotates about the vertical axis that bisects the coil. The model field **B** is the constant field between north and south poles of ferromagnets. As the coil rotates, the flux through it varies sinusoidally as a function of time. Then there is an induced alternating EMF around the coil.

The model generator in Fig. 7 has the ends of the coil in contact with two slip rings. The voltage difference between the rings is equal to the EMF ε around the coil. Thus the device generates an alternating voltage V (shown in the inset graph) with constant frequency equal to the rotational frequency of the coil. If the rings are connected by a wire with resistance R, an alternating current (AC) I will occur in the wire with $V = IR$ by Ohm's law.

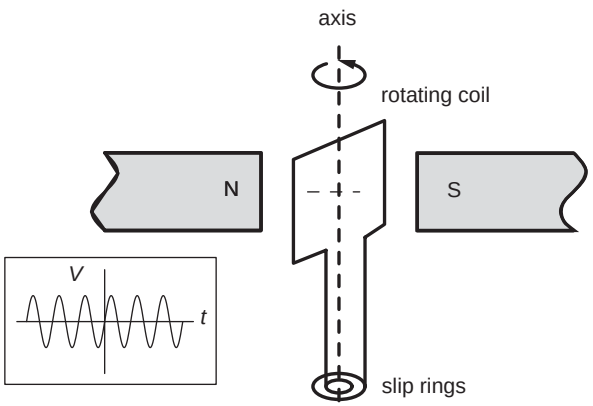


FIGURE 7 The design principle of an electric generator. The magnetic flux through the coil varies as the coil rotates and an EMF is induced in accord with Faraday's law. The voltage difference V between the slip rings varies sinusoidally in time (inset graph).

An external torque must be applied to the coil of the generator to maintain the constant rotation. If the slip rings are connected to an electric appliance, then energy is supplied by the generator because the induced EMF drives an alternating current in the appliance. But energy cannot be created—only converted from one form to another. If no external torque is applied, the generator coil will slow down and stop as its rotational kinetic energy is converted to the work done by the appliance. The slowing of the rotation is an example of magnetic braking: a conductor moving in a magnetic field experiences a force opposite to the motion. To maintain a constant rotation of the generator coil (and hence AC power to the appliance), the coil must be driven by a torque exerted by an external device. In a portable generator, the external torque is produced by a gasoline-powered engine. In an electric power plant, the external torque is produced by a turbine driven by a flow of hot steam; the steam is produced in a boiler (e.g., from burning coal in a coal-fired power plant or from the heat of nuclear reactions in a nuclear reactor).

The electric generators producing electric power around the world are large and complex machines; but the fundamental physics in these devices is simple: Faraday's law and the design principle in Fig. 7.

5.2 The Transformer

A transformer takes an input alternating EMF V_1 (e.g., from an electric generator) and makes an output EMF V_2 that might be supplied to an electric appliance (see Fig. 8). The frequencies of V_1 and V_2 are equal, but the amplitudes are different. A step-down transformer has $V_2 < V_1$; for example, a transformer might take in 110 volts (alternating) from a wall socket and put out 10 volts to a small electric motor in an appliance. A step-up transformer has $V_2 > V_1$; high-voltage applications require such a transformer.

The physical phenomenon acting in a transformer is electromagnetic induction. Figure 8 shows a simple transformer: two coils of conducting wire are wrapped around a toroidal ferromagnetic core. The alternating current driven in the primary coil by the input EMF $V_1(t)$ creates a strong alternating magnetic field $\mathbf{B}(t)$, according to Ampère's law. Because of the ferromagnetism, $\mathbf{B}$ extends around the toroid. Then the changing magnetic flux through the secondary coil induces an EMF by Faraday's law, $\varepsilon = -d\Phi/dt$; this EMF is the output voltage V_2. If N_1 is the number of turns of wire in the primary coil and N_2 is the number of turns in the secondary coil, then the ratio of the voltages is $V_2/V_1 = N_2/N_1$.

Transformers are a crucial part of the electric power grid. An electric power plant creates a three-phase alternating EMF supplied to the grid. But the electric appliances that will use the power are located far from the power plant. Therefore a current must occur over very long transmission lines. The amount of lost power P_{lost}, dissipated in resistance of the transmission wires, is given by the formula $P_{\text{lost}} = I^2R$ where $I =$ current and $R =$ resistance of the line. The power supplied to users is $P_{\text{used}} = IV$, where V is the transmitted EMF. The ratio of power lost to power used is

$$\frac{P_{\text{lost}}}{P_{\text{used}}} = \frac{(P_{\text{used}}/V)^2 R}{P_{\text{used}}} = \frac{P_{\text{used}} R}{V^2},$$

that is, it is inversely proportional to the square of the line voltage. Therefore the transmission of electric power is most efficient if the voltage V is high.

A typical power-plant generator might produce an alternating EMF with amplitude 168 kV (i.e., root mean square EMF 120 kV). But that voltage would be changed in a step-up transformer at the power plant to a very high voltage (e.g., 345 kV). High-voltage transmission lines carry the electric power across large distances. Then step-down transformers are required at the end of a transmission line to reduce the voltage for applications. A transformer substation might reduce the voltage to 4.8 kV for overhead lines into residential areas. A final transformer mounted on the utility pole reduces the voltage to 110 V for the wires to a single house.

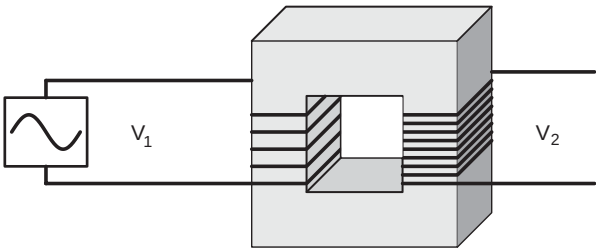


FIGURE 8 The design principle of a transformer. Two coils of conducting wire are wrapped around an iron core. The number of turns is different in the two coils, so the output EMF V_2 is different from the input EMF V_1.

6. THE MAXWELL EQUATIONS OF THE ELECTROMAGNETIC FIELD

The mathematical theory of electromagnetism was developed and published in 1864 by James Clerk Maxwell. He described the known electric and magnetic effects in terms of four equations relating the electric and magnetic fields and their sources—charged particles and electric currents. The development of this theory was a supreme achievement in the history of science. Maxwell's theory is still used today by physicists and electrical engineers. The theory was further developed in the 20th century to account for the quantum theory of light. But even in quantum electrodynamics, Maxwell's equations remain valid although their interpretation is somewhat different from the classical theory. In any case, Maxwell's theory continues today to be an essential part of theoretical physics.

A knowledge of calculus and vectors is necessary for a full understanding of the Maxwell equations. However, the essential structure of the theory can be understood without going into the mathematical details. Each equation is expressed most powerfully as a partial differential equation relating variations of the electric and magnetic fields, with respect to variations of position or time, and the charge and current densities in space.

Gauss's Law. Gauss's law, written as a field equation, is $\nabla \cdot \mathbf{E} = \rho/\varepsilon_0$. (The symbol $\nabla\cdot$ denotes the divergence operator.) Here $\rho(\mathbf{x}, t)$ is the charge per unit volume in the neighborhood of $\mathbf{x}$ at time t; ε_0

is a constant of nature equal to $8.85 \times 10^{-12}\,C^2/Nm^2$. Gauss's law relates the electric field $\mathbf{E}(\mathbf{x}, t)$ and the charge density. The solution for a charged particle q at rest is $\mathbf{E}(\mathbf{x}) = kq\mathbf{e}/r^2$, where r is the distance from the charge to $\mathbf{x}$, $\mathbf{e}$ is the direction vector, and $k = 1/(4\pi\varepsilon_0)$; this is the familiar inverse square law of electrostatics. Electric field lines diverge at a point charge.

Gauss's Law for Magnetism. The analogous equation for the magnetic field is $\nabla \cdot \mathbf{B} = 0$. There are no magnetic monopoles, particles that act as a point source of $\mathbf{B}(\mathbf{x}, t)$. Unlike the electric field lines, which may terminate on charges, the magnetic field lines always form closed curves because magnetic charges do not exist. There is no divergence of magnetic field lines.

Faraday's Law. The field equation that describes Faraday's law of electromagnetic induction is $\nabla \times \mathbf{E} = -\partial\mathbf{B}/\partial t$. The quantity $\nabla \times \mathbf{E}$, called the curl of $\mathbf{E}(\mathbf{x}, t)$, determines the way that the vector field $\mathbf{E}$ curls around each direction in space. Also, $\partial\mathbf{B}/\partial t$ is the rate of change of the magnetic field. This field equation expresses the fact that a magnetic field that varies in time implies an electric field that curls around the change of the magnetic field. It is equivalent to Faraday's statement that the rate of change of magnetic field flux through a surface S is equal to an electromotive force (EMF) around the boundary curve of S.

The Ampère-Maxwell Law. In a system of steady electric currents, the magnetic field is constant in time and curls around the current in directions defined by the right-hand rule. The field equation that expresses this field $\mathbf{B}$ (Ampère's law) is $\nabla \times \mathbf{B} = \mu_0\mathbf{J}$, where $\mathbf{J}(\mathbf{x})$ is the current per unit area at $\mathbf{x}$ and μ_0 is a constant equal to $4\pi \times 10^{-7}\,Tm/A$. (The units are T = tesla for magnetic field and A = ampere for electric current). But Ampère's law is incomplete, because it does not apply to systems in which the currents and fields vary in time. Maxwell deduced from mathematical considerations a generalization of Ampère's law,

$$\nabla \times \mathbf{B} = \mu_0\mathbf{J} + \mu_0\varepsilon_0\frac{\partial \mathbf{E}}{\partial t},$$

in which the second term on the right side is called the displacement current. The displacement current is a necessary term in order for the system of four partial differential equations to be self-consistent. The Ampère-Maxwell law implies that $\mathbf{B}$ curls around either electric current ($\mathbf{J}$) or changing electric field ($\partial\mathbf{E}/\partial t$). The latter case is analogous to electromagnetic induction but with $\mathbf{E}$ and $\mathbf{B}$ reversed; a rate of change in one field induces circulation in the other field.

Maxwell's introduction of the displacement current was a daring theoretical prediction. At that time there was no experimental evidence for the existence of displacement current. Laboratory effects predicted by the displacement current are very small, and their observation was not possible with the apparatus available at that time. However, the Maxwell equations, including the displacement current, make a striking prediction—that light consists of electromagnetic waves. The fact that Maxwell's theory explains the properties of light, and other forms of electromagnetic radiation, provides the evidence for the existence of the displacement current.

7. ELECTROMAGNETIC WAVES

Electromagnetism includes phenomena involving charges, currents, magnets, and the fields $\mathbf{E}$ and $\mathbf{B}$. Light is also an electromagnetic phenomenon. Light and other forms of radiation are described by field theory as electromagnetic waves. Therefore optics—the science of light—is a part of electromagnetism.

The Maxwell equations describe the behavior of the electric field $\mathbf{E}(\mathbf{x}, t)$ and magnetic field $\mathbf{B}(\mathbf{x}, t)$. The time-dependent fields influence each other, even in a vacuum where no charge or current is present, through Faraday's law ($\nabla \times \mathbf{E} = -\partial\mathbf{B}/\partial t$) and the displacement current ($\nabla \times \mathbf{B} = \mu_0\varepsilon_0\partial\mathbf{E}/\partial t$). The four field equations are all satisfied in vacuum if both $\mathbf{E}(\mathbf{x}, t)$ and $\mathbf{B}(\mathbf{x}, t)$ are waves traveling in space. However, the electric and magnetic waves are codependent (i.e., intricately related in several ways).

Figure 9 shows a snapshot in time of the fields for a section of an electromagnetic wave in vacuum. The field vectors are shown at points along a line parallel to the direction of propagation of the wave (z axis). The wavelength is constant (at least for the section of wave shown). The field vectors $\mathbf{E}$ and $\mathbf{B}$ vary in magnitude as a function of position along the line of propagation. The magnitudes are related by $B = E/c$, where c is the wave speed. The directions of $\mathbf{E}$ and $\mathbf{B}$ are orthogonal, in the x and y directions, respectively. The field components oscillate sinusoidally with z. An electromagnetic wave is a transverse wave because the fields oscillate in directions orthogonal to the direction of propagation.

Figure 9 shows the field vectors along only a single line parallel to the direction of propagation $\hat{\mathbf{z}}$. The full wave fills a three-dimensional volume. In a

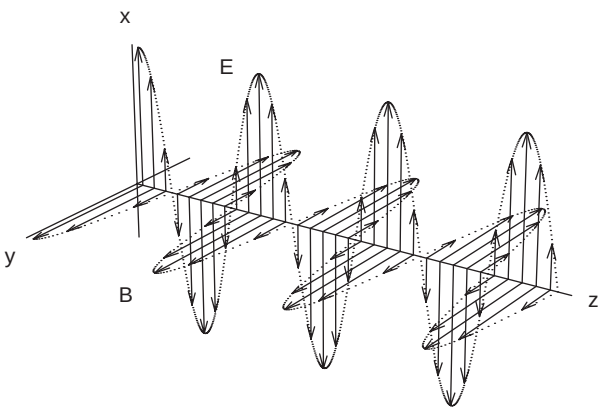


FIGURE 9 Electric and magnetic field vectors in a section of a linearly polarized electromagnetic wave.

linearly polarized plane wave the field vectors are constant on planar wave fronts orthogonal to $\hat{\mathbf{z}}$—that is, with **E** and **B** pointing in the $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ directions, respectively, uniformly over the wave front. All properties of polarized light are then described by the wave theory. As another 3D example, a short radio antenna radiates a spherical wave; far from the antenna **E** and **B** are orthogonal to one another on spherical wave fronts.

A wave is a structure that is extended in space. The parts of the wave also vary in time. Figure 9 is a snapshot showing the wave as a function of position at an instant of time. As time passes, the fields at any fixed position **x** will change. For a linearly polarized plane wave, **E** at **x** oscillates in time between orientations parallel and antiparallel to a fixed polarization direction ($\hat{\mathbf{x}}$), which is orthogonal to the line of propagation ($\hat{\mathbf{z}}$); **B** varies similarly in a direction ($\hat{\mathbf{y}}$) orthogonal to both **E** and the line of propagation. The combined variations in space and time imply that the wave structure moves as a whole in the direction of propagation. So a snapshot at a later time would look just the same as Fig. 9 except translated by some distance in the direction of propagation. In other words, the shape is constant but the positions of the nodes (points where $\mathbf{E} = 0$ and $\mathbf{B} = 0$) move in space.

7.1 The Electromagnetic Spectrum

Electromagnetic waves have an infinite range of wavelength λ. Visible light has λ from 400 nm (violet) to 700 nm (red). Ultraviolet and infrared light have shorter and longer wavelengths, respectively. Microwaves have λ from 1 mm to 1 m, with important technological uses (e.g., communications and radar). X-rays have very short wavelengths, $\lambda < 1$ nm. All these forms of electromagnetic radiation have the same wave speed in vacuum, which can be evaluated from the field theory, $c = 1/\sqrt{\mu_0 \varepsilon_0} = 2.998 \times 10^8$ m/s.

8. SUMMARY

Our scientific understanding of electricity and magnetism has developed over a period of centuries. Phenomena that we understand today as examples of electricity, magnetism, or electromagnetism have been observed since ancient times: For example, lightning is an electric discharge in the atmosphere; a lodestone attracts iron; light consists of electromagnetic waves. Such common phenomena, while easily observed, were unexplained until the scientific revolution. Today we are able to accurately describe and explain all variety of electromagnetic phenomena, based on the mathematical theory of fields. With this knowledge has come the ability to create technologies, such as the distribution of electric power, that affect our daily lives.

SEE ALSO THE FOLLOWING ARTICLES

Electrical Energy and Power • *Electricity Use, History of* • *Electric Motors* • *Magnetic Levitation* • *Magnetohydrodynamics*

Further Reading

Griffiths, D. J. (2000). "Introduction to Electrodynamics." 3rd ed. Prentice-Hall, Engelwood Cliffs, NJ.

Jackson, J. D. (1998). "Classical Electrodynamics." 3rd ed. John Wiley, New York.

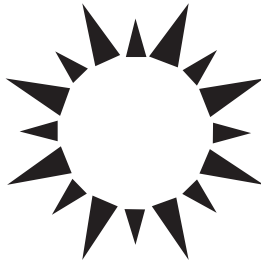
Lorrain, P., Corson, D. R., and Lorrain, F. (2000). "Fundamentals of Electromagnetic Phenomena." W. H. Freeman, New York.

Meyer, H. W. (1972). "A History of Electricity and Magnetism." MIT Press, Norwalk, CT.

Pollack, G. L., and Stump, D. R. (2002). "Electromagnetism." Addison-Wesley, San Francisco.

Whittaker, E. T. (1960). "A History of the Theories of Aether and Electricity." Harper, New York.

Emergy Analysis and Environmental Accounting



MARK T. BROWN
University of Florida
Gainesville, Florida, United States
SERGIO ULGIATI
University of Siena
Siena, Italy

1. **Introduction**
2. **Emergy Evaluation of the Biosphere and its Processes**
3. **Unit Emergy Values for Fuels and Some Common Products**
4. **Case Studies: Evaluation of Energy Conversion Systems**
5. **Summary**

Glossary

emdollar (or EM$) A measure of the money that circulates in an economy as the result of some process. In practice, to obtain the emdollar value of an emergy flow or storage, the emergy is multiplied by the ratio of total emergy to gross domestic product for the national economy.

emergy An expression of all the energy used in the work processes that generate a product or service in units of one type of energy. Solar emergy of a product is the emergy of the product expressed in the equivalent solar energy required to generate it. Sometimes it is referred to as energy memory.

emjoule The unit of measure of emergy, "emergy joule." It is expressed in the units of energy previously used to generate the product; for instance, the solar emergy of wood is expressed as joules of solar energy that were required to produce the wood.

empower Emergy per unit time (sej/time)

energy Sometimes referred to as the ability to do work. Energy is a property of all things that can be turned into heat and is measured in heat units (BTUs, calories, or joules)

nonrenewable emergy The emergy of energy and material storages like fossil fuels, mineral ores, and soils that are consumed at rates that far exceed the rates at which they are produced by geologic processes.

production Production measured in emergy is the sum of all emergy inputs to a process.

renewable emergy The emergy of energy flows of the biosphere that are more or less constant and reoccurring, and that ultimately drive the biological and chemical processes of the earth and contribute to geologic processes.

transformity The ratio obtained by dividing the total emergy that was used in a process by the energy yielded by the process. Transformities have the dimensions of emergy/energy (sej/J). A transformity for a product is calculated by summing all of the emergy inflows to the process and dividing by the energy of the product. Transformities are used to convert energies of different forms to emergy of the same form.

This article presents a brief review of the concepts of energy hierarchy and definitions of emergy and related quantities. Tables of data are given on global emergy flows from which emergy and transformities of most products and processes of the biosphere are calculated. Then tables of transformities for many products are provided. Finally, several case studies of evaluations of energy technologies are given. In a summary discussion emergy analysis (EMA) and embodied energy analysis (EEA) are compared describing the strengths and weaknesses of each approach related to the evaluation of energy systems.

1. INTRODUCTION

1.1 Energy and the Emergy Hierarchy

Emergy is a universal measure of the work of nature and society made on a common basis. The work of nature and society results in energy transformations

 © 2004 Elsevier Inc. All rights reserved.

that when viewed in totality are interconnected in webs of energy flow. All energy transformations of the geobiosphere can be arranged in an ordered series to form an energy hierarchy. For example, many joules of sunlight are required to make a joule of organic matter, many joules of organic matter to make a joule of fuel, several joules of fuel are required to make a joule of electric power, and so on. Since different kinds of energy are not equal in their contribution to processes, work is made comparable by expressing each form of energy in units of one form. To accomplish this, the available energy of different forms of energy is evaluated by means of the energy of one type previously required to produce it. This quantity of one type of energy previously used is called emergy (spelled with an "m").

1.2 Definitions

The following paragraphs contain definitions and a brief explanation of emergy concepts. A more complete introduction can be found in H. T. Odum's 1996 text, *Environmental Accounting: Emergy and Environmental Decision Making*.

Emergy is the availability of energy (exergy) of one kind that is used up in transformations directly and indirectly to make a product or service. The unit of emergy is the emjoule, a unit referring to the available energy of one kind consumed in transformations. For example, sunlight, fuel, electricity, and human service can be put on a common basis by expressing them all in the emjoules of solar energy that is required to produce each. In this case, the value is a unit of solar emergy expressed in solar emjoules (abbreviated sej). Although other units have been used, such as coal emjoules or electrical emjoules, in most cases, all emergy data are given in solar emjoules.

Unit emergy values are calculated based on the emergy required to produce them. There are three main types of unit emergy values as follows:

- Transformity is one example of a unit emergy value and is defined as the emergy per unit of available energy (exergy). For example, if 4000 solar emjoules are required to generate a joule of wood, then the solar transformity of that wood is 4000 solar emjoules per joule (abbreviated sej/J). Solar energy is the largest but most dispersed energy input to the earth. The solar transformity of the sunlight absorbed by the earth is 1.0 by definition.
- Specific emergy is the unit emergy value of matter defined as the emergy per mass, usually expressed as solar emergy per gram (sej/g). Solids may be evaluated best with data on emergy per unit mass for its concentration. Because energy is required to concentrate materials, the unit emergy value of any substance increases with concentration. Elements and compounds not abundant in nature therefore have higher emergy/mass ratios when found in concentrated form since more work was required to concentrate them, both spatially and chemically.
- Emergy per unit money is a unit emergy value used to convert money payments into emergy units. Since money is paid to people for their services and not to the environment, the contribution to a process represented by monetary payments is the emergy that people purchase with the money. The amount of resources that money buys depends on the amount of emergy supporting the economy and the amount of money circulating. An average emergy/money ratio in solar emjoules/$ can be calculated by dividing the total emergy use of a state or nation by its gross economic product. It varies by country and has been shown to decrease each year. This emergy/money ratio is useful for evaluating service inputs given in money units where an average wage rate is appropriate.

Emergy accompanying a flow of something (energy, matter, information, etc.) is easily calculated if the unit emergy value is known. The flow expressed in its usual units is multiplied by the emergy per unit of that flow. For example, the flow of fuels in joules per time can be multiplied by the transformity of that fuel (emergy per unit energy in solar emjoules/joule), or the mass of a material input can be multiplied by its specific emergy (emergy per unit mass in solar emjoules/gram). The emergy of a storage is readily calculated by multiplying the storage quantity in its usual units by the emergy per unit.

Unit emergy values are a kind of efficiency measure, since they relate all the inputs to an output. The lower the transformity or specific emergy, the more efficient the conversion. It follows from the second law that there are some minimum unit emergy values for processes, which are consistent with maximum power operations. While there is no way to calculate them directly, the lowest transformity found in long-operating systems is used as an approximation. When estimating a theoretical potential of some system, it is appropriate to use the best (lowest) transformity known.

Empower is a flow of emergy (i.e., emergy per time). Emergy flows are usually expressed in units of solar empower (solar emjoules per time).

1.3 Emergy Evaluation Procedure

Emergy accounting uses the thermodynamic basis of all forms of energy, materials, and human services but converts them into equivalents of one form of energy. Emergy accounting is organized as a top-down approach where first a system diagram of the process is drawn to organize the evaluation and account for all inputs and outflows. Tables of the actual flows of materials, labor, and energy are constructed from the diagram and all flows are evaluated. The final step of an emergy evaluation involves interpreting the quantitative results. In some cases, the evaluation is done to determine fitness of a development proposal. In others, it may be a question of comparing different alternatives. The evaluation may be seeking the best use of resources to maximize economic vitality. So the final step in the evaluation is to calculate several emergy indices that relate emergy flows of the system being evaluated with those of the environment and larger economy within which it is embedded and that allow the prediction of economic viability, carrying capacity, or fitness.

This evaluation process has been termed emergy synthesis. Synthesis is the act of combining elements into coherent wholes. Rather than dissect and break apart systems and build understanding from the pieces upward, emergy synthesis strives for understanding by grasping the wholeness of systems. Emergy is a systems concept, context driven, and cannot be fully understood or utilized outside of systems. By evaluating complex systems using emergy methods, the major inputs from the human economy and those coming "free" from the environment can be integrated to analyze questions of public policy and environmental management holistically.

1.3.1 Left-Right Energy Systems Diagram

Systems diagrams are used to show the inputs that are evaluated and summed to obtain the emergy of a resulting flow or storage. The purpose of the system diagram is to conduct a critical inventory of processes, storages, and flows that are important to the system under consideration and are therefore necessary to evaluate. Components and flows within diagrams are arranged from left to right reflecting more available energy flow on the left, decreasing to the right with each successive energy transformation. For example, abundant solar energy is utilized in successive transformations in ecological, agricultural, and technoeconomic subsystems to support a small amount of high-quality energy of humans, their government, and information. A simple diagram of the global system including humans is shown in Fig. 1. The left-to-right organization also corresponds to increasing scale of territory and turnover time. As illustrated in Fig. 1, every energy transformation box has more than one input, including larger energy flows from the left, lesser amounts from units in parallel, and small but important controlling energies feeding back from the right.

1.3.2 Preparation of an Emergy Evaluation Table

Tables of the actual flows of materials, labor, and energy are constructed from the diagram. Raw data on flows and storage reserves are converted into emergy units and then summed for a total emergy flow to the system. Inputs that come from the same source are not added, to avoid double counting. Only the larger input is accounted for. If the table is for the evaluation of a process, it represents flows per unit time (usually per year). If the table is for the evaluation of reserve storages, it includes those storages with a turnover time longer than a year.

Separate tables are constructed for evaluations of flows and storage. Tables are usually constructed in the same format, as given by the column headings and format shown in Table I.

- Column 1 is the line item number, which is also the number of the footnote found below the table where raw data sources are cited and calculations are shown.
- Column 2 is the name of the item, which is also shown on the aggregated diagram.
- Column 3 is the raw data in joules, grams, dollars or other units. The units for each raw data item are shown in column 4.
- Column 5 is the emergy per unit used for calculations, expressed in solar emergy joules per unit. Sometimes, inputs are expressed in grams, hours, or dollars, therefore an appropriate conversion ratio is used (sej/hr; sej/g; sej/$).
- Column 6 is the solar emergy of a given flow, calculated as raw input times the transformity (column 3 times column 5).
- Column 7 is the emdollar value of a given item for a given year. This is obtained by dividing the emergy in Column 6 by the emergy-to-money ratio (EMR) for the country and selected year of the evaluation (units are sej/$). The EMR is calculated independently. The resulting values in this column express the amount of economic activity that can be supported by a given emergy flow or storage.

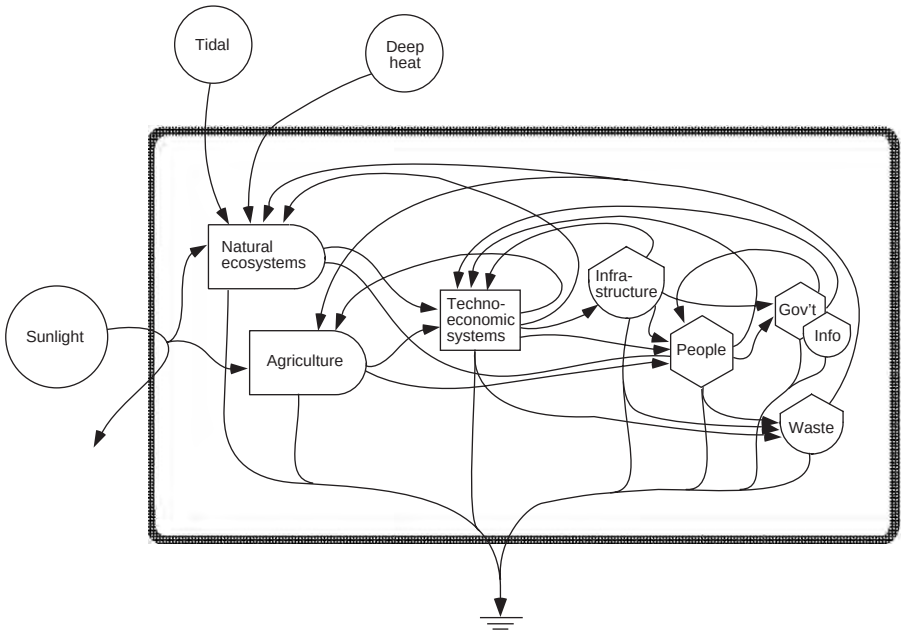


FIGURE 1 Successive energy transformations supporting human civilization. Large energy flows from the right support smaller and smaller amounts of high-quality energy. With each succeeding energy transformation, some energy is degraded some is passed on to the next level (to the left) and some is fed back in control actions. Explanations of symbols may be found in Odum (1996).

TABLE I

Example Emergy Evaluation Table

1	2	3	4	5	6	7
Note	Item	Data	Units	Emergy/unit (sej/unit)	Solar emergy (E + 15 sej/year)	em$ Value (1998 em$/year)
1.	First item	xx.x	sej/year	xxx.x	xxx.x	xxx.x
2.	Second item	xx.x	g/year	xxx.x	xxx.x	xxx.x
..						
..						
n.	nth item	xx.x	sej/year	xxx.x	xxx.x	xxx.x
O.	Output	xx.xx	sej or g/year	xxx.x	$\sum_n^1 Em$	xxx.x

1.3.3 Emergy of Storages

When calculating the emergy of stored quantities (storages)—for instance, wood biomass in a forest ecosystem or buildings in a city—it is necessary to sum the emergy of each of the inputs for the time of its contribution. Input emergy inflows are multiplied by the time it takes to accumulate the storage and exported yield, if any.

1.3.4 Evaluations Based on Averaged Inputs

Apparently all systems pulse with time intervals and pulse strength that increase with scale. To evaluate a process on one scale of time and space usually means using averages for each of the inputs from smaller scales where pulses are of high frequency. For example, for an evaluation of phenomena on the scale of human economy, yearly averages are often appropriate. On this scale average solar energy and average tidal energy are used. For calculations of global processes over a longer scale of time, an average of inputs from the deep earth may be used. Thus, for many purposes emergy evaluations are made with averaged inputs. The result is emergy evaluation as if the system was in a steady state. If the table is an evaluation of a storage, the result is in solar emjoules. If the table is an evaluation of a

product flow, the result is in solar emjoules per time (or empower).

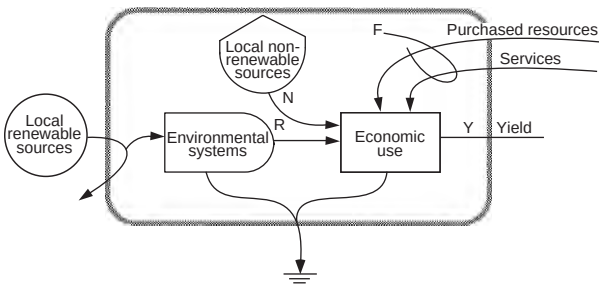
Dynamic emergy evaluations of storages and flows are also used. When a storage is growing, decreasing, or oscillating, its stored emergy contents and transformity are changing as well. Where a system is adequately represented by a dynamic model, an algorithm can be added to the model which calculates a time sequence of emergy and transformity. A simulation model of this type produces a running calculation according to the emergy definition, summing the inputs to storages, multiplying these by their unit emergy, and subtracting emergy removals.

1.3.5 Calculating Unit Emergy Values

After a table is prepared that evaluates all the inputs, unit emergy values of products can be calculated. The output or product (row "O" in Table I) is evaluated first in units of energy, exergy, or mass. Then the input emergy is summed and unit emergy value for the product calculated by dividing the emergy by the units of the output. The unit values that result are useful for other emergy evaluations. Thus, any emergy evaluation generates new emergy unit values.

1.3.6 Performance Indicators

The systems diagram in Fig. 2 shows nonrenewable environmental contributions (N) as an emergy storage of materials, renewable environmental inputs (R), and inputs from the economy as purchased (F) goods and services. Purchased inputs are needed for the process to take place and include human service and purchased nonrenewable energy and material brought in from elsewhere (fuels, minerals, electricity, machinery, fertilizer, etc.). Several ratios, or indices are given in Fig. 1 that are used to evaluate



Yield (Y) = R + N + F
Emergy yield ratio = Y/F
Emergy investment ratio = F/(R + N)
Environmental loading ratio = (F + N)/R
Empower density = (R + N + F)/area

FIGURE 2 Emergy indices used as performance indicators.

the global performance of a process as follows:

Emergy yield ratio. The ratio of the total emergy (local and imported) driving a process or system to the emergy imported. The ratio is a measure of the potential contribution of the process to the main economy, due to the exploitation of local resources.

Environmental loading ratio. The ratio of nonrenewable and imported emergy use to renewable emergy use. It is an indicator of the pressure of a transformation process on the environment and can be considered a measure of ecosystem stress due to a production (transformation activity).

Emergy sustainability index. The ratio of the emergy yield ratio to the environmental loading ratio. It measures the potential contribution of a resource or process to the economy per unit of environmental loading.

Emergy investment ratio. The ratio of emergy fed back from outside a system to the indigenous emergy inputs (both renewable and nonrenewable). It evaluates if a process is a good user of the emergy that is invested, in comparison with alternatives.

Empower density. The ratio of total emergy use in the economy of a region or nation to the total area of the region or nation. Renewable and nonrenewable emergy density are also calculated separately by dividing the total renewable emergy by area and the total nonrenewable emergy by area, respectively.

Several other ratios are sometimes calculated depending on the type and scale of the systems being evaluated.

Percent renewable emergy (%Ren). The ratio of renewable emergy to total emergy use. In the long run, only processes with high %Ren are sustainable.

Emprice. The emprice of a commodity is the emergy one receives for the money spent. Its units are sej/$.

Emergy exchange ratio. The ratio of emergy exchanged in a trade or purchase (what is received to what is given). The ratio is always expressed relative to one or the other trading partners and is a measure of the relative trade advantage of one partner over the other.

Emergy per capita. The ratio of total emergy use in the economy of a region or nation to the total population. Emergy per capita can be used as a measure of potential, average standard of living of the population.

1.3.7 Understanding Emergy Indicators: What They Indicate

Lengthy discussion of the performance indicators were given in a previous publication by the authors titled "Emergy Based Indices and Ratios to Evaluate Sustainability—Monitoring Economies and Technology towards Environmentally Sound Innovation." Additional background and further discussion follows.

Transformity only measures how much emergy it takes to generate one unit of output, regardless of whether or not the input is renewable. It is a measure of efficiency on the global spatial and timescale of the biosphere. It indicates the hierarchical position of an item in the thermodynamic scale of the biosphere and can be regarded as a quality factor from the point of view of biosphere dynamics.

The ELR = (N + F)/R is designed to compare the amount of nonrenewable and purchased emergy (N + F) to the amount of locally renewable emergy (R). In the absence of investments from outside, the renewable emergy that is locally available would have driven the growth of a mature ecosystem consistent with the constraints imposed by the environment and would be characterized by an ELR = 0. Instead, the nonrenewable imported emergy drives a different site development, whose distance from the natural ecosystem can be indicated by the ratio (N + F)/R. The higher this ratio, the bigger the distance of the development from the natural process that could have developed locally. In some ways, the ELR is a measure of the disturbance to the local environmental dynamics, generated by the development driven from outside. The ELR is clearly able to make a difference between nonrenewable and renewable resources, thus complementing the information that is provided by the transformity (Brown and Ulgiati, 1997). From past experience gained in previous case studies investigated, it appears that low ELRs (around two or less) are indicative of relatively low environmental impacts (or processes that can use large areas of a local environment to "dilute impacts"). ELRs between three and ten are indicative of moderate environmental impacts, while ELRs ranging from ten up to extremely high values indicate much higher environmental impacts due to large flows of concentrated nonrenewable emergy in a relatively small local environment.

The emergy yield ratio, EYR = (R + N + F)/F, is a measure of the ability of a process to exploit and make available local resources by investing outside resources. It provides a look at the process from a different perspective, its openness. It provides a measure of the appropriation of local resources by a process, which can be read as a potential additional contribution to the economy, gained by investing resources already available. The lowest possible value of the EYR is 1, which indicates that a process delivers the same amount of emergy that was provided to drive it, and that it is unable to usefully exploit any local resource. Therefore, processes whose EYR is 1 or only slightly higher do not provide significant net emergy to the economy and only transform resources that are already available from previous processes. In so doing, they act as consumer processes more than creating new opportunities for system's growth. Primary energy sources (crude oil, coal, natural gas, uranium) usually show EYRs greater than 5, since they are exploited by means of a small input from the economy and return much greater emergy flows, which have been generated by previous geologic and ecological activities that accumulated these resources over past millennia. Secondary energy sources and primary materials like cement and steel show EYRs in the range from 2 to 5, indicating moderate contribution to the economy.

If the ELR (nonrenewable versus renewable emergy flows) and EYR (outside versus local emergy flows) are combined, a sustainability index is created—that is, an aggregated measure of the potential contribution to the larger system (EYR) per unit of loading imposed on the local system (ELR). This indicator, called emergy index of sustainability (EIS) is usefully applicable to measure openness and loading changes occurring over time in both technological processes and economies. In principle, the lowest possible value of the EIS is zero (when EYR = 0 and ELR $< \infty$ or when EYR $\neq 0$ and ELR $\rightarrow \infty$), while the theoretical upper limit ($\rightarrow \infty$) is only possible for untouched, mature ecosystems. According to the results of several case studies investigated, EIS's lower than 1 appear to be indicative of consumer products or processes and those greater than 1 indicative of products that have net contributions to society without heavily affecting its environmental equilibrium. As it relates to economies, an EIS lower than 1 is indicative of highly developed consumer-oriented systems, EISs between 1 and 10 have been calculated for what have been termed "developing economies," while EISs greater than 10 indicate economies that have not yet significantly started any industrial development.

Finally, the emergy density, ED, measures the amount of emergy invested on one unit of land and suggest land be a limiting factor to any kind of development or process. Higher EDs characterize

city centers, information centers such as governmental buildings, universities, and research institutions, and industrial clusters, while lower EDs are calculated for rural areas and natural environments.

2. EMERGY EVALUATION OF THE BIOSPHERE AND ITS PROCESSES

Figure 3 shows an aggregated system diagram of the energy transformation network of the biosphere arranged with decreasing energy from left to right. In Fig. 3, all the components interact and are required by the others. As a result, the total emergy driving the biosphere (the sum of solar, tidal, and deep heat) is required by all processes within, and thus the emergy assigned to each of the internal pathways is the same. After millions of years of self-organization, the transformations of the driving energies by the atmosphere, ocean, and land are organized simultaneously to interact and contribute mutual reinforcements. Therefore, the energy flow of each jointly necessary process is the sum of the emergy from the three sources.

2.1 Annual Budget of Emergy Flow (Empower) Supporting the Geobiosphere

An emergy evaluation table of the main inputs to the geobiosphere of the earth (omitting, for the moment, the emergy use from nonrenewable resources) is given in Table II. The annual budget of emergy flow (empower) supporting the geobiosphere (atmosphere, ocean, and earth crust) includes solar energy, tidal energy, and heat energy from the deep earth. Other inputs from space, such as the high-energy radiation of solar flares, cosmic rays, meteorites, and stellar dust, are not evaluated. All of these vary with oscillations and pulses, and their emergy values vary with their intensities.

Total emergy contributions to the geobiosphere are about 15.83 E24 sej/year based on a reevaluation

TABLE II

Annual Emergy Contributions to Global Processes (after Odum et al., 2000)*

Input	Units	Inflow units/year	Emergy/unit sej/unit	Empower E24 sej/year
Solar insolation,	J[a]	3.93 E24	1.0	3.93
Deep earth heat,	J[b]	6.72 E20	1.20 E4	8.06
Tidal energy,	J[c]	0.52 E20	7.39 E4	3.84
Total		—	—	15.83

Abbreviations: sej = solar emjoules; E24 means multiplied by 10^{24}.

*Not including nonrenewable resources.

[a] Sunlight: solar constant 2 gcal/cm^2/min = 2 Langley per minute; 70% absorbed; earth cross section facing sun 1.27 E14 m^2.

[b] Heat release by crustal radioactivity 1.98 E20 J/year plus 4.74 E20 J/year heat flowing up from the mantle (Sclater *et al.*, 1980). Solar transformity 1.2 E4 sej/J based on an emergy equation for crustal heat as the sum of emergy from earth heat, solar input to earth cycles, and tide (Odum, 2000a).

[c] Tidal contribution to oceanic geopotential flux is 0.52 E20 J/year (Miller, 1966). Solar transformity of 7.4 E4 sej/J is based on an emergy equation for oceanic geopotential as the sum of emergy from earth heat, solar input to the ocean, and tide following Campbell (1998) (Odum, 2000a).

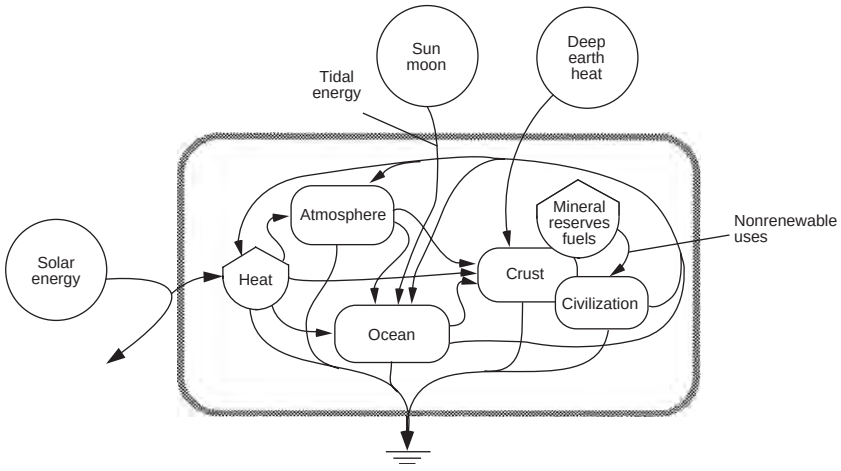


FIGURE 3 Energy transformation network of the biosphere. The interconnected pathways of energy flow, inputs, and all main components are necessary to all the other components and processes.

and subsequent recalculation of energy contributions done in the year 2000. Prior to that date, the total emergy contribution to the geobiosphere that was used in calculating unit emergy values was 9.44 E24 sej/year. The increase in global emergy reference base to 15.83 E24 sej/year changes all the unit emergy values that directly and indirectly were derived from the value of global annual empower. Thus, unit emergy values calculated prior to that year are multiplied by 1.68 (the ratio of 15.83/9.44).

2.2 Average Emergy Unit Values for Main Global Processes

Table III calculates unit emergy values for some main flows of the earth. The total emergy input to the geobiosphere in solar emergy (15.83 E24 sej/yr from Table II) is divided by each of the global product's ordinary measure (number of joules or grams). The unit values that result are useful for other emergy evaluations where global averages can be used.

2.3 Temporary Emergy Inputs to the Geobiosphere

In the 19th and 20th centuries, the production and consumption processes of human civilization using the large emergy in the geologic stores of fuels and minerals reached a scale with global impact. Because these storages are being used much faster than they are being generated in geologic cycles, they are often called nonrenewable resources. They are actually very slowly renewed resources. Table IV summarizes these additional components of the global emergy budget.

At present, the emergy contribution through the human civilization to the geobiosphere from the slowly renewed resources is greater than the inputs from renewable sources. A symptom of this surge of temporary emergy inflow is the carbon-dioxide accumulation in the atmosphere that causes greenhouse effects that change the pattern and intensity of weather.

It is important to note that the unit emergy values given in Table III are average values for global processes. It is well understood that there is no single

TABLE III

Emergy of Products of the Global Energy System (after Odum et al., 2000)

Product and units	Emergy* E24 sej/year	Production units/year	Emergy/unit sej/unit
Global latent heat, J[a]	15.83	1.26 E24	12.6 sej/J
Global wind circulation, J[b]	15.83	6.45 E21	2.5 E3 sej/J
Global precipitation on land, g[c]	15.83	1.09 E20	1.5 E5 sej/g
Global precipitation on land, J[d]	15.83	5.19 E20	3.1 E4 sej/J
Average river flow, g[e]	15.83	3.96 E19	4.0 E5 sej/g
Average river geopotential, J[f]	15.83	3.4 E20	4.7 E4 sej/J
Average river chem. energy, J[g]	15.83	1.96 E20	8.1 E4 sej/J
Average waves at the shore, J[h]	15.83	3.1 E20	5.1 E4 sej/J
Average ocean current, J[I]	15.83	8.6 E17	1.8 E7 sej/J

*Main empower of inputs to the geobiospheric system from Table II not including nonrenewable consumption (fossil fuel and mineral use).

[a] Global latent heat = latent heat of evapotranspiration 1020 mm/year (1020 mm/year) (1000 g/m^2/mm) (0.58 kcal/g) (4186 J/kcal) (5.1 E14 m^2) = 1.26 E24 J/year.

[b] Global wind circulation, 0.4 watts/m^2 (Wiin-Nielsen and Chen, 1993) (0.4 J/m^2/sec) (3.15 E7 sec/year) (5.12 E14 m^2/earth) = 6.45 E21 J/year.

[c] Global precipitation on land = 1.09 E11 m^3/year (Ryabchikov, 1975) (1.09 E14 m^3) (1 E6 kg/m^3) = 1.09 E20 g/year.

[d] Chemical potential energy of rainwater relative to seawater salinity (1.09 E20 g/year) (4.94 J Gibbs free energy/g) = 5.19 E20 J/year.

[e] Global runoff, 39.6 E3 km^3/year (Todd, 1970) (39.6 E12 m^3/year) (1 E6 g/m^3) = 3.96 E19 g/year.

[f] Average river geopotential work; average elevation of land = 875 m (39.6 E12 m^3/year) (1000 kg/m^3) (9.8 m/sec^2) (875 m) = 3.4 E20 J/year.

[g] Chemical potential energy of river water relative to seawater salinity (3.96 E19 g/year) (4.94 J Gibbs free energy/g) = 1.96 E20 J/year.

[h] Average wave energy reaching shores, (Kinsman, 1965) (1.68 E8 kcal/m/year) (4.39 E8 m shore front)(4186 J/kcal) = 3.1 E20 J/year.

[i] Average ocean current: 5 cm/sec (Oort *et al.*, 1989); 2-year turnover time (0.5) (1.37 E21 kg water) (0.050 m/sec) (0.050 m/sec)/(2 year) = 8.56 E17 J/year.

TABLE IV

Annual Emergy Contributions to Global Processes Including Use of Resource Reserves (after Brown and Ulgiati, 1999)

Inputs and Units	Inflow (J/year)	Emergy/unit* (sej/unit)	Empower E24 sej/year
Renewable inputs[a]	—	—	15.8
Nonrenewable energies released by society			
Oil, J[b]	1.38 E20	9.06 E4	12.5
Natural gas (oil eq.), J[c]	7.89 E19	8.05 E4	6.4
Coal (oil eq.), J[d]	1.09 E20	6.71 E4	7.3
Nuclear power, J[e]	8.60 E18	3.35 E5	2.9
Wood, J[f]	5.86 E19	1.84 E4	1.1
Soils, J[g]	1.38 E19	1.24 E5	1.7
Phosphate, J[h]	4.77 E16	1.29 E7	0.6
Limestone, J[I]	7.33 E16	2.72 E6	0.2
Metal ores, g[j]	9.93 E14	1.68 E9	1.7
Total nonrenewable empower			34.3
Total global empower			50.1

Abbreviations: sej = solar emjoules; E3 means multiplied by 10^3; t = metric ton; oil eq. = oil equivalents.

*Values of solar emergy/unit from Odum (1996) and modified to reflect a global resource base of 15.83 E24 sej/year.

[a] Renewable Inputs: Total of solar, tidal, and deep heat empower inputs from Odum (1996).

[b] Total oil production = 3.3 E9 Mt oil equivalent (British Petroleum, 1997) Energy flux = (3.3 E9 t oil eq.) (4.186 E10 J/t oil eq.) = 1.38 E20 J/year oil equivalent.

[c] Total natural gas production = 2.093 E9 m^3 (British Petroleum, 1997) Energy flux = (2.093 E12 m^3) (3.77 E7 J m^3) = 7.89 E19 J/year.

[d] Total soft coal production = 1.224 E9 t/year (British Petroleum, 1997) Total hard coal production = 3.297 E9 t/year (British Petroleum, 1997) Energy flux = (1.224 E9 t/year) (13.9 E9 J/t) + (3.297 E9 t/year)(27.9 E9 J/t) = 1.09 E20 J/year.

[e] Total nuclear power production = 2.39 E12 kwh/year (British Petroleum, 1997); Energy flux = (2.39 E12 kwh/year)(3.6 E6 J/kwh) = 8.6 E18 J/year electrical equivalent.

[f] Annual net loss of forest area = 11.27 E6 ha/year (Brown et al., 1997) Biomass = 40 kg m^2; 30% moisture (Lieth and Whitaker, 1975) Energy flux = (11.27 E6 ha/year) (1 E4 m^2/ha) (40 kg m^2)(1.3 E7 J/kg)(0.7) = 5.86 E19 J/year.

[g] Total soil erosion = 6.1 E10 t/year (Oldeman, 1994; Mannion, 1995) Assume soil loss 10 t/ha/year and 6.1 E9 ha agricultural land = 6.1 E16/g/year (assume 1.0% organic matter), 5.4 kcal/g Energy flux = (6.1 E16 g)(.01) (5.4 kcal/g)(4186 J/kcal) = 1.38 E19 J/year.

[h] Total global phosphate production = 137 E6 t/year (USDI, 1996) Gibbs free energy of phosphate rock = 3.48 E2 J/g Energy flux = (137 E12 g)(3.48 E2 J/g) = 4.77 E16 J/year.

[i] Total limestone production = 120 E6 t/year (USDI, 1996) Gibbs free energy of phosphate rock = 611 J/g Energy flux = (120 E12 g)(6.11 E2 J/g) = 7.33 E16 J/year.

[j] Total global production of metals 1994: Al, Cu, Pb, Fe, Zn (World Resources Institute, 1996): 992.9 E6 t/year = 992.9 E12 g/year.

unit emergy value for any given product, since no two processes are alike. This also holds for the processes of the biosphere. For instance, there are many transformities for rain depending on location and even time of year. Precipitation varies with altitude, is affected by mountains, and depends on the weather systems in complex ways. The evaluations in Table III are for the whole earth with 70% ocean. If the land is regarded as a higher level in the hierarchical organization of the geobiosphere, then rain over the land represents a convergence of oceanic resources as well as those of the continents, and calculation of continental rain transformity includes all geobiosphere driving energy (see rows 3 and 4 in Table III). As a result, continental rainfall has a higher transformity compared to the global average. To carry this idea even farther, the rainfall in any particular location may have a higher or lower transformity depending on the source area and intensity of the solar energy driving the cycles that produce it.

3. UNIT EMERGY VALUES FOR FUELS AND SOME COMMON PRODUCTS

Unit emergy values result from emergy evaluations. Several tables of unit emergy values for some common materials and energy sources follow. In Table V, unit emergy values are given for primary nonrenewable energy sources along with the emergy yield ratio (EYR). In some cases, the EYR is based on only one

TABLE V

Unit Emergy Values for Primary Nonrenewable Energy Sources (After Odum, 1996, Updated)

Item	Transformity (Sej/J)	Sej/g	EYR
Plantation pine (*in situ*)	1.1 E4	9.4 E7	2.8
Peat	3.2 E4	6.7 E8	3.0
Lignite	6.2 E4		6.8
Coal	6.7 E4		8–10.5
Rainforest wood (chipped, trans.)	6.9 E4	4.1 E8	2.2
Natural gas	8.1 E4		6.8–10.3
Crude oil	9.1 E4		3.2–11.1
Liquid motor fuel	1.1 E5		2.6–9.0
Electricity	3.4 E5		—

TABLE VI

Unit Emergy Values for Some Common Products (After Odum, 1996, Updated)

Item	Transformity (Sej/J)	Specific emergy (Sej/g)
Corn stalks	6.6 E4	
Rice, high energy[a]	7.4 E4	1.4 E9
Cotton	1.4 E5	
Sugar (sugarcane)[b]	1.5 E5	
Corn	1.6 E5	2.4 E9
Butter	2.2 E6	
Ammonia fertilizer	3.1 E6	
Mutton	5.7 E6	
Silk	6.7 E6	
Wool	7.4 E6	
Phosphate fertilizer	1.7 E7	
Shrimp (aquaculture)	2.2 E7	
Steel[b]	8.7 E7	7.8 E9

[a] After Brown and McKlanahan (1996).
[b] After Odum and Odum (1983).

evaluation for instance plantation pine. In other cases, several evaluations have been done of the same primary energy but from different sources and presumably different technology, so a range of values are given. For instance, the range of values for natural gas (6.8 to 10.3) represent the difference between offshore natural gas (6.8) and on natural gas produce in the Texas field on shore (10.3). Obviously each primary energy source has a range of values depending on source and technology. By using data from typical production facilities (and actually operating facilities), the unit emergy values represent average conditions and can be used for evaluations when actual unit values are not known. If it is known that conditions where an evaluation is being conducted are quite different than the averages suggested here, then detailed evaluations of sources should be conducted.

Table VI lists the unit emergy values for some common products in the order of their transformity. Only a few products are given here, while many more evaluations leading to unit emergy values have been conducted and are presented in a set of emergy folios, published by the Center for Environmental Policy at the University of Florida (www.ees.ufl.edu/cep).

Calculations of emergy production and storage provide a quantitative basis for making choices about environment and economy. The unit energy values done previously aid new evaluations, since they may save time and energy of authors attempting other evaluations because they do not need to reevaluate all inputs from scratch. In light of this, a series of emergy folios have been published, and others planned, that provide data on emergy contents, the computations on which they were based, and comparisons (www.ees.ufl.edu/cep/emergydownloads.asp). To date there are five folios by as many authors, who take the initiative to make new calculations or assemble results from the extensive but dispersed literature in published papers, books, reports, theses, dissertations, and unpublished manuscripts. The tabulating of unit emergy values and their basis is the main purpose of the handbooks.

4. CASE STUDIES: EVALUATION OF ENERGY CONVERSION SYSTEMS

Several case studies of conversion systems are evaluated using embodied energy analysis and emergy analyses (hereafter EEA and EMA, respectively). The results from case studies are discussed jointly in the following sections of this article. Comparisons are made to highlight synergisms that support each other and differences that require EMA be used for a deeper understanding of energy systems performance. Evaluations of the case studies resulted from a joint research project between the authors with funding from the Italian Energy Agency (ENEA).

4.1 Description of the Power Systems

4.1.1 Oil-Powered Thermal Plant

A conventional oil-powered thermal plant (Fig. 4) mainly consists of a fuel storage area, boilers for

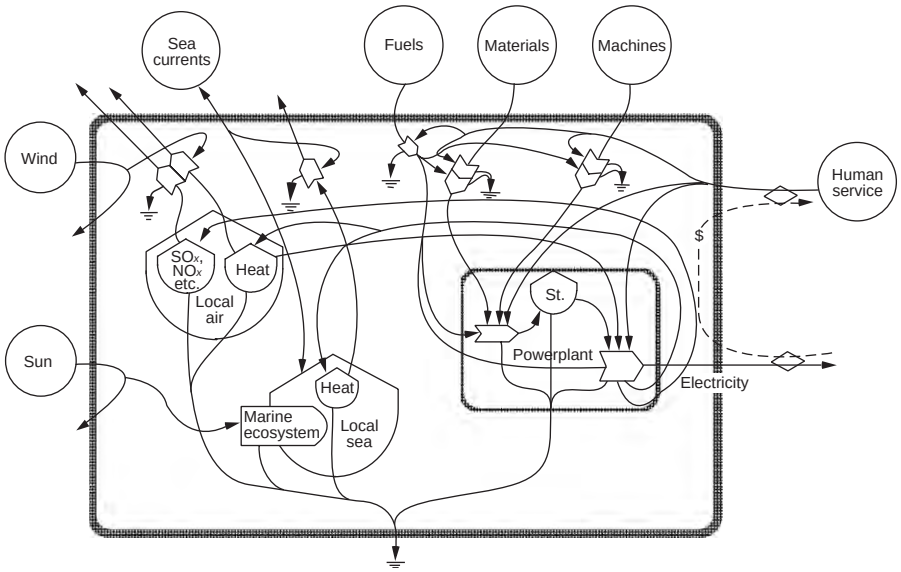


FIGURE 4 System diagram of a thermal power plant an its environment showing the use of nonrenewable energy as well as the use of the environment for cooling.

steam generation, steam turbines, electric generators, a set of electricity transformers from low to high voltage for easier transport to final users, one or more chimneys for the release of combustion gases to the atmosphere, and finally a set of heat exchangers and pumps to cool the plant off by means of river or sea water (44 m^3/s) and condense the steam. Filters and scrubbers for partial uptake of pollutants at the chimney mouth as well as fuel desulfurization devices are also used. The data for the evaluation were provided by the Italian National Electric Company (ENEL) and relate to the actual average performance of the Piombino 1280 MWe power plant (Tuscany, Italy). The plant consists of four 320 Mwe boiler-turbine-generator groups, cooled by the seawater from the nearby Mediterranean Sea. All inputs were listed and grouped according to their material components (steel, copper, etc.), then suitable energy and emergy intensity coefficients were used, according to previous studies available in the scientific literature. Labor and services are also accounted for in the emergy analysis only. Waste heat and chemicals released to the environment (seawater and atmosphere) were quantified and the amount of environmental services (cooling water, wind, etc.) needed for dispersal, dilution and abatement is calculated by means of the emergy accounting procedure. Fig. 4 shows the direct and indirect annual environmental flows supporting the construction of the plant structure (assuming a lifetime of 25 years) as well as its daily operation. The plant started to be fully operative in the year 1979, which means that it is not far from the end of its life cycle and may require a significant upgrade investment or final decommissioning and replacement by a more modern technology.

4.1.2 Wind Turbine

The investigated case study (Fig. 5) consists of a 2.5 MW wind powered field, composed with 10 single-blade, 250 kW, wind turbines M30-A, sited in southern Italy (Casone Romano). The wind field, installed in the year 1995 by the Riva Calzoni Company, is presently operated by ENEL. The distance between two generators is 150 m. Each turbine is located on the top of a 33 m support tower. Operating wind speeds are between 4 and 25 m/s. For wind speeds outside of these extremes, the turbine automatically stops. The wind field does not require any local labor for its daily activity and is directly managed from the operating company by means of remote control. The energy systems diagram of a wind power plant is shown in Fig. 5.

4.1.3 Geothermal Plant

To optimize the exploitation of the rich geothermal fields available in central Italy (Tuscany), the National Electric Company designed a 20 MW power module to be efficiently and cost-effectively installed in several sites of the geothermal area (Fig. 6). The main reason supporting the development of such a module is the need for the reduction of the time between the discovery of a new geothermal site and

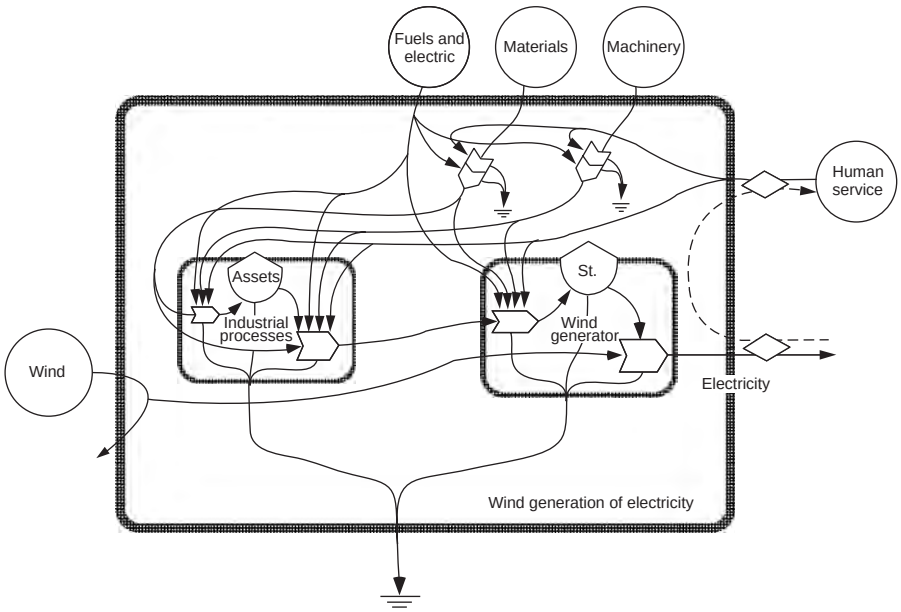


FIGURE 5 System diagram of wind power plant showing the industrial processes required to produce the generator and associated structure and the use of wind energy in the production of electricity.

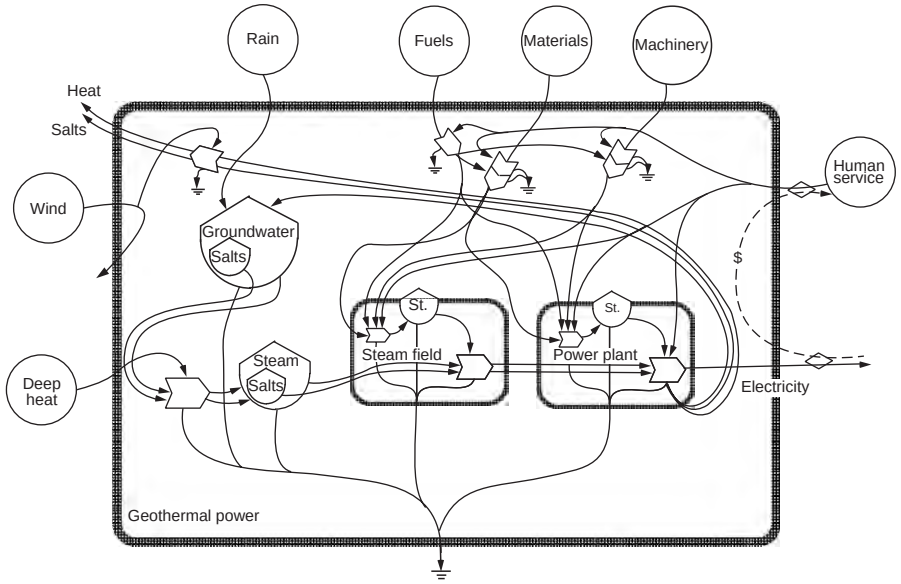


FIGURE 6 System diagram of a geothermal power plant showing the use of nonrenewable energies in developing the steam field and power plant, the use of geothermal steam, and release of heat and salts to the larger environment.

its actual exploitation. The module essentially consists of a turbine-generator group, receiving high-enthalpy fluids from nearby wells. A fraction of the extracted steam is then released to the atmosphere, through three cooling towers. The remaining fraction is, instead, condensed and reinjected into the underground reservoir to avoid their water depletion. The steam pipelines are made with steel and are insulated with rockwool and aluminum. Instead, the reinjection lines are made with glass fiber. The data used for the case study were provided by ENEL and relate the Cornia 2 power plant, sited close to Pisa and installed in the year 1994. No local labor is required. The plant is 24 hours per day distance-operated by the

company's operative station sited in Larderello, the site where the first electric plant running on geothermal heat was installed in 1904.

Geothermal fluids carry and release significant amounts of hydrogen sulphide, carbon dioxide, radioactive radon, arsenic, mercury, nitrogen, and boron. This involves plant corrosion problems as well as pollution problems in the area. For this reason, several devices have been designed for the uptake and abatement of the most dangerous chemicals. The plant as well as the environmental and technological flows supporting its activity are described by the systems diagram drawn in Fig. 6.

4.1.4 Hydroelectric Plant

The case study (systems diagram shown in Fig. 7) concerns the reservoir-dam electric plant Pollino Nord, sited in southern Italy. Its activity started in the year 1973. The reservoir has a capacity of 1000 m^3 and is located at about 600 m above sea level. Water flows out of the dam through an iron pipeline reaching the turbines located downstream, at about 50 m above sea level. Two 51 MW turbines with vertical axis convert the water energy into electricity, generating a total actual electric power of about 85 MW. The plant started its activity in 1973. No labor is required for plant operation. It is distance-operated from the ENEL headquarters in Catanzaro (southern Italy).

4.1.5 Bioethanol Production

Figure 8 describes the main steps of the process: corn production, harvesting and transport to plant, and industrial processing. Corn is processed to ethanol and distillers dried grains with solubles (DDGS), to be used as animal feed. Agricultural residues were considered as a partial source of process heat (to substitute for some of the coal), so that a lower fossil energy input can be charged to the industrial phase. However, harvesting agricultural residues depletes the amount of nutrients in soil and requires significant energy expenditure for their replacement. Soil erosion is also accelerated, with significant consequences on soil fertility. Therefore, the fraction of agricultural residues that can be actually harvested is small and depends on a variety of parameters that are not easy to determine. The evaluation was conducted assuming that approximately 70% of agricultural residues were harvested, taking some of the above mentioned consequences (loss and replacement of nutrients in soil) into account. Corn production data refer to Italian agricultural standards, while corn-to-ethanol industrial conversion refers to average available conversion technologies.

4.1.6 Hydrogen from Steam Reforming of Natural Gas

Steam reforming of hydrocarbons (mainly natural gas) has been the most efficient, economical, and

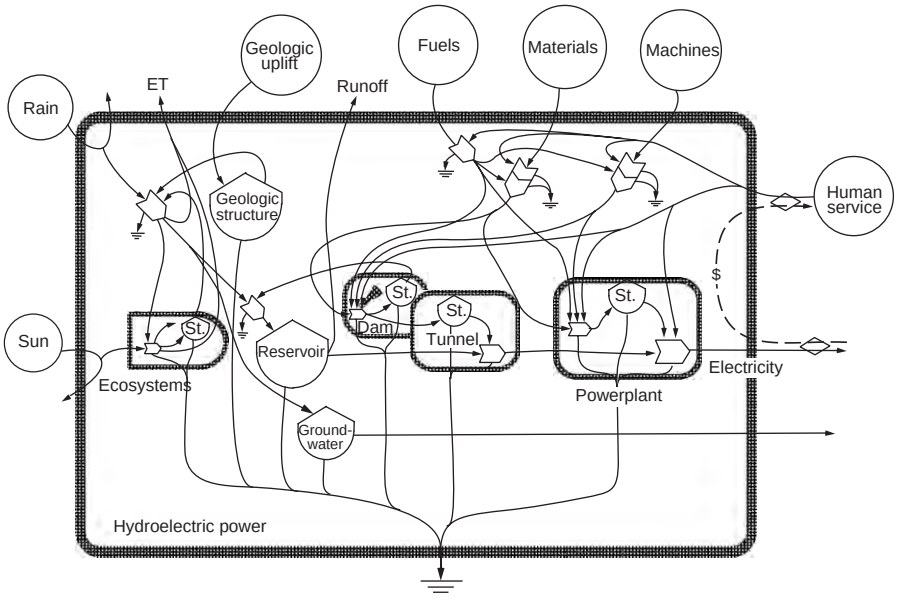


FIGURE 7 Systems diagram of a hydroelectric power plant showing the use of nonrenewable energy materials machines and human service in the construction operation and maintenance of the plant as well as the reliance of the system on geologic formation of the watershed and input of rain.

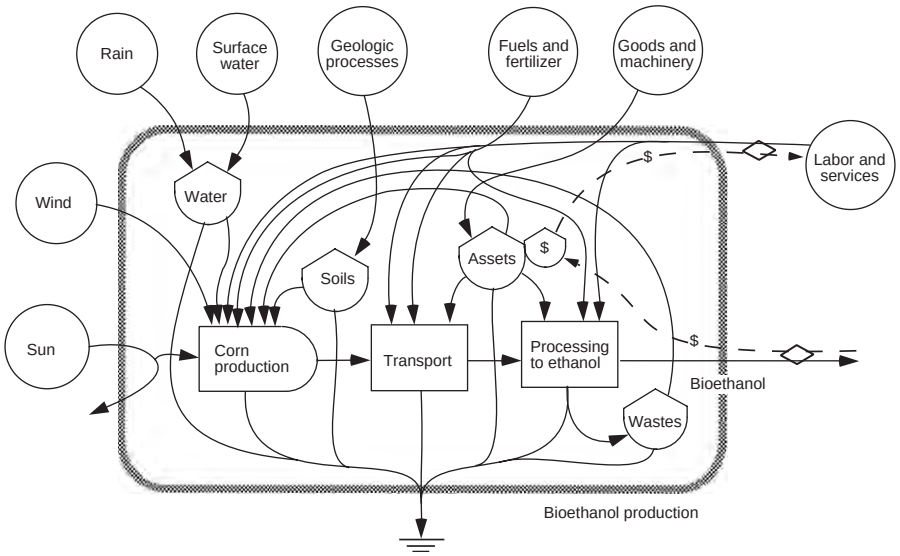


FIGURE 8 Systems diagram of bioethanol production showing the use of nonrenewable resources in the construction of the processing plant, transportation, and corn production as well as the environmental, renewable energies of sun, wind, and water.

widely used process for hydrogen production. An aggregated energy systems diagram of a conventional steam reforming process is shown in Fig. 9A. The feedstock (i.e. natural gas) is mixed with process steam and reacts over a nickel-based catalyst contained inside a system of alloyed steel tubes. To protect the catalyst, natural gas has to be desulfurized before being fed to the reformer. The following reactions take place in the reformer:

$$CH_4 + H_2O \rightarrow CO + 3H_2$$
$$CO + H_2O \rightarrow CO_2 + H_2.$$

The output of the reformer is fed to a series of shift reactors and a pressure swing adsorption unit, where hydrogen is separated and purified to over 99% purity. The reforming reaction is strongly endothermic and energy is supplied by combustion of natural gas or fuel oil. The reaction temperature is usually in the range 700 to 900°C.

4.1.7 Hydrogen from Water Electrolysis

Figure 9B shows an aggregated view of the hydrogen/water electrolysis process. Alkaline water electrolysis is a well-established industrial process in which electricity is used to split water into its component elements, generating hydrogen in a 1:1 molar ratio, with cell efficiencies in the order of 80 to 90%. The following reactions take place at the electrodes of an electrolysis cell (called an electrolyzer) filled with a suitable electrolyte (aqueous solution of KOH, NaOH, or NaCl) upon the application of a potential:

Cathode reaction: $2H_2O\ (l) + 2e^- \rightarrow H_2(g) + 2OH^-\ (aq)$

Anode reaction: $2OH^-(aq) \rightarrow \frac{1}{2}O_2(g) + H_2O(l)$

Overall cell reaction: $H_2O\ (l) \rightarrow H_2(g) + O_2\ (g)$

An electrolysis plant can operate over a wide range of capacity factors and is convenient for a wide range of operating capacities, which makes this process interesting for coupling with both renewable and nonrenewable energy sources. The needed electricity can be purchased from the grid (an interesting option in the case of excess electricity production from hydroelectric plants, for example) or produced on site. In this article, the electrolysis is assumed driven by the electricity generated by the oil-powered and wind-powered plants described earlier.

4.1.8 Ocean Thermal Energy Conversion (OTEC)

OTEC is an energy technology that converts indirect solar radiation to electric power (Fig. 10). OTEC systems use the ocean's natural thermal gradient—the fact that the ocean's layers of water have different temperatures—to drive a power-producing cycle. Commercial ocean thermal energy conversion plants must be located in an environment that is stable enough for efficient system operation. The temperature of the warm surface seawater must differ about 20°C (36°F) from that of the cold deep water that is no more than about 1000 meters (3280 feet) below the surface. The natural ocean thermal gradient necessary for OTEC operation is generally

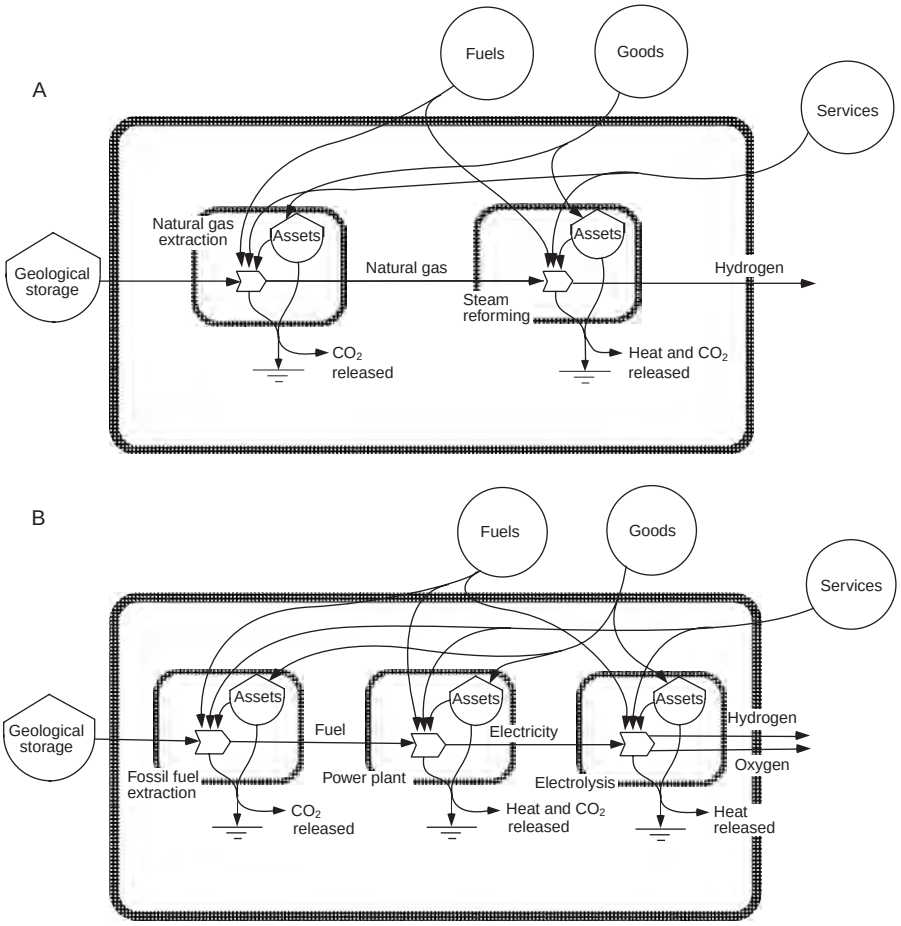


FIGURE 9 (A) Aggregated systems diagram of hydrogen production from steam reforming and (B) electrolysis.

found between latitudes 20 °N and 20° S. Within this tropical zone are portions of two industrial nations—the United States and Australia—as well as 29 territories and 66 developing nations. Of all these possible sites, tropical islands with growing power requirements and a dependence on expensive imported oil are the most likely areas for OTEC development. Electricity generated by plants fixed in one place can be delivered directly to a utility grid. A submersed cable would be required to transmit electricity from an anchored floating platform to land. Moving ships could manufacture transportable products such as methanol, hydrogen, or ammonia on board. The OTEC data for the evaluation are from Odum. He investigated the OTEC technology, performing both an energy and emergy assessment. His conclusions, based on emergy accounting, were that although OTEC shows a small net emergy yield it is unlikely to be competitive with fossil powered technologies for electricity, due to the excessive dilution of solar energy resulting into a very small heat gradient of marine water.

4.2 Methods

Apart from the OTEC plant, the case studies represent real plants, evaluated on the basis of the electricity actually generated and the fuel actually used. Therefore, calculated performance indicators may appear worse than those theoretically obtainable on the basis of nameplate power and efficiency, although the order of magnitude does not change significantly.

4.2.1 Data and Calculations

Unless specified differently, a lifetime of 25 years is assumed for each of the investigated plants or assets. Data used include construction materials, source energy, and labor in construction, as well as operation and maintenance, environmental inputs,

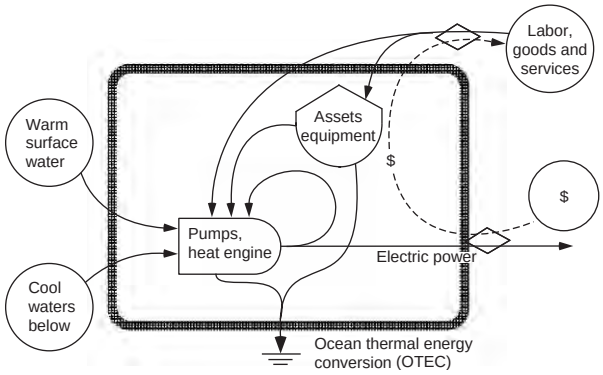


FIGURE 10 Aggregated systems diagram of ocean thermal energy conversion showing the use of labor, goods, and services (nonrenewable) and differential in surface and deep ocean water temperature.

and process outputs including electricity and pollutants. Input data for main plant components were kindly supplied by the companies producing and operating them.

All plants are analyzed using the same procedures as outlined earlier. Input data for each item (row) in the evaluation tables are the actual operational data of each plant and are converted into appropriate units using specific calculation procedures, most of which are described by Odum in his text *Environmental Accounting*. Input data to each investigated process are multiplied by suitable conversion coefficients (unit oil equivalents, g_{oil}/unit of item, for energy evaluation; transformities, sej/unit of item, for emergy evaluation), to yield energy and emergy values associated with each input item. Energy values (expressed as grams of oil equivalent) are then converted into grams of released carbon dioxide, by means of an average oil-to-carbon dioxide conversion coefficient derived from typical reaction stoichiometry. Energy, emergy, and CO_2 flows are then summed, yielding a total that is divided by the amount of the product of each plant to yield its energy cost, its transformity, and its associated carbon emissions. Finally, data are used to calculate several other energy-, emergy-, and carbon-flow indicators to obtain an overall picture of plant performance. All inputs to the electric production systems are expressed on a yearly basis. Fixed capital equipment, machinery, buildings, and so on are divided by their estimated useful life (10 to 30 years, depending on the case). Outputs are also expressed on an annual basis. CO_2 production is calculated for direct fuel combustion in thermal plants and estimated for goods and human services using multipliers of indirect energy consumption for their production. These multipliers are derived from a standardized comparison of literature in this field.

4.2.2 Performance Indicators

EEA typically evaluates energy sources and the outputs from process in their heat equivalents. Two indicators commonly used in EEA are calculated for all the investigated case studies. The first one is related to CO_2 production and global warming potential and is the amount of CO_2 released per megawatt of energy delivered. The second indicator is the output/input energy ratio (also called energy return on investment, EROI, according to Hall *et al.*). It is calculated as the ratio of heat equivalent energies of outputs and input flows and is considered as a measure of the first law energy efficiency of a process. The total input energy in this study is derived by multiplying all goods and materials consumed in the construction (divided by the average life span of the structure and equipment) and the annual maintenance of plants by energy equivalent multipliers, and adding the annual energy consumed directly and indirectly.

The output as electricity is converted to joules using the standard conversion of 3.6×10^6 J/kWh, while the masses of bioethanol and hydrogen are multiplied by their higher heating value (J/g). CO_2 releases are evaluated by multiplying the total input energy (sum of energy used directly and indirectly) by a standard stoichiometric conversion, 3.2 g CO_2/g oil equivalent. Renewable energy plants (wind, hydro, and geothermal) obviously do not burn fossil fuels directly, so their direct CO_2 emissions are lower (zero for the hydro and wind plants, but not negligible for the geothermal power plant due to underground CO_2 contained in the geothermal fluid extracted).

EMA indices are calculated in a similar way. Each item in Table VII (bioethanol) and Table VIII (wind turbine) as well as in all the other tables related to the investigated case studies is multiplied by an appropriate transformity to get the emergy input from the item to the process. The total emergy (a measure of the total environmental support or ecological footprint to the process) is then calculated by summing all inputs according to the previously described emergy algebra. Finally, a new transformity for the process output is calculated by dividing the total emergy by the exergy (sometimes the mass) of the product. Further emergy-based indicators (emergy yield ratio, environmental loading ratio, emergy index of sustainability, emergy density) are also calculated.

4.3 Results and Discussion

Comparing EEA- and EMA-based performance indicators for each kind of energy conversion device makes it possible to ascertain what pieces of information can be obtained from each one of the two approaches and to better understand the space-time scale difference between them.

A careful reading of Tables VII to X provides a rich set of information. Tables VII and VIII are given as examples of how emergy calculations are performed. Table VII refers to bioethanol production from corn (Fig. 8), while Table VIII refers to a 2.5 MW wind electricity production field (Fig. 5). Tables IX and X provide a comparison among energy-, carbon-, and emergy- based indicators in all the conversion patterns investigated. Table IX refers to electricity generation, while Table X refers to selected energy carriers other than electricity. Conversion factors used and references are clearly listed in each table.

4.3.1 Energy and Emergy Evaluations

Emergy evaluations of bioethanol production from corn (Table VII) and electricity generation from wind (Table VIII) are given as examples for a clear description of the emergy approach. These two case studies do not focus on the best available technologies and it is possible that other wind turbines or biomass-to-ethanol processes exist showing better performances. These two cases are used as more or less average-technology case studies. The same rationale applies to the other case studies summarized in Tables IX and X but not shown in detail here.

Renewable flows to bioethanol production from corn (Table VII) account for about 15.3% of the total emergy driving the process. They can be identified as the rain driving the agricultural phase and the wind dispersing the pollutants that come out of the industrial step. Electricity and fuels (including coal) account for about 17.4%, goods and machinery for 42.6%, while labor and services represent 24.7% of the total emergy input. Goods and machinery are therefore the main emergy input to the process, due to the environmental support embodied in their manufacture from raw materials.

Environmental flows as well as labor and services are not included in the conventional energy accounting, therefore a comparison between EEA and EMA requires the percentages be calculated taking these flows out of the total emergy. In so doing, fuels and electricity come out to be respectively the 29% of total emergy and the 45% of total energy, while goods and machinery account for 71% of total emergy and 55% of total energy. This means that EEA overestimates the role of direct energy inputs and underestimates the role of energy flows embodied in goods and machinery. A similar calculation can be performed for the wind turbine by using data presented in Table VIII. As far as wind plant is concerned, it can be noted that wind emergy accounts for about 82% of the total emergy input, while labor and services (4%) and goods and machinery (14%) play a much smaller role. A comparison with energy data is less significant here, since no direct energy inputs are required. It is crystal clear that EMA offers a more comprehensive picture of the whole process by taking into account the hidden environmental flows that support a given process and by evaluating their role on the scale of the biosphere, where they are compared by means of a common measure of environmental quality.

4.3.2 Embodied Energy Performance Indicators

The EEA results are synoptically shown in Tables IX and X. The EROIs of electricity generation range from a low 0.30 of the oil-powered plant to a high 23.7 of the hydroelectric plant. The geothermal and wind conversions also show high energy returns (respectively 20.7 and 7.7), while the OTEC plant only shows an energy return equal to 3.8, due to the low thermal gradient between deep and surface water temperatures. Similarly, Table X shows EROIs in the range 0.24 to 6.1 for energy carriers different than electricity (bioethanol and hydrogen). As expected, the lower value refers to hydrogen production via electrolysis with electricity generated by an oil-powered, thermal plant similar to the one described in Table IX. Instead, the highest EROI is shown by hydrogen generated via electrolysis by means of wind-powered electricity.

The net CO_2 release shows similar patterns, with higher values when direct fossil fuel combustion is involved and lower values when renewable sources drive the process, as expected. In the latter case, only indirect emissions can be calculated, linked to fuel use for the production of components and for machinery construction and operation. The surprising high emission from geothermal electricity is due to the CO_2 content of geothermal fluids.

The better performance of processes driven by renewables is clearly due to the fact that the EROI compares only commercial energies (i.e., calculates the ratio of the thermal energy delivered to the fossil or fossil equivalent energy that is provided to the

TABLE VII

Emergy Analysis of Bioethanol Production

Item	Unit	Amount (unit/ha*year)	Unit emergy values (seJ/unit)	Ref. for UEV	Solar emergy (seJ/ha*year)
Agricultural phase (corn production)					
Renewable inputs					
Sunlight	J	5.50E + 13	1.00E + 00	[1]	5.50E + 13
Rain water (chemical potential)	J	3.07E + 10	3.06E + 04	[2]	9.39E + 14
Earth cycle	J	3.00E + 10	1.02E + 04	[2]	3.05E + 14
Nonrenewable inputs					
Organic matter in topsoil used up	J	3.24E + 09	1.24E + 05	[2]	4.02E + 14
Nitrogen fertilizer	g	1.69E + 05	6.38E + 09	[2]	1.08E + 15
Phosphate fertilizer	g	8.20E + 04	6.55E + 09	[2]	5.37E + 14
Insecticides, pesticides, and herbicides	g	5.38E + 03	2.49E + 10	[3]	1.34E + 14
Diesel	J	6.67E + 09	1.11E + 05	[2]	7.40E + 14
Lubricants	J	1.64E + 08	1.11E + 05	[2]	1.82E + 13
Gasoline	J	3.00E + 00	1.11E + 05	[2]	3.33E + 05
Water for irrigation	J	1.36E + 09	6.89E + 04	[3]	9.37E + 13
Electricity for irrigation	J	2.00E + 09	2.52E + 05	[4]	5.04E + 14
Agricultural machinery (mainly steel)	g	1.36E + 04	1.13E + 10	[3]	1.53E + 14
Seeds	g	1.62E + 04	5.88E + 04	[7]	9.53E + 08
Human labor	years	1.30E-02	6.32E + 16	[6]	8.21E + 14
Annual services	US $	8.90E + 02	2.00E + 12	[6]	1.78E + 15
Additional inputs for harvest of 70% residues to be used as process energy					
Nitrogen in residues harvested	g	7.88E + 04	6.38E + 09	[2]	5.03E + 14
Phosphate in residues harvested	g	1.82E + 04	6.55E + 09	[2]	1.19E + 14
Diesel for residues	J	9.03E + 03	1.11E + 05	[2]	1.00E + 09
Machinery for residues	g	2.59E + 03	1.13E + 10	[3]	2.92E + 13
Labor for residues	years	1.39E-03	6.32E + 16	[6]	8.78E + 13
Additional services	$	8.90E + 01	2.00E + 12	[6]	1.78E + 14
Industrial phase (conversion of corn to bioethanol)					
Renewable inputs for the dilution of airborne and waterborne pollutants					
Wind	J	2.20E + 11	2.52E + 03	[2]	5.54E + 14
Nonrenewable inputs					
Diesel for transport	J	6.76E + 08	1.11E + 05	[2]	7.50E + 13
Transport machinery (mainly steel)	g	1.22E + 04	1.13E + 10	[3]	1.37E + 14
Plant machinery (mainly steel)	g	4.41E + 04	1.13E + 10	[3]	4.96E + 14

continues

Table VII continued

Item	Unit	Amount (unit/ha*year)	Unit emergy values (seJ/unit)	Ref. for UEV	Solar emergy (seJ/ha*year)
Cement in plant construction	g	7.84E + 04	3.48E + 09	[5]	2.73E + 14
Additional coal for hot water/steam generation	J	8.10E + 07	6.72E + 04	[2]	5.44E + 12
Process electricity	J	1.41E + 09	2.52E + 05	[4]	3.55E + 14
Process and cooling water	J	7.99E + 07	6.89E + 04	[3]	5.50E + 12
Gasoline denaturant	J	4.89E + 08	1.11E + 05	[2]	5.42E + 13
Ammonia	g	3.56E + 01	6.38E + 09	[2]	2.27E + 11
Lime	g	9.27E + 00	1.68E + 09	[2]	1.56E + 10
Electricity running the waste water treatment plant	J	4.35E + 07	2.52E + 05	[4]	1.10E + 13
Labor	years	1.68E-03	6.32E + 16	[6]	1.06E + 14
Annual capital cost and services	US $	2.22E + 02	2.00E + 12	[6]	4.44E + 14
Main product of industrial phase					
Ethanol produced, without services	J	5.65E + 10	1.28E + 05	[7]	7.22E + 15
Ethanol produced, with services	J	5.65E + 10	1.73E + 05	[7]	9.78E + 15

process). Energies that are not purchased, but instead are provided for free by nature (e.g., the solar energy driving the wind or the solar energy that drives the photosynthesis for corn production) are not accounted for by EEA and therefore are not reflected in the EROI or in the CO_2 release.

4.3.3 Emergy-Based Performance Indicators

A series of emergy performance indicators are given in Tables IX and X for selected energy conversion devices. Some of these processes, driven by intensive use of primary nonrenewable resources, are secondary energy conversions, while others (wind, hydro and geothermal) should be considered primary energy exploitation patterns. Their values can be usefully compared to those of the primary fuels (oil, coal, natural gas) shown in Table X. Electricity generation in Table IX shows transformities that are, as expected, process specific and range from the low value of about 1.1×10^5 seJ/J for hydro- and wind-electricity up to much higher values around 3.5×10^5 seJ/J for geothermal and oil-powered electricity. The role of labor and services in these processes can be quantified as about 5 to 10% of the transformity. This means that the main contributions to any electricity generation process are the driving forces of fuels or renewable sources, followed by the emergy associated with plant components. As a consequence, a significant lowering of the total emergy demand cannot be expected by simply moving the plant to countries with lower cost of labor, as has become very common with goods production. The OTEC plant combines very small inputs in the form of solar energy via the temperature gradient of the oceans and technological equipment and, instead, a significant input in the form of financial investment, labor, and services, to which a huge emergy is associated. This makes this plant too dependent on nonrenewable, purchased emergy inputs, very high compared to the locally available renewable flow. The transformity of OTEC electricity is around 1.7×10^5 seJ/J, but declines by more than one order of magnitude when labor and services are not accounted for.

Transformities of energy delivered in the form of energy carriers other than electricity (Table X) show similar orders of magnitude. Hydrogen from water electrolysis shows higher or lower transformities depending respectively on oil- or wind-generated electricity. Hydrogen from steam reforming of

TABLE VIII

Emergy Accounting of Wind Electricity Production in Italy[*]

Item	Unit	Amount	Unit emergy value (seJ/unit)	Ref. for UEV	Solar emergy (seJ)
Direct renewable inputs					
Wind	J	4.85E + 13	2.52E + 03	[2]	1.22E + 17
Nonrenewable and purchased inputs					
Concrete (basement)	g	5.62E + 06	2.59E + 09	[1]	1.45E + 16
Iron (machinery)	g	2.65E + 05	2.50E + 09	[2]	6.63E + 14
Steel (machinery)	g	8.61E + 05	5.31E + 09	[2]	4.57E + 15
Pig iron (machinery)	g	1.56E + 05	5.43E + 09	[2]	8.47E + 14
Copper	g	8.76E + 04	3.36E + 09	[3]	2.94E + 14
Insulating and miscellaneous plastic material	g	1.01E + 04	2.52E + 09	[4]	2.55E + 13
Lube oil	J	3.10E + 09	1.11E + 05	[5]	3.44E + 14
Labor	years	8.33E-02	6.32E + 16	[6]	5.26E + 15
Services	US $	2.94E + 01	2.00E + 12	[6]	5.88E + 13
Electricity generated					
Electricity, with labor and services	J	1.35E + 12	1.11E + 05	[7]	1.49E + 17
Electricity, without labor and services	J	1.35E + 12	1.06E + 05	[7]	1.44E + 17

[*]Data on a yearly basis; 2.5 MW plant sited at Casone Romano, Foggia, Italy.

References for transformities (Tables VII and VIII):

[1] Brown and Buranakarn (2003).
[2] Bargigli and Ulgiati (2003).
[3] Lapp (1991).
[4] Average estimate based on selected case studies.
[5] Odum (1996).
[6] Ulgiati (2003).
[7] From calculation performed in this work.

natural gas shows a relatively low transformity. Unfortunately, only the value without services was calculated in the steam-reforming case study. However, since steam-reforming is a well-known technology, not particularly labor intensive, a dramatic increase of its transformity should not be expected when adding the emergy value of labor and services. The comparison with the transformities of primary sources in Table IV clearly shows that transformities depend on the length of the manufacturing chain. More process steps require more emergy inputs and may decrease the actual energy delivered at the end of the chain. Oil, coal, natural gas, hydro- and wind-electricity, and hydrogen from steam reforming all show transformities on the order of 10^4 seJ/J. However, when technologically heavy steps are added to further process the primary energy source, transformities grow by an order of magnitude to 10^5 seJ/J. It can be clearly understood that if bioethanol were used to generate electricity or if electricity from oil were used to refine metals, transformities would become even greater.

The other emergy indicators complement this picture, by showing low EYRs for oil-powered and OTEC electricity as well as for bioethanol and hydrogen from thermal electricity. Not surprisingly, all the other energy conversions have EYRs around 5. Similarly, high ELRs characterize oil-based patterns, as well as bioethanol and OTEC electricity, while very low ELRs (below 1) are shown by the processes relying to a larger degree on renewables. As a consequence, the aggregated sustainability measure expressed by the EIS ranges from a low 0.06 for oil-based processes up to a significantly high 25 for wind-based ones. Again, the OTEC plant shows a very low performance, quantified by an EIS lower than 0.01. However, let's recall that the evaluation of OTEC is based on a feasibility study and not on an actually existing plant.

4.3.4 Comparison of EEA and EMA Results

Results presented in Tables IX and X allow further considerations in addition to the already discussed features of the two approaches. Both EEA and EMA

TABLE IX

Energy and Emergy-Based Indicators for Selected Electricity Power Plants

	Item	Unit	Oil plant (1280 MWe)[a]	Wind (2.5 MWe)[a]	Geothermal (20 MWe)[a]	OTEC 4.57 MWe[b]	Hydro (85 Mwe)[a]
Energy indicators							
E_{out}	Total electric energy delivered per year	J/year	2.35E + 16	1.35E + 13	3.28E + 14	1.19E + 14	3.94E + 14
E_{In}	Total energy invested per year	J/year	7.84E + 16	1.76E + 12	1.58E + 13	3.12E + 13	1.66E + 13
GWP	CO_2 released/unit of energy delivered[c]	g/MJ	260.22	10.17	182.32	20.43	3.29
ER	Energy return on investment (energy output/input)		0.30	7.67	20.76	3.82	23.73
Emergy indicators							
R_1	Renewable input from the local scale	seJ/year	3.43E + 20	1.22E + 18	5.63E + 19	1.13E + 17	2.84E + 19
R_2	Renewable input from the global scale[d]	seJ/year	1.72E + 21	0.00E + 00	2.64E + 19	0.00E + 00	0.00E + 00
N	Locally nonrenewable input[e]	seJ/year	0.00E + 00	0.00E + 00	7.74E + 18	0.00E + 00	7.47E + 18
F	Purchased plant inputs, including fuel	seJ/year	5.71E + 21	2.13E + 17	1.55E + 19	7.73E + 17	4.42E + 18
L	Labor and services[f]	seJ/year	5.36E + 20	5.33E + 16	4.17E + 18	1.95E + 19	3.73E + 18
Y	Yield ($R_1 + R_2 + N + F$), without labor and services	seJ/year	7.78E + 21	1.44E + 18	1.06E + 20	8.85E + 17	4.03E + 19
Y_L	Yield ($R_1 + R_2 + N + F + L$), with labor and services	seJ/year	8.32E + 21	1.49E + 18	1.10E + 20	2.04E + 19	4.40E + 19
Indices (including services for fuel supply and plant manufacturing							
Tr_1	Solar transformity of electricity, with labor and services	sej/J	3.54E + 05	1.10E + 05	3.35E + 05	1.71E + 05	1.12E + 05
Tr_2	Solar transformity of electricity, w/out labor and services	sej/J	3.31E + 05	1.06E + 05	3.23E + 05	7.44E + 03	1.02E + 05
EYR	Emergy yield ratio, $EYR = (Y_s)/(F + L)$		1.33	5.59	5.60	1.01	5.41
ELR	Environmental loading ratio, $ELR = (R_2 + N + F + L)/R_1$		23.26	0.22	0.96	180.00	0.55
ED	Emergy density, $ED = (Y_L)$/area of plant	sej/m^2	4.33E + 15	2.12E + 12	3.54E + 14	n.a.	2.84E + 13
EIS	EYR/ELR		0.057	25.64	5.87	0.006	9.84

n.a., not available.

[a] Brown and Ulgiati, 2002, modified.

[b] Data from H. T. Odum (2000b), partially modified by the authors.

[c] Global warning potential (GWP), includes direct CO_2 release from fuel combustion (if any), indirect CO_2 release embodied in inputs, and (for geothermal plant only) CO_2 from underground water, released through the cooling towers.

[d] Indirect environmental inputs from the larger scale, needed to dilute heat and chemicals that are released by the plant.

[e] Includes only locally extracted fuel delivered to plant, as well as underground water and topsoil used up (if any).

[f] Labor accounts for hours actually worked in the investigated process. Services include human labor in all steps prior than plant use, quantified as the emergy associated to the monetary cost (cost of item, $, x country emergy/GNP ratio, seJ/$).

TABLE X

Energy and Emergy-Based Indicators for Selected Nonelectric Energy Carriers

	Item	Unit	Bioethanol from corn[a]	Hydrogen from water electrolysis (oil powered)[b]	Hydrogen from water electrolysis (wind powered)[b]	Hydrogen from steam reforming of natural gas[c]
Energy indicators						
E_{out}	Total energy delivered per year	J/year	5.65E + 10	1.89E + 16	1.08E + 13	6.32E + 15
E_{in}	Total energy invested per year	J/year	3.77E + 10	7.84E + 16	1.77E + 12	8.18E + 15
GWP	CO_2 released/unit of energy delivered[d]	g/MJ	52.05	323.56	12.78	100.96
ER	Energy return on investment (energy output/input)		1.50	0.24	6.10	0.77
Emergy indicators						
R_1	Renewable input from the local scale	seJ/year	9.37E + 14	3.43E + 20	1.22E + 18	
R_2	Renewable input from the global scale[e]	seJ/year	1.08E + 15	1.73E + 21	0.00E + 00	
N	Locally nonrenewable input[f]	seJ/year	4.02E + 14	0.00E + 00	0.00E + 00	
F	Purchased plant inputs, including fuel	seJ/year	5.33E + 15	5.70E + 21	2.12E + 17	6.69E + 20
L	Labor and services[g]	seJ/year	3.65E + 15	5.48E + 20	6.85E + 16	
Y	Yield ($R_1 + R_2 + N + F$), without labor and services	seJ/year	7.75E + 15	7.77E + 21	1.43E + 18	6.69E + 20
Y_L	Yield ($R_1 + R_2 + N + F + L$), with labor and services	seJ/year	1.14E + 16	8.32E + 21	1.50E + 18	6.69E + 20
Indices (including services for fuel supply and plant manufacturing)						
Tr_1	Solar transformity of product, with labor and services	sej/J	2.02E + 05	4.40E + 05	1.39E + 05	
Tr_2	Solar transformity of product, w/out labor and services	sej/J	1.37E + 05	4.11E + 05	1.33E + 05	1.06E + 05
EYR	Emergy yield ratio, $EYE = (Y_s)/(F + L)$		1.27	1.33	5.36	
ELR	Environmental loading ratio, $ELR = (R_2 + N + F + L)/R_1$		11.16	23.26	0.23	
ED	Emergy density, $ED = (Y_L)$/area of plant	sej/m^2	1.12E + 12	4.33E + 15	2.13E + 12	
EIS	EYR/ELR		0.11	0.057	23.41	

[a] Ulgiati (2001), modified. Use of agricultural residues as source heat and an energy credit for coproducts (DDGS) accounted for.

[b] Calculated in this work. Performance coefficients for oil and wind powered electricity production from Table II. Water electrolysis performance characteristics from Bargigli *et al.* (2002).

[c] Bargigli *et al.* (2002). Only the step of hydrogen generation has been investigated, therefore emergy costs may be underestimated. Since this step has not direct environmental inputs, the environmental loading ratio and other indicators cannot be calculated.

[d] Global warming potential (GWP), includes direct CO_2 release from fuel combustion, indirect CO_2 release embodied in inputs, and (for geothermal plant only) CO_2 from underground water, released through the cooling towers.

[e] Indirect environmental inputs from the larger scale, needed to dilute heat and chemicals that are released by the plant.

[f] Includes only locally extracted fuel delivered to plant, as well as underground water and topsoil used up (if any).

[g] Labor accounts for hours actually worked in the investigated process. Services include human labor in all steps prior than plant use, quantified as the emergy associated to the monetary cost (cost of item, $, x country emergy/GNP ratio, seJ/$).

calculate bad performance indicators for oil-based energy systems. EEA simply indicates that the ability of the investigated devices to operate the conversion chemical energy to work is low, as expected for any engine subject to the second law and the Carnot factor. EMA adds an indication of poor sustainability in both the global dynamics of the economy and the biosphere. The same consideration applies to the investigated bioethanol from corn. In this case, however, EEA indicates a small thermal energy return, which supports several claims of bioenergy supporters in favor of increased use of photosynthesis for energy. On the other hand, EMA jointly evaluates the environmental work diverted from its natural pattern in order to provide free renewable inputs, to replace fertile soil eroded by the intensive tilling, to supply fuels, goods and machinery, to support labor and services, and concludes that the global sustainability of this product is so low that it cannot be usefully pursued as an alternative energy source.

Much lower environmental work is required to deliver energy via hydro- and wind-powered turbines, which simply means that less environmental activity is diverted from its original patterns to provide electricity to the economy. This ensures that these technologies are not withdrawing too many resources from the global biosphere dynamics. Renewable energies (sun, wind, rain and deep heat) already have a role in nature's self-organization processes, so that diverting to much of them to support an economic process reduces their input to other natural processes. EEA assigns a more favorable energy return to hydro- than to wind-electricity, while EMA-based sustainability is much greater for wind conversion, due to equal EYRs and much lower ELR of wind compared to hydro.

Geothermal energy return is surprisingly high according to EEA, since the huge amount of deep heat (high-quality flow) is not accounted for as an input. EMA indicators still suggest a good performance for geothermal electricity, but the fact that a large fraction of high-quality deep heat is dispersed increases the transformity of the output. In addition, the load on the environment due to the release of chemicals extracted with the underground fluid increases the ELR and decreases the global EIS of the process. None of these considerations could be obtained by the traditional energy analysis.

A final remark about OTEC is also illuminating. The process was evaluated as a net energy supplier according to EEA, due to an energy return equal to 3.8. However, this favorable result cannot be obtained without supplying a significant amount of resources in the form of capital, labor, and services. These resources, generated by previous investment of primary emergy sources, decrease the global sustainability of the process to values that make its viability very unlikely, more than its actual feasibility.

4.3.5 EEA and EMA: Scale and Energy Quality

Emergy analysis and embodied energy analysis treat the conceptual issues of scale and quality of energy very differently. From these differences, significant discussion has arisen over the years (see, for instance, the article in *Ecological Economics* titled "Embodied Energy Analysis and Emergy Analysis: A Comparative View" by Brown and Herendeen). In the following discussion, two important issues are elucidated: temporal and spatial scale, and energy quality.

4.3.5.1 Issues of Scale It should be kept in mind that EEA and EMA have different windows of interest (although they also share several common features) and therefore are used to investigate different questions about a given process.

In particular, EEA provides a measure of the overall commercial (oil equivalent) energy invested to support a production process. Its outcome is the amount of oil equivalent energy required to drive a process or to generate a product. Its spatial scale is the actual scale of the process, although it may in principle include the spatial scale of the previous commercial processes from which the input flows came from (for instance, if a mineral from a faraway country is fed to the process, transport must also be accounted for). Its timescale is the time required to extract the raw materials and to actually make the product. In principle, time is only accounted for in order to calculate the fraction of embodied energy of assets and machinery that are used up in the process and that therefore should be assigned to the product. Since machinery and assets have little embodied energy relevance, this time and the related amount of assets are often disregarded by EEA analysts.

Instead, the outcome of EMA is a quantification of the environmental support provided by nature to a process, which may be or may be not under human control. The spatial scale of EMA is larger than the actual scale of the process, since the evaluation also includes free environmental flows from - the larger scale of the biosphere that are not accounted for in energy analysis. In a similar way, the EMA timescale is the total time it took to make a resource via all natural processes involved (generation of minerals, fossil fuels, topsoil, water storages, etc.):

embodied in the transformity of a resource is the global environmental input required over space and time to drive the geological and biosphere processes, which converge to build a resource storage or to support a resource flow. Emergy-based indicators therefore have a built-in memory of the process history—that is, the real "trial and error" pathway that was followed by environment and humans to generate the product, including those pathways discarded by natural selection for maximum power output. The latter sentence may not completely apply to short timescale of technological processes and human economies, but results can be managed in such a way as to take into account the further uncertainty related to patterns that are not yet optimized.

4.3.5.2 Issues of Quality Energy has been defined as the ability to do work, based on the physical principle that work requires energy input. Energy is measured in units of heat, or molecular motion, the degree of motion resulting in expansion and quantified in degrees of temperature. All energies can be converted to heat at 100% efficiency; thus, it is relatively easy and accurate to express energies in their heat equivalents. The basic units of energy are the amount of heat required to raise a given amount of water a given number of degrees of temperature. Thus, the calorie is the amount of heat required to raise 1 cm^3 of water 1°Celsius. A joule is equal to 4.187 calories.

Heat-equivalent energy is a good measure of the ability to raise water temperature. However, it is not a good measure of more complex work processes. Processes outside of the window defined by heat engine technology do not use energies that lend themselves to thermodynamic heat transfers. As a result, converting all energies of the biosphere to their heat equivalents reduces all work process of the biosphere to heat engines. Human beings, then, become heat engines and the value of their services and information is nothing more than a few thousand calories per day. Obviously, not all energies are the same and methods of analysis need reflect this fact.

Different forms of energy have different abilities to do work, not only in terms of amounts of work but also in terms of kind of work. It is therefore necessary to account for these different abilities if energies are to be evaluated correctly. A joule of sunlight is not the same as a joule of fossil fuel or a joule of food, unless it is being used to power a steam engine. Sunlight drives photosynthesis. A system organized to use concentrated energies like fossil fuels cannot process a more dilute energy form like sunlight, joule for joule. Evaluation of energy sources is system dependent. The processes of the biosphere are infinitely varied and are more than just thermodynamic heat engines. As a result, the use of heat measures of energy that can only recognize one aspect of energy, its ability to raise the temperature of things, cannot adequately quantify the work potential of energies used in more complex processes of the biosphere. As in thermodynamic systems where energies are converted to heat to express their relative values, in the larger biosphere system as a whole energies should be converted to units that span this greater realm, accounting for multiple levels of system processes, ranging from the smallest scale to the largest scales of the biosphere, and accounting for processes other than heat engine technology. The ability of driving processes other than engine-like ones is a new quality aspect of resource flows, which is accounted for by emergy analysis.

Net energy evaluations of energy sources and transformation processes are designed to provide information concerning efficiency and potential yield under the engine-like point of view cited previously. This is something that energy analysis does appropriately. It is able to offer at least two different results:

1. As embodied energy analysis, it provides a measure of the commercial energy cost of a product (MJ or grams of oil equivalent per unit of product). When the product is energy, this measure is better expressed as EROI (joules of energy delivered per unit of energy provided to the process).

2. As exergy analysis at the process scale, it provides a measure of thermodynamic efficiency, indicates possible optimization patterns, and finally ranks the quality of the product from the user-side point of view.

Neither of these two energy analysis patterns is able to provide any significant insight into the quality of each input and output flow in the larger scale of the biosphere, as pointed out earlier. Neither of them takes into account the role of the environmental work supporting both ecosystems and human societies, in particular the role of unmonied inputs provided by nature to human economies. This is something that requires a scale expansion and a change of perspective, from local to global, i.e., the emergy approach.

5. SUMMARY

This article is a brief synopsis of the emergy accounting methodology. Emergy-based case studies of energy conversion systems are presented to allow both an increased understanding of how the method is applied as well as a critical comparisons with embodied energy analysis. Much like a short intensive course given in a limited time frame, this article has touched on what is most important. There is much literature explaining the theory and concepts of the emergy approach as well as applications of the method to a large number of case studies, ranging from ecosystems to socioeconomic systems, to evaluations of the biosphere, including energy processing, agriculture, industry, tourism, wastewater management, recycling patterns, generation and storage of culture and information, development of cities as emergy attractors, and scenarios for future societal development. References to emergy literature are given throughout. Starting on 1999, an international emergy conference takes place every other year. The books of proceedings of the first two emergy conferences provide a significant set of theoretical and applied papers for further reading. In all cases, the common thread is the ability to evaluate all forms of energy, materials, and human services on a common basis by converting them into equivalents of one form of energy, solar emergy, a measure of the past and present environmental support to any process occurring in the biosphere.

SEE ALSO THE FOLLOWING ARTICLES

Earth's Energy Balance • Ecosystem Health: Energy Indicators • Ecosystems and Energy: History and Overview • Exergy • Exergy Analysis of Energy Systems • Global Energy Use: Status and Trends • Human Energetics • Thermodynamics, Laws of

Further Reading

Bargigli, S., Raugei, M., and Ulgiati, S. (2002). Thermodynamic and environmental profile of selected gaseous energy carriers. Comparing natural gas, syngas and hydrogen. *In* "ECOS 2002," Book of Proceedings of the 15th International Conference on Efficiency, Costs, Optimization, Simulation and Environmental Impact of Energy Systems–Berlin (Germany), July 2–5 (G. Tsatsaronis, M. J. Moran, F. Cziesla, and T. Bruckner, Eds.), pp. 604–612.

Bargigli, S., Raugei, M., and Ulgiati, S. (2003). A multi-scale and multi-criteria analysis of aluminium "greenness." *J. Cleaner Production,* submitted.

Bargigli, S., and Ulgiati, S. (2003). Emergy and life-cycle assessment of steel production. *In* Proceedings of the Second Biennial Emergy Evaluation and Research Conference, Gainesville, FL.

British Petroleum (1997). "BP Statistical Review of World Energy, 1997." The British Petroleum Company, London.

Brown, L. R., Renner, M., and Flavin, C. (1997). "Vital Signs, 1997: The Environmental Trends That Are Shaping Our Future." W. W. Morton & Co., New York.

Brown, M. T., and Arding, J. (1991). Tranformities Working Paper, Center for Wetlands. University of Florida, Gainesville, FL.

Brown, M. T., Brandt-Williams, S., Tilley, D., and Ulgiati, S. (eds.). (2001). "Emergy Synthesis. Theory and Applications of the Emergy Methodology." Book of Proceedings of the First International Emergy Research Conference, Gainesville, FL, September 2–4, 1999. The Center for Environmental Policy, University of Florida, Gainesville, FL.

Brown, M. T., and Buranakarn, V. (2003). Emergy indices and ratios for sustainable material cycles and recycle options. *Resources, Conservation and Recycling* **38,** 1–22.

Brown, M. T., and Herendeen, R. A. (1996). Embodied energy analysis and emergy analysis: A comparative view. *Ecological Economics* **19,** 219–235.

Brown, M. T., and McClanahan, T. (1996). Emergy Analysis Perspectives for Thailand and Mekong River Dam Proposals. *Ecological Modeling* **91,** 105–130.

Brown, M. T., Odum, H. T., Tilley, D., and Ulgiati, S. (eds.). (2003). "Emergy Synthesis. Theory and Applications of the Emergy Methodology–2." Book of Proceedings of the Second International Emergy Research Conference, Gainesville, FL, September 20–22, 2001. The Center for Environmental Policy, University of Florida, Gainesville, FL.

Brown, M. T., and Ulgiati, S. (1997). Emergy based indices and ratios to evaluate sustainability–Monitoring economies and technology towards environmentally sound innovation. *Ecological Engineering* **9,** 51–69.

Brown, M. T., and Ulgiati, S. (1999). Emergy evaluation of the biosphere and natural capital. *Ambio* **28**(6), 486–492.

Brown, M. T., and Ulgiati, S. (2002). Emergy evaluation and environmental loading of electricity production systems. *J. Cleaner Production* **10,** 321–334.

Campbell, D. (1998). Emergy analysis of human carrying capacity and regional sustainability: An example using the state of Maine (Appendix). *Environmental Monitoring and Assessment* **51,** 531–569.

Hall, C. A. S., Cleveland, C. J., and Kaufman, R. (1986). "Energy and Resource Quality: The Ecology of the Economic Process." Wiley/Interscience, New York.

Huang, S. –L., Lai, H. –Y., and Lee, C. –L. (2001). Energy hierarchy and urban landscape system. *Landscape and Urban Planning* **53,** 145–161.

Kinsman, B. (1965). "Wind, Waves, Their Generation and Propagation on the Ocean Surface." Prentice-Hall, Englewood Cliffs, NJ.

Lapp, C. W. (1991). Emergy analysis of the nuclear power system in the United States. M. S. thesis, University of Florida, Department of Environmental Engineering Sciences, Gainesville, FL.

Mannion, A. M. (1995). "Agriculture and Environmental Change: Temporal and Spatial Dimensions." John Wiley & Sons, New York.

Miller, G. A. (1966). The flux of tidal energy out of the deep oceans. *J. Geophysical Res.* **71,** 2485–2489.

Odum, H. T. (1996). "Environmental Accounting. Emergy and Environmental Decision Making." John Wiley & Sons, New York.

Odum, H. T. (2000a). Handbook of Emergy Evaluation: A Compendium of Data for Emergy Computation Issued in a Series of Folios. Folio #2—Emergy of Global Processes. Center for Environmental Policy, Environmental Engineering Sciences, University of Florida, Gainesville, FL.

Odum, H. T. (2000b). Emergy Evaluation of an OTEC Electrical Power System. *Energy* **25,** 389–393.

Odum, H. T., Brown, M. T., and Ulgiati, S. (1999). Ecosystems as energetic systems. *In* "Handbook of Ecosystem Theories" (S. E. Jorgensen and F. Muller, Eds.), pp. 281–302. CRC Press, New York.

Odum, H. T., Brown, M. T., Whitefield, L. S., Woithe, R., and Doherty, S. (1995). Zonal Organization of Cities and Environment: A Study of Energy System Basis for Urban Society. A Report to the Chiang Ching-Kuo Foundation for Intergational Scholarly Exchange, Center for Environmental Policy, University of Florida, Gainesville, FL.

Odum, H. T., Brown, M. T., and Williams, S. B. (2000). Handbook of Emergy Evaluation: A Compendium of Data for Emergy Computation Issued in a Series of Folios. Folio #1–Introduction and Global Budget. Center for Environmental Policy, Environmental Engineering Sciences. University of Florida, Gainesville, FL.

Odum, H. T., and Odum, E. C. (2001). "A Prosperous Way Down: Principles and Policies." University Press of Colorado, Boulder, CO.

Odum, H. T., and Odum, E. C. (eds.). (1983). Energy Analysis Overview of Nations. Working Paper WP-83–82. International Institute for Applied Systems Analysis, Laxenburg, Austria.

Oldeman, L. R. (1994). The global extent of soil degradation. *In* "Soil Resilience and Sustainable Land Use" (D. J. Greenland and I. Szabolcs, Eds.). CAB International, Wallington, United Kingdom.

Oort, A. H., Ascher, S. C., Levitus, S., and Peixoto, J. P. (1989). New estimates of the available potential energy in the World Ocean. *J. Geophysical Research* **94,** 3187–3200.

Ryabchikov, A. (1975). "The Changing Face of the Earth." Translated by J. Williams. Progress, Moscow.

Scienceman, D. M. (1987). Energy and emergy. *In* "Environmental Economics" (G. Pillet and T. Murota, Eds.), pp. 257–276. Roland Leimgruber, Geneva, Switzerland.

Sclater, J. F., Taupart, G., and Galson, I. D. (1980). The heat flow through the oceanic and continental crust and the heat loss of the earth. *Rev. Geophys. Space Phys.* **18,** 269–311.

Tabacco, A. M., Bargigli, S., Raugei, M., and Ulgiati, S. (2003). Life cycle, exergy and emergy assessment of metallic chromium production from ore. *In* "Advances in Energy Studies: Reconsidering the Importance of Energy" (S. Ulgiati, M. T. Brown, M. Giampietro, R. A. Herendeen, and K. Mayumi, Eds.). SGE, Padova, Italy.

Todd, D. K. (ed.). (1970). "The Water Encyclopedia," 2nd ed. Lewis, Chelsea, MI.

Ulgiati, S. (2001). A comprehensive energy and economic assessment of biofuels: When "green" is not enough. *Critical Reviews in Plant Sciences* **20,** 71–106.

Ulgiati, S. (2003). "Emergy Analysis of Italy in the years 1984, 1989, 1991, 1995, and 2000." Unpublished manuscript, Energy and Environment Research Unit, Department of Chemistry, University of Siena, Italy (ulgiati@unisi.it).

USDI (1996). "Mineral Commodity Summaries, January 1997." U. S. Department of Interior, Washington, DC.

Wiin-Nielsen, A., and Chen, T. (1993). "Fundamentals of atmospheric energetics." Oxford Press, New York.

World Resources Institute (1996). "World resources 1996–97." Oxford University Press, Oxford.

Energy Development on Public Land in the United States

BRIAN BLACK
Pennsylvania State University, Altoona College
Altoona, Pennsylvania, United States

1. Introduction
2. Energy for All?
3. General Phases of Public Land Administration
4. Establishing the Legal Framework
5. Making Legal Exception: Native American Communities
6. History of Energy Development on Public Land
7. Debating Ownership: The Sagebrush Rebellion
8. A Symbolic Contest: Alaska
9. Battle Lines in the Current Debate

Glossary

application for permit to drill (APD) An application to drill a well; submitted by a lessee or operator to the Bureau of Land Management.
Bureau of Land Management (BLM) The Department of the Interior agency responsible for managing most federal onshore subsurface minerals.
environmental impact statement (EIS) A written analysis of the impacts on the natural, social, and economic environment of a proposed project or resource management plan.
lease An authorization to possess and use public land for a period of time sufficient to amortize capital investments in the land.
public lands The lands owned by the federal government, including parks, monuments, wilderness areas, refuges, underground mineral reserves, marine sanctuaries, historic parks, forests, and seashores.
withdrawal An action that restricts the disposition of public lands and that holds them for specific public purposes.

Public lands in the United States have provided supplies of raw energy resources for over a century. The story of the harvest and administration of energy from natural resources reveals the public debate and contest that often results when resources are owned in common. Recent developments suggest that future use of these resources will continue to be hotly debated.

1. INTRODUCTION

When Woody Guthrie wrote the song *This Land Is Your Land*, his rough-toned voice instructed Americans to recall that "this land was made for you and me." Nowhere is this American ideal more obvious than in the nation's publicly owned lands—or so it would seem. Yet the spirit of Guthrie's idealism has rarely penetrated the administration of the more than 600 million acres of land that belong to each citizen. One-third of all national land is administered by the federal government and is owned collectively by the people of the nation. For a variety of reasons, including climate and the late entry of Euro-American settlers, most of the federal land can be found in the American West.

Public lands include parks, monuments, wilderness areas, refuges, underground mineral reserves, marine sanctuaries, historic parks, forests, and seashores. Throughout American history, the administration of this great national resource has been tied to politics and the powerful elements of economic development. The use of public land sites and particularly the harvest of natural resources existing on them have been consistently debated in recent years. Should these lands be viewed as vast storehouses of resources on which our nation depends? Or are federal lands intended more for preservation, regardless of valuable resources that might be found within their borders? And, of course, when development takes place, who should gain the financial benefit?

 © 2004 Elsevier Inc. All rights reserved.

As a matter of existing law, mineral resources on public lands exist as national property, primarily due to specific strategies to foster development in the western United States. Outside variables, including the needs of war or domestic crisis or even presidential preference, influence the perceptions of mineral resources, particularly those resources related to energy production. Although the use and availability of such resources vary with the philosophy of each presidential administration, each American leader has viewed these energy sources as an important tool for economic development of communities in the western United States. The harvest of resources on public lands carries economic benefits locally and nationally (see Table I).

Although wood must be included as an energy resource that has been harvested from federal lands, the major sources in the 20th century have included: coal, oil, natural gas, shale, uranium, and geothermal. In recent years, there has additionally been a concerted effort to develop alternative energy resources (such as solar and wind) on public lands. By the end of the 20th century, the energy industry estimated that Americans received 30% of their fuels for energy production from public lands.

TABLE I

Disposition of Public Lands, 1781–1997[a]

Type of disposition	Acres
Unclassified[b]	303,500,000
Granted or sold to homesteaders[c]	287,500,000
Total unclassified and homestead	591,000,000
Support of comom schools	77,630,000
Reclamation of swampland	64,920,000
[c]Construction of railroads	37,130,000
Support of miscellaneous institutions[d]	21,700,000
Purposes not elsewhere classified[e]	117,600,000
Canals and rivers	6,100,000
Construction of wagon roads	3,400,000
Total granted to states	328,480,000
Granted to railroad corporations	94,400,000
Granted to veterans as military bounties	61,000,000
Confirmed as private land claims[f]	34,000,000
Sold under timber and stone law[g]	13,900,000
Granted or sold under timber culture law[h]	10,900,000
Sold under desert land[i]	10,700,000
Total miscellaneous dispositions	224,900,000
Grand total	1,144,380,000

[a] Data estimated from available records.

[b] Chiefly public, private, and preemption sales, but includes mineral entries, scrip locations, and sales of town sites and town lots.

[c] The homestead laws generally provided for the granting of lands to homesteaders who settled on and improved vacant agricultural public lands. Payment for the lands was sometimes permitted, or required, under certain conditions.

[d] Universities, hospitals, asylums, etc.

[e] For construction of various public improvements (individual items not specified in the granting acts), reclamation of desert lands, construction of water reservoirs, etc.

[f] The government has confirmed title to lands claimed under valid grants made by foreign governments prior to the acquisition of the public domain by the United States.

[g] The timber and stone laws provided for the sale of lands valuable for timber or stone and unfit for cultivation.

[h] The timber culture laws provided for the granting of public lands to settlers on condition that they planted and cultivated trees on the lands granted. Payments for the lands were permitted under certain conditions.

[i] The desert land laws provided for the sale of arid agricultural public lands to settlers who irrigated them and brought them under cultivation.

2. ENERGY FOR ALL?

The remaining oil, natural gas, and mineral deposits in the continental United States are concentrated in the American West, where they were created between 40 million and a few hundred million years ago. The abundance of such resources on public lands is a historical confluence of variables: lands that were unsettled and therefore taken by the federal government happen to have been the site, thousands of years ago, of ecosystems that today result in abundant energy resources. Western sites such as Fossil Butte National Monument (Wyoming) and Dinosaur National Monument (Utah) contain the fossilized remains that demonstrate abundant life many yeas ago, both in and out of the sea. Of course, as these ancient plants and animals decomposed over millions of years, extreme heat and pressure transformed them into resources that can be mined and burned to provide energy.

Such nonrenewable sources are just a fraction of the available energy resources. Today, the open space available on much of the federal lands has also become a resource for the development of another type of energy: renewable, including wind, solar, and water power. The U.S. Department of the Interior (USDI) oversees the leasing of lands for this development, including mining. Within USDI, the Bureau of Land Management (BLM) manages the development of all fossil fuels and minerals on nearly all of the lands, excluding areas of special jurisdiction such as the Naval Petroleum Reserves

TABLE II

Public Lands under Exclusive Jurisdiction of the Bureau of Land Management, Fiscal Year 1997[a]

	Vacant public lands (acres)[b]			Reserved lands (acres)		
State	Outside grazing districts	Within grazing districts	Total	Lu[c]	Other	Grand total
Alabama	3117	—	3117	—	107,846	110,963
Alaska	86,908,060	—	86,908,060	—	—	86,908,060
Arizona	1,516,765	10,093,234	11,609,999	32,321[d]	2,610,458	14,252,778
Arkansas	2059	—	2059	—	289,107	291,166
California	7,363,543	1,725,343	9,088,886	—	5,467,188	14,556,074
Colorado	480,331	6,781,321	7,262,065	—	998,241	8,296,512
Florida	1512	—	1512	—	23,765	25,277
Idaho	422,341	10,733,321	11,155,662	72,276	619,390	11,847,328
Illinois	3	—	3	—	224	227
Iowa	—	—	—	—	378	378
Louisiana	4351	—	4351	—	305,260	309,611
Michigan	47	—	47	—	74,807	74,854
Minnesota	6044	—	6044	—	145,211	151,255
Mississippi	1240	—	1240	—	55,971	57,211
Missouri	67	—	67	—	2094	2161
Montana	1,148,321	4,940,802	6,089,123	1,801,171	170,088	8,060,382
Nebraska	6580	—	6580	—		6,580
Nevada	3,140,726	44,493,239	47,633,965	3127	207,299	47,840,569
New Mexico	1,355,577	11,047,165	12,402,742	229,500	138,327	12,770,569
North Dakota	59,536	—	59,536	—	181	59,717
Oklahoma	2142	—	2142	—		2142
Oregon	585,675	12,455,100	13,040,775	78,124	3,104,840	16,233,739
South Dakota	272,277	—	272,277	—	7592	279,869
Utah	—	21,155,026	21,155,026	45,083	1,677,604	22,877,663
Washington	366,921	—	366,921	—	3189	370,110
Wisconsin	2521	—	2521	—	157,616	160,137
Wyoming	3,910,677	11,273,811	15,184,488	10,434	3,189,004	18,383,926
Total	107,560,433	134,698,775	242,259,208	2,308,242	19,355,680	263,929,258

[a] Lands under the exclusive jurisdiction of the Bureau of Land Management (BLM) (and resources) managed solely by the BLM. The table includes all Uperfected entries except for those involving reclamation and forest homesteads. Differences in fiscal year 1997 acreage from the previous year are due to land acreage acquired by the BLM; exchange land acreage received by the BLM; exchange BLM acreage (surface estate only) patented out by the BLM; withdrawn land acreage restored to BLM's jurisdiction; withdrawals, rights-of-way, acreage adjustments in carryover data.

[b] The following types of surveyed and unsurveyed public and ceded Native American lands are included: areas withdrawn under the executive orders of November 26, 1934, and February 5, 1935 (43 CFR 2400.0-3 *et seq.*); areas embraced in mineral withdrawals and classifications; area withdrawn for survey; and areas restored to entry within national forests (Act of June 11, 1906, 34 Stat. 233, 16 U.S.C.,§§ 506–509), within reclamation projects (Act of June 17, 1902, 32 Stat. 388) and within power site reserves (Act of June 10, 1920, 41 Stat. 1063; 16 U.S.C.,§§ 791). These lands are not covered by any nonfederal right or claim other than permits, leases, rights-of-way, and unreported mining claims.

[c] Land Utilization project lands, purchased by the federal government under Title III of the Bankhead–Jones Farm Tenant Act and subsequently transferred by various executive orders between 1941 and 1958 from jurisdiction of the U.S. Department of Agriculture to the U.S. Department of the Interior, now administered by the Bureau of Land Management. Acreages are within grazing districts unless otherwise noted.

[d] Includes acreage on the Barry M. Goldwater Air Force Range.

and hydroelectric watershed development areas that are administered by the Army Corps of Engineers (see Table II).

The sheer amount of energy resources on federal lands is staggering. Of the 75.6 million acres of coal-bearing public lands, only 1% is currently being

mined. As for oil and gas, energy companies are actively working between 18,000 and 25,000 wells on public lands, with tens of thousands more sites held in reserve. A rarer mineral, uranium, is processed to create nuclear energy, which provides electricity to millions of homes, schools, and businesses. About one-third of all oil and gas comes from public lands. In addition, public lands are estimated to contain about 68% of all undiscovered oil reserves and 74% of all undiscovered gas reserves in our country. Nearly all reserves of uranium are found on federal land. Public lands are also used for harvesting hydroelectric power, although these areas are administered by the Army Corps of Engineers.

Regardless of the resource, the story of the harvest and administration of this energy supply reveals a theme that recurs in many publicly owned areas: public debate and contest. If studied over time, the use and management of the energy resources on public lands demonstrate important changes in the ethics guiding the management of federal lands. In addition, though, a survey of public land use demonstrates a public controversy that promises to rage deep into the 21st century.

3. GENERAL PHASES OF PUBLIC LAND ADMINISTRATION

Broader philosophical ideas, called ethics, toward land-use inform the administration of public lands. For the purposes of this discussion, the administration of public lands can be separated into four distinct eras or shifts in these larger, governing ethics. Such ethics, of course, are formed by a complex set of cultural, social, and political variables. As a publicly administered entity, public lands are subject to the influence of larger patterns within American society. In the case of energy development on public lands, the predominant cultural variable has been the ideological shifts caused by the evolution of the modern environmental movement in the 20th century. The ethic of the environmental movement has provided a nonutilitarian impulse toward the administration of the public domain.

Although the overall progression of public land use has moved toward incorporating an environmental ethic into the general management of the land, there have been a few crucial junctures in which the overall trajectory of this policy shift has been interrupted by a backlash of regressive efforts at development. After introducing the overall phases of public land administration, the focus here will be to trace the general chronology of legal restrictions on public land policy, to explore how the harvest of specific resources has been affected by these laws, and, finally, to shed light on a few of these interruptions in the overall evolution of public lands policy. The data in Table III detail how public lands were allocated in 1996.

3.1 Settlement Phase

The settlement phase of land use grew out of the nation's emphasis on settling or reclaiming nearly 1.8 billion acres of original public domain. Hundreds of laws were passed between 1785 and 1878 that were designed to encourage settlement and development of the timber, mineral, and forage supplies. Typically, such policies said little about the native peoples who were being displaced by this disposal of land or about the fashion in which the lands were to be developed. This phase of land use had little to do with energy development, because it predominantly concerned an era in which there was little use of fossil fuels. Growing out of this approach to land use, however, Congress created the General Mining Law of 1872, which remains in effect to govern mining on public lands. The General Mining Law grants free access to individuals and corporations to prospect for minerals in public domain lands. Once anyone discovers a location, they can stake a claim on the mineral deposit. This claim provides the holder the right to develop the minerals and may be "patented" to convey full title to the claimant. In this stage of land use, the federal government clearly prioritized the stimulation of new use over any other consideration.

3.2 Resource Conservation Phase

Growing out of the American conservation movement of the late 1800s, the resource conservation era in public land use emphasized progressive ideals, particularly those expressed during the administration of President Theodore Roosevelt (1901–1909) by the president and his chief advisor on land management, Gifford Pinchot. This era marked the beginning of the nation's first period of public land stewardship. President William Henry Harrison previously began the forest reserve system with land appropriations before solidifying it the Organic Act of 1897.

Roosevelt and Pinchot emphasized forest management, expanding the reserves to 148 million acres. Energy resources, other than wood, attracted the attention of progressives with the Coal Lands

TABLE III

Applications, Entries, Selections, and Patents, Fiscal Year 1996

Type by state	Applications, entries, and selections				Patents issued		Purchase money (U.S. dollars)[a]
	Allowed		Approved				
	Number	Acres	Number	Acres	Number	Acres	
Homestead							
Montana					1	160	
Wyoming					2	304	
Total homestead					3	464	
Desert land							
Idaho					2	520	520
Nevada					8	2530	3162
Utah			1	39	1	39	40
Total desert land			1	39	11	3089	3722
Land sales (FLPMA)[b]							
Alabama					2	80	1050
Arizona					8[c]	280	213,681
California					2	10	550
Colorado					2	38	64,906
Idaho					2	100	15,456
Nebraska					7	349	1600
Minnesota					2	232	
Nevada					9	444	2,717,375
New Mexico					9	18	60,892
Oregon					9	1744	203,459
South Dakota					2	1	1000
Utah					10	875	71,050
Wyoming					5	246	10,775
Total land sales					69	4417	3,361,794
Mineral							
Arizona					6	487	
Total mineral					6	487	
Native American new trust							
California					1	13	
Minnesota					6	275	
Wyoming					4	22	
Total Native American new trust					11	310	
Native American fee and reissue trust							
Colorado			1	52	1	52	
Idaho					7	640	
Montana			145	27,161	145	27,161	
Minnesota					2	48	
Nebraska					3	38	
North Dakota			35	6903	35	6903	
Oklahoma					25	1519	
Oregon			24	2091	24	2091	

continues

Table III continued

Type by state	Applications, entries, and selections				Patents issued		Purchase money (U.S. dollars)[a]
	Allowed		Approved				
	Number	Acres	Number	Acres	Number	Acres	
South Dakota			76	16,696	76	16,676	
Washington			49	2779	49	2779	
Total Native American fee and reissue trust			330	55,682	367	57,907	
Miscellaneous cash sales							
Color of title							
Alabama					2	80	1050
Arkansas					2	194	12,172
Colorado			1	1	1	1	3276
New Mexico					5	2	6
North Dakota			1	40	1	40	430
Oregon			1	40	1	40	3760
Total color of title			3	81	12	357	20,694
Recreation and public purposes							
Arizona			6	1465	8[d]	1465	10,138
California					7	2548	523
Colorado			3	333	3	333	3558
Florida					1	1	
Minnesota					1	16	
Missouri					1	40	
Montana			1	160	1	160	
Nevada					8	460	14,812
New Mexico					2	84	50
Oklahoma					1	11	14
Oregon			1	3	1	3	50
Utah			6	314	6	314	67,800
Wyoming					2	243	6096
Total recreation and public purposes			17	2275	42	5678	103,041
Other							
Nevada					1[e]	2	69,502
Washington					1[f]	414	294,276
Total other					2	416	363,778
Total miscellaneous cash sales			20	2356	56	6451	487,513
Conveyance of federally owned mineral estates							
Arizona					7	5998	
California					1	561	
Colorado							50
Florida					2	7	225
Georgia					1	106	2000

continues

Table III continued

Type by state	Applications, entries, and selections				Patents issued		Purchase money (U.S. dollars)[a]
	Allowed		Approved				
	Number	Acres	Number	Acres	Number	Acres	
Minnesota					1	331	547
Montana					1	2	2600
New Mexico					1	1	1650
Utah					21	1443	584,798
Wyoming							50
Total conveyance of federally owned mineral estates					35	8449	591,920
Exchanges[g]							
Arizona					3	150	184,000[b]
Arkansas					1	40	
California					42	17,981	23,500
Colorado					48	6208[i]	—[j]
Idaho					32	7831	
Michigan					3	128	
Montana					11	12,058	
Nevada					20	45,039	10,360
New Mexico					10	80,476	
Oregon					17	11,541	74,000
South Dakota					3	707	
Utah					5	3963	
Washington					11	1925	150,000
Wyoming					3	5244	8311
Total exchanges					209	193,291	450,171
State grants							
Florida					1	175	
Total state grants					1	175	
Corrections							
Alabama					3	508	
Arkansas					1	320	
California					2	6548	
Colorado					3	74	
Indiana					1	64	
Louisiana					1	685	
Michigan					1	640	
Minnesota					3	380	
Mississippi					4	376	
Missouri					2	572	
Montana					1	15	
New Mexico					5[k]		
North Dakota					2	640	
Ohio					4	560	
Oregon					2	91	
Utah					5	471	
Wisconsin					2	280	

continues

Table III continued

Type by state	Applications, entries, and selections: Allowed Number	Allowed Acres	Approved Number	Approved Acres	Patents issued Number	Patents issued Acres	Purchase money (U.S. dollars)[a]
Wyoming					1	160	
Total corrections					43	12,384	
Other							
Alaska					364[l]	600,204	400
Colorado					1[m]	1	
Wyoming					2[n]	256	
Total other					367	600,461	400
Grand total			351	58,077	1178	817,594[o]	4,895,520

[a] Purchase money includes all revenues collected during the reporting fiscal year at allowed and approved stages from all ongoing cases leading to patent; however, this money is not always received in the same year that the application is allowed or approved, or in the same year that the patent is issued. Purchase money enters into the land exchange process because exchanges must result in equal value given and received by both parties; this means that cash payments are sometimes made or received by the Bureau of Land Mangement to ensure an equitable exchange.

[b] FLPMA, Federal Land Policy and Management Act of 1976.

[c] Patents issued under Sections 203 and 209 of the Federal Land Policy and Management Act of 1976 (43 U.S.C. §§ 1713, 1719).

[d] Includes two Certificate of Transfer/Change of Use patents.

[e] Public land sale (2.5 acres), Santini–Burton Act of December 23, 1980 (94 Stat. 3881; Public Law 96-586).

[f] Public land sale patent under the Act of September 19, 1964 (78 Stat. 988; Public Law 88-608).

[g] Includes National Forest Service forest exchanges, Act of March 20, 1922 (42 Stat. 465; 16 U.S.C. § 485) as amended; Federal Land Policy and Mangement Act exchanges, Act of October 21, 1976 (90 Stat. 2756; 43 U.S.C. § 1716; and other exchanges.

[h] The U.S. Forest Service received a cash equalization payment under the Sikes Act of 1974 (16 U.S.C. § 484a); no land received by the United States.

[i] Does not include 960 acres of mineral estate only that was patented.

[j] The U.S. Forest Service paid $666,010 in equalization payments to the exchange proponents.

[k] Recordable Disclaimers of Interest in Lands, Section 315 of the Federal Land Policy and Management Act of 1976 (43 U.S.C. § 1745); no patents issued.

[l] Eight homesites (37 acres), Act of March 3, 1927 (44 Stat. 1364; 43 U.S.C. § 687a), as amended; headquarters Sites (15 acres), Act of March 3, 1927 (44 Stat. 1364; 43 U.S.C. § 687a), as amended; one trade and manufacturing Site (66 acres), Act of May 14, 1898 (30 Stat. 413), as amended August 23, 1958 (72 Stat. 730; 43 U.S.C. § 687a); 39 tentative approvals (355, 355 acres), Alaska Statehood Act of July 7, 1958 (72 Stat. 339),as amended; 31 interim conveyances (212, 333 acres), Alaska Native Claims Settlement Act of December 18, 1971 (43 U.S.C. § 1601); 281 native allotment certificates (32, 396 acres), Act of May 17, 1906 (34 Stat. 197), as amended August 2, 1956 (70 Stat. 954; 43 U.S.C. §§ 270–1 to 270–3); 1 highway lot (0.35 acres), Act of June 14, 1926 (44 Stat. 741; 43 U.S.C. § 971a).

[m] Clear Creek County, Colorado Public Lands Transfer Act of 1993 (108 Stat. 674), approved May 19, 1994.

[n] Airport patents under Section 516 of the Airport and Airway Improvement Act of September 3, 1982 (49 U.S.C. § 2215).

[o] This total does not include 57,907 acres in Native American fee and reissue trust patents, and 12,384 acres in correction of patents, for a total of 70,291 acres.

withdrawal in 1906 that allowed the government to reserve mineral rights on 66 million acres. Initially, this reservation included oil and gas, though it would be further expanded to include many other minerals. Roosevelt also withdrew other public lands from settlement, using them to create national parks, monuments, wildlife refuges, and military bases. Within such sites, this era saw an increase in control over the use of the resources on public lands, including legislation such as the Mineral Leasing Act of 1920 and the Taylor Grazing Act of 1934. Pinchot's conservation ethic was reflected in laws that promoted sustainable use of natural resources. These ideas were also of interest to President Franklin Delano Roosevelt, who used the Civilian Conservation Corps (CCC) during the 1930s to carry out many of these ideas on public lands.

3.3 Post-World War II Phase

During World War II and for most of two decades following it, the federal management of public lands went along with nationwide initiatives to stimulate national growth and international security. In 1946, the Grazing Service merged with General Land Office to form the BLM, the entity that oversees

the majority of the public lands in the western United States. The pressure to harvest resources from these lands intensified with the need for materials for war and the expansion of American consumption with the massive growth of its middle class.

With the beginnings of environmentalism in the late 1950s, many Americans grew more concerned that public lands were not subject to environmental assessment prior to use. This shift in overall ethics is generally accepted to have grown from the effort to prevent the damming of the Green River in Dinosaur National Park in 1954. Although the environmental movement influenced land use in national parks to a much greater degree, the public lands shifted more toward a compromise philosophy of "multiple use."

3.4 Environmental Protection Phase

Although the era of environmental protection has met with significant interruptions, the overall pattern from 1960 to the present has been toward a more environmentally sustainable pattern of use on public lands. The Multiple Use Sustained Yield Act in 1960 was followed by the public outcry related to Rachel Carson's publication of *Silent Spring* in 1962. The ensuing flurry of environmental legislation culminated in 1969 with the National Environmental Policy Act (NEPA), which created the Environmental Protection Agency (EPA) and required that any development on federally owned lands needed to be preceded by an environmental impact statement. Ecological concepts, such as ecosystems and watersheds, provided administrators with a new way of viewing public lands. No longer simply a storehouse of potential resources, each of these sites could now be seen as part of larger, complex natural systems. This awareness crystallized during a serious period of contestation over the use of public lands in the late 20th century.

4. ESTABLISHING THE LEGAL FRAMEWORK

The basic tool for defining each of the eras in public land use has been laws or regulations. Throughout the history of public lands, the primary rationale for their administration was their usefulness. If settlers had found reasons to build communities in certain areas, it is likely that those areas would not have become public lands in the first place. Vast tracts of the territories that became states in the late 1800s were simply unwanted by settlers and developers. Similar to the Homestead Act, the General Mining Law of 1872 provided incentives to settlement and development (see Table IV). Typical of legislation of this era, the General Mining Law emphasizes individual rights with minimal government management. The law prioritizes free and open access to the land and relatively unlimited opportunity to explore and develop resources. Developers were offered access to permanent, exclusive resource rights and were exempted from paying royalties to the federal government. Clearly, the law is based on the simple motive to spur development.

The success of the effort of the General Mining Law to stimulate western development relied on allowing open exploration. During the late 19th century, little was done to stop exploration from growing rapidly into mine development and full-blown extraction. In these early years, miners were not required to share production, to create joint ventures, or to pay taxes or royalties. Many mining expeditions proceeded largely unnoticed on public lands. Today, the Mining Law continues to provide the structure for much of the western United States mineral development on public domain lands. The unfettered basis of use and ownership established by the 1872 law continues to fuel the debate over how to harvest energy and mineral resources from public lands. The basis for such law was the assumption that there were no alternative criteria for land valuation than the resources that it held. The basis for much of the debate at the end of the 20th century revolved around the simple question of what would happen when, for a variety of reasons, this land became desirable.

These debates in the 20th century slowed the unlimited development of the earlier years. Many of the public lands had acquired additional importance, independent of energy and mineral resources. Additionally, such development now rarely proceeded unnoticed. Some Americans began calling for energy companies to share revenues with the land's owner through royalties. Other observers reflected a growing call for more stringent oversight of the extraction process. This call resulted in the 1920 Minerals Leasing Act. For the first time, this act closed some federal lands to open access. Most often, such lands then became available for leasing. The 1920 act established a broad framework for leasing and for the payment of royalties. Lands were to be designated as "known mineral areas" (KMAs) and then administered by one of two methods: (1) preference right leasing, which provided exploration permits that would expire if the search was unsuccessful, or

TABLE IV

Patents Issued with Minerals Reserved to the United States through Fiscal Year 1996[a]

State	Type of mineral reservation (acres)						
	All minerals	Coal	Oil and gas	Phosphate	Oil and gas plus other minerals	Miscellaneous minerals and combinations	Total
1909–1948[b]							
Alaska	6501	10,823	1095		773		19,192
Alabama	4412	63,586	9563		2889		80,450
Arizona	2,547,517	4403	27,497			101,880	2,681,297
Arkansas	1107	1520	15,043	85	40		17,795
California	2,352,070	3005	156,783		23	1864	2,513,745
Colorado	4,271,042	1,348,288	215,423		38,494		5,873,247
Florida	1154		2304	71,259	520		75,237
Idaho	1,291,163	11,749	4940	70,036	216,060		1,793,948
Illinois	634	120					754
Iowa	359						359
Kansas	54,384		1421				55,805
Louisiana	1223		17,105		3844		22,172
Michigan	1935		3261				5196
Minnesota	235			8			243
Mississippi	974		10,231				11,205
Missouri	166						166
Montana	3,993,640	6,658,554	987,472	11,290	17,788	150	11,668,894
Nebraska	72,964		3253				76,217
Nevada	242,717	960	1119		80	40	244,916
New Mexico	6,378,118	614,779	112,995	680	70,673	2,092,091	9,269,336
North Dakota	134,578	4,636,851	11,915		1164	40	4,784,548
Ohio	38					744	782
Oklahoma	48,781		10,917				59,698
Oregon	1,639,742	5598	14,369		480		1,660,189
South Dakota	1,565,802	187,722	6328				1,759,852
Utah	856,083	215,528	98,922	1576	8157	1680	1,201,946
Washington	262,444	14,535	2518		384	400	280,281
Wisconsin	1546						1546
Wyoming	9,541,179	2,297,363	376,906	32,037	17,341	257	12,265,083
Total through 1948	35,272,508	16,075,384	2,091,380	406,971	378,710	2,199,146	56,424,099
Fiscal period[c]							
1949–1950	363,466	24,868	18,287	1138	7662	4235	419,656
1951–1960	1,258,347	64,204	650,841	5075	109,323	36,586	2,124,376
1961–1970	2,103,335	44,578	893,114	559	498,973	30,663	3,571,222
1971–1980	2,914,664	9344	186,689	340	185,764	49,784	3,346,585
1981–1990	2,694,395	16,609	294,416	439	159,006	166,558	3,331,423
1991	38,353		1603		3990	5534	49,480

Continues

Table IV continued

State	Type of mineral reservation (acres)						
	All minerals	Coal	Oil and gas	Phosphate	Oil and gas plus other minerals	Miscellaneous minerals and combinations	Total
1992	49,341	698	71,773		2486	2261	126,559
1993	41,431	1788	3034		8963	4096	59,312
1994	30,322		11,781		11,061	1354	54,518
1995	37,612	1680	38,630		21,265	27,033	126,215
1996	36,311	2271	57,483		3532	1988	101,585
Total, 1949–1996	9,567,577	166,040	2,227,651	7551	1,012,025	330,092	13,310,931
Grand total	44,840,085	16,241,424	4,319,031	414,522	1,390,735	2,529,238[d]	69,735,030

[a] Data prior to 1949 were tabulated from incomplete records. Data include patents issued on both BLM and Forest Service lands.
[b] Minerals began to be reserved with the Agricultural Coal Lands Act of March 3, 1909; acts expanding on this and covering other minerals were suvsequently enacted.
[c] As of June 30 through 1976; thereafter, as of September 30 each year.
[d] Does not include 165,712 stockraising homesteads comprising 70,362,925 acres that reserved all minerals to the United States. Some of these homesteads were entered but not patented.

(2) competitive bidding for designated tracts, which used lotteries, oral auctions, sealed bids, or other formats to disperse the rights to KMAs and transfer them into known leasing areas (KLAs).

Eventually, the Competitive Bidding Act would be made even more specific. For coal, the law differentiated between new or maintenance tracts, but still required bids for existing operators to continue mining an existing site. Oil and gas development used reservoir studies to determine the pattern of leasing and lottery systems. In 1987, lottery systems were discontinued for oil and gas and development and were replaced by a familiar pattern in which industry representatives nominated tracts, which were then assessed for their environmental acceptability.

Neither the 1872 Mining Law nor the 1920 Mineral Leasing Act contained any direct environmental controls, but mining claims are subject to all general environmental laws as a precondition for development. These restrictions and requirements grew out of additional legislation, particularly the seminal environmental policies of the 1960s and 1970s, including the Multiple Use Sustained Yield Act, Wilderness Act, National Forest Management Act, National Environmental Policy Act (NEPA), and Federal Land Policy and Management Act. To varying degrees, each of these policies addressed environmental protection, multiple use, and management of federal land generally. By imposing new requirements on agency actions, and by withdrawing some federal lands from development, these acts have significantly influenced mineral development on public lands.

By far, the most critical policies relating to the administration related to the formation and structure of the BLM. When it was formed by merging two portions of the U.S. Department of the Interior (the General Land Office and the Grazing Service) in 1946, the BLM became entrusted primarily with the management of the remaining public domain and railroad/wagon road grant lands. The BLM also has the primary responsibility for the management oversight of the mineral domain that underlies the public lands. As with the sister agencies, the BLM is heavily involved in forest and rangeland planning, as well as keeping the official land status records of all public federal lands. Over time, the BLM followed the lead of the U.S. Department of Agriculture (USDA) Forest Service and adopted policies of multiple use and, eventually, of ecosystem management. In many cases, BLM lands are adjacent to or intermingled with other public lands, particularly National Forest land. The BLM essentially operated under these previously existing acts until congressional discussion over forest management in the Forest Service came to a head in the mid-1970s. This debate resulted in the passage of the Federal Land Policy and Management Act of 1976 (FLPMA), which provided the BLM with its "organic act"—albeit three decades after its founding.

FLPMA brought the greatest change to the administration of public lands in American history. Through FLPMA, BLM acquired an unequivocal

statutory basis for its public land management policy; however, debate continued over what exactly this meant. The 1934 Taylor Act had provided for management of natural resources on public land "pending its final disposal." To some observers, the phrase spoke to the old policy of transfer of public lands into private ownership; to others, it provided an argument against federal investment in land administration or rehabilitation. FLPMA, however, clearly prescribed retention of public lands in federal ownership as the rule and made transfer a carefully circumscribed exception. FLPMA established or amended many land and resource management authorities, including provisions on federal land withdrawals, land acquisitions and exchanges, rights-of-way, advisory groups, range management, and the general organization and administration of BLM and the public lands. FLPMA also called for public lands and their resources to be periodically and systematically inventoried. The BLM was empowered to use the data to create a plan for each site's present and future. Additionally, after many years of debate, FLPMA also specified that the United States must receive fair market value for the use of public lands and their resources.

Finally, BLM was required to administer public lands in a manner that protected the quality of scientific, scenic, historical, ecological, environmental, air and atmospheric, water resource, and archeological values. Where appropriate, BLM was also empowered to preserve and protect certain public lands in their natural condition, particularly when such preservation would provide food and habitat for fish and wildlife. Of course, BLM also remained responsible for providing sites for outdoor recreation and human occupancy and use of such sites. In short, FLPMA proclaimed multiple use, sustained yield, and environmental protection as the guiding principles for public land management. Thanks to FLPMA, BLM took on the responsibility of administering public lands so that they are utilized in the combination that will best meet the present and future needs of the American people for renewable and nonrenewable natural resources. Although this remains the role of BLM today, there have been many exceptions to this jurisdiction.

5. MAKING LEGAL EXCEPTION: NATIVE AMERICAN COMMUNITIES

There remain areas of federal lands that do not cohere to these legal regulations. For instance, it is reported that 80% of the nation's uranium reserves, 33% of the nation's low-sulfur coal, and 3–10% of the recoverable oil and gas are located on Native American reservation land. Problems of jurisdiction and ownership have highlighted the history of resource extraction on native lands. More important, though, are the health, social, and environmental impacts that have been noted where such development has been allowed to take place.

Outside sources approached Native American reservations for energy resources as early as 1920. For instance, the Navajo's prolonged involvement with oil development began with the influence of New Mexico's own Secretary of the Interior Albert B. Fall (who will be discussed later for his pivotal role in the management of energy resources on federal land). With one of Interior's charges being to administer Native American reservations, Fall soon looked toward reservation land as a vast, untapped (and unprotected) resource. Native peoples, Fall wrote, were "not qualified to make the most of their natural resources." Using the General Leasing Act of 1920, Fall attempted to clarify the status of treaty lands for leasing so that the government could do the leasing for the Native Americans. Although Fall's perspective grew from a desire to battle waste, public sentiment after World War I held that the United States would run out of oil within 10 years. Fall saw to it that oil companies won leases on reservation land, particularly that of the Navajos in New Mexico. In addition, he conspired with oil companies to provide native occupants with ridiculously low royalty rates. Historian Kathleen Chamberlain notes that oil intensified factions within the native community. In sum, she writes, "oil also expedited a shift from subsistence and barter to a wage economy." Although oil generated some jobs for the community, it never neared the promised levels.

Throughout the history of energy development on Native American lands, the development has been found to undermine tribal and family values and identity. For instance, in Hopi and Navajo experiences, wealth from royalties was spread unevenly and the large influx of wealth disturbed traditional hierarchies within the culture. Additionally, the non-Native American workers, who were necessary to energy companies for reservation development of energy resources, represented a persistent threat to tribal sovereignty. Possibly the most invasive change that energy development has wrought on Native American groups is on the traditional political process. In nearly every case, the debate over leasing lands for extraction has bitterly divided tribal leadership,

ultimately resulting in long-term impacts on the political structure of the group. With such breakdown, the reservations become what scholars have called a "colonial reservation economy." Some of the most troubling instances involved negotiations with the U.S. federal government. During World War II and through the Cold War, uranium was mined from Hopi and Navajo reservations in Arizona. The mine remnants as well as the remaining waste have contributed to significant rises in cancer and other health problems. Additionally, many Hopi and Navajo men worked in the mines and have therefore contracted a variety of ailments, particularly lung cancer.

Since the 1970s, Native American groups have attempted to use the courts to acquire fair royalty rates and recompense for health hazards. This has been an uphill battle, with various successes starting in the late 1980s. The Native American community's lessons about the community impact of energy development blossomed into an organization, the Council of Energy Resource Tribes (CERT), which helps each tribe to complete independent management of its energy resources. This has helped many tribes to acquire fair royalty rates as well as to initiate extensive development of alternative sources on western reservation land.

6. HISTORY OF ENERGY DEVELOPMENT ON PUBLIC LAND

The laws and ideas relating to energy development on public lands have changed due to shifts in larger ethics within American society, but also due to the type of resource being extracted. Primarily, the pressure to mine energy resources grew as resources such as petroleum became more and more critical to American life and national security. The most significant change followed the increased use of petroleum for transportation after 1900. For instance, by World War I, the energy reserves on public land had been transformed from an attraction for potential settlement to a matter of national security. In World War I, fossil fuels became directly linked to the nation's ability to protect itself and maintain global security. Petroleum, however, is only one example. In this section, three specific energy resources and the time period that they most influenced are considered.

6.1 The Beginnings with Coal

The first energy resource to be granted unique status on public lands was coal. The Coal Act of 1864 gave the U.S. President authority to sell coal beds or fields for no less than $20/acre. Lands not sold in this fashion could be available to the general public at the minimum price under the general land statutes. In 1873, Congress authorized citizens to take vacant coal lands of not more than 160 acres for a minimum of $10/acre. This law governed the disposal of coal lands until 1920. Precious metals were governed much more freely. In all probability, the rationale behind this distinction lay in the relative ease of locating coal compared to that of most precious metals.

The importance of the coal supply to national development would make it a major political issue in the early 20th century. In rapid succession, the Roosevelt administration proposed controlling the federal grazing lands, increasing the national forests, and withdrawing the coal, oil, and potash lands from sale. In this same period, the administration also reserved potential waterpower sites and set out to reclaim the arid lands of the West. In 1906, Roosevelt faced up to the rumors of random exploitation of coal supplies on federal lands by withdrawing coal-bearing lands from public availability. He responded to industry's complaint that the acreage limitations in the 1873 statute impeded the development of certain types of coal. As a result, many coal and railroad companies evaded these limits by using dummy entries or by entries processed under the agricultural land statutes. To keep certain companies from acquiring a monopoly and to conserve mineral supplies, the withdrawals began in November 1906. Located in Colorado, North Dakota, Montana, Oregon, Washington, Utah, Wyoming, and the Territories of New Mexico, 66 million acres of land with supplies of "workable coal" were withdrawn. In his autobiography, Roosevelt explained his logic as follows:

> *The present coal law limiting the individual to 160 acres puts a premium on fraud by making it impossible to develop certain types of fields….It is a scandal to maintain laws which sound well, but which make fraud the key without which great natural resources must remain closed. The law should give individuals and corporations under proper government regulation and control…the right to work bodies of coal large enough for profitable development. My own belief is that there should be provision for leasing coal, oil and gas rights under proper restrictions.*

Many western congressmen reacted angrily to Roosevelt's withdrawal. They claimed that Congress and not the Chief Executive possessed the right to dispose of the public domain in such a manner. Some

referred to Roosevelt's use of conservation as "state socialism." Roosevelt revised his argument slightly in 1907 to include mineral fuels with forests and navigable streams as "public utilities." Between 1906 and 1909, public coal lands sold for between $75 and $100 per acre. During these same years, over 4 million acres were segregated, though many would be returned to the public domain.

6.2 Contested Development: Petroleum and Teapot Dome

Following the 1859 strike of the first oil well in Pennsylvania, development in the United States moved to the Midwest. As oil exploration moved westward, it became evident that the public lands held great stores of "black gold." By the 1870s, oil fever had swept specifically into the public domain of California. In the early 1890s, oil was discovered near Coalinga, California and in 1892, E. L. Doheny drilled the first successful well in Los Angeles. The Placer Act of 1870 included "all forms of deposit" except veins of quartz. Though the Department of Interior debated whether petroleum qualified, Congress passed an 1897 Placer Act that upheld that oil lands were administered by placer location. In 1909, President Taft temporarily withdrew 3 million acres of potential oil lands in the public domain.

Similar to Roosevelt's coal withdrawals, Taft's act garnered a great amount of debate—particularly in the West. Congress then passed the Pickett Act in 1910 to allow presidents to withdraw public lands for examination and reclassification for various "public purposes." During the 1910s, this act enabled the President to withdraw nearly all of the known oil lands. In order to conserve an adequate supply of oil for the Navy, two naval reserves were established by executive order in 1912; one was in Elk Hills, California and the other was in Buena Vista Hills, California. Similar to a forest reserve, the oil supply, administrators believed, was best preserved if left in the ground. These withdrawals were then followed in 1915 by reserves at Teapot Dome, Wyoming and in 1916 by the naval oil shale reserves in Colorado and Utah. In 1923, a similar reserve was created in Alaska.

Although most of the public debate centered on the right of the executive branch to make such withdrawals, other issues emerged as well, including whether the land should be sold or leased to oil companies, whether land previously patented to railroad companies could be withdrawn by the federal government as well, and whether oil companies had any standing if the federal government wished to withdraw land that they had already leased. These issues, of course, took on greater importance as the United States observed and then participated in World War I at the end of the decade. Petroleum had many military uses, including powering many of the ships of the new Navy.

Historian John Ise was one of the loudest critics of such government policy. His argument, however, differed from those of most of the critics. He wrote in 1926 that "during this time there was always overproduction of oil from privately owned lands and there was never any need or justification for opening any public lands." Although this argument shifted to the fore by the end of the 20th century, there was little public support for the idea of conservation. Western lawmakers focused on ensuring federal supplies of petroleum on public lands when it passed the Mineral Leasing Act of 1920, otherwise known as the Smoot Bill. Congress intended this act to bring relief to some of these issues and to offer more equal treatment to interests of the western United States. In the naval reserves, leases were given only on producing wells, unless the President chose to lease the remainder.

Within 6 months after the passage of the Mineral Leasing Act, the Secretary of the Navy was granted extensive jurisdiction over the naval reserves. In 1921, Albert B. Fall, a senator from New Mexico, was appointed Secretary of the Interior by Warren G. Harding. In short order, Fall convinced Secretary of the Navy Edwin Denby to relinquish control of the reserves and had Harding sign a secret presidential order to this effect. Within a few months, word leaked to the public that Fall had leased reserves in California and Wyoming to his friends in the oil business. The episode that transpired in full view of the public became the primary test of jurisdiction and law regarding energy development and the West.

In the scandal known as Teapot Dome, a congressional committee found that Fall had acted secretly and without competitive bidding to lease the Teapot Dome to Harry Sinclair and the California tract to E. L. Doheny. Allowing a private company to develop naval oil reserves, Fall went outside the boundaries of existing law. As a result, he was one of the first members of a presidential cabinet to serve a jail term; he was convicted of bribery and ordered to serve a year in prison as well as to pay a $100,000 fine.

Fall represented the western interests who felt betrayed by the overly zealous control of the federal government over the development of public lands. Before Warren Harding appointed him as Secretary of the Interior, Fall had earned his spurs in farming and

ranching. In New Mexico, he owned a ranch that spanned more than 55 miles long and 35 miles wide. He had been a judge and a sheriff. Throughout his public life in New Mexico, Fall referred to the territory as "corporation country." He hated government interference in business, and he did not have to struggle with torn loyalties when he became a federal employee. "I have long believed," Fall said in 1918, "that the laws should protect…capital and should be so enforced as to offer an inducement for its investment." In giving business the advantage of governmental favors, Fall had, however, acted as a man of a certain faith. The core of his faith was simple: because nature's resources were unending, there was no need for humanity to restrain its ambitions and desires. "All natural resources should be made as easy of access as possible to the present generation," Fall declared. "Man cannot exhaust the resources of nature and never will." This, of course, included those resources falling within the public domain.

When Americans began to hear about Teapot Dome in news reports in the 1920s, much of the public participated in a growing appreciation of nature. Fall's feelings were well known to many observers. On his appointment in 1921, one newspaper wrote that "To us Fall smells of petroleum and we know what that means among...politicians." Another paper, also prior to the Teapot Dome scandal, described the new administrator of federal lands as a man who found "wise conservation of natural resources...hateful to his every instinct." Although Fall would be held accountable for his abuse of power on the federal lands, his basic idea of unfettered development and unending abundance remains on the table. This has been a major portion of the debate over petroleum development on federal lands. Although Congress would act in 1928 to give the Secretary of the Navy exclusive jurisdiction of the petroleum reserves, the matter would continue to be debated throughout the 20th century.

6.3 Energy and National Security: Uranium

Similar debate followed the post-World-War II search for uranium to be used in nuclear weapons and eventually in nuclear power generation. The demand for fissionable source materials after the war and the urgency of a government monopoly in them prompted the enactment of the Atomic Energy Act of 1946, which reserved all such ores in public domain land to the United States and required that they also be reserved in any mineral patent.

The uranium boom reached its height around 1954, when prospectors flooded western lands in fashion similar to the California gold rush. Private prospectors made millions of dollars leasing and locating uranium supplies through the end of the 1950s. Applicable to both uranium and oil and gas deposits, the Multiple Mineral Development Act of 1954 clarified how lessees were to deal with dormant mining claims that may cloud their own title. The new act provided a procedure in which title examiners may be used. After the lessee filed in the proper county office, he or she then waited for 90 days before filing with the BLM.

7. DEBATING OWNERSHIP: THE SAGEBRUSH REBELLION

Picking up on the ideas of Albert Fall, a group of 11 western states rallied in the 1970s to gain control of the development of the resources on their public lands. In what has become known as the "Sagebrush Rebellion" and the "Wise Use Movement," these groups have gained national attention and have often been able to bypass FLPMA and other government regulations.

In the late 1970's, the movement flared up with the support of Adolph Coors and other westerners. Prominent in this legal effort, the Rocky Mountain Defense Fund employed talented lawyers James Watt and Ann Gorsuch. Prior to his election, Ronald Reagan espoused support for the Sagebrush Rebellion. When he was elected in 1980, Reagan wasted no time in appointing Watt the Secretary of the Interior and Gorsuch to head the Environmental Protection Agency. Watt pressed his ideas forward and was able to open thousands of acres of public land to coal mining. However, with his forced resignation in the 1983, the movement lost a great deal of its support from the Reagan White House.

The same ideas, however, have found new life in the administration of George W. Bush. Guided by leaders such as the Assistant Secretary of the Interior J. Steven Griles, who served under Watt and then spent years as a consultant to the energy industry, the Bush administration pushed to open up vast new tracts of federal land to oil, gas, and coal development.

8. A SYMBOLIC CONTEST: ALASKA

With the growing strength of the environmental movement after 1970, the utilitarian view of public

lands met public debate at nearly every juncture. This has been most evident in an ongoing symbolic fight over federal lands in Alaska.

The energy crisis of the 1970s, which was exacerbated by the 1973 Oil Embargo carried out by members of the Organization of Oil Exporting Countries (OPEC), caused a significant shift in feelings about oil development in federally owned areas of Alaska. The stalled debate over drilling in northern Alaska and constructing a trans-Alaska pipeline (TAP) to Valdez gathered congressional support and was ultimately carried out in the late 1970s. American leaders assured the public that the TAP would be a step toward energy independence. Specifically, the Nixon administration assured Americans that a TAP would cause oil imports from the Middle East to fall from 27% of the U.S. supply to 20% by 1990.

President Jimmy Carter's willingness to allow these developments was predicated on his effort to formalize restrictions on further such activities in undeveloped areas of Alaska. From 1978 to 1980, President Jimmy Carter established the Alaska National Wildlife Refuge (ANWR) and set aside 28% of Alaska as wilderness. Carter announced this watershed legislation as follows: "We have the imagination and the will as a people to both develop our last great natural frontier and also preserve its priceless beauty for our children." With the passage of the Alaska National Interest Lands Conservation Act (ANILCA) in 1980, the BLM was ordered to oversee 11 million acres of Alaska as wilderness. This included the Alaska National Wildlife Refuge.

Debate on further development in these areas of Alaska awaited the administration of George W. Bush. As part of the administration's energy plan, new priority was placed on developing energy resources on federal lands. Secretary of the Interior Gale Norton specifically zeroed in on ANWR in 2000. Congress rebuffed the first effort to open ANWR to oil development, but other efforts are sure to follow. With the Republican victory in the 2002 congressional elections, most environmentalists foresee little to stand in the way of energy developments in ANWR.

9. BATTLE LINES IN THE CURRENT DEBATE

Thanks to FLPMA, most new energy extraction requires an environmental impact statement (EIS), which is a detailed study of the proposed plan and how it will affect wildlife, plant life, and the quality of the water, air, and soil. The EIS also examines how the resource extraction will affect other uses of the public lands, such as recreation, cultural resource conservation, or grazing. If the impacts of the proposed well, dam, or mine seem too great, then alternatives are examined. These range from scrapping the plan entirely to using advanced technology to extract the resource, such as directional drilling, whereby the well is sited in a less vulnerable area but the drill still reaches the underground reserves.

The debate over ANWR crystallizes many of the issues that now make up the larger debate about how energy development should proceed on federal lands. During the administration of President William Clinton, oil and gas leasing increased considerably over previous administrations, but with new environmental restrictions on industries that sought to develop energy resources on public lands (see Table V). President Clinton also designated 21 western national monument areas and 56 million acres of roadless area. Such changes, however, were susceptible to any changes in national politics. That change came in the election of 2000.

"Environmental issues today," instructed Bush's Interior Secretary Gale Norton, "are more complex and subtle than the ones we faced in the 1960s and 70s. With the growth of our population and expansion of our economy, pressures have increased correspondingly on our undeveloped land, water resources, and wildlife." Norton intends to put "collaboration ahead of polarization, markets before mandates…and to transcend political boundaries." In order to do so, she and George Bush coined the phrase "new environmentalism" to meet the U.S. need "for a vibrant economy and energy security—while at the same time protecting the environment." Thus far, environmental organizations have been disappointed by Bush and Norton's "new environmentalism." "We're dismayed that Pres. Bush wants to turn the oil and gas industry loose on virtually all the lands in our national forests, national wildlife refuges, and other public systems," said Wilderness Society representative Bill Meadows. "We're beginning to wonder if James Watt is whispering in his ear," he said.

In addition to ANWR, federal lands along the Rocky Mountain front in northwestern Montana to southern Wyoming's Red Desert and Colorado's Vermillion Basin are being explored for energy development. The BLM collected data on oil and gas deposits and their availability throughout the

TABLE V

Continuing Oil and Gas Activities on Federal Lands as of September 30, 1996

State	Number of units in effect[a]	Number of producible and service holes[b]	Number of producible and service completions[b]	Number of producible leases[a]	Number of acres in producing status[a]
Alabama	4	40	43	18	6222
Alaska	9	150	161	38	67,350
Arkansas	2	311	334	158	73,353
California	26	5693	6122	350	79,102
Colorado	136	3098	3331	2122	1,397,557
Florida	—	4	4	2	3468
Illinois	1	13	14	5	964
Kansas	8	476	512	436	110,464
Kentucky	—	135	145	43	24,132
Louisiana	12	386	415	95	38,289
Maryland	—	9	10	—	—
Michigan	20	70	75	57	29,900
Mississippi	7	195	210	85	45,268
Montana	50	3318	3568	1307	679,084
Nebraska	6	36	39	20	6507
Nevada	16	91	98	57	44,300
New Mexico	231	24,367	26,201	6122	3,512,145
New york	—	2	2	4	1010
North Dakota	32	691	743	490	307,774
Ohio	—	308	331	125	19,162
Oklahoma	19	269	289	775	121,376
Pennsylvania	—	54	58	67	6967
South Dakota	7	181	195	76	48,921
Tennessee	—	10	11	3	2366
Texas	14	207	223	163	68,931
Utah	112	1919	2063	1236	914,203
Virginia	—	30	32	9	5126
West Virginia	—	280	301	174	104,065
Wyoming	518	13,984	15,037	6049	2,514,920
Total	1230	56,327	60,567	20,086	10,514,920

[a] Data from the Bureau of Land Management Case Recordaion System.

[b] Service holes and completions are not neessarily located on producible leases. Data from the Bureau of Land Management Automated Inspection Record System.

public lands in order to support the prodevelopment stance of the Bush administration. Many local landowners and environmentalists have fought these efforts. Norton has requested new financing to streamline permits and study sites to drill on federal lands. The Bush administration has also shown some interest in increasing the development of alternative energy sources on federal lands as well. The future of the public lands is certain to contain open debate over the domain's role in America's energy future.

Acknowledgments

The author thanks the Center of the American West, University of Colorado, Boulder, and the William and Flora Hewlett Foundation for assistance in preparing this article.

SEE ALSO THE FOLLOWING ARTICLES

Arid Environments, Impacts of Energy Development in • Development and Energy, Overview • Ecological

Risk Assessment Applied to Energy Development • Environmental Injustices of Energy Facilities • Oil-Led Development: Social, Political, and Economic Consequences • Polar Regions, Impacts of Energy Development • Population Growth and Energy • Sustainable Development: Basic Concepts and Application to Energy • Wetlands: Impacts of Energy Development in the Missippi Delta

Further Reading

Carstensen, V. (1963). "The Public Lands; Studies in the History of the Public Domain." University of Wisconsin Press, Madison.

Center of the American West (2003). "What Every Westerner Should Know about Energy." http://www.centerwest.org/

Chamberlain, K. (2000). "Under Sacred Ground: A History of Navajo Oil, 1922–1982." University of New Mexico Press, Albuquerque.

Clawson, M. (1987). The Bureau of Land Management, 1947–1953. *Cruiser* [Newsletter of the Forest History Society]. **10**(3), 3–6.

Gates, P. W. (1968). "History of Public Land Law Development." U.S. Government Printing Office for the Public Land Law Review Commission, Washington, D.C.

Gerard, D. (1997). "1872 Mining Law: Digging a Little Deeper." Political Policy Research Center (PERC) Policy Series PS-11. PERC, Bozeman, Montana.

Gordon, R., and VanDorn, P. (1998). "Two Cheers for The 1872 Mining Law." CATO Institute, Washington, D.C.

Hage, W. (1989). "Storm Over Rangelands: Private Rights in Federal Lands." Free Enterprise Press, Bellevue, Washington.

Ise, J. (1926). "The United States Oil Policy." Yale University Press, New Haven.

Leshy, J. D. (1987). "The Mining Law: A Study in Perpetual Motion." Resources for the Future, Washington, D.C.

Muhn, J., and Stuart, H. R. (eds.). (1988). Opportunity and Challenge: The Story of Blm. U.S. Government Printing Office, Washington, D.C.

National Research Council, Committee on Hardrock Mining on Federal Lands. (1999). "Hardrock Mining on Federal Lands." National Academy Press, Washington, D.C.

Robbins, R. M. (1976). "Our Landed Heritage: The Public Domain, 1776–1970." University of Nebraska Press, Lincoln.

Rohrbough, M. J. (1968). "The Land Office Business the Settlement and Administration of America's Public Lands, 1789–1837." Oxford University Press, New York.

Voight, W., Jr. (1976). "Public Grazing Lands: Use and Misuse by Industry and Government." Rutgers University Press, New Brunswick, New Jersey.

Williams, T. T. (2002). "Red: Passion and Patience in the Desert." Vintage Press, New York.

Energy Efficiency and Climate Change

ARTHUR ROSENFELD, PAT McAULIFFE, and JOHN WILSON
California Energy Commission
Sacramento, California, United States

1. **Introduction**
2. **Improving Trends in Energy Efficiency**
3. **The Combined Impact of Complimentary Efficiency Improvements**
4. **Efficiency Improvements Reduce Generation Requirements**
5. **Historic Primary Energy Trends in the United States**
6. **Energy Intensity Trends in the United States and Other Regions**
7. **Future Primary Energy Trends for the World**
8. **The Effect of Efficiency Improvements on Primary Energy Demand**
9. **Pathways to Improving Energy Intensity**

Glossary

energy efficiency Useful output divided by energy input.
energy intensity Energy consumption per unit of gross domestic or gross world product.
final energy Energy available for use in homes, factories, and for transportation after accounting for losses in conversion (for example, waste heat from an electric power plant).
global warming Increases in the earth's surface temperature brought about by the release of greenhouse gases.
greenhouse gases Gases in the atmosphere that tend to trap heat and contribute to global warming (for example, carbon dioxide).
gross domestic product (GDP) Total value of goods and services in a country.
gross world product (GWP) Total of gross domestic product for all countries.
kWh Unit of electrical energy equal to 3,413 Btus.
million tons of oil equivalent (Mtoe) 40 million Btus.
primary energy Energy available from sources such as petroleum and coal.
quads 10^{15} British thermal units.

Large quantities of fossil fuels are burned to meet the world's energy needs. The combustion of these fuels produces carbon dioxide, contributing to increasing concentrations of carbon dioxide in the earth's atmosphere. According to the Intergovernmental Panel on Climate Change, the increasing concentrations of carbon dioxide and other greenhouse gases in the atmosphere due to human activity are causing the average surface temperature of the earth to increase. Such human influences are expected to continue to contribute to rising greenhouse gas emissions and additional global warming through this century.

1. INTRODUCTION

Humans combust these fuels to provide energy for needs such as heating, transportation, refrigeration, and industrial processes. The efficiency with which fossil fuels can be converted to meet these needs determines how much primary energy will be combusted. The greater the end-use efficiency, the less fuel is needed and the less the impact that such human activities will have on global warming, all other factors remaining the same.

Over the past century, the efficiency of energy use has improved. For example, lighting has changed from the burning of candles that produced 1 lumen for every 6 watts of candle or wax burned, to incandescent electric lights that produced 17 lumens with each watt of electric input, to fluorescent lighting that produces 100 lumens per watt. So the energy intensity index, defined as the amount of primary energy needed to produce a unit of gross domestic product, continues to improve. If this index can continue to improve at a rate of 2% per year, the

 © 2004 Elsevier Inc. All rights reserved.

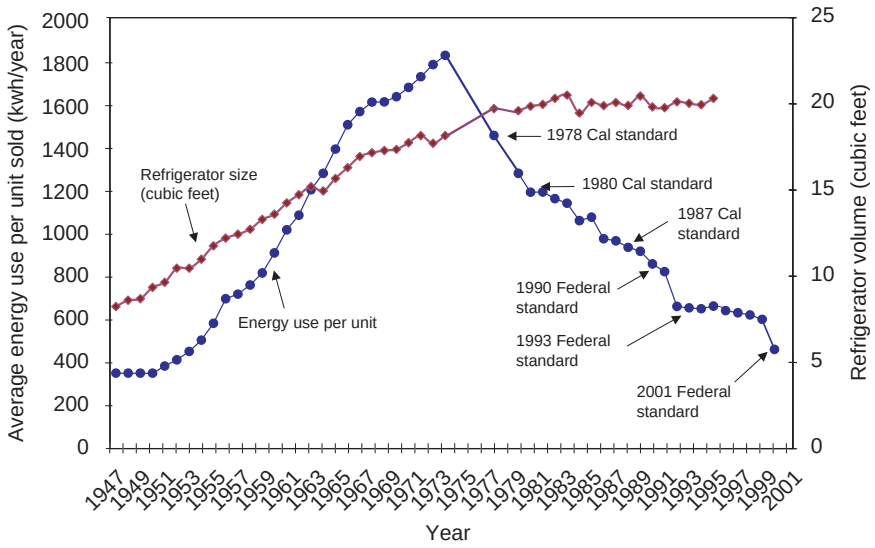


FIGURE 1 United States refrigerator use versus time. Annual drop from 1974 to 2001 is 5% per year. From David Goldstein of the Natural Resources Defense Council.

primary energy needs of the world can be limited even while increasing access to goods and services. Such an improvement in energy intensity also will lead to a reduction in carbon dioxide emissions and limit the extent of global warming.

2. IMPROVING TRENDS IN ENERGY EFFICIENCY

The efficiency of energy use is an important determinant of how much energy will be required to warm a home to a particular temperature or drive a car a certain distance. In our daily lives we often encounter the idea of energy efficiency when figuring how many miles per gallon our automobiles use while driving in town or on a freeway. Clearly, the more miles a car can travel on a gallon of gasoline, the less gasoline will be needed. The same concept applies to all uses of energy.

Since 1973, when the United States first experienced rapid increases in fossil fuel prices, considerable attention has been paid to improving the efficiency of energy use of various appliances and other products. The most effective path toward energy efficiency has been standards for autos, buildings, appliances, and equipment. Figure 1 shows the remarkable gains in refrigerators. Even while increasing in size, refrigeration has improved efficiency 5% every year since the late 1970s.

Figure 2 shows the results of these improvements. With an estimated 150 million refrigerators and freezers in the United States, the difference in consumption at 1974 efficiency levels compared to 2001 levels is the equivalent to avoiding 40 GW of power plants. A typical unit at a nuclear power plant is 1 GW, while the typical natural gas-fired combined cycle plant is 300 MW. This is based on saving 1400 kWh per year with the more efficient appliance and assuming power plant operation of 5000 hours per year.

Significant gains in efficiency also have been achieved with improvements in other appliances. Figure 3 illustrates improvements in three difference appliances (gas furnaces, central air conditioning, and refrigerators) compared to their respective consumption in 1972. This figure is taken from

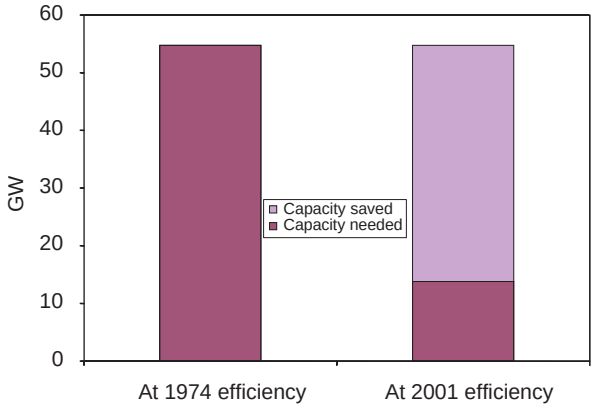


FIGURE 2 Electricity generating capacity for 150 million refrigerators and freezers in the United States.

Steven Nadel in the ECEEE 2003 Summer Study. Compared to consumption in 1972, new gas furnaces use 75%, central air conditioning units use 60%, and refrigerators only 25%. The figure also indicates that early standards were adopted by states. United States' national standards were not adopted until the late 1980s. Although these trends are encouraging, some other uses of energy have shown little improvement in efficiency since the 1980s. For example, average fuel efficiency for cars and truck remains at the same level as in 1985, even though there are no engineering obstacles blocking significant improvements in fleet efficiency. This was after significant improvement in CAFE standards from 18 miles per gallon for model year 1978 cars to 27.5 mpg for 1985 cars. However, even consideration of increasing CAFE standards remains a controversial topic.

3. THE COMBINED IMPACT OF COMPLIMENTARY EFFICIENCY IMPROVEMENTS

Efficiency improvements in appliance and building design and construction can often be combined to result in even larger energy savings. Figure 4 provides an example of this combined impact. Air conditioning operates to cool a home or a portion of a home based on the temperature setting of the

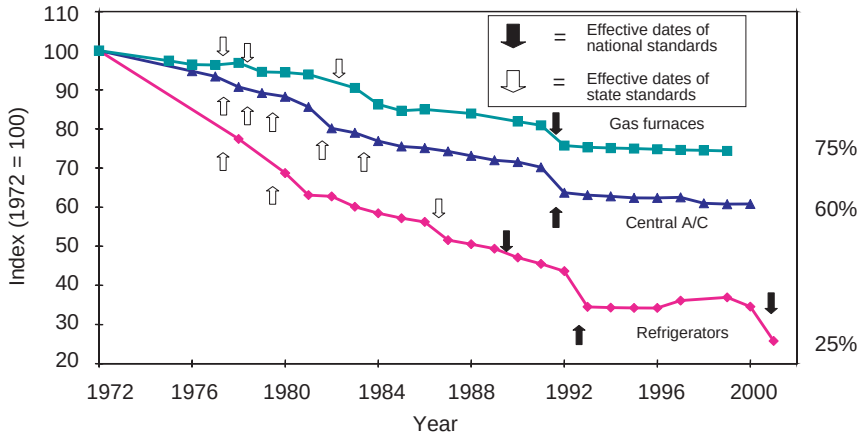


FIGURE 3 Impacts of standards on efficiency of three appliances. From S. Nadel, ACEEE, in ECEEE 2003 Summer Study, www.eceee.org.

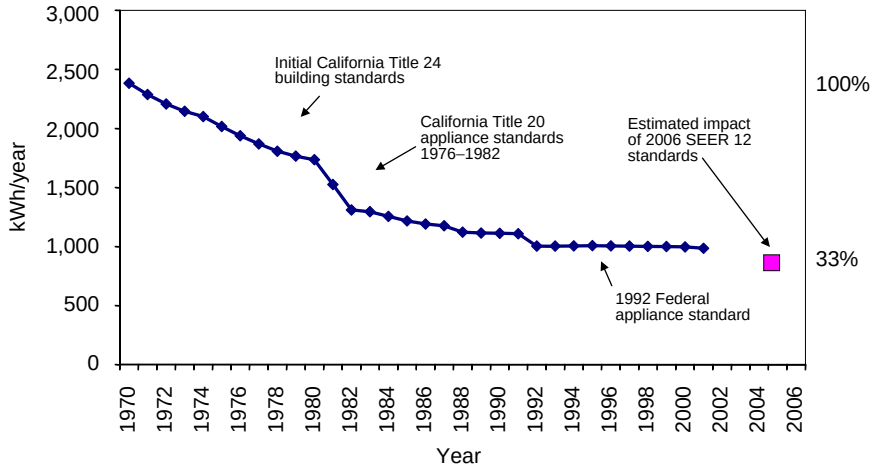


FIGURE 4 Annual usage of air conditioning in new homes in California. Annual drop averages 4% per year. From CEC Demand Analysis Office.

thermostat. The efficiency of the air conditioning equipment (for example, the compressor and air handlers) itself is known at the time of purchase and is required to meet a specific federal standard. However, how much it actually runs is also dependent on the characteristics of the building or home where the equipment is located. Prior to the 1970s many homes, especially in milder climates, had little or no insulation and single pane windows. As building standards were introduced, new homes were built to more stringent standards, resulting in lower requirements for heating and cooling to maintain the same temperature within the structure. As indicated in Fig. 4, the effect of combining both air conditioning standards and building standards has been to reduce energy use in a new California home to 40% of the usage in the early 1970s. With the implementation of newer standards in 2006, use is expected to decline to 33% of the 1970s levels.

4. EFFICIENCY IMPROVEMENTS REDUCE GENERATION REQUIREMENTS

Since the early 1970s, efficiency improvements have lowered energy demands and the need for peaking electrical power. To highlight the magnitude of such savings, we now consider how the U.S. demand for peak electrical power has been reduced, compared to what it might have been in the absence of efficiency improvements. The requirement for peaking power to meet air conditioner loads in the United States is about 250 GW. Standards only apply to the 200 GW of residential central systems and commercial rooftops air conditioning systems. By moving to current efficiency standards, equivalent to an energy efficiency ratio (EER) of 11, the United States has avoided about 135 GW of peak electricity load. This is about twice the peak load of the entire state of California and roughly equal to the total nuclear capacity in the United States. In addition, white roofs could save an additional 20 GW.

Enhancements in refrigerator and freezer efficiency save 200 billion kWh per year compared to consumption if refrigerators had remained at 1974 efficiency levels. That amount of electricity is nearly equal to annual electricity consumption in the state of California, about twice the amount of annual generation from existing renewables (not including hydroelectric) in the United States, and about one-third the quantity of electricity generated at nuclear stations in the United States in a year. Since the value of electricity at the point of consumption is worth nearly three times its value at the point of production, the value of improvements in refrigerator and freezer efficiency is worth the same as all the nuclear generation in the United States in a typical year.

Clearly, efficiency improvements have reduced the requirements for all forms of energy, even though some sectors, such as automobile transport, lag in this regard. Emissions of greenhouse gases have been proportionally reduced due to these efficiency gains. Yet it is generally accepted that the earth's surface temperature is on the rise, in great part due to human activity involving the combustion of fossil fuels. We now turn to past and estimated future trends in primary energy requirements.

5. HISTORIC PRIMARY ENERGY TRENDS IN THE UNITED STATES

Even with efficiency gains, energy requirements continue to increase. Figure 5 illustrates this trend. Primary energy use (not just electricity) increased by a factor of three over the time period of 62 years. Carbon dioxide emissions increased at nearly the same rate, as the vast majority of primary energy is provided from fossil fuels with only limited contributions from nuclear, hydroelectric, and other renewable energy sources.

Figure 6 shows a steady increase in per capita energy consumption from 1949 until the early 1970s. The flattening in 1974 is the impact of the OPEC oil embargo and high oil prices until the collapse of OPEC in 1985. Comparing Figures 5 and 6, we can see that much of the increase in primary energy is due to an expanding population, especially during the 1990s when per capita consumption was relatively flat.

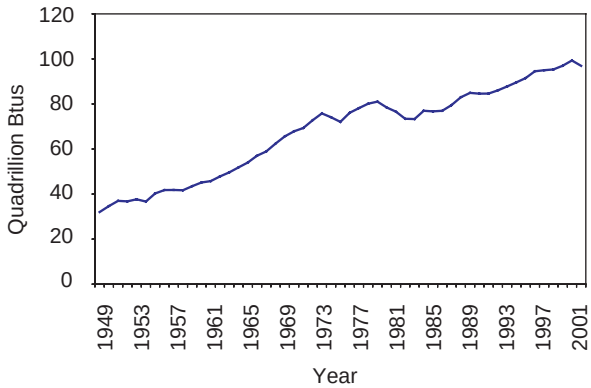


FIGURE 5 United States energy consumption, 1949 to 2001. From Table 1.5, *Annual Energy Review;* data for 2001 is preliminary.

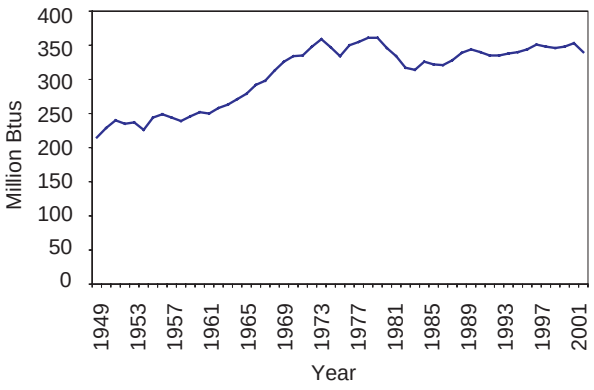


FIGURE 6 Energy consumption per person, 1949 to 2001. From Table 1.5, *Annual Energy Review*; data for 2001 is preliminary.

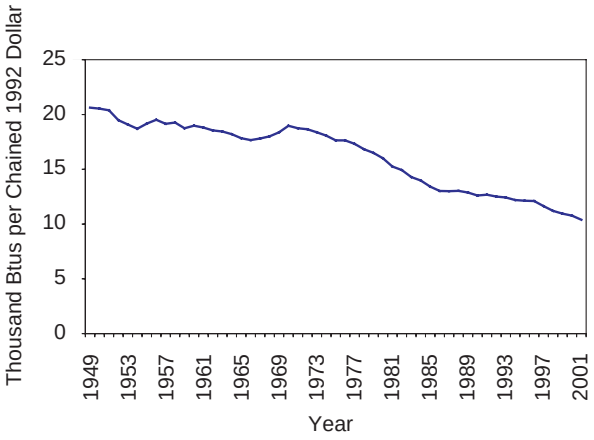


FIGURE 7 Energy consumption per dollar of gross domestic product, 1949 to 2001. From Table 1.5, *Annual Energy Review*; data for 2001 is preliminary.

6. ENERGY INTENSITY TRENDS IN THE UNITED STATES AND OTHER REGIONS

Another measure of interest to economists and energy analysts is how much energy is used per dollar of gross domestic product (GDP). This measure is referred to as energy intensity. The lower the number, the more efficient in terms of energy per dollar of GDP. Figure 7 shows that the United States improved its energy intensity at a rate of about 1.3% per year through this entire period. However, the rate of improvement is by no means uniform, as can be seen in the figure. Beginning in the early 1970s, the slope of the line changes as the rate of improvement in energy intensity increases.

Figure 8 provides similar data on energy intensities by geographic region. Five regions plus the world are included. Changes in energy intensity are not uniform throughout the world. Western Europe and North America have shown continual improvement over the time period, while other regions have remained flat or increased. China has shown dramatic and sustained improvement.

The next four figures illustrate annual rates of change in energy intensity data available from the United States Energy Information Agency (EIA). The EIA uses market exchange rate to compare gross domestic products across different countries. Market exchange rate is based on trading currency in international monetary markets.

Figures are provided for the United States, Western Europe, China, and the entire World. The

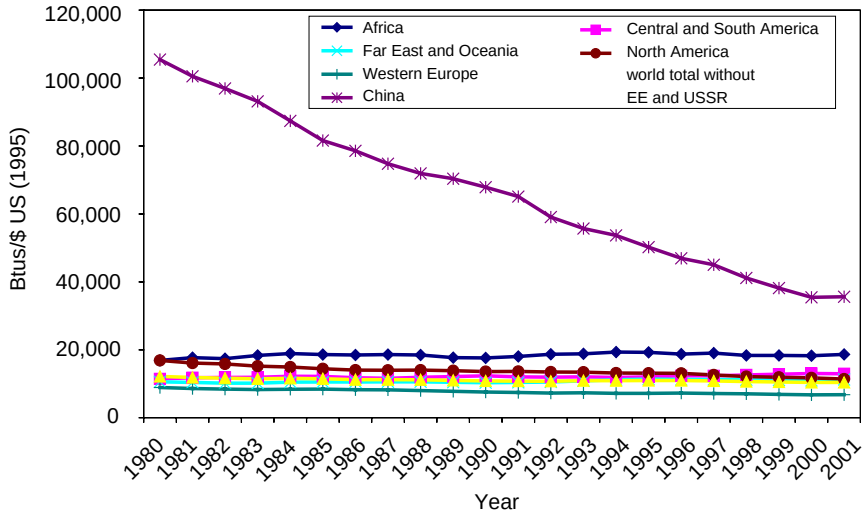


FIGURE 8 Energy intensity by geographic region including China, 1980 to 2001 (Btus/$US 1995). From Energy Information Agency (EIA) data.

vertical lines divide the time period (1981 to 2001) into three eras. Era 1 covers the latter years of high OPEC oil prices. Era 2 starts soon after the collapse of OPEC and the consequent stagnation of corporate automobile fuel economy (CAFE) standards. Era 3 begins with the significant impact of information technologies on business and the economy. The numbers written above each era reflect the average rate of change in E/GDP for that era. We introduce each of these figures and discuss trends of interest as illustrated in the figure.

Figure 9 tracks year to year changes in energy intensity and shows different rates of improvement in E/GDP in each of the three eras discussed earlier. In the early 1980s, the United States experienced improvements averaging −3.4% per year (improvements are indicated as negative numbers on these figures) and only −0.7% between 1986 and 1996. Improvements from 1997 through 2001 averaged −2.7%. Although many factors are at work, a number of analysts argue that the era improvements are the result of structural changes in the economy. The United States has moved away from energy intensive industry, such as steel production, and moved toward service and information-based firms, especially computer and internet firms. Such changes may portend not only less energy use for processes but improvements in scheduling, shipping, and other aspects of business practice due to the additional information availability.

Figure 10 shows a nearly constant improvement in the range of −1.3% per year and much less variability than in the United States. Europe already is considerably more efficient than the United States (see Fig. 8). In Fig. 11, we find nearly constant improvements in E/GDP in the range of −5.0% per year. Such improvement is dramatic. According to IEA's World Energy Outlook 2002, China's GDP has grown at about 8% per year since the 1970s to become the world's second largest consumer of primary energy. Finally, in Fig. 12 we find improvements in energy intensity averaging about −1.0% since the 1970s. Many factors contribute to the energy intensity of a county or a region within a country. As indicated by Figs. 8 to 12, the rate of change of energy intensity varies significantly throughout the world.

For energy intensity (E/GDP) to improve—that is, get smaller—GDP needs to increase more rapidly than primary energy, presuming that both are increasing. For example, primary energy use in the United States grew by 2.4% between 1999 and 2000. GDP grew by 3.7%, resulting in an improvement in energy intensity of 1.3% from 1999 to 2000. This article uses the convention of showing this as −1.3%. From the perspective of using energy better, the United States improved. Yet carbon emission also increased since the majority of primary energy comes in the form of fossil fuel—primarily coal, natural gas, and petroleum. If one is concerned about global warming, it seems that only three options are available to limit carbon emissions: (1) reduce the amount of primary energy requirements or limit the rate of growth in primary energy demand; (2) find

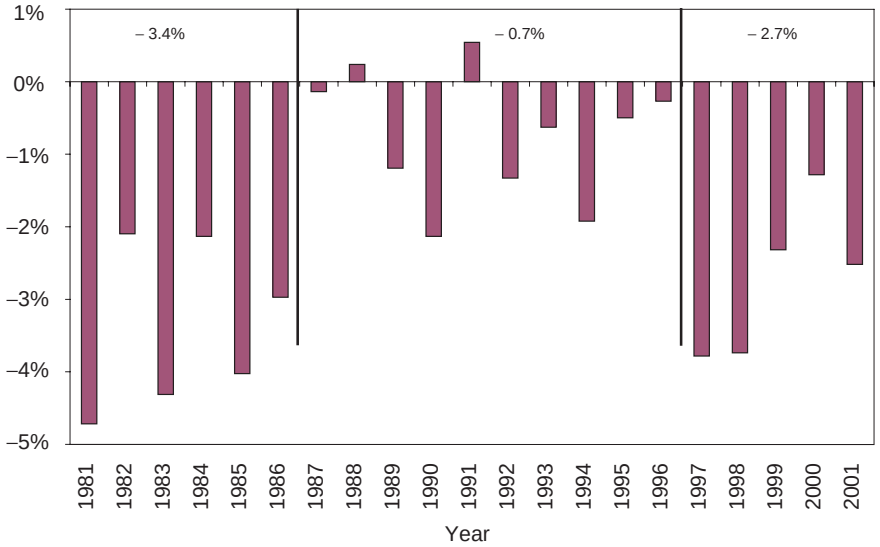


FIGURE 9 Annual rate of change in energy/GDP for the United States.

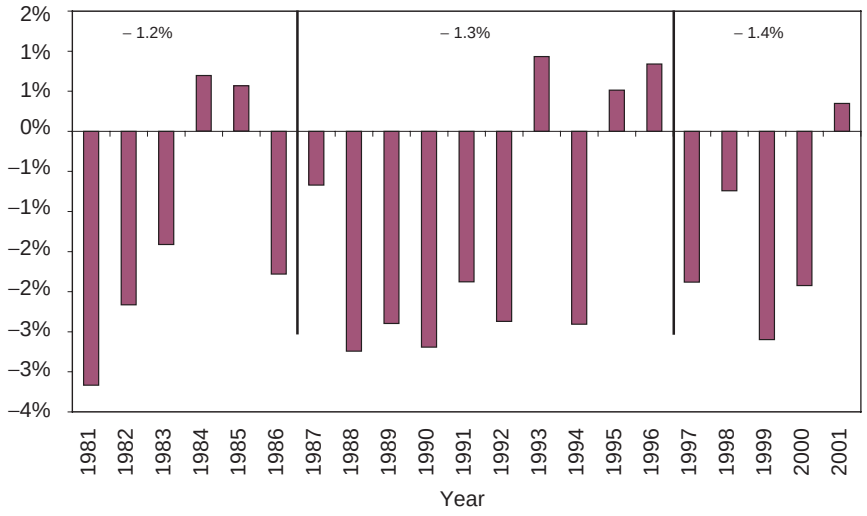


FIGURE 10 Annual rate of change in energy/GDP for Europe.

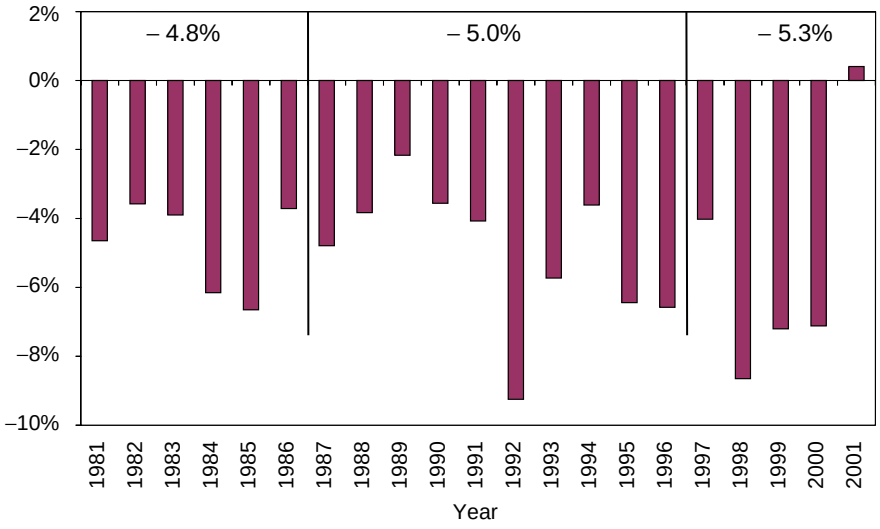


FIGURE 11 Annual rate of change in energy/GDP for China.

primary fuel sources that are not carbon based; and (3) sequester the carbon that is emitted. Not only are these options not mutually exclusive, but we expect all might well be necessary if we are to stem the tide of rising carbon concentrations in the atmosphere and global warming.

Even with improving E/GDP, concerns remain regarding primary energy demands for the world through the coming century and the amount of carbon that may be released into the atmosphere as fossil fuels are burned to meet these needs.

7. FUTURE PRIMARY ENERGY TRENDS FOR THE WORLD

Many different organizations have addressed the issue of how much energy might be needed in the future. Some have tried to forecast this amount; others have developed a host of scenarios to attempt to bound the magnitude of the problem. Estimates of future primary energy requirements vary considerably. Figure 13 illustrates the range of such scenarios. The scenarios all start at current levels on world energy demand, about 400 quads. By 2100, the range of

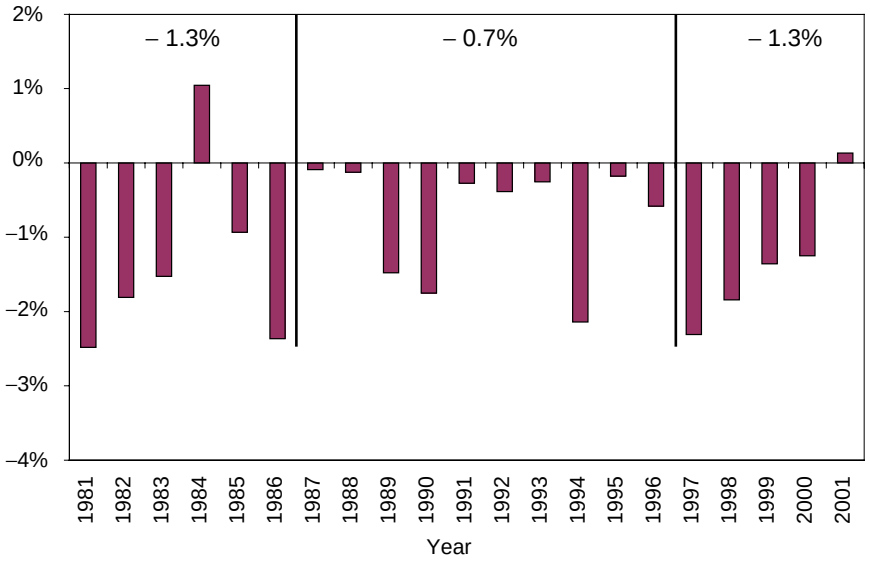


FIGURE 12 Annual rate of change in energy/GDP for the world. Note that Russia not included.

world primary energy demand is from 500 quads all the way to 2600 quads per year. These scenarios are predicated on different estimates of population growth, economic development, and energy intensity.

8. THE EFFECT OF EFFICIENCY IMPROVEMENTS ON PRIMARY ENERGY DEMAND

Efficiency improvements can drastically alter primary energy demand. We now discuss other cases that differ primarily in assumptions regarding the rate of improvement of energy intensity. Using the IPCC's IS 92a as the base case, Fig. 14 compares this case with two alternative trajectories of the world's primary energy consumption. The IS 92a case assumes E/GDP improving at −0.8% per year through 2020 and −1.0% per year thereafter through 2100. The second trajectory is taken from EIA's Annual Energy Outlook 2003. In the third case, energy intensity is assumed to improve at −2.0% annually through 2100. In assuming this improvement, total gross domestic product (or gross world product) is held at the same level as in the IS 92a case, and primary energy is set so that E/GDP improves by −2.0% per year.

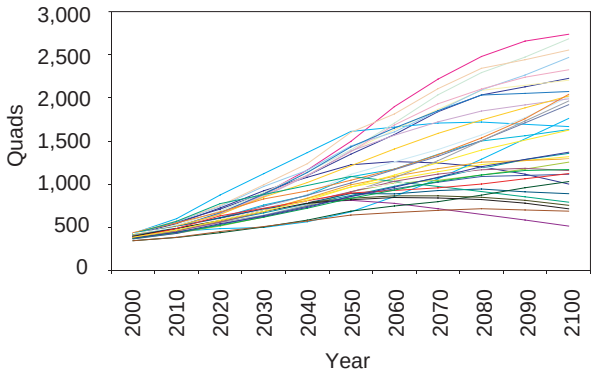


FIGURE 13 Primary energy 2000 to 2100. This figure is taken from data provided in IPCC Special Report on Emissions Scenarios 2001, Appendix VII.

Since the early 1990s when the IS 92a case was developed, primary energy demand has grown more slowly than anticipated. This explains the differences between the base case assumptions and the EIA International Energy Outlook as shown in Fig. 14. These differences result in a reduction of forecasted primary energy demand in 2025 of about 7.5%. However, if we assume that world energy intensity improves at a rate of −2% per year, world energy demand rises to only 510 quads by 2025 and then stays nearly flat through the end of the century. World demand is 400 quads per year.

The impact of improvements in energy intensity is dramatic. Primary energy demand is cut by more than 60% compared to the IS 92a base case. Put more dramatically, IS 92a requires, by 2100, in units of today's energy infrastructure, growth from 1.0 to 3.5, whereas the 2% per year scenario requires growth only from 1.0 to 1.3, at great savings of

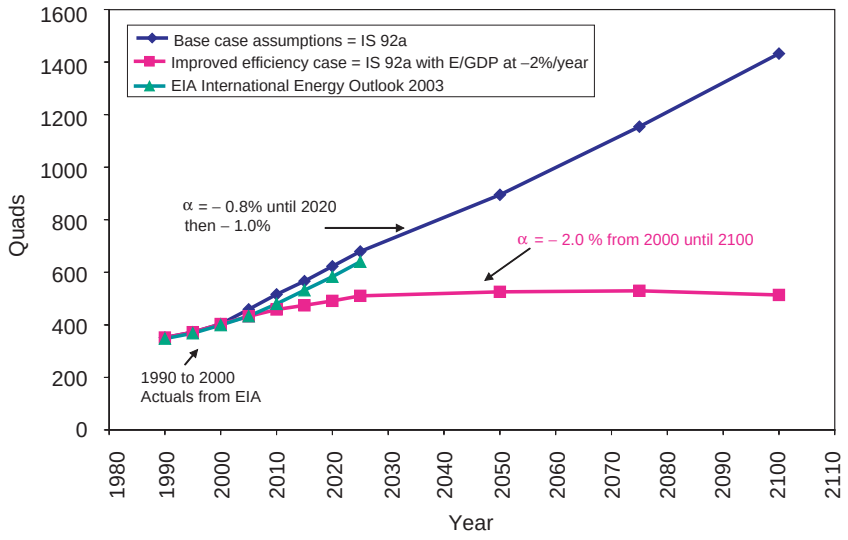


FIGURE 14 Primary energy demand in IPCC base case compared to other cases. α is the annual growth in energy intensity, E/GWP negative α indicates reduced energy intensity.

economic and environmental cost. While all this is occurring, per capita income rises to a worldwide average of nearly 25,000 (in 1995 US$). This is a level consistent with standards in developed European countries and Japan.

9. PATHWAYS TO IMPROVING ENERGY INTENSITY

We have seen that world energy intensity (E/GWP) is spontaneously dropping 1.3% per year, and that gaining another 1% per year for the coming century will cut demand and supply by another factor of 2.7. This extra 1% per year can by achieved by adopting and enforcing building, appliance, and vehicle fuel economy standards, and by ploughing about 1% of utility revenues back into efficiency programs designed to beat the standards. Examples of such programs are rebates for the purchase of super-efficient ("energy star") appliances, buildings, cars and trucks, and technical assistance to industries, architects, building owners and operators, and so on.

Standards are usually designed to reduce utility bills fast enough to repay a slightly higher first cost in 5 to 10 years, which is typically half the service life of the product. We call this a simple payback time (SPT) of 5 to 10 years. Efficiency measures to retrofit existing buildings and industry will also have SPTs that are safely shorter than the service life of the products installed. In contrast, investment in energy supply—power plants, transmission lines, refineries, dams, mines, and wells—typically has an SPT (from sale of energy) of 20 to 40 years, so efficiency is a better investment (although, of course, we need both), and the savings from efficiency, ploughed back into the economy, can stimulate jobs and growth.

Developed countries, which already have low energy intensity, can improve further by tightening and enforcing standards and extending their scope and by expanding efficiency programs. Reductions of CO_2 will then be an attractive by-product.

But later developing countries (LDCs), all of them lacking adequate energy supply, tend to focus on expanding their supply, and indeed rich countries should help them finance new supply. But first, for their own economic development and to reduce CO_2, more developed countries should attach conditions. Before financing supply with a 30-year payback time, developed countries should require that the host country develop standards and programs to capture all measures with a payback time of 3 to 5 years. Also, rich countries can help LDCs with technical assistance and training. These conditions should be adopted worldwide by donor countries, international banks, and host countries and their utilities. This is the cheapest way for the rich countries to delay climate change.

SEE ALSO THE FOLLOWING ARTICLES

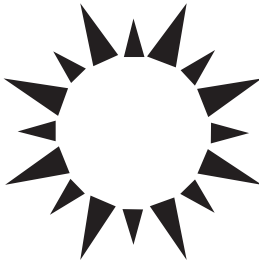
Carbon Taxes and Climate Change • Climate Change and Energy, Overview • Climate Change and Public Health: Emerging Infectious Diseases • Climate Change: Impact on the Demand for Energy • Climate Protection and Energy Policy • Consumption, Energy, and the Environment • Economics of Energy Efficiency • Energy Efficiency, Taxonomic Overview • Environmental Change and Energy • Greenhouse Gas Emissions from Energy Systems, Comparison and Overview • Obstacles to Energy Efficiency

Further Reading

"Climate Change 2001: The Scientific Basis." (2001). Contributions of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK.

Stewart, R. B., and Jonathan B. W. (2002). "Reconstructing Climate Policy Beyond Kyoto." AEI Press, Washington, DC.

Energy Efficiency, Taxonomic Overview



AMORY B. LOVINS
Rocky Mountain Institute
Snowmass, Colorado, United States

1. **Definition and Importance of Energy Efficiency**
2. **Benefits of Energy Efficiency**
3. **Engineering vs Economic Perspectives**
4. **Diminishing vs Expanding Returns to Investments in Energy Efficiency**
5. **Market Failures and Business Opportunities**
6. **Old and New Ways to Accelerate Energy Efficiency**

Glossary

conversion efficiency The physical ratio of desired output to total input in an energy conversion process, such as converting fuel to electricity or fuel to heat. Undesired or nonuseful outputs are excluded, making the ratio less than unity for most devices (except in such cases such as heat pumps and heat-powered chillers, where it can exceed unity and is called a coefficient of performance). The definition is thus based partly on what each individual finds useful. Synonymous with the thermodynamic concept of First Law efficiency, but counts only the quantity of energy, not also the quality, and hence differs from Second Law efficiency, which counts both.

customer The ultimate recipient of energy services, regardless of intermediate transactions.

delivered energy Secondary energy provided at the place where it is used to provide the desired energy service (e.g., electricity or fuel entering the end-use device that performs that final conversion). Further losses between that device and the customer may or may not be included. Delivered energy is net of distribution efficiency (1) to the end-use device, but may or may not be net of distribution efficiency (2) between end-use device and customer.

distribution efficiency (1) The fraction of centrally supplied energy shipped out (such as electricity from a power station, petroleum products from a refinery, or natural gas from a treatment plant) that is delivered to the end-use device, net of energy "lost" or "consumed" in its delivery. For electricity, this conversion into unwanted forms, chiefly low-temperature heat, comprises transmission as well as distribution losses, and is conventionally measured from the generator's busbar to the customer's meter. (2) The fraction of energy services produced by the end-use device that reaches the customer (e.g., the fraction of the heat produced by a furnace that provides warmth in a room, net of nonuseful heat escaping from pipes or ducts).

end use (1) The category of desired physical function provided by an energy service, such as heating, cooling, light, mechanical work, electrolysis, or electronic signal processing. (2) The physical quantity of such end use delivered to the customer, whether or not it is useful energy.

end-use device Equipment converting delivered energy into energy service.

end-use efficiency The physical ratio of end use (2) provided to delivered energy converted in the end-use device.

energy conservation An ambiguous term best avoided; considered by some as synonymous with increased energy efficiency but to many others connoting the opposite: privation, curtailment, and discomfort, i.e., getting fewer or lower quality energy services. The degree of confusion between these meanings varies widely by individual, culture, historic period, and language spoken. Some analysts, chiefly outside the United States, embrace energy conservation as an umbrella term for energy efficiency plus changes in personal habits plus changes in system design (such as spatial planning or design for product longevity and materials recovery/reuse).

energy efficiency Broadly, any ratio of function, service, or value provided to the energy converted to provide it. Herein, energy efficiency and its components all use (a) physical rather than economic metrics and (b) engineering, not economic, definitions (this physical convention can, however, become awkward with multiple inputs or outputs). Energy efficiency may or may not count thermodynamic quality of energy (ability to do work); see the distinction between First Law and Second Law efficiency.

energy intensity Energy (primary, delivered, or otherwise defined) "used" per unit of service or value provided.

 © 2004 Rocky Mountain Institute. Published by Elsevier Inc. All rights reserved.

Intensity can be expressed in economic or in physical terms (e.g., the very crude and aggregated metric of primary energy consumption per dollar of real gross domestic product).

energy service The desired function provided by converting energy in an end-use device (e.g., comfort, mobility, fresh air, visibility, electrochemical reaction, or entertainment). These functions, which together contribute to a material standard of living, are generally measured in physical units, not by energy used or money spent. Because diverse services are incommensurable, they cannot be readily added to express a meaningful "total end-use efficiency" for a person, firm, or society. Economic surrogates for such totals are seldom satisfactory.

extractive efficiency The fraction of a fuel deposit that is recovered and sent out for processing and use, net of energy employed to conduct those extractive processes.

hedonic (functional) efficiency How much human happiness or satisfaction (equivalent to economists' welfare metrics) is achieved per unit of energy service delivered. Some analysts similarly distinguish the task (such as delivering heat into a house) from the goal it seeks to achieve (such as the human sensation of indoor comfort in winter).

primary energy (1) Fossil fuel extracted and then converted, typically at a central facility, into a secondary form (e.g., crude oil into refined products or coal into electricity) for delivery to end-use devices. (2) Nuclear, hydroelectric, and renewable energy captured for such delivery; if electric, conventionally expressed as the imputed amount of fossil fuel used to produce that much electricity in a typical thermal power station. (3) The quantity of energy described in (1) or (2). Most analysts exclude from primary energy consumption the metabolic energy in human food, but include the nonsolar energy needed to grow, process, deliver, and prepare it.

secondary energy Any processed, refined, or high-quality form of useful energy converted from primary energy, such as electricity, refined petroleum products, dry natural gas, or district heat. Excludes undesired and nonuseful conversion products.

Second Law efficiency The ratio of First Law thermodynamic efficiency to its maximum theoretically possible value; equivalently, the ratio of the least available work that could have done the job to the actual available work used to do the job. For a device that produces useful work or heat (not both), such as a motor, heat pump, or power plant, Second Law efficiency is the amount of output heat or work usefully transferred, divided by the maximum possible heat or work usefully transferable for the same function by any device or system using the same energy input. This maximum is defined by the task, not the device: to maximize how much heat is delivered from fuel into a building, an ideal fuel cell and an ideal heat pump would be used. Second Law efficiency thus measures the effectiveness of a device in approaching the constraints of the First and Second Laws of thermodynamics. First and Second Law efficiencies are nearly equal for a power plant, but are very different when high-quality energy is converted into a low-energy useful form: a 60%-efficient (First Law) furnace delivering 43°C heat into a house in a 0°C environment has a Second Law efficiency of only 8.2%.

service substitution Providing a desired energy service by a different means (e.g., providing illumination by opening the curtain in the daytime rather than turning on an electric light).

useful energy The portion of an energy service that is actually, not just potentially, desired and used by customers (e.g., lighting an empty room, or overheating an occupied room to the point of discomfort, is seldom useful).

Efficient use of energy is in all countries the most important, economical, prompt, underused, overlooked, and misunderstood way to provide future energy services. It is rapidly becoming even larger, faster, and cheaper as technologies, delivery methods, and integrative design improve. Whole-system design can indeed make very large energy savings cost less than small ones. But capturing energy efficiency's remarkable potential requires careful terminology, prioritization, attention to engineering details and to market failures, and willingness to accept measured physical realities even if they conflict with economic theories. If well done, such energy efficiency can displace costly and disagreeable energy supplies, enhance security and prosperity, speed global development, and protect Earth's climate—not at cost but at a profit.

1. DEFINITION AND IMPORTANCE OF ENERGY EFFICIENCY

Energy efficiency is generally the largest, least expensive, most benign, most quickly deployable, least visible, least understood, and most neglected way to provide energy services. The 39% drop in U.S. energy intensity (primary energy consumption per dollar of real gross domestic product) from 1975 to 2000 represented, by 2000, an effective energy "source" 1.7 times as big as U.S. oil consumption, three times net oil imports, five times domestic oil output, six times net oil imports from Organization of Petroleum Exporting Countries (OPEC) members, and 13 times net imports from Persian Gulf countries. It has lately increased by 3% per year, outpacing the

growth of any source of supply (except minor renewables). Yet energy efficiency has gained little attention or respect for its achievements, let alone its far larger untapped potential. Physical scientists, unlike engineers or economists, find that despite energy efficiency's leading role in providing energy services today, its profitable potential has barely begun to be tapped. In contrast, many engineers tend to be limited by adherence to past practice, and most economists are constrained by their assumption that any profitable savings must already have occurred.

The potential of energy efficiency is increasing faster through innovative designs, technologies, policies, and marketing methods than it is being used up through gradual implementation. The uncaptured "efficiency resource" is becoming bigger and cheaper even faster than oil reserves have lately done through stunning advances in exploration and production. The expansion of the "efficiency resource" is also accelerating, as designers realize that whole-system design integration (see Section 4) can often make very large (one or two order-of-magnitude) energy savings cost less than small or no savings, and as energy-saving technologies evolve discontinuously rather than incrementally. Similarly rapid evolution and enormous potential apply to ways to market and deliver energy-saving technologies and designs; research and development can accelerate both.

1.1 Terminology

Efficiency unfortunately means completely different things to the two professions most engaged in achieving it. To engineers, efficiency means a physical output/input ratio. To economists, efficiency means a monetary output/input ratio, though for practical purposes many economists use a monetary output/physical input ratio. Only physical output/input ratios are used here, but the common use of monetary ratios causes vast confusion, especially to policymakers using economic jargon.

Wringing more work from energy via smarter technologies is often, and sometimes deliberately, confused with a pejorative usage of the ambiguous term energy conservation. Energy efficiency means doing more (and often better) with less—the opposite of simply doing less or worse or without. This confusion unfortunately makes the honorable and traditional concept of energy conservation no longer useful in certain societies, notably the United States, and underlies much of the decades-long neglect or suppression of energy efficiency. However, deliberately reducing the amount or quality of energy services remains a legitimate, though completely separate, option for those who prefer it or are forced by emergency to accept it. For example, the 2000–2001 California electricity crisis ended abruptly when customers, exhorted to curtail their use of electricity, cut their peak load per dollar of weather-adjusted real gross domestic product (GDP) by 14% in the first 6 months of 2001. Most of that dramatic reduction, undoing the previous 5–10 years of demand growth, was temporary and behavioral, but by mid-2002, the permanent and technological fraction was heading for dominance. Even absent crises, some people do not consider an ever-growing volume of energy services to be a worthy end in itself, but seek to live more simply—with elegant frugality rather than involuntary penury—and to meet nonmaterial needs by nonmaterial means. Such choices can save even more energy than can technical improvements alone, though they are often considered beyond the scope of energy efficiency.

Several other terminological distinctions are also important. At least five different kinds of energy efficiency can be measured in at least five different stages of energy conversion chains; these are discussed in Section 1.2. Also, technical improvements in energy efficiency can be broadly classified into those applicable only to new buildings and equipment, those installable in existing ones (retrofitted), those addable during minor or routine maintenance (slipstreamed), and those that can be conveniently added when making major renovations or expansions for other reasons (piggybacked).

Efficiency saves energy whenever an energy service is being delivered, whereas "load management" (sometimes called "demand response" to emphasize reliance on customer choice) changes only the time when that energy is used, either by shifting the timing of the service delivery or by, for example, storing heat or coolth so energy consumption and service delivery can occur at different times. In the context chiefly of electricity, demand-side management, a term coined by the Electric Power Research Institute, comprises both of these options, plus others that may even increase the use of electricity. Most efficiency options yield comparable or greater savings in peak loads; both kinds of savings are valuable, and both kinds of value should be counted. They also have important but nonobvious linkages: for example, because most U.S. peak electric loads are met by extremely inefficient simple-cycle gas-fired combustion turbines, saving 5% of peak electric load in 2000 would have saved 9.5% of total natural gas

consumption, enough to reequilibrate high 2003 gas prices back to ~$2/GJ lower historic levels.

Conflating three different things—technological improvements in energy efficiency (such as thermal insulation), behavioral changes (such as resetting thermostats), and the price or policy tools used to induce or reward those changes—causes endless misunderstanding. Also, consider that the theoretical potential for efficiency gains (up to the maximum permitted by the laws of physics) exceeds the technical potential, which exceeds the economic potential based on social value, which exceeds the economic potential based on private internal value, which exceeds the actual uptake not blocked by market failures, which exceeds what happens spontaneously if no effort is made to accelerate efficiency gains deliberately; yet these six quantities are often not clearly distinguished.

Finally, energy statistics are traditionally organized by the economic sector of apparent consumption, not by the physical end uses provided or services sought. End uses were first seriously analyzed in 1976, rarely appear in official statistics even a quarter-century later, and can be difficult to estimate accurately. But end-use analysis can be valuable because matching energy supplies in quality and scale, as well as in quantity, to end-use needs can save a great deal of energy and money. Supplying energy of superfluous quality, not just quantity, for the task is wasteful and expensive. For example, the United States now provides about twice as much electricity as the fraction of end uses that economically justify this special, costly, high-quality form of energy, yet from 1975 to 2000, 45% of the total growth in primary energy consumption came from increased conversion and grid losses in the expanding, very costly, and heavily subsidized electricity system. Much of the electric growth, in turn, provided low-temperature heat, a physically and economically wasteful use of electricity.

Many subtleties of defining and measuring energy efficiency merit but seldom get rigorous treatment, such as the following losses, services, or metrics:

- Distribution losses downstream of end-use devices (an efficient furnace feeding leaky ducts yields costlier delivered comfort).
- Undesired or useless services, such as leaving equipment on all the time (as many factories do) even when it serves no useful purpose.
- Misused services, such as space-conditioning rooms that are open to the outdoors.
- Conflicting services, such as heating and cooling the same space simultaneously (wasteful even if both services are provided efficiently).
- Parasitic loads, as when the inefficiencies of a central cooling system reappear as additional fed-back cooling loads that make the whole system less efficient than the sum of its parts.
- Misplaced efficiency, such as applying energy-using equipment, however efficiently, to a task that does not need it—say, cooling with a mechanical chiller when groundwater or ambient conditions can more cheaply do the same thing.
- Incorrect metrics, such as measuring lighting by raw quantity (lux or footcandles) unadjusted for its visual effectiveness, which may actually decrease if greater illuminance is improperly delivered.

To forestall a few other semantic quibbles, physicists (including the author) know energy is not "consumed," as the economists' term "consumption" implies, nor "lost," as engineers refer to unwanted conversions into less useful forms. Energy is only converted from one form to another; yet the normal metaphors are clear, common, and adopted here. Thus an 80%-efficient motor converts its electricity input into 80% torque and 20% heat, noise, vibration, and stray electromagnetic fields; the total equals 100% of the electricity input, or roughly 30% of the fuel input at a classical thermal power station. (Note that this definition of efficiency combines engineering metrics with human preference. The motor's efficiency may change, with no change in the motor, if changing intention alters which of the outputs are desired and which are unwanted: the definition of "waste" is as much social or contextual as physical. A floodlamp used to keep plates of food warm in a restaurant may be rather effective for that purpose even though it is an inefficient source of visible light).

More productive use of energy is not, strictly speaking, a physical "source" of energy but rather a way to displace physical sources. Yet this distinction is rhetorical, because the displacement or substitution is real and makes supply fully fungible with efficiency. Also, energy/GDP ratios are a very rough, aggregated, and sometimes misleading metric, because they combine changes in technical efficiency, human behavior, and the composition of GDP (a metric that problematically conflates goods and services with "bads" and nuisances, counts only monetized activities, and is an increasingly perverse measure of well being). Yet the two-fifths drop in U.S. energy intensity and the one-half drop in oil

intensity during the period 1975–2001 reflect mainly better technical efficiency. Joseph Romm has also shown that an important compositional shift of U.S. GDP—the information economy emerging in the late 1990s—has significantly decreased energy and probably electrical energy intensity, as bytes substituted for (or increased the capacity utilization of) travel, freight transport, lit and conditioned floorspace, paper, and other energy-intensive goods and services.

The aim here is not to get mired in word games, but to offer a clear overview of what kinds of energy efficiency are available, what they can do, and how best to consider and adopt them.

1.2 Efficiency along Energy Conversion Chains

The technical efficiency of using energy is the product of five efficiencies successively applied along the chain of energy conversions: (1) the conversion efficiency of primary into secondary energy, times (2) the distribution efficiency of delivering that secondary energy from the point of conversion to the point of end use, times (3) the end-use efficiency of converting the delivered secondary energy into such desired energy services as hot showers and cold beer. Some analysts add another term at the upstream end, (4) the extractive efficiency of converting fuel in the ground or power from wind or from sun in the atmosphere, etc. into the primary energy fed into the initial conversion device, and another term at the downstream end, (5) the hedonic efficiency of converting delivered energy services into human welfare. (Delivering junk mail with high technical efficiency is futile if the recipients did not want it.) Counting all five efficiencies permits comparing ultimate means, the primary energy tapped, with ultimate ends, the happiness or economic welfare created. Focusing only on intermediate means and ends loses sight of what human purposes an energy system is to serve. Most societies pay attention to only three kinds of energy efficiency: extraction (because of its cost, not because the extracted fuels are assigned any intrinsic or depletion value), conversion, and perhaps distribution. End-use and hedonic efficiency are left to customers, are least exploited, and hence hold the biggest potential gains.

They also offer the greatest potential leverage. Because successive efficiencies along the conversion chain all multiply, they are often assumed to be equally important. Yet downstream savings—those nearest the customer—are the most important. Figure 1 shows schematically the successive energy conversions and losses that require about 10 units of fuel to be fed into a conventional thermal power station in order to deliver one unit of flow in a pipe. But conversely, every unit of flow (or friction) saved in the pipe will save approximately 10 units of fuel, cost, and pollution at the power station. It will also make the pump's motor (for example) nearly two and a half units smaller, hence cheaper. To save the most primary energy and the most capital cost, therefore, efficiency efforts should start all the way downstream (see Section 4.2), by asking: How little flow can actually deliver the desired service? How small can the piping friction become? How small, well matched to the flow regime, and efficient can the pump be made? Its coupling? Its motor? Its controls and electrical supplies?

Analyses of energy use should start with the desired services or changes in well being, then work back upstream to primary supplies. This maximizes the extra value of downstream efficiency gains and

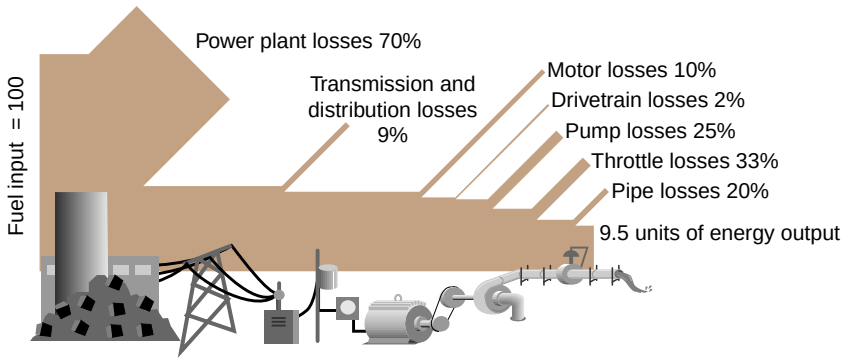


FIGURE 1 To deliver one unit of flow in the pipe requires about 10 units of fuel at the power plant, thus those 10-fold compounding losses can be turned around backward, yielding 10-fold compounding savings of fuel for each unit of reduced friction or flow in the pipe. From the E SOURCE "Drivepower Technology Atlas," courtesy of Platts Research & Consulting.

the capital-cost savings from smaller, simpler, cheaper upstream equipment. Yet it is rarely done. Similarly, most energy policy analysts analyze how much energy could be supplied before asking how much is optimally needed and at what quality and scale it could be optimally provided. This wrong direction (upstream to downstream) and supply orientation lie at the root of many if not most energy policy problems.

Even modest improvements in efficiency at each step of the conversion chain can multiply to large collective values. For example, suppose that during the years 2000–2050, world population and economic growth increased economic activity by six- to eightfold, in line with conventional projections. But meanwhile, the carbon intensity of primary fuel, following a two-century trend, is likely to fall by at least two- to fourfold as coal gives way to gas, renewables, and carbon offsets or sequestration. Conversion efficiency is likely to increase by at least 1.5-fold with modernized, better-run, combined-cycle, and cogenerating power stations. Distribution efficiency should improve modestly. End-use efficiency could improve by four- to sixfold if the intensity reductions sustained by many industrial countries, when they were paying attention, were sustained for 50 years (e.g., the United States decreased its primary energy/GDP intensity at an average rate of 3.4%/year from 1979 to 1986 and 3.0%/year from 1996 to 2001). And the least understood term, hedonic efficiency, might remain constant or might perhaps double as better business models and customer choice systematically improve the quality of services delivered and their match to what customers want. On these plausible assumptions, global carbon emissions from burning fossil fuel could decrease by 1.5- to 12-fold despite the assumed six- to eightfold grosser world product. The most important assumption is sustained success with end-use efficiency, but the decarbonization and conversion-efficiency terms appear able to take up some slack if needed.

1.3 Service Redefinition

Some major opportunities to save energy redefine the service being provided. This is often a cultural variable. A Japanese person, asked why the house is not heated in winter, might reply, "Why should I? Is the house cold?" In Japanese culture, the traditional goal is to keep the person comfortable, not to heat empty space. Thus a modern Japanese room air conditioner may contain a sensor array and moveable fans that detect and blow air toward people's locations in the room, rather than wastefully cooling the entire space. Western office workers, too, can save energy (and can often see better, feel less tired, and improve esthetics) by properly adjusting venetian blinds, bouncing glare-free daylight up onto the ceiling, and turning off the lights. As Jørgen Nørgård remarks, "energy-efficient lamps save the most energy when they are turned off"; yet many people automatically turn on every light when entering a room. This example also illustrates that energy efficiency may be hard to distinguish from energy supply that comes from natural energy flows. All houses are already $\sim$98% solar-heated, because if there were no Sun (which provides 99.8% of Earth's heat), the temperature of Earth's surface would average approximately -272.6°C rather than $+15^{\circ}$C. Thus, strictly speaking, engineered heating systems provide only the last 1–2% of the total heating required.

Service redefinition becomes complex in personal transport. Its efficiency is not just about vehicular fuel economy, people per car, or public transport alternatives. Rather, the underlying service should often be defined as access, not mobility. Typically, the best way to gain access to a place is to be there already; this is the realm of land-use management—no novelty in the United States, where spatial planning is officially shunned, yet zoning laws mandate dispersion of location and function, real-estate practices segregate housing by income, and other market distortions maximize unneeded and often unwanted travel. Another way to gain access is virtually, moving just the electrons while leaving the heavy nuclei behind, via telecommunications, soon including realistic "virtual presence." This is sometimes an effective alternative to physically moving human bodies. And if such movement is really necessary, then it merits real competition, at honest prices, between all modes—personal or collective, motorized or human-powered, conventional or innovative. Creative policy tools can enhance that choice in ways that enhance real-estate value, saved time, quality of life, and public amenity and security. Efficient cars can be an important part of efficient personal mobility, but also reducing the need to drive can save even more energy and yield greater total benefit.

1.4 Historic Summaries of Potential

People have been saving energy for centuries, even millennia; this is the essence of engineering. Most savings were initially in conversion and end use:

preindustrial households often used more primary energy than modern ones do, because fuelwood-to-charcoal conversion, inefficient open fires, and crude stoves burned much fuel to deliver sparse cooking and warmth. Lighting, materials processing, and transport end uses were also very inefficient. Billions of human beings still suffer such conditions today. The primary energy/GDP intensities in developing countries average about three times those in industrialized countries. But even the most energy-efficient societies still have enormous, and steadily expanding, room for further efficiency gains. Less than one-fourth of the energy delivered to a typical European cookstove ends up in food, less than 1% of the fuel delivered to a standard car actually moves the driver, U.S. power plants discard waste heat equivalent to 1.2 times Japan's total energy use, and even Japan's economy does not approach one-tenth the efficiency that the laws of physics permit.

Detailed and exhaustively documented engineering analyses of the scope for improving energy efficiency, especially in end-use devices, have been published for many industrial and some developing countries. By the early 1980s, those analyses had compellingly shown that most of the energy currently used was being wasted—i.e., that the same or better services could be provided using severalfold less primary energy by fully installing, wherever practical and profitable, the most efficient conversion and end-use technologies then available. Such impressive efficiency gains cost considerably less than the long-run, and often even the short-run, marginal private internal cost of supplying more energy. Most policymakers ignore both these analyses, well known to specialists, and less well-known findings show even bigger and cheaper savings from whole-system design integration (see Section 4).

Many published engineering analyses show a smaller saving potential because of major conservatisms, often deliberate (because the real figures seem too good to be true), or because they assume only partial adoption over a short period rather than examining the ultimate potential for complete practical adoption. For example, the American Council for an Energy-Efficient Economy estimates that just reasonable adoption of the top five conventional U.S. opportunities—industrial improvements, 40-mile per gallon (U.S. gallons; = 4.88 liters/100 km) light-vehicle standards, cogeneration, better building codes, and a 30% better central-air-conditioning standard—could save 530 million T/year of oil equivalent—respectively equivalent to the total 2000 primary energy use of Australia, Mexico, Spain, Austria, and Ireland. But the full long-term efficiency potential is far larger, and much of it resides in innumerable small terms. Saving energy is like eating an Atlantic lobster: there are big, obvious chunks of meat in the tail and the front claws, but a similar total quantity of tasty morsels is hidden in crevices and requires some skill and persistence to extract.

The whole-lobster potential is best, though still not fully, seen in bottom-up technological analyses comparing the quantity of potential energy savings with their marginal cost. That cost is typically calculated using the Lawrence Berkeley National Laboratory methodology, which divides the marginal cost of buying, installing, and maintaining the more efficient device by its discounted stream of lifetime energy savings. The levelized cost in dollars of saving, say, 1 kWh, then equals $Ci/S[1-(1+i)^{-n}]$, where C is installed capital cost (dollars), i is annual real discount rate (assumed here to be 0.05), S is energy saved by the device (kilowatt-hours/year), and n is operating life (years). Thus a \$10 device that saved 100 kWh/year and lasted 20 years would have a levelized "cost of saved energy" (CSE) of 0.8¢/kWh. Against a 5¢/kWh electricity price, a 20-year device with a 1-year simple payback would have CSE = 0.4¢/kWh. It is conventional for engineering-oriented analysts to represent efficiency "resources" as a supply curve, rather than as shifts along a demand curve (the convention among economists). CSE is methodologically equivalent to the cost of supplied energy (e.g., from a power station and grid): the price of the energy saved is not part of the calculation. Whether the saving is cost-effective depends on comparing the cost of achieving it with the avoided cost of the energy saved. (As noted in Section 2, this conventional engineering–economic approach actually understates the benefits of energy efficiency.)

On this basis, the author's analyses in the late 1980s found, from measured cost and performance data for more than 1000 electricity-saving end-use technologies, that their full practical retrofit could save about three-fourths of U.S. electricity at an average CSE $\sim$ 0.6¢/kWh (1986 dollars)—roughly consistent with a 1990 Electric Power Research Institute analysis in which the differences were mainly methodological rather than substantive. Similarly, the author's analysis for Royal Dutch/Shell Group found that full use of the best 1987–1988 oil-saving end-use technologies, assuming turnover of vehicle stocks, could save about 80% of U.S. oil use at an average levelized CSE below \$2.5/barrel (1986 dollars). Both analyses have proven systematically

conservative: today's potential is even larger and cheaper. (The analyses explicitly excluded financing and transaction costs, but those would only slightly affect the results. There is also a huge literature accurately predicting and rigorously measuring the empirical size, speed, and cost of efficiency improvements delivered by actual utility and government programs.) Such findings are broadly consistent with equally or more detailed ones by European analysts: for example, late-1980s technologies could save three-fourths of Danish buildings' electricity or half of all Swedish electricity at $0.016/kWh (1986 dollars), or four-fifths of German home electricity (including minor fuel switching) with a ~40%/year aftertax return on investment. Such findings with ever greater sophistication have been published worldwide since 1979, but have been rejected by nontechnological economic theorists, who argue that if such cost-effective opportunities existed, they would already have been captured in the marketplace, even in planned economies with no marketplace or mixed economies with a distorted one. This mental model—"don't bother to bend over and pick up that banknote lying on the ground, because if it were real, someone would have picked it up already"—often dominates government policy. It seems ever less defensible as more is learned about the reality of pervasive market failures (see Section 5) and the astonishing size and cheapness of the energy savings empirically achieved by diverse enterprises (discussed in Section 3). But by now, the debate is theological—about whether existing markets are essentially perfect, as most economic modelers assume for comfort and convenience, or whether market failures are at least as important as market function and lie at the heart of business and policy opportunity. To technologists and physical scientists, this seems a testable empirical question.

1.5 Discontinuous Technological Progress

This engineering/economics divergence about the potential to save energy also reflects a tacit assumption that technological evolution is smooth and incremental, as mathematical modelers prefer. In fact, although much progress is as incremental as technology diffusion, discontinuous technological leaps, more like "punctuated equilibrium" in evolutionary biology, can propel innovation and speed its adoption, as with 5×-efficiency light vehicles (see Section 4.1).

Technological discontinuities can even burst the conventional boundaries of possibility by redefining the design space. Generations of engineers learned that big supercritical-steam power plants were as efficient as possible (~40% from fuel in to electricity out). But through sloppy learning or teaching, these engineers overlooked the possibility of stacking two Carnot cycles atop each other. Such combined-cycle (gas-then-steam) turbines, based on mass-produced jet engines, can exceed 60% efficiency and are cheaper and faster to build, so in the 1990s, they quickly displaced the big steam plants. Fuel cells, the next innovation, avoid Carnot limits altogether by being an electrochemical device, not a heat engine. Combining both may soon achieve 80–90% fuel-to-electric efficiency. Even inefficient distributed generators can already exceed 90% system efficiency by artfully using recaptured heat.

As another example, most authors today state that the theoretical efficiency limit for converting sunlight into electricity using single-layer photovoltaic (PV) cells is 31% (~50% using multicolor stacked layers; the best practical values so far are around 25 and 30%). This is because semiconductor bandgaps have been believed too big to capture any but the high-energy wavelengths of sunlight. But those standard data are wrong. A Russian-based team suspected in 2001, and Lawrence Berkeley National Laboratory proved in 2002, that indium nitride's bandgap is only 0.7 eV, matching near-infrared (1.77 μm) light and hence able to harvest almost the whole solar spectrum. This may raise the theoretical limit to 50% for two-layer and to ~70% for many-layer thin-film PVs.

Caution is likewise vital when interpreting Second Law efficiency. In the macroscopic world, the laws of thermodynamics are normally considered ineluctable, but the definition of the desired change of state can be finessed. Ernie Robertson notes that when turning limestone into a structural material, one is not confined to just the conventional possibilities of cutting it into blocks or calcining it at ~1250°C into Portland cement. It is possible instead grind it up and feed it to chickens, in which ambient-temperature "technology" turns it into eggshell stronger than Portland cement. Were we as smart as chickens, we would have mastered this life-friendly technology. Extraordinary new opportunities to harness 3.8 billion years of biological design experience, as described by Janine Benyus in *Biomimicry*, can often make heat-beat-and-treat industrial processes unnecessary. So, in principle, can the emerging techniques of nanotechnology using molecular-scale self-assembly, as pioneered by Eric Drexler.

More conventional innovations can also bypass energy-intensive industrial processes. Making artifacts that last longer, use materials more frugally, and are designed and deployed to be repaired, reused, remanufactured, and recycled can save much or most of the energy traditionally needed to produce and assemble their materials (and can increase welfare while reducing GDP, which swells when ephemeral goods are quickly discarded and replaced). Microfluidics can even reduce a large chemical plant to the size of a watermelon: millimeter-scale flow in channels etched into silicon wafers can control time, temperature, pressure, stoichiometry, and catalysis so exactly that a very narrow product spectrum is produced, without the side-reactions that normally require most of the chemical plant to separate undesired from desired products. Such "end-run" solutions (rather like the previous example of substituting sensible land-use for better vehicles, or better still, combining both) can greatly expand the range of possibilities beyond simply improving the narrowly defined efficiency of industrial equipment, processes, and controls. By combining many such options, it is now realistic to contemplate a long-run advanced industrial society that provides unprecedented levels of material prosperity with far less energy, cost, and impact than today's best practice. The discussion in Section 4.1, drawn from Paul Hawken *et al.*'s synthesis in *Natural Capitalism* and Ernst von Weizsäcker *et al.*'s earlier *Factor Four*, further illustrates recent breakthroughs in integrative design that can make very large energy savings cost less than small ones; and similarly important discontinuities in policy innovation are summarized in Section 6.

In light of all these possibilities, why does energy efficiency, in most countries and at most times, command so little attention and such lackadaisical pursuit? Several explanations come to mind. Saved energy is invisible. Energy-saving technologies may look and outwardly act exactly like inefficient ones, so they are invisible too. They are also highly dispersed, unlike central supply technologies that are among the most impressive human creations, inspiring pride and attracting ribbon-cutters and rent- and bribe-seekers. Many users believe energy efficiency is binary—you either have it or lack it—and that they already did it in the 1970s. Energy efficiency has relatively weak and scattered constituencies, and major energy efficiency opportunities are disdained or disbelieved by policymakers indoctrinated in a theoretical economic paradigm that claims they cannot exist (see Section 3).

2. BENEFITS OF ENERGY EFFICIENCY

Energy efficiency avoids the direct economic costs and the direct environmental, security, and other costs of the energy supply and delivery that it displaces. Yet the literature often overlooks several key side-benefits (economists call them "joint products") of saving energy.

2.1 Indirect Benefits from Qualitatively Superior Services

Improved energy efficiency, especially end-use efficiency, often delivers better services. Efficient houses are more comfortable; efficient lighting systems can look better and help you see better; efficient motors can be more quiet, reliable, and controllable; efficient refrigerators can keep food fresher for longer; efficient cleanrooms can improve the yield, flexibility, throughput, and setup time of microchip fabrication plants; efficient fume hoods can improve safety; efficient supermarkets can improve food safety and merchandising; retail sales pressure can rise 40% in well-daylit stores; and students' test scores suggest 20–26% faster learning in well-daylit schools. Such side-benefits can be one or even two more orders of magnitude more valuable than the energy directly saved. For example, careful measurements show that in efficient buildings, where workers can see what they are doing, hear themselves think, breathe cleaner air, and feel more comfortable, labor productivity typically rises by about 6–16%. Because office workers in industrialized countries cost about 100 times more than office energy, a 1% increase in labor productivity has the same bottom-line effect as eliminating the energy bill, and the actual gain in labor productivity is about 6–16 times bigger than that. Practitioners can market these attributes without ever mentioning lower energy bills.

2.2 Leverage in Global Fuel Markets

Much has been written about the increasing pricing power of major oil-exporting countries, especially Saudi Arabia, with its important swing production capacity. Yet the market power of the United States—the Saudi Arabia of energy waste—is even greater on the demand side. The United States can raise its oil productivity more and faster than any oil exporter can adjust to reducing its oil output. This was illustrated

from 1977 to 1985, when U.S. GDP rose 27% while total U.S. oil imports fell by 42%, or 3.74 million barrels (bbl)/day. This was largely due to a 52% gain in oil productivity, causing U.S. oil consumption to fall by 17% and oil imports from the Persian Gulf to fall by 91%. That took away an eighth of OPEC's market. The entire world oil market shrank by a tenth; OPEC's share was slashed from 52 to 30%, and its output fell by 48%. The United States accounted for one-fourth of that reduction. More efficient cars, each driving 1% fewer miles on 20% less gasoline—a 7-mile per (U.S.) gallon gain in 6 years for new American-made cars—were the most important single cause; 96% of those savings came from smarter design, with only 4% deriving from smaller size.

Those 8 years around the second oil shock (1979) demonstrated an effective new source of energy security and a potent weapon against high oil prices and supply manipulations. The United States showed that a major consuming nation could respond effectively to supply disruptions by focusing on the demand side and boosting oil productivity at will. It could thereby exercise more market power than suppliers, beat down their prices (breaking OPEC's pricing power for a decade), and enhance the relative importance of less vulnerable, more diversified sources of supply. Had the United States simply continued its 1979–1985 rate of improvement of oil productivity, it would not have needed a drop of oil from the Persian Gulf after 1985. That is not what happened, but it could be if the United States chose to repeat and expand its previous oil savings.

2.3 Buying Time

Energy efficiency buys time. Time is the most precious asset in energy policy, because it permits the fullest and most graceful development and deployment of still better techniques for energy efficiency and supply. This pushes supply curves downward (larger quantities at lower prices), postpones economic depletion, and buys even more time. The more time is available, the more information will emerge to support wiser and more robust choices, and the more fruitfully new technologies and policy options can meld and breed new ones. Conversely, hasty choices driven by supply exigencies almost always turn out badly, waste resources, and foreclose important options. Of course, once bought, time should be used wisely. Instead, the decade of respite bought by the U.S. efficiency spurt of 1979–1985 was almost entirely wasted as attention waned, efficiency and alternative-supply efforts stalled, research and development teams were disbanded, and underlying political problems festered. From the perspective of 2004, that decade of stagnation looked like a blunder of historic proportions.

2.4 Integrating Efficiency with Supply

To the first order, energy efficiency makes supply cheaper. But second-order effects reinforce this first-order benefit, most obviously when efficiency is combined with onsite renewable supplies, making them nonlinearly smaller, simpler, and cheaper. Consider the following examples:

- A hot-water-saving house can achieve a very high solar-water-heat fraction (e.g., 99% in the author's home high in the Rocky Mountains) with only a small collector, so it needs little or no backup, partly because collector efficiency increases as stratified-tank storage temperature decreases.
- An electricity-saving house needs only a few square meters of PVs and a simple balance-of-system setup (storage, inverter, etc.). This can cost less than connecting to the grid a few meters away.
- A passive-solar, daylit building needs little electricity, and can pay for even costly forms of on-site generation (such as PVs) using money saved by eliminating or downsizing mechanical systems.
- Such mutually reinforcing options can be bundled: e.g., 1.18 peak MW of photovoltaics retrofitted onto the Santa Rita Jail in Alameda County, California, was combined with efficiency and load management, so at peak periods, when the power was most valuable, less was used by the jail and more was sold back to the grid. This bundling yielded a internal rate of return of over 10% including state subsidies, and a customer present-valued benefit/cost ratio of 1.7 without or 3.8 with those subsidies.

2.5 Gaps in Engineering Economics

Both engineers and economists conventionally calculate the cost of supplying or saving energy using a rough-and-ready tool kit called "engineering economics." Its methods are easy to use but flawed, ignoring such basic tenets of financial economics as risk-adjusted discount rates. Indeed, engineering economics omits 207 economic and engineering considerations that together increase the value of decentralized electrical resources by typically an order of magnitude. Many of these "distributed benefits," compiled in the author's *Small Is Profitable*, apply as much to decentralized efficiency as to

generation. Most of the literature on the cost of energy alternatives is based solely on accounting costs and engineering economics that greatly understate efficiency's value. Properly counting its benefits will yield far sounder investments.

Efficient end use is also the most effective way to make energy supply systems more resilient against mishap or malice, because it increases the duration of buffer stocks, buying time to mend damage or arrange new supplies, and it increases the share of service that curtailed or improvised supplies can deliver. Efficiency's high "bounce per buck" makes it the cornerstone of any energy system designed for secure service provision in a dangerous world.

3. ENGINEERING VS. ECONOMIC PERSPECTIVES

Engineering practitioners and economic theorists view energy efficiency through profoundly different lenses, yet both disciplines are hard pressed to explain the following phenomena:

1. During the period 1996–2001, U.S. aggregate energy intensity fell at a near-record pace despite record low and falling energy prices. (It fell faster only once in modern history, during the record high and rising energy prices of 1979–1985.) Apparently, something other than price was getting Americans' attention.
2. During the period 1990–1996, when a kilowatt-hour of electricity cost only half as much in Seattle as in Chicago, people in Seattle, on average, reduced their peak electric load 12 times as fast, and their use of electricity about 3640 times as fast, as did people in Chicago—perhaps because Seattle City Light encouraged savings while Commonwealth Edison Co. discouraged them.
3. In the 1990s, DuPont found that its European chemical plants were no more energy efficient than its corresponding U.S. plants, despite long having paid twice the energy price—perhaps because all the plants were designed by the same people in the same ways with the same equipment; there is little room for behavioral change in a chemical plant.
4. In Dow Chemical Company's Louisiana Division during the period 1981–1993, nearly 1000 projects to save energy and reduce waste added $110 million/year to the bottom line and
yielded returns on investment averaging over 200%/year, yet in the latter years, both the returns and the savings were trending upward as the engineers discovered new tricks faster than they used up the old ones. (Economic theory would deny the possibility of so much "low-hanging fruit" that has fallen down and is mushing up around the ankles: such enormous returns, if real, would long ago have been captured. This belief was the main obstacle to engineers' seeking such savings, then persisting after their discovery.)
5. Only about 25–35% of apartment dwellers, when told that their air conditioner and electricity are free, behave as economists would predict—turning on the air conditioner when they feel hot and setting the thermostat at a temperature at which they feel comfortable. The rest of the apartment dwellers show no correlation between air-conditioning usage and comfort; instead, their cooling behavior is determined by at least six other variables: household schedules, folk theories about air conditioners (such as that the thermostat is a valve that makes the cold come out faster), general strategies for dealing with machines, complex belief systems about health and physiology, noise aversion, and wanting white noise to mask outside sounds that might wake the baby. Energy anthropology reveals that both the economic and the engineering models of air-conditioning behavior are not just incomplete but seriously misleading.
6. The United States has misallocated $1 trillion of investments to about 200 million refrigerative tons of air conditioners, and 200 peak GW (two-fifths of total peak load) of power supply to run them, that would not have been bought if the buildings had been optimally designed to produce best comfort at least cost. This seems explicable by the perfectly perverse incentives seen by each of the 20-odd actors in the commercial real-estate value chain, each systematically rewarded for inefficiency and penalized for efficiency.
7. Not just individuals but also most firms, even large and sophisticated ones, routinely fail to make essentially riskless efficiency investments yielding many times their normal business returns: most require energy efficiency investments to yield roughly six times their marginal cost of capital, which typically applies to far riskier investments.

Many economists would posit some unknown error or omission in these descriptions, not in their theories. Indeed, energy engineers and energy economists seem not to agree about what is a hypothesis and what is a fact. Engineers take their facts from tools of physical observation. Three decades of conversations with noted energy economists suggest

to the author that most think facts come only from observed flows of dollars, interpreted through indisputable theoretical constructs, and consider any contrary physical observations aberrant. This divergence makes most energy economists suppose that buying energy efficiency faster than the "spontaneous" rate of observed intensity reduction (for 1997–2001, 2.7%/year in the United States, 1.4%/year in the European Union, 1.3%/year worldwide, and 5.3%/year in China) would require considerably higher energy prices, because if greater savings were profitable at today's prices, they would already have been bought; thus the engineers' bottom-up analyses of potential energy savings must be unrealistically high. Economists' estimates of potential savings at current prices are "top-down" and very small, based on historic price elasticities that confine potential interventions to changing prices and savings to modest size and diminishing returns (otherwise the economists' simulation models would inconveniently explode). Engineers retort that high energy prices are not necessary for very large energy savings (because they are so lucrative even at present prices) but are not sufficient either (because higher prices do little without enlarged ability to respond to them).

Of course, engineering-based practitioners agree that human behavior is influenced by price, as well as by convenience, familiarity, fashion, transparency, competing claims on attention, and many other marketing and social-science factors—missing from any purely technological perspective but central to day-to-day fieldwork. The main difference is that they think these obstacles are "market failures" and dominate behavior in buying energy efficiency. Most economists deny this, and say the relatively slow adoption of efficiency must be due to gross errors in the engineers' claims of how large, cheap, and available its potential really is. This theological deadlock underlies the debate about climate protection. Robert Repetto and Duncan Austin showed in 1997 that all mainstream climate-economics models' outputs are hard-wired to the input assumptions, and that realistic inputs, conforming to the actual content of the Kyoto Protocol and its rules, show that climate protection increases GDP. Florentin Krause has shown that the main official U.S. government analyses, taken together, concur. Yet the official U.S. position at the end of 2003 was still that climate protection, even if desirable, cannot be mandated because it is too costly.

In fact, climate protection is not costly but profitable; its critics may have the amount about right, but they got the sign wrong. The clearest proof is in the behavior and achievements of the smart companies that are behaving as if the United States had ratified the Kyoto Protocol, because energy efficiency is cheaper than the energy it saves. For example, large firms such as DuPont, IBM, and STMicroelectronics (ST; one of the world's largest chipmakers) have lately been raising their energy productivity by 6%/year with simple paybacks of a few years. DuPont expects by 2010 to cut its greenhouse gas emissions by 65% below the 1990 level; ST expects to achieve zero emissions (despite making 40 times more chips). British Petroleum announced that its 10% reduction by 2010 had been achieved 8 years early at zero net cost; actually, the 10-year net-present-valued saving was $650 million. Other examples abound; the Web sites www.cool-companies.org and www.pewclimate.org contain examples of the achievements of actively engaged businesses. The companies involved, many of them well known in the business world, are hardly naïve or deluded. Anyone ignoring this market reality is mistaking the econometric rearview mirror for a windshield. Econometrics measures how human populations behaved under past conditions that no longer exist and that it is often a goal of energy policy to change. Where price is the only important explanatory variable, econometrics can be a useful tool for extrapolating history into how price may influence near-term, small, incremental changes in behavior. But limiting our horizons to this cramped view of technical possibilities and human complexity rules out innovations in policies, institutions, preferences, and technologies—treating the future like fate, not choice, and thus making it so.

4. DIMINISHING VS. EXPANDING RETURNS TO INVESTMENTS IN ENERGY EFFICIENCY

Among the most basic, yet simply resolved, economic/engineering disagreements is whether investing in end-use efficiency yields expanding or diminishing returns. Economic theory says diminishing: the more efficiency we buy, the more steeply the marginal cost of the next increment of savings rises, until it becomes too expensive (Fig. 2). But engineering practice often says expanding: big savings can cost less than small or no savings (Fig. 3) if the engineering is done unconventionally but properly.

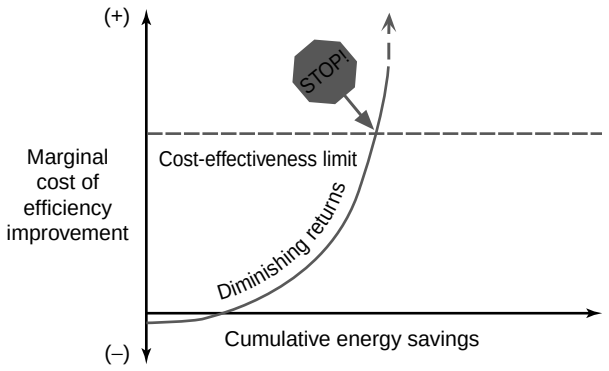


FIGURE 2 Diminishing returns (greater energy savings incur greater marginal cost) can be true for some (not all) components, but need not be true for most systems.

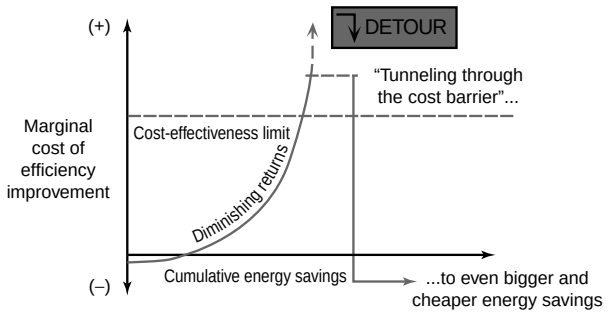


FIGURE 3 Optimizing whole systems for multiple benefits, rather than isolated components for single benefits, can often "tunnel through the cost barrier" directly to the lower-right-corner destination, making very large energy savings cost less than small or no savings. This has been empirically demonstrated in a wide range of technical systems.

4.1 Empirical Examples

Consider, for example, how much thermal insulation should surround a house in a cold climate. Conventional design specifies just the amount of insulation that will repay its marginal cost out of the present value of the saved marginal energy. But this is methodologically wrong, because the comparison omits the capital cost of the heating system: furnace, ducts, fans, pipes, pumps, wires, controls, and fuel source. The author's house illustrates that in outdoor temperatures down to –44°C, it is feasible to grow bananas (28 crops at this writing) at 2200 m elevation in the Rocky Mountains with no heating system, yet with reduced construction cost, because the superwindows, superinsulation, air-to-air heat exchangers, and other investments needed to eliminate the heating system cost less to install than the heating system would have cost to install. The resulting ~99% reduction in heating energy cost is an extra benefit.

Similarly, Pacific Gas and Electric Company's Advanced Customer Technology Test for Maximum Energy Efficiency (ACT^2) demonstrated in seven new and old buildings that the "supply curve" of energy efficiency generally bent downward, as shown schematically in Fig. 3. For example, an ordinary-looking new tract house was designed to save 82% of the energy allowed by the strictest U.S. standard of the time (1992 California Title 24); if this design were widely implemented rather than restricted to a single experiment, it was estimated by PG&E that it would cost about $1800 less than normal to build and $1600 less (in present value) to maintain. It provided comfort with no cooling system in a climate that can reach 45°C; a similar house later did the same in a 46°C-peak climate. Another example, the 350-m^2 Bangkok home of architecture professor Suntoorn Boonyatikarn, provided normal comfort, with 10% the normal air-conditioning capacity, at no extra construction cost. These examples illustrate how optimizing a house as a system rather than optimizing a component in isolation, and optimizing for life-cycle cost (capital plus operating cost, and preferably also maintenance cost), can make a superefficient house cheaper to build, not just to run, by eliminating costly heating and cooling systems. Similarly, a retrofit design for a 19,000-m^2 curtainwall office building near Chicago found 75% energy-saving potential at no more cost than the normal 20-year renovation that saves nothing, because the $200,000 capital saving from making the cooling system four times smaller (yet four times more efficient), rather than renovating the big old system, would pay for the other improvements.

In a striking industrial example, a pumping loop for heat transfer originally designed to use 70.8 kW of pumping power was redesigned to use 5.3 kW, 92% less, with lower capital cost and better performance. No new technologies were used, but only two changes in the design mentality:

1. Use big pipes and small pumps rather than small pipes and big pumps. The friction in a pipe falls as nearly the fifth power (roughly −4.84) of its diameter. Engineers normally make the pipe just fat enough to repay its greater cost from the saved pumping energy. This calculation improperly omits the capital cost of the pumping equipment—the pump, motor, inverter, and electricals that must overcome the pipe friction. Yet the size and roughly the cost of that equipment will fall as nearly the fifth power of pipe diameter, while the cost of the fatter pipe will rise as only about the second power of

diameter. Thus conventionally optimizing the pipe as an isolated component actually pessimizes the system! Optimizing the whole system together will clearly yield fat pipes and tiny pumps, so total capital cost falls slightly and operating cost falls dramatically.

2. Lay out the pipes first, then the equipment. Normal practice is the opposite, so the connected equipment is typically far apart, obstructed by other objects, at the wrong height, and facing the wrong way. The resulting pipes have about three to six times as much friction as they would have with a straight shot, to the delight of the pipefitters, who are paid by the hour, who mark up a profit on the extra pipes and fittings, and who do not pay for the extra electricity or equipment sizing. But the owner would do better to have short, straight pipes than long, crooked pipes.

Together, these two design changes cut the measured pumping power by 12-fold, with lower capital cost and better performance. They also saved 70 kW of heat loss with a 2-month payback, because it was easier to insulate short, straight pipes. In hindsight, however, the design was still suboptimal, because it omitted seven additional benefits: less space, weight, and noise; better maintenance access; lower maintenance cost; higher reliability and uptime; and longer life (because the removed pipe elbows will not be eroded by fluid turning the corner). Properly counting these seven benefits would have saved not 92% but nearer 98% of the energy, and cost even less, so about a factor-four potential saving was left uncaptured.

Other recent design examples include a 97% reduction in air-conditioning energy for a California office retrofit, with attractive returns and better comfort; lighting retrofit savings upward of 90% with better visibility and a 1- to 2-year payback; an energy cost reduction $>40\%$ with a 3-year payback in retrofitting an already very efficient oil refinery; $\sim 75\%$ electrical savings in a new chemical plant, with $\sim 10\%$ lower construction time and cost; $\sim 89\%$ in a new data center at lower cost; and ~ 70–90% in a new supermarket at probably lower cost. The obvious lesson is that optimizing whole systems for multiple benefits, not just components for single benefits, typically boosts end-use efficiency by roughly an order of magnitude at negative marginal cost. These enormous savings have not been widely noticed or captured because of deficient engineering pedagogy and practice. Whole-system design integration is not rocket science; rather, it rediscovers the forgotten tradition of Victorian system engineering, before designers became so specialized that they lost sight of how components fit together.

It is not even generally true, as economic theory supposes, that greater end-use efficiency costs more at the level of components. For example, the most common type of industrial motor on the 1996 U.S. market, the 1800-revolution per minute (rpm) totally enclosed fan-cooled (TEFC) motor (National Electrical Manufacturers' Association Design B), exhibited no empirical correlation whatever between efficiency and price up to at least 225 kW. (Premium-efficiency motors should cost more to build because they contain more and better copper and iron, but they are not priced that way.) The same is true for most industrial pumps and rooftop chillers, Swedish refrigerators, American televisions, and many other products. But even if it were true, artfully combining components into systems can definitely yield expanding returns.

Perhaps the most consequential example is in light vehicles. A small private company (see its Web site at www.hypercar.com) completed in 2000 the manufacturable, production-costed virtual design of a midsize sport utility vehicle (SUV) concept car that is uncompromised, cost-competitive, zero emission, and quintupled efficiency. It is so efficient (2.38 liters/100 km, 42 km/liter, 99 mpg, U.S. gallons) not just because its direct-hydrogen fuel cell is about twice as efficient as a gasoline-fueled Otto engine, but also because it is so lightweight (but crashworthy) and so low in aerodynamic drag and rolling resistance that it can cruise at highway speed on no more power to the wheels than a conventional SUV uses on a hot day just for air conditioning. This design suggests that cars, too, can "tunnel through the cost barrier," achieving astonishing fuel economy at no extra cost and with many other customer and manufacturing advantages. With aggressive licensing of existing intellectual property, such vehicles could start ramping up production as early as 2008. All major automakers have parallel development programs underway, totaling $\sim$\$10 billion of commitments through 2000, since the basic concept was put into the public domain in 1993 to maximize competition.

A full U.S. fleet of such light vehicles of various shapes and sizes would save about as much oil as Saudi Arabia produces (~ 8 million bbl/day); a global fleet would save as much oil as OPEC sells. Moreover, such vehicles can be designed to plug into the grid when parked (which the average car is $\sim 96\%$ of the time), acting as a power station on wheels, selling back enough electricity and ancillary

services to repay most of its cost. A U.S. fleet of light vehicles doing this would have ~5–10 times as much electric generating capacity as all power companies now own. This is part of a wider strategy that combines hydrogen-ready vehicles with integrated deployment of fuel cells in stationary and mobile applications to make the transition to a climate-safe hydrogen economy profitable at each step, starting now (beginning with ultrareliable combined heat and power in buildings). The resulting displacement of power plants, oil-fueled vehicles, and fossil-fueled boilers and furnaces could decrease net consumption of natural gas, could save about $1 trillion of global vehicle fueling investment over the next 40 years (compared with gasoline-system investment), and could profitably displace up to two-thirds of CO_2 emissions. It could also raise the value of hydrocarbon reserves, in which hydrogen is typically worth more without than with the carbon. It is favored by many leading energy and car companies today, and it is not too far off: over two-thirds of the fossil fuel atoms burned in the world today are already hydrogen, and global hydrogen production (~50 MT/year), if diverted from its present uses, could fuel an entire fleet of superefficient U.S. highway vehicles.

4.2 The Right Steps in the Right Order

Breakthrough energy efficiency results need not just the right technologies but also their application in the right sequence. For example, most practitioners designing lighting retrofits start with more efficient luminaires, improving optics, lamps, and ballasts. But for optimal energy and capital savings, that should be step six, not step one. First come improving the quality of the visual task, optimizing the geometry and cavity reflectance of the space, optimizing lighting quality and quantity, and harvesting daylight. Then, after the luminaire improvements, come better controls, maintenance, management, and training. Likewise, to deliver thermal comfort in a hot climate, most engineers retrofit a more efficient and perhaps variable-speed chiller, variable-speed supply fans, etc. But these should all be step five. The previous four steps are to expand the comfort range (by exploiting such variables as radiant temperature, turbulent air movement, and ventilative chairs); reduce unwanted heat gains within or into the space; exploit passive cooling (ventilative, radiative, ground coupling); and, if needed, harness nonrefrigerative alternative cooling (evaporative, desiccant, absorption, and hybrids of these). These preliminary steps can generally eliminate refrigerative cooling. If refrigerative cooling is still nonetheless desired, it can be made superefficient (e.g., system coefficient of performance 8.6 measured in Singapore), then supplemented by better controls and by coolth storage. Yet most designers pursue these seven steps in reverse order, worst buys first, so they save less energy, pay higher capital costs, yet achieve worse comfort and greater complexity.

Whole-system engineering optimizes for many benefits. There are, for example, 10 benefits of superwindows and 18 benefits of premium-efficiency motors or dimmable electronic lighting ballasts, not just the one benefit normally counted. (The arch that holds up the middle of the author's house has 12 different functions but is paid for only once.) Superwindows cost more per window, but typically make the building cost less because they downsize or eliminate space-conditioning equipment. Similarly, 35 retrofits to a typical industrial motor system, properly counting multiple benefits, can typically save about half its energy (not counting the larger and cheaper savings that should first be captured further downstream, e.g., in pumps and pipes) with a 16-month simple payback against a 5¢/kWh tariff. The saving is so cheap because buying the correct seven improvements first yields 28 more as free by-products. Such motor-system savings alone, if fully implemented, could save about 30% of all electricity worldwide. Such design requires a diverse background, deep curiosity, often a transdisciplinary design team, and meticulous attention to detail. Whole-system design is not what any engineering school appears to be teaching, nor what most customers currently expect, request, reward, or receive. But it represents a key part of the "overhang" of practical, profitable, unbought energy efficiency absent from virtually all official studies so far.

5. MARKET FAILURES AND BUSINESS OPPORTUNITIES

In a typical U.S. office, using one-size-fatter wire to power overhead lights would pay for itself within 20 weeks. Why is that not done? There are several answers: (1) The wire size is specified by the low-bid electrician, who was told to "meet code," and the wire-size table in the National Electrical Code is meant to prevent fires, not to save money. Saving money by optimizing resistive losses requires wire

about twice as fat. (2) The office owner or occupant will buy the electricity, but the electrician bought the wire. An electrician altruistic enough to buy fatter wire is not the low bidder and does not win the job. Correcting this specific market failure requires attention both to the split incentive and to the misinterpretation of a life-safety regulation as an economic optimum. This microexample illustrates the range and depth of market failures that any skilled practitioner of energy efficiency encounters daily. A 1997 compendium, *Climate: Making Sense and Making Money*, organizes 60–80 such market failures into eight categories and illustrates the business opportunity each can be turned into. Some arise in public policy, some are at the level of the firm, and some are in individuals' heads. Most are glaringly perverse. For example, in all but two states in the United States, regulated distribution utilities are rewarded for selling more energy and penalized for cutting customers' bills, so naturally they are unenthusiastic about energy efficiency that would hurt their shareholders. Nearly all architects and engineers, too, are paid for what they spend, not for what they save; "performance-based fees" have been shown to yield superior design, but are rarely used. Most firms set discounted-cashflow targets for their financial returns, yet tell their operating engineers to use a simple-payback screen for energy-saving investments (typically less than 2 years), and the disconnect between these two metrics causes and conceals huge misallocations of capital. When markets and bidding systems are established to augment or replace traditional regulation of energy supply industries, negawatts (saved watts) are rarely allowed to compete against megawatts.

In short, scores of market failures—well understood but widely ignored—cause available and profitable energy efficiency to get only a small fraction of the investment it merits. Thus most of the capital invested in the energy system is being misallocated. The most effective remedy would be to put systematic "barrier-busting" atop the policy agenda, turning obstacles into opportunities and stumbling-blocks into stepping-stones, so market mechanisms could work properly, as economic theory correctly prescribes.

Using energy in a way that saves money is not only a perquisite of the rich, it is also arguably the most powerful key to global economic development for the poor. Using quintupled-efficiency compact fluorescent lamps in Bombay or installing superwindows in Bangkok takes about a thousand times less capital compared to expanding the supply of electricity to produce the same light and comfort via inefficient lamps and office/retail air conditioners. The efficiency investment is also repaid about 10 times faster. The resulting $\sim$10,000-fold decrease in capital requirements could turn the power sector, which now uses about one-fourth of global development capital, into a net exporter of capital to fund other development needs. This is also true at the microscale of a rural village: a package of photovoltaics and superefficient end-use devices (lamps, pumps, mobile phone, water purification, vaccine refrigerator, etc.), with normal utility financing and no subsidy, often incurs debt service lower than what the villagers were already paying for lighting kerosene, candles, and radio batteries, so they have a positive cash flow from day one. Conversely, when Chinese authorities imported many assembly lines to make refrigerators more accessible, the saturation of refrigerators in Beijing households rose from 2 to 62% in 6 years, but the refrigerators' inefficient design created unintended shortages of power and of capital to generate it (an extra half-billion dollars' worth). A Chinese Cabinet member said this error must not be repeated: energy and resource efficiency must be the cornerstone of the development process. Otherwise, resource waste will require supply-side investment of the capital meant to buy the devices that were supposed to use those resources. This realization contributed to China's emphasis on energy efficiency (halving primary energy/GDP elasticity in the 1980s and nearly re-halving it since), laying the groundwork for the dramatic 1996 initial shift from coal to gas, renewables, and efficiency. This greatest contribution of any nation so far to reducing carbon emissions was a by-product of two other domestic goals: eliminating the bottleneck in China's development and improving public health.

6. OLD AND NEW WAYS TO ACCELERATE ENERGY EFFICIENCY

6.1 Old but Good Methods

In the 1980s and 1990s, practitioners and policymakers greatly expanded their tool kits for implementing energy efficiency. During the period 1973–1986, the United States doubled its new-car efficiency, and from 1977 to 1985, cut its oil use 4.8%/year. From 1983 to 1985, 10 million people served by Southern California Edison Company were cutting the decade-ahead forecast of peak load by

$8\frac{1}{2}$%/year, at ~1% of the long-run marginal cost of supply. In 1990, New England Electric System signed up 90% of a pilot market for small-business retrofits in 2 months. In the same year, Pacific Gas and Electric Company (PG&E) marketers captured a fourth of the new commercial construction market for design improvements in 3 months, so in 1991, PG&E raised the target, and got it all in the first 9 days of January.

Such impressive achievements resulted from nearly two decades of refinement of program structures and marketing methods. At first, utilities and governments wanting to help customers save energy offered general, then targeted, information, and sometimes loans or grants. Demonstration programs proved feasibility and streamlined delivery. Standards knocked the worst equipment off the market. (Congress did this for household appliances without a single dissenting vote, because so many appliances are bought by landlords, developers, or public housing authorities—a manifest split incentive with the householders who will later pay the energy bills.) Refrigerator standards alone cut electricity usage by new U.S. units by fourfold from 1975 to 2001 (5%/year), saving 40 GW of electric supply. In Canada, labeling initially did nearly as well. Utilities began to offer rebates—targeted, then generic—to customers, then to other value-chain participants, for adopting energy-saving equipment, or scrapping inefficient equipment, or both. Some rebate structures proved potent, such as paying a shop assistant a bonus for selling an energy-efficient refrigerator but nothing for selling an inefficient one. So did leasing (20¢ per compact fluorescent lamp per month, so that it is paid for over time), paying for better design, and rewarding buyers for beating minimum standards. Energy-saving companies, independent or utility owned, provided turnkey design and installation to reduce hassle. Sweden aggregated technology procurement to offer "golden carrot" rewards to manufacturers bringing innovations to market; once these new products were introduced, competition quickly eliminated their modest price premia. These engineered-service/delivery models worked well, often spectacularly well. Steve Nadel's 1990 review of 237 programs at 38 U.S. utilities found many costing <1¢/kWh (1988 dollars). From 1991 to 1994, the entire demand-side management portfolio of California's three major investor-owned utilities saved electricity at an average program cost that fell from about 2.8 to 1.9 current ¢/kWh (1.2¢ for the cheapest), saving society over $2 billion more than the program cost.

6.2 New and Better Methods

Since the late 1980s, another model has been emerging that promises even better results: not just marketing negawatts (saved watts), maximizing how many customers save and how much, but making markets in negawatts, i.e., also maximizing competition in who saves and how, so as to drive quantity and quality up and cost down. Competitive bidding processes let saved and produced energy compete fairly. Savings can be made fungible in time and space; transferred between customers, utilities, and jurisdictions; and procured by "bounty hunters." Spot, futures, and options markets can be expanded from just megawatts to embrace negawatts too, permitting arbitrage between them. Property owners can commit never to use more than x MW, then trade those commitments in a secondary market that values reduced demand and reduced uncertainty of demand. Efficiency can be cross-marketed between electric and gas distributors, each rewarded for saving either form of energy. Revenue-neutral "feebates" for connecting new buildings to public energy supplies (fees for inefficiency, rebates for efficiency) can reward continuous improvement. Standardized measurement and reporting of energy savings allow savings to be aggregated and securitized like home mortgages, sold promptly into liquid secondary markets, and hence financed easily and cheaply (see the Web site of the nonprofit International Performance Measurement and Verification Protocol, Inc., www.ipmvp.org). Efficiency techniques can be conveniently bundled and translated to "vernacular" forms, which are easily chosen, purchased, and installed. Novel real-estate value propositions emerge from integrating efficiency with on-site renewable supply (part of the revolutionary shift now underway to distributed resources) so as to eliminate all wires and pipes, the trenches carrying them, and the remote infrastructure to which they connect. Performance-based design fees, targeted mass retrofits, greater purchasing aggregation, and systematic barrier busting all show immense promise. Aggressively scrapping inefficient devices, paying bounties to destroy them instead of reselling them, could both solve many domestic problems (e.g., oil, air, and climate in the case of inefficient vehicles) and boost global development by reversing "negative technology transfer."

Altogether, the conventional agendas for promoting energy efficiency—prices and taxes, plus regulation or deregulation—ignore nearly all the most effective, attractive, transideological, and quickly spreadable methods. And they ignore many

of the new marketing "hooks" just starting to be exploited: security (national, community, and individual), economic development and balance of trade, protection from disruptive price volatility, avoiding costly supply overshoot, profitable integration and bundling with renewables, and expressing individual values. Consider, for example, a good compact fluorescent lamp. It emits the same light as an incandescent lamp but uses four to five times less electricity and lasts 8–13 times longer, saving tens of dollars more than it costs. It avoids putting a ton of carbon dioxide and other pollutants into the air. But it does far more. In suitable numbers (half a billion are made each year), it can cut by a fifth the evening peak load that causes blackouts in overloaded Mumbai, it can boost profits of poor American chicken farmers by a fourth, or it can raise the household disposable cash income of destitute Haitians by up to a third. As mentioned previously, making the lamp requires 99.97% less capital than does expanding the supply of electricity, thus freeing investment for other tasks. The lamp cuts power needs to levels that make solar-generated power affordable, so girls in rural huts can learn to read at night, advancing the role of women. One light bulb does all that. It can be bought at the supermarket and self-installed. One light bulb at a time, the world can be made safer. "In short," concludes Jørgen Nørgård, by pursuing the entire efficiency potential systematically and comprehensively, "it is possible in the course of half a century to offer everybody on Earth a joyful and materially decent life with a per capita energy consumption of only a small fraction of today's consumption in the industrialized countries."

6.3 Deemphasizing Traditionally Narrow Price-Centric Perspectives

The burgeoning opportunities for adoption of energy efficiency approaches suggest that price may well become the least important barrier. Price remains important and should be correct, but is only one of many ways to get attention and influence choice; ability to respond to price can be far more important. End-use efficiency may increasingly be marketed and bought mainly for its qualitatively improved services, just as distributed and renewable supply-side resources may be marketed and bought mainly for their distributed benefits. Outcomes would then become decreasingly predictable from economic experience or using economic tools. Meanwhile, disruptive technologies and integrative design methods are clearly inducing dramatic shifts of, not just along, demand curves, and are even making them less relevant by driving customer choice through nonprice variables. Ultralight cars, for example, would do a complete end run around two decades of trench warfare in the U.S. Congress (raising efficiency standards vs. gasoline taxes). They would also defy the industry's standard assumption that efficient cars must trade off other desirable attributes (size, performance, price, safety), requiring government intervention to induce customers to buy the compromised vehicles. If advanced cars can achieve not incremental but fivefold fuel savings as a by-product of breakthrough design integration, yet remain uncompromised and competitively priced, then the energy-price-driven paradigm becomes irrelevant. People will prefer such vehicles because they are better, not because they are clean and efficient, much as most people now buy digital media rather than vinyl phonograph records: they are simply a superior product that redefines market expectations. This implies a world in which fuel price and regulation become far less influential than today, displaced by imaginative, holistic, integrative engineering and marketing. In the world of consumer electronics—ever better, faster, smaller, cheaper—that world is already upon us. In the wider world of energy efficiency, the master key to so many of the world's most vexing problems, it is coming rapidly over the horizon. We need only understand it and do it.

Acknowledgments

The author is grateful to many colleagues, especially Dr. David B. Brooks, Dr. Jørgen S. Nørgård, and Dr. Joel N. Swisher, for their insightful comments, and to the William and Flora Hewlett Foundation for supporting the preparation of this article.

SEE ALSO THE FOLLOWING ARTICLES

Conservation Measures for Energy, History of • *Consumption, Energy, and the Environment* • *Economics of Energy Efficiency* • *Energy Efficiency and Climate Change* • *Industrial Energy Efficiency* • *Market Failures in Energy Markets* • *Obstacles to Energy Efficiency* • *Potential for Energy Efficiency: Developing Nations*

Further Reading

American Institute of Physics (AIP). (1975). "Efficient Use of Energy." AIP Conf. Procs. No. 25. AIP, New York.

Benyus, J. M. (1997). "Biomimicry: Innovation Inspired by Nature." William Morrow, New York.

Brohard, G. J., Brown, M. W., Cavanagh, R., Elberling, L. E., Hernandez, G. R., Lovins, A., and Rosenfeld, A. H. (1998). "Advanced Customer Technology Test for Maximum Energy Efficiency (ACT^2) Project: The Final Report. Proceedings, Summer Study on Energy-Efficient Buildings." American Council for an Energy-Efficient Economy, Washington, D.C.

Daly, H. E. (1996). "Beyond Growth—The Economics of Sustainable Development." Beacon Press, Boston.

Drexler, K. E. (1992). "Nanosystems: molecular machinery, manufacturing, and computation." John Wiley & Sons, New York.

E SOURCE (Platts Research & Consulting). (2002). "Drivepower Technology Atlas" and "Electronic Encyclopedia." E SOURCE, Boulder, Colorado. [Available through their Web site at www.esource.com.]

Fickett, A. P., Gellings, C. W., and Lovins, A. B. (1990). Efficient use of electricity. *Sci. Am.* **263**(3), 64–74.

Hawken, P. G., Lovins, A. B., and Lovins, L. H. (1999). "Natural Capitalism: Creating the Next Industrial Revolution." Little Brown, New York. [See also the Web site at www.natcap.org.]

International Project for Sustainable Energy Paths (IPSEP). (1989–1999). "Energy Policy in the Greenhouse." IPSEP, El Cerrito, California. [www.ipsep.org]

International Project for Sustainable Energy Paths (IPSEP). (2001). "Cutting Carbon Emissions at a Profit: Opportunities for the U.S." IPSEP, El Cerrito, California. [www.ipsep.org]

Johansson, T. B., Bodlund, B., and Williams, R. H. (eds.). (1989). "Electricity." Lund University Press, Lund, Sweden.

Lovins, A. B. (1992). "Energy-Efficient Buildings: Institutional Barriers and Opportunities." Strategic Issues Paper III. E SOURCE (Platts Research & Consulting), Boulder, Colorado. [www.esource.com]

Lovins, A. B. (1994). Apples, oranges, and horned toads. *Electricity J.* **7**(4), 29–49.

Lovins, A. B. (1995). The super-efficient passive building frontier. *ASHRAE J.* **37**(6), 79–81.

Lovins, A. B. (2003). "Twenty Hydrogen Myths." Available from the Rocky Mountain Institute Web site at www.rmi.org.

Lovins, A. B., and Cramer, D. R. (2004). Hypercars®, hydrogen, and the automotive transition. *Int. J. Vehicle Design* (in press).

Lovins, A. B., and Gadgil, A. (1991). "The Negawatt Revolution: Electric Efficiency and Asian Development." RMI Publication #E91-23. Rocky Mountain Institute, Snowmass, Colorado.

Lovins, A. B., and Lovins, L. H. (1991). Least-cost climatic stabilization. *Annu. Rev. Energy Environ.* **16,** 433–531.

Lovins, A. B., and Lovins, L. H. (1996). Negawatts: Twelve transitions, eight improvements, and one distraction. *Energy Policy* **24**(4), 331–344.

Lovins, A. B., and Lovins, L. H. (1997). "Climate: Making Sense *and* Making Money." Rocky Mountain Institute, Snowmass, Colorado.

Lovins, A. B., and Lovins, L. H. (2001). Fool's gold in Alaska. *Foreign Affairs* **80**(4), 72–85.

Lovins, A. B., and Williams, B. D. (1999). "A Strategy for the Hydrogen Transition." Proceedings, 10th Annual Hydrogen Meeting., 8 April. National Hydrogen Assoc., Washington, D.C.

Lovins, A. B., Datta, E. K., Feiler, T., Rábago, K. R., Swisher, J. N., Lehmann, A., and Wicker, K. (2002). "Small Is Profitable: The Hidden Economic Benefits of Making Electrical Resources the Right Size." Rocky Mountain Institute, Snowmass.

Lovins, A. B., Koomey, J. G., Datta, E. K., Bustnes, O.-E., and Swisher, J. N. (2004). "Out of the Oil Box: A Roadmap for U.S. Mobilization." Rocky Mountain Institute, Snowmass, Colorado.

Lovins, A. B., Lovins, L. H., Krause, F., and Bach, W. (1982). "Least-Cost Energy: Solving the CO_2 Problem." Brick House, Andover, Massachusetts. [Reprinted 1989 by Rocky Mountain Institute, Snowmass, Colorado.]

Nadel, S. (1990). "Lessons Learned." Report #90-08. New York State Energy R&D Authority, Albany, with NY State Energy Office and Niagra Mohawk Power Corp. American Council for an Energy-Efficient Economy, Washington, D.C. [www.aceee.org]

Nørgård, J. S. (2002). "Energy Efficiency and the Switch to Renewable Energy Sources. Natural Resource System Challenge II: Climate Change. Human Systems and Policy" (A. Yotova, ed.). UNESCO Encyclopedia of Life Support Systems. EOLSS Publ. Co., Oxford, U.K.

Reddy, A. K. N., Williams, R. H., and Johansson, T. B. (1997). "Energy After Rio: Prospects and Challenges." United Nations Development Program, New York.

Repetto, R., and Austin, D. (1997). "The Costs of Climate Protection: A Guide to the Perplexed." World Resources Institute, Washington, D.C. [www.wri.org]

Romm, J. J., and Browning, W. D. (1994). "Greening the Building and the Bottom Line: Increasing Productivity through Energy-efficient Design." Rocky Mountain Institute, Snowmass, Colorado.

Romm, J. J., Rosenfeld, A. H., and Herrmann, S. (1999). "The Internet Economy and Global Warming: A Scenario of the Impact of E-Commerce on Energy and the Environment." Center for Energy and Climate Solutions, Washington, D.C.

Swisher, J. N. (2002). "The New Business Climate." Rocky Mountain Institute, Snowmass, Colorado.

Swisher, J. N., Jannuzzi, G., and Redlinger, R. (1998). "Tools and Methods for Integrated Resource Planning: Improving Energy Efficiency and Protecting the Environment." UNEP Collaborating Centre on Energy and Environment, Denmark.

von Weizsäcker, E. U., Lovins, A. B., and Lovins, L. H. (1995/1997). "Factor Four: Doubling Wealth, Halving Resource Use." Earthscan, London.

Wilson, A., Uncapher, J., McManigal, L., Lovins, L. H., Cureton, M., and Browning, W. (1998). "Green Development: Integrating Ecology and Real Estate." John Wiley & Sons, New York. [See also "Green Developments," (2002) a CD-ROM published by the Rocky Mountain Institute, Snowmass, Colorado.]

Energy Futures and Options

JOHN ELTING TREAT
Pinnacle Partners
San Francisco, California, United States

1. Introduction
2. Criteria for Successful Futures Markets
3. Exchanges and Their Contracts
4. Futures Present
5. Trading Futures: A Primer
6. Future Directions
7. Summary

Glossary

forward contract A supply contract between a buyer and seller, whereby the buyer is obligated to take delivery, and the seller is obligated to provide delivery, of a fixed amount of a commodity at a predetermined price on a specified future date. Payment in full is due at the time of, or following, delivery.

futures contract A contract between a buyer and seller, whereby the buyer is obligated to take delivery, and the seller is obligated to provide future delivery, of a fixed amount of a commodity at a predetermined price at a specified location. Futures contracts are most often liquidated prior to the delivery date and are generally used as a financial risk management and investment tool rather than for supply purposes.

margin The amount of money or collateral deposited by a customer with a broker, or deposited by a broker with a clearing member, or by a clearing member with the clearinghouse, for the purpose of insuring the broker or clearinghouse against adverse price movement on open futures contracts.

spread (futures) The simultaneous purchase and sale of futures contracts for different months, different commodities, or different grades of the same commodity.

spread (options) The purchase and sale of options that vary in terms of type (call or put), strike prices, and/or expiration dates. May also refer to an options contract purchase (sale) and the simultaneous sale (purchase) of a futures contract for the same underlying commodity.

Energy futures trading in the 1980s focused on growth of the liquid petroleum markets for crude oil, natural gas liquids, and the major refined products of gasoline, heating oil, and fuel oil. In the 1990s, the boundaries of the energy complex expanded to include, first, natural gas (in 1990), then electricity (in 1996), and to add a wide range of options contracts based on the underlying futures markets.

1. INTRODUCTION

In the second half of the 19th century, a "petroleum exchange" flourished in New York. Again, in the early 1930s, when market discipline was briefly disrupted by the explosive growth of oil production in Oklahoma and Texas and oil prices fell dramatically, an oil futures market (in West Texas Intermediate crude oil) was established in California. It soon collapsed as a formidable alliance of "big oil" and "big government" restored discipline to the marketplace. Nearly 40 years of relative price stability ensued, leaving little incentive for the emergence of an oil futures market.

Only with the traumatic price increases accompanying the Arab oil embargo of late 1973 was another attempt made, this time back in New York at the Cotton Exchange. The contract called for Rotterdam delivery (to avoid the constraints of U.S. price regulations). That attempt was stillborn, however, doomed by continuing U.S. government price controls and a skeptical oil industry. In the decade that followed, the commercial realities and—equally important—the perceptions of those realities by the international oil industry had gradually changed; the point at which oil futures could fulfill the industry's need for risk management and could provide an outlet for the speculative impulses of investors had been reached. Investors' interest in oil had been captured by the commodity's new prominence in daily headlines and nightly newscasts.

The emergence of oil futures markets and their remarkable growth was a natural, indeed inevitable,

 © 2004 Elsevier Inc. All rights reserved.

consequence of three concurrent, but only partly interrelated, trends in petroleum, financial, and commodity markets. By far the most important determinant was the structural change in oil markets. The nationalization of production by members of the Organization of Petroleum Exporting Countries (OPEC) and non-OPEC governments alike, and the subsequent pressures to eliminate large third-party crude resales, resulted in a disintegration of an oil market that had been highly integrated since the days of J. D. Rockefeller. In the 10 years following the 1973 Arab oil embargo, the crude oil available to the major companies fell by nearly 50%, from about 30 million barrels (bbl)/day to just over 15 million bbl/day. Equity oil available to the majors fell even more sharply, by some 75%. The net result was a drop in the major's share of internationally traded oil from 62% to just 37%.

In addition to the newly created national oil companies, a host of oil trading firms and independent refineries entered the picture. The links that had traditionally tied upstream and downstream (vertical) integration were weakened to the breaking point. The reduction of horizontal integration (as large third-party crude sales were curtailed and joint venture production was nationalized) further eroded the ability of the largest oil companies to exercise control over markets. Simply put, there were too many actors with divergent commercial and political interests to guarantee market stability. The consequences were not long in coming (see Table I).

After a decade of virtually universal confidence that oil prices would rise, prices began to weaken and fluctuate over an ever-wider range, climaxing in the dramatic events of 1986, when prices fell from nearly $30/bbl to less than $10/bbl in a period of only 9 months. Although many market observers feel stability returned in 1987, it is interesting to note that prices oscillated between $15 and $22/bbl between September and December of that year alone. The fact is that stability has yet to rear its hoary head in current-day oil markets.

Along with the structural change that was reshaping oil markets during the decade, a second important trend was emerging from the financial markets. High interest rates (along with high oil prices) at the beginning of the 1980s were making inventory maintenance very expensive, causing oil company management teams to rethink traditional approaches to inventory and risk management. Also, hedging of financial risks was increasingly becoming a fact of life in foreign currency and interest rate markets. These trends ensured that oil companies were increasingly receptive to the hedging potential of the fledgling oil future markets. Finally, the third important factor that set the stage for the ultimate success of energy futures was the general growth and diversification of futures contracts in a wide variety of new markets, and the growing sophistication with which they were being used and modified to offer an ever wider range of hedging tools. For almost 100 years, futures markets (then commonly called commodity markets) were largely confined to the

TABLE I

Crude Oil Price Trends[a]

Year	Price
1950	$2.51
1960	2.88
1961	2.89
1962	2.90
1963	2.89
1964	2.88
1965	2.86
1966	2.88
1967	2.92
1968	2.94
1969	3.09
1970	3.18
1971	3.39
1972	3.39
1973	3.89
1974	6.74
1975	7.56
1976	12.17
1977	13.24
1978	13.30
1979	20.19
1980	32.27
1981	35.10
1982	32.11
1983	27.73
1984	27.44
1985	25.83
1986	12.52
1987	16.69
1988	13.27
1989	15.50

[a] Average annual prices (U.S. dollars). Data from the U.S. Energy Information Administration (EIA), *Monthly Energy Review*, March 1989 (published June 1989). Prices shown for 1950–1975 are from the EIA's "U.S. Wellhead." After 1975, the EIA series is "FOB Cost of Imports" into the United States; 1989 prices are based on the first quarter.

traditional agricultural products (especially grains). In the past two decades, however, there has been an explosion in the variety of products served by these markets. The first waves of expansion brought in new agricultural contracts (especially meats) and precious metals. The second phase, starting in the 1970s, saw the introduction of financial instruments, including currency, interest rate, and stock index contracts. A third phase brought in oil and a number of other industrial products. The fourth stage saw the introduction and rapid acceptance of options on futures contracts. The fifth stage saw an explosion of trading in over-the-counter (OTC) derivatives, often in direct competition with exchange-traded instruments. The introduction and success of oil futures was a product of the first three trends. The growing volatility and loss of confidence in the future stability of oil prices demanded the emergence of new market structures and institutions. One obvious sign of the change was the rapid growth of spot markets and the trading companies that thrived on price volatility.

Prior to 1979, less than 5% of internationally traded crude moved at spot prices, that is, outside of official term supply contract arrangements. By the end of 1985, virtually all crude moved at some sort of market-related pricing, and experts estimated that oil companies were acquiring anywhere from 30 to 50% of their supplies on a spot, noncontract basis. Although the proportion of oil sold on a purely spot basis has subsequently shrunk, the price risks remain, because term contracts today almost universally call for market-related prices. The trading companies, independent refineries, and increasingly the larger companies developed trading techniques to cope with the growing price volatility of these markets. The first response was to create informal, forward markets, at first only 30 days out, then 60 days out, and more recently 90 days out. A second response was an explosive growth in the demand for rapid, often real-time, pricing and other market information. Against this backdrop, futures became inevitable: a time-proven and efficient technique for coping with broad market instability.

Although the breadth of energy markets has expanded, their fundamental purposes remain the same. Futures markets, basically spot markets for standardized forward contracts, serve three functions:

1. Price discovery, giving an instantaneous reading of marginal price movements.
2. Risk management, allowing companies to hedge their price risks for limited periods of time (however, the hedging opportunity rarely extends more than 6 months forward as a result of market illiquidity in the more distant months).
3. Speculative opportunity, attracting additional risk capital to the market from outside the oil industry (low margin requirements—lower than in equity markets—enhance the attraction of futures as a vehicle for speculation).

These are the necessary but not sufficient conditions for a successful contract. In reality, however, new futures contracts often fail. The reason is that the criteria for a successful futures contract are simply too stringent, with too few physical markets that can actually meet those criteria.

2. *CRITERIA FOR SUCCESSFUL FUTURES MARKETS*

Assessing the suitability of any commodity/market for futures trading requires analysis of a number of conditions. These are considered in the following sections.

2.1 Price Volatility

Price volatility is perhaps the single most important criterion for assessing futures trading. It provides the basic economic justification for futures trading, which is to provide protection to the hedger against adverse price fluctuations. If a commodity is characterized by a relatively stable—or at least predictable—price, there would be little associated risk and there would be no need for a futures market. Price volatility is also necessary to attract risk capital from speculators and is essential to ensure sufficient liquidity to maintain the market.

Quantitative indicators: Variations of $\pm 20\%$ per annum are assumed to be the minimum necessary to sustain futures trading. In general, the greater the degree of volatility, the more likely a futures market will survive.

2.2 Uncertain Supply and Demand

Uncertainty in supply and demand is generally the cause of price volatility and therefore is generally present when price volatility is found.

Quantitative indicators: In energy markets, which typically display a rather high inelasticity of price demand, variations of $\pm 10\%$ during a 2-year period should be sufficient to sustain futures trading.

2.3 Sufficient Deliverable Supplies

Sufficiency of supply is the Catch 22 of futures trading. If sufficient deliverable supplies of the commodity meeting the quality specification of the contract are not available, futures trading will fail. However, there must be some uncertainty about the sufficiency of supplies if the previous conditions are to be met. In the United States, this dilemma is heightened by the regulatory requirements of the Commodity Futures Trading Commission (CFTC), whose fear of market squeezes at times forces exchanges to overstate deliverable supplies in order to gain government approval.

Quantitative indicators: Storage capacity equal to at least 30 days average demand is highly desirable.

2.4 Product Homogeneity

Product homogeneity is another prerequisite. Futures contracts are traded on the premise that product taken on a delivery will meet certain quality specifications. The commodity must therefore have certain key characteristics that are quantifiable, allowing the clear differentiation of product from other grades. Standardized tests and generally accepted procedures are essential. For example, the various American Petroleum Institute (API), German Institute for Standardization, and American Society for Testing and Materials (ASTM) standards generally provide the necessary references for oil. In addition, the existence of generally trusted independent inspection agencies or inspectors to administer these tests is an important aspect. A range of different products (for example, several types of crude oil) may be suitable for delivery, if the price differences between the various grades are relatively stable and if the technical characteristics of the various deliverable grades are sufficiently close to one another. This is often a difficult aspect of contract design, because the price variation between various grades of products fluctuates from time to time. For example, it may be desirable to allow several grades to be deliverable, perhaps with price adjustments for quality, in order to ensure sufficient deliverable supplies. However, if buyers are uncertain of what grades they will receive and if they place different values on the quality differences among the grades, they may be deterred from trading.

Quantitative indicators: The quality of the product must be capable of being described by objective, quantifiable standards.

2.5 Product Perishability

The perishability of a product can be a deterrent to trading. In general, a product should have a shelf life sufficiently long enough to permit storage and delivery as called for under the contract. In addition, the maintenance of inventories of the commodity will both facilitate deliveries and provide a ready pool of potential hedgers. While perishability is not usually a major concern in oil and natural gas markets, the stability of some oil product blends is an issue. Long storage of gasoline, for example, can result in separation of blended product.

Quantitative indicators: Products should have a minimum shelf or stock life of 6–12 months.

2.6 Market Concentration

Market concentration is a difficult factor to quantify. A successful futures market is a highly competitive market, marked by a large number of buyers and sellers. No one market participant, or plausible combination of market participants, should possess sufficient market power to exert unilateral control either on the supply or the demand for the commodity, either in the short or medium term. In oil markets, however, the existence of OPEC has not prevented the emergence of highly successful futures markets. The answer lies in the inability of OPEC to act decisively, and in the availability of alternative sources of supply and stocks that seriously limit OPEC's ability to achieve its stated objectives. However, the concentration of producers and/or consumers can be a serious obstacle in specific regional oil markets. Thus, a U.S. West Coast gasoline market would be risky, given the relative concentration of production in the hands of a small number of refiners. Similarly, an East Coast residual fuel market might be too heavily dominated by the demand from a small number of very large utilities to sustain liquid futures trading.

Quantitative indicators: In general, the market share of the top five firms should be less than 50%, and the top 10 firms should have less than 80%.

2.7 Availability of Price Information

Readily available price information is critical to market success. It should be noted that the opening of a futures market will probably, in turn, stimulate a rapid growth of price information services. However, at the outset, market participants must have a sufficiently broad base of price information to permit evaluation of spot prices and their relationship to

futures prices. Convergence between these two prices as the delivery period approaches is essential. A market in which all products are traded on the basis of long-term contracts for which prices remain undisclosed would be a very difficult market in which to establish futures trading.

Quantitative indicators: Daily cash market prices should be available from at least two independent sources.

2.8 Uniqueness

The uniqueness of a trading opportunity is another key factor. If an existing market for a commodity has reasonable liquidity and is serving its customers well, it is extremely difficult to launch a copycat contract. Inertia, habit, and personal relationships will tend to keep the traders loyal to the preexisting market. In addition, even if there is no active market at present, recent failures of similar contracts can be a substantial (but not fatal) deterrent.

Quantitative indicators: The ideal candidate would be a commodity that is not currently traded on any futures exchange in the world and has not been the subject of a failed attempt in the previous 5 years. However, special circumstances may override these concerns.

2.9 Market Timing

Market timing and blind luck are often critical to the success or failure of a contract. However, forecasting is often impossible. Ideally, contracts should be introduced to coincide with periods of high volatility and high levels of cash market activity. For example, a heating oil or natural gas contract would be best introduced in the fall months, when physical trading is at its yearly high. Conversely, a gasoline contract would be best introduced in the spring, prior to an anticipated surge of summer driving.

Quantitative indicators: Contracts should be introduced to coincide with high levels of cash market activity, to the extent that this is predictable. Alternatively, an astrologer might just as well be consulted.

3. EXCHANGES AND THEIR CONTRACTS

Two exchanges currently dominate trade of energy futures contracts: the New York Mercantile Exchange and the London International Petroleum Exchange. In addition several smaller exchanges also have offered energy futures contracts, albeit with limited and often transitory success, including the Singapore International Monetary Exchange (SIMEX; currently known as the Singapore Exchange Ltd.) and the Kansas City Board of Trade. The future of open-outcry trading on exchange floors is increasingly being challenged by the advent of electronic trading. Although the exchanges and their floor traders still have the upper hand, the growth of electronic trading and the success of all-electronic marketplaces (such as the Internet) pose serious challenges for the exchanges and their trading floors.

3.1 New York Mercantile Exchange

Founded more than 100 years ago, the New York Mercantile Exchange (NYMEX), or the "Merc," as it is often called, has enjoyed a diverse and colorful history. Evolving from a lower Manhattan produce exchange that dealt in contracts for butter, eggs, and even apples, NYMEX began to diversify some two decades ago, adding precious metals and—briefly—even stocks to its portfolio. Nevertheless, by the end of the 1970s, it was one of the smallest exchanges in the United States, outpaced by the growth of the large Chicago markets and by most other New York exchanges as well. After an abortive attempt to start a residual fuel oil contract in 1974, the Merc launched its first successful energy futures contract—New York heating oil—in 1978. Trading grew slowly but steadily while a companion contract for residual fuel was stillborn. With the addition of a leaded gasoline contract for New York delivery in 1981 (later replaced by an unleaded version of the same contract) and West Texas Intermediate crude oil in 1983, NYMEX achieved international prominence in the 1980s; its energy contracts grew at spectacular rates. The subsequent addition of options on crude oil, heating oil, and gasoline as well as futures contracts on propane, natural gas, and electricity added another dimension and further impetus to the growth of the NYMEX energy complex.

Since its merger with the Commodities Exchange (COMEX) in the mid-1990s, NYMEX has been known as the NYMEX Division of the merged exchange. Today, the NYMEX Division is the leading energy futures exchange and the third largest futures exchange in the world, following only the Chicago Board of Trade and the Chicago Mercantile Exchange. Although contracts for platinum and

palladium still survive on the Merc's trading floor, energy contracts regularly account for more than 90% of its turnover. NYMEX, for most of its life, was a nonprofit organization. However, in 2000, its members voted to "demutualize" and the exchange became a for-profit corporation, the first New York commodity exchange to do so.

3.2 International Petroleum Exchange

An independent outgrowth of London's loosely linked futures markets, the International Petroleum Exchange (IPE) was created in 1981 by a diverse coalition of oil traders and commodity brokerage firm members who saw the emerging success of the NYMEX heating oil contract in New York and were determined to build an energy futures market on the other shore of the Atlantic. Early success with a gas oil contract was followed by a series of unsuccessful attempts to trade crude oil. Finally, in 1988, the right set of circumstances and contract terms allowed IPE to launch a Brent crude oil market that has established good liquidity and a substantial following. Although IPE occupies second place among the world's energy exchanges, the gap between IPE and NYMEX is very large, with London's turnover averaging well under half of New York's volume. Following the NYMEX demutualization, the IPE too converted to for-profit status in 2001 and subsequently merged with the U.S.-based Intercontinental Exchange.

3.3 Singapore Exchange Ltd

The Singapore International Monetary Exchange was created in 1983 as a restructuring of the former Gold Exchange of Singapore (founded in 1978) with the strong support of the Chicago Mercantile Exchange (CME). CME and SIMEX offered customers an innovative mutual offset agreement, whereby positions on one exchange can be offset against positions on the other. Initially, the exchange concentrated on financial futures. but in February 1989 SIMEX launched a high-sulfur residual fuel oil contract that enjoyed a brief success. In the mid-1990s, they traded Brent crude oil for several years.

In 1999, SIMEX merged with the Stock Exchange of Singapore to form the Singapore Exchange Ltd. (SGX). Their futures trading is organized under the SGX Derivatives Trading Division (SGX-DT), which has recently launched a contract on Dubai crude oil—the second attempt to trade Middle Eastern crude. Like the other major exchanges, the SGX has also converted to a for-profit corporation.

3.4 Other Exchanges

From time to time, other futures exchanges, most recently Kansas City, have launched energy futures contracts. However, none of these has succeeded in building the liquidity needed to attract significant trading.

4. FUTURES PRESENT

Since 1974, there have been over 60 attempts to launch energy futures and options markets. The success rate has averaged about 20%, typical of the experience in other commodity markets. In spite of thorough research by the exchanges, often excruciating governmental reviews (particularly in the United States), and extensive marketing campaigns, roughly 80% of all future markets opened fail to reach the critical mass needed for takeoff (commonly defined as reaching an average open interest of 5000 contracts). Table II lists the various attempts to launch new futures markets in energy from 1970 through 2002.

When first introduced in late 1978, heating oil futures attracted smaller, independent marketers and refiners, who turned to the New York Merc as an alternative source of supply. Physical deliveries were initially quite high, as these smaller firms sought alternatives in a marketplace dominated by the larger companies. These initial participants were quickly joined by the spot oil traders and by a growing number of pure speculators on and off the trading floor, drawn from other financial and commodity markets. This phase lasted until well into 1983. Then, with the introduction of crude oil futures and the increasing instability of prices, the larger refiners and integrated companies reluctantly entered the market.

By 1984, more than 80% of the 50 largest companies were using futures. Larger end-users, such as airlines and other major energy consumers, also appeared. In addition, a far wider range of speculators entered the scene, as trading volume and open interest rose high enough to meet the minimum liquidity requirement of the commodity funds. Finally, another phase, dating from 1986, brought in almost all the remaining holdouts among the large U.S. companies, more foreign participation, and a new group of traders, the Wall Street refiners. These were companies such as Morgan Stanley and Bear Stearns, which were attracted by the rising volatility of oil prices and the speculative opportunities presented by that price instability, particularly relative to other markets. As one Bear Stearns trader

TABLE II

Energy Futures Markets Authorized, Opened, or Proposed, 1970–2002[a]

Year	Commodity[b]	Delivery point	Exchange[c]	Status
1974	Crude oil	Rotterdam	NYCE	Failed
	Residual fuel	Rotterdam	NYMEX	Failed
1978	Heating oil	New York	NYMEX	Active
	Residual fuel	New York	NYMEX	Failed
1981	Gas oil	Rotterdam	IPE	Active
	Heating oil	Gulf Coast	NYMEX	Failed
	Leaded gasoline	New York	NYMEX	Replaced
	Leaded gasoline	Gulf Coast	NYMEX	Failed
	Unleaded gasoline	Gulf Coast	CBOT	Failed
	Unleaded gasoline	Gulf Coast	NYMEX	Unopened
	Propane	Texas	NYCE	Failed
1983	WTI crude oil	Crushing, OK	NYMEX	Active
	LLS crude oil	St. James, LA	CBOT	Failed
	Brent crude oil	Rotterdam	IPE	Failed
	Heating oil	Gulf Coast	CBOT	Failed
1984	Unleaded gasoline	New York	NYMEX	Active
	Leaded gasoline	Gulf Coast	CME	Failed
	Unleaded gasoline	Gulf Coast	CME	Unopened
1985	Brent crude oil	North Sea	IPE	Failed
1986	WTI crude oil options	Cushing, OK	NYMEX	Active
1987	Heating oil options	New York	NYMEX	Active
	Brent crude oil options	Rotterdam	EOE	Failed
1988	Brent crude oil	Cash settlement	IPE	Active
	Propane	Texas	NYMEX	Active
1989	Residual fuel oil	Singapore	SIMEX	Active
	Unleaded gasoline options	New York	NYMEX	Active
	Brent crude oil options	Cash settlement	IPE	Active
	Gas oil options	Rotterdam	IPE	Active
	Brent crude oil	Rotterdam	ROEFEX	Failed
	Gas oil	Rotterdam	ROEFEX	Failed
	Residual fuel oil	Rotterdam	ROEFEX	Failed
	Residual fuel oil	Cash settlement	IPE	Lightly traded
	Residual fuel oil	Gulf coast	NYMEX	Failed
1990	Natural gas	Henry Hub, LA	NYMEX	Active
	Dubai crude oil	Cash settlement	SIMEX	Failed
	Dubai crude oil	Cash settlement	IPE	Failed
1992	Sour crude	Gulf Coast	NYMEX	Failed
	Gulf Coast gasoline	Gulf Coast	NYMEX	Failed
	Natural gas options	Henry Hub, LA	NYMEX	Active
1994	NY gasoline/crude crack spread	New York	NYMEX	Active
	NYH heating oil/crude crack spread	New York	NYMEX	Active
1996	COB electricity	COB	NYMEX	Failed
	Palo Verde electricity	Palo Verde	NYMEX	Failed
	COB electricity options	COB	NYMEX	Failed
	Palo Verde electricity options	Palo Verde, CA	NYMEX	Failed
	Permian basin natural gas	Permian Basin	NYMEX	Failed
	Alberta natural gas	Alberta, Canada	NYMEX	Failed

continues

Table II continued

Year	Commodity[b]	Delivery point	Exchange[c]	Status
1998	Cinergy electricity	Ohio	NYMEX	Failed
	Cinergy electricity options	Ohio	NYMEX	Failed
	Entergy electricity	Arkansas	NYMEX	Failed
	Entergy electricity options	Arkansas	NYMEX	Failed
	PJM electricity	New Jersey	NYMEX	Failed
1999	Middle East crude	United States	NYMEX	Failed
	Mid-Columbia electricity	United States	NYMEX	Failed
2001	Central Appalachian coal	United States	NYMEX	Failed
	Brent crude options	United Kingdom	NYMEX	Failed
2002	Middle East crude	Asia	SIMEX	Active

[a] Year first traded or, if never opened, year first authorized. Does not include various contract modifications or innovations in trading, e.g., in 2001 NYMEX began to offer "strip" trading in natural gas contracts across several months and introduced Exchanges of Futures for Swaps (EFS).

[b] Abbreviations: WTI, West Texas Intermediate; LLS, Light Louisiana Sweet; NYH, New York Harbor; COB, California/Oregon Border; PJM, Pennsylvania/New Jersey/Maryland.

[c] Abbreviations: NYCE, New York Cotton Exchange; NYMEX, New York Mercantile Exchange; IPE, International Petroleum Exchange; CBOT, Chicago Board of Trade; CME, Chicago Mercantile Exchange; EOE, European Options Exchange; SIMEX, Singapore International Monetary Exchange; ROEFEX, Rotterdam Energy Futures Exchange.

put it, "Plywood was dead, so we looked around for some better action and found it in oil." The low internal cost of capital for margin maintenance and a built-in trading infrastructure made these new entrants formidable competitors for the older oil trading and supply companies. However, even today, participation by independent producers and smaller end-users remains limited. Participation of the producers is limited by the lack of liquidity in the more distant months; participation of the end-users is limited by ignorance of how the markets operate, the high management cost of setting up a futures trading department, and, for a number of domestic as well as international companies, a very real basis-risk problem.

Futures trading has thus survived adolescence and entered a period of youthful maturity. But growth in the coming years will have to come from an expansion of futures trading opportunities in the form of new contracts rather than from bringing in new participants. In other words, to continue to grow, the exchanges will have to offer a bigger and more diverse menu, not just put more seats around the table. The recent success of options would seem to confirm this point of view. Today, after the smoke has settled from various attempts by exchanges in New York, Chicago, London, and Singapore, there are six well established energy markets in New York and London: (1) NYMEX West Texas Intermediate (WTI) crude oil, (2) NYMEX natural gas, (3) IPE Brent crude oil, (4) NYMEX heating oil, (5) NYMEX unleaded gasoline, and (6) IPE gas oil. In addition, options contracts have been successful as extensions of those future markets. The latest innovations by the exchanges include the introduction of trading in smaller contracts (the e-miNY at NYMEX), the expansion of electronic and after-hours trading, and the introduction of OTC trading and clearing services. Trading volumes of the major contracts are shown on Table III.

5. TRADING FUTURES: A PRIMER

Many readers of this article will be thoroughly familiar with the basic mechanisms and concepts of futures trading. This section is not for them. However, for those who are new to any type of futures trading, it is important to understand a few fundamentals about futures markets.

Futures markets offer hedgers or commercials (i.e., those who use a particular commodity in their business) and speculators the opportunity to buy or sell standardized contracts for a given commodity. In many cases, the same exchanges also offer options contracts on those same commodities. Options contracts as presently traded are options on the futures contract for the same commodity, which is often called the underlying futures.

5.1 Contract Identification

Both futures and options contracts are identified not only by the particular type of commodity being

TABLE III

Growth in Energy Futures and Options Trading (Daily Average Volume of Contracts Traded (000))[a]

Contract	Exchange	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
Futures																						
Crude oil																						
WTI	NYMEX	—	1.7	7.4	15.9	33.3	57.9	74.8	81.8	94.4	83.0	83.8	99.5	106.8	94.4	93.6	98.3	121.1	151.4	148.1	149.0	182.7
Brent	IPE	—	0.0	0.0	0.0	0.0	0.0	2.2	6.6	16.2	20.8	24.5	35.1	40.0	38.8	42.4	40.9	53.6	66.5			
Brent	SIMEX	—	—	—	—	—	—	—	—	—	—	—	—	—	0.5	0.2	0.1	0.1	—	—	—	—
MidEast crude	SIMEX	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0.1
Products																						
Heating oil	NYMEX	6.9	7.4	8.4	8.8	13.1	17.0	19.6	22.9	25.4	26.4	31.8	34.5	35.8	33.1	33.2	33.2	35.3	36.8	38.6	41.7	42.8
Gasoil	IPE	2.5	2.4	2.1	2.0	3.7	4.4	6.2	7.8	10.3	11.3	13.7	14.3	15.0	17.8	17.3	16.0	19.5	25.5			
Unleaded gas	NYMEX	—	—	—	0.5	1.8	8.2	13.1	17.9	20.7	21.8	26.5	29.6	29.8	28.3	25.1	29.7	31.9	34.8	34.7	38.0	43.9
Resid	SIMEX	—	—	—	—	—	—	—	4.6	0.8	0.7	1.1	1.2	0.7	0.0	0.0	—	—	—	—	—	—
Propane	NYMEX	—	—	—	—	—	0.2	0.1	0.1	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.1	0.1	0.1
Natural gas																						
Natural gas	NYMEX	—	—	—	—	—	—	—	—	0.5	1.7	7.6	18.7	25.3	32.3	35.1	47.3	63.6	76.6	71.8	47.5	97.4
Natural gas	IPE	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0.3	1.3	1.1			
Electricity																						
Palo Verde	NYMEX	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0.1	0.6	0.6	0.2			
California Oregon border	NYMEX	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0.3	0.5	0.6	0.3			
Cinergy	NYMEX	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0.4	0.1			
Entergy	NYMEX	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0.4	0.1			
Coal																						
Central Appalachian coal	NYMEX	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—		
Options/Spreads																						
Crude oil																						
WTI	NYMEX	—	—	—	—	4.5	12.4	21.7	22.7	20.9	19.6	26.0	28.6	22.6	15.9	21.0	23.0	28.9	32.6	30.0	27.8	80.7
Brent	IPE	—	—	—	—	—	—	—	0.2	0.6	0.9	3.1	4.2	2.1	2.3	1.5	1.0	11.8	9.3			
Products																						
Heating oil	NYMEX	—	—	—	—	—	1.1	0.5	1.2	1.6	3.4	5.0	3.2	2.8	2.8	4.4	4.8	2.6	2.8	5.6	4.2	2.4
Gasoil	IPE	—	—	—	—	—	0.0	0.1	0.1	0.4	0.4	0.8	0.9	0.5	0.5	0.4	0.3	7.4	3.5			
Gasoline	NYMEX	—	—	—	—	—	—	—	1.3	1.7	2.3	3.4	2.6	2.3	3.1	2.6	4.1	3	2.4	4.1	4.5	2.9
Natural gas																						
Natural gas	NYMEX	—	—	—	—	—	—	—	—	—	—	0.3	1.4	2.0	3.7	4.9	8.3	10.3	15.4	21.4	15.7	43.9
Spreads																						
Gasoline–Crude oil spread	NYMEX	—	—	—	—	—	—	—	—	—	—	—	—	0.2	0.3	0.1	0.2	0.1	0.2	0.1	0.1	0.1
Heating oil–Crude oil spread	NYMEX	—	—	—	—	—	—	—	—	—	—	—	—	0.2	0.3	0.2	0.1	0.1	0.2	0.2	0.2	0.1
Electricity																						
Palo Verde	NYMEX	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0.1	0.1	0.1	—	—	—	—
California Oregon border	NYMEX	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0.3	0.1	—	—	—	—	—
Cinergy	NYMEX	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Entergy	NYMEX	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—

[a] Annual averages, except for years on contract introduction, when averages computed on actual number of days traded. Minimal trading volumes entered as 0.1.

traded (e.g., heating oil, unleaded gasoline, Brent crude oil), but also by the delivery month called for in the contract. In practice, traders often abbreviate the names of months, so that it is common to hear references to "Feb Brent" or "Jan gas." Options contracts are further identified by their strike prices and whether they are options to buy (call) or sell (put). Thus an options trader will talk about "Jan gas 55 puts," meaning options to sell January unleaded gasoline futures contracts at 55 ¢/gallon. A buyer of a commodity contract is said to be long while he holds that contract. A seller is said to be short. The contracts traded are highly standardized with respect to volume, quality, and delivery terms. Exchanges do, however, change the terms of these contracts from time to time, to keep pace with changes in the physical market. The appropriate exchange can provide a copy of the latest contract terms.

5.2 Placing Orders

Except in the case of exchange members operating on their own account, all transactions must be conducted through a member of the exchange, who must also be registered to accept customer orders by the Commodities Futures Trading Commission in the United States or its counterparts in other countries. Orders can be placed anytime a broker is willing to answer the telephone, but exchange trading hours tend to fall between 9:00 a.m. and 5:00 p.m., with the New York exchanges closing earlier and the London exchanges closing later.

A buyer normally places an order by telephone to a broker, who may be located anywhere in the world. The broker in turn executes this order by telephone through exchange members on the floor of the appropriate exchange. Buyers can place various conditions on their orders, including price limits or time limits, and they may also simultaneously request a broker to close out the position if losses exceed a certain amount. Although brokers will generally accept such conditions on orders, they usually offer no guarantees that they can execute the order as given. Only a market order, in which the buyer (or seller) agrees to accept the prevailing market price, is virtually guaranteed for execution. Brokers will most often execute orders through the employees of their own firm on the floor. But in order to camouflage larger orders, execution will sometimes be shared with independent floor brokers, who execute orders on behalf of others.

On the exchange floor, all trading must be by open outcry, giving all present—at least in theory—an equal opportunity to take the other side of the trade. Assuming a willing seller is then found to meet the buyer's order, the trade is posted with the exchange. In practice, exchanges publish price quotations over the various electronic information services to provide an up-to-date record of pricing trends even before the official record of the transaction is entered. The actual trade is usually entered into the exchange's computer within a few minutes of the transaction. The buyer and seller, however, are not matched permanently. At the end of the day, each broker is assigned an appropriate long or short position with the exchange's clearing house. Thus, although there must be an equal number of buyers and sellers each day, their respective positions are maintained totally independently, and each buyer and seller is free to close his/her position at any time. To do so, the buyer will simply sell his/her contract back into the market, effectively clearing his position from the exchange's books.

5.3 Spreads

In addition to straightforward orders to buy or sell a single commodity for a single month, many traders take spread positions, which are positions in several different contracts to profit from the relative price movements between those contracts. For example, spreads can be placed between contracts for different delivery months for a single commodity. Or they can cover different commodities for delivery in the same month. Or they can cover different commodities and different months. One popular type of spread position in the energy contracts is the "crack spread," in which a position in crude oil is balanced against positions in both gasoline and heating oil, to approximate the refining process (in which crude oil is transformed by catalytic cracking into refined products). NYMEX has offered option contracts of such spreads since 1994. A newer type of spread trading is the "spark spread," which pairs positions in natural gas with those in electricity to approximate the gross margin of power generation using natural gas.

5.4 Margins and Clearing Houses

Exchanges collect margins (or deposits, as they are called in England) from each broker on behalf of the customer (and in most cases, the broker in turn collects similar funds from the customer). These margins, which are usually in the range of 5–10% of the contract's total face value, are normally designed to be equal to the average daily fluctuation in value

of the contract being traded. Exchanges will therefore tend to lower margins in times of low price volatility and to raise them in times of high price volatility. Every night, based on the final closing or settlement price, the exchange calculates the effect of that price on each position and either requests additional margin or pays excess margin to each broker. Thus, if prices go up from one day to the next, a buyer's margin is credited with a gain and a seller's margin is debited. The rules of margin maintenance by customers to brokers vary considerably from country to country.

The exchange and its clearinghouse are therefore always in a very strong position to guarantee all outstanding positions. Moreover, the clearinghouse holds its member brokers—not the ultimate customer—responsible for performing under the contracts. In the unlikely event that an individual broker is unable to perform as called for under the contract, all the members of the clearinghouse are called upon to guarantee performance. Futures markets therefore offer several levels of financial performance guarantees.

Prices on an exchange are freely determined by the interplay of buyers and sellers. However, exchanges do place certain limits on both the minimum and maximum amounts of fluctuation that can occur in a given time period. The minimum price change is referred to as a tick and, in the oil contracts, is typically equal to 1.00¢/bbl, 25.00¢/MT, or 0.01¢/gallon in New York. In all cases, there are no limits on the spot contract, which is the contract that is next scheduled to go to delivery, but all other contracts face limits. In New York, these limits are typically $1/bbl, $15/MT, or 2.00¢/gallon. In a given day, no trades may take place outside these ranges. However, if a limit is reached on one day, the limits are expanded by 50% for the next day's trading, and so on, up to a maximum of $2/bbl or 4.00¢/gallon. In London, the limits do not apply for a full day, but rather trigger cooling-off periods before trading is resumed.

5.5 Delivery

These markets should always be thought of primarily as financial markets, being used in parallel with physical movement of oil and natural gas. Nevertheless, delivery does take place and serves to ensure that the prices on futures markets remain closely linked to the real world. The standardization of contracts and their delivery terms are often unnecessarily rigid for the commercial participants, who prefer greater flexibility in their day-to-day operations. As a consequence, delivery typically occurs in only about 2% or less of all futures contracts. In simplest form, all those holding positions in a given contract at the closing bell on the last day of trading for a given contract are automatically required to take or make delivery of the specified commodity. The timing and methods of delivery are clearly spelled out in each contract and in the exchange's rules. The exchanges' staffs match buyers and sellers and the matched companies then must meet their respective obligations. Exchanges have found it useful, however, to permit several variations of this simple process. Prior to the exchange-matching process (typically the day after the end of trading), any two market participants may agree to an exchange for physicals (EFP), transferring title to oil (or natural gas) by mutual agreement in lieu of closing out their position on the exchange. EFPs can also be used to establish future positions by mutual agreement. In fact, in the U.S. crude markets, this mechanism is widely used as a routine means of buying and selling crude, because it has the attraction of the exchange's financial performance guarantees. A recent innovation, introduced by NYMEX, is an Exchange of Futures for Swaps (EFS), which adds additional flexibility.

Once trading in a given contract has ended and participants are matched, the two matched companies may elect to use an alternative delivery procedure (ADP), which also allows the two to make alternative arrangements. In the case of both EFPs and ADPs, the exchanges are relieved of any responsibility for guaranteeing performance.

5.6 Regulation

Exchanges are self-regulating and (since 2000) for-profit corporations, owned by their members/shareholders. The degree of governmental oversight and regulation has traditionally been most extensive in the United States (by the Commodities Futures Trading Commission) and least intrusive in the United Kingdom.

The exchanges maintain active compliance and market surveillance programs to enforce trading rules and to detect any evidence of market manipulation. Traders caught violating rules are typically fined, and in relatively infrequent instances, barred from trading. If evidence of market manipulation is uncovered, exchanges possess a wide range of powers to remedy the situation. These powers include the right to order a given participant to reduce or even eliminate his or her position, to substitute alternative delivery points or additional

supplies (i.e., by broadening quality specifications), or even to impose a cash settlement in place of physical delivery (assuming the contract calls for such delivery). These powers are not often used, but their very existence serves as a powerful disincentive to would-be market manipulators.

Perhaps the most controversial aspect of futures trading (particularly in the United States) is the permitting of dual trading; that is, allowing the same individuals to trade for their own account while simultaneously executing orders for customers as a floor broker. Many critics have argued that this practice provides opportunities for floor brokers to jump ahead of large customer orders, profiting from the market movements that those large orders are likely to provoke. Although such actions are a clear violation of exchange rules, detection is not always easy. Exchanges counter with the argument that dual trading promotes liquidity and that exchange enforcement activities are sufficient to prevent serious abuses. As periodic prosecutions in both Chicago and New York have demonstrated, there will always be temptations. Clearly, exchanges can improve their rules and surveillance. At a minimum, exchanges that want to allow dual trading have an obligation to create clear audit trails so that violations are easier to eventually detect. It seems likely however, that dual trading will be prohibited in futures trading as it is in securities trading. But the exchanges and their floor communities can be expected to resist this development until the bitter end.

6. FUTURE DIRECTIONS

The world oil market is relatively mature, but will continue to evolve as the supply/demand balances shift. On the supply side, this implies an increasing dependence on the Middle East and the Caspian Basin, creating pressures for the evolution of new marker crudes to replace the gradually diminishing supplies of WTI and Brent. On the oil products side, the inevitable growth of demand in the developing world will also challenge the traditional role of Rotterdam and the U.S. Gulf of Mexico as the pricing reference points. This may in turn increase the importance of Singapore and eventually lead to a more favorable market for the evolution of futures and options in that city.

Natural gas is still an emerging market. The growth of liquid natural gas (LNG) and interregional pipeline trade in the world should provide new opportunities for futures markets. Electricity continues to be a challenge for the futures exchanges. As of 2003, the attempts to launch and sustain electricity futures markets had been hampered by a variety of obstacles, especially the Enron scandal. But it seems likely that the exchanges will continue to seek opportunities to serve the electricity trading market. In addition, price volatility seems unlikely to retreat, so the need for risk management instruments, both on and off exchanges, will remain a constant in these markets for decades to come.

7. SUMMARY

The oil markets led the way for the entire futures complex. This was no accident. As the largest and most global energy source, oil is the natural leader. In addition, government regulations were first removed from oil before deregulation hit the natural gas and power markets. Although the natural gas and power markets have distinct characteristics, there is no doubt that the experience of oil markets in deregulation provides an important indicator of future directions for all energy markets. However, the rapid growth of e-commerce in natural gas and power has propelled those markets into a period of rapid growth, temporarily leaving oil the laggard. The explosive growth of electronic commerce is rapidly closing the gap between physical trading and futures. Within the first years of the 21st century, the convergence of wet and paper markets will be completed as the old "analog"(open-outcry) physical markets evolve to "digital" (Web-based) exchanges. At some point in the not too distant future, risk management tools will simply be a mouse click or a pull-down menu away, as companies and traders make their deals and fulfill them on the net.

Among energy markets, natural gas and electricity have leapt ahead to embrace e-commerce, leaving the oil patch to catch up. Ironically, the early deregulation of most national oil markets and the far-flung nature of global oil markets brought innovation to those markets in the 1970s and 1980s. However, the relatively quick deregulation of natural gas and electricity markets, the high level of market inefficiency spawned by decades of government economic regulation, and the greater homogeneity of product quality have encouraged those markets to move more quickly into the Internet era. Electronic exchanges blossomed in the late 1990s, and energy was one of the largest categories of business-to-business e-commerce. Electronic commerce will have a growing impact on the energy industry, including reducing

costs and improving service all along the supply chain, enhancing existing businesses, especially in the area of customer service, transforming existing business models, especially through the creation of "infomediaries" (companies that provide and analyze market information for other participants), and creating new businesses.

By lowering entry costs and facilitating access, Web-based exchanges can expand the range of energy products traded. The process has already started in the chemical industry, and Web sites (e.g., Chemdex, ChemConnect, and e-Chemicals) have been launched to compete for business. There will, however, remain two challenges for these sites. First, markets will remain focused on the liquidity of new products. Traders have notoriously short attention spans and will "tune out" if sites do not offer sufficient volume to move reasonable quantities quickly. Second, as these sites grow in importance and their impact on pricing grows, they will inevitably come under regulatory scrutiny, at least by the United States Government and possibly by the European Commission and other jurisdictions as well. The regulation (and taxation) of Internet transactions will likely be a major political and economic battle. Third, as these markets grow, there will be significant pressure to develop new risk management instruments as integral offerings. The growth of electronic trading should lower the hurdle rates for new risk management instruments. However, the launching of new futures and options instruments will always depend on their ability to attract liquidity early—as mentioned, traders are notoriously impatient and will not support illiquid markets. It is probable that the future instruments will therefore include a modest expansion of formal futures and options markets and a continued rapid expansion of off-exchange derivatives, perhaps with more standardized settlement terms and, eventually, common clearing mechanisms. The exchanges have realized the opportunity and NYMEX has been especially innovative in trying to capture more of these traditionally off-exchange activities.

Ideally, the market would like to have a complete suite of energy markets serving both the Atlantic and Pacific basins, which would imply the development of additional futures and options markets for crude oil, products, gas, and electricity markets, especially in Asia. Obviously, the growth in interconnections of gas and power grids would tend to accelerate that trend, as would the continued expansion of international LNG trade. There seems little risk (or hope?) that volatility will soon depart the energy scene. And as long as there is volatility, there will be risk and a desire to manage that risk. However demand and production will shift, trade patterns will change, and product specifications will evolve, all requiring changes in contract design and continuous innovation.

Acknowledgments

Portions of this article originally appeared in *Energy Futures, Trading Opportunities* (John Elting Treat, ed.; PennWell Books, 2000). Excerpted by permission of the author.

SEE ALSO THE FOLLOWING ARTICLES

Business Cycles and Energy Prices • Economics of Energy Demand • Economics of Energy Supply • Energy Services Industry • Inflation and Energy Prices • Innovation and Energy Prices • Oil Price Volatility • Prices of Energy, History of • Stock Markets and Energy Prices

Further Reading

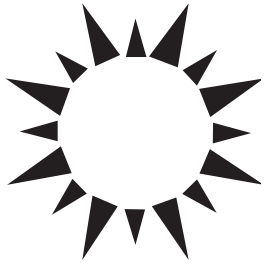
Errera, S., and Brown, S. L. (2002). "Fundamentals of Trading Energy Futures and Options." 2nd ed. PennWell Books, Tulsa, Oklahoma.

Eydeland, A., and Wolyniec, K. (2002). "Energy and Power Risk Management: New Developments in Modeling, Pricing and Hedging." John Wiley & Sons, New York.

Fusaro, P. C. (ed.). (1998). "Energy Risk Management: Hedging Strategies and Instruments for the International Energy Markets." McGraw-Hill, New York.

Fusaro, P. C., and Wilcox, J. (eds.). (2000). "Energy Derivatives: Trading Emerging Markets." Global Change Associates, Inc., New York.

Energy in the History and Philosophy of Science



ROBERT P. CREASE
Stony Brook University
Stony Brook, New York, United States

1. **The Puzzle of Energy**
2. **The Word Energy**
3. **The Concept of Energy**
4. **Energy Prior to 1800?**

Glossary

argumentative analogy An analogy in which the secondary term is used to bring to bear an already-organized set of equations or network of concepts on the primary term, so that the primary term becomes a mode of the secondary.

constructivism The view that that nature is how we represent it and that the real depends on intersubjective agreement.

context of discovery The manner in which a theory or datum was historically discovered.

context of justification The manner in which a theory or datum is justified by reference to all the available evidence on hand.

corporeal template The use of the first-person experiences of a sentiently felt body to forge or model new concepts.

energeia A term in Aristotle's metaphysics constructed out of ergon, meaning act or deed, used to refer to the activity of tending toward or enacting a goal (literally, en-act-ment).

filtrative analogy An analogy in which the secondary term calls attention to certain of its conventionally understood features in the primary term.

paradigm A framework that provides the conditions, criteria, and examples for explaining problems and phenomena of a standard type.

realism The view that nature has a fixed structure prior to the way human beings represent it.

vis viva Living force, later formulated as mv^2 and thus an ancestor of the modern notion of kinetic energy, that some early modern philosophers argued existed in addition to quantity of motion or dead force (formulated as mv).

One important and controversial issue in the history and philosophy of science concerns whether the phenomena named by concepts precede the development of the concepts themselves. While realists claim that nature has a fixed structure and only concepts have histories, constructivists claim that, inasmuch is nature is how we represent it, natural phenomena are themselves histories. The history of energy sheds some light on this issue, exposing the ambiguities that lead to each position, by highlighting the significance of the fact that the environment in which scientists work is not simply a human-environment relation but one mediated by technology.

1. THE PUZZLE OF ENERGY

Few things in the history of science raise more hackles than the issue of whether the phenomena named by concepts precede the development of the concepts themselves. The natural and commonsensical realist assumption, which it is fair to say is shared by most people and nearly all scientists, is that nature has a fixed structure, and that while scientific concepts have histories—they arise, are transformed, and occasionally pass away—the phenomena to which these concepts refer do not. Conceptual evolution merely testifies to the weakness of the science of the past, and to the success of contemporary science at obtaining a better grip on the world.

The opposite constructivist extreme is the view that nature is how we represent it, and the real depends on intersubjective agreement. Sociologist Bruno Latour, for instance, has provocatively argued that things, not just words, have histories; has claimed that microbes (and not just the concept of a microbe) did not exist before Pasteur; and has remarked that Pharaoh Ramses II could not have

 © 2004 Elsevier Inc. All rights reserved.

died of tuberculosis (as now thought) because the bacillus had no real existence before Robert Koch discovered it in 1882.

The history of energy is an intriguing episode in the history of science that sheds some light on this issue. It does so because the phenomenon of energy nicely illustrates the presence of factors such as the technological transformation of the world, and even the role of more metaphysical considerations such as the changing patterns of thought that sometimes have to occur for a concept to be taken as applying to nature.

2. THE WORD ENERGY

The word energy has a well-charted and uncontroversial history. Energy comes from the Greek energeia, or activity. For practical purposes one may say that the word was all but invented by Aristotle, who treats it as a noun form of a nonexistent verb. The word was constructed out of ergon, meaning act or deed; en-erg-eia might be literally translated as "en-act-ment." In his substance ontology, Aristotle uses it to refer to the activity of tending toward or enacting a goal, a meaning much different from today's. Aristotle saw bodies as composites of matter (hyle) and form (eidos). But for him, matter was only the possibility of a body, with no inherent structure and no perceptible or corporeal properties of its own—only a body can be soft or hard, hot or cold—and obtains a structure only via its form. Inextricably related to this distinction is another related to a body's activity. Every existing thing, he said, has an energeia, which maintains it in being and is the activity of tending toward or enacting a goal, its end or function (telos) to which it is naturally fitted. Aristotle has various names for how that activity is related to its end. A body's sheer potential or capacity for action he calls dynamis (potency), while its being at work en route to or at its telos he calls its en-ergeia. Every existing thing, then, has some activity or energeia, which maintains that thing in being. Aristotle seems only to have sensed, but not explicitly described, a difference between brute mechanical power and accumulated energy, or what we might call force and energy.

As philosophers such as Stephen Toulmin have emphasized, Aristotle's view was a product of the everyday situations with which Aristotle was specifically concerned. When he did concern himself with situations involving what we would call force and energy, they were mechanical horse-and-cart like situations in which a body (such as a horse) moves against a resistance (sources such as friction) to keep a body (cart) in motion. Aristotle compared other situations he was interested in to this type of example in formulating explanations of them. His analysis was therefore based on what Stephen Toulmin rather loosely (and prior to Thomas Kuhn's more elaborate and technical account) called a particular paradigm. It would not have occurred to Aristotle, Toulmin notes, to explain a familiar horse-and-cart situation with reference to "idealized, imaginary states-of-affairs that never in practice occur, ideals to which even the motions of the planets can only approximate." Unless we understand this context in which Aristotle was making and justifying explanations, we not only misunderstand him, but risk coming away thinking that he was guilty of elementary, unscientific blunders.

Aristotle's conception of energeia is, of course, worlds away from the modern conception of energy as the sheer capacity to do work. Energeia is not found in a specific place, but suffused everywhere and in different forms. It cannot be created and stored, converted or dissipated; it has a directionality toward forms to be realized and cannot be detached from these ends. The notion that energeia could be set up independently over and against the objects on which it does work, or stored apart from them, or collected or harnessed, would not have made sense to Aristotle.

Energy, the eventual English anglicization of energeia, came to lose the technical meaning Aristotle had given it and, for a while, virtually any value as a scientific term. In the mid-18th century, philosopher David Hume, for instance, complained that the words power, force, and energy are virtually synonymous, and "obscure and uncertain." As late as the seventh edition of the *Encyclopaedia Britannica* (1842), the word only merits a short entry: "ENERGY, a term of Greek origin, signifying the power, virtue, or efficacy of a thing. It is also used figuratively, to denote emphasis of speech."

3. THE CONCEPT OF ENERGY

How, then, did the word come to refer to a precise technical concept? The word would become a technical term only by being linked, at the beginning of the 19th century, with a concept that itself had undergone a complex evolution, and this evolution was the product of a number of factors.

One factor was a metaphorical extension of the subjective experience individuals have of themselves as a bodily center of action, which was then projected into nonhumans as one of their properties. This

illustrates what philosopher Maxine Sheets-Johnstone calls "the living body serv[ing] as a semantic template," a process which, she points out, is key to the emergence of many early scientific concepts. It is a classic case of the use of analogical thinking, or the use of the familiar to understand the unfamiliar. What is remarkable in this instance is that the familiar has its basis in the tactile-kinesthetic experiences of bodily life; thus, in a corporeal template.

Yale science historian Stanley Jackson, for instance, has written about how such a subjective experience was analogously extended into the concepts of early modern physics in the works of scientists such as Johannes Kepler and others. Kepler, like many scientists of his time, analogically projected a secular version of a soul-like animistic force into his mechanics. "If we substitute for the word 'soul' the word 'force' then we get just the principle which underlies my physics of the skies," Kepler wrote, adding that though he now rejected such souls, he had come "to the conclusion that this force must be something substantial—'substantial' not in the literal sense but... in the same manner as we say that light is something substantial, meaning by this an unsubstantial entity emanating from a substantial body."

What, then, was this unsubstantial entity? In the 17th century, this question sparked a furious metaphysical and scientific controversy on the existence, nature, and measure of force. René Descartes spoke of quantity of motion (mass times velocity, or mv), while Gottfried Leibniz argued that the force was not just a quantity but a quality of matter, called living force or *vis viva*, whose proper measure was mv^2. The *vis viva* controversy, as it was known, continued through the 18th century and was the subject of young Immanuel Kant's first essay. In 1807, Thomas Young, lecturing to the Royal Institution on "Collisions," said that "the term energy may be applied, with great propriety, to the product of the mass or weight of a body, into the square of the number expressing its velocity." He thereby tied the word, apparently for the first time, to its modern concept. But Young's "energy" was not ours, only what we call kinetic energy, and not even our formulation of it ($\frac{1}{2}$ mv^2).

The first half of the 18th century witnessed an intense discussion of the proper way to understand notions of activity, work, force, and *vis viva* among engineers and scientists including James Joule, Sadi Carnot, Hermann von Helmholtz, James Mayer, Emmanuel Clausius, William Thomson, and others. One of the most important motivations for this activity was to achieve a formulation adequate to the phenomena that scientists and engineers of the day were occupying themselves with: most notably, steam and heat engines, which were revolutionizing the engineering and industry of the day. In particular, scientists and engineers were concerned with the question of how to evaluate and measure the efficiency of steam and heat engines.

A second, closely allied factor involved what might be called a metaphysical shift, which opened up the possibility of seeing phenomena that look and act very differently as but different modes of one force. Thomas Kuhn emphasizes this point in an article on energy conservation, saying that such a metaphysical shift helps to explain why "many of the discoverers of energy conservation were deeply predisposed to see a single indestructible force at the root of all natural phenomena." The metaphysical shift in question itself may perhaps be specified more fully as the working out of another kind of analogical thinking, in this case, the use of an extended argumentative or scientific metaphor. Analogies and metaphors come in many different kinds, one common form being filtrative. In filtrative analogies or metaphors ("man is a wolf," "love is a rose") the secondary term (wolf, rose) calls attention to certain conventionally understood features of these things in the primary term (man, love). The focus is therefore on the primary term, and the secondary term only provides an assist. In argumentative or scientific analogies or metaphors, on the other hand, the priority of the two terms is reversed, and the secondary term is used to bring to bear an already-organized set of equations or network of concepts in order to understand the primary term—and the secondary term grows into the salient, technically correct term, often expanding its meaning in the process, with the primary term being merely one mode of the secondary. One classic illustration of an argumentative or scientific analogy occurred when Thomas Young and others began to call light (the primary term) a wave (the secondary term). Waves originally referred to a state of disturbance propagated from one set of particles in a medium to another set, as in sound and water waves. When Young and other early investigators called light a wave, they assumed that it, too, moved in some medium—but this proved erroneous, and scientists were soon saying that waves propagated in the absence of a medium. Despite this change, the same concepts and equations nevertheless governed light, which was now just another mode of an expanded conception of wave. Wave had grown into the technically correct term for characterizing light,

though in the process it became transformed and expanded (and the evolution would continue when waves turned up in quantum mechanics).

This kind of argumentative or scientific metaphor was at work in the beginning of the 19th century, in allowing engineers and scientists to see the presence of energy in quite different phenomena, from falling water and expanding air to heat and electricity. Only the kind of metaphorical sensitivity allowed the early investigators to realize the significance of the conversion processes that were so critical to the development of the concept of energy. It was doubtless key in allowing William Thomson, in a famous footnote (1849) to a work addressing a dispute between Joule and Carnot over conceptual and experimental problems in the heat formation of heat engines, to intimate that something involving both work and *vis viva* was conserved, though not yet visible "in the present state of science."

The specific changes in state that then made the conserved something visible are discussed at length in other entries. If any date marks the emergence of this new conception of energy into visibility, it is 1868, with the publication of the mechanics textbook, *A Treatise on Natural Philosophy*, by William Thomson and Peter Tait. The full articulation of this insight, involving the recognition that heat was energy—and only one of many forms—revolutionized science, and by the time of the ninth edition of the *Britannica* (1899), the entry on energy was six pages long and full of technical terms and equations.

The factors just mentioned are extremely significant for the history and philosophy of science, for they illustrate an inversion of the traditional conception of the relationship between theory and practice, as well as the fact that the environment in which scientists work is not simply a static human-environment relation, but a dynamic, ever-changing one mediated by technology. The practical coping with technologies can precede and even help give birth to theoretical reflection. Technologies therefore can become a kind of philosophical springboard for scientists: familiarity with them reshapes the theoretical imagination, allowing scientists to perceive and project new kinds of concepts or extend them to larger terrains.

4. ENERGY PRIOR TO 1800?

These developments help to shed light on the puzzle, mentioned at the outset, concerning whether the phenomena named by concepts precede the development of the concepts themselves and how the two vastly different positions mentioned can arise.

The puzzle is posed here by the question: Was there energy prior to 1800? This question would surely make most people, including and especially scientists, roll their eyes. Such people would interpret the question as asking whether the phenomenon named by the word existed. In this commonsense view, the answer is surely yes. Energy, after all, was not discovered or invented. It has powered the sun for billions of years, made organisms grow for millions, and driven industrial machines for hundreds. Our opinions, ideas, and representations of energy may have evolved, but energy apart from these—what we are trying to get at through them—does not. But if the question is interpreted as asking whether the abstract notion of a capacity to do work that comes in various forms convertible into each other existed prior to 1800, then the answer to the question is just as surely no.

Traditional philosophers of science would try to resolve the puzzle by appeal to the extremely influential distinction between the context of discovery, or the manner in which a theory or datum was historically discovered, and the context of justification, or the manner in which that theory or datum is justified by reference to all the available evidence on hand—and say that the convoluted, fortuitous, and often idiosyncratic process by which a discovery or concept enters science does not matter. There is some truth to this. But a term like energy cannot be justified in a bare environment, but only within a network or horizon of other historically given concepts (force, work, heat, etc.) and technologies (steam engines, instruments to measure temperature, etc.). These phenomena lead us to interpret other things in their light, changing our view of world and its findings, structures, and laws, leading us to project them as elements of the world—past, present, and future. This does not make scientific concepts arbitrary, an artifact of language or culture, but nor does it legitimate us thinking that these concepts refer to things apart from time and history.

The various interpretations of the question arise because the formulation of concepts such as energy is indebted both to a set of purely theoretical considerations and to a practical, technologically shaped world. This latter world is constantly changing, giving rise to new kinds of phenomena to be explained and related to phenomena that already appear. The network of theory and technologically shaped world forms a horizon in which scientific claims can be tested and be judged true or false.

If one emphasizes the values which permeate the horizon at the expense of the theoretical considerations, and the dependence on them that the invariance that phenomena such as energy show—if one emphasizes let us say the role of the development of steam and heat engines, the social and economic interest in creating more efficient types of such engines, and so forth; and if one emphasizes that a concept such as energy could not have emerged without such interests—it promotes a position such as Latour's.

On the other hand, if one emphasizes the theoretical considerations at the expense of the practical and technologically rich horizon thanks to which phenomena such as energy can show themselves as being invariant in myriad guises—if one emphasizes let us say the tightness and interdependence of the connection between concepts such as energy, power, work, and so forth—it promotes the view that science represents an ahistorical reality apart from a worldly context. But because the horizon is relatively stable, we tend to forget that it, too, can change, and that we justify things only in it.

So was there energy before 1800? If this means an abstractly defined concept of a quantity conserved in closed systems and related in precise quantitative ways to other concepts like power and work, no; if this means this is how we currently understand the world and its phenomena, yes. The history of the term energy is an episode in the history of science that helps to reveal the complexities of the history of science.

The scientific tradition as it has developed up to the present has shaped the contemporary notion of energy such that everything is a modality of energy. In cosmology, for instance, starting with radiant energy, some radiant energy becomes particles (thus mass energy), some particles become atoms, which in turn become molecules and condensed matter, and so forth. Energy is, as it were, the dynamic existential materia prima of the universe. This has both a theory side that gives to the present an illusion of wild nature that exceeds all representations and is beyond human creation, and a practical, representable, and controllable side that gives to the present an illusion of being merely socially constructed.

SEE ALSO THE FOLLOWING ARTICLES

Conservation of Energy Concept, History of • Early Industrial World, Energy Flow in • Economic Thought, History of Energy in • Forms and Measurement of Energy • Geographic Thought, History of Energy in • Thermodynamic Sciences, History of • Work, Power, and Energy.

Further Reading

Crease, R. (2002). What does energy really mean? *Physics World,* July.

Jackson, S. (1967). Subjective experiences and the concept of energy. *Perspect. Biol. Med.* **10,** 602–626.

Kuhn, T. (1977). Energy conservation as an example of simultaneous discovery. *In* "The Essential Tension: Selected Studies in Scientific Tradition and Change" (Thomas, Kuhn, Ed.), pp. 66–104. University of Chicago Press, Chicago.

Schönfeld, M. (2000). The *vis viva* debate. *In* "The Philosophy of the Young Kant" (M. Schönfeld, Ed.), pp. 17–35. Oxford University Press, New York.

Sheets-Johnstone, M. (1990). "The Roots of Thinking." Temple University Press, Philadelphia.

Toulmin, S. (1961). "Foresight and Understanding." Indiana University Press, Bloomington, IN.

Energy Ladder in Developing Nations

RICHARD H. HOSIER
United Nations Development Program (UNDP)
New York, New York, United States

1. Introduction
2. Conceptual Foundation
3. The Energy Transition: Drivers and Constraints
4. Empirical Evidence for the Energy Ladder Transition
5. Criticisms of the Concept
6. Conclusions

Glossary

commercial fuels Fuels for which every step of the supply chain involves a purchase or sale involving a monetary transaction.

dung Animal fecal matter—typically from cattle, buffalos, or camels—that is frequently burnt as fuel.

end use The final purposes to which energy is put. In the household sector, the end uses can also be called energy services and are typically considered to be cooking, lighting, and heating.

fuelwood Firewood, or woody biomass burnt as fuel.

interfuel substitution The substitution of one fuel for another in carrying out the same end use or fulfilling the same need.

logit analysis A statistical technique intended to predict the choice that would be made between a variable with only two possible values. Multinomial logit analysis is a multiequation version of logit analysis designed to predict the choice between a number of different options.

modernization The process by which a society becomes modern, current, or up to date. It involves economic development, industrialization, and urbanization among other forces.

noncommercial fuels Typically, traditional fuels that are not bought or sold in the marketplace.

Tobit analysis A statistical technique named for the economist James Tobin who developed it to factor into regression analyses dependent variables where a large portion of the sample being used had a value of zero (0) for the dependent variable.

traditional fuels Fuels that have been used by households since time immemorial (such as firewood or crop residues or dung). They are frequently gathered by individuals but are occasionally bought and sold in markets.

The energy ladder is a concept that has been used to describe the transition of households from utilizing traditional energy carriers for their energy service needs to utilizing more modern, technologically sophisticated energy carriers to meet those needs. It draws an analogy between household fuel choice and a ladder. Households using traditional fuels—such as firewood or dung—are assumed to be on the lower rungs of the ladder. Households using modern, commercial fuels—such as LPG, natural gas, or electricity—are assumed to be poised on the upper rungs of that ladder. Although this transition is of most concern to households in developing countries, a similar process is assumed to have taken place in developed countries as they went through the process of industrialization. The idea is that as a household unit's status changes—either through an increase in income, a shift in status, or a change in the society around it—it will begin to utilize a different mix of energy carriers to perform its energy needs. A household that chooses to begin utilizing kerosene or liquefied petroleum gas (LPG) for cooking instead of fuelwood is said to be moving "up" the energy ladder. In contrast, a household unit that makes the reverse decision—shifts from utilizing kerosene or LPG to using fuelwood—is said to make a move "down" the energy ladder.

1. INTRODUCTION

Since it became used commonly in the early 1980s, this analogue model has become a tool to studying

 © 2004 Elsevier Inc. All rights reserved.

the energy-use patterns prevalent in developing countries. Researchers and policy analysts have used the concept to better understand household fuel choice and to better design energy policies by anticipating their impacts on energy-using households. Typically, poorer households and rural households have limited access to energy resources and fuels and occupy lower rungs on the energy ladder. Wealthier households and urban households generally have greater access to energy resources and occupy higher rungs on the energy ladder.

The same distinctions that have been made initially at the household level can also be applied at the level of a region or a nation. In fact, underlying much of the conceptual background on the energy ladder at the household level is the understanding that as societies as a whole develop, they not only begin to utilize a greater quantity of energy (as measured in calories, toe, or joules) but their energy mix consists of a greater percentage of modernized fuels, such as petroleum, natural gas, and electricity. While this direct analogue of the individual household unit to the national level may introduce certain fallacies of composition, as individual energy consumption patterns sum up to provide a national energy consumption pattern, the individual-national analogue is quite reasonable: national energy use patterns do reflect, *inter alia*, the sum total of the individual energy-use patterns of their citizens.

As most research on the energy ladder in developing countries has focused on either confirming the existence of the household-level energy transition or specifying its form in various contexts, the household level discussion will form the focus for most of this article. However, when the discussion reverts to the use of cross-national comparisons and analyses to generalize using national-level data, the jump from the household or agent level to the national level will be clearly delineated.

2. CONCEPTUAL FOUNDATION

The energy ladder concept is built on the idea that every fuel that can be used by households has its own set of advantages and disadvantages depending on the various end uses in which it is deployed. Modern fuels are considered to have more advantages than traditional fuels and are therefore considered to be higher rungs on the ladder. In developing countries, the primary end uses or energy services that households need to fulfil are cooking, space heating, water heating, and lighting. Household decision making then must consider the qualities of all the available energy sources in carrying out these end uses in order to arrive at the final household decision regarding which fuels to use for which purposes. In many instances, the decisions are quite simple as the opportunity set facing the households is limited. In other instances—notably the peri-urban and urban areas of developing countries—households face a large range of options that make the household energy use decisions quite complex.

Leach argued that our understanding of this energy transition is based on energy surveys in developing countries that confirm "that most families have an ideal fuel preference 'ladder' running from biomass fuels up through kerosene to bottled gas (LPG), natural gas where it is available and electricity" (p. 118). Traditional fuels, such as firewood and dung, are relatively inconvenient to collect and store, require constant management while in use, and emit large quantities of smoke that has a detrimental impact on the health of the users. Modern fuels, such as natural gas and electricity are more convenient, cleaner to use, have power outputs that are easily controlled, and can be delivered directly to the kitchen (once the basic transmission and distribution infrastructure is built). The energy ladder or energy transition embodies the idea that as household income rises through the development process, they choose to utilize modern fuels and choose not to utilize the traditional fuels.

Each household makes energy-use decisions based on its unique circumstances, its preferences, and its environment. When aggregated across households, the resulting observed energy consumption patterns can become quite complex. In general, lower-income groups within a given area tend to rely more heavily on traditional biomass fuels for cooking. If they are able to afford lighting services at all, they will typically use kerosene or candles for lighting. Wealthier groups within that same rural area may also rely on traditional biomass fuels for cooking, but may supplement that biomass with kerosene or liquefied petroleum gas (LPG). For lighting needs, wealthier households may rely on electricity if their household is connected to the grid, but if not, they will also utilize kerosene, LPG, candles, or even battery-powered flashlights or torches. Thus, poorer and wealthier households in both urban and rural areas of developing countries demonstrate different energy use patterns resulting from their different decision made with respect to their energy consumption patterns.

Figure 1 illustrates diagrammatically the range of end uses and fuels faced by households at different

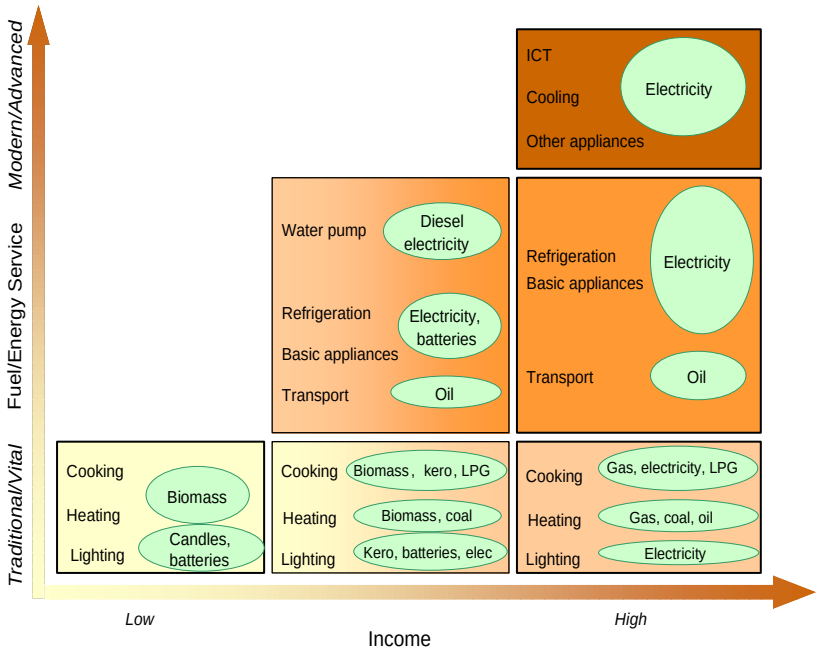


FIGURE 1 End uses and fuels faced by households at different income levels. From World Energy Outlook (2002), International Energy Agency (IEA). Used with permission.

income levels in developing countries. The horizontal axis represents income and the vertical axis represents the fuel used to fulfill an energy service need, moving from the most traditional to the most modern. As represented in this diagram, all households at all economic status levels must utilize energy for cooking, heating and lighting. In poorer households, the options normally used for these purposes are traditional biomass for cooking, and either candles, kerosene, or batteries for lighting. In middle-income households, the options to fulfill these services expand to include biomass, kerosene, and LPG for cooking; biomass, coal, or charcoal for heating; and kerosene, batteries, and candles for lighting. However, at middle-income levels, another set of end uses or energy services begin to be considered. These additional end uses include refrigeration, transport, basic appliances such as radio and television for entertainment, and even water pumping. The fuel options available for providing these services are petroleum-based fuels, batteries, and electricity. Finally, for upper-income households, the options for cooking extend to natural gas, LPG, and electricity, and the options for heating may include gas, coal, oil, or electricity. At this higher income level, electricity will provide in all likelihood provide all of the lighting needs. At the highest income level, electricity will likely provide all of the refrigeration and appliance needs while petroleum-fuels continue to fill the transport needs. In addition, the higher-income households begin to pursue a range of additional energy needs, such as air conditioning, for which the only option is electricity.

Thus, as household incomes increase, the household tends to move "up" the energy ladder to more sophisticated fuels. Not only does the choice of fuels change, but the absolute quantity of energy effectively utilized by the household increases as it begins to pursue a broader range of energy end uses or energy services. For nations as a whole, as the economy develops and the population becomes wealthier, its citizens begin to use a larger share of modern fuels, especially electricity, to provide a larger number of energy services. This results generally in a larger consumption of energy at the national level—hence, the well-documented traditional link between energy consumption and gross national product.

3. THE ENERGY TRANSITION: DRIVERS AND CONSTRAINTS

The resource allocation decisions resulting in a household's fuel-use patterns and, therefore its position on the energy ladder, are determined not just by income but also by a large number of other

factors. Because these factors influence the household's decision-making process, they also are considered to be important either as policy levers or as indicators of policy impact. In the following section, the factors driving or constraining the household's energy-use decisions are briefly discussed with respect to both their impact on the energy ladder decision and their importance as policy measures or as indicators of likely policy success. Some of these may be considered to be supply considerations; others may be characteristics of the fuels; and still others may have to do with the either the physical or economic environment faced by the household in a particular market.

3.1 Household Income

As has already been discussed, household disposable income is one of the most important factors determining household energy decisions in all countries, both developing and developed alike. In general, total energy use is positively correlated with income. In the context of the energy ladder discussion, wealthier households will show a greater predilection to using more sophisticated fuels than will poorer households. Expressed in microeconomic terms, the income elasticity of demand for most modern fuels (electricity, natural gas, LPG) is positive whereas for traditional fuels (over a wide range of incomes) it tends to be negative.

Income becomes an important policy discussion concerning household energy use. As incomes increase, so will the demand for modern fuels, other things being equal. Likewise, households with falling incomes are liable to move down the energy ladder to more traditional fuels. Perhaps a bigger concern for policymakers in developing countries revolves around the energy needs of the extremely poor household—that is, households living on less than $2 per day. Whether or not the incomes and other resources available to these households will enable them to obtain adequate energy remains a concern for energy policymakers. Unfortunately, in such cases the policy options may be so limited that policymakers are unable to influence the energy mix used by these extremely poor households.

3.2 Urbanization

Urbanization is an important indirect influence on a household's energy-use decisions. Households living in urban areas face a different opportunity set than those living in rural areas in two major respects. First, their ability to collect firewood or traditional fuels from their surrounding will tend to be more limited than that of their rural counterparts. Second, because their ability to engage in subsistence agriculture is more limited than that of their rural counterparts, they require a steady cash income from paid employment to be able to meet their budgetary needs. As a result, their monetary income tends to be higher than that of rural households and they will demonstrate a higher reliance on commercial or semicommercial fuels than their rural counterparts. In addition, urban infrastructure is normally more advanced than rural infrastructure meaning that households in urban areas tend to have better access to the electric or natural gas distribution grids as well as to markets for LPG bottles, kerosene, and even commercialized traditional fuels, such as charcoal. Therefore, an urban household will tend to occupy a higher rung on the energy ladder or use a larger share of modern, commercialized energy sources to meet its energy needs than will a rural household of comparable status.

Urbanization is an important consideration when discussing household energy policy and situation with respect to the energy transition. In most developing countries, more policy options have relevance to urban areas because all urban households are engaged in market activities for their livelihoods, supply-focused options are easier to design and implement, and price controls or subsidies are easier to enforce. For example, electrification of urban households or the extension of a natural gas distribution grid to urban households tends to be a far more cost-effective policy option to enhance household access to modern sources of energy sources than is applying those same policies to rural populations. In fact, given the low population densities of most rural areas in developed countries, grid extension as a policy intervention is often too expensive and cost-ineffective a policy option to consider pursuing at all.

3.3 Smoke Emissions and Exposure

One of the reasons that households in developing countries are presumed to prefer modern fuels over traditional ones is that the former are cleaner, in terms of both storage and handling and emissions. Traditional fuels, such as dung, crop residues, and fuelwood, emit noxious fumes when burned contributing respirable particulates to the household environment. Many households in developing countries burn these fuels for cooking indoors and often have no effective

means of venting the emissions from these fuels. This leaves the cooks—typically women—and other household members exposed to extremely high levels of particulate concentrations. Smith has documented that the rural indoor environment in many developing countries may represent the world's single most pervasive and intensive air pollution problem in terms of actual human exposure to particulates. In terms of total human exposure to pollutants, exposure to biomass (or coal) smoke from indoor cooking in developing countries poses a far more serious health threat than second-hand tobacco smoke or any source of ambient industrial emissions. Evidence from developing country households shows that concentrations of respirable particulates can run from 500 to more than 15,000 $\mu g/m^3$ in many indoor environments in developing countries. This level of pollution concentration from the indoor use poses very serious health threats to the exposed population. Although the medical records linking exposure to disease are rather weak in most developing countries, acute respiratory infections in children and chronic obstructive lung disease and *cor pulmonale* in adults exposed to smoke from biomass cooking have all been linked to exposure to cooking smoke in developing countries. Although other anticipated health effects from exposure to biomass cooking smoke—such as adverse pregnancy complications and increased cancer rates—have not yet been sufficiently documented through medical research, they must be assumed to be significant.

Smoke pollution clearly provides a significant incentive for households to move away from using these dirty fuels—such as firewood and coal—toward using more modern, cleaner fuels with fewer negative health impacts. Although many rural households and poorer urban households in developing countries may have no alternative to burning biomass fuels or coal in unventilated cooking devices, for those who do have some alternative, it provides a significant incentive to move "up" the energy ladder to cleaner, more modern fuels. For policymakers, the likely negative health impacts of continued reliance on these traditional fuels should create an incentive to find cooking alternatives that do not threaten the health of women and children in developing countries. One logical policy response might be to devise programs to encourage the use of LPG, natural gas, or electricity for cooking in areas where they are available and affordable to the local population. Government policies to support moves "up" the energy ladder in most developing countries could certainly be justified on the basis of the negative health impacts of continued traditional fuel use.

3.4 Appliance Costs

One of the obstacles to adopting modern fuels is the fact that a specialized piece of equipment—an appliance—is required to utilize them. The cost of these appliances to use more modern fuels may prove to be prohibitive for many poorer households in developing countries. Such costs serve as a transaction barrier to the adoption of more modern fuels by households in developing countries. The cost of LPG bottles, kerosene cooking stoves, LPG stoves, or electrical stoves may represent a major cost obstacle to households adopting and utilizing the more modernized fuels instead of the traditional biomass-based fuels found on the lower rungs of the energy ladder. For many households accustomed to cooking with wood over a traditional three-stone hearth, the additional cost of a fuel-efficient woodstove may prove prohibitive, even in cases where the fuel savings more than justify the additional costs.

Policies may be devised to assist some households in overcoming the financial barrier to purchasing appliances for modern fuels. While financing might help households amortize appliance cost over a longer period of time, such payments will still increase the costs of the more modern fuels, placing it at a disadvantage to the traditional fuel for which no improved appliance is required. Some governments have tried subsidizing or providing tax credits for the purchase of appliances designed to utilize modern fuels, thereby encouraging households to adopt and use the more modern fuels. Several African countries claim to have successfully encouraged a move away from fuelwood and charcoal by subsidizing LPG bottles and cooking devices while encouraging suppliers to provide better distribution networks for this relatively clean and modern fuel. These represent some of the few cases of desirable moves up the energy ladder that have been stimulated by deliberate government policy.

3.5 Relative Fuel Costs and Prices

Relative prices per delivered energy unit play a crucial role in determining what fuels a household will use as well as how much of those fuels to use to meet its own energy needs. In the case of traditional fuels, such as fuelwood, the costs may be measured in terms of the time spent gathering the fuel rather than in monetary units. How a household may choose to make this trade-off—between labor and monetary resources—involves a complex calculation made more difficult because of the different roles found

within the household unit: women may contribute the labor to gather the firewood, but men often control the money required to purchase the fuels. In general, the relative fuel costs and prices will be reflected in the household energy use patterns and will determine where on the energy ladder a household determines itself.

Policymakers have frequently utilized price to steer household energy use decisions. The most common case involves subsidized electricity prices: governments have frequently regulated electricity prices to be less than their actual costs in order to encourage its use and obtain favor with the electricity-consuming populace. By subsidizing connection fees and controlling energy charges at levels below their actual costs, governments encourage households to utilize more electricity than is economically sustainable. The net result is that more households move up the energy ladder than should, consuming more electricity than they can actually afford to purchase. In such cases, the electricity utility's balance sheet suffers.

A second example of governments using prices would be the kerosene subsidy. Many governments have sought to control by regulatory fiat the price at which kerosene can be sold in the market, either through price or controls a cross-subsidy provided from a surcharge on other petroleum fuels. While the goal of such a policy is admirable as it is intended to ensure that poorer households have access to a relatively modern and clean fuel—kerosene—for lighting and cooking, it also has proven to be problematic. Studies have shown that most of the benefits of these subsidies accrue to middle and upper-income families, not to the poor. In addition, the surcharge may be administratively cumbersome and the controlled price quickly becomes unsustainable. The costs of such policies frequently outweigh the benefits and governments should be aware of this prior to implementing such systems.

A strikingly different example of using pricing policy to shape household fuel-use decisions is the case of the stumpage fee imposed on trees being harvested for fuel use. Many governments in Africa have been convinced that there is inadequate control over the harvesting of firewood for sale to households and for the production of charcoal for sale to peri-urban and urban households. The proper response in such cases is to impose a tax or a stumpage fee to compensate for the fact that the harvester has not paid for the planting, protection, and caring for the tree over its lifetime. If a government imposes such a stumpage fee either at the point of harvest, the point of production, or the point of importation into an urban area, it has two relevant effects. First, it raises funds to be able to replace the tree. Second, it serves to raise the price of the fuel creating an incentive for the household to either utilize less wood or charcoal or to find another alternative, one either lower or higher on the energy ladder. Thus, pricing policies can stimulate movements either up or down the energy ladder. Such movements may be intentional or unintentional, depending on the situation.

3.6 Fuel Availability

Perhaps the single most important supply-side determinant of household fuel choice for traditional fuels has to do with fuel availability. While the availability of fuel sounds like a relatively straightforward matter to determine, it actually becomes quite a complex matter to evaluate in any given location. It may be determined by the forest cover, the predominant land-use system, the nature of land tenure system, cropping patterns, and the distance traveled to collect fuel. For rural households accustomed to using firewood, even convenient surrogates such as the distance traveled to collect firewood or the amount of time required to collect a headload may be a misleading measure of fuel availability. While much of the literature on the household energy use in developing countries originally emerged in response to concerns regarding increasing deforestation and the "fuelwood crisis," scientific work has delinked fuelwood use by rural households from deforestation. This does not mean, on the one hand, that in localized areas, households do not experience a severe shortage of fuelwood for domestic purposes or, on the other hand, that fuelwood use has never contributed to tree harvesting that is detrimental to the environment. Rather it means that energy and forest planners must maintain balance in their view and proposed plans for rural resource management and needs. Fuelwood availability, or rather fuelwood's limited availability in certain rural areas of developing countries, may create an incentive for households to choose to utilize fuels either higher (i.e., kerosene, LPG) or lower (crop residues, dung) on the energy ladder.

For those fuels that are lower on the energy ladder than fuelwood, such as crop residues or dung, they present their own challenges of availability. Crop residues are normally only available to rural households at certain post-harvest seasons during the year. As they are frequently used for fodder in rural areas,

their use for cooking may detract from the ability to raise livestock. Dung availability is dependent on livestock ownership in a given area and is easier to collect when livestock are penned in a small corral instead of being allowed to wander freely. As is the case with crop residues, dung has an alternative use as fertilizer, and using significant quantities of dung for cooking fuel will eventually contribute to soil exhaustion.

In many rural areas, modern fuels may simply be unavailable. While kerosene is fairly easy to distribute through formal and informal channels, LPG bottles are often viewed as cumbersome to manage meaning that LPG is not always available for household use. A decision to use electricity is entirely dependent on grid connections and the availability of power. While the availability of traditional fuels may be more a question of ecological status, the availability of modern fuels will be determined by the state of market development.

The question of fuel availability not only influences household fuel-use decisions in rural areas but also helps determine the markets for household fuels in urban and peri-urban areas. In many African countries, urban households rely on charcoal for the domestic cooking needs as firewood is relatively bulky and economically inefficient to transport and charcoal stoves are frequently more efficient than firewood stoves. An added advantage is that while charcoal combustion emits carbon monoxide, it does not emit the same type of noxious, respirable, particulate-laden smoke that firewood does. Hence, urban households using charcoal avoid much of the dangerous ground-level emissions from biomass smoke to which their rural counterparts using firewood are exposed.

In reviewing the question of availability of biomass fuels and its relevance to household fuel choice decisions, Barnes and Floor developed a transition model encompassing three stages that cities pass through in this transition. Stage 1 represents situations where the urban dwellers still rely heavily on wood use—encompassing situations where wood may still be free to the case where it is purchased and its price is fairly high. Stage 2 differs depending on local resource availability. In some parts of Africa (e.g., Senegal, Kenya, Tanzania) and the Caribbean (e.g., Haiti, Dominican Republic, Jamaica), the second stage encompasses heavy dependence on charcoal. For other cases (e.g., Indonesia and China), the second stage involves heavy reliance on coal and kerosene. The final stage, Stage 3, occurs when income in the city is fairly high and virtually all households are dependent on LPG, natural gas, or electricity. Virtually all cities in developed countries fit into this third category, and increasingly so do many cities in middle-income developing countries. The progression of urban areas through this urban energy transition corresponds well to the movements of individual households in those cities up and down the energy ladder.

On the policy front, for rural households, much effort has focused on supporting the planting of trees either in woodlots, on farmland, or in forest reserves. Policy efforts have also focused on ensuring proper ownership of and access to trees on public and private lands to guarantee fuel availability. While these options are intended to ensure continued sustainability of traditional woody biomass for use as a cooking fuel, there is a need to maintain balance between the use of trees for fuel and their use for nonenergy purposes. Trees have many more functions in the rural ecology than just supplying cooking fuel needs.

Other policy efforts in both rural and urban areas have sought to guarantee that households have access to modern fuels. Governments have lent support to the development of marketing and distribution networks for LPG and kerosene to ensure that households have an expanded array of modern fuel options. Rural electrification and urban electrification (i.e., connecting previously unconnected urban households to the electric grid) are both policies designed to provide households with access to electricity, the fuel typically at the top of the energy ladder. Although results demonstrate that households newly connected to the grid own neither appliances nor the wherewithal to use electricity for all of their domestic purposes, gradual accumulation and consolidation of household status is presumed to expand future household use of electricity.

4. EMPIRICAL EVIDENCE FOR THE ENERGY LADDER TRANSITION

The concept of the energy ladder has been used as a foundation for many of the studies of household energy use in developing countries that have been undertaken since the early 1980s. While not all of them have provided confirmation for the linear transformation envisioned in the narrowest interpretation of the energy ladder concept, most of them do demonstrate that households undergoing the development process do shift their energy use

patterns in response to both internal and external conditions. The following sections focus on small fraction of the empirical literature on household energy use and its documentation of ladder-like household energy transitions. Section A discusses the results from various surveys that have been implemented, and Section B examines some of the statistical modeling that has been undertaken to model household energy behavior.

4.1 Survey Evidence

Household surveys undertaken in rural Africa have demonstrated that fuelwood remains the predominate fuel for rural household energy use throughout the continent. However, the determinants of fuel switching or movements up or down the energy ladder may be somewhat difficult to predict in the rural context. A rural household survey undertaken in Kenya in 1981 shows that fuelwood is by far the single most important fuel used by households for cooking. When it was unavailable, households would be forced to switch either to crop residues or kerosene. Fuelwood availability is an extremely localized phenomenon and may be linked to diverse factors, such as land tenure, population density, settlement density, and crop types. On the national level, income may prove less useful as a predictor of household energy use patterns than might be other factors such as level of engagement with the cash economy. Subsistence farmers showed a relatively heavier reliance on firewood and a lesser consumption of kerosene, LPG, and electricity than did cash-crop farmers or even farm workers or schoolteachers. As most of rural Kenya remains unelectrified, the final rung in the energy ladder was occupied by a very small fraction of households. Rural household energy use patterns remain somewhat problematic and difficult to predict given the wide range of other confounding variables found to influence household fuel choice.

A survey undertaken by Davis in several South African rural areas again confirms that fuelwood is the predominant fuel of choice for all households at all income levels. There is a tendency to shift from collecting firewood to purchasing it as incomes increase, but firewood remains relatively constant as the most important cooking fuel in rural areas. Kerosene again emerges as a transitional fuel for cooking. Candles, kerosene, and electricity are the predominant fuels used for lighting in rural areas. For households connected to the grid, there is a decline in expenditures on kerosene and candles for lighting. But there is also a tendency for the lowest income households to continue purchasing and using candles even after they have grid electricity. This is attributable either to limited indoor wiring or to limited cash-flows prohibiting the purchase of cards for pre-payment meters. What does emerge from the South African evidence is the relatively large number of households that use multiple fuels for a single end use. For example, 25% of the electrified households and 56% of the unelectrified ones used both kerosene and fuelwood for cooking. In short, the authors conclude that a household energy transition from traditional to modern fuels is taking place in rural South Africa. However, they express reservations about whether it will result in a simple transition to electricity use or an increase in the use of multiple fuels for a single end use.

In urban areas across Africa, the situation appears to be more predictable. Hosier and Kipondya present data from urban Tanzania showing that charcoal is the predominant fuel used in urban households across all income groups. Although this predominance of charcoal may vary according to the city, its role varies little per income category. Kerosene consumption demonstrates the role of a transitional fuel whose use rises with small increases in income and then decreases at higher income levels. Electricity use increases across all income levels. The urban household energy transition appears to be proceeding slowly in Tanzania, despite the country's rather sluggish economic performance during the period under examination. Policy efforts—especially cross-product subsidies to kerosene and artificially low electricity prices—might have played an important role in ensuring that more urban households were able to move up the energy ladder than down during the period of examination.

A detailed ESMAP-sponsored survey of rural household energy use in six counties of China demonstrated that most households used fuelwood, agricultural residues, or coal for their domestic cooking needs. Households in the poorer counties tended to show greater reliance on firewood while those in the more industrialized counties used very little firewood and more coal. Coal was a major household fuel in four of the six counties, especially where fuelwood was scarce. In other locations where fuelwood was scarce, straw and crop waste were used. Electricity use was again positively related to income. On balance, income is seen to play an important role in driving households to use electricity, but fuelwood shortages—frequently very locally determined—tend to drive households either up or down the energy ladder to other fuels.

From an ESMAP-sponsored analysis of surveys of household energy use in six states of rural India, the evidence shows that there is a slow transition to modern fuels brought on both by increasing economic status of households and by an increasing scarcity of fuelwood. In a cross-sectoral analysis, higher-income households clearly used more electricity and more LPG than did lower-income households. Upper-income households also spent a larger absolute amount of money on their household energy needs, but due to the higher incomes, this represented a smaller fraction of their total income. While a higher fraction of upper-income rural households also utilized electricity for lighting and other purposes, the utilization of biomass energy appeared to decline only at the highest income levels in the rural areas. However, the time devoted to collecting fuelwood in rural areas declined with income as wealthier households would tend to purchase firewood instead of gathering it. From 1980 to 1996, over 8% of rural households acknowledged switching from using firewood for their domestic cooking chores to using some other fuels. Over 60% of these households moved to kerosene and LPG for cooking. The remaining households acknowledged fuel switches either to biogas or down the energy ladder to crop residues or dung. In both urban and rural areas, kerosene is seen to fulfill the role of "first step" up the energy ladder as many households begin using it to cook with prior to moving on to LPG or other modern energy carriers. In urban India, LPG has become the fuel of choice for cooking and has begun to be widely utilized. This has been attributable not just to an increased supply of the fuel, but also to the breaking up of the near-monopoly on LPG distribution maintained by the national petroleum companies. In general, these data from India show that there is a movement among both rural and urban households to more sophisticated modern fuels from traditional, biomass fuels.

In a review of energy use surveys in urban areas in Asia, Sathaye and Tyler presented household energy consumption patterns from cities in India, China, Thailand, the Phillipines, and Hong Kong. While each country and indeed each city within a country demonstrated very different energy use patterns, their findings showed several tendencies. First, as incomes increased, households tended to rely less on biomass and more on modern fuels. Somewhat surprisingly, modern fuels may have been preferred over traditional biomass fuels not only because of their greater convenience but also because of they are cheaper when measured in terms of their cost per unit of energy delivered. Second, as incomes increased, households not only made use of a larger quantity of electricity, but they also made use of electricity for a larger set of end uses. Urban households at all levels of income began purchasing and using electrical appliances, from light bulbs to refrigerators to color television sets. This relatively rapid dissemination of appliances throughout urban Asia would indicate that household electricity demand is likely to grow in the same rapid way as it has in the more developed, Organization for Economic Cooperation and Development (OECD) countries. It also poses an opportunity for electricity conservation: if these appliances can be made more efficient, then household standards of living can continue to rise with a less significant increase in electrical requirements. Finally, kerosene use—which is considered to be an inferior modern fuel—declined in importance through time and as incomes increased. While not confirming the existence of a rigid, linear, and universal energy ladder, the evidence presented from urban Asia confirmed that households do move to more modern fuels across time and as income allows.

4.2 Statistical and Econometric Evidence

Given the complexity of household fuel-use patterns and the large number of factors influencing household fuel choice, tables of descriptive statistics do not go very far in understanding household fuel choice. To push beyond the understanding available from using descriptive statistics, a number of authors have used regression-based and econometric analyses to analysis household fuel choice in developing countries. These analyses are important as they provide insights both into the factors that appear to determine household fuel-choice decisions and into the validity of the energy ladder transition itself.

From the analysis of a number of household energy surveys based in south Asia, Leach developed multiple regression equations designed to explain household energy utilization. Because household size is closely related to income, he chose to use per capita household energy consumption figures as the dependent variable. His analysis concluded that income, household size, and settlement size are the key factors determining household fuel use. Although the elasticities that he estimates demonstrate the expected sign in most cases, he argued that they are not much help to energy planners as the important independent variables are not under the planners' control. Other factors that are under more direct

control of energy policymakers—such as relative fuel prices, the cost of appliances, and the availability of commercial fuels—are still not shown to be crucial determinants of household fuel-use patterns, partly because of the inadequacy of the survey techniques deployed in gathering the data.

Using nationwide household energy data from Zimbabwe, Hosier and Dowd developed a model of household fuel choice. In the aggregate, their data showed that fuelwood and kerosene use decreased with income while electricity use increased with income. They then formulated a multinomial logistic model of fuel choice, where the dependent variable represented the probability of a household choosing one fuel over another to cook with. Their data made a distinction between gathered wood and purchased wood as being distinct alternatives. They also allowed for a separate "transitional" category for households that used both wood and kerosene or electricity. Their system of four equations was able to predict nearly 90% of the household fuel choices correctly. Their results demonstrated that an increase in income and localization in an urban area will both increase the probability that a household will make a move up the energy ladder. They conclude that while energy policies might have an influence on household fuel decisions in the urban areas, fuel decisions in the rural areas appear to be insensitive to readily manipulated policy instruments.

Using urban household energy survey data from Java, Indonesia, Fitzgerald, Barnes, and McGranahan undertook two distinct sets of econometric estimations for cooking and lighting. For the analysis of cooking fuels, they first estimated the demand for energy for cooking from the three fuels used (wood, kerosene, LPG) for cooking in Java using a Tobit analysis to account for the fact that many households did not use at least one of the fuels for cooking. The analysis included household size, the prices of all fuels, the availability of all fuels, urban size, and an index measure of cooking size to estimate the amount of each fuel used for cooking. Their results show that firewood tends to be treated as an inferior good: as household income rises, fuelwood use tends to decline. However, wherever fuelwood is plentiful, all households—even wealthier ones—tend to use it for cooking. The results also show that LPG tends to be the fuel of choice for the highest income households surveyed, but again, its availability limits its distribution. Kerosene again serves as a bridging fuel—it is used sparingly by lower and upper-income households, but fairly heavily among the middle-income groups. Their second estimation examined the demand for energy (measured in MJ) in all forms attempting to isolate the substitution effects between the various fuels. These estimates show that actual substitution effects are slightly different than those derived from technical evaluations of fuels and efficiencies. For lighting, the authors repeated similar analyses of lighting demand and total residential fuel demand. The fuel demand equations demonstrate that electricity is the fuel of choice for all households in urban Java. The lighting substitution analysis demonstrates that a household using kerosene for lighting will have to utilize 10 times as much energy (heating value) for lighting as a similar household lighting with electricity. In addition, a household switching from kerosene to electricity can easily cut its overall lighting expenditures in half while obtaining nearly six times the amount of lighting. Overall, the analysis shows that urban residents of Java show household fuel preferences that resemble those that would be predicted by an energy ladder, with the exception being perhaps that firewood is used across all income groups if it is locally plentiful.

For the cities of Dakar, Senegal, and Abidjan, Cote d' Ivoire, Koumoin studied household energy demand using cross-sectional, time-series data for the years 1978 to 1992 taken from the Food and Agriculture Organization (FAO) production yearbook, the electric utilities in the countries, and a set of World Bank studies of the region. A system of eight equations representing the demand for energy in the urban households of the two cities was estimated in the form of a generalized logit model. His results show that the demand for charcoal is relatively price and income inelastic, meaning that it is used by households at all income levels. Rather than seeing immediate interfuel substitution as might be expected on the basis of the energy ladder concept, the evidence seems to support a gradual transition process with multifuel utilization at the household level being an initial step toward the anticipated substitution. Because of the relative price inelasticity of charcoal, electricity demand may be that sector of the overall household energy consumption which is most sensitive to pricing policies and variations. Local considerations, such as household size, migrant labor, and fuel availability, also seem to be important in shaping the profile of household energy demand. The importance of fuel availability can be seen in that LPG demonstrates a higher than expected penetration because of long-standing policy efforts to promote it as a cooking fuel. In such an economically marginal economic environment, the author concludes that policy efforts might well focus

on ensuring stability in the market for traditional fuels so as not to further threaten the economic base of the population. Interfuel substitution and movements up the energy ladder will likely take place over an extended period of multiple-fuel use given the nature of the prevalent economic and environmental conditions in the cities under examination.

Using a 4-year long study of a village in Mexico supplemented by a household energy survey in four Mexican states, Masera *et al*. discovered that rather than households moving in a straightforward, linear fashion along the energy ladder, in fact, their transition may be from a simple, traditional fuel use to a situation where multiple fuels are used simultaneously. Wealthier households continued to cook with fuelwood indoors even when they owned and also used LPG cooking devices. Multiple fuel use appears to be pursued for two reasons. First, it is a risk-minimization strategy so that if the LPG supply fails, the household will face only minor disruption as they are still accustomed to cooking with firewood. Second, cooking with both wood and LPG is an accepted practice because of preferences for a diet of tortillas—the traditional staple food—which cannot be cooked efficiently on LPG appliances. This cooking pattern has a number of implications. First, households that adopt the more modern fuel, LPG, rarely cease to cook with wood. As a result, their firewood consumption levels drop by only 30% once they begin cooking with LPG. Second, these multifuel using households typically are wealthier members of the community, but they willingly face exposure to nearly twice as many particulates as lower-income households. Fuel stacking, the process by which households continues to use firewood (a traditional fuel) while also adopting LPG (a modern fuel), is seen as a logical resultant of cultural and risk factors found in Mexico, and it appears to continue despite awareness of the potential negative health impacts. Identifying policy solutions to the type of nonlinear, straightforward interfuel substitution found in these areas of Mexico will require much greater thought, study, and a paradigm slightly more complex than that of simplified energy ladder.

5. CRITICISMS OF THE CONCEPT

While it is clear that households facing the development transition will also undergo some form of transformation of their household energy use, the energy ladder concept has been criticized from a number of different perspectives. As it developed from a general understanding of household budgetary decisions, a knowledge of the development process, and a familiarity with the differences in energy use patterns between more developed and less developed countries, it serves as a loose conceptual framework that can neither be proved or disproved, but that rather informs or underlies a discussion of the transition in household energy use patterns in a particular context. Nevertheless, several criticisms of the concept deserve mentioning as they might provide a foundation for conceptual models that will prove even more useful in understanding the household energy transition.

The first such criticism is that the energy ladder fails to account for the complexities of actual household fuel usage, assuming that all households—or at least most households—will move progressively from traditional fuels to more modern fuels in a predetermined succession. It fails to account for households that might skip rungs by jumping from firewood to LPG without using what appears to be the ubiquitous transitional fuel: kerosene. It also fails to account for multiple fuel use as anything other than a transitional option between two rungs even though many households even in developed countries use more than one fuel for cooking, both as risk-minimization strategies and as cultural preferences. Multiple fuel-use strategies might even result in the most efficient overall household energy-use profiles, as fuels are used for those activities for which they are most efficient. Multiple fuel use may be a logical long-term household fuel choice given cultural preferences, local fuel availability or unavailability, system-wide energy efficiency, and other location-specific considerations. Pursuing multifuel use further might lead to the concept of bifurcated household energy ladder or a different stage model incorporating platforms being maintained at various stages rather than merely a simple ladder whose only options are for remaining stationary, ascending, or descending. A more complex model would do a better job of explaining reality, but would also run the risk of sacrificing the main strength of the energy ladder concept: its intuitive appeal.

A second criticism centers on the value-laden nature of the energy ladder. Value-laden judgments regarding the modernity of society and the fuels associated with a society at a particular "stage" of development underpin much of the thinking behind the use of the energy-ladder concept. Modernization theory—wherein all countries are engaged in the homogenizing conveyor belt called "modernization"

resulting in their becoming more like developed countries—has been discredited as a model for development thinking for quite some time. Yet the energy ladder still builds on this paradigm. Because households in developed or OECD countries have followed a certain evolution in their energy-use patterns, households in developing countries have been assumed to fall into the same—or at least—a very similar pattern. It ignores questions about the conspicuous consumption or lifestyle issues that must be raised in the context of developed countries. Clearly, if all of the people in developing countries used energy in the same profligate manner that those in developed countries do, the earth's energy resources would be rapidly depleted. The energy ladder assumes that there exists but one final end point for the household energy transition, and that end point is the same for developed and developing countries alike. It thereby ignores the possibility that developing countries might pursue paths that are more environmentally friendly; more sustainable over the longer term; and consistent with their own cultural traditions.

A third criticism focuses on the fact that the energy ladder has been built mostly on household energy surveys providing a snapshot of energy consumption at a particular point in time. Its basis is cross-sectional in nature, building more on differences between households at one point in time rather than on longitudinal evidence documenting changes in time among the same households. Generalizing from cross-sectional data to draw time-series conclusions may lead to a misrepresentation of trends and relationships, including the underestimation of the importance of price and cost to influence household decision making. More careful analysis of trends over time would clearly help to strengthen the case for the household energy transition as a whole.

Fourth, the energy ladder concept makes no allowance for the dynamics of power and control within each household unit. In rural households, for example, it may not be unusual for women to be charged with collecting firewood and cooking, while men control financial resources and therefore make decisions requiring financial allocations. For the sake of convenience and reducing exposure to toxic smoke, women might want to move up the energy ladder, but are prevented from doing so by an unwilling husband who controls the household's financial resources. Thus, the predictive capability of the model may be limited in cases where such situations occur. Some would argue that using the energy ladder as a concept informing the analysis may preclude the consideration of gender roles in the household energy sector altogether. As earlier discussions have shown, household energy discussion clearly could benefit from greater analysis of the disaggregated nature of household units and the different decision-making frameworks of the individual agents involved.

Finally, the energy ladder yields more predictive insights into the household energy transition are more consistent in urban areas rather than rural areas. By and large, rural household energy-use patterns appear to be overwhelmingly determined by local fuel availability rather than other factors that are readily predicted on the basis of economic theory or factored into regression equations. For the most part, policymakers exert relatively little control over rural household fuel-use patterns. In urban areas, the problem may seem more comprehensible and tractable using the microeconomic theory of consumer behavior. That is not to deny the many instances of documented upward or downward shifts on the energy ladder by rural households. Rather, it merely means that the complexity of rural household energy use may be more difficult to reduce to a simple analogue model than are urban energy-use patterns.

6. CONCLUSIONS

The concept of the energy ladder has informed energy studies and policies in Africa, Asia, and Latin America. Household fuel choices are influenced by a number of factors that are both internal and external to the households making the decisions. Statistical models have attempted to identify these factors and define the strength of their influence in determining household fuel choices. Although it has been used as a conceptual basis to inform much of the work on household energy in developing countries, criticisms have brought to light several limitations. To date, no one has posited an alternative conceptual framework describing household fuel choice and decisions in the process of economic development. As a result, the energy ladder concept will likely continue to be used as an analogue model informing discussions of household energy transitions in developing countries. Its strength lies in its analytical clarity and simplicity. Given the importance of energy to economic development and the continued challenge of providing energy for both the subsistence and development needs of the urban and rural poor in developing

countries, the issues raised in the discussion of the energy ladder will likely continue to be important for energy and development professionals for the foreseeable future.

SEE ALSO THE FOLLOWING ARTICLES

Cultural Evolution and Energy • Development and Energy, Overview • Economic Growth and Energy • Equity and Distribution in Energy Policy • Global Energy Use: Status and Trends • International Comparisons of Energy End Use: Benefits and Risks • Population Growth and Energy • Sustainable Development: Basic Concepts and Application to Energy • Technology Innovation and Energy • Transitions in Energy Use • Women and Energy: Issues in Developing Nations • Wood in Household Energy Use

Further Reading

Barnes, D. F., and Floor, W. (1999). Biomass energy and the poor in the developing world. *J. International Affairs* **53**(1), 237–259.

Davis, M. (1998). Rural household energy consumption: The effects of access to electricity–evidence from South Africa. *Energy Policy* **26**(3), 207–217.

DeWees, P. A. (1989). The woodfuel crisis reconsidered: Observations on the dynamics of abundance and scarcity. *World Development* **17**(8), 1159–1172.

Eberhard, A., and van Horen, C. (1995). "Poverty and power: Energy and the South African state." Pluto Press, East Haven, CT.

Energy Sector Management Assessment Program (ESMAP) (1996). "Energy for Rural Development in China: An Assessment Based on a Joint Chinese/ESMAP Study for Six Counties." UNDP/World Bank ESMAP Program, Washington, DC.

Energy Sector Management Assessment Program (ESMAP) (2002). "Energy Strategies for Rural India: Evidence from Six States." United Nations Development Program/World Bank ESMAP Program, Washington, DC.

Fitzgerald, K., Barnes, D., and McGranahan. G. (1990). "Interfuel Substitution and Changes in the Way Households Use Energy: The Case of Cooking and Lighting Behavior in Urban Java." Industry and Energy Department Working Paper, Energy Series Paper Number 29. The World Bank Industry and Energy Department, PRS, Washington, DC.

Hosier, R. H. (1985). "Energy Use in Rural Kenya: Household Demand and Rural Transformation." Energy, Environment and Development in Africa, Volume 7. Beijer Institute and Scandinavian Institute of African Studies, Stockholm.

Hosier, R. H., and Dowd, J. (1987). Household fuel choice in Zimbabwe: An empirical test of the energy ladder hypothesis. *Resources and Energy* **9,** 337–361.

Hosier, R. H., and Kipondya, W. (1993). Urban household energy use in Tanzania: Prices, substitutes and poverty. *Energy Policy* **21,** 454–473.

International Energy Agency (2002). "World Energy Outlook 2002," especially Chapter 13, "Energy and Poverty." Organization for Economic Cooperation and Development and IEA, Paris.

Koumoin, M. C. (1998). Global carbon initiative and CO_2 mitigation in urban Sub-Saharan Africa: Misreading the empirical evidence? *World Resources Rev.* **10**(4), 607–632.

Leach, G. (1987). "Household Energy in South Asia." Elsevier Applied Science, London and New York.

Leach, G. (1992). The Energy Transition. *Energy Policy* **20**(February), 116–123.

Masera, O. R., Saatkamp, B. D., and Kammen, D. M. (2000). From linear fuel switching to multiple cooking strategies: A critique and alternative to the energy ladder model. *World Development* **28**(12), 2083–2103.

Sathaye, J., and Tyler, S. (1991). Transitions in household energy use in urban China, India, the Philippines, Thailand, and Hong Kong. *Annual Rev. Energy and the Environment* **16,** 295–335.

Smith, K. R. (1987). "Biofuels, Air Pollution, and Health: A Global Review." Plenum, New York.

Smith, K. R. (1993). Fuel combustion, air pollution exposure, and health: The situation in developing countries. *Annual Rev. Energy and the Environment* **18,** 529–566.

Energy Services Industry

EDWARD L. VINE
Lawrence Berkeley National Laboratory
Berkeley, California, United States

1. **Introduction to the Energy Services Industry and Energy Service Companies**
2. **Origins and Growth of the U.S. ESCO Industry**
3. **Future Prospects for the U.S. ESCO Industry**
4. **International ESCO Activities**
5. **Conclusions**

Glossary

demand-side management (DSM) bidding An auction in which a utility generally solicits proposals from energy service companies interested in achieving specified amounts of DSM savings (e.g., 1000 kW of demand reduction).

energy service company (ESCO) A company that is engaged in developing, installing, and financing comprehensive, performance-based projects, typically 5–10 years in duration, centered around improving the energy efficiency or load reduction of facilities owned or operated by customers.

guaranteed savings The ESCO guarantees to the customer that the project's realized savings will equal (or exceed) the payments made (or financing costs absorbed) by the customer over the term of financing. Also known as chauffage in Canada and Europe.

investment-grade audit A detailed audit that is used as the basis for financing a project as well as for projecting energy and other savings on a performance contract.

non- and limited-recourse project financing Based on payments made exclusively from the cash flow (revenues) of the project and not by the other nonproject assets of the customer or ESCO.

off-balance sheet financing The building or facility owner can receive investments in his or her facility but with no financial liability recorded on the organization's balance sheet.

performance-based projects The ESCO's compensation, and often the project's financing, are meaningfully tied to the amount of energy actually saved, and the ESCO assumes the risk in linking its compensation directly to results.

shared savings projects The ESCO's payments are typically structured as a percentage of the energy savings actually realized by the end user over the length of the contract.

super ESCOs ESCOs that provide traditional energy services and supply gas and/or electricity (and/or other fuels) to customers.

third-party financing Funding of energy savings investments by an outside company, using energy savings to pay for the project.

The energy services industry represents an important group for promoting energy efficiency in developed and developing countries. This article provides an overview of energy service companies (ESCOs), starting with a description of their basic characteristics, and then tracing their origin and growth in the United States. After examining the future prospects for the U.S. ESCO industry, the article ends by briefly examining the ESCO industry in other countries.

1. INTRODUCTION TO THE ENERGY SERVICES INDUSTRY AND ENERGY SERVICE COMPANIES

1.1 ESCOs and Performance Contracting

An energy service company (ESCO) is a company that is engaged in developing, installing, and financing comprehensive, performance-based projects, typically 5–10 years in duration, centered around improving the energy efficiency or load reduction of facilities owned or operated by customers. Projects are performance-based when the ESCO's compensation, and often the project's financing, are meaningfully tied to the amount of energy actually saved, and the ESCO assumes the risk in linking its compensation directly to results. Performance contracting is another term for the performance-based approach.

 © 2004 Elsevier Inc. All rights reserved.

Energy performance contracts can be distinguished from two contracting approaches commonly utilized by other types of energy-efficiency service providers: (i) contracts for installation of efficient equipment for a specified fee, which is typically backed by a manufacturer's equipment warranty contract, and (ii) energy consulting services, such as building audits and advice on potential energy-efficiency improvements. Contracting based on performance does not necessarily have to be undertaken by an ESCO, but in practice ESCOs have been the pioneers and major users of performance contracting for energy-efficiency projects. ESCOs used performance contracting primarily as a method for assuring their clients that they would confidently achieve energy and cost savings—there were no other companies in the marketplace taking this risk. For this reason, ESCOs are fundamentally different from consulting engineers and equipment contractors specializing in efficiency improvements; the former are typically paid a fee for their advice, whereas the latter are typically paid a fee for their equipment, and both accept no project risk. Such risk creates the highest possible motivation to properly specify, design, engineer, install, and maintain savings over the length of the contract.

1.2 ESCO Diversity

Each ESCO differs in terms of ownership, target markets, technology focus/expertise, in-house capabilities, etc. The following are some of the key areas in which ESCOs can be differentiated:

Ownership: ESCOs may be privately owned (independent) companies, utility subsidiaries or affiliates, nonprofits, joint ventures, manufacturers, or manufacturers' subsidiaries.

Target market: ESCOs typically focus on market niches, by subsector (e.g., hospitals, schools, and government) and by project size (e.g., large customers).

Service specialization: Some ESCOs perform project installation using in-house expertise, whereas others specialize in engineering design and analysis and subcontract other services (e.g., equipment installation).

Technology: Most ESCOs have a technological bias [e.g., lighting or heating, ventilation, and air-conditioning (HVAC) controls], whereas others are comprehensive by seeking to achieve energy savings from the widest possible array of cost-effective measures in a given facility. Also, some ESCOs focus more on electrical end uses than those using other forms of energy (primarily because most projects have involved lighting), but comprehensive ESCO projects almost always involve thermal measures.

Geographic preference: Some ESCOs have established a preference to conduct business in specific geographic regions of one country, whereas others are national and a few have done some international work.

Project financing: ESCO financing capabilities vary with the financial strength of the ESCO. Those with "deep pocket" parent companies typically have a greater capacity to offer off-balance sheet financing. Some ESCOs have significant, well-established financing capabilities, some have limited in-house financing capabilities, and others have none at all and must arrange financing through lenders (the typical case). However, all ESCOs rely to some extent on third-party financing since most ESCOs do not have internal financing capabilities.

1.3 ESCO Capabilities and Skills

ESCOs are diverse: There is no real prototypical ESCO. However, most ESCOs typically have the following capabilities and skills: project development, engineering and design, feasibility analysis, energy analysis, general contracting, the ability to finance directly or arrange third-party financing, project and construction management, purchase and installation of equipment, risk management, monitoring and verification (M&V) of savings, training, operations and maintenance services for the installed equipment, and administrative services. Most of these capabilities reside among the staff of the company (in-house), but some of these skills are contracted out (e.g., installation of equipment). In addition, ESCOs may provide additional services, such as cogeneration and alternative power production.

Some additional comments on these activities are warranted. First, ESCOs use working capital for general corporate purposes such as marketing. They use project financing to construct a project; they also reimburse themselves through project financing for the amount of working capital expended to develop a particular project. All or most of the activities that ESCOs conduct for a project are bundled into that project's cost, which is usually financed entirely with debt and lease financing (i.e., no equity contribution is required to be made by the customer since payments come only from realized savings); project financing using ESCO or third-party equity contributions is rare. In sum, ESCOs attempt to remove the first-cost disincentive and technical uncertainty by

providing outside sources of capital and guaranteeing the performance of the equipment. The first-cost disincentive is a real barrier to investments in energy efficiency, whereas the perception of technical uncertainty is a more formidable barrier than the actual performance of energy-saving measures. The cost of the ESCO service is typically paid for from the stream of energy cost savings resulting from the project.

Second, because ESCO projects are performance based, ESCOs must measure and verify the energy savings, and these savings must be monitored over the project's life to make sure the performance-based savings are accurate and persist. M&V is typically conducted by periodically comparing a customer's energy bills with an established pre-project baseline and/or through energy metering.

Third, maintenance is usually the final component of most performance-based energy-efficiency projects. The maintenance is for all (or some portion) of the new equipment over the life of the contract. The cost of this ongoing maintenance may be folded into the overall cost of the project. Thus, throughout the life of the performance contract, the customer receives not only the benefit of reduced energy costs but also the benefit of reduced maintenance costs. The maintenance component is also important since it enables the ESCO to ensure that the equipment is maintained in a way that preserves optimal performance. As an additional service on most contracts, the ESCO also provides any specialized training needed so that the customer's maintenance staff is equipped to take over at the end of the performance contract period (the period may be 5, 10, or 15 years, and sometimes longer).

Fourth, each ESCO entering into an energy-savings performance contract tries to educate its customer on energy use and seeks to bring that customer into an energy-efficiency partnership whose purpose is to achieve the maximum energy savings available from the customer's facility. This is not only good for the customer but also beneficial for the ESCO as it seeks to maintain a long-term contract with the customer and obtain additional business.

In sum, although none of the discrete skills that an ESCO employs are particularly unique, the added value an ESCO brings is its ability to integrate a wide variety of skills and apply them efficiently to projects ranging from several hundred thousand dollars to tens of millions of dollars. Thus, ESCOs view themselves as project developers and integrators.

1.4 ESCO and Customer Responsibilities in Performance Contracting

In an energy performance contract, project responsibilities are divided between the ESCO and the customer. Once a customer has been contacted, an ESCO has the following responsibilities:

- Conducts walk-through energy audit and preliminary engineering study.
- Prepares engineering proposal and project development agreement based on audit results.
- Presents proposal and agreement to customer for comments and approval.
- Conducts an investment-grade audit.
- Finances purchase of energy-saving and measurement equipment (unless financed by customer or by third-party bank or leasing company).
- Installs equipment.
- Maintains equipment over life of contract (if maintenance provided).
- Delivers agreed level of energy savings.
- Monitors, measures, and verifies energy savings.

Customer responsibilities include the following:

- Reviews and approves engineering proposal (if proposal is not approved, customer pays ESCO a fee).
- Operates host facility as specified.
- Compensates ESCO as appropriate to type of financing used.
- Pays value of share of energy savings (if ESCO financing secured by shared savings).
- Permits ESCO to maintain and measure savings.
- Pays for services (if third-party financing).
- Maintains equipment over life of contract (if maintenance not provided).

1.5 Financing of ESCO Projects

Most ESCOs provide financing for projects from their own funds (working capital); however, they generally are only able to finance the initial stages of a project. Even though the ESCO eventually has no equity investment in the project, the ESCO does provide its equity to fund the initial risk equity by investing its working capital to develop and finance a paid-from-savings project for the customer. Thus, ESCOs seeking to use performance contracting must develop or arrange sources of debt and equity to finance their projects. More often, the ESCO's role is to arrange

financing for its customers with leasing companies, institutional investors, and commercial banks.

The customer usually enters into separate agreements for energy services and for financing. Typically, the ESCO may arrange financing but is not a party to the finance agreement. The end-user credit risk is assumed by the financier, not the ESCO. The ESCO and the end user enter into a turnkey contract or energy services agreement whereby the ESCO provides engineering, equipment installation, and other services such as operations, maintenance, and savings verification. If the savings do not materialize as projected, the end user has recourse against the ESCO.

Many ESCO projects are financed using nonrecourse or limited-recourse financing. Since no capital is required up front and financiers are repaid through energy savings, ESCOs and customers are allowed to develop projects that are much larger than their net worth. Although nonrecourse financing has important advantages (e.g., a positive cash flow immediately for the customer), it also introduces significant business risks into the financing and higher project costs.

The two types of project financing commonly used by ESCOs in the United States are guaranteed savings and shared savings. In a guaranteed savings agreement, the ESCO guarantees to the customer that the project's realized savings will equal (or exceed) the payments made (or financing costs absorbed) by the customer over the term of financing. In guaranteed savings contracts, ESCOs typically guarantee approximately 80–85% of the savings they expect to get out of a project. In these projects, the customer assumes the obligation to repay the project costs to a third-party financier (usually introduced by the ESCO), which is very often a commercial bank or leasing company. If the customer finances the project itself, the customer repays itself by saving energy costs.

If the savings realized fall short of the payments made by the customer due to a failure of the project to perform as guaranteed, the ESCO pays the customer the difference. If, however, actual savings exceed the payments, excess savings are normally returned to the customer. In some cases, the customer and ESCO negotiate the sharing of savings in excess of the amount guaranteed.

In shared savings projects, the ESCO's payments are typically structured as a percentage of the energy savings actually realized by the end user over the length of the contract. The percentage is predetermined, but the percentage of savings allocated to each party can vary over the period of the contract. The ESCO must ensure that the proceeds it receives will cover the costs it has incurred in implementing the project.

In exchange for the ESCO providing project financing and for the ESCO's assumption of performance and credit risk, the customer agrees for a fixed term to pay the ESCO the value of a set percentage of those savings. The portion of savings paid to the ESCO that is attributed to debt service is always higher for shared savings than for guaranteed savings projects, reflecting the ESCO's significantly greater risk and expense. The ESCO uses its payments from the customer to repay debt associated with the project, conduct equipment maintenance, measure and monitor savings, and earn a return. As with guaranteed savings projects, the ESCO measures energy savings throughout the term of the project agreement.

2. ORIGINS AND GROWTH OF THE U.S. ESCO INDUSTRY

2.1 ESCO Driving Forces

The three key driving forces affecting the ESCO industry in the United States during the past 15 years have been energy prices, tax policies, and utility demand-side management (DSM) bidding programs. The U.S. ESCO industry began in the late 1970s and early 1980s, a period during which energy prices were dramatically rising in response to the 1973 oil embargo and the Iranian revolution in 1979, which presented the opportunity for fledgling ESCOs to make a business out of reducing customers' growing energy costs. For example, when energy prices were high (1980–1983), the annual growth rate in ESCO investments during this period was estimated to be 35%; the rising electricity prices from 1989 to 1994 led to another period of increased growth rate (25%) in ESCO investments.

During its formative years, changes in tax policy were also critical in shaping the ESCO industry. For example, the energy tax credit (available for energy-efficiency investments from 1980 to 1983) and the investment tax credit (for general capital investment, including efficiency improvements, from 1980 to 1986) affected the rate of ESCO project investment between 1980 and 1986. After the energy tax credits ended in 1983, ESCO investments decreased by 10% in 1984.

Finally, utility DSM bidding programs significantly impacted the ESCO industry in its early years. In a DSM bidding program, the utility enters into a

long-term conservation purchase agreement with the provider (usually an ESCO) to reduce demand or provide firm energy savings during specified periods. The utility pays for these savings on a unit basis, often simply per kilowatt-hour, much as it would pay for private power under a purchase agreement. The first utility DSM bidding program began in late 1987, initiating a period of substantial utility contributions to the pool of investment capital underwriting performance-based efficiency projects. Most observers believe, however, that the influence of these utility programs on the U.S. ESCO industry peaked in 1994: Some utilities froze or canceled DSM programs in the atmosphere of uncertainty that commenced with the California restructuring proposal in April 1994.

2.2 Types of ESCOs

Four main groups of ESCOs dominate the U.S. ESCO industry: original ESCOs, engineer-based ESCOs, manufacturer-based ESCOs, and utility-based ESCOs. Whatever their origins, all ESCOs are characterized by their project development capabilities and their assumption of performance risk for their projects.

The original ESCOs entering the marketplace were start-up ventures established specifically to pursue energy performance contracting. The engineer-based ESCOs grew out of businesses providing engineering services for a fee. Most of these consulting engineers specialized in performing detailed energy audits of clients' buildings, proposing designs for new energy systems, and suggesting improvements to operating practices. Some engineering firms fundamentally altered their business strategy and changed from fee-based energy services consulting to project development. These firms were willing and able to back their designs with performance guarantees and to secure financing for projects with the savings these projects were guaranteed to produce.

Manufacturer-based ESCOs are manufacturers of building control systems designed to regulate energy-using equipment in buildings and facilities. The earliest firms offered control systems specifically designed to reduce their customers' energy bills. Like the consulting engineers, these manufacturers extended their traditional business activity (e.g., controls) to assume the risks associated with project development. In the early 1980s, these companies established energy service divisions to develop performance-based projects centered around their control systems. These companies further evolved by developing comprehensive projects that include the full range of available energy-efficiency measures.

Utility-based ESCOs include those ESCOs formed in response to the growth of utility DSM programs, either by the utilities or by private investors. Numerous utilities have launched DSM programs that complement ESCO projects: For example, utility customers have used rebate programs to improve the cost-effectiveness and comprehensiveness of their performance-based projects. Similarly, some utilities have offered DSM bidding programs in which ESCOs have been encouraged to participate. Nevertheless, the utility-created ESCO has historically been rare. Utility entry into the ESCO market or acquisition of existing ESCOs have accelerated significantly as utilities seek to position themselves to compete effectively in a retail energy services environment.

2.3 The Maturation and Growth of the ESCO Industry

The U.S. ESCO industry has matured in the past decade. The more positive image of the industry and its resultant growth stem from a combination of factors, including (i) the trend toward guaranteed energy savings, (ii) the development of standardized monitoring and verification protocols, (iii) the involvement of the National ESCO Industry Association (NAESCO) in accreditation and M&V protocols, (iv) DSM bidding programs, (v) the facilitation of energy performance contracting in the federal government, and (vi) improvements in performance and reductions in costs of energy-efficiency measures.

There are no firm data on the ESCO industry's size in the United States, and traditional financial measures are difficult to apply in an industry composed primarily of privately held companies and large public firms that do not report the financial results of their efficiency divisions separately. Nevertheless, there are more than 60 national and regional ESCOs actively operating in the United States. Approximately 12 companies account for 75% of the industry revenues. ESCOs typically develop projects in the $600,000 to $2 million range, and they install between $800 million and $1 billion worth of energy-efficient equipment (including labor and profit) each year. Since 1990, the installed cost of all projects implemented by ESCOs in the United States is $6–7.2 billion. During the late 1990s, the ESCO industry grew by approximately 25% per year, mainly driven by rising electricity prices during

from 1984 to 1994. In 2002, ESCO industry revenues for energy-efficiency-related services in the United States ranged from $1.8–2.1 billion.

2.4 ESCO Markets and Marketing

ESCOs have established themselves in well-defined niche markets using either geographic or industry-specific market segmentation strategies. For example, ESCOs that have been selected in DSM bidding programs naturally target certain types of customers (e.g., hospitals and small retail) in specific utility service territories. ESCOs have also tended to focus on regions in which retail electricity prices are high in order to maximize differences between project cost and benefits to customers from energy-efficiency investments. Other ESCOs, particularly those affiliated with the large controls and equipment companies, utilize a decentralized and extensive branch sales force that targets certain industries or sectors (e.g., schools and state and local governments).

The focus of most ESCOs has been on medium to large commercial and institutional customers: Local and state government, schools, and universities account for approximately 65–75% of overall ESCO activity. ESCOs have also achieved success among other groups of commercial-sector customers (e.g., owner-occupied office buildings, shopping malls, hotels, and a few national franchises). The federal government, for which military projects have huge energy bills, is also targeted by many ESCOs.

Industrial customers have hosted fewer performance-based efficiency projects, but their share of projects has increased slightly over time as energy services are becoming integrated with other services. In the industrial sector, ESCOs have attempted to differentiate themselves through specialized expertise in certain technologies (e.g., adjustable-speed drives and controls) or processes (e.g., pulp/paper and automotive).

Because the typical ESCO project costs more than $350,000, small commercial and industrial companies and residential customers are generally not served by ESCOs. Some ESCOs do work in certain niche markets (e.g., large multifamily projects and water and wastewater facilities), but these are rare.

The so-called MUSH institutional markets (municipal governments, universities, schools, and hospitals) have been the focus of ESCO marketing because they tend to be very capital constrained; have a longer investment horizon; are faced with an aging infrastructure and equipment in need of modernization; and often have limited technical expertise to develop, coordinate, and manage a comprehensive energy-efficiency project. Furthermore, ESCOs have been able to arrange relatively attractive financing terms with the MUSH markets because financial institutions have regarded these customers as relatively good credit risks (i.e., they are unlikely to go out of business or relocate).

2.5 Energy Measures in ESCO Projects

In 1980, approximately 66% of ESCO investment was in electrical efficiency measures, and the rest was in thermal energy measures. From 1980 to 1987, ESCOs gradually invested in more thermal energy measures and fewer electricity measures, until the investment ratio in 1987 was 55% of investments in electricity measures and 45% of investments in thermal energy measures. Beginning in 1988, when utility DSM bidding programs were predominantly sponsored by electric utilities, investments in electricity measures increased, resulting in an ESCO investment ratio of 75% electrical versus 25% thermal. Recently, thermal energy measure investments have increased so that the relative shares in 1997 were approximately 65% electrical versus 35% thermal.

A number of technologies have come to form the core of the typical performance-based energy-efficiency project, including high-efficiency lighting, high-efficiency HVAC equipment, efficient motors and variable-speed drives, and centralized energy management systems. Along with the core technologies are a full array of weatherization services where appropriate and, sometimes, more innovative applications, such as integrated cogeneration or renewable energy technologies.

2.6 Measurement and Verification and Protocols for Measurement and Verification

All ESCOs measure and verify the energy saved by their projects over the projects' lives—as demanded by the customer and as the essence of energy performance contracting. The purpose of monitoring energy savings over time is to ensure that those savings persist and to identify immediately any anomalies in consumption patterns that may occur. It also serves to ensure the quality and effectiveness of ongoing maintenance, entailed in most ESCO projects.

The development of standardized procedures for the measurement and verification of savings has helped both end users and the financial community to understand this key dimension of performance contracting. ESCOs have led the effort to verify, rather than estimate, energy savings, particularly when public and utility ratepayer funds are paying for them.

U.S. ESCOs have adopted the U.S. Department of Energy's (DOE) International Performance Measurement and Verification Protocol (IPMVP) as the industry standard approach to measurement and verification. States ranging from Florida to New York require the use of the IPMVP for state-level energy-efficiency retrofits. The U.S. federal government, through the Department of Energy's Federal Energy Management Program, uses the IPMVP approach for energy retrofits in federal buildings.

2.7 NAESCO

NAESCO was established in 1983 to bring credibility to the ESCO industry. Its members include independent performance contracting ESCOs; ESCOs affiliated with electric and natural gas utilities; energy-efficiency equipment suppliers, manufacturers, and other energy service providers; and a "miscellaneous" group (e.g., consulting firms, financial companies, and public-sector members). The primary activities of NAESCO include conferences, regulatory intervention (i.e., lobbying), accreditation, and publications.

NAESCO has a model energy performance contract that has been in use for 10–15 years and that forms the basis for most energy performance contracts today. NAESCO also serves an important policing service, ensuring that ESCOs are doing business in a fair and equitable fashion. The NAESCO board holds its members to a high standard of performance, as reflected in the accreditation process. To receive the accreditation, an ESCO goes through a rigorous review of credentials, references, technical capabilities, financial capabilities, project history, and interviews of former and existing clients. Certification must be renewed every year.

2.8 Utility Programs and ESCOs

Utility DSM bidding programs have played a critical role in the development of the ESCO industry. Although not all ESCOs have participated in utility DSM bidding programs and some ESCO participants in these programs have negative views of their experience, most observers of the ESCO industry agree that these programs have helped to promote the ESCO industry.

2.9 Energy Performance Contracting in the Federal Government

In the past 25 years, both federal and state governments have adopted legislation for improving energy efficiency. Although most of this legislation has had indirect impacts on ESCOs (e.g., the promotion of energy efficiency), the real impact has been on prompting governments as customers to seek assistance from ESCOs. Governmental facilities are prime candidates for ESCO projects: These facilities tend to use energy inefficiently, the owners of the buildings have constrained budgets, and there are many buildings.

ESCOs can contract directly with the government, mainly through a vehicle called an energy savings performance contract (ESPC). An ESPC can be a type of shared savings contract, with no government capital required up front. ESCOs can finance ESPC with their own internal funds or with third-party financing. Under an ESPC, an ESCO typically provides the government with design, installation, and ongoing maintenance services as well as the capital needed to finance efficiency improvements. Payments are made to the ESCO based on a share of savings that resulted from the energy-savings measures implemented (i.e., a shared savings contract).

The Federal Energy Management Program (FEMP) of the U.S. DOE developed the broad-area FEMP Energy Saver Performance Contract (Super ESPC). The first broad-area contract, covering eight states in DOE's western region, was awarded in 1997. Under these contracts, federal agencies award delivery orders from existing contracts and reduce the lead time and resources to award individual contracts. Performance-based energy services are available to federal agencies with government-owned facilities within a region. These contracts involve the competitive selection of a small number of contractors (i.e., ESCOs) that are allowed to conduct negotiations with individual facilities for the awarding of delivery orders under the contracts. The selected contractors are able to add work to their contracts for a limited time without the need for additional extensive, time-consuming competition and evaluation.

3. FUTURE PROSPECTS FOR THE U.S. ESCO INDUSTRY

3.1 Opportunities and Challenges in a More Competitive Utility Environment

The future prospects for ESCOs and other energy service providers will be affected by both the length of the transition period to a more competitive electricity industry and the ultimate organization and institutional structure of bulk power and retail service markets. The issues surrounding electricity industry restructuring are enormously complex and multifaceted. The pace of industry restructuring is likely to vary state by state given regional differences in electricity prices and resource endowments and the influence of state regulation. Because the pace and ultimate outcome of industry restructuring cannot be determined, ESCOs and other energy service providers face a retail market environment with significant uncertainties, many of which are regulatory (rather than market) driven.

Many observers believe that the restructuring of the utility industry will lead to lower power prices, resulting in a reduction in energy-efficiency programs and projects (e.g., energy-efficiency projects will be less attractive and paybacks will be longer if energy is sold at 5 rather than 9¢/kWh). Short-term impacts have already been felt: (i) Energy is receiving the attention of top management in businesses as well as utilities, (ii) gas and electricity savings due to lower prices are seen as more important to end users than energy-efficiency savings (i.e., low-cost energy providers are expected to prevail), (iii) energy is primarily seen as a commodity (rather than as a service), (iv) more emphasis is being placed on short-term contracts, and (v) customers are requiring shorter paybacks and shorter returns on projects. The end result is that customers will be less motivated to implement energy-efficiency projects.

On the other hand, despite the expectation of lower prices, there is a great opportunity for energy efficiency and new types of energy-efficiency services. In a competitive marketplace, as price differences among power suppliers narrow, value-added products and services (such as energy efficiency) will be offered to customers by various ESCOs and other types of energy service providers. In addition, state adoption of a system benefits charge will encourage and fund energy-efficiency activities during the transition to a competitive electric industry.

In summary, in a restructured electricity industry, it is likely that support for energy efficiency will continue, although most likely on a diminished basis. For ESCOs that have historically targeted larger institutional and C&I customers and that have depended on utility DSM funding, ratepayer funding for energy efficiency in the form of financial incentives (e.g., rebates) may be unavailable or available only for a short period (e.g., 2–5 years).

3.2 New Customer Needs

Customers have an array of needs that ESCOs will be asked to address. Customers are also quickly realizing that they will have more options and choice in a restructured electricity industry and will demand more products and services. Furthermore, customers will want changes in the way ESCOs do business with them, such as more flexible contracts, less risk, more warranties, simpler and clearer agreements, more customer control and involvement in projects, more convenience, and bundling or unbundling of products and services.

3.3 New ESCO Products and Services

ESCOs have the potential to interact with customers with more creativity, and it is likely that ESCOs will offer packages of comprehensive energy services, different fuels, varieties of pricing plans, and other creative ideas (including energy efficiency) as the market develops and grows. Occasionally, ESCOs will need to unbundle products and services for clarity and competitiveness, whereas at other times they will need to bundle products and services for convenience. Some ESCOs may manage a customer's fuel and energy needs through bill consolidation, analysis, and payment. The more innovative ESCOs may also be asked to provide nontraditional, performance-based customer service, such as providing "total comfort" or lighting based on $/ft^2$ or $/load.

Hence, ESCOs may find more opportunities in providing assistance in other areas besides utility-related issues, including facilities management and operations, capital equipment budgets, environmental concerns, and compliance with government regulations. For example, some ESCOs are considering guaranteeing environmental compliance as well as energy savings when they change a process to improve the environment (e.g., indoor air quality). They will be responsible for any mistakes (e.g., if the facility does not comply with environmental regulations or if savings do not occur), and they will fix the mistakes.

In sum, today's ESCO may become tomorrow's super ESCO—a fully integrated (full-service) energy

services company that brokers power and offers energy efficiency and other energy services. The ESCOs that respond to these challenges may emerge stronger than before and ahead of their new competitors as they provide "energy services" in its broadest sense.

3.4 Changes in ESCO Practices

In addition to providing new services, ESCOs will need to change their business practices to remain competitive. Many of them will form partnerships, joint ventures, mergers, and strategic alliances (both regionally and nationally). All of them will take more risks in providing new services, developing new alliances, being entrepreneurial and agile developers, energy brokering and marketing, and integrating energy efficiency with power supply and communication services. More ESCOs will be asked by customers to act as problem solvers to address multiple and complex customer demands and to identify resource efficiency opportunities. Accordingly, ESCOs will be looked upon as technology integrators, resource managers, and networkers. Finally, ESCOs are expected to continue to target large (especially institutional) customers and wholesale markets (federal government) that need engineering and financial assistance. However, they will also be aggregating smaller customers (small C&I and residential) to strategically position themselves at the forefront of the ESCO industry.

3.5 New ESCO Competitors (or Partners?)

The restructuring of the electric and natural gas industries will change the nature of competition and competitors in the energy-efficiency performance contracting market. In a more competitive electricity industry, the competitors to ESCOs will likely include power marketers, utilities, utility affiliates, financial institutions, equipment suppliers, property managers, cable companies, telecommunication companies, building control companies, high-technology companies, independent power producers, and energy traders.

ESCOs will need to determine whether they want to compete with these entities or form partnerships or joint ventures. Most non-comprehensive ESCOs will continue to seek to partner with other firms to develop business and to implement large-scale projects. Full-service providers will also be expected to partner with other firms to compete for customers by offering a range of potential energy services, making it more difficult for specialized or small energy providers to develop their own work independently. Partnering or consolidation of full-service providers with utilities or large, well-capitalized firms can be expected to increase the level of competition and the effectiveness of competition for customers.

As they prepare for competition, many utilities are using the regulated distribution business unit as an incubator to test various customer and energy services, whereas others have located these activities in unregulated retail service companies. Increasingly, the markets sought and the services offered by utility distribution companies and ESCOs are becoming indistinguishable. This presents existing ESCOs with both a competitive challenge and a partnering opportunity, from ad hoc marketing agreements to formal joint ventures and acquisitions. In addition to retention of electricity and/or gas commodity sales, these utility/ESCO partnerships can sell comfort, light, power, reliability, and end-use commodities (e.g., refrigeration, steam, chilled water, and compressed air). They can also offer a variety of building-related services that enhance the initial energy-efficiency services, such as facilities management, operations and maintenance services, and energy management/control.

The benefits to ESCOs of these alliances are primarily reductions in the transaction costs of lead generation, qualification, and closing. The benefits to utilities are defensive (e.g., customer retention), offensive (e.g., value added to wholesale and eventually retail offers), and potential revenue enhancement (through technology applications, fuel switching, customer growth, etc.).

4. INTERNATIONAL ESCO ACTIVITIES

The ESCO industry has been developing in countries besides the United States, particularly since the mid- to late 1990s. In most cases, the number of ESCOs are small (less than 10) and revenue from the projects is generally less than $10 million. The impediments to the successful development of an ESCO industry in other countries are similar to those encountered in the United States. In particular, the following appear to be formidable:

- The concept of ESCOs and energy performance contracting is new to nearly all of the parties that need to be familiar with it (i.e., there is a lack of

awareness): financiers, attorneys, contractors, insurers, utilities, government agencies, and customers.
- Government policies regarding energy performance contracting are uncertain.
- Legal and contractual issues specific to performance contracting are unclear.
- Economic conditions are adverse (e.g., stagnant economic development) for investments in energy efficiency in many countries.
- There is a lack of capital (equity financing or equity investments) for establishing ESCOs in many countries.

Nevertheless, several countries have developed an ESCO industry and have started to promote the development of a vibrant ESCO industry. Several countries have established associations of energy service companies (e.g., China, Egypt, Japan, Korea, and the Ukraine). Revenue from the ESCO industry in a few countries exceeds $50 million (e.g., Brazil, Canada, China, and Japan). In 2001, ESCO industry revenues for energy-efficiency-related services outside the United States ranged from $560–620 million.

5. CONCLUSIONS

The ESCO industry is expected to grow significantly in the United States and in selected countries in the near future. There will be numerous new entrants in the ESCO market, including supply-side bidders (e.g., independent power producers) and utility subsidiaries. In the United States, ESCOs with technical strengths and deep pockets (or access to sizable funding) will survive; small ESCOs that specialize will disappear in a period of consolidation, some of which may be acquired by utilities. In the United States, the consolidation of ESCOs and the trend for utilities to acquire or develop full-service ESCOs to compete for other utilities' customers (particularly large customers) have become more prominent during the past few years. These are expected to also occur overseas, where the energy industry is being restructured.

SEE ALSO THE FOLLOWING ARTICLES

Corporate Environmental Strategy • Derivatives, Energy • Economics of Energy Demand • Economics of Energy Supply • Energy Futures and Options • Investment in Fossil Fuels Industries • Trade in Energy and Energy Services

Further Reading

Cudahy, M., and Dreessen, T. (1996). "A Review of the ESCO Industry in the United States." National Association of Energy Service Companies, Washington, DC.

Dayton, D., Goldman, C., and Pickle, S. (1998). The energy services company (ESCO) industry: Analysis of industry and market trends. *In* "Proceedings of the 1998 ACEEE Summer Study," Vol. **6**, pp. 29–45. American Council for an Energy-Efficient Economy, Washington, DC.

Goldman, C., and Dayton, D. (1996). Future prospects for ESCOs in a restructured electricity industry. *In* "Proceedings of the 1996 ACEEE Summer Study," Vol. **10**, pp. 59–69. American Council for an Energy-Efficient Economy, Washington, DC.

Goldman, C., Jurgens, P., Fowlie, M., Osborn, J., and Kawamoto, K. (2000). Historical performance of the U.S. ESCO industry: Results from the NAESCO project database. *In* "Proceedings of the 2000 ACEEE Summer Study," Vol. **5**, pp. 89–102. American Council for an Energy-Efficient Economy, Washington, DC.

Goldman, C., Osborn, J., Hopper, N., and Singer, T. (2002). "Market Trends in the U.S. ESCO Industry: Results from the NAESCO Database Project." LBNL-49601. Lawrence Berkeley National Laboratory, Berkeley, CA.

Osborn, J., Goldman, C., Hopper, N., and Singer, T. (2002). Assessing U.S. ESCO industry performance and trends: Results from the NAESCO database project. *In* "Proceedings of the 2002 ACEEE Summer Study," Vol. **5**, pp. 233–246. American Council for an Energy-Efficient Economy, Washington, DC.

Shippee, G. (1996). Energy service companies: New trends in a changing industry. *In* "Proceedings of the 1996 ACEEE Summer Study," Vol. **10**, pp. 153–156. American Council for an Energy-Efficient Economy, Washington, DC.

U.S. Department of Energy (2000). "International Performance Measurement and Verification Protocol." U.S. Department of Energy, Washington, DC.

Vine, E., Murakoshi, C., and Nakagami, H. (1998). International ESCO business opportunities and challenges: A Japanese case study. *Energy Int. J.* **23,** 439–447.

Vine, E., Nakagami, H., and Murakoshi, C. (1999). The evolution of the U.S. energy service company (ESCO) industry: From ESCO to super ESCO. *Energy Int. J.* **24,** 479–492.

Entrainment and Impingement of Organisms in Power Plant Cooling

TODD P. CALLAGHAN
Massachusetts Office of Coastal Zone Management
Boston, Massachusetts, United States

1. **Importance of Entrainment and Impingement**
2. **Collecting Entrainment and Impingement Data**
3. **Sources of Mortality to Entrained and Impinged Organisms**
4. **Quantifying Impacts to Local Populations of Organisms**
5. **Reducing Impacts**

Glossary

air cooling The process by which radiator-type coils containing steam transfer heat to air passing over the coils; water is added to the cooling system only for periodic maintenance and cleaning (~200,000 gallons/day).

closed-cycle cooling The process by which cooling water cycles through a generating facility, absorbs heat from steam passing through a condenser, and then evaporates the heat to the atmosphere through cooling towers before it recycles through the facility; water is added to the cooling system to replace that lost by evaporation (~8 million gallons/day at a large facility).

condenser An apparatus used to convert steam coming from electricity-generating turbines back to water before the water is discharged from or recycled through a facility.

cooling water intake system (CWIS) The combination of intake cribs, screens, pressurized washes, debris and fish return sluices, and circulating water pumps used to withdraw cooling water.

delta T The temperature change of water passing through a cooling water system (delta T = discharge temperature − intake temperature).

diurnal cycle A response to daily changes in light intensity (e.g., the daily movement of aquatic organisms up and down in the water).

entrainment The uptake of small organisms or life stages during the process of water withdrawal.

fish return system An apparatus consisting of a pressurized wash that removes impinged fish from traveling screens and a flume that conveys the fish back to the water body from which they were removed.

impingement The entrapment of organisms on intake screens used to prevent debris from entering cooling water intake structures.

once-through cooling The process of pumping water across a heat exchanger to condense steam from electricity-generating turbines; the heat-laden water is then discharged back to the surface water source, and the condensed steam is recycled to the facility's boilers (millions to billions of gallons of water/day are required).

traveling screen A mechanical device used to exclude organisms and debris from entering pump systems and condensers; screens are hung in front of a cooling water intake, partially above the water level, and are designed to rotate vertically to allow impinged organisms and debris to be removed when sprayed with pressurized wash water.

young-of-the-year A juvenile fish observed in the same calendar year as its hatching.

Many electricity-generating facilities withdraw cooling water from surface waters to absorb and transport waste heat. Incidental to the withdrawal of water is the concurrent uptake or entrapment of the organisms that live in those waters. Entrainment refers to the uptake of small organisms or life stages (e.g., eggs, larvae) during the process of water withdrawal, whereas impingement is the entrapment of larger organisms on screens used to prevent debris from entering cooling water intake structures.

1. IMPORTANCE OF ENTRAINMENT AND IMPINGEMENT

Entrainment and impingement susceptibility is contingent on an organism's size relative to the mesh

 © 2004 Elsevier Inc. All rights reserved.

covering the intake (i.e., smaller organisms become entrained rather than impinged). The typical screen mesh size of a cooling water intake structures (CWIS) is 0.375 inch; therefore, entrained organisms are usually smaller than this, whereas impinged organisms are usually larger than 0.375 inch.

1.1 Factors Affecting Entrainment and Impingement Rates

Entrainment and impingement rates generally increase with increasing intake volumes and velocities unless the CWIS is equipped with a barrier to prevent organisms from approaching the CWIS. Entrainment rates are several orders of magnitude higher than impingement rates because smaller organisms and younger life stages (e.g., eggs, some larvae) are more abundant and have less capacity to overcome intake velocities compared with larger organisms. Ultimately, entrainment and impingement rates are influenced by a combination of the facility's intake flow, the location of the CWIS relative to the water body's physical characteristics, and the temporal and spatial distribution of organisms in the water body relative to intake rates.

Although entrainment and impingement can occur with any type of water withdrawal, the potential for the largest effects is at electricity-generating facilities that use once-through cooling because they typically withdraw water in significantly larger volumes than do other industries. Other cooling technologies (e.g., closed-cycle cooling, air cooling) require significantly less water; therefore, electricity-generating facilities using these technologies have less potential for impacts to resident and migrating aquatic species.

In addition to intake flow rates, the siting of power plants and their CWIS affects the incidence of entrainment and impingement (Fig. 1). Most CWIS are located in nearshore areas of lakes, rivers, estuaries, and marine coasts. Whether in fresh or salt water, these areas have great ecological importance. Their shallow waters, submerged and emergent vegetation, and sheltered inlets provide feeding, resting, spawning, and nursery habitat for a wide variety of vertebrate and invertebrate species. In addition, when rock structures such as jetties are used to protect cooling water intakes, fish with habitat preferences for these structures (e.g., cunner, tautog) experience high rates of entrainment. CWIS have also been known to impinge waterfowl, seals, turtles, and crustaceans. Not surprisingly, when power plants that use once-through cooling withdraw their water directly from the shores of these productive waters, entrainment and impingement rates can be very high. On the other hand, offshore waters are typically cooler, provide less vegetation for shelter, and may be lower in dissolved oxygen, especially in deep, slow-moving freshwater systems. Although these offshore waters also harbor many species, most of them seek shallow nearshore areas to spawn. Consequently, offshore intakes may have lower entrainment and impingement rates.

Annual schedules of cooling water withdrawal also influence rates of entrainment. Egg and larvae

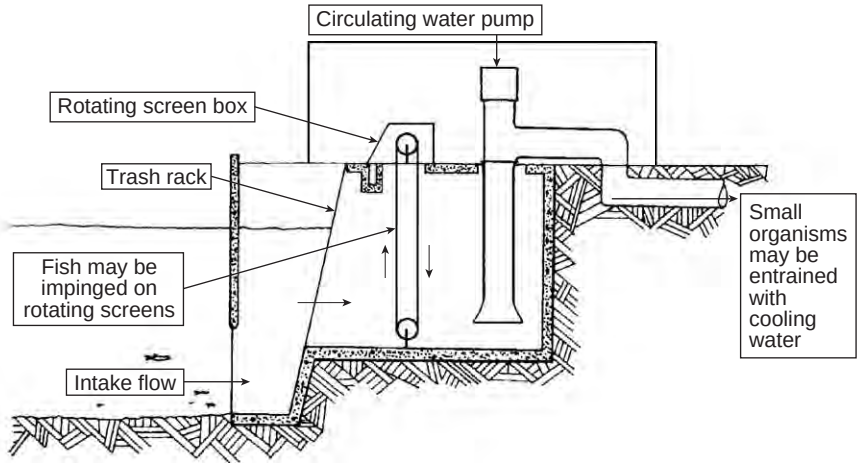


FIGURE 1 Schematic side view of a cooling water intake structure (CWIS). Circulating water pumps draw water into the CWIS. The trash rack and rotating screens keep large objects (including fish) from entering the cooling water system. Organisms that pass through the screens may be entrained into the plant's cooling system. Modified from Coutant, C., and Bevelheimer, M. (2003). Induced sweeping flows at CWIS for reducing fish impingement. *In* "Proceedings: Symposium on Cooling Water Intake Technologies to Protect Aquatic Organisms." U.S. Environmental Protection Agency, Washington, DC.

abundances are greatest in the water column during the spring when most organisms spawn. Consequently, facilities that withdraw large quantities of cooling water at this time are likely to have high rates of entrainment during the spring. Some facility managers have scheduled their annual maintenance outages (when cooling water withdrawal is minimal) in March or April to minimize entrainment during this biologically active period.

1.2 Importance of Measuring Entrainment and Impingement

Quantifying entrainment and impingement rates is an important step in understanding the losses to biological populations that can occur due to cooling water withdrawal. Large power plants can entrain and impinge a significant proportion of young organisms or life stages that would otherwise recruit to local biological populations. Although predators, meteorological events, fishing, and disease all contribute to mortality rates, the additional population reduction caused by power plants may contribute to the destabilization of natural oscillations in population abundances, potentially resulting in significant long-term reductions. Because entrainment and impingement rates are controllable variables, knowing the magnitude of losses due to power plants can lead to a variety of management decisions aimed at minimizing losses and maintaining biological populations.

In the United States, the Environmental Protection Agency (EPA) and state regulatory bodies are charged with protecting biological resources. The Federal Water Pollution Control Act of 1977 and its amendments (known collectively as the Clean Water Act) require that the location, design, construction, and capacity of a CWIS reflect the best technology available for minimizing adverse environmental impact. Because the law does not clearly define the term "adverse environmental impact," vigorous and sometimes contentious debates have emerged as to the impact of cooling water withdrawal on local populations of aquatic organisms and how government agencies should regulate these impacts. The implementation of the Clean Water Act and other regulations may require facility owners to upgrade to withdrawal technologies with lower impacts (e.g., closed-cycle, air cooling). These technological upgrades or operational changes may cost tens or hundreds of millions of dollars. In some cases, implementation costs may prevent the construction of new intakes altogether. Under Section 316(b) of the Clean Water Act, a company faced with attaining regulatory requirements can assert that the investment is "wholly disproportionate" to the environmental benefits to be gained from the modifications. Furthermore, some industry representatives contend that significant upgrades will affect the availability and price of energy across regional grid structures. It is the responsibility of the EPA to evaluate the balance of information and to determine whether technological upgrades are warranted. Clearly, managing entrainment and impingement impacts has important consequences for natural resources as well as for the energy industry.

2. COLLECTING ENTRAINMENT AND IMPINGEMENT DATA

With such important economic and resource protection decisions dependent on entrainment and impingement values, quantifying these parameters reliably is critical. However, there is no standard methodology or procedure for collecting entrainment and impingement data. Methods for estimating entrainment and impingement are determined on a site-by-site basis by the characteristics of the facility's CWIS, the flow rate through the facility, the life history of the species at risk, the characteristics of the water body, the priorities of the regulatory community, and the costs of the various methodologies.

2.1 Entrainment Sampling

Because of the patchy distribution of larvae and eggs, achieving the sampling effort necessary to calculate a reliable estimate of abundances can be challenging. To quantify entrainment impacts, estimates of the number of larvae and eggs going through the plant must be compared with estimates of the number of larvae and eggs of each species in the source water body adjacent to the plant. These ratios will provide species-specific estimates of the cropping of biological populations by the power plant.

Entrainment rates are highly dependent on the spatial and temporal distribution of larvae and eggs. Water currents, diurnal cycles, time from hatching, and species type all can affect entrainment estimates by influencing densities of larvae and eggs adjacent to intakes. In addition, flow rate and CWIS configuration influence the proportion of larvae and eggs adjacent to the intake that are drawn into a facility's cooling system. For example, increasing flow rate usually increases entrainment rate; however, larvae may be

able to avoid becoming entrained by avoiding the part of the water column from which a plant withdraws its cooling water. Because of the multitude of influences on entrainment rates, data collections need to occur frequently enough so that they are representative of the various combinations of conditions present in the source water body. Moreover, interannual variability is often great enough to warrant replication across years as well as within years.

Early efforts to quantify the mortality of larvae within cooling water systems used conical nets suspended in both the intake structures and discharge plumes of electricity-generating facilities. (It should be noted that larval sampling techniques also capture floating eggs; thus, entrainment of eggs and larvae are often assessed simultaneously.) However, one of the problems with this method is that the survival of larvae is influenced by the velocity of water flowing through the net, with higher velocities leading to increased mortality. Higher velocities also increase the risk of extrusion through the net and increase the limitations of the net to catch larvae, resulting in an incomplete sample. Furthermore, discharge plumes often have much greater velocities than do intakes. As a consequence, larvae collected in discharges may have higher mortality due to the force of the discharge water trapping them in collection nets. In short, traditional net sampling can add another source of mortality when only the mortality caused by the cooling system is of interest.

To reduce sampling-related mortality, some researchers have used a flume-like sampling station, floated in front of the intake or discharge, onto which water can be pumped. Unlike net sampling, a larval table equalizes sampling stresses for both intake and discharge samples by withdrawing water at a constant rate in both areas. However, the pump used to bring larvae onto the sampling table can cause a significant amount of mortality to the larvae.

To avoid pump-related mortality, samples may be taken using a pump that is downstream of the sampling screen. In this way, water and entrained organisms are pulled rather than pushed onto the sampler and its screens; therefore, no organisms have to go through the pump system. Other sampling efforts have eliminated the pump altogether, using the velocity of water in the discharge to bring water into the floating sampling station. However, this method does not work in intake structures where the velocity of water is not great enough to force water onto the floating sampling station.

Although innovations in sampling techniques have produced methods that increase the reliability of larval mortality estimates, conical nets and pumps are still commonly used due to cost or design issues associated with sampling in intakes and discharges. For example, only one of the nine entrainment studies conducted for power plants along the coast of Massachusetts between 1994 and 2003 reported the use of a larval table, whereas all of the others used conical net sampling.

2.2 Impingement Sampling

Impingement is easier to measure than is entrainment because the presence or absence of larger organisms is more obvious and, therefore, easier to quantify. Impingement samples can be collected directly from traveling screens or from collection pens built into a fish return system. Often, impingement rates are assessed for only part of a 24-h period and long-term rates are extrapolated from the short-term assessment intervals. Impingement rates are usually reported as number of fish per year or number of fish per volume of flow through the power plant.

The most frequently impinged fish are usually young-of-the-year fish, although older juveniles and adults are also impinged. Schooling juvenile fish (e.g., menhaden) often follow shoreline margins and can become impinged in large numbers, especially during the late summer and fall. Impingement rates for juvenile schooling fish can exceed 1000 fish per hour and tens of thousands weekly.

3. SOURCES OF MORTALITY TO ENTRAINED AND IMPINGED ORGANISMS

Entrained and impinged organisms can incur several types of mortality associated with a CWIS. The risks of mortality arise due to physical, chemical, or thermal stresses within the cooling water system. In addition, there is the possibility of latent mortality after an organism has been discharged from the cooling water system.

3.1 Physical Stress

Physical stresses result in direct damage to the tissues of entrained organisms as they pass through the pumps and pipes of a cooling system. Within the system, they are subject to abrasion, compression, and shear stress. Physical impacts to impinged organisms may include suffocation while pressed

against the intake screens and tissue damage from pressurized screen washes and the velocity of the return sluice. Some sluices have been known to drop fish several vertical feet into the receiving water, whereas others have sent fish to rocky areas that are dry at low water. Such poorly designed sluices may stun or injure fish, leaving them more susceptible to predation. For example, birds have been known to frequent the outfalls of fish return systems, plucking the returned fish from the water. Furthermore, some sluices deliver fish directly to the discharge channel, where they experience turbulence as well as elevated temperatures that may contribute to mortality. In addition, sluices that return fish too close to intake structures may contribute to reimpingement, especially if the fish are stunned by the physical stress of impingement, high-pressure spray, and a turbulent return sluice.

3.2 Chemical Stress

Chemical mortality results from deleterious physiological effects induced by chemicals added to the cooling water as it flows through the facility. Biocides (typically chlorine) are used to keep the cooling water system free of fouling organisms such as algae, barnacles, and mussels. Biocides injected directly into cooling water will kill entrained larvae and eggs. Ultimately, these biocides are discharged to the receiving water body and, therefore, may also affect the organisms in the vicinity of the outfall. If biocides are used to keep traveling screens free of fouling organisms, impinged fish can also be dosed with toxic chemicals.

3.3 Thermal Stress

Thermal stress to entrained organisms occurs as the water that surrounds them absorbs heat from condenser tubes. Excessive temperatures cause the breakdown of proteins and can result in developmental problems or death. The delta T of a plant directly influences entrainment mortality. Minimizing delta T will increase the rate of survival because most organisms have the physiological ability to withstand small increases in temperature. However, for all organisms, there is a temperature threshold above which acclimation to higher temperatures will not occur and mortality will follow. The duration of exposure to elevated temperatures also affects survival. Therefore, the risk of mortality to organisms that have survived being entrained will increase as they spend increasing amounts of time in a heated discharge or if they are reentrained into the power plant.

Increased temperatures may affect impinged organisms if a facility is designed to clean its screens by backflushing, which involves using heated water to dislodge debris and fouling organisms attached to the intake screens. This method can also physically damage impinged fish while subjecting them to excessive temperatures. In addition, impinged organisms can be thermally stressed if the return sluice deposits them in the plant's thermal effluent.

3.4 Latent Mortality

The combination of physical, chemical, and thermal stresses experienced during entrainment or impingement can make organisms more susceptible to pathogens, infections, or predation after they are returned to the water body. The mortality associated with these combined effects is called latent mortality. Latent mortality is difficult to quantify because long-term effects might not manifest themselves during the time frames of most studies. However, some researchers have attempted to quantify latent mortality by holding recently entrained larvae and impinged fish in tanks and monitoring their survivorship. Still, in practice, most assessments conservatively assume 100% through-plant mortality in the interest of expediency and cost savings, obviating the need for a latent mortality estimate.

4. QUANTIFYING IMPACTS TO LOCAL POPULATIONS OF ORGANISMS

The simplest method of evaluating cooling water withdrawal effects is to estimate the annual absolute losses of all life stages of species entrained or impinged. The magnitude of the impact can then be assessed relative to management criteria or historical trends in species abundances. Beyond calculating absolute losses, there are a number of methods for quantifying power plant impacts due to entrainment and impingement. These include equivalent adult analysis, calculating production forgone, calculating indirect losses to predators due to mortality to forage species, calculating conditional mortality, comparing equivalent adults or production forgone to commercial landings, and modeling populations.

4.1 Entrainment Impact Assessments

The equivalent adult method of assessing entrainment and impingement effects provides managers with the number of individuals (eggs, larvae, or juveniles) that would have survived to reproductive age if they had incurred only those mortality pressures independent of the generating facility. Equivalent adult losses can be calculated for those species whose life stage-specific mortality rates are known. Typically, these values are known for only a select few commercially or recreationally important fish species.

A second means of calculating impacts is called the production forgone method. Production forgone is similar to calculating equivalent adults except that it is used to determine the biomass (measured in pounds or kilograms) that would have been produced if the entrained and impinged organisms had experienced sources of mortality only in the absence of the plant. Equivalent adult analysis is usually applied to species that are part of a commercial or recreational fishery, whereas production forgone is usually used to measure impacts to nonfished species.

Another impact assessment method takes into account the indirect losses caused by the mortality of species that are prey to species that are part of a managed fishery. This assessment occurs in three steps. First, production forgone is calculated for the prey species. Next, for each predator species in the fishery, production forgone of their prey is multiplied by a food conversion efficiency factor that accounts for the biomass of prey consumed by the predators that results in an increase in predator biomass. Finally, the loss of predator biomass due to entrainment of young life stages of their prey is calculated by apportioning the biomass of prey to individual predator species based on the specific proportions of each prey species in a predator's diet and the relative biomass of each predator in the area of interest. In effect, this method converts biomass of prey species to biomass of predator species that have known commercial or recreational value.

Yet another assessment method, and perhaps the most controversial because it is the most subjective, is to determine conditional mortality rates. Conditional mortality is the fractional reduction of a population of organisms due only to entrainment and impingement. To estimate conditional mortality, one must (1) define the boundaries of the local populations of organisms entrained and impinged, (2) estimate the abundance of organisms in these populations, and (3) compare entrainment and impingement data for individual species with population abundances adjacent to the intake. In the absence of genetic and behavioral information that could truly bound a population of interbreeding organisms, impact assessments rely on subjective semi-arbitrary boundaries defined by geographical features, hydrodynamics, and policy decisions by resource agencies. Increasing population boundaries, and thus a population's size, decreases the magnitude of impacts. Given that the magnitude of impacts can lead to expensive plant upgrades, spirited discussions can surround the definition of population boundaries.

A similar approach has been used to determine the percentage of larvae flowing past a plant that is entrained. In this assessment, the rate of larvae flowing through a plant (i.e., the density in the discharge multiplied by the plant's flow rate) is divided by the flux of larvae flowing past the intake structure (i.e., the density in the water body multiplied by the net current flow). This method provides an estimate of the plant's ability to affect the pool of larvae in its source water. The reliability of this estimate of impact is dependent on empirical data and accurate measurements of flow rate and larval density. In addition, because ocean currents and some rivers can flow in opposite directions depending on the tide, the opportunity for larvae to be entrained is dependent on how many times they flow back and forth in front of a CWIS.

Some impact assessments have compared entrainment and impingement losses with annual commercial landings of affected fish species. However, this estimate of impact can be misleading if the commercial landings data are from an area that is larger than the area of interest immediately adjacent to an intake structure. Another reason why commercial landings are not a good point of comparison is that they are influenced by fishing effort, management policies restricting effort, and the market price of various fish species. Because of the variability in these factors, commercial landings are not a reliable measure of species abundance.

After bounding the populations of interest and estimating current impacts, further analysis may involve predicting future impacts to populations using mathematical models. Models may be used to project how a population will respond to various operating or mitigation scenarios.

In models, the distribution of larvae and eggs is often represented as uniform, with their densities averaged across a section of water column. This simplification keeps model projections manageable while decreasing run time and financial costs.

However, in reality, the distribution of larvae and eggs is heterogeneous and the risk of higher density patches of organisms becoming entrained will vary over time. To take advantage of the simplicity of projection models and account for the real-world patchiness of eggs and larvae, cross-sectional averages of organism densities determined by models can be converted to more realistic densities by multiplying by a ratio called the withdrawal factor (W). W is the ratio of the density of entrained organisms in the intake structure to the density of entrainable organisms found in the adjacent water body.

Some models include a recirculation factor to account for the proportion of entrained organisms that reenter the CWIS because the discharge is close enough to the intake to reentrain discharge water. Recirculation can be determined through modeling or dye studies of the movement of the discharge plume. The amount of recirculation is site specific and may be as great as 25%. Incorporating recirculation of flow into a model tends to decrease estimates of impacts because a portion of the organisms in the cooling water is assumed to have already been killed by the plant's cooling system.

When the goal of an impact assessment is to quantify the specific sources of mortality to entrained organisms, also known as the entrainment mortality factor (f), the individual mortality rates due to the physical, chemical, or thermal stresses associated with passing through the cooling water system are estimated. Estimating f entails separating out mortality incurred through the sampling process (i.e., physical stresses due to pumps, nets, and handling time) from the mortality caused solely by the cooling water system. Because there are so many potential sources of mortality once eggs and larvae are in a cooling system, and because quantifying them reliably to the satisfaction of regulatory agencies can be an onerous and costly task, many impact assessments will conservatively estimate that 100% of all entrained eggs and larvae suffer mortality.

When the 100% entrainment mortality assumption is not used, through-plant mortality (f) is important to modeling operational procedures needed to reduce entrainment mortality because trade-offs exist between delta T and flow rate. When the flow of water passing through a cooling water system decreases, the relative amount of heat per volume of cooling water increases, thereby increasing delta T. Under certain operational scenarios, the potential increase in mortality to entrained larvae and eggs associated with an increased delta T can offset reductions in entrainment achieved by flow reductions. However, when through-plant mortality is assumed to be 100%, flow rate becomes the most important factor because delta T has no effect on impact assessment. A further consideration is that there is also a trade-off between flow and thermal output to the receiving waters. Increasing flow through the plant can decrease delta T and reduce thermal impacts within the cooling system and in the receiving water.

4.2 Impingement Impact Assessments

Impacts due to impingement are relatively easier to quantify than are entrainment impacts. Representative samples of impinged organisms can be collected directly from traveling screens by special sampling trays or can be sampled from the return sluiceway. Although fish are usually the most frequently collected organisms, crustaceans, turtles, birds, and seals can also be impinged. However, in practice, most impact assessments focus on the number of fish impinged. In the overall assessment of a plant's impact, losses due to impingement are added to equivalent adult, production forgone, and conditional mortality estimates.

5. REDUCING IMPACTS

The most effective way of reducing impacts caused by cooling water withdrawal is to avoid the use of cooling water altogether by initially designing or upgrading a facility to utilize air cooling. However, in many cases, this is cost-prohibitive. Alternatively, closed-cycle cooling costs significantly less than does air-cooling and requires substantially less water than do once-through cooling systems; therefore, it has lower entrainment and impingement impacts. When air cooling or closed-cycle cooling systems are not an option due to design, siting, or cost constraints, power plant owners can reduce entrainment and impingement impacts through low-impact siting of a new CWIS, operational changes, technological upgrades, and/or biological mitigation. The efficacy of each of these methods has been debated within the CWIS permitting process and has been shown to be dependent on the site-specific characteristics of an individual facility and the distribution of organisms around its intake structure(s). Finally, the implementation of new and existing regulations can potentially reduce entrainment and impingement impacts on a regional scale.

5.1 CWIS Siting

When the use of once-through cooling is the only option, one way of minimizing impacts is through the careful siting of new intake structures after consideration of the temporal and spatial distribution of organisms in the source water body. Identification and avoidance of spawning runs, egg-laying sites, hydrodynamically low-energy or sheltered areas where larvae may collect, juvenile nurseries, feeding areas, and overwintering refuges can greatly reduce potential impacts. To achieve this, some intakes may be resited to offshore locations to remove them from the most biologically productive habitats.

Intake structures raised slightly above the bottom will aid in keeping bottom-dwelling organisms from entering the intakes (but it will do little to deter planktonic or swimming organisms). Although the careful siting of intakes outside the most productive habitats will reduce impacts, there will always be some level of impacts due to CWIS unless some physical characteristic of the water (e.g., consistently low dissolved oxygen) prevents organisms from living there.

5.2 Operational Changes

Changes to operational procedures at the facility can also reduce impacts. For example, flow reductions may reduce impacts because entrainment and impingement rates are often directly correlated with flow rates. One way of reducing flow is to use variable-speed pumps (instead of on/off pumps) to reduce water intake during months when biological activity is highest. Another method is to operate in what is called "piggyback" mode, where the discharge from one unit is used to cool another unit. This cooling water recycling measure increases the risks to organisms that are initially entrained (because they may go through the cooling system several times) but decreases the overall amount of new cooling water needed, thereby decreasing impacts. Another operational procedure that can reduce impingement impacts is the continuous rotation of screens in front of the intake. Typically, screens are rotated only when excessive blockage of the screens (by vegetation, debris, or organisms) decreases intake rates beyond the threshold of condenser efficiency. Continuous rotation of screens during months when impingement is known or expected to be highest can reduce the amount of time that impinged organisms spend on the screens, thereby increasing the probability of survival if there is an adequate return sluice.

5.3 Technological Upgrades

In addition to carefully planned CWIS siting and operational protocols, technological modifications and additions provide an opportunity for reducing impacts.

The upgrade to modified traveling screens with a well-designed fish return system can enhance survivorship of many impinged organisms. Improvements to traveling screens and fish return systems to minimize the physical stress on fish before they return to the receiving water body include modified fish buckets on the screens and low-pressure spray washes, variable speed drives for rotating the screens (3–30 feet/min), and separate fish and debris return troughs (Fig. 2). Furthermore, designing return sluices so that they deliver fish back into the receiving water body at all water levels (i.e., not allowing returned fish to drop on the shore at low water levels), and far enough away from the intake so as not to be reimpinged, may also increase survival rates of returned fish.

When biocides are necessary to reduce biofouling on intake screens, the location of biocide injection

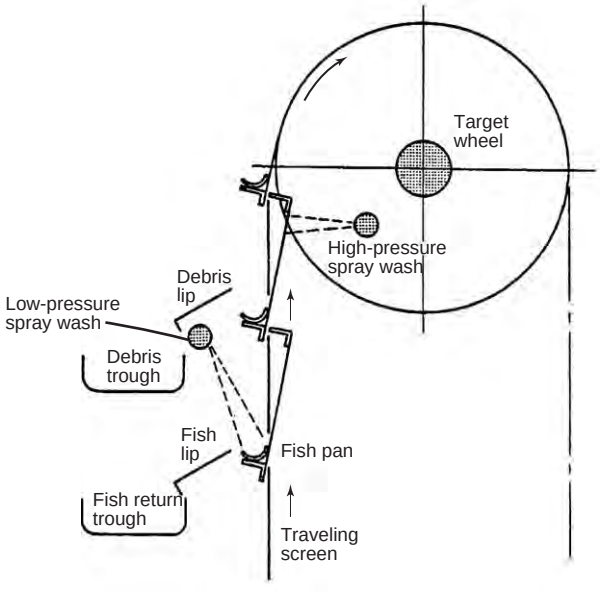


FIGURE 2 Schematic side view of a modified traveling screen with separate fish and debris troughs and a low-pressure spray to remove impinged fish. From Johnson, I. C. (2003). Fish return system efficacy and monitoring studies for JEA's Northside Generating Station. *In* "Proceedings: Symposium on Cooling Water Intake Technologies to Protect Aquatic Organisms." U.S. Environmental Protection Agency, Washington, DC.

can be important to the survival of impinged organisms. Positioning the biocide injectors so that they do not dose impinged organisms (i.e., aiming injectors toward the descending portion of the traveling screen) may help to increase survival rates. Furthermore, using targeted biocides (i.e., those that kill only biofouling organisms and not fish) will also increase the survival of impinged organisms.

Other technological upgrades include physical and behavioral barriers. Filter fabric curtains, fixed panel screens, or wedge wire screens placed in front of the CWIS create a physical barrier between organisms in the water body and the intake structure (Fig. 3) These barriers, if installed and maintained properly, reduce intake velocity and can prevent impingement and reduce entrainment (although the smallest eggs and larvae may still pass through the barrier and into an intake). Other systems that use angled screens or vertical panels placed at 90 degrees to the flow of water create flows that divert fish to a bypass on the outside of the screens that prevents them from entering the cooling system (Fig. 4). These systems have been shown to reduce impingement but not entrainment. In a similar way, localized turbulent flows generated at points near an intake can "sweep" fish away from intakes (Fig. 5). This method has been shown to be successful in preventing impingement of certain species of fish.

Placing the intake structure under gravel or behind a porous cobble dike is another upgrade that can potentially reduce entrainment and impingement if the intake velocity across the gravel or cobble is less than it would be across an unprotected intake. However, the

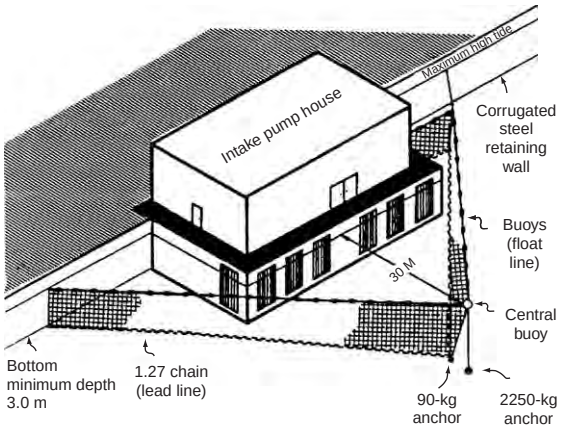


FIGURE 3 Schematic of a barrier net for reducing impingement at Bowline power plant in New York. From U.S. Environmental Protection Agency. (1994). "Preliminary Regulatory Development Section 316(b) of the Clean Water Act Background Paper Number 3: Cooling Water Intake Technologies." EPA, Washington, DC.

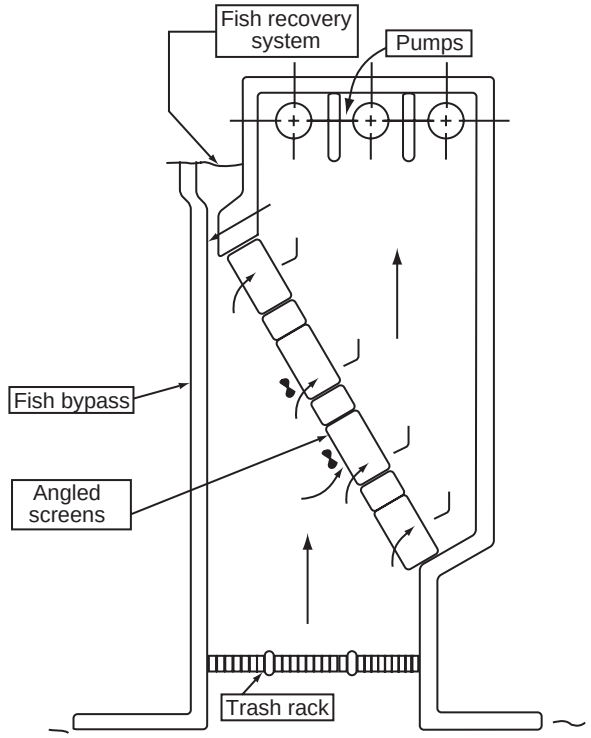


FIGURE 4 Schematic of angled screens located upstream from a cooling water intake to redirect fish toward a bypass around the intake structure. From U.S. Environmental Agency. (1994). "Preliminary Regulatory Development Section 316(b) of the Clean Water Act Background Paper Number 3: Cooling Water Intake Technologies." EPA, Washington, DC.

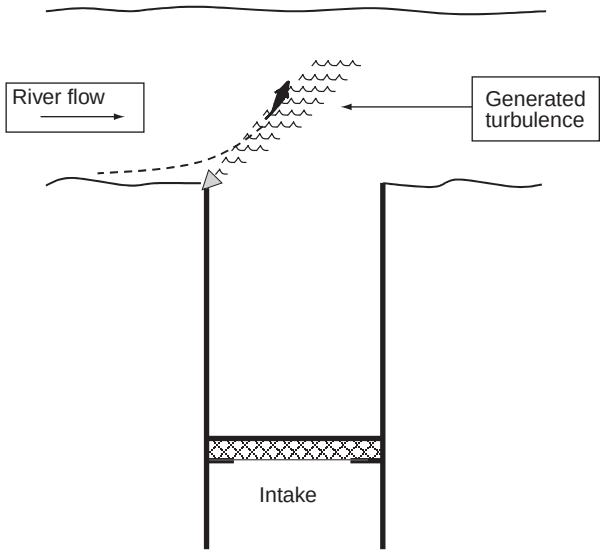


FIGURE 5 Diagram of how induced turbulent flows can sweep fish away from an intake structure. From Coutant, C., and Bevelheimer, M. (2003). Induced sweeping flows at CWIS for reducing fish impingement. *In* "Proceedings: Symposium on Cooling Water Intake Technologies to Protect Aquatic Organisms." U.S. Environmental Protection Agency, Washington, DC.

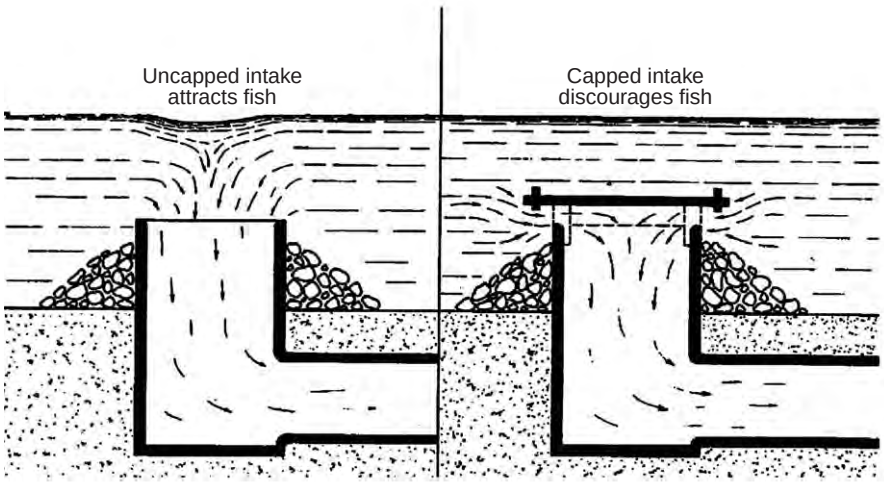


FIGURE 6 Schematic of the difference in intake velocity between capped and uncapped cooling water intakes. From U.S. Environmental Protection Agency. (1994). "Preliminary Regulatory Development Section 316(b) of the Clean Water Act Background Paper Number 3: Cooling Water Intake Technologies." EPA, Washington, DC.

efficacy of an under-gravel or porous dike barrier may decrease due to clogging by debris and sediment or fouling by colonizing aquatic organisms.

Other methods of impact reduction are behavioral barrier systems. One behavioral system that works on the premise that fish avoid rapid changes in horizontal flow is called a velocity cap. A velocity cap is placed on an offshore vertical intake and converts vertical flow into horizontal flow (Fig. 6). Like the panel and angled screen systems, this device has only been shown to reduce impingement but not entrainment impacts, and it is more effective for some species than for others. Acoustical or optical (strobe light or air bubble) deterrents have also been used as behavioral barriers to keep fish from approaching CWIS, but the efficacy of these techniques has been shown to be limited and species specific.

5.4 Biological Mitigation

An innovative but still unproven method for reducing impacts is biological mitigation. The theory behind biological mitigation is to accept that a certain number of losses will occur at a given facility and that these losses can be balanced by adding more individuals to the affected biological populations.

Some biological mitigation strategies propose to conserve or restore habitat so as to increase the abundance of certain life stages of affected populations. To facilitate this process, the EPA created a "habitat-based replacement cost" method of valuing biological resources lost to entrainment and impingement. This valuation system identifies the costs of habitat restoration required to replace biological resources lost to entrainment and impingement. One of the strengths of this method is that it recognizes that all organisms, and not just those that are recreationally or commercially important, are valuable to an aquatic ecosystem. An intensive search for habitat restoration opportunities for each affected species and a system of prioritizing the options based on their expected benefits to the affected species is required. Once implemented, only intensive, long-term, and carefully controlled studies can determine whether habitat remediation has led to increases in the abundance of species life stages affected by power plants.

Another form of biological mitigation is restocking. In restocking plans, the sum total of equivalent adult fish entrained plus the number of adult fish that have been impinged is replaced by hatchery-raised fish of the same species. One of the challenges with this method is to ensure that the hatchery-raised fish have the same genetic makeup and ecologically important behavioral responses (e.g., habitat selection, prey capture, predator avoidance) as do the fish they are meant to replace. An additional consideration is that restocking often occurs without respect to restocking the same life stages that are experiencing the greatest impacts. For example, juvenile fish grown in a hatchery and released as mitigation for entrainment impacts will not have the same ecological role as will the larvae they are meant to replace.

Biological mitigation is a field that requires more research before its efficacy can be ascertained.

5.5 New Regulations to Reduce Impacts

At the regulatory level, the EPA has established requirements to reduce entrainment and impingement impacts at new facilities using once-through cooling. On December 18, 2001, it issued its "final rule" implementing Section 316(b) of the Clean Water Act for new facilities. The national requirements for new facilities establish a two-track system. Under track I, facilities using more than 10 million gallons of intake water per day must meet capacity, velocity, and capacity- and location-based proportional flow requirements. Facilities using more than 2 million but less than 10 million gallons of intake water per day do not have to meet the capacity limitation but must use technologies to reduce entrainment. More specifically, the track I standards require the following:

- Intake flow must be commensurate with that of a closed-cycle system (although facilities with 2–10 million gallons/day intake flow are exempt).
- The maximum through-screen intake velocity is 0.5 feet per second.
- If an intake is sited in a freshwater river or stream, flow must be less than 5% of the annual mean flow.
- If an intake is sited in a lake or reservoir, withdrawal from a cooling water intake must not disrupt the natural thermal stratification of the source water body.
- If an intake is sited in a tidal river or an estuary, withdrawal can be no greater than 1% of the volume of the water column within the area centered around the intake, with the diameter defined by the distance of one tidal excursion at mean low water.

Under track II, a facility has the option to demonstrate that alternative technologies or operations chosen by the manager of the facility will reduce entrainment and impingement mortality to a level comparable to that required under track I. Also under track II, flow levels are not limited to those commensurate with closed-cycle cooling, and intake velocity does not need to be less than 0.5 feet per second; however, the other withdrawal constraints are the same. In addition, facility operators choosing track II need to undertake a comprehensive biological evaluation demonstrating that the chosen technologies and operations have an adverse environmental impact level commensurate with that achievable under track I. The EPA has been evaluating requirements to reduce entrainment and impingement effects at existing facilities that use cooling water in order to meet the goals of track II.

SEE ALSO THE FOLLOWING ARTICLES

Aquaculture and Energy Use • Electricity, Environmental Impacts of • Fisheries and Energy Use • Hydropower, Environmental Impact of • Thermal Pollution • Wetlands: Impacts of Energy Development in the Mississippi Delta

Further Reading

Barnthouse, L. W., Klauda, R. J., Vaughan, D. S., and Kendall, R. L. (eds.). (1988). "Science, Law, and Hudson River Power Plants: A case study in environmental impact assessment monograph 4." American Fisheries Society, Bethesda, MD.

Boreman, J., Goodyear, C. P., and Christensen, S. W. (1981). An empirical methodology for estimating entrainment losses at power plants sited on estuaries. *Trans. Am. Fish. Soc.* **110,** 255–262.

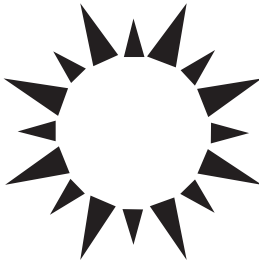
Environmental Protection Agency. (2001a). Efficacy of cooling water intake structure technologies. www.epa.gov/waterscience/316b/technical/ch5.pdf.

Environmental Protection Agency. (2001b). Proposed Section 316(b) phase II existing facilities rule case study analysis. www.epa.gov/waterscience/316b/casestudy/.

Van Winkle, W., Christensen, S. W., and Suffern, J. S. (1979). "Incorporation of Sublethal Effects and Indirect Mortality in Modeling Population-Level Impacts of a Stress, with an Example Involving Power-Plant Entrainment and Striped Bass," ORNL/NUREG/TM-288. Oak Ridge National Laboratory, Oak Ridge, TN.

Entropy

CARLO BIANCIARDI and SERGIO ULGIATI
University of Siena
Siena, Italy



1. **Introduction**
2. **Entropy as a Thermodynamic State Function**
3. **Statistical Meaning of Entropy and Boltzmann's H-Theorem**
4. **Entropy and Information Theory**
5. **The Free-Energy Concept**
6. **Conclusion**

Glossary

energy Energy is usually defined as the capacity for doing mechanical work—that is, displacing an object from a position to another by means of the application of a force. This textbook definition is sometimes criticized for being reductive and not always easily applicable to processes other than purely physical ones. A broader definition for energy could be the ability to drive a system transformation, which clearly includes all kinds of physical, chemical, and biological transformations.

enthalpy Enthalpy, H, is a thermodynamic property expressing the total energy content of a system. It is defined as $H = U + pV$, where U is the internal energy of the system (any kind of energy form of its molecules) and pV is the system's energy in the form of expansion work (in fact pV, the product of pressure and volume, has the physical dimensions of energy). In a chemical reaction carried out in the atmosphere the pressure remains constant and the enthalpy of reaction (ΔH) is equal to $\Delta H = \Delta U + p\Delta V$. Enthalpy is a state function, because it only depends on the initial and final states of the system, no matter how the process develops.

macroscopic/microscopic Any system is composed of parts. Parts contribute to the functioning of the whole. Matter is composed of molecules, an economy is made up of production sectors and firms, for example. The dynamics of the whole system may be or appear to be different from the behavior of the component parts. A car may be stopping in a parking lot, yet its molecules possess a chaotic motion, which cannot be seen when we look at the car. A national economy may be in good health, on average, yet some sectors or local firms may face a difficult survival. Macroscopic therefore concerns the system's dynamics, considered as a whole, while microscopic applies to the behavior of the lower scale components.

perfect gas A perfect or ideal gas is a gas the molecules of which are infinitely small and exert no force on each other. This means that their motion is completely chaotic and does not depend on any mutual interaction, but only on completely elastic collisions, which cause a transfer of energy from a molecule to another and a change of their direction and velocity.

probability The probability of an event is the ratio of the number of times the event occurs to a large number of trials that take place. The number of trials (all events that may occur, including those different from the ones fitting the request) must be very large for the probability to be reliable. As an example, assume we cast a die. If the die is perfectly constructed, the probability of a five is 1 out of 6 possible different castings. However, in a real case, we must cast it many times to ensure that a five occurs out of the other possible events in a 1:6 ratio.

reversible/irreversible A reversible change in a process is a change that can be reversed by an infinitesimal modification of any driving force acting on the system. The key word infinitesimal sharpens the everyday meaning of the word reversible as something that can change direction. In other words, the state of the system should always be infinitesimally close to the equilibrium state and the process be described as a succession of quasi-equilibrium states. A system is in equilibrium with its surroundings if an infinitesimal change in the conditions in opposite directions results in opposite changes in its state. It can be proved that a system does maximum work when it is working reversibly. Real processes are irreversible, therefore there is always a loss of potential work due to chemical and physical irreversibility.

self-organization All systems in nature self-organize—that is, use the available energy and matter inputs in order to achieve higher levels of structure and functioning under natural selection constraints. In so doing, they may maintain their present level of complexity or become more and more complex by also increasing the number of components, functions, and interactions.

statistic The science dealing with sampling and recording events, which occur in a different way in dependence of a given parameter (time, temperature, etc.). It estimates

 © 2004 Elsevier Inc. All rights reserved.

events by recording and analyzing a given behavior verified by an appropriately large number of random samplings. The statistical analysis may lead to hypotheses about the behavior of a given set of events.

temperature Temperature is a property of an object that tells us the direction of the flow of heat. If the heat flows from A to B, we can say that the temperature of the body A is higher than that of the body B. In some a way it measures the concentration of heat within matter and can be linked to the average kinetic energy of the molecules in a body. A thermometric scale (centigrade or Celsius scale of temperature) can be defined by dividing by 100 the temperature interval between the temperature of melting ice (0°C) and the temperature of boiling water (100°C) at sea level. Other existing scales depend on the reference parameters chosen as the zero level. A scale of absolute temperatures (Kelvin) defined as described in this article is linked to the Celsius temperature, t, by means of the relationship: $T\ (K) = t\ (^{\circ}C) + 273.15$. A unit absolute degree (K) coincides with a unit centigrade degree (°C).

work The transfer of energy occurring when a force is applied to a body that is moving in such a way that the force has a component in the direction of the body motion. It is equal to the line integral of the force over the path taken by the body, $\int F \times ds$. Its meaning is different from the usual life work, often associated with activities not always involving visible motion. Physical work requires a body motion under the action of the force.

Entropy, a measure of energy and resource degradation, is one of the basic concepts of thermodynamics, widely used in almost all sciences, from pure to applied thermodynamics, from information theory to transmission technology, from ecology to economics. In spite of the mathematical difficulty of its definition and application for the nonspecialists, the word entropy is often used in the everyday life, where it assumes less constrained, looser meanings than in the scientific community, being generally associated to disorder, lack of organization, indefiniteness, and physical and social degradation. In this article, we first give the mathematical definition of entropy and discuss its meaning as a basic thermodynamic function in reversible, ideal processes. The relationship of the entropy concept with real irreversible processes, statistical thermodynamics, information theory, and other sciences is then reviewed.

1. INTRODUCTION

The development of thermodynamics as well as the introduction of the concept of entropy is deeply rooted into the fertile technological ground of the Industrial Revolution, which took place in England between the beginning of the 18th and the second half of the 19th century. The development of steam-powered machines (first used to pump water out of coal mines and then to convert into work the heat generated by coal combustion) amplified both technological and purely scientific research, in an interplay of mutual reinforcement. The "miner's friend," an engine built by Thomas Savery in 1698, had a very low efficiency. Other more efficient engines (such as Newcomen's engine in 1712) opened the way to Watt's steam engine (1765), the efficiency of which was much higher than Newcomen's engine. The Scottish technician James Watt had to solve several technological and theoretical problems before his machine could effectively replace the previous steam engines. However, at the beginning of 19th century, hundreds of Watt's machines were used in Europe and constituted the basis for the significant increase of production that characterized the Industrial Revolution.

The conversion of energy from one form to another, and in particular the conversion of heat into work, has been investigated by many scientists and technicians in order to deeper understand the nature of heat toward increased efficiency. These studies yielded a general framework for energy conversion processes as well as a set of laws, the so-called laws of thermodynamics, describing the main principles underlying any energy transformation. It is worth noting that the investigation about thermal engines and the limits to the conversion of heat to mechanical work (second law, Carnot, 1824, "Réflexions sur la puissance motrice du feu et sur les machines propres à developper catte puissance") preceded the statement of the first law of thermodynamics. This is because heat conversion to mechanical work (steam engine, second half of 18th century) generated first an attention to the existing constraints and then a much clearer understanding of the nature of heat and a quantitative equivalence between heat and mechanical work (middle of the 19th century).

Clausius then provided Carnot's "Réflexions," with a stronger mathematical basis and introduced the concept of entropy, S, in so developing a clear statement of the second law. The entropy concept is perhaps the most complex one in physical science, also because it was gradually charged by many meanings belonging not only to thermodynamics, but also to a variety of other fields and disciplines. A further definition of the entropy concept came from L. Boltzmann (1844–1906) in the second half of the 19th century, by means of the so-called H-theorem,

in which the statistical meaning of entropy was definitely assessed.

2. ENTROPY AS A THERMODYNAMIC STATE FUNCTION

A consequence of the first and second laws of thermodynamics is that the efficiency, η_R, of any reversible machine only depends on the two temperatures, T_1 and T_2, between which the machine operates. In short, after adequately choosing the thermometric scale in order to have $T = t + 273.14$ (t is Celsius temperature), it is

$$Q_2/Q_1 = T_2/T_1 \tag{1}$$

and

$$\eta_R = (T_1 - T_2)/T_1 \tag{2}$$

independently on the fluid or the amount of heat exchanged. The temperature T is called absolute temperature and measured in absolute degrees (or Kelvin degrees, K). Since it does not depend on the particular thermometric substance used, it therefore coincides with the one obtained by using an ideal fluid as the thermometric substance.

Equation (1) can be rewritten as

$$Q_1/T_1 - Q_2/T_2 = 0 \tag{3a}$$

and generalized to the case of n sources in a reversible cycle, as

$$\sum Q_i/T_i = 0, \tag{3b}$$

where Q_i has a positive sign if heat is provided to the fluid and a negative sign if heat is released. Eq. (3b) can be further generalized to the case of infinitesimal amounts of heat δQ exchanged by each source in any reversible process with continuous temperature changes:

$$\int_{\text{cycle}} \delta Q/T = 0. \tag{4a}$$

For any *irreversible* process S, it would be

$$\int_{\text{cycle}} \delta Q/T < 0 \tag{4b}$$

because $\eta_s < \eta_R$

Let us now consider the most general case of a reversible cycle as the one shown in Fig. 1, with continuous temperature changes. It can always be simplified as the succession of two reversible transformations from A to B (along the pathway 1) and vice versa (along the pathway 2).

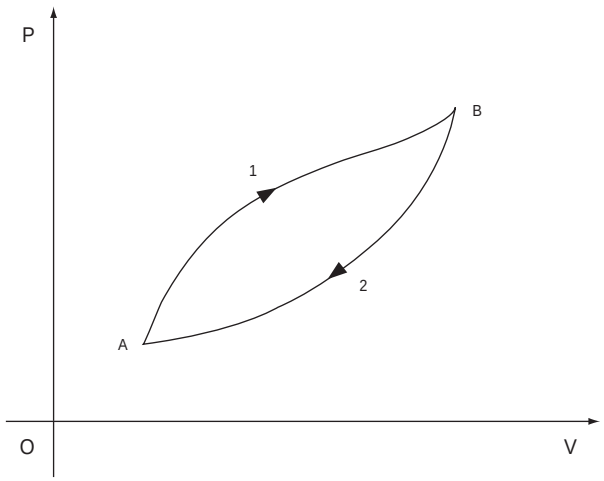


FIGURE 1 Diagram of a reversible cycle, with continuous temperature changes, simplified as the succession of two reversible transformations from A to B (along the pathway 1) and vice versa (along the pathway 2).

According to Eq. (4a), we may write

$$\int_A^B \delta Q/T|_1 + \int_B^A \delta Q/T|_2 = 0, \tag{5}$$

where the subscripts indicate the paths from A to B and vice versa. Due to the reversibility of the two transformations $A \to B$ and $B \to A$, we may also write

$$\int_A^B \delta Q/T|_1 = -\int_B^A \delta Q/T|_2 = \int_A^B \delta Q/T|_2. \tag{6}$$

This result is of paramount importance, because it demonstrates that the integral does not depend on the specific reversible transformation from A to B, but only on the actual states A and B. We are therefore able to define a new function only depending on the system state (such as the internal energy U).

Clausius (1850) called this function entropy, defined as follows:

> *Given an arbitrary initial system state O, an arbitrary final state A and an arbitrary reversible transformation from O to A, the value of the integral*
>
> $$S(A) = \int_O^A \delta Q/T + S(O) \tag{7}$$
>
> *calculated over any reversible transformation $O \to A$ is a function only of the states O and A. This value is called entropy.*

Entropy is measured in cal/K or J/K. It clearly appears that S(A) is defined but for an arbitrary constant S(O), which has no practical utility, because we are always interested in entropy differences between states and not in absolute values. We can

therefore equate this constant to zero. For any reversible transformation from a state A to a further state B, the entropy variation dS will be

$$dS = S(B) - S(A) = \int_A^B \delta Q/T \qquad (8)$$

independent on S(O), the entropy of the arbitrary initial state O. Unlike dS, an exact differential independent on the path, δQ is path dependent and it is therefore indicated by means of the δ symbol. The expression at the right side of Eq. (8) is called the Clausius integral. As a consequence, in order to calculate the variation of entropy between the initial state A and the final state B of a system undergoing a real and irreversible process we must first consider the corresponding reversible transformation of the system between the same two states and calculate dS over this pathway. If we then consider a globally irreversible cycle (Fig. 2), made up with an irreversible transformation from A to B and a reversible transformation from B to A again, it must be according to Eq. (4b) and Eq. (8):

$$\int_A^B \delta Q/T|_{irr} + \int_B^A \delta Q/T|_{rev} < 0 \qquad (9a)$$

$$\int_A^B \delta Q/T|_{irr} + S(A) - S(B) < 0 \qquad (9b)$$

$$S(B) - S(A) > \int_A^B \delta Q/T|_{irr}. \qquad (9c)$$

This famous inequality, due to Clausius and named after him, states that the variation of entropy between two states for an irreversible transformation is always higher than Clausius integral. From Eqs. (8) and (9c) we have that in a generic transformation,

$$S(B) - S(A) \geq \int_A^B \delta Q/T, \qquad (10)$$

where the equality applies only to reversible processes. If a system is isolated (i.e., it has no exchange of energy [or matter] with the surrounding environment), we have

$$S(B) - S(A) \geq 0 \qquad (11a)$$

$$S(B) \geq S(A) \qquad (11b)$$

that is, "If an isolated system undergoes a transformation from an initial state A to a final state B, the entropy of the final state is never lower than the entropy of the initial state." This statement is the so-called entropy law.

Therefore, when an isolated system reaches the state of maximum entropy consistent with the constraints imposed upon it, it cannot undergo any further transformation. The state characterized by the maximum entropy is the most stable equilibrium state for the system. From a strictly thermodynamic point of view, the increase of a system's entropy measures the amount of energy no longer usable to support any further process evolution. When all the energy of a system becomes unusable (degraded heat with no temperature gradients to drive heat flows), no more system transformations are possible. In fact, heat cannot be further transformed into other forms of energy nor can it be transferred from a body to another within the system, because all bodies have the same temperature.

It is important to underline that the law of entropy states in an unequivocal way that in an isolated system all real transformations proceed irreversibly towards the state of maximum entropy. The whole universe is assumed to be an isolated system, and therefore we may expect it to reach the maximum entropy state in which its energy is completely converted into heat at the same temperature and no further transformation can be supported. This state would be the state of maximum equilibrium for the universe and has been defined as its thermal death.

Finally, unlike in classical mechanics, no time symmetry is observed in thermodynamics. While the movement of a point (i.e., a virtual object) may always occur from A to B and vice versa (time symmetry between t and $-t$ over the time axis), a real

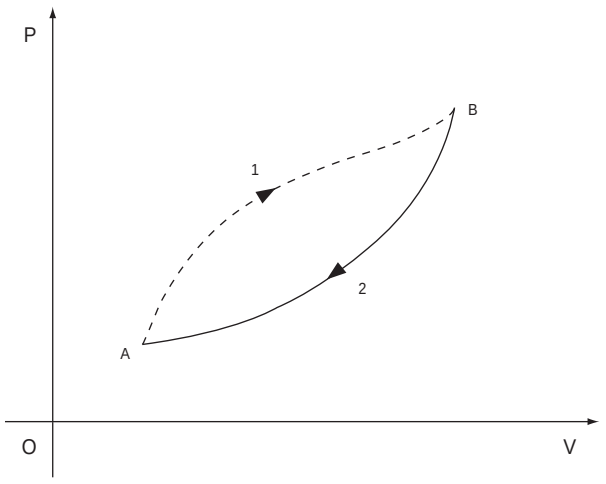


FIGURE 2 Diagram of a globally irreversible cycle, with continuous temperature changes, simplified as the succession of one irreversible transformation from A to B (along the pathway 1, dotted line) and one reversible transformation (along the pathway 2).

process in an isolated system cannot reverse its evolution if this change is not accompanied by an increase of entropy.

In 1906, Nerst demonstrated that the entropy of any thermodynamic system at the absolute temperature of 0 K is equal to zero, whatever is the state of the system (Nerst's theorem). As a consequence of Nerst's statement, also known as the third law of thermodynamics, it is convenient to choose one of these states as the initial state O and to define the entropy of state A (at a temperature $T \neq 0$ K) as

$$S(A) = \int_O^A \delta Q/T, \qquad (12)$$

where the arbitrary constant S(O) of Eq. (7) is equal to 0, in order to be in agreement with Nerst's theorem, and where the integral must be calculated over any reversible transformation from O (at $T = 0$ K) to A (at $T \neq 0$ K).

It is worth noting that, at $T = 0$ K, it is impossible to have a mixture of finely ground substances. In fact, if so, it would be possible to diminish the entropy of the system by separating them out of the mixture. This yields an *equivalent* statement of the third law: The entropy of a pure substance is equal to zero at $T = 0$ K. As a consequence, the substances that are in the mixture will spontaneously separate from each other when their temperature drops in the proximity of $T = 0$ K.

3. STATISTICAL MEANING OF ENTROPY AND BOLTZMANN'S H-THEOREM

Dice players know very well that if we cast two dice it is much easier to get a seven than a twelve. This is because a seven can be obtained as $1+6$, $2+5$, $3+4$, $4+3$, $5+2$, and finally $6+1$, where the first number is the value obtained from the first die, and the second number is the value of the second die. There are therefore six ways to obtain a seven, and only one $(6+6)$ to obtain a twelve. We may state that the macroscopic configuration seven has six microscopic states accomplishing it, and is therefore more probable than the configuration twelve.

A similar result can be achieved if we consider a box divided into two compartments (Fig. 3A). There are several ways of putting four balls in it. We may put four balls in its left compartment leaving the right compartment empty. Or we may do the opposite. The two states are identical from the microscopic and statistical points of view. In fact, if we try to extract one ball from the box, in the dark, we have a 50% probability of putting our hand in the empty part and a 50% probability of putting our hand in the compartment with the four balls. The probability is the same for both compartments, which are therefore perfectly equivalent from a statistical point of view. Therefore, the macroscopic state "four balls in a compartment and no balls in another" can be achieved by means of two microscopic configurations perfectly equivalent.

Things are different if we want to distribute the four balls in both compartments, two in each one of them. We have now six possible ways for distributing our balls, each of them equivalent from a statistical point of view (Fig. 3B). The macroscopic state having two balls per compartment has more ways to occur (more microstates), which means that it is more probable that in Fig. 3A.

It is easy to understand now that a macroscopic configuration having three balls in a compartment and only one in the other is more probable than the two previously considered configurations (Fig. 3C) since it can be achieved by means of eight different microstates, all perfectly equivalent.

Let us now consider a macroscopic state of a perfect gas, defined by given values of the quantities P (pressure), V (volume), and T (temperature). It may result from a given number N of microstates—that is, a given number N of different positions and velocities of the billion molecules which compose it. In each microstate there will be different values of the position and the velocity of each molecule, but the global result of the behavior of all the molecules will be exactly that macrostate characterized by the measured values of P, V, and T. Of course, the larger the number of possible equivalent microstates N, the higher the probability of the corresponding macrostate. We know from the previous section that an isolated perfect gas tends to and will sooner or later achieve the state characterized by the maximum entropy consistent with the energy U of the gas. This state is also the macrostate characterized by the largest number of equivalent microstates. Ludwig Boltzmann (1844–1906) understood the strong link between the thermodynamic and the statistical interpretation of the entropy concept. In 1872 he published a milestone paper in which he introduced the basic equations and concepts of statistical thermodynamics.

As the first step, Boltzmann assumed that all the n molecules constituting the gas had the same velocity and divided the volume V into a very large number of

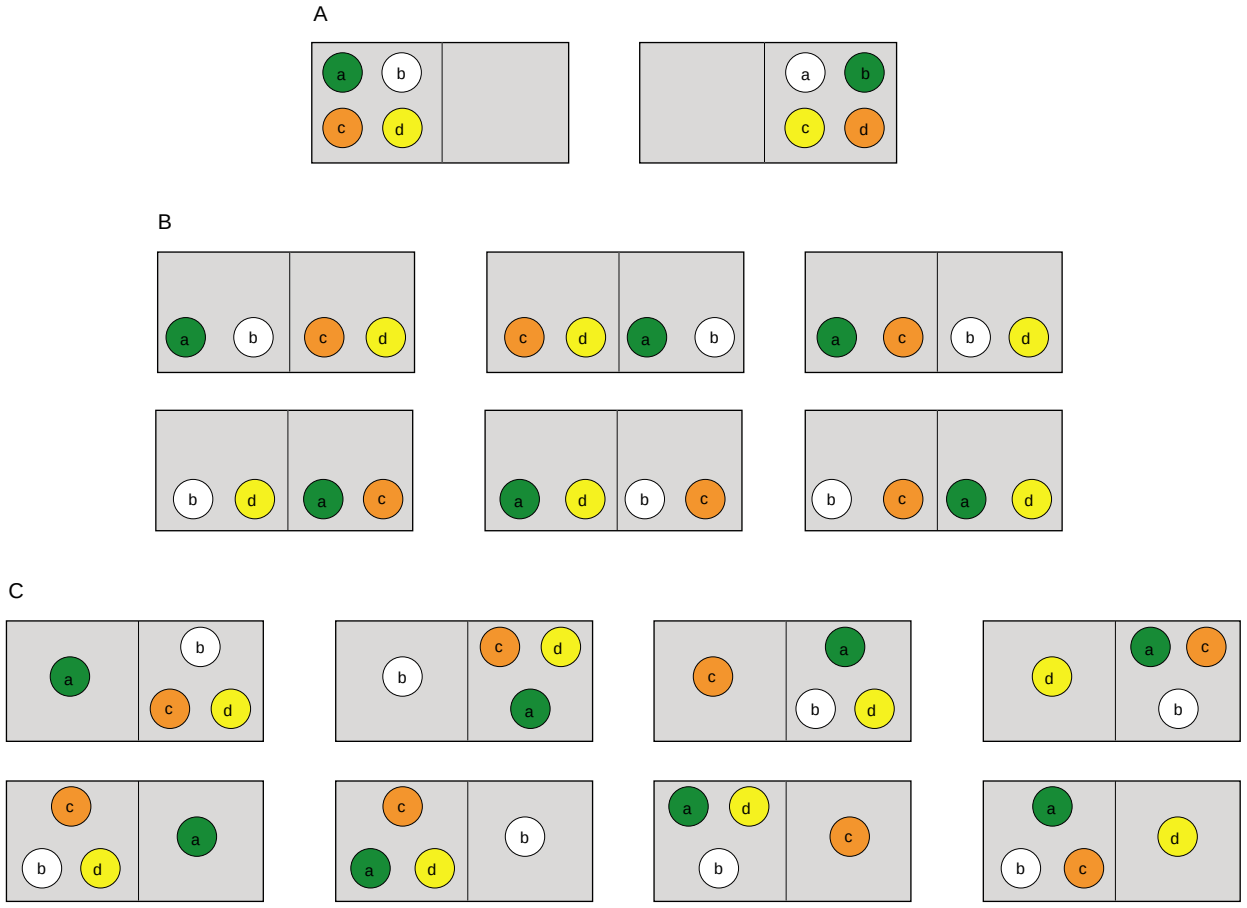


FIGURE 3 Macrostates and microstates in a system made up of four balls in two separated compartments. Each macrostate may correspond to different numbers of equivalent microstates. Figures A, B, and C show the different number of equivalent microstates that may correspond to a given macrostate, characterized by a measurable temperature, pressure, and volume.

cells, M, each having the smallest experimentally measurable size. It is then easy to demonstrate that the number of distinct microstates N corresponding to the macrostate of volume V is

$$\mathrm{N} = \frac{n!}{n1.....n_M!,} \tag{13}$$

where n_i is the number of molecules contained in the i-th cell and $n = \sum_i n_i$ Boltzmann also assumed that the entropy of the state is linked to the number N of microstates by the logarithmic expression

$$\mathrm{S} = k \log N, \tag{14}$$

where $k = 1.38 \times 10^{-23}$ joule/K, the so-called Boltzmann's constant. The logarithm base used is arbitrary, since the conversion from a base to another one would simply require the introduction of a multiplicative constant. For a reason, which will clearly appear in the next section, it is preferable to use the logarithms with base two.

According to Nerst's theorem and to Eq. (14), it must be $N = 1$ when $\mathrm{T} = 0\,\mathrm{K}$, in order to have $\mathrm{S} = 0$—that is, at the absolute temperature $\mathrm{T} = 0\,\mathrm{K}$, a system's state is always composed by only one microstate.

By applying to Eqs. (13) and (14) the properties of natural logarithms and the Stirling formula for the mathematical expansion of $n!$, Boltzmann derived the expressions

$$S = k \log N = k \log\left(\frac{n!}{n_1!.....n_M!}\right) = k\, H \tag{15a}$$

$$H = -\sum_i p_i \log p_i \quad \text{with} \sum_i p_i = 1, \tag{15b}$$

where p_i is the probability to find one molecule in the i-th cell. It is not difficult to show that the maximum entropy state is that state for which all the p_i's have the same value (i.e., that state in which all the cells are occupied by the same number of molecules). It can also be said that this is the most disordered

state, since the distribution of the molecules among the cells is the most uniform and therefore the one showing the lower level of organization.

As the second step, Boltzmann dealt with the most general case in which the positions and velocities of the molecules are all different. He introduced a distribution function, $f(x,y,z; v_x,v_y,v_z)$, to account for the different positions and velocities and expressed the p_i's as functions of $f(x,y,z; v_x,v_y,v_z)$. He also transformed the summation $\sum$ of Eq. (15b) into an integral, due to the very high number of molecules:

$$H = -\int f \log f \, dx \, dy \, dz \, dv_x \, dv_y \, dv_z. \quad (16)$$

Boltzmann finally demonstrated in 1872 his well-known H-theorem—that is, that it must be

$$dH/dt \geq 0. \quad (17)$$

According to the Eq. (17), H (and therefore $S = k\,H$ as well) never decreases over time. Actually Boltzmann defined H without the minus sign in Eq. (16) and then demonstrated that $dH/dt \leq 0$. As a consequence, Eq. (15a) became $S = -k\,H$. We prefer, in agreement with other authors, the introduction of the minus sign in order to have H increasing over time according to Eq. (17). In so doing, H and S show exactly the same trend.

Boltzmann therefore restated by means of statistical considerations the law of entropy in isolated systems—that is, he gave a statistical reading of the second law of thermodynamics and the entropy concept. He also demonstrated that his velocity distribution function f coincided with the one hypothesized by Maxwell in 1859, from which the most probable state corresponding to the maximum entropy could be derived.

Boltzmann's seminal work generated a heated debate about mechanical reversibility and thermodynamic irreversibility. Several critics questioned that thermodynamic irreversibility could be described and explained in terms of mechanical reversibility. It is not the goal of this article to go into the details of this fascinating debate. The latter soon became a philosophical investigation about the mechanistic conception of reality as well as the so-called arrow of time (i.e., the spontaneous direction of natural processes as suggested by their irreversibility).

The most difficult part of this concept is trying to figure out two enormous numbers—that is, the number of molecules constituting the system (the ideal gas or any other) and the number of microstates contributing to a given macrostate of the system. We would like to underline that Boltzmann only stated the statistical irreversibility of the evolution of an isolated system. After a sufficiently (infinitely) long time, the system will certainly achieve the most stable equilibrium state, characterized by the maximum entropy, although among the billion microstates contributing to a given macrostate in any instant, there is a percentually small (but different from zero) number of microstates capable of driving the system to a macrostate at lower entropy. This means that future lower entropy states are not impossible for the system, but only extremely unlikely.

4. ENTROPY AND INFORMATION THEORY

Information theory is relatively recent. It moved its first steps in 1948, when C. E. Shannon published his seminal paper, "A Mathematical Theory of Communication." This paper was the starting point of modern communication technologies, which have so deeply modified all aspects of our life. Shannon's problem was how to measure and transmit a message from a source (transmitter) to a recipient (receiver) through a transmission channel generally characterized by a background noise capable of altering the message itself. We will focus in this paper on the problem of information measure.

First of all, it is important to underline the fact that the concept of message is very broad. However, the set of messages that each source is able to transmit must be specified without any possible ambiguity. A letter of the alphabet, the ring of a bell, a word or a particular gesture are all examples of possible messages. Shannon considered the n elementary messages that a source could transmit (i.e., the basic alphabet of the source itself). He first assumed that these messages were equally probable (i.e., that the probability for each of the n options to be transmitted was $p = 1/n$). If two consecutive elementary messages are transmitted, the possible options (i.e., the number of possible different sequences of two messages out of n available ones) are n^2. In general, if m messages are transmitted in a row, the possible different sequences become $N = n^m$ (i.e., the number of all the possible permutations of the n elementary messages in groups of m elements, also allowing for repetitions). At the same time, the information transmitted (i.e., the events or descriptions that characterize the source and that are not yet known by the recipient) should be twice, three times,…, m times the one transmitted by means of only one elementary message. In any case, this

information should be somehow proportional to the number m of messages that make up the transmitted sequence of messages. Shannon assumed it was proportional to the logarithm of the possible options N, which easily yielded

$$\begin{aligned} \mathrm{I} &= \log N = \log n^m = m \log n \\ &= -m \log 1/n = -m \log p. \end{aligned} \tag{18a}$$

The information conveyed by each elementary message is therefore

$$i = -\log p = \log 1/p. \tag{18b}$$

From Eq. (18b) it appears that the information conveyed by a message is proportional to the log of the reciprocal of its probability. If $p<1$, it is $i>0$. This also means that the less probable the message, the higher the information content transferred from the source to the recipient. If we already know that a given message will be transmitted ($p=1$), it will not carry any additional information. In fact, we would not switch the TV on if we already knew which news are going to be announced. Instead, an unlikely and therefore unexpected message carries a significant piece of information. Finally, no information is carried by a message that cannot be transmitted ($p=0$).

Shannon then generalized his equation to the case of messages showing different transmission probability. He assumed that the source could transmit n elementary messages in groups of m elements, with probabilities p_i, where $i=1,\ldots, n$. If m is large enough, the i-th message with probability p_i will appear $m\ p_i$ times and the total information carried by the sequence of m messages will be

$$\mathrm{I} = -\sum_i m\ p_i \log p_i. \tag{19a}$$

As a consequence, the average information conveyed by each of the m messages will be

$$\mathrm{H} = -\sum_i p_i \log p_i \quad \text{with} \quad \sum_i p_i = 1 \tag{19b}$$

Equation (19b) has the same mathematical form as Eq. (15b), thus indicating a surprising similarity between Boltzmann's entropy and Shannon's information. For this reason, the same symbol H is used and Shannon's expression is very often called entropy of a source. Actually, the entropy of the source should be indicated as in Eq. (15a), where k is Boltzmann's constant and H would be Shannon's expression for the average information of a message. The entropy H of a source, intended as the average information content of a given message, also indicates the recipient's ignorance about the source, or in other words his a priori uncertainty about which message will be transmitted. This fact should not be surprising, since it is exactly by receiving the message that the uncertainty is removed.

In 1971, M. Tribus and E. C. McIrvine provided an interesting anecdote about Shannon, who is reported to have said:

> *My greatest concern was what to call it. I thought of calling it "information," but the word was overly used, so I decided to call it "uncertainty." When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, "You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage."*

True or not, the story points out the links between the two concepts. Equation (19b) also allows the definition of a unit to measure information. Let us assume that the source releases information by means of a binary alphabet—that is, it is able to transmit only two elementary messages, both having a probability $p=\frac{1}{2}$. It is common habit to indicate these messages by means of the two digits, 0 and 1, which therefore become the basis of the so-called digital alphabet and transmissions. The average information carried by an elementary message is, in such a case, from Eq. (19b):

$$\begin{aligned} \mathrm{H}_1 &= -1/2 \log 1/2 - 1/2 \log 1/2 \\ &= -1/2 \log_2 1/2 - 1/2 \log_2 1/2 \\ &= -1/2 \times (0-1) - 1/2 \times (0-1) \\ &= 1/2 + 1/2 = 1. \end{aligned} \tag{20}$$

This unit quantity is called a bit (acronym for binary digit), which indicates a pure number, with no physical dimensions. One bit of information is, by definition, the information carried by each of the two elementary messages released by a binary source with equal probability 1/2. The reason of the choice of logarithms in base two is now evident.

For the sake of completeness, we may also add that another very common unit is the byte (1 byte = 8 bit). Eight bits allow for the construction of $2^8 = 256$ different sets of elementary messages, or different 8-bit strings, to each of which we may for instance assign a given symbol on the keyboard of a personal computer.

4.1 Maxwell's Demon

The link between S and H was deeply investigated soon after the introduction of Shannon's theory. Each isolated system can be considered either a source of information or vice versa. For example, an isolated gas with totally thermal energy and uniform temperature is in a high-entropy, disordered state (from a thermodynamic point of view). It is also in a macrostate to which a high number of microstates may correspond and therefore it is a highly probable state (from a statistical point of view). If so, the uncertainty about the actually occurring microstate is huge and a large number of bits (much information) would be needed to ascertain it (from the point of view of Information theory).

A clear example of the link between entropy and information is the so-called Maxwell's demon, a problem that fascinated scientists for a long time. Assume that we have two rooms at the same temperature, due to the fact that in each room there is the same distribution of molecular velocities and kinetic energies. If we could separate the slowly moving molecules from the fast ones, the temperatures of the two rooms should vary. Maxwell assumed that this task could be performed by a demon quickly opening or closing a frictionless window in the separation wall, in order to pass fast (hot) molecules to a room and slow (cold) molecules to the other one (Fig. 4). The final effect would be the separation of hot from cold molecules and the generation of a temperature gradient (one room would be at higher temperature than the other one), without spending energy from outside, due to the absence of friction in the process. In so doing, the entropy of the system would decrease and the second law would be violated. This paradox was only solved in 1956 by L. Brillouin. He pointed out that the demon needed information about each approaching molecule, in order to decide about opening or not the window. He could, for instance, light them up. The process of generating information would cause an increase of the overall entropy of the system + environment, larger or equal to the entropy decrease of the system only.

From Eq. (15a) and the result of Eq. (20), we have

$$S = k\,H = 1.38 \times 10^{-23}\ \text{joule/K} \times 1\ \text{bit} = 1.38 \times 10^{-23}\ \text{joule/K} \times 1\ \text{bit}. \quad (21)$$

Therefore, 1 bit of information requires an entropy generation equal to 1.38×10^{-23} joule/K. This entropy cost of information is so low, due to the small value of k, that people thought information was free, as for Maxwell's demon. Of course, if H is calculated by means of natural logarithms, then 1 bit corresponds to $k \ln 2$ entropy units.

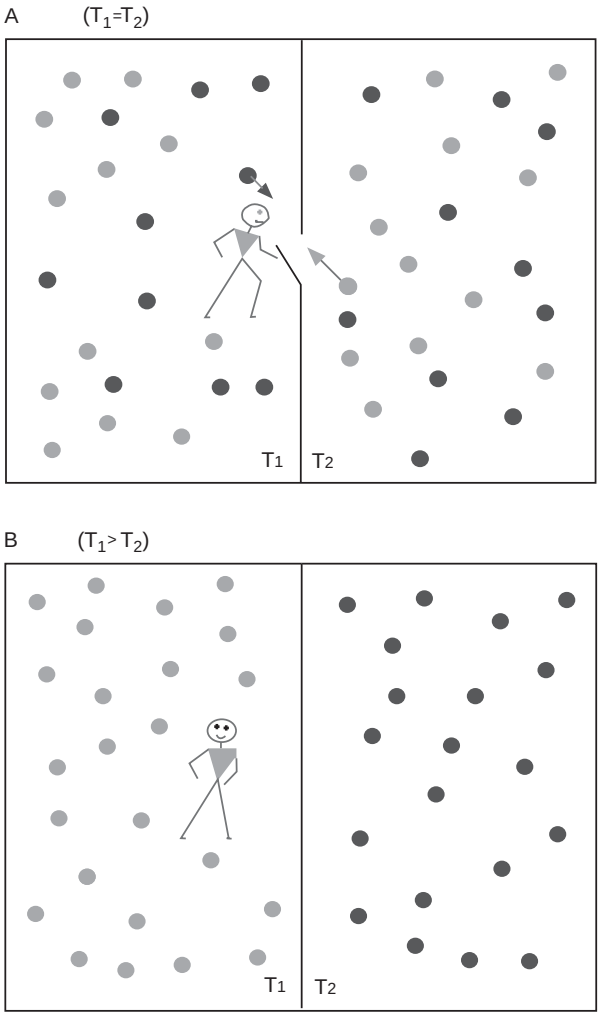


FIGURE 4 Maxwell demon. The demon seems to be able to violate the second law of thermodynamics, by selecting the molecules and sorting the hot (gray) from the cold ones (black). The two compartments are initially at the same temperature (A). After some time, all hot molecules migrated to the left compartment and cold ones to the right compartment. As a consequence of the selection, the temperature of the left side is higher than the temperature of the right side (B).

5. THE FREE-ENERGY CONCEPT

What happens when entropy increases? What does it mean in practical terms? Lots of pages have been written, linking entropy increase with disorder, disorder with less organization, less organization with lower quality, lower quality with lower usefulness.

Sometimes links have been a little arbitrary, because things are different with isolated or open systems. How can we quantify these concepts?

We know from Section 2 that if an isolated system undergoes a transformation from a state A to a state B, its entropy never decreases. The entropy law states in an unequivocal way that in an isolated system all real transformations tend irreversibly towards a state of maximum entropy. This statement applies to any kind of energy conversion, not only to heat-to-work ones. An entropy increase means that a fraction of the energy provided to the system is transformed into degraded heat, no longer usable to support any further process evolution. The energy actually usable is less and less work can be done. Many feel confused by using increasing entropy to mean decreasing ability to do work, because we often look to increasing quantities with positive feelings in our mind. For the entropy concept it is the opposite. The chemist J. W. Gibbs (1839–1903) actually "solved" this problem by introducing a new quantity, which he called available energy, linked to the entropy concept by means of a fundamental equation named after him. Available or free energy was then recognized as being an important property not only in chemistry but also in every energy transformation, in which the dissipation of a gradient (temperature, pressure, altitude, concentration, etc.) yields a product or delivers work as well as releasins some degraded energy or matter.

Several slightly different expressions exist for available energy, so that we will only refer to one of them, the so-called Gibbs energy, or Gibbs free energy, defined for reversible system transformations at constant pressure. This is the most frequent case, because the majority of processes on Earth occur at constant pressure.

5.1 Gibbs Energy

In general, the available energy is that fraction of a system's energy that is actually usable to drive a transformation process. It can be calculated by subtracting from the total system's energy the fraction that becomes unavailable due to the entropy increase. Assuming a process in which heat is transferred at constant pressure p, let us first define the Gibbs-energy function, G, as

$$G = H - TS, \tag{22a}$$

where

$$H = U + pV \tag{22b}$$

is the system's enthalpy, a measure of the total heat content of a system, while

$$U = Q + W \tag{22C}$$

is the well-known expression of the first law of thermodynamics. For a measurable change of the system's state at constant temperature T, from Eq. (22a) we have

$$\begin{aligned} G_f - G_i &= \Delta G = \Delta H - T\Delta S \\ &= (H_f - H_i) - T(S_f - S_i), \end{aligned} \tag{23}$$

which quantifies the amount of the available energy ΔG, after enthalpy and entropy changes have been accounted for. The subscripts i and f indicate the initial and final states.

This result shows the importance of the entropy concept also in practical applications. It indicates that the energy of a system is not always totally available to be transformed into work. Part of this energy is not converted but instead is released as high entropy waste (degraded heat), no longer able to support change of state of the system. The terms available and free can be now understood, as opposed to the bound, unavailable fraction of energy present in the entropy term of Eq. (23).

A process with $\Delta G < 0$ (i.e., a process whose initial state has higher available energy than the final state) may evolve spontaneously towards the latter, and deliver work. This result has a general validity and clearly shows the importance of the entropy-related, free energy concept.

Heat conversion into work and chemical reactions are clear examples of how the second law and the entropy concept work. In both cases we have spontaneous processes: (1) the heat flows from a source at higher temperature to a sink at lower temperature, and (2) the reaction proceeds from the higher to the lower free energy state. In both cases we have a process driven from a gradient of some physical quantity (respectively, a difference between two temperatures and a difference between two free energy states) and some work delivered to the outside at the expenses of the free energy of the system. In general terms, whenever a gradient of some physical quantity (altitude, pressure, temperature, chemical potential, etc.) exists, it can drive a transformation process. The free energy of the system is used whenever these spontaneous processes occur. The result is a lowered gradient, and an increased total entropy of the process + environment system, in addition to the work done.

6. CONCLUSION

The entropy concept is fascinating and rich of significance. At the same time, it appeared scaring and confusing to many. Many focused on the gloomy picture of increasing disorder and thermal death, characteristic of equilibrium thermodynamics and isolated systems. In this regard, entropy is used to point out and measure the decreasing availability of high quality resources (i.e., resources with low entropy and high free energy), the increase of pollution due to the release of waste, chemicals, and heat into the environment, the increase of social disorder due to degraded conditions of life in megacities all around the world, the "collapse" of the economy, and so on.

Others identify entropy as the style of nature: since the biosphere is an open system supported by a constant inflow of solar energy, its structures and life phenomena undergo a continuous process of self-organization. Geologic processes, atmospheric systems, ecosystems, and societies are interconnected through a series of infinitely different and changing relationships, each receiving energy and materials from the other, returning same, and acting through feedback mechanisms to self-organize the whole in a grand interplay of space, time, energy and information. During this self-organization process, entropy is produced and then released towards the outer space. Living structures do not violate the second law, but feed on input sources, which continuously supply low entropy material and energy (free energy) for system development.

While the first point of view calls for increased attention to avoid misuse of resources and prevent degradation of both natural and human environments (i.e., prevent the loss of stored information), the second point of view focuses on the creation of organization and new structures out of disordered materials, thanks to a flow of resources from outside. This point of view calls for adaptation to the style of nature, by recognizing the existence of oscillations (growth and descent) and resource constraints, within which many options and new patterns are however possible. Both points of view are interesting and stimulating and account for the richness of the entropy concept. The latter, first introduced in the field of energy conversion, acquired very soon citizenship in several other fields. We already mentioned the information theory in Section 4, but many other examples can be easily provided.

Environmentally concerned economists focused on the entropy tax—that is, on the fact that the development of economic systems is always dependent on the degradation of natural resources, the amount of which is limited. Some pointed out that the notion of entropy helps "to establish relationships between the economic system and the environmental system," as noted by Faber *et al*. in 1983. In 1971, Nicholas Georgescu-Roegen underlined that "The Entropy Law…emerges as the most economic in nature of all natural laws."

The ecologist Ramon Margalef introduced in 1968 a quantity, called ecological diversity, D, as a measure of the organization level of an ecosystem. This quantity is defined exactly in the same way as entropy, namely,

$$\mathrm{D} = -\sum_i p_i \log p_i \qquad (24)$$

where p_i is the probability for an organism to belong to a given species. The higher the ecological diversity, the lower the production of entropy per unit of biomass, because resources are better utilized and support the growth of the whole spectrum of ecosystem hierarchy. Instead, simplified ecosystems with fewer species (such as agricultural monocultures) do not use available resources and coproducts as effectively, resulting into a higher production of entropy per unit of time and biomass.

Finally, entropic styles have been adopted by several artists to express in their works their desire of novelty and their rejection of previous, static views on life. Some understood and meant entropy alternatively as disorder, indefiniteness, motion, rebellion—the opposite of the ordered and well-defined, Newtonian structures of Renaissance artists. In 1974, Rudolf Arnheim published an interesting assay on disorder and order in arts as well as on the need for complexity, which moves artists toward investigating new forms to express their creativity.

In the end, we would do a disservice to the entropy concept, by limiting its significance to the concept of disorder, which is, however, one of its meanings, or even to the practical application of the concept of free energy. Entropy, a word now very common in the everyday language, may be suggested as the other side on the coin of life itself, with its development, its richness and complexity. Life processes (such as photosynthesis) build organization, add structure, reassemble materials, upgrade energy, and create new information by also degrading input resources. Degraded resources

become again available for new life cycles, ultimately driven by the entropic degradation of solar energy.

SEE ALSO THE FOLLOWING ARTICLES

Conservation of Energy Concept, History of • Energy in the History and Philosophy of Science • Entropy and the Economic Process • Exergy • Exergy Analysis of Energy Systems • Exergy Analysis of Waste Emissions • Exergy: Reference States and Balance Conditions • Thermodynamics and Economics, Overview • Thermodynamic Sciences, History of • Thermodynamics, Laws of

Further Reading

Arnheim R. (1974). "Entropy and Art. An Assay on Disorder and Order." University of California Press, Berkeley, California.

Boltzmann, L. (1872). Further studies on the Thermal Equilibrium of Gas Molecules (from *Sitzungsberichte der kaiserlichen Akademie der Wissenschaften*), Vienna.

Brillouin, L. (1962). "Science and Information Theory." Academic Press, New York.

Clausius, R. G. (1850). Ueber die bewegende Kraft der Wärme. *Annalen der Physik und Chemie* **79**, 368–397, 500–524 [translated and excerpted in William Francis Magie, *A Source Book in Physics*, McGraw-Hill, New York, 1935].

Faber, M., Niemes, H., and Stephan, G. (1983). "Entropy, Environment, and Resources. An Assay in Physico-Economics." Springer-Verlag, Berlin.

Georgescu-Roegen, N. (1971). "The Entropy Law and the Economic Process." Harvard University Press, Cambridge, MA.

Gibbs, J. W. (1931). "The Collected Works of J. Willard Gibbs." Longmans, Green and Co, London.

Margalef, R. (1968). "Perspectives in Ecological Theory." The University of Chicago Press, Chicago.

Maxwell, J. C. (1871). "Theory of Heat." Longmans, Green, & Co., London.

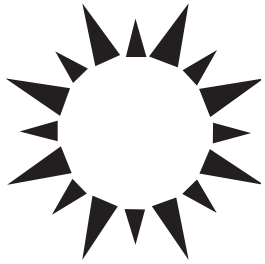
Nernst, W. (1906). Ueber die Berechnung chemischer Gleichgewichte aus thermischen Messungen. *Nachr. Kgl. Ges. Wiss. Gött.* **1**, 1–40.

Prigogine, I. (1978). Time, structure, and fluctuations. *Science* **201**, 777–785. Text of the Lecture delivered in Stockholm on 8 December 1977, when Prigogine was awarded the Nobel Prize in chemistry.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27**, 379–423.

Tribus, M., and McIrvine, E. C. (1971). Energy and information. *Scientific American* **225**(3), 179–184.

Entropy and the Economic Process



GABRIEL A. LOZADA
University of Utah
Salt Lake City, Utah, United States

1. Entropy
2. The Laws of Thermodynamics
3. Economic Implications of Entropy to Production Theory
4. Entropy and Neoclassical Economic Analysis
5. Interpretations and Misinterpretations of Entropy
6. Conclusion

Glossary

entropy change The amount of heat absorbed by a material divided by the material's absolute temperature.
isolated system A collection of materials that exchanges neither matter nor energy with its surroundings.
macroeconomics The study of the economy as an aggregated whole, generally ignoring the behavior of individual industries or consumers and instead concentrating on economy-wide magnitudes such as unemployment, inflation, interest rates, the trade deficit, and growth of gross national product.
microeconomics The study of the behavior of one or a few firms, industries, or groups of consumers in an economy.
natural resources Exhaustible resources such as coal, renewable resources such as fish, and ecosystem capacities useful to humans such as the ability to cleanse dirty water.
neoclassical economics The school of economic thought which postulates that consumers act in a way that maximizes their well-being subject to their income and possibly other constraints, that firms act in a way that maximizes their profits subject to the technological constraints of production, and in general that economic agents act in a way that maximizes an objective function that can in principle be rigorously and completely mathematically described, subject to constraints that can also in principle be so described.
second law of thermodynamics Also known as the entropy law, it states that in an isolated system, entropy cannot decrease; it can either rise or remain constant.
neoclassical theory of value The theory that market prices are determined by the intersection of demand curves and supply curves, which in turn arise from the behavior of self-interested atomistic economic agents; these "curves" may be more complicated than simple functions over Euclidian space.

To explain why a capitalist economy produces certain commodities and how the prices of those commodities are determined, a description must be given of consumers' desire and ability to buy the commodity and of how firms produce that commodity. The description of how firms produce a commodity is in many ways a description of the engineering opportunities open to the firm, although describing the allocation and supervision of labor is also important, as is the firm's access to diverse forms of financing. In describing the engineering opportunities open to the firm, economists usually have to work with a simplified version of all the potentially available technologies in order for the problem to be tractable. However, it is important for the simplifications to obey fundamental physical laws, or at least not flagrantly violate them, so economists studying production should have a basic knowledge of those laws. The writers who first emphasized the relevance of the laws of thermodynamics to production economics were Frederick Soddy (the 1921 Nobel laureate in chemistry), Nicholas Georgescu-Roegen, and Herman E. Daly. The second law of thermodynamics, called the entropy law, has been the focus of most of the work in this field.

1. ENTROPY

The first law of thermodynamics states that in an isolated system, energy is neither created nor destroyed. This law cannot explain why an ice cube placed on a sidewalk on a hot summer day always melts, because the first law would permit heat to flow from the ice cube as long as an equal amount of heat

 © 2004 Elsevier Inc. All rights reserved.

flows to the air. Entropy is the concept used to explain why ice cubes on a hot sidewalk melt and in general why heat always flows from hotter objects to colder ones, as well as why spontaneous physical processes are spontaneous.

The definition of entropy is straightforward. Let T stand for absolute temperature, that is, temperature in degrees Kelvin, denoted K. (Kelvin temperature is 273.15 degrees higher than Celsius temperature.) Let Q stand for the flow of heat into a material (or system, which is a collection of materials). The metric unit of heat Q, as well as that work, is the joule, denoted 'J'; the English unit is the calorie or the BTU. Then the change in the entropy S of the material (or system) is defined to be

$$dS = dQ_{\text{rev}}/T, \tag{1}$$

where the subscript rev stands for a reversible process; physicists call a process reversible if it involves no dissipative effects such as friction, viscosity, inelasticity, electrical resistance, or magnetic hysteresis. If a system undergoes an irreversible process between an initial equilibrium state i and a final equilibrium state f, the entropy change of the system is equal to the integral from i to f of dQ/T, taken over any reversible path from i to f. No integral is taken over the original irreversible path.

Suppose one has an isolated system containing two bodies, one hot and one cold, placed in thermal contact. When a given quantity of heat, Q_0, flows from the hotter body to the colder one, the change in entropy of the system is

$$\Delta S = \frac{-Q_0}{T_{\text{hot}}} + \frac{+Q_0}{T_{\text{cold}}}.$$

Since $T_{\text{hot}} > T_{\text{cold}}$, $\Delta S > 0$. If heat were to flow in the opposite direction, away from the colder body and toward the warmer body, then the Q_0 terms in the previous equation would change sign and ΔS would be negative. Because it is the former and not the latter which is always observed, 19th-century physicists postulated that in an isolated system, entropy never decreases. This is one form of the second law of thermodynamics.

2. THE LAWS OF THERMODYNAMICS

The laws of thermodynamics are as follows:

- *Zeroth law*. Two systems that are in thermal equilibrium with a third are in thermal equilibrium with each other.
- *First law*. During a process in which no heat is exchanged with the environment, the work done is only a function of the initial and final states of the system, not of the path. Furthermore, during *any* process, heat flow, Q, is equal to $U_f - U_i + W$ where U_f and U_i are the final and initial internal energies of the system and W is the net work done by the system.
- *Second law, Kelvin-Planck statement*. No process is possible whose *sole* result is the absorption of heat from a reservoir and the conversion of this heat into work.
- *Second law, Clausius statement*. No process is possible whose *sole* result is the transfer of heat from a cooler to a hotter body.
- *Second law, entropy statement*. In an isolated system, entropy is nondecreasing.
- *Third law, unattainability statement*. It is impossible to reach absolute zero by any finite number of processes.
- *Third law, Nerst-Simon statement*. In the limit as temperature goes to zero degrees Kelvin, the entropy change of any reaction is zero.

Proofs that the three statements of the second law are equivalent can be found in standard texts, as can proof that the two statements of the third law are equivalent.

3. ECONOMIC IMPLICATIONS OF ENTROPY TO PRODUCTION THEORY

3.1 Heat Engines

The first application of entropy was to "heat engines"; these are devices, such as internal combustion engines and power-plant steam engines (in both conventional and nuclear power plants), that use temperature differences to do their work. (A car's motor, for example, will not work in a room hotter than the spontaneous ignition temperature of gasoline.) The efficiency of a heat engine is defined to be its work output divided by its heat input. A surprising and important implication of the second law of thermodynamics is that the maximum efficiency of any heat engine operating between the temperatures T_{low} and T_{high} is $1-(T_{\text{low}}/T_{\text{high}})$. This is less than 1 because T_{low} is strictly greater than zero due to the unattainability statement of the third law. A felicitous explanation of the maximum-efficiency

equation appears in Sir Oliver Lodge's *Energy*. Lodge also explains what it has to do with entropy:

> *If the quantity of heat energy Q is lost by a body at temperature T, and gained by a body at temperature T′, then the entropy lost is Q/T, and the [entropy] gained is Q/T′. The gain is greater than the loss, if T′ is less than T. But if some of the heat is utilised,—converted into mechanical or other forms of energy by means of some kind of engine,—so that only Q′ is imparted to the colder body (the work done corresponding to the difference Q−Q′) then it may be that Q/T = Q′/T′; and in that case the entropy of the system remains constant. This is the condition for a perfect engine conducting a reversible operation. Irreversible operations increase entropy or dissipate energy, for the lost availability cannot be recovered. The efficiency of an engine means the ratio between the portion of heat utilised and the heat taken from the source. It is (Q−Q′)/Q; and we see now that in a perfect engine this would be (T−T′)/T, because the Q's and the T's are then proportional. (pp. 39–40)*

Lodge's "perfect" heat engine is called a Carnot engine; on a temperature-versus-entropy diagram, the cycle of this engine traces out a rectangle, called a Carnot cycle.

This second law limit on the maximum efficiency of heat engines is striking because it establishes a boundary beyond which technological progress will never cross. This has the salutary effect of throwing scientific doubt on the predictions of technological Panglossians who believe that technical progress is unbounded and can eventually remedy any problem facing industrial civilization. On the other hand, not all engines are heat engines (fuel cells, for example, are not), and the second law maximum-efficiency limitation only applies to heat engines.

All engines—and indeed all uses of energy—are, however, subject to another limitation due to the entropy law. Lord Kelvin's principle of the degradation of energy states that whenever an irreversible process takes place, the effect on the universe is the same as that which would be produced if a certain quantity of energy were converted from a form in which it was completely available for work into a form in which it is completely unavailable for work. This amount of energy is the temperature of the coldest heat reservoir at hand times the entropy change of the universe brought about by the irreversible process. As a consequence of this principle, it is never desirable to recycle energy (because accomplishing such recycling requires using up more energy than would be regenerated).

3.2 Chemical Processes

To discuss implications of entropy to chemical rather than mechanical processes—for example, if one wished to analyze fuel cells—it is useful to rewrite Eq. (1) as

$$\Delta S = \int_i^f \frac{C_p(T)}{T}\, dT, \tag{2}$$

where i denotes some convenient initial reference state, f denotes the state of interest, and C_p denotes the heat capacity at constant pressure of the chemical substance. (The heat capacity of a substance, also known as its specific heat, is the amount of heat or work needed to raise the temperature of 1g of the substance by 1 degree Kelvin [or Celsius]. A substance's heat capacity changes slightly depending on its initial temperature.) In a laboratory, Eq. (2) can be approximated to any degree of accuracy by replacing the integral with a sum, as in

$$\Delta S \approx \sum_{j=i}^{f} \frac{C_p(T_j)}{T_j}\,(1\mathrm{K}), \tag{3}$$

where the temperature range from T_i to T_f has been broken down into j parts. Using Eq. (3), the entropy of one gram of a substance is experimentally determined in the following way: Start with the substance at a convenient initial temperature and pressure. Measure the heat needed to raise its temperature from T_1 to T_2, then divide this figure by $T_2 - T_1$ to obtain the specific heat (at constant pressure). Divide this specific heat by the initial temperature; this is $C_p(T_1)/T_1$. Next measure the heat needed to raise the temperature from T_2 to T_3, and divide that by $T_3 - T_2$, then divide that by the temperature; this is $C_p(T_2)/T_2$. Continue until T_j is reached. The change in entropy of the substance between the initial and final states is approximately equal to the sum of the $C_p(T_j)/T_j$ terms, times 1 degree Kelvin to adjust the units.

In this way, the entropy change of different materials can be experimentally determined. The results of such experiments are given—for example, in freshmen chemistry textbooks—in tables of entropies for many substances. For instance, the standard entropy of one mole of copper (63 grams) is 33 J/K = 8 cal/K, where "cal" denotes calories, and where standard (or absolute) entropies are defined with an initial condition of 0 K and one atmosphere pressure and a final condition of 300 K and one atmosphere pressure.

Armed with such knowledge, and the second law's principle that entropy in an isolated system never decreases, the equilibrium proportions of different chemicals reacting with each other can be determined merely by calculating, on paper, which proportions maximize the entropy of the mixture. In other words,

once entropy values have been tabulated, the equilibrium direction and extent of any chemical reaction can be determined without further laboratory work. There is a caveat: while the equilibrium direction and extent of any chemical reaction can be determined, not all chemical systems to go equilibrium; some get stuck in a metastable state in which an activation energy barrier blocks their path to the equilibrium. This barrier can be broken by a catalyst, which works by supplying the activation energy, which is then returned in full to the catalyst.

3.3 General Processes

The entropy law thus governs the efficiency of heat engines, the recycling of energy, and the direction and extent of chemical reactions. More generally, it sets the ground rules within which mechanical and chemical systems operate. Even more generally, it constitutes a fundamental principle of irreversibility in the following sense: Imagine a system that is isolated. By the second law, entropy always increases in such a system (it could remain the same, but only if there were no friction, only perfect machines, and so forth). Choose two dates t_1 and t_2 with $t_1 < t_2$. Then the entropy of the system at t_1 (call it S_1) is less than the entropy of the system at t_2 (call it S_2): $S_1 < S_2$. Then it is in principle impossible to reverse all of the changes that have happened in this isolated system between t_1 and t_2—assuming the system remains isolated—because such a reversal would imply a decrease of entropy. This means that perpetual motion machines are impossible. It means that complete recycling of all a process's components is impossible; some fresh inputs will always be needed for the process to occur again. *A fortiori*, the economy will always require natural resource input of some sort from nature. The economy will never be a perpetual motion machine.

Nicholas Georgescu-Roegen wrote that "the Economic Process has a unidirectional irrevocable evolution" (from low to high entropy), and, in fact, all processes have such an evolution. The activities of an isolated mine, factory, or household—activities that ultimately turn valuable inputs into waste products—are all physically irrevocable. As Georgescu-Roegen and Herman Daly have emphasized, this replaces the typical freshman textbook conceptualization of the economy as a circular flow with an alternative conceptualization of the economy as entailing a one-way flow of matter and energy from usefulness to uselessness.

The type of economic growth occurring over the past four centuries has greatly accelerated this flow toward the ultimate maximum-entropy equilibrium from which no change could ever be made. In speeding up this final end, modern economic growth has produced not just more waste but also wonderful dissipative structures, which, as we learn from the work of Ilya Prigogine, can be the vehicles through which entropy is transformed from low to high. Dissipative structures include machines, human beings, and amazing new technologies like R. Dean Astumian's molecular motors, which function precisely because they are not in thermodynamic equilibrium and can use the resulting chaotic turbulence to do work.

Important as the conceptualization of the economy as a one-way process always dependent on fresh inputs from nature is, two caveats are in order. The first is that the word "isolated" is an important qualifier. A farm, for example, is not thermodynamically isolated from the sun. To construct a proper isolated system for agriculture, the sun has to be included as part of the system. The second caveat then comes into play: while a system as a whole cannot be returned to its previous state, a component of the system—such as a farm's soil—can be returned to its previous state as long as some other component of the system (such as the sun) suffers an increase in entropy sufficient to offset the decrease in entropy of the component that has returned to its previous state. In principle, this means it may be possible to perfectly recycle the materials of any production process that involves or could be made to involve the sun (until the sun dies).

Nicholas Georgescu-Roegen believed such perfect matter recycling to be impossible. Since such an impossibility does not follow from any of the scientifically accepted laws of thermodynamics, he proclaimed this impossibility as a new "fourth law of thermodynamics." Ecologists often believe that nutrient recycling in ecosystems is 100% complete and thus dispute the fourth law's validity. By contrast, economics Nobel laureate Paul Samuelson, in a forward to a book about bioeconomics, wrote that Georgescu-Roegen's "Law of Inevitable Dissipation of Useful Concentrated Matter … is good-sense certainty …: only if all useful copper and gold and helium could be 100 per cent recycled would a *perpetuum mobile* be possible. If pigs could fly, if my Aunt Sally were a stagecoach, if … and if …." So the truth of the fourth law remains in dispute.

As an aside, not only does $t_1 < t_2$ in a closed system imply $S_1 \leq S_2$, but $S_1 < S_2$ in a closed system also

implies $t_1 < t_2$. The second statement has been used to explain the direction of time, "time's arrow." For example, it has been proposed that since we live in a universe whose expansion causes the universe's entropy S to change, the reason we perceive time to increase in the direction it does is that that causes us to conclude that the universe's S is increasing, in accordance with the second law of thermodynamics, not decreasing, which would violate the second law.

4. ENTROPY AND NEOCLASSICAL ECONOMIC ANALYSIS

Neoclassical microeconomic theorists have no difficulty incorporating scientific constraints into their work. Indeed, as far back as Tjalling Koopmans' and Gerard Debreu's work of the 1950s, general equilibrium theorists have explicitly incorporated an irreversibility assumption into their work, as well as a more basic assumption that there cannot be outputs without inputs. Unfortunately, general equilibrium analysis is conducted at a level of abstraction that makes it almost useless for policy applications. (For example, applied general equilibrium analysis of trade policy assumes full employment, which begs the question of free trade's effects on employment.)

Neoclassical analysis of long-run economic growth has been conducted, not using microeconomic general equilibrium, but instead using a macroeconomic framework in which aggregate economic output is assumed to be a function $f(L, K)$ of the economy's stock of labor L and physical capital K, or sometimes a function $f(L, K, R)$ including natural resources R. Inspired by the work of Robert U. Ayres, one notices that elementary mass balance (conservation of matter) is lacking in $f(L, K)$—for where does K come from? Surely not from L, and most physical capital nowadays is not alive and able to reproduce itself! — and it is also lacking in $f(L, K, R)$ in the common formulation where K can be maintained at a constant or even increasing level while R goes to zero.

Neoclassical resource economists have constructed a useful body of microeconomic knowledge about how profit-maximizing firms exploit the environment, as well as the implications of different types of government regulation of that activity. However, these results usually do not treat the overall issue of natural resources in long-run growth.

Fundamentally, no economic technique will be able to definitively foretell the abundance and accessibility of particular natural resources far in the future, nor the need for those resources in the future. Educated guesses are possible, but novelty, contingency, and lack of predetermination mean that it is not possible to predict the future. There is no basis to confidently assume that the future will resemble past centuries, in which technological progress made it possible and desirable to exploit vast, hitherto useless stocks of many different materials found in nature.

5. INTERPRETATIONS AND MISINTERPRETATIONS OF ENTROPY

Equations (1), (2), and (3) are completely straightforward in their literal meaning. However, they give no intuitive insight into what entropy is. Other famous formulas in physics, such as $F = ma$ or even $E = mc^2$, deal with concepts we have at least some everyday familiarity with or understanding of, such as force, mass, acceleration, energy, and the speed of light. By contrast, Q/T or C_P/T are remote and hard to grasp. This has led scientists to come up with interpretations or explanations of entropy. Some of these interpretations are extremely valuable to scientific specialists in the field. Others are unfortunate misinterpretations. Neither the useful ones nor the useless ones have much importance to economists, but their continuing great influence on the historical development of the field makes it important to understand them.

5.1 Disorder, Randomness, and Probability

One of the most surprising consequences of Eq. (1) comes from analyzing the mixing of two ideal gases in a container whose walls allow no heat to flow. (An ideal gas has molecules that do not interact with their neighbors.) At first sight, the formula textbooks give for the resulting change of entropy—which is

$$\Delta S = -R\left[n_1 \ln\frac{n_1}{n_1+n_2} + n_2 \ln\frac{n_2}{n_1+n_2}\right], \qquad (4)$$

where n_1 and n_2 are the moles of the gases and R is the universal gas constant (8.31 J/(K · mole)—seems wrong, because since no heat can flow into or out of the container, it seems from Eq. (1) that the system's entropy cannot change. This doubt about Eq. (4)'s validity is incorrect. Entropy change is not $\int_i^f dQ/T$

(which is zero in this case) but rather $\int_i^f dQ/T$ over a reversible path. The mixing of two ideal gases is not reversible. The only way to calculate the entropy change accompanying the mixing of two ideal gases is to replace the original irreversible path linking the states before and after the mixing with a reversible path linking those same states, then performing the integration. Because entropy change is a function of the initial and final states only, entropy change along the new reversible path, which can be calculated using Eq. (1), must be the same as the entropy change along the original irreversible path. Along the reversible path, heat does flow, and the entropy change turns out to be Eq. (4), which is greater than zero.

The remarkable result that mixing two gases increases entropy has led observers to think that there is more to the idea of entropy than expressed by Eq. (1)—that entropy measures increasing mixing, or spatial disorder, or even some general notion of order per se. This is incorrect. Increasing entropy does not mean increasing disorder in any intuitive, layperson sense. For example, the spontaneous (i.e., entropy-increasing) evolution of an oil/water mixture is toward oil on top, water on the bottom, which is hardly increasing disorder according to any intuitive meaning of the word disorder.

Confusing the issue is that while increasing entropy does not mean increasing disorder in any intuitive, layperson sense, it does mean increasing disorder in an abstract, highly technical sense defined by the Boltzmann equation

$$S = k\ln W + S_0, \tag{5}$$

where k is Boltzmann's constant (1.38×10^{-23} J/K), where the constant S_0 is, by tradition but not by necessity, set equal to zero, and where W is defined by Eq. (6), where N is the number of particles in the system and n_i is the number of particles in energy level i:

$$W = \max_{n_0,n_1,\ldots,n_k} \frac{N!}{n_0!n_1!\cdots n_k!} \text{ such that} \tag{6}$$

$$\sum_i n_i = N \text{ and} \tag{7}$$

$$\sum_i n_i \cdot (\text{Energy level } i) = \text{Total system energy.} \tag{8}$$

(Note that $N!/(n_0!\, n_1!\ldots n_k!)$ is the number of ways that N distinguishable objects can be put into groups of n_0, n_1, …, n_k objects.)

While W defined in Eq. (6) is not a random variable and comes from combinatorics not statistics, Boltzmann christened it the thermodynamic probability, which has led observers to incorrectly think that entropy has something to do with randomness. On the other hand, another interpretation of Boltzmann's ideas does involve randomness. In this interpretation, the maximization operator in Eq. (6) is dropped, and W becomes a random variable, which takes on the value $N!/(n_0!n_1!\ldots n_k!)$ with probability

$$\frac{N!/(n_0!n_1!\cdots n_k!)}{\sum_{n_0,n_1,\cdots,n_k} N!/(n_0!n_1!\cdots n_k!)}.$$

The problem with this interpretation is that it allows the second law to be violated with nonzero probability, and such a violation has never been observed.

Boltzmann's ideas have grown into the important science of statistical mechanics, by which the entropies of various substances are explained, from a molecular level; the older classical thermodynamics described how to measure entropy but could not explain why it should take on a particular value for a particular substance. For extremely simple physical systems, statistical mechanics does imply that lower entropy states are those which are intuitively more ordered and less random. Thus, one mole of a substance in the form of a simple solid has lower entropy than if that mole is in the form of a simple liquid, and that in turn has lower entropy than if that mole is in the form of a simple gas. Also, random mixtures of simple substances have higher entropy than if those substances are ordered into unmixed, pure components, as seen in Eq. (4). However, most substances are not simple in the sense necessary for this intuition to be valid, as the oil and water example showed. Treatments that ignore this caveat—such as Erwin Schrödinger's famous 1944 essay "What Is Life"—should be read with the greatest caution.

5.2 Absolute Entropy

Notice that Eq. (1) defines only changes in entropy; absolute amounts of entropy are not defined. It is sometimes asserted that, by contrast, in statistical mechanics absolute entropy is defined—equivalently, that the constant S_0 in Eq. (5) must be zero. Modern treatments of entropy and its physical meaning show that this is not true. This controversy has no bearing on laboratory entropy measurements, which always measure entropy differences anyway.

5.3 Thermal versus Configurational Entropy

Occasionally, scientists find it useful to distinguish the contribution to entropy change represented by Eq. (4)

from other contributions to entropy change. When this is done, the component of the entropy change given by Eq. (4) is called configurational entropy or the entropy of mixing, and the other components of the entropy change are together called thermal entropy. The choice of the term thermal entropy is unfortunate, because it can mislead one into thinking that configurational entropy has nothing to do with thermal changes and hence that configurational entropy has nothing to do with Eq. (1). Configurational entropy is actually just as much a component of Eq. (1) as thermal entropy is.

5.4 Entropy and Free Energy

To calculate the entropy change when an event occurs, the resulting entropy change of the system's surroundings must be added to the entropy change undergone by the system itself. Under conditions of constant pressure and temperature, the former is the energy released by the system into its environment (traditionally represented by $-\Delta H$) times $1/T$, and the latter is ΔS. Adding these two entropy changes together yields a total entropy change of $\Delta H/T - \Delta S = (1/T)(\Delta H - T\Delta S) = (1/T)\Delta G$, where ΔG is known as the Gibbs Free Energy. Tables of Gibbs Free Energies of common substances are readily available, and such a table is all one needs to determine how entropy changes when an event occurs under conditions of constant pressure and temperature. This makes Gibbs Free Energy extremely valuable for chemists, but it should not obscure the fact that entropy is the fundamental physical quantity and that Gibbs Free Energy is a derived quantity of no independent physical importance. Similarly, Helmholtz Free Energy $F = U - TS$ (where U is internal energy) is useful in situations where volume and temperature T are constant, but it is not a fundamental quantity either.

5.5 Information

An equation that looks like Eq. (5) with Eq. (6) appears in the field of information theory, with, of course, entirely different meanings for the symbols. This does not imply any connection between entropy and information, any more than the similarity between the formula for the sum of scores on two student examinations ($e_1 + e_2$) and the formula for the sum of the energy of two hotel elevators ($e_1 + e_2$) implies any connection between student examination scores and hotel elevators. The entropy of information theory, while no doubt quite useful to practitioners of that field, has nothing to do with the second law of thermodynamics.

5.6 Minimum Energy Requirements

Near the end of Section 3.1 it was shown that the amount of energy E degraded during an irreversible process is the temperature of the coldest heat reservoir at hand (call it T_0) times the entropy change of the universe brought about by the irreversible process:

$$E = T_0 \Delta S. \tag{9}$$

Suppose we desire to obtain pure copper from oxide copper ores of the type often found in Africa. Because CuO is thermodynamically stable at room temperatures, breaking up CuO into copper and oxygen would cause an entropy decrease, so it cannot occur spontaneously. Some authors have calculated the amount of the entropy decrease, then used Eq. (9) to assert that T_0 times that amount is a minimum amount of energy necessary to liberate the copper from the CuO. (Their general proposition is that there is a definite theoretical limit to the minimum energy input to a process and that there is no substitute for this minimum energy.) There may be some sense in which this is true, but no amount of an energy commodity (such as coal or electricity) is needed to obtain copper from CuO. What is done is simply to introduce sulfuric acid into the environment of the copper oxide; a typical reaction is $CuO + H_2SO_4 \rightarrow Cu^{++} + SO_4^{--} + H_2O$, which increases entropy and is thus spontaneous. This example shows that there is no minimum "energy commodity" requirement for physical, chemical, and hence industrial processes.

5.7 Entropy and Energy

There is no such thing as low-entropy energy. Entropy is a characteristic of matter, and a substance's entropy changes when the substance reversibly absorbs or emits energy (or when an irreversible process involves the substance and a reversible process linking the same initial and final states would entail the substance's absorption or emission of energy); see Eq. (1).

5.8 Entropy Theory of Value

A glance at a table of standard entropies for different substances will show the error of believing that the lower an object's entropy, the higher its economic value. Similarly, each time one heats a kettle of water

in a kitchen, the water becomes more useful at the same time as its entropy increases. Adding impurities to iron to create steel simultaneously increases the entropy of the metal and increases its economic value. The entropy of oceanic pollutants is decreased by their concentration in the flesh of oysters, but this decrease of entropy makes the oysters unsafe to eat, thus less valuable to humans. So there is no general connection between low entropy and high economic value, nor between high entropy and low economic value. It is true that the process of production and consumption turns valuable inputs into useless waste, with the useless waste having, as a total assembly, higher entropy than the valuable inputs. However, this generalization only applies to assemblies as a whole, not to their component parts.

Furthermore, what is valuable will always be partially determined by psychological, historical, sociological, and in any case nonphysical traits unrelated to entropy; a personal computer made in 1987 was quite valuable then, but would be considered junk now, even if its physical traits are unchanged. Even a simple container of oxygen, whose entropy decreased when it was extracted from air, can be useful and valuable in one situation—helping hospital patients breathe—and be an unwelcome hazard in another situation, such as near a stove.

6. CONCLUSION

The entropy law implies that economic production will never be entirely self-contained or self-replicating, let alone self-expanding. Production will always require fresh inputs from nature, and nature's size, both of terrestrial stocks and of the rate of solar flow, is finite.

SEE ALSO THE FOLLOWING ARTICLES

Aggregation of Energy • Economic Systems and Energy, Conceptual Overview • Economic Thought, History of Energy in • Entropy • Evolutionary Economics and Energy • Exergoeconomics • Exergy • Exergy Analysis of Energy Systems • Thermodynamics and Economics, Overview • Thermodynamic Sciences, History of • Thermodynamics, Laws of

Further Reading

Atkins, P. W. (1984). "The Second Law." Scientific American/W. H. Freeman, New York.

Beard, T. R., and Lozada, G. A. (1999). "Economics, Entropy and the Environment: The Extraordinary Economics of Nicholas Georgesçu-Roegen." Edward Elgar, Cheltenham, UK.

Biswas, A. K., and Davenport, W. G. (1994). "Extractive Metallurgy of Copper." 3rd ed. Pergamon Press, Oxford.

Daly, H. E. (1993). The steady-state economy: Toward a political economy of biophysical equilibrium and moral growth. *In* "Valuing the Earth: Economics, Ecology, Ethics" (H. E. Daly and K. N. Townsend, Eds.). The MIT Press, Cambridge, MA.

Dugdale, J. S. (1996). "Entropy and Its Physical Meaning." Taylor & Francis, Bristol, PA.

Faber, M., Proops, J., and Speck, S. (1999). "Capital and Time in Ecological Economics: Neo-Austrian Modelling." Edward Elgar, Cheltenham, UK.

Gaskell, D. R. (1981). "Introduction to Metallurgical Thermodynamics." 2nd ed. Hemisphere, New York.

Georgescu-Roegen, N. (1971). "The Entropy Law and the Economic Process." Harvard University Press, Cambridge.

Goldstein, M., and Goldstein, I. F. (1993). "The Refrigerator and the Universe: Understanding the Laws of Energy." Harvard University Press, Cambridge, MA.

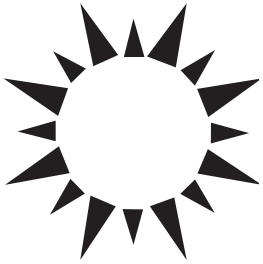
Lozada, G. A. (1999). The role of entropy and energy in natural resource economics. *In* "Bioeconomics and Sustainability: Essays in Honor of Nicholas Georgescu-Roegen" (J. M. Goudy and K. Mayumi, Eds.). Edward Elgar, Hants, England.

Toman, M. A., Pezzey, J., and Krautkraemer, J. (1995). Neoclassical economic growth theory and 'sustainability.'. *In* "Handbook of Environmental Economics" (D. W. Bromley, Ed.). Basil Blackwell, Cambridge, MA.

Wicken, J. S. (1987). "Evolution, Thermodynamics, and Information: Extending the Darwinian Program." Oxford University Press, New York.

Zemansky, M. W. (1968). "Heat and Thermodynamics: An Intermediate Textbook." 5th ed. McGraw-Hill, New York.

Environmental Change and Energy



I. G. SIMMONS
University of Durham
Durham, United Kingdom

1. Human Access to Energy
2. Conceptual Frameworks
3. Historical Stages of Access to Energy
4. Overview from Today's Perspectives

Glossary

agro-ecology Crop production as a component of an ecosystem that includes not only the desired species but also, for example, the weeds, water, and nutrient relations; soil characteristics; and human inputs.

cosmogenesis The creation of the whole cosmos in a single event. In some societies it has been seen as an annual renewal of nature.

developed countries Those nation-states with an industrial economy and generally high gross demestic product and commercial energy use per capita, though not necessarily evenly distributed.

extrasomatic Outside the body. In using a bow and arrow, the energy deployed is somatic but the technology used to channel it is extrasomatic.

heterotrophic Nourished from outside—that is, of all organisms incapable of photosynthesis, which constitute most living things except green plants.

intensification Delivery of more product per unit area per unit time, for example, double-cropping of rice or shelling with heavier ordnance.

less developed countries (LDCs) Those nation states with a low level of industrial development and usually a low consumption of commercial energy. Most have high rates of population growth. Sometimes referred to as low-income economies (LIEs).

moa The large flightless herbivorous bird of New Zealand, equivalent to the ostrich, emu, and cassowary. Extinct.

noosphere Concept of Soviet biologist V. I. Vernadsky that there would be a global envelope of thought equivalent to the lithosphere and biosphere.

organochlorine A class of chemical developed in and after World War II with biocidal properties; the most famous is DDT. Without natural breakdown pathways, they have accumulated in ecosystems. Still used in LDCs.

overkill A popular but graphic term for the mass extinction of genera of animals. Originally conceived for late Pleistocene times but extended to the disappearance of the dinosaurs and held in reserve for a "nuclear winter."

pastoralism A mode of life involving herds of domesticated animals pastured over large areas, with seasonal migrations that track the availability of water and animal food.

permanent cultivation The replacement of a preexisting ecosystem with a permanently manipulated agroecology in order to produce crops as frequently as climate and soil conditions allow.

photosynthesis The chemical process of the basis of life, in which the energy of the sun is transformed into the tissues of plants.

precautionary principle A commandment of environmental management in which actions are avoided if there is a probability that they will lead to environmental degradation. Like certain other Commandments, breaches are frequent.

savanna A regional-scale ecosystem of tall grasses with interspersed trees. It usually experiences fire at least once a year and in Africa may well have coevolved with the genus *Homo*.

shifting cultivation Crop production in which the land is allowed to revert to a wilder state once the soil is exhausted or weeding is too onerous. Much derided by colonial administrators but ecologically appropriate in times of low population growth.

teratogenic Of fetuses (usually but not exclusively applied to humans) that are deformed by the mother's ingestion of certain chemicals. Dioxins (of which some are more toxic than others, such as 2,3,4,7,8-TCDD) have been detected as a major cause.

1. HUMAN ACCESS TO ENERGY

In the planet's ecological webs, humans are heterotrophic: they cannot fix energy from sunlight like

 © 2004 Elsevier Inc. All rights reserved.

plants. So in its 4-million year perspective, human history can be seen as a set of stages of access to energy sources. Early hominids relied on recently captured energy as edible plant and animal material, and at one stage (whose chronology is disputed but may have been as long ago as 0.5 my) gained control of fire at landscape scale. This reliance on solar energy as biomass is also characteristic of preindustrial or solar-based agriculture, but here the concentration of energy is enhanced by making it available in delineated fields or herded animals. Hence, many more people could be fed off a smaller area, especially if water control could be added to the technological repertoire. Once fossil fuels were exploitable for steam generation, solar power could be subsidized by energy from the stored photosynthesis as represented by coal, oil and natural gas. Energy concentration in these fuels was higher than more recent sources, and this trend is continued by the harnessing of the energy content of the nuclei of uranium and hydrogen.

2. CONCEPTUAL FRAMEWORKS

A number of concepts are embedded in this overview of human history. The first is that of a stage of simply garnering the usufruct in the form of plant and animal material that could be gathered, hunted, or scavenged. Most early human societies were probably mobile, so that these early species of the genus *Homo* could be said to have lived in an environmentally determined fashion. Thus, another concept is that of control and the rest of human history can be viewed as a desire to increase the degree of mastery over environmental components, of which energy has been arguably the most important. The first key step was regulation of fire away from the hearth and into the landscape, with the multiple benefits thus brought about, as amplified later in this article. Fire is also at the center of the developments of industrialization, when the combustion of coal released energy that could be controlled via a steam-using technology; its concentration was such that even mobile devices could be pioneered. Direction of the energy of the atomic nucleus has been achieved for fission but for fusion, only the awesome threat of the hydrogen bomb remains available.

Another major concept is that of technology. Apart from eating it as food, humans must access the usefulness of most energy sources via an extrasomatic device. They must learn to preserve and transport fire, for example; crop harvesting is enhanced by the use of an iron blade produced by smelting metal ores with charcoal; the steamship carried goods, ideas, and troops from one end of the earth to the other. The energy made available to growing populations in the 19th century from the potato comes not only from the carbohydrates inherent in the tubers but from its better growth when encouraged by bagged fertilizers. Electricity is a flexible carrier of energy that has facilitated all kinds of environmental change, not least in the transmission of ideas about how to control the environment. In all these devices, the hope is that the group will improve their net energy yield—that is, the ratio of usable energy harvested to energy invested as labor, draft animals' food and the making of the technology itself. The major stages in human access to energy before the 20th century and the quantitative results are summarized in Table I.

Technology, however, is not simply a set of inventions. It is the product of, and embedded within, human societies. This means that the strict rationality and "scientific" qualities associated with it are not necessarily found in every application. To begin with, much technology was prescientific, in the sense that it was the product of practical-minded men (sic?), who happened on a better way of doing things. Only after the 17th century did the theory precede the technology in truly scientific procedures. (It has been argued that the steam engine, for example, was pioneered by practical men with no knowledge of the theories of physics.) Certain possibilities follow: it is feasible to reject technology on religious, economic or simply prejudicial grounds, as medieval Islam is said to have regarded the wheel

TABLE I

Energy Consumption through Time

Economy	Energy consumption (GJ cap day)	Sources
Staying alive only	0.0001	Plant and animal foods
Hunter-gatherers	2.0	Wild plant and animal foods plus fire
Preindustrial agriculture	20–40	Domesticated foods plus fire, charcoal
Eve of industrialization	70–100	Foods plus fire and plentiful fuels such as charcoal
Britain in 1800 CE	20	Foods plus wood and a little coal
Britain in 1900 CE	116	Foods plus developed coal-based economy

and sensible nations have regarded atomic weapons. Hence the distribution of energy technologies is likely to be uneven, as is shown by the access to commercial energy today, or by the way in which the adoption of fossil-fuel based technologies took off in the West after ca. 1750 AD but not elsewhere. So access to, and use of, energy in any society is mediated by technologies that are themselves subject to the organization of that society, with all its potentials for, especially, inequalities of distribution.

The relationship of all these kinds of energy use to environment has of course varied according to the intensity of available energy and the number (and density) of people, but a few historical generalizations can be made. The first is that the environment is a source of energy that may be altered as the energy is garnered. An animal population severely depleted by intensive hunting is an obvious example. A savanna fired frequently to drive out game will undergo species' shift as fire-intolerant plants are extirpated. A strip mine changes the ecology in a radius measured in tens of kilometers. Second, while many of these changes may be deliberate and desired by a society, there may well be side effects which are undesired and degradatory. The smoke from unregulated coal burning has been the source of much human ill health in cities since the 13th century, for instance; the fallout from the nuclear accident at Chernobyl in 1986 sterilized large areas of formerly productive land. The possibility of negative consequences of human actions in matters of energy access and use was pointed out in the myth of Prometheus, which retains a certain resonance today.

3. HISTORICAL STAGES OF ACCESS TO ENERGY

None of these stages, with exception of the last, can be given precise dates. Thus the chronology must be read as having wide margins for different places. Agriculture, to quote an obvious example, may have been dominant in southwest Asia by 8000 BCE but only arrived in Australia with European colonization in the 19th century. Coal was used to console Roman soldiers on Hadrian's Wall but is only a significant global influence from about 1800 AD/CE.

3.1 Hunters and Gatherers (4 MY–8000 BCE)

This stage encompasses (1) human biological evolution in the transitions from early African forms of *Homo* to the anatomically modern human of *Homo sapiens* at *ca* 35 ky with a worldwide distribution and (2) cultural development through the various levels of the Palaeolithic and Mesolithic to hunter-gatherers described at and after European contact, usually described as near-recent groups. There are a few remaining groups but a way of life that comprised 100% of the human population in 10,000 BCE now accounts for something less than 0.001% and is shrinking. The term "hunting and gathering" also includes fishing and scavenging; it may also be written in the reverse to suggest the dominance of plant material in most of their life ways.

The environmental relations of hunter-gatherers are dominated by the practice of food collection rather then food production. This is not to say that there is no environmental manipulation in the cause of gaining nutrition or status, but that it is practiced in a context where there is (1) a low population density, (2) no conscious attempt to manipulate the genetics of biota with the exception of the dog, and (3) a mythological context of nonseparation from the natural world and of annual cosmogenesis. For many years, academic interpretation of these relationships postulated a "child of nature" construction in which hunter-gatherers lived in complete harmony with the natural world, taking little from it and exhibiting complete reverence. Yet the development of an ever more efficient technology was characteristic of that age, as it is of industrial and post-industrial times. Devices such as the bow, spear thrower, and blowpipe harnessed the energy of a group of human muscles to direct a missile down a narrow channel. Net energy returns were of the order of 1:5-15 and might rise to 1:40. Controlled fire seems to have been used event in wet places like the lowland tropical forests,!QJ; the Californian Coast redwoods, and the moorlands of England and Wales.

Some examples of hunter-gatherer life indeed fitted a Golden Age view. Work in the 1950s and 1960s on groups like the !Kung San Bushmen of the Kalahari seemed to show that with relatively little effort, a well-nourished population could persist even in an unforgiving environment. It is certainly true that given a large and preferably diverse territory, an abundant staple food (the case of some of the !Kung it was the mongongo nut), and a slow rate of population growth, hunter-gatherers could exert relatively little effect on their environment. Some indeed adopted conservationist attitudes to their resources. Nevertheless,

many historical examples have been shown to have brought about considerable environmental changes.

The most thoroughgoing of such manipulations has been the discovery that the introduction of humans into regions hitherto lacking their presence has resulted in the extinction of large groups of animals, with an emphasis on the loss of large (>50 kg adult body weight) herbivore mammals and flightless birds. The initial scene for this revelation was the High Plains of North America in the late Pleistocene (12-11 ky) but the same phenomenon has been detected in islands in the Pacific and Indian Oceans at later dates, such as New Zealand/Aotearoa (900 CE), Hawaii (600 CE), and Madagascar (700 CE). By contrast, no such overkill phases have been noted for Africa, where humans evolved and spread gradually. Another relationship between hunters and animals sees massive kills but no long-term effects on the populations of the latter since they are so massive (or so rapid at reproduction) that humans at hunter-gatherer densities cannot bring about extinction. This seems to have been the case before European contact on the High Plains of North America. Here, as in the northern forests of Canada, such equilibria did not survive contact with the more energy-intensive economies, whether these were basically agricultural as with the Hudson's Bay Company and its fur traders or the railroads with their hungry and bored laborers. Such interrelationships have been found in recent food-collecting societies: the BaMbuti of the Zaire Basin have traded meat outside their forests to the detriment of duiker populations, for example. Some groups have internalized industrial technology, as with the Inuit, the rifle, and the snowmobile, for example, leading to controversy about their access to traditional prey such as the bowhead whale. Interestingly, no such contentions seems to have arisen over plant life, suggesting that the mythic role of dead animal flesh is deeply rooted in human societies: an aspect of control, perhaps.

At an early stage, hominid energy use must have been totally metabolic in nature, confined to that needed for growth, movement, and reproduction. As cultures became more complex, demands for additional energy grew, with any orientation toward meat consumption likely to cause more impacts than reliance on plant material. Notable technological developments seem oriented to the procurement of meat rather than plants, and in the great artistic monuments of the Upper Palaeolithic in the caves of France and Spain there are no paintings of roots and berries.

3.2 Pre-Industrial or Solar-Powered Agriculture (8000 BCE–1750 CE)

In energy terms, agriculture represents an *intensification* of access. To collect enough energy for survival, an average food collector needed some 26 km^2 of territory as a support system. Subsistence farmers with adequate water supplies can manage on 0.1 ha per head and achieve net energy returns of 1:20. Environmentally, this is achieved by breeding plants and animals in cultural images, by altering ecosystems so as to direct their energy flows at human societies and by coping with the unforeseen consequences of these alterations of the previous patterns. The actual biological channels through which these new patterns were directed varied in time and space. In southwest Asia in 9000 to 8000 BCE, wheat, barley, sheep, and goats were central; in southeast Asia at the same time it was rice above all; in the Americas somewhat later, the potato, tomato, maize, and llama were key elements. Yet the energetics had much in common, as had the environmental consequences.

Compared with hunter-gatherers, the energy is garnered closer to home. There is usually a permanent settlement and the food energy is produced so as to minimize expenditure on travel—hence the development of the field, where a crop is kept within bounds so that its maintenance and yield is under the eye of the farmer, be it wheat or cattle. Energy has to be invested in sowing, weeding, fencing, predator control, storage, and myriad other tasks so that the nearer at hand, the greater the probability of a positive energy balance at dinnertime. The domestication of animals meant that a seasonal round of tapping plants not edible by humans and turning them into milk, meat, blood, and skins could develop on the arid or mountainous fringes of cultivable areas.

Not surprisingly, humans enlisted help in the new diversity of tasks required by the novel economy. Draft animals are not really energy efficient in the sense that they consume food even when not working but they can deliver concentrated power at key times, such as spring ploughing or carrying metal ores to the smelter. They relieve humans of wearisome tasks and may furnish symbols of status. Equally important ameliorations of work came with the harnessing of solar energy in the forms of water and wind power. Perhaps the most important was the realization that precipitation (itself a consequence of solar energy distributions) could be stored and then released at times critical for crop growth or animal watering. Coupled with energy investment in, for

example, terracing, crop production was not only intensified but made more reliable. Since storage and distribution often meant that water had to be lifted, the application of wind power to the chore was another gain. The transformation of energy in wind and falling water to rotary motion also facilitated the technologies of the mill, useful for grinding corn and metal ores as well as lifting water. So as with the bow and arrow in the Upper Palaeolithic, a whole range of extrasomatic technology is brought into action to achieve both survival and wider cultural aims.

Food production is thus the core of preindustrial agriculture and at times of Malthusian pressure, virtually any environmental manipulation becomes acceptable in order to increase the quantity of crop. There are three main types of agroecology: shifting cultivation, permanent cultivation, and pastoralism. The first of these, sometimes regarded as a primitive system, is improved by conversion to permanence, but this is probably a colonialist view of a sophisticated piece of adaptation to forest and savanna environments. Its ecology focuses on the clearance of a patch of near-natural vegetation and the take-up of the nutrient capital into crops. Cleared vegetation may be burned to add to the available stock of nutrients. As fertility falls or weeds become tiresomely prevalent, the patch is abandoned and another opening replaces it. Revegetation then builds up the nutrient stock and smothers the weeds. The crop plants in fact much less efficient at energy capture than the original vegetation, but that loss is traded for the cultural acceptability of the food, herb, spice, and medicinal plants that replace it. Moreover, a cleared patch in a matrix of near-natural vegetation does not readily leak nutrients to runoff, so energy expenditure on adding manure or other sources of nutrients is limited compared with permanently cleared areas. The system is not easily intensified, however, and has to be transformed when population growth necessitates a return to a cultivable patch before the vegetation has restored the nutrient levels. Though generally thought of as a tropical system (especially in forests and savannas), it was used in the deciduous forests of Europe in the Neolithic (7000–5500 BCE) and as late as the early 19th century CE in the Boreal forests of Finland.

The historical energetics of food production are without doubt dominated by systems of permanent cultivation, found all over the world except in polar and subpolar regions. Along with sedentary settlement, they represent one of the truly revolutionary shifts in human-environment relations and the energy relationships reflect the new conditions. The basic agroecology is that of the permanent replacement of the preexisting ecology with crop plants and animals and of progressive intensification of output as populations grow. The energy flow of the land is directed toward the crop biota and all else is labeled as weeds or pests and if possible killed or driven away. As with shifting cultivation, maintenance of nutrient levels is at the heart of the system, and both energy and materials have to be used in their upkeep. Cropping is, of course, a major factor in nutrient loss from the fields, but bare soil is vulnerable to losses from runoff and many systems have adopted strategies for minimizing such losses: cover plants are one such, but the most universal is the terrace, found in a broad belt from the equator to about 55°N and the equivalent southern latitudes.

To harvest the energy in a crop, there has to be considerable energy investment. At some time, there has to be breeding of the best strains of plants and animals for local conditions, with no necessary guarantee of success in prescientific times. More obviously, consider the average field in a preindustrial society. The soil has to be turned over for a start. A spade or digging stick may suffice, or draft animals may be harnessed to a plough with a wooden or metal-tipped ploughshare. The energy pathways can be immediately seen: human energy in the first and animal energy in the second. The animals have to be fed and may thus form part of the energy flows taken by the post-harvest crop. Or indeed they may compete for energy with the human population by needing special croplands: "horses eat people" was a proverb in early modern Hungary. If metal is used then there is an ancillary need for energy in smelting and working the iron. Sowing and weeding need investment, as does the fencing, which keeps the wrong animals out of the crop or the wolves from the sheep at night. Children are very useful for shepherding and bird scaring but have to be fed. Harvesting of plants is generally labor intensive, and only careful built storage will keep at least some of the rats out. Water management also requires energy investment, whether the problem is too much of it or too little. Lastly, many plant foods require cooking since the ability of the human gut to digest cellulose is restricted. Our species is better at processing animal tissues, but the thermodynamics of conversion of plant material to animal tissue ensure that meat in preindustrial societies is a luxury. Yet in many parts of the world, such systems were stable environmentally, nutritionally adequate and even provided surpluses.

On the fringes of settled areas, environments not conducive to permanent agriculture were often

found. Characteristically, these were semiarid in nature, with long periods without rainfall and no permanent surface drainage. Mountains form another category. These could nevertheless be tapped for human use via flocks of domesticated animals in the system known as pastoralism. The fundamental agroecology was the cropping of cellulose in plants inedible to humans by domesticates such as sheep, goats, and llamas (and many more species adapted to particular environments, such as the yak), whose sustainable products could be stored for later consumption and trade. Many dairy products have their origin in such systems and surplus animals might be sold off for meat (and skins) though eaten only rarely by the pastoral communities themselves, who, however, sometimes tapped blood from the live beasts as a source of protein. The energetics of this system are more like hunting and gathering than permanent cultivation in the sense that there has to be continual movement, in this case so as not to overcrop the vegetation. One form of the life way involves the movement up mountains in the summer. This transhumance was common in Europe until the 19th century, and a form of it persisted among people of Basque origin in Nevada until very recently. The energy flow of this system is generally completed by the selling off of most of the embedded energy in the animals and their products to sedentary societies, and there is no investment in nutrient levels. Once the pastures are visited often, then the photosynthetic efficiency declines as the species' mix shifts toward the arid tolerant and more bare ground is seen.

Fishing was often an efficient method of garnering energy since the investment came more or less free as river water, lakes, or wind in sails. Energy investment in construction of boats, weirs, and traps was generally long lasting: only nets required frequent renewal. Yet fish were usually an adjunct to diets: only in a very few places were agricultural societies dominated by their consumption, though as an addition to the palate of flavors they have long been popular.

One outcome of these systems was that they could often provide a surplus. This could be used to fuel, so to speak, population growth in the producing society. Alternatively, it could be mulcted off to feed a superior class who did not have to toil in order that they might eat. Instead they might effect environmental change in other directions. One direction might be that of developing certain nonagricultural skills full time: itinerant metalsmiths are postulated for prehistoric Europe, and presumably the astronomers of ancient Babylon did not grow their own barley. Likewise, the management of forests could become a specialized occupation, with a society recognizing that the maintenance of its chief energy source apart from food was a key element in its survival. This was especially so in the great ages of empire building and trade in wooden ships, when ship timber became a key instrument of national policy for countries such as Great Britain and France. At a less material level, philosophers and university faculties might not have to bake their own bread, though interestingly monastic houses and rural clergy were usually required to be self-sufficient to some degree, even though the latter might be entitled to a tithe of local crops. The most visible product of diverting energy into the hands of an aristocracy (who in a fascinating piece of etymology in English wielded "power") was pleasure and conquest. The latter came by feeding large numbers of soldiers, usually by local communities and on an involuntary basis, diverting the energy stored after the last harvest into the army's day-to-day maintenance. The former can be seen in landscape relics such as hunting parks and gardens. Many a European city park was once the playground of the King or the Archbishop: St. James's and Hyde Parks in London had such a royal origin. The gardens of the Alhambra Palace in Granada were the work of the governors of the moorish province of Andalucia. Lastly, energy might be invested in the afterlife: only very productive irrigated cereal-growing provided enough surplus to feed the laborers who built the Giza pyramids in Egypt.

Above all, perhaps, surpluses made possible the city. Although small by today's standards, they represented concentrations of stored energy in the form of structures. They lived heterotrophically off imported energy as food and fuel, which they transformed into all kinds of wastes. (The basic ecology of today's cities is not very different.) Though gardens might produce food and water mills some energy, their citizens were mostly dependent on their hinterlands, for which a certain nostalgic reverence turns up as early as classical Rome. Some of the energy and matter was however turned into knowledge and stored not only in people's heads but in writing. It is difficult to see how the growth of abstract thought could have taken place without the concentrations made possible by the energy surpluses allowed by well-conducted agriculture. Even before the steamship, the control of wind power allowed the improvement of food production by its facilitation of transferring new crops to new places. The role of the potato in industrializing Europe followed on from its earlier spread out of South America, for example; the penetration of all kinds of European practices into

the Americas can scarcely be contemplated without the horse. So good energy trappers were taken around the world. It might be arguable that the major takeoff of human population growth usually dated to about 1650 ce was in part fuelled by improved energy capture and that better crops of all kinds were implicated in the improved nutrition. (It can be added that in some countries the use of energy to boil water to make tea and coffee has been suggested as a health-producing measure that added to survival rates. Maybe).

The theme of surplus energy produced by preindustrial agriculture has one apogee: it fed the people who conceived and built the pyramids, the Gothic cathedrals, Angkor Wat, and the Ryoan-ji. Somewhat imperfectly, it must be admitted, Mozart got fed as a result.

3.3 Industrialism (1750–1950 CE)

This is sometimes called the age of fossil fuels. Without doubting the accuracy of the label, note should be made of the exploitation of these energy sources in the years before their full flowering. Natural gas, for example, was piped in hollow bamboos from natural seepages into Chinese towns in medieval times and lit the streets. Coal exports from the port of Newcastle in northeast England fuelled breweries and laundries in London in the 14th century CE. But all this was as nothing compared with the explosion of coal use (followed by oil in the late 19th century and natural gas in the mid-20th century) once the potential for combustion of these fuels to produce steam, kinetic energy of all kinds, heat, and electricity had been realized. Further, once past the mid-18th century, the technology might follow a science in that hypothesis testing and theory building preceded the construction of the machines.

The concentrated energy content of mineralized photosynthesis is central to the idea of surpluses. The environmental relations are characterized by the impacts of getting the fuels and then of using them via, as before, the mediation of technology. The key change from the previous era is the highly positive ratio of energy investment to energy yield at, for example, the pit head and the well head, with ratios of 1:20–30. Even when postextraction processing is complex, as with oil (though not with coal, the earliest fuel of this kind to be used on a massive scale), the ratio is still highly favorable. Only with the large heat losses encountered in the generation of electricity in conventional power stations do the ratios begin to drop toward preindustrial magnitudes.

The energy-environment story of the 18th and 19th centuries is that of the adoption of fossil fuels as the basis of transport, communications, and manufacturing in a set of regions that became what are now called the developed countries (or sometimes the north or the high-income economies [HIEs] or simply the rich) and the failure of other parts of the world to undergo the same transitions. They became the less developed countries (euphemized as the developing countries, or sometimes referred to as the south, the low-income economies [LIEs], or the poor). The reasons for both are complex and outside the scope of this work. The energy available in classical Greece, in 17th century Europe, and China in 1880 CE was about the same but the adoption of fossil fuels brought about radical change. The core areas in, say 1850 CE, were western and central Europe together with eastern North America. Soon to be added were eastern Europe and western Russia, Japan, a few parts of the northern shore of the Mediterranean, and other patches of North America. On the map, the sign of the times was the railroad. Those heartlands are still significant in the sense that they acted as core areas for whole nations to adopt enough of the new technologies to ensure their future access to fossil energy supplies, so the present-day situation in which access to energy (usually measured as per capita consumption of commercial energy) replicates roughly the precedence established in the 19th century. Though wood as a fuel has declined greatly in relative terms (its energy content is about 14 MJ kg, compared with hard coal at 29 MJ kg), in absolute terms its use is still rising due to population growth in the LIEs (Fig. 1).

The myriad ways in which these fuels are extracted and utilized exert considerable environmental impacts, and many books have been written about each of them. There is a basic ecology common to all in which the hydrocarbons are lifted from the earth, processed, and stored, and then transformed into their end uses. These latter may be direct (as heat, for example, to power boilers) or as embedded energy as in fertilizers and biocides. All kinds of machines mediate at every stage and millions of different items are manufactured. The environmental impact of some is obvious and direct: the bulldozer is a good instance. Others are more subtle: stored electricity makes possible battery-operated TVs in remote Asian forests whose people are then able to see HIE consumerism at work, or indeed at play or at prayer. The bulldozer is the culmination of earth-moving equipment that started with the steam crane; microelectronics carry on the tradition of the electric telegraph. In that sense,

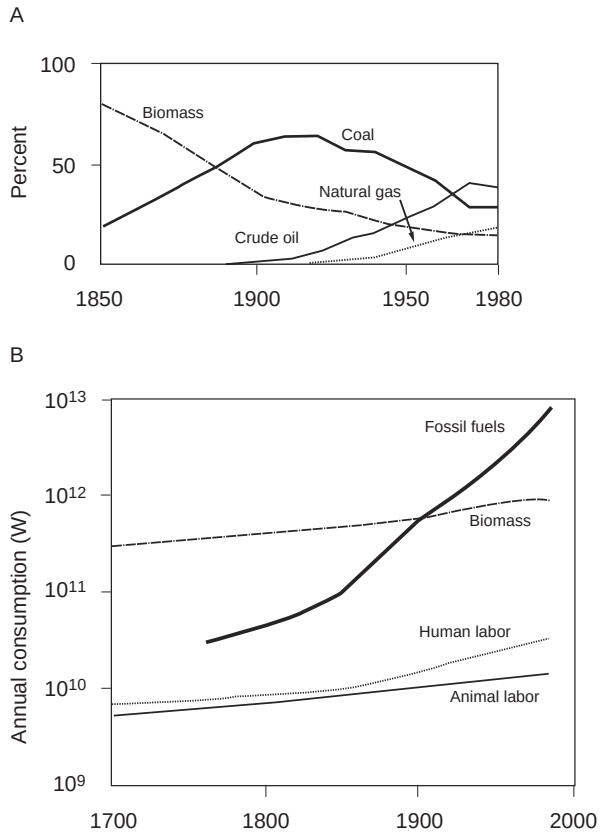


FIGURE 1 The transitions to fossil fuels. (A) shows the global proportions of primary energy for the period 1850–1980 CE and (B) the absolute amounts over a longer time period. Though biomass has declined in relative terms, the rising populations of LDCs (especially in Africa and Asia) have increased actual demand for wood fuel. From V. Smil (1991), "General Energetics." Wiley-Interscience, New York and Chichester, Fig. 10.2 (part), page 185, and Fig. 14.6, page 303. Copyright © 1991. This material is used by permission of John Wiley & Sons, Inc.

the latter exemplifies the way in which energy as electricity above all carries ideas, which include those devoted alike to changing and protecting the environment. It has without doubt made possible the global coalescence of ideas which have on the one hand allowed corporations quickly to find states willing to accept GM crops but on the other hand pioneered agreements like the Kyoto protocols, which all civilized nations have agreed to ratify.

The historical effects of industrialization on food production are seen in many environmental changes in the 19th century. Many temperate grasslands were ploughed to grown cereals for export to the rapidly growing populations of industrializing regions. The demand for meat also allowed vegetation change in other grasslands as grazing densities increased. All such changes were facilitated by the takeover of areas previously used by preindustrial people (nowadays labeled as "indigenous"), whether officially colonized (as in South Asia where "empty" lands were converted to tea and coffee by British colonial authorities) or internally reallocated, as on the High Plains of the United States. In the case of meat, the technology of refrigeration was critical, as was the reliable transport offered by railroad and steamship. Although to be more a feature of the post-1950 period, the disappearance of the seasons in the rich world began, as asparagus, strawberries, and flowers consumed transportation energy in their journeys from the warm and sunny to the still-emerging-from-winter shops of London, Paris, and New York. The loss of the local, as energy for transport plummeted in price, has its origins in the 19th century, though clearly not confined to that period. One short-lived counter-trend came about in the 1890s in Paris, where there was enough horse-dung to fertilize and heat an export trade in winter salad crops within the city boundaries.

The theme of increased material production can be carried beyond that of food. One result of immense environmental significance was the burgeoning appetite for material possessions, leading to environmental manipulations of all kinds. This was exacerbated by the craving for the new, so that reuse and recycling (a feature of poor societies everywhere) became far less common. Again, the introduction of electricity (as late as the post-1945 in many rural parts of Europe) impelled many authorities to speed up hydropower development, with its many environmental consequences. This is just one of many environmental outcomes of the extraction of energy sources: huge open-cast coal mines in central Europe and oil wells in Mesopotamian deserts were all visible before 1950. The overall impression is of self-amplifying systems in which the extraction of energy is followed by the development of technology to use it and this in turn spawns further environmental change. The motor vehicle is a case in point, where impacts include not only the emissions to the atmosphere but also the paving over of soils, with implications for runoff rates and flood levels on lower ground. Hence, all industrial societies are replete with embedded energy: obvious in the Chicago coal yards but less visible in the suit from Jermyn Street.

As in the previous energy era, opportunities for pleasure were increased. The form of travel now called tourism was made possible by the steamship and the railroad: before the 19th century, a pleasure visit from western Europe to Egypt was too great an

undertaking for almost everybody except the aristocracy. After about 1860, Britons might holiday in Switzerland, beginning a love-affair with the mountains denied them in their own small-scale islands, one which took them to the summit of Everest in 1953. (Actually it was a New Zealander and a Nepalese who first reached it, but imperial attitudes were still strong. The relevance of industrialization is in the specialized gear now used, and the fact that an expedition has had to be mounted to clear up the "litter" from such climbers including some bodies.) The ability to travel to relatively remote places led to conflicts over environmental change: the U.S. National Parks post-1945 slogan of "parks are for people" led to developments that people of later worldviews have found difficult to demolish, in all senses of the word. In a secondary frame, urban electricity made possible all kinds of entertainment, which may well help to inculcate environmental attitudes, whether these be of the exploitive or the conservative kind. Think of oil company advertisements on television or of documentaries on whales.

Only recently has the application of industrial energy to warfare been examined in an environmental context. At one level, there is the way in which calorie production must be stepped up, for soldiers need feeding well, as do the industrial workers who make their munitions. Horses, too, consume cereals and were important as late as 1945. So many grasslands are ploughed up to make way for wheat, oats, and rye crops, with downstream aftermaths for wildflower populations and soil losses. The opposite happens when relict landmines prevent land being used for agriculture: a World War II problem in Poland to this day. But the main impact of industrial warfare before 1950 was seen on the Western Front in World War I. This was in effect a huge concentration of energy use in the form of soldiers, horses, shells, and bullets. The environmental alterations were profound and almost total, as can be sensed from the paintings of John Nash, for example. Yet the land was quite quickly retransformed to agricultural use, and only a few consciously preserved remnants of trench systems are left, though the whole of them can be seen from the air at the end of summer like an archaeological crop mark. The less static warfare of 1939 through 1945 left fewer long-tern environmental changes in Europe but transformed the vegetation, soils, and indeed landforms of some Pacific islands.

If some summative expression of industrialization and environment is sought, then two phenomena stand out. The first is that enough carbon was emitted into the global atmosphere to start the rise in concentration of (especially) carbon dioxide, which underpins today's concerns about climatic change. The mid-18 c is usually taken as the base-level of 280 ppm, to compare with recent values of 360 ppm. The second is the explosion of the city, parasitic on the rest of the world for energy in all forms and for oxygen. The carbon loading has been mostly caused by the rich, but the city's ecology is shared by LIE and HIE alike.

3.4 Post-Industrial Economies (post–1950 CE)

After 1950, most of the indicators of environmental change start to rise in magnitude. Graphs such as those in B.L. Turner's 1990 compilation show sharp upward trends in resource use, waste production, emissions to air and water, and land use transformations. The increases are seen in Table II, where some shift toward an increased proportion being devoted simply to movement can be detected. Most of these are outworkings of increased use of oil, whose ascendancy over coal in this period is unchallenged. So post-industrial economies are in large measure the old fossil-fuel-powered systems writ large in many ways (not least in the geopolitics of oil) but in some of its energy-environment relationships there is something different about the post-1950 world.

For once it is possible to date the genesis of the new world quite accurately: 2.30 p.m. on December 2, 1942, in Chicago, where Enrico Fermi conducted the

TABLE II

Energy Use in Industrial and Postindustrial Societies

Society	Total energy use (10^3 kcal cap day)	Plant and animal (%)	Home and commerce (%)	Industry and agriculture (%)	Transportation (%)
Postindustrial	230	4.3	28.6	39.5	27.3
Industrial	77	9.0	41.5	31	18

first controlled fission of atomic nuclei and thus made possible both military and civilian nuclear power. (The energy content of pure ^{235}U is 5.8×10^5 MJ kg, compared with oil at 43 MJ kg.) The military applications have had some environmental effects, most notably in the fallout of radioactive material from atmospheric weapons testing before the ban in 1963. Levels of long-lived radionuclides (such as ^{137}Cs and ^{90}Sr) built up in high latitudes and have accumulated in the bodies of, for example, the Saami people. The potency of thermonuclear weapons was such that in the 1970s scientists constructed scenarios for a 5000 MT exchange that would produce at least 12 months of virtual darkness in the Northern Hemisphere ("nuclear winter"). This knowledge is sometimes credited with helping to bring about the end of the Cold War. The potency of radionuclides to affect human tissues on all timescales from the immediate to the 240,000-year half-life of ^{240}Pu has accompanied the development of civilian nuclear power where very strict international agreements control the discharges to air and water. In spite of these, nations like Norway and Ireland object vociferously to the marine releases of the UK plant at Sellafield on the coast of Cumbria. Of greater concern, however, is the possibility of accidents resulting in the meltdown of the core of a reactor, as has happened at Sellafield in 1957, at Three Mile Island (United States) in 1979 and at Chernobyl (Ukraine) in 1986. Only the latter was truly disastrous environmentally and to human health but the memory is very vivid all over Europe. Yet of even more extensive importance is the fate of high-level radioactive wastes. Nobody yet knows really what to do with them since they must be sequestered from the environment for about one-quarter million years, a commitment rather like that of the Netherlands in keeping out the sea. One further long-term consequence of this era has been the formulation of chemicals with long-term toxicities, which can accumulate in food chains. The organochlorine pesticides and dioxin are the best known examples with the former causing high avian mortality and the latter implicated in teratogenic fetuses after its distribution in Vietnam as a contaminant of herbicides used in the war of liberation in the 1960s and 1970s.

The adoption of nuclear power in HIEs has mostly been motivated by a fear that fossil fuels might run out, either physically, economically, or politically. The HIE lifestyle is now dependent on relatively cheap and totally constant energy supplies: to interrupt gasoline flow at the pumps for a couple of days is virtually to bring about insurrection, whereas a coal miners' strike for 3 months has virtually no effect. In a similar vein, it takes little imagination to fashion scenarios of a western world without constant electricity. Since capital formation is now a matter of colored numbers on screens, then all kinds of projects to alter the environment might come to a standstill, and the standing of the small screen as a major source of authority in western society (especially if it shows a man (*sic*) in a white coat) would be greatly diminished—no more documentaries about global warming for a while at least.

Unlike the fossil fuel era, it is difficult to designate any particular piece of technology with direct environmental effects as characteristic only of the post-1950 period. In more ways, there has been more indirect effects through capital accumulation in ways that have spun off into environmental projects like the very large dams of the late 20th century planned for China, India, and possibly Turkey. One of the diagnostic features of the post-industrial society, however, is its appellation as a knowledge economy, and there is no quibbling with the fact that the environment is a big absorber of space on hard disks the world over. Cheap and ubiquitous energy makes possible the gathering of environmental data, its storage, and even its transmission in windowless, air-conditioned classrooms.

The present is always impossible to assess in historically oriented accounts and it may well be that the post-1950 phases is not really a new era but simply an intensification of the one that started in 1750. It could equally be the start of a transition to something totally new, in which case the energy-environment linkages would be a central feature along with one of their outgrowths already in place: the creation of a truly global network of instant communication (a noosphere in fact, as envisaged by Vernadsky in 1945) such as there has never been.

4. OVERVIEW FROM TODAY'S PERSPECTIVES

In the perspective of the last 10 ky, access to energy can be seen as having produced tensions in a world where access is highly differential (Table III). In a positive view, many developments have been facilitated, without which today's humans would not even be able to aspire to the cultural richness and diversity that have evolved. By contrast, access to energy has allowed impacts that if foreseen by some omniscient world power imbued with the precautionary principle would likely have been prohibited.

TABLE III

Energy Consumption at the End of the 20th Century, GJ Cap Year

Region	Commercial energy	Biofuels
World	60	3.7
N & C America	243	2.0
Europe	134	1.1
Africa	12	7.2

Looked at positively, access to energy has been one consequence of agriculture, which has brought about the possibility of adequate nutrition even for human populations of more than 6 billion: if they do not get it, then the causes are political, not to do with energy capture. In the wake of the shifts to more intensive sources of food than offered by hunting and gathering, a quite wonderful diversity of human cultures came about. (It is interesting that the more up-market the tourism, the further back along this spectrum people want to go, albeit only on vacation.) There are still elites who think that imperialism was a positive outcome of energy made available by gunpowder, the breech-loading rifle, the steamship, and by their latter-day equivalents. Travel has been one further outreach into which energy can be channeled. For humans, there is tourism and space travel (think of the rocket as an updated blowpipe), for instruments gathering data there is the satellite and the remote weather station on a buoy in the Pacific. To sum it all up, there is the city and its urbanized fringes, where the CBD stacks the decision makers in high towers (always a visible emblem of power, i.e., of personal access to energy: recollect as well the personal jet and the LDC leaders' Mercedes) and the suburbs spread out the expenditure of energy and money.

But change inevitably brings loss. Most of these have been irreversible, like the diminution of biodiversity caused by habitat change as, for example, agricultural intensification works toward monoculture or the demand for wood products devastates forests. In the period 1700–1980, forests and woodlands decreased globally by about 19%, grassland and pasture by about 1%, but croplands increased by 466%. Industrialized warfare is better at destruction of all kinds, including life in all its forms, than a few armies raging seasonally across selected terrain. The chemical industry has produced many quality-enhancing pharmaceuticals but has also thoughtlessly allowed out persistent substances that have engendered problems for wild species, crops, and human health alike. One effect has also been to introduce uncertainty: models of climate change are in the end models and can be disregarded on those grounds by those wanting to emulate the moa rather than the ostrich. Though the multiplicity of environmental changes that are problematic for somebody or something is bewildering, the main two for humans can probably be summed up as meat and motors. The energy going into both produces all kinds of patterns that could be rearrayed to the greater good of many more inhabitants of the planet.

If, though, there is only one allowable view of the history of energy and environment, it must be that access to the energy flows and stores of the planet, especially after the 18th century, has allowed the growth of the human population—from about 720 million in 1750 to 6.1 billion of the early 21st century. Whether regarded historically, culturally, technologically, or politically, that view fills the lens.

SEE ALSO THE FOLLOWING ARTICLES

Climate Change and Energy, Overview • Development and Energy, Overview • Ecological Risk Assessment Applied to Energy Development • Energy Efficiency and Climate Change • Environmental Injustices of Energy Facilities • Industrial Agriculture, Energy Flows in • Sustainable Development: Basic Concepts and Application to Energy • Technology Innovation and Energy • Transitions in Energy Use

Further Reading

Bormann, F. H. (1976). An inseparable linkage: Conservation of natural ecosystems and the conservation of fossil energy. *BioScience* **26,** 754–760.

Cottrell, F. (1955). "Energy and Society." Greenwood Press, Westport, CN.

Debeir, J. C., Deléage, J-P., and Hémery, D. (1991). "In the Servitude of Power. Energy and Civilization through the Ages." Zed Books, London.

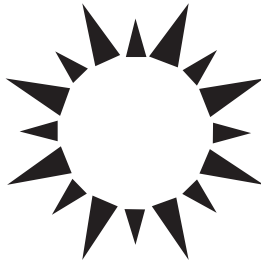
Hall, C. A. S., Cleveland, C. J., and Kaufmann, R. K. (1986). "Energy and Resource Quality. The Ecology of the Resource Process." Wiley-Interscience, New York and Chichester.

McNeill, J. R. (1991). "Something New Under the Sun: An Environmental History of the Twentieth-Century World." Allen Lane, New York and London.

Simmons, I. G. (1996). "Changing the Face of the Earth: Culture, Environment, History." 2nd ed. Blackwell, Cambridge, MA, and Oxford.

Smil, V. (1991). "General Energetics: Energy in the Biosphere and Civilization." Wiley-Interscience, New York and Chichester.
Smil, V. (1993). "Global Ecology: Environmental Change and Social Flexibility." Routledge, London, and New York.
Turner, B. L. (ed.). (1990). "The Earth as Transformed by Human Action." Cambridge University Press, Cambridge and New York.
Vernadsky, V. I. (1945). The biosphere and noosphere. *Am. Scientist* **33,** 1–12.

Environmental Gradients and Energy



CHARLES A. S. HALL
State University of New York
Syracuse, New York, United States

1. **Introduction**
2. **An Integrative Approach Based on Bioenergetics and Five Important Principles about Energy and Ecology**
3. **Methodology for Measuring Energy Balance within *N* Gradients**
4. **Energy Balance and Population Dynamics**
5. **Summary**

Glossary

abundance Number of a species found per area.
distribution Range of occurrence, as in brook trout are distributed in the Appalachian Mountains from Maine to Georgia.
gradient A gradual change in some condition from low to high. For example, there is a gradient in temperatures in the Luquillo Mountains of Puerto Rico from 22° at the top of the mountain to 30° at the bottom.
opportunity cost When an organism uses some of its surplus energy for one activity that energy is no longer available for some other activity.
population dynamics Changes in species' population abundance over time.
succession More or less orderly replacement of one species by another over time in an ecosystem following the relaxation of a disturbance.

The examination of an organism's energy costs and gains along environmental gradients, such as temperatures varying over a terrain, can help establish a central issue for all levels of ecology, which is what determines the distribution and abundance of species and populations.

1. INTRODUCTION

Population ecology is concerned with the distribution of individuals of a particular species over space and time, community ecology examines the distribution of species congregations over space and time, and ecosystem science considers materials transfer and other processes which in turn influences the collective distribution and abundance of organisms within defined environments over time. These issues are equally important to plant and animal ecologists, as well as to those who study behavior and other interactions.

1.1 Observed Patterns in the Distribution and Abundance of Biota

How are species distributed over space and time? There are several important generalities. Usually individuals of a species are more abundant near the middle of their geographic range and become progressively rarer toward the edges, a pattern that approaches a normal curve, or sometimes a skewed normal curve. This generality relates to the assumption that distribution and abundance of a species reflects the probability density distribution of the species' required environmental resources and biotic interactions. A corollary is that populations of species are often found in a greater variety of habitat types near the middle of their geographic range compared to the edges. A second major pattern is that in any given environment some species are relatively abundant, but most are relatively rare. Furthermore, species that are abundant in the center of their range tend to have widespread distributions, and the rarer species may be rare everywhere. Relative abundance also is influenced by differential survival and reproduction, although factors that

 © 2004 Elsevier Inc. All rights reserved.

cause those differences are often not readily apparent and are rarely quantified.

A third important generality is that species or population abundances usually do not follow repeatable patterns over long time periods in a given environment, even though the number of species may persist. While examples of relatively constant populations certainly exist, such as Spartina grass in many coastal environments and brown trout in small streams, most populations that have been studied over long time periods seem to fluctuate widely in abundance, even when plotted on a log scale. The dynamics of populations tend to be, or at least appear, stochastic and should not be described or modeled by a deterministic or logistic equation. In most environments, populations are almost continually reduced by random environmental disturbances of the many abiotic (e.g., temperature, moisture, light, nutrients) and biotic (e.g., predation, competition, parasitism, dispersal) factors that influence population dynamics.

1.2 The Need for an Integrative Approach

Many investigators have remarked on the general lack of an integrative approach to assessing the many potentially important factors effecting organismal distribution and abundance, as well as the need for more comprehensive approaches of understanding the impact of more than one resource or factor. In fact, many studies have shown that all factors examined were clearly important in determining the distribution and abundance of the populations in question. However, the difficulties and costs associated with measurement of multiple factors in ecological studies are great, especially over the long time periods required to demonstrate interannual variation, which may be essential to inferences. The lack of multifactor empiricism in ecology is also rooted in the lack of a general, unifying theory that (1) integrates the myriad of biotic and abiotic factors that influence species distributions and abundances and (2) provides a mechanism for explaining how organisms or species identify resources and evaluate (or trade off) the costs associated with either obtaining them or adapting to alternatives.

1.3 Multiple Factors and Tradeoffs

Maguire examined the growth response of a phytoplankter as a function of two (or, in theory, more) environmental gradients (e.g., variability in temperature and a nutrient). Maguire then examined how these resources changed over time as the season progressed and temperatures warmed while nutrients were reduced. By integrating the environmental changes in relation to concommittant changes in growth of that species over the year, or between years, he was able to show how abundances and distributions changed as environmental conditions waxed and waned. This seems substantially superior to assuming a constant rate of growth and a fixed carrying capacity, as is the case in the more generally used logistic models.

Many other studies have identified the tradeoffs associated with different resources or adaptations in relation to biotic distributions and abundances. Some very good examples pertain to plant life history strategies, where investments in chemical defense versus growth, fixation of carbon versus nitrogen, efficiency of water use versus carbon uptake, or growth rate versus structural integrity have been quantified. A particularly interesting example is one pertaining to nitrogen concentration of leaves. There is a remarkably linear relation between leaf nitrogen concentration and growth rates of plants. Aspen *(Populus tremuloides)* and other plants with high nitrogen concentrations have high growth rates, probably due to the presence of high levels of carbon fixation enzymes. However, leaves high in nitrogen are also more desirable food for many heterotrophs, from microbes to moose. But this means that high growth rates may also be associated with high loss rates through herbivory, and high rates of herbivory can induce plants to produce chemical defenses. Our point is that approaches emphasizing tradeoffs are important when trying to understand biotic distributions and population dynamics because essentially every resource can be obtained only at the expense of other resources. Smith and Huston proposed a general theory linking vegetation patterns to species-specific physiological and life history traits in relation to energetic constraints on the capture and use of resources. Their approach emphasized fundamental tradeoffs facing plants, such as CO_2 uptake versus water retention (at the level of stomates as well as the whole plant), root versus shoot allocations in relation to the costs of acquiring water or nutrients versus sunlight, and allocation of photosynthate to leaves versus the whole plant. Plants are very plastic with respect to these tradeoffs, both among and within species, by virtue of adaptations.

Any holistic scheme that attempts to provide a comprehensive approach must satisfy at least three

criteria. First, it must analyze, and give procedures for assessing, the tradeoffs that exist among more than two or three factors. Second, explicit explanations must be given as to why these tradeoffs should exist or of the mechanisms by which they operate relative to the observed distribution and abundance of a population or species. Third, they must start from first principles, an approach that tends to work well in other disciplines. Energy is the common currency of ecology and is necessary for any biophysical process to occur. Thus, energy should be fundamental in the development of an ecological theory that attempts to explain the distribution and abundance of animals and plants in relation to the myriad of ecological factors that impinge on all populations.

2. AN INTEGRATIVE APPROACH BASED ON BIOENERGETICS AND FIVE IMPORTANT PRINCIPLES ABOUT ENERGY AND ECOLOGY

The essence of the approach developed here is to integrate energy costs and gains along all appropriate environmental gradients to yield net energy gain in the form of biomass elaboration and reproduction. In this way, the observed distribution and abundance relationship may be explained in terms of the net energy balance or budget of the organism, population, or species. Our central tenet is that a sufficient, positive energy balance must be sustained through time and space such that a significant number of individuals reproduce effectively. If this does not occur, then eventually the population and ultimately the species will decline toward extinction or, through the process of natural selection, adapt some other energy strategy that will permit survival.

It is necessary first to identify five important principles about energy and ecology. First, as identified in many review articles, energy contributes to survival and reproduction of organisms in accordance with our contemporary understanding of basic evolutionary. We believe that the most appropriate way to express this relationship is energy per time, or power, as developed in many publications by Howard Odum and also others. In the words of Ware, "Optimal life history is concerned with power allocation." Energy is vital to organisms in that, other things being equal, the more energy an organism can sequester per unit time, the larger its energy reserves will be to help meet stress and contingencies. Therefore, if an organism can maximize energy intake, it is more likely to reach age of reproduction rapidly and maximize the number of propagules it can propel into the future to perpetuate its genome. Orians and Solbrig summarize this view: "a plant capable of increasing photosynthesis within the constraints of its available resources and the physical environment should gain advantages in competition with other plants, defenses against herbivores, and should have more energy to devote to reproduction."

Second, an examination of an organism's ongoing energy balance allows a less circular definition of fitness than is possible through the more commonly used criteria of reproduction itself. Although at various times in an organism's life its ability to accumulate energy can be satiated, in nature this is very rare. This is especially true with respect to high-quality food items, whose availability is usually limited.

Third, energy balance is important from the perspective of how much energy an organism must divert from growth and reproduction to overcoming contingencies such as adverse environmental conditions. In general, organisms reproduce only from the energy surplus after meeting other contingencies in their life, including growth to a reproductive size. However, when conditions become so extreme that the organism itself has a high probability of death, all available resources often are put into reproduction.

Fourth, the importance of energy as it effects organisms encompasses far more than just the environmentally available or stored (i.e., trophic or elaborated) energy that is usually emphasized. It includes, among other things, metabolic energy drains imposed on organisms by nonoptimal environmental conditions or added stresses, thermal energy gained or lost by behavioral modifications and biotic interactions as discussed later. Most important, however, it includes the energy costs of respirational requirements as well as energy required for an organism to live in and to exploit particular environments, especially suboptimal environments. For example, a rainbow trout is isotonic to water of about one-quarter the salt content of seawater. In fresh or saltwater, that fish must use at least 30% more energy for basal metabolism simply for osmotic adjustments. This is probably one reason that estuaries are such important fish nurseries. Fish can grow better in environments where their basal metabolism is lower. Fifth, in order to grow and ultimately reproduce, organisms require resources and must invest stored energy reserves to get them.

Moreover, all resources vary in quantity and quality. The most important point about energy and its relation to other resources is not that everything can be reduced to energy (which is false) but rather that every material (and most nonmaterial) resource has an associated acquisition energy cost. Most resources become limiting in part because their acquisition may require too great an energy investment on the part of the organism. Thus, high-quality resources are often those that require less energy expenditure per unit of resource obtained. However, with sufficient energy reserves and as constrained by its genome, an organism can divert energy to the acquisition or mobilization of essentially whatever other resources are in short supply, even when their quality is low. Likewise, quality also reflects the value of the energy per kcal of heat content so that, for example, a kcal of food has more utility to an animal than a kcal of incident solar energy and food with a high nitrogen content may be much more useful. We define high-quality resources as those of highest utility that are available to the organism with relatively low energy investment.

Bioenergetics can provide a comprehensive and explicitly testable set of mechanisms for predicting where species will be found and for determining why they are not found in other locations, although it cannot account for conveyance of propugules to new sites. While it is true that organisms need resources other than energy—for example, various nutrients—these other resources can be examined from the perspective of their acquisition energy cost.

2.1 Net Energy Balance: The Elusive Integrator

Each factor that impinges on an organism has energy costs and, in many cases, potential energy gains. The life history strategy (phenology) of an organism is the outward expression of the genetic codification of past successful trials of an organism's energy return on investment. At each point in an organism's life it must invest energy for obtaining more energy while decreasing energy drains and converting tradeoffs about which contingency is most deserving of whatever energy reserves might be available. Tradeoffs occur because most energy gains are obtained only by increasing energy costs—that is, by increasing energy investments. Energy investments by plants to grow leaves versus roots is one obvious example. Another is the energy investment by animals in motility for the purpose of optimizing energy intake from food resources.

Life is a continuous balance of energy investments, each with potential gains but also real and potential costs. As one simple example, plants have energetic tradeoffs whereby the advantage gained by early reproduction tends to be at the cost of future reproduction. For another example, longer-lived but slower growing trees often put large energy investments into energy-intensive structural fibers that generally allow them to survive storms and other contingencies, but at the cost of reduced growth and reproduction. Similarly large, long-lived trees that are easy for insects to find tend to be more heavily defended with secondary compounds that operate to discourage insect feeding. Generally, plants that are exposed to high levels of infestations increase their investments in secondary compounds in subsequent years. In summary, this energy balance perspective derives from the fundamental energetics ideas of earlier investigators as detailed at the end of this entry and involves examination and integration of all pertinent energy costs and gains throughout the life history of the organism. Populations (or species) may be examined in the same context, if one accepts that the net energy balance of the population results from the cumulative energy balance of individuals within the population (or species). This gives a less circular and more mechanistic set of tools to assess tradeoffs facing an organism or population, and ultimately the relative success of its genome, than by using fitness (potential reproductive output) as a direct function of multiple environmental gradients. Ideally both approaches would give similar answers to the same question.

2.2 Energy Costs and Gains in Relation to Environmental Gradients

As noted, population densities tend to be higher in the middle of the species' geographic range, and we can extend that idea to say it is because physical, chemical, and biotic resource gradients are optimal there, which allow maximal energy profits. For plants, light, water, and nutrients are clearly important in this context, and for animals, food derived from photosynthate and (generally) oxygen are essential. For both, temperature, often considered the master regulator, is critical. The concept and data of gradient analysis, as developed initially by Ramensky, Gleason, and Whittaker, demonstrate this fundamental ecological pattern. In terms of our energy balance approach, individuals of a species will be distributed over space and time in a way that may

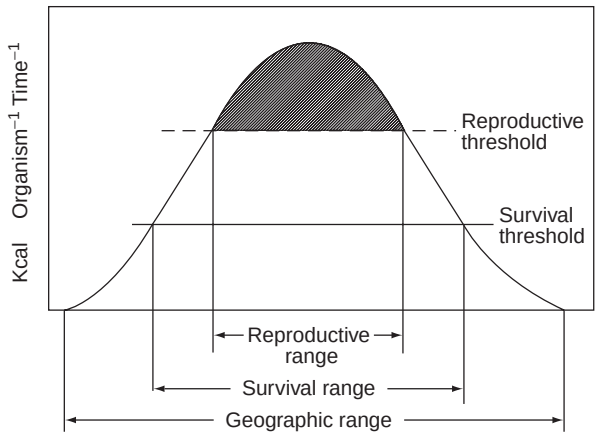


FIGURE 1 Net energy balance of a species over its geographic range. The horizontal solid line represents zero net energy gain for the species (survival threshold)—that is, conditions where energy gain is just sufficient to compensate for basal metabolic energy, plus the energy required to capture food and overcome stress. Below that line, organisms can exist only by consuming their own energy resources. The broken line indicates the larger amount of energy intake needed to generate sufficient energy surplus for reproduction (reproductive threshold). The hatched area of the curve is the only part of the range in which the species can exist in the long term. From Hall, C. A. S., Stanford, J., and Hauer, R. (1992). The distribution and abundance of organisms as a consequence of energy balances along multiple environmental gradients. *Oikos* **65**, 377–390. Used with permission.

be idealized as a more or less normal curve representing energy return on investment along a gradient (Fig. 1). Two important points derive from Fig. 1. First, individuals near the center of the collective distribution enjoy a high energy profit, which allows a large net gain of energy, growth, and either numerous offspring per effort or frequent reproduction. Individuals on the margins of their physiological, territorial, or nutritional range with respect to the gradient are unable to generate a sufficient energy profit to produce population growth. Competition may also be more important at the margins of the range because it raises the cost of resource acquisition and creates more of an impact on individual energy gain in locations on the gradient where the net gain is already small. Hence, if the gradient is favorable and temporally stable, the species will be abundant and widespread, which is consistent with the first generality of observed patterns of distribution discussed earlier. The second point is that rare species are those whose particular morphologies and behaviors do not allow them to be abundant anywhere because they never make a sufficient energy profit. This is consistent with the second generality of observed patterns of distribution discussed previously.

This approach emphasizes that all abiotic variables that may be important for bioproduction of any species may be examined in the context of the energy costs and gains to that species. This is regardless of whether resource limitation is examined in the context of availability of the resource or the ability of the species to obtain the resource. For example, at very low nutrient concentrations, such as the 1 ug/l of bioavailable phosphorus that commonly occurs in oligotrophic lakes (where phytoplankton production is often phosphorus limited), a 10 um plankton cell is surrounded by roughly 100 atoms of phosphorus in the 10 ul of water adjacent to the cell. Even though it may seem that phosphorus atoms are relatively abundant, bioavailability of phosphorus atoms is more influenced by the large concentration gradient that exists between the inside and the outside of the cell membrane. Uptake of phosphorus by the cell to mediate growth requires energy expenditure (active transport). Therefore, nutrient uptake at low concentrations is influenced critically by light. A low concentration of nutrients acts as an energy drain to the phytoplankter, which in turn reduces the reproductive output of the plant and contributes to their low concentrations found where nutrients are low.

Finally, various biotic interactions may also be considered as gradients in our energy balance context. For example, natural selection operates to optimize the risks of predation versus the benefits of exposure to predators in such a way that net energy gains are optimized. Thus, it is possible to view predation energetically from a "Lloyd's of London view." In the old days of sailing ships, Lloyd's knew that a certain proportion of the ships they insured would founder in any given year; but the company management also knew that if no ships left port, no profits would be forthcoming to the shipping industry (or to Lloyd's). On the other hand, captains who loaded too much sail trying to maximize profits usually foundered (or otherwise lost more than they gained) and were selected against. As an analogous example, Gilliam and Fraser found that the risk taken by minnows could be manipulated by changing the energetic return of a foray into predator-infested waters, and that the behavior of minnows in accepting risk was predictable from an energetic model of energy gains and costs. Likewise, it costs energy to avoid predation, including both the direct costs of avoiding food rich areas that contain predators or of manufacturing defenses; there are also the indirect costs of decreased motive efficiency. Using similar reasoning, competition (both inter- and

intraspecific) usually reduces the availability of a resource for one or both of the competitors, and hence may reduce the range or move the optimum location of an organism along the resource gradient. Hence, competition for resources may reduce the net energy profits, unless the organism is a very good competitor, in which case net gains may indeed be enhanced.

2.3 Some Examples of Energy Tradeoffs in Response to Resource Gradients

An important aspect of any analysis of energy balance in relation to resource gradients is that there are opportunity costs to being a particular species or genome. This simply refers to the fact that energy investment in life history strategies that produce prosperity at one place along an environmental gradient (opportunity costs) precludes adaptation to other conditions. Thus, inherent in the concept of phenological energy balance in response to an environmental gradient is the concept of tradeoffs. At any given time, the energy available to adjust to contingencies or to reproduce effectively usually is limited and must be invested wisely.

Often the relations defining costs and gains (tradeoffs) along a gradient are nonlinear. Such nonlinearities, at least in principle, can produce many of the classic hump-shaped growth curves along a gradient found in the literature. In the case of the growth of a particular aquatic insect or fish as a function of temperature, foraging rates increase, often linearly at least up to a point. But respiration increases more rapidly as temperature rises, because of its Q10-type response (Fig. 2A). More complicated relations can exist, especially at the extremes of the gradient, but even so the maximum net gain is normally near the middle of the gradient, even for homeotherms including humans. The net energy return to an organism is the difference between energy gains and energy costs. Thus, the observed growth curve versus temperature relation can be explained as the difference between energy gains and costs along a temperature gradient (Fig. 2). A clear tradeoff exists between temperature and growth for many organisms. Vannote and Sweeney elegantly showed how this idea applied to distribution and fecundity of different species of mayflies.

Bunce and his colleagues measured photosynthesis and respiration of eight plant species that had different distributions over the elevational gradient of moisture on a mountain in Arizona and calculated

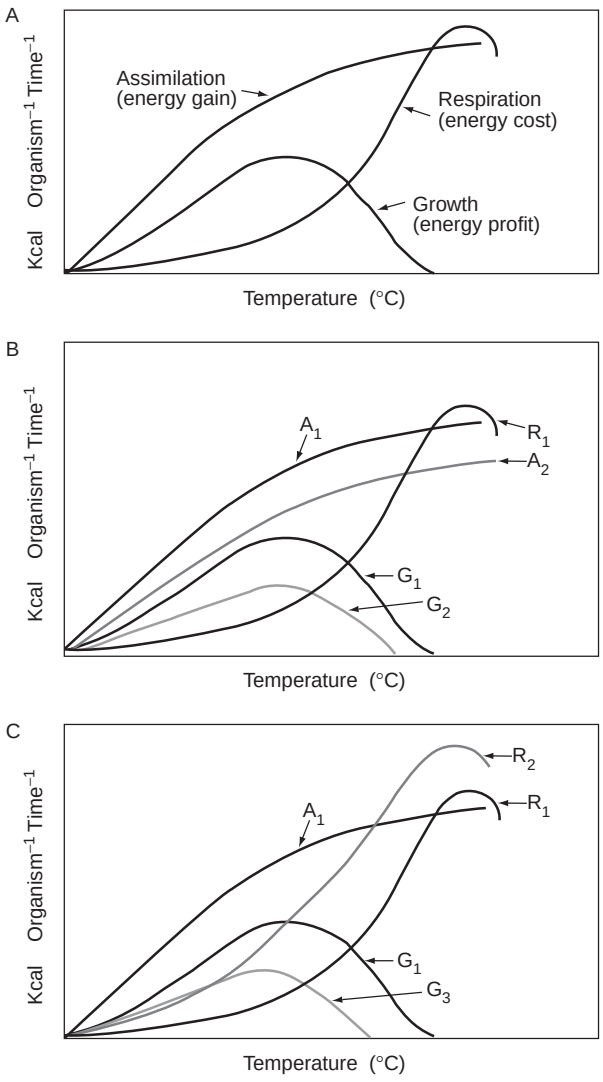


FIGURE 2 (A) Growth of a poikilotherm in response to an environmental gradient (in this case temperature). We hypothesize that net energy profit, which equals growth plus reproduction, is the result of changing energy costs and gains along the gradient. The commonly observed bell-shaped growth curve is the net difference between metabolic energy expenditures (respiration) and gains (assimilation), realized as individual growth (energy profit). (B) Factors that reduce energy assimilation (curve A_1 changes to curve A_2) will result in a decrease in individual growth rate (G_1 to G_2). Decreased quantity or quality of food assimilated may be the result of abiotic factors (e.g., nutrient availability, disturbance) or biotic factors (e.g., competition or predation pressure). (C) Factors that result in increased respiration (R_1 increases to R_2) will also decrease individual growth rate (G_1 to G_2). Increased respiration may be the result of changing abiotic conditions (e.g., caused by disturbance) or manifestation of biotic interactions (e.g., increased search time for food due to competition, increased activity caused by predation pressure). From Hall, C. A. S., Stanford, J., and Hauer, R. (1992). *Oikos* **65**, 377–390. Used with permission.

the carbon balance of each species as a function of moisture and hence elevation. They found that the lower limit of each of the species was approximately 200 m above the zero carbon balance elevation predicted from their field measurements. In other words, each species would not be able to make a significant carbon, and hence energy, profit very much beyond the measured limits of their range. The actual mechanisms for drought tolerance or avoidance were complex, but the net effect was well predicted by their model. Thus, tradeoffs along the moisture gradient were apparent between species. Some had mechanisms, such as a relatively large area of open stomates, favoring high carbon gain caused by high rates of photosynthesis, but they were also prone to high water losses. Other species were more conserving of water but had lower carbon gains.

Trout in a river are principally drift feeders. If they are able to maintain position in rapidly flowing water, the current brings more food to them, but the energy costs of swimming rapidly to maintain position may be too great to maintain a positive energy balance in the long term. Bachman developed a general theory of the energy costs and gains of trout feeding in waters of different velocities. Dominant or fastest-growing fish were found approximately where the difference between the costs and gains were greatest. Thus, dominant trout grew faster where the current was optimal relative to food delivery (energy gains), swimming speed, and duration (energy costs). Young steelhead trout were also distributed along the velocity gradient of a stream approximately as predicted from this model. Brook trout change the area they defend, and their antagonistic displays, as a function of changes in water velocity in a way that maximized the energy efficiency with which they defended their territories. The same kind of result held for pelagic fish and their swimming speeds. Young bleak (a British minnow) voluntarily swam at velocities ($107\,\mathrm{mm\,s^{-1}}$), similar to velocities that theoretically optimized their net energy gains ($111\,\mathrm{m\,s^{-1}}$). In a more general sense, organisms that defend larger territories get more food for their young but must expend more energy in that defense, resulting in a bell-shaped net return curve.

2.4 Molecular and Cellular Tradeoffs

While it is not difficult to find organisms responding to gradients, there is not a well-developed or recognized procedure for *why* organisms should do well at certain parts of gradients and not well at others, or for why this should affect energy gains and costs. There are probably many reasons, depending on the gradient considered, but an important point is that many enzymes controlling growth, maturation, and reproduction tend to have great specificity. Moreover, this specificity is greatly influenced by the environment, especially temperature. Enzymes exhibit thermal optima that correspond closely to body temperatures in the case of homeotherms or to the preferred habitat temperatures in poikilotherms. Organisms tend to seek temperatures where their particular enzyme variants are most effective; therefore, the distributions of closely related species may be explained by variable enzymatic responses within an environmental gradient. At temperatures below the thermal optima, enzymes function slowly. This is due partially to low levels of kinetic energy and partially to the constraints on flexibility because of the specificity of temperature-mediated activities, such as folding, at the active site. Above the thermal optima, the weak chemical bonds that are responsible for formation of transition state structures, and ultimately catalysis, are weakened and break. Various physiological mechanisms have evolved to broaden the functional range of temperatures and lower Q10 values. Clearly, however, there are limits to the functional range of a given protein molecule. In addition, the number of isozymes with different thermal optima available to an organism necessarily is limited to the number of alleles for that enzyme, a maximum of two in the case of diploids (assuming no genetic duplication). There are exceptions. Rainbow trout (*Oncorhynchus mykiss*), a fish with a relatively broad thermal tolerance among salmonids, are tetraploid. As a consequence they have twice as many alleles to generate enzymes as other trouts. Presumably this allows various races of the rainbow to be adapted more specifically to warm or cool water.

It is probably efficient for organisms to have enzymes with a relatively narrow temperature range for metabolic activity, because enzymes that operate well at a given temperature appear to operate less well at other temperatures. This has been demonstrated empirically for different subpopulations of the killifish *(Fundulus heteroclitus)*. Each subpopulation of the fish had an optimal operating temperature where metabolic efficiency was maximized. Moreover, an associated optimal temperature for maximum kinetic performance of lactate dehydrogenase B was observed, which related to swimming efficiency and hatching success. Such tradeoffs along gradients are also important for the structural

components of cells. Membrane lipids, for example, are subject to functional alteration at different temperatures.

2.5 Extension to *N* Gradients

Not all the gradients affect an organism's welfare equally, because some, such as temperature and food quantity and quality, have a large effect on metabolism whereas others, such trace metals, often have minimal effect. But the cumulative effects will be reflected in the organism's integrated net energy balance in relation to each relevant gradient. For example, a tree species may grow progressively well or poorly from the bottom of a mountain to the top, due to differences in temperature. Likewise, it may do well or poorly as a function of sunlight or moisture gradients from one side of the mountain to the other. Shelford introduced and Hutchinson developed the idea that the realized niche of an organism could be determined empirically by plotting the boundaries of that organism's existence as a function of the *N* factors that influenced the distribution of the organism. The resulting *N*-dimensional hyperspace is the definition of the range (and sometimes niche) of that organism. A species is the aggregate genetic code, and its phenotype is the organism that can generate a sufficient energy profit to reproduce at some place along each possible environmental gradient. Since it apparently does not pay to exist along the entirety of each environmental gradient, natural selection of each species is focused on the part of each gradient that allows a sufficient profit for the species to survive and reproduce. This may be a strong factor favoring speciation. Widespread species are those that have evolved mechanisms for making significant energy profits within a large portion of the hyperspace created by all gradients that impinge on the phenology of the species.

Hutchinson's concept of *n*-dimensional hyperspace can be extended such that all relevant resource gradients, and the energy costs and gains for an organism to live along each of those gradients are integrated graphically and mathematically (Fig. 3). The well-being, growth, and eventual reproductive success of the organism is defined by the interpretation of energy costs and gains on all gradients that influence the organism's phenology. For purposes of illustration, these energy profit gradients can be envisioned as a Chinese egg set of *N* dimensions, with the exterior egg representing the limits of the organism's range. Where individuals

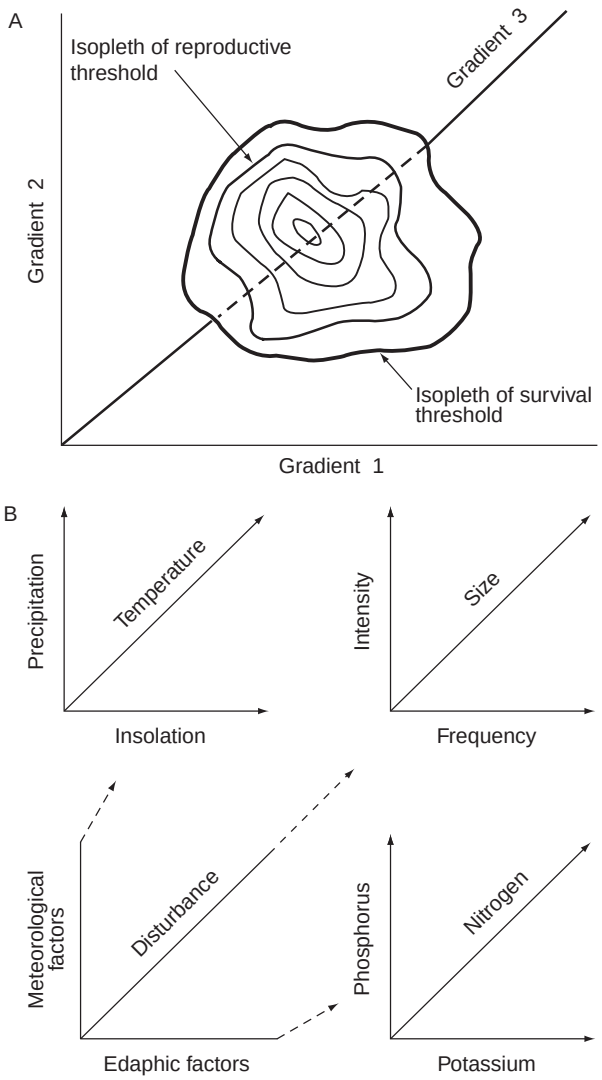


FIGURE 3 (A) Isopleths of net energy profits, defined as energy gains minus costs, of a species as a function of three environmental gradients. Net energy profits and, hence, species distribution and abundance respond to the influence of each environmental gradient involved in the energy balance of the species. Smaller energy profits therefore are shown as isopleths occurring successively outward from the point of optimum profit in relation to these three gradients. Although drawn hypothetically here in just three dimensions, in reality energy profits are influenced by many (*N*) environmental gradients and should be visualized in the context of an *N*-dimensional hyperspace (see text). (B) One way to display the effects of more than three gradients simultaneously. In this case, the figure is showing nine gradients (three each: edaphic, disturbance, and meteorological) thought to influence the distribution and growth of plants and animals in the Luquillo Mountains of Puerto Rico. Edaphic gradients include soil nutrient (e.g., N, P, K) concentrations, disturbance gradients include intensity, size, and duration of hurricanes, landslides, and tree-falls, and meteorological gradients include the relation between precipitation, temperature, and solar insolation at different altitudes in the mountains. From Hall, C. A. S., Stanford, J., and Hauer, R. (1992). *Oikos* **65**, 377–390. Used with permission.

of a species range beyond where they can survive for long periods, the next interior egg represents the survival range, and so on. Reproduction takes place within some interior egg, which represents an even greater optimization within the hyperspace. Maximum reproduction takes place at a center of gravity representing simultaneously the midrange of all impinging gradients. Thus, many gradients can be represented and analyzed as in Fig. 3.

2.6 Niches, Succession, and Competition

In the context of this energy balance perspective, a Gaussian niche is the volume of all possible energy profits for a species within the hyperspace of all gradients to which the organism might be exposed. A Hutchinsonian niche is the volume of all realized energy profits for a species and is at least partly dependent on the geographical history of the species. Succession, then, may be caused in large part by the changing energy profit regimes of the environment in response to the presence or absence of various species and the changes produced on that site by various species over time. It is not clear whether species generally affect other species by any other mechanism than by changing the quality and hence energy cost of resources. For example, trees influence understory plants by shading or lowering available nutrients. The conditions favoring a positive energy balance for particular species change over successional time.

Competition often has been invoked as the principal mechanism that causes the distribution patterns of organisms, but that view seems to be increasingly called into question. The data of Whittaker, for example, show plant species distributed along elevational gradients in more or less random patterns, rather than in avoidance patterns that would be more consistent with competition. This indicates that the species are not causing each other to move "away" from one another on the gradient. Although few ecologists would argue with the existence of competition in nature, there is a very large difference between establishing its presence and attributing to this one factor the fundamental role in structuring communities. Where competition is demonstrated to be operational, it can be factored easily into the energy balance analysis promoted herein. As noted earlier, exploitive or interference competition can alter the availability of a resource and impact the energy balance of the competitors. For example, light-intolerant shrubs are aided by the shading of other plants (exploitation) or one species may produce a substance that is toxic or stressful to another (interference). Such influences can be treated as additional gradients (axes) that impinge on the organism's energy balance. If enough axes are examined, the concept of competition may not be as relevant in explaining the displacement of different species along resource gradients. Each species simply does not have the morphological, behavioral, or, especially, physiological (biochemical) wherewithal to provide an energy balance that allows it to exist along the full length of all important gradients.

3. METHODOLOGY FOR MEASURING ENERGY BALANCE WITHIN N GRADIENTS

Species-specific responses of individuals at locations within many environmental gradients result in a complex pattern of energy costs and gains. This seemingly intractable pattern can be analyzed and predicted by using simulation modeling, if the basic input data for each gradient exist. First, an appropriate time loop is generated. Then, the environmental gradients are established mathematically according to historical or simulated patterns. Geographical specificity can be included; for example, procedures have been developed for projecting meteorological information from a single station to a montane landscape. Next, the individuals are located mathematically along each relevant environmental gradient for each time unit (month or day or hour) as a function of the species' known preferences or behavior in response to each gradient. Data on the energy costs and gains are needed for each gradient as per Fig. 2. Then, all costs and gains are summed across gradients for the time period of interest or as the various state variables (gradients) are changed.

4. ENERGY BALANCE AND POPULATION DYNAMICS

A final point is that the multifactorial energy balance approach given here should be useful in attempting to understand population dynamics. Very simply, the net energy balance of a species changes as environmental conditions change. Temporal variations are inherent in resource gradients, and phonological energy balances will likewise vary. Therefore,

populations will be dynamic in relation to net energy profits invested in offspring and the energy balance of those offspring as environmental conditions become more or less favorable over space and time (Fig. 4). This replaces the dogma that survival is the only determinant of future population levels and leads to mechanistic, explicitly testable hypotheses that can be applied to any individual or population. One good example of where the energy balance approach might be useful in producing a more realistic explanation of year-to-year variation in abundance of a species is in regard to recruitment variability and the dynamics of commercial marine fisheries. For years, fisheries managers have focused harvest goals on density-dependent (logistic) responses in considering controls on stock-recruitment dynamics. However, in the multispecies, environmentally complex world of most exploited marine fisheries, the implications of the analysis in Fig. 4 seems to be much more appropriate based on the rationale of this paper.

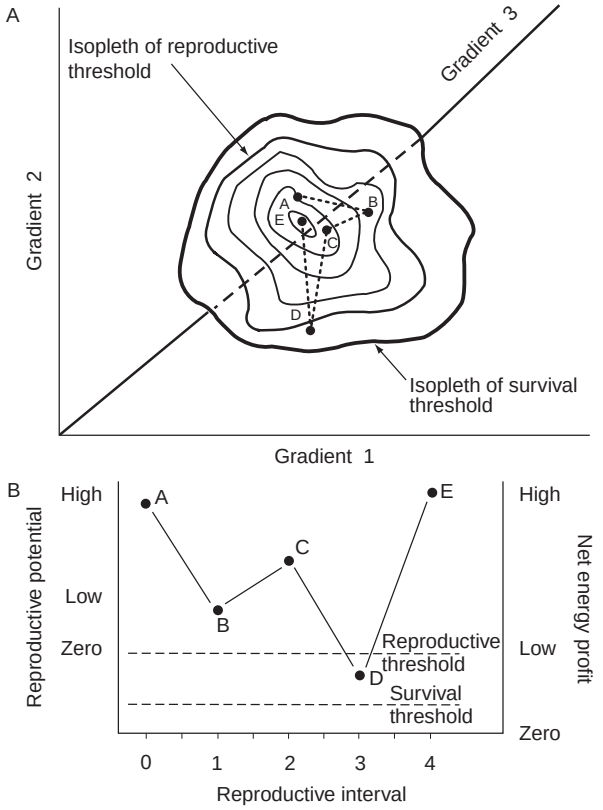


FIGURE 4 (A) Net energy balance (isopleths) of a population existing at A–E different positions within the *N*-(in this case three dimensional) hyperspace determined by important ecological gradients that influence growth throughout the life history. (B) Temporal reproductive potential of the population represented in Fig. 4A. In this example, the cumulative energy gains of individuals exceed losses and the population survives, although investments in reproduction are variable over time. For genomes that die after a single reproductive effort and do not have multiple cohorts (e.g., many insects), population extinction would occur at D. For organisms that live many years (e.g., most commercially important fishes), such year-to-year fluctuations in reproduction and similar changes in death rates lead to the population dynamics that we observe. From Hall, C. A. S., Stanford, J., and Hauer, R. (1992). *Oikos* **65**, 377–390. Used with permission.

5. SUMMARY

This article began with a discussion of generally observed patterns of biotic distributions. It concludes that these patterns can be explained as a consequence of energy balance along multiple environmental gradients due to species-specific energy costs and gains (energy balance) in response to the many (Hutchinsonian *N*-dimensional) environmental or resource gradients. Competition, predation, and other biotic interactions operate principally by increasing energy costs to the species and can be included in our energy balance methodology as additional environmental gradients of energy costs. Persistence of a population and, ultimately, the species in a given locality will occur only where energy return on investment allows a significant energy profit in the form of propugules. At any given time, energy investment in physiological, behavioral, or morphological adaptations that optimize abundance at one point within the *N*-dimensional hyperspace of environmental gradients precludes the investment of the same energy elsewhere. Within the spatial and temporal range where the species can reproduce, abundance is related to the stochastic nature of the important environmental forcing functions because energy balance is ultimately controlled by the sum of organismal responses to all environmental influences. Populations tend to be most abundant where their genome allows the optimization of energy profits. Therefore, population dynamics reflect the cumulative effect of environmentally induced shifts in organismal energy balance.

Holistic energy perspectives in the past have been complicated and controversial and often based on very theoretical considerations. But when an energy perspective is integrated with the vast bulk of appropriate physiological and ecological analysis, as done here, the energy approach can generate an elegant and powerful syntheses. The goal here is to

show that energy analysis can be used to integrate formerly disparate components of ecology using an approach that is straightforward enough to require little in the way of going beyond most traditional views in biology. If the energy balance theory presented here is robust, it has the potential of synthesizing ecology at many levels.

Acknowledgments

This article was derived from Hall, C. A. S., Stanford, J., and Hauer, R. (1992). The distribution and abundance of organisms as a consequence of energy balances along multiple environmental gradients. *Oikos* **65,** 377–390.

SEE ALSO THE FOLLOWING ARTICLES

Ecological Footprints and Energy • Ecological Risk Assessment Applied to Energy Development • Ecosystem Health: Energy Indicators • Ecosystems and Energy: History and Overview • Environmental Change and Energy

Further Reading

Bunce, J. A., Chabot, B. F., and Miller, L. (1979). Role of annual leaf carbon balance in the distribution of plant species along an elevational gradient. *Bot. Gaz.* **140,** 288–294.

Fry, F. E. J. (1971). The effect of environmental factors on the physiology of fish. *In* "Fish Physiology" (W. S. Hoar and D. J. Randall, Ed.), Vol. VI, pp. 1–87.Academic Press, New York.

Gilliam, J. F., and Fraser, D. F. (1987). Habitat selection under predation hazard: test of a model with foraging minnows. *Ecology* **68,** 1856–1862.

Hochachka, P. W., and Somero, G. N. (1984). "Biochemical adaptation." Princeton University Press, Princeton, NJ.

Hutchinson, G. E. (1965). The Niche: An abstractly inhabited hypervolume. *In* "The Ecological Theory and the Evolutionary Play," pp. 26–78. Yale University Press, New Haven, CT.

Krebs, C. J. (1985). "Ecology: The Experimental Analysis of Distribution and Abundance." 3rd ed. Harper & Row, New York.

Odum, H. T., and Pinkerton, R. C. (1955). Times speed regulator, the optimum efficiency for maximum output in physical and biological systems. *Am. Sci.* **43,** 331–343.

Shelford, V. E. (1951). Fluctuations of forest animal populations in east central Illinois. *Ecol. Monogr.* **21,** 183–213.

Vannote, R. L., and Sweeney, B. W. (1980). Geographic analysis of thermal equilibria: A conceptual model for evaluating the effect of natural and modified thermal regimes on aquatic insect communities. *Am. Nat.* **115,** 667–695.

Ware, D. M. (1982). Power and evolutionary fitness in teleosts. *Can. J. Fish. Aquat. Sci.* **39,** 3–13.

Environmental Injustices of Energy Facilities

NANCY IRWIN MAXWELL
Boston University School of Public Health
Boston, Massachusetts, United States

1. Coal and Coke Production
2. Petroleum Production
3. Coal- and Oil-Fired Electric Power Plants
4. The Nuclear Fuel Cycle
5. Large Hydroelectric Dams
6. Home Use of Biomass Fuels

Glossary

biomass fuels Any plant or animal matter used as fuel, including animal dung, crop residues, wood, charcoal, or peat.

respirable particulates Airborne particles with diameter less than 10 microns.

vectorborne disease A disease that affects more than one host species and that is transmitted by one host (the vector) to other host species. Many vectors are insects or arthropods; mosquitoes, for example, are the vector that transmits malaria.

working level (WL) One WL is defined as any combination of radon progeny in one liter of air that ultimately releases 1.3×10^5 MeV of potential alpha energy.

working level month (WLM) One WLM is equal to 170 WLs (that is, exposure to an average of 1 WL for a working month of 170 hours).

The energy sector relies on a broad range of facilities to extract and distribute resources, produce energy, and handle wastes. All these facilities benefit those who use the energy produced. But many facilities also create environmental pollution and have negative impacts on human health—impacts that may fall on populations who do not benefit from the energy produced. In some cases the environmental exposures or health impacts of energy facilities, like those of other industrial facilities, fall disproportionately on certain geographic regions or nations, due to the concentration of resources or facilities in certain locations. In other situations, disproportionate burdens may fall on particular subpopulations, including disenfranchised groups such as low-socioeconomic status communities, communities or workers of color, less developed nations, women or men (male workers, for example), children or other age groups, future generations, or susceptible populations. This article describes such environmental injustices, or inequities, associated with energy facilities. It begins with energy production activities related to fossil fuels (coal and oil) and electric power generated from these fuels, and then the article considers in turn the nuclear fuel cycle, large hydroelectric dams, and home use of biomass fuels. International statistics on energy production and consumption are taken from the World Energy Council.

1. COAL AND COKE PRODUCTION

Nearly all the world's coal production—and its environmental and health consequences—is concentrated in Europe, North America, and Asia. Asia accounted for 38% of 1999 global production (China 24%, India 7%). The United States and Europe each produced 23% of the world's coal. The global distribution of coal consumption mirrors that of production rather closely.

1.1 Environmental Impacts of Underground and Surface Mining of Coal

Coal mining traditionally took place underground, with men and equipment following seams of coal deep within mountains. Since the 1960s, surface mining methods have become more common, and today they account for more than half of the coal

 © 2004 Elsevier Inc. All rights reserved.

produced in the United States. In this approach, also known as strip mining, oversized bulldozers remove the land surface until coal is exposed, producing large amounts of waste rock—and dust—in the process.

Historically, destruction of forests often accompanied underground coal mining in the United States. The usual practice in the 1800s was to cut down all trees with commercial value on the overlying land before beginning to mine. Much of the timber was used for props in the coal mine shafts and for railroad ties as track was laid to carry away the coal.

Both underground mining and surface mining of coal cause acid mine drainage. This problem occurs when the mineral iron disulfide is exposed to water and air, leading to the formation of sulfuric acid. Large volumes of acidic liquid drain into local streams, killing fish as well as aquatic plants.

An extreme form of strip mining now being practiced in the Appalachian region of the United States is known as mountaintop removal. As the name suggests, several hundred feet are sheared off the tops of mountains that are close to one another, exposing the coal, which is removed. The waste rock is dumped into valleys so that a set of neighboring mountains in effect become a single plateau. The resulting valley fill buries streams and may become unstable as it becomes saturated with water; its impacts on local groundwater transport are unknown.

1.2 Community Burdens of Coal Mining

Coal production in the United States is highly concentrated in the Appalachian region. The leading coal-producing states, West Virginia and Kentucky, together accounted for 43% of the value of U.S. coal produced in 1997, as reported by the U.S. Economic Census. Pennsylvania and Virginia together accounted for an additional 17%, bringing the total for the Appalachian region to 61% of the value of U.S. coal production. Wyoming, Colorado, and Alabama together account for another 16%. The leading coal states, West Virginia and Kentucky, are also two of the poorest states in the United States, ranking 50th and 44th, respectively, on median household income in 1999, and 4th and 7th on the proportion of their population who lived in poverty (18% and 16%).

In Appalachian communities where mountaintop removal is under way, residents endure dynamite blasts, falling rock, and dust. Property values decline, and many residents move away. Studies of communities near ordinary surface mining in the United Kingdom indicate that particulate concentrations in air are somewhat elevated in these areas, and children in mining communities make somewhat more visits to their doctors for respiratory symptoms such as wheezing and coughing.

On a global scale, community impacts of coal mining in less developed nations are likely to be more pronounced than in the United States and the United Kingdom, which can afford more stringent regulatory controls. For example, the study of surface mining of coal in the United Kingdom reported mean concentrations of respirable particulates (particulates less than 10 microns in diameter, known as PM_{10}) ranging from 14.4 to 22.3 $\mu g/m^3$ (micrograms particulates per cubic meter of air) in nearby communities. A study of surface mining of coal in one of India's major coal fields reported concentrations of respirable particulates of roughly 70 $\mu g/m^3$ in nearby residential areas, as well as concentrations of total suspended particulates (i.e., including larger particulates) of 200 to 500 $\mu g/m^3$ in these areas.

1.3 Impacts to Workers

Communities also bear the burden of disease that occurs among their workers. Coal miners, traditionally men, are at high risk of physical injury from machinery and from strenuous, sometimes repetitive, physical labor. Many coal miners suffer hearing loss from the noise of machinery. Coal workers' pneumoconiosis, also known as black lung disease, is a chronic lung disease caused by exposure to coal dust. As the lungs become progressively scarred with continuing exposure, they become stiffer and lung function is impaired. There is no cure, and the disease is sometimes fatal.

Some coal is converted to coke for use as fuel in blast furnaces in the steel industry. Coke itself is made in a coke oven, where gaseous components of the coal are driven off using extreme heat, leaving the coke as a residue. Coke oven workers, especially those working in undesirable jobs at the top of the oven where fumes are strongest, are exposed to high levels of chemicals known as coal tar pitch volatiles.

A classic 1975 occupational health study by Mazumdar and colleagues, which focused on lung cancer risk to coke oven workers, gives a window onto historical racial inequities in this work setting. The study estimated the men's cumulative exposures to coal tar pitch volatiles based on their tenure in various job categories. As shown in Fig. 1, the proportion of nonwhite workers increases steadily from lower to higher categories of cumulative exposure. This pattern reflects the historical practice

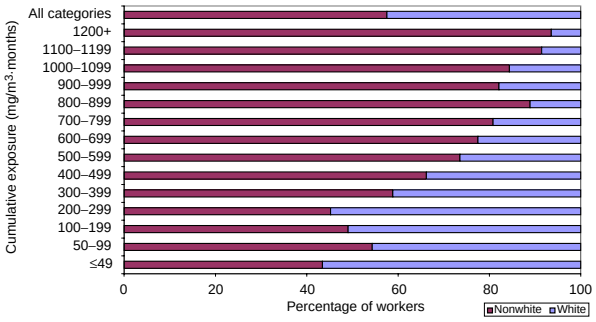


FIGURE 1 White and nonwhite coke oven workers, by exposure category. From Mazumdar *et al.* (1975).

of differentially assigning African-American workers to the high-risk "topside" jobs, while white workers were more likely to be assigned safer positions away from the top of the coke oven.

2. PETROLEUM PRODUCTION

The regions that consume the largest share of the world's petroleum production are not those that produce the largest shares. North America is the largest consumer (30% of global consumption) and is a major producing region (19%), but not the foremost. Similarly, proportional consumption outstrips production in Asia and Europe. The Middle East, by contrast, produced 30% of the world's oil and natural gas in 1999 but consumed a mere 9%.

The environmental and health impacts of oil production can be substantial; the compensating benefits that accrue to producing regions are not only access to the energy produced, but also financial gain from oil sales and political power gained through controlling resources and production.

Where foreign interests control oil production, the host country may bear the environmental health impacts while a multinational corporation realizes the financial gain. Such a situation now exists in Nigeria, which produced about 30% of Africa's oil in 1999. British and U.S. oil companies, in joint ventures with the Nigerian state oil company, have done most of the oil exploration and extraction in the country. Regional populations within Nigeria have bitterly protested the exploitation of local oil resources, with profits departing from the region and environmental contamination left behind.

2.1 Oil Refineries

Most chemical releases from oil refineries are to air. In the U.S. Environmental Protection Agency's (EPA's) Toxic Release Inventory data for 2000, ammonia and sulfuric acid were released in the largest amounts; also prominent were volatile aromatic hydrocarbons that are found in both crude oil and petroleum products (e.g., benzene, toluene, xylene), as well as *n*-hexane and the gasoline additive methyl *tert*-butyl ether.

Like crude oil resources, oil refineries are geographically concentrated, so their environmental and health impacts fall on local populations. Within the United States, the 1997 U.S. Economic Census identifies 244 petroleum refineries nationwide (excluding small establishments). Three states—Texas, California, and Louisiana—host more than 40% of those refineries, representing more than half the total value of U.S. petroleum shipments (Table I).

Petroleum facilities may also be highly concentrated on a more local scale. In Louisiana, for example, data from the U.S. EPA's Toxics Release Inventory show that the 29 reporting petroleum facilities are located in just 19 of the state's 525 five-digit zip codes—mostly near Baton Rouge, in Calcasieu and Caddo parishes on the Texas border, and in parishes near New Orleans. Of all the chemical releases from these facilities, half are from a single ExxonMobil refinery in East Baton Rouge, whose releases are almost five times greater than those of the second-ranking refinery. As shown in Table II, this immense refinery is located in a zip code whose population is 89% nonwhite—about three times higher than the state average. The third-ranked refinery in Louisiana, located in Caddo parish, is in a zip code whose population is 90% nonwhite. The stretch of the Mississippi River from Baton Rouge to New Orleans, known locally as "cancer alley," is home not only to several refineries but also to a much larger number of petrochemical factories. Activists have long protested the concentration of these petrochemical industries in poor, nonwhite communities.

2.2 Oil Storage Facilities

The U.S. EPA's Toxics Release Inventory reported releases from 474 petroleum bulk terminals in 2000. Not surprisingly, some of the largest of these transport hubs are located at ports. For example, five facilities, representing 7% of releases nationwide, are coastal facilities in Greater Boston, and five facilities, representing 4% of releases nationwide, are located in Los Angeles County. The chemicals released in the largest amounts by weight are some of the same volatile organic chemicals released by oil refineries, though the total releases from bulk terminals amount to only about 5% of those from refineries.

TABLE I

Geographic Concentration of Oil Refineries in the United States

Location	Number of refineries	Percentage of U.S. refineries	Value of shipments ($1000)	Percentage of value of shipments
United States	244	—	157,934,872	—
Texas	49	20	44,412,926	28
California	31	13	19,296,030	12
Louisiana	25	10	25,795,152	16
Top three states	105	43	89,504,108	57

Source: U.S. Economic Census (1997).

TABLE II

Minority and Poverty Status of Louisiana Zip Codes Hosting Three Refineries with Largest Environmental Releases

Location		Total on- and off-site reported releases (lbs)	Percentage nonwhite	Percentage in Poverty
Louisiana (statewide)		10,420,308	32	20
Louisiana (median of 525 zip codes)		—	27	20
Zip Code	**Parish**			
70805	East Baton Rouge	5,169,845	89	34
70669	Calcascieu	1,084,032	16	15
71109	Caddo	835,951	90	33

Source: U.S. EPA, Toxics Release Inventory (2000); U.S. Census (2000).

Petroleum bulk terminals, as large industrial installations, are often located in lower income neighborhoods. Four of the five Boston-area "tank farms," for example, are located in some of the most poverty-stricken communities in the eastern part of the state. In addition to routine chemical releases and the potential for spills or attacks, communities that host petroleum bulk facilities endure the exhaust, noise, and accident hazards of the many trucks that roll through their neighborhoods each day.

3. COAL- AND OIL-FIRED ELECTRIC POWER PLANTS

Inequitable environmental impacts of burning fossi fuels in electric power plants include local air pollution burdens on disadvantaged populations, regional and international effects of acid deposition, and climate change on a worldwide scale.

3.1 Disparate Impacts of Local Air Pollution

Coal- and oil-fired electric power plants contribute to local air pollution, especially particulates, sulfur dioxide, and nitrogen oxides. In the United States, industrial sources of air pollution tend to be located in lower-income urban areas, and so local impacts often fall most heavily on poor and minority populations. The prevalence of asthma, which may be exacerbated by exposure to air pollutants, has been rising in the United States in recent decades and is particularly elevated among inner-city children.

3.2 Transboundary Acid Deposition

Through a series of complex chemical reactions in the atmosphere, sulfur dioxide and nitrogen oxides emitted from coal- and oil-fired power plants may be converted to sulfuric acid and nitric acid, respectively. These chemicals may be transported hundreds of miles by wind and ultimately be dissolved in precipitation—rain, snow, or fog—and deposited far from their point of origin. Such long-distance transport was brought about by the installation of tall smokestacks in the 1970s; intended to alleviate local pollution, the tall stacks placed air pollution within reach of higher level air currents.

As a result, acid deposition as an environmental impact is very often inequitable, since the receiving area is typically far from the source of the pollution.

For example, acid deposition in the northeastern United States and parts of eastern Canada is due largely to sources in the Midwestern United States. Similarly, burning of fossil fuels in Germany and the United Kingdom contributes to acid deposition in Scandinavia, and coal-fired plants in China were the original source of acid deposition in Japan. Acid deposition can acidify drinking water sources, potentially mobilizing toxic metals such as lead from the sediments of drinking water reservoirs or from drinking water distribution pipes. Acid deposition also harms aquatic life and damages buildings and statues.

3.3 Inequities Related to Global Climate Change

The very designation of "developed" and "less developed" nations identifies an enormous disparity in material security and public health status between these groups of countries. It is inherently unfair that much of the economic growth in the developed nations occurred before the impacts of carbon fuels on global climate were appreciated, whereas the less developed nations must take account of these impacts as they seek economic growth. The dilemma of defining an equitable approach to controlling emissions will not easily be resolved.

Some of the anticipated impacts of global climate change are expected to fall more heavily on less developed countries simply as a matter of geography. For example, warming climates will increase the ranges of mosquitoes and other vectors of infectious disease, and this impact will be felt most strongly in parts of the globe where these diseases, such as malaria, are already endemic. Similarly, much larger numbers of people in the developing world live in low-lying areas that will be threatened by rising sea levels. Small island nations from the Pacific to the Caribbean are experiencing increased temperatures and are uniquely vulnerable to a rising sea level.

The Intergovernmental Panel on Climate Change (IPCC) in its report *Climate Change 2001: Impacts, Adaptation, and Vulnerability* projects that many areas in the developing world are more vulnerable to the effects of global climate change, whereas impacts will generally be more moderate, or more mixed, in North America and Europe. In the more arid regions of Africa, for example, global climate change is expected to exacerbate water scarcity and increase desertification. Food insecurity, already a problem in much of the continent, is expected to worsen due to water stress in agricultural areas and impacts of climate change on fisheries. Increases in temperature and rainfall will increase the range of vector-borne diseases and will also make it more difficult to maintain sanitation, with a resulting increase in waterborne illness. Increased food insecurity and infectious disease are also projected for areas in Asia. In parts of Asia that are already vulnerable to typhoons or cyclones, droughts, and floods, such extreme weather events are projected to increase. IPCC projects that with a 100 cm rise in sea level, 5763 km^2 in India would be submerged, putting a population of 7.1 million at risk. Under the same scenario, Bangladesh could lose 29,846 km^2 of land area, putting 14.8 million people at risk (13.5% of the country's population). Even a sea-level rise of 45 cm is projected to displace 5.5 million in Bangladesh.

The IPCC projects that global climate change will bring more difficult conditions for the southern and Arctic areas of Europe than for other areas of the continent. Pressures on water supply will increase. Crop yields in northern Europe are projected to increase somewhat, though yields in southern Europe are likely to be somewhat reduced. The major direct risks to human health are anticipated from episodes of extreme heat, especially in urban areas, from some increase in vector-borne disease, and from flooding. For North America, a mix of positive and negative impacts is projected. More extreme weather events are anticipated, such as episodes of extreme heat or cold, storm surges, and floods. Overall agricultural production is not threatened under moderate climate change scenarios, though some regions will gain and others will lose in agricultural productivity, and greater warming could cause more disruption of farming. Some increase in vector-borne diseases is also projected. The strong social, economic, and public health infrastructures in Europe and North America are expected to mitigate effects somewhat.

In addition to placing an inequitable burden on developing nations that are less responsible for the problem and less able to manage its impacts, climate change will differentially affect future generations in all regions of the world. More than any other environmental health problem, climate change is global in the scale of its impacts and uncertain in the scope and nature of its intergenerational effects.

4. THE NUCLEAR FUEL CYCLE

Like other energy resources, uranium reserves are unevenly distributed around the globe. The United

States and Canada have produced the most uranium over the long term. U.S. production has fallen off dramatically since the 1980s, but Canada remains the leading producer of recent years. Historically South Africa and Germany have also been major producers of uranium; today Australia, Niger, Namibia, and several countries of the former Soviet Union are producing increasing shares of global uranium output. In contrast, about half of the world's nuclear generating capacity is currently concentrated in Europe and about one-third in North America (mostly in the United States); Japan is the largest user of nuclear power in Asia.

4.1 Environmental Health Hazards of the Nuclear Fuel Cycle

The nuclear fuel cycle, with its unusual suite of environmental health hazards, raises issues of environmental inequity distinct from those of fossil fuels. Nuclear power plants do not produce air pollution in the normal sense of the term, nor do they contribute to global climate change. The hazards associated with nuclear power are those of ionizing radiation, and include the potential for catastrophic reactor failure as well as the unique challenges of disposing of wastes that remain radioactive for thousands or millions of years. Until the emergence of global climate change as a threat to human health, nuclear power was unique in shifting to future generations much of the burden of health risk associated with today's energy production.

Short-term exposure to high levels of ionizing radiation, as in a nuclear accident, may be rapidly fatal or may cause radiation sickness—a syndrome that includes effects on the blood-forming system, the gastrointestinal system, and the central nervous system. Exposure to lower levels of ionizing radiation over a period of time leads to an increased risk of leukemia and of other cancers including cancers of the thyroid, bone, lung, and breast. Birth defects including microcephaly, growth retardation, and mental retardation have also been associated with low-level exposure to ionizing radiation *in utero*.

4.2 Distribution of Nuclear Energy's Benefits and Risks

The distribution of benefits and public health risks associated with nuclear power is inherently inequitable, simply in geographic terms. The "front end" of the nuclear fuel cycle (uranium mining and milling, and fabrication of reactor fuel) offers no benefit to local populations, other than employment, and workers in uranium mines and mills are exposed to elevated concentrations of radon gas. When radon gas is inhaled, exposure to its radioactive breakdown products, known as radon progeny or radon daughters, increases lung cancer risk. People living in nearby communities are also exposed to radioactive dust from the freestanding piles of waste rock (called tailings) left near mines and mills. In some communities, tailings have been used as construction materials for roads or even homes, resulting in higher radiation exposures to residents.

The hazards associated with mining and milling are usually located far from the people who benefit from the electric power produced by a nuclear reactor. Those populations, often in densely populated areas, do, however, bear the uncertain risks associated with routine low-level radioactive emissions and would be hardest hit in the event of a severe accident—though the incident at Chernobyl demonstrated how geographically widespread such impacts can be. At the "back end" of the nuclear fuel cycle, specially constructed waste repositories are ideally located in isolated areas, which would typically be far from the beneficiaries of the power plants that produced the wastes. This strategy reduces the size of the population at risk, but shifts the burden of risk to people who got no benefit from nuclear power. In the absence of such a repository, spent fuel may be held for long periods at the power plants, as currently in the United States, or disposed of in a less secure manner.

In the United States, most uranium mines and mills are in the West—predominantly in Arizona, Colorado, New Mexico, Utah, and Wyoming. The 103 nuclear power plants operating in the United States in 2003, by contrast, are mostly located east of Kansas and Nebraska. Indeed, if a map of uranium mill tailings piles were overlaid on a map of nuclear power plants, the tailings piles would drop neatly into an area not served by nuclear power. Yucca Mountain, in the southernmost corner of Nevada, is the long-planned site for permanent disposal of U.S. nuclear power plant wastes. The wastes are to reach this remote location by rail lines that crisscross densely populated areas that benefit from nuclear power, regions with no nuclear activities at all, and areas already heavily affected by uranium mining and milling.

4.3 Profile: Differential Impacts on Native Americans of the Nuclear Fuel Cycle in the United States

4.3.1 Early Uranium Mining in the United States

The Four Corners region—where Colorado, New Mexico, Arizona, and Utah meet at right angles—is home to rich reserves of carnotite, an ore containing oxides of both uranium and vanadium. The Navajo Nation is here, located mostly in Arizona but extending into New Mexico and Utah. Carnotite was first mined on a small scale in the early 1900s for the orange color it gave to ceramic pottery glazes. During World War I, radium was in demand for painting instrument dials, and between the world wars vanadium—used to harden steel—was much sought after. But uranium mining began in earnest in the early 1940s with the start of the Manhattan Project and the loss of Czechoslovakian uranium reserves to the Nazi invasion of that country. It became a true boom industry in the late 1940s when the U.S. Atomic Energy Commission (AEC) began to ratchet up development of atomic bombs. For national security reasons, the AEC was at this time the sole buyer of the area's uranium, though mining operations were run by private companies. The AEC would soon begin to press the development of nuclear power as well, further increasing the demand for uranium.

When the mining peaked in 1955–1956, there were hundreds of active uranium mines on the Navajo reservation, as well as several mills where the ore was crushed and the uranium extracted. Jobs were scarce on the reservation, and by the time the boom ended, some 2500 to 3000 Navajos had worked in the mines, either on the reservation or in nearby areas. Smaller numbers of Hispanic men and non-Hispanic white men also worked in the mines. These early mines were known for their poor working conditions. Until the early 1960s, there was no ventilation and no protective gear. Large-scale uranium mining in the region ceased in the late 1960s.

Miners' exposure to radon progeny, which emit radiation as alpha particles, is measured in units called working levels (WL). The WL is a measure of the concentration of potential alpha particles in air. (One WL is defined as any combination of radon progeny in one liter of air that ultimately releases 1.3×10^5 MeV of potential alpha energy.) A working level month (WLM) is equal to 170 WLs (that is, exposure to an average of 1 WL for a working month of 170 hours). The federal exposure limit for miners was set in 1971 at 4 WLMs per year.

4.3.2 Evolving Understanding of the Health Risks of Uranium Mining

As early as 1879, a lung disease called *Bergkrankheit* ("mountain illness") was described in uranium miners in the Erz Mountains in Europe. By the time of World War I, lung cancer had been identified as an important cause of death in these miners. Uranium ore had been known for some time to be radioactive, and radon gas had also been identified. However, scientific confusion over whether and how radon might cause lung cancer continued until the role of radon progeny was elucidated in the early 1950s. Still, the link between occupation and illness was clear, and both Czechoslovakia and Germany had begun to compensate lung cancer in uranium miners as an occupational disease by the early 1930s.

The U.S. Public Health Service (PHS) undertook the first study of radon exposure and lung disease in U.S. uranium miners, many of whom were Navajo, through medical surveys during the 1950s. Both the health agency (the PHS) and the regulatory body (the AEC) have been strongly criticized for failing to obtain informed consent from the miners before using them as research subjects and for recording miners' exposures and health outcomes without providing ventilation and protective gear to reduce exposures already known to be harmful.

More than four decades of published research now documents the experience of the early U.S. uranium miners as this cohort has aged, accumulating illnesses and deaths. Exposure to radon progeny has been found to be associated with increased risk of both malignant and nonmalignant lung disease. Further, because the ore in this region is found in formations of quartz sandstone, workers are exposed to silica dust and are at risk for silicosis. Smoking enhances the lung cancer risk of radon exposure—that is, the lung cancer risk of smoking and radon exposure combined is greater than the sum of their separate risks.

In 1990, Congress passed the Radiation Exposure Compensation Act (RECA). Miners who worked during the years that the U.S. government was the sole purchaser of uranium (1942–1971) may qualify for one-time compensation in the amount of $100,000 for primary lung cancer and certain nonmalignant respiratory diseases including pulmonary fibrosis, pneumoconiosis, and silicosis. To qualify, a miner (or his survivors) must provide documentation of a designated medical condition and must also be able to document a certain total radon exposure.

4.3.3 Differential Impacts of Uranium Mining on Navajo Miners

Because most U.S. uranium mines were located on or near the Navajo Nation, the Navajos were vastly overrepresented among the miners, compared to their representation in the U.S. population as a whole. Navajos were more likely to work underground than other miners, partly because they were less likely to hold supervisory positions. It appears that in some small Navajo-run mines, working conditions were even worse than in the larger corporate mines. Some Navajos did not speak English, inhibiting their access to any health information that might have been available to the miners. Thus the burden of the government's ongoing failure to regulate working conditions in mines, and of its conduct of research violating accepted ethical standards, fell heavily on the Navajo.

Although early publications on the U.S. uranium mining cohort included some data on mortality among Navajos, a full analysis of the lung cancer risk among Navajo uranium miners was not published until 1984. This landmark study, led by Jonathan Samet of the New Mexico Tumor Registry, identified all cases of lung cancer among Navajo men in New Mexico that had been identified by the tumor registry and compared the men's work histories to those of a control group of Navajo men not diagnosed with lung cancer. Because every Navajo with lung cancer had a history of uranium mining, it was not possible to calculate the relative risk for this exposure (with a zero denominator, the relative risk is infinity), but the lower bound risk was statistically estimated to be a 14-fold increase. That is, the study found that a history of uranium mining carried at least a 14-fold increased risk of lung cancer, and probably a much higher risk.

A second milestone in documenting the differential impacts of the uranium era on Navajo miners came in a 1997 publication by Douglas Mapel and others at the University of New Mexico comparing the *non*malignant lung disease experience of Native American, Hispanic, and non-Hispanic white uranium miners. The study clearly documented some expected contextual patterns. Nearly two-thirds of the miners (63.4%) were Native American, 24.7% were Hispanic, and 11.9% were non-Hispanic white. A higher proportion of Native American miners had worked only in *under*ground uranium mining (84.3% versus 59.4% Hispanic and 62.4% non-Hispanic white), and a lower proportion of Native American miners had worked only in *above*ground uranium mining (3.7% versus 7.5% Hispanic and 9.3% non-Hispanic white). The study group reflected the well-known differences in smoking patterns—a much larger percentage of Navajos in the study had never smoked (72.9% versus 34.6% Hispanic and 22.2% non-Hispanic white). Further, those Navajos who did smoke had smoked a much smaller number of cigarettes (Hispanic and non-Hispanic white miners had accumulated more than five times as many "pack-years" of cigarettes smoked). Statistical analysis revealed that this difference in smoking behavior had an important consequence: although all three groups of miners experienced obstructive lung disease, among Hispanic and non-Hispanic white miners this condition was largely attributable to smoking, whereas among the Navajo (most of whom had never smoked), obstructive disease was mostly attributable to uranium mining. Other research has documented a similar effect for lung cancer, identifying a situation—perhaps unique in the annals of public health—in which most of the lung cancer in a population is attributable to a single occupational exposure.

To receive compensation under RECA for nonmalignant lung conditions, such as obstructive pulmonary disease or restrictive pulmonary disease, miners must meet certain diagnostic criteria—for example, lung function must be reduced by a certain percentage below normal on a standard lung function test. RECA specified the standards to be used, and they were based on the general U.S. population of non-Hispanic whites. Mapel and colleagues showed that 24% of Navajo miners who would be classified as having restrictive lung disease on the basis of a Native American-specific formula were excluded using the RECA standard. That is, some Navajo miners have been denied compensation under RECA although ethnic-specific criteria indicated that they were sufficiently impaired to warrant compensation. Amendments to the RECA in 2000 adjusted some eligibility criteria.

4.3.4 Differential Impacts of Uranium Mining on Navajo Communities

The risks of large-scale uranium mining were also brought to the Navajo at home, because much of the uranium ore so highly valued by the broader U.S. society was located within the boundaries of the Navajo Nation. The U.S. government, which held reservation lands in trust for Native Americans, freely exploited those resources, with profound impacts to the natural environment and risks to the health of local populations. Especially in the early decades, uranium was mined disproportionately on

native lands. The sense of injustice felt by Native Americans from this experience is reflected in the use of terms such as "internal colonialism," to describe U.S. government exploitation of uranium resources on Native American lands, and "national sacrifice area," to describe the environmental devastation that resulted.

Because groundwater seeped into both underground shaft mines and the deeper open-pit mines, the water had to be continuously pumped out so that operations could continue. This "dewatering" released large volumes of water. A single site might dump tens of thousands of gallons of water per day into local streams—water contaminated with radionuclides and heavy metals. Uranium mills, many located on or near reservation land, left behind enormous piles of tailings, an estimated 140 million tons altogether in seven western states. These heaps of pulverized waste rock are an ongoing source of radioactive dust. In 1979, an earthen dam collapsed, releasing more than 1000 tons of tailings sludge and almost 100 million gallons of water that were carried downstream into a Navajo community. Even the introduction of vents in mines sometimes resulted in community exposures, if vents were placed in populated areas above mines. All these burdens fell on communities that already lived with more poverty, poorer health, and a lower life expectancy than the general U.S. population.

4.3.5 Potential Siting of Nuclear Waste Facilities on Native American Lands

Since the 1980s, the federal government has planned to build a permanent repository for high-level radioactive wastes from nuclear power plants at Yucca Mountain, Nevada, but the expected opening year for the facility has receded steadily to 2010. As the permanent repository has been delayed, pressure has built for an interim storage facility (known as monitored retrievable storage, or MRS), so that spent fuel can be moved from storage sites at power plants. During the 1990s, the U.S. Department of Energy actively sought communities and Native American tribes to host such an MRS facility. The department worked to create an interactive process through which safety provisions and financial or social benefits would be negotiated such that the community would be satisfied that the potential negative impacts of the site had been offset. Though dozens of communities and tribes began the process, extended negotiations were held only with Native American constituencies—the Mescalero Apache of New Mexico and the Skull Valley Band of the Goshute in Utah. Both negotiations ultimately foundered. The Skull Valley Band has since entered negotiations with a private utility to host an MRS, resulting in legal battles that are not yet resolved. Although these Native American groups have participated voluntarily in siting negotiations, the process has nevertheless raised the question of an environmental injustice against disadvantaged populations whose economic need made them vulnerable.

5. LARGE HYDROELECTRIC DAMS

Although hydropower is a renewable energy technology and is often seen as relatively benign in terms of pollution, its facilities may have very different impacts on different populations.

5.1 Hydropower and Environmental Equity

Hydropower is a well-developed technology that is relatively accessible to developing nations. At present, five nations—the United States, Canada, China, Brazil, and the Russian Federation in that order—are global leaders in hydropower capacity, together representing almost 40% of the world's operating capacity. Dam projects now under construction worldwide, including the massive Three Gorges Dam in China, are expected to increase current global operating capacity by 16%. Taking account of capacity currently under construction or in the planning stages, China and Brazil are expected in coming years to push the United States and Canada to the third and fourth ranks, and India is expected to move up into sixth place behind the Russian Federation.

Thus, at a global scale, hydropower is becoming more equitably distributed between richer and poorer nations. On the smallest scale, too, when the benefits and burdens of small dams are shared by the same people, hydropower's impacts may be equitable. But to a great extent the very large dam projects that emerged in the latter 20th century and continue into the 21st deliver benefits and burdens to different populations, creating persistent inequities.

Hydropower is a relatively inexpensive form of electric power after the initial investment in construction, so the principal benefit of a large dam is cheap electricity for some population. Large dams may offer other benefits related to management of a waterway, such as flood control, improved navigation, or increased opportunities for irrigation. On the

other hand, new reservoirs may displace large numbers of people, often disenfranchised populations, and there may be detrimental downstream effects on fishing or farming. Finally, those located downstream of a dam are at risk in the event of a catastrophic failure, though this is generally considered an unlikely event.

Although this discussion focuses on the outward environmental impacts of hydropower facilities on different demographic groups, a historical episode from the occupational setting deserves mention. In 1930–1931, the 3-mile Hawk's Nest Tunnel was bored through a mountain in West Virginia to divert water from its natural riverbed to a hydropower plant being built for an industrial facility. The rock was high in silica, and dust was thick in the tunnel, but no measurements were taken, and attempts to ventilate the tunnel or damp down the dust were minimal. Although chronic silicosis was a recognized occupational illness in miners at the time, it was the fast-paced construction of the Hawk's Nest Tunnel that revealed *acute* silicosis, an illness that killed many workers within months of their tunnel labors. The workers, many of whom were brought in from outside the county for the tunnel project, were about two-thirds African American. Of the white laborers, 44% worked inside the tunnel at any time, and 22% worked *only* inside the tunnel (Fig. 2). By contrast, among the African-American laborers, fully 70% worked at some point inside the tunnel, and 35% worked *only* inside the tunnel. Thus the episode is noteworthy not only for revealing the acute form of silicosis but also for the striking racial inequities in workforce exposures to the silica-laden dust of the tunnel.

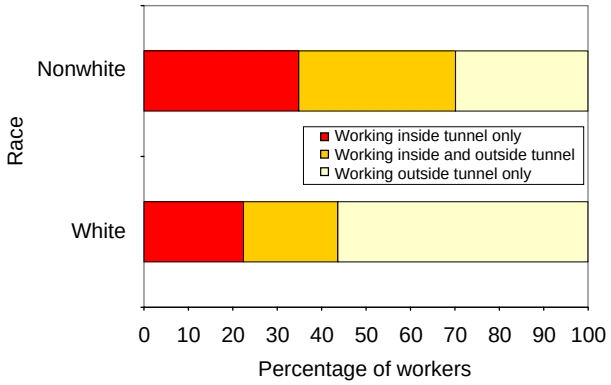


FIGURE 2 Exposure of white and nonwhite workers in building the Hawk's Nest Tunnel. Data from Cherniak, M., *et al.* (1986). "The Hawk's Nest Incident." Yale University Press, New Haven, Connecticut.

5.2 Profile: The Three Gorges Dam on the Yangtze River in China

The Three Gorges Dam now being built in Hubei province in China will be the largest dam ever constructed. The high dam is to create a narrow reservoir 600 kilometers (370 miles) long filling the Qutang, Wu, and Xiling Gorges of the Yangtze River. Construction began in 1994; by 1997 a diversion dam was completed so that work could begin on the main dam. Flooding and power generation began in 2003, and the project is to be completed in 2010.

5.2.1 Benefits

The dam is expected to produce 18,000 MW of electric power, and some 15 million people in five provinces in Central China will get inexpensive electricity from the project. By replacing coal-burning power plants, the dam is also expected to result in cleaner air in the Yangtze Valley.

In addition to its energy benefits, the dam is intended to control flooding, a major problem on the Yangtze River—although many tributaries enter the river below the dam site. It is also expected to smooth navigation of the river above the dam, though critics argue that locks needed to carry ships around the dam may stall river traffic.

5.2.2 Displacement of Residents

The Three Gorges Dam will flood nearly 20 cities, hundreds of towns, and more than a thousand villages. Estimates of the number of people to be displaced by the dam range from 1.3 million to 1.9 million. To date, more than 600,000 have been moved from their homes. Some 30,000 hectares (74,000 acres) of farmland will be flooded, and perhaps half of the people who will ultimately be displaced are farmers.

Because so many people must be moved within such a short time, the competition for resources is intense. Many of the displaced face hunger or food insecurity, occurring in the context of a disrupted social fabric. Although the government has compensated those displaced by the dam project, the compensation has been inadequate and many of the displaced become migrants. Resentment against the government is widespread, as is resentment against the displaced populations in the areas that receive them.

Because the Three Gorges Dam is such a high-profile project, institutional pressure to continue with it is intense, despite any problems or challenges that may occur. In this context, the large-scale removal and resettlement of people creates troubling

potential for human rights abuses, including punishment of those who refuse to leave—or who leave but later return to their homes.

A survey of displaced persons by Li Heming and colleagues gave insight into the experience of the dispossessed. The resettling of farmers offers special challenges because of their intense attachment to the land. Many of those who have refused to leave, or who have returned, are farmers. For the most part, the resettlement program has sought to move rural residents to other rural locations not too far distant from their original homes. But farmers have been given less land than they owned in the area to be flooded, and land of lower quality—for example, with poor soil on a steep slope—or suited for growing different crops from the ones they are familiar with.

Other displaced residents, urban and rural, have been moved to urban settings far away from the Three Gorges Area. Here housing costs often outstrip the compensation provided. Some of the displaced are not given jobs in their new locations, and some of those who are given jobs lose them. The resistance to living in a distant location is strong. The most successful relocations appear to take place when elderly persons, who are no longer working, move in with relatives. Many displaced persons believe that the project has been tainted by corruption.

5.2.3 Other Detrimental Impacts

The Three Gorges Dam will destroy the habitat of a number of indigenous species, including some found only in China and a species of sturgeon found only in the Yangtze river. The dam will flood three gorges famous for their natural beauty, along with the many antiquities located there. Some critics of the project have suggested that upstream pollution sources will degrade the quality of the water in the reservoir. Downstream concerns include a loss of fertility in farmlands that are no longer regularly flooded by the river and loss of wells because groundwater will not be replenished.

Finally, there is some concern about a catastrophic failure of the dam in the event of an earthquake, especially if engineering decisions or construction quality have been compromised. Some critics even fear that the sheer weight of the reservoir could trigger an earthquake along a fault line that lies under it.

6. HOME USE OF BIOMASS FUELS

A total of 1.4 billion tons of fuelwood was produced worldwide in 1999, and a roughly equal amount is consumed. Global production and use of fuelwood is concentrated in Asia (46%) and Africa (30%). India alone accounts for 14% of global fuelwood production; India, China, and Indonesia together account for almost one-third of the world's production.

The use of wood for fuel declined rapidly in the developed world with the increasing availability of oil and natural gas after World War II. In the United States, wood was used to heat almost one-quarter of all homes in 1940, but only 1% in 1970 (Fig. 3). Wood heat enjoyed something of a resurgence in the 1970s and 1980s, especially among middle- to upper-income households in New England and in the Northwest, but fell off again in the 1990s.

Global peat production amounted to only about 1% by weight of fuelwood production in 1999. Peat remains predominantly a fuel of Europe, where 95% of the world's peat was produced. Finland accounts for 37% of the world's peat production, Belarus and the Russian Federation 15% each, and Ireland 14%.

6.1 Health Effects of Domestic Woodsmoke in Developed Countries

Woodsmoke's constituents form a profile of products of combustion, including respirable particulates less than 10 microns in diameter (PM_{10}), carbon monoxide, nitrogen oxides, and sulfur oxides, polycyclic aromatic hydrocarbons (PAHs), formaldehyde, and various volatile organic compounds (VOCs). Exposure to woodsmoke or to its constituent respiratory irritants (e.g., particulate matter, nitrogen dioxide, sulfur dioxide) has been associated with inhibition of the cilia that line the respiratory tract and sweep foreign objects up and out; constriction of airways due to swelling of the membranes that line them

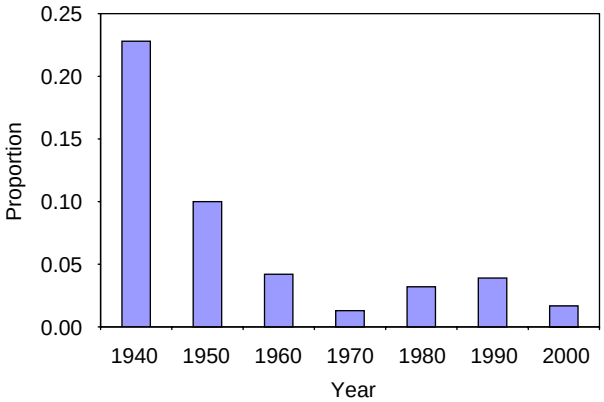


FIGURE 3 Proportion of U.S. housing units using wood for heart, 1940–2000. Data from 2000 U.S. Census.

(chronic airway resistance); a decrease in the volume of air that can be deliberately expelled from the lungs after inhaling (vital capacity); an increase in such respiratory symptoms as coughing, wheezing, and shortness of breath; and increased risk of infectious respiratory disease in children. Formaldehyde is a respiratory irritant that exacerbates asthma and causes headaches; in addition, chronic exposure is associated with increased risk of nasal cancer. Some PAHs have also been shown to be carcinogenic or to have immunosuppressive effects.

6.2 Profile: Effects of Home Heating and Cooking in Developing Nations

It is estimated that approximately half of the world's population—around 3 billion people—use biomass fuels to heat their homes and cook their foods. In rural locations in developing countries, the great majority of households use biomass fuels to meet domestic needs. Biomass fuels are derived from plants or animals, including not only wood and peat, but also animal dung and crop residues. Among these fuels, animal dung and crop residues are the dirtiest to burn but also least expensive; wood and charcoal burn somewhat more cleanly but cost more. Oil, gas, and electricity are the cleanest and most expensive fuels, out of reach of much of the world's population.

Biomass fuels are typically burned in primitive stoves with very incomplete combustion, in an open fire, or even using more than one stove, or some combination of open fire and stove. Depending on the time of year, the type of fuel and stove being used, and the nature of the structure, pollutant concentrations in indoor air may be many times those found in homes in the developed world. The U.S. EPA sets a standard for respirable particulates (particulate matter less than 10 microns in diameter, or PM_{10}) of 50 $\mu g/m^3$ as an annual average and 150 $\mu g/m^3$ as a daily average. (These small particles penetrate deeply into the lung; in fact, particles less than 2.5 microns in diameter penetrate more deeply still.) In a recent study of health impacts of wood stoves in Quebec, Canada, 9% of the homes examined had 24-hour average concentrations greater than 50 $\mu g/m^3$; 67% had 24-hour average concentrations less than 11 $\mu g/m^3$. These measurements contrast sharply with those taken in homes in developing countries where biomass fuels are used for heating and cooking. Here, researchers report typical 24-hour average concentrations in homes ranging from 200 to 5000 $\mu g/m^3$; and peak concentrations of more than 50,000 $\mu g/m^3$ recorded near stoves.

6.2.1 Assessing Exposures to Domestic Biomass Fuels

An individual's cumulative exposure to the products of combustion depends on a number of factors in addition to the type of fuel and the combustion efficiency of the stove. For example, the fire may be burning brightly or smoldering. The use of the stove is affected by season and time of day. The degree to which pollutants are evenly dispersed through the house is affected by whether the cooking and living areas are in the same room or are separate from one another. Ventilation varies widely across homes.

However, research by Majid Ezzati and others has shown that individual patterns of behavior are the most important determinants of exposure. How much time is spent in the cooking area? How much time is spent bending over the stove or fire, adjusting the fuel or stirring the food? Studies have shown that women aged 16 to 50 and girls aged 6 to 15 spend more time in the cooking area, and more time engaged in high-exposure activities related to cooking, than do men or boys in the same age groups. As a result, women's cumulative exposures to indoor air pollution are higher than men's, and the proportion of their exposure that comes from cooking is also higher. Similar gender differences in exposure are not documented for children under 6, who have not yet taken on many household duties. However, young children's exposures are higher than those of older boys or of men aged 16 to 50, reflecting time spent in their mother's company.

6.2.2 The Global Burden of Disease from Domestic Use of Biomass Fuels

Exposure to indoor air pollution in developing countries has been most clearly associated with increased risk of acute lower respiratory infection, middle ear infection (otitis media), and chronic obstructive pulmonary disease. An estimated 1.5 to 2 million deaths each year in developing countries are due to indoor air pollution from the burning of biomass fuels, with perhaps 1 million of these due to acute lower respiratory infections in children under age 5. In sum, the burden of disease from domestic burning of biomass fuels rests most heavily on disenfranchised groups: the world's poorest countries, lower income populations within those countries, women, and young children.

SEE ALSO THE FOLLOWING ARTICLES

Climate Change and Public Health: Emerging Infectious Diseases • Economic Geography of Energy • Environmental Change and Energy • Equity and Distribution in Energy Policy • Geopolitics of Energy • Global Energy Use: Status and Trends • Public Reaction to Nuclear Power Siting and Disposal • Women and Energy: Issues in Developing Nations

Further Reading

Ezzati, M., and Kammen, D. M. (2002). The health impacts of exposure to indoor air pollution from solid fuels in developing countries knowledge, gaps, and data needs. *Environmental Health Perspectives* **110**(11), 1057–1068.

Heming, L. (2001). Reservoir resettlement in China: past experience and the Three Gorges Dam. *Geographical J.* **167**(3), 195–212.

Intergovernmental Panel on Climate Change. Technical Summary: Climate Change 2001: Impacts, Adaptation, and Vulnerability. A report of Working Group II of the Intergovernmental Panel on Climate Change.

Mapel, D. W., Coultas, D. B., James, D. S., Hunt, W. C, Stidley, C. A., and Gililand, F. D. (1997). Ethnic differences in the prevalence of nonmalignant respiratory disease among uranium miners. *Am. J. Public Health* **87**(5), 833–838.

Mazumdar, S., Redmond, C., Sellecito, W., and Sussman, N. (1975). An epidemiological study of exposures to coal tar pitch volatiles among coke oven workers. *J. Air Pollution Control Assn.* **25,** 382–389.

Samet, J. M., Kutvirt, D., Waxweiler, R. J., and Key, C. R. (1984). Uranium mining and lung cancer in Navajo men. *New England J. Medicine.* **310**(23), 1481–1484.

World Energy Council. Survey of Energy Resources (1999). www.worldenergy.org.

Environmental Kuznets Curve

DAVID I. STERN
Rensselaer Polytechnic Institute
Troy, New York, United States

1. **Introduction**
2. **Theoretical Background**
3. **Econometric Framework**
4. **Results of EKC Studies**
5. **Theoretical Critique of the EKC**
6. **Econometric Critique of the EKC**
7. **Other Evidence**
8. **Decomposing Emissions**
9. **Conclusions**

Glossary

cointegration Classical regression analysis assumes that the error term in a regression is normally and independently distributed. An extreme violation of this condition is when the error term is a random walk. Such a regression is termed a spurious regression and the results are not reliable. It usually indicates that a required variable with random walk behavior has been omitted from the model or that the variables in the model follow random walks and are not related to each other.

consistent estimation May show bias and inefficiency, but both are reduced toward zero as the sample size increases.

developed economies Countries with high income levels; typically have an income per capita greater than U.S. $10,000, life expectancy greater than 70 years, and literacy rates greater than 95%.

developing economies Countries with low- and middle-income levels; equivalent to the term "Third World."

econometrics The specialized branch of mathematical statistics applied to the analysis of economic data.

economic growth An increase in economy-wide economic production usually measured by an increase in gross domestic product; also, the process of the economy growing over time.

environmental degradation Any human-caused deterioration in environmental quality.

externality When a production or consumption activity has unintended damaging effects, such as pollution, on other firms or individuals and no compensation is paid by the producer or consumer to the affected parties (negative externality); when activities have beneficial effects for others, such as freely available research results, and the recipients do not pay the generator of the effect (positive externality).

heteroskedasticity When the variance of the error term in a regression model varies systematically across observations or cases.

income elasticity The percentage increase in some variable when income increases by 1%.

income per capita Usually computed as the gross national product of a country for one year divided by the total population.

integrated variable A random walk is an integrated variable—the current value is equal to the previous period's value plus a random shock. This means that the effect of these shocks on the future does not decline over time. Rather, the current value of the variable is an integration of all past shocks.

omitted variables bias Occurs when variables that should be in a regression model are omitted and they are correlated with those variables that are included.

panel data A data set that includes observations of a number of individuals, countries, firms, and the like over a number of time periods such as months or years.

simultaneity bias When there is mutual feedback between two variables and ordinary regression analysis cannot provide consistent estimates of model parameters.

sustainable development If the current average well-being of people could be maintained into the indefinite future.

technological change The invention and introduction of new methods of production as well as new products.

The environmental Kuznets curve is a hypothesized relationship among various indicators of environmental degradation and income per capita. During the early stages of economic growth, degradation and pollution increase, but beyond some level of income per capita (which will vary for different indicators) the trend reverses, so that at high income levels economic growth leads to environmental improvement. This implies that the environmental

 © 2004 Elsevier Inc. All rights reserved.

impact indicator is an inverted U-shaped function of income per capita. Typically, the logarithm of the indicator is modeled as a quadratic function of the logarithm of income.

1. INTRODUCTION

An example of an estimated environmental Kuznets curve (EKC) is shown in Fig. 1. The EKC is named for Simon Kuznets, who hypothesized that income inequality first rises and then falls as economic development proceeds. Emissions of various pollutants, such as carbon dioxide, sulfur, and nitrogen oxides, are tightly coupled to the use of energy. Hence, the EKC is a model of the relationship among energy use, economic growth, and the environment.

The EKC is an essentially empirical phenomenon, but most of the EKC literature is statistically weak. It is very easy to do bad econometrics, and the history of the EKC exemplifies what can go wrong. The EKC idea rose to prominence because few paid sufficient attention to the relevant diagnostic statistics. Little or no attention has been paid to the statistical properties of the data used such as serial dependence and random walk trends in time series, and few tests of model adequacy have been carried out or presented. However, one of the main purposes of doing econometrics is to test which apparent relationships are valid and which are spurious correlations.

When we do take such statistics into account and use appropriate techniques, we find that the EKC does not exist. Instead, we get a more realistic view of the effect of economic growth and technological changes on environmental quality. It seems that most indicators of environmental degradation are monotonically rising in income, although the "income elasticity" is less than 1.0 and is not a simple function of income alone. Time-related effects, intended to model technological change common to all countries, reduce environmental impacts in countries at all levels of income. However, in rapidly growing middle-income countries, the scale effect, which increases pollution and other degradation, overwhelms the time effect. In wealthy countries, growth is slower and pollution reduction efforts can overcome the scale effect. This is the origin of the apparent EKC effect.

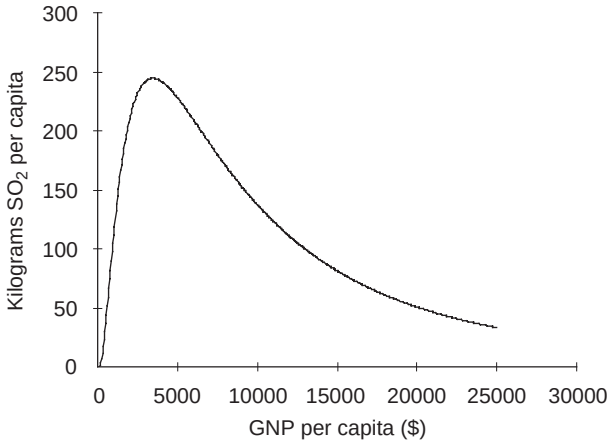


FIGURE 1 Environmental Kuznets curve for sulfur emissions. Data from Panayotou (1993) and Stern *et al.* (1996).

The econometric results are supported by recent evidence that, in fact, pollution problems are being addressed and remedied in developing economies.

This article follows the development of the EKC concept in approximately chronological order. Sections II and III review in more detail the theory behind the EKC and the econometric methods used in EKC studies. Sections IV to VI review some EKC analyses and their critiques. Section VII discusses the recent evidence from Dasgupta and colleagues and others that has changed the way in which we view the EKC. The final two sections discuss an alternative approach, decomposition of emissions, and summarize the findings.

2. THEORETICAL BACKGROUND

The EKC concept emerged during the early 1990s with Grossman and Krueger's pathbreaking study of the potential impacts of the North American Free Trade Agreement (NAFTA) and Shafik and Bandyopadhyay's background study for the *World Development Report 1992*. However, the idea that economic growth is necessary for environmental quality to be maintained or improved is an essential part of the sustainable development argument promulgated by the World Commission on Environment and Development in *Our Common Future*.

The EKC theme was popularized by the International Bank for Reconstruction and Development's (IBRD) *World Development Report 1992,* which argued that "the view that greater economic activity inevitably hurts the environment is based on static assumptions about technology, tastes, and environmental investments" and that "as incomes rise, the demand for improvements in environmental quality will increase, as will the resources available for investment." Others have expounded this position even more forcefully, with Beckerman claiming that "there is clear evidence that, although economic

growth usually leads to environmental degradation in the early stages of the process, in the end the best—and probably the only—way to attain a decent environment in most countries is to become rich." However, the EKC has never been shown to apply to all pollutants or environmental impacts, and recent evidence challenges the notion of the EKC in general. The remainder of this section discusses the economic factors that drive changes in environmental impacts and that may be responsible for rising or declining environmental degradation over the course of economic development.

If there were no change in the structure or technology of the economy, pure growth in the scale of the economy would result in a proportional growth in pollution and other environmental impacts. This is called the scale effect. The traditional view that economic development and environmental quality are conflicting goals reflects the scale effect alone. Panayotou, a proponent of the EKC hypothesis, argued,

> *At higher levels of development, structural change towards information-intensive industries and services, coupled with increased environmental awareness, enforcement of environmental regulations, better technology, and higher environmental expenditures, result in leveling off and gradual decline of environmental degradation.*

Therefore, at one level, the EKC is explained by the following "proximate factors":

1. Scale of production implies expanding production, with the mix of products produced, the mix of production inputs used, and the state of technology all held constant.
2. Different industries have different pollution intensities, and the output mix typically changes over the course of economic development.
3. Changes in input mix involve the substitution of less environmentally damaging inputs to production for more damaging inputs and vice versa.
4. Improvements in the state of technology involve changes in both (a) production efficiency in terms of using less, *ceteris paribus,* of the polluting inputs per unit of output and (b) emissions-specific changes in process that result in less pollutant being emitted per unit of input.

These proximate variables may, in turn, be driven by changes in underlying variables such as environmental regulation, awareness, and education. A number of articles have developed theoretical models about how preferences and technology might interact to result in different time paths of environmental quality. The various studies make different simplifying assumptions about the economy. Most of these studies can generate an inverted U-shape curve of pollution intensity, but there is no inevitability about this. The result depends on the assumptions made and the value of particular parameters. It seems fairly easy to develop models that generate EKCs under appropriate assumptions, but none of these theoretical models has been empirically tested. Furthermore, if in fact the EKC for emissions is monotonic as recent evidence suggests, the ability of a model to produce an inverted U-shaped curve is not necessarily a desirable property.

3. ECONOMETRIC FRAMEWORK

The earliest EKCs were simple quadratic functions of the levels of income. However, economic activity inevitably implies the use of resources, and by the laws of thermodynamics, use of resources inevitably implies the production of waste. Regressions that allow levels of indicators to become zero or negative are inappropriate except in the case of deforestation where afforestation can occur. This restriction can be applied by using a logarithmic dependent variable. The standard EKC regression model is

$$\ln(E/P)_{it} = \alpha_i + \gamma_t + \beta_1 \ln(\mathrm{GDP}/P)_{it} + \beta_2[\ln(\mathrm{GDP}/P)]^2_{it} + \varepsilon_{it}, \quad (1)$$

where E is emissions, P is population, GDP is gross domestic product, ε is a random error term, and ln indicates natural logarithms. The first two terms on the right-hand side are intercept parameters that vary across countries or regions i and years t. The assumption is that although the level of emissions per capita may differ over countries at any particular income level, the income elasticity is the same in all countries at a given income level. The time-specific intercepts are intended to account for time-varying omitted variables and stochastic shocks that are common to all countries.

The "turning point" level of income, where emissions or concentrations are at a maximum, can be found using the following formula:

$$\tau = \exp[-\beta_1/(2\beta_2)]. \quad (2)$$

The model is usually estimated with panel data. Most studies attempt to estimate both the fixed effects and random effects models. The fixed effects model treats the α_i and γ_t as regression parameters, whereas the random effects model treats them as components of the random disturbance. If the effects

α_i and γ_t and the explanatory variables are correlated, the random effects model cannot be estimated consistently. Only the fixed effects model can be estimated consistently. A Hausman test can be used to test for inconsistency in the random effects estimate by comparing the fixed effects and random effects slope parameters. A significant difference indicates that the random effects model is estimated inconsistently due to correlation between the explanatory variables and the error components. Assuming that there are no other statistical problems, the fixed effects model can be estimated consistently, but the estimated parameters are conditional on the country and time effects in the selected sample of data. Therefore, they cannot be used to extrapolate to other samples of data. This means that an EKC estimated with fixed effects using only developed country data might say little about the future behavior of developing countries. Many studies compute the Hausman statistic and, finding that the random effects model cannot be estimated consistently, estimate the fixed effects model. But few have pondered the deeper implications of the failure of this test.

Tests for integrated variables designed for use with panel data find that sulfur and carbon emissions and GDP per capita are integrated variables. This means that we can rely only on regression results that exhibit the cointegration property. If EKC regressions do not cointegrate, the estimates will be spurious. Very few studies have reported any diagnostic statistics for integration of the variables or cointegration of the regressions, and so it is unclear what we can infer from the majority of EKC studies.

4. RESULTS OF EKC STUDIES

The key features differentiating the broad variety of EKC studies can be displayed by reviewing a few of the early studies and examining a single indicator (sulfur) in more detail. This section also discusses EKCs for total energy use, seen by some as a proxy indicator for all environmental impacts.

The early EKC studies appeared to indicate that local pollutants, such as smoke and sulfur emissions, were more likely to display an inverted U-shaped relation with income than were global impacts such as carbon dioxide. This picture fits environmental economics theory; local impacts are internalized in a single economy or region and are likely to give rise to environmental policies to correct the externalities on pollutees before such policies are applied to globally externalized problems.

Grossman and Krueger produced the first EKC study as part of a study of the potential environmental impacts of NAFTA. They estimated EKCs for SO_2, dark matter (fine smoke), and suspended particles (SPM) using the GEMS data set. This data set is a panel of ambient measurements from a number of locations in cities around the world. Each regression involved a cubic function in levels (not logarithms) of purchasing power parity adjusted (PPP) per capita GDP and various site-related variables, a time trend, and a trade intensity variable. The turning points for SO_2 and dark matter are at approximately \$4000 to \$5000, whereas the concentration of suspended particles appeared to decline even at low income levels.

Shafik and Bandyopadhyay's study was particularly influential in that the results were used in the *World Development Report 1992*. They estimated EKCs for 10 different indicators using three different functional forms. Lack of clean water and lack of urban sanitation were found to decline uniformly with increasing income and over time. Deforestation regressions showed no relation between income and deforestation. River quality tended to worsen with increasing income. However, local air pollutant concentrations conformed to the EKC hypothesis, with turning points between \$3000 and \$4000. Finally, both municipal waste and carbon emissions per capita increased unambiguously with rising income.

Selden and Song estimated EKCs for four emissions series—SO_2, NO_x, SPM, and CO—using longitudinal data from World Resources. The data were primarily from developed countries. The estimated turning points all were very high compared with those in the two earlier studies. For the fixed effects version of their model, the estimated turning points were as follows (converted to 1990 U.S. dollars using the U.S. GDP implicit price deflator): SO_2, \$10,391; NO_x, \$13,383; SPM, \$12,275; and CO, \$7,114. This study showed that the turning point for emissions was likely to be higher than that for ambient concentrations. During the initial stages of economic development, urban and industrial development tends to become more concentrated in a smaller number of cities that also have rising central population densities, with the reverse happening during the later stages of development. So, it is possible for peak ambient pollution concentrations to fall as income rises, even if total national emissions are rising.

Table I summarizes several studies of sulfur emissions and concentrations, listed in order of estimated

TABLE I

Sulfur EKC Studies

Author(s)	Turning point (1990 U.S. dollars)	Emissions or concentrations	Purchasing power parity	Additional variables	Data source for sulfur	Time period	Countries/ cities
Panayotou (1993)	3,137	Emissions	No	—	Own estimates	1987–1988	55 developed and developing countries
Shafik and Bandyopadhyay (1992)	4,379	Concentrations	Yes	Time trend, locational dummies	GEMS	1972–1988	47 cities in 31 countries
Torras and Boyce (1998)	4,641	Concentrations	Yes	Income inequality, literacy, political and civil rights, urbanization, locational dummies	GEMS	1977–1991	Unknown number of cities in 42 countries
Grossman and Krueger (1994)	4,772–5,965	Concentrations	No	Locational dummies, population density, trend	GEMS	1977, 1982, 1988	Up to 52 cities in up to 32 countries
Panayotou (1997)	5,965	Concentrations	No	Population density, policy variables	GEMS	1982–1984	Cities in 30 developed and developing countries
Cole *et al.* (1997)	8,232	Emissions	Yes	Country dummy, technology level	OECD	1970–1992	11 OECD countries
Selden and Song (1994)	10,391–10,620	Emissions	Yes	Population density	WRI: primarily OECD source	1979–1987	22 OECD and 8 developing countries
Kaufmann *et al.* (1997)	14,730	Concentrations	Yes	GDP/Area, steel exports/GDP	UN	1974–1989	13 developed and 10 developing countries
List and Gallet (1999)	22,675	Emissions	N/A	—	U.S. EPA	1929–1994	U.S. states
Stern and Common (2001)	101,166	Emissions	Yes	Time and country effects	ASL	1960–1990	73 developed and developing countries

income turning points. On the whole, the emissions-based studies have higher turning points than do the concentrations-based studies. Among the emissions-based estimates, both Selden and Song and Cole and colleagues used databases that are dominated by, or consist solely of, emissions from Organization for Economic and Cooperative Development (OECD) countries. Their estimated turning points were $10,391 and $8,232, respectively. List and Gallet used data for 1929 to 1994 for the 50 U.S. states. Their estimated turning point is the second highest in Table I. Income per capita in their sample ranged from $1,162 to $22,462 (in 1987 U.S. dollars). This is a wider range of income levels than was found in the OECD-based panels for recent decades. This suggests that including more low-income data points in the sample might yield a higher turning point. Stern and Common estimated the turning point at more

than $100,000. They used an emissions database produced for the U.S. Department of Energy by ASL that covers a greater range of income levels and includes more data points than do any of the other sulfur EKC studies.

The recent studies that use more representative samples of data found that there is a monotonic relation between sulfur emissions and income, just as there is between carbon dioxide and income. Interestingly, an estimate of a carbon EKC for a panel data set of OECD countries found an inverted U-shaped EKC in the sample as a whole. The turning point was at only 54% of maximal GDP in the sample. A study using a spline regression implied a within-sample turning point for carbon for high-income countries. All of these studies suggest that the differences in turning points that have been found for various pollutants may be due, at least in part, to the different samples used.

In an attempt to capture all environmental impacts of whatever type, a number of researchers estimated EKCs for total energy use. In each case, they found that energy use per capita increases monotonically with income per capita. This result does not preclude the possibility that energy intensity (i.e., energy used per dollar of GDP produced) declines with rising income or even follows an inverted U-shaped path.

A number of studies built on the basic EKC model by introducing additional explanatory variables intended to model underlying or proximate factors such as "political freedom," output structure, and trade. On the whole, the included variables turned out to be significant at traditional significance levels. However, testing different variables individually is subject to the problem of potential omitted variables bias. Furthermore, these studies did not report cointegration statistics that might indicate whether omitted variables bias was likely to be a problem or not. Therefore, it is not really clear what can be inferred from this body of work.

The only robust conclusions from the EKC literature appear to be that concentrations of pollutants may decline from middle-income levels and that emissions tend to be monotonic in income. Emissions may decline over time in countries at many different levels of development. Given the likely poor statistical properties of most EKC models, it is hard to come to any conclusions about the roles of other additional variables such as trade. Too few quality studies of other indicators apart from air pollution have been conducted to come to any firm conclusions about those impacts as well.

5. THEORETICAL CRITIQUE OF THE EKC

A number of critical surveys of the EKC literature have been published. This section discusses the criticisms that were made of the EKC on theoretical, rather than methodological, grounds.

The key criticism of Arrow and colleagues and others was that the EKC model, as presented in the *World Development Report 1992* and elsewhere, assumes that there is no feedback from environmental damage to economic production given that income is assumed to be an exogenous variable. The assumption is that environmental damage does not reduce economic activity sufficiently to stop the growth process and that any irreversibility is not severe enough to reduce the level of income in the future. In other words, there is an assumption that the economy is sustainable. However, if higher levels of economic activity are not sustainable, attempting to grow fast during the early stages of development, when environmental degradation is rising, may prove to be counterproductive.

It is clear that the levels of many pollutants per unit of output in specific processes have declined in developed countries over time with technological innovations and increasingly stringent environmental regulations. However, the mix of effluent has shifted from sulfur and nitrogen oxides to carbon dioxide and solid waste, so that aggregate waste is still high and per capita waste might not have declined. Economic activity is inevitably environmentally disruptive in some way. Satisfying the material needs of people requires the use and disturbance of energy flows and materials. Therefore, an effort to reduce some environmental impacts might just aggravate other problems. Estimation of EKCs for total energy use is an attempt to capture environmental impact regardless of its nature.

Various critics argued that if there was an EKC-type relationship, it might be partly or largely a result of the effects of trade on the distribution of polluting industries. The Hecksher–Ohlin trade theory, the central theory of trade in modern economics, suggests that, under free trade, developing countries would specialize in the production of goods that are intensive in the production inputs they are endowed with in relative abundance: labor and natural resources. The developed countries would specialize in human capital and manufactured capital-intensive activities. Part of the reductions in environmental degradation levels in developed countries and part of the increases in environmental degradation levels in

middle-income countries may reflect this specialization. Environmental regulation in developed countries might further encourage polluting activities to gravitate toward developing countries.

These effects would exaggerate any apparent decline in pollution intensity with rising income along the EKC. In our finite world, the poor countries of today would be unable to find other countries from which to import resource-intensive products as they themselves become wealthy. When the poorer countries apply similar levels of environmental regulation, they would face the more difficult task of abating these activities rather than outsourcing them to other countries. There are no clear solutions to the impact of trade on pollution from the empirical EKC literature.

Some early EKC studies showed that a number of indicators—SO_2 emissions, NO_x, and deforestation—peak at income levels around the current world mean per capita income. A cursory glance at the available econometric estimates might have led one to believe that, given likely future levels of mean income per capita, environmental degradation should decline from this point onward. However, income is not normally distributed but rather very skewed, with much larger numbers of people below mean income per capita than above it. Therefore, it is median income, rather than mean income, that is the relevant variable. Both Selden and Song and Stern and colleagues performed simulations that, assuming that the EKC relationship is valid, showed that global environmental degradation was set to rise for a long time to come. More recent estimates showed that the turning point is higher; therefore, there should be no room for confusion on this issue.

6. ECONOMETRIC CRITIQUE OF THE EKC

Econometric criticisms of the EKC concern four main issues: heteroskedasticity, simultaneity, omitted variables bias, and cointegration issues.

Heteroskedasticity may be important in the context of cross-sectional regressions of grouped data. Data for a country are a sum or mean of observations for all of the regions, firms, or households within that country. Smaller residuals may be associated with countries with higher total GDPs and populations. Taking heteroskedasticity into account in the estimation stage seems to significantly improve the goodness of fit of globally aggregated fitted emissions to actual emissions.

Simultaneity may arise because increasing energy use drives increasing income, but increasing income may also increase the demand for energy in consumption activities. In addition, if environmental degradation is sufficiently severe, it may reduce national income, whereas the scale effect implies that increased income causes increased environmental degradation. A Hausman-type statistical test for regressor exogeneity can be used to directly address the simultaneity issue. No evidence of simultaneity has been found using such a test. In any case, simultaneity bias is less serious in models involving integrated variables than in the traditional stationary econometric model. The Granger causality test can be used to test the direction of causality between two variables. In tests of causality between CO_2 emissions and income in various individual countries and regions, the overall pattern that emerges is that causality runs from income to emissions or that there is no significant relationship in developing countries, whereas causality runs from emissions to income in developed countries. However, in each case, the relationship is positive, so that there is no EKC-type effect.

Three lines of evidence suggest that the EKC is an incomplete model and that estimates of the EKC in levels can suffer from significant omitted variables bias: (1) differences between the parameters of the random effects and fixed effects models tested using the Hausman test, (2) differences between the estimated coefficients in different subsamples, and (3) tests for serial correlation. Hausman-type test statistics show a significant difference in the parameter estimates for estimated random effects and fixed effects models in the majority of EKC studies. This indicates that the regressors—the level and square of the logarithm of income per capita—are correlated with the country effects and time effects, indicating that the regressors are likely correlated with omitted variables and that the regression coefficients are biased. As expected, given the Hausman test results, parameter estimates turn out to be dependent on the sample used, with the global and developing country estimates of emissions EKCs for sulfur and carbon showing a turning point at extremely high income levels and the developed country estimates showing a within-sample turning point. In the cases where serial correlation in the residuals was tested, a very high level was found, indicating misspecification in terms of either omitted variables or missing dynamics.

Tests for integrated variables designed for use with panel data indicate that sulfur emissions per capita, carbon emissions per capita, and GDP per capita are

integrated. This means that we can rely only on regression results that exhibit the cointegration property. Results for cointegration are less clear-cut than those for integration in the individual time series. For sulfur emissions in 74 countries from 1960 to 1990, approximately half of the individual country EKC regressions cointegrate, but many of these have parameters with "incorrect signs." Some panel cointegration tests indicate cointegration in all countries, and some accept the non-cointegration hypothesis. But even when cointegration is found, the forms of the EKC relationship vary radically across countries, with many countries having U-shaped EKCs. A common cointegrating vector in all countries is strongly rejected.

7. OTHER EVIDENCE

Dasgupta and colleagues present evidence that environmental improvements are possible in developing countries and that peak levels of environmental degradation will be lower than those in countries that developed earlier. They present data that show declines in various pollutants in developing countries over time. They show that although regulation of pollution increases with income, the greatest increases occur from low- to middle-income levels. In addition, diminishing returns to increased regulation would be expected, although better enforcement at higher income levels would also be expected. There is also informal or decentralized regulation in developing countries. Furthermore, liberalization of developing economies over the past two decades has encouraged more efficient use of inputs and less subsidization of environmentally damaging activities. Multinational companies respond to investor and consumer pressure in their home countries and raise standards in the countries in which they invest. Furthermore, better methods of regulating pollution, such as market instruments, are having an impact even in developing countries. Better information on pollution is available, encouraging government to regulate and empowering local communities. Those who argue that there is no regulatory capacity in developing countries would seem to be wrong.

Much of Dasgupta and colleagues' evidence is from China. Other researchers of environmental and economic developments in China came to similar conclusions. For example, China is adopting European Union standards for pollution emissions from cars with an approximately 8- to 10-year lag. Clearly, China's income per capita is far more than 10 years behind that of Western Europe. Furthermore, China has reduced sulfur emissions and even carbon emissions during recent years. Ambient air pollution has been reduced in several major cities, and with the exception of some encouragement of road transport, the government is making a sustained effort in the direction of environmentally friendly policies and sustainable development.

8. DECOMPOSING EMISSIONS

As an alternative to the EKC, an increasing number of studies carry out decompositions of emissions into the proximate sources of emissions changes described in section II. The usual approach is to use index numbers and detailed sectoral information on fuel use, production, emissions, and so on that, unfortunately, is unavailable for most countries. Econometric decomposition models may require less detailed data. Decomposition models represent the relations between energy use and the environment more explicitly than do EKC models.

The conclusion from the studies conducted so far is that the main means by which emissions of pollutants can be reduced is by time-related technique effects, particularly those directed specifically at emissions reduction, although productivity growth or declining energy intensity has a role to play, particularly in the case of carbon emissions where specific emissions reduction technologies do not yet exist. Estimates of emissions-specific reductions for sulfur range from approximately 20% globally over a 20-year period to 50 to 60% in West Germany and The Netherlands during the 1980s alone.

Although the contributions of structural change in the output mix of the economy and shifts in fuel composition may be important in some countries and at some times, their average effect seems less important quantitatively. Those studies that include developing countries find that technological changes are occurring in both developing and developed countries. Innovations may first be adopted preferentially in higher income countries but seem to be adopted in developing countries with relatively short lags. This is seen, for example, regarding lead in gasoline, where most developed countries had substantially reduced the average lead content of gasoline by the early 1990s but many poorer countries also had low lead contents. Lead content was much more variable at low income levels than at high income levels.

9. CONCLUSIONS

The evidence presented in this article shows that the statistical analysis on which the EKC is based is not robust. There is little evidence for a common inverted U-shaped pathway that countries follow as their incomes rise. There may be an inverted U-shaped relation between urban ambient concentrations of some pollutants and income, although this should be tested with more rigorous time series or panel data methods. It seems unlikely that the EKC is a complete model of emissions or concentrations.

The true form of the emissions–income relationship is likely to be monotonic, but the curve shifts down over time. Some evidence shows that a particular innovation is likely to be adopted preferentially in high-income countries first with a short lag before it is adopted in the majority of poorer countries. However, emissions may be declining simultaneously in low- and high-income countries over time, *ceteris paribus,* although the particular innovations typically adopted at any one time could be different in various countries.

It seems that structural factors on both the input and output sides do play a role in modifying the gross scale effect, although on the whole they are less influential than time-related effects. The income elasticity of emissions is likely to be less than 1.0 but not negative in wealthy countries, as proposed by the EKC hypothesis.

In slower growing economies, emissions-reducing technological change can overcome the scale effect of rising income per capita on emissions. As a result, substantial reductions in sulfur emissions per capita were observed in many developed countries during the past few decades. In faster growing middle-income economies, the effects of rising income overwhelmed the contribution of technological change in reducing emissions.

SEE ALSO THE FOLLOWING ARTICLES

Development and Energy, Overview • Economic Growth and Energy • Economic Growth, Liberalization, and the Environment • Energy Ladder in Developing Nations • Environmental Change and Energy • Sustainable Development: Basic Concepts and Application to Energy

Further Reading

Arrow, K., Bolin, B., Costanza, R., Dasgupta, P., Folke, C., Holling, C. S., Jansson, B-O., Levin, S., Mäler, K-G., Perrings, C., and Pimentel, D. (1995). Economic growth, carrying capacity, and the environment. *Science* **268,** 520–521.

Beckerman, W. (1992). Economic growth and the environment: Whose growth? Whose environment? *World Dev.* **20,** 481–496.

Cole, M. A., Rayner, A. J., and Bates, J. M. (1997). The environmental kuznets curve: An empirical analysis. *Environ. Dev. Econ.* **2,** 401–416.

Dasgupta, S., Laplante, B., Wang, H., and Wheeler, D. (2002). Confronting the environmental kuznets curve. *J. Econ. Perspect.* **16,** 147–168.

Grossman, G. M., and Krueger, A. B. (1994). Environmental impacts of a North American Free Trade Agreement. *In* "The U.S.–Mexico Free Trade Agreement" (P. Garber, Ed.). MIT Press, Cambridge, MA.

International Bank for Reconstruction and Development. (1992). World Development Report 1992: Development and the Environment. Oxford University Press, New York.

Kaufmann, R. K., Davidsdottir, B., Garnham, S., and Pauly, P. (1997). The determinants of atmospheric SO_2 concentrations: Reconsidering the environmental Kuznets curve. *Ecol. Econ.* **25,** 209–220.

List, J. A., and Gallet, C. A. (1999). The environmental Kuznets curve: Does one size fit all? *Ecol. Econ.* **31,** 409–424.

Panayotou, T. (1993). "Empirical Tests and Policy Analysis of Environmental Degradation at Different Stages of Economic Development, working paper WP238." International Labor Office, Geneva, Switzerland.

Panayotou, T. (1997). Demystifying the environmental Kuznets curve: Turning a black box into a policy tool. *Environ. Dev. Econ.* **2,** 465–484.

Perman, R., and Stern, D. I. (2003). Evidence from panel unit root and cointegration tests that the environmental Kuznets curve does not exist. *Austr. J. Agric. Resource Econ.* **47,** 325–347.

Selden, T. M., and Song, D. (1994). Environmental quality and development: Is there a Kuznets curve for air pollution? *J. Environ. Econ. Environ. Mgmt.* **27,** 147–162.

Shafik, N., and Bandyopadhyay, S. (1992). "Economic Growth and Environmental Quality: Time Series and Cross-Country Evidence," background paper for World Development Report 1992. The World Bank, Washington, DC.

Stern, D. I. (1998). Progress on the environmental Kuznets curve? *Environ. Dev. Econ.* **3,** 173–196.

Stern, D. I. (2002). Explaining changes in global sulfur emissions: An econometric decomposition approach. *Ecol. Econ.* **42,** 201–220.

Stern, D. I., and Common, M. S. (2001). Is there an environmental Kuznets curve for sulfur? *J. Environ. Econ. Environ. Mgmt.* **41,** 162–178.

Stern, D. I., Common, M. S., and Barbier, E. B. (1996). Economic growth and environmental degradation: The environmental Kuznets curve and sustainable development. *World Dev.* **24,** 1151–1160.

Torras, M., and Boyce, J. K. (1998). Income, inequality, and pollution: A reassessment of the environmental Kuznets curve. *Ecol. Econ.* **25,** 147–160.

World Commission on Environment and Development. (1987). "Our Common Future." Oxford University Press, Oxford, UK.

Equity and Distribution in Energy Policy

ADAM ROSE
The Pennsylvania State University
University Park, Pennsylvania, United States
SNORRE KVERNDOKK
Ragnar Frisch Centre for Economic Research
Oslo, Norway

1. **Introduction**
2. **Welfare Analytics**
3. **Illustrative Examples**
4. **Conclusion**

Glossary

competitive equilibrium The point, defined in prices and quantities, at which demand for goods and services equals the supply. All producers and consumers are so small that they cannot influence the prices in the market.

equity A normative concept related to the distribution of rights and goods. Equity seems to be a broader concept than fairness, while justice is sometimes taken to be an umbrella term that incorporates all dimensions of evaluation besides efficiency.

global justice The solution to a comprehensive redistributive scheme of society-wide justice, where global does not necessarily mean international.

libertarianism A theory for the organization of society emphasizing private property rights and individual freedom. The baseline is that individual freedom prevails except where others may be harmed.

market failure When the markets fail to allocate resources efficiently. Reasons for failure are imperfect competition, natural monopolies, externalities, and public goods. The existence of market failures requires market intervention in the form of supplementing, supplanting, or strengthening the market.

Pareto efficiency A situation characterized by no waste of resources. As long as resources can be allocated so that we experience an increase in production/utility of one firm/individual without increasing other firm's/individual's production/utility, we experience a Pareto improvement. An allocation is Pareto optimal if nobody can be made better off without reducing the well-being/production of others.

social welfare function A representation of the welfare of the society based on individual objectives. The most widely used analytical version is a function of utility of the individuals, in terms of both absolute and relative levels.

utilitarianism A theory for the organization of society. The utilitarian aim is to distribute goods so as to maximize the total utility of members of the society, where "goods" are interpreted broadly to include economic goods, rights, freedom, and political power.

utility possibility frontier A graphical representation of the maximum utility combinations two individuals or groups can attain, given the efficient allocations of finite resources in society.

welfare economics A normative theory for allocating scarce resources and distributing output and income.

The ideal goals of energy policy are to achieve efficiency in production and consumption, as well as an acceptable distribution, of energy goods and services. Efficiency is achieved without policy intervention if there are no imperfections in the market for energy and if all prices reflect the true value of resources. However, it is often necessary to correct prices (e.g., by taxes or subsidies) and to ensure competition (e.g., by anti-trust laws, trade liberalization, or deregulation). Still, even efficient markets can have undesirable distributional consequences for households and producers. A policy that leads to an efficiency gain may not be overall welfare improving if some are made worse off, or it may be viewed as inequitable if it widens the gap between rich and poor. Thus, the distribution of energy goods is also an important part of energy policy.

 © 2004 Elsevier Inc. All rights reserved.

1. INTRODUCTION

1.1 Overview

Given the prominence of energy in many realms of society, as well as the several unique considerations surrounding its extraction, refinement, transportation, and end use, energy has long been the focus of policy making. This has ranged from minor tweaking of market allocation through taxes and subsidies to outright government control. Energy policy began with concern about market failure due to ordinary and natural monopoly but is now increasingly oriented to reducing externalities, primarily environmental pollution. It has oscillated between laissez-faire, government intervention, and deregulation. Allocation issues have dominated but distributional issues have long been important. Moreover, nearly all policies tended to improve the allocation of energy resources also have significant implications for distribution and its assessment in terms of equity or fairness.

This section first briefly summarizes the range of root causes of energy policy and how they relate to distribution and equity. It then offers some basic definitions of equity. Section 2 provides an overview of welfare economics, the theoretical underpinnings for energy policy, and provides some illustrations of key concepts. Section 3 provides some major examples of distribution and equity in energy policy. They are typical but not necessarily representative of the broad range of issues. Rather, they represent the authors' research interests. Finally, Section 4 provides a brief summary.

1.2 Pervasiveness of Energy Policy

1.2.1 Basic Issues

A combination of factors make energy policy pervasive throughout the world. All of these factors have important direct or indirect distributional implications. Energy is a basic necessity for individual, business, and national survival everywhere. In many places, energy is a major commodity of local production and external trade. All stages of the energy cycle are subject to market failures, which typically require policy intervention. Large rents accrue from energy, making it a ripe target, including for purposes of redistribution. Energy is not evenly distributed on a geographic basis, so there is a spatial mismatch between sources and uses. Many energy resources are finite and concern is mounting over the legacy we leave to our descendants. Energy, especially renewable energy, is a key to sustainable economic development, which appears to require much policy guidance. Energy is also the focus of major technological breakthroughs that often require large R&D funding. Finally, energy has become a major political tool and even a weapon.

1.2.2 Relationship to Distribution and Equity

Because energy is a necessity, there is concern that all consumers have access to it and at an affordable price. This has led to price controls or subsidies for lower income groups in many countries. The survival of business customers, however, is often left to the market, but there are some exceptions for small businesses or those in lagging regions. Other cases in which distributional considerations affect access to energy include the aftermath of an earthquake or terrorist attack, or in the face of an electricity shortage during a war or other crisis, where customers are prioritized for rationing. Interestingly, residential customers often receive higher priority than many businesses.

Several countries, and portions of many others, are highly dependent on energy for their livelihood directly or indirectly. Some, such as Appalachia, are relatively poor, while others such as Texas or the Middle East, are relatively wealthy. In essence, every major energy producing region has an exportable surplus that serves as a stimulus to its economy, though the competitiveness of the relevant energy market has significant effects on how much income is derived from this activity and how it is shared between the various factors of production. The situation has also been especially ripe for protectionist policies, especially at the national level, that often favor one segment of society over another. Environmental, technology, and trade policies have and will continue to cause major shifts in the relative economic vitality of these regions.

The uneven geographic distribution of energy has been a root cause of concentration of energy production, political tensions, and redistributional issues. This is most vivid in the international oil arena and several portions of the world where oil has been used as a bargaining chip or even a weapon. A massive redistribution of wealth has taken place in the more obvious case of the OPEC cartel, but also in the more subtle ways because comparative advantage is highly dependent on relative resource endowments.

One of the most prominent factors that require energy policy tension relates to market failure, the

inability of markets to allocate resources efficiently without policy intervention (including market strengthening). All four major categories of failure discussed below have strong implications for the distributional consequences of energy policy. With respect to externalities, defined as a by-products of activities not reflected in the market price, fossil energy use is probably the greatest single source of air pollution, and the production and transportation of fossil energy are major sources of water pollution. Moreover, these types of pollution have transboundary features, which means one geographic area that benefits from energy production or use imposes problems on other regions. Corrective measures, such as pollution taxes or tradeable permits, involve large sums of money. The imposition of a tax may be on a producer, but through economic interdependencies and tax shifting, its incidence is often spread much more broadly. The relative progressiveness and regressiveness of these policy instruments is a long-standing issue in terms of socioeconomic entities (income classes, racial/ethnic groups, age groups). At international levels, this has become a serious issue in terms of burden-sharing between industrialized and developing countries.

The major consideration in the area of public goods is what is often referred to as the subset of common property resources, where the good is nondivisible but does have nonexcludability/nonappropability (of payment) characteristics. One example of this subset may be oil and gas fields. Conventional markets lead to degenerate competition in terms of over-production and plummeting prices. Solutions such as unitization involve difficult decisions on the distribution of rents.

Ordinary monopoly, or market power in general, is one of the most visible market failures in the energy area. This includes U.S. oil companies, state-owned hydropower in many countries, and the world crude oil market. Not only does this cause a misallocation of resources but also a transfer of wealth in the direction of the monopolist. In fact, many public policies (e.g., excess profits taxes) stem from distributional rather than efficiency concerns of this type of market failure.

Natural monopolies arise because scale economies and other factors imply the efficiency of single providers. In addition to the conundrum of attempting to find efficient prices in this context, there is the concern that deviations from competitive levels will come at the expense of consumers. Over the past decade, there has been a reassessment of the extent to which natural monopolies do in fact exist. Deregulation or restructuring have in some cases been poorly designed and have had dramatic distributional consequences, most vividly illustrated by the crisis in California.

Not only does the market often fail to allocate resources efficiently but even more often it fails to distribute them, on the proceeds from them, fairly. Essentially, the market is blind to equity, and extra-market means are usually necessary to steer it in the right direction.

Large rents (profits) accrue from energy production making it a prime target for policy machinations to capture them. These rents are often viewed as indicating monopoly profits or as unfair; however, in many cases, they relate to scarcity rents, returns to high risk in mineral extraction, or to superior quality deposits. The existence of large rents has led to aggressive policies or outright regulation. Also, in the United States and other countries, vast amounts of energy resources are located on public lands, so that the distributional issues are in the forefront. Public officials must often make hard choices between constituencies. This typically has to be accomplished in the absence of market signals, though an increasing number of market institutions have been developed to address the issue (e.g., bonus-lease bidding).

Many energy resources are finite, and there is concern about utilizing them today and leaving fewer or higher cost deposits for future generations (intergenerational equity). This has also caused energy policy to become intertwined with sustainability issues, which have both normative and positive (e.g., public participation) distributional consequences. This has also shifted the emphasis to renewable energy as discussed later.

Energy is key to economic development, especially of the sustainable type. Market forces alone will not promote development, so this is another, and very broad sweeping, area of policy, replete with distributional consequences and equity considerations. Economic development is not simply pursued to raise the absolute standard of living but also to help close the economic gap between countries.

The form of energy has major distributional implications. Fossil fuels are geographically concentrated and subject to natural monopoly considerations, especially in their transportation. Renewable energy, such as solar, wind, dry-rock geothermal, and biomass, are relatively ubiquitous. Hence, they have the potential for decentralizing energy production and even eliminating a great deal of the transportation/transmission needs. These alternate energy resources thus have the ability to lessen the impact

of OPEC and large producers of energy in general and thus to redistribute wealth.

Energy is the subject of major technological advances. This arises because fuels and conversion lend themselves to technical refinement. Major breakthroughs can cause major geographic and socioeconomic redistributions. Much of the research is undertaken or subsidized by government. Issues abound concerning industry favoritism. A "level playing field" is absent here, as well as in most areas of energy policy.

Energy is a political tool. Wealth has always been a major basis for political power, and the concentration of energy deposits, and the nature of energy refining and transportation, have been a major source of strength and tension. The Arab oil embargo is a vivid example, and many contend that the Gulf War, as well as the invasion of Iraq, have been motivated by the politics of oil. The situation is complicated in many parts of the world as a battle between the haves and have-nots or as a remnant of colonialism, one of the ultimate distributional issues.

1.3 Definition of Equity

The concepts of efficiency and equity are central in policy analyses. While Pareto optimality, a situation characterized by no waste of resources, is an accepted efficiency principle, there is not a consensus on a best equity principle. As a result, most economic analyses, also those related to energy, have concentrated on efficiency problems, and equity often plays a secondary role in economic policy making, even if it will always play an important role as a principle of all social interactions.

Distinguishing between the related concepts of equity, fairness, and justice may in general be difficult, and we do not attempt to make extensive distinctions. However, fairness often has a specific meaning in economic theory and philosophical theories of justice. One example is to define fairness as nonenvy. Equity seems to be a broader concept than fairness, while justice is sometimes taken to be an umbrella term, incorporating all dimensions of evaluation besides efficiency.

2. *WELFARE ANALYTICS*

2.1 Welfare Maximization

One branch of mainstream economics studies equity by formalizing the concept of a social welfare function (SWF), in which it was possible to separate efficiency and equity considerations though tradeoffs between the two objectives still prevail. We can illustrate this by a simple diagram containing two major constructs. First is a utility possibility frontier, or UPF, which traces the maximum combination (m) of the utility (U) two individuals or groups (1,2) can attain, given all the efficient allocations of finite resources in society (see curve $U_m^1 U_m^2$ in Fig. 1). Interestingly, all the points on the UPF are Pareto optimal, (i.e., one person cannot be made better off without another person being made worse off). Thus, points such as C and D in Fig. 1 are not Pareto optimal. Note that there are an infinite number of Pareto optimal points on the UPF. We can be assured, under ideal conditions, that the market will lead to a point on the UPF, thus illustrating the first fundamental theorem of welfare economics. But from the standpoint of efficiency alone, the points on the UPF cannot be ranked.

However, if we introduce equity, then we have a way of preferring some to others. In fact, that is the essence of the definition of a social welfare function—a complete ranking of all states (resource allocations and product or income distributions) of the economy. Although many alternative equity concepts exist, the one most prevalently used in welfare economics is "vertical" equity, which holds that equity increases as utility (or income) disparities between individuals decrease. Generally, equity is

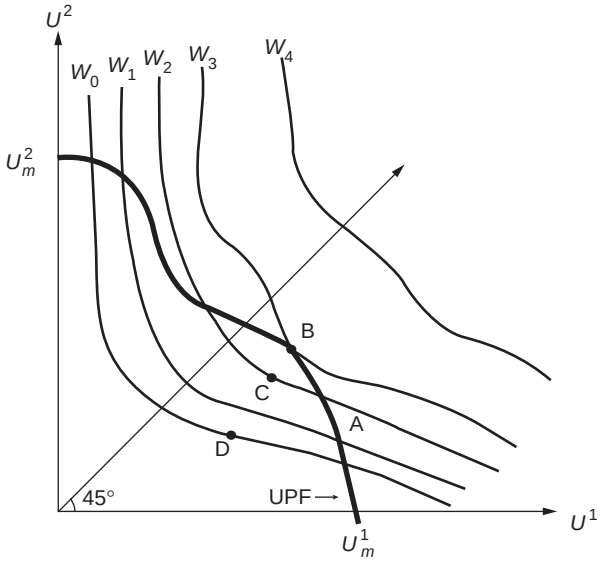


FIGURE 1 Welfare maximization and efficiency-equity trade-offs.

enhanced as joint utility allocations in Fig. 1 move closer to the 45° line.

A social welfare function that embodies both efficiency and equity can be superimposed on Fig. 1. The welfare contours (W_0, W_1, W_2, W_3, W_4) are analogous to indifference curves (i.e., they are horizontal slices of a three-dimensional surface). The shape of these contours reflects equity; the more bowed toward the center, the greater the emphasis on equity.

The welfare optimum is given by point B. Again, there is no guarantee that a market economy will lead to this outcome, and, in fact, given the infinity of possible Pareto optimal outcomes, it is likely that it will not. Moreover, while theoretically society can move from a competitive equilibrium that is not a maximum welfare outcome, such as A, to point B through the use of various types of transfers, there may be some limitations in practice, as discussed in the following section.

For some, one major justification for downplaying equity is the efficiency-equity tradeoff. This refers to the possibility that transfers to promote equity will create a work disincentive to the transferring party (usually the well-to-do), with the ensuing inefficiency limiting society to a lower level of welfare even when utilities (or incomes) are equal or made more equal. Still, this is an empirical question, though most studies indicate the disincentive effect is rather small.

We can use Fig. 1 to illustrate this potential problem. Previously we noted the ideal transfer would move society from point A to point B. However, if the disincentive effect exists, and is weak, the move would be from point A to point C, whereby there is an efficiency-equity tradeoff but an unequivocal welfare gain. But it is also possible that the disincentive is strong and will result in a move from point A to point D, whereby overall welfare is even lower than in the original state. The threat of moving to point D, and even the possibility of moving to point C (note the discussion of benefit cost analysis below), causes many economists to be nervous about pursuing equity goals. We can be more certain of achieving equity with the minimum of efficiency loss if we utilize lump sum transfers, or if we avoid price-distorting policies. The difficulty is that there are few lump sum transfers (i.e., transfers that are not based on effort).

2.2 Social Justice Theories

So far we have discussed distribution within the framework of a social welfare function, which is the mainstream approach in economic theory. These welfaristic theories (meaning theories that focus on welfare outcomes) are in the tradition of utilitarianism. We will now take a brief look at this theory, and also consider some nonwelfaristic approaches.

A theory of justice is, as theories of rationality and ethics, a normative theory. Such a theory has two aspects. First, it will regulate individual rights (and duties); second, it will propose or evaluate a distribution of goods (and burdens). Different theories of justice may weight these two aspects differently. They are, for instance given equal weight by utilitarians, while John Rawls gives political freedom and rights a lexicographical priority over economic distribution.

In the following, we will consider philosophical theories of justice (or global justice theories)—that is, theories that are centrally designed for the whole society and are intended to compensate people for various sorts of bad luck that may result in low levels of income. This means that we will not consider theories for decentralized distribution decisions (or local justice theories), such as who shall perform military service, who shall receive organs for transplantation, and so on, but theories that provides suggestions to society-wide problems such as income distributions or the social security system. These theories are also relevant for energy policy, as we will see in Section 3.

2.2.1 Utilitarianism

The utilitarian aim is to distribute goods so as to maximize the total utility of members of the society, where "goods" are interpreted broadly to include economic goods, rights, freedom, and political power. A utilitarian welfare function is usually defined as the sum of the utility of all the members (i.e., all individuals have an equal weight). One problem with this function as well as for other forms of SWFs is the interpersonal comparison of utility, which is not a straightforward issue. We will, however, not examine these problems in this article.

Even though utilitarianism does not explicitly address equity, its welfare maximization objective does have distributional implications. Different versions of the theory exist. One important distinction is the difference between act-utilitarianism and rule-utilitarianism. The former enjoins us to perform the act that will maximize the utility on every single occasion, while the latter tells us to act so that we will maximize total utility over time. These two forms of utilitarianism may conflict, as many actions

provide immediate pleasure but long-run costs. For example, a high level of fossil fuel consumption today may maximize current welfare, but due to the accumulation of greenhouse gases it will cause a reduction of maximum attainable welfare in the future.

2.2.2 Rawlsian Theory

The theory of John Rawls can basically be seen as a critique of utilitarianism. According to Rawls, utilitarianism has no respect for individuals. A person is not seen as valuable and worth protecting on his or her own right. A classical but grotesque example is if you kill a healthy person and use his or her organs to save the life of five other human beings. This may increase the utility of the society, but it does not respect individual rights. Rawls argues that a theory of justice should respect the individuals as ends in themselves.

The methodological starting point of Rawls theory is the original position. Most persons would argue for what is just or unjust depending on their own position in the society. Therefore, as a starting point to decide the basic structure of the society, which according to Rawls is the primary subject of justice, we have to think about a hypothetical and idealized world where all individuals sit behind a veil of ignorance; the original position. They do not know their abilities, sex, race, or position in the society. All they know is that they are going to live in the society about which they are going to decide the basic structure. In this hypothetical situation, Rawls argues that they will agree on certain principles:

First principle. Each person is to have an equal right to the most extensive total system of equal basic liberties compatible with a similar system of liberty for all.

Second principle. Social and economic inequalities are to be arranged so that they are both (1) to the greatest benefit of the least advantaged and (2) attached to office and positions open to all under conditions of fair equality of opportunities.

The first principle, the principle of greatest equal liberty, is about individual rights. As this principle has a lexicographic priority, none of the basic liberties should be traded against material advancement. However, it is the first part of the second principle, the difference principle, that is mostly referred to. One common misinterpretation in economics is that Rawls argues for maximizing the utility of the least advantaged. However, Rawls theory is not utilitarian, rather an alternative theory and Rawls argues for distributing "primary goods"—that is, goods everybody needs to realize his or her plans of life, independent of what the plan is. While the first principle distributes one subset of the total set of primary goods, the basic liberties, the difference principle distributes another subset including wealth, income, power, and authority. The last part of the second principle, the principle of fair equality of opportunity, requires that we go beyond formal equality of opportunity to insure that persons with similar skills, abilities, and motivation enjoy equal opportunities. Finally it should be noted that Rawls's theory is a theory of intragenerational distributive justice, and the principles cannot easily be extended to intergenerational concerns.

One element of Rawls's theory of justice, which may be particularly relevant when discussing energy policy, is what he defines to be morally relevant and morally arbitrary. The distribution of goods should not be based on morally arbitrary components. To illustrate, consider two elements that are often claimed to contribute to economic growth. One is the human-cooperative activity, and the other is the natural component. The second element is components present on the earth's surface, which are not due to human production. One example is natural resources or assets that are arbitrarily distributed on the earth's surface. In the literature, these are compared to the distribution of natural endowments, defined as such contingencies that are inescapable and present from birth, and may contribute to material advancement, or economic growth because of the implications for human cooperative activity. Examples are nationality, genetic endowments, race, gender, and social class. According to Rawls, natural endowments are "neither just or unjust; nor is it unjust that men are born into society at any particular position. These are simply natural facts. What is just or unjust is the way that institutions deal with these facts." The distribution of natural assets and endowments are, in general, assumed to be morally arbitrary and unacceptable as standards for distribution, since they are beyond the control of human beings. One interpretation is that the geographical distribution of natural resources like oil and gas are simply facts, in the same way as individual talent is a fact. This distribution is, therefore, not just or unjust. However, the income generated from these resources should be considered when discussing justice. Thus, this may be an argument for taxing this income as a part of a distributional policy.

2.2.3 Libertarian Theory

Probably the most widely used decision rule in economic theory is the Pareto criterion, which is most often applied to efficiency and political feasibility (e.g., in bargaining theory). It has also been applied to equity, for instance by a common-sense conception of justice. In libertarian theory, the baseline is that individual freedom prevails except where others may be harmed. Thus, this ethic is in the same line as Pareto superiority saying that there is an improvement in welfare if one or more persons are made better off due to change in resource use as long as the other persons are at least as well off as before. A strict interpretation of libertarian theory is that an act is only immoral if anyone is worse off because of it.

Robert Nozick's theory is the best-known statement of libertarian thought. The theory can be summarized in three principles: justice in appropriation, justice in transfer, and justice in rectification. A distribution of goods is just if it is the end result of an unbroken chain of just transfers, beginning from a just original appropriation. If these conditions are not satisfied, justice in rectification requires that we should establish the distribution that would have occurred if all unjust links in the chain had been replaced by just ones.

The first principle is essentially a finder's keepers principle, where the basic idea is that anyone has the right to appropriate, exploit, and enjoy the fruits of any unowned piece of nature. The principle of just transfers says that the outcome of any voluntary transaction between two or more individuals is just. It is assumed that there is no coercion. If individuals agree on a contract that will benefit all, there is no reason to stop the contract. The only reason to invalidate the contract is if anyone uses its power to make the nonagreement state worse for other parties than it would otherwise have been. The last principle is probably the main weakness of the theory, as identifying the point in time where the earliest violation occurred and thereafter the counterfactual chain of just transfers may be rather indeterminate.

2.3 Link to Energy Policy

Despite the prevalence of distributional and equity considerations in energy policy, much less attention has been paid to them in contrast to the attention to efficiency, policy instrument design, and various industrial organization topics. Several reasons exist for this situation but paramount is the fact that distribution nearly always, and equity by its very essence, are in the realm of normative economics. Energy policy, of course, is also inherently normative, but most analysts feel more comfortable promoting an objective such as economic efficiency, for which the definition has almost universal consensus. Moreover, most interpretations of efficiency call for making all parties better off (or at least not making anyone worse off). On the other hand, there is no consensus on the best definition of equity, as illustrated by the different theories described earlier. Furthermore, the promotion of this objective (in a static context, as opposed to a case like economic development) is often viewed as a zero-sum game, where improvements for one entity must emanate from another one.

Of course, in the real world, few policies do not at least have some adverse effect on some parties. However, this is finessed by benefit-cost analysis (BCA), which evaluates only the overall net gain (and ignores the fact that some incur net cost). In essence, BCA stipulates that benefits and costs be evaluated equally no matter to whom they accrue. Some attempts have been made over the years, however, to introduce distributional issues into BCA by suggesting winners compensate losers and by assigning weights to individuals. However, when choosing a SWF for doing a BCA analysis, ethical views and other specific assumptions are made at the same time. Thus, another approach is to provide sufficient welfare indicators (i.e., information that can be used as input in a BCA analysis). Instead of providing information that is the output from a certain BCA analysis, the analyst can provide factual information about different population groups (such as their social state, size, or characteristics, as well as their changes in income and access to public goods).

As will be shown later, in the primary theoretical underpinnings of economic policy, welfare economics, equity receives considerable attention. Moreover, direct redistribution is viewed as a sound and even ideal, policy option to promote equity, while not adversely affecting efficiency. The classic example is the debate over natural gas well-head price ceilings in the United States. In 1953, a U.S. court ruled that natural gas extraction had monopoly elements and imposed price caps at the wellhead. Opposition to this policy increased over the next two decades because of the bureaucratic backlog of legitimate cost adjustments and instances of major gas shortages. The counter-argument, however, began to hinge less on the monopoly condition and more on concern that, if wellhead activity were deregulated, the ensuing fly-up of gas prices would have an

especially adverse effect on the poor. Thus, the debate hinged on the relative merits of two approaches: (1) maintaining price caps to make low-cost natural gas available to the needy, while distorting market signals potentially causing shortages, and (2) deregulating gas prices and providing some form of tax relief to low income groups. Natural gas wellhead prices were in fact deregulated beginning in 1978. No major legislation was passed to provide subsidies or tax relief for low income groups, though attempts were made to lessen the potential fly-up by a slow phasing in of deregulation, and hence market price increases. Fortunately, large price increases did not take place, primarily due to the drop in world oil prices and economic recession of the early 1980s. This is an example showing that efficiency and equity can be in conflict, though external circumstances can modify the tension. Also, ideal policies do not always take place (an example of political failure analogous to market failure), and distributional issues may have a lower priority than overall efficiency.

3. ILLUSTRATIVE EXAMPLES

3.1 Carbon Taxes

A number of studies have examined the income distribution impacts of carbon taxes or carbon emission permits. The complexity and international nature of the issue is one reason, an increasing number of analyses are performed in a general equilibrium context.

Gbadebo Oladosu and Adam Rose begin an examination of this issue by summarizing the three special features most emphasized to distinguish the impacts of these policies in contrast to the incidence of taxes in general. First, although the initial focus is on a few but very prominent sectors that emit carbon (coal/oil/gas extraction, transportation, and refining), the fundamental role of these products, however, means that carbon reduction policies will eventually ripple throughout the economy, with possibly surprising outcomes. This is one of the major reasons computable general equilibrium models are used (i.e., numerical simulation models based on behavioral responses by producers and consumers to price signals in interconnected markets).

Second, fossil energy products and most energy-intensive processed goods (food, housing, automobiles) are necessities, making it relatively more difficult to substitute away from them. Spending on necessities is inversely related to income and, hence, all other things being equal, carbon taxes would lean toward being regressive in partial equilibrium terms.

Third, unlike most existing taxes, carbon taxes are not aimed primarily at raising revenue. Moreover, they do not create a distortion in the price system but are intended to correct one. The factors have important implications for the disposition of carbon tax revenues (or revenues from the auction of carbon emission permits), including the possibility of using carbon tax revenues for tax relief that promises to reduce the distortionary nature of the preexisting tax system. This revenue recycling can take a number of forms (reductions in personal income taxes, corporation income taxes, etc.), with different distributional impacts (both regressive and progressive). Again, however, the final impacts of these alternatives are not a priori obvious when one allows for general equilibrium considerations.

Overall, a large number of other factors, both unique to carbon taxation and applicable to tax policy in general, can have a major bearing on the relative unevenness of impacts. It is also important to note several factors that affect the size of the aggregate impact, since it will also have a bearing on the degree to which the baseline income distribution changes. Of course, the size of the aggregate impact can affect the distribution of impacts in highly nonlinear models or where such factors as income elasticities (i.e., the percentage change in demand following from a 1% increase in income of demand vary strongly across income groups). Major factors include the following:

1. The energy-intensity of the economy. The higher this factor, the larger the overall impact and the more profound income inequalities of impacts can become in relation to the baseline.

2. The magnitude of the carbon tax or emission permit price, for reasons similar to factor 1.

3. The unit upon which the tax is based (e.g., energy equivalent, carbon emissions, or carbon content). This bears on the relative bluntness or precision of the policy and hence its cost-effectiveness and overall impacts.

4. The narrowness or breadth of products or entities on which the tax is imposed. For example, a tax on energy as a whole may be regressive but a tax on motor fuels may be progressive.

5. The point of initial imposition of the tax (i.e., upstream suppliers of energy or downstream users of energy). This is related to whether a tax on production or consumption is specified. Progressivity

is likely to be greater if imposed upstream and if tax shifting is difficult.

6. The ability to shift the tax forward onto customers or backward onto factors of production. In terms of the latter, shifts on to labor are likely to be more regressive than shifts onto capital.

7. The extent of factor mobility, as in the regressivity of transitional effects on coal miners having to find jobs in other industries.

8. The extent to which general equilibrium effects are taken into account. While this typically exacerbates the level of impacts, it mutes their unevenness as they are spread out over the economy as a whole.

9. The extent of production/income distribution/consumption interactions. For example, a large decrease in coal production may have a disproportionate effect on income of high-wage unionized miners, but the decrease in their consumption may be for products that are characterized by a predominant number of low-wage earners.

10. The extent to which dynamic effects are taken into account (e.g., with respect to savings and investment). The income distribution has an effect on economic growth, which in turn affects future distribution. Here progressivity is often thought to have a detrimental effect on future growth, though the effect on future income distribution is ambiguous.

11. The use of annual income versus lifetime income as a reference base (e.g., as proxied by consumption). The latter is the more appropriate measure given the long-run nature of the issue.

12. The extent to which demographic considerations pertaining to household composition are taken into account; related to this is the demarcation of income groups, especially at the highest and lowest levels.

13. The type of revenue recycling (including lump-sum transfers) and in contrast to alternatives such as budget deficit reduction and individual and corporate tax relief. The latter is usually considered the most regressive.

14. Asset market considerations, such as the extent to which financial returns are affected. This factor bears primarily on higher income groups.

15. The degree to which the impacts result in unemployment. Those already in lower income groups are less able to withstand the shocks of both temporary and long-term unemployment.

16. Basic parameters and assumptions of the analytical model (especially price elasticities of demand and supply, elasticities of substitution with respect to inputs and imports, market structure, etc.). For example, the higher the elasticities of substitution between energy and other factors, the less the negative impacts. Also, the greater the variation in price and income elasticities of demand, the greater the potential progressivity or regressivity.

If we try to summarize studies of carbon taxes on households, we find that the distributional effects of a carbon tax can often be regressive unless the tax revenues are used either directly or indirectly in favor of the low-income groups. Consequently, there seems to be a conflict between efficiency and equity goals, although some analyses are less conclusive. Due to the many causal factors involved and the relative lack of analyses on this topic, it is difficult to draw conclusions about country differences. The same applies to the impact of model type or size of the carbon tax.

3.2 OPEC and Environmental Policy

A carbon tax also has distributional consequences for producers of energy. One important issue in the international context is how a carbon tax affects the producers of fossil fuels. In the ongoing international climate negotiations, the possible impacts on fossil fuel producers and producing countries have played an important role, and both coal producers and OPEC countries have shown a very reluctant position in the negotiations.

A study by Elin Berg, Snorre Kverndokk, and Knut Einar Rosendahl shows the impacts on fossil fuel producers and especially oil producers, as well as the distribution of burden between producers and consumers, of a global carbon tax. This is analyzed by an intertemporal equilibrium model for fossil fuels, where the dynamic aspects of resource extraction is taken into account. According to their results, a global carbon tax of $10 per barrel will hurt OPEC more than other producers. The reason is that the cartel finds it optimal to reduce its production in order to maintain a high price level when the tax is introduced. Hence, the crude oil price is almost unchanged initially, and, in the first 40 years, the tax is almost completely shifted to the consumers. As the producer price of oil grows slower compared to the reference case, non-OPEC accelerates its oil production. Thus, while the market structure is important in explaining OPEC's behavior in the model, the dynamic aspect explains the behavior of the fringe. The oil wealth of non-OPEC is reduced by merely 8%, whereas OPEC's wealth is reduced by more than 20%. Thus, if we do not consider the environmental

benefits, the main losers under the carbon tax regime will be the consumers and OPEC, while the tax revenue is collected by the governments in the consuming countries.

The importance of market power is demonstrated by the results of a competitive model—that is, a model where all consumers and producers behave competitively (no market power). Here low-cost producers (OPEC countries) supply significantly more initially than in the model with market power, which implies a much lower oil price, making oil production unprofitable for high-cost producers (non-OPEC) in the first period. In this situation, a carbon tax will have a relatively much higher impact on the non-OPEC oil wealth compared to that noted earlier (i.e., about 40%), as the low-cost producers will no longer act to maintain the price. However, this reduction is less dramatic for non-OPEC than the original change in market structure, which reduces the oil wealth by more than 70%. For the OPEC countries the relative loss due to the carbon tax is about the same as mentioned previously.

The results may explain why OPEC countries are reluctant to introduce global carbon taxes or other restrictions on carbon emissions. In contrast, other oil producers seem to have less to fear if such an agreement were reached, and this seems to be partly consistent with their positions toward carbon restrictions. Therefore, economic factors may be important in explaining attitudes of oil producers in global warming negotiations.

3.3 Distribution of Tradable Emission Permits

Several alternative criteria have been put forth for the interregional and international analysis of the equity implications of environmental policy. Nine equity principles, a general operational rule emanating from each, and a corresponding rule applicable to the allocation of tradable emission permits at the regional level are presented in Table I.

Of course, there are other levels of application of equity criteria, including sectors and households. In some ways, these units are superior to regions, because they address the issue of relative impacts in terms of welfare of the individual (the basic unit in a democracy). Two reasons are offered for orienting the analysis to the regional level, however. First, are limitations on the availability of data for production and household delineations according to welfare levels (e.g., income brackets or profit rates). Second, political realities indicate that regions are important. This is reinforced by the "regional" character of legislative bodies and several major economic sectors (e.g., oil and gas production, steel), which have translated into regional influence in previous environmental policy making, such as U.S. acid rain legislation. Several other important regional dimensions of the issue relating to natural and social science considerations.

To be operational, an equity criterion must be applied to a reference base, essentially a metric or index, which itself has no ethical content. Consider the following six permit allocation formulas (consisting of an equity criterion and an associated reference base):

1. Sovereignty (emissions based)
2. Egalitarian (population based)
3. Economic activity (GRP based)
4. Ability to pay (inverse-GRP based)
5. Sovereignty (energy-use based)
6. Sovereignty (energy-production based)

Note that there is not a one-to-one correspondence between criteria and reference bases. For example, the sovereignty criterion can be implemented according to more than one reference base, and gross regional product (GRP) can serve as a reference base for more than one criterion. Reference base distinctions are important for several reasons. First, as in the case of three bases applied to the sovereignty criterion, they can reflect alternative energy policy positions. For example, Formula 5 simulates a downstream administered program on fossil energy end users, while Formula 6 simulates an upstream program on fossil energy producers. Most important, the welfare implications of various references bases for the same equity criterion will differ as well.

The results of the application of the six equity formulas to the sharing of the Kyoto protocol target within United States by Adam Rose and Zhong Zhang are presented in Fig. 2. Some interesting results arise.

First, for some of the formulas (e.g., Sovereignty/Energy Use), there is very little variation of welfare gains across regions. Second, for some regions (e.g., Southwest, Mid-Atlantic, New England), there is little difference across equity formulas. For some regions, the equity formula does matter greatly (e.g., south central, north central). Note also the results for some equity criteria differ greatly from their application in the international domain. For example, the egalitarian (per capita) criterion leads

TABLE I

Alternative Equity Criteria for Greenhouse Gas Mitigation Policy

Criterion	Basic definition	General operational rule	Operational rule for CO_2 permits
Allocation based			
Sovereignty	All regions have an equal right to pollute and to be protected from pollution	Cut back emissions in a proportional manner across all regions	Distribute permits in proportion to emissions, energy-use, land area, etc.
Egalitarian	All people have an equal right to pollute or to be protected from pollution	Allow emissions in proportion to population	Distribute permits in proportion to population
Ability to pay	Mitigation costs should vary directly with regional economic well-being	Richer regions should shoulder a higher proportion of gross cost of abatement[a]	Distribute permits inversely to GRP
Econ Activity	All regions should be allowed to maintain their standard of living	Richer regions should not be penalized	Distribute permits in proportion to GRP
Outcome based			
Horizontal	All regions should be treated equally in terms of changes in welfare	Equalize net welfare change across regions (net loss as proportion of GRP equal for each region)[b]	Distribute permits to equalize net welfare change (net loss as proportion of GRP equal for each region)[b]
Vertical	Welfare gains should vary inversely with regional economic well-being; welfare losses should vary directly with GRP	Progressively share net welfare change across regions (net loss proportions directly correlated with per capita GRP)[b]	Progressively distribute permits (net loss proportions directly correlated with per capita GRP)[b]
Compensation	No region should be made worse off	Compensate net losing regions	Distribute permits so no region suffers a net loss of welfare
Process based			
Rawl's Maximum	The welfare of the worst-off region should be maximized	Maximize the net benefit to the poorest regions	Distribute largest proportion of net welfare gain to poorest regions
Consensus	The negotiation prices is fair	Seek a political solution promoting stability	Distribute permits in a manner that satisfies the (power weighted) majority of regions

[a] Gross cost refers to mitigation cost only and does not include benefits or permit transactions.
[b] Net welfare change is equal to the sum of mitigation costs + permit sales revenues − permit purchase expenditures.

to the relatively greatest cost burden being incurred by one of the regions of the United States with the lowest per capita income (south central states). This is just the opposite of the result in the international domain, where countries such as China and India stand to gain the most from application of this criterion.

4. CONCLUSION

Energy is a necessity in both consumption and production. Thus, energy policies may have large consequences on income and profit. The ideal goals of this policy are to achieve efficiency, as well as an acceptable distribution of energy goods and services. However, even efficient markets can have undesirable distributional consequences for households and producers. A policy that leads to an efficiency gain may not be overall welfare improving if some are made worse off, or it may be viewed as inequitable if it widens the gap between rich and poor. This makes the distribution of energy goods an important part of energy policy.

This article briefly summarizes the range of root causes of energy policy and how they relate to distribution and equity, and it provides an overview of welfare economics and social justice theories, the theoretical underpinnings for energy policy. It also provides some illustrations. As our survey shows, there is no single best equity principle nor one that is universally agreed upon. Therefore, it is difficult to develop unambiguously good policy proposals with respect to the distributional effects of energy policy

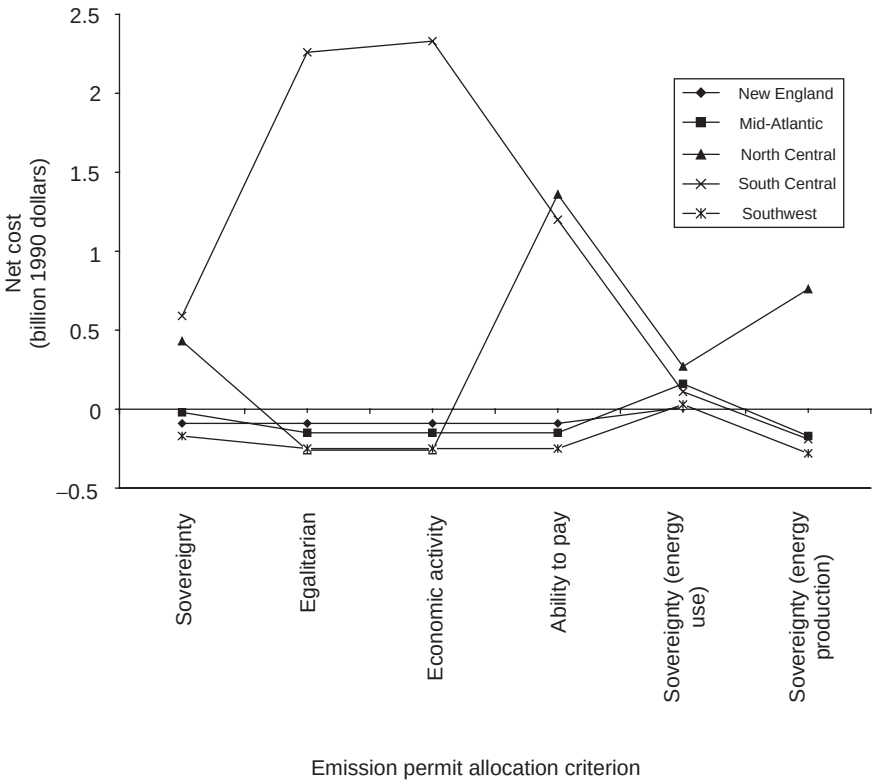


FIGURE 2 Net cost of meeting U.S. Kyoto protocol commitments in selected regions under alternative permit allocation criteria.

in particular. However, policymakers should be aware of the distributional effects before a policy is implemented in general.

SEE ALSO THE FOLLOWING ARTICLES

Coal Industry, Energy Policy in • Cost–Benefit Analysis Applied to Energy • Development and Energy, Overview • Economic Growth, Liberalization, and the Environment • Electric Power Reform: Social and Environmental Issues • Energy Ladder in Developing Nations • Environmental Injustices of Energy Facilities

Further Reading

Adelman, M. A. (1972). "The World Petroleum Market." Johns Hopkins University Press, Baltimore, MD.

Berg, E., Kverndokk, S., and Rosendahl, K. E. (1997). Market power, international CO_2 taxation and petroleum wealth. *Energy J.* **18**(4), 33–71.

Brennan, T., Palmer, K., Kopp, R., Krupnick, A., Stagliano, V., and Burtraw, D. (1996). "A Shock to the System: Restructuring American's Electricity Industry." Resources for the Future, Washington, DC.

Bye, B., Kverndokk, S., and Rosendahl, K. E. (2002). Mitigation costs, distributional effects, and ancillary benefits of carbon policies in the Nordic countries, the UK, and Ireland. *Mitigation and Adaptation Strategies for Global Change* **7**(4), 339–366.

Daniziger, S., Haveman, R., and Plotnick, R. (1986). How income transfer programs affect work, savings, and income distribution. *J. Economic Literature* **19**, 1–36.

Gordon, R. L. (1985). Energy policy issues. *In* "Economics of the Mineral Industries" (W. Vogely, ed.), 4th ed. AIME, New York.

Goulder, L. (1995). Effects of carbon taxes in an economy with prior tax distortions: An intertemporal general equilibrium analysis. *J. Environmental Economics and Management* **29**, 271–295.

Harrison, D. (1994). "The Distributive Effects of Economic Instruments for Environmental Policy." OECD, Paris.

McAvoy, P., and Pindyck, R. (1975). "The Economics of the Natural Gas Shortage (1960–1980)." North-Holland, Amsterdam.

Munasinghe, M. (1993). "Energy Policy Analysis and Modeling." World Bank, Washington, DC.

Oladosu, G., and Rose, A. (2003). "The Income Distribution Impacts of Climate Change Mitigation Policy in the Susquehanna River Basin Economy." Paper presented at the 14th Annual Conference of the European Association of Environmental and Resource Economics, Bilbao, Spain.

Rawls, J. (1971). "A Theory of Justice." Oxford University Press, Oxford.

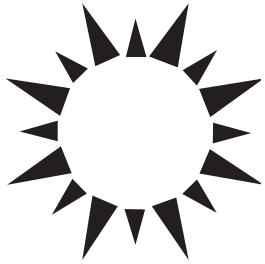
Roemer, J. (1996). "Theories of Distributive Justice." Harvard University Press, Cambridge, MA.

Rose, A., Stevens, B., and Davis, G. (1988). "Natural Resource Policy and Income Distribution." Johns Hopkins University Press, Baltimore, MD.

Rose, A., and Zhang, Z. X. (2003). Interregional burden-sharing of greenhouse gas mitigation in the United States. *Mitigation and Adaptation Strategies for Global Change* (forthcoming).

Ethanol Fuel

CHARLES E. WYMAN
Dartmouth College
Hanover, New Hampshire, United States



1. **Use as a Fuel**
2. **Ethanol Production Approaches**
3. **Production from Sugar Crops**
4. **Production from Starch Crops**
5. **Production from Cellulosic Biomass**
6. **Directions and Opportunities for World Ethanol Production**
7. **Conclusions**

Glossary

anhydrous ethanol Essentially pure ethanol containing very little, if any, water.

biomass Plant matter of recent (nongeologic) origin or materials derived therefrom.

cellulase A combination of enzymes that catalyze the reaction of water with cellulose to release shorter chains and ultimately soluble glucose sugar.

cellulose A polymer of up to about 10,000 glucose molecules joined by beta 1-4 linkages in long, straight cellulose chains that can align parallel to one another and join by hydrogen bonds in crystalline regions that typically contain much of the cellulose. About 40 to 50% of the structural portions of plants is cellulose.

cellulosic biomass Biomass composed primarily of inedible fibrous plant parts having cellulose as a prominent component. Examples include grass, wood, and agricultural and forestry residues. Also known as lignocellulosic biomass.

corn A grain composed of about 70% starch, 10 to 11% crude protein, 4.5 to 6.0% oil, 6% hemicellulose, 2 to 3% cellulose, 1% lignin, and 1% ash.

E85 A mixture of 85% ethanol with 15% gasoline.

E95 A mixture of 95% ethanol with 5% gasoline.

enzymes Proteins originating from living cells and capable of catalyzing chemical changes in organic substances.

ethanol A chemical compound with the formula C_2H_5OH that is widely used in beverages, solvents, cosmetics, and other commercial products and is also an excellent transportation fuel. Ethanol is readily metabolized and biodegradable and does not persist in the environment.

hemicellulose An amorphous and noncrystalline polymer usually containing arabinose, galactose, glucose, mannose, and xylose sugars and typically including lower amounts of other compounds such as hexuronic acids and methoxyl, acetyl, and free carboxyl groups. Hemicellulose represents about 25 to 30% of the overall dry weight of cellulosic biomass and works with lignin to cement the cellulose fibers into a composite material of great strength.

hydrolysis The reaction of water with a compound to form other compounds.

hydrous ethanol Often a mixture of 95.5% ethanol and 4.5% water as used in Brazil.

lignin A complex phenylpropanoic acid polymer that is chemically linked with hemicellulose to cement cellulose together. Lignin represents about 15 to 20% of the dry weight of cellulosic biomass and is not fermentable at an appreciable rate.

starch A polymer of glucose molecules joined predominately by alpha 1-4 linkages.

sugar A family of simple, often sweet, compounds consisting of carbon, hydrogen, and oxygen obtained in the juice of many plants and particularly from sugarcane and sugar beets. Also refers specifically to sucrose, a particular type of sugar. Other sugars typically contained in biomass include arabinose, galactose, glucose, mannose, and xylose.

Ethanol provides a valuable liquid fuel alternative to a transportation sector that is disproportionately dependent on one politically volatile and finite energy source: petroleum. In 2000, the United States consumed almost 100 quads of energy (a quad is 10^{15} BTUs or 1 EJ), with the largest single fraction, almost 40%, coming from petroleum. Furthermore, while the domestic supply and demand for all other energy sources are nearly balanced in the United States, close to two-thirds of petroleum is imported, with unstable regions of the world controlling the largest sources, and imports continue to increase as domestic well productivity drops. This high dependence on imported oil results in strategic vulnerability to price and availability disruptions such as

 © 2004 Elsevier Inc. All rights reserved.

occurred in the "energy crises" of the 1970s. Imported petroleum also accounts for about 45% of the net trade deficit of the United States and was the single largest contributor at $59.2 billion in 1999. Transportation uses over two-thirds of petroleum consumed in the United States and is almost totally dependent on oil, with over 96% of transportation energy from this one source. Furthermore, vehicle miles and total numbers of vehicles continue to rise, increasing petroleum demand more. Thus, abundant alternative domestic energy sources for transportation are vital. Although other countries may not have the high per capita transportation energy use of the United States, many suffer from similar oil-related strategic and economic vulnerabilities.

Our transportation fuel situation presents other important challenges. First, petroleum is a finite resource that cannot be sustained indefinitely, and its price must increase and availability decrease over time. Petroleum also represents a valuable resource for chemicals and materials that benefit from the unique complexity of petroleum far better than burning it to carbon dioxide and water. Additionally, burning petroleum contributes about 43% of the carbon dioxide released into the atmosphere each year in the United States, and this large quantity makes carbon dioxide the most potent greenhouse gas that could cause global climate change. Furthermore, because of its high dependence on petroleum, transportation contributes about 33% of the carbon dioxide released in the United States annually, more than for residential, commercial, or industrial applications. The United States is the largest contributor to release of greenhouse gases accounting for about 25% of the world total, and new fuels that dramatically reduce such emissions are essential to change this situation. Better transportation fuels and utilization systems can also substantially improve urban air quality and reduce environmental effects of handling and consuming fuels.

Ethanol has favorable fuel, resource, and environmental attributes as a transportation fuel. Ethanol is a liquid fuel that readily integrates with our infrastructure, is widely accepted for blending with gasoline, and can be used in nearly pure form. Large quantities of this fuel are now made from starch in the United States and sugar in Brazil as well as lesser amounts in other countries. Ethanol can also be made from abundant, sustainable sources of low-cost cellulosic biomass such as agricultural and forestry residues, portions of municipal solid waste, and herbaceous and woody crops that can dramatically reduce petroleum consumption and greenhouse gas emissions. For the latter, costs could potentially be reduced to be competitive as a pure fuel. This article summarizes some of the key attributes of this beneficial fuel and outlines how it can be produced from these resources.

1. USE AS A FUEL

1.1 Background

Ethanol is also known as ethyl alcohol or fermentation alcohol. Although ethanol is often referred to as just "alcohol," particularly as a beverage, the term "alcohol" technically refers to a broader family of compounds with the chemical formula ROH in which R contains one or more alkyl groups containing carbon and hydrogen atoms and OH refers to the hydroxyl group containing oxygen and hydrogen atoms, respectively. Ethanol is a member of the alcohol family and has the chemical formula C_2H_5OH in which C, H, and O refer to carbon, hydrogen, and oxygen atoms, in that order.

Ethanol is a colorless, clear liquid that looks like water and is completely miscible with it. Ethanol has a somewhat sweet flavor when diluted with water; a more pungent, burning taste when concentrated; and an agreeable ether-like odor. It is widely used in medicines, lotions, tonics, colognes, rubbing compounds, and solvents and also for organic synthesis. Beverage ethanol dates back to ancient times, with Sumerians and Babylonians making beer before 6000 BC. It was also among the earliest compounds investigated with the study of ethanol from distillation of wine likely in 11th-century Italy.

1.2 Transportation Fuel in Internal Combustion Engines

Ethanol is more volatile than water, flammable, burns with a light blue flame, and has excellent fuel properties for spark ignition internal combustion engines. Its research octane number (RON) is 109, its motor octane number (MON) is 90, and its average octane is 99 while the latter for regular gasoline is typically about 88. The latent heat of vaporization of 663.4 kJ/liter (2,380 Btu/gal) is higher than that of gasoline at about 250 kJ/liter (900 Btu/gal). The energy content of ethanol is 21.2 mega joules/liter (76,000 Btu/gallon), as the lower heating value, about two-thirds that of gasoline at about 30.1 mega joules/liter (115,000 Btu/gallon). Thus, somewhat more ethanol is needed to achieve the same

range as gasoline. On the other hand, engine efficiency can be improved by running at higher compression ratios that take advantage of the better fuel properties of pure ethanol, and ethanol can achieve about 75% or more of the range of gasoline on a volume basis for an optimized ethanol engine. The boiling point of ethanol is 78°C (172°F) while that of gasoline is about 27°C (81°F), and the Reid vapor pressure (closed vessel pressure at 38°C (100°F)) for ethanol is only 16 kilopascals (2.3 psi) versus about 71 kilopascals (10.3 psi) for gasoline. Thus, evaporative emissions that cause smog formation are less for pure ethanol than for gasoline, but because ethanol is a polar molecule that resembles water more than gasoline, low levels of ethanol blended with gasoline increase vapor pressure. Although vapor pressure drops as more ethanol is added, U.S. vehicle manufacturers do not warranty conventional vehicles with these blends, and fuel providers have been reluctant to reduce the vapor pressure of gasoline to compensate. The lower vapor pressure of pure ethanol also means that starting is more difficult in cold weather. Ethanol has a very low toxicity, particularly in comparison to other fuels, and is readily biodegradable in water and soils, reducing penetration of plumes from leaks and consequences of spills compared to MTBE and petroleum-based fuels.

In Brazil, ethanol, mostly from cane sugar, is produced as either anhydrous ethanol that contains 99.6% ethanol and 0.4% water for use in 22% blends with gasoline or as hydrous ethanol containing 95.5% by volume ethanol and 4.5% water that is burned directly as a pure fuel. Brazil originally promoted ethanol use through creating a state-owned company, Petrobras, which guaranteed sales to companies and sold the fuel to customers at a discount relative to gasoline. Economic incentives were also offered to ethanol producers. As a result, annual production of hydrous ethanol fueled vehicles climbed from only 3114 in 1979 to more than 697,000 at the peak in 1986. Hydrous ethanol use peaked at about 10.88 billion liters (2.88 billion gallons) in 1989 with another 1.72 billion liters (0.46 billion gallons) added to gasoline as 22% ethanol blends during that year, as shown in Fig. 1 However, incentives were then reduced, and hydrous vehicle manufacture virtually ceased by 1999 due to sugar pricing and ethanol supply problems. Although production of some hydrous ethanol vehicles resumed recently and the Brazilian government approved increasing ethanol content in blends to 26% for gasoline and 3% for diesel, Fig. 1 shows that total ethanol production slipped to about 11.15 billion liters (2.95 billion gallons) in 2001 with only about 5.14 billion liters (1.36 billion gallons) used as hydrous fuel as ethanol vehicles have been replaced with those that use anhydrous/gasoline mixtures.

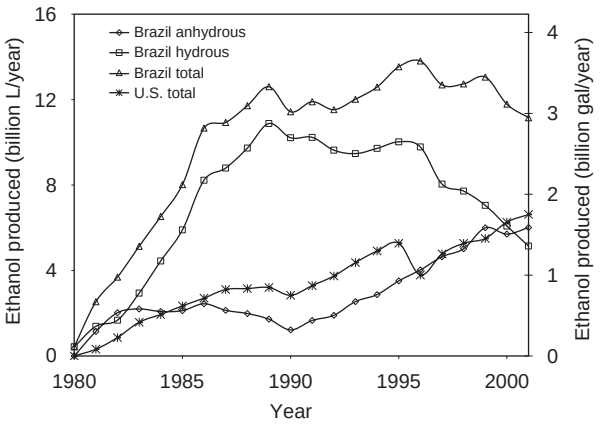


FIGURE 1 Annual production of fuel ethanol in Brazil and the United States since 1980. Ethanol is made from cane sugar in Brazil and from starch crops (mostly corn) in the United States. In Brazil, anhydrous ethanol is sold as 22% blends with gasoline, while hydrous ethanol, containing about 4.5% water, is used as a pure fuel. Almost all the fuel ethanol sold in the United States is used as a 10% blend in gasoline.

In the United States, fuel ethanol production grew from virtually nothing in 1980 to about 6.63 billion liters (1.75 billion gallons) by 2001, as shown in Fig. 1. Most of this ethanol is from corn starch and is used as 10% ethanol blends, reducing gasoline consumption and extending petroleum supplies. Ethanol also improves octane, decreasing the need for other octane boosters such as benzene, which are toxic and often carcinogenic, and oxygen in ethanol reduces emissions of unburned hydrocarbons and carbon monoxide, particularly for older vehicles that tend to burn rich.

Special vehicles are sold in the United States that can burn any fuel containing from 0 to 85% ethanol in gasoline, the latter being designated as E85, with 15% gasoline being used to promote cold starting. Only 172 of these vehicles were on the road in 1992, but the number rose to 48,022 in 2001, with 13,342 net new vehicles added that year. E85 engines consumed about 26.2 million liters (6.9 million gallons) of ethanol in 2001. These flexible fueled vehicles have in line sensors that measure ethanol levels and adjust the air-fuel ratio to maintain proper combustion. Although the compression ratio cannot be adjusted to fully capitalize on the favorable fuel properties of ethanol in high-level blends as with hydrous ethanol in Brazil, the technology ensures

fuel availability. Catalytic converters are needed to minimize tailpipe emissions of aldehydes, the primary pollutant resulting from incomplete ethanol combustion, and NO_x formation could increase for high-compression engines designed to maximize fuel utilization efficiency.

Ethanol has been used in compression ignition engines, but the lubrication properties and cetane of ethanol are not matched to conventional diesel technology. Thus, the engines must be modified or additives used to improve compatibility. On the other hand, adding ethanol to diesel fuel reduces particulate emissions, a key need, and formulations have been developed to stabilize dispersion of ethanol in diesel fuel.

1.3 Fuel Cells

Fuel cells are attractive because of low emissions and good efficiencies. Fuels cells react hydrogen with oxygen to make water at controlled conditions that minimize formation of pollutants such as NO_x and produce no unburned hydrocarbon emissions. In 2002 nearly every major automaker had large teams working on vehicle fuel cells, with most focused on operation with either gasoline or hydrogen. Because storing hydrogen onboard vehicles is challenging, using a liquid fuel such as ethanol as a hydrogen source is highly desirable but more complex to implement. Reformer technologies being developed for gasoline in fuel cells would generally be capable of using ethanol, but experience is limited on releasing hydrogen of acceptable quality from either one. Reforming methanol to hydrogen for fuel cells in buses and cars was developed and demonstrated in the United States, Germany, and Japan before environmental concerns about methanol limited this option. Reforming ethanol requires more steps, and this technology has been demonstrated in the laboratory but not onboard vehicles. Ethanol can also be partially oxidized to hydrogen, carbon monoxide, and carbon dioxide followed by reaction of carbon monoxide to hydrogen, and a partial oxidation reactor coupled to a shift reactor and a polymer electrolyte membrane (PEM) fuel cell followed by heat recovery achieved an efficiency of 43% with undetectable emissions of hydrocarbons, carbon monoxide, and nitrous oxides. Thus, ethanol fuel cells are very promising, but considerable research and development would enhance the readiness of this not widely studied technology.

2. ETHANOL PRODUCTION APPROACHES

Ethanol is manufactured by two primary paths: (1) fermentation of sugars derived from sugar, starch, or cellulosic materials or (2) reaction of ethylene with water. The former is favored for production of fuel. The latter has been used to make industrial grade ethanol for solvents, cosmetics, medicines, and so on, but purification of fermentation ethanol is displacing ethylene-derived ethanol for these applications. Some new processes are under development for fermentation or catalytic conversion of synthesis gas (carbon monoxide and hydrogen) from wood, natural gas, or coal, but synthesis gas from sustainable resources such as biomass is expensive. This article focuses on fermentation ethanol because it can potentially meet fuel cost and volume needs for large-scale fuel production, while ethylene is too expensive and supplies would likely be stretched.

3. PRODUCTION FROM SUGAR CROPS

3.1 Processing Sugar to Ethanol

Brazil produced about 18.5 million metric tonnes (20.4 tons) of processed sugar in the 2001/2002 harvest, with approximately 9.45 million metric tonnes (10.4 tons) used domestically and the rest exported. Brazilian sugar is mostly derived from sugarcane, a drought-tolerant tropical and subtropical crop containing about 12 to 17% sugars (90% sucrose, 10% glucose) and 68 to 72% moisture. Brazil grew about 272 million metric tonnes (300 million tons) of sugarcane in the 2001/2002 season, making it second to India in world cane production that year. In 1975, Brazil adopted a Pro-Alcool Program to convert sugar to ethanol to reduce dependence on petroleum imports that were damaging the economy, and Fig. 2 summarizes the sugar extraction and fermentation approach. In dedicated ethanol plants, cane juice is heated to about 105 to 110°C (221 to 230°F) to reduce microbial contamination and then fermented. In combined ethanol-sugar facilities, sucrose is removed by crystallization and the remaining molasses containing up to about 65% sugars is fermented. In either approach, sugar, $C_6H_{12}O_6$, is readily fermented to ethanol, C_2H_5OH, by common yeast such as *Saccharomyces cerevisiae* according to a reaction carried out for

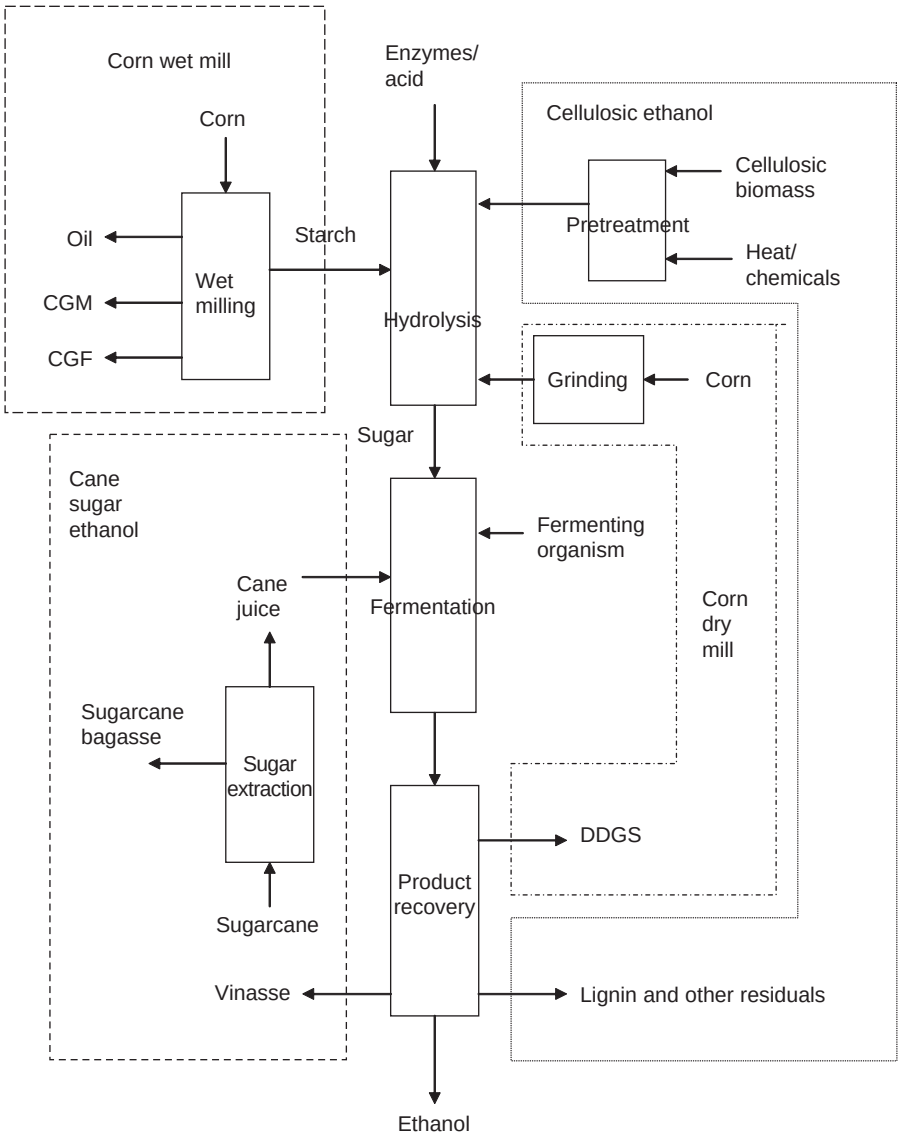


FIGURE 2 A simplified process flow diagram for production of ethanol from cane sugar, corn, and cellulosic biomass. All have similar fermentation and ethanol recovery operations but use different approaches to release sugars and generate different coproducts. Sugar can be directly extracted from sugarcane, and the residual bagasse is used as a boiler fuel to provide much of the energy for the extraction and ethanol production and recovery operations. In a corn dry mill, corn is ground, and enzymes and heat are added to hydrolyze starch to sugars for conversion to ethanol ,while the oil, protein, and fiber in corn are recovered after fermentation as an animal feed known as DDGS. Wet mills first fractionate corn to separate corn oil, corn gluten meal (CGM), and corn gluten feed (CGF) to capture value for food and animal feed, and the starch can then be hydrolyzed to sugars for fermentation to ethanol. For cellulosic biomass, heat and acids or enzymes hydrolyze the hemicellulose and cellulose portions to release sugars that can be fermented to ethanol, and the lignin and other remaining fractions can be burned to provide all the process heat and electricity for the conversion step with excess electricity left to export.

millennia:

$$C_6H_{12}O_6 \rightarrow 2C_2H_5OH + 2CO_2 \qquad (1)$$

in which CO_2 is carbon dioxide released during this process. About 70% of Brazilian ethanol facilities, often called distilleries, use batch operations in which sugar, nutrients, and other ingredients are added at the beginning of the fermentation or fed batch processes in which one or more of these ingredients are added as fermentation proceeds. However, continuous processes in which ingredients are constantly added and product removed from

fermentation vessels are also used. Yeast is typically concentrated and separated from the fermentation effluent, washed with sulfuric acid to reduce bacterial contamination, and recycled to fermentors, resulting in high cell densities that shorten fermentations to 6 to 10 h at temperatures of about 33 to 35°C with high ethanol yields (90 to 92% of theoretical) and concentrations (8 to 11% on a volume basis). According to Eq. (1), no more than 51.1 kg of ethanol can be obtained from fermentation of 100 kg of sugar, with yeast achieving about 90 to 92% of this. Thus, about 460 to 470 kg (920 to 940 pounds) or 580 to 590 liters (139 to 142 gallons) of ethanol result from a metric tonne (ton) of sugar. Following yeast removal, the fermentation broth containing ethanol, water, nutrients, and some yeast passes to distillation to make 95.5% "hydrous" ethanol, which can be taken to 99.6% "anhydrous" ethanol in a dehydration section, if desired. The remaining dissolved solids, called vinasse, are used as a fertilizer and in irrigation. About 5170 liters of ethanol could be obtained from a hectare of land in Brazil in 1996, and 13.9 billion liters (3.7 billion gallons) of ethanol in addition to 13.5 million metric tonnes (14.9 million tons) of sugar were made from 273 million metric tonnes (300 million tons) of sugarcane that year.

3.2 Costs

The costs of making ethanol in Brazil are estimated to be about \$0.006 to 0.007/liter (\$0.023 to 0.026/gallon) for labor, \$0.004 to 0.006/liter (\$0.015 to 0.023/gallon) for maintenance, \$0.002/liter (\$0.008/gallon) for chemicals, \$0.002 to 0.003/liter (\$0.008 to 0.011/gallon) for energy, \$0.004/liter (\$0.015/gallon) for other items, and \$0.127 to 0.134/liter (\$0.481 to 0.507/gallon) for sugar. Thus, the total cash cost of ethanol is on the order of \$0.167 to 0.185/liter (\$0.632 to 0.700/gallon). The overall capital investment is about \$0.52/annual liter (\$2.00/gallon) of capacity for a process that operates for 150 days/year making about 240,000 liters (900,000 gallons) of ethanol per day. Allowing another \$0.04 to 0.06/liter (\$0.15 to 0.23/gallon) for interest on capital, capital recovery, and other fixed costs, the total cost including return on capital as applied in Brazil is about \$0.21 to 0.25/liter (\$0.79 to 0.95/gallon) at the plant gate without transportation, distribution, taxes, and other costs. More on-stream time would greatly reduce the unit capital cost charged to ethanol. Hydrous costs about 7% less than anhydrous ethanol because less energy and capital are needed to meet final product specifications. Costs in Brazil dropped at about a 30% rate with each doubling of experience through 1990 and continue to decline at about a 10% rate due to learning curve effects resulting from commercial experience. Opportunities have been identified to reduce costs by about 23%, mostly by reducing the intensity of growing cane but by also improving conversion operations.

3.3 Energy Balance

About 197 to 222 mega joules of fossil energy are used for providing seeds, labor, power for agricultural operations, agricultural equipment, fertilizers, lime, herbicides, insecticides, and cane transportation to grow a metric tonne of sugarcane (169,000 to 191,000 Btu/ton). However, much of the energy for sugar extraction and fermentation is provided by burning the bagasse left after processing, but some fossil energy is used for electricity generation, chemicals and lubricants manufacture, erection of buildings, and manufacture of equipment, a total of about 40 to 70 mega joules/metric tonne (34,000 to 60,000 Btu/ton) of cane. Overall, about 237 to 292 mega joules of fossil energy are consumed (204,000 to 251,000 Btu/ton) compared to 1707 to 1941 mega joules of energy in the ethanol from a metric tonne (1.470×10^6 to 1.669×10^6 Btu/ton). Based on these values, the energy balance ratio of ethanol energy output to fossil fuel input is between about 5.9 to 8.2, not including the 175 to 328 mega joules/metric tonne of cane (150,000 to 282,000 Btu/ton) left from burning bagasse after sugar has been extracted, a very favorable energy balance.

3.4 Greenhouse Gas Emissions

In simple terms, accumulation of methane, carbon dioxide, nitrous oxide, and other gases that trap infrared radiation are predicted to cause global climate change, and carbon dioxide has the greatest impact because of the huge quantities released when fossil fuels are burned. Burning coal, petroleum, and natural gas, in order of decreasing impact, release previously trapped carbon into the atmosphere, resulting in a net accumulation of carbon dioxide. On the other hand, if fuels are made with renewable biomass such as wood, carbon dioxide released during processing and burning can be recaptured by photosynthesis when more biomass is grown to replace that harvested. While biomass is carbon dioxide neutral, fossil inputs to plant, grow, harvest,

transport, convert, and distribute biomass and its products contribute to greenhouse gas accumulation and must be considered.

Burning bagasse can provide 100% of the thermal energy and 92% of the electricity for processing of sugar to ethanol, and less than about 8% comes from fossil sources. Based on Brazilian experience that 1.2 liters of ethanol can replace 1.0 liters of gasoline, the net contribution to carbon dioxide accumulation by vehicles running on hydrous ethanol is less than about 10% of gasoline, and substituting surplus bagasse or other renewables for fossil fuel to generate electricity could improve the balance further. However, intensive sugarcane agricultural practices raise serious concerns about its sustainability, and low and no-till and other practices will be needed for cane sugar to become more viable.

4. PRODUCTION FROM STARCH CROPS

4.1 Starch Processing

Corn grain contains about 70% starch, 10 to 11% crude protein, 4.5 to 6.0% oil, 6% hemicellulose, 2 to 3% cellulose, 1% lignin, and 1% ash and is the dominant choice for current fuel ethanol production by dry or wet milling operations in the United States. In a typical dry mill, grain is milled to a powder, heated with water addition to about 85°C, mixed with alpha-amylase enzyme, held for up to an hour, heated further to 110 to 150°C to liquefy the starch and reduce bacteria levels, cooled back to about 85°C, and held with more alpha amylase for about 1 h. Then, the stream is cooled further, and glucoamylase enzyme added to complete conversion to sugars known as dextrose. This overall saccharification operation occurs by the hydrolysis reaction:

$$(C_6H_{10}O_5)_n + nH_2O \rightarrow nC_6H_{12}O_6 \qquad (2)$$

in which $C_6H_{12}O_6$ is a glucose sugar molecule formed when the alpha bonds linking n units of $C_6H_{10}O_5$ in long chains of starch are broken and combined with n molecules of water, H_2O. Yeast then ferment glucose to ethanol and carbon dioxide, just as in Eq. (1). The fermentation can be in batch operations held until the fermentation is complete in about 48 h or by continuous processes with ongoing sugar addition and takeoff of fermented broth known as beer. The beer containing about 12 to 14% or more ethanol is typically distilled to around 95% purity, close to the azeotropic composition at which further enrichment by ordinary distillation ceases, and this mixture is dehydrated with molecular sieves to anhydrous ethanol. The solids containing the remaining protein, oil, and fiber are dried to a 27% protein product known as distillers dried grains with solubles (DDGS) or just distillers dried grains (DDG), depending on whether process syrup is combined with the solids or not, and used in animal feeds. The carbon dioxide can be sold for carbonated beverages, dry ice, and other applications. Overall, about 460 to 490 liters (110 to 118 gallons) of ethanol weighing 365 to 390 kg (730 to 780 lbs), 357 to 335 kg (714 to 670 lbs) of DDGS, and 350 to 370 kg (700 to 740 lbs) of carbon dioxide result from each metric tonne (per ton in parentheses) of corn processed—close to equal masses of each coproduct. Figure 2 presents a simplified block flow diagram for dry milling.

Although the greatest number of facilities are dry mills run by smaller producers, more ethanol is made in larger wet milling operations generally owned by larger companies. In wet mills, corn is first steeped and milled to separate the starch, protein, fiber, and germ, again as summarized in Fig. 2. Corn oil is removed from the germ for food, and the remaining germ meal is combined with the fiber and water from steeping as a 21% protein animal feed known as corn gluten feed. A portion of the corn protein is sold as corn gluten meal, a 60% protein product added to chicken and other animal feeds. Starch can be used directly for industrial uses and food or broken down by enzymes as for dry milling to make dextrose that can be enzymatically converted to high fructose corn syrup for a beverage sweetener or fermented to ethanol. One metric tonne (1 ton in parentheses) of corn can yield approximately 34 kg (67 lbs) of oil, 52 kg (105 lbs) of corn gluten meal, 283 kg (567 lbs) of corn gluten feed, 470 liters (112 gallons) or 370 kg (740 lbs) of ethanol, and 354 kg (708 lbs) of carbon dioxide. Again, all corn components are utilized, with protein sold for animal feed, the major market for corn.

4.2 Costs

Although ethanol costs tend to be site specific and business sensitive, recent surveys provide a useful picture. The average cost of all feedstocks for dry and wet mills was about $0.23/liter ($0.87/gallon) of ethanol with dry mills varying from $0.25/liter ($0.94/gallon) for small plants to $0.21/liter ($0.79/gallon) for medium plants and $0.22/liter ($0.82/gallon) for large facilities and wet mills being about $0.024/liter ($0.09/gallon) greater than the overall

average, apparently because of the greater quantities of corn required. The multiple wet mill coproducts provided revenue of $0.11/liter ($0.43/gallon) of ethanol versus about $0.07/liter ($0.28/gallon) for dry mills. The difference between the cost of corn and coproduct revenues, termed the net cost of corn, is about $0.14/liter ($0.53/gallon) for dry mills and $0.13/liter ($0.48/gallon) for wet mills but varied between $0.095/liter ($0.36/gallon) and $0.24/liter ($0.91/gallon) for wet mills and $0.12/liter ($0.44/gallon) and $0.25/liter ($0.95/gallon) for dry mills over the ten years ending in 2000.

Other costs are also incurred. Averages for electricity and fuel were about $0.030/liter ($0.112/gallon) for wet mills and $0.035/liter ($0.131/gallon) for dry mills for an overall weighted average of $0.031/liter ($0.118/gallon). Operating labor, repairs, and maintenance were lower for dry mills at about $0.029/liter ($0.109/gallon) compared to wet mills at about $0.033/liter ($0.124/gallon). Yeast, enzymes, chemicals, and other ingredients contributed another $0.024/liter ($0.090/gallon) for dry mills and $0.030/liter ($0.114/gallon) for wet mills. Averages for management, administration, insurance, and taxes were similar at $0.009/liter ($0.037/gallon) for dry mills and $0.010/liter ($0.038/gallon) for wet mills. All other costs averaged about $0.019/liter ($0.072/gallon) for wet and $0.013/liter ($0.051/gallon) for dry mills. Thus, total cash costs for other than corn were about $0.12/liter ($0.46/gallon) for wet mills and $0.11/liter ($0.42/gallon) for wet mills, and combined with net feedstock costs, the average cost is about 0.25/liter ($0.94/gallon), with a less than 1% advantage to wet mills. This gate cost does not include transportation, distribution, taxes, and other costs to the consumer.

The difference between ethanol revenues and cash costs of production must be weighed against capital costs to determine return on investment. Capital costs are estimated to be between $0.283/annual liter ($1.07/annual gallon) to $0.631/annual liter of capacity ($2.39/gallon), varying with facility type, size, and technology. Wet mills tend to cost more than dry mills, and larger dry mills generally cost less than smaller ones per volume of ethanol made because of economies of scale. Capital and operating costs have dropped over time through learning curve effects and research.

Several considerations apply to starch ethanol costs, many of which apply to ethanol from sugar. Costs and coproduct revenues vary with location and economic climate, and many costs components do not always work together. For example, the value of ethanol usually drops with gasoline prices. In addition, higher corn or sugar prices or softer coproduct markets increase the net feedstock cost. In this vein, increasing corn ethanol production would eventually drive up corn prices and drive down coproduct selling prices, making further capacity increases uneconomic. On the other hand, technology improvements enhance profitability.

Tax incentives in the United States promote blending with gasoline and were essential to initiating ethanol production in the late 1970s and early 1980s in response to petroleum shortages and price spikes. Over time, this excise tax exclusion dropped to $0.0140/liter ($0.053/gallon) of 10% blended fuel, reducing ethanol costs by $0.140/liter ($0.53/gallon) of ethanol used. Alternatively, the tax on blended gasoline is reduced by $0.0110/liter ($0.0416/gallon) for 7.7% ethanol and $0.0081 ($0.0308/gallon) for gasoline containing 5.7% ethanol. This exemption is not paid to farmers but is an avoided tax for blenders of ethanol from any form of biomass with gasoline. Blenders could apply an income tax credit instead of the excise tax exemption. The exemption is scheduled to be further cut to $0.0137/liter ($0.052/gallon) of 10% blended fuel in 2003, $0.0135/liter ($0.051/gallon) in 2005, and eliminated on December 31, 2007. Many states, particularly in the Midwest, also provide incentives to encourage ethanol use. Although subject to controversy, tax and other assistance policies are vital for introducing new technology, particularly for commodity markets, and have supported emergence of new industries such as the large petrochemical sector that grew out of World War II.

4.3 Energy Balance

Incorrect and outdated data are periodically used to spur controversy about whether making corn ethanol consumes more or less energy than contained in the product. However, up-to-date facts show continual improvements, with reductions in energy demand as the technology matures resulting in favorable energy benefits. Accounting for the variability of corn productivity over time and with location, an input of 2789.3 mega joules/dry metric tonne (57,476 Btu/bushel or 2,415,100 Btu/dry ton) of corn was determined for all the energy required to grow and harvest corn and haul it to storage. The energy to transport corn from storage to ethanol production was estimated to be about 292.3 mega joules/dry metric tonne (6020 Btu/bushel or 253,000 Btu/dry ton) or 690.6 J/liter (2.263 Btu/gallon) of ethanol made. Recent surveys showed that about 13.594 mega

joules of heat and electricity are needed to make one liter (48,772 Btu/gallon) of ethanol via dry mills and about 15.118 mega joules/liter (54,239 Btu/gallon) for wet mills. However, new plants generally achieve on the order of 11 mega joules/liter (40,000 Btu/gallon) or less. The survey showed about 460 and 467 liter of ethanol/dry metric tonne (2.64 and 2.68 gallons/bushel or 111 and 113 gallons/dry ton) were derived from corn in dry and wet mills, respectively. Finally, about 442.6 kilojoules/liter (1,588 Btu/gallon) of ethanol was required to distribute ethanol to end users. Combining these estimates, dry and wet mills required about 20.75 and 22.16 mega joules/liter (74,447 and 79,503 Btu/gallon) of ethanol, respectively, with the average weighted by number of facilities being about 21.525 mega joules/liter (77,228 Btu/gallon).

Some energy should be assigned to coproducts from corn milling, with distribution by energy inputs, energy content, relative market values, relative weight, and the amount of energy needed to replace them analyzed. However, the latter based on the energy to produce equivalent feed and food as from coproducts is often favored, with growing soybeans an appropriate alternative. On that basis, the final estimate is a net energy displacement of about 6.307, 5.369, and 5.883 mega joules/liter (22,629, 19,262, and 21,105 Btu/gallon) for dry, wet, and weighted average facilities, respectively, when ethanol is made from corn compared to the higher heating value for ethanol of 23.402 mega joules/liter (83,961 Btu/gallon).

This analysis means that one volume of ethanol generates about 37%, 30%, and 34% more energy than needed to manufacture it by dry, wet, and the overall weighted average of production facilities, respectively. Furthermore, because about 6.34 units of liquid fuel energy result from 1.0 energy unit of petroleum, corn ethanol provides even greater advantages for displacing imported liquid fuels, a major concern for economic and political stability in the United States.

4.4 Greenhouse Gas Emissions

The positive energy balances suggest that corn ethanol has a lower impact on global climate change than gasoline. Again, if the greenhouse gas emissions that would result if soybeans were grown to meet markets served by corn ethanol coproducts are subtracted, burning ethanol from starch crops as a mixture of 85% ethanol with 15% gasoline (E85) reduces carbon dioxide emissions by about 40% and emissions of all greenhouse gases including methane and nitrous oxides by about 20 to 30% relative to pure gasoline. Because a significant portion of greenhouse gases comes from the gasoline in E85, greater reductions would be possible if technology were developed to use hydrous ethanol as in Brazil. If corn stover or other renewable energy sources were substituted for fossil fuels to power dry and wet mills as in Brazil, the greenhouse gas benefits of corn ethanol would be better yet.

5. PRODUCTION FROM CELLULOSIC BIOMASS

5.1 Cellulosic Biomass

Ethanol can also be made from sugars contained in the structural portion of plants. In particular, about 40 to 50% is cellulose—up to about 10,000 glucose molecules joined by beta 1-4 linkages as long linear molecules that can align parallel to one another and join by hydrogen bonding in crystalline regions that typically contain much of the cellulose. Another sugar polymer known as hemicellulose represents about 25 to 30% of the dry plant mass; unlike cellulose, hemicellulose is amorphous. Hemicellulose is typically made up of the five sugars arabinose, galactose, glucose, mannose, and xylose, along with lower amounts of other compounds including hexuronic acids and methoxyl, acetyl, and free carboxyl groups. Lignin represents about another 15 to 20% by dry weight and is a complex phenylpropanoic acid polymer. Lignin and hemicellulose are joined by covalent bonds and cement cellulose fibers into microfibril bundles. Plants also contain lesser amounts of minerals, free sugars, proteins, and other components that vary with species. Together, this composite of cellulose, hemicellulose, lignin, and other generally lesser ingredients is called lignocellulosic, or just cellulosic, biomass and has great strength to support the aerial portion of plants.

Cellulosic biomass is abundant with prominent examples including agricultural residues (e.g., corn stover, sugarcane bagasse), forestry residues (e.g., forest thinnings), industrial wastes (e.g., paper sludge), major portions of municipal solid waste (e.g., waste paper), herbaceous plants (e.g., switchgrass), and woody species (e.g., poplar). In addition to its abundance, cellulosic biomass can be grown with low energy inputs and is comparable in cost to

petroleum. For example, biomass costing about $44/dry metric tonne ($40/ton) is competitive with petroleum at about $6.00/barrel based on equal mass or about $12.70/barrel based on equivalent energy content. Some cellulosic wastes such as rice straw can be obtained for lower prices, and in some cases, such as paper sludge, users would be paid to dispose of them (i.e., tipping fee). Waste conversion to ethanol can also help solve serious disposal problems.

5.2 Processing of Cellulosic Biomass

As outlined in Fig. 2, the overall process for making cellulosic ethanol is conceptually similar to processing corn but with distinctions that have important energy balance and greenhouse gas emission implications. Sulfuric and other acids can convert both hemicellulose and cellulose into fermentable sugars. Alternatively, enzymes known as cellulase hydrolyze cellulose to glucose while a collection of enzymes loosely termed hemicellulase could breakdown hemicellulose. The remaining lignin cannot be converted to sugars but can provide boiler fuel for process heat and electricity or be reacted into coproducts.

Biological conversion of cellulose to glucose by cellulase and fermentation of glucose to ethanol realize high yields vital to success, and the advances of the evolving field of modern biotechnology promise significantly lower costs. In this case, mills or other devices reduce the feed material size, facilitating transfer of heat and chemicals later, and the material is pretreated to expose cellulose to cellulase to realize high glucose yields. Currently, pretreatment is often by dilute acid hemicellulose hydrolysis because good yields of sugars are realized from hemicellulose during pretreatment and from cellulose in the subsequent enzymatic hydrolysis and acid costs are relatively low. In this approach, cellulosics are reacted in 0.5 to 1.0% sulfuric acid at about 160 to 190°C for approximately 10 minutes, releasing most hemicellulose sugars by a hydrolysis reaction similar to Eq. (2) but including up to five different sugars as well as other compounds in the cellulosic matrix. The contents are flash cooled to stop the reaction as the temperature quickly drops to about 100°C and also to remove some acetic acid, furfural, and other fermentation inhibitors that are released from biomass or produced by degradation reactions during pretreatment. Next, the solid and liquid fractions are separated and ion exchange employed in some designs to remove a portion of acetic acid and much of the sulfuric acid from the liquid. Lime is added to neutralize the remaining sulfuric acid and then increase the pH to about 10 in an "overliming" operation that removes toxics to downstream biological steps.

A small portion of the solids alone or including treated liquid is fed to make cellulase by an organism such as the fungus *Trichoderma reesei*. The effluent is then combined with the bulk of the pretreated solid to hydrolyze cellulose to glucose, and an organism such as yeast is added to convert the glucose to ethanol soon after release in a simultaneous saccharification and fermentation or SSF approach. Because SSF rapidly converts glucose and the two glucose molecule cellobiose to ethanol, the concentration of both of these powerful inhibitors of cellulase activity are reduced. Thus, the rates, yields, and concentrations for ethanol production are all improved via the SSF approach compared to sequential hydrolysis and fermentation even though the temperature has to be reduced from optimum levels for hydrolysis alone to accommodate less temperature tolerant yeast. The SSF configuration also eliminates about half of the fermentation vessels, cutting costs, and ethanol accumulation reduces the likelihood of successful invasion by unwanted organisms that thrive if sugar accumulates. Alternatively, conditioned liquid hemicellulose hydrolyzate can be added to SSF in an approach labeled simultaneous saccharification and cofermentation (SSCF) if the organism used ferments hemicellulose sugars and glucose to ethanol. Of course, hemicellulose sugar fermentation can be performed in a separate vessel by such an organism, albeit at the expense of more equipment and somewhat higher costs.

Distillation concentrates approximately 5% fermentation beer streams to about 95% ethanol, and this azeotropic mixture is often passed to molecular sieves to recover a nearly 100% ethanol product suitable for blending with gasoline or use as a pure anhydrous fuel. For use as a pure fuel without gasoline addition, the 5% water might be left in a la Brazilian experience. The solids and solubles remaining following SSF/SSCF pass out in the distillation column bottoms, are concentrated, and then are burned to generate steam that can provide all of the process heat and electricity with some excess electricity left to export. Water from the system is treated by anaerobic digestion, and the resulting methane is burned for steam generation.

5.3 Technical and Economic Progress

The United States and other counties have funded research to improve cellulosic biomass conversion to

ethanol, and despite modest and sporadic funding, the total estimated cost at the plant gate including capital charges dropped from about $1.05 to 1.32/liter ($4.00 to 5.00/gallon) of ethanol in 1979 to approximately $0.29 to 0.32/liter ($1.10 to 1.20/gallon) now, a value competitive with corn ethanol. For a facility making about 197.6 million liters/year (52.2 million gallons/year) of ethanol, the capital costs were projected to be about $234 million, the feedstock cost about $19.3 million/year, the other variable operating costs about $5.7 million/year, and the fixed operating costs about $7.5 million/year. Such cost projections depend strongly on location, estimation methods, available data, key performance parameter assumptions, technology applied, and financing arrangements and vary considerably. Furthermore, although corn and sugar conversion costs are based on commercial experience, no cellulosic facilities are yet operational. Thus, while useful for guiding research, cellulosic ethanol costs are approximate indicators of actual costs. However, the fact that several firms are working to commercialize the first cellulosic ethanol facilities confirms the economic merits, and their biggest challenge is financing the substantial capital cost for first-of-a-kind technology in light of its perceived risk.

Progress on two major fronts has fed the cost reductions outlined. First, important improvements have been made in overcoming the recalcitrance of biomass. Dilute acid pretreatment has been optimized to increase sugar yields from both the hemicellulose and cellulose fractions, and better cellulase formulations have achieved higher yields with lower doses. SSF improved economics substantially, with enzymes or organisms that utilize cellobiose, a polymer of two glucose molecules, proving particularly important by reducing the concentration of this powerful cellulase inhibitor.

The second hurdle, overcoming the diversity of sugars, was a major breakthrough vital to economic viability. Native organisms cannot ferment all five sugars typically contained in cellulosic biomass to ethanol with high yields, a necessity to competitiveness. However, in the late 1980s, researchers genetically engineered bacteria to make ethanol from all the sugars at close to theoretical yields by inserting two genes from the bacteria *Zymomonas mobilis* into other host bacteria to produce the enzymes pyruvate decarboxylase and alcohol dehydrogenase that transform pyruvate, an intercellular compound in the host, to ethanol instead of the products ordinarily made. This accomplishment increases revenue while reducing costs for disposal of unused sugars.

5.4 Energy Balance

Burning solids and concentrated solubles after distillation, solids from aerobic treatment, and methane from anaerobic treatment can generate enough steam to meet all process heat and power demands. In addition, about 10.9 MW of surplus power is left for sale for a currently envisioned ethanol plant fed 2000 metric tonnes/day (2,205 tons/day) of cellulosic biomass, resulting in about 0.4 energy units of ethanol and close to 0.1 energy units of electricity per unit of energy in the biomass feed. The amount of ethanol will improve as yields are increased through learning curve effects and better technology, and power exports to the grid will increase as (1) cellulosic ethanol technology improves to use less heat and power and (2) more efficient cycles are used for making electricity from steam. The key is that no fossil fuel is needed to convert cellulosic biomass into ethanol, and significant amounts of excess electricity are left for sale into a market compatible with large-scale fuel production.

It has been estimated that about 6 to 15% of the ethanol fuel value is required to grow biomass, another 1 to 4% to transport it to the conversion facility, and 1 to 5% for fuel distribution. If all of these energy needs are met from fossil sources, about 4.0 to 6.7 times as much ethanol energy is released as fossil fuel used for the overall cellulosic ethanol process, but the ratio increases to about 4.3 to 7.2 if the thermal value of exported electricity is included. More comprehensive analyses project that current cellulosic ethanol technology provides a net energy output from about 16.7 to 21.2 mega joules/liter (60,000 to 76,000 Btu/gallon) of ethanol compared to its lower heating value of 21.2 mega joules/liter (76,000 Btu/gallon), depending on the feedstock employed. Overall, these low fossil inputs result from the low-energy demands to supply cellulosics, generation of all heat and power for conversion from lignin and other biomass residues, and to a lesser extent, export of excess power.

Given its favorable energy balance, cellulosic ethanol could be substituted for gasoline or diesel fuel to harvest and transport cellulosics to the process and distribute ethanol to market. In addition, methane produced by anaerobic digestion of process residues could be converted into ammonia to reduce fossil fuel use for growing biomass. The result would be that virtually no fossil fuels are needed to grow, convert, and transport ethanol when it is made from cellulosics while both fuel and power are produced, and the ratio of energy outputs to fossil energy inputs

would become very large. Advances in ethanol production and power generation would improve this picture.

A recent study showed that fossil fuel inputs for cellulosic ethanol were lower than for a large number of options including gasoline, other petroleum-based fuels, natural gas–based fuels, and electrolysis hydrogen. Yet the total energy required to make cellulosic ethanol was toward the high end. This apparent contradiction is readily explained because burning lignin and other residues totally meets the large process energy demand without fossil fuels. The fact that most of the more efficient technologies in the study are practiced commercially while cellulosic ethanol is not suggests that significant improvements will evolve for cellulosic ethanol through the commercial learning curve, and the likely outcome is that the energy balances will improve dramatically, increasing the energy displacement benefits far more.

5.5 Greenhouse Gas Emissions

As shown for ethanol from sugar and corn, carbon dioxide generation by burning fossil fuels has a large impact on greenhouse gas emissions, and the low fossil energy inputs for cellulosic ethanol production are expected to have important implications for global climate change. In fact, studies have shown that using 85% cellulosic ethanol with gasoline, E85, reduces GHG emissions by about 65% compared to straight gasoline, and that running E95 in vehicles would decrease GHG emissions by about 75 to 78%. A large part of these emissions comes from the gasoline, and even greater reductions in GHGs of about 85% or more are possible if pure ethanol were used. Advances in both ethanol production and power-generation technologies should improve yields, reduce energy inputs, and generate more electricity, reducing GHG emissions even more.

5.6 Some Other Technologies for Processing Cellulosic Biomass

A favorable pretreatment must produce highly digestible cellulose, recover hemicellulose sugars with high yields, have a low capital cost, require limited energy inputs, minimize degradation, use low-cost chemicals, minimize problems with chemical additives, and integrate well with the overall process. Many studies with physical pretreatments such as size reduction and radiation and biological pretreatments with various fungi have been unsuccessful, and only a few chemical pretreatments other than dilute sulfuric acid currently promise to be cost-effective. In the ammonia fiber explosion (AFEX) process, ammonia added to biomass reduces cellulose crystallinity, opens up the structure for enzymes, can be mostly recovered and recycled, provides fermentation nutrients, and produces a readily fermentable hydrolyzate without post treatment. A near neutral pH approach favors soluble oligomer release from hemicellulose to keep formation of sugar monomers that can degrade low. Ammonia circulation through biomass removes lignin, improving cellulose hydrolysis yields. Lime also removes lignin, but the reaction is slow. Dilute nitric acid reduces materials of construction costs compared to sulfuric but is itself more expensive. More research is needed to better understand the performance and costs of these and other options that promise to meet the demanding criteria for effective pretreatment.

Acids can be used to breakdown cellulose. Concentrated acids realize high yields vital to economic success, but the cost of required acid recovery operations must be significantly reduced. Substantial experience exists for dilute acid hydrolysis, but yields are not high unless practical countercurrent flow configurations can be developed.

5.7 Opportunities for Further Improvements

The pretreatment step and combined biological operations for cellulase production, cellulose hydrolysis, and sugar fermentation are the most expensive in the currently defined process. *Trichoderma reesei* and other fungi generally used for manufacturing cellulase arose out of World War II research to combat decay of cotton gear in the tropics. Commercial cellulase is estimated to cost over $0.26/liter ($1.00/gallon) of ethanol produced, and making it on site is estimated to cost about $0.092/liter ($0.35/gallon). Although now expensive, cellulase offers near theoretical glucose yields from cellulose, and the evolving tools of modern biotechnology promise reduced costs by improving its effectiveness in hydrolyzing cellulose and increasing the rate of making enzyme. Better pretreatment could also reduce cellulase use, and development of continuous processes could cut equipment idle time and eliminate expensive propagation systems.

Although subsidies are needed for ethanol at current costs to compete for blending with gasoline, the cost of cellulosic biomass is competitive with

petroleum, and ethanol promises to be competitive as a pure fuel without subsidies if yields are increased, energy used more efficiently, operating costs reduced, and capital costs brought in line with those for corn processing. Less energy for size reduction, lower cost materials of construction, lower chemical use, and reduced process energy while maintaining or enhancing yields are all important pretreatment targets. Combining the operations for cellulase production, cellulose hydrolysis, and fermentation in a continuous operation now referred to as Consolidated BioProcessing (CBP) promises to dramatically reduce operating and capital costs for biological steps. Together such improvements would drop total cash and capital recovery costs to about $0.13/liter ($0.50/gallon) based on performance judged to be achievable for mature technology or $0.09/liter ($0.34/gallon) if near theoretical performance results. Thus, research targeted at advanced pretreatment and biological process integration could lead to a cost-competitive sustainable fuel with major benefits. Alternative uses for lignin, minerals, organic acids, and other components in cellulosic feedstocks and chemicals made from the low-cost sugars could increase profitability.

Enhanced cellulosic crop productivity will reduce land requirements for a given fuel production, promote economies of scale, and reduce feedstock costs. Greater impact for a given land use can be achieved with more efficient vehicles (e.g., fuel cells, hybrid vehicles). Because feed for livestock is grown on most agricultural land, recovering protein from fast growing crops such as switchgrass also used to produce ethanol could support making both fuel and food.

Several steps in cellulosic conversion are well positioned for commercial applications. Recombinant organisms are effective now for conversion of all five sugars in cellulosic biomass to ethanol. Although erroneous information may suggest otherwise, ethanol purification is not expensive or energy intensive. Rather, attention should be focused on significantly advancing pretreatment and cellulose conversion to make cellulosic ethanol costs competitive.

6. DIRECTIONS AND OPPORTUNITIES FOR WORLD ETHANOL PRODUCTION

Brazil is the leading producer of fuel ethanol in the world and is increasing the ethanol in blends from 22 to 26% and promoting 3% ethanol in diesel fuel. However, replacement of ethanol vehicles with vehicles that burn low level blends will cut ethanol consumption unless large-scale purchase of hydrous ethanol fueled vehicles resumes.

Ethanol use is growing in the United States. The Clean Air Act Amendments of 1990 require 2% oxygen in gasoline to reduce vehicle tailpipe emissions of unburned hydrocarbons and carbon monoxide, and about 16.0 billion liters (4.2 billion gallons) of methyl tertiary butyl ether (MTBE) and 6.63 billion liters (1.75 billion gallons) of ethanol were used as oxygenates in 2001 to meet these requirements. Because MTBE works its way into groundwater from underground storage tanks, is a persistent contaminant, has a turpentine-like odor and solvent-like taste, and is a suspected carcinogen, U.S. legislation mandates phase out of MTBE. Ethanol does not have these problematic features and is the leading alternative. A renewable fuels standard (RFS) that bans MTBE, eliminates the oxygen requirement in gasoline, enhances clean air rules, and establishes a minimum renewable fuels standard that would grow to 5 billion gallons by 2012 appears likely. Passage of this or related legislation could triple ethanol use after 10 years and result in the United States overtaking Brazil as the world's largest producer of fuel ethanol. Enhanced ethanol production would reduce petroleum imports, the trade deficit, and greenhouse gas emissions, and would create domestic employment.

Other countries have begun or are about to begin ethanol production. Canada produced about 238 million liters (62.9 million gallons) of ethanol in 2001, with plans to more than double that output soon. Canada provides a federal excise tax exemption of $0.10/liter ($0.38/gallon) for the ethanol portion of blended fuels. Although most Canadian ethanol is from wheat and corn starch, Canada produces about 17 million liters (4.5 million gallons) of ethanol from a wood processing waste, spent sulphite liquor, but most of it is sold for industrial uses.

France has been the largest ethanol producing country in Europe, making about 115 million liters (30 million gallons) per year in 2000, mostly from beet sugar for conversion to ethyl tertiary butyl ether (ETBE) for use as an oxygenate in gasoline. However, Spain now has two grain plants with a total capacity of about 325 million liters (86 million gallons) per year, and a third plant planned to reach 526 million liters (139 million gallons) in capacity. Sweden made about 25 million liters (6.7 million gallons) of ethanol in 2000. Recently, the European Commission encouraged increased use of ethanol and biodiesel for

transportation with the goal of substituting 20% of diesel and gasoline by alternative fuels that reduce greenhouse gas emissions by 2020.

China planned to introduce 380 million liters (100 million gallons) of ethanol production capacity in 2001 and achieve a total annual ethanol production capacity of about 2.0 billion liters (534 million gallons) by late 2002. Mostly grain is used to relieve structural surpluses while also reducing oil imports, and blending ethanol will help address air pollution problems in major cities. India has a major program in progress to use some of their large sugar stocks to make ethanol for 5 and 10% blends with gasoline and plans for more than 30 facilities making up to about 380 million liters (100 million gallons) of fuel ethanol annually by the end of 2002. In addition, Thailand is pursuing production of ethanol from cane sugar and tapioca, corn, and rice starch with a near term goal of achieving an annual capacity of about 500 to 730 million liters (130 to 190 million gallons).

This expansion of ethanol production creates an immense opportunity for ethanol from cellulosic biomass. Cellulosic ethanol processes can be the low-cash-cost producer of ethanol, and the transition to cellulosics will become even more attractive as large-scale ethanol production from grains puts upward pressure on grain prices and likely reduces the selling price of coproducts. However, the perceived risk and higher capital costs of first-of-a-kind facilities present serious challenges to implementing this important technology. Advanced technologies that realize the low-cost potential of cellulosic feedstocks could lower capital investments and improve returns but delay implementation. Coupling development of advanced technologies with a commercial learning curve effect would accelerate the emergence of cellulosic ethanol as a competitive pure fuel without subsidies.

7. CONCLUSIONS

Considerable experience has been gained with low-level ethanol blends (10 and 22% ethanol in gasoline) and nearly pure (95.5% ethanol with the rest water and 85% with gasoline) in spark-ignition engines in Brazil and the United States that demonstrates its excellent fuel properties, and ethanol use is spreading. Ethanol can also be added to diesel fuel with additives and could undoubtedly provide hydrogen for fuel cells. Environmental attributes such as biodegradability and low toxicity avoid many problems with gasoline and MTBE.

When made from domestic biomass, ethanol provides an important alternative for a transportation sector almost totally dependent on economically, politically, and strategically vulnerable petroleum imports that continually climb. Furthermore, use of such resources can dramatically reduce emissions of greenhouse gases that trap radiation and cause global climate change.

Brazil led ethanol use as a pure fuel and for blending with gasoline, and a substantial fleet of dedicated ethanol vehicles evolved. However, world sugar demands and the phase down of incentives impacted pricing and availability, and dedicated vehicle production virtually stopped, hurting ethanol consumption. Because residues are burned to provide process heat and electricity, limited fossil fuel is needed to convert cane sugar into ethanol, and net emissions of carbon dioxide are much less than for gasoline. The cost of Brazilian cane sugar ethanol has dropped significantly through learning curve effects, but sugar prices and cultivation practices likely limit competitiveness with gasoline for large-scale uses.

In the United States, ethanol is produced from starch from corn and other crops, and the remaining components are used for feed and food to realize value from all of the grain. Energy consumption has dropped dramatically over time via technology advances and commercial learning curve effects, and corn ethanol now generates about one-third more energy than used to make it. Also, corn ethanol replaces about six times its energy content in petroleum, substantially reducing oil imports. Overall, the result is a reduction of about 40% in carbon dioxide emissions and 20 to 30% in total greenhouse gas emissions for E85 fuel relative to burning just gasoline with even greater reductions possible for pure corn ethanol use and powering the process by renewable resources. Further cost reductions would allow greater impact without tax incentives but become challenging if corn costs rise and coproduct prices drop as corn ethanol production expands.

Cellulosic biomass is abundant and competitive in price with petroleum. Costs for processing this recalcitrant, heterogeneous feedstock to ethanol have been reduced considerably, and companies are striving to introduce cellulosic technology commercially. Because burning lignin and other residues can provide the energy for conversion with electricity left for export and fossil energy requirements to provide biomass feed and distribute ethanol are low, at least four to more than seven times as much energy is contained in products than fossil energy inputs, and cellulosic ethanol would have less than 10 to

15% of the greenhouse gas emissions of gasoline for current technology. Once commercialized, learning-curve effects are expected to dramatically improve costs and environmental benefits, making it vital to successfully introduce cellulosic ethanol technologies soon. Costs and environmental attributes would further benefit through development of advanced technologies that reduce total production costs to compete with gasoline without subsidies and novel technologies that recover protein and other valuable cellulosic components. Ultimately, high productivity cellulosic crops would reduce land use and costs, while efficient transportation technologies would bring fuel demand in line with sustainable supplies. However, substantially more research is needed on advanced technologies to achieve these goals, and attention must be focused on overcoming perceived risk and capital investment challenges that now inhibit commercializing first-of-a-kind cellulosic ethanol processes before their major benefits can result.

Acknowledgments

This work was made possible through the support of the Thayer School of Engineering at Dartmouth College, for which the author is grateful. The Feldberg Library at Dartmouth College and particularly Karen Sluzenski are recognized for rapidly gathering valuable data. Hosein Shapouri of the United States Department of Agriculture and Michael Wang at Argonne National Laboratory provided important cost and energy balance information, and Luiz Carlos Correa Carvalho from UNICA in Brazil furnished Brazilian ethanol facts. The author appreciates input from several experts, including Jeffrey Bentley, Michael Gaylor, Donald Johnson, Lee Lynd, Jack Saddler, and Hosein Shapouri, who provided helpful suggestions. The National Renewable Energy Laboratory is commended for readily available, well-documented, detailed reports on cellulosic ethanol processes and economics.

SEE ALSO THE FOLLOWING ARTICLES

Biomass, Chemicals from • Biomass Combustion • Biomass for Renewable Energy and Fuels • Biomass Gasification • Biomass: Impact on Carbon Cycle and Greenhouse Gas Emissions • Biomass Resource Assessment • Greenhouse Gas Emissions, Alternative Scenarios of • Transportation Fuel Alternatives for Highway Vehicles

Further Reading

General Motors Corporation, Argonne National Laboratory, BP, ExxonMobil, and Shell. (2001). "Well-to-Wheel Energy Use and Greenhouse Gas Emissions of Advanced Fuel/Vehicle Systems: A North American Analysis, Volume 2, Main Report," Argonne National Laboratory, Report No. ANL/ES/RP-104528, Argonne, Illinois, June.

Goldemberg, J., Monaco, L. C., and Macedo, I. C. (1993). The Brazilian fuel-alcohol program. *In* "Fuels and Electricity from Renewable Resources" (R. H. Williams, T. B. Johansson, H. Kelly, and A. K. N. Reddy, Eds.), United Nations Solar Energy Group for Environment and Development (SEGED), 1992 UN Conference on Environment and Development in Brazil, Chapter 20, pp. 841–863.

Lynd, L. R. (1996). Overview and evaluation of fuel ethanol from cellulosic biomass: Technology, economics, the environment, and policy. *Ann. Rev. Energy Environ.* **21,** 403–465.

Lynd, L. R., Cushman, J. H., Nichols, R. J., and Wyman, C. E. (1991). Fuel ethanol from cellulosic biomass. *Science* **251,** 1318–1323.

Moreira, J. R., and Goldemberg, J. (1999). The alcohol program. *Energy Policy* **27,** 229–245.

Shapouri, H., Duffield, J.A., and Wang, M. (2002). "The Energy Balance of Corn Ethanol: An Update," U. S. Department of Agriculture, Agricultural Economic Report No. 813, Washington, DC.

Shapouri, H., Gallagher, P., and Graboski, M. (2002). "USDA's 1998 Ethanol Cost-of-Production Survey," U. S. Department of Agriculture, Agricultural Economic Report No. 808, Washington, DC.

Wang, M., Saricks, C., and Santini, D. (1999, January). "Effects of Fuel Ethanol Use on Fuel-Cycle Energy and Greenhouse Gas Emissions," ANL/ESD-38, Center for Transportation Research, Argonne National Laboratory, Argonne, IL.

Wheals, A. E., Basso, L. C., Alves, D. M. G., and Amorim, H. V. (1999). Fuel ethanol after 25 years. *TIBTECH* **17,** 482–487.

Wooley, R., Ruth, M., Glassner, D., and Sheehan, J. (1999). Process design and costing of bioethanol technology: A tool for determining the status and direction of research and development. *Biotechnol. Progress* **15,** 794–803.

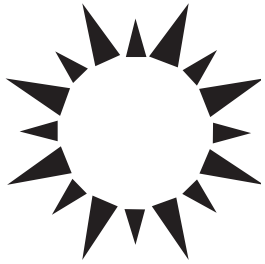
Wooley, R., Ruth, M., Sheehan, J., Ibsen, K., Majdeski, H., and Galvez, A. (1999, July). "Lignocellulosic Biomass to Ethanol Process Design and Economics Utilizing Co-Current Dilute Acid Prehydrolysis and Enzymatic Hydrolysis: Current and Futuristic Scenarios," National Renewable Energy Laboratory, Report NREL/TP-580–26157 Golden Colorado.

Wyman, C. E. (1999). Biomass ethanol: Technical progress, opportunities, and commercial challenges. *Ann. Rev. Energy Environ.* **24,** 189–226.

Wyman, C. E. (Ed.) (1996). "Handbook on Bioethanol: Production and Utilization." Applied Energy Technology Series, Taylor and Francis.

Wyman, C. E., Hinman, N. D., Bain, R. L., and Stevens D. J. (1993). Ethanol and methanol from cellulosic biomass. *In* "Fuels and Electricity from Renewable Resources" R. H. Williams, T. B. Johansson, H. Kelly, and A. K. N. Reddy, Eds.), United Nations Solar Energy Group for Environment and Development (SEGED), 1992 UN Conference on Environment and Development in Brazil, Chapter 21, pp. 865–924.

European Union Energy Policy



FELIX C. MATTHES
Oeko Institute, Institute for Applied Ecology
Berlin, Germany

1. **European Union**
2. **Areas of Regulation Concerning Energy Policy**
3. **Energy Policy Structures**
4. **Energy Policy, Competition, and The Single Market**
5. **Energy, Environment, and Climate Protection Policy**
6. **Nuclear Energy and Energy Policy**
7. **Energy Research Policy**
8. **Security of Energy Supply**
9. **Energy Policy and EU Enlargement**
10. **Energy Policy and External Relations**

Glossary

acquis communautaire The entire body of European laws. This includes all the treaties, regulations, and directives passed by the European institutions as well as judgments laid down by the European Court of Justice.

European Community (EC) The generic term for three different European communities: the European Coal and Steel Community (Treaty of Paris, 1951), the European Atomic Energy Community (Treaty of Rome, 1957), and the European Economic Community (Treaty of Rome, 1957). The EC formed the umbrella organization for these three communities following the signing of the fusion treaty in 1967.

European Economic Area (EEA) An agreement in 1992 between the European Community and three of the four members of the European Free Trade Association (i.e., Norway, Iceland, and Liechtenstein) to form a free trade area in all markets but agriculture; previously called the European Economic Space.

European Free Trade Association (EFTA) An organization founded in 1961; EFTA aims at supporting free trade among its members and increasing the liberalization of trade on a global basis, but particularly within Western Europe. Member states are Iceland, Liechtenstein, Norway, and Switzerland.

European Union (EU) The collective term for the European Community, the Common Foreign and Security Policy, and cooperation in the areas of justice and home affairs formed by the European Union Treaty (Maastricht Treaty, 1992).

European Union burden sharing The sharing out of emissions allowances among the 15 member states under the EU umbrella. The agreement on burden sharing was reached in June 1998 and is legally binding as part of the EU's instrument of ratification of the Kyoto Protocol.

single market (internal market) An economic area without internal restrictions, in which the four fundamental freedoms (free movement of goods, persons, services, and capital) can be realized. The Community's internal market project ("Europe 1992") was initiated during the 1980s and is the main reason for the huge movement toward integration at the end of the 1980s and beginning of the 1990s.

The European Union (EU) is one of the largest economic entities in the world. Although there is a significant and fast integration of many relevant policies of the EU member states, the development of a common EU energy policy in its different dimensions (single market, environment, security of energy supply, research, external relations, etc.) is a complicated process. Furthermore, the enlargement of the European Union creates new challenges for energy policy and the energy policy arena in Europe.

1. EUROPEAN UNION

With 15 states, a population of approximately 375 million, and economic output of roughly 8500 billion euros (€8500 billion), the European Union comprises one of the largest economic areas in the world. In 2004, 10 more states, primarily in Central and Eastern Europe, will join the European Union. The population of the EU will then increase to approximately 450 million, although the economic output will rise to only about €9000 billion as a result of the enlargement. The economic output is expected to rise to approximately €13,000 billion by

 © 2004 Elsevier Inc. All rights reserved.

2020, although the population is predicted to decrease slightly for demographic reasons.

The primary energy consumption of the current 15 member states of the European Union (EU-15) is about 1450 million metric tons of oil equivalents; this has increased by ~1% annually in the past decade (see Table I). Oil is the largest source of primary energy in the EU-15, accounting for over 40% of energy consumption, followed by natural gas, which accounts for over 20%, and coal, accounting for ~15%. The dependency of the EU on energy imports has increased steadily in the past few decades. Over 70% of oil consumption in the EU is covered by imports, and natural gas and coal imports account for roughly half of the total demand.

Sources of energy in the EU that can be used at competitive costs are very unevenly distributed. The main countries that have large natural gas reserves are the United Kingdom, The Netherlands, and Denmark; it is expected that production in all three countries will decrease. Out of all the member states of the EU-15, only the United Kingdom, Germany, and Spain have large coal reserves, and coal production can only be sustained with massive subsidies. Substantial lignite volumes are produced in Germany and Greece (Table I). Nuclear power plays a very different role in each of the EU member

TABLE I

Production and Consumption of Fossil Fuels in the European Union[a]

Member state	Hard coal	Lignite	Crude oil	Natural gas (TJ)[c]
	($\times 10^3$ metric tons)[b]			
Austria	—	1.3→ (97%)	1.1↓ (13%)	71.3↑ (23%)
Belgium	—	—	—	0.1→ (0%)
Denmark	—	—	17.8↑ (221%)	344.8↑ (167%)
Finland	—	5.6↓ (84%)	—	—
France	2.0↓ (15%)	0.3↓ (84%)	1.7↓ (2%)	69.6↓ (4%)
Germany	30.4↓ (58%)	175.4→ (95%)	3.2↓ (3%)	735.0↑ (22%)
Greece	—	67.0↑ (99%)	0.3↓ (1%)	2.0↓ (2%)
Ireland	—	—	—	44.6↓ (28%)
Italy	—	—	4.6↓ (5%)	633.7↓ (24%)
Luxembourg	—	—	—	—
Netherlands	—	—	2.4↓ (4%)	2414.6↓ (150%)
Portugal	—	—	—	—
Spain	14.5↓ (40%)	8.7→ (101%)	0.2↓ (0%)	6.9↓ (1%)
Sweden	—	—	—	—
United Kingdom	32.5↓ (52%)	—	126.2↑ (147%)	4542.6↑ (112%)

[a] Production for 2000/2001 and consumption trends for the past decade.
[b] The consumption trend is indicated by the arrow position; the percentage of gross domestic consumption is shown in parenthesis.
[c] Gross caloric value.

states. Only eight member states of the EU-15 operate nuclear power stations, and nuclear power accounts for ~33% of the total electricity volume overall within the EU-15, a proportion that remained more or less constant throughout the 1990s. The output of nuclear power stations has risen from ~117 to 125 gigawatts (GW) since 1990.

The framework for energy policy in the European Union is therefore based on a common economic area of considerable size, on the one hand, and on highly varied (1) energy industry structures in relation to primary energy volumes, (2) dependency on energy imports, and (3) orientation to specific energy sources, on the other hand. Ultimately, the basic conditions for European Union energy policy will change substantially as a result of the addition of new members in 2004.

2. AREAS OF REGULATION CONCERNING ENERGY POLICY

Since the beginnings of the European communities, the energy sector has formed a constituent element. Two of the three European treaties (European Coal and Steel Community and European Atomic Energy Community) particularly relate to the energy sector or to specific segments of the energy industry, even if the main reasons for concluding the treaties were primarily connected to security policy rather than to energy policy. The creation of the Single European Market, which was initiated in 1986 and includes the energy sector, lent a new quality to the energy policy and energy industry dimensions of the European Community (EC). However, the European Union has no direct jurisdiction over a European energy policy, and energy policy is the sole responsibility of the member states. Despite several attempts, all efforts so far to incorporate an energy chapter, which could be used to establish EU jurisdiction over energy policy, into the European Community Treaty have failed. Nevertheless, many of the areas of regulation that concern the energy sector directly or indirectly are influenced considerably by the European Union, even though, in practice, policy integration beyond the fundamental jurisdiction of the EU has been furthered in the different areas to a varying extent. The development of the European Single Market is very advanced. The liberalization of the electricity and gas markets is providing increasing dynamism for the energy sector. The development of trans-European networks belongs to the areas of energy policy that are explicitly within the jurisdiction of the EU. In the area of competition policy, the strict aid regulations concerning the reduction of state subsidies have, in particular, had a considerable effect on energy policy.

Integration has made relatively substantial progress in traditional environmental policy regulation areas. Above all, regulations for reducing conventional pollutant emissions are relevant to the energy sector in this respect. A common policy on climate protection is in the initial stages, but is becoming increasingly dynamic. In the area of nuclear energy, there are specific Community regulations that concern the area of research and development, the provision of beneficial financing, and the ownership and control of fissionable materials, whereas regulations concerning nuclear safety are strictly within the jurisdiction of the member states. A whole range of Community approaches to policy concern the area of research, development, and demonstration, albeit with very different equipment, deadlines and continuity.

The dialogue with energy-producing states is gaining in importance in the European Union's external relations. The strong relevance to energy policy of many EU regulations in related policy areas (single-market policy, competition policy, environment policy, etc.), despite the EU's lack of energy policy jurisdiction, leads to the situation that the principle of unanimity is claimed for policies and measures that have a significant effect on the energy sector (e.g., some fields of environmental policy). In reality, and above all in the 2004 new-member process, the veto option made possible with this principle is only of secondary importance.

3. ENERGY POLICY STRUCTURES

The European Commission has a key role in European energy policy. The Commission has the sole right of initiative for the legislation process and has a multitude of implementation, arbitration, and monitoring functions and the like. The Commission's Directorate-Generals for Energy and Transport, Competition, the Internal Market, and Environment are particularly relevant to energy policy. The governments of the member states participate in the decision-making processes through the Council of Ministers of the EU. The course of important compromises in Community policies is set here. The Council of Ministers for Energy, the Council of Ministers for the Environment, and the Council of Ministers for Finance are the main decision-making bodies important for energy policy.

The role of the European Parliament, which is directly elected by EU citizens, has become much more significant over the past few years. For most, but not all, legislation processes in the EU, the European Parliament must be involved on a graded basis. This involvement may take the form of codecision, assent, or hearing. For activities concerning energy policy, the codecision process is most important: three readings take place in both the Parliament and the Council of Ministers and a Conciliation Committee procedure may be initiated if necessary. The most important committees for energy policy in the European Parliament are the Committee on Industry, External Trade, Research, and Energy, as well as the Committee on the Environment, Public Health, and Consumer Policy.

European Union law arises from the founding treaties of the Community (primary law) and from secondary law, i.e., the legal instruments produced by the institutions of the Community. European Union regulations are wholly binding and are directly valid in each member state. European Union directives are binding for each member state with respect to the result to be achieved, and the directives must be translated into domestic law. However, the member states do have some freedom to choose the form and methods. The Court of Justice of the European Communities has gained in importance for energy policy. It checks the legal instruments of the EU and the member states to ensure that they are consistent with the treaties, and it decides on the requests of national courts concerning the interpretation or application of European Union law.

4. ENERGY POLICY, COMPETITION, AND THE SINGLE MARKET

The development of the European single market for energy forms a crucial basic condition for European energy policy. On the basis of the directives for liberalization of the electricity and gas sector that came into effect in 1997, the member states had to open their electricity and gas markets to competition by 1999 and 2000, respectively, and had to meet specific minimum requirements for regulation (decartelization, network access provisions). However, the minimum requirements for opening the markets (33% through 2003) and for network access transparency (possibility of negotiated network access), as well as the unbundling requirements for the companies, remained low.

Although the creation of a European single market was a priority during the liberalization, the electricity and gas markets essentially remained national markets, despite the fact that they were opened for competition. Energy prices became particularly dynamic, even on the national markets, and dropped considerably, especially at wholesale level. Even 4 years after the markets opened, changes in supplier are the exception rather than the rule in most member states, both for large customer groups and other consumer groups. The directives were revised in 2003 after the various member states found liberalization to be sluggish in reality. Accordingly, all member states must amend their national legislation by 2007 such that the electricity and gas markets are fully open to competition. The minimum requirements for companies (more rigorous rules for unbundling of generation, transmission, distribution, and sales), for regulation of the electricity and gas markets, and for consumer information are being intensified. The conditions for cross-border energy supply within the European Union are regulated directly by a corresponding European Union regulation. In addition, the EU has identified a range of infrastructure projects for electricity and natural gas, with which network bottlenecks are to be eliminated and cross-border trade intensified.

In contrast to the grid-bound final customer markets, the energy supply companies have become Europe oriented relatively quickly. A large number of company transactions in which the major European companies (EDF Energy, E.ON Energie, RWE Solutions, Vattenfall, etc.) played a particularly active role have been performed within and between the member states. These companies have stopped pursuing a large number of activities in the nonsupply industry and are instead pursuing massive growth strategies in the core business of energy through mergers and acquisitions. Non-European supply companies are rarely involved. Due to differences in timing of implementation of the market opening process, and also to a number of national conditions, the process within the European Union has involved in part a strong increase in market concentration. Trends toward oligopolies, however, are emerging on some electricity and gas supply submarkets in Europe. Expansion of the dominant market position of some large European suppliers was also evident during the privatization process in Central and Eastern Europe and the Baltic states.

State aid for the energy sector is strictly forbidden in the European Union, unless specific assent has been given. This has particular significance for state-

subsidized coal production in the EU, which, at €6.2 billion annually, still reaches a substantial level; German coal production is awarded the largest proportion (over 65%) of these subsidies. In the past, coal subsidies have steadily been reduced, primarily due to pressure from the European Commission, and they will continue to decline in the future. The reduction of regulations specific to energy sources in the European single market is also evident from the fact that one of the EU founding treaties—the treaty establishing the European Community for Coal and Steel (ECCS)—was not extended when it expired after 50 years on 23 July 2002. After it expired, coal and steel gained the status of normal industrial products. Via a special regulation of the Council, degressive aid for securing resources, for reducing production activity, and for extraordinary financial burdens (inherited burdens) has been declared permissible, at least until 2007, without it having to be approved in individual cases. If state aid is to be used for the production of renewable energy, it is regulated by Community guidelines governing state aid for environmental protection.

Discussions within the European Commission on the admissibility of so-called national energy bases (in other words, specific proportions of primary energy consumption for which advanced subsidy options could be made available to the member states) have been very controversial.

5. ENERGY, ENVIRONMENT, AND CLIMATE PROTECTION POLICY

Based on the Community's jurisdiction over environmental protection, the European Union is implementing a whole range of policies that have considerable effects on the energy sector. The range of policies extends from EU measures with immediate effect, to directives involving limited degrees of freedom for the member states when they are implemented, to the specification of goals that can be realized by the member states in any way they wish.

Climate protection is one of the strategic priorities of the EU in relation to sustainable development. A Community strategy on climate protection is inevitable because the EU has ratified the Kyoto Protocol and has therefore undertaken a double commitment. On the one hand, the EU, as an independent party to the agreement, has promised to reduce greenhouse gas emissions by 8%, from 1990 levels, in the initial period of 2008–2012; on the other hand, the member states have undertaken varied commitments as part of EU burden sharing, ranging from promises to make reductions (Luxembourg, 28%; Germany and Denmark, 21%; United Kingdom, 12.5%) to achieving growth restrictions (Spain, +15%; Greece, +25%; Portugal, +27%). For this purpose, the EU has developed a European Climate Change Program (ECCP), which comprises a number of measures that are also relevant to the energy sector. They include a directive on the promotion of electricity produced from renewable sources (2001), a directive on the promotion of cogeneration (2003), a directive on energy performance of buildings (2003), and a directive on energy services (planned for 2004). In contrast to the directives on liberalization, these directives generally contain only basic restrictions and definitions, and in some cases indicate target goals for expanding specific technologies. For example, in the field of renewable energy, the target is to increase the contribution of renewable energy to total primary energy consumption from 6 to 12% by 2010 and to increase its contribution to electricity consumption from 11 to 22%. Because implementation of the directives allows member states considerable freedom, it cannot be assumed that the targets will actually be achieved. However, more binding specifications result from a range of other measures. These include, for example, Commission directives for labeling electric appliances with energy efficiency data (1999), for adopting the Energy Star system for office equipment from the United States (2001), and for creating minimum requirements for the energy efficiency of electrical devices (refrigeration equipment, 1996; washing machines, tumble driers, ovens, water heaters, television sets, video equipment, electric motors, pumps, air compressors, etc., per agreements negotiated with industry at the end of the 1990s).

The initiatives pursued in the European Climate Protection Programme and in an independent Community strategy for promoting energy efficiency serve to realize the goal of improving energy productivity in the EU-15 annually by an additional 1% compared with the usual levels. However, the most prominent basic conditions for European Union energy policy result from an EU-wide emissions trading scheme for greenhouse gases, the introduction of which is planned for 2005. CO_2 emissions should be restricted for all industrial plants with a thermal input capacity greater than 20 MW and made flexible using a Europe-wide system of tradable certificates. In addition to a range of Europe-wide

specifications (participating sectors, coverage of carbon dioxide only, incorporation of the flexible mechanisms of the Kyoto Protocol, Joint Implementation and Clean Development Mechanism, free allocation of most of the emission allowances at first, etc.), the member states have considerable freedom to shape the system, particularly when distributing the emission rights to the different plants on the basis of the national allocation plans.

Considerable progress has been made in the reduction of conventional pollutant emissions in the energy sector. Sulfur dioxide emissions from power plants in the EU-15 dropped by almost 60% in the 1990s, and the corresponding nitrogen oxide emissions decreased by roughly 25%. Particularly with the directive on the limitation of atmospheric emissions of certain pollutants from large combustion plants (2001), similar demands have been established at EU level for ecological standards for power plants in terms of conventional pollutants, and hence a consistent ecological framework has been created for competition in the field of thermal power plants. New power plants must meet strict standards, and existing plants must observe defined emission ceilings and targets. The directive on national emission ceilings for certain atmospheric pollutants (2001), which sets ambitious reduction targets for the entire spectrum of conventional atmospheric pollutant emissions within the respective member states, forms a framework beyond the energy industry in the strictest sense, but still has an important formative influence on the energy sector. One of the main targets of this directive is to halve, within a reasonable period, the number of areas in which the critical load of acidifying substances is exceeded.

6. NUCLEAR ENERGY AND ENERGY POLICY

Nuclear energy plays an important role in the energy supply of the European Union. Eight of the fifteen EU member states operate a total of 145 nuclear reactors with a net output of roughly 120,000 MW, which corresponds to about one-fifth of the entire power plant capacity in the EU-15. The most recent nuclear power plant in the EU-15 was commissioned in France in 1999.

In the EU member states that operate nuclear power plants, the proportion to power supply contributed by nuclear energy varies. In The Netherlands, ~4% of all electricity is generated by nuclear power plants, and in Finland, Germany, Spain, and the United Kingdom, the contribution is between 26 and 33%; in Sweden, Belgium, and France, it is 47, 58, and 75%, respectively. Three member states (Belgium, Germany, and Sweden) pursue an active policy of withdrawing from nuclear energy. Two member states (Finland and France) pursue a policy of setting up new nuclear power stations, although an actual project is underway in Finland alone. In the remaining three countries (The Netherlands, Spain, and the United Kingdom), it is expected from today's perspective that the use of nuclear energy will be phased out. Out of the 10 new EU member states, five operate nuclear power plants, which are predominantly of Russian design.

The treaty on establishing the European Atomic Energy Community (Euratom Treaty), which was signed in 1957 and is valid for an indefinite period, created the basis for a range of specific regulations on the use of nuclear energy; unlike other areas of regulation in the EU, these regulation should actually be part of secondary legislation. First, these special regulations relate to specific problems in the use of nuclear energy. This includes monitoring nuclear fuel in the EU and a range of other regulations on radiation protection. It must be emphasized in this respect that the nuclear fuels used in the EU are officially the property of the Community and the nuclear power operators have only rights of use. Second, there is a whole range of regulations concerning the promotion of nuclear energy. They include the promotion of research specifically concerning the use of nuclear energy, as well as special financing possibilities for nuclear energy projects. Funds for research into nuclear fission and nuclear fusion amounted to between €1.2 billion and €1.3 billion for each 4- or 5-year Euratom research program (see Section 7) in the past few years. In the framework of Euratom loans, the EU provides investments for nuclear projects, the (very favorable) refinancing conditions of which are based on those for public borrowers. In the period 1977–2002, a total of €3.2 billion was available through Euratom loans, which were also provided for nuclear projects in Central and Eastern European states, as well as in the states of the former Soviet Union, in order to commission new nuclear power plants, among other things. In 2002, the European Commission presented a proposal to increase the Euratom loan ceiling by a further €2 billion, which led to heated debates on the legitimation of Euratom loans, e.g., before the background of market distortions in favor of nuclear energy. Third, other regulations focus on the

Community control of fuel procurement. Based on the original assumption that diversification is a necessity due to a shortage of fuel, nuclear fuel procurement is predominantly handled by a separate institution, the Euratom Supply Agency, and strict checks are carried out to ensure adherence to specific import quotas (for example, from Russia). This has led to intervention in private contracts of individual power plant operators. Finally, the legislation procedures in the Euratom Treaty differ greatly from those in other EU regulations. In particular, this affects the participation of the European Parliament, whose opportunities for involvement and influence in the context of Euratom fall considerably short of the procedures provided for in the EU treaty.

With regard to nuclear energy, the concentration of binding regulations at the EU level is comparatively low, except in the fields of proliferation and health-related radiation protection. However, in the important and—in terms of the common market and competition provisions—cost-related regulation areas of disposal and liability, the regulations of the member states dominate. Therefore, the service specifications, implementation periods, and reserve variants that are relevant to fuel and plant disposal vary greatly among the EU member states, leading to significant differences in the disposal cost burdens of nuclear power generation, in addition to varying implementation efficiency. Companies operating nuclear power plants in EU member states must also make liability provisions to a widely varying extent, which is also connected to differing cost burdens. With regard to safety standards, the lack of coordinated regulations, particularly in the context of EU enlargement, has proved to be a major failing. Against this background, attempts have been made since 1999 to develop a common approach to reactor safety and corresponding regulation procedures in the Western European Nuclear Regulators Association (WENRA); however, the objective of establishing coordinated and binding safety standards for the use of nuclear power has not yet been achieved.

7. ENERGY RESEARCH POLICY

The European Union promotes energy research through specific programs. A distinction is made between nuclear and nonnuclear research in this respect. Nonnuclear energy research is implemented through the EU's general framework program for research. The sixth and most recent EU Framework Programme for Research and Technological Development (2002–2006) has a total budget of €17.5 billion, €2.12 billion of which is earmarked for sustainable development, global changes, and ecosystems. This also includes the field of sustainable energy systems, for which €810 million has been appropriated. Compared with the corresponding research field in the fifth Euratom framework program for research (1998–2002), this constitutes a reduction in funds (€1.042 billion was employed for nonnuclear energy research previously). The fourth Euratom framework program for research (1995–1998) had a budget of €1.03 billion in the field of research, development, and demonstration concerning nonnuclear energy technologies.

Renewable energy systems (including integration into energy supply systems), energy saving technologies, hydrogen and fuel cells, CO_2 separation and dumping, and socioeconomic research in the energy sector are promoted above all in this regard. Fossil energy generation technologies play a much smaller role in the current research program than in previous promotion programs. Research specifically for the coal sector was promoted until 2002 through the research and development programs of the European Community for Coal and Steel (ECCS). It is stipulated in the agreements on the expiry of the ECCS in July 2002 that income from the remaining assets of the ECCS will be placed in a research fund for coal and steel; 27.2% of this is earmarked exclusively for the coal sector.

The promotion of nuclear energy research is implemented in a special program in the scope of the Euratom Treaty. A total of €1.23 billion is provided for in the sixth Euratom research program for the period 2002 to 2006. The majority of the budget (€750 million) is earmarked for fusion research, the International Thermonuclear Experimental Reactor (ITER) project being the focus of the research. Research in the field of radioactive waste, radiation protection, plant safety, and new reactor designs is supported by the remaining funds. The budget for the promotion of research by Euratom has remained largely constant in comparison with previous promotion programs. In the fifth Euratom framework program for research (1998–2002), a total of €1.26 billion was employed, whereas the fourth Euratom framework program for research (1995–1998) had €1.336 billion.

Beyond research, the EU establishes framework programs lasting several years for supporting the field of energy research. The total budget for the energy promotion program from 1998 to 2002 was €175 million, the majority of which was earmarked

for energy saving (SAVE program; €66 million) and renewable energy (ALTENER program €77 million). The "Intelligent Energy for Europe" program (2003–2006) has a total budget of €215 million, which is to be used primarily in the areas of energy efficiency (continuation of SAVE), renewable energy (continuation of ALTENER), corresponding activities in the international field (COOPENER), and energy matters in the transport sector (STEER).

8. SECURITY OF ENERGY SUPPLY

The major oil price fluctuations in 1999/2000 led to discussions on the security of energy supply in the European Union being conducted with renewed intensity. The European Commission has drawn up a well-heeded discussion framework for this, as presented in the Green Paper, "Towards a European Strategy for the Security of Energy Supply," The particular challenges of energy supply security for the first two to three decades of the 21st century are a result of the EU's increasing dependency on imports (from the current 50% to an estimated 70% in 2030), the necessary reinvestment in the European energy industry on a large-scale, and increased environmental and climate protection requirements.

The main features of a long-term European Union strategy on energy policy are as follows:

- An increase in energy efficiency on the demand side of the market is to form an important core of future EU energy policy, because an increased energy supply from the EU member states cannot be developed under acceptable conditions.
- The precedence of climate-friendly energy sources, such as new and renewable energy sources for generating electricity and heat as well as for the transport sector; a further discussion on the use of nuclear energy in Europe is also required for this purpose in the view of the European Commission.
- The coordinated increase of strategic stocks of crude oil and natural gas.
- The development of new sources for energy imports and new import channels for crude oil and natural gas.

Even though the debate initiated by the European Commission on security of energy supply in the EU is clearly driven by a whole range of ulterior motives (fight for more jurisdiction for EU institutions over energy policy, attempts to reevaluate nuclear energy, etc.), and even though, despite an extensive discussion process, the Green Paper has continued to be merely discussion input from the European Commission (and other EU institutions, such as the European Parliament, have not become involved), the debate on energy supply security for EU energy policy will form an important framework in the future.

9. ENERGY POLICY AND EU ENLARGEMENT

Essential changes to the basic political conditions for energy policy in Europe will emerge from enlargement of the European Union in 2004. In addition to the accession of 10 states in mid-2004 (Malta, Cyprus, the Czech Republic, Hungary, Poland, Slovakia, Slovenia, Estonia, Latvia, and Lithuania), negotiations are continuing with other states (Bulgaria and Romania). The acceptance of Turkey into the EU will remain a highly controversial issue in the next few years. Other potential applicant states are those in the Western Balkans. The number of EU member states will therefore rise from 15 (EU-15) to 25 (EU-25) in the short to medium term. In the medium to long term, the number of member states may increase to over 30 (EU-30 +). In addition, the European Free Trade Association (EFTA) regulations and bilateral regulations with the non-EU states of Norway and Switzerland also form a specific European framework for energy policy. The decision-making structures in a European Union with 25, 30, or more member states will (have to) change considerably. Majority decisions and a clearer allocation of tasks will play a larger role if the capacity of the Community and member states to act is to be retained and expanded. At the same time, competition for available funds will increase.

The European Union's enlargement to the east is the enlargement with the greatest economic divide compared with previous enlargement phases. In the four enlargement phases up to 1995, the economic strength of the new member states was comparable to, or even greater than, that of the European Union. The states to be incorporated into the EU in 2004 have an average gross domestic product (GDP) level per resident of only ~20% of that of the EU today. The adjustment of the economic strength of the accession states to the level of the EU requires strong growth processes in these countries on the one hand, and a large rise in the cost efficiency requirements of EU policies on the other hand. With respect to their energy markets, the

future EU member states face a threefold challenge that none of the current EU members had to confront. They must (1) liberalize energy prices and the energy markets in a relatively short space of time and at the same time must (2) renovate and modernize their plants, and (3) they have much less economic strength and purchasing power compared to the EU as it is today.

The restructuring of the energy sector is progressing to a varying degree in the different states and is generally also closely associated with the privatization of the energy industries. The motives for privatization are varied, but are often mainly due to the generation of income as well as the necessity and expectation of a considerable influx of capital from international investors. In the privatization procedures that have taken place so far, the companies that were already strongest in the EU were particularly successful, so privatizations in the EU accession states have already contributed to a further concentration of companies and to a substantial increase in the market power of individual energy suppliers in the European energy industry. This results in competition-related and liberalization-related challenges for forming the basic conditions of the energy markets.

In many of the accession states, energy efficiency is much lower compared to the current level in the EU-15. Although the primary energy consumption per resident in the EU-15 is mostly (much) higher than it is in the accession states, energy productivity (expressed as primary energy consumption per unit GDP on the basis of purchasing power parities) is poorer in most of the accession states, compared to the EU-15 average, by a factor of roughly two. The degree of macroeconomic stabilization in the accession states will, in particular, play a large role in the creation of stable basic conditions for investment, particularly in the area of energy efficiency. Economic stability and a high degree of legal security are central requirements for corresponding investment. In view of the restricted purchasing power and rising energy prices, primarily for households and small consumers as a result of liberalization, increasing energy efficiency also has a strong social element.

The situation concerning emissions corresponds to the initial condition of energy productivity. There is a much higher specific output of the main pollutants (carbon dioxide and sulfur dioxide) in most states than there is in the EU-15. This means substantial exertions and expenditure are necessary for the states to observe the EU environmental standards quickly; in part, the air pollution control measures have already progressed relatively far, however.

In principle, at the date of accession, the entire body of European legislation (acquis communautaire) will be applicable to the accession countries, obliging them to fulfill the specific requirements of each piece of legislation. Against this background, EU enlargement creates for all applicant countries a process, in the course of which significant progress has to be achieved before official accession. The interaction of this basic requirement with economic restructuring has led to significant improvements in economic efficiency and in the quality of the environment. For example, total sulfur dioxide emissions from power plants in the Czech Republic have been reduced by 88% in the period from 1990 to 2000. Relevant reductions can also be observed for greenhouse gas emissions. Due to the provisions of the Kyoto Protocol, however, greenhouse gas reductions of the accession countries cannot be taken into account in compliance on the part of the European Union until 2012. The primary energy structure of an expanded EU will not change much on the whole as a result of enlargement. Nevertheless, in some of the future member states, substantial moves toward mineral oil and natural gas are to be expected. There are additional challenges for some countries that have a very specific energy industry structure.

Poland, the largest producer of coal in Europe today, will become a member of the EU. Although coal production has almost halved since the end of the 1980s, Polish coal production in 2001, at $\sim$100 million t, was 31% higher than the entire coal production in the EU. Due to the unfavorable geological conditions, Polish coal extraction has been classified as being just as uncompetitive as Western European production, so the massive reduction in subsidies deemed necessary by the EU will result in a considerable drop in production. The discussion on coal subsidies will therefore become more explosive in terms of European politics. Among the new and potential EU member states, some countries are extensive lignite producers (e.g., Poland, Czech Republic, Hungary, and Bulgaria). Because only Germany, Greece, and Spain (although the latter is cutting back) lay claim to significant lignite production in the EU, the role of lignite will change in the short, medium, and long term in an enlarged EU.

Substantial growth in oil and natural gas consumption is expected in most accession states in the

next few years. On the one hand, this places considerable demands on the inclusion of these states in the stockpiling regime of the International Energy Agency (IEA). On the other hand, the natural gas markets in most of the states applying for EU membership are still strongly oriented to procuring natural gas from Russia. The diversification of natural gas supplies for states in Central and Eastern Europe will therefore form an important factor in energy policy in the medium term.

With the enlargement of the EU, the Community is faced with specific safety problems concerning substantial nuclear power plant capacity (mostly Soviet-designed plants), even though nuclear power is a less significant energy source in most of the accession states than it is in the EU today. Because no generally binding safety standards apply at the EU level, only a comparison with the practices and regulations applicable in all member states was used as a measure for demands for high safety standards in the field of nuclear power in the EU accession process. An enlarged EU will then be confronted with a whole range of reactors that, in the general view, do not satisfy the requirements and practices commonly accepted in the EU and have (serious) design faults. Therefore, in the course of EU accession, the closure of reactors has been agreed with some states, although the implementation of this agreement has not yet been fully guaranteed. Other reactors need to be retrofitted urgently, but there is a dispute about whether and to what extent "acceptable safety levels" can be achieved through retrofitting measures, particularly in light of the fact that construction weaknesses and quality shortcomings have arisen when plants have been commissioned. Many questions concerning the ultimate disposal of highly radioactive waste and the financing of nuclear plant closures have not been answered at all by representatives of most of the new EU member states.

District heating accounts for a large proportion of the total energy supply (up to 25%) in most of the Central and Eastern European accession states, and in the Baltic accession states. This figure is much higher than the EU-15 average ($\sim$2%), and is also higher than the figure in states with a high proportion of district heating (Denmark, 15%; Finland, $\sim$11%; Germany, 4%). The renovation and modernization of district heating systems and the introduction of highly efficient combined heat and power (CHP) production is not only an area for action concerning the energy industry in the Eastern European EU applicant states, but also has far-reaching social effects.

10. ENERGY POLICY AND EXTERNAL RELATIONS

Energy policy is a feature in the external relations of the European Union to an increasing extent. This is mainly because of two processes. First, the EU set the target of creating a European single market that includes other countries, in addition to the EU member states. Second, there is also a need for stronger cooperation with some energy-exporting regions in light of the EU enlargement, and also due to developments on the global energy markets in view of energy supply security. An essential determinant for the EU's relations with its immediate neighbor states is the aim of developing network infrastructures in such a way that at least 10% of the respective domestic electricity consumption will be covered by cross-border supply.

The European Union is also trying to influence the competition structures in neighbor states. For example, the bargaining cartel of Norwegian gas producers for natural gas supplies in the EU was broken up due to pressure from the European Union. Establishment of comparable market access and specification of equal standards (especially in the field of nuclear energy) are at the forefront of single-market activities in relation to third-party states.

The EU has institutionalized three fields of external relations and has also defined further areas of particular interest in relationship to Russia, the Mediterranean countries, and southeast Europe. In October 2000, the EU agreed on a regular European Union–Russia Energy Dialogue proposal to develop a strategic energy partnership. Because Russia is the largest single partner of the EU for energy supplies, long-term natural gas supply agreements, the establishment of legal and investment security for energy projects, trade in nuclear materials, and environment and climate protection issues (with the Kyoto Protocol to the Framework Convention on Climate Change in mind) are at the forefront. Previously, in 1997, the European–Mediterranean Energy Forum was established in order to increase cooperation with a total of 12 Mediterranean countries (Egypt, Israel, Jordan, Lebanon, Malta, Morocco, the Palestinian National Authority, Syria, Tunisia, Turkey, and Cyprus, with observer status for Libya and Mauritania). The free trade area planned for 2010, energy supply security issues, and environmental projects in the energy industry are at the forefront here. Finally, since March 2002, the European Commission has been submitting proposals for setting up a regional electricity and natural gas single market

in southeast Europe, and it has signed corresponding agreements (Albania, Bosnia and Herzegovina, the Federal Republic of Yugoslavia, Bulgaria, the former Yugoslav Republic of Macedonia, Greece, Turkey, Croatia, Romania, Kosovo, with observer status for the Republic of Moldavia, Slovenia, Austria, Hungary, and Italy). Cooperation and planning work has been undertaken by the different state and industrial institutions in the so-called Athens Forum.

In addition to these comparatively strongly institutionalized external relations, the European Union has defined three areas of particular interest for EU energy policy. These include, first, the so-called Northern Dimension, which aims to increase cooperation between Baltic states and Russia (especially in the field of nuclear security and waste disposal). Second, the European Union has defined particular interests in relation to the oil and gas reserves in the Caspian Basin. Facilitating the transportation of Caspian resources to Europe is a key issue here. For this purpose, the development of transport infrastructure across Turkey and Iran is under consideration; great importance is also attached to cooperation with Azerbaijan and Kazakhstan, however. Third, Ukraine plays a particular role in EU energy policy; it is the most important country in the transit of Russian natural gas to Western Europe. However, development of a cooperative relationship is at a very early stage here.

Particularly with Russia and the other states of the former Soviet Union in mind, the Energy Charter may become very significant, increasing investment security and access to network infrastructures. The extent to which this instrument may affect multilateral cooperation between the energy industry and energy policy will mainly be decided by Russia's attitude toward this agreement, and this is somewhat unpredictable.

SEE ALSO THE FOLLOWING ARTICLES

Equity and Distribution in Energy Policy • Fuel Economy Initiatives: International Comparisons • Geopolitics of Energy • National Energy Policy: Brazil • National Energy Policy: China • National Energy Policy: India • National Energy Policy: Japan • National Energy Policy: United States • Nationalism and Oil • Research and Development Trends for Energy

Further Reading

Cames, M., Matthes, F. C., *et al*. (2000). "Instruments and Options for Environmental Policy during the Accession Process of EU Associated Countries in the Area of Environment and Energy." UBA Texte 42-01. Federal Environmental Agency, Berlin.

Commission of the European Communities (CEC). (2000). Action plan to improve energy efficiency in the European Community. "Communication from the Commission to the Council, the European Parliament, the Economic and Social Committee and the Committee of the Regions." COM (2000) 247 final, 26.04.2000. CEC, Brussels.

Commission of the European Communities (CEC). (2003). "Communication from the Commission to the Council and the European Parliament on the Development of Energy Policy for the Enlarged European Union, Its Neighbours and Partner Countries." COM (2003) 262 final, 13.05.2003. CEC, Brussels.

European Commission (EC). (2001). Green Paper. "Towards a European Strategy for the Security of Energy Supply." EC, Luxembourg.

European Environment Agency (EEA). (2002). "Energy and Environment in the European Union." Environmental Issue Report No. 31. EEA, Luxembourg.

European Parliament (EP). (2001). "The European Parliament and the Euratom Treaty: Past, Present and Future." Energy and Research Series, ENER 114 EN. EP, Luxembourg.

Matthes, F. C., and Timpe, C. (2000). "Sustainability and the Future of European Electricity Policy." Heinrich Boell Foundation Documentations, Papers and Reports No. 4, Berlin.

Evolutionary Economics and Energy

QUENTIN DUROY and JOHN GOWDY
Rensselaer Polytechnic Institute
Troy, New York, United States

1. **Energy and the Evolution of Human Society**
2. **Prices, Energy, and Scarcity**
3. **Energy and Economic Growth**
4. **Energy and Evolutionary Economics**
5. **Conclusion**

Glossary

energetics The study of the sources of energy, direct and indirect, available for mechanical work.
entropy In a closed system concentrations tend to disperse, structure tends to disappear, and order becomes disorder.
exergy The amount of energy converted into useful work.
exosomatic evolution The evolution by humans of outside-the-body tools and technology to supplement the limitations of our physically evolved bodies.
neoclassical economics The dominant school of economics whose core is the notion of general equilibrium and Pareto efficiency.
Pareto efficiency Exists when no further trading can make one party better off without making another worse off.
punctuated equilibrium The pattern of evolution tends to be one of long periods of stasis punctuated by periods of rapid change.
self-organization The tendency of an organic system, such as an ecosystem or the human economy, to actively shape its evolutionary environment.

Economic evolution has not been smooth or continuous. The history of the human occupation of Earth is characterized by long periods of stasis interrupted by periods of rapid movement to another system of socioeconomic organization. Major punctuations in human history include the use of fire, agriculture as the dominant mode of production, and the industrial revolution. All these changes involved the intensification of energy use and all of them signaled dramatic changes in social and economic organization. The past 100 years or so have seen the most rapid and dramatic increase in the human population and in the planetary impact of our species in the 100,000-year history of *Homo sapiens*. The 5-2 fold increase in population and the approximately 40-fold increase in the value of economic output since 1850 were made possible by the era of petroleum. Harnessing fossil fuels has led to unimaginable material wealth for a large percentage of the world's population. However, the orgy of economic growth has left billions in poverty and has also brought humans to the limits of Earth's life-support systems. The tools of mainstream welfare economics have proved to be of limited value in understanding the role of energy in the evolution of the economy. Standard neoclassical economics is a theory of the allocation of scarce resources among alternative ends. It has a lot to tell us about the nature of market exchange, but it does not contain a useful theory of economic evolution or a notion of humans as a biological species inhabiting a finite planet. Fortunately, other tools are available to economists in addition to the general equilibrium model of constrained optimization. An evolutionary approach to energy use can give us an appreciation of the uniqueness of energy as a primary input to economic activity and of the importance of the fossil fuel age in reshaping the planet we inhabit.

1. ENERGY AND THE EVOLUTION OF HUMAN SOCIETY

Two major energy-induced milestones in human cultural and technological evolution were the use of fire and the use of fossil fuels. The earliest evidence of the use of fire is associated with a human ancestor known as *Homo erectus*, who first appeared

 © 2004 Elsevier Inc. All rights reserved.

approximately 1.9 million years ago (there is controversy here, and some classify the earliest fire user as another species called *Homo ergaster*). Fire made possible the expansion of human occupation of the planet from Africa to the cooler areas of Europe and Asia. *Homo erectus* is arguably the most successful evolutionary human experiment in terms of having a sustainable technology (called the *Acheulean*) that lasted, essentially unchanged, for approximately 1 million years. Later humans used fire to dramatically change local ecosystems and landscapes. The Aborigines of Australia and Native Americans used fire to create habitats for large mammals that could be hunted. Much of the landscape encountered by the first Europeans in Australia and North America was the product of human alteration by fire.

Although coal has been used for heating and cooking for millennia, the age of fossil fuels was ushered in by the invention and refinement of the steam engine by Thomas Savery (1698), Thomas Newcome (1712), and James Watt (1769). In the 19th century, coal began to be heavily used in the booming industrialization of western Europe and the United States, especially in iron smelting. This development led to a dramatic increase in coal mining, forcing mining companies to dig deeper into the ground. Eventually, extraction activities became severely constrained by groundwater flooding. To remedy this problem, the mining industry relied on steam pumps. The use of these pumps was the first industrial application of the steam engine. The subsequent development and domination of steam and heat engines in the late 19th and early 20th centuries can be attributed, in retrospect, to an evolutionary accident that led to a technological lock-in and the current dependence of most nations on fossil fuels. What future historians will call the oil era will last only approximately 250 years, and we are now closer to the end of this era than to the beginning. The use of fossil fuel to power the industrial revolution not only had a dramatic effect on the level of economic output but also changed human society and the relationship between humans and the rest of the planet.

In the 19th century, Lewis Henry Morgan and Edward Tylor were among the first social scientists to apply an evolutionary approach to the study of traditional societies. However, it was the work of Leslie White and Fred Cottrell in the 1940s and 1950s that highlighted the central role of energy in the evolution of social systems. White argued that the evolution of human culture is conditioned by the ability of humans to harness and control energy: "Culture evolves as the amount of energy per capita per year is increased, and/or as the efficiency of the instrumental means of putting the energy to work is increased." He proposed a systemic analysis of cultural evolution. Human society, in his view, is organized around three systems: technological, sociological, and ideological. Via exosomatic evolution (the emergence of technology and material artifacts), humans managed to successfully harness and control different forms of energy. An example of exosomatic evolution is one of the most important "energy revolutions"—the development of the spear thrower, or atlatl, approximately 30,000 years ago. The additional power that this invention imparted to the human body revolutionized the hunting of big game animals in the Upper Pleistocene. According to White, the evolution of technology inevitably led to the creation of social institutions that, in turn, engendered a specific worldview or ideology (these three systems coevolved). White argued that cultural evolution was marked by two revolutions: the agricultural (harnessing and controlling cultivation and domestication) and the industrial (harnessing and controlling fossil fuels). Each revolution was subsequently characterized by an increase in technology that eventually reached a plateau due to inner systemic limits. A given technological system reaches its limit in terms of innovation, given the type of energy used, and the social system is no longer adequate to deal with the rapid growth of social units. Further cultural advance requires harnessing and controlling different energy forms, thus leading to a new revolution.

Much like White, Cottrell focused his analysis on the level of energy use as one of the key determinants in the emergence of complex cultural forms. Cottrell's historical account of energy use and social evolution and his methodology are similar to those of Julian Steward and subsequent ecological anthropologists. Steward criticized the unilinear aspect of White's evolutionary theory for failing to capture the complexity and diversity of human cultures. For Steward and Cottrell, there was no systemic reason for every culture to follow a given universal developmental path. Cottrell saw institutions and ideology as limiting factors toward the creation of a single world culture. Steward believed that regularities existed in cultural change but that local variations prevented the occurrence of universal stages of development. Nevertheless, beyond their theoretical and methodological differences, White, Cottrell, and Steward understood the importance of the laws of thermodynamics and the biophysical

constraints these laws impose on the availability and use of energy. More generally, their work emphasized the need for interdisciplinary approaches to the study of the role of energy in sociocultural evolution.

As documented by Cleveland, Daly, and Martinez-Alier, the development of energetics—the study of the role of energy in economic production—was an important contribution of the social sciences in the 19th and early 20th centuries. However, it was the growing concern over increasing environmental degradation and energy shortages in the 1970s that made the work of Patrick Geddes, Frederick Soddy, and others all the more relevant. The environmental movement led scholars, across disciplines, to focus on the role of energy in social and economic systems. In the 1970s, Adams, Ayres, and Georgescu-Roegen, among others, wrote extensively about the significance of the laws of thermodynamics to social and economic evolution. These authors taught us that socioeconomic processes are subject to the first and second laws of thermodynamics. The first law states that the total matter and energy for any system and its surroundings does not change over time. Matter and energy are conserved. The second law states that entropy always increases in an isolated system. Low-entropy (concentrated and useful) energy is irreversibly transformed in a closed system into high-entropy energy (dispersed and unusable). For Adams, exosomatic evolution provides human beings with the capacity to capture and degrade energy more rapidly and more effectively. Consequently, human activities speed up the march of entropy.

For Georgescu-Roegen, the unfolding of the entropy law is what the economic process is fundamentally about. Although there is little doubt that social and economic evolution is subject to the laws of thermodynamics, many issues are still unresolved. Georgescu-Roegen pushed the theoretical boundaries of the relevance of thermodynamics to economics by proposing a fourth law, which states that complete recycling is impossible (i.e., perpetual motion of the third kind is not possible). Ayres contested this proposition, arguing that although the creation of a "wastebasket" is inevitable due to the second law, this wastebasket, over a long period of time, can be a source of usable energy and matter. As Kozo Mayumi pointed out, the fourth law, as a consequence, needs to be qualified as follows: On an economic timescale, given the state of current technology and the limits to energy inputs, complete recycling is impossible.

With the growing literature and debates over the relevance of thermodynamics to social sciences during the past three decades, a consensus has developed about the issues of concern to environmental, energy, and ecological economists. First, economic production and consumption processes are constrained by biophysical limits and therefore are subject to the laws of thermodynamics. Second, increased human activity (economic growth) speeds up the entropic process (increases in heat and material wastes) and puts more pressure on the natural environment. Third, if unchecked, this will eventually limit the potential for economic and social evolution.

2. PRICES, ENERGY, AND SCARCITY

The relationship between energy and economic value has intrigued economists since the inception of economics as an academic discipline. Members of the first recognized school of economics, the 18th-century French physiocrats, saw solar energy as being the ultimate source of wealth in agricultural production. Later, Sardi Cournot experimented with the steam engine to show how much useful work could be obtained from a given amount of energy. William Jevons recognized the importance of coal to the British economy and warned of the consequences of its increasing scarcity. In the late 19th and early 20th centuries, a number of physical scientists, including Geddes and Soddy, wrote extensively about economic issues and warned of the consequences of ignoring the importance of energy to economic value. Geddes is particularly interesting because he is perhaps the first to criticize—in a letter to Leon Walras in 1883—the general equilibrium approach of mathematical economics for ignoring the physical nature of production and the biological nature of the human species. Geddes did not advocate an energy theory of value, but he did point out that there is more to economic value than mere "feelings." Coal, for example, was not just another commodity but rather an essential input for producing food and other necessities for human existence. Geddes and Soddy pioneered the field of energetics–that is, calculating the amount of energy required to produce various economic goods.

In the 1960s and 1970s, the work of Kenneth Boulding and Nicholas Georgescu-Roegen was tremendously influential in turning the attention of economists to biophysical constraints on economic activity. Geogrescu-Roegen's monumental work, "The Entropy Law and the Economic Process," inspired a generation of economists to consider the environmental consequences of economic growth. Georgescu-Roegen was emphatic in rejecting market

prices as a means of allocating irreplaceable natural resources or irreducible pollution. Since future generations cannot bid in energy markets, the allocation of these resources between generations should be a matter of public policy, not isolated individuals bidding at a particular point in time in ill-informed markets.

Kenneth Boulding's classic article, "The Economics of the Coming Spaceship Earth," captured the spirit of the new environmental movement by emphasizing that pollution and resource exhaustion were part of any economic process. Except for a few isolated voices, however, most economists were silent on the energy issue until the oil crisis in 1973. Even after energy prices quadrupled in a few months, mainstream economists such as Robert Solow downplayed the role of energy, stressing the ability of the economy to substitute other inputs for energy in the face of relative price changes. The energy economy debate frequently led to ludicrous statements on both sides. The economist Julian Simon stated that there is enough energy to fuel continual economic growth for the next 7 billion years. Although not in the same league as Simon in terms of fantasy, many environmentalists made overly pessimistic predictions, such as those of the Club of Rome warning of "a rather sudden and uncontrollable decline in both population and industrial capacity" because of resource shortages.

The relationship between energy prices and resource scarcity has been the subject of intense debate among economists. In the 1930s, Harold Hotelling described the ideal exhaustion path of an absolutely scarce resource. Hotelling assumed, among other things, that economic actors are aware of the total amount of the scarce resource, of present and future extraction costs, and, implicitly, of future extraction technology. Given these assumptions, rising prices will reflect the increasing scarcity of this resource. Numerous writers (e.g., Barnett and Morse) have incorrectly used Hotelling's result to suggest that since real energy prices have not significantly increased during the past 100 years, energy is not a scarce resource. Norgaard pointed out that this argument is logically flawed since it goes from Hotelling's conclusion (economic indicators will reflect scarcity) to his premise (if resources are scarce) while ignoring his caveat that resource allocators must be fully informed of future scarcity.

After more than 30 years of debate, punctuated by real-world energy crises, it is clear that changes in energy availability can have drastic consequences for the economy. It is also clear that market economies have a remarkable ability to adjust to energy shocks and, at least in the short term, to substitute technology for increasingly scarce resources. However, the role of essential inputs in production is one of the major unresolved issues in economics. We know that fossil fuel energy is an essential factor of production and we know that the supply of fossil fuels is finite. We also know that higher prices will call forth substitutes and more efficient technologies. So far at least, neoclassical theory has not provided much guidance to understanding the relationship between energy and economic growth. Crude estimates of the marginal rates of substitution between "energy" and "nonenergy" are methodologically flawed and do not tell us much about the ability of the economy to adjust to resource exhaustion. Evolutionary economics can give us guidance about the role of energy in the economy and suggest some policies to perhaps ensure a smoother transition to non-fossil fuel energy sources.

3. ENERGY AND ECONOMIC GROWTH

Among world leaders, economic growth is seen as the unique source of present and future well-being. Since energy is crucial to economic growth, one would think that its role should be an integral part of all growth theories. However, energy fails to appear as a variable in most mainstream models of economic growth. The traditional Solow model explains economic growth through the changes in two factors of production: capital (K) and labor (L). According to the model, output (Q) is generated through a production mix of K and L, whose physical properties are described by a Cobb–Douglas production function:

$$Q = AK^{\alpha}L^{\beta}, \tag{1}$$

where α and β are the partial output elasticities (and output shares) of K and L, respectively. In this model, one factor of production can potentially be completely replaced by another. Empirical estimations of the Solow model failed to explain most of the variation in economic growth. The part of the growth unaccounted for by capital and labor (approximately 87% of the total) was attributed by Solow to technical progress: The "Solow residual" (A in Eq. 1) represents a "catch-all" category dubbed technological change. Under the pressure of criticism from within and from outside the field, economists have developed more sophisticated growth models.

Recent models are based on endogenous growth and consider a broader array of relevant variables, including the Solow residual (hence the endogenous character of the theory), a research and development component, human capital, and natural resources. Even though the models possess a greater explanatory power, they still assume potential perfect substitution between all inputs. Moreover, in empirical estimations, a large part of the variation in economic growth remains unexplained.

To resolve some of the problems encountered in more conventional models, and to anchor them in biophysical reality, Charles Hall and colleagues developed economic growth models including energy as a factor of production. Hall's approach is to include energy in the production function as a full-fledged factor of production in a traditional Solow model. However, he explicitly rejects the assumption of perfect substitution because the unlimited substitution of energy by another form of capital would clearly violate the laws of thermodynamics. In a biophysical sense, energy cannot disappear during the production process (as the first law states) and economic production cannot take place without systematic inputs of low-entropy energy and subsequent dissipation of high-entropy energy (as the second law states). Hall's model successfully reproduced economic growth over a 30-year period for three industrialized countries—Japan, the United States, and West Germany. For Japan and West Germany, the model leaves unexplained less than 10% of the variation in output, even though the two countries underwent significant structural changes as a consequence of the two oil crises in the 1970s. For the United States, 98.3% of the variability in production is explained by the variables for the 1960–1977 period. The results are less conclusive for the 1978–1993 period. However, in this model, technological change is accounted for by a simple variable (creativity) that fails to capture the complex structural changes that resulted from the rapid development of energy-conservation technologies beginning in the 1970s. Hall's results clearly illustrate the prominent role played by energy in economic production.

A model by Ayres and Warr focuses on the concept of exergy, a measure of the conversion of energy into useful work. Exergy measures how much physical work can potentially be done by a system. Unlike energy, exergy is not conserved: It is used up by the system. Ayres and Warr introduce a measure of "useful work" (U) in the Solow model and they find that, since 1900, improvements in energy conversion into physical work explain most of the variation in technical progress; the unexplained residual is only 12%.

These findings emphasize the importance of the role of energy in economic growth. To a large extent, they constitute empirical evidence of the theoretical work of White, Cottrell, and other earlier writers on the role of energy in economic and cultural evolution. They also reflect an intuitive sense that increased energy use and efficient energetic conversion lead to increased growth and development. Figure 1 shows a clear correlation between gross domestic product (GDP) and energy consumption in the United States. The problem of causality remains, however, since the figure does not indicate whether increased GDP leads to increased energy use or vice versa. Cutler Cleveland and colleagues demonstrated, taking into account energy quality, that energy use is indeed the independent variable and economic growth the dependent one. Growing energy use and improving energy efficiency fueled the growth in GDP and provoked dramatic changes in cultural evolution.

Figure 2 illustrates the significant differences between developing and advanced nations in terms of energy use per capita. Industrialized countries have undergone major technological developments and a major improvement in their standard of living during the past 150 years. They are also the countries with very high levels of energy use. Of course, a contributing factor in the energy/GNP relationship is the ability of high-income countries to keep the energy flowing to them. Figures 1 and 2 suggest an essential relationship between energy use and economic activity—a relationship difficult to determine using traditional economic models.

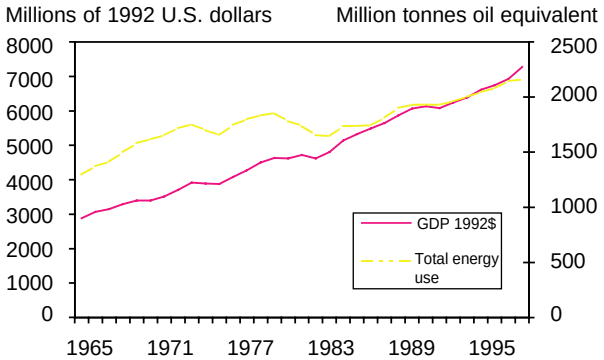


FIGURE 1 Evolution of real GDP and total primary energy use in the United States, 1965–1997. Data from the U.S. Department of Commerce and "BP Statistical Review of World Energy."

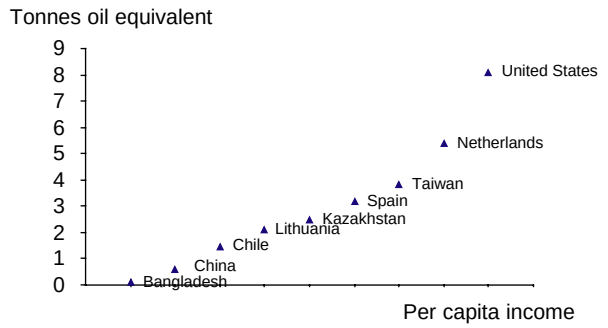


FIGURE 2 Energy use per capita in tonnes oil equivalent for selected countries, 2000. Data from the Energy Information Administration and "BP Statistical Review of World Energy."

4. ENERGY AND EVOLUTIONARY ECONOMICS

Evolutionary economics is not a well-defined field of economic theory. Economists who call themselves "evolutionary" hold a wide variety of views about the pace of evolutionary change, the nature of economic equilibrium, and the role of prices in economic theory and policy. Nevertheless, a number of useful concepts from evolutionary economics can be applied to the study of the role of energy in the human economy. Most of the debate within economics about resources and the economy has taken place within the framework of neoclassical growth theory, which merely extends the optimal factor allocation framework to the allocation of factor growth rates. In contrast, an evolutionary framework can examine discontinuous change, exogenous as well as endogenous factors, and biophysical limits. Three useful concepts from evolutionary economics are the notion of punctuated equilibrium, the existence of multiple equilibria, and the phenomenon of self-organization.

Punctuated equilibrium as a description of the fossil record was first proposed by Niles Eldredge and Stephen Jay Gould in 1972. They argued that evolution proceeds not only by gradual changes at the margin but also by quantum leaps from one form to another. Therefore, the evolutionary history of a particular species is characterized by periods in which it remains virtually unchanged, punctuated by abrupt events leading to a rapid transformation into another species. As discussed previously, the human epoch can be seen as a series of fairly stable regimes interrupted by periods of rapid change. The major transitions in human physical evolution 2.4 million years ago (*Homo habilis*), 1.9 million years ago (*H. erectus*), and 100,000 years ago (*H. sapiens*) were abrupt in terms of the fossil record. Likewise, human cultural evolution is also characterized by abrupt changes from one technology to another. The transition to agriculture was essentially complete within in a few thousand years and the transition to the industrial age was even more rapid. Georgescu-Roegen identified three "Promethean" technologies as fundamental qualitative changes in human history: fire, agriculture, and the steam engine. These were fundamental bioeconomic mutations that redefined the human presence on the planet.

What caused these punctuations? Increasing evidence suggests that evolutionary changes, including human evolution, have been driven by abrupt climate events. Good evidence exists linking the origin of the genus *Homo* to the onset of glaciation. The appearance of *H. erectus* coincided with a dramatic change to a cooler and drier climate in Africa. The transition to agriculture that occurred throughout the world after the end of the last ice age ushered in a warmer and uniquely stable climate. The industrial age coincided with an unusually warm period in Europe. Another punctuation will likely occur in the near future as the temperature of the planet adjusts to the massive pumping of carbon into the atmosphere by the burning of fossil fuels.

Evolution involves two critical features: variation and selection. It is increasingly accepted that Darwinian selection can occur at several levels, from the gene to the organism and the group. In contrast to the evolution of most other species, human traits may be culturally selected through various systems of rewards and punishments. This accounts for the incredible variety of human cultures. The existence of cultural "filters" for human traits means that institutions and institutional change are critical to economic evolution. The availability and use of energy sources depend on consumer attitudes, history and historical lock-in, as well as political intrigue. The experience of the energy price shocks of the 1970s holds two particularly valuable lessons. First, these shocks demonstrated the vulnerability of the global economy to changes in energy availability. Second, they showed the ability of markets to substitute energy efficiency and other sources of energy for petroleum.

The ability of the economy to adjust to energy shortages should not be overestimated. The short-term energy disruptions of the 1970s and early 1980s caused much economic damage and, being politically driven, they did not signal impending fossil fuel exhaustion. Another lesson is that the ability of policy intervention to change patterns of energy consumption should not be underestimated. Policies

promoting energy efficiency played a major role in softening the impact of the energy shocks, especially in Japan and Europe. The energy crises demonstrated, as careful commentators have been noting for more than a century, that energy is not just another commodity. It is essential to the production process. Cheap energy has been the ultimate source of the unprecedented economic growth of the past 150 years and the transition to a non-fossil fuel economy will be another punctuation in human history. The most important by-product of energy use, increased carbon in the atmosphere, will most likely cause another punctuation as the climate continues to respond to the perturbation caused by the fossil fuel age.

Multiple equilibria also characterizes the evolution of the human economy. Ecologists tell us that there is nothing inevitable about the evolution of a particular ecosystem. Small changes can cause an ecosystem to change to another equilibrium that might look quite different than the original one. Ecologists no longer accept the notion of a "climax forest," an inevitable and unique final form that would automatically evolve in a given area. Natural systems show that many equilibrium states are possible with any given set of initial conditions. Likewise, the market economy—contrary to the model of welfare economics that implicitly assumes a single, unique, global equilibrium—can have a large variety of stable equilibria. Human social systems have even more possibilities than natural ecosystems since cultural systems of rewards and punishment can create any conceivable social equilibrium. Also, there is no reason to say that one system is better than another.

The implications of multiple equilibria for energy use are both positive and negative. On the one hand, change can be abrupt and unpredictable. On the other hand, human institutions, including the market, can adjust rapidly to suddenly changing parameters. The global economy can most likely adjust to new energy shocks and probably even to the end of the fossil fuel age, but what the next economic system will look like is anyone's guess. Likewise, many surprising outcomes are possible as a result of the current episode of global climate change. The planet could continue to warm gradually or suddenly change to another state (much colder or much warmer).

In terms of policy choices, the existence of multiple equilibria implies that we need not be afraid of moving away from the one true optimum. The notion of a unique and stable Pareto optimum toward which the economy naturally moves is an illusion. Likewise, Pareto efficiency in allocation is a goal chosen by politicians and abetted by neoclassical economists, but there is nothing scientific or sacred about it. As Daniel Bromley argues, efficiency as the most important policy goal is an ideology, not a science-based conclusion. Rather than searching for the one true optimum, a more fruitful research project would be determining what categorizes different kinds of equilibria. What keeps a system in equilibrium in the face of fairly severe changes in the environment? Can we anticipate when a system is about to change to another state?

Self-organization refers to the process of the emergence of order in organic systems. It is a development process that is cumulative, discontinuous, irreversible, and creative. It incorporates processes at one level and creates new structures at higher levels. According to Stuart Kaufman, much of the order we see in the biological world may not be the result of natural selection at all but rather the spontaneous order of self-organized systems. Likewise, the (usual) order we see in economic systems may not be the result of the invisible hand of market selection but rather a spontaneous order of self-organized systems having little to do with prices or efficiency. The relationship between self-organization and thermodynamics is obvious. Self-organizing systems prosper by drawing down energy and information from surrounding systems. The global market economy expands by drawing down stocks of low-entropy fossil fuels, stocks of renewable resources such as forest and ocean systems, and, arguably, human cultural diversity.

The fossil record shows that the evolution of species is driven by interplay between internal and external forces. On the one hand, species compete according to the laws of Darwinian natural selection in self-organizing and adaptive systems. On the other hand, evolutionary outcomes are sometimes dominated by rapid and unpredictable external changes. Evolutionary systems cannot be understood without an appreciation of the interplay between external and internal, micro and macro, order and disorder. Like biological systems, the economy is an evolving complex system buffeted by constant changes in boundary conditions.

The current self-organization of the world economy is driven by the world's dowry of fossil fuels. Increases in economic activity mean increases in energy use, which also increase the dissipation of high-entropy energy, putting irreversibly more pressure on the natural environment and speeding up the emergence of scarcity in both renewable (e.g., via

pollution) and nonrenewable resources (via depletion and deterioration). Therefore, the increased dependence on fossil fuels may eventually limit the potential for diversity. As fewer resources are available for human beings to harness, control, and transform, fewer potential developmental paths are available. With fewer developmental paths to choose from, evolution becomes limited. To a large extent, evolution depends on diversity and variation, and if diversity is low, sorting mechanisms such as natural selection have less raw material with which to work.

5. CONCLUSION

The dependence of the market economy on fossil fuel energy, together with the fact that fossil fuels are being rapidly exhausted, paints a disturbing picture of the human prospect. Added to this is the fact that the major punctuations in human physical and cultural evolution have apparently been driven by climate change—occurring today because of carbon emissions from fossil fuels. On the other hand, our evolutionary history shows that we are an adaptable species, capable of making drastic changes in lifestyles and technology to meet new challenges. From an evolutionary perspective, market economies are not as stationary or as fragile as the standard economics model shows them to be. The global economy withstood a number of monumental shocks and it eventually adjusted to the new conditions. Also, there is no reason to believe that a "free" market equilibrium is superior. A number of stable states are possible, and a number of policy tools are available to make the adjustment to the end of the fossil fuel age smoother. We should not hesitate to use them out of fear of disturbing a "natural" free market system. The end of fossil fuels does not mean the end of civilization.

The technology already exists to enable us to move to a non-fossil fuel economy. With a little help from sound energy policy initiatives, the prospects for changing to another Promethean technology are good. Technological adjustment need not mean using ever more sophisticated technologies to hang on to the bitter end to an unsustainable system.

SEE ALSO THE FOLLOWING ARTICLES

Business Cycles and Energy Prices • Cultural Evolution and Energy • Economic Growth and Energy • Economic Thought, History of Energy in • Entropy and the Economic Process • Exergoeconomics • Exergy • Prices of Energy, History of • Transitions in Energy Use

Further Reading

Adams, R. (1981). Natural selection, energetics, and cultural materialism. *Curr. Anthropol.* **22,** 603–608.

Ayres, R., and Warr, B. (2004). Accounting for growth: The role of physical work. *Struct. Change Econ. Dynamics*, in press.

Ayres, R., d'Arge, R., and Kneese, A. (1970). "Aspects of Environmental Economics: A Material Balance, General Equilibrium Approach." Johns Hopkins Univ. Press, Baltimore.

Bergh, J. C., and van den, J. M. (2002). Evolutionary analysis of economic growth, environment and resources. *In* "Scarcity and Growth in the New Millennium" (R. Ayres, D. Simpson, and M. Toman, Eds.). Resources for the Future, Washington, DC.

Bromley, D. (1990). The ideology of efficiency: Searching for a theory of policy analysis. *J. Environ. Econ. Management* **19,** 86–107.

Calvin, W. (2002). "A Brain for all Seasons: Human Evolution and Abrupt Climate Change." Univ. of Chicago Press, Chicago.

Cleveland, C. (1999). Biophysical economics: From physiocracy to ecological economics and industrial ecology. *In* "Bioeconomics and Entropy: Essays in Honor of Nicholas Georgescu-Roegen" (K. Mayumi and J. Gowdy, Eds.). Elgar, Cheltenham, UK.

Daly, H. (1977). "Steady State Economics." Freeman, San Francisco.

Georgescu-Roegen, N. (1971). "The Entropy Law and the Economic Process." Harvard Univ. Press, Cambridge, MA.

Hodgson, G. (1993). "Economics and Evolution: Putting Life Back into Economics." Univ. of Michigan Press, Ann Arbor.

Martinez-Alier, J. (1987). "Ecological Economics: Energy, Environment and Society." Blackwell, Oxford, UK.

Potts, J. (2000). "The New Evolutionary Microeconomics." Elgar, Cheltenham, UK.

Stern, D. (2000). A multivariate cointegration analysis of the role of energy in the U.S. macroeconomy. *Energy Econ.* **22,** 267–283.

Exergoeconomics

ENRICO SCIUBBA
University of Roma 1 "La Sapienza"
Rome, Italy

1. **Introduction**
2. **The Concept of Exergy**
3. **The Economics of a Commodity-Producing Process**
4. **Second vs First Law Optimization**
5. **Structural Representation of Energy Conversion Systems**
6. **Thermoeconomics**
7. **Exergoeconomics and Extended Exergy Accounting**
8. **Conclusions**

Glossary

commodity In economics, a physical object produced by a man-made process. In this article, commodity also includes dematerialized services.

objective function In optimization procedures, a mathematical expression depending on process parameters, whose extremum (minimum or maximum) corresponds to the "most convenient" combination of production factors for the assigned output.

Exergoeconomics (ee) is the branch of applied thermodynamics that deals with second law-based process synthesis and optimization methods. Per se, EE does not include optimization techniques: It provides the objective function, but the actual optimization of process performance (i.e., the achievement of the specified output with the most convenient combination of production factors) is carried out with one of the available mathematical techniques. Therefore, this articles does not deal with the numerical procedures that lead to the optimized system but rather describes and discusses the engineering background of the methods that lead to the formulation of the objective function.

1. INTRODUCTION

Exergoeconomics (EE) is a relatively new technique. It originated in the 1960s, first in Europe, where Wolfgang Fratscher (Germany), Jan Szargut (Poland), and Valeriy Brodyanskii (USSR) published papers on the "economic value of the concept of exergy," and then in the United States, where Myron Tribus, Yehia El-Sayed, and Robert Evans published a series of papers on a mathematical cost-optimization procedure based on exergy (which they called availability). The term thermoeconomics (TE) was first used in Tribus's MIT course notes in 1960 and in Evans' doctoral thesis in 1961. The concept was developed both in breadth and in depth in academic circles, but it did not impact process engineers, industrial managers, and energy planners until approximately 20 years later, mostly due to the efforts of El-Sayed, Richard Gaggioli, Tadeusz Kotas, and Michael Moran in the early 1980s and Antonio Valero and George Tsatsaronis in the late 1980s. In a series of papers from 1986 to 1989, Valero derived a general and completely formalized costing theory to calculate the exergetic cost of a product from the exergetic input into the process and the structure of the production process. In the 1990s, TE was extended (mostly by Christos Frangopoulos and Michael von Spakowsky) by adapting it to off-design conditions and time-dependent problems and by including environmental effects. The literature on TE is abundant, and a comprehensive review has not been attempted. Regarding EE, there is a qualitatively very interesting debate, expressed by the works of Valero, Göran Wall, Szargut, and Enrico Sciubba, that seems to lead to an exergetic theory of value on which to base societal planning (especially with respect to the much-debated issues of sustainable-development and limited-growth models).

In this article, the concept of exergy and cost optimization is briefly discussed, followed by separate treatments of TE and EE. A caveat is necessary: Some authors, most notably Tsatsaronis and von Spakowsky, use TE and EE as synonyms and consider irrelevant whether the optimization is carried using money or exergy as a quantifier. Here,

 © 2004 Elsevier Inc. All rights reserved.

TE is considered as a special case of EE, and treatment of these is developed with this guideline.

2. THE CONCEPT OF EXERGY

Energy is considered an a priori concept in thermodynamics: We are taught that "energy is an extensive property of a system such that its change in value between two states is equal to the adiabatic work between these states." However, energy is known to manifest itself in several forms, each endowed with a different quality. The concept of quality can be fully explained only by referring to the second law, but in general the quality of an energy flux is a measure of its capacity to cause change. We can identify "ordered" or "high-quality" energy forms (potential, kinetic, mechanical, and electrical) that can be ideally converted into each other with 100% efficiency so that even after a number of transformations, as shown in Fig. 1A, the output flux of useful energy, En_{out}, is quantitatively equal to the flux of useful input, En_{in}. However, other forms (internal energy, chemical energy, thermal radiation, and turbulent kinetic energy) cannot be converted into each other, much less into high-quality energy, without an intrinsic efficiency loss, so the useful output En_{out} is now always lower than the useful input En_{in} (Fig. 1B). Since energy is conserved (this is a basic postulate in all physical sciences), the difference $En_{in}-En_{out}$ must be accounted for: Indeed, we invariably find that this difference has been dissipated into a low-temperature flux absorbed or provided by the environment. The "environment" is meant here in its usual layman sense—that is, as a system so large that its properties (pressure, temperature, chemical composition, kinetic or potential energy, etc.) are not affected by the interaction with the man-made system. These considerations, though formally incomplete from a scientific standpoint, convey the correct information in an easily understandable form.

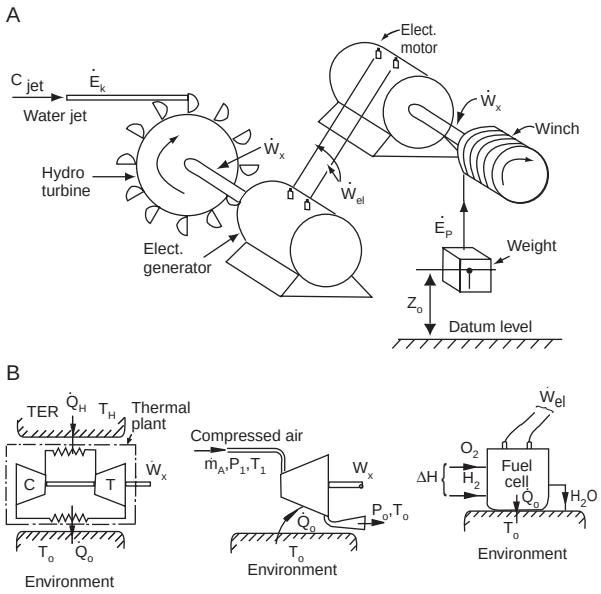


FIGURE 1 (A) A chain of ideal energy conversions of high-quality energy forms into each other. (B) Examples of conversion of low-quality into high-quality energy forms.

The concept of exergy has been developed to provide a congruent and coherent quantification to the quality of an energy form. Imagine that at a certain time t_o we place a system S identified by a certain set of thermodynamic properties (V, z_B, p, T, c, m, μ, etc.) in contact with the environment E and arrange things in such a way that S may exchange mass and energy only with E (Fig. 2). Our intuition tells us and experiments confirm that after some time t_{relax}, which depends on the extension of S and on the difference between its initial properties and those of E, S will come to equilibrium with E; that is, even possible dynamic interactions will occur at a macroscopically stationary state and will result in no net change in the mass and energy contents of either system. Exergy is defined as the maximum work developed in this ideal process. Since the interaction is between E and S, and we assume that all processes can be described by a succession of quasiequilibrium states (so that the time interval t_{relax} drops out from the problem formulation), exergy will be a function only of the initial and the final state of S (the final state being equal to that of the environment) so that it is an attribute of the pair (S,E) and not of S alone. S and E exchange energy fluxes under different forms: kinetic (all parts of S in relative motion with respect to E come to rest), potential (the barycentric position of S reaches a fixed elevation in the gravity field of E),

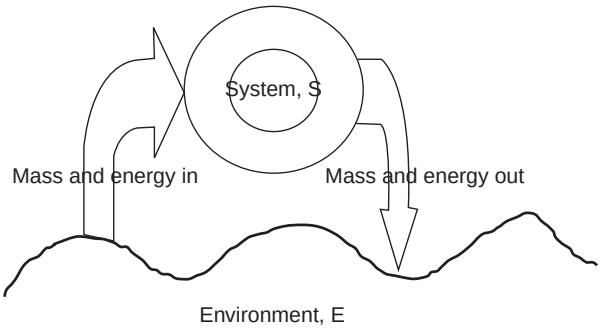


FIGURE 2 A system *S* and its environment *E*.

thermal (heat flows from S to E or vice versa depending on the initial temperature difference $T_S - T_E$), work interaction (E performs or receives work from its boundary interactions with S), and mass exchange (mass fluxes from S to E and vice versa carry different energies). Let us consider here only stationary interactions between S and E (this is the so-called control volume approach) in which flows of energy and matter from S to E are continuous and (on average) steady in time. By applying energy and entropy balances to the complex system composed of S and E, we obtain

$$\sum_j \left(1 - \frac{T_E}{T_j}\right) Q_j - W + \sum_{\text{in}} m_i e_i - \sum_{\text{out}} m_{\text{out}} e_{\text{out}} = E_\lambda, \quad (1)$$

where the first term represents the thermal exergy flow, the third and fourth the exergy in- and outflows due to material exchanges, and E_λ is the exergetic destruction. In Eq. (1), we make use of the definition of specific exergy:

$$e = h - h_E - T_E(S - S_E) + \Sigma_k(c_k \mu_k - c_E \mu_{k,E}) + 0.5V^2 + gz. \quad (2)$$

From Eqs. (1) and (2), we can draw several conclusions:

1. The initial kinetic and potential exergies of S are entirely "recovered" into useful work; in fact, they might be included in the enthalpic term by considering a baroclinic total enthalpy, $h_{\text{tot}} = h + 0.5V^2 + gz$.

2. Work exchanges between S and E are entirely recovered in the exergy balance. This derives from the fact that work is exergy and, as a consequence, the kinetic and potential energy contents are entirely equivalent to exergy. If we define a quality or exergetic factor as the ratio between the energy and the exergy content of an energy flux, and since electrical energy is equivalent to work, all these high-quality forms are seen to have an exergetic factor equal to 1.

3. Thermal energy has an exergetic factor equal to its associated Carnot efficiency: This reflects the fact that, under the given assumptions, the maximum work that can be extracted from a quantity of heat Q available at a certain temperature T is equal to $W_Q = (1 - T_E/T)\ Q$.

4. If S is initially at a lower temperature than E, thermal energy will flow from E to S, with a corresponding exergy flow equal to $E_Q = (1 - T/T_E)\ Q$. This amount is always positive.

5. For a simple substance whose states are defined by temperature and pressure (e.g., most gases for most practical engineering purposes), exergy is always positive, except for particular combinations of p_S/p_E and T_S/T_E.

6. The chemical potential of the elements in S cannot be entirely converted to work: Since S must come to equilibrium with E, the most we can recover for the generic kth component is the difference between the values of its Gibbs function in S and in E, each weighted by the respective concentration. In other words, we can ideally transform only a portion of the initial ΔG_S into useful work because the products of reaction must be at their respective environmental concentrations and chemical potentials at T_E and p_E.

7. Exergy is not conserved: In every real process there is an exergy destruction caused by irreversibility effects. In the simplest case (work $\Rightarrow$ work transformation), this exergy destruction is equal to the irreversible entropy generation due to frictional effects. This can be seen by considering that S exchanges only adiabatic work with E, and the loss is equivalent to frictional heat. If there is only one medium participating in the exchange, from Eqs. (1) and (2) we obtain

$$\begin{aligned} E_\lambda &= -W + m(e_S - e_E) \\ &= -W + m[h_S - h_E - T_E(s_S - s_E)] \\ &= -T_E(s_S - s_E)[W]. \end{aligned}$$

Since $s_S - s_E$ must be negative (S may spontaneously evolve only toward higher entropies), the destroyed exergy flux E_λ is positive and equal to the "lost work" defined by Gouy and Stodola.

8. From Eqs. (1) and (2), we obtain the general exergy equation for flow systems:

$$e_S = e_{\text{ph}} + e_{\text{ch}} + e_{\text{k}} + e_{\text{p}} + e_j[\text{J/kg}], \quad (3)$$

where the suffixes indicate the usual denomination of the various terms: physical, chemical, kinetic, and potential exergy, respectively. If other forms of energy fluxes are involved (magnetic, nuclear, etc.), one needs only to identify the suffix j and apply the corresponding definition.

9. Exergy is an extensive property so that $E_{(S1+S2)} = E_{S1} + E_{S2}$.

10. If two systems S_a and S_b interact, the maximum work obtainable from their exclusive interaction is $E_a - E_b$. Furthermore, if a stream a undergoes a series of transformations $i, j, k, \ldots, z$ in which it receives the exergies $E_i \ldots E_z$, its final exergy content is the sum of $E_a + E_i + E_j + E_k + \ldots E_z$. Exergy is thus an additive property. For engineering

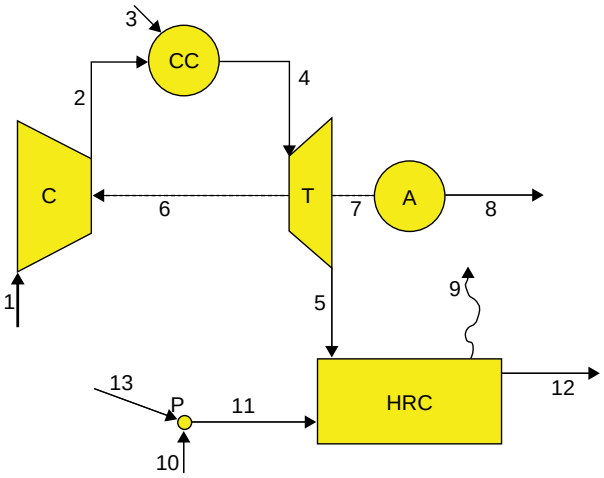


FIGURE 3 Schematic representation of a cogeneration plant.

(and EE) calculations, it is convenient to define a standard environment to which all exergy values can be referred, but this is by no means necessary, as Gaggioli and coworkers have shown.

Practical applications of the exergy concept are important: It can be shown that a second law efficiency defined on the basis of exergy fluxes provides more information than the first law efficiency based on energy fluxes. For example, consider the cogeneration plant shown in Fig. 3. From the design data given in Table I, the efficiencies shown in Table II can be calculated and the following considerations apply:

1. First law efficiency attains a value of 0.655: This means that 65% of the fuel chemical energy is converted into useful energy output. This is misleading, however, because it does not account for the temperature-dependent quality of the heat flux Q_{cog}. A exergetic analysis, in which all inputs and outputs are expressed in exergetic terms, provides an η_{II} of 0.347, correctly indicating that a substantial portion of the chemical exergy of the fuel has been "degraded" into heat (i.e., into a low-quality energy form).

2. Component exergetic efficiencies (η_{II}) also provide interesting information, especially if compared with their energetic counterparts (η_I). For example, whereas the quasi-adiabatic compressor and turbine display similar values for η_I and η_{II}, both the heat-recovery boiler and the condenser have $\eta_{II} < \eta_I$. This indicates that it is useless to invest excessive resources to improve boiler performance because this component is very efficient in transmitting the heat flux it is required to transmit

TABLE I

Design Data for the Cogeneration Plant Shown in Fig. 3

$T_{env} = 300\,K$	$p_{env} = 0.1\,MPa$	$\Delta p_{2-4} = 0.02\,p_2$	$\eta_{p,C} = 0.85$
$T_1 = 300\,K$	$p_1 = 0.1\,MPa$	$\Delta p_{5-9} = 0.08\,p_9$	$\eta_{p,T} = 0.9$
$T_4 = 1300\,K$	$p_2 = 0.8\,MPa$	$\Delta p_{11-12} = 0.1\,MPa$	$\eta_{el} = 0.98$
$T_9 = 353\,K$	$p_9 = 0.1\,MPa$		$\eta_{mec} = 0.95$
$T_{10} = 673\,K$	$p_{10} = 0.1\,MPa$	$c_p = 1.1\,J/(kgK)$	$\eta_{CC} = 0.98$
	$P_{11} = 1\,MPa$	$R_{air} = 287\,J/(kgK)$	$\eta_P = 0.8$
		$\kappa = 1.4$	$\eta_{HRB} = 0.9$

TABLE II

Calculated Process Parameters for the Cogeneration Plant Shown in Fig. 3

$T_2 = 603\,K$	$p_4 = 0.798\,MPa$	$\eta_{1,C} = 0.85$	$\eta_{2,C} = 0.76$
$T_5 = 777\,K$	$p_5 = 0.108\,MPa$	$\eta_{1,T} = 0.9$	$\eta_{2,T} = 0.87$
$T_{12} = 403\,K$		$\eta_{1,CC} = 0.94$	$\eta_{2,CC} = 0.59$
		$\eta_{1,HRB} = 0.81$	$\eta_{2,HRB} = 0.68$
$\eta_{plant,el} = 0.2$			
$\eta_{1,plant} = 0.658$	$\eta_{2,plant} = 0.34785$		

($\eta_{I,B} = 0.81$). The boiler low exergetic efficiency is intrinsic and not due to design errors: It degrades a high-temperature, high-exergy heat flux into a low-temperature, low-exergy one. Similar considerations apply to the combustor: It is the degradation of the relatively high-quality chemical energy into heat, and not its energy-transfer capacity, that lowers its exergetic efficiency: Energywise, the performance of a combustor is more than satisfactory (≈ 0.94). This is the rationale for the use of fuel cells, which exploit the ΔG of the fuel by means of chemical reactions, without combustion.

3. THE ECONOMICS OF A COMMODITY-PRODUCING PROCESS

Consider a process that produces a certain commodity (a good or a service) (Fig. 4), and assume that both its internal structure and its operating parameters are known so that all of its outputs **O** can be calculated if all of the inputs **I** are known (the transfer function Π of process P is known, where $\mathbf{O} = \Pi \times \mathbf{I}$). To produce a cumulative amount O_1 of the commodity over a period of time T, a certain amount of inputs $I_1, I_2, \ldots, I_N$ are used together with a certain amount of process exergy E_{in}. In general,

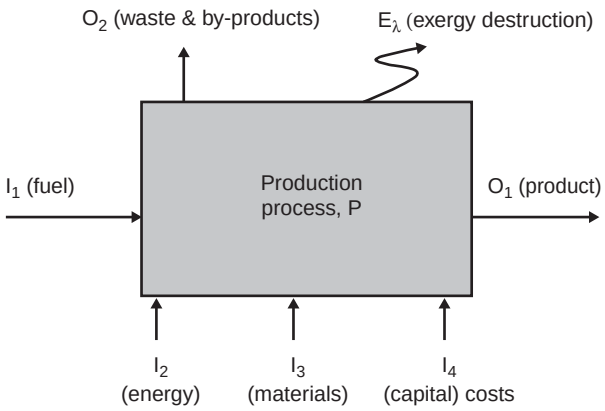


FIGURE 4 A commodity-generating process.

some waste O_3 and some by-product O_2 are also produced. Two additional production factors, namely labor and capital, must also be accounted for. From the process mass and energy balances we can compute the amounts of $m_{I1}, m_{I2}, \ldots, m_{IN}$ and the corresponding exergies $e_{I1}, e_{I2}, \ldots, e_{IN}$ necessary to produce one unit of the desired commodity O_1 (1 kg, one piece, or 1 kJ, depending on the type of product). We can also compute the unit expense for labor (man-hours for unit output) and for capital (€ of rate of return on investment in the equipment per unit output). Over the time period t_p (usually 1 year), we can thus compute O_1 production cost as the ratio of the sum of all expenses (materials, exergy, labor, and capital) to the cumulative amount of production of O_1, both computed in t_p:

$$c_{O_1} = \frac{Cumulative\ expenses}{Cumulative\ production} = \int_{t_p} \left(\frac{C_I + C_E + C_L + C_C}{O_1} \right) dt \quad [€/kg,\ €/unit,\ or\ €/kWh] \tag{4}$$

where O_1 is measured in proper units (kg/s, units/s, or kW). Such a production cost is dependent on both the structure of the production line and its efficiency. If we improve the efficiency, we can obtain a larger amount of O_1 with the same amount of inputs or, conversely, the same O_1 with a reduced input. If we decrease the number of components in the plant, we can lower the capital expenditures for the same amount of output. In practice, the functional relation between efficiency and capital cost is such that higher efficiency components have a higher initial cost, and thus lead to higher capital expenditures. However, the relationship is not linear, and this indicates the possibility of performing process optimization (i.e., searching for process structures and/or operating parameters that lead to a lower production cost C_{O_1}). A whole branch of engineering, called engineering economics, deals with these issues: We can explore some of its implications by examining a simple example consisting of an elementary energy conversion device, a heat exchanger.

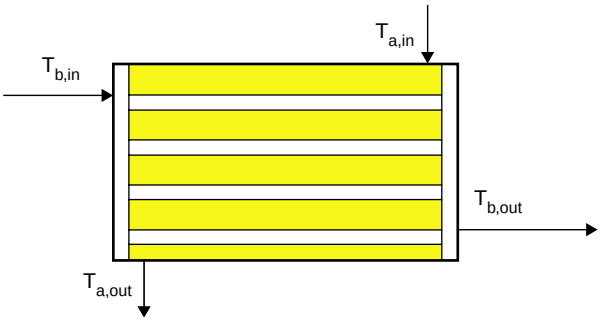


FIGURE 5 Schematic representation of a heat exchanger.

The function of a heat exchanger is to transfer a quantity of energy Q to a medium at lower temperature. In Fig. 5, fluid a (the heat donor) transfers a portion of its sensible heat, $m_a c_{pa} (T_{a,in} - T_{a,out})$ to fluid b (heat acceptor), which increases its temperature from $T_{b,in}$ to $T_{b,out}$. The process can be described by the following equations:

$$\text{Energy balance}: m_a c_{pa}(T_{a,in} - T_{a,out})\eta_{exchanger} = m_b c_{pb}(T_{b,out} - T_{b,int}) \tag{5}$$

$$P_{pump,tot} = \frac{m_a \Delta p_a}{\rho_a \eta_{pump,a}} + \frac{m_b \Delta p_b}{\rho_b \eta_{pump,b}} \tag{6}$$

$$\text{Heat transfer at the interface}: Q_{a\to b} = m_b c_{pb}(T_{b,out} - T_{b,int}) = UA_{HE} LMTD \tag{7}$$

In Eq. (7),

$$LMTD = \frac{(T_{a,in} - T_{a,out}) - (T_{b,in} - T_{b,out})}{\ln\left(\frac{(T_{a,in} - T_{b,out})}{(T_{a,out} - T_{b,in})} \right)}$$

is the log-mean temperature difference calculated under the following assumption: For fixed mass flow rates, the amount of heat transferred from a to b increases with LMTD. Since $T_{a,in}$ and $T_{b,out}$ are specified, this means that a smaller terminal temperature difference, $TTD = T_{a,in} - T_{b,out}$, is desirable. However, to lower the TTD more, heat exchange (HE) area must be used, and this leads to an increase in the pressure losses Δp_a and Δp_b in the heat exchanger passages and to a higher pump work. Here, Reynolds effects on the overall heat-exchange coefficient U and on the friction factor f are neglected, both of which are assumed to remain

fixed throughout. The design problem can be rephrased as follows:

1. The effectiveness of the HE increases with an increase in the active HE surface so that more output $[m_b \times c_{p,b} \times (T_{b,out} - T_{b,in})]$ is generated by the same input $[m_a \times c_{p,a} \times (T_{a,in} - T_{a,out})]$.
2. The parasitic losses that the pumps and fans must compensate for also increase with the active HE surface so that a proportionally larger amount of electrical energy must be purchased.

Let us formulate a costing problem on the following bases: The cost of the unit of heat acquired by stream b is given by the rate of cash outflow c_{HE} corresponding to the installed cost C_{HE} of the exchanger (for simplicity, only the active area A_{HE} is considered here) plus the cost of stream a c_a plus the cost of the electricity c_{el} needed to power the pumps. The following expressions may be used:

$$c_{HE} = CRFC_{HE} \tag{8}$$

$$\mathrm{CRF} = \frac{i(1+i)^N}{(1+i)^N - 1}, \tag{9}$$

where $i = 0.05$ is the yearly interest rate and $N = 15$ years is the assumed technical life of the HE:

$$C_{HE} = K_{HE}\left(\frac{Q_{a\to b}}{U \times \mathrm{LMTD}}\right)^x = K_{HE}A^x,$$
$$\text{with } K_{HE} = 2000\,€/m^2 \text{ and } x = 0.6. \tag{10}$$

where $K_{HE} = 2000\,€/\mathrm{m}^2$ and $x = 0.6$.

$$c_a = K_a \times m_a \times c_{p,a}(T_{a,in} - T_E) \tag{11}$$

$$c_{el} = 7.7\,¢€/kWh \tag{12}$$

$$\Delta p_b = f \times m_b^2, \text{ with } f = 0.01 \quad 1/(\mathrm{m} \times \mathrm{kg}) \tag{13}$$

where $f = 0.01[1/(m \times \mathrm{kg})]$. For a fixed c_a and assuming a zero cost $c_{b,in}$ for stream b at the entrance of the heat exchanger, the specific cost of the heating of stream b is given by

$$c_{b,out} = \frac{c_{HE} + c_{el}}{m_b} = \mathrm{CRF}\,\frac{K_{HE}A_{HE}^x}{m_b} + \frac{c_{el}P_{pump,tot}}{m_b}. \tag{14}$$

Since m_b increases almost linearly with A (Eq. 7), whereas P is proportional to the square of m_b (Eq. 13), the area-related cost term in Eq. (14) will decrease with an increasing heat exchange area and the pressure-loss term will increase linearly (because f is constant). Figure 6 depicts the situation: The total cost $c_{b,out}$ exhibits a minimum that can be found by optimizing, under the given assumptions and constraints, Eq. (14) with respect to A_{HE}. By manipulating the costing equations, the following

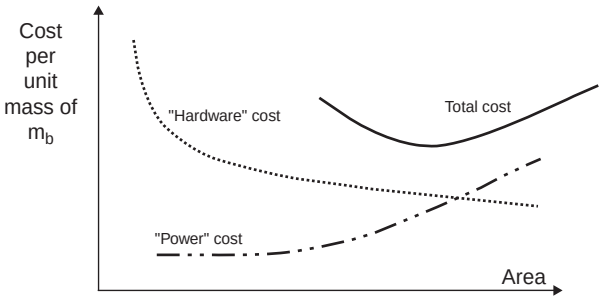


FIGURE 6 Determination of the optimal cost of the heated stream.

(approximate) expression is obtained:

$$c_{b,out} \approx C_{area} \times A_{HE}^{-0.4} \times C_{power,1} \times A_{HE}^{-1} + C_{power,2} \times A_{HE}^{2} \tag{15}$$

which derived w.r.t. A_{HE} and equated to 0 provides an expression for the optimal heat exchange area A_{opt}:

$$-0.4 \times C_{area} \times A_{opt}^{1.4} - C_{power,1} \times A_{opt}^{-2} + 2 \times C_{power,2} \times A_{opt} = 0 \tag{16}$$

where

$$C_{area} = \frac{\mathrm{CRF} \times K_{HE} \times c_{p,b} \times (T_{b,out} - T_{b,in})}{U \times \mathrm{LMTD}}$$

$$C_{power,1} = \frac{c_{el} \times m_a \Delta p_a \times c_{p,b} \times (T_{b,out} - T_{b,in})}{\rho_a \times \eta_{pump,a} \times U \times \mathrm{LMTD}} \textit{and}$$

$$C_{power,2} = \frac{c_{el} \times f}{\rho_b \eta_{pump,b}}\left[\frac{U \times \mathrm{LMTD}}{c_{p,b}(T_{b,out} - T_{b,in})}\right]^2.$$

The approximation here is due to the fact that the constants depend on LMTD and $(T_{b,out} - T_{b,in})$, which in turn vary with A.

The method of engineering economics is employed in process and design engineering to assess the solution to a design problem not only from a thermodynamic standpoint but also from an economic one. As demonstrated in the previous example, this method is based on the following assumptions:

1. Mass and energy balances exist for the process under consideration.
2. A suitable cost equation can be derived for all products of interest.
3. The coefficients in this cost equation may be expressed in terms of process parameters.
4. The specific cost of all incoming streams is known.

Of course, in real engineering practice problems are not so crisply posed as purported here. The

following are the most common difficulties encountered in industrial cost analysis:

1. Only approximate or incomplete mass and energy balances exist. In this case, good technical common sense must be used to approximately close all balances.
2. Cost equations are not known for all production factors (material, energy, labor, and capital), or they are known only for plants of a different scale. The missing data must be judiciously extrapolated from the existing ones or prudently estimated on the basis of known data in related fields.
3. The cost of incoming streams is not known since they originate from the environment or are cogenerated in the same process. For the flows of raw materials of unknown origin, usually transportation costs may be used as a good approximation. For water or air at ambient temperature, the cost is usually assumed to be equal to zero. In this case, however, a better choice is to charge the incoming stream with the (real or virtual) effluent treatment costs necessary to restore its original environmental quality.
4. Labor costs are site dependent so that the same process may have different economic optima in different locations. The location becomes a design parameter.
5. Capital costs (interest rate, social costs, taxation level, etc.) are site dependent. Furthermore, they are also time dependent. Economic optimization must be performed independently for each site. Regarding the time dependency, levelization methods exist that allow the engineer to average out both market fluctuations and changing financial conditions by referring the calculations to a fixed economic state of reference (current money and constant money formulations).
6. There are more than one useful outputs, and their production lines share some of the components, materials, labor, and energy inputs. Allocation decisions must be made about what share of capital, material, labor, and energy costs must be allocated to which product. The optimal point is substantially affected by the allocation criteria.

Engineering economics encompasses more complex topics than those described here: the treatment of cash flows, the allocation of equipment depreciation, the inclusion of inflation and taxes in the economic calculations, etc. Interested readers may consult the book by Bejan, Tsatsaronis, and Moran.

The limits of engineering economics, in the formulation presented here, depend on the fact that it entirely neglects second law considerations. The efficiencies that appear in its formulation are first law efficiencies and as such are incapable of assessing the thermodynamic quality of participating streams. This is particularly inconvenient when the inputs and/or outputs of the process under consideration have different qualities, as in the frequent case of combined or cogenerating power plants. In these cases, engineering economics provides the wrong answer: Both the optimum it suggests and the relative cost ratio between different products are incorrect. It is from a critical analysis of these weaknesses that TE was first formulated.

4. SECOND VS FIRST LAW OPTIMIZATION

Suppose we wish to find the optimal pressure ratio β for the turbogas of the cogeneration plant shown in Fig. 3. This is indeed a credible design task. An increase in β leads to higher efficiency, but also to a lower air mass flow rate, and for a fixed turbine inlet temperature T_4, it leads to a decrease in T_5 so that for a fixed water return temperature T_{10}, the amount of generated steam decreases as well. A decrease in β, on the other hand, reduces the efficiency of the gas turbine cycle and, for a fixed required power output, increases the mass flow rate at compressor inlet, produces a higher T_5, and produces a higher steam mass flow rate. Figure 7 displays the sensitivity of the plant outputs (P, Q_{12}, and E_{Q12}) to the variation of β. First and second law efficiencies are shown as well. It is clear that η_I and η_{II} peak for different values of β. Which objective function is correct—the one based on η_I or the one based on η_{II}? From a purely thermodynamic standpoint, η_{II} provides the correct solution, and the reason is apparent if we look at the energy and exergy flow diagrams of the process, which are called the Grassmann (for exergy) and

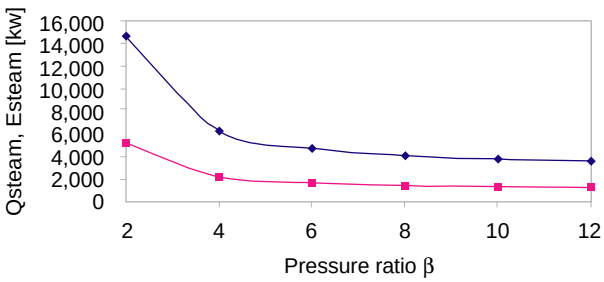


FIGURE 7 Cogeneration plant: sensitivity of the global performance indicators to β.

Sankey (for energy) diagrams (Fig. 8). These diagrams constitute a very effective pictorial representation of the energy and exergy flow within the process: They consist of boxes interconnected by arrows, with each arrow representing a stream and its width being proportional to the relative energy or exergy flux associated with it (usually normalized with respect to the main fuel input). Each box represents a component. For the plant under consideration, the energy diagram indicates that the losses, measured as percentage of the total input, are approximately the same for all components: 20% for the pump, 15% for the compressor and the boiler, 10% for the turbine, and 5 and 2% for the alternator and combustor, respectively. On the basis of this information, if we wanted to improve the design, we would concentrate on the pump first, then on the compressor and boiler, and so on. The diagram provides a ranking of the relative importance of each component in the energy dissipation during the process. The exergy diagram tells a different story, however: The highest losses (approximately 40%) are in the combustor and the boiler. This reflects more accurately the physics of the phenomena: In these two components, we enforce strongly irreversible transformations of high-quality into low-quality energy, irrevocably destroying exergy. A more accurate analysis (not performed here) would show that these low exergetic efficiencies are almost independent of the process parameters: They cannot be cured by improving, for instance, the thermal insulation of the combustor or by increasing the global heat transfer coefficient in the boiler. The irreversibility is introduced by the type of transformation these two components enact, and to remove the corresponding exergy destruction we would have to select a different device (e.g., a fuel cell instead of the combustor and a back-pressure steam turbine instead of the boiler). Thus, the proper objective function to use in the thermodynamic optimization is the one based on second law concepts.

Caution should be exercised here because energy conversion systems display a relatively high degree of interdependency: Components interact in complicated ways with each other, and an improvement in the efficiency of one component may lead to increased exergy destruction in other units, resulting in no net change or even a decrease in η_{II}. Furthermore, the installation cost of a component depends in a nonlinear way on its efficiency, and it may also affect the magnitude of other production factors, such as labor or material input. Therefore, an exergy analysis by itself does not tell the entire story. Methods are needed that combine exergy

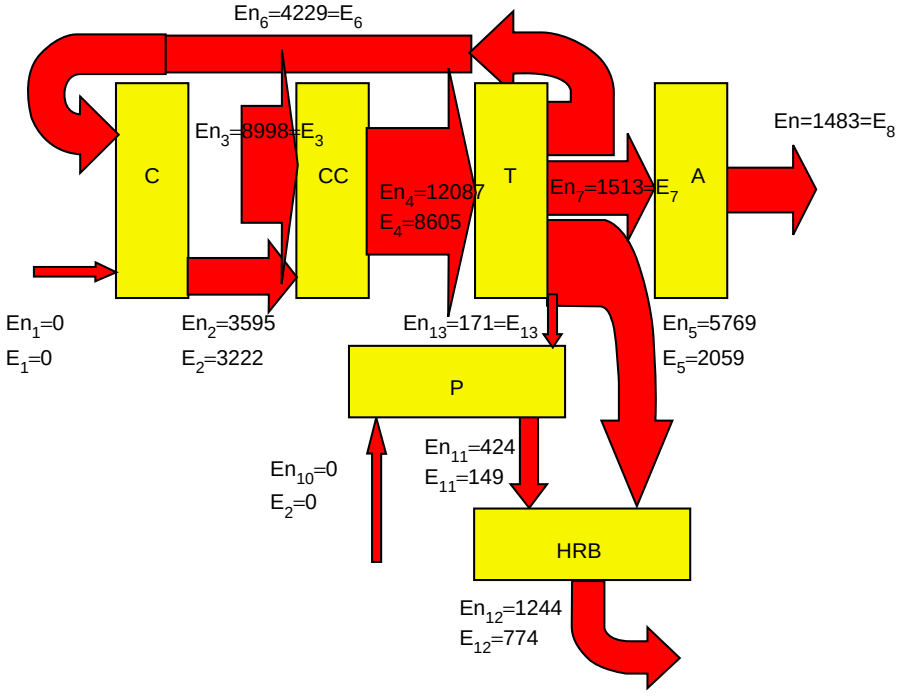


FIGURE 8 Energy (Sankey) and exergy (Grassmann) flow diagrams for the cogeneration plant shown in Fig. 3. En, energy flux; E, exergy flux (both in kW).

analysis and costing, as engineering economics does with energy analysis and costing. The simplest of these method is TE.

An entirely different type of optimization problem encountered in practice is the so-called design synthesis problem, in which the goal is to find the process structure (interconnectivity) that yields a maximum of the objective function. This is a much more difficult task, and some of the possible solution methods are discussed by Frangopoulos. In general, second law analysis provides more insight than its first law counterpart. Especially for synthesis problems, it is important to correctly assess the contribution to the overall process performance of each component, and the exergy flow diagrams offer an immediate glimpse into the real causes of the irreversibilities occurring in the process.

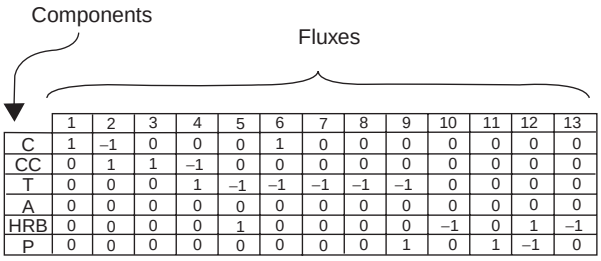
5. STRUCTURAL REPRESENTATION OF ENERGY CONVERSION SYSTEMS

One of the major difficulties in the assessment of the optimal performance of an energy conversion system regards the complex interconnection of its components. As a result, the response of the system to variation in the performance of one of its subunits is not easily predictable: It may well happen that an increase in the efficiency of one component leads to an overall performance deterioration. Several methods have been proposed to represent these cross-effects: The symbolic exergoeconomic method, first formalized by Valero and coworkers in 1988 and 1989, is discussed in detail here, complemented by a formalism adapted from the book by Kotas.

Consider again the cogeneration process of Fig. 3, which we refer to as **P** in this section, and define a matrix CM (connectivity matrix) constructed according to the following algorithm:

1. The jth column ($j = 1, \ldots, N$) of CM represents the jth process streams.
2. The ith row of CM ($i = 1, \ldots, M$) represents the ith component.
3. The generic element CM_{ij} is 0 if flux j does not interact with component i, 1 if flux j enters component i, and -1 if flux j exits component i.

For $M = 6$ and $N = 13$, Fig. 9 represents the ***CM***(6,13) for the plant shown in Fig. 3. This matricial representation of a process can also be derived from an application of graph theory (the Leontief input–output model) and is the basis for a compact and elegant computational procedure. By examining the streams in the cogeneration plant, we see that some of them transport mass and exergy (1–5 and 9–12) and some only exergy (6–8 and 13). We can correspondingly define a mass vector **vm**(13), every element of which represents the mass flow associated with the corresponding stream, and two energy and exergy vectors **vh**(13) and **ve**(13), which contain the energy and exergy values of each stream, respectively. As a further convention, vm contains a 0 in all positions corresponding to a "massless exergy carrier" (a flux of mechanical or electrical energy). It turns out that the mass and energy balances as well as the exergy accounting of the process are described by the scalar product of the matrix CM with the three vectors **vm**, **vh**, and **ve**, respectively.



	1	2	3	4	5	6	7	8	9	10	11	12	13
C	1	–1	0	0	0	1	0	0	0	0	0	0	0
CC	0	1	1	–1	0	0	0	0	0	0	0	0	0
T	0	0	0	1	–1	–1	–1	–1	–1	0	0	0	0
A	0	0	0	0	0	0	0	0	0	0	0	0	0
HRB	0	0	0	0	1	0	0	0	0	–1	0	1	–1
P	0	0	0	0	0	0	0	0	1	0	1	–1	0

FIGURE 9 The connectivity matrix associated with the cogeneration plant shown in Fig. 3.

ve is a vector that contains the exergy destructions in each component. On the basis of this formalism, Valero developed a procedure to symbolically compute the general transfer function of the process—that is, the matricial function Π that premultiplied by any property vector of the input streams (the fuels in Valero's terminology) provides the resulting property vector for the output streams:

$$\begin{aligned} \Pi \times \mathbf{mv}_\mathrm{I} &= \mathbf{mv}_\mathrm{O} \\ \Pi \times \mathbf{hv}_\mathrm{I} &= \mathbf{hv}_\mathrm{O} \\ \Pi \times \mathbf{ev}_\mathrm{I} &= \mathbf{ev}_\mathrm{O} \end{aligned} \tag{17}$$

There is a direct relationship between CM and Π, the second being a sort of augmented form of the former, but with the 0 and 1 substituted by influence coefficients. Note that Π is necessarily square, whereas CM is not. This transfer function is of course a function of the internal characteristics of each component, particularly the efficiencies. Once a closed-form expression for Π is known, we can study the sensitivity of the process performance to that of each one of its components. It is convenient here to adopt the formalism of the influence coefficients, first

proposed independently by Beyer and Brodyanskii in the 1970s and later systematized by Kotas. For each component, we thus define a coefficient of structural bonds σ_{ij} and a coefficient of influence π_{ij}:

$$\sigma_{k,\,i} = \frac{\left(\partial E_{\lambda,\,P}/\partial x_i\right)}{\left(\partial E_{\lambda,\,k}/\partial x_i\right)} \tag{18}$$

$$\pi_{j,\,i} = \left(\frac{\partial E_{in}}{\partial E_j}\right)_{x_i=\text{variable}}. \tag{19}$$

In Eq. (18), $E_{\lambda,P}$ and $E_{\lambda,k}$ are the exergy destructions of the entire process and of the kth component, respectively, linked to the efficiency by

$$\eta = 1 - \frac{E_\lambda}{E_{\text{input}}}. \tag{20}$$

If the ith thermodynamic variable x_i is varied, it produces in general a series of modifications in the efficiencies of some of the components: The sensitivity of each component is measured by $\partial E_{\lambda,k}/\partial x_i$. At a higher aggregation level, the coefficient of structural bonds $\sigma_{k,\,i}$ measures the sensitivity of the process to a variation of the efficiency of the kth component. However, the variation of the efficiency of components that operate on quantitatively small exergy flows is not as important for the process as that of a major components that elaborate the main fluxes of exergy: This aspect is quantified by the influence coefficient π_{ij}. If a variation of the ith thermodynamic variable x_i causes a variation in the magnitude of a certain exergy flux E_j, π_{ji} measures the influence on the overall input of the process.

Once we know σ and π, we can formally analyze (i.e., without to a numerical process simulation) the influence of a variation in a design or operating parameter of the plant. The problem, of course, is that to obtain these coefficients we must perform either an extensive experimental campaign or numerical process simulations.

6. THERMOECONOMICS

Thermoeconomics, in Tsatsaronis' definition, is exergy-aided cost minimization. It is a relatively new branch of engineering sciences that combines engineering economics and exergy analysis to answer the question of the "optimal" process structure and design. The major assumption in TE is that exergy is the only rational basis for constructing a costing function. The difference between TE and EE lies in the treatment of the monetary production factors: TE considers a monetary product cost optimized in terms of exergy considerations, whereas EE transforms monetary expenses into equivalent exergy fluxes and performs an optimization operating solely on exergy fluxes.

In TE, a cost is associated with each stream according to its exergy content. For the generic component shown in Fig. 10, a cost balance shows that the sum of the cost rates associated with the output flows is equal to that of the cost rates associated with the input flows plus the cost rates due to the nonenergetic production factors (capital, labor, and environmental costs). The latter are computed exactly as they would be in engineering economics, taking into account interest rates, inflation, taxes, plant depreciation, etc., and they are denoted here by Z (€/s). The cost balance equation for the component of Fig. 10 (a boiler) is

$$\sum_{\text{in},j}\left(c_{\text{in}} m_{in} e_{\text{in}}\right)_j + c_q E_q + c_w E_w + Z = \sum_{\text{out},i}\left(c_{\text{out}} m_{\text{out}} e_{\text{out}}\right)_i, \tag{21}$$

where the mass flow rates m_j, the corresponding exergy values e_j, and the heat (E_q) and work (E_w) exergy flows are known from previous process calculations. If there are N inputs and M outputs, there will be $N+M$ exergetic cost coefficients c_k. For the analysis to be feasible, the N coefficients relative to the input flows must be known, either from the design specifications or from the analysis of upstream components. We are left with M unknowns, and we need $M-1$ auxiliary equations to supplement Eq. (22). These auxiliary equations are derived on the basis of thermodynamic cost-allocation considerations. In essence, these auxiliary relations express the designer idea of the purpose of that particular component. Note that they are needed to make the transfer function matrix Π square. In our case, the reasoning is as follows:

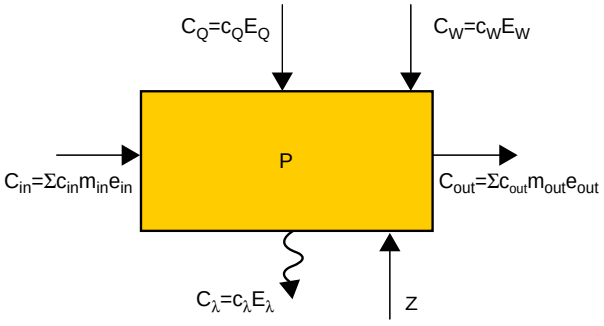


FIGURE 10 Thermoeconomic representation of a process *P*. All flows in €/s.

1. Since the flue gas must pass through a gas-treatment unit before being discharged into the environment, the cost rate associated with the installation and operation of this unit must be charged to this stream:

$$c_3 = -Z_{\text{Desox}}/m_3e_3. \tag{22}$$

2. Similarly, the ash must be charged with the costs of the ash-handling and -disposing equipment:

$$c_4 = -Z_{\text{inert}}/m_4e_4. \tag{23}$$

3. Regarding the boiling and the superheating sections of the tubes, we can consider that the boiler must provide both streams with the respective exergy increase at the same time; therefore, for both streams the cost increase is proportional to the exergy increase with the same constant of proportionality:

$$\frac{m_6c_6 - m_5c_5}{m_6e_6 - m_5e_5} = \frac{m_8c_8 - m_7c_7}{m_8e_8 - m_7e_7}. \tag{24}$$

Bejan, Tsatsaronis, and Moran provide examples of commonly used auxiliary equations for various components.

Once we have M equations in M unknowns, we can solve the TE problem and compute the cost of each stream, particularly that of the final products. If we want to execute an optimization, however, we need additional functional relationships to link the exergetic efficiency of each component to its installation cost (which is reflected in the capital flow factor Z). In the past decade, data have been published by several authors: we recommend the 1998 compilation by El-Sayed.

TE is the most advanced and powerful technique for the cost optimization of complex production systems. It requires an understanding of the exergy method of system analysis and a working knowledge of engineering economics: Both topics are included in most engineering curricula. In every monthly issue of every energy-related technical journal, one may find original applications of TE or a revisitation of an old application. Its impact on the engineering community has been profound because it has promoted the shift, advocated by exergy analysis methods, toward the consideration of irreversibilities and of their real causes in industrial processes. However, it has two weak points. First, being in essence a revised version of engineering economics, it suffers from the adoption of two separate quantifiers: one (exergy) to assess the technical worth of a process, and the other (money) to assess the profitability of the product. There is something inherently awkward in the "pricing" of a physical stream that makes one consider whether it would not be more convenient to base decision tools on purely physical parameters. Second, as a consequence of the first weakness, a correct assessment of environmental issues is problematic in TE. Despite some attempts, for instance, by Frangopoulos and von Spakowsky and independently by Szargut, the problem of attributing a cost to environmental damage remains an open and controversial issue. The current solution is to use the results of large-scale medical studies to attribute a "social cost" to the unit of pollutant discharged into the environment and then charge this cost to the product(s) of the plant. Since pollution effects are inherently nonlinear, and the translation of "a tolerable amount of health risk" in monetary terms is somewhat perplexing, this issue has been intensely debated, resulting in the formulation of a method for process costing based solely on exergy. This method should be called exergoeconomics, and it is described as such here. However, since many engineers use TE and EE as interchangeable terms with the same meaning, it has been published under the name of extended exergy accounting (EEA).

7. EXERGOECONOMICS AND EXTENDED EXERGY ACCOUNTING

If we want to unify the TE costing method and express all expenses by means of a single quantifier, how can we convert monetary expenses into exergy fluxes? The answer has been provided by a controversial energy-accounting theory, emergy (energy memory) analysis, formulated in the 1970s by Howard Odum and mainly used to study the energy cascade of biological systems. Odum maintains that the monetary capital that circulates in a society (a town, county, or nation) is entirely supported by, and therefore equivalent to, the amount of solar energy that maintains life in the environmental niche in which that society thrives. This idea, not developed to its full potential in emergy theories, has been echoed in Herman Daly's critical assessment of the economic theories of growth: An economist by trade, Daly (echoing and developing concepts formulated by others, such as Nicholas Georgescu-Roegen, Lotka, and Robert Costanza) distinguishes between natural capital and monetary capital and demonstrates that the two are not completely substitutable for each other but rather complementary. The

conclusion is that it is the amount of resources we avail ourselves of, and not our capital intensity, that is the key to our survival as an industrial society on Earth. From a completely different point of view, Szargut reached similar conclusions. He devised a procedure to track all amounts of exergy used in the production of a commodity from the mining of the raw materials to the final distribution and disposal, and he proposed to associate a cumulative exergy content to each good or service to express its value in purely exergetic terms. However, both Szargut and Daly left unanswered the problem of how to convert nonenergetic expenditures into resource consumption indices. Extended exergy accounting is the first theory to provide a unified treatment of these issues, along the guidelines discussed here.

EEA adopts the standard exergy accounting method of Szargut to embody into a product all of the exergetic expenditures incurred during its production: Extraction, refining, transportation, preprocessing, final processing, distribution and disposal activities are computed in terms of exergy consumption. Exergy, unlike energy, is not conserved, and in each step of the extended production line a portion of it is irrevocably destroyed in irreversible entropy generation. Thus, the exergetic production factors can be accounted for in full. Regarding the three nonenergetic externalities (labor, capital, and environmental remediation costs), EEA proposes the treatment discussed in the following sections.

7.1 Capital

With regard to including capital and general financial costs in the balance sheet of an energy conversion system, there seems to be a general agreement: We employ money and its time value as the sole quantifier for goods, services, labor, resources, etc. This paradigm of money as a basic metric for energy-related activities is so deeply rooted in our culture that most industrial energy balances are expressed in monetary price per unit of product rather than per energetic (not to mention exergetic) content. There are two reasons to support the necessity of maintaining a purely monetary criterion:

1. The society within which energy conversion systems operate consists of a complex network of related systems, and all exchanges among them are already assessed on a monetary basis.
2. On an average world basis, the contribution of energy-related costs to the global gross national product varies between 2 and 9%; therefore, money (capital), and not energy, is the main quantifier of human activities.

It is perfectly logical to reverse this point of view by considering that economic systems are ecosystems that function only because of the energy and material fluxes that sustain human activities. All agricultural, industrial, and economic activities can only exist as long as they exploit (use) biophysical resources taken from a reservoir of noninfinite mass capacity but of practically infinite exergy capacity. From this point of view, it clearly appears that exergetic content, and not capital, is the correct measure for the value of a commodity or a service, and that the monetary price should reflect this new scale of values.

Naturally, to support the cancellation of monetary prices would be unrealistic and destabilizing: The very structure in which we live and function demands the conservation of the "price-tag" concept in everyday economic activities and on a worldwide scale. EEA advocates that this tag be calculated on the basis of the extended exergetic content (EEC) of a good or service. The numerical relation between the EEC and the price of a product is also immediately clear: The conversion factor is the ratio of some measure of the monetary circulation to the global exergetic input; just as the monetary flux is the driving agent of societal development according to the current paradigm, the exergy flow is its physical driving force. The choice of a proper indicator for the monetary flux, of course, is somewhat arbitrary: EEA adopts the absolute measure of the global monetary circulation ($M2$), which for each country is computed and published by the central bank (or other organization):

$$e_{\text{Cap, Sector}} = \frac{E_{\text{in, Sector}}}{M2} \quad [D\text{J}/€]. \tag{25}$$

When processes are analyzed, it is the exergetic equivalent of the financial expenditures that must be inserted in the balance as an input, and the same applies to the revenues. It is noteworthy that for most state-of-the-art processes the substitution of the equivalent exergetic value for the monetary price causes a significant discrepancy in the exergy balance: the reason is clearly the over- or underestimation of the real extended exergetic content of materials, feed stocks, labor, and energy flows. For the EEA method to be optimally applicable, it is necessary that the economic and exergetic values become locally consistent in the long term, allowing the two value scales to reach a sort of local fixed parity. The same definition of the exergy equivalent implies that different countries may have different

e_{Cap} due to their different productive and economic structures and lifestyles, and that for a country e_{Cap} may vary over time due to an evolving social structure.

7.2 Labor

The capability of attaching a properly computed exergetic value to a labor input (taken here to include all service-related, blue- and white-collar human activities) is perhaps the most relevant novelty of EEA. In all practical applications of engineering cost accounting, including TE, labor is accounted for on a purely monetary basis, but very reasonable alternative approaches have been formulated. Odum's previously mentioned emergetic approach, for instance, computes a value for labor on the basis of the total time-averaged embodied energy content of the human life in the particular area where a worker lives and works. The problem is that the attribution of a purely monetary value to labor is unsatisfactory because it assigns a higher weight to market conditions and financial considerations than to social, technical, and environmental issues that although properly valued may so distort the objective function that new solutions emerge.

EEA proposes to assign labor, and human services in general, in each portion of a society an exergetic value computed as the total (yearly averaged) exergetic resource input into that portion divided by the number of working hours generated therein:

$$e_{L,Sector} = \frac{E_{in,Sector}}{n_{work\ hours,Sector}} \quad [J/man\text{-}hour]. \qquad (26)$$

7.3 Environmental Impact

All methods for assigning a monetary cost to environmental damage in essence suggest that once a substance is acknowledged to be (in some sense) harmful, it become regulated (i.e., legal upper limits are set for its free release into the environment). This is equivalent to setting, for a given technology, an upper price for the cleanup costs for that particular effluent. For many pollutants, however, the environmental situation has deteriorated to a point at which the overall balance of the biosphere has already been affected, and even a small amount of each one of the pollutants is, strictly speaking, intolerable. Therefore, to make a nonzero limit acceptable, the risks to humans are assessed in terms of monetary health- and life-expectancy parameters, and an upper bound is set a priori to the expenditures in such a way so as to remain below a certain statistical probability of incurring that risk. It is claimed that the consequent economical burden is transferred both ways in the economy: downwards, affecting the price of resources that are perceived to be more or less clean, and upwards, increasing the price of products seen as less environmentally friendly. The fallacy of this assertion may be exposed by considering that willingness to pay and the attitude toward a sustainable resource exploitation are different in different countries and may well vary in time. Therefore, in reality, the disturbances caused by a local modification in the environmental regulations affect regions possibly far away from the point of origin, and the method actually promotes an unfair transfer of not only the pollution but also the health risks from a region to another. A solution to this situation has been sought by linking the monetary structure of the environmental levies to energetic considerations: This is the rationale behind pollution commodity trading and the exergy tax. These are remedial measures, aimed at a fairer redistribution of the environmental pressure on a global scale, and they do not address the issue of how high the actual environmental cost is (all methods use the currently regulated values as a basis for their calculations).

EEA advocates a substantially different approach. Consider a process P (Fig. 11A), and assume that its only effluent is a stream that contains hot chemicals, some of which are not at their standard environmental concentration. To achieve a zero environmental impact, these chemicals would have to be brought to both thermal and chemical equilibrium with the surroundings. Thus, the real exergetic cost of the zero impact is proportional to the physical exergy of the effluent because it corresponds to the extended exergy (the sum of the physical exergy spent in the cleanup process plus the invested exergy—labor and capital—required by the installation and operation of the effluent cleanup devices) ideally required to cool the effluent to T_0 and break it up into its constituents such that each one is in equilibrium conditions with the surroundings. A possible representation of such an effluent treating process is shown in Fig. 11B: The additional process P_t requires an energetic input—possibly some auxiliary materials, labor, and invested exergy—but its output will have a zero physical exergy. The additional exergetic expenditures required by P_t must be charged to the effluent O_2, whose extended exergy will now be higher than its original one. Therefore, the overall conversion efficiency of the joint process $(P + P_t)$ is decreased. There may be

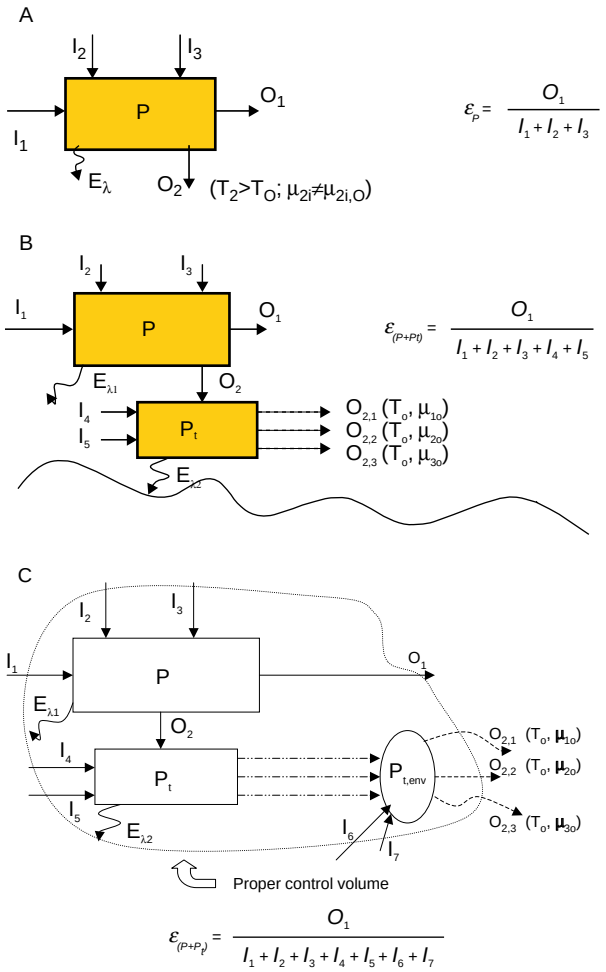


FIGURE 11 Environmental cost assessment in EEA. (A) The effluent O_2 is not at reference conditions. (B) Treatment of the effluent O_2. Each final effluent is at its reference conditions. (C) Only a portion of the cleanup is performed by man-made treatment processes; the remainder occurs spontaneously by biodegradation in the immediate surroundings.

effluents for which some of the chemical decomposition reactions take place spontaneously, in a short time and in the immediate surroundings of the emitting source. In such cases (Fig. 11C), the reactions must draw on some exergy source within the environment (a certain particular chemical catalyst, oxygen, water, solar radiation, or even a biological system), and this exergy flow must be accounted for as well.

EEA thus allows for a consistent incorporation of the effects of effluent treatment in the extended exergetic balance of a process, and it provides an absolute order-of-magnitude estimate of the minimum exergy consumption necessary to achieve zero impact. If an acceptable level of pollutant is specified, then the minimum exergetic expenditure will be proportional to the difference between the values of the physical exergies of the effluent stream between the point of its release and the regulated state point. Here, we use one of the desirable features of environmental cost estimates and at the same time effectively avoid the considerable effort required to determine what the "tolerable environmental impact limit" for a certain pollutant would be.

Once all fluxes (which can be equally regarded as production factors) have been computed in terms of exergy equivalents, EEA makes use of the same structured representation of TE to compute the costs of the final products and to study their dependence on a variation of process parameters. With all terms now expressed in terms of a uniform quantifier, the procedure is somewhat simpler and, more important, absolutely independent of external factors such as market fluctuations or time-varying currency exchange levels.

8. CONCLUSIONS

Current engineering practice uses a mixed energy–monetary formulation to study the cost-formation process of industrial products. TE represents a substantial improvement because its exergy–monetary procedures allow for an exact treatment of irreversibilities and for a better (more precise) accounting of the real causes of exergetic waste, which is equivalent to production cost (if we could find abundant amounts of a certain finished product in nature, its cost would be zero). EE, or EEA, broadens the objective and bases its treatment on a purely exergetic paradigm. TE is becoming increasingly popular, and modern production lines are designed according to its dictates. It is hoped that in the near future, EEA will constitute the basis for our cost-accounting procedure. When (and if) this occurs, we will be able to assess our industrial processes on performance indicators that correctly reflect the resource to final use (including disposal, in a omnicomprehensive cradle-to-grave approach) of our exergy resources.

SEE ALSO THE FOLLOWING ARTICLES

Economic Thought, History of Energy in • *Entropy and the Economic Process* • *Evolutionary*

Economics and Energy • Exergy • Thermodynamics and Economics, Overview • Thermodynamics, Laws of

Further Reading

Ahern, J. E. (1980). "The Exergy Method of Energy Systems Analysis." Wiley, New York.

Baehr, H. D. (1962). "Thermodynamik. Eine Einführung in die Grundlage und ihre technischen Anwendungen." Springer-Verlag, Berlin.

Bejan, A., Tsatsaronis, G., and Moran, M. J. (1996). "Thermal Design and Optimisation." Wiley, New York.

Bosnjakovic, F. (ed.). (1961). "Fachheft Exergie, Die praktische Bedeutung der Exergie," Vol. 13, No. 11.

Daly, H. (1996). "Beyond Growth: The Economics of Sustainable Development." Beacon, Boston.

El-Sayed, Y. (1998). A short course in thermoeconomics, Notes for summer school, Ovidius University, Constantza, Romania.

Georgescu-Roegen, N. (1971). "The Entropy Law and the Economic Process." Harvard Univ. Press, Cambridge, MA.

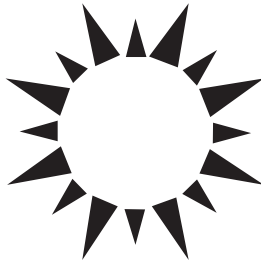
Moran, M. J. (1982). "Availability Analysis: A Guide to Efficient Energy Use." Prentice-Hall, Englewood Cliffs, NJ.

Odum, H. T. (1971). "Environment, Power and Society." Wiley, New York.

Sciubba, E., and Wall, G. (2003). Exergy—A brief commented history. *Int. J. Exergy.*

Exergy

GÖRAN WALL
Exergy Consultant
Mölndal, Sweden



1. **Thermodynamics and Exergy**
2. **Common Forms of Exergy**
3. **Exergy Tools**

Glossary

energy Motion or ability of motion (e.g., the disordered motions of the hot molecules in a cup of coffee); measured in joules (J).
entropy A physical concept of disorder; measured in joules per kelvin (J/K).
exergy Work (i.e., ordered motion) or ability of work (e.g., the motion of a falling object); measured in joules.
exergy analysis Calculation of the total input and output of exergy of a product or service.
exergy destruction Exergy that vanishes in an irreversible process; also called irreversibility.
information Physical capacity of carrying a message; measured in binary units (bits).
irreversible Real processes are always irreversible (i.e., exergy is destroyed) in order to occur.

The exergy concept has its roots in the early work of what would later become thermodynamics. In 1824, N. L. Sadi Carnot stated that the extractable work of a heat engine is proportional to the temperature difference, which led to the definition of the second law of thermodynamics. Willard Gibbs introduced the notion of available work, including the diffusion term, in 1873. He simply stated the following: We will first observe that an expression of the form

$$-\varepsilon + T\eta - Pv + M_1m_1 + M_2m_2 + \ldots + M_nm_n \qquad (1)$$

denotes the work obtainable by the formation (by a reversible process) of a body of which ε, η, v, m_1, $m_2,\ldots,m_n$ are the energy, entropy, volume, and the quantities of the components, within a medium having the pressure P, the temperature T, and the potentials M_1, M_2, …, M_n. (The medium is taken to be so large that its properties are not sensibly altered in any part by the formation of the body.)

Equation (1), originally referred to as Eq. (54) in Gibbs' work, is in exact correspondence with the present definition of exergy. Exergy is a thermodynamic potential; it is a general measure of "difference" or contrast. Thus, two states must be defined—the actual state and an equilibrium or a reference state, often defined by the environment. In 1953, the term exergy was first expressed by Z. Rant to denote "technical working capacity." In 1961, Tribus linked classical thermodynamics with information theory through the entropy concept. In 1977, a concise theory of exergy was offered and exergy was introduced as a useful concept not only for engineering but also for resource accounting in order to reduce environmental destruction. The term exergy is gaining in popularity and gradually replacing terms such as availability, available energy, essergy, utilizable energy, work potential, available work, and convertible energy.

1. THERMODYNAMICS AND EXERGY

Thermodynamics provides the concepts of temperature T, pressure P, heat Q, work W, internal energy U, entropy S, and four laws of thermodynamics. First, the zeroth law establishes the concept of temperature. Second, the first law defines energy as a conserved quantity (i.e., added heat ΔQ is equal to the increase of internal energy ΔU and the extracted work ΔW):

$$\Delta Q = \Delta U + \Delta W. \qquad (2)$$

The extracted work is force times distance, or pressure times volume change (i.e., PdV). Pressure is an intensive variable (i.e., independent of the amount of matter). Volume is an extensive quantity that obviously depends on size or amount of matter:

$$dW = PdV, \qquad (3)$$

 © 2004 Elsevier Inc. All rights reserved.

TABLE I

Energy versus Exergy

Energy	Exergy
The first law of thermodynamics	The second law of thermodynamics
Nothing disappears.	Everything disperses.
Energy is motion or ability to produce motion.	Exergy is work (i.e., ordered motion) or ability to produce work.
$\Delta Q = \Delta U + \Delta W$ (2)	$E = T_0 \Delta S^{\text{tot}}$ (6)
where ΔQ is the total heat supplied to the system, ΔU is the total increase in the internal energy U of the system, and ΔW is the total increase in the external energy of the system or the total work done by the system.	where E is exergy, T_0 is the temperature of the environment, and ΔS^{tot} is the change of the entropy of the total system S^{tot} (i.e., the system and the environment).
$U = mc^2$ (7)	$E = k \ln 2 T_0 I$ (8)
Energy U and matter m are "the same thing."	Exergy and information I is "the same thing."
Everything is energy.	Contrast is exergy.
Energy is always conserved (i.e., in balance); it can neither be produced nor consumed.	Exergy is only conserved in a reversible process but partly consumed in an irreversible process, (i.e., real processes). Thus, exergy is never in balance for real processes.
Energy is a measure of quantity.	Exergy is a measure of quantity and quality.

where V is the volume. Third, the second law defines the concept of entropy through heat:

$$dS \geq 0,\ dS_{\text{reversible}} = 0 \tag{4}$$

$$dQ = TdS, \tag{5}$$

analogous to Eq. (3). Finally, the third law defines the zero level for entropy.

Thermodynamics only treats reversible processes (i.e., nonreal processes) and systems in equilibrium states. Thus, dynamics is not treated and a more proper name is thermostatics. Even so, thermodynamics remains one of the most successful pieces of physics, particularly in engineering.

The first law (i.e., energy is conserved) generally fails to identify losses of work and potential improvements or the effective use of resources (e.g., in an adiabatic throttling process). The second law of thermodynamics shows that for some energy forms, only a part of the energy is convertible to work (i.e., the exergy). Thus, exergy is a very useful tool in engineering. Table I summarizes the main differences between energy and exergy.

1.1 Calculation of Exergy

The exergy of a system is expressed by the formula

$$E = U + P_0 V - T_0 S - \sum_i \mu_{i0} n_i, \tag{9}$$

where U, V, S, and n_i denote extensive parameters of the system (energy or internal energy, volume, entropy, and the number of moles of different chemical materials i), and P_0, T_0, and μ_{i0} are intensive parameters of the environment (pressure, temperature, and chemical potential). This expression is exactly what Gibbs stated in 1873 (Eq. 1). The chemical potential μ_{i0} may be generalized to also include exergy from electromagnetic forces, stress, nuclear forces, etc. This makes exergy the most general concept of work or ability of work. The situation is illustrated in Table II. System A is surrounded by the environment A_0 that is assumed to be much larger than the system (i.e., almost infinite in size). These two systems define two unique thermodynamic states—a difference or a contrast (Table II).

TABLE II

System A in a Much Larger Environment A_0

Intensive quantities	Extensive quantities
T_0 = temperature	U = internal energy
P_0 = pressure	V_0 = volume
μ_{i0} = chemical potential for substance i	S_0 = entropy
	n_{i0} = number of moles of substance i

System A_0 is assumed to be unchangeable in order to initially also define the final state (i.e., the equilibrium state), when both systems have interacted in such a way that the contrast is exhausted. In

the real world, the situation is often like this. The ambient is usually so immense that it may be regarded as unchangeable. Each system is described by intensive and extensive quantities. The intensive quantities are actual measures of the contrast, such as the temperature T, the pressure P, and the chemical potential μ for every substance i. These quantities are not affected by the size of the system but only by the level of contrast. The extensive quantities, on the other hand, are directly proportional to the size of the system. However, it has been determined that human action (e.g., from combustion of fossil fuels) has a destructive impact on the environment in terms of changing its intensive quantities, such as temperature and composition, through the so-called greenhouse effect. The unpredictable dynamics of the environment makes this a dangerous activity.

The external energy P_0V, energy as "ambient heat" T_0S, and energy as "ambient substances" $\mu_{i0}n_i$ are the product of an intensive quantity of the environment and an extensive quantity of the system. This is the expression of the available work from system A when it is brought to equilibrium with the environment, system A_0—that is, when all contrasts have been reversibly transformed into work. This expression is composed of four different terms. First, U is the energy carried within the system, which is also called internal energy. However, part of this energy is useless as work (i.e., exergy). The value of this useless part is handled by the last two terms. P_0V is the work as a volume V that occupies a space of the environment of pressure P_0. This is pure work. Thus, by reducing the volume of the system A work can be extracted. These two terms are positive contributions to the work. The energy or internal energy unfortunately also carries two parts of useless energy or disorder, which must be deducted. T_0S is the part of the energy U that is useless due to lack of order, disorder, or entropy S (i.e., heat at ambient temperature T_0). Similarly, the last term is another useless part of the energy U; in this case, as substances completely identical to substances in the environment (i.e., substances at ambient states). Obviously, these two terms are not in contrast to the environment; thus, they cannot generate work. The energy that is carried by substances can only be used "down to" the level that is given by the environment (cf. with the available potential energy of a waterfall or the carrying capacity of a ship, which is the total capacity minus the ballast).

In summary, exergy expresses the maximum available work, in this case with reference to a specified environment (i.e., to what extent the contrast that is carried by the system can be used as work). Work and heat are concepts related to a process (i.e., that something is happening). They are not state variables. They are phenomena of time; that is, the existence of time is necessary in order to experience these phenomena. When it is said that work and heat are stored in a body, this means that the ability to create work and heat is stored. What is stored is exergy as some form of time-independent contrast.

By using the so-called Gibbs relation,

$$U = ST - VP + \sum_i n_i\mu_i, \tag{10}$$

the following expression is derived from Eq. (9):

$$E = S(T - T_0) - V(P - P_0) + \sum_i n_i(\mu_i - \mu_{i,0}), \tag{11}$$

from which it is seen that E vanishes at equilibrium with the environment; that is, when $T = T_0$, $P = P_0$, and $\mu_i = \mu_{i0}$. This is the same as saying that with no contrast there is no exergy and no potential of action.

Analogously, the exergy of a flow can be written as

$$E = H - T_0S - \sum \mu_{i,0}n_i, \tag{12}$$

where H is the enthalpy.

Exergy is the most general expression of work. Other, less general but more common expressions of thermodynamic potentials (e.g., Gibbs and Helmholz free energies and enthalpy) are special cases of exergy as described in Table III. From Eq. (9) and the limitations specified in Table III, the equivalences

TABLE III

Relations between Differences in Exergy and in Other Thermodynamic Potentials

Limitations	$\Delta E = \Delta X$	Definition of X	Usually named
$\Delta n_i = 0$, $P = P_0$, $T = T_0$	$\Delta E = \Delta G$	$G = U + PV - TS$	Gibbs free energy
$\Delta n_i = 0$, $\Delta V = 0$, $T = T_0$	$\Delta E = \Delta F$	$F = U - TS$	Helmholtz free energy
$\Delta n_i = 0$, $\Delta S = 0$, $P = P_0$	$\Delta E = \Delta H$	$H = U + PV$	Enthalpy

between the difference in exergy and the difference in these other potentials can be derived.

1.2 Reference States

The major conceptual issue is the choice of reference state to which the chemical reactant under consideration is to be compared. Usually, the standard state of each element is defined as its pure elementary state, such as C for carbon, O_2 for oxygen, N_2 for nitrogen, and so on. The pure elements are normally assigned zero values in Gibbs free energy.

The standard exergies of chemical substances are often calculated on the assumption that the environment consists of a number of reference substances, one of each chemical element, with standard concentrations based on the average concentration in the natural environment. The reference substances selected are classified into three groups: gaseous compounds of the atmosphere, solid substances from the lithosphere, and both ionic and nonionic substances from the oceans. These groups are not in thermodynamic equilibrium, neither internally nor externally. Minor deviations of the actual environment from the standard environment only occasionally introduce significant errors to the exergy values.

It should also be noted that exergy may well be defined with reference to any other state (i.e., not necessarily a reference state). Thus, a locally defined environmental state is sometimes preferred. However, such a definition will imply a less general applicability of its use.

2. COMMON FORMS OF EXERGY

2.1 Work

Since exergy is defined as the maximum work potential, they are equivalent in every respect:

$$W = E - T_0 \Delta S. \tag{13}$$

It is stated in the second law that ΔS is always positive. Thus, from Eq. (13) it is obvious that exergy gives an upper limit to the work that is extractable from the process. The maximal amount of work is achieved when ΔS is minimal.

2.2 Pressure

Exergy is work, which is force F times distance S (FS) or pressure P times volume V (PV). This is illustrated by the movement of a piston in a cylinder in Fig. 1.

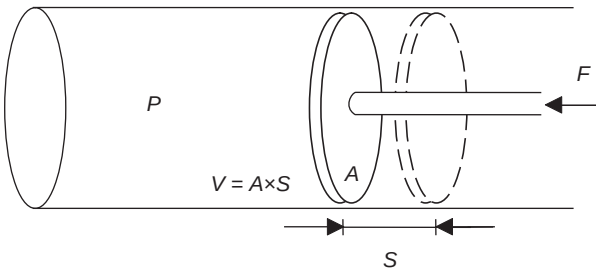


FIGURE 1 Pressure P is acting on a piston with area A that is moved the distance S by applying the force F.

2.3 Heat and Cold

An expression can be derived for the exergy of systems at different temperatures than that of the environment, both higher and lower. This will generate heat transfer when they are allowed to interact with each other or with the environment.

In Fig. 2, heat Q is transferred between two reservoirs at temperatures T and T_0, and $T > T_0$. A reversible heat engine (i.e., a Carnot engine) working between the temperatures T and T_0 gives the maximal available work. By assuming that the heat is completely lost (i.e., maximum entropy production), the exergy being lost (i.e., the available exergy that is able to be utilized) can be calculated. The ambient temperature is T_0 (Fig. 2).

When $T > T_0$, the following familiar relation holds:

$$E = T_0 S^{\text{tot}} = T_0 \left(\frac{Q}{T_0} - \frac{Q}{T} \right) = Q \left(1 - \frac{T_0}{T} \right). \tag{14}$$

If the temperature $T < T_0$, this only changes the direction of the heat flow Q, which then goes from T_0 to T; the exergy becomes

$$E = T_0 \left(\frac{Q}{T} - \frac{Q}{T_0} \right) = Q \left(\frac{T_0}{T} - 1 \right). \tag{15}$$

The ratio between the exergy E and the energy of the transferred heat Q is the exergy factor E/Q. Thus, exergy is obtained by multiplying the energy by the exergy factor. The exergy factor of heat from a heat reservoir (i.e., at a constant temperature T that differs from the ambient temperature T_0) is

$$\frac{E}{Q} = \left| \frac{T - T_0}{T} \right|. \tag{16}$$

This is a generalization of Eqs. (14) and (15) that holds at both $T > T_0$ and $T < T_0$. This expression is

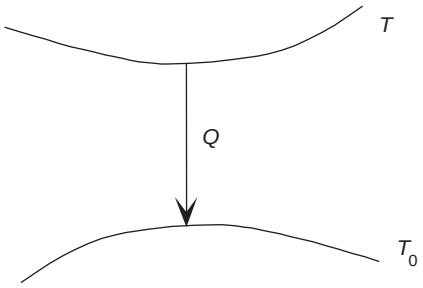


FIGURE 2 Heat Q is transferred between two reservoirs at temperatures T and T_0.

valid for heat from phase changes, i.e.,

$$E = Q\left|\frac{T_P - T_0}{T_P}\right| = L\left|\frac{T_P - T_0}{T_P}\right|, \qquad (17)$$

where T_P is the temperature during the phase change and L is latent heat, which depends on the substance and type of phase change.

Figure 3 shows how E/Q varies with temperature T according to Eq. (16). When $T < T_0$, energy is often regarded as negative, but the exergy is still positive. Figure 3 shows that E/Q strongly increases with low temperature. When T tends to absolute zero temperature (i.e., 0 K or -273.15°C), E/Q tends to infinity. This explains the difficulty in reaching absolute zero temperature. At high temperature, the E/Q comes closer to 1, which means that the exergy (E) and energy (Q) of the heat become almost the same. Higher temperatures (i.e., larger temperature differences) generate more work in a heat engine, which was stated by Carnot in 1824. This equation is useful to describe the exergy needed to maintain the temperature in hot spaces (e.g., furnaces, stoves, ovens, or houses) as well as in a cold storage (e.g., a refrigerator or a freezer) or to keep a pleasant indoor climate by air-conditioning on a hot summer day. Exergy reflects better than energy that heat or cold are more "expensive" when they are needed most. Thus, it is easier and less costly to generate cold in winter and heat in summer, and the storage of ice from winter to summer was once a profitable business. Also, waste heat to the environment carries a higher exergy together with a higher environmental impact in winter than in summer.

The exergy of a body with temperature T (i.e., exergy as sensible heat) is slightly more complicated to calculate since the temperature of the body as well as the delivered heat decrease when heat is transferred. Assume that the heat capacity of the body is C

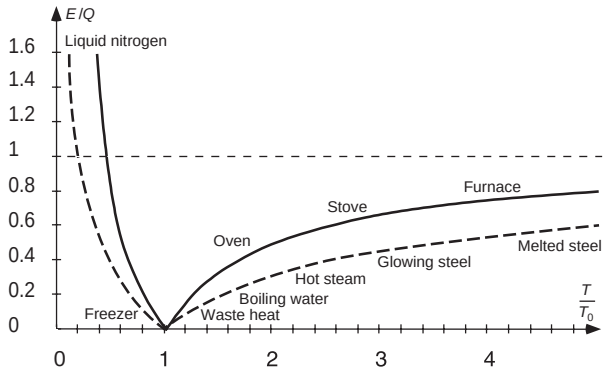


FIGURE 3 E/Q of heat as a function of its temperature; the solid line relates to Eq. (16), and the dashed line relates to Eq. (18).

(i.e., constant); the exergy becomes

$$E = C(T - T_0)\left(1 - \frac{T_0}{T - T_0}\ln\frac{T}{T_0}\right)$$

$$= C\left(T - T_0 - T_0\ln\frac{T}{T_0}\right). \qquad (18)$$

This relation is plotted in Fig. 3, and a number of examples are given, such as waste heat and melted steel. In many cases, such as for melted steel, the total exergy at different temperatures is a combination of latent heat due to phase changes and sensible heat (Eqs. 16 and 18, respectively). Assume that a power station produces 1000 MW of electricity and 2000 MW of waste heat as energy power at a temperature of 310 K by using cooling water at a local ambient temperature of $T_0 = 300$ K. From Eq. (18), the exergy of the waste heat in the cooling water is

$$E = Q\left(1 - \frac{T_0}{T - T_0}\ln\frac{T}{T_0}\right)$$

$$= 2000\left(1 - \frac{300}{310 - 300}\ln\frac{310}{300}\right)$$

$$\approx 2000 \times 0.0165 = 33\ \text{MW}.$$

The discharge of 33 MW waste heat as slightly warmed water causes an increase in the local environmental temperature, which gradually changes the local ecosystem. Most likely, the more exergy that is discharged into the environment, the more damage will be done to the local environment. Thus, exergy is also a suitable environmental indicator.

2.4 Light

The exergy per unit area of light (i.e., black body radiation, which is radiating from a body of temperature T and received by a body of temperature

T_0) can be derived as

$$\dot{e} = \dot{u}\left[1 + \frac{1}{3}\left(\frac{T_0}{T}\right)^4 - \frac{4}{3}\frac{T_0}{T}\right], \tag{19}$$

where $\dot{u}$ is energy emission per unit area and $\dot{u} = \sigma T^4$, where $\sigma = 5.67 \times 10^8\,\mathrm{WK^{-4}\,m^{-2}}$; this is the Stefan–Boltzmann constant.

Sunlight supports life on Earth. The temperature of the surface of the sun is estimated to be approximately $T_{\mathrm{sun}} = 6000\,\mathrm{K}$ and that of the earth approximately $T_{\mathrm{earth}} = T_0 = 300\,\mathrm{K}$. The exergy factor of the sunlight on the earth (i.e., the ratio of the exergy flux to the total energy flux) is

$$\left(\frac{\dot{e}}{\dot{u}}\right)_{\mathrm{sunlight}} = 1 + \frac{1}{3}\left(\frac{300}{6000}\right)^4 - \frac{4}{3} \times \frac{300}{6000} \approx 0.933.$$

Similarly, by considering the temperature of space that is assumed to be 3 K, the exergy factor of the earth's radiation to space is approximately 0.987. Strangely, more exergy leaves the earth with reference to space than is received on the earth as sunlight. However, to our knowledge, this radiated exergy is not part of any life-support system. Transmission through the atmosphere changes the spectral distribution of solar radiation due to the selective absorption of radiation. It also varies with the composition of the atmosphere and the length of the path through the atmosphere, which is determined by the position of the sun. The total solar energy and exergy power to the earth are approximately 1.7×10^{17} and $1.6 \times 10^{17}\,\mathrm{W}$, respectively; that is, the solar energy and exergy power at right angles to the inflow are approximately 1367 and $1275\,\mathrm{W/m^2}$, respectively, before entering the atmosphere.

If the ambient temperature on the earth increases due to the greenhouse effect (i.e., from the accumulation of gases such as carbon dioxide in the atmosphere), this implies a loss of solar exergy. For every higher degree in ambient earth temperatures the loss is approximately 1.5 times the total exergy use in society or approximately $2.5 \times 10^{13}\,\mathrm{W}$. This exergy will now, instead, add to the exergy of the heat radiation from the earth into space.

2.5 Material Substances

Equation (11) gives the following relation if the pressure P and the temperature T are constant (i.e., $P = P_0$ and $T = T_0$):

$$E = \sum_i n_i(\mu_i - \mu_{i,0}). \tag{20}$$

Furthermore, assume that the substance only departs from the environment in terms of chemical potential μ and concentration c. Also, assume that the chemical potential can be written as

$$\mu_i = \mu_i^0 + RT_0 \ln c_i, \tag{21}$$

where μ_i^0 is the chemical potential for the material i in relation to its standard state, obtained from chemical tables, usually the pure element in solid, liquid, or gas phase. Then the exergy becomes

$$E = \sum n_i(\mu_i^0 - \mu_{i0}^0) + RT_0 \sum n_i \ln\frac{c_i}{c_{i0}}, \tag{22}$$

where μ_{i0}^0 is the chemical potential for the material in the environment in relation to its standard state. Data on the chemical potential of different substances are found in the chemical literature. For only one material, the following relation holds:

$$E = n\left[\mu^0 - \mu_0^0 + RT_0 \ln\frac{c}{c_0}\right]. \tag{23}$$

Chemical reactions occurring in combustion processes often result in extensive exergy destruction. Evidently, no energy losses occur in the reaction. Only exergy calculations show that the production of entropy can cause a large percentage of potential work to be destroyed in the reaction. This indicates that an electrochemical membrane reactor, such as a fuel cell, could reduce exergy losses dramatically.

For pure reference components, which exist in the environment, exergy consists of the exergy that can be obtained by diffusing the components to their reference concentration. For gases and when the ideal gas law is applicable, this can be written as

$$E = nRT_0 \ln\frac{P_i}{P_{i0}}, \tag{24}$$

where P_i and P_{i0} are the partial pressures of the gas in the emission and in the environment, respectively.

Assume a power plant is fueled by pure carbon, C, with the exergy content of $410\,\mathrm{kJ\,mol^{-1}}$. Thus, at complete combustion and with no exergy losses besides waste heat, approximately $7.3\,\mathrm{kmol\,s^{-1}}$ of CO_2 is produced as exhaust gases. From Eq. (24), this emission of CO_2 carries an exergy into the environment as waste of

$$E = nRT_0 \ln\frac{P_i}{P_{i0}} \approx 7300 \times 8.31 \times 300 \times \ln\frac{1}{300 \times 10^{-6}} \approx 148\,\mathrm{MW},$$

where the concentration of CO_2 in the atmosphere is assumed to be 300 ppm. Thus, this is a considerable part of the exergy use that also reduces the efficiency

of the process. This emission will cause environmental effects such as global warming by influencing the heat radiation from the earth. When a substance does not exist in the reference environment, it often reacts to reference substances in order to get in equilibrium with the environment.

2.6 Nuclear Fuel

When the nuclear particles (i.e., nucleons) regroup into lower energy levels, exergy is released. In fission, this occurs by splitting heavy atoms, and in fusion it occurs by combining light atoms to form heavy atoms. Since a part of the transformed exergy in most cases appears as neutrinos, which can be regarded as invisible and noninteractive (i.e., not able to convert to work), nuclear exergy may be written as

$$E = \Delta mc^2 - E_{\text{neutrino}}, \tag{25}$$

where c is the speed of light in a vacuum (approximately $3 \times 10^8\,\text{m}\,\text{s}^{-1}$). The exergy of the neutrinos is often approximately 5% of the totally transformed exergy; thus, the exergy factor of nuclear energy is approximately 0.95.

2.7 Information

Exergy appears in daily life basically as three different phenomena: energy, matter, and information. Of these, energy and matter dominate in terms of amount of exergy, whereas information is exergy of a different magnitude, often several magnitudes lower than that of the energy. The exergy of a sheet of paper that carries a written message consists of the exergy of the paper and the exergy of the message (i.e., the information that it represents). The exergy of the paper (i.e., a material substance that may be combusted) amounts to approximately 50 kJ for an A4 sheet weighing 3 g (i.e., 17 MJ/kg). If this sheet carries a message of 5000 alphabetic characters (e.g., an English text), the exergy of this information is approximately 1.5×10^{-20} kJ (Eq. 8) since each character in reality only carries approximately 1 bit of information due to redundancy, which will be explained later. Exergy as material substances amounts to more than 10^{21} times the exergy as information on this sheet of paper. Thus, if it were possible to give an exact figure for the total exergy of this printed paper, the exergy of the information would not appear until in the 21st digit. Thus does not account for the exergy of the ink, which will also exceed that of the information by several magnitudes. This indicates that exergy of information should be treated differently than exergy of energy and matter. Therefore, special attention is given to the exergy of information.

Thermodynamics deals with the physical laws of systems containing many identical particles. From a macroscopic standpoint, the systems may be as small as $10^{-21}\,\text{m}^3$, but from a microscopic standpoint they are still large. $10^{-21}\,\text{m}^3$ contains approximately 10^9 atoms. Thus, it is impossible to have a detailed knowledge of the motion of each individual particle. In statistical mechanics, which provides the theoretical basis for the understanding of thermodynamics, advantage is taken of the large number of particles and statistical methods are used to describe macroscopic properties as average properties resulting from the motion and interaction of many particles.

Information theory deals with systems of incomplete information. The general problem of how to take into account, in an unambiguous way, the available and incomplete information was successfully dealt with by Jaynes in 1957. As Jaynes showed, statistical mechanics is a special case of information theory. Conversely, powerful concepts and methods, developed in statistical mechanics, may be extended into other fields through information theory. In 1961, Tribus bridged the gap to thermodynamics in his classic book, *Thermostatics and Thermodynamics*. Thus, the theory of thermodynamics survived the breakthrough of quantum physics and was incorporated as a special branch of information theory through statistical mechanics. Reality will now be approached from this microscopic view (i.e., where everything is built up from discrete energy quanta). Thus, a system can only appear in a limited number of discrete states, and if energy is added to the system this will occur step by step with not less than one energy quantum at the time.

Assume that the total number of possible states of the system is Ω. The probability that the system is in the jth state is indicated by P_j and the sum of all probabilities is normalized to 1:

$$\sum_{j=1}^{\Omega} P_j = 1. \tag{26}$$

In statistical mechanics, the entropy S is defined as

$$S = -k \sum_{j=1}^{\Omega} P_j \ln P_j, \tag{27}$$

where k is the Boltzmann constant. This gives a completely new insight into the concept of entropy.

From classical thermodynamics, entropy is just a state variable with no physical interpretation, which for generations of students made entropy a mysterious concept. Now it may be regarded as disorder and related to a lack of information or knowledge.

If the system is known to be in a specific state, the probability for this state is equal to 1. Then, the probability for all other states must be equal to 0 since the sum of all probabilities is set at 1 according to Eq. (26). The knowledge of the system is complete and from Eq. (27) the entropy becomes 0: $S=0$ $[0\times\ln(0)=0$ and $1\times\ln(1)=0]$. The state of the system is known. On the other hand, suppose nothing is known about the possible states in which the system is in—the knowledge is nil. All possible states are then equally probable. Since the number of possible states is Ω, then the probability of each state must be $1/\Omega$ in order to sum up to 1, according to Eq. (26). The entropy of the system now reaches its maximum value (i.e., $S=k\ln\Omega$). From our knowledge, the system is in complete disorder, like a library in which all the books are placed randomly. In 1896, Ludvig Boltzmann realized this expression of entropy for a thermodynamic system in equilibrium. Tragically, he never gained recognition for this fundamental relation during his lifetime.

The probabilities at equilibrium P_j^0 are those that maximize entropy S, under whatever constraints on the system that may be given

$$S_{\text{eq}}=S_{\max}=-k\sum_{j=1}^{\Omega}P_j^0\ln P_j^0. \tag{28}$$

The entropy difference between a given state and the equilibrium state of a system then becomes

$$S-S_{\text{eq}}=-k\left(\sum_{j=1}^{\Omega}P_j\ln P_j-\sum_{j=1}^{\Omega}P_j^0\ln P_j^0\right), \tag{29}$$

which then becomes a negative number since $S\le S_{\text{eq}}$. Therefore, the concept of negentropy is sometimes defined as $S_{\text{eq}}-S$, which instead becomes positive and increases with order.

According to information theory, the information content or information capacity of a system is, in binary units (bits),

$$I=\frac{1}{\ln 2}\left(\sum_{j=1}^{\Omega}P_j\ln P_j-\sum_{j=1}^{\Omega}P_j^0\ln P_j^0\right). \tag{30}$$

From Eqs. (29) and (30), it is seen that entropy, or rather negentropy, is very closely linked to information:

$$S_{\text{eq}}-S=k'I, \tag{31}$$

where $k'=k\ln 2\approx 1.0\times10^{-23}$ J/K. Thus, 1 bit of information is equivalent to 1.0×10^{-23} J/K of negentropy.

Assume a system of N different particles with two possible states each (e.g., 0 or 1). Then, $\Omega=2^N$. If there are no other restrictions, then all P_j^0 must be 2^{-N}; see Table IV, in which different particles are indicated by boxes in a row that can carry either 0 or 1.

Total information about the system (i.e., when one of P_j equals 1 and all the others equal 0) gives $I=N$, according to

$$I=\frac{1}{\ln 2}\left[0-\sum_{j=1}^{2^N}2^{-N}\ln(2^{-N})\right]$$
$$=\frac{1}{\ln 2}\left[-2^N\frac{1}{2^N}(-N)\ln 2\right]=N.$$

For every particle, there is information corresponding to a "yes" or "no" to a specific question. Every such answer corresponds to 1 bit of information. Equation (6) (i.e., $E=T_0\Delta S^{\text{tot}}$) gives

$$E=T_0(S_{\text{eq}}^{\text{tot}}-S^{\text{tot}}), \tag{32}$$

where tot is the total system, which in this case is only the system. Equations (31) and (32) give the following relation between exergy and information:

$$E=k'T_0I. \tag{33}$$

Thus, $k'T_0\approx 2.9\times10^{-21}$ J is the amount of exergy connected to 1 bit of information at room temperature. It should be noted that information is a measure not necessarily of actual meaningful information but rather of information capacity. It can also be thought of as a measure of order or structure.

The net flow of information capacity to the earth from the sun may now be calculated from Eq. (33) and the fact that the net inflow of exergy as sunlight (i.e., after reduction for reflection) is approximately

TABLE IV

Number of Possible States and Probabilities of a System of N Different Particles with Two Possible States Each

N	Boxes	Possible states	$\Omega=2^N$	P_j^0
1	□	0, 1	$2^1=2$	$\frac{1}{2}$
2	□□	00, 01, 10, 11	$2^2=4$	$\frac{1}{4}$
3	□□□	000, 001, 010, 100, 011, 101, 110, 111	$2^3=8$	$\frac{1}{8}$
etc.	…	…	…	…

1.1×10^{17} W. This gives an enormous amount of information, or

$$E = \frac{I}{k'T_0} = \frac{1.1 \times 10^{17}}{2.9 \times 10^{-21}} \approx 3.8 \times 10^{37} \text{ bits/s}.$$

Of this, man only uses a tiny fraction, approximately 5×10^{13} bits/s. Thus, the relation between used and available information capacity is approximately 10^{-24}. The same relation for energy conversion is approximately 10^{-5}. Therefore, most of the information flow from the sun (i.e., the exergy of sunlight) remains to be used.

3. EXERGY TOOLS

3.1 Exergy Losses

Energy is always in balance; however, for real processes exergy is never in balance. This unbalance is due to irreversibilities, which are usually called exergy destruction. The exergy output consists of the used output and the nonused output (i.e., exergy of waste output E_{waste}). It is very important to distinguish between exergy destruction caused by irreversibilities and exergy waste due to unused exergy (i.e., exergy flow to the environment). Both represent exergy losses, but irreversibilities have, by definition, no exergy and, thus, no environmental effects. However, due to high exergy losses from irreversibilities, more exergy is used to compensate for this loss, and this in turn may have an environmental impact (i.e., an indirect effect). In the literature, exergy destruction is commonly referred to as availability destruction, irreversibility, and lost work.

Exergy destruction is related to the entropy production by

$$E_{\text{in}}^{\text{tot}} - E_{\text{out}}^{\text{tot}} = T_0 \Delta S^{\text{tot}} = \sum_i (E_{\text{in}} - E_{\text{out}})_i, \qquad (34)$$

where ΔS^{tot} is the total entropy increase, $E_{\text{in}}^{\text{tot}}$ is the total exergy input, $E_{\text{out}}^{\text{tot}}$ is the total exergy output, and $(E_{\text{in}} - E_{\text{out}})_i$ is the exergy destruction in process i.

By definition, an exergy balance only exists for reversible processes. Thus, for real processes [i.e., irreversible processes ($\Delta S^{\text{tot}} > 0$)], exergy is never in balance because the total exergy input always exceeds the total exergy output (i.e., $E_{\text{in}}^{\text{tot}} > E_{\text{out}}^{\text{tot}}$).

By calculating the exergy loss (i.e., destruction and waste), possible process improvements are visualized. In general, when the exergy loss is high, this part should be considered for improvement first. However, this "tackle the biggest loss first" approach is not always appropriate because every part of the system depends on the others, so an improvement in one part may cause increased losses in other parts. As such, the total losses in the modified process may in fact be equal to or even larger than those in the original process configuration. Also, the use of renewable and nonrenewable resources must be considered. Therefore, the problem needs a more careful approach, which is described later.

3.2 Exergy Efficiencies

Exergy efficiency, which is also called second law efficiency, effectiveness, or rational efficiency, is usually defined as utilized exergy divided by used exergy. This must be a number between 0 and 1 since all real processes involve exergy destruction. This is in distinction to energy efficiency, which may well exceed 1. However, there are several ways to define the utilized exergy and used exergy, which add ambiguity to the application of this concept. Also worth noting is that exergy efficiency is often also defined as utilized exergy divided by the exergy that is theoretically possible to utilize due to, for example, time constraints. The faster a process operates the more exergy is lost due to increased irreversibility.

A simple definition of efficiency expresses all exergy input as used exergy and all exergy output as utilized exergy. So the exergy efficiency $\eta_{\text{ex},1}$ becomes

$$\eta_{\text{ex},1} = \frac{E_{\text{out}}}{E_{\text{in}}} = 1 - \frac{E_{\text{in}} - E_{\text{out}}}{E_{\text{in}}}, \qquad (35)$$

where the definition of exergy destruction given previously is added. However, this efficiency does not always provide an adequate characterization of the thermodynamic efficiency of processes, such as heat transfer, separation, and expansion. Often, there exists a part of the output exergy that is unused (i.e., an exergy waste E_{waste} to the environment). Thus, the utilized exergy is given by $E_{\text{out}} - E_{\text{waste}}$, which is called the exergy product E_{pr}. The output consists of two parts:

$$E_{\text{out}} = E_{\text{pr}} + E_{\text{waste}}. \qquad (36)$$

The exergy efficiency $\eta_{\text{ex},2}$ now instead becomes

$$\eta_{\text{ex},2} = \frac{E_{\text{out}} - E_{\text{waste}}}{E_{\text{in}}} = \frac{E_{\text{pr}}}{E_{\text{in}}} = \eta_{\text{ex},1} - \frac{E_{\text{waste}}}{E_{\text{in}}}. \qquad (37)$$

Sometimes, part of the exergy going through the system is unaffected. This part has been called transit exergy E_{tr} (Fig. 4).

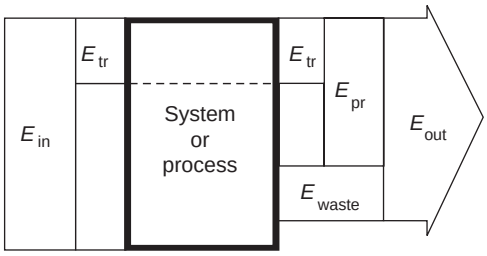


FIGURE 4 The input and output of exergies for a system.

If the transit exergy E_{tr} is deducted from both the input and the output exergy (or, rather, from the exergy product), the exergy efficiency $\eta_{ex,3}$ becomes

$$\eta_{ex,3} = \frac{E_{out} - E_{waste} - E_{tr}}{E_{in} - E_{tr}} = \frac{E_{pr} - E_{tr}}{E_{in} - E_{tr}}. \quad (38)$$

Due to the difficulties of calculating transit exergy and its lack of recognition, the exergy efficiency $\eta_{ex,2}$ is normally used:

$$\eta_{ex} = \eta_{ex,2} = \frac{E_{out} - E_{waste}}{E_{in}} = \frac{E_{pr}}{E_{in}}. \quad (39)$$

These latter definitions are compared by applying them to a system with two different processes A and B (Fig. 5). The exergy efficiencies for process A are $\eta_{ex,2} = 91\%$ and $\eta_{ex,3} = 10\%$, and those for process B are $\eta_{ex,2} = \eta_{ex,3} = 50\%$. Thus, determining which is the most efficient process is a matter of defining efficiency. However, from Fig. 5, it is seen that the exergy destruction of process A is larger than that of process B (9 vs 5); thus, process A should really be improved first.

From this comparison it is seen that a better insight is offered by using exergy flow diagrams. From an exergy flow diagram, one can learn

- the exergy efficiencies of the various parts of a system
- the different exergy inputs and outputs
- where the various exergy flows come from and go to
- the amount of transit exergy
- how much exergy is destroyed in each processes

This often leads to a better insight and understanding of the system and its characteristics, which implies a better basis for improvement considerations.

3.3 Exergy Flow Diagrams

From the previous discussion, it is known that ambiguity reduces if an exergy flow diagram is used to demonstrate an exergy transfer instead of a ratio. In engineering, these diagrams are often used to describe the energy or exergy flows through a process (Fig. 6).

Figure 6 shows a typical power plant, its main components, and the energy and exergy flows of the plant. This diagram shows where the energy and exergy losses occur in the process and also whether exergy is destroyed from irreversibilities or whether it is emitted as waste to the environment. This information is very useful when considering the ecological effects of an activity. It also gives a hint

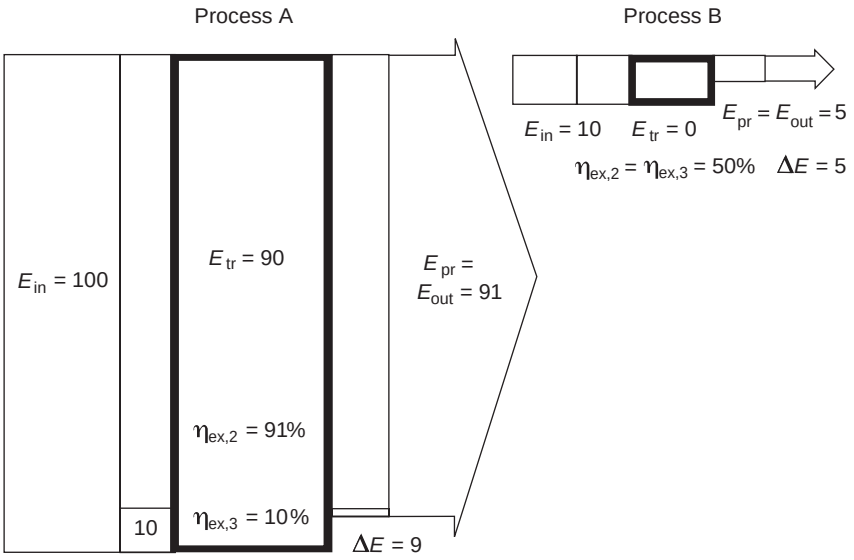


FIGURE 5 Comparison of efficiencies of exergy flows for two processes, A and B.

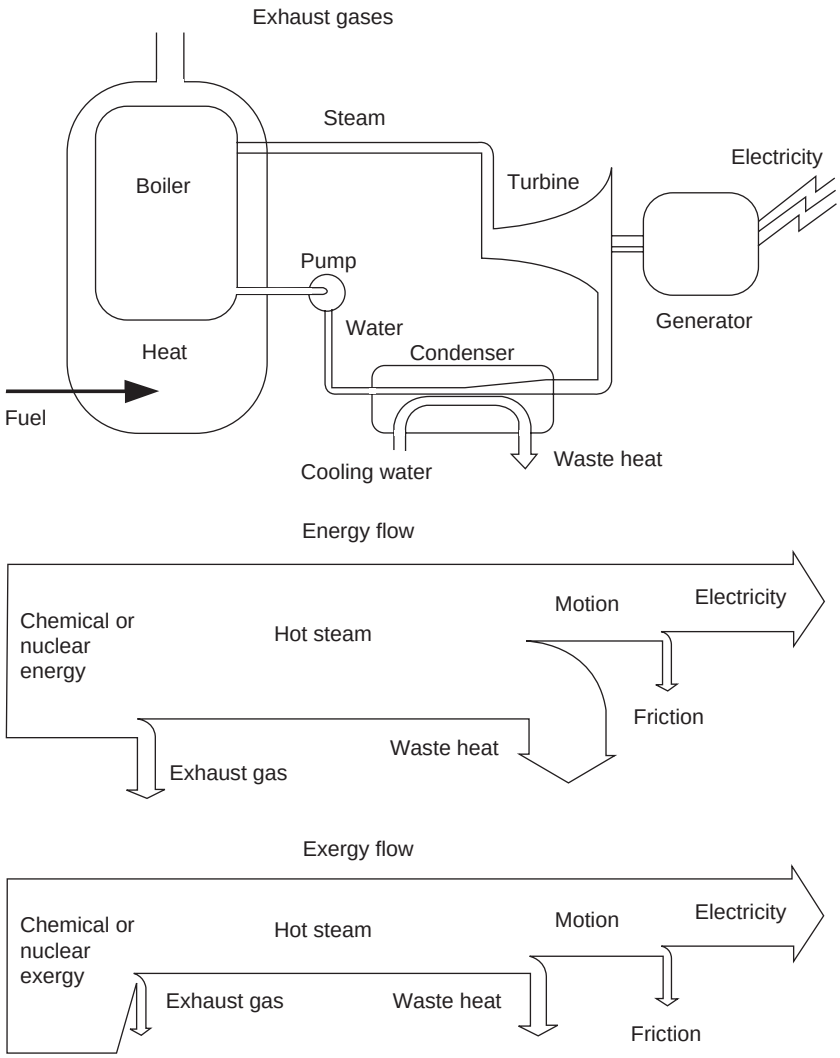


FIGURE 6 Energy and exergy flow diagrams of a condensing power plant.

about the possibilities of improving the process and where to direct efforts for improvement. In the energy flow diagram, energy is always conserved, as stated by the first law of thermodynamics. The waste heat carries the largest amount of energy into the environment, far more than is carried by the exhaust gases. However, in the exergy flow diagram the temperature of the waste heat is close to ambient so the exergy becomes much less. The exergy of the exhaust gas and that of the waste heat are comparable. Energy and exergy losses from friction have little direct impact on the environment. The ecological effects are more related to the exergy flows than to the energy flows, which makes exergy flows better ecological indicators.

The widths of the inflows and outflows in the energy and the exergy diagrams are almost equal because both the inflows and the useful outflows are very high-quality energy forms, and for them energy and exergy are almost equal. For electricity, the energy and the exergy values are identical since electricity does not carry any entropy. The losses in the two diagrams, however, are quite different. First, there are losses in the furnace, where fuel is converted into heat. In an oil- or coal-fueled power plant the flame temperature is approximately 2000°C, but in a nuclear power plant the temperature is only approximately 300°C due to safety regulations. The heat is then transmitted through heat exchangers to boil water (i.e., produce steam in what is often called a steam generator). The pressure is high to make the water boil at high temperatures. Through the walls of the furnace, and through pipes, heat is lost to the environment. Heat is also lost with the exhaust gases,

usually in order to keep the chimneys free from condensed water. These losses represent only a minor part of the total energy flow. However, from the exergy diagram, we see that more than one-third of the exergy is lost at this stage. We also see that exergy disappears and only a small fraction leaves the plant. The exergy flow literally shrinks. Large quantities of entropy are created because the steam that leaves the boiler has a relatively low temperature and pressure due to limitations in the fatigue strength of the components in the process, principally the boiler and the turbine blades. In a nuclear power plant, more than half the converted exergy is lost during this stage of the process.

The energy and exergy of hot steam in Fig. 6 differ a great deal. In the energy diagram, major losses occur in the condenser as waste heat in the cooling water. The waste heat in this case, however, is heat at a temperature slightly above the ambient temperature and its exergetic value is therefore very low. This is obvious from the exergy diagram, in which the waste heat is a minor flow. Thus, for a conversion of waste heat into work by an ideal process (i.e., a Carnot engine), only the amount indicated in the exergy diagram would become work, and the loss would be waste heat at ambient temperature (i.e., exergy equals zero). It is not possible to produce any more work or electrical energy from the steam, or from any other energy form, than what is determined by its exergy content.

When converting mechanical energy into electrical energy, both of which have an exergy factor of 1, small losses arise from friction. These losses are on the whole equally small in both diagrams. Part of the friction losses is due to mechanical fatigue (i.e., wearing out of shafts and bearings).

Figure 7 illustrates the energy and exergy flows of an oil furnace, an electric heater, an electric heat pump, and a combined power and heat plant (i.e., a cogeneration plant). In the oil furnace, the energy efficiency is assumed to be approximately 85%, with losses due mainly to the hot exhaust gases (Fig. 7). The exergy efficiency is very low (approximately 4%) because the temperature difference is not utilized when the temperature is decreased to a low of approximately 20°C, a comfortable indoor climate.

Electric heating by short-circuiting in electric resistors has an energy efficiency of 100% by

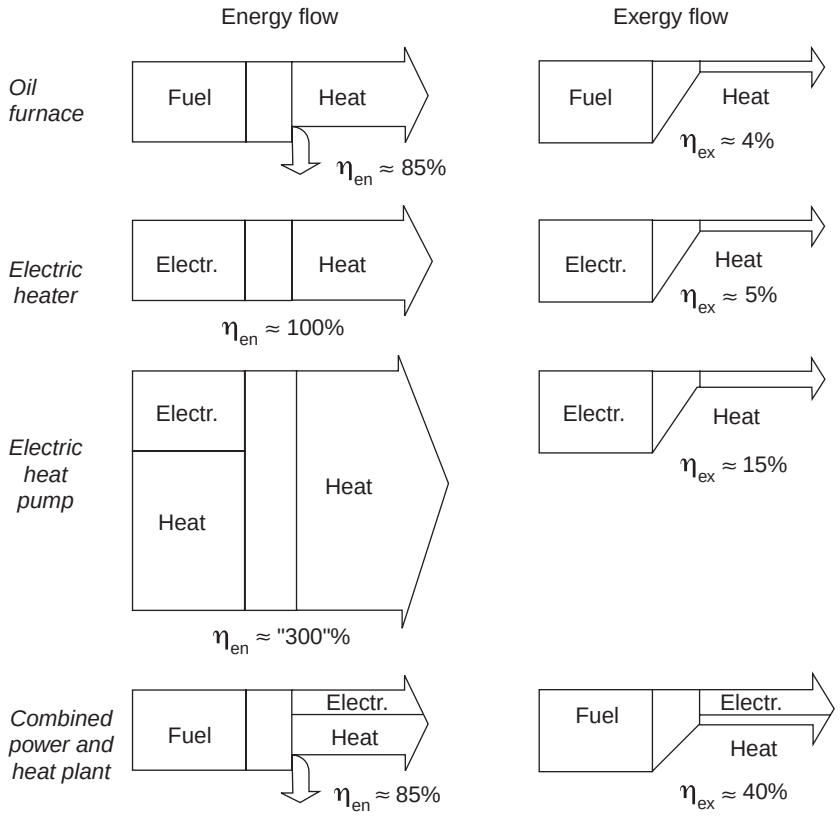


FIGURE 7 Energy and exergy flows through some typical energy systems.

definition of energy conservation. The energy efficiency of an electric heat pump is not limited to 100%. If the heat originating from the environment is ignored in the calculation of the efficiency, the conversion of electrical energy into indoor heat can be well over 100% (e.g., 300% as in Fig. 7). The exergy flow diagram of the heat pump looks quite different. The exergy efficiency for an electric heater is approximately 5% and for the heat pump 15%.

In Fig. 6, the energy and exergy efficiencies are the same because the inflow of fuels and the outflow of electricity both have an exergy factor of approximately 1. For a combined power and heat plant (Fig. 7), the exergy efficiency is about the same as that for a condensing power plant (Fig. 6). The main exergy loss occurs in the conversion of fuel into heat in the boiler. Since this conversion is practically the same in both the condensing and the combined power plants, the total exergy efficiency is the same (i.e., approximately 40%). The maximum energy efficiency of an ideal conversion process may be more than 100%, depending on the definition of efficiency. The exergy efficiency, however, can never exceed 100%.

Exergy flow diagrams may also include the total use of physical resources (i.e., energy and other resources for a system). Figure 8 shows the main conversions of energy and materials in Swedish society in 1994. The flows of resources go from left to right in the diagram (i.e., from the resource base to the consumption sector). Thus, the diagram basically represents the resource supply sector in which resources, such as crops and minerals, are turned into consumer goods, such as food, transport, and thermal comfort. The width of the flows is defined by their exergy content and the unit of the flows is Joules. The accuracy of the flows varies a great deal between the different areas. For the electricity system, the accuracy is quite high, whereas for

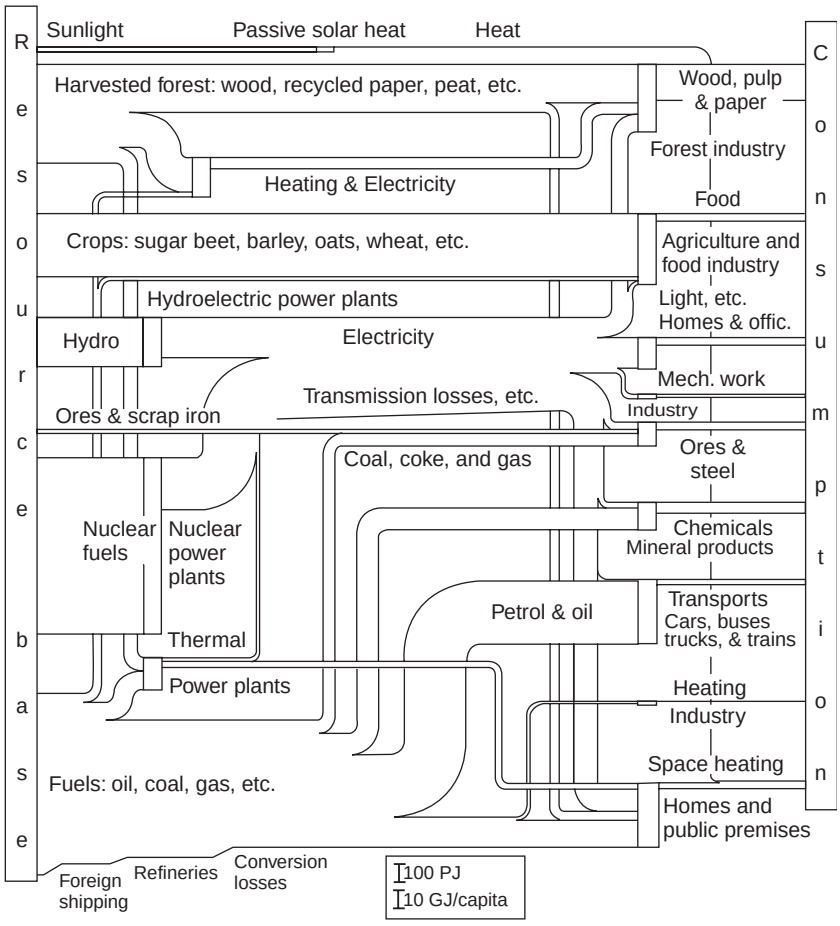


FIGURE 8 The exergy conversion system in the Swedish society in 1994. Total input was approximately 2720 PJ or 310 GJ per capita, and total output was approximately 380 PJ or 43 GJ per capita.

sectors related to agriculture and forestry it is not. In order to not make the diagram too complicated, only exergy flows exceeding 5 PJ are included. The inflows are ordered according to their origins. Sunlight is thus a renewable natural flow. Besides a minor use of wind power, far less than 5 PJ, this is the only direct use of a renewable natural flow. Harvested forests, agricultural crops, and hydropower are renewable exergy flows derived from funds, which of course are founded on the renewable natural flow of sunlight. Iron ore, nuclear fuels, and fossil fuels are non-renewable exergy flows from deposits, which are exhaustible and also carry toxic substances. The unfilled boxes represent exergy conversions, which in most cases represent a huge number of internal conversions and processes. The resources actually demanded by society appear as outflows on the right side of the diagram. The total inflow of resources during 1994 was approximately 2720 PJ or 310 GJ per capita, and the net output was 380 PJ or 40 GJ per capita. Thus, the overall efficiency of the supply sector can be estimated at less than 15%, which must be regarded as poor. As can be seen, some sectors are extremely inefficient. Some resource conversion systems have a ridiculously poor efficiency. For nuclear fuel to space heating through short-circuit heaters the utilization is less than 0.025% if the physically available nuclear exergy is considered.

Exergy is a concept to be used by society at large, e.g., in town and country planning. In addition to having an important pedagogical value, it bridges gaps between different areas of science (e.g., thermodynamics, information theory, economics, and ecology). Thus, exergy will play a central role in disciplines that are dedicated to sustainable development.

SEE ALSO THE FOLLOWING ARTICLES

Entropy • Entropy and the Economic Process • Exergoeconomics • Exergy Analysis of Energy Systems • Exergy Analysis of Waste Emissions • Exergy: Reference States and Balance Conditions • Thermodynamics and Economics, Overview • Thermodynamics, Laws of • Thermodynamic Sciences, History of

Further Reading

Encyclopedia of Life Support Systems, http://www.eolss.com and http://www.eolss.net.

Exergy, an International Journal, http://www.exergyonline.com.

Tribus, M. (1961). *Thermostatics and Thermodynamics*. Van Nostrand, New York.

Wall, G. (1986). Exergy—A useful concept, Ph.D. thesis. Chalmers University of Technology, Sweden; http://www.exergy.se/goran/thesis/index.html.

Wall, G. (1997). *Exergetics*. http://www.exergy.se/ftp/exergetics.pdf.

Exergy Analysis of Energy Systems

MARC A. ROSEN
University of Ontario Institute of Technology
Oshawa, Ontario, Canada

1. **Exergy**
2. **Exergy Analysis**
3. **Survey of Applications of Exergy Analysis to Energy Systems**
4. **Examples of Applications of Exergy Analysis to Energy Systems**
5. **Simple Device: Electrical Resistance Space Heater**
6. **Simple System: Thermal Energy Storage System**
7. **Complex System: Coal-Fired Electrical Generating Station**
8. **Macrosystem: The Aggregate Energy Systems of a Country**
9. **Conclusions**

Glossary

chemical exergy The maximum work obtainable from a substance when it is brought from the environmental state to the dead state by means of processes involving interaction only with the environment.

dead state The state of a system when it is in thermal, mechanical, and chemical equilibrium with a conceptual reference environment, which is characterized by a fixed pressure, temperature, and chemical potential for each of the reference substances in their respective dead states.

environmental state The state of a system when it is in thermal and mechanical equilibrium with the reference environment—that is, at the pressure and temperature of the reference environment.

exergy (i) A general term for the maximum work potential of a system, stream of matter, or a heat interaction in relation to the reference environment as the datum state. Also known as available energy, availability, essergy, technical work capacity, usable energy, utilizable energy, work capability, work potential, and xergy. (ii) The unqualified term exergy or exergy flow is the maximum amount of shaft work obtainable when a steady stream of matter is brought from its initial state to the dead state by means of processes involving interactions only with the reference environment.

exergy analysis An analysis technique, based on the second law of thermodynamics, in which process performance is assessed by examining exergy balances, losses, and efficiencies.

exergy consumption The exergy consumed or destroyed during a process due to irreversibilities within the system boundaries. Also known as dissipation, irreversibility, and lost work.

physical exergy The maximum amount of shaft work obtainable from a substance when it is brought from its initial state to the environmental state by means of physical processes involving interaction only with the environment. Also known as thermomechanical exergy.

reference environment An idealization of the natural environment that is characterized by a perfect state of equilibrium—that is, the absence of any gradients or differences involving pressure, temperature, chemical potential, kinetic energy, and potential energy. The environment constitutes a natural reference medium with respect to which the exergy of different systems is evaluated.

reference state A state with respect to which values of exergy are evaluated. Several reference states are used, including the environmental state, dead state, standard environmental state, and standard dead state. Also known as ground state.

thermal exergy The exergy associated with a heat interaction—that is, the maximum amount of shaft work obtainable from a given heat interaction using the environment as a thermal energy reservoir.

Energy analysis is based on the first law of thermodynamics, which embodies the principle of the conservation of energy. Exergy analysis is a thermodynamic analysis technique for systems and processes that is based on the second law of thermodynamics and has several advantages over energy analysis.

 © 2004 Elsevier Inc. All rights reserved.

First, more meaningful efficiencies are evaluated with exergy analysis since exergy efficiencies are always a measure of the approach to the ideal. Second, the inefficiencies in a process are better pinpointed with exergy analysis in that the types, causes, and locations of the losses are identified and quantified. For these reasons and others, exergy analysis has been increasingly applied during the past several decades. In this article, the role of exergy analysis in the assessment and improvement of energy systems is discussed. First, exergy and its use as a tool through exergy analysis are briefly described. Then, the range of energy systems that have in the past been assessed using exergy analysis are surveyed. Finally, several examples of applications of exergy analysis are presented, ranging from simple energy devices and systems to complex energy systems and large macrosystems.

1. EXERGY

Exergy is defined as the maximum amount of work that can be produced by a stream or system as it is brought into equilibrium with a reference environment, and it can be thought of as a measure of the usefulness or quality of energy. Exergy is consumed during real processes due to irreversibilities and conserved during ideal processes.

Exergy quantities are evaluated with respect to a reference environment. The intensive properties of the reference environment in part determine the exergy of a stream or system. The exergy of the reference environment is zero. The reference environment is in stable equilibrium, with all parts at rest relative to one another. No chemical reactions can occur between the environmental components. The reference environment acts as an infinite system and is a sink and source for heat and materials. It experiences only internally reversible processes in which its intensive state remains unaltered (i.e., its temperature T_o, pressure P_o, and the chemical potentials μ_{ioo} for each of the i components present remain constant).

Energy and exergy balances can be written for a general process or system. An energy balance for a system may be written as

$$\text{Energy input} - \text{energy output} = \text{energy accumulation}. \quad (1)$$

Energy input and output refer respectively to energy entering and exiting through system boundaries. Energy accumulation refers to buildup (either positive or negative) of the quantity within the system. Clearly, energy is subject to a conservation law.

By combining the conservation law for energy and nonconservation law for entropy (which states that entropy is created during a process due to irreversibilities), an exergy balance can be obtained:

$$\text{Exergy input} - \text{exergy output} - \text{exergy consumption} = \text{exergy accumulation}. \quad (2)$$

Exergy is consumed due to irreversibilities. Exergy consumption is proportional to entropy creation. Equations (1) and (2) demonstrate an important difference between energy and exergy: Energy is conserved, whereas exergy, a measure of energy quality or work potential, can be consumed.

2. EXERGY ANALYSIS

Exergy analysis involves the application of exergy concepts, balances, and efficiencies to evaluate and improve energy and other systems. Many engineers and scientists suggest that devices can be well evaluated and improved using exergy analysis in addition to or in place of energy analysis. Exergy analysis has been applied to a wide range of processes and systems. This article focuses on applications to energy systems.

2.1 Exergy Analysis and Limitations on Increased Efficiency

Increases in efficiency are subject to two constraints, which are often poorly understood: (i) theoretical limitations, which establish the maximum efficiency theoretically attainable for a process by virtue of the laws of thermodynamics, and (ii) practical limitations, which further limit increases in efficiency.

First, consider practical limitations on efficiency. In practice, when selecting energy sources and utilization processes the goal is not to achieve maximum efficiency but rather to achieve an optimal trade-off between efficiency and such factors as economics, sustainability, environmental impact, safety, and societal and political acceptability. This optimum is dependent on many factors controllable by society; furthermore, these factors can be altered to favor increased efficiency (e.g., governments can offer financial incentives that render high-efficiency technologies economically attractive and/or provide disincentives for low-efficiency alternatives through special taxes and regulations).

Now, consider theoretical limitations on efficiency. To assess the potential for increased efficiency, the theoretical limits must be clearly understood. In the past, lack of clarity on this issue has often led to confusion, in part because energy efficiencies generally are not measures of how close the performance of a process or device approaches the theoretical ideal. The consequences of such confusion can be very significant. For example, extensive resources have occasionally been directed toward increasing the energy efficiencies of devices that in reality were efficient and had little potential for improvement. Conversely, at other times devices have not been targeted for improved efficiency, even though the difference between the actual and maximum theoretical efficiencies, which represents the potential for improvement, has been large.

The difficulties inherent in energy analysis are also attributable to the fact that it considers only quantities of energy and ignores energy quality, which is continually degraded during real processes. Exergy analysis, as discussed earlier, overcomes many of the problems associated with energy analysis.

2.2 Applications of Exergy Analysis beyond Thermodynamics

Exergy concepts can be applied beyond thermodynamics. One area in which applications of exergy are increasing is that of environmental impact. Many suggest that the impact of energy resource utilization on the environment and the achievement of increased resource-utilization efficiency should be addressed by considering exergy, among other factors. Although the exergy of an energy form or a substance is a measure of its usefulness, exergy is also a measure of its potential to cause change. The latter point suggests that exergy may be, or provide the basis for, an effective measure of the potential of a substance or energy form to impact the environment.

Another area in which applications of exergy are increasing is that of economics. In the analysis and design of energy systems, techniques are often used that combine scientific disciplines (especially thermodynamics) with economics to achieve optimum designs. For energy systems, costs are conventionally based on energy. Many researchers, however, have recommended that costs are better distributed among outputs, based on exergy. Methods of performing exergy-based economic analyses have evolved and are referred to as thermoeconomics, second law costing, and exergoeconomics. These analysis techniques recognize that exergy, more than energy, is the commodity of value in a system, and they assign costs and/or prices to exergy-related variables. These techniques usually help determine the appropriate allocation of economic resources so as to optimize the design and operation of a system and/or the economic feasibility and profitability of a system (by obtaining actual costs of products and their appropriate prices).

2.3 Steps for Energy and Exergy Analyses

A simple procedure for performing energy and exergy analyses involves the following steps:

- Subdivide the process under consideration into as many sections as desired, depending on the depth of detail and understanding desired from the analysis.
- Perform conventional mass and energy balances on the process, and determine all basic quantities (e.g., work and heat) and properties (e.g., temperature and pressure).
- Based on the nature of the process, the acceptable degree of analysis complexity and accuracy, and the questions for which answers are sought, select a reference environment model.
- Evaluate energy and exergy values, relative to the selected reference environment model.
- Perform exergy balances, including the determination of exergy consumptions.
- Select efficiency definitions, depending on the measures of merit desired, and evaluate the efficiencies.
- Interpret the results, and draw appropriate conclusions and recommendations relating to such issues as design changes, retrofit plant modifications, etc.

3. SURVEY OF APPLICATIONS OF EXERGY ANALYSIS TO ENERGY SYSTEMS

Many applications of exergy analysis to energy systems have been reported in the literature, particularly during the past several decades. Some extensive bibliographies have been compiled, including one by Goran Wall (http://exergy.se). In addition, a journal devoted to exergy matters, *Exergy: An International Journal*, has been established.

The types of applications of exergy methods to energy systems that have been reported in the past are extremely varied and include the following:

- Electricity generation using both conventional devices, such as fossil and nuclear power plants, and alternative devices, such as fuel cells and solar energy systems
- Engines of various types
- Transportation systems for land, air, and water transport
- Heating and cooling systems for building systems and industrial applications
- Cogeneration systems for producing heating and electrical needs simultaneously
- Petrochemical processing and synthetic fuels production
- Chemical processes, such as sulfuric acid production, distillation, and water desalination
- Metallurgical processes, such as lead smelting
- Combustion technologies and systems
- Energy storage systems, such as batteries, pumped storages, and thermal energy storages

4. EXAMPLES OF APPLICATIONS OF EXERGY ANALYSIS TO ENERGY SYSTEMS

Examples of applications of exergy analysis are presented in the following four sections. Because the range of energy systems to which exergy has been applied is great, the four examples described here were selected to cover the breadth of complexity. Consequently, we start with a simple device (an electrical resistance space heater) and then consider a simple system (a thermal energy storage system), where simple is taken here to imply that the system has a very limited number of components. Next, we examine a more complex system (a coal-fired electrical generating station). Finally, we investigate the application of exergy analysis to a "macrosystem"—that is, a large conglomerate of energy systems that cover a region such as a city, province, or country.

5. SIMPLE DEVICE: ELECTRICAL RESISTANCE SPACE HEATER

An electrical resistance space heater converts electricity to heat at a temperature suitable for keeping a room at a comfortable temperature. It is illustrated in Fig. 1.

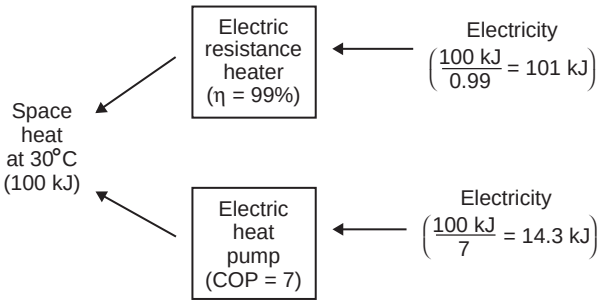


FIGURE 1 Comparison of the quantity of electricity required to provide a fixed amount of space heat (100 kJ) using two different heating devices: an electric resistance heater and an electric heat pump. η, energy efficiency; COP, coefficient of performance.

The energy efficiency of electric resistance space heating is often said to exceed 99%. The implication is that the maximum possible energy efficiency for electric resistance heating is 100%, corresponding to the most efficient device possible.

This understanding is erroneous, however, because energy analysis ignores the fact that in this process high-quality energy (electricity) is used to produce a relatively low-quality product (warm air). Exergy analysis recognizes this difference in energy qualities and indicates that the exergy of the heat delivered to the room is approximately 5% of the exergy entering the heater. Thus, based on exergy, the efficiency of electric resistance space heating is approximately 5%.

We therefore obtain useful information from the exergy results. Since thermodynamically ideal space heating has an exergy efficiency of 100%, the same space heating can theoretically be achieved using as little as 5% of the electricity used in conventional electric resistance space heating. In practical terms, one can achieve space heating with a greatly reduced electricity input using a high-efficiency electric heat pump (Fig. 1), using 15% of the electricity that electric resistance heating would require for a heat pump with a coefficient of performance of 7.

6. SIMPLE SYSTEM: THERMAL ENERGY STORAGE SYSTEM

A thermal energy storage (TES) system receives thermal energy and holds the energy until it is required. TESs can store energy at temperatures above or below the environment temperature, and there are many types of systems (e.g., tanks, aquifers, ponds, and caverns).

The evaluation of a TES system requires a measure of performance that is rational, meaningful, and practical. The conventional energy storage efficiency is an inadequate measure. A more perceptive basis for comparison is needed if the true usefulness of thermal storages is to be assessed and thus permit maximization of their economic benefit. Efficiencies based on ratios of exergy provide rational measures of performance since they can measure the approach of the performance of a system to the ideal.

That the energy efficiency is an inappropriate measure of thermal storage performance can best be appreciated by a simple example. Consider a perfectly insulated thermal storage containing 1000 kg of water, initially at 40°C. The ambient temperature is 20°C, and the specific heat of water is taken to be constant at 4.2 kJ/kg K. A quantity of 4200 kJ of heat is transferred to the storage through a heat exchanger from an external body of 100 kg of water cooling from 100 to 90°C. This heat addition raises the storage temperature 1.0°C to 41°C. After a period of storage, 4200 kJ of heat is recovered from the storage through a heat exchanger that delivers it to an external body of 100 kg of water, raising the temperature of that water from 20 to 30°C. The storage is returned to its initial state at 40°C.

For this storage cycle the energy efficiency, which is the ratio of the heat recovered from the storage to the heat injected, is 4200/4200 kJ = 1, or 100%. However, the recovered heat is at only 30°C and of little use, having been degraded even though the storage energy efficiency was 100%. The exergy recovered in this example is evaluated as 70 kJ and the exergy supplied as 856 kJ. Thus, the exergy efficiency, which is the ratio of the thermal exergy recovered from storage to that injected, is 70/856 = 0.082, or 8.2%—a much more meaningful expression of the achieved performance of the TES. Consequently, a device that appears to be ideal on an energy basis is correctly shown to be far from ideal on an exergy basis, clearly demonstrating the benefits of using exergy analysis for evaluating TESs.

7. COMPLEX SYSTEM: COAL-FIRED ELECTRICAL GENERATING STATION

Energy and exergy analyses are applied to a coal-fired electrical generating station. The coal-fired Nanticoke Generating Station is selected as the representative station for the analysis. The station has a net unit electrical output of approximately 500 MWe and is operated by the provincial electrical utility, Ontario Power Generation (formerly Ontario Hydro), in Ontario, Canada. This example clearly illustrates how exergy analysis allows process inefficiencies to be better determined than does an energy analysis and efficiencies to be more rationally evaluated. Consequently, the example identifies areas in which the potential for performance improvement is high and trends that may aid in the design of future stations.

7.1 Process Description

A detailed flow diagram for a single unit of the station is shown in Fig. 2. The symbols identifying the streams are described in Table I, in which data are provided for the streams. The diagram in Fig. 2 is divided into four main sections:

- *Steam generation*: Heat is produced and used to generate and reheat steam. In the station, eight pulverized coal-fired natural circulation steam generators each produce 453.6 kg/s steam at 16.89 MPa and 538°C and 411.3 kg/s of reheat steam at 4.00 MPa and 538°C. Air is supplied to the furnace by two 1080-kW, 600-rpm motor-driven forced-draft fans. Regenerative air preheaters are used. The flue gas passes through an electrostatic precipitator rated at 99.5% collection efficiency and exits the plant through two multiflued, 198-m-high chimneys.
- *Power production*: The steam produced in the steam generation section is passed through a series of turbine generators that are attached to a transformer. Extraction steam from several points on the turbines preheats feedwater in several low- and high-pressure heat exchangers and one spray-type open deaerating heat exchanger. The low-pressure turbines exhaust to the condenser at 5 kPa. Each station unit has a 3600-rpm, tandem-compound, impulse-reaction turbine generator containing one single-flow high-pressure cylinder, one double-flow intermediate-pressure cylinder, and two double-flow low-pressure cylinders. Steam exhausted from the high-pressure cylinder is reheated in the combustor.
- *Condensation*: Cooling water from Lake Erie condenses the steam exhausted from the turbines. The flow rate of cooling water is adjusted so that a specified temperature increase in the cooling water is achieved across the condenser.
- *Preheating*: The temperature and pressure of the feedwater are increased in a series of pumps and feedwater–heater heat exchangers.

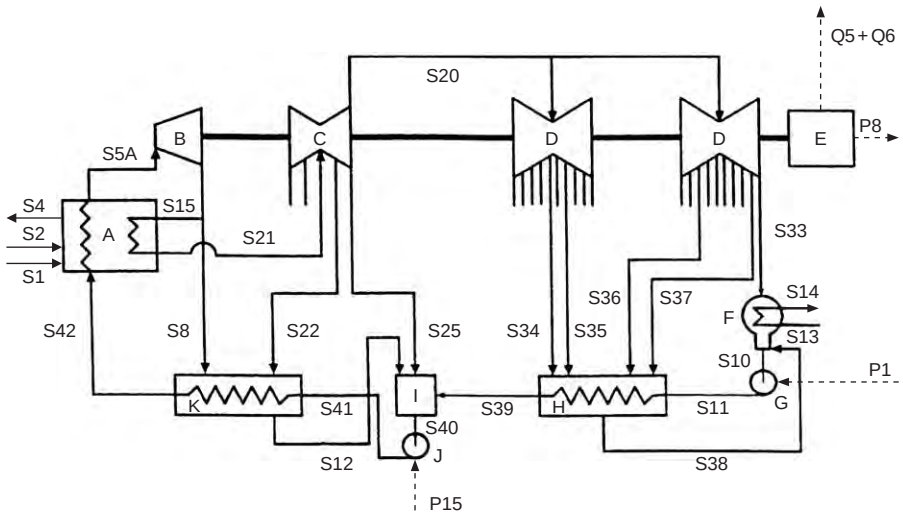


FIGURE 2 A unit of the coal-fired electrical generating station. Lines exiting the turbines represent extraction steam. The station has four main sections: Steam generation (A), power production (B–E), condensation (F), and preheating (G–K). The external inputs for device A are coal and air, and the output is stack gas and solid waste. The external outputs for device E are electricity and waste heat. Electricity is input to devices G and J, and cooling water enters and exits device F. A, steam generator and reheater; B, high-pressure turbine; C, intermediate-pressure turbine; D, low-pressure turbines; E, generator and transformer; F, condenser; G, hot well pump; H, low-pressure heat exchangers; I, open deaerating heat exchanger; J, boiler feed pump; K, high-pressure heat exchangers.

7.2 Analysis Assumptions and Simplifications

The reference environment model used here has a temperature of 15°C (the approximate mean temperature of the lake cooling water), a pressure of 1 atm, a chemical composition consisting of air saturated with water vapor, and the following condensed phases at 15°C and 1 atm: water (H_2O), gypsum ($CaSO_4 \cdot 2H_2O$), and limestone ($CaCO_3$). For simplicity, heat losses from external surfaces are assumed to occur at the reference environment temperature of 15°C.

7.3 Results and Discussion

Energy and exergy values for the streams identified in Fig. 2 are summarized in Table I. Exergy consumption values for the devices are listed, according to process diagram sections, in Table II. Figures 3 and 4 illustrate the net energy and exergy flows and exergy consumptions for the four main process diagram sections described in the legend to Fig. 2.

Overall energy and exergy efficiencies are evaluated as

$$\text{Energy efficiency} = \frac{(\text{net energy output with electricity})}{(\text{energy input})} \qquad (3)$$

and

$$\text{Exergy efficiency} = \frac{(\text{net exergy output with electricity})}{(\text{exergy input})}. \qquad (4)$$

Coal is the only input source of energy or exergy for the station, and the energy and exergy efficiencies are 37 and 36%, respectively. The small difference in the efficiencies is due to the fact that the specific chemical exergy of coal is slightly greater than its energy. Although the energy and exergy efficiencies for the station are similar, these efficiencies differ markedly for many station sections.

In the steam generation section, exergy consumptions are substantial, accounting for 659 MW (72%) of the 916 MW total exergy losses for the station. Of the 659 MW of exergy consumed in this section, 444 MW is due to combustion and 215 MW to heat transfer. The energy and exergy efficiencies for the steam generation section, considering the increase in energy or exergy of the water as the product, are 95 and 49%, respectively. The steam generation section thus appears significantly more efficient on an energy basis than on an exergy basis. Physically, this discrepancy implies that although 95% of the input energy is transferred to the preheated water, the energy is degraded as it is transferred. Exergy analysis highlights this degradation.

TABLE I

Flow Data for a Unit of the Coal-Fired Electrical Generating Station

Stream	Mass flow rate (kg/s)[a]	Temperature (°C)	Pressure (N/m^2)	Vapor fraction[b]	Energy flow rate (MW)	Exergy flow rate (MW)
S1	41.74	15.00	1.01×10^5	Solid	1367.58	1426.73
S2	668.41	15.00	1.01×10^5	1.0	0.00	0.00
S3[c]	710.15	1673.59	1.01×10^5	1.0	1368.00	982.85
S4	710.15	119.44	1.01×10^5	1.0	74.39	62.27
S5A	453.59	538.00	1.62×10^7	1.0	1585.28	718.74
S8	42.84	323.36	3.65×10^6	1.0	135.44	51.81
S10	367.85	35.63	4.50×10^3	0.0	36.52	1.20
S11	367.85	35.73	1.00×10^6	0.0	37.09	1.70
S12	58.82	188.33	1.21×10^6	0.0	50.28	11.11
S13	18,636.00	15.00	1.01×10^5	0.0	0.00	0.00
S14	18,636.00	23.30	1.01×10^5	0.0	745.95	10.54
S15	410.75	323.36	3.65×10^6	1.0	1298.59	496.81
S20	367.85	360.50	1.03×10^6	1.0	1211.05	411.16
S21	410.75	538.00	4.00×10^6	1.0	1494.16	616.42
S22	15.98	423.23	1.72×10^6	1.0	54.54	20.02
S25	26.92	360.50	1.03×10^6	1.0	88.64	30.09
S33	309.62	35.63	4.50×10^3	0.93	774.70	54.07
S34	10.47	253.22	3.79×10^5	1.0	32.31	9.24
S35	23.88	209.93	2.41×10^5	1.0	71.73	18.82
S36	12.72	108.32	6.89×10^4	1.0	35.77	7.12
S37	11.16	60.47	3.45×10^4	1.0	30.40	5.03
S38	58.23	55.56	1.33×10^4	0.0	11.37	0.73
S39	367.85	124.86	1.00×10^6	0.0	195.94	30.41
S40	453.59	165.86	1.00×10^6	0.0	334.86	66.52
S41	453.59	169.28	1.62×10^7	0.0	347.05	77.57
S42	453.59	228.24	1.62×10^7	0.0	486.75	131.93
Q5					5.34	0.00
Q6					5.29	0.00
P1					0.57	0.57
P8					523.68	523.68
P15					12.19	12.19

[a] The composition of all streams is 100% H_2O, except that on a volume basis the composition of S1 is 100% carbon, that of S2 is 79% N_2 and 21% O_2, and that of both S3 and S4 is 79% N_2, 6% O_2, and 15% CO_2.

[b] Vapor fraction is listed as 0.0 for liquids and 1.0 for superheated vapors.

[c] Stream S3 (not shown in Fig. 2) represents the hot product gases for adiabatic combustion.

A large quantity of energy (775 MW for each unit) enters the condensers, of which nearly 100% is rejected, and a small quantity of exergy (54 MW for each unit) enters, of which approximately 25% is rejected and 75% is internally consumed. Thus, energy analysis results lead to the erroneous conclusion that almost all losses in electricity-generation potential for the station are associated with the heat rejected by the condensers, whereas exergy analysis demonstrates quantitatively and directly that the condensers are responsible for little of these losses (Fig. 4). This discrepancy arises because heat is rejected by the condensers at a temperature very near that of the environment.

In the power production and preheating sections, energy losses were found to be very small (< 10 MW total), and exergy losses were moderately small (118 MW in the power production section and 23 MW in the preheating section). The exergy losses

TABLE II

Breakdown by Section and Device of Exergy Consumption Rates for a Unit of the Coal-Fired Electrical Generating Station

Section/device	Exergy consumption rate (MW)
Steam generation section	
Steam generator (including combustor)	659.0
Power production section	
High-pressure turbine	26.4
Intermediate-pressure turbine	22.3
Low-pressure turbines	59.2
Generator	5.3
Transformer	5.3
Total	118.5
Condensation section	
Condenser	43.1
Preheat section	
Low-pressure heat exchangers	10.7
Deaerating heat exchanger	5.1
High-pressure heat exchangers	6.4
Hot well pumps	0.1
Boiler feed pumps	1.1
Total	23.4
Total consumption rate	844.0

are almost completely associated with internal consumptions.

In assessing the thermodynamic characteristics of a coal-fired electrical generating station, several insights into the performance of such stations have been obtained. First, although energy and exergy efficiencies are similar for the station, energy analyses do not systematically identify the location and cause of process inefficiencies, whereas exergy analyses do. Specifically, energy losses are associated with emissions (mainly heat rejected by condensers) and exergy losses primarily with consumptions (mainly in the combustors). Second, since devices with the largest thermodynamic losses have the largest margins for efficiency improvement, efforts to increase the efficiencies of coal-fired electrical generating stations should focus on the combustors. For instance, technologies capable of producing electricity without combustion (e.g., fuel cells) or utilizing heat at high temperatures could increase efficiencies significantly. This suggestion, of course, is overly simplistic because such decisions require consideration of other technical and economic factors in addition to thermodynamic efficiency.

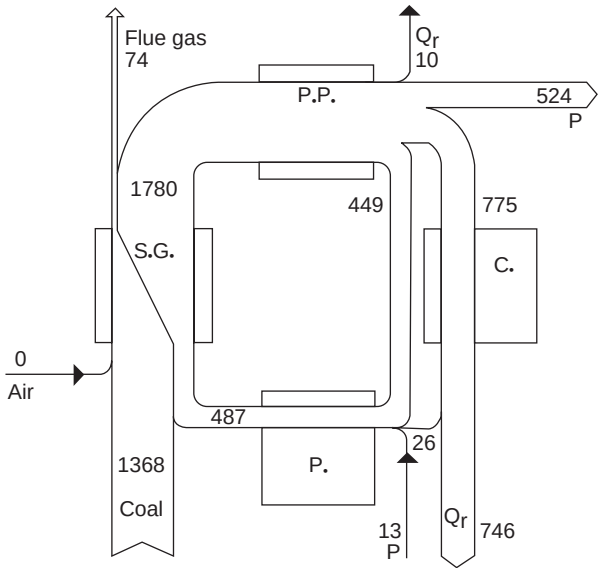


FIGURE 3 Simplified energy diagram for a unit of the coal-fired electrical generating station, indicating net energy flow rates (MW) for streams. Stream widths are proportional to energy flow rates. Station sections shown are steam generation (S.G.), power production (P.P.), condensation (C.), and preheating (P.). Streams shown are electrical power (P), heat input (Q), and heat rejected (Q_r).

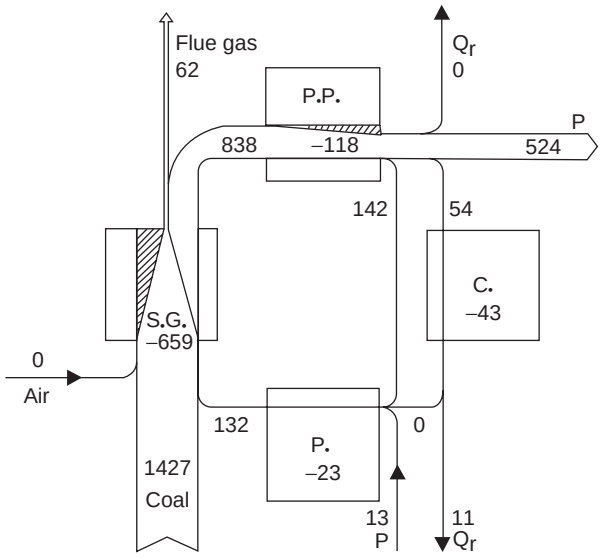


FIGURE 4 Simplified exergy diagram for a unit of the coal-fired electrical generating station, indicating net exergy flow rates for streams and consumption rates (negative values) for devices (MW). Stream widths are proportional to exergy flow rates, and shaded regions are proportional to exergy consumption rates. Other details are as in the legend to Fig. 3.

Third, the use of heat rejected by condensers only increases the exergy efficiencies by a few percent.

8. MACROSYSTEM: THE AGGREGATE ENERGY SYSTEMS OF A COUNTRY

The energy-utilization efficiency of a macrosystem, such as a region or country, can be assessed beneficially using exergy analysis. This section illustrates how the efficiency of energy utilization in a particular macrosystem, Canada, can be examined using both energy and exergy analyses. This example is based on a 1992 publication by the author titled "Evaluation of Energy Utilization Efficiency in Canada Using Energy and Exergy Analyses," which used Canadian energy data for 1986.

The relation between energy resources and sectors of a macrosystem can be modeled as in Fig. 5. The energy resources considered here are coal, petroleum, natural gas, nuclear energy, and hydraulic energy. Renewable energy resources (e.g., wind, solar, wave, and tidal) are neglected since the quantity used in Canada is minor relative to other resources. Four sectors are considered. The residential–commercial sector, which is driven mainly by fossil fuels and electricity, includes residential, public administration, commercial, and other institutional buildings and involves as principal devices space heaters, water heaters, cooking appliances, and clothes dryers. The transportation sector is primarily driven by fossil fuels and includes rail, marine, air, road and urban transport, pipelines, and retail pump sales. The industrial sector has as its main tasks work production and heating and includes mines, steam generation, pulp and paper mills, cement, iron and steel mills, petroleum refining, smelting and refining, other manufacturing, agriculture, chemicals, and nonenergy uses. The utility sector includes processes for electricity generation from coal, petroleum, natural gas, and nuclear and hydraulic energy sources. The same reference environment model as used in the previous section is used here but with a reference environment temperature of 25°C (except where noted).

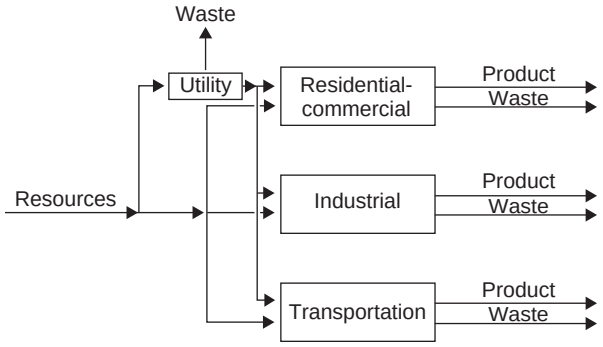


FIGURE 5 Illustration of the energy flows in a macrosystem.

8.1 Principal Types of Processes and Their Efficiencies

Energy and exergy efficiencies are based on the following general definitions:

$$\text{Energy efficiency} = \frac{(\text{energy in products})}{(\text{total energy input})} \quad (5)$$

and

$$\text{Exergy efficiency} = \frac{(\text{exergy in products})}{(\text{total exergy input})}. \quad (6)$$

The main types of processes encountered, for which energy and exergy efficiencies are evaluated, are as follows:

- Heating: Electric and fossil fuel heating processes are taken to generate product heat at a constant temperature.
- Work production: Electric and fossil fuel work production processes produce shaft work.
- Electricity generation: Electricity is generated from fossil fuel, nuclear, and hydraulic sources. The hydraulic source is taken to be the change in kinetic and potential energy of flowing water (even though coal equivalents are often used to represent hydraulic energy inputs). The nuclear source is taken to be nuclear-derived heat at a temperature so high that the energy and exergy of the nuclear-derived heat are equivalent. The efficiency definitions for nuclear plants used here follow nuclear industry conventions (although these efficiency definitions are inadequate because they are based on the heat released from the nuclear fuel rather than its energy or exergy content).
- Kinetic energy production: Fossil fuel-driven kinetic energy production processes occur in some transportation-sector devices (e.g., turbojet engines) and increase the kinetic energy of matter.

8.2 Sector Analyses

8.2.1 Transportation Sector

In the first four columns of Table III, a breakdown is presented by mode of transport of sector energy use, whereas the last two columns show corresponding energy efficiencies. Since vehicles generally are not operated at full load, a distinction is made between

TABLE III

Transportation Sector Data

Mode of transport	Main fuel types	Energy used in the sector, by mode of transport		Energy efficiences (%)	
		PJ	%	Rated load	Estimated operating
Rail	Diesel	75.39	4.6	35	28
Air	Jet fuel	131.2	8.1	35	28
Marine	Diesel, gasoline	74.82	4.6	—	15
Road and urban transit	Gasoline, diesel, propane, electricity	164.3	10.1	28	22
Retail pump sales	Gasoline, diesel, propane	1093	67.2	25	15
Pipelines	Natural gas	86.21	5.3	—	29

TABLE IV

Electricity-Generation Data for the Utility and Industrial Sectors

Mode of generation	Electrical energy generated (PJ)[a]	Energy input (PJ)	Energy efficiency (%)
Hydraulic[b]	1000	1176	85
Nuclear	230.1	767.0	30
Fossil	293.0	915.6	32

[a] Hydraulic and nuclear generation is entirely within the utility sector, whereas 262.3 PJ of fossil generation is within the utility sector and the remainder is within the industrial sector. The contribution by different fossil fuels is 91.2% coal, 6.7% petroleum, and 2.1% natural gas.

[b] The hydraulic source is taken to be the change in kinetic and potential energy of flowing water and not a fossil fuel equivalent.

TABLE V

Energy Consumption Data for the Residential–Commercial Sector

Devices	Energy used in the sector, by type (%)		
	Electrical	Fuel	Total
Space heaters	15.0	79.5	66.0
Water heaters	14.0	15.0	10.0
Cooking appliances	5.0	4.5	1.7
Clothes dryers	5.0	1.0	1.3
Others[a]	61.0	0.0	21.0

[a] This category includes fans, pumps, air conditioners, refrigerators, freezers, washers, lights, etc.

rated load (full load) and operating (part load) efficiencies. The energy efficiencies in Table III are equal to the corresponding exergy efficiencies. Weighted means are obtained for the transportation-mode energy efficiencies in Table III, where the weighting factor is the fraction of the total sector energy input that supplies each transportation mode (fourth column in Table III). Thus, the weighted mean overall efficiency for the transportation sector based on energy, which is equal to that based on exergy, is 18%.

8.2.2 Utility Sector

For the different types of electricity generation, Table IV lists energy efficiencies and the quantities of electricity generated and energy inputs. The energy efficiencies in Table IV are equal to the corresponding exergy efficiencies. Using data from this table, the mean overall energy and exergy efficiencies for the utility sector are found to be equal at 53%.

8.2.3 Residential–Commercial Sector

Energy end-use quantities for this sector are listed in Table V. Operating data for the principal devices in the sector are listed in Table VI and are used to determine the device exergy efficiencies. Weighted mean energy and exergy efficiencies are calculated for the overall residential–commercial sector using a three-step process. First, weighted means are obtained for the electrical energy and exergy efficiencies for the devices in Table VI, where the weighting factor in both cases is the ratio of the electrical energy input to the device to the total electrical energy input to all of the devices (as given in Table V). These energy and exergy efficiencies are 89 and 20%, respectively. Second, weighted mean efficien-

TABLE VI

Device Data for the Residential–Commercial Sector[a]

	Product-heat temperature (K)		Heating efficiency (%)			
			Electrical		Fuel	
Devices	Electrical	Fuel	Energy	Exergy	Energy	Exergy
Space heaters	328	328	100.0	17.1	65.0	11.1
Water heaters	350	374	93.0	25.4	62.0	14.0
Cooking appliances	394	—	80.0	22.5	—	—
Clothes dryers	350	350	50.0	9.6	50.0	10.3

[a] Corresponding reference environment temperatures are 272 K for space heating and 283 K for all other devices.

cies for the fossil fuel-driven devices in Table VI are similarly determined as 64 and 12%, respectively. Third, overall weighted means are obtained for the energy and exergy efficiencies for the electrical and fossil fuel processes together, where the weighting factor is the ratio of the total energy (either fossil fuel or electrical) input to the residential–commercial sector to the total energy input to the sector. Since 34% of the 2129 PJ input to the residential–commercial sector is electrical energy and the remaining 66% is fuel energy, the weighted mean overall energy and exergy efficiencies for the sector are 73 and 14%, respectively. These means are representative of the sector since the devices in Table VI account for more than half of the total sector energy consumption.

8.2.4 Industrial Sector

The industrial sector is composed of many industries, each of which encompasses a wide variety of complex plants. The five most significant industries were identified, and a single representative commercial process for each was examined. The identified industries with the selected commercial processes are as follows:

- Pulp and paper (integrated paper mill plant)
- Iron and steel (blast furnace and steel mill plant)
- Cement (hydraulic cement factory)
- Petroleum (petroleum refining plant)
- Chemical [NEC (salt) chemical preparation plant]

These five industries are taken to be representative of the sector since they account for approximately half of the sector energy use.

Only heating and mechanical processes are considered since these processes account for 81% of the industrial-sector energy consumption (66% for heating and 15% for mechanical drives). Following the temperature categories in Table VII, process heat data for the five selected commercial processes are listed in Table VIII, with the percentages of the heat in each range supplied by fossil fuels and electricity. Canadian energy consumption in the form of electricity and fossil fuels for each of the five main industries is listed in Table IX, with a breakdown of the total energy demand. The values are obtained by averaging the percentage of electrical and fuel energy demands for each of the industries shown. Subsequently, steam (from utilities) is treated as an industrial-sector product and thus passes through the sector unchanged, and electricity generation in the industrial sector is included in the utility sector.

TABLE VII

Process Heating Temperature Range Categories and Energy Efficiencies for the Industrial Sector

Product heat temperature category	Product heat temperature range (°C)[a]	Heating energy efficiency (%)	
		Electrical	Fuel
Low	<121	100	65
Medium	121–399	90	60
High	>399	70	50

[a] Temperature ranges were determined subjectively after examining many commercial processes.

Three steps are used in evaluating industry efficiencies. First, energy and exergy efficiencies are obtained for process heating for each of the temperature range categories in Table VIII using data from Tables VII and VIII. Second, mean heating energy and exergy efficiencies for each of the five industries (see the fifth and sixth columns of Table IX) are evaluated using a two-step procedure: (i)

TABLE VIII

Process Heating Data for All Categories of Product Heat Temperature for the Industrial Sector

Industry	Product heat temperature category	Mean product heat temperature (°C)	Energy used in each product heat temperature category, by type (%)	
			Electricity	Fuel
Pulp and paper	Low	67	100.0	0.0
	Medium	160	0.0	83.0
	High	732	0.0	17.0
Iron and steel	Low	45	4.2	0.0
	Medium	—	0.0	0.0
	High	983	95.8	100.0
Cement	Low	31	91.7	0.9
	Medium	163	0.0	9.0
	High	1482	8.3	90.1
Petroleum	Low	57	10.1	13.8
	Medium	227	9.4	22.6
	High	494	80.4	63.6
Chemical	Low	42	62.5	0.0
	Medium	141	37.5	100.0
	High	—	0.0	0.0

TABLE IX

Process Data for the Industrial Sector

Industry	Energy used in sector, by type			Energy and exergy efficiencies (%)			
				Heating		Total	
	Electrical (PJ)	Fuel (PJ)	Total (%)	Energy	Exergy	Energy	Exergy
Pulp and paper	181.4	169.6	16.8	80.0	20.4	82.0	33.0
Iron and steel[a]	155.7	284.5	17.2	58.0	43.4	64.0	52.0
Cement	5.38	52.29	1.4	56.0	37.5	62.3	47.0
Petroleum	16.45	87.81	2.9	65.0	29.4	70.0	41.0
Chemical	63.00	140.8	7.4	71.0	18.5	74.5	32.0

[a] This category includes smelting and refining industrial operations.

Weighted mean efficiencies for electrical heating and for fuel heating are evaluated for each industry in Table VIII using as weighting factors the values listed in the last two columns of Table VIII, and (ii) weighted mean efficiencies for all heating processes in each industry are evaluated with these values using as weighting factors the ratio of the industry energy consumption (electrical or fuel) to the total consumption of both electrical and fuel energy (Table IX). Third, weighted mean overall (i.e., heating and mechanical drive) efficiencies for each industry are evaluated using the values in Table IX, assuming mechanical drive energy efficiencies of 90% and using as weighting factors the fractions of the total sector energy input for both heating and mechanical drives (as noted earlier).

Weighted means for the weighted mean overall energy and exergy efficiencies (listed in the last two columns of Table IX) for the five principal industries in the industrial sector are obtained, using as weighting factors the fraction of the total industrial energy demand supplied to each industry (given in the fourth column of Table IX), as 73 and 41%, respectively.

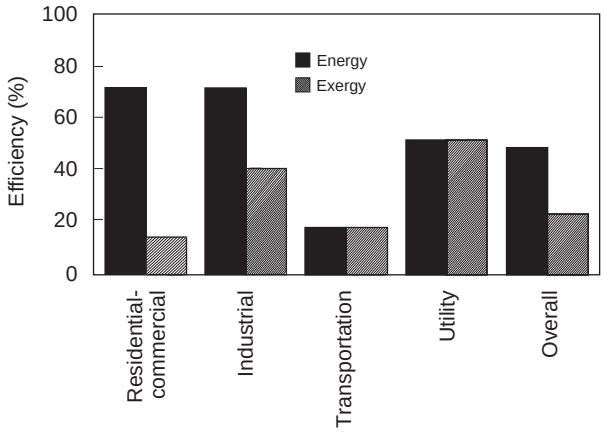


FIGURE 6 Energy and exergy efficiencies for the four main sectors of the Canadian economy and for the overall economy.

8.2.5 Sector Comparison

The efficiencies for the four sectors and for the overall Canadian economy are compared in Fig. 6. The energy efficiencies range from 18 to 73% and the exergy efficiencies from 14 to 53%. The most efficient sectors on energy and exergy bases are different. The residential–commercial sector exhibits the greatest differences for energy and exergy evaluations. The most significant differences between energy and exergy efficiencies are attributable to heating processes. For example, the overall exergy efficiency of the residential–commercial sector (14%) is much lower than the overall energy efficiency (73%) because high-quality energy (i.e., high exergy) sources are used for low-quality energy (i.e., low exergy) demands. The higher exergy efficiency of the industrial sector is mainly attributable to the high temperatures required for most industrial heating ($\approx$400°C) compared to residential–commercial heating ($\approx$25°C). The industrial sector thus utilizes relatively more of the quality or work potential of fuels.

8.3 Analyses for the Overall Macrosystem

Figures 7 and 8 illustrate, respectively, the flows of energy and exergy in Canada for the year considered. The left sides of the figures consist of the total energy or exergy inputs to all sectors by fuel type. These quantities are then branched according to sector demand. Fuels such as coal, gas, and petroleum are routed to "fuel for electricity" and include fuels used to produce electricity by utilities as well as by industry. Similarly, "fuel for end uses" represents the net fuel requirement for all sectors and is routed together with electricity and steam flows to the sectors. The final branching into useful and waste energy is achieved by multiplying the energy for end use (fuel, steam, and electricity) in each sector by the corresponding efficiency (see the values summarized in Fig. 6).

The energy flow diagram indicates that nearly 50% of the total energy consumed in Canada is

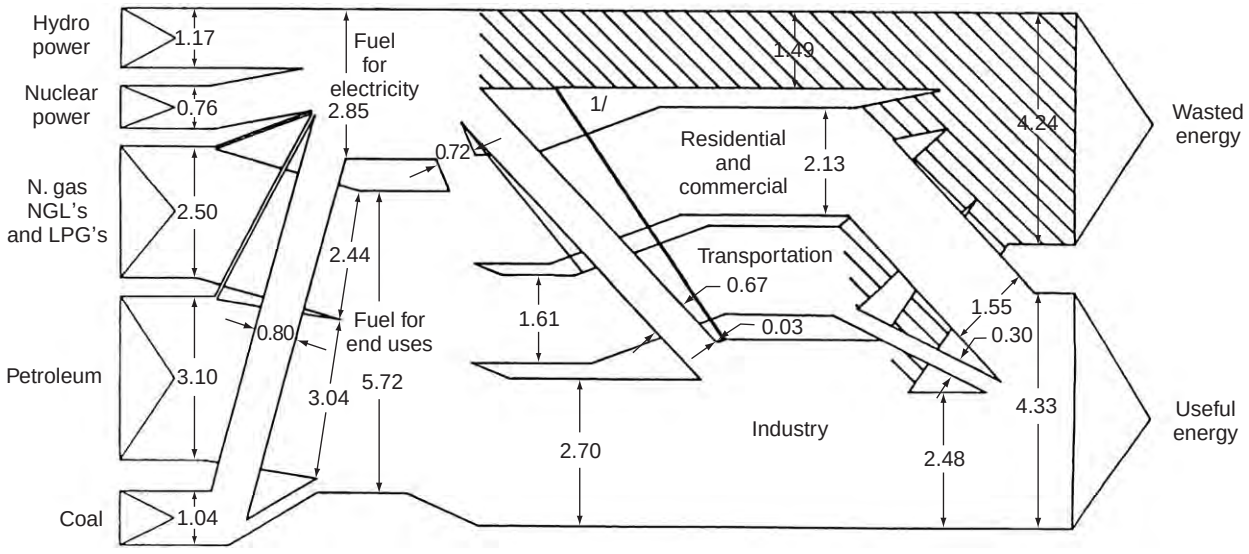


FIGURE 7 Flows of energy (EJ) in Canada. Widths of lines are proportional to the magnitude of the quantity represented. Hydraulic energy is reported in kinetic energy equivalent. The portions of coal, petroleum, and natural gas used for industrial-sector electricity generation are routed to the utility sector and then back to the industrial sector in the form of electricity. The flow associated with the symbol 1/represents steam produced in the utility sector for the end-use sectors.

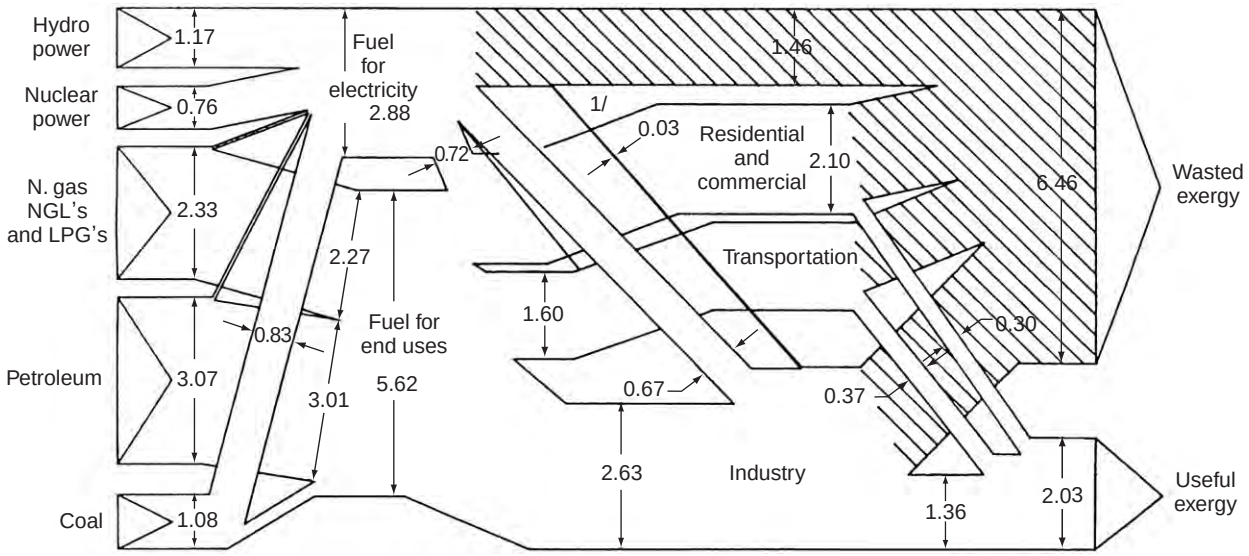


FIGURE 8 Flows of exergy (EJ) in Canada. Details are as given in the legend to Fig. 7.

converted to useful energy for end uses, and that the transportation and utility sectors produce the most waste. The exergy flow diagram differs markedly from the energy diagram in that only 24% of Canadian exergy consumption is converted to useful exergy for end uses, with the residential–commercial and industrial sectors producing the greatest waste. Thus, energy analysis indicates a more limited margin for efficiency improvement than exergy analysis. However, only exergy analysis indicates the margin for improvement that actually exists. Therefore, by using exergy rather than energy analysis, energy utilization efficiency in the macrosystem is more clearly illuminated, and more rational assessments are obtained for allocations of energy-related research and development (R&D).

9. CONCLUSIONS

Exergy analyses of energy systems provide useful information that can have direct implications on designs, improvements, and application decisions. Furthermore, exergy analyses, more than energy analyses, provide insights into better directions for R&D. Here, "better" is loosely taken to mean "more promising for significant efficiency gains." There are two main reasons for this statement:

- Exergy losses represent true losses of the potential that exists to generate the desired product from the given driving input. This is not true in general for energy losses. Thus, if the objective is to increase efficiency while accounting for energy degradation, focusing on exergy losses permits R&D to focus on reducing losses that will affect the objective.
- Exergy efficiencies always provide a measure of how closely the operation of a system approaches the ideal, or theoretical upper limit. This is not generally true for energy efficiencies. By focusing R&D on those plant sections or processes with the lowest exergy efficiencies, effort is directed to those areas that inherently have the largest margins for efficiency improvement. By focusing on energy efficiencies, on the other hand, one can inadvertently expend R&D effort on areas for which little margins for improvement, even theoretically, exist.

Exergy analysis results typically suggest improvement efforts should concentrate more on internal rather than external exergy losses, based on thermodynamic considerations, with a higher priority for the processes having larger exergy losses. Note that this statement does not indicate that effort should not be devoted to processes having low exergy losses because simple and cost-effective ways to increase efficiency by reducing small exergy losses should certainly be considered when identified.

In general, application and R&D decisions should not be based exclusively on the results of energy and exergy analyses, even though these results provide useful information to assist in such decision making. Other factors must also be considered, such as economics, environmental impact, safety, and social and political implications.

SEE ALSO THE FOLLOWING ARTICLES

Conservation of Energy, Overview • Entropy • Entropy and the Economic Process • Exergoeconomics • Exergy • Exergy Analysis of Waste Emissions • Exergy: Reference States and Balance Conditions • Thermodynamic Sciences, History of • Thermodynamics, Laws of

Further Reading

Brodyanski, V. M., Sorin, M. V., and Le Goff, P. (1994). "The Efficiency of Industrial Processes: Exergy Analysis and Optimization." Elsevier, London.

Dincer, I., and Rosen, M. A. (2002). Energy and exergy analyses of thermal energy storage systems. *In* "Thermal Energy Storage Systems and Applications," pp. 411–510. Wiley, London.

Kotas, T. J. (1995). "The Exergy Method of Thermal Plant Analysis," Reprint. Krieger, Malabar, FL.

Moran, M. J. (1989). "Availability Analysis: A Guide to Efficient Energy Use." Rev. ed. American Society of Mechanical Engineers, New York.

Moran, M. J., and Sciubba, E. (1994). Exergy analysis: Principles and practice. *J. Eng. Gas Turbines Power* **116,** 285–290.

Moran, M. J., and Shapiro, H. N. (2000). "Fundamentals of Engineering Thermodynamics." 4th ed. Wiley, New York.

Rosen, M. A. (1992). Evaluation of energy utilization efficiency in Canada using energy and exergy analyses. *Energy Int. J.* **17,** 339–350.

Rosen, M. A. (1999). Second law analysis: Approaches and implications. *Int. J. Energy Res.* **23,** 415–429.

Rosen, M. A. (2001). Energy- and exergy-based comparison of coal-fired and nuclear steam power plants. *Exergy Int. J.* **1,** 180–192.

Rosen, M. A., and Dincer, I. (1997). Sectoral energy and exergy modeling of Turkey. *ASME J. Energy Resources Technol.* **119,** 200–204.

Rosen, M. A., and Horazak, D. A. (1995). Energy and exergy analyses of PFBC power plants. *In* "Pressurized Fluidized Bed Combustion" (M. Alvarez Cuenca and E. J. Anthony, Eds.), pp. 419–448. Chapman & Hall, London.

Szargut, J., Morris, D. R., and Steward, F. R. (1988). "Exergy Analysis of Thermal, Chemical and Metallurgical Processes." Hemisphere, New York.

Yantovskii, E. I. (1994). "Energy and Exergy Currents: An Introduction to Exergonomics." Nova Science, New York.

Exergy Analysis of Waste Emissions

MARC A. ROSEN
University of Ontario Institute of Technology
Oshawa, Ontario, Canada

1. **Exergy and Exergy Analysis**
2. **Energy and the Environment**
3. **Exergy and the Environment**
4. **Relations between Exergy and Environmental Impact**
5. **Environmental Impact of Exergy Losses**
6. **Comparison of Exergy and Other Indicators for the Environmental Impact of Waste Emissions**
7. **Conclusions**

Glossary

chemical exergy The maximum work obtainable from a substance when it is brought from the environmental state to the dead state by means of processes involving interaction only with the environment.

dead state The state of a system when it is in thermal, mechanical, and chemical equilibrium with a conceptual reference environment (having intensive properties pressure P_o, temperature T_o, and chemical potential μ_{ko} for each of the reference substances in their respective dead states).

environmental pollution cost (EPC) The environmental cost of a pollutant, based on such factors as the cost to society for the correction or compensation of environmental damage and the cost incurred to prevent a harmful emission from escaping into the environment.

environomics A methodology for analyzing and improving energy-related systems by taking into account energy, exergy, economic, and environmental factors.

environmental state The state of a system when it is in thermal and mechanical equilibrium with the reference environment (i.e., at pressure P_o and temperature T_o of the reference environment).

exergy (i) A general term for the maximum work potential of a system, stream of matter, or a heat interaction in relation to the reference environment as the datum state. Also known as available energy, availability, essergy, technical work capacity, usable energy, utilizable energy, work capability, work potential, and xergy. (ii) The unqualified term exergy or exergy flow is the maximum amount of shaft work obtainable when a steady stream of matter is brought from its initial state to the dead state by means of processes involving interactions only with the reference environment.

exergy analysis An analysis technique, based on the second law of thermodynamics, in which process performance is assessed by examining exergy balances, losses, and efficiencies.

exergy consumption The exergy consumed or destroyed during a process due to irreversibilities within the system boundaries. Also known as dissipation, irreversibility, and lost work.

reference environment An idealization of the natural environment that is characterized by a perfect state of equilibrium—that is, the absence of any gradients or differences involving pressure, temperature, chemical potential, kinetic energy, and potential energy. The environment constitutes a natural reference medium with respect to which the exergy of different systems is evaluated.

reference state A state with respect to which values of exergy are evaluated. Several reference states are used, including environmental state, dead state, standard environmental state, and standard dead state. Also known as ground state.

removal pollution cost (RPC) The cost of removing a pollutant from a waste stream prior to discharge into the environment.

waste emission Release to the environment of a material that is not considered usable and is therefore being discarded.

Exergy analysis is a thermodynamic technique for assessing and improving systems and processes, which is similar but advantageous to energy analysis, in large part because it is based on the second law of thermodynamics. The exergy of an energy form or a substance is a measure of its usefulness. Exergy can be applied beyond thermodynamics to understanding waste emissions and reducing their environmental

 © 2004 Elsevier Inc. All rights reserved.

impact because exergy can also be viewed as a measure of the departure of a substance from equilibrium with a specified reference environment, which is often modeled as the actual environment. Furthermore, exergy is a measure of the potential of a substance to cause change. The exergy of an emission to the environment, therefore, is a type of measure of the potential of the substance to change or impact the environment. The greater the exergy of an emission, the greater its departure from equilibrium with the environment and the greater may be its potential to change or impact the environment. The exergy of an emission is zero only when it is in equilibrium with the environment and thus benign. These points suggest that exergy may be, or provide the basis for, one effective indicator of the potential of an emitted substance to impact the environment.

This article describes both the relation between exergy and the environment and the exergy of waste emissions. Such knowledge can permit the development of a better understanding of, and more rational approaches to mitigating, the environmental impact associated with waste emissions. First, exergy and its use as a tool through exergy analysis are briefly described. Then, energy and the environment and exergy and the environment, and some relations between them, are discussed. Finally, exergy is compared with other indicators for the environmental impact of waste emissions.

1. EXERGY AND EXERGY ANALYSIS

Exergy is defined as the maximum amount of work that can be produced by a flow of matter or energy as it comes to equilibrium with a reference environment. Exergy is a measure of the potential of a flow to cause change, as a consequence of being in disequilibrium with the reference environment. For evaluations of exergy, the state of the reference environment, or the reference state, must be specified completely, normally by specifying its temperature, pressure, and chemical composition. The reference environment used in exergy evaluations is discussed further in Section III.B. Exergy is not subject to a conservation law. Rather, exergy is consumed or destroyed due to the irreversibilities in any process.

Exergy analysis is a method that uses the conservation of mass and conservation of energy principles together with the second law of thermodynamics for the design and analysis of energy systems. The exergy method is useful for attaining more efficient energy-resource use because it enables the locations, types, and true magnitudes of wastes and losses to be determined. Therefore, exergy analysis can reveal whether, and by how much, it is possible to design more efficient energy systems by reducing the inefficiencies in existing systems. Applications of exergy analysis to many processes and systems are described elsewhere in this encyclopedia.

Exergy concepts can also be applied to disciplines outside of thermodynamics. One area in which applications of exergy are increasing is that of economics. Another area in which applications of exergy are increasing is that of environmental emissions and impact, and this is the topic of the current article.

2. ENERGY AND THE ENVIRONMENT

Energy production, transformation, transport, and end use have significant impacts on the earth's environment, and environmental costs are usually associated with the emissions—whether they are of a thermal, chemical, nuclear, or other nature—from processes and technologies. One means of reducing the environmental impact of emissions is to increase the efficiency of resource utilization. However, increasing efficiency generally entails greater use of materials and energy resources, and the environmental burdens associated with these can offset the environmental gains of improved efficiency.

Data on the current world distribution of per capita energy consumption suggest that a close correlation exists between a country's energy consumption and economic development. However, there does not appear to be a simple correlation between a country's energy consumption and the decay of its natural environment. As environmental concerns became major issues in the 1980s, increasing attention was paid to the relationship between energy use and the environment. However, there are limited works on the relationships between the exergy and environmental concepts even though to appropriately reduce energy consumption and the related environmental impact, engineers, scientists, and policymakers may benefit from such knowledge.

One of the major challenges of designing technologies so as to reduce their environmental impact involves determining an environmentally optimal configuration or selecting the most appropriate one from competing options. This determination or

selection is made difficult by the complex relationship between the technology being considered and the characteristics of the residuals produced. Process or technology changes generally affect the flow rate and composition of all resulting effluent streams. Therefore, evaluating alternatives often requires one to compare the relative environmental merits of distinctly different residual streams. Existing systems for performing such analyses have focused primarily on subjective ranking techniques.

3. EXERGY AND THE ENVIRONMENT

There is agreement among some researchers that a less subjective metric capable of providing greater insight may exist in the form of the thermodynamic quantity exergy. Exergy analysis provides information about energy and material quality that is both descriptive and physically significant. Because of its origins within the thermodynamic community, however, few applications and investigations of the potential of exergy methods for reducing environmental impact have been performed.

Some work is nevertheless being carried out. For instance, the Consortium on Green Design and Manufacturing at the University of California, Berkeley, is carrying out a project titled "Exergy as an Environmental Indicator" in an effort to increase the practical application of exergy analysis to rectifying the problems associated with material and energy flows in industry. This work has focused on developing a generalizable technique to calculate exergy in an industrial setting, exploring the significance of environmental ground states and how they might be utilized to leverage results, and establishing the requisite databases for performing a wide range of highly diverse analyses. In addition, several projects have been carried out at the Delft University of Technology and University of Twente to determine whether exergy analysis can be used as an instrument for environmental policy development, especially for the comparison of alternative production chains. These studies are investigating if the linkages of exergy with phenomena such as pollution and dispersion can be converted into a reliable tool on which policy decisions can be based, and they are exploring how the environmental effects of processes can be linked to or expressed in terms of exergy changes. Also, a methodology known as environomics has been developed for analyzing and improving energy-related systems by simultaneously taking into account energy, exergy, economic, and environmental factors.

The exergy contents of waste emissions are more meaningful than the corresponding energy contents as measures of potential for environmental impact. By definition, material and energy flows only possess exergy when in disequilibrium with a reference environment. The exergy associated with waste emission streams has the potential to cause environmental damage, particularly when it is released in an unconstrained manner into the environment. Some believe that by considering the exergy content of a waste emission, rational and meaningful assessments can be made of the environmental impact potential of the emission.

3.1 Reducing Environmental Impact with Exergy

Efforts to increase energy and exergy efficiency can reduce environmental impact by reducing energy and exergy losses. Increased efficiency frequently contributes to achieving energy security in an environmentally acceptable way, partly by reducing emissions. Increased efficiency also reduces the requirement for new facilities for the production, transportation, transformation, and distribution of the various energy forms; these additional facilities all cause some environmental impacts. To further reduce environmental pollution, efficiency-improvement actions often need to be supported by pollution-control technologies and/or fuel substitution. Regional and national actions to improve exergy efficiency, in particular, can have major impacts on environmental protection.

One study investigated thermodynamics and sustainable development and noted that one tool likely to be important for obtaining sustainable development is exergy analysis. That study stated that whereas energy can never be destroyed, exergy can be destroyed and this exergy destruction (irreversibility) should be appropriately minimized to make development more sustainable. This study also showed that environmental effects associated with emissions and resource depletion can be expressed in terms of one indicator, which is based on physical principles.

Several researchers have suggested that a potentially useful way to link the second law and environmental impact is through exergy because it is a measure of the departure of the state of a system from that of the environment. The magnitude of the

exergy of a system depends on the states of both the system and the environment. This departure is zero only when the system is in equilibrium with its environment. Eminent thermodynamicist Myron Tribus suggested in 1971 that performing exergy analyses of the natural processes occurring on the earth could form a foundation for ecologically sound planning because it would indicate the disturbance caused by large-scale changes.

An understanding of the relations between exergy and the environment may reveal some of the underlying fundamental patterns and forces affecting changes in the environment and help researchers to deal better with environmental damage. Before examining such relations, however, it is useful to discuss the reference environment used in exergy analysis.

3.2 Exergy and the Reference Environment

The exergy Ex contained in a system may be written as

$$Ex = S(T - T_o) - V(P - P_o) + N_k(\mu_k - \mu_{ko}), \qquad (1)$$

where the intensive properties are temperature T, pressure P, and chemical potential of substance k, μ_k, and the extensive properties are entropy S, volume V, and number of moles of substance k, N_k. The subscript o denotes conditions of the reference environment. It is evident from this equation that the exergy of a system is zero when it is in equilibrium with the reference environment (i.e., when $T = T_o$, $P = P_o$, and $\mu_k = \mu_{ko}$ for all k).

Exergy analysis requires a reference environment model. Exergy is evaluated with respect to the reference environment model, with the exergy of a flow and system being dependent on the intensive properties of the reference environment. The exergy of the reference environment is zero, and the exergy of a stream or system is zero when it is in equilibrium with the reference environment. There are several theoretical characteristics of the reference environment. The reference environment is in stable equilibrium and has all parts at rest relative to one another. No chemical reactions can occur between the reference environment components. Furthermore, the reference environment acts as an infinite system, is a sink and source for heat and materials, and experiences only internally reversible processes in which its intensive state remains unaltered (i.e., its temperature, pressure, and chemical potentials for each of the components present remain constant).

The natural environment does not possess the theoretical characteristics of a reference environment. The natural environment is not in equilibrium, and its intensive properties vary spatially and temporally. Many chemical reactions in the natural environment are blocked because the transport mechanisms necessary to reach equilibrium are too slow at ambient conditions. Thus, the exergy of the natural environment is not zero because work could be obtained if it were to come to equilibrium. In developing reference environment models for exergy analysis, a compromise is often made between the theoretical requirements of the reference environment and the actual behavior of the natural environment.

Several reference environment models have been proposed. The most significant classes of these are the following:

Natural environment subsystem models: These models attempt to realistically simulate subsystems of the natural environment. A model consisting of saturated moist air and liquid water in phase equilibrium was proposed in 1963. An extension of that model that allows sulfur-containing materials to be analyzed was proposed subsequently. The temperature and pressure of this reference environment (Table I) are normally taken to be 25°C and 1 atm, respectively, and the chemical composition is taken to consist of air saturated with water vapor, with the following condensed phases at 25°C and 1 atm: water (H_2O), gypsum ($CaSO_4 \cdot 2H_2O$), and

TABLE I

Reference Environment Model

Temperature, T_0	25°C	
Pressure, P_0	1 atm	
Composition	(i) Atmospheric air, saturated with H_2O at T_0 and P_0, having the following composition:	
	Air constituents	**Mole fraction**
	N_2	0.7567
	O_2	0.2035
	H_2O	0.0303
	Ar	0.0091
	CO_2	0.0003
	H_2	0.0001
	(ii) The following condensed phases at T_0 and P_0: Water (H_2O) Limestone ($CaCO_3$) Gypsum ($CaCO_4 \cdot 2H_2O$)	

limestone ($CaCO_3$). The stable configurations of C, O, and N, respectively, are taken to be those of CO_2, O_2, and N_2 as they exist in air saturated with liquid water at T_o and P_o (the temperature and pressure for the reference environment); that of hydrogen is taken to be in the liquid phase of water saturated with air at T_o and P_o; and those of S and Ca, respectively, are taken to be those of $CaSO_4 \cdot 2H_2O$ and $CaCO_3$ at T_o and P_o.

Reference substance models: With this model, a reference substance is selected for every chemical element and assigned zero exergy. A model in which the reference substances are selected as the most valueless substances found in abundance in the natural environment was proposed by Jan Szargut. The criterion for selecting such reference substances is consistent with the notion of simulating the natural environment but is primarily economic in nature, and it is vague and arbitrary with respect to the selection of reference substances. Part of this environment is the composition of moist air, including N_2, O_2, CO_2, H_2O, and the noble gases; gypsum (for sulfur); and limestone (for calcium). A related model, in which reference substances are selected arbitrarily, has also been proposed. This model is not similar to the natural environment. Consequently, absolute exergies evaluated with this model do not relate to the natural environment and cannot be used rationally to evaluate efficiencies or environmental impact.

Equilibrium and constrained equilibrium models: A model in which all the materials present in the atmosphere, oceans, and a layer of the crust of the earth are pooled together and an equilibrium composition is calculated for a given temperature was proposed by Joachim Ahrendts in 1980. The selection of the thickness of crust considered is subjective and is intended to include all materials accessible to technical processes. Thicknesses varying from 1 to 1000 m and a temperature of 25°C were considered. For all thicknesses, the model differed significantly from the natural environment. Exergy values obtained using these environments are significantly dependent on the thickness of crust considered and represent the absolute maximum amount of work obtainable from a material. Since there is no technical process that can obtain this work from materials, the equilibrium model does not give meaningful exergy values when applied to the analysis of real processes. Ahrendts also proposed a modified version of his equilibrium environment in which the calculation of an equilibrium composition excludes the possibility of the formation of nitric acid (HNO_3) and its compounds. That is, all chemical reactions in which these substances are formed are in constrained equilibrium, and all other reactions are in unconstrained equilibrium. When a thickness of crust of 1 m and temperature of 25°C are used, the model is similar to the natural environment.

Process-dependent models: A model that contains only components that participate in the process being examined in a stable equilibrium composition at the temperature and pressure of the natural environment was proposed by Bosnjakovic in 1963. This model is dependent on the process examined and is not general. Exergies evaluated for a specific process-dependent model are relevant only to the process. They cannot rationally be compared with exergies evaluated for other process-dependent models or used in environmental assessments.

4. RELATIONS BETWEEN EXERGY AND ENVIRONMENTAL IMPACT

Three relationships between exergy and environmental impact that may help explain some of the patterns that underlie changes in the environment are discussed in the following sections.

4.1 Order Destruction and Chaos Creation

The destruction of order, or the creation of chaos, is a form of environmental damage. Fundamentally, entropy is a measure of chaos. A system of high entropy is more chaotic or disordered than one of low entropy. For example, a field with papers scattered about has a higher entropy than the field with the papers neatly piled. Conversely, exergy is a measure of order. Relative to the same environment, the exergy of an ordered system is greater than that of a chaotic one.

The difference between the exergy values of the two systems containing papers described previously is a measure of the minimum work required to convert the chaotic system to the ordered one (i.e., in collecting the scattered papers). In reality, more than this minimum work, which only applies if a reversible cleanup process is employed, is required. The exergy destroyed when the wind scatters a stack of papers is a measure of the order destroyed during the process.

4.2 Resource Degradation

The degradation of resources found in nature is a form of environmental damage. A resource is a

material, found in nature or created artificially, that is in a state of disequilibrium with the environment. Resources have exergy as a consequence of this disequilibrium. For some resources (e.g., metal ores), it is their composition that is valued. Many processes exist to increase the value of such resources by purifying them (i.e., by increasing their exergy). This is done at the expense of consuming at least an equivalent amount of exergy elsewhere (e.g., burning coal to produce process heat for metal ore refining). For other resources (e.g., fuels), it is normally their reactivity that is valued (i.e., their potential to cause change or "drive" a task or process).

By preserving exergy through increased efficiency (i.e., degrading as little exergy as necessary for a process), environmental damage is reduced. Increased efficiency also has the effect of reducing exergy emissions, which, as discussed in the next section, also play a role in environmental damage.

The earth is an open system subject to a net influx of exergy from the sun. It is the exergy (or order states) delivered with solar radiation that is valued; all the energy received from the sun is ultimately radiated out to the universe. Environmental damage can be reduced by taking advantage of the openness of the earth and utilizing solar radiation instead of degrading resources found in nature to supply exergy demands. This would not be possible if the earth were a closed system because it would eventually become increasingly more degraded or entropic.

4.3 Waste Exergy Emissions

The exergy associated with process wastes emitted to the environment can be viewed as a potential for environmental damage. Typical process wastes have exergy, a potential to cause change, as a consequence of not being in stable equilibrium with the environment. When emitted to the environment, this exergy represents a potential to change the environment. In some cases, this exergy may cause a change perceived to be beneficial (e.g., the increased rate of growth of fish and plants near the cooling water outlets from thermal power plants). More often, however, emitted exergy causes a change that is damaging to the environment (e.g., the death of fish and plants in some lakes due to the release of specific substances in stack gases as they react and come to equilibrium with the environment).

Emissions of exergy to the environment can also interfere with the net input of exergy via solar radiation to the earth. The carbon dioxide emitted in stack gases from many processes changes the atmospheric CO_2 content, affecting the receiving and reradiating of solar radiation by the earth.

The relation between waste exergy emissions and environmental damage has been recognized by several researchers. By considering the economic value of exergy in fuels, in 1970 Gordon Reistad developed an air pollution rating in which the air pollution cost for a fuel was estimated as either the cost to remove the pollutant or the cost to society of the pollution (i.e., the tax that should be levied if pollutants are not removed from effluent streams). Reistad claimed the rating was preferable to the mainly empirical ratings then in use.

4.4 An Apparent Paradox

The reader may at this point detect a paradox. It was pointed out that exergy in the environment—in the form of resources—is of value, whereas exergy in the environment—in the form of emissions—has a "negative" value because it has a potential to cause environmental damage.

This apparent paradox can be resolved if the word constrained is included in the discussions (Fig. 1). Constrained sources of exergy in the environment are of value. Most resources found in nature are constrained. Unconstrained emissions of exergy to the environment have negative value. In fact, if emissions to the environment are constrained (e.g., separating out sulfur from stack gases), then there are two potential benefits: (i) The potential for environmental damage is blocked from entering the environment, and (ii) the now-constrained emission potentially becomes a valued commodity, a source of exergy.

4.5 Qualitative Illustration

The decrease in the environmental impact of a process, in terms of the three measures discussed in this section, as the exergy efficiency increases for the process is

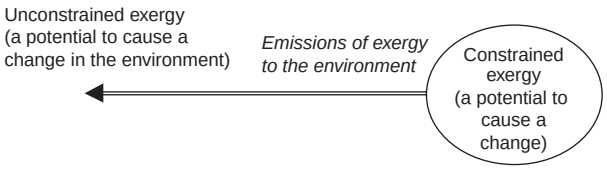


FIGURE 1 Comparison of constrained and unconstrained exergy illustrating that exergy constrained in a system represents a resource, whereas exergy emitted to the environment becomes unconstrained and represents a driving potential for environmental damage.

illustrated qualitatively and approximately in Fig. 2. Exergy methods also correlate with sustainability if, for simplicity, we consider sustainability to indicate the degree to which human activity can continue without facing a lack of resources or negatively changing the environment. To better illustrate the relation between exergy and sustainability and environmental impact, Fig. 2 can be expanded as shown in Fig. 3. In this figure, sustainability is seen to increase and environmental impact to decrease as the exergy efficiency of a process increases. The two limiting efficiency cases in Fig. 3 are significant:

- As exergy efficiency approaches 100%, environmental impact approaches zero since exergy is converted from one form to another without loss (either through internal consumptions or through waste emissions). Also, sustainability becomes increasingly great because the process approaches reversibility (i.e., no losses occur so the process can go forward and backward indefinitely).
- As exergy efficiency approaches 0%, sustainability approaches zero because exergy-containing resources are used but nothing is accomplished. Also, environmental impact becomes increasingly great because to provide a fixed service, an ever-increasing quantity of resources must be used and a correspondingly increasing amount of exergy-containing wastes are emitted.

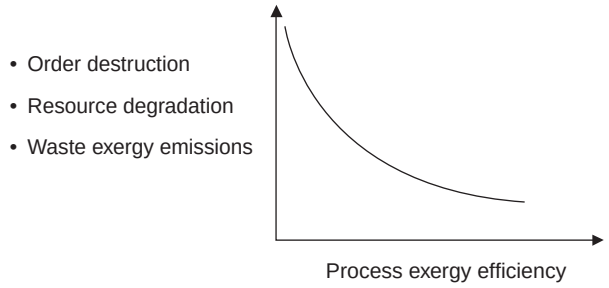


FIGURE 2 Qualitative illustration of the relation between the exergy efficiency of a process and the associated environmental impact in terms of order destruction, resource degradation, or waste exergy emissions.

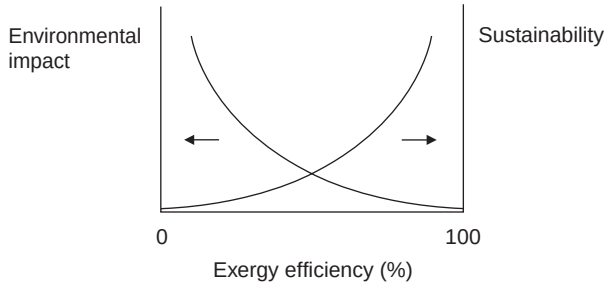


FIGURE 3 Qualitative illustration of the relation between the environmental impact and sustainability of a process and also its exergy efficiency.

4.6 Illustrative Example

The ideas discussed in this section are demonstrated for a unit of a typical 500-MWe coal-fired electrical generating station. Overall balances of exergy and energy for the station are illustrated in Fig. 4. This station was examined in the article titled "Exergy Analysis of Energy Systems." For this station, each of the three relationships between exergy and environmental impact described in the previous subsections are demonstrated:

- Waste exergy emissions from the plant occur with (i) the gaseous wastes exiting through the stack, (ii) the solid wastes exiting the combustor, and (iii) waste heat released both to the lake, which supplies condenser cooling water, and to the atmosphere. Contained within each of these emissions is a potential to impact on the environment. The societal

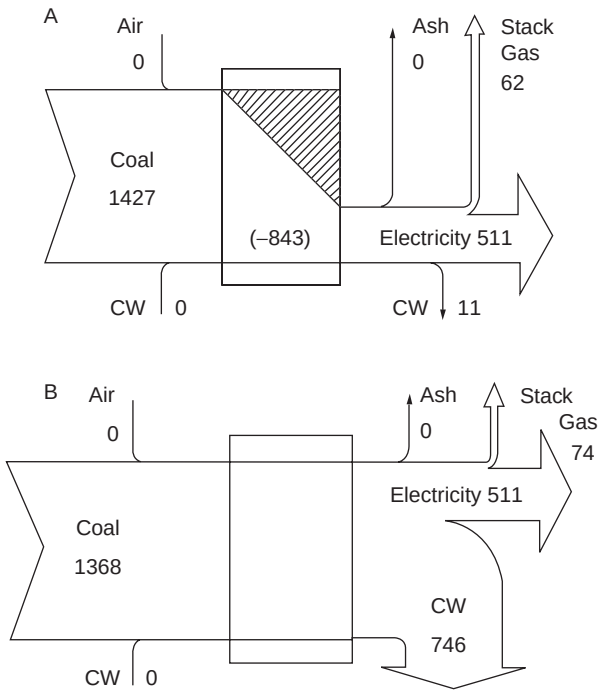


FIGURE 4 Overall energy and exergy balances for a unit of a typical 500-MWe coal-fired electrical generating station. The rectangle in the center of each diagram represents the station. Widths of flow lines are proportional to the relative magnitudes of the represented quantities. CW, cooling water. (A) Exergy balance showing flow rates (positive values) and consumption rate (negative value, denoted by hatched region) of exergy (in MW). (B) Energy balance showing flow rates of energy (in MW).

consensus regarding emissions of harmful chemical constituents in stack gases and "thermal pollution" in local bodies of water indicates that the potential for impact of these emissions is already recognized but not from an exergy perspective.

- During the electricity generation process, a finite resource (i.e., coal) is degraded as it is used to drive the process. Although a degree of degradation is necessary for any real process, increased process efficiency can reduce this degradation for the same services or products. In the extreme, if the process considered in the example were made ideal, the exergy efficiency would increase from 37 to 100%, and coal use as well as related emissions would decrease by more than 60%. These insights are provided by the exergy, but not energy, considerations.
- The more abstract concepts of order destruction and chaos creation are also illustrated. Certainly the degradation of coal to stack gases and solid wastes represents a destruction of order, especially since less exergy is present in the products of the process. The emissions of wastes to the environment increase the chaos of the world by allowing the constrained material and heat products of combustion to be released without constraint into the environment.

5. ENVIRONMENTAL IMPACT OF EXERGY LOSSES

As mentioned earlier, exergy losses, which are made up of exergy destruction and waste exergy emissions, have a significant effect on environmental impact. Exergy destruction, in particular, can be used as one of the most significant criteria for assessing the depletion of natural resources. Exergy analysis can assist efforts to minimize the use of natural resources because it can indicate where the work potential of natural resources in relation to the surrounding environment is lost (i.e., where irreversibility, or exergy destruction, occurs).

The exergy in a flow can only be entirely converted to products in a fully reversible process since there are no exergy destructions within the system and no waste exergy emissions to the environment. In real-world processes, which are irreversible, both destruction of the exergy of resources and waste exergy emissions occur. Much effort is being expended on reducing resource exergy destructions and eliminating waste exergy emissions—often by converting them into useful products. The reversible process noted earlier represents a theoretical ideal that we can strive toward but never actually achieve.

A waste emission possesses exergy as a result of its being in a state of mechanical, thermal, and/or chemical disequilibrium with the reference environment. It is useful here to follow the standard separation of exergy into two components—physical and chemical. In this article of exergy and the environment, the exergy of an emission attributable to mechanical and thermal disequilibrium is not considered significant because the potential impact on the environment of physical exergy is limited. That is, pressure differences between an emission and the environment normally dissipate soon after the emission reaches the environment, and temperature differences are normally localized near the emission source (e.g., thermal pollution in the region of a lake near the cooling water discharge of a thermal power plant) and can be controlled. The exergy of an emission due to chemical disequilibrium (i.e., the chemical exergy), on the other hand, is often significant and not localized. The chemical exergy of emissions is normally of more concern.

6. COMPARISON OF EXERGY AND OTHER INDICATORS FOR THE ENVIRONMENTAL IMPACT OF WASTE EMISSIONS

The exergy of waste emissions has been compared to some other measures to assess or control the potential environmental impact of emissions, including air emission limits established by the government of Ontario, and two proposed quantifications of "environmental costs" for emissions from fossil fuel combustion. These comparisons are carried out to detect trends and patterns that may permit the exergy of a substance to be used as a useful predictor of potential environmental impact and consequently as a tool in establishing emission limits that are rationally based rather than formulated by trial and error.

6.1 Background on Pollution Limits and Costs

Air pollution limits in Ontario, Canada, are covered by that province's principal legislation governing environmental quality, the Environmental Protection Act. The overall goal of the legislation is to ensure environmental conditions such that human health

and the ecosystem of the earth are not endangered. In Ontario, the Ministry of the Environment is the governing body that oversees the development and implementation of environmental legislation for industry. For the control of air pollution, allowable air emission limits (i.e., allowable mass of pollutant per volume of air averaged over a specified time period) that must be achieved prior to discharge (e.g., prestack) are provided for numerous substances. The potential of a substance to impact on the environment is evaluated by 10 parameters: transport, persistence, bioaccumulation, acute lethality, sublethal effects on mammals, sublethal effects on plants, sublethal effects on nonmammalian animals, teratogenicity, mutagenicity/genotoxictiy, and carcinogenicity. With this information, point of impingement (POI) air emission limits are determined through a method known as the best available pollution control technology.

6.2 Environmental Costs of Emissions to Air

Two methods for developing environmental costs for emissions to air are considered here:

- For air emissions resulting from the combustion of fossil fuels, consider the cost of removing the pollutants from the waste stream prior to discharge into the environment. This cost can be related to the exergy of the pollution and is referred to as the removal pollution cost (RPC). The removal costs for waste emissions are evaluated as the total fuel cost per unit fuel exergy multiplied by the chemical exergy per unit fuel exergy and divided by the exergy efficiency of the pollution removal process. The exergy efficiencies for the removal of pollutants from waste streams vary. Some sources indicate that exergy efficiencies are less than 5% when the removal process involves mechanical separation. For simplicity, exergy efficiencies of 1% for all pollutants are used here.
- Estimate environmental costs of pollutant, which are referred to here as environmental pollution costs (EPCs). Such work is most advanced for emissions released into the atmosphere resulting from the combustion of fossil fuels (e.g., carbon dioxide and sulfur dioxide). Environmental costs for some of these emissions have been estimated for Canada (Table II). Values for EPCs are based on quantitative and qualitative evaluations of the cost to society for the correction or compensation of environmental damage and/or the cost incurred to prevent a harmful emission from escaping into the environment.

TABLE II

Environmental Pollution Costs (EPCs) for Selected Pollutants

Pollutant	EPC (1998 Can. $/kg pollutant)[a]
Particulates[b]	4.11
CO	3.71
NO_x	2.91
SO_2	2.65
CH_4	0.97
Volatile organic compounds	0.45
CO_2	0.030

[a] The EPCs are based on values from 1990, with an adjustment applied to the dollar values to account for inflation in Canada between 1990 and 1998 (as measured by the Consumer Price Index for all products). Statistics Canada reports the adjustment factor as 1.164.

[b] Including heavy metals (e.g., lead, cadmium, nickel, chromium, copper, manganese, and vanadium).

6.3 Relations among the Measures

POI air emission limits, standard chemical exergies, RPCs, and EPCs for emissions have been studied and some preliminary relations discerned. For instance, evidence suggests that EPC increases with increasing standard chemical exergy. Also, EPCs generally appear to increase at a decreasing rate with increasing percentages of pollution emission exergy. These general trends are expected and logical.

Overall, the two measures considered here for the environmental cost of pollutants (RPCs, which are based on the cost to remove a pollutant from the environment, and EPCs, which are based on the costs attributable to the negative impact of the pollutant on the environment), although based on different principles and evaluated differently, are of the same order of magnitude for a given pollutant. The RPC methodology is based on a simple, theoretical concept, whereas the EPC methodology relies on subjective interpretations of environmental impact data. Thus, exergy-based measures for environmental impact may provide a foundation for rational environmental indicators and tools.

7. CONCLUSIONS

The relations between environmental impact and exergy, and between the chemical exergy values of

substances found in waste emissions and measures of environmental impact, could contribute to the development of simple, rational, and objective procedures for assessing the harmful effects on the environment and predicting the potential for environmental impact for a substance. This view is primarily based on the general premise that a substance has exergy due to its disequilibrium with respect to a reference environment, and therefore the exergy of unconstrained waste emissions has the potential to impact on the environment. The potential usefulness of exergy analysis in addressing and solving environmental problems is significant, but further work is needed if utilizable, objective exergy-based measures of the potential of a substance to impact on the environment are to be developed.

SEE ALSO THE FOLLOWING ARTICLES

Entropy • Exergy • Exergy Analysis of Energy Systems • Exergy: Reference States and Balance Conditions • Thermodynamic Sciences, History of • Thermodynamics, Laws of • Waste-to-Energy Technology

Further Reading

Ayres, R. U., Ayres, L. W., and Martinas, K. (1998). Exergy, waste, accounting, and life-cycle analysis. *Energy Int. J.* **23,** 355–363.

Berthiaume, R., Bouchard, C., and Rosen, M. A. (2001). Exergetic evaluation of the renewability of a biofuel. *Exergy Int. J.* **1,** 256–268.

Connelly, L., and Koshland, C. P. (2001). Exergy and industrial ecology. *Exergy Int. J.* 1, 146–165, 234–255.

Creyts, J. C., and Carey, V. P. (1997). Use of extended exergy analysis as a tool for assessment of the environmental impact of industrial processes. *In* "Proceedings of the ASME Advanced Energy Systems Division" (M. L. Ramalingam, J. L. Lage, V. C. Mei, and J. N. Chapman, Eds.), AES Vol. 37, 129–137. American Society of Mechanical Engineers, New York.

Edgerton, R. H. (1982). "Available Energy and Environmental Economics." Heath, Toronto.

Frangopoulos, C., von Spakovsky, M. R. (1993). A global environomic approach for energy systems analysis and optimization (parts 1 and 2). *In* "Proceedings of the International Conference on Energy Systems and Ecology, Cracow, Poland," pp. 123–144.

Gunnewiek, L. H., and Rosen, M. A. (1998). Relation between the exergy of waste emissions and measures of environmental impact. *Int. J. Environ. Pollut.* **10**(2), 261–272.

Kotas, T. J. (1995). "The Exergy Method of Thermal Plant Analysis." Reprint ed. Krieger, Malabar, FL.

Moran, M. J. (1989). "Availability Analysis: A Guide to Efficient Energy Use." Rev. ed. American Society of Mechanical Engineers, New York.

Reistad, G. M. (1970). Availability: Concepts and applications, Ph.D. dissertation, University of Wisconsin, Madison.

Rosen, M. A. (1999). Second law analysis: Approaches and implications. *Int. J. Energy Res.* **23,** 415–429.

Rosen, M. A., and Dincer, I. (1997). On exergy and environmental impact. *Int. J. Energy Res.* **21,** 643–654.

Rosen, M. A., and Dincer, I. (1999). Exergy analysis of waste emissions. *Int. J. Energy Res.* **23,** 1153–1163.

Rosen, M. A., and Dincer, I. (2001). Exergy as the confluence of energy, environment and sustainable development. *Exergy Int. J.* **1,** 3–13.

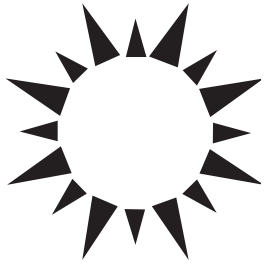
Sciubba, E. (1999). Exergy as a direct measure of environmental impact. *In* "Proceedings of the ASME Advanced Energy Systems Division" (S. M. Aceves, S. Garimella, and R. Peterson, Eds.), AES Vol. 39, pp. 573–581. American Society of Mechanical Engineers, New York.

Szargut, J., Morris, D. R., and Steward, F. R. (1988). "Exergy Analysis of Thermal, Chemical and Metallurgical Processes." Hemisphere, New York.

Tribus, M., and McIrvine, E. C. (1971). Energy and information. *Sci. Am.* **225**(3), 179–188.

Wall, G., and Gong, M. (2001). On exergy and sustainable development. *Exergy Int. J.* **1,** 128–145, 217–233.

Exergy: Reference States and Balance Conditions



ROBERT U. AYRES
Center for the Management of Environmental Resources
Fontainebleau, France
Chalmers Institute of Technology
Gothenburg, Sweden

ANDREA MASINI
London Business School
London, United Kingdom

1. **Standard Formalism for Chemical Exergy Calculations**
2. **Sink-Specific Reference Species**
3. **Exergy Balance as a Tool for Analysis**
4. **Exergy as a Measure of Resource Availability or Losses**

Glossary

exergy The maximum amount of useful work that can be extracted from a material as it approaches thermodynamic equilibrium with its surroundings by reversible processes. The second law of thermodynamics implies that it is never possible to utilize all available or potentially useful work for practical purposes. There is some waste or loss of exergy in every process. Lost exergy corresponds to increasing entropy.

Gibbs energy of formation The total amount of energy that is either used up or released during a chemical reaction. Gibbs free energy $(\Delta G) = (\Delta H) - t\ (\Delta S)$, where (ΔH) is the change in enthalpy; t is the temperature at which the reaction took place; and (ΔS) is the change in entropy, or amount of disorder, that occurs in the molecules involved during the reaction.

reference species One or more of the common chemical forms that characterize the final equilibrium state of a given "mix" of chemical elements, taking into account relative atomic abundance. The equilibrium state is what would finally result from free, spontaneous, and uninhibited interactions with all the other elements in the mix.

The standard chemical exergy, per mole, of any pure compound involving chemical elements can be computed by means of a simple formula. To calculate the chemical exergies of mixtures and composite materials, such as ores and alloys, it is only necessary to have detailed composition data and do the sums. To compile generic exergy tables, however, it is necessary to adopt a general convention on reference states.

1. STANDARD FORMALISM FOR CHEMICAL EXERGY CALCULATIONS

The standard chemical exergy, per mole, of any pure compound involving these elements can be computed by means of a simple formula, namely

$$B = G + \sum_j n_j b_j, \qquad (1)$$

where G is the standard Gibbs free energy of formation of the compound, the n_j are molar fractions of the jth chemical element, and the index j runs over the elements in the compound. The G values are tabulated in standard reference works, such as the *Handbook of Physics and Chemistry*. The b_j values have been calculated for all the important elements. To calculate the chemical exergies of mixtures and composite materials, such as ores and alloys, it is only necessary to have detailed composition data and do the sums.

To compile generic exergy tables, however, it is necessary to adopt a general convention on reference

 © 2004 Elsevier Inc. All rights reserved.

states. This could be a daunting task if a very local definition were adopted for each case. However, noting that there are three major environmental media—or "sinks"—a reasonably general definition has been proposed. In the case of chemical elements found in the atmosphere as such (oxygen, nitrogen, inert gases) and carbon—because carbon reacts with atmospheric oxygen to yield a gaseous oxide (CO_2)—the composition of the atmosphere is taken to be the reference state. In the case of chemical elements whose oxides or chlorides are soluble in water, the composition of the ocean is taken to be the reference state. For all other elements, mainly metals whose oxides are insoluble solids, the average composition of the earth's crust down to some specified depth is taken to be the reference state. (The technical problem here is that the earth's crust is not in chemical equilibrium with either the atmosphere or the oceans. In particular, the earth's crust contains significant quantities of hydrocarbons and sulfides, of biological origin, that are potentially reactive with oxygen but that are not exposed to the atmosphere. Obviously in the very long run, in the absence of biological activity, these materials would be oxidized and all of the free oxygen in the atmosphere would ultimately be combined with nitrogen, carbon, or sulfur. Thus, the essence of the convention is to assume that the three media [air, oceans, and crust] are in chemical equilibrium in the short run—that is, in terms of the time scales of interest.) The latter definition is somewhat arbitrary, to be sure, but it turns out that calculated values in most cases are not extremely sensitive to the crustal depth assumption.

A procedure has been developed that results in a computed "standard" exergy for each material that is independent of the actual sink. This is done by assuming each element is obtained by successive reactions from a relatively stable chemical form that is chosen once and for all for each element and identified as its reference species. The reference species should be one of the more common chemical forms that would characterize the final equilibrium state of a given "mix" of chemical elements, taking into account relative atomic abundance. (The equilibrium state is what would finally result from free, spontaneous, and uninhibited interactions with all the other elements in the mix. This idealized situation is sometimes called a Gibbs reactor.) The reference state for most elements is taken to be either its most oxidized or chlorinated form, depending on volatility and solubility. To carry out this procedure the final compounds must be identified, along with their thermodynamic parameters, and specified at standard temperature (298.15°K) and atmospheric pressure (101.325 kPa).

It is normal practice in the thermodynamics of closed systems to assign a zero value for enthalpy of formation of pure elements. However, this convention makes calculation awkward for open systems. Instead, Szargut has proposed that the zero level should be assigned to the (unique) reference species. (Szargut *et al.* have identified 10 gaseous species for 10 chemical elements, 31 ions dissolved in seawater for 31 chemical elements and other reference species for the same number of elements.) Then the standard exergy for any chemical will be defined as the maximum work that can be extracted when that chemical is transformed by successive reactions into the reference species in the standard state of temperature and pressure.

There is also an adjustment for the reference species themselves, since their natural states are not the same as the standard state. For instance, the natural state is often one with a much lower vapor pressure due to low concentration. Thus, the standard exergy of the reference species in the natural state is the work that can be extracted in going from the standard state to the natural state. Thus, the standard chemical exergy of oxygen is bigger than the chemical exergy of atmospheric oxygen, because the latter is at lower (partial) pressure. Calculation of the chemical exergy for reference species depends on whether the species is found in the atmosphere, in seawater or in the earth's solid crust.

Implementation of the preceding scheme for a compound requires the following steps:

1. Calculation of the standard exergy of each of the elements $b^0_{ch,el}$ from the corresponding equation representing its formation (by a reversible reaction) from the reference species. (These equations are equations of formation of the reference species from elements, expressed in reverse.) Since there is just one reference species per element, the system of equations has a unique set of solutions. As noted already, reference species should be common components of the environment. Szargut has selected 10 reference species found in the atmosphere, 31 in the ocean, and 13 species in the earth's crust. The equation is as follows:

$$b^0_{ch} = \Delta G^0_f + \sum_{el} n_{el} b^0_{chel}, \tag{2}$$

where G_f is the Gibbs energy of formation of the target compound, b^0_{ch} is the chemical exergy of the target compound per mole, b^0_{chel} is the chemical exergy

of the component element (previous tabulated), and n_{el} is the number of atoms of the element in the compound.

2. To correct for differences between standard conditions and actual conditions and abundance in the three main media, we use one of the following three equations. For atmospheric species,

$$b^0_{ch} = -RT_n ln \frac{p_n}{p_{0n}}, \tag{3}$$

where p_n is the normal pressure (101.325 kPa) and p_m is the partial pressure of the compound in the atmosphere.

For reference species in the earth's crust,

$$b^0_{ch} = RT_n ln x_{\text{in}}, \tag{4}$$

where x_{in} is the average concentration (mole fraction) of the ith substance in the earth's crust.

In the case of reference species dissolved in seawater, the approach is slightly different. The following formula allows the direct calculation of the standard chemical exergy of the pure element contained in the reference species:

$$\begin{aligned} b^0_{ch} = & -j\Delta_f G^0 + \tfrac{1}{2} jz b^0_{ch\ H_2} \\ & - j\sum n_k b^0_{ch\ k} - 2.303 RT_n jz(pH) \\ & - jRT_n \ln m_n \gamma, \end{aligned} \tag{5}$$

where $\Delta_f G^0$ is the standard normal free energy of formation of the reference ion, j is the number of reference ions derived from one molecule of the element under consideration, z is the number of elementary positive charges in the reference ion, n_k is the number of moles of additional elements in the reference ion, m_n is the conventional molarity of the reference species in seawater, and γ is the activity coefficient of the reference species in seawater.

2. SINK-SPECIFIC REFERENCE SPECIES

Szargut's simplified approach is convenient for many purposes. However, it may be inadequate for purposes of assessing potential environmental harm from pollutants, where localized sink-specific reactions are typically important. For instance, the standard chemical exergy of sulfur in Szargut's scheme is based on the assumption that the end product is the SO_4^- ion (in water), independent of whether the sulfur was originally emitted into air or water. It is important to recall that the atmosphere is not in equilibrium with the ocean (nor is the ocean in chemical equilibrium with the earth's crust). The potential harm due to atmospheric emissions must be related to the exergy of the pollutant with respect to atmospheric reference species only. Clearly, the real situation is somewhat more complex than Szargut's simplified scheme, at least for some of the major biologically active elements and compounds found in pollutants.

The starting point for a *gedanken* equilibration process is a homogeneous mixture of pure chemical elements in a reactor. The chemical composition of the final "soup" of compounds depends on the elementary composition of the original mixture. This is the primary difference between the atmosphere, hydrosphere, and lithosphere. Each chemical element in the starting mix must be used up in the final soup of compounds. For instance, hydrogen tends to combine with oxygen forming water (H_2O) whenever possible. Thus, most of the hydrogen in the earth's crust—including the oceans and atmosphere—is in the form of water (including its ionized forms H^+ and OH^-). Similarly, most of the carbon is also combined with oxygen. This is in the form of gaseous CO_2 in the air, carbonic acid H_2CO_3 (in water), or one of its metallic salts such as calcium or magnesium carbonate in the earth's crust. (The only hydrogen and carbon that are normally found on the surface of the earth [i.e. in the atmosphere, oceans, sedimentary, or metamorphic rocks] that are not on one of these forms are in carbohydrates, fats, or hydrocarbons resulting [directly or indirectly] from photosynthesis. Fossil hydrocarbons, including coal, are derived from previously living matter. Graphite is mostly of metamorphic origin, from coal or other sedimentary hydrocarbon deposits. Hydrogen and carbon are very rare in igneous rock, although some hydrogen sulfide and ammonia [presumably dissolved under pressure in magma] are emitted by volcanoes.)

To get a feel for the situation, suppose the elements were mixed together initially in a gaseous soup but at a low enough temperature to permit stable molecules to survive. Assume that various reactions occurred in sequence, rather than simultaneously. The first step in the sequence might be for the two most mutually attractive elements to combine preferentially with each other. This reaction would continue until the supply of one of the reactants is exhausted. Then, the two most reactive of the remaining elements would combine until one of them is used up, and so on. The procedure continues, in principle, until every possible

(exothermic) reaction has occurred. The final result is a chemical equilibrium in which at least one and possibly several of the most reactive elements are found with only one partner, although that partner may combine preferentially with several others. (Thus, both hydrogen and carbon are found almost exclusively with oxygen, except when they combine with each other. Oxygen—being more abundant—is also the preferred partner of many other elements.) In a true equilibrium it makes sense (in most cases) to choose the most abundant compound of each element as the reference species for that element.

However, this simplified mental model of a sequential reactor fails to allow for a number of real-world factors. In the first place, one of the major combinations resulting from the vapor-phase reactions is water, which condenses at a relatively high temperature, whereas quite a few other light molecules, including hydrogen, methane, ammonia, hydrogen sulfide, and carbon dioxide, as well as molecular nitrogen, oxygen, and the rare gases, remain in vapor phase at the same pressures and temperatures. Thus, all known planets have volatile atmospheres as well as solid cores. The presence of liquid water on earth is a result of the specific combination of temperatures and pressures that make it possible. It also depends on the peculiar fact that frozen water is slightly less dense than liquid water and therefore floats rather than sinking. Absent this unique characteristic of water, polar ice would sink and accumulate on the Arctic and Antarctic ocean floors, gradually cutting off ocean currents. This would eliminate the major mechanism for cooling the tropics and warming the polar regions. The result would be hotter tropics, colder polar regions, and much more violent storms resulting from the temperature gradient. The storms would carry water vapor from the tropics to the poles, where it would condense as ice. It would also make the remaining tropical oceans more and more saline (like the Dead Sea). The end result would be a stable world of glaciers and salt.

Water is an extraordinarily powerful ionic solvent for some alkali metals and halogens, as well as sulfates and nitrates. But many other molecular combinations, especially of metals, do not dissolve to a significant degree in water. Nor do some of the major families of hydrocarbons and lipids. In short, the liquid water on the earth's surface constitutes a kind of filter, which efficiently separates soluble from insoluble species. The latter, of course, build up into sedimentary layers.

The composition of earth's atmosphere is unlike that of the dead planets. Both molecular oxygen and molecular nitrogen exist in the earth's atmosphere only because of biological processes. The role of photosynthesis in maintaining the carbon cycle is very well known. The corresponding role of denitrification organisms is less well understood. But without them, the atmospheric nitrogen would gradually combine with oxygen (thanks to electrical storms) and would be dissolved in the ocean as nitric acid.

Evidently the three major sinks (the atmosphere, the oceans, and the sedimentary layers of the earth's crust) are self-organized systems in a stable state that is far from thermodynamic equilibrium. Such systems exist thanks to complex biological and geo logical processes driven by a continuous flux of solar exergy, via photosynthesis and denitrifying bacteria.

However, there are problems with the simple theory. An example will help to clarify some points. When another element or compound (say sulfur) is injected into the atmosphere, it reacts with oxygen until it is entirely consumed. The usual product is SO_2. However, the SO_2 either reacts with water vapor to form sulfurous acid H_2SO_3 or it oxidizes to SO_3 and subsequently reacts with water to form sulfuric acid H_2SO_4. These acids either wash out in rain or react with ammonia to form ammonium sulfite or sulfate, which also washes out in rain. The acid rain causes corrosion on metal surfaces and changes the pH of soil and freshwater lakes. Eventually most of the sulfur does indeed end up as SO_4^- ion in the seawater. (But seawater is not a permanent residence for sulfur: the sulfate ion gradually combines with calcium or magnesium ions and precipitates into sediments.) Yet much of the potential environmental harm occurs in intermediate environments where the sulfur *enters* as sulfate ion and the most likely final (reference) state is probably calcium sulfate.

Curiously enough, the most abundant sulfur compound in the atmosphere is not sulfur dioxide, as one might expect; it is an organic compound carbonyl sulfide (*COS*), probably of biological origin. Yet there is a possibility that this compound is produced in the atmosphere by a reaction with methane such as $SO_3 + CH_4 \rightarrow COS + 2H_2O$. Unfortunately one cannot easily perform the relevant exergy calculations, due to unavailability of data on Gibbs energy of formation of *COS*.

In the case of sulfur, one might propose not just one but three reference states, depending on the medium of interest. In the atmosphere itself the

reference state would be either SO_2 or possibly COS; in seawater it would be SO_4^{2-} ion, and in the (sedimentary) lithosphere it would be calcium sulfate, probably in the form of the mineral gypsum $CaSO_4 \cdot 2H_2O$. In the case of sulfur (and other elements) another complication arises, namely that the lithosphere is anything but homogeneous. In particular, there are vast differences between the sedimentary or metamorphic component of the earth's crust (the part that has resulted from weathering and has been exposed to the atmosphere at some stage) and the unweathered igneous rock left over from the earth's original solidification. Most sulfur in nonsedimentary igneous rocks, however, is combined with metals as sulfides (MS, MS_2, etc.). As these are exposed to water or air (e.g., by mining operations), the sulfides gradually oxidize to sulfurous acid or sulfur dioxide.

2.1 Standard Theory

The standard theory calculates standard chemical exergy of sulfur by adopting SO_4^- dissolved in water as the reference species. Calculations are based on Szargut's analysis and give the following result:

$$b^0_{ch}[\mathrm{S}] = 609.6\ \mathrm{kJ/mole}. \tag{6}$$

2.2 Sink-Specific Approach

Three independent measures may be computed for the atmosphere, the seawater, and the lithosphere as follows:

Atmosphere. Apart from COS (mentioned earlier), the most abundant compound containing sulfur in the atmosphere is sulfur dioxide, SO_2. Its mean volume concentration in the atmosphere has been estimated as about $2 \cdot 10^{-8}$. From Eq. (2) we obtain for the reference species

$$b^0_{ch}[\mathrm{SO_2}] = 40.25\ \mathrm{kJ/mole}. \tag{7}$$

The formation of sulfur from SO_2 (or vice versa) can be written as follows

$$\mathrm{S} + \mathrm{O_2} = \mathrm{SO_2}. \tag{8}$$

Exergy does not have to be recalculated for oxygen since the original reference species selected by Szargut was atmospheric O_2. Thus, applying an exergy balance one obtains

$$b^0_{ch}[\mathrm{S}] + b^0_{ch}[\mathrm{O_2}] + \Delta G^0_r = b^0_{ch}[\mathrm{SO_2}], \tag{9}$$

and

$$\begin{aligned} b^0_{ch}[\mathrm{S}] &= b^0_{ch}[\mathrm{SO_2}] - b^0_{ch}[\mathrm{O_2}] - \Delta G^0_r \\ &= 107.96\ \mathrm{kJ/mole}. \end{aligned} \tag{10}$$

Mean residence time of sulfur dioxide in the atmosphere is fairly low (about 10^{-2} years), thus choosing this substance as a reference species is quite inappropriate. It is reasonable to assume that sulfur-containing compounds easily evolve toward other chemical forms, which produce SO_4^- dissolved in seawater.

Seawater. The chosen reference species is the ion SO_4^- As noted previously, Szargut calculated $b^0_{ch}[\mathrm{S}] = 609.6$ kj/mole However, Szargut assumed that the standard chemical exergy for oxygen in seawater was equal to that of atmospheric oxygen. Taking into account the lower concentration of oxygen in seawater, standard chemical exergy of O_2 increases to 21.11 kJ/mole. In principle, the exergy of sulfur should be recalculated. However, for some compounds it seems reasonable to assume (as Szargut does) that reactions actually occur in the vapor phase, even when considering ions dissolved in seawater or solid compounds.

Lithosphere. Szargut's chosen reference species is gypsum ($CaSO_4 \cdot 2H_2O$). Unfortunately, no data on the mass fraction of this mineral are available for calculating its standard chemical exergy as a reference species. However, if an estimated average concentration of 10^{-4} is assumed, we obtain (from Eq. 3)

$$b^0_{ch}[\mathrm{CaSO_4 \cdot 2H_2O}] = 22.83\ \mathrm{kJ/mole}. \tag{11}$$

The following reaction can be identified for the formation of sulfur from gypsum:

$$\begin{aligned} &\mathrm{S} + \tfrac{3}{2}\mathrm{O_2} + \mathrm{CaCO_3} + 2\mathrm{H_2O} \\ &= \mathrm{CaSO_4 \cdot 2H_2O} + \mathrm{CO_2}. \end{aligned} \tag{12}$$

$CaCO_3$ is another reference species for the lithosphere and its chemical exergy is 1.0 kJ/mole. Assuming that oxygen and carbon dioxide participate in the reaction in gaseous form and that a liquid reference state may be assumed for water, one obtains

$$\begin{aligned} b^0_{ch}[\mathrm{S}] &= b^0_{ch}[\mathrm{CaSO_4 \cdot H_2O}] \\ &\quad + b^0_{ch}[\mathrm{CO_2}] - \tfrac{3}{2} b^0_{ch}[\mathrm{O_2}] \\ &\quad - b^0_{ch}[\mathrm{CaCO_3}] - \Delta G^0_r \\ &= 171.27\ \mathrm{kJ/mole}. \end{aligned} \tag{13}$$

Thus, Szargut calculates the standard exergy for sulfur as 609.6 kJ/mole (assuming sulfate SO_4^{2-} in sea

water as the reference state), but a sink-specific approach yields very different figures, namely 107.96 kJ/mole to atmospheric SO_2 (not a good choice, however, because of its short life) and 171.27 kJ/mole to gypsum in the lithosphere.

The preceding example can be extended to many other elements. For instance, most of the silicon in the sedimentary part of the earth's crust is also combined with oxygen as silica (SiO_2). This is the logical reference species for silicon. In nonsedimentary rock, it is usually in one of the reduced metallic silicate forms $MSiO_3$ or M_2SiO_3. These silicates spontaneously react with carbonic acid H_2CO_3 to form silica SiO_2 and corresponding hydroxide MOH or $M(OH)_2$. The latter reacts further with any acid in the environment to form a metallic carbonate, sulfate, or chloride. Carbonates obviously predominate. Aluminum is always combined with oxygen (as alumina, Al_2O_3), even in igneous rock, so this is the logical reference species. Alumina and silica are sometimes combined in a single mineral (kyanite, Al_2SiO_5).

Sodium, potassium, and calcium in the lithosphere are normally oxidized as Na_2O, K_2O, and CaO. These are often combines in more complex minerals, such as feldspars. These are major rock-forming minerals that constitute up to 60% of many igneous rocks metal. They include potassium feldspars ($KAlSi_3O_8$), sodium feldspars (albite, $NaAlSi_3O_8$), calcium feldspars (anorthite, $CaAl_2Si_2O_8$), and all combinations. The oxides are convenient reference species for all three metals.

Phosphorus in the lithosphere is only found in one mineral, apatite [$Ca_3(PO_4)_2$], of biological origin. Apatite is subject to weathering by carbonation, much the same way in which silicates weather. This process releases phosphoric acid $H_3PO_4^-$ or HPO_4^{2-}. These would be natural reference species in water. However, in soil much of this phosphoric acid is bound (occluded) into complexes with iron and aluminum, so the latter would be the logical reference species in the lithosphere. Iron is mostly mined from oxide deposits, also of biological origin, but iron is often partially oxidized (ferrous iron) even in igneous rock. The likely (most abundant) reference species is Fe_2O_3. Among the other major metals (chromium, copper, lead, manganese, nickel, zinc, etc.), the heavier metals are found mainly as sulfides in igneous rock; however, Szargut chose reference species in the lithosphere that are all oxides.

The halogens (F, Cl, Br, I) combine preferentially with the alkali metals, sodium, potassium, magnesium, or calcium. The halogenated sodium, potassium, and magnesium compounds are quite soluble in water and are mostly accumulated in the oceans, except for evaporite deposits. The logical reference species for the first three are simply the ionic forms; F^-, Cl^-, and Br^-. For iodine it is IO_3^-. In igneous rock, these salts are only found in trace percentages. However, for chlorofluorocarbons (CFCs), which are accumulating in the atmosphere, a specialized sequence of reactions—involving diffusion into the stratosphere and ionization by UV—may suggest a different exergy calculation.

Nitrogen is a special case. Molecular nitrogen in the atmosphere results only from biological processes, namely bacterial denitrification of organic nitrogen compounds. Nitrates are stored in sediments as organic material. They are found in igneous rock only in trace percentages. In principle, molecular nitrogen tends to combine with oxygen as nitrogen oxide (NO), nitrogen dioxide (NO_2), and eventually nitric acid (HNO_3), which reacts with ammonia in the atmosphere or dissolves in water and washes out as acid rain. The acid, in turn, reacts with most metallic hydroxides (in water) to form the corresponding nitrates. However, as mentioned already, the oxidation reactions in air are negligibly slow at normal temperatures and pressures. (They do occur more rapidly at high temperatures and pressures).

In the circumstances, it probably makes sense from an environmental standpoint to assume that one of the alkali metal nitrates [$Ca(NO_3)_2$, $NaNO_3$ or KNO_3] is also the reference species for nitrogen, contrary to Szargut's view, despite the fact that other compounds of these elements are much more abundant in the lithosphere. This may lead to some confusion, but as long as it is understood to apply only to the case of nitrogen compounds (and not to other compounds of the alkali metals, it should be possible to maintain reasonable consistency.

To summarize, it can be argued that a substance with a long—but not infinite—residence time in the atmosphere (or the ocean) might be a candidate for treatment *as if* it were a permanent resident. The argument is that when a compound remains within an environmental medium for a significant time, but subsequently disappears from that medium (whether by transformation, diffusion, precipitation, or washout), it may nevertheless be considered as a reference species within that medium. Candidates for atmospheric reference species might include nitrous oxide (150 years), methane (10 years), carbonyl sulfide (5 years), especially in comparison with carbon monoxide (60 days), sulfur dioxide (14 days), or hydrogen sulfide (4 days).

3. EXERGY BALANCE AS A TOOL FOR ANALYSIS

Unlike energy, exergy is not a conserved variable. It can be gained or lost in physical processes. However, exergy can be accumulated or depleted. It can also be stored in mineral ores or fossil fuels, for instance. Finally, as noted previously, exergy inflows and outflows to and from any subsystem are definable and measurable. (Computational aspects are discussed earlier.)

In the case of hydrocarbon fuels and carbohydrates such as cellulose and sugars, the exergy content of a fuel is very closely related to the usual measure (heat-of-combustion or enthalpy). However, as noted already, the term exergy is much more widely applicable. Given the previously stated conventions, it is definable for any material substance whatsoever, whether combustible or not. For instance, the exergy content of a metal ore (i.e., the work required to separate ore of that grade from the earth's crust) reflects the amount of natural "low entropy" that is stored in the ore. This would be lost if the ore were dissipated (mixed) uniformly into the earth's crust or into the ocean. The better the quality of the ore, the greater its exergy content.

Exergy, $\boldsymbol{B}$ as defined previously, is not only a natural measure of the resource inputs to an economic system. It is also a measure of the material outputs. Exergy is lost in all processes, mostly—but not entirely—as low temperature heat. But some exergy is lost in the form of chemically or physically reactive materials. These production or consumption wastes are dissipated into the environment, where the insertion of unfamiliar chemical species (i.e., chemical potentials) in delicately balanced biogeochemical cycles can cause trouble. At the microscale, even small trace amounts of some chemicals are enough to disrupt life processes. In fact, there is a general label for such disruptive chemicals: they are toxins.

To calculate the exergy content of waste materials, in practice, there are two approaches. Ideally they should be used together for mutual verification. The first approach requires a detailed knowledge of the chemical composition of the waste stream and the Gibbs free energies of formation of the components. Once these are known, the calculation of exergy proceeds component by component. The difficulty is obvious: for many chemical and metallurgical processes it is difficult to obtain reliable data on the chemical composition of the wastes. Nevertheless, the composition of the waste stream can be estimated approximately, in many cases, if the basic chemical reactions, temperatures, pressures, and yields are known. Indeed, commercially available software suitable for use in a desktop computer is capable of making such calculations at minimal cost.

The second, alternative approach involves using the exergy "balance" equation:

$$B_{in} = B_{product} + \Delta B_{process} + B_{waste}, \qquad (14)$$

which can be rewritten

$$B_{waste} = B_{in} - \Delta B_{process} - B_{product}$$

Here $\boldsymbol{B_{process}}$ is essentially a balancing term. It represents internal exergy lost in the process. Evidently, if $\boldsymbol{B_{process}}$ is known, the exergy content of wastes can be determined directly from the composition of the process inputs (including utilities) and that of the main products. The exergies of process inputs and product outputs are computed, as before, using Eq. (13), when Gibbs free energies of formation are known. However, as a practical matter, the exergy loss (or gain) in the process is typically much larger than the exergy content of the waste. Thus, to use Eq. (15) it is necessary to calculate process losses precisely. This is sometimes feasible on a case-by-case basis, but it is rarely easy.

4. EXERGY AS A MEASURE OF RESOURCE AVAILABILITY OR LOSSES

The use of exergy as a measure of resource stocks and flows for purposes of resource accounting was first proposed by Wall. The use of chemical exergy (or its equivalent, potential entropy) as a measure of potential harm from process wastes was a more recent idea. The latter notion also suggests the possibility that chemical exergy may be a useful indicator of sustainability.

However, while the research in this area so far is suggestive, there is more to be done before the application methodology can be regarded as complete and unambiguously successful. The problem is that the computational methodology (developed by Szargut) assumes that the exergy of a given chemical element is independent of the composition of the environmental sink into which it is injected. This implies that the exergy is independent of what specific reactions are induced in the actual sink. This implication seems reasonable enough if, for instance,

exergy is being used for purposes of resource accounting.

However, if exergy is to be used as a measure of potential environmental harm, it is unreasonable to assume that the chemical composition (and other characteristics) of the particular environmental sink are irrelevant. Hence, this simplifying assumption should be replaced by a methodology for calculating sink-specific measures. In particular, critics have emphasized that waste exergy, as calculated in terms of the three gross sinks—atmosphere, oceans, and lithosphere—is not a valid indicator of environmental toxicity. It is possible that a better measure could be developed if the internal chemical environment of living cells were considered as a fourth major sink. However, considerable research would have to be done to verify this conjecture.

However, although waste exergy is not (yet) a reliable measure of environmental toxicity, exergy analysis can be used to evaluate and compare waste flows and aggregate emissions from period to period or country to country. The exergy content of wastes is not necessarily proportional to the potential environmental harm the wastes may cause, but the exergy content of a waste stream is nevertheless a rough measure of its reactivity in air or water (i.e., its tendency to initiate spontaneous uncontrolled chemical reactions in environmental media). In this regard, one can say that although the exergy content of a particular waste stream is not a measure of human or animal toxicity, it is certainly a better indicator of its potential for causing harm than is its total mass, for instance.

SEE ALSO THE FOLLOWING ARTICLES

Complex Systems and Energy • Entropy • Exergoeconomics • Exergy • Exergy Analysis of Energy Systems • Exergy Analysis of Waste Emissions • Thermodynamics and Economics, Overview • Thermodynamics, Laws of

Further Reading

Ayres, R. U., Martinás, K., and Ayres, L. W. (1998). Exergy, waste accounting and life cycle analysis. *Energy—The Int. J.* **23**(5), 355–363.

Brobst, D. A., and Pratt, W. P. (1973). United States mineral resources. Professional Paper: 820. United States Geological Survey, Washington, DC.

Glansdorff, P., and Prigogine, I. (1971). "Thermodynamic Theory of Structure, Stability and Fluctuations" (Series). Wiley-Interscience, New York.

Nicolis, G., and Prigogine, I. (1977). "Self-Organization in Nonequilibrium Systems" (Series). Wiley-Interscience, New York.

Prigogine, I. (1976). Order through fluctuation: Self-organization and social system. *In* "Evolution and Consciousness" (E. Jantsch and C. W. Waddington, Eds). Addison-Wesley, New York.

Szargut, J., Morris, D. R., and Steward, F. R. (1988). "Exergy Analysis of Thermal, Chemical, and Metallurgical Processes" (Series). Hemisphere, New York.

Experience Curves for Energy Technologies

CHRISTINE WOERLEN
Boston University
Boston, Massachusetts, United States

1. **Definition of Experience Curves for Energy Technologies**
2. **Experience Curves**
3. **Reasons for Cost Reduction and Consequences for the Formulation of Experience Curves**
4. **Problems of Data Identification and Statistical Inference**
5. **Some Case Studies**
6. **Conceptual Modifications**

Glossary

diffusion Increase in market share of a \ew technology after a successful innovation.
economies of scale Decrease in unit cost of production through increase in scale of production.
experience curves The relationship between unit cost and cumulative deployment as represented by an exponential decay function.
innovation First commercial application of a newly developed technology.
interindustry spillover Transfer of knowledge, technologies, materials, and the like from one industrial application to another (e.g., the concept of jet engines from aeronautics to gas turbines).
intraindustry spillover Transfer of formerly exclusive knowledge and experience from one competitor to another (e.g., by labor mobility).
learning by doing Increase in skill level of labor through increasing experience.
learning curves The relationship between falling unit cost of production and increasing skill level of labor caused by an increase in experience, represented by an exponential decay function.

Experience curves are used to assess the future cost of a technology as a function of cumulative installations. The underlying rationale is that it is very costly to invent, develop, and prototype a new technology, process, or product. After an innovation—the first commercial application—a successful technology or product will spread from niche applications to larger markets. In this process of diffusion, unit costs will constantly decrease. This reduction of unit costs over time is described by experience curves.

1. DEFINITION OF EXPERIENCE CURVES FOR ENERGY TECHNOLOGIES

The concept of experience curves was derived from empirically observed learning curves in manufacturing. Learning curves describe the cost reduction effects by learning by doing. Through individual, organizational, and inter- and intraindustrial learning, the input of labor necessary to produce a certain product decreases as experience is gained. On the other hand, experience curves describe the cost reduction of a product or technology in general, disregarding the reasons for total cost reduction. Possible reasons for the reduction of unit costs during technology diffusion are economies of scale in all input factors (e.g., intermediates, transaction costs, other overhead), higher rates of production, changes in the production processes, and (particularly in the case of energy technologies) efficiency gains and changes in the product itself.

In the last years, experience curves have drawn considerable attention in the energy field due to a variety of technological challenges. Energy systems in many developing countries are in a phase of expansion and growth. Energy systems in most developed countries go through some restructuring, inducing a change in investor behavior. Both phenomena increase interest in present and future investment costs. Legislators, electricity generators,

 © 2004 Elsevier Inc. All rights reserved.

and other energy producers are also confronted with environmental demands (e.g., climate change, pollution, nuclear disposal and safety) that require the increased use of new cleanup technologies and new clean energy conversion technologies. Most of these new technologies are in early stages of their diffusion and considerable future cost reduction is expected.

Experience curves also play an important role in energy policy analysis. They are used to assess the "buy-down" cost of a new technology (Fig. 1). To optimize the size of a support program for a specific new technology, experience curves are used to determine how much additional investment is necessary to bring the cost of the technology down to levels at which it is economically competitive with other technologies. To that end, the incremental cost of the new technology and the speed of the cost reduction are determined from empirical data. The integral under the experience curve is equal to the investment required to buy down the cost of the technology between two arbitrary points on the curve.

Experience curves are also widely used in energy scenarios calculations. Investment costs and conversion efficiencies of energy technologies are important driving factors for future energy demand and supply. The most important energy system models, such as those of the U.S. Energy Information Administration and those used in the Intergovernmental Panel on Climate Change's (IPCC) Special Report on Emission Scenarios, use experience curves for the extrapolation of future energy costs because future investment costs for energy conversion and use are needed as inputs to these models. Because of their low data requirement and relatively robust nature, experience curves are very well suited to generate these inputs.

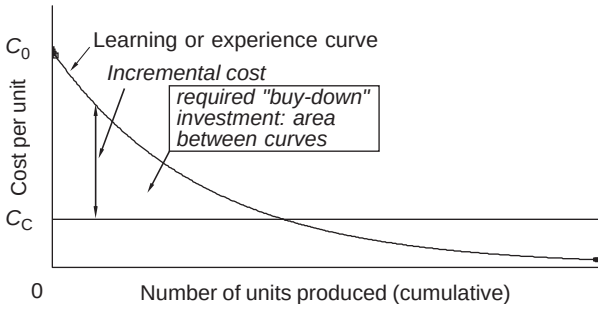


FIGURE 1 Experience curve for an equipment unit of a new energy technology. The unit cost for a piece of equipment falls with increasing cumulative deployment of the technology. C_0 is the cost of a technology at innovation. C_c is the cost of a conventional backstop technology. The integral over the incremental cost up to the point of intersection ("buy-down") indicates how much additional investment is necessary to make the new technology economically competitive with the established technology.

The more sophisticated energy system models can include dynamic specifications into the model code such that the prices used for optimization of the model are generated during the model runs and depend, in a dynamic way, on the respective deployment figures of the current simulation.

Management consulting firms (e.g., Boston Consulting Group) use experience curves for consulting clients on strategic price-setting behavior. Experience curves describe cost reduction as a "black box." The actual reasons for the cost reduction are completely ignored in the model. This makes experience curves seemingly simple to use but also imposes some caveats on the use of the concept. The data requirements for an estimation by linear regression are seemingly very low, and extrapolation into the future seems almost trivial. This article first gives more detail on the concept and then discusses the reasons why cost reduction for the whole product or technology might occur. It explains how these reasons should be incorporated into the model and what kind of data problems result. Several examples for cost reduction in various energy technologies and some extensions to the model are discussed.

2. EXPERIENCE CURVES

2.1 The Concept

Experience curves assume a logarithmic relationship between the experience from cumulative installation and cost. Typically, this is represented by the following functional form:

$$C_{\text{cum}} = C_0 * CUM^b, \tag{1}$$

where CUM is the cumulative output of the industry, C_{cum} is the cost of a unit product after output of CUM, b is the "learning index," and C_0 is the cost at the starting point of the data.

This functional form has some very convenient mathematical properties. For example, the parameter estimates are independent of the starting point of the data used for the estimation. The functional form is completely independent of the stage of development or diffusion of the technology. It implies that the cost reduction per additional unit of output decreases, so decreasing marginal reduction in costs is factored in. However, the functional form also prescribes that there is no limit to cost reduction. Some authors embrace this because they think that human ingenuity and competition will supply the opportunity for continually decreasing cost.

Other authors suggest including a minimum level of cost in the functional form by adding a simple linear term.

The functional form of the experience curve is estimated with linear regression after logarithmic transformation:

$$\log C_{\text{cum}} = \log C_0 + b \log CUM. \qquad (2)$$

Empirical studies usually estimate the coefficients of this equation using ordinary least squares and usually report highly significant correlations for the logarithmic form. However, highly significant correlations are to be expected in a logarithmically transformed linear regression of integrated variables and should not lead the investigator to conclusions that are not actually supported by the data. In fact, econometrically tested and statistically significant estimations of the experience curves are hard to find in the literature.

The data requirement for the estimation of Eq. (2) is very small. The data necessary for production cost forecasts (i.e., cumulative installations) are easily available from energy scenario calculations. The high level of aggregation of the data does not require any explicit knowledge about future changes in technological details that are particularly hard to predict. However, some researchers prefer to modify the parameters for time periods in the far future or distant past rather than assuming homogenous cost reduction over several decades.

2.2 The Progress Ratio

The "progress ratio" (PR) describes the relative cost after a doubling of cumulative installation. It is a very explicit and obvious benchmark for cost reduction effects. It is calculated from the learning index b estimated with the experience curve:

$$PR = 2^b. \qquad (3)$$

A progress ratio of 85% means that costs are reduced by 15% with each doubling of installed capacity. This ratio is constant. It is also independent of the rate at which the experience is accumulated, of the developmental stage of the technology, and of possible qualitative changes in the product or technology.

The progress ratio has been calculated for other, non-energy-related industries (e.g., products of the chemical industry or airplane manufacturing). Typically, progress ratios lie between 75 and 85%.

2.3 Scope

Experience curves have been estimated for single firms but are more common for a whole sector or whole industry (e.g., the wind turbine industry, all makers of combined cycle gas turbines) or for cumulative experience in all varieties of one technology (e.g., solar photovoltaics). In a globalized world, and specifically in energy technology markets, it is reasonable to assume that learning processes draw from a worldwide base of experience. Therefore, experience curves for energy technologies should reflect global experience, and global data should be used for their estimation.

3. REASONS FOR COST REDUCTION AND CONSEQUENCES FOR THE FORMULATION OF EXPERIENCE CURVES

Experience curves, as a typical black box model, do not make any assumptions about the reasons for cost reduction. However, for the applications of experience curves, these may not be ignored. This section explains the reasons for cost reduction and implications for the interpretation of progress ratios and cost reduction effects. In energy technologies, there are five reasons for cost reduction and improvements in the competitiveness of a technology: learning by doing, economies of scale in production of the equipment, improvements in the efficiency and quality of the equipment, standardization, and market conditions.

3.1 Learning by Doing

Learning by doing results in cost reductions whenever a firm or any other type of organization improves its productivity of labor through day-to-day operations. Wright studied the development in labor productivity in the U.S. airframe industry and showed that the number of hours required to produce an aircraft declines with the cumulative number of aircrafts produced. Arrow introduced the learning by doing hypothesis into economic mainstream and built a theory of macroeconomic growth on it.

Cost reduction through cumulative experience can be achieved within the firm, but firms can also profit from intra- and interindustry spillover effects. Mobility of "human capital" and copycat behavior are possible causes of intraindustry spillover and can be treated jointly with learning by doing even when looking at a

single firm. Argote and Epple showed that "organizational forgetting," which would reverse the cost reduction through learning by doing, occurs as well.

Cost reduction through learning by doing has been estimated in the past using the same functional form that later served as the blueprint for the experience curve. Learning by doing reduces the overall cost of a project to the degree that labor is used as a factor of input to production. The share of the total cost for this input factor is not likely to remain constant over time. Thus, if learning by doing is not the only factor contributing to cost reduction, theory does not fully support the use of the same functional form for both curves even if this is common practice.

3.2 Economies of Scale in Production and in Input Factors

The classic reason for cost reduction is economies of scale in the production process and in the procurement of input factors. When the output of an industry grows from a few hand-manufactured units per year to mass production, unit costs can go down by several orders of magnitude. Input materials for large-scale production can be purchased at better prices, and overhead can be reduced. Administrative effort and transaction cost per unit are reduced.

Upscaling a given production facility happens in an incremental way. Thus, even the cost of the aggregate production of a whole industry, particularly a small industry such as any national wind turbine industry, might not follow a smooth decay but rather have kinks. Just like in the case of learning by doing, there are natural limits to the size of factories, after which economies of scale might actually disappear and turn into diseconomies of scale in the form of increases in unit costs. The experience curve is a simplifying and generalizing representation of this growth process.

Economies of scale refer to the scale of current production, that is, to the current output per time. Because the classic experience curve concept does not control for time or the rate of production, estimating the equation can result in the same progress ratios for a given cumulative output at low and high rates of production. The experience curve concept assumes that the experience contribution of the xth unit depends only on how many units of experience have been gained before and not, for example on the cumulative sales in monetary terms. This points to the problem of defining the exact metrics to be used in the curve (e.g., whether the cumulative output should really be expressed in units produced or in a different metric), discussed in Section 4.

3.3 Improvements in Efficiency and Usefulness of the Product

Energy equipment can also exhibit qualitative changes over time that lead to higher efficiency of conversion without increasing unit cost in proportion. For example, bigger wind energy converters do not usually cost more in proportion as they grow larger, but their specific (marginal) cost per megawatt of capacity decreases. In addition, the efficiency of fuel use typically increases with changes in quality and size of the generator, so that the specific cost of the energy service goes down even more.

Efficiency changes can be taken into account in experience curves, but only if the delivered energy service (e.g., the kilowatt-hours of electricity) are the measure of experience as well as the unit for which costs are considered. Just like in the case of diseconomies of scale, equipment might evolve into an application suited for specific niche markets and, as a result of these qualitative changes, specific costs might be higher in later stages of technology diffusion.

3.4 Diffusion and Standardization

A fourth reason for the reduction of cost with increasing maturity of a technology is the standardization that a successful technology will undergo during its diffusion from niche markets into general markets. Standardization drives down overhead and compatibility costs. A logistic rather than exponential curve seems to be the appropriate representation of this phenomenon.

3.5 Market Conditions

All phenomena discussed so far might not have any effect on unit cost of production or on prices of the technology if the market structure is not sufficiently competitive. Not only might monopolists and oligopolists be able to charge prices that are far higher than the costs of developing, producing, and distributing a product or technology, but a noncompetitive market structure also relieves market participants from the pressure to improve products and production processes. Thus, learning and efficiency gains might be realized more slowly than in competitive situations. Other market conditions, such as the maturity of the market on the demand side and the opportunity of

spillover effects, might affect the speed of learning and improvements of the industry standard in economic or technological terms and, thereby, have indirect effects on the cost reduction over time.

In addition, the market situation in input factor markets (e.g., labor, materials) can play an important role in both improving and reducing cost reduction and experience effects. However, in an abstract and very general model of cost reduction with increased experience, it is very hard to account for these effects.

4. PROBLEMS OF DATA IDENTIFICATION AND STATISTICAL INFERENCE

Considering the reasons for a possible decrease in unit cost of energy technologies, it becomes clear that the variables representing experience as well as cost need to be chosen with care. Some of the changes (e.g., improved efficiency of conversion, improved reliability) only reduce the cost of the delivered output of the technology, whereas others (e.g., economies of scale in production, learning by doing) reduce the cost of equipment manufacturing and will have effects on both possible metrics: the cost of capacity and the cost of the energy service. Other identification problems and problems of statistical inference are discussed in this section.

4.1 The Dependent Variable

4.1.1 Cost per Capacity or per Energy Service

Most of the newer studies of energy technologies, particularly those in the field of renewable and intermittent energy sources, concentrate their analyses on the cost per unit capacity of the technology. These data are readily available for some technologies. Data on the cost of the energy itself (e.g., the power output for some technologies) are harder to obtain for several reasons. The actual energy output depends on local load factors and the supply with fuel inputs. Fluctuations in fuel prices and operation and maintenance (O&M) costs introduce additional uncertainties. Long-term trends in fuel prices or O&M costs might even offset any cost reductions through cumulative experience if measured in prices per energy service. Still, when experience curves are used for a comparison between different energy technologies, the assessment should always refer to the cost per unit of energy (e.g., kilowatt-hour) because this metric accounts for internal changes in the energy technology as well as in the other conditions that influence the relative costs of different technologies. Only this metric, if any, allows for the assessment of buy-down costs (as defined earlier as the amount of investment needed until the costs reach the level of any backstop technology). To do so, it is recommended to define standard (e.g., average) conditions for load factors and fuel input prices and quantities. To extrapolate cost reduction into the future, the standard situations (e.g., load factors, efficiency, fuel prices) need to be extrapolated into the future as well. This apparent abstraction is a reminder of the character of experience curves as a highly abstract tool for long-term technology assessment rather than as a planning tool for the design of an energy system on short and intermediate time horizons.

Related to this problem is the question of how to include O&M costs. These might be subject to a long-term cost reduction of their own as experiences with use of the new technology are collected and as the design of new equipment is improved for ease of use and reliability.

4.2.2 Costs or Prices

Another problematic aspect of the dependent variable is the question of whether cost or price data are used. Experience curve theory relates to the production cost for equipment or the energy product. Most empirical applications of experience curves to energy technologies use prices because they are easier to obtain than are cost data. When using price data as a proxy for costs, researchers implicitly make very restrictive assumptions on the market structure. Because experience curves include observations from many years, structural changes in the markets for energy technologies should be excluded if price data are used. Another major reason for various trends in prices and costs is changing market transformation policies and subsidy schemes for new energy technologies.

4.2 The Independent Variable: The Measure of Experience

4.2.1 Cumulatively Installed Capacity versus Cumulative Production of Energy Service

The measure of the accumulation of experience should be chosen in close analogy to the measure of cost for which the experience curve is estimated. For the reasons mentioned previously—economies of scale, market fluctuations, and qualitative changes in

the technology—cumulative energy services delivered would be the relevant measure for experience when thinking about the competitiveness of an energy technology, but data access and international aggregation are easier for cumulative capacity in megawatts. Current research investigates the usability of other measures of experience (e.g., cumulative investments, cumulative units of installation).

4.2.2 The Specification of the Technology

Whenever spillover effects from other industries are present, the estimation of cost reduction is possibly flawed. This is often problematic when technological concepts originate in other industries (e.g., gas turbine having its roots in jet engines). Also, the design of a technology might change in the course of product diversification and the establishment of a technology in new (e.g., distributed) applications.

The attribution of learning and accumulating experience also depends on which technology one actually examines. For example, the outcome when the experience collected with solar photovoltaic power generation is aggregated over all semiconductor materials that have been used for the purpose is drastically different from that when just the currently prevalent material is included in the experience measure. This phenomenon is highly technology specific, and no general recommendation can be given.

4.2.3 Global or National Scale

Typically, in a globalized world, the cumulative experience worldwide should be used as the driver for cost reduction through increased experience. With the exception of nuclear power, the markets for energy technologies are typically global markets. However, if there are specific local dynamics in a technology, the time series could become inconsistent. Furthermore, global data can sometimes be very hard to access, particularly for new technologies that are developed in several countries in parallel.

4.3 The Functional Form: Expression of the Speed of Cost Reduction

The assumption of exponential decay of costs reflects the fact that marginal improvements in the production processes diminish with increased experience. For experience curves, this form has been inferred from the idea of the learning curve. Fast learning with the first batches of production leads to rapid cost reduction initially, which can slow down during later stages of higher industry or business maturity. The traditional logarithmic shape of assumed cost reduction has very pleasant mathematical properties that technically allow for the estimation by simple linear regression and ordinary least squares methods. However, there is no reason to assume that the reduction in production cost for any piece of energy technology can be indefinitely continued, as implied by this functional form. Common sense suggests that the material cost is a natural lower limit for the cost. This means that the function should at least contain a constant price floor. The smooth logarithmic function also does not account for radical changes, implying that the curve should not be used for monopolies or monopsonies.

Furthermore, as discussed previously, learning and the other cost-reducing phenomena strongly depend on the rates of production. Thus, not all influences are captured by the classic functional form, and a multivariate approach might be much more appropriate even at the expense of the simplicity of the approach. In addition, market conditions and the circumstances of development and production of the equipment are likely to change during the time in which experience is accumulated, so that even at constant rates of deployment or constant rates of growth of production, the velocity of learning and improvements in the technology will not remain constant. Although the simple functional form has its own merits, it might not be able to account for the full story. This once again underlines the black box character of the approach. Experience curves are a tool for generating smooth curves for future projections on a large scale and for the long term, and they are not necessarily suitable for the assessment of historic trends or detailed short-term predictions. Because of this limited ability to reflect the actual processes, experience curves might be suitable only for very coarse assessments.

4.4 Spurious Regression

The experience curve concept implies a statistical relationship between a cumulative variable (e.g., cumulative deployment) and a noncumulative variable (e.g., unit cost). A statistical analysis of such a pair of variables needs to account for the autocorrelation and integration characteristics of the data. Typically, the cost at time t_{x+1} after the deployment of CUM_{x+1} will not be independent of the cost in time t_x after the deployment of CUM_x but can be represented by $C_{x+1} = C_x - \Delta C + \varepsilon$ (i.e., it should resemble a random walk with a downward drift term). Cumulative deployment CUM_{x+1} can be represented by $CUM_{x+1} = CUM_x + \Delta CUM$, where

ΔCUM equals the amount deployed during the period of observation. If the period of observation is a regular interval and the technology exhibits a typical growth pattern of a diffusion phase, the amount deployed during each period ΔCUM itself can be represented by $\Delta CUM_{x+1} = \Delta CUM_x + \Delta(\Delta CUM_x) + \varepsilon$ (i.e., it a random walk process with a upward drift). Thus, CUM_{x+1} does not fulfill the requirements for a stochastic variable. Rather, the classic concept requires an ill-defined regression of a random walk with drift with the integration of a random walk with drift. This vulnerates the standard assumptions of ordinary least squares regression, and in most cases a logarithmic transformation does not change the inherent characteristics of the data, so that the regressions based on the classic concept are most likely to be spurious. Extensive testing of the data and the error term should be conducted to prevent conclusions that are not supported by the data. However, proper statistical treatment and testing is the exception rather than the rule.

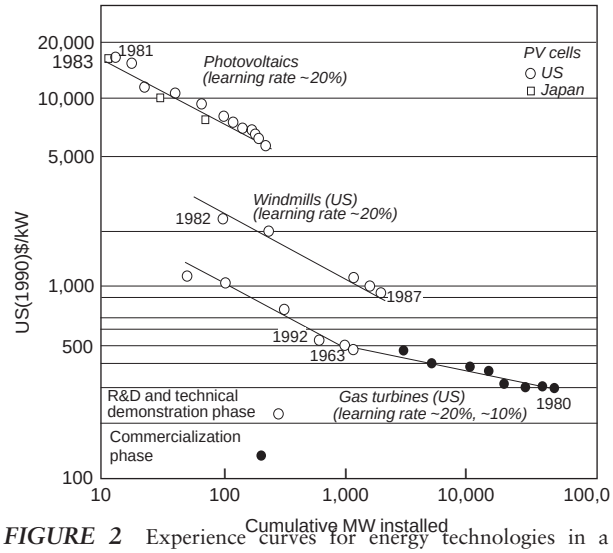


FIGURE 2 Experience curves for energy technologies in a variety of International Institute for Applied Systems Analysis estimates. Figure courtesy of Nakicenovic, N., Gruebler, A., and MacDonald, A. (Eds.) (1998). "Global Energy Perspectives." Cambridge University Press, Cambridge, UK.

5. SOME CASE STUDIES

Recently, the classic experience curve concept has been applied to the area of new energy technologies, mostly renewable energy technologies but also combined cycle gas turbines. In this section, several technology-specific applications are discussed. These case studies demonstrate possible problems with the concept and new energy technologies and offer interesting ideas for improvement. Several large-scale research efforts on experience curves were aggregated in the International Energy Agency (IEA)-led International Collaboration on Experience Curves for Energy Technology Policy (EXCETP). All case studies have in common that where more than one estimate has been made, the estimated progress ratios span wide ranges where, as described previously, the estimate should be robust with regard to the selection of data. This fact is a clear warning to pay close attention to selecting the metrics and the database for the estimation and to testing the estimates for statistical validity, as explained previously. Figure 2 contains a classic display of several examples of estimates of experience curves that have been partly included in the following literature survey.

5.1 Wind Energy Converters

Several studies have applied the concept to various data sets with varying success. Junginger reviewed 16 studies on cost reduction in wind turbine production and found that they used different cost units and different proxies for the cumulative experience—and got very different results. Most of these studies interpreted the results of the classic function with a very specific set of reasons. These studies hardly ever accounted for the data, identification, or statistical problems mentioned previously. Most studies indicated that progress ratios for wind energy converters have been much higher (the cost reduction has been slower) than the classic rule-of-thumb progress ratios of 75 to 85%. The range of results was 62 to 96%.

5.2 Solar Photovoltaics, Niche Markets, and the Influence of Market Transformation Programs and Policies

Solar photovoltaics has been studied extensively with experience curves. Progress ratios between 68 and 82% have been estimated. The technology was characterized by high initial investment cost and comparatively high maintenance cost due to changing micro-design. The countries with the highest installed capacity of solar panels are Japan and Germany, where the installation was subsidized. Other market transformation initiatives have been launched for developing countries. Using these examples, the effect of market introduction subsidies on the cost of the equipment has been estimated. The induced cost reduction effects have then been

factored into the benefit/cost ratio of the policies to justify the intervention. The experience curves were the basis for the assessment of the necessary amount of subsidy.

Even if the fuel is free and even in good locations, electricity from photovoltaics currently is more expensive than standard grid power. However, in certain niche markets, either sunny rural locations remote from the grid or sunny urban locations where no backup power supply is available, photovoltaics can be a cost-effective solution. From this starting point, the current costs of photovoltaic systems, the market size of the niche markets, and the progress ratio can be used to calculate the point in time at which solar photovoltaics will break even with grid-connected power technologies.

5.3 Gas Turbines and the Influence of the Diffusion Stage of the Technology

Nakicenovic and colleagues estimated a two-stage learning process for gas turbines. After a steep decay in cost with a progress ratio of approximately 80% in the airline industry, a lower rate of cost reduction follows with a progress ratio of approximately 90% (Fig. 2).

According to Colpier and Cornland, who analyzed the economics of the natural gas-fired combined cycle gas turbine (CCGT), the specific investment price for larger CCGTs has decreased by as much as 35% in real terms during the past decade. Compared with experiences in other industries, this is a very high decrease. CCGT is not a new technology in itself but rather a combination of the established technologies of gas and steam turbines that has been used since the 1970s. To find an explanation for the high rates of cost reduction, Colpier and Cornland expanded the classic theory by taking into account strategic pricing behavior. They hypothesized that a strategically pricing firm sells at prices lower than marginal cost during the development phase of a product. The firm develops into a market leader, reduces its own marginal cost with increased experience and can sell under a "price umbrella," that is, at prices higher than marginal cost and constant until competitors have caught up. In the subsequent "shake-out phase," prices will drop sharply and approach marginal cost. The authors claimed that this phenomenon has been observed for many different technologies. Their analysis indicated that the prices for CCGT technology have been dropping in a way that looks like a margin shake-out rather than the less drastic decrease of unit construction cost through experience and economies of scale. This could be supported by the fact that the CCGT market is dominated by only four manufacturers and their licensees. The authors concluded that the decline of the specific investment prices is likely to level off in the future.

More generally, pricing behavior is linked not only to the developmental stages of a technology but also to the rate of deployment and the rate of technological change, both of which affect the rate of learning and of cost reduction in different ways.

5.4 Gas Infrastructure

Zhao investigated historical cost development of gas pipelines. She ignored O&M costs and focused on the construction cost per cubic meter of pipeline capacity. In the United States, these costs increased and decreased twice between 1985 and 1998. Zhao observed no learning effects in onshore pipeline construction and found no significant cost reductions or progress ratios. For offshore pipeline installation, she measured cost in U.S. dollars per cubic meter of pipeline volume and days required to lay 1 km of pipeline. Her data, in fact, indicated that costs go down, although no statistical significance could be achieved. She concluded that because pipeline construction is not a replicative process, learning effects are probably small. Economies of scale can be achieved (e.g., through increased capacities, i.e., pipe diameter), although it is unclear how she derived these results. Experience curves seem to be an inappropriate tool for the analysis of the cost of gas infrastructure.

6. CONCEPTUAL MODIFICATIONS

Quantitative estimates for experience curves should be used with caution because the theoretical foundation, as well as the empirical support, is rather weak. Nevertheless, the occurrence of cost reduction with increasing diffusion and market penetration of a technology is undisputed. To get better estimates for the velocity of cost reduction, more research is under way.

One school of modifications amends the factors that are ignored by the traditional formulations by a simple addition of the logarithm of cumulative expenses for research and development (R&D) in Eq. (2). This formulation implies that cumulative R&D expenses are a perfect substitute for commercial investments in the technology after its innovation

in their effect on cost reduction. Attempts to estimate this specification have not resulted in satisfactory results, due mainly to the statistical and conceptual problems discussed previously.

Other studies have tried to decompose the cost reduction effect into its components, as explained in section II. Here, too, the statistical problems of serially correlated and integrated regression residuals and unit roots have not been solved. However, future research into such multivariate approaches is necessary and promising.

In this respect, it is also useful to work on temporal analyses of economic cost functions. Isoard and Soriat took a first step in this direction when they combined cost functions and learning curves to capture the simultaneous econometric estimation of learning by doing and returns to scale, endogenizing technical progress. A test for Granger causality establishes the causal relationship between deployment and cost reduction.

Other extensions are possible and should be explored. In theory, experience curves can be conceived for aspects other than energy unit costs. Other parameters that improve with cumulative deployment through continuous changes in the average technological design could be the efficiency of energy conversion or the emissions of pollutants from a conversion process. For these, too, the average improvement over time could be modeled with exponential decay functions. Linking these results back to the unit cost can give additional insights into the nature of cost reduction processes.

For future research, it is important to strengthen the foundation of the concept. One such avenue is linking the concept to the life cycle analysis of technologies. The life cycle of a technology starts with the invention of a new technological concept. The concept is simulated and prototyped before the actual innovation. The innovation itself consists of the first commercial application of the technology. For energy technologies, this often happens in niche markets (e.g., aeronautics). The innovation is followed by a diffusion phase, when the market penetration of the technology increases, often along a similar (logistic) path as followed by the sinking cost of the technology. The forces effecting such a diffusion pattern can theoretically be subdivided into supply–push and demand–pull forces. Therefore, market forces and the economic environment should be included in deriving a better founded concept for the cost reduction of new energy technologies.

SEE ALSO THE FOLLOWING ARTICLES

Cost–Benefit Analysis Applied to Energy • Environmental Kuznets Curve • Resource Curse and Investment in Energy Industries • Technology Innovation and Energy

Further Reading

Argote, L., and Epple, D. (1990). Learning curves in manufacturing. *Science* **247,** 920–924.

Arrow, K. J. (1962). The economic implication of learning by doing. *Rev. Econ. Studies* **29,** 155–173.

Colpier, U. C., and Cornland, D. (2002). The economics of the combined cycle gas turbine: An experience curve analysis. *Energy Policy* **30,** 309–316.

Duke, R., and Kammen, D. M. (1999). The economics of energy market transformation programs. *Energy J.* **20**(4), 15–64.

International Energy Agency. (2000). "Experience Curves for Energy Technology." Organization for Economic Cooperation and Development/IEA, Paris.

Isoard, S., and Soriat, A. (2001). Technical change dynamics: Evidence from the emerging renewable energy technologies. *Energy Econ.* **23,** 619–636.

Junginger, M. (2000). "Experience Curves in the Wind Energy Sector: Use, Analysis, and Recommendations, report NWS-I-2000-29." Department of Science, Technology, and Society, Utrecht University, Netherlands.

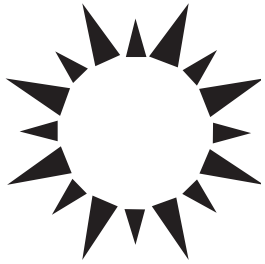
Woerlen, C. (2003). Economic and technological aspects of the market introduction of renewable power technologies. Ph.D. dissertation, Boston University.

Wright, T. P. (1936). Factors affecting the cost of airplanes. *J. Aeronautical Sci.* **3,** 122.

Zhao, J. (2000). "Diffusion, Costs, and Learning in the Development of International Gas Transmission Lines, IR-00-054." International Institute for Applied Systems Analysis, Laxenburg, Austria.

External Costs of Energy

DARWIN C. HALL
California State University
Long Beach, California, United States



1. Economic Theory of Externalities
2. Categories of Externalities: National Security and Pollution
3. Estimates of External Energy Costs
4. Policy Choices
5. Conclusions

Glossary

British thermal unit (btu) The average amount of energy per degree Fahrenheit needed to raise the temperature of 1 pint of water from freezing to boiling at sea level, which is equal to the energy needed to raise the temperature by 1°F, from 60° to 61°.

common property It has exclusions on use and allocation of a benefit stream determined collectively by community agreement or through government.

e-folding time This is for exponential decay what doubling time is for exponential growth. Suppose that a stock can be described over time by exponential decay to zero. At initial time, the amount is X_0 and over time the amount is given by $X_t = X_0 \exp(-rt)$. The time derivative is given by $dX/dt = -rX_t$. At $t = 1/r$, $X_t = X_0/e$, where e is the exponential number, approximately 2.71828. The ratio $1/e$ equals approximately 37%, so the e-folding time is the time at which the stock declines by 63% of the original value X_0. For an annual decay rate, for example, if $r = 2\%$, the e-folding time is 50 years.

NO_x Oxides of nitrogen, such as nitrogen dioxide (NO_2) or nitrogen monoxide (NO).

private property This has exclusions on use and enjoyment of a stream of benefits that are determined by a private owner or owner's agent.

public goods or bads These are not depletable by use and not excludable.

Vickery auction named after Noble Laureate William Vickery, this refers to an auction designed to avoid strategic bidding and to encourage small incremental bids that reveal the marginal cost curves of the bidders. The total amount is specified in advance. Each bidder may submit multiple bids for incremental amounts, and the lowest bids are accepted up to the point at which the total amount is reached. All winning bidders are paid the same price, not the amount they bid, equal to the lowest of all the losing bids.

willingness to accept (WTA) A schedule of prices and quantities that gives the minimum amount a consumer would accept in exchange for giving up an additional unit, as a function of the number of units the consumer holds.

willingness to pay (WTP) A schedule of prices and quantities that gives the maximum amount a consumer would pay in exchange for obtaining an additional unit, as a function of the number of units the consumer holds.

The external costs of energy are the costs not reflected in the market price, well over 100% for some energy sources. This large divergence between the social cost of energy and the price occurs because of environmental externalities associated with conventional energy sources and national security subsidies for imported oil and for nuclear power. As we shift to imported energy supplies, the remaining investment in domestic energy supply—extraction, transportation, and transformation—is 8% of all investment in the U.S. economy (measured by fixed assets in 2001). Whether domestic or international, many of these energy investments (power plants, pipelines, refineries, and oil and coal extraction) last for 50 years or more. Market prices that omit externalities are encouraging consumption of fossil fuels, investments in military preparedness, international investments in fossil fuel supply, and domestic investments in low-fuel-economy vehicles. At the same time, the same market prices discourage investment in alternative energy technologies, discourage investment in high-fuel-economy vehicles and vehicles that use alternative fuels, and discourage investment in energy efficiency. External costs of energy provide incentives for long-term investments resulting in gross and lasting economic inefficiency.

External costs occur at each stage, from exploration to production, transportation, conversion, consumption, and waste disposal. National security

 © 2004 Elsevier Inc. All rights reserved.

external costs accrue to oil because of national defense expenditures to secure oil that are paid by general taxes rather than taxes on oil imports. National security risks from the use of nuclear power range from the risk of attacks on power plants to attacks on waste transport and storage and the spread of nuclear technology and material. External costs of fossil fuels result from pollution through environmental media—soil, oceans, freshwater, the troposphere, and the stratosphere. Damage occurs to human health, there is a reduction of ecosystem services to the economy, and there is loss of the intrinsic value of nature. Economic theory helps to focus on marginal (rather than total) external costs and to separate external costs from costs that have been internalized through the market or regulation. Economic theory faces formidable challenges due to joint pollutants, nondepletable catalytic pollutants, and nonconvex, subtle, cascading effects among physical, biological, and social systems. Overly narrow applications of economic theory, particularly marginal benefits and marginal costs of externality abatement, are harmful to the formulation of well-designed policy options. The limits of the economic theory of externalities justify policies, such as the Clean Air Act, that do not require a balancing of marginal benefits and cost to set ambient air quality standards. The challenge is to design, implement, and improve policies that address unintended, harmful effects of energy development, production, and consumption.

1. ECONOMIC THEORY OF EXTERNALITIES

Economic efficiency is defined as occurring when it is not possible to make someone better off without making others worse off, which occurs when the marginal social benefit of consumption equals the marginal social cost of production. In theory, with a specific definition of "perfect competition," markets achieve economic efficiency. Some of the assumptions are the absence of market power, markets that provide costless and accurate information, rational behavior narrowly defined, the absence of public goods (or public bads), and the absence of externalities. Externalities occur whenever decisions of economic agents cause incidental costs or benefits borne by others that are not reflected in market prices. In particular, external costs drive a wedge between social and private marginal benefits and costs (Fig. 1).

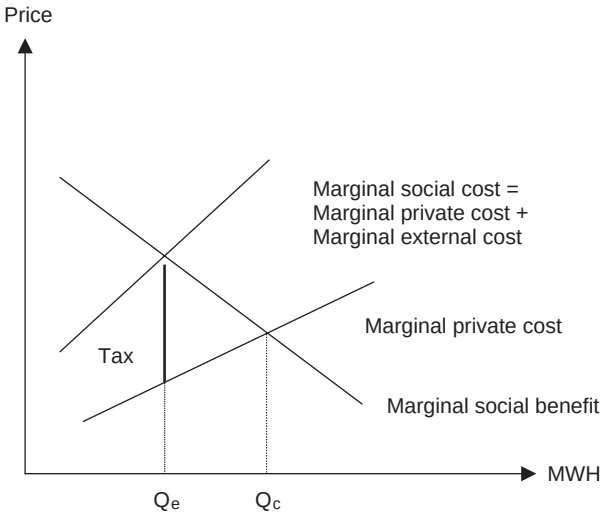


FIGURE 1 Marginal social benefit and cost.

If accurate information were costless, markets would be unnecessary. Markets provide imperfect information at a cost. All markets, whether for pollution credits, future prices of wholesale natural gas or electricity, or commodities sold to consumers, require government regulation to establish and enforce property rights. Property rights specify the right to pollute or, conversely, the right to a clean environment. Both markets and government enforcement of property rights that regulate markets are costly. Broadly, the idea of economic efficiency can inclusively consider the costs of markets and their regulation, including the costs of providing public goods and regulating externalities. In this sense, economic efficiency is the criterion most economists favor to evaluate policy changes. Although there is reason to be sympathetic toward the criterion of economic efficiency, the fundamental assumptions of economic efficiency are inconsistent with most applications to energy externalities. Finally, the criterion of economic efficiency is not adequate to assess strategies to reduce macroeconomic risks or technology forcing policies designed to alter the course of technological change.

1.1 Pigouvian Tax on Output versus Pollution Tax

The earliest analysis of externalities focused on external costs of production, for example, megawatt-hours (MWH) of electricity, where the marginal external costs are added to the marginal private costs to obtain the marginal social costs (Fig. 1). The amount consumed and produced by a competitive

market equals Q_c, and the economically efficient amount is Q_e. In order to achieve economic efficiency, government imposes a tax on output equal to the external cost at Q_e. This approach erroneously assumes that the externality is produced in fixed proportions so that, for example, a tax to account for NO_x emissions from a natural gas-fired power plant would also be applied—in error—to a solar thermal power plant.

Modern environmental economics poses the externality problem in terms of the marginal benefit and cost of externality reduction (Fig. 2). The marginal cost of reducing NO_x is given by listing all methods of pollution abatement in ascending order of incremental cost. The marginal cost is negative until the curve crosses the abscissa at N_f because investments in electricity conservation per megawatt-hour are cheaper than the incremental cost of building and operating new power plants. N_f is the level of air pollution abatement that is actually a free lunch. The marginal benefit of externality abatement depends on the property right—whether the polluter has the right or the public trust establishes a right to a level of a clean environment. The marginal benefit is given either by the damaged party's willingness to pay (WTP) for abatement or the damaged party's willingness to accept (WTA) payment in exchange for enduring the damage (Fig. 2).

The WTA is greater than the WTP due to an income effect and a substitution effect. If the polluter holds the right, the damaged party's budget constraint limits the WTP, but if the right is held by the public there is no budget constraint on the WTA; this is the income effect. The substitution effect can be understood using the example of air pollution; there are no close substitutes for cleaner air, so the WTA is much larger than the WTP. If the correct marginal benefit curve is WTP, then the optimal amount of abatement equals N_P and the optimal pollution tax equals T_P. If the correct marginal benefit curve is WTA, then the optimal amount is larger, equal to N_A, and the tax equals T_A. There is no unique solution.

Both the WTA and the WTP curves emanate from the point N_c in Fig. 2. The maximum possible reduction in NO_x emissions is equal to the total emitted so that N_c corresponds to zero emissions to the environment. Consequently, whatever the property right, the social demand for externality abatement reaches the abscissa at the point of abatement where the negative externality is zero. In order to simplify the figures and discussion that follow, a single marginal benefit curve (MB) replaces the WTA and WTP curves.

1.2 Demand for Externality Abatement over Time

Demand for externality abatement shifts from MB_1 to MB_2 to MB_3 in Fig. 3 for two reasons—public goods and concentration of the externality. Public goods, such as increased national security or air pollution abatement, have one key property: They are nondepletable in consumption. The social demand for nondepletable goods is found by vertically adding individual demand, in contrast to

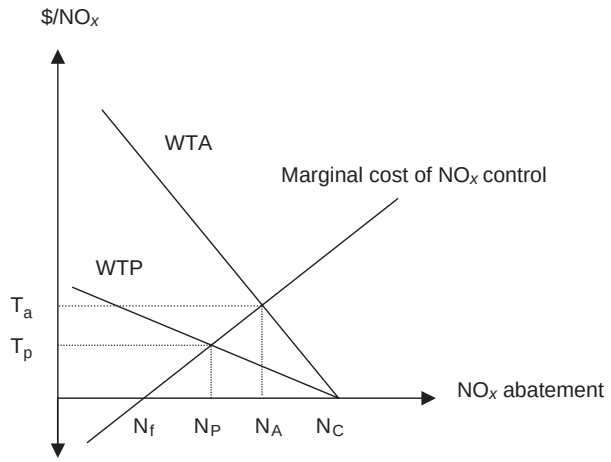


FIGURE 2 Marginal benefit and cost of externality reduction.

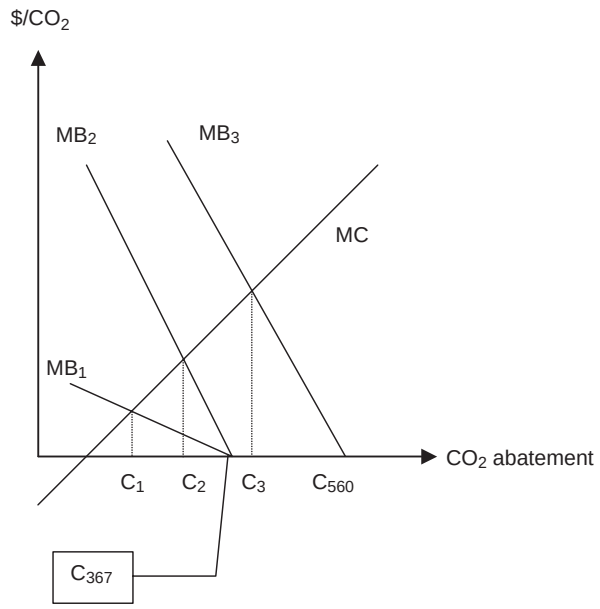


FIGURE 3 Optimal abatement over time.

the horizontal addition for private goods. Consequently, as population grows over time, the demand for externality abatement shifts from MB_1 to MB_2, holding pollution concentration constant.

An important air pollutant from fossil fuels is CO_2. CO_2 emissions have increased the ambient concentration of CO_2 since the industrial revolution. The ambient concentration was approximately 280 parts per million (ppm) in the late 1800s, and increased to 367 ppm by 1999. In the next 50 to 60 years, CO_2 concentration is forecast to reach 560 ppm, doubling the concentration since the preindustrial revolution. Within 200 to 400 years the concentration will increase 4 to 10 times unless we significantly abate emissions and sequester carbon from the atmosphere. In Fig. 3, let C_{367} represent the abatement necessary to return to 280 ppm in year 1999. Over time, the intersection of the demand for abatement and the abscissa will shift to the right, from C_{367} to C_{560} by the year 2055, resulting in the marginal benefit curve MB_3.

Over time, due to the public good nature of CO_2 abatement and population growth, and the growing concentration of the externality, the optimal externality abatement will increase from C_1 to C_2 to C_3.

1.3 Regulation, Internalization, and Incremental Change

Regulation internalizes the negative externality from pollution. For example, the local air pollution regulatory authority in the Los Angeles area established a set of stationary-source air pollution regulations requiring controls for power plants and oil refineries, among other pollution sources, to reduce particulates and precursors to ozone. The pollution control cost of electricity and refined petroleum products in southern California is reflected in the prices of these products. In 1990, the air basin was significantly out of compliance with the federal ambient air pollution standards—a point such as O_{1990} in Fig. 4. The federal standards for ozone, 0.12 ppm, exceed the state standards, 0.08 ppm, shown as $O_{0.12}$ and $O_{0.08}$, respectively. If the situation is correctly depicted in Fig. 4, the marginal benefit of abatement, B_{1990}, is greater than the marginal cost of abatement, C_{1990}, and additional abatement is efficient.

It is not possible, however, to estimate the marginal benefit curve for reasons discussed later. In the peer-reviewed literature, a team of economists, chemists, airshed modelers, and health experts estimated the incremental benefit to the Los Angeles region of attaining the federal standards for both ozone and particulate simultaneously. The estimate of incremental benefit equals approximately \$10 billion for federal attainment and \$12 billion (1990 \$) for state attainment, so the incremental benefit of moving from the federal to state standards was \$2 billion. If ozone and particulate were a single, homogeneous pollutant, and pollution abatement benefits were convex and met other technical assumptions in economic analysis, the incremental benefit of \$2 billion could be shown as the area under the marginal benefit curve in Fig. 4. Unfortunately, the standard economic theory of externalities shown in Fig. 4 is incorrect.

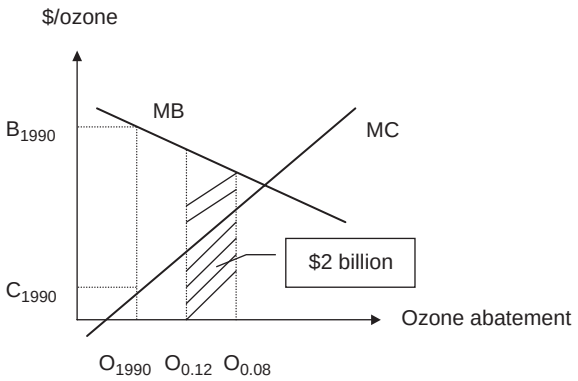


FIGURE 4 Partially internalized by regulation.

1.4 Joint Pollutants, Nonconvexities, and Nondepletable Catalytic Pollutants

Chemical reactions do not follow the assumptions of economic models. Ozone (smog) is created by the reaction of hydrocarbons and NO_x. Unburned hydrocarbons are released when we produce, process, and consume petroleum to get energy. Nitrogen, approximately 78% of the atmosphere, is stable at low temperatures, but fossil fuel consumption releases heat that creates NO_x. The ozone isopleths in Fig. 5A, similar to isoquants, give the combinations of reactive organic gases (ROGs) and NO_x that create a given amount of ozone. Because the isopleths include regions with positive slopes, in these regions decreases in NO_x actually increase the amount of ozone, holding ROGs constant, as shown by the arrow in Fig. 5A. The marginal benefit of NO_x abatement first rises and then falls, as shown in Fig. 5B, which is an example of nonconvexity.

Optimal abatement does not occur where the curves cross in Fig. 5B. First, if area A is less than B, it is better to do nothing (*ceteris paribus*). Second, emission control options typically reduce both NO_x and ROGs (*mutatis mutandis*), shifting the marginal

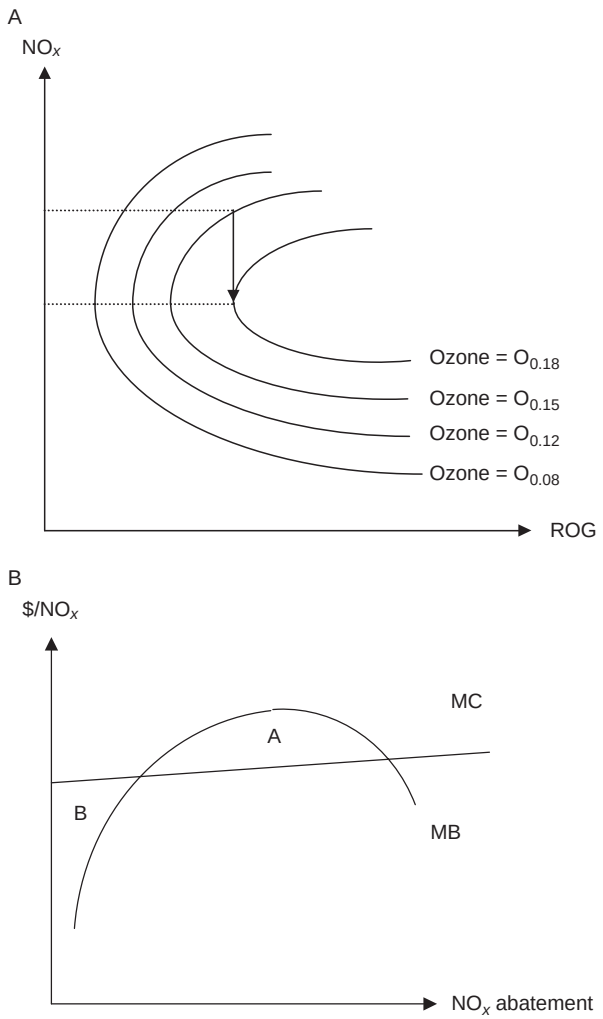


FIGURE 5 (A) Ozone isopleths. (B) Marginal benefit curve (MB) derived from isopleths.

benefit curve. Third, nitrate compounds from the nitrogen oxides are particulates with an aerodynamic diameter $<10\,\mu m$, small enough to slip through the nasal barriers and lodge into the lungs. The human body's immune system reacts in ways that cause irritation and exacerbate respiratory disease. Consequently, NO_x causes damage directly as well as indirectly through ozone. Since NO_x acts as a catalyst in the formation of ozone, the same molecule can later contribute to particulate. The analogy in economic theory is a public "bad", which is nondepletable in use. Hence, the marginal benefit of abatement is found by adding the demand for particulate abatement vertically to the demand for NO_x abatement derived from the demand for ozone abatement. These complexities render obsolete the economic theory of externalities for application to energy from fossil fuel.

1.5 Nonconvex, Subtle, Cascading Effects among Physical, Biological, and Social Systems

Consider the policy of planting shade trees and painting surfaces white to lower the urban heat island effect, reducing the demand for air-conditioning to conserve electricity. Electric utilities change the dispatch order, altering the amount of electricity generated from power plants within and outside the urban airshed. The location of power plants and the emission control equipment determine the change in emissions of ROGs and nitrogen oxides, each of which separately causes damage to human health and the environment, that together form ozone in the presence of sunlight and an inversion layer. Surface roughness, affected by planting trees, alters mixing and transport within the airshed. Different types of trees have different emissions of ROGs, further adding to the complexity of impacts. Financial impacts include the dollar values of fuel, pollution credits, electric system reliability and associated timing of power plant investments, tree maintenance, and painting, in addition to the external costs of changes in emissions of greenhouse gases and carbon sequestration. This level of complexity further renders useless the modern economic analysis of externalities depicted in marginal benefit and marginal cost curves for abatement.

1.6 Measuring Policy Options: Partial vs General Equilibrium

The portrayal of economic efficiency in Figs. 2–4 is based on partial equilibrium analysis. General equilibrium analysis models the interrelationships among markets, for example, accounting for impacts of changes in the price of oil that shift the demand for natural gas, shift the demand for vehicles, and shift the supply of transportation services, all of which in turn have effects that ripple throughout the economy.

Petroleum is a minor energy source in the residential and commercial sectors in the United States. Petroleum supplies 24% of the industrial sector, with ample room for substitutes, but petroleum supplies 97% of the transportation sector. Fuel cell technology is close to commercial viability and could be developed as a substitute for petroleum for transportation. Fuel cells would also substantially reduce urban air pollution. A policy to require that a fraction of sales be ultra-low-emission vehicles would speed the development of the infrastructure needed to commercialize fuel cells. A partial equilibrium analysis would deem a successful policy as a failure.

Whether based on hydrogen or natural gas-derived methanol, suppose a successful fuel cell program in 2002 results in transportation costs comparable to the price at that time—$30 per barrel of oil. Falling oil consumption would likely result in the collapse, or at least the diminishment, of the OPEC cartel market power. The resulting decline in oil prices could easily reach less than $10 per barrel, placing the fuel cells program politically at risk. Oil interests could commission partial equilibrium studies that show the fuel cells program to be an economic disaster at $30 per barrel, when in fact oil is at $10. Of course, the correct counterfactual for evaluating the fuel cells program would hypothesize oil prices in the context of a transportation sector without a substitute, requiring a computable general equilibrium (CGE) model with a cartel sector.

CGE models are routinely applied to environmental regulations to gauge the economic cost of policies to abate pollution. However, CGE models have not been modified to measure lost productivity from mortality, and ill-health effects of pollution causing worker absences or impairing human development, nor have they been modified so that increased medical expenses are measured as counterproductive rather than as adding to economic activity. Of course, the correct counterfactual for evaluating pollution abatement policy would correct CGE models to account for harm to human health, damage to agriculture and fisheries and forests, damage to vehicles and structures such as buildings and bridges, recreational losses, public good losses measured by contingent valuation studies, and so on.

2. CATEGORIES OF EXTERNALITIES: NATIONAL SECURITY AND POLLUTION

Although national security and pollution are the two most important negative externalities from energy production and consumption, neither should be estimated in isolation from the other. Most energy policies affect both national security and pollution.

2.1 Oil and National Security

Of the 68 million barrels of oil produced per day (mbd) in 2000, 30% were from the three top producers: Saudi Arabia (8.4 mbd), Russia (6.5 mbd), and the United States (5.8 mbd). This is changing because Iran, Iraq, Kuwait, Qatar, Saudi Arabia, and the United Arab Emirates hold 65% of the world's proven reserves, whereas the United States holds 2%. The United States consumes 19.7 mbd while producing 5.8 mbd, on net importing 70%. Of the total petroleum used in the United States, transportation use equals 68%. The transportation sector has virtually no substitutes for petroleum, which supplies 97% of all energy in that sector.

If externality abatement is subsidized through general taxes, the marginal cost of abatement is not necessarily equal to the optimal tax on the externality. As shown in Fig. 4, optimal policy design requires knowledge of both the demand for externality abatement and the cost of abatement. At the federal standard for ozone of 0.12 ppm, Fig. 4 depicts a marginal cost of abatement less than the marginal benefit. Another example is the national security costs of importing oil.

Broadman and Hogan identify three parts of an optimal oil import tariff: high oil prices that reflect monopsony power, adverse macroeconomic effects of higher oil prices on terms of trade, and the adverse impact of oil price shocks. Because the United States consumes 25% of world oil, the price depends on the amount consumed (monopsony power). The U.S. macroeconomy does not perform as well when oil prices are high, changing the terms of trade and requiring greater exports for the same amount of imports. The adverse impacts of oil price shocks are the basis of a 'security tariff' that can be calculated as the expected loss, equal to the product of the probability of an oil price shock from disruption of supplies multiplied by the negative impact of an oil price shock. The negative impact of an oil price shock can be estimated from the magnitudes experienced since 1970. The security tariff is interpreted as the willingness to pay for military protection of oil supplies, estimated in 1985 dollars to be equal to $7.00 per barrel (conditioned on the price of oil at $15/bbl). The axes in Fig. 2 could be relabeled so that the ordinate is in dollars per barrel and the abscissa is reduction in national security; thus, point N_C is complete security. Then the optimal tax, T_P, is $7.00 per barrel in 1985 dollars. The marginal cost curve in Fig. 2 would be the cost of additional security, including the cost of additional military preparedness and the cost of the strategic petroleum reserve.

Prior to the invasion of Kuwait by Iraq, the defense expenditures per barrel of imported oil plus the cost of the strategic petroleum reserve (SPR) were approximately $8.50 to $9.00 (in 1985 dollars) per barrel. I calculated the lost interest from storing oil, equal to the interest that would have accrued by investing the

amount of money paid for the oil that was stored in the SPR; in 1985 $ the lost interest equals $1.15 per barrel of imported oil, and there is an additional $0.60 per barrel for the increase in the size of the optimal reserve from its inception to the mid-1980s. I also performed a time series analysis of defense spending and oil imports during the period from 1968 to 1989. After removing time trends and random walks, the data reveal a correlation between defense spending in each year to oil imports lagged two years, where one year of the lag accounts for the time delay between publication of oil import data and Congressional appropriations, and the second year of the lag accounts for the delay between appropriations and actual budget years. The result is an econometric estimate equal to $7.30 of defense spending per barrel of imported oil in 1985 dollars. Consequently, the marginal benefit calculated by Broadman and Hogan is close to the marginal cost I calculated, and one may conclude that in the 1980s the United States provisioned close to the optimal amount of national security related to importing oil. The price of oil did and does not include the national security cost, as it was and still is paid through general taxes rather than a tax on oil imports, and this national security subsidy for oil has the unfortunate consequence of economic inefficiency.

Certainly, the national security cost of oil imports is higher today. Macroeconomic effects of oil price shocks include less productivity, greater unemployment, and an increase in general inflation. Some oil revenue flows to terrorists and states acquiring weapons of mass destruction. Recessions have reverberating impacts on presidential elections. Presidents Carter and George H. W. Bush and Vice President Gore lost reelection following oil price spikes. In all three cases, the Federal Reserve Board responded to general inflation caused by oil price spikes by increasing the discount rate. Whether the recessions were caused directly by the oil price spikes or indirectly by the misdirected response of the Federal Reserve Board, the effect is the same: a recession during a presidential election.

2.2 Nuclear Power and National Security

Three types of terrorism are plausible. First, nuclear power plants and research reactors in the United States are potential targets for terrorists determined to release radioactivity. Second, increased trade in nuclear fuel and waste products provides more opportunities for sophisticated or government-sponsored terrorists to obtain the materials for a nuclear device. Third, dispersion of radioactive material throughout a city requires little sophistication and may not cause a large amount of much human and physical damage but would have a substantial psychological impact.

2.2.1 Attacks on Nuclear Power Plants

Attacks from the inside and outside are technically feasible, and the risk is growing. For example, Iran threatened on Tehran radio, "U.S. centers and nuclear reactors can be more vulnerable than the missile bases of the Islamic Republic of Iran." Approximately one-half of terrorism incidents occur in the Middle East, followed by Europe and Latin America, particularly Colombia (drug related) and Peru (the Marxist Sendero Luminoso). A substantial portion of international terrorist attacks are on U.S. interests. More than 300 metric tons of plutonium is generated annually, 45 of which is separated and suitable for weapons. In comparison, 6 kg of plutonium was used in the bomb dropped on Nagasaki. Plutonium separated from spent fuel is shipped to waste disposal sites. Some of this transport occurs on commercial airlines and by truck, train, and ship.

2.2.2 Diverting Nuclear Material: Nuclear and Dirty Bombs

Nuclear material is a threat if used to make a nuclear bomb or if spread by a conventional bomb. Nuclear bombs can be made from plutonium or highly enriched uranium. Plutonium is produced in plants that reprocess waste from conventional reactors. Natural uranium contains approximately 0.7% of the fissile isotope U-235. Conventional reactors use uranium slightly enriched to approximately 5% U-235. Low-enriched (to 20%) U-235 is used in research reactors. Highly enriched U-235 is a category higher than 20%, and weapons-grade U-235 is enriched to 90%.

India and Pakistan both produce enriched uranium; India also produces plutonium. These countries and others publicly embarked on programs to produce nuclear power, transferred the technology, and developed parallel programs to develop nuclear weapons. North Korea and Iran are openly pursuing this path.

2.2.3 Diverting Nuclear Waste: Poisoning Cities

Nuclear waste from breeder reactors includes plutonium-239, which is extremely toxic in minute quantities and has a half-life of 24,400 years. Low-level nuclear waste from conventional nuclear power plants includes cesium-137 and strontium-190, which

are less toxic and have half-lives of approximately 30 years. Twenty half-lives is equivalent to 600 years, the time during which the waste poses a risk.

Unsophisticated terrorists, willing to accept exposure and death, could use nuclear waste to poison a city. The level of health risk from low-level waste may not be as great as that from an attack on a nuclear plant or from a bomb. However, the terror of exposure to nuclear radiation, no matter how low the dose, would still be considerable.

2.3 Pollution

Pollution occurs at every link in the chain from production to consumption: exploration, extraction, transportation, conversion, processing, consumption, and disposal. Pollution travels through environmental media, including tropospheric and stratospheric air pollution, surface and ground freshwater pollution, ocean pollution, and soil pollution. Damages vary across energy sources and uses, including human mortality and morbidity, reduction of ecosystem services to the economy, immiserization of the human spirit, and diminution of the intrinsic value of nature.

The objective of estimating the benefits of lower emissions is to determine, in economic terms, what damage society avoids by selecting one set of policy options over others. To determine this, the potential damages of emissions must be calculated. This involves the following:

1. Estimate the changes in emissions for all pollutants affected by policy changes.
2. Estimate the changes in ambient pollutant concentrations in all relevant environmental media predicted to result from changes in emissions.
3. Estimate the pollution transformation and transport to target.
4. Establish the exposure–dose/response relationships between ambient changes and health effects and impacts on horticulture, visibility, buildings/structures, etc. For human mortality and morbidity, this requires the additional intervening step of estimating changes in exposure and dose in the damaged population that would result from changes in ambient concentrations, where dose depends on body size, age, activity, and location affecting intake and internal concentration.
5. Calculate the level of effects projected.
6. Attach dollar values to the estimated effects.
7. Attribute dollar values to the pollutants that caused the damages or to the policy changes that reduced the damages.

There are various methods for accomplishing each of these steps, but each must be completed to calculate the economic value of a policy change. Each step represents a complex calculation that often violates economic assumptions, such as differentiability and convexity. All of the links in this chain present challenges and pose obstacles to consensus, as the following examples of urban air pollution and global warming illustrate.

2.3.1 Estimating Reductions in Emissions from Policies

Step 1 requires an inventory of all the types of pollutants affected by the policy options and the appropriate counterfactual with which to compare the policy. An example is regional air pollution caused by electricity generation. A seemingly straightforward policy is to hold a Vickery auction for renewable electricity generation of a certain number of megawatts, such as the auction held by the California Public Utilities Commission in 1994 (and rescinded the following year on the grounds that deregulation was impending and markets would best decide how much and what type of additional capacity to build). An electricity dispatch model is needed to calculate the change in emissions.

Power plants are typically dispatched in increasing order of short-run marginal cost until the system demand is met. As demand changes from off-peak to peak periods, the least expensive of the remaining plants is dispatched to meet load, subject to technical constraints, spinning reserve requirements to meet fluctuations in load, and a reserve margin to cover maintenance outage and unscheduled outages. Some plants cannot be shut down during the weekend and some have a minimum operating level. Some are slow-start plants and others are quick-start plants. Technical constraints determine the type of load each can serve.

The addition of power plants to a system alters all of the previous calculations. By accounting for the location of the units, one can calculate emissions for each airshed. The calculation of emissions is a direct function of the calculated dispatch order. Consequently, the emission estimates depend on the reordering of dispatch due to a change in the combination of power plants.

2.3.2 *Estimating Changes in Pollution Concentration*

Step 2 requires the application of a model to project how emission changes will alter ambient pollution concentration. This really has two parts; the baseline pollution concentration is forecast in the absence of the policy, and the altered pollution concentration is forecast conditional on the change in emissions. Fig. 6 provides an example of a baseline forecast of ambient CO_2, showing historical levels during the past 600 million years and projecting what could happen during the next 400 years based on consumption of economically available fossil fuels.

The Intergovernmental Panel on Climate Change uses 120 years as the e-folding time atmospheric residence for a pulse increase of CO_2 sufficient for $RCO_2 = 2$. RCO_2 is the ratio of CO_2 to the ambient concentration prior to the industrial revolution in the late 1800s. The e-folding time is the number of years at which the biosphere and ocean removes 63% of the increase from pre-industrial levels; so that if CO_2 were to instantaneously double and no additional emissions occurred, after 120 years ambient CO_2 concentrations would still be 37% higher. For increases greater than $RCO_2 = 2$, the e-folding time increases dramatically. Carbon uptake by plants is limited because biological material decays, releasing carbon. For example, at the Triassic–Jurassic boundary approximately 200 million years ago, CO_2 was so high that climate effects led to the disappearance of most species. Almost all models used by economists make the assumption of linearity, seriously underestimating ambient concentrations for a tripling ($RCO_2 = 3$) or more. At three times preindustrial concentrations, the e-folding time would be between 380 and 700 years. To model the atmosphere beyond a doubling of CO_2, the models used by economists are misspecified. For a tripling ($RCO_2 = 3$) or larger, the natural science literature assumes that 50% of the emissions remain in the atmosphere, as presented in three alternative baseline forecasts in Fig. 6B.

Three economic models, combined with the previously discussed assumption, predict baseline RCO_2 increasing to 9, 11, and 16. To get a sense of the scale involved, Fig. 6A shows atmospheric concentrations measured in RCO_2 during the past 600 million years. The macro models forecast that economic forces will cause, within the next 325–350 years, the atmosphere to revert to an era never experienced by most plants and animals living today—levels not experienced by Earth at any time during the past 375 million years.

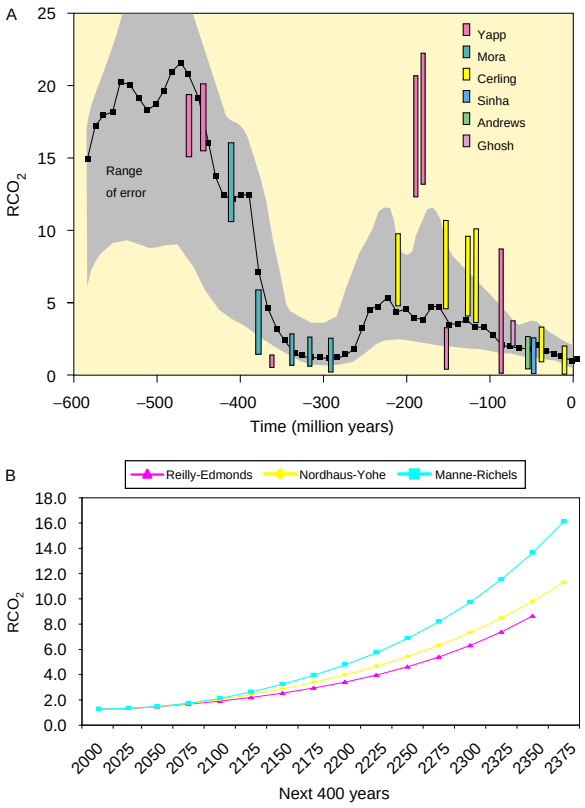


FIGURE 6 (A) Past atmospheric CO_2. Atmospheric CO_2 versus time for the Phanerozoic (past 550 million years). The parameter RCO_2 is defined as the ratio of the mass of CO_2 in the atmosphere at some time in the past to that at present (with a preindustrial value of 300 parts per million). The heavier line joining small squares represents the best estimate from GEOCARB II modeling, updated to have the effect of land plants on weathering introduced 380–350 million years ago. The shaded area represents the approximate range of error of the modeling based on sensitivity analysis. Vertical bars represent independent estimates of CO_2 levels based on the study of paleosols. Reprinted with permission from Berner, R. A. (1997). The rise of plants and their effect on weathering and atmospheric CO_2. *Science* **276**(25), 544–546. Copyright American Association for the Advancement of Science. (B) Baseline forecasts of RCO_2. The rate of emissions over time is based on three macroeconomic models that link forecasts of future economic activity with the potential use of fossil fuels. In descending order of the growth rate for CO_2 emissions, the macro models are labeled MR, NY, and RE. MR refers to Manne, A., and Richels, R. (1990). *Energy J.* **11**(2), 51–74. NY refers to Nordhaus, W., and Yohe, G. (1983). *In* "Changing Climate," pp. 87–153. National Research Council, National Academy Press, Washington, DC. RE refers to Reilly, J., Edmonds, J., Gardner, R., and Brenkert, A. (1987). *Energy J.* **8**(3), 1–29. Accounting for all fossil fuels, the economically available fossil fuels vary between 8 and 17 metric teratons. The MR forecast assumes 17 metric teratons, NY assumes 12, and RE assumes 8. Reproduced with permission from Hall, D. C. (2001). *In* "The Long-Term Economics of Climate Change: Beyond a Doubling of Greenhouse Gas Concentrations" (D. C. Hall and R. B. Howarth, Eds.). Elsevier, Amsterdam.

2.3.3 Estimating Pollution Transformation and Transport

To continue the example of global warming, ultraviolet sunlight passes through CO_2, warming the earth's surface. Infrared heat radiates back into the atmosphere, warming greenhouse gases (GHGs) such as CO_2 and radiating back to the earth (the greenhouse effect). As the concentration of GHGs rises, so does the mean global temperature. Coupled general circulation models of the atmosphere and oceans project increases in temperature and changes in rainfall frequency and intensity as GHGs rise.

Natural scientists expect the ambient concentration of GHGs to double within the next 100 years (most likely by 2060). Natural scientists use a doubling as a convenience in the thought experiment: If concentrations double, what will be the impact on global average temperature? The answer, from coupled general circulation models (GCMs), is 1.5–4.5°C. For easier cross-discipline comparison, economists use the same benchmark, $RCO_2 = 2$.

If we confine the baseline forecast to a doubling of concentration, it may be reasonable to assume that Earth's systems return to equilibrium, as is done in one famous model (DICE, RICE, and derivatives). Most economists assume that the ocean captures a significant amount of heat, with an e-folding time of 500 years, so increases in deep ocean temperature decay back toward the preindustrial revolution level outside the time frame of any model. This specification is inconsistent with findings of ocean surface water temperatures. More problematic for analysis of $RCO_2 > 2$, the Atlantic Ocean thermohaline circulation may end, resulting in a new stable equilibrium for ocean and atmospheric circulation.

Figure 7 presents baseline forecasts of mean global temperature from a model that specifies an ocean thermal lag, where one-half of the radiative forcing is stored in the ocean for 50 years and then released. The three baseline forecasts show how global temperature forecasts depend on key assumptions regarding the sensitivity of the climate to increases in CO_2 and the amount of economically available fossil fuels. The ocean thermal lag is consistent with findings that from 1948 to 1998, the heat content of the top 3000 m of the oceans increased by 2×10^{23} J, corresponding to a warming rate of 0.3 w/m^2 or approximately one-half of the warming predicted by GCMs for the past century.

Based on a comparison among three GCMs, for a doubling of CO_2 equivalent gases, the models predict 12% greater warming for the United States than the global mean because the United States is at a high latitude. As the earth warms, the largest temperature changes will occur near the poles.

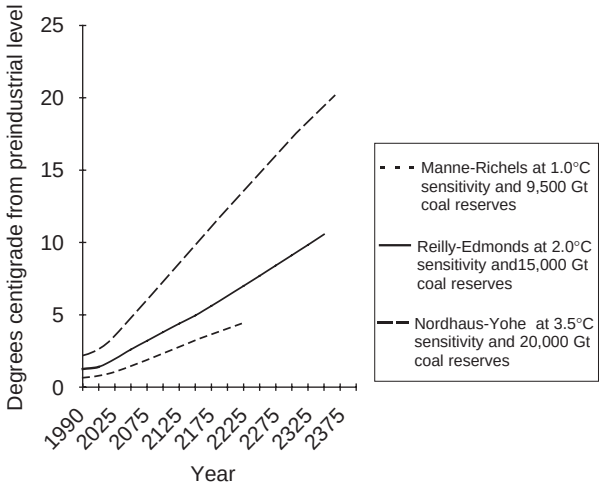


FIGURE 7 Mean global temperature increase.

2.3.4 Establishing Exposure–Dose/Response Relationships

Models of exposure–dose/effect relationships for air pollution are increasingly complex. Much of the past economics literature simply correlates air pollution concentrations with morbidity and mortality. One recent model calculates dose taking into account indoor, outdoor, and in-transit exposure; activity level that increases respiration; and body mass that changes with age. Exposure is measured in the REHEX model by time-weighted and location-specific ambient concentrations relative to indoor, in-transit, and outdoor activity for each age group and work status. Doses are integrated measures of pollutant intake. REHEX translates exposure into dose by accounting for activity and age. Dose is dependent on activity (e.g., work type) and age (children have higher metabolism and elderly people have lower metabolism).

Individual ill effects are identified in the health effects literature, with dose–response estimates that vary by age and health status. Effects range from eye irritation to hospitalization and death.

2.3.5 Calculating the Level of Effects

For some examples, such as climate change, the level of effects depends on the time frame for the analysis. Moderate climate change, based on a doubling of GHG concentrations, benefits midlatitude and upper latitude countries with milder weather; milder weather benefits agriculture, forestry, shipping, and transport; milder weather lowers heating bills and makes surroundings more conducive to human activity.

Damage to midlatitude countries includes more frequent and severe droughts and storms, hurricanes, cyclones, rapid snowmelt, and flooding; increased costs for water transport, storage, and flood control; migration of pests and increases in agricultural pest damage and control costs; migration, spread, and ill effects of disease and increases in health care costs; higher cooling bills; damage to coastal cities; reduced aquaculture yields; damage to coastal ecosystems and wetlands; and beach erosion. Damages to equatorial countries will be severe, including immersion of island nations and coastal properties, river delta flooding and damage to wetlands, damage to agriculture, loss of life, the destruction of coral reefs worldwide and the fisheries and tourism dependent on them, and severe droughts and floods.

Tripling or quadrupling of GHG concentrations would have severe consequences for the global economy, human health, and just about every aspect of human endeavor. Frozen deposits of methane in ocean sediments could thaw, instantaneously releasing massive amounts of GHGs into the atmosphere. There may be several global equilibria of GHG concentrations in the atmosphere and oceans. Tripling or quadrupling GHG concentrations could cause Earth's climate to shift to a new equilibrium and to never return to the climate of today. A tripling of GHG concentrations within 150 years is consistent with several models of economic activity and concomitant fossil fuel consumption. The range of uncertainty for a doubling of GHG concentration is between 1.5 and 4.5 degrees warming of mean global temperature, but we are ignorant of the range of temperature increase if GHG concentrations were to triple, quadruple or quintuple. We have modeled the range of temperature increase that would result would result in a shut-down of the Atlantic Ocean currents that keep the climate mild in northern Europe, but we are ignorant of the GHG concentration necessary to warm intermediate ocean water and release catastrophic quantities of methane to the atmosphere.

2.3.6 Attaching Dollar Values

Valuing the adverse effects associated with energy involves estimation of the value people place on avoiding the identified effects. Many of the effects are nonmarket goods, such as damage to human health and loss of life. Dollar values for nonmarket goods can be derived indirectly from hedonic studies or directly through surveys using contingent valuation methods or their derivatives. Hedonic studies use market data to obtain values such as differences in wages to measure wage premiums paid in exchange for higher risk jobs. From these studies, economists calculate the value of statistical life (VSL). Contingent valuation surveys ask respondents how much they are willing to pay to reduce risks or avoid ill effects.

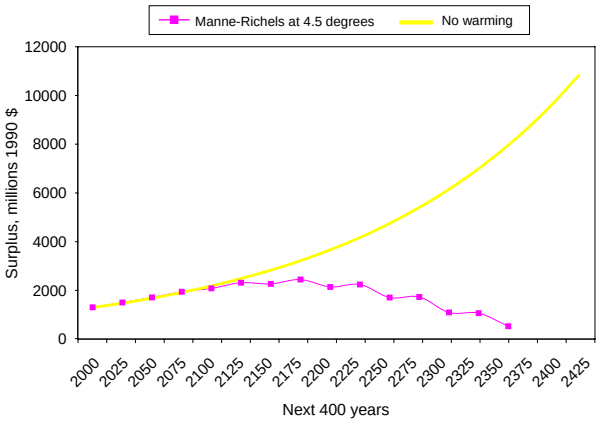


FIGURE 8 Agricultural surplus for wet climate.

VSL studies have recently been used to estimate the value of one life year. Assuming that the elderly are primarily at risk means that premature death from particulate matter (PM) is heavily discounted. When these biased estimates are corrected by including the risk of death to children, the estimated value of reducing particulates will increase, unless of course we use income stream over a lifetime and discount it. Then, the value of a child's life is much smaller than that of an adult at a prime earning age.

Valuation also requires the specification of a discount rate. With a rate greater than 3% (which is the minimum that many economists use), costs and benefits 100 years or more into the future are discounted to virtually zero in today's dollars and thus have virtually no impact on a benefit–cost calculation. For example, Fig. 8 shows an estimate of the time path of U.S. agricultural producer and consumer surplus in a scenario of climate change in which initially there are minor benefits from warming but ultimately agriculture in the United States collapses. With a discount rate of 5%, the present value of warming is greater than that of avoiding warming because the early minor benefits count more than the eventual collapse. If the discount rate is set to 1%, however, then economic valuation of the future years has significant weight to conclude in favor of avoiding collapse.

2.3.7 External Costs of Pollutants and Benefits of Policies

Step 6 requires knowledge of the precise relationships between ambient levels of all pollutants that

derive from the initial emissions and the ability to separate the contribution of each pollutant to ambient changes.

A study estimated the value of meeting federal ozone and PM10 standards in the South Coast Air Basin of California. For this purpose, steps 2–6 were completed. Step 1 was omitted because the question was what the benefits would be of meeting the standards, so the assumption was that emissions would be reduced enough to meet that objective. Step 7 was also omitted because there is no scientific or economic foundation for allocating inputs among joint products or for attributing products to nondepletable inputs. Another study based on the same methodology followed steps 2–5 for the purpose of projecting the pattern of exposure as emissions are reduced. These studies recognized the fundamental difficulties of evaluating a change in emissions relative to eventual damages, including the difficulties presented by the need to estimate air quality changes from emissions changes and to allocate damages back to emissions, and by the vast gap between what is known about the health effects of ozone and other pollutants and what can be quantified and valued in economic terms.

3. ESTIMATES OF EXTERNAL ENERGY COSTS

To understand the magnitude of negative external costs of energy, it is useful to aggregate and average over energy sources, comparing fossil fuels (coal, crude oil, and natural gas) with nuclear power, with solar power separately, and with solar power combined with natural gas to be operated as a baseload power plant. Aggregation in this sense can be easily misinterpreted as a set of estimated pollution taxes that would result in economic efficiency. This is not the case, as explained previously.

In Table I, the estimated externality values are in Year 2000 dollars, updated from previous studies using the implicit price deflator. The national security costs and carbon emissions can be directly attributable to specific energy sources. The national security costs of oil are the addition of national defense expenditures attributable to importing oil and the cost of the strategic petroleum reserve. Note that the national security cost of oil has not been updated to reflect events since 1990, so this estimate is low. The estimates for coal and nuclear power are specific to those energy sources. Numerous externalities have been omitted, particularly for coal, including water pollution, waste disposal, and wholesale destruction of ecosystems at mining sites. The air pollution damages from natural gas and oil, however, represent a sum total that has been attributed to the energy sources linearly, based on the proportions of specific air pollutants from each source.

Table I does not provide optimal taxes that will internalize the externalities. The economic theory of externalities, depicted in Fig. 1, shows a tax on output that coincides with the units of measurement in Table I, whereas Fig. 2 shows a tax on externalities that would have units of measurement in dollars per pollutant. Even if the values were reported in dollars per pollutant, such calculations are illustrative only because of nonconvexities (e.g., NO_x is a nondepletable catalytic pollutant that converts into both ozone and particulate) and the inability of models to capture subtle effects that cascade between natural and social systems.

TABLE I

External Costs from Pollution and National Security[a]

Energy source	Private cost market price 2000 $	External marginal cost					Social marginal cost
		Subsidies	Carbon emissions	Air pollution	Other externalities	National security	
Natural gas ($/mmbtu)	5.80	—	0.24	0.39	—	0	6.46
Crude oil ($/bbl)	29.50	—	1.92	14.88	—	12.70	59.20
Coal ($/ton)	31.70	—	9.05	17.98	—	0	58.73
Nuclear($/MWh)	6.75	1.80	0	0	1.80	0.12	8.68
Solar thermal ($/MWh)	9.51	0	0	0	—	0	9.51
Baseload solar and gas ($MWh)	5.80	0	0.14	0.22	—	0	6.16

[a] Source: Calculated from Hall (1990, 1992a,b) using the implicit price deflator for the Year 2000 relative to previous estimates.

Table I demonstrates that external costs of coal and oil are at least double the price; natural gas and nuclear power have much lower external costs than coal and oil; and the major problem with nuclear power is the private cost of building, maintaining, upgrading, and operating power plants and waste disposal. It is also clear from Table I that in the southwest United States thermal solar power with natural gas backup as a baseload unit has a lower social cost than nuclear power.

4. POLICY CHOICES

There are five aspects to good economic policy analysis. One is to consider all aspects of economic efficiency that policies affect. The second is to recognize when static economic efficiency arguments are not applicable due to misspecification of the alternatives, nonconvexities, nondepletable pollutants, uncertainty, timescale, and ignorance. The third is to incorporate the dynamic efficiency criterion in the context of state dependency of technology. The fourth is to consider policy strategies that increase flexibility and diversity to reduce macroeconomic risk. Finally, the fifth is to analyze wealth distribution effects of policies that determine political viability.

4.1 Evaluating All Market Failures That Policy Choices Affect

Externalities include benefits and damage from climate change. Many policies that would reduce emissions of GHGs would also reduce other forms of pollution caused by energy production and consumption, and some policies would also alter national security risks. Similarly, many policies that reduce regional air pollution also reduce water pollution, national security risks, and externalities from global warming. Policies to encourage investment in energy efficiency also address other forms of market failure, such as asymmetric information. Policies with joint benefits require benefit estimation that is inclusive and not exclusive; this is obvious, but it is rarely done in the economics literature.

Standard incentive-based policies include pollution taxes; tradable permits; hybrid combinations of long-term tradable permits with repeated, short-term permit sales; and permits tied to direct regulation of emission limits. These policies are evaluated with the static economic efficiency criterion but do not address some of the market failures associated with underinvestment in energy-efficiency technologies.

Studies show internal rates of return $>20\%$ on commercial applications of energy-efficient lighting ballasts, whereas interest rates are much lower. These market imperfections—adverse selection, high transaction costs, principle agent problems, and intrafirm information flows—result in economic inefficiency. Specifically, asymmetric information exists between ballast sellers and potential buyers for improvement in quality, resulting in a lemons market and adversely selecting the lower quality (at the lower price). Because separate meters do not exist for each use of electricity, and a single meter is used to calculate electricity bills, businesses have a high cost of evaluating the efficiency of appliances. One principle agent problem occurs when commercial buildings are leased, the owner is the agent for the lessee (the principle) in the purchase of fluorescent lighting, and asymmetric information exists between the principle and the agent regarding the replacement costs. Another principle agent problem occurs when the renter is not separately metered, but rather the owner pays the electric bill and the renter and purchases the lower priced and inefficient appliance.

There are three additional sources of market failure for energy efficiency investments. First, electricity customers face higher borrowing rates for energy efficiency investments than the interest rates faced by electricity producers to invest in additional supply. Second, retail rate design for electricity and natural gas is based upon short-run marginal costs rather than long-run marginal cost rate, so that rates encourage investment in energy efficiency only up to the point that the capital and operating cost of the energy efficiency investment is cheaper than the operating cost of energy production. Third, there are negative externalities in energy production that are avoided by energy conservation. If the borrowing costs are lower for electricity producers than electricity efficiency investments by electricity consumers, or if the electricity is priced equal to short-run marginal cost of production, then at the margin electricity consumers will not invest in efficiency even when it is less expensive than the marginal cost of building and operating additional power plants.

4.2 Policy Options and the Static Efficiency Criterion

Policy options include direct regulation (or pejoratively, command and control regulation) and incen-

tive-based approaches, including pollution taxes and tradable permits allocated to polluters. Direct regulation includes requirements for pollution permits that polluters install the best available control technology when rebuilding power plants. Pollution taxes have fallen out of political favor, and may have constitutional challenges because the optimal pollution tax may vary over time and each tax adjustment could legally require congressional approval. Tradable permits give the holder the right to pollute a certain amount, to buy permits if emissions increase, and to sell permits if emissions decrease. One version of tradable permits is to lease the permit at a market auction. The so-called property rights solution is to allocate tradable permits in accordance with past levels of pollution, in essence appropriating common property and giving the property right to the polluter, without a corresponding property tax or any recognition of the economic rent.

If the property right to cause the externality resides with the party causing the damage, then the WTP is the correct marginal benefit curve. If the property right resides with the damaged party, the WTA is the marginal benefit curve (see Fig. 2). Solutions proposed by economists imply different allocations of property rights. Pollution taxes retain the property right for the public. Tradable permits, whose allocation is based on past levels of pollution, allocate the property right to the polluter. If tradable permits are leased at auctioned prices, the public retains the property right. When the U.S. Environmental Protection Agency changed direct regulations so that dirty power plants no longer had to be retrofitted to reduce pollution, the property right shifted from the pollutee to the polluter, and the marginal benefit curve was reduced by fiat, from the WTA to the WTP in Fig. 2. There is no unique optimal level of externality abatement without assuming who should own the property right, a normative issue.

Modern economic analysis favors the property rights solution to negative externalities. This selection is based on the criterion of static economic efficiency applied to Fig. 2. The solution requires setting the permitted amount of externalities equal to $N_C - N_P$, and polluters will trade permits if they can reduce pollution at a marginal cost less than the price of the permit. The result will be N_P abatement and the equilibrium price of the permit equal to the optimal pollution tax, T_P. This solution avoids the cost of direct regulation inherent in having regulators determine the cost of technology—a job best left to the market.

The property rights solution is an obvious misnomer since the other solutions—taxes, direct regulation, and auctioned leases of tradable permits—all have property rights retained by the public. Because the tradable permits are not subject to property taxes, it is also a misnomer to refer to the right to pollute as private property. An alternative is to lease tradable permits, retaining the property right for the public; this is referred to as the common property solution, favored by a small but growing group of influential economists. The common property solution—leasing tradable permits to pollute $N_C - N_A$—results in N_A abatement and an equilibrium price equal to T_A in Fig. 2.

Both the property rights tradable permits and the common property tradable permits options do not result in efficiency unless the amount of permitted pollution depreciates over time so that abatement increases in accordance with the analysis in Fig. 3, with the marginal benefit of abatement increasing over time due to the public good nature of abatement and population growth, and the increase in the pollution concentration shifting the intercept of the marginal benefit.

More problematic for the tradable permit approaches is that they do not acknowledge the joint pollutants, nonconvexities, and subtle cascading effects among natural and economic systems that render marginal benefit and marginal cost curves irrelevant.

4.3 Dynamic Efficiency and State Dependence of Technology

Some new technologies show promise for reducing externalities from energy. If CO_2 can be economically captured, it could be stored in several ways. One idea is to release CO_2 at the bottom of the oceans, but ocean currents could recirculate it. A second approach is to inject CO_2 into depleted oil and gas wells, raising pressure to capture remaining oil and natural gas and leaving the CO_2 in an impermeable reservoir; Ontario Power Generation has purchased the rights to GHG emission credits of 9 million tons of carbon and plans to inject the gas into old oil fields in the United States. Another idea is to inject CO_2 into coal mines, displacing and capturing methane gas. A Norwegian oil firm is injecting CO_2 into a saline aquifer under the North Sea, where the carbon is sequestered by reacting with the salt.

The integrated gasifier combined cycle (IGCC) is a technology that might lead to economically capturing CO_2 by burning pulverized coal in the presence of steam, generating CO and hydrogen gas. The hydrogen can be used to operate fuel cells or burned

to generate steam for electricity, and the CO can be catalytically converted to CO_2 and captured.

Fuel cells can chemically generate electricity from either hydrogen gas or methane gas. Some buses, trucks, and smaller vehicles are powered by fuel cells. Fuel cells can also be strategically placed to stabilize a transmission and distribution grid, avoiding urban air pollution emissions.

Infrastructure investments for energy supply and transportation—dams, power plants, pipelines, roads, and transportation networks—all have lifetimes in decades, even centuries. We can identify irreversible environmental damage of staggering magnitude lasting centuries, but natural science is unable to pinpoint how much pollution the earth's systems can sustain prior to triggering cataclysmic events. What is certain is that eventually we need to redirect the path of technological change.

In the 1960s, economists argued correctly that as rising market prices reflected scarcity of fossil fuels, the result would be substitution, new discoveries of fossil fuels, and technological change that reduced the cost of discovering, extracting, transporting, and converting fossil fuels to useful purposes. It is the success of these arguments that makes it clear that policies that encourage technologies such as the fuel cell, IGCC, solar thermal power, solar electric power, biomass, geothermal energy, ocean thermal energy conversion, wind power, and other substitutes for polluting fossil fuels will also lead to further improvements in fossil fuel technology because owners of fossil fuels have extraordinary wealth that can only be protected by research and development that furthers fossil fuel supply and use. Economic forces and institutional and political infrastructure lead to technological path dependency.

Direct regulation has an advantage over tradable pollution permits when evaluated by the dynamic efficiency criterion for selecting among policy options. Direct regulation can force technological change and can do so quite flexibly. An example is to require a fraction of sales to be zero- or ultra-low-emission vehicles, allowing vehicle manufacturers to trade the requirement among each other. Another example is to require a fraction of the power purchased by a utility to be generated by zero or lower polluting technologies, called a "set-aside" for renewable energy; this can be efficiently accomplished with a Vickery auction —such as the 1994 auction held by the California Public Utilities Commission and discussed above. Energy efficiency standards for building design, electric motors, appliances, heating, ventilation, refrigeration, and air-conditioning are also examples. Technology-forcing policies include pricing and tax subsidies of renewable electricity technologies, and renewable energy tax credits; rebates, low-interest loans, and subsidized electricity rates for efficient appliance purchases; and mileage standards for vehicles (CAFE standards).

4.4 Flexibility and Diversity to Reduce Risk

Policies that reduce macroeconomic risk can be designed by applying the concepts of flexibility and diversity to hedge against adverse outcomes. Consider that an electric utility sets spinning reserve requirements at least as large as the amount supplied by the largest power plant in operation. Grid operation derates transmission capacity to account for loop-flow feedbacks that would cause the entire grid to fail if the largest transmission line were to short. Backup capacity is always on call in the amount equal to the largest single link on which the system relies. Our macroeconomic performance has hinged on a single supplier of oil, Saudi Arabia, with no backup capacity except the strategic petroleum reserve.

Solar thermal with natural gas backup, wind power, and other renewable technologies provide both flexibility and diversity. The small size of renewable electricity generators, coupled with the speed with which they can be built, permits a close matching of supply and demand, reducing the need for substantial reserve margins for electricity supply. Technology-forcing policies for energy efficiency and alternative energy sources result in more available substitutes in the event that fossil fuel prices increase.

4.5 Political Viability and Wealth Redistribution

The argument against pollution taxes is that the size of wealth redistribution makes taxes politically unfeasible. This argument is simply wrong and has been proven so. The argument against marginal cost water rates was that it is politically unfeasible—witness the Tucson, Arizona, city council voted out en mass 1 year after implementing such rates in the 1970s. However, a two-part increasing block tariff for water rates, based on long-term marginal cost, is politically feasible for the city of Los Angeles. Similarly, a two-part pollution tax could be levied with a low or zero tax up to some amount, and a higher tax thereafter, providing three policy tools: the initial tax, the breakpoint, and the higher tax. Furthermore, pollu-

tion taxes could be designed to be revenue neutral, potentially leading to greater efficiency in energy use and in other sectors of the economy. It is easy to imagine pollution taxes offsetting tax credits for technology development, for example.

Other pollution-reducing policies meet similar resistance. Auctioning pollution permits would be more efficient than allocating them to existing polluters, the latter model being the most often followed in the United States. Any effort to implement such an auction scheme is met with effective political resistance. Ironically, allocation with no auction is a large-scale wealth-distribution scheme that favors polluters.

Wealth redistribution clearly affects the political feasibility of policy options. Dating back at least to "The Logic of Collective Action," economists have recognized that the interests of small, well-organized groups regularly outweigh the larger public interest. The argument against pollution taxes is often couched in terms of the regressive effect on the average family. This is a disingenuous means of diverting attention from the very substantial economic rents that fossil fuel producers and marketers currently collect while placing health and other environmental risks on the families.

5. CONCLUSIONS

Economic analysis is useful for informing policy choices but not capable of proving that any single policy choice is more efficient than another. To do so would require the analyst to have all the information about costs and benefits that precludes central planners from replacing markets in the allocation of resources. The laws of chemistry and physics result in violations of assumptions in the theory of externalities. Furthermore, economic efficiency as measured in benefit–cost analysis is a narrow measure of net human welfare that indicates nothing about the equity of one option compared to another. It is the most inclusive method economists have developed to date, but it still is not appropriate as the primary means of choosing among policies that have far-reaching implications for this and future generations.

This should not be taken as an argument to abandon economic analysis but rather to work closely with other social and natural scientists to improve our ability to more accurately, transparently, and fully represent and compare alternative energy futures.

SEE ALSO THE FOLLOWING ARTICLES

Climate Change and Energy, Overview • Cost–Benefit Analysis Applied to Energy • Electric Power Generation: Valuation of Environmental Costs • Geopolitics of Energy • Goods and Services: Energy Costs • Greenhouse Gas Abatement: Controversies in Cost Assessment • Motor Vehicle Use, Social Costs of

Further Reading

Broadman, H. G., and Hogan, W. W. (1988). Is an oil tariff justified? An American Debate: The numbers say yes. *Energy J.* **9**(3), 7–29.

Cifuentes, L., Borja-Aburto, V. H., Gouveia, N., Thurston, G., and Davis, D. L. (2001). Hidden health benefits of greenhouse gas mitigation. *Science* **293**, 1257–1259.

Daily, G. (ed.). (1997). "Natures Services: Societal dependence on natural ecosystems." Island Press, Washington, DC.

Hall, D. C. (ed.). (1990). Contemporary policy issues [Special edition]. *Social Private Costs Alternative Energy Technol.* **8**(3). [Reviewed in *Rev. Gen. Thermique* **350**, 1991]

Hall, D. C. (1992a). Social cost of CO_2 abatement from energy efficiency and solar power in the United States. *Environ. Resour. Econ.* **2**, 21–42.

Hall, D. C. (1992b). Oil and national security. *Energy Policy* **20**(11), 1089–1096.

Hall, D. C. (ed.). (1996). Advances in the Economics of Environmental Resources: Marginal Cost Rate Design and Wholesale Water Markets. JAI, Greenwich, CT. [Reviewed by G. L. Carriker (1998). *Am. J. Agric. Econ.* **80**(4).]

Hall, D. C. (1998). Albedo and vegetation demand-side management options for warm climates. *Ecol. Econ.* **24**, 31–45.

Hall, D. C., and Hall, J. V. (1997). Estimating the benefits of emissions reductions in complex atmospheres. *Int. J. Global Energy* **9**(4–6), 286–298.

Hall, D. C., and Howarth, R. B. (eds.). (2001). The long-term economics of climate change: Beyond a doubling of greenhouse gas concentrations. *Adv. Econ. Environ. Resour.* **3**.JAI

Hall, D. C., Win, M. S., and Hall, J. V. (1995). Air pollution impacts from demand side management. *Energy Int. J.* **20**, 27–35.

Hall, J., Winer, A., Kleinman, M., Lurmann, F., Brajer, V., and Colome, S. (1991). Valuing the health benefits of clean air. *Science* **255**(5046), 812–817.

Hamilton, J. D. (1996). This is what happened to the oil price–macroeconomy relationship. *J. Monetary Econ.* **38**, 215–220.

Hanemann, W. M. (1991). Willingness to pay and willingness to accept: How much can they differ? *Am. Econ. Rev.* **81**(3), 635–647.

Koomey, J. G., Sanstad, A. H., and Shown, L. J. (1996). Energy-efficient lighting: Market data, market imperfections, and policy success. *Contemp. Econ. Policy* **14**(3), 98–111.

Kothavala, Z., Oglesby, R. J., and Saltzman, B. (1999). Sensitivity to equilibrium surface temperature of CCM3 to systematic changes in atmospheric CO_2. *Geophys. Res. Lett.* **26**, 209.

Levitus, S., Antonov, J. I., Boyer, T. P., and Stephens, C. (2000). Warming of the world ocean. *Science* **287**(5461), 2225–2229.

McElwain, J. C., Beerling, D. J., and Woodward, F. I. (1999). Fossil plants and global warming at the Triassic–Jurassic boundary. *Science* **285,** 1386–1390.

Rosenfeld, A., Atkinson, C., Koomey, J., Meier, A., Mowris, R., and Price, L. (1993). A compilation of supply curves of conserved energy for U.S. buildings. *Contemp. Policy Issues* **11,** 45–61.

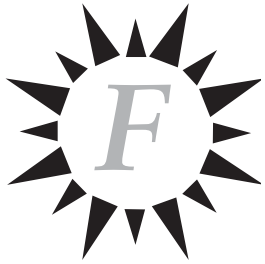
U.S. Environmental Protection Agency (1999). "Benefits and Costs of the Clean Air Act 1990–2020." U.S. Environmental Protection Agency, Washington, DC.

U.S. Environmental Protection Agency (2001). "Benefits and Costs of the Clean Air Act 1990–2020: Draft Analytical Plan for EPA's Second Prospective Analysis." Office of Policy Analysis and Review, Washington, DC.

Viscusi, W., Magat, W., Carlin, A., and Dryfus, M. (1994). Socially responsible energy pricing. *Energy J.* **15**(2), 23–42.

Walton, A., and Hall, D. (1990). Solar power. *Contemp. Policy Issues* **8**(3), 240–254.

Fire: A Socioecological and Historical Survey



JOHAN GOUDSBLOM
University of Amsterdam
Amsterdam, The Netherlands

1. The Original Domestication of Fire
2. Agrarianization
3. Industrialization
4. Symbolic and Religious Aspects
5. Conclusion: The Eclipse of Fire?

Glossary

agrarianization The formation and spread of a socio-ecological regime structured around the domestication of plants and animals.

anthroposphere That part of the biosphere that has been affected by humans.

domestication The "taming" and incorporation into human society of originally "wild" nonhuman sources of energy.

industrialization The formation and spread of a socio-ecological regime structured around the exploitation of fossil fuels.

The ability to handle fire is, along with language and the use of tools, a universal human attainment: no human society in historical times is known to have lived without it. The ability to handle fire is also exclusively human: whereas other animals use rudimentary signals and tools only humans have learned to handle fire. From the outset, mastery of fire has strengthened the position of humans in the biosphere. At the same time, it has deeply affected the relations among humans themselves. Appreciation of the control over fire is therefore essential for an understanding of both the external expansion and the concomitant long-term internal transformation of the anthroposphere. The control over fire has been essential in the emergence of agriculture and stock raising, as well as the rise of modern industry. Its implications, including both benefits and costs, are so far-reaching as to often remain unnoticed and taken for granted. This entry intends to convey the enormous importance of the fire regime for human history and its profound impact on the biosphere.

1. THE ORIGINAL DOMESTICATION OF FIRE

1.1 Origins

Human life, like all life, consists of matter and energy structured and directed by information. All life is part of an ecosystem, and all ecosystems together constitute the biosphere—the total configuration of living things interacting with each other and with nonliving things. Every form of life continuously affects, and is affected by, its ecosystem.

As participants in the biosphere, the early hominids and their only surviving species, *Homo sapiens,* have gradually strengthened their position—at first slowly and almost imperceptibly, later at an increasingly more rapid pace with ever more striking consequences. In the process, they engrossed increasingly more terrain and incorporated increasingly more nonhuman resources into their groups: first fire, then, much later, certain selected plants and animals, and, later again, fossil fuels. As they incorporated more energy and matter into their societies, those societies grew in size, strength, and productivity, while at the same time also becoming more complex, more vulnerable, and more destructive.

During the first stage in human history and "prehistory," known in archaeology as the Paleolithic or Old Stone Age, which lasted for thousands of millennia, the overall pace of social and cultural development was slow in comparison with later

 © 2004 Elsevier Inc. All rights reserved.

stages. Yet some momentous changes took place, with great consequences for the relationships between humans and the natural environment. Humans began not only to use but also to make tools of their own, and they learned to control fire. The combination of tools and fire enabled groups of humans to leave their original habitat, the savannas of East Africa, and to migrate into other parts of the world, penetrating first into remote corners of Eurasia and then also into Australia and the Americas. The Paleolithic can thus be seen as a long run up, which was later followed by an enormously accelerating sprint of which our present age is the latest episode. It was the scene of the incipient expansion of the human realm, the anthroposphere.

A basic trend in all human history, and certainly during its earliest phases (often designated as prehistory), has been the increasing differentiation between humans and all closely related animals in terms of their behavior, power, and general orientation or attitude—their habitus. Thanks to the flexibility acquired in the course of evolution, humans were able to learn a large repertory of new forms of behavior. Especially successful were those innovations in behavior that added to human power vis-à-vis other large animals, both predators and competitors. Transmitted by learning from generation to generation, those innovations entered into the human habitus and became second nature.

The primary condition for the process of differentiation in behavior, power, and habitus has been, and continues to be, the inborn human capacity for culture as manifested in technology, organization, and civilization, each of which represents the results of social learning. Indeed social learning is the crux of culture: gathering information and passing it on to others—at first in direct interaction, at a later stage in history through written texts and today by other audio-visual means as well. The stores of cumulated information (or "cultural capital") have enabled people to tap increasingly larger and more varied flows of matter and energy and to integrate those flows into their societies.

The domestication of fire was the first great human interference with natural processes. It had numerous far-reaching implications, stretching from the first hesitant beginnings to our contemporary fuel-driven economy. It demands our attention because it reveals something of the socioecological infrastructure of the contemporary world.

In several ways, the original fire regime may be seen as a paradigm for the socioecological regimes that were developed later. It presents a paradigm in a double sense. First, in actual practice the regime by which humans learned to extend their care for and control over fire could serve as a model for subsequent forms of care for and control over other forces in nonhuman nature such as plants and animals. Second, we may regard the domestication of fire as a model case in a more theoretical fashion, since it brings out the strong link between such apparently contradictory tendencies as increases in control and dependency, in robustness and vulnerability, and in the potential for production and for destruction.

1.2 First Impact

Fire, like all natural forces, has a history. Chemically fire is a process of highly accelerated oxidation of matter (fuel) induced by heat (ignition). Three conditions are therefore necessary for it to occur: oxygen, fuel, and heat. During the first eons in the history of the earth, at least two of these, oxygen and fuel, were absent. Oxygen did not become available until, after at least a billion years, life emerged. And it was only less than half a billion years ago, during the Devonian geological age, that life assumed the form of plants, providing matter suitable for burning. From then on, most places on earth with seasonally dry vegetation were regularly visited by fire, ignited on rare occasions by falling rocks, volcanic discharges or extraterrestrial impacts, but mostly by lightning.

Its domestication by humans opened an entirely new episode in the history of fire. Humans thoroughly altered the frequency and intensity of fires. They brought fire to regions of the planet where it seldom or never burned spontaneously, and they tried to banish it from places where without human interference it would have burned repeatedly. Thus, increasingly, "natural" fire receded and made way to human or, more precisely, anthropogenic fire.

Wherever humans migrated, they took their fire along. The presence of humans with fire deeply altered the landscape, including flora and fauna. The human impact is amply documented (though still controversial) for Australia—a continent that was colonized by humans rather late. Everywhere on the planet, areas such as rain forests, deserts, and the polar regions, which were not receptive to fire proved to be hard to penetrate for humans too.

Humans are the only species that has learned to manipulate fire. Control over fire has become a species monopoly, with an enormous impact on other species, both animals and plants. It provides us with

an excellent example of how new forms of behavior could change power balances—in this case, between humans and all other animals, ranging from primates to insects—and how shifts in power balances could engender changes in habitus, both among the humans who gained greater self-confidence from the presence of fire in their groups and among animals that might be bigger and stronger than humans but learned to respect and fear their agility with fire.

Control over fire, in addition to having become exclusively human, has also become universally human. We know of no human society of the past 100,000 years that lacked the skills needed to control fire.

The original domestication of fire was a dramatic transition. Palaeoanthropologists are still debating when exactly it took place. The estimates range from as many as 1.5 million to a mere 150,000 years ago. It is still an open and fascinating question whether the first steps to control over fire coincided with other changes in early human development.

In retrospect, the initial domestication of fire was an event of momentous import. A wild force of nature—blind, capricious, and hazardous—was now tended, cared for, and supplied with fuel. Our early ancestors went to all this trouble, not altruistically, but because it served them well. They put the potentially destructive and essentially purposeless force of fire to work for their own productive purposes. They managed to make fire regularly available. They no longer had to hunt for it, hoping to find somewhere the smoldering remains of a natural blaze; they made it a part (and even the center) of their own group, and they revered it as a symbol of eternal life.

The domestication of fire made humans less directly dependent on natural forces that continued to be beyond their control such as the alternation of day and night or the cycle of the seasons. It made the contrast between dark and light, between warm and cold, or between wet and dry more amenable to manipulation, and thus gave humans a greater margin of freedom from the grip of nature. It increased their power, defined as the capacity to influence the outcome of an interaction. Armed with fire, humans were able to open up impenetrable tracts of bush and to drive away animals much fiercer and stronger than they were. The gain in power made their lives more comfortable and secure. The possibilities of heating, lighting, and cooking all contributed to what we would now call a higher standard of living.

1.3 Long-Term Consequences

"Wherever primitive man had the opportunity to turn fire loose on a land, he seems to have done so from time immemorial." This statement by the American geographer Carl Sauer may sound exaggerated; but it still fails to convey the full impact that the domestication of fire has had, both on the larger biosphere and, within the biosphere, on human society itself.

The most immediate effect of the domestication of fire on the biosphere in general was an increase in the frequency with which fires occurred. Prior to its human mastery, fire was ignited mostly by lightning. From now on another source was added: even before they learned to make fire themselves, humans were able to preserve it in their hearths and to apply it wherever they saw fit. Consequently, as the number of anthropogenic fires increased, the proportion of natural fires diminished. The geologist Peter Westbroek has suggested, in an admittedly speculative manner, that the earliest human fire use may have affected the planetary atmosphere and caused some of the major climate changes in the Pleistocene. Substantive evidence indicating modification of the landscape by human foragers equipped with fire has been brought forward for Australia, where most of the indigenous forests were burned down in the millennia following the arrival of the first Aborigines.

From the very beginning, humans used fire in two basic forms: the hearth and the torch. The hearth was the original site at which a fire was kept, usually at a cave entrance where it could be protected against the rain and still receive some air circulation. Since it had to be tended and since fuel had to be brought to it, it served almost naturally as a center for group life, providing heat, light, and a common focus. From the hearth developed in the course of time a variety of fire containers such as crucibles, stoves, kilns and furnaces, and, in our day, the mobile engines of cars and airplanes.

The hearthlike uses of fire have always had two kinds of environmental side effects. First of all, fuel had to be supplied. As long as human communities were small and living in areas with abundant wood, this did not cause much of a problem. When greater numbers of people started living in large urban concentrations, however, the need for fuel became a strong contributing factor to deforestation over big areas and, in our own age, to depletion of fossil resources. The second side effect of hearthlike fire consists of its waste products: ashes and smoke. Although smoke might be useful in driving away

insects and other insidious creatures, it has always been considered mostly as a nuisance to be got rid of. As long as people lived in isolated caves or huts, this was relatively easy. Urbanization and industrialization have seriously aggravated the problem.

While the functions of the hearth were originally primarily turned inward, the torch was a more outward directed implement. It was used not only to provide light at night, but also by day to set fire to shrubs and grasses—an effective way to destroy obstacles for foraging and to rout animals, both predators and prey, and thus to extend the human domain. Torches undoubtedly contributed to deforestation: wood was burned wholesale, regardless of its possible value as timber or fuel. In the age of agriculture, the torch was used for slash and burn and other techniques for clearing land, and it served as the model for a whole array of fire weapons culminating in our own time in rocket-propelled missiles.

Surveying the entire trajectory of the human use of fire from its earliest beginnings, we can distinguish three stages. During the first stage, there were no groups possessing fire; there were only groups without fire. Then, there must have been a second stage, when there were both groups with fire and groups without fire. We do not know how long that stage lasted, nor do we know how often it may have recurred. All we know is that it has come to an end. It was a transitional stage, leading up to the stage in which humankind has now been living for thousands of generations: the stage when there are no longer any groups without fire. All human groups are groups with fire.

Although we lack empirical evidence for the first two stages, that very lack leaves us no choice but to accept the conclusion that societies with fire were in the long run obviously more fit to survive than societies without fire. If we then ask why it was that societies without fire disappeared, there seems to be only one plausible answer: because they had to coexist with societies with fire, and apparently in the long run such coexistence proved impossible.

This may sound like a dismal conclusion suggesting fierce contests ending in the elimination of the losers. If such contests did indeed take place, they have left no trace of empirical evidence; we only have the highly imaginative evocations of what might have happened in books and films like *The Quest for Fire,* directed by Jean-Jacques Annaud. However, we can also view the fact that possession of fire has become a universal attribute of all human societies as an important example of the general rule that changes in one human group lead to changes in related other groups. If group A had fire, and the neighboring group B did not, group B had a problem. It could either try to minimize contact with group A and perhaps move away, or do as group A had done and adopt a fire regime, which should not pose insurmountable difficulties as long as the capacity to learn from the others was sufficient. In the latter case, instead of a zero-sum elimination struggle there would have been what American freelance author and scientist Robert Wright calls a "nonzero" situation, with an outcome from which both parties benefited.

The rule that changes in one human group lead to changes in other related groups may sound like a rather tautological explanation for social change, but it is not. It is a highly generalized empirical observation, similar to an observation we can make about fire: fire generates fire and, in a similar fashion and more generally, change generates change, and social change generates social change.

Such is the nature of the dynamics of human society and culture. After the original domestication of fire, it was never humans alone who interacted with other human groups and with nonhuman nature. It was always humans with fire, equipped with fire and with the products of pyrotechniques: cooked food, pointed spears and arrows, earthenware, metal tools and weapons. Their presence put an end to humans without fire.

Another general conclusion to be drawn from these observations is the following: changes in climate and precipitation have never ceased to be important causes for humans to change their way of life. Humans are no different from other species in that they will always have to accommodate the basic conditions of earthly nature, such as the alteration of day and night or of monsoons and seasons. In the course of human history, however, in addition to these overridingly powerful extra-human conditions, conditions brought about by humans themselves have become increasingly more important to the extent that, in our contemporary world, humanity has become a major agent of ecological change.

1.4 Regimes

The domestication of fire meant that people tamed a strong and potentially destructive natural force and made it into a regularly available source of energy. In so doing they initiated changes in the natural environment, in their social arrangements, and in their personal lives. These three aspects (ecological,

sociological, and psychological) are all part of the changing human relationship with fire.

In its ecological aspects, the domestication of fire affected the relations between humans and the nonhuman world so deeply that we may call it the first great ecological transformation brought about by humans, which was followed much later by the second and third of such transformations, generally known as the agricultural and industrial revolutions and better characterized by the long-term processes of agrarianization and industrialization.

Each of the three transformations spelled the formation of a new socioecological regime: the fire regime, the agrarian regime, and the industrial regime, marked by, respectively, the utilization of fire and elementary tools, the rise and spread of agriculture and animal husbandry, and the rise and spread of large-scale modern industry. The later regimes have not made the earlier regimes obsolete; rather, they have absorbed them and, in so doing, transformed them. Each new regime brought an expansion of the anthroposphere within the biosphere.

Defining the three regimes jointly in similar terms is helpful in order to better understand each of them separately as well as in their interrelations. A common conceptual model invites and facilitates comparison. The comparison allows us to explain the sequence in the emergence of the regimes and to perceive not only their similarities and differences but also their interlocking.

2. *AGRARIANIZATION*

2.1 Continuities

The history of the past ten thousands years may be read as a series of events accompanying the process of agrarianization of humankind, a process in the course of which humanity extended the domain of agriculture and animal husbandry all over the world and in so doing made itself increasingly more dependent on this very mode of production.

The initial transition from gathering and hunting to agriculture and animal husbandry was not necessarily abrupt. A group that started to cultivate a few crops would not have to give up its older ways altogether. There would have been only few, if any, agrarian societies from which gathering and hunting disappeared completely at once. However, the proportion of products acquired in the older way inevitably diminished as agriculture and animal husbandry advanced.

From the very beginning, the process of agrarianization was linked closely to the domestication of fire. It is hard to imagine how people could have begun to cultivate plants and to domesticate animals had the art of handling fire not already been thoroughly familiar to them. For one thing, they needed a hearth fire to cook on. The first crops cultivated on any large scale were cereal grains such as wheat, rice, and maize, which, owing to their high nutritional value and their capacity to withstand storage for long periods, formed very appropriate staple foods for a human community; to serve this purpose, however, they had to be made more easily digestible with the help of fire.

A second and very different reason why the control of fire formed a precondition for agrarianization was the human predominance over all other mammals, which was grounded partly in the use of fire. The human species' monopoly over fire was so solidly established by the time agriculture began, and is today so easily taken for granted, that it is seldom given separate attention in this context. Yet it deserves mention. Their hegemony in the animal kingdom enabled people not only to bring certain species, such as goats and sheep, under direct control, but also—at least as important—to keep most of the remaining "wild" animals at a distance from their crops and herds.

Third, experience in controlling fire may have furthered plant and animal domestication in another, even more intangible way, which our distance in time makes it difficult to assess precisely but which we are also, for that very reason, likely to underestimate. The time-honored practice of handling fire could not have failed to prepare humans for the many chores involved in agriculture and animal husbandry. It taught them that extending care upon something nonhuman could be well worth the trouble and thus made it more acceptable to them to accommodate the strains of an agrarian life, full of self-imposed renunciation for the sake of a possible future yield.

The most immediately visible link between early agriculture and ancient use of fire lay in the custom of burning off land with an eye to food production. Of old, foraging peoples were wont to apply their torches in order to keep the land open for gathering and hunting. Even in recent times, those firing practices were continued in some parts of the world, as in Australia where the Aborigines' judicious use of fire in keeping their land open for kangaroos and humans has become known as "firestick farming," a term suggesting a form of proto-agrarianization.

2.2 Sequences

The rise of agriculture was in many respects remarkably similar to the domestication of fire. Again, humans added new sources of energy to their own, this time by adopting certain plants and animals and integrating them into human societies. Plants that were formerly wild now began to be cultivated, wild animals were tamed and used for food or other purposes such as traction, and all these species were made part of the human domain, of the anthroposphere, which thus increased in size and complexity.

The transition from foraging to agriculture did not automatically make people happier and healthier. Agrarian life brought new hardships. Diets became more monotonous. Sedentary life in villages increased the susceptibility to disease. It allowed for a rise in fertility, but it also caused higher mortality. The result was large numbers of children, many of whom never reached adulthood. Not surprisingly, research on skeletons reveals that after agrarianization, human lives tended to become briefer, and bodies shorter. It is equally understandable that the texts of the great religions of Eurasia all exhibit a nostalgia for the lost paradise of a pre-agrarian era.

The very nature of agrarian life foreclosed a return to foraging, however. There are a few known cases of such a return, but those are exceptions. Almost everywhere, with so many people, and so little land, the only option for the survival of an agrarian population was cultivation. "Work" was writ large over the agrarian world.

If, at a very early stage, the domestication of fire had made human groups more productive but also more vulnerable (because from now on they had to rely on their fire), the rise of agriculture had the same twofold effect. Being able to grow more food, human groups grew more populous and became more dependent on their crops, and thus more vulnerable to failure or loss of their harvests. As the expansion of agriculture and pastoralism left increasingly less land available for foraging, the opportunities to escape from this vicious circle dwindled.

The first stage in the process of agrarianization necessarily involved clearing the land, removing any existing vegetation that would compete with the planted crops. In many cases, the most efficient procedure to accomplish this was by means of fire. As long as land remained plentiful, continued recourse was often taken to fire in a system practiced throughout the world and known under various regional names that are usually subsumed in the standard literature under the label "shifting cultivation." Shifting cultivation implies that a piece of primary forest is first cleared by the method of slash and burn and is then used for one or more harvests of crops. When, after a while, crop nutrients in the soil become exhausted and undesired plants (weeds) begin to dominate, the farmers temporarily abandon the land and turn to an adjacent lot, burn the vegetation down, and bring it under cultivation, again until they find harvesting unrewarding. Eventually they return to their first plot, which by then has become secondary forest or secondary bush, and resume their activities of burning, planting, and harvesting there. The duration of the entire cycle may vary as to time and place, but the principle remains the same.

In many parts of the world, in the course of time, burning the land and letting it lie fallow for a number of years or seasons ceased to be a regular practice and was replaced by more intensive methods of working the soil, requiring greater investments of labor but yielding a larger output per acre and thus making it possible to feed more mouths. The most common means of accomplishing higher yields by harder work were irrigation and ploughing.

2.3 Fire in Settled Agrarian Societies

In all agrarian societies with permanent settlements, new uses of fire and new attitudes toward fire developed. During the long first stage of human fire use, the main concern always was to keep the communal fire burning. From now on, however, the main concern became almost the opposite: to keep the fires that were lit for numerous purposes from spreading and from running out of control.

The uses of fire became increasingly more varied. Specialized pyrotechnic crafts emerged, such as blacksmiths and potters. Among the growing urban populations, fire was regarded with greater anxiety, for several reasons: with the proliferation of fires, the risks of conflagrations increased, and with the accumulation of property, people had more to loose.

Of course, it was fire as a natural force with its self-generating destructiveness that was feared. But more and more, this natural force manifested itself almost exclusively in the guise of anthropogenic fire. With all the fires burning in a city, one moment of carelessness might cause a conflagration. People had to rely on other people's caution. They had to oppress attempts at deliberate fire setting. And they had to reckon with the very worst danger: the organized form of murder and arson known as war.

A common problem in all advanced agrarian societies was the prevention of uncontrolled fire in

cities. This problem may have been less urgent where the major construction material was stone or brick; but throughout the agrarian era we find major cities in which most houses were built from timber. Rome, Constantinople, Moscow, London, and Delhi as well as the capitals of China, Korea and Japan all suffered conflagrations; all of them faced problems of fire prevention.

Today we may tend to conceive of problems of fire prevention primarily in terms of technology: of building materials and technical equipment. However, the problems were (and still are) at least as much civilizational, or social. The crucial issue was human behavior, care, consideration. According to a well-known story, a temple in ancient Greece burned down because the servant was careless; similar events must have occurred all over the world. Everywhere it was the human factor that counted most of all.

This was also reflected in the presence or absence of fire brigades. Today, it may seem self-evident that a city should have a fire brigade. However, the power relations in earlier societies could be such that the authorities would forbid the organization of fire brigades by citizens. A dramatic example of a ban put on urban fire brigades is provided by two letters that were exchanged in the second century CE between the governor of a province in the eastern part of the Roman empire, Pliny the Younger, and his emperor, Trajan. In the first letter, Pliny informed the emperor that the capital city Nicomedia was completely ruined by a fire; people had watched the blaze passively, unable to do anything. He therefore asked permission to establish a fire brigade of citizens in order to prevent such disasters in the future. The emperor denied the request bluntly, for—as Pliny should have known—such an organization of citizens could easily turn into a club with subversive political activities.

It is still unknown to what degree Trajan's reply was typical of rulers in agrarian empires. The scattered evidence provided by local histories suggests that the situation was very different in the towns of medieval Europe, where the members of various guilds were required to take a share in fighting fires. This issue is a promising subject for comparative historical research.

3. INDUSTRIALIZATION

3.1 The Industrial Regime

While the basic theme of human history over the past 10,000 years has been the agrarianization of the world, the history of the past 250 years has unfolded under the aegis of industrialization. In the process, the anthroposphere has become one global constellation, extending all over the planet, and its impact on the biosphere has become more and more intense.

In line with the approach to the domestication of fire and agrarianization, industrialization can be defined as the formation of a socioecological regime, structured around a new source of energy: fossil fuel—first in the form of coal, later also of oil and gas. The nature of this new energy source has made the industrial regime different in a basic way from the earlier socioecological regimes. Unlike plants and animals, and even wood, fossil fuel is not directly connected to the continuous flow of solar energy. It is a residue of solar energy from a remote past contained in geological formations. The energy stocks are not diffuse like sunlight but concentrated in particular locations from where they can be extracted through concerted human effort. They have two seemingly contradictory properties: they are abundant and finite.

The abundance is indeed great. Coal, oil, and gas represent the remains of the nonoxidized biomass—in other words, unburned fuel—of over 3 billion years. When people began exploiting those enormous reserves they entered as it were, in the words of the environmental historian Rolf Peter Sieferle, a "subterranean forest" of inconceivably large dimensions, which, moreover, in the course of 250 years of exploitation and exploration proved to contain far more riches than was originally expected.

Yet no matter how large, the hidden stocks are also finite. In contrast to plants, which are the direct autotrophic products of the photosynthetic conversion of solar energy, and to animals, which are the heterotrophic consumers of plant and animal food, geological stocks of fossil fuel do not partake in any living metabolism. They are incapable of growth or reproduction, irreplaceable, unrenewable and, as is becoming more and more evident, their use generates combustion products that enter the biosphere.

3.2 Origins and Preconditions

The genesis of industrialization raises a whole array of intriguing questions concerning its preconditions and its immediate causes as well as its functions and side effects. These are in principle the same questions that can be raised about the domestication of fire and the origins of agriculture. Because industrialization began only recently, we have the benefit of far more and far more precise empirical evidence; never-

theless, the problems remain puzzling. Any answer that comes to mind can only be tentative and subject to caveats.

One thing is certain: industrialization was not triggered by a major change in climate. At most, its beginnings more or less coincided with the end of the most recent secular dip in temperature in Europe, from 1550 to after 1700, known as the little Ice Age; but no relationship with this climatological event was evident. Industrialization was a completely anthropogenic transformation, brought about by humans in societies that were fully equipped with fire and agriculture.

The beginning of industrialization heralded a new era in human history. Yet it was clearly a continuation of the fire regime, for every industrial process rested on controlled use of fire. Industrialization also presupposed agrarianization: it could never have started without a strong agrarian basis that was able to feed numerous people who were not engaged in agriculture themselves and who could therefore be recruited for work in industry and trade. From the outset, the industrial world coexisted and coevolved with the agrarian world, in close interdependence and symbiosis, culminating in our time in forms of industrial farming or agroindustry that lean as heavily on fossil fuel as any other form of industrial production.

3.3 Further Parallels

Just as agrarianization must have begun in small farming enclaves, carved out in an environment that continued to be the domain of foragers, industrialization started with single steam-powered factories—often called mills as if they still were driven by wind or water—standing apart in the agrarian landscape. To sketch their initially semi-isolated position, Sieferle's image of industrial archipelagos is appropriate.

From the very start, however, even as islands in an agrarian landscape, the factories had an ecological impact stretching beyond the land on which they stood. To begin with, considerable amounts of matter and energy went into building them. The usual construction material was partly timber, but to a larger extent brick, the manufacturing of which involved huge quantities of fuel. The engines were made largely of metal, also in a highly fuel-intensive fashion. Then, brick and timber were needed for the houses to accommodate the workers. So even before the factories began to operate, their tentacles were already reaching into the environment.

Once in operation, the factories had to be supplied with a continuous input of material, such as iron and cotton to be processed and of fuel to keep the engines going. The need for fuel explains the location of the early industrial plants: close to coal mines and, in the case of heavy industry, sites of iron ore deposits. Of course, a nearby sea or river port facilitated transportation; failing that, canals were built, and later railways, connecting the islands of industrial production with the areas where the raw materials were found and with the markets where the products were sold. In 19th-century Britain, an ever more extensive and intricate network of canals and railways was formed through which the industrial regime spread its reach over the entire country.

3.4 Industrialization as Intensified Land Use

Our earliest human ancestors treated the land they lived in as essentially all of the same nature—a territory for both collecting and hunting, for foraging food as well as fuel. In agrarian societies, most of the land was parceled up into three distinct parts with distinct functions: fields or arable planted with crops; pastures or meadows where cattle and horses grazed; and heath and wood land providing fodder for pigs and sheep and, even more important, fuel and timber. All three areas were controlled by humans, who were able to make the land thus divided up more productive for their own specific purposes, but who also found themselves faced with a continuously decreasing wild acreage.

Industrialization offered an escape from this predicament. The exploitation of coal was a highly intensified form of land use, which relieved the pressure to use large areas extensively for growing wood for fuel. The virtual expansion of land attained through the use of coal helps to explain the unstoppable advance of industrialization, once it was underway. Like the agrarian regime, the industrial regime made offers that were disagreeable in many ways to a great many people, but in the long run for all of them impossible to ignore or to refuse. The huge gains in energy could be transformed into economic, political, and military advantages that proved to be irresistible. All over the world, societies without fossil energy industry made way for societies with fossil energy industry, a process which could not fail to have profound consequences for the biosphere.

In our own age, each year more than 100,000 square kilometers of remaining wild land in Africa, Asia, and South America are subjected to human

exploitation. One of the causes is the need for fuel in the Third World; in many densely populated areas, fuel for the fire to cook on has become as proverbially scarce as the food to be cooked. In addition, there is a rising demand for timber and pulp from the most highly industrialized countries. The greatest threat to the continued existence of natural forests does not lie in felling the trees, however, but in the indiscriminate burning down of entire tracts in order to clear the ground for raising crops and cattle for commerce. The world economy generates a rising demand for products of tropical agriculture. To meet this demand, garden plots are converted into plantations, and the small farmers have to leave. At the same time, population continues to grow, so that pressure on land becomes even greater. As a result, the contradictory situation arises that in a world suffering from severe shortage of wood, each year many tens of thousands of hectares of forest are set on fire.

In Australia, exceptionally, anthropogenic fire has diminished since the establishment of a European colonial regime. From the time of their first arrival, the British had a sensation of the Aborigines going about burning the land incessantly. The British took every measure they could to repress these burning practices, with such success that rainforest has been regaining ground again in Australia over the past 200 years. An unintended side effect has been the increasing incidence of large wildfires during dry seasons, for the many fires of limited scope lit by the Aborigines prevented the vast fires that now, when they are raging, constitute a serious threat to the metropolitan suburbs.

3.5 Current Uses and Manifestations of Fire

The 20th and 21st centuries have seen a common shift in preference for energy sources all over the world. Wherever electricity is now available, it is the favorite means of generating light, and combustion engines are, along with electricity, the favorite means of generating motion.

The shift from steam engines to electricity and internal combustion is reflected in the changed industrial landscape, which in the Western world is no longer dominated by endless rows of factory chimney stacks. Yet the methods of production in modern industry, and in agriculture as well, continue to be highly fuel-intensive. Most of the energy consumed, including most electricity, is derived from the fossil fuels coal, oil, and gas. Combustion processes thus still play a central role, but they are relegated to special containers so that most people are not directly confronted with them. Soot, smoke, and fire risks are reduced to a minimum. The furnaces and combustion chambers in which enormous heat is concentrated remain cool on the outside.

Typical products of modern fuel-intensive industry are cars, with engines designed to be propelled by finely tuned and minutely controllable combustion processes. Indeed, the car may almost serve as a symbol of the highly complex and differentiated ways in which, in our day, thermal energy is being used. Cars are set in motion by burning fossil fuel. They are made of steel, plastic, and glass—materials that are produced and processed at high temperatures. Yet no one who gets into his or her vehicle and turns on the electrical ignition to start the engine needs be consciously aware of using fire and products of fire. When driving, people do not perceive the processes of combustion that keep their car going: they do not see the petrol gas burning under the bonnet, nor have most of them even remotely sensed the fire in the factories and power plants without which their cars would never have been produced at all.

A very different example to the same effect is farming. In the early 19th century, when Britain was already beginning to industrialize, practically all the energy consumed on the farm was still produced within the confines of the farm and its fields in the form of human and animal labor; the open fire that was burning in the hearth was fuelled with wood from the immediate surroundings. By the turn of the 21st century, the situation has become very different, with practically all the energy used now brought in from outside the farm, in the form of fertilizer, oil and petrol, and electricity.

A major advantage of the new sources of energy is their flexibility. The fuels are easier to transport and to distribute than wood or coal, and the combustion can be regulated more precisely. Given the technical facilities, gas, oil, and electricity provide for very even and accurately controllable flows of energy. Electricity has the additional advantage of being totally "clean" at the place of destination. Domestically, a few simple actions and a negligible risk suffice to provide people with an immense array of services: some substituting for old chores such as cleaning the floor and washing dishes, others based on entirely new appliances such as television sets and computers. Industrially, the same advantages apply at a much larger scale, permitting a far more diversified use of power than was possible with steam power.

The impact of electricity and internal combustion engines makes itself felt in every sector of social life: in agriculture, industry, traffic and transportation, domestic work and leisure. Everywhere it is possible to mobilize large quantities of energy with very little physical effort. The result is to make life more comfortable in many respects, enhancing the sense that physical processes can be mastered and also, at times, fostering the illusion of independence.

An illusion it clearly is. Regardless of whether people can avail themselves of energy in the form of a petrol engine, a battery, or a connection to an electric circuit or a gas main, in each case they are able to do so only because they are part of a complex and far reaching network of social interdependencies, connecting them eventually to the energy stored in fossil fuels. As long as the supply lines are functioning, and as long as people are able to meet their financial obligations, they do not need to bother much about the entire constellation. They are immediately confronted with it, however, the moment something goes wrong with any of the conditions.

In this way the exploitation of the new sources of energy clearly continues a trend that has begun with the first domestication of fire. Dependence on the forces of nature has become more indirect (which is not to say less important!), and by the same token dependence on cultural and social resources has increased. A complicated technical and organizational apparatus is needed to continuously maintain the supply of energy. Most of this apparatus is located behind the scenes of industrial society, invisible to the ordinary consumer. Energy is made available in such a convenient fashion that it is easy to forget the social effort required to produce it.

That social effort is spent, first of all, at the drilling wells and in the mines where the energy is won and, next, during the operations of processing it into consumable gas, oil, or electricity, and of transporting and distributing it. While the many provisions needed for the undisturbed flow of energy are often taken for granted, they cannot fail to exert permanent pressures on those who benefit from it as customers. The bills have to be paid, financially and otherwise.

The permanent availability of electricity, at every hour, in many parts of the world, has led to a diminution of the contrast between day and night. By the middle of the 19th century, the large investments made in their factories impelled many owners, under the pressures of competition, to let the engines run day and night. Gaslight illuminated the workplace. In the 20th century, night life has steadily extended, especially in the cities. Water mains, sewage, gas, electricity, telephone, fax, internet, radio, police, fire brigade, hospitals are all generally expected to operate day and night. International interdependencies never come to a halt. This is one of the reasons why many people turn on the news as soon as they wake up in the morning; before resuming their daily activities they wish to learn what has happened while they were asleep—in their own country, where it was night, and elsewhere, where it was day time.

Once in a while there is a hitch. Sometimes the local supply of electricity breaks down, as happened for a number of hours in the blackout in New York on July 13, 1977, and for longer periods at a regional scale in California in 2001 and in the northeastern United States and southern Canada in 2003. In New York the failure was due to technical causes; in California its primary causes lay in disturbances in the financial sector. A combination of economic and political complications brought about the international oil crisis of 1973, when the majority of oil-producing countries jointly managed to enforce a drastic increase in the world price of crude oil.

Disturbances are remarkably rare, in view of the prodigious growth of energy consumption since 1950. The industrial economy is a fuel economy, revolving around the regular supply of fuel that can be easily converted into energy. Light, warmth, motion, and even coolness are produced with fuel. The rising supply of all these fuel-intensive amenities in turn constantly stimulates demand from customers, who are eager to enhance both their physical comfort and their social status.

In a large part of the world, access to the benefits of modern technology is still restricted to a small upper crust of society. Moreover, while in the western world and parts of Asia effective measures have been taken to reduce the polluting effects of industrial and domestic fuel use in densely populated areas, the quality of the air has only deteriorated in the rapidly growing megacities of Africa, Asia, and Latin America. As a consequence, according to an estimate of the World Health Organization in 1997, air pollution killed about 400,000 people worldwide annually.

Still, economic growth has continued in the poor parts of the world, and this is bound to affect the rich countries as well. Wealth attracts poverty; history abounds with examples of this general rule. When the opportunity offers itself, many people from poorer regions will try to migrate to regions with a higher standard of material comfort.

Meanwhile, global fuel consumption continues to rise. In rich countries, advanced technologies permit

customers to enjoy the use of their appliances and vehicles without any physical inconvenience caused by the combustion processes. Poorer people generally are prepared to put up with noise and smell to profit from fossil fuel energy.

3.6 Fire and War

With growing interdependence among people, dependence on natural forces has become more indirect: longer and more ramified social chains are formed between the production of things and their use. Even the threat of violent destruction of lives and property comes far less often from natural forces than from forces unleashed by one human group against another. The most powerful groups are those that command the organizational and technical means to mobilize huge amounts of energy and matter against their enemies.

In our day, the search for the most effective means of violence has led to the exploitation of a new source of energy: nuclear fission. For the first time in human history, a new source of energy was first applied at a large scale in war, with the sole intent of massive destruction. It was the result of an enormously concentrated accumulation of technical, scientific, and economic capital, invested in a single one-purpose enterprise: the production of an atomic bomb.

But no matter how single-focused the effort, the invention of the atomic bomb followed a general rule in the history of technology and produced unintended side effects. One of these was that the development of the bomb by the United States became the first step in a process of escalation and proliferation, which has become the greatly enlarged version of the paradigmatic case *par excellence* of an arms race. Another unintended consequence has been the rising level of risk of fatal consequences in case of a breakdown in any of the nuclear plants that were built in the second half of the 20th century.

War has always been an important factor in the relations between humans and their physical environment. In most cases, it was directed at destruction of a part of the anthroposphere: especially that part in which the organizational basis of the enemy was supposed to be most vulnerable. The more investments a group made in controlling its environment, the more susceptible it became to losses through violence. Advanced agrarian communities could suffer a severe setback in case their rice fields, vineyards, or terraces were destroyed. Many cities in agrarian societies underwent drastic reduction or were even totally annihilated after military surrender: the greater the concentration of physical wealth, the more irreparable the destruction.

In industrial society, enormous means of destruction were developed even before the invention of the atomic bomb. During World War II, air raids brought devastation to a great many cities. Global industrial society proved to be sufficiently resilient and affluent, however, so that after the war every bombed city was rebuilt at its original site. Such were the powers of recuperation of the industrialized parts of the anthroposphere.

4. SYMBOLIC AND RELIGIOUS ASPECTS

4.1 The Blessings of Fire

We may safely assume that fire, having played such a crucial role in the development of human group life, has always kindled the imagination and given rise to rites and beliefs. Unfortunately, for the earliest stages we have only material remains to go by, and it is very difficult to reconstruct the symbolic and religious meanings attached to fire solely on that basis. Since the invention of writing, however, many facts relating to these intangible aspects have been recorded. More recently, anthropological fieldwork has yielded a profusion of data and insights.

Thus, from all over the world myths have been collected relating how man first came to master fire. Typically these myths represent the initial conquest of fire not as a gradual long-term process but as a single adventurous event, often marked by cunning and deceit. According to most myths, fire was at first possessed by the gods until, at a given moment, they gave it to humans or, alternatively, it was stolen from them, either by some animal who then passed it on to humans or by a semidivine Promethean culture hero.

All these myths convey a sense of the preciousness of fire. Fire is regarded as a unique possession, which separates humans from the animal world. As a corollary of this conception, many peoples used to tell that other peoples did not have fire and, consequently, could not cook their food so that they had to live like animals. Needless to say, fire has been a universal element of culture ever since the emergence of *Homo sapiens*, and all these stories are spurious and slanderous.

Directly related to such beliefs is the idea that fire steered by human hands has purifying effects. The domain of human culture is seen as coinciding with the range of domesticated fire. Thus, cooked meat is

regarded as edible for humans, raw meat as only fit for beasts. Similarly, land cleared by burning is trusted as belonging to the human domain, whereas uncleared bush and forest are feared as wild and dangerous, a hideout for predators, serpents, insects, and evil spirits.

The sense of preciousness of fire also found expression in rites. The communal fire would be the center of group life, and people would treat it as sacred, surrounding it with ceremony and taking great care that it would never cease burning. The Roman cults of the hearth of Vesta, the fire goddess, show how such ancient rites could persist in an urban setting. In many societies the perpetual burning of the communal fire would be interrupted once a year; all fires were then ritually extinguished and replaced by new, pure fire that had to be solemnly ignited by the chief priest.

4.2 The Curses of Fire

When, after the rise of agriculture, people increasingly came to live in villages and cities, the number and variety of fires also increased, and so did the risk that any one of these fires would cause a conflagration. Fire continued to be a cherished necessity, but since it was so abundantly available, the sense of its preciousness was often superseded by anxiety and fear of the dangers it posed.

The transition was reflected in some of the great religions that emerged in Eurasia during the past 3000 years. The veneration of fire was most strongly preserved in Hinduism, with its cult of the fire god Agni, and in Zoroasterism, with its cult of Atar. Both religions kept alive the age-old worship of fire as a primeval force of vital importance to humans; but this was combined with awe and anxiety for its destructive powers.

Christianity and Islam stand much further away from the ancient worship of fire. Some old traditions persist, as in the burning of incense and candles, including the renewal at Easter in Roman Catholic churches. Fire no longer plays any part in beliefs about the origins of human society, however. Instead, it is strongly associated with evil. Sinners and unbelievers are threatened with a life after death in which they will be tormented by eternal burning. It seems as if the agonies of city conflagrations as described by several pagan authors from ancient Rome have found a religious symbolization in the vision of hell.

Fire festivals, such as those held over Europe at Saint John's day when things considered impure would be committed to the flames, were occasions of combined joy and horror. Modern literature and art abound with evocations of the terrors of fire. When novelists describe the nightmare of a conflagration, they attribute the disaster, quite realistically, not to natural forces such as lightning but to human action—either to individual negligence, to madness or malice, or to collective hostilities in war.

4.3 Fire as an Element

A more neutral way of regarding fire, not in terms of its good or bad effects on human beings but as a natural phenomenon in its own right, was first proposed by philosophers of ancient Greece. They developed, out of various antecedents, a generally accepted doctrine according to which the universe consisted of four elements: earth, water, air, and fire. In China and India, similar systems of thought were worked out. With only minor modifications the worldview in which fire constituted one of the major elements remained the dominant cosmology for natural scientists and physicians well into the modern era.

Alchimists and chemists, as the direct successors of the metallurgists and fire masters of a previous age, made the study of fire into their central concern. Just as modern mechanical industry in its initial stages owed a great deal to the steam engine, the rise of modern physical science would have been inconceivable without the numerous efforts of investigators experimenting with and theorizing about fire.

5. CONCLUSION: THE ECLIPSE OF FIRE?

The controlled use of fire has been a human species monopoly since time immemorial. It constitutes an integral part of the apparatus with which humans have established and continue to maintain their dominance over other species.

Ever since the earliest stages of its domestication, the same phenomenon has recurred again and again in the human relations with fire: deliberately sought advances in control lead to unintended increases in dependency, and as the dependency sinks deeper into the social and economic infrastructure, it tends to become increasingly less clearly perceptible. Like agrarianization, industrialization began with conspicuously huge applications of fire. As industrial production became more specialized and highly organized, so did the use of fire. In everyday life in highly industrialized contemporary societies, flames

are visibly present only in such highly domesticated guises as cigarette lighters, candles for ceremonial use, or wood fires intended to create an atmosphere of comfort and relaxation. When, on the other hand, a fire is shown on television or in the press, it almost always spells war and disaster. The regularly controlled combustion processes on which industrial production largely rests are now mostly relegated away from public view.

In accordance with this apparent eclipse, natural scientists and philosophers no longer regard fire as one of the four elements out of which the world is composed. The very concept of fire has disappeared from scientific literature, to be replaced by the more abstract concept of energy—something that cannot be directly seen, heard, smelled, or felt. Similarly, many a day may pass in the life of the average citizen of a modern industrial state during which he or she does not perceive any physically observable fire. Yet it continues to be an integral part of the anthroposphere.

SEE ALSO THE FOLLOWING ARTICLES

Cultural Evolution and Energy • Early Industrial World, Energy Flow in • Forest Products and Energy • Hunting and Gathering Societies, Energy Flows in • Women and Energy: Issues in Developing Nations • Wood Energy, History of • Wood in Household Energy Use • World History and Energy

Further Reading

De Vries, B., and Goudsblom, J. (eds.). (2002). "Mappae Mundi. Humans and Their Habitats in a Long-Term Socio-Ecological Perspective." Amsterdam University Press, Amsterdam.

Elias, N. (2000). "The Civilizing Process. Sociogenetic and Psychogenetic Investigations." Rev ed. Blackwell, Oxford.

Goudsblom, J. (1992). "Fire and Civilization." Allen Lane, London.

Goudsblom, J., Jones, E., and Mennell, S. (1996). "The Course of Human History. Economic Growth, Social Process, and Civilization." M. E. Sharpe, Armonk, NY.

Grübler, A. (1998). "Technology and Global Change." Cambridge University Press, Cambridge, UK.

McClellan, J. E. III, and Dorn, H. (1999). "Science and Technology in World History." Johns Hopkins University Press, Baltimore.

Pyne, S. J. (1991). "Burning Bush. A Fire History of Australia." Henry Holt and Co., New York.

Pyne, S. J. (2001). "Fire. A Brief History." University of Washington Press, Seattle, WA.

Rossotti, H. (1993). "Fire." Oxford University Press, Oxford, UK.

Sieferle, R. P. (2001). "The Subterranean Forest. Energy Systems and the Industrial Revolution." The White Horse Press, Cambridge, UK.

Simmons, I. G. (1996). "Changing the Face of the Earth. Culture, Environment, History," 2nd ed. Blackwell, Oxford.

Turner, B. L., II, Clark, W. C., Kates, R. W., Richards, J. F., Marhews, J. T., and Meyer, W. B. (eds.). (1990). "The Earth as Transformed by Human Action Global and Regional Changes in the Biosphere over the Past 300 Years." Cambridge University Press, Cambridge, UK.

Wright, R. (2000). "Nonzero. The Logic of Human Destiny." Pantheon, New York.

Yergin, D. (1991). "The Prize: The Epic Quest for Oil, Money and Power." Simon and Schuster, New York.

Fisheries and Energy Use

PETER TYEDMERS
Dalhousie University
Halifax, Nova Scotia, Canada

1. **Introduction**
2. **Overview of Global Fisheries**
3. **Energy Inputs to Fisheries**
4. **Energy Performance of Fisheries for Human Consumption**
5. **Energy Performance of Fisheries for Reduction**
6. **Conclusions**

Glossary

active fishing gear Equipment used for fishing in which the aquatic organisms being targeted encounter the gear primarily through movement of the gear itself. Examples include cast nets, dredges, harpoons, jigs, and all forms of seine and trawl fishing.

animate energy Energy dissipated through the application of human or animal muscles to do work.

artisanal fisheries Small-scale fisheries that are typically executed by members of fishing households (as opposed to commercial companies) in which relatively small amounts of capital and energy and relatively small, if any, fishing vessels are employed.

culturally mediated energy All forms of energy dissipated through the application of technologies in the process of human activities.

energy intensity The amount of culturally mediated energy required to provide a given quantity of a product or service of interest. In the current context, energy intensity is expressed either as the total joules of energy required to land a live weight or "round" tonne of fish or shellfish harvested, or in terms of fuel use intensity, the total liters of fuel burned directly on fishing vessels per tonne of fish or shellfish landed.

energy return on investment (EROI) ratio A dimensionless ratio calculated by dividing the amount of useful energy provided by a given activity by the culturally mediated energy dissipated in providing it. In the case of food production systems, a common energy output used to calculate the EROI is the edible protein energy yield from the system being evaluated.

fishing Any activity that results in the catching, collecting, or harvesting of fish, and or aquatic invertebrates for any purpose other than scientific research. In the current context, the harvesting of aquatic plants and mammals is not considered as fishing, nor is aquaculture.

industrial fisheries Those fisheries typically undertaken by commercial companies in which relatively large quantities of capital and energy are deployed.

net primary productivity (NPP) The difference between the total amount of carbon taken up by plants via photosynthesis and the amount of carbon lost by living plants via respiration.

passive fishing gear Equipment used for fishing in which the aquatic organisms being targeted encounter the gear primarily through their own movement or as a result of the movement of the surrounding waters. Examples include all forms of hook and line gears, drift nets, gillnets, traps, and weirs.

Fishing is one of the most ancient and widespread of human endeavors. Contemporary fisheries harvest an enormous variety of aquatic organisms from virtually every aquatic environment on the planet using a diverse array of technologies. Reflecting the diversity of global fisheries, culturally mediated energy inputs also vary widely in both form and magnitude. At one extreme, traditional artisanal fisheries typically rely on relatively small inputs of exogenous energy, mostly in the form of wind, and animate energy to propel vessels and haul nets. Indeed, among subsistence fisheries in which animate energy inputs predominate, the nutritional value of the catch must routinely exceed the energy expended by the human muscles engaged in the fishery if it is to remain viable. Although such traditional, low-input fisheries persist in many parts of the world, high-input, industrialized fisheries now account for the majority of global landings. Among these fisheries, particularly those targeting high-value species, it is now common for direct fossil fuel energy inputs alone to exceed the nutritional energy embodied in the catch by at least an order of magnitude.

 © 2004 Elsevier Inc. All rights reserved.

1. INTRODUCTION

As with all human activities, fishing entails the dissipation of energy in support of its primary activity, the harvesting of aquatic organisms. While the energetic cost of fishing is less obvious, and consequently receives much less attention than the direct impact that fishing has on targeted stocks and associated marine ecosystems, it is precisely the availability of abundant energy, in particular fossil energy, that enables many contemporary fisheries to continue even when stocks are in decline. Consequently, analyses of the forms and quantities of energy dissipated in fisheries, and in particular changes in energy use over time, can provide a powerful measure of the biophysical scarcity of exploited populations.

2. OVERVIEW OF GLOBAL FISHERIES

2.1 Scope and Extent

Data compiled by the Food and Agriculture Organization (FAO) of the United Nations indicate that, globally, approximately 90 million tonnes of fish and shellfish, representing more than 130 genera and 750 species, are harvested from the wild each year (Fig. 1). Although in excess of 200 countries report landings of fish and shellfish, the top 10 fishing nations account for over half of the total (Fig. 2). Similarly, while fishing is conducted in virtually every aquatic environment on the planet, over 90% of global landings are derived from marine waters, with most of this coming from highly productive coastal shelf and upwelling ecosystems.

2.2 Importance of Fisheries

One of the oldest forms of food production, fishing currently provides two-thirds of the fish and shellfish consumed by people worldwide, with aquaculture accounting for the balance. As a result, fishing directly contributes approximately 10% of the total animal protein intake by humans. In addition, although wild-caught food fish tend to contain less fat than terrestrial animals, the chemical composition of marine-sourced lipids is of significant nutritional importance. In particular, many species of wild-caught fish are relatively rich sources of Omega-3, or n-3, fatty acids, compounds that are believed to

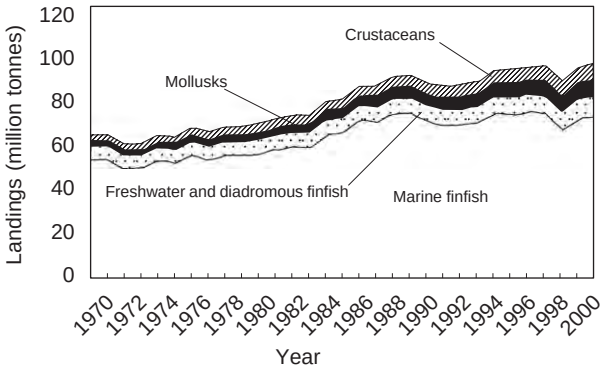


FIGURE 1 Global capture fisheries landings: 1970–2000. Data from FAO statistical database Fishstat+.

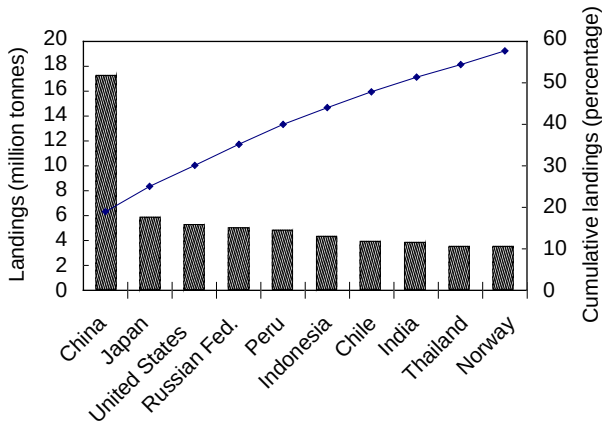


FIGURE 2 Absolute and cumulative fisheries landings by the top 10 fishing nations in 1998. Data from FAO (2000).

substantially decrease the risk of heart disease, inflammatory processes, and certain cancers.

It is important to note, however, the role of fish in human diets varies widely among countries and regions (Fig. 3). This variation not only reflects differences in resource availability and cultural preferences with respect to the seafood consumption but also, to a certain extent, development status. Of the 30 countries whose citizens are most dependent on fish as a source of protein, over 80% are less developed countries.

Of the approximately one-third of global landings not consumed directly as human food, the vast majority is destined for reduction to fish meal and oil. Currently, both of these reduction coproducts are directed overwhelmingly to livestock and aquaculture feeds, with smaller volumes of fish oil being consumed directly by people, either as edible oil products or nutritional supplements. Traditionally, however, fish oils have also been used in a wide

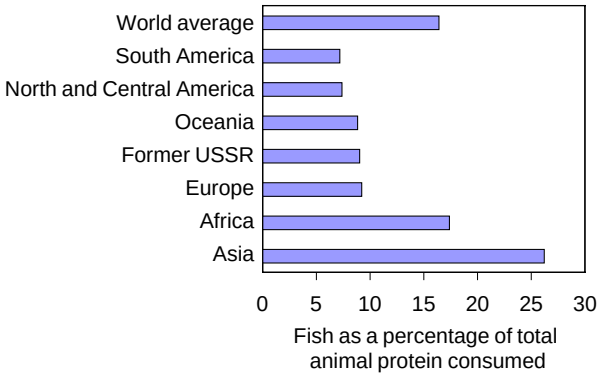


FIGURE 3 Importance of fish and shellfish from all sources to human diets.

range of industrial applications or burned as a source of fuel and illumination. While representing relatively small total landings, other nonfood fisheries are conducted to provide an array of high-value medicinal, decorative, and luxury products including shark cartilage, pearls, and live fish for the aquarium trade.

Reflecting their scale and diversity, it is not surprising that fisheries can play an important role in national and regional economies. Globally, the FAO estimates that as of the late 1990s, fisheries directly employed approximately 25 million people and yielded products whose first sale value exceeded $80US billion annually. The socioeconomic impact of fisheries is most pronounced, however, in many less industrialized countries and rural coastal communities throughout the world, where alternative employment opportunities are often limited.

2.3 Fishing Technologies

Based on archaeological evidence from central Africa, humans have been using dedicated fishing technologies, in this early instance a fishing spear, for at least 90,000 years. As the scope and range of fisheries has expanded over the intervening millennia, so too has the diversity of methods used to harvest fish and shellfish. While it is impractical here to explore the immense range of fishing techniques used around the world, a few broad generalizations are useful: fisheries in which some form of fishing vessel is used account for the vast majority of global landings, and most of the world's catch is taken using active fishing gears, particularly trawling and purse seining techniques in which large nets are respectively either towed behind fishing vessels or used to encircle schools of fish.

3. ENERGY INPUTS TO FISHERIES

Energy flows have been used to evaluate the performance of food production systems for over a hundred years. However, it was not until the first oil price shock of 1973 and the resulting concern regarding the dependency of industrialized food production systems on fossil fuels that systematic analyses were undertaken. One of the first major products of this period of research was Gerald Leach's handbook "Energy and Food Production" published in 1976. While it dealt primarily with agricultural systems, it included, for the first time, data on the major culturally mediated energy inputs to six fisheries from four continents. Since then, a small number of researchers in various parts of the world have continued to evaluate the energetics of fisheries from a variety of perspectives.

From an energetic perspective, fishing is a process in which a variety of culturally mediated forms of energy are dissipated in order to capture edible chemical energy embodied in the carcasses of fish and shellfish. The ultimate source of both the chemical energy embodied in the harvested organisms and most forms of culturally mediated energy is the sun. Consequently, the amount of solar energy required to sustain global landings and fishing activities is enormous. However, to date very few fishery-specific energy analyses have systematically attempted to account for solar energy inputs. As a result, other than a brief discussion of the ecosystem support associated with growing the biomass of animals harvested by global fisheries, solar energy inputs are not a focus of this review. Instead, this work deals primarily with the major forms and quantities of culturally mediated energy inputs to fishing.

3.1 Ecosystem Support of Fisheries

While not a direct measure of the solar energy flux underpinning fish and shellfish harvests, estimates have been made of the scale of ecosystem support appropriated by global fisheries. Typically, this is done by estimating the fraction of global aquatic net primary productivity (NPP)—or the amount of carbon fixed by plants via photosynthesis minus that lost through respiration—required to sustain the mass of fish and shellfish harvested. In 1995, Daniel Pauly and Villy Christensen, fisheries scientists currently based at the University of British Columbia, provided the most robust analysis of this measure to date. Their work indicates that as of the

mid-1990s, global fisheries landings and discards appropriated approximately 8% of the total NPP available in the world's aquatic ecosystems. More striking, however, were their findings regarding those highly productive freshwater and near-shore marine ecosystems that sustain nearly 85% of total fisheries landings yet spatially represent less than 10% of the planet's total aquatic environment. Within the context of these most productive aquatic ecosystems, fisheries alone appropriate between 24 and 35% of the NPP available.

3.2 Forms of Direct Energy Inputs

Culturally mediated energy inputs to fishing activities can be categorized into direct and indirect types. Common indirect inputs, often referred to as embodied energy inputs, are those associated with building and maintaining fishing vessels and providing fishing gear, bait, and ice. In contrast, in most fisheries, direct energy inputs are typically those required to propel fishing vessels and deploy fishing gears. The three dominant forms of energy dissipated to these ends include animate, wind, and fossil fuel energy, each of which are briefly described next.

3.2.1 Animate Energy

Animate energy is the one form of culturally mediated energy common to all of the world's fisheries regardless of their technological sophistication. In many traditional artisanal fisheries, human muscles still provide the bulk of the energy used to deploy gears, handle the catch, and, where they are employed, propel vessels. However, while animate energy inputs may represent the dominant form of energy applied in a fishery, in absolute terms, the quantities involved are typically small. For example, among subsistence fisheries in which animate energy inputs predominate, the nutritional value of the catch must routinely exceed the energy expended by the human muscles engaged in the fishery if it is to remain viable. In the case of most contemporary fisheries, however, while human animate energy inputs remain part of the production equation, they are generally dwarfed by the inputs of wind or fossil fuel energy.

Unlike preindustrial agriculture wherein a wide variety of animals were domesticated to provide important secondary sources of animate energy, relatively few fisheries have systematically employed animals. Examples, however, include the traditional use of trained fishing cormorants in parts of China and Japan and the now largely outmoded European practice of using otters and diving ducks to drive fish into traps.

3.2.2 Wind

For as long as people have sailed, it is likely that wind energy has been used to support fishing activities. The integration of wind energy into fisheries not only allowed fishing vessels to be propelled farther and faster than would otherwise be possible, it facilitated the development of fishing techniques that would otherwise have been impractical, if not impossible, to conduct from a rowed boat. Specifically, various trawl or dragger fisheries in which nets or dredges are towed either through the water or along the seafloor were almost all first developed within the context of sail fisheries.

Although little if any research has been conducted to quantify the wind energy inputs to sail-assisted fisheries, prior to the introduction of fossil fuels, it likely accounted for a large proportion of the culturally mediated energy inputs to global fisheries.

3.2.3 Fossil Fuels

As with most other food production sectors, fossil fuels have become the dominant form of energy used in fishing in a relatively short period of time. The process began in England in the late 1800s, when coal-fired steam engines were first installed on trawl fishing vessels to provide power for both propulsion and net hauling. With the advantages of increased speed and power, together with the ability to operate regardless of the wind, steam trawling expanded rapidly. Although coal- and oil-fired steam engines remained in use on fishing vessels throughout much of the 20th century, their significance was ultimately eclipsed by internal combustion engines.

Gasoline- and diesel-fueled internal combustion engines were first adapted for use on fishing boats in the early 1900s. Technological advances made during and after the Second World War, however, greatly accelerated their integration into fisheries. Consequently, over the past 50 years not only has the size of the global fishing fleet increased but so has its power. The trend to larger, more powerful fishing vessels is exemplified by the emergence of so-called supertrawlers, vessels that can exceed 100 m in length with propulsive engines well in excess of 10,000 horsepower. At the same time, more and more relatively small engines are introduced annually into small-scale fisheries around the world. Given these twin trends of increasing prevalence and size of engines, it is not surprising that the bulk of the

world's catch is now harvested using vessels propelled by fossil fuels.

3.3 The Relative Importance of Direct and Indirect Energy Inputs

Among modern industrial fisheries in which the major direct and indirect culturally mediated energy inputs have been systematically analyzed using either process analysis or input-output techniques, a consistent pattern has emerged. Direct fuel energy inputs typically account for between 75 and 90% of the total culturally mediated energy inputs, regardless of the fishing gear used or the species targeted (Fig. 4). Depending on the character of the fishery and the scope of the analysis conducted, the remaining 10 to 25% is generally composed of energy inputs associated with vessel construction and maintenance, and the provision of labor, fishing gear, bait, and ice if used.

On most fishing vessels, direct fuel inputs are used primarily for vessel propulsion. In some fisheries, however, secondary energy-consuming activities, including onboard processing, refrigeration, and freezing, can account for a nontrivial portion of the fuel burned. Squid jig fisheries are interesting and fairly extreme examples of fisheries in which a relatively large proportion of fuel inputs are used for activities other than vessel propulsion. In these fisheries, vessels typically employ batteries of high-intensity lamps, automated jigging machines, and freezers, all powered by diesel-fuelled generators to attract, hook, and preserve the catch while at sea. As a result, these nonpropulsion energy demands can, in combination, account for a much as 40% of the total fuel burned (Fig. 5).

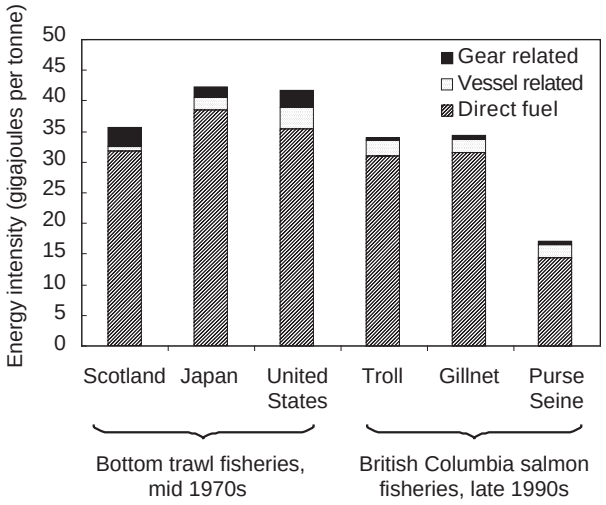


FIGURE 4 Relative contribution of direct and indirect energy to various fisheries. Data sources: Scotland: Edwardson, W. (1976). The energy cost of fishing. *Fishing News Intl.* **15**(2), p. 36–39. Japan: Watanabe and Uchida (1984); United States: Rawitscher (1978); British Columbia salmon fisheries: Tyedmers (2000).

Among indirect energy inputs, those associated with building and maintaining the fishing vessels themselves regularly account for the largest fraction (Fig. 4). This is particularly the case when a vessel's major components (i.e., its hull, superstructure, decks, and holds) are fabricated primarily from relatively energy-intensive materials such as aluminum and steel as opposed to wood or fiberglass. For example, an analysis of the direct fabrication and major embodied energy inputs to build and equip two equivalent 10 m long salmon gillnet fishing vessels, one constructed largely of fiberglass and wood and the other of aluminum, found that the embodied energy inputs were over 50% higher in the case of the aluminum-hulled vessel (Table I). Moreover, when the electricity required to fabricate the aluminum-hulled vessel is included in the analysis, the total embodied energy costs of this vessel rise to 2.5 times that of its fiberglass counterpart.

3.4 Comparing the Energy Performance of Fisheries

Most energy analyses of fisheries have focused largely, if not exclusively, on evaluating the direct fuel inputs to fishing. This relatively narrow focus

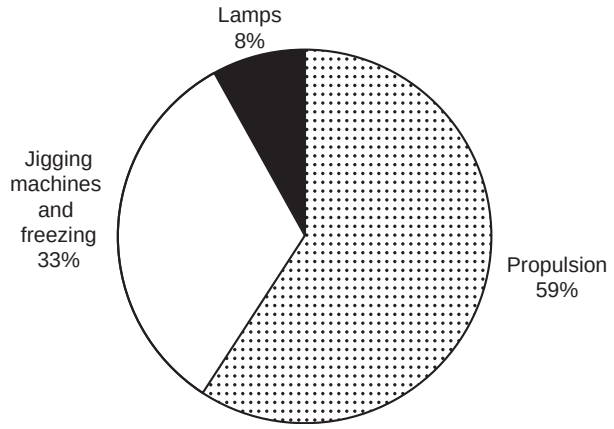


FIGURE 5 Functional fuel utilization on a Japanese long-distance squid fishing vessel. Data from Ishikawa, *et al.* (1987). A case study of energy analysis of long distance squid angling. *Bull. Jap. Soc. Sci. Fish.* **53**(9), 1525–1531.

TABLE I

Direct and Indirect Energy Inputs to Build a 10-Meter-Long Salmon Gillnet Fishing Vessel Fabricated from Either Fiberglass or Aluminum

	Material and embodied energy inputs			
	Fiberglass hull		Aluminum hull	
Inputs	kg	GJ	kg	GJ
Glass				
Matting	1500	15		
Sheet	100	1	100	1
Wood (m^3)	5	21		
Fiberglass resin	1200	90		
Steel				
Engine and transmission	1110	28	1110	28
Fishing equipment	400	10	400	10
Anchor, winch and chain	200	5	200	5
Hardware and other	1050	26	550	14
Lead (batteries)	100	2	100	2
Mixed metals (controls, etc.)	390	10	390	10
Aluminum				
Fuel and water tanks	540	75	540	75
Fishing equipment	330	46	330	46
Hull, decks, holds, and cabin			2250	315
Total embodied energy		329		506
Electricity to weld hull, etc.				324
Total		329		830

not only reflects the fact that these inputs dominate the energy profile of fishing, but detailed analyses of indirect inputs tend to be labor intensive and time consuming. Therefore, in order to facilitate the comparison of as wide a range of fisheries as possible, throughout the balance of this review only the direct fuel energy inputs to various fisheries is considered. Consequently, the energy intensities of various fisheries is expressed in terms of the liters of fuel burned per "round" or live weight tonne of fish or shellfish landed.

However, as food fisheries can also be thought of as food energy "producing" systems, it is also useful to calculate their energy efficiency. Traditionally within analyses of agriculture, aquaculture, and fisheries systems, this has been done by calculating either their industrial energy input to edible food energy output ratio or its inverse, the edible energy return on investment (EROI) ratio. As the nutritional importance of fish and shellfish is primarily a function of their protein content, edible protein energy output is the most meaningful basis of comparison. Therefore, edible protein energy return on investment (protein EROI) ratios for various fisheries are also presented.

4. ENERGY PERFORMANCE OF FISHERIES FOR HUMAN CONSUMPTION

To date, most energy analyses of fisheries for human consumption have concentrated on relatively large-scale industrial fisheries that are conducted for commercial gain. Moreover, few analyses have been undertaken on fisheries based in the Southern Hemisphere. As a result, much of the discussion that follows reflects these biases.

Given the diversity of fisheries for human consumption, both in terms of species harvested and fishing gears used, it is helpful to focus on broad subsets when considering their energy performance. Consequently, the following discussion is organized into four major divisions: fisheries for demersal species, or those fish that spend much of their life on or near the ocean floor; fisheries for pelagic species, or those fish that spend most of their life swimming in the water column; fisheries for invertebrates; and subsistence fisheries.

4.1 Fisheries for Demersal Species

Fisheries for demersal finfish species are some of the largest, most economically important fisheries in the world. Each year, upwards of a hundred species of demersal finfish may be harvested in fisheries globally. Of these, the most important include the various gadoid, or codlike species, and flatfish species. Given that, by definition, demersal species live on or near the ocean floor, fisheries for these species are undertaken in waters ranging from under 10 m to many hundreds of meters deep. As a result, many different types of fishing gears are used in their capture. By far the largest tonnages, however, are

harvested each year using trawl and to a lesser extent, long-line fishing gear.

As a result of their importance, fisheries for demersal species have been the subject of numerous energy analyses over the last three decades (Table II). Some of the most recent work to systematically evaluate the fuel consumption of a wide variety of demersal finfish fisheries examined 29 fisheries involving more than 3000 vessels based in Atlantic Canada, Iceland, Norway, and Germany. As of the late 1990s, these fisheries together harvested approximately 1.2 million tonnes of fish annually from ecosystems around the North Atlantic. While fuel use intensities varied by an order of magnitude among these 29 fisheries, from lows of just under 300 l/t to more than 2000 l/t, on average they consumed just over 500 liters of diesel per tonne of fish landed.

TABLE II

Energy Performance of Industrial Fisheries for Direct Human Consumption

Main fishery targets	Gear	Time frame	Location of fishery	Fuel use intensity (liters/tonne)	Edible protein EROI
Demersal fisheries					
Redfish spp.	Trawl	Late 1990s	North Atlantic	420[a]	0.11
Cod/flatfish spp.	Danish seine	Late 1990s	North Atlantic	440[a]	0.10
Cod/haddock	Longline	Late 1990s	North Atlantic	490[a]	0.091
Cod/saithe	Trawl	Late 1990s	North Atlantic	530[a]	0.084
Alaskan pollock	Trawl	Early 1980s	North Pacific	600[b]	0.052
Flatfish spp.	Trawl	Early 1980s	NW Pacific	750[b]	0.066
Croakers	Trawl	Early 1980s	NW Pacific	1500[b]	0.029
Flatfish spp.	Trawl	Late 1990s	NE Atlantic	2300[a]	0.019
Pelagic fisheries					
Herring/mackerel	Purse seine	Late 1990s	NE Atlantic	100[a]	0.56
Herring	Purse seine	Early 1990s	NE Pacific	140[c]	0.36
Herring/saithe	Danish Seine	Late 1990s	NE Atlantic	140[a]	0.35
Salmon spp.	Purse seine	1990s	NE Pacific	360[c]	0.15
Salmon spp.	Trap	Early 1980s	NW Pacific	780[b]	0.072
Salmon spp.	Gillnet	1990s	NE Pacific	810[c]	0.068
Salmon spp.	Troll	1990s	NE Pacific	830[c]	0.067
Herring	Purse seine	Early 1980s	NW Pacific	1000[b]	0.051
Skipjack/tuna	Pole and line	Early 1980s	Pacific	1400[b]	0.053
Skipjack/tuna	Purse seine	Early 1980s	Pacific	1500[b]	0.049
Swordfish/tuna	Longline	Late 1990s	NW Atlantic	1740[a]	0.042
Salmon spp.	Gillnet	Early 1980s	NW Pacific	1800[b]	0.031
Swordfish/tuna	Longline	Early 1990s	Central Pacific	2200[d]	0.027
Tuna/billfish	Longline	Early 1980s	Pacific	3400[b]	0.022
Shellfish fisheries					
Abalone/clams	Hand gathering	Early 1980s	NW Pacific	300[b]	0.11
Crab	Trap	Late 1990s	NW Atlantic	330[a]	0.057
Scallop	Dredge	Late 1990s	North Atlantic	350[a]	0.027
Shrimp	Trawl	Late 1990s	North Atlantic	920[a]	0.058
Shrimp	Trawl	Early 1980s	North Pacific	960[b]	0.056
Norway lobster	Trawl	Late 1990s	NE Atlantic	1030[a]	0.026
Crab	Trap	Early 1980s	NW Pacific	1300[b]	0.014
Spiny lobster	Trawl	Early 1980s	NW Pacific	1600[b]	0.017
Squid	Jig	Early 1980s	NW Pacific	1700[b]	0.033
Shrimp	Trawl	Late 1990s	SW Pacific	3000[d]	0.019

Notes to sources: [a]Tyedmers (2001); [b]Watanabe and Okubo (1989); [c]Tyedmers (2000); [d]unpublished data.

Much of the variability in fuel use intensity among the 29 North Atlantic fisheries appears to be a function of two factors. Not surprisingly, the relative abundance and catchability of various targeted species has a major influence on resulting energy performance. Evidence of this is can be seen by comparing average fuel use intensities of various trawl fisheries targeting very different species assemblages in different parts of the North Atlantic (Table II). Specifically, Canadian and Icelandic trawl fisheries targeting redfish species experienced average fuel use intensities of just over 400 l/t while trawl fisheries for cod and related species from across the North Atlantic averaged 530 l/t. At the other extreme, German trawl fisheries targeting high value flatfish species in the North Sea consumed approximately 2300 liters of diesel per tonne of fish landed.

A second factor that affects the energy performance of fisheries is the type of fishing gear employed. This can be illustrated by focussing on those North Atlantic fisheries that targeted Atlantic cod and related species (Table II). On average, cod fisheries in which Danish seine nets were used burned approximately 440 liters per tonne of fish landed. In contrast, long-line fisheries averaged 490 l/t, while trawl fisheries for cod averaged 530 l/t.

Looking beyond the North Atlantic, a comprehensive analysis of the fuel use intensity of all Japanese fisheries as of the early 1980s suggests that North Pacific fisheries for demersal species may be more energy intensive than many of their North Atlantic counterparts (Table II). Specifically, Japanese trawl fisheries targeting either Alaskan pollock, various flatfish species or yellow croaker burned, respectively 600, 750, and 1500 liters of diesel per tonne of fish landed. It is important to keep in mind, however, when comparing the performance of these Japanese fisheries with those described previously from the North Atlantic, that almost 20 years has elapsed between when the two data sets were acquired. Over that period of time much can have changed that might influence the energy performance of any specific fishery. As we shall see, however, evidence from a variety of fisheries suggests that there has been a general tendency for the energy performance of industrial fisheries to decline over time.

From an energy efficiency perspective, the performance of fisheries targeting demersal species also ranges widely. As measured in terms of their edible protein EROI, the 29 North Atlantic fisheries discussed previously together averaged 0.095. This means that the edible protein energy content of all the fish that they landed amounted to just under 10% of the fuel energy that they burned. More generally, edible protein EROI values for individual fisheries targeting demersal species can range from approximately 0.02 to over 0.1 (Table II). While efficiencies in this range may seem low, they are broadly in keeping with the energy performance of other industrial fisheries and, in fact, compare favorably with other animal protein production systems (Table III).

4.2 Fisheries for Pelagic Species

Pelagic fish species harvested for human consumption encompass a diverse variety of animals. They range from the small, densely schooling, largely planktivorous herrings, to large, top-level carnivores including various tuna and billfish species. What they generally have in common, however, is that all are typically harvested using fishing gears deployed at or near the ocean's surface.

TABLE III

Edible Protein EROI values from Livestock and Aquaculture Production Systems

Production system (locale)	Edible protein EROI
Carp—unspecified culture system (Indonesia)	0.70
Tilapia—unspecified culture system (Africa)	0.11
Mussel—longline culture (Scandinavia)	0.10
Carp—unspecified culture system (Israel)	0.084
Turkey (United States)	0.077
Milk (United States)	0.071
Tilapia—unspecific culture (Israel)	0.066
Tilapia—pond culture (Zimbabwe)	0.060
Swine (United States)	0.056
Eggs (United States)	0.038
Catfish—intensive pond culture (United States)	0.030
Chicken (United States)	0.029
Tilapia—intensive cage culture (Zimbabwe)	0.025
Atlantic salmon—intensive cage culture (Canada)	0.025
Shrimp—semi-intensive culture (Colombia)	0.020
Chinook salmon—intensive cage culture (Canada)	0.020
Lamb (United States)	0.020
Atlantic salmon—intensive cage culture (Sweden)	0.020
Beef—feedlot (United States)	0.019
Seabass—intensive culture (Thailand)	0.015
Shrimp—intensive culture (Thailand)	0.014

For sources, see Tyedmers (2001).

Reflecting the diversity of their quarry, fisheries for pelagic species display the widest range of energy intensities of fisheries for human consumption (Table II). At one extreme are those that target herring and associated species using purse seine or Danish seine nets. For example, recent analyses of herring fisheries from both the Atlantic and Pacific oceans indicate that they typically consume between 100 and 150 liters of diesel per tonne of fish landed. It is interesting to note, however, that an analysis of a herring fishery based in Japan, conducted in the early 1980s, reported a fuel use intensity of 1000 l/t. It is unclear from the data available why such a large difference should exist between this fishery and comparable fisheries in other parts of the world.

In sharp contrast, analyses of fisheries for relatively high value, large pelagic species like tuna and billfish report fuel use intensities ranging from 1400 to 3400 l/t (Table II). Among these fisheries, those that use long-line fishing gear are often the most energy intensive.

Between these two energy performance extremes, lie fisheries for Pacific salmon. Reported fuel use intensities of salmon fisheries range from just under 400 l/t to 1800 l/t (Table II). In a comparative analysis of the three fishing gears used to harvest salmon in British Columbia, Canada, it was found that purse seiners burned, on average, just under half the amount of fuel, per tonne of salmon landed, than did either gillnet or troll fishing boats.

Largely as a result of the major differences in the amount of fuel that they burn, the energy efficiency of fisheries for pelagic species, measured in terms of their edible protein EROI ratios, range from as low as 0.02 (or 2%) to over 0.5 (or 50%).

4.3 Invertebrate Fisheries

While fisheries for invertebrates account for a relatively small proportion of global landings (Fig. 1), they are some of the most lucrative fisheries in the world. It is not surprising therefore that in many instances they can be highly energy intensive. For example, fisheries targeting various species of shrimp and lobster often burn upwards of 1000 liters of fuel per tonne of landings (Table II). Interestingly, however, not all relatively high value invertebrate fisheries are this energy intensive. Some, including the valuable North Atlantic scallop and crab fisheries, consume between 300 and 350 liters of fuel per tonne of shellfish landed.

As measured in terms of their edible protein EROI ratios, many invertebrate fisheries are relatively energy inefficient. This not only reflects rather high fuel use intensities, but the fact that many species of invertebrates yield relatively small quantities of edible meat. For example, the adductor muscle in scallops generally only represents 10 to 12% of the live weight of the animal. Similarly, the edible portion of many crab and lobster species normally does not exceed 30% of their live weight. In contrast, edible muscle usually constitutes between 50 and 60% of most finfish species.

4.4 Subsistence Fisheries

To date, little attention has been paid to the energy performance of small-scale fisheries. However, an analysis of a northern Canadian aboriginal subsistence fishery targeting primarily pike and pickerel in rivers and lakes provides some intriguing and potentially worrying insights into their performance. From data collected through the late-1970s, human ecologist Fikret Berkes estimated that Cree fishermen, based in the community of Fort George, Quebec, burned, on average, approximately 1400 liters of gasoline per tonne of fish landed in the course of tending their fishing nets. Remarkably, this level of fuel use intensity is on the same scale as some of the most energy intensive industrial fisheries for human consumption in the world. It also highlights the extraordinary vulnerability of fossil fuel-dependent artisanal fisheries, and ultimately the communities that depend on them, to disruptions in the supply or increases in the price of fuel.

4.5 Temporal Trends in Energy Performance

Numerous analyses have been undertaken to evaluate how the energy performance of fisheries for human consumption has changed through time. Although the time periods and types of fisheries analyzed differ, when their results are expressed in a consistent format, say in terms of their edible protein EROI ratios, a regular and troubling pattern emerges (Fig. 6). Over time, the energy performance of virtually every fishery analyzed has declined, and in some cases quite dramatically. For example, the edible protein EROI ratio of the entire New Bedford, Massachusetts-based fishing fleet experienced a five-fold declined from just over 0.15 (or 15%) to under 0.03 (or 3%) over the period from 1968 to 1988 (Fig. 6A). Similarly, over the course of a decade, beginning in the late 1960s, the edible protein EROI

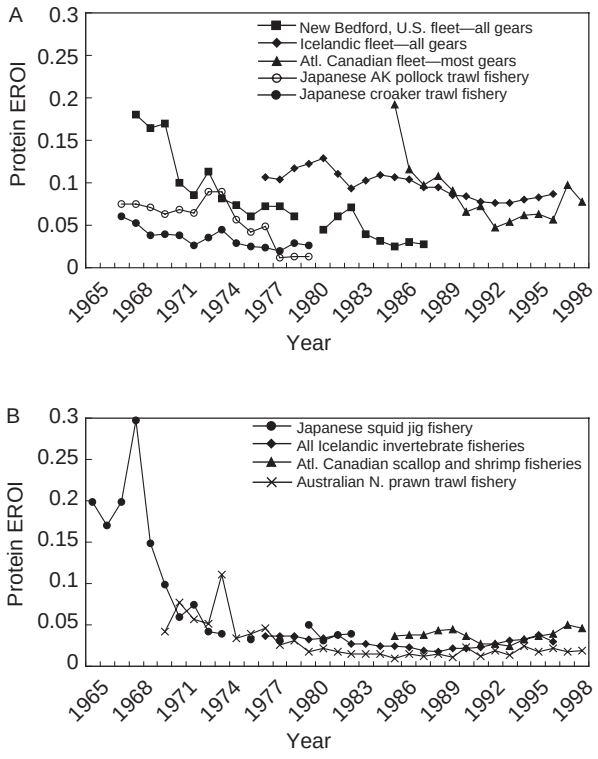


FIGURE 6 Changes in edible protein EROI ratios of fisheries for human consumption: 1965–1999. (A) Demersal finfish fisheries. (B) Shellfish fisheries. Data compiled from Watanabe and Uchida. (1984); Sato, K. *et al.* (1989). Chronological change of energy input for squid angling in Japan 1956–1983. *Bull. Jap. Soc. Sci. Fish* **55**(**11**), 1941–1945; Mitchell and Cleveland (1993). Resource scarcity, energy use and environmental impact: A case study of the New Bedford, Massachusetts, USA Fisheries. *Environ. Management* **17**(**3**), 305–317; Tyedmers (2001); and unpublished data.

ratio of the entire Japanese-based squid fishing fleet declined from over 0.20 (or 20%) to consistently below 0.04 (or 4%) (Fig. 6B).

While their relative importance may vary between fisheries, the two factors that likely contribute most to the general trend to poorer energy performance over time include decreases in relative abundance or proximity of targeted fisheries resources, and increases in the size, power, and technical sophistication of both fishing vessels and fleets.

5. ENERGY PERFORMANCE OF FISHERIES FOR REDUCTION

As many as 50 countries routinely report some production of fish meal and oil. However, less than a dozen account for the lion's share of total annual global output. The most significant of these include Peru, Chile, Iceland, Norway, Denmark, Japan, the United Kingdom, the United States, and the countries of the former Soviet Union.

TABLE IV

Fuel Energy Intensities of Reduction Fisheries as of the Late 1990s

Main fishery targets	Gear	Location of fishery	Fuel use intensity (liters/tonne)
Capelin/herring	Purse seine	NE Atlantic	20
Menhaden	Purse seine	NW Atlantic	32
Capelin/herring	Trawl	NE Atlantic	80
Blue Whiting	Purse seine	NE Atlantic	85
Sand Eels/herring	Trawl	NE Atlantic	95
Herring/sand Eels	Purse seine	NE Atlantic	100
Herring/mackerel	Trawl	NE Atlantic	110

Data source: Tyedmers (2001).

Although they account for 30% of global landings, most of the world's fish meal and oil production is derived from large tonnage landings of a relatively small number of species. In the waters off Peru and Chile the major species harvested for reduction include the Peruvian anchoveta, Inca scad, and the South American pilchard. In the northeast Atlantic, the main species destined for reduction include capelin, blue whiting, Norway pout, European sprat, various species of sand eel together with smaller quantities of Atlantic herring and Atlantic mackerel. In U.S. waters, Atlantic and Gulf menhaden account for virtually 100% of landings destined entirely to reduction.

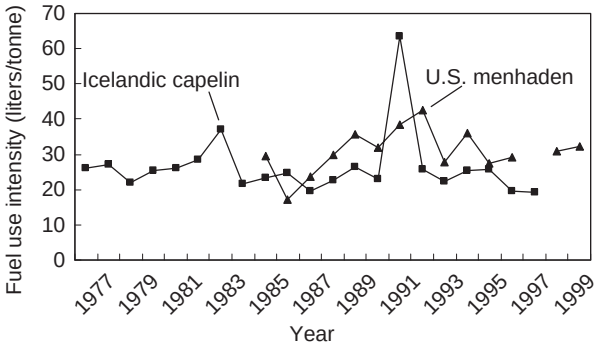


FIGURE 7 Changes in fuel use intensity of fisheries for reduction, 1977 to 1999. Data from Tyedmers, 2001, and unpublished data.

Recent analyses of various North Atlantic fisheries for reduction indicate that in most instances, they consume less than 100 liters of diesel per tonne of fish landed (Table IV). Remarkably, it is not unusual for some of the largest of these fisheries to routinely burn well below this mark. For example, time series data from both the Icelandic and American purse seine fisheries, targeting capelin and menhaden respectively, indicate that in recent decades, they have seldom consumed more than 40 liters of fuel per tonne of fish landed (Fig. 7). Such low and seemingly stable fuel use intensities would seem to suggest that in both instances, fish stocks and fleet effort are being carefully managed.

While the data available do not support direct head-to-head comparisons, from the spectrum of fuel use intensities associated with various North Atlantic fisheries for reduction, it would appear as if those employing purse seine nets are routinely less energy intense than those using trawl fishing gear (Table IV).

6. CONCLUSIONS

Global fisheries are highly varied. They harvest an enormous diversity of aquatic animals using a variety of mostly vessel-based fishing gears. Not surprisingly, therefore, the major forms and absolute quantities of culturally mediated energy inputs dissipated in their execution also vary widely. However, in the case of most fisheries, particularly large-scale industrial fisheries over the past half-century, fossil fuel inputs have come to dominate their energy profile.

Of concern, though, is evidence from a number of fisheries from around the world of a general tendency to poorer energetic performance over time. This pattern of ever-increasing apparent biophysical scarcity reinforces the oft expressed concerns regarding both the declining status of many of the world's fish stocks and the growing size and power of the global fishing fleet. In spite of the ongoing declines in their energy performance, however, many industrial fisheries for human consumption remain energetically competitive with other animal protein–producing systems.

SEE ALSO THE FOLLOWING ARTICLES

Aquaculture and Energy Use • Diet, Energy, and Greenhouse Gas Emissions • Ecosystem Health: Energy Indicators • Entrainment and Impingement of Organisms in Power Plant Cooling • Food System, Energy Use in • Industrial Agriculture, Energy Flows in • Livestock Production and Energy Use • Marine Transportation and Energy Use • Ocean, Energy Flows in • Wetlands: Impacts of Energy Development in the Mississippi Delta

Further Reading

Food and Agriculture Organization of the United Nations. (2000). The state of the world fisheries and aquaculture 2000. FAO, Rome, Italy.

Leach, G. (1976). "Energy and Food Production." IPC Science and Technology Press, Surrey.

Mitchell, C., and Cleveland, C. J. (1993). Resource scarcity, energy use and environmental impact: A case study of the New Bedford, Massachusetts, USA Fisheries. *Environ. Management* **17**(3), 305–317.

Pauly, D., and Christensen, V. (1995). Primary production required to sustain global fisheries. *Nature* **374**, 255–257.

Pimentel, D., Shanks, R. E., and Rylander, J. C. (1996). Bioethics of fish production: Energy and the Environment. *J. Agricult. Environ. Ethics* **9**(2), 144–164.

Rawitscher, M. A. (1978). "Energy Cost of Nutrients in the American Diet." Unpublished doctor of philosophy thesis. University of Connecticut, Storrs, United States.

Tyedmers, P. (2000). "Salmon and Sustainability: The Biophysical Cost of Producing Salmon through the Commercial Salmon Fishery and The Intensive Salmon Culture Industry." Unpublished doctor of philosophy thesis. University of British Columbia, Vancouver, Canada.

Tyedmers, P. 2001. Energy consumed by North Atlantic Fisheries. *In* "Fisheries Impacts on North Atlantic Ecosystems: Catch, Effort and National/Regional Datasets" (D. Zeller, R. Watson, and D. Pauly, Eds.), Fisheries Centre Research Reports **9**(3), 12–34.

von Brandt, A. (1984). "Fish Catching Methods of the World." Avon Litho Ltd., Stratford-upon-Avon, Warwichshire.

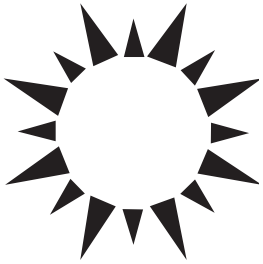
Watanabe, H., and Okubo, M. (1989). Energy input in marine fisheries of Japan. *Bull. Jap. Soc. Sci. Fish.* **53**(9), 1525–1531.

Watanabe, H., and Uchida, J. (1984). An estimation of direct and indirect energy input in catching fish for fish paste products. *Bull. Jap. Soc. Sci. Fish.* **50**(3), 417–423.

Wilson, J. D. K. (1999). "Fuel and Financial Savings for Operators of Small Fishing Vessels." FAO Fisheries Technical Paper 383. FAO. Rome, Italy. 46.

Flywheels

JOHN R. HULL
Argonne National Laboratory
Argonne, Illinois, United States



1. Introduction
2. Rotor
3. Bearings
4. Power Input and Output
5. Containment
6. Applications

Glossary

flywheel Rotating object, such as a wheel, often used for opposing and moderating by its inertia any fluctuation in speed in the machinery with which it revolves; also capable of storing mechanical energy.

high-temperature superconductor (HTS) A superconductor with a critical temperature above the boiling point of nitrogen.

motor/generator (M/G) Combined motor and generator used to convert between the mechanical energy of the flywheel and electrical energy.

rotational kinetic energy Energy in the form of rotational motion given by the quantity $KE = 1/2\ I\omega^2$, where KE is the kinetic energy (units of joules), I is the moment of inertia (units of kgm^2, and ω is the rotational speed (units of rad/s).

specific energy Energy stored divided by mass of object (units of m^2/s^2).

The flywheel, often in the form of a spinning disk or cylinder, stores rotational kinetic energy. As an integral part of the spindle whorl and the potter's wheel, the flywheel is among the most ancient mechanical devices known to humans. The use of the wheel for storing energy seems to be older than its use as a means of supporting vehicles. The flywheel is a fundamental element of many machines, opposing and moderating by its inertia any fluctuation of speed in the machinery with which it revolves. The flywheel principle is also used in several toys such as the top and the yo-yo. More recently, flywheels have been used as mechanical batteries to store energy for electrical applications. If a motor/generator (M/G) is attached to a flywheel, rotational kinetic energy may be transformed easily and efficiently into electrical energy and vice versa.

1. INTRODUCTION

The essential components of the flywheel are the rotor, bearings, power conversion, and containment. The rotor is the rotating mass in which the energy is stored. The bearings connect the rotating part of the flywheel to a nonrotating platform. Conversion of the stored energy to mechanical power may be accomplished with a mechanical linkage such as a transmission gear. When the flywheel is used as a mechanical battery, a motor/generator (M/G) converts energy between rotating kinetic and electrical. Flywheels are capable of attaining high specific energy, and tip speeds of high-performance flywheels are several times the speed of sound in air. Such flywheels are contained in a vacuum vessel, and an additional safety-related structure often surrounds the flywheel to contain possible debris in case of failure.

The kinetic energy E of a rotating rigid body is

$$E = 1/2\ I\omega^2,$$

where I is the moment of inertia of the body about the axis of rotation and ω is the angular frequency of rotation. The flywheel increases its rotational speed as energy is accumulated. The flywheel slows when it discharges energy for use.

A distinctive feature of all flywheel energy storage (FES) systems is their very high power density. They can be charged at very high rates and deliver very high powers, with the only limit being the rating of the transmission system and the torque the flywheel itself can withstand. Power densities greater than 500 W/kg are commonplace. The specific energy of a

 © 2004 Elsevier Inc. All rights reserved.

flywheel is comparable to electrochemical batteries in the range of 10 to 50 Wh/kg.

The discussion in this article is limited to the flywheel used for energy storage. Flywheels have several advantages over chemical batteries. First, flywheel life is nearly independent of the depth of discharge. The power is limited by the power electronics and the size of the M/G rather than by electrochemistry. Second, the state of the charge of a flywheel battery is readily determined from its rotational velocity, whereas determination of the state of charge for an electrochemical battery is more difficult. The flywheel generates no emissions, makes little noise, and is made of materials with negligible disposal problems.

2. ROTOR

The rotor is the rotating part of the flywheel. It consists of components of the bearing, the M/G, and material whose only purpose is to provide the moment of inertia.

The maximum amount of kinetic energy that can be stored in a rotating solid is determined by its geometry and the strength of the materials that compose the solid. For a rotor composed of a single isotropic material, which is the case for most metals, the maximum specific energy e is given by

$$e = K\sigma_u/\rho,$$

where σ_u is the ultimate burst strength, ρ is the density, and K is a geometric factor of unity or less. The geometric factor K depends only on the shape of the rotor, so the specific energy is independent of size. Because specific energy is proportional to the square of the velocity, a consequence of this equation is that, for a fixed geometric proportion, the rim speed at which a rotor will burst is independent of size. Thus, rim speed is a better figure of merit for comparing rotor materials than is rotational speed because the rotational speed at which the rotor bursts is inversely proportional to size; that is, smaller rotors can attain higher rotational speeds.

In general, when calculating the burst speed, one must take into account tangential hoop stress and radial stress independently. For most geometries with isotropic materials, hoop stress dominates. For operating flywheels, a factor of safety is introduced so that none of the components ever approaches burst speed. Moderate-strength steel, such as 4130, is often used for metallic rotors because of low cost and manufacturing experience. Maraging steel and titanium alloys are used for metallic rotors with higher specific energy. Circumferential stress must also be taken into account when the flywheel is accelerating or decelerating. However, in nearly all cases, the stress induced by acceleration is negligible when compared with the stress induced by centrifugal force.

The specific energy of flywheels can increase dramatically if anisotropic materials such as fiber composites are used to create the rotor. Although many combinations and designs of composites are possible, the most common designs contain long strands of high-strength filaments embedded in an epoxy matrix. The strands are wound predominantly in the circumferential direction. An example of such a rotor is shown in Fig. 1. E-glass, S-glass, graphite, and Kevlar are typically used fibers. Graphite fibers are the strongest and may be 3 to 10 times stronger than, and approximately one-fifth as dense as, steel. Graphite is a crystalline form of carbon. The diameter of an individual graphite fiber ranges from approximately 25 to 800 μm, and the strength of a fiber can range from approximately 2 to 8 Gpa. Usually, the smaller the diameter, the stronger the fiber.

In addition to providing higher specific energy, fiber composites provide an advantage over metals in the area of confinement. When a metallic rotor breaks, its debris usually includes several large pieces moving at high speed. A fiber composite rotor tends to fail with small shard-like pieces or even small grains.

Fiber-based composites exhibit low shear stress, which is related to delamination or separation of the epoxy from the fiber. Thus, hoop-wound designs that contain fiber composites are very strong in the hoop

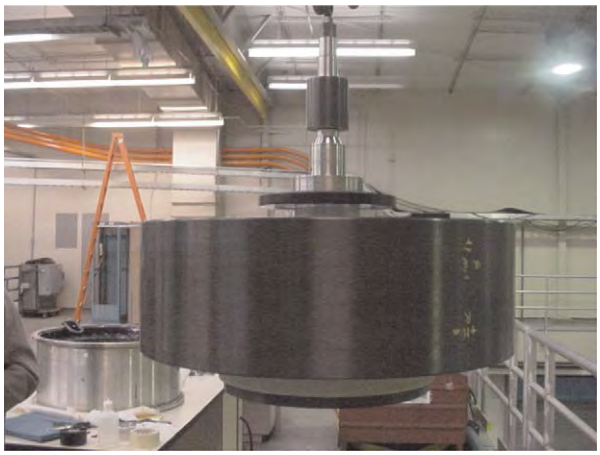


FIGURE 1 Photo of a 1-kWh flywheel with graphite fiber rotor. (Courtesy of Boeing.)

direction but relatively weak in the radial direction. Radial stress is the result of the tendency of the faster moving outer layers to expand more under centrifugal force than the slower moving inner layers. Radial strength can be increased by introducing fibers oriented in the radial direction in a small fraction of the rotor volume.

The radial growth problem also manifests itself in designs that position bearings and power conversion equipment on shafts or hubs at relatively small radii. Connection to a rim with a much larger radius must be made in a manner that takes into account the difference in radial expansion. This is often accomplished by using spokes, flexibly shaped hubs, or connectors that accommodate the radial expansion with a contraction along the axial direction of rotation.

With composite rotors, the kinetic energy in a (large) frisbee-sized flywheel would be approximately 1 kWh, and a person-sized flywheel could store 20 to 40 kWh. Metallic rotors are typically limited to rim velocities of 300 to 500 m/s, whereas composite rotors can easily achieve speeds above 1000 m/s. The highest reported rim velocity at the time of this writing was 2087 m/s. A commonly used rotor design calls for lower strength, lower cost glass fibers for the inner section of a wheel and higher strength, higher cost graphite fibers for the outer section.

It is desirable that the rotor remain rigid over all of its operating speed. Bending modes of the rotor may become resonant vibrations if the rotational speed is at the same frequency as the bending mode. Fibers may be strategically placed along the axial direction of the rotor or in some other direction to change the frequency of a bending mode.

Flywheel design must also take into account the influence of rotor dynamics. A prominent example of this influence is the use of flywheels as mechanical batteries on vehicles. A spinning flywheel has a gyroscopic moment that is proportional to the rotational velocity and acts to keep the axis of rotation always in the same direction. If the vehicle changes direction in a way that changes the axis of rotation, a large torque can be exerted on the vehicle. To overcome this tendency, the flywheel can be suspended in a special gimbaled mounting, or two counter-rotating flywheels can be paired so that the net gyroscopic moment is zero.

3. BEARINGS

The bearings connect the rotating part of the flywheel with a stationary housing and must support the weight of the rotor, keep it in position, and allow it to rotate freely. These functions must accommodate any motion associated with both the normal rotation of the flywheel and any change in balance of the rotating components. The bearings must resist any external influence that exerts a torque on the wheel that is resisted by the wheel's gyroscopic moment. Bearings may also be divided into thrust bearings and journal bearings. A thrust bearing provides most of its force along the axis of rotation, whereas a journal bearing provides most of its force perpendicular to the axis of rotation.

The bearings of a flywheel must provide stability against motion in any of the 5 degrees of freedom associated with a rotating rigid body. From a mechanical viewpoint, the bearings must provide levitation force, stiffness, and damping. The bearings provide stiffness, which may be thought of in terms of a set of springs that connect the rotor to a massive foundation. A suspended system of this sort will exhibit resonant frequencies at which oscillation amplitudes can be large if an exciting force exists at this frequency. If the rotor is rotating at a speed close to one of these resonant frequencies, it can excite this oscillation. Such rotational speeds are called critical speeds or criticals. The three fundamental modes of oscillation for a rotor with an axis of rotation in the vertical direction are shown in Fig. 2. The oscillation along the axis of rotation (vertical mode in Fig. 2) is generally not harmful. However, a mode transverse to this axis (radial mode) can lead to large imbalance forces as the center of mass rotates at some radius about the system's geometrical center, defined by the bearings. This imbalance force increases with oscillation amplitude and with the square of the rotational frequency. If the criticals are insufficiently damped, the imbalance forces may be large enough

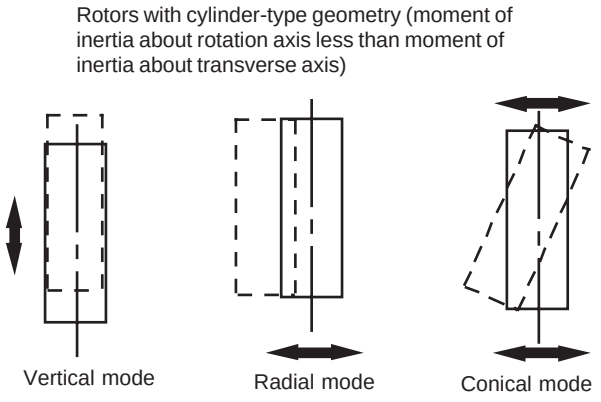


FIGURE 2 Schematic of rigid body critical modes for rotor with vertical rotational axis.

TABLE I

Influence of Bearing Type on the Design and Performance of Flywheels with Stored Energy in the 1- to 10-kWh Range

Bearing type	Levitation pressure (MPa)	Intrinsic loss fraction of KE_{max}/h	Parasitic loss fraction of KE_{max}/h	Total loss fraction of KE_{max}/h	Criticals	Sensitive to rotor imbalance
Mechanical	>10	0.1	0.02	0.12	Design out of operating range	Most
Magnetic	0.6	<0.001	0.01	0.01	Tunable	Some
Superconducting	0.3	<0.0001	0.001	0.001	Far below operating range	Least

to overcome the restoring force of the bearings and destroy the flywheel.

If the rotor runs in vacuum, some of the bearing components must run in vacuum. Typically, an auxiliary water cooling system is used for mechanical bearings and active magnetic bearings (AMBs). Superconducting bearings require a cryogen. The energy that is needed to operate the refrigerator of the pumps for the coolant is a parasitic loss.

Three general types of bearings are used in flywheels: mechanical, magnetic, and superconducting. Table I provides information for an approximate comparison of the influence of bearing type on flywheel design and performance of flywheels with stored energy in the 1- to 10-kWh range.

Table I shows that the losses of mechanical bearings render these bearings suitable for FES for storage times on the order of minutes to perhaps an hour. Losses of magnetic bearings render these bearings suitable for storage times up to perhaps a day. Superconducting bearings are suitable for FES with storage times up to many days.

FES systems with mechanical bearings have been sold commercially. Prototypes of FES devices that contain magnetic bearings and superconducting bearings are being field-tested. Development continues in all of these bearing technologies.

3.1 Mechanical Bearings

Mechanical bearings can be divided into several subcategories: rolling-element bearings, hydrodynamic journal bearings, gas bearings, squeeze-film bearings, and hydrostatic bearings.

Rolling-element bearings are probably the most common type of bearing used in machinery. They are relatively compact, can transmit heavy loads of various forms, and can be installed and serviced easily. Several specific design types with a range of stiffness are available. Damping in rolling-element bearings is very small but may be enhanced by squeeze-films.

Journal bearings consist of a circular section of shaft (the journal) rotating inside a bearing "bush," which is nominally circular. The clearance space between the two is partially filled by the lubricating fluid, which is pressurized by the motion. In addition to reducing the friction in rotational motion, this hydrodynamic film behaves like a complicated arrangement of springs and dampers and so influences the critical speeds and imbalance response. If designed improperly, the film can contribute to rotor instability, which is the result of self-excited vibration. Hydrostatic bearings differ from hydrodynamic bearings in that the lubricant pressure required to separate the bearing surfaces is supplied from some external pressure source and is not a result of journal rotation.

Gas bearings are similar to oil-lubricated bearings, but because the gas is compressible, the behavior differs. Compared with hydrodynamic bearings of similar surface area, gas bearings carry much smaller loads and generally operate with smaller film clearances. The advantage of gas bearings is that they tend to have much smaller rotational losses because the viscosity of a gas is much smaller than that of a liquid. Several types of gas bearings are available: self-acting, externally pressurized, porous wall, and flexible vane.

Squeeze-film bearings are a special case of hydrodynamic bearings. This type of bearing consists of a conventional rolling-element bearing and an external structure for damping radial motion. The outer race of the rolling-element bearing is surrounded by a damper casing. A small radial clearance between the outer race and the inner bore of the casing accommodates a thin lubricating film. The fit is so close that the outer race of the conventional bearing does not rotate. However, the squeezed film is pushed from one region to another by the radial force of the

outer race. The lubrication pressure is built up purely by the action of squeezing the film in this clearance. Friction between the lubricant and the races, and inertia in the lubricant itself, produces a radial stiffness.

If the rotor of the flywheel is rotating in air, any of the preceding bearings can be used. If the rotor operates in vacuum, the number of bearing choices is limited and the vapor pressure of the lubricant must be low. Solid-lubricated ceramic bearings (e.g., those made of silicon nitride [SiN] and lubricated with molybdenum disulfide or other proprietary solid films) are often used in this situation.

3.2 Active Magnetic Bearings

In AMBs, levitation is achieved by the attractive force between an energized magnet and a ferromagnetic body. Typically, the magnet is a stationary electromagnet and the rotor is a ferromagnet. Such a system is inherently unstable, so active feedback is used to modulate the field of the electromagnet depending on the position of the rotor. Such bearings have been used for decades in industrial applications and have been used recently in flywheel prototypes.

In the basic setup of an AMB, position signals from gap sensors are used by a controller power amplifier to set the appropriate currents and voltages of the electromagnets in such a way that stable levitation takes place. In the example of an iron rotor below a ferromagnet, if the rotor starts to fall so that the gap becomes too large, the current in the magnet is increased to increase the attractive force. Similarly, if the gap becomes too small, the current is decreased. Control is effective on bearing activation and continues for all rotational speeds.

A distinctive advantage of the AMB is that it can adapt itself to operating conditions and can communicate with its environment. Within certain limits, the AMB can be set to operate with arbitrary stiffness and damping, independent of the rotor position. The air gap (i.e., the position of the rotor) can be controlled in real time, independently of the stiffness. As a consequence, it is possible to maintain a fixed rotor equilibrium position for various loads. A rigid rotor can be made to rotate about its principal axis without transmitting vibrations to the foundation. By changing the stiffness and damping, the rotor is capable of easily crossing the critical speeds of the bearings. In addition, bending vibrations of an elastic rotor can be significantly reduced through an appropriately designed active control loop. Damping and stiffness are usually mutually dependent.

Early AMBs were powered by analogue controllers; however, the tendency now is toward digital controllers. The design of early AMB systems was based on proportional–integral–differential control algorithms. In general, the feedback law is chosen to be linear and is developed on the basis of the system linearized about the static equilibrium position. Recent investigations have shown that there is performance merit in the use of certain nonlinear feedback laws.

Typically, switching amplifiers are used to power the coils. These amplifiers use transistor switches that connect the coils to a DC voltage. The transistors are either on or off and produce on the coil a square wave voltage, which yields a sawtooth current. By controlling the on and off times, most current waveforms can be produced in the coils. The disadvantage of the switching amplifiers is that the oscillations in the current cause remagnetization loss in the magnetic bearing. However, the shorter the switching period, the weaker are these oscillations.

One of the fundamental limitations of an AMB is that the bearing force slew rate is limited because the magnet coils are inductive and the power supply to the driving amplifier is subject to a maximum voltage. The slew rate limitation causes the bearing force to change more slowly than the control signal demands and, thus, introduces a phase lag. Another way in which to state this effect is that the inductance of the coils and the maximum voltage provide an upper limit to the stiffness of the system.

The position sensor typically uses a high-frequency carrier and provides a linear signal versus the rotor location. Optical sensors, inductive displacement sensors, capacitance sensors, Hall effect sensors, and eddy current sensors are commonly used. The sensors must be contact free and able to measure a rotating surface. The surface quality of the rotor and the homogeneity of the material at the sensor will influence the measurements. A bad surface will produce noise disturbances, and geometry errors may cause disturbances with the rotational frequency or with multiples thereof. “Training” of the sensors in place may alleviate some of these noise problems. In this technique, the rotor is rotated at some fixed speed, the surface is measured over many revolutions, and the average of the measurements for a given set of angles is stored in a computer memory so that the sensor “remembers” the surface. The data set serves as a background that is subtracted from the measurement during regular operation.

A new type of AMB, which is under development, does not use sensors. In place of the usual current

amplifier, a voltage amplifier is used. The current is measured in the amplifier, and a control algorithm can be developed so that the rotor can be stabilized from this measurement.

In principle, the current requirements in the coils of an AMB can be made arbitrarily small if resolution of the sensors is high enough and the sensor and power electronics are fast enough to respond. That is, it takes less force (and therefore less current) to return a rotor from a 1-micron deviation from equilibrium than it does to return it from a 1-mm excursion. As sensors become faster and their resolution becomes higher, and as power electronics becomes faster, one may expect that the losses in AMBs will decrease. At some point, the homogeneity of the rotor and the magnet system will provide the determining limit of these losses.

3.3 Superconducting Bearings

In its simplest form, a superconducting levitation system includes a permanent magnet (PM) levitated in a stable position over a superconductor. Because the superconductor is not a magnet and exhibits special properties, its presence enables the system to bypass Earnshaw's theorem, a general theorem in magnetics that states that any magnetic system will be statically unstable. One of the fundamental properties of superconductors is their tendency to exclude magnetic flux from their interiors. This exclusion of magnetic flux, called the Meissner effect, makes them behave like a strong diamagnet. Accordingly, a superconductor with a PM positioned close above it develops a shielding current, which excludes flux in such a way that the actual magnet "sees" its mirror image.

More specifically, if the magnetization of the PM is in the vertical direction with its north pole down, the image will also be vertically magnetized but with its north pole up, exerting a repulsive force on the real magnet. The closer the magnet gets to the superconductor, the stronger the repulsive force; the farther away the magnet, the weaker the force. This setup can yield a levitation that is stable in the vertical direction.

The levitation phenomenon that is the result of the diamagnetic response is greatly enhanced by additional phenomena that are a result of flux pinning. A flux pinning center is a nonsuperconducting region such as an inclusion, a crack, or some other crystalline defect. Because the superconducting region that surrounds the nonsuperconducting center is strongly inclined to exclude magnetic flux, a magnetic flux line through the center tends to become trapped (pinned) there. With a sufficient number of flux lines trapped in the superconductor, the PM will remain stably levitated in position, even over a flat surface. The flux lines between the PM and the superconductor act in an imperfect analogy to mechanical springs with attachments on the PM and in the superconductor. If the magnet is moved vertically or horizontally, the "springs" will tend to pull the magnet back to its equilibrium position. If the flux pinning is sufficiently strong, the magnet may be stably suspended below the superconductor or even along its side.

The potential for superconductor levitation in various applications received greatly increased attention with the discovery of high-temperature superconductors (HTSs) that exhibited critical temperatures exceeding the boiling point of nitrogen. The current material of choice for superconducting bearings is Y–Ba–Cu–O (YBCO). This choice is due to its large superconducting current density at liquid nitrogen temperatures and the ability to grow large grains. In bulk YBCO, the grains can grow to diameters of several centimeters when made by a process called melt texturing. In the current state-of-the-art, the upper limit to this process appears to be a diameter of approximately 10 cm.

If the suspended PM is pulled hard enough to the side or vertically, it is possible to move one or more of the trapped flux lines into new pinning centers and so change the equilibrium position. Such a change leads to hysteresis in the levitation force and an associated energy loss if it occurs in a cyclic pattern.

So long as the magnetic flux distribution in the superconductor does not change, rotation of the PM encounters no resistance. In reality, hysteresis loss in the superconductor will act to decrease the rotation rate. If one picks a radius on the surface of the PM and travels in a circle (azimuthally) on the surface at that radius, one sees a large DC magnetic field with a small AC component. The AC component produces a changing magnetic field at the surface of the superconductor and produces a hysteresis loss. The ideal PM would exhibit a magnetization that is perfectly symmetric about its axis of rotation; that is, it would exhibit a high degree of azimuthal homogeneity. Real PMs contain defects that are a result of the manufacturing process. These defects are mainly a misorientation in individual grains of the easy magnetization from the intended direction. Typically, in PMs with the best homogeneities, at a fixed radius above the rotating surface, the amplitude of the AC component of the magnetic field is of the order of 1%

of the average field at that radius. This inhomogeneity, although small, is sufficient to cause a detectable decay in rotation rate when the magnet spins in vacuum.

Even if the PM were perfectly homogenous, hysteresis loss would still occur due to horizontal and vertical vibrations that are transmitted to the system from the outside world. However, in practical superconducting bearings, the loss due to PM inhomogeneity is much larger than that due to vibration transmission.

As discussed in section II, the bearing gives rise to oscillation modes with corresponding critical frequencies associated with these resonances, with the most important being radial resonance. (If the rotor is a long cylinder, conical resonance is also important, but the discussion is essentially identical to that for radial resonance.) Accordingly, one may divide the loss behavior into three regions: below the resonance, at the radial resonance, and above the resonance. Below the resonance, the losses are due to the inhomogeneity of the magnetic field of the PM. The resonance region typically occurs at low frequency, that is, much lower than the operating range of FES systems. Above the resonance, the losses are influenced by an additional factor, caused by the rotation of the magnet about its center of mass rather than about its center of geometry or center of magnetism. Above and below the resonance, the losses are velocity independent.

Because of the size limitation imposed during production of high-performance bulk superconductors, a large bearing system will need an array of superconductors. In the case of a single magnet levitated over an array of superconductors, an additional loss term arises due to the magnetization of the individual superconductors on levitation. The magnetization of the array leads to an AC magnetic field that is seen by the rotating PM and results in eddy current losses that occur in the PM and depend on the electrical conductivity of the PM.

The low stiffness of the HTS bearing has profound consequences in rotor dynamic behavior of flywheels that is not exhibited with other types of bearings. Rigid rotor bearing resonances, due to the magneto-mechanical interactions in the PM/HTS system, occur at very low frequencies that are on the order of several hertz. Thus, the operating speed regime of disk or ring-type rotors is in the supercritical range, that is, well above the bearing resonant frequencies. Therefore, the operating speed range between the bearing resonances and the frequencies at which bending modes start to appear in the rotor structure is usually large. Furthermore, the low-stiffness bearing allows the rotor to self-center at low speeds, with relatively low out-of-balance forces at these resonances, so that rotor balancing is less critical than with other bearing designs. The HTS bearings often contain large gaps between rotating and stationary surfaces and can tolerate radial oscillations of several millimeters in amplitude.

3.4 Magnetic Bias

In any of the bearing systems discussed in the preceding sections, considerable advantage in terms of reduced losses can be achieved by using a magnetic bias to take up most of the bearing load. For example, augmentation may take the form of an Evershed-type design, in which a PM pair, one on the stator and one on the rotor, are in attractive levitation (Fig. 3). The spacing of this magnet pair is adjusted so that, at the equilibrium position, the attractive force is between

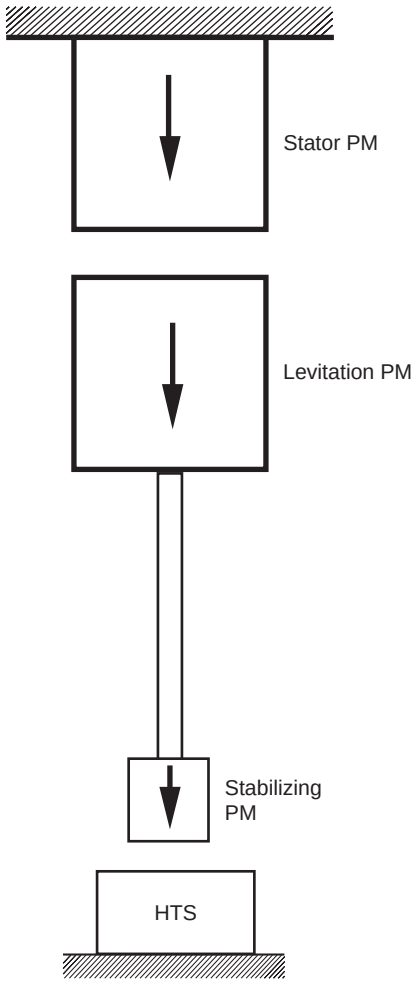


FIGURE 3 Evershed-type magnetic bearing with HTS bearing providing stability.

99 and 100% of the rotor weight. With this magnet pair, the system is stable radially but unstable vertically. An auxiliary bearing system (mechanical, active magnetic, or superconducting) is used to provide the remaining less than 1% of the rotor weight plus the required stiffness. The rotational losses of a PM pair generally consist of eddy currents induced in one PM by the azimuthal field inhomogeneity of the other PM. These losses are typically smaller than those associated with the auxiliary bearing that makes up the stabilizing part of the system. Low-loss magnetic biasing can also be accomplished with a PM and ferromagnet combination.

Because superconducting bearings provide a stabilizing stiffness in all directions, the magnetic bias can be designed with low stiffness. Because an AMB typically provides stable stiffness in only one direction, it may be appropriate to design the magnetic bias with a rather large stiffness that provides stability in all directions but one. Then, only one AMB controller is required to stabilize against movement in the unstable direction.

4. POWER INPUT AND OUTPUT

When the flywheel is used as a mechanical battery, energy must be converted from rotational kinetic energy to usable electrical power and vice versa. This conversion usually involves two major components: an M/G and power electronics. The M/G converts energy between mechanical and variable-frequency electrical. Power electronics then converts between variable-frequency AC and DC electricity. In some cases, the DC is the output of the flywheel. More commonly, additional power electronics converts between the DC and either the external AC electrical grid or a critical load.

Many types of M/Gs are possible. In one commercial flywheel design, the flywheel is on a coaxial shaft that is separate from, and spins faster than, the main generator. When power is required, a clutch is engaged to transfer torque to the constant-speed generator. The motor and generator could be separate components, but typically the M/G is one unit, with the function of operating as motor or generator determined by the power electronics.

A common design for the motor is the brushless synchronous type. In this design, one or more PMs, attached to the rotating part of the flywheel, move in close proximity to stator coils. The power electronics equipment supplies a voltage to the stator coils that is timed synchronously with the rotational angle of the rotor. When operated as a generator, the rotating PMs induce a voltage on the stator coils, and the power electronics regulate how much current flows. Both the output frequency and the output voltage of the generator depend on the angular velocity of the flywheel; both increase as velocity increases.

The primary function of the M/G is to provide torque to the flywheel. An undesirable side effect in most M/G designs is the production of radial and axial forces that tend to destabilize the flywheel. These forces must be accommodated by the bearings, and for magnetic or superconducting bearings, which are of inherently low electromechanical stiffness, the destabilizing tendency can be significant. Special M/G designs, with slightly lower specific power, have been used in flywheels with these bearings. In one method, a toothless back iron is used in the stator coils. A second method uses a special PM geometry called a Halbach array, which produces an internal dipole magnetic field and very little radial or axial force, even when the rotor part of the M/G is not perfectly centered relative to the ironless stator coils.

Progress in power electronics, particularly in high-power insulated gate bipolar transistors and field effect transistors, underlies higher power flywheel operation. Variable-speed motor drives can be used for the power input section. Several motor operating modes are possible, with a common one providing fixed-voltage pulses to the stator coils at various pulse widths. As the flywheel speed increases and the back electromotive force (EMF) on the stator coils builds, the pulse lengths become longer. Because the generator output voltage decreases with decreasing speed, there is a lower limit on the speed at which power can be withdrawn at the rated value. As speed and voltage decrease, if power stays the same, the current must increase. This means the power electronics must be designed for the lowest operating speed of the flywheel, leading to larger equipment than would otherwise be required for the higher speeds. As an example, a flywheel might be designed to provide full power anywhere from maximum to half-maximum speed. Because stored energy is proportional to the square of the velocity, the available stored energy would be only 75% of the maximum rotational kinetic energy.

5. CONTAINMENT

Flywheels are often contained in some type of housing to maintain the flywheel in a vacuum and to provide personal safety from contact or, in the case

of failure, from debris. Early flywheels were not enclosed, and this is true today of some modern low-performance flywheels such as those used in presses.

Housing flywheels in evacuated containers reduce aerodynamic drag, with the level of vacuum being determined by whatever value of windage loss is tolerable in a particular application. Today's systems typically operate somewhere between atmospheric pressure (100 kPa) and 1 mPa, with the highest speed devices operating at the lowest pressure. Lower speed flywheels are larger but can be operated in air with acceptable losses. Reduction of aerodynamic drag is important for high-performance flywheels not only to reduce the loss rate of stored energy but also to prevent the flywheels from overheating.

A potential parasitic loss is the power required to run the vacuum pumps. It is believed that the rotor containment chambers of commercial versions of high-performance flywheel systems will be exhausted and then hermetically sealed. Vacuum will be maintained over time by a getter system such as barium or activated charcoal. Considerable experience in this technology has been gained from the fabrication of vacuum-insulated containers such as thermos bottles and superconducting magnet dewars. If the rotor is metal, few problems should be encountered in accomplishing this. A properly cured composite should exhibit little outgassing at ambient temperature, but this remains a technical issue for high-performance rotors.

Good design makes the probability of a flywheel failure extremely small. However, for high-performance flywheels, where operating experience is still limited, a safety containment structure is often part of the vacuum housing. For systems that use magnetic bearings, the containment often includes a set of mechanical touchdown bearings. The safety containment of a flywheel is otherwise similar to that of aircraft gas turbines. It must avoid any fragment penetrations and avoid excess deformation; limit the forces, moments, and vibrations transferred to the outside; and avoid the emission of dangerous or pollutant substances.

When a flywheel breaks up, the total angular momentum of the debris cloud remains the same as before the breakup, and this moment must be accommodated by the containment structure. The motion of individual debris fragments is translational and rotational. For very small particles, the energy is nearly entirely translational. Composite material rims usually produce many tiny fragments. Depending on the type of failure, they may take the form of long needle-like shards or small grains. Any PM or metallic component usually forms larger fragments in the debris cloud. These fragments may provide strong local loading when they hit the containment shell.

Containment structures typically employ a suite of strategies. The energy of the particle pieces is often absorbed in plastic deformation of a collapsible structure inside the containment wall. Turnaround rings and deflection shields are often used to redirect particle momentum. Rotatable liners are sometimes employed to accommodate the angular momentum of the debris cloud.

6. APPLICATIONS

Flywheels have been used in numerous applications. They have provided the prime energy for turning the propellers of torpedoes and for numerous transportation applications in electric buses and railways, both onboard and as wayside power providers. They have been used in mine locomotives where explosion risk is present, and they have been used to power city buses. Flywheels are under consideration to replace chemical batteries in the International Space Station and for use as combined energy storage and momentum wheels in earth-orbiting satellites. Tiny flywheels are used as the universal prime movers in some toys. Flywheels with very high discharge rates, called compulsators, have been used to power electromagnetic rail guns.

Flywheels continue to be used in all reciprocating engines and in all machines that require very high power for a small part of their working cycles. All mechanical presses use flywheels, usually of the disk or rim type; these flywheels are quite massive for the quantity of energy they store. The ability of flywheels to deliver very high power for a short time has been used in inertial friction welding and in some machines such as inertial starters. Other applications of flywheels can be found in devices that must maintain their angular velocity in case of a power failure or during the shutoff of a plant; typical examples include lubrication or cooling pumps and emergency systems.

Flywheels used as mechanical batteries are often integral parts of hybrid electric vehicles. Here, the flywheel is used for times when high power is required, thereby allowing the internal combustion engine to run at a fixed rate where efficiency can be optimized and exhaust pollutants can be minimized. The engine runs an alternator that continuously charges the flywheel. The vehicle uses both an

internal combustion engine and an electric motor for propulsion. The basic idea is that the average power needed to propel the vehicle should be supplied by the engine, which then can operate at a nearly constant optimal speed, thereby reducing fuel consumption, air and noise pollution, and engine maintenance requirements as well as extending engine life. Short bursts of power, for climbing hills and acceleration, are taken from an energy store (a battery of one sort or another) that is replenished directly by the engine or by regenerative braking when the vehicle is slowed down.

Electric utilities have a great need for inexpensive energy storage systems, such as flywheels, because their inexpensive baseload capacity is typically underused at night and they must use expensive generating sources to meet their peak loads during the day. The availability of a distributed network of diurnal storage could also make use of underused capacity in transmission lines at night as well as add robustness to the electric grid. Efficient energy storage would also be beneficial to renewable energy technologies such as photovoltaics and wind turbines.

The most practical FES option assumes arrays of many individual flywheel units. Although FES units with stored energy as low as 2 to 10 kWh may have some applicability, electric utilities are most likely to desire units with approximately 5 MWh of storage capacity. Units with such capacity, composed of modules of individual 10-kWh flywheels, are approximately the largest that can be factory assembled and truck mounted for delivery to substations or energy storage facilities. With a fleet of FES units that handle real-time automatic grid control, a utility could manage generation variability as well as load variability. Such units could also provide uninterruptible power for factories.

SEE ALSO THE FOLLOWING ARTICLES

Batteries, Overview • *Electrical Energy and Power* • *Electric Motors* • *Hybrid Energy Systems* • *Mechanical Energy* • *Storage of Energy, Overview*

Further Reading

Ehrich, F. F. (1992). "Handbook of Rotordynamics." McGraw–Hill, New York.

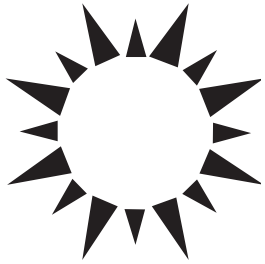
Genta, G. (1985). "Flywheel Energy Storage." Butterworths, London.

Goodwin, M. J. (1989). "Dynamics of Rotor-Bearing Systems." Unwyn Hyman, London.

Hull, J. (2000). Superconducting Bearings. *Supercond. Sci. Technol.* **13,** R1–R14.

Schweitzer, G., Bleuler, H., and Traxler, A. (1994). "Active Magnetic Bearings." Hochschulverlag, Zurich, Switzerland.

Food Capture, Energy Costs of



JERRY V. MEAD
State University of New York, College of Environmental Science and Forestry
Syracuse, New York, United States

1. Introduction
2. Bioenergetics
3. Food Capture, Maximum Power, and Evolution
4. Environmental Impact on Energy Cost of Food Capture
5. Feeding Strategies and the Costs of Food Capture
6. Summary

Glossary

abiotic Pertaining to nonliving environmental factors, including light, temperature, water, nutrients, soil, and atmospheric gases.
assimilation efficiency The fraction of ingested food that an organism is able to absorb.
bioenergetics The study of the flow and transformation of energy within and between living organisms and their environment.
biotic Of or having to do with life or living organisms.
ectotherm An organism that regulates its body temperature largely by exchanging heat with its surroundings.
endotherm An organism that generates heat from food energy to maintain its body temperature, typically above the temperature of its surroundings.
homeotherm An organism, such as a mammal or bird, that maintains a constant body temperature, largely independent of the temperature of its surroundings.
maximum power hypothesis A concept stating that there will be selection for rate of energy use that yields the most useful work per unit time.
poikilotherm An organism, such as a fish or reptile, having a body temperature that varies with the temperature of its surroundings.

The amount of energy expended by an organism to collect a unit of energy is referred to as the energy costs of food capture. The ability of a species to capture and utilize energy is key to its survival. Therefore, all organisms expend energy to maintain basic body functions and require additional energy to feed, grow, and reproduce. The energy costs for capturing food include energy expenditure for finding, handling, digesting, and assimilating the food. Both biotic and abiotic factors influence the profitability of a given method of feeding, and the energy gained from feeding minus the energy costs of capturing that energy is an organism's net gain of energy that can be used to increase the fitness of a population.

1. INTRODUCTION

The influence of the environment on the success of food capture can be mediated by behavioral and physiological adaptations. However, such adaptations are not made by the organism, but are products of natural selection. The theory of evolution through the process of natural selection states that traits in organisms that produce the most descendants have been selected for over time. For an organism to create descendants, an organism must have a surplus of energy to invest in reproduction and/or the rearing of offspring. In a sense, organisms are players in a game in which there are fees, in terms of energy and nutrients, to remain a player and additional fees that must be invested into activities that may win a player surplus energy. Surplus energy can then be used to perpetuate their genes. In this game, the rules are determined by physical, chemical, and biological processes. Successful organisms are those that, by making profitable investments, acquire surplus energy per unit time of playing the game and then use this energy to perpectuate their genes.

Why are there so many species on Earth? This is one of the most fascinating questions asked by ecologists. The main reason for the richness of species may be that a diverse set of organisms is more effective than a limited set is at capturing available food energies. This idea is illustrated in the following discussions by describing the current understanding, for both plants and animals, of the

 © 2004 Elsevier Inc. All rights reserved.

energy costs of maintaining life and capturing food. The focus is on net energy gain, although it is recognized that resources other than energy, such as nutrient availability, can influence the profitability of a foraging strategy. How the biota minimize the energy costs of food capture using physiological and behavioral adaptations is also described.

2. BIOENERGETICS

Understanding the costs related to food capture requires a basic knowledge of an organism's demand for energy and the profitability of captured energy in terms of growth and reproduction. Bioenergetics is the study of the flow and transformation of energies used within and by living systems. A thorough review of bioenergetics is not the purpose in this section; instead, the focus is on an organism's basic needs for energy and how these basic needs relate to the environment, followed by a discussion of how an organism can invest its energy into capturing food to yield a net profit.

2.1 Energy Flow in Organisms

The pathways of energy flow in biota are similar among species. Species vary primarily in terms of rate or efficiency of processes, method of food capture, and types of foods consumed. Once food has been collected, it is defined as consumed. A common misconception is that everything consumed enters an organism. Sometimes an organism loses pieces or fluids of the food being consumed before it is utilized. The portion of food consumed that passes into an organism is defined as ingested material. Ingested food then passes through the internal energy pathways of the organism.

Ingested food must be absorbed or assimilated into the organism for food to be useful to do work. For plants, light energy that has excited a photosynthetic pigment is converted into internal available energy. In animals, food in the digestive system is assimilated through absorption across the digestive tract. Saprophytic organisms such as fungi assimilate foods by absorbing it directly into cells through diffusion. The efficiency of assimilation varies among biota and within species (Table I). These differences in efficiencies are due to many factors, such as rate of food ingestion (related to temperature, incidentally), food quality, and organism physiology. Energy that is not assimilated is lost through egestion in animals, through light reflectance and heat loss in plants, and through leaching in saprophytes. In animals, there is another loss associated with moving ingested materials through the body, i.e., specific dynamic action. Other organisms have smaller costs for internal transport of foods and rely mainly on diffusion and chemical processes.

Assimilated food energy is the actual quantity of energy available to fuel various life processes. In most cases, assimilated energy is used first to fuel metabolism. Total metabolism refers to the sum of energy used to rebuild and fuel the basic functions of life (basal metabolism) and the energy used to perform activities (active metabolism) such as swimming, flying, or walking.

2.2 Energy Costs and Demands for Food Capture

2.2.1 Energy Cost of Metabolism

The energy and time costs of capturing and finding food include both basal and active metabolism. Basal metabolism increases exponentially as a function of environmental temperature until the temperature reaches the optimal temperature for the organism and then declines with increasing temperature. Optimal temperatures for basal metabolic rate can differ among species and among individuals of a given type of species; this is primarily due to enzyme reactivity. In addition, basal metabolism decreases per unit body weight as body weight increases, and varies depending on whether an organism regulates its internal body temperature (Fig. 1). Time spent searching or waiting for food requires energy for basal metabolism. The increased physical energy used in active metabolism can be one of the largest costs in food capture and is difficult to measure or predict because it depends largely on behavior and, like basal metabolic rate, can be controlled by environmental conditions. For example, a tuna fish can increase its metabolic rate when foraging and when the tuna seeks prey in warmer surface waters.

2.2.2 Energy Costs for Locomotion

Time spent searching for food can be costly in energy for an individual. These costs can be large due to energy-intensive activities such as walking or the time required locating prey. The energy costs of various activities are often difficult to quantify. Scientists have developed elaborate techniques to measure active metabolism, including assessing metabolic tracers (e.g., oxygen), using sensors that measure muscle contractions, and analyzing movement using video equipment or mechanical devices

TABLE I

Assimilation Efficiencies for Poikilothermic and Homeothermic Species[a]

Species	Diet	Assimilation efficiency (%)	Production efficiency (%)
Aquatic poikilotherms			
Megalops cyprinoidea (fish)	Fish	92	35
Nanavax sp. (predatory gastropod)	Invertebrates	62	45
Hyallela azteca (amphipod)	Algae	15	15
Ferrissia rivularis (stream limpet)	Algae	8	19
Tegula funebralis (herbivorous gastropod)	Algae	70	24
Hydropsyche occidentalis (filter-feeding caddis fly)	Algae	55	10
Littorina sp. (periwinkle)	Detritus	45	14
Terrestrial poikilotherms			
Saltmarsh grasshopper	*Spartina*	27	37
Melanoplus species (grasshopper)	Lespedeza	37	37
Mites (Oribatei)	Detritus	25	22
Sugarbeet leaf (plant)	Light	16	—
Terrestrial homeotherms			
Least weasel	Mice	96	2.3
Little brown bat (*Myotis lucifugus*)	Insects	88	—
Marsh wren	Insects	70	0.5
Grasshopper mouse	Insects	78	5.7
Kangaroo rat	Seeds	87	6
Vole	Vegetation	70	3
Jackrabbit sp. (*Lepus*)	Vegetation	52	5.5
Jackass penguin (*Spheniscus demersus*)	Anchovies	78	—
Cottontail rabbit (*Sylvilagus* sp.)	Vegetation	52	6
Cattle	Vegetation	38	11
African elephant	Vegetation	32	1.5

[a] Compiled by J. Mead using data from Colinvaux (1986) and other sources.

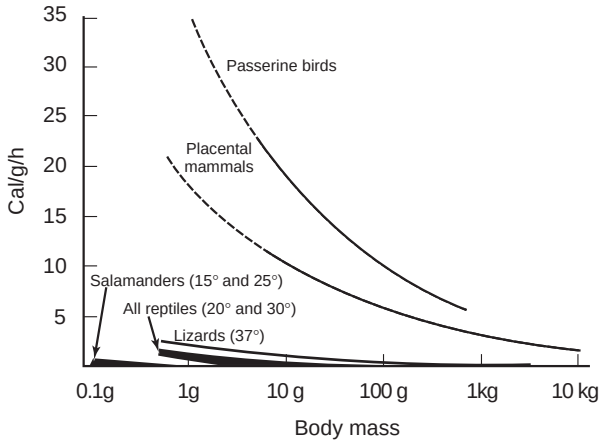


FIGURE 1 Resting metabolism as a function of body size and temperature (in °C). Ectothermic animals can be small because their energy consumption for maintenance is low. A bird weighing 1 g would use about 30 times the energy used by a reptile of the same weight. Not surprisingly, 1-g reptiles exist but 1-g birds do not. Reproduced with permission from Pough (1980), with permission from the University of Chicago Press.

(e.g., treadmills). For example, the field metabolic rate (FMR), the total metabolism for an organism in an environment, can be estimated for both terrestrial and amphibious animals by injecting "heavy" water into an animal's body. Heavy water has more neutrons than protons; the deuterium isotope, which replaces many of the hydrogen atoms in the water molecules, is detectable using laboratory techniques. Because the metabolic processes of most plants and animals depend on water, the rate of metabolic processes such as basal and active metabolism can be quantified by tracing the rate of uptake and release of the heavy deuterium isotope in the injected water. The field metabolic rate for many organisms has been estimated (Table II).

Movement is often the largest energy cost of food capture. Swimming, flying, and running are the basic types of movement. The energy efficiency of movement (distance traveled/unit weight/unit energy expended) differs with the medium that the

TABLE II

Measured Field Metabolic Rate and Energy Cost of Food Capture for Selected Ectotherms and Endotherms[a]

Species	Locomotion used	Feeding strategy	FMR[b] (kJ/g/day)	Cost of food capture (kJ captured:kJ cost)
Ectotherms				
Alder tree (*Alnus viridis*)	Nonmotile	Capture light energy	0.0005	10:1
Norway Spruce tree (*Picea abies*)	Nonmotile	Capture light energy	0.00275	9.1:1
Mussels	Nonmotile	Filter feeder		5:1
Komodo dragon (*Varanus komodoensis*)	Walking	Sit-and-wait predator	0.054	
Sand lizard (*Eremia lineoocellata*)	Walking	Sit-and-wait predator	0.166	1.6:1
Black and yellow sand lizard (*Eremia lugubris*)	Walking	Stalking predator	0.210	2:1
Endotherms				
Anna's hummingbrid (*Calypte anna*)	Flying	Collector	7.1	1.7:1
Little brown bat (*Myotis lucifugus*)	Flying	Stalking predator	2.2	1.9:1
Jackass penguin (*Spheniscus demersus*), breeding	Swimming	Stalking predator	0.64	2.1:1
Jackass penguin (*Spheniscus demersus*), nonbreeding	Swimming	Stalking predator	0.64	6.8:1
Commercial fishing boat (shrimp, scallops, herring, cod)	Propeller	Filter feeding		0.03:1–0.14:1
Modern agriculture	Tractor	Omnivore	5.0–60	1.4:1–60:1
Humans, foraging	Walking	Omnivore	0.001	3:1–6:1
Humans, pastoralism	Walking	Omnivore	0.01	3:1–5:1
Shifting agriculture	Walking	Omnivore	0.4–1.5	17:1–25:1
Traditional farming	Walking	Omnivore	0.5–2.0	17:1–20:1

[a] Data compiled by J. Mead.
[b] FMR, field metabolic rate.

organism is traveling through (such as water or air) and the morphology, body weight, and speed of movement of the organism. The rate of movement that is most energy efficient is defined as the minimum cost of transport. Most organisms have a rate of movement that has the highest energy efficiency, but there are some exceptions (such as starlings), when energy efficiency increases with speed.

For each mode of locomotion, the energy cost of transport decreases per unit weight of organism with increasing body weight (Fig. 2). Large animals such as blue whales may expend much energy per unit time for movement, but the energy cost to move a gram of whale is relatively low compared to the energy cost of moving a gram of mouse the same distance. The relation between body weight and the cost of transport helps explain the profitability of movement in foraging energetics. The high cost of locomotion for small organisms is the reason why most migratory animals have body weights greater than 1 kg. It is simply too expensive for small animals to migrate long distances without additional forces such as wind and ocean currents. The present explanation for the increase in energy efficiency in locomotion with size is that, compared to smaller organisms, larger organisms have greater inertial forces that aid in locomotion, and frictional forces (such as gravity) are much less due to lower ratios of surface to body weight.

Swimming by fish is the most energy-efficient method of locomotion on Earth, followed by powered flight and then crawling and running (Fig. 2). Each species must have structures and physiologies suited to the type of environment in which they live. Swimming and flying are more

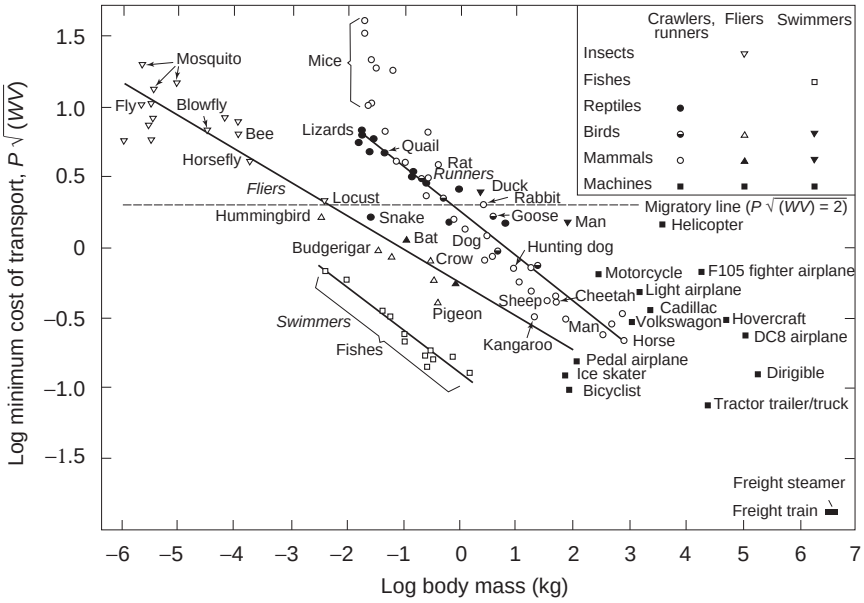


FIGURE 2 Energy costs of swimming, flying, and running, illustrating the energetic costs for searching for food. The data suggest that animals suited to swimming are more energy efficient at locomotion compared to flying and running species. Reproduced with permission from Tucker (1975).

efficient methods of movement than is running or walking because muscle efficiency is less in running or walking, compared other methods of transport. In running or walking, a muscle must contract and stretch to complete a cycle of movement such as swinging a leg. Contracting a muscle pulls on tendons that move parts of the body, and, in walking, muscles contract in the leg to push the body forward. Once the leg has contracted and performed work by pushing the body forward, the muscle must be stretched to maintain the body's center of gravity and to move the leg back into position for contraction. The energy expended to stretch a muscle is lost as heat in walking or running and performs negative work in that the stretch of the muscle must displace the appendage used in locomotion in a direction that is not the direction of travel. The main reason swimming and flying are a more energy-efficient means for locomotion compared to walking or running is that muscles are stretched less often. As a means for movement, however, swimming, not flying, is more energy efficient; until recently, scientists could explain only one-seventh of the difference in efficiency between these two methods of locomotion. Several aspects of swimming explain its efficiency: (1) muscles are typically stretched little in swimming; (2) most fish can regulate buoyancy using a swim bladder and/or their body density is similar to that of the surrounding medium, decreasing the energy used in maintaining a vertical position (flyers and runners must invest energy in maintaining their vertical position in the air); (3) vortices generated from the thrust of a foil, such as a tail, allow a swimmer to recover some of the energy lost in the movement of the tail; and (4) swimmers can utilize vortices produced by the environment, giving them the ability to achieve energy efficiencies greater than 100%. Machines such as submarines or boats have low energy efficiencies in movement compared to fish because their rigid hulls restrict the vessels' ability to utilize energy found in vortices in the environment. In addition, human mechanisms for converting mechanical energy into a jet or propulsive force do not generate vortices that are useful for recovering energy lost during thrust. Propeller-produced propulsion generates vortices that are perpendicular to the direction of movement and offer little potential for recovering energy. In this case, natural selection still seems to be a more efficient engineer than humans are.

Birds, bats, and insects are the only animals that have the ability for self-powered flight. Flying and swimming are similar in that lift and thrust are the two principal forces important to locomotion. Thrust is defined as the force applied by the organism to generate forward movement, whereas lift is the force exerted by an organism to change or maintain vertical position in a fluid. At a given body weight, flying is a less energy-efficient means of locomotion compared to swimming because more energy is

needed by flyers to generate lift. Both swimmers and flyers that use foils, such as wings and tails, generate vortices that can produce efficient jets of fluids. The jet produced by movement of a foil produces motion from thrust. Without a medium such as water or air, the movement of a tail or wing would not create a propulsive jet.

Walking and running organisms can recover some of the energy lost in stretching muscles by having elastic structures that store the energy used in stretching or by having many legs (e.g., centipedes), which reduces the energy expended in maintaining the center of gravity. The low energy efficiency of locomotion for small-bodied species has yet to be explained, but larger organisms are believed to have elastic tendons that can recover some of the energy lost in stretching muscles. As body size increases, animals have more tendon biomass to absorb energy used in stretching muscle. The recovery of energy used in stretching muscle is similar to the deformation of a bouncing ball. A ball bounces because the energy in the force of the ball hitting the floor is stored briefly in the deformed ball and then is released when the ball structure reforms, which pushes the ball away from the floor. Like a bouncing ball, biota can store in the animal's body some of the energy in the force of the animal's appendage hitting the ground, and then use this stored energy to maintain the center of gravity or for movement, by redirecting the energy into a force that can perform work.

The wheel has increased energy efficiency in locomotion from that of walking and running by reducing the number of muscles that are stretched. Imagine a person walking on a level surface and a person riding a bicycle next them; the walker's body dips during stride and then is pushed upward, but the bicyclist's body does not dip or bob during motion. A wheel maintains a person's center of gravity during stride and eliminates the bobbing action of a walker and the associated stretching of muscles. Each force applied to a pedal by bicyclist results in a perpendicular forward force using a gear. The ability of biota to perform certain types of locomotion has a large influence on the energy costs of finding and capturing food, but the environment sets the stage for the need for movement.

Once the energy cost of total metabolism has been paid, surplus energy and materials are used in foraging, growth, or reproduction. If surplus energy does not exist, life functions can be performed only by using the finite energy supply stored by the organism.

3. FOOD CAPTURE, MAXIMUM POWER, AND EVOLUTION

3.1 Optimal Foraging Theory and Maximum Power

Optimal foraging theory states that traits of an individual will be selected for that maximize the surplus energy gained per unit time from foraging. Optimal foraging theory is used to predict the most profitable feeding strategy for an organism given the energy costs for capturing food versus the energy gained from food capture. For example, as shown in Eq. (1), Holling's disk equation (so named because C. S. Holling used paper disks in his model of functional response) can be used to predict an optimal feeding strategy:

$$\frac{\text{Surplus energy}}{\Delta\text{Time}} = \frac{(\text{Energy gain} - \text{Cost of capture})}{(\text{Search time} + \text{Capture time})} \quad (1)$$

where surplus energy is food energy captured that is available for maintenance, activities, growth, or reproduction; energy gain is energy assimilated by capturing food; cost of capture is energy expended while searching for and capturing food; search time is time spent searching for food; and capture time is time spent capturing food. Equation (1) illustrates several components of a feeding strategy that can be altered to maximize the amount of surplus energy gained from feeding. An organism can reduce the time hence the energy costs to search for or capture food. In addition, the energy gained from consuming food can be maximized by selecting for the highest quality food items.

Optimal foraging theory fits neatly within a more universally applicable concept, Howard Odum's maximum power concept. The maximum power concept predicts that the optimal strategy would be to invest the amount of potential energy that yielded the largest quantity of useful work per unit time, which for organisms is surplus energy per unit time. An organism must invest an amount of potential energy into capturing food that yields the most work in terms of growth or reproduction per unit time, which may not be the most energy-efficient food capture strategy available to them.

3.2 Influence of Time and Space on Foraging Energetics

Factors that influence the cost of food capture can be spatially and temporally complex. Food supply may be patchy in distribution and these patches of food may move, shrink, or grow. In addition, physical conditions

may vary over space and time, influencing the cost of food capture. An organism can increase the profitability of a foraging strategy by investing energy and time into migration. For example, zooplankton production in the North Pacific Ocean changes over a season because the spatial and temporal variability of the environment affects the profitability of foraging strategy. Fish that remain in one location and consume the zooplankton may not maximize net energy gain because the food abundance shifts with the season. Even if animals find a good patch of food to feed on, it may be more profitable to move to a new location after a period of feeding than to remain in one place. Movement to new areas for foraging can be risky, in that it is uncertain whether an animal will find better foraging conditions, the animal may expose itself to predation risks, and there is an energy cost for locomotion.

Average-rate-maximizing foraging models use patch models to predict optimal feeding strategies by determining the most profitable investment of energy and time spent searching for food in a patch versus moving between patches. If the entire region available for foraging is thought of as a mosaic of patches of feeding conditions that vary in profitability, the mean profitability of patches can be used as the benchmark to determine the relative profitability of feeding in any given patch. Patch models predict the amount of time an organism should feed in a given patch by determining the profitability of foraging in a patch relative to the mean profitability of foraging in all patches and the amount of time required to move to a new patch. In conditions in which the relative profitability of a patch and the time required to move between patches are high, the model predicts that the forager will remain in the patch for a long period of time. Foragers are thought to move between patches often when foraging conditions are relatively poor when compared to other patches, and energy and time costs for travel are small. Foraging models aid in understanding the role of migration in foraging energetics, but these models are only formalizations of knowledge about these systems and hence are incomplete.

4. ENVIRONMENTAL IMPACT ON ENERGY COST OF FOOD CAPTURE

4.1 Environmental Factors Influencing the Energy Cost of Food Capture

Many factors in the environment can influence the energy cost of food capture, and understanding their individual effects is difficult because many factors interact. Abiotic and biotic conditions that influence foraging energetics include the quantity and quality of food, environmental temperature, light, fluid dynamics, nutrient availability, and risk of injury or mortality.

4.1.1 Quantity and Quality of Food

Quantity of food influences the amount of time and energy spent searching for food; food quality relates to the profitability of a food item for growth and reproduction. In many foraging systems, the amount of food available for foraging is not directly proportional to the energy cost of food capture. As food abundance increases, the amount of food captured per unit time reaches an asymptote (Fig. 3) and the energy costs of food capture decrease.

4.1.2 Environmental Temperature

Temperature is considered as the master regulator of metabolic rate. Knowing how a species responds to temperature is useful in understanding the ability of an organism to perform feeding activities and the associated energy costs. The response of an organism to environmental temperature is described as homeothermic, poikilothermic, ectothermic, or endothermic, based on the physiology of the organism. Homeotherms are species with internal body temperatures that vary little over time; poikilotherms have varying internal body temperatures. Humans are homeotherms that maintain internal body temperature at 98.6°F. Insects are examples of poikilotherms because their body temperatures may vary

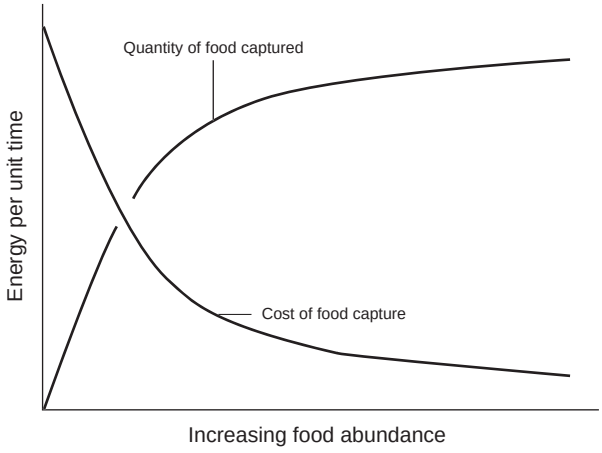


FIGURE 3 Energy cost of food capture as a function of food abundance, illustrating that in many feeding systems, the ability to capture food reaches an asymptote with increasing food supply.

greatly over time. Ectotherm body temperature is regulated by the external environment, whereas endotherms regulate their internal body temperature using internal energy. A homeotherm is a specialized endotherm that has a relatively constant internal body temperature, whereas endotherms do not necessarily need to have constant body temperatures. Each of these factors influences the efficacy and efficiency of energy capture.

Endotherms have a feeding advantage over ectotherms in cold climates, because the metabolism of ectotherms, and hence feeding, is restricted in a cold environment. Despite this restriction on feeding ability, ectotherms may not starve because the basal metabolic demand for energy is also less in cold weather. The endotherms have an advantage over ectotherms in these cold climates because they can generate internal body heat, increase their metabolism, and increase their foraging rate at an energy cost. Large size decreases heat loss through decreasing surface-to-volume ratios, thus polar bears in the Arctic reduce energy costs of maintaining elevated body temperature by being large and have higher metabolisms for foraging. A potential disadvantage to the homeothermic design is that, when compared to poikilotherms, it is a high-energy demand, low-energy efficiency system that requires more food for survival compared to a poikilotherm of similar body weight (Fig. 1; Tables I and II). If food is in low supply, the homeotherm may starve quickly because it must use its energy reserves to maintain an internal body temperature.

Robins (*Erithacus rubecula*) are a good example of the utility of increasing metabolism in a cold environment. The robin is homeothermic, but its prey (earthworms and insects) is both poikilothermic and ectothermic. The prey of the robin has a much lower total metabolism compared to a robin in a cold environment; the higher active metabolism of the robin enables it to catch prey easily because it can move much faster than the prey. The difference in active metabolism between prey and predator can decrease the predator's cost of food capture, as in the robin's case, or it can increase the cost of prey capture when the prey, compared to the predator, has much higher active metabolism.

The Komodo dragon (*Varanus komodoensis*) found in Indonesia is an example of a predator having difficulty running down and capturing its prey because it has a much lower active metabolism. The Komodo dragon preys on birds, rodents, and large mammals (such as deer). Deer, the primary prey item of dragons, are homeotherms, whereas the Komodo is a poikilotherm. It is difficult for the Komodo to run down a deer and physically overpower it because of the Komodo body structure and relatively low active metabolism compared to a deer. However, the dragon does have a good chance at injuring or subduing a deer through surprise attacks. Even if the prey escapes in a surprise attack, if the dragon has bitten the prey, the dragon often captures it eventually because the prey has been poisoned with a deadly bacteria (*Staphylococcus*) found in the Komodo's saliva.

4.1.3 Light

The quality and quantity of light affects the growth rates of plants and also the success of many foragers at capturing food. Light intensity influences prey detection, rate of photosynthesis, and indirectly several other factors that impact foraging systems. The amount of energy received from light per unit time is related to the frequency of photons that reach a surface. Frequency of light is inversely proportional to the wavelength traveled by photons. Wavelength also defines the type of light energy present in an environment. Wavelengths of visible light have certain colors. Light energy used for plant growth—photosynthetic active radiation (PAR)—ranges in wavelength from 400 to 700 nm. Red light, at a wavelength around 700 nm, has the lowest energy for photosynthesis. Blue light, at a wavelength of around 400 nm, is the highest energy PAR.

In aquatic communities, the amount of energy in various wavelengths determines how far light is transmitted through water and hence the availability of light to plants. If a photon of light is conceptualized as an arrow shot from a bow, the arrow that has the most initial energy, in momentum, will travel deepest into a body of water before frictional forces deplete the arrow's momentum. Thus green light is absorbed or reflected near the surface, whereas blue light, which has the most energy, will travel further than green light. The clarity of a gas volume or a body of water, which relates to the amount of materials in suspension, will influence how deeply light penetrates. As result of differences in the photons' ability to be transmitted through water, relative abundances of photons traveling across different wavelengths change with distance from the surface of a body of water.

Plants and fungi must invest energy to capture food energy by building structures and chemicals (e.g., chlorophyll). Adaptations in the structure of a forest canopy and in the composition of an ocean phytoplankton community appear to maximize surplus

energy captured from the sun. The structure of a forest canopy represents layer upon layer of vegetation in space whereby upper layers shade lower layers from sunlight. A fraction of the light intercepted by the overhanging canopy is reflected. Sunlight that penetrates into the forest canopy has a larger relative abundance of high-energy frequencies of light, but is lower in total energy compared to light at the top of the forest canopy, but still contains energy for photosynthesis. In a dense forest, the amount of leaves will increase per unit area the closer the canopy layer is to the forest floor, until the addition of leaf biomass does not yield a net gain from energy capture.

Natural selection has even used ultraviolet and infrared light for food capture. Rattlesnakes and blowflies are just two examples of the ways species sensing infrared or ultraviolet light can reduce the energy costs of finding food. Rattlesnakes are poikilotherms that have low active metabolisms in cold conditions (Table II). Thus, like the Komodo dragon, rattlesnakes have difficulty chasing and then subduing their prey and are more effective at capturing food as sit-and-wait predators that use poisons. The pit organs on the rattlesnake's head are receptors of infrared radiation that assist in locating food. Blowflies, in contrast, forage on carrion. Carrion is an ephemeral source of food and is patchy in distribution. Carrion is often a good source of energy and nutrients, creating a strong selection pressure for organisms that can find and utilize this resource quickly. Ultraviolet radiation is emitted at certain stages in carrion decay. If a light frequency matches that emitted from carrion decay, the fly investigates; otherwise, the fly conserves its energy for more likely sources of food. The blowfly's energy cost of food capture would be much larger without the ability to sense the presence of such an ephemeral food source.

The direct relation between light and the success of predator prey systems is apparent, but other indirect effects of light on these systems must not be overlooked. Most of life on Earth is fueled by solar energy. Temperature of the surrounding environment depends on energy in light, which in turn not only regulates the metabolism of organisms but determines other physical properties of the environment that affect the energy cost of food capture, such as fluid dynamics.

4.1.4 Fluid Dynamics

Fluid dynamics influence the cost of food capture and movement by controlling food availability. For example, caddis fly larvae and black flies feed on food transported in fluids. Species of caddis fly larvae live on stream bottoms and depend on food that has deposited in the stream bottom. Water flow may move food away.

On a larger scale, fluid dynamics generates hurricanes, tornadoes, and flooding that can greatly alter environments. Changes in foraging energetics for a species associated with a disturbance can be beneficial or nonbeneficial for growth and reproduction and the impacts on foraging may be long term or short term. Many foraging organisms have a high "energy returned on investment" (EROI) because they are exploiting the large energy flow of their environment.

Fluid dynamics also impacts the thermodynamics of biota. The movement of fluids increases the rate of thermal conductivity, causing phenomenon such as the wind-chill effect. The profitability to a homeotherm of foraging can be altered by changes in thermal conductivity because of changes in energy costs of maintaining body temperature. Similarly, by influencing thermal conductivity, fluid dynamics alters the ability of poikilotherms to perform foraging activities.

4.1.5 Nutrient Limitation

In plants, nutrient limitations increase the cost of capturing light energy by limiting the plant's ability to build the necessary structures and chemicals for photosynthesis. Nitrogen fixation is an energy cost for food capture that can compensate for nutrient limitation. Bacteria and blue-green algae are the only organisms capable of nitrogen fixation and are either free living or form symbiotic associations with plants or other organisms (e.g., termites, protozoa, and legumes). Nitrogen fixation is an energy-intensive method for fixing atmospheric nitrogen to compensate for nutrient limitation [Eq. (2)]. Nitrogen fixation is only profitable in nitrogen-depleted environments.

$$\begin{aligned} N_2\ (\text{atmospheric}) + 160\ \text{kcal} &\rightarrow 2N \\ 2N + 3H_2 &\rightarrow 2NH_3 + 12.8\ \text{kcal} \end{aligned} \qquad (2)$$

An interesting example of the energetic constraint that relative nutrient availability can have on energy cost of food capture is found in a trout pond ecosystem located below a pasture in western Manitoba, British Columbia. Several of the eutrophic Manitoba ponds studied by David Schindler (of the University of Alberta) had phytoplankton populations composed mainly of nitrogen-fixing, blue-green algae that were encouraged by the high concentration of phosphorus

in water from cow manure. Blue-green algae are a very poor source of food for upper trophic levels and in addition are poisonous, so trout died. Nitrogen was added to the trout ponds, which changed the algal assemblage to predominately green algae, because nitrogen fixing was no longer advantageous. Green algae are a high-quality food source for zooplankton and resulted in an increase in trout production.

4.1.6 Foraging Energetics in Relation to Risk of Mortality

The energy cost of food capture can be influenced by the risk of mortality by restricting the time and/or location that an organism spends foraging. For example, Barbara Peckarsky (of Cornell University) found that the presence of insectivorous fish (trout) gave a chemical signal for mayflies to forage mainly at night, rather than when trout could see them. The hiding of mayflies under stones during the day in response to the presence of trout increased the energy cost of food capture for both predator and prey.

Disease or injury can also increase the energy cost of food capture for plants. The tar spot fungi (*Rhytisma* sp.) infect the leaves of trees maple trees, creating large, black, tarlike spots on leaves. Sometimes the tar-spot infection is so profuse that the tree will drop all of its leaves and then grow new ones. Growing a new set of leaves costs the tree much energy and nutrients and the energy contained in leaf biomass is large enough that it is energetically profitable for a tree to reabsorb leaf energy and nutrients before shedding them. If the tar-spot disease affects the tree late enough in the growing season, the energy cost of developing new leaves is larger than the energy returned from photosynthesis in the remaining part of the growing season, and thus a tree will often not foliate until next spring.

5. FEEDING STRATEGIES AND THE COSTS OF FOOD CAPTURE

5.1 Individual Strategies

5.1.1 Net-Building Filter-Feeding Caddis Flies

In ecosystems that are made up of flowing water, organisms can minimize energy investment in capturing food by constructing nets. Filter feeding can yield low energy costs for food capture, but there are several physical and biological constraints. These constraints include hydraulic forces, variability in the physical conditions, abundance of drifting food, quality of drifting food, filter net design, general lack of net portability, and the ability and energy costs for an organism to maintain its position in a flow regime. Within these constraints are trade-offs when caddis flies exploit different habitats and types of food.

Often the most abundant and consistently available source of drifting food is in high flow-rate environments. The process of natural selection has produced several species of caddis flies that can utilize these environments. Adaptations include terminal hooks for securing the invertebrate to the substrate, silken safety cords to prevent drift when dislodged, and portable cases that act as ballasts in high water flow. Additional adaptations are found in net design because the fly must build a net that has a size, shape, and mesh size that functions most effectively at capturing the particle sizes available. As a result, filter feeding by caddis flies does not always yield high-energy returns. Net design is important in that hydraulics regulates the size of food that is transported in water column. As flow rate increases, the size of particles transported by water increases through increased kinetic energy. Larger food items are typically of higher quality (animal in origin) and are usually preferred by filterers, but the combination of large particles and high flow can destroy the net. Once energy has been invested in a net, the environmental conditions may leave a fly with a net that does not capture food effectively. This fly would then have to take an energetic loss and either abandon the habitat for better conditions, create a new net, or continue filtering with a net of marginal quality. Filter feeders have a large potential to minimize the cost of food capture, but the fly may not reach this potential in an environment that is temporally and spatially dynamic.

5.1.2 Filter-Feeding Mussels

Filter feeding using energy to pump water into and out of clams and mussels can also be a very profitable feeding strategy. To examine this, the basal metabolic rate of mussels can be estimated in the laboratory during a time when they are not feeding. Small plastic beads are then added to an aquarium container containing mussels; the mussels start to filter out the particles, increasing their oxygen consumption. Real food particles added to the chamber in place of plastic beads allow measurement of growth rate and food consumption. The quantity of food energy consumed per unit energy invested in feeding is then calculated by difference. The highest energy return in investment into feeding is estimated at about 500% (Table II).

5.1.3 Drift Feeding and Fishes

Some fish feed on suspended food in the water column. Alewives and other herrings have netlike or filter systems whereas other fish have an evolutionary modification in the buccal cavity that reduces energy cost in food capture. Energy costs for foraging by fish and by mussels and caddis flies are different in that fish have a much lower cost for powered locomotion compared to a mussel or caddis fly; this enables fish to adjust to environments with varying conditions.

Many fish have specialized buccal cavities that can literally suck up their food. This modification in the cavity increases capture success and reduces the amount of energy invested in movement of the fish to swallow prey. Sliding cartilaginous mouth structures underlie the mechanism that allows these fish to suck up their prey; the volume of the buccal cavity can be expanded when the mouth structures slide into position. When the cavity is expanded, the negative pressure created inside the mouth causes external water and food to flood inside the cavity.

A classic study by Kurt Fausch (of Colorado State University) on territoriality and foraging in salmonines shows how stream fishes use adaptations in the buccal cavity to minimize the cost of food capture. Fausch observed salmonines in a stream consume drifting food items. Stream hydraulics over space can be very complex. Rocks on the bottom generate shielding to water flow, creating pockets of slow-moving water that are in close proximity to water moving at a high rate of speed. A fish downstream of such a structure invests little energy to remain in position while having access to a nearby environment that has a good supply of drifting food. The mosaic of food availability and locations of such slow-water feeding habitats creates feeding locations of differing quality. Fausch found that fish defend these feeding zones, resulting in a hierarchy of fish occupying habitats of differing quality.

5.1.4 Territoriality Reduces Energy Cost of Food Capture

Territoriality within a species' population can maximize surplus energy yield from feeding according to what is called Allee's principal. At first, this principal appears to conflict with optimal foraging theory in that it would be seen that the increased cost for active metabolism in defending a territory would decrease surplus energy yielded from feeding. However, studies have shown that the energy cost in defending a territory is small in relation to the surplus energy gained by protecting the availability of resources. Additionally, Allee's principal says that the population would gain from an established "pecking" order, so that energy is not wasted in conflict. If populations grew beyond available space and resources, the mean individual fitness would decrease. A large population of trout may have such a high demand for energy that the available resources are used mainly for metabolism, leaving little or no surplus energy for growth and reproduction. An optimal number of fish relative to their food supply is one that creates a maximum amount of surplus energy.

5.1.5 Phytoplankton Migration and Nutrient Limitation

Some types of phytoplankton, such as dinoflagellates, decrease the energy cost of capturing light energy by investing energy into movement. A stationary phytoplankton can experience a decrease availability of local concentrations of nutrients needed for photosynthesis and growth. Phytoplankton that invest energy into movement are able to exploit water with higher nutrient concentrations.

5.1.6 Olfaction and Predation

Having the ability to detect prey through chemoreception or olfaction increases the probability of food capture by giving the predator the element of surprise over its prey or by preparing the predator for an encounter. Thus, this strategy reduces the energy costs for finding prey.

5.1.7 Cottrell's Analogy: Humans, Machines, Fossil Fuels, and Agriculture

Machines using fossil fuels have enabled humans to pay the energy costs of food capture. The use of machines to exploit resources is not exclusive to humans. Cottrell gives an analogy between the human invention of the steam engine and the evolution of the water pump in filter feeders. Both the steam engine and the filter-feeder's water pump have increased the power to perform food-collecting activities, the result being an increase in populations of humans and filter feeders. The fundamental difference between the humans and filter feeders is that humans gained the ability to subsidize the costs of food capture using an external source of energy, and the filter feeders costs are subsidized by individual active metabolism.

Subsidizing the costs of food capture using fuels can lead foragers to foraging strategies that deviate from what optimal foraging theory would predict, because the energy cost for food capture is no longer important. Modern ocean fishing techniques

illustrates how humans have increased their power for capturing food using fossil fuels. The quantity of ocean fish humans harvest in a year is increasing while the energy cost for capturing fish increases even as fish abundances decline. Humans continue to invest more energy into fishing, even though the energy profitability of catching fish is declining.

Cuban offshore fishing industry and agriculture are clear examples of the level of human dependence on paying the energy costs of food capture with fossil fuels. Cuban offshore fishing collapsed when cheap fossil fuel supplied by the former Soviet Union was no longer available (Fig. 4) to run large commercial fishing boats. In addition, the production of important crops such as sugar cane and tobacco decreased because the country could not support the energy cost of running farming equipment.

As human population has increased, so has food production. Some attribute this increase in food production to large technical leaps in food harvesting efficiency (Table II), but most of the increase is due to elevated rates of production made possible by fossil fuel subsides. In fact, data on the energy cost of food capture for various methods used in agriculture indicate that as farming practices evolved from shifting agriculture to traditional farming to modern agriculture, the rate of production increased but energy efficiency decreased (Table II). Earlier methods used for food capture, such as foraging and pastoralism, have much lower energy returned on investment compared to other methods and cannot be performed at high rates. Although shifting agriculture is one of the most profitable investments of energy into food capture for humans, it is insufficient to feed all of the human population worldwide because yield per hectare, when averaged over the entire area, is very low. Modern humans have subsidized the energy costs of food capture to the point where the EROI is far less important than yield per unit time or area.

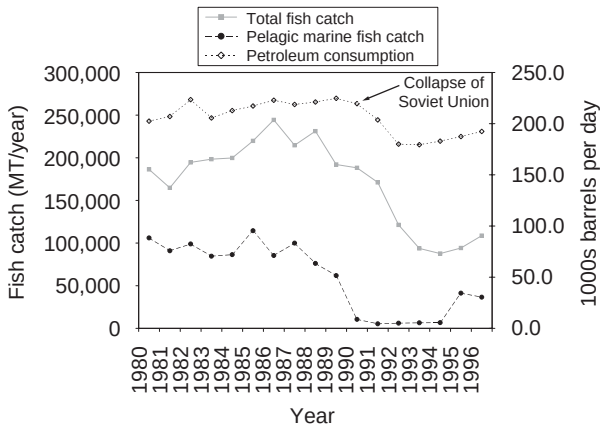


FIGURE 4 Fish catch in relation to quantity of oil consumption in Cuba. Cuban fisheries dependent on oil subsidies collapsed when the low-cost Soviet oil supply was no longer available. Compiled by J. Mead and P. Thompson (SUNY-ESF) using data from the Food and Agricultural Organization.

The foraging strategies described here are clearly a subset of the many others found in the world. For example, species have developed the use of sonar to find food (e.g., bats) to exploit food resources, and other species produce light to attract their prey (e.g., angler fish). The utility of studying energy costs in understanding feeding strategies is quite apparent.

5.2 Group Behaviors That Increase EROI in Food Capture

Foraging as a group of individuals can decrease the energy cost of food capture. Food may be located more effectively by groups than by individuals (e.g., dolphins, wolves), and food may be herded into locations or conditions that make capture much easier than otherwise, or the group can recover energies that are normally lost by an individual.

5.2.1 Dolphins and Whales: Herding Preys Using Bubbles

Herding of prey by a group of predators is not exclusive to humans. Whales and dolphins decrease the energy cost of food capture by herding their prey close to the water surface, where capture is easier. Small to medium-sized fish such as sardines school into tight groups of moving fish called "bait balls" when they detect a predator. Bait balls act as defense mechanism for many single-predator encounters because predators are typically much more effective at capturing individual fish. However, groups of cetaceans (whales, dolphins, etc.) take advantage of this defense mechanism by swimming under the bait ball, separating out individual fish from the outer perimeter of the ball and herding them up, toward the surface, where they are easy to capture.

5.2.2 Honeybees and the Waggle Dance

The time and energy needed to find food can be reduced if a population can focus its feeding on regions where food is abundant. Communication among members of a population allows an individual to instruct or lead other members of its population to a food source it has located. Honeybees have evolved communication skills to do just that. When a

foraging honeybee finds a good source of nectar, it collects a full load and then returns to the hive. On arrival at the hive, the bee performs an elaborate "dance" and emits a series of sounds using its wings, informing the other bees in the hive where the food is located. The movements and sounds the bee communicates with its wings are known as the waggle dance. Researchers have found that the communication of the foraging bee is very detailed, and includes direction, distance, and elevation of the food source.

6. SUMMARY

Thinking about foraging species in terms of energetics, environmental gradients, and foraging theory helps to comprehend and synthesize the performance of species in the world around us. Many factors influence the energy cost of food capture, but to understand the foraging strategy of an organism requires integration of these factors. The world has a fascinating array of different living systems, all of which depend on energy. There are low-energy, high-energy efficiency ectothermic organisms and high-energy, low-energy efficiency endothermic organisms. Organisms vary greatly in body size and physiological and behavioral adaptations; when considered collectively, these adaptations enable a strategy for capturing food that ultimately attempts to further the persistence of a species' traits. Based on the data in Table II, plants seem to be well adapted to capture a little more energy per unit energy invested in food capture compared to other nonhuman species. Also, humans have the ability to gain more energy from their energy investment in food capture compared to other species, but often sacrifice energy efficiency for rate of production, following Odum's maximum power concept.

SEE ALSO THE FOLLOWING ARTICLES

Conversion of Energy: People and Animals • Food System, Energy Use in • Heterotrophic Energy Flows • Human Energetics • Migration, Energy Costs of • Photosynthesis and Autotrophic Energy Flows • Reproduction, Energy Costs of • Thermoregulation

Further Reading

Alexander, R. M. (1999). "Energy for Animal Life." Oxford Univ. Press, Oxford.

Colinvaux, P. (1986). "Ecology." John Wiley & Sons, New York.

Hall, C. A. S. (1995). "Maximum Power: The Ideas and Applications of H. T. Odum." University Press of Colorado, Boulder.

Keragosian, M., Ringler, N. H. (2004). Foraging ecology of fishes: *In* "Fishes and Forestry—Worldwide Watershed Interactions and Management." Blackwell Publ., Oxford, in press.

Nagy, K. A. (1987). Field metabolic rate and food requirement scaling in mammals and birds. *Ecol. Monogr.* **57**(2), 11–128.

Peters, R. (1983). "The Ecological Implications of Body Size." Cambridge Univ. Press, Cambridge.

Pough, F. H. (1980). The advantages of ectothermy for tetrapods. *Am. Natural.* **115**, 92–112.

Schmidt-Nielsen, K. (1997). "Animal Physiology: Adaptation and Environment." Cambridge Univ. Press, Cambridge.

Smil, V. (1991). "General Energetics: Energy in the Biosphere and Civilization." John Wiley & Sons, New York.

Stephens, D. W., and Krebs, J. R. (1986). "Foraging Theory." Princeton Univ. Press, New Jersey.

Tucker, V. A. (1975). The energetic cost of moving about. *Am. Sci.* **63**(4), 413–419.

Cost–Benefit Analysis Applied to Energy

J. PETER CLINCH
University College Dublin
Dublin, Ireland

1. **Definition of Cost–Benefit Analysis Applied to Energy**
2. **History of Cost–Benefit Analysis**
3. **Rationale for Cost–Benefit Analysis Applied to Energy**
4. **Cost–Benefit Methodology**
5. **Valuation of Energy Externalities**
6. **Further Issues in the Application of Cost–Benefit Analysis to Energy**

Glossary

benefits transfer A process of taking environmental valuation estimates from existing studies and applying them outside of the site context in which they were originally undertaken.
discounting Adjusting a cost or benefit in the future so that it may be compared to costs and benefits in the present.
energy externality A cost or benefit that arises from an energy project but is not borne or appropriated by those involved in the energy project (e.g., pollution from energy generation).
environmental valuation The process of placing monetary values on environmental effects.
revealed preference methods Environmental valuation methods that use data from actual behavior to derive values for environmental assets.
sensitivity analysis A presentation of various possible outcomes of a future project or policy under conditions of uncertainty.
shadow pricing The process of adjusting a distorted market price to better reflect the true opportunity cost to society of using that resource.
stated preference methods Direct environmental valuation methods that ask people in a survey to place a value on an environmental asset.
total economic value A categorization of the various elements of value of a resource, project, or policy.
value of statistical life The translation of the value of a small reduction in the risk of death into a value for a project that would involve the equivalent of one life saved or lost.
welfare criterion The criterion in a cost–benefit analysis to assess whether a project or policy benefits society.
welfare measure A measurement tool to assess the strength of consumers' preferences for or against a project or policy in order to assess its costs and benefits.

Cost–benefit analysis applied to energy is the appraisal of all the costs and all the benefits of an energy project, policy, or activity, whether marketed or not, to whomsoever accruing, both present and future, insofar as possible in a common unit of account.

1. DEFINITION OF COST–BENEFIT ANALYSIS APPLIED TO ENERGY

Cost–benefit analysis (CBA), also known as benefit–cost analysis, is rooted in applied welfare economics. It is a way of organizing and analyzing data as an aid to thinking. It provides a set of procedures for comparing benefits and costs and is traditionally associated with government intervention and with the evaluation of government action and government projects. The underlying rationale for CBA is rational choice; that is, a rational agent will weigh the costs and benefits of any proposed activity and will only undertake the activity if the benefits exceed the costs.

A confusion that sometimes exists relates to the difference between social cost–benefit analysis and cost–benefit analysis. However, to an economist, social impacts and environmental impacts are economic impacts, so any analysis purporting to be a CBA must address social and environmental impacts.

 © 2004 Elsevier Inc. All rights reserved.

In this way, CBA differs from financial analysis in that it assesses the costs and benefits to the whole of society rather than to a portion of society. For example, a financial analysis of a proposed new power plant by a private energy supplier will not take into account the costs imposed on wider society (such as environmental impacts), whereas a CBA, by definition, will do so. In addition, there are various other subtle but important differences between financial analysis and CBA related to the fact that market prices do not always reflect the true value of goods and services to society. This provides the basis for a technique known as shadow pricing, which adjusts distorted market prices such that they may be used to value costs and benefits appropriately.

As the previous definition implies, a CBA must include all the costs and all the benefits to whomsoever accruing. Thus, the definition instructs that whether the costs or benefits are marketed or not makes no difference. In other words, even if some of the costs or benefits are not exchanged in markets at all and so do not have a price, they may still be of value. This provides the basis for the inclusion of environmental and wider social impacts in CBAs of energy projects and policies. The common unit of account in CBA is money. Therefore, insofar as is possible and reasonable, all impacts, including environmental and social, should be evaluated in monetary terms so that they can be compared with one another. This provides the basis for a technique known as nonmarket or environmental valuation—that is, placing monetary values on goods and services not exchanged in markets, such as environmental improvements and damages.

In many cases, the costs and benefits of a project or policy will arise at different times. This is the case in relation to most energy activities, where the consequences of actions may not be known until far into the future, particularly in the case of environmental impacts such as global warming. This provides the basis for a technique known as discounting, which allows future costs and benefits to be compared with those in the present.

CBA has been used *inter alia* to estimate the environmental costs of power generation, the net social benefit of different power supply options, the net benefit of the use of renewable energy sources (e.g., solar, wind, and hydroelectric power), the net benefits of alterative modes of transport, the benefits of energy efficiency programs in the commercial and domestic sectors, as well as for policy analysis of energy efficiency regulations such as energy rating and labeling.

2. HISTORY OF COST–BENEFIT ANALYSIS

Detailed histories of CBA can be found elsewhere, but the essence is as follows. In 1808, U.S. Secretary of the Treasury Albert Gallatin gave what is believed to be the first recommendation for the employment of CBA in public decision making when he suggested that the costs and benefits of water-related projects should be compared. In 1936, the U.S. Flood Control Act deemed flood control projects desirable if "the benefits to whomsoever they may accrue are in excess of the costs." In 1950, after a debate on the foundations of CBA in welfare economics, the U.S. Federal Inter-Agency River Basin Committee published a guide to CBA known as the "Green Book." From then on, there was considerable interest in CBA and the development of cost–benefit rules, with seminal contributions by a number of authors, particularly during the 1970 s. Although the distortions created by government intervention in the 1970 s provided much room for the application of CBA, the technique went through a lull in the early 1980s when structural reform replaced piecemeal intervention. However, CBA has made a comeback since then, mainly as a result of increased public awareness of environmental degradation and the associated demand that this cost be accounted for in analyses of projects and policies. However, for a number of years, the economist's view that environmental impacts could be assessed in monetary terms had little impact on policy, with many environmentalists having a particular problem with this concept. For example, in the U.S. Clean Air Act (1970), it was forbidden to use CBA, and the U.S. Clean Water Act (1972) set the objective of the elimination of all discharges into navigable waters. However, in the 1980 s, it became a requirement that the U.S. Environmental Protection Agency perform CBAs on all environmental directives to assess their economic efficiency and significant resources were devoted to researching cost–benefit techniques and methods of placing monetary values on the environment. Although it was not until the Single European Act was introduced in 1987 that a legal basis for European Union (EU) environmental policy was provided, the nature of environmental policy reflected the origins of the EU being an economic one. EU policy that "the potential benefits and costs of action or lack of action must be taken into account in the preparation of policy on the environment" provides an obvious platform for the use of CBA.

3. RATIONALE FOR COST–BENEFIT ANALYSIS APPLIED TO ENERGY

The need for CBA applied to energy results from the existence of market failure. If there were no market failure and all energy suppliers were privately owned, there would be no need for CBAs of energy projects and policies. If market signals were appropriate, those activities and projects that are in the best interests of society would be implemented automatically and financial analyses by firms would simply replace any need for CBAs. However, the market does not work perfectly. Market failure necessitates CBA of energy projects, policies, and activities as a result of the existence of externalities, the associated need for environmental policy instruments, state ownership of energy generation and supply, assessment of sustainability, and "green" national accounting.

3.1 Energy Externalities

One of the most significant areas of environmental concern today relates to the environmental impacts of energy generation from fossil fuels, particularly acidification precursors, greenhouse gases, and localized air quality impacts. Much of these emissions result from power generation and the transport sector. These environmental impacts occur at local, national, regional, and global levels.

The environment can be defined as those parts of our physical and psychological endowment that we somehow share, which are "open access." An economic analysis shows that it is precisely this shared nature of many environmental endowments that threatens their quality and character. In a market system, goods are allocated by the price mechanism. In a free market, the price of a good is determined by the demand for, and supply of, the good. Price is therefore a reflection of society's willingness to pay for, or the valuation it places on, the good in question. However, the shared nature of many environmental assets, such as air, the atmosphere, and water resources, means that they are not owned and so do not have a price. When goods are seen as free they are overused. Thus, the market fails to protect environmental assets adequately.

This failure of the market can be best demonstrated with a simple example: If a power plant produces SO_2 emissions resulting in acid rain, which causes forest damage and damage to buildings, in a free market these costs will not be reflected in the cost of power generation. Thus, the price of the energy will not reflect the true cost to society of generating it. In this case, more energy will be generated and more of the environmental asset depleted than is in the best interest of society. When a cost is imposed on those other than the person or business that produces the cost, this is known as an external cost. Of course, many energy projects will reduce such externalities, such as through renewable energy sources and improved heating and insulation systems.

When assessing the social efficiency of energy activities using CBA, it is necessary to include these externalities and other wider societal impacts. However, because by definition such costs and benefits are not traded in markets, there is no price to reflect their value. This provides the basis for the tools of environmental valuation.

3.2 Calibration of Environmental Policy Instruments

Unless some mechanism is found to internalize these externalities, they will cause a misuse of society's scarce resources. This provides the basis for government intervention in the market. The notion of externalities was addressed by economists as far back as the 1930s. Pollution was seen as the consequence of an absence of prices for scarce environmental resources and so the prescription for solving the problem was to introduce a set of surrogate prices using environmental (Pigouvian) taxes and charges. A large literature has developed on these so-called market-based instruments. One of the more controversial concepts in environmental economics is that there is an optimal level of pollution. Pollution is usually the by-product of a useful activity, such as the production of goods and services; in other words, it is an input into the production process. Reducing pollution usually involves a cost in terms of output forgone and/or the cost of the technical process of abatement. Therefore, there is a trade-off between the environmental improvement and the costs of pollution reduction. The optimal level of pollution occurs where the benefit of another unit of abatement (the marginal benefit) equals the (marginal) cost, with the (marginal) cost being higher after that point. In the case of an extremely hazardous pollutant, the optimal level of pollution will be zero or close to it, but in most cases the optimal level of pollution will be positive. CBA provides the tools to assess these trade-offs. In reality, few economic instruments are actually based on such valuations. Research on so-called target setting using

environmental valuation is still in a preliminary stage. However, it could be used to assess the optimal level of an energy tax.

3.3 State Ownership of Energy Generation and Supply/State Support of Energy Efficiency Initiatives

In some countries, energy supply companies are owned by the state. In some cases, the original rationale for such ownership was that no private company would bear the large fixed costs of developing an energy supply system. In addition, it is sometimes suggested that state companies can perform as natural monopolies and thereby exploit economies of scale while producing at minimum average cost and subsidizing energy for certain sections of society. CBA has a role to play in assessing the merits of such state initiatives and is particularly relevant when assessing the merits of government intervention to improve energy efficiency, for example, by subsidizing wind power or a domestic energy efficiency program.

3.4 Sustainability Assessment and Green National Accounting

CBA is useful in assessing sustainability; that is, what is the value of our resource stock, how is it being depleted, and is man-made capital that is replacing it of equal value? Green accounting adjusts traditional measures of economic performance such as gross national product to reflect environmental damage and the depletion of resources. In relation to energy, CBA may be used to assess the social costs and benefits of depleting energy resources such as fossil fuels.

4. COST–BENEFIT METHODOLOGY

4.1 Welfare Criterion

In using CBA to assess the costs and benefits of a project or policy to society, it is first necessary to define the social efficiency criterion. The most frequently used criterion in microeconomic theory is Pareto optimality. Using this criterion, a project would be deemed to pass a CBA if it made at least one person better off without making anybody else worse off. For an investment to pass the Pareto optimality test, compensation must be paid to those bearing the costs so as to leave them indifferent between their welfare *ex post* and *ex ante*. Most policies and investments leave somebody worse off and thus most would fail to meet the Pareto criterion. Frequently in CBA, the Kaldor–Hicks test is used. This test embodies the potential compensation principle. Kaldor categorized an action as efficient if those who gain from the action could compensate the losers and still be better off. Hicks framed the question slightly differently and labeled the action efficient if the potential losers from the action could compensate the potential beneficiaries of the action for not initiating the action. In both cases, the action is an improvement regardless of whether the compensation is actually paid. Thus, if the CBA shows the project to result in a welfare improvement under the Kaldor–Hicks test, this is essentially a potential Pareto improvement since there is the potential, in theory, to pay compensation to the losers such that nobody would be worse off. This will be the case as long as the benefits of the project outweigh the costs (i.e., the benefit:cost ratio is greater than 1).

Whether or not this potential payment is enough to show that a project is good for society has been the matter of much debate. Some authors have suggested that despite the shortcomings of the Kaldor–Hicks criterion, it is often useful to measure the effects of a change in the total value of output independently of the distribution of output; however, in judging the social efficiency of a project, equity aspects should also be considered. A two-part test may be used in which a project or policy is put to the Kaldor–Hicks test and then tested to determine if it improves (or at least does not disimprove) the distribution of income. Approaches that consider questions of equity therefore allow the possibility of acceptance of a project or policy that returns a negative figure for the sum of individual welfare changes if the gains from income distribution outweigh these losses. However, when considering equity the difficulty is in the assignment of weights to individual welfare changes. Consideration of equity issues and distributional consequences are particularly relevant in the case of projects that affect the price of energy and/or health. Energy price increases may be regressive unless poverty-proofing measures are put in place because energy is a necessity and therefore requires a larger proportion of the budget of less well-off people. On the other hand, improvements in domestic energy efficiency may bring significant health and comfort benefits to the elderly, who in some countries have been shown to be more likely to live in energy-inefficient homes and therefore more likely to be subject to winter morbidity and mortality.

4.2 Welfare Measures

In order to assess the costs and benefits of a project or policy, it is necessary to examine the strength of consumers' preferences for or against the project. The standard approach is to measure the benefit an individual obtains from a project or policy by the maximum willingness to pay on his or her part for such a benefit or the maximum willingness to accept compensation for not receiving such a benefit. Similarly, a cost can be measured by the maximum willingness to pay on the part of an individual to avoid such a cost or the maximum willingness to accept compensation for such a cost being imposed on the individual. Such welfare measures may be approximated using shadow pricing and/or direct valuation methods from surveys.

4.3 Temporal Considerations

Account must be taken of the fact that costs and benefits arise at different times. Thus, future costs and benefits must be discounted so that they can be compared with present costs and benefits. The rate of return for an investment can be strongly affected by the discount rate if there is a long time span involved. Therefore, the choice of discount rate is very important. In some estimates of damages from global warming resulting from the burning of fossil fuels, a small change in the assumed discount rate will have significant implications for climate change policy. However, despite the enormous literature on discounting, it is a difficult task to calculate the social rate of discounting. The discount rate is measured by some combination of a social rate of time preference and a measure of the opportunity cost of capital. Economists have been unable to provide a definitive figure for the social rate of discount, and so a rate tends to be chosen rather than measured.

In an application of CBA to any project or policy, a range of discount rates should be used. Often, for a government project, a published government test discount rate is used for making policy recommendations. These vary by jurisdiction. For example, UK treasury guidelines for the opportunity cost of public investment take the rate of return to marginal private sector investment as 6%. For pricing output sold commercially, public sector agencies are expected to achieve a required average rate of return of 8% to be consistent with the private sector. In the United States, recommended rates for investment projects vary. The Office of Management and Budget recommends a rate of 10%, whereas the General Accounting Office prefers a rate of 2.5% based on government bond yields. The World Bank has traditionally used an 8% rate in project analysis.

4.4 Uncertainty and Risk

In *ex ante* CBAs, investments with long time horizons are subject to uncertainty, particularly in relation to future prices. In addition, when evaluating costs and benefits *ex ante*, there may be considerable uncertainty regarding the magnitude of various parameters. This is particularly the case in relation to potential environmental impacts and their valuation. In cases in which there is a high degree of uncertainty in relation to the coefficients being used to give the monetary value of outputs of a project, lower bound estimates are sometimes used because, if the results are to be the basis for the recommendation of a change in policy, it is thought by some authors that it is better for the benefits to be underestimated rather than overestimated. It is important to note that lower bound results do not necessarily imply the lowest possible results. Rather, they denote the lower limits of a range of reasonable estimates.

Sensitivity analysis, whereby results are calculated for various scenarios, should be used in all CBAs. In addition, confidence intervals should be presented where possible. A general rule is that the more results the better while giving guidance and reasons for the particular estimates chosen by the author(s) for the purposes of recommendations based on the CBA results. Risk assessment may be incorporated by including probability estimates for various scenarios.

4.5 Shadow Prices and Wages and Community Stability

Shadow prices must be used when there are distortions within an economy that render certain prices invalid as measures of the social opportunity costs and benefits of using resources in the project in question. In the real world, market prices may not reflect the social value of goods or services because of distortions in the market, such as monopoly power (which inflates price), indirect taxes and subsidies (unless they are correcting some other distortion), and unemployment. Thus, when calculating each value of a project, it is necessary to assess the appropriateness of using the market value and, where necessary, these values should be adjusted for market distortions. This can be a problem in energy projects for which a particular piece of technology may be supplied by one

company and/or estimates must be made regarding how price would decline should such a piece of technology be widely adopted, for example, in a national domestic energy efficiency retrofit.

The approach to valuing labor is worth detailing. In reports purporting to be CBAs, labor or "employment" is sometimes treated as a benefit rather than a cost. Labor is an input into a project and, therefore, a cost even though it may have some additional benefits such as community stability, as discussed later. However, that cost may be less than the market wage would suggest. If labor markets clear, the shadow price of labor will equal the market wage. Any increase in employment in one sector of the economy will merely reduce the availability of labor to another sector. This is likely to be the case when the jobs are highly skilled and/or there is a shortage of such skills. However, for a country with a high unemployment rate, it could be argued that an increase in the demand for labor in one sector of the economy will not necessarily displace a job in another sector; that is, there may be employment additionality whereby if the new job is filled by a person who was previously unemployed, a very low cost (not zero, unless it is considered that the unemployed's spare time is totally unproductive) in terms of output forgone is imposed on society. In this case, the shadow price of labor would be close to zero. It has been the practice of some government CBAs to assume a shadow price of labor of zero. However, this is against the international practice of setting the shadow price of labor at most a fraction below the market wage. For example, in Canada, the shadow wage is usually 95% of the market wage, and in the United Kingdom they are usually considered equal. In many energy projects, specialized skills are required and so a shadow wage close to the market wage will usually be appropriate. However, in some household energy efficiency improvement schemes, long-term unemployed people have been trained how to fit various sorts of insulation, for example. In this example, a case could be made for using a lower shadow wage rate.

A related value that should be considered is whether an energy project may contribute to community stability. For example, would the provision of a new power plant in a rural area provide jobs such that depopulation of that area would halt? This value is not the value of the job per se. Rather, it is the added value of a job being in a particular place. These values are difficult to assess but are real. For example, it is often observed that employees will be willing to accept lower salaries to live in their hometown.

4.6 Valuing Changes in Energy Use

When calculating the benefits of reductions in energy use, the price of energy may be used to value such savings. However, it may be necessary to shadow price such savings. If the state supplies energy, an assessment must be made of whether there is any form of subsidy involved; if so, this should be removed. Where there is monopoly power in the energy market, either locally or, for example, through the monopolistic behavior of the Organization of Petroleum Exporting Countries, the price of energy is unlikely to reflect the true cost to society of its generation.

Energy supply is considered so essential that a premium may be placed on its supply from domestic sources if this reduces the risk of supply interruption. Most countries have assessed the importance of such supply, and this information may be used to assess any premium that should be applied. If the increase in supply that would be generated by the energy project under consideration is marginal to the total supply in a country, it is unlikely that such a premium would be significant.

In an *ex ante* analysis, a sensitivity analysis should be undertaken for alternative scenarios regarding the possible future changes in energy prices. These are difficult to predict because there are a number of offsetting effects. Possible deflationary pressures include the development of more energy-efficient machinery and technology, improved availability (and therefore reduced prices) of non-fossil fuel energy sources and systems, and any change in the regulation of the energy sector. Possible inflationary pressures include rapid economic growth and a consequent increased demand for energy, the exercising of market power in oil production, future scarcity of fossil fuels, and the introduction of carbon taxes and/or tradable permits in light of the Kyoto Protocol. However, if the price of energy is inflated by a carbon tax/emissions-trading system set to reflect the external costs of energy use, it is important to avoid summing external costs estimated by the techniques described later with the energy savings using this price because if the price already reflects the external effects, it would be inappropriate to include them a second time.

4.7 Externalities

Shadow prices adjust for one form of market failure. However, a CBA must also take into account the potential existence of externalities and public goods, as explained previously. The consideration of such

costs and benefits is extremely important in CBAs because all of the costs and all of the benefits to the whole of society are being considered. The difficulty is that these external costs and benefits are not exchanged within markets, so there is no price to reflect their value. For this reason, it is necessary to use nonmarket or environmental valuation approaches to place monetary values on these externalities.

5. VALUATION OF ENERGY EXTERNALITIES

Most energy projects have some impact on the environment that must be factored into a CBA. For example, a new coal-fired power plant will have implications for climate change, localized air pollution, and the production of acidification precursors. A cost–benefit analyst assessing the merits of a renewable energy project or an energy efficiency program will wish to factor in the value of reductions in such environmental effects.

5.1 Economic Value Concepts

A useful framework for assessing the effects of environmental impacts on humans is known as the total economic value approach. This expands the traditional concepts of economic value to include option value and existence value. In essence, an environmental asset is considered to have both use and nonuse value.

Use value can be broken down into actual use value and option value. Suppose we are estimating the value of damage caused to woodland from acid rain resulting from power generation. Actual use value is derived from the actual use of the asset and is composed of direct use value, such as timber from the woodland, and indirect use value, such as the value that would be yielded to inhabitants downstream should a forest reduce flooding.

Option value can be broken down into a number of components: the value individuals place on the preservation of an asset so that they may have the option of using it in the future (value of future options for the individual) for recreational purposes, for example; the value they attach to preserving the asset because it can be used by others (vicarious value); and the value they attach to preserving the asset so that future generations have the option of using it (bequest value). Thus, total use value is defined as

$$\begin{aligned}\text{Total use value} = {} & \text{direct use value} \\ & + \text{indirect use value} \\ & + \text{option value},\end{aligned}$$

where

$$\begin{aligned}\text{Option value} = {} & \text{value of future options for the} \\ & \quad \text{individual} \\ & + \text{vicarious value} \\ & + \text{bequest value}.\end{aligned}$$

Nonuse values are existence or passive-use values. Existence values reflect the benefit to individuals of the existence of an environmental asset, although the individuals do not actually use the asset. These values may exist due to sympathy for animals and plants and/or some human approximation of the intrinsic value of nature (i.e., this value reflects a sense of stewardship on the part of humans toward the environment). Existence value is classed as a nonuse value such that total economic value (TEV) is defined as

$$\begin{aligned}\text{TEV} = {} & \text{actual use value} + \text{option value} \\ & + \text{existence value}.\end{aligned}$$

It is important to note that the TEV approach attempts to encapsulate all aspects of a resource that enter into human's utility functions (i.e., it is anthropocentric). Thus, it does not attempt to find an intrinsic value for the environment, where intrinsic means the value that resides in something that is unrelated to humans. However, it does attempt to include the intrinsic value that humans bestow on nature. There is no agreement on the exact breakdown of the values in the TEV framework; however, the important point in CBA is to account for all these values as best as possible.

5.2 Environmental Valuation Methods

There are two broad categories of nonmarket valuation methods: revealed preference methods and stated preference methods.

5.2.1 Revealed Preference Methods

Revealed preference studies use data from actual behavior to derive values for environmental assets. Suppose the value of a marketed good depends in part on an environmental good. These methods try to isolate the value of the environmental good from the

overall value of the marketed good. Such techniques include the following:

Hedonic pricing method: The value of a house depends, in part, on its surrounding environment. This method examines the determinants of the house's value and then isolates the effect that the environmental feature (e.g., a nearby park) has on the house. This is used as an estimate of the value of the environmental asset in question.

Travel cost method: The value of the park can be estimated by the willingness to pay to attend the park in terms of costs of travel and entrance to the park.

Production function approaches: These value a decrease (increase) in environmental quality as a result of some activity by the loss of (addition to) the value of production due to that activity. For example, the cost of acid rain resulting from pollution from fossil fuel-based power plants might be measured by the lost output of timber from forests (dose–response functions), the cost of restoring the damage to buildings (replacement cost), or the cost of defensive expenditures.

The production function approaches may seem more applicable to estimating externalities of energy generation, but the hedonic pricing method and travel cost method may be used to estimate the value of an amenity such as a forest park. This value could then be used to calculate the cost of damage to the park from air pollution.

5.2.2 Stated Preference Methods

Direct valuation methods ask people in a survey to place a value on the environmental asset in question by rating (contingent valuation), ranking (contingent ranking), or choosing trade-offs between various policy alternatives (choice procedures). These direct methods via questionnaires are the only approaches that can estimate existence value. The best known and most controversial of these methods is contingent valuation.

The contingent valuation method (CVM) collects information on preferences by asking households how much they are willing to pay for some change in the provision of an environmental good or the minimum compensation they would require if the change were not carried out. Since the mid- to late 1980 s, this method has become the most widely used approach for valuing public goods and there has been an explosion in the number of CV studies.

The CVM approach is to choose a representative sample, choose an interview technique (face-to-face, telephone, or mail), and then construct a questionnaire containing:

- A detailed description of the good being valued and the hypothetical circumstance in which it is made available to the respondent.
- Questions that elicit the respondents' willingness to pay for the good being valued: open ended (How much would you pay?), bidding (Will you pay x, if not, will you pay $x-y$), payment card (Which of these amounts would you choose?), and binary choice (If it costs x, would you pay it? Yes/no).
- Questions about the respondents' characteristics (e.g., age, education, and income), their preferences relevant to the good being valued (e.g., how concerned or interested are they in the environment), and their use of the good.

The mean or median willingness to pay from the sample of households surveyed can be multiplied up to the population level to give a total valuation.

5.3 Valuing Environmental Damage and Improvements from Energy Projects

The most significant environmental effects of energy projects that are included in CBAs tend to be the contribution of such projects to global warming from changes in CO_2 emissions and acid rain resulting from the production of NO_x and SO_2. In addition, many energy projects have implications for the production of particulate matter that may have consequences for human health and visibility.

The impacts of climate change are as follows: storm and flood damage (increased in some regions, such as the Caribbean and Pacific, but not others), increased mortality and morbidity, effects on small islands (sea levels rise due to heating water and melting ice caps), increasing salinity of estuaries, damage to freshwater aquifers, and effects on agriculture and forestry. It is estimated that a substantial portion of the existing forested area of the world will undergo major changes.

There are three approaches to placing a monetary value on the benefits of greenhouse gas reductions. The damage-avoided approach values 1 ton of carbon not emitted by the cost of the damage that would have been done by global warming in the event that it had been emitted. The offset approach measures the value of not emitting 1 ton of carbon using one method compared to the next cheapest alternative method. The avoided cost of compliance

approach measures the 1 ton of saved carbon by the avoided cost of compliance with a global/regional CO_2 emissions reduction agreement. A variety of estimates have been produced regarding the benefits of reducing CO_2, including studies by the United Nations Intergovernmental Panel for Climate Change.

Several major research projects have endeavored to place monetary values on all the environmental effects related to energy. Two EU-funded projects, one on the external costs of power generation (ExternE) and the other on European environmental priorities, are very useful sources of information on such values.

The ExternE project used an "impact pathway" methodology. Under this methodology, the fuel cycle, or life cycle of the fossil fuel used for generation, is first divided into stages, such as extraction, transport, combustion, and waste disposal. The environmental burdens imposed at each stage and the likely impact of these burdens on the environment and the public are isolated. The impacts are then valued using a variety of economic valuation methods, such as contingent valuation, direct cost of damage, or cost of mitigation. These costs are then normalized to give a cost per unit energy output and a value per ton of a pollutant reduced.

The "priorities" project endeavors to derive values for various environmental externalities suitable for use in a European context. The impacts of SO_2 on acute mortality, morbidity, materials, and the fertilization effect are all included; however, some impacts are omitted, such as forest damage, adverse effects on ecosystems and cultural heritage, and the dampening effect that SO_2 has on global warming. The main areas of uncertainty relate to the value of mortality. NO_x emissions are dealt with under two headings, acidification and tropospheric ozone. The fertilization effect of NO_x is examined under acidification. For damage to health and crops, the effects of NO_x on tropospheric ozone are used. Again, the bias is toward underestimation because damage caused to forests, biodiversity, non-crop vegetation, and materials is excluded.

Care should be taken when using unit damage values such as those in the studies mentioned previously. First, the confidence intervals are large (and should always be quoted in a CBA). In addition, there is confusion about which emissions cause which adverse effects. For this reason, it is very easy to double count when using estimates derived from different studies. For example, in some studies negative health effects are included in SO_2 damage estimates, whereas in others the most significant health effects result from particulate matter emissions. Finally, the damage caused by 1 ton of most air pollutants (excluding CO_2) will vary depending on the geographical origin of the emission. However, most estimates are derived from an exercise known as benefits transfer.

5.4 Benefits Transfer

To reduce costs and where there is no alternative procedure, it is standard practice in CBAs to apply existing monetary valuation studies outside of the site context in which they were originally undertaken. This is known as benefits transfer. There are a number of rules when using this method, the most obvious being that the contexts of the studies should be very similar. In addition, appropriate statistical techniques should be used, including meta-analyses of best practice studies; the studies should have similar population characteristics; and property rights should be similarly distributed. The method is most accurate when the impacts are not affected by local characteristics (e.g., income and population size). The most comprehensive studies of energy externalities referred to previously involve significant benefits transfer.

5.5 Valuing Mortality and Morbidity

Placing monetary values on life and death is perhaps the most controversial area of economics. There are two major approaches for carrying out this task. The first values life directly through livelihood. This is known as the human capital method and is the oldest and most easily applied. It equates the value of life of an individual as the present value of future lost output of the individual, which is usually estimated using income. However, it is important to note the substantial drawbacks of this approach. Most obvious, it values livelihood rather than life per se. It is quite improbable that there would be a 1:1 relationship between individuals' earnings and their demand for risk reduction. In other words, the human capital approach makes no reference to how humans value risks to their own lives. One of the most glaring flaws of this approach, however, is the fact that it places a zero value on the retired population and those who do not "produce" marketed output (e.g., those who work in the home).

The appropriate way of valuing reductions in mortality is not to value a life per se but rather to value people's willingness to pay for a reduction in

the risk of mortality. These data derive primarily from one of three study formats: the wage risk model (also known as the hedonic wage model), the avertive behavior method, and the contingent valuation method. Although space does not permit a detailed explanation of these methods, the general approach is to value a small reduction in risk and then translate this into the equivalent of a life saved. This is the value of a statistical life (VSL).

There has been controversy regarding this approach in relation to the estimation of the costs of power generation. The VSL estimates would be expected to be considerably lower for a developing country because the willingness of people to pay to reduce risk of death is partially dependent on income. Thus, the VSL used for a developing country (e.g., for deaths from coal mining) would be expected to be lower than the VSL used for a developed area (e.g., for deaths from air emissions in Europe by the burning of such coal). It is notable that the ExternE project steered clear of controversy by using the same VSL for each area.

The value of morbidity may be assessed using a cost of illness approach. In the case of air pollution from energy generation, this would be the sum of the cost, both public and private, of hospital stays and drugs for the treatment of associated illnesses. More positively, the health benefits of improving domestic energy efficiency would be the associated avoided costs of illnesses. Other costs avoided by wider society include the loss of productivity from those who are ill and unable to work. The avoided cost to the individual beyond drugs and hospitalization direct costs (if any) is more difficult to calculate. One approach is to estimate individuals' willingness to pay to avoid "restricted-activity days." The priorities study referred to previously contains such estimates.

6. FURTHER ISSUES IN THE APPLICATION OF COST–BENEFIT ANALYSIS TO ENERGY

When undertaking or evaluating a CBA of an energy project, there are a number of additional issues that must be considered. The tool is not without its controversies. These arise because of its use of money as the common unit of measurement. First, assessing welfare changes using willingness to pay and/or willingness to accept is perfectly valid in theory if one believes that income is distributed optimally. However, care must be taken when the population(s) whose welfare is being assessed has a high income variance. Poor people will be willing to pay less, all else being equal, simply because of their budget constraint. In addition, because the values and market prices used in a CBA of an energy project or policy are in part a product of the existing income distribution, if that distribution changes, prices will change, and so will market and nonmarket values. Second, distributional effects tend to be ignored in CBAs and yet energy policies often have very different consequences for different socioeconomic groups. Because energy is a necessity, energy efficiency schemes may benefit those on lower incomes more than average. However, without poverty-proofing measures, energy taxes will comprise a much higher proportion of their income than average. Third, not all support attempts to place monetary values on environmental assets, and the value of the statistical life concept is particularly controversial. Some find it unpalatable to consider that the environment might be traded with other goods and services let alone "lives." In addition, as in all public policy or human actions, the environmental values used are by definition anthropocentric.

Finally, the environmental valuation methods are limited in their ability to assess values. Many environmental outcomes assessed by natural scientists are subject to large confidence intervals even before economists endeavor to value them. Also, the valuation methods vary in their ability to assess environmental values. Economists have identified a number of criteria by which environmental valuation methods should be evaluated: validity (the degree to which it measures the total economic value of a change in the environment), reliability (the extent to which the variability in nonmarket values is due to random sources), applicability (assesses where the techniques are best applied), and cost (the practicality and viability of carrying out the valuation study). In general, revealed preference methods (e.g., productivity loss and replacement cost) are widely accepted valuation methods that provide valid and often reliable estimates of use values. However, they are limited in that they only provide estimates of use values, and it is often difficult to find suitable and reliable links between market goods and environmental amenities. The estimated values are often sensitive with respect to modeling assumptions, reducing their reliability in such cases. In general, however, revealed preference methods are applicable for valuation of policies that primarily affect use values and can provide very reliable results. Stated

preference methods such as contingent valuation are capable of measuring both use and nonuse (existence/passive-use) values. They are applicable to a larger array of environmental issues than are revealed preference methods, and they are often the only applicable methods for benefit estimation. However, they do have a number of limitations and may be subject to a number of biases.

There are some simple rules for using cost–benefit techniques. Simpler environmental valuation methods should be used where possible (e.g., production function approaches) while recognizing these are primarily of interest when the policy changes will mainly influence use values. A high degree of reliability in stated preference methods may be achieved by adhering to sound principles for survey and questionnaire design and if the method is applied to suitable environmental goods. The methods work best for familiar environmental goods and for salient environmental issues. The methods are likely to be least reliable where the environmental good being valued is complex (e.g., in the case of biodiversity).

Finally, in CBAs of energy initiatives, it is important to note that the estimates of values are made at the margin. The plausibility of the techniques described depends in part on the fact that most things remain the same, with the implicit assumption that there is no paradigm shift in values, power, income, and the like during the time period of assessment.

SEE ALSO THE FOLLOWING ARTICLES

Decomposition Analysis Applied to Energy • *Ecological Risk Assessment Applied to Energy Development* • *Energy Development on Public Land in the United States* • *Environmental Change and Energy* • *Equity and Distribution in Energy Policy* • *External Costs of Energy* • *Green Accounting and Energy* • *Value Theory and Energy*

Further Reading

Clinch, J. P., and Healy, J. D. (2001). Cost–benefit analysis of domestic energy efficiency. *Energy Policy* **29**(2), 113–124.

Commission of the European Communities—DG Environment (2003). European environmental priorities Web site: An integrated economic and environmental assessment. Available at europa.eu.int/comm/environment. ExternE Web site: http://externe.jrc.es.

Hanley, N., and Spash, C. (1993). "Cost–Benefit Analysis and the Environment." Elgar, Aldershot, UK.

Layard, R., and Glaister, S. (eds.). "Cost–Benefit Analysis," 2nd ed. Cambridge Univ. Press, Cambridge, UK.

Zerbe, R., and Dively, D. (1994). "Benefit Cost Analysis in Theory and Practice." HarperCollins, New York.

Food System, Energy Use in

CHRIS E. DUTILH
Unilever
Rotterdam, The Netherlands
ANITA R. LINNEMANN
Wageningen University
Wageningen, The Netherlands

1. **Definition and Boundary of the Food System**
2. **On-Farm Production**
3. **Transport**
4. **Processing, Packaging, and Storage**
5. **Consumption**
6. **Conclusions**

Glossary

conversion efficiency The efficiency with which energy from one component (feed) is converted into another component (e.g., meat, milk) expressed as a $kg\,kg^{-1}$.

direct energy The energy requirement for a specific process, expressed as megajoules per kilogram ($MJ\,kg^{-1}$ or $10^6\,J\,kg^{-1}$).

embodied energy Sum of the energy requirements in all preceeding production steps required to generate a food product or intermediate.

energy requirement The energy input required in a certain step in the production of food.

indirect energy All energy requirements not considered as "direct"; for example, auto fuel is usually considered direct energy, whereas energy to build roads is considered indirect.

nutritional value A measure of the amount of energy that becomes available to the body after ingestion of a food product, expressed as kilojoules per kilogram ($kJ\,kg^{-1}$ or $10^3\,J\,kg^{-1}$); the nutritional value of a product is not related to the energy use in food production.

Energy use of food production systems starts with photosynthesis. Incoming sunlight is absorbed by chlorophyll in plant leaves. Water and carbon dioxide, which are required for the reaction, are taken up by the cell wall, and oxygen is emitted to the environment. The $(CH_2O)_n$ ("sugar") produced is the building block that plants use to make carbohydrates, proteins, and fats. This conversion process requires chemical energy.

1. DEFINITION AND BOUNDARY OF THE FOOD SYSTEM

Energy use of food production systems starts with photosynthesis, that is, at the interception of sunlight by plants and the subsequent transformation of $CO_2 + H_2O$ into $(CH_2O)_n$ and O_2. The incoming sunlight is absorbed by chlorophyll in plant leaves. Water and carbon dioxide, which are required for the reaction, are taken up by the cell wall, and oxygen is emitted to the environment. The $(CH_2O)_n$ ("sugar") produced is the building block that plants use to make carbohydrates, proteins, and fats. This conversion process requires chemical energy, which is obtained by the respiration of some of the products formed during photosynthesis: $CH_2O + O_2 \rightarrow CO_2 + H_2O +$ energy. For example, from 1 g glucose ($C_6H_{12}O_6$), a plant can produce 0.9 g other carbohydrates.

The transformation of sugars into other plant components is accompanied by more weight loss. For the production of amino acids, this loss depends on the source that supplies the nitrogen. For example, 1 g of glucose can be transformed into 0.7 g of amino acids when nitrogen is available as ammonia (NH_3). However, the usual source of nitrogen is NO_3^-, and in that case relatively much energy is needed for the reduction of nitrate into ammonia, according to the following pathway: $NO_3^- + 8\,H \rightarrow NH_3 + 2H_2O + OH^-$. This pathway results in the production of 0.5 g of amino acids from 1 g of glucose. The production of fat by plants is most energy demanding. For example, 1 g of glucose is required to

 © 2004 Elsevier Inc. All rights reserved.

generate 0.4 g of fat. The low conversion rate from sugars into fats on a weight basis is one of the reasons why the dry matter yield of oil-producing crops such as soybeans are significantly lower than the yields of starch-producing crops such as potatoes, rice, and grain, even though the lengths of their growing periods are not much different (Table I).

On average, plants use 1 g of glucose to produce 0.7 g of plant material that is made up of 25% protein, 55% cellulose, 6% fat, 6% sugars, and 8% minerals. Carbohydrates, proteins, and fats from plants can be used for human nutrition, but they may also be used as a feed for animals to produce eggs, dairy products, and meat.

Besides sunlight, various other sources and forms of energy are required during the subsequent stages of the food production system. That energy is directly needed to convert raw materials into new products or ingredients; therefore, it is called a direct energy requirement. Apart from energy input, there also is a loss of raw materials during each of the stages. Because raw material represents an amount of embedded energy, the materiel loss can also be described as an indirect energy requirement for the production stage. The total energy requirement for each stage is the sum of direct and indirect energy. For example, during the consumption stage at home, the direct energy is used to buy, store, and prepare the food. The indirect energy requirement at home is represented by the substantial amount of food (up to 15% of all food purchased) that is not consumed and ends up in the waste bin.

For a complete assessment of the embodied energy of food products, the entire life cycles of those products must be considered. These life cycles start at the farm, where the on-farm production takes place. Crops are sown, grown and harvested, whereas animals are bred, fed, and slaughtered. Food products end their life cycles at the moment of consumption. Some food products are consumed after a short and simple life cycle, for example, an apple that comes from a consumer's own backyard garden. On the other hand, shops offer a large variety of food products that are the result of sometimes long and complex life cycles. Such food products may contain exotic ingredients or ingredients that have undergone various processing steps, for example, a pizza with a topping of tropical fruits and processed meat.

Irrespective of the complexity of the life cycles of food products, several general handling steps can be discerned. The first step, already mentioned, is on-farm production. Following the harvest, agricultural products are transported either directly to a shop or to a factory for further processing (e.g., washing, cutting, cooking, mixing, packaging). Processed products are stored (either chilled or at ambient temperatures) and subsequently are transported to a retailer, where they are offered for sale. After purchasing the food, a consumer will carry the products home from the shop. At home, the products may be stored for a while, after which time they will be prepared for consumption. Alternatively the consumer can go to a restaurant to eat the food there. In that case, the restaurant owner/employees will have bought, stored, and prepared the food. Each of these steps, for which various alternatives exist, requires energy. In the end, the sum of all energy requirements in each step determines the total embodied energy of a certain food product. In what follows, this article looks in some detail at each step in the production chain and indicates potential differences in energy requirements between eating out and eating at home.

TABLE I

Yield per Hectare for Various Industrial Food Crops

	Yield (tonnes/ha)	Dry matter content (percentage)
Tomatoes	40–75[a]	5
Potatoes	30–70[a]	25
Vegetables	1–90[b]	5–20[b]
Grain	3–10[b]	95
Oil-producing crops(e.g., soybeans)	1–4	98

[a] Depending on agricultural conditions.
[b] Depending on the type of crop.

2. ON-FARM PRODUCTION

The embodied energy of food products (as distinct from nutritional or caloric value) differs among various food categories, and substantial differences also exist within a specific food category. Table II presents average embodied energy for some fresh ingredients from plant and animal origin under Dutch farm production circumstances. Crops grown in heated glasshouses, as is common in several northern countries, require substantial higher amounts of energy. That increase may be up to 40 MJ kg^{-1}. In some regions, glasshouses (or plastic equivalents) are used without heating just to make sure that heat from the sun is being collected

during the day and not lost too easily during the night. That practice saves energy and enhances the growth rate.

To raise the yield per hectare, fertilizers can be used. The production of fertilizers requires approximately 10 to 25 MJ kg^{-1}. The application of fertilizers may add up to 5 MJ kg^{-1} to the total energy requirement of the on-farm stage. However, in most cases, it is substantially lower than that. For crop protection, various chemical compounds are used (e.g., herbicides, pesticides, fungicides). The production of these chemicals requires between 100 and 1000 MJ kg^{-1}, and the applications vary substantially. The energy requirement for plant protection may rise up to the same level as for fertilizers. However, that figure can vary from year to year and from crop to crop because the application of crop protection aids depends very much on weather and local circumstances.

The embodied energy for products from animal origin is approximately a factor 10 higher than that for products from plant origin (Table II). Within the category of products from animal origin, fish–meat has a relatively low embodied content. The low energy requirement can be explained by the fact that fish are cold-blooded and, therefore, require less energy to produce body warmth (thermoregulation). There are two main sources of fish: sea fish and fish from aquaculture. The overall energy requirements for the production of these two do not differ much, but the buildups are rather different. For sea fish, a lot of energy is required for the ships at sea—between 2 and 50 MJ kg^{-1} depending on the type of fish. There are many variables that make an accurate calculation of the energy requirement for fishing very complex (e.g., the amount of by-catch, the efficiency of fish filleting, the use of by-products). In aquaculture, the fish are fed with fish meal, which originates from dry sea fish residues.

Feed-to-meat conversion ratios partly explain the differences in energy requirements between products from plant and animal origin. Table III shows the conversion efficiencies for chicken, pork, and beef based on long-term records of feeding efficiency ratios by the U.S. Department of Agriculture (USDA). Cumulative feed/gain ratios for chicken are between 1.8 and 2.0 for chickens in the most common 2.0- to 2.5-kg range. That ratio is raised by at least 10% to account for the feed requirements of breeder hens and cockerels that do not directly contribute to the meat production, as well as for the fact that some animals die prematurely. Ratios between 2.0 and 2.2 represent the standard of recent good performance. Typical rates from long-term statistics are somewhat lower (Table III).

TABLE II

Embodied Energy for Some Fresh Foods from Plant and Animal Origin Produced in The Netherlands

Product	Embodied energy (MJ kg^{-1})
Plant origin	
Potatoes	1
Vegetables	1–4
Fruit	2–5
Pulses	6–10
Animal origin	
Milk	10
Eggs	20
Fish (caught at sea)	20–40
Butter and cheese	50–60
Meat (chicken, pork, beef)	30–70

TABLE III

Typical Efficiencies for Meat Production

	Chicken	Pork	Beef
Feed (kg kg^{-1} live weight)	2.5	3.5	3–5[a]
Edible weight (percentage of live weight)	55	55	40
Feed (kg kg^{-1} edible weight)	4.5	6.4	7.5–12.5
Nutritional value (kcal kg^{-1})	1800	1700–3300	1200–3300[b]
Average energy conversion efficiency[c]	11.5	11.2	6.0–9.5

[a] Depending on the inclusion of milk production.

[b] Depending on fat content of the meat (average nutritional value 2500 kcal kg^{-1}).

[c] Percentage of the embodied energy in feed that is retained in the embodied energy of meat, assuming the average embodied energy for cattle feed of 3500 kcal kg^{-1}.

Pork is the least feed-intensive red meat due to pigs' inherently low basal metabolism. The overall feed/gain ratio for North American pigs from weaning to slaughter ranges between 2.5 and 3.5 with *ad libitum* intake of feed. In The Netherlands, on average 320 kg of feed is required for 115 kg of live weight pork. The addition of feeding costs of the breeding stock and adjustments (e.g., for premature mortality) can raise the overall feed/gain ratio significantly. From the early 1900s onward, the nationwide feed/gain ratio for pigs has declined only slightly, in spite of the substantially improved regimes, due to the selection of less lardy pigs. The efficiency of metabolizable energy conversion in pigs peaks at approximately 45%, whereas conversion to fat can be as efficient as 75%.

Cattle are the least efficient converters of feed to meat. Their basal metabolism is higher than that in pigs, and their large body mass and long gestation and lactation periods mean that feed requirements of breeding cattle claim at least 50% more energy than those of pigs and nearly three times as much as those of chickens. The USDA's historic feed/meat data for all of the nation's cattle and calves imply that as little as 3% of gross energy in feed is converted into energy in food.

From the energy makeup point of view, there is a rather poor energy conversion efficiency for meat products, although the situation is not as bad as the conversion efficiency factors in Table III may suggest. A large part of the cattle feed consists of materials that are left over from the human food chain, and cows can convert plant material that humans cannot digest (e.g., grass) into highly valuable food components such as milk and meat. However, we can certainly say that eating animal protein requires significantly more energy and requires more land than does eating vegetable food.

3. TRANSPORT

Crops that are not cultivated under controlled environmental conditions, such as those in heated glasshouses, can be grown only in a certain region during a limited period of the year. However, consumer demand for food is not restricted to these periods. This discrepancy can be overcome in two ways: conservation of the harvested produce and import of products from other regions. The latter option requires energy for transportation, which varies largely with the means and distances of transport (Table IV). As a consequence, major differences in the energy requirements will exist for products that are transported over a long distance: In general, the energy requirement for transport by airplanes is very high, whereas the energy requirement for transport by vessels is relatively low.

TABLE IV

Energy Requirement for Transportation

Means of transport	Energy requirement[a] ($MJ\,kg^{-1}\ 1000\,km^{-1}$)
Seagoing vessel	0.1–0.2
River vessel	0.4–0.7
Train	0.8–1.0
Truck	2–5
Airplane	10–25
Car	50–250[b]

[a] The energy requirement depends on size and load factor.
[b] Depending on petroleum use as well as the amount of products carried (20–40 kg).

TABLE V

Energy Requirements for Some Industrial Preservation Techniques

Preservation technique	Energy requirement ($MJ\,kg^{-1}$ end product)
Drying[a]	0–15
Freezing	5
Heat treatment[b]	5–10

Source. 1996 data from Unilever operations.
[a] The energy requirement ranges from nearly zero for sun drying to approximately 15 MJ kg^{-1} for hotair drying.
[b] Blanching, pasteurization, sterilization.

4. PROCESSING, PACKAGING, AND STORAGE

Many agricultural products undergo some industrial processing before they reach the consumer. Advantages of processed foods over fresh agricultural products include convenience and prolongation of shelf life. In general, industrial food production includes cleaning, mixing, preservation, and packaging operations. Preservation can be achieved by drying, freezing, or heat treatment (Table V). All processing steps require energy, but mechanical handling (interfactory transportation, mixing, and shaping) generally constitutes only 10 to 20% of the total energy requirement. The bulk of the energy

requirement in food processing is caused by heating and cooling required for food preservation.

Several options exist for the packaging of food products. Some food products may require specific packaging materials not only to safeguard food integrity and food safety but also for marketing purposes. Table VI presents data on the energy requirement for the production of the most commonly used packaging materials. When interpreting these data, one should take into account that the packaging for a particular product will require different amounts (weights) of the various packaging materials, thereby complicating a direct comparison of the data in the table. Moreover, the overall energy requirement related to the use of packaging depends on the waste treatment in a particular country due to the energy recovery potential during waste incineration. In case all waste is incinerated with energy recovery, the total energy requirement is reduced as shown in Table VI. Ranges for the overall embodied energy for the ready-to-eat meals, preserved and packed in different ways, are presented in Table VII.

Ready-to-eat meals are prepared in the factory. This preparation requires relatively little energy per portion because the production is large scale. The final product needs to be stored and transported at low temperatures; hence, it requires more energy. Preparation of food at home usually starts with fresh ingredients. In the chain from farm to home, there usually is a significant amount of product loss that does not occur with factory production. In restaurants, the energy requirement and product losses usually are higher than for the two other routes. But in general, the total energy requirements (including the embodied energy of losses in the chain) for food prepared at home and ready-to-eat meals do not differ a great deal.

Finished products are finally stored and distributed to a point of sale. This requires energy for transportation (Table IV). In addition, some food products have to be kept cool or frozen. Average energy requirements for cooling and freezing vary substantially (Table VIII). The most efficient is storage in a well-isolated warehouse, whereas storage in a small open freezing cabinet is very energy demanding. The latter option is often used in shops to stimulate consumers to buy such products. In those cases, priority is given to easy access and not to efficient energy use.

There seems to be a direct relationship between embodied energy and product cost, particularly for commodity food products, as indicated in Fig. 1. Only a few products are situated above the line, representing those products produced in greenhouses with subsidized energy. There are many products below the line, mainly branded and seasonal food products. In those cases, the price that is being paid in the market is higher than the energy equivalent because there also is an emotional value to the product.

TABLE VI

Energy Requirements for the Production of Some Packaging Materials

Material	Energy requirement (MJ kg^{-1})	Recovery potential (MJ kg^{-1})
Glass		—
Virgin	13	—
Recycled (100%)	10	—
Steel		—
Virgin	35	—
Recycled (100%)	20	—
Paper/Cardboard		—
Virgin	45	18
Recycled (75%)	35	18
Plastic	85	45
Aluminum	—	—
Virgin	195	—
Recycled (50%)	100	—

TABLE VII

Embodied Energy for Some Ready-to-Eat Meals

Type of ready-to-eat meal	Embodied energy (MJ kg^{-1})
Cooled, plastic packaging	45–70
Ambient, canned	75–110
Frozen, cardboard packaging	80–115

TABLE VIII

Energy Requirements for Cooled and Frozen Storage

Location	Energy requirement (MJ kg^{-1} $week^{-1}$)
Warehouse	0.01
At home[a]	2–5
Shop[a]	1–10

Source. 1996–1997 data from Unilever analyses.

[a] The wide ranges are due to the variety of cooling and freezing cabinets, ranging from well-isolated facilities to small open boxes.

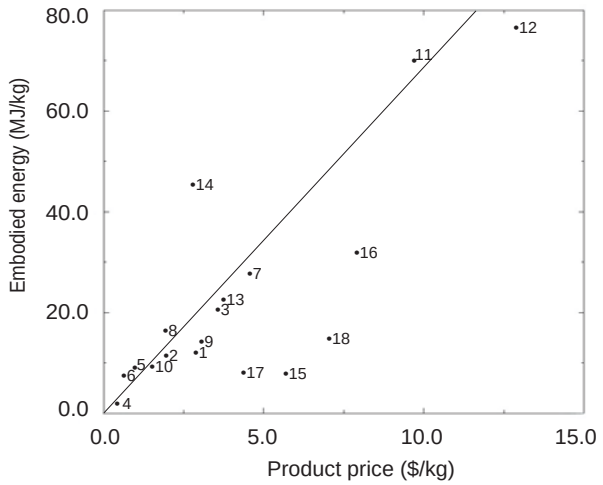


FIGURE 1 Relationship between embodied energy and product price for the following food products: (1) whole-meal bread, (2) wheat flour, (3) rice, (4) potatoes, (5) carrots, (6) onions, (7) jam, (8) sugar, (9) margarine, (10) milk, (11) butter, (12) cheese, (13) eggs, (14) greenhouse tomatoes, (15) beer, (16) wine, (17) fresh vegetables, (18) strawberries.

5. CONSUMPTION

After purchasing the food, a consumer will add to the overall embodied energy of the product by using a car to take the product home, by storing it in a refrigerator, by cooking it, by washing the dishes, and (ultimately) by disposing of the nonconsumed waste. In most countries in the Western world, on average, shopping adds 0.5 to 3.0 MJ kg^{-1} because a substantial part of all shopping is done by car. Nearly 50% of all Dutch consumers use a car to do their supermarket shopping at a distance of approximately 3 km from their homes, buying on average approximately 20 kg of products each time. Dutch consumer habits seem to be comparable to those in other countries, although the distances to the supermarkets are longer in some countries (e.g., France, the United States) and the proportion of motorized shopping differs to some extent.

In-house, some food products can be kept at ambient temperature, whereas others require energy for cooled or frozen storage (Table VIII). The amount of energy for storage depends on the energy efficiency of the refrigerator and/or freezer, its storage capacity, the temperature installed, the amount of product stored, and the frequency of opening. The storage capacity that is required for a household is related to the frequency of shopping. This relation is demonstrated in Table IX, where the energy requirement for four different situations has been calculated. In each case, the same amount of food is purchased per week and the shopping trips are done by car. Cars use petroleum (46 MJ/kg) at an average rate of 6.5 L (4.8 kg or 215 MJ) per 100 km.The distance to the shop has been varied (2 or 5 km), as has the shopping frequency (two or three times/week). Our calculations demonstrate that for small distances between house and shop, the total energy requirement for shopping and storage is not very much influenced by the shopping frequency. At longer distances, the energy requirement for transporting the goods is so high that it pays to have a large refrigerator to store the goods.

Ready-to-eat meals are prepared in a factory. This preparation requires relatively little energy per portion because the production is large scale. In many cases, the final product needs to be stored and transported at low temperatures; hence, it requires more energy. Preparation of food at home often starts with fresh ingredients. In the chain from farm to home, there usually is a significant amount of product loss due to material deterioration and processing inefficiencies. Those losses are usually substantially lower in the case of factory production

TABLE IX

Energy Requirements for Shopping and Storage of Food at Different Shopping Distances and Frequencies

Shopping distance (km)	Shopping frequency ($week^{-1}$)	Energy requirement for shopping (MJ $year^{-1}$)	Storage capacity (L)	Energy requirement for storage (MJ $year^{-1}$)	Total energy requirement for shopping + storage (MJ $year^{-1}$)
2	2	895	235[a]	1235	2130
2	3	1341	155[b]	1020	2361
5	2	2236	235[a]	1235	3471
5	3	3354	155[b]	1020	4374

[a] As 210 L for cooled storage and 25 L for frozen storage; energy requirement 0.4 kWh per day per 100 L.
[b] As 140 L for cooled storage and 15 L for frozen storage; energy requirement 0.5 kWh per day per 100 L.

TABLE X

Energy Requirements for Cooking

Situation	Energy requirement (MJ kg^{-1})
Home	
Homemade	5–7
Ready-to-eat	1–2
Restaurant	10–15

TABLE XI

Energy Requirements for Dishwashing in an Average Dutch Household (2–3 Persons)

Method	Energy requirement per event (MJ)	Daily energy requirement (MJ)
Manual dishwashing		
Water heated by gas	1.0	1.0
Water heated by electricity	2.5–3.0	2.5–3.0
Dishwasher	4.0–5.5	2.5–3.5

due to better controlled logistics. In some restaurants, the energy requirement and product losses are higher than those for eating at home, but that depends very much on the number of customers and the amount of food that has to be discarded. In general, the total embodied energies (including the embodied energy of losses in the chain) for food prepared at home and ready-to-eat meals do not differ a great deal (Table X).

The amount of energy for cooking depends on the method of preparation, the energy source (with gas being more efficient than electricity), the energy efficiency of the cooking device, and the type of recipe. A major additional factor is the number of portions being prepared at the time. In general, the amount of energy required per portion diminishes at higher numbers. For very small numbers, cooking by microwave is to be preferred; for larger numbers, cooking with gas turns out to be most energy efficient.

After a meal, the dishes have to be washed. The energy requirements for this activity depend on the energy needed to heat the water and on whether dishwashing is done manually or with a dishwasher. Dishwashers vary in their energy efficiency. Households that use electrically heated water may save on total energy requirements by using an energy-efficient dishwasher (Table XI). This is due to the fact that the production of electricity as an energy source has an efficiency of only 40%. The data in Table XI are based on 17 L as the average amount of water used for manual dishwashing. The energy used to produce the dishwasher itself is negligible in comparison with the amount of energy that is needed for the actual dishwashing during the life span of the machine (10–15 years).

Finally, at the end of the food chain, consumers dispose of inorganic waste such as packaging materials and organic waste (e.g., leftovers, parts of the foods that were removed during the cleaning and cutting of the ingredients of the meal). On average, some 10 to 15% of the total energy makeup of the foods that are purchased by consumers are not used for consumption but rather discarded. This loss equals an amount of approximately 1000 MJ/year/person.

6. CONCLUSIONS

For all food categories, there is a wide variety of energy requirements over the production chain that are due predominantly to the following:

- *Time and place of consumption in relation to the period and region of production.* Foods that need to be stored for a long period and/or transported over long distances inevitably have higher energy requirements than do foods that are consumed during the season of production close to the area where they originate. However, in some (warm) regions, food can be cultivated much more efficiently than in others, so production of food away from home does not necessarily always lead to increased energy requirements.
- *Production method.* Cultivation under controlled environmental conditions in glasshouses requires additional energy for heating.
- *Degree of processing.* All preservation steps inevitably add to the energy requirements.
- *Consumer food preference.* Consumers can opt for protein sources with high overall energy requirements, such as meat and meat products, or protein from plants with much lower energy requirements.
- *Consumer behavior.* Inadequate shopping, storage, and preparation of food products may add substantially to the energy use of the food system.

SEE ALSO THE FOLLOWING ARTICLES

Aquaculture and Energy Use • Conversion of Energy: People and Animals • Diet, Energy, and Greenhouse Gas Emissions • Fisheries and Energy Use • Food Capture, Energy Costs of • Human Energetics • Industrial Agriculture, Energy Flows in • Lifestyles and Energy • Livestock Production and Energy Use

Further Reading

Bouwman, M. (2000). "Tracking Transport Systems: An Environmental Perspective on Passenger Transport Modes." Ph.D. thesis, University of Groningen, The Netherlands.

Dutilh, C. E., and Kramer, K. J. (2000). Energy consumption in the food chain. *Ambio* **29**(2), 98–101.

European Commission. (1997). "Harmonization of Environmental Life Cycle Assessment for Agriculture," final report, concerted action AIR3-CT94-2028, appendix 3. European Commission, Brussels, Belgium.

Hendrickson, J. A. (1997.). Energy use in the U.S. food system: A summary of existing research and analysis. *Sustainable Farming* **7**(4).

Kramer, K. J. (2000). "Food Matters: On Reducing Energy Use and Greenhouse Gas Emissions from Household Food Consumption." Ph.D. thesis, University of Groningen, The Netherlands.

Smil, V. (2002). Worldwide transformation of diets, burdens of meat production, and opportunities for novel food proteins. *Enzyme Microbial Technol.* **30,** 305–311.

Forest Products and Energy

BRYNHILDUR DAVIDSDOTTIR
Boston University
Boston, Massachusetts, United States

1. **The Forest Products Industry: Organization and Main Processes**
2. **Forest Products and Energy Consumption**
3. **Forest Products and Energy Generation**
4. **The Future: Energy Conservation and Self-Generation Potentials**
5. **Conclusion**

Glossary

black liquor Spent pulping liquor.
charcoal A black porous solid wood fuel that consists mainly of elemental carbon; produced from heating wood in a vessel devoid of oxygen.
cogeneration A process that produces both thermal and electrical energy.
disembodied technological change Technological change that does not require substantial capital investments; influences only the input efficiency of already installed vintages.
embodied technological change Technological change that requires substantial capital investment; influences only the input efficiency of the most recent vintage of the capital stock.
hogged fuels Forest residues from extracting industries such as woodchips that are used to produce energy.
pulpwood Forest products that are used to produce pulp such as woodchips and logs.
wood fuels Forest products that are used for energy generation; include black liquor, fuelwood, charcoal, and residues.
wood pulp Pulp that is produced from virgin forest fibers.

Wood fuels, or energy derived from forest products, are an important contributor to renewable energy supply worldwide. In developing countries, the household sector is the largest consumer of wood fuels. In many developing countries, total consumption of wood fuels is increasing but with a declining fractional share of wood fuels in total national energy supply. Of course, substantial regional and subregional differences exist. In the developed world, industrial consumption of wood fuels has mostly replaced traditional household wood fuel use. The pulp and paper industry is largely responsible for this development through its use of chemical pulping processes. The pulp and paper industry is one of the most energy-intensive manufacturing industries. Over the past half-century, it has gone from primarily purchasing the process energy required to produce pulp and paper toward self-generation from wood fuels often more than half of its energy needs. Now in the beginning of the 21st century, some paper companies are even net energy exporters, primarily due to cogeneration capabilities. Continued transition to self-generation of energy contains important implications for material use in the industry and is influenced by economic, organizational, and technological issues. Consequently, an in-depth analysis of the pulp and paper industry is warranted to understand the prospects for a continued increase in the share of energy supplied from forest products in developed countries.

1. THE FOREST PRODUCTS INDUSTRY: ORGANIZATION AND MAIN PROCESSES

The forest products industry produces two major classes of products: (1) lumber and wood products and (2) paper and allied products. The lumber and wood products sector produces whole logs, lumber, woodchips, hardwood flooring, furniture parts, and the like. It operates primarily on purchased energy, and total energy consumption in this sector is relatively low. In the United States, it accounts for 16% of the total energy consumed by the forest products industry as a whole. Thus, this article

 © 2004 Elsevier Inc. All rights reserved.

focuses on the other half of the forest products industry—the paper and allied products industry.

The paper and allied products industry, or the pulp and paper industry, produces products such as pulp, newsprint, printing and writing papers, and paperboard. The United States currently produces 38% of all pulp produced in the world, followed by Canada, which produces 16%. Correspondingly, the United States produces 36% of all paper and paperboard produced in the world, followed by Japan, which produces 10%. The industry accounts for 0.7% of gross domestic product (GDP) in the United States, but in Finland and Sweden the industry contributes approximately 3.4 and 4.5%, respectively.

The organizational structure of the industry can be divided into three broad sectors: pulp mills, paper and paperboard mills, and finishing. The first two sectors account for more than 97% of all energy use and 100% of all self-generation of energy. In the United States, the industry is mostly competitive and is vertically integrated both upstream and downstream. For instance, it is common for paper companies to own the timber resources they use as a source of raw materials; for example, the three largest paper companies own more than 47% of all privately held timberland. In the United States, it is also common for paper and pulp mills to be located at the same site—what is called an "integrated mill." The capacity of the mills that are integrated tends to be much larger than that of the mills that are not. On average, an integrated mill has a capacity of 1000 tons/day, whereas a paper mill that is not integrated has a capacity of 200 tons/day.

Internationally, capacity utilization rates are extremely high and have been over 90% in the United States for the past quarter-century. Capacity utilization rates are defined as total annual production divided by total annual production capacity. The reasons for those historically high capacity utilization rates are the immense capital intensity of the industry and low profit margins. To remain profitable, the industry must run its machinery very close to full capacity. Because it is too expensive to shut down the equipment or let it lay idle, it is difficult to retrofit already installed machinery, for example, to improve energy efficiency. Consequently, most capital that is installed is used for a long time. This manifests itself in a slow rate of capital turnover and implies that technological improvements occur only incrementally. New improvements tend to be those that can be retrofitted without major fixed-capital expenditures and extensive downtime or those that are considered cobenefits to process or product improvements. This has significant implications for energy use and energy conservation in the industry. The production of paper and paperboard closely follows the organizational structure of the industry and is discussed next.

1.1 Pulp and Paper Manufacture

The two main processes in pulp and paper manufacturing are pulp and paper/paperboard production.

1.1.1 Pulping

Pulping is the process where fibers are separated and treated to produce pulp. The wet pulp is then converted into paper at an integrated pulp and paper mill or is dried and transported from the pulp mill to a paper mill. Different pulping processes are used depending on the fiber material and the desired end product. Fiber materials consist mostly of virgin or waste fibers. Virgin fibers are derived from pulpwood such as from woodchips or logs. Virgin fibers are used to produce a variety of pulps, with the main categories broadly characterized as chemical pulp (where the main process is called kraft pulping), semi-chemical pulp, and mechanical pulp. Pulp mills may also produce pulp from plant materials other than wood, collectively known as nonwoods. Nonwood fiber sources include agricultural residues and industrial fiber crops such as hemp. Wastepaper pulping took off during the past decade. For instance, in Denmark and the United States, waste fibers account for close to 100% and 38% of total fiber use, respectively (Fig. 1).

Pulp mills can use only virgin fibers, waste fibers, or a mix of both. When producing pulp from virgin

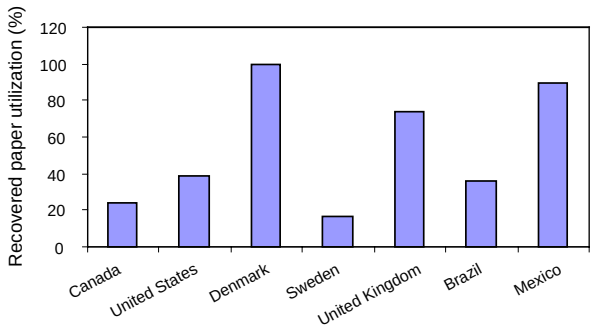


FIGURE 1 Use of wastepaper in selected countries in 2001. Recovered paper utilization rate is defined as the amount of wastepaper used in the production of paper and paperboard divided by the total amount of paper and paperboard produced. From the Food and Agricultural Organization, www.fao.org/forestry.

fibers, the raw materials are first prepared, for example, by debarking and chipping. The next step is the actual pulping process. The primary purpose of pulping is to free the fibers from the lignin that binds the fibers together and/or to separate the fibers in water. Pulp with longer fibers and less lignin is considered to be of highest quality and produces the strongest paper. Mechanical pulping is used for lower grade paper such as newsprint and is the process used for recycled fibers. The advantage of this process is that it gives much higher pulp yields (90–95%) than do other processes but at the same time produces weaker paper because the method yields shorter fibers. Also, if virgin fibers are mechanically pulped, the paper ages faster due to the presence of lignin. Mechanical pulping of virgin fibers is energy intensive and requires 15 million to 16 million BTUs per ton of pulp produced and does not provide the pulping liquors necessary for the self-generation of energy from black liquor. However, the pulping of waste fibers is much less energy intensive and requires approximately 4 million BTUs per ton of pulp.

Chemical pulping is by far the most common method for pulping wood both in Europe and in the United States. In the United States, it accounts for more than 80% of all pulp produced. Chemical pulps have a low yield (40–55%), but the pulp produced is of very high quality. The most common process, the kraft process, accounts for more than 95% of all chemical pulp produced both in the United States and in Europe. In this process, the virgin materials are mixed with a chemical solution, called white liquor, in a digester, where they are pressurized and heated to dissolve the lignin from the wood. After being dissolved, the virgin material is separated into individual fibers and the chemical mixture, now called black liquor, is concentrated in an evaporator. Then, it is burned in a recovery boiler, where the energy is recovered, often cogenerating both steam and electricity, which is then used as process energy. Kraft pulping requires 5 to 7 million BTUs per ton of pulp, but chemical recovery requires an additional 4 to 7 million BTUs per ton of pulp.

Semi-chemical pulping is a combination of chemical and mechanical pulping, where the fibers are chemically treated before they are mechanically pulped. Following any of the pulping processes, the fibers either move on to the bleaching phase to reach a certain level of brightness or move on directly to the papermaking phase. The bleaching phase requires 5.0 to 7.5 million BTUs per ton of pulp.

1.1.2 Papermaking

Different types of pulp are used for the production of various paper products. Given the different energy requirements and opportunities for energy self-generation, the output mix, and thus demand for paper, has major implications for energy use in the industry.

Mechanical pulp is most often used to produce newsprint. Sometimes, chemical pulps are added, for example, to improve strength. Bleached kraft pulp is used mainly in quality printing and writing papers, and unbleached kraft is used to produce linerboard and sack paper. Semi-chemical and chemimechanical pulps typically are used to produce boxboard and sanitary products.

The papermaking stage is made up of paper and paperboard mills and consists of three main processes: stock preparation, paper production, and finishing. First, the pulp is dispersed in water and combined with additives. Then, this slush of pulp, water, and various additives is fed to a paper machine, and eventually dry bulk rolls or sheets of paper or paperboard are produced. The drying process is the most energy-intensive part of papermaking, but the total specific energy use ranges from 3.5 to 10.0 million BTUs per ton of paper depending on the paper type being produced.

2. FOREST PRODUCTS AND ENERGY CONSUMPTION

2.1 Total Energy Use and End Use Efficiency

Internationally, the pulp and paper industry is a major energy consumer and one of the most energy-intensive manufacturing industries. In the United States, only petroleum and coal products production rival its energy intensity (Fig. 2).

The American Paper Institute (API) and later the American Forest and Paper Association (AF&PA) have gathered and analyzed energy consumption in the industry since 1972. According to their data, the industry consumes approximately 12% of total manufacturing energy use. As the industry has grown, total energy consumption has increased on average 1.2% per year. At the same time, the intensity of energy use has declined considerably, from 39.3 million BTUs per ton of paper and paperboard in 1972 to 29.32 million BTUs per ton in 1998. This decline amounts to on average 0.9% per year decline in energy intensity (Fig. 3).

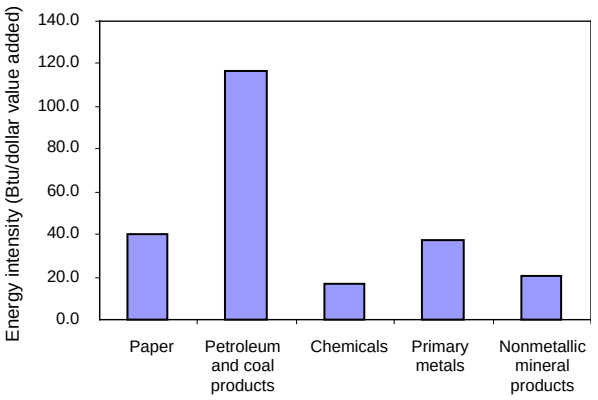


FIGURE 2 Total energy consumption (thousands of BTUs) per dollar of value added by industry. From the U.S. Energy Information Administration, 2002 Energy Consumption Survey, 1998 Annual Survey of Manufacturers.

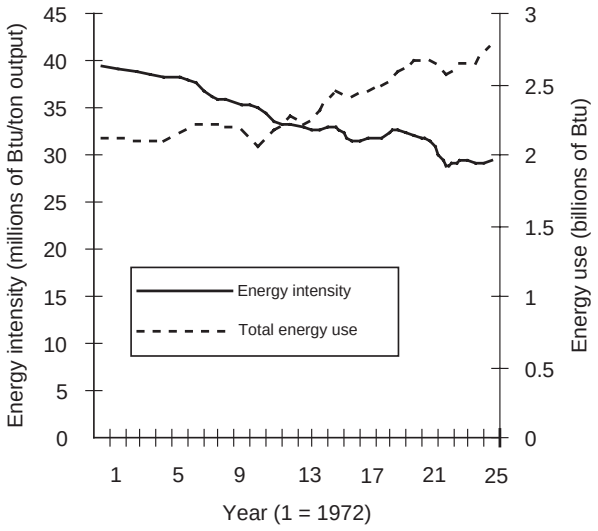


FIGURE 3 Total energy use (billions of BTUs) and energy use per output (millions of BTUs/ton output) in the U.S. paper and allied products industry. From the American Forest and Paper Association, various years.

Despite substantial improvements in energy intensity, total energy use has continued to increase because production growth on average has increased faster than have improvements in energy intensity. As depicted in Table I, which demonstrates the decline in energy intensity in selected countries, energy intensity has declined the most in the United Kingdom and the least in Sweden.

The data in Fig. 2 and Table I depict national annual averages. However, major geographical and temporal differences exist, both in terms of absolute quantities and in terms of changes in total fuel use and energy intensity. Geographical differences rest on

TABLE I

Relative Changes in Energy Intensity Expressed as Annual Changes from 1973 to 1991 (Percentages)

Country	Annual change
United States	−0.9
Japan	−1.9
Germany	−1.4
France	−1.5
United Kingdom	−2.6
The Netherlands	−2.0
Sweden	−0.8
Australia	−1.4
Average	−1.6

Source. Farla *et al*. (1997). Adjusted with permission from Elsevier.

the variation in processes that are used and, thus, on the products produced. They also are influenced by regionally heterogeneous input prices, input availability, public policies, and the like. Studies have shown that the factors having the largest impact on changes in energy intensity and total energy use are changes in output levels, changes in output (and thus process) mix, and changes in production efficiency.

By definition, an increase in output levels, without a change in output mix, increases the use of energy. But based on the relative energy intensities of the processes in use, the total amount of energy used by the industry can either increase or decrease due to a change in the mix of output given that each output category can somewhat be associated with a particular pulping and papermaking process.

The difference in energy intensities of various products rests mainly on three factors: the pulping process that is used (and thus whether chemical recovery is required), the question of whether to bleach the paper or not, and the papermaking process (Table II).

Table II illustrates the relative energy intensities of various paper and pulp manufacturing processes that are in widespread use relative to the least intensive pulping process—wastepaper pulping. The data demonstrate that the energy intensity of different pulping processes can vary by more than 400%.

In addition to the varying energy requirements based on output and process mix, variations in total energy use and energy intensity originate in the amount of energy required to heat air and ground water that is used in the production process. For instance, according to the AF&PA, total energy requirements per ton of output decrease significantly

TABLE II

Relative Energy Intensities of Major Processes Using Current Technologies versus Wastepaper Pulping

Process	Process/wastepaper pulping
Wood preparation	0.12
Pulping	
Kraft pulping	1.92
Sulfite pulping	1.99
Groundwood and thermomechanical	4.09
Dissolving and special alpha	3.71
Semihemical pulping	2.28
Wastepaper pulping	1
Market pulp drying	0.97
Bleaching	1.74
Chemical recovery	
Kraft recovery	2.33
Sulfite recovery	1.75
Semichemical recovery	1.40
Papermaking	
Newsprint	1.90
Printing and writing paper	3.21
Packaging and industrial paper	3.28
Tissue paper	2.60
Paperboard	3.16
Auxiliary	0.35

Source. Energetics (1990). Adjusted with permission from Energetics.

during the summer months in the northern United States by a difference of 15 to 20%.

2.2 Energy-Saving Technological Change

Energy efficiency also has changed due to embodied and disembodied technological change. Energy-saving disembodied technological change is defined as low- or no-cost operational changes in the energy efficiency of the existing capital stock. Embodied technological change requires substantial capital investments through which new technologies diffuse into the capital stock. Such diffusion requires either expansion or replacement investment. Expansion investment, by definition, expands the existing capital stock, enabling increased production levels and increased total energy use, albeit usually at a slightly lower level of energy intensity, assuming new capital is more efficient than existing capital. Replacement investment directly replaces retired capital and includes the investment spent on retrofitted capital. Consequently, replacement investment does not expand the capital stock. By definition, it does not increase total energy use; thus, it enables a faster decrease in energy intensity than does expansion investment, assuming new capital is more efficient than old capital.

Embodied technological change is considered a more permanent change in energy efficiency, even though disembodied change accounts for a substantial part of energy savings, as is the case in the pulp and paper industry. This is because embodied changes are directly imbedded into the capital stock, whereas disembodied gains are more easily reversible. Disembodied changes occur, for example, due to the accumulated experience of employees (learning by doing). Consequently, if the rate of employee turnover increases, the rate of disembodied change slows down.

Historically, energy-saving technological change in the pulp and paper industry has been incremental and a combination of embodied and disembodied changes, an expected feature of a mature industry. This incremental change in energy intensity has provided a continued counter to the increase in output levels and has resulted primarily from gradual process improvements, a shift to wastepaper pulping, and better housekeeping practices. Again, the historical low rate of capital turnover is one of the main culprits in keeping the rate of change incremental. Other reasons include, for instance, that research and development in the industry is very much geared toward improving existing production technologies, as is the case in most mature industries. In the United States, private research and development (R&D) expenditures in the pulp and paper industry are the lowest among all manufacturing industries—and are spent by companies that specialize in consumer products. This suggests that much of R&D expenditures are directed toward incremental product development rather than toward energy intensity improvements. What this indicates is that the incremental development of the industry's knowledge base over time favors already existing processes and that this acts as a barrier to the adoption of new technologies. This is called technology lock-in. Thus, as the industry becomes more mature, it becomes less inclined to update its technologies and, as a result, closely follows a specific predetermined technology trajectory. Second, the total cost of energy is on average lower than that of other inputs such as fibers, and fossil fuel prices in the United States are low and fluctuating. Thus, other investments besides energy-saving investments may seem more profitable and returns less uncertain. Third, the required rate of capital recovery for an investment in energy-saving

equipment—the payback period—is approximately 2 years. Consequently, if an energy conservation project does not offer benefits other than energy cost savings, the payback period must be less than 2 years.

Combining those factors, and taking into account the low profit margins and the necessity to maintain a high capacity utilization rate, makes it hard to justify and execute substantial investments in energy-saving equipment leading to embodied changes in energy efficiency. As a result, the norm in the pulp and paper industry is a gradual decline in energy intensity through efficiency improvements that generally result from improved housekeeping practices or that are side effects of other process- or capacity-related changes.

2.3 Energy Mix

The pulp and paper industry uses two types of end use energy: thermal and electrical. Electrical energy is used primarily to drive pumps and fans. Thermal energy, which is in the form of steam, is used, for example, in digesters, in evaporators, and in the drying of pulp. The industry is in the enviable position of being able to generate itself a large proportion of both the required thermal energy and the required electrical energy. For instance, in the United States, it generates more than 55% of the sum of required electrical and thermal energy, and more than 80% of that energy is generated directly as a by-product of chemical pulping. As depicted in Fig. 4, the remaining energy consumption (collectively called purchased energy) is divided mostly among purchased electricity, coal, natural gas, and residual fuel oil, with a small fraction being allocated to distillate fuel oil and liquid petroleum gas.

The fuel mix in the U.S. pulp and paper industry has changed substantially since 1972. The industry has slowly increased the use of self-generated energy, with an annual average increase of 3.27% for the past 30 years or so. At the same time, the fractional share of purchased fuels has declined. Of purchased fuels, residual fuel oil is slowly being phased out, and the fractional share of natural gas is increasing. The fractional share of coal has remained relatively constant since the early 1980s.

Between 1972 and 1983, the U.S. pulp and paper industry rapidly switched from purchased fuels to self-generated energy, as indicated by the decline in purchased energy intensity and the increase in the intensity of self-generated energy use (Fig. 5).

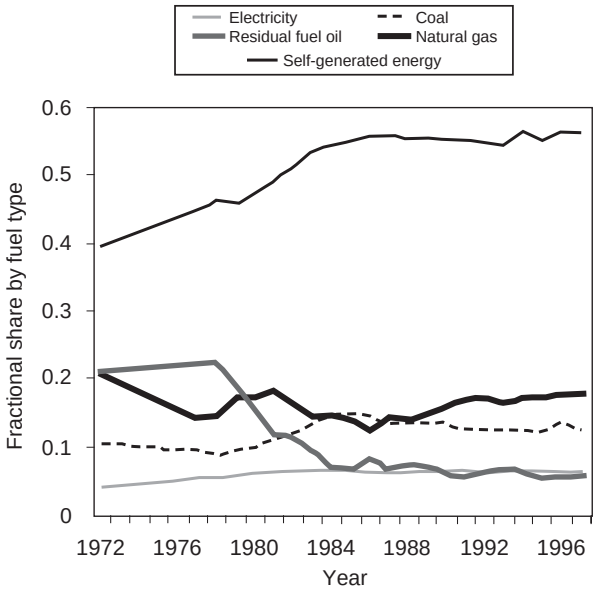


FIGURE 4 Fractional share of fuels by type in the United States pulp and paper industry. From the American Forest and Paper Association, various years.

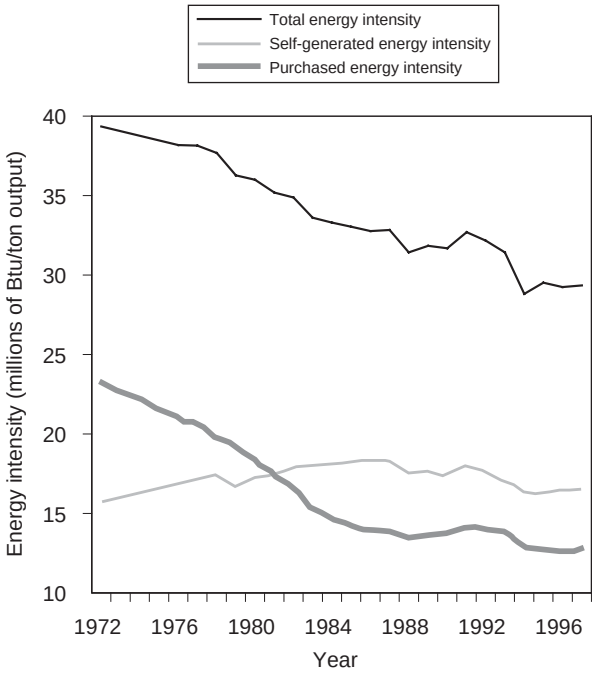


FIGURE 5 Total energy intensity, intensity of purchased energy, and intensity of self-generated energy in the U.S. pulp and paper industry. From the American Forest and Paper Association, various years.

This expansion occurred in part as a result of higher energy prices during the 1970s and a change in the Public Utility Regulatory and Policies Act of 1978 (PURPA). Those two events made cogeneration

of steam and electricity economically efficient. Lower energy prices and the close link between material and energy use reduced the rate of cogeneration after 1983. As previously discussed, more than 80% of the energy that is self-generated is a direct by-product of chemical pulping of virgin fibers. A continued increase in the fractional share of self-generated energy either requires an increase in the share of the chemical pulping of virgin fibers (which has important implications for the use of recycled fibers and the output mix) or requires technological change that will increase the energy and economic efficiency of cogeneration. The former requirement is a major obstacle to an increase in the share of self-generated energy and vividly illustrates the close interconnection between materials and energy cycles in the industry. The latter requirement is a smaller obstacle given that such technology already exists; however, the character of the industry presents a formidable challenge to the ability of this new technology to diffuse into the capital stock. This issue is discussed later in section IV, where the potential for energy conservation and continued expansion into self-generation is explored.

3. FOREST PRODUCTS AND ENERGY GENERATION

As illustrated in section II, and in Figs. 4 and 5, the pulp and paper industry enjoys substantial flexibility in its choice of energy inputs. This historical fuel flexibility of the industry results from its direct access to wood fuels and its heavy use of steam for process heat. Because steam systems can use a wide variety of fuels, the industry has been able to shift with relative ease from residual fuel oil to gas and to substitute bark and forest residues for fossil fuels. In addition, the industry contributes to energy supply indirectly by supplying some of the materials that contribute to methane generation in landfills that can be used for electric power generation and the like.

3.1 Direct Use of Forest Products as Energy

3.1.1 Wood Fuels: Categories and Importance

Wood fuels are a renewable and climate-friendly energy source because they are considered carbon neutral if used and produced in a sustainable and efficient manner. Wood fuels consist of four main commodities: fuelwood (also called firewood), charcoal, black liquor, and wood waste.

Fuelwood is used extensively in most developing countries and is collected from the forest, often as fallen branches or twigs. In domestic applications, it is converted into heat in simple ovens, often for cooking, baking, or drying, and/or is used for space and water heating. The fuel gases are also used for the smoking of food products. Contrary to popular belief, the gathering of fuelwood is not responsible for large-scale deforestation; however, it can contribute to localized forest and soil degradation. Larger issues associated with the traditional use of fuelwood are the health effects of using inefficient stoves in poorly ventilated areas. This leads to emissions of pollutants such as carbon monoxide, methane, and nitrogen oxides that carry serious health risks. For instance, those pollutants are known to cause respiratory infections, adverse pregnancy outcomes, and tuberculosis. In addition, the collection of fuelwood is hard physical labor. Because women are mostly responsible for collecting fuelwood and cooking, women and small children are most adversely affected. Commercial fuelwood is also used as an industrial source of energy in many developing countries. Typical industrial uses include curing and drying, food processing, and lime and brick burning. Although the total amount of wood consumed by the industrial sector is much smaller than that consumed by households, it is nonetheless significant. For example, in Kenya and Sri Lanka, 26 and 18% of fuelwood is used by industry, accounting for more than 70 and 65% of all industrial energy use, respectively.

Charcoal is produced from wood through carbonization or pyrolysis. The wood is heated in a closed container that is devoid of oxygen. The absence of oxygen prevents the wood from burning to ashes; thus, the wood decomposes to charcoal, among other residues. Charcoal plays an important but a much smaller role than does fuelwood in both industrial and household sectors of most developing countries. In households, it is used for tasks similar to those of fuelwood, but a switch from fuelwood to charcoal provides numerous benefits. Those include the fact that charcoal generally does not decompose in storage, it burns without smoke for the most part, it does not create dangerous flames while cooking, and it requires a simple stove whose heat output is relatively easy to control. As a result, a switch to charcoal often represents an increase in the quality of life. The disadvantages of charcoal use are related primarily to the inefficiencies inherent in its produc-

tion and use. Production efficiency in most developing countries ranges between 8 and 28% and is a function of moisture content and type of equipment. Charcoal is also inefficient in use. Although charcoal stoves are more efficient than fuelwood stoves (20–35% vs 10–25%), charcoal stoves are much less efficient than modern fuel stoves such as kerosene (35–50%) and electric stoves (75–85%). The combined production and use inefficiencies have important consequences for the abundance of local wood resources given that cooking with charcoal often uses three to four times more wood than does cooking with fuelwood. Charcoal is also used in industrial settings in both developing and developed countries, for instance, in the production of iron and steel, in metallurgy, and in the chemicals industry.

The other two categories of wood fuels are derived from secondary sources and consist of black liquors from pulp production and wood waste from forest products industries and other products from society such as wastepaper.

Wood fuels account for approximately 7% of the world's total energy supply and are an extremely important source of energy in many developing countries. According to the Food and Agricultural Organization (FAO), developing countries consume approximately 77% of the world's supply of wood fuels, and this in turn accounts for 15% of total energy consumption in the developing world. Consumption in developed countries accounts for the remaining 23%, but this share amounts to 2% of total energy consumption in the developed world. Of course, substantial regional and national differences exist. Developing countries rely mostly on fuelwood and charcoal, but in most developed countries the pulp and paper industry is instrumental in supplying and using such energy. For instance, in 1997 in New Zealand, black liquor accounted for nearly 100% of all energy derived from forest products, and in Japan, Portugal, Sweden, Norway, and Finland, it accounted for 95, 87, 81, 79, and 80%, respectively (Table III). Forest products provided approximately 17% of total national energy supply in Finland and Sweden.

In the United States, black liquor alone contributes 63% of all energy derived from forest products, accounting for slightly over 2% of total energy supply.

TABLE III

Wood Fuel Consumption in 1997 by Selected Countries and Types (Petajoules)

Country	Fuelwood	Charcoal	Black liquor	Total
Bangladesh	307.39	—	0.95	318.34
Burkina Faso	110.26	1.07	—	111.33
Belgium	5.55	0.51	4.75	10.81
Cameroon	135.67	2.68	—	138.35
Finland	40.80	0.12	162.36	203.28
Kenya	193.89	42.17	1.53	237.59
Mexico	233.39	3.74	7.01	244.14
New Zealand	—	—	15.44	15.44
Sweden	38.81	0.49	171.20	210.50
United States	698.64	24.68	1225.50	1948.82

Source. Food and Agricultural Organization (www.fao.org/forestry). Adjusted with permission from the Food and Agricultural Organization of the United Nations.

3.1.2 Wood Fuel Use in the Pulp and Paper Industry

The pulp and paper industry uses not only black liquors but also solid biomass residues such as woodchips, forest residues, wastepaper, and bark. When using solid biomass residues, the industry uses so-called "wood-fired" power boilers (which are similar to coal-fired boilers), where the forest residues are burned to generate steam and/or electricity. Also, it is common for boilers to have multifuel capabilities, enabling the aforementioned fuel flexibility in the industry. Wood has an advantage over natural gas, coal, and oil in that the industry has a cheap supply in the form of wood wastes, bark, and hogged fuels, giving a substantially higher thermal energy content per dollar spent when compared with both oil and coal. Furthermore, wood waste can be burned alongside coal. Disadvantages in use of solid biomass residues are the often high moisture content in the wood waste and low volumetric energy density. Despite those disadvantages, the use of hogged fuels, bark, and wood waste accounts for approximately 20% of self-generated energy in the U.S. industry.

The process of using black liquor to produce energy begins with a process called chemical pulping. In that process, sodium compounds called white liquor are used to cook the virgin wood to release wood fibers from the lignin in the wood. After the woodchips are cooked, they are separated from the lignin/sodium solution, washed, and screened after which they proceed on to the papermaking process. However, the lignin/sodium solution, which now is called black liquor, is transported to a condenser, where it is evaporated to create concentrated strong black liquor. Then, the strong black liquor is transported to a recovery boiler. The most

common model is the Tomlinson recovery boiler. In the recovery boiler, the black liquor is combusted to yield recoverable thermal energy and an inorganic solution of sodium compounds. The inorganic chemicals are extracted as smelt at the bottom of the boiler, giving green liquor. Then, the green liquor is converted to white liquor and reused. The thermal steam energy from the boiler is used directly as process energy, such as for the evaporation of black liquor, and also in driving steam turbines to produce electrical energy, which can be either used on site or sold. This double use of waste heat as steam in the production process and the simultaneous production of electricity is called cogeneration. The industry is the leading manufacturing industry in the United States in the cogeneration of steam and electricity, having produced in 1998 39.4% of all cogenerated electricity.

3.2 Indirect Use of Forest Products as Energy

Indirect generation of energy also occurs as a result of paper waste management practices. In 1995, the U.S. Environmental Protection Agency (EPA) estimated that 49.8 million tons of paper and paperboard were discarded as municipal solid waste in the United States, representing a tremendous amount of usable matter that is buried every year. There are at least two means by which this energy can be harnessed: through the direct incineration of wastepaper (preventing the wastepaper from entering the landfill) or through the capture of methane (which forms in landfills as a result of paper decomposition).

Decomposition of paper and wood products in landfills occurs in a series of stages. The first stage of decomposition is characterized by the removal of oxygen from the waste by aerobic bacteria. Then, the environment becomes oxygen limiting and the bacteria shift to fermentative metabolism. During the second stage, fermentative bacteria hydrolyze polymers into sugars, and those are then fermented further into short-chain carboxylic acids, carbon dioxide, and hydrogen. As acids are produced, the acidity of the refuse increases but then stabilizes due to the removal of acetate and the release of metallic cations that dilute the hydrogen ions. Eventually, this leads to the development of a pH level that is close to neutral. A neutral pH is a necessary prerequisite for the landfill to enter the third stage of decomposition. During this final stage, acetogens further oxidize the remaining short-chain carboxylic acids into acetate, carbon dioxide, and hydrogen. Because the acetogens cannot perform this final stage of decomposition unless the hydrogen ions are removed, this process must occur parallel to the growth of hydrogen-consuming bacteria such as those that produce methane, hence the production of methane.

TABLE IV

Methane Generation Potential of Various Paper Products (Gram Methane/Gram Refuse)

Material	Methane potential
Printing and writing	0.098
Newsprint	0.054
Tissue	0.131
"Other"	0.108

Source. Micales and Skog (1997). Adjusted with permission from Elsevier.

Because lignin inhibits decomposition, the methane generation capacity of paper products varies as a function of the amount of lignin that remains present (Table IV).

Newsprint is produced using mechanical pulp that retains most of the initial lignin; thus, the methane generation capacity is lower for newsprint than for other paper categories. Because tissue is produced using chemical processes, all of the lignin has been removed, hence the highest methane generation potential. Currently in the United States, approximately 10% of the methane generated in landfills is captured for energy purposes, indicating a substantial future potential.

4. THE FUTURE: ENERGY CONSERVATION AND SELF-GENERATION POTENTIALS

The question remains, will the pulp and paper industry be able to become even more self-sufficient in terms of energy generation and even become a net energy exporter and, thereby, increase the role of wood fuels as a source of energy? Will the industry simultaneously be able to reduce total and purchased energy intensity? This section addresses those issues by highlighting technological and economic potentials and the looming barriers to change. A short conclusion follows.

4.1 Technical Potentials

This subsection examines whether technical potentials exist to (1) increase the share of self-generation

of energy and (2) reduce energy intensity. The discussion is based on a study performed at the Lawrence Berkeley National Laboratory by Ernst Worrell, Nathan Martin, and colleagues on U.S. industry potential to increase its energy efficiency and energy self-generation.

4.1.1 Increase in Energy Self-Generation

Black liquor gasification is arguably the most promising new technology to increase self-generation capabilities in the industry. As discussed previously, the cogeneration of electricity and steam usually is performed by recovery boiler/steam-driven turbines. However, the new biomass cogeneration technologies use gas turbines rather than steam turbines. Compared with conventional steam turbines, gas turbine combined cycle cogeneration offers high power to steam ratios and high thermal cycle efficiencies. For instance, power output is three times higher than what is possible with the conventional Tomlinson recovery boiler. Thus, black liquor gasification can potentially have a revolutionary impact on the industry by allowing kraft mills to become self-sufficient in both thermal and electrical energy as well as to become net electricity exporters. Despite high initial capital costs, it is believed that in the long run black liquor gasification systems will prove to be economically efficient.

4.1.2 Improvements in Energy Efficiency

Improvements in efficiency are technologically feasible at every level of pulp and paper production. The discussion in this subsection is a nonexhaustive account of some of the potential improvements in energy efficiency and is intended to illustrate the potential for change. Keep in mind that the average energy intensity of the U.S. pulp and paper industry is close to 29 million BTUs per ton of output. When preparing raw materials, the already available energy-saving technologies, such as ring-style debarkers and cradle debarkers, would save approximately 0.03 million to 0.05 million BTUs per ton of output, and improved screening processes could save up to 0.5 million BTUs per ton of output. Those three investments have a payback period of less than 1 year given U.S. energy prices. In mechanical pulping, biopulping could reduce energy intensity by 3.4 million BTUs per ton of output and could reduce various refiner improvements by 1.6 million BTUs per ton of output. However, the payback period is estimated to be 30 years for biopulping, whereas it is only 3.4 years for refiner improvements. In thermomechanical pulping, potential improvements range from 1 million to 7.5 million BTUs per ton of output via improved heat recovery with a payback period of 4.7 years. Efficiency of chemical pulping can increase by 8.4 million BTUs per ton of output by using continuous digesters with a payback period of 7.7 years. The efficiency of papermaking can increase the most by improving drying technologies such as through the use of Condabelt drying. Energy savings are estimated to be up to 2.40 million BTUs per ton of output with a payback period of 3.8 years. Dry sheet forming, or production of paper without the addition of water, will save up to 5.5 million BTUs of energy per ton of output with a payback period of 88 years. Numerous general measures can also be applied, including installing energy efficient lighting, which will save up to 0.1 million BTUs per ton of paper with a payback period of 3.7 years, and pinch analysis, which optimizes thermal energy flows throughout the plant. This measure is likely to save up to 2.5 million BTUs per ton of paper with a payback period of 1 year. Substantial improvements can also be made to make the production and distribution of steam more efficient, including improved boiler maintenance, which will save approximately 1.7 million BTUs per ton of output with a payback period of 0 years. Other measures include leak repair and steam trap maintenance, with energy savings of up to 0.7 and 2.5 million BTUs and with payback periods of 0.1 and 0.3 years, respectively. Finally, the industry can reduce total primary energy consumption by increasing the use of recycled fibers. This switch might not reduce purchased energy requirements unless the waste pulp is replacing mechanical or thermomechanical pulp.

Clearly, significant economically efficient engineering potentials exist for the pulp and paper industry to increase process energy efficiency. An investment is deemed economically efficient if the payback period is 0 to 3 years. According to the aforementioned study from the Lawrence Berkeley National Laboratory, such cost-effective energy savings amounted to 16% of total primary energy use in the industry in the United States in 1994, and including the expansion of recycling increased that potential to 22%.

4.2 The Barriers to Energy-Saving Potentials

The first culprit, as normally is the case with very capital-intensive mature industries, is the low rate of capital turnover. Because it is extremely expensive to

replace old capital, not just due to capital costs but also due to the required downtime of equipment, capital is kept as part of the capital structure for up to 100 years. Second, energy-related investments compete with other investments for capital, and the relatively low fuel and electricity prices that have prevailed since the mid-1980s have reduced the profitability of energy-saving investments. Combining that with fluctuating energy prices, uncertain returns, and a very short payback time (less than 2 years) explains much of the unused potential for saving energy. Third, the close link among the input materials, output types, and amount and type of energy that the industry can use acts as a limit to the extent to which the industry can switch over to self-generated energy. Fourth, energy savings often seem far from the radar screen of decision makers, who focus more often on improving and extending existing product lines.

5. CONCLUSION

Wood fuels that are derived from forest products remain an important source of renewable, potentially carbon-neutral energy. As countries developed, the role of wood fuels declined and shifted from primarily household use to industrial use. In developed countries, the pulp and paper industry, through its use of chemical pulping, is the main user and provider of wood fuels, thereby keeping them on the horizon as a significant source of energy. The pulp and paper industry is one of the most energy-intensive manufacturing industries, but in the United States the industry self-generates from wood fuels (mostly black liquor) more than 55% of its energy needs. Further extension of energy self-generation partially depends on the rate of diffusion of new technologies, yet the character of the industry poses formidable challenges to technological change, as do economic factors such as low and fluctuating energy prices.

An increase in the share of self-generated energy is also closely tied to the use of specific input materials and output demand because black liquor is a by-product of chemical pulping of virgin fibers. If paper companies increase the use of pulping processes that rely on virgin fibers, they continue to be able to cogenerate steam and electricity from black liquor. An increase in the use of recycled fibers conserves forest resources. This may, on the one hand, be more energy intensive in terms of purchased energy if the waste pulp replaces chemically created pulp. On the other hand, the replacement of recycled pulp for mechanically produced pulp conserves a significant amount of purchased energy.

Because methane has 21 times the global warming potential of carbon dioxide, important trade-offs also exist regarding greenhouse gas emissions. Those trade-offs are linked directly to waste management. If wastepaper is put into landfills, it will generate methane and carbon dioxide. If it is recycled and the waste pulp replaces nonbleached chemical pulps, it may increase purchased energy consumption and, thus, increase carbon dioxide emissions. If it is incinerated, it will not contribute significantly to methane emissions but will not play a role in the conservation of forest resources. It seems that in this particular industrial system, it is challenging to conserve one resource without using more of another or to reduce emissions of one greenhouse gas without emitting more of another.

Despite formidable barriers to change, the future is not bleak for the pulp and paper industry. Major potentials exist for improving energy efficiency, and new cogeneration technology can transform a substantial part of the industry into a net energy exporter—completely reliant on renewable self-generated energy. In the future, the industry could derive large profits from selling energy or, alternatively, could let the utilities sector invest in and operate the new cogeneration technology. Either way, the potential exists for an increase in the energy efficiency of the pulp and paper industry as well as for an increase in the role of forest products, and thus wood fuels, in the total supply of energy in the developed world.

SEE ALSO THE FOLLOWING ARTICLES

Biomass, Chemicals from • *Biomass for Renewable Energy and Fuels* • *Biomass Resource Assessment* • *Coal, Fuel and Non-Fuel Uses* • *Cogeneration* • *Fire: A Socioecological and Historical Survey* • *Wood Energy, History of* • *Wood in Household Energy Use*

Further Reading

American Forest and Paper Association. (various years). "Statistics of Paper, Paperboard, and Woodpulp." AF&PA, Washington, DC.

Biermann, C. J. (1996). "Handbook of Pulp and Papermaking." Academic Press, San Diego.

Energetics. (1990). "Industrial Technologies: Industry Profiles." Energetics, Columbia, MD.

Farla, J., Blok, K., and Schipper, L. (1997). Energy efficiency developments in the pulp and paper industry. *Energy Policy* **25,** 745–758.

Food and Agricultural Organization of the United Nations. (2001). Recovered paper data 2001. Rome, Italy.

Food and Agricultural Organization of the United Nations. (2002). Woodfuel consumption 1997. Rome, Italy.

Gilbreath, K. R., Gupta, A., Larson, E., and Nilson, L. J. (1995). Background paper on energy efficiency and the pulp and paper industry. Prepared for the ACEEE 1995 Summer Study on Energy Efficiency in Industry, Grand Island, NY.

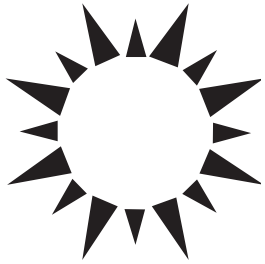
Herzog, H., and Tester, J. W. (1991). Energy management and conservation in the pulp and paper industry. *In* "Energy and the Environment in the 21st Century" (J. Tester, D. O. Wood, and N. Ferrari, Eds.), pp. 437–447. MIT Press, Cambridge, MA.

Martin, N., Anglani, N., Einstein, D., Khrushch, M., Worrell, E., and Price, L. K. (2000). "Opportunities to Improve Energy Efficiency and Reduce Greenhouse Gas Emissions in the U.S. Pulp and Paper Industry," LBNL-46141. Lawrence Berkeley National Laboratory, Berkeley, CA.

Micales, J. A., and Skog, K. E. (1997). The decomposition of forest products in landfills. *Intl. Biodeterioration Biodegradation* **39**(2–3), 145–158.

Smith, M. (1996). "The U.S. Paper Industry and Sustainable Production: An Argument for Restructuring." MIT Press, Cambridge, MA.

Forms and Measurement of Energy



TREVOR M. LETCHER
Natal University
Durban, South Africa
ARTHUR WILLIAMSON
University of Canterbury
Christchurch, Canterbury, New Zealand

1. **Introduction**
2. **The History of the Concept of Energy**
3. **Kinetic and Potential Energy**
4. **Heat and Work**
5. **Measurement of Energy in Terms of the SI System of Units**
6. **Forms of Energy**
7. **Conclusion**

Glossary

energy The capacity to do work.
heat The transfer of energy as a result of a temperature difference.
kinetic energy Energy by virtue of movement.
potential energy Energy by virtue of position.
radiant energy Energy due to electromagnetic vibrations.
work The transfer of energy into a useful form.

An understanding of the physical world requires the concept of energy, or the capacity to do work. Our modern society depends on energy for heating, cooling, transport of all kinds, lighting, and communications, to name but a few vital needs. Thousands of years ago, human energy requirements were limited to the body's own energy and that was fueled by food—the carbohydrates and fats of a hunter–gatherer's diet. Today, the total energy required to ensure the survival and well-being of a person in a modern society is many hundreds of times greater than that of a person in a primitive society.

1. INTRODUCTION

Energy is defined as "the capacity to do work." It can take many forms. We talk of kinetic energy (obtained from ocean waves and from wind), gravitational energy (obtained from falling water), electrical energy (obtained from batteries and turbines), chemical energy (obtained from exothermic reactions such as combustion of diesel and gasoline), thermal energy (obtained from burning coal or wood), radiant energy (from sunlight), and nuclear energy (obtained from fission or uranium atoms or fusion of hydrogen nuclei). Some of these energy forms are more useful than others and many can be transformed from one form to another. For example, the energy from a nuclear reaction can be used to heat water, which in turn can be used to do work by turning a turbine to make electrical energy to heat our houses.

A fundamental law of thermodynamics (the study of the transformations of energy) is that energy is conserved in any process. This implies that when energy transfer does take place, no energy is lost in the process. Usually in the transfer process, a significant fraction of the energy is lost to useless heating of the environment. The efficiencies of conversions can be as low as 10 or 20%. This law of conservation of energy was a major breakthrough in the understanding of energy and has never been violated. One of its implications is that energy cannot be created. In other words, perpetual motion is impossible. This, however, has not stopped hundreds of would-be millionaires from sending their plans to patent offices around the world.

2. THE HISTORY OF THE CONCEPT OF ENERGY

The concept of energy took a long time to develop and was hindered by incorrect and unsubstantiated

 © 2004 Elsevier Inc. All rights reserved.

ideas of Aristotle and many others. The turning point came when Galileo and then Newton, in the 17th century, properly defined the force, f, acting on a body of mass, m, moving with an acceleration, a, as

$$f = ma. \quad (1)$$

It took another two hundred years before energy was correctly formulated by William Thomson in 1851. It was the steam engine that catalyzed the concept. Energy was defined as the work done in moving a body a distance $\times$ against a force, f,:

$$E = w = fx. \quad (2)$$

But it was not until the middle of the 20th century that the separate units for heat (the calorie) and work (the erg) were dropped in favor of the joule and the link between energy, heat and work was finally made.

During the early part of the twentieth century, it was shown that energy can only occur in discrete packets or quanta, and the concept of the quantization of energy was born. For most forms of energy, these packets are so very close to each other that they effectively form a continuum and the sum of all the quantized energies approximates to the energy as defined by Newtonian physics. The quantization of energy becomes important when dealing with small energies related to small particles such as electrons and atoms. In particular, vibration, rotation, and electronic energies are affected and quantum effects must be considered.

3. KINETIC AND POTENTIAL ENERGY

We tend to think of energy in its many manifestations, some of which are given in the introduction. However, in reality, there are only three types of energy: kinetic, potential and radiant energy. The kinetic energy of the body of mass, m, moving with a velocity, v, is defined as

$$E_K = 1/2mv^2. \quad (3)$$

It is the energy a body has as a result of its movement.

The potential energy of a body, on the other hand, is the capacity of the body to do work by virtue of its position. A body of mass, m, dragged up a height, h, against gravity, has the potential energy

$$E_{PE} = mgh, \quad (4)$$

where g is the acceleration due to gravity, which on Earth is about $9.8\,m\,s^{-2}$. This body has potential energy as it rests in position. From that position it can be made to do useful work, such as if it is allowed to fall under gravity while linked to a pulley which lifts a weight.

The link between kinetic and potential energy and other forms of energy is not always clear. Electrical energy, for example, is really the potential energy of electrons in the presence of positive charges. Chemical energy is a sum of the potential energy of electrons in chemical bonds and the kinetic energy of the bonding electrons as they move around an atom or a molecule. Nuclear energy is due to the potential energy and kinetic energy of the subatomic particles within an atomic nuclei.

4. HEAT AND WORK

The terms heat and work are synonymous with energy, and both are simply ways of transferring energy. Heat refers to the transfers of energy as a result of a temperature difference (ΔT). The energy transferred from a hot body of mass, m (in contact with a cold body), as the body's temperature drops from T_1 to T_2 is given by

$$\Delta E_{HEAT} = mS\Delta T, \quad (5)$$

where $(T_1 - T_2) = \Delta T$ and S is the specific heat of the body. The specific heat is defined in terms of Eq. (5) and is the energy required to heat 1 g of the material of the body, through 1 K. For pure water it is $4.18\,J\,g^{-1}\,K^{-1}$. (This is equivalent 1 calorie per gram per Kelvin.)

Work, on the other hand, refers to the transfer of energy into a useful form (i.e. lifting a weight). For example, the work done, w, in lifting a mass, m. through a height, h, against gravity is

$$w = mgh. \quad (6)$$

Work is an agent of transfer and can take many forms. Types of work include expanding a surface, expanding a volume of a gas, pushing electrons along a wire (electricity) and lifting weights.

Heat and work are best understood in terms of the movement of molecules. The molecules of a stationary body at a temperature greater than 0 K (-273.15 K) move in a chaotic manner. If the molecules all moved in the same direction, the body would move! These molecules may rotate, vibrate, or translate. The latter refers to kinetic motion. These movements are collectively known as thermal motion. Molecules in a hot body have a greater thermal motion than the molecules in a cold body. The

transfer of energy from a hot body to a cold body is a result of the stimulation of molecules in the cold body to move more vigorously by the molecules in the hot body. Heat can thus be considered as the transfer of energy that makes use of the chaotic motion of the molecules (see Fig. 1A).

When work is done, the molecules or electrons all move in the same direction in an organized way. When a weight is lifted against gravity, all the molecules of the weight move in the same direction—upward against gravity. When electrical work is done, the working electrons all move in the same direction. Work must therefore be considered as the transfer of energy that makes use of organized motion (Fig. 1B).

Looking back at the historical use of energy, it was relatively easy for early humans to use heat to transfer some forms of energy to thermal motion. For many thousands of years, humans used burning logs of wood and presumably sunlight (radiant energy) on sun-facing caves to warm their caves. The transfer of energy so that work is done is a very much more difficult process and was not really done by machines until the advent of steam engines, which heralded the Industrial Revolution less than 200 years ago. Work, other than that done by human muscles in lifting, pushing, and pulling, is a recent phenomenon.

5. MEASUREMENT OF ENERGY IN TERMS OF THE SI SYSTEM OF UNITS

The measurement of energy involves no less than six of the seven fundamental quantities and units that make up the Système International d'Unités. The quantities and units together with their symbol are given in Table I. All other quantities can be derived from these fundamental ones. For example, acceleration can be written in terms of distance and time (L/T^2), force is the product of mass, and acceleration and energy can be written as the product of force and distance ($ML/^2T^2$); this combination is often called the dimensionality of the quantity. All energy forms have the same dimensionality, as do all units of pressure and all units of force and so on. The unit size of a quantity in the SI is the product of the fundamental units of which the quantity is composed (Table II).

The Newton meter^{-2} (Pascal) is a rather small unit of pressure for many everyday purposes and one more commonly uses the bar (100 000 Pa), which is very close to atmospheric pressure (1 standard atmosphere = 101 325 Pa).

The SI has a set of conventional multipliers, which are used for quantities larger and smaller than the basic units (Table III). The rate of energy flow is the Watt (W)

$$1\ \text{Watt} = 1\ \text{J s}^{-1}.$$

Electrical energy is often metered and measured in kilowatt hours (kWh):

$$1\ \text{kWh} = 3\,600\,000\ \text{J}.$$

The distinction between power (rate of flow of energy) and energy is often badly made especially in

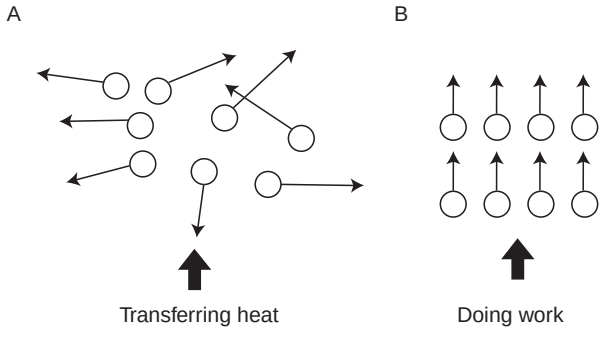


FIGURE 1 (A) Heat is the transfer of energy that makes use of the chaotic motion of molecules. (B) Work is the transfer of energy that makes use of the organized motion of molecules.

TABLE I

The Seven SI Units

Quantity	Name	Symbol
Length	Meter	m
Mass	Kilogram	kg
Time	Second	s
Temperature	Kelvin	K
Amount of substance	Mole	mol
Electric current	Ampere	A
Luminous Intensity	Candela	cd

TABLE II

Some Secondary SI Units

Quantity	Definition	Name	Symbol
Energy	$\text{kg m}^2\,\text{s}^{-2}$	Joule	J
Force	kg m s^{-2}	Newton	N
Pressure	$\text{kg m}^{-1}\,\text{s}^{-1}$		
	N m^{-2}	Pascal	Pa
Power	$\text{kg m}^2\,\text{s}^{-3}$	Watt	W

TABLE III

Prefixes Used in the SI System of Units

Multiplication	Multiple or submultiple	Symbol
10^{12}	tera	T
10^{9}	giga	G
10^{6}	mega	M
10^{3}	kilo	k
10^{2}	hecta	h
10^{1}	deca	da
Base unit		
10^{-1}	deci	d
10^{-2}	centi	c
10^{-3}	milli	m
10^{-6}	micro	μ
10^{-9}	nano	n
10^{-12}	pico	p
10^{-15}	femto	f
10^{-18}	atto	a

reporting electrical power/energy. This confusion frequently leads to the use of the term "kilowatts per hour," which is neither an energy unit nor a unit of power, and kilowatts when kilowatt hours is meant.

Extensive properties, those that are proportional to the amount of substance such as heat capacities are in SI units expressed in terms of energy per mole. Except in chemical practice the mole is little used and extensive properties are frequently expressed as specific properties—that is, properties per unit mass. For example, the molar heat capacity of copper is $24\,\mathrm{J\,mole^{-1}\,K^{-1}}$ or in specific terms $378\,\mathrm{J\,kg^{-1}\,K^{-1}}$.

It is to be noted that if all quantities in a calculation are expressed in SI units then the result will also be in SI units. This coherence is of considerable practical use in ensuring that the results of calculations do not come out in strange units.

6. FORMS OF ENERGY

6.1 Mechanical and Gravitational Energy

Mechanical forms of energy can be considered as those that relate to movement and position. The most important type is translation or kinetic energy and it can be calculated from Eq. (3). Kinetic energy can be found in the energy of the wind and ocean waves.

Wind energy and its transformation into electricity have the advantage of being environmentally friendly. There are no emission of toxic substances, and it is a renewable resource. It is however, not a steady source and furthermore, it is a secondary energy form as it depends on the radiant energy from the sun. All this means that wind energy is a dilute and weak form of energy. In spite of this, a large number of wind turbines are being erected in most continents, and wind energy could be an important renewable energy source in the future.

The amount of energy reaching a wind machine can be calculated from Eq. (3), given the wind speed, the wind collecting area of the vanes, and the density of the air. For example, if the wind velocity is $10\,\mathrm{m\,s^{-1}}$, the wind vane collecting area is $10\,\mathrm{m^2}$, and the density of air is $1.2\,\mathrm{kg\,m^{-3}}$, then the kinetic energy, E_K, per second is

$$E_K = 1/2\,mv^2 = (0.5 \times 10 \times 10 \times 1.3 \times 10^2)\mathrm{J}$$

$$\text{Power} = 6.5 \times 10^3\,\mathrm{J\,s^{-1}}.$$

The harnessing of ocean waves is severely hampered by the corrosive nature of the sea, and very few such projects remain in the world after much enthusiasm in the 1980s and 1990s. The ocean waves are dependent on the moon-earth gravitational attraction and also on wind direction. Wave energy is thus also a secondary energy source and hence is rather dilute compared to primary sources.

The gravitation potential energy between two bodies of mass m, and m_2, separated by a distance r, depends on the force between the bodies, f_G:

$$f_G = \frac{Gm_1m_2}{r^2}, \tag{7}$$

where G is the universal gravitational constant and has a value of 6.673×10^{-11} Newton meter $\mathrm{kg^{-1}}$. The potential energy E_G is thus equal to

$$E_G = \frac{Gm_1m_2}{r}. \tag{8}$$

The acceleration, g, due to gravity on any mass, due to m_2, is given by

$$g = \frac{Gm_2}{r^2}, \tag{9}$$

and on earth, g has a value of about $9.8\,\mathrm{m\,s^{-2}}$. (The radius of the earth is about $6.37 \times 10^6\,\mathrm{m}$ and the mass of the earth is about $6.0 \times 10^{24}\,\mathrm{kg}$.)

The gravitational potential energy of falling water can be considerable if the flow of water and the fall is large. There are many very large power stations

around the world that utilize hydroelectric power. It is a clean energy source, free of chemical pollutants. Unfortunately, there is a limit to the number of places on earth where large volumes of falling water can be found. Hydroelectric power contributes about 2% of the total amount of energy used in the world (a 1994 figure).

There are other types of mechanical energy forms that are scientifically interesting and important but are not on a scale large enough for exploitation as an energy source. These include vibrational, rotational, and electronic energies. The potential energy of a spring is given by

$$E_{VIB} = 1/2kx^2, \tag{10}$$

where x is the displacement of the spring from its rest position and k is the spring constant. Such energies are important in understanding the bonding that takes place between atoms. These vibrational energies are quantized and the available energies are small and given by

$$E_{VIB} = (v + 1/2)h\nu \; v = 0, 1, 2, 3\ldots, \tag{11}$$

where h is Planck's constant (6.626×10^{-34} J s) and ν is the frequency of the vibration This frequency can be calculated from the spring constant and the masses of the two atoms involved in the bond.

Rotational energy is found in molecules and in planets. The rotational energy of molecules is even smaller than the vibration energy of the bonds between atoms. Rotational energy is also quantized and the energies confined to

$$E_{ROT} = \frac{J(J+1)h}{4\Pi I} \; J = 0, 1, 2, \ldots, \tag{12}$$

where I is the moment of inertia about an axis.

Electronic energy is the extra energy an outer electron needs to move to a higher electronic orbit. It is also the energy given out when an excited outer electron moves back to its ground state. This energy is radiant energy and sometimes the emission is in the visible region of the electromagnetic spectrum, and we can see it as a color. When sodium atoms are excited (NaCl in a bunsen-burner flame or electrically excited sodium vapor in a street lamp), they emit an intense yellow flame. Electronic energy is quantized, and the energies are related to the frequency, ν, of the emitted electromagnetic energy by

$$E_{ELEC} = h\nu. \tag{13}$$

The energies of translating atoms and molecules are also quantized but the allowed energy levels are so close together that the energies can be considered as a continuum, and the equations of Newtonian physics apply to these particles. The allowed energy levels for a particle of mass, m, in a one-dimensional box of length L is given by

$$E_{TRANS} = \frac{n^2h^2}{8mL^2} \; n = 1, 2, 3, \ldots. \tag{14}$$

There is another type of mechanical energy that is related to steam engines and to expanding gases or vapors. The work done by a gas or vapor when it expands (with a volume change of ΔV), against a pressure p, is

$$w = p\Delta V. \tag{15}$$

As an example, let us consider the work done by a mole of water at its boiling point, as it turns into a vapor (with a volume change of $\Delta V = 30 \times 10^{-3}$ m^3), against atmospheric pressure (101325 Pa). The work done is

$$w = (101325 \times 30 \times 10^{-3})J = 3040 \text{ J}.$$

This is the energy that drives steam engines.

To put the other larger mechanical energies into perspective, let us consider the kinetic energy of a moving tennis ball and the potential energy of water about to fall.

The energy of a tennis ball (m = 60 g) hit by Serena Williams, traveling at a velocity of 100 km s^{-1}, has a kinetic energy of

$$E_{KE} = 1/2mv^2 = (1/2 \times 0.060 \times 100^2)J = 300 \text{ J}.$$

The potential energy of 1.0 kg of water, about to fall 100 m under gravity (g = 9.8 m s^{-2}) is

$$E_{PE} = mgh = (1 \times 9.8 \times 100)J = 980 \text{ J}.$$

Both energies are rather small.

6.2 Chemical Energy

Chemical energy is stored in molecules through the bonds that bind atoms, and some of this energy can be released when chemical reactions take place. In this process, some bonds are broken (involving an input of energy) and some bonds are made (involving a release of energy). The overall effect is either endothermic or exothermic. However, for most reactions, the process is exothermic and energy is released. This release is usually in the form of thermal energy.

The thermal energy associated with a chemical reaction is always measured at constant pressure and is known as the enthalpy of the reaction, $\Delta_r H^\theta$, and relates to conditions of 1 bar pressure (10^5 Pa)

and 298.15 K. As an example, let us consider the reaction between hydrogen and oxygen:

$$2H_2(g) + O_2(g) = 2H_2O(g). \quad (16)$$

This reaction involves breaking two moles (or molecules) of H–H bonds, one mole (or molecule) of O–O bonds and making four moles of O–H bonds. As the values of the bond energies are known ($\Delta H^{\theta}_{H-H} = 436\,kJ\,mol^{-1}$, $\Delta H^{\theta}_{O-O} = 497\,kJ\,mol^{-1}$ and $\Delta H^{\theta}_{O-H} = 463\,kJ\,mol^{-1}$), it is possible to determine the enthalpy of the reaction, $\Delta_r H^{\theta}$:

$$\begin{aligned}(\Delta_r H^{\theta}) &= (-4 \times 463 + 497 + 2 \times 436)kJ \\ &= -483\ kJ\ (\text{the negative sign implies an exothermic process}).\end{aligned}$$

The experimental enthalpy value for the formation of two moles of $H_2O(g)$ from two moles of $H_2(g)$ and a mole of $O_2(g)$ is indeed -483 kJ.

It is important to know that the sum of the bond energies in a molecule do not make up the total energy of the molecule. The bond energies relate only to chemical reactions. The total energy involves all types of energies, including the energy of binding protons to neutrons and inner electrons and so on. The total energy of a molecule is difficult, if not impossible, to determine.

Chemical energy is the most important type of energy on Earth. It accounts for 96% of all energy sources. These include oil (37%), coal (24%), natural gas (21%), and biomass (14%). These are 1994 figures. Chemical energy supplies the energy for all modes of transport, whether it be ships, cars, trucks, buses, or planes, through the combustion of hydrocarbons such as gasoline and diesel. Chemical energy also supplies energy to most of the electricity power stations in the world through the combustion of natural gas (CH_4), oil, or coal. Typical combustion reactions are

$$\underset{\text{(gasoline)}}{C_8H_{18}(l)} + O_2(g) \rightarrow CO_2(g) + H_2O(l)$$
$$\Delta_r H^{\theta} = -5461\,kJ\,mol^{-1} \quad (17)$$

and

$$\underset{\text{(methane)}}{CH_4(g)} + O_2(g) \rightarrow CO_2(g) + H_2O(l)$$
$$\Delta_r H^{\theta} = -890\,kJ\,mol^{-1}. \quad (18)$$

These reactions involve a relatively large amount of energy for a small mass of fuel. The gasoline reaction (Eq. 17) produces 47900 J per gram of n-octane and the natural gas reaction (Eq. 18) produces 55600 J per gram of methane. The steam engine, by comparison, produces 169 J per gram of H_2O.

Biomass energy refers to the use of plant material as an energy resource. It includes the growing of sugar cane and corn for alcohol production (a viable exercise in Brazil and also in the mid-west United States) and the burning of wood for heat and energy. Much of the energy used in third world countries is of the latter type. Another interesting and useful biomass process is the exploitation of methane gas from landfills. The decomposition of organic material, under anaerobic conditions, produces an equal volume of methane and carbon dioxide gases. The methane gas may be used for heating or to generate electricity. Energy from landfill projects are now commonplace in many countries of the world, notably in the United Kingdom and the United States. The energy is a renewable resource and exploitation does reduce the amount of methane gas (a greenhouse gas, which is 27 times more effective that carbon dioxide) that enters the atmosphere.

The energy used by humans for mechanical activities such as running, lifting weights, pulling, and so forth comes from the combustion of food (carbohydrates, fats, etc.) in much the same way as does the energy of an internal combustion engine. Not a very nice thought. The combustion process in the human body is done through a series of complex reactions, and the overall reaction looks similar to that of burning wood or diesel! Typical carbohydrate and fat reactions are

$$\underset{\text{(carbohydrate)}}{C_6H_{12}O_6(s)} + 6O_2(g) \rightarrow 6CO_2(g) + 6H_2O(l)$$
$$\Delta_r H^{\theta} = -2803\,kJ\,mol^{-1} \quad (19)$$

and

$$\underset{\text{(fat)}}{C_{45}H_{86}O_6(s)} + 63\tfrac{1}{2}O_2(g) \rightarrow 45CO_2(g) + 43H_2O(l),$$
$$\Delta_r H^{\theta} = -27820\,kJ\,mol^{-1}. \quad (20)$$

The carbohydrate (glucose) and the fat produce $15.6\,kJ\,g^{-1}$ and $38.5\,kJ\,g^{-1}$, respectively. The fat produces more than twice the energy per gram that does the carbohydrate. You might have seen these energy values in health food literature as 3.7 and 9.2 kcal per gram, respectively. The calorie is an old-fashioned unit for energy and 4.18 J = 1 calorie. You might also have seen these energy values written as 3.7 and 9.2 Calories per gram, respectively. The Calorie is a health food unit for a kilocalorie.

Chemical energy is measured in calorimeters and the simplest type is known as a "bomb" calorimeter. The bomb is made of thick steel, and is of the order of 100 mm high with an internal volume of about $300\,cm^3$. The only type of reactions that can be investigated in these calorimeters are combustion reactions. The inside of the bomb contains a sample holder and contacts to attach a high-resistance fuse wire. The contacts are connected to a cable through gas-tight seals, to a battery and a switch, outside the bomb. The bomb is also fitted with an oxygen inlet pipe and a pressure valve, rated at 30 atmosphere. The sample (e.g., a few grams of glucose) is weighed out (m_1) and placed in the sample holder, and the fuse wire is connected to the contacts and bent so that it passes through the sample. Oxygen is pumped into the bomb until the pressure is about 20 atmospheres. The bomb is then placed in a tank of water, which has been weighed (about 3 kg) (m_2) and a thermometer and stirrer placed in the tank. Once the bomb and water has reached equilibrium (30 minutes), the temperature of the water is noted (T_1), the fuse is fired through the external switch and battery, and the combustion reaction (Eq. 19) takes place. The heat of combustion heats the bomb, which in turn warms the water. When an equilibrium has been reached (another 30 minutes), the temperature of the water is noted (T_2). The bomb and the water should have the same change of temperature i.e., $(T_2–T_1) = \Delta T$.

The energy given out by the glucose sample is balanced by the energy absorbed by the water and the bomb. If the energy per mole of glucose combustion reaction is $\Delta_r U^{\theta}$, then

$$[(m_g/MM)\Delta_r U^{\theta}]_{glucose} = [m_2 s_2 \Delta T]_{water} + [m_1 s_1 \Delta T]_{bomb}, \quad (21)$$

where m_g is the mass of glucose, MM is the molar mass of glucose, s_2 is the specific heat of the water ($4.18\,J\,g^{-1}\,K^{-1}$), and m_1 and s_1 are the mass and specific heat of the calorimeter. The product ($m_1 s_1$) is unique for each calorimeter (it is called the bomb's heat capacity) and can be determined from calibration experiments using a standard substance such as benzoic acid. The value of $\Delta_r U^{\theta}$ is calculated from Eq. (21). This energy refers to a constant volume experiment. The value of $\Delta_r H^{\theta}$ is usually very close to $\Delta_r U^{\theta}$. The exact value can be calculated from

$$\Delta_r H^{\theta} = \Delta_r U^{\theta} + \Delta n RT, \quad (22)$$

where Δn is the change in the amount of gas in the reaction. For this reaction (see Eq. 19), there are six moles of $O_2(g)$ on the reactant side and six moles of $CO_2(g)$ on the products side. For this combustion reaction $\Delta n = 0$ and $\Delta_r H^{\theta} = \Delta_r U^{\theta}$.

There is another type of chemical energy that should be considered and that is the energy of a phase change. By phase change we mean a solid changing to a liquid or a liquid changing to a vapor. A common example is the change of liquid water to steam. This type of energy has been called the latent heat and is due to intermolecular interactions which bind molecules together as a result of the attractions of dipoles and fluctuating dipoles that are created by the electrons wizzing around molecules. The larger the molecule, the more electrons in the molecule and the greater will be this interaction and the greater will be its latent heat or its enthalpy of phase change. For water, the enthalpies of fusion (melting) at 273.15 K and vaporization at 373.15 K are 6.0 and $40.7\,kJ\,mol^{-1}$, respectively. It obviously takes a lot more energy to separate the liquid molecules to form a vapor than it does to separate the molecules in solid water to form a liquid. Much of the energy (coal) used to fire up a steam engine is used to overcome the enthalpy of vaporization of water and create steam from liquid water.

It is possible to store energy using the reversible property of phase equilibria. For example, solar energy can readily be used to melt Glauber salts ($Na_2SO_4.10H_2O$) because of its low melting temperature (306 K). This molten material will revert to a solid when the temperature drops below the melting point (at night) and will release the stored energy, which is the enthalpy of fusion.

6.3 Electrical Energy

Electrical energy is the most useful of energies as it can be readily transported along copper or aluminum cables for long distances. It can also be conveniently switched on and off. Electrical energy is not unlike the gravitational energy of falling water with the electrons moving under a potential difference replacing water falling under the influence of gravity.

The electrical energy of electrons moving under a potential difference of V with a current i for a time t is

$$E_E = Vit. \quad (23)$$

Electrical energy is readily converted from chemical energy, creating a flow of electrons. Such devices are batteries and one such battery is the Daniel cell. The natural reaction between zinc and a solution of copper sulfate to form zinc sulfate and metallic

copper can be demonstrated by placing a piece of metallic zinc in a solution of copper sulfate whereupon the zinc will begin to dissolve and red metallic copper will be precipitated. The overall reaction can be represented as the displacement of the copper ions from solution as metallic copper by the formation of zinc ions from the zinc in the process of which electrons are transferred from the zinc to the copper ions:

$$Zn(s) + Cu^{2+}(aq) \rightarrow Zn^{2+}(aq) + Cu(s). \quad (24)$$

This reaction can be made to take place in a more controlled way in an electrochemical cell (the Daniel cell), in which zinc and copper electrodes are each immersed in a solution of its sulfate and are separated by a membrane which allows the diffusion of ions. If an external metallic connection (a piece of wire) is made between copper and the zinc rods, an electric current (electrons) can flow between the two rods; zinc metal will dissolve to form zinc ions and the copper ions will deposit as metal on the copper rod. The completion of the flow of electric charge will be achieved by the movement of zinc ions through the membrane and the movement of copper ions toward the copper rod. This electrochemical cell is a prototype of the many different kinds of battery that are used to power various devices from hearing aids to video cameras and flashlights in daily use.

6.4 Radiation of Electromagnetic Energy

Radiation energy is a unique form of energy. It is characterized by a wave with a wavelength and a frequency:

$$c = \lambda \nu. \quad (25)$$

where c is the speed of light ($3 \times 10^8\,m\,s^{-1}$). The magnitude of radiant energy is given by

$$E_R = h\nu. \quad (26)$$

Radiation energy is emitted from an object in the form of electromagnetic waves, which consist of electric and magnetic fields whose amplitudes vary with time. These waves propagate out from the object at the speed of light. Familiar examples of electromagnetic radiation include gamma rays, X-rays, ultraviolet waves, visible waves, infrared waves, microwaves, TV waves, and radio waves. The full range of the electromagnetic spectrum spans cosmic rays at the short wavelength and radio waves at the long wavelength end (see Fig. 2). It is the wavelength that categorizes the different types of radiation. Only a small region of the spectrum is visible to the human eye. That part is called the visible spectrum. The radiant energy that causes sunburn and possible cancer is the ultraviolet rays. These are of shorter wavelength (hence greater energy) than the visible rays and the infrared rays. It is the latter that gives us the sensation of warmth. These rays have similar energies to the vibration energies of the molecules in our bodies and hence are readily absorbed.

The main source of electromagnetic radiation on Earth is the sun. This radiation involves only the UV, visible and IR parts of the electromagnetic spectrum. Sunlight is the mainstay and support for all plant and animal life on earth. The total amount of sunlight reaching the earth each day is 3.11×10^{23} J. This is an enormous amount of energy and 11 days of sunshine over the whole earth is equivalent to all the fossil fuel (gas, oil, coal) in the earth. However, the earth's surface is large and the amount of sunshine reaching the earth peaks at about $1.2\,kW\,m^{-2}$ on the equator at midday. This is really a small amount of energy. A solar energy collector of almost $8\,km \times 8\,km$ is needed to create a reasonably large electricity generating power station of 1.0 GW (assuming a conversion factor of 10%). Even this is unrealistic as it assumes all the solar collecting panels are dust free and clean and that the sun shines brightly 8 hours every day. Solar energy is a dilute form of energy, but it is a renewable and clean energy source and will certainly be an important source in the future as the earth's fossil fuel resources become exhausted.

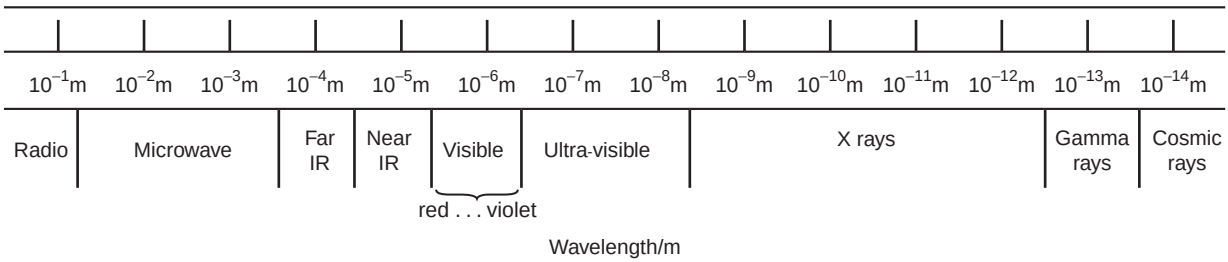


FIGURE 2 The electromagnetic spectrum.

6.5 Nuclear Energy

Nuclear energy comes from reactions involving the nuclei of atoms. In these reactions, the subatomic particles, neutrons and protons of the reacting species, find a more stable arrangement in the products. For nuclear reactions to take place, the reacting species must penetrate the negative electron shield surrounding the nucleus and then penetrate the positively charged nucleus. Thus, the reacting species must have large kinetic energies and preferably should have zero charge. That is why neutrons (zero charge) are often used. In chemical reactions, by contrast, the nucleus is never part of the process and the rearrangement that takes place relate only to the outer electrons of the atoms.

A simplified version of a typical nuclear reaction taking place in our sun is the reaction between two hydrogen isotopes (deuterium) to form an isotope of helium:

$$ {}^{2}_{1}H + {}^{2}_{1}H \rightarrow {}^{3}_{2}He + {}^{1}_{0}n. \qquad (27) $$

Reactions of this type produce large amounts of energy. This particular reaction produces 3×10^8 kJ per gram of deuterium. This is significantly greater (by about six orders of magnitude) than the magnitude of the chemical reactions we met earlier (Eqs. 16 through 20). The deuterium reaction in Eq. (27), involving a fusion of nuclei, is called a fusion reaction. The energy that is released in this reaction and in other fusion reactions comes in many forms: radiation energy (δ-radiation, U-V, IR, and visible light), kinetic energy, and thermal energy.

Another type of nuclear reaction involves the splitting of a nuclei by high-energy neutrons. Such reactions are referred to as fission reactions and the material that undergoes the splitting is called fissile material. Uranium-235, uranium-233, and plutonium-239 are examples of fissile material and all of these are used in nuclear power plants. To ensure fission, the kinetic energy of the neutrons must be high enough to enter the nucleus and split it. For many isotopes of many elements, the arrangement of neutrons and protons in their nuclei is sufficiently stable to ensure that the bombardment by high-energy neutrons has no effect.

A typical fission reaction is

$$ {}^{235}_{92}\mathrm{U} + \mathrm{n} \rightarrow {}^{141}_{56}\mathrm{Ba} + {}^{92}_{36}\mathrm{Kr} + 2{}^{1}_{0}\mathrm{n} $$
$$ \Delta \mathrm{E} = -8 \times 10^{10}\ \mathrm{J\,mol^{-1}}. \qquad (28) $$

The amount of energy is again more than six orders of magnitude greater that typical chemical reactions used in power generation.

If the number of neutrons and the concentration of uranium-235 nuclei is high enough, the nuclear reaction becomes self-sustaining. The minimum mass is known as the critical mass. If the reaction is too vigorous, an explosion will occur. This is what happened in Nagasaki and Hiroshima in 1945 and in Chernobyl in 1986. Nuclear reactors for power generation have strict mechanical control over the uranium-235 enriched rods and moderators to reduce the possibility of runaway nuclear reactions (chain reactions).

The release of energy in nuclear reactions always involves a loss of mass. The relationship between the energy output, ΔE, and the loss of mass, Δm, is given by Einstein's famous equation:

$$ \Delta \mathrm{E} = \Delta \mathrm{m} c^2, \qquad (29) $$

where c is the speed of light, $3.00 \times 10^8\ \mathrm{m\,s^{-1}}$.

In a nuclear reactor, the energy heats a heat exchanger (e.g., water or sodium), which transfers heat to water to make steam, which is then used to turn a turbine and create electricity. Nuclear energy makes up 2% of all the energy used in the world. France has taken the lead here, and 80% of its electricity is produced from nuclear energy. In 1995 there were 400 nuclear powerstations in more than 30 countries. This represented about 10% of the world's electricity.

There are two major problems with fission nuclear power: (1) the products are radioactive and need to be contained for hundred of years because of their long half-lives and (2) there is the possible danger of a runaway chain reaction resulting in an explosion and the subsequent contamination of the environment with radioactive fallout. Solutions to these problems will bring about a resurgence of nuclear power production worldwide, especially as the world's fossil fuel reserves are beginning to run down and the temptation to extract large amounts of nuclear energy from small masses of fuel becomes irresistible.

Fusion reactions would, in theory, be the saving grace of the nuclear industry because the products are essentially free of radioactivity and, moreover, the fuel (deuterium) is available from seawater. Unfortunately, the fusion process is difficult to achieve in practice because of the difficulty in supplying the deuterium molecules with enough energy for the nuclei to strike each other with enough force to allow them to interact and for the neutrons and protons to rearrange. Much work is currently being done in a number of laboratories

around the world, but the process is costly and technically challenging.

Geothermal energy (the energy that heats natural hot-water springs) is due to energy from radioactive sources deep inside the earth. This energy is being tapped in Iceland and in Italy.

The measurement of the energy from nuclear reactions can be determined from a special type of calorimeter, but it can also be calculated from a detailed analysis of the relative atomic masses of the products and reactants, followed by the application of Einstein's equation (Eq. 29).

7. CONCLUSION

A knowledge of the forms of energy available on our earth (and a knowledge of how to measure and calculate the magnitudes of each type of energy) gives us the necessary data to begin to make the important decisions as to which type of energy can be exploited in the future. This is particularly pertinent as the world's fossil fuel reserves begin to dwindle.

SEE ALSO THE FOLLOWING ARTICLES

Combustion and Thermochemistry • Conservation of Energy Concept, History of • Electrical Energy and Power • Electromagnetism • Energy in the History and Philosophy of Science • Entropy • Heat Transfer • Mechanical Energy • Nuclear Power, History of • Temperature and Its Measurement • Thermodynamic Sciences, History of • Work, Power, and Energy

Further Reading

Atkins, P. (2003). "Galileo's Finger—The Ten Great Ideas of Science." Oxford University Press, Oxford.

Atkins, P., and de Paula, J. (2002). "Physical Chemistry." Oxford University Press, New York.

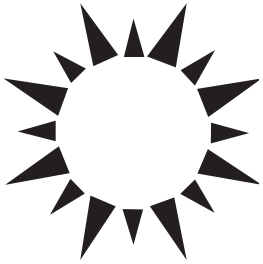
Atkins, P., and Friedman, R. S. (1997). "Molecular Quantum Mechanics." Oxford University Press, Oxford.

Boyle, G. (ed.). (1996). "Renewable Energy: Power for a Sustainable Future." Oxford University Press, Oxford.

Ramage, J. (1997). "Energy, a Guide Book." Oxford University Press, Oxford.

Fuel Cells

NIGEL BRANDON
Imperial College
London, United Kingdom



1. **What Is a Fuel Cell?**
2. **Fuel Cell Types**
3. **Fuel Selection**
4. **Fuel Cell Applications**
5. **Fuel Cell Costs**
6. **Environmental Benefits**

Glossary

anode The fuel electrode, where electrons are donated as the fuel is oxidised. The negative electrode in a fuel cell.
balance of plant All the components that make up a power system, other than the fuel cell stack, for example, the control system, reformer, power conditioner, or compressor.
bipolar plate A dense, electronic (but not ionic) conductor that electrically connects the anode of one cell to the cathode of another. It also distributes fuel or air to the electrodes.
cathode The air electrode, where electrons are accepted and oxygen is reduced. The positive electrode in a fuel cell.
electrocatalyst A catalyst incorporated into both the anode and the cathode to promote the electrode reactions.
electrolyte A dense ionic (but not electronic) conductor. Each fuel cell type is distinguished by the nature of the electrolyte used within it.
interconnect Another term for the bipolar plate (see above).
membrane electrode assembly (MEA) An assembled anode, cathode, and electrolyte.
open circuit voltage The voltage from the fuel cell when no current is being drawn from it.
overpotential The voltage loss within an operating fuel cell due to electrode kinetics, mass transport limitations, and component resistance.
positive electrolyte negative (PEN) The assembled cathode, electrolyte, and anode of a solid oxide fuel cell.
reformer The fuel processor that converts the fuel to a hydrogen-rich gas suitable for the fuel cell.
stack An assembly of many individual fuel cells, complete with gas manifolds and electrical outputs.

A fuel cell is a device for directly converting the chemical energy of a fuel into electrical energy in a constant temperature process. Fuel cells operate on a wide range of fuels, including hydrogen, and are seen as a clean, high-efficiency power source and an enabling technology for the hydrogen economy.

1. WHAT IS A FUEL CELL?

The fuel cell can trace its roots back to the 1800s when a Welsh-born, Oxford-educated barrister, Sir William Robert Grove, realized that if electrolysis, using electricity, could split water into hydrogen and oxygen, then the opposite would also be true. Grove subsequently built a device that would combine hydrogen and oxygen to produce electricity—the world's first gas battery, later renamed the fuel cell.

It was another British scientist, Francis Thomas Bacon, who was the first to develop a technologically useful fuel cell device. Bacon began experimenting with alkali electrolytes in the late 1930s, using potassium hydroxide instead of the acid electrolytes used by Grove. Bacon's cell also used porous "gas-diffusion electrodes" rather than Grove's solid electrodes. These increased the surface area over which the electrochemical reactions occurred, improving power output. In the early 1960s, Pratt and Whitney licensed the Bacon patents and won the National Aeronautics and Space Administration contract for the Apollo spacecraft utilizing onboard liquid hydrogen and oxygen to provide both power and water. Innovation and development has continued since that time with pioneering work by, for example, Westinghouse, Union Carbide, and General Electric, among others.

Today, potential applications for fuel cells range from battery replacement in consumer goods and portable computers, through residential scale combined heat and power (CHP), to distributed energy generation. However, the key problem limiting the

 © 2004 Elsevier Inc. All rights reserved.

significant commercial uptake of fuel cells is cost, and it is cost reduction, together with the need to demonstrate high levels of reliability and durability, that are the primary concerns of fuel cell developers today.

As demonstrated by Grove's gas battery, a fuel cell is analogous to a battery, but one that is constantly being recharged with fresh reactants. In a similar manner to a battery, each fuel cell comprises an electrolyte, which is an ionic conductor, and two electrodes (the negative anode and positive cathode), which are essentially electronic conductors.

The nature of the ion transfer varies between the different types of cell, but the principle shown in Fig. 1 for a polymer electrolyte membrane fuel cell (PEMFC) is representative. In this case, hydrogen is fed to the anode of the cell where it splits into a proton and electron, the former passing through the electrolyte and the latter forced around an external circuit where it drives a load. The proton and electron combine with oxygen from the air at the cathode, producing pure water and a small amount of heat. The overall reaction is thus

$$H_2 + 0.5O_2 \Leftrightarrow H_2O. \quad (1)$$

The anode is then the negative electrode of the device, and the cathode the positive.

The fuel cell differs from a conventional heat engine (such as the internal combustion engine or the gas turbine), in that it does not rely on raising the temperature of a working fluid such as air in a combustion process. The maximum efficiency of a heat engine is subject to the Carnot efficiency limitation, which defines the maximum efficiency that any heat engine can have if its temperature extremes are known:

$$\text{Carnot efficiency} = (T_H - T_L)/T_H, \quad (2)$$

where T_H is the absolute high temperature and T_L is the absolute low temperature. In contrast, the theoretical efficiency of a fuel cell is related to the ratio of two thermodynamic properties, namely the chemical energy or Gibbs free energy (ΔG^0) and the enthalpy (ΔH^0) of the fuel oxidation reaction:

$$\text{Maximum fuel cell efficiency} = \Delta G^0/\Delta H^0. \quad (3)$$

Figure 2 provides an illustration of the theoretical efficiency possible from a fuel cell running on hydrogen and air as a function of temperature and compares this to the Carnot efficiency of a heat engine at the same temperature, assuming a low temperature of 25°C. As the Gibbs free energy for reaction (1) falls with increasing temperature, while the enthalpy remains largely unchanged, the

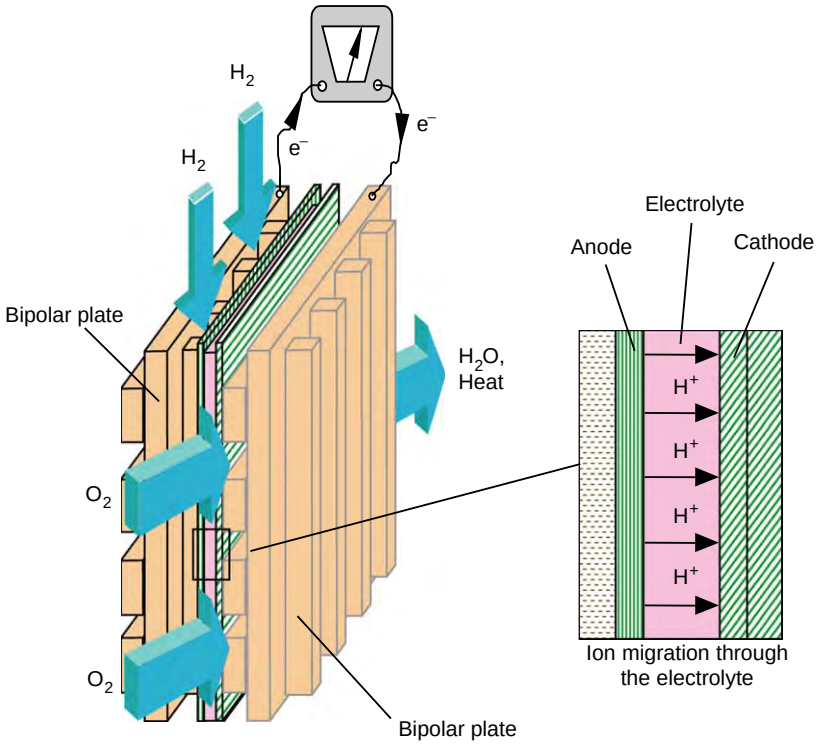


FIGURE 1 Schematic illustration of a polymer electrolyte fuel cell.

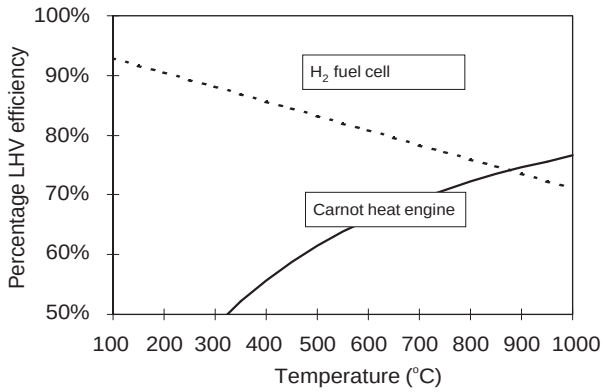


FIGURE 2 Maximum efficiency (on a lower heating value basis) of a hydrogen/air fuel cell, and a heat engine limited by the Carnot cycle, as a function of temperature.

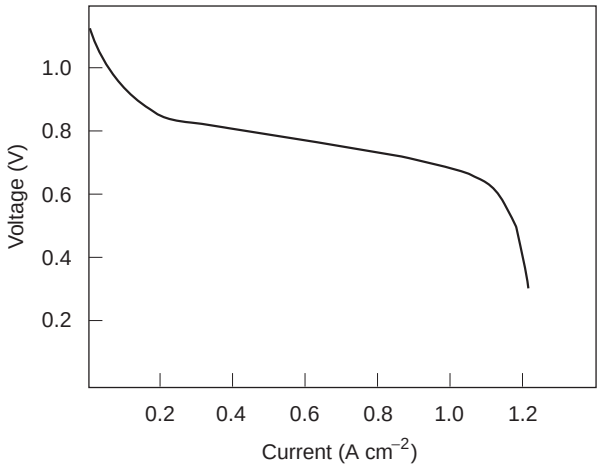


FIGURE 3 Schematic illustration of a typical fuel cell current-voltage curve.

theoretical efficiency of the fuel cell falls with increasing temperature. Indeed, at high temperatures the theoretical efficiency of a heat engine is higher than that of a hydrogen driven fuel cell. However, because of the need for motion in a heat engine, either rotary or linear, significant materials issues are associated with operating them at high temperatures, from the perspective of both durability and cost. Fuel cells do not have moving parts operating at high temperatures and thus are less susceptible to this problem.

However, other factors play a role in determining the actual efficiency of an operating fuel cell, in particular operating temperature, fuel type, and materials selection. For example, losses associated with the kinetics of the fuel cell reactions fall with increasing temperature, and it is often possible to use a wider range of fuels at higher temperatures. Equally, if a fuel cell is to be combined with a heat engine, for example, in a fuel cell/gas turbine combined cycle, then high fuel cell operating temperatures are required to maximize system efficiency. All these factors mean that there is considerable interest in both low-temperature and high-temperature fuel cells, depending on the application.

Figure 3 illustrates the shape of the current-voltage characteristics that would be expected from a typical fuel cell. When no current is being drawn from the fuel cell, the cell voltage is at a maximum, termed the open circuit voltage (E), which increases with the partial pressures of the fuel and air gases and decreases with increasing temperature, according to the Nernst equation:

$$E = E^0 - RT/nF \ln\left(pH_2O/pH_2\ pO_2^{0.5}\right). \quad (4)$$

where E^0 is related to the Gibbs energy for the reaction via:

$$\Delta G^0_{298} = -nE^0F, \quad (5)$$

where n is the number of electrons involved, 2 for reaction (1), and F is the Faraday constant ($96495\,C\,mol^{-1}$). As the value of ΔG^0_{298} for the reaction of hydrogen with oxygen to form water is $-229\,kJ\,mol^{-1}$, then an open circuit voltage of around 1.2 V would be expected from a hydrogen/air fuel cell operating at near ambient temperatures under standard conditions.

As current is drawn from the cell, additional irreversible losses result in a decrease in the cell voltage (E_{cell}), according to:

$$E_{cell} = E - iR - \eta a - \eta c, \quad (6)$$

where iR refers to ohmic losses within electrodes, interconnects, and current take-off's due to the finite resistance of the materials used, ηa refers to the overpotential at the anode, reflecting losses due to both electrode kinetics and mass transport limitations, and ηc refers to the equivalent overpotential at the cathode. Different loss terms are reflected in different regions of the current-voltage curve, with the initial fall in voltage reflecting electrode kinetics, the central linear region being dominated by iR losses, before mass transport limitations dominate at high current densities.

Hence, under load a single cell produces a reduced cell voltage due to the losses highlighted here. It is the task of the fuel cell designer to minimize these losses by judicious selection of materials and cell geometry. Electrodes, for example,

are required to be porous to enable gas transport to and from the active catalyst region adjacent to the electrolyte, yet also to be good electronic conductors. As such, there is considerable research into electrode structures and materials, so as to minimize overpotentials. Another choice facing the fuel cell system designer is the operating cell voltage.

At any point other than very close to the open circuit voltage (where parasitic loads become significant), then, the higher the cell voltage, the higher the cell efficiency—but the lower the power density. In general, a cell voltage of about 0.6 to 0.7 V is taken as the operating point. This is a compromise between efficiency (a high-efficiency reducing fuel consumption) and power density (a low-power density increasing stack size and weight, and hence capital cost). To achieve a useful output power, individual cells are connected together in a "stack". This is achieved using an interconnect or bipolar plate, which joins the anode of one cell to the cathode of the next cell. The interconnect also separates and often distributes the fuel and oxidant. An example of a PEMFC stack is shown in Fig. 4. In this case, the interconnect is known as a flow field plate, and the combination of electrolyte and electrodes as the membrane electrode assembly (MEA). Other fuel cell types use different terminology (positive electrolyte negative or PEN for an SOFC, for example) for these components, but their function remains the same.

The fuel cell stack is then incorporated into a system, which meets the demand of a specific application. Fuel cell systems can include a range of balance of plant (BoP), depending on the fuel used, the application, and the fuel cell type. BoP can include a fuel processor, compressor, power conditioner, and control system. Together, these can add significant cost to the overall system. As well as high efficiency and low emissions levels (discussed further in Section 6), a number of further advantages are cited for fuel cells:

1. They are quiet and involve few moving parts, other than some fans or a compressor to blow air into the device, and hence do not require much maintenance.
2. They are modular, such that several can be coupled together to increase the capacity of a system, but can be mass-manufactured to reduce cost.
3. They exhibit an increase in efficiency at low loads, unlike a heat engine, which normally only exhibits maximum efficiency around the design point for the device.

2. FUEL CELL TYPES

There are five main classes of fuel cell, each with differing characteristics, and differing advantages and disadvantages. The five types are summarized in Table I, with each taking the name of the electrolyte used in its fabrication. These five classes of fuel cell can essentially be further grouped into one of two classes, distinguished as either low-temperature fuel cells (AFC, PEMFC, PAFC), or high-temperature fuel

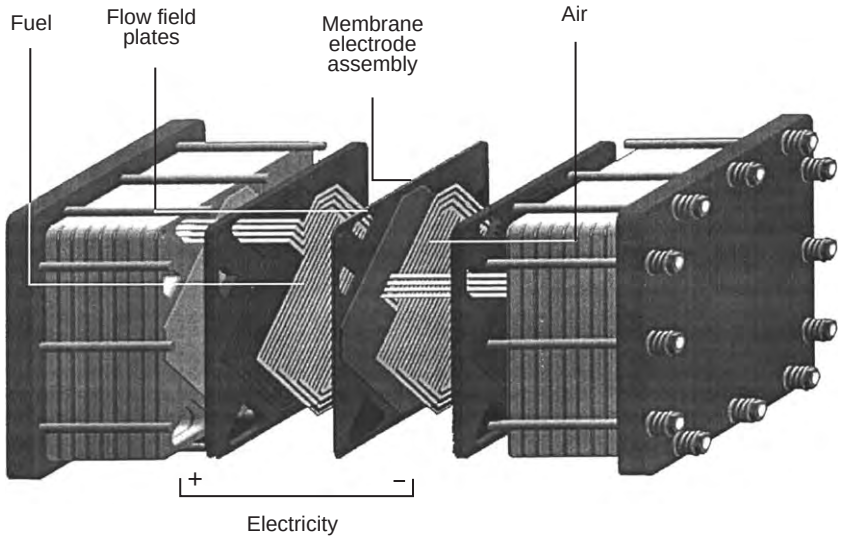


FIGURE 4 Example of a polymer electrolyte fuel cell (PEMFC) stack.

TABLE I

Summary of the Five Main Fuel Cell Types

Fuel cell type	Alkaline	Polymer	Phosphoric acid	Molten carbonate	Solid oxide
Acronym	AFC	PEMFC	PAFC	MCFC	SOFC
Operating temperature	60–250°C	80–120°C	150–220°C	600–700°C	600–1000°C

cells (MCFC, SOFC). The low-temperature fuel cells can be distinguished by the following common characteristics:

- They require a relatively pure supply of hydrogen as a fuel (e.g., AFCs are sensitive to carbon dioxide, PEMFCs to carbon monoxide). This usually means that a fuel processor and some form of gas cleanup is required, adding cost and reducing system efficiency.
- They incorporate precious metal electrocatalysts to improve performance.
- They exhibit fast startup times.
- They are available commercially (AFC, PAFC) or are approaching commercialization (PEMFC).

In contrast, high-temperature fuel cells can be classed as having the following general features:

- They have fuel flexibility; they can be operated on a range of hydrocarbon fuels at high efficiency.
- Their increased operating temperature reduces the need for expensive electrocatalysts.
- They can generate useful "waste" heat and are therefore well suited to cogeneration applications.
- They exhibit slow startup times.
- They can require expensive construction materials to withstand the operating temperature, particularly in the balance of plant.
- Reliability and durability are concerns, again due to the operating temperature.
- They are suitable for integration with a gas turbine, offering very high efficiency combined cycles.
- They are further from commercialization, though a significant number of demonstrators are in operation.

The previous summary is, of course, a generalization. For example, a significant amount of work is ongoing to develop intermediate temperature SOFCs (IT-SOFCs), which operate at temperatures toward 600°C and which aim to overcome a number of the disadvantages of high-temperature fuel cells cited here. Similarly, work also is being carried out to develop low-temperature fuel cells that operate directly on methanol rather than clean hydrogen, eliminating the need for a fuel processor. Nonetheless, it is reasonable to distinguish between the low-temperature and high-temperature variants as being broadly best suited to transportation and stationary power applications, respectively, applications that place differing requirements on the fuel cell stack and system. Each of the main fuel cell types is discussed in more detail in the following sections.

2.1 Alkaline Fuel Cells (AFCs)

The alkaline fuel cell has a long history in the space program. It is still used in the space shuttle in an expensive guise, producing power for the onboard systems by combining the pure hydrogen and oxygen stored in the rocket-fuelling system.

The electrolyte is concentrated (85 wt%) potassium hydroxide (KOH) in AFCs operated at 250°C, or less concentrated (35 to 50 wt%) KOH for operation below 120°C. The electrolyte is retained in a porous matrix (typically asbestos), and electrocatalysts include nickel and noble metals. The main difficulties with this fuel cell type are (1) carbon monoxide, always found in hydrogen produced by reforming hydrocarbon or alcohol fuels (see Section III), is a poison to the precious metal electrocatalysts, and (2) carbon dioxide (in either fuel or air) will react with the KOH electrolyte to form potassium carbonate. As such, applications are essentially limited to those where either pure oxygen and hydrogen can be used.

2.2 Polymer Electrolyte Fuel Cells (PEMFCs)

PEMFCs have high-power density, rapid startup, and low-temperature operation (around 80 to 120°C), and so are ideal for use in applications such as transport and battery replacement. The electrolyte used is a proton conducting polymer. This is typically a perfluorinated polymer, though other hydrocarbon-based membranes are under development in an attempt to reduce cost or to enable

operation at temperatures approaching 200°C. The catalytically active layer sits adjacent to the membrane, supported on a PTFE treated carbon paper, which acts as current collector and gas diffusion layer. For operation on pure hydrogen, platinum is the most active catalyst, but alloys of platinum and ruthenium are used when higher levels of carbon monoxide are present (CO is a poison in all low-temperature fuel cells). Water management in the membrane and electrodes is critical for efficient performance, because the membrane must be hydrated, while avoiding flooding of the electrode pores with water.

The present cost of PEMFC systems is high. While mass-manufacturing techniques will help bring down those costs, further technical innovation is also needed. The potential for replacing diesel standby generators that are often noisy, expensive, and unreliable is engendering great interest. As many countries move toward liberalization of their energy systems, the opportunities for decentralized generation are also growing. Demonstrator plants of 250 kWe are also being produced and operated. These run on natural gas using a reformer, for the main part, and offer very low emissions but high efficiency as their benefit.

2.3 Phosphoric Acid Fuel Cells (PAFCs)

Phosphoric acid fuel cells have been very successful in fuel cell terms over the past 5 years as stationary cogeneration plants, with more than 220 commercial power plants delivered. The PAFC fleet has demonstrated over 95% + availability, and several units have passed 40,000 hours operation. PAFCs operate at 150 to 220°C, using a 100% phosphoric acid electrolyte retained in a silicon carbide matrix. Generally platinum electrocatalysts are used in both anode and cathode.

Typical applications lie in hospitals, where the waste heat can be used in laundry and other areas and where consistent and reliable power is required; in computer equipment power provision, where the absence of power surges and spikes from the fuel cell enables systems to be kept running; and in army facilities and leisure centers that have a suitable heat and power requirement. While the PAFC is by far the most commercial of the fuel cells to date, it may well be superseded in the longer term by PEMFC plants that can potentially be produced more cheaply and by SOFC or MCFC plants that have more useful heat output and that operate at higher efficiencies on hydrocarbon fuels such as natural gas.

2.4 Molten Carbonate Fuel Cells (MCFCs)

The electrolyte in the MCFC is usually a combination of alkali carbonates, retained a ceramic $LiAlO_2$ matrix. The temperature of operation is 600 to 700°C, where the alkali carbonates form a highly conducting molten salt, with carbonate ions providing the means for ionic conduction. The increased temperature of operation means that precious metal electrocatalysts are not needed, and generally nickel anodes and nickel oxide cathodes are used. A design constraint with the MCFC is the need for CO_2 recirculation, meaning that it is difficult to operate below the 100 kWe's scale, removing the market in micro combined heat and power (micro-CHP).

2.5 Solid Oxide Fuel Cells (SOFCs)

Solid oxide fuel cells operate at elevated temperatures, generally above 800°C for the all ceramic high-temperature variant and in the range 600 to 800°C for metal-ceramic intermediate-temperature solid oxide fuel cells. The electrolyte is a dense ceramic, usually yttria stabilized zirconia (YSZ), which is an oxide ion conductor at elevated temperatures. The cathode is typically a perovskite material such as strontium doped lanthanum manganite, often mixed with the YSZ in the form of a composite. The anode is a cermet of nickel and YSZ. The main difference between the high temperature SOFC and the IT-SOFC lies in (1) the thickness of the electrolyte, which tends toward 20 μm thick films for IT-SOFCs to reduce ionic resistance, and (2) the interconnect material, with stainless steel being used at the lower temperatures of the IT-SOFC, whereas more expensive high chrome alloys, or oxides such as lanthanum chromite, are needed at higher temperatures.

SOFCs lend themselves to applications in which their high-temperature heat can be used. This heat can be used in two basic ways: for heating processes such as those in industry or in homes or for integration with turbines in hybrid cycles for very high efficiency electricity production. Recent advances in microturbines have led to the concept of a combined power plant of 250 kWe with an efficiency approaching 70%. Specific applications in which SOFCs may be used are in decentralized electricity generation of 250 kWe to 30 MWe, industrial cogeneration of 1–30 MWe, or domestic applications of 1–5 kWe. Intermediate temperature SOFCs are also of interest for vehicle auxiliary power unit (APU) applications, operating on diesel or gasoline. Carbon

monoxide is not a poison for SOFCs, meaning that a wide range of fuels can be used, together with a simpler, and therefore cheaper, fuel processor. It is also possible to recuperate heat from the fuel cell within the fuel reformer, improving system efficiency when compared to low temperature fuel cells when operating on hydrocarbon fuels.

3. FUEL SELECTION

All fuel cells can run on hydrogen as a fuel. This is combined with oxygen, normally fed into the fuel cell as air, to form water, as shown in reaction (1). However, high-temperature fuel cells can also run on other fuels, especially hydrogen-rich gases, and some low-temperature systems are able to run on specific liquid fuels. The application of the fuel cell system often determines on which fuel it will run. Although fuel cells require comparatively clean fuels, they are also flexible in which ones they will accept. High-temperature fuel cells will operate directly on hydrogen-rich gases such as methane and are being successfully tested on forms of biogas. They, and their lower temperature cousins, will also accept any form of gas produced from liquid or solid fuels, provided that it contains a high percentage of hydrogen. This allows for great flexibility according to location and enables the efficient conversion of many indigenous or waste resources. This is discussed in more detail in the following for two applications, stationary power generation and transportation.

3.1 Stationary Power Generation

The common fuel for those commercial stationary systems that exist (mainly PAFC based) is natural gas, which is reformed by a separate steam reformer before the hydrogen is fed into the fuel cell stack. Demonstration PEMFC systems also use this method. In contrast, high-temperature fuel cells are able to operate directly on some gas streams, though they may need to be cleaned up by the removal of sulfur and higher hydrocarbons. The advantage is that high-temperature cells are able to reform the fuel (methane, for example) directly on the anode of the fuel cell because they have sufficient thermal energy and catalyst present to do so. Both steam reforming and partial oxidation reforming can be used. The former is preferred on the grounds of efficiency, though partial oxidation may be needed in an external reformer or for startup:

Steam reforming:

$$C_nH_m + nH_2O \Leftrightarrow nCO + (n + m/2)H_2 \qquad (7)$$

Partial oxidation:

$$C_nH_m + (n/2)O_2 \Leftrightarrow nCO + (m/2)H_2. \qquad (8)$$

Good thermal integration by incorporating the reformer within a high-temperature fuel cell stack offers significant efficiency gains over an external reformer. However, the internal reforming process can cause severe temperature gradients across the stack, and many designers prefer to use a separate catalyst bed, distributed within the fuel cell stack. These processes are called direct and indirect internal reforming, respectively. As well as the reformer, low temperature fuel cells also require a high- and low temperature shift reactor to carry out the water gas shift reaction of CO to CO_2:

$$CO + H_2O \Leftrightarrow CO_2 + H_2, \qquad (9)$$

together with selective oxidation, to reduce the CO content to parts per million levels.

Fuel cells will also run on reformate from other fuel sources, and there are successful demonstration PAFC systems using gas from landfill sites or sewage farms.

3.2 Transportation

The fuelling problem in transport is more complex than in stationary systems, because there are far more severe size, weight, cost, and performance constraints. Vehicles could be fuelled using pure hydrogen, stored onboard the vehicle. Indeed, this is how most demonstration buses are supplied, making them true zero-emission vehicles. However, there is rarely provision for hydrogen fuelling, so if fuel cells are to be used in cars there may have to be another solution.

Like stationary systems, transport fuel cells (usually PEMFC) can operate on a hydrogen-rich reformate. This is less efficient than using pure hydrogen but may be a pragmatic approach to the fuelling issue. Methanol has been suggested as a good compromise fuel. Although it also requires a supply infrastructure to be developed, it is a liquid and can be handled in a similar way to gasoline, though concerns have been raised regarding its toxicity level. It is also comparatively easy to process into hydrogen. Gasoline can also be reformed using a sequence of autothermal, partial oxidation, and water gas shift reformers, though the resulting reformate stream contains less hydrogen than from methanol and reduces the system performance somewhat more.

The fuelling question remains significant. If methanol turns out to be the fuel of choice, then not only will methanol reformers with sufficiently high performance and low cost have to be integrated into all fuel cell cars, but the problem of supplying methanol to the consumer will also have to be addressed. However, the consensus if opinion in this sector seems to be moving toward hydrogen as the fuel of choice. For this to become a reality, hydrogen storage systems need to be improved, though it is possible to use compressed hydrogen in cylinders designed for compressed natural gas. Equally important are the standards and regulatory decisions that have to be made to allow hydrogen vehicles on the road.

In some cases, fuels other than hydrogen can be used within the fuel cell without reforming into hydrogen. Significant effort, for example, is being applied to the development of a direct methanol fuel cell that can use methanol as a fuel directly without any preprocessing. While this decreases system complexity, it makes the fuel cell more difficult to design and could result in lower efficiency. However, this route is particularly attractive for microfuel cells, which aim to replace battery technology.

4. FUEL CELL APPLICATIONS

There are many areas in which fuel cells could potentially be used to replace conventional power equipment, discussed under the broad headings of stationary power generation, transport, and battery replacement.

4.1 Stationary Power Generation

Even within stationary applications there are a number of distinct divisions, though the most important ones have to do with temperature and the amount of waste heat that can be used. In general, plant sizes of several hundred kWe's (enough for a leisure center or small office block) down to 1 to 5 kWe (a single house or a portable power unit) are under development.

High-temperature systems can be used in more demanding applications where larger systems are required or additional heat is useful. Such systems are also more efficient when operating on fuels such as natural gas, as they enable internal reforming. While heat of 80°C or more can be used for space heating, hot water and possibly absorption chillers to allow for cooling, industrial steam raising, or gas turbine bottoming cycles require temperatures of at least 500°C. This not only allows for the potential of generating extra electrical power and thus improving the overall system efficiency to nearly 70%, but also the possibility of using cogenerated heat and increasing total energy efficiency to 90%. Either of these options brings down the cost per unit of energy even if the capital cost of the system is high, but can only be effectively based around a high-temperature fuel cell variant.

4.2 Transport

Transport applications tend to demand rapid startup and instant dynamic response from fuel cell systems, so a high-temperature fuel cell is unlikely to be competitive as the main engine in applications such as cars and buses. The prime candidate for these vehicle propulsion systems is the PEMFC, which exhibits both of the above characteristics while also having very high power density. This is important as it must also occupy a similar amount of space to an internal combustion engine. Of recent interest has been the development of auxiliary power units for vehicles, in which the fuel cell meets the onboard electric load of the vehicle. Both PEMFCs and ITSOFCs are under development for this application.

AFC systems have traditionally been used in space applications by NASA—in the *Gemini*, *Apollo*, and space shuttle programs—but are also being investigated for certain transport applications, mainly in vehicles with limited duty cycles such as delivery vehicles and fork lifts.

Of course, transport is not confined to the car and aerospace markets—locomotives, ships, scooters, and a whole host of other applications offer potential for a variety of fuel cell systems. For ships and trains, where the application is almost akin to having a stationary power plant running all the time, startup and system dynamics are less important than noise, emissions, fuel consumption, and vibration. Serious investigations are under way as to the benefits of installing SOFCs, for example, on ships to remove the marine diesel—traditionally a source of heavy pollution, very high noise levels, and damaging vibration. However, in these cases the preferred fuel would be heavy fuel oil, one of the most difficult to process for a fuel cell application.

PEMFC systems have also been commissioned for a number of the world's navies—Ballard and Siemens have each been active in their respective countries putting PEMFCs into submarines.

4.3 Battery Replacement

An area in which it is possible that fuel cells may break through to commercialization in the very near future is in the replacement of conventional batteries. Battery power for laptop computers, mobile phones, and many other devices is expensive and often inconvenient if recharging is required every few hours. A small fuel cell with an equally small fuel source could potentially operate for longer than a battery, but with refueling only taking a few minutes instead of many hours. Batteries cost much more for the amount of power they can supply than almost any other application, and this gives the fuel cell a good chance of entering the market, although it is still expensive. The ideal fuel cell for these applications is the PEMFC, which not only has the high power density required for miniaturization, but also has the least challenging electrolyte management issues of the low-temperature fuel cells.

4.4 The Fuel Cell Industry

The recent upsurge of interest in fuel cell systems has led to rapid growth in industries developing fuel cell technology around the world. As well as fuel cell technology developers, there are an even larger number of component suppliers, which are not dedicated to the fuel cell industry but have a significant role to play, for example, in the supply of compressors, pumps, valves, and so forth. Indeed, it is estimated that typically only around one-third of the cost of a fuel cell system is the actual fuel cell stack, with around one-third being balance of plant, and one-third integration, installation, and commissioning costs. Cost issues are discussed in more detail in the following section.

5. FUEL CELL COSTS

The true costs associated with fuel cells are not yet clear—either from a capital or operating perspective. Present costs are well above conventional technologies in most areas, though this depends slightly on the type of fuel cell and the market area in which it may play a part.

The economics of fuel cell systems are also different in different market niches. The fuel cell has the potential to usurp many traditional technologies in a variety of markets, from very small batteries and sensors to multimegawatt power plants. Each system has very different characteristics and will accept very different prices. For example, a laptop battery substitute that could run for 20 h instead of 2 h could command a high price, especially if it could be refueled in seconds from a canister rather than recharged over several hours. At the other end of the scale, the potential for building modular power plants in which maintenance can be carried out on each module without shutting down the system is worth a significant amount of money to the owner.

Traditional economic calculations have suggested that the fuel cell system for large-scale power generation needs to be less then $1500/kWe before it will be competitive, while the fuel cell system for automobiles and mass production must be competitive with the internal combustion engine at $50/kWe or below. That said, it must be remembered other drivers exist for the technology, including environmental benefits, and an issue of increasing strategic importance for many counties, namely a reduced reliance on oil. Some fuel cell systems will sell themselves at $10,000/kWe, however, if they can be installed where there is currently no available technology capable of meeting requirements.

However, it is clear that all fuel cell costs at present—and these are estimated at anything between 500 and 10,000 dollars per kilowatt (a mature technology such as a gas turbine costs about $400/kWe)—are high because they are representative of an emerging technology. Once in mass production, recent estimates predict costs of $40 to $300/kWe for PEMFCs for transport applications, depending on assumptions regarding technology development. It is clear that both further technical innovation as well as mass manufacture will be needed to compete solely on a cost basis with the internal combustion engine.

High-temperature systems tend to be more expensive as they require significant investment in associated balance of plant, but should still be able to be manufactured for close to $600 per kilowatt, not far from the current price for a gas turbine or gas engine.

6. ENVIRONMENTAL BENEFITS

One of the areas in which fuel cells should demonstrate significant advantages is in their potential for minimal environmental impact. As well as offering a high theoretical efficiency, all fuel cells emit low levels of pollutants such as the oxides of sulfur (SO_x) and nitrogen (NO_x). SO_x emissions are

low because low sulfur fuels such as methanol or desulfurized natural gas are used. NO_x emissions are low because even the high-temperature fuel cells such as the molten carbonate fuel cell and the solid oxide fuel cell operate at temperatures well below those needed to form NO_x by the thermal combination of nitrogen and oxygen. The formation of NO_x at high temperatures is a problem for those trying to push up the efficiency of heat engines by increasing their maximum operating temperature. Fuel cell systems are already exempt from permitting requirements in some U.S. states, including New York, meaning that they can bypass one of the planning stages (in itself a useful bonus in reduced cost and time).

The high efficiency of the electrochemical process is an added advantage, since less fuel is required to produce a given amount of power than would otherwise be needed. This means that less fuel is used and, of increasing importance, that less CO_2 is released. Estimates suggest that a fuel cell power plant running on a traditional cycle could produce 20 to 30% less CO_2 than traditional power plants, and that vehicles powered by fuel cells would provide similar benefits.

In the case of vehicles, it is particularly important to examine the fuel on which they run. While fuel cell vehicles powered on gasoline reformer technology will have little or no benefit in reduced greenhouse gas emissions, methanol-powered cars may produce 25% less CO_2. Using hydrogen produced from natural gas will result in cuts of up to 40%, while hydrogen from renewably generated electricity, such as wind or solar, emits no greenhouse gas at any stage of the process.

In the ultimate renewable/zero-emission scenario, fuel cells can be run on pure hydrogen produced from renewable energy, using electrolysis of water. This ensures that there are no polluting emissions at any stage of the fuelling process (though there will inevitably be some from the manufacture and construction of the plant). This scenario is applicable both to stationary energy and to transport, and would enable countries to reduce their dependence on imported energy in various forms, in addition to being very clean. Hydrogen can also be produced from a whole variety of primary resources, including biomass and fossil fuels.

SEE ALSO THE FOLLOWING ARTICLES

Alternative Transportation Fuels: Contemporary Case Studies • Batteries, Overview • Batteries, Transportation Applications • Cogeneration • Fuel Cell Vehicles • Fuel Economy Initiatives: International Comparisons • Hydrogen, End Uses and Economics • Hydrogen, History of • Hydrogen Production • Transportation Fuel Alternatives for Highway Vehicles

Further Reading

Fuel Cell Today. www.fuelcelltoday.com. Accessed on November, 2003.

Kordesch, K., and Simader, G. (1996). "Fuel Cells and Their Applications." VCH, Weinheim, Germany.

Larminie, J., and Dicks, A. (2000). "Fuel Cell Systems Explained." Wiley, Chichester, United Kingdom.

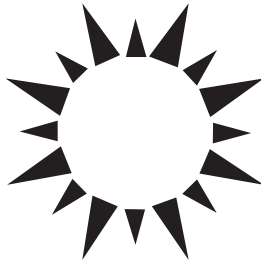
Minh, N. Q., and Takahashi, T. (1995). "Science and Technology of Ceramic Fuel Cells." Elsevier, Amsterdam.

Steele, B. C. H., and Heinzel, A. (2001). Materials for fuel cell technologies. *Nature* **414,** 345–352.

U.S. Department of Energy (2000). "Fuel Cell Handbook." 5th Ed. Produced under contract DE-AM26–99FT40575.

Fuel Cell Vehicles

JOHN M. DECICCO
Environmental Defense
Dexter, Michigan, United States



1. **Introduction**
2. **Factors Shaping Automobile Propulsion Technology**
3. **Fuel Cell Technologies**
4. **Vehicular Fuel Cell Systems**
5. **Fuel Supply, Storage, and Processing**
6. **Performance and Cost**
7. **Commercialization Issues**

Glossary

autothermal A characteristic of fuel processor design that combines exothermic and endothermic reactions so that the resulting system is close to thermally neutral.

criteria air pollutant An airborne substance regulated according to public health-based standards; key criteria pollutants pertinent to motor vehicles include carbon monoxide, various toxic substances, ozone and fine particulate matter, and their precursors including reactive hydrocarbons and nitrogen oxides.

electrolyte A substance that dissociates into ions, so that it can conduct electricity by means of ion transport (as opposed to electron conduction as in a metallic conductor, for example).

energy density Energy content of a fuel or energy storage device per unit mass.

fuel processor A device that converts a fuel into hydrogen-rich gas and removes impurities so that the resulting gas can supply the anode of a fuel cell; also termed a *reformer.*

platform The primary load-bearing structural assembly that determines the basic size of a motor vehicle and supports its powertrain, body panels, and suspension components.

powertrain The interconnected set of components that drives the wheels to propel a motor vehicle, commonly including engine, transmission, driveshaft and differential, or transaxles.

POX Partial oxidation, a combustion process in which the reactants are only partially oxidized, allowing hydrogen to be dissociated from a fuel without being completely consumed.

PrOx Preferential oxidation, a catalyzed chemical reaction in which species other than hydrogen are oxidized so as to yield a product stream containing maximum amounts of hydrogen with minimum amounts of undesirable constituents (such as CO).

reformate The gaseous output stream from a reformer or fuel processor, containing hydrogen for supplying a fuel cell anode.

reformer A device that converts a fuel into hydrogen-rich gas and removes impurities so that the resulting gas can supply the anode of a fuel cell; also termed a *fuel processor.*

regenerative braking The process of slowing a vehicle by drawing kinetic energy into a generator (typically an electric motor operating "in reverse"), thereby exerting a rotational drag on the wheels.

specific energy Energy content of a fuel or energy storage device per unit volume.

stack A set of individual fuel cells arranged next to each other and connected in series.

ultracapacitor An electrochemical device that stores electricity through capacitive charging and is capable of providing high power in a compact size and mass.

1. INTRODUCTION

A fuel cell is an electrochemical device that directly converts a fuel to electricity by means of reactions on the surfaces of electrodes and transport of ions through an electrolyte. A fuel cell can be thought of as a chemical battery whose reactants are fed from external sources rather than packaged as part of the battery. A key feature of a fuel cell is transformation of the chemical potential energy of a fuel directly into electricity, a high-value form of energy that can be put to many uses. The fuel cell's direct energy conversion process occurs without an intermediate step of heat generation, as involved in combustion engines. Fuel cells designed to operate at relatively low temperatures, as suitable for automotive propul-

 © 2004 Elsevier Inc. All rights reserved.

sion, need catalysts such as platinum that enable reaction rates sufficient to provide adequate power.

A fuel cell engine couples a fuel cell to an electric motor to provide mechanical power. A fuel cell vehicle is a motor vehicle having a fuel cell engine as its prime mover. Thus, a fuel cell vehicle has the advantages of electrical propulsion while retaining the advantages of carrying its own chemical fuel. Electrical propulsion is quiet and offers high efficiency, through avoidance of many mechanical frictional losses. Another advantage is its controllability, through fully programmable "tuning" of the vehicle's driving characteristics for responsiveness and smoothness. An electric drive system can further increase a vehicle's efficiency by recovering energy during braking, energy that otherwise is lost as low-grade heat in friction brakes. Advantages of an onboard chemical fuel include the relatively low mass occupied by the stored fuel, good driving range between refueling, and ease of refueling. Such advantages are related to a chemical fuel's energy density, or amount of energy per unit mass. Liquid fuels (such as hydrocarbons and alcohols) have higher energy densities than gaseous fuels (such as natural gas and hydrogen), and chemical fuels typically have higher energy densities than rechargeable batteries.

In addition to the fuel cell itself, a fuel cell vehicle requires a variety of additional components. Some are the auxiliary parts needed for a complete fuel cell power system. Others include fuel storage and processing systems, an electric motor, battery, associated electronics, and control systems. The major functional components are illustrated schematically in Fig. 1. The stack consists of a series of individual fuel cells, each consisting of electrodes, electrolyte, plates with channels for supplying the reactants and removing the exhaust products, and connections. Auxiliaries include blowers, pumps, heat exchangers, and other items needed to regulate the heat and mass flows from the stack. Fuel is supplied from the onboard tank, holding a liquid fuel or compressed hydrogen. Some designs entail a fuel processor, or reformer, to convert the fuel to a hydrogen-rich gas for feeding the fuel cell; that process may create pollutants requiring emissions controls for cleanup. The stack produces direct current that is regulated by a power electronics package, which in turn supplies electricity to one or more electric traction motors for driving the vehicle. The motor may recapture braking energy that is fed back through the power electronics to recharge the battery. At minimum, a fuel cell vehicle has a battery for startup; a more powerful battery or other high-power storage device is used to provide supplemental traction power in a hybrid fuel cell vehicle. Thus, a fuel cell vehicle is a complex machine, requiring numerous, high-technology components that must be integrated together and manufactured at the low cost and high reliability expected of mass-produced automobiles.

2. FACTORS SHAPING AUTOMOBILE PROPULSION TECHNOLOGY

Fuel cell vehicles are being developed because they promise to meet the requirements expected of automobiles in a market increasingly constrained by environmental and resource limitations. Air pollution and oil dependence have been persistent challenges for vehicles powered by petroleum fuels (gasoline and diesel). Global warming presents a new challenge in the need to limit carbon dioxide (CO_2) emissions from fossil fuel combustion. It is unclear whether automobiles powered by combustion engines running

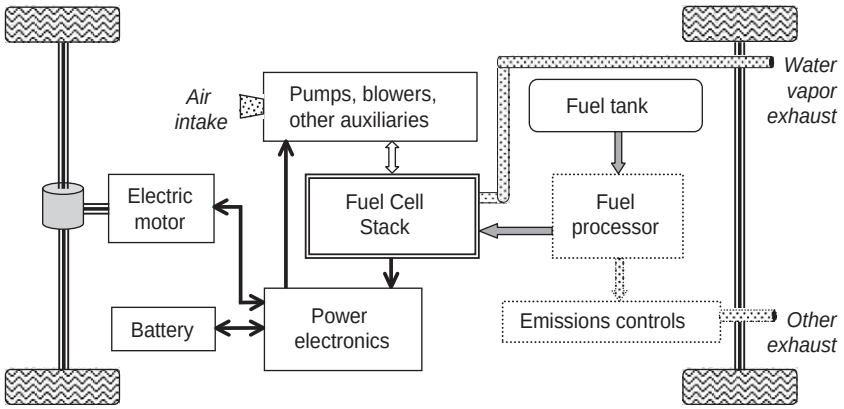


FIGURE 1 Major components of a fuel cell vehicle.

on conventional fossil-derived fuels will be able to meet these challenges. Moreover, electric propulsion has its attractive features of quiet, precisely controllable operation and high efficiency for both powertrain and accessories. The growing number of electrical amenities in automobiles further raises the value of a propulsion system that provides ample electric power capacity. In short, eliminating the internal combustion engine has great appeal and motivates the development of fuel cell vehicles.

Fuel cell vehicles are not the only option for addressing future automotive expectations. Improvements to internal-combustion engines can partially address the challenges, and further refinements can provide additional progress. Hybrid combustion-electric vehicles extend the efficiency horizon of petroleum-fueled designs while boosting onboard electric power capacity. Advanced strategies for controlling criteria air pollutants generated by gasoline vehicles (including sophisticated onboard combustion control and cleanup systems as well as reformulated fuels) make air quality concerns less a driver for technology change than the global warming issue, which presents a more fundamental challenge. Effective control of criteria air pollution is a greater hurdle for diesel engines, which otherwise have an energy-efficiency advantage. Conventional alternative fuels (such as alcohols, ethers, plant oils, and natural gas) also offer partial solutions, but their benefits have not been great enough to overcome the high costs of transition. Using low-CO_2 producing fuels in hybrid vehicles may enhance the prospects for some alternatives. Rechargeable battery electric vehicles have cost and range limitations that inhibit extensive use with existing transportation infrastructure. What makes fuel cell vehicles attractive is their promise for affordably addressing the multiple design requirements likely to be faced by future automobile propulsion systems.

3. FUEL CELL TECHNOLOGIES

Fuel cells are typically classified according to type of electrolyte. While many varieties of fuel cells have been demonstrated in the laboratory, five major types are seeing development for commercial applications:

- Polymer electrolyte membrane (PEM) cells use a plastic (polymer) membrane that becomes electrically conducting when hydrated (saturated with water); they operate near 100°C.
- Alkaline fuel cells use a caustic electrolyte such as potassium hydroxide (KOH); they also operate near 100°C.
- Phosphoric acid fuel cells use that acid as the electrolyte and operate around 200°C.
- Molten carbonate fuel cells use electrolytic mixtures of metal (Li, Na, K) carbonates and operate at 600°C or higher; they are highly efficient and can reform a fuel within the cell.
- Solid oxide fuel cells use zirconium oxide as a solid-state electrolyte and operate at 600°C or higher, and also have advantages of high efficiency and in-cell fuel reforming.

As fuel cell technologies mature, they may see a variety of uses, including battery replacement for portable electronics, large-scale electric power generation, distributed generation, combined heat and power, and transportation. Both PEM and alkaline fuel cells have been used on spacecraft. The first fuel cell vehicle was a farm tractor demonstrated in 1959 using an alkaline cell. Phosphoric acid fuel cells are in limited commercial use for specialized power generation but are too expensive and difficult to manage for vehicle use. Molten carbonate and solid oxide fuel cell technologies are being developed for stationary electric power generation. The need to bring the relatively large quantities of these electrolytes up to a high operating temperature makes these fuel cells unsuitable as a prime mover for motor vehicles, although small solid oxide fuel cells are being investigated as auxiliary power units.

The PEM fuel cell is the focus of automotive development efforts because of its low operating temperature and potential for low-cost mass production. Only PEM cells running on hydrogen have been developed enough to power practical motor vehicles. However, PEM cells that run on methanol are being investigated. It is also premature to dismiss alkaline fuel cells as a long-run option. Figure 2 illustrates the functional elements of a hydrogen PEM fuel cell. Fuel (hydrogen or a hydrogen-rich gaseous mixture) is supplied to the anode (negative electrode) where it diffuses to a catalyst layer. There it dissociates into electrons, supplied to the external electrical load, and hydrogen ions (protons), which are conducted across the polymer electrolyte membrane (also known as a proton exchange membrane) to the cathode. At the cathode, electrons from the returning current and oxygen from the air complete the reaction at a second catalyst layer, forming water vapor that is exhausted from the fuel cell.

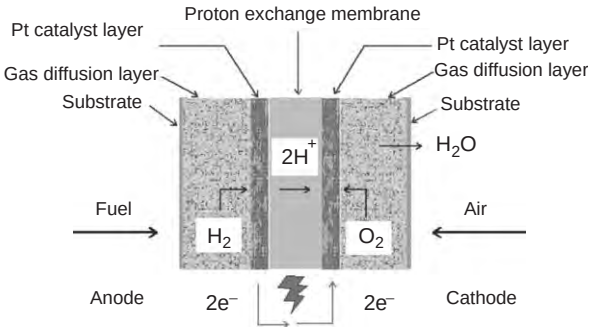


FIGURE 2 Schematic diagram of a polymer electrolyte membrane (PEM) fuel cell. Courtesy of Johnson Matthey Fuel Cells Today Web site (www.fuelcelltoday.com), with permission.

The ideal, open-circuit voltage of a hydrogen-air cell at typical operating conditions is approximately 1.2 volts. Under practical load conditions, a single hydrogen PEM cell might produce 0.7 volts. The voltage produced in part determines the cell's energy efficiency; at 0.7 volts, a hydrogen PEM cell is roughly 60% efficient in converting the chemical potential energy of hydrogen into electricity. Research and development on fuel cell design seeks to obtain a maximum amount of current from a given cell surface area while maintaining voltage, which determines the efficient power output of the fuel cell. A stack of cells connected in series is used to produce the power levels needed for vehicle propulsion.

4. VEHICULAR FUEL CELL SYSTEMS

The components of a fuel cell vehicle, as shown in Fig. 1, draw on a variety of technologies, not all of which are at the same state of development. The particular requirements for fuel cell vehicle systems depend on the configuration. Three categories of components are common to all fuel cell vehicle systems: the fuel cell stack itself, auxiliary equipment that supports the operation of the stack, and the electric drive system. Fuel storage and processing components depend on the choice of onboard fuel; these systems are discussed in Section V of this article.

4.1 Fuel Cell Stacks

It is progress in the development of hydrogen-air PEM stacks that has made fuel cells a contender for powering automobiles of the future. For many years, the energy and power densities of PEM cells were so low and the amount of platinum catalyst required was so high that most commercial applications seemed out of the question. For example, the platinum requirements for the PEM cells used on Gemini space missions of the 1960s were on the order of 100 g/kW, for a cost factor of $1500/kW (assuming a platinum cost of $15/g). A typical automotive fuel cell stack would be 80 kW, implying a cost of $120,000 for the catalyst material alone. By comparison, current automotive catalytic converters require roughly 0.05 g/kW of platinum-group metals, costing on the order of $100 for an average car. More stringent emissions standards are pushing precious metal requirements higher, so that future gasoline vehicles may need 0.1 to 0.2 g/kW of platinum group metals.

During the 1980s, research on electrode catalyst deposition led to large reductions in the amount of platinum needed. By the early 1990s, PEM fuel cell platinum requirements had dropped below 10 g/kW; as of 2002, state-of-the art designs needed less than 2 g/kW, with further reductions still in sight. Such breakthroughs have motivated engineering development work on the other critical elements of a fuel cell stack, such as the membrane, electrode assemblies, and flow fields that supply hydrogen and air and exhaust the reaction products from the cells. Platinum needs remain a cost challenge for PEM fuel cells, however, and ongoing research targets another order-of-magnitude reduction in platinum catalyst requirements as needed to place the technology in the range of affordability and precious metal availability that would make widespread use feasible.

Fuel cell stacks can be designed as compact modules, with connections for their electrical output as well as for feeding fuel and air into the stack and removing exhaust, cooling the stack, and monitoring its operating state. Since PEM cell stacks are built in modular form, a great deal of flexibility will be available to designers in terms of how stacks are placed and arranged in the vehicle. Using current designs, a typical fuel cell stack for automotive applications, delivering 80 kW of peak power, occupies a volume of about 80 liters and weighs roughly 90 kg. However, even more compact designs are likely to evolve as experience is gained. Stacks could be built in modules of specified sizes, so that multiple modules would be arranged and connected to supply the differing power requirements of different sized vehicles while permitting enhanced economies of scale in mass production. A remaining engineering challenge for fuel cells is achieving sufficient membrane durability. Other developmental issues include

freeze protection and further cost reduction for electrodes, membranes, and separator plates.

The latter serve the multiple functions of separating adjoining cells, providing the flow fields that transport fuel-air reactants and exhaust products to and from the cells, and removing waste heat. Early PEM cells have used graphite plates, but these are costly to produce. Coated metallic and composite materials are being developed for separator plates, with goals of meeting low-cost mass production targets while achieving performance and durability requirements in the highly reactive fuel cell environment.

4.2 Fuel Cell Engine Auxiliaries

Numerous other parts are needed for a complete fuel cell engine suitable for motor vehicles. These auxiliary components include many familiar pieces of equipment, such as blowers, pumps, heat exchangers, sensors, and controls. However, fuel cells have specialized performance requirements, so substantial engineering effort is needed to develop auxiliary systems suitable for fuel cell vehicles.

A blower or compressor is needed to supply air to a fuel cell stack, and air compression represents the largest parasitic load in the system. The pressure of the air supplied to the cells influences the stack efficiency, but a trade-off is involved. The higher output that comes with higher pressure entails an energy penalty for running the compressor. Designers are experimenting with various pressure regimes, and although pressure may vary depending on operating conditions, most designs operate at no more than three atmospheres. Compression power requirements approach 10% of stack output at peak power, while other loads (pumps, cooling, controls, etc.) are generally in the 1 to 2% range. Optimized designs can hold the average air compression penalty to 3 to 5% of stack output. Energy in the fuel cell exhaust stream is typically captured in a turbine expander to help power the compressor.

Another critical auxiliary component is an integrated system for moisture control and thermal management. This system maintains membrane moisture, exhausts water vapor, removes heat from the stack, and manages the flow of heat among various components for optimum efficiency. Fuel cell membranes need to be precisely hydrated to function properly. Too much water accumulation blocks the reactions, but the membrane can fail or degrade if it becomes too dry. Fuel cells need to start in very cold weather; motor vehicles are typically designed to be able to start at −40°C. Therefore, special "anti-freeze" hydrating solutions are being developed that do not poison the stack; mechanisms are needed to address the water vapor that is present in the exhaust stream and other fuel cell subsystems as well as quickly bring a stack up to operating temperature.

Heat removal presents its own special challenges, since PEM cells operate at low temperatures and so need to efficiently reject low-trade heat. Ambient pressure designs operate near 80°C. This level is much lower than the over 2000°C flame temperature and roughly 500°C exhaust gas temperature of a gasoline combustion engine. Pressurized PEM fuel cells use only relatively low pressures, so temperatures are under 200°C in any case, offering but a small difference to ambient temperature for driving heat transfer. Even at its high, roughly 50% efficiency, a 80 kW stack must dissipate 5 to 10 kW of waste heat under average loads and as much as 40 kW without risking membrane degradation under peak loads. Some of this heat is carried away in the exhaust stream, but much must still be handled by directly cooling the stack and transferring heat to the ambient air. Thus, sizable heat exchangers and fans are needed.

4.3 Electric Drive Systems

The primary electric drivetrain components for fuel cell vehicles are the same as those for any electric vehicle: traction motors, power electronics, and batteries. Electric drive components require their own sets of auxiliaries and management systems, for control and cooling of the equipment. A fuel cell vehicle may have a hybrid powertrain, in which the fuel cell is sized at less than the vehicle's peak power requirement and additional power is supplied from an electricity storage device, such as a high-power battery or ultracapacitor. Although a hybrid powertrain is more complex than one using a fuel cell alone, it offers advantages of regenerative braking, which helps vehicle efficiency, and the availability of greater power for startup and auxiliary loads. Hybridization also would have an economic advantage if the cost of peak power from the electricity storage device is less than that of meeting the peak power requirement with a fuel cell alone.

The electric drive system for a fuel cell vehicle is comparable to that for a battery electric vehicle or series hybrid, with the traction motor and power electronics scaled to meet the full power needs of the vehicle. This situation contrasts with electric drive system for a parallel hybrid as used for contemporary models (such as the Toyota Prius and Honda Civic Hybrid), in which the electric drive components

supply only a portion of the tractive power with the rest mechanically provided from the combustion engine. The specific components used in an fuel cell vehicle, such as type of motor, depend on the particular electric drive system design, which is itself a new and evolving area of engineering.

Although more developed than fuel cells themselves, electric drive technologies are not mature by automotive standards. As of 2002, hybrid electric vehicle production volumes were still an order of magnitude lower than those of typical mainstream vehicle models, and costs remain high. For fuel cell vehicles, electric drive system costs compound those of the core fuel cell system and its fuel storage or processing components. The cost of electric drive components is expected to fall substantially, with some estimates as low as $30/kW (retail-equivalent, excluding batteries). Time and production experience will be needed for costs to fall to such levels, which are likely to be essential for widespread commercialization of any type of electric vehicle, including fuel cell vehicles. As experience is gained and research progresses, the types of electric drive components used in fuel cell vehicles will also evolve. For example, instead of traditional electric traction motors, compact wheel motors might be used as suggested by GM's Autonomy fuel cell vehicle concept.

5. FUEL SUPPLY, STORAGE, AND PROCESSING

Depending on the fuel they carry, fuel cell vehicles entail various requirements for fuel supply (including production and distribution), fuel storage on the vehicle, and on-vehicle fuel processing that might be needed.

A fuel cell vehicle operated on gasoline may require minimal changes in existing fuel supply infrastructure, but faces special processing requirements onboard the vehicle. A fuel cell vehicle operated directly on hydrogen, on the other hand, does not require a fuel processor, but needs a new fuel supply infrastructure and workable technology for onboard fuel storage. Methanol is also being considered for fuel cell vehicles, since it has the convenience of liquid while being easier to process into hydrogen than gasoline. Table I summarizes the issues and technical status of these three key fuel options for fuel cell vehicles. Ethanol has processing requirements similar to those for gasoline and a long-term potential for production from renewable resources. Natural gas has processing requirements similar to those for methanol and is already widely distributed, but faces the challenges associated with gaseous storage.

The ultimate fuel supply, storage, and processing configuration of fuel cell vehicles will be shaped by how well various fuel-related technologies progress through research and development. The configuration will also be influenced by the broader changes in transportation infrastructure (such as increasing use of intelligent systems) that will occur over the time frame needed to bring these technologies to states of workability and affordability that would enable widespread commercialization.

5.1 Fuel Supply

The auto and oil industries are among the world's largest, and the extensive infrastructure that supplies gasoline and diesel motor fuel is built of years of investment in both equipment and know-how. If practical ways can be found to power fuel cells with

TABLE I

Fuel Options for Fuel Cell Vehicles

	Hydrogen	Methanol	Gasoline
Fuel economy gain	High	Modest	Small/nil
CO_2 reduction (near-term)[a]	Modest	Low	Small/nil
Emissions control need	Zero	Low	Low
Fuel supply	Needs major development	Very limited	Already exists
On-board storage	Needs major development	Readily feasible	Already exists
Fuel processing	Not needed	Under development	Under development
Added car cost/weight	Modest	High	High

[a] Near-term CO_2 reduction assumes hydrogen and methanol derived from natural gas; the long-term potential is greater for all options as renewable resources or carbon sequestration systems for fossil fuels become available.

existing hydrocarbon fuels, or variants of them, adoption of fuel cell vehicles could proceed more quickly than if it were constrained by the rate at which infrastructure can be changed. On the other hand, fuel cells powered directly by hydrogen are already technically workable. Thus, another commercialization path could be niche vehicles that can accept the range restrictions of compressed hydrogen storage using new infrastructure that is constructed incrementally.

Gasoline-like hydrocarbon fuels for powering fuel cell vehicles would have properties tailored to the new technology, just as today's motor fuels are formulated for optimum performance in combustion engines. Impurities such as sulfur would have to be essentially eliminated; this change is already under way as sulfur levels in conventional fuels are being lowered for pollution control reasons. The mix of particular hydrocarbons would also change. But the key characteristics of being an energy-dense, convenient-to-distribute liquid would remain, and that is the attraction. Hydrocarbon fuel supply is also quite energy efficient. The "well-to-tank" efficiency—that is, the amount of fuel energy delivered to a vehicle per unit of primary energy—is typically over 80% for gasoline and over 90% for diesel.

The major problems for use of hydrocarbons in fuel cell vehicles are (1) the fact that hydrocarbons still carry carbon to the vehicle and (2) the difficulties of fuel processing (discussed later). All known hydrocarbon fuel technologies for motor vehicles involve releasing the carbon as CO_2. Practical ways to "scrub" or capture carbon from vehicles do not exist. Thus, even with the fuel cell vehicle's potentially higher efficiency and the possibility of offsets through carbon sequestration, there will be limits to the effectiveness of fossil-derived fuels for addressing global warming. Thus, hydrocarbons supplied by an evolved version of the current motor fuel supply infrastructure are viewed as a possible transitional pathway rather than the ultimate fuel. Hydrocarbons derived from renewable resources could offer a zero or very low net release of CO_2 to the atmosphere (and of course, such fuels could also be used in combustion engine vehicles).

Hydrogen is the ideal fuel for PEM cells and any other low-temperature fuel cell technology. Currently and for the near-term, natural gas is the most economical source of hydrogen. It can be produced by steam reforming at central plants and then distributed to filling stations, or natural gas could be locally reformed at filling stations to provide the hydrogen. The former approach is how most hydrogen is now made. The primary energy efficiency for supplying compressed hydrogen in this manner is typically 70% (that is, the energy content of the compressed hydrogen at its point of distribution is about 70% of the energy content of the natural gas at the well). Local reforming would take advantage of existing natural gas distribution infrastructure and is generally viewed as the most economical way to begin building a hydrogen fuel distribution system. Hydrogen can also be made by reacting coal with water; this approach holds economic promise because coal is so plentiful. Research is under way to find ways to sequester the resulting CO_2 and such a system could enable the use of coal to generate hydrogen for a variety of applications, including vehicles, with lower CO_2 emissions than existing direct use of fossil fuels.

The ultimate attraction of hydrogen is that it is a carbon-free energy carrier which, if produced from renewable resources, offers the potential of an energy system with zero greenhouse gas emissions. If vehicular hydrogen supply systems were initially developed using natural gas, for example, they ultimately might be converted to hydrogen supplied from renewable resources.

Other fuel options have their own supply characteristics, which differ in the near and long term. Methanol is now made from natural gas and that would be the most economical source for the foreseeable future. In the long run it could be produced from coal or biomass, with the latter resource option offering very low greenhouse gas emissions. But methanol also has toxicity issues that are difficult to manage. Ethanol is now made from sugar or starch-bearing crops in an energy- and resource-intensive agricultural system, making it costly and of dubious net environmental benefit even though it is in some sense renewable. However, ethanol can be made through efficient, enzymatic conversion of cellulose from organic wastes and specialized crops, thereby offering near-zero greenhouse gas emissions. Alcohols still carry carbon to the vehicle, so achieving their potential for zero or very low greenhouse gas emissions would require both truly renewable, zero-fossil-fuel input energy crops plus essentially zero-emissions processing, refining, and distribution systems.

Finally, from a long-term fuel supply perspective, it cannot be discounted that the most effective way to use fuel cells to power vehicles might be indirectly, at least for many vehicles in urban and near-urban areas. In this case, the fuel cells are kept off the vehicle, which would be powered by advanced

batteries charged through an intelligent, electrified infrastructure to which electricity is supplied by stationary fuel cells. This approach would represent an even closer coupling of motor vehicles to a power grid than other options for grid-connected vehicles, including vehicles with onboard fuel cells, that could supply power to the grid but would still need a separate fuel infrastructure.

5.2 Fuel Storage

Fuel storage onboard an automobile is straightforward and well established for liquids, but presents cost, range, and packaging challenges for any gaseous fuel. The challenge is particularly acute for hydrogen, which is the least dense of substances. Thus, a fundamental breakthrough in onboard hydrogen storage technology is needed for widespread use of direct hydrogen fuel cell vehicles to become feasible. Even if fuel cells and their auxiliary systems achieve low costs, and hydrogen supply and infrastructure systems become competitive, all known options for storing hydrogen have limitations that would inhibit commercialization of fuel cell vehicles for personal automobiles as used today.

Hydrogen is the least dense of all fuels in spite of its high specific energy of 142 MJ/kg, compared to 46 MJ/kg for gasoline. All vehicular storage options for hydrogen present difficulties, rooted in the specific energy and energy density values shown in Table II. Options include compressed gas, liquid hydrogen and metal or chemical hydrides (which involve chemical reactions to store and release the hydrogen). Devices that react iron oxide with water to produce hydrogen have been explored. Adsorption in carbon nanotubes and high-pressure confinement in glass microspheres have been proposed but remain largely speculative.

Compressed gas storage is well known through experience with natural gas vehicles; most fuel cell demonstration vehicles use compressed hydrogen. Problems with gaseous storage include onboard space requirements and costs inherently higher than those of liquid fuel tanks. Packaging challenges exist even for natural gas, which has an energy density nearly four times higher than that of hydrogen. Development is under way on very-high pressure tanks made of carbon-fiber composites that offer greater shape flexibility than metal cylinders. Automakers have successfully tested tanks rated at 5000 psi, which is higher than the 3000 psi range now typical for compressed gas storage, and designs in the 10,000 psi range are being developed.

TABLE II

Onboard Fuel Storage System Characteristics

Fuel storage option	Specific energy (MJ/kg)	Energy density (MJ/L)
Gasoline (liquid in steel tank)	36	32
Methanol (liquid in steel tank)	18	18
Hydrogen		
Compressed gas (5000 psi/340 atm)	7	2
Cryogenic liquid	15	5
Metal hydride	3	12
Organic hydride	9	7
Sponge iron	5	6
Nanotubes	50	100

Note: Comparisons are made on a higher heating value (HHV) basis: 46 MJ/kg for gasoline, 23 MJ/kg for methanol, and 142 MJ/kg for hydrogen, counting the mass and size of fuel system components but not counting powertrain energy conversion losses. Values for hydride and sponge iron options are fairly uncertain but representative of what is expected through research; the nanotube option is still considered speculative.

Storing hydrogen as a liquid requires very cold, cryogenic temperatures. At atmospheric pressure, hydrogen liquefies at 20 K (−253°C or −423°F); the saturation temperature rises by only 10 K going to 10 atmospheres. Special systems are needed to maintain such low temperatures on a vehicle and safety manage the hydrogen boil-off that occurs while the vehicle is not in use. Liquid hydrogen storage systems have been demonstrated in compact packages suitable for vehicles, along with automated refueling systems that approach gasoline in terms of speed and convenience. A serious handicap is high energy cost of liquefying hydrogen, amounting to roughly one-third of the fuel's energy content. Because it involves unrecoverable low-grade heat rejection, this energy loss is fundamental and degrades the environmental benefits and economics no matter what resource is used for hydrogen production. Further losses occur during distribution, refueling, and storage (due to boil-off), so that the net supply efficiency of liquid hydrogen may be no more than 50%.

Hydrides of various forms are being investigated for storing hydrogen; they are relatively compact but have a high weight penalty. Metal hydrides store roughly 7% hydrogen by weight, not counting

containment, but not all of that is recoverable in practical storage systems (by comparison, gasoline is about 16% hydrogen by weight). Energy penalties are involved in charging (filling) and discharging hydride systems. Some metallic systems (such as magnesium alloys) are costly but compact; other materials (such as sodium borohydride) are less costly but face other technical challenges. Another system under investigation is lithium hydride in an organic mineral oil slurry, and other organo-metallic hydrides may hold promise, although the charging and discharging reactions involve elevated temperatures and catalysts to effectively release the hydrogen. Some systems would have to be refilled by simultaneous pumping of fresh, "charged" hydride into the vehicle while pumping out the hydrogen-depleted chemical for reprocessing. In general, hydride systems are still in early stages of development and have not yet been shown to be workable for automobiles.

5.3 Fuel Processing

The challenges of supplying hydrogen and storing it onboard a vehicle and have prompted efforts to develop onboard fuel processors (reformers) for converting liquid hydrocarbon or alcohol fuels into a hydrogen-rich anode feed gas. In general, such fuel processing involves two major steps: (1) dissociating the hydrogen from the fuel and (2) purifying the resulting gaseous stream to have maximum hydrogen levels and minimal constituents such as CO that can poison the stack. Each step involves multiple reactions and processes and a preliminary cleanup stage is needed if sulfur or other detrimental substances are present in the fuel.

Methanol can be processed by steam reforming, a catalytic process that is endothermic (requiring heat input). Steam reforming is a standard process for large-scale production of hydrogen from natural gas. Low-cost catalysts, such as copper/zinc oxides on alumina for methanol and nickel on alumina for natural gas, can be used for steam reforming. Two reactions are involved: (1) the fuel and water are converted to a synthesis gas containing CO and H_2, and (2) a water-gas shift reaction converts H_2O plus CO into CO_2 plus additional H_2. The resulting reformate is rich in H_2 and low in CO. Since air input is not used in steam reforming, the reformate is not diluted with nitrogen. But because it is not possible to run a complete water-gas shift reaction, further CO cleanup is needed using a catalyzed preferential oxidation ("PrOx") processor (that is, a device that preferentially oxidizes CO rather than H_2). A precious metal catalyst is required for the PrOx processors shown to be effective to date.

Because methanol is a liquid that can be steam-reformed at a relatively low temperature of about 250°C, some researchers consider methanol-based systems more feasible than systems based on hydrocarbons that require reforming temperatures in excess of 650°C. However, steam reforming involves a trade-off between fast, responsive reaction rates needed for driving and the slower reaction rates that are limited by the heat transfer characteristics of the reformers. Slow startup and high cost also remain difficult issues for methanol processing.

Any hydrocarbon can be converted into a H_2 reformate by using a partial oxidation ("POX") process followed by a water-gas shift reaction and then PrOx cleanup of the residual CO. This approach is being developed for gasoline as well as flexible-fuel reformers that could handle alcohols as well as other fuels. Partial oxidation is exothermic, since it involves burning some of the fuel, and the POX process does not itself require a catalyst. However, most hydrocarbon fuel processors use a catalyst to lower the reaction temperature, improve hydrogen yields, and avoid soot formation. Many are designed to be autothermal, using the heat released from the POX as input for the steam reforming process, so that the resulting system is close to thermally neutral. This approach helps the fuel processor achieve a high efficiency.

Startup and transient response are less problematic for POX and autothermal reformers than for steam reformers. However, they face a greater CO cleanup challenge, since their raw reformate can have a CO concentration as high as 10% (from gasoline, for example). Also, nitrogen in the input air dilutes the H_2 in the reformate, thereby lowering the efficiency of the fuel cell, and can produce ammonia, adding to the cleanup requirements. Providing a very clean H_2 reformate is critical, because impurities in the reformate not only poison the fuel cell catalyst, but also cause the fuel cell membrane to degrade. To achieve high effectiveness and durability, most fuel processing catalysts contain noble metals, leading to high catalyst costs.

All approaches to onboard liquid fuel processing have problems of cost, size and weight, startup and dynamic response, and overall workability that need to be resolved before practical designs can be developed. Although the basic processes involved are well known, combining them into an affordable device that would fit into a car is a formidable

challenge. Although size and weight are a lesser concern in nonautomotive applications, many similar technical issues must be resolved even for fuel cells in stationary power generation.

6. PERFORMANCE AND COST

The ultimate appeal of fuel cell vehicles rests on whether they can provide superior environmental performance (low emissions of air pollutants and greenhouse gases) while meeting or exceeding the driving performance and other consumer expectations of competing technologies. At the same time, fuel cell vehicles will need to be affordable in terms of providing such features with customer value sufficient to be marketed successfully.

A fuel cell vehicle's environmental performance advantage is determined largely by the operating efficiency of the stack and associated components. Part of the fuel cell's attraction is its inherently high energy conversion efficiency, exceeding 50% over much of the load range, compared to the 35% or 45% efficiency achieved by gasoline or diesel engines, respectively, only in their ideal operating range. However, net fuel cell system efficiency is strongly influenced by auxiliary systems, particularly air compression. For liquid fueled systems, fuel processing penalties also come into play. Overall vehicle efficiency also depends the balance-of-vehicle parameters that determine mass, aerodynamic drag, tire rolling resistance, and accessory loads. The level of greenhouse gas emissions also depends on the source of the fuel, as it does for all vehicle propulsion options.

The performance and cost of fuel cell vehicles and key competing advanced vehicles is illustrated in relative terms in Fig. 3. While numerous other alternative fuels (such as natural gas, alcohols, synthetic fuels) could also enter into such a comparison, much of what is known can be inferred by comparing direct hydrogen and gasoline-based fuel cell vehicles to evolved versions of gasoline and diesel internal combustion engine vehicles, including hybrid-electric vehicles that retain internal combustion engines. The comparisons of Fig. 3 are normalized, showing both retail price increases and greenhouse gas emissions reductions in percentage terms. The estimates are referenced to a contemporary gasoline car, represented by "Today's Vehicles" at origin of the axes, rated at 28 miles per gallon (mpg) on an average of U.S. EPA regulatory city and highway driving tests. Cost and performance depend on numerous assumptions regarding

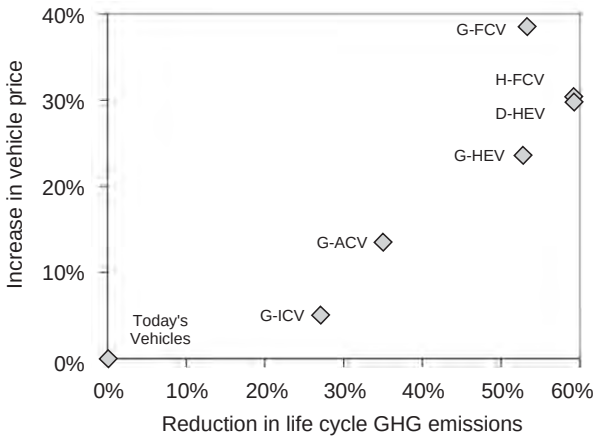


FIGURE 3 Greenhouse gas emissions reductions and retail prices increases of fuel cell vehicles relative to other advanced vehicles. All fuels assume oil or natural gas feedstocks and projections are given relative to today's vehicles as represented as a midsize car rated at 28 miles per gallon on the U.S. EPA composite test cycle. Key: G, gasoline; D, diesel; H, hydrogen, ICV/ACV, improved/advanced conventional vehicle; HEV, hybrid electric vehicle (on advanced platform); FCV, fuel cell vehicle (on advanced platform).

future technologies; the values could change depending on R&D progress. Nevertheless, the ordinal ranking reflects current knowledge of the engineering potential of fuel cell cars compared to other future technologies.

Even using fossil fuel feedstocks (oil for the gasoline and diesel vehicles, natural gas for methanol and hydrogen), fuel cell vehicles and other options all show lower greenhouse gas emissions than today's cars. A good part of this reduction is associated with an advanced vehicle platform having reduced mass, better aerodynamics, and other reductions in the "load," or amount of energy that needs to be delivered by the vehicle's power plant. The comparisons for the fuel cell vehicles, hybrid-electric vehicles, and advanced conventional vehicles assume a uniform advanced platform. This platform has substantial load reduction compared to today's cars, reflected in how the gasoline-powered advanced conventional vehicle (G-ACV) in Fig. 3 achieves a greater GHG emissions reduction than the improved conventional vehicle shown for reference (G-ICV).

Among fuel cell vehicles, the direct hydrogen design (H-FCV) provides the greatest greenhouse gas reduction and has the lowest expected cost, since it avoids the cost and other penalties associated with onboard fuel processing. The H-FCV illustrated in Fig. 3 assumes compressed gas hydrogen storage

which, while technically feasible, may not be practical for mass-market vehicles. The gasoline fuel cell vehicle (G-FCV) represents the highest cost option. Although it may not reduce greenhouse gas emissions more than a hybrid-electric combustion engine vehicle, the G-FCV is being researched because it offers a transition path to hydrogen fuel cell vehicles that make use of existing infrastructure.

The comparisons shown here assume hydrogen produced from natural gas. Substantially lower or even zero greenhouse gas emissions are hypothetically possible with hydrogen produced from solar energy, other renewable resources, or nuclear power. Very low greenhouse gas emissions could be obtained with fossil-derived hydrogen if ways were found to sequester the carbon, since direct use of hydrogen avoids carrying carbon atoms to the vehicle. Similarly, very low greenhouse gas emissions would be feasible with methanol or other fuels that might be processed into onboard hydrogen if they were derived from renewable resources or other feedstocks enabling offset of the carbon.

Because of diesel fuel's high efficiency in piston engines and low energy requirements for production from crude oil, diesel vehicles have lower greenhouse gas emissions than similar gasoline vehicles. A diesel hybrid-electric vehicle (D-HEV) offers lifecycle greenhouse gas emissions reductions as great as those of a H-FCV using hydrogen derived from natural gas. However, it is not yet known whether diesel engines can attain adequately low levels of criteria air pollutants, as already demonstrated for gasoline internal combustion engines and likely to be feasible for any fuel cell vehicle. Costs of an advanced D-HEV are expected to be similar to those of a H-FCV, due to the significant costs premium of diesel engines and their emissions controls compared to gasoline engines.

To put costs into perspective, the average real (inflation-adjusted) price of new cars in the United States has been rising linearly for many years, as a result of consumer features as well as regulatory changes to address safety and environmental hazards. Between 1980 and 2000, average new car real prices rose 34%. Continuing a linear trend over the next 20 years—a plausible early horizon for mainstream commercialization of fuel cell vehicles—implies a further 21% price increase. Current expectations, as reflected in Fig. 3, suggest fuel cell vehicle costs exceeding such a trend. Nevertheless, a combination of research progress (lowering costs) and greater value placed on the environmental or other attributes offered by the technology could yield a situation where fuel cell vehicles are competitive in future market and regulatory contexts. Indeed, the prospects for such a business case underpin the research and development investments now being made by automakers in fuel cell technologies.

7. COMMERCIALIZATION ISSUES

The commercialization prospects for fuel cell vehicles depend not only on their performance and cost, but also on how well they can compete with other technology options that address similar market and policy needs. While market forces have not traditionally motivated design change for reasons of environmental performance, customer values and expectations can evolve and such characteristics could grow in importance. However, inherent market conservatism will favor less disruptive ways to address evolving needs, which might be met by improved gasoline and diesel vehicles, including hybrid-electric versions. Yet looking over the long-run, particularly the need to substantially reducing greenhouse gas emissions, hydrogen fuel cells may well provide a solution that is superior to other alternatives.

While the use in mass-market private vehicles is the prospect that motivates extensive fuel cell R&D, cars and light trucks may not be the initial market for fuel cells vehicles. Fuel cells seem likely to become competitive sooner in urban buses and delivery vehicles. Centrally managed fleets of such vehicles could be serviced by a limited hydrogen refueling infrastructure; onboard hydrogen storage (e.g., with compressed gas) is also more feasible with such vehicles. Engines for buses and other heavy vehicles are produced at lower volumes with higher unit costs than those for light-duty vehicles, enabling fuel cells to compete sooner in this more limited market. Urban operation is well suited for electric drive and causes higher population exposures to exhaust pollution, providing a higher value on the very low or zero emissions offered by fuel cells. The higher cost tolerance, lower infrastructure barriers, and higher value of a very clean technology suggest that buses and other urban vehicles offer greater opportunities for initial fuel cell vehicle commercialization than light duty vehicles. Nevertheless, improvements in diesel emissions control and the emergence of diesel electric hybrid designs will still provide notable competition for fuel cell vehicles.

For light vehicles, the most difficult barriers faced by fuel cell vehicles are those associated with the fuel,

namely, the dual challenges of onboard hydrogen storage and liquid fuel reforming. Both of these technical hurdles present fundamental R&D challenges. Until such challenges are overcome, it will not be possible to confidently project when commercialization of fuel cell cars might proceed. Moreover, electric drivetrain components (motors, controllers, high-power batteries) are just now beginning to see mass production for hybrid vehicles. It is likely to take over a decade to gain experience and reduce costs for such components sufficiently to enable fuel cell vehicles to be affordable to the mass market. Thus, although automakers began introducing fuel cell demonstration vehicles in 2002, widespread commercialization will not be possible for at least another 15 or more years. Even over a 15- to 20-year time frame, building a business case for fuel cell vehicles will require the convergence of three pathways: (1) engineering development and maturation of fuel cell stacks and auxiliary components, (2) cost reduction for electric drive components, and (3) resolution of the barriers related to the right fuel for fuel cell vehicles.

Given such the long time horizon, non-energy-related needs could begin to reshape the personal transportation system. These other needs include issues such as congestion, safety, personal security, and convenience, and the opportunities associated with telecommunications and computer technology. It is not obvious that any "drop in" new fuel and technology, including fuel cells, that does little to address such other needs will provide benefits compelling enough to displace evolved versions of the internal combustion engine running on hydrocarbon fuels. Nevertheless, given the fuel cell's promise for meeting the long-term challenges for motor vehicle energy use, research on the technology has captivated the interests of industry and governments around the world, so that fuel cells may well become the prime mover of choice in a future transportation system whose details are yet to be imagined.

SEE ALSO THE FOLLOWING ARTICLES

Alternative Transportation Fuels: Contemporary Case Studies • Batteries, Transportation Applications • Biodiesel Fuels • Ethanol Fuel • Fuel Cells • Fuel Cycle Analysis of Conventional and Alternative Fuel Vehicles • Fuel Economy Initiatives: International Comparisons • Hybrid Electric Vehicles • Hydrogen, End Uses and Economics • Transportation Fuel Alternatives for Highway Vehicles

Further Reading

Angrist, S. W. (1976). "Direct Energy Conversion." 3rd Ed. Allyn & Bacon, Boston.

Appleby, A. J. (1999, July). The electrochemical engine for vehicles. *Sci. Am.* **28,** 74–79.

DeCicco, John, M. (2001). "Fuel Cell Vehicles: Technology, Market and Policy Issues." Society of Automotive Engineers, Warrendale, PA.

Borroni-Bird, C. E. (1996). Fuel cell commercialization issues for light-duty vehicle applications. *J. Power Sources* **61,** 33–48.

Koppel, T. (1999). "Powering the Future." Wiley, Toronto.

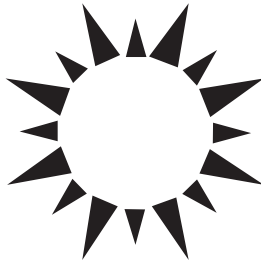
Oei, D-G. (1995). Direct hydrogen-fueled proton exchange membrane (PEM) fuel cell for transportation. SAE Paper No. 952763. Society of Automotive Engineers, Warrendale, PA.

Stobart, R. K. (1999). Fuel cell power for passenger cars—what barriers remain? SAE Paper No. 1999–01–0321. Society of Automotive Engineers, Warrendale, PA.

Thomas, C. E., James, B. D., Lomax, F. D., and Kuhn, I. F. (2000). Fuel options for the fuel cell vehicle: Hydrogen, methanol, or gasoline? *Int. J. Hydrogen Energy* **25,** 551–567.

Weiss, M. A., Haywood, J. B., Drake, E. M., Schafer, A., and AuYeung, F. F. (2000, October). On the road in 2020: A life-cycle analysis of new automobile technologies. Energy Laboratory Report No. MIT EL00–003. Massachusetts Institute of Technology, Cambridge, MA.

Fuel Cycle Analysis of Conventional and Alternative Fuel Vehicles



MICHAEL WANG
Argonne National Laboratory
Argonne, Illinois, United States

1. **Introduction**
2. **Alternative Transportation Fuels and Alternative Fuel Vehicles**
3. **Fuel Cycle Analyses of Vehicles and Transportation Fuels**
4. **Fuel Pathways and Vehicle/Fuel Combinations**
5. **Energy and Emissions: Items of Interest**
6. **Issues Involved in Transportation Fuel Cycle Analyses**
7. **Completed Studies**
8. **Fuel Cycle Energy and Emission Results of Vehicle/Fuel Systems**
9. **Conclusions**

Glossary

advanced vehicle technologies The engineering and design processes that lead to vehicles with high energy efficiencies and low emissions; include direct-injection, hybrid electric, fuel cell, and battery-powered electric vehicle systems.

alternative transportation fuels Energy sources other than petroleum gasoline and diesel for motor vehicle applications. These fuels can be produced from nonpetroleum sources and may have energy and emission benefits.

feedstock stage The first stage of a transportation fuel cycle; the feedstock stage includes fuel feedstock recovery, storage, and transportation.

fuel cycle analysis A procedure used to analyze energy use and emissions during the cycle of energy feedstock production and transportation; transportation fuel production, transportation, and distribution; and transportation fuel use in vehicles (also called well-to-wheels analysis).

fuel stage The second stage of a transportation fuel cycle; the fuel stage includes fuel production, storage, transportation, and distribution, in addition to activities such as fuel liquefaction and compression for long-distance transportation and onboard vehicle storage.

greenhouse gases The by-products of energy use that are believed to have the potential to cause global warming. Although there are many greenhouse gases, the three major greenhouse gases that result from motor vehicle fuel use are carbon dioxide, methane, and nitrous oxide.

life cycle analysis The process of evaluating energy use and emissions of a product from "cradle to grave" (including raw material recovery, material fabrication, product manufacture, product use, product recycling, and disposal). This process is typically used for consumer products; fuel cycle analysis (well-to-wheels analysis) is usually used for transportation vehicles/fuels.

pump-to-wheels stage The second stage of a well-to-wheels analysis (also called vehicle operation stage). It begins with refueling of vehicle fuel tanks. Emissions at this stage may come from fuel combustion or conversion, fuel evaporation, and brake and tire wear.

vehicle operation stage The last stage of a transportation fuel cycle; this stage begins with refueling of vehicle fuel tanks. Emissions at this stage may come from fuel combustion or conversion, fuel evaporation, and brake and tire wear.

well-to-pump stage The first stage of the well-to-wheels analysis (also called the upstream stage), it covers all activities from the wells where fuel feedstock is recovered, to fuels available at the pumps of vehicle refueling stations. In particular, it includes activities associated with the feedstock stage and the fuel stage of a transportation fuel cycle analysis.

well-to-wheels analysis Also called a fuel cycle analysis, this process assesses all activities from the wells where fuel feedstock is recovered, to vehicle wheels where useful energy is generated and used to move vehicles. The analysis focuses in particular on energy use and emissions of energy feedstock production and transportation; transportation fuel production, transportation, and distribution; and transportation fuel use in vehicles.

 © 2004 Elsevier Inc. All rights reserved.

The spark-ignition and compression-ignition engine technologies that are currently employed in motor vehicles were developed more than 100 years ago. These conventional vehicle technologies are fueled by petroleum-derived gasoline and diesel fuels (the so-called conventional fuels). Over the past 100 years, the conventional technologies have been dramatically improved, reducing cost and increasing performance. Humanity enjoys the mobility provided by these technologies more than ever. However, their use causes negative effects, such as the exclusive dependence on petroleum by the transportation sector, air pollution (especially urban air pollution), and large amounts of greenhouse gas (GHG) emissions. Alternative transportation fuels to replace conventional fuels and advanced vehicle technologies (such as electric drive technologies) to increase vehicle energy efficiency have long been studied and promoted worldwide. Alternative transportation fuels and advanced vehicle technologies could help solve urban air pollution problems, reduce GHG emissions, and reduce U.S. dependence on imported oil. To evaluate the energy and emission effects of alternative fuels and vehicle technologies accurately and adequately, researchers must consider emissions and energy use from fuel production processes as well as from vehicle operations (so-called fuel cycle analysis). This is especially important for technologies that employ fuels with distinctly different primary energy sources and fuel production processes because the emissions and energy use associated with fuel production and distribution can be significantly different.

1. INTRODUCTION

Fuel cycle analysis is requisite when comparing different vehicle technologies powered by different fuels. Figure 1 shows the significantly different greenhouse gas (GHG) emission results obtained by considering only vehicle tailpipe emissions (pump to wheels) versus considering both fuel production (well to pump) and tailpipe emissions (fuel cycle). A number of studies have been conducted to estimate fuel cycle emissions and energy use associated with various transportation fuels and vehicle technologies. The results of those studies were influenced by the assumptions made by individual researchers regarding technology development, emission controls, primary fuel sources, fuel production processes, and many other factors. Because different methodologies and parametric assumptions were used by different researchers, it is difficult to compare and reconcile the results of different studies and to conduct a comprehensive evaluation of fuel cycle emissions and energy use. Computer models for calculating emissions and energy use are needed to allow analysts and researchers to test their own methodologies and assumptions and make accurate comparisons of different technologies.

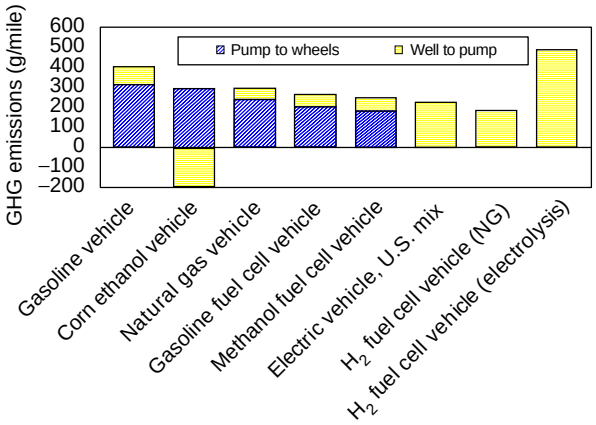


FIGURE 1 Tailpipe vs fuel cycle greenhouse gas emissions of selected vehicle/fuel systems. NG, natural gas.

The effects on energy use and emissions of using advanced vehicle/fuel systems can differ considerably because (1) the production technologies and distribution pathways for the fuels are significantly different and (2) the fuel economy and emissions performance associated with the vehicles powered by these fuels vary. If the energy and environmental effects of vehicle/fuel systems are to be thoroughly examined, the fuels must be evaluated on the basis of a full fuel cycle, including both fuel production and in-vehicle usage.

2. ALTERNATIVE TRANSPORTATION FUELS AND ALTERNATIVE FUEL VEHICLES

2.1 Alternative Transportation Fuels

At present, in the United States and worldwide, motor vehicles are fueled almost exclusively by petroleum-based gasoline and diesel fuels. Since the first oil price shock in 1973, efforts have been made to seek alternative fuels to displace gasoline and diesel fuels and achieve energy and environmental benefits. Some of the alternative fuels that have been researched and used are liquefied petroleum gas (LPG), compressed

natural gas (CNG), liquefied natural gas (LNG), methanol (MeOH), dimethyl ether (DME), Fischer–Tropsch diesel (FTD), hydrogen (H_2), ethanol (EtOH), biodiesel, and electricity. Production processes associated with gasoline, diesel, and each of these alternative fuels differ.

For fuel cycle analyses, gasoline often serves as the baseline fuel to which other fuels are compared. Because of emission concerns, requirements for gasoline quality are in place in many countries to help reduce motor vehicle emissions. In the United States, so-called reformulated gasoline, compared to the gasoline used in the 1970s, has much better quality in terms of emission reduction potential. Oxygen additives such as methyl *tert*-butyl ether (MTBE) or ethanol are often mixed with gasoline. To control emissions of nitrogen oxides (NO_x) and particulate matter of size smaller than 10 μm (PM_{10}) with after-combustion treatments for diesel vehicles, regulations are in place or are being developed to maintain the sulfur content of diesel fuels at a very low level. The demand for both low-sulfur gasoline and low-sulfur diesel requires significant desulfurization efforts in petroleum refineries.

Liquefied petroleum gas, primarily propane, is produced in either petroleum refineries or natural gas (NG) processing plants. In the United States, about 40–50% of LPG is produced in petroleum refineries and the remaining 50–60% is produced in NG processing plants. Natural gas can be delivered to vehicle refueling stations, where it is compressed to above 3600 psi for storage on CNG vehicles, or it can be liquefied in central liquefaction plants and transported to refueling stations for LNG vehicle applications. Compared with compressed natural gas vehicles (CNGVs), liquefied natural gas vehicles (LNGVs) have one distinct advantage, a longer driving range per refueling. However, cryogenic storage of LNG onboard a vehicle presents technical and cost challenges.

Methanol is produced through synthesis of a gaseous mixture of H_2, carbon monoxide (CO), and carbon dioxide (CO_2). Although methanol can be produced from biomass, coal, heavy oil, naphtha, and other feedstocks, the availability of reasonably cheap NG feedstock makes steam methane reforming (SMR) technology an economical way to produce methanol. SMR is a mature technology and is widely used in existing methanol plants. Methanol plants may be located near remote gas fields to take advantage of cheap, remote NG supplies.

Dimethyl ether, which has physical properties similar to those of LPG, has been proposed and tested as an alternative to diesel fuel in compression-ignition (CI) engines. Use of DME in diesel engines offers NO_x and PM emissions reduction benefits. DME is now used predominantly as an aerosol propellant and is produced from methanol through a dehydration process. Research and development efforts have focused on technologies to produce DME directly from coal and NG.

The Fischer–Tropsch (FT) process produces middle distillates containing no sulfur and virtually no aromatics; the process also produces naphtha and wax. Using FT diesel in CI engines helps reduce NO_x and PM emissions. The FT reaction process was used by Germany during World War II to produce diesel fuel from coal and by South Africa during the oil embargo against that country's apartheid. Currently, several major companies are actively pursuing the production of FTD from remotely located ("stranded") natural gas.

Although H_2 can be produced from many feedstocks, most H_2 is currently produced from NG via SMR technologies. Several alternative schemes, such as partial oxidation (POX), autothermal reforming (ATR), and plasma reforming, have been developed and used commercially to produce H_2. Hydrogen plants can generate a significant amount of steam; some of the steam produced is used as process steam within the plant, and the remainder can be exported to nearby chemical plants. The gaseous H_2 ($G.H_2$) produced can be compressed to above 5000 psi for storing on board hydrogen-powered vehicles, or the $G.H_2$ can be liquefied to produce liquid H_2 ($L.H_2$) and used in hydrogen-powered vehicles. Besides production in central plants, H_2 can be produced in refueling stations to avoid construction of expensive hydrogen distribution infrastructure. In addition, H_2 can be produced via electrolysis of water with electricity that is generated by using different energy sources.

Biomass can be used to produce ethanol and biodiesel. Ethanol fuel is currently produced from grains such as corn and sugarcane. Research and development efforts to develop and commercialize technologies to produce ethanol from cellulosic biomass are ongoing. Methyl or ethyl esters that are produced from vegetable oils and animal fats are commonly called biodiesel (BD). Biodiesel, which can be used in diesel engine vehicles, is produced via a transesterification process from natural vegetable oils (soybean oil, cotton oil, and rapeseed oil). Outside the United States production is mainly from rapeseed oil; in the United States, biodiesel is produced primarily from soybeans.

Electricity can be used directly as a transportation fuel in battery-powered electric vehicles (EVs); in

combination with an onboard engine, electricity can supply the motive power for hybrid electric vehicles (HEVs). Also, electricity can be used to produce H_2 via electrolysis of water. The energy use and emission effects of electricity as a transportation fuel are primarily determined by the type of fuels (e.g., coal, NG, or oil) used and the type of fuel combustion technologies (e.g., steam boilers, simple-cycle gas turbines, or combined-cycle gas turbines for NG-fired power plants). Because the electric generation mix is so different from region to region, the energy and emission effects of electricity also vary greatly.

2.2 Alternative Fuel Vehicles

Figure 2 shows the numbers and types of alternative fuel vehicles on the road in the United States. Besides these conventional vehicle technologies, advanced vehicle technologies such as HEVs and fuel cell vehicles (FCVs) are also under development. Past fuel cycle analyses have covered both conventional and advanced vehicle technologies. Some of the alternative fuels (such as methanol, ethanol, and biodiesel) can be blended with gasoline or diesel for vehicle applications.

3. *FUEL CYCLE ANALYSES OF VEHICLES AND TRANSPORTATION FUELS*

The fuel cycle for a given transportation fuel includes the following processes: energy feedstock (or primary energy) production; feedstock transportation and storage; fuel production; fuel transportation, storage, and distribution; and vehicle operations that involve fuel combustion or other chemical conversions (Fig. 3). The processes that precede vehicle operations are often referred to as the well-to-pump (WTP) stage, the vehicle operations are referred to as the pump-to-wheels (PTW) stage, and the entire fuel cycle is referred to as the well-to-wheels (WTW) cycle. Various models have been developed that allow researchers to conduct fuel cycle analyses of vehicle/fuel systems. The GREET (greenhouse gases, regulated emissions, and energy use in transportation) model, developed by Argonne National Laboratory, is in the public domain and, so far, more than 1000 users worldwide have used the model to predict the energy use and emissions associated with alternative fuels and vehicles. The GREET model and its documents are posted on the Internet at http://greet.anl.gov.

Mark Delucchi of the University of California at Davis has developed another transportation fuel cycle models. The model has been used by Delucchi for applications in the United States, Canada, and Europe. Another analytic tool, the TEAM model, developed by Ecobalance, is a commercial model used primarily for life cycle analysis of consumer products. In certain configurations, the model can be used to analyze vehicle/fuel systems. The National Renewable Energy Laboratory in the United States has applied the model to analyze bioethanol and biodiesel. Shell (Royal Dutch Petroleum Company) recently applied the model to analyze FT diesel.

For fuel cycle analysis of vehicle/fuel systems, the results need to be per distance traveled. If the analysis is conducted for each unit of energy produced and used, the energy efficiency and emission differences during the pump-to-wheels (vehicle operation) stage may be excluded, resulting in incomplete analysis of

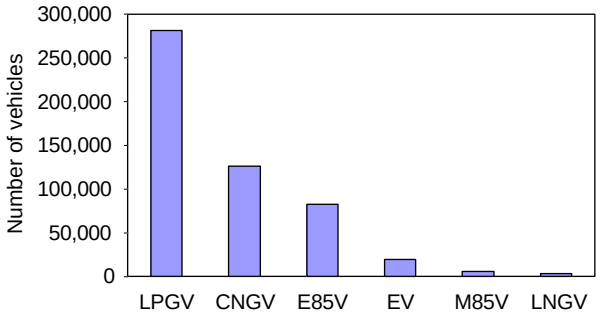


FIGURE 2 Stocks of alternative fuel vehicles in the United States in 2002. LPGV, Liquefied petroleum gas vehicles; CNGV, compressed natural gas vehicles; E85V, vehicles fueled by 85% ethanol/15% gasoline (by volume); EV, electric vehicles (battery powered); M85V, vehicles fueled by 85% methanol/15% gasoline (by volume); LNGV, liquefied natural gas vehicles. Data from the Energy Information Administration (U.S. Department of Energy) Web site at http://www.eia.doe.gov, accessed on Nov. 12, 2002.

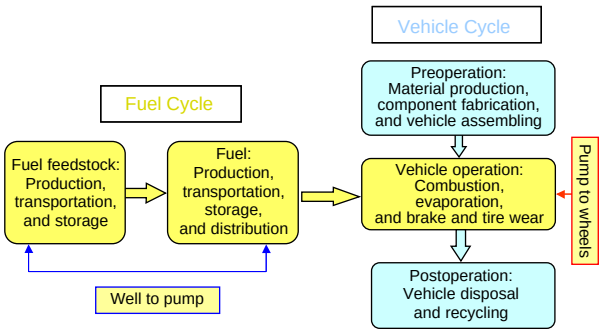


FIGURE 3 Coverage of a fuel cycle analysis.

the systems. To estimate well-to-wheels energy use and emissions, the well-to-pump energy use and emissions per unit of fuel available at refueling pumps are estimated first. The WTP estimates take into account energy uses (including use of process fuels and energy losses during WTP activities) and emissions associated with the activities, as discussed in reports prepared by Delucchi and Wang about their models. Estimates of WTW emissions usually include emissions of greenhouse gases [primarily CO_2, methane (CH_4), and nitrous oxide (N_2O)] and criteria pollutants such as volatile organic compounds, CO, NO_x, PM_{10}, and sulfur oxides (SO_x). Emissions occurring during WTP activities include those resulting from the combustion of process fuels and from noncombustion processes such as chemical reactions, fuel leakage, and evaporation. Of these pollutants, CO_2 emissions are relatively easy to estimate because they depend almost exclusively on the carbon content of the fuels used. Emissions of other pollutants are affected by combustion technologies, as well as the amounts and the types of fuels used. To a large extent, emissions estimates for these other pollutants rely heavily on emission factors (in grams per unit of fuel used) for different combustion technologies fueled by different process fuels. These emission factors can be obtained from databases such as the AP-42 document created by the U.S. Environmental Protection Agency (EPA).

Sulfur oxide emissions are calculated by assuming that all sulfur contained in process fuels is converted into sulfur dioxide (SO_2) for combustion technologies fueled with process fuels except for coal and residual oil. The SO_x emissions for these two process fuels are usually subject to after-combustion control devices. PTW energy use and emissions are calculated on a per-mile basis. Usually, energy use per distance traveled is calculated from vehicle fuel economy values, which are available from measurements or vehicle simulations. Vehicle emissions values are available from emission certifications, estimations of on-road vehicle operations, and laboratory tests.

Although fuel economy and emissions for benchmark gasoline vehicles and diesel vehicles are presented as actual values (e.g., miles per gallon for fuel economy and grams per mile for emissions), fuel economy and emissions for alternative fuel vehicles (AFVs) are usually presented as ratios to benchmark vehicles. In this way, differences in vehicle operating conditions can be reduced to some extent, as long as the fuel economy and emission ratios are for similar vehicle types and under similar operating conditions. For criteria pollutants, location (e.g., urban, rural) is also important, and fuel cycle analyses need to be conducted with a special focus on where emissions occur. In the GREET model, criteria pollutant emissions are separated into "total emissions" (those that occur everywhere) and "urban emissions" (those that occur within metropolitan areas, where large populations are exposed to the pollutants).

4. FUEL PATHWAYS AND VEHICLE/FUEL COMBINATIONS

Table I shows different types of vehicle propulsion systems and the transportation fuels that have been studied for their potential to power the vehicles. Gasoline, CNG, LNG, LPG, methanol, ethanol, and hydrogen can be used in vehicles equipped with conventional spark-ignition (SI) engines. Interest in developing efficient, low-emission, spark-ignition direct-injection (SIDI) engine technologies has heightened in recent years. Although vehicles with SIDI engines can be fueled efficiently with liquid fuels (gasoline, methanol, ethanol, and liquid hydrogen), gaseous fuels do not appear to offer inherent fuel economy benefits when used in these engines. Compression-ignition direct-injection (CIDI) engines, which can be fueled by diesel, DME, FT diesel, and biodiesel, help increase vehicle fuel

TABLE I

Potential Combinations of Vehicle Technologies and Transportation Fuels

Vehicle technology[a]	Potential fuels for use[b]
Conventional SI engine	Gasoline, CNG, LNG, LPG, MeOH, EtOH, H_2
SIDI engine	Gasoline, MeOH, EtOH, $L.H_2$
CIDI engine	Diesel, dimethyl ether, Fischer–Tropsch diesel, biodiesel
SIDI engine HEV	Gasoline, CNG, LNG, LPG, MeOH, EtOH, $L.H_2$
CIDI engine HEV	Diesel, dimethyl ether, Fischer–Tropsch diesel, biodiesel
Battery-powered EV	Electricity
FCV	Hydrogen, MeOH, gasoline, EtOH, diesel, CNG, LNG, LPG, naphtha

[a] Abbreviations: SI, spark ignition; CIDI, compression ignition, direct injection; HEV, hybrid electric vehicle; EV, electric vehicle; FCV, fuel cell vehicle.

[b] Abbreviations: CNG, compressed natural gas; LNG, liquefied natural gas; LPG, liquefied petroleum gas; MeOH, methanol; EtOH, ethanol; $L.H_2$, liquid hydrogen.

economy substantially, although they face the challenge of reducing NO_x and PM emissions.

Hybrid electric vehicles achieve significant fuel economy gains compared with conventional vehicle powertrain configurations. HEVs can be equipped with conventional SI, SIDI, or CIDI engines. HEVs can also be designed to obtain electricity from the electric grid and combine it with power from onboard engines. Battery-powered EVs, charged with grid electricity, offer large fuel economy gains and zero tailpipe emissions. The limited driving range, limited performance characteristics, and high battery costs of EVs limit their market penetration, but EVs continue to be promoted.

Major automakers and the government have programs aimed at development of fuel cell vehicles. Although the ultimate fuel for FCVs is hydrogen, hydrocarbon fuels are being considered, at least for a transition period, before expensive hydrogen production and distribution infrastructure is in place. In theory, any hydrocarbon fuel (e.g., methanol, gasoline, ethanol, diesel, CNG, LNG, LPG, and naphtha) can be used to produce hydrogen on board FCVs. Use of hydrocarbon fuels in FCVs requires onboard fuel processors to produce H_2.

5. ENERGY AND EMISSIONS: ITEMS OF INTEREST

A typical life cycle analysis includes analysis of the inventory of environmental releases, evaluation of potential environmental impacts, and interpretation of the inventory analysis and impact evaluation. Environmental releases include emissions to air, water, and solid waste emissions; these are defined in document 14040 of the International Standards Organization (ISO), *Environmental Management, Life-Cycle Assessment—Principles and Framework*. Prior fuel cycle analyses of vehicle/fuel systems have focused primarily on air emissions and energy use. For air emissions, studies often include the following criteria:

1. Greenhouse gas emissions: carbon dioxide (CO_2), methane (CH_4), nitrous oxide (N_2O), and several other gases, including hydrofluorocarbons, perfluorocarbons, and sulfur hexafluorides.
2. Criteria pollutant emissions: volatile organic compounds (VOCs), carbon monoxide (CO), nitrogen oxides (NO_x), particulate matter [usually smaller than 10 μm (PM_{10})], and sulfur oxides (SO_x).
3. Emissions of air toxics: acetaldehyde, benzene, 1,3-butadiene, and formaldehyde.
4. Energy use parameters: total energy use, fossil energy use, and petroleum use.

To estimate CO_2-equivalent GHG emissions, the GHGs are often combined with their global warming potentials (GWPs) as described by the Intergovernmental Panel on Climate Change (IPCC) in its 2001 report, *Climate Change 2001: The Scientific Basis, Technical Summary of the Group I Report*. The three major GHGs for evaluating vehicle/fuel systems are CO_2, CH_4, and N_2O. The GWPs of the greenhouse gases vary with the time horizon used to calculate them. With a time horizon of 100 years, GWPs are 1, 23, and 296 for CO_2, CH_4, and N_2O, respectively.

Emissions of criteria pollutants are regulated in the United States by the EPA (the term "criteria pollutants" comes from the EPA's establishment of criteria for ambient concentrations and emissions of these pollutants). These pollutants primarily cause adverse health effects. Thus, the locations in which they are emitted as well as their amounts are important. Total emissions are emissions that occur everywhere, including unpopulated or sparsely populated regions, and urban emissions are those that occur in urban areas with large populations.

Air toxics are hydrocarbons identified by EPA as potential cancer-causing species. Acetaldehyde, benzene, 1,3-butadiene, and formaldehyde are the major air toxics emitted by motor vehicles. Air toxic emissions of various vehicle/fuel combinations have been assessed in a study at Argonne National Laboratory by Winebrake and colleagues.

All complete fuel cycle analyses of energy impacts include total energy use; which means that all energy sources, such as renewable energy and nonrenewable energy (i.e., fossil energy), are taken into account. Some analyses further analyze fossil energy impacts; in these studies, assessments take into account fossil energy, including petroleum, natural gas, and coal. Because fossil energy resources are finite, analyses of fossil energy use reveal the impacts of particular vehicle/fuel systems on energy resource depletion. This aspect of life cycle analysis is especially important for biofuels such as ethanol and biodiesel, for which analysis of total energy impacts may not provide adequate information because the energy in ethanol and biodiesel comes from solar energy during crop growth, and use of solar energy per se is not a concern. Some studies also include analysis of petroleum displacement effects by vehicle/fuel

systems. This is an important issue for countries that rely on imported oil and seek to reduce their oil consumption.

6. ISSUES INVOLVED IN TRANSPORTATION FUEL CYCLE ANALYSES

There are several key issues that transportation researchers need to keep in mind when conducting and using the data that result from life cycle analysis studies.

6.1 Usefulness of Fuel Cycle Analysis Results

ISO document 14040 identifies four areas for using life cycle analysis (LCA) results: (1) identifying opportunities to improve the environmental aspects of products, (2) helping industry, governments, and nongovernmental organizations make sound environmental decisions, (3) selecting relevant environmental indicators for measuring the environmental performance of products, and (4) marketing claims of the environmental performance of products. For transportation fuel cycle analyses (FCAs), the first two areas are important. The results of FCAs of vehicle/fuel systems are helpful in comparing the energy and emissions performance of different systems on a complete basis. Such comparisons help society identify and pursue the technologies that offer the greatest potential energy and emissions benefits; for example, they can be used to rank technologies in relative terms rather than estimating the absolute amount of energy use and emissions for given technologies. FCA results may be less useful for estimating total energy use and total emissions for all the activities associated with a vehicle/fuel system.

As ISO 14040 states, the results of LCAs in general, and transportation FCAs in particular, can be very helpful in identifying opportunities to improve energy performance and environmental conditions. That is, FCA results should be used in a dynamic way to identify the subsystems that can contribute to improved energy and environmental performance of a vehicle/fuel system and to identify potential system improvements, and not in a static way to show simply which systems are better than others.

6.2 Allocation Methods

For processes that produce multiple products, energy and emission burdens have to be allocated to individual products. ISO 14040 advocates reducing the need to allocate by expanding the system boundary. That is, when comparing two systems, the system boundaries for both can be expanded so that all products from both systems are within a newly expanded single system. In transportation FCAs, the system boundary expansion is also called the displacement method (for a full discuss ion, see the 1999 report by Wang, *Transportation Fuel-Cycle Model, Volume 1: Methodology, Development, Use, and Results)*. In practice, however, the system expansion approach can become time and resource intensive. Also, in some cases, systems may be expanded subjectively without considering marketplace effects. Thus, allocation of energy and environmental burdens is still widely used in transportation FCAs. The burdens can be allocated based on shares of output products in terms of energy content, market value, or mass. Of these three options, the energy content basis is appropriate if all products are energy products. The market value basis could be used if the market values of products will drive market behaviors, which can, in turn, affect technologies. The mass basis usually is the least favorable method for transportation FCAs. In any event, all products need to be taken into account in evaluating multiple-product processes and systems. Ignoring coproducts, as some past studies have done, results in unfair treatment of the primary product of a process or a system. Also, the fact that the coproducts indeed have market values is ignored.

6.3 Technology Maturity Levels

In comparative transportation FCAs, different fuels and vehicle technologies are often analyzed together. Comparisons of technologies should take into consideration different commercial maturity levels; some fuels and technologies are already mature and in use, whereas others may still be in the research and development stage. Because FCAs generally do not address technology maturity and commercial readiness, it is important to remember that FCAs and their results cannot be used to assess the commercial viability of fuels and vehicle technologies. FCAs simply provide information on energy and emission effects that would occur if the evaluated fuels and technologies were to be introduced into the marketplace.

The level of technology maturity also affects FCA data quality; For technologies at the research and

development stage, few or no actual operational data are available. Sources of data for these technologies are often limited to engineering analyses, predictions, and simulations. There is typically an abundance of reliable performance data for fuels and vehicle technologies already in the marketplace.

6.4 Uncertainties in Input Parameters and Output Results

Because FCAs cover many activities from wells to wheels, there is a demand for data. Data associated with different activities can come from a wide variety of sources of differing quality, and different engineering designs and facility operation can vary significantly. To address the reality that the data obtained are subject to a large degree of uncertainty, many FCAs have been conducted using sensitivity analyses that change key input parameters. Many researchers now incorporate Monte Carlo simulations in their studies to address uncertainties on a systematic basis. With Monte Carlo simulations, inputs parameters and output results are associated with probability distribution functions.

6.5 Transparency

Transportation FCA results are determined by assumptions regarding input parameters, simulation methodologies, and system boundaries. These assumptions need to be presented explicitly in transparent ways to allow studies to be compared and to make clear whether general conclusions can be drawn from individual studies. Sensitivity analyses can be included in FCA studies to demonstrate the importance of some of the assumptions in determining results. Sometimes, seemingly similar studies are combined to generalize conclusions.

7. COMPLETED STUDIES

Some of the results of past studies are summarized in the following sections. Although many of the studies estimated both (1) absolute energy use and emissions of vehicle/fuel combinations and (2) energy use and changes in emissions of alternative vehicle/fuel systems relative to benchmark technologies (usually gasoline-fueled light-duty vehicles), the summary here presents relative changes only, because they are more comparable than absolute amounts across studies.

Of the vehicle types that use alternative transportation fuels, most completed studies evaluated only passenger car applications. However, many alternative transportation fuels and advanced vehicle technologies can also be applied to heavy-duty trucks. The benchmark vehicles for heavy-duty vehicle comparisons are usually diesel trucks. Because heavy-duty vehicle technologies can be considerably different from light-duty vehicle technologies and because heavy-duty vehicle driving patterns are very different from those for passenger cars, the relative FCA results summarized here are generally not applicable to heavy-duty vehicles.

7.1 Kreucher, Ford Motor Company (1998)

The study by W. M. Kreucher included evaluations of well-to-wheels emissions, energy use, and costs. Vehicle fuel economy for a Ford Escort was simulated by using Ford's proprietary model on the federal test procedure (FTP) cycle. Well-to-pump energy and emissions were simulated with a model developed by Arthur D. Little for the U.S. Department of Energy (DOE). Emissions analysis included CO_2 and criteria pollutants.

7.2 Weiss *et al.*, Massachusetts Institute of Technology (2000)

Weiss and others at the Energy Laboratory at the Massachusetts Institute of Technology (MIT) conducted a study to compare several vehicle propulsion systems fueled by alternative transportation fuels. In conducting the comparison, the researchers projected vehicle and fuel production technologies from 1996 to 2020. Their analysis was primarily for North American applications. Greenhouse gas emissions in that study include CO_2 and CH_4. A midsize passenger car, such as a Toyota Camry, was used for vehicle simulations. Well-to-pump data were based on a critical review of published data. Vehicle fuel economy results were simulated by using an MIT vehicle computer simulation model. One of the study's strengths is vehicle characterization and simulations. The study also presents some detailed information on vehicle performance and technology advancement.

7.3 (S&T)^2 Consultants, Inc., for Methanex (2000)

The (S&T)^2 study was conducted primarily for Canadian well-to-pump conditions, but the results

are somewhat applicable to U.S. WTP conditions. The study included GHG emissions only (CO_2, CH_4, and N_2O) and employed Canadian oil recovery efficiencies and carbon emissions, which are high because a significant amount of oil is recovered from oil sands in Canada. Also, the study assumed low refinery efficiency and high GHG emissions for petroleum refineries. Emissions from vehicle materials cycles were also included. A derivative of the model developed by Delucchi was used for simulating well-to-wheels GHG emissions. The study is unique in its collection and use of Canadian WTP data on GHG emissions and fuel production efficiencies.

7.4 Ishitani, University of Tokyo (2000)

The study by H. Ishitani was conducted assuming Japanese conditions for both WTP activities and vehicle technologies. Vehicle fuel economy was simulated with a model that the author developed. The study included well-to-wheels energy efficiencies only. It appears that well-to-pump and pump-to-wheels efficiency assumptions were more optimistic in this study than in the other studies reviewed here.

7.5 Louis, Shell Global Solutions (2001)

The Shell study estimated well-to-wheels GHG emissions and energy use for 24 vehicle/fuel systems. Greenhouse gas emissions included CO_2 and CH_4. The vehicle technologies that were evaluated were internal combustion engine vehicles (ICEVs), HEVs, and FCVs. The analysis for WTP activities was conducted for the United Kingdom primarily with Shell data. A Mercedes A-class passenger car was chosen for comparison. The GREET model developed by Argonne National Laboratory was used in the study to conduct WTW simulations.

7.6 Delucchi, University of California at Davis (2001)

The study by M. A. Delucchi evaluated motor vehicles, forklifts, and space and water heaters. It included emissions of GHGs and criteria pollutants, but not energy use, although the Delucchi model includes estimation of energy use. The CO_2-equivalent GHG emissions addressed in the study were CO_2, CH_4, N_2O, nonmethane organic compounds (NMOCs), CO, NO_x, SO_x, PM, and hydrofluorocarbon (HFC)-134a and chlorofluorocarbon (CFC)-12. Delucchi included both fuel and vehicle cycles in his analysis. The results summarized in Fig. 3 are fuel cycle results. Delucchi showed that vehicle cycle GHG emissions could be 16% of vehicle tailpipe GHG emissions.

7.7 Arthur D. Little (2002)

Arthur D. Little (ADL) has conducted studies for the DOE and the state of California since the early 1990s. The report cited here was prepared for the DOE to evaluate fuel choices for FCVs. The study considered energy efficiencies, GHG emissions, costs, and safety for use of the following fuels in FCVs, ICEVs, and battery-powered EVs: hydrogen, methanol, ethanol, gasoline, and diesel. The analysis was conducted for year 2010. ADL developed a vehicle model to simulate vehicle fuel economy under different driving cycles. A modified GREET version was used for WTW simulations.

7.8 Summary

Tables II and III summarize changes in WTW energy use (Table II) and GHG emissions (Table III) by vehicle/fuel system relative to baseline gasoline vehicle data from the studies described in the preceding sections. Energy use results from five of the seven studies (the Methanex and Delucchi studies did not address energy use) indicate that, of the nine SI engine vehicle/fuel systems, use of cellulosic ethanol, corn ethanol, methanol, and H_2 in SI engines could lead to increased WTW energy uses, whereas use of LPG and CNG results in reduced energy use. Of the three CIDI engine vehicle/fuel systems, use of FT diesel could result in increased WTW energy use, because of large energy losses during FT diesel production. Use of all of the eight HEVs results in reductions in energy use. Among the FCVs, use of electrolysis H_2, cellulosic ethanol, and corn ethanol could result in increased energy use, because of very large energy losses during WTP activities for these fuels. Of the four studies analyzing gasoline FCVs, only the MIT study concluded that gasoline FCVs could have higher WTW energy use compared to baseline gasoline vehicles. Of the five studies that addressed EVs, two showed increases in energy use by EVs, but three showed reductions.

The WTW GHG emission changes presented in Table III for six of the reviewed studies (the study by the University of Tokyo did not present GHG emission results) reveal that, of the nine fuels for SI engine vehicle applications, NG-based L.H_2 could

TABLE II

Changes in Well-to-Wheels Total Energy Use by Vehicle/Fuel Systems Relative to Gasoline Vehicles[a]

	Research group[c]				
Study[b]	Ford	MIT	Univ. Tokyo	Shell	ADL
SI engine vehicles		1−			
Advanced gasoline ICEV			3−	1−	
LPG ICEV	1−				
CNG ICEV	−				
Corn EtOH ICEV	4+				
Cellulosic EtOH ICEV	3+				
Methanol ICEV	2+				
NG G.H_2 ICEV				1+	1+
NG L.H_2 ICEV				2+	
NG MH H_2 ICEV				1−	
CIDI engine vehicles					
Diesel ICEV	2−	2−		2−	2−
FTD ICEV		2+		1+	
DME ICEV				1−	
HEVs					
Gasoline HEV		2−		2−	2−
CNG HEV		3−		2−	
NG G.H_2 HEV				2−	
NG L.H_2 HEV				1−	
NG MH H_2 HEV				2−	
Diesel HEV		3−		3−	2−
FTD HEV		1−		1−	
DME HEV				2−	
FCVs and EVs					
NG G.H_2 FCV		2−	3−	2−	3−
kWh G.H_2 FCV					2+
NG L.H_2 FCV				1−	2−
NG MH H_2 FCV				2−	3−
Gasoline FCV		1+	3−	2−	3−
Methanol FCV		1−	2−	2−	2−
Corn EtOH FCV					1+
Cellulosic EtOH FCV					1+
CNG FCV				3−	
FTD FCV				1−	
DME FCV				2−	
EV	4+	2−	3−	2+	2−

[a] Change in use: 4+, increase of 60–80%; 3+, increase of 40–60%; 2+, increase of 20–40%; 1+, increase of 0–20%; 1−, decrease of 0–20%; 2−, decrease of 20–40%; 3−, decrease of 40–60%; 4−, decrease of 60–80%; 5−, decrease of greater than 80%.

[b] Abbreviations: ICEV, internal combustion engine vehicle; LPG, liquefied petroleum gas; CNG, compressed natural gas; EtOH, ethanol; NG, natural gas; G.H_2, L.H_2, MH H_2, gaseous, liquid, and metal hydride hydrogen; CIDI, compression ignition, direct injection; FTD, Fischer–Tropsch diesel; DME, dimethyl ether; HEV, hybrid electric vehicle; FCV, fuel cell vehicle; EV, electric vehicle.

[c] MIT, Massachusetts Institute of Technology; ADL, Arthur D. Little.

result in increased WTW GHG emissions. For methanol, one study showed increased GHG emissions and another study showed reductions. Use of all of the other seven fuels results in GHG emissions reductions. Of the three fuels for CIDI engine vehicle applications and eight fuels for HEV applications, all

TABLE III

Changes in Well-to-Wheels Greenhouse Gas Emissions by Vehicle/Fuel Systems Relative to Gasoline Vehicles[a]

	Research group					
Study	Ford	MIT	Methanex	Shell	Delucchi	ADL
SI engine vehicles		1−				
Advanced gasoline ICEV						
LPG ICEV	1−				2−	
CNG ICEV	1−			2−	2−	
Corn EtOH ICEV	2−				1−	
Cellulosic EtOH ICEV	3−				3−	
Methanol ICEV	1+				1−	
NG G.H_2 ICEV				1−	2−	1−
NG L.H_2 ICEV				1+		
NG MH H_2 ICEV				2−		
CIDI engine vehicles						
Diesel ICEV	1−	2−		2−	2−	2−
FTD ICEV		1−		1−		
DME ICEV				2−		
HEVs						
Gasoline HEV		2−		2−		2−
CNG HEV		3−		3−		
NG G.H_2 HEV				3−		
NG L.H_2 HEV				2−		
NG MH H_2 HEV				3−		
Diesel HEV		3−		3−		2−
FTD HEV		2−		2−		
DME HEV				3−		
FCVs and EVs						
NG G.H_2 FCV		2−	3−	3−		3−
kWh G.H_2 FCV			1−			1+
NG L.H_2 FCV			3−	2−		3−
NG MH H_2 FCV	3−			3−		4−
Gasoline FCV		1+	2−	2−		2−
Methanol FCV		2−	3−	3−		2−
Corn EtOH FCV						3−
Cellulosic EtOH FCV						5−
CNG FCV				3−		
FTD FCV			2−	2−		
DME FCV				2−		
EV	2−	2−		2−	2−	2−

[a] Change in use: 3+, increase of 40–60%; 2+, increase of 20–40%; +, increase of 0–20%; 1−, decrease of 0–20%; 2−, decrease of 20–40%; 3−, decrease of 40–60%; 4−, decrease of 60–80%; 5−, decrease of greater than 80%.

result in GHG emissions reductions. All of the 12 FCV and EV types are shown to have GHG emissions reduction benefits, except for electrolysis H_2 and gasoline for FCV applications, for which two studies showed increases (the four studies showed decreases for these two options).

8. FUEL CYCLE ENERGY AND EMISSION RESULTS OF VEHICLE/FUEL SYSTEMS

Because different studies have different system boundaries and parametric assumptions, the studies

described in Section 7 resulted in different magnitudes of changes in energy use and GHG emissions. The GREET model was used to develop the results presented in this section. In conducting the analyses with the GREET model, key assumptions were specified by taking into account the results from various completed studies. Because some of the vehicle/fuel systems are emerging technologies and are not in the marketplace yet, all technologies were analyzed for the time frame around 2010, so that both emerging technologies and technologies already in the marketplace can be compared within the same period.

8.1 Specifications of and Key Assumptions for Fuel Production Pathways

Table IV presents specifications of and key assumptions of fuel production pathways for North American applications. For each fuel pathway, the table presents the fuel pathway (including feedstocks), key parametric assumptions, and main pathway characteristics. Note that the petroleum gasoline pathway is the baseline pathway to which other fuel production pathways are compared.

8.2 Vehicle Fuel Economy Assumptions

Table V lists the fuel economy of baseline gasoline vehicles (GVs) and the fuel economy ratios of alternative vehicle/fuel systems relative to the baseline GV. The values are based on research at Argonne National Laboratory of various vehicle technologies and fuel economy simulations from studies completed at Argonne and elsewhere. The baseline GV fuel economy [27.5 miles per gallon (mpg)) represents the on-road fuel economy value for a typical midsize passenger car produced in 2010 (there is a considerable penalty from laboratory-tested mpg to on-road actual mpg). The mpg ratios for alternative vehicle/fuel systems are based on test results and vehicle simulations. These values were used in the GREET model to conduct WTW simulations.

8.3 Well-to-Pump Energy Efficiencies

Figure 4 shows the energy efficiencies of various fuel production pathways from well to pump. The efficiencies shown are defined as the energy in a given fuel (available at pumps in vehicle refueling stations) divided by total energy inputs during all WTP activities, including the energy content of the fuel. One way to interpret the WTP efficiencies in the figure is as the difference between 100% and the energy efficiencies, which roughly represent energy losses during WTP stages for making a given fuel available at the pump. As stated in Section 3, WTP activities include feedstock production; feedstock transportation and storage; fuel production; and fuel transportation, storage, and distribution.

The reformulated gasoline (RFG) pathway is considered to be the baseline pathway to which other pathways are compared. Figure 4 shows that the baseline RFG pathway has a WTP efficiency of about 80%. There are four fuel pathways, LPG, CNG, LNG, and low-sulfur (LS) diesel, that have WTP efficiencies higher than that of the baseline RFG pathway. These pathways have fuel production efficiencies higher than the gasoline refining efficiency (see Table IV).

Fuel pathways in the next group have energy efficiencies of 50–70%. These pathways include biodiesel from soybeans, DME from NG, methanol from NG, FT diesel from NG, G.H_2 from NG, and ethanol from corn. For renewable biodiesel and ethanol, energy losses from photosynthesis during soybean and corn plant growth are not included because solar energy is abundant. However, energy is consumed for soybean and corn farming and production of biodiesel and ethanol. These effects result in lower WTP efficiencies for these two renewable fuel pathways, compared to the baseline RFG pathway. The NG-based pathways for DME, methanol, FT diesel, and G.H_2 have lower WTP efficiencies because they are subject to considerable energy losses during fuel production.

The next five fuel pathways have WTP energy efficiencies of 30–50%. They include European Union (EU) electricity generation, ethanol from cellulosic ethanol, L.H_2 from NG, U.S. electricity generation, and G.H_2 from EU electricity via electrolysis. Electricity generation with the EU/U.S. generation mix has low WTP efficiencies because electricity generation efficiencies are usually low (see Table IV). Note that for nuclear power generation, an energy efficiency of 100% is assumed here, even though nuclear power plant efficiencies are usually in the range of 35–40%. This is because the energy content in uranium used for power generation is not a concern, because a unit of mass of uranium can generate a large amount of electricity. In a way, nuclear energy could be treated in a manner similar to renewable energy. Because of the way nuclear power and renewable power (such as hydropower)

TABLE IV

Specifications and Key Assumptions for Fuel Production Pathways[a]

Fuel pathway	Key assumptions	Main pathway characteristics
Petroleum to gasoline	Petroleum refining efficiency of 85.5%	For U.S. refineries; 60% crude is imported and 40% domestic; gasoline with sulfur content of 30 ppm and with EtOH as the oxygenate.
NG to CNG	NG compression efficiency of 93%	NG is recovered in North America
NG to LNG	LNG production efficiency of 90.5%	NG is recovered outside of North America; LNG is produced outside of North America and transported to North America
Petroleum and NG to LPG	NG to LPG production efficiency of 96.5%; petroleum to LPG production efficiency of 93.5%	Of the total LPG produced for North American use, 60% is from NG and 40% from petroleum
NG to MeOH	MeOH production efficiency of 68%	Methanol is produced outside of North America and transported to North America
Corn to EtOH	Energy use for ethanol production of 45,500 Btu/gal; ethanol yield of 2.6 gal per bushel of corn	Both dry and wet milling EtOH plants are included; coproducts (primarily animal feeds) are considered
Herbaceous biomass to EtOH	EtOH yield of 92 gal per dry ton of biomass	Cogenerated electricity in cellulosic ethanol plants is considered
Petroleum to LSD	Petroleum refining efficiency of 87%	For U.S. refineries; 60% crude is imported and 40% domestic; diesel with sulfur content of 15 ppm
NG to DME	DME production efficiency of 70%	DME is produced outside of North America and transported to North America
NG to FTD	FTD production energy efficiency of 63% and carbon efficiency of 80%	FTD is produced outside of North America and transported to North America
Soybeans to biodiesel	Biodiesel yield of 1.36 gal per bushel of soybeans	Coproducts (animal feeds and glycerine) are considered
U.S. average electricity	Coal-fired power plant efficiency of 36.5%; NG-fired power plant efficiency of 41%; oil-fired power plant efficiency of 35%	The U.S. electricity generation mix: 54% coal; 15% NG; 1% oil; 18% nuclear; 12% others
European union average electricity	Coal-fired power plant efficiency of 36.5%; NG-fired power plant efficiency of 41%; oil-fired power plant efficiency of 35%	The European Union electricity generation mix: 25% coal; 18% NG; 7% oil; 34% nuclear; 16% others
NG to G.H_2	G.H_2 production efficiency of 71.5%; gaseous H_2 compression efficiency of 85%	G.H_2 is produced in North America with North American NG
NG to L.H_2	G.H_2 production efficiency of 71.5%; H_2 liquefaction efficiency of 71%	L.H_2 is produced outside of North America and transported to North America
Electricity to G.H_2 via electrolysis	G.H_2 production efficiency of 71.5%; G.H_2 compression efficiency of 86%	G.H_2 is produced at refueling stations
Electricity to L.H_2 via electrolysis	G.H_2 production efficiency of 71.5%; H_2 liquefaction efficiency of 66%	L.H_2 is produced at refueling stations

[a] Fuels are for use in North America. Abbreviations: EtOH, ethanol; NG, natural gas; CNG, compressed natural gas; LNG, liquefied natural gas; LPG, liquefied petroleum gas; MeOH, methanol; LSD, low-sulfur diesel; DME, dimethyl ether; FTD, Fischer–Tropsch diesel; G.H_2, L.H_2, gaseous and liquid hydrogen.

are treated here, the EU generation mix has a higher WTP efficiency compared to the U.S. generation mix (the EU generation mix has higher shares of nuclear power and hydropower). L.H_2 from NG has an energy efficiency of about 42% because this pathway is subject to large energy losses during H_2 production and liquefaction. The G.H_2 from EU electricity via electrolysis has an energy efficiency of 32% because of the energy losses during electricity generation and H_2 production.

The last three pathways are electrolysis-based H_2 production pathways (G.H_2 from U.S. electricity, L.H_2 from EU electricity, and L.H_2 from U.S. electricity). They have WTP efficiencies of 23–30%. These pathways are subject to large energy losses during electricity generation, H_2 production, H_2

TABLE V

Gasoline Vehicle Fuel Economy and Fuel Economy Ratios of Alternative Vehicle/Fuel Systems

Vehicle[a]	Fuel economy ratio (relative to baseline GVs)
Baseline GVs (mpg)	27.5
CNGV	1.05
LNGV	1.05
LPGV	1.08
M90 vehicle	1.08
E90 vehicle	1.08
LS diesel CIDI vehicle	1.21
DME CIDI vehicle	1.21
FTD CIDI vehicle	1.21
BD20 CIDI vehicle	1.21
Gasoline HEV	1.43
LS diesel HEV	1.63
EV	3.50
H_2 FCV	2.36
Methanol FCV	1.60
Gasoline FCV	1.50

[a] Abbreviations: GV, gasoline vehicle; CNGV, compressed natural gas vehicle; LNGV, liquefied natural gas vehicle; LPGV, liquefied petroleum gas vehicle; E90, blend of 90% ethanol/10% gasoline (by volume); LS, low sulfur; CIDI, compression ignition, direct injection; DME, dimethyl ether; BD20, blend of 20% biodiesel/80% petroleum diesel (by volume); EV, electric vehicle.

compression (for $G.H_2$), and H_2 liquefaction (for $L.H_2$). The WTP energy efficiency of 23% for $L.H_2$ from U.S. electricity implies that, if FCVs are to be powered with $L.H_2$ produced this way, they need to achieve vehicle efficiencies at least four times as high as conventional GVs in order to have equal energy effects—a formidable challenge.

8.4 Well-to-Wheels Total Energy and Fossil Energy Results

The combination of well-to-pump results and the energy use and emissions associated with vehicle operation (also called the pump-to-wheels stage) results in full well-to-wheels energy and emission results for vehicle/fuel systems. The energy efficiencies of 19 fuel pathways were presented in the preceding section. The well-to-wheels energy use and emissions results for 23 representative vehicle/fuel systems, combined with some of the vehicle technology options listed in Table I, are now considered.

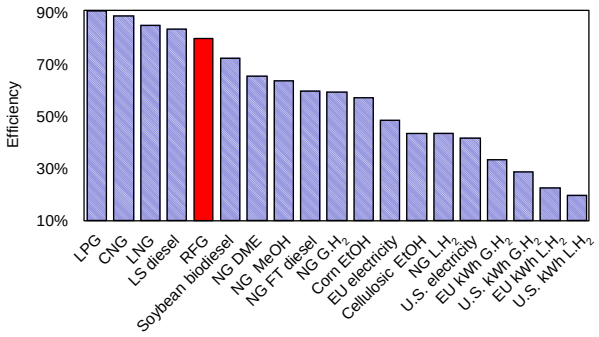


FIGURE 4 Well-to-pump energy efficiencies of fuel production pathways. LPG, Liquefied petroleum gas; CNG, compressed natural gas; LNG, liquefied natural gas; LS diesel, low-sulfur diesel; RFG, reformulated gasoline; NG DME, natural gas (dimethyl ether); NG MeOH, natural gas (methanol); NG FT diesel, natural gas (Fischer–Tropsch diesel); NG $G.H_2$, natural gas (gaseous hydrogen); corn EtOH, corn-derived ethanol; EU, European Union; cellulosic EtOH, cellulosic-derived ethanol; NG $L.H_2$, natural gas (liquid hydrogen).

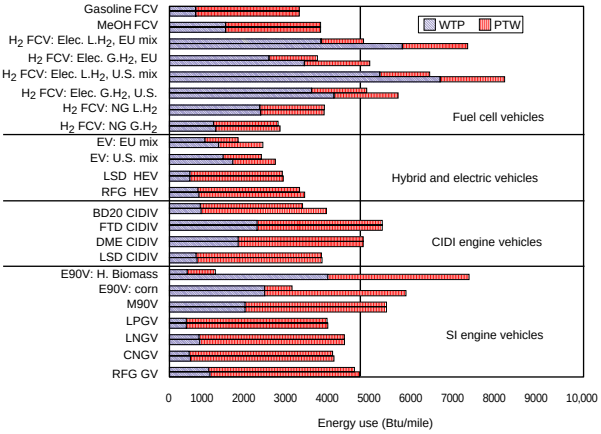


FIGURE 5 Well-to-wheels total energy use and fossil energy use of vehicle/fuel systems. WTP, Well to pump; PTW, pump to wheels; FCV, fuel cell vehicle; LSD, low-sulfur diesel; HEV, hybrid electric vehicle; BD20, blend of 20% biodiesel/80% petroleum (by volume); CIDIV, compression-ignition, direct-injection vehicle; SI, spark ignition; FTD, Fischer–Tropsch diesel; E90 V, vehicles fueled by 90% ethanol/10% gasoline (by volume); M90 V, vehicles fueled by 90% methanol/10% gasoline (by volume); GV, gasoline vehicle; see legend to Fig. 4 for other abbreviations.

Figure 5 shows WTW results for total energy use and fossil energy use. Total energy use includes both fossil and renewable energy sources. Fossil energy use includes three fossil energy sources, petroleum, NG, and coal. In this and the following six figures, the 23 vehicle/fuel systems are separated into four groups by vehicle technologies: SI engine vehicles, CIDI engine vehicles, HEVs and battery-powered EVs, and FCVs. For each system, the bottom bar represents total energy use and the top bar represents

fossil energy use. Also, for each bar, the results are separated into well-to-pump and pump-to-wheels components.

For total energy use, of the seven SI engine vehicle options, use of methanol and ethanol could increase total energy use relative to baseline RFG vehicles. The increases are caused by high amounts of energy use during the WTP stage. On the other hand, use of CNG, LNG, and LPG results in reduced total energy use, because of lower WTP energy use. Among the four CIDI vehicle options, use of DME, and especially FT diesel, could lead to increases in total energy use. But, because of high CIDI engine efficiencies, use of LS diesel and a blend of LS diesel and biodiesel (BD20; a blend of 20% biodiesel and 80% LS diesel) results in reduced total energy use. All four HEV and EV options reduce total energy use, because HEVs and EVs have much higher vehicle efficiencies compared to vehicles with SI engines (see Table V). The eight FCV options have diversified WTW energy results. The four electrolysis-based H_2 pathways actually increase total energy use, even though FCVs are much more efficient compared to SI engine vehicles. This is because electrolysis-based H_2 pathways have much lower efficiencies than the baseline RFG fuel pathway (Fig. 4) has. The two NG-based H_2 pathways and the two hydrocarbon fuels (gasoline and methanol) achieve energy benefits.

For fossil-energy-based fuels (i.e., fuels produced from petroleum and gasoline), WTW fossil energy results are similar to total energy results. However, renewable fuels (ethanol and biodiesel) achieve great fossil fuel benefits, even though they have potential total energy use penalties. Also, H_2 produced from both EU and U.S. electricity has much smaller increases in fossil fuel use because a significant portion of electricity is generated from nonfossil energy sources.

8.5 Well-to-Wheels GHG Emissions

Figure 6 shows WTW GHG emissions of the 23 vehicle/fuel systems. For each system, the bottom bar represents CO_2-equivalent emissions of CO_2, CH_4, and N_2O, combined with their global warming potentials (GWPs). The top bar represents CO_2 emissions only. For the two ethanol pathways (corn and cellulosic ethanol), there are some negative emissions. They represent carbon uptake during biomass growth. The carbon is eventually emitted to the air during ethanol combustion. Thus, the net results from the two ethanol pathways are the difference between the bars on the right of the axis

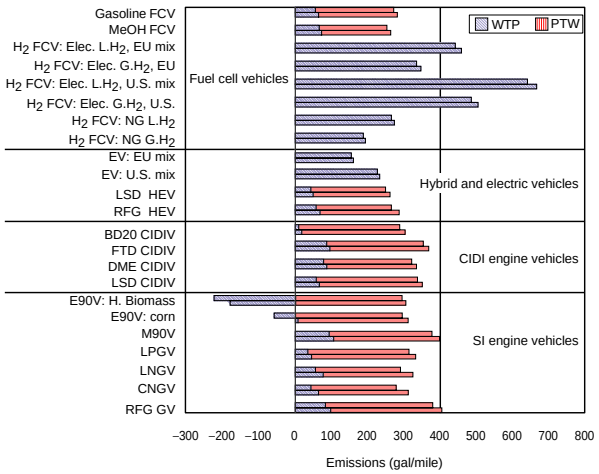


FIGURE 6 Well-to-wheels greenhouse gases and CO_2 emissions of vehicle/fuel systems; see Figs. 4 and 5 for abbreviations.

and the bars on the left. SI engine vehicles fueled with cellulosic ethanol have the lowest GHG emissions because a large amount of carbon is sequestered during biomass growth. Furthermore, biomass, not fossil fuels, is used in cellulosic ethanol plants to provide heat and steam for plant operations.

The group with the next lowest GHG emissions includes battery-powered EVs and FCVs fueled with $G.H_2$ made from NG. In the EV case, the very high efficiency overwhelmingly outweighs the low efficiency of electricity generation. In addition, some of the electricity used by EVs is generated from zero-emission sources such as nuclear power and hydropower. In the FCV case, the vehicle efficiency outweighs the inefficiency of $G.H_2$ production and compression. The third group with low GHG emissions includes HEVs, FCVs fueled with $L.H_2$ from NG, FCVs fueled with methanol and gasoline, and CIDI vehicles fueled with BD20. These vehicle/fuel systems achieve moderate GHG emissions reductions either because of efficiency gains of the vehicle technologies or because of reduced WTP GHG emissions for fuel production and distribution. The fourth group achieves some GHG emission reductions. This group includes SI engine vehicles fueled with CNG, LNG, LPG, M90 (a blend of 90% methanol and 10% gasoline, by volume), and E90 (a blend of 90% ethanol and 10% gasoline, by volume) with ethanol from corn; CIDI vehicles fueled with LS diesel, DME, and FT diesel; and FCVs fueled with $G.H_2$ from EU electricity. The last group includes FCVs fueled with $L.H_2$ from EU and U.S. electricity and $G.H_2$ from U.S. electricity. These vehicle/fuel systems actually increase GHG emissions because of

the high energy losses during electricity generation, H_2 production, H_2 compression (for gaseous H_2), and H_2 liquefaction (for liquid H_2).

For most vehicle/fuel systems, CO_2 emissions are slightly lower than GHG emissions, because CH_4 and N_2O emissions are excluded in the former. In a few cases, such as with ethanol and NG-based fuels, the differences between CO_2 and GHG emissions could be significant. In these cases, exclusion of CH_4 and N_2O emissions could give the evaluated fuel options unwarranted benefits.

8.6 Well-to-Wheels VOC Emissions

Figure 7 shows total VOC emissions of the evaluated vehicle/fuel systems. For each technology option, the bottom bar represents total VOC emissions (emissions that occur everywhere), whereas the top bar represents urban emissions (emissions that occur within urban areas). Urban emissions, a subset of the total emissions, are determined by information on locations of facilities within the United States. For vehicle operation emissions, urban emissions are estimated by using the vehicle distance traveled within U.S. urban areas. The separation of criteria pollutant emissions is intended to provide information for potential human exposure to the air pollution caused by these pollutants. Because the major concern with criteria pollutants is human health effects, the results of urban emissions are discussed here, even though there are large differences between total and urban emissions for many technology options.

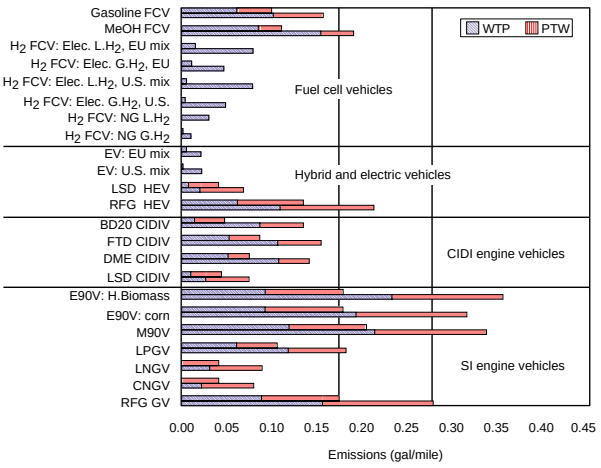


FIGURE 7 Well-to-wheels total and urban volatile organic compound emissions of vehicle/fuel systems; see Figs. 4 and 5 for abbreviations.

Of the seven SI engine technology options, CNGVs, LNGVs, and LPGVs achieve large urban VOC emissions reductions. These fuels generally have lower vehicle emissions. The E90 vehicles have VOC emissions equivalent to those of baseline GVs; M90 vehicles have a little higher VOC emissions, mainly because of high emissions from WTP stages. The four CIDI engine vehicle options have much lower urban VOC emissions compared to baseline GVs. Of the two HEV options, RFG HEVs have somewhat lower VOC emissions. Low-sulfur diesel HEVs (equipped with CIDI engines) have much lower VOC emissions. EVs and FCVs fueled by H_2 have the lowest urban VOC emissions: these vehicles are zero-emission vehicles. WTP urban VOC emissions from these options are usually small. FCVs fueled with methanol and gasoline do not achieve VOC emissions reductions as high as those offered by EVs and H_2-powered FCVs, although they offer considerable emissions reductions. This is because WTP urban VOC emissions for methanol and gasoline are relatively large, and production of H_2 from methanol and gasoline used to fuel FCVs generates VOC emissions.

8.6 Well-to-Wheels CO Emissions

Figure 8 presents the WTW CO emissions of the evaluated systems. Internal combustion engines produce a large quantity of CO emissions. Consequently, the technologies powered by these engines have high CO emissions. Also, the two FCV options with onboard fuel processors produce a significant amount of CO emissions. EVs and direct H_2 FCVs

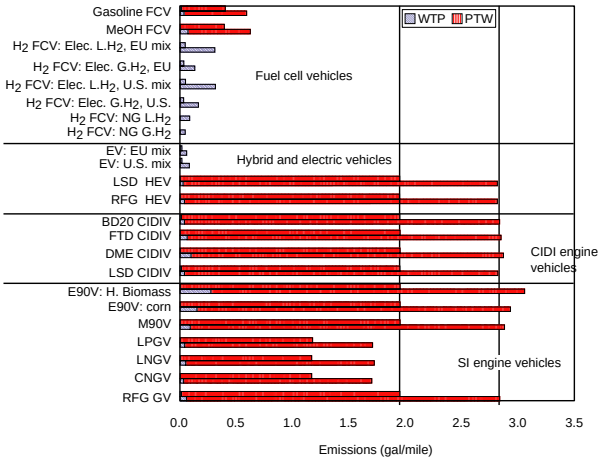


FIGURE 8 Well-to-wheels total and urban CO emissions of vehicle/fuel systems; see Figs. 4 and 5 for abbreviations.

have much lower CO emissions than the baseline GVs do.

8.7 Well-to-Wheels NO_x Emissions

Figure 9 shows NO_x emissions for the 23 vehicle/fuel systems. The differences between total and urban NO_x emissions for all technology options are large. Of the seven SI engine vehicle options, all options fueled by alternative fuels achieve significant urban NO_x emissions reductions, primarily because of low WTP urban NO_x emissions for these fuels. This is due to the plan to locate production facilities for alternative fuels outside of U.S. urban areas, in contrast to the siting of a large proportion of U.S. petroleum refineries within U.S. urban areas.

The five CIDI engine technology options and the two EV options (including LS diesel HEVs) offer small or no urban NO_x emissions reduction benefits. These technology options have high vehicle emissions. Their lower WTP NO_x emissions are not enough to offset high tailpipe NO_x emissions to achieve large WTW NO_x emissions reduction benefits. In the two EV cases, because some electric power plants are located in urban areas, NO_x emissions from these power plants cause urban NO_x emissions similar to those of baseline GVs.

FCVs fueled with H_2 from NG and with methanol and gasoline have lower urban NO_x emissions as compared to baseline GVs. The nonzero urban NO_x emissions for these vehicle options are caused primarily by their WTP urban NO_x emissions. The four electrolysis H_2 production options with EU and U.S. electricity actually increase urban NO_x emissions, even though FCVs do not generate any NO_x emissions.

8.8 Well-to-Wheels PM_{10} Emissions

Figure 10 shows PM_{10} emissions for the evaluated vehicle/fuel systems. The majority of urban PM_{10} emissions are from vehicle operations. Of these, a large proportion is from brake and tire wear. This will be especially true when future stricter tailpipe PM_{10} emission standards take effect. On the other hand, for total PM_{10} emissions, the technologies relying on electricity as the ultimate fuel have much higher PM_{10} emissions because of the high PM_{10} emissions from coal-fired power plants.

8.9 Well-to-Wheels SO_x Emissions

Figure 11 presents SO_x emissions of the evaluated systems. Because SO_x emissions are a regional as well as local air pollutant, both total and urban SO_x

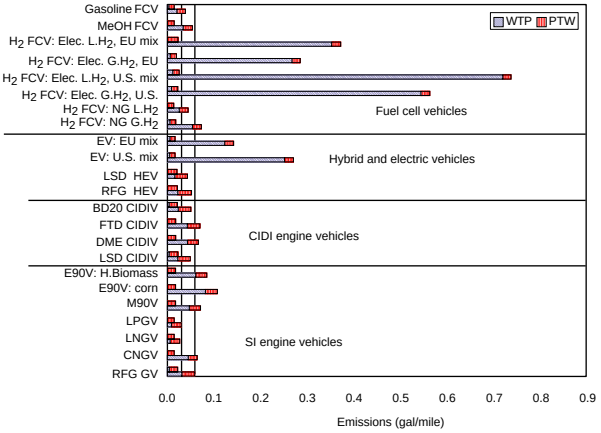


FIGURE 10 Well-to-wheels total and urban PM_{10} emissions of vehicle/fuel systems; see Figs. 4 and 5 for abbreviations.

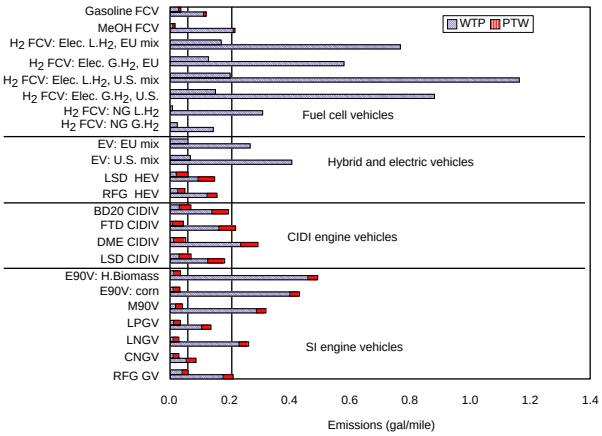


FIGURE 9 Well-to-wheels total and urban NO_x emissions of vehicle/fuel systems; see Figs. 4 and 5 for abbreviations.

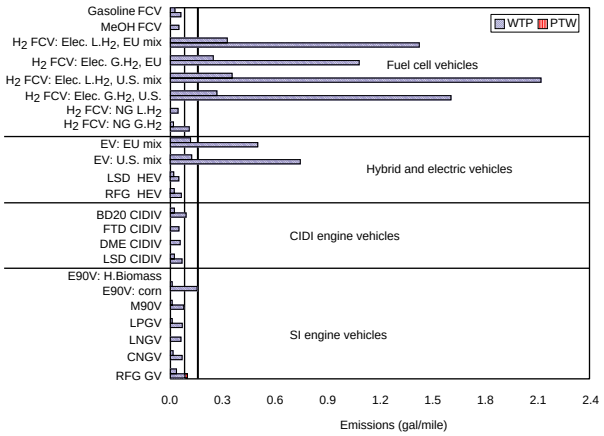


FIGURE 11 Well-to-wheels total and urban SO_x emissions of vehicle/fuel systems; see Figs. 4 and 5 for abbreviations.

emissions could be important. All the alternative vehicle/fuel systems achieve SO_x emissions reduction benefits because RFG (used in the baseline vehicle) has a higher sulfur content than do other fuels, resulting in high tailpipe SO_x emissions. Also, petroleum refineries generate considerable SO_x emissions.

All the options that rely on electricity generate very high total SO_x emissions. Corn-based ethanol and FCVs fueled with $G.H_2$ from NG generate somewhat high SO_x emissions. In the case of corn ethanol, this is because a significant amount of steam in ethanol plants is generated from coal combustion.

9. CONCLUSIONS

To evaluate the energy and emissions impacts of advanced vehicle technologies and new transportation fuels fully and accurately, researchers must conduct well-to-wheels analyses for two reasons: (1) fuel production pathways can be significantly different for new transportation fuels, resulting in different WTP energy and emission effects, and (2) vehicle tailpipe emissions, especially emissions of criteria pollutants, continue to be controlled and could become a smaller portion of the WTW emissions associated with vehicle/fuel systems.

In general, advanced vehicle/fuel systems achieve energy and emissions benefits relative to the benchmark gasoline vehicle. However, certain fuel pathways must be avoided to achieve the energy and emission benefits of some vehicle/fuel systems. Emission controls for criteria pollutants need to be considered for well-to-pump activities, as well as pump-to-wheels stages, so that advanced vehicle/fuel systems can help solve urban air pollution problems.

SEE ALSO THE FOLLOWING ARTICLES

Alternative Transportation Fuels: Contemporary Case Studies • Biodiesel Fuels • Ethanol Fuel • Fuel Cell Vehicles • Fuel Economy Initiatives: International Comparisons • Hybrid Electric Vehicles • Internal Combustion Engine Vehicles • Transportation Fuel Alternatives for Highway Vehicles • Vehicles and Their Powerplants: Energy Use and Efficiency

Further Reading

Arthur D. Little. (2002). "Guidance for Transportation Technologies: Fuel Choice for Fuel Cell Vehicles." Final report, Cambridge, Massachusetts, prepared for the U.S. Department of Energy, Feb. 6. Arthur D. Little, Inc., Cambridge, Massachusetts.

Delucchi, M. A. (2001). A life-cycle emissions analysis: Urban air pollutants and greenhouse gases from petroleum, natural gas, LPG, and other fuels for highway vehicles, forklifts, and household heating in the U.S. *World Resource Rev.* **13**(1), 25–51.

Energy Information Administration (EIA). (1998). "International Energy Annual 1996." DOE/EIA-0219(96), Energy Information Administration, Office of Integrated Analysis and Forecasting. U.S. Department of Energy, Washington, D.C.

Intergovernmental Panel on Climate Change (IPCC). (2001). "Climate Change 2001: The Scientific Basis, Technical Summary of the Group I Report." Cambridge University Press, Cambridge, U.K.

International Standard Organization (ISO). (1997). "ISO 14040, Environmental Management, Life Cycle Assessment—Principles and Framework." ISO, Barrow, U.K.

Ishitani, H., Baba, Y., and Kobayashi, O. (2000). Well-to-wheel energy efficiency of fuel cell electric vehicles by various fuels. [Translated from the original paper published in Japanese in *Journal of Energy and Resources*.]

Kreucher, W. M. (1998). "Economic, Environmental and Energy Life-Cycle Inventory of Automotive Fuels." SAE technical paper 982218. Society of Automotive Engineers, Warrendale, Pennsylvania.

Louis, J. J. J. (2001). "Well-to-Wheel Energy Use and Greenhouse Gas Emissions for Various Vehicle Technologies." SAE technical paper 2001-01-1343. Society of Automotive Engineers, Warrendale, Pennsylvania.

$(S\&T)^2$ Consultants, Inc. (2000). "Assessment of Emissions of Greenhouse Gases from Fuel Cell Vehicles." Prepared for Methanex Corporation, June 5. $(S\&T)^2$ Consultants, Inc., Delta, British Columbia, Canada.

U.S. Environmental Protection Agency (EPA). (2002). "Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990–2000." EPA 430-R-02-003, U.S. Environmental Protection Agency, Office of Atmospheric Programs. EPA, Washington, D.C.

Wang, M. (1999). "GREET 1.5—Transportation Fuel-Cycle Model, Volume 1: Methodology, Development, Use, and Results" ANL/ESD-39, Vol. 1, Center for Transportation Research. Argonne National Laboratory, Argonne, Illinois.

Wang, M., and Huang, H. (1999). "A Full Fuel-Cycle Analysis of Energy and Emissions Impacts of Transportation Fuels Produced from Natural Gas." Center for Transportation Research. Argonne National Laboratory, Argonne, Illinois.

Wang, M. Q., Saricks, C. L., and Wu, M. (1997). "Fuel-Cycle Fossil Energy Use and Greenhouse Gas Emissions of Fuel Ethanol Produced from U.S. Midwest Corn." Prepared for the Illinois Department of Commerce and Community Affairs by the Center for Transportation Research. Argonne National Laboratory, Argonne, Illinois.

Wang, M., Saricks, C. L., and Santini, D. (1999). "Effects of Fuel Ethanol Use on Fuel-Cycle Energy and Greenhouse Gas Emissions." ANL/ESD-38. Center for Transportation Research. Argonne National Laboratory, Argonne, Illinois.

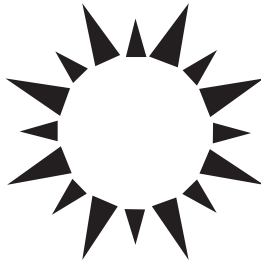
Wang, M. Q., Saricks, C. L., and Wu, M. (1999). Fuel ethanol produced from U.S. Midwest corn: Help or hindrance to the vision of Kyoto? *J. Air Waste Manage. Assoc.* **49**, 756–772.

Wang, M., Saricks, C. L., and Santini, D. (1999). Greenhouse gas emissions of fuel ethanol produced from corn and cellulosic biomass. *Environ. Mgr.*, October, pp. 17–25.

Weiss, M. A., Heywood, J. B., Drake, E. M., Schafer, A., and AnYeung, F. F. (2000). "On the Road in 2020, a Life-Cycle Analysis of New Automobile Technologies." Energy Laboratory Report # MIT EL 00-003. Massachusetts Institute of Technology, Cambridge, Massachusetts.

Winebrake, J., He, D., and Wang, M. (2000). "Fuel-Cycle Emissions for Conventional and Alternative Fuel Vehicles: An Assessment of Air Toxics." ANL/ESD-44, Center for Transportation Research. Argonne National Laboratory, Argonne, Illinois.

Fuel Economy Initiatives: International Comparisons



STEVEN PLOTKIN
Center for Transportation Research, Argonne National Laboratory
Washington, D.C., United States

1. Overview
2. Fuel Economy Standards
3. Fuel Taxes
4. Vehicle Tax Incentives Promoting Higher Fuel Economy
5. Information Programs to Promote Higher Fuel Economy
6. Government Sponsorship of Research and Development

Glossary

driving cycle A carefully measured trip that is defined by the vehicle speed during every second of the trip over a specified distance; using a driving cycle allows different vehicles to be tested and compared in a consistent manner.

fuel economy A measure of vehicle fuel efficiency expressed as distance traveled divided by fuel consumed (e.g., miles per gallon, kilometers per liter).

hybrid drivetrains (hybrid vehicles) Generally refers to vehicles that combine a conventional fuel-driven engine with an electric drivetrain, that is, a battery (or other means of storing electricity) and a motor/controller.

light-duty vehicles Automobiles, compact passenger vans, sport utility vehicles, and pickup trucks; in the United States, statutes regulating light-duty vehicles generally restrict these to vehicles less than 8500 pounds "gross vehicle weight" (i.e., curb weight plus carrying capacity of passengers plus cargo).

top runner A means of setting vehicle efficiency, emissions, or other standards by identifying representative vehicles that excel in the characteristic being targeted and using their performance as the basis for setting standards for an entire class of vehicles.

This article describes initiatives aimed at improving the fuel economy of light-duty passenger vehicles—automobiles and light trucks (sport utility vehicles, pickup trucks, and small passenger vans). The article's primary focus is on direct measures that create fuel economy targets for national and/or corporate fleets of new vehicles.

1. OVERVIEW

Reducing the growth of fuel consumption and greenhouse emissions from light-duty vehicles (LDVs) is a critical component of attempts to prevent global warming and to reduce the vulnerability of national economies to the adverse effects of oil price volatility. In the United States, where travel by LDVs is most intensive, LDVs consume nearly 8 million barrels of oil per day, or 40% of U.S. oil consumption, and account for nearly 17% of U.S. emissions of carbon dioxide. According to the U.S. Department of Energy's Energy Information Administration, energy use and greenhouse emissions from LDVs in the United States are expected to grow by nearly 2% per year over the next few decades. Although the rest of the world is, on average, far behind the United States in vehicle ownership and use (nearly 800 vehicles/1000 population in the United States vs $\sim$120 vehicles/1000 population for the world average and $\sim$30 vehicles/1000 population for the developing world), growth rates of both vehicle ownership and fuel use are much higher outside the United States and especially in the developing world. It is unlikely that U.S. levels of vehicle ownership and per capita personal travel will become the global norm, but it appears inevitable that vehicle ownership and personal travel will continue to grow rapidly throughout the world.

Restraining the growth of greenhouse emissions and oil use of LDVs can be accomplished by improving vehicle efficiency (fuel economy), substituting nonpetroleum fuels (with lower lifecycle greenhouse emissions), reducing travel demand

 © 2004 Elsevier Inc. All rights reserved.

(e.g., by raising the cost of travel or promoting compact development), and promoting shifts in travel to other modes that are more efficient and/or use nonpetroleum fuels (e.g., rapid transit). All of these approaches are worth pursuing.

In addition to reducing oil consumption and greenhouse emissions, improvements in fuel economy will affect emissions that directly affect human and ecosystem health. The primary positive effect is through the reduced need to produce, refine, and distribute vehicle fuels, yielding a proportional reduction in so-called "upstream" emissions and fewer hydrocarbon emissions from vehicle refueling. However, the effect on vehicle tailpipe emissions is decidedly mixed and may well be negative. On the one hand, reductions in fuel consumption generally yield a proportional reduction in emissions from the engine. On the other hand, this may make little practical difference in actual tailpipe emissions if emissions regulations are written in terms of maximum allowed weight of emissions per unit of distance driven (as they are in the United States, Europe, and Japan). In this case, manufacturers are likely to use fewer emissions controls on those vehicles with inherently lower emissions, yielding the same emissions outcome—with the important caveat that, if vehicle emissions controls malfunction, the more efficient vehicles will then tend to have lower tailpipe emissions than their less efficient counterparts. A decidedly negative impact is associated with any increase in travel caused by the lower per mile fuel costs associated with improved efficiency (the so-called "rebound effect"); this will increase emissions. The overall net impact on health-related emissions of these several effects is not clear.

The three major markets that currently have active fuel economy initiatives—the United States, Japan, and the European Union (EU)—have approached the pursuit of higher levels of fuel economy in different ways. The United States has pursued what some consider to be a contradictory policy, combining an economic policy that maintains low fuel prices with a regulatory policy that demands attainment of fixed fuel economy levels. U.S. standards have remained at 27.5 mpg for passenger cars and between 20 and 21 mpg for light trucks since 1985 (although there is a pending proposal to increase the light truck standard to 22.2 mpg by model year 2007). At the same time, U.S. gasoline prices have been half or less than half the prices prevailing in Europe and Japan (due to much lower tax rates in the United States), and this has provided minimum economic incentive for U.S. motorists to seek out vehicles with high fuel economy. On the other hand, until recently neither Japan nor the EU had used formal standards to stimulate higher levels of fuel economy, apparently relying on high fuel prices (generally $3.00/gallon and up) to maintain efficient fleets. However, Japan recently introduced a set of "weight class standards" that require attainment of fuel economy targets in each of nine weight categories, with an estimated fleet fuel economy improvement (from 1995 levels) of about 23%—to approximately 35 mpg—by 2008. Also, the EU accepted a voluntary commitment from its vehicle manufacturers associations for a reduction in "per vehicle" CO_2 emissions of 25% percent—to 140 g CO_2/km—by 2010 (2009 for the Japanese and Korean associations), also from a 1995 baseline. If fuel composition does not change, this corresponds to a fuel economy improvement of 33%—to about 41 mpg.

In addition to fuel taxes and fuel economy standards, approaches to improving LDV fuel economy include

- vehicle purchase and ownership taxes, tax credits, and direct purchase subsidies rewarding high efficiency;
- government requirements for public access to information about fuel economy and fuel costs, allowing vehicle purchasers to appropriately incorporate these factors into their purchase decisions;
- government sponsorship of research and development (R&D) on higher efficiency vehicles;
- government requirements for fleet purchase of high-efficiency vehicles.

Although this article does not address regulations controlling health-related vehicle emissions, many developing nations are beginning to adopt emissions standards based on those of the industrialized nations, especially recent EU standards. Complying with these standards may require adoption of improved engine controls and fuel injection, both of which should yield improved fuel efficiency.

2. *FUEL ECONOMY STANDARDS*

2.1 Overview

The most direct way in which to improve LDV fuel economy is to establish clear targets for new vehicle fuel economy or CO_2 emissions (although the latter

can theoretically be attained partly, or even solely, by switching to lower carbon fuels). In 1974, U.S. automakers voluntarily agreed to achieve a 40% improvement in fuel economy by 1980. This voluntary agreement was superseded by mandatory standards promulgated in 1975, aiming for still higher gains by 1985. Canada followed the U.S. example with a set of voluntary fuel economy targets mimicking the United States' mandatory ones. Within the past few years, both the EU and Japan have established new targets, the EU through a voluntary commitment with its automobile manufacturers associations and Japan through a set of mandatory fuel economy targets based on vehicle weight classes.

Setting fuel economy or CO_2 emissions targets for future vehicles or fleets raises a number of important concerns. Some of the key concerns include the following:

- *Economic efficiency and impacts.* Many in the environmental community have championed new fuel economy standards because they appear to guarantee results and perhaps because the U.S. standards appear to have been quite successful. However, economists have generally not supported fuel economy standards because they fear that such standards may force manufacturers to produce vehicles that consumers will not want, they do not address the demand for vehicle travel (and actually will tend to boost travel by reducing fuel costs, thereby offsetting some of the fuel savings associated with the increased efficiency—the rebound effect), and they inevitably create strong incentives for manufacturers to "game play" to take advantage of quirks in the regulations. A number of studies have compared fuel economy standards with other means of reducing oil consumption and greenhouse gas emissions, and others have examined the economic impacts of new standards, particularly the impact on employment. A thorough examination of these studies would require a separate discussion, but they can be summarized with extreme brevity as follows. First, standards generally are less efficient than direct economic incentives, in particular fuel taxes, which will suppress travel as well as stimulate higher fuel economy. However, there is little agreement about the magnitude of the difference in economic efficiency among the different approaches due to the wide range in estimates of the fuel price sensitivity of travel and fuel economy and similar disagreements about the actual costs of new standards. Second, employment impact studies sponsored by the vehicle manufacturers generally conclude that new standards will reduce employment, whereas studies sponsored by the environmental community conclude that new standards will increase employment. These differences are driven primarily by very different underlying assumptions about the success and costs of new standards as well as about the approach that manufacturers will take to comply with the standards.
- *Selection of targets and assumptions about future technological and market development.* Unless fuel economy targets are quite modest and short term, the process of developing reasonable standards requires a set of assumptions about how automotive technology will change within the time frame of the standards, how consumer tastes will change (or not change), and how manufacturer market behavior will change (e.g., whether companies will go after new markets, whether companies will produce new classes of vehicles). Analyses leading to new standards have tended to avoid these questions by assuming little or no change, for example, no change in vehicle attributes such as weight and power (except possibly to address existing emissions and safety standards). All aspects of the vehicle marketplace have changed substantially over time, with strong impacts on fuel economy improvement potential.

Those who might wish to apply any of the existing standards to their own markets, or who simply wish to understand how ambitious the targets are, should pay careful attention to a variety of factors that affect the fuel economy potential of a particular market. The more important factors include the following:

- *Fuel taxes.* All else equal, higher fuel taxes and prices will mean that vehicle purchasers will have greater incentive to buy more fuel-efficient vehicles and manufacturers will have greater incentive to make them. Although raising fuel prices may be a desirable strategy for increasing fleet fuel efficiency (and reducing vehicle miles driven), higher fuel taxes might not be politically viable. In any case, attaining a given fuel economy target should generally be easier in markets with high fuel prices.
- *Driving cycles and conditions.* The driving cycle over which fuel economy is measured can strongly influence the performance of fuel economy technologies such as improved engines and use of lightweight materials. In general, regulators try to measure fuel economy on a driving cycle that corresponds well to actual driving conditions in their market. For example, Japanese driving conditions are best represented by congested city driving, and

the official Japanese "10.15" driving cycle used to measure fuel economy for regulatory purposes is a slow, stop-and-go cycle. (10.15 refers to the two sets of speed profiles in the cycle: one with 10 and the other with 15 separate driving modes—acceleration, deceleration, or constant speed driving. The 10-mode portion of the cycle is driven three times, followed by one 15-mode portion.) The U.S. regulatory cycle uses much faster speeds, with the European cycle somewhat in between, and these cycles represent average driving conditions in their markets reasonably well. One effect of these differences is to reward certain key technologies that focus on improving "low-load" efficiency—variable valve timing, hybrid drivetrains, idle–off, and the like—with significantly higher fuel economy "boosts" on the Japanese cycle than on the U.S. cycle. In contrast, improvements in aerodynamics are well rewarded on the U.S. cycle, and are nearly irrelevant on the Japanese cycle, because aerodynamic drag forces are much more important at high speeds than at low speeds. A critical effect of differences in driving cycles among the different markets is that the fleet fuel economy value and targets in each market are not directly comparable to those in the other markets. In general, if the current fuel economy levels in each market were to be measured on the U.S. driving cycle, the European fleet would have its measured fuel economy levels raised a bit (from their value on the European cycle) and the Japanese fleet would be raised even more. However, the percentage changes needed to "normalize" the fleets to a single driving cycle will change over time as technology changes.

- *Emissions regulations.* Emissions regulations can affect the performance and cost of key engine technologies and in extreme circumstances can prevent certain technologies from being used. Currently, emissions standards in the three markets with fuel economy standards are strictest in the United States and most lenient in Japan. Unless excellent progress is made in lean burn NO_x control technology, strict U.S. NO_x standards might serve to ban or significantly limit diesel use in the LDV fleet and reduce the fuel economy performance of direct injection gasoline engines. In contrast, European and Japanese standards allow for such "lean technologies" and, moreover, do not hold diesel-fueled vehicles to NO_x emissions levels as stringent as those for gasoline-fueled vehicles.
- *Vehicle taxes.* High vehicle taxes can serve to make new fuel economy technologies more expensive because they apply to the price of the total vehicle. For example, EU countries have "value-added" taxes that apply to the total price of the vehicle. On the other hand, countries may impose taxes on vehicle sales and/or ownership that vary with fuel efficiency or with factors that are likely to affect fuel efficiency, especially engine power or displacement, and these taxes would provide a positive incentive for high-efficiency vehicles. Several EU countries and Japan have such taxes. Also, some countries may give tax credits or deductions for certain kinds of high-technology vehicles, increasing the cost-effectiveness of these vehicles and making it somewhat easier for fleets to achieve high fuel economy. For example, the United States gives tax deductions to purchasers of hybrid electric vehicles.
- *Vehicle miles driven.* More driving means higher annual fuel savings from new technologies, thereby improving the cost-effectiveness of these technologies. For example, LDVs are driven farther in the United States ($\sim$13,000 miles/year for the first 5 years) than in Europe (8000–9000 miles/year), and much farther than in Japan ($\sim$6000 miles). This tends to raise the cost-effectiveness of new technologies in the United States versus that in the other markets.
- *Average vehicle efficiency.* In considering how high a percentage increase in current fuel economy might be sought, a high current fuel economy will mean that a particular percentage increase will save less fuel than would be saved if current fuel economy was low. In other words, a 20% increase would save more fuel in the United States than in Europe or Japan because the current U.S. fuel economy level is lower than in the other markets.

2.2 U.S. Corporate Average Fuel Economy Standards

In the United States, all major automakers are subject to Corporate Average Fuel Economy (CAFE) standards requiring their new vehicle fleets to meet specific fuel economy targets. The standards were established by the Energy Policy and Conservation Act of 1975 in response to the 1973 oil embargo and subsequent oil price shock. Under the act, each manufacturer's fleet of new vehicles is subdivided into passenger cars and light trucks (including pickup trucks, vans, and sport utility vehicles) and the passenger car fleet is further subdivided into domestic and import vehicles, where vehicles with less than 75% U.S./Canadian content are considered to be imports. Each individual fleet must then meet a separate target. In 2002, each manufacturer's domestic and import passenger car fleet had to average at

least 27.5 mpg and each manufacturer's light truck fleet had to average at least 20.7 mpg. Manufacturers that do not comply with these standards are liable for a civil penalty of $5.50 for each tenth of 1 mpg its fleet falls below the standard, multiplied by the number of vehicles in the noncomplying fleet. The U.S. government has enforced the standards; for example, in calendar year 2001, manufacturers paid civil penalties totaling more than $33 million for failing to comply with the standards in model years 1999 and 2000.

Additional important characteristics of the standards are as follows:

- *Low-volume manufacturers.* The standards can be waived for low-volume manufacturers, that is, those producing fewer than 10,000 passenger cars annually worldwide.
- *Vehicle weight limits.* The standards apply only to vehicles whose gross vehicle weight rating (GVWR) is 8500 pounds or less. A number of large sport utility vehicles have GVWRs above this level and are not subject to the standards.
- *Setting and amending standards.* The U.S. Congress set the original schedule for passenger car standards and authorized the National Highway Traffic Safety Administration (NHTSA), in the Department of Transportation, to set the light truck standards. Manufacturers were required by Congress to comply with the 27.5-mpg standard for passenger cars by 1985, and the standard has remained unchanged since (except for a brief reduction from 1986 to 1989). NHTSA is authorized to amend the passenger car and light truck standards.
- *Testing responsibility.* Manufacturers are responsible for measuring their own fleet fuel economy. They test preproduction prototypes of their new vehicle models and submit the results to the U.S. Environmental Protection Agency (EPA). EPA retests approximately 10% of the tested vehicles to confirm the manufacturer's tests.
- *Test procedure.* The procedure for testing vehicle fuel economy consists of having a professional driver drive the vehicle on a dynamometer through two driving cycles designed to simulate city and highway driving. Each vehicle is driven in the same manner on the same type of equipment, so that results can be compared with confidence. The city test begins with a cold start (as if the vehicle had been parked overnight) and simulates an 11-mile, stop-and-go trip at mostly low speeds (although the maximum speed is 56 mph), averaging 20 mph. The trip takes 31 min with 23 stops and a total of about 5.5 min of idle time. The highway test begins with the engine warmed up and simulates a 10-mile combined highway/rural road trip at an average speed of 48 mph and a maximum of 60 mph. The resulting two fuel economy values are averaged by assuming that 55% of driving is city driving and 45% is highway driving, with the average calculated by the following formula:

$$1/\text{vehicle mpg} = 0.55/(\text{city mpg}) + 0.45/(\text{highway mpg}).$$

Calculating CAFE. Each manufacturer's CAFE is the sales-weighted average of the tested fuel economy levels of its vehicles. This average is calculated in a manner similar to the preceding formula, that is, the inverse of the fuel economy level (the fuel consumption level) of every vehicle in the manufacturer's fleet is added, with the sum divided by the total number of vehicles. To clarify, if a manufacturer sold two types of vehicles with annual sales of 100,000 and 300,000, its fleet average fuel economy (CAFE) would be calculated using the following formula:

$$1/CAFE = [100,000/(\text{mpg of vehicle 1}) + 300,000/(\text{mpg of vehicle 2})] \div 400,000.$$

Credits. Manufacturers can earn credits if they exceed the standards, and these credits can be applied to any of the subsequent or preceding 3 years to offset shortfalls. Also, a manufacturer failing to meet the standard for a given model year can avoid paying the civil penalties by submitting a compliance plan to NHTSA demonstrating that it will earn enough credits within the next 3 model years to offset the penalties. Manufacturers also earn credits by producing and selling alternative fuel vehicles, with greater credits awarded for dedicated vehicles than for dual-fuel vehicles (which can burn either gasoline or an alternative fuel). Although each alternative-fuel vehicle obtains a high fuel economy value based on the theoretical amount of oil-based fuel it consumes (theoretical because dual-fuel vehicles may burn gasoline exclusively), total credits awarded to a manufacturer cannot increase its official CAFE value by more than 1.6 mpg.

- *Original schedule.* Although the passenger car CAFE targets have remained constant since 1990 and the light truck targets have remained constant since 1996, the original legislation was crafted to allow companies to raise their fleet fuel economy levels gradually, taking into account the impossibility of redesigning entire corporate fleets simultaneously as

well as the time needed to develop improved technology. Thus, for passenger cars, the 1978 standard was 18.0 mpg, with yearly increases thereafter (to 19.0 mpg in 1979, 20.0 in 1980, 22.0 in 1981, 24.0 in 1982, 26.0 in 1983, and 27.0 in 1984), leading to the 27.5-mpg target in 1985.

Although there have been arguments about whether increases in U.S. LDV fuel economy after 1975 were caused mostly by the CAFE regulations or were instead caused largely by market forces, the U.S. fleet has generally complied with the regulations and has maintained fuel economy values above the required levels despite low gasoline prices and shifts in consumer tastes toward larger, more powerful vehicles. However, since attaining the standards, U.S. fleet fuel economy has essentially stopped improving, and the combined car and truck fuel economy has actually declined due to increasing market shares of (less efficient) light trucks. Over the past two decades, the U.S. fleet has used a continuing influx of new technology to allow fleet fuel economy levels to be maintained while other factors—weight, acceleration time, luxury features, and the like—moved in directions that would otherwise have reduced fuel economy. EPA has estimated that the tested fuel economy (not adjusted downward for onroad conditions) of the model year 2001 fleet of cars and trucks, which was 23.9 mpg, would have instead been 30.3 mpg had the weight mix and 0-to-60 mph acceleration times of the fleet been the same as that of the 1981 model year. In other words, the fleet traded away between 6 and 7 mpg for higher weight and better acceleration during that 20-year period.

A number of controversies have arisen over the CAFE standards. Probably the most important is the claim that size and weight reductions demanded by the standards have had an adverse impact on overall vehicle safety. This claim has tended to dominate discussions about raising future CAFE standards. The National Research Council's (NRC) Committee on the Effectiveness and Impact of Corporate Average Fuel Economy Standards concluded in 2002 that the change in size and weight of the fleet between 1976 and 1993 caused about 2000 additional traffic fatalities in 1993. There was a strong dissent to the committee's conclusions from within the committee itself as well as from other analysts. Furthermore, an earlier national academy panel concluded that the available analyses showing vehicle weight/traffic fatality correlations should not be used to project fatalities caused by fleetwide weight changes because the analyses could not adequately control for various confounding factors such as driver age. The issue is complicated by the fact that fuel economy can be improved without weight reduction and that the distribution of weight within the fleet is likely to be more important to safety than the average weight. A reanalysis of traffic fatality data sponsored by Honda, using more recent data but the same analytical method cited by the NRC committee, found essentially no impact on traffic fatalities from a 100-pound reduction in the average weight of the fleet.

Manufacturer responses to some provisions of the standards have caused some controversy. For example, the division into domestic and import fleets has created incentives for manufacturers to alter the domestic "content" of individual models to move them into one fleet or the other depending on circumstances. The original purpose of the division was to create an incentive for manufacturers to produce small cars within the United States so as to boost domestic employment (manufacturers needed the smaller cars to provide a counterweight to larger, less efficient models that previously had been their major market focus for domestic production). The division succeeded in stimulating domestic small car production, but it also created an incentive to add import content to larger domestic models so as to get them out of the domestic fleet and into a fleet of more efficient import models that could more easily offset their low fuel economy. For example, Ford temporarily made its full-size (and relatively low-fuel economy) Crown Victoria model into an "import" to get it out of the domestic fleet, thereby aiding that fleet in complying with its CAFE standard.

In 2002, the State of California passed legislation (Assembly Bill 1493) that requires the California Air Resources Board to establish greenhouse gas standards for vehicles in model year 2009 and beyond, applying to corporate fleet averages. Although automakers will be able to partially comply with the standards by reducing greenhouse gas emissions from nonvehicle sources and by shifts in fuels, the standards will essentially have an effect similar to that of fuel economy standards. The bill has been challenged in the courts by the vehicle manufacturing industry on the basis that the standards will be fuel economy standards and that only the federal government can set such standards.

2.3 Japanese Weight Class Standards

The Japanese government has established a set of fuel economy standards for gasoline- and diesel-

powered light-duty passenger and commercial vehicles, with fuel economy targets based on vehicle weight classes. Table I shows the vehicle targets for gasoline- and diesel-powered passenger vehicles, as measured on the Japanese 10.15 driving cycle. The targets for gasoline-powered vehicles are to be met by 2010, and 2005 is the target year for diesel-powered vehicles. Figure 1 illustrates the percentage improvements required for each weight class of gasoline-powered vehicles as well as the market share for each class at the time the regulations wave created. The original regulations demanded that the targets be met by each automaker for each weight class; that is, automakers could not average across weight classes by balancing a less-than-target vehicle in one class with a better-than-target vehicle in another class. Revisions to the regulations in January 2001 allow automakers to accumulate credits in one weight class and use them in another class, with important limitations.

Assuming no change in vehicle mix, these targets imply a 22.8% improvement in gasoline passenger vehicle fuel economy (15.1 km/L in 2010 vs 12.3 km/L in 1995) and a 14.0% improvement in diesel passenger vehicle fuel economy (11.6 vs 10 km/L) compared with the 1995 fleet. In other words, compliance with

TABLE I

Japanese Weight Class Fuel Economy Standards for Passenger Vehicles

	Vehicle weight class (kg)								
	750	875	1000	1250	1500	1760	2000	2250	2500
A. Gasoline-Fueled Passenger Vehicles for 2010									
Vehicle curb weight (kg)	–702	703–827	828–1015	1016–1265	1266–1515	1516–1765	1766–2015	2016–2265	2266–
Vehicle curb weight (pounds)	1548	1550–1824	1826–2238	2240–2789	2791–3341	3343–3892	3894–4443	4445–4994	4997–
Fuel economy target (km/L)	21.2	18.8	17.9	16.0	13.0	10.5	8.9	7.8	6.4
Fuel economy target (mpg)	49.9	44.2	42.1	37.6	30.6	24.7	20.9	18.3	15.0
B. Diesel-Powered Passenger Vehicles for 2005									
Vehicle weight (kg)	–1015	1016–1265	1266–1515	1516–1765	1766–2015	2016–2265	2266–		
Vehicle weight (pounds)	–2238	2240–2789	2792–3341	3343–3892	3894–4443	4445–4994	4997		
Target standard (km/L)	18.9	16.2	13.2	11.9	10.8	9.8	8.7		
Target standard (mpg)	44.5	38.1	31.0	28.0	25.4	23.0	20.5		

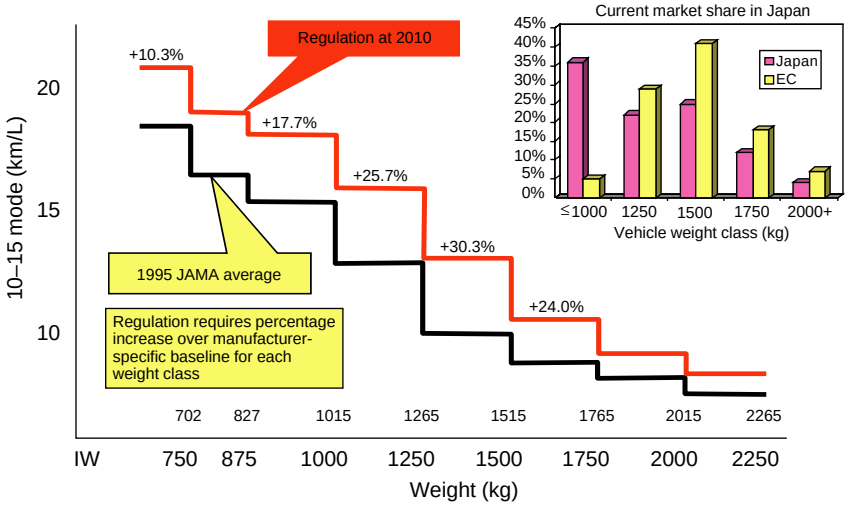


FIGURE 1 New Japanese fuel economy regulations. (Source: Japan Ministry of Transport.)

these standards will produce a Japanese gasoline-fueled, light-duty passenger vehicle fleet of 35.5 mpg and a diesel fleet of 27.3 mpg by 2010 and 2005, respectively, as measured using the Japanese 10.15 driving cycle (and assuming no changes in the market shares of the various weight classes). The Japanese 10.15 driving cycle is substantially slower than the combined U.S. city/highway cycle, and the U.S. equivalent miles per gallon for this fleet would be significantly higher.

The regulations call for civil penalties if the targets are not met, but these penalties are very small. The formal enforcement procedure includes a government assessment of whether the automakers have taken the necessary steps to comply with the targets, recommendations by the regulating agencies about steps they might take, and another assessment by the agencies of the adequacy of the automakers' compliance with the recommendations, with the procedures taking more than 1 year. Realistically, enforcement will come from pressure from the government and the auto companies' desire to avoid public embarrassment, not from the financial penalties.

The fuel economy targets were selected by identifying representative "best in class" fuel economies in each weight class and demanding that the average new vehicle meet that level in the target year. The Japanese call this the "top-runner" method of selecting fuel economy targets. The actual top-runner fuel economy target might not be the highest in the weight class. As described in the next subsection, some vehicles will not be candidates for top-runner status so as to avoid saddling the automakers with impractical targets based on exotic technology or nonrepresentative characteristics.

As noted previously, manufacturers may accumulate credits in a weight class by achieving levels of fuel economy higher than the target levels. These credits, measured as (difference in fuel consumption between the vehicle and the target fuel consumption rate) * (number of vehicles sold), can be applied to vehicles in another weight class at a 50% rate; that is, two credits are needed to offset one unit of underperformance. However, the credits can be applied only if the automakers' performance in the failed weight class meets the threshold values of fuel efficiency shown in Table II (for gasoline vehicles).

The fuel economy regulations have additional requirements over and above the actual fuel economy targets. These are as follows:

1. For new vehicles, fuel economy and major efficiency technologies onboard must be recorded in catalogs and displayed at exhibits.
2. Government is charged with providing education and other incentives for vehicle users and manufacturers, making sure that fuel economy regulation proceeds in harmony with other regulations (especially new emissions and fuel quality standards), reviewing manufacturers' efforts to improve fuel economy, and trying to harmonize this regulation with similar efforts in Europe and the United States.
3. Manufacturers are expected to develop new efficiency technologies, design vehicles of outstanding efficiency, and help to educate the public.
4. The public is charged with selecting efficient vehicles and using them in an efficient manner.

Although the Japanese standards do not take effect until 2010, some concerns have already arisen. First, changes in the weight composition of the Japanese fleet may change the degree of improvement in fuel economy demanded by the standards. Figure 2 shows how the weight distribution of the fleet has changed between 1995 and 2000. Primary changes include a large reduction in the lightest (750-kg) market segment and substantial increases in the heavier segments. If the year 2000 weight distribution is applied to the fuel economy standards, the expected improvement in fleet fuel economy over the 1995 baseline would be 18.0% versus the 22.8% percent projected by the Japanese government based on the 1995 weight distribution. Conceivably, future changes in the fleet's weight distribution might drive down the expected improvement even further.

TABLE II

Fuel Economy Thresholds for Use of Credits (Japanese Weight Class Standards)

Weight class (kg)	750	875	1000	1250	1500	1760	2000	2250	2500
Fuel economy threshold (km/L)	20.2	18.5	17.1	14.1	12.6	9.8	8	6.8	6.1
Target fuel economy (km/L)	21.2	18.8	17.9	16	13	10.5	8.9	7.8	6.4

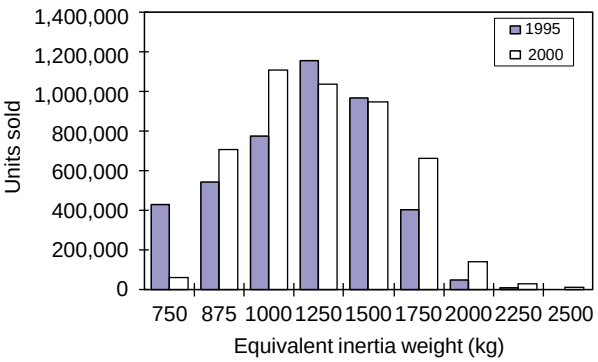


FIGURE 2 Weight distribution of the Japanese new LDV fleet, 1995–2000. (Source: Kiyoyuki Minato, Japan Automobile Research Institute.)

Another concern is the limitation on averaging fuel economy across weight classes. This allowed "50% averaging" still leaves manufacturers with far less flexibility than if full averaging were allowed. This lack of flexibility particularly affects European and U.S. manufacturers that export into the Japanese market and that are likely to have limited numbers of models, and possibly only one model, in a weight class.

2.3.1 The Japanese Top-Runner Selection Process

The new Japanese weight class standards for LDV fuel economy are based on identifying top runners—vehicles with exemplary fuel economy—in each weight class and requiring future vehicles in that class to attain similarly high fuel economy. According to Japanese industry sources, the top-runner selection process is more nuanced than simply selecting the highest fuel economy vehicle in each class and establishing that as the standard. Instead, the regulators—the Ministry of International Trade and Industry (MITI) and the Ministry of Transportation (MOT)—with the consultation of the Japanese Auto Manufacturers Association (JAMA), select top-runner targets based on their judgment of the relevance of class-leading vehicles and their technology to the class as a whole. For example, because the large majority of personal vehicles use automatic transmissions, only vehicles with automatic transmissions are considered as top-runner candidates for personal vehicles. For commercial vehicles, separate goals are established for vehicles with automatic and manual transmissions. Top-runner candidates that have very small sales volumes and/or very expensive technology are not likely to be considered because the regulators are interested in technologies that have been shown to be practical for the class as a whole. For example, the hybrid electric vehicles currently in the Japanese marketplace were not used as top runners.

2.4 ACEA Voluntary Commitment to the EC on Vehicle Carbon Emissions

The European Automobile Manufacturers Association (ACEA) has offered, and the European Commission (EC) has accepted, a voluntary commitment to reduce the CO_2 emissions from new light-duty passenger vehicles, with firm fleetwide targets of 140 g CO_2/km (~41 mpg for gasoline) by 2008, measured under the new European test cycle (Directives 93/116/EC and 99/100/EC). This represents approximately a 25% reduction from the 1995 average of 187 g/km (~30 mpg) on this cycle. The European cycle is likely to produce lower fuel economy ratings than the U.S. combined urban/highway cycle, so the "U.S. equivalent" miles per gallon ratings of the year 2008 European fleet will likely be higher than 41 mpg if the targets are met.

Note that the 140-g CO_2/km goal is a collective target, not a target for each individual company. The ACEA participants in the voluntary commitment—BMW, Fiat, Ford of Europe, GM Europe, DaimlerChrysler, Porsche, PSA Peugeot Citroen, Renault, Rolls-Royce, Volkswagen, and Volvo—have not publicly committed to individual objectives.

The voluntary commitment applies to light passenger vehicles classified as M1 in European Council Directive 93/116/EC, which includes vehicles with no more than eight seats in addition to the driver. The commitment includes a promise to introduce some models emitting 120 g/km (~48 mpg) or less by 2000 and a nonbinding 2003 target range of 165 to 170 g/km (~34–35 mpg). In addition, the commitment was to be reviewed in 2003 with the aim of moving toward a fleet goal of 120 g/km by 2012. Finally, ACEA agrees to monitor compliance with the commitment jointly with the EC.

In exchange for their commitment to meet the 2008 CO_2 emissions goal, industry members insisted that the voluntary commitment be contingent on a number of conditions, including the following:

- *Clean fuels availability.* Because the industry believes that direct injection engines will play a key role in achieving the targets, the commitment requires "full-market availability" of clean fuels needed for this technology by 2005—gasoline with 30 ppm sulfur content and less than 30%

aromatics, diesel fuel at 30 ppm sulfur and cetane number greater than or equal to 58. The low sulfur level is needed to avoid poisoning the catalysts needed for NO_X control.
- *Protection against unfair competition.* Non-ACEA members must commit to similar goals, and the European Community will agree to try to persuade other car manufacturing countries to embrace equivalent efforts. The latter is designed to protect ACEA members from suffering in world market competition for their European efforts.
- *Regulatory ceasefire.* There are no new regulatory measures to limit fuel consumption or CO_2 emissions.
- *Unhampered diffusion of technologies.* The companies assume that the EC will take no action that would hamper the diffusion of efficiency technologies, particularly direct injection gasoline and diesel engines. Presumably, this could include tighter emissions standards on NO_x and particulates. ACEA assumes that up to 90% of the new vehicle fleet will use direct injection engines by the 2008 compliance date, and tighter standards for these pollutants might interfere with the diffusion of this technology.

The voluntary commitment also incorporates a statement that might be seen as an escape hatch for the companies:

> *On the basis of the outcome of the monitoring, or if the impacts of this commitment on the European automotive industry, particularly its employment situation and its global competitive environment, are detrimental, ACEA and the [European] Commission will review the situation and make any necessary adjustments in good faith.*

Both JAMA and the Korean Automobile Manufacturers Association (KAMA) have agreed to a revised version of the ACEA targets—achievement of the 2008 target levels in 2009.

A concern with the voluntary commitment is that it does not specify individual corporate targets, nor does it define mechanisms to guarantee attainment of the fleet target. The flexibility afforded by the lack of specified individual targets theoretically could lead to an efficient market solution if companies could agree among themselves to attain levels of fuel economy improvement that represent approximately equal marginal net value (value of fuel savings minus technology cost). However, it is not clear that there is any agreement among the companies defining their individual targets, and there is no guarantee that all companies will make a fair effort toward meeting the fleet target. The lack of specified corporate targets might tempt companies to provide less than their full effort and might make it difficult for the EU to assess responsibility if the industry fails to meet the target.

3. FUEL TAXES

3.1 Overview

Economists have overwhelmingly favored fuel taxes over fuel economy standards as a means to reduce fuel consumption because taxes give maximum flexibility to both vehicle manufacturers and purchasers and because they influence fuel consumption through both fuel demand and vehicle supply by making travel more expensive (thereby reducing vehicle miles traveled) and by creating an economic incentive for manufacturers to build efficient vehicles and for consumers to purchase them. However, it is important to note that the presumed superior effectiveness of fuel taxes over CAFE standards depends on the assumption that markets for fuel economy and travel are very efficient. This assumption about an efficient market, especially as it pertains to the market for fuel economy, is debatable, and the interested reader is urged to examine the considerable literature on this subject.

Because analyses of future light-duty transportation energy use require estimates of the impact of fuel prices on travel and fleet fuel economy, estimates of the fuel price elasticity of travel, fuel economy, and fuel consumption are ubiquitous in the literature. However, there is substantial disagreement about the magnitude of these elasticities because travel volumes, fuel economy, and fuel consumption are dependent on several variables other than fuel price and because fuel prices have tended to be volatile during the past few decades, thereby complicating attempts to estimate long-run elasticities. Thus, the magnitude of the effect of changes in fuel taxes on fleet fuel economy and on travel volumes and fuel consumption is also subject to considerable disagreement. For the United States, recent studies by D. L. Greene and others imply that a 10% gasoline price increase would cause at most approximately a 2% increase in average fuel economy, a 2 or 3% decrease in travel, and a 4% decline in gasoline consumption (compared with a baseline, and increasing, level) over the long term. However, others conclude that the decline in gasoline consumption would be higher—as much as 6 or 8%. The effect elsewhere will be different. For example, in Europe and Japan, fuel prices make up a larger share of the total cost of

travel, implying that fuel price elasticities there will likely be higher than those in the United States.

3.2 United States

In the United States, motor vehicle fuel taxes have generally been viewed as "user taxes" at the federal and state levels. In other words, the federal government and most state governments set their per gallon tax rate at a level that will pay for a substantial share of road-building requirements. In addition, states and localities add sales taxes as a percentage of fuel prices. Federal taxes in 2000 were $0.184/gallon for gasoline and $0.244/gallon for diesel fuel. Gasohol, a blend of 90% gasoline and 10% ethanol by volume, was taxed at only $0.131/gallon. State taxes averaged about $0.20/gallon plus sales taxes, the latter of which vary between 1 and 7%. In addition, 12 states allow localities to add their own taxes.

The result is that taxes on gasoline and diesel fuel in the United States have been quite low in comparison with those in Europe and Japan. In 2000, the average U.S. gasoline price (less local taxes) was $1.48/gallon, of which 28% was federal and state taxes. The average total gasoline price in 2001 was $1.63/gallon, and the average diesel price was $1.52/gallon.

3.3 Japan and Europe

Europe and Japan have traditionally taxed gasoline at high rates, so that gasoline prices have been much higher than those in the United States. These higher prices, along with differences in land use patterns and other factors, play an important role in the strong differences in average per vehicle miles traveled among Europe, Japan, and the United States; over their first 5 years of life, LDVs are driven approximately 13,000 miles in the United States, 8000 miles in Europe, and 6000 miles in Japan.

In 2001, Japanese gasoline prices averaged $3.80/gallon as compared with $3.68/gallon in France, $3.74/gallon in Germany, and $4.63 in the United Kingdom. With the exception of the United Kingdom, diesel prices were substantially lower—$2.70/gallon in Japan, $2.80/gallon in France, and $2.91/gallon in Germany. U.K. diesel prices were virtually the same as gasoline prices—$4.42/gallon.

In Europe, the price differences among the various countries and the gasoline/diesel price difference are due nearly entirely to differences in taxation. Table III displays the gasoline and diesel taxes in the several European countries. As shown by the table, every European country but the United Kingdom taxes diesel fuel substantially less than it does gasoline on a volumetric basis despite diesel fuel's higher volumetric energy content. This differential has stimulated sales of diesel-powered vehicles, which have soared in Western Europe—to nearly 40% of new vehicle sales during the first 6 months of 2002.

TABLE III

European Taxes on Gasoline and Diesel Fuel as of 2001 [(Euros/Liter)/(Euros/Gallon)]

Country	Gasoline	Diesel	Value-added tax (percentage)
Austria	0.408/1.55	0.283/1.07	20.0
Belgium	0.507/1.92	0.290/1.10	21.0
Denmark	0.524/1.98	0.367/1.39	25.0
Finland	0.552/2.09	0.300/1.14	22.0
France	0.571/2.16	0.374/1.42	19.6
Germany	0.593/2.25	0.409/1.55	16.0
Greece	0.297/1.13	0.245/0.93	18.0
Ireland	0.348/1.32	0.249/0.94	20.0
Italy	0.520/1.97	0.382/1.45	20.0
Luxembourg	0.372/1.41	0.253/0.96	15.0
The Netherlands	0.402/1.52	0.339/1.28	19.0
Portugal	0.289/1.09	0.246/0.93	17.0
Spain	0.372/1.41	0.270/1.02	16.0
Sweden	0.528/2.00	0.356/1.35	25.0
United Kingdom	0.765/2.90	0.765/2.90	17.5

Figure 3 shows the average gasoline prices in Canada, France, Germany, Japan, the United Kingdom, and the United States for the years 1990 and 2001, illustrating the importance of taxes in the overall prices.

4. VEHICLE TAX INCENTIVES PROMOTING HIGHER FUEL ECONOMY

4.1 Overview

A number of countries have tax programs designed to reward purchasers and owners of fuel-efficient vehicles (or to penalize purchasers and owners of fuel-inefficient vehicles) by tying vehicle purchase and/or ownership taxes either directly to fuel economy or to vehicle features associated with fuel economy (e.g., engine displacement). Aside from taxes aimed specifically at fuel economy or associated vehicle characteristics, all countries in the major auto markets have vehicle taxes that are based on the total purchase

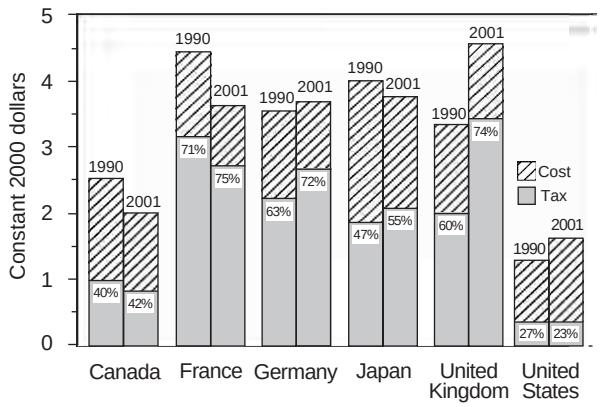


FIGURE 3 Gasoline prices for selected countries, 1990 and 2001, showing effect of taxes. From Oak Ridge National Laboratory. (2002). "Transportation Energy Data Book," 22nd ed.

price of the vehicle (e.g., sales taxes in the United States, value-added taxes in Europe). This latter form of taxes can actually serve as an impediment to higher fuel economy to the extent that increased efficiency is attained with high-cost technology, such as continuously variable automatic transmissions, hybrid electric drivetrains, and variable valve controls, and such taxes add to the cost of these technologies. Of course, such taxes, by making vehicles more expensive, reduce sales and so limit fuel consumption in this manner.

4.2 United States

The United States has two distinct tax-based programs aimed at improving vehicle efficiency. The first is the well-known "gas guzzler tax," established by the Energy Tax Act of 1978, which adds a substantial purchase tax to the price of passenger cars (light-duty trucks are not subject to the gas guzzler tax) whose fuel economy falls below an EPA test value of 22.5 mpg. The current tax rate begins at $1000 U.S. for cars with fuel economy between 21.5 and (less than) 22.5 mpg, and it escalates to $7700 U.S. for cars with fuel economy of less than 12.5 mpg. For the 2001 model year, there were 35 models subject to the tax, and all of these were expensive luxury or performance cars such as the Dodge Viper, Mercedes SL500, Rolls-Royce Corniche, and BMW 750IL. Most analysts believe that mass market manufacturers have tried hard to keep their larger models above the tax's minimum fuel economy, not only to avoid the payment but also to avoid having these vehicles subject to the stigma attached to being identified as "guzzlers."

The second U.S. program is a substantial tax deduction from federal income taxes of up to $2000 U.S. (i.e., $2000 U.S. may be deducted from reported net income before calculating taxes) for hybrid vehicles. The tax credit was originally established as a "Clean Fuel" program designed to promote the use of alternative nonpetroleum fuels, but the U.S. government's Internal Revenue Service determined that the program applies to hybrids. This program is scheduled to phase out between 2004 and 2006, although Congress might extend the deadline.

4.3 Japan

Japan has instituted a number of tax incentives for higher fuel economy or for characteristics that are associated with higher fuel economy. They are as follows (with U.S. dollar values calculated at 120 yen/dollar):

- Regarding taxes directly promoting hybrid vehicles, the vehicle acquisition tax is reduced to 2.2% for hybrid vehicles (the national acquisition tax for conventional vehicles is 5%). For example, the purchaser of a Toyota Prius would save about $400. Second, there is better tax treatment of corporate income tax for company purchases (this applies to all vehicles meeting 2010 fuel economy targets and generating 25% less emissions than current standards). Third, there are purchase incentives for private individuals and special government loans for company purchases. The private incentive is half the price difference between the hybrid and equivalent conventional vehicles. For a Prius, the incentive is worth about $1800.
- The annual tax of vehicles meeting both low emissions standards and fuel economy standards is reduced. A Toyota Prius owner would save approximately $120/year for the first 2 years of ownership.
- The vehicle tax is based on weight and assessed at inspection (6300 yen or ~$53/half ton).

4.4 Europe

As in Japan, many members of the EU have passenger car taxes that either directly reward high fuel efficiency or reward vehicle characteristics that are associated with high fuel efficiency. Table IV shows the variety of taxes on both vehicle purchase and ownership in the 15 European countries that have such taxes. Of particular interest are yearly taxes in Denmark that vary inversely with fuel

economy—at approximately $30/vehicle/year for gasoline vehicles with fuel economy above 20 km/L (47 mpg), approximately $330/vehicle/year for vehicles at about 23 mpg, and more than $1100/vehicle/year for vehicles below 10.6 mpg.

The United Kingdom has introduced a number of economic incentives for higher fuel economy, including a vehicle purchase tax based on CO_2 emissions ranging from 60 to 160 pounds (~$95 to $250 U.S. at 2002 exchange rates), tax credits for companies purchasing vehicles with 120 g/km or less CO_2 emissions, and 1000-pound (~$1600 U.S.) grants to consumers purchasing hybrid vehicles.

Another example of a tax program based directly on fuel economy is the tax rebate program adopted by the Netherlands in January 2002 that awarded a purchase tax rebate of 1000 euros for cars with an "A" fuel economy rating (at least 20% more efficient than other cars of similar size) and 500 euros for cars with a "B" rating (10–20% higher fuel economy than cars of similar size). However, this program ended after its first year in existence due to a change in governments.

TABLE IV

Taxes on Vehicle Purchase and Ownership Based on Fuel Economy-Related Variables: European Union and United Kingdom

Country	Sales tax	Ownership tax
Belgium	Based on cc	Based on cc
Denmark		Fuel economy, weight
Germany		cc
Spain	<1.6 L: 7% >1.6 L: 12%	Horsepower
Finland		
France		
Greece		cc
Ireland	<1.4 L: 22.5% 1.4–2 L: 25% >2 L: 30%	cc
Italy		Horsepower
Luxembourg		cc
The Netherlands		Weight
Austria	Fuel economy Maximum 16%	Horsepower
Portugal	cc (e.g., 1.6 L): 5032 euros	cc
Sweden		Weight
United Kingdom		CO_2 emmissions

Note: cc, cubic centimeters of engine displacement.

4.5 Measures in Other Countries

We know of no studies that have examined vehicle taxation programs elsewhere in the world, but it seems likely that there are a number of tax regimes that affect vehicle fuel economy, albeit indirectly. For example, Brazil reduces its 25% tax on industrialized products to 10% for automobiles with engines less than 1 L in size, greatly encouraging sales of low-powered vehicles that are likely to be fuel efficient.

5. INFORMATION PROGRAMS TO PROMOTE HIGHER FUEL ECONOMY

5.1 Overview

A variety of information programs exist in the major automobile markets to assist potential purchasers of new cars in comparing fuel economy characteristics (and other attributes) among competing vehicles. However, there is no evidence that the existence of such systems plays a significant role in consumer choice of vehicles. For example, in 1995, the light-duty fleets in the two European countries with fuel economy labeling systems (Sweden and the United Kingdom) were the least efficient in Europe (according to International Energy Agency data).

5.2 United States

The fuel economy information program in the United States is the oldest and most highly developed of the major vehicle markets. The current system includes a label on the window of all new cars and light trucks available for sale that lists the approximate fuel economy the average driver can expect to obtain in city and highway driving (obtained by adjusting the fuel economy test values obtained on the official EPA city and highway test cycles downward by 10% for the city cycle and 22% for the highway cycle) and other relevant information. More specifically, the label includes the following:

- a description of the vehicle sufficient for the customer to distinguish it from other similar models, including the identical model with a different engine or transmission;
- the (adjusted) city and highway fuel economy values;
- the statement of the likely range of fuel economy values obtained by the majority of drivers,

estimated at 15% below to 15% above the city and highway fuel economy values;
- a statement of the range of fuel economy levels obtained by comparable vehicles;
- an estimate of the annual fuel cost.

The second component of the system is a fuel economy guide, available at automobile dealers and on a Web site (www.fueleconomy.gov), that offers an extensive menu of information about fuel economy and other vehicle attributes. Main features include the following:

- for every vehicle marketed, identified by manufacturer, model, engine option, and transmission type: city and highway fuel economy, annual fuel cost (assuming 15,000 miles/year [45% highway driving and 55% city driving]), annual greenhouse gas emissions (with the same assumptions), and an EPA air pollution score;
- on the Web site, a search engine that allows selection by market or size class, by manufacturer, or by fuel economy;
- on the Web site, capability to display competing vehicles side by side;
- identification of "best" and "worst" vehicles, on the basis of fuel economy, for each size class;
- compilation of advice on driving more efficiently;
- simple information about advanced technology vehicles (e.g., hybrid electric vehicles).

EPA also offers a "Green Vehicle Guide" at its Web site (www.epa.gov/autoemissions) that lists, for each new vehicle, its fuel economy value and a ranking from 0 to 10 based on its emissions of air pollutants.

In addition to the consumer information available in the Fuel Economy Guide and on the Web site, EPA publishes a yearly compilation of fuel economy trends that provides guidance to policymakers and analysts about national progress in fuel efficiency. The 2001 version of the report, *Light-Duty Automotive Technology and Fuel Economy Trends 1975 through 2001,* is available on the Web (www.epa.gov/oms/cert/mpg/fetrends/rdtrnd01.htm). The report highlights key fuel economy and technology trends in the U.S. LDV fleet, including trends in vehicle attributes, such as horsepower and weight, that affect fuel economy. Fuel economy information is broken down in a variety of ways (e.g., total fleet, all cars, all light trucks) by size and weight class, by manufacturer, by marketing group (e.g., Ford Group—Ford, Mazda, Volvo, Jaguar, Land Rover), and the like. Also, the report shows the difference in fleet fuel economy using the standard EPA assumption of 55% city driving and 45% highway driving and shows actual measured values of city and highway driving (the "actual" fuel economy values are about 2% lower than the estimated values using the 55/45 split). Trends in technology (e.g., four-wheel drive, four valves per cylinder) and vehicle characteristics (e.g., horsepower, weight) are shown in tables and figures. One interesting feature is a series of analyses that show how the fleet fuel economy would change if all of the vehicles in each size class or weight class attained the fuel economy of the best vehicles in that class. Finally, the report includes an analysis of what fleet fuel economy would have been had there been no change in average vehicle weight and acceleration capabilities over the past 10 or 20 years. This analysis shows the impact on fuel economy of changes in vehicle weight and acceleration capability. The analysis is then repeated with size class substituted for weight.

5.3 Japan

The government of Japan does not require fuel economy labels on new cars, although it has started to consider introducing such labeling. However, the Ministry of Land, Infrastructure, and Transport does publish a fuel economy guide (and maintains the guide on its Japanese-language Web site) that includes information for each vehicle on fuel economy and CO_2 emissions and highlights vehicles that already meet 2010 fuel economy standards. Also, manufacturers must list fuel economy values in sales brochures and vehicle specification sheets at car dealerships.

5.4 Europe

Two European countries, Sweden and the United Kingdom, have had vehicle fuel economy labeling systems in place for the past two decades or so. The United Kingdom's system has been mandatory, with fines set for noncompliance, but the Swedish system generally has not been enforced.

The EC has established (in Directive 1999/94/EC) consumer information requirements for new passenger cars that demand the following:

- fuel economy labels on all new cars that include both fuel economy and CO_2 emissions rates;
- ready availability of a fuel economy guide containing fuel economy ratings of all new vehicles sold in that market;

- posters in new vehicle showrooms listing the fuel economy ratings of all vehicles for sale at those locations;
- incorporation of fuel economy/fuel consumption values in advertising material.

The official labeling program in the Netherlands complies with the EC directive in that labels contain fuel consumption and CO_2 emissions values, and it goes a step further in giving each vehicle a CO_2 emissions rating, from "A" (>20% more efficient) to "G" (>30% less efficient), that compares the car with others of similar size (defined as car width × car length) as well as with the average of all cars. The rating is 75% dependent on the comparison with cars of similar size and 25% dependent on the comparison with the average of all cars. This feature was adopted as a compromise, on the one hand, to ensure that small cars that are rather inefficient for their size are not always labeled as if to appear superior to larger cars that are quite efficient for their size and, on the other, to avoid giving large cars the same rating as small cars of similar technical efficiency even though the large cars have much higher CO_2 emissions.

6. GOVERNMENT SPONSORSHIP OF RESEARCH AND DEVELOPMENT

6.1 Overview

Although the automotive industry is a vigorous sponsor of research and development—worldwide, nearly $20 billion U.S. worth in 1997—governments throughout the world sponsor additional automotive R&D, both separately from and in partnership with the industry. This work focuses primarily on four areas: emissions, safety, fuels, and fuel economy. Within the past few years, government sponsorship of automotive R&D has moved sharply in the direction of attempting to advance the performance and cost-effectiveness of automotive fuel cells, which address three of the four areas: emissions, fuels, and fuel economy. Aside from large programs in the three major market areas—the United States, Japan, and Western Europe—there is a significant fuel cell program in Canada and there are small efforts in South Korea and elsewhere.

Aside from direct sponsorship, governments promote R&D in the automotive industry (and elsewhere) through generous tax credits for R&D. Although the actual amount of tax credits taken by automakers in the various markets is proprietary information, Japanese vehicle companies (excluding parts producers) had approximately $190 million U.S. available to them in 1999 and $122 million in 2000. U.S. companies (including parts producers) had approximately $540 million available to them in 1997 and $617 million available in 1998 in U.S. R&D tax incentives.

6.2 United States

The U.S. government has sponsored R&D in vehicle efficiency technology for decades, primarily through the Department of Energy, whose Electric and Hybrid Vehicle Program, begun in 1976, spent about $500 million from its inception to the beginning, in 1992, of the Partnership for a New Generation of Vehicles (PNGV). PNGV was a joint R&D program between the federal government and the U.S. Council for Automotive Research (USCAR), an R&D partnership of Ford, General Motors, and DaimlerChrysler. For approximately a decade, PNGV was the centerpiece of the government's vehicle research efforts, with one of its key goals being the demonstration of a family car that could obtain 80 mpg while meeting all safety, emissions, performance, and affordability requirements. Research areas included engine powertrain systems (including lean-burn engines and emissions controls), advanced materials, electric energy storage and electric motors and controllers (for electric hybrids), and vehicle manufacturing. Federal research dollars were spread among the USCAR companies and their suppliers, the national laboratories, and other R&D performers such as universities.

In January 2002, the FreedomCAR Partnership superseded the PNGV, with the research focus shifting to the development of hydrogen as a primary fuel for cars and trucks. In fiscal year 2002, the Department of Energy spent approximately $150 million on FreedomCAR-related research, with $60 million focused on fuel cell and hydrogen infrastructure R&D. Despite the shift in primary focus, FreedomCAR's research goals retain some of PNGV's focus on materials and engine powertrain systems.

6.3 Japan

The Japanese government is sponsoring extensive research activity in high-efficiency vehicle technology for both conventional and fuel cell drivetrains. The 2001 Prime Minister's Action Plan calls for a $155

million R&D program (among other initiatives such as vehicle purchase subsidies, tax incentives, and low-interest loans) aimed at putting 10 million low-emissions vehicles (including fuel cell, hybrid, and alternative-fuel vehicles) on the road by 2011, with 50,000 of these being fuel cell vehicles. The plan further calls for 5 million fuel cell vehicles to be on the road by 2020. More recent recommendations from senior ministers of the key transportation and environment ministries call for increased fuel cell R&D funding, more rapid vehicle deployment than is called for in the action plan, and additional measures to achieve global fuel cell leadership.

Prior to the new focus on fuel cells, the major government/industry R&D project for high-efficiency vehicles was the Advanced Clean Energy Vehicle (ACE) Project. The ACE Project's primary goal was to develop vehicles using clean fuels that would double energy efficiency with ultra-low emissions vehicle (ULEV) emissions. The total government plus private funding over the project's lifetime was approximately $175 million, but the LDV part of this was substantially smaller (the project included development of technologies for buses and cargo trucks).

6.4 Europe

Both European auto manufacturers and the EU responded vigorously to the U.S. PNGV program, the manufacturers by forming the European Council for Automotive R&D (EUCAR) to coordinate R&D among the companies and the EU by initiating the Car of Tomorrow program. EU funding for EUCAR starting with the initiative was $2.3 billion for 5 years.

Fuel cell funding by European governments was approximately $60 million in 2001, with half from the EU and half from individual countries. However, the majority of these funds is aimed at transit projects. Key research areas are component cost reduction, development of a cost-effective hydrogen supply infrastructure, and development of vehicle systems. At the national government level, France and Germany have the most robust programs at approximately $11.5 million and $8 million, respectively (for fiscal year 2000). The Netherlands, Spain, Denmark, Italy, and the United Kingdom have programs at or above $2 million per year, with Switzerland, Sweden, and Norway having smaller programs. In addition, individual governments have initiated more generalized efficiency R&D programs. For example, the United Kingdom's Foresight Vehicle Initiative, with nearly $20 million of government money invested, has focused on hybrid, electric, and alternative-fuel vehicles; advanced materials and structures; and powertrain technology—a program quite similar to that of the PNGV. However, even this program's focus is shifting more to fuel cell technology.

6.5 Canada

The Canadian government's R&D support has focused strongly on fuel cell research, in which it is a global leader. Canadian fuel cell research began in 1983 at Ballard Power Systems through a defense contract to further develop General Electric/National Aeronautics and Space Administration technology developed for the U.S. space program. Canadian government fuel cell support was approximately $40 million in 2001, about one-third of the U.S. government's effort.

Acknowledgments

This article was produced by the University of Chicago as operator of Argonne National Laboratory under Contract W-31-109-ENG-38 with the U.S. Department of Energy. The U.S. government retains for itself, and others acting on its behalf, a paid-up, nonexclusive, irrevocable worldwide license in the article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the government.

SEE ALSO THE FOLLOWING ARTICLES

Alternative Transportation Fuels: Contemporary Case Studies • European Union Energy Policy • Fuel Cycle Analysis of Conventional and Alternative Fuel Vehicles • Hybrid Electric Vehicles • National Energy Policy: Japan • National Energy Policy: United States • Taxation of Energy • Transportation Fuel Alternatives for Highway Vehicles • Vehicles and Their Powerplants: Energy Use and Efficiency

Further Reading

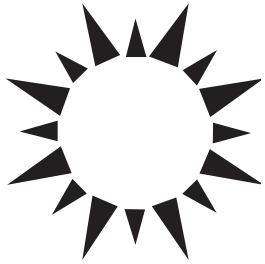
Greene, D. L. (1996). "Transportation and Energy." Eno Transportation Foundation, Landsdowne, VA.

International Energy Agency. (2001). "Saving Oil and Reducing CO_2 Emissions from Transport: Options and Strategies." OECD/IEA, Paris.

National Research Council. (2002). "Effectiveness and Impact of Corporate Average Fuel Economy (CAFE) Standards." National Academy Press, Washington, DC.

Plotkin, S. (1999). Technologies and policies for controlling greenhouse gas emissions from the U.S. automobile and light truck fleet. *Transport. Q.* **53**(2), 7–30.

Fuzzy Logic Modeling of Energy Systems



BJØRN LUDWIG
Technowledgement Consulting
Göttingen, Germany
Technical University of Clausthal
Clausthal, Germany

1. Introduction
2. Methodical Aspects
3. Modeling Approach with Soft Computing
4. Application to the Energy Supply
5. Results and Discussion

Glossary

fitness function A mathematical expression transforming a certain combination of several input values into one output value, representing the evaluation of the input, often in relation to the optimum.

fuzzy logic A soft computing method to carry out logic operations; uses verbal expressions described by fuzzy sets generated from membership functions.

genetic algorithms A soft computing method that uses biological processes of evolution for optimization.

linguistic variable A variable for which the values are linguistic terms (for instance, "small") instead of numbers. Fuzzy sets generated by membership functions build the connection between the linguistic expression and the numerical information.

membership function A mathematical expression that represents to what degree an element belongs to a mathematical set, considering that these memberships do not have to be total memberships.

operator A mathematical manipulation on numbers or terms. Classical examples are addition, multiplication, differentiation, etc. In soft computing, other operators are needed, e.g., fuzzy or genetic operators such as aggregation, accumulation, crossover, mutation, and reproduction.

soft computing Mathematical methods of calculating, developed mainly within the past 30 years, and not necessarily operating on classical functions and numbers, but taking into account the human kind of information processing and principles of biological operating.

sustainable development A goal for the future development of mankind under intragenerational and intergenerational justice. The negotiated road map for future economical, ecological, and societal sound development is the Agenda 21, from the United Nations Conference for Environment and Development held in Rio de Janeiro in 1992.

technology assessment An interdisciplinary scientific and societal approach to the questions of what impacts technology application and the technological progress have on mankind, and what technologies are needed and preferred.

The energy supply system is crucial to the further development of humans. Evaluations concerning the impact of energy conversion technologies on this development are difficult due to the complexity of the problem. This leads to different conclusions about how to proceed. Using soft computing techniques, both technically measurable quantities and nontechnical evaluation criteria can be taken into account to obtain an objective evaluation, to discover hidden evaluation criteria, and to formulate a goal for a future sustainable development.

1. INTRODUCTION

Current global development increasingly calls attention to the realities of population growth, extended mobility, and increasing energy demands and their contribution to environmental problems. This collision between civilization and the ecosystem illuminates two dilemmas: (1) The goals of maintaining society's prosperity and protecting the environment are not independent. Rather, they are competing goals, because technological applications increasingly

 © 2004 Elsevier Inc. All rights reserved.

require extensive use of nonrenewable resources. (2) The globally increasing demand for nonrenewable resources necessitates even more technology application and development. Thus, guided by the vision of sustainability, the concept of technology assessment (TA) must deal with the question of technology's ability to enable sustainable development. To avoid increasing both resource use and risk, new technologies should be appropriate to address developmental needs, yet ecologically "reasonable." Designing such technologies requires improved information on and knowledge of technical systems. This knowledge encompasses development and application of technical systems and the connections among economic, social, and political systems, as well as impacts on the environment. Such knowledge should be used when evaluating both existing and new technologies within decision support systems. Problems characteristically encountered in making such evaluations involve multicriteria decision making, aggregation of facts, weighting of goals, uncertain information, and integrating people into the decision-making process. The overall problem is not a lack of data, but the difficulty in generating knowledge from a vast amount of partially uncertain, partially fuzzy data and information, often structured in such a way as to be rather useless. The global energy supply system, for example, is structured to consist of different primary energy carriers. This leads inherently to different evaluations of the system (for instance, due to the different ranges or durations of the primary energy carriers, or to the differences in the dangers of operating different conversion technologies), thus the present structure cannot be the most desirable one. An essential question for the future, concerning technology assessment studies, is how to achieve a sustainable energy supply. Under the present energy supply structure, evaluations for each applied energy conversion technology are needed.

Approaches to this entire problem are possible using a hybrid method of the soft computing techniques, fuzzy logic and genetic algorithms, first introduced by the author. This approach of recursive optimization considers different technical and nontechnical evaluation criteria to achieve optimum primary energy percentages that form the best energy mix. The aim is to reduce the multidimensional decision-making problem to a comparison of aggregated one-dimensional numbers. This should be achieved by combining the advantages of human thought-process structures and processes of biological development with the logical and analytical accuracy of computers.

2. METHODICAL ASPECTS

Examining such problems analytically requires cognitive information processing. The desired systematic procedures require application of methods to organize and to extend our knowledge on complex systems. Generally, problem-specific combinations of methods are applied. Methodical problems and deficits usually emerge due to the complexity of the processes or of the impacts. Problems in describing complex dynamical systems are due to nonlinear impacts and boundary conditions, to boundary conditions that are time dependent and/or solution dependent, and to different time constants within the subsystems.

2.1 Deficit Analysis

Knowledge about technical systems is commonly gained by analyzing the underlying interactions. This is mainly characterized by the building of models. These are mainly mathematical models that allow systematic simulations to identify potential reactions of the system.

In the case of a holistic and systemic view, the system border is extended to all participating and impacting areas. Thus, interactions and impacts have to be considered, but often resist analysis using a description with classical mathematics—say, a formalistic handling. These entities have one of the following properties: unique basic variables are not identifiable, unique causalities are not recognizable among them, or measurable indicators cannot be determined in case of known interactions. This plays a considerable role in the discussion of sustainability, because the entities could be impacts such as risk, quality of life, or diversity. These are hardly measurable quantities. Mainly they are considered in models by boundary conditions or verbal discussion. Neglecting these entities or disaggregating them to other available variables leads, in any case, to an information loss, so the complexity of the system is not represented completely by the model.

2.2 Deficit Synthesis

Complex systems in practice are handled, or at least are made more accessible, by composite numbers or indicators. This is familiar, for instance, in process or chemical engineering. In contrast to the information content of single-measurement data, composite indicators generate new knowledge by aggregating single data and information. The procedure of

aggregation varies depending on the indicators. Nevertheless, indicators of different aggregation levels always provide the same information content. Indicators are especially important in environmental discussions. Assessing or evaluating a situation or guiding a decision requires differently aggregated indicators that characterize a process or a state.

The aggregation problem points out the following dilemma: the methodical deficits of assessments are due to problems of comparison because of the transformation of indicators to highly aggregated assessment measures. On the other hand, this procedure is necessary, which becomes obvious, for instance, if entities such as environmental pollution are to be quantified. This property allows less aggregated indicators to be used at the beginning of systems analysis (for instance, within an inventory analysis of modern environmental management approaches) and enhances transparent aggregation procedures.

2.3 Deficit Assessment

A multidimensional assessment problem can be considered as a logical measurement operation. The assessment problem then is the optimal arrangement of a set of possible alternatives, with respect to the relevant aims and to the preferences of the decision makers, considering given restrictions. The aim of an assessment is to determine a particular value for any alternative so that its relative advantages are represented in only a single expression. The solution of this selection problem will be especially difficult if the problem structure is complex (that is, when many objectives are to be considered, different assessment scales emerge, objectives are weighted differently, information is at least partially uncertain, the problem is time-dependent, many participants are to be involved in the decision-making process, or no unique criterion exists for decision making). Contributing variables could be factors such as risk or diversity, which represent heterogeneous fuzzy expressions that cannot be broken down any further, often with many different impacts. This makes comparability of these variables difficult, especially when factors of different aggregation degree are involved, and makes any evaluation subjective. Nevertheless, to describe real systems, these influences are indispensable quantities.

2.4 Goal

Designing innovative technologies, considering widespread aspects such as sustainable development, requires environmental information for design, assessment, and decision processes. The difficulties mainly emerge during recording and assessing of entities. The problem is caused by aggregated system variables that are incomparable. More adequate methodical approaches maintain the complexity of these variables. Necessary subjective inputs during a system analysis then have to be made absolutely clear to maintain transparency, and they have to be introduced as early as possible in an algorithmic procedure to avoid a conscious manipulation of the solution.

3. MODELING APPROACH WITH SOFT COMPUTING

At a level of thinking that does not have to be conscious, humans use a kind of knowledge that is not sharp and is hard to formulate explicitly. This fuzziness is not a property of reality, but of perception. This kind of knowledge is characterized by the handling of incomplete and not well-structured patterns and information. Here the question emerges as to how and with what language we can model such uncertainties. The methodical deficits mentioned previously can be softened by application of the techniques of soft computing. Apart from artificial neural nets, these are primarily fuzzy logic and genetic algorithms; their usefulness is due to their potential to integrate complexity into calculation.

3.1 Fuzzy Logic

Within the context of this discussion, only fuzzy set theory, one branch of fuzzy logic, is of interest for application to technology assessment, because the important questions of TA do not deal with maintaining states of operation, the subject of control theory, but with fuzzy information processing. Fuzzy logic enables the handling of verbal and indeterminate information using exact mathematics.

Traditional binary logic is part of fuzzy logic, as a special case operating only with two values of interpretation. A set is fuzzy limited if not all the members of the set are assigned a value of 1, or 100%, which means total membership. A fuzzy set is defined by the generalized characteristic function, called the membership function μ. This real function can take on any values, but usually it is normalized into the interval [0, 1].

The key notion when modeling with fuzzy logic is the linguistic variable. In contrast to mathematically precise quantifiable equations, verbal descriptions of

activities contain terms of properties with smooth gradation transitions. Linguistically formulated variables are generally more intelligible, but have a higher aggregated information content. These entities are therefore more difficult to quantify without an information loss. The values of a linguistic variable are verbal expressions, called linguistic terms (for instance, the term "small"). The content of each linguistic term is identified with one fuzzy set and is assigned to the related numerical scale of the basic variable by a particular membership function. Thus, the fuzzy sets build the connection between the linguistic expression and the numerical information.

To process fuzzy formulated knowledge, several linguistic variables must be linked by linguistic operators. The connecting rules represent the knowledge, which is stored in a rule base or knowledge base, similar to expert systems. The procedure consists of the sequential steps fuzzification, inference, and defuzzification.

3.2 An Assessment Method Using Fuzzy Logic

Along with the preceding remarks, a general assessment method using fuzzy logic has been developed. An assessment is a determination of the goodness of a considered issue. This will be made by a function that systematically integrates knowledge into the procedure of modeling and also handles the aggregation procedure. The assessment procedure is based on a multidimensional vector of criteria containing all knowledge input, which characterizes the matter to be evaluated related to the chosen degree of aggregation. The aim is to get statements of the form "*s*th is good, fair, satisfying, to introduce, desirable, etc." This kind of statement can be understood as a result of logic operations.

The concept of linguistic variables enables the formulation of scales of values with linguistic terms as gradations. The essential and critical process is the formalistic assignment of human knowledge to fuzzy sets. However, through this procedure, existing subjective preferences are integrated transparently at each stage, so a conscious manipulation has only a limited impact on the final result.

The general five-step top-down procedure is as follows:

1. Definition of a scale for evaluation.
2. Definition of subordinate criteria for ranking onto this scale. The question concerning which of the possibly independent criteria and impacts will influence the superior scale from step 1 has to be answered.
3. Stepwise disaggregation of these basic criteria. The subordinate impacts are disaggregated to further arbitrary substages, depending on the given knowledge. The result is an aggregation hierarchy of the different impacts. Generally, steps 2 and 3 answer the question of which of the less aggregated and possibly independent impacts are impacting on the next higher aggregation level. Depending on the character of an impact, it can be described by a measurable quantity on a less aggregated level. Otherwise, it keeps the character of an aggregated entity. Both kinds of characters can be considered simultaneously.
4. Fixing the specifications of the fuzzy system parameters. This includes fixing the numbers of linguistic terms and the verbal connotations of all attendant variables, the shape of the membership functions (default: triangular), and the logic operators for aggregation (default: minimum) and for accumulation (default: maximum), as well as the defuzzification method (default: center of gravity). It is recommended to choose an odd number of linguistic gradations for each input variable (default: three) to ensure a certain resolution, with the additional possibility of naming a medium term. The number of linguistic terms of all superior aggregated variables then can be taken as determined in order to get an objective rule base (see step 5).
5. Formulation of the rule base. The phenomenological dependencies between the variables are stored in the rule base. The rules state which combination of linguistic input terms is connected with which linguistic output term. To achieve a definite assignment in the rule base, the number of linguistic terms of a superior aggregated variable then is determined by all possible combinations of input linguistic terms of all related subordinate variables.

This is carried out by a numerical codification. The N linguistic terms of a subordinate variable are assigned to numbers in the interval $[-(N-1)/2,...,-1, 0, 1, ..., (N-1)/2]$ ($N = 3, 5, 7, ...$), where the medium term is assigned to zero. For instance, a variable with $N_{In} = 3$ linguistic terms (small, medium, large) would be assigned to $(-1, 0, 1)$. Thus, we can determine the number N_{Out} of linguistic terms of an output variable from its M input variables ($M = 1, 2, 3, ...$) with their $N_{In,1}, ..., N_{In,M}$ linguistic terms as follows [Eq. (1)]:

$$N_{Out} = 1 + 2\sum_{i=1}^{M} \frac{N_{In,i-1}}{2} \tag{1}$$

For instance, an output variable aggregated from $M = 2$ input variables with $N_{In,1} = 3$ and $N_{In,2} = 5$

linguistic terms would then have $N_{Out} = 1 + 2(1 + 2) = 7$ linguistic terms.

The rules can be derived from the following procedure. The conditions of all rules that are the permutations of all linguistic input terms can be written numerically codified. The conclusion is then represented by the linguistic term of the output variable (numerical codification is assigned by the sum of the numerical input codifications). For instance, considering one three-term input variable with the terms (small, medium, large) codified as (−1, 0, 1), and a second input variable with five terms (very small, small, medium, large, very large) codified as (−2, −1, 0, 1, 2), the particular rule may be written as follows:

$$\text{IF Input}_1 = \text{small AND Input}_2 = \text{medium}$$
$$\text{THEN Output} = ? \quad (2)$$

which is codified as

$$\text{IF Input}_1 = -1 \text{ AND Input}_2 = 0$$
$$\text{THEN Output} = -1 + 0 = -1 \quad (3)$$

According to Eq. (1), the output variable has seven terms codified as (−3, −2, −1, 0, 1, 2, 3). Therefore, the conclusion Output = −1 means that the result is the linguistic output term that is located one term smaller than the medium term, whatever the verbal connotation may be.

3.3 Genetic Algorithms

The idea of technology assessment to point out alternatives to any problem solution can be taken as an optimization problem, where optimization of an application is aimed at, for instance, its energy efficiency, its material intensity, or its sustainability. The application of usual optimization procedures is limited. Restraints can be nonlinear problems or problems with many parameters or many restrictions due to the programming effort and memory demand.

Genetic algorithms, instead, refer to a behavior that would be similar to a biological process of evolution. These techniques do not have requirements for linearity, differentiation, or a certain structure, and often require less programming effort. Problem solutions are kept or changed due to the principle of survival of the fittest. The point is that the search for more capable solutions takes place with a high probability at sites of the solution space, where an enhancement higher than average is required. These features are used in genetic algorithms to evaluate systematically created solutions. A genetic algorithm can be defined as an iterative procedure maintaining a population of structures that are candidate solutions to specific domain challenges. Each candidate solution is called a chromosome. As in biological systems, a chromosome can make a copy of itself. The copy might be slightly different from its parent. During each temporal increment, called a generation, the structures in the current population are rated for their effectiveness as domain solutions, and on the basis of these evaluations, a new population of candidate solutions is formed using specific "genetic operators," which are used by most genetic algorithms, such as reproduction, crossover, and mutation:

1. *Reproduction:* Through reproduction, genetic algorithms produce new generations of improved solutions by selecting parents with higher fitness ratings or by giving such parents greater probability to be contributors.
2. *Crossover:* Usually genetic algorithms use binary strings to represent solutions. Crossover means choosing a random position on the binary string (that is, after a certain digit) and exchanging the segments either to the right or to the left of this point with another string partitioned similarly.
3. *Mutation:* This genetic operator is an arbitrary change of a single bit of a chromosome. It is needed to keep the algorithm from getting stuck. This operator changes a 1 to 0 or a 0 to 1.

The natural optimization techniques can be distinguished by their main operators. In case of genetic algorithms, the main operator is crossover, which is chosen with a probability of 80–100%. Mutation, in contrast, is a very subordinate operator. Such a bit change occurs at a low probability of 1–10%. Genetic algorithms use a three-step iterative process: (1) test a solution to see how good it is, (2) select the best parents, and (3) generate offspring. Genetic algorithms provide a set of efficient, domain-independent search heuristics. Information that enables rejection of inferior solutions and accumulation of good ones is received. Only genetic algorithms and their derivatives provide methods powerful enough for optimizing complex problems. Traditional optimizing tools are not applicable.

4. APPLICATION TO THE ENERGY SUPPLY

Decreasing resources of fossil primary energy carriers and globally increasing mean per capita energy demand and population growth bring to view the

question how and by which technologies the future energy supply can be guaranteed. Of primary concern here is the evaluation of the sustainability of the energy supply system. An energy supply system provides energy via different conversion technologies. In the present global energy supply system, 90% of the demand is still provided by fossil fuel-based primary energy carriers. Increasing energy demands and increasingly urgent global impacts require development of a new supply system, keeping in mind future generations and threshold countries with vastly growing economies; this is referred to as intergenerational and intragenerational justice. Due to the requirement for sustainability of a new energy supply system, one of the most important questions to be answered concerns which primary energy carriers can be and should be used, and to what extent in consideration of limited resources and potential risks. This question concerns the optimal structure of percentages of the different primary energy carriers, of the different energy conversion technologies, or the optimal energy mix. The applied recursive optimization is a two-step iterative procedure. First, a given energy supply structure must be evaluated, considering the technologies involved, with regard to certain evaluation criteria. Second, the input structure will be varied to find the best rated energy mix.

4.1 Assessment: The Fuzzy Fitness Function

To evaluate a given primary energy structure, the fuzzy logic assessment method described previously is used. The assessment procedure is based on a multidimensional vector of criteria that characterizes a conversion technology. The method is applied as follows, according to the previous description:

1. Definition of a scale for evaluation. The interesting entity is the sustainability of an energy conversion technology, here reduced to the related primary energy carrier.

2. Definition of subordinate criteria for the ranking onto the scale sustainability. In case of energy conversion technologies, the independent criteria duration of the primary energy carrier and ecological relevance of the technology are chosen.

3. Stepwise disaggregation of these subordinate criteria. In this case (Fig. 1), the ecological relevance could be considered as being made up from the basic criteria climate impact, area demand, and risk. Area demand is composed of the value and the size of the used area. Duration, for instance, may depend on the level of technology, on the reserve of the primary energy carrier considered, and on social factors. The climate impact can be specified with technical quantities such as emissions and global warming potential. However, both are not differentiated further here. Risk in this approach should specify the potential danger in case of an accident.

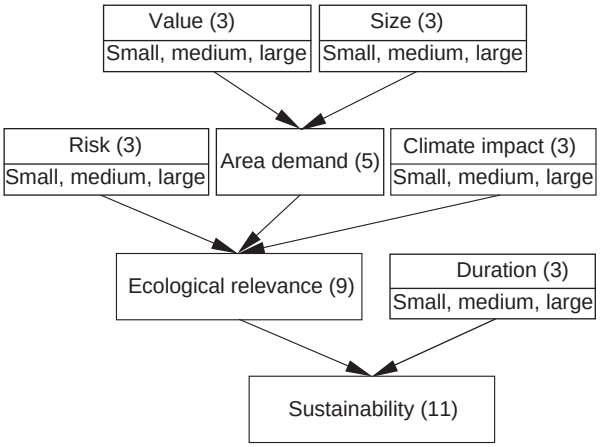


FIGURE 1 Aggregation hierarchy of the assessment function. The number of linguistic terms of each variable determined with Eq. (1) is in parentheses. Reproduced with permission from the Society of Automotive Engineers, SAE 1999-01-2708. Copyright 1999, SAE.

4. Fixing the specifications of the fuzzy system parameters. Here, for all specifications, the default standard parameters are chosen. Each input variable has been aligned to three linguistic terms (small, medium, large). According to Eq. (1) and the hierarchical aggregation structure (Fig. 1), the numbers of linguistic terms for the superior variables given in Fig. 1 are obtained. All fuzzy sets at the edges of the intervals of the basic variables are performed in such a way that defuzzification of the minimum (maximum) linguistic term would provide the value of 0 (1) of the basic variable. Therefore, the membership functions of these linguistic terms of the input basic variables have to be defined symmetrical to 0 (1).

5. Formulation of the rule base. With the preceding procedure, the objective rule bases for the three aggregation levels (for area demand with 9 rules, for ecological demand with 45 rules, and for sustainability with 27 rules) are obtained. For example, the aggregation levels would have the following formulations:

$$\begin{aligned} &\text{First aggregation level:}\\ &\text{IF Value} = \text{medium}\\ &\text{AND Size} = \text{small}\\ &\text{THEN Area Demand} = \text{small} \end{aligned} \tag{4}$$

Second aggregation level:
IF Risk = small AND
Area Demand = small
AND Climate Impact = small
THEN Ecological Relevance = very very small (5)

Third aggregation level:
IF Ecological Relevance = very very small
AND Duration = large
THEN Sustainability = very very very large. (6)

Each primary energy carrier being considered can now be characterized by the vector of criteria (elements are the values of the inputs, each on a specific aggregation stage). The assessment result can be a verbal or a numerical value of the highest aggregated output variable, "sustainability." In case of an energy mix, different technologies are considered with their relative percentages. An input value for the mix then has to be built by the sum of the values of the single technologies weighted with the percentage within the mix. This assessment procedure operates as the fitness function of the genetic algorithm.

4.2 The Optimization Procedure

The optimization problem consists of determining the percentages of each considered technology that participates in supplying the energy demand. During the optimization procedure, the following fitness function F has to be optimized:

$$F = F(\alpha, \beta, \gamma) \rightarrow \max \quad (7)$$

where α represents the contribution of fossil-fuel-based primary energy carriers, β represents that one of the regenerative (renewable)-energy-based primary energy carriers, and γ represents the contribution of nuclear-energy-based primary energy carriers. The boundary condition, of course, is $\alpha + \beta + \gamma = 1$, or 100%. First, the aggregated inputs to the fuzzy block are determined (Fig. 2). Therefore, each single, specific evaluation of a technology taken from Table I is weighted by its percentage and then summarized. This resulting vector of criteria for a considered energy structure serves as input vector for the fuzzy fitness function that determines the total goodness of the structure. The genetic algorithm now varies the input vector of percentages of a given energy structure until the maximum fitness is achieved (Fig. 2).

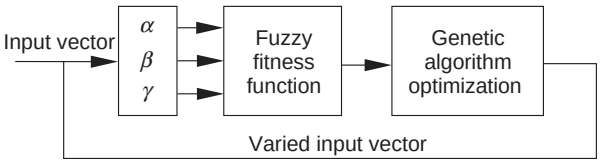


FIGURE 2 Recursive optimization, where the contributions of primary energy carriers are α, fossil fuels; β, regenerative (renewable) energy; and γ, nuclear energy. Reproduced with permission from the Society of Automotive Engineers, SAE 1999-01-2708. Copyright 1999, SAE.

5. RESULTS AND DISCUSSION

Vectors of criteria for different technologies and sustainability results are shown in Table I. The current energy mix has been determined twice, first considering the nuclear percentage based on uranium-235 (mix 1) and second based on the breeder technology (mix 2). A separate control calculation with the analytical hierarchy process method has resulted in the same ranking tendency. Although fossil energy carriers use oil and gas in addition to coal, oil and gas have not been differentiated here, which should be done for more precise results. Also, in reality, hydro energy is the most significant

TABLE I

Vectors of Criteria for Different Primary Energy Carriers

	Criteria[a]					
Vector	CI	Value	Size	Risk	D	S
Test cases[b]						
Max	0	0	0	0	1	1
Min	1	1	1	1	0	0
Technologies[c]						
Fossil (coal)	1.0	1.0	0.01	0.01	0.3	0.439
Nuclear (uranium-235)	0.1	1.0	0.01	0.95	0.1	0.400
Nuclear (breeder)	0.1	1.0	0.01	1.0	0.8	0.510
Regenerative (solar)	0.1	0.01	1.0	0.01	1.0	0.780
Mix 1995[d]						
(1)	0.906	0.968	0.042	0.079	0.308	0.444
(2)				0.082	0.359	

[a] Data from SAE 1999-01-2708. CI, climate impact; D, duration; S, sustainability.
[b] The two test cases verified the algorithm.
[c] The single values for the technologies were determined from the literature.
[d] Evaluation of the 1995 energy structure in Germany: (1), standard nuclear energy; (2), fast breeder technology.

contributor to regenerative energy supplies, but the percentage in Table I has been taken as consisting entirely of solar energy (photovoltaics). Therefore, the values in Table I for climate impact, value, and size of area are not calculated properly. Nevertheless, the results allow the following statements:

1. The assessment procedure provides correct results for given vectors of criteria.
2. Maximum sustainability cannot be achieved.
3. The traditional nuclear energy conversion technology has been ranked worst because of the high risk and the simultaneous short duration.
4. The current energy mix is debatable, considering this evaluation.
5. The best sustainability is achieved for regenerative energies, and is strongly dependent on the value of the size of area used.

Not surprisingly, the optimization procedure provides 100% regenerative energy carriers as the best solution if risk is considered, and 100% nuclear energy if not. Far more interesting are results evaluated as being a little worse than the pure regenerative solution, but better than today's mix. In this case, the dominant regenerative part decreases with decreasing evaluation values, which also start to oscillate. Simultaneously, the percentages of fossil-fuel-based and nuclear-based primary energy carriers increase alternately, also with increasing oscillations. However, there are obviously several solutions with the same evaluation result.

Using a linguistic notion of arrangement, fuzzy logic allows simultaneous consideration of entities and technically measurable quantities by a verbal description of interactions. The evaluation results for the current energy mix and for certain other possibilities with respect to meeting the worldwide future energy demand show that this kind of consideration is useful to discover hidden evaluation criteria and to formulate a goal for a future sustainable development.

SEE ALSO THE FOLLOWING ARTICLES

Computer Modeling of Renewable Power Systems • *Modeling Energy Supply and Demand: A Comparison of Approaches* • *Neural Network Modeling of Energy Systems* • *Population Growth and Energy*

Further Reading

Gore, A. (1993). "Earth in the Balance." Plume, New York.
Grefenstette, J. (1982). Optimization of control parameters for genetic algorithms. *IEEE Transact. Syst. Mgmt. Cybernet.* **16**(1).
International Standards Organization (ISO). (1997). "Environmental Management—Life Cycle Assessment." ISO 14040. Beuth, Berlin.
Ludwig, B. (1995). "Methoden zur Modellbildung in der Technikbewertung." Papierflieger, Clausthal-Zellerfeld, Germany.
Ludwig, B. (1997). On the sustainability of future energy systems. *Energy Convers. Mgmt.* **38**(15–17), 1765–1776.
Ludwig, B. (1997). The concept of technology assessment—An entire process to sustainable development. *Sus. Dev.* **5**(3), 111–117.
Ludwig, B. (1997). "Optimization of Energy Systems Under the Aspect of Sustainability." Proceedings, 32nd Intersociety Energy Conversion Engineering Conference (IECEC), Honolulu, July 27–August 1, 1997, pp. 2070–2075.
Ludwig, B. (1998). Fuzzy logic applications in technology assessment studies. *J. Intelligent Fuzzy Syst.* **6**, 375–388.
Mitchell, G. (1996). Problems and fundamentals of sustainable development indicators. *Sus. Dev.* **5**, 1–11.
Saaty, T. L. (1980). "The Analytic Hierarchy Process." McGraw-Hill, New York.
United Nations. (1997). "Yearbook of World Energy Statistics." United Nations, New York.
Zadeh, L. A. (1965). Fuzzy sets. *Inform. Control* **8**, 338–353.
Zimmermann, H.-J. (1991). "Fuzzy Set Theory and Its Applications." Kluwer Academic Publ., Boston.

Gas Hydrates

TIMOTHY S. COLLETT
United States Geological Survey
Denver, Colorado, United States

1. Implications for Gas Hydrates
2. Gas Hydrate Technical Review
3. Energy Resource Potential of Gas Hydrates

Glossary

clathrate A class of compounds characterized by cages of molecules that encapsulate other molecules; from the *Latin clathratus*, meaning to encage.
dissociation A chemical process by which temperature and/or pressure changes cause a group of molecules to be separated into simpler groups.
methane A colorless flammable gas consisting of carbon and hydrogen.
methanol Methyl alcohol; a colorless, flammable liquid derived from natural gas.
permafrost Ground (soil or rock) that remains at or below 0°C for at least 2 years.
seismic reflectors An interface between two media with different elastic properties as shown on a seismic profile.
well log A graphic record of the measured physical characteristics of a subsurface rock section as encountered in a well.

Gas hydrates, composed of water and gas, are crystalline substances in which solid lattices of water molecules trap gas molecules in a cagelike structure, or a clathrate. Gas hydrates occur in sedimentary deposits under conditions of pressure and temperature present in permafrost regions and beneath the sea in outer continental margins. Despite the fact that relatively little is known about the ultimate resource potential of natural-gas hydrates, it is certain that they are a vast storehouse of natural gas. There is much to learn about these compounds before this potentially enormous resource can be turned into an economically producible reserve.

1. IMPLICATIONS FOR GAS HYDRATES

The combined information from Arctic gas hydrate studies shows that, in permafrost regions, gas hydrates may exist at subsurface depths ranging from about 130 to 2000 m. Anomalous seismic reflectors, known as bottom-simulating reflectors (BSRs), have been mapped at depths below the seafloor ranging from about 100 to 1100 m. The presence of gas hydrates in offshore continental margins has been inferred mainly from these BSRs. Although methane, propane, and other gases can be included in the clathrate lattice structure of gas hydrates, methane hydrates appear to be the most common in nature. The amount of methane sequestered in gas hydrates is probably enormous, but estimates of the amounts are speculative and range over three orders of magnitude, from about 3114 trillion to 7,634,000 trillion cubic meters, which represents a very large potential energy resource.

Gas hydrates also represent a significant drilling and production hazard. Russian, Canadian, and American drilling engineers have described numerous problems associated with gas hydrates, including blowouts and well-bore casing failures. Seafloor stability is also an important issues related to gas hydrates. Seafloor stability refers to the susceptibility of the seafloor to collapse and slide as the result of gas hydrate disassociation. It is possible that both natural and anthropogenic changes can contribute to *in situ* gas hydrate destabilization, which may convert a hydrate-bearing sediment to a gassy water-rich fluid, triggering seafloor subsidence and catastrophic landslides. Evidence implicating gas hydrates in triggering seafloor landslides has been found along the Atlantic Ocean margin of the United States. However, the mechanisms controlling gas hydrate-induced seafloor subsidence and landslides are not well understood.

Recent studies indicate that atmospheric methane, a greenhouse gas, is increasing at a rate such that the current concentration will probably double in the next 50 years. Because methane is 21 times more radiatively active than carbon dioxide, it is predicted that methane will surpass carbon dioxide as the predominant atmospheric greenhouse gas in the second half of the next century. The source of this atmospheric methane is uncertain; however, numerous researchers have suggested that destabilized natural-gas hydrates may be contributing to the buildup of atmospheric methane.

2. GAS HYDRATE TECHNICAL REVIEW

Under appropriate conditions of temperature and pressure (Fig. 1), gas hydrates usually form one of two basic crystal structures (Fig. 2). Each unit cell of Structure I gas hydrate (Fig. 2A) consists of 46 water molecules that form two small dodecahedral voids and six large tetradecahedral voids. Structure I gas hydrates can hold only small gas molecules such as methane and ethane, with molecular diameters not exceeding 5.2 Å. The chemical composition of a Structure I gas hydrate can be expressed as $8(Ar,CH_4,H_2S,CO_2)46H_2O$ or $(Ar,CH_4,H_2S,CO_2)5.7H_2O$. The unit cell of Structure II gas hydrate consists of 16 small dodecahedral and 8 large hexakaidecahedral voids formed by 136 water molecules. Structure II gas hydrates may contain gases with molecular dimensions in the range of 5.9 to 6.9 Å, such as propane and isobutane. The chemical composition of a Structure II gas hydrate can be expressed as $8(C_3H_8,C_4H_{10},CH_2Cl_2,CHCL_3)136H_2O$ or $(C_3H_8,C_4H_{10},CH_2Cl_2,CHCL_3)17H_2O$. At conditions of standard temperature and pressure (STP), one volume of saturated methane hydrate (Structure I) will contain as much as 189 volumes of methane gas; because of this large gas storage capacity, gas hydrates are thought to represent an important source of natural gas.

The geologic occurrence of gas hydrates has been known since the mid-1960s, when gas hydrate accumulations were discovered in Russia. Cold surface temperatures at high latitudes on Earth are conducive to the development of onshore permafrost and gas hydrates in the subsurface (Fig. 3). Onshore gas hydrates are known to be present in the West Siberian Basin and are believed to occur in other permafrost areas of northern Russia, including the

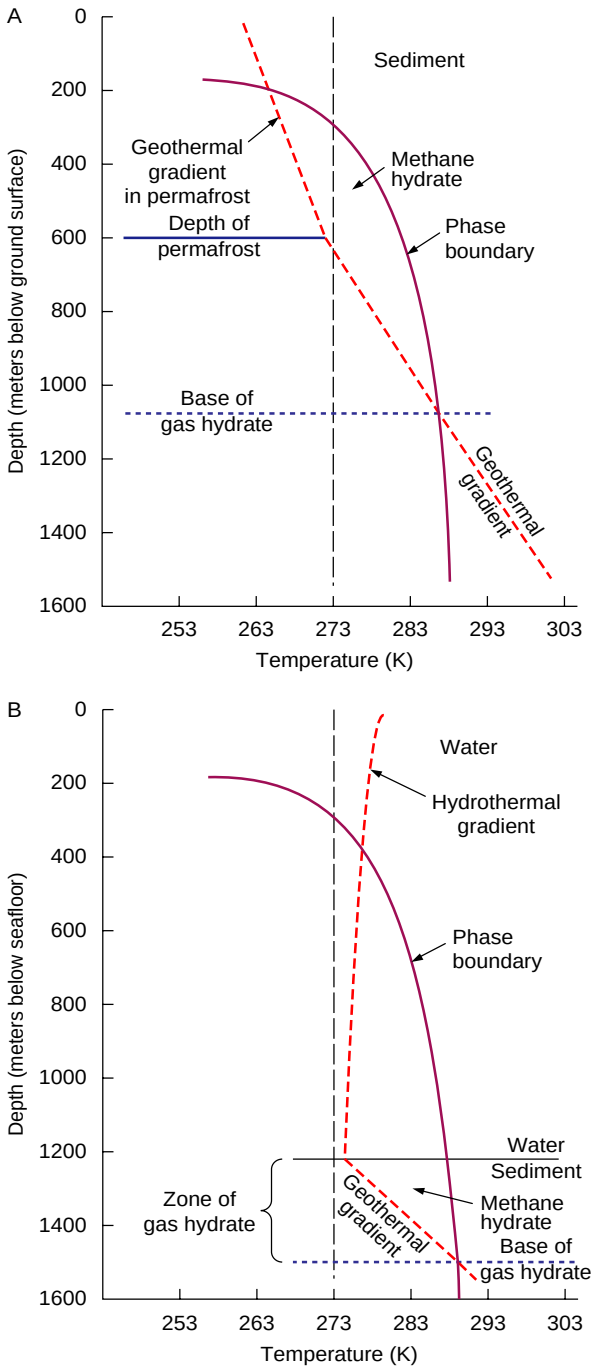


FIGURE 1 Graphs showing the depth–temperature zone in which methane hydrates are stable in (A) a permafrost region and (B) an outer continental margin marine setting.

Timan-Pechora province, the eastern Siberian Craton, and the northeastern Siberia and Kamchatka areas. Permafrost-associated gas hydrates are also present in the North American Arctic. Direct evidence for gas hydrates on the North Slope of Alaska comes from a core test in a gas hydrate exploration well, and

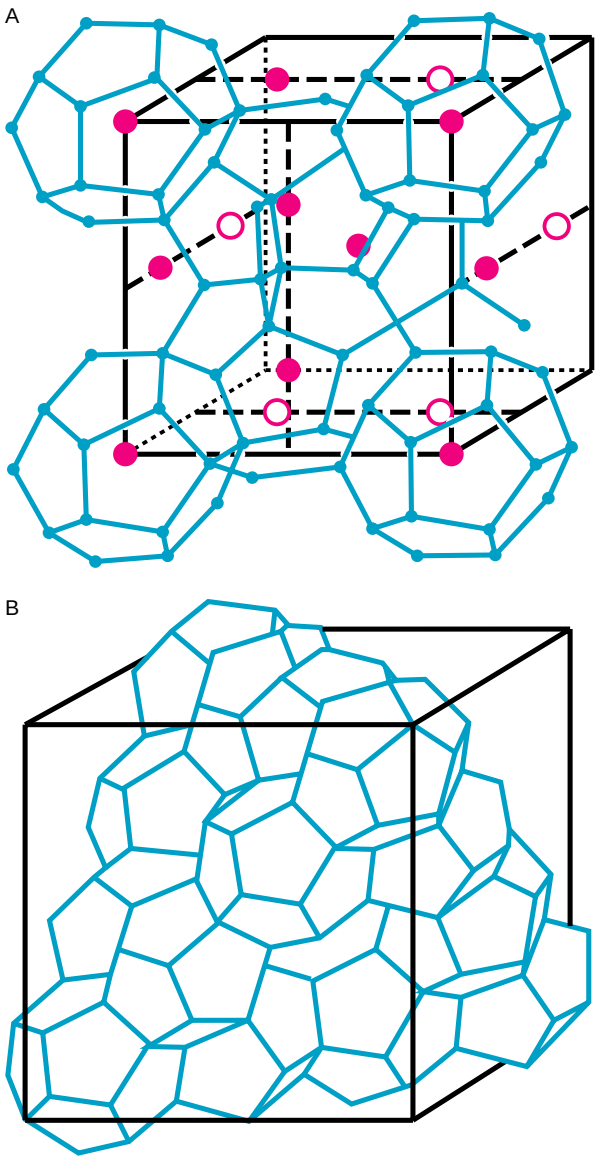


FIGURE 2 Two gas hydrate crystal structures: (A) Structure I and (B) Structure II.

indirect evidence comes from drilling and open-hole industry well logs that suggest the presence of numerous gas hydrate layers in the area of the Prudhoe Bay and Kuparuk River oil fields. In Canada, well-log responses attributed to the presence of gas hydrates have been obtained in about one-fifth of the wells drilled in the Mackenzie Delta, and more than half of the wells in the Arctic Islands are inferred to contain gas hydrates. The recently completed Mallik 2L-38 gas hydrate research well confirmed the presence of a relatively thick, highly concentrated, gas hydrate accumulation on Richards Island in the outer portion of the Mackenzie River Delta.

In marine environments, gas hydrates have been recovered in gravity cores within 10 m of the seafloor in the sediment of the Gulf of Mexico, the offshore portion of the Eel River Basin of California, the Black Sea, the Caspian Sea, and the Sea of Okhotsk. Also, gas hydrates have been recovered at greater subbottom depths during research coring along the southeastern coast of the United States on the Blake Ridge, in the Gulf of Mexico, in the Cascadia Basin near Oregon, in the Middle America Trench, on offshore Peru, and on both the eastern and western margins of Japan.

3. *ENERGY RESOURCE POTENTIAL OF GAS HYDRATES*

World estimates for the amount of natural gas in gas hydrate deposits range from 14 trillion to 34,000 trillion cubic meters for permafrost areas and from 3100 trillion to 7,600,000 trillion cubic meters for oceanic sediments. Current estimates of the amount of methane in the world's gas hydrate accumulations are in rough accord at about 20,000 trillion cubic meters. If these estimates are valid, the amount of methane in gas hydrates is almost two orders of magnitude larger than the estimated total remaining recoverable conventional methane resources, estimated to be about 250 trillion cubic meters.

The 1995 National Assessment of United States Oil and Gas Resources, conducted by the U.S. Geological Survey, focused on assessing the undiscovered conventional and unconventional resources of crude oil and natural gas in the United States. This

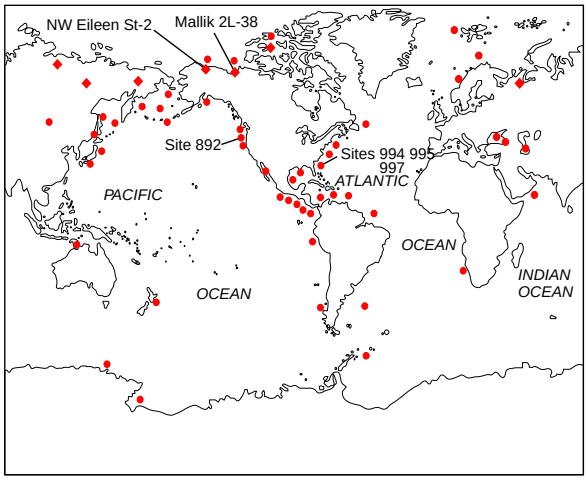


FIGURE 3 Locations of known and inferred gas hydrate occurrences.

assessment included for the first time a systematic resource appraisal of the in-place natural gas hydrate resources of the onshore and offshore regions of the United States. In this assessment, 11 gas hydrate plays were identified within four offshore and one onshore gas hydrate provinces. The offshore gas hydrate provinces assessed lie within the U.S. Exclusive Economic Zone adjacent to the lower 48 states and Alaska. The only onshore province assessed was the North Slope of Alaska. In-place gas resources within the gas hydrate regions of the United States are estimated to range from about 3200 trillion to 19,000 trillion cubic meters of gas, at the 0.95 and 0.05 probability levels, respectively. Although this wide range of values shows a high degree of uncertainty, it does indicate the potential for enormous quantities of gas stored as gas hydrates. The mean in-place value for the entire United States is calculated to be about 9000 trillion cubic meters of gas, which is a very large volume of gas, when it is considered that the United States has only about 7 trillion cubic meters of conventional natural gas reserves.

Even though gas hydrates are known to occur in numerous marine and Arctic settings, little is known about the technology necessary to produce gas hydrates. Most of the existing gas hydrate "resource" assessments do not address the problem of gas hydrate recoverability. Proposed methods of gas recovery from hydrates (Fig. 4) usually deal with dissociating or "melting" gas hydrates *in situ* by (1) heating the reservoir beyond hydrate formation temperatures, (2) decreasing the reservoir pressure below hydrate equilibrium, or (3) injecting an inhibitor, such as methanol or glycol, into the reservoir to decrease hydrate stability conditions. Gas recovery from hydrates is hindered because the gas is in a solid form and because hydrates are usually widely dispersed in hostile Arctic and deep marine environments. First-order thermal stimulation computer models (incorporating heat and mass balance) have been developed to evaluate hydrate gas production from hot water and steam floods; the models have shown that gas can be produced from hydrates at sufficient rates to make gas hydrates a technically recoverable resource. However, the economic cost associated with these types of enhanced gas recovery techniques would be prohibitive. Similarly, the use of gas hydrate inhibitors in the production of gas from hydrates has been shown to be technically feasible, but the use of large volumes of chemicals such as methanol comes with a high economic and environmental cost. Among the various techniques for production of natural gas from *in situ* gas hydrates, the most economically promising method is considered to be the depressurization technique. However, the extraction of gas from a gas hydrate accumulation by depressurization may be hampered by the formation of ice and/or the reformation of gas hydrate due to the endothermic nature of gas hydrate dissociation.

The Messoyakha gas field in the northern part of the Russian West Siberian Basin is often used as an example of a hydrocarbon accumulation from which gas has been produced from *in situ* natural-gas hydrates. Production data and other pertinent geologic information have been used to document the presence of gas hydrates within the upper part of the Messoyakha field. It has also been suggested that the production history of the Messoyakha field demonstrates that gas hydrates are an immediately producible source of natural gas, and that production can be started and maintained by conventional methods. Long-term production from the gas hydrate part of the Messoyakha field is presumed to have been achieved by the simple depressurization scheme. As production began from the lower free-gas portion of the Messoyakha field in 1969, the measured reservoir pressures followed predicted decline relations; however, by 1971, the reservoir pressures

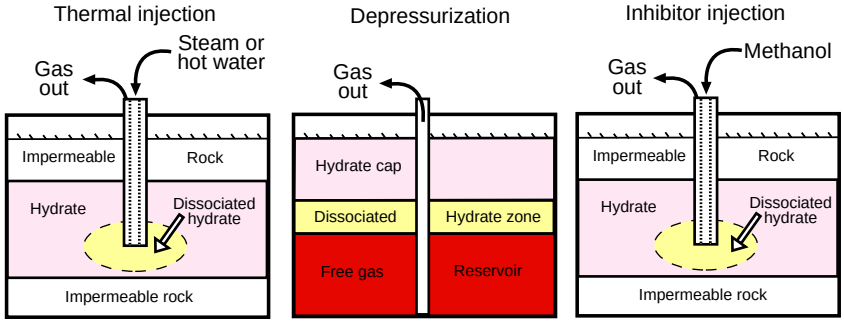


FIGURE 4 Schematic of proposed gas hydrate production methods.

began to deviate from expected values. This deviation has been attributed to the liberation of free gas from dissociating gas hydrates. Throughout the production history of the Messoyakha field, it is estimated that about 36% (about 5 billion cubic meters) of the gas withdrawn from the field has come from the gas hydrates. However, several recent studies suggest that gas hydrates may not be significantly contributing to gas production in the Messoyakha field.

It should be noted that the current assessment of proposed methods for gas hydrate production does not consider some of the more recently developed advanced oil and gas production schemes. For example, the usefulness of downhole heating methods such as *in situ* combustion, electromagnetic heating, or downhole electrical heating have not been evaluated. In addition, advanced drilling techniques and complex downhole completions, including horizontal wells and multiple laterals, have not been considered in any comprehensive gas hydrate production scheme. Gas hydrate provinces with existing conventional oil and gas production may also present the opportunity to test relatively more advanced gas hydrate production methods. For example, in northern Alaska, existing "watered-out" production wells are being evaluated as potential sources for hot geopressured brines that will be used to stimulate gas hydrate production thermally.

It has been suggested that some of the methane liberated from a producing gas hydrate field could be lost to the atmosphere; methane is believed to be a significant greenhouse gas. It is possible that large-scale production of gas from hydrates could contribute methane to the atmosphere; however, the rates of gas release from a hydrate accumulation would probably be similar to those associated with conventional natural gas fields. In addition, any modern gas hydrate field completion plan would attempt to monitor and remediate any potential accidental release of gas.

Because of uncertainties about the geologic settings and feasible production technology, few economic studies have been published on gas hydrates. The National Petroleum Council (NPC), in its major 1992 study of gas, published one of the few available economic assessments of gas hydrate production (Table I). This information assessed the relative economics of gas recovery from hydrates using thermal injection and depressurization. It also benchmarked the cost of gas hydrate production with the costs of conventional gas production on Alaska's North Slope. Since the 1992 NPC report, the cost of conventional gas production has declined. However, within countries with considerable production of cheaper conventional natural gas, hydrates appear not to be an economically viable energy resource in a competitive energy market. However, in certain parts of the world with unique economic and/or political motivations, gas hydrates may become a critical sustainable source of natural gas within the foreseeable future.

TABLE I

Economic Study of Gas Hydrate Production

Economic factor[a]	Thermal injection[b]	Depressurization	Conventional gas
Investment ($\times 10^6$\$)	5084	3320	3150
Annual cost ($\times 10^6$\$)	3200	2510	2000
Total production (MMcf/year)[c]	900	1100	1100
Production cost (\$/Mcf)	3.60	2.28	1.82
Break-even wellhead price (\$/Mcf)	4.50	2.85	2.25

[a] Mcf, thousand cubic feet; MMcf, million cubic feet.

[b] Assumed process: injection of 30,000 barrels/day of water at 150°C.

[c] Assumed reservoir properties: h (reservoir thickness) = 7.3 m, ϕ (reservoir porosity) = 40%, k (reservoir permeability) = 600 md.

SEE ALSO THE FOLLOWING ARTICLES

Natural Gas Resources, Global Distribution of • Natural Gas Resources, Unconventional • Oil and Natural Gas Exploration • Oil and Natural Gas: Offshore Operations • Oil and Natural Gas Resource Assessment: Geological Methods • Oil Sands and Heavy Oil • Oil Shale • Polar Regions, Impacts of Energy Development • Remote Sensing for Energy Resources

Further Reading

Cherskiy, N. V., Tsarev, V. P., and Nikitin, S. P. (1985). Investigation and prediction of conditions of accumulation of gas resources in gas-hydrate pools. *Petrol. Geol.* **21,** 65–89.

Collett, T. S. (1993). Natural gas hydrates of the Prudhoe Bay and Kuparuk River area, North Slope, Alaska. *Am. Assoc. Petrol. Geol. Bull.* **77,** 793–812.

Collett, T. S. (1995). Gas hydrate resources of the United States. *In* "1995 National Assessment of United States Oil and Gas

Resources" (D. L. Gautier, G. L. Dolton, K. I. Takahashi, and K. L. Varnes, Eds.). Available on CD-ROM from the U.S. Geological Survey, Digital Data Series 30.

Dallimore, S. R., Uchida, T., and Collett, T. S. (1999). Scientific results from JAPEX/JNOC/GSC Mallik 2L-38 gas hydrate research well, Mackenzie Delta, Northwest Territories, Canada. *Geol. Surv. Can. Bull.* **544,** 1–403.

Holbrook, W. S., Hoskins, H., Wood, W. T., Stephen, R. A., and the ODP Leg 164 Scientific Party. (1996). Methane hydrate, bottom-simulating reflectors, and gas bubbles: Results of vertical seismic profiles on the Blake Ridge. *Science* **273,** 1840–1843.

Kvenvolden, K. A. (1988). Methane hydrate—A major reservoir of carbon in the shallow geosphere? *Chem. Geol.* **71,** 41–51.

Kvenvolden, K. A. (1993). Gas hydrates as a potential energy resource—A review of their methane content. *In* "The Future of Energy Gases" (D. G. Howell, Ed.), pp. 555–561. U.S. Geological Survey Professional Paper 1570. USGS, Reston, Virginia.

Kvenvolden, K. A., and Barnard, L. A. (1983). Hydrates of natural gas in continental margins. *In* "Studies in Continental Margin Geology: American Association of Petroleum Geologists, Memoir 34" (J. S. Watkins and C. L. Drake, Eds.), pp. 631–641. AAPG, Tulsa, Oklahoma.

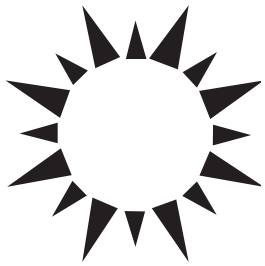
MacDonald, G. T. (1990). The future of methane as an energy resour. *Annu. Rev. Energy* **15,** 53–83.

Makogon, Y. F. (1981). "Hydrates of Natural Gas." PennWell Publ. Company, Tulsa, Oklahoma.

Max, M. D. (2000). "Natural Gas Hydrate in Oceanic and Permafrost Environments." Kluwer Academic Publ., London.

Sloan, E. D. (1998). "Clathrate Hydrates of Natural Gases." 2nd Ed. Marcel Dekker, New York.

Gasoline Additives and Public Health



SERAP ERDAL
University of Illinois at Chicago
Chicago, Illinois, United States

1. Introduction to Gasoline Additives
2. Human Health Effects of Methyl Tertiary-Butyl Ether (MTBE)
3. Human Health Effects of Other Ether Compounds
4. Human Health Effects of Ethanol
5. Conclusions and Recommendations

1. INTRODUCTION TO GASOLINE ADDITIVES

Since 1979, oxygenates as gasoline additives have been used in limited areas of the United States as octane enhancers to replace lead at levels around 2 to 8% by volume. During the 1980s, oxygenates came into wider use as some states implemented oxygenated gasoline programs for the control of carbon monoxide air pollution in cold winter. Oxygenates were added to conventional gasoline nationally at higher percentages with the passage of the 1990 Clean Air Act (CAA), which required that oxygenates be added either seasonally (15% MTBE by volume) or year-round (11% MTBE by volume) to gasoline in specific parts of the country where carbon monoxide in the winter or concentrations of ozone in the summer exceed their respective National Ambient Air Quality Standards (NAAQS). Title II of the 1990 CAA addresses mobile sources and mandates the following two types of oxygenated fuel programs:

- Section 211 (m) of the CAAA of 1990 directed the EPA to mandate that states establish a wintertime oxygenated fuel program in 42 geographic areas designated as moderate or serious CO nonattainment areas (i.e., areas with CO concentrations exceeding the CO NAAQS) by November 1, 1992. The oxygenated fuel program was mandated during the winter months of November through February based on the fact that gasoline combustion is less efficient at colder temperatures, resulting in increased CO emissions, and that pollutants stay closer to ground level in winter. During these months, the oxygen content requirement for gasoline sold in the designated areas is a minimum of 2.7 percent.
- Title II of the 1990 CAA under Section 211(k) specified that only reformulated gasoline (*RFG*) with a 2% oxygen weight requirement can be sold in metropolitan areas with populations more than 250,000, classified as extreme or severe for ozone pollution. These included Los Angeles, New York, Baltimore, Chicago, Milwaukee, Houston, San Diego, Sacramento, Philadelphia, and Hartford. Less-polluted nonattainment areas for ozone opted to implement the RFG program since its initiation in January 1995. During Phase I of the RFG program, refiners were required to reduce vehicle emissions of ozone-forming volatile organic compounds (VOCs) and air toxics by 15 to 17% annually, compared to 1990 baseline averages. The benzene and the aromatic content of RFG could not exceed 1.0% and 25% by volume, respectively. Nitrogen oxides (NO_x) emissions were not allowed to increase relative to conventional "baseline" gasoline. During Phase II, a 25 to 29% reduction in VOC was required, as well as a 20 to 22% reduction in air toxics and a 5 to 7% reduction in NO_x emissions from the 1990 baseline.

Among the available oxygenates in the marketplace, methyl tertiary butyl ether (MTBE) is the most widely used because of its low vapor pressure, compatibility with the blending and distribution system for gasoline, high octane rating, availability

 © 2004 Elsevier Inc. All rights reserved.

and ease of production, and low cost. MTBE had the second highest volume of all chemicals produced in the United States in 1997.

On the other hand, ethanol is the predominant oxygenate used in gasoline in some areas, particularly in the Midwest. Other fuel oxygenates that are in use or may potentially be used to meet the 1990 CAA oxygenated fuel requirements include ethyl tertiary butyl ether (ETBE), tertiary-amyl methyl ether (TAME), diisopropyl ether (DIPE), tertiary-butyl alcohol (TBA), and methanol. This article assesses the human health effects of MTBE, ETBE, TAME, DIPE, and ethanol based on potential for exposure to these compounds as a result of using oxygenates in the U.S. gasoline supply. The priority is given to human studies in analyzing health effects information related to each of these compounds in gasoline.

In the 1992–1993 CO season, MTBE-blended gasoline accounted for approximately two-thirds of the demand, with ethanol providing the remainder. According to the EPA, ethanol is the predominant oxygenate used in the oxyfuel program (about 7.5% and 15% by volume for ethanol and MTBE, respectively) to meet the 2.7% by weight oxygen requirement. On the other hand, MTBE at 11% by volume is the primary oxygenate used in RFG to meet the 2% by weight oxygen requirement. The EPA's Office of Mobile Sources estimated in 1994 that approximately 30 million people live in areas using wintertime oxygenated gasolines and about 45 million live in high ozone RFG areas. Later estimates by SRA Technologies suggest that 70 and 57 million people live in areas using oxyfuel and RFG, respectively. It is also reported that about 109 million Americans live in counties where MTBE is used as oxygenate in gasoline. There is some overlap between the areas implementing the winter oxyfuel program and the areas implementing the year-round RFG program. Estimated oxygenated volume consumed in RFG areas in the United States is 237 (MTBE), 0.3 (ETBE), 14.6 (TAME), and 24.2 (ethanol) thousand barrels per calendar day. These estimates suggest that a large fraction of the U.S. population may get exposed to oxygenates on a daily basis.

2. HUMAN HEALTH EFFECTS OF METHYL TERTIARY-BUTYL ETHER (MTBE)

Acute health symptoms attributed to methyl tert-butyl ether (MTBE) were reported in various parts of the United States soon after the addition of oxygenated compounds to gasoline during the winter of 1992–1993. Such complaints were not anticipated but have subsequently focused attention on possible health risks due to oxygenated gasoline. The health complaints were accompanied by complaints about reduced fuel economy and engine performance and by controversial cancer findings in long-term animal studies.

2.1 Epidemiological Evidence for Air Exposures

Following the introduction of MTBE-oxygenated gasoline in Fairbanks, Alaska, in November 1992, about 200 residents reported headaches, dizziness, irritated eyes, burning of the nose and throat, coughing, disorientation, nausea, and other acute health symptoms. The Alaska Department of Health and Social Services and the Center for Disease Control (CDC) conducted a number of field epidemiology studies to investigate the potential relationship between MTBE oxyfuel exposure and acute health symptom responses. A two-phase study was performed in Fairbanks, with Phase I occurring while MTBE-oxyfuel was in use and Phase II occurring 6 weeks after MTBE-oxyfuel use had been terminated. In this study, the relationship between MTBE oxyfuel exposure and blood levels of MTBE and a major metabolite of MTBE, tertiary butyl alcohol (TBA), was also investigated. The occupational and non-occupational exposure groups answered a questionnaire inquiring about the presence of certain symptoms, including the seven key complaints of headache, eye irritation, burning of nose/throat, cough, nausea, dizziness, and spaciness. Occupationally exposed individuals reported a greater prevalence of health complaints during Phase I compared to Phase II. Overall, the CDC's analysis of Fairbanks data showed that there was a statistically significant difference between the number of cases among the individuals who were exposed to higher concentration of MTBE (e.g., taxi drivers) compared to people exposed to lower exposure levels (e.g., students). The CDC cautioned that the results did not provide definitive evidence that the use of MTBE oxyfuels caused the reported symptoms due to potential confounding factors (e.g., negative publicity about MTBE-oxyfuel effects, small sample size, strong odor associated with MTBE, and lack of controls) and called for further research to determine whether there actually is an increase in acute health symptoms and, if so, whether they are causally associated

with MTBE-oxyfuel. However, the governor of Alaska suspended the program in Fairbanks within 2 months of the introduction of oxyfuel and a switch from MTBE-oxyfuel to ethanol-oxyfuel occurred in following years.

A desire to conduct a study similar to the Fairbanks study in the absence of confounding factors observed in Alaska (especially negative publicity) led to a field study in another part of the country. The CDC compared the prevalence of acute health symptom reports in Stamford, Connecticut, in April 1993 (where MTBE-oxyfuel was used) and Albany, New York, in May 1993 (with no oxyfuels). There were no known consumer health complaints in either location. For the Stamford study, CDC in cooperation with the State of Connecticut Health Department selected 221 individuals from four groups of workers and administered a questionnaire similar to that used in Fairbanks. The CDC reported that the 11 individuals with the highest blood MTBE levels were statistically significantly more likely to report one or more key symptoms than the other 33 people with lower blood levels. However, there was no apparent relationship between symptom prevalence by occupational category and the median blood MTBE or the TBA concentrations associated with each category. CDC again cautioned that the small numbers of subjects studied, the wide range of blood MTBE and TBA levels within each category, and the wide variability of response prevented a definitive conclusion. For the Albany study, 264 individuals were nonrandomly selected and categorized into three groups presumed to have tiered levels of exposure. Overall, the results indicated that the Albany residents with no MTBE-oxyfuel exposure had symptom prevalence as high as those in Stamford residents with MTBE-oxyfuel exposure. It was suggested that this outcome could potentially be attributed to the high prevalence of symptoms due to allergies and viruses at the time of the study in Albany.

In New Jersey, the EOHSI evaluated 237 state-employed garage workers from two groups in 1993: (1) 115 northern New Jersey workers sampled during the wintertime oxyfuel program and (2) 122 southern New Jersey workers sampled 10 weeks after the phase-out of the program in that area. All participants were asked to complete questionnaires about symptoms and their severity over the previous 30 days and about preshift and postshift symptoms. Essentially no differences in symptom reports were found between the northern (high-exposure) and southern (low-exposure) groups. In a separate study, 11 workers who pumped MTBE-oxyfuel for more than 5 hours per day were matched with 11 workers from the south who had low or nondetectable exposures to MTBE. No differences in symptom prevalence were found between the groups.

Another EOHSI study attempted to identify a sensitive subpopulation for the effects of MTBE-oxyfuel exposure. This involved 13 subjects with multiple chemical sensitivities (MCS), five subjects with chronic fatigue syndrome (CFS), and six healthy controls. Through a symptom questionnaire, these subjects were asked to assess symptoms associated with situations in which they could be exposed to MTBE such as refueling and driving. On a total symptom score range of 0 (no discomfort) to 28 (severe discomfort), the MCS subjects scored 5.3, compared to 4.8 for the CFS subjects and 1.2 for the healthy subjects. It was not possible to draw any definitive conclusions from this study due to (1) lack of evidence to support that an unusually high rate of symptoms or an increase of symptoms was occurring uniquely where MTBE was most prevalent, (2) small sample sizes, and (3) inability to distinguish between those subjects using and not using MTBE oxyfuel.

In response to variety of health concerns of Milwaukee citizens after the beginning of the RFG program in January 1995, the state of Wisconsin conducted a random-digit-dial telephone survey of 500 residents of the Milwaukee metropolitan area and equal numbers of residents in Chicago and in non-RFG areas of Wisconsin. In Milwaukee, oxygenates in the RFG samples were 49% MTBE, 38% ethanol, and 14% ETBE. Symptom prevalence was significantly higher in Milwaukee than in either Chicago or elsewhere in Wisconsin (23% compared to 6%). In Milwaukee, persons were more likely to report symptoms if they had experienced a cold or flu, smoked cigarettes, or were aware that they had purchased RFG. Because overall symptom prevalence was essentially equivalent in Chicago and non-RFG areas of Wisconsin, the authors concluded that factors other than RFG use such as an increase in price of gasoline, odors of MTBE, and negative media coverage significantly contributed to the differences in symptom prevalence between Milwaukee and the other two areas. However, due to study design limitations, the study could not "rule out subtle effects of RFG exposure, or the possibility that a relatively small number of individuals may have a greater sensitivity to RFG mixtures."

Hakkola and his coworkers investigated neurological effects reported by 101 Finnish male tank drivers with exposure to gasoline containing 10%

MTBE. Drivers with longer exposure scored higher for fatigue and hostility changes as compared to drivers with shorter exposure. In addition, 20% of drivers reported the following symptoms in the order of decreasing frequency: headache, nausea, increased salvation, dizziness, and dyspnea.

2.2 Controlled Human Inhalation Exposure Chamber Studies

Assessment of humans exposed directly in a chamber to MTBE was hoped to answer whether there was a direct causal relation between MTBE exposure and acute health symptoms, particularly in view of the inconclusive evidence gathered from the field epidemiology studies. In an EPA investigation, 37 healthy nonsmoking individuals were exposed for 1 h to both clean air and 1.4 ppm (5 mg/m^3) "pure" MTBE in air at 75° F and 40% humidity. The endpoints selected, based on observations from Alaska, were the types of symptoms associated with organic solvent exposure: (1) nonspecific symptoms including headache, nasal irritation, throat irritation, cough, eye irritation, odor quality, and dizziness; (2) measures of behavioral response (neurobehavioral evaluation system test battery); (3) indicators of upper airways inflammation (nasal lavage); and (4) indicators of eye irritation. The study found that MTBE exposure had no statistically significant effects on any of the endpoints. Yale University researchers replicated the EPA study, with few minor modifications—that is, 43 subjects, slightly higher MTBE concentration of 1.7 ppm (6 mg/m^3), and an additional experiment involving a 1-h exposure to a complex mixture of 16 VOCs commonly found in gasoline at a concentration of 7.1 ppm without MTBE. Similar to the EPA study, it was found that MTBE exposure had no statistically significant effects on any of the endpoints studied as compared to clean air exposure. The VOC exposure caused an increase in inflammatory cells in the nasal lavage on the day following exposure when the VOC exposure compared to clean air exposure. Taken together, the EPA and Yale investigations showed that controlled exposure of a small number of healthy humans to MTBE in air under the conditions studied did not cause increased symptoms or measurable responses. Exposure concentrations utilized in these controlled exposure studies (1.4 and 1.7 ppm) did not represent the worst-case exposure conditions based on measurements from the customer breathing zone samples taken in air during refueling (median exposure range: 0.3 to 6 ppm, occasionally higher). Nor were subjects chosen who were seemingly susceptible to MTBE based on their complaints.

Another controlled exposure chamber study exposed normal subjects to much higher concentrations of MTBE (5, 15, and 25 ppm compared to 1.4 and 1.7 ppm) and a longer exposure duration (2 hours) than the EPA or the Yale study (1 hour). This study did not include neurobehavioral assessments. No effects associated with MTBE exposure were identified. In a separate study, 13 healthy male subjects were exposed to 25 and 75 ppm MTBE in air in an exposure chamber. Subjects with 75 ppm exposure developed irritation of mucous membranes and nonspecific symptoms indicative of effects on mood.

A controlled exposure study was conducted at the EOHSI to investigate the symptoms, psychophysiologic and neurobehavioral effects of exposure to the MTBE. The exposures were conducted at concentrations similar to those found while refueling a vehicle using both control subjects and those individuals self-described as sensitive to MTBE. In the double-blinded repeated-measures controlled exposure study, subjects were exposed for 15 min each to clean air, gasoline alone, gasoline with 11% MTBE, and gasoline with 15% MTBE. Relative to the control group ($n = 19$), sensitive individuals ($n = 12$) reported significantly more total symptoms only when exposed to gasoline with 15% MTBE and not to gasoline with 11% MTBE or clean air. However, the sensitive group did not exhibit significant differences in psychophysiologic responses from the control group at any exposure level. This study did not identify a dose-response relationship for MTBE exposure and symptoms were specific to MTBE. The EOHSI research group also studies the driving habits and vehicle maintenance patterns of these two study groups. Results showed that the individuals who had self-reported heightened sensitivity to oxygenated gasoline vapors drove their vehicles more often and fueled their vehicles more frequently than the control group. In addition, sensitive individuals more frequently drove their vehicles with compromised body integrity that could lead to higher exposures.

2.3 Health Effects Associated with Oral Exposure to MTBE

In the mid-1990s, the focus on MTBE began to shift from questions about air quality-related health risks and benefits to questions about water quality risks. The change was triggered by a number of

factors including the USGS studies demonstrating contamination of the nation's water supply with MTBE, and discovery of underground storage tank (UST) leaks resulting in high concentrations of MTBE in water causing the closure of drinking water wells in Santa Monica, South Tahoe, and San Jose, California. The emergence of widespread reports of MTBE-impacted drinking water resources in other states (e.g., New Jersey, Maine, Vermont, New York, Virginia) heightened interest in protection of water resources from MTBE. The presence of MTBE in ground water and in reservoirs used for drinking water in other parts of the country is now relatively widespread.

Faced with an acute potential public health problem and degradation of water quality (e.g., strong odor and taste associated with MTBE) resulting in loss of water supplies for consumer consumption, the California legislature enacted a number of bills calling to address issues pertaining to oxygenates including risks and benefits to human health and the environment of MTBE and its combustion byproducts in air, water, and soil and alternate oxygenates like ethanol and ETBE. In response to these legislative bills, the Office of Environmental Health Hazard Assessment (OEHHA) of the California Environmental Protection Agency proposed a Public Health Goal (PHG) of 14 ppb for MTBE in drinking water in June 1998 to protect consumers against health risks over a lifetime of exposure. The proposed PHG for MTBE was based on carcinogenic potency of 1.7×10^{-3} (mg/kg-day)$^{-1}$, which is the geometric mean of the cancer slope factor estimates obtained from three cancer sites: combined male rat kidney adenomas and carcinomas in the 1992 inhalation study of Chun *et al.*, the male rat Leydig cell tumors and the leukemia, and lymphomas in female rats in the 1995 oral study by Belpoggi *et al.* In July 1998, the Department of Health Services (DHS) adopted a secondary maximum contaminant level (SMCL) for MTBE of 5 ppb based on available data on odor and taste thresholds. Another product was a University of California (UC) report on "Health and Environmental Effects of MTBE," submitted to the governor and legislature of the state of California on November 12, 1998. The UC report states "For the general population, the risk of exposure to MTBE through ingestion of MTBE-contaminated water is currently low." However, because MTBE is highly soluble in water, poorly adsorbs to soil, readily transfers to groundwater from LUSTs, and pipelines, travels much further than plumes commonly associated with gasoline spills, and is highly resistant to biodegradation, the report concluded that "There are significant risks and costs associated with water contamination due to the use of MTBE." Most importantly, the UC scientists recommended phasing out MTBE over an interval of several years instead of an immediate ban. Based on significant risk to California's environment associated with the continued use of MTBE in gasoline, Governor Davis, on March 25, 1999, directed appropriate state regulatory agencies to devise and carry out a plan to begin immediate phase out of MTBE from California gasoline, with 100% removal achieved no later than December 31, 2002.

Faced with another MTBE problem, in 1996, the EPA began working on revising the 1992 drinking water lifetime health advisory for MTBE of 20–200 ppb to meet the needs of states that were struggling to endorse safe levels that ensure both consumer acceptability of MTBE-laden drinking water and public health protection against potential adverse health effects of ingested MTBE. Meanwhile, in the absence of enforceable federal drinking water standards, individual states began to adopt action levels or cleanup levels for MTBE in water. In December 1997, the EPA's Office of Water published "Drinking Water Advisory: Consumer Acceptability Advise and Health Effects Analysis on Methyl Tertiary-Butyl Ether (MTBE)." The advisory recommended that keeping MTBE levels in the range of 20 to 40 ppb or below to protect consumer acceptance of the water resource would also provide a large margin of safety from potential adverse health effects. The EPA could not develop a RfD for MTBE due to the persisting limitations of health effects data (e.g., lack of a pharmacokinetic model for converting inhalation exposure data for MTBE to oral exposure, limited reported pathology and historical tumor incidence, no toxicity studies with oral exposure to water). A blue-ribbon panel convened by the United States Environmental Protection Agency (EPA) recommended that Congress should act to provide clear federal and state authority to regulate or eliminate the use of MTBE and other gasoline additives that threaten drinking water supplies.

2.4 Carcinogenic Effects of MTBE

There are no human data on which carcinogenicity of MTBE can be evaluated. However, the evidence from chronic exposure by inhalation and oral exposure routes demonstrate that MTBE is

carcinogenic in rats and mice. Two chronic animal cancer bioassays with inhalation exposure to MTBE resulted in rare kidney tumors and testicular tumors in the mid- and high-dose groups of male rates, and liver tumors in mice at the high-dose group. Sprague-Dawley rats exposed to MTBE in olive oil as the vehicle for oral gavage for 2 years showed dose-related increases in lymphomas and leukemias in females and in interstitial cell tumors of the testes in males. The biological mechanism responsible for MTBE cancer production in animals is not known. The lack of mechanistic information limits our ability to determine whether the MTBE bioassay results are relevant to humans. In 1998, IARC found insufficient evidence to establish that MTBE poses cancer risks to humans. The EPA has also concluded that potential for MTBE's carcinogenicity for humans is of concern and requires further study.

3. HUMAN HEALTH EFFECTS OF OTHER ETHER COMPOUNDS

The other aliphatic ethers, ethyl tertiary butyl ether (ETBE), tertiary amyl methyl ether (TAME), and diisopropyl ether (DIPE) are proposed as potential alternatives to MTBE as an oxygenate in gasoline. In particular, TAME was considered to be an attractive alternative due to its lower vapor pressure as compared to MTBE despite its lower octane content. These compounds are being used in varying percentages in gasoline in the United States (e.g., 19.90% TAME in Sussex County, Delaware, 1.6% TAME in San Diego, California, in 1999).

Although a number of toxicity studies in animal models for other ether oxygenates of ethyl tertiary butyl ether (ETBE), tertiary amyl methyl ether (TAME), and diisopropyl ether have been conducted and published, human relevance of these findings is not known. There appears to be qualitative similarities to the MTBE health effects, with kidney and liver being the most sensitive target organs. Biotransformation and pharmacokinetics of these compounds in animals and human volunteers have been studied except DIPE. A summary of animal studies and pharmacokinetic studies in animals and humans can be found in Ahmed's 2001 research.

The EPA has not been able to establish any quantitative cancer potency factor or noncancer toxicity values (i.e., reference dose [RfD] or reference concentration [RfC]) to date due to inability to establish dose-response relationships from the animal data in a scientifically defensible manner. To fill this data gap, the EPA mandated the manufacturers to perform toxicity testing of evaporative emissions of these compounds under the CAA Section 211(b) by 2004. The EPA also called for "comprehensive evaluation of multimedia, life-cycle environmental impacts" of different fuels to assess viability of TAME, ETBE, and DIPE to be a safe replacement for ethanol or MTBE. A number of scientific bodies convened to evaluate health effects of oxygenated gasolines more than two decades also repeatedly called for more research into understanding the health effects of these compounds alone and in combination with gasoline. The California DHS proposed to add ETBE and TAME to the list of unregulated chemicals for which monitoring is required by public water systems.

4. HUMAN HEALTH EFFECTS OF ETHANOL

Ethanol used as oxygenate in gasoline is produced from grains or other renewable agricultural and forestry feedstocks. The U.S. Congress has approved a 5.4 cents/gallon federal subsidy for this biomass-based source of fuel in nations' gasoline supply. Ethanol production in 2000 was 1.5 billion gallons. Approximately 30% of this amount is used in RFG during the summer in ozone nonattainment areas. About 20% is used in the winter oxygenate program to control CO air pollution. The remainder 50% is used in conventional gasoline in the United States to enhance octane and extend fuel supplies. Ethanol is highly volatile when spilled; it will evaporate five times faster than MTBE. However, it is commonly used in many modern day products including pharmaceuticals, mouthwash products, alcoholic beverages, and cleaning products. Because humans frequently ingest fermented beverages containing 12% ethanol by volume, ingestion-related health effects of ethanol are well known. The Health Effects Institute (HEI) in its assessment of health effects of oxygenates stated that our knowledge on health effects of ethanol emanates from ingestion exposures. Ethanol was shown to be associated with increased rates of tumors of digestive track with oral exposure to moderate levels of alcohol when other carcinogens were present. Ethanol is classified as a Class I carcinogen by IARC based on human data for alcoholic beverage consumption. There is no data of such association for inhalation exposures. However,

ethanol, when blended in gasoline increases volatility of other compounds, results in an increase in evaporative emissions of VOCs, including benzene. Many of these VOCs are known or suspected carcinogens and also have adverse health effects associated with noncancer health endpoints. Ethanol exposure is also associated with cancers in the liver, esophagus, oral cavity, pharynx, and larynx. There have been reports of adverse reproductive and developmental effects of associated with exposure to ethanol in animal and human studies. Irvine analyzed the relevance of the developmental toxicity of ethanol in occupational settings and concluded that while developmental toxicity may result from drinking alcoholic beverages for which threshold levels have yet to be determined there was no evidence for developmental toxicity in occupational environment.

With the executive order to phase out of MTBE in gasoline distributed in California by December 31, 2002, Governor Davis also directed CARB and the State Water Resources Board to conduct an environmental fate and transport analysis of ethanol in air, surface water, and groundwater and the OEHHA to prepare an analysis of health risks associated with exposure to ethanol in gasoline. In terms of air quality effects, the substitution of ethanol for MTBE in CA's fuel supply was found to have not significantly impact air quality as long as RFG regulation in CA addressed the potential for ethanol to increase evaporative emissions and to cause more rail and truck traffic to bring ethanol to the state. The CA analysis found no substantial differences in the air-related public health impacts of different non-MTBE fuel formulations including ethanol. Although the replacement of MTBE by ethanol is expected to have some benefits in terms of water contamination due to ethanol's relative toxicity and rapid biodegradation, these could not be quantified. The overall estimated health risks for exposure scenarios with ethanol, MTBE, and nonoxygenated fuel did not differ between fuel formulations, but showed a significant improvement for 2003 in comparison to 1997.

The Northeast States for Coordinated Air Use Management (NESCAUM) also prepared an analysis in response to potential phase-out of MTBE in Northeast States due to adverse water quality impacts of MTBE and the need to protect environmental resources. The NESCAUM estimated that the use of ethanol instead of MTBE would increase emission rates of acetaldehyde, a combustion by-product, into the atmosphere in the Northeast by 50 to 70%. Acetaldehyde is also a probable human carcinogen. Ambient levels of acetaldehyde in the Northeast exceed the health-protective levels of acetaldehyde at majority of the air monitoring locations throughout the northeastern states. NESCAUM estimates that ethanol at 10% by volume in conventional gasoline would increase the VOC emissions significantly. Therefore, NESCAUM stressed the importance of studying health effects of ethanol alone, in combination with gasoline, but importantly, its lifecycle, its degradation products, and concurrent VOC emission increases. Schifter *et al*. also estimated an increase of about 80% in acetaldehyde emissions with 6% by volume ethanol as a substitution for 5% by volume MTBE accompanied by 16% decrease in CO and a 28% decrease in formaldehyde motor vehicle emissions in Mexico City. However, according to the CARB analysis on impact of use of ethanol-containing gasoline on air quality in Denver, Colorado, Albuquerque, New Mexico, Brazil, and other areas, acetaldehyde levels were substantial only in Brazil, where the fuels contained either pure ethanol or 22% ethanol, much greater than the maximum 10% ethanol allowed in CA gasolines.

There are no EPA toxicity values established for cancer (cancer slope factor) or noncancer (RfD or RfC) health endpoints for ethanol. There is also no safe drinking water standard for ethanol. The lack of toxicity values for ethanol for various exposure routes and lack of exposure information in various exposure media do not enable us to quantitatively estimate health risks associated with exposure to ethanol under different exposure conditions for public exposed to ethanol due to its circulation in the environment related to its widespread use in gasoline. Such analysis should not only include ethanol, but also all environmental degradation products of ethanol, and health risks and benefits analysis focusing on multimedia exposures is essential for not repeating the history of MTBE episode.

5. CONCLUSIONS AND RECOMMENDATIONS

There have been significant air quality benefits since the implementation of the oxygenated fuels program. Although some of the improvement in air quality indices may be simply due to fleet turnover, increase in fuel efficiency over time, and improvements in vehicle testing and maintenance programs, the oxygenates have also been credited for air quality

benefits. For example, 10 to 14% reduction in ambient CO concentrations has been attributed to use of oxygenated in winter months in CO nonattainment areas. Although the National Research Council concluded that little air quality impact in terms of ozone reduction can be gained from the use of oxygenates in RFG, the EPA's blue ribbon panel acknowledged major reductions in VOCs, especially, some of the carcinogenic air toxics (e.g., benzene, 1,3-butadiene) in many urban environments where RFG is used during the summer months. However, health complaints associated with exposure to MTBE containing gasoline and discovery of degradation in water quality in national scale with contamination of groundwater, surface water, and drinking water resources with MTBE overshadowed the potential air quality benefits and associated air-quality-related health benefits in the decade since the enactment of the oxygenated fuel programs by the Congress. The course of the national debate on oxygenates has shifted to potential alternatives to MTBE. We know more about the potential human health effects associated with exposure to MTBE alone or in combination with gasoline than about its potential substitutes. Quantitative assessment of the public health inhalation risks of MTBE-oxyfuel in comparison to conventional gasoline has been difficult, in part due to uncertainties and data deficiencies in toxicity combined with the paucity of the population-specific human exposure data. Although both short-term field epidemiological studies and a few controlled human laboratory studies of healthy adults provide a better understanding of human exposures to MTBE and MTBE-oxyfuel, the large differences in terms of confounding parameters, study design, exposure conditions, and exposure parameters greatly complicate the interpretation of the results of these few studies. Although more animal toxicity data for MTBE have become available and have increased our scientific understanding of the toxicity of MTBE (especially the pharmacokinetics of MTBE), there are insufficient data to answer questions about the relevance of these animal data to humans. Additionally, in the absence of oral toxicity studies, health effects of ingested MTBE cannot be determined. Although neither the epidemiological studies nor the controlled human exposure studies of normal and sensitive subjects have been, in majority, inconclusive, they cannot rule out the possibility that for a small percentage of the population, those claiming that they were affected by MTBE, there is a causal relationship between their exposure and their symptoms. In conclusion, the human health effects of MTBE are still unclear. For some other oxygenates, animal testing has not been thoroughly conducted and published yet. Any alternatives to MTBE, including TAME, ETBE, DIPE, and ethanol, should be studied with well-designed experiments to assess potential environmental and public health implications of using these compounds in large scale. This requires a substantial commitment and well-coordinated research effort by government and industry. The EPA scientists have evaluated retrospectively the governmental decision making with MTBE and concluded that

> *The most valuable lesson to be gained from the MTBE experience may be the importance of obtaining comprehensive understanding of the many and varied risks and benefits potentially offered by any given fuel option. This should entail a consideration of all facets of the life cycle of a fuel or additive, including the environmental fate of any emissions or releases, because the by-products of a chemical may be of greater toxicological or environmental significance than the parent compound itself.*

Given the fact that ethanol has increasingly gained oxygenate market share in the absence of such comprehensive evaluation should remind the EPA and the scientific community about the importance of critical findings gleaned through the MTBE experience.

SEE ALSO THE FOLLOWING ARTICLES

Air Pollution from Energy Production and Use • Air Pollution, Health Effects of • Clean Air Markets • Climate Change and Public Health: Emerging Infectious Diseases • Consumption, Energy, and the Environment • Ethanol Fuel • Hazardous Waste from Fossil Fuels • Motor Vehicle Use, Social Costs of • National Energy Policy: United States • Suburbanization and Energy

Further Reading

Adler, K., Schaffer, S., and Peeples, J. E. (1993). Oxygenated fuel season starts smoothly. *Fuel Reformulation* **62,** November/December.

Ahmed, F. E. (2001). Toxicology and human health effects following exposure to oxygenated or reformulated gasoline. *Toxicol. Lett.* **123,** 89–113.

Anderson, H. A., Hanrahan, L., Goldring, J., and Delaney, B. (1995). An investigation of health concerns attributed to reformulated gasoline use in Southeastern Wisconsin. Wisconsin Department of Health and Social Services, Division of Health, Bureau of Public Health, Section of Environmental Epidemiology and Prevention. Madison, WI.

Beller, M., Schloss, M., and Middaugh, J. (1992). "Potential illness due to exposure to oxygenated fuel in Fairbanks, AK." State of Alaska, Department of Health and Social Services, Anchorage, AK.

Belpoggi, F., Soffritti, M., and Maltoni, C. (1995). Methyl-Tertiary-Butyl Ether (MTBE)—A gasoline additive—causes testicular and lympho-hematopietic cancers in rats. *Toxicol. Industrial Health* **11**(2), 119–149.

Cain, W. S., Leaderer, B. P., Ginsberg, G. L., Andrews, L. S., Cometto-Muḫiz, J. E., *et al.* (1996). Acute exposure to low-level methyl tertiary-butyl ether (MTBE): Human reactions and pharmacokinetic responses. *Inhalation Toxicol.* **8,** 21–48.

Center for Disease Control and Prevention (CDC) (1993). An investigation of exposure to methyl tertiary butyl ether in oxygenated fuel in Fairbanks, Alaska. National Center for Environmental Health, Atlanta, GA.

Center for Disease Control (CDC) (1993). An investigation of exposure to methyl tertiary butyl ether among motorists and exposed workers in Stamford, Connecticut. U.S. Department of Health and Human Services, National Center for Environmental Health, Atlanta, GA.

Center for Disease Control (CDC) (1993). An investigation of exposure to MTBE and gasoline among motorists and exposed workers in Albany, New York. U.S. Department of Health and Human Services, National Center for Environmental Health, Atlanta, GA.

Chandler, B., and Middaugh, J. (1992). "Potential illness due to exposure to oxygenated fuel in Anchorage, Alaska." State of Alaska, Department of Health and Social Services, Bull. No. 1, Anchorage, AK.

Chun, J. S., Burleigh-Flayer, H. D., Kintigh, W. J. (1992). Methyl Tertiary Butyl Ether: Vapor inhalation oncogenicity study in Fischer 344 rats. BRRC Report 91N0013B. Union Carbide. Bushy Run Research Center. Export, PA.

Davis, J. M. (2002). Fuel oxygenates: How do their health effects compare? Presented at the ASTSWMO State Symposium on Fuel Oxygenates: Analysis, Assessment, and Remediation. October 21. Arlington, VA.

Davis, J. M., and Farland, W. H. (2001). The paradoxes of MTBE. *Toxicolog. Sci.* **61,** 211–217.

Environmental Protection Agency (EPA) (1992). Methyl-t-Butyl Ether. Drinking Water Health Advisory. Health and Ecological Criteria Division. Office of Science and Technology. Office of Water, Washington, DC.

Environmental Protection Agency (EPA) (1993). Assessment of potential health risks of gasoline oxygenated with methyl tertiary butyl ether EPA/600/R-93/206. Office of Research and Development, Washington, DC.

Environmental Protection Agency (EPA) 1994. Health risk perspectives on fuel oxygenates. Report No. EPA 600/R-94/217. Ann Arbor, Michigan, Office of Research and Development.

Environmental Protection Agency (EPA) 1994. National air quality and emissions trends report, 1993. EPA 545/R-94–026. Washington, DC.

Environmental Protection Agency (EPA) (1995). Proceedings of the Conference on MTBE and Other Oxygenates: A Research Update, July 26–28, 1993, Fairview Park Marriott, Falls Church, Virginia. EPA/600/r-95/134. U.S. Environmental Protection Agency. National Center for Environmental Assessment, Research Triangle Park, NC.

Environmental Protection Agency (EPA) (1996). Oxyfuel research needs. Report No. EPA 600/R-96/069. Office of Research and Development, Washington, DC.

Environmental Protection Agency (EPA) (1997). Drinking water advisory: consumer acceptability advice and health effects analysis on methyl tertiary butyl ether (MTBE) EPA-822-F-97–009. Health and Ecological Criteria Division, Office of Science and Technology, Office of Water, Washington, DC.

Environmental Protection Agency (EPA) 1999. Achieving Clean Air and Clean Water: The Report of the Blue Ribbon Panel on Oxygenates in Gasoline. EPA420-R-99–021. September 15.

Environmental Protection Agency (EPA) 1999. SB 521 Research Project Review. Executive Summary, Review of Volumes I–IV. "Health and Environmental Assessment of MTBE." Office of Mobile Sources, Office of Air and Radiation, Washington, DC.

Environmental Protection Agency (EPA) 2000. Control of MTBE in gasoline. Office of Transportation and Air Quality. March

Erdal, S., and Goldstein, B. D. (2000). Methyl tert-butyl ether as a gasoline oxygenate: Lessons for environmental public policy. *Annu. Rev. Energy Environ.* **25,** 765–802.

Federal Register (1998). National Toxicology Program; Call for Public Comments on 11 Agents, Substances, Mixtures and Exposure Circumstances Proposed for Listing in or Delisting from the Report on Carcinogens, 9th ed. **63**(239), 68783–68785.

Fielder, N., Kelly-McNeil, K., Mohr, S., Lehrer, P., Opiekun, R. E., *et al.* (2000). Controlled human exposure to methyl tertiary butyl ether in gasoline: Symptoms, psychophysiologic and neurobehavioral responses of self-reported sensitive populations. *Environ. Health Persp.* **108,** 753–763.

Fielder, N., Mohr, S. N., Kelly-McNeil, K., and Kipen, H. M. (1994). Response of sensitive groups to MTBE. *Inhalation Toxicol.* **6,** 539–552.

Hakkola, M., Honkasalo, M. L., and Pulkkinen, P. (1997). Changes in neuropsychological symptoms and moods among tanker drivers exposed to gasoline during a work week. *Occupational Med.* **47**(6), 344–348.

Irvine, L. F. H. (2003). Relevance of the developmental toxicity of ethanol in the occupational setting: A review. *J. Applied Toxicol.* **23,** 289–299.

Johanson, G., Hihlen, A., and Lof, A. (1995). Toxicokinetics and acute effects of MTBE and ETBE in male volunteers. *Toxicol. Lett.* **82/83,** 713–718.

Lidderdale, T. (1999). Oxygenate demand in reformulated and oxygenated gasoline control areas. Report. Department of Energy. Energy Information Administration. August 6.

Lioy, P. J., Weisel, C. P., Jo, W., Pellizzari, E., and Raymer, J. H. (1994). Microenvironmental and personal measurements of methyl-tertiary butyl ether (MTBE) associated with automobile use activities. *J. Exposure Anal. Environ. Epidemiol.* **4,** 427–441.

Mohr, S. N., Fiedler, N., Weisel, C., and Kelly-McNeil, K. (1994). Health effects of MTBE among new jersey garage workers. *Inhalation Toxicol.* **6,** 553–562.

Moolenaar, R. L., Hefflin, B. J., Ashley, D. L., Middaugh, J. P., and Etzel, R. A. (1994). Methyl tertiary butyl ether in human blood after exposure to oxygenated fuel in Fairbanks, Alaska. *Arch. Environmental Health* **49**(5), 402–409.

National Research Council (NRC) (1996). Toxicological and Performance Aspects of Oxygenated Motor Vehicle Fuels. National Academy Press, Washington, DC.

National Research Council (NRC) (1999). Ozone-Forming Potential of Reformulated Gasoline. Committee on Ozone-Forming Potential of Reformulated Gasoline. Commission on Geosciences, Environment, and Resources, National Academy Press, Washington, DC.

National Science and Technology Council (1997). Interagency Assessment of Oxygenated Fuels. National Science and

Technology Council, Committee on Environment and Natural Resources (CENR) and Interagency Oxygenated Fuels Assessment Steering Committee. White House Office of Science and Technology Policy (OSTP) through the CENR of the Executive Office of the President, Washington, DC.

Newman, A. (1995). MTBE detected in survey of urban groundwater. *Environ. Sci. Tech.* **29**(7), 305A.

Office of Science and Technology Policy (OSTP) (1996). Air Quality Benefits of the Winter Oxyfuel Program. Prepared for Interagency Oxygenated Fuel Assessment. Coordinated by Office of Science and Technology Policy, Executive Office of the President, Washington, DC.

Opiekun, R. E., Freeman, N. C. G., Kelly-McNeil, K., Fiedler, N. L., and Lioy, P. J. (2001). Effect of vehicle use and maintenance patterns of a self-described group of sensitive individuals and nonsensitive individuals to methyl tertiary-butyl ether. *J. Exposure Anal. Environ. Epidemiol.* **11,** 79–85.

Peeples, J. E. (1994). EPA finalizes reformulated gasoline regulations—Finally! *Fuel Reformulation,* March/April.

Prah, J. D., Goldstein, G. M., Devlin, R., Otto, D., Ashley, D., *et al.* (1994). Sensory, symptomatic, inflammatory and ocular responses to and the metabolism of methyl tertiary butyl ether in a controlled human exposure experiment. *Inhalation Toxicol.* **6,** 521–538.

Rihimaki, V., Matikainen, E., Akila, R., Teravainen, E., Mutanen, P., *et al.* (1998). "Central Nervous System Effects of the Gasoline Additive Methyl-Tert-Butyl Ether (MTBE)." Finnish Institute of Occupational Health, Finland.

Schifter, I., Vera, M., Diaz, L., Guzman, E., Ramos, F., and Lopez-Salinas, E. (2001). Environmental implications on the oxygenation of gasoline with ethanol in the metropolitan area of mexico city. *Environ. Sci. Technol.* **35,** 1893–1901.

Squillance, P. J., Pope, D. A., and Price, C. V. (1995). "Occurrence of the Gasoline Additive MTBE in Shallow Ground Water in Urban and Agricultural Areas. U.S. Geological Survey. Fact Sheet FS-114–95." U.S. Geol. Surv, Rapid City, SD.

UCRL (1999). Health and Environmental Assessment of the Use of Ethanol as a Fuel Oxygenate. Prepared by the California Air Resources Board, the State Water Resources Control Board, and the California Environmental Protection Agency's Office of Environmental Health Hazard Assessment December 30.

White, M. C., Johnson, C. A., Ashley, D. L., Buchta, T. M., and Pelletier, D. J. (1995). Exposure to methyl tertiary butyl ether from oxygenated gasoline in Stamford, Connecticut. *Arch. Environ. Health.* **50,** 183–189.

Zogorski, J. S., Delzer, G. C., Bender, D. A., Baehr, A. L., Stackelberg, P. E., *et al.* 1998. MTBE: Summary of Findings and Research by the U.S. Geological Survey. Proceedings of the 1998 Annual Conference of the American Water Works Association.

Geographic Thought, History of Energy in

BARRY D. SOLOMON
Michigan Technological University
Houghton, Michigan, United States
MARTIN J. PASQUALETTI
Arizona State University
Tempe, Arizona, United States

1. **Early Research on the Coal Industry, 1934–1955**
2. **Fossil Fuels and Electric Power Generation, 1955–1971**
3. **The Birth of Commercial Nuclear Energy, 1951–1973**
4. **Modeling the Spatial Development and Transportation of Fossil Fuels and Power Plant Siting**
5. **Hazards, Risks, and Technological Diffusion: Behavioral Research by Geographers Since the TMI Accident**
6. **Development of Sustainable Energy Resources**

Glossary

econometric analysis A test of economic theory with linear or nonlinear regression and other multivariate statistical methods using socioeconomic and technical data.

geography of energy The study of energy development, transportation, markets, or use patterns and their determinants from a spatial, regional, or resource management perspective.

location theory Spatial theory developed to solve optimization problems for selecting a preferred zone or location for a facility, which involves trading off transportation costs with production and processing costs, production balances, and market-delivery levels.

mathematical programming model An operations research technique designed to determine an optimal solution to a well-defined problem by maximizing or minimizing an objective function subject to technical, economic, and often environmental constraints.

regional input–output analysis An analysis of the economy at the urban, state (provincial), or multistate (provincial) levels based on a fundamental identity that equates supply and demand using fixed proportions of intermediate inputs.

Three Mile Island (TMI) The name of the nuclear power plant near Harrisburg, Pennsylvania, that was the site of the most serious nuclear accident in the United States on March 28, 1979.

Since the field of geography encompasses broad study of interactions between humans and the environment, the concept of energy has been operationalized and studied in several ways. These approaches have included energy balance studies in geographical climatology, energy resource availability and conservation in applied climatic research, cultural geography, and economic geography of energy resources. In the past few decades, new geographic perspectives on energy have been introduced from spatial analysis and modeling to applied, political, transportation, hazards, geographic information systems, global environmental change, and ecological economics. Some of these areas are even considered separate fields.

Understanding and modeling solar radiation and the energy budget of the climate at a variety of spatial scales have long been central concerns of several types of climatological studies: modeling, synoptic and dynamic climatology, climate change, and applied climatology. Analytical approaches have become increasingly mathematical over time, and this work has been somewhat separate from other branches of geography. It is consequently more convenient to trace the evolution of the energy concept in geographic thought by focusing on the contributions of human geography, which initially considered energy from economic and cultural perspectives. The article begins with a discussion of pre-World War II studies of the coal industry.

 © 2004 Elsevier Inc. All rights reserved.

Subsequent sections consider the rise of petroleum as a global political force, commercial nuclear power, and additional work on the location and spatial distribution of fossil fuels. A new period in geographic thought followed the Three Mile Island nuclear power plant accident in 1979, when geographers gave attention to technological hazards and risks of energy systems, the back end of the nuclear fuel cycle, the diffusion of energy conservation technologies, and the increasing scarcity of fossil fuels. The article concludes with a discussion of the growing need to develop sustainable energy resources in the context of concerns about global climatic change and rising energy demands in developing nations.

1. EARLY RESEARCH ON THE COAL INDUSTRY, 1934–1955

Coal, as the fossil fuel of greatest abundance and widest distribution, has long attracted attention from geographers. In the early years, this attention focused on the location of deposits and their transport and much of it mirrored the major coal fields of Britain, Germany, Poland, and the United States. The accounts in geography journals were largely descriptive, economic, and cartographic. Little or no interest was evident regarding the environmental costs. Later, when the public mood shifted and impacts were widespread, more attention was paid to environmental impacts, especially air pollution and its dispersal, impacts of acid rain, landscapes altered by strip mining and waste heaps, and the forced relocation of residents. This emphasis came with the increasing knowledge not only of the damage that was created by our demand and use of coal but also that we should and could do something to reduce or reverse the impact it was having on air, land, and water. From this point forward, research articles were much more widely dispersed, within a range of environmental journals and books, and the type of writing was more analytical.

2. FOSSIL FUELS AND ELECTRIC POWER GENERATION, 1955–1971

Parallel to the growth and development of the oil and natural gas industries in the United States and other advanced economies, geographers began to give more attention to the other fossil fuels, their development, and use patterns in the mid-1950s. Another exciting new energy source gained attention with the commercialization of atomic power plants in Britain, Russia, and the United States during this period. Nuclear power development attracted considerable attention among geographers in North America and Europe. This topic is considered separately since the economics, spatial diffusion, risks, social acceptance, waste problems, and other dimensions of nuclear power were highly uncertain in the 1950s and thus it was treated differently.

Although a quantitative revolution in geography first appeared in the mid-1950s there was little evidence of a change in methods in energy research until the early 1960s. Indeed, most geographical publications on energy issues in the 1950s were studies on the regional geography of energy (i.e., "areal differentiation"). Although important, this work was overly descriptive and often handicapped by scant understanding of the complex interplay between resources and technical, economic, and social factors in energy markets. Several detailed analyses were completed on the oil, gas, and electric power industries of the United States, Britain, Europe, and the former Soviet Union during this time. A few researchers went beyond the single-region scale and assessed the development of power interconnections in Europe, North America, and the Soviet Union. A 1960 paper addressed the conflict between hydroelectric power generation and salmon fisheries in the Pacific Northwest, which is still relevant today. This study suggested the use of (but did not apply) optimization criteria to help resolve the conflict. The same year marked the inaugural publication of *Soviet Geography: Review and Translation*, which was to become a major scholarly outlet for energy studies of the former Soviet Union and its various regions.

Neoclassical studies of the regional geography of energy continued throughout the 1960s at a variety of spatial scales. Several books and articles on the economic geography of the oil and gas industries analyzed national and international patterns of location, economic growth, development, and trade but did not suggest general market patterns. The first truly international energy study, a book on the economic geography of the world oil industry by the British geographer Peter Odell, was published in 1963, just a few years after the formation of the Organization of Petroleum Exporting Countries (OPEC). Articles on coal geography also assessed economic growth and development but considered

declining or less prosperous coal fields and their socioeconomic effects as well. Political factors (i.e., national policies, government regulation, political stability, etc.) began to be mentioned more frequently in this literature but were still secondary to economic considerations.

Of special significance was the work by Gerald Manners in the 1960s to incorporate location theory into energy geography. General locational principles were identified and applied to the oil, gas, and electric power sectors. Transportation costs were shown to be of fundamental importance because they determine the energy facility location between a resource market(s) and final demand center(s). Also important are factor costs of production, resource availability, weight and quality, scale economies, technical efficiency, modal and interfuel competition, and the load factor(s). Similar principles were used to explain the changing pattern of pit location within a coal field. The first such analytical, general approach for studying the industrial geography of energy was published in book form by Manners in 1964. (A 1950 book on the geography of energy by Pierre George received little notice outside of France.) Owing to data availability and the scale of the economies, his research focused on Western Europe and North America. National studies and atlases for the United States and United Kingdom were to come later.

As the 1960s closed, a few prominent authors completed seminal works on the international patterns of energy geography. A landmark book by Odell addressed oil and world power; this book was first published in 1970 and updated seven times. The monograph considered this leading industry's production, economic growth, and trade in the context of its prominent role in national and international political affairs. By analyzing the global structure of the oil industry along with regional perspectives and company–nation state relationships, Odell provided the most complete story of this vital resource. Three more general energy papers by geographers in the September 1971 issue of *Scientific American* showed the close connection between energy resource type and technology at the various life-cycle stages and the pattern of energy flows in primitive and advanced industrial economies.

Two final (although brief) energy books from the perspective of world economic geography closed out this period, both of which were written by Nathaniel Guyol. One of these monographs provided, for the first time, a system of energy accounting that made it possible to compare energy statistics in multiple countries on a consistent basis, a subject he had worked on since the late 1940s. An accounting worksheet, suggestive of a first cut of a simple input–output table, was developed for The Netherlands as an example. The other volume was narrower but more detailed: a systematic survey of the electric power industry in 162 nations. Important observations were made by Guyol on the regional variations in electricity supply and use.

3. THE BIRTH OF COMMERCIAL NUCLEAR ENERGY, 1951–1973

The use of commercial nuclear power grew up as the stepchild of the military program. After spending an enormous amount of money to develop and construct the three bombs that were detonated in 1945, there was a concerted effort to turn this research to peaceful use. There followed almost a decade of research and development that led to the construction of electrical generating stations at Calder Hall, England, Shippingport, Pennsylvania, Eureka, California, and eventually more than 400 other commercial units.

Geographers showed little interest in the dispersed siting of the nuclear research program of the war years, but afterwards they addressed the possibility of using nuclear bombs to reshape the earth's surface, the location of nuclear fuels, and the siting of nuclear power plants. It was electricity generation that was to achieve the greatest momentum as commercial enterprises began appearing in several countries: Britain, the United States, the Soviet Union, and later France, Germany, and Japan. Researchers during this period published a steady, if thin, series of geographical studies, only rarely turning their attention to the spatial pattern of the fuel cycle or dispersal of the technology.

For more than 15 years, there seemed little uncertainty or public misgivings about the wisdom of a nuclear power program, and the geographic literature did not contradict this trend. Industry, government, and the public seemed to all be moving in the same direction, one that looked to nuclear power to generate a growing, perhaps dominant, proportion of the great amounts of new generating capacity that many countries were demanding. With U.S. growth rates for electricity exceeding 7% per year in the 1960s, nuclear power not only seemed like the most likely supplement to coal-fired power plants but also seemed to be the eventual replacement

technology as concerns grew over the pollution that such fossil power plants produced. Moreover, it was recognized that by themselves coal stations would not be able to keep ahead of the tidal wave of demand for electricity. The public, bolstered by the reassurances of government and contractors, regarded the use of nuclear power as the only way to meet the steep rise in need.

During the early civilian period, power plant siting was influenced mostly by seismic stability and the availability of cooling water. Little consideration was given to isolating power plants from urban areas; indeed, many were located close to load centers, the better and more conveniently to service their needs. Thus, the pattern that was emerging favored the densely settled northeastern United States and Chicago. Gradually, additional power plants appeared in the southeast. In Britain, all but one (in Wales) were sited along coasts. France used both coastal and riverine sites, starting mostly with the latter locations, and in contrast to the United States and the United Kingdom, the distribution of nuclear power plants had no obvious locational concentration. However, the benign public reaction to nuclear power was to be short-lived.

4. MODELING THE SPATIAL DEVELOPMENT AND TRANSPORTATION OF FOSSIL FUELS AND POWER PLANT SITING

By the early 1970s, the quantitative revolution was firmly entrenched in energy geography. Researchers began to apply several mathematical and statistical techniques to problems of spatial energy resource development and transportation, power plant siting, and socioeconomic and environmental impact analysis. These important research developments, spanning the 1970s through the mid-1990s, are discussed in turn.

4.1 Resource Development and Transportation

Beginning in the 1970s, geographers developed and applied a variety of simulation, mathematical programming, and econometric models to oil and gas issues. Among the early work was a rather extensive analysis of North Sea oil development through 2080. A less ambitious study assessed the optimal oil pipeline network in the British sector of the North Sea. North American oil industry studies also used mathematical programming models, with the site selection and storage levels at the Strategic Petroleum Reserve in Louisiana as an example. Since the mid-1980s, a series of related papers have used econometric methods to update the oil industry models of geologist M. King Hubbert. This work often differentiated resource quality and used more than 50 years of historical data to assess the bleak future prospects of the U.S. oil industry. Somewhat less attention has been given to oil development in other regions. A few econometric studies have analyzed foreign investment, locational analysis, and resource-based industrialization in the Middle East, other developing countries, and the former Soviet Union. Additional studies have highlighted the main result of these developments for the United States: growing dependence on foreign sources of oil.

Mathematical programming methods also began to dominate research on the development and marketing of natural gas in North America and the Soviet Union. Since the production of this fossil fuel generally lagged that of petroleum in the United States and was decontrolled starting in the late 1970s, new research opportunities were pursued for modeling greater development, especially in Alaska, and trade with Canada and Mexico. Similarly, natural gas emerged as the fuel of choice in the Soviet Union in the 1980s. In order to advance this research area, it was necessary to improve upon previous models of the pipeline network and capacities. Following the decline in oil and gas prices and plentiful gas supplies in the United States since 1986 this research area has been largely abandoned. Only one major analysis of U.S. natural gas development has been completed since then, an econometric analysis of the declining yield per effort and resource scarcity.

Not surprisingly, the quantitative research methods that were first used in the 1970s for oil and gas studies also were applied to coal problems, along with the use of cluster analysis and spatial interaction models. Given the wide distribution of coal reserves in the United States and their often long distances from markets, transportation bottlenecks have occasionally arisen. Such problems are ideal cases for the use of operations research methods, such as linear, nonlinear, mixed integer, and multiobjective programming. An early 1970s study of the Great Lakes region was a landmark in the modeling of the whole coal resource delivery system. This was followed by a series of coal studies: regional industry analyses, such

as for Britain, western U.S. states, and Pennsylvania; transshipment problems involving railroads and ports; coal exports; competition for water resources; other environmental constraints; and structural and technological change.

Geographers, along with scholars in other fields, have also studied the development potential of synthetic fuels from fossil fuels. This work has analyzed the spatial distribution patterns and development potential of coal gasification and liquefaction plants, oil shale, tar sands, and alcohol fuels. The huge water and other resource requirements and environmental effects were given prominent attention. These studies tracked the government policy interest and short-lived massive subsidies to potential synthetic fuels industries in North America in the late 1970s and early 1980s, which were briefly considered a viable alternative to the growing dependence on foreign oil supplies. Among these resources, only alcohol fuels (from ethanol and methanol) and the Canadian tar sands industry have established commercial viability.

4.2 Power Plant Siting

The siting patterns of electric power generating stations have been a major research area among energy geographers. Although early papers on this subject were published in 1960, the vast majority of these studies were done in the mid-1970s through the mid-1980s. Perhaps surprisingly, a retrospective discriminant analysis of U.S. power plant siting patterns from 1912 to 1978 showed that fuel choice was more important than regional differences in the siting process. Development of site suitability models for fossil-fueled and nuclear power plant location analysis was pioneered at the Oak Ridge National Laboratory in the late 1970s. These and later siting models were applicable to efforts to avoid or mitigate inevitable locational conflicts surrounding proposals for many new energy facilities. Geographers developed land use screening procedures to generate local site suitability scores by estimating a land use compatibility index and using importance weights offered by decision makers. A large range of engineering, ecological, and socioeconomic criteria were analyzed. These models were applied to the Mid-Atlantic states. Despite the value of these methods for initial site screening, it has been noted that the approach largely failed to be implemented because of a lack of need for new power plant sites at that time. On the other hand, the screening models amount to a stepwise heuristic and can select inferior sites. Improvements to the method were made through the development and application of optimal transshipment and multiobjective programming models, which were also applied to the Mid-Atlantic region. Another study considered the challenges in siting coal-fired power plants in western states, where there are stricter air quality standards.

The siting and safety of nuclear power plants received the most attention by geographers of energy during the early to mid-1980s. Ironically, following the Three Mile Island (TMI) accident in 1979, North American energy geographers shifted their research from power plant siting to reactor decommissioning and waste disposal, whereas British colleagues focused on nuclear plant siting. Computer programs were written in England to analyze alternative power plant siting criteria while emphasizing remote areas and the potential human dose from radioactive releases during accidents. Government policy was shown to downplay the siting criterion of proximity to human population centers, and thus arguably safety, despite the existence of ample remote sites in Britain. Cluster analysis was used to show that local population density received greater attention with regard to siting in the United States than in Britain. Broadening this discussion from a Canadian perspective, it was shown that existing nuclear power expansion plans in the mid-1980s were not as robust as construction of small-scale hydroelectric stations. Expansion of cheap hydroelectric power in eastern Canada, whether small or large, fits in well with plans for increased electricity trading with the United States. Indeed, further study has highlighted the parallel socioeconomic and environmental challenges accompanying the recent "outsiting" of nonnuclear power plants in Mexico and Canada that have been dedicated to U.S. markets.

Since U.S. electric utilities had a greatly diminished need for power plant siting in the late 1980s as independent power producers and demand-side management picked up the slack, energy geographers abandoned regional siting analysis and developed broader, more detailed, and sophisticated industry resource planning models. A few major, large linear programming (LP) models were built with a full range of technical options for meeting future electricity demand under environmental constraints. One of these used multicriteria decision making and applied it to demand-side management planning in British Columbia, Canada. Most of the LP models

were applied to sulfur dioxide (SO_2) emissions control. There was some limited regional detail to them, which overlapped with the work of regional scientists and regional economists.

4.3 Socioeconomic and Environmental Effects

Research by geographers on the mixed effects of coal mining dates back to at least the early 1950s. Initially, this work focused on the eastern United States, but by the late 1970s and 1980s it had shifted to the arid West (along with an increasing share of the coal production). This research analyzed the potential socioeconomic impacts of "boomtown" development of coal or synthetic fuels. The declining economic fortunes of some of the eastern coal mining areas such as anthracite towns of Pennsylvania continued to receive attention in the 1980s. Regional input–output (IO) analysis was demonstrated to be an especially useful tool for analyzing economic effects of spatial shifts in the coal industry. Similarly, most coal research in Britain during this period focused on the uneven economic development or shift between declining and developing regions and a short-lived effort to revive the industry. When U.S. western energy resource development plans were scaled back, research emphasis shifted to the economic and ecological effects and control of acid rain emissions of SO_2 and NO_x. This subject continued to receive the attention of a few geographers through the 1990s as the Acid Rain Program of the Clean Air Act Amendments was implemented, including (interstate) SO_2 allowance trading.

Many energy geographers have examined the socioeconomic and environmental effects and planning issues surrounding petroleum development in disparate regions such as the Middle East, Scotland, Norway, and the U.S. Gulf Coast. The U.S. studies have also considered impact mitigation in the context of the National Environmental Policy Act and the macroeconomic effects of a future large supply disruption by OPEC. Econometric analyses by geographers for the United States have shown since 1989 that policies to expand domestic oil production would have adverse economic and ecological effects. Although less sweeping in its findings, a regional economic analysis in Canada of potential oil resource development in the nearby Beaufort Sea found that little or none of the net economic benefits would accrue to Canada's Northwest Territories. This research used a multiregional IO model.

5. HAZARDS, RISKS, AND TECHNOLOGICAL DIFFUSION: BEHAVIORAL RESEARCH BY GEOGRAPHERS SINCE THE TMI ACCIDENT

Public disinterest in nuclear power was shaken by the successful campaign to raise awareness about its purported safety risks. It was a movement that slowly gained momentum, but by 1976 an antinuclear referendum in California received ample public notice and approximately 40% of the votes. Although the referendum failed, it was a watershed for the nuclear industry because it raised many more questions than the industry could successfully answer about the safety and the financial feasibility of the technology. In addition, it prompted more restrictive and demanding siting criteria, and more than anything else it put nuclear power on the map of regular public debate. Like colleagues from other disciplines, geographers became more involved, and it was a natural topic of interest because most of the issues were spatial in nature.

The increased attention by geographers first manifested itself in questions about the adequacy of the siting criteria that had been unquestioned during the previous decade. Since much of the impetus for the nuclear debate had come from California, it was not surprising that the strongest opposition to power plant siting appeared there as well. Geographers played a quiet role at first, mostly related to siting, but their participation grew as the public debate expanded to other spatial issues, such as radioactive waste transport and disposal, dispersal of radioactivity by air currents, and identification of evacuation routes that might be needed in the event of an accident. These developments were all part of a simmering debate about nuclear power in the United States and elsewhere, but they would come to a boil with the TMI Unit 2 accident near Harrisburg, Pennsylvania, on March 28, 1979. Ironically, Pennsylvania had enthusiastically commissioned the earliest U.S. nuclear power plant 22 years earlier.

It was the TMI accident that firmly placed geographers within the nuclear power debate. For several years, many geographers turned their attention to the myriad questions that the accident raised, including siting criteria, the dispersal pattern of possible contamination, evacuation behavior, and the adequacy of emergency planning and preparedness. The effects that the accident had on nuclear power were fundamentally geographical. Most

noticeably, no new operating licenses were granted afterward, largely because of the elevated public concerns about the appropriateness of virtually every suggested site. Emergency requirements that were promulgated became so demanding that power plants already completed (e.g., the Shoreham plant on Long Island) were unable to develop an approvable evacuation plan.

Geographers participated in these studies and were most heavily involved in evaluating and measuring public risk perceptions. They asked questions about relative risk, the level of risk required to trigger evacuation, and the use of distance as a buffer to risk. These issues came into play because of the obvious inadequacy of emergency preparations at TMI, but the interests of geographers were not limited to them. Issues that had long been beneath the surface came into full view, and they included all forms of "nuclear geography," including a striking increase in the attention paid to radioactive waste transport and disposal.

Although attention to these matters was initially concentrated in the United States, it did not remain within its borders. Geographers in other countries, particularly Britain, heard the call as well, and geographers there had been making similar intellectual preparation to participate. There was a different attitude in the United Kingdom, however, in large part because their system of government had not been as receptive to an open debate on nuclear power as in the United States. As it soon became clear, its citizens had more practical reason for concern. Great Britain is a country smaller than Arizona, without the vast distances and isolation possible in the United States. Moreover, when the inquiry into the 1957 Windscale accident in northwest England confirmed public suspicions about nuclear hazards and ill-considered official secrecy, a phalanx of geographers was to lead the charge for a much wider nuclear debate.

All the nuclear plants that currently exist in the United States had already been sited by 1979, but this was not the case everywhere. In the United Kingdom, for instance, great attention was paid to siting and safety issues of the planned addition to the Sizewell nuclear station on the Suffolk coast. There was already a Sizewell A, but Sizewell B was to be a pressurized-water reaction of American design, ostensibly safer than the Magnox and Advanced Gas-Cooled reactors that comprised all the early British reactors. In response to the work of many groups, and especially geographers in the United States and Britain, the government conducted the longest and most detailed public inquiry in the history of the country.

Like the TMI accident, the Sizewell inquiry galvanized geographical attention. Soon, geographers in England were writing about waste transport through London, waste disposal under the Irish Sea, and elevated leukemia frequencies in Cumbria. Perhaps due to the smaller size of the country, much of the work of geographers in Britain attracted more attention than similar work did in the United States. Geographers regularly appeared on TV and in various newspapers and magazines, which gave their comments more visibility. Eventually, geographers from Britain and the United States began collaborating on articles and books. Even France, which operates nuclear power plants that are closer to London than those in Wales, was not immune from criticism, and although the debate was more subdued there, geographers tracked and reported on various acts of public opposition.

Attention of geographers to nuclear issues peaked in the early 1980s and then began to subside, only to be shaken into a higher degree of agitation 7 years later by the accident at the Chernobyl nuclear power plant in Ukraine. This was a far more serious accident, and it had global implications. Climatologists tracked the movement of radioactive materials, but access to data was difficult owing to long-standing policies of the Soviet Union. Nevertheless, several geographers focused their attention on the disaster, with some examining issues of the power plant's location adjacent to the vulnerable Pripyat Marshes and 90 miles upwind from the heavy concentration of humans around Kiev, others observing how land had to be reshaped and forests leveled for safety reasons, and still others focused on the forced evacuation of more than 135,000 people who lived near the plant. Others were intent on observing the various political ramifications, including how the accident helped hasten the collapse of communism.

The Chernobyl accident also pounded several more nails in the nuclear coffin. In 1988, Swedish voters reaffirmed that nation's decision to phase out nuclear power, and Germans voted to take much the same direction. The RBMK design used at Chernobyl was largely abandoned, and many of those that had been built in the former Soviet Union and satellite countries have been closed. Moreover, attention to the various elements of nuclear power plant decommissioning gained increasing momentum.

Although it was obvious that nuclear plants would eventually be "turned off," the questions were when, by what process, and to what effect. There had been

little public attention to these questions prior to 1985, but geographers helped bring attention to them during the next decade, emphasizing postoperational land use, transport and storage of the resulting low-level radioactive waste, the spatial and temporal sequence of targeted plants, and the socioeconomic consequences of the decommissioning process.

At approximately the time that interest in nuclear power was on the decline among geographers, attention was increasing in the area of energy conservation and efficiency. It was not causative, but when problems began appearing and proliferating with nuclear power, several electric utility demand-side management strategies were waiting in the wings. The logic was that if the electricity that nuclear power had been expected to supply was not actually going to be available, shortfalls would occur unless demand could be cut. It was cut, across the board and across the country, in every end-use sector, with the most noticeable changes occurring in the commercial and industrial parts of the economy and gradual progress in the residential sectors as well. Geographers worked with other professions to identify where these changes could be made and evaluated program effectiveness. There were several obvious uses to emphasize, including increased efficiency in lighting, insulation, appliances, and transportation.

6. DEVELOPMENT OF SUSTAINABLE ENERGY RESOURCES

Although physical geographers have long included solar energy in its various manifestations in their work, only since the late 1970s have geographers actively researched renewable forms of energy. Even so, this research interest tapered off in the late 1980s, seemingly coincident with the crash in world oil prices in 1986. This hiatus was short-lived because by the early 1990s, there was renewed interest in fossil fuel scarcity, namely that of oil and gas. Geographers also began to explore the energy policy implications of the growing concern regarding global climate change, with an emphasis on energy efficiency and renewable energy development.

6.1 Oil and Gas Scarcity

Although acknowledgment of the ultimate scarcity of petroleum resources dates back to the 1956 work of M. King Hubbert, who predicted with startling accuracy the ultimate size and production peak of oil resources for the United States and the world, there is much greater concern with its "net energy" from production rather than ultimate depletion. Cutler Cleveland and Robert Kaufmann, among others, in a series of papers since 1991 have convincingly forecast using econometric analyses that the amount of oil added to proven reserves per effort from additional well drilling (yield) in the United States will cease to be a net energy source early in this century. This line of research builds on the work of Hubbert, who was the first to recognize that the historical production pattern would follow a classic bell-shaped curve, reflecting the interplay of resource depletion, real oil prices, technical innovation, and political decisions on long-term production costs. Consequently, domestic oil discovery, production, and proven oil reserves are on an inexorable downward path. These findings call into question energy policy decisions based on assumptions that these geophysical realities of the U.S. resource base can be altered.

With the conventional oil resource base increasingly scarce and dependence on it unsustainable, the cleaner option of natural gas has received more attention, especially since estimates of its proved reserves were greatly increased in the early 1990s. Here too, analysis by Cleveland and Kaufmann is instructive. Using a double-log econometric model of yield per effort of gas drilling in the United States, they found an exponential decline for gas discoveries not associated with oil fields in the lower 48 states from 1943 to 1991. Their model accounts for cumulative drilling, real gas price, changes in drilling effort, and shifts between onshore and offshore gas. Thus, natural gas development lags oil but is subject to similar factors controlling its scarcity.

6.2 Greenhouse Gas Reduction and Renewable Energy Development and Use

Coal, the third fossil fuel, is well-known to have the highest emission rate of carbon dioxide (CO_2), the principal greenhouse gas. By the early 1990s, a growing cadre of geographers began to actively work on energy policy responses to the threat of planetary warming and climate change. Indeed, they recognized that many preferred policy options also reduce conventional air pollutants and have a variety of other ancillary benefits. These energy options include switching to gas, energy efficiency and energy conservation, and rapid development of renewable and possible nuclear energy sources (although no

major energy geographers have promoted the latter option through their research).

Three major types of greenhouse-related energy studies have dominated the work of energy geographers in the past few years: detailed technology assessments at the U.S. national level, sectoral-specific studies, and foreign case studies. One leading study provided an integrated assessment of the potential benefits of advanced energy technologies for greenhouse gas reductions in the United States. The authors found that large CO_2 reductions are possible at incremental costs below the value of energy saved, especially efficiency measures. Many sectoral-specific studies by geographers have highlighted the technological improvements and potential roles of energy efficiency, wind, solar, and solar–hydrogen technologies in buildings, transportation, manufacturing, and electric power. Much of this research has been conducted at the Oak Ridge National Laboratory.

Since global warming and climatic changes do not respect political boundaries, case studies of other nations can be valuable. For instance, several geographers have studied the broad range of policies that would be required to significantly reduce CO_2 emissions in Germany, Canada, and China. These studies not only highlight the great difficulty of significant emission reductions but also emphasize the necessity to implement policy reforms outside of the energy sectors (e.g., addressing urban design and transportation policies). To accelerate the process of emissions reduction in developing countries and economies in transition, the 1992 Framework Convention on Climate Change approved a pilot program in joint implementation of the treaty, whereby two or more nations cooperatively reduce or sequester emissions in their countries. Implementation of this pilot program was slowed, however, by concerns among developing countries that national sovereignty would be violated and that Western nations or Japan would gain a competitive economic advantage. Geographers have generally been sympathetic to these concerns, questioning the policy's fairness and cost-effectiveness, while identifying constructive solutions that involve compromises by the applicable countries.

Renewable energy technologies have different spatial characteristics than those of fossil or nuclear fuels (e.g., with respect to availability, transportability, reliability, and storage), which have yet to be fully explored. Among conventional energy resources, the two renewable sources that are well established are hydroelectricity and biomass. Hydroelectric installations, however, are not always sustainable because gradual siltation of their reservoirs may limit facility lifetime to 50–200 years. Since the 1990s, hydroelectric power has attracted increasingly negative attention in the United States and elsewhere, where the large part of its potential has already been tapped. Identified shortcomings include forced migration of people, displacement and destruction of salmon stocks, and downstream ecological changes.

Geographers have written a modicum of papers and monographs on hydropower facility construction in the past decade, which have focused on trends in North America and a few mammoth dam projects in developing countries such as Brazil, China, and India. Research emphasis in the United States has shifted from dam construction to dam removal and downstream restoration. Several dams have been suggested for demolition, especially smaller structures. Of the large dams, the most attention has been given to the proposed removal or "neutralization" of Glen Canyon Dam, located upstream from the Grand Canyon.

Biomass has long been the major (potentially) renewable energy source in developing nations, and it has received the most attention among the renewables from geographers. It has also received research attention in Scandinavia, where it is commonly used but somewhat controversial in Sweden. Along with its relative abundance, biomass also has challenging limitations owing to its wide variety of forms, ranging from wood to crops and garbage.

Wind energy has been the fastest growing energy source in the world in the past decade, and in the late 1990s Germany overtook the United States as the global leader. An important book printed in 2002 includes 10 chapters by geographers that compare and critique the varying prospects and policy approaches to renewable energy development, especially wind, biomass, and solar energy, in the United States and northern Europe. Wind energy use and public perception in California, the location of its most extensive U.S. development, have been examined by geographers in articles and a 2002 book that examines the issue of aesthetic intrusion on the landscape. Extensive study of wind energy has also been conducted by climatologists. For example, a 1997 paper analyzed the influence of climatic variability on wind resources in the U.S. Great Plains, the area of greatest wind power potential.

Like wind and hydropower, immobility presents geographic challenges for the development of geothermal energy, particularly the pressure it puts

on natural resources. Geographers have given the least research attention among the renewable energy sources to geothermal (because this research area is dominated by geologists and engineers), after significant contributions were made by Pasqualetti in the late 1970s and 1980s.

6.3 Energy in the Developing World

Faster than other professions, geography is predisposed to grasp differences between the developed and the developing world—differences that are particularly apparent in considerations of energy. The developed world relies on fossil fuels, whereas the developing world relies on renewables. The developed world searches for energy to maintain a high standard of living, whereas the developing world must sell its resources to stave off poverty. The developed world worries about the environmental impacts of energy supply, whereas the developing world worries about having any energy supply at all. The developed world has achieved a high standard of living by consuming prodigious amounts of energy, whereas the developing world resents any attempts to curtail their upward mobility by doing the same, even though the environmental impacts are now so much more obvious, troubling, and dangerous. The developed world is in the minority, whereas the developing world is in the majority. Such differences form the basis for much of the work geographers have pursued in developing countries.

Although many developing countries share a need to trade environmental quality for energy supply, they are in other ways profoundly different. For example, China and Mexico are industrial countries, whereas Haiti and Ghana have primarily agrarian economies. In the former, government concentrates on how to increase gross domestic product; in the latter, government tends to worry more about matters of day-to-day survival. The work of geographers has focused on issues of environmental quality, economic growth, equity, and sustainable development and in a more speculative mode the consequences of bringing per capita energy consumption in these countries closer to that of the United States or the European Union.

The work of energy geographers in developing countries is often closely tied to environmental considerations. China, for example, is intent on using its great reserves of coal to help fuel its climb toward greater economic prosperity. As a consequence of pursuing this goal, air pollution has increased in many locations within the country, as has concern about the possible impact of such additional coal burning on global warming. Geographers, among many scientists, have monitored such trends and have expressed opinions ranging from how one might minimize such impacts to suspecting that the impacts on climate change are insignificant.

Other topics of interest to geographers include energy trade and transportation; the use and scarcity of biomass; hydropower development; and the development of geothermal, wind, and solar power. Of these topics, energy transport and trade have had perhaps the most natural and longest interest. The focus of such attention by geographers evolved from matters of shipping routes and port facilities to the location of lines for the movement of energy by rail, pipe, and wire. Shipping routes tend to change infrequently, and then usually as the result of military conflicts and new discoveries. Modern examples include the closing of the Suez Canal and the development of oil fields in the Gulf of Campeche, both in the 1970s.

Geographers continue to track the discovery and development of new sources of energy, although such activities trend toward the expansion of existing areas of extraction rather than new developments. Exceptions to this rule include oil developments in African countries such as Chad and South American countries such as Ecuador, Peru, and Bolivia. In the latter country, interest has been drawn especially to natural gas fields on the eastern side of the Andes, with the prospect of piping through Chile to the Pacific Coast for liquefaction and transport to markets. One market is the United States, and Bolivia is considering shipping to a planned regasification facility in Baja, California, and then piping the gas to California. In so doing, they will be avoiding the environmental encumbrances of trying to locate regasification plants within the United States while still having a U.S. market. In all cases, these and similar projects hold the hope for improving the economic prosperity of citizens of the producing countries, although the usual questions exist as to whether a significant proportion of the extracted wealth will remain local. Geographers have noted this ethical issue and in a few cases have proffered recommendations.

Although these examples rest on matters of resource sales and environmental impacts, geographers have sustained an interest in the transportation routes, including the economic costs and benefits of the developing countries that are physically touched along the way. Chad needs to move its oil through

Nigeria, for example, and the vast resources of the landlocked Caspian Sea must cross through several countries to gain access to the sea, which is a diplomatic challenge. Locally, Mexico and the United States have been working to establish closer ties through increased land-based trade of electricity, coal, oil, and natural gas, in both directions, across their common border. This has caught the attention of geographers in both countries, who have pointed to the shared advantages. They have also identified some of the shared environmental costs when the impacts of energy development cross international borders, such as when air pollution wafts northward from power plants in the Salinas district south of Eagle Pass, Texas.

Hydropower projects always hold a particular appeal for developing countries. Not only do they provide electricity, flood protection, and, often, irrigation water but also they create jobs and national pride. In developing countries, however, they also produce some problems that have not been as publicized elsewhere. Often, these problems have been associated with the displacement of large numbers of people as reservoirs fill. In the past, the most notable examples of this pattern have been in Africa and Brazil, but currently world attention has focused on the Yangtze River in China, the Naranda River in India, and the Euphrates River in Turkey. Even in places where few people live, hydroelectric projects are often controversial, either for reasons of environmental disruption, such as on the border of Uruguay and Brazil, or for reasons of native land claims (as with the huge projects on James Bay in Canada). Although all these hydro projects have inherently geographic applicability, only a few geographers are involved in studying them.

The oldest and most traditional energy resource is biomass, usually in the form of wood from shrubs and trees, crops, and animal dung. Biomass can be used for heating, cooking, and transportation. Wood fuel use and sustainable development have been extensively researched in case studies of India, Pakistan, Tanzania, Kenya, Zimbabwe, Ghana, and Haiti. Brazil, in contrast, has supplemented wood and charcoal use with sugarcane and bagasse for ethanol fuel and cogeneration. The heat content of biomass is low by weight compared to coal and other fossil fuels, so for areas of quickly growing population many geographers have reported how such dependency quickly transforms landscapes, which often do not recover their original appearance or productivity. The relentless and growing dependency on such resources has secondary impacts on soil erosion, microclimates, and other aspects of local and regional ecological balance, which geographers have long noted and studied, especially as it pertains to the creation and implementation of government policies.

Developing countries are among the most logical candidates for the development of renewable energy resources, especially direct solar and wind power. From a geographical perspective, direct solar power, either for hot water or for electricity, holds the greatest advantage because it is not site specific; photovoltaic panels can be installed and used with little need for local technical competence or development of a delivery infrastructure. Geographers have written about the relative advantages of such sources of power for isolated people in India and several other developing countries in Latin American and Africa, and the experiences in these locations can be transferred to just about any place in the tropics. Regarding wind power, some developing countries, such as Mongolia, have great potential, but wind (like geothermal) is spatially concentrated and immobile. The number of studies should increase in the coming decades as the quest for sustainable development continues.

SEE ALSO THE FOLLOWING ARTICLES

Conservation of Energy Concept, History of • Economic Geography of Energy • Economic Thought, History of Energy in • Energy in the History and Philosophy of Science • Geopolitics of Energy • Nuclear Power: Risk Analysis • Public Reaction to Energy, Overview • Sustainable Development: Basic Concepts and Application to Energy

Further Reading

Cleveland, C. J., and Kaufmann, R. K. (1991). Forecasting ultimate oil recovery and its rate of production: Incorporating economic forces into the model of M. King Hubbert. *Energy J.* **12**(2), 17–46.

Cook, B. J., Emel, J. L., and Kasperson, R. E. (1990). Organizing and managing radioactive-waste disposal as an experiment. *J. Policy Analysis Management* **9**(3), 339–366.

Cook, E. F. (1976). "Man, Energy, Society." Freeman, San Francisco.

Cuff, D. J., and Young, W. J. (1986). "The United States Energy Atlas." 2nd ed. Macmillan, New York.

Elmes, G. A., and Harris, T. M. (1996). Industrial restructuring and the changing geography of the United States coal-energy system, 1972–1990. *Ann. Assoc. Am. Geogr.* **86**(3), 507–529.

Guyol, N. B. (1971). "Energy in the Perspective of Geography." Prentice-Hall, Englewood Cliffs, NJ.
Hosier, R.H. (Ed.). (1993). Urban energy and environment in Africa [Special issue]. *Energy Policy* **21**(5), 434–558.
Jackson, E. L., Kuhn, R. G., Macey, S. M., and Brown, M. A. (1988). Behavioral energy research in geography. *Can. Geogr.* **32**(2), 162–172.
Majumdar, S. K., Miller, E. W., and Panah, A. I. (Eds.). (2002). "Renewable Energy: Trends and Prospects." Pennsylvania Academy of Science, Easton, PA.
Manners, G. (1971). "The Geography of Energy." 2nd ed. Hutchinson, London.
Odell, P. R. (1986). "Oil and World Power." 8th ed. Penguin, Harmondsworth, UK.
Pasqualetti, M. J. (ed.). (1990). "Nuclear Decommissioning and Society: Public Links to a New Technology." Routledge, London.
Pasqualetti, M. J. (2000). Morality, space, and the power of wind-energy landscapes. *Geogr. Rev.* **90,** 381–394.
Sagers, M. J., and Green, M. B. (1986). "The Transportation of Soviet Energy Resources." Rowman & Littlefield, Totowa, NJ.
Smil, V. (1988). "Energy in China's Modernization." Sharpe, Armonk, NY.
Smil, V. (1998). "Energies: An Illustrated Guide to the Biosphere and Civilization." MIT Press, Cambridge, MA.
Solomon, B. D., and Ahuja, D. R. (1991). International reductions of greenhouse-gas emissions: An equitable and efficient approach. *Global Environ. Change* **1,** 343–350.
Wilbanks, T. J. (1994). Sustainable development in geographic perspective. *Ann. Assoc. Am. Geogr.* **84**(4), 541–556.
Zeigler, D. J., and Johnson, J. H. Jr. (1984). Evacuation behavior in response to nuclear power plant accidents. *Professional Geogr.* **36**(2), 207–215.

Geopolitics of Energy

AMY MYERS JAFFE
Rice University
Houston, Texas, United States

1. Introduction
2. The Stubborn Reality of the Middle East
3. World Demand Trends
4. Enter Russia
5. The Importance of U.S. Growth to Oil Producers
6. OPEC Capacity Expansion
7. Consumer Country Power in a Deregulated Setting
8. Consumer Country Strategies

Glossary

Anthropologic Belonging to the scientific study of the origin, behavior, and physical, social, and cultural development of humans.
biomass Plant material, vegetation, or agricultural waste used as a fuel or an energy source.
carbon intensive Having a high content of carbon, a primary element that when burned produces carbon dioxide, which in great amounts can produce changes to the global climate.
externality An intended or unintended cost resulting from a government policy or market operation, coming or obtained from an independent source.
fossil fuel Fuel formed from the geological past and dug out of the earth.
monopoly Exclusive possession in the trade of some commodity.
monopsony Near exclusive possession in the trade of some commodity held by a small group.
Organization of Petroleum Exporting Countries (OPEC) A cartel of oil-producing countries that work together as a group to regulate oil market prices.
renewable Able to be renewed; not depletable.
sustainable productive capacity The daily amount of oil that an individual oilfield or a group of oilfields can produce at a rate that can be sustained for more than 90 days.

World energy use is expected to increase dramatically over the next three decades, reaching some 15.3 billion tons of oil-equivalent (toe) by 2030, up from 9.2 billion toe in 2000, according to the International Energy Agency's *World Energy Outlook 2002*. This increase is the equivalent of two-thirds of current demand and will be driven by strong economic growth in the developing world, especially Asia, and by population growth, which is expected to expand by one-third during the same time period. The developing countries' share of world demand will rise from roughly one-third today to 40% by 2030. Meeting such huge requirements will be a major challenge, requiring massive investment at every stage of the energy supply chain, much of it in developing countries that currently lack adequate resources to meet their current needs. Mobilization of this investment will be a major test for policymakers into the future.

1. INTRODUCTION

More than a quarter of the world's population has no access to electricity, and close to half of the world's people rely on traditional biomass—wood, crop residues, and animal waste—as their main cooking and heating fuels. This situation is unlikely to change, barring the development of a cheap breakthrough energy technology. Current projections are that fossil fuels will remain the primary sources for energy, at approximately 90% of the increase of demand, for at least the next three decades, potentially with dire environmental consequences. Energy emissions of carbon dioxide (CO_2) are expected to grow roughly in line with energy demand unless there is faster deployment of more efficient and cleaner technologies or a larger than expected, policy-driven shift toward less carbon-intensive fuels.

According to projections made in 1993 by the Australian Bureau of Agriculture and Research Economics (ABARE), the growth in CO_2 emissions

 © 2004 Elsevier Inc. All rights reserved.

for the developing world is projected to increase by five times that for the developed world in the next decade or so. Currently, the developed countries emit approximately 60 to 70% of total anthropologic CO_2 emissions and are also responsible for 80% of current concentrations. During the coming decades, the contribution to emissions from the developing world will gain pace, even in the event that Western industrial countries begin to curb emissions in line with the Kyoto accords. In particular, the per capita emissions of four key developing countries—China, India, Indonesia, and Brazil—are expected to grow rapidly during the coming decade and beyond. In 1995, the per capita emissions of these four countries represented only 5 to 10% of the per capita emissions of the United States and were significantly lower than the world average. Over the long term, rising energy use in these countries can be expected to have a dramatic effect on global greenhouse gas (GHG) emissions (Table I).

The desire to wean the global economy—industrial and developing worlds alike—off of fossil fuels to safeguard the environment and provide a fuel source affordable across the globe poses many barriers, including technical cost, available infrastructure, and long lead times for new investments. Many forecasts predict a lowering of the growth in coal demand during the coming three decades as some societies, such as China and India, favor a switch to other sources (e.g., natural gas) for environmental reasons. But coal remains in ample supply in much of the developing world, and production costs are low. World coal reserves are enormous and, compared with oil and natural gas, are widely dispersed. Faced with the insecure nature of imported oil supply and potentially rising prices, even countries seeking environmental improvement may find themselves unable to move away from heavy reliance on coal in power and industrial sectors. As natural gas and crude oil prices have risen during the past few years, a reemergence of rising coal use has been registered. So, although it is expected that coal use will wane during the coming decades for environmental reasons, such forecasts may be deemed too optimistic (Fig. 1).

Environmental advocates would like to see global energy politics move in the direction of renewable fuels such as solar, wind, and hydropower. According to the International Energy Agency's (IEA) *World Energy Outlook 2002* (*WEO 2002*), whereas natural gas is expected to see its world share rise from 23% today to close to 30% by 2030, renewable energy is unlikely to capture a large part of the world energy market during the coming two decades or so (Fig. 2).

The reality is that even with massive investment, renewable resources are unlikely to overtake fossil fuels to any significant extent during the coming decades. Within the "green" community, advocates would like to see the relationship between oil and renewable energy flip during the next 50 years or so such that oil becomes a marginal fuel and renewable energy becomes the major supply source (Fig. 2). But such a green scenario would require massive government intervention and incredible investments of capital to level the current playing field, where fossil fuels maintain a pivotal price advantage in most markets. Thus, without major changes from technological advances or policy intervention, investment in renewables, although highly beneficial to environmental concerns, is unlikely to emerge at large enough rates to make a significant dent in rising demand for oil and natural gas.

TABLE I

Projected Emission Levels in 2010 (Millions of Metric Tons of Carbon)

Country	Year	Residential/ Commercial	Transport	Industrial and other	Final consumption	Primary requirement	Per capita metric tons of CO_2
China	1995	131.77	36.97	487.90	656.64	798.17	2.44
	2010	222.42	92.55	826.24	1141.21	1387.48	3.77
India	1995	18.14	32.77	85.95	136.86	228.28	0.90
	2010	39.14	60.23	153.78	253.15	422.49	1.37
Indonesia	1995	9.91	12.55	14.48	36.94	59.39	1.13
	2010	17.68	21.64	24.47	63.79	102.59	1.59
Brazil	1995	4.27	25.70	26.97	56.94	64.68	1.49
	2010	8.17	43.26	42.03	93.46	110.24	2.13

Source. Hartley, Medlock, Warby: Rice University, Baker Institute.
Note. U.S. per capita emissions in metric tons of CO_2 were 19.13 in 1992 and 19.69 in 1996.

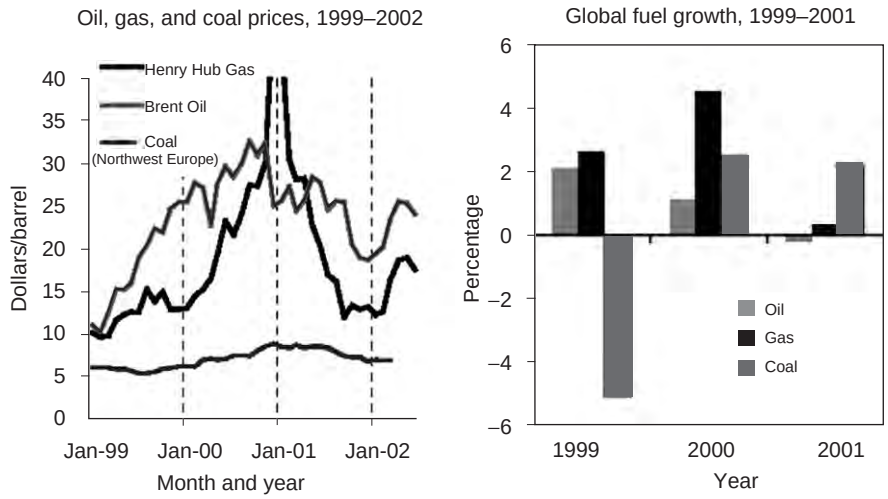


FIGURE 1 Graphs illustrating the reemergence of coal through the process of rising demand. From BP.com.

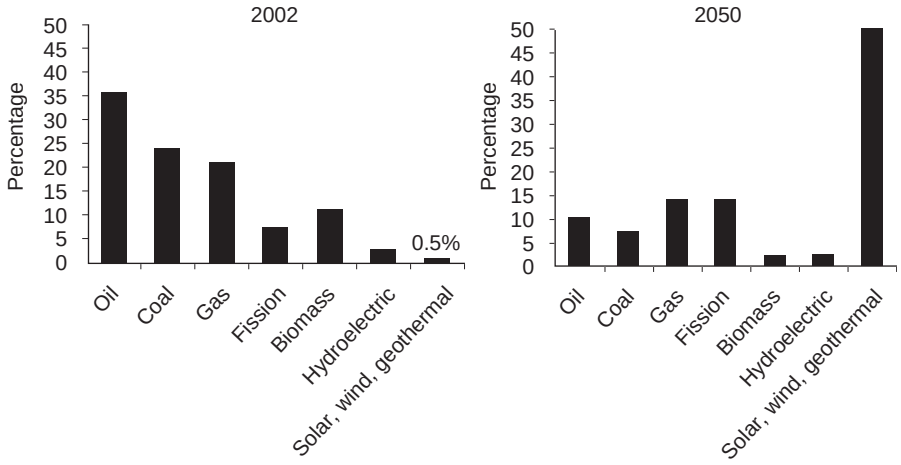


FIGURE 2 Bar graphs depicting the environmental vision for 2050. From Rice University.

For example, the share of nonhydro renewable resources in Organization for Economic Cooperation and Development (OECD) electricity generation, despite relatively proactive green policies by world political standards, is expected to rise from 2% in 1997 to only 4% in 2020 under current programs and is likely to achieve only a 9% share with significant government intervention. According to the IEA's *World Energy Outlook 2001* (*WEO 2001*), to accomplish this 9% share would require new investments of $228 billion and would result in a reduction of CO_2 emissions of only 6% over the period. Wind power capacity has grown by more than 70% since 1997 but still represents less than 1% of the world's total installed electricity capacity, and 69% of world wind facilities are located in Europe.

Nuclear power could be a significant contributor to the long-term global sustainable energy mix, but to play this role successfully, nuclear power will have to overcome serious economic and political challenges, including concerns over nuclear waste, safety, and nonproliferation. In many countries where nuclear power plays a major role, governments are exiting the generation business, with some exceptions (e.g., Finland, South Korea). As electricity markets become increasingly deregulated, nuclear facilities will have to compete on commercial grounds with generators using other fuel sources, and this may be increasingly difficult given the long lead times for siting, design, and other regulatory approvals and grassroots plant costs.

Although third-generation nuclear reactors will continue to play a major role for the next decade or so, the future rests with new fourth-generation designs that might be able to provide safer, more reliable nuclear power through smaller modular units that

will better conform to commercial needs and competitive pressures. In the meantime, until nuclear power is able to offer a competitive challenge to traditional fuels, it is assumed that the bulk of the increase in energy demand will be met with a rise in oil and natural gas production. According to *WEO 2001,* global oil demand is expected to rise by approximately 1.6% per year, from 77 million barrels/day in 2002 to 120 million barrels/day by 2030. Near three-quarters of this increase in demand will come from the transport sector, where renewable energy and nuclear energy are not expected to play a significant role, barring a major technological breakthrough.

2. THE STUBBORN REALITY OF THE MIDDLE EAST

Much ink has been spilled chronicling the world's rising dependence on Middle East oil. According to *WEO 2001,* more than 60% of the world's remaining conventional oil reserves are concentrated in the Middle East. A quarter of these reserves sit in Saudi Arabia alone. The Middle East is currently supplying more than one-third of world oil demand (Fig. 3).

This percentage could rise significantly in the future, depending on policies in consumer countries and on the pace of development of new resources and technologies. The IEA forecasts that the Middle East share of world oil supply will rise to more than 40% by 2020 under its *World Energy Outlook 2000* (*WEO 2000*) reference case. The U.S. Department of Energy, in one forecast, even predicts that the need for OPEC oil could rise from 28 million barrels/day in 1998 to 60 million barrels/day in 2020, with the majority of supply coming from the Middle East, especially Saudi Arabia.

Iran, Iraq, Syria, Sudan, and Libya currently produce approximately 8 million barrels per day or roughly 10% of the world oil supply. Saudi Arabia alone is responsible for nearly 10% of the world oil supply and holds a unique position in oil markets. It maintains the largest share of spare idle production capacity of any other nation in the world. The kingdom is the only oil producer in the world that can single-handedly replace, within a short period of time, the total loss of exports for any other oil producer on the globe. No other nation currently has enough spare capacity to claim this role. Saudi Arabia is also the world's largest exporter, in past years selling nearly 100% more than its next largest export competitor, Russia (Fig. 4).

Saudi Arabia's cushion of spare capacity, which represents more than two-thirds of all global spare capacity, has provided security and stability to world oil markets for two decades. The kingdom derives its international clout from this custodial role and is unlikely to relinquish it, especially in the face of weakening oil market trends. The kingdom must worry about market share considerations in the face of rising investment in Russia, Africa, Iran, and Canada as well as long-term economic considerations. Oil revenues are critical to the health and well-being of the Saudi economy. Therefore, Saudi Arabia can ill afford to jeopardize its coveted position as a stable and reliable supplier from which its geopolitical influence and long-term economic interests derive.

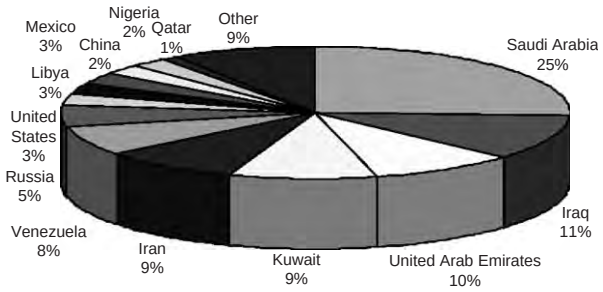


FIGURE 3 Pie chart depicting countries containing proven world resources in the year 2000.

3. WORLD DEMAND TRENDS

World oil demand is expected to rise at a rate of roughly 1.5 to 2.0% per annum over the next two decades, from about 76.4 million barrels/day in 2001

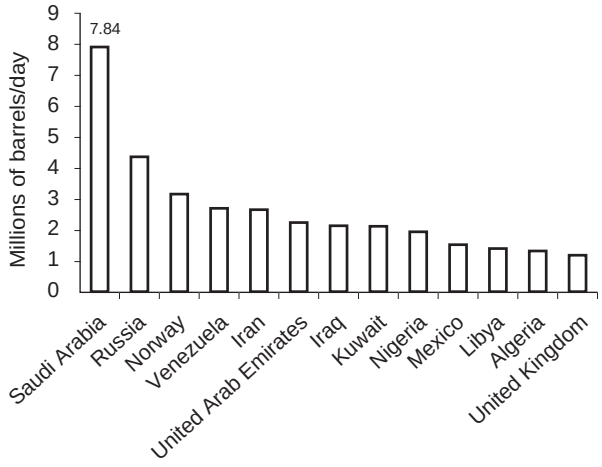


FIGURE 4 Bar graph depicting top world oil net exporting countries for the year 2000.

to 89 million barrels/day in 2010 and 104 million barrels/day by 2020, according to *WEO 2002*. During this same period, Asian demand is anticipated to increase from 28 to 35% of total world demand. Asia's oil consumption, at more than 19 million barrels/day, already exceeds that of the United States. Currently, only approximately 40% of Asia's oil supplies are produced inside the region.

By 2010, total Asian oil consumption could reach 25 million to 30 million barrels/day, of which 18 million to 24 million barrels/day will have to be imported from outside the region. Asian oil demand averaged 21.5 million barrels/day during the first half of 2003, whereas local oil production totaled only some 6.2 million barrels/day, leaving a gap of more than 15.3 million barrels/day that was met by imports, mainly from the Middle East and to a lesser extent from Africa, according to the *Petroleum Intelligence Weekly* database. This is up from a gap of more than 11 million barrels/day in 1998. As Asia develops, oil demand is expected to grow two to three times more rapidly than in the industrialized West, reaching approximately 29 million barrels/day by 2010, according to the "business as usual" forecast scenario of the IEA. Regional oil supplies are not expected to keep pace in any corresponding fashion. The IEA forecasts that Asian oil production will average only 6.4 million barrels/day by 2010, only a modest gain from current levels. In a study by Jaffe and Barnes, analysts expected somewhat larger gains of 1 million barrels/day or so but far less than the 6 million to 11 million barrels/day projected gain in demand.

More than 60% of Middle East exports already go to Asia and nearly 70% of all Asian imports come from Middle East oil producers. Some 84% of all crude oil refined in Singapore comes from the Middle East, whereas 78% processed in Japan originates from that region. The IEA projects that the Asia Pacific region will be importing 20 million to 24 million barrels/day from the Middle East by 2020. These oil linkages could present new political and strategic challenges in light of the emerging geopolitics of the post-September 11, 2001, period. The defense of oil supplies from the Persian Gulf is now achieved nearly single-handedly by the United States. Although it is hard to predict how the U.S. "War on Terror" will affect this role during the coming decades, expanding oil ties between major Asian powers and Middle East exporters will certainly influence East Asian strategic thinking, especially in relation to regional sea lanes issues. The result is likely to be some alteration in big power relations in East Asia during the coming years that might test traditional relationships.

Saudi Arabia and other Middle East oil producers have already grasped the growing importance of the Asian markets and have been building bridges to the region. In July 2001, Saudi Arabia overtook the United Arab Emirates as Japan's primary supplier of oil. Saudi Arabia's exports to Japan grew by 20.7% to 27.40 million barrels that month, whereas the United Arab Emirates' exports to Japan dropped by 19.7% to 26.25 million barrels, according to a *Kyodo News Service* report. Pricing flexibility will be key to the kingdom's strategy to win over any market to which it lays claim.

China's current refinery configurations are currently not suitable to refine large amounts of the lower quality supplies that are produced in Persian Gulf countries such as Iraq, Iran, Saudi Arabia, and Kuwait. This explains the relatively small amount of crude imports from these countries (Fig. 5). Without refinery upgrades, China is projected to have only a little more than 1.35 million barrels/day of capacity to process this lower quality Persian Gulf oil by 2005 (Table II).

However, low-sulfur supplies from Yemen and Oman are expected to fall during the coming years as production from older reservoirs begins to decline. Thus, China is seeking light crude oil supplies from other sources, including Africa, and is seeking to make new upstream investments to lock up such supplies. The state oil company, CNOOC, secured Indonesian supplies recently with its $585 million purchase of the Widuri and Cinta fields in Indonesia from Repsol-YPF, thereby adding 100,000 barrels/day of light crude oil that can be processed quickly through China's existing refinery capacities.

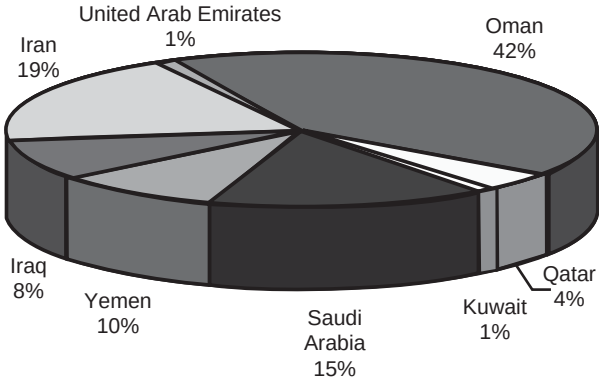


FIGURE 5 Pie chart depicting Chinese oil imports from Middle East countries for the year 2000.

TABLE II

China Refinery Capacity by Crude Tolerance (Thousands of Barrels/Day)

Crude Type	1998	2002	2005
Sweet: <0.9% S	4350	4050	3700
Medium sweet: 0.90–1.15% S	160	550	750
Sour: 2% S or higher	240	600	750–1350
Total	4750	5200	5200–5800

Source. Asia Pacific Energy Consulting. "China's Oil Strategy towards the Middle East." Working paper, APEC.

SINOPEC is expected to follow suit and has been negotiating for an exploration deal in Nigeria. Chinese state firms are also seeking exploration and production deals in other African states as well as Libya.

To protect its piece of the growing Chinese market, Saudi Arabia has shown a willingness to help China upgrade its refining capacity to use more Saudi crude oil. In late 2001, Saudi Aramco, ExxonMobil, and Fujian Petrochemical took another step toward the ultimate completion of a 240,000-barrels/day upgrade of a major Fujian refinery by signing an agreement for a joint feasibility study.

In addition, Saudi Aramco is now the largest shareholder in the Thalin refinery and has been negotiating with SINOPEC to expand the refinery at Maoming. Also, China has made a deal for 10 million tons of Saudi oil annually for a 50-year period, with SINOCHEM processing part of the oil at the new joint venture Qingdao refinery and the rest at its Dalian facility, signaling a close and growing import/export relationship between the two countries.

4. ENTER RUSSIA

Mirroring an important international trend, Saudi and other Middle East producers will find competition for growing Asian markets from Russia, which in the aftermath of the September 11 attacks has made clear its intentions to increase oil and gas production with an eye to enhancing exports to the West and Northeast Asia.

Russian firms have been able to marshal capital on their own to revive the Russian oil industry, and Russian oil production rose by more than 640,000 barrels/day from September 2000 to September 2001. The following year, Russian production rose nearly 600,000 barrels/day, from 6.95 million barrels/day in December 2001 to 7.55 million barrels/day in August 2002. Russian production continued to rise to 8.385 million barrels/day by September 2003. Moscow's political leaders, as well as leaders from Russia's corporate oil and gas sector, are portraying the country's oil firms as stable sources of supply, willing to add output to the market to sustain reasonable prices. In the view of Russian leaders, this new geopolitics of energy "can help Moscow gain both economically and politically," according to Morse and Richard.

If concrete plans of the Russian oil firms and their Central Asian counterparts come to fruition during the coming years, rising oil exports from the former Soviet Union could rise by another 2 million barrels/day during the next 5 years or so and move closer to rivaling those of Saudi Arabia. This threat has grabbed the imagination of Middle East policymakers inside OPEC, who called on Russia late in 2001 to cease and desist from stealing OPEC's market share. Moscow initially responded with conciliatory talk toward the end of 2001 and again in March 2002, but Russian oil production rose 600,000 barrels/day during the first three quarters of 2002 versus 2001 levels. And Russian officials have publicly announced Moscow's intentions to increase its oil production during the coming years to help the West diversify its oil supply. Russian government officials and private company chairmen also posited the possibility of Russia establishing an international strategic stockpile of Russian oil that could be used to serve countries that are not currently able to amass strategic stocks. The Russian stockpile would be used to supply markets during times of market emergency and to dampen price run-ups.

Four Russian oil giants have announced plans to build a proposed $2 billion pipeline and deepwater year-round port at Murmansk, in the north of Russia, that could be used to export oil in very large crude carriers (VLCCs) to distant markets such as the United States. A feasibility study of the port project, which is slated to handle up to 1 million barrels/day of exports, is being financed by Lukoil, Yukos, Sibneft, and TNK. The project has been considered for U.S. Exim Bank support. Russian companies see the U.S. sales as a means to diversify its markets so as to avoid oversupplying Europe with growing sales from the former Soviet Union.

Traditionally, Russian oil exports have been bound for Europe and have not confronted Middle East oil producers' efforts to build market share in the United States and Asia. But more recently, there has been a political push in both Moscow and Beijing

to facilitate growing cooperation between Russian and Chinese firms in oil and gas investments and trade. A similar warming of relations on energy trade issues has taken place between Russia and Japan.

The view from Beijing now is that Russia is a key future hydrocarbon supplier. China hopes to receive supplies from new investments by Western firms at Sakhalin and from new field developments within Russia, especially by Yukos. A natural gas pipeline plan involves some 9.5 billion cubic meters (bcm)/year of hydrocarbon exports to Japan and potentially Korea and China from the Sakhalin Islands. Within the next 5 years or so, Russian supply to East Asia from the Sakhalin oilfields under current development are also likely to provide upward of 500,000 barrels/day of incremental supply.

Beyond Sakhalin, there are other prospects from East Siberia, where Yukos, the Russian firm holding perhaps the largest potential reserves in the area, is already in discussion with CNPC/Sinopec on joint projects.

TABLE III

Importance of U.S. Growth to Oil Producers

	1991	2000
U.S. and Asia Pacific share of world oil demand		
United States	24%	25%
Asia Pacific	22%	28%
U.S. net oil imports as a share of global oil trade		
Total	32.34 million barrels/day	42.67 million barrels/day
Net U.S. imports	6.79 million barrels/day	10.20 million barrels/day
U.S. share of increase in global oil trade, 1991–2000	33.02%	
U.S. share of increase in OPEC's output, 1991–2000	55.58%	

Source. Morse, E. (2002, February 21). "The Price of Crude Oil: Markets, Politics, and Geopolitics." Lecture, Columbia University.

5. THE IMPORTANCE OF U.S. GROWTH TO OIL PRODUCERS

So much attention has been paid to the increase expected from Asian markets that analysts often forget to note the continued importance of the U.S. market to OPEC. U.S. net oil imports rose from 6.79 million barrels/day in 1991 to 10.2 million barrels/day in 2000, while global oil trade (i.e., oil that was exported across borders from one country to another) rose from 32.34 million to 42.67 million barrels/day. The U.S. share of the increase in global oil trade over the period was a substantial 33%. In OPEC terms, the U.S. import market was even more significant—more than 50% of OPEC's output gains between 1991 and 2000 (Table III).

Saudi Arabia, in particular, has been keenly aware of the importance of the U.S. market and careful to keep nagging political issues between the United States and the kingdom on a separate track from oil policy. The kingdom has been diligently maintaining its U.S. market share during recent years and threw down the gauntlet in 1997 to punish Venezuela for challenging its grip on the important U.S. market. Following the oil price collapse of 1997–1998, newly elected Venezuelan President Hugo Chavez rescinded state oil firm PDVSA's plans to expand capacity to 7 million barrels/day by 2007. Instead, Venezuela announced revised plans to grow its production more gradually, to 5.5 million barrels/day by 2006. However, prolonged political strife inside the country and a major oil workers strike in 2003 not only discouraged further investment by foreign companies (which were already discouraged by rising tax rates) but also left much of Venezuela's oil industry infrastructure damaged, lowering its potential to produce and refine crude oil. Venezuela's oil production capacity was already dropping during recent years given constraints on investment capital and a tough geological decline curve at aging fields, but damage sustained during the oil workers strike and near civil war in the country led to a more rapid reduction in sustainable capacity to 2.2 million barrels/day at the end of 2003, down from a peak of 3.7 million barrels/day in 1998.

Saudi Arabia may face a new challenge to its U.S. market share from a regional exporter, but this time north of the border. Analysts for the kingdom's oil industry say that they are watching with concern the rising output profile for Canada. The start-up of new tar oil sands projects and other offshore projects has put Canada on track for the largest jump in oil output in several years, despite lower capital spending by industry that is often cited as evidence that production will decline. Canada's production has increased from 2.155 million barrels/day in 2000 to 2.3 million barrels/day in mid-2003.

Companies have spent more than $7 billion in oil sands projects since 1996 and are expected to

allocate an additional $25 billion by 2007. Production, which has already reached 800,000 barrels/day, could rise to 1.5 million barrels/day or more by 2010 if currently proposed projects meet their targets (Table IV).

The potential for Canadian heavy crude and oil sands development is huge. Of the 2.5 trillion barrels of crude bitumen (oil sands) resources in place in Canada, approximately 12% (or 300 billion barrels) is thought to be recoverable, a figure comparable to the proven reserves of Saudi Arabia. This is in addition to extra heavy oil resources. In Venezuela, where some U.S. oil companies are working on heavy crude upgrading projects, more than 1.2 billion barrels of bitumen are thought to exist.

During past decades, the high costs of mining and processing tar sands and extra heavy crude oil were somewhat prohibitive at approximately $21 per barrel. But engineering and technological advances are lowering costs rapidly. Analysts estimate that oil sands projects "all in" costs (i.e., investment and operating) have fallen to as little as $8 to $9 per barrel and could fall to $6 to $7 per barrel during the coming years as industry continues to make new advances. However, cost variables in Canada have been somewhat unstable given a tight labor market, and cost overruns have been seen at several of Canada's top existing oil sands projects, including Syncrude, Athabasca, and Millennium.

African supply may also become a key focus of efforts by the West to diversify away from the Middle East. Of the 14.5 million barrels/day of exports projected to be needed from Africa or the Middle East by 2010 in the Atlantic Basin market to fill its supply gap, at least half will be able to be sourced from Africa and North Africa. Angola, a non-OPEC producer, is projected to see increased production, from 750,000 to more than 2 million barrels/day by 2010, as new fields are developed in its deep offshore zone at full-cycle costs of between $3 and $5 per barrel. Chinese national oil companies are also showing interest in African upstream investment, with CNPC already firmly planted in Sudan and Chinese companies showing expanded interest in Nigeria, Chad, Niger, Equatorial Guinea, Algeria, Tunisia, and Libya.

TABLE IV

Approved Canadian Oil Sands Projects

Project	2001	2002	2005	2010
Syncrude	225	250	350	465
Suncor	130	220	260	400
Shell/Chevron	—	155	155	155
Albian/Koch/T.N.			95	190
Canadian Natural Resources				300
Gulf Resources				100[a]

Source. Petroleum Argus newsletter.
[a] Proposed for 2013.

6. OPEC CAPACITY EXPANSION

Over the past few years, there has been a dramatic shift in the internal politics of OPEC, reflecting changes of political leadership at the highest levels and of broader policies of key members. The result has been greater cohesion inside the producer group and a clearer articulation and implementation of goals and aspirations.

The mid-1990s were characterized by OPEC disunity and overproduction, so much so that the cartel members themselves were questioning the future viability of the organization. Venezuela's ambitious campaign to increase oil productive capacity from 2.8 million barrels/day in 1991 to 7 million barrels/day by 2010, combined with the gradual rise in Iraqi oil export rates through the auspices of the UN Oil for Food program, made it nearly impossible for the producer group to agree on a workable production sharing agreement.

During this period of the mid-1990s, market share considerations were paramount for most OPEC countries with spare capacity. Saudi Arabia, in particular, had to concern itself with both the short- and long-range implications of Venezuela's market expansion plans, which were expected to have a significant bearing on the kingdom's ability to maintain its sales to the United States. Venezuela's government was committed to supplementing state revenue by increasing oil exports and in 1992 began a new policy to allow international oil companies (IOCs) to invest in the country's once nationalized oilfields. This campaign contributed to a rise in Venezuelan output, from roughly 2.5 million barrels/day through the 1980s to between 3.5 million and 3.7 million barrels/day by 1997. Venezuela refused to comply with any OPEC production sharing agreements, greatly damaging the producer group's ability to manage oil markets and providing a strong disincentive for Saudi Arabia to contribute a large cutback in output to defend oil prices.

This political stumbling hit a dramatic snag in during 1997–1998 when an unexpected economic

meltdown hit much of Asia and precipitated an unexpected drop in Asian oil use. Asian oil demand fell to 18.18 million barrels/day in 1998, down from 18.53 million barrels/day, a drop of 1.9%. This compared with a 4.5% rise in Asian oil demand between 1996 and 1997. The slowdown in Asia contributed to a major change in the global oil supply–demand balance, limiting the growth in worldwide oil demand to 74.24 million barrels/day in 1998, up only 0.5% from 1997 compared with more typical annual growth rates of 2 to 3% during the mid-1990s (Table V).

TABLE V

Oil Demand per Region, 1996–1998 (Million Barrels/Day)

	1996	1997	Percentage change	1998	Percentage change
World	72.03	73.9	2.6	74.24	0.5
United States	18.25	18.62	2.0	18.7	0.4
European Union	13.03	13.10	0.5	13.39	2.2
Asia	17.74	18.53	4.5	18.18	−1.9
China	3.55	11.5	11.5	3.95	−0.3

By 1997–1998, these circumstances led oil prices to collapse below $10, requiring extraordinary efforts by both OPEC and non-OPEC countries such as Mexico and Norway. The intense financial suffering of nearly all OPEC countries and key changes in government leadership in Venezuela paved the way for a major agreement among oil producers to trim output and propel oil prices back above $22 per barrel. OPEC set a price band of $22 to $28 per barrel as its target that remained in place as of this writing.

OPEC's new, more unified dynamic is rooted in several factors:

- A rise in democratization, freedom of the press and political debate, and a growing tide of anti-Americanism are bringing a greater concern for popular opinion inside OPEC countries, especially in the Middle East Gulf region, than was the case in the past. This new concern for popular sentiment is restricting the options of regional leaders to accommodate Western interests. Populations, as well as some leaders, remain bitter about the suffering that took place when oil prices collapsed during the late 1990s.
- Rising populations and economic stagnation in many OPEC countries indicate that revenue pressures have been taking precedence over other considerations.
- Lack of investment in infrastructure and oilfields over the years due to tight state treasuries and rising social pressures has greatly curtained OPEC's spare productive capacity, rendering it much easier for cartel members to agree to restrain output, at least in the short term.

Given this new dynamic and colored in part by the rise of Chavez as president of Venezuela, OPEC rhetoric during recent years has taken a more strident turn, with oil ministers defending their choices to attain a "fair" price for OPEC oil despite ramifications for global economic growth rates. The debate has become one of economic struggle for "rents" between oil producers who demand high revenues and major consumers whose economies can grow faster with low oil prices.

OPEC's rhetoric has generally been directed at OECD consumer governments that capture rents from oil sales through high national energy taxes. OPEC's anti-tax, anti-Western rhetoric comes against the backdrop of popular domestic sentiment inside OPEC countries that their governments are not doing enough to deliver economic benefits to a substantial portion of their populations. Leaders in OPEC countries cannot be seen as delivering benefits to Western consumers at the expense of their own citizens because such perceptions would leave regimes more vulnerable to public attack and more susceptible to the efforts of opposition parties and groups.

Although political factors have contributed to OPEC's unity over the past 2 years, the more limited amount of spare capacity that has to be held back by most member states is also a supportive factor. Indeed, oil prices tend to track changes in OPEC spare capacity quite closely (Fig. 6).

In 1985, when oil prices collapsed, OPEC was estimated to have some 15 million barrels/day of shut-in production capacity, equal to perhaps 50% of its theoretical capacity (Iran and Iraq were at war with one another at the time) and 25% of global demand. By 1990, when Iraq invaded Kuwait, spare capacity globally was still 5 million to 5.5 million barrels/day, the amount of oil taken off the market by the UN embargo. That was approximately 20% of OPEC's capacity, and 8% of global demand, at the time. During the winter of 2003, before OPEC's seasonal cuts, spare capacity was a negligible 2% of global demand, a level that left oil markets highly susceptible to disruption and supported a run-up in prices to more than $30 per barrel.

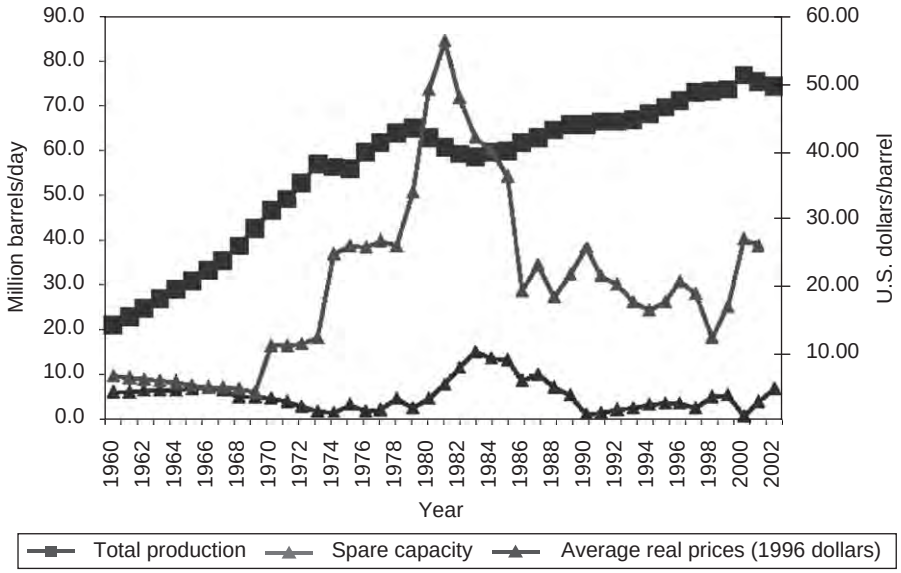


FIGURE 6 Line graph depicting the relationship between oil prices and the OPEC spare capacity. From Hetco Trading.

OPEC has been helped during recent years by irregularity in Iraqi production rates given the regime in Baghdad's aggressive posture to use the withdrawal of its UN-monitored oil exports as a political tool. In 2003, the U.S. military campaign in Iraq greatly curtailed the country's oil exports when facilities were looted and sabotaged in the immediate aftermath of the war. Iraq's export rates fell from just over 2 million barrels/day prior to the U.S. campaign to approximately 500,000 barrels/day in late 2003. In addition, other OPEC members have also benefited from a drop in capacity in Venezuela and Kuwait, both of which have been struggling against technical problems and natural declines in key fields. Venezuelan oil production capacity has fallen from 3.7 million barrels/day in 1998 to approximately 2.2 million barrels/day currently. Kuwait has lost capacity in certain western and northern fields such as Raudhatain, limiting its output capacity by several hundreds of thousands of barrels per day. Sustainable Kuwait capacity has recently been pegged at 2 million barrels/day.

However, the challenge to OPEC will increase during the coming years as OPEC member countries begin to anticipate an expected restoration of Iraqi capacity over time. Several individual OPEC countries have a backlog of new fields that can be brought on line soon. Several countries, such as Nigeria, Algeria, and Iran, have been quietly expanding capacity and anticipate possible production rises over the next few years. Countries such as Algeria and Nigeria, whose new fields involve foreign oil company investment, will increasingly be under additional pressure from these IOC investors to obviate OPEC agreements and allow new fields to move forward at optimum production rates. Meeting these pressures will be difficult for OPEC if market demand for its oil does not increase significantly as expected due to a slowdown in economic growth or other inhibiting factors.

Iran is forging ahead with capacity expansions and could reach more than 5 million barrels/day by 2008 based on investments through buy-back agreements with IOCs and National Iranian Oil Company investment. However, this ambitious plan will be dependent on the success of major IOC field developments, including Azadegan, South Pars, Sirri, Gagh Saran, and the Ahwaz area. Nigeria has plans to expand production in its offshore area, and this could (if successful) raise production to more than 3.5 million barrels/day by 2006.

Prior to the U.S. campaign in early 2003, Iraq had commercial export capacity of approximately 2.2 million barrels/day via Turkey and the Gulf, with additional smuggling capacity through Syria, Iran, and Jordan. U.S. officials were hopeful that Iraq would be able to resume exports at prewar levels by late 2003. Iraq has expressed a desire to expand its productive capacity to 6 million barrels/day, and analysts believe that Iraq could quickly raise its oilfield productive capacity back to 3.6 million barrels/day if security can be restored across the

country and Iraq's upstream oilfield sector were opened to foreign oil company investment. However, for Iraq to be able to export these higher volumes, extensive repairs will be needed to both its oilfields and its export infrastructure.

There will be pressure inside OPEC to bring Iraq back into the production quota system as its production capacity is restored. Key members may insist that Iraq be initially limited to its historical allocation of a quota at parity with that of Iran and then hold its export increases to levels in line with market demand. However, Iraq will have pressing economic reconstruction needs to take into consideration as well, raising questions, depending on the future politics inside the country, as to whether it will be in a position to opt for policies that emphasize solidarity with OPEC.

OPEC's current capacity is estimated at 29.5 million barrels/day but could rise to as much as 36.4 million barrels/day by 2005 if new fields under development come on-line as expected. If OPEC were to expand access to foreign IOC participation, particularly in key countries now off limits to Western investment, capacity could expand to as much as 44 million barrels/day by 2010, leaving very little room for growth from non-OPEC supplies to take place without pressuring oil prices over the next 8 years or so.

7. CONSUMER COUNTRY POWER IN A DEREGULATED SETTING

As discussed previously, the United States is a major buyer of imported crude oil, and its imports represent a large share of the market for internationally "traded" oil. Given the large scale of U.S. purchases, incremental U.S. acquisitions of oil affect the overall international market price of oil. Stated another way, the cost of each marginal barrel is higher than the price paid for that barrel given that this additional purchase affects the costs of all oil consumed. From the perspective of the United States, this constitutes an externality in the sense that consumers will base their decisions regarding the quantity of oil (and oil products) purchased on the average, rather than the marginal, cost of oil. On the other hand, the fact that the United States faces a rising supply curve for oil gives it monopsony power. To the extent that the United States—or a group of consuming countries that include the United States or is of a comparable scale—take concrete actions to reduce the size of its purchases, it can lower the market price of oil.

It has been well established that OPEC frequently changes its price targets in response to changes in market demand. Discussion of the size of a monophony (power) wedge that is the "difference between the current price of oil and the marginal cost of adding a barrel of demand" was the subject of an Energy Modeling Forum working group and other exercises.

OPEC's response to efforts by the OECD to reduce oil demand has varied over time. Initially, during the 1980s as the OECD demand receded, OPEC maintained its output path, letting oil prices decline. This result produced the highest buying power wedge size and reinforced the viability of consumer country policies that sought to exercise this monopsony power. However, OPEC can also maintain its price preference path and allow its market share to decline. This appears to be its strategy during recent years. But the continuation of the production sharing discipline needed to maintain this approach may become harder to sustain in the future.

Consuming countries have implemented two key approaches to dealing with OPEC given the increasing trend toward oil price deregulation inside the OECD. One is that governments have organized to hold strategic stocks of oil. The second is the imposition of consumer taxes on oil, thereby reducing demand for oil. As discussed previously, OPEC has engaged in a public relations battle against Western consumer taxes on oil.

Prior to deregulation, during the 1970s, crude oil was sold mainly through exclusive, long-term, fixed-priced contract arrangements with a handful of major suppliers. Crude prices throughout the world were regulated by national governments at fixed levels. Under this system, oil sellers could pose a clear threat to any individual buyer. A clear thread linked a disruption in supply to the shortfall of actual deliveries. A particular barrel lost in exports was neatly matched to a particular barrel lost to a particular buyer that had a contract for that barrel. Little recourse was available. Barrels were moved around the world like pieces on a chessboard, and there was no other *deus ex machina* to mitigate or alter the impact of any supply cutoff if one's particular barrels were lost. Lucky buyers whose supply chains remained intact could continue to receive their oil at prices locked by long-term fixed-price contracts at precrisis levels. In a word, some buyers were safe and others were not. This system created a psychology that enhanced OPEC's monopoly power and left individual consuming nations vulnerable to the possibility of a supply cutoff. The

oil crisis of 1973 confirmed fears that oil producers might use this leverage to adverse political ends.

Given the high costs of a cutoff under this fixed price system (i.e., regulated price system), large consumers, oil distributors, and middlemen had direct incentives to carry ample inventories of spare supplies. The possibility that oil from any particular supplier could be disrupted without recourse was reason enough to pay the costs to carry extra supplies on hand. Prices that could be passed on to consumers were in many cases inflexible and regulated, and competitors that were not subject to a cutoff would have a clear cost advantage.

Today's world is radically different. Crude oil prices have been "deregulated," that is, set by market forces through market auction practices that allocate supplies to the highest bidder. Price arbitrage will ensure that a price increase for one buyer or location will be a price increase for all buyers and locations.

In the new free-wheeling marketplace for oil, buyers cannot insulate themselves from price jumps by holding long-term contracts given that prices in these contracts are pegged to the spot market. But in today's market, one can always get the oil one desires, even in a major supply disruption, by simply paying a higher price.

At the end of the day, as the market clears, such as it did during the Iraqi invasion of Kuwait in 1990, buyers with imports from countries that were not subject to disruption are in no better position than buyers that had all of their imports subject to interruption. In an unfettered spot market, shortages will manifest themselves not in a physical inability to find supplies but rather in the rising price of oil. The market will "clear" or allocate available physical supplies to the highest bidder.

In this new commodicized market, commercial players such as distributors and middlemen have less incentive to carry inventory, especially if supply disruptions are infrequent. Higher costs during times of supply cutbacks can simply be passed on to end users, reducing the business risks that motivated the accumulation of commercial inventories previously. Although the existence of a large spot market trading at unregulated prices certainly appears to work against OPEC's ability to "blackmail" any particular buyer, structural reductions in commercial inventories can augment the producer group's power to the extent that alternative supplies from countries outside of OPEC are reduced.

The "public goods" aspect of oil market stability elevates the accumulation of oil inventories to a domain beyond the commercial industry. Individual players in the oil market will not take into account the externalities of low commercial stocks such as increased energy price volatility and its negative distributive effects on poorer consumers or its aggregate negative macroeconomic effects. Thus, government plays a role in ensuring that adequate inventories are on hand to maintain orderly markets and to counter the temptation of suppliers with monopoly power from taking advantage of short-term tightness in oil markets. To achieve these ends, government-held stock levels must be credible to convince oil producers that efforts to exploit temporary market tightness by further cuts in production to achieve even higher oil prices will not be successful. Such attempts to extract additional rents from consuming countries would be countered by the release of sufficient inventories to offset any cuts in production contemplated by producers. The larger the government-held stocks and the more consuming governments that participate in such a stock holding program, the more effective it is likely to be in serving as a deterrent to OPEC's monopoly power in deregulated markets.

Market deregulation has left consumers more susceptible to the ill effects of energy price volatility. However, consuming country governments have been able to reduce the negative effects of price variability by increasing energy efficiency and reducing dependence on oil through the use of hefty consumer taxes on oil. The net effect of such taxes is to discourage wasteful use of energy by consumers while collecting some of the rents that would otherwise accrue to oil producers. Furthermore, large oil consumption taxes can force OPEC to accept lower prices, as occurred throughout most of the late 1980s and 1990s.

When OPEC's monopoly power strengthens due to short-term market tightening, the incentive to exploit that power is tempered by the fact that increases in monopoly rents will not accrue entirely to producers but rather must be shared with consuming countries that have high energy taxes.

8. CONSUMER COUNTRY STRATEGIES

OPEC's ability to raise prices in the short run to levels that may be damaging to the economies of major consuming countries poses a major policy challenge to the governments of consuming countries. Rising oil prices threaten consumer country national interest in several ways. Oil price volatility

can inhibit investment, inhibit economic growth, and spur inflation in major economies that purchase consumer country exports.

The burden of rising energy import costs also threatens social stability in key regional consuming countries such as India and Pakistan and the Southeast Asia region. Moreover, supply constraints also make it easier for governments or subnational groups to threaten vital interests of the United States, Japan, and their allies.

Thus, consuming countries have a clear interest in undertaking policies that will undermine OPEC's short- and long-term ability to act as a cartel to inflate oil prices. Policies undertaken in conjunction with other consuming nations are likely to be more effective than policies undertaken individually by increasing the strength of the monopsony wedge.

Policy options include the following:

- Expand and restructure the IEA to better reflect the change in consumer country demand trends.
- Encourage investment in diverse non-OPEC oil resources.
- Develop and deploy alternative energy technologies.
- Develop federal policies to continue enhancing energy efficiency.
- Maintain or expand taxes on oil and gasoline.

The IEA was created a quarter of a century ago as a mutual protection society of OECD countries. Designed as a political grouping to prevent any oil-producing countries from using oil exports as a political instrument to influence the foreign policies of IEA members, the IEA was formed at a time when the OECD countries dominated global energy consumption. Today, the IEA excludes the most rapidly growing energy-consuming countries in the world, including China, India, and Brazil. And, as a result, these new consumers become vulnerable economically during times of disruptions and become potentially vulnerable to political pressures of producers.

During recent years, there has been discussion about increasing the number of member countries inside the IEA, and South Korea has joined the organization. Other countries, such as China and India, are investigating creating national strategic oil stockpiles. The IEA invited both China and India to participate as observers in recent meetings and is pursuing options for finding mechanisms for major non-IEA oil-consuming countries to participate in joint stockpiling emergency programs.

However, the extent of effectiveness of the IEA system will depend on oil market developments, including Asian demand trends. The member countries of the IEA now represent a smaller portion of the oil market than they did at the time of the IEA's formation in 1977. As oil demand growth in Asia expands during the coming decade, new strains could come to the international system if new policies are not put into place. The omission of key consumer countries from Asia in the global emergency stockpiling system will increasingly put pressure on the effectiveness of limited existing stocks in the OECD countries. Moreover, tensions created by Asian "free riding" or possible "hoarding" actions during a crisis could hinder the IEA's ability to stabilize international oil markets in the future.

The OECD countries comprising the IEA accounted for 42.3 million barrels/day out of a total world oil use of 60.6 million barrels/day in 1977 or approximately 70% of world oil demand. The United States alone consumed 30% of the world's oil used in 1977. Asia Pacific demand at that time, at 10.1 million barrels/day or roughly 16% of world oil demand, was a less critical component in the world oil use situation.

By 2001, the OECD share of world oil use had declined to 62% of total world demand, while Asia Pacific use had grown to 28%, overtaking the U.S. share of 25%. Asian economic powers Japan, South Korea, Australia, and New Zealand are OECD members and, as such, are part of the IEA system now. A few Asian countries, such as Pakistan and China, have announced plans to create strategic stocks, but other key Asian oil consumers, such as India, Taiwan, Thailand, and The Philippines, are not. As their share of world oil demand grows, this disconnect between Asia's size and importance as a consumer region and its lack of energy policy coordination with other large oil-consuming countries (and/or the IEA) will create new problems and challenges for international oil markets and the international economic system. In particular, it is important that large Asian consuming countries not purchase and "hoard" oil during an IEA stockpile release because this activity would reduce the effectiveness of a stock release to calm markets and prevent oil supply shortages. The best possible win–win scenario would be for new links to develop between the IEA and other large consumer countries or consumer country groups.

Despite the obvious benefits from an Asian stockpiling group or system, Asian countries have been reluctant to commit to the expense involved in creating such a system. Although an Association of Southeast Asian Nations (ASEAN) delegation visited

Japan in 2000 to study the industrialized country's emergency stockpiling system, the cost per capita was considered "high" by many ASEAN delegates. Press reports indicate that ASEAN energy ministers have focused on security of sea lanes and international gas pipelines, rather than on joint stockpiling arrangements, at recent meetings.

Meetings between IEA officials and Chinese state planning officials have similarly failed to produce an agreement of coordination. The Chinese government has announced its intention to build a strategic petroleum reserve, but so far the government has failed to define the reserve's size or possible location publicly. It is considering requiring state oil companies to hold stocks comparable to 10% of their import levels. There has also been no official verification that China would like to work cooperatively with its Asian neighbors or OECD countries to develop measures to safeguard against disruptions in oil supplies from the Middle East. In March 2001, the State Council and the State Development Planning Commission included a "strategic petroleum reserve" as a key project in the 10th 5-year plan, but continued debate about its size, location, and funding among party leaders, State Council agencies, National People's Congress deputies, and the three central-owned state oil and gas companies has called into question the quick construction of such a reserve.

Although informal programs to encourage stockpiling by developing world countries would have a positive impact, such efforts cannot replace the more effective tool of centralized coordination with the IEA. Centralized efforts are needed so that international norms and standards can be met during a crisis. The United States and Japan are reviewing ways in which the IEA can work with key countries that are not members of the IEA to encourage them to define their strategic oil stockpile requirements and to build strategic stocks (or create minimum inventory requirements for industry). Russia has also made pronouncements recently, offering to establish a Russian strategic oil stockpile to be used to supply countries lacking in their own strategic stocks. This concept was aired at the U.S.–Russia Energy Summit in Houston, Texas, in October 2002.

In the post-September 11 climate, consumer governments are increasingly discussing enhancing development of backstop technologies or promoting alternative energy sources that can serve to reduce the need for fossil fuel. In this practice, backstop technologies create an incentive for oil producers to avoid oil price shocks and supply disruptions for fear that the new technologies would be released and used, permanently eliminating sales markets. Alternative energy supplies provide ready substitutes if an increase in the price of oil is too extreme, and they can shield the economy from the negative impact from disruption of any one fuel source.

The deployment of improved car technology could have a dramatic effect on future oil demand trends as well as play a major role in lowering CO_2 emissions by advancing fuel efficiency. The benefits to the role that enhanced energy efficiency has played in protecting the domestic Japanese economy from oil price variability are well known. Japan did not experience a severe recession after the 1979–1980 price shock, whereas the United States, United Kingdom, and Germany, all of which were less energy efficient at the time, did experience a painful loss in economic output.

It has been shown that the lower a country's energy consumption to gross domestic product (GDP) ratio or the shorter the period that oil prices will remain higher, the lower the cost of the trade-off between inflation and GDP loss. New technologies on the horizon could allow more gains in energy efficiency. Such technologies may include microturbines for distributed power markets, improved car technologies, and household solar technologies. OECD governments should encourage the deployment of these technologies in the marketplace through tax incentives or other vehicles in an effort to reduce their individual exposure to OPEC's monopoly power.

SEE ALSO THE FOLLOWING ARTICLES

Economic Geography of Energy • Environmental Injustices of Energy Facilities • European Union Energy Policy • External Costs of Energy • Geographic Thought, History of Energy in • Global Energy Use: Status and Trends • Nationalism and Oil • OPEC, History of • World Environment Summits: The Role of Energy

Further Reading

Bohi, D. R., and Toman, M. A. (1996). "The Economics of Energy Security." Kluwer Academic, Boston.

Chinese energy majors in long march upstream. (2002). *Petroleum Intelligence Weekly,* January 28, p. 3.

Fujian Petrochemical, ExxonMobil, Aramco sign agreement for submission of Fujian project joint feasibility study. (2001). *Business Wire*, November 15.

Gately, D. (1982). OPEC and the buying power wedge. *In* "Energy Vulnerability" (J. L. Plummer, Ed.), pp. 37–57. Ballinger, Cambridge, MA.

Hartley, P., Medlock, K., and Warby, M. (n.d.) "First Things First: Development and Global Warming." Working paper. Baker Institute, Rice University.

Huntington, H. G. (1998). Crude oil prices and U.S. economic performance: Where does the asymmetry reside? *Energy J.* **19**(4), 107–132.

Jaffe, A.M., and Barnes, J. (n.d.). "Energy Trade in Asia." Policy forum paper, Johns Hopkins University, School of Advanced International Studies.

Jaffe, A. M., and Lewis, S. W. (2002). China's oil diplomacy. *IISS Survival* **44**(1), 115–133.

Jaffe, A. M., and Soligo, R. (2002). The role of inventories in oil market stability. *Q. Rev. Econ. Finance* **42,** 401–415.

Japan's crude oil imports fall for 4th month in July. (2001, August 24). *Kyodo News Service.*

Manning, R. (2000). "The Asian Energy Factor." Palgrave, New York.

Maugeri, L. (2003). Not in oil's name. *Foreign Affairs* **82**(4), 165–174.

Morse, E. L., and Richard, J. (2002). The battle for energy dominance. *Foreign Affairs* **81**(2), 16–31.

Ng, E. (2001, November 7). ASEAN not in favour of "expensive" oil stockpile. *Agence France Presse.*

Russia's TNK pushes for new markets, oil reserve sites for urals. (2002). *Oil Daily,* October 4.

Saudi Arabia's petrochemical industry diversifies to face challenges. (1999). *Oil and Gas Journal,* August 16, p. 65.

Yergin, D. (1991). "The Prize." Simon & Schuster, New York.

Yergin, D., and Hillenbrand, M. (Eds.). (1982). "Global Insecurity." Houghton Mifflin, Boston.

Geothermal Direct Use

JOHN W. LUND
Geo-Heat Center, Oregon Institute of Technology
Klamath Falls, Oregon, United States

1. Utilization
2. Equipment
3. Economic Considerations
4. Future Developments

Glossary

agribusiness application Agriculture and aquaculture developments; in this article, it includes the heating of the various applications, such as greenhouses and fish ponds.
aquaculture pond/raceway heating Heat energy provided to ponds and raceway in order to optimize the growing of fish, shrimp, alligators, or other aquatic species.
balneology The use of mineral waters for therapeutic purposes.
direct use The use of the heat energy rather than its conversion to other forms such as electricity; for example, a furnace used for space heating.
district heating Energy provided to a group of buildings or even to an entire community, generally in the form of hot water or steam supplied from a central plant to customers through a pipeline distribution system.
energy cascading The multiple use of an energy source by various applications, with each successive one using an increasingly lower temperature in order to maximize the efficiency of the system.
geothermal energy Energy that can be extracted from the earth's internal heat. The heat is produced by the radioactive decay of thorium, potassium, and uranium that exist in the earth's molten core.
greenhouse heating Heat energy provide to a glass or plastic enclosure in order to optimize the growing of tree seedlings, flowers, and vegetables.
heat pumps A device used to transfer heat from a low-temperature resource to a high-temperature reservoir, thus providing the higher temperature for space heating. The process can be reversed to provide space cooling.
industrial application Manufacturing, machining, assembling, and producing products for consumption; in this article, it includes the heating of the various applications.
space heating Energy provided to a room and building for heating purposes.

Direct or nonelectric utilization of geothermal energy refers to the immediate use of the heat energy rather than to its conversion to some other form such as electrical energy. The primary forms of direct heat use are for swimming, bathing, and balneology (therapeutic use); space heating and cooling, including district heating; agriculture (mainly greenhouse heating and some animal husbandry), aquaculture (mainly fish pond and raceway heating), and industrial processes; and heat pumps (for both heating and cooling). In general, the geothermal fluid temperatures required for direct heat use are lower than those for economic electric power generation.

Most direct use applications use geothermal fluids in the low to moderate temperature range between 50 and 150°C, and in general, the reservoir can be exploited by conventional water well drilling equipment. Low-temperature systems are also more widespread than high-temperature systems (above 150°C), so they are more likely to be located near potential users. In the United States, for example, of the 1350 known or identified geothermal systems, only 5% are higher than 150°C; whereas 85% are lower than 90°C. In fact, almost every country in the world has some low-temperature systems, whereas only a few have accessible high-temperature systems.

1. UTILIZATION

Traditionally, direct use of geothermal energy has been on a small scale by individuals. Recent developments involve large-scale projects, such as district heating (Iceland and France), greenhouse complexes (Hungary and Russia), or major industrial use (New Zealand and the United States). Heat exchangers are also becoming more efficient and better adapted to geothermal projects, allowing use of lower temperature water and highly saline fluids. Heat pumps utilizing very low-temperature fluids have extended geothermal developments into

 © 2004 Elsevier Inc. All rights reserved.

TABLE I

Summary of Regional Geothermal Use in 2000

	Direct use			
Region	MWt	%	GWh/year	%
Africa	125	0.8	504	1.0
Americas	4,355	28.8	7,270	13.7
Central	5		38	
North[a]	4,308		7,013	
South	42		219	
Asia	4,608	30.4	24,232	45.7
Pacific Rim	3,489		18,242	
Central/Mid-East[b]	1,119		599	
Europe	5,715	37.7	18,905	35.7
Central/east	1,284		4,405	
West/north[c]	3,872		11,307	
CIS[d]	559		3,463	
Oceania	342	2.3	2,065	3.9
Total	15,145		52,976	

[a] Includes Mexico.
[b] Includes Turkey.
[c] Includes Azores and Guadeloupe.
[d] Includes Russia, Georgia, and Armenia.

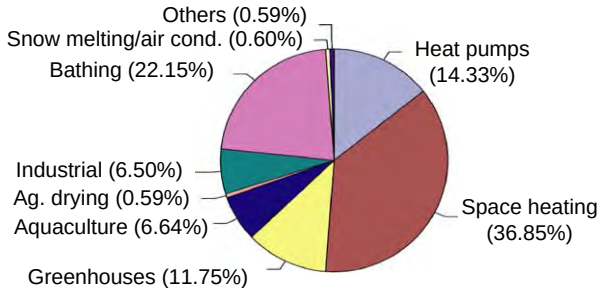


FIGURE 1 Distribution of geothermal energy use throughout the world.

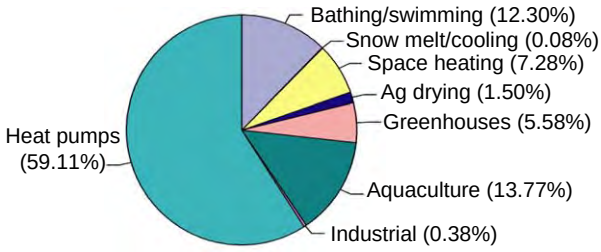


FIGURE 2 Distribution of geothermal energy use in the United States.

traditionally nongeothermal countries, such as France, Switzerland, and Sweden, as well as areas of the midwestern and eastern United States. Most equipment used in these projects is of standard, off-the-shelf design and needs only slight modifications to handle geothermal fluids.

Worldwide in 2002, the installed capacity of direct geothermal utilization was 15,145 MWt and the energy use was approximately 190,699 TJ/year (52,976 GWh/year) utilizing at least 52,746 kg/s of fluid distributed among 58 countries. A summary by region is presented in Table I. This amounts to saving an equivalent of 13.5 million tonnes of fuel oil per year (TOE). The distribution of the energy use among the various types of use is shown in Fig. 1 for the world, and for comparison that of the United States is shown in Fig. 2. The installed capacity in the United States (2000) is 4000 MWt and the annual energy use is 20,300 TJ (5640 GWh), saving 3.94 million TOE. Internationally, the largest uses are for space heating (37%) (three-fourths of which is due to district heating) and for swimming, bathing, and balneology (22%), whereas in the United States the largest use is for geothermal heat pumps (59%). In comparison, Iceland's largest geothermal energy use (77%) is for space heating [15,600 TJ/year (4,334 GWh/year)], primarily with district heating systems.

The Lindal diagram, named after Baldur Lindal, the Icelandic engineer who first proposed it, indicates the temperature range suitable for various direct use activities (Fig. 3). Typically, agricultural and aquacultural uses require the lowest temperatures, from 25 to 90°C. The amounts and types of chemicals in the geothermal water, such as arsenic and dissolved gases such as boron, are a major problem with regard to plants and animals; thus, heat exchangers are often necessary. Space heating requires temperatures in the range of 50–100°C, with 40°C useful in some marginal cases and ground-source heat pumps extending the range to 4°C. Cooling and industrial processing normally requires temperatures higher than 100°C. In terms of market penetration, the leading user of geothermal energy is Iceland, where more than 86% of the population enjoys geothermal heat in their homes from 26 municipal district heating services and 50% of the country's total energy use is supplied by direct heat and electrical energy derived from geothermal resources.

1.1 Historical Development

Although the direct use of geothermal energy has a much longer history than that of electric power generation, the data on utilization has been under

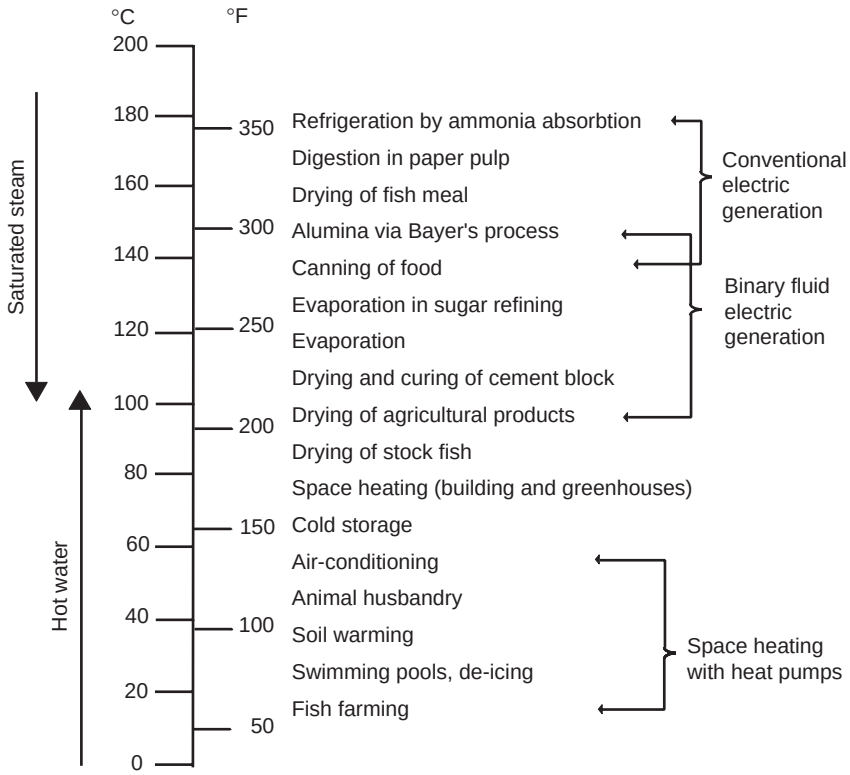


FIGURE 3 Lindal diagram.

reported. In fact, it is difficult to compare installed capacity and annual use due to the inclusion or exclusion of bathing, swimming, and balneology data. This has not been consistent; in the past, this use was not included, but in current reports it is included but not in a consistent manner. Also, values prior to 1970 were not summarized and up to 1980 could only be estimated from country descriptions in rapporteur reports. The early reports did not include China, a large user of geothermal energy for direct use, due to the political situation at the time, and they also did not include the United States even though a geothermal district heating system had been installed in Boise, Idaho, in 1892 and individual wells had been utilized in Klamath Falls, Oregon, since the 1930s for home heating. Finally, since many direct uses are small and not concentrated in one place, they are often overlooked by researchers reporting on their own country.

As a result, the 1961 United Nations (UN) conference in Rome reported only developments in Iceland, New Zealand, Italy, Japan, and Kenya. This report described district heating of 45,000 houses in Reykjavik; use of 1000 wells in Rotorua, New Zealand, for space heating; heating of 95,000 m^2 of greenhouses in Iceland; production of 21,000 tons/year of salt in Japan; the pulp and paper plant at Kawerau, New Zealand; the chemical industry at Larderello, Italy; pig raising in New Zealand; and chicken hatching in Kenya. The 1970 report of the UN meeting in Pisa included descriptions from Hungary, Iceland, Italy, Japan, New Zealand, and the USSR. As mentioned previously, China and the United States were not included. The data in Table II are based on information from the 1970 UN conference in Pisa, a report by Lawrence Livermore Laboratory in 1975, the second UN conference in San Francisco, papers by Lund in 1979 and 1982, and reports from the Geothermal Resources Council (GRC) annual meetings in 1985 and 1990, the 1995 World Geothermal Congress in Italy, and the 2000 World Geothermal Congress in Japan. Starting in 1995, geothermal heat pumps (ground-source heat pumps) were included in the reports and now comprise a significant percentage of the totals.

The large increase in installed capacity between 1980 and 1985 is due to the inclusion of pool heating at spas in Japan (onsen) and the first available data from China. The annual growth rate (based on MWt) from 1970 to 1980 was 9.3%, from 1980 to

TABLE II

Worldwide Development of Geothermal Direct Heat Utilization

	Installed energy			
Year	MWt	GWh/year	No. of countries	Participants reporting
1960			5	Iceland, Italy, New Zealand, Japan, and Kenya
1970	800 est.	2,200 est.	6	+ Hungary, and USSR
1975	1300 est.		10	+ France, the Phillipines, Turkey, and the United States
1980	1950		14	+ Austria, Czechoslovakia, Germany, and Taiwan
1985	7072	23,960	24	+ Australia, Canada, China, Columbia, Denmark, Mexico, Poland, Romania, Switzerland, and Yugoslavia
1990	8064		30	+ Algeria, Belgium, Bulgaria, Ethiopia, Greece, Guatemala, Thailand, and Tunisia; some countries not reporting
1995	8664	31,236	30	Argentina, Georgia, Israel, Macedonia, Serbia, Slovakia, and Sweden; some countries not reporting
2000	15,145	52,976	58	See Lund and Freeston (2001)

1990 it was 15.2% (which was strongly influenced by data from Japan and China), and from 1990 to 2000 it was 6.5%. The overall growth rate during the past 30 years has averaged 10.3% annually. The large increases from 1970 to 1990 (average annual increase of 12.2%) and the reduction from 1990 to present were influenced by the availability of cheap fossil fuels and the economic slowdown in Southeast Asia.

1.2 Swimming, Bathing, and Balneology

Romans, Chinese, Ottomans, Japanese, and central Europeans have bathed in geothermal waters for centuries. Today, more than 2200 hot springs resorts in Japan draw 100 million guests every year, and the "return-to-nature" movement in the United States has revitalized many hot spring resorts.

The geothermal water at Xiaotangshan Sanitarium, northwest of Beijing, has been used for medical purposes for more than 500 years. Today, the 50°C water is used to treat high blood pressure, rheumatism, skin disease, diseases of the nervous system, ulcers, and generally for recuperation after surgery. In Rotorua, New Zealand, at the center of the Taupo Volcanic Zone of North Island, the Queen Elizabeth Hospital was built during World War II for U.S. servicemen and later became the national hospital for the treatment of rheumatic disease. The hospital has 200 beds, an outpatient service, and a cerebral palsy unit. Both acidic and basic heated mud baths treat rheumatic diseases.

In Beppu on the southern island of Kyushu, Japan, hot water and steam meet many needs: heating, bathing, cooking, industrial operations, agriculture research, physical therapy, recreational bathing, and even a small zoo. The waters are promoted for "digestive system troubles, nervous troubles, and skin troubles." Many sick and crippled people come to Beppu for rehabilitation and physical therapy. There are also eight Jigokus ("burning hells") in town that show various geothermal phenomena that are used as tourist attractions.

In the area of the former Czechoslovakia, the use of thermal waters has been traced back before the occupation of the Romans and has had a recorded use of approximately 1000 years. Today, there are 60 spa resorts located mainly in Slovakia that are visited by 460,000 patients annually, usually for an average stay of 3 weeks each. These spas have old and well-established therapeutic traditions. Each sanitarium is designated for the treatment of specific diseases depending on the chemical composition of the mineral waters and spring gas, availability of peat and sulfurous mud, and climatic conditions. The therapeutic successes of these spas are based on centuries of healing tradition (balneology), systematically supplemented by the latest discoveries of modern medical science.

Bathing and therapeutic sites in the United States include Saratoga Springs, New York; Warm Springs, Georgia; Hot Springs, Virginia; White Sulfur Springs, West Virginia; Hot Spring, Arkansas; Thermopolis, Wyoming; and Calistoga, California. These sites were originally used by Indians for bathing and recuperating from battle. There are more than 115 major geothermal spas in the United States with an annual energy use of 1500 TJ.

1.3 Space Conditioning

Space conditioning includes both heating and cooling. Space heating with geothermal energy has widespread application, especially on an individual basis. Buildings heated from individual wells are popular in Klamath Falls, Oregon; Reno, Nevada; and Taupo and Rotorua, New Zealand. Absorption space cooling with geothermal energy has not been popular because of the high temperature requirements and low efficiency. Geothermal heat pumps (groundwater and ground coupled) have become popular in the United States, Sweden, and Switzerland for both heating and cooling.

An example of the use of space heating and cooling with low- to moderate-temperature geothermal energy is the Oregon Institute of Technology in Klamath Falls, Oregon (Fig. 4). Here, 11 buildings (approximately 62,000 m^2 of floor space) are heated with water from three wells at 89°C. Up to 62 liters/s of fluid can be provided to the campus, with the average heat utilization rate more than 0.53 MWt and the peak at 5.6 MWt. In addition, a 541-kW (154 tons) chiller requiring up to 38 liters/s of geothermal fluid produces 23 liters/s of chilled fluid at 7°C to meet the campus cooling base load (recently decommissioned).

1.4 District Heating

District heating originates from a central location and supplies hot water or steam through a network of pipes to individual dwellings or blocks of buildings. The heat is used for space heating and cooling, domestic water heating, and industrial process heat. A geothermal well field is the primary source of heat; however, depending on the temperature, the district may be a hybrid system, which would include fossil fuel and/or heat pump peaking.

Geothermal district heating systems are in operation in at least 12 countries, including Iceland, France, Poland, Hungary, Turkey, Japan, and the United States. The Warm Springs Avenue project in Boise, Idaho, dating back to 1892 and originally heating more than 400 homes, is the earliest formal project in the United States. The Reykjavik, Iceland, district heating system (Fig. 5) is probably the most famous. This system supplies heat for approximately 160,000 people. The installed capacity of 830 MWt is designed to meet the heating load to approximately −10°C; during colder periods, the increased load is met by large storage tanks and an oil-fired booster station.

In France, production wells in sedimentary basins provide direct heat to more than 500,000 people from 61 projects (200,000 housing units). These wells provide 40–100° C water from depths of 1500–2000 m. In the Paris basin, a doublet system (one production and one injection well) provides 70°C water, with the peak load met by heat pumps and conventional fossil fuel burners (Fig. 6).

1.5 Agribusiness Applications

Agribusiness applications (agriculture and aquaculture) are particularly attractive because they require

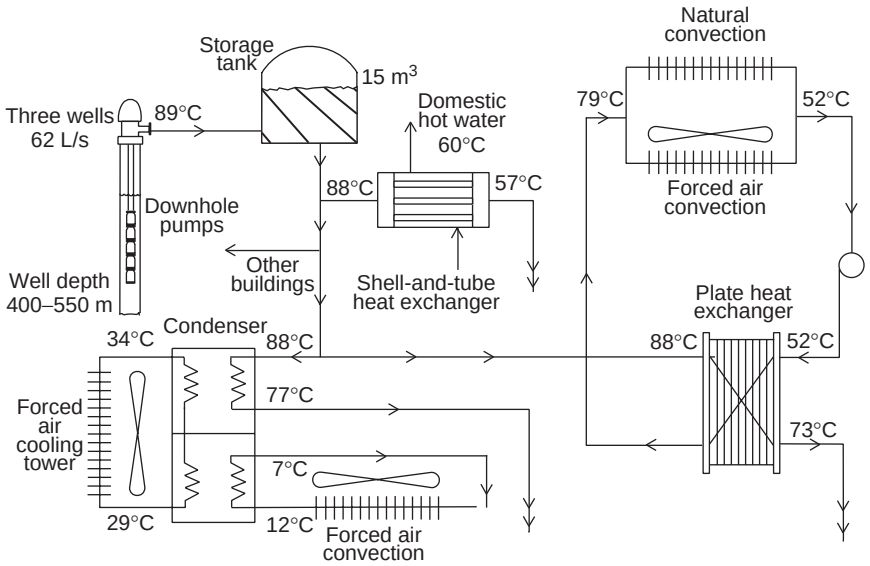


FIGURE 4 Oregon Institute of Technology heating and cooling system.

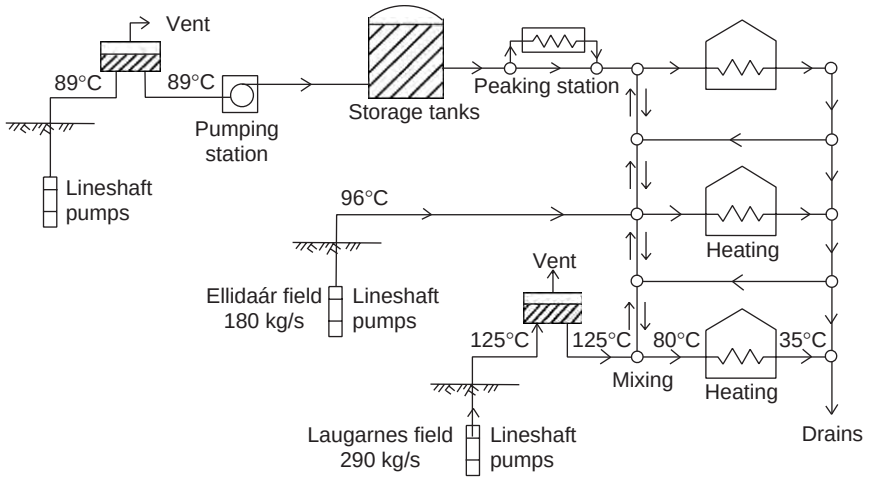


FIGURE 5 Reykjavik district heating.

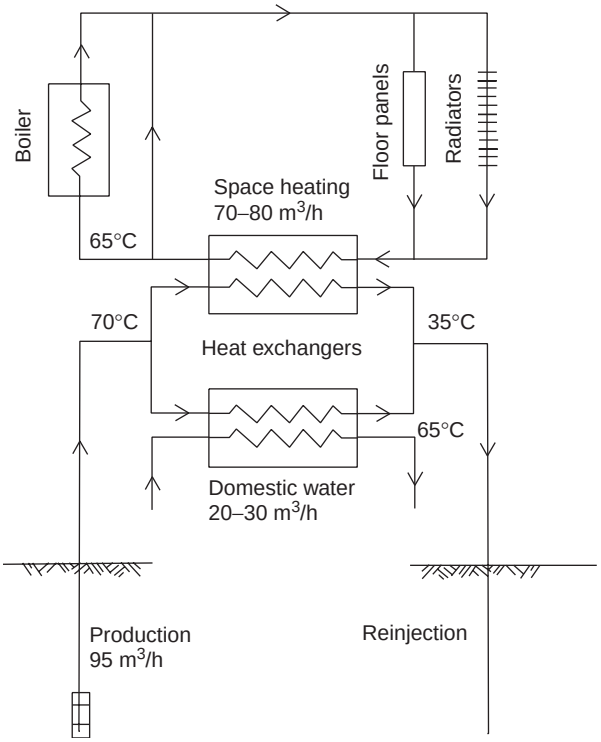


FIGURE 6 Melun l'Almont (Paris) doublet heating system.

heating at the lower end of the temperature range, for which there is an abundance of geothermal resources. Use of waste heat or the cascading of geothermal energy also has excellent possibilities. A number of agribusiness applications can be considered: greenhouse heating, aquaculture and animal husbandry, soil warming and irrigation, mushroom culture, and biogas generation.

Numerous commercially marketable crops have been raised in geothermally heated greenhouses in Tunisia, Hungary, Russia, New Zealand, Japan, Iceland, China, and the United States. These include vegetables such as cucumbers and tomatoes, flowers (both potted and bedded), house plants, tree seedlings, and cacti. Using geothermal energy for heating reduces operating costs (which can account for 35% of the product cost) and allows operation in colder climates where commercial greenhouses would not normally be economical.

The use of geothermal energy for raising catfish, shrimp, tilapia, eels, and tropical fish has produced crops faster than by conventional solar heating. Using geothermal heat allows better control of pond temperatures, thus optimizing growth (Fig. 7). Fish

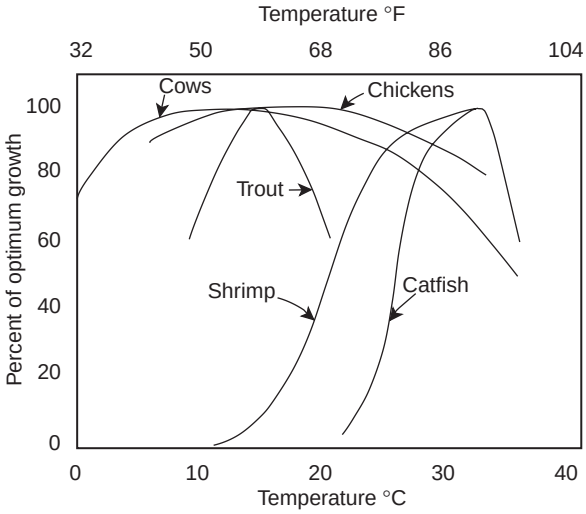


FIGURE 7 Effect of temperature on animal and fish growth.

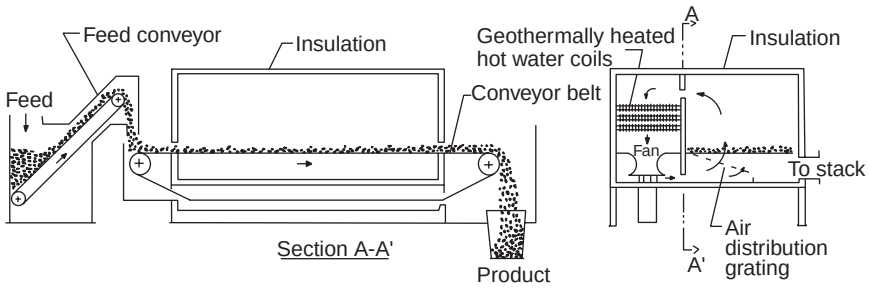


FIGURE 8 Continuous belt dehydration plant.

breeding has been successful in Japan, China, and the United States. A very successful prawn-raising operation, producing 400 tons of giant Malaysian freshwater prawns per year at $17–27/kg, has been developed near the Wairakei geothermal field in New Zealand. The most important factors to consider are the quality of the water and disease. If geothermal water is used directly, concentrations of dissolved heavy metals, flurorides, chlorides, arsenic, and boron must be considered.

Livestock-raising facilities can encourage the growth of domestic animals by a controlled heating and cooling environment. Use of an indoor facility can lower the mortality rate of newborns, enhance growth rates, control diseases, increase litter size, make waste management and collection easier, and, in most cases, improve the quality of the product. Geothermal fluids can also be used for cleaning, sanitizing, and drying of animal shelters and waste as well as assisting in the production of biogas from the waste.

1.6 Industrial Applications

Although the Lindal diagram shows many potential industrial and process applications of geothermal energy, the world's uses are relatively few. The oldest industrial use is in Larderello, Italy, where boric acid and other borate compounds have been extracted from geothermal brines since 1790. Today, the two largest industrial uses are the diatomaceous earth-drying plant in northern Iceland and a pulp, paper, and wood processing plant in Kawerau, New Zealand. Notable U.S. examples are two onion dehydration plants in northern Nevada and a sewage digestion facility in San Bernardino, California. Alcohol fuel production has been attempted in the United States; however, the economics were marginal and thus the industry has not been successful.

Drying and dehydration are important moderate-temperature uses of geothermal energy. Various vegetable and fruit products are feasible with the use of continuous belt conveyors (Fig. 8) or batch (truck) dryers with air temperatures from 40 to 100°C. Geothermally drying alfalfa, onions, pears, apples, and seaweed is an example of this type of direct use. A new development in the use of geothermal fluids is the enhanced heap leaching of precious metals in Nevada by applying heat to the cyanide process. Using geothermal energy increases the efficiency of the process and extends the production into the winter months.

2. EQUIPMENT

Standard equipment is used in most direct-use projects, provided allowances are made for the nature of geothermal water and steam. Temperature is an important consideration, as is water quality. Corrosion and scaling caused by the sometimes unique chemistry of geothermal fluids may lead to operating problems with equipment components exposed to flowing water and steam. In many instances, fluid problems can be designed out of the system. One such example concerns dissolved oxygen, which is absent in most geothermal waters except perhaps the lowest temperature waters. Care should be taken to prevent atmospheric oxygen from entering district heating waters, for example, by proper design of storage tanks. The isolation of geothermal water by installing a heat exchanger may also solve this and similar water quality-derived problems. In this case, a clean secondary fluid is then circulated through the user side of the system as shown in Fig. 9.

The primary components of most low-temperature, direct-use systems are downhole and circulation pumps, transmission and distribution pipelines, peaking or backup plants, and various forms of heat-extraction equipment (Fig. 9). Fluid disposal is either surface or subsurface (injection). A peaking

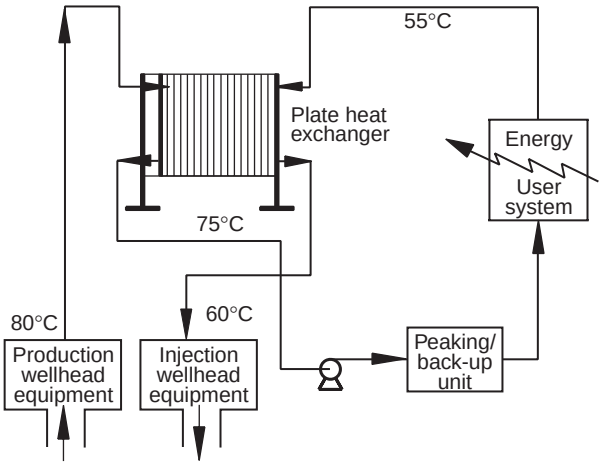


FIGURE 9 Geothermal direct-utilization system using a heat exchanger.

system may be necessary to meet maximum load. This can be done by increasing the water temperature or by providing tank storage (such as is done in most of the Icelandic district heating systems). Both options mean that fewer wells need to be drilled. When the geothermal water temperature is warm (<40°C), heat pumps are often used.

The equipment used in direct-use projects represents several units of operations. In the following sections, the major units are described in the order in which they are used to transform geothermal waters for district heating.

2.1 Downhole Pumps

Unless the well is artesian, downhole pumps are needed, especially in large-scale, direct utilization systems. Downhole pumps may be installed not only to lift fluid to the surface but also to prevent the release of gas and the resultant scale formation. The two most common types are lineshaft pump systems and submersible pump systems.

The lineshaft pump system (Fig. 10) consists of a multistage downhole centrifugal pump, a surface-mounted motor, and a long driveshaft assembly extending from the motor to the pump. Most are enclosed, with the shaft rotating within a lubrication column, which is centered in the production tubing. This assembly allow the bearings to be lubricated by oil since hot water may not provide adequate lubrication. A variable-speed drive set just below the motor on the surface can be used to regulate flow instead of simply turning the pump on and off.

The electric submersible pump system (Fig. 11) consists of a multistage downhole centrifugal pump, a downhole motor, a seal section (also called a protector) between the pump and motor, and electric cable extending from the motor to the surface electricity supply.

Both types of downhole pumps have been used for many years for cold water pumping and, recently, in geothermal wells (lineshafts have been used on the Oregon Institute of Technology campus in 89°C water for 45 years). If a lineshaft pump is used, special allowances must be made for the thermal expansion of various components and for oil lubrication of the bearings. Lineshaft pumps are preferred over the submersible pumps in conventional geothermal applications because lineshaft pumps cost less and they have a proven track record. However, for setting depths exceeding approximately 250 m, a submersible pump is required.

2.2 Piping

The fluid state in transmission lines of direct-use projects can be liquid water, steam vapor, or a two-phase mixture. These pipelines carry fluids from the wellhead to either a site of application or a steam–water separator. Thermal expansion of pipelines heated rapidly from ambient to geothermal fluid temperatures (which may vary from 50 to 200°C) causes stress that must be accommodated by careful engineering design.

The cost of transmission lines and the distribution networks in direct-use projects is significant. This is especially true when the geothermal resource is located a great distance from the main load center; however, transmission distances of up to 60 km have proven economical for hot water (e.g., the Akranes project in Iceland, where asbestos cement covered with earth has been successful).

Carbon steel is the most widely used material for geothermal transmission lines and distribution networks, especially if the fluid temperature is higher than 100°C. Other common types of piping material are fiberglass-reinforced plastic (FRP) and asbestos cement. The latter material, used widely in the past, cannot be used in many systems today due to environmental concerns; thus, it is no longer available in many locations. Polyvinyl chloride (PVC) piping is often used for the distribution network and for uninsulated waste disposal lines in which temperatures are well below 100°C. Conventional steel piping requires expansion provisions, either bellows arrangements or by loops. A typical piping installation has fixed points and expansion points approximately every 100 m. In addition, the

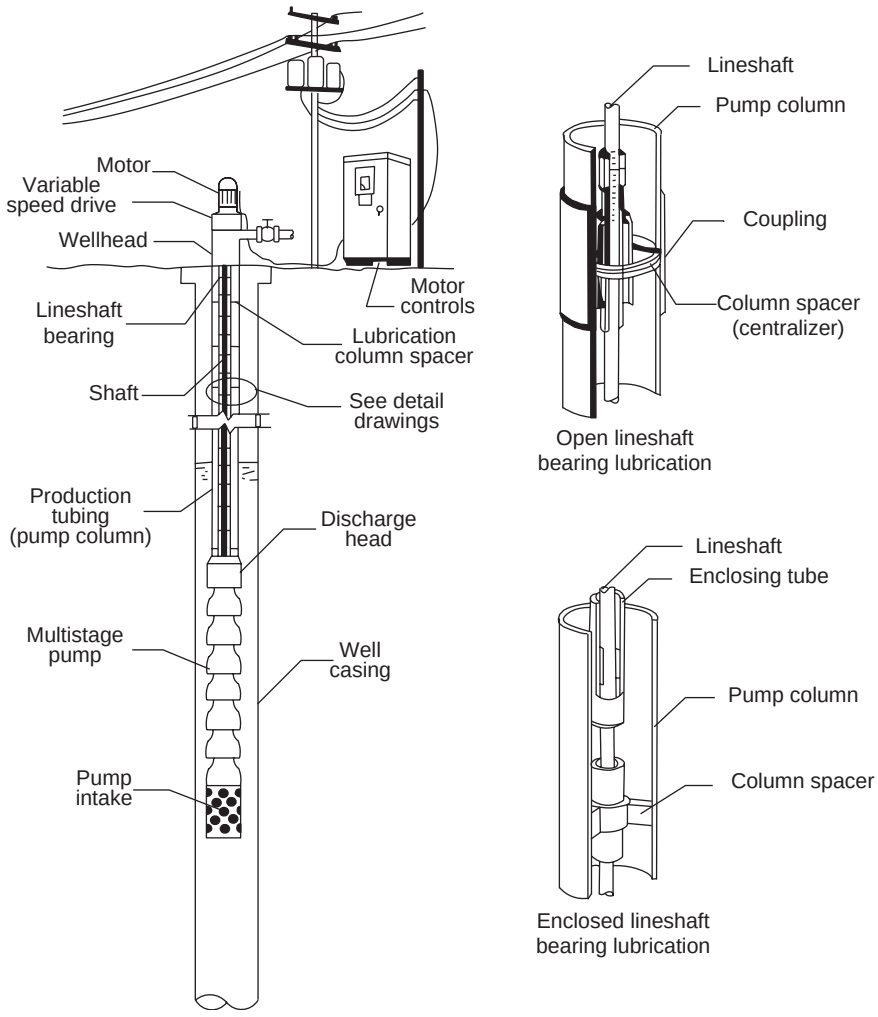


FIGURE 10 Lineshaft pump.

piping has to be placed on rollers or slip plates between points. When hot water pipelines are buried, they can be subjected to external corrosion from groundwater and electrolysis. They must be protected by coatings and wrappings. Concrete tunnels or trenches have been used to protect steel pipes in many geothermal district heating systems. Although expensive (generally more than $300 per meter of length), tunnels and trenches have the advantage of easing future expansion, providing access for maintenance and a corridor for other utilities, such as domestic water, waste water, electrical cables, and phone lines.

Supply and distribution systems can be either single-pipe or two-pipe systems. The single-pipe system is a once-through system in which the fluid is disposed of after use. This distribution system is generally preferred when the geothermal energy is abundant and the water is pure enough to be circulated through the distribution system. In a two-pipe system, the fluid is recirculated so that the fluid and residual heat are conserved. A two-pipe system must be used when mixing of spent fluids is necessary and when the spent cold fluids need to be injected into the reservoir. Two-pipe distribution systems typically cost 20–30% more than single-piped systems.

The quantity of thermal insulation of transmission lines and distribution networks depends on many factors. In addition to minimize the heat loss of the fluid, the insulation must be waterproof and watertight. Moisture can destroy any thermal insulation and cause rapid external corrosion. Aboveground and overhead pipeline installations can be considered in special cases. Considerable insulation is achieved by burying hot water pipelines. For example, burying

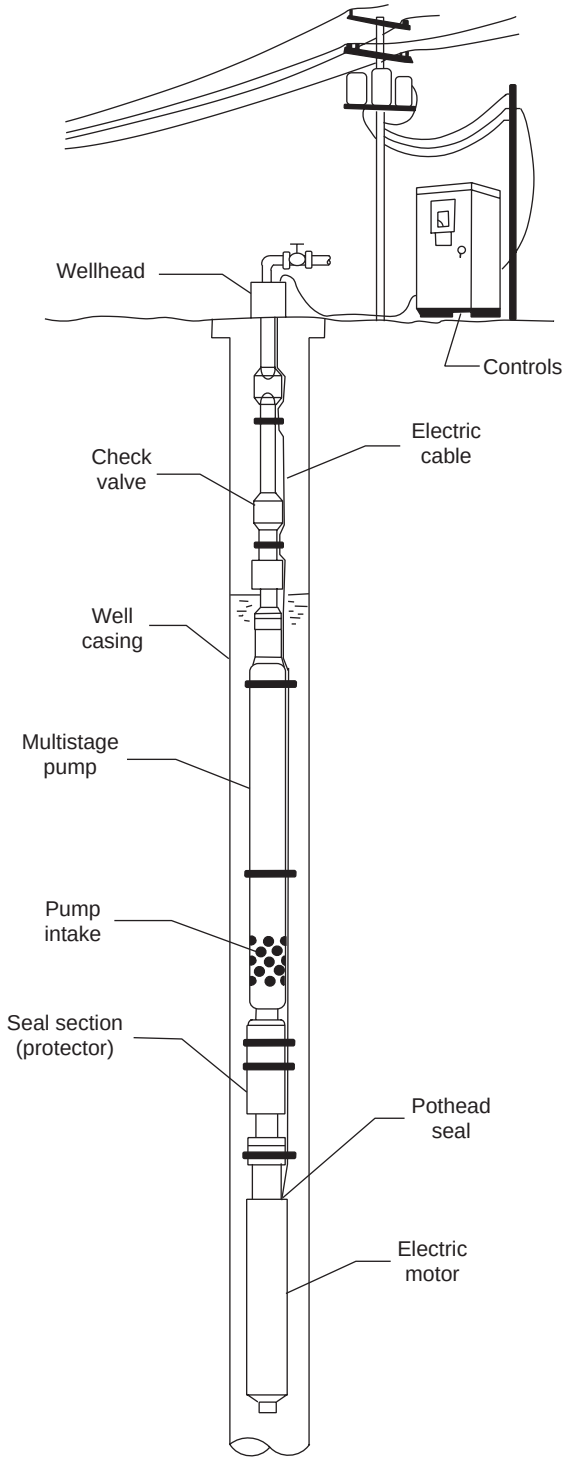


FIGURE 11 Submersible pump.

bare steel pipe results in a reduction in heat loss of approximately one-third compared to aboveground steel pipes in still air. If the soil around the buried pipe can be kept dry, then the insulation value can be

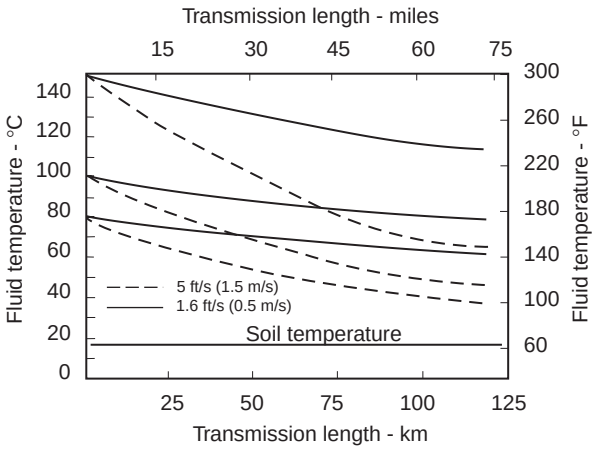


FIGURE 12 Temperature drop in hot water transmission line.

retained. Carbon steel piping can be insulated with polyurethane foam, rock wool, or fiberglass. Belowground, such pipes should be protected with PVC jacket; aboveground, aluminum can be used. Generally, 2.5–10 cm of insulation is adequate. In two-pipe systems, the supply and return lines are usually insulated, whereas in single-pipe systems only the supply line is insulated.

In flowing conditions, the temperature loss in insulated pipelines is in the range of 0.1–1.0°C/km, and in uninsulated lines the loss is 2–5°C/km (in the approximate range of 5–15 liters/s flow for 15-cm diameter pipe). It is less for larger diameter pipes; for example, a <2°C loss is experienced in the aboveground, 29-km long and 80- to 90-cm-diameter line (with 10 cm of rock wool insulation) from Nesjavellir to Reykjavik, Iceland. The flow rate is approximately 560 liters/s and 7 h are required to cover the distance. Uninsulated pipe costs approximately half that of insulated pipe and thus is used where temperature loss is not critical. Pipe material does not have a significant effect on heat loss; however, the flow rate does. At low flow rates (off-peak), the heat loss is higher than that at greater flows. Figure 12 shows fluid temperatures, as a function of distance, in a 45-cm-diameter pipeline insulated with 50 mm of urethane.

Several examples of aboveground and buried pipeline installations are shown in Fig. 13. Steel piping is used in most cases, but FRP or PVC can be used in low-temperature applications. Aboveground pipelines have been used extensively in Iceland, where excavation in lava rock is expensive and difficult. In the United States, belowground installations are more common to protect the line from vandalism and to eliminate traffic barriers.

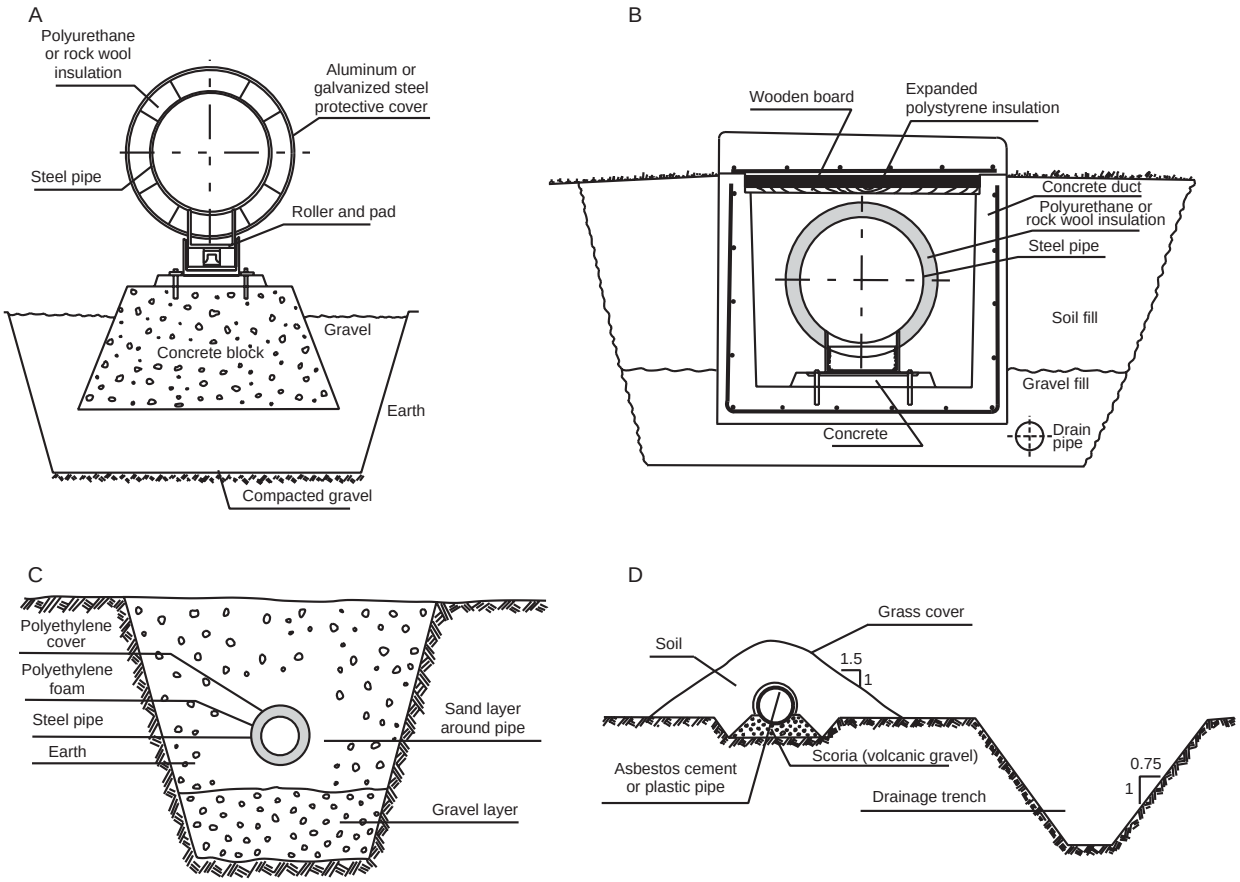


FIGURE 13 Examples of above- and belowground pipelines: (A) aboveground pipeline with sheet metal cover; (B) steel pipe in concrete tunnels; (C) steel pipe with polyurethane insulation and polyethylene cover; and (D) asbestos cement pipe with earth and grass cover.

2.3 Heat Exchangers

The principal heat exchangers used in geothermal systems are the plate, shell-and-tube, and downhole types. The plate heat exchanger consists of a series of plates with gaskets held in a frame by clamping rods (Fig. 14). The countercurrent flow and high turbulence achieved in plate heat exchangers provide for efficient thermal exchange in a small volume. In addition, compared to shell-and-tube exchangers, they have the advantages of occupying less space, are easily expandable when addition load is added, and cost 40% less. The plates are usually made of stainless steel, although titanium is used when the fluids are especially corrosive. Plate heat exchangers are commonly used in geothermal heating situations worldwide.

Shell-and-tube heat exchangers may be used for geothermal applications but are less popular due to problems with fouling, greater approach temperature (the difference between incoming and outgoing fluid temperature), and larger size.

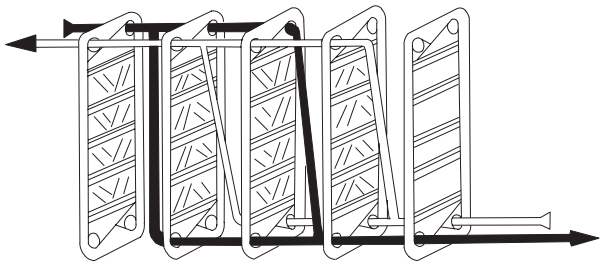


FIGURE 14 Plate heat exchanger.

Downhole heat exchangers eliminate the problem of disposing of geothermal fluid since only heat is taken from the well. However, their use is limited to small heating loads, such as heating of individual homes, a small apartment house, or business. The exchanger consists of a system of pipes or tubes suspended in the well through which secondary

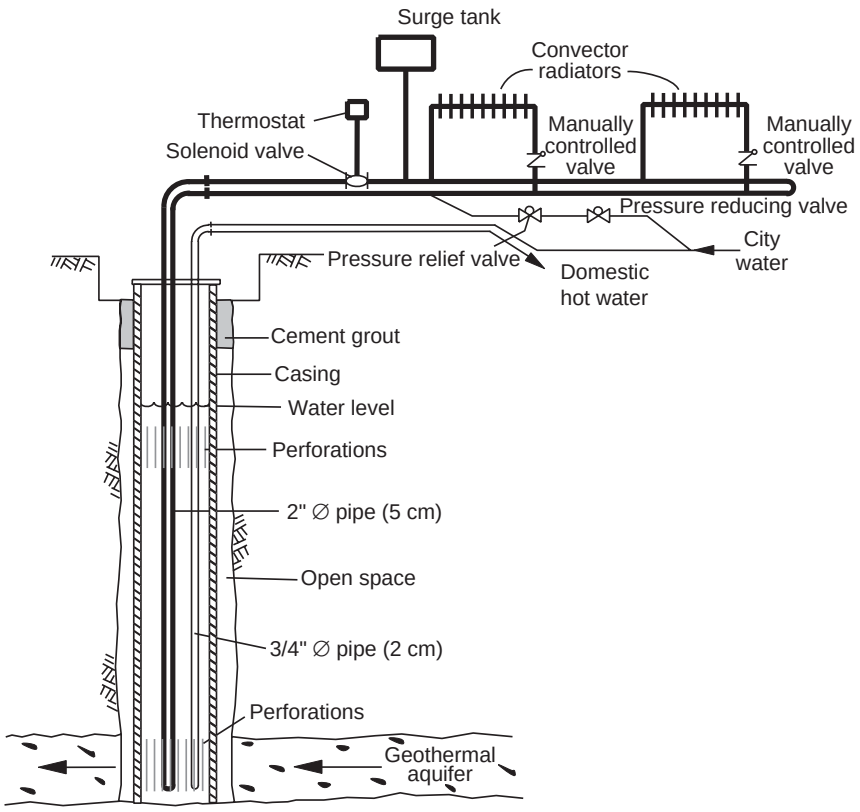


FIGURE 15 Downhole heat exchanger (typical of Klamath Falls, OR).

water is pumped or allowed to circulate by natural convection (Fig. 15). In order to obtain maximum output, the well must have an open annulus between the wellbore and casing, and perforations above and below the heat exchanger surface. Natural convection circulates the water down inside the casing, through the lower perforations, up into the annulus, and back inside the casing through the upper perforations. The use of a separate pipe or promoter to increase the vertical circulation has proven successful in older wells in New Zealand.

2.4 Heat Pumps

Ground-coupled and groundwater (often called ground-source or geothermal) heat pump systems are being installed in great numbers in the United States, Sweden, Switzerland, and Germany. Groundwater aquifers and soil temperatures in the range of 5–30°C are being used in these systems. Ground-source heat pumps utilize groundwater in wells or by direct ground coupling with vertical heat exchangers (Fig. 16). Nearly every state in the United States, especially in the midwestern and eastern states, utilizes these systems, which are in part subsidized by public and private utilities. It is estimated that approximately 70,000 groundwater systems, and more than 210,000 closed-loop vertical and 170,000 horizontal systems, are in use.

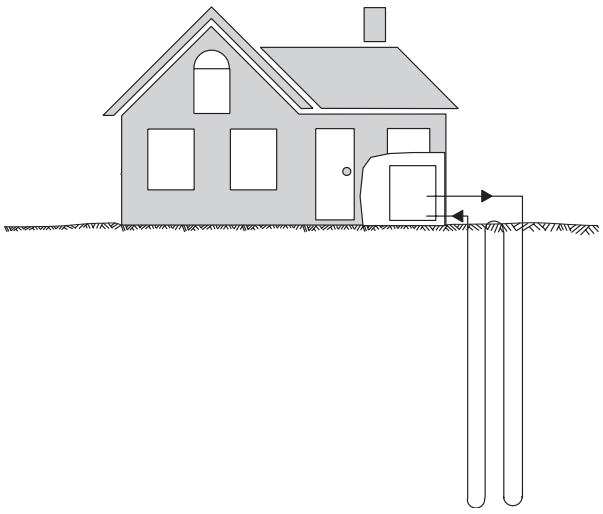


FIGURE 16 Typical ground-source heat pump installation.

Like refrigerators, heat pumps operate on the basic principle that fluid absorbs heat when it evaporates into a gas and likewise gives off heat when it condenses back into a liquid. A geothermal heat pump system can be used for both heating and cooling. Water-to-air and water-to-water heat pumps are adaptable to geothermal energy. Heat pumps with heating capacities ranging from less than 3 kW to more than 1500 kW are available.

2.5 Convectors

Heating of individual rooms and buildings is achieved by passing geothermal water (or a heated secondary fluid) through heat convectors (or emitters) located in each room. The method is similar to that used in conventional space heating systems. Three major types of heat convectors are used for space heating: forced air, natural air flow using hot water or finned tube radiators, and radiant panels (Fig. 17). All these can be adapted directly to geothermal energy or be converted by retrofitting existing systems.

2.6 Refrigeration

Cooling can be accomplished from geothermal energy using lithium bromide or ammonia absorption refrigeration systems. The lithium bromide system is the most common because it uses water as the refrigerant. However, it is limited to cooling above the freezing point of water. The major application of lithium bromide units is the supply of chilled water for space and process cooling. They may be either one- or two-stage units. The two-stage units require higher temperatures (approximately 160°C), but they also have high efficiency. The single-stage units can be driven with hot water at temperatures as low as 77°C (such as at Oregon Institute of Technology; Fig. 4). The lower the temperature of the geothermal water, the higher the flow rate required and the lower the efficiency. Generally, a condensing (cooling) tower is required, which will add to the cost and space requirements.

For geothermally driven refrigeration below the freezing point of water, the ammonia absorption system must be considered. However, these systems are normally applied in very large capacities and have seen limited use. For the lower temperature refrigeration, the driving temperature must be at or higher than approximately 120°C for a reasonable performance. Figure 18 illustrates how the geothermal absorption process works.

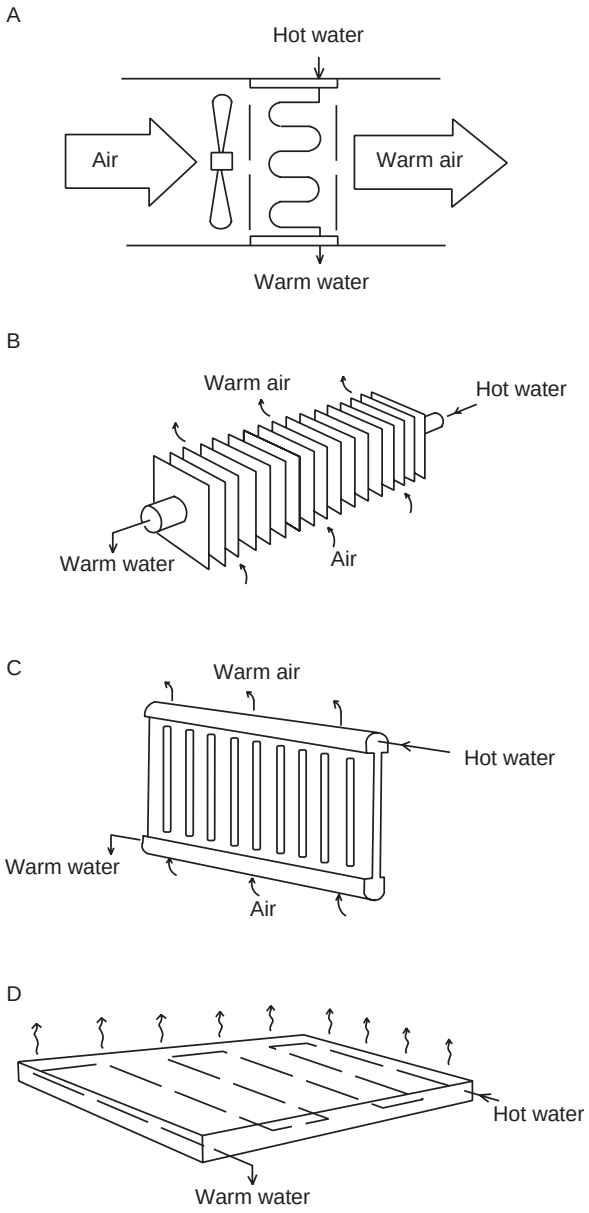


FIGURE 17 Convectors: (A) forced air; (B) material convection (finned tube); (C) natural convection (radiator); and (D) floor panel.

3. ECONOMIC CONSIDERATIONS

Geothermal projects require a relatively large initial capital investment, with small annual operating costs thereafter. Thus, a district heating project, including production wells, pipelines, heat exchangers, and injection wells, may cost several million dollars. In contrast, the initial investment in a fossil fuel system includes only the cost of a central boiler and distribution lines. The annual operation and

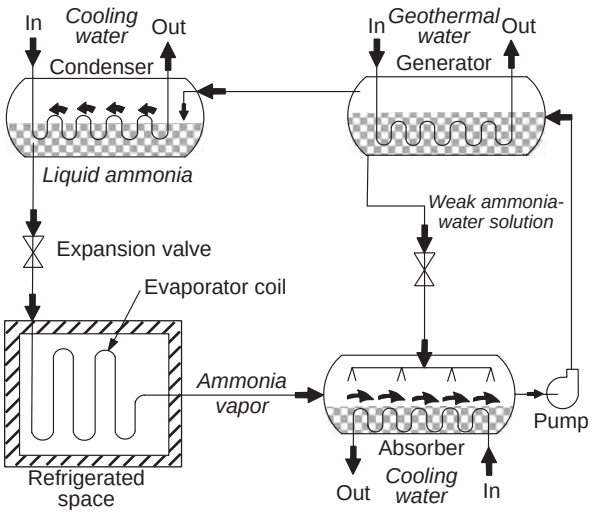


FIGURE 18 Geothermal absorption refrigeration cycle.

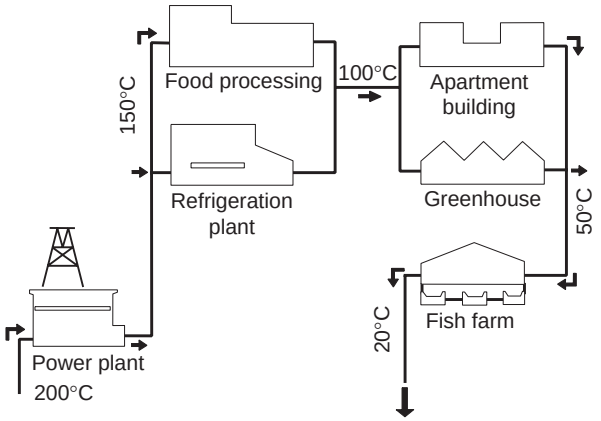


FIGURE 19 An example of cascading.

maintenance costs for the two systems are similar, except that the fossil fuel system may continue to pay for fuel at an ever-increasing rate, whereas the cost of the geothermal fuel is stable. The two systems, one with a high initial capital cost and the other with high annual costs, must be compared.

Geothermal resources fill many needs, including power generation, space heating, greenhouse heating, industrial processing, and bathing. Considered individually, however, some of the uses may not promise an attractive return on investment because of the high initial capital cost. Thus, one may have to consider using a geothermal fluid several times to maximize benefits. This multistage utilization, where increasingly lower water temperatures are used in successive steps, is called cascading or waste heat utilization. A simple form of cascading employs waste heat from a power plant for direct-use projects (Fig. 19).

Geothermal cascading has been proposed and successfully attempted on a limited scale throughout the world. In Rotorua, New Zealand, after geothermal water and steam heat a home the owner often uses the waste heat for a backyard swimming pool and steam cooker. At the Otake geothermal power plant in Japan, approximately 165 tonnes per hour of hot water flows to downstream communities for space heating, greenhouses, baths, and cooking. In Sapporo, Hokkaido, Japan, the waste water from the pavement snow-melting system is retained at 65°C and reused for bathing.

Current district heating costs are 0.23–0.42 ¢/1000 kcal (0.27–0.49 ¢/kWh) in Turkey compared to 3.4 ¢/kWh for natural gas and 11.2 ¢/kWh for electricity-based heating. The Klamath Falls district heating system charges 1.6–2.0 ¢/kWh. This is 50–80% of the natural gas cost, depending on the efficiency of the gas conversion, and the comparable cost for electricity in the city is 5.5 ¢/kWh. Construction costs for heating in Turkey are $850–1250/kW and the cost per residence is approximately $2000, an investment that is amortized over 5–10 years. Stefansson reported that an average consumer heating cost in 1995 for four European countries was 2.4 ¢/kWh.

Other estimates (1990 data) of the capital cost for various direct-use projects in the United States are as follows:

Space heating (individual): $463/kW of installed capacity
District heating: $386/kW of installed capacity
Greenhouses: $120/kW of installed capacity
Aquaculture: $26/kW of installed capacity

International data indicate a rate of $270/kW of installed capacity for all projects reported, with a range of $40–1880/kW. In the United States, the annual operation and maintenance cost is estimated to be 5% of the installed cost.

4. FUTURE DEVELOPMENTS

There appears to be a major potential for the development of low to moderate enthalpy geothermal direct use throughout the world, which is not being exploited due to financial constraints and the low price of competing energy sources. Given the right environment, and as gas and oil supplies dwindle, the use of geothermal energy will provide a competitive, viable, and economic alternative source of renewable energy.

Future development will most likely occur under the following conditions:

1. Co-located resource and uses (within 10 km).
2. Sites with high heat and cooling load density (>36 MWt/km^2).
3. Food and grain dehydration (especially in tropical countries, where spoilage is common).
4. Greenhouses in colder climates.
5. Aquaculture to optimize growth, even in warm climates.
6. Ground-coupled and groundwater heat pump installation (for both heating and cooling).

SEE ALSO THE FOLLOWING ARTICLES

Cogeneration • District Heating and Cooling • Geothermal Power Generation • Ground-Source Heat Pumps • Ocean Thermal Energy • Solar Thermal Power Generation • Thermal Energy Storage

Further Reading

Boyd, T. L. (1999). The Oregon Institute of Technology Geothermal Heating System—Then and now. *Geo-Heat Center Q. Bull.* **20**(1), 10–13.

Geo-Heat Center (1997.). Geothermal direct-use equipment. *Geo-Heat Center Q. Bull.* **19**(1).

Geo-Heat Center (1999.). Downhole heat exchangers. *Geo-Heat Center Q. Bull.* **20**(3).

Geo-Heat Center Web site: http://geoheat.oit.edu.

Gudmundsson, J. S., and Lund, J. W. (1985). Direct uses of Earth heat. *Energy Res.* **9**, 345–375.

Kavanaugh, S., and Rafferty, K. (1997). "Ground-Source Design of Geothermal Systems for Commercial and Institutional Buildings." ASHRAE, Atlanta, GA.

Lund, J. W. (1996a). Balneological use of thermal and mineral waters in the USA. *Geothermics* **25**(1), 103–148.

Lund, J. W. (1996b). Lectures on direct utilization of geothermal energy, United Nations University Geothermal Training Programme, Report 1. Orkustofnun, Reykjavik, Iceland.

Lund, J. W., and Boyd, T. L. (2000). Geothermal direct-use in the United States, update: 1995–1999. *In* "Proceedings of the World Geothermal Congress 2000, Japan." International Geothermal Association, Pisa, Italy.

Lund, J. W., and Freeston, D. H. (2001). Worldwide direct uses of geothermal energy 2000. *Geothermics* **30**(1), 29–68.

Lund, J. W., Lienau, P. J., and Lunis, B. C. (eds.). (1998). "Geothermal Direct-Use Engineering and Design Guidebook." Geo-Heat Center, Klamath Falls, OR.

Muffler, L. P. J. (ed.). (1979). "Assessment of geothermal resources of the United States—1978, USGS Circular No. 790." U.S. Geological Survey, Arlington, VA.

Rafferty, K. (2001). "An Information Survival Kit for the Prospective Geothermal Heat Pump Owner." Geo-Heat Center, Klamath Falls, OR.

Ragnarsson, A. (2000). Iceland country update. *In* "Proceedings of the World Geothermal Congress 2000, Japan," pp. 363–376. International Geothermal Association, Pisa, Italy.

Geothermal Power Generation

KENG CHOON LEE
Geopower NZ Limited
Mount Roskill, Auckland, New Zealand

1. **Geothermal Resources**
2. **Geothermal Power Cycles**
3. **Geothermal Drilling**
4. **Geothermal Steam Production Fields**
5. **Geothermal Power Stations**
6. **Environmental impacts and other technical problems**
7. **Commercial development of geothermal power projects**

Glossary

Annubar A device (trade name) that measures steam flow rate in a pipe.
bleed A small flow through a side valve below the master valve to keep a well hot.
casing A steel pipe that forms a well.
casing head flange The first flange that is attached to a well casing above ground.
directional (deviated) drilling Drilling a well away from the vertical line.
drain pot A short pipe at the bottom of a steam pipe to collect condensate flowing in it.
drawdown Downhole pressure loss causing a well unable to sustain constant discharge.
exergy The maximum thermodynamic work that can be converted from heat.
feed zone The location where geothermal fluids enter a well.
Kuster gauge A device (trade name) that measures pressure and/or temperature in a well.
master valve The first shut-off valve above the casing head flange of a well.
nozzle A short pipe attached to a pressure vessel or pipe.
reinjection The process of putting spent or unused geothermal fluids back into the ground.
sinker bar A long metal bar that provides counterweight when an instrument is lowered into a flowing well.
spinner A device that measures fluid flow direction and/or velocity in a well.
stimulation The process of removing the cold column of water sitting in a well so that the well can self-discharge.
superficial steam velocity Steam velocity assuming that the steam phase of a two-phase fluid occupies the entire cross section of the pipe.
tracer A chemical injected into a well to detect the water flow path.
turbine nozzle A wheel of fixed blades to direct steam flow to a wheel of rotating turbine blades.
wellhead The area around the casing head flange.
wellhead pressure (WHP) Fluid pressure in a well just below the master valve.

Energy can be converted from one form to another. Geothermal power generation is the production of electrical energy from the heat (thermal) energy of the earth (geo). Electricity is a much more convenient, clean, and useful form of energy than is heat. Geothermal energy has been used directly for its heat and therapeutic effect for thousands of years by the Romans, Chinese, and Turks. Its indirect use was realized in 1904 when Prince Piero Ginori Conti of Italy successfully produced electricity in an experimental "indirect cycle" using pure steam produced from a heat exchanger. This led to the first 250-kWe geothermal power station in 1913 at Larderello in Tuscany, Italy. Nowadays, geothermal energy is used directly for many purposes, but it is the generation of electricity that provides humankind with the most benefits and challenges. Whereas Larderello is the first dry steamfield that was developed for power, Wairakei in New Zealand is the first geothermal power station that was developed from a wet field. Energy conversion is subjected to the laws of thermodynamics. Geothermal power generation is similar to the conventional fossil fuel-fired power generation using steam turbines. The main difference is the lack of a manmade boiler in a geothermal power station; hence, steam supply requires more careful scientific assessment. However, geothermal power generation has other advantages. It has environmental benefits due to its low carbon dioxide emissions; it is "renewable" if managed in a sustained manner; it has cultural, social, and economic benefits in terms of tourism and recreational values; and because it is indigenous, it lessens the dependence on

 © 2004 Elsevier Inc. All rights reserved.

imported oil. Developing countries such as The Philippines, El Salvador, and Kenya have much of their electricity generated by geothermal energy, saving these countries a great deal in terms of oil imports.

1. GEOTHERMAL RESOURCES

The earth is an enormous sphere with more than a 6000-km radius. A solid brittle outer shell approximately 100 km thick, consisting of the continental crust (5–50 km thick) and the outermost part of the mantle, floats on a dense liquid metal core. Between the solid shell and the liquid core is the plastic layer of the mantle. The innermost core is believed to be solid metal due to the enormous pressure there. The continental crust is not one piece; rather, it is made up of many pieces known as tectonic plates. The earth has a heat source that is mainly from within the core due to radioactive materials. Heat is also generated by friction in the continental crust and the outermost mantle as the various tectonic plates move relative to each other. The consequences of this are the often catastrophic earthquakes and volcanic eruptions.

The temperature on the earth's surface ranges from about -50°C at the poles to about 50°C in the deserts. The earth's innermost core is believed to be approximately 5000°C ($\pm 1000^{\circ}$C). However, the earth's temperature gradient is not linear from the surface to the center of the earth but rather increases very steeply with depth from the earth's surface to approximately 1000°C at 100 km. To put things in perspective, we are looking at roughly 10 km of the continental crust for geothermal resources.

Geothermal resources can be defined as all of the heat energy stored in the earth's crust that is accessible locally. The word "locally" is important because a 20°C resource is usable at a location with -20°C ambient temperature but not 20°C (unless a heat pump is used). This is because heat transfer requires a temperature difference between a heat source and a heat sink. Geothermal resources from which energy could be extracted economically in the near future (<100 years) are parts of the useful accessible resource base. Geothermal resources that can be extracted economically today are defined as geothermal reserves. These identified reserves are further divided into possible, probable, and proven reserves depending on how much scientific work (geological, geophysical, geochemical, and drilling) has been carried out on them.

Whether a geothermal resource is technically suitable for geothermal power generation depends very much on the type and quality of the geothermal system and on the available technology. Currently, the geothermal system that is suitable for power generation is the hydrothermal type. This consists of a reservoir that contains hot water under high pressure. When this water is accessed by a drillhole, it ascends and some flashes to steam as the pressure releases.

The thermal water is mainly meteoric, so geothermal energy can be considered "sustainable" if the quantity of water withdrawn does not exceed recharge and if the heat added to the replenished water is at least equal to that extracted. However, geothermal energy is technically not strictly "renewable" because it is heat in the rock that heats the water, and this is not renewable. Most geothermal systems have lifetimes of less than 1 million years. Also, in general, heat that is extracted with the water from wells is much higher than that replenished from the reservoir rock because the heat capacity of water is much greater than that of the rock.

Hydrothermal systems are commonly divided into vapor- or steam-dominated and liquid- or water-dominated systems. In a steam-dominated system, the fluid in the reservoir is such that steam is the continuous phase that controls the pressure. When this fluid reaches the surface, the steam produced is normally superheated. In a water-dominated system, water is the continuous pressure-controlling phase. When this fluid reaches the surface, less than 50% by weight is steam. Although much of the world geothermal power generation comes from vapor-dominated systems, they are rare. The most famous are Larderello in Italy and The Geysers in California. Most geothermal systems are water dominated and are one-tenth the size of The Geysers in electrical power capacity. Examples of famous water-dominated systems include Wairakei in New Zealand and Cerro Prieto in Mexico.

Geothermal resources are commonly classified according to the average fluid temperatures in their reservoirs. Unfortunately, the demarcation lines are not agreed on (Table I). Exergy is also used to classify geothermal resources according to their specific exergy index,

$$SE \times I = (h - 273s)/1192,$$

where h is the specific enthalpy in J/g and s is the specific entropy in J/gK. Geothermal resources with $SE \times I > 0.5$ are high exergy or quality, those with $SE \times I < 0.05$ are low quality, and those in between are medium resources (Fig. 1).

TABLE I

Classification of Geothermal Resources by Temperature (°C)

	(a)	(b)	(c)	(d)
Low enthalpy	<90	<125	<100	<150
Intermediate	90–150	125–225	100–200	
High enthalpy	>150	>225	>200	>150

Source. Dickson and Fanelli (1990).

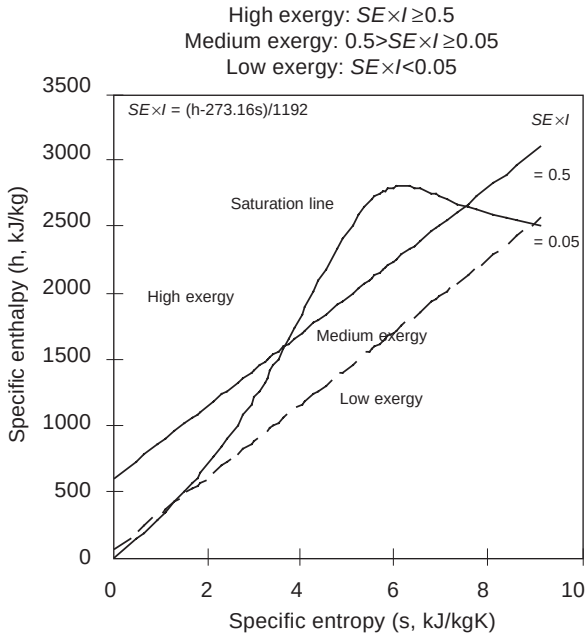


FIGURE 1 Classification of geothermal resources by exergy. Reproduced from Lee (2001).

The average earth's temperature gradient for the top 10 km from the surface is approximately 30°C/km. Places with gradient greater than 40°C/km are designated as thermal areas. Thermal fields are thermal areas with subsurface permeability that allows a fluid to convey deep-seated heat to the surface. Some thermal areas have 90°C/km gradient. The gradient can be higher in active volcanic areas.

Indicators of geothermal resources are surface thermal manifestations. However, a lack of surface manifestations does not imply a lack of geothermal resources (e.g., hot dry rock [HDR] system, geopressured system). Surface manifestations can be divided into active and inactive manifestations. Active manifestations are associated with heat and fluid discharges, whereas inactive ones are cold-altered ground and sinter deposits. Examples of active manifestations include hot springs, geysers, fumaroles, solfataras, mud pools, hot pools, warm ground, steaming ground, hydrothermal eruption craters, and volcanoes. Altered ground has rocks whose minerals have changed chemically due to interactions with thermal fluids. Sinter deposits are chemicals precipitated out of the thermal fluids and persist when the thermal fluids stop discharging.

An ideal geothermal system consists of a heat source and a reservoir of geothermal fluid in a confined space accessible by drilling. Heat is transferred from the heat source to the reservoir, which in turn transfers heat to the earth's surface. The heat source is normally a magmatic intrusion at shallow depths (5–10 km). The reservoir is a volume of hot rock that is permeable so that water can flow in and out of it. Because of the high pressure and temperature in the reservoir, the water tends to absorb gases such as carbon dioxide and hydrogen sulfide and to dissolve solid chemicals such as silica, calcite, and salts. When the geothermal fluid is withdrawn from the reservoir, pressure in the reservoir is likely to decrease, and this will cause more cold water to flow into the reservoir to be heated up again. This is the ideal renewable or sustainable geothermal system.

An HDR geothermal system contains a reservoir of hot rock that is not permeable. Research has been carried out to create permeability in the rock by hydrofracturing with water at very high pressure after drilling into the reservoir. Temperatures of 200°C have been measured in HDR at a depth of 3 km, and HDR systems of up to 10 km deep are being researched.

A geopressured geothermal system consists of a reservoir that is totally confined and contains water at very high pressure. There is no recharge of the reservoir water, which is usually saline. The presence of a geopressured system is most likely in a deep sedimentary basin. Pressures of 400 to 1000 bar abs and temperatures of 100 to 200°C at a depth of approximately 3 km are said to be typical.

A geothermal resource is normally "discovered" by geoscientific exploration. The exploration program normally follows the sequence of reconnaissance, prefeasibility, and feasibility studies. Geologists identify any geothermal surface manifestations and make geological and hydrological studies. These studies are important for all subsequent work, including siting of exploration and production wells. Geochemists study the chemistry of geothermal fluids on the surface to determine the type of geothermal system and estimate the reservoir fluid temperature. Geophysicists measure physical parameters such as heat flow, temperature, electrical conductivity, gravity, and

magnetism of the rocks. The information can be used to deduce the shape, size, depth, and capacity of the deep geothermal reservoir. Exploration wells are then drilled to confirm the power potential of the geothermal resource.

In a geothermal power project, the most important factor is the capacity and life of the resource. Unlike an oil reservoir that is exhausted when the oil is completely extracted, a geothermal reservoir can have its reservoir fluids replaced. This recharge of the geothermal reservoir fluids is an important process that controls the behavior of the reservoir. The study of recharge behavior is by means of geothermal reservoir engineering. Geological and well test data are normally analyzed using reservoir modeling computer software to estimate the capacity and life of the resource.

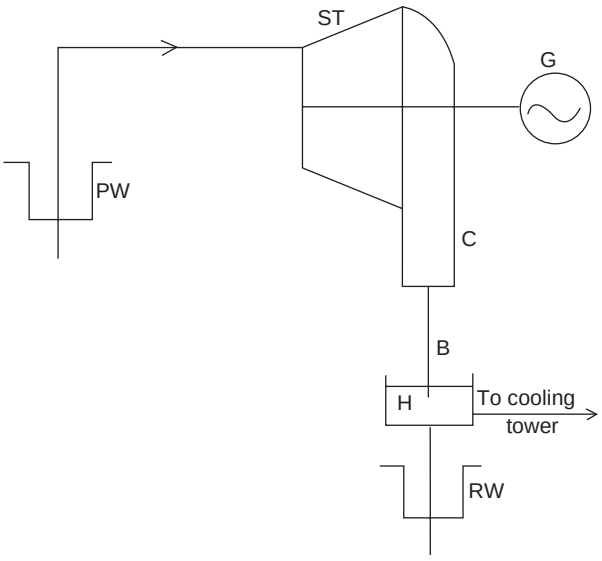


FIGURE 2 Power cycle for dry steam systems with condenser. B, barometric leg; C, condenser; G, generator; H, hot well; PW, production well; RW, reinjection well; ST, steam turbine.

2. GEOTHERMAL POWER CYCLES

When a geothermal resource has been deemed suitable for power generation, a decision is required to select a suitable power cycle. A power cycle consists of a series of thermodynamic processes that converts heat energy to electricity. The most common geothermal power cycle is the Rankine cycle, consisting of two isothermal and two adiabatic processes. More than one power cycle can be used for a geothermal resource. Indeed, other direct uses can be combined with power generation.

Geothermal power cycles can be classified into five major types: steam Rankine cycle, organic Rankine cycle (ORC), total flow cycle, Kalina cycle system (KCS), and trilateral flash cycle (TFC). ORC, KCS, and TFC are binary cycles that use geothermal fluids as the primary heat source to vaporize a secondary fluid for power generation. Currently, nearly all geothermal power plants use steam or ORC or a combination of the two. There are several total flow machines, but they are not widely used. The first pilot Kalina geothermal power plant was commissioned in Iceland in 2000. TFC is still at the experimental stage.

The following subsections examine the various power cycles based on the types of geothermal resources.

2.1 Dry Steam (Vapor-Dominated) Systems

Dry steam systems are rare, but they have high power capacity and are currently producing approximately 50% of total world geothermal electrical power.

Conceptually, the power cycle for dry steam geothermal systems is the simplest: saturated or slightly superheated dry geothermal steam from wells is piped directly to condensing steam turbines (Fig. 2). The condensate is reinjected for environmental reasons. Other problems associated with solids and gases carried in the steam have to be addressed. Solids are removed using a trapping device before they reach the turbines, and gases are removed from the condensers by a gas extraction system. The gases are mainly carbon dioxide ($>90\%$) and some hydrogen sulfide. Depending on the environmental protection law, the hydrogen sulfide might need to be treated by a chemical process.

2.2 Two-Phase (Liquid-Dominated) Systems

Two-phase systems are liquid-dominated systems. In two-phase geothermal systems, the geothermal fluids contain both water and steam at roughly an equal weight ratio (i.e., dryness, $x \sim 0.5$) at the wellhead. The water phase needs to be removed by a separator, and the steam is piped to the steam turbines (Fig. 3). The separated water can be reinjected directly or, depending on the fluid pressure, can be flashed to produce a lower pressure steam for power generation before being reinjected (Fig. 4).

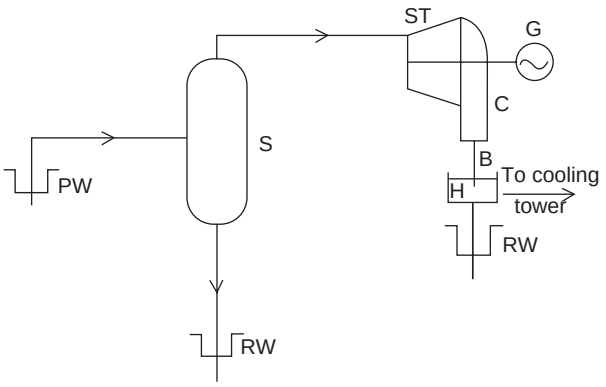


FIGURE 3 Power cycle for single-pressure hot water or two-phase geothermal systems. B, barometric leg; C, condenser; G, generator; H, hot well; PW, production well; RW, reinjection well; S, separator; ST, steam turbine.

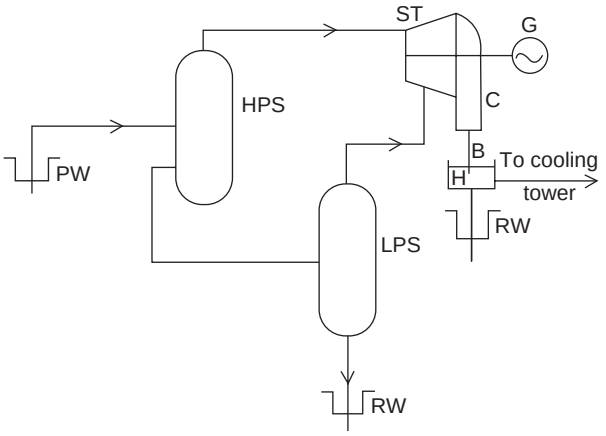


FIGURE 4 Power cycle for dual-pressure hot water or two-phase geothermal systems. B, barometric leg; C, condenser; G, generator; H, hot well; HPS, high-pressure separator; LPS, low-pressure separator; PW, production well; RW, reinjection well; ST, steam turbine.

2.3 Hot Water (Liquid-Dominated) Systems

Hot water systems are the most common geothermal systems suitable for power generation. In a hot water system, both the liquid and steam phases appear at the wellhead because the water flashes as it ascends the wells. The dryness fraction at the wellheads is normally less than 0.3. Depending on the wellhead pressures, the power cycle selected can be a single-pressure steam system (Fig. 3) or a dual-pressure one (Fig. 4). It is common to flash the water phase to produce more steam in a low-pressure flasher or separator. In practice, there is no real distinction between a flasher and a separator. Both are pressure

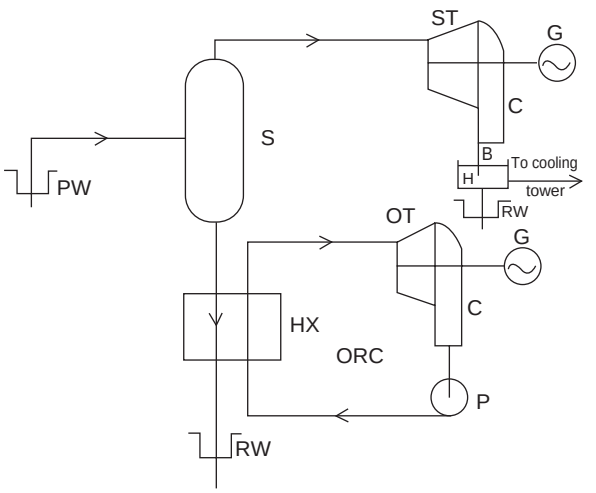


FIGURE 5 Power cycle for hot water or two-phase geothermal systems with binary ORC. B, barometric leg; C, condenser; G, generator; H, hot well; HX, heat exchanger; OT, organic vapor turbine; P, pump; PW, production well; RW, reinjection well; S, separator; ST, steam turbine.

vessels, and both have the function of separating the liquid from the steam phases. The flashing is normally done by a throttling orifice or valve installed upstream of the flash vessel. However, to maintain the large volume of low-pressure flashed steam, a flasher is normally larger than a high-pressure separator. Wairakei originally had a three-pressure steam system, but it was down-rated to a two-pressure system in 1983 because the steamfield drawdown could not sustain the high-pressure steam.

Nowadays, it is common to use the separated water for an ORC (Fig. 5). Exhaust steam from a back pressure turbine can also be used for a separate ORC (Fig. 6). For a hot water system that has a low dryness fraction, high gas content, or corrosive fluids, the steam Rankine cycle is unsuitable. A total flow or binary cycle (ORC, KCS, or TFC) may be used.

In a total flow cycle, the entire geothermal fluids from the wells are fed to a total flow machine that is robust enough to accept two-phase fluids (Fig. 7). The materials of construction can be selected to withstand the corrosive fluids. The helical screw expander is the first total flow machine to be developed and is now also used to generate electricity from waste heat in other industries. Other total flow machines are the biphase turbine and the Robertson engine.

In an ORC, the geothermal fluids are used to vaporize a low-boiling point (30–40°C) fluid, such as isopentane or isobutane, to very high pressure (15–20 bars), and this fluid is then used to provide the motive force for a vapor turbine. The advantage

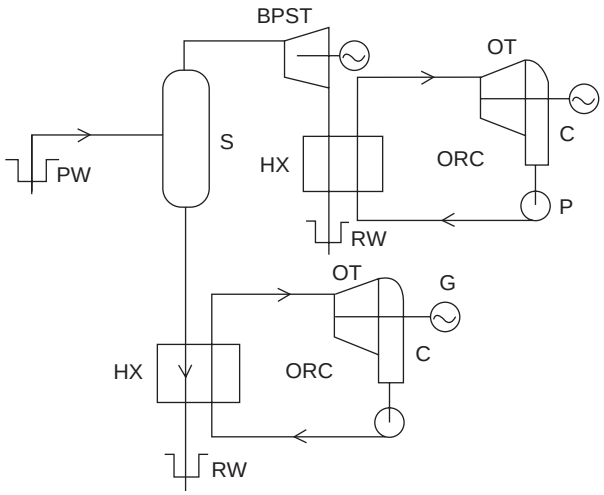


FIGURE 6 Power cycle for hot water systems with back pressure steam turbine (BPST) and binary ORC. See Fig. 5 caption for abbreviations.

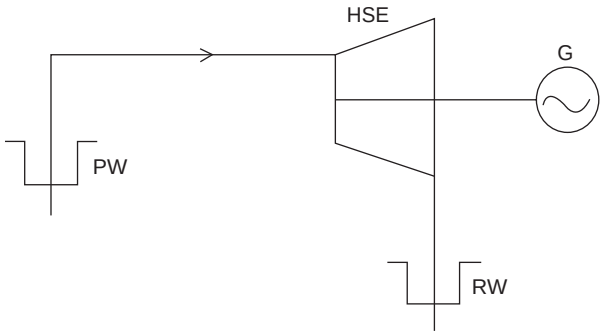


FIGURE 7 Total flow power cycle for hot water geothermal systems. G, generator; HSE, helical screw expander; PW, production well; RW, reinjection well.

of the binary ORC is that geothermal fluids with a temperature as low as 80°C can be used to generate electricity (Fig. 8).

A Kalina power cycle makes use of a mixture of two fluids in the secondary loop of the power cycle. It is considered a family of power cycle systems because not only can the ratio of the mixture of fluids vary with the temperature of the heat source, but different fluids also can be used for the mixture. An ammonia/water mixture appears to be the most suitable for geothermal applications. A pilot Kalina cycle geothermal power plant was commissioned in Iceland in 2000 and tested successfully using a geothermal fluid temperature of 121°C.

TFC, developed in 1993, is the latest binary power cycle. The concept of TFC is to improve the heat transfer process between the geothermal fluid and the secondary working fluid by keeping the working fluid

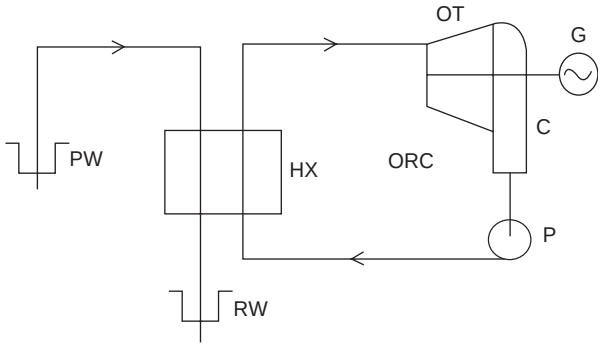


FIGURE 8 Binary ORC power cycle for hot water geothermal systems. See Fig. 5 caption for abbreviations.

as pressurized liquid, which is then used to power a helical screw expander. The best working fluid is still being researched, but low hydrocarbons appear to be most suitable.

2.4 Hot Dry Rock

HDR geothermal technology is still in the research stage. The concept is to drill two wells into hot rocks at a depth of 5 to 10 km, where the temperature gradient is high (70–90°C/km). Cold water is pumped into one well and pressurized to cause fractures in the hot rocks so that pressurized hot water flows up the other well. This is equivalent to an artificial liquid-dominated geothermal system. Power cycles suitable for natural hot water systems would be suitable for the HDR system.

2.5 Geopressured Geothermal System

A geopressured system of 1000 bar abs and 100°C is equivalent to an average hot water system in its ability to produce electricity. Because this is still at an early research stage, no power cycles have been proposed for a geopressured system, although the total flow power cycle using a helical screw expander appears to be most suitable.

3. GEOTHERMAL DRILLING

Geothermal drilling technology is adapted from the oil industry. Most of the geothermal wells are drilled with a rotary-type drilling rig. Because of the hard, abrasive, and high-temperature volcanic rocks of geothermal systems, the drilling bits require hard abrasive resistant teeth such as diamond or tungsten carbide. Hence, geothermal drilling is an expensive

activity in geothermal development. Most geothermal wells are drilled 1 to 2 km deep to intersect faults and fractures so as to increase the success rate of production from permeable zones. This is enhanced by directional or deviated wells as opposed to vertical wells of earlier development. Directional drilling also allows drilling of more than one well on a single well pad, especially in difficult terrains. A geothermal well typically costs U.S. $1 million and increases exponentially with depth. Geothermal drilling may be divided into three types: exploration, production, and reinjection.

Exploration drilling is carried out to assist earth scientists investigating a geothermal resource. Geologists will study the drill cuttings and drill cores of exploration wells that are normally drilled vertically and in small diameters (100–150 mm or 4–6 inches). It is not uncommon to make continuous drill cores from exploration wells so that the subsurface geology of the geothermal area can be confirmed. Exploration wells may be productive, and geothermal fluids discharged from the wells will confirm the chemistry of the fluids from earlier geochemistry studies.

Production wells are drilled after data from exploration wells have proven that a geothermal resource is economically viable in terms of its expected generating capacity and life. Geothermal fields of earlier development, such as Wairakei and Cerro Prieto, have vertical wells drilled to approximately 1 km deep. The fields have flat terrains, and wells are spaced as close as 20 m from each other. Nowadays, it is common to have several production wells drilled from a single well pad using directional drilling. Modern production wells are typically 2 km deep and may be deviated 1 km away from the well pad. Typical production wells have 250-mm ($9^5/_8$-inch) diameter production casing. Some large geothermal fields in The Philippines have large (500-mm or 20-inch) diameter production casing. (Note: As drilling is adapted from the oil industry dominated by the United States, the unit used is imperial.)

For environmental reasons, modern practices reinject most, if not all, geothermal fluids produced from the production wells back into the ground after the energy has been used. This is achieved by piping the spent geothermal fluids to reinjection wells. Reinjection wells are drilled similarly to the production wells. Good reinjection wells require as much permeability as do production wells to swallow the reinjected fluids. Hence, a good production well makes a good reinjection well. The reinjection may require pumping to achieve the required injection rate. However, reinjection of cooler geothermal fluids back into the ground may cool the hot fluids from production wells. Hence, reinjection wells are normally located either downstream of the flow direction of the geothermal fluids in the reservoir or at the edges of the production field so that reinjected fluids flow away from production zones.

A well is drilled by a rotating drilling bit (commonly a tricone bit with a tungsten carbide insert), with an additional load applied to it from the weight of a thick wall tube called the drill collar. The rotating motion comes from a rotating table with a square hole that transmits torque to a square tube known as the kelly. The kelly is connected to the drill collar by drill pipes to form the drill string. Drill pipes have a thinner wall than does the drill collar. The drill string is smaller in diameter than the hole being drilled, allowing a drilling fluid to be pumped down inside the drill string, clearing the cuttings around the drill bit, and carrying the cuttings up via the annulus between the hole and the drill string.

The drilling of a typical production well (Fig. 9) commences by drilling a large diameter hole (24–26 inches) up to approximately 50 m. A 20- to 22-inch outer diameter pipe (casing) is then cemented in the hole. The cementation is achieved by pumping a cement slurry down inside the casing and flowing up the annulus. The cement inside the casing is then displaced by water to just above the bottom of the casing (shoe). When the cement has set, a drilling wellhead with blowout preventer (BOP) is erected, controlling accidental discharge of geothermal fluid during drilling. Drilling continues with a smaller diameter hole (17.5 inches) for a $13^3/_8$-inch casing from the top of the cement near the casing shoe to the next casing depth. This sequence is repeated until three or four casings are cemented to a depth assessed to be the top of the desired production zone in the reservoir. The final casing is called the production casing, usually $9^5/_8$ inches in a $12^1/_4$-inch hole. In general, the shallow outer casings are required for drilling operations, whereas the inner casings are for production purposes. These casings not only control the geothermal fluids flowing inside the production casing but also prevent cold fluids outside from entering the well.

When the final section of the hole (8.5 inches) is completed in the production zone, a 7-inch casing with holes (slotted liner) is usually used to line the well. Unlike other casings, the liner is not cemented into place. It can be sitting at the bottom of the well or hanging from the production casing shoe. The liner prevents the production formation at the hole from collapsing and allows the geothermal fluids to

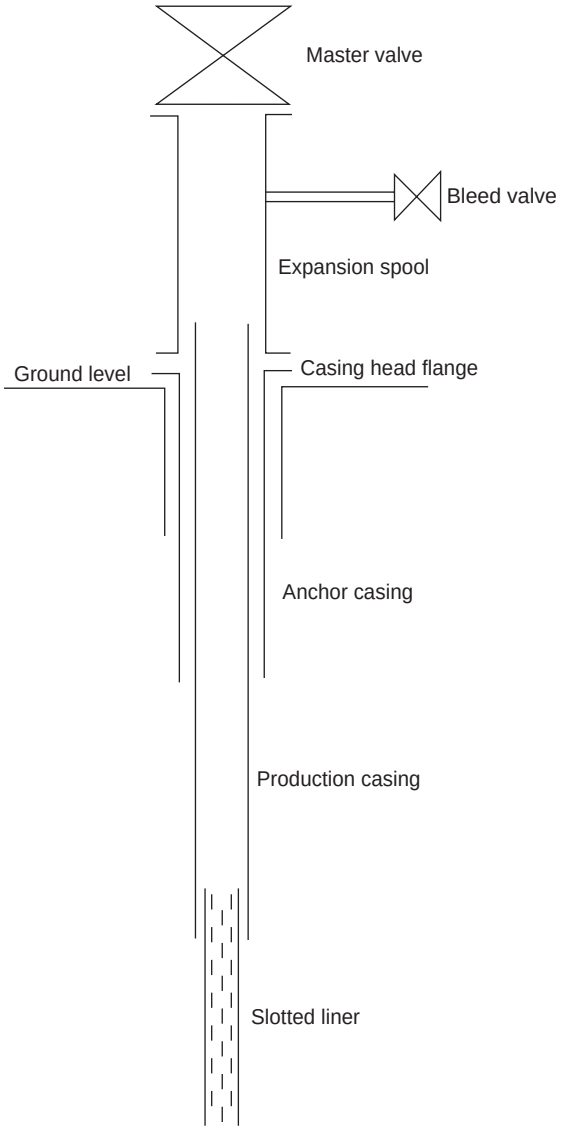


FIGURE 9 A typical well casing layout.

enter the well and flow up inside the production casing. A wellhead consisting of flanges, valves, and pipe spools is usually attached to the top of one of the intermediate cemented casings. The production casing is allowed to expand into a wellhead pipe spool during production of the hot fluids.

Because the safety of the drilling rig and the completed well can be at risk for an unstable well site, a complete geotechnical investigation might be necessary before site construction. Of particular concern is the possibility of a shallow blowout during drilling in areas with shallow permeable rock formation. To prevent this, consolidated grouting of the drilling site is carried out. This involves drilling and cementing slim holes to depths of 10 to 30 m around the drilling site. This will direct any accidental blowout discharge of geothermal fluids away from the drilling rig.

A reliable supply of water is essential for safe drilling operation. Water is required for the drilling fluid—usually mud, cementing, quenching, or controlling a well from discharging. Water is also used as the drilling fluid when total loss of circulation is encountered, especially in the production zone, to prevent sealing of the production zone by drilling mud. A typical production geothermal well requires a reliable water supply of 2 m^3/min.

An important factor in drilling a geothermal well is to prevent the well from discharging during drilling. This is achieved by having the column of drilling fluid create a pressure greater than the fluid pressure in the reservoir formation. To increase the density of the drilling fluid, bentonite is commonly added to the drilling fluid to form drilling mud. There are other advantages in drilling with mud. It can keep drill cuttings in suspension in the event of stoppage during the drilling operations such as connection of additional drill pipes. However, drilling with mud also has its drawback. When a production zone is hit, the drilling mud, having higher pressure than the formation, will flow into the production zone fractures, effectively sealing the production zone. Hence, alternative drilling fluids that will not seal the production zone should be used when drilling in the production zone. These can be aerated water, foam, or air. In such situations, the practice of balance drilling is best such that the drilling fluid pressure balances the formation pressure.

4. GEOTHERMAL STEAM PRODUCTION FIELDS

Geothermal wells need to undergo a test program before they are used. This is so that the baseline conditions of both the wells and the geothermal aquifers that they tap can be determined. This baseline data are critical because all future information is compared against them.

After drilling has been completed and before the initial discharge, the well downhole conditions are measured. The temperature and pressure are measured by using a clockwork Kuster gauge or electronic logging tools. Standard practice is to initially do an injection (or completion) test, that is, sometime at a series of flows, where the temperature, pressure, and possibly flows using a spinner tool are

measured at intervals down the open hole section of the well. The well is then shut in and allowed to fully heat up. This can take less than 1 day (e.g., for a steam well) to more than 6 months (e.g., for a deep well with poor permeability). The downhole conditions during heat-up are also measured at intervals. The well is shut in to minimize circulation so that conditions in the cased-off section can be measured and so that the gas column can depress the liquid in the well to the upper feed zone. The downhole measurements will reveal the production zones in a well because hot geothermal fluids circulating at the production zones will cause these areas to heat up faster. The pressure profile of a well can reveal the water level in the well. The water level can also be determined by lowering a float into the well.

Once the new well is fully heated, it needs to be discharged to rid it of drilling mud, fluid, and debris before the well can be put into service. The initial discharge is normally a vertical one due to the erosive nature of the materials. The well is then tested to determine its output characteristics, that is, mass flow rate and enthalpy as a function of wellhead pressure (Fig. 10). Chemical sampling is also done during this output test. These data are very important because they comprise the baseline information.

The well output can be measured by several methods. For a small well discharge, a simple calorimeter can be used. The well output is discharged into a large vessel containing a mass of cold water. A heat balance will determine the output.

A method unique to the geothermal well test is the James lip pressure method. This method makes use of discharging two-phase fluid at choked flow conditions through an open pipe. By measuring the static "lip" pressure at the open end of the pipe, the flow rate can be determined from a correlating equation. The equation now is in several forms. The original equation in imperial units is

$$G = 11400\ P^{0.96}/h^{1.102},$$

where G = flow in pounds/(s feet^2) of the lip pressure discharge pipe cross-section area, P = lip pressure in psia, and h = fluid enthalpy in Btu/pound.

In metric units, the equation is

$$G = 0.184\ P^{0.96}/h^{1.102},$$

where G = tonnes/(s cm^2), P = bar absolute, and h = kJ/kg.

The lip pressure method requires knowledge of the fluid enthalpy. However, the equation is relatively insensitive to enthalpy, so it can be estimated if not known. From experience, it is known that the discharge shape of the fluid at the lip pressure pipe will give some indication of the enthalpy. An estimate can also be made if the downhole conditions are known. If the fluid is known to be saturated water at a certain pressure during drilling, the enthalpy is also known. If the discharge is made into a silencer and the water flow rate is measured by a weir plate, the total mass flow can be determined. A quick estimate of the heat discharge or power potential of a well can be made by the lip pressure method during the first well vertical discharge. This number is useful for planning of the surface equipment for further tests.

A well output is most accurately measured by the separator method. The well is discharged into a separator. The steam phase is measured in a steam pipe using an orifice plate or Annubar, and the water is measured by a weir plate in a weir box at the silencer. The water flow rate needs to be corrected for that flashed in the silencer. However, this method is expensive. Production well surface equipment is normally designed to allow regular output tests with minimum disruption.

There are other well tests, such as downhole well pressure and temperature measurements, that are normally done with the well on bleed. Once a well has been initially discharged, it should not be completely shut off because a column of cold water may form in the well, preventing the well from discharging again without stimulation.

The discharge of a well may interfere with the flow of a nearby well. Interference tests can be done by either measuring the downhole pressures or measuring the discharge from a well while a nearby well is discharging at different flow rates.

After enough wells have been tested to confirm the geothermal field's capacity and suitability for the production of electricity, steamfield design needs to be carried out. It can be carried out in parallel with

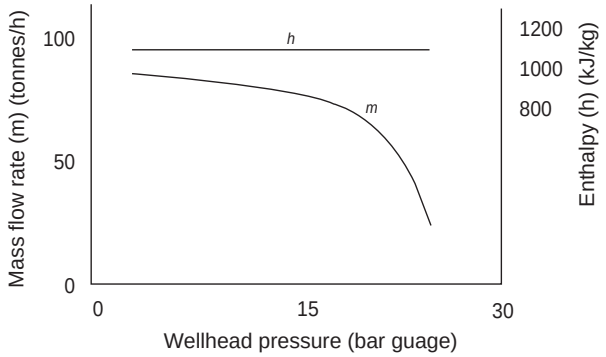


FIGURE 10 A typical geothermal well output characteristics.

further production well drilling. This is to determine the best method for collecting the geothermal fluids from the wells and processing them, especially when the geothermal fluids are two-phase ones. In general, a geothermal steamfield consists of the following major plants and equipment: the wells, the wellhead, the flasher/separator for a wet field, the silencer, fluid transmission pipelines, and a reinjection system.

4.1 The Wells

The use of a geothermal resource depends entirely on the output from the wells. A well's output can vary significantly from that of a typical well (Fig. 10). An example is the cycling wells that have fluctuating mass output with a cycle time as short as a few minutes. This type of well is not suitable for power production unless it can be balanced by stable output wells. Another type of problem well is one that continuously discharges solids with the well fluids. Sometimes this problem can be solved by throttling, but generally a dirty well is not used for production.

4.2 The Wellhead

A typical wellhead consists of several valves to provide service functions required (Fig. 11). The master valve is for completely shutting off the well. A small side valve is installed below the master valve to "bleed" the well when the main flow is shut off. It is not a good practice to completely shut in the well because the well may cool down and a column of cold water may sit in the well, preventing the well from discharging when the master valve is reopened. There may also be casing and cement damage from the thermal cycling. The well can flow again only when the column of water is removed by stimulation. Well stimulation is normally carried out by pumping air at the bottom of the well to airlift the flow. It can also be done by pumping steam or air into the well to depress the water column and then releasing the pressure quickly. Bleeding allows the well to stay hot.

It is difficult and costly to replace a master valve. A procedure must be in place to reduce damage during operation. A master valve is normally replaced by plugging the well below the master valve level, and this process normally requires a drilling rig. Some wells can be quenched by cold water injection through the bleed valves, allowing the valve to be changed. Because the master valve is so critical, it should be opened or shut only when there is no flow and should always be in either a completely shut or a fully open position. A sacrificial service valve is installed downstream of the master valve. The service valve can be used for throttling, but long-term throttling of a well is normally done using a throttling orifice.

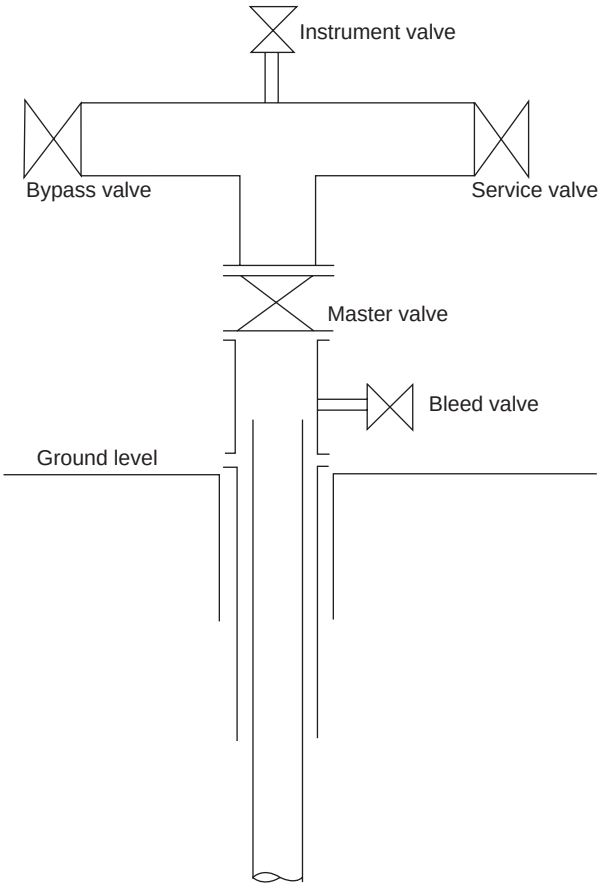


FIGURE 11 A typical wellhead and its valves' arrangement.

The service valve is commonly installed vertically above the master valve. It can be used to admit an instrument such as a Kuster gauge. The instrument is first lowered to sit above the shut master valve. The instrument is hung on a stainless steel wire that runs through a gland packing. The master valve is opened only after the gland packing is in place. The flanged spool with the gland system is called the recovery tube.

It is not common to make downhole measurements while a well is flowing. The instrument needs a long sinker bar to counter the buoyancy of the discharge. This means that a long recovery tube is needed. The flowing survey data are generally meaningful only if the well has a liquid feed zone of good permeability.

The bypass valve is required when the output of the well is not required for power generation for a short period, or for other reasons (e.g., tests, emergencies), and the well is not shut off due to possible interference. The flow is diverted through the bypass valve to a

silencer. For environmental reasons, discharge to waste via the silencer is not done unless necessary.

The arrangement of the wellhead valves normally results in the top service valve being several metros above the casing head flange. Earlier designs were to install the casing head flange in a deep cellar or at ground level, and a platform was built to access the service valve. Nowadays, a deep cellar is not preferred because hydrogen sulfide and carbon dioxide, being heavier than air, can accumulate in such a cellar, causing a lethal hazard. The preferred option is to install the casing head flange at ground level and have a tee piece just above the master valve and the service valve on the horizontal pipework. A small instrument valve or a blind flange is installed on the tee piece directly above the master valve.

4.3 The Flasher/Separator

Although sometimes a flasher is differentiated from a separator, the two are the same pressure vessel in practice. A flasher allows high-pressure geothermal water to flash at a lower pressure. In practice, the lower pressure is achieved by a throttling orifice or valve upstream of the flasher/separator. The vessel then separates the two phases by centrifugal and gravity effects.

The modern separators differ in many aspects from those first used at Wairakei during the 1950s. They differ in size, location, body arrangement, separated water vessel, two-phase inlet nozzle, water outlet nozzle, steam outlet nozzle, and efficiency. The original Wairakei separators were wellhead separators. Each wellhead has its own small separator handling the capacity of a single well. It can be as small as 750 mm diameter and 2 m high, handling about 20 tonnes/h of steam. A modern separator can be larger than 3 m in diameter and 15 m high, and it operates as a separation plant, accepting well output of several large wells. It can handle 600 tonnes/h of steam and is rated at more than 60 MWe.

The earliest separators at Wairakei were the Woods type, and some are still operating there today. A Woods separator has a U-bend-type separator operating in series with a top steam outlet cyclone separator that has a tangential two-phase inlet nozzle. The later design was the significantly cheaper Weber type with a bottom steam outlet (Fig. 12). A modern separator has a spiral inlet nozzle. This spiral inlet design increases the capacity of the separator by increasing the inlet superficial steam velocity from 20 to 40 m/s before the separation process breaks down. The water outlet nozzle is installed at the bottom of the body exiting from the side. This is to improve the efficiency of water discharge to the water vessel, minimize erosion if the water is laden with solids, and minimize the risk of debris entering or blocking the water outlet pipe. The Weber separator has an efficiency that produces steam dryness exceeding 99.9% at design flow rates.

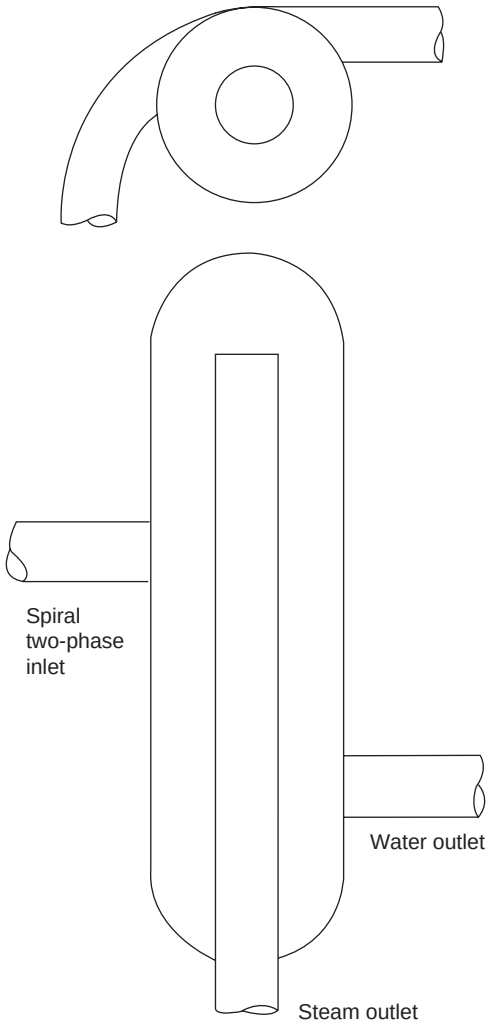


FIGURE 12 A common Weber-type geothermal separator.

A modern separator design is an integral design that does not have a separate water vessel. Some modern designs incorporate an integral horizontal water vessel to reduce the height. To reduce height and earthquake effect further, large horizontal separators can be used.

4.4 The Silencer

A geothermal silencer is like a separator operating at atmospheric pressure. It is normally connected directly to a well and a separation plant for bypass

or emergency discharge. The silencer reduces the high-frequency noise discharging from an open pipe to a low-frequency rumbling noise that is more bearable to human ears. The earlier Wairakei silencer design consisted of twin barrels, and the body was made of timber. The modern silencer design consists of a single barrel made of corrosion-resistant steel, sometimes internally lined with concrete (Fig. 13). Fiberglass barrels worked well for silencers not in constant use. Some new silencer designs have a cone fitted at the top of the barrel to increase the steam velocity so that it is dispersed farther away from the silencer area. The noise can be further attenuated with an annulus full of scoria rock in the silencer barrel.

For a dry steamfield, a silencer is commonly just a pit full of rocks. A bypass pipeline discharges the dry steam at the bottom of the pit.

4.5 Fluid Transmission Pipelines

Many kilometers of pipelines are required in a geothermal steamfield to transport geothermal fluids from the wells and ultimately to the power plant. The geothermal fluids can be a two-phase steam/water mixture, separated steam, or separated water. All of the pipelines together with the separators, valves, and other plants and equipment in a geothermal steamfield have been collectively known as the energy gathering system (EGS), steam gathering system (SGS), steam collection system (SCS), or fluid distribution and collection system (FDCS).

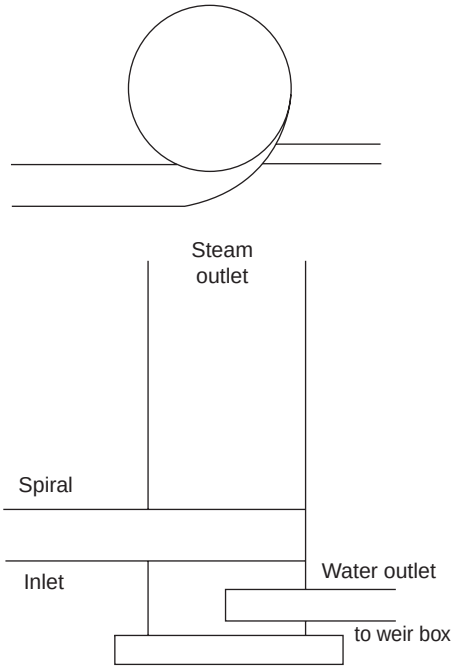


FIGURE 13 Modern single-barrel silencer.

When wet steamfields, such as Wairakei and Cerro Prieto, were first developed, geothermal two-phase flow was not well understood and two-phase pipelines were kept to a minimum. So, separators were installed at the wellheads, and separated water and steam were transported as single-phase fluids from the separators to the desired destinations. So, each wellhead had a complicated pipe network.

Nowadays, two-phase pipelines are used extensively to transport geothermal fluids from several wells to feed a separation plant. This design greatly simplifies the pipe network at each wellhead and the whole steamfield because it allows the use of fewer large diameter pipelines in the steamfield. Another more important advantage is that large diameter pipes cause less pressure loss of the fluid being transported, thereby conserving energy.

The location of a separation plant is important because it affects well outputs and the total pressure loss and, ultimately, the economics of the power project. Two other factors affect the location of the separation plant: steam quality and steam pressure control.

The efficiency of a separator can be 99.99%. For a large separation plant, there may be 500 kg/h of water carryover in the separated steam. This water carryover needs to be removed as much as possible before it reaches the steam turbines because the water carryover in droplets traveling at high velocity erodes the turbine blades or solids precipitated from the water deposit in the turbine nozzles, ultimately choking the turbine output. There are two ways in which to remove the water carryover. One is to install a demister or scrubber at the power station end. The other is to install drain pots along the steam pipelines. The drain pots do not work if the steam velocity is high and the water carryover travels as droplets with the steam. There is also a minimum number of drain pots spaced at a minimum distance apart for the drain pots to function properly. This means that there must be a minimum length of steam pipeline downstream of the separation plant to allow for the scrubbing of the steam phase.

The second factor is that a minimum pipe length, and hence volume, is also required for the pressure control system to function properly.

4.6 Reinjection System

A major problem with a liquid-dominated steamfield is the large quantity of separated water that requires

disposal. When Wairakei was developed roughly 50 years ago, environmental concerns were not an issue and the separated water system was designed to simply drain into the nearby Waikato River. Nowadays, the only acceptable method is reinjection (unless the field is situated near the sea). Either the water flows down the well by gravity or, more often, a pump is required to increase the injectivity of the well. A control system is generally required to stop flooding or steam carryover.

The location of a reinjection well is important. A reinjection well within a steamfield can replenish the geothermal fluid that has been withdrawn, but cooler reinjected fluid may break through to the production wells and cool the production fluids. Nowadays, the strategy is to reinject at the perimeter of the steamfield to minimize the risk of the cooling effect. A problem with this strategy is that the permeability on the margins is generally poor, so more wells are needed. The aquifer flow paths can also change under exploitation so that reinjection wells can have rapid returns to production areas where initial tracer tests showed no flow. In the steam-dominated field of The Geysers, waste sewage water is reinjected to replenish the drawdown of the field.

TABLE II

World Installed Geothermal Power Generating Capacity in 2000

Country	MWe
United States	2898.0
The Philippines	1909.0
Italy	785.0
Mexico	755.0
Indonesia	589.5
Japan	546.9
New Zealand	437.0
Iceland	170.0
El Salvador	161.0
Costa Rica	142.5
Nicaragua	70.0
Kenya	57.0
Guatemala	33.4
China	32.0
Russia	25.0
Turkey	20.4
Portugal (Azores)	16.0
Ethiopia	8.5
France (Guadeloupe)	4.2
Thailand	0.3
Australia	0.2
Total	**8661.0**

Source. Institute of Geosciences and Earth Resources, Pisa, Italy.

5. GEOTHERMAL POWER STATIONS

The total installed geothermal power generating capacity in the world is approximately 9000 MWe from 21 countries, with the United States leading at nearly 3000 MWe and The Philippines with nearly 2000 MWe (Table II). Other major countries are Italy, Mexico, Indonesia, Japan, and New Zealand, with between 400 and 800 MWe each.

Geothermal power stations have very high availability, load, and capacity factors ($>90\%$). They are most suitable as base load power stations. Liquid-dominated reservoirs are not suitable for peak load generation because two-phase flow and the separation process make their operations and control at changing flows difficult.

A geothermal power station that uses steam turbines to generate electricity is similar to a conventional fossil fuel-fired power station. The main differences are:

- Geothermal steam is usually saturated steam at less than 10 bar abs and 200°C (occasionally superheated up to $\sim$50 bar abs and 300°C), but in a fossil fuel-fired power station, the steam can be supercritical to as high as 300 bar abs and 600°C.
- Geothermal steam normally contains impurities of noncondensable gas (NCG) at up to a few percentage points by weight of steam (mainly carbon dioxide [$>90\%$] and a few percentage points of hydrogen sulfide) and dissolved solids (mainly silica or calcite) that precipitated from flashing of the liquid phase.
- A fossil fuel-fired power station has a boiler and the steam/water is in a closed loop, whereas a geothermal power station does not have a boiler and the steam/water is in an open loop.

Because of the low-pressure saturated steam, a geothermal steam turbine unit is limited to a maximum capacity of approximately 55-MWe condensing unit due to the maximum length of the last-stage turbine blades limited by the material metallurgy. Two of these turbines can be installed on a single shaft to a 110-MWe generator.

The NCG is allowed to pass through the steam turbines but must be removed from the condensers so

that the low subatmospheric pressure in the condensers can be maintained. A gas extraction system is used to remove the NCG from the condensers. A steam jet ejector is suitable for low gas contents of less than 1% weight of steam. A steam jet ejector has no moving parts, so it is easy to operate and has a low maintenance cost. However, the steam flow used to drive it is high (~five times the NCG flow rate). Hence, steam jet ejectors are suitable only for low NCG levels. The NCG removed from the condensers is normally exhausted to the atmosphere.

For economic reasons, a gas exhauster is required when the gas contents exceed 2%. This is normally a multistage centrifugal compressor with an intercooler. Recirculation valves between the stages are necessary for control. It is typical to drive the compressor by electric motor, but the compressor can be driven by the steam turbine.

A liquid ring pump can be used to replace the final-stage steam jet ejector to improve efficiency and economy. Liquid ring pump systems are used in fossil fuel-fired power stations to maintain condenser vacuum. They have a small capacity, so they are suitable only for low gas contents.

The condensers mentioned previously are direct-contact condensers. Shell and tube condensers can also be used to better control NCG emissions. However, shell and tube condensers are expensive and less efficient.

When a geothermal resource cannot produce a sufficient quantity of clean enough steam to drive steam turbines, other modes of power generations need to be used. If the steam contains too much NCG (>10% weight of steam), a condensing turbine unit becomes uneconomical. In such a case, a back pressure turbine that does not have an after-condenser can be used. The steam is exhausted directly to the atmosphere, but this mode of generation is unlikely to be acceptable nowadays for environmental reasons. Either total flow or binary mode of generation needs to be used.

Many geothermal ORC binary plants are in operation around the world. A heat exchanger (shell and tube type) is used to transfer heat from the geothermal fluids to a low-boiling point secondary fluid (isopentane or isobutane). The secondary fluid superheated vapor drives a vapor turbine that exhausts to an air-cooled shell and tube condenser. The secondary fluid is in a closed loop, and the spent geothermal fluids are reinjected.

Whereas an ORC binary plant usually uses an air-cooled condenser, a typical condensing steam turbine unit uses a direct-contact condenser. The resultant condensate/cooling water mixture is normally removed from the condenser via a barometric leg or a pump. This condensate mixture flows to a holding pond known as the hot well, and from there it is pumped to a cooling tower to be cooled for recirculation to the condensers. As the geothermal steam from the steamfield eventually becomes a part of the recirculating cooling water, the total volume of the recirculating water tends to increase, even though some is evaporated at the cooling tower. Hence, some of the condensate mixture is reinjected from the hot well. Because this condensate mixture is warm (~40°C) and oxygenated, it is quite corrosive to carbon steel. Stainless steel pumps, pipes, and well casing are required for condensate reinjection.

Two types of cooling towers are commonly used for a geothermal power station: natural draft and mechanical draft cooling towers. For economic reasons, a natural draft cooling tower is built for an approximately 100-MWe capacity geothermal power station. Natural draft cooling towers also function better in a cooler climate. For warm climates, mechanical draft cooling towers are more suitable. Because the fan of a mechanical draft cooling tower is limited by the length of the its blades due to the strength of the construction material, a mechanical draft cooling tower has a small capacity. So, mechanical draft cooling towers are normally built as rows of cells.

6. ENVIRONMENTAL IMPACTS AND OTHER TECHNICAL PROBLEMS

Geothermal power stations have their own unique developmental, operational, technical, and environmental problems. Many geothermal resources occur in or near national parks or remote uninhabitable locations because they are normally associated with volcanoes and have beautiful surface manifestations and so become public reserves and tourism hot spots. Most of these geothermal resources do not get developed for power generation. Examples include Yellowstone Park in the United States and Orakei Korako Geyserland in New Zealand. However, geothermal resources that exhibit little surface manifestation in national parks do get developed. Examples include Hatchobaru and Otake in Japan.

Volcanoes, and hence geothermal resources, are also associated with mythical power and beliefs. The development of these geothermal resources requires much consultation and negotiation with

the natives inhabited in the areas. The final development of these resources normally includes some forms of partnership and reparation for the natives. Examples include Mount Apo in The Philippines and Ohaaki in New Zealand.

The development of a geothermal power project will have environmental impacts on the physical surroundings as well as on the socioeconomics and health of the surrounding communities. The physical environment includes the land, water, air, flora, and fauna. Most countries have strict environmental legislation such as the National Environmental Policy Act (NEPA) in the United States and the Resource Management Act (RMA) in New Zealand. The most important requirement of the environmental laws is the environmental impact report (EIR). This report is also called the environmental assessment (EA), environmental impact assessment (EIA), environmental impact statement (EIS), and assessment of environmental effects (AEE). The acts mandate processes that must be followed for the assessment and suggestions for any preventive and mitigating measures. Any disagreement may be redressed in an environment court. The EIR normally follows a checklist of environmental impacts. The positive result of an EIR is an environmental permit or consent, approved and issued based on the technical, environmental, and social acceptability of the proposed project.

Every geothermal resource is unique; hence, the actual physical environmental impacts of a geothermal power project development and its subsequent operation vary from one project to another. Some examples of the common physical environmental impacts are described next.

The land may be unstable, so any development may trigger landslide or soil erosion that could be made worse in areas of heavy rainfall. This is particularly true for many geothermal systems in difficult terrain associated with volcanic environments. Examples of geothermal systems with difficult terrain include The Geysers in the United States and Leyte in The Philippines. The actual land area required for a geothermal power plant is relatively small compared with that needed for other types of power plants when the areas of the resources are taken into account. A 100-MWe geothermal power plant requires approximately 1 km^2 of land; a coal-fired power plant requires approximately 4 km^2, including the coal mining area; a hydropower station requires roughly 10 to 30 km^2; and a wind power station requires approximately 40 to 70 km^2. The land uses that are unique to a geothermal power plant are the pipeline corridors and geothermal well pads. Some geothermal areas are used for farming (e.g., Mokai and Rotokawa in New Zealand), but geothermal power plants and steamfields can coexist very well with other land uses such as farming and agriculture. Indeed, some development of geothermal systems provides aquaculture (e.g., prawn farm in Wairakei) and recreation (e.g., swimming pool at Svartsengi, Iceland).

Land subsidence is a common occurrence in geothermal development. Wairakei, after more than 40 years of operation, has a general subsidence of 10 to 30 mm/year over the whole steamfield. The maximum subsidence is approximately 250 mm/year, creating a small lake 18 m deep. Fortunately, the area of maximum subsidence is confined to a small area at the perimeter of the steamfield, causing minimal effects. Subsidence has also occurred at the Ohaaki after more than 20 years. It is difficult to predict land subsidence before the development of a geothermal steamfield. So, a monitoring program of the ground levels is required from the start of the development to ensure the safe operation of the steamfield as well as for environmental reasons. In general, the withdrawal of liquid from wet fields appears to cause greater subsidence than does withdrawal of steam only from dry fields.

The fluid production from a steamfield may also cause the surface manifestations to change, causing hydrothermal eruptions and creating new mud pools, fumaroles, and ground steam discharges. The changes to the "Craters of the Moon" near Wairakei are believed to be the result of fluid production at Wairakei. The withdrawal of geothermal fluids may also cause local micro earthquakes.

Gaseous discharge from the operation of a geothermal power plant is relatively small compared with that from the operation of a fossil fuel-fired power station. In general, the combustion of fossil fuels to generate electricity produces 3 to 20 times more carbon dioxide (kg/kWh) than that produced by a geothermal power plant. Although geothermal steam normally contains a small quantity of hydrogen sulfide that is not present in fossil fuels, the combustion of fossil fuels produces much more sulfur dioxide and nitrogen oxides that are not present in geothermal steam.

Water is required for the well drilling and construction phases. A circulating water system is required for the power plant during operation. Wet geothermal fields produce much separated geothermal water that requires disposal. The takings and discharges of water during the development,

construction, and operation stages all must be assessed for environmental impacts to the sources and sinks of water. All of these require water rights that are reviewed and monitored at regular periods.

Because the amount of land required for a geothermal power plant is relatively small, the effects on flora and fauna are also relatively small. However, being unique in nature, geothermal areas may also contain unique forms of flora and fauna. These flora (mostly fungi) and fauna (small organisms) may survive only in the warm areas of the geothermal activities. Their ecosystems might need to be protected and maintained.

Noise pollution of a geothermal power project will occur during the well drilling, construction, and operation phases. Because many geothermal areas are in remote areas where the background noise is very low ($\sim$40 dBA), any additional noise from geothermal activities will be heard by the local residents immediately. Impacts of noise on animals are reported to affect their behavior, physical size, and reproductive activities. Well drilling produces noise of 80 to 110 dBA, and a well discharging without a silencer produces 120 to 125 dBA of noise, which is at the threshold of pain for the human ear. The noise of a well discharging into a silencer is reduced to approximately 90 dBA.

Visual effects from a geothermal development include the tall drilling rig during drilling phases and the tall ($\sim$100-m) natural draft cooling tower if adopted. Wet mechanical draft cooling towers are much lower, but their plumes are much more visible. The plumes can be minimized with a wet and dry cooling tower. Visual impact of geothermal pipelines has been minimized by painting them the color of the background.

Socioeconomic impacts are of most interest to the federal and local governments and the surrounding communities. The costs, benefits, and impacts on the local and wider communities will ultimately determine the success or failure of a geothermal project. Since the 1992 Earth Summit of the UN Conference on Environment and Development (UNCED) in Rio de Janeiro, Brazil, the emphasis has been on sustainable development, defined by the World Commission on Environment as "development that meets the need of present generation without compromising the need of future generations to meet their needs."

A fundamental purpose of a socioeconomic impact study of a geothermal development is to identify the three key sectors that are affected by the project: investor, beneficiary, and victim. The final outcome of the project must have the equity of benefit and compensation to all affected sectors. The main aim of a socioeconomic study is to determine the changes in conditions within, adjacent to, and in proximity to the project area as well as the direct and indirect impacts of these changes. This requires past and current baseline data, usually going back 10 years. These data are gathered from records, surveys, and consultations with the local communities and other interest groups. Social acceptability of projects are normally obtained if no extreme changes result and the affected communities can adjust their lifestyles to the new benefits that are gained. These benefits may include employment, lower electricity prices, and funding for local charities.

The most difficult problem that a geothermal power plant faces is the reliability of the steam supply. Well characteristics can and do change with time after a period of production. Many geothermal power stations (e.g., Wairakei, The Geysers) have never been loaded to their full capacities due to the shortfall of well output. The initial logical solution to this problem was to drill more wells, but this practice tends to worsen the problem with rapid drawdown. This problem is due to the overestimation of the reservoir power potential. The solution to this problem is staged development. The expected drawdown was incorporated during the design stage at Ohaaki, but the drawdown was faster than predicted.

A major operational problem that a geothermal power plant faces is the solid deposition of silica or calcite in wells and equipment exposed to the geothermal fluids. Although the chemistry of the deposition can be predicted with reasonable accuracy, the solution to the deposition problem is not always clear-cut. The common solution is to inject a chemical inhibitor into the problem wells. Suitable calcite inhibitors include polyacrylate, polymaleic acid, and phosphonate, depending on the application temperature. Many acids or alkali are suitable silica inhibitors.

The inhibitor is injected just below the fluid flashing point where the solid precipitates out of the liquid phase during flashing (Fig. 14). The inhibitor injection is very effective if applied correctly, although it does not actually stop deposition altogether. Physical removal of the solid deposits is still required once every few years instead of a few times each year without inhibitor. Not only is there a loss of production due to well shutdown to remove the deposits, but the cleaning may require a drilling rig or expensive acids.

Another hazard associated with a geothermal power plant is earthquake and/or volcanic eruption.

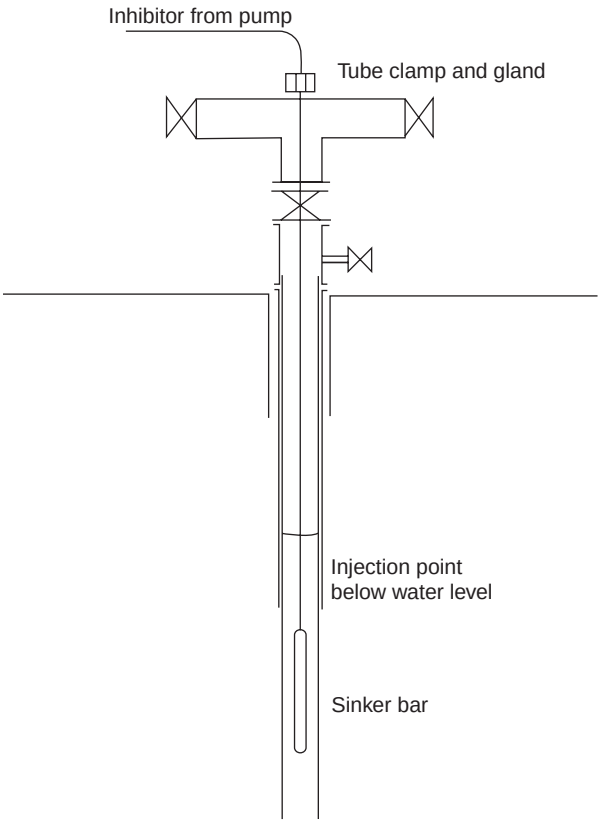


FIGURE 14 A typical inhibitor injection arrangement in a production well.

All plants and equipment can be easily designed to resist the expected seismic forces. Design for volcanic hazard is relatively easy if the discharge is slow-moving lava because it is simply a matter of locating the plant outside the lava flow path. It is much more difficult if the discharge is of the ash type.

7. COMMERCIAL DEVELOPMENT OF GEOTHERMAL POWER PROJECTS

In most countries, geothermal resources are classified as natural resources belonging to the government. Since the 1980s, there has been a worldwide trend to corporatize and/or privatize government commercial activities. Whereas the geothermal resources in the United States have been commercially developed by the private sector since the 1960s at The Geysers, other countries, notably The Philippines, Indonesia, and New Zealand, did not started private commercial geothermal projects until the 1980s. These private power developers are known as independent power producers (IPPs). There is a perception that the private sector is more efficient than the public sector. The efficiency arises from the increased flexibility of the private sector and its accessibility to commercial funding.

There are several types of contracts in the commercial development of geothermal power projects. In a BOT (build, operate, and transfer) contract, a private developer builds the whole power station (including the steamfield), operates the power station for an agreed period of time to recoup its investment, and then transfers the power station back to the host country. In a BOO (build, own, and operate) contract, the developer builds and operates the power station as its own business. In a BOOT (build, own, operate, and transfer) contract, the developer transfers the power station back to the host country after an agreed period.

Developing a geothermal resource carries higher commercial risks than does building and operating a power plant. From a private developer's viewpoint, it is more favorable to just have an energy conversion agreement (ECA) contract in which the private developer builds and operates a power plant, converting the geothermal steam supplied by the host country and producing electricity for the host country (Fig. 15). The host country develops the geothermal resource. An ECA contract will involve a steam sales agreement (SSA) and a power purchase agreement (PPA). In a more traditional EPC (engineer, procure, and construct) contract, an engineering consulting company designs and builds the power plant and/or steamfield for a fee.

When private developers first entered the commercial development of geothermal power projects, they developed the resources, operated the steamfields, and sold steam to power companies operated by government agencies. The classic IPP operation is for the private sector to develop both the geothermal resource and the power plant and to sell electricity to the host utility that deals with the electricity market. The latest trend is the ECA operation.

An IPP, like any other investor, looks for a good return on its investment. To minimize the investment risk, the IPP wants certainty in costs and revenues over the investment period. A geothermal power project is like a hydropower project—capital intensive but with a low running cost. A geothermal project is commercially viable only if it creates a positive net present value (NPV) over its economic life for the investors. This requires a guaranteed stream of revenue in the future from the power sales, and the risks in the steam supply from the

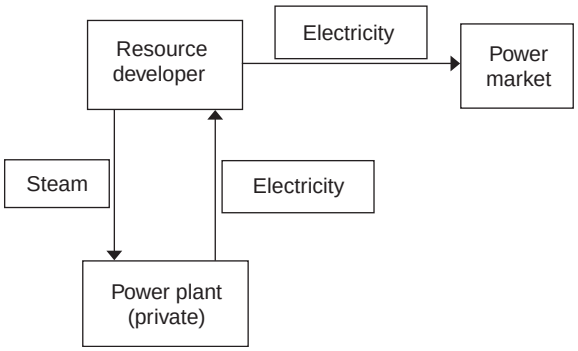


FIGURE 15 ECA project structure.

geothermal resource must be manageable. Therefore, the selection of the exploitation process, power cycle, and power plant is critical to minimize capital cost and maximize power output.

Risks are inherent in the resource and power plant development. Resource exploration risk is relatively high, with a corresponding higher return potentially at moderate investment. This type of risk is familiar to the mining and oil exploration companies. The steamfield and power plant development risk is low. Most IPPs and bankers are familiar with this risk. Risks are best managed by those who are best at controlling them. A developer is best responsible for the project implementation, and contractors are best at managing risks within their own areas of work. Some risks, such as resource and steamfield process risks, must remain with the owner. Risk allocation has driven the move toward EPC contracts.

There is an apparent conflict between the financial goal of maximizing return with a rapid development program and the risk mitigation strategy to minimize risk with a slow program. The solution to this apparent dilemma is to adopt a staged development.

The production drilling stage is a major expenditure phase, with each well costing U.S. $750,000 to $3 million. Well outputs are highly variable, from a dry well to one that may produce 40 MWe. Once production drilling starts, the aim is to produce electricity, and hence return on investment, as soon as possible. Cost controls are paramount during the production drilling stage.

A decision is required for the generating capacity of the power plant. For a staged development, the capacity of the first stage can be 25% of the estimated power potential. Because there are economies of scale in turbo-generator unit size, the tendency is to go for the largest single unit of approximately 55 MWe if possible. However, this requires optimization against operational flexibility and power demand. In general, the power plant costs approximately 50 to 55% of the total cost of a geothermal power project, the production drilling costs 25%, and the steamfield costs 15%, with the remaining 5 to 10% of costs due mainly to exploration drilling. Currently, a geothermal power project costs between U.S. $1000 and $2000/kWe, depending on the quality of the steam. An electricity unit cost of U.S. 5 to 8 cents/kWh is achievable. This is very competitive compared with conventional thermal and hydropower stations.

Acknowledgments

I thank the following people for their support and assistance: D. H. Freeston, P. R. L. Browne, C. Morris, L. Bacon, K. L. Brown, and C. Z. Lee. I acknowledge the following permissions: M. Dickson and M. Fanelli for Table I, International Geothermal Association for environmental course materials convened by K. L. Brown (Geothermics) for Fig. 1, Geothermal Energy Association for Table II, H. Hole for drilling lecture notes, J. B. Randle for commercial development lecture notes, K. Koorey for Fig. 13, and K. C. Foong for steamfield lecture notes. I also thank the Geothermal Institute for office space.

SEE ALSO THE FOLLOWING ARTICLES

Geothermal Direct Use • Ground-Source Heat Pumps • Ocean Thermal Energy • Thermal Energy Storage

Further Reading

Armstead, H. C. H. (1983). "Geothermal Energy: Its Past, Present, and Future Contributions to the Energy Needs of Man," 2nd ed. E & FN Spon, London.

Clark, I. F., and Cook, B. J. (Eds.). (1992). "Geological Science: Perspectives of the Earth." Australian Academy of Science, Canberra.

Dickson, M. H., and Fanelli, M. (Eds.). (1990). "Small Geothermal Resources: A Guide to Development and Utilization." UNITAR/UNDP Center on Small Energy Resources, Rome.

Goff, F., and Janik, C. J. (2000). Geothermal systems. *In* "Encyclopedia of Volcanoes" (H. Sigurdsson, Ed.). Academic Press, San Diego.

Grant, M. A., Donaldson, I. G., and Bixley, P. F. (1982). "Geothermal Reservoir Engineering." Academic Press, New York.

Hochstein, M. P., and Browne, P. R. L. (2000). Surface manifestations of geothermal systems with volcanic heat sources. *In* "Encyclopedia of Volcanoes" (H. Sigurdsson, Ed.). Academic Press, San Diego.

James, R. (1970). Factors controlling borehole performance. *In* "Proceedings of UN Symposium on Development and Utilisation of Geothermal Resources" (Vol. 2, pt. 2, pp. 1501–1515). United Nations, Pisa, Italy.

Kestin, J. (Ed.). (1980). "Sourcebook on the Production of Electricity from Geothermal Energy." Brown University Press, Providence, RI.

Lee, K. C. (1997). Geothermal power technologies for different types of geothermal resources. *In* "New Energy Technology for Asia Pacific" (World Energy Council Asia Pacific Regional Forum, Vol. 2, pp. 189–203). WEC China National Committee, Beijing.

Lee, K. C. (2001). Classification of geothermal resources by exergy. *Geothermics* **30,** 431–442.

Morris, C. (2003). "Operation and Maintenance Experiences of Wairakei Geothermal Power Station" (preprint). Contact Energy Ltd., New Zealand.

Smith, I., Stosic, N., and Aldis, C. (1995). Trilateral flash cycle system: A high efficiency power plant for liquid resources. *In* "Proceedings of the World Geothermal Congress" (Vol. 3, pp. 2109–2114). WGC, Florence, Italy.

Glass and Energy

CHRISTOPHER W. SINTON
Alfred University
Alfred, New York, United States

1. Introduction
2. Types of Glass
3. Introduction to Commercial Glass Manufacturing
4. Energy

Glossary

container glass The largest glass sector; formed by a two-stage process of initial forming by pressing followed by blowing to obtain the finished hollow shape.
cullet Scrap or waste glass that is added to the batch materials.
float glass Flat glass that is made by floating molten glass on a pool of molten tin.
glass A material with a three-dimensional network of atoms that forms a solid lacking the long-range periodicity typical of crystalline materials.
optical fibers Very high-purity silica glass fibers used to transmit light for telecommunications.
oxy-fuel firing The replacement of the combustion air with oxygen in a glass furnace.
recuperator A continuous heat exchanger that preheats combustion air with the heat from the exhaust gases.
regenerator A stack of refractory material that absorbs the heat from exhaust gases. There are two regenerators in a regenerative furnace that are cycled over a specific time period.

Glass is a material that enjoys widespread use in many commercial applications, including buildings, beverage containers, automobiles, and telecommunications. However, the manufacture of glass is a very energy-intensive undertaking and the industry is constantly working to reduce its energy consumption. This article describes the use and flow of energy in commercial glass production. In order to understand the relationship between energy and glass manufacturing, this articles first defines what glass is and then describes how it is produced.

1. INTRODUCTION

The atomic or molecular structure of solid materials can generally be found in two states: ordered and disordered. If the atoms of a material are ordered in a repeated array throughout, the structure will be an ordered three-dimensional pattern. This is a crystalline material. Solid materials that lack long-range atomic order are noncrystalline or amorphous. When a material is cooled from a liquid state to a solid, its formation into a crystalline or amorphous solid depends on how easily the random atomic structures of the liquid can organize into an ordered state while cooling. Factors such as composition and cooling rate are important. Rapid cooling of a liquid enhances the formation of an amorphous structure since the material is given little time to reach an ordered state. This is a glass. Figure 1 shows a comparison of the ordered arrangement of crystalline silicon dioxide and the disordered arrangement of a silicon dioxide glass. A common definition of a glass was given by the physicist W. H. Zachariasen, who stated that glass is an extended, three-dimensional network of atoms that forms a solid lacking the long-range periodicity (or repeated, orderly arrangement) typical of crystalline materials.

Manufactured glass products are made by melting crystalline raw materials into a molten state. The melt is subsequently shaped into the desired product and cooled to make a rigid form. The composition of the glass depends on the desired properties of the glass, which in turn define which raw materials are needed. Conceptually, there are thousands of different combinations of raw materials that will produce a glass, but in practice there are a limited number of formulations that are used to make the common glass products that are used on a daily basis. These limitations are typically dictated by practical and economic considerations.

 © 2004 Elsevier Inc. All rights reserved.

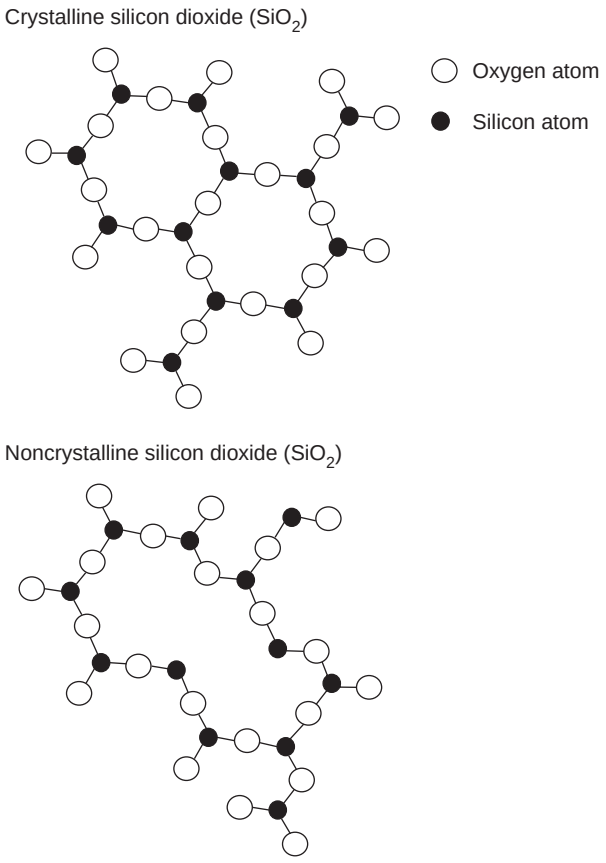


FIGURE 1 Schematic diagram showing the ordered arrangement of crystalline silicon dioxide and the disordered arrangement of a silicon dioxide glass.

2. TYPES OF GLASS

Glasses can be found in the natural environment, mostly in areas of volcanism where erupted magma is quickly cooled before crystallization can occur. Perhaps the best known example is obsidian, a black glass that consists mainly of oxides of silicon, aluminum, calcium, and sodium and/or potassium.

Commercially produced glasses differ widely in chemical composition and in physical and optical properties. Various materials can form a glassy or vitreous structure, such as the oxides of silicon and boron. Most commercial glasses are based on silica (SiO_2), which is the most abundant material on the earth's crust and is an excellent glass former. There are some nonsilicate glasses, but they make up only a very small fraction of commercially produced glass.

Pure vitreous silica glass can be made by melting quartz sand and cooling it quickly. Although silica makes an excellent glass for many applications, its high melting temperature ($>2000^{\circ}C$) makes it too expensive for use in commodity glass products such as bottles and windows. The melting temperature can be reduced to a practical range by the addition of fluxes, the most common of which are alkali oxides (Li_2O, Na_2O, and K_2O). However, the addition of too much alkali can lead to the degradation of properties, such as the chemical durability of a glass (i.e., it will begin to dissolve when exposed to water). This can be compensated for by the addition of property modifiers, such as alkaline earth oxides (CaO and MgO) and aluminum oxide (Al_2O_3). The addition of these glass modifiers changes the glass structure, resulting in changes in the physical and chemical properties of the glass.

The composition used for a specific glass product depends on the desired final product performance, manufacturing characteristics (e.g., melting and forming), environmental considerations, and raw material/fuel costs. Although there are many different combinations of elements that will make a glass, there are some general compositional types that are widely used in commercial glass making. A glass made of silica, alkali, and alkaline earth elements is commonly called a soda-lime-silica glass and is the most common type. It has a relatively low melting temperature (begins melting at $\sim 800^{\circ}C$ and is usually processed up to $1400^{\circ}C$), has the proper viscosity characteristics for continuous production, and uses raw materials that are relatively inexpensive. It is the basis for making flat glass, containers, and consumer ware (e.g., drinking glasses).

Borosilicate glasses contain boron, which decreases thermal expansion, lowers viscosity, and allows faster fabrication speeds. Borosilicate glasses are used when resistance to thermal shock is desired, such as for cookware laboratory glass. Fiberglass products that are used for thermal insulation and structural reinforcement of composites contain boron.

Lead crystal and crystal glass are silicate glasses with added lead. The lead decreases viscosity and, hence, decreases the temperatures needed to work the glass and enhances the dissolution of refractory particles. Lead increases the refractive index of the glass, imparting a brilliance to the finished piece.

Specialty glasses include a variety of products, such as cathode ray tubes (CRTs), lighting glass (tubes and bulbs), laboratory and technical glassware, optical fibers, ceramic glasses (cookware), and glass for the electronics industry (e.g., LCD panels). Optical fibers are a specialty glass that is made of extremely high-purity silica. Optical fiber glasses cannot attain the needed purity levels produced by melting silica sand, so the glass is manufactured by precipitating chemically purified silica vapor.

3. INTRODUCTION TO COMMERCIAL GLASS MANUFACTURING

Glass manufacturing is essentially a commodity industry, with the majority of product sold to other industries, such as beverage, construction, and automobile manufacturers. In 1999, the U.S. glass manufacturing industry produced more than 18 million metric tons, worth approximately $17.6 billion. In the European Union, 1996 production was approximately 29 million metric tons (Table I).

3.1 Raw Materials

The overall goal of glass manufacturing is to convert crystalline raw materials into a homogeneous, flowing liquid that is free of visible defects that can be formed into a final product. This needs to be accomplished as quickly and as economically as possible and, at the same time, comply with all environmental regulations. Figure 2 shows a schematic diagram of the process of commercial glassmaking as described in this section. There are several methods of making and forming glass, but all begin with the mixing of raw materials or "batching." For most glasses, the most important glass former is silica. The raw material for this is quartz sand usually derived predominantly from sandstone deposits. Such deposits must contain few impurities, such as iron, which will affect the color of the glass. Glass property modifiers are introduced with a mix of limestone ($CaCO_3$) and dolomite ($CaMg[CO_3]_2$). The alkali flux is usually soda ash (Na_2CO_3), which comes from either a mined mineral ore called trona or the Solvay process of reacting sodium chloride with limestone. The soda and lime materials are carbonates so that upon heating the carbonate is driven off as volatile CO_2, leaving the metal oxide behind. Feldspar or nepheline syenite are aluminosilicate minerals that can be used in the batch to add alumina (Al_2O_3) and some alkali. Borosilicate glass batches can have 5–10% borate minerals, such a borax ($Na_2O \cdot 2B_2O_3 \cdot 5H_2O$) and colemanite ($Ca_2B_6O_{11} \cdot 5H_2O$) added to them. Mineral or rock wool is often made from slag that is a by-product from metal smelting.

Recycled glass, or cullet, is added to the batch to reuse waste glass from both within the plant and outside sources. Cullet also provides an early melting liquid phase that enhances reaction rates. The effect of cullet on energy use is discussed later. The main problem associated with cullet is the potential for contamination, particularly from pieces of metal, ceramic, or glass-ceramic. These contaminants can, at best, lead to rejected product and, at worst, damage the expensive refractory lining.

Minor additives include colorants, reducing agents, and fining agents. Colorants are added to control or add color and are usually oxides of transition metals or rare earth elements, such as cobalt, cerium, iron, and chrome. Fining agents are used to promote the removal of bubbles from the melt that, if not removed, will result in flaws in the final product. Common fining agents include sulfates,

TABLE I

Glass Sectors and Annual Production (Million Metric Tons)

Sector	European Union[a]	United States[b]
Flat glass	6.4	4.5
Container glass	17.4	8.7
Fiber	2.5	2.3
Speciality/Pressed/Blown	2.7	2.7
Total	29.0	18.3

[a] 1996 estimates.
[b] 1999 estimates.

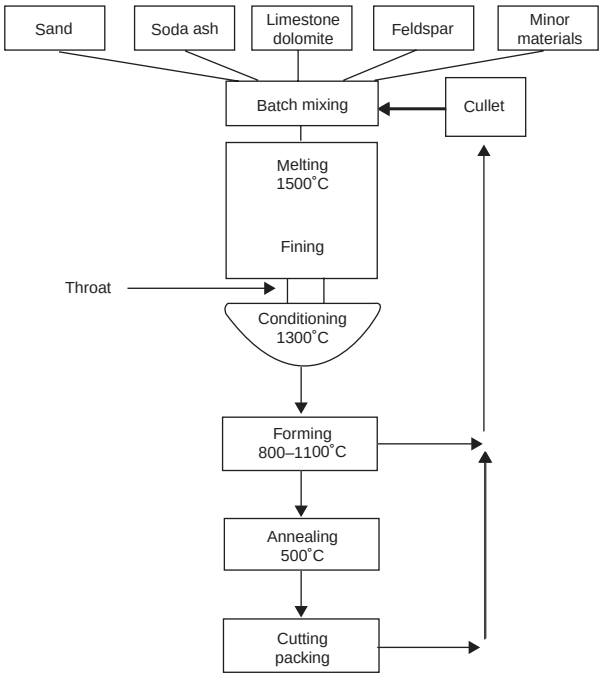


FIGURE 2 Schematic diagram of the process of commercial soda–lime–silica glassmaking showing raw materials, process sections, and approximate operating temperatures.

fluorides, nitrates, and oxides of arsenic or antimony. The most common combination for soda–lime–silica glasses is sodium sulfate and a reducing agent such as carbon.

The batching process consists of transferring raw materials directly from storage silos into a hopper and subsequently mixing the raw batch in a large-scale mixer. Cullet may be added after mixing. The mixed batch is placed in storage hoppers located directly adjacent to the "dog house," or the end of the glass furnace at which the batch is introduced and melting commences. The batch may be moistened with water or sometimes caustic soda to reduce particle segregation during transport to the furnace.

TABLE II

Advantages and Disadvantages of Different Furnace Types

Furnace type	Advantages	Disadvantages
Regenerative	Long furnace life	Higher NO_x levels
	Long-term experience	Higher capital cost
Direct fired (mostly oxygen/fuel)	Reduction in fuel	Refractory corrosion
	NO_x reduction	Short-term experience
	Lower capital cost	Cost of oxygen production
	Particulate reduction	
Electric	Efficient	High electricity cost
	No one-site pollution	Short furnace life

3.2 Glass Melting Furnaces

There are several types of melting furnaces used in the glass industry depending on the final product, raw materials, fuel choice, the size of operation, and economic factors. A melter can be either periodic or continuous. The latter is most widely used in large-scale operations. Continuous melters maintain a constant level by removing the glass melt as fast as raw material is added. Table II shows the three major types of furnaces and their advantages and disadvantages. The energy for melting comes from either the combustion of fossil fuels or electricity. The temperature necessary for melting ranges between 1300 and 1550°C. All furnaces are lined with high-temperature refractory materials to keep the corrosive melt from escaping and to insulate the melter.

Regenerative combustion furnaces recover heat from the exhaust stream to preheat the incoming combustion air by alternatively passing the exhaust and combustion air through large stacks of latticework refractory brick (regenerators or checkers). There are two sets of regenerators, so that as one is being preheated by the exhaust gases the other is transferring heat to the incoming combustion air (Fig. 3). The cycle is reversed approximately every

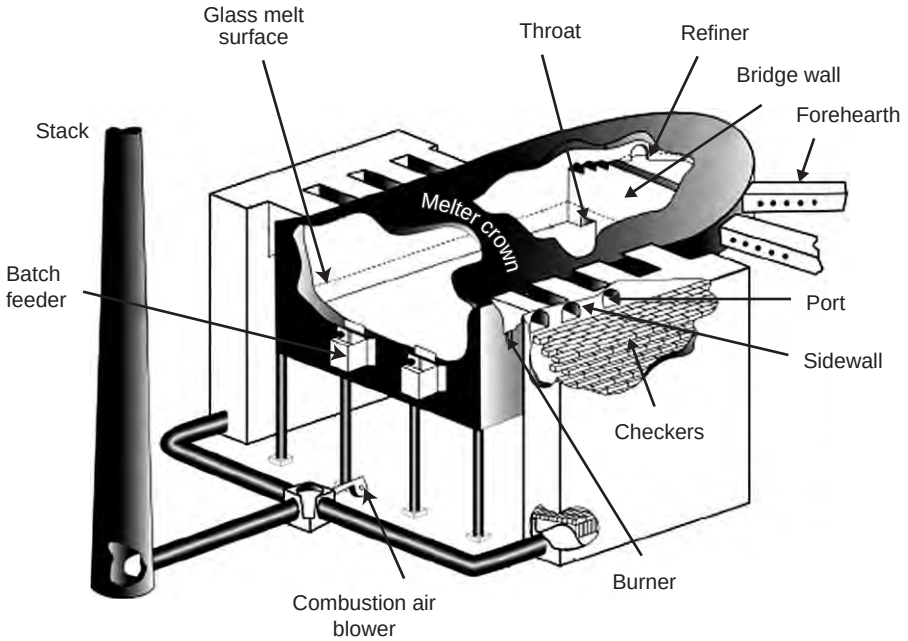


FIGURE 3 Diagram of a cross-fired regenerative furnace. This type is often used in large-volume production of container and flat glass. From U.S. Department of Energy (2002). "Energy and Environmental Profile of the U.S. Glass Industry."

20 min. Most glass-container plants have either end-fired (burners at each end) or cross-fired (burners on each side) regenerative furnaces, and all flat glass furnaces are cross-fired with five or six ports on each side with two burners for each port. Combustion air preheat temperatures of up to 1400°C may be attained, leading to very high thermal efficiencies. A variant of the regenerative furnace is the recuperator, in which incoming combustion air is preheated continuously by the exhaust gas through a heat exchanger. Recuperative furnaces can achieve 800°C preheated air temperatures. This system is more commonly used in smaller furnaces (25–100 tons per day). For large-capacity installations (>500 tons per day), cross-fired regenerative furnaces are almost always used. For medium-capacity installations (100–500 tons per day), regenerative end-port furnaces are most common.

A direct-fired furnace does not use any type of heat exchanger. Most direct-fired combustion furnaces use oxygen rather than air as the oxidizer. This is commonly called oxy-fuel melting. The main advantages of oxy-fuel melting are increased energy efficiency and reduced emission of nitrogen oxides (NO_x). By removing air, nitrogen is removed, which reduces the volume of the exhaust gases by approximately two-thirds and therefore reduces the energy needed to heat a gas not used in combustion. This also results in a dramatic decrease in the formation of thermal NO_x. However, furnaces designed for oxygen combustion cannot use heat-recovery systems to preheat the oxygen. Initially, furnaces that use 100% oxy-fuel were used primarily in smaller melters (<100 tons per day), but there is a movement toward using oxy-fuel in larger, float glass plants.

An electric furnace uses electrodes inserted into the furnace to melt the glass by resistive heating as the current passes through the molten glass. These furnaces are more efficient, are relatively easy to operate, have better on-site environmental performance, and have lower rebuild costs compared to the fossil-fueled furnaces. However, fossil fuels may be needed when the furnace is started up and are used to provide heat in the working end or forehearth. These furnaces are most common in smaller applications because at a certain size, the high cost of electricity negates the improved efficiency.

Some regenerative furnaces use oxygen enrichment or electric boost to optimize the melting process. Electric boosting adds extra heat to a glass furnace by using electrodes in the bottom of the tank. Traditionally, it is used to increase the throughput of a fossil fuel-fired furnace to meet periodic fluctuations in demand without incurring the costs of operating a larger furnace. It is also used to enhance the pull rate (the output at the forming end) of a furnace as it nears the end of its operating life.

3.3 The Melting Process

Regardless of the furnace type used, there are four basic steps to melting glass:

1. Conversion of the batch into a molten liquid free of undissolved crystalline material (called stones)
2. Removal of bubbles, or fining
3. Homogenizing the liquid so that it is chemically and thermally uniform
4. Conditioning the melt to bring it to a state in which it can be formed into the desired product

For a continuous furnace, these four steps occur simultaneously. The average residence time, from introduction of the batch to forming, is on the order of 24 h for container furnaces and up to 72 h for float glass furnaces. Glass melting tanks can contain 1500 tons or more of molten glass.

The melting of the batch consists of several steps from the moment it is introduced into the tank. Due to the low thermal conductivity of the batch materials, the initial melting process is quite slow. As the materials heat up, the moisture evaporates, some of the raw materials decompose, and the gases trapped in the raw materials escape. The first reaction is the decomposition of the carbonates at approximately 500°C, which releases CO_2. The fluxes begin to melt at approximately 750°C and begin to react with the silica. At 1200°C, the silica sand and other refractory materials dissolve into the molten flux. As the remaining silica dissolves into the melt, its viscosity increases rapidly.

Bubbles are generated by the decomposition of raw materials, from reactions between the melt and the glass tank, and from air trapped between batch particles. These are removed as they rise buoyantly through the melt to the surface or as gases dissolve into the melt. Removal of the last few bubbles is the most difficult. Fining agents enlarge these bubbles to assist in their removal. Fining occurs as the melt moves forward through the throat of the melter to the conditioner, where melt homogenization and cooling take place.

Without homogenizing the melt, linear defects called cords (regions of chemical heterogeneity) will be seen in the final product. Furnaces are designed to promote recirculating convective currents within the melt to homogenize the glass. Convective mixing can

be enhanced by physically stirring the melt with bubbles or mechanical stirring devices.

3.4 Forming Processes

After melting, fining, and homogenization, the melt passes through the throat of the furnace (Fig. 3) into a conditioning chamber, where it is cooled slowly to a working temperature between 900 and 1350°C. It is then ready to be delivered as a viscous mass through the forehearth to be formed into a specific product. The role of the forehearth is to remove temperature gradients and transfer the molten glass to forming operations. It consists of an insulated refractory channel equipped with gas-fired burners or electric heaters and an air cooling system.

Flat glass is most commonly formed by the float process, whereas patterned and wired glass are formed by rolling. Float glass and rolled glass are produced almost exclusively with cross-fired regenerative furnaces. The float process was developed by the Pilkington Brothers Glass Company of England in the 1950s, and it is the technique of pouring the melted glass onto a pool of molten tin. The glass floats on the tin and spreads out to form a uniform thickness. The glass is then mechanically stretched or constrained to attain the desired ribbon width and thickness. The tin bath is surrounded by a reducing atmosphere of nitrogen and hydrogen to keep the bath from oxidizing. The glass ribbon is pulled and modulated by water-cooled pulleys and exits the bath into the annealing lehr. Annealing is the process of heating and cooling the glass at a controlled rate to remove any residual stresses that could cause catastrophic failure while in use. After exiting the lehr, the glass is automatically cut into the desired sizes. The float method uses much less energy than preceding plate and sheet glass forming due to the use of more efficient furnaces and the elimination of the need for surface polishing and grinding. Rolled glass is formed by squeezing molten glass at approximately 1000°C between water-cooled steel rollers to produce a ribbon with a surface pattern or embedded wire.

Containers such as bottles and jars are made using a two-stage pressing and blowing process. This process is fully automated and begins with molten glass flowing from the furnace forehearth to a spout or set of spouts. The flow of glass is controlled by a mechanical plunger that produces a piece or gob of molten glass. The gob is then conveyed to the forming process, which produces a primary shape in a blank mold using compressed air or a metal plunger. The primary shape is placed into the finish mold and given its final shape by blowing compressed air into the container. During the forming process, the glass temperature is reduced by as much as 600°C to ensure that the containers are sufficiently rigid when placed on a conveyor. The cooling is achieved with high volumes of blown air against the molds. After forming, the containers are annealed by reheating to 550°C in a lehr and then cooled under controlled conditions. Larger operations can produce 500 bottles per minute.

Glass fibers can be found in nature where fountaining lavas are wind blown to produce long strands known as Pele's tears, after the Hawaiian goddess of fire. The manufacture of commercial glass fiber is similar, with some exceptions. Most fiber glass melts are produced with cross-fired recuperative furnaces, although there are several oxy-fuel fired furnaces in Europe. Short-strand glass wool is produced by spinning the melt in a crucible lined with holes. As the molten glass is ejected through the holes, high-velocity air blows the glass into fibers. For continuous glass fibers, the melt flows from the furnace through a series of refractory-lined channels to bushings. Bushings have several hundred calibrated holes or bushing tips that are electrically heated to precisely control the temperature and, thus, the viscosity and flow rate of the melt. The melt is drawn through the bushing tips by a high-speed drum to form continuous filaments. Once drawn from the bushing, the filaments are quickly cooled and then coated with a polymer binder.

4. ENERGY

Based on the previous description of glassmaking, it is evident that glass manufacturing is a very energy-intensive process. Energy accounts for 15–25% of the cost of glass products, depending on the sector and the country of manufacture. In 1999, the U.S. glass industry spent more than $1.3 billion on approximately 295 quadrillion joules (PJ) [280 trillion British thermal units (Btu)], including products from purchased glass. Taking into account electricity losses due to generation, transmission, and distribution, this figure increases to 417 PJ (395 trillion Btu) of total energy use.

By far the most energy-intensive components of glass manufacture are melting and refining, which together account for 70–80% of the total energy requirements. The remaining energy is used in the forehearths, tin bath (for float glass), forming/cutting processes, annealing, and other plant services. For

example, in container manufacturing, the furnace accounts for 79% of the total energy used. The other areas of energy use are the forehearth (6%), compressed air (4%), mold cooling (2%), lehr (2%), and others (7%).

4.1 Sources of Energy

The sources of energy used to produce glass are natural gas, electricity, and, to a lesser extent, fuel oil. Fuel oil has been historically important, but today natural gas is the dominant fossil fuel used in glassmaking. Natural gas has higher purity, is relatively easy to control, and does not require on-site storage facilities. The higher purity of natural gas results in a reduction of sulfur dioxide emissions relative to fuel oil. Natural gas accounts for approximately 80% of energy use and more than 98% of purchased fuels in the U.S. glass industry. Although natural gas is mainly used to melt the raw materials, it is also used in emission control either to reduce NO_x in the exhaust stream or to incinerate toxic emissions from glass fiber production. Electricity is used in smaller resistance melters, fiberglass bushings, tin bath heating, annealing, and for powering motors to produce compressed air and operate conveyers and blowers.

4.2 Theoretical vs Actual Energy Use in Melting

The theoretical minimum energy required to transform the crystalline raw materials to a molten liquid is approximately 2.68 GJ/metric ton for a soda–lime–silica glass formulation. The calculation assumes all available heat is fully utilized and has three components:

- The heat of fusion to form the melt from the raw materials
- The heat, or enthalpy, required to raise the temperature of the glass from 20 to 1500°C
- The heat content of the gases released during melting.

The actual amount of energy required can be from 2 to 15 times the theoretical minimum, although for the majority of large-scale melters it is generally less than 8 GJ/ton. The amount of energy per unit of product produced can be referred to as the specific energy or energy intensity. As a comparison to other materials, the values for glass are similar to those of cement (5.4 GJ/ton) and lower than those of steel and iron (25.4 GJ/ton) and pulp and paper products (32.8 GJ/ton).

In addition to the specific product, the type of furnace and the desired quality of the glass affect the specific or per unit energy use. For instance, because flat glass needs to be completely free of imperfections, it generally requires longer residence times in the furnace than does container glass and therefore uses more energy per ton of glass produced. Table III shows the range of specific energy values for several different fossil fuel furnaces and glass products. Efficiencies in large regenerative furnaces can be as high as 55% and electric furnaces are 70–90% efficient (not considering the 25–30% efficiency of off-site electricity generation). Electric melters are more efficient because there are no heat losses in the stack or regenerator. Table IV shows the range of efficiencies for several types of furnaces.

The large differences between the theoretical and the actual energy used can be attributed to heat losses and other inefficiencies. The majority of heat loss is

TABLE III

Energy Consumption for Melting and Refining

Furnace type	U.S. average specific energy (GJ/metric ton)	EU average specific energy (GJ/metric ton)[a]
Flat glass		
Regenerative side-port	8.4	6.3
Electric boost side-port	6.2	—
Container glass		
Large regenerative	7.5	4.2
Electric boost	5.3	—
Oxy-fuel fired[b]	4.5	3.3
Small electric melter	2.7	—
Pressed and blown glass		
Regenerative	5.3	—
Oxy-fuel fired[b]	3.5	—
Electric melters	9.9	—
Recuperative	—	6.7
Regenerative TV tube	—	8.3
Insulation fiber glass		
Electric melters	7.2	—
Recuperative melters	6.7	4.3
Oxy-fuel fired[b]	5.4	—
Textile fiber		
Recuperative melters	10.1	—
Oxy-fuel fired[b]	5.4	—

[a] EU values assume 70% cullet for container glass, 40% cullet for television tube and tableware, and 20% cullet for flat.

[b] Does not include energy used in oxygen production.

TABLE IV

Comparison of Energy Efficiency Characteristics

Recuperative furnaces	20–40%
Regenerative furnaces	35–55%
Electric furnaces[a]	70–90%
Oxy-fuel[b] furnaces	40–55%
Regenerative w/batch/cullet preheater	50–65%
Oxy-fuel w/batch/cullet preheater[b]	50–65%

[a] Does not include 25–30% efficiency of off-site electricity generation.

[b] Does not include energy used in oxygen production.

from the furnace structure and through the removal of exhaust gases for fossil fuel melters. As described previously, modern melters that burn fuel employ a regenerator or recuperator to recover heat from the exhaust gases and preheat the incoming combustion air. This reduces the overall energy consumption by 50–70% compared to if it were direct-fired. The energy source, heating technique, and heat recovery method are central to the design of the furnace. The same choices also affect the environmental performance and energy efficiency of the melting operation.

4.3 Energy Reduction

The glass industry is constantly striving to close the gap between the theoretical and actual energy needed to melt raw materials into glass. Because melting and refining use the majority of energy in glass making, it is this part of the process that must be examined initially in any attempt to reduce per unit energy consumption. There are two general approaches toward reducing energy consumption: optimizing the existing system or completely replacing the current methods with and advanced melting system.

To reduce the amount of energy in conventional melting furnaces, the following may be addressed:

- Improvement of the combustion efficiency
- Reduction of the residence time in the tank
- Improvement of the insulation around the structure
- More effective use of the exhaust gas heat
- Oxy-fuel firing
- Increased cullet use

The melting technique chosen by a manufacturer is largely determined by economic considerations, and this can have a major effect on the energy efficiency. Although energy efficiency is an important aspect of the operating costs, these must be balanced by the other capital and operating costs, the desired production rate, and environmental performance (although energy efficiency and environmental performance are related).

Improvements in combustion efficiency have been driven by the need to reduce NO_x emissions. Reengineering of the burners and the control systems has also resulted in overall energy savings. Low NO_x burners reduce the volume of combustion air to close to stoichiometric levels, such that only the amount of air needed for combustion is introduced. Air that is introduced but not used in the combustion process would simply be heated and would carry that heat away in the exhaust.

The melt needs time (residence time) in the tank to produce a defect-free glass. However, the longer the melt remains in the tank, the more energy is expended in the process. Any reduction in the amount of time needed to melt, fine, or homogenize the melt will result in direct energy reductions. Residence times depend on the furnace temperature, composition of the batch, grain size of the batch ingredients, amount and grain size of cullet, and homogeneity of the batch. Improved batching techniques include ensuring optimal grain sizes and the addition of cullet. Because of the low viscosity of the initial melting fluxes, there may be a tendency for the batch to "unmix" and require excessive homogenizing times. Pelletizing or pre-reacting some of the batch components may reduce this effect. Furnace temperature can be increased to reduce melting and refining times, although this is done at the expense of the life of the refractory lining. Electric boost can be used in fossil fuel melters to increase the pull without changing the furnace size. In general, the electricity required to increase the pull rate by 1 ton is approximately 22–28 kW. The degree of homogeneity depends on the quality requirements, which vary considerably between and within glass sectors. Bubbling is a method of decreasing homogenization time in which air is injected through several nozzles installed at the bottom of the melting chamber, agitating the melt and promoting mixing.

As much as 30% of energy can be lost through the furnace structure. Structural heat losses are inversely proportional to the furnace size because of the change in melter surface area-to-volume ratio. Thus, larger furnaces have inherently lower heat losses. Regardless of the size of the melter, improvements in the insulation of the structure will increase energy efficiency. Most glass contact refractories that are used to line the furnace are dense, fusion cast materials that have high thermal conductivity and

need to be insulated. The application of insulation to a furnace depends on the area of the furnace and the operating conditions, and not all parts can be insulated. Insulation increases the operating temperature of the refractory lining, which tends to shorten its life span. Overheated refractory material can also degrade and shed material into the glass, creating defects. Therefore, more efficient insulation needs to be balanced with the quality of the glass and lifetime of the refractories.

Advances in refractory material engineering have allowed furnaces to operate longer between rebuilds at higher levels of insulation. For example, superduty, high-purity silica and alumina–zirconia–silica bricks have been developed for the crown to improve insulation. Sprayed fiber insulation can be applied to the regenerator structure, resulting in up to a 50% reduction in regenerator structural heat losses.

Up to 30% of the heat from melting is expelled through the exhaust gases. Furnaces equipped with regenerators or recuperators recover some of this heat, with end-fired regenerative furnaces generally more efficient than cross-fired regenerative furnaces. However, combustion control and furnace size are more limited for end-fired furnaces. The amount of energy recovered by regenerators may be increased by increasing the quantity of refractory bricks in the checkers using enlarged regenerator chambers or in separate but connected structures. However, as one approaches the physical maximum efficiency, there are limitations on the cost of the extra refractory bricks and the limitation of available space. Improved materials and designs such as corrugated cruciform bricks can improve heat transfer.

Gases exiting the regenerators or recuperators are 300–600°C, and there are other potential areas of use for the remaining waste heat, although too much heat loss in the exhaust stream will adversely affect the ability of the plant to vent the exhaust up the stack. One heat-recovery technique is to pass exhaust gases through a boiler to generate steam. The steam can be used for heating, on-site electricity generation, or to drive air compressors. Potential problems are fouling of the boiler tubes, high capital costs, and interconnection with the utilities (in the case of electricity generation).

Another potential use for the heat is preheating the batch or cullet, which is normally fed cold into the furnace. This can be accomplished by direct or indirect exposure of the batch to the exhaust gases. Direct preheating can be done by directing the hot exhaust gas opposite the batch as it falls from a feeder. This can raise the batch or cullet temperature up to 400°C. Indirect preheating systems use a plate heat exchanger and may consist of blocks of horizontal waste gas flow and vertical material funnels. This can lead to heating of the cullet or batch to 300°C. The exhaust gases exiting the preheaters will be cooled by approximately 270–300°C. Potential problems are increased particulate emissions and size sorting, although this is not as important in preheating cullet.

As mentioned previously, oxy-fuel firing can be more energy efficient than air-fuel furnaces. Most glass companies have incorporated some oxy-fuel firing into their operations, either as oxygen-enriched air or as 100% oxygen. A furnace that is equipped with oxygen-enriched air can produce the same amount of glass as with air combustion but at lower fuel input because of the reduced exhaust volume. However, when regenerative furnaces are converted to oxy-fuel firing, the heat-recovery checkers are eliminated. In addition, oxygen is expensive and to make oxy-fuel systems economical, a cost-effective and reliable source of oxygen is required. Several on-site systems can be used to fractionate oxygen from ambient air. Liquified oxygen can be used, but its use is limited by transportation costs. The energy required to use oxygen ranges from 250 to 800 kWh per ton of oxygen.

4.4 Cullet and Energy

Cullet has a lower melting energy requirement than the virgin raw materials because the heat of fusion of converting crystalline material to a liquid is not needed and its mass is 20% lower than that of the equivalent batch materials. General energy savings are 0.15–0.30% for each percent of cullet added to the batch. For example, a melter that uses 50% cullet will use 7.5–15% less energy than a batch with no cullet.

Most plants routinely recycle all internal cullet—that is, waste glass produced within the production process. For flat glass production, approximately 10–20% of the produced glass returns as internal cullet due to edge trimming, defects, product changes, and breakages. External sources of cullet (from consumer or external industrial sources) are also used, with the amount used depending on the product, specifications, availability, and price. The composition of external cullet can vary and the presence of impurities such as metal and ceramics may limit its use. Stringent final product quality requirements can restrict the amount of foreign cullet a manufacturer can use. Container glass manufacturing benefits from bottle recycling schemes and, depending on the

country, this sector can consistently use 80% cullet in the batch. It is more difficult for flat glass manufacturers to find external cullet, although they are able to obtain scrap from the industries that purchase their product, such as architectural and automotive window manufacturers. All manufacturers keep a store of cullet on site in the event that the raw material batching system is offline and they need to melt at 100% cullet.

4.5 Advanced Melters

The next generation of glass melters needs to be more energy efficient than current melters, but they also must operate as rapidly and economically to produce the highest quality product. Currently, the most efficient glass melters are the large regenerative furnaces, but they are also the most capital intensive, have high rebuild costs, and are very inflexible in terms of varying or reducing output.

There are potential technologies that can make melting more efficient by making heat transfer to the batch more efficient. Submerged combustion melting uses natural gas and an oxidizer injected directly into the melt, which improves heat transfer and promotes rapid mixing. Plasma, which is a partially ionized gas that conducts electricity, can be very efficient at transferring heat directly to batch materials, but it is difficult to apply on a large scale. Microwaves can heat without direct contact, but different components of the batch have different susceptibilities to heating by microwaves.

One obvious approach to reducing the energy needs of the commercial glass industry is to reduce the residence time of the melt in the furnace. Rapid glass melting systems would allow the raw materials to be melted and homogenized without the need for the relatively slow, convective mixing that is currently practiced. The industry would like the current average 24- to 48-h residence time of a melt reduced to 1 h.

Combined melting and mixing can be achieved using a variety of mechanisms. Higher temperatures reduce viscosity and decrease mixing times, but the reduced time must be balanced by the increased energy input. Higher temperatures also adversely affect the refractory lining and can increase emissions of NO_x and volatile components of the batch. Improved mechanical mixing will decrease homogenization times. Decreased fining times can be achieved by controlling the atmosphere surrounding the melt or applying a vacuum.

There are several conceptual designs of fully reengineered methods of melting glass. Many of these segment the furnace into distinct sections for melting, fining, and homogenization. One segmented melting concept takes advantage of the different melting requirements of virgin batch materials and cullet. Virgin batch is charged into an all-electric premelting furnace that converts 75% of the raw material into glass. The premelted batch then moves into an enlarged doghouse where cullet is added (cullet comprises at least 60% of the raw material). The batch/cullet mix then enters the second melting chamber that uses oxy-fuel burners. The potential benefits are up to 25% improved energy efficiency and lower emissions. Another concept under development, referred to as the advanced glass melter, injects all batch materials into the flame of a natural gas-fired furnace, heating them rapidly and discharging them into the melt chamber. Because of lower flame temperatures, this system has potential for low NO_x emissions.

Acknowledgments

C.W.S. thanks Jim Shelby (Alfred University), Tony Longobardo (Guardian Industries), and Ernst Worrell (Lawrence Berkeley National Laboratory) for their assistance and reviews.

SEE ALSO THE FOLLOWING ARTICLES

Aluminum Production and Energy • Cement and Energy • Industrial Energy Use, Status and Trends • Plastics Production and Energy • Steel Production and Energy

Further Reading

Glass Manufacturing Industry Council (2001). "Glass Melting Technologies of the Future." Glass Manufacturing Industry Council, Westerville, OH.

Integrated Pollution Prevention Control (2000). "Reference Document on Best Available Practices in the Glass Manufacturing Industry." European Commission, Institute for Prospective Technological Studies, Seville.

Ruth, M., and Dell'Anno, P. (1997). An industrial ecology of the U.S. glass industry. *Resour. Policy* **23**, 109–124.

Shelby, J. E. (1997). "Introduction to Glass Science and Technology." Royal Society of Chemistry, Cambridge, UK.

U.S. Department of Energy (2002). "Energy and Environmental Profile of the U.S. Glass Industry." U.S. Department of Energy, Washington, DC.

Varshneya, A. K. (1994). "Fundamentals of Inorganic Glasses." Academic Press, San Diego.

Woolley, F. E. (1991). Melting/fining. *In* "Engineered Materials Handbook: Ceramics and Glasses," Vol. 4, pp. 386–393. ASM International, Materials Park, OH.

Global Energy Use: Status and Trends

LINDA E. DOMAN
U.S. Department of Energy, Energy Information Administration
Washington, D.C., United States

1. **Current Issues in Energy Demand**
2. **Today's Energy Use**
3. **Resources and Reserves**
4. **Drivers of Energy Demand**
5. **Regional Energy Use**
6. **Summary**

Glossary

American Petroleum Institute (API) gravity A measure of specific gravity (in degrees) of crude oil or condensate. An arbitrary scale expressing the gravity or density of liquid petroleum products.

barrel A unit of volume equal to 42 U.S. gallons.

British thermal unit The quantity of heat required to raise the temperature of 1 lb of liquid water by 1°F at the temperature at which water has its greatest density (approximately 39°F).

fossil fuel An energy source formed in the earth's crust from decayed organic material. The common fossil fuels are petroleum, coal, and natural gas.

gas to liquids (GTL) A process that combines the carbon and hydrogen elements in natural gas molecules to make synthetic liquid petroleum products, such as diesel fuel.

generator capacity The maximum output, commonly expressed in megawatts, that generating equipment can supply to the system load, adjusted for ambient conditions.

liquefied natural gas (LNG) The liquefied product derived from natural gas (primarily methane) that has been reduced to a temperature of −260°F at atmospheric pressure.

megawatt One million watts of electricity.

oil resources The conventional oil resource base is defined by three categories: remaining reserves (oil that has been discovered but not produced); reserve growth (increases in reserves resulting mainly from technological factors that enhance a field's recovery rate); and undiscovered (oil that remains to be found through exploration).

pig iron The crude, high-carbon iron produced by reduction of iron ore in a blast furnace.

proved energy reserves The estimated quantities of energy sources that analysis of geologic and engineering data demonstrates with reasonable certainty are recoverable under existing economic and operating conditions. The location, quantity, and grade of the energy source are usually considered to be well established in such reserves.

tar sands The naturally occurring bitumen-impregnated sands that yield mixtures of liquid hydrocarbon and that require further processing other than mechanical blending before becoming finished petroleum products.

watt The unit of electrical power equal to 1 amp under a pressure of 1 V; 1 W is equal to 1/746 horsepower.

watt-hour The electrical energy unit of measure equal to 1 W of power supplied to, or taken from, an electric circuit steadily for 1 hr.

World energy use has increased steadily over the past several decades. Much of the growth in global energy consumption has been concentrated on the use of fossil fuels (oil, natural gas, and coal). This trend is expected to continue over the foreseeable future. Industrially mature nations will continue to rely on fossil fuels to meet their energy needs for all end uses, but the greatest rate of energy use is projected to occur in the emerging economies of the developing world. In these developing nations, fossil fuel use will be needed to feed growing industrial machinery, particularly the energy-intensive energy end uses that are needed to build industrial infrastructure, such as cement and steel-making, paper and pulp, and other energy-intensive processes. In this article, the trends of worldwide energy use by region are discussed. Also considered are the drivers of increased energy use, and the ramifications for future energy demand. These include the impact of increased energy demand on the environment and the role that advancing technologies may play on the world's future energy mix.

1. CURRENT ISSUES IN ENERGY DEMAND

Non-fossil-fuel energy sources (i.e., nuclear power and renewable energy sources) face difficulties in competing with fossil fuels. In the case of nuclear power generation, the industry has encountered political, economic, and social difficulties in many parts of the world, beginning with the nuclear accidents at Three Mile Island in the United States in 1979 and at Chernobyl in the Soviet Union in 1986. There is strong public sentiment against nuclear power in much of the industrialized world, based on concerns about plant safety, radioactive waste disposal, and the proliferation of nuclear weapons. Some growth in nuclear power may be expected among the developing countries of the world, particularly in developing Asia, where 17 of the 35 nuclear power reactors currently under construction worldwide may be found. However, the industrialized world is by and large expected to see a decline in nuclear power generation over the long term, as aging reactors are retired and are not replaced by new nuclear generators.

Renewable energy sources enjoy a great measure of popularity with the public—both because these sources of energy are cleaner burning compared to other energy sources and because they are not subject to resource depletion, perceived to be a problem for fossil fuels. However, renewable energy sources cannot yet compete economically with fossil fuel resources. With fossil fuel prices likely to remain moderate over the next quarter century, renewable energy sources will probably fail to gain market share relative to the fossil fuels in a business-as-usual scenario. In the absence of significant government policies aimed at reducing the impacts of carbon-emitting energy sources on the environment, it will be difficult to encourage a wide-scale expansion of renewable resources.

It should also be noted, however, that the renewable energy forecasts presented here deal only with commercial renewable energy sources, i.e., those projects connected to national electricity grids. The Paris-based International Energy Agency (IEA) has estimated that noncommercial, traditional biomass resources are the primary sources of heating and cooking for 2.4 billion people in the developing world. Unfortunately, sources of comprehensive data on the use of noncommercial fuels are not readily available on a consistent, worldwide basis, and so cannot be included in projections of global energy consumption. The IEA expects the developing populations of the world to begin to switch away from traditional fuels to commercial energy sources, but the reliance on the noncommercial fuels will clearly remain strong in the mid-term future given the high levels of current use.

Another trend that is expected to continue is a move toward improving energy intensity. As time goes by and energy technologies become more efficient and widespread, energy use becomes more and more productive. That is, by employing more efficient technologies, less energy is needed to produce the same amount of output. One part of the world where this is particularly obvious is among the nations of the former Soviet Union. Rich with fossil fuel resources and mired in dysfunctional political and economic systems, Russia and the Soviet satellite nations used energy very inefficiently for much of the past half century. With the fall of the Soviet Union in the early 1990s, these economies began to transition into a market-based economic system, and already the old, Soviet-era machinery and technologies are being replaced by more energy-efficient capital stock. This has meant an improvement in the region's energy intensity since the mid-1990s. As these trends continue, the transitional economies of Eastern Europe and the former Soviet Union will begin to see patterns of energy intensity that resemble those of the industrialized world.

The real growth in energy use is projected to occur in the developing economies of the world. Economic expansion is being accompanied by consumer demand for energy-consuming products. Growing demand for energy to fuel industrial end uses, as well as energy demand in the residential, commercial, and especially transportation and electric utility sectors, will have profound implications for the world's geopolitical relationships. There is tremendous opportunity for large increments of energy consumption in these economies, as nations strive to advance standards of living to match those of the industrialized world. In particular, many nations that are emerging economically will see consumer demand increasing for personal motor vehicles and a wide assortment of home appliances, including televisions and personal computers. National electrification has become an important goal of many of these nations, to improve both the productivity and the standards of living of their populations.

2. TODAY'S ENERGY USE

In the year 2001, 404 quadrillion British thermal units (Btu) of energy were consumed worldwide.

Petroleum products accounted for the largest share (39%) of the world's total energy consumption (Fig. 1). Oil products are particularly important for the transportation sector, and this will probably be true for the next several decades. The world's transportation infrastructure relies heavily on the combustion engine, which, in turn, largely relies on petroleum to operate. Currently, there are no truly commercially competitive alternatives to petroleum products in the transportation sector. Short of a substantial technological advance—such as hydrogen-fueled vehicles—that would offer competition to traditional petroleum-product-fueled automobiles, trucks, and airplanes, it will be difficult for any other fuel type to replace oil in this sector.

Although oil is predominant in the transportation sector, alternative fuels have displaced petroleum product use in other end-use sectors. In many parts of the world, for example, other fuels are increasingly used to fuel electricity generation and the importance of oil in this end-use sector has been decreasing over the past several decades. This trend is expected to continue into the future. Natural gas and nuclear power have been two energy sources that have taken market share away from oil in the electric utility sector, especially since the 1970s. Nuclear power generation advanced strongly in the 1970s. If it were not for the public concerns about the safety of nuclear power, because of accidents at Three Mile Island in 1979 (United States) and at the then-Soviet Chernobyl in 1986, nuclear generation may well have continued to experience significant growth and might have gained substantial market share among the electricity-generating energy sources.

Natural gas is now seen as the fuel of choice for new electricity generation worldwide. Combined cycle gas turbines have been demonstrated to be very efficient electricity generators. What is more, natural gas is widely regarded as a more cleanly burning fossil fuel compared to either coal or oil. In an age in which environmental concerns are increasingly important to the world's governments, natural gas is a desirable alternative to coal- or oil-fired power plants, and gas does not come with the negative aspects of nuclear power, such as concerns about safety and radioactive waste disposal. Where infrastructure and resources are available, natural gas has already gained significant share of the electricity sector. For instance, natural gas supplies more than half (54% in 2000) of the electricity generation in Russia, which, not coincidentally, possesses 31% of the world's total natural gas proved reserves.

Coal is the predominant source of electricity generation in many countries. In the United States, coal accounts for 52% of the country's total electricity-generating capacity. Although it is a relatively "dirty" form of energy (the combustion of coal emits more carbon dioxide than the same amount of natural gas and of most petroleum products), reserves of coal are also fairly widespread across the world and are plentiful. By some estimates, at current production levels, the world has enough coal to last for another 200 years or more. In two of the world's fastest growing developing economies, China and India, coal is an extremely important energy resource. These two countries alone are expected to account for more than 85% of the world increment in coal use over the next two and one-half decades (Fig. 2). Although incremental coal use over most of the world is expected to be for the generation of electricity, in China coal is also an important resource for

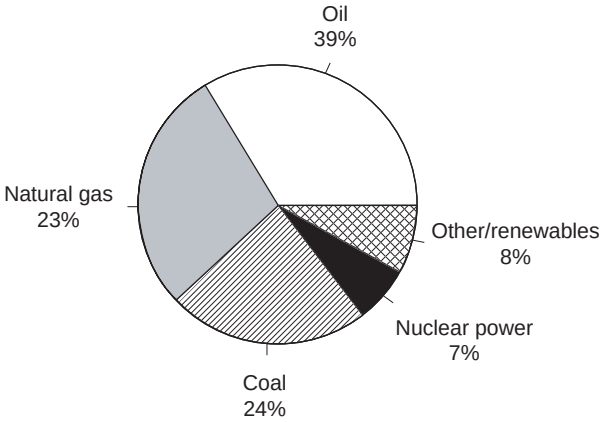


FIGURE 1 World energy use by fuel type, 2001. Data from Energy Information Administration (2003).

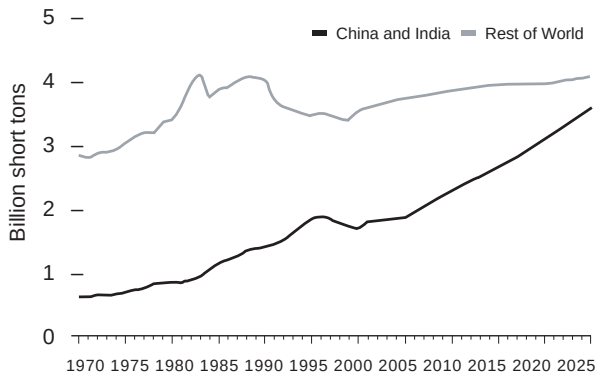


FIGURE 2 World coal use, 1970–2025. Data from Energy Information Administration (2003).

industrial purposes. China is the world's largest producer of both steel and pig iron. The country also uses coal for steam and direct heat for industrial applications, primarily in the chemical, cement, and paper and pulp industries. This heavy reliance on coal may lessen somewhat over the next two decades, but because coal is such a plentiful resource in China, it is difficult to envision a time when its importance will be eclipsed by any other currently available energy source.

As previously noted, nuclear generation showed great promise in the 1970s. Between 1970 and 1986, nuclear generation increased by an estimated 14% per year (Fig. 3). However, the nuclear accidents at Three Mile Island in 1979 and Chernobyl in 1986, in conjunction with the high capital costs associated with bringing nuclear generation into operation relative to coal- or natural-gas-fired generation, slowed the development of nuclear power. Despite problems associated with the disposal of highly radioactive waste, many analysts believed nuclear power would become an increasingly important source of the world's electric power mix. However, public wariness about the safety of nuclear power largely halted the growth of this sector in the 1990s, especially among the industrialized nations and the nations of Eastern Europe and the former Soviet Union. In fact, no orders for new nuclear generators have occurred in Austria, Hungary, Italy, Mexico, the Netherlands, Switzerland, or the United States since 1973. The only industrialized countries that still have plans to expand nuclear generation with new builds are Finland, France, and Japan.

Most of the growth in new nuclear generation is expected to occur in developing Asia (Fig. 3). Of the 35 reactors under construction in 2003, nearly half

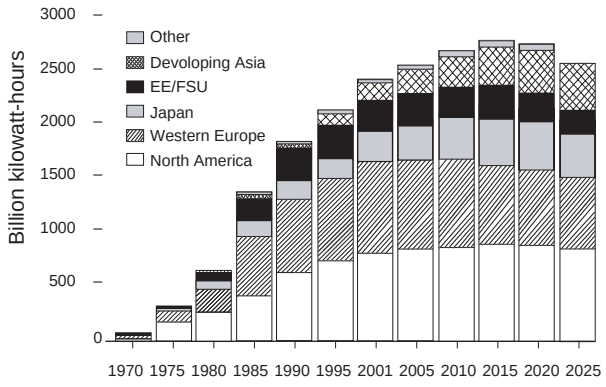


FIGURE 3 World nuclear power consumption by region, 1970–2025. EE/FSU, Eastern Europe/former Soviet Union. Data from Energy Information Administration (2003).

are located in the countries of developing Asia. This includes 8 reactors in India alone, 4 in China, 2 each in South Korea and Taiwan, and 1 in North Korea. These new reactors are expected to add some 12,000 MW of installed nuclear capacity to the existing 28,000 GW on-line in the region as of January 2003. Nuclear generation remains an attractive alternative to other forms of electricity in these countries, either because it will allow them to diversify their electricity-generating sources, such as in India and China, or because there are few indigenous resources that can be used to fuel electricity demand growth, such as in South Korea and Taiwan.

Many countries would like to expand the use of renewable energy resources. In the industrialized world, many governments view renewable energy resources (with the notable exception of large-scale hydroelectric power) as an attractive alternative to expanding electricity generation because of the lack of negative environmental impact, and renewable energy resources are often cited as a way for countries to reduce or limit emissions of carbon dioxide. In western Europe, many government incentive programs are in place to spur the growth of renewable resources, such as wind, solar, geothermal, and small hydroelectric generators, among others. Germany, Denmark, and Spain have made great strides in developing their wind generation capacity through government programs. That said, wind power and other nonhydroelectric forms of on-grid, commercial renewable energy resources remain fairly small portions of the electric generation mix of these nations.

As with any energy form, renewable energy does not come without its controversies. Sometimes populations faced with wind farms view them as eyesores, despoilers of the surrounding landscape, and there can be complaints about the noise that the whirring turbine blades can produce. Procuring site licenses has been particularly difficult in the United Kingdom. The Non Fossil Fuel Obligation—a program used in the United Kingdom to finance the development of renewable energy resources between 1989 and 2001, when it was replaced by the Renewable Obligation of the National Electricity Trading Arrangements program—approved power purchase contracts of wind capacity amounting to 2676 MW, but by the end of 2001 only 468 MW of wind capacity had been installed.

Hydroelectricity remains the largest and widest form of renewable energy worldwide. In the United States alone, hydroelectric power accounts for

between 8 and 10% of the nation's total electricity generation, depending on the amount of rainfall in any given year. Many developing nations rely quite heavily on hydro power for their electricity supplies. In South America, Brazil, Peru, and Chile use hydroelectricity for 80% or more for their electricity needs. Large-scale hydroelectric facilities are often controversial because of their impact on the environment and because it is often necessary to displace surrounding populations in order to create the reservoirs needed to supply the hydro turbines with enough water to generate needed electricity.

Because most of the hydroelectric capacity in industrialized countries has either already been exploited or is located far from the population centers that could use the hydro-generated electricity, much of the growth in hydroelectricity is expected to be in the form of large-scale hydro projects in the developing countries of the world, especially those in developing Asia. Although the average size of a thermal electricity generator (i.e., oil, coal, or natural gas) is about 300 MW, China is currently constructing an 18,200-MW hydroelectric project on the Yangtze River, the Three Gorges Dam project. In Malaysia, the government is pressing forward to construct a 2400-MW Bakun Dam project. These projects and others in Vietnam, India, and Laos, among others, have met with staunch opposition from the international community because of their potentially detrimental impact on wildlife, plant life, and the indigenous populations that will be disrupted by the dams' construction. Still, governments persist with the developments of large-scale hydro projects because of the strong growth in demand for electricity and in an effort to diversify electricity-generating energy sources.

Noncommercial biomass is often not discussed as part of the world's energy use. Although the burning of wood and other plant life to cook and heat water is considered an important component of energy use in many rural parts of the developing world, it is also difficult to quantify. The International Energy Agency has been attempting to measure the amount of biomass used worldwide since 1994. Their estimates indicate that some 2.4 billion people in the developing world still rely on traditional, noncommercial biomass to meet their heating and cooking needs. Even when people in the developing world have access to electricity to meet their lighting needs, they often still rely on biomass for other residential uses. The heavy reliance on surrounding flora is accompanied by many problems. Because developing countries typically have high population growth rates relative to the industrialized world, there is a tendency to need more and more plant life to satisfy the populace's energy needs. The result is that a locale's biomass is often overharvested, resulting in the danger of deforestation, which can degrade the environment and make sustaining life difficult, if not impossible. The Tata Energy Research Institute of India noted that many women in rural South Africa carry 20 kg of fuel wood an average of 5 km each day. The effort required to carry out this task accounts for a large share of the gatherer's daily caloric intake, and as deforestation requires the person to go farther and farther to achieve the requisite fuel wood, the task will eventually not be sustainable.

The use of biomass, particularly fuel wood, dung, and charcoal, can also negatively impact the health of the people utilizing these products. People dependent on biomass resources are consistently exposed to indoor particulate and carbon monoxide levels that are far in excess of those the World Health Organization has deemed safe. Traditional stoves that burn dung or charcoal often emit large amounts of carbon monoxide and other noxious gases, resulting with acute respiratory illnesses and a shortened life span. Although the International Energy Agency has noted that the use of biomass has been declining steadily since biomass surveys were initiated, the agency also believes that it will be 30 years or more before the world will see a substantial reduction in the reliance that developing nations' populations have on this energy source.

3. RESOURCES AND RESERVES

In the 1970s, in an environment of oil shortages caused by the oil embargo staged by the Organization of Petroleum Exporting Countries (OPEC) cartel members, many analysts were concerned with estimating the point at which oil supplies and, indeed, all fossil fuels would be depleted. Gas rationing and high world oil prices associated with this time period further fueled this concern. In fact, work done by the U.S. geologist M. King Hubbert suggested that world oil production would ramp up and peak over only a few short decades and then just as rapidly decline.

Technology has helped to increase the volume of proved oil reserves over the past few decades. Three-dimensional drilling and seismic imaging have been enormously helpful in improving the ability of the oil industry to produce heretofore unrecoverable oil. In part as a result of these enhancements, most

analysts today have a more measured view of the potential for world fossil fuel resources. In its most recent assessment of oil's long-term production potential, the U.S. Geological Survey identified at least 3 trillion barrels of ultimately recoverable conventional oil—that is, oil that can be recovered with current prices and technology—worldwide. Historically, it has been estimated that only about 25% of that oil that has been classified as "ultimately recoverable" has already been produced; rough calculations indicate that a likely peak in worldwide conventional oil production would occur at some point after 2020.

Analysts would agree that fossil fuels are a limited resource and, as such, are subject to depletion. Further, depletion would lead to scarcity and higher prices. There is, however, the matter of nonconventional resources. Nonconventional resources are defined as those resources that cannot be produced economically at current prices using current technology. This categorization shifts as the prices for these resources continue to rise, and the gap between conventional and nonconventional thus narrows. As a result, the combination of escalating prices and technological advancements will transform the nonconventional into the conventional. Much of the perceived pessimism regarding the finite quantities of resources has focused only on conventional resources as defined today. However, it was demonstrated in the 1990s that technological advances can help lower the costs of producing liquid fuels from several nonconventional sources, including heavy oils, tar sands, and natural gas.

Heavy oils generally have an American Petroleum Institute (API) gravity of less than 24° and will not flow on their own at standard reservoir temperatures. Tar sands are similar, but have more viscous oil that is located closer to the surface. More than 3.3 trillion barrels (oil in place) of heavy oil and tar sands are estimated to exist worldwide. The most significant deposits of these resources are found in Canada and Venezuela. There are two distinct ways to recover these resources. For deposits that are located close to the surface, mining can be used. For deposits located deeper under the surface, steam injection could be used to heat the oil and allow it to flow more easily, so that it behaves more like conventional resources. Although it is true that these heavy oils and tar sands will have to be cleaned and upgraded before they behave like conventional refinery feedstocks, higher prices for conventional sources would make the additional costs associated with the additional processes economical. A sustained world oil price of between $23 and $25 per barrel could make these nonconventional resources economical to develop and produce.

Significant portions of the world's natural gas resources are located in remote areas or in small deposits. Development of such sites is usually discouraged, because delivery by way of pipeline or in the form of liquefied natural gas (LNG) is not economical. However, natural gas molecules can be reconstituted as liquid synthetic petroleum products by employing an updated version of a technology that has been available since World War II, the so-called Fischer–Tropsch method. This gas-to-liquids (GTL) technology presents an attractive alternative for the small, remote deposits, because the infrastructure for petroleum products is already in place. GTL technology is also versatile enough to accommodate smaller gas deposits economically. By some estimates, GTL technology could be economically viable with sustained world oil prices of between $26 and $28 per barrel. It is possible to envision a scenario in which oil prices could climb even higher than those that would make nonconventional or GTL technologies economical. For instance, OPEC producers could adopt a conservative productive capacity strategy that might result in a doubling of prices, in real terms, over the next few decades. Sustained high world oil prices might further alter the supply side of the market even beyond those suggested by heavy oil, tar sands, and GTL. This might lead to the development of other nonconventional resources. Coal-to-liquids technologies and even shale oil could then economically be developed. Shale oil has enormous potential reserves that would dwarf those of conventional resources.

As with petroleum reserves, natural gas resources have in general increased in every year since the 1970s. As of January 1, 2003, proved natural gas reserves—that is, those quantities that can be recovered under present technologies and prices—were estimated by the *Oil & Gas Journal* at 5501 trillion cubic feet. Most of the increases in natural gas reserves in recent years have occurred in the developing world, and about 71% of the world's natural gas reserves are found in the Middle East and in the former Soviet Union (Fig. 4). The remaining reserves are spread fairly evenly among other regions of the world. Despite the high rates of increase in the use of natural gas worldwide, most regional reserves-to-production ratios have remained high. Worldwide, the reserves-to-production ratio is estimated at 62 years, but the former Soviet Union has a reserves-to-production ratio estimated at nearly 80 years; this

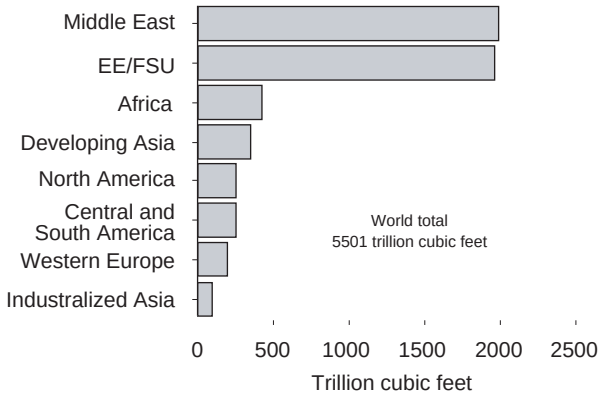


FIGURE 4 World natural gas reserves by region as of January 1, 2003. EE/FSU, Eastern Europe/former Soviet Union. Data from *Oil & Gas Journal*, Vol. 100, No. 52, December 23, 2002, pp. 114–115.

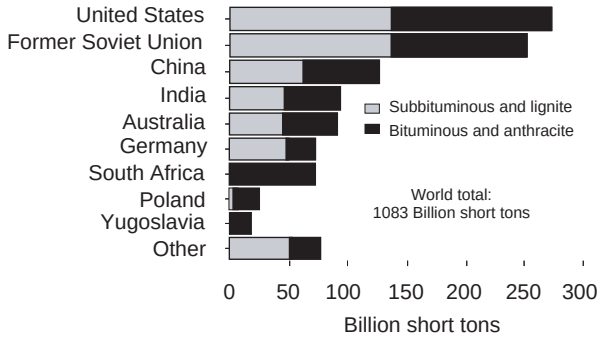


FIGURE 5 World recoverable coal reserves. Data from World Energy Council London, UK (2001) (for all countries except the United States; data for the United States provided by the Energy Information Administration); also based on unpublished data of the Coal Reserves Data Base (February 2002).

ratio in Africa and in the Middle East is about 90 years and over 100 years, respectively.

According to the U.S. Geological Survey's most recent assessment of world natural gas resources, there is a significant amount of natural gas that remains to be discovered. The mean estimate for worldwide undiscovered natural gas is 4839 trillion cubic feet. Of the natural gas resources that are expected to be added over the next 25 years, reserve growth accounts for 2347 trillion cubic feet. It is estimated that one-fourth of the undiscovered natural gas is located in undiscovered oil wells. Thus, it should come as no surprise that more than one-half of the mean undiscovered natural gas is expected to come from the Middle East, former Soviet Union, and North Africa. Although the United States has produced more than 40% of its total estimated natural gas endowment and carries less than 10% as remaining reserves, in the rest of the world reserves have been largely unexploited. Outside the United States, the world has so far produced less than 10% of its total estimated natural gas endowment and carries more than 30% as remaining reserves.

World coal reserves are also considered to be quite plentiful. Total recoverable reserves of coal are those quantities of coal that geological and engineering data indicate with reasonable certainty can be extracted in the future under existing economic and operating conditions; around the world, total recoverable reserves are estimated at 1083 billion short tons. At current consumption levels, there are still 210 years worth of sufficient coal reserves. Although coal deposits are widely distributed, 60% of the world's recoverable reserves is located in three regions; the United States holds an estimated 25% of world reserves, the former Soviet Union holds 23%, and China holds 12%. Another four countries, Australia, India, Germany, and South Africa, collectively account for another 29% (Fig. 5).

4. DRIVERS OF ENERGY DEMAND

There are several important drivers associated with the growth in energy demand. One driver, a nation's income, which can be measured by gross domestic product (GDP), can indicate how much energy a given nation will require to support the expanding economy. Potential growth in population will influence the demand for energy to support an expanding population. Another factor is the rate of efficiency gains a country makes in its energy consumption. Other factors, such as environmental policies, taxes, and subsidies, and other influences can also have an impact on the development of energy markets from country to country. An examination of these factors and how they are manifested in various parts of the world can help to explain the regional trends in energy consumption.

Although economic growth is certainly a key factor to determining the growth in energy demand, how energy is used relative to income can be equally important. Economic growth and energy demand are linked and the strength of that link varies among regions and their stages of economic development. In the industrialized world, history has shown the link to be a relatively weak one, with energy demand lagging behind economic growth. In developing countries, demand and economic growth have been

closely correlated in the past, with demand growth tending to track the rate of economic expansion.

The historical behavior of energy intensity in the former Soviet Union (FSU) is problematic. Since World War II, the levels of energy intensity in the Soviet Republic economies have been higher than they have been in either the industrialized or the developing nations (Fig. 6). In the FSU, however, energy consumption grew more quickly than GDP until 1990, when the collapse of the Soviet Union created a situation in which both income and energy use were declining, but GDP fell more quickly; as a result, energy intensity increased. As the nations of the FSU continue to recover from the economic and social problems of the 1990s, energy intensity should begin to decline once again, as new, more efficient capital stock replaces inefficient Soviet-era equipment and GDP continues to expand.

The stage of economic development and the standard of living of individuals in a given region strongly influence the link between economic growth and energy demand. Advanced economies with high standards of living have relatively high per capita energy use, but they also tend to be economies in which per capita energy use is stable or changes very slowly, and increases in energy use tend to correlate with employment and population growth. In the industrialized world, there is a high penetration rate of modern appliances and motorized personal transportation equipment. To the extent that spending is directed to energy-consuming goods, it involves more often than not purchases of new equipment to replace old capital stock. The new stock tends to be more efficient than the equipment it replaces, resulting in a weaker link between income and energy demand. In the developing world, standards of living, while rising, tend to be low relative to those in more industrially advanced economies.

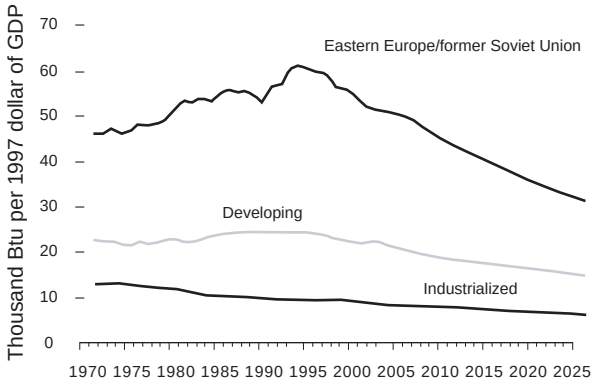


FIGURE 6 World energy intensity by region, 1970–2025. GDP, Gross domestic product. Data from Energy Information Administration (2003).

It is worth noting that, although economic growth and energy intensity impact the increases in energy demand, other factors may also influence the development of energy use. Individual countries may employ many devices to encourage consumer behavior. For instance, in the industrialized countries of western Europe, taxes are often levied in an effort to get consumers to use less energy. Motor gasoline and diesel fuel are heavily taxed in many western European countries. The government tries to use this disincentive to consume to lower congestion in urban areas and to lower the impact of motor vehicle emissions on the environment. In many industrialized countries, tax incentives, such as the wind production tax credit in the United States, are used to encourage consumers to use a more environmental, but not yet economically competitive, form of energy, in this case, wind. In the developing country, price subsidies are sometimes employed to make it possible for a population to increase the penetration of modern technologies. Many oil-producing countries, e.g., Indonesia and Nigeria, heavily subsidize their motor fuels to support fledgling motor vehicle ownership.

5. REGIONAL ENERGY USE

The current mix of the world's energy markets, the drivers associated with estimating increases in energy consumption, and the potential resources available to meet future demand have been considered; the regional trends in world energy use are now assessed.

At present, the industrialized world is responsible for the lion's share of the global energy use. The United States alone currently accounts for 25% of the total world energy use. In contrast, China, a country with a population of about four times that of the United States, presently consumes only 8% of the world's total energy production. Energy use among the industrialized countries is expected to grow, though more slowly in the future than it has over the past several decades. The industrialized countries will probably maintain the largest regional share of world energy consumption (Fig. 7). However, the growth in consumption is expected to be low relative to other parts of the world. Many industrialized nations—notably Japan and many of the countries of western Europe—are faced with slow population growth rates and thus energy markets will not need to accommodate a great deal of new populace. Further, most, if not all, of the energy end-use sectors

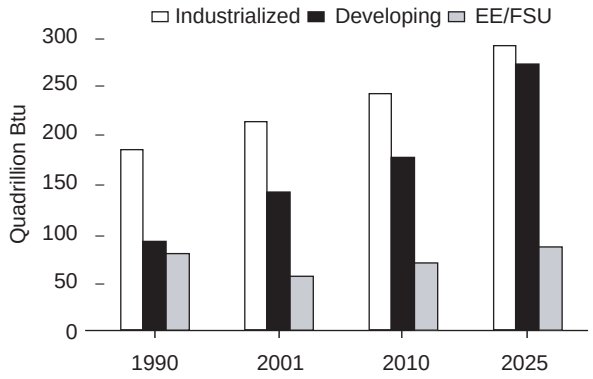


FIGURE 7 World energy consumption by region, 1990–2025. EE/FSU, Eastern Europe/former Soviet Union. Data from Energy Information Administration (2003).

TABLE I

Projected Growth in World Population by Region, 2001–2025[a]

Region	Average annual increase (%)
North America	0.9
Western Europe	−0.1
Industrialized Asia	0.2
Eastern Europe/former Soviet Union	−0.2
Developing Asia	1.1
Middle East	2.0
Africa	2.3
Central and South America	1.2
Total worldwide	**1.1**

[a]Data from United Nations (2001).

will not see large increments in energy demand. Residential and industrial consumers already possess well-established infrastructures, so most of the future energy demand will utilize more efficient equipment.

In comparison to the industrialized nations, there is enormous potential energy demand from burgeoning populations in the developing world. Africa, the Middle East, Central and South America, and developing Asia are all expected to experience average annual population increases of over 1.0% per year between 2001 and 2025 (Table I), with many individual countries among these regions seeing population growth in excess of 2% per year. A corresponding increment in energy consumption will be needed to ensure the prosperity of these economies. As these developing economies continue to emerge, there will be increasing consumer demand for electricity to run modern appliances such as refrigerators, stoves, ovens, washers, dryers, and even personal computers, to name a few, and also for personal motor vehicles. Motor vehicle ownership is considered to be among the first symbols of emerging prosperity in many developing countries, and as the industrial base expands, so too will the need to move goods and services to market.

The fastest growing energy markets will likely be centered in developing Asia. In the Energy Information Administration's *International Energy Outlook 2003*, economic growth in this region is projected to increase 5.1% per year between 2001 and 2025. This increment in the region's income will require a corresponding increase in energy use. The Energy Information Administration (EIA) expects developing Asian energy demand to grow by 3.0% per year over this time period. Many of the nations of this region have a relatively immature infrastructure. In China, for instance, there is great potential for the transportation sector to expand. In 1999, there were an estimated 12 motor vehicles per 1000 persons, compared to the 777 vehicles per 1000 persons recorded for the United States. Although a 2002 analysis by the EIA suggests that motorization levels in China might grow to as many as 52 vehicles per 1000 persons by 2020, China would still only have levels that are 6% that of the current U.S. motorization levels.

Growing transportation energy use could have a profound impact on the demand for world oil. With no economic alternatives to oil presently available, the pressures on world oil supplies will be great. In the case of China, should transportation energy use grow at a robust rate of, say, 5.0% per year between 2001 and 2025, an increment of 4.4 million barrels of oil per day would be required to fuel the Chinese transportation market. This is substantial, considering that total oil consumption in China is projected to grow by 5.9 million barrels per day over this time period. The geopolitical ramifications of such growth in demand could be significant. China became a net importer of oil in 1993, and in 2001 the country was importing 2.0 million barrels per day of oil, with 0.9 million barrels per day coming from the OPEC Persian Gulf. With the demand increases projected by the EIA, China's oil import dependence on the OPEC nations is apt to increase dramatically over the next two and one-half decades. In the EIA reference case, China will import 7.8 million barrels per day by 2025, most of which is expected to come from the Persian Gulf nations.

Energy demand could grow even more dramatically than analysts assume if developing markets

manage to reach the consuming patterns of the industrialized countries more quickly than conventional wisdom suggests. This is entirely possible. Consider the case of South Korea, where energy consumption grew rapidly over the past three decades, by nearly 12% per year. South Korea embraced a market-based economy and has made tremendous strides in becoming an industrialized nation. Standards of living have grown substantially. In 1970, South Korea had a per capita GDP of about $2000 per person (in real 1997 U.S. dollars), but by 2001 this had climbed to $11,908 per year. Per capita energy use correspondingly increased, from 21 million Btu per person in 1970 to 172 million Btu in 2001. In contrast, China's per capita energy use was 31 million Btu in 2001. Should China's per person energy use achieve the current South Korea per capita energy use levels by 2025, another 163 quadrillion Btu of energy (or 80% more than projected for China in 2025 according to the EIA reference case forecast) would be needed to meet the additional demand. A similar paradigm could be made for many other developing countries.

The potential for new energy demand is also fairly high in the transitional economies of the former Soviet Union. Although energy demand plummeted in the early 1990s after the dissolution of the Soviet regime, by the late 1990s and into the 21st century, the economies of these nations began to recover. The growth in energy use is not expected to be as dramatic as that in the developing world, mostly because the rate of population growth is set to decline or will remain slow in most of these nations, and because energy, in the past, has been used very inefficiently. As these economies transition from planned economies to more market-based economies and as foreign investment in the private sectors begins to take hold, energy will be used more efficiently and with new, less energy-intensive equipment.

6. SUMMARY

World energy use is poised to increase substantially over the next few decades. How energy markets develop and what kinds of political, economic, or social changes might influence these changes are impossible to determine with any degree of certainty. Current understanding of fossil fuel resources suggests that there will be ample supplies of fossil fuels available to meet the growing demand for the foreseeable future, at relatively low consumer costs. It is also true that government intervention to encourage conservation or to support the consumption of nonfossil alternatives may decrease the need to exploit all of these resources. Further, there is an opportunity for technological innovation that may advance the use of other energy forms in the future, but this cannot yet be predicted. After all, the share of world energy use of nuclear power generation grew from less than 0.5% in 1970 to almost 7% in 1999. A similar advancement, say in automobiles that could economically run on hydrogen, could profoundly change the world reliance on petroleum products, albeit the penetration of such an advancement might require more than only a few decades without some incentive to make such a dramatic change to the infrastructure.

SEE ALSO THE FOLLOWING ARTICLES

Cultural Evolution and Energy • Development and Energy, Overview • Earth's Energy Balance • Economic Geography of Energy • Economic Growth, Liberalization, and the Environment • Energy Ladder in Developing Nations • Geopolitics of Energy • Nationalism and Oil • Population Growth and Energy • Sustainable Development: Basic Concepts and Application to Energy • Technology Innovation and Energy • Transitions in Energy Use

Further Reading

American Automobile Manufacturers Association. (1997). "World Motor Vehicle Data." AAMA, Detroit, Michigan.

Department of Trade and Industry. (2001). "New and Renewable Energy: Prospects for the 21st Century." Web site of the government of the United Kingdom of Great Britain and Northern Ireland, available at http://www.dti.gov/uk/renewable/consultations.htm.

Energy Information Administration. (1999). "International Energy Outlook 1999," DOE/EIA-0484(99). U.S. Department of Energy, Washington, D.C.

Energy Information Administration. (2002). "International Energy Outlook 2002," DOE/EIA-0484(2002). U.S. Department of Energy, Washington, D.C.

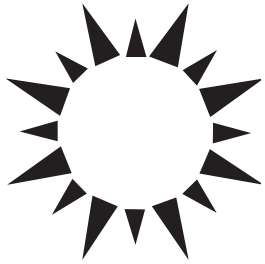
Energy Information Administration. (2003). "International Energy Outlook 2003," DOE/EIA-0484(2003). U.S. Department of Energy, Washington, D.C.

Energy Information Administration. (2003). "Annual Energy Outlook 2003," DOE/EIA-0383(2003). U.S. Department of Energy, Washington, D.C.

International Atomic Energy Agency. (2003). "Reference Data Series 2, Power Reactor Information System." Available on the Internet at http://www.iaea.org/.

International Energy Agency. (2002). "World Energy Outlook 2002." International Energy Agency, Paris, France.
International Energy Agency and PT Communications. (2002). "IEA Wind Energy Annual Report 2001." International Energy Agency and PT Communications, Boulder, Colorado.
National Energy Board. (2000). "Canada's Oil Sands: A Supply and Market Outlook to 2015." National Energy Board of Canada, Calgary, Alberta.
Porter, E. (1995). "Are We Running Out of Oil?" American Petroleum Institute Discussion Paper No. 081. American Petroleum Institute, Washington, D.C.
United Nations. (2001). "World Population Prospects: The 2000 Revision, Volume 1, Comprehensive Tables." United Nations, New York.
U.S. Geological Survey. (2000). "World Petroleum Assessment 2000." Web site of the U.S. Geological Survey. Available at http://greenwood.cr.usgs.gov/.
World Energy Council. (2001). "Survey of Energy Resources 2001." World Energy Council, London.
Worldwide look at reserves and production. (2002). *Oil Gas J.* **100**(52), 114–115.

Global Material Cycles and Energy



VACLAV SMIL
University of Manitoba
Winnipeg, Manitoba, Canada

1. Material Cycles
2. Geotectonic Cycle
3. Water Cycle
4. Carbon Cycle
5. Nitrogen Cycle
6. Sulfur Cycle

Glossary

carbon The sixth element of the periodical table and the key structural element of all living matter.
geotectonics The process of reforming the earth's surface that is driven primarily by creation and subduction of the ocean floor.
nitrogen The seventh element of the periodical table and an indispensable ingredient of proteins, the principal metabolic vehicles of life.
sulfur The 16th element of the periodical table whose presence in disulfide bonds makes the proteins three-dimensional.
water A triatomic molecule of hydrogen and oxygen with many unique physical and chemical properties that makes life possible.

The earth is a constant recipient of an enormous flux of solar radiation as well as considerable infall of the cosmic debris that also arrives periodically in huge, catastrophe-inducing encounters with other space bodies, particularly with massive comets and asteroids. Nothing else comes in, and the planet's relatively powerful gravity prevents anything substantial from leaving. Consequently, life on the earth is predicated on incessant cycling of water and materials that are needed to assemble living bodies as well as to provide suitable environmental conditions for their evolution. These cycles are energized by two distinct sources: Whereas the water cycle and the circulations of carbon, nitrogen, and sulfur, the three key life-building elements, are powered by solar radiation, the planet's grand tectonic cycle is driven by the earth's heat. For hundreds of thousands of years, our species had no discernible large-scale effect on these natural processes, but during the 20th century anthropogenic interferences in global biospheric cycles became a matter of scientific concern and an increasingly important, and controversial, topic of public policy debates.

1. MATERIAL CYCLES

Two massive material cycles would be taking place even on a lifeless Earth: (i) a slow formation and recycling of the planet's thin crust and (ii) evaporation, condensation, and runoff of water. As weathering breaks down exposed rocks, their constituent minerals are transported by wind in terrigenic dust and in flowing water as ionic solutions or suspended matter. Their journey eventually ends in the ocean, and the deposited materials are either drawn into the mantle to reemerge in new configurations 10^7–10^8 years later along the ocean ridges or in volcanic hot spots or they resurface as a result of tectonic movements as new mountain ranges. This grand sedimentary–tectonic sequence recycles not just all trace metals that are indispensable in small amounts for both auto- and heterotrophic metabolism but also phosphorus, one of the three essential plant macronutrients.

On human timescale (10^1 years), we can directly observe only a minuscule segment of the geotectonic weathering cycle and measure its rates mostly as one-way oceanward fluxes of water-borne compounds. Human actions are of no consequence as far as the

 © 2004 Elsevier Inc. All rights reserved.

geotectonic processes are concerned, but they have greatly accelerated the rate of weathering due to deforestation, overgrazing of pastures, and improper crop cultivation. Incredibly, the annual rate of this unintended earth movement, 50–80 billion metric tons (Gt), is now considerably higher than the aggregate rate of global weathering before the rise of agriculture (no more than 30 Gt/year).

In addition, we act as intentional geomorphic agents as we extract, displace, and process crustal minerals, mostly sand, stone, and fossil fuels. The aggregate mass moved annually by these activities is also larger than the natural preagricultural weathering rate. Our activities have also profoundly affected both the quantity and the quality of water cycled on local and regional scales, but they cannot directly change the global water flux driven predominantly by the evaporation from the ocean. However, a major consequence of a rapid anthropogenic global warming would be an eventual intensification of the water cycle.

In contrast to these two solely, or overwhelmingly, inanimate cycles, global circulations of carbon, nitrogen, and sulfur are to a large extent driven or mediated by living organisms, largely by bacteria, archaea, and plants. Another characteristic that these cycles have in common is their relatively modest claim on the planet's energy flows. Photosynthesis, carbon cycle's key terrestrial flux, requires only a tiny share of incident solar radiation, and the cycling of nitrogen and sulfur involves much smaller masses of materials and hence these two cycles need even less energy for their operation: Their indispensability for Earth's life is in renewing the availability of elements that impart specific qualities to living molecules rather than in moving large quantities of materials. Similarly, only a small fraction of the planet's geothermal heat is needed to power the processes of global geotectonics.

Although the basic energetics of all these cycles is fairly well-known, numerous particulars remain elusive, including not only specific details, whose elaboration would not change the fundamental understanding of how these cycles operate, but also major uncertainties regarding the extent and consequences of human interference in the flows of the three doubly mobile elements (carbon, nitrogen, and sulfur), whose compounds are transported while dissolved in water as well as relatively long-lived atmospheric gases. For millennia these interventions were discernible only on local and regional scales, but during the 20th century their impacts increased to such an extent that the effects of these interferences now present some historically unprecedented challenges even in global terms.

2. GEOTECTONIC CYCLE

The earth's surface is composed of rigid oceanic and continental plates. The former are relatively thin (mostly 5–7 km), short-lived [mostly less than 140 million years (Ma)], and highly mobile (up to 10–20 cm/year). The latter ones are thicker (only 35–40 km), long-lived (10^9 years), and some, most notably Asian and African plates, are virtually stationary. Plates ride on the nearly 3000-km-thick solid but flowing mantle, which is the source of hot magma whose outpourings create new seafloor along approximately 72,000 km of the earth-encircling ocean ridge system. Old seafloor is eventually recycled back into the mantle in subduction zones marked by deep ocean trenches (Fig. 1). There are substantial uncertainties concerning the actual movements of the subducted plates within the mantle: The principal debate is between the two-layer models and the whole-mantle flows.

Today's oceans and continents are thus transitory features produced by incessant geotectonic processes that are energized by three sources of the earth's heat: energy conducted through the lithosphere from the underlying hot mantle, radiogenic decay of heat-producing crustal elements, and convective transport by magmas and fluids during orogenic events. The relative importance of these heat sources remains in dispute, but the heat flow at the earth's surface can be measured with high accuracy. Its total is approximately 44 trillion watts (TW), prorating to approximately 85 mW/m^2 of the earth's surface. The average global heat flux of approximately 85 mW/m^2 is equal

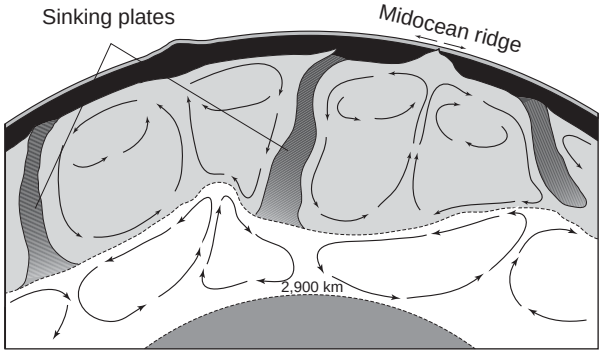


FIGURE 1 A cross section of the earth shows tectonic plates, an ocean ridge, a subduction trench, and, in this model, the circulation in mantle that is separated into two cells.

to a mere 0.05% of solar radiation absorbed by surfaces (168 W/m^2), but acting over huge areas and across long timespans, this flow is responsible not only for creating new ocean floor, reshaping continents, and building enormous mountain chains but also for energizing earthquakes and volcanic eruptions.

Mean oceanic heat flow is approximately 100 mW/m^2, whereas the continental flow is only approximately half as high. The ocean floor transmits approximately 70% of the earth's total geothermal flux and nearly one-third of the oceanic heat loss occurs in the South Pacific, where the spreading rates, particularly along the Nazca Ridge, are faster than anywhere else on Earth. By far the highest heat flows are along midocean ridges, where approximately 3–3.5 km^2 of new ocean floor is created by hot magma rising from the mantle. The total heat flux at the ridges, composed of the latent heat of crystallization of newly formed ocean crust and of the heat of cooling from magmatic temperatures (approximately 1200°C) to hydrothermal temperatures (approximately 350°C), is between 2 and 4 TW.

Divergent spreading of oceanic plates eventually ends at subduction zones, where the crust and the uppermost part of the mantle are recycled deep into the mantle to reappear through igneous process along the ridges. Every new cycle begins with the breakup of a supercontinent, a process that typically takes approximately 200 Ma, and is followed first by slab avalanches and then by rising mantle plumes that produce juvenile crust. After 200–440 Ma, the new supercontinent is broken up by mantle upwelling beneath it.

During the past billion years (Ga), the earth has experienced the formation and breakup of three supercontinents: Rodinia (formed 1.32–1.0 Ga ago and broken up between 700 and 530 Ma ago), Gondwana (formed between 650 and 550 Ma ago), and Pangea (formed 450–250 Ma ago and began to break up approximately 160 Ma ago). Pangea spanned the planet latitudinally, from today's high Arctic to Antarctica; its eastern flank was notched by a V-shaped Tethys Sea centered approximately on the equator. Unmistakable signs of continental rifting can be seen today underneath the Red Sea and along the Great Rift Valley of East Africa: The process of continental breakup is still very much under way.

Inevitably, these grand geotectonic cycles had an enormous impact on the evolution of life. They made it possible to keep a significant share of the earth's surface above the sea during the past 3 Ga, allowing for the evolution of all complex terrestrial forms of life. Changing locations and distributions of continents and oceans create different patterns of global oceanic and atmospheric circulation, the two key determinants of climate. Emergence and diversification of land plants and terrestrial fauna during the past 500 Ma years have been much influenced by changing locations and sizes of the continents. The surface we inhabit today has been, to a large extent, fashioned by the still continuing breakup of Pangea. Two of the earth's massive land features that have determined the climate for nearly half of humanity—the Himalayas, the world's tallest mountain chain, and the high Tibetan Plateau—are its direct consequences. So is the northward flow of warm Atlantic waters carried by the Gulf Stream that helps to create the mild climate in Western Europe. Volcanic eruptions have been the most important natural source of CO_2 and, hence, a key variable in the long-term balance of the biospheric carbon. Intermittently, they are an enormous source of aerosols, whose high atmospheric concentrations have major global climatic impacts.

3. WATER CYCLE

Most of the living biomass is water, and this unique triatomic compound is also indispensable for all metabolic processes. The water molecule is too heavy to escape the earth's gravity, and juvenile water, originating in deeper layers of the crust, adds only a negligible amount to the compound's biospheric cycle. Additions due to human activities—mainly withdrawals from ancient aquifers, some chemical syntheses, and combustion of fossil fuels—are also negligible. Both the total volume of Earth's water and its division among the major reservoirs can thus be considered constant on a timescale of 10^3 years. On longer timescales, glacial–interglacial oscillations shift enormous volumes of water among oceans and glaciers and permanent snow.

The global water cycle is energized primarily by evaporation from the ocean (Fig. 2). The ocean covers 70% of the earth, it stores 96.5% of the earth's water, and it is the source of approximately 86% of all evaporation. Water has many extraordinary properties. A very high specific heat and heat capacity make it particularly suitable to be the medium of global temperature regulation. High heat of vaporization (approximately 2.45 kJ/g) makes it an ideal transporter of latent heat and helps to retain plant and soil moisture in hot environments. Vaporization of 1 mm/day requires 28 or 29 W/m^2, and with average daily evaporation of approximately

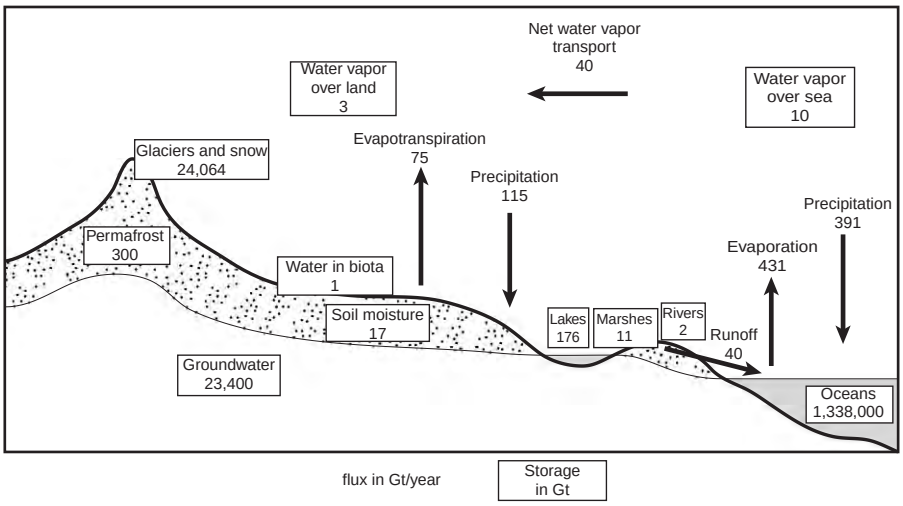


FIGURE 2 Annual flows of the earth's global water cycle. The totals are all in billion of metric tons.

3 mm, or an annual total of 1.1 m, the global latent heat flux averages approximately 90 W/m^2. This means that approximately 45 quadrillion watts (PW), or slightly more than one-third of the solar energy absorbed by the earth's surface and the planet's atmosphere (but 4000 times the amount of annual commercial energy uses), is needed to drive the global water cycle.

The global pattern of evaporation is determined not only by the intensity of insolation but also by the ocean's overturning circulation as cold, dense water sinks near the poles and is replaced by the flow from low latitudes. There are two nearly independent overturning cells—one connects the Atlantic Ocean to other basins through the Southern Ocean, and the other links the Indian and Pacific basins through the Indonesian archipelago. Measurements taken as a part of the World Ocean Circulation Experiment show that 1.3 PW of heat flows between the subtropical (latitude of southern Florida) and the North Atlantic, virtually the same flux as that observed between the Pacific and Indian Oceans through the Indonesian straits.

Poleward heat transfer of latent heat also includes some of the most violent low-pressure cells (cyclones), American hurricanes, and Asian typhoons. Oceanic evaporation (heating of equatorial waters) is also responsible for the Asian monsoon, the planet's most spectacular way of shifting the heat absorbed by warm tropical oceans to continents. The Asian monsoon (influencing marginally also parts of Africa) affects approximately half of the world's population, and it precipitates approximately 90% of water evaporated from the ocean back onto the sea. Evaporation exceeds precipitation in the Atlantic and Indian Oceans, the reverse is true in the Arctic Ocean, and the Pacific Ocean flows are nearly balanced.

Because of irregular patterns of oceanic evaporation (maxima up to 3 m/year) and precipitation (maxima up to 5 m/year), large compensating flows are needed to maintain sea level. The North Pacific is the largest surplus region (and hence its water is less salty), whereas evaporation dominates in the Atlantic. In the long term, oceanic evaporation is determined by the mean sea level. During the last glacial maximum 18,000 years ago, the level was 85–130 m lower than it is now, and satellite measurements indicate a recent global mean sea level rise of approximately 4 mm per year.

The difference between precipitation of evapotranspiration has a clear equatorial peak that is associated with convective cloud systems, whereas the secondary maxima (near 50°N and °S) are the result of extratropical cyclones and midlatitude convection. Regarding the rain on land, approximately one-third of it comes from ocean-derived water vapor, with the rest originating from evaporation from nonvegetated surfaces and from evapotranspiration (evaporation through stomata of leaves). Approximately 60% of all land precipitation is evaporated, 10% goes back to the ocean as surface runoff, and approximately 30% is carried by rivers (Fig. 2). Because the mean continental elevation is approximately 850 m, this riverborne flux implies an annual conversion of approximately 370×10^{18} J of potential energy to kinetic energy of flowing water, the principal agent of geomorphic denudation shaping the earth's surfaces. No more than approximately

15% of the flowing water's aggregate potential energy is convertible to hydroelectricity.

The closing arm of the global water cycle works rather rapidly because average residence times of freshwater range from just 2 weeks in river channels to weeks to months in soils. However, widespread construction of dams means that in some river basins water that reaches the sea is more than 1 year old. In contrast to the normally rapid surface runoff, water may spend many thousands of years in deep aquifers. Perhaps as much as two-thirds of all freshwater on Earth is contained in underground reservoirs, and they annually cycle the equivalent of approximately 30% of the total runoff to maintain stable river flows. Submarine groundwater discharge, the direct flow of water into the sea through porous rocks and sediments, appears to be a much larger flux of the global water cycle than previously estimated.

4. CARBON CYCLE

The key structural element of life keeps cycling on several spatial and temporal scales, ranging from virtually instantaneous returns to circulations that take 10^8 years to complete. At one extreme are very rapid assimilation–respiration flows between the plants and the atmosphere and less rapid but still fast assimilation–decomposition–assimilation circuits that move the element from the atmosphere to plant tissue of various ecosystems and return it to the atmosphere after carbon's temporary stores in dead organic matter (in surface litter and soils) are mineralized by decomposers (Fig. 3). The other extreme involves slow weathering of terrestrial carbonates, their dissolution, riverborne transport, and eventual sedimentation in the ocean, and the element's return to land by geotectonic processes. Human interest in the cycle is due not only to the fact that carbon's assimilation through photosynthesis is the foundation of all but a tiny fraction of the earth's life but also to the cycle's sensitivity to anthropogenic perturbations.

Photosynthesis, the dominant biospheric conversion of inorganic carbon from the atmospheric carbon dioxide (CO_2) to a huge array of complex organic molecules in plants, proceeds with very low efficiency. On average, only approximately 0.3% of the electromagnetic energy of solar radiation that reaches the earth's ice-free surfaces is converted to chemical energy of new phytomass. Short-term maxima in the most productive crop fields can be close to 5%, whereas even the tropical rain forests, the planet's most productive natural ecosystems, do not average more than 1%. Plant (autotrophic) respiration consumes approximately half of the gross photosynthesis, and it reduces the global net primary productivity [i.e., the new phytomass available every year for consumption by heterotrophs (from bacteria to humans)] to approximately 60 Gt of carbon on land and approximately 50 Gt of carbon in the ocean.

Virtually all this fixed carbon is eventually returned to the atmosphere through heterotrophic respiration, and only a tiny fraction that accumulates in terrestrial sediments forms the bridge between the element's rapid and slow cycles. After 10^6–10^8 years of exposure to high pressures and temperatures, these sediments are converted to fossil fuels. The most likely global total of biomass carbon that was transformed into the crustal resources of coals and hydrocarbons is in excess of 5 Tt of oil equivalent, or 2.1×10^{23} J. However, some hydrocarbons, particularly natural gas, may be of abiotic origin. Much more of the ancient organic carbon is sequestered in kerogens, which are transformed remains of buried biomass found mostly in calcareous and oil shales.

Intermediate and deep waters contain approximately 98% of the ocean's huge carbon stores, and they have a relatively large capacity to sequester more CO_2 from the atmosphere. The exchange of CO_2 between the atmosphere and the ocean's topmost (mixed) layer helps to keep these two carbon reservoirs stable. This equilibrium can be disrupted by massive CO_2-rich volcanic eruptions or abrupt climate changes. However, the rate of CO_2 exchange between the atmosphere and the ocean is limited because of the slow interchange between the relatively warm mixed layer ($>25^\circ$C in the tropics) and cold (2–4°C) deep waters. As is the case on land, only a small fraction of organic carbon is not respired and goes into ocean sediments (Fig. 3).

Analysis of air bubbles from polar ice cores shows that the atmospheric concentration of CO_2 has remained between 180 and 300 parts per million (ppm) during the past 420,000 years and 250–290 ppm since the rise of the first high civilizations during the 5000 years preceding the mid-19th century. Humans began changing this relatively stable state of affairs first by clearing natural vegetation for fields and settlements, but until the 19th century these releases of CO_2 did not make any marked difference. These conversions became much more extensive after 1850, coinciding with the emergence of fossil-fueled civilization.

Land use changes dominated the anthropogenic mobilization of carbon until the beginning of the 20th

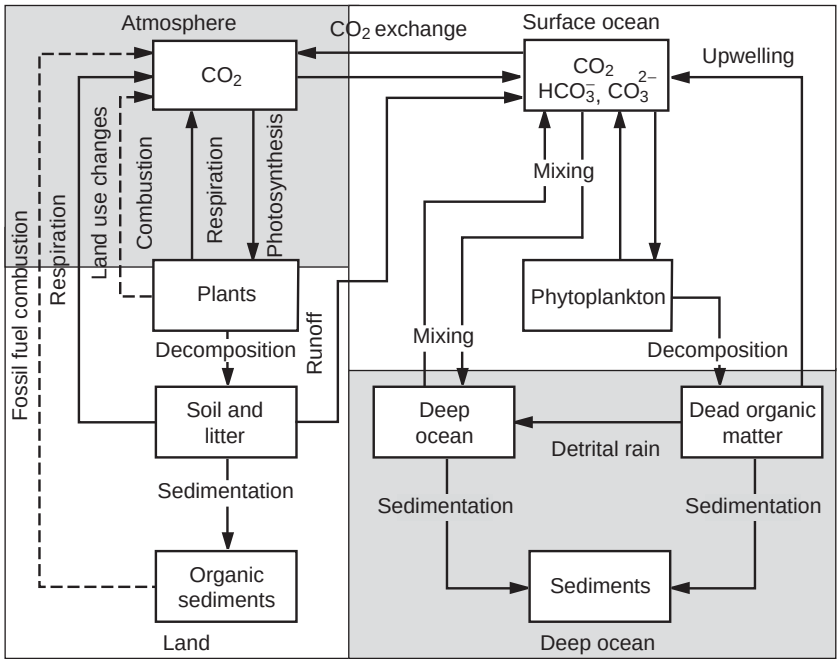


FIGURE 3 Principal reservoirs and flows of the biospheric carbon cycle.

century and they are now the source of 1.5–2 Gt of carbon per year to the atmosphere. Fossil fuel combustion had accelerated from an annual conversion of 22 EJ and emissions of less than 0.5 Gt of carbon in 1900 to more than 300 EJ and releases of more than 6 Gt of carbon by 2000. Every year, on average, approximately half of this anthropogenic flux of approximately 8 Gt of carbon stays in the atmosphere. As a result, atmospheric CO_2 levels that have been constantly monitored at Hawaii's Mauna Loa observatory since 1958 increased from 320 ppm during the first year of measurement to 370 ppm in 2000.

A nearly 40% increase in 150 years is of concern because CO_2 is a major greenhouse gas whose main absorption band of the outgoing radiation coincides with Earth's peak thermal emission. A greenhouse effect is necessary to maintain the average surface temperature approximately 33°C above the level that would result from an unimpeded reradiation of the absorbed insolation, and energy reradiated by the atmosphere is currently approximately 325 W/m^2. During the past 150 years, anthropogenic CO_2 has increased this flux by 1.5 W/m^2 and other greenhouse gases have added approximately 1 W/m^2.

The expected temperature increase caused by this forcing has been partially counteracted by the presence of anthropogenic sulfur compounds in the atmosphere, but another century of massive fossil fuel combustion could double the current CO_2 levels and result in rapid global warming. Its effects would be more pronounced in higher latitudes, and they would include intensification of the global water cycle that would be accompanied by unevenly distributed changes in precipitation patterns and by thermal expansion of seawater leading to a gradual rise in sea level.

5. NITROGEN CYCLE

Although carbon is the most important constituent of all living matter (making up, on average, approximately 45% of dry mass), cellulose and lignin, the two dominant molecules of terrestrial life, contain no nitrogen. Only plant seeds and lean animal tissues have high protein content (mostly between 10 and 25%; soybeans up to 40%), and although nitrogen is an indispensable ingredient of all enzymes, its total reservoir in living matter is less than 2% of carbon's huge stores. Unlike the cycles of water or carbon that include large flows entirely or largely unconnected to biota, every major link in nitrogen's biospheric cycle is mediated by bacteria.

Nitrogen fixation—a severing of the element's highly stable and chemically inert molecule (N_2) that makes up 78% of the atmosphere and its incorporation into reactive ammonia—launches this great cascade of biogenic transformations. Only a very

limited number of organisms that possess nitrogenase, the enzyme able to cleave N_2 at ambient temperature and at normal pressure, can perform this conversion. Several genera of rhizobial bacteria live symbiotically with leguminous plants and are by far the most important fixers of nitrogen on land, whereas cyanobacteria dominate the process in freshwaters as well as in the ocean.

There are also endophytic symbionts (bacteria living inside stems and leaves) and free-living bacterial fixers in soils. The current annual rate of biofixation in terrestrial ecosystems is no less than 160 million metric tons (Mt) of nitrogen, and at least 30 Mt of nitrogen is fixed in crop fields. Biofixation is a relatively energy-intensive process, with 6–12 g of carbon (delivered by the plant to bacteria as sugar) needed for every gram of fixed nitrogen. This means that the terrestrial biofixation of approximately 200 Mt of nitrogen requires at least 1.2 Gt of carbon per year, but this is still no more than 2% of the net terrestrial primary productivity.

Subsequently, there are four key links in the nitrogen cycle (Fig. 4). After nitrogen fixation is nitrification, the conversion of ammonia to more water-soluble nitrates, which is performed by a small number of bacterial species living in both soils and water. Assimilation is the incorporation of reactive nitrogen, be it as ammonia (NH_3) or as nitrate (NO^{-3}), by autotrophic organisms into amino acids that are then used to synthesize proteins. Ammonification is enzymatic decomposition of organic matter performed by many prokaryotes that produces ammonia for nitrification. Finally, denitrification is the closing arm of the cycle in which bacteria convert nitrates into nitrites and then into N_2.

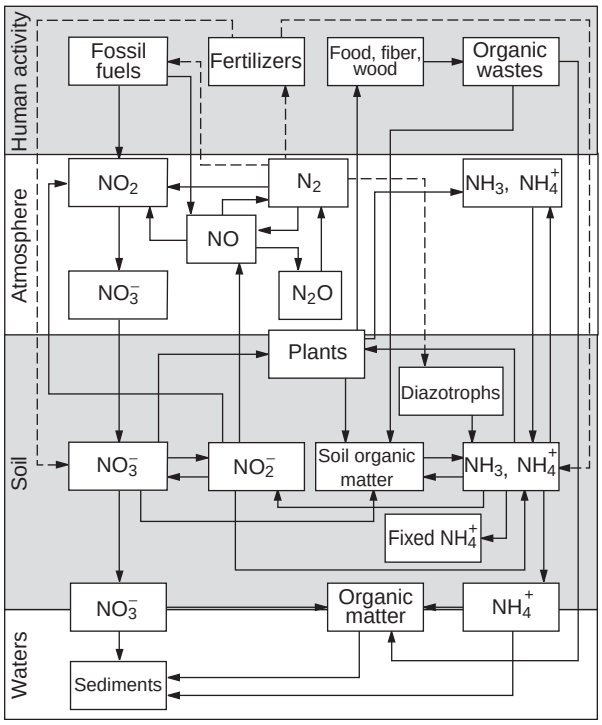


FIGURE 4 Biogeochemical nitrogen cycle centered on plants.

Human interference in the nitrogen cycle is largely due to the fact that heterotrophs, including people, cannot synthesize their own amino acids and hence must digest preformed proteins. Because nitrogen is the most common yield-limiting nutrient in crop cultivation, all traditional agricultures had to resort to widespread planting of leguminous crops and to diligent recycling of nitrogen-rich organic wastes. However, the extent of these practices is clearly limited by the need to grow higher yielding cereals and by the availability of crop residues and human and animal wastes. Only the invention of ammonia synthesis from its elements by Fritz Haber in 1909 and an unusually rapid commercialization of this process by the BASF company under the leadership of Carl Bosch (by 1913) removed the nitrogen limit on crop production and allowed for the expansion of the human population supplied by adequate nutrition (both chemists were later awarded Nobel Prizes for their work). Currently, malnutrition is a matter of distribution and access to food, not of its shortage.

These advances have a considerable energy cost. Despite impressive efficiency gains, particularly since the early 1960s, NH_3 synthesis requires, on average, more than 40 GJ/t, and the conversion of the compound to urea, the dominant nitrogen fertilizer, raises the energy cost to approximately 60 GJ/t of nitrogen. In 2000, Haber-Bosch synthesis produced approximately 130 Mt of NH_3 (or 110 Mt of nitrogen). With production and transportation losses of approximately 10%, and with 15% destined for industrial uses, approximately 82 Mt of nitrogen was used as fertilizer at an aggregate embodied cost of approximately 5 EJ, or roughly 1.5% of the world's commercial supply of primary energy.

The third most important human interference in the global nitrogen cycle (besides NH_3 synthesis and planting of legumes) is the high-temperature combustion of fossil fuels, which released more than 30 Mt of nitrogen, mostly as nitrogen oxides (NO_x), in 2000. Conversion of these oxides to nitrates, leaching of nitrates, and volatilization of ammonia from fertilized fields (only approximately half of the nutrient is actually assimilated by crops) and from animal wastes are the principal sources of nitrogen leakage into ecosystems. Excess nitrogen

causes eutrophication of fresh and coastal waters, and it raises concerns about the long-term effects on biodiversity and productivity of grasslands and forests. NO_x are key precursors of photochemical smog, and rain-deposited nitrates contribute to acidification of waters and soils. Nitrous oxide, a potent greenhouse gas, released during imperfect denitrification is another long-term environmental concern.

6. SULFUR CYCLE

Sulfur is an even rarer component of living molecules than is nitrogen. Only 2 of the 20 amino acids that make up proteins (methionine and cysteine) have the element as part of their molecules, but every protein needs disulfide bridges to make long three-dimensional polypeptide chains whose complex folds allow proteins to be engaged in countless biochemical reactions. The largest natural flux of the element is also the most ephemeral: Sea spray annually carries 140–180 Mt of sulfur into the atmosphere, but 90% of this mass is promptly redeposited in the ocean. Some volcanic eruptions are very large sources of SO_2, whereas others have sulfur-poor veils; a long-term average works out to approximately 20 Mt of sulfur per year, and dust, mainly desert gypsum, may contribute the same amount (Fig. 5).

Biogenic sulfur flows may be as low as 15 Mt and as high as 40 Mt of sulfur per year, and they are produced on land by both sulfur-oxidizing and sulfate-reducing bacteria present in water, mud, and hot springs. Emitted gases are dominated by hydrogen sulfide, dimethyl sulfide (DMS), propyl sulfide, and methyl mercaptan. DMS is also generated in relatively large amounts by the decomposition of algal methionine in the ocean, and these emissions have been suggested to act as homeostatic controllers of the earth's climate. Higher concentrations of DMS would increase albedo by providing more condensation nuclei for increased cloudiness, the resulting reduced insulation would lower planktonic photosynthesis, and diminished DMS emissions would provide fewer condensation nuclei and let in more solar radiation. Later studies showed that not just the magnitude but also the very direction of this feedback are questionable.

Human intensification of the sulfur cycle is due largely to the combustion of coals (containing typically approximately 2% sulfur) and crude oils (sweet crudes have less than 0.5% and sour ones more than 2% sulfur). Some natural gases are high in sulfur, containing it mainly as hydrogen sulfide (H_2S), but it is easily removed before they enter pipelines. Fossil fuel combustion accounts for more than 90% of all anthropogenic sulfur emissions, whose global total increased from 5 Mt in 1900 to approximately 80 Mt in 2000. The remainder is emitted largely during the smelting of color metals (mainly Cu, Zn, and Pb), and a small share originates in chemical syntheses.

Our interest in the sulfur cycle is mainly due to the fact that sulfur dioxide (SO_2) emitted from fossil fuel combustion and nonferrous metallurgy is rapidly oxidized to sulfates, whose deposition is the leading source of acid deposition, both in dry and in wet forms. Before their deposition, sulfur compounds remain up to 3 or 4 days in the lowermost troposphere (average, <40 h), which means that they can travel several hundred to more than 1000 km and bring acid deposition to distant ecosystems. Indeed, the phenomenon was first discovered when SO_2 emissions from central Europe began to affect lake biota in southern Scandinavia. Because of the limits on long-distance transport (atmospheric residence times of DMS and H_2S are also short-lived), sulfur does not have a true global cycle as does carbon (via CO_2) or nitrogen (via denitrification), and the effects of significant acid deposition are limited to large regions or to semicontinental areas.

Deposited sulfates (and nitrates) have the greatest impact on aquatic ecosystems, especially on lakes with low or no buffering capacity, leading to the decline and then the demise of sensitive fish, amphibians, and invertebrates and eventually to profound changes in a lake's biodiversity. Leaching of alkaline elements and mobilization of toxic aluminum from forest soils is another widespread problem in acidified areas, but acid deposition is not the sole, or the principal, cause of reduced forest productivity. Since the early 1980s, concerns about the potential long-term effects of acidification have led to highly successful efforts to reduce anthropogenic SO_2 emissions throughout Europe and in eastern North America.

Reductions of sulfur emissions have been achieved by switching to low-sulfur fuels and by desulfurizing fuels and flue gases. Desulfurization of oils, although routine, is costly, and removing organic sulfur from coal is very difficult (pyritic sulfur is much easier to remove). Consequently, the best control strategy for coal sulfur is flue gas desulfurization (FGD) carried out by reacting SO_2 with wet or dry basic compounds (CaO and $CaCO_3$). FGD is effective,

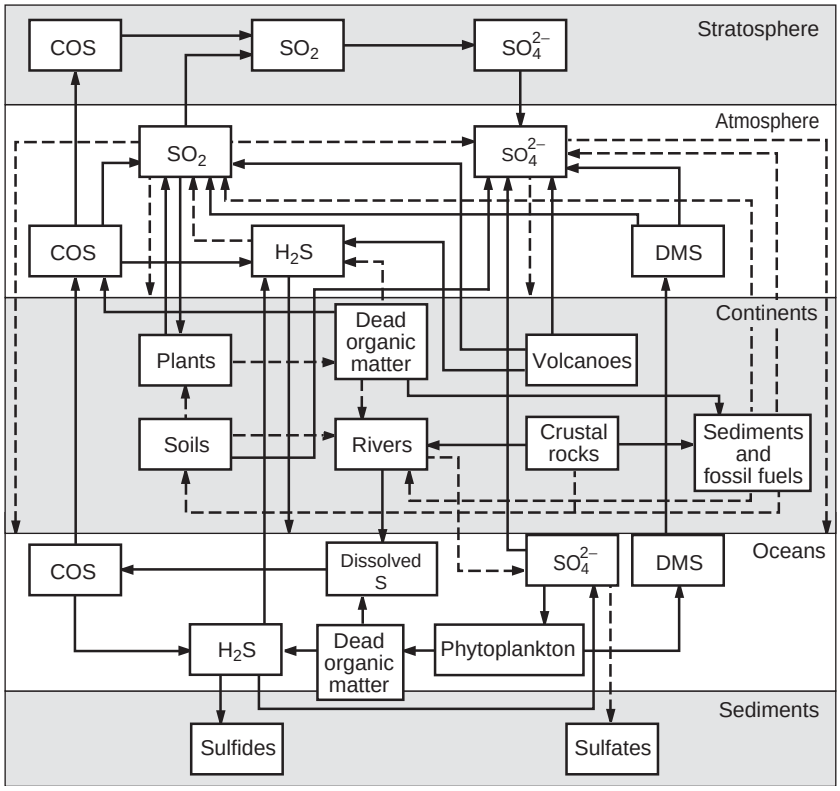


FIGURE 5 Sulfur's biogeochemical cycle extends from the stratosphere to marine sediments.

removing 70–90% of all sulfur, but the average U.S. cost of $125/kW adds at least 10–15% to the original capital expense, and operating the units increases the cost of electricity production by a similar amount. Moreover, FGD generates large volumes of wastes for disposal (mostly $CaSO_4$).

Airborne sulfates also play a relatively major role in the radiation balance of the earth because they partially counteract the effect of greenhouse gases by cooling the troposphere: The global average of the aggregate forcing by all sulfur emissions is approximately $-0.6\,W/m^2$. The effect is unevenly distributed, with pronounced peaks in eastern North America, Europe, and East Asia, the three regions with the highest sulfate levels.

A 16-fold increase in the world's commercial energy consumption that took place during the 20th century resulted in the unprecedented level of human interference in global biogeochemical cycles. Thus, we must seriously consider the novel and a very sobering possibility that the future limits on human acquisition and conversion of energy may arise from the necessity to keep these cycles compatible with the long-term habitability of the biosphere rather than from any shortages of available energy resources.

SEE ALSO THE FOLLOWING ARTICLES

Biomass: Impact on Carbon Cycle and Greenhouse Gas Emissions • Conversion of Energy: People and Animals • Diet, Energy, and Greenhouse Gas Emissions • Earth's Energy Balance • Economic Geography of Energy • Ecosystem Health: Energy Indicators • Environmental Gradients and Energy • Land Requirements of Energy Systems • Sun, Energy from • War and Energy • World History and Energy

Further Reading

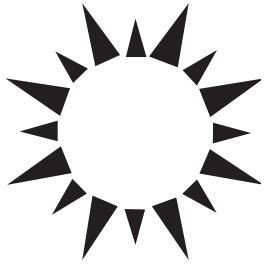
Browning, K. A., and Gurney, R. J. (eds.). (1999). "Global Energy and Water Cycles." Cambridge Univ. Press, New York.

Condie, K. C. (1997). "Plate Tectonics and Crustal Evolution." Oxford Univ. Press, Oxford, UK.

Davies, G. F. (1999). "Dynamic Earth: Plates, Plumes and Mantle Convection." Cambridge Univ. Press, Cambridge, UK.

Galloway, J. N., and Cowling, E. B. (2002). Reactive nitrogen and the world: 200 years of change. *Ambio* **31,** 64–71.
Godbold, D. L., and Hutterman, A. (1994). "Effects of Acid Precipitation on Forest Processes." Wiley–Liss, New York.
Houghton, J. T., *et al.* (eds.). (2001). "Climate Change 2001. The Scientific Basis." Cambridge Univ. Press, New York.
Irving, P. M. (ed.). (1991). "Acidic Deposition: State of Science and Technology." U.S. National Acid Precipitation Assessment Program, Washington, DC.
Smil, V. (1985). "Carbon Nitrogen Sulfur." Plenum, New York.
Smil, V. (1991). "General Energetics: Energy in the Biosphere and Civilization." Wiley, New York.
Smil, V. (2000). "Cycles of Life." Scientific American Library, New York.
Smil, V. (2001). "Enriching the Earth: Fritz Haber, Carl Bosch, and the Transformation of World Food Production." MIT Press, Cambridge, MA.
Smil, V. (2002). "The Earth's Biosphere: Evolution, Dynamics, and Change." MIT Press, Cambridge, MA.

Goods and Services: Energy Costs



ROBERT A. HERENDEEN
Illinois Natural History Survey
Champaign, Illinois, United States

1. Introduction
2. Methods to Determine Energy Cost
3. Illustrations/Applications of Energy Cost
4. Conclusions

Glossary

embodied energy The energy consumed "upstream" to facilitate a flow of goods or services (units = energy).
energy cost of living The energy required to facilitate a household's consumption of goods and services (units = energy/year).
energy intensity (ε) The energy required to facilitate a flow of one unit of a specified good or service (units = energy/gloof).
gloof Generic term to cover the range of goods and services; for example, a pound of fertilizer, the dollar value of an airline ticket, or a liter of water.
indirect effect Necessary costs not considered "direct." For example, auto fuel is usually considered direct energy, whereas energy to build roads is considered indirect. It depends on the definition and system boundary.
input–output (I–O) analysis A subdiscipline of economics that explicitly determines indirect effects.
net energy analysis A comparison of the energy costs and the energy produced by an energy technology such as a coal mine or photovoltaic panel.
power Energy per unit time (units = energy/time).
system boundary The limit up to which costs, benefits, impacts, and consequences are considered; can refer to spatial, temporal, or conceptual issues.
trophic position (TP) An indicator of dependence of an organism or group of organisms on solar-driven photosynthesis in an ecosystem. If the system has a linear food chain, it is called trophic level (dimensionless).
vertical analysis (also known as process analysis) A method to determine energy cost by tracing back through the production process.

Energy analysis determines the total energy required to produce a good or service, including the indirect effects through the chain of production and delivery. This energy is commonly called energy cost, although it does not mean the monetary cost of energy. Once the energy cost is known, the energy requirements of different patterns of production/consumption of goods and services can be analyzed. In this article, the imprecise term energy is used, although the preferred, thermodynamically correct term is free energy.

1. INTRODUCTION

Most of us know intellectually that energy—solar and fossil—supports all of our life-support systems, from agriculture to transportation, commerce, and medicine. Some of us know this more viscerally as well; ecologists and farmers see every day that "all flesh is grass." Readers may remember the U.S. east coast blackout of 1965, the oil crisis of 1973 (long waiting lines at gas stations), or even the California electricity shortages during the summer of 2001. I first "got it" while sitting in traffic in 1971 in Oak Ridge, Tennessee. Ahead of me was a car towing a trailer on which there was an aluminum pleasure boat. Because there was increasing controversy in the Tennessee Valley regarding electricity powered by strip-mined coal and nuclear reactors and also from dams, energy was on my mind. I knew that aluminum is an energy-intensive metal, and I realized that I was looking at "embodied" energy, and hence embodied environmental damage and controversy.

On that day, I started 30 years of work in energy analysis. When I see bread, autos, baby carriages, lawn sprinklers, hospitals, playgrounds, orchestral concerts, haircuts, blue jeans, orthodontic braces, street artists, aircraft carriers, or banjoes, I see energy. The question was and is the following: How much embodied energy? In this article, I present methods used to quantify energy cost and illustrate how it is used in several applications.

 © 2004 Elsevier Inc. All rights reserved.

Energy cost is an example of an indirect effect. Once we realize that a good or service "embodies" an input that is not evident, we are free to calculate the energy cost of an orange, the orange cost of energy, the water cost of hamburger, or how much pollution is released when a Sierra Club book (or this encyclopedia) is published. However, then we are vexed with how far to go in the calculation. For example, how much energy is required to allow a car to travel 1 mile? Here are possible answers to the question:

1. The fuel burned
2. Plus the energy to extract, refine, and transport the fuel
3. Plus the energy to manufacture the car (prorated to 1 mile's use)
4. Plus the energy to produce tires, replacement parts, etc.
5. Plus the energy to build and maintain roads.
6. Plus the energy to maintain auto repair shops, government regulation and registration services, etc.
7. Plus the energy to produce and maintain that portion of the health system used to care for accident victims and sufferers of auto-related health problems.
8. Plus …

Where to stop depends on how we bound the issue, and in the end this decision should be made by the users of the result.

Also, there are additional steps. A similar expanding wave of concern and impact ripples out from, for example, the energy to make the car, which could include

1. The energy consumed at the assembly plant
2. Plus the energy used to make the steel, glass, rubber, etc.
3. Plus the energy used at the iron mine, sand pit, and sulfur mine
4. Plus …

even including the cars used by the iron mine so that the process runs in circles, although successive steps become smaller.

2. METHODS TO DETERMINE ENERGY COST

2.1 Vertical Analysis

The previous process is called vertical analysis or process analysis; I call it following your nose. Fig. 1 indicates the stepwise process of producing a product such as a car. The auto factory is the box on the far left. To produce an auto, the factory receives inputs from three other sectors—for example, steel, glass, and rubber. Each of these receives inputs from other sectors, and so on. Vertical analysis traces this web of inputs backward. At each node, it evaluates the amount of the desired input (e.g., energy) at that point and prorates it on the basis of that sector's output. Given the complexity of a modern economy, vertical analysis is labor-intensive and data-demanding.

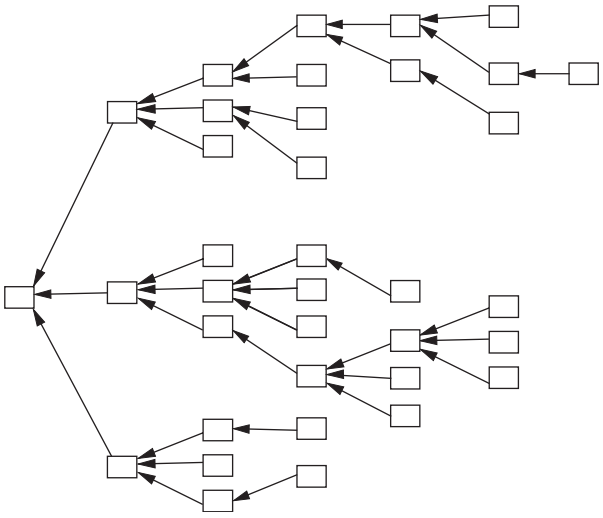


FIGURE 1 Schematic representation of vertical analysis (also called process analysis) to determine energy cost.

2.2 Input–Output Analysis

Economic input–output (I–O) analysis is a well-established method to trace indirect effects in monetary terms. The U.S. Department of Commerce prepares detailed information on the dollar transactions among 350–500 sectors covering the U.S. economy. Many other nations have similar programs. Using several assumptions, one can use the analytical machinery of I–O to combine these data with supplementary information on energy use to determine energy cost.

We start with an economy with N sectors, each with an output flow measured in units of gloof/time. Gloof can be almost anything: tons of steel, pounds of butter, hours in a dentist's chair, dollars worth of day care, etc. It can even vary from compartment to compartment (as in the following example). The assumptions are as follows:

1. Every flow in the system that we wish to count (units = gloof/time) has an associated energy

intensity according to its source compartment (units = energy/gloof).

2. Embodied energy flows everywhere there are flows that we consider important, and the magnitude of the embodied energy flow (units = energy/time) is the product of the energy intensity (units = energy/gloof) times the original flow (units = gloof/time). This assumes that the flow variable is a good surrogate for embodied energy, which is a judgment call by the user.

3. Embodied energy is conserved in each compartment—that is, the amount "in" equals the amount "out."

Assumption 3 is stated by Eq. (1) and illustrated in Fig. 2.

$$\sum_{i=1}^{N} \varepsilon_i X_{ij} + E_j = \varepsilon_j X_j, \tag{1}$$

where X_{ij} is the actual gloof flow from compartment i to compartment j (units = gloof$_i$/time), X_j is the sum of all output flows from compartment j (units = gloof$_j$/time), E_j is the actual energy input flow to compartment j from outside the system (units = energy/time), and ε_j is the energy intensity of output of compartment j (units = energy/gloof$_j$). Equation (1) is assumed to hold for each of the N compartments, yielding N equations in N unknowns. We know the X_{ij}'s and E_j's, so we can solve for the ε_j's. Equation (1) formalizes the allocation of indirect effects in a self-consistent manner and can be used to allocate indirect anything, not just energy.

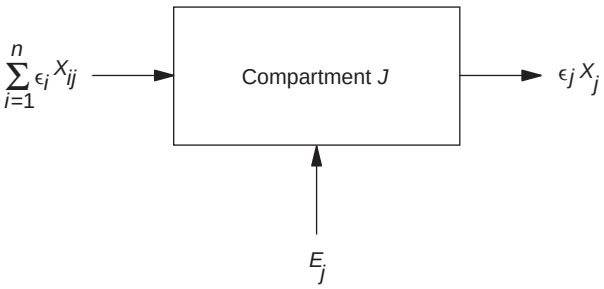


FIGURE 2 Input–output approach to determine energy cost. It is assumed that the embodied energy in = embodied energy out. E_j is the energy flow into sector j (units = energy/time), X_{ij} is the flow of product i to sector j (units = gloof/time), X_j is the total output of sector j (units = gloof/time), and ε_j is the energy intensity of product j (units = energy/gloof).

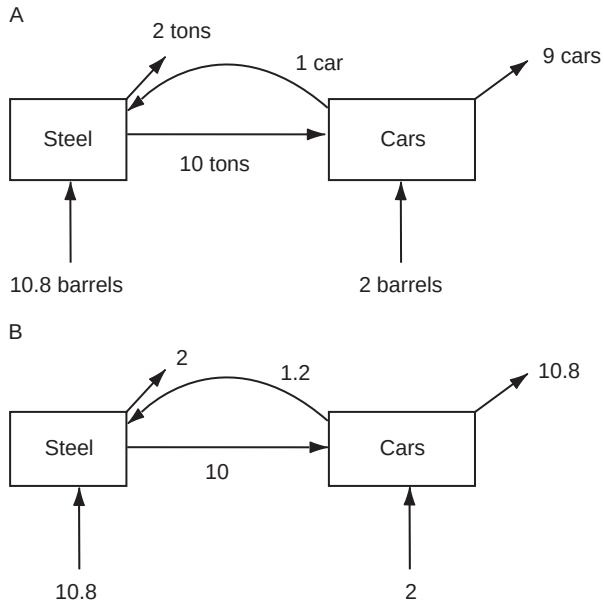


FIGURE 3 A two-sector economy illustrating the calculation of energy cost. (A) Original flows per unit time. Diagonal arrows represent shipments to final consumption. Energy inputs are in barrels of oil. (B) Embodied energy flows (units = barrels/unit time). All flows are assumed steady over time.

2.3 An Example of the Calculation of Energy Cost

Figure 3A shows a two-sector economy burning oil to produce steel and cars. We use Eq. (1) to determine the energy intensities:

$$\text{Steel sector: } 10.8 + 1\varepsilon_{\text{car}} = 12\varepsilon_{\text{steel}}$$
$$\text{Car sector: } 10\varepsilon_{\text{steel}} + 2 = 10\varepsilon_{\text{car}}$$

Solving these two equations gives $\varepsilon_{\text{steel}} = 1$ barrel oil/ton steel and $\varepsilon_{\text{car}} = 1.2$ barrel oil/car. The embodied energy flows are obtained by multiplying the flows in Fig. 3A by the intensities, giving the result shown in Fig. 3B. Each sector is in embodied energy balance, as we assumed in Eq. (1). In addition, the whole system is in balance: 12.8 barrels of oil enters and is burned, and 12.8 barrels of embodied oil is contained in the shipments of steel and cars to final consumption. Furthermore, only 2/12 of the energy to produce a car is used directly (i.e., burned in the auto factory).

This is the I–O approach for two sectors. Calculation for systems with more sectors is operationally the same but tediously complicated—a job easily handled by computers. Typical results are shown in Table I. In addition, one may show that a vertical analysis of Fig. 3A yields the same results as the I–O approach. Because of the feedback of cars to steel, the analysis has an infinite number

TABLE I

Energy Intensities for Personal Consumption Categories, 1972–1973[a]

Consumption category	Energy intensity (103 Btu/$1972)	Consumption category	Energy intensity (103 Btu/$1972)
Food at home	54	Auto insurance	19
Food away from home	44	Auto registration and fees	0[a]
Alcoholic beverage	51	Local bus and train, commuting	56
Tobacco	30	Local bus and train, school	56
Rented dwelling	14	Local bus and train, other	56
Owned dwelling	12	Transportation to school away from home	132
Other shelter, lodgings	21	Plane, trip outside commuting area	167
Fuel oil and kerosene	890	Train and bus, trip outside commuting area	68
Coal and wood	795	Ship, trip outside commuting area	129
Other fuels	795	Limousine and taxi, trip outside	56
Gas main	915	commuting area	
Gas in bottles and tanks	999	Car rental	38
Electricity	616	Health insurance	19
Water, sewage, trash	26	Health care	47
Telephone, telegraph, cable	24	Personal care	48
Domestic service	0[a]	Owned vacation home	12
Household textiles	59	Other transportation costs on vacation	91
Furniture	45	All-expense tours, summer camps	31
Floor coverings	71	Other vacation expenses	33
Major appliances	66	Boats, aircraft, motorcycles	65
Small appliances	65	Television	48
Housewares	57	Other recreation	41
Miscellaneous house furnishings	48	Reading	43
Dry cleaning, laundry	44	Private education	35
Clothing	44	Public education	37
Vehicle purchase	68	Miscellaneous consumption expenditure	41
Vehicle purchase finance charges	27	Personal insurance and pensions	20
Gasoline and oil	443	Gifts and contributions	41
Tires and lubrication	69	Increase in savings	47
Batteries	59	Housing purchase and improvement	40
Auto repair and service	37	Increase in investment	47

[a] Labor and government services are assumed to have zero energy intensity relative to the consumer. For labor, this is done to avoid double counting. For government, the energy requirement is not lost but rather charged against government expenditures.

of steps, but a computational trick makes this manageable.

For typical energy studies, the following assumptions have usually been made: Energy inputs are coal, petroleum, natural gas, and a "fossil equivalent" of hydro, wind, photovoltaic, and nuclear electricity. Solar energy input to silviculture and agriculture is usually considered a free input and not counted for economic systems. However, solar energy is usually counted in energy analysis of ecosystems.

3. ILLUSTRATIONS/APPLICATIONS OF ENERGY COST

3.1 Auto Manufacture and Use

Energy analysis showed that approximately 12% of the energy to produce and market a U.S. automobile was consumed in the automobile sector vs. 37% in the primary metals sectors. In addition, the fuel in the tank was only approximately 60% of the energy to provide auto transport; the other 40% was car

TABLE II

Energy Cost of Corn Production in Illinois, 1996

Type of agiculture:	Conventional
Crop:	Yellow corn
Field size:	76 acres
Yield:	168 bushels/acre
Calorific energy of crop	5.13×10^9 Btu (402×10^3 Btu/bushel)
Crop energy/input energy:	8.4

	$1996 Cost or Btu	Energy intensity (10^3 Btu/$1996 or Btu/Btu)	Energy (10^6 Btu)	% of total
Seed	$2284	26.2	59.8	9.8
Fertilizer	$4725	769.7	363.7	59.6
Herbicide	$2179	26.2	57.0	9.3
Pesticide	0	26.2	0	0.0
Machines	$2017	10.7	43.0	7.0
Farm fuel	52×10^6 Btu	1.20[a]	62.3	10.2
Custom application	$2129	11.6	24.7	4.0
Total			610.4	100.0

[a] Twenty percent extra energy is consumed in extracting, refining, and transporting fuel to point of use. This is an experimental plot; in commercial practice, typically some pesticides would be applied.

manufacture (8%), energy cost of energy (9%), tires, maintenance, road construction, parking, insurance, etc. This is an old result from the 1970s, but it is still compelling.

3.2 Grain Production

The ecologist Howard T. Odum said we eat "potatoes made of oil," a reflection of the energy demands of modern mechanized agriculture. Table II shows typical energy requirements for a corn crop. In the rich soil and favorable climate of Illinois, the fossil energy inputs sum to only one-eighth of the crop's calorific value (the sun is not counted). Note that fertilizer and herbicides account for 2/3 of total energy input, while farm fuel is only 10%. It is debatable whether fossil fuel inputs should be compared with the calorific content of food. More compelling are studies that show that the fossil energy input per unit of food output has increased during the past several centuries. Likewise, if we move to the right on any of these spectra—good→marginal soil and climate, grain→flesh, low→high processing and packaging, near→distant—the energy inputs typically increase. For the average food as delivered to American mouths, the fossil energy input is 5–10 times the calories delivered.

3.3 Household Energy (Energy Cost of Living)

Once obtained, energy intensities can be used to link everyday personal consumption patterns with their impact in resource use, as indicated in Eq. (2):

$$\text{Energy} = \sum_{i=1}^{N} (\text{energy intensity of consumer good } i) \times (\text{consumption of consumer good } i). \quad (2)$$

As with much energy analysis, effort peaked in the 1970s and 1980s, but here I present an update.

3.3.1 Energy Intensities

Very few vertical analyses have been performed on specific products, such as automobiles and agricultural chemicals. The remaining energy intensities have been calculated for approximately 350 sectors covering the full range of economic expenditures in the U.S. I–O economic tables. One assumes that dollars are a satisfactory numeraire for allocating embodied energy. Energy data compatible with I–O data are provided by the U.S. Department of Energy, Energy Information Administration. These are used with a version of Eq. 1.

One immediate question is whether the energy intensities do vary among sectors. If not, the energy

TABLE III

Energy Intensities of Consumer Expenditures, Updated (by Approximations) to 1999

Expenditure category	Energy intensity (Btu/$1999)
Food	9095
Alcohol and tobacco	5457
Housing	5457
Residential energy	105000
Gasoline and motor oil	115000
Auto purchase, maintenance	10914
Public transportation	20000
Apparel	8025
Health, personal care	8560
Entertainment and communication	5457
Education and reading	7276
Miscellaneous	7490
Contributions	7490
Insurance, pension	3638
Asset change	8560

consequences of spending a dollar for any good or service would be the same, and the details of expenditures would be irrelevant. We already know that the primary metals industry is energy intensive compared with a service industry, but we are interested in consumption options open to individuals and households, which generally covers finished products. Even so, there is a significant variation, as shown in Table I. For example, the energy intensity of gasoline was 443,000 Btu/$ 1972, whereas that of private education was 35,000 Btu/$ 1972, a variation of a factor of 13. Among nonenergy, nontransportation categories, there was a variation of a factor of 5, from rental dwellings (14,000 Btu/$ 1972) to floor coverings (71,000 Btu/$ 1972). In Table III, the consumption categories have been aggregated to 15 and updated to 1999.

3.3.2 Household Consumption Data

The U.S. Bureau of Labor Statistics (BLS) periodically performs detailed surveys of consumption "market baskets" of U.S. households. A total of 19,000+ households were surveyed in 1972 and 1973, with expenditures broken down into 61 categories as given in Table I. In the updated analysis here, I use results of a smaller BLS survey for 1999. The system boundary thus is defined by what the household purchases.

Combining expenditures and intensities using Eq. (2) yields Fig. 4, which shows the energy cost taken back to the mine mouth or wellhead, of the market basket of average households in lowest and highest expenditure quintiles and for the overall average household. Of the 15 consumption categories, 2 are direct energy expenditures (residential fuel and electricity and auto fuel). For the average households, these amount to 5% of monetary expenditures (Fig. 4A) but 44% of the household's total energy cost (Fig. 4B). Figure 4B also shows that for the lowest quintile, the direct energy requirement is approximately 57% of the total, and for the top quintile it is approximately 33%. There seems to be an upper limit on how much auto fuel a household can use but not on football tickets, clothing, furniture, plane tickets, second homes, etc. For the average expenditure level and above, one would miss more than half of the energy cost of living by examining only direct energy.

Under certain assumptions, the energy cost of living can be used for two more speculative purposes: to predict the economic effects of energy price increases on different households (this has been done), and in the analysis of sprawl issues (because more auto use and larger residences are associated with living further from urban centers).

3.4 Appliances: Energy to Operate vs Energy to Produce/Maintain/Dispose

What fraction of the energy use for a refrigerator or a toaster is used to manufacture, market, maintain, and mash it, and what fraction is used to operate it? The answer is useful in evaluating different strategies to reduce energy use, such as increasing operational efficiency, decreasing energy use in manufacture, or increasing the device's lifetime. This is an example of life cycle cost in energy terms. Consider all the energy associated with an appliance:

$$E_{\text{life}} = \text{manufacturing energy} + \text{maintenance energy} + \text{disposal energy} + \text{operational energy}.$$

The average total power is the lifetime energy divided by the lifetime, T:

$$p_{\text{tot}} = \frac{E_{\text{life}}}{T} = \left(\frac{\text{manufacturing} + \text{maintenance} + \text{disposal}}{E_{\text{life}}} + 1\right) \times p_{\text{operation}}. \quad (3)$$

For electric appliances, energy is measured in kilowatt-hours, and power is measured in kilowatts.

Table IV summarizes energy use for selected household appliances. Most appliances tend to have

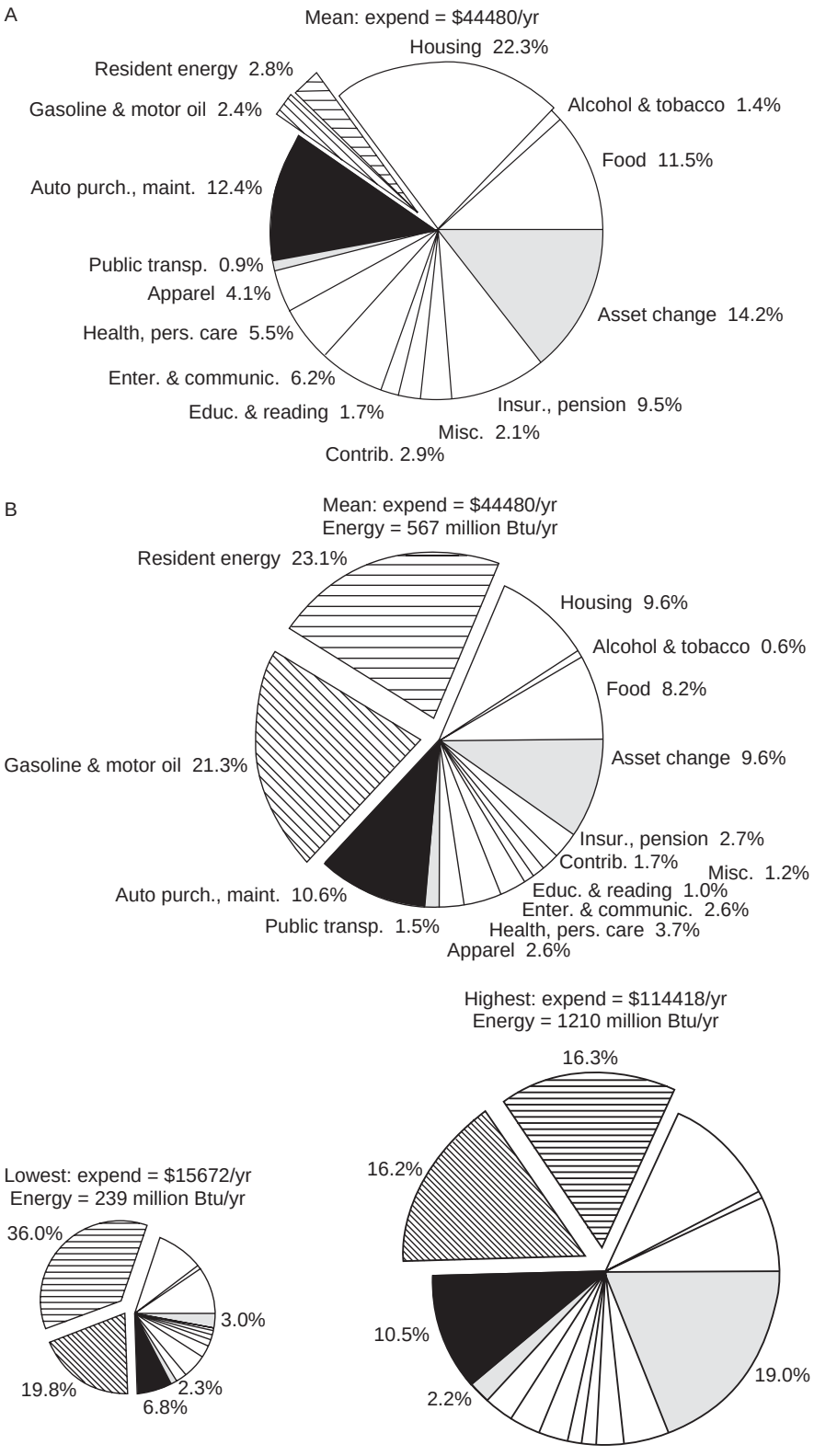


FIGURE 4 Energy cost of living, 1999. (A) Annual expenditures for average U.S. household. (B) Energy costs of expenditures in Fig. 4A and for lowest and highest expenditure quintiles as well. Figure 4B is obtained from Fig. 4A by multiplying energy intensities by expenditures (Eq. 2). In Fig. 4B, the area of the circle is proportional to the total household energy impact.

TABLE IV

Comparison of Operating and (Manufacuring + Maintenance + Disposal) Energy for Selected Appliances

Electric appliance	$p_{operation}$ (kWh/year)	T(year)	Indirect fraction (manufacturing–maintenance + disposal) as fraction of life energy	Classification ($p_{operation}$/indirect fraction)[b]
Blender	17	14	0.34	Low/high
Mixer	14	14	0.35	Low/high
Refrigerator	750	14	0.13	High/low
Water heater	4700	8	0.01	High/low
75-W incandescent bulb[b]	657	0.1	0.02	High/low
17-W compact fluorescent bulb[b]	146	1.14	0.08	High/low
Coffee maker	83	8	0.07	Low/low
Toaster	43	8	0.19	Low/low
Personal computer (at work)[c]	160	7	0.5	High/high

[a]Classification: $p_{operation}$: <100 kWh/year = low; ≥ 100 kWh/year = high. Indirect fraction: <0.25 = low; ≥ 0.25 = high.
[b] Assumes bulbs are on continuously. These two bulbs provide approximately equal light levels.
[c] Assumes the computer is used 8 h/day, 200 days/year.

high operation power and a low indirect energy fraction [i.e., (manufacturing + maintenance + disposal)/$E_{life} < 1$] or vice versa. For the appliances usually considered large energy users, such as refrigerators, freezers, water heaters, and light bulbs, the fraction is ≤ 0.25. For smaller appliances, it is ≥ 0.25. This implies, but does not prove, that to save energy one should concentrate on operational energy efficiency for the first group, whereas for the second there is likely potential in both operational and manufacturing efficiency.

The personal computer, if used 8 h a day, 5 days a week, has a high operation power (160 kWh/year) and a high indirect fraction (≈ 0.5).

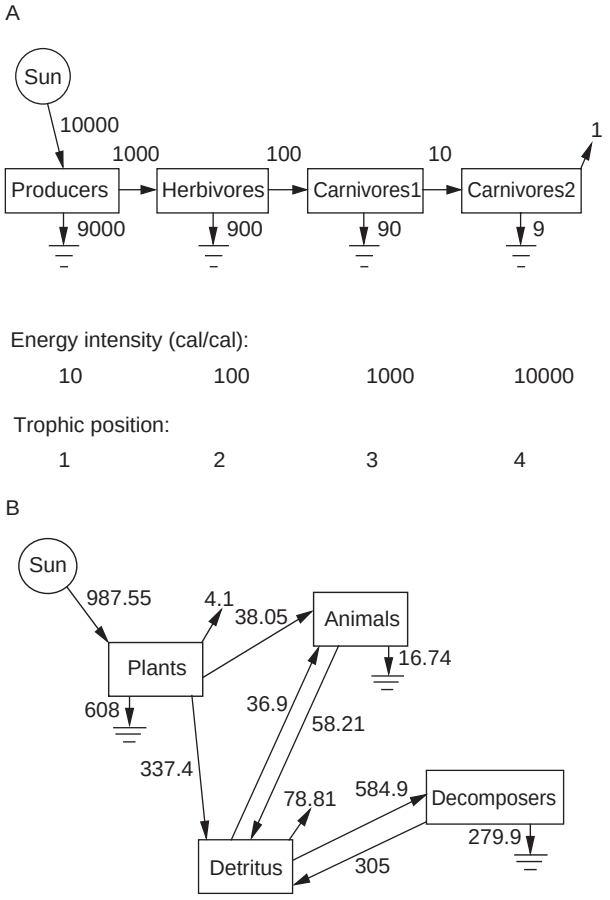
3.5 Energy Cost in Ecosystems

There are several indicator quantities that summarize indirect energy effects in ecosystems. Two that are especially appropriate for quantifying the solar dependence of all life (with the negligible exceptions of deep-sea vent communities, which are based on geochemical energy) are energy intensity and trophic position. Trophic structure refers to the number of energy "transactions" separating the sun (which is assigned trophic position 0) from the compartment in question. Originally, energy flow in ecosystems was visualized in terms of straight food chains, for which trophic positions (called trophic levels in this case) are successive integers (Fig. 5A). However, most ecosystems have web-like energy flows (Fig. 5B). Equation (1) applies to webs, so calculating energy intensities requires no new technique. On the other hand, trophic position needs to be defined. The trophic position of a compartment, TP, is the energy-weighted sum of TPs of inputs $+1$. Using the language we used for calculating energy intensities, for each compartment,

$$\mathrm{TP}_j = \sum_{i=1}^{N} \left(\frac{X_{ij}}{\sum_{i=1}^{N} X_{ij}} \right) \mathrm{TP}_i + 1. \qquad (4)$$

Equation (4) is similar to Eq. (1), but it has two important differences. First, the flows X_{ij} must be in terms of biomass energy because trophic ecology is by definition concerned with energy flow. Second, the factors in the bracket sum to 1 because they are normalized with respect to the input flows, whereas in Eq. (1) normalization is with respect to total output and the factors X_{ij}/X_j need not sum to 1. Equation (4) represents N equations in N unknowns.

For the (idealized) food chain in Fig. 5A, the weighting factors are all 1, and Eq. (4) gives the trophic positions 1–4. For the (real) food web in Fig. 5B, TPs are 1, 3.43, 3.90, and 4.90. Because plants receive input only from the sun, $\mathrm{TP}_{plants} = 1$. Because decomposers gets all their input from detritus, $\mathrm{TP}_{decomposers} = \mathrm{TP}_{detritus} + 1$. Animals and detritus have more than one input, and their TPs are mixtures. In this accounting, decomposers (e.g., bacteria) are on top of the energy pyramid and the food web.



Compartment	Energy intensity (cal/cal)	Trophic position
Plants	2.60	1.00
Animals	9.56	3.43
Detritus	12.40	3.90
Decomposers	23.80	4.90

FIGURE 5 Energy intensities and trophic positions for food chain and web. (A) Idealized food chain. (B) Boreal bog ecosystem food web. Units for both = g fixed carbon m^{-2} $year^{-1}$. Flows to ground symbols are metabolic losses. Detritus is dead material the origin of which is not visually identifiable. Decomposers include bacteria.

4. CONCLUSIONS

Energy is important, and the idea of energy cost makes intuitive sense. The details of obtaining and using energy cost complicate its usefulness, but when done transparently, it can provide insight and help in decision making. To do it transparently, we require at the least a procedure, a system boundary, and many assumptions. For brevity, I have not discussed most of the assumptions, and I have included only a few applications of energy cost. Others include comparing energy embodied in imported and exported goods and services (energy balance of trade), comparing embodied energy and labor costs to determine the employment effects of energy policy, determining equity effects of energy taxes, and comparing the energy cost with the energy output of an energy source technology such as a power dam. The latter, called net energy analysis, is presented in a separate article in this encyclopedia.

SEE ALSO THE FOLLOWING ARTICLES

Cost–Benefit Analysis Applied to Energy • Energy Services Industry • External Costs of Energy • Global Energy Use: Status and Trends • Industrial Ecology • Industrial Energy Efficiency • Industrial Energy Use, Status and Trends • Input–Output Analysis • Leisure, Energy Costs of

Further Reading

Bullard, C., and Herendeen, R. (1975). The energy costs of goods and services. *Energy Policy* **3,** 268–278.

Bullard, C., Penner, P., and Pilati, D. (1978). Net energy analysis: Handbook for combining process and input–output analysis. *Resour. Energy* **1,** 267–313.

Finn, J. T. (1976). Measures of ecosystem structure and function derived from analysis of flows. *J. Theor. Biol.* **56,** 115–124.

Hall, C., Cleveland, C., and Kaufmann, R. (1986). "Energy and Resource Quality: The Ecology of the Economic Process." Wiley, New York.

Herendeen, R. (1988). Net energy considerations. *In* "Economic Analysis of Solar Thermal Energy Systems" (R. West and F. Kreith, Eds.), pp. 255–273. MIT Press, Cambridge, MA.

Herendeen, R. (1998). "Ecological Numeracy: Quantitative Analysis of Environmental Issues." Wiley, New York. [See Chap. 8].

Herendeen, R., and Fazel, F. (1984). Distributional aspects of an energy conserving tax and rebate. *Resour. Energy* **6,** 277–304.

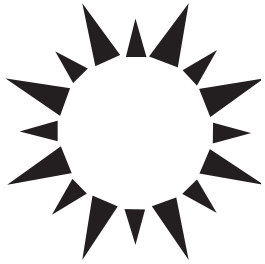
Herendeen, R., Ford, C., and Hannon, B. (1981). Energy cost of living, 1972–1973. *Energy* **6,** 1433–1450.

Odum, H. T. (1996). "Environmental Accounting." Wiley, New York.

Schulze, P. (ed.). (1999). Measures of Environmental Performance and Ecosystem Condition. National Academy of Engineering, Washington, D.C.

Slesser, M. (ed.). (1974). Energy analysis workshop on methodology and conventions, August 25–30, Guldsmedhyttan, Sweden. International Federation of Institutes of Advanced Study, Stockholm, Sweden. [Reprinted in Energy accounting as a policy analysis tool, Committee print for the 94th Congress, 2nd session, Serial CC, 68–391. U.S. Government Printing Office, Washington, DC, 1976]

Green Accounting and Energy



PETER BARTELMUS
Wuppertal Institute for Climate, Environment and Energy
Bergische Universität Wuppertal, Germany

1. **Accounting for Sustainability**
2. **The Concepts and Methods of Integrated Environmental and Economic Accounting**
3. **Accounting for Energy and Related Impacts**
4. **Outlook: Policy Use of Accounting Indicators**
Appendix: Monetary Valuation in the SEEA

Glossary

capital maintenance Keeping produced and natural capital intact as a principle of sustainability of economic activity, applied in environmental accounting.

dematerialization Reduction of material inputs into, and residual outflows from, the production and consumption of goods and services as a means of diminishing environmental pressure from economic activity.

energy accounts Accounts and balances, which measure energy inputs, transformations and uses; sometimes extended to include material flows in embodied energy values.

energy theory of value Rationale for the quantitative evaluation of natural resources and goods and services in terms of energy expended in their production.

environmental accounting (1) National accounting: short for integrated environmental and economic accounting, as advanced by the System of Environmental and Economic Accounts (SEEA). (2) Corporate accounting: usually refers to environmental auditing but also includes environmental costing and life-cycle (cradle-to-grave) assessments of products.

environmental valuation Different techniques of estimating a monetary value for natural assets or asset changes in environmental accounting, including market valuation of natural resources, maintenance costing of environmental impacts, and damage valuation of environmental effects.

environmentally adjusted (net) capital formation (ECF) Environmental accounting indicator that deducts the cost of natural capital consumption from net capital formation.

environmentally adjusted (net) domestic product (EDP) Environmental accounting aggregate, obtained by deducting the cost of natural capital consumption from net domestic product.

green accounting Popular term for environmental accounting, including material flow accounts.

material flow accounts Physical accounts and balances, which present inputs of materials into the economy, their accumulation, and their outflow to the natural environment and other economies; usually measured in mass (weight) units.

natural capital Natural assets in their role of natural resource supply and provision of environmental services (notably of waste/residual disposal) for economic production and consumption.

System of Environmental and Economic Accounts (SEEA) Extended national accounts system, advanced by the United Nations and other international organizations; incorporates natural assets and asset uses to measure the interactions between environment and economy in physical and monetary units.

total material requirement (TMR) Key indicator of material flow accounts that measures the total material base of the economy; includes besides direct material inputs, "hidden" movements of materials associated with imports and extraction of natural resources.

Green accounting is a popular term for environmental accounting or, more precisely, integrated environmental and economic accounting at national and corporate levels. At both levels, the purpose is to capture the elusive notion of long-term sustainability of economic performance, hampered by environmental impacts of production and consumption. Long lists of environmental and sustainable development indicators failed to assess sustainable economic growth and development in a comprehensive and transparent manner. The more systematic greening of the national accounts succeeds in defining and measuring the environmental sustainability of economic activity. This article describes the concepts and methods of environmental accounting at the national level and addresses specifically the role of energy in accounting and sustainability analysis.

 © 2004 Elsevier Inc. All rights reserved.

1. ACCOUNTING FOR SUSTAINABILITY

1.1 Concepts and Indicators of Sustainable Development

Conspicuous pollution incidents in the 1960s and the energy crises of 1973 and 1974 brought about environmental gloom. Well-known titles like Carson's "silent spring" or Loraine's "death of tomorrow" alerted to risks for human survival. The Club of Rome presented in 1972 a computerized and hence seemingly objective world model, which predicted the collapse of the industrial system and Malthusian population decline within this century. Nonrenewable resource depletion was the main reason for this business-as-usual scenario.

William Kapp can be credited with alerting mainstream economists to the social costs of environmental concerns. Neoclassical economics had largely ignored environmental phenomena as external to economic planning and budgeting. In Kapp's view the staggering social costs of environmental damage had reached a magnitude beyond the capacities of welfare analysis, national income accounting, and compartmentalized social and natural sciences. Based on a similar assessment of compartmentalized institutions, the World Commission on Environment and Development (WCED) advanced in 1987 the notion of sustainable development. The commission's report on "Our Common Future" described the new development paradigm as an integrative approach to tackling the interlocking environmental and developmental crises. Two Earth Summits, convened by the United Nations in 1992 in Rio de Janeiro and 2002 in Johannesburg, confirmed sustainable development as a global paradigm of equal importance for industrialized and developing countries.

The WCED also offered the popular definition of sustainable development as "development that meets the needs of the present without compromising the ability of future generations to meet their own needs." The definition is vague: it does not specify the categories of human needs, does not give a clear time frame for analysis and does not indicate particular roles for the environment or social concerns in long-term development. The paradigm needs to be operationalized, and indeed numerous indicators for and of sustainable development have since been advanced.

Indicators *for* sustainable development typically include long lists of measures of environmental, social, economic, and institutional aspects of development. Relatively loose frameworks generate some useful categorization but do not in general link or aggregate these indicators. The main drawback of indicator lists is the difficulty of comparing and evaluating data expressed in different units of measurement. Such indicators may support the management of particular areas of environmental, social, or economic concern but are less useful for the overall planning and implementation of sustainable development.

Proposals of indicators, or indices rather, *of* sustainable development has been the response to the comparability and aggregation problem. The approach is to average or otherwise combine selected indicators, or to correct conventional economic aggregates of wealth, income, and consumption. Compound measures of sustainable or human development or social progress deserve our attention since they all try to take the environment "into account." Well-known examples are the following:

- Human Development Index (HDI), developed by the United Nations Development Programme. The index is an average of per capita income, life expectancy, and education, intended to provide a more complete (than GDP) assessment of human achievements.
- Environmental Sustainability Index (ESI), initiated by the World Economic Forum. An average high performance of 20 indicators, made up of 68 variables, defines environmental sustainability.
- Genuine Progress Indicator (GPI) for the United States, later applied and modified in other countries. The GPI subtracts defensive expenditures (of environmental protection, defense, and other welfare maintaining—rather than increasing—expenses) and adds or subtracts positive and negative environmental and other social costs and benefits.
- World Bank's National Wealth aggregate of produced assets, natural capital, and human resources.

All these indices suffer from a more or less arbitrary selection of partially correlated indicators, a questionable mix of weighting and valuation techniques, or inconsistency with standard accounting concepts and conventions. The idea seems to be to demonstrate the nonsustainability of economic growth, rather than to support policy making. For instance, the authors of the GPI present their

indicator explicitly as a means to discredit GDP as a measure of social progress.

1.2 Green Accounting: Assessing the Environmental Sustainability of Economic Growth

The elusive notion of sustainability needs to be operationalized in a more transparent and systematic manner. To this end, natural science offers basic principles for the nature-economy interface, and economics provides the tool of accounting for this interface. Thermodynamic laws of matter and energy conservation and dissipation govern the use of natural resources. Formal double-entry accounting can then be applied to assess the use (input) and dispersion (output) of these resources from/to the natural environment.

Figure 1 illustrates the "throughput" of materials (including energy carriers) through the economy. The figure also shows the two basic functions of the environment: (1) resource supply of raw materials, space, and energy to the economy (source function) and (2) waste assimilation (sink function). The latter represent an important ecological service. Other ecosystem services such as habitat for species, soil protection, or local climate control are sometimes categorized as indirect benefits for humans; they are probably better assessed by regionalized environmental and ecological statistics.

The fortunate coincidence of physical laws that ensure the equality of inputs and outputs of energy and materials, and input-output based economic accounting points to an obvious approach of applying the latter to the former. The result is green accounting—that is, the extension of conventional economic accounts into the physical world of the use and abuse of nature by economic activity.

The application of accounting tools requires aggregation of physical environmental data by means of a common measuring rod. Environmentalists and economists—to make a crude distinction between ecological and economic views of the human environment—disagree on whether to use physical measures, or prices and costs, for weighting the importance of environmental impacts. Environmental economists seek to "internalize" the costs of environmental impacts (environmental "externalities") into the decisions of households and enterprises. Environmentalists, on the other hand, refute such commodification of an indivisible (public) good on whose value markets should not have a say.

Two operational sustainability concepts can thus be distinguished according to the economic and ecological outlook:

- Economic sustainability makes use of the established requisite for economic growth, capital maintenance, and extends the (produced) capital concept to include nonproduced natural capital.
- Ecological sustainability considers material throughput through the economy as a pressure on the carrying capacities of natural systems, and aims to reduce this pressure to tolerable levels by dematerializing the economy.

Two accounting systems attempt respectively to capture these notions of sustainability. Material flow accounts (MFA) and related physical input-output tables (PIOT) measure material flows in physical—weight—units. The United Nations System for integrated Environmental and Economic Accounting (SEEA) embraces both physical and monetary accounts, seeking compatibility with the worldwide adopted System of National Accounts (the SNA). Considering the integrative power of the SEEA, the

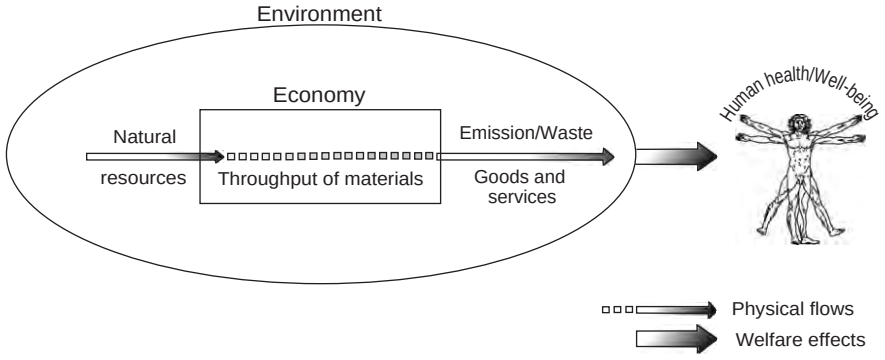


FIGURE 1 Environment-economy interaction and effects. Flows of material throughput from the environment into the economy and out of the economy into the environment. Outputs from the economy and the environment affect human welfare. From Bartelmus (2001) Fig. 1, p.1723, with permission from EOLSS publishers.

1992 Earth Summit called for its application by all member States of the United Nations in its Agenda 21. An international group of experts and accountants has now revised the SEEA, which the United Nations and other international organizations will issue as "Integrated Environmental and Economic Accounting 2003."

The narrow focus on environmental (at the expense of social and other concerns) sustainability of economic activity by both types of accounts does not do justice to the much broader paradigm of sustainable development. Development itself encompasses, besides economic objectives, social, ecological, cultural, and political ones. A corresponding broad sustainability notion would therefore have to incorporate all these objectives. Measurability is the main problem in accounting for these concerns, even when more narrowly focusing on the costs of human, social, and institutional capital formation and use in economic growth.

1.3 Energy: The Ultimate Limit?

Several energy analysts and ecological economists have deplored the neglect of energy in mainstream economic analysis. In their view, energy can be seen as the ultimate source of life and at the same time as its ultimate limit. The reasons for this neglect include a relatively low share of energy costs in production due to incomplete energy accounting and the underpricing of energy resources. Prices that ignore the social costs of energy depletion and pollution from energy production and use create a delusion of abundance and might undermine the sustainability of economic activity. Subsidies, notably of labor-intensive fossil fuel production, may increase employment but aggravate underpricing and environmental impacts. Addressing the pressures on and scarcities of natural resources, the WCED seems to concur that the ultimate limits to global development might indeed be determined by the impacts of energy use.

The significance of energy production and consumption for the overall sustainability of economic development can, however, only be assessed in a broader context of accounting, comprehensively and comparably, for economic benefits, natural resource use, and waste generation by different economic sectors. Section 2 describes the general concepts and methods of such accounting.

Section 3 then indicates how environmental accounting applies to energy production and consumption and related environmental impacts. It does not elaborate on the different techniques of narrowly defined energy balances and analyses. Rather, the section shows how energy fits into the environmental accounting frame and how such accounting may contribute to assessing the sustainability of economic performance and growth.

2. THE CONCEPTS AND METHODS OF INTEGRATED ENVIRONMENTAL AND ECONOMIC ACCOUNTING

2.1 Material Flow Accounts: Assessing the Physical Metabolism of the Economy

Figure 2 can be viewed as the accounting interpretation of the generic environment-economy interface of Fig. 1. It represents the material flow accounts (MFA), which Robert Ayres and others have characterized as picturing the physical metabolism of society. The MFA show material inputs from abroad and the natural environment, their accumulation as stock in infrastructures and other durable goods, and their output of wastes and residuals into the environment and (as exports) to other countries. The MFA were developed in Austria and Germany and mostly applied in industrialized countries. The statistical office of the European Union, Eurostat, prepared guidelines on economy-wide MFA for improving the international comparability of the accounts.

Adding up the material flows (in tons) obtains various indicators at different stages of throughput as shown in acronyms in Fig. 2. Table I elaborates on the definition and contents of these indicators. Perhaps best known is the total material requirement (TMR). The indicator is defined as the sum of primary material inputs, including their ecological rucksacks. These rucksacks are hidden flows, which do not become a physical part of any product but are associated with its production, use, and disposal (e.g., soil erosion from agriculture or forestry, or gangue from mining).

Scholars of the German Wuppertal Institute, who focused their work on compiling and analyzing material flow accounts, have argued that their indicators measure actual and potential, and possibly still unknown, environmental impacts. A downward trend of the TMR, together with the usual upward trend of economic output (GDP), would thus indicate a decoupling of environmental pressure from economic growth. This can be seen as a process of dematerialization of the economy, and hence of ecological sustainability, as defined previously.

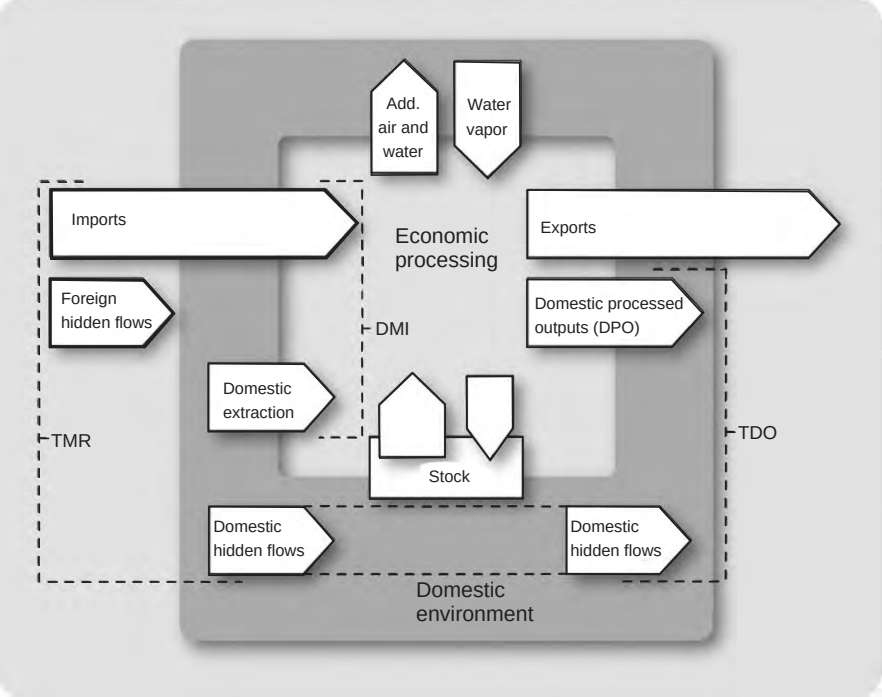


FIGURE 2 Material flow accounts and indicators. Adding up material flows at various stages of throughput through the economy obtains different indicators of material input, accumulation, and output. From Matthews *et al.* (2000), "The Weight of Nations: Material Outflows from Industrial Economies," Fig. 1, p. 5, World Resources Institute, Washington, DC, with permission.

The question is, how much dematerialization makes economic growth sustainable? Since physical material flows are not expressed in monetary units and are therefore not directly comparable to economic output or income, we can answer this question only by setting standards or targets for desirable rates of economic growth and necessary throughput reduction. The former president of the Wuppertal Institute, Ernst von Weizsäcker advanced the "Factor 4" standard of halving material input into the economy while doubling wealth over a period of 3 to 5 decades. Factor 4 and other standards of even a higher order can be interpreted as dematerialization targets for global sustainable development.

Overall indicators like the TMR are relatively crude measures of environmental impact since they use the weight of material inputs for weighting the significance of potential environmental effects. Table I shows therefore also various output indicators of wastes and emissions. They do weigh impacts by weight but are closer to specific environmental pressures and impacts. Their decline is indicative of the detoxification of the economy as they address the sink/pollution concern of sustainability.

2.2 Linking Physical and Monetary Accounts

The MFA were not developed in full consistency with economic concepts and measures. Their purpose was, at least initially, to provide an alternative to misleading—with regard to the human quality of life—accounting aggregates such as GDP. As a result, the MFA differ from standard and extended (see Section 2.3) economic concepts and indicators, notably with regard to the following:

- The narrow focus on material flows, excluding stocks and stock changes of produced and natural assets
- A corresponding inability to assess permanent losses (depletion and degradation beyond regeneration) of natural assets by flow indicators such as the TMR—an essential concern of sustainability
- A focus on different categories of materials by the MFA, rather than on the economic sectors responsible for the consumption of natural resources and the emission of residuals
- The assumption that material inputs reflect actual and potential environmental impacts, as opposed

TABLE I

Material Flow Balance and Derived Indicators[a]

Inputs (origin)	Outputs (destination)
Domestic extraction	Emissions and wastes
• Fossil fuels (coal, oil, etc.)	• Emissions to air
• Minerals (ores, gravel, etc.)	• Waste landfilled
• Biomass (timber, cereals, etc.)	• Emissions to water
	Dissipative use of products
Imports	• (Fertilizer, manure, compost, seeds, etc.)
Direct material inputs (DMI)	Domestic processed output (DPO) to nature
Unused domestic extraction	Disposal of unused domestic extraction
• From mining/quarrying	• From mining/quarrying
• From biomass harvest	• From biomass harvest
• Soil excavation	• Soil excavation
Total material input (TMI)	Total domestic output (TDO) to nature
	Exports
	Total material output (TMO)
Upstream flows associated with imports	
Total material requirement (TMR)	Net additions to stock (NAS)
	• Infrastructures and buildings
	• Other (machinery, durable goods, etc.)
	Upstream flows associated with exports

Source. Bringezu (2002). Material flow analysis—unveiling the physical basis of economies. *In* "Unveiling Wealth—On Money, Quality of Life and Sustainability." (P. Bartelmus, Ed.), Table IV. 2, p.119. Kluwer Academic, Dordrecht, Boston, and London.

[a]Excludes water and air flows (unless contained in other materials).

to the (cost value of) actual impacts recorded for a specific accounting period

- The use of physical (mass) units of measurement rather than the monetary (price and cost) values of economic accounts.

Facing mounting criticism of the conventional accounts and pressure to modify (to "green") these accounts for environmental impacts, national accountants took first steps toward incorporating readily available environmental statistics into their accounting frameworks. To this end, they made use of—physical, Leontief-type—input-output tables whose monetary counterparts are already a part of the national accounts system. The result is a relatively weak integration of environmental data either into a purely physical input-output table (PIOT) or as an add-on of physical indicators into the conventional accounts. The latter generates a hybrid physical-monetary National Accounting Matrix including Environmental Accounts (NAMEA).

German statisticians compiled a comprehensive PIOT for 1990 and updated the tabulations for 1995 (but did not publish them so far). The German PIOT is basically a sectoralization of the compound MFA: it specifies the material and substance flows as inputs and outputs in a detailed sectoral breakdown. Stock/asset accounts of produced capital goods and natural resources are omitted, as are difficult-to-assess-and-allocate ecological rucksacks. Supplementary subtables present energy balances in calorific values (terajoules) and in terms of inputs, outputs and transformations of energy carriers.

The Dutch NAMEA goes further by linking physical with monetary data. It stops short of any monetary valuation of environmental impacts by juxtaposing environmental statistics and indicators next to the economic activities responsible for environmental impacts. Like the PIOT, the NAMEA incorporates impacts on the source and sink functions of the environment in its supply (output) and use (input) accounts, as shown in Table II. In addition, the NAMEA presents some further weighting (beyond the use of tons) and aggregation for national environmental themes. The themes include energy (in calorific values), acidification, eutrophication, greenhouse effect, ozone layer depletion, and loss of natural resources (of gas and oil).

Table II is a simplified structure of a hybrid accounting system, based on the NAMEA approach. It presents the conventional economic accounts of production, consumption, capital accumulation, income generation, and international trade. Below and to the right of these accounts (separated by dashed lines), material and substance flow accounts are added to the economic transaction rows and columns of the economic accounts. The NAMEA can thus be seen as a PIOT linked to a monetary input-output table with some further supplementary indicators for selected environmental policy themes.

2.3 SEEA: Integrating Physical and Monetary Accounts

Physical accounts do not have the integrative power of the monetary national accounts. Weighting

TABLE II

Simplified Structure of a NAMEA

	Monetary accounts				Physical accounts	
Use / Supply	Industries 1, 2, . . .	Households (public and private)	Capital	Rest of the world (ROW)	Material flows (residuals, natural resources)	Policy themes, loss of natural resources
Products	Intermediate consumption	Household consumption	Gross capital formation	Exports	Emissions from industries	
Consumption of households					Emissions from households	
Capital	Capital consumption			Capital transfer from ROW	Other volume changes (discoveries etc.)	
Income generation	Value added, NDP					
Rest of the world (ROW)	Imports		Capital transfer to ROW	Balances	Transboundary pollutants to ROW	
Material flows (residuals, natural resources)	Natural resource inputs		Contributions to policy themes, loss of natural resources	Transboundary pollutants to ROW		Contributions to policy themes, loss of natural resources

Source. Adapted from Keuning and de Haan (1998), Table 1, pp. 146–147.

environmental impacts by weight of substance and resource flows may alert us to trends of overall environmental pressures and serve the management of particular natural resources; physical accounts cannot, however, compare the significance of environmental problems with the environmental and economic costs and benefits, as perceived by economic agents. Calculating ratios of resource productivity (GDP/TMR) or ecoefficiency (environmental impacts per GDP) is the most one can do to relate physical impacts to economic output. Monetary valuation appears to be the only possibility of fully integrating environmental concerns into the economic accounting system while ensuring consistency and comparability of "greened" with conventional economic indicators.

Pricing "priceless" environmental source and sink functions that are not traded in markets has its own problems. National accountants, especially from industrialized countries, have therefore been quite recalcitrant to implementing environmental accounts in monetary terms. They seem to believe that they might lose some of their longstanding goodwill if they let in new controversial concepts and valuations even through supplementary, so-called satellite accounts. While some accountants now favor the incorporation of the cost of depletion of economic natural resources (like timber, oil, or fish; see the appendix to this article) into the conventional accounts, many still consider the costing of environmental externalities a matter of modeling.

Nonetheless, the relatively new system for integrated environmental and economic accounting (SEEA), advanced by the author and his colleagues at the United Nations Statistics Division, did tackle the valuation problem. The objective was to incorporate environmental concerns into a comprehensive, integrated national accounts system. The basic methods of market valuation, maintenance costing, and contingent (damage) valuations are briefly described in the appendix. Their objective is to overcome major flaws of the conventional (monetary) accounts, in particular the following:

- Neglect of new scarcities of natural resources, which are a threat to sustained productivity of the economy
- Exclusion of environmental degradation as an externality of economic activity
- Accounting for environmental protection expenditures as national income and product when such outlays could be considered as a maintenance cost of society.

Figure 3 shows how environmental functions can be consistently integrated in the national accounts by doing the following:

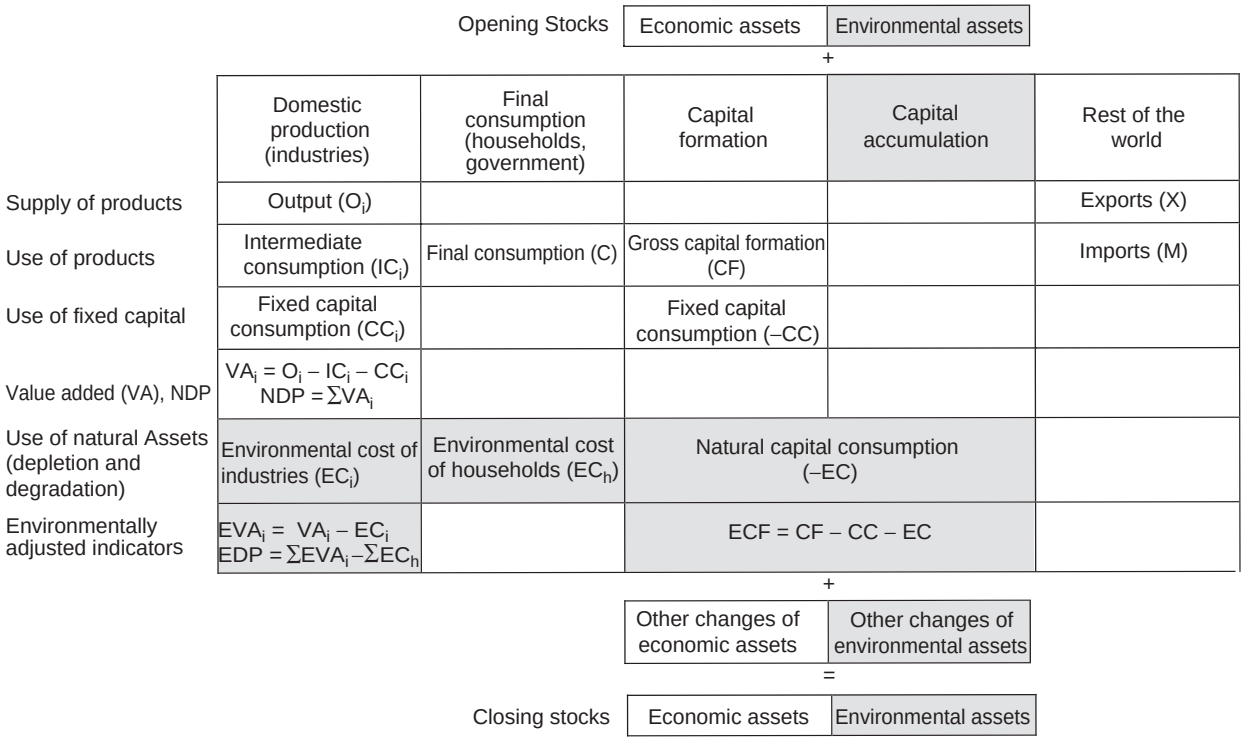


FIGURE 3 SEEA: structure and main indicators. The integrated accounting system incorporates environmental and natural – economic assets into the conventional flow and stock accounts and defines new environmentally adjusted ("green") economic indicators. From Bartelmus (1999), "Greening the National Accounts, Approach and Policy Use," DESA Discussion Paper No. 3, Fig. 3, p.7. United Nations, New York.

- Incorporating environmental assets as natural capital—in addition to produced (fixed) economic capital—into the asset accounts of the SNA
- Costing natural capital consumption (EC) in addition to and in consistency with fixed capital consumption (CC) in both the asset, and supply and use accounts
- Maintaining the accounting identities for the definition and calculation of the key economic (value added, VA, net domestic product, NDP, capital formation, CF), and environmentally adjusted (EVA, EDP, and ECF) indicators.

Environmental protection expenditures could be shown explicitly as "thereof" items of the supply and use of products in Fig. 3. The SEEA identifies these expenditures in appropriate classifications but does not suggest their deduction from GDP as a defense against welfare losses.

Accounting for natural capital consumption and maintenance (through reinvestment) extends the sustainability notion, as already built into the conventional net indicators of value added and capital formation, into the field of environmental services for the economy. In analogy to the wear and tear (i.e., the ultimate destruction) of capital goods in the production process, natural capital loss is defined as the permanent loss of natural resource stocks and waste absorption capacities. This definition clarifies the content of physical depletion and degradation—beyond natural regeneration and replenishment: the regeneration of nature can be seen as a cost-free natural repair process rather than as a capital loss.

Moreover, the costing of capital maintenance avoids controversial damage (to health and ecosystems) valuation (see the appendix). It can be argued, from an economic, utility maximizing optimality point of view, that such welfare effects are the correct values for environmental cost internalization into the budgets of households and enterprises. However, the purpose of national accounting is not welfare analysis or modeling of optimal behavior but the ex-post assessment of economic performance.

The revised SEEA takes a different, somewhat ambiguous view on welfare and other valuations. On the one hand, the monetization of the use of environmental assets is deemed to be a case of modeling a so-called green GDP; on the other hand,

damage valuation, as applied in cost-benefit analyses, is extensively discussed. As a consequence, the revised SEEA trades much of its original systemic character for flexibility in a modular, issue-oriented approach.

3. ACCOUNTING FOR ENERGY AND RELATED IMPACTS

Energy resource scarcities and environmental impacts from energy production and use are sometimes presented as the principal environmental problem; however, they do not tell the whole sustainability story. The previously described, accounting systems tell a more comprehensive story. This section illustrates, therefore, how the part of energy can be assessed within the full network of economic-environmental interactions. A detailed technical description of energy statistics and balances is thus beyond this article and would also be more relevant for the management of selected energy categories than for overall environment and sustainability policies.

3.1 Physical Energy Accounts

Physical energy balances are an established tool of energy statistics. They present the origin and destination of energy flows in calorific values or other energy equivalents. In accounting terms, such balances are "supply and use accounts," which can be made compatible with the national accounts. Typically they ignore, however, a number of environmental concerns, including the following:

- Important new and renewable energy sources, as well as nutritional energy flows, needed for getting the full picture of what Helmut Haberl calls "society's energetic metabolism"
- Environmental and social impacts of energy production and use such as resource depletion, emissions, land use changes, resettlement for hydropower development, and accidents.

To some extent, the previously described physical accounts incorporate the depletion of energy sources and the environmental impacts of energy production and use by appropriate coverage and classifications of energy carriers and substance flows. Perhaps best suited for detailed energy accounting in this regard are the physical input-output tables (PIOT). Table III illustrates the approach taken in supplementary energy tables of the PIOT. The table can also be interpreted as an elaboration of the physical part of the hybrid NAMEA (Table II) with regard to energy inputs and outputs, measured in calorific values. However, the PIOT do not relate emissions to energy supply and use but list them separately for the standard economic sectors. As in the NAMEA, the flows of residual substances could of course be added as outputs of energy production and consumption and as inputs into nature and the rest of the world.

The flow accounts of the PIOT do not show stocks of produced or natural capital, but stock changes only, as accumulation. As a consequence, periodic inputs and outputs balance only in the nonaccumulative flows for industries and households. The last row of Table III indicates an imbalance for produced and nonproduced natural assets and foreign trade (i.e., a net accumulation in the inventories of energy products and cultivated natural assets) and for nonproduced energy raw materials. The trade balance is negative, reflecting a net import of goods and services.

Physical flow accounts do not measure the ultimate objective of energy use, namely the production of energy services like heating, transportation or work. As a proxy, the energy PIOT includes various categories of final, or useful, energy. Most energy accounts exclude, however, the nutritional energy supply to humans and animals, needed for work and drive power, as well as hidden energy flows generated outside the boundary of the particular energy system. Inclusion of these flows would permit a comprehensive assessment of the energetic metabolism of societies as a complement to the industrial metabolism measured by material flow accounts.

3.2 Integrated Physical and Monetary Energy Accounts

Section 2.3 and Fig. 3 described the integration of physical and monetary accounts in the SEEA. The revised SEEA focuses largely on the physical side, at least as far as (noneconomic) environmental assets (see the appendix) are concerned. Its flow accounts thus cover energy supply and use in a similar fashion as PIOT and NAMEA. Both systems play in fact a dominant part in the revised SEEA. Considering, however, that the main reasons for greening the (monetary) national accounts have been environment-related deficiencies of the established accounts, the following brief discussion of energy stocks and flows is based on the more integrative original SEEA and its operational handbook.

TABLE III

PIOT for Energy Balances[a]

Supply	Industries 1, 2 etc.	Accumulation		Household consumption	Rest of the world (ROW)
		Produced assets	Nonproduced natural assets		
Inputs					
Raw materials[b]	x[c]				
Products[d]	x	x[e]		x	x
Waste for incineration	x				
Useful energy to nature[f]			x		
Energy losses to nature			x		
Outputs					
Raw materials[b]			x		
Products[d]	x	x[g]			x
Waste for incineration					
Useful energy [f]	x			x	
Energy losses	x			x	
Inputs minus outputs		x(+)	x(+)		x(−)

Source. Based on a case study for Germany (Stahmer, Kuhn, and Braun, 1998, Tables 4 and 5).
[a] Unit of measurement: calorific value, for example, joule or ton of oil equivalent (toe).
[b] Energy carriers and hydropower.
[c] Energy production (electricity, mining).
[d] Fuelwood, electricity, gas, coal, oil, natural gas, chemical and petroleum products, etc.
[e] Change in stocks.
[f] Light, heat.
[g] Cultivated natural assets (forestry, fishery).

The main advantages of such accounting over physical energy balances are the following:

- Coverage, definition, and valuation of energy stocks (oil deposits, standing fuel wood, etc.)
- Clear definition and costing of changes in stocks, notably the depletion of renewable (biotic and aquatic) and nonrenewable (fossil) energy resources as capital consumption
- Comparability of energy stocks and flows (stock changes)
- Comparability of conventional economic indicators of the energy sector (capital formation, value added, cost, income, etc.) with environmentally (cost-)adjusted indicators.

Assessing the current and future availability of energy resources, and hence of energy stocks and consumption, is a key concern of any sustainability analysis from the "capital maintenance" point of view (see Section 1.2). A number of categories and definitions used in the SEEA cater to this concern and are particularly relevant to energy carriers. For mineral and energy (subsoil) resources, the so-called McKelvey box categorizes physical stocks according to the probability of their existence and the economic feasibility of their extraction. The national—economic—accounts include proven reserves only (as opposed to probable, possible, or speculative resources), which are deemed profitable to exploit. The SEEA adopts the same definition of proven reserves to ensure consistency in (depletion) cost accounting and corresponding net indicators. Separate tables could of course present asset stocks weighted by probability (of existence) factors.

Subsoil energy carriers are nonrenewable resources. Any extraction reduces by definition the total availability of the resource, regardless of any potential discoveries of new deposits, and hence depletes the resource. The U.S. Bureau of Economic Analysis (in a one-time case study) disagreed, however, and treated discoveries as production (development) of a capital good. This approach offsets largely the extraction of the nonrenewable resource, turning it statistically into a renewable one. As a result, the total value of stocks of mineral

reserves has not changed much in the United States since 1958 (in constant prices). Further noteworthy issues, related to the accounting for nonrenewables and dealt with in the forthcoming revised SEEA, are the treatment of decommissioning costs (e.g., of nuclear power plants) as anticipated ("leveled-off") capital consumption, and of subsidized assets (with negative rents) as noneconomic, zero-value stocks.

The assessment of the depletion of renewable (energy) resources is more complex, since processes of natural regrowth of biotic (fuel wood) and replenishment of cyclical (water) resources, as well as capital formation (afforestation) need to be accounted for. The idea is to obtain the net amount of the permanent loss of the resource—beyond its sustainable yield—which should be costed as nonsustainable capital consumption. The determination of sustainable yield requires complex modeling for specific yield uses such as fuel wood from timber production, or water use for hydropower or as a coolant. In practice, the overall change in the physical amount and value of a resource between opening and closing stocks is frequently taken as a proxy for its depletion in physical and monetary terms. For example, the sustainable cut of forests is usually measured as the net increase in forest biomass, accounting for felling, natural losses from disease or fire, natural regrowth, and reforestation. This simplification neglects, however, changes in the composition and age structure of the forest that may affect future yields and economic returns.

Similarly, changes in the assimilative capacity of environmental media, owing to natural and human-made restoration, should be taken into account when estimating the net emissions—from energy production and use—that cannot be safely absorbed by natural systems. Only these emissions overload nature's existing assimilative capacities and should be costed as (nonsustainable) environmental externalities (see the appendix).

As discussed earlier, the physical and monetary environmental accounts consider only the permanent loss of natural capital as capital consumption. Most new and renewable energy sources like wind, sun, or ocean, have not experienced such loss, and their *(in situ)* economic value can be assumed to be zero. Of course, this does not mean that there are no investment and operating cost for harnessing these energy sources. However, these are conventional expenses (capital formation, capital consumption, and operating cost) that are covered by the established national accounts. The SEEA, and notably its revised version, identifies these environmental activities as natural resource management and exploitation by means of additional regrouping and classification. These environmental expenditures are important for the management of the different resources but do not assess the core environmental concerns of depletion and degradation.

The purpose of monetary estimates of economic (private) and environmental (social) costs is to assess the environmental sustainability of economic activity. However, some energy analysts and ecological economists appear to share Malcolm Slesser's view that monetary valuation is "after all nothing more than a highly sophisticated value judgement." Instead, they suggest the use of energy values or potentials for measuring the natural or absolute value of goods and services—independent of volatile market preferences. Different valuation principles for material and energy flow accounting include the following:

- Embodied energy, the direct and indirect energy required to produce goods and services
- Exergy, the potential work content of materials, energy carriers, and residuals
- E*m*ergy, the energy memory of the total amount of energy of one kind, required directly and indirectly to make a product or service (advanced in the writings of the late Howard Odum)

The main critique of these energy valuations has been, in turn, on the neglect of individual value judgements or preferences of economic agents. Additional, more practical problems are the data and knowledge requirements for a myriad of different production and transformation processes that use energy and generate emissions and discharges. Energy value accounting does therefore not seem to be capable of capturing the overall interaction of environment and economy. Such accounting could provide useful supplementary information for energy policies and management; it is less likely, though, to become a tool of overall assessments of economic or ecological sustainability.

4. OUTLOOK: POLICY USE OF ACCOUNTING INDICATORS

Rapther than reviewing the wide range of uses of a multipurpose accounting system, this section leads us back to the initial main concern of green accounting—the assessment of sustainability.

Table IV provides a framework for sustainability accounting and analysis with regard to the two key operational, economic and ecological, sustainability concepts. The table sets out from the rationale and strategies of reducing material throughput through dematerialization and of sustaining economic growth through capital maintenance. The corresponding physical and monetary accounting tools and indicators measure relatively strong and weak sustainability, respectively. Compliance with minimum standards for the preservation of critical (irreplaceable) capital is to attain strong sustainability. On the other hand, monetary accounts and their adjusted economic indicators imply the substitutability of different capital categories and production factors. They aim to assess the—weak—sustainability concept of overall capital maintenance.

Reducing material flows into the economy, as expressed in particular by their sum total, TMR, aims at decoupling economic growth from the generation of environmental impacts. A collaborative study, published by the World Resources Institute, shows that TMR per capita has been leveling off for selected industrialized countries (Germany, Netherlands, the United States) at about 80 tons per annum, except for Japan at 45 tons (because of its low per capita energy use). Given that GDP per capita has been rising in the high-TMR countries, there is some—relative—delinkage from growth, albeit far from the sustainability targets of Factor 4 (see Section 2.1). At least in these countries, ecological sustainability has not been in sight during the accounting periods (1975–1993).

Unfortunately, there does not seem to be an automatic solution to delinking environmental impacts or pressures from economic growth as suggested by the Environmental Kuznets Curve hypothesis. The corresponding inverted U-curve correlates low environmental impacts with low levels of economic development (per capita income), increasing impacts with industrialization and decreasing impacts with affluence. The possible transition to a dematerialized service economy might explain the automaticity, but empirical confirmation (except for a few pollutants) is elusive. Environmental regulation, possibly supported by fiscal incentives (subsidies for resource-saving technologies and disincentives such as material consumption taxes) will therefore need to be applied to ensure the preservation of critical capital.

Costing produced and natural capital consumption is a means to reserve funds for reinvestment and hence capital maintenance. At the macroeconomic level, the deduction of these costs from GDP obtains EDP, a net indicator whose decrease would indicate nonsustainable economic growth. Case studies of the SEEA, as presented, for instance, by Kimio Uno and Peter Bartelmus, did not show a reversal of (upward) GDP trends by EDP. However, this does not necessarily confirm sustainability because of the relatively short time series available.

A more pertinent way of looking into the sustainability of economic performance would be

TABLE IV

Concepts, Indicators, and Policies of Environmental Sustainability

	Ecological sustainability: Dematerialization/detoxification	Economic sustainability: Produced and natural capital maintenance
Rationale	Reducing throughput below carrying capacities	Sustaining economic growth
Strategy	Decoupling economic growth from environmental pressure	Maximizing economic growth while keeping produced and natural capital intact
Accounting tools	Material flow accounts (MFA), physical input-output table (PIOT), national accounting matrix including environmental accounts (NAMEA)	System for integrated environmental and economic accounting (SEEA)
Key indicators	TMR, TDO, DMI, TMC, resource productivity: GDP/TMR	"Green" economic aggregates: EDP, EVA, ECF, total wealth
Strength of sustainability	Strong: reduction of material and substance flows to meet sustainability standards (Factor X targets), maintenance of critical capital	Weak: overall capital maintenance, allowing substitution between produced and natural capital and other production factors
Policy instruments	Regulation and standard setting (Factor X, emission standards), fiscal (dis)incentives for material-saving innovation	Use of green indicators and variables in economic analysis and policy; market instruments of environmental cost internalization

Source. Adapted from Bartelmus, P. (2003), abridged version of Table 1, p.68, with permission.

to measure a nation's ability to generate new capital after taking produced and natural capital consumption into account. The Environmentally adjusted Net Capital Formation (ECF) indicator showed a significant difference between some developing and industrialized countries. For example, Indonesia, Ghana, and Mexico at times exhibited nonsustainability by using up more capital than generating new one (i.e., they showed an overall disinvestment in produced and natural assets). Industrialized countries, on the other hand, did increase their capital build-up and met thus one necessary condition for sustaining their economies.

Past overall capital maintenance or increase hide, however, the fact that in the long run complementarities of natural capital might make it impossible to maintain production and consumption patterns and growth rates. The empirical testing of this assumption should be an important field of sustainability research.

At the microeconomic level, one important use of environmental (social) cost accounting is to determine the level at which market instruments of cost internalization such as eco-taxes should be initially set for different economic agents according to polluter/user-pays-principles. The ultimate effects of environmental cost internalization on production and consumption patterns and overall economic growth and employment need, of course, be modeled with the usual restrictions and assumptions. The results of such modeling might point to the need of resetting the original levels of market instruments according to their predicted efficacy.

Greened national accounts and PIOTs are a particularly useful tool for integrative environmental-economic modeling since they generate a high degree of transparency with regard to the underlying data and indicators. This applies also to the manifold energy scenarios (and related climate change models), which have predicted or denied reaching the limits of energy supply and, possibly, economic growth. They have also generated a wide range of policy options according to differing coverage, modeling techniques, and policy targets. The International Institute for Applied Systems Analysis (IIASA) and the World Energy Council have presented concise overviews of energy scenarios on their Web sites.

As already mentioned, greened national accounts can only capture the immediate interaction between environment and economy. A broader assessment of sustainability would have to take account of social, institutional, and political aspects of sustainable development. In the field of energy, the disparities in the distribution of and access to scarce nonrenewable resources have triggered aggressive policies of securing affordable energy supply. Assessing natural resource availability, ownership, and trends of consumption in different regions and nations could alert us to looming resource scarcities and to possible international conflicts in resource use. It might also be the first step toward more rational policies of ensuring energy security, and perhaps security, for all.

APPENDIX: MONETARY VALUATION IN THE SEEA

Market valuation uses prices for natural assets, which are observed in the market. It is usually applied to economic assets of natural resources. The SNA defines economic assets as "entities (a) over which ownership rights are enforced … and (b) from which economic benefits may be derived by their owners." Figure 3 displays therefore part of natural capital consumption under the column of economic assets.

Where market prices for natural resource stocks, such as oil deposits or fish in the ocean are not available, the economic value of these assets can be derived from estimates of the—discounted—sum of life-long net returns, obtained from their potential use in production. It is at this value that a natural asset would be traded if a market existed for the asset. Market valuation techniques are also applied to changes in asset values, caused in particular by depletion (i.e., their nonsustainable use). These value changes represent losses in the income-spinning capacity of an economic asset. Depletion cost allowances reflect thus a weak sustainability concept, calling for the reinvestment of environmental cost in any income-generating activity.

Maintenance valuation permits the costing of the degradation of noneconomic—environmental—assets, in particular from emissions. Typically these externalities are not traded in markets. The SEEA defines maintenance cost as those that "would have been incurred if the environment had been used in such a way as not to have affected its future use." Maintenance costs refer to best-available technologies or production processes with which to avoid, mitigate, or reduce environmental impacts. They are used to weight actual environmental impacts, generated during the accounting period by different

economic agents. As with depreciation allowances for the wear and tear of produced capital, such costing can be seen as the funds required for reinvesting in capital maintenance.

The SEEA also discusses contingent and related damage valuations. These valuations are applied in cost-benefit analyses of particular projects and programs but are hardly applicable at the national level. They refer to the ultimate welfare effects (damages) of environmental impacts, which are incompatible with the market price and cost conventions of the national accounts. The original SEEA considered damage valuation in environmental accounting as experimental, and its operational handbook discourages its application.

SEE ALSO THE FOLLOWING ARTICLES

Cost–Benefit Analysis Applied to Energy • Development and Energy, Overview • Economic Growth and Energy • Environmental Change and Energy • Environmental Kuznets Curve • Goods and Services: Energy Costs • Sustainable Development: Basic Concepts and Application to Energy • World Environmental Summits: The Role of Energy

Further Reading

Ayres, R. U., and Ayres, L. W. (Eds.). (2002). "A Handbook of Industrial Ecology." Edward Elgar, Cheltenham, United Kingdom, and Northampton, MA.

Bartelmus, P. (2001). Accounting for sustainability: Greening the national accounts. *In* "Our Fragile World, Challenges and Opportunities for Sustainable Development" (M. K. Tolba, Ed.), Forerunner to the Encyclopedia of Life Support Systems, Vol. II. EOLSS, Oxford.

Bartelmus, P. (2003). Dematerialization and capital maintenance: Two sides of the sustainability coin. *Ecol. Econ.* **46**, 61–81.

Eurostat (2001). "Economy-wide Material Flow Accounts and Derived Indicators, a Methodological Guide." European Communities, Luxembourg.

Haberl, H. (2001). The energetic metabolism of societies, part I: Accounting concepts. *J. Industrial Ecology* **5**(1), 11–33.

Kapp, K. W. (1971, and previous editions: 1950 and 1963). "The Social Costs of Private Enterprise." Schocken Books, New York.

Keuning, S. J., and deHaan, M. (1998). Netherlands: What's in a NAMEA? Recent results. *In* "Environmental Accounting in Theory and Practice" (K. Uno and P. Bartelmus, Eds.). Kluwer Academic, Dordrecht, Boston, and London.

Odum, H. T. (1996). "Environmental Accounting, Energy and Environmental Decision Making." Wiley, New York.

Slesser, M. (1975). Accounting for Energy. *Nature* **254**, 170–172.

Stahmer, C., Kuhn, M., and Braun, N. (1998). "Physical Input-output Tables for Germany 1990." Eurostat Working Paper No. 2/1998/B/1. Eurostat, Luxembourg.

United Nations (1993). "Integrated Environmental and Economic Accounting." United Nations, New York.

United Nations (2000). "Integrated Environmental and Economic Accounting—An Operational Manual." United Nations, New York.

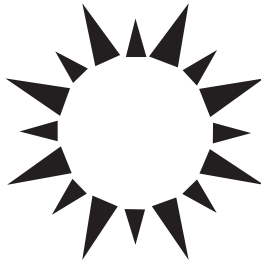
United Nations, *et al.* (1993). "System of National Accounts 1993." United Nations and others, New York and other cities.

United Nations, *et al.* (in prep.). "Integrated Environmental and Economic Accounting 2003." United Nations and others, New York and other cities.

Uno, K., and Bartelmus, P. (Eds.). (1998). "Environmental Accounting in Theory and Practice." Kluwer Academic, Dordrecht, Boston, and London.

World Resources Institute (WRI), *et al.* (1997). "Resource Flows: The Material Basis of Industrial Economies." World Resources Institute, Washington, DC.

Greenhouse Gas Abatement: Controversies in Cost Assessment



MARK JACCARD
Simon Fraser University
Vancouver, British Columbia, Canada

1. **Introduction**
2. **Alternative Definitions of Cost**
3. **Uncertainties and Assumptions about Technology and Preference Change**
4. **Assessment**

Glossary

autonomous energy efficiency index (AEEI) The rate at which the economy becomes more or less energy intensive, with the price of energy remaining constant relative to the prices of other inputs. Indicates productivity improvements in energy use that are not caused by increases in energy prices.

bottom-up analysis (model) An analysis (or model) that focuses on the financial costs of technologies that, if widely adopted to meet the energy service needs of firms and households, would lead to reductions in greenhouse gas emissions. Tends to estimate low costs of greenhouse gas abatement.

computable general equilibrium (CGE) model A top-down model of the economy that includes all of its major components (consumers, producers, government, external agents) and markets (capital, labor, materials, energy, interindustry trade, final goods and services, government revenues and expenditures, international trade) and the relationships between them. Such a model can compute the new general equilibrium that results after a perturbation to one component or market, as would be caused by a greenhouse gas tax that affected the price of energy and perhaps other goods and services.

consumers' surplus A premium that consumers are willing to pay for a good or service above what they actually pay for it (its market price).

economies of learning The reductions in financial costs (capital and operating) of technologies (especially newer technologies) as manufacturers and operators gain experience with them with greater production. The rate of learning, and thus cost decline, is sometimes referred to as an "experience curve."

economies of scale The reductions in financial costs (capital and operating) of technologies as larger units are built or a greater number of units are manufactured on a regular basis. Can occur with old and new technologies.

elasticities of substitution (ESUB) Changes in the relative use of inputs (capital, labor, energy, materials) in response to changes in their relative costs. A high elasticity implies an easy ability to switch away from an input, such as energy, as its relative price increases, and a low elasticity implies the opposite.

financial costs The monetary costs that are used by engineers and financial planners to estimate the initial and ongoing capital outlay required by an investment. Unless preceded by a term such as "expected," financial cost estimates exclude consideration of financial risks, or intangible costs and benefits (e.g., consumers' surplus).

hybrid analysis (model) An analysis (or model) that combines the technological explicitness of the bottom-up approach with estimation of consumer and firm behavior of the top-down approach.

option value The increase in expected net benefits resulting from the decision to delay an irreversible investment under uncertainty. While this definition is predominant in energy efficiency literature, there are alternative definitions for this term.

top-down analysis (model) An analysis (or model) that focuses on aggregate relationships between inputs and outputs of the economy, and the interplay of these through price feedback mechanisms. The relative costs and uses of inputs (capital, labor, energy, and materials) are portrayed in production functions for firms and in consumption or utility functions for households. A "computable general equilibrium model" (CGE) is a type of top-down model. Top-down models tend to estimate high costs of greenhouse gas abatement, although this depends on the values of their elasticities of substitution and autonomous energy efficiency index parameters.

 © 2004 Elsevier Inc. All rights reserved.

In deciding how and by how much to reduce greenhouse gas emissions, policy makers need to know the potential cost. But there are alternative definitions of cost, the evolution of technologies and preferences is a complex and highly uncertain process that can profoundly influence estimates of long-run cost. For these and other reasons, independent researchers and interest groups have produced widely divergent estimates of costs, and policy makers do not know whom to believe. The different analytical and modeling approaches are discussed here in order to explain their relative contribution to the differences in greenhouse gas reduction cost estimates. Each contrasting cost definition and contrasting assumption of technical and preference evolution informs the estimates. The focus is on energy-related technologies, because these produce the greatest anthropogenic contribution to rising atmospheric concentrations of greenhouse gases, but the issues are equally relevant to all technologies and management practices that reduce greenhouse gas emissions, whether involving forestry, agriculture, treatment of municipal solid waste, or capture and storage of CO_2.

1. INTRODUCTION

In 1991, M. Grubb referred to "the missing models" in arguing for the development of a new generation of models that would better inform policy makers of the critical issues in estimating greenhouse gas (GHG) reduction costs. To be useful to policy makers, models for costing GHG reduction need to explain the key factors behind divergent cost estimates. But over a decade later, cost estimates still diverge widely, in part because of different assumptions about technology and preference evolution (which is perhaps inevitable) but also because of how costs are defined and depicted in GHG costing models. "Bottom-up" models tend to ignore or underrepresent important constraints to the adoption of new equipment, whereas "top-down" models incorporate these constraints but inadequately represent the potential for technological innovation and preference changes that can reduce costs in the long run. Newer "hybrid" models attempt to incorporate key strengths of both top-down and bottom-up approaches in order to be more useful to policy makers who face critical decisions about the magnitude and timing of GHG reduction targets.

2. ALTERNATIVE DEFINITIONS OF COST

An action to reduce GHG emissions is defined as a change in equipment choice, equipment use rate, lifestyle, or resource management practice that changes GHG emissions from what they otherwise would be. Examples are: choosing more efficient light bulbs, turning off unused lights, telecommuting some days of the week, and practicing forestry or agriculture differently. The cost of an action is the difference in costs between the business-as-usual scenario and a world in which the action is undertaken. Unfortunately, analysts apply alternative definitions of these costs.

2.1 Bottom-Up Cost Definition and Critique

Bottom-up analysis, applied frequently by engineers, physicists and environmental advocates, estimates how changes in energy efficiency; fuel type; emission controls on equipment, buildings, and infrastructure; and even land-use practices might lead to different levels of GHG emissions. Technologies (furnaces, light bulbs, electric motors, vehicles) that provide the same energy service (space heating, lighting, industrial motive force, personal mobility) are generally assumed to be perfect substitutes except for differences in their financial costs and their emissions of GHGs and other pollutants. When their financial costs (capital and operating) in different time periods are converted into present value using a social discount rate, many current and emerging technologies available for reducing GHG emissions appear to be profitable or just slightly more expensive relative to existing equipment. Bottom-up analyses often show, therefore, that substantial GHG emission reduction can be profitable or low cost if the low-emission technologies were to increase from their small market share to achieve market dominance.

Although they are all characterized by a focus on technologies and by cost definitions that are restricted to financial costs, bottom-up models can vary in terms of how they depict the interconnections between different decisions in the economy. A simple bottom-up analysis would assess each energy service individually (home heating, commercial lighting) and conduct a financial analysis to determine the least-cost technology for each. A fully integrated, optimizing bottom-up model would determine the least-cost technology for all services simultaneously. This

integrated approach may indicate a very different outcome if the model shows, for example, that dramatic reductions in GHG emissions from the electricity sector are inexpensive. If these are pursued, then a drop in the GHG content of electricity will reduce the GHG cost-effectiveness of electricity conservation actions in the end-use sectors. A well-known, integrated bottom-up linear programming model used in North America, Europe, and increasingly elsewhere is the MARKAL (market allocation) optimization model.

An indication of the bottom-up approach is provided by researchers assessing the costs to the United States of achieving the target it negotiated (but has not ratified) under the Kyoto Protocol of 1997. The U.S. government agreed to reduce its GHG emissions to 7% below their 1990 level by 2010. According to forecasts at the time, this represented a reduction of almost 30% from expected emission levels for 2010. Brown and coauthors at five major energy research laboratories funded by the U.S. government found that the United States could reduce to its 1990 level of emissions with no net cost to the U.S. economy, and thus virtually no impact on gross domestic product (GDP). This could be achieved with a host of policies, including a GHG tax of no more than $25/ton of carbon ($25/t C). All of the reductions would be attained domestically, even though the United States has the capability under the Kyoto agreement to purchase reduction credits from other countries.

Many economists, however, criticize the bottom-up approach in general, and this study in particular, for the assumption that a single, *ex ante* (anticipated) estimate of financial cost differences indicates the full social cost of switching technologies. Economists such as A. Jaffe and R. Stavins argue that technologies may differ to consumers and businesses in ways that are not represented by this single financial value. The mundane example of purchasing a light bulb illustrates these differences.

- New technologies usually have a higher chance of premature failure than do conventional technologies and therefore pose greater financial risk. There is some probability that *ex poste* (realized) financial costs will exceed *ex ante* financial costs for new technologies. New compact fluorescent light bulbs have exhibited higher rates of premature failure compared to conventional incandescent light bulbs, requiring a higher than expected financial outlay because of early replacement in some cases.
- Technologies with longer payback periods (relatively high capital costs) are riskier if the cumulative probability of failure or accident, or undesired economic conditions, increases over time. Because of the higher purchase cost of compact fluorescent light bulbs, the chance of accidental breakage prior to paying back the initial investment is higher than for an incandescent light bulb.
- Two technologies may appear to provide the same service to an engineer but not to the consumer. Many people find compact fluorescent light bulbs to be less than perfect substitutes for incandescent light bulbs in terms of attractiveness of the bulb, compatibility with fixtures, quality of light, and timing to reach full intensity, and would pay more to maintain high levels of these nonfinancial attributes.
- Not all firms and households face identical financial costs: acquisition, installation, and operating costs can vary by location and type of facility. Compact fluorescents may be more expensive to acquire in nonmetropolitan areas where the market is smaller. This heterogeneity means that a comparison of single-point estimates of financial costs may exaggerate the benefits of market domination by a low-GHG technology.

The first two differences relate to "option value," which is the expected gain from delaying or avoiding an irreversible investment while waiting for new information that might lead to a better decision. The third difference refers to what economists call "consumers' surplus," or the extra value that consumers realize above the financial cost of a particular technology. The fourth difference simply acknowledges "market heterogeneity," i.e., that a technology may be cheaper than its competitor in one circumstance but more expensive in another.

When consumers and businesses are asked, induced, or forced to switch away from a technology they would otherwise have chosen, economists say that the social cost of this switch is the difference in financial costs plus or minus any intangible costs related to option value, consumers' surplus, and market heterogeneity. By ignoring these values, bottom-up analysts may overestimate the willingness of firms and households to take actions that reduce GHG emissions, and thus underestimate the social cost. Politicians, in contrast, seem to be instinctively aware of these other values, and tend to question claims that GHG emissions can be reduced at little or no cost. In providing only part of the cost picture, the bottom-up approach is less helpful to policy makers than it could be.

Another challenge with the bottom-up approach is that its technology-specific focus hinders the ability to assess broader macroeconomic effects, notably the trade and structural repercussions resulting from changes in energy prices and costs in the economy. In this sense, bottom-up models usually provide only a "partial equilibrium analysis" of the response to GHG reduction policies. If, however, the cost of GHG reduction is as low as this approach often concludes, or if governments are unwilling to pursue rapid and costly decreases in GHG emissions, then macroeconomic feedbacks may be small enough that this partial equilibrium assessment is sufficient. But even if energy-focused GHG policies are too modest to elicit a significant macroeconomic change, they may cause unintended responses in energy markets that are outside the purview of a conventional bottom-up model. For example, subsidized or voluntary decisions to acquire more energy-efficient equipment (cars, lights, appliances, industrial machinery) reduce energy service operating costs, and may thus encourage greater consumption of energy services. The magnitude of this "rebound effect" is disputed, but certainly all sides agree that at some point lower operating costs foster higher use. Returning to the lightbulb example, the installation of outdoor decorative and security lighting, often using high-efficiency lightbulbs, is increasing in North America. Most bottom-up models do not, however, link the demand for energy services to the cost of the service, and thus would not simulate a rebound effect.

2.2 Top-Down Cost Definition and Critique

The alternative, top-down analysis of economists usually relies on historical and cross-sectional market data to estimate aggregate relationships between the relative costs and the relative market shares of energy and other inputs to the economy, and links these to sectoral and total economic output in a broader, equilibrium framework. Elasticities of substitution (ESUB) indicate the substitutability between any two pairs of aggregate inputs (capital, labor, energy, materials) and between energy forms (coal, oil, gas, renewables) within the energy aggregate. Another key parameter in such models, the autonomous energy efficiency index (AEEI), indicates the rate at which price-independent technological evolution improves energy productivity. Relatively high parameter values for AEEI and for energy-related ESUB (a large degree of substitutability between energy and capital, and between GHG-intensive and non-GHG-intensive forms of energy) equate to a relatively low cost of GHG emission reduction. Because these parameters are estimated from real market behavior, as energy prices and energy consumption have changed historically, they are said to reveal the actual preferences of consumers and businesses—and therefore implicitly incorporate losses or gains in option value and consumers' surplus, as well as reflect the market heterogeneity of real-world financial cost conditions.

With these key response parameters estimated, economists simulate the economy's reaction to an economic signal (a GHG tax, tradable GHG permits, technology subsidies, or some combination of these) that increases the relative cost of GHG-intensive technologies and behavior. The magnitude of the economic signal necessary to achieve a given GHG reduction target provides the implicit marginal cost of that target. Because top-down models usually integrate energy supply and demand within a total economy model, this marginal cost includes the feedback between energy supply and demand, on the one hand, and between these components and the economy as a whole, on the other. Top-down approaches are, therefore, more often associated with a "general equilibrium analysis" of policy effects; models designed expressly to close all key economic feedback loops are referred to as computable general equilibrium models (CGE models). Because movement away from current technologies includes all intangible value losses, top-down analysis usually generates substantially higher estimates of GHG reduction costs.

The top-down parameters, ESUB and AEEI, are usually estimated from historical data. Historical data generate relatively low values for the economy's willingness to substitute (ESUB) away from GHG-intensive technologies and activities, in part because full long-run price responses are difficult to detect in an ever-changing world. Experts can, however, judgmentally change the AEEI or ESUB parameter values to generate low-cost estimates, simply by portraying consumers and businesses as more prone to switch technologies in response to changes in financial costs—in effect, reducing the estimated option value losses, consumers' surplus losses, and market heterogeneity that otherwise explain technological inertia. But setting these parameters judgmentally undermines to some extent the very argument in favor of the top-down approach, by forgoing its real-world empirical basis. The future value of AEEI, being highly uncertain, is hotly disputed.

Top-down models are relatively easy to construct, and this may explain why there are so many variants in contrast to the dominance that MARKAL has attained among integrated bottom-up modelers. Some of the top-down models in the United States are G-Cubed, MERGE, SGM, and MIT-EPPA; all of these are computable general equilibrium models. Top-down models have also been applied to estimating the costs to the United States of achieving its Kyoto target, and these provide a contrast with the bottom-up estimate. Weyant and Hill summarized the results of a multimodel evaluation of the Kyoto protocol by several top-down models, especially the CGE type. Of the 11 participating models, 8 models found that the United States would require a carbon tax of at least $150/t C, and of these, 4 models required a tax higher than $250/t C (one higher than $400/t C). GDP impacts ranged from modest levels to the loss of over 3% of economic growth.

Like the bottom-up approach, however, the top-down approach is also vulnerable to the criticism of being unhelpful to policy makers. Substantial reduction of GHG emissions, in order eventually to stabilize atmospheric concentrations, is a long-run objective that requires dramatic technological change over a lengthy period in which consumer preferences are also likely to evolve. Policy makers need to know if and how their policies can influence long-run technological change, and thus the *ex ante* and *ex poste* financial costs of technologies, as well as the long-run evolution of consumer preferences, in ways that reduce these intangible losses from switching to less GHG-intensive technologies.

Even if the confidence intervals of the estimated ESUB and AEEI parameters are narrow, meaning that they are relatively robust at explaining past changes in energy demand, there is no guarantee that parameter values derived from past experience will remain valid into the future. Financial costs and consumer preferences change over time, and the magnitude and direction of this change may be influenced by policy. For example, until recently, there was no incentive to design and commercialize low-GHG technologies. Now, such technologies are under development worldwide. Azar and Dowlatabadi note that as research, development, and production expand, aided by government policy, "economies of scale" and "economies of learning" reduce financial costs and can reduce option values, perhaps dramatically. The market penetration of these new technologies and their falling costs imply higher future values for ESUB and AEEI, reflecting the economy's increasing ability to reduce GHG emissions at lower costs.

A similar logic applies to historical estimates of consumers' surplus losses. A typical pattern in new technology commercialization is an initial phase of slow market penetration, limited to acquisition by early adopters, followed by a market transformation phase in which expanding consumer awareness and acceptance precipitate a feedback loop of product improvement, intensified marketing effort, and accelerating dissemination. This shift in consumer preferences means lower consumers' surplus losses from technology switching, implying that the long-run social costs of GHG reduction may be partially influenced by policy. Again, this would cause higher long-run values for the all-encompassing ESUB and AEEI parameters of top-down models, and thus lower costs of GHG reduction. Returning to the light bulb example, the growing cumulative production of compact fluorescent light bulbs has reduced their production cost and market price over the past decade, which in turn has shortened investment paybacks and decreased the financial risks from premature breakage. Manufacturers have also improved the reliability and attractiveness of compact fluorescents (appearance, size, hue) so that they are now closer substitutes for incandescent light bulbs. Does this mean that the household market for light bulbs is on the verge of a market transformation in which compact fluorescent light bulbs overtake and marginalize incandescent light bulbs? It might. But an important caution is that low-GHG technologies are not guaranteed lower financial costs and growing consumer acceptance. Experienced marketers of new products are well aware of the difficulty of predicting financial cost reductions and future consumer preferences; the experience of every new technology is different. If the general trend of technological evolution, however, is for new technologies to increase the ability and willingness of consumers and businesses to opt for low-GHG choices, as is likely, then AEEI and ESUB values will increase and the long-run costs of GHG emission reduction will fall.

The importance of technological change for estimating future costs raises another challenge for the top-down approach in terms of its usefulness to policy makers. Because conventional top-down analysis conceives of technological change as an abstract, aggregate phenomenon, characterized by ESUB and AEEI parameter values, it only helps policy makers assess top-level policy instruments, such as taxes and tradable permits. But policy makers often prefer technology- and building-specific policies in the form of tax credits, subsidies, and regulations because these policies avoid the dramatic

(and unpopular) increases in energy prices that would otherwise be required to cause the degree of technological transformation for substantial GHG emission reductions. A model is thus more useful to policy makers if it can assess the combined effect of these focused policies with perhaps relatively modest economy-wide applications of taxes and tradable permits, but conventional, aggregated top-down models are unable to do this.

2.3 Compromise Hybrid Costing Approaches

In summary, the top-down and bottom-up approaches offer competing cost definitions, both of which have their failings in terms of usefulness to policy makers. The bottom-up approach ignores option values and consumer preferences. The top-down approach incorporates these cost factors, at an aggregate level, but ignores the potential for long-run changes to them as technology-specific and economy-wide policies stimulate technical change and perhaps shifting consumer preferences.

Researchers have tried to address these different strengths and weaknesses of top-down and bottom-up analysis by designing policy costing models that are both technologically explicit and behaviorally realistic—simulating the economy's long-run building and equipment stock turnover as a function of policies that target technological innovation and business and consumer preferences, while incorporating feedbacks between energy supply and demand, and between the energy system and the structure and output of the economy. Bohringer and others refer to this as a "hybrid modeling" approach, because it attempts to bridge the top-down/bottom-up schism. An example of a so-called hybrid model is the National Energy Modeling System (NEMS) of the U.S. government. Like a bottom-up model, NEMS is highly disaggregated, having substantial technological detail in its representation of energy supply and of the individual energy demand sectors (transportation, residential, commercial/institutional, and industry). But, unlike a typical bottom-up model, technology choices within these sectors are simulated based on behavioral parameters that are mostly estimated from real-world empirical data. Alternative vehicle choice in the NEMS transportation module, for example, is based on parameters estimated from surveys that asked consumers about vehicle options. Also, the model is close to general equilibrium in that a policy simulation equilibrates energy supply and demand, on the one hand, and this supply–demand composite with the macroeconomy as a whole, on the other hand.

With the hybrid modeling approach, the policy analyst can help the policy maker sort out the importance of differing definitions of cost for the estimates of GHG reduction costs. A hybrid model can first be applied in bottom-up mode, focusing just on financial costs and treating these in a deterministic manner, ignoring risk factors and intangible costs such as consumers' surplus losses. Then, its parameters can be adjusted to incorporate the estimated effects of these other factors. The policy maker has the opportunity to see the importance of these different cost definitions for the different policy cost estimates, and to test the cost assumptions that seem most plausible. For example, the U.S. Energy Information Administration applied NEMS in 1998 to estimate the costs of U.S. compliance with the Kyoto Protocol. It estimated that a carbon tax of $294/t C was required to reach the target with domestic actions alone and that this would cause a 3% loss in GDP growth. These results are within the range of top-down models in the multimodel evaluation, but distant from the bottom-up study of the U.S. national research laboratories. This is not surprising given that the 12-year period between 1998 and the Kyoto deadline year of 2010 provides insufficient time for policies to influence the evolution of new technology costs or consumer preferences. Because hybrid models and top-down models have the same definition of cost, they are likely to show similar costs of GHG reduction over a short-period simulation. Over a 20- or 30-year time frame, however, the cost estimates from a hybrid model could diverge significantly from those of a typical top-down model, if the hybrid included dynamic influences on technology costs and consumer preferences, as is explained in the following section.

3. UNCERTAINTIES AND ASSUMPTIONS ABOUT TECHNOLOGY AND PREFERENCE CHANGE

Even if the definition of costs were not in dispute, the challenge of GHG emission reduction presents enormous uncertainties that can result in widely divergent cost estimates. For a given country, three broad categories of uncertainty stand out.

- Uncertainty about external factors creates profound uncertainty for the GHG reduction costs of a given country. Will other countries make and keep

emission reduction targets, and if so at what cost and with what implication for competitive trade conditions? To what extent will a given country be allowed to utilize international permits and international actions to meet its commitment? What international emission permit trading mechanism will exist and what will be the price of permits?

- Uncertainty about technological innovation and its implications for future financial costs increases as the time horizon is extended. Over a 20- to 30-year time frame, dramatic technological change may occur with equally dramatic implications for the financial cost of GHG reduction policies. For example, innovations that reduce the cost of converting fossil fuels into hydrogen, while sequestering safely underground the CO_2 by-product, raise the prospect of accelerated commercialization of GHG-free, hydrogen-based technologies such as fuel cells. The financial costs of such technologies are not independent of public policy, although the relationship is highly uncertain.
- Uncertainty about consumer and business preferences increases as the time horizon is extended. How might these preferences change as environmental awareness, income, available technologies, and policy alter over that same 20- to 30-year time frame? Thus, policies that mandate the development and commercialization of hydrogen-based fuel cell vehicles are likely to influence the marketing efforts of vehicle manufacturers, with a possible effect on the preferences of consumers. Again, the magnitude of this effect is highly uncertain.

The importance of external uncertainty is usually assessed by constructing and simulating scenarios that show contrasted but plausible future values for uncertain factors, such as the price of internationally traded GHG emission permits or crude oil. If, however, a country has a significant influence on the market of internationally traded energy commodities or GHG permits, as is the case for the United States, it should also model the relationship between its domestic policies and these international prices. The NEMS model of the U.S. government has modules that link developments in the U.S. market to international energy prices and could link to a module of internationally traded GHG permits. Top-down models, being more aggregated, have dominated the exploration of this uncertainty so far, although there have been significant advances in recent years in linking together the MARKAL bottom-up models from several countries for this type of analysis.

Given the long-run nature of an environmental challenge such as climate change risk, and the profound technological change that is required to achieve substantial GHG emission reductions, the uncertainties associated with technological innovation and consumer and business preferences are of great significance for the estimation of costs. For addressing these uncertainties, bottom-up, top-down, and hybrid approaches offer different capabilities.

3.1 Top-Down Approaches to Long-Run Uncertainty, and Critique

Although top-down models have not traditionally been applied to extremely long-run periods in which the technological context is likely to change, some economists have used them for 50- and even 100-year simulations of the policy response to climate change risk. But as the time frame extends further into the future, the character of technical change plays an increasingly important role in determining the costs of GHG reduction, yet technical change in top-down models is simply portrayed via the ESUB and AEEI parameters. These parameters are typically estimated from past technological and preference changes associated with price changes for energy and other inputs.

How confident can the modeler, and the policy maker, be in the stability of these parameter values in future time periods? For example, how does the top-down modeler translate new information about the cost of renewable electricity generation or high-efficiency light bulbs into their parameter estimates for AEEI and ESUB? How does the modeler show the policy maker the relationship between technology-focused research and development policies and these same parameters? To be useful to policy makers, models need to indicate the direction, probability, and impact of technological changes that influence both the financial costs and consumers' surplus costs of GHG emission reduction—and the extent to which these changes might be influenced by policy. Parameters estimated from historical conditions may have little ability to anticipate the future technology and preference dynamics in response to policies seeking to induce long-run, profound technological change.

3.2 Bottom-Up Approaches to Long-Run Uncertainty, and Critique

In recognition of this challenge, bottom-up modelers have focused on applying lessons from past technological innovations to new technologies, linking cumulative manufacture and expanded application

of a technology to reductions in its financial costs—the "economies of learning" or "experience curves" discussed earlier. For example, as global production of windmills has grown over the past 20 years, the cost of windmill-generated electricity fell quickly at first and today continues gradually down toward an asymptote, possibly in the range of (U.S.) $0.04/kWh or even lower. Researchers suggest that zero-emission, hydrogen-fueled vehicles could exhibit a similar declining cost pattern over the next 20–30 years.

In this bottom-up framework, the analyst can make projections about the declining financial costs of emerging technologies as they penetrate the market, assume a rate of diffusion, and then calculate the resulting GHG reductions and financial costs. Bottom-up modelers can also link the future financial costs of these new technologies to policy initiatives, such as research and development subsidies, in what Grubler, Nakicenovic, and Victor, among others, refer to as "endogenous modeling of induced technical change." But the very criticisms of bottom-up modeling that motivated the development of hybrid modeling approaches apply equally to the exploration of long-run technical change. Uncertainty about emerging technologies concerns not just the diversity of views on their future *ex ante* financial costs (the slope of their experience curves), but also uncertainty about the willingness of consumers and businesses to embrace any particular technology as its financial costs fall and awareness of it increases.

An example is the choice of technology for providing personal mobility in the urban areas of industrialized countries. Emerging substitutes for the conventional vehicle driven by a gasoline-fueled internal combustion engine include replacing gasoline with ethanol, adopting battery–gasoline vehicles, adopting hydrogen-fuel-cell vehicles, and switching to public transit. Bottom-up modeling of induced technical change could show a substantial decrease in the *ex ante* financial cost of public transit relative to the alternatives, but this would mislead policy makers if it had dramatically underestimated the reluctance of urbanites in wealthy countries to switch away from personal vehicles. A similar concern applies to many of the new technologies that are now being applied to estimate the long-run costs of GHG reduction.

3.3 Hybrid Efforts to Address Long-Run Uncertainty

Uncertainty internal to the system being modeled, therefore, is related both to the long-run dynamic between technology commercialization and financial costs, on the one hand, and between technology commercialization and preferences, on the other. To be useful to policy makers, simulations of policies to induce long-run technical change should provide an indication of how reductions in the *ex ante* financial costs of technology options might translate into rates of market penetration, the critical issue overlooked by most bottom-up modeling.

To address this uncertainty, researchers of technological change policy, such as Train, increasingly apply "discrete-choice models" to relate past preferences to certain technology attributes in what is called "revealed preference research." Estimated confidence intervals provide an indication of uncertainty. But for new technologies there is no past market behavior, so discrete-choice researchers must instead ask consumers and businesses to make choices between new or even hypothetical technologies with different bundles of attributes, what is called "stated preference research." An example is to ask consumers to make a choice from among a set of vehicles that vary in terms of acceleration, refueling time, fuel type, emissions, and range. Confidence intervals from this research indicate uncertainty, but may be biased or imprecise given that the choices are hypothetical.

Hybrid models are well suited for incorporating the estimates from discrete-choice modeling into their behavioral parameters. First, discrete-choice research is focused at the technology level, the same level of resolution at which bottom-up and hybrid models operate. Second, hybrid models have behavioral parameters at the level of detail required for simulating technology choices—decision-specific discount rates, and technology-specific costs that can be adjusted to reflect consumers' surplus differences. Research in this direction is still at an early stage, but recent advances are showing a good deal of promise. The transportation module of the NEMS hybrid model, for example, already contains in its behavioral parameters the results of revealed and stated preference research on alternative vehicle choices. Other approaches to addressing technical uncertainty involve subjective adjustment of top-down model parameters (ESUB and AEEI) to incorporate lessons from technology-focused research, and the introduction of behavioral elements into the technology-choice algorithms of bottom-up models.

4. ASSESSMENT

Bottom-up models can be useful to policy makers by showing the technological possibilities for advancing

toward an environmental target, such as reducing GHG emissions. However, these models can also be misleading in that their definition of cost is restricted to a single *ex ante* estimate of financial costs that ignores possible losses of option value and consumers' surplus by businesses and consumers. Most bottom-up models are also not able to estimate the overall effect of energy-impacting policies on the macroeconomy. Top-down models, in contrast, can be useful to policy makers because they incorporate the preferences of businesses and consumers—in that their parameters are usually estimated from real-world behavior—and connect the energy system to the macroeconomy. Bottom-up models tend to estimate low costs of GHG reduction relative to top-down models. Different types of comparative analysis, some of it using hybrid models that combine technological explicitness with behavioral realism, can help policy makers understand the role of these different definitions of cost on the divergence in cost estimates between the two approaches.

A second key challenge for GHG reduction cost estimation is due to uncertainties, and hence different assumptions, about technology and preference changes over very long periods. Recent research with bottom-up models attempts to explore how financial costs may decrease as new technologies experience the benefits of learning and perhaps mass production. However, these models lack the behavioral component that is critical in assessing the extent to which the decreasing cost of a new technology will drive its adoption by consumers or firms. Models with hybrid characteristics of technological explicitness and behavioral realism offer some hope to improve the usefulness of models for decision makers and thus address Grubb's challenge of more than a decade ago to develop "the missing models."

SEE ALSO THE FOLLOWING ARTICLES

Biomass: Impact on Carbon Cycle and Greenhouse Gas Emissions • Bottom-Up Energy Modeling • Diet, Energy, and Greenhouse Gas Emissions • External Costs of Energy • Goods and Services: Energy Costs • Greenhouse Gas Emissions, Alternative Scenarios of • Greenhouse Gas Emissions from Energy Systems, Comparision and Overview • Modeling Energy Markets and Climate Change Policy • Multicriteria Analysis of Energy

Further Reading

Azar, C., and Dowlatabadi, H. (1999). A review of technical change in assessment of climate policy. *Annu. Rev. Energy Environ.* **24,** 513–544.

Bohringer, C. (1998). The synthesis of bottom-up and top-down in energy policy modeling. *Energy Econ.* **20,** 233–248.

Brown, M., Levine, M., Romm, J., Rosenfeld, A., and Koomey, J. (1998). Engineering-economic studies of energy technologies to reduce greenhouse gas emissions: Opportunities and challenges. *Annu. Rev. Energy Environ.* **23,** 287–385.

Grubb, M. (1991). Policy modeling for climate change: the missing models. *Energy Policy* **21,** 203–208.

Grubler, A., Nakicenovic, N., and Victor, D. (1999). Modeling technological change: Implications for the global environment. *Annu. Rev. Energy Environ.* **24,** 545–569.

Jaccard, M., Nyboer, J., Bataille, C., and Sadownik, B. (2003). Modeling the cost of climate policy: Distinguishing between alternative cost definitions and long-run cost dynamics. *Energy J.* **24,** 49–73.

Jaffe, A., and Stavins, R. (1994). Energy-efficiency investments and public policy. *Energy J.* **15,** 43–65.

Train, K., and Atherton, T. (1995). Rebates, loans, and customers' choice of appliance efficiency level: Combining Stated- and revealed-preference data. *Energy J.* **16,** 55–69.

U.S. Energy Information Administration. (1998). "Impacts of the Kyoto Protocol on U.S. Energy Markets and Economic Activity." U.S. Department of Energy, Washington, D.C.

Weyant, J., and Hill, J. (1999). Introduction and overview (special issue on the costs of the Kyoto Protocol: A multi-model evaluation). *Energy J.* vii–xliv.

Greenhouse Gas Emissions, Alternative Scenarios of

LEO SCHRATTENHOLZER and KEYWAN RIAHI
International Institute for Applied Systems Analysis
Laxenburg, Austria

1. **Greenhouse Gas Emissions Scenarios**
2. **The Difference between Scenarios, Projections, and Forecasts**
3. **Scenarios, Uncertainty, and Probability**
4. **Energy–Economy–Environment Scenarios**
5. **Why Build Scenarios of Greenhouse Gas Emissions?**
6. **How are Scenarios of Greenhouse Gas Emissions Built?**
7. **Driving Variables of Scenarios**
8. **Scenarios and E3 Models**
9. **Model Inputs and Outputs**
10. **Published Scenarios**
11. **Overall Ranges of Greenhouse Gas Emissions and Their Driving Forces**
12. **Three Groups of Scenarios**
13. **Integrated Assessment of Climate Change**
14. **Scenarios and Policymaking**

Glossary

E3 models Energy–economy–environment models.
final energy The form of energy that is bought and sold in the market to consumers.
greenhouse gas (GHG) A gas that can trap heat in the atmosphere by absorbing the longwave radiation emanating from Earth's surface.
Kyoto Protocol An agreement between industrialized countries (the so-called Annex B countries), obliging them, individually or jointly, to ensure that their average annual carbon dioxide equivalent emissions of six greenhouse gases (CO_2, CH_4, N_2O, SF_6, CF_4, and halocarbons) will be 5% below 1990 levels in the commitment period 2008 to 2012.
primary energy Naturally occurring form of energy.
reference energy system (RES) A schematic and aggregated representation of all energy conversion technologies and the flows between them. A RES is not uniquely defined, but rather depends on the level of aggregation considered in a particular analysis.
scenario A qualitative and/or quantitative description of (the development) of a system, based on a set of internally consistent assumptions concerning the main driving forces.
stochastic models Representations that include uncertainty by replacing certainty (of parameters and functional relations) with probability distributions.
storyline In connection with a scenario, a storyline is a qualitative, textual description of how the future might evolve.

This article reports on global carbon emissions projected with long-term energy–economy–environment (E3) scenarios. Many of the prominent emission scenarios have been published under the auspices of the Intergovernmental Panel on Climate Change (IPCC). After briefly discussing methodological aspects of scenario building—in particular its driving forces and other inputs—the article proceeds to systematically discuss scenario results for three groups of scenarios, that is, high-impact, CO_2 mitigation, and sustainable development scenarios. For these groups, ranges are given for the most important inputs and outputs. For mitigation scenarios, typical mitigation costs are reported.

1. GREENHOUSE GAS EMISSIONS SCENARIOS

In the context of this article, the term "scenario" means the qualitative and/or quantitative descriptions of (the development of) a system, based on a set of internally consistent assumptions concerning the

 © 2004 Elsevier Inc. All rights reserved.

main driving forces. A scenario of greenhouse gas (GHG) emissions then depicts the development of GHG emissions over time for a given geographical region. As the greenhouse gas problem is a global problem, the focus here is describing global scenarios.

In addition to including a geographical specification, the system used to describe GHG emissions usually also includes so-called driving forces. These driving forces comprise socioeconomic, demographic, and technology variables that determine the evolution of the energy, agricultural, and other systems, the major sources of anthropogenic GHG emissions. GHG emission scenarios include one or more greenhouse gases, but in most cases, one of the greenhouse gases included is carbon dioxide (CO_2). For most GHG emission scenarios, the six gases included in the Kyoto Protocol are the maximum number of gases considered. Analogous to the glass roofs of actual greenhouses, greenhouse gases are restricted in the narrow sense to mean gases that trap the heat that emanates from Earth's surface, but other gases and aerosols (such as SO_x, water vapor, black carbon, and others) influence the heat balance of Earth (for instance, but not always, by shading, which has a cooling effect). The so-called precursor gases are yet another important factor for the dynamic heat balance of the atmosphere. They are not greenhouse gases, but they are involved in atmospheric chemical reactions that produce or absorb greenhouse gases.

The most important use of greenhouse gas emission scenarios is to serve as inputs for models that calculate the climate impact of given global emission pathways. Conceptually, the climate impact of GHG emissions is described by the chain: emissions, atmospheric concentrations, radiative forcing (a measure of the amount of heat trapped per unit area, e.g., per square meter), temperature change, and specific impacts (e.g. land losses due to sea level rise). Consecutive links of this chain are connected through a complex system that is not yet completely understood. Any evaluation of the causal chain between GHG emissions and impact is therefore surrounded by considerable uncertainty. To give an example, according to recent estimates from the Intergovernmental Panel on Climate Change (IPCC), the temperature increase that is expected to result from a doubling of atmospheric GHG concentrations (relative to preindustrial levels) may range between 1.7 and 4.2 K.

2. THE DIFFERENCE BETWEEN SCENARIOS, PROJECTIONS, AND FORECASTS

It is important to distinguish a scenario from forecasts and, even more so, from predictions, the most important difference being that forecasts and (to an even higher degree) predictions are meant to portray particularly likely future developments. In contrast, scenarios may very well include elements that may not be considered the most likely development. For example, the task of perhaps the most prominent efforts to generate GHG emission scenarios, the IPCC Special Report on Emission Scenarios (SRES), edited by Nakićenović and Swart in 2000, was to portray the entire range of plausible greenhouse gas emissions during the 21st century. Another example of scenarios that are not constructed for the purpose of portraying particularly likely developments are those scenarios that describe a consistent set of circumstances that lead to a desired outcome. If models are used to generate such a scenario, the model is often run "backward." Scenarios designed to generate a particular outcome are called "normative." In contrast, scenarios that describe the consequence of assuming alternative future states of the world, which usually are meant to be particularly likely, are called "descriptive." This distinction is conceptual, and borderline cases exist that could be classified either way.

3. SCENARIOS, UNCERTAINTY, AND PROBABILITY

A common way of treating uncertainty is to define probability distributions, according to which an uncertain event is distributed. Many probability distributions are well understood, and the interpretation of medians, percentiles, variances, and other indicators is a common practice in research. Perhaps less common is the insight that probabilities such as those describing uncertainties about the evolvement of the climate system are subjective and that they therefore cannot be uniquely determined. As a practical definition of probability (in contrast to a strictly formal mathematical definition), de Finetti (1974) proposed "degree of belief." Of course, belief is always in the eyes of the beholder, and it is therefore consistent with this definition to postulate that all probabilities are subjective and that probability theory is the science of analyzing and

processing given (subjective) probability distributions.

This way of looking at probabilities is far from commonplace, and the perception and interpretation of probabilistic statements by the interested public are therefore unpredictable. Moreover, it can easily happen that two scientists attach different degrees of belief to one and the same scenario, even when they have produced that scenario jointly. Publishing different probabilities for the same scenario would then almost certainly cause confusion. This was the reason that Nakićenović and Swart, and the other authors of the IPCC SRES report, refrained from estimating probabilities or the relative likelihood of each of the 40 GHG emission scenarios that they generated and published. Of course, there are scientific disciplines in which subjective probabilities of most scientists match or in which it is acceptable to work with probability distributions such as Gaussian (normal) distributions, the parameters of which have been estimated from empirical data. In such cases, the use of probability distributions and, possibly, of stochastic models is well established, and the interpretation of the results is straightforward, at least conceptually. This is not the case for GHG emission scenarios, for which subjective probabilities of different scientists and policymakers alike may vary quite substantially. Instead of hiding these differences and then wondering how it is possible that different conclusions are drawn from identical scenarios, it would appear worth trying to be more specific about differences of subjective probabilities, particularly of the catastrophic consequences of climate change. Doing so could be a way of transferring disagreement on the mitigation measures to disagreement on probabilities, which could lead to agreement on how to manage uncertainty surrounding many aspects of climate change.

4. ENERGY–ECONOMY–ENVIRONMENT SCENARIOS

Many scenarios of greenhouse gas emissions belong to the family of energy–economy–environment (E3) scenarios. There is a close relationship between the three systems. E3 scenarios usually focus on the energy system, including the economy as a driving force and the environment to map the environmental impact of energy production, conversion, and end use. Although global E3 scenarios now usually include the climate issue as one of the most important global environmental problems, the global climate need not be the main theme of such scenarios. More and more scenario builders address the issue of sustainable development, which is a more general concept that includes the issue of climate change.

5. WHY BUILD SCENARIOS OF GREENHOUSE GAS EMISSIONS?

Generally speaking, GHG emission scenarios are built to analyze the consequences of policy action (or non-action) on the development of greenhouse gas emissions, the costs of their abatement, and the impact on the economy. National scenarios can serve the purpose of designing national policies to fulfill a GHG emission reduction commitment such as the targets set by the Kyoto Protocol. Global scenarios can serve as inputs to climate models. In this way, many GHG emission scenarios can collectively describe the range of plausible emissions, with climate models describing the range of possible consequences on climate. It is still too early for climate models to calculate numerical estimates of climate impacts within reasonably narrow bounds. Should this be possible at some future date, GHG emission scenarios can be the first step of an analysis that compares abatement costs with avoided damage. Until that date, damage scenarios will have to serve as placeholders for accurate quantitative estimates.

6. HOW ARE SCENARIOS OF GREENHOUSE GAS EMISSIONS BUILT?

The process of scenario building depends on the subject of the scenarios and the issues to be analyzed with them. An illustration of a typical and widespread method to build long-term GHG emission scenarios is the method that was used to build the scenarios described in the IPCC Special Report on Emission Scenarios (SRES).

As already mentioned, the purpose of the SRES was to gauge the range of plausible global GHG emissions in the 21st century. This required thinking of limiting cases on the high and low ends of the plausible range. The first step in defining the limiting cases—and, in fact, defining any case—was to think of general circumstances that would be consistent with and lead to the targeted outcome in terms of

GHG emissions. High levels of emissions, for example, could be the consequence of high rates of population growth and economic growth, high carbon-intensity of gross domestic product (GDP), poor technological progress, and other factors. Such qualitative assumptions are then formulated in the "storyline." A storyline is a qualitative, textual description of how the future might evolve. It gives the rationale and narrative behind a specific combination of driving forces aiming at directing the development path of aggregate variables of a scenario (for instance, relative to their past trends). For example, "high economic growth rate" means that assumptions of economic growth in a scenario should be on the high side of past trends. Such qualitative characterizations included in storylines are later quantified and transformed into model input numbers.

7. DRIVING VARIABLES OF SCENARIOS

The most important of the "driving variables" of energy demand, which in turn is a major determinant of carbon emissions, are GDP and energy intensity of GDP. Disaggregating one step further, GDP can be thought of as the composite of population and GDP per capita. Energy intensity, together with the carbon intensity of energy supply, can be aggregated to the carbon intensity of GDP. Looking only at the logic of the variables and the problem, the carbon intensity of GDP might seem to be the most obvious driving variable to determine carbon emissions of an economy. However, the point of disaggregating, creating even more variables, is eventually to arrive at those variables that can be most plausibly projected into the future. This criterion depends on the expertise of the scenario builders, and different groups therefore use different driving variables for the projection of carbon emissions.

8. SCENARIOS AND E3 MODELS

The route from storylines, via driving forces, to scenarios is through formal (mathematical) models. Although this discussion is not on modeling, it appears appropriate to give a description of some relevant modeling terms here.

There are several criteria to classify models. One way is to classify them by the overall purpose of model usage, into normative and descriptive. This distinction is particularly important for models that project carbon emissions. For instance, scenarios projecting low-level carbon emissions are often built for the purpose of specifying a set of conditions under which certain environmental goals can be reached. In contrast, descriptive models typically describe the development of a system as the consequence of few assumptions, which are often meant to be particularly likely. An important class of descriptive scenarios is the group of "business-as-usual" scenarios, which aim at projecting past trends in a straightforward way into the future. Sometimes, such scenarios are also referred to as "dynamics-as-usual" scenarios if the point is to be made that systematic changes of trends have been incorporated in them and that incorporating them leads to distinctly different results, compared to mere trend extrapolations.

The core of energy models aiming at projections of greenhouse gases usually incorporates an abstract image of energy flows from primary-energy to final consumption. This image is usually called a reference energy system (RES). Conceptually, the RES has primary energy carriers on its input end and energy demands (and by-products such as pollutants) on the output end. The two ends are connected by a set of technologies that extract, convert, transport, and distribute various forms of energy. One important characterization of an energy model is the degree of detail with which its RES is described in the model. Another way of characterizing models is by its so-called system boundaries. For energy models, one important issue with respect to the system boundary of the model is whether energy demand is substitutable by other factors of production, such as capital and labor, and whether it is price responsive. This issue concerns the degree of detail in which the overall economy is described in the energy model.

9. MODEL INPUTS AND OUTPUTS

As previously stated, the driving forces determine model output such as greenhouse gas emissions. A most important qualifier to this general statement is the amendment "within the given constraints." Using the concept of the reference energy system, such constraints concern primary-energy recoverability, the availability and performance of technologies, and energy demand. Of these groups, energy demand is the one that is shaped the most by the driving forces discussed thus far, because it reflects GDP,

population, GDP per capita, and energy intensity. Energy intensity is also related to the conversion efficiencies of the technologies included in the reference energy system. So is the carbon intensity, which, in addition to being related to the performance of non-carbon-emitting technologies, is also related to the model assumptions on the recoverability of hydrocarbon energy resources.

Generally speaking, technological progress can therefore be considered as another driving force of greenhouse gas emissions. It overlaps with other driving forces, but in many studies, the role of technological development is central. In comparison to the other driving forces, this may well be regarded by policymakers as one of the driving forces that can be steered particularly effectively (for instance, by supporting research and development of energy technologies). As it will be substantiated later, during the discussion of the policy relevance of long-term GHG emission scenarios, it can be argued that such a policy would be aiming at influencing the driving forces in a much more sustainable way, compared to energy taxes or "command-and-control" measures. This, of course, is not meant to say anything against the latter, just to emphasize the importance of technology policy as an important additional way of driving greenhouse gas emissions in the right direction. In fact, the mitigation of greenhouse gases to volumes that would lead to a "stabilization of greenhouse gas concentrations in the atmosphere at a level that would prevent dangerous anthropogenic interference with the climate system," as postulated by the United Nations Framework Convention on Climate Change, would suggest combining all available policies and measures for the given purpose.

10. PUBLISHED SCENARIOS

One recent comprehensive effort to explore alternative GHG emission paths projected for the 21st century led to the publication of the IPCC Special Report on Emission Scenarios (SRES). The SRES scenarios were developed to represent the range of driving forces and emissions in the scenario literature so as to reflect current understanding and knowledge about underlying uncertainties. Hence, plausibility, size, and location of the emissions range were defined by emission scenarios that had been published by the time the SRES work began. Note that the SRES scenarios do not include additional climate initiatives, which means that no scenarios are included that explicitly assume the implementation of climate policies, such as the United Nations Framework for Climate Change (UNFCC) or the emissions targets of the Kyoto Protocol.

The first step of SRES work thus consisted of compiling what has become also known as the SRES database, produced by Morita and Lee in 1998. This database includes the results of some 400 E3 scenarios, which are described in terms of the most important variables characterizing the long-term development of the E3 system either globally or for major world regions. The variables include population, economic growth, energy demand, carbon emissions, and others. Although not all scenarios in the database report on all variables, the scenarios included can be regarded as representative of the range of possibilities still regarded as plausible by the global modeling community 5 years after they were collected in the SRES database.

One way of interpreting the ranges of the projected carbon emissions and their driving forces would be to see them as quantifications of the uncertainty surrounding the evolvement of the variables during the course of the present century. To what extent the ranges and frequency distributions quantify the uncertainty in any reliable way (i.e., to what extent they can be interpreted as probabilities) is open for speculation and individual judgment. In the overview of scenario results given in the following section, such an interpretation of frequencies as probabilities is withheld, and the focus is instead on the description of statistical indicators, such as medians, variances, the minimum, and the maximum.

11. OVERALL RANGES OF GREENHOUSE GAS EMISSIONS AND THEIR DRIVING FORCES

For the summary presentation of ranges of greenhouse gas emissions and their driving forces, a common way of graphically presenting values and ranges of a number of variables is to use a regular polygon and (zero-based) axes between the center and each vertex. A heptagon representation with seven variables is chosen to summarize the global scenarios of the SRES database; five variables describe the values of scenario variables in the year 2100, one describes a cumulative figure up to that year, and one describes a growth rate. The five variables with values for 2100 are (1) CO_2 emissions, expressed in billions of tons (Gt, gigatons; 10^9 tons)

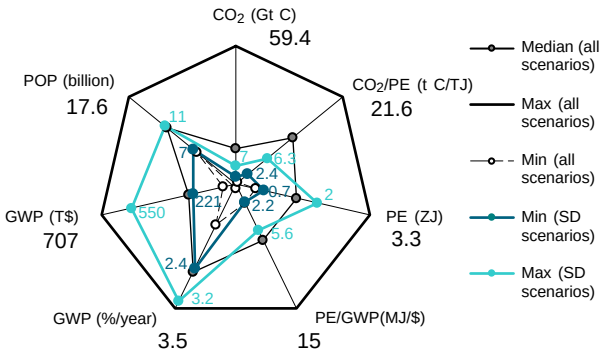


FIGURE 1 Global carbon emissions and their main driving forces. The minimum and maximum of the values for the Special Report on Emissions Scenarios database, as well as for the sustainable-development (SD) scenarios only, are shown on seven axes of the heptagon, and they are connected. The seven axes show ranges for indicators across the scenarios in 2100 (see text for discussion). PE, Primary energy; GWP, gross world product; T$ trillions of dollars; POP, population. IIASA, International Institute for Applied Systems Analysis.

of carbon, (2) specific carbon emissions per unit of primary energy (PE), expressed in tons of carbon per terajoule, (3) total primary energy demand up to the year 2100, expressed in zetajoules (10^{21} joules), (4) specific primary energy consumption of economic output, expressed as gross world product (GWP) in megajoules per U.S. dollar (1990 purchasing power), and (5) gross world product, expressed in trillions (10^{12}) of U.S. dollars (1990 purchasing power); (6) population, expressed in billions (10^{9}) of people; and (7) growth of gross world product, expressed as the average annual growth rate of the GWP between 1990 and 2100. The resulting heptagon is presented in Fig. 1.

The outer, symmetric heptagon in Fig. 1 represents the maximum values for the group of all SRES database scenarios for each of the seven variables. The smallest irregular heptagon inside this envelope represents the respective minimum values for all SRES database scenarios. Another irregular heptagon shows the respective seven median values for all SRES database scenarios. Preempting the later discussion of sustainable development (SD) scenarios, the figure also shows, for comparison, two heptagons that connect the seven minimum and maximum values for the group of SD scenarios for each of the seven variables.

In interpreting the heptagons, it should be remembered that any one of them is unlikely to connect points that belong to the same scenario. For instance, it would not appear logical if either a minimum or a maximum heptagon connected values for the same scenario. Take economic growth and primary-energy intensity, for example. One common assumption in E3 scenarios is that with higher economic growth, technological progress can occur faster. As a consequence, the maximum primary energy intensity should not coincide with maximum economic growth. Figure 1 also illustrates that CO_2 emissions in the year 2100 cover a range between zero, i.e., no net carbon emissions, to almost 60 billion tons (gigatons) of carbon (Gt C), with the median at approximately 18 Gt C, significantly below halfway. In contrast, world GDP growth and population both have a "bias" toward the maximum. For economic output, this means that the SRES database does not include scenarios with zero or low economic growth during the 21st century.

12. THREE GROUPS OF SCENARIOS

The overall ranges of global GHG emissions and their driving forces provide a first overview of possibilities that have been considered by modelers. To better understand the sometimes considerable differences between GHG emissions (and between other variables) projected in different scenarios, three separate groups of scenarios are now described: high-impact, CO_2 mitigation, and sustainable-development scenarios. The description of these three groups also includes scenarios that have contributed to neither SRES scenarios nor the SRES database. The reason is that the mandate for the authors of the SRES was to survey carbon emissions in scenarios that did not include explicit climate policies. Although scenarios that included more general environmentally benign strategies, such as sustainable development, were included, explicit climate mitigation scenarios were not included in the SRES process. Rather, they were included in another IPCC publication, the Third Assessment Report, compiled by Metz *et al.* in 2001.

Another preface is required before we can meaningfully describe SD scenarios. We follow a definition of SD E3 scenarios as proposed by the Environmentally Compatible Energy Strategies (ECS) Project at IIASA. According to this definition, all scenarios that satisfy the following four criteria will be referred to as sustainable-development scenarios:

1. Economic growth (in terms of GDP per capita) is sustained throughout the whole time horizon of the scenario.
2. Socioeconomic inequity among regions, expressed as the world-regional differences of gross domestic product per capita, is reduced significantly

over the 21st century, in the sense that by 2100, the per capita income ratios between all world regions are reduced to ratios close to those prevailing between Organization for Economic Cooperation and Development (OECD) countries today.

3. Long-term environmental stress is mitigated significantly. In particular, carbon emissions at the end of the century are approximately at or below today's emissions. Other GHG emissions may increase, but total radiative forcing, which determines global warming, is on a path to long-term stabilization. Other long-term environmental stress to be mitigated includes impacts to land use (e.g., desertification). Short- to medium-term environmental stress (e.g., acidification) may not exceed critical loads that threaten long-term habitat well being.

4. The reserves-to-production (R/P) ratios of exhaustible primary-energy carriers do not decrease substantially from today's levels.

The proponents of this definition argue that it has been inspired by and is consistent with the "Brundtland spirit," which is quantitative enough to serve the purpose of classifying long-term energy–economic scenarios, and that it comes close to the definition of weak economic sustainability, complemented with some environmental and social constraints.

The quantitative characterizations of the three groups of E3 scenarios in this section are based on 34 scenarios that were developed and published by the IIASA's ECS Project since 1998. These include scenarios developed with the World Energy Council (WEC) in 1998, scenarios developed for the IPCC Special Report on Emissions Scenarios, as well as scenarios developed for the impact assessment for the Working Group III of the IPCC Third Assessment Report. As to the ranges of carbon emissions and driving forces, the 34 IIASA ECS scenarios are representative of the scenarios in the SRES database.

12.1 High-Impact Scenarios

Most scenarios of the SRES database include a high-level environmental impact in at least one respect. For the quantitative summary in this subsection, the eight IIASA ECS scenarios that are representative of these high-impact scenarios have been selected. One class of scenarios in this group consists of the business-as-usual scenarios. In particular, the SRES-B2 scenario, which was designed in such a way that its driving variables stay in close proximity to the medians of the SRES database, is classified as high impact. Another important function of high-impact scenarios is to serve as a reference, relative to which mitigation and other policies can be quantified. One example is the SRES-A1C scenario ("C" stands for coal), in which atmospheric CO_2 concentrations in the year 2100 reach 950 parts per million by volume (ppmv), more than three times the preindustrial level. SRES-A1C is also characterized by a high rate of economic growth. The projected global GDP in 2100 is 550 trillion (10^{12}) U.S. dollars (in 1990 prices). Relative to global GDP in 1990, 20.9 trillion, this corresponds to an average annual growth rate of just below 3%. The lowest economic growth rate in this group of eight scenarios is projected in the IIASA WEC B scenario, one of the business-as-usual scenarios. There the projected average annual growth rate is 2%.

Global population does not show characteristic features in the high-impact scenarios. Projections range from 7.1 billion (after an earlier peak) to 15.1 billion in the year 2100. Global primary-energy consumption in the high-impact scenarios ranges from 1500 to 2700 exajoules (EJ) in 2100. Relative to the global consumption in 1990, 350 EJ, this range implies annual growth rates between 1.3 and 1.9%.

12.2 CO_2 Mitigation Scenarios

Mitigation scenarios explore cases in which the global atmospheric CO_2 concentration is stabilized at different levels. The following description is based on 19 IIASA ECS scenarios. These scenarios were constrained to stabilize atmospheric CO_2 concentrations at levels of 450, 550, 650, and 750 ppmv in 2100, and they used high-impact scenarios as references. The criterion that classifies scenarios in this group raises the question as to whether constraining atmospheric CO_2 concentrations in this way leads to sustainable-development scenarios. The answer is "not necessarily." Of the 19 scenarios included in this group, only 8 would also qualify as SD scenarios. Although this observation is based on a sample that might be considered small, it suggests that sustainable development is a more general strategic goal compared to climate mitigation.

Making a reference-mitigation pair of scenarios comparable suggests leaving crucial assumptions on general technological development and resource availabilities the same in both scenarios. At the same time, it could be argued that it would appear implausible to assume exactly the same kind of technological progress as in the reference and stringent emission mitigation, and at the same time. There is no clear-cut solution to this issue, but to

keep rules simple, the same set of technology data and the same set of resource availabilities as for the corresponding baseline scenarios were assumed for the mitigation cases. To respond to the "technological response" argument, sustainable-development scenarios were built (these are described in the following subsection).

The main mechanisms to reduce emissions relative to the reference cases can therefore be summarized as supply shifts toward primary-energy carriers, with less carbon content and price responsiveness of final energy demand. As a general characterization of the carbon emission trajectories in the mitigation cases, it can be said that they are similar to those derived by other modelers for the same concentration limits, in particular the popular "WRE" trajectories described by Wigley, Richels, and Edmonds in 1996. Very often, the first question about mitigation concerns costs. Choosing typical examples of mitigation aiming at atmospheric CO_2 concentrations of 550 ppmv (corresponding to roughly twice the preindustrial level), and quantifying the difference in global GDP in 2100 between the mitigation scenarios and the corresponding reference cases, obtains reductions of 7 trillion U.S. dollars for the SRES-A2 scenario (236 instead of 243) and 8 trillion for the SRES-A1C scenario (542 instead of 550). These correspond to relative reductions by 3 and 1.5%, respectively. At the average annual growth rates of economic output assumed in the two scenarios, this would mean delaying the achievement of the given output by 1.5 and 0.5 years for the A2 and A1C scenarios, respectively. In comparison with the reductions of global economic output in the mitigation cases, the difference in primary-energy consumption is relatively higher. In A2-550, primary-energy consumption in 2100 is 18% lower than that of the reference A2 scenario. In A1C, this difference is lower (6%).

The comparison of these two pairs of scenarios clearly shows that the relative economic impact of mitigation is higher where lower growth had been projected in the corresponding reference case. Nonetheless, reaching the mitigation goal of 550 ppmv in the higher growth scenario A1C reduces atmospheric CO_2 concentrations in the year 2100 by a larger amount (from 950 ppmv), compared to the lower growth scenario A2 (from 780 ppmv).

12.3 Sustainable-Development Scenarios

By the definition used here and in comparison to most other scenarios, sustainable-development scenarios project low levels of GHG emissions. The following description is based on seven IIASA ECS scenarios. Three distinct general qualitative strategic ideas are reflected in this group of scenarios. One idea can be dubbed "small is beautiful," and it leads to the scenarios of the SRES-B1 family. These scenarios are characterized by low energy intensity of GDP, which is dominated by the service sector. Energy supply is highly decentralized and relies to a high degree on renewable energy. In contrast, the SRES-A1T scenario describes a high-technology interconnected world in which technological progress (in particular, in the field of carbon-free sources, which include nuclear energy) is key to creating an environmentally compatible global energy system.

Built some years earlier than the SRES scenarios, the two sustainable-development scenarios of the IIASA WEC set (IIASA-WEC C1 and IIASA-WEC C2) describe a "global family" in which the industrialized countries transfer substantial amounts of resources to the developing countries and regions, to aid development in a way that preserves the global environment. Even without involving the assumption of massive international transfers of wealth, all scenarios in this group assume a fully cooperative world with efficient markets and without constraints such as trade barriers, monopolies, and the like. This assumption can be regarded mostly normative in the sense of the distinction between normative and descriptive, and substantial amounts of goodwill would be required to implement any of the sustainable-development scenarios.

Comparing projected GDP of sustainable-development scenarios with the projections of all scenarios in the SRES database (Fig. 2) shows that projected GDP in SD scenarios tends to cover the higher part of all GDP projections. The crosshatched area in Fig. 2 includes average annual GDP growth rates between 2 and 3%. This suggests that sustainable development tends to require wealth rather than to prevent it. According to the definition of sustainable-development scenarios, projected global carbon emissions in this group in 2100 are below today's emissions. Accordingly, atmospheric CO_2 concentrations in the year 2100 are between 445 and 560 ppmv. Figure 3 illustrates the carbon emission reductions projected in sustainable-development scenarios, comparing the carbon intensity of primary energy in SD scenarios with that of all scenarios in the SRES database. Figure 3 shows that the highest projected average decarbonization rates in the database are near 3.3% per year between 1990 and 2100.

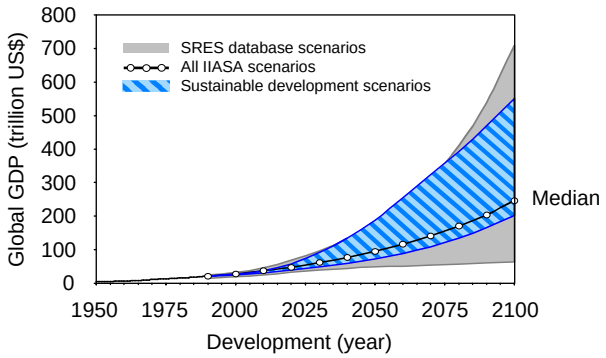


FIGURE 2 Global economic product (GDP) in trillions of 1990 U.S. dollars; development from 1950 to 1990 and in the scenarios [193 of which are from the Special Report on Emission Scenarios (SRES) database] from 1990 to 2100. Historical data are from two United Nation studies published in 1993; the database is from Morita and Lee (1998). IIASA, International Institute for Applied Systems Analysis.

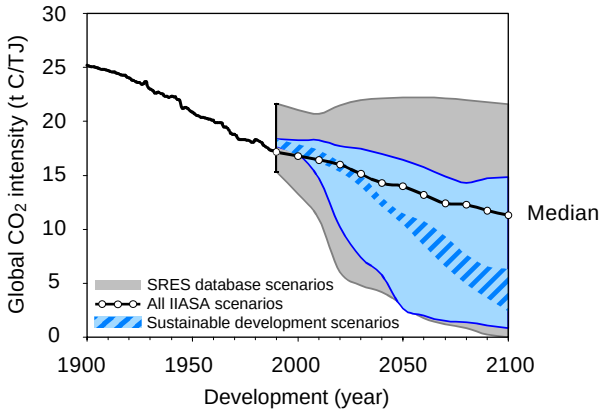


FIGURE 3 Global carbon intensity of primary energy; actual development from 1900 to 1990 and in the scenarios [140 of which are from the Special Report on Emission Scenarios (SRES) database] from 1990 to 2100. Historical data are from Nakićenović (1996); the database is from Morita and Lee (1998). Note that net carbon emissions are used to calculate the carbon intensities (i.e., bioenergy is excluded). The differences in the carbon intensities for the base year are due to different accounting conventions and data problems in general. IIASA, International Institute for Applied Systems Analysis.

This means a reduction of energy-related carbon emissions per unit of primary energy by a factor of 40 during this time horizon, which, in turn, leads to an energy system with almost no carbon emissions. Decarbonization in the sustainable-development scenarios leads to projected carbon intensities for the year 2100 between 2.4 and 6.3 C/TJ, both down from 18 C/TJ in 1990.

13. INTEGRATED ASSESSMENT OF CLIMATE CHANGE

Analyses of emission reduction, explicitly or implicitly, mainly address the cost component of climate abatement. For a more complete view of the problem, however, it is necessary also to get an estimate of the benefits of emission mitigation in terms of avoided damage. This is the task addressed by models for the integrated assessment of climate change. In view of the substantial uncertainties surrounding the impacts of climate change, estimates of avoided damage vary within wide ranges, even for given emissions. Still, a range of estimates can be better than no estimate, and any damage estimate is good enough to analyze the cost-effectiveness of policies and measures (that is, the cost of a unit of emission reduction). This aspect of cost-effectiveness was one motivation for the ongoing efforts to study more greenhouse gases than CO_2 only, although CO_2 is the most important anthropogenic greenhouse gas and therefore has been studied most intensively. The most obvious non-CO_2 greenhouse gases to analyze in addition to CO_2 are those included in the Kyoto Protocol of the UNFCC (CH_4, N_2O, SF_6, CF_4, and halocarbons). The six "Kyoto gases" and their abatement are the subject of the Stanford-based Energy Modeling Forum study (EMF-21), which involves modelers from North America, Europe, and Asia. The study is still ongoing in 2003, but the first preliminary results show that considering GHGs other than CO_2 leads to significantly reduced mitigation costs.

14. SCENARIOS AND POLICYMAKING

Mitigation scenarios incur mitigation costs in terms of GDP lost, but such costs for sustainable-development scenarios have not been reported here. The reason is that the mitigation scenarios discussed herein were defined relative to reference cases, whereas SD scenarios were defined in absolute terms. This may seem just like a curious coincidence, but it is not, and it appears worth emphasizing the policy implications of this difference, which is not just a methodological peculiarity, but by deliberate design. This difference means to express the idea that a strategy aiming directly at sustainable development could be more effective than conceptually constraining the evolvement of a system that is implicitly

assumed to be preferable, if not for the risk of catastrophic consequences of climate change. This has been recognized in the past for strategies to mitigate local pollution in cases in which the issue is to limit pollution either by so-called end-of-pipe technologies or by avoiding it from the beginning. It has been argued, for instance, that given enough notice (time), progress in automobile technology could have achieved the same emission reduction as catalytic converters, by improving the internal combustion engine instead of treating the exhaust gases. It is the idea of sufficient notice that may provide for an analogy to GHG emissions. If it holds, this analogy suggests that aiming for sustainable development might be more promising than imposing carbon constraints, given that sufficient time is available for action. In turn, this means that action should be initiated sooner rather than later.

The projections of GHG emissions and their driving forces reported here vary over wide ranges. Carrying the analysis further, to atmospheric GHG concentrations, radiative forcing, temperature change, and impacts, in a sense increases the size of these ranges, as a consequence of the uncertainties surrounding the evolvement of the global climate, which is not understood well enough to predict the consequences of given GHG emissions within narrow limits. Whatever the overall strategy to avoid catastrophic consequences of climate change, this uncertainty will play a major role in policymaking. What will be called for is uncertainty management at unprecedented scales, and flexibility of strategies will be an obvious plus.

SEE ALSO THE FOLLOWING ARTICLES

Biomass: Impact on Carbon Cycle and Greenhouse Gas Emissions • Carbon Capture and Storage from Fossil Fuel Use • Carbon Taxes and Climate Change • Climate Change and Energy, Overview • Climate Protection and Energy Policy • Diet, Energy, and Greenhouse Gas Emissions • Greenhouse Gas Abatement: Controversies in Cost Assessment • Greenhouse Gas Emissions from Energy Systems, Comparison and Overview

Further Reading

De Finetti, B. (1974). "Theory of Probability," Vol. 1. John Wiley & Sons, London and New York.

Houghton, J. T., Ding, Y., Griggs, D. J., Noguer, M., van der Linden, P. J., and Xiaosu, D. (eds.). (2001). "Climate Change 2001: The Scientific Basis." Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change (IPCC). Cambridge Univ. Press, Cambridge, UK.

Klaassen, G., Miketa, A., Riahi, K., and Schrattenholzer, L. (2002). Targeting technological progress towards sustainable development. *Energy Environ.* **13**(4/5), 553–578.

Metz, B., Davidson, O., Swart, R., and Pan, J. (eds.). (2001). "Climate Change 2001: Mitigation." Contribution of Working Group III to the Third Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge Univ. Press, Cambridge UK.

Morita, T., and Lee, H.-C. (1998). "IPCC SRES Database, Version 0.1." Emission Scenario Database, prepared for IPCC Special Report on Emissions Scenarios. Available on the Internet at http://www.cger.nies.go.jp.

Nakićenović, N. (1996). Freeing energy from carbon. *Daedalus* **125**(3), 95–112.

Nakićenović, N., Grübler, A., and McDonald, A. (eds.). (1998). *Global Energy Perspectives*. Cambridge University Press, Cambridge UK, pp. 299, ISBN 0521645697.

Nakićenović, N., and Swart, R. (eds.). (2000). "Emissions Scenarios." Special Report of the Intergovernmental Panel on Climate Change (IPCC). Cambridge Univ. Press, Cambridge UK.

Rao, S., and Riahi, K. (2003). "Long-term Multigas Mitigation Strategies Using MESSAGE." Presentation at the International Energy Workshop, 25 June. International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria.

Riahi, K., and Roehrl, R. A. (2000). Greenhouse gas emissions in a dynamics-as-usual scenario of economic and energy development. *Technol. Forecast. Social Change* **63**(3), 195–205.

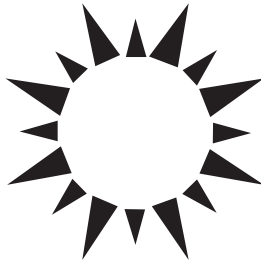
Riahi, K., and Roehrl, R. A. (2000). Energy technology strategies for carbon dioxide mitigation and sustainable development. *Environ. Econ. Policy Stud.* **3**(2), 89–123.

United Nations (UN). (1993). "Macroeconomic Data Systems, MSPA Data Bank of World Development Statistics, MSPA Handbook of World Development Statistics." MEDS/DTA/1 and 2 June. UN, New York.

United Nations (UN). (1993). "UN MEDS Macroeconomic Data Systems, MSPA Data Bank of World Development Statistics." MEDS/DTA/1 MSPA-BK.93, Long-Term Socioeconomic Perspectives Branch, Department of Economic and Social Information and Policy Analysis. UN, New York.

Wigley, T. M. L., Richels, R., and Edmonds, J. A. (1996). Economic and environmental choices in the stabilization of atmospheric CO_2 concentrations. *Nature* **379**, 240–243.

Greenhouse Gas Emissions from Energy Systems, Comparison and Overview



ROBERTO DONES, THOMAS HECK, and STEFAN HIRSCHBERG
Paul Scherrer Institut
Villigen, Switzerland

1. Introduction
2. Fuel-Specific Greenhouse Gas Emission Factors
3. Energy Chain-Specific Greenhouse Gas Emissions
4. Comparison of Greenhouse Gas Emissions

Glossary

Central European Electricity Interconnection (CENTREL) Regional group of four transmission system operator companies of the Czech Republic, Hungary, Poland, and Slovak Republic.

energy chain All stages associated with energy supply by means of specific energy carriers including mining, processing, and conversion of an energy carrier, as well as the treatment and disposal of the produced waste.

mixed oxides fuel (MOX) Oxides of plutonium and uranium.

Organization for the Nordic Electric Cooperation (NORDEL) Includes Denmark, Finland, Norway, Sweden, and Island.

Union for the Coordination of Transmission of Electricity (UCTE) Members in year 2002 are Austria, Belgium, Bosnia-Herzegovina, Czech Republic, Croatia, Denmark (associated member), France, Germany, Greece, Hungary, Italy, Luxembourg, Macedonia, the Netherlands, Poland, Portugal, Slovak Republic, Slovenia, Spain, Switzerland, and Serbia and Montenegro. The CENTREL countries Czech Republic, Hungary, Poland, and Slovak Republic have officially joined UCTE in 2001.

According to the Intergovernmental Panel on Climate Change, the earth's climate system has changed, globally and regionally, with some of these changes being attributable to human activities resulting in emissions of greenhouse gases. Energy supply and the fossil systems in particular are the dominant contributors to the emissions of these gases. This article provides an overview and comparison of greenhouse gas emissions associated with fossil, nuclear, and renewable energy systems. In this context, both the direct technology-specific emissions as well as the contributions from full energy chains within the life cycle assessment framework are considered. This discussion also provides examples illustrating the differences between industrialized and developing countries and explores the impact of technological advancements.

1. INTRODUCTION

1.1 Global Greenhouse Gas Emissions from Energy Systems and Other Sources

According to the Intergovernmental Panel on Climate Change (IPCC), the anthropogenic greenhouse gas (GHG) emissions contributing most clearly to global warming in terms of relative radiative forcing are CO_2, CH_4, halocarbons, and N_2O. Radiative forcing is the change in the net vertical irradiance (expressed in watts per square meter: Wm^{-2}) at the boundary between the troposphere and the stratosphere. Compared to the preindustrial era 250 years ago, additional radiative forcing due to increases of the well-mixed greenhouse gases is estimated to be $2.43\,Wm^{-2}$. CO_2 contributes most ($1.46\,Wm^{-2}$), followed by methane ($0.48\,Wm^{-2}$), halocarbons ($0.34\,Wm^{-2}$), and N_2O ($0.15\,Wm^{-2}$).

Other possible factors influencing the global climate are less well understood and thus quantitatively

 © 2004 Elsevier Inc. All rights reserved.

more uncertain. Among them are stratospheric ozone (cooling), tropospheric ozone (warming), sulphate (cooling), black carbon and organic carbon (warming or cooling), biomass burning (cooling), mineral dust (warming or cooling), aerosol indirect effects (cooling), land use (change of albedo, i.e., share of reflected sun light), and solar variation.

Table I shows the global emissions of the major greenhouse gases and the contribution of anthropogenic sources for the late 1990s. CO_2 emissions originate mainly from combustion of fossil fuels (i.e., from energy systems). These CO_2 emissions are quite well known. For CH_4 and N_2O, the total emission rates are much more uncertain and it is difficult to assess global emissions from single sources. Halocarbons are molecules containing carbon and either chlorine, fluorine, bromine, or iodine. Among the halocarbons are several refrigerants that are used as working fluids for cooling or heating (i.e., for transport of thermal energy). Many refrigerants interesting for energy uses (in refrigerators, heat pumps, air conditioners, etc.) have very high global warming potentials (GWP) on a per kg emitted basis. Energy scenarios for reduction of CO_2 emissions often include increased use of heat pumps substituting for fossil fuel heating systems. It has to be kept in mind that refrigerant emissions due to leakages counteract to a certain extent in the total GHG balance. Halocarbon emissions are almost completely made by humans. The table does not include refrigerants like CFC-11 or CFC-12, which have been banned because of their high ozone-depleting potential; they show low emission rates but are still abundant in the atmosphere due to emissions in the past.

TABLE I

Annual Emissions of Important Greenhouse Gases in the Late 1990s

	Annual emissions (Mt/year)	**Lifetime (years)**	**GWP 100-year**	**CO_2-equiv. 100-year (Mt/year)**
II. CO_2	29,000		1	29,000
Fossil fuels	*22,400*			*22,400*
Cement production	*700*			*700*
Land use, etc.	*6000(3000–9000)*			*6000*
III. CH_4	600	8.4–12	23	13,800
Energy	*100(89–110)*			*2300*
Biomass burning	*40(23–55)*			*900*
Other anthropogenic sources	*230*			*5300*
Natural sources	*230*			*5300*
IV. N_2O	26	120	296	7700
Automobiles	*0.3(0.2–0.4)*			*90*
Industry, including energy	*2.0(1.1–2.8)*			*600*
Other anthropogenic sources	*8.5(7.6–9.6)*			*2500*
Natural sources	*15*			*4500*
V. HFC refrigerants				
HFC-23	0.007	260	12,000	84
HFC-134a	0.025	13.8	1300	33
HFC-152a	0.004	1.4	120	0.5
VI. Other halocarbons				
Perfluoromethane (CF_4)	0.015	$>50{,}000$	5700	86
Perfluoroethane (C_2F_6)	0.002	10,000	11,900	24
Sulphur hexafluoride (SF_6)	0.006	3200	22,200	133

Emissions are given in real mass per year and in CO_2-equiv. mass per year. GWP, global warming potential; HFC, hydrofluorocarbon.
Source. IEA (1999); rounded values.

1.2 Methodological Basis and Scope of Comparisons

The most straight-forward accounting of GHG emissions can be based on emission factors associated with combustion of the various fossil fuels. This approach can be used for the establishment of the overall national emission inventories but is not practical when the goal is to account fully for emissions associated with the use of specific technologies. While uses of nuclear and renewable energy sources exhibit practically negligible emission levels for GHGs at the stage of power generation, the same is not necessarily true for other stages of the corresponding energy chains. In addition, emissions may arise when manufacturing the components for the plants, transporting fuels and other materials, or in the decommissioning stage.

Life cycle assessment (LCA), an approach utilizing process chain analysis specific to the types of fuels used in each process, allows for the full accounting of all such emissions, also when they take place outside of the national boundaries. Thus, LCA considers not only emissions from power plant construction, operation, and decommissioning but also the environmental burdens associated with the entire lifetime of all relevant upstream and downstream processes within the energy chain. This includes exploration, extraction, processing, and transport of the energy carrier, as well as waste treatment and disposal. The direct emissions include releases from the operation of power plants, mines, processing factories, and transport systems. In addition, indirect emissions originating from manufacturing and transport of materials, from energy inputs to all steps of the chain, and from infrastructure are covered.

An alternative, not process-oriented approach, is the input/output (I/O) method; it divides an entire economy into distinct sectors and based on the inputs and outputs between the sectors generates the energy flows and the associated emissions. A hybrid approach is also frequently employed, combining LCA and I/O methods; the I/O method is then used exclusively for assessing processes of secondary importance.

This article provides comparisons of GHG emissions, based on several approaches mentioned earlier. The largest part of the data cited here has been derived using a full scope LCA implementation. Due to consistency and other practical reasons it is not feasible to account for results from all major published studies addressing the GHG emissions. As the main reference work, the most recent results of the comprehensive Swiss study addressing LCA-based environmental inventories in Europe has been used. This is supplemented by some results from studies carried out by researchers in and for other countries. More details on the methodological aspects of LCA as applied in the Swiss study will be provided in Section 3.

2. FUEL-SPECIFIC GREENHOUSE GAS EMISSION FACTORS

For national accounting of direct CO_2 emissions from fuel combustion, the International Energy Agency (IEA) within OECD uses the IPCC methodology. Neither GHGs other than CO_2 nor non-energy emission sources are included in the IEA statistics. Table II summarizes the fuel-specific emission factors.

The direct emissions based on the factors provided in Table II can be combined with appropriate contributions from the production chain of fuels. Table III shows the emission factors per unit of mass from the upstream chain of selected fuels in the Swiss LCA study. The large variations for coal products depend on the differences between the producing regions (mainly the energy needed for mining and the coalbed methane emissions) and between supply mixes (transport energy). The variations for coke and coke by-products depend on the lower average efficiency assumed for the world average production versus coke making in Central Europe.

A Japanese LCA study addressed burdens from fossil fuel extraction, processing, and transportation, associated with uses of the fuels in Japan. The supplying countries were addressed separately. In Table IV, only the average emission factors are shown. In the Japanese study, CO_2 and CH_4 emissions are given separately; here they are summed using for methane the GWP according to IPCC. The average values for coal are in the lower end of those estimated by the Swiss study, partly because of the different distribution of the origin of the coal and the transport distances, partly due to the different assumptions on the energy uses at mine and methane emissions. The Japanese study did not address the refinery step, which explains the differences for oil. The results for CO_2 emissions for a single supplying country may change by a factor of four for coke, three for steam coal, two for crude oil and LPG, but only by about 20% for LNG.

TABLE II

Carbon Conversion Factors for Fuels[a]

	Fuel	Carbon emission factor (tC/TJ)	Carbon emission factor (tCO_2/TJ)
Liquid fossil	Primary fuels		
	Crude oil	20	73
	Orimulsion	22	81
	Natural gas liquids	17.2	63
	Secondary fuels/products		
	Gasoline	18.9	69
	Jet kerosene	19.5	72
	Other kerosene	19.6	72
	Shale oil	20	73
	Gas/diesel oil	20.2	74
	Residual fuel oil	21.1	77
	LPG (liquid petroleum gas)	17.2	63
	Ethane	16.8	62
	Naphta	20	73
	Bitumen	22	81
	Lubricants	20	73
	Petroleum coke	27.5	101
	Refinery feedstocks	20	73
	Refinery gas	18.2	67
	Other oil	20	73
Solid fossil	Primary fuels		
	Anthracite	26.8	98
	Coking coal	25.8	95
	Other bituminous coal	25.8	95
	Subbituminous coal	26.2	96
	Lignite	27.6	101
	Oil shale	29.1	107
	Peat	28.9	106
	Secondary fuels/products		
	Patent fuel and brown coal/peat briquettes (BKB)	25.8	95
	Coke oven/gas coke	29.5	108
	Coke oven gas	13	48
	Blast furnace gas	66	242
Gaseous fossil	Natural gas (dry)	15.3	56
Biomass	Solid biomass	29.9	110
	Liquid biomass	20	73
	Gas biomass	30.6	112

[a] The factors are given in terms of tonne C and tonne CO_2 to facilitate application. Values are rounded.
Source. Revised 1996 IPCC Guidelines for National Greenhouse Gas Inventories (1997), IPCC/OECD/IEA, Paris.

TABLE III

Greenhouse Gas Emissions from the Upstream Chains of Fossil Fuels Used in Europe Calculated in the Swiss LCA Study

		kg CO_2-equiv./kg fuel	
		Minimum	Maximum
Hard coal	At producing region[a]	0.040	0.343
	Country-specific supply mix[b]	0.188	0.322
	Supply mix to UCTE	0.270	
	Coke[c]	0.492	0.655
	Coke oven gas[c]	0.612	0.814
	Benzene[c]	0.685	0.911
	Tar[c]	0.637	0.847
	Briquettes	0.336	
Lignite	At mine	0.017	
	Briquettes	2.223	
	Pulverized	0.812	
Oil[d]	Heavy fuel oil	0.423	
	Light fuel oil	0.480	
	Bitumen	0.402	
	Diesel	0.482	
	Diesel low sulphur	0.495	
	Kerosene	0.477	
	Naphta	0.417	
	Petroleum coke	0.488	
	Petrol low sulphur	0.769	
	Propane/butane	0.573	
	Refinery gas	0.558	
Natural gas	Western European high-pressure grid	0.491	

[a] Minimum: South America; maximum: Western Europe.
[b] Minimum: Poland; maximum: Germany.
[c] Coke oven gas, benzene, and tar are by-products of coke making. Minimum: Central Europe; maximum: estimation for world average.
[d] Values for Western Europe. The analysis performed for the Swiss-related chain resulted in greater values by a factor 1.2 to 1.4.
Source. Dones *et al.* (2003).

TABLE IV

LCA Greenhouse Gas Emissions from the Upstream Chain of Fossil Fuel Mixes Used in Japan

Fuel	kg CO_2-equiv./kg fuel
Steam coal	0.168
Coke	0.118
LNG	0.640
Crude oil (for refinery)	0.122
Crude oil (for power plant)	0.153
LPG total[a]	0.319
NGL total[b]	0.258

[a] Weighted over the two processes producing liquid petroleum gas (LPG) from liquefied natural gas (LNG) (0.560 kg CO_2-equiv./kg) and producing LPG at oil fields (0.258 kg CO_2-equiv./kg).
[b] Weighted over the three processes producing natural gas liquids (NGLs) during LNG liquefaction (0.214 kg CO_2-equiv./kg to 0.481 kg CO_2-equiv./kg) and directly at the oil field (0.254 kg CO_2-equiv./kg).
Source. Kato *et al.* (2000).

3. ENERGY CHAIN-SPECIFIC GREENHOUSE GAS EMISSIONS

This section provides chain-specific results obtained in the Swiss LCA study for Switzerland and Europe, supplemented by representative result examples from a number of published analyses by researchers in other countries. Some additional basic features of the LCA methodology as applied in the Swiss applications are summarized here; most of these principles also apply to other state-of-the-art studies, which, however, may differ in terms of scope, level of detail, specific assumptions, and methodology applied (some results are based on hybrid approaches). The most important features are as follows:

- Energy systems, transport systems, material manufacturing, production of chemicals, waste treatment and disposal, as well as agricultural products, have been assessed using detailed process analysis developed under consistently defined common rules.
- Electricity inputs were modeled using production technology or supply mix as close as feasible to the actual situation. In case of lack of specification, the UCTE mix was used as an approximation.
- Allocation criteria were developed for multipurpose processes.

The results provided here focus on electricity supply, but selected results are also given for heat generation and for cogeneration systems.

3.1 Fossil Energy Chains

3.1.1 Coal

3.1.1.1 Hard Coal In the Swiss study European country-specific average power plants operating around year 2000 were analyzed. For the estimation

of the infrastructure of the plants, two power levels of 100 and 500 MW were considered; a mix with a share of 10 and 90%, respectively, was defined for the reference plant. The reference unit was assumed to be used for middle load with 4000 hours of operation per year at full capacity, and operate 150,000 hours during its lifetime. Emissions of CO_2 were estimated on the basis of the coal characteristics (e.g., lower heating values in the range of 22.1 to 26.6 MJ/kg), and the average net efficiencies of single units operating in Europe (29 to 40%), while emissions of CH_4, N_2O, and CO are averages taken from the literature.

Average coal from eight supplying regions was considered: West and East Europe, North and South America, Australia, Russia, South Africa, and Far East Asia. Specific average data were used for coal characteristics (e.g., upper heating values in the range of 20 MJ/kg to 28 MJ/kg), share of open pit and underground mines, methane emissions, land use, and energy uses for each of these regions. Import mixes in year 2000 for all European countries with coal plants have been defined.

The average methane emissions from coal mining in the eight modelled regions range between 0.16 g CH_4/kg (typical emission from open pit coal mines in the United States) and 13.6 g CH_4/kg product coal (West Europe).

The results for GHG GWP 100a from the chains associated with the average hard coal in European countries is between 949 g CO_2-equiv./kWh for the NORDEL countries and 1280 g CO_2-equiv./kWh for the Czech Republic (including several cogenerating plants, where the emission is entirely allocated to the electricity generation). The average for UCTE countries (excluding CENTREL) in year 2000 is 1070 g CO_2-equiv./kWh. Methane contributes nearly 7%, and N_2O 0.8% to the total GHG emissions for the UCTE average hard coal chain; CO_2 emissions practically cover the rest.

The upstream chain contributes between 8% (Portugal) to 12.5% (Germany) to total GHG. The total GHG associated with production regions varies between 0.04 kg CO_2-equiv./kg coal at production in South America up to 0.288 kg CO_2-equiv./kg coal at production in Russia.

Figure 1 provides a comparison between the range of average estimates of normalized GHG emissions for UCTE-countries, the average for Japan and United States, the range obtained for the coal chain in the Province Shandong in China, and the range according to a worldwide survey carried out in 1997. The Japanese and U.S. results are on the lower side of the ranges for UCTE countries and in the 1997 survey; the same applies to the lower range of the estimates from the study for China. The higher range from the China study reflects the low efficiencies characteristic in particular for small units in China as well as large contributions from mining; the potentially substantial but difficult to estimate additional GHG emissions from uncontrolled coal fires have not been included.

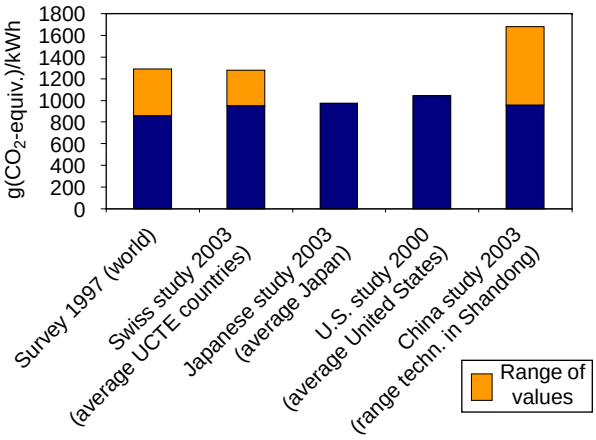


FIGURE 1 Greenhouse gas emissions from coal power plants. The figure shows ranges of greenhouse gas emissions in CO_2-equivalents per kWh electricity from coal power plants and associated fuel cycles according to different studies. From Survey (1997), Van de Vate, J. F., *Energy Policy* 25, pp. 1–6; Swiss study (2003), Dones *et al.*; Japanese study (2003), Hondo (2003); U.S. study (1999), Spath *et al.* (1999); China study (2003), Dones *et al.* (2003).

3.1.1.2 Lignite The reference plant used for lignite in the Swiss LCA study has similar characteristics as for hard coal but a larger share of plants with low rated power was used. The reference unit is assumed to be used for base load with 6000 hours of operation per year at full capacity, and work 200,000 hours during its lifetime. Emissions of CO_2 are estimated on the basis of the characteristics of the average lignite (e.g., lower heating values in the range of 5.2 to 16.6 MJ/kg) and the average efficiencies of single units operating in Europe (range for country-average is between 23 and 40%), while emissions of CH_4, N_2O, and CO are UCTE averages taken from the literature.

Considering that lignite plants are mine-mouth, only an average European open-pit mine has been modeled on the basis of information limited to only few mines. Only 0.23 g CH_4/kg lignite are assumed to be emitted during mining operations.

The results for GHG GWP 100a from the chains associated with the average lignite power plants in

European countries is between 1060 g CO_2-equiv./kWh for Austria and 1690 g CO_2-equiv./kWh for the Slovak Republic. The average for UCTE countries (excluding CENTREL) in year 2000 is calculated as 1230 g CO_2-equiv./kWh. Methane contributes about 0.6%, and N_2O 0.5% to total GHG emission for the UCTE average lignite chain; CO_2 emissions practically cover the rest. Mining contributes marginally between 0.9% (France) to 2.6% (Greece) to total GHG.

3.1.2 Oil

Since the role of oil in electricity generation is decreasing, only few key facts are provided from the Swiss study. The European country averages for GHG emissions of oil chains range from 519 to 1200 g CO_2-equiv./kWh 100-year. The UCTE average around the year 2000 was about 880 g CO_2-equiv./kWh 100-year, of which about 88% or 775 g/kWh are the direct emissions during power plant operation.

For the fuel oil supply, emissions occur during crude oil exploration, long-distance transport (e.g., in transoceanic tankers), oil processing in refineries, and local distribution. For an average oil power plant in Europe, the highest contributions to the upstream GHG emissions occur in the oil exploration phase and in the heavy oil processing in refineries.

3.1.3 Gas

3.1.3.1 Natural Gas Chain For natural gas, like for the other fossil fuel fired electricity or heating systems, the dominant contributor to GHG emissions are the CO_2 emissions from the power plant or boiler. Natural gas is transported in pipelines over long distances. Since natural gas consists mainly of methane (i.e., natural gas itself is a greenhouse gas), leakages in the pipelines can contribute significantly to the total GHG emissions. For European countries, the methane emissions can make up to about 10% of the total GHG emissions in the full chain depending on the location of the natural gas power plant or the boiler. Together with CO_2 and other GHG emissions in the upstream chain, the emissions other than directly from the power plant can constitute more than 10% of the total GHG emissions for European natural gas power plants (about 17% for the UCTE average in year 2000).

The country averages of the full chain GHG emissions of natural gas power plants in Europe range from 485 to 991 g CO_2-equiv./kWh 100-year. The UCTE average around the year 2000 was about 640 g CO_2-equiv./kWh 100-year. Therein are about 530 g direct CO_2-equiv./kWh emissions during operation of the power plants.

For the modeling of the best technology combined-cycle gas power plant, data from the new 400 MW power plant Mainz-Wiesbaden (Germany) were used. According to the operators, this is the natural gas power plant with the highest net electric efficiency (58.4%) worldwide (year 2001). Because the efficiency depends also on the local conditions (the Mainz-Wiesbaden plant is located directly at the river Rhine, which yields good cooling conditions), it was assumed that a comparable plant at an average location in Europe would have a net efficiency of about 57.5%. The full chain GHG emissions of the best combined-cycle power plant (about 420 g/kWh) are much lower than those of an average gas power plant.

Figure 2 compares the range of average estimates of normalized GHG emissions from the gas chain for UCTE-countries, with the average for LNG in Japan, results for combined cycle plants in Europe, Japan and the United States, and the range according to the worldwide survey carried out in 1997. The upper range in the survey probably represents a plant using a gas mix rather than pure natural gas.

3.1.3.2 Industrial Gas Industrial gas covers blast furnace gas from steel production and coke oven gas.

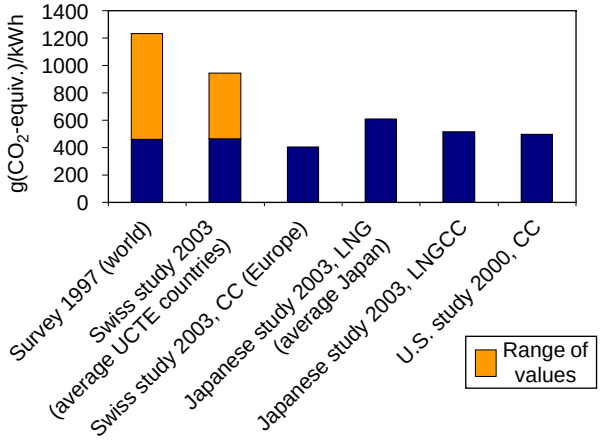


FIGURE 2 Greenhouse gas emissions from gas power plants. The figure shows ranges of greenhouse gas emissions in CO_2 equivalents per kWh electricity from gas power plants according to different studies. From Survey (1997), Van de Vate, J. F., *Energy Policy* **25**, pp. 1–6; Swiss study (2003) Dones *et al.*; Japanese study (2003), Hondo (2003); U.S. study (2000), Spath, P. L., and Mann., M. K., "Life Cycle Assessment of a Natural Gas Combined-Cycle Power Generation System," National Renewable Energy Laboratory, Golden, CO.

The results cited here are based on the mix of blast furnace gas and coke oven gas used for electricity production in UCTE around year 2000. Due to its high CO_2 content, blast furnace gas has high CO_2 emission factors of 90 to 260 g/MJ burned gas. Emission factors of coke oven gas range between 40 and 90 g/MJ gas burned. This leads to exceptionally high total GHG emissions, exceeding 1700 g CO_2-equiv./kWh for the European average industrial gas mix. The methane contribution comes mainly from the coke oven gas production chain.

3.1.4 Heating and Cogeneration

3.1.4.1 Heating Two hard coal heating systems were modeled in the Swiss LCA study: an industrial furnace with capacity in the range 1 to 10 MW, and a stove of about 5 to 15 kW. The thermal efficiency of the furnace is 80%, the one of the stove 70%. The industrial furnace is assumed to be fuelled with the average Western European hard coal supply mix, the stove either with coke or with briquettes.

Assuming that all CO is oxidized into CO_2, CO contributes about 10% to the total GHG emissions associated with the stove. Direct CH_4 emissions from burning briquettes are 20 times higher than from burning coke in the stove. In case of the briquette, direct methane emissions from the stove are about 50% of the total methane emissions calculated for the chain. Due to a lower carbon content per unit of energy, burning briquettes has lower direct CO_2 emissions than burning coke.

Condensing gas and oil boilers use the heat from combustion and the heat from condensation of the water in the flue gas. Modern condensing natural gas boilers can reach annual net efficiencies of about 102%, modern oil boilers about 98%. (The ratio refers to the lower heating value of the fuel therefore efficiencies of more than 100% are possible for condensing boilers.) High efficiency reduces the CO_2 emissions. Direct CO_2 emissions of a modern condensing natural gas boiler <100 kW are about 56 g/MJ, the GHG emissions of the full chain for a condensing natural gas boiler in Central Europe in year 2000 add up to about 71 g/MJ CO_2-equiv. 100-year. For a 10 kW condensing nonmodulating oil boiler with direct CO_2 emissions of about 74 g/MJ, the full chain GHG emissions for location Central Europe are about 89 g/MJ CO_2-equiv. 100-year.

Figure 3 shows a comparison of GHG emissions for the energy chains based on various heating technologies of different capacity, utilizing hard coal, natural gas, and oil. The lowest emissions are for the natural gas systems, followed by oil systems.

3.1.4.2 Cogeneration Figure 4 shows a comparison of CO_2 emissions per kWh_e from modern small cogeneration plants of different capacity and technology, located in Switzerland. Allocation of emissions to the products is in this case based on exergy. The higher the capacity, the higher the electric efficiency and the lower the CO_2 emissions for electricity. The total efficiency is approximately constant for the different plants shown. The CO_2 emissions per MJ fuel burned are the same for all natural gas plants but higher for the diesel plant because of the higher emission factor of diesel oil.

3.2 Nuclear Energy Chain

The amount of GHG emissions from the nuclear chain associated with light water reactors (LWRs) is controlled by few parameters: the nuclear cycle considered, the average enrichment and burnup at discharge of the fuel, the lifetime of the plants, especially of the power plant, and, most important, the enrichment process used and the electricity supply to the enrichment diffusion plant (if its services are used).

The Swiss LCA study on energy systems addresses the Swiss, French, German, and average UCTE nuclear chains, separately modeling boiling water reactor (BWR) and pressurized water reactor (PWR) power plant technologies with reference to the two 1000 MW class units installed in Switzerland.

The chain was decomposed into several steps: uranium mining (open pit and underground), milling, conversion, enrichment (diffusion and centrifuge), fuel fabrication, power plant (light water reactors), reprocessing, conditioning (encapsulating spent fuel), interim storage of radioactive waste, and final repositories (geological for high and intermediate level waste).

The study assumes partial or total reprocessing of the fuel, according to the conditions in West European countries. In particular, 40% of the total spent fuel produced at Swiss and German plants during their lifetime was assumed to be reprocessed, whereas for France this is most likely the fate of 100% of the fuel; for UCTE a weighted average of 80% of total spent nuclear fuel was assumed to undergo reprocessing. Mixed oxides fuel (MOX) was modeled in the study as MOX is

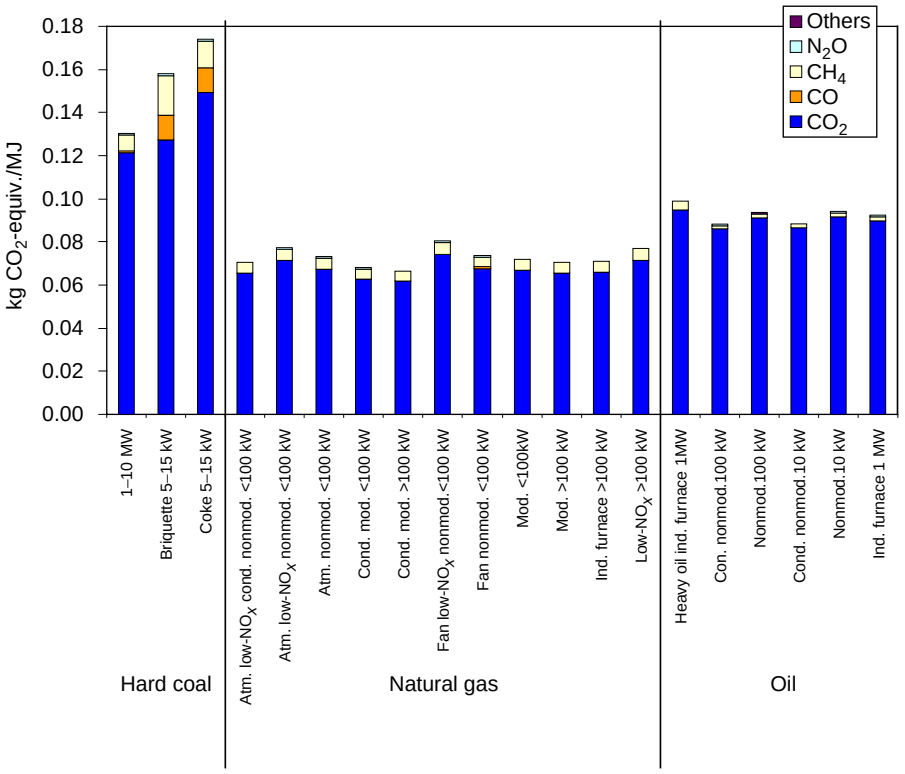


FIGURE 3 Greenhouse gas emissions from fossil fuel heating systems. CO_2-equivalent emissions per MJ heat are shown for different fuels and different technologies. Atm., atmospheric boiler; Fan, fan burner; Mod., modulating boiler; Cond., condensing boiler; Ind. furnace, industrial furnace. Low-NO_x refers to boilers built in early 1990s. All emission factors estimated for boilers operating around year 2000 in Central Europe; modulating boilers refer to the most modern technology on the market. From Dones *et al.* (2003).

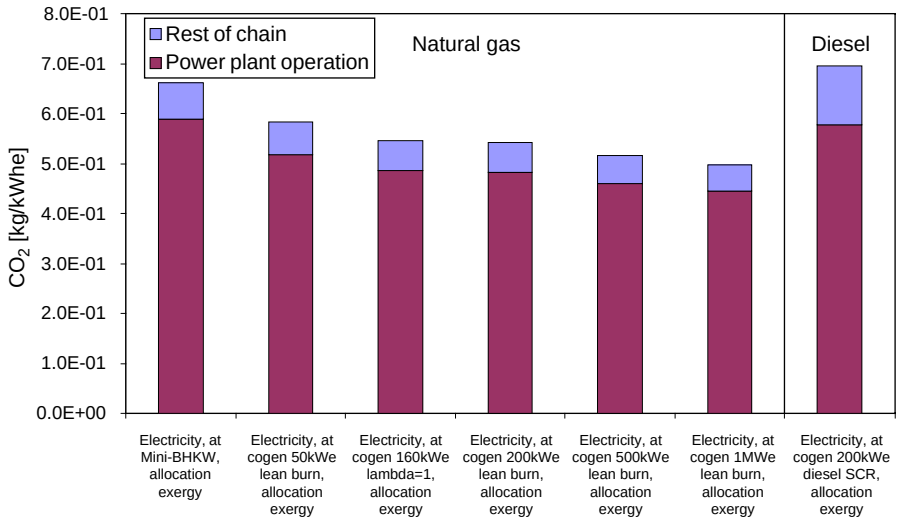


FIGURE 4 CO_2 emissions per kWhe from modern small cogeneration plants of different capacity and technology. Lambda = 1: cogeneration unit with three-way catalyst. Mini-BHKW is a small device comparable to a boiler for a small house with about 2 kWe. SCR, selective catalytic reduction. The allocation to electricity is based on the exergy of electricity and heat assuming an upper temperature of about 80°C for heat. From Dones *et al.* (2003).

widely used in Western Europe, for which recycled plutonium is mixed with depleted uranium from enrichment.

Lifetime of 40 years was assumed for the power plants. Average U-235 enrichment ranges between 3.8% for French PWR and Swiss BWR and 4.2% for Swiss PWR. The corresponding average burnup of spent fuel ranges between nearly 43 MW_{th} day/kgU for the French plants and 52 MW_{th} day/kgU for the Swiss PWR. The study includes a description of enrichment supplies and relevant electricity input, based on a literature search and assumptions on the worldwide enrichment market and country-specific enrichment services.

The total emissions of GHG for the modeled nuclear chains are calculated between approximately 6 and 11 g CO_2-equiv./kWh, where the minimum is estimated for the French nuclear chain, assuming 10% MOX, and 100% use of enrichment at the diffusion Eurodif plant in Tricastin (supplied by nuclear plants and using no CFCs as refrigerants but water). The maximum was calculated for Germany, under the assumption of about 13% MOX and a mix of enrichment services including 10% of the USEC diffusion plant which is assumed to be supplied by coal power plants; nearly 70% of the enrichment services are assumed to be supplied by URENCO facilities (one is based in the north of Germany), based on centrifugal technology, which is about 65 times less energy intensive than the USEC plant.

Calculated total GHG is prevalently from CO_2 emissions, between 90 and 93%. Methane contributes between 3 and 6%, whereas N_2O is between approximately 1 and 3%. The hydrofluoro- and hydrochlorocarbon emissions from the enrichment stage are below 5%.

The GHG associated with the power plant ranges between approximately 1 and 1.3 g CO_2-equiv./kWh, and the waste management (back-end or downstream) between approximately 0.6 and 1 g CO_2-equiv./kWh. The upstream chain makes the rest, which may substantially change according to the main assumptions for the cycle.

For comparison, the total GHG emissions for the Chinese reference chain were estimated at 9 g CO_2-equiv./kWh using centrifuge technology for all enrichment services. Conversely, taking the extreme assumption of only diffusion enrichment powered by coal plants, the highest GHG emission was calculated at nearly 80 g CO_2-equiv./kWh. When electricity mixes with mixed fuels (also including gas and nuclear) are used, this amount could be halved to about 45 g CO_2-equiv./kWh.

3.3 Renewable Energy Chains

3.3.1 Biomass

For biomass burning boilers and cogeneration systems modeled in the Swiss LCA study, only untreated wood was considered (i.e., no waste wood combustion was accounted for). The emission of direct, biogenic CO_2 from combustion was calculated with a carbon content in dry wood of 0.494% for all types of wood fuels. All carbon absorbed by the trees and contained in the wood that is eventually burned was assumed to be emitted during the combustion, either as CO_2 or as CO. Furthermore, all CO is assumed to be fully oxidized to CO_2 in the atmosphere and as such accounted in the total GHG.

Infrastructure and emission data are corresponding to average operation of modern wood boilers, available on the central European market around year 2000. The wood fuel supply chain represents average central European conditions.

Several classes of heating technologies were modeled, namely 6, 30, and 100 kW for wood log furnace, and 50, 300, and 1000 kW for wood chips furnace, using wood directly from forest or residual wood from industrial processes. Boiler operation was modeled for hardwood (mainly beech), softwood (mainly spruce), and the Swiss commercial residual wood mix (72% softwood, 28% hardwood). Wood pellet furnaces of 15 and 50 kW were also modeled.

For the industrial wood chips, practically all burdens of the processes have been allocated to commercial wood products on the basis of an economic evaluation, rather than to the residues eventually burned in furnaces. This zero-allocation assumption might not be applicable if wood chips would become more important for heat production.

Figure 5 shows the comparison of the results of the Swiss study with an Austrian LCA study, the latter analyzing wood heating systems in Central Europe. Only the systems with high efficiency modeled in the Austrian study are used here, as these have comparable efficiencies as assumed in the Swiss study.

The capacities of the logs and chips boilers used for this comparison are between 6 and 1000 kW for the Swiss study and 10 to 50 kW for the Austrian study. The pellet boilers have similar capacities in the Swiss and Austrian studies. The electric power rates of the CHP plants are between 335 and 400 kWe for the Swiss study and 210 kWe to 36 MWe for the Austrian one.

The total efficiencies of log/chips and pellet boilers are between 68 and 85%. The electric efficiencies of

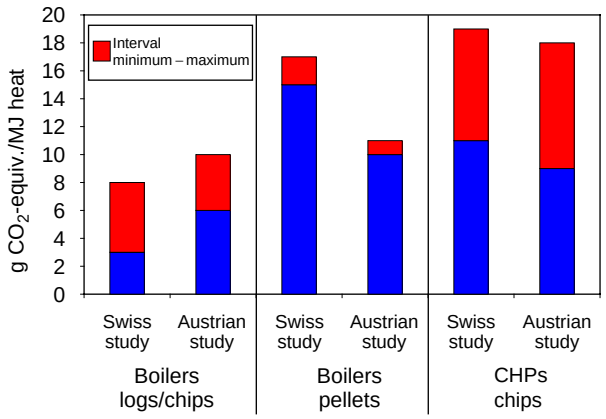


FIGURE 5 Greenhouse gas emissions for wood heating systems. The figure shows the comparison of results of two studies, Swiss and Austrian, analyzing wood heating systems in Central Europe, using LCA. From Swiss study, Dones *et al.* (2003); Austrian study, Jungmeier *et al.* (1999), Treibhausgasbilanz der Bioenergie. Vergleich der Treibhausgasemissionen aus Bioenergie-Systemen und fossilen Energiesystemen. Joanneum Research, Institut für Energieforschung, Graz. Bericht Nr.: IEF-B-06/99.

the cogeneration (CHP) plants in the Swiss study are between 0.9 and 2.3%, thermal efficiencies are between 76 and 78%. The Austrian study reports electric efficiencies of the CHP plants between 8 and 39%, and thermal efficiencies between 51 and 80%.

All values for net GHG emissions of the compared systems are within a relatively small range between 3 g CO_2-equiv./MJ and 19 g CO_2-equiv./MJ. There is no obvious correlation between system capacity and net GHG emissions. Direct CH_4 and N_2O emissions are usually lower at a higher capacity, but the indirect GHG emissions may be higher.

In general, the net GHG emissions from wood chips and wood logs heating systems are not significantly different. Due to incomplete combustion, burning wood logs in automatically controlled furnaces produces more CO and CH_4 than burning chips. Compared to all other heating systems, the (modern) 6 kW fireplaces are emitting relatively high amounts of CO and CH_4.

Compared to other wood boilers, pellet furnaces have higher net GHG emissions, between 15 and 17 g CO_2-equiv./MJ. More than 90% are CO_2 emissions of fossil origin. About 65% of these are due to material and energy uses in the fuel supply chain, and 22% are due to material and energy uses in fuel transport. The main reason is that energy requirements for the pellet production are higher than for chips and logs. Additionally, the transport distances of the pellets to the boilers are usually higher than for chips and logs.

Two wood cogeneration systems of 6400 and 1400 kW_{th} installed in Switzerland and burning industrial chips have been assessed in the Swiss LCA study, each of them with two alternative modes, with multicyclone filter and with enhanced emission control (electric filter + selective noncatalytic reduction (SNCR), assumed to operate with urea). Both systems are mainly designed for heat production (thermal efficiency of nearly 77%); the electricity (efficiency between 1 and 3%) is mostly used for self-consumption.

Allocating all emissions to heat production, 11 g CO_2-equiv./MJ are calculated for the technology with multicyclone-filter and 19 g CO_2-equiv./MJ for the technology with emission control. The reason for the higher GHG emissions of the second technology is that the urea in the SNCR is partially transformed to N_2O. However, the emission factor of this N_2O emission is quite uncertain, and only little information is available in the literature. The interval of N_2O emissions is between 4 and 41 mg/MJ wood input.

3.3.2 Hydro

3.3.2.1 Swiss and European Hydro Power Plants To reflect the situation in Switzerland and Europe, two main types of hydropower plants (storage and run-of-river) and additionally pumped storage plants were analyzed in the Swiss LCA study. A representative sample of four run-of-river power plants in Switzerland and one in Austria was used. Lifetime was assumed to be 80 years for the fixed structures and 40 years for other parts. Net average efficiency is 82% (best efficiency can be 88%).

A representative sample of 52 Swiss concrete dams with a height of more than 30 m was considered for calculating the average. Lifetime was assumed to be 150 years for the dam, 40 years for turbines and pipes, and 80 years for the rest of materials. Net average efficiency, including pipe losses, is 78% (best efficiency can be 84%). The data refer to plant construction of a mix of dam types built between 1945 and 1970; therefore they might not be fully representative for more modern construction, or for a specific dam type, or for an individual unit. The data have been extrapolated to preliminary describe dam mixes in other alpine countries (Austria, Italy, and France) and nonalpine European countries.

No studies seem to exist for CH_4 and N_2O emissions from hydroelectric storage plants in the Alpine region. Hence, the methane emission has been roughly extrapolated from data for natural Swiss lakes, which has been estimated as $4 \cdot 10^{-3}$ kg m^{-2} a^{-1}. Applying

this value as first guess to all Swiss reservoirs, the emission factor would be calculated at 0.014 g CH_4/kWh or 0.32 g CO_2-equiv./kWh. Measurements for N_2O releases from a natural lake in Switzerland have been used for a rough estimation of this emissions from reservoirs, given at $7.7 \cdot 10^{-5}$ g N_2O/kWh or 0.023 g CO_2-equiv./kWh. CO_2 emissions were not accounted for, due to the lack of information on the net balance (see below) for alpine regions.

The Swiss LCA study used available figures for direct GHG emissions from Norwegian and Swedish reservoirs, averaging around 6 g CO_2-equiv./kWh, also for all non-Alpine regions except Finish reservoirs, for which 30 g CO_2-equiv./kWh were assumed.

The results for GHG GWP 100a show that for alpine regions the construction of the dams contributes the most to the total emission of 4 to 5 g CO_2-equiv./kWh. The material (mostly concrete) contributes about 70% to total. With the above assumptions on direct CH_4 emissions, other European reservoirs may exhibit average emissions of 10 g CO_2-equiv./kWh and Finland around 34 g CO_2-equiv./kWh. By increasing GHG fluxes, the emissions associated with the construction of the dam decrease in importance. In case of run-of-river, results are of the order of 3 g CO_2-equiv./kWh.

A survey of several LCA studies performed in 1997 concluded that the indirect contributions approximately range 1 to 10 g CO_2-equiv./kWh for reservoir as well as for run-of-river plants, the difference depending on the plant and site characteristics (type of dam, height/width of dam, capacity of reservoir, location, installed electric capacity, load factor, dedicated transmission lines). However, small hydro may have somewhat higher GHG emission factors.

Pumped storage environmental burdens are a function of the input electricity source and the total efficiency, assumed to be 70%. Hence, results for GHG are strongly depending on the assumption for the input electricity mix. Results of the Swiss LCA study span from the lowest value for Norway (0.027 kg CO_2-equiv./kWh—mix based prevalently on hydro) to 1.62 kg CO_2-equiv./kWh for Poland (mix based mostly on coal). However, these results should be compared only with other systems providing the same service as pumped storage (i.e., peak load).

3.3.2.2 Issues Related to Direct GHG Emissions from Hydroelectric Reservoirs Direct GHG emitted during the operation of a reservoir due to the decomposition of organic matter should be counted for additionally to indirect emissions related to the construction, maintenance, and decommissioning of the plant, in order to compare hydropower with other electricity systems.

Since the late 1980s, research has been invested in the assessment of direct GHG emissions from the air-water interface of reservoirs. Key parameters mentioned for the GHG emission factors are climate (in tropical reservoirs the biodegradation is faster); amount of flooded biomass; nature of the flooded soils; depth of reservoir, which controls methane oxidization; and the ratio energy production to surface.

The state-of-the-art in 2003 for the estimation of GHG emissions from reservoirs comprises the consideration of the emissions from flooded land area only, with contributions from CO_2 diffusive emissions for the initial 10-year period after first filling, CH_4 bubbling and diffusive emissions, N_2O diffusive emissions, and degassing emissions for all previously mentioned species. Ongoing research is aiming at modeling the net emissions, including the entire catchment area and the various phases in the life of the reservoirs in the boreal, temperate, and tropical regions, including pre-impoundment natural GHG emissions, which may be substantial.

Diffusive flow is generally leading to greater GHG emissions than bubbling flow. It is reported that about 99% of CO_2 is released through diffusion, whereas for CH_4 this mechanism releases from 14 to 90% of the total flow. Canadian researchers report that diffusive CO_2 accounts for 80% of total emissions in boreal and temperate sites, whereas methane bubbling makes 60% of total emissions in shallow tropical sites but less than 10% in deep reservoir. In general, N_2O contributes only a few percentage points.

Canadian researchers estimated total GHG emissions (using IPCC 1996 GWP) of 265 g CO_2-equiv. $m^{-2} a^{-1}$ (± 150) from boreal and temperate reservoirs on the basis of measurements of diffusive emission fluxes at the water-air interface of 15 different hydroelectric developments. Data for six tropical reservoirs were summarized giving 4300 g CO_2-equiv. $m^{-2} a^{-1}$ (± 2300) for reservoirs with large shallow water areas (average depth below 10 m) and 1700 g CO_2-equiv. $m^{-2} a^{-1}$ (± 400) for deeper reservoirs (average depth >25 m). The Canadian researchers concluded that reservoirs in tropical regions emit between 5 and 20 times more GHG than in boreal and temperate regions.

Brazilian researchers reported the results of two surveys of measurements in several Brazilian

reservoirs located in different parts of the country. By extrapolating these two surveys to 1-year emissions, they obtained for total CO_2 plus CH_4 the approximate range 1050 to 3930 g CO_2-equiv. $m^{-2} a^{-1}$ (using IPCC 1996 GWP). Methane contributed 7 to 86% of the total GHG emissions. Using the average capacity factor for Brazilian hydroelectric plants of approximately 50%, an interval of 12 to 2077 g CO_2 equiv./kWh from direct emissions in the year of the measurements can be derived for different reservoirs. The average GHG emission factor for direct CO_2 and CH_4 emissions from seven plants, weighted by the energy generation, would be around 340 g CO_2 equiv./kWh. However, it would be inappropriate to use such estimation for the lifetime of the plants, because variations of the fluxes may occur.

On the basis of measurements by several teams of direct emissions in boreal and tropical reservoirs of direct emissions, and of the expected electricity production assuming a load factor of 0.6 and a lifetime of 100 years, Canadian researchers have estimated ranges of GHG emission factors. Boreal and temperate reservoirs would have average GHG emission factors between 0.01 and 0.06 kg CO_2 equiv./kWh, whereas emissions from tropical reservoirs would range between 0.2 and 3 kg CO_2 equiv./kWh. The highest specific emissions were found in shallow (average depth between 4 and 10 m) tropical hydroelectric developments where the ratio of energy production to flooded area is low. A 1997 survey reported for the La Grande Complex in Canada a direct GHG emission of 34 g CO_2-equiv./kWh, based on measurements; this would reduce to 15 g CO_2-equiv./kWh including all sources, assuming gradual return to zero net emissions after 50 years of operation.

Besides decomposition of organic matter, there is evidence of the strong influence of the flooded peat soil for methane emissions in Finn reservoirs, especially if they are shallow. Finnish sources based on direct measurements reported an average range of 65 to 72 g CO_2-equiv./kWh for such reservoirs.

The issue of the net GHG emissions deserves further research in various regions of the world considering the potential developments of hydropower.

3.3.3 Wind

The production of electricity by wind turbines generates no direct GHG emissions. Indirect emissions are following the construction and assembling of the wind power plant, the production and assembling of its materials, transports of materials, and waste disposal processes.

The Swiss LCA study analyzed several wind turbines for Swiss-specific and European average conditions. Plants with 30 kW, 150 kW, 600 kW, and 800 kW were modeled for onshore, and a unit of 2 MW for offshore (based on data from the wind park Middelgrunden, near to Copenhagen). Actual capacity factors in Switzerland range from 8 to 14%. The average for Europe was assumed 20% for onshore and 30% for offshore. The wind mix in Switzerland is mostly composed of 600 and 800 kW turbine, making 57 and 40% of wind electricity in 2002, respectively.

The GHG emissions are decreasing with increasing installed capacity of the turbine. For Swiss conditions, one 800 kW unit emits 20 g CO_2-equiv./kWh, while for average European conditions it decreases to 14 g CO_2-equiv./kWh, for both onshore and offshore. Other studies estimated a typical release of GHG associated to wind turbines in areas with optimal wind speed on the order of 10 g CO_2-equiv./kWh.

For the 800 kW plant, about 72% of GHG emissions stems from the production of materials; about 15% are caused by material processing, 8% by waste disposal, and the rest by transports, final assembling, and installation. The fixed parts (foundation and tower) cause 23% of the CO_2 emissions, the moving parts the rest. For the 2 MW offshore plant, about 81% of the emissions originate from the production of materials. The fixed parts make 44% of the total CO_2 emissions.

3.3.4 Solar

Mono- (mc-Si) and polycrystalline (pc-Si) silicon cell technologies utilized in photovoltaic (PV) panels were analyzed in the Swiss LCA study for Swiss conditions (yield of 885 kWh $kW_{peak}^{-1} a^{-1}$). The plants considered were grid-connected 3 kW_{peak} units installed on buildings in two configurations (i.e., integrated and nonintegrated slanted-roof); also façade and flat-roof were considered. The production chain has been decomposed into detailed steps. The efficiency of the PV cells was assumed to be 16.5% for mc-Si and 14.8% for pc-Si.

Use of electronic-grade silicon from the European electronic industry was assumed as input to the PV industry. The production of metallurgical grade silicon is assumed to occur in Norway (hydro-based electricity mix) and the purification in Germany, with use of a mix of hydropower and a highly efficient combined heat and power plant. To reflect market

conditions, besides electronic grade also a 50% share of off-grade silicon has been considered for the production of the wafers. Energy uses for purification were allocated on the basis of economic value of these two coproducts plus silicon tetrachloride. Changing these assumptions may lead to substantial change of the calculated emissions. In the future, solar grade silicon will probably be used, with associated reduced energy inputs for purification. This production was also modeled in the Swiss study, assuming an increase of cell efficiencies to 17.5% for the mc-Si and 15.7% for the pc-Si systems. Thirty years lifetime was considered for all conditions.

These assumptions lead to a decrease of total GHG emissions compared to previous analyses. Nonintegrated slanted-roof mc-Si panels exhibit about 73 g CO_2-equiv./kWh, whereas for pc-Si technology this value is 59 g CO_2-equiv./kWh. Integrated types have slightly lower LCA emissions. For future conditions, the emissions may decrease to 46 g CO_2-equiv./kWh and 39 g CO_2-equiv./kWh for the mc-Si and pc-Si technologies, respectively.

3.3.5 Heat Pumps

An attractive alternative to fossil fuel fired boilers is the use of heat pumps. Heat pumps can gain heat either from the ambient air, from the ground, or from groundwater. The seasonal performance factor (SPF) of an electric heat pump is the ratio between the annual heat energy output and the annual electric energy input. The SPF depends on the efficiency of the heat pump (coefficient of performance, COP), but also on the local climatic conditions and on the integration of the heat pump into the building (e.g., low- or high-temperature hydronic heat distribution system). Field measurements showed that the average SPF in Switzerland in the year 1998 was about 3.2. Air/water heat pumps using the ambient air had an average SPF of 2.8, brine/water heat pumps using ground heat had an average SPF of 3.9.

The total GHG balance of an electric heat pump depends strongly on the SPF of the heat pump and on the electricity mix used for operation. A heat pump is almost CO_2-free if an almost CO_2-free source of electricity (renewables or nuclear) can be used. But heat pumps can also improve the GHG balance of heating when electricity from fossil fuels is used. A combination of a heat pump with a seasonal performance factor of 3.5 and a cogeneration plant with an electric efficiency of 35%, a total efficiency of about 90%, and electric transmission losses of 2.5% yields about 174% heat relative to the fossil fuel input. The most modern natural gas combined-cycle power plants in Europe can reach a net electric efficiency of 57 to 58%. A combination of a 57.5% CC power plant with an SPF 3.5 heat pump yields a heat efficiency of about 200%.

On the negative side, there are the emissions of refrigerants from heat pumps (besides the indirect emissions from material production and maintenance valid also for other systems). The HFC refrigerant emissions from heat pumps over the lifetime are roughly 2 to 5 mg/MJ. For a heat pump with refrigerant R134a (HFC-134a), which has a 100-years global warming potential of 1300, this corresponds to GHG emissions of about 3–6 g/MJ CO_2-equiv. 100-year. Compared to the CO_2 emissions of a natural gas boiler in the order of 60 g/MJ, heat pumps still can provide a GHG reduction potential if the electricity supply is favorable from the GHG point of view.

Assuming a market penetration of 30% in the building sector using presently available technology, the IEA Heat Pump Centre estimated the CO_2 reduction potential of heat pumps being about 6% of the total global CO_2 emissions from fossil fuels in 1997. This does not include a full LCA analysis and insofar has to be viewed as a rough estimate.

4. COMPARISON OF GREENHOUSE GAS EMISSIONS

The energy sector plays a dominant role with regard to worldwide emissions of GHGs, in particular CO_2. About 77% of yearly CO_2 emissions worldwide originate from fossil fuels combustion. The contribution of the energy sector to the emission of other most important GHGs (i.e., CH_4 and N_2O) is much smaller in relative terms but still important. Based on the total GWP of all GHGs emitted yearly worldwide, the overall share of the energy sector in emissions relevant for global warming is of the order of 50%.

4.1 Current Technologies

The results presented in Section 3 for the individual electricity systems are summarized here. This focuses on the results of the Swiss LCA study for Switzerland and Europe, which cover a broad spectrum of options in a consistent manner. Some conclusions are provided. For a number of detailed comparisons of results for specific chains with other studies and for special issues, refer to Section 3.

Figure 6 shows the overall results with contribution of the various GHGs. In Figure 7 the same overall results are provided but with differentiation between contributions from power plant and rest of the chain. Figures 6 and 7 are complemented by the ranges of GHG emissions from European country-specific energy systems provided in Table V.

1. *Electricity systems ranking based on GHG emissions.* GHG emissions per unit of generated electricity are typically highest for industrial gas, followed by lignite, hard coal, oil, and natural gas. Hydro exhibits very low GHG emissions in boreal and temperate regions, in most cases two orders of magnitude lower than coal. However, hydroelectric developments in tropical regions may emit during

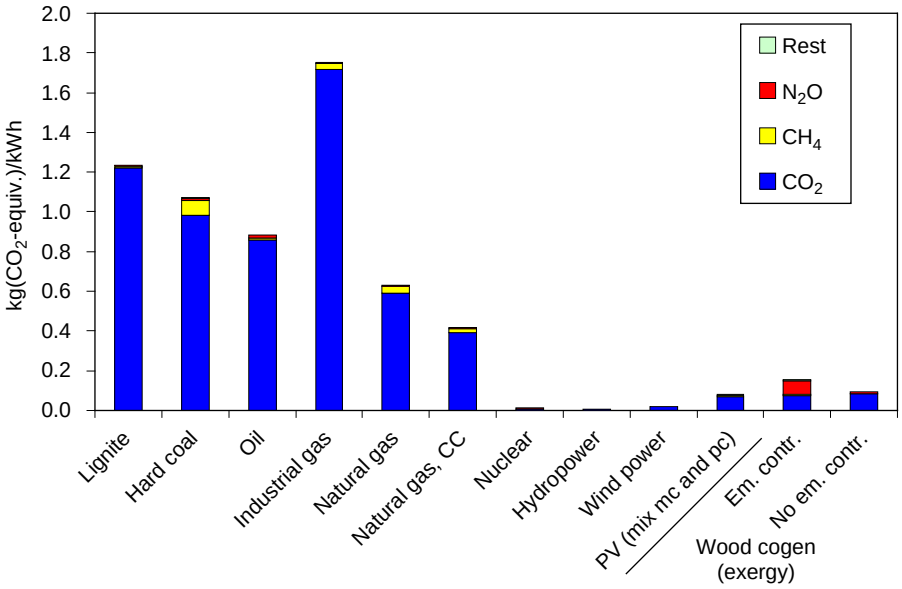


FIGURE 6 Overview of full chain greenhouse gas emissions for different electricity generation systems. The figure shows the contributions of the different greenhouse gases in kg CO_2-equivalents (100 years) per kWh_e. Lignite, hard coal, oil, industrial gas, natural gas, and nuclear refer to energy systems associated with average UCTE power plants around year 2000. Natural gas CC: Combined-cycle power plant of currently best technology, gas supplied by the European high-pressure grid. The renewables (hydropower, wind power, photovoltaic, wood) refer to location Switzerland. From Dones *et al.* (2003).

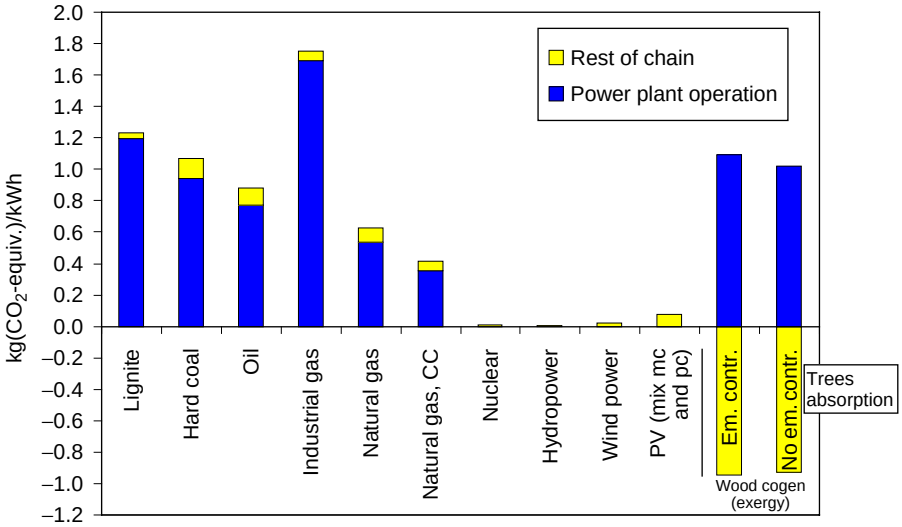


FIGURE 7 Greenhouse gas emissions from power plant operation and from the rest of chain for different electricity generation systems. For explanations see Fig. 6 and text.

TABLE V

Ranges of GHG Emissions per kWh Electricity from European Country-Specific Energy Sources, using GWP 100 years

	Minimum (kg CO_2-equiv./kWh)	Maximum (kg CO_2-equiv./kWh)
Lignite	1.060	1.690
Hard coal	0.949	1.280
Oil	0.519	1.190
Industrial gas	0.865	2.410
Natural gas	0.485	0.991
Nuclear power	0.008	0.011
Hydropower[a]	0.003	0.027
Wind power[b]	0.014	0.021
PV (mix mc and pc)[c]	0.079	—
Wood cogeneration[d]	0.092	0.156

[a] Mix of reservoirs and run-of-river plants.

[b] Calculated for average conditions of Switzerland (maximum) and Western Europe (minimum).

[c] Calculated for average Swiss conditions.

[d] 6400 kW_{th}/400 kW_e plant; allocation exergy; no emission control (minimum); emission control SNRC (maximum).

Source. Dones *et al*. (2003).

operation between 5 and 20 times more GHGs, which at the higher range is comparable to emissions from fossil sources. Taking into account full energy chain contributions including also the grey ones, GHG emissions from nuclear and wind energy (under favorable wind conditions) are in the same low range as typical for hydro. The corresponding net emissions for biomass are in the middle range (i.e., one order of magnitude lower than coal and one order of magnitude higher than nuclear and wind). The same applies to older solar photovoltaic systems in countries with climate conditions such as in central Europe, while current photovoltaic technologies exhibit even smaller differences compared to nuclear and wind; still better performance can be achieved in southern countries.

2. *Differences between industrialized and developing countries*. Based on quite limited comparative material for fossil systems, a pattern can be distinguished, namely the GHG emissions are significantly higher in the developing countries. This is due to use of technologies characterized by lower efficiency as well as distinct features of fuel cycles (relatively large methane emissions from mines and from gas pipelines; coalfield fires in China).

3. *Dominant GHGs*. CO_2 dominates GHG emissions from energy systems. In the case of hard coal and natural gas, the methane contribution in terms of CO_2-equivalents may reach the level of 10% due to high leakage rates from coal mines and gas pipelines. N_2O may be of high relative importance for wood cogeneration, if pollutant emission controls employing catalysts generating N_2O are applied.

4. *Role of GHG emissions other than from power plant*. For fossil systems power plant GHG emissions are clearly dominant. It can be noted that in the case of major pollutants, the situation may be quite different—that is, emissions from power plants are higher than the other contributions, but the latter may be very substantial. For European conditions the upstream part of the chain typically contributes about 10% of the total GHG emissions for the hard coal chain, slightly more than 10% for oil chain, and can approach 20% for conventional natural gas systems. For nuclear, wind and solar photovoltaic, the GHG emissions other than from power plant are totally dominant in relative terms. The same applies to typical hydro plants, with the exception of dams operating under special conditions resulting in very large direct emissions from reservoirs.

For comparisons of heating and cogeneration systems, refer to Figs. 3, 4, and 5. Gas boilers are as expected better performers than oil boilers when modern systems are compared. Biomass exhibits best GHG performance along with heat pumps, provided that electricity input to the latter comes from zero or low carbon sources. Cogeneration based on biomass is clearly favorable in terms of GHG emissions compared to the corresponding gas and oil systems. Evaluation of cogeneration is subject to variability, depending on which allocation scheme is employed.

4.2 Electricity Mixes

Calculating and presenting GHG emission data for single technologies is important for comparisons and decision making. However, it is also essential to generate the total burdens associated with the electricity mixes. These can then be used in LCA studies of products.

Two approaches are used to generate the relevant mixes (i.e., marginal and average mix approach). The marginal approach suggests to use the last available technology or mix of technologies (e.g., the cheapest available for base load, or the technology[ies], which will most probably cover an increase in demand, or the specific energy technology which will be chosen for a process/industry), whereas with the average

approach only average mixes supplying the grid are considered.

The Swiss LCA study addressed national production mixes for all European countries and the three major associations of electric transmission operators UCTE, CENTREL, and NORDEL in year 2000. GHG results for these electricity mixes are summarized in Table VI.

Depending on the shares of technologies, the total GHG emissions can vary by several orders of magnitude from the hydro-based Norwegian mix to the coal-based Polish mix. Additionally, the study addressed the issue of the supply mix. This was modeled by merging all imports to the domestic production mix and thus recalculating the share of technologies to be used both for the domestic supply mix as well as for the export mix. This model may somewhat overestimate the effective import shares and may not describe seasonal exchanges properly. The results for Switzerland, a country with intense exchanges of electricity with neighboring countries due to its central position in Europe, are illustrated in Table VII.

The GHG-free production mix based on hydro and nuclear (58 and 38%, respectively, in year 2000) changes to a supply mix with quite significant fossil components due to imports from Germany. Another issue concerns the influence of the grids (losses and material use) to the GHG associated to the electricity services. The analysis performed for the Swiss conditions leads to increases of GHG by 2.5, 5, and 18% with respect to the supply mix at the busbar of plants in the case of high, medium, and low voltage levels, respectively. With equal network losses, the increases may be lower in percent in case of country supply mixes with higher fossil share.

A Japanese study covered the electricity production mixes for 10 Japanese electric companies. Besides direct emissions from fossil plants, the authors included the emissions from the upstream chains, but excluded construction and decommissioning of the various facilities and transport systems. Only CO_2 and CH_4 were considered. Using GWP from IPCC 2001, the results are in the range of 296 to 795 g CO_2-equiv./kWh for nine companies, whose average is calculated at 468 and 1079 g CO_2-equiv./kWh for Okinawa. Remarkably, the UCTE (2000) and the Japanese production mixes have practically the same GHG intensity.

TABLE VI

LCA-Based GHG Emissions for European Electric Associations and Countries

Electricity mix	kg CO_2-equiv./kWh
Production mix (at busbar)	
UCTE	0.470
CENTREL	0.932
NORDEL	0.132
Country-specific production mix[a]	0.007–1.13
Country-specific supply mix[a,b]	0.010–1.11

[a] Minimum: Norway; maximum: Poland.

[b] The used model adds all electricity imports to the domestic production, recalculates the shares of technologies on this basis, and applies this portfolio to domestic supply and exports.

Source. Dones *et al*. (2003).

4.3 Advanced Technologies and Sequestration

4.3.1 Advanced Technologies

The LCA methodology has been originally developed and primarily applied for operating systems. Consequently, the input is normally based on the actual experience. Furthermore, the standard approach is static and its application to future systems requires extensions, extrapolations, and a number of additional assumptions on technological advancements and structural changes.

Few studies have implemented extensions of LCA applications to future systems, based on literature as well as direct information from the industry and on expert judgment. The result-driving parameters in the LCA study are emissions, efficiencies, material intensities (for construction and operation), and transportation requirements. The relative importance of these parameters varies significantly between energy chains.

The most important expected changes towards improvements of ecological performance identified in

TABLE VII

LCA-Based GHG Emissions for Swiss Electricity Mixes

Electricity mix	kg CO_2-equiv./kWh
Production mix (at busbar)	0.018
Supply mix (with import/export)[a]	0.120
High voltage	0.123
Medium voltage	0.126
Low voltage	0.142

[a] See note *b* in Table VI.

Source. Dones *et al*. (2003).

the 1996 Swiss study on evolutionary future electricity generating technologies were as follows:

- *Gas systems*. Reduction of gas leakage, improvements of power plant burner performance characteristics and of overall power plant efficiency.
- *Hard coal systems*. Increased methane recovery in underground mining, improvements in power plant abatement technology, and of overall power plant efficiency.
- *Nuclear systems*. Reductions of electricity consumption in enrichment by replacement of diffusion by centrifuges or laser technologies, power plant improvements (particularly extended life time and increased burn-up).
- *Hydro systems*. Overall power plant efficiency improvements (turbine).
- *PV systems*. Improvements in the manufacturing of mono-crystalline-silicon (m-Si) and amorphous-silicon (a-Si) solar cells (yield, electricity consumption), and in cell efficiencies.

Some of the improvements mentioned have already been implemented in the best technologies available around year 2000, which are not yet widely disseminated.

For electricity inputs needed for the LCA modules external to Switzerland, a European mix for year 2010 was defined, based on a forecast by the International Energy Agency. The mix reflects the expansion of gas, reduction of oil shares, and a relatively small but significantly increased contribution of renewables other than hydro. Coal, hydro, and nuclear remain at about the same level percent. Since the new systems generally show better performance and lower emissions than the old ones, assumptions needed to be made with respect to the market penetration of the new ones. This was done individually for each energy source, taking into account their specificity and the expected developments reflected by the assumed overall contribution of each energy carrier to the mix.

Future electricity generation systems based on fossil chains have the potential to reduce GHG emissions by 20 to 40% in comparison with the corresponding present average European fossil systems. However, they remain the largest GHG producers: two orders of magnitude higher than hydro in Alpine regions, wind and nuclear, and one to two orders higher than photovoltaic.

4.3.2 Sequestration

Since CO_2 emissions into the atmosphere should be reduced, storage of CO_2 in geological reservoirs or in the ocean is a possible option under discussion. The sequestration procedure includes capture of CO_2 from the fuel combustion device, dehydration, compression, transport, and injection into the final reservoir.

The fate of CO_2 in ocean is not yet fully understood. Storage in geological reservoirs seems to be a more promising method, but also in this case there are open issues to be investigated like the permanency of the storage. Possible environmental risks of storage in saline aquifers include leakages, catastrophic releases (although considered unlikely), changes in geochemistry, water quality, and ecosystem health caused by pH shifts due to carbonic acid, or contamination of potable water due to displacement of brine.

Table VIII shows estimates of the storage capacities of possible geological reservoirs. Depleted gas and oil fields could store only about 40% of the CO_2 emissions expected between year 2000 and year 2050 (though this does not include fields that are not yet operating). The actual storage capacity of saline aquifers under reasonable economic conditions is highly uncertain.

The main techniques for CO_2 capture are postcombustion capture, precombustion capture, and oxy-fuel combustion. Postcombustion capture of CO_2 from the flue gas usually involves chemical solvents like monothanolamine (MEA). In precombustion capture, the fuel is converted into CO_2 and hydrogen; the hydrogen is then combusted. Oxy-fuel combustion means an increase of CO_2 concentration in the flue gas by using concentrated oxygen (which is, e.g., produced by cryogenic air separation) for combustion.

CO_2 capture reduces the net efficiency of a power plant and thus increases the fuel consumption per kWh delivered to grid. Using current technology, the

TABLE VIII

Global CO_2 Storage Capacities of Geological Reservoirs

CO_2 emissions and storage capacities	Gt CO_2
Emissions 2000–2050 (IPCC business as usual scenario)	2000
Theoretical storage capacities of geological reservoirs	
Depleted gas fields	690
Depleted oil fields	120
Deep saline aquifers	400–10,000
Unminable coal seams	40

Source. Gale (2002). (Storage capacities based on injection costs of up to 20 US $/ton CO_2; gas and oil fields exclude fields that are not yet producing.)

changes would be significant. For a natural gas combined cycle power plant, the fuel consumption is expected to increase by 16 to 28% due to CO_2 capture. The respective figures for coal power plants have been estimated to be 22 to 38% for pulverized coal and 16 to 21% for integrated gasification combined cycle (IGCC). Capital costs would increase by 300 to 800 US$/kW, or about 30 to 50% for IGCC, about 70 to 80% for pulverized coal, and about 80 to 100% for natural gas CC.

SEE ALSO THE FOLLOWING ARTICLES

Acid Deposition and Energy Use • Biomass: Impact on Carbon Cycle and Greenhouse Gas Emissions • Carbon Capture and Storage from Fossil Fuel Use • Carbon Sequestration, Terrestrial • Climate Change and Energy, Overview • Diet, Energy, and Greenhouse Gas Emissions • Ecosystem Health: Energy Indicators • Greenhouse Gas Abatement: Controversies in Cost Assessment • Greenhouse Gas Emissions, Alternative Scenarios of

Further Reading

Bartlett, K. B., *et al.* (1990). Methane flux from the Amazon River Floodplain. *J. Geophys. Res.* Sept. 20, 16773–16788.

Dones, R., Zhou, X., and Tian, C. (2003). Life cycle assessment. *In* "Integrated Assessment of Sustainable Energy Systems in China—The China Energy Technology Program" (Eliasson, B., and Lee, Y. Y., Eds.), pp. 320–344. Kluwer.

Dones, R., Bauer, C., Bolliger, R., Burger, B., Faist Emmenegger, M., Frischknecht, R., Heck, T., Jungbluth, N., and Röder A. (2003). Life cycle inventories of energy systems: Results for current systems in Switzerland and other UCTE countries. Final report ecoinvent 2000 No. 5. Paul Scherrer Institut, Villigen, and Swiss Centre for Life Cycle Inventories, Duebendorf, Switzerland. Online version under www.ecoinvent.ch.

Dones, R., Gantner, U., Hirschberg, S., Doka, G., and Knoepfel, I. (1996). Environmental inventories for future electricity supply systems for Switzerland. PSI Report No. 96-07, Villigen, Switzerland.

Duchemin, E., Lucotte, M., St-Louis, V., and Canuel, R. (2002). Hydroelectric reservoirs as an anthropogenic source of greenhouse gases. *World Resource Rev.* **14**(3), 334–353.

Fearnside, P.M. (2002). Greenhouse gas emissions from a hydroelectric reservoir (Brazil's Tucuruí dam) and the energy policy implications. *Water, Air, and Soil Pollution* 133, 69–96.

Gale, J. (2002). Proceedings of "Workshop on carbon dioxide capture and storage," Nov. 18–21. IPCC. Regina, Canada.

Hondo, H. (2003). Life cycle GHG emission analysis of power generation systems: Japanese case. Submitted to *Energy—The International Journal.*

International Energy Agency (IEA) (1999). CO_2 emissions from fuel combustion 1971–1997. OECD/IEA, Paris.

Intergovernmental Panel on Climate Change (IPCC) (2001). "Climate Change 2001: The Scientific Basis." Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change (Houghton, J. T., *et al.*, Eds.). Cambridge Univ. Press, Cambridge, UK.

Jungmeier *et al.* (1999). Jungmeier, G., Canella, L., Spitzer, J., and Stiglbrunner, R. Treibhausgasbilanz der Bioenergie. Vergleich der Treibhausgasemissionen aus Bioenergie-Systemen und fossilen Energiesystemen. Joanneum Research, Institut für Energieforschung, Graz. Bericht Nr.: IEF-B-06/99.

Kato, Y., Inaba, A., and Mayasuki, S. (2000). The CO_2 emissions for fossil fuels from producing countries to Japan. Proceedings of APEC/AIST/NEDO symposium "LCA for APEC Member economies," H12.11.1-2, pp. 87–92. Tsukuba, Japan.

Matsuno, Y., and Betz, M. (2000). Development of life cycle inventories for electricity grid mixes in Japan. *Int. J. LCA* **5**(5), 295–305.

Pinguelli Rosa, L., dos Santos, A., Matvienko, B., and Sikar, E. (2002). Carbon dioxide and methane emissions from Brazilian hydroelectric reservoirs. First Brazilian Inventory of Anthropogenic Greenhouse Gas Emissions—Background Reports, Alberto Luiz Coimbra Institute for Graduate Studies and Research in Engineering COPPE, Brazilian Ministry of Science and Technology, Brasilia.

Spath, P., Mann, M., and Kerr, D. (1999). Life-cycle assessment of coal-fired power production, National Renewable Energy Laboratory, NREL/TP-570-25119, Golden, CO.

Van de Vate, J. F., and Gagnon, L. (1997). Greenhouse gas emissions from hydropower—The state of research in 1996. *Energy Policy* **25**(1), 7–13.

Van de Vate, J. F. (1997). Comparison of energy sources in terms of their full energy chain emission factors of greenhouse gases. *Energy Policy* **25**(1), 1–6.

Ground-Source Heat Pumps

ARIF HEPBASLI
Ege University
Izmir, Turkey

1. Ground-Source Heat Pumps
2. Worldwide Ground-Source Heat Pump Installations
3. Main Components of a Ground-Source Heat Pump System
4. Design Types of Ground Heat Exchangers
5. Soil Characteristics
6. Groundwater Heat Pumps
7. Surface Water Heat Pumps
8. Hybrid Ground-Source Heat Pump Systems
9. Performance Evaluation of Ground-Source Heat Pumps

Glossary

Air Conditioning and Refrigeration Institute (ARI) The nonprofit organization that sets the standards for testing and comparisons for the heating, air-conditioning, and refrigeration industries.
aquifer A water-bearing geological formation; a large body of underground water.
coefficient of performance (COP) The ratio of heat energy delivered or extracted to work to operate the equipment.
compressor The device that, through compression, increases the pressure and temperature of entering refrigerant, and discharges a hot dense gas.
condenser A heat-exchange coil within a mechanical refrigeration system; used to reject heat from the system.
energy efficiency ratio (EER) The ratio of the rated cooling capacity (in British thermal units per hour) divided by the amount of electrical power used (in watts) for any given set of conditions.
entering water temperature (EWT) The temperature of the heat transfer fluid leaving the ground heat exchanger.
evaporator A heat-exchange coil within a mechanical refrigeration system; used to absorb heat into the system.
ground heat exchanger (GHE) A device that exchanges ground heat; thermal energy is extracted from (or rejected into, depending on mode of operation) the ground.
heat exchanger A device used to transfer heat from a fluid (liquid or gas) to another fluid; the two fluids are physically separated (usually by metal tubing).
heating seasonal performance factor (HSPF) The ratio of the total amount of heat energy supplied to the space for the heating season with respect to the total electrical energy required by the heating system for the same season.
heat pump A device that uses a refrigerant cycle to take low-quality heat, concentrate it to a higher quality heat, and move that heat to another location in the cycle.
heat sink The area or media where heat is deposited (inside a building, etc.).
heat source The area or media from which heat is removed (water, air, etc.).
heat transfer fluid A fluid (water, or a mixture of water and antifreeze) that is circulated through the ground heat exchanger via a circulating pump.
refrigerant A fluid (liquid or gas) that picks up heat by evaporating at a low temperature and pressure.
reversing valve A valve that changes the direction of refrigerant flow in a heat pump.
seasonal energy efficiency ratio (SEER) The average energy efficiency ratio over the entire cooling season.

Ground-source heat pumps (GSHPs), also known as geothermal heat pumps (GHPs), use electricity to move heat, not to generate it, and are recognized to be outstanding heating, cooling, and water-heating systems. The primary benefit of GSHPs compared to other systems is the increase in operating efficiency, which translates to reduced heating and cooling costs, but there are additional advantages. One notable benefit is that GSHPs are classified as renewable energy technology, although they are electrically driven. According to the U.S. Environmental Protection Agency (EPA), well-designed and properly installed high-efficiency GSHP systems produce less environmental harm as compared to any other alternative space-conditioning technology currently available. On a full fuel cycle basis, these systems are the most efficient technology available, with the lowest CO_2 emissions for minimum greenhouse warming impact.

 © 2004 Elsevier Inc. All rights reserved.

1. GROUND-SOURCE HEAT PUMPS

Ground-source heat pumps are not a new idea. Patents on the technology date back to 1912, in Switzerland. One of the oldest GSHP systems in the United States, in the United Illuminating headquarters building in New Haven, Connecticut, has been operating since the 1930s.

There are two main types of GSHP systems: ground-coupled (vertical or horizontal) closed loop and water-source (groundwater) open loop, as shown in Figs. 1 and 2, respectively. Ground-coupled heat pumps (GCHPs) are known by a variety of names. These include ground-source heat pumps, earth-coupled heat pumps (ECHPs), earth energy heat pumping systems, earth energy systems, ground-source systems, geothermal heat pumps, closed-loop heat pumps, solar energy heat pumps, geoexchange systems, geosource heat pumps, and a few other variations. In marketing ground-source heat pumps, to cope with the diverse terminology, sales personnel may wish to connect GCHPs to renewable energy sources (solar, geothermal), to connect them to environmental awareness (earth energy), and to dissociate them from air heat pumps (ground-source systems). However, two terms are commonly used to describe the technology in general: geothermal heat pumps and ground-source heat pumps. The former term is typically used by individuals in marketing and government, and the latter is used by engineering and technical types. In addition, ground-coupled, groundwater, and surface water heat pumps are referred to as subsets of ground-source heat pumps. The systems will be referred to as ground-source heat pumps throughout this text.

The closed-loop systems use a buried earth coil as the ground heat exchanger (GHE) through which a heat transfer fluid, typically an antifreeze solution, is circulated. The GHE, installed either vertically in borings or horizontally in trenches, exchanges heat with the ground (Fig. 1). In open-loop systems, the groundwater is pumped into the heat pump unit, where heat is extracted from (or rejected into, depending on mode of operation) the water; the water is then disposed of in an appropriate manner. If possible, releasing the water into a stream, river, lake, pond, ditch, or drainage tile, known as open discharge, is the easiest and least expensive disposal method. A second means of water discharge is using a return well that returns the water to the ground aquifer (Fig. 2).

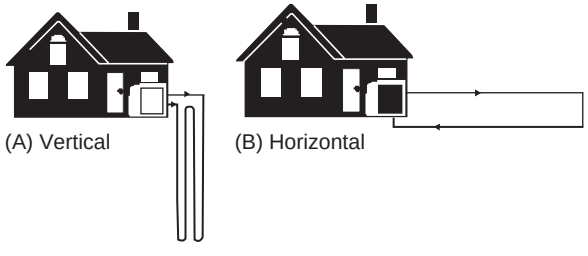


FIGURE 1 Ground-coupled (closed-loop) systems. Reproduced by permission from the Geo-Heat Center, Oregon Institute of Technology.

2. WORLDWIDE GROUND-SOURCE HEAT PUMP INSTALLATIONS

Among the worldwide applications of geothermal energy for direct utilization, GSHPs have had the largest rate of growth since 1995, almost 59%, representing 9.7% annually. Most of this growth has occurred in the United States and Europe, though interest is developing in other countries, such as Japan and Turkey. The installed capacity is 6875 thermal-derived megawatts (MW_t), and the annual energy use is 23,287 terajoules (TJ)/year, or 6453 gigawatt-hours (GWh)/year, at the beginning of 2000 in 27 countries, as shown in Table I. It is estimated that the actual number of installed units is around 500,000, and the equivalent number of 12-kW units installed is slightly over 570,000. It is estimated that 450,000 units are presently installed in the United States; thus, this rate of installation would add an additional 1.1 million units for a total of about 1.5 million units by 2010.

The world's largest GSHP installation is the Galt House East Hotel and Waterfront Office Buildings in Louisville, Kentucky. A 4700-ton GSHP system is used to meet the heating and cooling needs of the complex. The 70,000 m^2 (750,000 ft^2) Galt House East Hotel was completed in 1984 and uses a

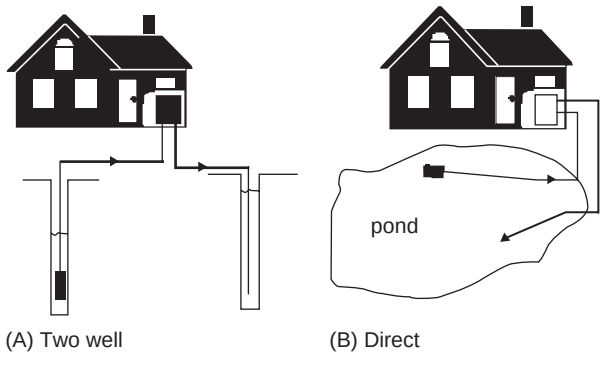


FIGURE 2 Groundwater (open-loop) systems. Reproduced by permission from the Geo-Heat Center, Oregon Institute of Technology.

TABLE I

Worldwide Geothermal Heat Pump Installations in 2000

Country	MW_t[a]	TJ/year	GWh/year	Actual no.	Equiv. no. (12 kW)
Australia	24	57.6	16.0	2000	2000
Austria	228	1094	303.9	19,000	19,000
Bulgaria	13.3	162	45.0	16	1108
Canada	360	891	247.5	30,000	30,000
Czech Republic	8.0	38.2	10.6	390	663
Denmark	3	20.8	5.8	250	250
Finland	80.5	484	134.5	10,000	6708
France	48	255	70.8	120	4000
Germany	344	1149	319.2	18,000	28,667
Greece	0.4	3.1	0.9	3	333
Hungary	3.8	20.2	5.6	317	317
Iceland	4	20	5.6	33	33
Italy	1.2	6.4	1.8	100	100
Japan	3.9	64	17.8	323	323
Lithuania	21	598.8	166.3	13	1750
Netherlands	10.8	57.4	15.9	900	900
Norway	6	31.9	8.9	500	500
Russia	1.2	11.5	3.2	100	100
Poland	26.2	108.3	30.1	4000	2183
Serbia	6	40	11.1	500	500
Slovak Republic	1.4	12.1	3.4	8	117
Slovenia	2.6	46.8	13.0	63	217
Sweden	377	4128	1146.8	55,000	31,417
Switzerland	500	1980	550.0	21,000	41,667
Turkey	0.5	4.0	1.1	23	43
UK	0.6	2.7	0.8	49	53
USA	4800	12,000	3333.6	350,000	400,000
Total	6875.4	23,286.9	6453.1	512,678	572,949

Source. Data from Lund (2000).
[a] Thermally derived megawatts.

1700-ton GSHP system. The installation cost was $1500 per ton. In comparison, a conventional system would have cost between $2000 and $3000 per ton. The hotel complex energy use is approximately 53% of a similar non-GSHP system in an adjust unit, and the system saves about $25,000 monthly in reduced energy costs. The Waterfront Office Buildings, built in 1994, add about 89,000 m^2 (960,000 ft^2) of office space and almost 3000 tons of GSHP capacity to the project, making this the world's largest commercial GSHP project. It is reported that Galt House East has been operating for 15 years with no system problems and that the GSHP system has performed even better than expected.

One of the largest commercial GSHP systems running today is at Stockton College in Pomona, New Jersey, where 63 GSHPs with a total capacity of 1655 tons (5825 kW) are connected to a GHE, which consists of 400 wells, each 129 m (425 ft) deep.

3. MAIN COMPONENTS OF A GROUND-SOURCE HEAT PUMP SYSTEM

GSHPs use the relatively constant temperatures, typically 7° to 21°C (45° to 70°F), of the soil and water beneath the frost line to provide efficient heating and cooling all year long. The efficiencies achieved by these systems are impressive and allow commercial users to save up to 50% over conventional heating and cooling systems, plus they reduce

maintenance costs. GSHPs use the stable temperature of the ground as a heat source to warm buildings in winter and as a heat sink to cool them in summer.

A GSHP system for space heating and cooling of houses consists of three distinct subsystems (see Fig. 3):

1. Earth connection or ground heat exchanger. Many types of GHE designs are possible, but for GSHP systems, only two basic design concepts have been implemented so far. They are direct-expansion GHEs, which consist of a buried copper piping network through which refrigerant is circulated, and secondary fluid circulating GHEs, through which a water or a water–antifreeze solution is circulated. The GHE is buried in the ground near the building to be conditioned. Heat transfer fluid is circulated through the GHE via a circulating pump (stage 1, in Fig. 3).

2. Heat pump unit. This device is composed of five primary components (Fig. 3): a compressor (stage 2), a refrigerant–air (or water) heat exchanger (stage 3), an expansion valve (stage 4), and a heat exchanger between the refrigerant and GHE fluid (stage 5); there is also a reversing valve (stage 6). In the heating mode, heat is extracted from the ground by the GHE. This heat is transferred to the refrigerant in the refrigerant–fluid heat exchanger, is upgraded in the heat pump cycle, and is supplied to the house by the refrigerant–air heat exchanger, which acts as the condenser. In the cooling mode, the process is reversed to extract heat from the house and reject it to the ground. The most widely used type of unit is a water-to-air heat pump, which circulates a water or water–antifreeze solution through a liquid-to-refrigerant coil and a GHE. Air to be heated or cooled is circulated through a conventional finned-tube air-to-refrigerant coil and duct system. Water-to-water heat pumps are also used by replacing the forced air system with a hydronic loop. Systems using water-to-air and water-to-water heat pumps are often referred to as GCHPs with secondary solution loops, to distinguish them from direct-expansion GCHPs.

3. Heat distribution. There are two basic ways to distribute the energy extracted from the earth and water, namely, forced-air systems and hydronic systems. The simplest way to cool a building is to circulate air by using fans to blow the air through coils. This kind of system is not suitable for heating, however, unless the earth or water source is usually warm. In addition to space conditioning, GSHPs can be used to provide domestic hot water when the system is operating.

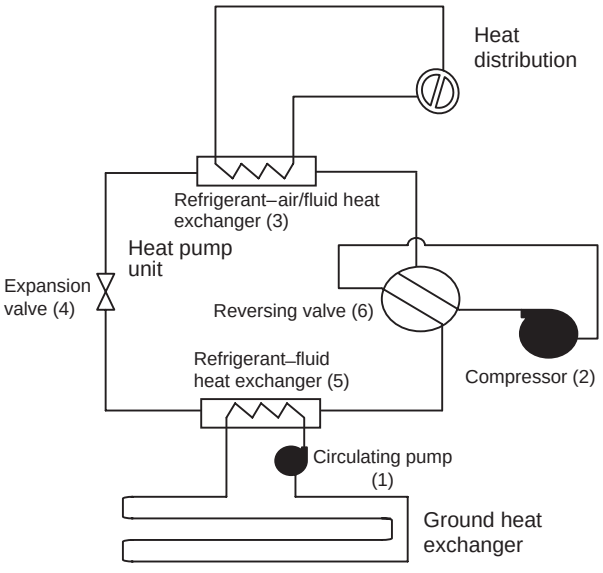


FIGURE 3 A schematic of a ground-source heat pump system, showing stages 1 through 6 (see text for discussion).

4. DESIGN TYPES OF GROUND HEAT EXCHANGERS

The type of GHE used will affect heat pump system performance (therefore, the heat pump energy consumption), auxiliary pumping energy requirements, and installation costs. Choice of the most appropriate type of GHE for a site is usually a function of specific geography, available land area, and life cycle cost economics. Primarily, only two types of GHE designs are used, vertical and horizontal.

4.1 Vertical Ground Heat Exchangers

The vertical ground heat exchanger (VGHE), illustrated in Fig. 1A, is generally considered when land surface is limited. The VGHE consists mainly of two small-diameter high-density polyethylene (PE) tubes that have been placed in a vertical borehole that is subsequently filled with a solid medium. The tubes are thermally fused at the bottom of the bore to a close-return U-bend. Vertical tubes range from 0.75 to 1.5 inches in nominal diameter, and bore depths vary from 15.2 to 183 m (50 to 600 ft), depending on local drilling conditions and available equipment. Multiple wells are typically required, with well spacing not less than 4.6 m (15 ft) in northern climates and not less than 6.1 m (20 ft) in southern climates to achieve the total heat transfer requirements. A minimum base separation distance of 6.1 m

(20 ft) is recommended when loops are placed in a grid pattern. This distance may be reduced when the bores are placed in a single row, when the annual heating load is much greater than the annual cooling load, or when vertical water movement mitigates the effect of heat buildup in the loop field.

4.2 Horizontal Ground Heat Exchangers

Horizontal ground heat exchangers, illustrated in Fig. 1B, are most often used for small installations where space is not a premium. Pipes are buried in trenches, typically at a depth of 1.2–3.0 m (4–10 ft). Depending on the specific design, as many as six pipes may be installed in each trench, with about a foot separating the pipes, and trenches typically being 3.7–4.6 m (12–15 ft) apart. The "slinky," a spiral loop comprising overlapping coils of pipe, can be used to maximize the number of feet of pipe buried per foot of trench, but slinkys require more feet of pipe per ton of thermal energy needed, compared to conventional horizontal systems.

Another variation of the spiral-loop system involves placing the loops upright in narrow vertical trenches. The spiral-loop configuration generally requires more piping, typically 43.3–86.6 m/kW per system cooling ton (500–1000 ft), but less total trenching as compared to the multiple horizontal-loop systems just described. For the horizontal spiral-loop layout, trenches are generally 0.9–1.8 m (3–6 ft) wide; multiple trenches are typically spaced about 3.7 m (12 ft) apart. For the vertical spiral-loop layout, trenches are generally 15.2 cm (6 inches) wide; the pipe loops stand vertically in the narrow trenches. In cases in which trenching is a large component of the overall installation costs, spiral-loop systems are a means of reducing the installation cost. As noted with horizontal systems, slinky systems are also generally associated with lower tonnage systems when land area requirements are not a limiting factor.

The key issue that makes GSHPs unique is the design of the GHE system. Most operational problems with GSHPs stem from the performance of the GHE system. There are several ways of sizing a GHE:

1. Using rules-of-thumb. In the residential sector, many systems have been designed using rules-of-thumb and local experience, but for commercial-scale systems, such practices are ill advised. For this purpose, heat extraction/rejection rates obtained from experience are used. For all but the most northern climates, commercial-scale buildings will have significantly more heat rejection than extraction. This imbalance in heat rejection/extraction can cause heat buildup in the ground to the point where heat pump performance is adversely affected and hence system efficiency and possibly occupant comfort suffer.
2. Short method 1. This method uses the building-block load, estimates of average equivalent full-load operating hours, average values of the equipment energy efficiency ratio (EER) and coefficient of performance (COP), and some values to estimate the temperature penalty due to long-term heat exchange in the GHE. Besides this, this computation also requires obtaining ground thermal properties and thermal resistances.
3. Short method 2. Besides the energy parameters for the heat pump (heating and cooling loads, and run time), the highest supply water temperature acceptable for cooling mode and the lowest temperature acceptable for heating mode must be known. These values, along with local ground temperature, establish the design temperatures for sizing a GHE. Using three simple tables and heat transfer equations, the required calculation may be easily done.
4. Using design software. The abbreviated computation described in short methods 1 and 2 is still very time consuming. More importantly, it is difficult to evaluate the impact of design alternatives, because most of the calculations have to be repeated. The procedure lends itself well to computer-aided design.

Proper design for commercial-scale systems almost always benefits from the use of design software. Today, software tools are available to support the design of any GHE system that meets the needs of designers and installers. Software for commercial-scale GSHP system design should consider the interaction of adjacent loops and predict the potential for long-term heat buildup in the soil. These tools are available from several sources, including the International Ground-Source Heat Pump Association (IGSHPA). In addition, several manufacturers have designed their own proprietary tools, and these are more closely tuned to particular system requirements.

5. SOIL CHARACTERISTICS

Soil characteristics are an important concern for any GHE design. The operation of the heat exchanger

induces a simultaneous heat and moisture flow in the surrounding soil. The transfer of heat between the GHE and adjoining soil is primarily by heat conduction and to a certain degree by moisture migration. Therefore, transfer depends strongly on the soil type, temperature, and moisture gradients. With horizontal GHEs, the soil type can be more easily determined because the excavated soil can be inspected and tested.

The soil thermal diffusivity (α_s) is a defined property and is the ratio of the thermal conductivity (k_s) and the heat capacity ($\rho_s C_s$). Therefore, the three soil properties, k_s, ρ_s, and C_s, must be known or estimated to predict the thermal behavior of GHEs. Obtaining accurate values for thermal properties of the soil requires a detailed site survey. In order to estimate the thermal properties of granular soils (sand, clay, silt) it is necessary to determine the sand and clay content, dry density, and moisture content. The following analytical equations are used to calculate the thermal properties of the soil based on soil dry density and moisture content.

The thermal conductivity of the soil, k, in watts/meter degree Kelvin, may be calculated from Eqs. (1) and (2):

$$k_s = 0.14423(0.9 \log \omega - 0.2)10^{0.000624\rho_{sd}} \quad (1)$$

for silt and clay soils

$$k_s = 0.14423(0.7 \log \omega + 0.4)10^{0.000624\rho_{sd}} \quad (2)$$

for sand soils where ω is moisture content in percentage by weight and ρ_{sd} is dry density in kilograms/cubic meter.

Thermal diffusivity is sensitive to moisture content because the specific heat of water (4.1868 kJ/kg K) is far greater compared to the typical values for soils and rock (0.84 to 1.05 kJ/kg K). Both the dry specific heat and the dry density of soil, C_{psd} and ρ_{sd}, must be corrected in order to determine α_s as follows:

$$\alpha_s = \frac{k_s}{C_{psc}\,\rho_{sc}} \quad (3)$$

with

$$C_{psc} = \left[\omega C_{pw} + (100 - \omega)C_{psd}\right]/100 \quad (4)$$

and

$$\rho_{sc} = \left[\omega \rho_{pw} + (100 - \omega)\rho_{sd}\right]/100 \quad (5)$$

where C_{psc} is the corrected specific heat of soil, ρ_{sc} is the corrected density of soil, C_{pw} is the specific heat of water (4.1868 kJ/kg K), and ρ_{pw} is the density of water (1000 kg/m^3).

6. GROUNDWATER HEAT PUMPS

Groundwater heat pump (open-loop) systems are the oldest and most well established of the GSHP systems, with the first examples of large commercial applications installed in the late 1940s. Common design variations include direct (groundwater used directly in the heat pump units), indirect (building loop isolated with a plate heat exchanger), and standing column (water produced and returned to the same well), as shown in Fig. 2A. Direct systems are typically limited to the smallest applications. Standing-column systems are employed in hard-rock geology sites where it is not possible to produce sufficient water for a conventional system.

Although seemingly simple in nature, these systems require careful consideration of well design, groundwater flow, heat exchanger selection, and disposal, even for very small commercial applications. The direct groundwater system is very susceptible to problems induced by poor water quality, the most common of which is scaling of the refrigerant-to-water heat exchangers. This design is recommended in only the smallest applications when practicality or economics precludes the use of an isolation heat exchanger and/or when groundwater quality is excellent (the determination of which requires extensive testing). The standing-column system has been installed in many locations in the northeast portion of the United States. Like the direct groundwater system, it, too, is subject to problems induced by poor water quality. In general, water quality in the area where most of the systems have been installed is extremely good, with low pH and low hardness (little scaling potential). Standing-column systems are used in locations underlain by hard-rock geology, where wells do not produce sufficient water for conventional open-loop systems and where water quality is excellent. Well depths are often in the range of 304.8 to 457.2 m (1000 to 1500 ft) and the systems operate at temperatures between those of open- and closed-loop systems.

In most commercial applications, the optimum design dictates a flow of 0.045–0.054 liters/sec per kW, or 2.5–3.0 gallons per minute (gpm)/ton, on the building-loop side of the exchanger and 0.018–0.045 liters/sec per kW (1–2.5 gpm/ton) on the groundwater side. As a result of this, the approach, or minimum, temperature difference between the two flows occurs at the building-loop return (heat pump "leaving" water) and groundwater leaving end of the exchanger. Groundwater below 60°F may be circulated directly through hydronic coils in series or in parallel with heat pumps.

There are two basic options for water disposal from an open-loop system: surface and injection. Both options are subject to regulatory oversight and permitting. Surface disposal, the most common method used in the past, is less expensive, but requires that the receiving body be capable of accepting the water over a long period. Injection is more complex and costly but offers the certainty that the groundwater aquifer will not be adversely affected (aquifer decline) by the operation of the system over the long term, because the water is "recycled." For surface disposal, it may be advisable to place a pressure-sustaining valve on the end of the system to maintain full piping when the pump is not operating. Some designers prefer simply to place a motorized valve at this point in the system and interlock it with the pump (through an end switch).

The optimum design of the groundwater heat pump (GWHP) is to realize optimal cooperation for all components, and to consider the running costs and first investment simultaneously under present technical and economic conditions.

7. SURFACE WATER HEAT PUMPS

Surface water heat pump (SWHP) systems can be either closed-loop systems similar to GCHPs or open-loop systems similar to GWHPs (Fig. 2B). However, the thermal characteristics of surface water bodies may be quite different from those of the ground or groundwater. Some unique applications are possible, but special precautions may be warranted.

Closed loops of pipe can be positioned at the bottom of a body of water at sites where the water is deep enough and has significant flow. Such sites utilize the circulating fluids to effect excellent heat exchange with the surrounding water, and limit the need for excavation. SWHP systems are a viable and relatively low-cost GSHP option. Lakes, streams, and bays can be very good heat sources and sinks for these systems, if properly utilized. Many successful systems are currently in operation.

Open systems are restricted for use in warmer climates or for buildings in colder climates with cooling-only applications (or very small heating loads). In colder climates, lake temperatures may range from 38° to 40°F. Pumping at a rate of 0.054 liters/sec per nominal kilowatt (3 gpm/nominal ton) will result in an outlet temperature 3.3°C (6°F) below the heat pump inlet. Even in warmer climates, caution is necessary to ensure that pumping rates are adequate and that the reservoir is large and deep enough to prevent the water temperature from falling below 5.6°C (42°F).

8. HYBRID GROUND-SOURCE HEAT PUMP SYSTEMS

Ground-source heat pump systems offer an attractive alternative for both residential and commercial heating and cooling applications due to their higher energy efficiency compared to conventional systems, but their higher first cost has been a significant drawback to wider acceptance of the technology. This is especially true in commercial and institutional applications, where the vertical GHE is commonly preferred. These types of buildings are generally cooling dominated and therefore reject more heat to the ground than they extract on an annual basis (Fig. 4). As a result, the required GHE length is significantly greater than the required length if the annual loads were balanced. One option to reduce the size of the GHE, and therefore the first cost of the system, is to balance the ground thermal loads by incorporating a supplemental heat rejecter into the system. GSHP systems that incorporate a supplemental heat rejecter have been referred to as hybrid GSHP systems.

Supplemental heat rejection can be accomplished with a cooling tower, fluid cooler, cooling pond, or pavement heating system. Currently suggested design methods for hybrid GSHP systems attempt to size the GHE based on the annual heating load and then size

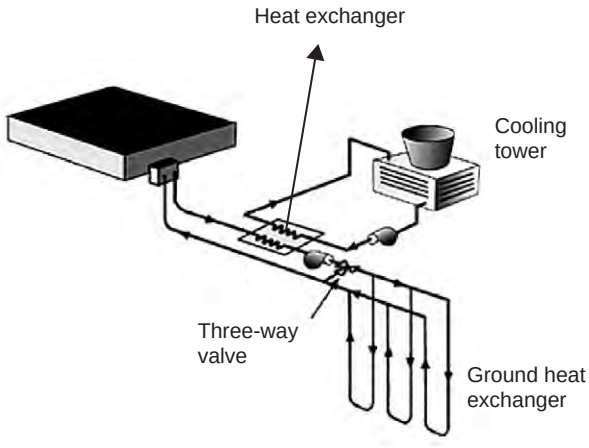


FIGURE 4 Cooling-tower supplemental for cooling-dominated loads.

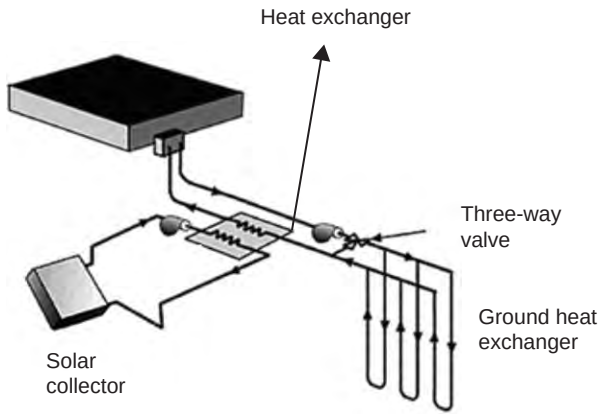


FIGURE 5 Solar-assisted system for heating-dominated loads.

the supplemental heat rejecter to balance the annual ground loads. However, the design of the system components also depends on the strategy used to control the supplemental heat rejecter. A smaller supplemental heat rejecter operated for more time may reject the same amount of heat as a larger supplemental heat rejecter operated for less time. Hence, a balance between the size of GHE, the size of the supplemental heat rejecter, and the control strategy is required to achieve the best economic alternative.

In northern climates, where the heating load is the driving design factor, supplementing the system with solar heat can reduce the required size of a closed-loop ground-coupling system. Solar panels, designed to heat water, can be installed into the ground-coupled loop (by means of a heat exchanger or directly), as illustrated in Fig. 5. The panels provide additional heat to the heat transfer fluid. This type of variation can reduce the required size of the ground-coupled system and increase heat pump efficiency by providing a higher temperature heat transfer fluid.

9. PERFORMANCE EVALUATION OF GROUND-SOURCE HEAT PUMPS

The energy performance of a GSHP system is influenced by three primary factors: the heat pump unit, the circulating pump or well pumps, and the GHE. Designers often devote more energy and resources to the GHE because it is a novel component to most heating, ventilating, and air-conditioning (HVAC) engineers. When the GHEs are correctly designed and installed and high-efficiency extended range heat pump units are specified, GSHPs are extremely efficient. Thus, the high demand and energy use of pumps that may result from unnecessary oversizing, restrictive piping, and poor control are especially noticeable.

9.1 Heat Pump Unit

The heat pump is the largest single energy consumer in the system. Its performance is a function of the efficiency of the machine and the water temperature produced by the GHE or well (either in the heating or cooling mode). The most important strategy is to start with an efficient heat pump. It is difficult and expensive to enlarge a GHE to improve the performance of an inefficient heat pump.

The purpose of rating the efficiency of a heat pump unit is to indicate the relative amount of energy needed to provide a specific heating/cooling output. The more efficient the equipment, the less energy will be used to do the same job. Many of the terms used in the heating, air-conditioning, and refrigeration industry convey performance and efficiency. Among these, four terms, the coefficient of performance, the energy efficiency ratio, the seasonal energy efficiency ratio (SEER), and the heating seasonal performance factor (HSPF), are widely used for these purposes in the United States and are currently rated by the American Refrigeration Institute (ARI). Although codes and standards identify minimum efficiencies, they do not fully communicate the energy efficiency that is achieved by today's heat pumps. A review of manufacturers' literature on commercially available equipment indicates that cooling efficiencies (EERs) of 12.0 to 16.8 Btu/Wh and heating efficiencies (COPs) of 3.0 to 4.3 are readily available (to convert the EER rating to the COP, divide the EER rating by 3.413). In the United States, heat pumps are rated on tonnage (i.e., 1 ton of cooling power produced by 1 ton of ice is equal to 12,000 Btu/hour, or about 3.51 kW).

When comparing equipment efficiencies, it is important to make an appropriate comparison. The efficiency of any heating and cooling equipment varies with application, load, and related heat-source and heat-sink temperatures. Furthermore, standard ratings are for specific conditions and operating conditions. It is highly recommended that the ratings systems be used only to compare one type of equipment to another. The performance of equipment rated in one standard should not be compared with the performance of units rated in another standard. The ratings of water-to-air heat pumps

should not be compared with equipment rated by the air-source heat pump standard, which is much more complex and optimistic.

The actual performance of the equipment is a function of the water temperature produced by the GHE, which is a function of the ground temperature, the pumping energy, and the design of the GHE. In other words, the entering water temperature (EWT) of the heat pump unit is perhaps the single most representative parameter of the GSH effectiveness and heat pump loading. For example, in a region where the local ground temperature is 16°C and the GHE is designed for the customary 11–14°C above ground temperature, a heat pump rated at an EER of 16.8 would actually operate at an EER of 14.2 under peak load conditions. A poorly designed GHE that forces the unit to operate at 17°C above ground temperature would reduce the value to less than 13.0. These are for the cooling operation, which is the dominant load in commercial applications. However, the same relationship holds for heating operations. Besides this, systems with inadequate GHE capacity will have excessively high EWTs in summer and excessively low EWTs in winter.

9.2 Circulating Pump

The energy performance of the system is also influenced by the pumping energy required to circulate the fluid through the heat pump and the GHP. In the design of GSHP systems, the benchmarks given in Table II are proposed for pumping power for commercial GCHP systems. These benchmarks are for judging the effectiveness of a pumping and piping system design for a minimum of 0.162 m^3/kWh of cooling, with optimum pumping flow rates ranging from 0.162 to 0.192 m^3/kWh of cooling. Higher rates require excessive pump power, and lower rates limit the heat pump. In addition, pumping energy should range between 6 and 7% of the total system energy used.

TABLE II

Benchmarks for GSHP System Pumping Efficiency[a]

Watts (input)		Performance	
Per ton	Per kilowatt	Efficiency	Grade
≤50	≤14	Efficient systems	A: Excellent
50–75	14–21	Acceptable systems	B: Good
75–100	21–28	Acceptable systems	C: Mediate
100–150	28–42	Inefficient systems	D: Poor
>150	>42	Inefficient systems	E: Bad

[a] Required pump power to achieve cooling capacity.

Source. Copyright 2003, American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc. Reprinted by permission from Kavanaugh *et al.* (1997).

To put the values given in Table II into perspective, consider an office building with a 175-kW cooling load and heat pump units selected to operate at an EER of 14 under peak conditions. With an efficient circulating pump design (10 W/kW), the energy demand of the circulating pump would amount to 175 kW × 10 W/kW = 1750 W. Combining the heat pump demand with the heat pump unit demand results in a system EER of 13.5. The same building and equipment coupled to a poorly designed pumping system consuming 34 W/kW would yield a system EER of only 12.2, thus compromising the premium paid for the higher efficiency equipment.

9.3 Ground Heat Exchanger

The design of GHEs is complicated by the variety of geological formations and properties that affect thermal performance. Proper identification of materials, moisture content, and water movement is important. Careful consideration of the interaction between these three components and their impact on system performance is necessary in order to minimize operating costs for the building owner.

The key parameter for GHE layout is the specific performance, i.e., the heat extraction rate during the heating season or the heat rejection rate during the cooling season in watts per meter of borehole (or meter of borehole per kilowatt) for vertical GHEs (or length of buried pipe loop for horizontal GHEs). The heat rejection rate ranges from 40 to 100 W/m, with a typical average of 55–70 W/m in mid-European countries. Based on a study that covered the 22 GSHP systems in the United States, for closed-loop systems, the GHE pipe length ranged from 20.5 to 52.0 m/kW_t, with an average of 39.3 m/kW_t. Of those with vertical bores, the range is 14.4–17.7 m of bore per kW_t.

The impact soil type and moisture content on the performance of GSHPs has recently been reported by some investigators. The length of a GHE can be reduced by using sand or thermally improved bentonite grout instead of standard bentonite grout. It can be concluded that using thermally enhanced materials instead of standard ones decreases the required GHE length by 15–20%. For applications in combination with other heat sources, the specific performance may differ considerably. Therefore, it is

essential to obtain a second parameter, i.e., the specific heat extraction per meter of borehole length for a year. This value ranges between 50 and 200 kWh/m per year for mid-European facilities.

GHEs and associated components have significantly different characteristics compared to other types of HVAC heat exchangers; these differences impact the selection of flow rates, allowable head losses, and piping materials. GSHP piping design may be affected by the following special factors:

1. The ground provides by far the greatest resistance to heat flow; thus, highly conductive pipe materials, extended surfaces, and high internal fluid velocities are of little value.
2. Because the fluid must travel through both the building and GHE, rust inhibitor solutions must be acceptable for in-ground use. Thus, a material that does not require toxic inhibitors is necessary.
3. When design guides are followed, high fluid velocities are unnecessary for good heat transfer in the GHE and head losses through the ground coils can be relatively small [<6 ft (2 m) of water].
4. Because high fluid velocities are unnecessary for good heat transfer in the GHE, moderate flow imbalances will have little impact on overall heat exchange.
5. High-efficiency, extended-range water-to-air heat pumps do not require precise water flow control (or high head loss) to operate near maximum efficiency. A $\pm 33\%$ change in water flow rate results in less than a $\pm 2\%$ change in cooling capacity.
6. Because the water coils of extended-range, high-efficiency heat pumps are much larger than water-loop heat pump coils, head loss is relatively small [<12 ft (4 m) of water].
7. Because the combined head losses in the ground coil and heat pump are small [<18 ft (5.5 m) of water], the lengths of the headers, control valve losses, and fitting restrictions have a significant impact on overall system head loss.
8. Because the cost of the recommended piping material (thermally fused high-density polyethylene) is low relative to labor costs, low friction losses can be economically justified as a method of minimizing pump head.
9. Corrosion- and leak-resistant piping material with a minimal number of in-ground joints is critical to high system reliability and life.

SEE ALSO THE FOLLOWING ARTICLES

Geothermal Direct Use • Geothermal Power Generation • Ocean Thermal Energy • Thermal Energy Storage

Further Reading

Bose, J. E., Parker, J. D., and McQuiston, F. C. (1985). "Design/Data Manual for Closed-Loop Ground-Coupled Heat Pump Systems." American Society of Heating, Refrigerating, and Air-Conditioning Engineers, Inc. Atlanta.

Federal Technology Alerts (FTA). (2001). "Ground-Source Heat Pumps Applied to Federal Facilities," 2nd Ed. Available on the Internet at http://www.pnl.gov.

Healy, P. F., and Ugursal, I. (1997). Performance and economic feasibility of ground source heat pumps in cold climate. *Int. J. Energy Res.* **21,** 857–870.

Hepbasli, A. (2002). Performance evaluation of a vertical ground-source heat pump system. *Int. J. Energy Res.* **26,** 1121–1139.

Huttrer, G. W. (1997). Geothermal heat pumps: An increasingly successful technology. *Renew. Energy* **10**(2/3), 481–488.

Kavanaugh, S. (1991). "Ground and Water Source Heat Pumps: A Manual for the Design and Installation of Ground Coupled, Ground Water and Lake Water Heating and Cooling Systems in Southern Climates." Energy Information Services, Tuscaloosa, Alabama.

Kavanaugh, S. P., and Rafferty, K. (1997). "Ground-Source Heat Pumps: Design of Geothermal Systems for Commercial and Institutional Buildings." American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc., Atlanta.

Lund, J. W. (2000). Ground-source (geothermal) heat pumps. *In* "Course on Heating with Geothermal Energy: Conventional and New Schemes Course textbook of World Geothermal Congress 2000" (P. J. Lienau, Ed.), pp. 209–236. Kazuno, Tohoku District, Japan.

Lund, J. W. (ed.). (2001). Geothermal heat pumps—An overview. *Q. Bull.* **22**(1), 1–2.

Lund, J. W. (2003). Direct-use of geothermal energy in the USA. *Appl. Energy* **74,** 33–42.

Rafferty, K. (1999). Design issues in the commercial application of GSHP systems in the U.S. *In* "Proceedings of International Summer School" (K. Popovski, J. Lund, D. Gibson, and T. Boyd, Eds.), pp. 153–157. Geo-Heat Center, Oregon Institute of Technology, Klamath Falls, Oregon.

Gulf War, Environmental Impact of

FAROUK EL-BAZ
Boston University
Boston, Massachusetts, United States

1. **Definition of War Impacts**
2. **General Setting**
3. **Preinvasion Conditions**
4. **Postwar Effects**
5. **Remediation Approaches**
6. **Future Monitoring**

Glossary

desert pavement Natural residual concentration of wind-segregated and closely packed pebbles and other rock fragments mantling a desert surface.

Gulf War The international military campaign of "coalition forces" led by the United States to liberate Kuwait, during January–November 1991, from the occupation by Iraq's forces that began in August 1990. "Gulf" refers to the Arabian Gulf, which is also known as the Persian Gulf.

Landsat A series of U.S. satellites designed to obtain and transmit images of the surface of the earth up to 15 m in resolution (detail).

sabkha Flat, low area covered by clay, silt, and sand and often encrusted with salt; it forms in restricted coastal areas of arid and semiarid environments.

sand dune An accumulation of loose sand heaped up by the wind, where there is an abundant source of sand and constant wind.

sand sheet A large plain of wind-blown sand lacking the discernible sides or faces that distinguish sand dunes.

Systeme pour L'Observation de la Terre (SPOT) A French satellite that obtains and transmits images of the earth up to 5 m in resolution (detail).

tarcrete A hardened surface layer composed of surface sand, gravel, and other rock fragments cemented together by petroleum droplets and soot from the plumes of oil well fires (in Kuwait).

Thematic Mapper An advanced imaging system of Landsat 4, 5, and 7 that collects surface radiation in multispectral bands up to 30 m in resolution (detail) and panchromatic (black-and-white) bands of 15-m resolution.

The Gulf War of 1990–1991 resulted in severe environmental effects, which were analyzed using satellite images followed by field observations. Major disturbances were caused by the disruption of the "desert pavement," a layer of pebbles and gravel that protects the fine-grained soil beneath. The disruption was caused by the activities of military personnel including the digging of trenches and pits, building of berms and sand walls, and the traffic of heavy military vehicles. The exposure of fine-grained soil caused the increase of dust storms and the formation of numerous sand dunes that encroached on roads, installations, and farms. In addition, the explosion of over 600 oil wells caused other major environmental hazards. These included the formation of 240 oil lakes in topographically low areas. Also, plumes from the well fires deposited oil droplets and soot, which hardened on the desert surface as a layer of "tarcrete," affecting desert animals and natural vegetation. Nearly one-third of the land surface of Kuwait was impacted by the war. Although some remediation was performed, long-term monitoring is required to fully understand and ameliorate the impacts on the desert environment.

1. DEFINITION OF WAR IMPACTS

Iraq invaded Kuwait on August 2, 1990. On January 17, 1991, international coalition forces led by the United States began a military campaign for the liberation of Kuwait that became known as the Gulf War. These events resulted in numerous environmental impacts on the desert surface of Kuwait. The impacts were identified based on the study and

 © 2004 Elsevier Inc. All rights reserved.

analysis of satellite images. The most significant of these impacts were the results of exploding the oil wells, most of which caught fire. Plumes of smoke and oil droplets were carried by the southward wind and deposited on the desert surface. The resulting layer of hardened tarcrete disrupted natural vegetation. Some of the exploded oil wells did not catch fire and the oil seeped to form vast lakes. The oil-soaked lake beds continue to be an environmental hazard more than a decade later.

In addition to oilfield-related environmental impacts, the desert surface was disrupted by military troop actions. Disruption of the surface layer occurred due to the passage of heavy military vehicles, digging of trenches, building of berms, etc. All these activities exposed fine-grained material from below the protective layer of pebbles known as desert pavement. Exposure of the fine-grained soil to wind increased dust storms as well as the number and size of sand dunes.

An understanding of the nature of the desert surface in Kuwait is necessary to appreciate these environmental impacts. For this reason, a description of the general setting of Kuwait and the characteristics of its surface is given, followed by an assessment of the prewar conditions based on comparisons of satellite images obtained before and after the Gulf War. The major impacts are described with this background in mind. Recommendations are given for monitoring the changes and remediation of the environmental problems.

2. GENERAL SETTING

Kuwait is an arid country that lies along the northwestern corner of the Arabian (Persian) Gulf, bordering Saudi Arabia to the south, Iraq to the north and west, and the Gulf to the east (Fig. 1). The surface topography of its 17,818 km is marked by flat to undulating gravel-covered plains that slope gently eastward from a maximum elevation of 284 m (above mean sea level) in the west.

During the past three decades, rainfall ranged from less than 25 mm in 1964 to 375 mm in 1972, with an average of 115 mm per year. The precipitation infiltrates rapidly, leaving no permanent surface water, although some wadis fill temporarily with water from winter rains. The largest of these is Wadi Al-Batin, which is 5 km in width at the western border

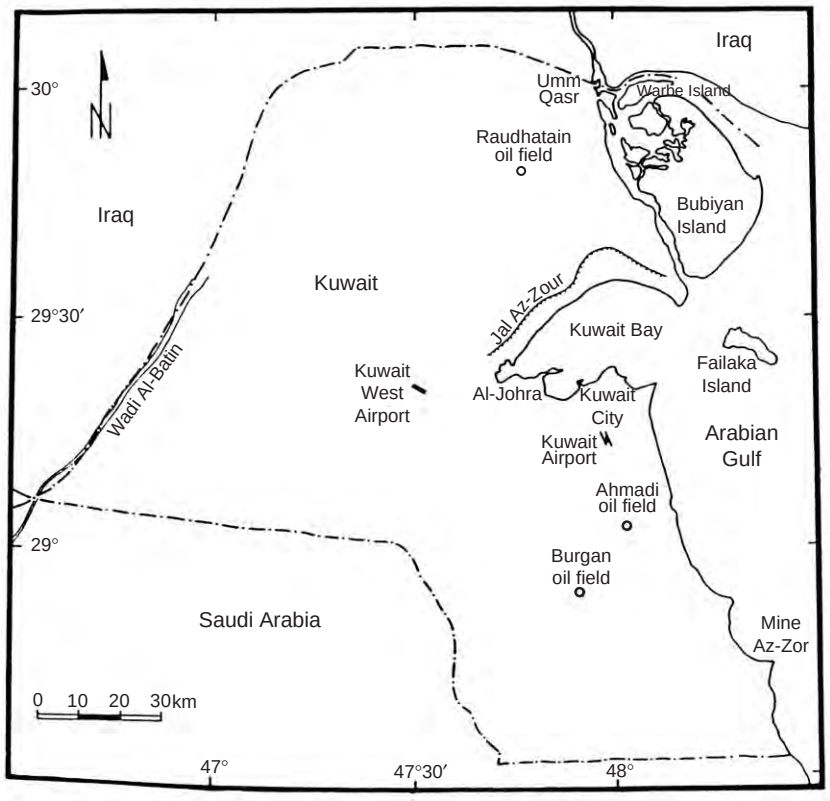


FIGURE 1 Sketch map of Kuwait showing its major features.

of the country. This wadi is part of a partially buried, water-dug channel whose tributaries drain part of the Hijaz Mountains 850 km to the southwest in Saudi Arabia. It was proposed to name the complete channel the Arabia River, whose delta deposits form much of the land surface of Kuwait (Fig. 2).

As mapped by the U.S. Geological Survey, Kuwait consists of flat-lying Tertiary rocks overlying gently folded Cretaceous and Jurassic formations. Exposed rock types include the Eocene Dammam formation, a white, fine-grained cherty limestone that shows some karst development at its contact with the overlying Miocene and Pliocene age Fars and Ghar formations, which are primarily calcareous sandstones, sandy limestones, clay, and sand. Completing the section above the Fars and Ghar formations is the Dibdiba formation, which is composed of sandstone. Quaternary deposits include lag gravels that blanket much of the interior of Kuwait and coastal deposits including unconsolidated marine sands, mudflats, and supratidal sabkha surfaces.

Structurally, Kuwait lies on the Arabian Shield, an area noted for its stability since Cambrian times. The Shield tilts slightly to the northeast, giving rise to sedimentation of the Arabian Shelf consisting of a sequence of laterally extensive but thin limestones, marls, shales, sandstones, and evaporites. North–south trending anticlinal structures, developed with gentle warping and uplift, serve as traps for Kuwait's oil reserves. In addition to this warping, there is evidence of salt diapirs with magnetic low signatures on the anticlinal structures.

Topographically, Kuwait is a flat sandy desert that may be divided into two parts. The north is a vegetated stone desert with shallow depressions and low hills running northeast to southwest and ending near an area of underground water storage. The principal hills in the north are Jal Az-Zor (145 m) and the Liyah Ridge. The southern region is a bare plain covered by sand; the Ahmadi Hill (125 m) is the sole exception to the flat terrain. In addition to Wadi Al-Batin, the only other valley of note is Ash-Shaqq, a portion of which lies within the southern part of the country and the rest in Saudi Arabia.

Morphologically, east of Wadi Al-Batin (along the borders between Kuwait and Saudi Arabia; Fig. 1) the desert surface is composed of a calcretic plain. This unit is composed of rock debris surrounding hills and gravel ridges. Often, parts of the unit are covered with a crust of saline or alkaline soils. The largest morphological unit of Kuwait lies east of the calcrete unit and northeast of Wadi Al-Batin. In this unit, the

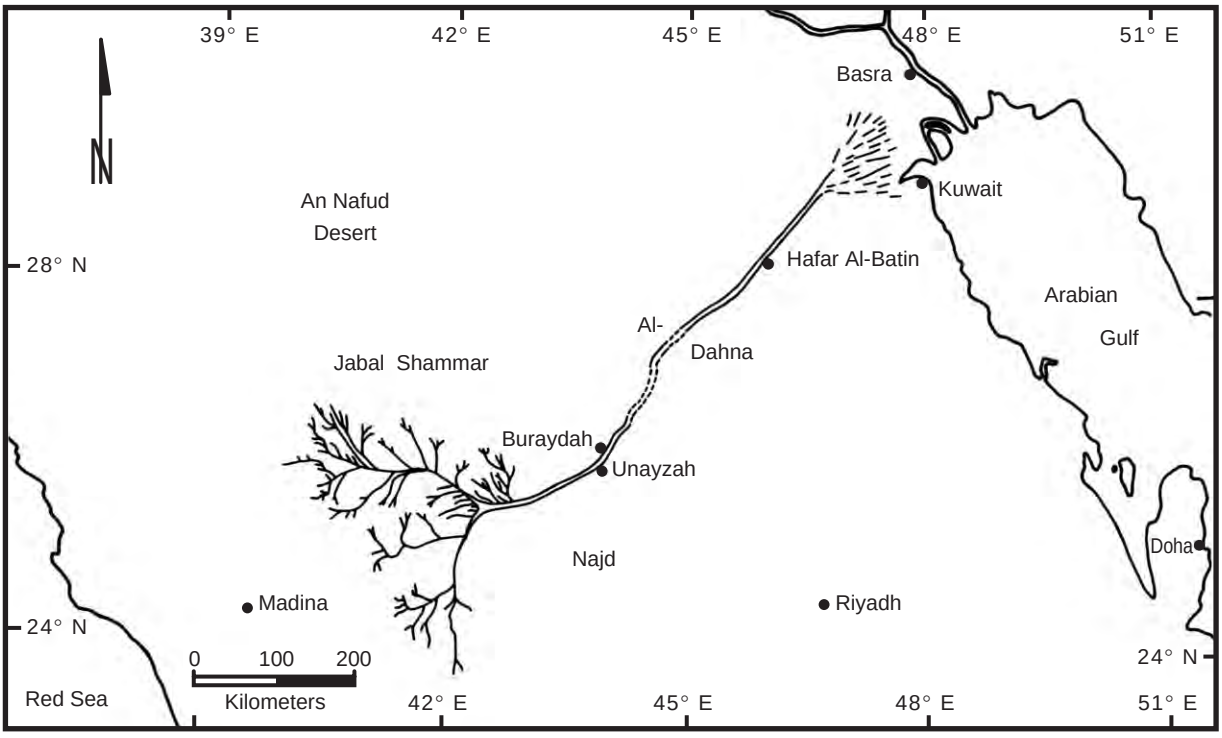


FIGURE 2 Schematic illustration of the tributaries and partly sand-buried channel of the "Arabian River." Note that western Kuwait is part of the ancient delta of the river.

terrain is rugged and composed mainly of a vegetated sand sheet, locally referred to as the Al-Huwaimliyah plain. The ruggedness is caused by the accumulation of sand in the lee of the wind behind desert shrubs. Spaces between the sand piles are composed of a thick bed of wind-blown sand. Therefore, destruction of the vegetation by extended droughts, overgrazing, or vehicular overuse results in the mobilization of sand in the otherwise stable sand sheet.

Northern Kuwait is covered by a vast plain with particles ranging in size from 4 to 10 mm in diameter. This unit is considered a desert pavement of lag gravels, which are fragments of rock too large to be moved by the wind and therefore "lag" behind the finer particles that move with the wind. Their concentration varies from scattered pebbles to densely packed accumulations of pebbles and granules. The predominance of the unit in northern Kuwait is largely attributed, in the literature, to the occurrence of the gravelly Dibdiba formation, which outcrops in northern Kuwait and southern Iraq. However, I propose that the gravel originated in western Saudi Arabia and was transported through the channel of the Arabia River through Wadi Al-Batin (Fig. 2). This is largely based on an apparent provenance of the gravel; rocks from which the gravel may have originated abound in the Higaz Mountains of western Arabia (Fig. 2).

From the characteristics of the desert pavement of Kuwait, it is clear that gravels were first moved by water currents and deposited in a sheet together with finer fractions, which were later blown away by deflation. As a result of the continuous deflation, pebbles and granules were concentrated on the surface until a deflation armor or desert pavement was acquired. The gravel ridges were formed by reversed topography as a result of the preferential deflation of flood plains dissected by fluvial channels in the fan-shaped delta of Wadi Al-Batin.

The coastal zone of Kuwait is composed mainly of sabkha deposits. These are flat areas of clay, silt, and sand that are often encrusted with salt. Along the coasts of Warba and Bubiyan islands, the supratidal flats are covered by thin laminae of sandy mud encrusted by a thin lithified crust. Large areas of sabkha are composed of extensive sheets of mud with almost no relief and flourishing mangroves.

3. PREINVASION CONDITIONS

During the two decades before 1990, developing land for agriculture had increased dramatically in Kuwait. This increase in agricultural activity resulted in the disruption of the desert pavement in developed areas and the exposure of soil to wind action. This, in turn, increased the amount of dust in the air and the movement of sand and its accumulation in mobile dunes. In addition, increases in dust and sand storms in the northern Gulf region were noticed in the 1970s due to major agricultural expansion in the Western Desert of Iraq. Satellite images clearly show that dust storms in southwestern Iraq spread over the eastern part of the Arabian Peninsula.

Similarly, growth of urban communities results in the disturbance of the stable desert surface. Most of the environmental effects of urbanization occur during digging, when soils are exposed to eolian erosion. Such efforts were particularly obvious during the expansion of Al-Jahrah, a community west of Kuwait City (Fig. 1).

Farther inland, disruption of the desert surface from the oil industry resulted from exploration and development activities, including geophysical surveys, drilling operations, building of roads, and laying of pipelines. These activities usually require a large number of vehicles, which criss-cross the desert surface affecting both the desert pavement and the scant desert vegetation.

For example, the aforementioned morphological unit of Al-Huwaimliyah plain, the vegetated sand sheet, extended into the Burgan oil field in south central Kuwait (Fig. 1). This area was badly degraded due to excessive vehicular activities and the desert surface was disturbed. Because disturbances of this type mobilize the sand, vehicular traffic in the open desert in Burgan field had been prohibited during the 6 years prior to the invasion of 1990. Due to this prohibition, vegetation began to flourish and the sand sheet was once again stabilized.

Military installations in the open desert also had some impact on the environment of Kuwait's desert prior to the Gulf War. Military bases have greatly suffered from the effects of destabilizing the desert surface in terms of having to constantly alleviate the effects of sand accumulations around buildings and their entrances. The study of sand deposits within Ali Al-Salem and Ahmad Al-Jabir bases provided insights into accumulations on roads. Three cases of pavement profiles on a slightly raised bed were considered: a flat pavement, one in which the underlying bed was slightly elevated on the upwind side of the road, and another with the elevation on the downwind side of the road (Fig. 3).

In all three cases, sand accumulated on either side of the raised roadbed; regardless of its slight elevation, a roadbed represents a topographic

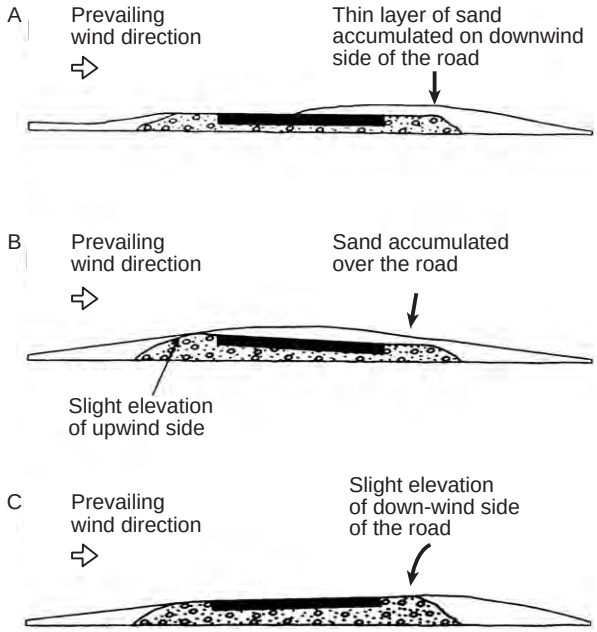


FIGURE 3 Deposition of sand along a road with a horizontal profile (A), an upwind elevation (B), and a downwind elevation (C). Modified from Fikry Khalaf.

impediment to the wind. In the case of the flat roadbed and pavement, sand accumulated on the downwind side of the road. The road with a slight elevation upwind was completely covered by sand because vorticity of the wind as it was confronted by the raised roadbed caused deposition of its sand load. In the case of the slight elevation of the roadbed and pavement on the downwind side, the road remained free of sand. In this case, the wind energy was increased as it moved upslope and carried the sand with it across the road to be deposited beyond the downwind elevation of the roadbed.

In the future, similar precautions and protective measures will have to be taken to safeguard installations in the open desert, such as roads, airport runways, helicopter pads, oil wells and storage tanks, checkpoints, and residential sites in the oil fields of Kuwait. Each site will have to be considered separately because remedies may differ due to local topographic characteristics, the nature of the desert surface, and the potential of particle transport by the wind.

4. POSTWAR EFFECTS

The physical effects of the Gulf War on the desert surface of Kuwait are numerous. Satellite images were used to identify these effects and to quantify their extent. NASA's Landsat Thematic Mapper images with a 30-m resolution were used in conjunction with Systeme pour L'Observation de la Terre (SPOT) images at 10-m resolution.

The variety and extent of environmental damage due to the Gulf War were characterized mainly through "change detection." In this process, an image of a specific area is compared to another image of the same area taken at a different time; the change from one image to another is computed. In this case, four Landsat images of Kuwait obtained in 1989 (prior to the invasion) were compared to four taken in 1992, upon cessation of the last oil well fire.

Most people are familiar with the horrific images of burning oil wells, which resulted from the wanton explosion of Kuwait's 700 oil wells by the Iraqi troops prior to their withdrawal. Naturally, this tragedy resulted in severe environmental effects, particularly spewing smoke and oil particles in the atmosphere for nearly 1 year. There were numerous other environmental effects, which are discussed in the following sections.

4.1 Vehicle Tracks

The movement of a great number of tanks and mechanized vehicles on the natural surface of Kuwait adversely impacted the desert pavement, particularly in the western part of the country. During the invasion, columns of Iraqi military equipment moved along parallel lines both inside and just west of the border of Kuwait. Such heavy equipment forces the surface cover of pebbles and cobbles deep into the sand and excavates large amounts of soil on either side of the tracks.

The net result of vehicle movement on a naturally packed surface is the decrease in the closeness of packing of sand grains. This phenomenon increases the potential of grain movement by the wind, resulting in the mobilization of previously protected fine particles. The spread of newly exposed soil on either side of the tracks allows for their detection on satellite images, even if the resolution of the images is less than the width of the tracks (Fig. 4).

During occupation, most of the troop movements were along paved roads. However, some areas were trampled by military vehicles, causing damage to the desert pavement. This is particularly true in the southwestern and northwestern parts of the country as Iraqi soldiers raised defensive fortifications. In areas covered by a sand sheet, such as in southeastern Kuwait, the tracks of military vehicles have already been covered by drifting sand. On the other hand, areas covered by a desert pavement remained deeply

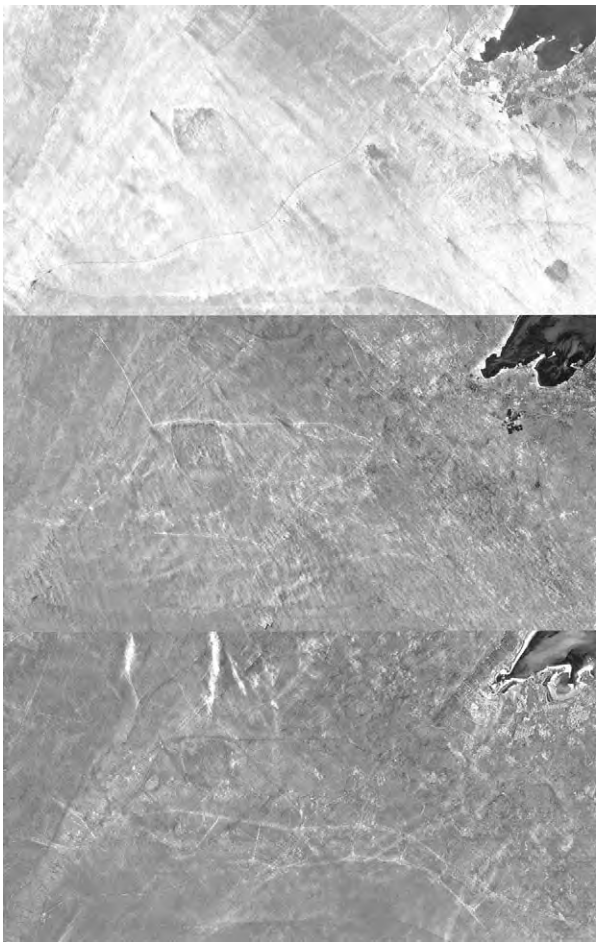


FIGURE 4 Vehicle tracks on the desert surface as depicted on three successive Landsat images of Kuwait's southwestern corner, just east of Wadi Al-Batin. Courtesy of Earthsat Corporation.

scarred, such as north of the Jal Az-Zor escarpment (Fig. 1).

4.2 Trenches and Pits

Iraq's armored forces spent much of their time during the fall of 1990 digging trenches in the desert of Kuwait (Fig. 5). This is particularly true along Kuwait's coastline. Beyond the trenches, pits were dug as hiding places for military equipment and storage areas of ammunition.

After the passage of a windy (shamal) season, most pits and trenches had been filled by drifting sand. The ammunition pits that were covered by sand represent a potential danger in the future. These may be identified by the use of ground-penetrating radar so they can be excavated and destroyed. Such procedures would again mobilize the sand for downwind movement if the dug-up areas are not stabilized, for example, by spreading pebbles on top of the exposed soil.

FIGURE 5 A trench dug in the desert surface north of Jal Az-Zor that was partly filled by sand less than 1 year from the end of the Gulf War.

4.3 Berms and Sand Walls

Berms were built in Kuwait's desert to slow the potential advance of liberation forces. The major berms were constructed in the south along the Saudi border, in the central part (south of the Burgam oil field), and in the northern part of Kuwait (north and west of the Jal Az-Zor escarpment). The berm-building activity continued throughout the occupation phase. Iraqi military strategists knew that the sand walls would be eroded by the consistent wind. Therefore, berms were sprayed with crude oil by pipelines that were laid specifically for this purpose from Kuwait's nearby oil fields. These berms represent a source of shifting sand in the long term. Currently, the soil in these berms is stabilized by the hardened oil crust (Fig. 6). However, this crust will gradually be dismantled by wind erosion and the fine-grained sand will be released regularly.

The effects of such berms on the spread of sand on the desert surface are exemplified by the case of the sand wall in Saudi Arabia just south of its border with Kuwait, which was constructed to stop potential smugglers, mostly of liquor. The berm

is less than 3 m high and a few meters wide at the base. However, its sand was mobilized and spread over such a large area (up to 1 km) so as to make it clearly visible on satellite images of 30-m resolution (Fig. 7).

In addition to the berms built by Iraqi military personnel, an extensive berm was constructed during the summer of 1992 by Kuwait's security forces in the northern part of the country. This berm was intended to stop a wave of smugglers and political agitators from Iraq following the cessations of hostilities. Since wind erodes the fine-grained material of the berm, it is constantly maintained by bulldozers. This berm is a source of drifting sand and must be monitored.

FIGURE 6 An oil-covered berm in northern Kuwait built by Iraqi forces. Crude oil was used to stabilize the sand.

4.4 Mines and Ordnance Detonation

The Iraqi military forces planted between 500,000 and 2.5 million land mines in the desert and coastal areas of Kuwait. The mining was done during the occupation phase to hamper the expected liberation of Kuwait by military action. Numerous defense lines were established particularly along the border of Saudi Arabia and the coastline of the Arabian Gulf. In addition, mines were laid in areas of high troop concentrations, particularly in the triborder area near the western corner of Kuwait.

The laying of land mines is a process that results in the disturbance of stabilized soil and exposure of fine-grained material to the action of wind. This is particularly obvious in a minefield along the northern coast of Kuwait Bay (Fig. 1). In this case, the mines

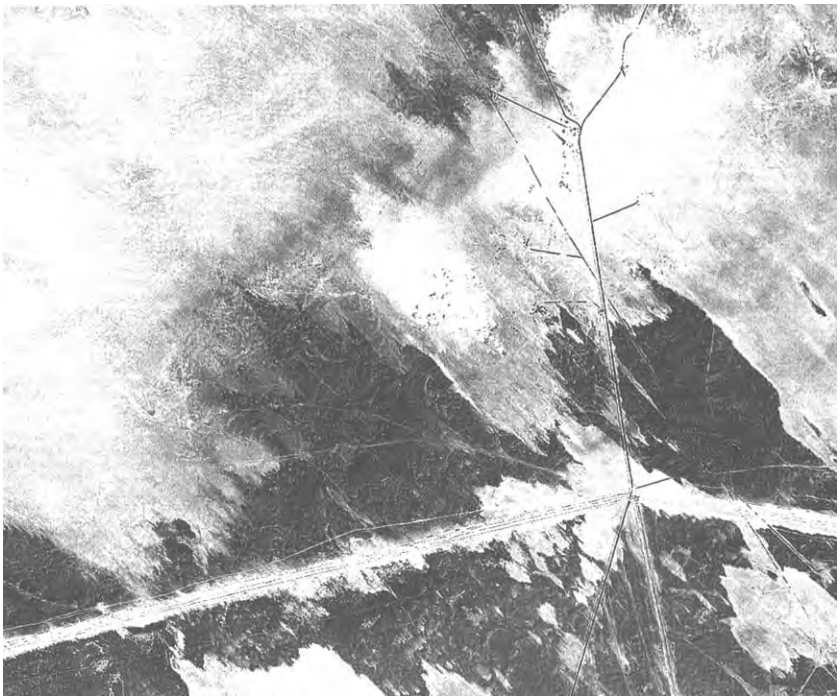


FIGURE 7 Effects of wind erosion on a berm in Saudi Arabia just south of its border with Kuwait. The berm is on the sharp northern edge of the white zone; the wind blowing to the southeast caused the jagged southern boundary due to sand transport downwind. This is part of a SPOT image acquired on May 25, 1992.

were laid in a soil that was stabilized by salt crust. The process of digging the land to plant the mines destroyed the drier crust, resulting in the deflation of the soil layer and exposure of the mines (Fig. 8).

During liberation, several passages were made through the minefield along the southern border of Kuwait. Breakthroughs were made by exploding the mines in groups. This process also disturbed additional stretches of desert pavement. In addition, bombing of Iraqi defensive positions by coalition forces resulted in cratering of the desert surface where bombs missed their targets. Ejecta from the craters exposed much fine-grained soil to the wind. However, because of the accurate targeting these craters are few and far between.

Much of the disruption of the desert surface in the minefields occurred during the postliberation phase as a result of mine detection and detonation. Crews of mine detectors were followed by mine sweepers and/or detonation equipment. The raking of the land and the explosion of mines and undetonated cluster bombs and other ordnance further disturbed the desert surface by exposing fine-grained soil to the wind.

4.5 Oil Crust

As shown in Fig. 9, deposits of oil and soot particles from the plumes of oil well fires were clearly visible in satellite images. When observed in the field, this deposit formed a hardened crust on top of the desert surface. It varied in thickness, with an average of 6 cm; the closer to the center of the deposit, the thicker the crust. Minor variations in the exact orientation of shamal winds caused the spread of the oil rain downwind of the well fires. Naturally, the thicker the deposit, the darker the surface. Minor variations in the spectral reflectance of such deposits were easily depicted on Landsat Thematic Mapper images, and it was possible to correlate the reflectance values to the thickness of the oil crust. Because it represented a new rock formation, this deposit was named tarcrete and defined as a conglomerate consisting of surface sand and gravel cemented together into a hard mass by petroleum droplets and soot.

The total surface area of Kuwait covered by tarcrete was approximately 943 km^2. Particulates in this area had been basically stabilized. Petroleum sprays had been used in the past to stabilize dune sands. Thus, from the standpoint of sand mobility, the tarcrete had a positive impact.

The effects of the tarcrete on plant life must be considered separately. Some plants died from the oil spray. Other species flourished but were not palatable to grazing animals. Furthermore, the salinity of the soil in areas covered by the tarcrete may change with time as sulfur compounds react with moisture to increase soil acidity. For this reason, monitoring of the pH values and other soil characteristics is recommended.

FIGURE 8 Mines exposed by wind erosion of sand along the coast of Kuwait. The minefield was supposed to prevent a marine landing of coalition forces (photograph by Jassim Al-Hassan).

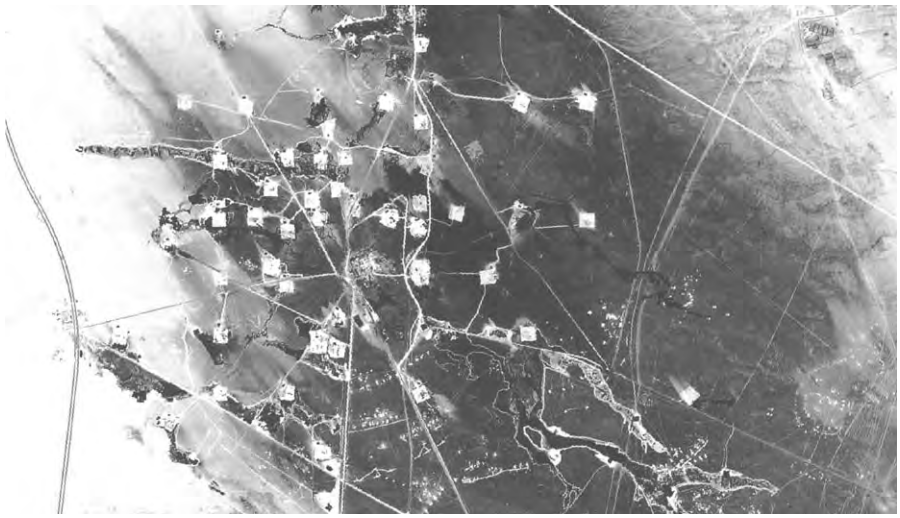


FIGURE 9 Part of a SPOT image acquired on March 25, 1992, of the Raudhatain oil field showing the new dirt roads (white lines) and sand pits around well heads (white squares) that were created during the firefighting operations. Note the oil lakes, which are oriented roughly east–west, within and along the western edge of the darkened surface.

4.6 Oil Lakes

Some of the exploded oil wells did not catch fire, and the seeped oil accumulated in low areas forming lakes (Fig. 9). Smaller oil lakes were also created during the firefighting phase. These lakes were identified using satellite images. The production of accurate maps of oil lakes necessitated the study of Kuwait by both Landsat Thematic Mapper and SPOT as well as maps that were based on field observations. Change detection techniques were used to study the characteristics of oil lakes in Thematic Mapper images. Also, geographic information system correlations of all data sets allowed their accuracy to be checked and the mapped lake boundaries to be confirmed. It was concluded that after a full understanding of the spectral characteristics of the oil lakes in Thematic Mapper images, their accurate mapping could be improved by using the higher resolution SPOT data.

Disregarding patches smaller than $1\,km^2$, the surface area covered by oil lakes in the desert of Kuwait was $35.4\,km^2$ in the fall of 1991, as mapped by the Kuwait Institute for Scientific Research. The total area of oil lakes was $27.5\,km^2$ as mapped from SPOT images obtained on March 25, 1992. The difference between the two measurements is easily ascribed to the shrinkage of the lakes due to the evaporation of aromatic compounds and the seepage of some of the oil into the underlying soil during the intervening months. The larger figure must be used in the estimation of remediation needs because all oil-saturated soil must be treated or removed, replaced, and stabilized.

5. REMEDIATION APPROACHES

It is clear from the previous discussion that the Gulf War has had severe short- and long-term effects on the desert surface of Kuwait and neighboring areas. Short term effects included the following:

1. Evaporation of aromatic compounds such as benzine from the surfaces of the oil lakes created a hazard to humans, animals, and plants.
2. Increased dust storm activity due to the release of fine silt and clay in areas previously protected by course desert pavement increased hazards to aviation in the whole Gulf region. In addition, mineral aerosols in the air constituted a hazard to humans if lodged in the throat, lungs, and eyes.
3. Movement of sand and its accumulation over minefields and ordnance pits hampered the detection of mines and unexploded bombs. In early 1992, several explosions caused severe injury to people and death to grazing animals.
4. The movements of heavy military vehicles as well as the oil and soot from the well fireplumes severely reduced plant cover in the open desert.

Long-term effects include the following:

1. An increase in sand storm activity due to the mobilization of sand-sized particles. Such storms

interfere with all modes of transportation and cause traffic accidents.

2. The formation and movement of sand dunes, which encroach on roads, farms, oil installations, and settlements in the desert. Once a dune is formed, it continues to move as long as the wind blows. For this reason, sand dune encroachment represents a longest term impact on the desert surface of Kuwait.

3. An increase in soil acidity due to the interaction of moisture with remnants of the oil on the desert surface. This increase hampers the rejuvenation of some natural plants.

4. Potential pollution of the groundwater resources from seepage, through fractures, of oil from the stagnating pools of crude petroleum. This represents a long-term impact particularly since agriculture in Kuwait depends solely on groundwater reservoirs.

For these reasons, mitigation of the effects of the Gulf War on the desert surface of Kuwait will require decades for remediation and monitoring of its effects. Remediation approaches include the following:

1. Filling of empty trenches using recently mobilized sand to limit the exposure of sand-sized particles to the action of wind.

2. Leveling of berms and other sand barriers and walls because these structures represent a source of sand for future activation and mobilization by wind.

3. Stabilizing the flattened surface either by increasing desert plant cover or, more efficiently, by spreading gravel on top of the soil to mimic the natural desert pavement.

4. Removing the oil-saturated soil from areas covered by oil lakes to limit potential long-term effects on soil acidity or groundwater pollution.

5. Replacing the removed soil with clean sand to return the land to the original flat contours.

6. Stabilizing the new surfaces by spreading gravel to form the armor against the erosive action of wind.

7. Seeding of areas of the desert that lost vegetation cover using, as much as possible, seeds of natural plants.

6. FUTURE MONITORING

The sources of environmental impacts on the desert surface due to the Gulf War are related to the invasion of Kuwait by Iraqi forces, the activities that persisted during the occupation phase, the initiation of liberating Kuwait by coalition forces and the events that accompanied the retreat of Iraqi troops, and the environmental impact of fighting the oil well fires and the spread of the oil lakes on the desert surface of Kuwait.

Some remediation of the environmental effects of the Gulf War was accomplished during the past decade. This is particularly true in the case of oil lakes. Most of the oil was pumped and bioremediation of the soil was accomplished. However, the more subtle impacts, such as those of sand movement and plant growth, have not been given equal priority. Therefore, the need exists for a long-term effort to monitor these factors and study the best ways to limit their impact on the natural environment.

These impacts can be monitored by use of satellite images, including National Oceanic and Atmospheric Administration satellites to monitor the initiation of dust storms, Landsat Thematic Mapper and SPOT images to monitor sand accumulation and movement, and higher resolution (1 m) satellite images to monitor changes to the topsoil. Interpretations of satellite image data need to be checked and confirmed in the field at specific field stations. At these stations, instruments should be installed to measure the activities discussed in the following sections.

6.1 Dust Storms

The exposure of vast amounts of soil to the action of wind causes an increase in the frequency and ferocity of dust storms. These hamper visibility on roads and at airports and cause respiratory ailments. The need exists to monitor the initiation of dust storms upwind for advance warning of the impending danger.

6.2 Sand Accumulation

The wind carries enormous amounts of sand in the flat terrain of Kuwait. The sand begins to accumulate where the wind regime is affected by topographic impediments. Such accumulations hamper transport and operations in the oil fields and other installations in the desert. Thus, a monitoring system should be established to alleviate the impact of encroaching sand dunes by stabilizing the moving sand with desert vegetation.

6.3 Soil Rejuvenation

The oil rain has caused much damage to the natural desert soils in Kuwait. It is feared that additional harm to the soil might be caused by an increase in acidity. A program needs to be instituted to monitor

soil conditions and characteristics so that a remedy to the potential problem can be carried out in a sustained manner. This can be accomplished by monitoring soil acidity and its treatment wherever pH levels require it.

6.4 Plant Growth

Degradation of the soil has caused much damage to natural desert plants in Kuwait. Some plants have died in areas covered by the oil soot and droplets, and other species have appeared that are not palatable to grazing animals. The growth of plants should be monitored in view of the drastic changes in the desert environment.

SEE ALSO THE FOLLOWING ARTICLES

Arid Environments, Impacts of Energy Development in • Crude Oil Releases to the Environment: Natural Fate and Remediation Options • Crude Oil Spills, Environmental Impact of • Hazardous Waste from Fossil Fuels • Nationalism and Oil • National Security and Energy • Oil Crises, Historical Perspective • War and Energy

Further Reading

Al Doasari, A. (2000). Analysis of the changes in the tarcrete layer on the desert surface of Kuwait using satellite imagery and cell-based modeling. Ph.D. dissertation, Boston University, Boston, MA.

Al-Hassan, J. M. (1992). "The Iraqi Invasion of Kuwait, an Environmental Catastrophe." Fahad & Al Marzouk, Kuwait.

El-Baz, F., and Al-Sarawi, M. (1996). Kuwait as an alluvial fan of a paleo-river. *Z. Geomorphol. N. F. Suppl.-Bd.* **103,** 49–59.

El-Baz, F., and Al Sarawi, M. (2000). "Atlas of the State of Kuwait from Satellite Images." Kuwait Foundation for the Advancement of Sciences, Kuwait.

El-Baz, F., and Makharita, R. M. (2000). "The Gulf War and the Environment." Gordon & Breach, Lausanne, Switzerland.

Holden, C. (1991). Kuwait's unjust desserts: Damage to its desert. *Science* **251,** 1175.

Khalaf, F. I., Gharib, I. M., and Al-Kadi, A. S. (1982). Sources and genesis of the Pleistocene gravelly deposits in northern Kuwait. *Sedimentary Geol.* **31,** 101–117.

Koch, M., and El-Baz, F. (1998). Identifying the effects of the Gulf War on the geomorphic features of Kuwait by remote sensing and GIS. *Photogrammetic Engineers Remote Sensing* **64**(7), 739–747.

United Nations (1991). Kuwait. Report to the secretary-general on the scope and nature of damage inflicted on the Kuwait infrastructure during the Iraqi occupation, Security Council Document No. S/22535. United Nations, New York.

U.S. Geological Survey (1963). Geologic map of the Arabian Peninsula. U.S. Geological Survey (Washington, D.C.) and Arabian-American Oil Company (Aramco, Dhahran, Saudi Arabia). [U.S. Geological Survey Misc. Geol. Inv. Map I-270; scale 1:2,000,000]

Hazardous Waste from Fossil Fuels

MAJID GHASSEMI
New Mexico Institute of Mining and Technology
Socorro, New Mexico, United States

PAUL K. ANDERSEN and ABBAS GHASSEMI
New Mexico State University
Las Cruces, New Mexico, United States

RUSSELL R. CHIANELLI
University of Texas at El Paso
El Paso, Texas, United States

1. **Overview of the Fossil Fuel Cycle**
2. **Wastes Associated with Petroleum and Its Products**
3. **Waste Associated with Coal and Its Production**
4. **Summary**

Glossary

asphaltenes The part precipitated by addition of a low-boiling paraffin solvent such as normal pentane and benzene-soluble fraction, derived from carbonaceous sources such as petroleum, coal, or oil shale.

boiler slag A black granular material that is coarser than conventional fly ash.

BTEX The BTEX chemicals (benzene, toluene, ethylbenzene, and xylenes) are volatile monoaromatic hydrocarbons that are commonly used in crude petroleum and petroleum products.

characteristic hazardous waste Waste solids, liquids, or containerized gases that exhibit ignitability, corrosivity, reactivity, or toxicity characteristics.

coal bottom ash The coarse, granular, incombustible by-product that is collected from the bottom of furnaces that burn coal.

coal combustion by-products (CCBs) Fly ash, bottom ash, boiler slag, and flue gas emission that are produced when coal is burned.

drilling mud Waste stream associated with drilling operations, including oil-based mud (OBM), synthetic-based mud (SBM), and water-based mud (WBM).

drilling wastes Wastes that are associated with oil exploration and production (E&P).

exploration wastes Wastes that are primarily related to drilling and well completion.

fly ash The fine powder formed from the mineral matter in coal, consisting of the noncombustible matter in coal plus a small amount of carbon that remains from incomplete combustion.

flue gas desulfurization (FGD) gypsum Also known as scrubber gypsum; it is the by-product of an air pollution control system that removes sulfur from the flue gas in calcium-based scrubbing systems.

fossil fuels The biomass that produces coal, gas, oil, and tar sands.

heavy crude Crude oil that contains a relatively high portion of residuum.

listed hazardous wastes Wastes from nonspecific sources and wastes from specific sources and discarded commercial chemicals.

methyl tertiary-butyl ether (MTBE) MTBE is a gasoline additive and is highly water soluble.

naphtha The petroleum fraction ranging from low-boiling C_4 hydrocarbons to those boiling as high as approximately 220°C.

polycyclic aromatic hydrocarbons (PAHs) Often referred to polynuclear aromatics (PNA); they are a class of very stable organic molecules composed of only carbon and hydrogen.

produced water Water produced in association with crude oil.

proppants/frac sand wastes Semisolid sludge consisting of aluminum silicate beads and formation sand.

residuum The petroleum fraction boiling above 343°C.

Superfund A commonly used name for the federal Comprehensive Environmental Response, Compensation and Liability Act (CERCLA).

 © 2004 Elsevier Inc. All rights reserved.

tank bottom wastes Sediment that accumulates in the bottom of oil field vessels and pipelines when fluid turbulence is low.

waste oil Oil arising as a waste product of the use of oils in a wide range of industrial and commercial activities.

workover and completion wastes Wastes from operations in which an oil well's head is partially open to the atmosphere and is filled with a water-base fluid that maintains pressure on the formation to prevent blowout.

The world uses two basic fossil fuels for everyday life. Coal production is used mainly for the production of electricity and petroleum products are mainly used for transportation. The population of the United States uses an amount of fossil fuel in one year that took nature approximately one million years to produce! There are major environmental problems involved with the production, transportation, and consumption of these fossil fuels. Petroleum production has environmental problems at every stage of production, transportation, and consumption. During production, the waste products include drilling mud, drill cuttings, chemical additives, and produced water. In fact, there are 7.5 barrels of produced water for one barrel of oil. The main environmental problem of petroleum refining and storage include oil process wastes and leakage of the storage tanks. Pipeline leaks are also among major environmental issues for transportation of petroleum products. The production of coal requires mining operations. The initial mining virtually ravages the land. The processing of the coal leaves large amounts of tailings and other by-products. Additionally, the use of coal causes major problems with air quality. The fact that the United States uses one billion tons of coal annually for electricity alone is an extremely large environmental problem.

1. OVERVIEW OF THE FOSSIL FUEL CYCLE

Energy fosters human activities and is a necessity for modern life. Energy heats and cools buildings, powers industries, fuels transportation system, etc. All energy on Earth comes from the sun, which contributes to the formation of biomass. The biomass produces decay, which becomes buried in sediments, thereby produces coal, gas, oil, and tar sands. These are called fossil fuels. Figure 1 illustrates a detailed fossil fuel cycle. Fossil fuels are nonrenewable. In 1 year, mankind expends an amount of fossil fuel that took nature approximately 1 million years to produce.

Energy production from fossil fuels can create major environmental problems at every stage of production, transportation, and consumption. Fossil fuels pose risks to the environment. In this article, we discuss the wastes that are associated with fossil fuel in its production, transportation, and consumption.

2. WASTES ASSOCIATED WITH PETROLEUM AND ITS PRODUCTS

Petroleum is derived from Latin *petra*, meaning rock or stone, and Latin *oleum*, meaning oil. Crude oil is a natural occurring material extracted from formations beneath the earth via established drilling production processes. Crude oil refining operations involve extracting useful petroleum products that include naphtha, middle distillate, and residuum. The petroleum fraction ranging from low-boiling C_4 hydrocarbons to those boiling as high as approximately 220°C is called naphtha, which is used primarily in gasoline production. The petroleum fraction boiling from approximately 150 to 343°C is called middle distillate and is used for jet fuel, diesel fuel, and home and commercial heating oils. The petroleum fraction boiling above 343°C is called residuum. The portion that boils from approximately 343 to approximately 566°C is called vacuum gas oil and has no natural market. Material boiling above 566°C is called vacuum residuum, which is a heavy and tarry substance containing a heavier asphaltene fraction.

If crude oil contains a relatively high portion of residuum, it is called heavy crude. Residuum is burned as an industrial fuel in power plants and in ship boilers or is asphalted to make road material. Since gasoline is in great demand, petroleum processes convert lighter fractions (reforming processes) and heavier fractions (cracking processes) to gasoline-range products.

The petroleum industry produces a large percentage of the United State's hazardous waste sites. Various wastes are generated during petroleum exploration, production, and refining. During exploration and production, large quantities of low-toxicity solid waste and wastewater are produced.

In the United States, petroleum production and transportation result in the leakage of approximately 280 million barrels of petroleum (crude oil) each, creating serious environmental problems. Petroleum refinery wastes result from processes designed to

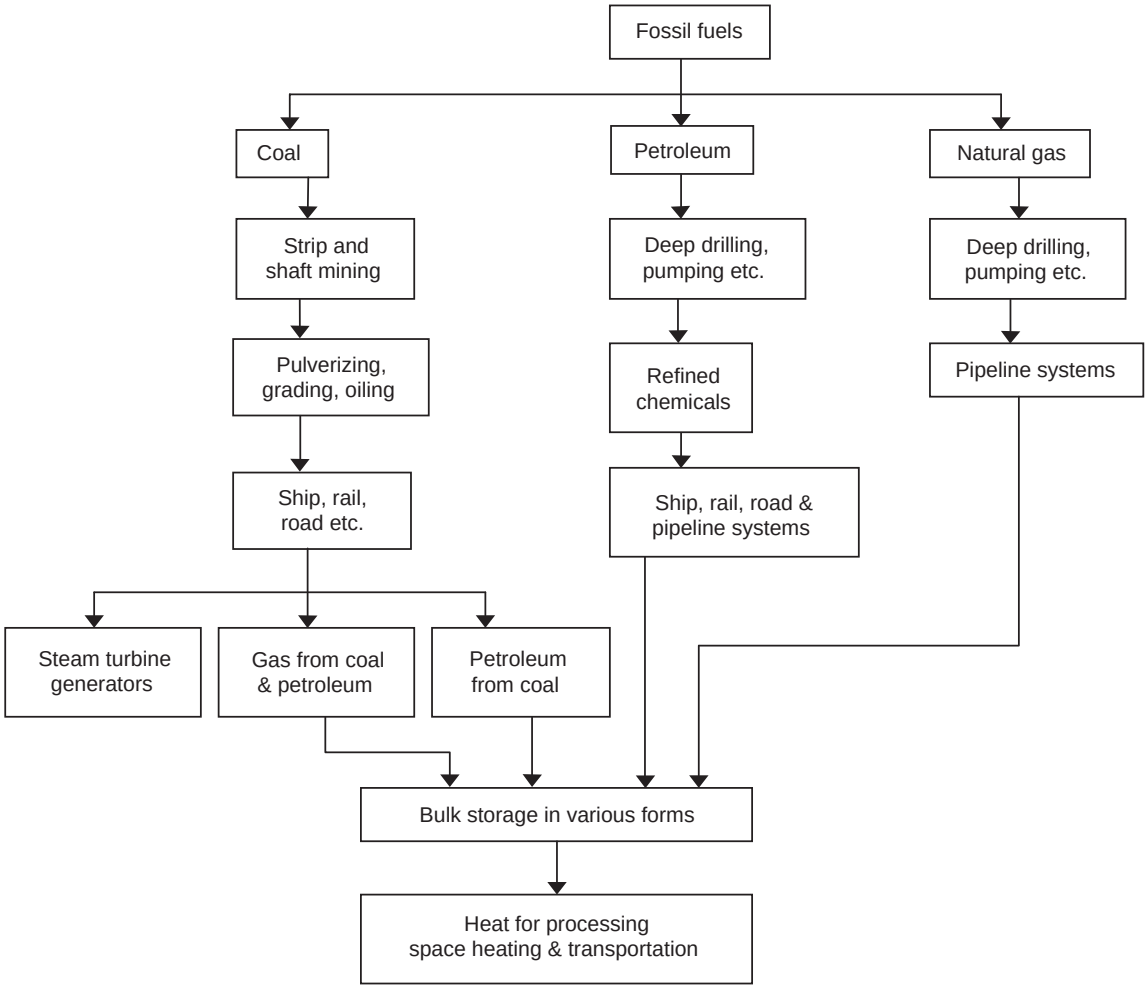


FIGURE 1 Detail of the fossil fuel cycle.

remove naturally occurring contaminants in the crude oil, including water, sulfur, nitrogen, and heavy metals. The waste streams from refineries are smaller but are considerably more toxic, containing listed hazardous waste and characteristic hazardous waste. In addition, poor housekeeping practices during exploration, production, and refining can produce excessive waste.

2.1 Waste from Exploration, Production, and Refining

Oil exploration and production (E&P) operations generate large volumes of drilling waste every year.

2.1.1 Waste from Petroleum Exploration

Exploration wastes are primarily related to drilling and well completion. Table I lists oil production, drilling, and exploration waste streams. Between 1986 and 1995, an average of 313 wells were drilled each year. Drilling operations generate more than 1000 tons of drilling waste per well. This means that more than 300,000 tons of drill cuttings may be produced each year. In addition, the American Petroleum Institute (API) survey of onshore and coastal operators in the United States estimates that approximately 149 million barrels (bbl) of drilling waste (17.9 million bbl of produced water and 20.6 million bbl of other associated wastes) were generated from E&P operations in 1995. The API has also estimated that the average volume of produced water was approximately 7.5 barrels of water per barrel of oil in 1995.

Drilling wastes include drilling mud, drill cuttings, wash water, and chemical additives (Table I). Drilling mud comprises the largest waste stream associated with drilling operation. There are three types of

TABLE I

Oil Production, Drilling, and Exploration Waste Streams[a]

Produced water	
Benzene	Hydrocarbons
Toluene	Radioactive pollutants Ra^{226} and Ra^{228}
Miscellaneous surface water	
Snowmelt and precipitation	Engine coolant
Runoff from cleaning vehicles and shops	Unused cement slurry
Rigwash	
Drilling mud	
Heavy metals	Chlorides (including potassium chloride, sodium chloride, calcium chloride, and magnesium chloride)
Polynuclear aromatic hydrocarbon components (OBM only)	
Barium, gypsum, barites	Clays (bentonite, kaolinities)
	Additives (biocides, scale inhibitors, corrosion inhibitors)
Drill cuttings	
Pieces of rocks and sand	Soil and wet sludge containing oil and water
Workover and completion waste	
Well completion, treatment and stimulation fluids	Fracturing media
Surfactants	Used filters
Weighting agents/viscosifiers	Biocides
Thinners	Packing fluids
Muds (WBM or OBM associated with workovers)	Detergents
Produced water (associated with workovers)	Defoamers
Crude oil	Paraffin
Acidizing agents	Sludges
Corrosion inhibitors	Pieces of downhole equipments
Gels	Intermaterial from downhole mechanical repair (produced sands, formation or pipe scale, cement cutting, and slurries)
Solvents	
Paraffin solvents and dispersants	
Proppants/frac sand	
Semisolid sludge containing the aluminum silicate beads along with formation sand	
Bottom waste	
Solids	Accumulated heavy hydrocarbons (asphaltic or paraffin)
Sands	Drilling mud
Emulsions	Semisolid waste
Pigging wastes	
Produced water	
Oily debris, spent filters, and spent filter media	
Rags	Sand
Sorbent materials	Diatomaceous earth
Water with surfactant (backwash)	Particulates
Oil	Socks
Coal	Cartridges
Gravel	Canisters
Hydrocarbon wastes	
Dirty diesel	Organic halides
Hydrocarbon-bearing soil	Hydrocarbon solids
Camp wastes	
Surplus paints	Wood cribbing supporting temporary pipelines
Batteries	Thread protectors from pipe ends
Light bulbs	Drums for chemical storage

continues

Table I continued

Radioactive smoke detector elements	Oil and air filters on vehicles
Wood pallets	Vehicle lube oil
Fabric pit liners	
Ozone-depleting chlorofluorocarbon	
Halons	Freon
Naturally occuring radioactive materials	
Spent filters	Produced water
Contaminated soil	Completion waste
Camp waste	Hydrocarbon waste

[a] Abbreviations used: OBM, oil-based mud; WBM, water-based mud.

drilling mud: oil-based mud (OBM), synthetic-based mud (SBM), and water-based mud (WBM). Most onshore wells are drilled with WBM or OBM. The key constituents of OBM, SBM, and WBM contaminated cuttings are summarized in Table I. The leachability, runoff, and seepage of heavy metals, chlorides, and the total polyaromatic hydrocarbons degrade the soil and contaminate groundwater, fish, and shellfish. Through the consumption of contaminated fish and shellfish, they can also endanger human health.

2.1.2 Waste from Petroleum Production

Production operations generate massive quantities of produced water. Produced water—water produced in association with crude oil—is by far the largest waste stream in most oil fields. The volume of produced water far exceeds the volume of drilling wastes (approximately 60 times more). The produced water is discharged into surface waters, injected underground, or transported to a commercial oil field waste disposal facility. The majority of produced water is disposed of underground through injection wells, which is allowed by the U.S. Environmental Protection Agency (EPA).

Produced water and drilling fluids could damage agricultural land, crops, ephemeral streams, and livestock and threaten endangered species, fish, and other aquatic and bird life. Metals and polycyclic aromatic hydrocarbon (PAH) discharge cause contamination of aquatic and bird life. Also, improper functioning of injection wells can degrade groundwater. In addition, oil field waste disposal facilities pose significant risk to migratory birds and other wildlife.

Workover and completion wastes result from operations in which an oil well's head is partially open to the atmosphere and is filled with a water-based fluid that maintains pressure on the formation to prevent blowout. Excluding the fluids that were disposed via on-site underground injection, the API calculates that U.S operators generated approximately 5.6 million bbl of completion and workover fluids in 1985. More current data are not available.

The environmental impact related to injection of workover and completion fluids is the potential migration of wastes to underground drinking water sources. Also, seepage of native brines from improperly plugged, unplugged, and abandoned wells could damage groundwater, agricultural land, and domestic and irrigation water.

Tank bottom wastes are a type of sediment that accumulates in the bottom of oil field vessels and pipelines when fluid turbulence is low. Tank bottom wastes are generated continuously at nearly all oil production operations. Oily debris saturated with crude oil comes from oil spill cleanups (minor and major) and spent filter and filter media waste comes from filtering crude oil. According to the API, U.S operators generated a total of 1.5 million bbl of bottom tank wastes and oily debris in 1985. From this total, 1,232,000 bbl (42 U.S. gallons per barrel) were bottom tank wastes and 1,261,000 bbl oily debris.

Potential environmental impact of tank bottom wastes and oily debris may come from landspreading, roadspreading, on-site pits, and on-site burial. The potential impact of landspreading and roadspreading may be a risk of off-site migration of constituents of concern due to surface water run-on and runoff. In addition, off-site transportation could result in high oil and grease concentration wastes to soil and/or vegetation. Also, on-site pits and on-site burial could potentially impact groundwater due to downward migration of constituents of concern and also surface water and soils.

Proppants/frac sand wastes are related to the aluminum silicate beads that are injected into wells to hold formation fractures open, along with formation sand, that together form a semisolid sludge. Hydrocarbon wastes such as dirty diesel fuel are contaminated from pressure testing pipelines. Camp wastes are equivalent in quantity to those of small municipalities.

Ozone-depleting chlorofluorocarbons (CFCs) are used in oil and gas production as a fire and explosion suppressant. Naturally occurring radioactive material is produced in solution with oil field brines and deposited as scale on the inside of oil field vessels and piping. Its radioactivity (rated as "low specific activity" by the Nuclear Regulatory Commission) poses a hazard if it is inhaled as a dry dust during vessel and pipe cleaning or repair.

2.1.3 Waste from Petroleum Refining

Table II shows the waste generation data for petroleum refineries as given by the API survey of U.S. refineries. The major sources of wastewater in petroleum refinery are cooling tower blowdown, boiler blowdown, oily wastes, and other process wastes. Solid wastes from refineries are also a significant problem.

In 1976, Congress enacted the Resource Conservation and Recovery Act (RCRA) for reduction of hazardous waste. This early hazardous waste legislation imposed "cradle to grave" liability on waste generators for the waste that they produce and dispose of. In 1984, Congress passed the Hazardous and Solid Waste Amendments (HSWA), which amended RCRA in an attempt to reduce groundwater contamination by banning or regulating landfill disposal of specific listed hazardous wastes. The refining sector of the petroleum industry generates both HSWA listed and other hazardous wastes. Five refinery solid waste streams are listed hazardous wastes by the EPA and subject to HSWA regulations: dissolved-air-flotation floating solids (K048), slop-oil-emulsion solids (K049), heat-exchanger-bundle cleaning solids (K050), API separator sludge (K051), and leaded-gasoline-tank bottoms (K052).

TABLE II

Refining Residual Stream

Oily sludge and other organic residuals	
American Petroleum Institute separator sludge	Primary sludge (F037)
Dissolved-air-flotation unit float	Primary sludge (F038)
Slop oil emulsion solids	Nonleaded tank bottoms
Leaded tank bottom	Waste oils/spent solvents
	Other oily sludge/organic residuals
Pond sediments	
Contaminated soils and solids	
Heat exchanger bundle cleaning sludge	Residual/waste sulfur
Contaminated soils/solids	Other contaminated solids
Residual coke/carbon/ charcoal	
Spent catalysts	
Fluid cracking catalyst or equivalent	Other spent catalyst
Hydroprocessing catalyst	
Aqueous residuals	
Biomass	Spent stretford solution
Oil-contaminated water (not wastewater)	Treatment, storage and disposal leachate (F039)
High-Ph/low-Ph waters	Other aqueous residuals
Spent sulfide solution	
Chemicals and inorganic residuals	
Spent caustics	Residual amines
Spend acids	Other inorganic residuals
Other residuals	

Several refinery, petrochemical, waste treatment, recycling, and disposal sites have been listed as Superfund sites. An example of a refinery and petrochemical Superfund site is the Sikes Disposal Pits Superfund site. According to the EPA, numerous refineries and petrochemical companies have disposed of chemical and oil-based wastes at the Sikes site. Other examples are the Baldwin Waste Oil site and the Falcon Refinery site, also known as the National Oil Recovery Corporation. The R&H Oil Company site, also known as the Eldorado Refining and Marketing, Inc. site and Tropicana Energy site, and the Dubose Oil Products Company Superfund site, which is a waste treatment, recycling, and disposal facility, are other examples of Superfund sites.

A major portion of refining sector hazardous wastes are not HSWA listed but are characterized as hazardous waste because testing has shown that they have one or more characteristics of hazardous waste (i.e., toxicity, ignitability, reactivity, and corrosivity). These wastes include primary treatment plant sludge, secondary treatment plant sludge, biological treatment sludge, cooling tower sludge, ion exchange regenerant, fluidized catalytic cracking catalysts, and hydrocracking catalysts.

Direct releases from a refinery, especially air emission, can pose significant risk to public, employee, and environmental health. These releases may involve process, storage, transport, fugitive, and secondary emissions.

2.2 Waste from Petroleum Products and Petroleum Storage

Oils originating from petroleum crude can be processed, blended, and formulated to produce numerous products for many applications. Users of

oils and oil products include not only industrial and commercial interests but also the individual citizen through the use of gasoline, lubricants, and fuel oils. These users of petroleum products produce large quantities of waste. Waste oils, due to the nature of their use, contain various contaminants, including heavy metals, combustion by-products, and substances arising from the original use (e.g., PCBs).

The principal uses of oils are as fuels, lubricants, hydraulic and transmission fluids, heat transfer fluids, and insulants. As fuels, oils cover a wider spectrum from light, highly mobile liquids to heavy fuel oils of such low viscosity at ambient temperatures that they must be heated to enable their use. The transportation industry is a major user of oil as a fuel for aircraft, shipping, and many land-based transport systems.

Petroleum products (gasoline, diesel fuel, crankcase oil, and waste oil) are a complex mixture of hundreds of chemicals with individual toxic effects. Those that are most often measured are referred to as BTEX (benzene, toluene, ethylbenzene, and xylenes). The BTEX chemicals are volatile monoaromatic hydrocarbons that are commonly found together in crude petroleum and petroleum products. They are considered one of the major causes of environmental pollution because of widespread occurrences of leakage from underground petroleum storage tanks and spills at petroleum production wells, refineries, pipelines, and distribution terminals. It has been estimated that 35% of the 1.4 million gasoline storage tanks in the United States are leaking. Contamination of groundwater with the BTEX compounds is difficult to remedy because these compounds are relatively soluble in water and can diffuse rapidly once introduced into an aquifer.

Some of the petroleum waste products are gasoline, oil, lubricants, machining or cutting oils, leakage from storage tanks, aqueous waste, sulfur dioxide, and PAHs.

2.2.1 Gasoline

When gasoline or another petroleum product is released into the environment, primary concerns generally include the possibility of contaminating soil and ground- and surface water. Gasoline spill and leaking underground storage tank contaminants are known to be a source of groundwater contamination. The United States uses approximately 16 million bbl of oil and oil by-products, such as gasoline, each day (245 billion gallons per year). Much of this material is stored in tanks at some point during transport from source to point of use. There are approximately 6 million tanks in the United States for storing petroleum, petroleum by-products, and petroleum-related chemicals. Many of the storage tanks are underground. According to the EPA, approximately 1 out of every 4 underground storage tanks may be leaking. Even a small gasoline leak of one drop per second can result in the release of approximately 400 gallons of gasoline into the groundwater in 1 year. A few quarts of gasoline in the groundwater may be enough to severely pollute a farmstead's drinking water.

As gasoline seeps through the soil and encounters groundwater, some components will dissolve and travel with the groundwater. After seeping underground, the petroleum products remain near the top of the water table because they are lighter than water. When wells draw water from the upper portions of the water table, they collect the contaminants during periods of low water availability and high consumption.

A plume of gasoline contaminants has been found in groundwater due to spills and leaking underground storage tanks. Petroleum fuels contain a number of potentially toxic compounds, including common solvents, such as benzene, toluene, and xylene, and additives, such as methyl tertiary-butyl ether (MTBE), ethylene dibromide (EDB), and organic lead compounds. The gasoline additive MTBE has contaminated many thousands of drinking water supplies and wells. MTBE causes widespread groundwater contamination due to its high water solubility and other unique properties. The strong, terpene-like odor of MTBE makes its presence easy to detect, even in minute quantities.

These contaminants (petroleum fuel toxic compounds) are all confirmed or potential carcinogens and could pose a significantly increased cancer risk. MTBE may cause cancer and myriad other health problems, such as asthma and respiratory illnesses. EDB and benzene are also considered carcinogens (cancer-causing).

2.2.2 Crude Oil (Petroleum)

Oils are not especially toxic, but contaminants such as additives, breakdown products, and other substances that may have become mixed with oils during their legitimate use may be much more so. Oils, however, do have considerable potential to cause environmental damage by virtue of their persistence and their ability to spread over large areas of land or water. Films or coverings of oil substances may reduce or prevent air from reaching life-forms of all types within an area of land or sea and can rapidly result in significant degradation of environmental quality in those media. Waste acid sludge arising

from the acid/clay process, a common method of recycling waste oils, represents a serious environmental threat.

2.2.3 Lubricants

Lubricants embrace an even wider spectrum of products with many different viscosities and viscosity ranges and the extensive use of performance-enhancing additives. A major source of waste oils is from lubrication. Lubricating oils need replacing at intervals, and although improved designs of engines and better oil grades may reduce the frequency of change, there is little prospect of the need for oil changes being avoided altogether. Major contributors to this market are cars, commercial vehicles, aircraft, railway locomotives, and other large pieces of machinery. Oils from these sources usually contain not only contamination from their use but also a variety of additives intended to improve performance. For example, lubricants are used for propulsion engines. Motor vehicles of modern design have several different non-fuel oil applications, such as engine crankshaft lubrication, gearbox lubrication, and hydraulic fluids for brake systems. Motor vehicles and their use impinge on most aspects of life.

2.2.4 Machining or Cutting Oils

Machining or cutting oils are used extensively for metal cutting, machining of components on lathes, and general engineering activities. In many instances, depending on the metal being machined, oils for this purpose are used as an emulsion in water with detergents, biocides, chlorinated paraffins, halogenated and nonhalogenated additives, and additives included to improve performance. The waste, which can be substantial in volume, comprises an emulsion of typically 2–5% oil in water. The waste may also contain metal dusts and similar contaminates.

Unintentional releases of process fluid from equipment are another source of waste. Any equipment that can potentially leak (e.g., pump, valve, and flanges) is a source of pollution. An oil refinery may have 250,000 such pieces of equipment.

2.2.5 Storage Tanks

Another source of pollution is leakage from storage tanks. At refineries and other oil/product handling facilities, there are many aboveground crude oil and refined product storage tanks in service. API Standard 653 requires tanks to be taken out of service periodically and to be visually inspected for leaks. However, this standard is often ignored since inspection can be very costly. Buried pipelines within refinery boundaries represent another source of leaks, and according to the API they have caused more groundwater contamination than the more visible storage tanks.

Tank washings and sludge from storage tanks and refineries, including those from the cleaning of oil storage tanks, can be a source of contamination since they contain breakdown products. In addition, tank washings can generate substantial volumes of waste, usually with medium to low oil contents. Typical sources are oily bilge water, oily ballast water, water from the jetting and cleaning of tanks both onboard ships and at land-based installations, and true tank-cleaning residues. The range of oils encountered in the generation of these wastes can be broad and can include accumulated oil sludge and sediment, tank scale, and even grit from grit blasting and related cleaning work.

2.2.6 Aqueous Waste

Many oil wastes are highly aqueous in nature and must first be subject to treatment to separate the waste into an oil-rich fraction that may be usable in some way and an aqueous waste stream that may require further treatment prior to final disposal. Processes of this type usually yield sludges that, along with other oil waste-based sludges, have no beneficial use and will require final disposal. Aqueous wastes are unlikely to be entirely free from oil and may need further treatment. Refining processes such as coking sulfur recovery, steam cracking, hydrocracking, and crude desalting create large volumes of oily wastewater. Wastewater facilities in particular are considered to be a source of significant atmospheric hydrocarbon emissions. Wastewaters constitute more than 80% of total hazardous waste mass at refineries. Also, aqueous wastes may contain other contaminants, depending on the original process, including surfactants, solvents, and toxic metals. PAHs are of particular concern because of their carcinogenic properties.

2.2.7 Sulfur Dioxide

All fossil fuels, including crude oil, contain organic sulfur compounds. When the fuels are burned, sulfur oxides are emitted. Sulfur oxides can cause respiratory problems, but emissions of these compounds are a concern primarily because they react in the atmosphere, forming acids. Over time, deposition of these acids through precipitation and diffusion can increase the acidity of natural bodies of water and change local ecosystems. Crude oil contains materials

such as soils, metals, and sulfur- and nitrogen-containing compounds. Removing these materials during refining results in large volumes of waste.

2.2.8 Polycyclic Aromatic Hydrocarbons

Polycyclic aromatic hydrocarbons, often referred to as polynuclear aromatics, are a class of very stable organic molecules composed of only carbon and hydrogen. These molecules are flat, with each carbon having three neighboring atoms, much like graphite. PAHs are released into the atmosphere as gases or particles during the incomplete combustion of organic material. PAHs have a number of sources, including cars, trucks, ships, and aircraft; industrial-power generation; oil refining and waste incineration; and primarily domestic combustion for heating and cooking, especially solid fuel heaters using wood and coal.

Of the 1410 hazardous waste sites listed on the National Priorities List (NPL) (Superfund or NPL sites), 352 have been contaminated by PAHs found in soil, groundwater, and sediment. Exposure to PAHs usually occurs by breathing air contaminated by wildfires or coal tar or by eating foods that have been grilled. The Department of Health and Human Services has determined that some PAHs may reasonably be expected to be carcinogens.

The smallest member of the PAH group is naphthalene, a two-ring compound that is a gas at normal temperatures. Three- to five-ring PAHs occur as either gases or particles in air. PAHs with five or more rings tend to be solids that attach to the surface of other particulate matter in the atmosphere.

2.3 Pipeline Leaks

Pipelines have historically been a primary source of oil spills both offshore and onshore. Pipeline leaks are not a rare occurrence. When they do occur somewhere within the 160,000 miles of liquid fuel-carrying pipelines that criss-cross the nation, they can have disastrous consequences. In 2000, gasoline leaking from an Olympic Pipe Line Company line in Bellingham, Washington, caused an explosion that killed three people. During the 1990s, major reportable spills, those involving at least 2100 gallons, occurred approximately four times a week and totaled approximately 6 million gallons per year. In 1996, a rupture in a 36-in. Colonial pipeline released more than 1 million gallons of diesel fuel into the Reedy River near Fork Shoals, South Carolina, creating a slick that spread for more than three-fourths of a mile down the waterway. A 1993 leak from a Colonial pipeline in Virginia dumped more than 300,000 gallons of oil into a tributary of the Potomac, forcing Fairfax to temporarily switch to an alternate water supply. A 1991 leak from a Colonial pipeline near Spartanburg, South Carolina, spewed an estimated 562,000 gallons of oil into a tributary of the Enoree River, forcing two towns to close their water-treatment plants. they Tables III and IV summarize the sources and causes of oil spills from the inland transportation and distribution system.

TABLE III

Sources of Oil Spills from Inland Transportation and Distribution System, 1987–1995

Inland source	No. of spills	Volume spilled (millions of gallons)	% of total volume spilled
Fixed facility	43,188	79.5	62
Pipeline	13,771	33.7	26
Highway transport	13,280	6.2	5
Rail transport	2483	3.7	3
Above ground storage tank	921	4.4	3
Underground storage tank	795	0.5	<1
Total	74,438	128	100

TABLE IV

Causes of Oil Spills from Inland Transportation and Distribution System, 1987–1995

Cause	No. of spills	Volume spilled (millions of gallons)	% of total volume spilled
Equipment failure	11,476	22.8	50
Operator error	4235	4.3	9
Transportation accident	3187	3.0	7
Natural causes	583	1.7	4
Dumping	2250	1.0	2
Other	2584	2.3	5
Unknown	934	10.4	23
Total	25,249	45.5	100

2.4 Noise Contamination

Noise in refineries originates from rotating machinery, such as cooler fans, turbines, compressors, engines, and motors. High-velocity flow of fluids

through valves, nozzles, and piping also contributes to the general noise level.

3. WASTE ASSOCIATED WITH COAL AND ITS PRODUCTION

Coal waste products are similar to those produced during petroleum exploration, production, refining, and use. Coal is a solid with a complex hydrocarbon structure and a higher carbon-to-hydrogen ratio than petroleum. Coal is the world's most plentiful fossil fuel energy resource, and its use is likely to quadruple by 2020. Coal has a wide range of moisture content (2–40%), sulfur content (0.2–8%), and ash content (5–40%), causing environmental problems from its use. Burning of coal creates approximately 60% of soot-creating sulfur dioxide emission in the United State. It is also a major source of smog-forming nitrogen oxide pollution, mercury contamination, and carbon dioxide, releasing more than 51 tons of mercury and 2 billion tons of carbon dioxide into the air every year. Sulfur dioxide and nitrogen oxide pollution also creates fine particles, which can cause significant increases in acute and chronic respiratory diseases. At every stage of production, transportation, and consumption, coal poses risks to the environment.

Coal is dirty fossil fuel: Mining it can ravage the land, and burning it can generate a large amount of carbon dioxide and other pollutants. The mining/cleaning of coal creates a significant amount of waste. Wastes that are associated with mine extraction (i.e., surface mining, open-pit mining, and underground mining) and processes of coal include liquid effluents (mine tailings containing fine materials that are discharged as slurry to impoundment areas), solid wastes (coarse materials typically >0.5 mm), and coal dust. According to a survey conducted by the U.S. Department of Energy, approximately 2.5–3 billion tons of fine coal has been deposited in tailing impoundments that typically have a life of up to 20 years.

Mining of coal has left large volumes of mining overburden throughout the United States. Mine overburden and waste soils (commonly referred to as mine tailings) are the waste products generated during mining and milling operations. Mine overburden is any material, consolidated or unconsolidated, excluding topsoil, that lies above a natural deposit of coal. It is also the material after removal from its natural state in the process of surface coal mining.

A complex set of chemical and biological reactions and conditions occur over time in mine overburden. Surface deposition of overburden waste products results in atmospheric exposure and subsequent weathering of sulfide minerals present in the rock, precipitation, and ground- or surface water. Oxidation of sulfide minerals by chemical and microbial processes produces sulfuric acid and heavy metal ions that find their way into surface and groundwaters. The release of these components into natural waters is a widespread and persistent problem in areas of the country that have a long history of coal or precious metal mining. Many of these tailings are causing problems with water quality and revegetation.

Mine overburden piles have acid-producing potential as well as acid-neutralization potential. Generally, the rate of acid production exceeds the rate of acid neutralization. Degradation of water quality of surface and groundwaters in contact with acid-producing mine tailings piles is the major environmental problem associated with acid mine drainage. The pH of the water is lowered due to the introduction of sulfuric acid as a result of sulfide mineral weathering. Acidic mine drainage laden with heavy metals and the precipitation of "yellow boy" has a devastating effect on the ecology of freshwater streams.

Heavy metal ions are released into the water as a result of mineral weathering. Depending on the composition of the mine overburden, the pollutants released into surface and groundwaters typically include any of the following heavy metal ions: iron, zinc, aluminum, manganese, cobalt, nickel, copper, arsenic, selenium, cadmium, and lead.

Mine overburden piles also contain *Thiobacillus ferrooxidans*. *Thiobacillus ferrooxidans* is responsible for the oxidation of iron and inorganic sulfur compounds in mine overburden areas. *Thiobacillus ferrooxidans* is acidophilic (acid loving, capable of surviving at low pH) and has a physiology that is well suited for growth in an inorganic mining environment. It obtains carbon by fixing atmospheric carbon dioxide and is strictly autotrophic. *Thiobacillus ferrooxidans* catalyzes the acid and heavy metal and produces reactions that occur in mine tailings resulting in the degradation of water quality downstream of mining operations.

Approximately 1 billion tons of coal is burned annually to generate electricity in the United States; as a result, more than 97 million tons of coal combustion by-products (CCBs) are produced. CCBs are considered a product by the American Coal Ash Association and are referred to as coal combustion products (CCPs). In 1998, CCPs ranked as the third largest nonfuel mineral commodity produced in the United States. Only one-fourth of these are utilized,

whereas the rest (80% CCBs) are disposed of mainly in landfills and surface impoundments (Table V).

With shrinking landfill space and increasing amounts of ash generation, there is an urgent need to utilize these materials in an environmentally safe manner. One potential use of these materials is waste stabilization. Intrinsic properties of these products allow immobilization of hazardous nuclear, organic, and toxic metal wastes for safe and effective environmental disposal. Additionally, coal ash contains plant nutrients and can be used to amend highly acidic soils.

CCBs produced during coal-fired electric power production are fly ash, bottom ash, boiler slag, and flue gas desulfurization (FGD) material. The physical characteristics and trace element composition of fly ash, bottom ash, boiler ash, and FGD are given in the Tables VI and VII, respectively.

3.1 Fly Ash

Fly ash is the finest of coal ash particles. It is called fly ash because it is transported from the combustion chamber by exhaust gases. Fly ash is the fine powder formed from the mineral matter in coal, consisting of the noncombustible matter in coal and a small amount of carbon that remains from incomplete combustion. Fly ash is generally light tan in color and consists mostly of silt-sized and clay-sized glassy spheres. Properties of fly ash vary significantly with coal composition and plant operating conditions. In the United States, approximately 50 million tons of fly ash is reused annually.

Fly ash is referred to as either cementations or pozzolanic. A cementations material is one that hardens when mixed with water. A pozzolanic material will also harden with water but only after activation with an alkaline substance such as lime. These cementations and pozzolanic properties make some fly ashes useful for cement replacement in concrete and many other building applications. Fly ash is used in concrete and cement products, road base, oil stabilizer, clean fill, filler in asphalt, metal recovery, and mineral filler.

TABLE V

Amount of Coal Combustion By-products (CCBs) Produced in the United States

Material	Metric tons $\times 10^6$
Crushed stone	1500
Sand and gravel	1020
CCBs	97.7
Cement	85.5
Iron ore	62

3.2 Coal Bottom Ash

Coal bottom ash and fly ash are quite different physically, mineralogically, and chemically. Bottom ash is a coarse, granular, incombustible by-product collected from the bottom of furnaces that burn coal for the generation of steam, the production of electric power, or both. Bottom ash is coarser than fly ash, with grain sizes ranging from that of fine sand to that of fine gravel. The type of by-product produced depends on the type of furnace used to burn the coal. Bottom ash is used as construction and railroad fill material, abrasive blasting grit, granules on asphalt roofing shingles, aggregate for concrete and masonry blocks, substitute for sand for traction on icy roadways, and soil amendment to increase permeability.

TABLE VI

Summary of Physical Characteristics and Engineering Properties of Coal Combustion By-products

			Flue gas desulfurization material	
Physical characteristic	Fly ash	Bottom ash/boiler ash	Wet	Dry
Particle size (mm)	0.001–0.1	0.1–10.0	0.001–0.05	0.002–0.074
Compressibility (%)	1.8	1.4		
Dry density (lb/ft^3)	40–90	40–100	56–106	64–87
Permeability (cm/s)	10^6–10^4	10^{-3}–10^{-1}	10^{-6}–10^{-4}	10^{-7}–10^{-6}
Shear strength	0–170	0		
Cohesion (psi)	24–45	24–45		
Angle of internal friction (°)				
Unconfined compressive strength (psi)			0–1600	41–2250

TABLE VII

Trace Element Composition of Coal Combustion By-products

	Fly ash							
	Mechanical		Esp/baghouse		Bottom ash/boiler slag		Dry flue gas desulfurization material	
Element (mg/kg)	Range	Median	Range	Median	Range	Median	Range	Median
Arsenic	3.3–160	25.2	2.3–279	56.7	0.50–168	4.45	44.1–186	86.5
Boron	205–714	258	10–1300	371	41.9–513	161	145–418	318
Barium	52–1152	872	110–5400	991	300–5789	1600	100–300	235
Cadmium	0.40–14.3	4.27	0.10–18.0	1.60	0.1–4.7	0.86	1.7–4.9	2.9
Cobalt	6.22–76.9	48.3	4.90–79.0	35.9	7.1–60.4	24	8.9–45.6	26.7
Chromium	83.3–305	172	3.6–437	136	3.4–350	120	16.9–76.6	43.2
Copper	42.0–326	130	33.0–349	116	3.7–250	68.1	30.8–251	80.8
Fluorine	2.50–83.3	41.8	0.4–320	29.0	2.5–104	50.0	—	—
Mercury	0.008–3.0	0.073	0.005–2.5	0.10	0.005–4.2	0.023	—	—
Manganese	123–430	191	24.5–750	250	56.7–769	297	127–207	167
Lead	5.2–101	13.0	3.10–252	66.5	0.4–90.6	7.1	11.3–59.2	36.9
Selenium	0.13–11.8	5.52	0.6–19.0	9.97	0.08–14	0.610	3.6–15.2	10.0
Silver	0.08–4.0	0.70	0.04–8.0	0.501	0.1–0.51	0.20	—	—
Strontium	396–2430	931	30–3855	775	170–1800	800	308–565	432
Vanadium	100–377	251	11.9–570	248	12.0–377	141	—	—
Zinc	56.7–215	155	14–2300	210	4.0–798	99.6	108–208	141

3.3 Boiler Slag

Boiler slag is coarser than conventional fly ash and is formed in cyclone boilers, which produce a molten ash that is cooled with water. Boiler slag is generally a black granular material with numerous engineering uses.

3.4 FGD Gypsum

FGD gypsum is also known as scrubber gypsum. FGD gypsum is the by-product of an air pollution control system that removes sulfur from the flue gas in calcium-based scrubbing systems. It is produced by employing forced oxidation in the scrubber and is composed mostly of calcium sulfate. FGD gypsum is most commonly used for agricultural purposes and for wallboard production.

4. SUMMARY

Worldwide energy demand increases even as knowledge of how fossil fuel use threatens the global environment grows. Environmental impacts arise at all stages in the life cycle of fossil fuel sources. These impacts are as follows:

- Fossil fuel extraction: coal mining and oil extraction
- Fossil fuel distribution: transport of fuels by road, rail, water, and pipelines; transmission of electricity
- Fossil fuel processing: petroleum refining and coal preparation
- Fossil fuel combustion in the domestic, commercial, industrial, and transport sectors
- Waste and residue management: coal ash and refinery residues

The environmental burdens associated with these activities are many and varied, including:

- Emissions of pollutants to air, mainly from fossil fuel combustion but also from methane leakage from coal and oil extraction and distribution
- Discharges of pollutants to rivers, estuaries, and seas from coal mining and oil extraction
- Groundwater pollution from coal mining
- Visual intrusion, noise, and loss of wildlife habitat

SEE ALSO THE FOLLOWING ARTICLES

Acid Deposition and Energy Use • Air Pollution from Energy Production and Use • Air Pollution, Health Effects of • Clean Air Markets • Clean Coal Technology • Crude Oil Spills, Environmental Impact of • Ecosystem Health: Energy Indicators • Ecosystems and Energy: History and Overview • Electricity, Environmental Impacts of • Gasoline Additives and Public Health • Gulf War, Environmental Impact of

Further Reading

Aldridge, C. L., and Bearden, R. (1985). "Studies on Heavy Hydrocarbon Conversion." Exxon Research and Development Laboratories, Baton Rouge, Louisianna.

Allen, D. T., and Rosselot, K. S. (1997). "Pollution Prevention for Chemical Processes." Wiley, New York.

Fries, M. R., Zhou, J., Chee-Sanford, J., and Tiedje, J. M. (1994). Isolation, characterization, and distribution of denitrifying toluene degraders from a variety of habitats. *Appl. Environ. Microbiol.* **60**(8), 2802–2810.

Gary, J. H., and Handwerk, G. E. (1994). "Petroleum Refining Technology and Economics." Dekker, New York.

Ghassemi, A. (2002). "Handbook of Pollution Control and Waste Minimization: Minimization and Use of Coal Combustion Byproducts (CCBs): Concepts and Applications." Dekker, New York.

Gibbons, J. H., Blair, P. D., and Gwin, H. L. (1990). "Managing Planet Earth; Strategies for Energy Use." Freeman, New York.

Harwood, C. S., and Gibson, J. (1997). Shedding light on anaerobic benzene ring degradation: A Process Unique to Prokaryotes. *J. Bacteriol.* **179**(2), 301–309.

Higgins, T. E. (1995). "Pollution Prevention Handbook." Lewis, New York.

Pacific Northwest Pollution Prevention Research Center (1993). "Pollution Prevention Opportunities in Oil and Gas Production, Drilling, and Exploration." Pacific Northwest Pollution Prevention Research Center. Seattle, Washington.

Salisbury Post (2001, June). Better monitoring needed—Pipeline spill no rare event.

Smith, M., Manning, A., and Lang, M. (1999, November). Research on the re-use of drill cuttings onshore, Report No. Cordah/COR012/1999 for Talisman Energy (UK) Limited.

U.S. Environmental Protection Agency (1995). "Data obtained from EPA Emergency Responsive Notification System," Inland spill data. U.S. Environmental Protection Agency, Washington, DC.

U.S. Environmental Protection Agency (2000, January). "Special Wastes: Crude Oil and Natural Gas: Associated Waste Report; Executive Summary." U.S. Environmental Protection Agency, Washington, DC.

World Bank Group (1998). "Pollution Prevention and Abatement Handbook." World Bank Group, Geneva.

Heat Islands and Energy

HAIDER TAHA
Altostratus, Inc.
Martinez, California, United States

1. **Introduction**
2. **Definition of Heat Islands**
3. **Causes and Effects of Urban Heat Islands**
4. **Typical Magnitudes of UHIs**
5. **Impacts of UHIs on Energy**
6. **Mitigation of UHI Effects to Save Energy**

Glossary

albedo (a) For a flat surface, albedo is generally defined as hemisphere- and wavelength-integrated reflectivity. Thus, whereas reflectivity is the ratio of the amount of reflected to total incident radiation at a particular wavelength, albedo can, by definition, encompass a range of wavelengths of interest. In this article, the interest is in solar albedo with an integral over the portion of the solar spectrum between approximately 0.28 and 2.8 μm. Effective albedo is the albedo resulting from geometrical effects (e.g., multiple reflections) in addition to the effects of the surface albedo. Of course, the definition of albedo is not restricted to surfaces.

boundary layer (BL) (e.g., planetary BL and canopy BL) A layer in the fluid (e.g., air) whose momentum, mass, or heat characteristics are affected or dictated by the fluxes of respective quantities at the underlying surface. Thus, generically, a BL is a layer in which the effects of the surface are "felt."

Bowen ratio (β) The ratio of sensible heat flux (H) to latent heat flux (λe) from a surface or plane, land cover, or region of interest: $\beta = H/\lambda e$. The definition can be applied to any scale.

cost of conserved energy (CCE) In this article, CCE is defined as the compounded additional cost per saved kWh that must be invested in order to implement an energy-saving strategy or feature. The CCE of interest is that of increased roof or urban albedo. $\mathrm{CCE} = [\Delta I/\Delta E] \times [i/\{1-(1+i)^{-n}\}$, where ΔI is additional investment ($), ΔE is the annual energy savings (kWh/year), i is interest or compounding rate (e.g., 6%), and n is the number of years.

degree-days (DD) A measure of the magnitude and duration of temperature deviation from a certain given threshold over a certain period of time (e.g., annual, seasonal, and monthly). Thus, for example, $\mathrm{DD} = \Sigma(T_{\mathrm{avg}} - T_{\mathrm{t}})$, where T_{t} is any threshold (reference) temperature [e.g., 18.3°C (65°F)], and T_{avg} is the 24-h average temperature. The summation can be taken over any period of interest (e.g., 1 year).

efficiency Coefficient of performance (COP), energy efficiency ratio (EER), and seasonal energy efficiency ratio (SEER) are measures of the ratio of work provided by a system (e.g., cooling or heating) to the power input to that system. COP is dimensionless, but EER and SEER typically have units of Btu/h W^{-1}.

evaporative cooling The process of cooling the air by evaporation of water, in which case the heat of vaporization is taken from the air. Thus, evaporative cooling is a constant enthalpy process.

view factor (e.g., sky view factor) The ratio of the amount of radiation (e.g., long-wave flux) leaving a surface (e.g., street level) through the sky aperture "seen" by the surface to the radiation leaving the surface in all directions (e.g., to the hemisphere). The view factor is thus a purely geometrical definition and can be qualitatively defined as the ratio of sky area or aperture (solid angle) seen from a certain surface (or point) to the area (solid angle) of the entire hemisphere (or sphere) seen by that same surface (or point).

Weather conditions and local meteorology, e.g., urban climates, have significant impacts on regional and local energy use intensities and patterns. An important phenomenon of urban climates, which has direct bearings on energy use, is the urban heat island (UHI). The UHI itself is partly a result of energy use in urban areas. When a UHI is present, air temperature can be several degrees higher than in the surrounds and in summer the urban-scale peak electric demand for cooling can increase by up to 10%.

1. INTRODUCTION

Urban heat islands (UHIs) and energy are interrelated in two complementary ways. First, energy

 © 2004 Elsevier Inc. All rights reserved.

conversion and use can contribute to creating UHIs via increased anthropogenic heating of ambient air by machinery, equipment, air conditioners, cooling towers, motor vehicles, and other sources of heat (e.g., combustion). Although this is only one of many UHI causative factors, it can be significant in urban areas, where such sources are highly concentrated. Energy use and conversion can also result in increased emissions of air pollutants, CO_2, nitrous oxide (N_2O), and water vapor (motor exhausts, refineries, chimneys, electricity generators, etc.), which can enhance local (e.g., urban) greenhouse effects. From a relatively more extreme point of view, it can also be argued that passive and active energy conversion systems (e.g., solar and, to a much lesser extent, wind systems) can also add to the UHI. This would occur, for example, if a vast array of photovoltaic cells were installed in an urban area such that the effective urban albedo is decreased and heat transfer from the cells to the air becomes larger. However, this effect is small since photovoltaics and other solar equipment (e.g., solar water heaters) are sparse and cover a small surface area. Thus, although the albedo of a photovoltaic cell may be approximately 0.05, the impact on overall urban albedo is very small. Also, most high-concentration solar energy conversion facilities (e.g., solar power plants and wind farms) are usually located in desert and rural areas, not in urban regions. Thus, they do not contribute to UHIs directly.

Second, UHIs directly affect energy use in buildings; heating, ventilating, and air-conditioning (HVAC) systems and equipment; motor vehicles; industrial processes; and so on. In buildings, more cooling and less heating are generally needed when UHIs are present. In addition, equipment efficiencies (coefficient of performance, energy efficiency ratio, and seasonal energy efficiency ratio) of HVAC, industrial, and other equipment are to a certain extent functions of ambient temperature and thus are affected by UHIs, even if only by small amounts.

2. DEFINITION OF HEAT ISLANDS

A heat island is simply an area, volume, or region in which the temperature is higher than that of the surroundings. This very general definition can obviously apply to any scale and environment, from micrometers to thousands of kilometers. That is, a heat island can exist on a scale of a few millimeters, as a bubble of hot air surrounded by colder air, or on a scale of thousands of kilometers, such as a high-pressure weather system moving across a relatively cooler air mass. It can also be thought of as a more global or universal phenomenon. Obviously, we would like to put some bounds on this definition (for the purpose of this article) and speak only of the UHI. Thus, a UHI is the difference between the temperature in a certain urban location and that at a given reference point in a nonurban location (e.g., an upwind rural location). Obviously, the choice of urban and reference points or a collection of points can affect the magnitude and the characteristics of a reported UHI.

Keeping the urban scale in mind, there can be a surface temperature heat island and an air temperature heat island (Fig. 1). The latter can be further broken down using various scales and criteria, but some common classifications include canopy-layer heat islands and the more generic boundary-layer heat islands. The former generally occur below effective roof level (quite qualitatively) and the latter extend up to the boundaries of the urban heat plume.

Figure 1 is a conceptual diagram of a 3.5°C UHI. In reality, UHI attributes are very site specific and meteorology dependent. Figure 1 shows a hypothetical temperature-difference profile (difference from a rural reference temperature), indicated with a thick dashed line. Of course, the intensity of the UHI (0–3.5°C) is purely for illustration purposes. Note the downwind displacement (advection) of the location of peak air temperature difference (i.e., the location of maximum UHI, denoted by "max").

3. CAUSES AND EFFECTS OF URBAN HEAT ISLANDS

In the simplest situations, air temperature is dictated by energy balance at the underlying surface. The following is the basic form of a surface energy balance equation:

$$\alpha_s Q + \alpha_s q + \alpha_1 L + Q_f = \varepsilon\sigma T_o^4 + h_c(T_o - T_a) + kdT_o/dz|_{z=0} + \lambda e,$$

where α_s and α_l are the absorptivities for short- and long-wave radiation, respectively; Q and q are the direct and diffuse short-wave radiative fluxes, respectively; L is incoming long-wave flux; Q_f is anthropogenic heat flux; ε is emissivity; σ is the Stefan–Boltzmann constant; T_o and T_a are surface and air temperatures, respectively; h_c is the convective heat transfer coefficient; k is thermal conductivity; λ is latent heat of vaporization; and e is

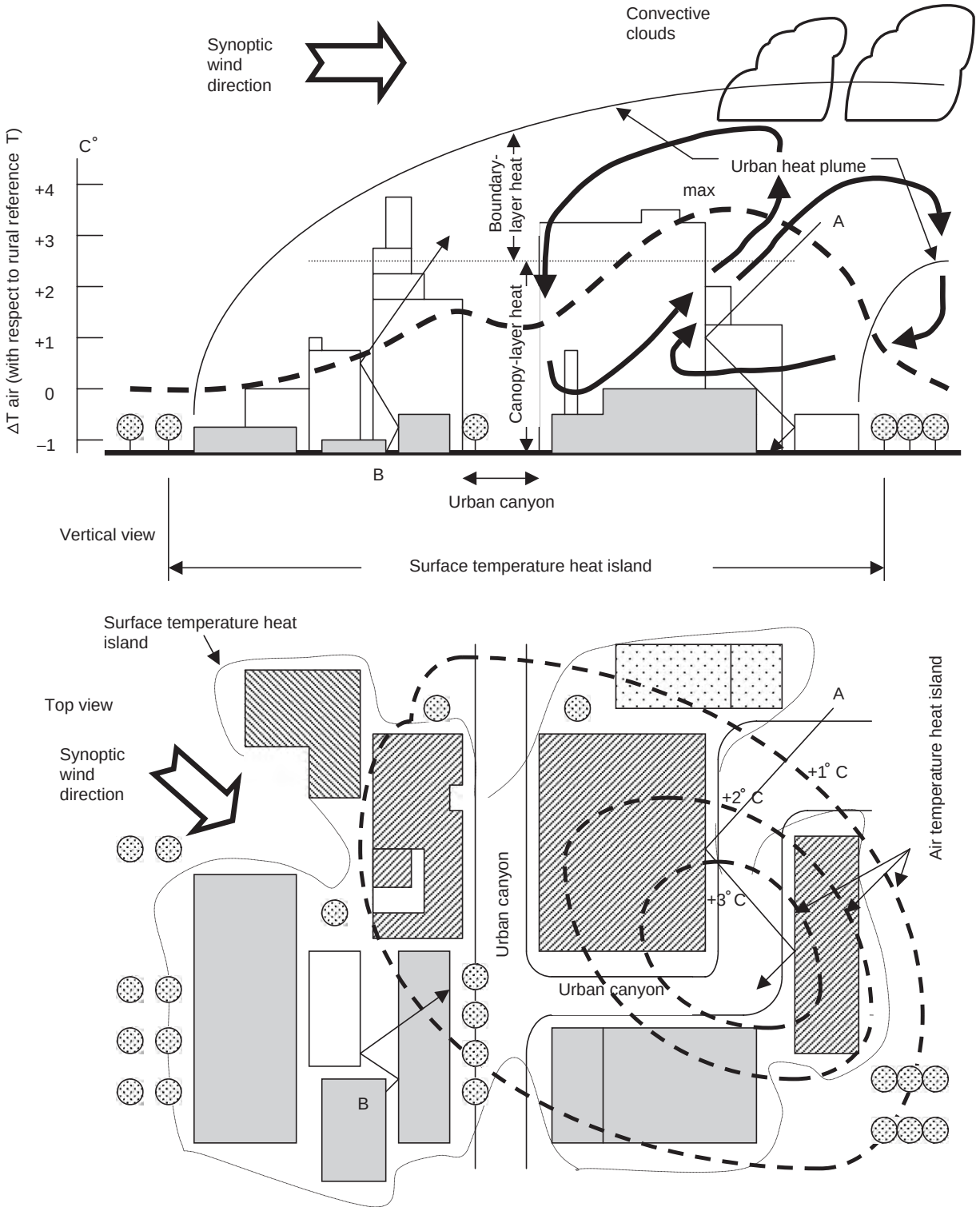


FIGURE 1 Schematic of an urban heat island: (top) vertical view (cross section); (bottom) corresponding projection top view. The thick broken lines show the air temperature UHI intensity read off the scale (°C) at left in the vertical view. The thick arrows show a UHI convective cell advected downwind of the urban center by the synoptic wind. The boundaries of an urban heat plume (also advected downwind) are shown. Other aspects shown are the boundaries of the surface temperature UHI, canopy and boundary layer UHI, convective clouds, incoming and outgoing radiative beams (A and B), and urban canyon.

evaporation rate (mass flux). Thus, the first and second terms on the left side represent the total absorbed short-wave (e.g., solar) radiation flux at the surface (both direct and diffuse), and the third term is the absorbed long-wave (e.g., near infrared) flux. The fourth term on the left is the anthropogenic heat flux (e.g., from motor vehicles, building equipment, and other heat sources) into the surface or near-surface atmospheric layer. On the right side of the equation, the first term represents long-wave radiative flux from the surface, the second term is the sensible heat flux, the third term is ground heat flux (heat conduction through the surface), and the last term is latent heat flux. Some of the terms in this equation are discussed directly or indirectly in the following sections.

3.1 Causes of UHIs

When the energy balance of an urban area is altered so that there is slower cooling relative to that in rural areas (e.g., after sunset), a classic nighttime UHI results. However, when there is significant heat storage in the urban canopy layer, or any additional sources of heat (e.g., anthropogenic), a UHI can arise at any time. Heat islands can appear as a result of a single causative factor or, more likely, a combination of several factors. The following are the main causative factors of the UHI:

Thermal capacity and urban geometry: Urban areas in general are abundant in materials that have high thermal capacity (e.g., concrete, brick, stone, pavements, and asphalt). In addition, urban regions have much more total surface area (exposed to the sun) per horizontal area than do rural regions [surface-to-area ratio (SAR)]. Rural areas tend to have a SAR close to 1. In residential neighborhoods, the SAR is on the order of 2 or 3, whereas in urban cores it is approximately 5 or higher. In cities such as Manhattan, the SAR can easily reach 10 or higher. As a result of the combination of these two factors (increased thermal capacity and increased SAR), solar radiation is captured and stored more efficiently in urban areas. The release of this extra heat, stored in all types of urban structures and materials, contributes directly to warming the air and the creation of a UHI. For example, see beams A and B in Fig. 1.

Sky view factor (SVF): The smaller SVF in urban areas, relative to that in open surrounds, helps create a UHI, especially the nighttime type. This is because radiative cooling of the urban surfaces (and canopy-layer air) to the cooler sky is hindered by the obstructions in urban areas (walls, structures, etc.). In Fig. 1, the beam at point B undergoes multiple reflections that decrease its intensity and warm the surfaces and air in contact with them. The number of reflections may increase when the urban "canyon" is surrounded by taller structures. As a result, a UHI is created due to the differential in cooling rates between the urban area and that of the rural surrounds.

Albedo and effective albedo: The albedo of building materials, pavements, and other urban structures may be lower than that of vegetation or barren land (which is abundant in rural areas). However, in some cases, the reverse is true. For example, low-altitude aircraft measurements have shown that urban albedo can reach 0.20 when the rural vegetated surroundings have an albedo of 0.15. In general, a UHI can arise when urban albedo is lower than that of the rural surrounding areas. In addition to the fact that many urban surfaces (roofs, walls, pavements, streets, structures, etc.) have low albedos, urban areas also create a lower effective albedo compared to that of rural or surrounding areas. The effect of geometry (i.e., urban canyons) is to increase multiple reflections of incident solar radiation and therefore increase the probability that a photon will be absorbed by the canyon surfaces (instead of escaping back), thus resulting in lower effective albedo. This can be seen by the simplified path of beam A in Fig. 1. After several reflections, the beam is less intense and some photons never make it to the canyon floor.

Bowen ratio: As a result of a relatively less dense vegetation cover in urban areas, the partitioning of incoming solar radiation into sensible and latent heat fluxes (among others) is altered. Thus, in urban areas, the Bowen ratio (β) is larger and this results in higher air temperatures. In rural areas or suburbs in which there is typically more vegetation, evaporation, and evaporative cooling, β is smaller, which can help keep air temperatures relatively lower. Typical values for β are approximately 4 or 5 in urban areas and approximately 0.8–1.5 in vegetative canopies. By comparison, β is approximately 0.1 for oceans and in the tropical forests it is approximately 0.2.

Anthropogenic heating: Urban areas alter the energy balance not only "passively," such as through effects of urban geometry, thermal capacity, decreased effective albedo, and lack of vegetation, but also "actively" via injection of heat directly into the air and, to a smaller extent, into the surface. This is typically referred to as furnace heat, anthropogenic heat, or man-made heat. Typical sources of

anthropogenic heat include motor vehicles, stacks and chimneys, HVAC in motor vehicles and buildings, industry, refineries and processing plants, and power plants. Because it is difficult to directly measure anthropogenic heat flux, it is usually indirectly estimated or derived from energy use profiles at sources or in areas of interest. Typical values of anthropogenic heat flux density in residential areas are on the order of 10–20 $W m^{-2}$ and in dense urban cores in the range of 50–100 $W m^{-2}$. Flux densities larger than 100 $W m^{-2}$ are thought to occur at certain times in extremely dense urban cores, such as Lower Manhattan (approximately 120–150 $W m^{-2}$). In Japan, flux densities of up to 500 $W m^{-2}$ have been reported, although this is highly uncommon. For comparison, the maximum incoming solar radiation flux density at street level on a cloudless summer day can range between 800 and 1000 $W m^{-2}$ around solar noon. Outside of the atmosphere, the incoming solar radiation flux density (solar constant) is approximately 1350 $W m^{-2}$.

Of course, a UHI can be caused by any one of the previously discussed causative factors or, most commonly, by a varying degree of combinations of such. As a result, the characters of UHIs and their extents can be very diverse and different from one location or time to another.

3.2 Effects of UHIs

UHIs can alter the meteorology, air pollutant emission rates, pollutant transport, and photochemistry in the area in which they exist. Of course, they also affect energy use. Briefly, the nonenergy effects of UHIs are

Creation of a wind pattern or circulation, which is typically a convective cell located in the UHI vicinity: Depending on the strength of the background winds (e.g., synoptic) relative to that of the UHI winds, the convective cell can be asymmetrical and be advected downwind (see, for example, the thick arrows in Fig. 1). In general, the larger and denser a city (larger population), the stronger its UHI circulation.

Increased cloudiness and rainfall as a result of enhanced convective activity within the heat plume associated with UHIs (see hypothetical clouds in Fig. 1, downwind of the UHI maximum and near the location of divergence of the UHI circulation): Of course, UHIs produce increased convective clouds or precipitation only when the right conditions are met (e.g., temperature, winds, moisture, and condensation nuclei). Long-term observations suggest that UHIs can cause up to an 8% increase in cloudiness, 14% more rainfall, and 15% more thunderstorms near urban areas compared to their surroundings.

Increased ambient temperatures, the most obvious effect of UHIs: On average, typical UHIs are on the order of 2°C, but extremes as high as 8°C have been recorded. Figure 1 shows a hypothetical UHI of 3.5°C downwind of the highest urban concentration. It is obvious that the choice of urban and rural reference points can dictate the "observed" UHI intensity.

Increased mixing of pollutants in the boundary layer (BL): Depending on stability, depth of the BL, and convective activity and strength, pollutants emitted into the urban atmosphere can be well mixed and dispersed within the BL, resulting in lower apparent concentrations compared to a situation in which the BL is shallower and pollutant mixing is inhibited.

Increased temperature-dependent emissions of air pollutants: For example, the emission of biogenic hydrocarbons (e.g., isoprene from vegetation) is highly temperature dependent. It can double or triple for every 6–10°C increase in air temperature in the range of 25–37°C. In addition, most fugitive emissions of hydrocarbons (e.g., from fuel storage tanks, on-road motor vehicles, motor engines hot-soak, and evaporative losses) are temperature dependent. Direct emissions of particulate matter (e.g., of interest PM_{10} and $PM_{2.5}$) are also temperature dependent to a certain extent. Increased precursor emissions generally lead to an increased rate of smog formation, decreased visibility, and poorer air quality.

Photochemical production of smog generally positively correlated with temperature: This is true at least for some of the most influential photochemical reactions. Thus, a UHI can act as a catalyst for accelerating the formation of tropospheric/urban ozone, which poses a significant health problem in major urban areas. Because higher temperatures can exacerbate ozone formation while at the same time increase the mixing of pollutants in the BL, there may appear to be a competing effect. However, most data and observations suggest that the net effect of a UHI in terms of ozone concentrations is a negative impact on air quality.

4. TYPICAL MAGNITUDES OF UHIs

Because UHIs form as a result of several factors that differ in space, time, location, geography, and with combinations of such, it is obvious that their

intensities differ not only in absolute terms but also in location, time, duration, and an array of spatio-temporal characteristics. For example, some UHIs peak at night (most common), others during the day, and others have multiple diurnal peaks. In infrequent cases, a UHI can also be roughly constant during a 24-h period. In addition to hourly and subhourly fluctuation in intensity, UHIs also have diurnal and seasonal fluctuation components.

Thus, it is difficult to generalize and characterize UHI intensities with one or two indicators. However, there may be some observed upper bounds and common (frequently observed) intensities. The upper bounds seem to be related to city size (population) in some ways. Urban heat islands of 1 or 2°C are typical of the North American continent (the United States and Canada), where they can peak up to 4°C or more. Observations and numerical simulations of a number of U.S. metropolitan areas suggest that UHIs on the order of 2°C are typical in Los Angeles, Chicago, Atlanta, Washington, DC, Philadelphia, and New York. UHIs of 1–1.5°C are found in Houston, Dallas, Phoenix, and Miami. In Japan and China, UHIs on the order of 3°C have been frequently reported. Larger UHIs have been reported in Brazil (as high as 8°C in Sao Paolo) and Britain (as high as 6 or 7°C in London). In Mexico, similar UHIs have been observed. As discussed later, a UHI of 1 or 2°C is a significant factor in energy use.

5. IMPACTS OF UHIs ON ENERGY

The impacts of UHI on energy can be seen on regional, urban, and building scales. In terms of energy supply, conversion, and use, these scales roughly correspond to utility service territory and power plants, city-scale grids, and building scales. Of course, urban and building energy use are not the only contributors to utility-scale loads, but most often they are a major component. As previously mentioned, a UHI can cause several meteorological perturbations, but perhaps the most relevant from an energy perspective is that of increased ambient air temperature. Temperature can also be thought of as a surrogate (implicit indicator) for other local weather parameters, such as available sunshine, cloudiness, precipitation, and winds.

The typical energy impact of UHIs is to increase cooling loads in summer and decrease heating loads in winter. The net effect, of course, depends on the characteristics of a region's climate (summer dominated, winter dominated, both, or none) and general meteorological features (e.g., wind patterns, available sunshine, precipitation, and topographically induced flows and phenomena). It also depends on local energy-specific aspects, such as population density, dominant sectors of energy use, types of buildings, age, distribution, general building envelope characteristics, saturation of HVAC equipment, and local cost of fuel and electricity rates. However, in general, a UHI has relatively larger impacts on (increasing) cooling loads than on (decreasing) heating loads. Also, the former is an unwanted effect, whereas the latter is generally a welcomed one. Thus, the remainder of this section focuses more on the cooling energy impacts of UHIs than on their heating energy implications.

A first estimate of the effects of UHIs on energy (e.g., heating and cooling loads) would involve quantifying their impacts on heating and cooling degree-days (HDD and CDD, respectively). If a certain reference temperature is selected [e.g., 18.3°C (65°F)], we find that most urban areas in the United States see a typical increase in CDD of 15–35% as a result of UHIs, with some extremes as high as 90%. A typical decrease in HDD is approximately 5–14%, with extremes as high as approximately 30%. Table I shows the impact of UHIs on HDD and CDD in selected regions.

However, one CDD is not equivalent to one HDD when converted to energy costs, for example, if cooling is done with electricity and heating mainly with natural gas. Thus, in addition to the fact that electricity and gas costs are different, the HDD/CDD weight also depends on how cooling and heating are achieved locally (i.e., HVAC systems, primary and secondary systems, power plants, and overall system

TABLE I

Impacts of UHI on Annual HDD and CDD in Selected Cities (Base 18.3°C)[a]

City	Impact of UHI on HDD (%)	Impact of UHI on CDD (%)
Los Angeles	−30	+90
Washington, DC	−5	+20
St. Louis	−5	+10
New York	−5	+25
Baltimore	−15	+35
Seattle	−15	+55
Detroit	−5	+15
Chicago	−5	+25
Denver	−10	+20

[a] Rounded to the nearest 5%.

efficiencies). Thus, the conversion from HDD/CDD to energy use and costs depends on local factors and energy rates. In the United States, the average electricity cost is approximately \$0.08/kWh and for gas approximately \$0.65/therm (1 therm = 10^5 Btu), but deviations from these averages (by geographical location and time/season) can be quite significant. Thus, the HDD/CDD data given in Table I are for qualitative comparison purposes.

At the building scale, the energy impacts of UHIs are rather straightforward to understand, even though building characteristics, thermal integrity, occupancy schedules, and thermal comfort preferences may be complicating factors. In general, residential or small nonresidential buildings are significantly affected by the UHI. In larger buildings, in which internal loads (heat gain from occupants, lighting, machinery, and equipment) are larger than the heat gain or loss through the envelope, the sensitivity to air temperature (UHI) is relatively smaller. Table II gives some orders of magnitudes for expected peak cooling electric demand change with temperature in North America. These ranges are for illustration purposes and not exact numbers.

Thus, for example, the effect of a 3°C UHI would be to increase peak demand for cooling on a summer afternoon by approximately 5% of the daily average demand in large office buildings (relative to whole-building electricity consumption). For small offices, the increase would correspond to approximately 10% and for residential buildings to approximately 15%.

The impacts of UHIs on building energy use have also been studied and analyzed with numerical simulations in an attempt to isolate the effect of UHIs from other factors. Computer modeling is needed in such sensitivity studies since it is nearly impossible in the real world to find exactly identical buildings some within and others outside of a UHI with everything else similar so that the impacts of UHIs can be isolated. Advanced computer simulations show that in terms of system cooling peak, UHIs can cause increases in the range of 5–40% depending on weather, location, and building type. In terms of heating peaks, the UHIs can cause a decrease on the order of 4 or 5%. The simulations also show that in some regions and climates, the effect of UHIs is small or negligible. In addition to HVAC-related energy use, there are other aspects in buildings that are indirectly related to UHIs, such as indoor thermal comfort, indoor environmental quality, natural ventilation, and the performance of passive energy systems.

TABLE II

Magnitudes for Typical Temperature Sensitivity of Peak Electric Demand for Cooling at the Building Scale

	Total floor area ($\times 1000\,m^2$)	$kW°C^{-1}$	Total electricity usage ($kWh\,m^{-2}\,year^{-1}$)
Large office buildings	50–100	50–110	100–130
Small commercial buildings	1–1.8	1.2–3	70–120
Residential buildings	0.1–0.4	0.2–0.35	50–100

At the urban scale, data suggest a sensitivity in the peak electric demand of 2–5%/°C in the United States. Values of 2.5–3%/°C are common in cities such as Los Angeles, Chicago, Washington, DC, and New York. Values of approximately 5%/°C are found in cities with relatively hotter climates, such as Atlanta, Houston, Dallas, Phoenix, and Miami. The increase in electric demand occurs when the daily maximum temperature is above a certain threshold (e.g., 20°C in summer). In terms of absolute values, these percentages translate into a range of effects. In areas such as the Los Angeles basin, this corresponds to approximately 130 MW/°C, whereas in hotter regions, such as California's Central Valley or the state of Texas, this is equivalent to approximately 300 MW/°C. Thus, depending on the UHI and the regional climate, the actual increase in demand can be significant. On the other hand, in some cold-weather regions, the sensitivity of the electric demand may be negative, indicating savings (rather than penalties) with increased air temperatures as a result of decreased electricity use for heating. This occurs in the northeastern United States and along the Canadian border.

For the purpose of this discussion, it is possible to identify four climate groups in the United States that correspond to certain ranges in the sensitivity of the peak electric demand to temperature. Such climate types are cold, moderate, warm, and hot, as shown in Table III.

According to various numerical and observational studies, UHIs are typically in the range of 1–3°C in North America. Thus, in Los Angeles, Chicago, Washington, DC, New York, and Phoenix, approximately 5% of the peak electric demand is a result of UHIs. In Houston and Dallas, this is approximately

TABLE III

Sensitivities of Peak Electric Demand by Climate Types in the United States

Climate	Sensitivity of peak (%/°C)	States
Cold	−1	Maine, Vermont, Montana
Moderate	1	Massachusetts, Virginia, the Dakotas, Michigan
Warm	2–3	Alabama, Tennessee, California
Hot	3–5	Florida, Louisiana, Texas, Arizona

8%, in Atlanta it is approximately 11%, whereas in Philadelphia it is approximately 3%. In general, it is estimated that 5–10% of the urban peak electric demand in summer-dominated large U.S. urban areas is a result of the UHI effect. Of course, the higher temperatures associated with a UHI not only increase the demand for cooling but also decrease the efficiency of the air-conditioning/cooling systems in meeting the loads.

Finally, UHIs also indirectly affect other aspects of energy use. Warmer conditions, for example, induce an increase in watering needs (e.g., for urban vegetation, nearby crops, and parks) and thus an increased need for pumping water and distributing it. Higher air temperatures also cause an increased need for air-conditioning use in motor vehicles. The additional air-conditioner load requires more energy (increased fuel consumption). For motor vehicle air-conditioner compressors, the "on time" fraction increases by approximately 5%/°C (i.e., the compressor is on 5% longer for every 1°C increase in ambient air temperature above a certain threshold). Another aspect of UHI is health and thermal comfort, which entail indirect use of energy as well. Heat waves, for example, can be exacerbated with UHIs and thus heat-related hospital admissions increase and health care becomes a serious concern during the summertime. Although difficult to quantify, these aspects involve increased energy usage in health facilities.

6. MITIGATION OF UHI EFFECTS TO SAVE ENERGY

Just as the UHI imparts effects on energy at various scales (e.g., building and urban scales), its mitigation (reversing or compensating for its effects) can also be addressed at such scales.

Thus, at the building scale, the goal is to cool the building envelope or cool the ambient air in the immediate vicinity or both. More accurately, the goal is to prevent the envelope and ambient air from heating up as much as they would without UHI control. The current thinking is that cooling the building envelope should be done using high-albedo (reflective) materials so that the amount of solar radiative flux absorbed, and hence the heat conducted through the structure, is reduced. Typical building materials (e.g., roofing materials) have albedos in the range of 0.08–0.25. For example, the majority of asphalt shingles typically have an albedo of 0.10–0.15. Roofing membranes, such as black single-ply roofing, have a typical albedo of 0.06 and gravel roofs between 0.12 and 0.34, depending on the color of the gravel, but most tend to be approximately 0.15.

A hypothetical example of conventional roofing materials with an albedo of 0.25 is shown by the lower (thick) line in Fig. 2. Roof albedos can be increased rather easily with high-albedo (reflective) materials or coatings so that the new albedos attain the range of 0.4–0.6 after accounting for effects of weathering, aging, and soiling. Some reflective materials have an initial albedo of up to 0.85 or higher but require maintenance to keep it at that high value. A hypothetical example of a high-albedo, light-colored material (with an albedo of 0.70) is shown by the upper (thin) line in Fig. 2. However, high-albedo materials need not be lightly colored or white. For example, a hypothetical material shown by the middle line (thick dashed) in Fig. 2 has an albedo of 0.5, but it will appear dark to the human eye because of its low reflectivity in the visible range.

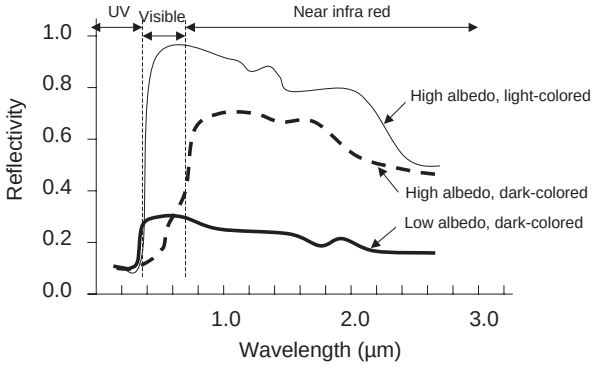


FIGURE 2 Hypothetical spectral reflectivities (albedo) of various materials. These hypothetical materials have albedos of 0.25, 0.50, and 0.70, respectively, from bottom to top.

Materials for increasing roof albedo exist in a variety of colors and textures. In the United States, their incremental cost over traditional roofing materials is \$0–2/m^2. Incremental cost is defined here as the extra costs incurred as a result of selecting a reflective roof instead of a traditional one. Of course, if the selection is done at the time of maintenance reroofing or during the initial construction phases, there may be no incremental costs. The implementation of high-albedo roofs can offset or more than offset the local effects of UHIs on energy use in buildings. Roof surfaces with higher albedo are typically 5–15°C (up to 30°C) cooler than their low-albedo counterparts. Reflective roofs warm up at approximately one-third the rate of more conventional roofs. Thus, whereas conventional roofs can be as much as 40–45°C warmer than the overlying air, high-albedo roofs are typically 10–15°C warmer than the air, assuming similar underlying construction and materials. During certain times, the surface temperature of reflective roofs can be very close to that of ambient air.

To cool the ambient air around a building, the current thinking is to increase vegetative cover so as to promote evaporative cooling. The tree canopy also has the beneficial effect of shading the building structure, thus significantly decreasing solar heat gain by the envelope. Another energy benefit of increased urban vegetation, particularly of the evergreen type, is wind shielding. Although this is only indirectly related to heat islands, it has a significant energy effect. In winter, for example, trees shield buildings from the cold winds, reducing the infiltration of colder air to the indoor spaces and thus helping to reduce heating energy needs, especially at higher latitudes. Of course, this effect is important only for one or two-storied buildings, for which trees can shield the wind from the structures. Upper floors in taller buildings do not benefit from ground-level tree canopies. Urban vegetation has other benefits, such as improving air quality by trapping and depositing pollutants from the air. It also has aesthetics values and qualities.

At the urban scale, the main concept is to implement high-albedo materials not only on buildings (e.g., roofs) but also on pavements, roadways, parking lots, sidewalks, and other built-up surfaces. The strategy would also include a reforestation program in which a large number of trees are planted throughout urban areas. The effects of such large-scale implementation of urban albedo and reforestation programs would be to reduce the urban heat island, completely offset it, or sometimes more than offset it. Simulations show that regional air temperature can be decreased by 4°C or more at certain locations and times as a result of increased urban albedo. These modeling studies also show that increasing vegetative canopy cover by 30% can potentially decrease air temperature by 3°C or more. However, a reduction in air temperature in urban areas of 1 or 2°C is more typical according to mesoscale meteorological modeling work. Thus, in most cities discussed previously, the UHI can be offset by large-scale implementation of high-albedo materials and urban forest. In terms of utility-scale impacts of such strategies, a decrease of 2°C in the UHI would be equivalent to savings of 4–10% from the afternoon peak electric demand in summer.

Extrapolation of numerical simulation results to regional and national scales suggests that the savings from implementing an aggressive reflective roof and pavement program can be significant, even after accounting for the wintertime heating energy penalties that can inadvertently result from increased reflectivity to incoming solar radiation. Typically, however, the heating penalties are smaller than the cooling benefits because in wintertime, solar radiation flux density is smaller (lower sun altitude angle), cloud cover is larger, overcast events are more frequent, and snow and rain can cover building and built-up surfaces (e.g., pavements and parking lots). Thus, the impacts of increased albedo is minimal under such conditions.

Table IV provides some simulation results obtained at the Lawrence Berkeley National Laboratory. The first row gives estimated net energy savings (in millions of dollars per year) from high-albedo roofs for each city. These results were based on the assumptions that residential reflective roofs have an albedo of 0.5 and that nonresidential reflective roofs have an albedo of 0.7 (both having an initial albedo of 0.2).

Table IV also gives the CDD and HDD (base 18.3°C) to show the climate traits for each region. Obviously, the regionwide savings also relate to city size and urban population density, not just to climate type. In addition, savings in cooling electricity and increases in natural gas needs for heating are given per 100 m^2 per year for these selected cities. The savings and penalties are given for both residential and nonresidential buildings and the net effect, integrated over the region, is shown. Extrapolating these results to nationwide scale, the net savings from an aggressive high-albedo roofs program in the United States is estimated to be on the order of \$700 million per year.

TABLE IV

Energy Savings from Complete Implementation of High-Albedo Roofs

	City							
	Los Angeles	Phoenix	Dallas	New Orleans	Atlanta	Miami	WDC/Baltimore	New York
\$M year^{-1} (net savings)	35	37	20	9	9	20	8	16
CDD, base 18.3°C (65°F)	940	3815	2415	2620	4920	4130	1045	1000
HDD, base 18.3°C (65°F)	1300	1155	2305	1680	3020	140	5235	5090
kWh year^{-1} 100 m^{-2}, residential	−182	−314	−166	−199	−153	−259	−137	−104
Therms year^{-1} 100 m^{-2}, residential	+4	+1	+3	+3	+4	0	+6	+10
kWh year^{-1} 100 m^{-2}, nonresidential	−350	−409	−224	−287	−239	−340	−221	−211
Therms year^{-1} 100 m^{-2}, nonresidential	+3	+2	+4	+2	+6	0	+9	+9

The cost of conserved energy (CCE) has been calculated for a range of building types, roof types, insulation levels, roof life spans, various levels of increased albedo, and a range of weather conditions. The target, of course, is to keep the CCE under \$0.08/kWh, which is the U.S. national average cost for electricity. To achieve this target, the incremental cost for reflective roofs should be kept under \$0.6 m^{-2} if it is assumed that the life spans of reflective and nonreflective roofs are equal. This is not true, however, since the life span of reflective roofs can be double that of nonreflective roofs, and if the cost of replacement for nonreflective roofs is factored in, then the cutoff incremental cost for reflective roofs can be much larger (to keep CCE under \$0.08/kWh). This is quite significant since other benefits of reflective roofs are not factored in (e.g., air quality and health benefits).

Increased vegetative cover cools the ambient air within and under a moist, evaporating canopy by up to 3°C. For example, parks are typically 0.5–1.5°C cooler than surrounding areas in higher latitudes. Observations in Canada show that urban parks can be 2.5°C cooler than their immediate built-up surroundings. In relatively lower latitudes (e.g., the southwestern United States), they have been observed to be 1–3°C cooler. Measurements in California's Central Valley show that orchards can be up to 6°C cooler than surrounding nonvegetated land but are typically only 2°C cooler. Studies in Japan show urban parks to be 1.5°C cooler than their surroundings in the summertime. However, it has also been shown that trees can sometimes increase air temperature within or above the canopy, especially if they are drier. The increase can be on the order of 1°C. Thus, the net effects of urban reforestation can be more complex to quantify than the effects of increased urban albedo, which are relatively more straightforward to understand.

For vegetation, the costs vary widely depending on tree species, age, size, planting location, etc. In the United States, trees can cost less than \$5 when small and young. However, some 5-m trees cost \$500 or more if purchased individually. Labor (planting) and maintenance (watering, pruning, and trimming) costs should also be factored in for an accurate analysis.

SEE ALSO THE FOLLOWING ARTICLES

City Planning and Energy Use • Ecological Footprints and Energy • Economic Growth and Energy • Land Requirements of Energy Systems • Population Growth and Energy • Suburbanization and Energy • Urbanization and Energy

Further Reading

American Society of Heating, Refrigerating and Air-Conditioning Engineers (1998). Energy savings of reflective roofs. *ASHRAE Tech. Data Bull.* **14**(2).

Arya, S. P. (1988). "Introduction to Micrometeorology." Academic Press, New York.

Department of Energy (1990). Energy and climate change, Report of the DOE Multi-Laboratory Climate Change Committee. Lewis, Chelsea, MI.

Landsberg, H. (1981). "The Urban Climate." Academic Press, New York.

Linder, K. P., and Inglis, M. R. (1989). The potential effects of climate change on regional and national demands for electricity. *In* "The Potential Effects of Global Climate Change on the United States," Report No. EPA-230–05–89–058, Appendix H, pp. 1–38. Environmental Protection Agency, Washington, DC.

Oke, T. R. (1987). "Boundary-Layer Climates." Methuen, London.

Pomerantz, M., Akbari, H., Berdahl, P., Konopakci, S., and Taha, H. (1999). Reflective surfaces for cooler buildings and cities. *Philos. Magazine B* **79**(9), 1457–1476.

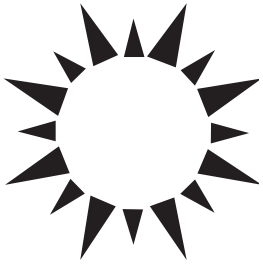
Taha, H. (ed.). (1997a). *Energy Buildings Special Issue Urban Heat Islands* **25**(2).

Taha, H. (1997b). Modeling the impacts of large-scale albedo changes on ozone air quality in the south coast air basin. *Atmos. Environ.* **31**(11), 667–1676.

Taha, H., Konopacki, S., and Gabersek, S. (1999). Impacts of large-scale surface modifications on meteorological conditions and energy use: A 10-region modeling study. *Theor. Appl. Climatol.* **62,** 175–185.

Heat Transfer

FABIO GORI
University of Rome "Tor Vergata"
Rome, Italy



1. **Historical Survey**
2. **Modality of Heat Transfer**
3. **Heat Conduction**
4. **Thermal Radiation**
5. **Heat Transfer by Convection**
6. **Dimensional Analysis**

Glossary

blackbody A body absorbing all the thermal radiation impinging on it.
boundary The geometric contour dividing the thermodynamic system from the exterior.
conservation equations Equations expressing the balance of a prescribed quantity (e.g., mass, momentum, and energy).
convection Heat transfer at the boundary between a solid wall and a moving fluid.
exterior The part of space external to the thermodynamic system.
first law of thermodynamics The physical principle relating heat transfer to mechanical work.
Fourier equation Basic constitutive equation of heat conduction relating heat transfer to temperature gradient.
heat conduction Heat transfer in a substance, solid, liquid, or gas without macroscopic movements of the matter (not involving electromagnetic waves).
heat transfer Energy exchange across the boundary of a thermodynamic system due to a temperature difference.
Newton equation Basic constitutive equation of a fluid relating shear stress to velocity gradient.
non-Fourier equation Constitutive equation of heat conduction for special phenomena (e.g., low temperatures).
non-Newtonian equation Constitutive equation for shear stress of special fluids.
scale of temperature In the International System, the unit of measurement of the temperature is the degree Kelvin (K), which is 1/273.16 of the temperature of the triple point of water, and the dimension is $[\theta]$.
second law of thermodynamics The physical principle establishing a limit in the conversion of heat into work.
temperature The fundamental quantity measured by a thermometer.
thermal radiation Heat transfer connected to electromagnetic waves.
thermodynamic system Part of space under investigation.

Heat transfer is the energy interaction between a thermodynamic system and the exterior, due to a temperature difference, across the boundary. Heat transfer is measured, according to the first law of thermodynamics, as the work exchanged by a closed system in a cyclic process. The unit of heat transfer is Joule (J) (i.e., equal to the work unit), and the dimensions are $[M L^2 T^{-2}]$. No instrument can directly measure heat transfer, and only the thermometer can measure temperature.

1. HISTORICAL SURVEY

Experiments on heat transfer were done in ancient times and several kinds of heat machines have been invented. The first proposal to use a thermoscope to measure temperature is attributed to Galileo Galilei. The Galileo scientific heritage was continued by the Cimento Academy, founded in Florence in 1657 by the Granduke Ferdinand II and Prince Leopold of Medici. The Academy promoted experiments that allowed modern science to make the jump from the heat transfer empiricism of the past to the modern heat transfer scientific design. The most significant scientific contributions of the Cimento Academy, in its 10 years of existence, were the proposal to use several instruments to measure temperature and experiments in barometry and vacuum, which allowed other physical quantities to be defined, such as pressure, which plays a significant role in heat transfer.

 © 2004 Elsevier Inc. All rights reserved.

2. MODALITY OF HEAT TRANSFER

There are three modalities of heat transfer: heat conduction, thermal radiation, and convection. Heat conduction is the kind of heat transfer, associated with the internal energy of matter, that occurs within a substance (gas, liquid, or solid) without macroscopic movement of its parts. No heat conduction is present in vacuum. Thermal radiation is the kind of heat transfer that is present also in vacuum. Convection is a combined type of heat transfer, occurring in a fluid (liquid or gas), which includes heat conduction and fluid motion.

2.1 Heat Conduction

This kind of heat transfer is present within a substance and is associated with the energy of the molecules, atoms, and their components. Heat conduction is related to the microscopic behavior of the matter. On the other hand, the basic, or constitutive, equation of heat conduction is related to the temperature gradient. The fundamental law of heat conduction must be consistent with the second law of thermodynamics—that is, heat must be exchanged between parts of a body or bodies at different temperatures. In contrast, no heat transfer is present between parts at the same temperature. The first law of thermodynamics, or the energy conservation equation, states that energy is conserved in the absence of heat sources and sinks. In general, the temperature at a certain point on the matter is a function of the space coordinate and time. An isothermal surface is a geometric surface on which temperature is the same throughout. If the temperature along this surface is constant in time, the isotherm can be said to be in steady state. Assume two isothermal surfaces with a temperature difference equal to dT distant in space dn, where dn is measured perpendicularly to the isothermal surface. The heat conduction per unit time and area, Q/A, has been proposed by J. B. Fourier to be proportional to a macroscopic physical property of the substance k according to the equation

$$q = Q/A = -k\ dT/dn = -k\nabla T. \tag{1}$$

The property k, called thermal conductivity of the substance, is defined by Eq. (1) and has been found to be a physical property (i.e., it is a function of temperature and pressure). The units of k are W/m K and the dimensions are $[M\ L\ T^{-2}\ \theta^{-1}]$. The thermal conductivity can be related to the microscopic behavior of the substance. The coupling of the Fourier equation, also called the constitutive equation for heat conduction, with the first law of thermodynamics, or the energy conservation equation, allows one to write a differential equation, in a three-dimensional rectangular domain, in unsteady state as

$$\rho c\ \partial T/\partial t = \partial/\partial x(k\ \partial T/\partial x) + \partial/\partial y\ (k\ \partial T/\partial y) + \partial/\partial z\ (k\ \partial T/\partial z) + q_g, \tag{2}$$

where q_g is the heat generated per unit time and volume, $k = f(x, y, z, t)$, $\rho = f(x, y, z, t)$, and $c = f(x, y, z, t)$. Equation (2) is valid for isotropic, heterogeneous media. The assumption of constant thermal conductivity allows Eq. (2) to be written as

$$\begin{aligned}&\partial T/\partial t\\ &= k(\rho c)\ (\partial^2 T/\partial x^2 + \partial^2 T/\partial\, y^2 + \partial^2 T/\partial\, z^2) + q_g/(\rho c)\\ &= \alpha\nabla^2 T + q_g/(\rho c).\end{aligned} \tag{3}$$

The new physical property of the substance, $\alpha = k/\rho c$, called thermal diffusivity, has been introduced because of the unsteady state; it has the units (m^2/s) and the dimensions $[L^2\ T^{-1}]$. Equation (3) has a parabolic form, which implies an infinite velocity of propagation of the effect or, in other words, that the effects of a thermal disturbance propagate instantaneously to an infinite distance from the source. Equation (3) has been verified for phenomena at ambient and high temperatures, and experimental results have been found to be in agreement with the theory. For very low-temperature phenomena, Eq. (1) has been modified and the following is one of the proposals:

$$\alpha/C^2\ \partial q/\partial t + q = -k\nabla T, \tag{4}$$

and the corresponding conservation equation becomes

$$1/C^2\partial^2 T/\partial^2 t + 1/\alpha\ \partial T/\partial t = \nabla^2 T, \tag{5}$$

where C is the velocity of propagation of heat transfer, given by $C^2 = \alpha/t_o$, and t_o is a relaxation time. The heat transfer with finite propagation velocity has been confirmed experimentally at liquid helium II temperatures. A similar effect of the liquid helium II has been found in materials of great purity and at low temperatures. Equation (5) is a hyperbolic equation.

2.2 Thermal Radiation

Thermal radiation is energy transfer in the form of electromagnetic waves. The microscopic mechanism can be related to the energy transport by photons

released from molecules and atoms. The physical parameters that describe thermal radiation are the photon or wave velocity, c, the wavelength, λ, and the frequency, ν. The photon energy is given by the relation

$$E = h\nu, \tag{6}$$

where h is the Planck constant, $h = 6.6256 \times 10^{-34}$ Js. Frequency and wavelength are related by

$$\lambda\nu = c. \tag{7}$$

The thermal radiation velocity in vacuum is equal to $c_0 = 2.997925 \times 10^8$ m/s. The wave velocity c in a medium is connected to c_0 by the relation

$$n_0 = c_0/c, \tag{8}$$

where n_0 is the refraction index of the medium traveled by the electromagnetic waves. Electromagnetic waves are classified by the wavelength λ, which is usually measured on the scale of 10^{-6} m $= 1\,\mu$m. Visible light is in the range 0.4–0.7 μm. The wavelengths generated by heated bodies are in the range 0.3–10 μm and are the thermal radiations of common interest. Radiation with a wavelength larger than the visible one is called infrared, whereas that with a wavelength smaller than the visible is called ultraviolet. In all energy applications except cryogenics, the characteristic dimensions of the system are large compared to the wavelengths of the thermal radiation. In cryogenics, because of the low temperatures, the radiation wavelengths are large. In the following, it is assumed that the energy transfer by radiation occurs along straight lines, excluding scattering or refraction. The radiant energy flux per unit time, $d\Phi$, can be evaluated for an area element dA along a direction with an angle β toward the surface normal. The radiant energy flux $d\Phi$ contained in a solid angle $d\omega$ within the frequency range $d\nu$ is given by

$$d\Phi = K_\nu \cos\beta\, dA\, d\omega\, d\nu = K_\lambda \cos\beta\, dA\, d\omega\, d\lambda, \tag{9}$$

where K_ν and K_λ are the monochromatic intensities of the radiant flux. In a nonemitting and nonabsorbing medium the intensity K_ν is constant along a ray, whereas it varies if it emits or absorbs radiation, and the radiant energy flux increase $d^2\Phi_e$, which is an order of magnitude smaller than $d\Phi$, is

$$d^2\Phi_e = j_\nu\, \rho dV\, d\omega\, d\nu, \tag{10}$$

where j_ν is the energy emitted per unit time and unit mass into a unit solid angle and within a unit frequency range. The energy flux absorbed along a path length ds is

$$d^2\Phi_a = \chi_\nu\, \rho ds\, d\Phi, \tag{11}$$

where χ_ν is the coefficient of absorption. The decrease in the radiant energy flux due to scattering is

$$d^2\Phi_s = -\sigma_\nu\, \rho ds\, d\Phi, \tag{12}$$

where σ_ν is the coefficient of scattering. Thermal radiation is present when there is local thermodynamic equilibrium in a medium. Application of the laws of thermodynamics to media in thermodynamic equilibrium allows the following conclusions:

The monochromatic intensity emitted, $K_{\nu e}$, is given by $K_{\nu e} = j_\nu/\chi_\nu$.

A radiation beam, traveling toward the interface between two media, with an angle β toward the surface normal is partly reflected with the same angle, with a ratio given by the reflectance or reflectivity, ρ_ν, and it partly penetrates, with an angle β' according to the law $\sin\beta'/\sin\beta = c'/c$.

The assumption of constant velocities, c' and c, allows one to find the final relation: $j_\nu c^2/\chi_\nu = j'_\nu c'^2/\chi'_\nu =$ constant, or a universal function, independent of the medium.

The absorbance of a medium, α_ν, is the ratio between the radiant energy flux absorbed and that approaching the interface.

The intensity of the radiation emitted by a medium to another medium, $i_{\nu e}$, is found in relation to the previous parameters as $i_{\nu e} = \alpha_\nu K_{\nu e}$.

A medium with $\alpha_\nu = 1$ is a blackbody, and no radiation impinging on that body is reflected or transmitted but is only absorbed. For a blackbody, the intensity $i_{\nu b}$ becomes $i_{\nu b} = K_{\nu e}$.

One of the conclusions of the electromagnetic theory of Maxwell was that the radiation reflected from an interface between two media exerts a pressure on that surface. The radiation pressure, which the solar radiation exerts on the surface of Earth, is very small (4×10^{-6} N/m^2). An important application of radiation pressure is associated with space flight, for which it has been proposed as a means of propelling space vehicles. The monochromatic intensity has been derived by Planck from the quantum theory as

$$i_{\nu b} = 2h\nu^3/(c_0^2(\exp(h\nu/\mathbf{k}T) - 1)), \tag{13}$$

where $k = 1.38054 \times 10^{-23}$ J K, the Boltzmann constant is universal. From Eq. (7), written in vacuum, it is found that

$$i_{\nu b} = 2C_1/[\lambda^5(\exp(C_2/\mathbf{k}T) - 1)], \tag{14}$$

where $C_1 = 0.59548 \times 10^{-16}$ W m^2 and $C_2 = 1.43879$ cm K. The maximum of $i_{\nu b}$ occurs according to the relation

$$\lambda_{\max} T = C_3, \tag{15}$$

where $C_3 = 0.28978$ cm K. The total heat radiated by a blackbody, into a unit solid angle, in the wavelength range from 0 to ∞ is

$$i_b = \int_0^\infty i_{\lambda b} d\lambda = \sigma T^4/\pi. \tag{16}$$

The emissive power e_b is found by integration over the whole hemispherical space (i.e., the solid angle ω),

$$e_b = i_b \int \cos\beta d\omega = \pi i_b = \sigma T^4, \tag{17}$$

where $\sigma = 5.6697 \times 10^{-8}\,\mathrm{W/m^2\,K^4}$. The previous equation is a good approximation of the behavior of the radiation emitted from a blackbody into a gas.

2.3 Heat Transfer by Convection

Convection occurs in a fluid during its flow at the boundary with a solid wall. The temperature of the fluid distant from the wall can be assumed to be constant and the gradient of temperature is mainly concentrated in a relatively narrow layer, δ, of fluid, close to the wall. A thin film of fluid adheres to the wall and temperature variation takes place across the layer δ. The temperature is constant outside the layer δ due to mixing of the fluid. In the layer δ of fluid, the heat is transferred by pure conduction as in a solid body. In the oversimplified case of a linear temperature variation across the layer, it is possible to write

$$Q = k_f A(T_f - T_w)/\delta, \tag{18}$$

where k_f is the thermal conductivity of the fluid, A is the surface involved in the heat transfer, T_f is the fluid temperature, and T_w is the wall temperature. The layer δ depends on the velocity of the fluid, the shape and the structure of the surface, and other factors. In the practical application, the ratio k_f/δ is substituted by a new coefficient h, the convective heat transfer coefficient, such that Eq. (18) becomes

$$Q = hA(T_f - T_w). \tag{19}$$

The coefficient h is not a property of the fluid, as is k in the constitutive equation of heat conduction, but it depends on the fluid dynamics conditions established in the fluid flow. It is important to use this coefficient in heat conduction problems as thermal boundary conditions. Later, the link between the heat transfer coefficient and the fluid flow is discussed. In convective studies, the concept of boundary layer, shear layer, and turbulence must be introduced. Several forces are involved in the movement of a fluid: body forces (e.g., gravity), electric or magnetic forces, centrifugal or Coriolis force, surface forces due to the influence of the surrounding fluid, and tensile and shear (or viscous) forces. The surface forces per unit area are the stresses. In a fluid, the tensile and shear stresses are related to the time rate of the deformation, and at rest only the tensile stress (i.e., pressure) is present. The shear stress must be related to the rate of deformation. For a simple flow situation in which the velocities, variation is only in the direction y, normal to the velocity, Newton proposed the proportional relation between the shear stress τ, in the flow direction, to the velocity gradient

$$\tau = \mu du/dy, \tag{20}$$

where the property μ is the dynamic viscosity, with units in kg/m s and dimensions $[M\ L^{-1}\ T^{-1}]$. The kinematic viscosity, ν, also called momentum diffusivity, is defined as

$$\nu = \mu/\rho, \tag{21}$$

with units $\mathrm{m^2/s}$ and the dimensions $[L^2\,T^{-1}]$. If the fluid obeys Eq. (20), it is Newtonian; otherwise, it is non-Newtonian. Several equations have been proposed for non-Newtonian fluids, but only some of them are shown here:

$$\text{Ostwald–de Waele fluid}: \ \tau = m(du/dy)^n, \tag{22}$$

where m is the consistency of the fluid and n is the index.

$$\text{Eyring fluid}: \ \tau = A \sin h^{-1}(du/dy\ 1/B), \tag{23}$$

where A and B are constants.

$$\text{Bingham plastic}: \ \tau = \mu du/dy + \tau_o \quad \text{if } \tau > \tau_o \text{ and } du/dy = 0 \text{ if } \tau < \tau_o. \tag{24}$$

Fluid particles in contact with a surface at rest have zero velocity, but velocity increases rapidly in the thin layer of thickness δ, called the boundary layer. The Reynolds experiment demonstrated the existence of two forms of flow—laminar and turbulent. In laminar flow, the individual streamlines run orderly side by side, whereas in turbulent flow they are interwoven with each other in an irregular manner. The motion of fluid particles in turbulent flow fluctuates around some mean flow path, and vortices of different size and orientation occur. The measurement of the velocity component, u, in turbulent flow at a location versus time results in the sum of a time-averaged value, $\bar{u}$, and an instantaneous value, u', of the timewise fluctuation. Assuming that the mean velocity is only in the x direction of u and that $\bar{v}$ and $\bar{w}$ are zero, the turbulent

intensity J is defined as

$$J = \sqrt{1/3(\bar{u}'^2 + \bar{v}'^2 + \bar{w}'^2)}/\bar{u}, \tag{25}$$

where each of the three terms is the time mean of the square of the velocity fluctuations. Heat transfer is influenced by the nature of the turbulence, as described by the length scale of the turbulent fluctuations and the vortices. A measure of the length scale of the turbulent fluctuations and vortices is given by the length L,

$$L = \int_0^\infty R(y)dy, \tag{26}$$

where the parameter R, given by

$$R = (\bar{u}_1' u_2')/\sqrt{\overline{u_1}'^2}\sqrt{\overline{u_2}'^2} \tag{27}$$

is obtained by hot-wire measurements, which determine the instantaneous velocity fluctuations simultaneously but at a varying distance y. Heat transfer by convection is enhanced by turbulent flow compared to the laminar one.

2.4 Heat Transfer Applications

There are many applications of heat transfer. The main applications are discussed here. Heat transfer is important in many practical applications relating to different branches of engineering but with opposite goals. Indeed, for mechanical and civil engineering, when energy must be conserved, and heat transfer to the exterior must be reduced, e.g., heat transfer losses in buildings during the winter. Energy savings from a building's heating and cooling systems can be obtained with the use of heating networks and cogeneration plants. One very important application is for very low-temperature situations such as cryogenics, which uses superconducting systems, in which heat transfer control with the exterior is crucial. On the other hand, in electrical and electronic engineering, because of the Joule effect, heat transfer to the exterior must be increased to maintain a constant temperature or to decrease the temperature of the system.

The problem of dissipating heat is an important issue in many electrical and electronic applications, such as in electrical wires and cables. Increased heat dissipation can be obtained with finned surfaces, which are of great importance in many energy systems, such as economizers in steam power plants, convectors for steam and hot water heating systems, electric transformers, and aircraft engines. In electrical machines, dissipation of electric energy requires an appropriate design, which is also required for chemical and nuclear fields in which heat sources are present. Heat transfer with change of phase is important in melting and solidification processes, combustion, frost penetration into the ground, and in aerodynamic ablation phenomena such as the protective devices of atmospheric reentry vehicles. The presence of a moving heat source or sinks in heat conduction is important in arc welding, surface hardening, continuous casting, or quenching. Radiation heat transfer phenomena are important in conventional, nuclear, and fusion power plants; gas turbines; and propulsion devices for aircraft, missiles, satellites, and space vehicles. Radiation pressure, in connection with space flight, has been proposed as a means of propelling space vehicles. Energy exchange by radiation is important in the calculation of heat transfer between surfaces in enclosures, as applied to satellites or space probes, moving in vacuum. The correct evaluation of the shape factor is important in several engineering applications, such as the radiation cooling surfaces in the furnace of a steam boiler or the tungsten spiral in an electric light bulb. Gas radiation is important in the heat transfer evaluation of furnaces and steam boilers. It is also important in plasma technology and the reentry phase of satellites or space vehicles through the atmosphere, when the gas is at very high temperatures. Flame radiation is important in the evaluation of heat transfer within tube bundles of water-tube boilers. Heat transfer analysis is useful for evaluating the temperature of a fluid flowing in a tube if the fluid temperature differs greatly from the ambient temperature. An additional systematic error in the temperature measurement is due to the heat exchange by radiation with the surrounding solid surfaces. Solar radiation evaluation is another important application of heat transfer because it can be employed for house heating and cooling.

3. HEAT CONDUCTION

3.1 Steady State

3.1.1 Composite Plane Wall

Consider a composite wall with three layers of different materials with thermal conductivities k_1–k_3, thickness x_1–x_3, and temperatures at the two extremes equal to T_{w1} and T_{w2}. The solution to Eq. (2), using the one-dimensional assumption

and steady state, gives the heat flux through the surface A:

$$Q/A = (T_{w1} - T_{w2})/(x_1/k_1 + x_2/k_2 + x_3/k_3). \quad (28)$$

The temperature variation within the layers is linear if the thermal conductivity is constant versus the temperature. If the boundary conditions at the two extreme walls are not given with assumed values of temperature but are defined according to Eq. (19) (i.e., the heat flux is specified in the form of the two convective heat transfer coefficients, h_i and h_e), and the two temperatures of the fluids are known as T_e and T_i, the heat flux is given by

$$\frac{Q}{A} = \frac{(T_i - T_e)}{(1/h_i + x_1/k_1 + x_2/k_2 + x_3/k_3 + 1/h_e)}. \quad (29)$$

3.1.2 Composite Cylindrical Wall

Consider a metallic tube with internal radius R_1, external radius R_m, and thermal conductivity k_m covered by an insulating material with external radius R_i and thermal conductivity k_i. Assume the internal wall of the metallic tube is T_{w1} and the external wall of the insulating material is T_{w2}. The solution to the conservation equation, obtained using the one-dimensional cylindrical assumption and the steady state, gives the heat transfer per unit length L of the cylinder:

$$\frac{Q}{L} = \frac{(T_{w1} - T_{w2})}{[\ln(R_m/R_1)/(2\pi k_m) + \ln(R_i/R_m)/2\pi k_i)]}. \quad (30)$$

If the boundary conditions are of specified heat fluxes, with the convective heat transfer coefficients h_i and h_e and the two fluid temperatures (the internal T_i and the external T_e), then the heat transfer is

$$Q/L = (T_i - T_e)/(1/(2\pi R_1 h_i) + \ln(R_m/R_1)/(2\pi k_m) + \ln(R_i/R_m)/(2\pi k_i) + 1/(2\pi R_i h_e)). \quad (31)$$

For a large radius of the insulation, R_i, the previous equation shows a decrease in heat transfer with the increase in the radius R_i. In the case of small values of R_i the previous equation presents a maximum depending on R_i. The radius of the maximum heat transfer is equal to

$$R_{crit} = k_i/h_e. \quad (32)$$

The practical application is for the insulation of electrical wires, where the objective is to provide an electric insulation and the maximum heat transfer for wire cooling.

3.1.3 Extended Surfaces

Assume that a solid body needs to dissipate more heat and the temperature of the cooling fluid cannot be modified. The only possibility is to increase the surface of heat transfer between the solid body and fluid. The simplest case is that of a thin rod with constant cross section A, perimeter C, and length L that is exchanging heat with a fluid of temperature T_f. With the assumptions that the temperature of the rod is only dependent on the distance from the body, is equal to T_0 at the base of the rod near the wall, is constant across the section A, and is in steady state, Eq. (3) can be solved to give the following temperature variation along the rod:

$$\frac{[T(x) - T_f]}{[T_0 - T_f]} = \frac{[\cosh(m(L-x) + (h_2/(mk))\sinh(m(L-x))]}{[\cosh(mL) + (h_2/(mk))\sinh(mL)]}, \quad (33)$$

and the heat flux at the boundary between rod and body is

$$\frac{Q}{A} = \frac{mk(T_0 - T_f)[h_2/(mk) + \tanh(mL)]}{[1 + (h_2/(mk))\tanh(mL)]}, \quad (34)$$

where $m^2 = hC/(kA)$, h is the convective heat transfer along the perimeter and h_2 at the final section of the rod, and k is the thermal conductivity of the metal rod. If the temperature of a fluid flowing in a tube is measured by a thermometer located in a well welded into the tube wall, and the tube wall has a lower temperature than the fluid, heat will be exchanged by conduction from the well to the tube wall and the indicated temperature will not be the true fluid temperature. The error can be calculated using the previous equations. Finned surfaces are of great importance in many energy systems, such as economizers in steam power plants, convectors for steam and hot water heating systems, electric transformers, and in aircraft engines. Practical finned surfaces differ from the thin rod, and the problem becomes more complicated from a mathematical standpoint. The finned surfaces of practical interest are rectangular fins, straight fins of triangular profile, and circumferential fins, for which the problems are the minimum weight, the fin effectiveness, and the fin arrangement.

3.1.4 Walls with Heat Sources

When heat sources are present in heat conducting bodies, it is necessary to calculate the temperature distribution within the bodies. In electrical machines and apparatus, the dissipation of electric energy

requires an appropriate design. In the chemical and nuclear fields, similar heat sources are present. In a plane wall $2L$ wide, with heat sources surrounded by a fluid at temperature T_e, Eq. (2) becomes, in steady state

$$d^2(T - T_e)/dx^2 + q_g/k = 0, \tag{35}$$

and the solution is

$$(T - T_e) = -q_g x^2/(2k) + C_1 x + C_2, \tag{36}$$

where the two constants C_1 and C_2 can be calculated by the boundary conditions imposed on the surfaces. With the convective heat transfer boundary condition at the two walls, and the convective heat transfer coefficient given by h, the final solution is

$$(T - T_e) = q_g(L^2 - x^2)/(2k) + q_g L/h. \tag{37}$$

The temperature distribution along the plane wall is parabolic with the maximum on the middle of the plane.

3.2 Unsteady State

The temperature of a body exchanging heat with a fluid of constant temperature is variable with time in unsteady state. Assume that the body has a large thermal conductivity compared to that of the fluid and the temperature difference is low in the body compared to that of the fluid. In this case, the first law of thermodynamics allows one to write the equivalence between the body internal energy variation (per unit time) and the heat transfer by convection with the fluid:

$$\rho c V dT/dt = hA(T_f - T), \tag{38}$$

where V is the volume and A is the surface of the body. With the initial condition of the body temperature, T_0, the following solution is found:

$$\begin{aligned}(T - T_f)/(T_0 - T_f) &= \exp - (hA/\rho c V t) \\ &= \exp - [(hL/k)(\alpha t/L^2)],\end{aligned} \tag{39}$$

where $L = V/A$ is a characteristic dimension that depends on the form of the body. The group of variables hL/k is called the Biot number and is represented by B_i. The group of variables $\alpha t/L^2$ is called the Fourier number and is represented by F_o.

3.3 Moving Boundaries

In many engineering heat transfer problems, there is a change of phase in the conducting medium. Examples include melting and solidification processes, combustion, frost penetration into the ground, and aerodynamic ablation applications such as in the protective devices of atmospheric reentry vehicles. The interface between the two regions of different physical properties moves as a function of time and latent heat is rejected or absorbed at its boundary. The formation of ice in water at 0°C is a problem that allows an exact solution, neglecting the heat capacity in the ice. The ice layer at time t is $x(t)$, the temperature in the liquid phase is equal to the phase change temperature T_{cp} and is constant throughout the liquid layer, and the solid phase temperature changes from the phase change value T_{cp} to that of the external surface T_s. The latent heat of fusion, Q_L, released by the change of phase of a layer dx has to be removed by conduction through the solid layer according to the following equation:

$$\rho Q_L dx/dt = k(T_{cp} - T_s)/x, \tag{40}$$

where ρ and k are relative to the solid. Imposing the boundary condition that the ice layer is zero at the beginning time, the analytical solution is

$$x = \sqrt{2k(T_{cp} - T_s)t/(\rho Q_L)}. \tag{41}$$

Consider the solidification of a liquid that has a temperature T_2 at a long distance from the surface. Heat is removed from the surface, which is maintained at the constant temperature T_1. In addition to the latent heat of fusion, solidification can take into account that the thermal conductivities, specific heats, and densities are different for the two phases. With the method developed by Neumann, it is found that the two temperatures, T_S and T_L, respectively in the solid (ice) and liquid (water) phases, are given by

$$\frac{(T_{cp} - T_L)}{(T_{cp} - T_1)} = 1 - \frac{\operatorname{erf}(x/2\sqrt{\alpha_1 t})}{\operatorname{erf}(K\beta/2)\sqrt{\alpha_1}} \tag{42}$$

$$\frac{(T_{cp} - T_s)}{(T_{cp} - T_2)} = 1 - \frac{\operatorname{erfc}(x/2\sqrt{\alpha_2 t})}{\operatorname{erfc}(K/2\sqrt{\alpha_2})}. \tag{43}$$

The interfaces move according to the laws

$$X_1 = K\beta\sqrt{t} \tag{44}$$

$$X_2 = K\sqrt{t}, \tag{45}$$

where $\beta = \rho_2/\rho_1$ takes into account the density decrease of ice and K is a constant to be determined numerically. In the water–ice system, an adequate approximation of K is

$$K_0 = \sqrt{2(T_{cp} - T_1)k_1/(Q_L\rho_1)} \quad \text{for } \beta = 1 \tag{46}$$

$$K_\beta = \sqrt{2(T_{cp} - T_1)k_1/(Q_L\rho_1\beta^2)} \quad \text{for } \beta > 1. \tag{47}$$

In the process of ablation, the melting fluid can be assumed to be completely removed by aerodynamic forces, the surface recedes with time, the surface temperature remains constant at the phase change temperature, and a temperature distribution exists only in the remaining solid.

Assuming the semiinfinite solid is heated by a constant heat flux, q_0, at the surface $x = 0$, the melted material is removed, the boundary of the solid and melt line is located at $X(t)$, the temperature distribution in the solid penetrates to a depth $\delta(t)$, and the temperature for $x > \delta(t)$ is constant and equal to T_S. The variation of the melt line X with time t is represented parametrically by

$$\Omega = -1/3[\zeta - 2 + 2(1+v)\ 1n\{(2(1+v) - \zeta)/(2v)\}], \tag{48}$$

where $\Omega = q_0^2 t/(\rho Q_L k \theta_P)$, $v = Q_L/(c\theta_P)$, $\zeta = q_0(\delta - X)/(\theta_P k)$, and $\theta_P = T_{cp} - T_S$, and also by

$$\lambda = -1/3[\zeta - 2 + 2v\ 1n\{(2(1+v)\text{-}\zeta)/(2v)\}], \tag{49}$$

where $\lambda = q_0 X/(K\theta_P)$.

The presence of a moving heat source or sinks in heat conduction is important in arc welding, surface hardening, continuous casting, or quenching. If a point source moves at velocity U in x direction in an infinite medium, with temperature T_f and a heat generated by the point source equal to q_g, the temperature distribution is given by

$$T - T_f = q_g \exp(-(U/(2\alpha))\ (r + \xi))/(4\pi k r), \tag{50}$$

where $r = \sqrt{(\xi^2 + y^2 + z^2)}$ and $\xi = x - Ut$.

4. THERMAL RADIATION

Several processes occur at high temperatures in engineering, and the proper knowledge of the radiation heat transfer phenomena is important for the correct design of equipment. Applications are relate to conventional, nuclear, and fusion power plants; gas turbines; and propulsion devices for aircraft, missiles, satellites, and space vehicles.

4.1 Radiation of Strongly Absorbing Media

For strongly absorbing media, the parameters can be related to the surface conditions neglecting the differences in the thermodynamic state within the thin layer, which contributes to absorption or emission. A new term, monochromatic emittance ε_λ, is introduced by the equation

$$\varepsilon_\lambda = i_\lambda / i_{\lambda b} = \alpha_\lambda, \tag{51}$$

which shows that the monochromatic absorptance is equal to the monochromatic emittance. A radiant beam, incident on a medium, can be absorbed in the medium, transmitted through it, or reflected. Reflectance, ρ_λ, is the ratio of the reflected to the incident energy; absorbance, α_λ, is the ratio of the absorbed to the incident energy, and transmittance, τ_λ, is the ratio of the transmitted to the incident energy. According to energy conservation, the following relation can be written:

$$\rho_\lambda + \alpha_\lambda + \tau_\lambda = 1, \tag{52}$$

which for a strongly absorbing medium, in which the transmitted energy is small, becomes

$$\rho_\lambda + \alpha_\lambda = 1. \tag{53}$$

Most solids absorb radiation very strongly and all the incident radiation is absorbed in a very thin layer of the surface. Exceptions are a few solid substances like glass, quartz, rock salt, and most liquids in the visible and near-infrared range. However, all the mentioned substances and liquids absorb very strongly in the bulk of the infrared.

4.1.1 Enclosures with a Nonparticipating Medium

Energy exchange calculations between surfaces in an enclosure with a nonparticipating medium can be applied to satellites or space probes moving in vacuum. The surfaces of the enclosure are assumed to be strongly absorbing media. The surfaces emit with an intensity independent of direction and with perfectly diffuse or specular reflection. The shape factor determines the influence of the local arrangement of the surfaces involved. Considering two area elements, dA_1 and dA_2, with dimensions small compared to the distance, s, between them, the shape factor of dA_1 toward dA_2, $dF_{dA_1 - dA_2}$, is given by the ratio between the heat flux incident on dA_2 to the heat flux leaving dA_1 as

$$dF_{dA_1 - dA_2} = \cos\beta_1 \cos\beta_2\ dA_2/(\pi s^2). \tag{54}$$

The shape factor depends only on the angular relationship describing the mutual position of the two area elements. It can be shown that

$$dA_1 dF_{dA_1 - dA_2} = dA_2 dF_{dA_2 - dA_1}, \tag{55}$$

which is called the reciprocity relation. The shape factor for a finite area A_2 is finally given by

$$F_{dA_1 - A_2} = \int_{A_2} (\cos\beta_1 \cos\beta_2\ dA_2/(\pi s^2)) \tag{56}$$

and the reciprocity relation for finite areas is

$$A_1 F_{A_1-A_2} = A_2 F_{A_2-A_1}. \quad (57)$$

The shape factor is 1 for two surfaces infinite in extent and for a surface inside an enclosure. For other configurations, the integral must be solved. For a surface A_i, completely surrounded by n surfaces, A_k, the following relation holds:

$$\sum_k F_{A_i-A_k} = 1. \quad (58)$$

The correct evaluation of the shape factor is important in several engineering applications, such as the radiation cooling surfaces in the furnace of a steam boiler or the tungsten spiral in an electric light bulb.

4.1.2 Enclosures with Black Surfaces

The radiation energy flux leaving the surface i, with A_i, is $A_i e_{bi}$, where $e_{bi} = \sigma T_i^4$. The radiation energy flux leaving a surface k in the direction toward surface i is $A_k F_{k-i} e_{bk}$; it arrives at surface i and is absorbed by it. The total energy flux absorbed by the surface i is $\sum_k A_k F_{k-i} e_{bk}$, and the net heat flux, exchanged per unit time, is

$$Q_i = A_i e_{bi} - \sum_k A_k F_{k-i} e_{bk} \quad (59)$$

Using the reciprocity law and Eq. (54), we obtain

$$Q_i = A_i \sum_k F_{i-k}(e_{bi} - e_{bk}), \quad (60)$$

and this equation can be written for each surface. The system of equations can be solved once the boundary conditions are prescribed, usually knowing the temperature, the emissive power, and the shape factor. Rapid solution of these equations can be carried out with analogy to the electric networks.

4.1.3 Enclosures with Gray Surfaces

Surfaces that emit radiation with an intensity independent of direction and that are perfect diffuse reflectors allow the introduction of radiosity, B, as the total heat flux leaving the surface per unit area and time. Radiosity is then independent of whether radiation is generated by emission or reflection. Irradiance, H, is the heat flux arriving at a surface per unit area and time. The surfaces are also assumed as gray and non-transparent, with $\alpha_i = \varepsilon_i$ and $\rho_i + \alpha_i = 1$. The heat loss of the surface by radiation is

$$Q_i = (B_i - H_i) A_i. \quad (61)$$

The heat flux H_i arriving from the surrounding surfaces is

$$H_i = \sum_k F_{i-k} B_k. \quad (62)$$

The radiosity B_i, due to emission and reflection, is given by

$$B_i = \varepsilon_i e_{bi} + \rho_i H_i. \quad (63)$$

The term H_i can be eliminated from the previous equations, with the assumption $\rho_i = 1 - \varepsilon_i$, giving

$$Q_i = \sum_k A_i F_{i-k}(B_i - B_k) \quad (64)$$

$$Q_i = A_i(e_{bi} - B_i)\varepsilon_i/(1 - \varepsilon_i). \quad (65)$$

Similar equations can be solved for each surface with the same boundary conditions as those for the enclosure with black surfaces, and the similar electrical analogy can be carried on. An additional resistance has been introduced into each connection of a node without influence on the potential of this node. In other words, the temperature of an adiabatic wall is independent of the absorption of the wall. The application of the procedure to two parallel gray surfaces of very large extent gives the heat transfer

$$Q_1/A_1 = (e_{b1} - e_{b2})/(1/\varepsilon_1 + 1/\varepsilon_2 - 1), \quad (66)$$

whereas for a surface A_1, enclosed in A_2, it gives

$$Q_1/A_1 = (e_{b1} - e_{b2})/(1/\varepsilon_1 + A_1/A_2(1/\varepsilon_2 - 1)). \quad (67)$$

If the surface 2 is much larger than 1, Eq. (67) becomes

$$Q_1/A_1 = (e_{b1} - e_{b2})\varepsilon_1. \quad (68)$$

4.2 Radiation of Weakly Absorbing Media (Gases)

Scattering is neglected in gases. Knowledge of the coefficient of absorption χ_λ is sufficient for the analysis because of the Kirchhoff equation,

$$j_\lambda = \chi_\lambda K_{\lambda e} = \chi_\lambda i_{\lambda b}. \quad (69)$$

The coefficient of absorption χ_λ can be measured from emission and absorption measurements in a medium of dimension S using the simplified equation of the monochromatic emissivity, ε_λ:

$$\varepsilon_\lambda = 1 - \exp(-\rho\chi_\lambda S) = \alpha_\lambda = 1 - \tau_\lambda. \quad (70)$$

Heat transfer by radiation can be calculated in an enclosure filled with an absorbing and emitting gas,

with the assumptions that the intensity of the walls of the enclosure are independent of direction and that they are diffuse reflectors and gray surfaces. The gas filling the enclosure is gray. The radiation heat flux leaving the surface k to the surface i is $A_iF_{i-k}B_k$, and the part arriving on surface i is $A_iF_{i-k}(1-\alpha_{i-k,g})B_k$, where $\alpha_{i-k,g}$ is the average absorbance between surfaces i and k. The radiation heat flux arriving at surface i is $A_i\varepsilon_{i,g}e_{bg}$, where $\varepsilon_{i,g}$ is the emissivity of the gas. The radiation heat flux leaving i is A_iB_i, and the net radiation heat loss of the surface i is

$$\begin{aligned} Q_i &= A_i[B_i - \sum_k F_{i-k}(1-\alpha_{i-k,g})B_k - \varepsilon_{i,g}e_{bg}] \\ &= A_i\varepsilon_{i,g}(B_i - e_{bg}) \\ &\quad + \sum_k A_iF_{i-k}(1-\alpha_{i-k,g})\,(B_i - B_k), \end{aligned} \tag{71}$$

and the net heat loss of the gas due to the radiation heat is

$$Q_g = \sum_i A_i\varepsilon_{i,g}(e_{bg} - B_i). \tag{72}$$

The last three equations are sufficient to calculate the heat loss for surfaces with known temperature, the temperature for adiabatic surfaces, and the heat loss of the gas in the enclosure. Gas radiation is important in the heat transfer evaluation of furnaces and steam boilers. It is also important in plasma technology and in the reentry phase of satellites or space vehicles through the atmosphere when gases are at very high temperatures. In this condition, radiation has its largest contribution from visible or near-ultraviolet wavelength range. Flame radiation is important in the evaluation of heat transfer within the convective surfaces of tube bundles in water-tube boilers.

5. HEAT TRANSFER BY CONVECTION

5.1 Conservation Equations

The conservation equations are written for a volume of fluid, in a system of Cartesian coordinates, of dimensions dx, dy, dz. The mass conservation for a fluid system without chemical reactions is

$$\partial\rho/\partial t + \partial(\rho u)/\partial x + \partial(\rho v)/\partial y + \partial(\rho w)/\partial z = 0, \tag{73}$$

and introducing the notation

$$D/dt = \partial/\partial t + u\partial/\partial x + v\partial/\partial y + w\partial/\partial z \tag{74}$$

it can be written as

$$D\rho/dt + \rho(\partial u/\partial x + \partial v/\partial y + \partial w/\partial z) = 0. \tag{75}$$

The conservation of momentum for the control volume states that the sum of all the forces acting on the mass must equal the time derivative of the momentum of the fluid mass. The fluid momentum changes due to the momentum flux, both out of and in the control volume, or due to the time change of the momentum of mass of the control volume. The surface forces per unit area, or the stresses, represent the action of the surrounding fluid on the control volume fluid. The stresses have two indices, as, for example, p_{xy} and p_{xz}. The first index indicates the direction normal to the surface and the second the direction of the stress component. The shear stresses are, for example, p_{xy} and p_{xz}. The normal stress $(p_{xx}-p)$ is composed of a viscous stress p_{xx}, which is a tensile stress that vanishes when the velocity becomes zero, and the fluid pressure p. The minus sign for the pressure takes into account the fact that it is assumed positive when it is compression stress (i.e., it acts from outside). The momentum conservation in the x direction is

$$\begin{aligned} &\partial(\rho u^2)/\partial x + \partial(\rho vu)/\partial y + \partial(\rho wu)/\partial z + \partial(\rho u)/\partial t \\ &= \partial(p_{xx} - p)/\partial x + \partial p_{yx}/\partial y + \partial p_{zx}/\partial z + \rho g_x. \end{aligned} \tag{76}$$

Using the mass conservation equation, it becomes

$$\begin{aligned} &\rho u\partial u/\partial x + \rho v\partial u/\partial y + \rho w\partial u/\partial z + \rho\partial u/\partial t \\ &= -\partial p/\partial x + \partial p_{xx}/\partial x + \partial p_{yx}/\partial y + \partial p_{zx}/\partial z + \rho g_x, \end{aligned} \tag{77}$$

and with notation similar to that of the mass conservation equation, that is,

$$Du/dt = \partial u/\partial t + u\partial u/\partial x + v\partial u/\partial y + w\partial u/\partial z, \tag{78}$$

it becomes

$$\begin{aligned} \rho Du/dt = &- \partial p/\partial x + \partial p_{xx}/\partial x \\ &+ \partial p_{yx}/\partial y + \partial p_{zx}/\partial z + \rho g_x. \end{aligned} \tag{79}$$

It is similarly obtained in the other two directions:

$$\begin{aligned} \rho Dv/dt = &- \partial p/\partial y + \partial p_{xy}/\partial x \\ &+ \partial p_{yy}/\partial y + \partial p_{zy}/\partial z + \rho g_y \end{aligned} \tag{80}$$

$$\begin{aligned} \rho Dw/dt = &- \partial p/\partial z + \partial p_{xz}/\partial x + \partial p_{yz}/\partial y \\ &+ \partial p_{zz}/\partial z + \rho g_z. \end{aligned} \tag{81}$$

At this point, the empirical relations connecting the stresses with the deformation of the fluid volume, also called constitutive equations, must be introduced, but only for Newtonian fluids. The shear

stresses for a Newtonian fluid have been derived by Stokes and are the following

$$p_{xx} = 2\mu\partial u/\partial x - \mu'(\partial u/\partial x + \partial v/\partial y + \partial w/\partial z) \quad (82)$$

$$p_{yx} = p_{xy} = \mu(\partial v/\partial x + \partial u/\partial y) \quad (83)$$

for the x direction and similarly for the other directions. The viscosity μ' has the value of $2\mu/3$ for monoatomic molecules and is the only one considered here. Introducing the previously mentioned constitutive equations, the three momentum conservation equations become

$$\begin{aligned}\rho Du/dt = &-\partial p/\partial x + 2\partial(\mu\partial u/\partial x)/\partial x \\ &+ \partial[\mu(\partial v/\partial x + \partial u/\partial y)]/\partial y \\ &+ \partial[\mu(\partial u/\partial z + \partial w/\partial x)/\partial z \\ &- \partial[\mu'(\partial u/\partial x + \partial v/\partial y + \partial w/\partial z)]/\partial x \\ &+ \rho g_x \end{aligned} \quad (84)$$

$$\begin{aligned}\rho Dv/dt = &-\partial p/\partial y + 2\partial(\mu\partial v/\partial y)/\partial y \\ &+ \partial[\mu(\partial v/\partial z + \partial w/\partial y)]/\partial z \\ &+ \partial[\mu(\partial v/\partial x + \partial u/\partial y)]/\partial x \\ &- \partial[\mu'(\partial u/\partial x + \partial v/\partial y + \partial w/\partial z)]/\partial y \\ &+ \rho g_y \end{aligned} \quad (85)$$

$$\begin{aligned}\rho Dw/dt = &-\partial p/\partial z + 2\partial(\mu\partial w/\partial z)/\partial z \\ &+ \partial[\mu(\partial w/\partial x + \partial u/\partial z)]/\partial x \\ &+ \partial[\mu(\partial w/\partial y) + \partial v/\partial z)]/dy \\ &- \partial[\mu'(\partial u/\partial x + \partial v/\partial y + \partial w/\partial z)]/\partial z \\ &+ \rho g_z. \end{aligned} \quad (86)$$

These equations are referred to as Navier–Stokes equations. The energy conservation equation is derived on the basis of the first law of thermodynamics applied to the fluid contained in the control volume. The time variation of the total energy contained in the control volume, $e_t = u + V^2/2$, which is the sum of the internal and kinetic energy, $dE_t/dt = d(e_tM)/dt$, is equal to the sum of the work transfer per unit time, dW/dt, of the energy exchange by convection, dE_c/dt, of the energy transfer by molecular movements, dE_m/dt (i.e., heat conduction) but also by diffusion:

$$dE_t/dt = d(e_tM)/dt = dW/dt + dE_c/dt + dE_m/dt. \quad (87)$$

For the control volume of dimensions dx, dy, dz, the total energy becomes

$$d(e_tM)/dt = \partial[(\rho e_t)dxdydz]\partial t. \quad (88)$$

The work exchanged in the x direction is positive when the force is acting in the flow direction and negative in the opposite direction and is

$$\begin{aligned}dW/dt = &[\partial(up)/\partial x - \partial(up_{xx})/\partial x - \partial(up_{yx})/\partial y \\ &- \partial(up_{zx})/\partial z]dxdydz + \rho ug_x dxdydz. \end{aligned} \quad (89)$$

The energy flux by convection in the x direction is

$$dE_c/dt = -\partial[\rho ue_c dxdydz]/\partial x. \quad (90)$$

The energy transfer by molecular movements in the x direction is

$$dE_m/dt = -\partial[e_{mx}dxdydz]/\partial x. \quad (91)$$

Using similar expressions for the other two directions, with the help of the mass conservation equation and the substantial derivative of the total energy, the total energy conservation equation is obtained as follows:

$$\begin{aligned}&\rho De_t/dt + \partial(up)/\partial x + \partial(vp)/\partial y + \partial(wp)/\partial z \\ &+ \partial(e_{mx})/\partial x + \partial(e_{my})/\partial y + \partial(e_{mz})/\partial z \\ &= \partial(up_{xx})/\partial x + \partial(up_{yx})/\partial y + \partial(up_{zx})/\partial z \\ &+ \partial(vp_{xy})/\partial x + \partial(vp_{yy})/dy + \partial(vp_{zy})/\partial z \\ &+ \partial(wp_{xz})/\partial x + \partial(wp_{yz})/\partial y + \partial(wp_{zz})/\partial z \\ &+ \rho ug_x + \rho vg_y + \rho wg_z. \end{aligned} \quad (92)$$

For a one-component fluid, the energy flux e_m is due to the Fourier equation.

5.2 Forced Convection

5.2.1 Laminar Flow

In a two-dimensional steady-state flow, the previous equations, restricted to the boundary layer, reduce to

$$\partial u/\partial x + \partial v/\partial y = 0 \quad (93)$$

$$\rho(u\partial u/\partial x + v\partial u/\partial y) = -\partial p/\partial y + \mu\partial^2 u/\partial y^2 \quad (94)$$

$$\partial p/\partial y = 0 \quad (95)$$

$$\rho c_p(u\partial T/\partial x + v\partial T/\partial y) = k\partial^2 T/\partial y^2, \quad (96)$$

considering negligible the influence of the gravitational body forces and assuming a fluid with constant properties.

5.2.2 Turbulent Flow

Considering a constant-property fluid in a two-dimensional flow and dividing the turbulent flow in a time-averaged value, $\bar{v}$ and a fluctuating value, v', for each variable, $u = \bar{u} + u'$, $v = \bar{v} + v'$, $p = \bar{p} + p'$, $T = \bar{T} + T'$, using the time average rules it is possible to obtain the following equations for the conservation of mass and momentum:

$$\partial\bar{u}/\partial x + \partial\bar{v}/\partial y = 0 \quad (97)$$

$$\rho(\partial\bar{u}/\partial t+\bar{u}\partial\bar{u}/\partial x+\bar{v}\partial\bar{u}/\partial y+\partial\overline{u'^2}/\partial x+\partial\,\overline{u'v'}/\partial y = -\partial\bar{p}/\partial x+\mu\partial^2\bar{u}/\partial y^2 \quad (98)$$

$$\rho\partial\,\overline{v'^2}/\partial y = -\partial\,\bar{p}/\partial y. \quad (99)$$

Integration of Eq. (99) gives $p_s = \bar{p} + \rho\overline{v'^2}$, which simplifies Eq. (98) giving the following equations, including the energy conservation equation:

$$\partial\bar{u}/\partial x + \partial\bar{v}/\partial y = 0 \quad (100)$$

$$\rho(\partial\bar{u}/\partial t + \bar{u}\partial\bar{u}/\partial x + \bar{v}\partial\bar{u}/\partial y) = -\partial p_s/\partial x + \mu\partial^2\bar{u}/\partial y^2 - \rho\partial\overline{u'v'}/\partial y \quad (101)$$

$$\rho c_p(\partial\bar{T}/\partial t + \bar{u}\partial\bar{T}/\partial x + \bar{v}\partial\bar{T}/\partial y) = k\partial^2\bar{T}/\partial y^2 - \rho c_p\partial\overline{T'v'}/\partial y. \quad (102)$$

The momentum equation differs from the laminar counterpart for the term containing the fluctuating quantities, which can be interpreted as resulting from turbulent stresses. The turbulent–boundary-layer equations contain too many unknowns and it is necessary to relate the turbulent stresses to the mean time values. Boussinesq first proposed the following:

$$\tau_t = -\rho\,\overline{u'}\,\overline{v'} = \rho\varepsilon_m\partial\,\bar{u}/\partial y, \quad (103)$$

where ε_m is the turbulent diffusivity of momentum. Similarly, the corresponding term of the energy equation has been proposed as given by

$$-\rho c_p\,\overline{T'}\,\overline{v'} = \rho c_p\varepsilon_h\partial\,\bar{T}/\partial y, \quad (104)$$

which introduces ε_h, the turbulent diffusivity of heat. The ratio of the two turbulent diffusivities has been defined as the turbulent Prandtl number:

$$Pr_t = \varepsilon_m/\varepsilon_h. \quad (105)$$

5.3 Natural Convection

The laminar boundary-layer equations under body force due to gravitational acceleration in low-velocity conditions, when the effect of viscous dissipation can be neglected, can be written for plane flow as well as for steady rotationally symmetric flow, as

$$\partial(\rho u r_0^n)/\partial s + \partial(\rho v r_0^n)/\partial y = 0 \quad (106)$$

$$\rho(u\partial u/\partial s + v\partial u/\partial y) = -dp/ds + \partial(\mu\partial u/\partial y)/\partial y + \rho g_s \quad (107)$$

$$\rho c_p(u\partial T/\partial s + v\partial T/\partial y = \partial(k\partial T/\partial y)/\partial y, \quad (108)$$

where s is the coordinate along the curved surface and y is normal to it. The exponent n is zero for two-dimensional flow and 1 for rotationally symmetric flow. Outside the boundary layer, the viscous effects are negligible and the momentum equation reduces to

$$dp/ds = \rho_e g_s - \rho_e u_e du_e/ds, \quad (109)$$

where e refers to the conditions outside the boundary layer. The momentum equation becomes

$$\rho(u\partial u/\partial s + v\partial u/\partial y) = g_s(\rho - \rho_e) + \rho_e u_e du_e/ds + \partial(\mu\partial\,u/\partial y)/\partial y, \quad (110)$$

and introducing the thermal expansion coefficient

$$\beta = 1/v(\partial v/\partial t)_p = -1/\rho(\partial\rho/\partial t)_p, \quad (111)$$

the following equation is obtained:

$$\rho(u\partial u/\partial s + v\partial u/\partial y) = -\beta\rho g_s(T - T_e) + \rho u_e du_e/ds + \partial(\mu\partial u/\partial y)/\partial y. \quad (112)$$

6. DIMENSIONAL ANALYSIS

For some heat transfer phenomena, the basic understanding of the physical processes is not enough to permit the writing of the fundamental equations. The alternative is to perform experiments and to generalize the results by dimensional analysis. For other heat transfer applications, an approximate knowledge of the equations is possible, and this allows one to start with these equations and make them dimensionless. One example of this approach is the case of flow separation on cylinders. The writing of the basic Navier–Stokes equations allows the conclusion that the velocity fields and pressure fields around cylinders are physically similar, provided the Reynolds number has the same value for all the considered cases. If the cylinder is kept at a temperature T_w and the fluid is at T_0, the Navier–Stokes equations can be written dimensionless and the dimensionless temperature can be found as

$$\theta^* = (T - T_0)/(T_w - T_0) = f(x/d,\ \mathrm{Re},\ Pr,\ Ec), \quad (113)$$

where x is the Cartesian coordinate and d is the cylinder diameter. The Reynolds number, Re, is defined as

$$\mathrm{Re} = \rho u_0 d/\mu, \quad (114)$$

where u_0 is the fluid velocity outside the boundary layer. The Prandtl number, Pr, is defined as

$$\mathrm{Pr} = \mu c_p/k. \quad (115)$$

The Eckert number, Ec, is defined as

$$\mathrm{Ec} = u_0^2/c_\mathrm{p}\theta_\mathrm{w}. \tag{116}$$

The convective heat transfer coefficient at the wall between fluid and cylinder, h, can be put in dimensionless form by introducing the Nusselt number, Nu, which is defined as

$$\mathrm{Nu} = hd/k = (\partial\theta^*/\partial(x/d)). \tag{117}$$

Considering also a fluid with variable properties, T_w/T_0, the final relation of the average Nusselt number on the total surface is

$$\mathrm{Nu}_\mathrm{average} = f(\mathrm{Re},\ \mathrm{Pr},\ \mathrm{Ma},\ T_\mathrm{w}/T_0, \alpha, \gamma), \tag{118}$$

where the Mach number, Ma, is defined as

$$\mathrm{Ma} = u_0/a_0, \tag{119}$$

where u_0 is the velocity of the fluid outside the boundary layer and a_0 is the velocity of sound for a gas with c_p, $\gamma = c_\mathrm{p}/c_\mathrm{v}$, $k = C_k T^\alpha$, and $\mu = C_\mu T^\alpha$:

$$a_0 = \sqrt{(\gamma - 1)c_\mathrm{p}T_0} = \sqrt{\gamma R\ T}. \tag{120}$$

The following relation holds between Mach and Eckert numbers:

$$\mathrm{Ec} = (\gamma - 1)\mathrm{Ma}^2/(T_\mathrm{w}/T_0 - 1). \tag{121}$$

In the practical applications, Eq. (118) is usually stated as the sum of exponential terms of the form

$$X = C_1 + C_2 Y^a Z^b W^c, \tag{122}$$

where the values of C_1, C_2, a, b, and c have to be found by specific experiments. The effect due to the rarefaction of gas is taken into account by the dimensionless Knudsen number, Kn, which is defined as

$$\mathrm{Kn} = \lambda/L = \mathrm{Ma}/\mathrm{Re}\sqrt{\gamma\pi/2}\ , \tag{123}$$

where λ is the mean free molecular path and L is a body dimension characteristic in the flow field. In natural or free convection flow (i.e., when the flow and heat transfer are present because of a constant body force due to the gravitational effect, g) two additional dimensionless numbers are defined: the Grashof number, Gr, and the Rayleigh number, Ra, defined as

$$\mathrm{Gr} = \beta g(T_w - T_0)L^3/\nu^2 \tag{124}$$

$$\mathrm{Ra} = \mathrm{GrPr}. \tag{125}$$

Another dimensionless number that is often defined in heat transfer problems is the Stanton number, St, defined as

$$\mathrm{St} = \mathrm{Nu}/(\mathrm{RePr}). \tag{126}$$

SEE ALSO THE FOLLOWING ARTICLES

Combustion and Thermochemistry • Conservation of Energy Concept, History of • Energy in the History and Philosophy of Science • Forms and Measurement of Energy • Mechanical Energy • Temperature and Its Measurement • Thermal Energy Storage • Thermodynamic Sciences, History of • Work, Power, and Energy

Further Reading

Carslaw, H. S., and Jaeger, J. C. (1970). "Conduction of Heat in Solids." Oxford Univ. Press, New York.

Eckert, E. R. G., and Drake, R. M., Jr. (1972). "Analysis of Heat and Mass Transfer." McGraw-Hill, New York.

Manuscripts (1666–1667). "Essays of Natural Experiences Done in the Cimento Academy." National Library of Florence, Florence, Italy.

Heterotrophic Energy Flows

KENNETH A. NAGY
University of California, Los Angeles
Los Angeles, California, United States

1. **Importance of Heterotrophic Energy Flow**
2. **Energy Budget of an Animal**
3. **Energy Intake**
4. **Energy Use**
5. **Efficiency**
6. **Regulation of Energy Metabolism**
7. **Summary**

Glossary

anaerobic metabolism Cellular production of energy-containing adenosine triphosphate in the absence of oxygen.
basal metabolic rate The lowest rate of heat production by an adult mammal or bird (endothermic or "warm-blooded" animal), measured when body temperature is normal and costs for activity, temperature regulation, digestion, and other expenses are low or zero.
chemical potential energy Energy that is contained in the foodstuffs and that is released as heat or as other forms of chemical energy (e.g., adenosine triphosphate) during cellular oxidative or anaerobic metabolism.
efficiency The portion of energy used or converted out of the total energy input, usually expressed as a ratio (out/in) or as a percentage (100 * out/in).
feeding rate The amount of food, usually in grams of dry matter of the diet, that an animal eats each day.
field metabolic rate The rate of heat production by an animal living undisturbed in its natural habitat.
homeostasis Regarding energy flow, the maintenance over time of energy balance, where in = out and energy storage is zero (none accumulated or depleted over time).
oxidative metabolism Cellular production of energy-containing adenosine triphosphate from foodstuffs via an oxygen-using process.
production The synthesis of new biochemicals that contain chemical potential energy and their accumulation as stored fatty tissue, new body tissue (growth), or reproductive tissues (e.g., embryos, eggs, milk).
standard metabolic rate The lowest rate of heat production by a reptile, amphibian, fish, or invertebrate (ectothermic or "cold-blooded" animal) measured at a specified body temperature and when costs for activity, temperature regulation, digestion, and other expenses are low or zero.

Autotrophic organisms, including most plants and some micro-organisms, "eat" energy in the form of photons (radiant energy as sunlight) or as energy-containing bonds in inorganic chemicals. Heterotrophs eat autotrophs, thereby obtaining their energy in the form of organic chemicals (carbohydrates, fats, and proteins). Nearly all animals are strict heterotrophs, dining either on plants, other animals, or the remains of these organisms. Thus, life on earth depends on the flow of energy, primarily from the sun, through plants, through herbivores, through carnivores, and finally through detritivores.

1. IMPORTANCE OF HETEROTROPHIC ENERGY FLOW

1.1 Money Analogy

Energy is intimately involved in living processes at every level of biological organization, from the organelles in cells to ecosystems. The central importance of energy as a common currency for individual animals and within ecosystems is somewhat analogous to the importance of money in human societies. The main goal of working each day is to obtain this resource (food or money) in sufficient quantities to meet the needs of living that day. Daily needs are not fixed but rather vary with lifestyle and especially with behavior. If income exceeds daily needs, the excess can be stored or used for growth or reproduction. Too little of this resource can be fatal.

1.2 Competition for Food, Selection, and Adaptation

Thomas Malthus observed that populations reproduce faster than their food supply, and Charles

 © 2004 Elsevier Inc. All rights reserved.

Darwin realized that this should cause intense competition for food within a species. The critical component of the food is presumed to be its energy. Hence, energy availability should limit population density. It follows that there should be intense selection favoring individuals that manage their energy budgets more effectively. Effective management may involve (1) getting more energy from the environment each day, (2) digesting food more thoroughly, (3) reducing daily metabolic expenditures, and/or (4) increasing the "efficiency" of reproduction by producing more successful offspring from the same energy intake. If energy is the principal limiting resource for wild animals, selection pressure on biological properties that are directly concerned with energy metabolism should be more intense than selection pressure on other properties of animals. Thus, energy should be evolutionarily paramount, and most of the important features of living animals should be understandable from the perspective of energetics.

Unfortunately, studies indicate that this paradigm may be too simplistic an explanation of the biology of wild animals given that only rarely can food supply be shown to be an important limiting resource. When food energy is in excess, there is no pressure to conserve energy. If food is abundant, animals can simply eat more and overcome any shortcomings of energy management or inefficiencies within their bodies. Food shortages probably do limit populations at times, but such bottlenecks may be decades or centuries apart, making them difficult to study. During intervals between energy bottlenecks, other causes of mortality, such as predation, may predominate. Nevertheless, the energy bottlenecks are probably quite severe when they occur and should have a large selective effect on the characteristics of subsequent generations. Thus, animals may display phenotypes that reflect adaptations to energy limitations even though energy might not be limiting at the time. If so, knowledge of the energetics of a species should yield a better understanding of its biology, ecology, and evolution in general.

1.3 Energy Costs of Living

Living animal cells are chemically different from their environment, which is the extracellular fluid within the animal. These differences across the cell membrane tend to disappear due to passive diffusion down concentration gradients, and maintenance of constant intracellular conditions requires a continuous expenditure of energy by the cells. Additional energy is used by groups of differentiated cells (organs and organ systems) in carrying out the higher level functions they perform within animals. At the level of the whole animal, more energy is used for integrated actions such as social interchanges, predator avoidance, and food acquisition. Population energy flow is the sum of energy use by individuals, and community energy flow incorporates the effects of interactions among species of producers, consumers, and detritivores. Whole animals also must spend extra energy, beyond that spent by their cells on intracellular maintenance, to maintain internal chemical homeostasis in their extracellular fluids. Additional energy is needed for animals to move, obtain and digest their food, interact socially, grow, and reproduce. A complete energy budget for an animal would include all of the avenues of heat gain and loss and chemical potential energy gain and loss. Such a budget could be given in units of daily energy intake, loss, and storage in kilojoules per day (kJ/day). These units are rates reflecting the dynamic nature of animal energetics. This article focuses on chemical potential energy intake as food and follows the fate of that food energy.

2. ENERGY BUDGET OF AN ANIMAL

Animals fall into two main metabolic categories: endotherms (primarily mammals and birds that are sometimes termed "warm-blooded") and ectotherms (fish, amphibians, reptiles, and invertebrates). Endotherms use their high metabolic rates to produce the internal heat that allows them to maintain relatively high and constant body temperatures compared with their surroundings. Ectotherms have quite low rates of metabolic heat production, too low to warm their bodies much at all. However, some species of ectotherms, especially lizards, can achieve and maintain high body temperatures for hours by using external heat sources such as sunlight and sun-warmed rocks. Endotherms have much more metabolic machinery (enzymes) in their cells, and this makes their cellular operating costs much higher than those of ectotherms. The ability of endotherms to keep themselves warm allows them to occupy colder habitats such as those at high latitudes and altitudes. Ectotherms, which have good fasting endurance due to their low energy needs, can persist in habitats such as deserts and the ocean's depths, where food may be scarce periodically or continuously.

The only source of chemical potential energy for most animals is their food. The food can be taken in by mouth, or nutrients can be absorbed across body walls as in some parasites and marine organisms. Some of the food energy is digested and absorbed (assimilated energy), and the indigestible food residues are voided by the animals as feces. Feces actually contain additional substances such as bacteria from the hindgut that grow and feed on the undigested dietary residues as well as sloughed off gut cells and proteins (e.g., mucus) that contain energy that was assimilated earlier but escaped redigestion and reassimilation. Thus, the term "apparent assimilation efficiency" (AE) is used to describe the following ratio: (food energy in) – (fecal energy out)/(food energy in). Some assimilated chemical potential energy leaves animals as urine, and this energy is also not available for metabolism. Most of the urinary energy is in nitrogen waste products (e.g., urea, uric acid), but some is in mucus. The difference between energy ingested as food and that excreted as urine and feces is called metabolizable energy, and "apparent metabolizable energy efficiency" (MEE) is used to describe the following ratio: (food energy in) – (feces and urinary energy out)/(food energy in) (Fig. 1).

Metabolizable energy is used primarily for oxidative metabolism in most animals, and all of this energy ultimately leaves the animal as heat (Fig. 1). If any metabolizable energy is left over after paying the energetic costs of the essential physiological processes, it can be used for production (growth, storage, or reproduction). The essential energetic costs of living include the minimum (idling) cost, called basal metabolic rate (BMR) in endothermic animals and standard metabolic rate (SMR) in ectotherms, along with the additional costs of obtaining, ingesting, digesting, absorbing, and biochemically processing food as well as the costs of being alert, interacting socially, maintaining body temperature, being active (e.g., locomoting), and responding to diseases or parasites. Energy savings resulting from reduced oxidative metabolism may occur in endotherms that can lower their body temperatures for hours or days (torpor) or weeks (hibernation). Total metabolic cost also may be lower than the sum of its (separately determined) components if an animal can substitute heat produced from one process (e.g., activity cost, digestion cost) for simultaneous requirements for heat needed for maintenance of body temperature (thermoregulation cost).

When an animal eats just enough food to provide the metabolizable energy it burns in oxidative metabolism each day, it is maintaining energy balance. The chemical potential energy content of its body will not be changing from day to day, although during a given day it will have a positive energy budget when absorbing its food and a negative energy budget when the assimilation rate

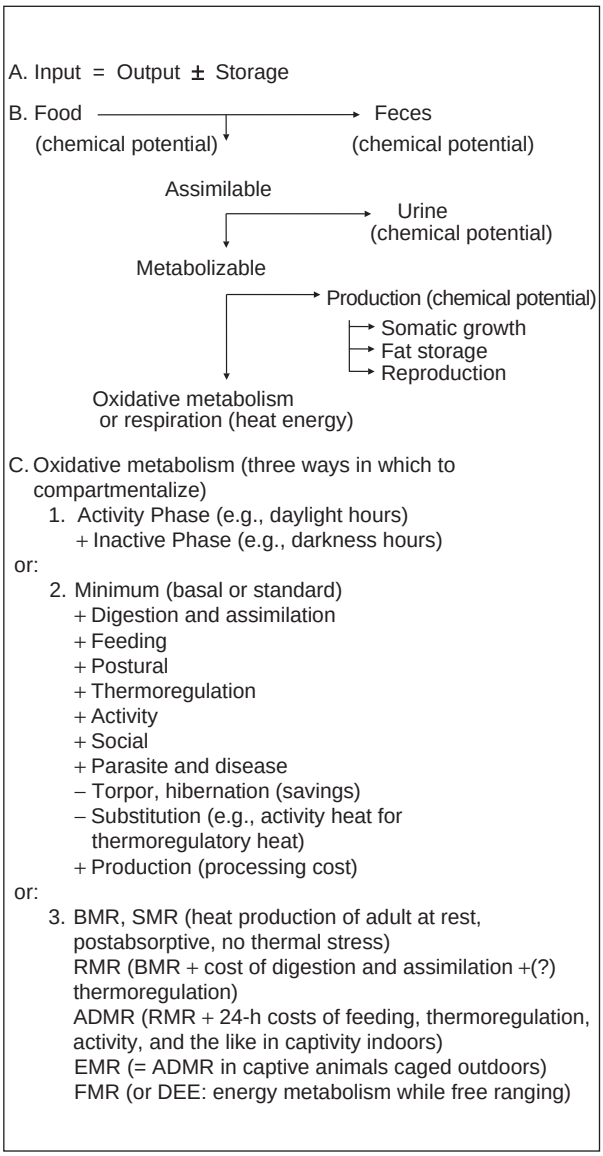


FIGURE 1 Three versions of a daily energy budget. Version A is in the form of a simple equation, and Version B accounts for the substances (food, feces, and urine) that move in and out of an animal. Version C shows three ways of subdividing the daily metabolic rate (daily heat production) portion of the energy budget of an animal, reflecting the different ways of measuring component parts. BMR, basal metabolic rate; SMR, standard metabolic rate; RMR, resting metabolic rate; ADMR, average daily metabolic rate; EMR, existence metabolic rate; FMR, field metabolic rate; DEE, daily energy expenditure.

declines below its metabolic rate. When an animal consumes more or less metabolizable energy than it burns in a day, its body will store the excess or contribute the deficit from body stores of energy. Most animals must have balanced energy budgets over relatively short periods of time (days, weeks, or even just hours for small endotherms such as some shrews and hummingbirds) to maintain life.

There are several ways in which to subdivide, or "itemize," an energy budget (Fig. 1). In theory, the various costs of living can be identified and listed as separate entries in the budget. Such a theoretical budget has intellectual value and appeal, but the practical difficulties of measuring each component of such a budget usually lead to its modification. For example, a theoretical budget might include the energetic expenses of hunting for and obtaining food (foraging cost), looking for and courting a mate (reproductive cost), and watching for and avoiding predators (antipredation cost). Using a rodent such as a desert kangaroo rat as an example, let us assume that it is in its burrow for 22 h and out of the burrow and active for 2 h each day. Let us also assume that we were able to measure the rat's daily metabolic expenditure and its metabolism only while it was in its burrow and that the difference between these expenditures (45 kJ/24 h–31 kJ/22 h) is 14 kJ/2 h active. The first problem in partitioning these 14 kJ into their theoretical budget categories is deciding whether basal metabolism, thermoregulatory and digestion costs, and the like, which the rat is also paying while it is outside its burrow doing other things, should be subtracted off to determine the pure cost of foraging, antipredation, and the like. This correction requires assumptions about energy allocation in active animals that have not yet been verified. The second problem is deciding, say from detailed observations of the animal's behavior while it was above ground, what portion of those 2 h the rat spent foraging, looking for mates, and avoiding predators and then subdividing the 14 kJ appropriately. This requires reading the mind of the kangaroo rat to determine whether it was looking for food, predators, or mates while it was sitting still in an alert, eyes-open posture during much of its time above ground. In fact, active animals probably do all three things simultaneously. The conservative choice is to label the 14 kJ simply as "metabolic cost while above ground." Creative experimentation will be required to increase the resolution of energy metabolism measurements of wild animals to evaluate the components of theoretical energy budgets. For now, researchers are settling for budget categories that reflect measurable parameters such as active and inactive periods of the day or basal (BMR), resting (RMR), average daily (ADMR), existence (EMR), and field (FMR) metabolic rates (Fig. 1).

Metabolizable energy that is not burned is available for production (synthesis of new biomass). Excess energy can simply be converted to fat and stored somewhere in the body, where it can be used for somatic growth (mostly as the protein and lipids needed for cell division and development) or for reproduction (growth of eggs or embryos as well as production of milk by mammals). Production involves formation of new biomass from foodstuffs. There are metabolic costs involved in these biochemical processes that add to the total rate of oxidative metabolism (heat production) of an animal producing new biomass. Animals are using energy continuously, and it is important to view energetics as an assemblage of processes, each having its own rate that varies through time. A useful analogy is that of a leaky bucket under an intermittently opened faucet. Various holes in the bucket can represent the components of oxidative metabolism that drain water (energy) out, lowering the water level. The intermittent faucet represents feeding events that add water (energy) to the bucket periodically. If more water is added than drains out each day, the volume of water in the bucket will increase (grow) over time. However, unlike the holes in the bucket, animals can influence the rates at which they use energy for various purposes, and they also influence their rate of energy intake (feeding rate).

We are accustomed to thinking about energy as packets, such as food items, that are measured in units of calories or joules. These are static units that do not represent rates of energy flow. The ratio of energy units to time units (joules/day or watts) does represent flow, and it is used herein. It is also important to pay attention to the time unit in discussions of animal energetics. Animals can spend energy very quickly over short time intervals (seconds or minutes, as during sprinting), and their maximum rates of energy expenditure can be far higher than their maximum rates of energy intake. However, over the course of a day, they may achieve energy balance by interspersing periods of activity with periods of resting, so that total energy expenditures do not exceed income. Likewise, some seasons may be difficult for animals and other seasons may be benevolent to animals, but over a whole year, the animals may achieve energy balance or even a positive energy budget that allows for growth or reproduction.

3. ENERGY INTAKE

The rate of metabolizable ("usable") energy intake that an animal achieves depends on many factors, including food availability in the habitat, feeding rate, diet composition and digestibility, capacities of the digestive tract, and appetite.

3.1 Food Availability

The availability of food in a habitat is different from food present in a habitat. Some of the potential food items are not available for a variety of reasons. For example, chemical potential energy in plant tissues might not be available to herbivorous animals because it is unreachable (underground or high on a shrub or tree), inedible (woody trunk of a tree or spine-covered stem or leaf), distasteful (containing antifeedant chemicals), toxic (containing poisons), or nutritionally inadequate (poor balance of nutrients or full of protein digestion inhibitors such as tannins). Some plant parts may be nutritionally poor or even toxic by themselves but can be edible and nutritious when eaten along with other plant parts that counteract toxins or make up for nutritional deficiencies. Similarly, plant items that are nutritious for an animal species having certain detoxification or digestive capabilities may be detrimental to another animal species that does not have those capabilities. Moreover, food availability also includes effects of patchy food distribution in space and time, and the ability of various animal species to exploit a given patch depends on the characteristics of the animals (e.g., small vs large, ability to catch insects). Thus, "food availability" must be defined from the perspective of the animal of concern and not simply as a fixed property of a habitat.

3.2 Food Composition and Digestibility

The chemical composition of diets that animals choose varies greatly. The plant matter in the diets of herbivores can consist largely of cell wall components that include cellulose, hemicellulose, and lignins. The chemical potential energy in these substances (collectively called "fiber") is generally unavailable to many herbivorous animals because they do not have the enzymes needed to digest fiber. Thus, less than half of the potential energy in the food may be available to these herbivores. On the other hand, animals that eat other animals are ingesting a food that has a chemical composition very similar to that of their own bodies, and most of those chemicals are digestible.

Diet digestibility, or digestive efficiency (DE), refers to the percentage of the dry matter (not energy, which is called assimilation efficiency or AE) of food that is digested and absorbed across the gut: DE = 100 * [(dry matter in) – (dry matter out as feces)/(dry matter in)]. DE varies greatly among different animals and diets. DE can be as low as 0.5% in earthworms (detritivores) eating soil with little organic matter in it (Table I) or as high as essentially 100% in hummingbirds or butterflies eating nectar, which is pure sugar dissolved in water. Herbivorous animals fall into two main categories: fermenters and nonfermenters. The most specialized fermenters are the ruminant mammals (e.g., cattle, antelope, sheep), which have a four-chambered stomach that contains an ecosystem of micro-organisms capable of digesting plant cell walls. The fiber is broken down by cellulases produced by microbes and is fermented by those microbes into volatile fatty acids that can be used by the animals for energy. The microbes use fiber fermentation and most of the soluble chemicals in the food (e.g., sugars, starch, proteins, lipids from the plant cells) for their own metabolism

TABLE I

Average Digestibilities and Energy Yields for Some Representative Diet/Consumer Combinations

Dietary category	Example species	Apparent dry matter digestibility (percentage)	Energy yield (metabolizable kJ/g dry matter)
Herbivore, monogastric	Gopher, jackrabbit	30–75	7–13
Herbivore, fermenter	Sheep, antelope	60–75	10–13
Carnivore	Insectivorous lizard	70–85	16–19
Granivore	Seed-eating rodent	90	16.9
Nectarivore	Hummingbird, butterfly	100	18
Soil eater	Earthworm	0.5–5.0	0.7

and production. Then, the microbes are digested and absorbed later in the digestive tract (beginning in the fourth stomach chamber, which is acidic). Thus, ruminants "cultivate" and "eat" microbes (mainly bacteria) and also eat plants. Nonruminant vertebrate herbivores ("monogastrics") have the standard acidic stomach and are able to digest only the soluble components in plant matter in their stomachs and small intestine. Thus, they obtain much less of the total energy in the diet than do fermenters (Table I), although in the field monogastrics often compensate by selecting young vegetation that has more soluble and less fiber components. However, many species of monogastric herbivores have some fermentation occurring somewhere in their gut, either in a somewhat modified portion of the stomach (e.g., macropod kangaroos, some primates) or in the large intestine (e.g., horses, many primates), and these species gain some of the energy locked up in plant fiber. Rabbits and some herbivorous rodents ingest their own feces (coprophagy) and thereby derive nutrition from the fermentative microbes bodies, as do ruminants. Carnivores digest most of the dry matter in their food (Table I). Granivores and nectarivores have the highest gain of energy from their food. They eat products of plants with which they have coevolved, wherein the plant benefits from pollination or seed dispersal and the animals benefit by obtaining food of high energetic quality. It is not surprising to find large benefits to consumers in circumstances where those producers that favor consumption have a selective advantage.

3.3 Stomach Capacity

Daily food consumption may be influenced by maximum stomach capacity. If predation or competitive pressures are high, animals that can obtain more food during a single foraging trip should have greater fitness. However, the information needed to test this hypothesis is largely unavailable. Early attempts to measure stomach volume by removing stomachs during autopsy, filling them with water, and weighing them are of questionable relevance to the situation in a live animal. In general, the capacities (volumes) of organs increase as a constant percentage of body mass as body size increases, but there is much room for interspecific differences. This is a topic for future research to address.

3.4 Appetite

Appetite is a major factor influencing energy intake rates, but little is known about appetite in wild animals. There is much literature on laboratory rats and mice, as well as on humans in relation to obesity and control of body mass, in both the physiological and psychological arenas. The scanty information on wild animals comes from measurements of feeding rate under various conditions rather than from measurements of hunger or satiation, which are the components of appetite, probably because of the difficulties in evaluating hunger and satiation directly in wild animals. Nutrient deprivation and hormone manipulation experiments indicate that food consumption rates of wild animals are determined primarily by energy needs rather than by needs for specific nutrients such as certain amino acids, fatty acids, vitamins, minerals, and/or carbohydrates or by needs for protein, carbohydrates, lipids, and/or vitamins in general. Animals eating diets deficient in specific vitamins, minerals, or amino acids generally do not increase food consumption beyond that needed to satisfy their energy requirements, but new foods become more attractive to them. However, animals deprived of water do develop an appetite for water (thirst), and sodium deprivation leads to "salt hunger" in some species. In general, appetite is influenced primarily by the energy requirements of the animal. This is also an area where research is needed.

3.5 Rates of Feeding

Feeding rates that are actually achieved in the field are influenced not only by intrinsic factors, such as endothermy or ectothermy, activity level, reproductive status, and digestive physiology, but also by extrinsic factors, such as season, food availability, competitive and social interactions with other individuals of the same or other species, and reduction of feeding opportunities due to the presence of predators or inclement environmental conditions (e.g., excessive midday heat, darkness, rainstorms). Over a period of several days or weeks, animals are usually able to compensate for short-term difficulties in getting food and are able to obtain approximately enough food to maintain energy balance. Body size is the most important determinant of feeding rate; bigger animals eat more food than do smaller animals. However, this is not a one-to-one relationship. An animal weighing 10 times more than another animal does not eat 10 times more food each day; rather, it eats only approximately 5 to 6 times more, in accordance with the scaling of metabolic rate. When differences in body mass are accounted for, feeding rates still vary by more than 25 times among species during their activity seasons. For example, a

representative insectivorous lizard that weighs 100 g consumes about 0.7 g of food (dry matter intake [DMI]) per day, whereas a typical 100-g bird living in a marine habitat consumes about 18 g of dry matter per day (Table II). Both may be living in the same seashore habitat, eating a similar diet (arthropods), and maintaining the same body temperature during the day, but the lizard's metabolic energy expenditures over a 24-h period are only about 4% those of the bird, so the lizard's food needs are proportionately lower as well. Within various groups of endothermic vertebrates, there is a 230% variation in feeding rates among same-sized animals. Desert-dwelling mammals and birds have relatively low feeding rates, and marine birds have relatively high feeding rates. Desert mammals and birds are known to have lower metabolic rates (basal and field) than do related species living in other habitats, so their food requirements should be correspondingly lower. The high feeding rates of herbivorous mammals, relative to other mammals, most likely result in part from their low MEE (Table I), meaning that relatively more plant food must be eaten to obtain a given daily rate of metabolizable energy intake.

TABLE II

Feeding Rates of Wild Vertebrates, Summarized in Allometric Equations Derived from Field Measurements of Energy Metabolism, for Various Groups of Mammals, Birds, and Reptiles

Animal group	*a*	*b*	Expected feeding rate (g dry matter/day) for a 100-g animal
Eutherian mammals (58 species)	0.299	0.767	10.2
Marsupial mammals (20 species)	0.483	0.666	10.4
Herbivorous mammals (26 species)	0.859	0.628	15.5
Desert mammals (25 species)	0.192	0.806	7.9
Birds (95 species)	0.638	0.685	15.0
Passerine birds (39 species)	0.630	0.683	14.6
Desert birds (15 species)	0.407	0.681	9.4
Marine birds (36 species)	0.880	0.658	18.2
Reptiles (55 species)	0.011	0.920	0.77
Herbivorous reptiles (9 species)	0.033	0.717	0.91
Insectivorous lizards (27 species)	0.011	0.914	0.73

Note. The equations have the following form: $y = ax^b$, where y = feeding rate (in grams dry matter consumed per day), x = body mass (in grams), a is the intercept at body mass = 1 g, and b is the allometric slope.

4. ENERGY USE

4.1 Metabolic Rate

Oxidative metabolism is the avenue of energy use that powers the daily costs of living (analogous to the overhead of human financial budgets such as housing, food, and transportation). This portion of the energy budget, also known as respiratory metabolism or metabolic rate, ultimately ends up as heat, which is produced primarily inside the animal during biochemical reactions or by friction of moving parts but also outside the animal, mainly by friction with, and work done on, the environment. Although most animals are capable of fueling these energy expenditures by anaerobic metabolism for short periods, either because of a lack of oxygen during breath holding (e.g., forced diving) or because the muscles are called on to spend adenosine triphosphate (ATP) energy faster than it can be produced by oxidative processes, the "oxygen debt" that is built up is paid off ultimately by aerobic processes. However, a few animals live most or all of their lives in anaerobic conditions (e.g., worms in anoxic mud and unstirred water in swampy areas, some intestinal parasites), and anaerobic metabolism is their main source of ATP energy. Nevertheless, most animals have access to oxygen and produce ATP primarily by oxidative metabolism.

Energy needs for oxidative metabolism apparently get the highest priority for metabolizable energy. When an animal is eating extra food, it will either store fat, grow in mass, or reproduce with the extra metabolizable energy (Fig. 2). However, as the feeding rate is diminished, these productive processes decline to zero while the metabolic rate remains nearly constant. At feeding levels below those needed for energy balance (where the rate of metabolizable energy intake = the rate of oxidative metabolic expenditure), production is "negative" in that the animal draws on chemical potential energy stored in its body to make up the deficit between intake from food and metabolic expenses (Fig. 2) and body mass declines. An animal's rate of oxidative metabolism does not remain exactly constant over a range of feeding rates. Metabolic rate increases as food intake increases above the weight maintenance level of intake, apparently due to greater costs of digestion and assimilation and to added

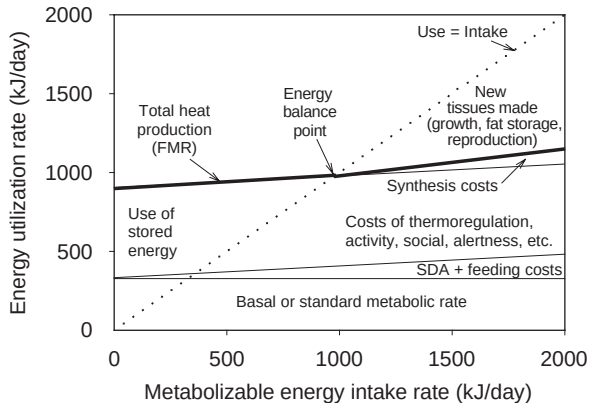


FIGURE 2 Fate of metabolizable energy input at various feeding rates in a representative 1-kg mammal. As daily feeding rate increases, metabolizable energy intake (dotted line) increases to a point where intake rate = utilization rate (energy balance point), and stored energy (e.g., fat) in the body is no longer used to make up the deficit (deficit indicated by the triangular space between dotted line [intake rate] and heavy solid line [metabolic rate] to the left of the energy balance point). At higher intake rates, extra energy intake (indicated by the triangular space above the heavy solid line to the right of the energy balance point) ends up in new biomass (production) either as fat, new somatic tissues, or reproductive tissues. Specific dynamic action (SDA, the cost of digesting food) and feeding costs increase as more food is consumed in a day, and costs of synthesizing new tissues increase as food consumption exceeds the energy balance point.

biosynthesis costs for production. The decline in metabolic rate as the food intake rate drops farther below the maintenance level is due to reduced digestion costs and probably to changes in behavior (reduced activity) and physiology (e.g., lower peripheral and even core body temperatures) that often accompany various degrees of starvation.

Rates of energy metabolism in free-living wild vertebrates are known as field metabolic rates (FMRs) or as daily energy expenditures (DEEs) and are expressed in units of kilojoules per day. There is much variation in FMRs of different animals, with measured values ranging from 0.2 kJ/day ($\sim$2.5 mW given that 1 W = 86.4 kJ/day) in a small African gecko to 67,700 kJ/d (784 W) in a male elephant seal weighing half a metric ton. This difference of nearly six orders of magnitude (10^6) in FMR is due largely to differences in body mass; bigger animals need more energy each day to pay their daily costs of living. FMRs increase exponentially (rather than linearly) with body mass, as do many other physiological and morphological variables. The slopes of the log–log (allometric) relationships for FMRs in various taxonomic, dietary, and habitat groupings of vertebrates range from 0.5 to near 1.0 and average approximately 0.7 (Table III). A slope of

TABLE III

Field Metabolic Rates of Wild Vertebrates, Summarized in Allometric Equations Derived from Field Measurements of Energy Metabolism, for Various Groups of Mammals, Birds, and Reptiles

Animal group	*a*	*b*	Expected FMR (kJ/day) for a 100-g animal
Eutherian mammals (58 species)	4.21	0.772	147
Marsupial mammals (20 species)	10.1	0.590	153
Rodents (30 species)	5.48	0.712	145
Desert rodents (15 species)	9.68	0.487	91
Herbivorous mammals (26 species)	7.94	0.646	156
Birds (95 species)	10.5	0.681	242
Passerine birds (40 species)	10.4	0.680	238
Marine birds (36 species)	14.25	0.659	296
Desert birds (15 species)	6.35	0.671	140
Reptiles (55 species)	0.196	0.889	11.8
Desert lizards (16 species)	0.177	0.935	13.1
Herbivorous reptiles (8 species)	0.232	0.813	9.8

Note. The equations have the following form: $y = ax^b$, where y = feeding rate (in grams dry matter consumed per day), x = body mass (in grams), a is the intercept at body mass = 1 g, and b is the allometric slope.

0.7 indicates that an animal that is 10 times larger than another animal will have an FMR approximately five times greater.

By solving the allometric equations in Table III for a 100-g animal, we can adjust FMRs for body size effects and then compare the FMRs of various taxa. It is obvious that birds and mammals are "high-energy" machines in comparison with reptiles. The endotherms need 15 to 25 times more energy per day to pay their costs of living (overhead) as do lizards of the same body mass. This enormous difference in the daily cost of living is due in part to the much higher concentrations and activities of metabolic enzymes, higher mitochondrial densities and cellular membrane surface areas, and higher thyroid activities in endotherms. It is also due to metabolic responses to the daily body temperature regimes of the two groups. Lizards become cool at night and their metabolism decreases, but endotherms stay warm and their metabolism may increase rather than

decrease at night to provide the needed heat. Among the endotherms, daily costs are similar for eutherian mammals and marsupial mammals but are higher for birds. One group of birds, those that live in marine habitats, has a very high cost of living, being about twice that of similar-sized mammals. As a group, desert-dwelling mammals and birds have relatively low FMRs, and this may benefit them in their sparsely provisioned habitat. FMRs also change with seasons, as indicated by the doubling in FMRs of desert kangaroo rats from summer to winter (Fig. 3A). This is probably related to the higher thermoregulatory costs during the cooler months (Fig. 3B). To the contrary, FMRs of desert tortoises decline during winter because the cold weather inhibits activity in these ectothermic animals and they hibernate during winter (Fig. 3).

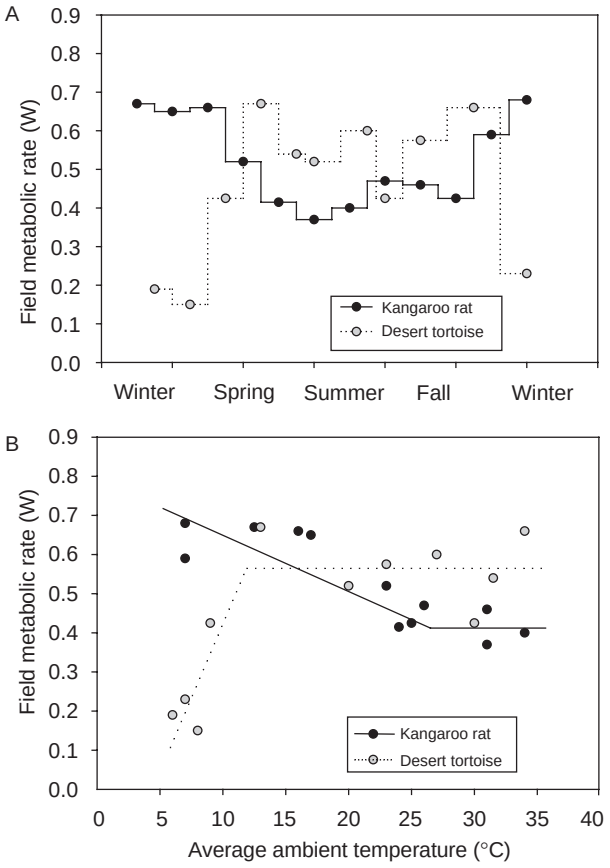


FIGURE 3 Field metabolic rates throughout a year in two desert animals: the kangaroo rat (*Dipodomys merriami*) and the desert tortoise (*Gopherus agassizii*). (A) The endotherm (rat) had higher metabolic rates during the cold seasons than during the warm seasons, but the reverse was true for the ectotherm (tortoise), which had higher FMRs during the warm seasons. (B) The ambient temperature effects are more easily seen when the FMR data are replotted versus average ambient air temperature. Interestingly, the annual energy expenditures of the two species are similar because the tortoise, a "low-energy" ectotherm, is about 30 times larger than the kangaroo rat, a "high-energy" endotherm. Thus, one individual of either species will have a similar impact on the food resources in its habitat.

What determines FMRs? Itemizing daily field energy expenses can reveal the mechanisms underlying daily costs of living and can provide explanations for differences in FMRs.

4.2 Production

Metabolizable energy consumed in excess of that required by oxidative metabolism can be used to create new biomass in the form of somatic growth, storage, or reproduction—or even two or all three simultaneously (Fig. 2). There are two issues here. The first concerns the conditions under which an animal consumes more energy than it needs to maintain energy balance, and the second concerns how extra ingested energy will be allocated. Both issues depend on species of the animal, its age, its sex, its endocrinological status, and the food supply in its habitat, among other things.

Many animal species (e.g., most reptiles, many invertebrates) seem to eat whenever food is available to them even though they may have already eaten enough to meet their immediate metabolic needs. These species have indeterminate growth and continue to increase in size throughout life via somatic growth. Although the extra energy is allocated mainly to growth, fat may be accumulated during times of great food abundance. For example, some snakes have greater fat contents just before winter begins as well as during occasional rodent population explosions, and filter-feeding bivalve mollusks have high fat contents during autumn as well as during algal blooms that may occur in some years and not others. These species do accumulate fat at certain times of the year, usually before seasons of low food availability, but the regulation of mode of production is poorly understood in many species. Some indeterminate growers, such as marine crustaceans, bivalves, and coelenterates, are able to reverse the growth process and become smaller while oxidizing their own somatic tissues during periods of food shortage.

Determinate growers, such as birds and many insects and mammals, have a fixed adult size range, and individuals do not grow much after reaching adult size. In fact, most birds reach adult size before they fledge and leave the nest. For adults of these species, appetite usually limits food intake to just that needed for oxidative metabolism even though food may be abundant in the habitat. In general, wild

vertebrates do not become obese, in contrast to some species of domestic animals and humans. Adults of species having determinate growth tend to have similar body fat contents in the field during a given season, and very fat individuals are rare. Many of these species are capable of increasing appetite and accumulating much adipose tissue, but they do not get fat until just before the period of need (e.g., winter hibernation, migration). One could speculate that obese animals would be quite attractive to predators due to their high energetic reward as well as to their lowered agility and consequent increased vulnerability. If so, selection would act to restrict periods of obesity to times of need only.

During the breeding season, both indeterminate and determinate growers allocate energy-containing chemicals to reproductive tissues. In many species, the males' energy investment in reproductive tissues (testes and accessory gland enlargement and sperm production) is minor compared with that of females. (An exception is found in male crickets and other insect species where males pass a large spermatheca to females, which use some of the chemical energy in the spermatheca to produce eggs.) Depending on species and body size, females may synthesize reproductive tissues (e.g., eggs, embryos, placentas, milk) at very high rates, and their feeding rates may double or even triple (in some species of small rodents) during some phases of the breeding period. For example, a female deer mouse may produce a litter that, at weaning time, may weigh several times her own body mass—all accomplished in a few weeks.

The chemicals needed for synthesizing the new biomass apparently are taken from the pool of substrates circulating in the blood. Hormones (i.e., growth hormone and thyroid hormones in the case of somatic growth, reproductive hormones in the case of reproductive tissues) trigger anabolic metabolism in certain tissues, and this apparently depletes blood substrates. This is presumably detected by the liver and other organs, which then add substrates to the blood to restore homeostasis. On the other hand, fat deposition appears to be a consequence of higher than normal blood levels of nutrients, favoring production and storage of lipids in adipose tissue. This process is influenced by the pancreatic hormones insulin and glucagon. It is not clear whether increased appetite or hormonal stimulation of fat deposition is the primary stimulus for the rapid deposition of fat in some small mammals just before they hibernate and in some birds just before they migrate.

It is possible for both "negative" and "positive" production to occur simultaneously in an animal. For example, lactating female mammals may draw on their own tissues to produce milk for their babies when food intake is inadequate. Similarly, female desert tortoises use protein and fats from their own soma, along with newly consumed foodstuffs, to synthesize eggs during spring.

It is important to recognize that neither food energy supply nor energy intake is the primary determinant of production in many animals. Food supply may be high, but that does not necessarily mean that an animal will consume more and then grow or reproduce. Although fat deposition may occur as a result of a high food intake in response to high food availability, somatic growth (in determinate growers) and reproduction are usually triggered by factors other than energetics, and it is these factors that increase appetite. Appetite seems to be determined largely by hormonal conditions in the animal that are responding to factors such as season, photoperiod, and age and sex of the animal as well as by energetic parameters in the animal and its environment.

In the field, an animal that is producing new biomass is paying not only additional energy for biosynthesis costs but also the extra costs associated with getting the additional food needed for growth. These costs include the extra activity involved in foraging longer, avoiding predators longer, thermoregulating longer, and interacting socially longer while getting the extra food as well as the higher specific dynamic action required to process the extra food. These added costs should also be considered as costs of production.

5. EFFICIENCY

Not all of the chemical potential energy in the food can be used by animals because some is indigestible and is lost as feces, and some that is digested and absorbed leaves the body as urinary wastes (which in some animals is mixed inseparably with feces when voided). The fraction of the food's energy that is absorbed across the gut and not lost in feces is called assimilated energy, and the percentage of dietary energy that this amounts to is the apparent AE. Apparent MEE, or the percentage of dietary energy that is usable for metabolism or production, is assimilated energy corrected for urinary energy losses. Production efficiency can be expressed in a variety of ways. Here, it is given as net production efficiency (NPE), which is the percentage of metabolizable energy that appears as chemical potential energy in new biomass (e.g., fat, growth, eggs, embryos, milk).

TABLE IV

Typical Energetic Efficiencies (Percentages) in Animals Having Various Diet

Diet type	Assimilation efficiency	Metabolizable energy efficiency	Net production efficiency
Herbivore, monogastric	30–77	22–73	20–55
Herbivore, ruminant	60–75	50–70	23
Herbivore, hindgut fermenter	60–72	55–68	23
Granivore	77–90	74–81	22–24
Carnivore	70–88	62–81	30–58
Omnivore	80	74	23–58
Lactivore	95	87	45
Detritivore	12–44	12–42	53
Parasites	77–100	71–95	50

Note. AE (assimilation efficiency) = [(I−L) / I]100; MEE (metabolizable energy efficiency) = [(I−L−U)/I)]100; NPE (net production efficiency) = [P/(I−L−U)]100; I, food energy intake; L, feces energy loss; U, urinary energy loss, P, energy allocated to production of biomass.

AE and MEE vary greatly among diets (Table IV). This variation is due mainly to the relatively indigestible organic components of the various diets: cellulose in plants, keratin in vertebrates, and chitin in arthropods. Within the herbivores, the type of digestive system an animal has also influences AE. For a given diet, herbivores that ferment foods, either at the beginning of the gut in specialized stomachs (e.g., ruminants) or at the end of the gut (cecum or colon fermenters), can digest and assimilate more of that food's energy. These fermenters obtain some of the energy in cellulose, which is broken down by cellulases produced by anaerobic microbes in their fore- or hindgut. This allows ruminants to live on lower quality plant foods, such as dry grasses (straw), and reduces competition for food among herbivore types. Monogastric herbivores often choose high-quality foods (e.g., young leaves, flowers, buds) when available, and they may attain high AE and MEE (up to 73–77% [Table IV]) by diet selection, thereby avoiding the low AE (22–30%) they would have if they ate dry grasses.

Young suckling mammals (lactivores), parasites, and animals that eat seeds (granivores), nectar (nectarivores), or other animals (carnivores) all have relatively high AE and MEE. Coevolution of plants along with the animals that pollinate their flowers and disperse their seeds probably has favored their cooperation and the ready satisfaction of those animals' energy needs from those plants' products. On the other hand, herbivores are essentially predators on plants, and evolution has favored plants that developed defenses against herbivory such as spines, woody tissues, poisons, crystalline inclusions (silica spines), and antifeedant chemicals (secondary plant compounds). Moreover, plant foods often are unbalanced in regard to the nutrient proportions (e.g., energy, protein, minerals, vitamins) required by the consumer. Thus, some herbivores have extraordinary modifications of their feeding and digestive apparati that improve their ability to survive on vegetation (e.g., plant toxin avoidance or resistance by some caterpillars and beetles, coprophagy by rabbits, rumination by sheep and antelope). Animals that eat other animals (carnivores and parasites) do not face the problems of food quality that the herbivores face because the composition of these animals' food is essentially the same as their own. However, carnivores often are challenged by problems of food quantity in that their prey may be hard to obtain.

Production efficiency can be measured only when an animal is producing new biomass, and this may be a fleeting event in the lives of some wild animals, especially birds and many mammals and insects. The values for NPE in Table IV are mainly from domestic, farm, or captive wild animals, and they represent near maximum efficiencies that occur during peak production periods. Efficiencies will be lower in animals that forage for their own food because extra energy is required to obtain the extra food needed to support growth. Nevertheless, some generalizations seem warranted from studies on domesticated animals. Ectotherms generally have higher peak NPE than do endotherms, apparently due to the higher metabolic expenditures for oxidative metabolism paid by endotherms. Where a range of values for NPE is given in Table IV, the highest efficiencies are for ectotherms and the lowest are for endotherms. Diet type has a large effect on NPE, with high-protein foods having lower efficiencies than high-fat

and high-starch diets. Finally, the type of production occurring influences NPE, with milk production and somatic growth having efficiencies substantially higher than those for fattening. There is still controversy about whether NPE declines at higher rates of food intake, so the "law of diminishing returns," derived from experiments on adding increasing amounts of fertilizer to crops, might not be applicable to animal energetics.

Studies on five species of farm animals, selected for good production capabilities, show a modest and surprisingly uniform maximum rate of food energy consumption (maximum ME intake = 4.4–5.7 BMR). A subsequent evaluation of maximum observed feeding rates in domestic and wild mammals (8 species) and birds (11 species) indicates that maximum ME intake ranges from three to six times BMR. Thus, animals may limit their food intake to rates that their guts can deal with effectively. The role of appetite is clearly critically important but still poorly understood for wild animals.

Energetic efficiency values are useful for evaluating the quality of various foods and for identifying feeding and digestive adaptations of animals eating a variety of food types. It is tempting to hypothesize that animals should be selected for having higher efficiencies, but this is unlikely for three reasons. First, efficiency is a ratio of two rate processes, and there is no evidence that natural selection responds to such an abstract derived variable as a ratio. Second, some important efficiencies are already near the maximum of 100%, leaving little variation on which selection might operate. Third, an animal can increase its daily rate of metabolizable energy intake much more easily and dramatically by eating more food and keeping AE constant than by keeping feeding rate constant and improving its digestive physiology enough to increase MEE substantially.

6. REGULATION OF ENERGY METABOLISM

Most physiological processes and systems are closely regulated, so it is reasonable to ask whether and how oxidative metabolism of animals is regulated. Metabolic rates of marine and aquatic invertebrates and water-breathing vertebrates may be regulated by oxygen availability in their environments. Metabolism of terrestrial and aquatic invertebrates and ectothermic vertebrates that live in cold climates may be reduced dramatically when these animals freeze during winter. For air-breathing vertebrates (birds, mammals, reptiles, and most amphibians), current knowledge indicates that aspects of energy budgets (e.g., thermoregulation, reproductive output) are regulated but that overall daily energy expenditure (FMR) itself is probably not directly controlled. Instead, the FMR appears to be a consequence of its component parts. For example, an ectothermic vertebrate has a total daily expense for its resting metabolic costs that will be determined by its daily body temperature excursion, which is most likely determined by the animal's minute-by-minute decisions to hide, forage, bask, display, or escape predators and not by some metabolic control center in its central nervous system. This animal may spend additional energy for activity associated with basking, feeding, interacting, and escaping. These costs are also consequences of other phenomena, as are the added costs of digesting and processing food, producing new biomass, and remaining alert and responsive while above ground. Thus, the total daily expense (FMR) seems to be the sum of costs incurred for tasks not directly related to energy metabolism. The situation for an endotherm is similar except that its minimum metabolic rate (BMR) is fixed and its thermoregulatory costs are determined by ambient temperatures.

Animals do some things that may reflect a crude ability to regulate (mainly reduce) FMRs. Ectotherms may choose to remain inactive in cool retreats for days, especially during periods of environmental stress (e.g., drought, low food availability), and this behavior reduces the daily use of energy and other nutrients. Similarly, some small endotherms may use torpor or hibernate, and their low body temperatures reduce FMRs greatly, but this also renders them relatively unresponsive and vulnerable to predation. However, for normally active animals, there is no known mechanism that might sense the rate of energy expenditure or the accumulated daily total energy expenditure and then limit the amount allocated to thermoregulation, activity, digestion, and the like during the rest of the day (regulate output). Instead, it seems that animals spend whatever energy is called for to get through the day and then regulate intake, by means of altered appetite, to achieve energy balance. Similarly, energy allocation to growth and reproduction appears to be a consequence of hormonal changes associated with regulation of those processes, with appetite being the "cart" following the "horses" of growth and reproduction.

Even though the rate of oxidative metabolism might not be under direct physiological control, there are upper limits that constrain energy expenditures.

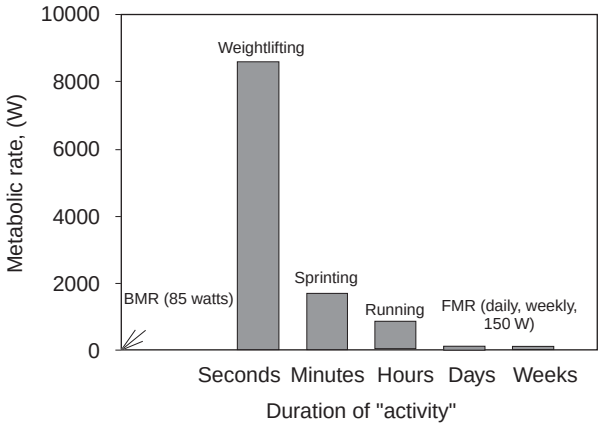


FIGURE 4 Rates of energy expenditure achieved by human athletes during bouts of activity of various kinds. Weightlifters can use ATP very quickly during a lift (100 times their BMR), but this cannot be sustained for very long due to buildup of oxygen debt. Sprinters also use anaerobic sources of ATP and cannot sustain sprinting for longer than a few minutes. Runners (at 20 times BMR) sustain their activity aerobically but are not maintaining energy balance while running because their intake is zero. Maximum sustainable metabolic rates, where humans are in energy balance, are only 5 times BMR. Typical FMRs of nonathlete humans are 1.5 to 2.5 times BMR.

The level of these limits depends on the amount of time involved in the determination of the limit. For example, a champion weightlifter can put out power during a lift that is more than 100 times his BMR, but this effort draws on and uses up energy-containing chemicals and oxygen stored in muscles, so it cannot be sustained for more than a few seconds (Fig. 4). Similarly, sprinters use aerobic as well as anaerobic metabolism to power their running, so sprinting cannot be sustained for more than a few minutes, but sprinters can run faster (use energy faster) than can long-distance runners, who are entirely aerobic. However, continuous running is not a sustainable activity either due to the needs to eat and sleep periodically. The highest sustainable FMR measured in humans is approximately 480 W (in Tour de France cyclists) or approximately 5.6 times BMR. Maximums measured in wild animals are approximately 6.9 times BMR in a small marsupial mammal, 5.0 times BMR in a rodent (eutherian mammal), 6.7 times BMR in a breeding seabird, and 4.6 times estimated field SMR in a small reptile.

7. SUMMARY

Chemical potential energy, in the form of food, is a centrally important resource for animals (heterotrophs) and has been termed the "primary currency of life." Knowledge of the ways in which animals gain, use, and store energy, and of their rates or "flows," may reveal much about the biology of various species, and such knowledge is essential in understanding energy flows in populations and through communities. The input side of a chemical potential energy budget is simple, being made up only of food intake. Energy outputs, corrected for loss of chemical energy in excreta, include heat resulting from oxidative metabolism and biomass (production of new tissues in the form of adipose tissue [fat storage], somatic growth, and reproductive tissues [sperm, embryos, eggs, placentas, and milk]). Different kinds of foods yield widely different amounts of energy; soil and detritus contain the least, plant tissues are intermediate, animal matter is higher, and seeds and nectar are highest. However, the availability of these food types in the field is not uniform in space or time. Endothermic animals require much more (15–25 times more) energy each day than do same-sized ectothermic animals. Large animals need more food energy than do small animals, but the increased food requirement is only about half that expected from the difference in body masses (a 10-fold mass increase requires only a 5- to 6-fold food intake increase). Most or all of the available food energy (metabolizable energy) is used for oxidative metabolism (metabolic rate), which has highest priority. Excess metabolizable energy can be used for production, but many animals do not consume excess food, despite its availability, except during certain seasons (e.g., breeding, premigratory, prehibernation). Daily energy needs appear to be influenced mainly by energy use for resting costs, temperature regulation, and activity. Although there appear to be upper and lower limits to rates of energy expenditure and food intake, actual daily energy needs seem to be determined more by lifestyle, season, and local circumstances than by precise regulation of energetic parameters themselves. Digestibility, AE, and MEE are determined mainly by diet type, although there is much variation among digestive capabilities of different herbivores. Net production efficiency is usually higher in ectotherms than in endotherms. Although daily energy expenditures by animals have operational upper and lower limits, actual expenditures do not seem to be regulated directly; rather, they are subservient to other aspects of the animals' biology such as thermoregulation needs, foraging expenses, reproductive and migration costs, and costs of interactions with other organisms (e.g., conspecifics, predators) in

their surroundings. Appetite, a poorly understood factor in wild animals, apparently matches income with expenditures, so that most animals are approximately in energy balance much of the time. The evolution and success of birds and mammals, which have extraordinarily high costs of living compared with those of ectotherms, still needs a compelling explanation.

SEE ALSO THE FOLLOWING ARTICLES

Conversion of Energy: People and Animals • *Earth's Energy Balance* • *Ecosystems and Energy: History and Overview* • *Food Capture, Energy Costs of* • *Human Energetics* • *Migration, Energy Costs of* • *Photosynthesis and Autotrophic Energy Flows* • *Reproduction, Energy Costs of* • *Thermoregulation*

Further Reading

Blaxter, K. (1989). "Energy Metabolism in Animals and Man." Cambridge University Press, Cambridge, UK.

Kleiber, M. (1975). "The Fire of Life." R. E. Krieger, Huntington, NY.

Louw, G. N. (1993). "Physiological Animal Ecology." Longman Scientific and Technical, Essex, UK.

McNab, B. K. (2002). "The Physiological Ecology of Vertebrates: A View from Energetics." Cornell University Press, Ithaca, NY.

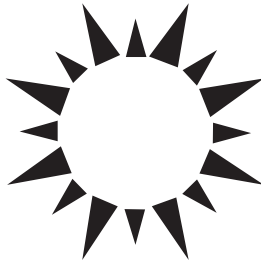
Nagy, K. A. (2001). Food requirements of wild animals: Predictive equations for free-living mammals, reptiles, and birds. *Nutr. Abstr. Rev.* **71,** R21–R32.

Nagy, K. A., Girard, I. A., and Brown, T. K. (1999). Energetics of free-ranging mammals, reptiles, and birds. *Annu. Rev. Nutr.* **19,** 247–277.

Robbins, C. T. (1993). "Wildlife Feeding and Nutrition." 2nd ed. Academic Press, San Diego.

Human Energetics

WILLIAM R. LEONARD
Northwestern University
Evanston, Illinois, United States



1. Introduction
2. Calorimetry
3. Sources of Food Energy
4. Determinants of Daily Energy Needs
5. Human Variation in Energy Dynamics

Glossary

aerobic respiration (metabolism) Energy producing reactions that require oxygen (O_2).
basal metabolic rate (BMR) The minimum level of energy required to sustain basic biological function of the body.
body mass index (BMI) A weight-for-height index used to assess physical nutritional health, calculated as weight(kg)/height(meters)^2.
calorimetry The measurement of heat transfer.
kilocalorie (kcal) The amount of heat required to raise the temperature of 1 kilogram (or 1 liter) of water by 1°C. This is the unit most commonly used to measure the energy content of food and human energy requirements.
macronutrient Chemical components of food that are used as sources of energy. These include carbohydrates, fats, and proteins.
physical activity level (PAL) The ratio of total daily energy expenditure to basal metabolic rate (TDEE:BMR).
total daily energy expenditure (TDEE) The total energy requirements of a person over the course of an entire day.

The dynamic relationship between energy intake and expenditure has important consequences for human health, survival, and reproduction. Human energy requirements are determined by a variety of factors, including age, sex, body size, and activity levels. Energy availability, on the other hand, is strongly shaped by one's social, economic, and physical environments. Worldwide variation in nutritional health directly reflects differences in long-term energy balance. Among rural populations of the developing world, conditions of negative energy balance and undernutrition are common due to high levels of daily energy expenditure and marginal food availability. Conversely, in the industrialized world rates of obesity and chronic metabolic diseases are increasing at an alarming rate due to the abundance of dietary energy and increasingly sedentary lifestyles.

1. INTRODUCTION

Sufficient dietary energy is essential to the survival and health of all animals. From the perspective of human biology and health, energy is particularly important for a number of reasons. First, food and energy represent critical points of interaction between humans and their environment. The environments in which we live determine the range of food resources that are available and how much energy and effort are necessary to procure those resources. Indeed, the dynamic between energy intake and energy expenditure is quite different for a subsistence farmer of Latin America than it is for an urban executive living in the United States. Beyond differences in the physical environment, social, cultural, and economic variations also shape aspects of energy balance. Social and cultural norms are important for shaping food preferences, whereas differences in subsistence behavior and socioeconomic status strongly influence food availability and the effort required to obtain food.

Additionally, the balance between energy expenditure and energy acquired has important adaptive consequences for both survival and reproduction. Obtaining sufficient food energy has been an important stressor throughout human evolutionary history, and it continues to strongly shape the biology of traditional human populations.

 © 2004 Elsevier Inc. All rights reserved.

This article examines aspects of energy expenditure and energy intake in humans. First it explores the principles of calorimetry, the methods used to measure energy expenditure in humans and other animals. The discussion then examines sources of energy in our diets, the macronutrients—carbohydrates, fats, and proteins. Next, the of the physiological basis of variation in human energy requirements is evaluated, specifically the different factors that determine how much energy a person must consume to sustain himself of herself. Finally, the discussion considers patterns of variation in energy expenditure and intake among modern human populations, highlighting the different strategies that humans use to fulfill their dietary energy needs and the health consequences associated with the increasingly sedentary lifeways of modern society.

2. CALORIMETRY

The study of energy relies on the principle of calorimetry, the measurement of heat transfer. In food and nutrition, energy is most often measured in kilocalories (kcal). One kilocalorie is the amount of heat required to raise the temperature of 1 kilogram (or 1 liter) of water, 1°C. Thus, a food item containing 150 kcal (two pieces of bread, for example) contains enough stored chemical energy to increase the temperature of 150 liters of water by 1°C. Another common unit for measuring energy is the joule or the kilojoule [1 kilojoule (kJ) = 1000 joules]. The conversion between calories and joules is 1 kcal is equal to 4.184 kilojoules.

Techniques for measuring energy expenditure involve either measuring heat loss directly (direct calorimetry) or measuring a proxy of heat loss such as oxygen consumption (O_2) or carbon dioxide (CO_2) production (indirect calorimetry). Direct calorimetry is done under controlled laboratory conditions in insulated chambers that measure changes in air temperature associated with the heat being released by a subject. Although quite accurate, direct calorimetry is not widely used because of its expense and technical difficulty.

Thus, methods of indirect calorimetry are more commonly used to quantify human energy expenditure. The most widely used of these techniques involve measuring oxygen consumption. Because the body's energy production is dependent on oxygen (aerobic respiration), O_2 consumption provides an accurate indirect way of measuring a person's energy expenditure. Every liter of O_2 consumed by the body is equivalent to energy cost of approximately 5 kcal. Consequently, by measuring O_2 use while a person is performing a particular task (for example, standing, walking, running on a treadmill), the energy cost of the task can be determined.

Figures 1 and 2 show alternative methods for measuring of oxygen consumption. Figure 1 shows the classic Douglas bag method for measuring O_2 uptake. With this technique, subjects breathe through a valve that allows them to inhale room air and exhale into a large collection bag. The volume and the O_2 and CO_2 contents of the collected air sample are then measured to determine the total amount of oxygen consumed by the subject. Recent advances in computer technology allow for the

FIGURE 1 Measurement of resting energy expenditure using the Douglas Bag method of indirect calorimetry. The subject breathes into the collection bag for a specified amount of time. The volume of the collected air sample and its oxygen and carbon dioxide content will then be measured to determine the amount of energy the subject has used.

FIGURE 2 Measurement of energy expenditure using a computerized metabolic system (Aerosport TEEM 100). As with the Douglas bag method, this system measures the volume of expired air along with its oxygen and carbon dioxide contents to determine rates of energy expenditure.

determination of O_2 consumption more quickly without directly collecting the expired air samples. Figure 2 shows an example of a computerized system for measuring oxygen consumption. As with the Douglas bag method, this system determines energy costs by measuring the volume and the O_2 and CO_2 concentrations of expired air samples.

The most recently developed method of indirect calorimetry is the doubly labeled water (DLW) technique that involves the use of stable isotopes. With the DLW method, subjects drink a small amount of water that has been enriched with isotopes of both hydrogen (^2H, deuterium) and oxygen (^{18}O). After the labeled water has been consumed, subjects then provide urine samples for the subsequent 10 to 14 days. Over this period, the concentrations of both the ^2H and ^{18}O isotopes decrease, because they are being lost in the products of aerobic respiration, H_2O and CO_2. The concentration of ^{18}O in the body decreases more rapidly than that of ^2H, because the oxygen is contained in both metabolic waste products (H_2O and CO_2), whereas hydrogen is contained in only one (H_2O). Figure 3 shows a graph of the rates of decline (washout rates) of the two isotopes over a 14-day period. The differences in the washout rates provide an accurate measurement of CO_2 production and O_2 consumption over the period of study. These estimates of CO_2 production and O_2 consumption are then converted to kilocalories to provide a measure of total energy expenditure of the subject over the entire study period.

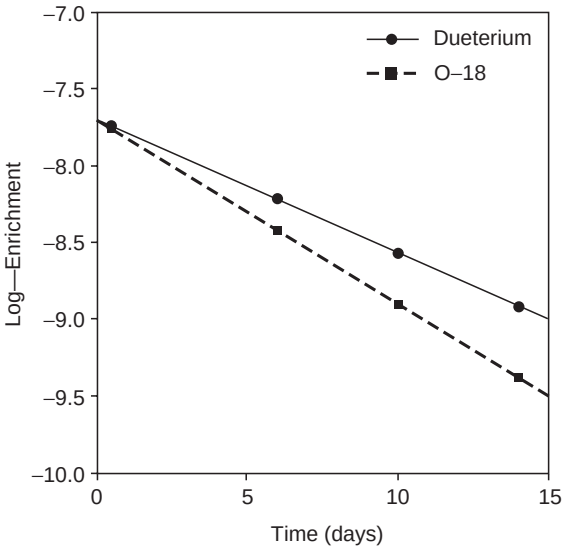


FIGURE 3 Graph showing the rates of decline (washout rates) in the two key isotopes (^{18}O and ^2H [deuterium]) used to measure daily energy expenditure in the "doubly labeled water" method. The difference in the washout rates of the two isotopes provides a measure of carbon dioxide (CO_2) production, which allows for the determination of energy expenditure.

Validation studies of the DLW method against direct calorimetry have shown that the technique is quite accurate (± 2 to 3% error for individuals). Hence, since the late 1980s, DLW has become the most accurate technique for measuring human energy expenditure in real-world (i.e., nonlaboratory) situations. The main limitations of the technique remain its expense, about $800 to $1000 per subject.

3. SOURCES OF FOOD ENERGY

The main chemical sources of energy in our foods are carbohydrates, protein, and fats. Collectively these three energy sources are known as macronutrients. Vitamins and minerals (micronutrients) are required in much smaller amounts and are important for regulating many aspects of biological function.

Carbohydrates and proteins have similar energy contents; each provides ~ 4 kilocalories of metabolic energy per gram. In contrast, fat is more calorically dense; each gram provides about 9 to 10 kilocalories. Alcohol, although not a required nutrient, also can be used as an energy source, contributing ~ 7 kcal/g. Regardless of the source, excess dietary energy can be stored by the body as glycogen (a carbohydrate) or as fat. Humans have relatively limited glycogen stores (about 375 to 475 grams), in the liver and muscles. Fat, however, represents a much large source of stored energy, accounting for approximately 13 to 20% of body weight in men and 25 to 28% in women.

3.1 Carbohydrates

The largest source of dietary energy for most humans is carbohydrates (~ 45 to 50% of calories in the typical American diet). The three types of carbohydrates are monosaccharides, disaccharides, and polysaccharides. Monosaccharides, or simple sugars, include glucose, the body's primary metabolic fuel; fructose (fruit sugar); and galactose. Disaccharides, as the name implies, are sugars formed by a combination of two monosaccharides. Sucrose (glucose and fructose), the most common disaccharide, is found in sugar, honey, and maple syrup. Lactose, the sugar found in milk, is composed of glucose and galactose. Maltose (glucose and glucose), the least common of the disaccharides, is found in malt products and germinating cereals. Polysaccharides,

or complex carbohydrates, are composed of three or more simple sugar molecules. Glycogen is the polysaccharide used for storing carbohydrates in animal tissues. In plants, the two most common polysaccharides are starch and cellulose. Starch is found in a wide variety of foods, such as grains, cereals, and breads, and provides an important source of dietary energy. In contrast, cellulose—the fibrous, structural parts of plant material—is not digestible to humans and passes through the gastrointestinal tract as fiber.

3.2 Fats

Fats provide the largest store of potential energy for biological work in the body. They are divided into three main groups: simple, compound, and derived. The simple, or neutral fats, consist primarily of triglycerides. A triglyceride consists of two component molecules: glycerol and fatty acid. Fatty acid molecules, in turn, are divided into two broad groups: saturated and unsaturated. These categories reflect the chemical bonding pattern between the carbon atoms of the fatty acid molecule. Saturated fatty acids have no double bonds between carbons, thus allowing for the maximum number of hydrogen atoms to be bound to the carbon (that is, the carbons are saturated with hydrogen atoms). In contrast, unsaturated fatty acids have one (monounsaturated) or more (polyunsaturated) double bonds. Saturated fats are abundant in animal products, whereas unsaturated fats predominate in vegetable oils.

Compound fats consist of a neutral fat in combination with some other chemical substance (for example, a sugar or a protein). Examples of compound fats include phospholipids and lipoproteins. Phospholipids are important in blood clotting and insulating nerve fibers, whereas lipoproteins are the main form of transport for fat in the bloodstream.

Derived fats are substances synthesized from simple and compound fats. The best known derived fat is cholesterol. Cholesterol is present in all human cells and may be derived from foods (exogenous) or synthesized by the body (endogenous). Cholesterol is necessary for normal development and function because it is critical for the synthesis of such hormones as estradiol, progesterone, and testosterone.

3.3 Proteins

Proteins, in addition to providing an energy source, are also critical for the growth and replacement of living tissues. They are composed of nitrogen-containing compounds known as amino acids. Of the 20 different amino acids required by the body, 9 (leucine, isoleucine, valine, lysine, threonine, methionine, phenylalanine, tryptophan, and histidine) are known as essential because they cannot be synthesized by the body and thus must be derived from food. Two others, cystine and tyrosine, are synthesized in the body from methionine and phenylalanine, respectively. The remaining amino acids are called nonessential because they can be produced by the body and need not be derived from the diet.

4. DETERMINANTS OF DAILY ENERGY NEEDS

A person's daily energy requirements are determined by several different factors. The major components of an individual's energy budget are associated with resting or basal metabolism, activity, growth, and reproduction. Basal metabolic rate (BMR) represents the minimum amount of energy necessary to keep a person alive. Basal metabolism is measured under controlled conditions while a subject is lying in a relaxed and fasted state.

In addition to basal requirements, energy is expended to perform various types of work, such as daily activities and exercise, digestion and transport of food, and regulating body temperature. The energy costs associated with food handling (that is, the thermic effect of food) make up a relatively small proportion of daily energy expenditure and are influenced by amount consumed and the composition of the diet (for example, high-protein meals elevate dietary thermogenesis). In addition, at extreme temperatures, energy must be spent to heat or cool the body. Humans (unclothed) have a thermoneutral range of 25 to 27°C (77 to 81°F). Within this temperature range, the minimum amount of metabolic energy is spent to maintain body temperature. Finally, during one's lifetime, additional energy is required for physical growth and for reproduction (for example, pregnancy, lactation).

In 1985, the World Health Organization (WHO) presented recommendations for assessing human energy requirements. The procedure used for determining energy needs involves first estimating BMR from body weight on the basis of predictive equations developed by the WHO. These equations are presented in Table I. After estimating BMR, the total daily energy expenditure (TDEE) for adults (18

TABLE I

Equations for Predicting Basal Metabolic Rate (BMR) Based on Body Weight (Wt, in Kilograms)

	BMR (kcal/day)	
Age (years)	Males	Females
0–2.9	60.9 (Wt)−54	61.0 (Wt)−51
3.0–9.9	27.7 (Wt)+495	22.5 (Wt)+499
10.0–17.9	17.5 (Wt)+651	12.2 (Wt)+746
18.0–29.9	15.3 (Wt)+679	14.7 (Wt)+496
30.0–59.9	11.6 (Wt)+879	8.7 (Wt)+829
60+	13.5 (Wt)+487	10.5 (Wt)+596

Source. FAO/WHO/UNU (1985).

TABLE II

Physical Activity Levels (PALs) Associated with Different Types of Occupational Work Among Adults (18 Years and Older)

	PAL			
Sex	Minimal	Light	Moderate	Heavy
Male	1.40	1.55	1.78	2.10
Female	1.40	1.56	1.64	1.82

Source. FAO/WHO/UNU (1985).

years old and older) is determined as a multiple of BMR, based on the individual's activity level. This multiplier, known as the physical activity level (PAL) index, reflects the proportion of energy above basal requirements that an individual spends over the course of a normal day. The PALs associated with different occupational work levels for adult men and women are presented in Table II. The WHO recommends that minimal daily activities such as dressing, washing, and eating are commensurate with a PAL of 1.4 for both men and women. Sedentary lifestyles (for example, office work) require PALs of 1.55 for men and 1.56 for women. At higher work levels, however, the sex differences are greater. Moderate work is associated with a PAL of 1.78 in men and 1.64 in women, whereas heavy work levels (for example, manual labor, traditional agriculture) necessitate PALs of 2.10 and 1.82 for men and women, respectively.

In addition to the costs of daily activity and work, energy costs for reproduction also must be considered. The WHO recommends an additional 285 kcal/day for women who are pregnant and an additional 500 kcal/day for those who are lactating.

TABLE III

Energy Constants and PALs Recommended for Estimating Daily Energy Requirements for Individuals Under the Age of 18 Years Old

Age (years)	Males	Females
	Energy constant (kcal/kg body weight)	
<1.0	103	103
1.0–1.9	104	108
2.0–2.9	104	102
3.0–3.9	99	95
4.0–4.9	95	92
5.0–5.9	92	88
6.0–6.9	88	83
7.0–7.9	83	76
8.0–8.9	77	69
9.0–9.9	72	62
	PAL	
10.0–10.9	1.76	1.65
11.0–11.9	1.72	1.62
12.0–12.9	1.69	1.60
13.0–13.9	1.67	1.58
14.0–14.9	1.65	1.57
15.0–15.9	1.62	1.54
16.0–16.9	1.60	1.52
17.0–17.9	1.60	1.52

Source. FAO/WHO/UNU (1985); James and Schofield (1990).

Energy requirements for children and adolescents are estimated differently because extra energy is necessary for growth and because relatively less is known about variation in their activity patterns. For children and adolescents between 10 and 18 years old, the WHO recommends the use of age- and sex-specific PALs. In contrast, energy requirements for children under 10 years old are determined by multiplying the child's weight by an age- and sex-specific constant. The reference values for boys and girls under 18 years old are presented in Table III.

5. HUMAN VARIATION IN ENERGY DYNAMICS

5.1 Energy Consumption

Compared to most other mammals, humans are able to survive and flourish eating a remarkably wide range of foods. Human diets range from completely vegetarian (as observed in many populations of South Asia) to ones based almost entirely on meat

and animal foods (for example, traditional Eskimo/Inuit populations of the Arctic). Thus, over the course of our evolutionary history, humans have developed a high degree of dietary plasticity. This ability to utilize a diverse array of plant and animal resources for food is one of the features that allowed humans to spread and colonize ecosystems all over the world.

Table IV presents information on the percentage of energy derived from plant and animal foods for subsistence-level (that is, food producing) and industrial human societies. The relative contribution of animal foods varies considerably, ranging from $<10\%$ of dietary energy in traditional farming communities of tropical South America, to more than 95% among traditionally living Inuit hunters of the Canadian Arctic.

Subsistence-level agricultural populations, as a group, have the lowest consumption of animal foods. Among hunting and gathering populations, the contribution of animal foods to the diet is variable, partly reflecting the environments in which these populations reside. For example, the !Kung San, who live in arid desert environments of southern Africa, have among the lowest levels of animal food consumption among hunter-gatherers. In contrast, hunters of the arctic rely almost entirely on animal foods for their daily energy. Foragers living in forest and grassland regions of the tropics (for example, the Ache and the Hiwi) have intermediate levels of animal consumption.

The amount of energy that humans consume also varies widely. Table V shows per capita energy consumption (kcal/person/day) for selected countries around the world. Daily energy intakes in the United States (3699 kcal) and France (3518 kcal) are more than double those found in Afghanistan (1745 kcal). Overall, individuals from industrialized countries consume about 55% more energy per day than their counterparts from the developing world (3311 versus 2140 kcal/day).

These broad worldwide differences in daily energy consumption are mirrored in differences in under- and over-nutrition. As discussed Section 5.3, under-nutrition remains a persistent problem among populations in the developing world, despite the fact that worldwide food production is sufficient to meet the energy needs of all people. In contrast, populations of the industrialized world, most

TABLE IV

Percentage of Dietary Energy Derived from Animal and Plant Foods in Selected Human Populations

Population	Percentage energy from animal foods (%)	Percentage energy from plant foods (%)
Hunter-gatherers		
!Kung (Botswana)	33	67
Ache (Paraguay)	56	44
Hiwi (Venezuela)	68	32
Inuit (Canada)	96	4
Pastoralists		
Turkana (Kenya)	80	20
Evenki (Russia)	31	69
Agriculturalists		
Quechua (highland Peru)	5	95
Coastal Ecuador	7	93
Yapú (lowland Colombia)	11	89
Industrial societies		
France	38	62
Japan	20	80
United Kingdom	31	69
United States	27	73

TABLE V

Per Capita of Dietary Energy Supply (kcal/Person/Day) for Selected Countries

Country	Per capita energy supply (kcal/day)
Developing countries	
Afghanistan	1745
Bolivia	2174
Botswana	2183
Columbia	2597
Ethiopia	1858
Haiti	1869
Pakistan	2476
Phillipines	2366
Tanzania	1995
Average	**2140**
Industrialized countries	
France	3518
Canada	3119
Israel	3278
Japan	2932
Norway	3357
Spain	3310
United Kingdom	3276
United States	3699
Average	**3311**

Source. Millstone and Lang (2003).

notably the United States, are now heavier and fatter than ever before. Just as there is a dramatic economic divide between rich and poor populations around the world, there is also a stark nutritional divide.

5.2 Energy Expenditure

Humans also show considerable variation in levels of energy expenditure. Recent comparative analyses indicate that daily energy expenditure in human groups typically ranges from 1.2 to 5.0 x BMR (that is, PAL = 1.2–5.0). The lowest levels of physical activity, PALs of 1.20 to 1.25, are observed among hospitalized, and nonambulatory populations. In contrast, the highest levels of physical activity (PALs of 2.5 to 5.0) have been observed among elite athletes and soldiers in combat training. Within this group, Tour de France cyclists have the highest recorded daily energy demands of 8050 kcal/day (a PAL = 4.68)!

Table VI presents data on body weight, total daily energy expenditure, and PALs of adult men and women from selected humans groups. Men of the subsistence-level populations (that is, foragers, herders, and agriculturalists) are, on average, 20 kilograms (45 pounds) lighter than their counterparts from the industrialized world and yet have similar levels of daily energy expenditure (2897 versus 2859 kcal/day). The same pattern is true for

TABLE VI

Weight (kg), Total Daily Energy Expenditure (TDEE; kcal/Day), Basal Metabolic Rate (BMR; kcal/Day), and Physical Activity Level (PAL) of Selected Human Groups

Group	Sex	Weight (kg)	TDEE (kcal/day)	BMR (kcal/day)	PAL (TDEE/BMR)
Industrialized populations					
18–29 years	M	75.6	3298	1793	1.84
	F	69.2	2486	1480	1.68
30–39 years	M	86.1	3418	1960	1.74
	F	67.9	2390	1434	1.67
40–64 years	M	77.0	2749	1673	1.64
	F	70.0	2342	1386	1.69
65–74 years	M	76.4	2629	1650	1.59
	F	60.2	2055	1267	1.62
75 and older	M	72.6	2199	1434	1.53
	F	48.3	1458	980	1.48
Average	M	77.5	2859	1702	1.67
	F	63.1	2146	1309	1.63
Subsistence populations					
!Kung foragers	M	46.0	2319	1383	1.68
	F	41.0	1712	1099	1.56
Ache foragers	M	59.6	3327	1531	2.17
	F	51.8	2626	1394	1.88
Inuit hunters	M	65.0	3010	1673	1.80
	F	55.0	2350	1305	1.80
Evenki pastoralists	M	58.4	2681	1558	1.75
	F	52.7	2067	1288	1.63
Aymara agriculturalists	M	54.6	2713	1355	2.00
	F	50.5	2376	1166	2.03
Highland Ecuador, agriculturalists	M	61.3	3810	1601	2.38
	F	55.7	2460	1252	1.96
Costal Ecuador, agriculturalists	M	55.6	2416	1529	1.58
	F	47.8	1993	1226	1.63
Average	M	57.2	2897	1519	1.90
	F	50.6	2227	1247	1.78

women; those from subsistence-level populations are 12.5 kilograms (~28 pounds) lighter than women of industrialized societies but have higher levels of daily energy expenditure (2227 versus 2146 kcal/day).

Thus, when we express daily energy needs relative to BMR, we find that adults living a modern lifestyle in the industrialized world have significantly lower physical activity levels than those living more traditional lives. Among men, PALs in the industrialized societies average 1.67 (range = 1.53 to 1.84), as compared to 1.90 (range = 1.58 to 2.38) among the subsistence-level groups. Physical activity levels among women average 1.63 in the industrialized world (range = 1.48 to 1.69) and 1.78 (range = 1.56 to 2.03) among the subsistence-level societies.

The differences in daily energy demands between subsistence-level and industrialized populations are further highlighted in Fig. 4, which shows daily energy expenditure (kilocalories/day) plotted relative to body weight (in kilograms). The two lines denote the best-fit regressions for both groups. These regressions show us that at the same body weight, adults of the industrialized world have daily energy needs that are 600 to 1000 kilocalories lower than those of people living in subsistence-level societies.

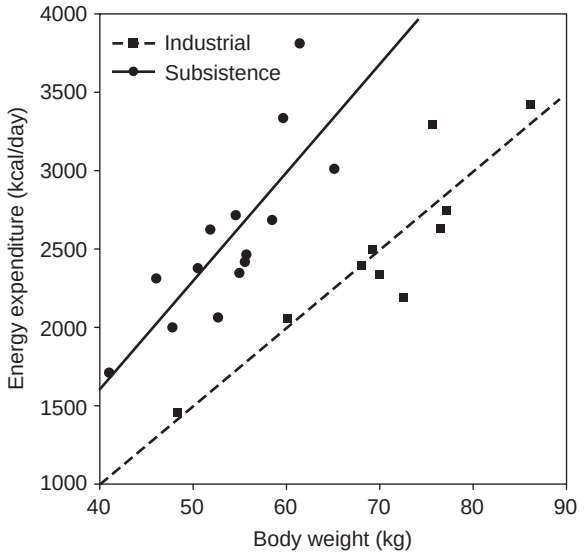


FIGURE 4 Daily energy expenditure (kcal/day) versus body weight (kg) for adult men and women of industrial and subsistence-level populations (from data presented in Table V). At a given body weight, individuals from subsistence-level groups have levels of energy expenditure that are 600 to 1000 kcal greater than those of the industrialized populations.

5.3 Energy Balance and Its Health Consequences

Ultimately, energy balance, the dynamic relationship between energy consumed and expended, is critical for determining nutritional health and well-being. In the industrialized world, low levels of physical activity coupled with relative abundance of food create a net surplus of energy, or a positive energy balance. Conversely, among rural food-producing societies of the developing world, limited energy availability and heavy workloads produce widespread under-nutrition (negative energy balance).

To assess long-term energy balance (physical nutritional status), nutritionists typically measure body weight and composition (i.e., fat and muscle). One of the most widely used measures of physical nutritional status is the body mass index (BMI). The BMI is a weight-for-height measure that is calculated as weight (in kilograms) divided by height (in meters) squared. Because the BMI is strongly correlated with levels of body fatness, it provides good long-term measure of energy balance.

Table VII shows the BMI cutoffs for different levels of nutritional health for adults (individuals 18 years and older). BMIs between 18.5 and 24.9 kg/m^2 are considered healthy because this range is associated with the lowest mortality risks. Individuals with BMIs under 18.5 kg/m^2 are considered to be underweight. Mild and moderate under-nutrition is associated with BMIs of 17.0 to 18.4 and 16.0 to 17.0 kg/m^2, respectively. BMIs of less than 16 kg/m^2 are indicative of starvation conditions (i.e., severe and acute under-nutrition). BMIs of 25 and above are indicative of over-nutrition. Individuals with BMIs between 25.0 and 29.9 kg/m^2 are considered overweight, and those with BMIs of 30.0 kg/m^2 or more are classified as obese.

In the United States, rates of obesity have increased dramatically (see Table VIII). As of 2001, 21% of the

TABLE VII

BMI Categories for Assessing Physical Nutritional Status in Adults

BMI Range	Classification
< 16.0	Severe under-nutrition
16.0–16.9	Moderate under-nutrition
17.0–18.4	Mild under-nutrition
18.5–24.9	"Healthy"
25.0–29.9	Overweight
≥ 30	Obese

TABLE VIII

Percentage of U.S. Adults Who Are Obese (BMI ≥ 30 kg/m^2), 1991–2001

	Percent obese (BMI ≥ 30)					
Sample	**1991**	**1995**	**1998**	**1999**	**2000**	**2001**
Men	11.7	15.6	17.7	19.1	20.2	21.0
Women	12.2	15.0	18.1	18.6	19.4	20.8
Race, ethnicity						
White, non-Hispanic	11.3	14.5	16.6	17.7	18.5	19.6
Black, non-Hispanic	19.3	22.6	26.9	27.3	29.3	31.1
Hispanic	11.6	16.8	20.8	21.5	23.4	23.7
Other	7.3	9.6	11.9	12.4	12.0	15.7
Educational level						
Less than High school	16.5	20.1	24.1	25.3	26.1	27.4
High school degree	13.3	16.7	19.4	20.6	21.7	23.2
Some college	10.7	15.1	17.8	18.1	19.5	21.0
College or above	8.0	11.0	13.1	14.3	15.2	15.7

Source. National Center for Chronic Disease Prevention and Health Promotion, Behavioral Risk Factor Surveillance System (1991–2001); www.cdc.gov/nccdphp/dnpa/obesity/trend/prev_char.htm.

adult American population was considered to be obese, up from ~12% in 1991. Obesity rates are particularly high in minority groups—31% in African Americans and 24% among Hispanics—linked to conditions of poverty, class, and level of education.

These rising rates of obesity in the United States are also associated with increased rates of other cardiovascular problems, including high cholesterol, high blood pressure, and diabetes. This cluster of chronic health problems is referred to as the metabolic syndrome, because all of the individual features reflect a surplus of energy.

In the developing world, under-nutrition has declined; however, it remains a persistent problem. As shown in Table IX, rates of under-nutrition (BMIs <18.5 kg/m^2) are relatively higher in Asia and Africa than in Latin America. In China and India, the levels are 12% and 49%, respectively whereas those in the Sahelian region of Africa fall between 15% and 20%. The prevalence of under-nutrition in Latin American nations is typically less than 10%. These national-level statistics can be misleading in that they mask large urban-rural differences in nutritional health. Sadly, it has generally been the impoverished rural populations of the developing world that have shown the smallest improvements in health and nutritional status.

While under-nutrition remains a severe problem in rural regions of the Third World, increasing rates of obesity are being found among urban populations of the developing world. This trend has been documented in North Africa (e.g., Morocco and Tunisia) and South America (e.g., Brazil and Chile) and is likely attributable to the Westernization of dietary habits and more sedentary lifestyles. As in the United States, diabetes and other cardiovascular diseases have increased as a consequence of the rise in obesity. It thus appears that much of developing world is now undergoing a rapid nutritional transition such that obesity and under-nutrition exist as parallel problems.

TABLE IX

Percentage Under-Nutrition (BMI < 18.5 kg/m^2) among Adults in Selected Countries

Region/Country	Percentage under-nutrition
Africa	
Ghana	20
Mali	16
Asia	
China	12
India	49
Latin America	
Brazil	6
Cuba	7
Peru	3

Source. Shetty and James (1994).

Addressing the emerging obesity epidemic in the United States and throughout the world will require changes in consumption patterns as well as broader lifestyle modifications. The wider availability of

inexpensive market foods that are high in fats and simple carbohydrates is certainly a primary cause of the obesity problem. However, the other half of energy balance equation—reduced energy expenditure associated with increasingly sedentary lifestyles—is also a critical and yet often overlooked part of the problem. Consequently, it is not surprising the most effective nutritional interventions are those that promote increased exercise and activity levels as well dietary modifications.

SEE ALSO THE FOLLOWING ARTICLES

Complex Systems and Energy • Conversion of Energy: People and Animals • Earth's Energy Balance • Easter Island: Resource Depletion and Collapse • Ecosystems and Energy: History and Overview • Emergy Analysis and Environmental Accounting • Food Capture, Energy Costs of • Food System, Energy Use in • Goods and Services: Energy Costs • Heterotrophic Energy Flows • Hunting and Gathering Societies, Energy Flows in • Thermoregulation

Further Reading

Black, A. E., Coward, W. A., Cole, T. J., and Prentice, A. M. (1996). Human energy expenditure in affluent societies: An analysis of 574 double-labelled water measurements. *European J. Clinical Nutrition* **50,** 72–92.

Durnin, J. V. G. A., and Passmore, R. (1967). "Energy, Work and Leisure." London: Heineman.

Food and Agriculture Organization, World Health Organization, and United Nations University (FAO/WHO/UNU) (1985). Energy and protein requirements. Report of Joint FAO/WHO/UNU Expert Consultation. WHO Technical Report Series No. 724. World Health Organization, Geneva.

Gibson, R. S. (1990). "Principles of Nutritional Assessment." Oxford University Press, Oxford.

James, W. P. T., and Schofield, E. C. (1990). "Human Energy Requirements: A Manual for Planners and Nutritionists." Oxford University Press, Oxford.

Kleiber, M. (1975). "The Fire of Life: An Introduction to Animal Energetics." 2nd Ed. Krieger. Huntington, NY.

Leonard, W. R. (2000). Human Nutritional Evolution. *In* "Human Biology: A Biocultural and Evolutionary Approach" (S. Stinson, B. Bogin, R. Huss-Ashmore, and D. O'Rourke, eds.), pp. 295–343. Wiley-Liss, New York.

Leonard, W. R. (2002). Food for thought: Dietary change was a driving force in human evolution. *Scientific American* **287**(6), 106–115.

McArdle, W. D., Katch, F. I., and Katch, V. L. (2001). "Exercise Physiology: Energy, Nutrition and Human Performance." 5th Ed. Lippincott Williams & Wilkins, Philadelphia.

McLean, J. A., and Tobin, G. (1987). "Animal and Human Calorimetry." Cambridge University Press, Cambridge, UK.

Millstone, E., and Lang, T. (2003). "The Penguin Atlas of Food." Penguin Books, New York.

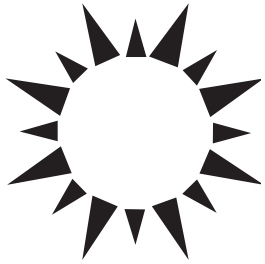
Montoye, H. J., Kemper, H. C. G., Saris, W. M. H., and Washburn, R. A. (1996). "Measuring Physical Activity and Energy Expenditure." Human Kinetics, Champaign, IL.

Prentice, A. (1990). "The Doubly-Labelled Water Method for Measuring Energy Expenditure: A Consensus Report by the IDECG Working Group." International Atomic Energy Agency, Vienna.

Shetty, P. S., and James, W. P. T. (1994). Body mass index: A measure of chronic energy deficiency. FAO Food and Nutrition Paper, no. 50. FAO, Rome.

Ulijaszek, S. J. (1995). "Human Energetics in Biological Anthropology." Cambridge University Press, Cambridge, UK.

Hunting and Gathering Societies, Energy Flows in



TIM BAYLISS- SMITH
University of Cambridge
Cambridge, United Kingdom

1. Introduction
2. Problems in Defining Hunting and Gathering
3. Problems in Generalizing from Modern HG Societies
4. Energy Flow in Contrasting HG Environments
5. Energy Flow and Optimal Foraging Theory
6. Conclusion

Glossary

collecting A hunter–gatherer strategy in which task groups move out from a residential base to harvest, process, and transport seasonal concentrations of resources for future use.

foraging A hunter–gatherer strategy for exploitation of resources in the vicinity of the residential base, to harvest and consume resources that are continuously available in low concentrations.

Holocene The period of Earth history that began 10,000 years ago, at the end of the last glacial period.

hominid A primate of the order Hominidae, which includes modern humans (*Homo sapiens sapiens*), earlier human subspecies, and their direct ancestors.

logistic mobility The pattern of spatial movements associated with collecting resources from a residential base.

megafauna Mammals, reptiles, and birds that belong to the largest size categories.

optimal foraging theory (OFT) Models of hunter–gatherer behavior that predict the selection of resources, foraging behavior within patches of resources, or the mobility of the residential base. OFT uses microeconomic theory to assess what outcome will be the most efficient for the individual or for society—for example, by maximizing energy gain.

Pleistocene A period of Earth history that began 2.5 million years ago, and characterized by repeated glacial/interglacial fluctuations in climate.

residential mobility The pattern of frequent moves of the residential base made by hunter–gatherer groups that exploit resources by foraging.

sedentism A sedentary settlement pattern, through which hunter–gatherers settle down to live in fixed locations and in more permanent dwellings.

Hunter–gatherer (HG) societies have existed in all continents, and today survive mainly in tropical forests, semidesert areas, boreal forests, the tundra, and the Arctic, all environments where agricultural colonization and other forms of displacement have not yet occurred. These societies are defined principally according to their dependence on nondomesticated resources, accessed through hunting, trapping, fishing, and gathering. In many societies, cultivation or herding coexists with the HG mode, and exchange often provides people with some items not available from local sources. Typically, these societies are small scale, egalitarian, and live at low population densities, although not in all cases. Their patterns of energy flow, including diet, labor use, mobility, and optimizing behavior, have been used as models for understanding past HG societies in areas that are now agricultural, and also to illuminate the behavior of other hominids, because HG subsistence has been dominant for about 90% of the history of *Homo sapiens sapiens* and about 99% of the time since hominids first evolved.

1. INTRODUCTION

Extrapolating data back in time from a few modern HG societies must be done cautiously. The geographical distribution of modern hunter–gatherers is highly skewed; modern groups live in environments marginal for farming or pastoralism and in places remote from colonialism, modernity, and the nation state. Because of the disappearance of megafauna,

 © 2004 Elsevier Inc. All rights reserved.

they exploit a range of resources that are much depleted compared to the Pleistocene norm. Hunter–gatherers encompass such a range of different adaptations, particularly in relation to variations in the seasonality in energy supply and variations in the spatial concentration of food energy, that generalizing from particular case studies is very hazardous. Problems in defining HG societies need first to be addressed (Section 2), before confronting this ecological diversity and achieving a meaningful classification of HG societies (Section 3). Only then is it possible to make valid generalizations about the patterns of energy flow in HG societies (Sections 4 and 5), to generate findings that might be applicable to other times and places.

2. PROBLEMS IN DEFINING HUNTING AND GATHERING

2.1 Origins of the Hunter–Gatherer Concept

The concept of "hunter–gatherer" became necessary when, in the 19th century, Western scholars began to classify the different modes of subsistence that existed in the world. Sedentary agriculturalists were distinguished from pastoralists, people primarily dependent on their herds of domesticated animals. But what of those societies whose food resources were not domesticated, who lived instead by hunting wild animals, fishing, and gathering wild plant foods? Such peoples lived in an immense range of different environments and in rather strongly contrasted societies, in places as remote from each other as African deserts, South American jungles, and the Arctic tundra. Even so, for convenience, all were put in the one category, hunter–gatherer, for reasons that perhaps were partly political. By seeing HG societies as a homogeneous category with a preagricultural and therefore "primitive" technology, it became easier to justify their displacement following the spread of European settlement and colonial expansion.

Despite its unsatisfactory origins, the HG concept became even more firmly established when the long prehistory of the genus *Homo* was demonstrated by comparison with the rather short history of plant and animal domestication. Whereas agriculture was recognized as extending no further back than the Neolithic, perhaps 6000 years ago, the origins of *Homo sapiens* lay a long way back in the Pleistocene. In 1966, a symposium was held in Chicago on the theme "Man the Hunter," at which Richard Lee and Irven DeVore claimed that "cultural man has been on earth for some 2,000,000 years, and for more than 90% of this time he has lived as a hunter–gatherer."

Although modern archaeology would not extend the origins of *H. sapiens sapiens* (anatomically modern humans) beyond about 130,000 years ago, research has also confirmed that no clear evidence exists for full-scale agriculture anywhere until well after the Holocene began 10,000 years ago. Nor is pastoralism any older; indeed, most pastoral societies postdate the formation of settled farming communities, with which they typically exchange products (e.g., cattle for grain). From various perspectives, therefore, it became important for scholars to understand more about the economies, societies, and cosmologies of these preagricultural and prepastoral HG societies, because they were now seen as representing such a large part of the human heritage. Yet we inherit many prejudices and assumptions about such societies. Prejudice in Western thought can be traced back at least to the time of Thomas Hobbes (1651), who typecast hunter–gatherers as not merely the primeval state of humanity, but people who lived lives that he famously described as "solitary, poor, nasty, brutish and short." The influence of this stereotype undoubtedly lingered until the mid-20th century, but it did so alongside another more romantic vision of hunter–gatherers. This alternative view can be traced back to John Dryden, who in 1670 depicted hunter–gatherers as living in a state of grace from which the rest of humanity had fallen, coining the equally famous phrase "the noble savage" to describe them.

The stereotypes of both Hobbes and Dryden influenced Western attitudes and policies towards HG societies during the colonial period, and were not finally dispelled until the 1960s when the results of scientific research became more widely known, including studies of the energy efficiency of hunting and gathering. Ironically, by then most HG societies were disappearing, displaced by colonial genocide and modernity or absorbed by the expansion of neighboring groups and by the nation state.

2.2 Sources of HG Food Energy

Hunter–gatherers are usually identified today by means of a "package" of traits, including economic, ecological, sociocultural, and ideological features, but of these the HG economy is the most important

characteristic. The *Cambridge Encyclopaedia of Hunters and Gatherers* offers the following definition of the HG mode: "Subsistence based on hunting of wild animals, gathering of wild plant foods, and fishing, with no domestication of plants and no domesticated animals except the dog." As well as their food energy, the fuels, fibers, and tools of hunter–gatherers also come from nondomesticated, noncultivated sources, although iron tools, pottery, and a few other items may be acquired by exchange.

In most HG societies, there are five categories of food energy procurement: (1) hunting, involving the active pursuit of mammals, reptiles, or birds; (2) trapping, which is the passive capture of prey (mammals, birds, or fish) by means of prior investments in pitfalls, traps, nets, weirs, and other technical means; (3) fishing, using a wide range of active techniques; and (4) gathering, which involves the acquisition of passive animal prey/products such as shellfish, snails, insect larvae, birds' eggs, and honey, and also plant foods of all kinds. The final category, (5) exchange with neighboring groups, should be added, in almost all modern cases, although the extent to which food energy is traded, rather than other goods, varies a lot.

The relative importance of the various sources of subsistence is highly variable between different HG societies, dependent on environmental opportunities and constraints. However, to prioritize hunting in the definition of such societies is misleading. Only in a minority of cases (for example, in the Arctic, in boreal forests, and in places where fish and sea mammals are especially abundant) do hunting, trapping, and fishing contribute more to the diet than does the gathering of plant foods. Meat is generally the preferred food in all societies, but gathered plant foods are more abundant in a majority of cases and, crucially, are more dependable. "Gatherer–hunters" might, therefore, be a more accurate term.

Other definitions of HG subsistence could be cited, but what they all have in common is the absence of direct human control over the sources of food energy on which people depend, including the reproduction, the distribution, and the abundance of the various plant and animal species. Lack of direct control in HG societies is contrasted to what agriculturalists achieve, including (1) a transformation of ecosystems using energy-intensive means and (2) a partial or total domestication of plants and animals, with either minor or major genetic changes to the various species involved.

2.3 Energy Foraging versus Energy Collecting

For understanding spatial behavior in HG societies, the archaeologist Lewis Binford made an important distinction between foraging and collecting food energy. These represent alternative strategies for solving problems of resource distribution. Foraging takes place where resources are widespread and continuously available, and where frequent residential mobility enables hunter–gatherers to move their camps to the resource they wish to use. In this way, they can forage for what they need by means of immediate return technologies and social practices. The storage of food energy is minimal in such societies, which are typical of HG societies in the semideserts and forests of tropical and midlatitudes (Fig. 1).

In contrast, logistical mobility takes place where collecting of resources is needed for future consumption. From the residential base, task groups are sent out to hunt or gather particular foods, undertake the initial processing, and return to base with a surplus that then provides a future energy store. This alternative has been called by James Woodburn a strategy of delayed return, and it is essential where periods of resource shortage can be anticipated. The northwest coast of North America, for example, was

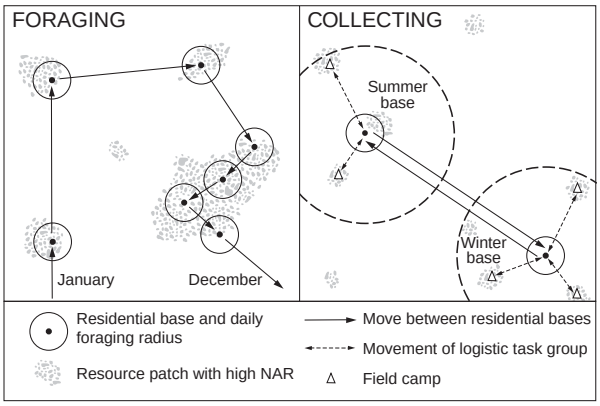


FIGURE 1 Schematic maps contrasting the foraging and the collecting strategies of hunter–gatherers. In both cases, residential bases are located in places with good access to resource patches with high net acquisition rates. A net acquisition rate (NAR) is defined as food energy output divided by work energy input, per hour of foraging/collecting activity. The residential bases of the foragers are each occupied for much shorter periods, compared to those of the collectors. In both cases, virtually no stored food energy is in the possession of the society at the end of the more difficult season of the year, but foragers consume energy on an immediate-return basis in the intervening period, whereas collectors have a delayed-return consumption pattern made possible by their logistic strategy.

an environment where salmon and waterfowl were superabundant in season, and where a single sturgeon, sea lion, or whale could provide enough food for many days or weeks of consumption. However, according to anthropologist Wayne Suttles, "abundance consisted only of certain things at certain times and always with some possibility of failure." To cope with abundance, in a situation where for some foods (e.g., salmon) more effort was needed to store the food than to collect it, complex technologies and social institutions had be developed that could guarantee delayed energy return, and so prevent the society from being destroyed by famine at the end of winter.

Most HG societies need to have a mix of foraging and collecting strategies, being residentially mobile in some seasons (e.g., summer or dry season) but in other seasons (e.g., winter or wet season) operating from fixed base camps by means of logistical mobility. A primary emphasis on delayed-return energy use seems to be associated with the emergence of more stratified ("complex") societies, such as the ranked societies of chiefs, freemen, and slaves that existed on the northwest coast of North America. Ian Hodder has argued that the ideologies associated with delayed-return HG subsistence paved the way for agriculture, but other archaeologists disagree.

Using these concepts, HG societies can be classified according to their adaptations to two critical aspects of their energy resources: (1) seasonal and year-to-year fluctuations, which control the scale of energy storage required, and (2) variations in the spatial concentration of critical resources (food, water, fuel), which permit different degrees of sedentism. The four main categories of HG society are shown in Table I, together with examples. In practice, it is rare today to find societies that depend solely on hunting, fishing, and gathering. For their subsistence, most documented communities engage in a limited amount of cultivation, herding, and/or trading, in addition to their main HG focus.

2.4 Use of Energy Subsidies

Unlike farmers, hunter–gatherers do not depend on energy subsidies (e.g., from fossil fuels), and their impact on the landscape is small-scale apart from localized fuelwood consumption. However, HG societies do exert indirect control over energy flows in the ecosystem, particularly by using fire to assist with hunting and to improve the grazing. In what is today the United States, driving game with fire was practiced almost everywhere in pre-Columbian times. Open areas of grassland found today in the apparent Yosemite wilderness are in fact the legacy of former HG hunting practices. In Australia, the use of fire by HG societies transformed vegetation over a much longer period (perhaps 50,000 years), reducing

TABLE I

A Classification of Hunter–Gatherer Societies[a]

Spatial concentration of main sources of food energy[b]	Different mobility patterns	Example: Seasonal and year-to-year fluctuations in resource abundance[c] — Low	Example: Seasonal and year-to-year fluctuations in resource abundance[c] — High
Low	Residential mobility	Inland Australian Aborigines Kalahari San (e.g., !Kung)	No examples known
Intermediate	Seasonal base camps with logistic mobility	Coastal Australian Aborigines (e.g., Gidjingali)	Sub-Arctic maritime societies (e.g., Coastal Sámi, Varanger Fjord, Norway)
		Arctic tundra (e.g., Tuluaqmiut Inuit, north Alaska)	Boreal forest hunters (e.g., Cree, northern Ontario)
		Sub-Arctic inland (e.g., Mountain Sámi, northern Sweden)	Torres Strait islanders, Australia
High	Sedentism	No examples known	Northwest Coast Indians (e.g., Kwakiutl, Salish, Chinook)

[a] Based on the extent of seasonal and year-to-year fluctuations in available food energy and the extent of spatial concentration in food energy, with ethnographic examples.

[b] Correlates with different mobility patterns.

[c] Correlates with (1) extent of capital investments in storage to safeguard future energy supplies and (2) with institutions for ownership and defense of territories.

evergreen rainforest to small patches and greatly enlarging the open woodlands dominated by *Eucalyptus*. Far from being "natural," *Eucalyptus* woodland is now interpreted as an anthropogenic fire-climax vegetation.

Some anthropologists studying hunting practices in Aboriginal Australia and among North American Indians are so impressed by the importance of fire as an indirect tool for ecosystem management that they use the phrase "firestick farming." In the savannas of northern Australia, for example, in the early part of the dry season when cool nights dampen the ground with dew, the Aborigines use small and controlled bush fires that clear away undergrowth but leave the trees undamaged. The growth of new grass in such areas attracts game and makes hunting much easier. Because the modern state has imposed a ban on firestick farming, infrequent but catastrophic and far more destructive forest fires take place.

2.5 Social Differences in Energy Use

A consequence of the low intensity of management of food energy sources by HG societies is that they live at very low population densities, often in small bands of 25–50 persons, with temporary or seasonal settlements and frequent long-distance mobility. In most respects, these societies are strongly egalitarian, with food sharing and an absence of social stratification and only weak institutions for political control. However, there is a strong gender division of labor, with males doing most of the hunting and females gathering the plant foods. Fishing and the collecting of shellfish and small game may be an activity of both sexes, and children participate in the food quest from an early age.

Hierarchy is most strongly developed in semisedentary HG societies where seasonally abundant resources make possible large-scale food energy storage, as among Northwest Coast Indians such as the Kwakiutl in British Columbia or the Chinook in Washington state. Lee and DeVore, in the classic text *Man The Hunter*, wrote "we make two basic assumptions about hunters and gatherers, (1) they live in small groups, and (2) they move around a lot." Perhaps these "simple," residentially mobile HG societies have always coexisted with more sedentary, more complex, and more hierarchical ones.

As regards their attitudes to energy consumption, these are not societies in which maximizing energy flow finds ideological support. Far more important is maximizing the security of energy supplies, simultaneously minimizing the energy expended in work so that more time is spent enjoying leisure and strengthening social bonds with others—itself, arguably, a useful risk-minimizing strategy.

3. PROBLEMS IN GENERALIZING FROM MODERN HG SOCIETIES

3.1 Can We Extrapolate from HG Societies to Hominids?

There has been much interest in the possibility that studies of energy flow in HG societies can be used to achieve an understanding of our hominid ancestors. As Robert Foley put it, "until very recently being a hominid, being a human and being a hunter–gatherer were very nearly the same thing." Do studies of the energetic efficiency of food-gaining among the !Kung San of the Kalahari, or the energy storage strategies of sub-Arctic maritime hunting/fishing communities, have general implications for the behavior of humans or hominids who are remote from these societies in time and space? Perhaps it is the intentionality of the human subject that distinguishes his or her behavior from the superficially similar food quest of the nonhuman HG groups, as well as the human capacity for a symbolic expression of hunting experiences, as shown in the cave art of the Upper Palaeolithic and the shamanism of modern HG groups.

Remember also that modern HG groups exploit a depleted set of resources compared to their Pleistocene ancestors, because of megafaunal extinctions. The capacity clearly existed in the past for rapid population growth and a relationship with nature that is far removed from the stable and rather harmonious one that HG societies have exhibited in modern times. Moreover, these societies are not a random sample of those that existed on Earth before the spread of agriculture. Not only are they restricted to environments less suited ecologically to agriculture or pastoralism, but also they are no longer "pure" in their HG lifestyle, interacting with farmers or herdsmen when opportunity arises.

3.2 The Skewed Distribution of HG Societies Today

Using ethnographic sources, we can map the distribution of 115 HG societies that existed in the 20th century (Fig. 2). Based on the environments they occupy, five groups can be distinguished:

FIGURE 2 Map of 115 hunter–gatherer societies that were ethnographically documented in the 19th and 20th centuries.

- Hot deserts and tropical savannas (19 societies).
- Tropical rain forests and coral reefs (30 societies).
- Boreal forest and tundra zone (38 societies).
- Cool temperate maritime regions (13 societies).
- Other environments (15 societies).

It is apparent that in modern times HG societies have been absent from almost all environments that are prime farming land. Therefore, those that still survive are a very biased sample of the HG mode of the past. The temperate deciduous forests, for example, have not been the domain of hunter–gatherers in Europe since the Neolithic 5000–6000 years ago. In eastern North America and Chile, such forests were used for a mixed HG/agricultural strategy before being taken over by European colonists some 400 years ago; the same process removed Aboriginal HG societies from the temperate forests of Tasmania in the 19th century. While being careful not to label the environments of modern HG societies as necessarily "marginal," which they are not except from the point of view of agriculture, we should also bear in mind the predominance of hot deserts, jungles, and seasonally frozen environments in the world list.

4. ENERGY FLOW IN CONTRASTING HG ENVIRONMENTS

4.1 Managing the Daily Risks of Energy Shortage

4.1.1 General Principles

In some societies, it might make sense to understand behavior in terms of profit or surplus (i.e., maximizing energy output), but most anthropologists now believe that for HG societies maximizing security (i.e., minimizing risk) makes more sense. Security is increased when the risks of food shortage are successfully managed. From her studies of the San in southern Africa, Polly Wiessner identified four general ways in which risk is reduced: (1) prevention of energy deficit, (2) transfer of deficit, (3) pooling of resources, and (4) storage strategies. To prevent deficits on a daily basis, hunter–gatherers need appropriate tools for killing prey or gathering plants; expert knowledge of the probabilities of success and how it can be achieved; and intragroup cooperation in tracking, processing, and transporting foods. The transfer of deficit is achieved by sharing within the

group and by exchange relationships, whereas the pooling of resources is another social strategy that provides for an individual some insurance against failure in food energy acquisition.

Storage strategies for food and water are also potentially useful, but except in the short term these are difficult to achieve in hot and humid climates. Fuel storage is essential only for HG societies coping with extreme winter conditions. Unless food storage can be achieved effectively, for the hunter or gatherer to maximize short-term gain by producing as much energy as possible is not a sensible strategy. Where energy storage is problematic, the social strategies of sharing, cooperation, and exchange are more effective, but all require some form of investment. Investing in friendships and exchange, activities that require leisure time, might prove to be a better strategy for risk reduction, compared to killing more and more animals.

4.1.2 Example: Anbarra Gidjingali Aborigines

The ways in which risk and productivity can interact are well demonstrated by the example of the Anbarra, a subgroup of about 40 people that are part of the Gidjingali (population 350) of Arnhem Land, northern Australia. The economy of the Anbarra was studied by Rhys Jones and Betty Meehan at a time (early 1970s) when a large part of the Anbarra diet still derived from HG resources. Anbarra territory is at the mouth of the Blyth river, in the subhumid tropical savanna zone. There is seasonal access to the rich resources of open woodlands, grassy plains, freshwater swamps and lagoons, sand dunes, and mangroves, as well as the open sea, using a semisedentary pattern of settlement. In the wet season, people live by the coast and depend largely on fish and shellfish. Later in the year, they range inland to take advantage of cycad fruits, wild yams, and small game, and they move further inland across the grassy plains for wallaby hunting in the latter part of the dry season. The seasonal availability of foods is one obvious constraint over their pattern of energy consumption.

Another important factor, however, is the probability of success in the food-gaining quest. There is an inverse relationship between the productivity of labor and the probability of a successful trip. Activities such as wallaby hunting provide on average a high yield of food energy per unit of time invested, but are also the most risky activities, with a 75% chance of complete failure of a hunting trip. At the other end of the spectrum are plant-gathering activities, for which success is virtually guaranteed but the returns to labor are considerably less attractive. The Anbarra's overall strategy is to combine high-yield, high-risk activities (e.g., wallaby hunting or spear fishing) with low-yield, low-risk tasks (e.g., collecting shellfish and gathering of plant foods such as yams and cycads). Table II shows data for risk and yield for four of the major Anbarra food resources.

A viable long-term strategy for energy acquisition therefore requires a blend of different activities. Hunting is men's work, and to ensure against their failure the women must engage in less rewarding but more secure tasks such as gathering cycads and other plants and collecting shellfish. In this way, Anbarra

TABLE II

The Trade-off between Productivity and Risk in Northern Australia[a]

Hunter–gatherer activity	Sex of persons involved[b]	Gross acquisition rate (MJ)[c]	Risk[d]	Importance[e]
Buffalo hunting	M	8400	0.001	8
Wallaby hunting	M	210	0.25	7
Fishing with spear	M	146	0.4	3
Trapping fish	M	146	0.55	5
Fishing with hook and line	M and F	25	0.65	1
Cycad gathering and processing	F	5.4	1.0	6
Yam gathering	F	5.0	0.9	4
Shellfish gathering	F	4.2	0.9	2

[a] The gross acquisition rate for various Anbarra activities during successful trips, compared to the probability of success per attempted trip. Calculations based on 1972–1973 data from Jones (1981).

[b] M, Male; F, female.

[c] Megajoules of food energy per person-hour on a successful trip.

[d] Probability of success per attempted trip.

[e] Ranked according to percentage of days in the year when the activity was attempted.

society maximizes utility in terms of risk avoidance, secures a balanced diet, and maintains social solidarity and mutual aid between the sexes.

4.2 Managing the Energy Cost of Mobility

4.2.1 General Principles

Rather than managing plants and animals to achieve a spatially focused concentration of food energy, which is the strategy of agriculturalists and pastoralists, hunter–gatherers must depend on the natural distribution patterns in time and space of the resources on which they depend. In circumstances in which there is a natural concentration of resources in time and space, sedentism is encouraged, so that people and their possessions do not have to keep shifting from place to place. Even if temporary, sedentism is likely to be attractive to HG societies for many reasons: (1) it confers many social benefits; (2) it allows long-term storage of food energy to be an option, thereby increasing future security; (3) it reduces the direct energy expenditures involved in walking, carrying infants, and transporting possessions from one residential base to another; and (4) it reduces the indirect costs of high mobility, such as higher infant mortality, less equipment, and smaller energy stores, if all food must be carried. However, unless food energy sources are spatially concentrated, sedentism also reduces the accessible territory to a local area, which is soon likely to become overexploited. In arid areas, water may also become scarce. The HG lifestyle can be seen as the outcome of a perpetual compromise between the various costs of mobility (many of them measurable as energy costs) and the costs of sedentism. The disadvantages of sedentism will become energy costs once carrying capacity is exceeded, because when hunting yields decline and other resources are depleted, labor productivity will fall. As a result, people must work for longer periods, and/or journey times from home base to food source must increase, so as to compensate.

4.2.2 Example: The Dobe!Kung

The Dobe subgroup of the !Kung San were studied by Richard Lee in the 1960s, and provide an example of these interacting constraints. For the Dobe, water is the overriding factor that dictates the location of camps. The camp serves as the home base for 30–40 people, who move out each morning in groups of 2 or 3 to collect plant foods (especially mongongo nuts) or hunt game. All groups must return before evening if they wish to share the common meal. After moving to a new camp, the food resources of the local area (defined by the Dobe as the area within a 2-hour hike) are exploited first. By fanning out in all directions and moving at about 5 km/hour with pauses for rest and work, an area within a 10-km radius can be worked in the space of a 6-hour day.

Once the low-productivity/low-risk resources within this local area have been depleted (i.e., the plant foods and easily collected small game), it is more efficient for the group to move to a new camp than it is for people to spend more and more time walking. A division of labor becomes evident, with younger, more active people making longer trips to mongongo woodlands and the older people and children staying close to the camp and searching for less desirable foods. Any who are not prepared to walk further for the staple mongongo must eat foods such as bitter melons, roots, acacia gum, and the heart of the ivory palm. Richard Lee showed that for mongongo nut foraging, the round-trip distance was the best measurement of the time and energy "cost" of obtaining this desirable food. Camps are moved quite frequently (five or six times in the year), but never very far. A rainy season camp in the mongongo woodland is seldom more than 16–20 km from the previous one, but this is sufficient to shift the spatial focus of exploitation to a new and almost unexploited area.

A 10-km radius is a relatively small territory for hunting the larger game animals, which can move over long distances. Potentially, Dobe hunters must be more mobile than the gatherers and small-game hunters, and the Dobe hunters do sometimes range up to 30 km from a home base using overnight stops, in pursuit of high-yield but high-risk game. However, to do so involves the carrying of much equipment, including water, and the return journey with an animal carcass can be equally arduous. Even for these hunters it is generally more energy efficient to focus on those game available within a 10-km radius.

4.2.3 The Original Affluent Society?

For HG groups, any risk of subsistence crisis will be reduced where (1) food energy derives from a diverse set plant and animals and (2) aggregate food resources are continuously available during the year, even if certain foods are more abundant in certain seasons. If a preferred food such as meat is temporarily scarce, then, through more intensive efforts and high spatial mobility, other less desirable foods (e.g., edible plants) can usually be accessed elsewhere. As a result of these strategies, basic

supplies of food energy can be secured throughout the year, and what is produced is almost immediately shared and consumed.

Both the Anbarra Gidjingali (Section 4.1.2) and the Dobe!Kung San (Section 4.2.2) are examples of HG societies that can exploit such circumstances, but with somewhat different mobility patterns. Dobe campsites must be moved frequently, whereas the Anbarra can be semisedentary, with long-term seasonal residence at the boundary between land and estuary. But even in their arid "desert wasteland" (judged from an agropastoralist standpoint), the Dobe!Kung strategy can be judged as highly efficient in maximizing leisure time. According to Lee's data, they devoted only 12–19 hours per week to food-gaining "work" to achieve a state of subsistence energy balance (nobody hungry, but no storage). The surprisingly low level of energy input resulted in the San (and, by extension, all HG societies) being characterized by Marshall Sahlins as "the original affluent society." Agriculturalists, by contrast, usually have to work harder, enjoy less adequate diets, and sometime suffer harvest failure.

In this situation, there is no urgent reason for an HG society to produce a food surplus even if circumstances permit, and there is also little incentive to invest in techniques for food storage. Such "hand-to-mouth" strategies characterize the relatively non-seasonal environments of tropical forests, midlatitude deserts, and productive high-latitude coasts such as the Aleutian islands, Tierra del Fuego, and Tasmania. HG societies that occupy such environments share a number of features in common. Robin Torrence has demonstrated that the technologies employed by such societies involve relatively few types of tools. Hunting focuses on the most rewarding methods of securing prey, and gathering is equally focused on plant foods with the highest returns per unit of gathering effort. New tools and new techniques are added to the HG repertoire in order to reduce the probability of failure, but this only tends to happen in more uncertain environments where the costs of concentrating efforts on only the "best" (i.e., most rewarding) food sources can be very high.

4.3 Coping with Extreme Seasonality in Energy Resources

4.3.1 General Principles

Higher levels of risk are typical of more seasonally fluctuating environments, especially where winter conditions of darkness and extreme cold make hunting difficult and the option of constant mobility becomes extremely arduous. Midwinter sedentism is virtually obligatory in sub-Arctic and Arctic environments, where food storage is not so technically difficult (stored food will not spoil in the freezing conditions), but where substantial quantities of food energy must be amassed if the group's winter survival is to be safeguarded. HG societies in such environments typically invest much energy in strategies that enable them to cope with extreme seasonality—in particular, mass harvesting and mass storage.

4.3.2 Mass Harvesting

To exploit seasonal gluts of animal food requires first an efficient technology and social organization, so that by a concentration of effort a huge surplus of food energy can be amassed. Hunting wild reindeer/caribou, for example, requires leadership, communal effort, and the creation of decoys, barriers, and pitfall traps covering extensive areas, so that many animals can be killed during the window of opportunity that opens up during the brief period of migration in the fall. At Rondane in central Norway, wild reindeer were hunted on a large scale during the period 1000–1200 AD. On the high plateaus, large trapping systems were designed using 3-km lines of wooden stakes to guide the animals into a steep gully, where they could be slaughtered in a funnel-shaped corral. Ørnulf Vorren has found and mapped the remains of 1700 stakes, each supported on a cairn of stones and requiring timber to be transported from pine forests located at least 10 km away. It was obviously hard work to set up and to maintain this system, and the hunt needed the coordinated efforts of many people, but the reward was a large stockpile of meat and furs. When the snows came, this stockpile could be transported to winter base camps on sledges, using tame reindeer as draft animals.

4.3.3 Mass Storage

The limits to energy storage capacity have often been set less by people's capacity to collect food than by their capacity to store it. For HG groups to take advantage of seasonal runs of fish, for example, they require large investments in weirs, nets, and traps, and these must be in place before harvesting begins. Equally important, however, are the investments in processing, transport, and secure storage of such foods. For groups such as the Chinook, Salish, or Kwakiutl, which occupy the big rivers of the northwest coast of North America, a heavy fall run of

salmon represented a crisis of abundance. How much food could be stored depended mainly on the number of hands available for the work of cutting and skewering the fish, how many drying racks had been prepared in houses, the availability of fuel for smoke, and the number of containers available. Preserving food was largely women's work, and the key role of women was reflected in contested rights over them, high bride prices, and the institution of slavery.

Not all fish keep equally well, and many salmon species cannot be preserved for the winter simply by drying. On the Columbia River, for example, the Chinook dried the filleted salmon and then pulverized it with berries. Tons of this preserved food was then stored in wooden boxes sealed with oil rendered from sea mammal fat. On the Great Plains, buffalo meat was prepared in a similar way with berries and grease to make pemmican, which, if made properly, could be stored for years, although most was consumed within a single year.

Techniques to preserve food energy, such as salting, drying, and smoking, all represent major investments. Like the technology of mass harvesting, these storage technologies represent forms both of stored energy and of social capital to provide the reserves without which winter survival would be impossible. A consequence of the investments in traps and storage is that a form of property is created by the group or individual responsible, and this property and the territories that serve as catchments (pastures, bays, rivers) may also be defended by groups against outsiders. In such circumstances, too, a more stratified and hierarchical HG society tends to emerge.

4.3.4 Example of Mass Harvesting/Storage: The Varanger Sámi

The Coastal Sámi of Varanger fjord in north Norway are today largely assimilated into the modern society and economy of Norway. They are farmers, commercial fishermen, and reindeer herders rather than hunters and gatherers. In the past, however, this large, treeless Arctic fjord, with its hinterland of stunted birch forests and mountain tundra, was the site of one of Europe's last hunter–gatherer societies. Their seasonal pattern of resource use has been reconstructed by the archaeologist Knut Odner (Fig. 3).

Until the 1700s, the main basis for both subsistence and lifestyle was fishing (cod, coalfish, and haddock), hunting wild reindeer, and hunting whales and seals. All of these foods could be plentiful in season, and a range of minor foods such as seabirds and their eggs, grouse, hare, and freshwater salmon

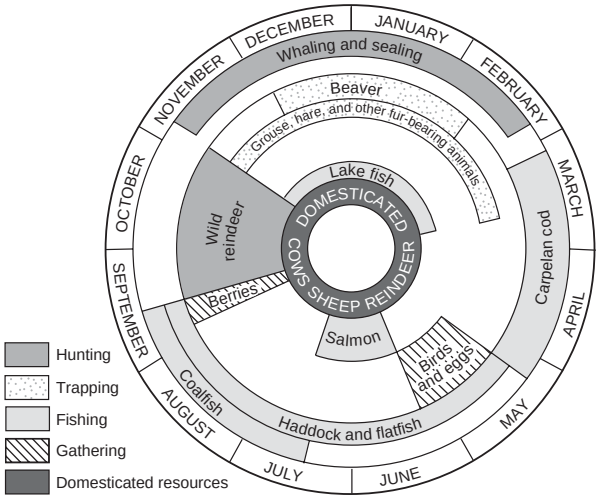


FIGURE 3 Seasonality of the food energy sources available the Varanger Sámi before circa 1700, as reconstructed by Knut Odner.

was also available for brief periods. The problem was to devise effective delayed-return strategies to enable the society to cope with fluctuations, particularly the very sparse resources available in the late-winter season, after the end of whaling and sealing in mid-February and before the arrival of shoals of cod at the end of March.

Of all the various resources, sea fishing for resident species was the most dependable year-round source of energy (low productivity but low risk). Migratory shoals also made possible a storable surplus, particularly the spring cod. On average, a fisherman could expect in one day to land about 100 kg of cod, and he could maintain this level of productivity for about 10 weeks. Each family might therefore expect to be able to dry about 7000 kg of fish in the season. Drying racks, women's labor, and access to boats, fish hooks, and lines were all essential for success. In an environment where fuel energy was scarce, air drying was the main preserving technique in spring, until warmer summer weather caused the fish to become fly-blown.

Until the wild reindeer became extinct, the Varanger Sámi employed various techniques for mass harvesting during their fall migration. The most effective were systems of fences and corrals using stones or poles stuck in the ground, and also lines of pitfalls. The remains of about 5000 hunting pits can be seen around Varanger fjord today. Probably all families contributed hunters to the task group, and the carcasses would have been equally divided. The frozen meat and the skins provided a source of food and clothing for the community in the winter months.

By 1700, growing trade and reindeer domestication made possible an increased sedentism, the animals playing a crucial role in transporting meat and fuel energy from the hinterland. A more sedentary lifestyle in turn made it possible for cattle and sheep to be kept on a small scale, despite the high energy costs that foddering and housing them entailed. The animals and their wool, milk, and cheese made a small but critical contribution to the seasonal energy gap that existed in the late winter, especially after the disappearance of wild reindeer as a significant resource. In these ways, the dependence of the Varanger Sámi on HG energy sources diminished, as they diversified to become semi-agropastoralists.

5. ENERGY FLOW AND OPTIMAL FORAGING THEORY

5.1 Resource Selection

Biologists would argue that we owe our current capacity to consume and enjoy a great variety of foodstuffs to the wide-ranging omnivory of our primate ancestors. That omnivory implies that hominid foragers consistently faced multiple choices in the subsistence quest. Almost all environments contain more species that are edible by humans than can be effectively be harvested by them. Resource selection models are designed to analyze this situation, by asking the question "which of the available resources will an efficient forager seek to harvest?" By comparing the predicted pattern of resource selection, which represents maximum efficiency, with the actual behavior and diet of HG groups, we can perhaps obtain some measure of how far the groups conform to efficiency predictions, rather than being led by other preferences.

The situation that is modeled is where a searching forager randomly encounters edible resources and is then faced with a decision. Should he/she take time to harvest (hunt, gather) this item, or, alternatively, is it better to continue searching in the hope of finding something more rewarding to harvest? Anthropologists have devised a measure of profitability termed the net acquisition rate (NAR), which is a measure of how much net energy is acquired (food energy output divided by work energy input) per hour of time devoted to the harvest. Some examples of the data necessary for these calculations are shown in Table II for the Anbarra. In the Anbarra case, the gross yield per person-hour for shellfish collecting averaged a modest 900 kcal, whereas spearfishing generated 14,000 kcal/hour and wallaby hunting much more. Knowing that fish suitable for spearing might be encountered later in the day, or a wallaby might be found, does the forager pause at the shellfish beds and harvest them, or does he or she move on in the hope of better things? The dilemma is that stopping to harvest what is encountered precludes harvesting something else—in other words, there are opportunity costs involved in making any choice. (In practice, mindful of the risks of failure in seeking high-risk resources with a high NAR, probably the woman forager will concentrate on harvesting the reliable but low-NAR shellfish and plant foods, whereas the male hunter will gamble on finding game.)

Bruce Winterhalder has used this approach to show how an optimal foraging strategy can be modeled. The model predicts the total set of resources that will be exploited if foragers have perfect knowledge (or gradually acquire it) regarding the benefits (the perceived NAR) of making certain choices. These modeling exercises suggest that resources with very low profitability, however abundant, will always be avoided if higher ranked species are available, because the opportunity costs of exploiting low-profit resources are too high. As Winterhalder says, "pennies in a field of sufficient dollars will always be ignored whatever their numbers." For Cree Indian foragers living in the boreal forests of northern Ontario, Canada, the model predicted that the introduction of steel traps, snares, and rifles should have led to an increase in profitability (energy gain per hour). As a result, this new technology should have expanded the range of species that the forager decided was worth harvesting. By contrast, the more recent introduction of the snowmobile so greatly increased the rate of encountering items that the efficient forager would now focus on a narrower set of resources. These predicted shifts were verified from the historical record, suggesting that this approach might have a wide potential for understanding the diversity of HG foraging behavior.

5.2 The Use of Patches

It is also possible to model foraging behavior within patches of resources. Again, the assumption (derived from microeconomics) is that foragers make decisions based on their knowledge of alternatives. A decision to stop consumption of one resource, or to cease one harvesting activity in favor of another one, is made by comparing the value of the last unit gained to the anticipated value of the alternative.

Thus the marginal value of harvesting one resource is compared to the average value of moving on.

Patches are discrete, localized concentration of resources, located such that a mobile forager might encounter several in the course of a day. Is it better to strip a patch of all the items that are encountered, or should the optimum forager move on the next patch quickly, having only taken the most accessible items? Maximum efficiency requires the forager to abandon a patch when its declining marginal rate of return equals the average NAR of many visits to many patches. This application of optimum foraging theory predicts that foragers will nearly always depart from a patch before it has been depleted of all its resources, which matches anecdotal evidence from HG practice. However, as with all these models, testing the predictions against real-world data is far from easy.

5.3 Central Place Foraging

Ethnographic studies show that few HG societies are wholly sedentary, and that residential relocation occurs with a frequency of up to 60 moves per year in extreme cases. The distance of relocation is normally in the 5- to 10-km range, but relocation is occasionally up to 70 km in one move. If we consider these base camps as central places, then can "central place theory" of human geography help to explain the diversity of HG residential mobility patterns? The theory predicts that to economize on the energy costs of resource acquisition, central places will be chosen either (1) adjacent to a critical resource that is rare and bulky, such as water, or more generally (2) at a spot chosen because it is at the center of gravity of the key resources that need to be transported there, such as water, fuel, and food. From this location, adjacent resources are exploited first. The net acquisition rate declines as the forager moves outward, reflecting the energy costs of walking to the resource and then transporting the items that are harvested back to base. Later, as the inner zone becomes depleted, the group experiences lower and lower NAR levels, until a point is reached when the marginal return from continuing to live in a central place is lower than the energy costs of moving elsewhere.

Winterhalder has shown that manipulation of the model predicts that camps will relocate more often (1) if relocation costs are low (for example, where the next suitable central place is not too far away); (2) if depletion of zones close to the central place is rapid, which will happen if patch density is low; and (3) if alternative sites offer a high initial rate of return. The very high mobility of foraging bands in the initial phase of colonization of North America at the end of the Ice Age is consistent with the predictions of the model, in a situation whereby the big game herds were abundant but easily depleted.

6. CONCLUSION

6.1 Energy and HG Theory

Energy provides a convenient way in which costs and benefits can be quantified, at a range of spatial and temporal scales. It is clear that the enormous diversity of HG societies makes it impossible to construct simple generalizations about energy flow, but on the other hand we can use concepts of efficiency to make some progress in understanding the various adaptations that can be observed. So far, theoretical work has focused on optimal foraging behavior, wherein "optimum" is defined in terms of maximizing energy yield or minimizing the direct energy expenditures involved in human labor.

Meanwhile, ethnographic studies of HG communities emphasize risk avoidance as well as energy efficiency as the dominant motives underlying economic behavior. The avoidance of energy deficit during critical seasons when resources are unavailable, and insurance against subsistence failure, are seen by anthropologists as key factors. These concepts can help to explain much of the variation that exists between different HG societies in their residential mobility, their focus on energy storage, and their defense of territories, and the social institutions that underpin these adaptations. Concepts of energy flow may help to provide a unifying theory of HG subsistence behavior.

6.2 The Future of HG Research

Ironically, as HG societies disappear or are transformed by modernity and the market economy, the opportunities for new research will become increasingly limited to the most marginal, and therefore the most atypical, HG groups. As our interest in mankind's own past grows, the desire to know more about energy flows in preagricultural societies is not likely to diminish. The surge of new information from palaeontology and human genetics about the remote origins of *Homo sapiens* is likely to arouse even more curiosity about this vanishing mode of food energy acquisition, given the dependence of all hominids on HG subsistence techniques. Only by placing modern HG societies firmly into both their ecological and historical contexts can we use 21st

century knowledge to illuminate patterns of ecosystem management in the remote Pleistocene era.

SEE ALSO THE FOLLOWING ARTICLES

Cultural Evolution and Energy • Early Industrial World, Energy Flow in • Easter Island: Resource Depletion and Collapse • Fire: A Socioecological and Historical Survey • Food Capture, Energy Costs of • Human Energetics • Industrial Agriculture, Energy Flows in • Transitions in Energy Use • Wood Energy, History of • World History and Energy

Further Reading

Harris, D. R. (1996). Introduction: themes and concepts in the study of early agriculture. *In* "The Origins and Spread of Agriculture and Pastoralism in Eurasia" (D. R. Harris, Ed.), pp. 1–9. UCL Press, London.

Ingold, T., Riches, D., and Woodburn, J. (eds.). (1987–1988). "Hunters and Gatherers. Vol. 1, History, Evolution and Social Change"; "Hunters and Gatherers. Vol. 2. Property, Power and Ideology." Berg, Oxford, and St Martin's Press, New York.

Jones, R. (1981). Hunters in the Australian coastal savanna. *In* "Human Ecology in Savanna Environments" (D. R. Harris, Ed.), pp. 107–146. Academic Press, London.

Kelly, R. L. (1995). "The Foraging Spectrum: diversity in Hunter–Gatherer Lifeways." Smithsonian Institution Press, Washington, D.C.

Lee, R. B., and DeVore, I. (eds.). (1968). "Man The Hunter." Aldine, New York.

Lee, R. B., and Daly, R. (eds.). (1999). "The Cambridge Encyclopaedia of Hunters and Gatherers." Cambridge Univ. Press, Cambridge.

Odner, K. (1992). "The Varanger Saami, Habitation and Economy AD 1200–1900." Scandinavian Univ. Press, Oslo.

Panter-Brick, C., Layton, R. H., and Rowley-Conwy, P. (eds.). (2001). "Hunter–Gatherers: An Interdisciplinary Perspective." Cambridge Univ. Press, Cambridge.

Winterhalder, B., and Smith, E. A. (eds.). (1981). "Hunter–Gatherer Foraging Strategies: Ethnographic and Archaeological Analyses." Univ. Chicago Press, Chicago.

Hybrid Electric Vehicles

JOHN M. GERMAN
American Honda Motor Company, Inc.
Ann Arbor, Michigan, United States

1. **Introduction: Effects of Hybrid Electric Vehicles**
2. **History**
3. **Hybrid Electric System Design**
4. **Real-World Hybrid Examples**
5. **Hybrid Advantages and Disadvantages**
6. **Future Technology Development**
7. **Future Hybrid Markets**

Glossary

energy density Energy storage per unit weight. Energy is the power rate multiplied by the length of time the power is maintained; measured as the amount of energy (such as kilowatt-hour) per unit of weight (such as kilogram or pound).

hybrid electric A system that combines a combustion engine (e.g., gasoline, diesel, or turbine) and an electric motor and energy storage (e.g., a battery or ultracapacitor).

hybrid propulsion A combination of two different power sources for vehicle propulsion. One source derives its power from fuel, such as an internal combustion engine or a fuel cell. The second source is a device that stores and reuses energy. This can be electrical energy or electromechanical energy, such as a hydraulic accumulator or a flywheel.

load Forces that must be overcome to accelerate a vehicle and maintain cruising speed. These include air drag, tire rolling resistance, mechanical resistances in the drivetrain, vehicle mass resistance to acceleration, and accessory loads from pumps, air conditioning, and power steering, for example.

power density The power delivered per unit weight; measured as the amount of power, such as kilowatts or horsepower, per unit of weight, such as kilograms or pounds.

regenerative braking The capture of energy from a vehicle's inertia when the vehicle slows down. A large amount of the energy that is used to accelerate a vehicle to operating speeds is normally lost when the vehicle slows down or brakes. Regenerative braking captures part of this energy and uses it to recharge an energy storage device.

ultracapacitors Higher specific energy and power versions of electrolytic capacitors; devices that store energy as an electrostatic charge.

The internal combustion engine has dominated the car and light-truck market for over 100 years. Although remarkable improvements have been made over the past 30 years to reduce air pollutants to nearly zero and to almost double vehicle efficiency, increasing concerns about global warming and energy security are pushing vehicles toward even greater efficiency. Adding a high-power electric motor and electric storage capacity to an internal combustion engine offers significant fuel savings. The engine can be shut off at idle to avoid wasting fuel, the motor can be driven in reverse when braking to capture energy that would otherwise be lost, and the boost from the electric motor allows use of a smaller, more efficient, engine. In the long term, the electrical power from the motor can be used to replace existing mechanical accessory drives with more efficient electrical devices, as well as to provide power for additional features desired by customers. Against these benefits must be weighed the additional complexity and costs of the hybrid electric power system.

1. INTRODUCTION: EFFECTS OF HYBRID ELECTRIC VEHICLES

In the early days of the automobile, around the turn of the 20th century, there was spirited competition between vehicles powered by electricity and by internal combustion engines. The internal combustion engine won primarily because of the high amount of energy in liquid fuels, which allows vehicles to travel farther without refueling than is possible on an electric charge. Ten gallons of gasoline weighs only about

 © 2004 Elsevier Inc. All rights reserved.

28 kg (62 lb), but contains about 330 kWh (kilowatt-hours) of energy (1.1 million Btu). By comparison, a modern lead-acid battery weighing the same 28 kg provides only about 1.1 kWh. This overwhelming energy advantage of liquid fuel has ensured the dominance of the internal combustion engine for the past 100 years, despite its relatively low efficiency.

Problems with the use of internal combustion engines and fossil fuels are well documented, including low efficiency, air pollution, fossil fuel use, energy security, dependence on foreign oil suppliers, lead poisoning, and leaking storage tanks. To combat these problems, many alternatives to the gasoline internal combustion engine have been proposed over the years, such as steam power, turbine engines, electric vehicles, and the use of alternative fuels such as methanol, ethanol, compressed natural gas, and propane. The latest contender is the fuel cell, powered by hydrogen. However, so far, the internal combustion engine has won over all proposed alternatives. It has accomplished this because the incremental advantages of switching to another propulsion system have been less than the cost of switching to an entirely new infrastructure. In addition, every time the internal combustion engine has been challenged, it has responded with enough improvement to keep the alternative off the market. For example, with the development of modern computer controls and catalysts, air pollutants have been reduced to levels undreamed of even 10 years ago. Fuel efficiency has also increased greatly due to advances in technology. The average car today achieves almost twice the fuel economy of the average car 30 years ago, in addition to accelerating much faster. However, even with the efficiency increases, the average efficiency of a gasoline internal combustion engine in typical in-use operation is still only about 15%. The other 85% is lost to engine heat, heated exhaust gases, aerodynamic drag, tire rolling resistance, driveline losses, and braking.

Adding an electric motor and energy storage to the internal combustion engine can significantly improve efficiency in a variety of ways, depending on how the system is designed:

- Engines are least efficient when operating at low loads. The electric motor can be used to supply part or all of the propulsion energy at low speeds and loads, minimizing engine use under inefficient conditions.
- The electric motor can assist the engine during acceleration. This allows use of a smaller engine without any loss in overall performance. For a given load, smaller engines have better efficiency, due to lower frictional and heat losses.
- The high-power electric motor allows rapid engine restarts. This allows the engine to be shut off at idle and avoids fuel consumption while the vehicle is stopped.
- The electric motor can be used to capture regenerative braking energy. This is done by using the vehicle's inertia to drive the electric motor in reverse while the vehicle is slowing down, thus creating free electric energy.
- The additional electric power from the motor/generator can be used to replace mechanical and hydraulic devices and pumps with more efficient electric versions.

The advantages of hybrid systems have long been recognized. The first hybrid vehicle was built in 1898 and several manufacturers sold hybrid vehicles in the early 1900s. However, hybrid vehicles also have significant problems. They require two propulsion systems, which take up room, add weight, and greatly increase the cost. Another problem is that careful coordination of the operation of the engine and the motor is necessary to achieve much of the efficiency benefits and to avoid driveability problems. This was not possible with mechanical controls, and the production of hybrid vehicles did not survive continued development of the internal combustion engine in the early 1900s.

Renewed interest in hybrid vehicles has coincided with development of computer controls and improved batteries, combined with increasing concerns about the contribution of carbon dioxide emissions to global warming. Sophisticated computer controls allow maximum efficiency benefits while providing smooth, seamless coordination of the two propulsion systems. Advanced batteries, such as those made with nickel–metal hydride (NiMH), provide higher energy density and much longer cycle life, which is important to customer acceptance of hybrid technology. Hybrid vehicles offer a way to reduce fuel use significantly, with corresponding reductions in global warming gases and fuel cost to consumers, plus modest reductions in criteria air pollutants. The primary concern with hybrid systems remains the cost of the additional components.

2. HISTORY

2.1 Early Hybrid Electric Vehicles

From about 1890 through 1905, electric vehicles, internal combustion-powered vehicles, and steam cars were all competitively marketed and sold in the

United States. Electric vehicles had an early development lead in the United States due to the work of electricity pioneers, such as Edison, Tesla, and Westinghouse. Also, the limited driving range of electric vehicles was sufficient for the small U.S. cities in the late 19th and early 20th centuries, and the great distances between cities had roads that were largely inadequate for vehicle travel. However, it was obvious from the beginning that batteries severely limited the range and utility of electric vehicles. Although most developers went straight to internal combustion-powered vehicles, some did try to marry the advantages of the electric vehicle and the internal combustion engine.

Justus Entz probably built the first hybrid vehicle in 1898 for the Pope Manufacturing Co. in Connecticut. The vehicle caught fire during the initial test drive and Pope Manufacturing abandoned the concept. The first production hybrid vehicles were made in Europe in the early 1900s. Camile Jenatzy demonstrated a parallel hybrid system at the 1901 Paris Auto Show, with a 6-hp engine and a 14-hp motor/generator. In 1903, the Lohner–Porshe group brought out a series hybrid system, using a 20-hp engine to drive a 21-kW direct current (DC) generator and an electric motor in each of the two front wheels. A few years later, the Mercedes Company teamed with Mixte to introduce the Mercedes-Mixte. To improve the limited range of electric vehicles, two well-known U.S. electric car manufacturers, Baker and Woods, independently developed hybrid vehicles in 1917. However, after some initial development and vehicle operation, both concluded that the hybrid approach added too much complexity, weight, and cost.

2.2 Modern Development of Hybrid Vehicles

Air pollution concerns in the late 1960s and early 1970s spurred renewed interest in electric vehicles and hybrid electric vehicles. At that time, many people believed the internal combustion engine was inherently highly polluting and that it would need to be replaced in order to clean up the air. Battery–electric vehicles were a logical candidate because they have no emissions while in operation, although there is some offset from increased emissions from electric power plants. However, the short range and high cost of batteries continued to be insurmountable problems. To extend the range of primarily electric vehicles, a substantial number of concepts were introduced at the time, with small gasoline engines that could be turned on when needed. However, the large battery packs required for this type of vehicle were too expensive and degraded performance too much to be commercially viable.

The oil crises of 1973 and 1979 turned the auto industry on its head. Gasoline shortages, fuel price spikes, and predictions of continued oil shortages and increasing fuel prices caused the public to suddenly demand higher efficiency vehicles. Congress responded in 1975 with the Energy Policy and Conservation Act (EPCA). The cornerstone of EPCA was Corporate Average Fuel Economy (CAFE) standards, which mandated that cars double their average fuel economy by 1985 and that the Department of Transportation set cost-effective standards for light trucks.

As fuel prices went down in the early 1980s and stayed down the entire decade, the buying public gradually forgot about fuel economy. By 1990, vehicle purchasers were back to demanding other attributes they valued more highly than fuel savings, such as luxury, performance, and utility. The average fuel economy of new cars and light trucks, combined, peaked in 1987 and has been slowly declining ever since. Against the declining public concern with fuel economy was increasing concern about a new problem, global warming. Every gallon of gasoline burned produces about 20 lb of carbon dioxide, the primary greenhouse gas accused of increasing global temperatures. In addition to global warming concerns, the Gulf War of 1991 provoked concerns about national energy security.

Faced with such concerns, several attempts were made in Congress to raise CAFE standards beyond the levels set during the early energy crisis years. However, in Congress, there was also significant opposition to raising the CAFE standards; resistance came from a coalition of members concerned about the health of the domestic auto industry and general antiregulatory interests, and from those who did not believe that global warming is a significant concern.

2.3 Supercar

Rather than try to push a mandatory program through Congress, President Clinton and Vice-President Gore struck a deal with General Motors (GM), Ford, and Chrysler. If the domestic automakers would promise to try to build ultraefficient cars, the Clinton administration would back off from its campaign promise to raise CAFE standards. In the

fall of 1993, the President announced that the U.S. government and the American auto industry would work together to build a family car that got 80 mpg, or about three times the mileage of an average car. Officially named the Partnership for a New Generation of Vehicles (PNGV), most people soon dubbed the program "Supercar." The partnership spent 9 years and about $3 billion, roughly half each from the government and industry, on developing a super-efficient hybrid vehicle. General Motors, Ford, and Chrysler each developed a working hybrid demonstration vehicle, as illustrated by GM's 80-mpg Precept concept vehicle (Fig. 1). Although the program spurred advances in materials and hybrid design, it did not produce a vehicle that the automakers felt could be mass-produced at a price consumers would be willing to pay.

The primary problem with the Supercar program was the arbitrary goal of 80 mpg. This goal forced the automakers to choose options that were not cost-effective or were not marketable, such as diesel engines, lightweight materials, and extreme aerodynamic improvements. However, the program did spur significant advances in hybrid design that are being incorporated into future hybrid vehicles.

The announcement of the PNGV program also helped spur Toyota into designing its own hybrid vehicle, the Prius, in secret. Toyota's goal was a more realistic 55 mpg, or twice the mileage of an average car, using a highly efficient gasoline–electric hybrid design. The initial design was tailored to congested urban driving conditions in Japan and performance was not adequate for the U.S. market. However, it was sold in Japan while Supercar was still establishing concepts. Toyota redesigned the hybrid system on the Prius 2 years later to increase performance to levels acceptable to U.S. customers. In the meantime, Honda developed its own hybrid vehicle, the Insight, and beat Toyota to the U.S. market in late 1999.

FIGURE 1 The GM Precept concept hybrid vehicle.

3. HYBRID ELECTRIC SYSTEM DESIGN

The amount of power needed to propel a vehicle varies greatly, depending on the driving requirements. The vehicle must be designed to meet the most demanding requirements, such as hard accelerations under fully loaded conditions. However, this amount of power is needed infrequently. The large majority of driving requires only a small fraction of the available power from the engine. With a hybrid electric system, a smaller, more efficient engine can be used, with an electric motor added to restore performance.

Other types of hybrid systems have been considered. For example, Chrysler built a prototype vehicle in the early 1990s that combined a combustion engine with a flywheel that stores mechanical energy and provides power to the wheels. More recently, Ford and the U.S. Environmental Protection Agency (EPA) cooperated on a hybrid that combines an internal combustion engine with a hydraulic/nitrogen gas energy storage system. This system can provide considerable power from a stop and may find an application for helping to launch heavy-duty trucks.

For the large majority of applications, only the hybrid electric vehicle offers improved fuel economy and adequate performance at a cost customers might be willing to pay. In addition, the high amount of electrical power available on hybrid electric vehicles can be used to meet the increasing electrical power demands associated with other vehicle features, potentially adding value beyond just higher power-train efficiency.

3.1 Types of Hybrid Electric Systems

Although the ways in which the electric motor and the engine can be designed, and how they interact, are virtually limitless, the two basic configurations are series and parallel hybrids.

3.1.1 Series Hybrid

In the series hybrid, the engine is connected to a generator, which in turn is used to supply the motor and/or charge the battery (Fig. 2). Only the electric motor is directly used to drive the wheels. The battery pack can be charged by the engine and generator set and by using the generator to capture energy normally lost to braking. The principal advantage of this configuration is that the engine can be run in more efficient operating zones and shut off when it is not needed. The system also eliminates the need for

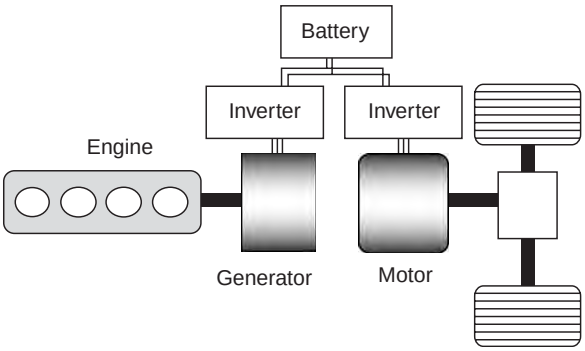


FIGURE 2 Series hybrid system diagram. Courtesy of Toyota Motor Company.

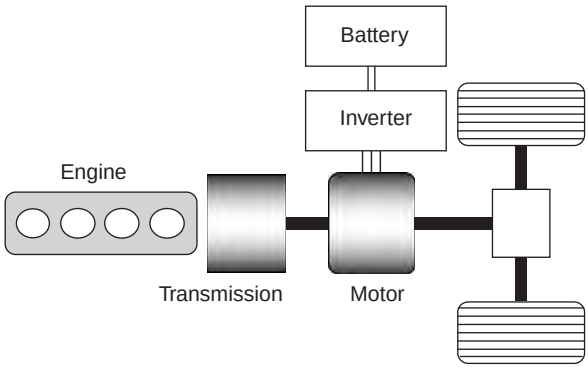


FIGURE 3 Parallel hybrid system diagram. Courtesy of Toyota Motor Company.

clutches and conventional transmissions. The need to convert the output of the engine to electricity before driving the motor creates significant inefficiencies if the engine is used as the primary power source, so this configuration is usually set up with a small engine that provides a range booster to a large battery pack. This configuration has very low emissions and has excellent efficiency during slow stop-and-go driving. However, it requires a large, heavy battery pack, which significantly increases cost and reduces vehicle performance. It is also much less efficient during high-speed driving, due to losses in converting the mechanical power from the engine to electricity and in charging and discharging of the battery. In addition, the internal combustion motor must still be large enough to meet minimum acceleration needs should the battery pack become depleted.

3.1.2 Parallel Hybrid

The parallel hybrid uses an engine and a motor, which are both connected directly to the drivetrain (Fig. 3). The highest efficiency range of each is selected and used, depending on operating condition. A parallel system is more complicated than a series system and there are additional challenges in integration of the two power sources. The major advantage is that large efficiency gains can be realized even with a relatively small, light battery pack. This minimizes the additional cost of the battery pack and the performance penalty from the weight of the batteries.

3.1.3 Variations on the Basic Design

The simplest way to arrange a parallel hybrid is to mount the electric motor in line with the engine. Another feasible arrangement is to power one set of wheels (either front or rear) with the conventional engine and the other set of wheels with an electric motor or motors. This system is referred to as a "through-the-road" parallel hybrid, because the two power sources are run in parallel by independently driving different wheels. The primary advantage of this system is that it adds part-time all-wheel-drive (AWD) capability. It also allows more flexibility in the mounting location of the electric motor. However, the system adds considerable complexity and cost. In addition, the AWD capability is limited to the energy in the battery pack and cannot be maintained indefinitely.

Another option is to add "plug-in" capacity. The basic hybrid system is self-contained, with all of the electricity generated by the on-board motor/generator while the vehicle is in operation. A vehicle with a battery pack that can be optionally recharged while the vehicle is parked, by plugging it into an electrical output, is commonly referred to as a grid-connected hybrid. Grid-connected hybrids can be used with either parallel or series design. The advantage of a grid-connected system is that it allows significant driving range in all-electric mode, without turning on the engine. Grid-connected hybrids have been promoted as a way to extend the range of a true electric vehicle, while still relying on the all-electric operation for urban operation with no combustion emissions at the vehicle. However, the value of electric-only operation has greatly diminished as ultraclean gasoline engines have become available. More importantly, the electric-only vehicle also suffers from many of the problems of a pure electric vehicle, such as the need for a large, heavy, and expensive battery pack and long recharge times.

3.2 Hybrid Components

3.2.1 Electric motor

Because the electric motor is used frequently, both for acceleration and for regenerative braking, efficiency

and high torque are important attributes. If the motor is placed between the engine and the transmission, then a compact design and high heat resistance are also important. There are currently two basic types of electric motors. One is an induction alternating current (AC) motor. This design uses part of the electric current to generate the magnetic field, within which the motor windings turn. The second design is a permanent magnet direct current motor. This type of motor generally uses three-phase windings and switching modules to control the flow of electricity sequentially to each phase. The AC design is simpler and cheaper. However, brushless DC (BLDC) permanent magnet motors are more efficient, more compact, and lower weight. The most efficient BLDC motors use magnets sintered with rare earth metals, such as neodymium. It is the cost of the magnet that makes DC motors more expensive.

Another important factor is the design voltage of the hybrid system. Higher motor power capacities are best met at higher voltages, because this reduces current and allows smaller wires and more efficient motors and inverters. The operating voltage can affect the choice of battery and the type of power electronics, which has cost implications. If the electric motor is located between the flywheel and the transmission clutch, it becomes very important to make the motor as thin as possible to minimize increases in powertrain length. This is especially important for transverse-mounted front-wheel-drive applications, in which the distance between the wheels is limited. Fig. 4. illustrates the placement of a compact motor design on the Civic Hybrid. The BLDC electric motor uses concentrated windings to conserve space (Fig. 5).

FIGURE 4 Electric motor placement on the Honda Civic Hybrid. Courtesy of Honda Motor Company.

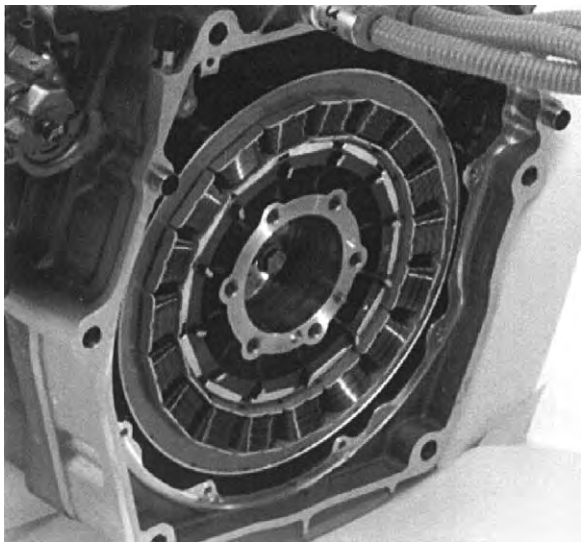


FIGURE 5 Thin electric motor designed to fit between the engine and transmission on the Honda Civic Hybrid. Courtesy of Honda Motor Company.

3.2.2 Energy Storage

Battery packs are the primary type of energy storage for hybrid electric vehicles. The only reasonable alternative is the ultracapacitor. Unless significant all-electric range is desired, energy requirements for hybrid electric vehicles are fairly modest. Energy storage is primarily used for acceleration, regenerative braking, and low-speed, low-power operation. At higher loads, it is more cost-efficient to use the engine to power the vehicle than it is to increase the size of the battery pack. Thus, hybrid vehicle electric power requirements are either brief or low load. Although some energy reserve is needed for extended hill climbs, energy storage demands are relatively small.

On the other hand, power requirements are substantial. If significant engine downsizing is to be achieved, the electric motor must provide at least a 50% torque boost at low engine speeds. Although the electric motor can be biased to provide its maximum power at relatively low engine speeds, this still requires an electric motor that has at least 15% of the power output of the engine. Even for a small car, the electric motor needs to be at least 10 kW (about 13 hp). The electric motor power requirements scale up with vehicle size and with the use of electric-only operation. Thus, batteries for hybrid vehicles must be able to provide or absorb high power rates to support acceleration and to recapture high amounts of energy during brief braking events.

Batteries developed for electric vehicles in the 1990s were optimized for energy storage, because they needed to maximize all-electric vehicle range. These batteries are not suitable for hybrid vehicles. Batteries were redesigned for hybrid vehicles to provide much higher power density, with the trade-off of lower energy density. Even with the redesign, most battery packs still must be sized to deliver maximum power without overheating. This results in more energy storage than is needed for the vast majority of driving conditions, which increases the cost and weight of the battery pack.

Ultracapacitors have very high power density. They excel at rapid transfer of power and offer the potential for higher durability and lower cooling requirements. They are ideal for assisting acceleration and at recapturing regenerative braking energy. Unfortunately, they have very low energy density and cannot store enough energy to cover longer load demands, such as hill climbs. The energy density is so low that no manufacturers are considering their use for hybrid vehicles, although some manufacturers are using them to provide momentary assist on fuel cell vehicles.

3.2.2.1 Lead-Acid Battery The primary advantage of the lead-acid battery is low cost. Being developed for hybrid vehicles are advanced lead-acid batteries with higher power, such as the valve-regulated lead-acid (VRLA) battery. However, even VRLA batteries have relatively low energy density and short cycle life. The short cycle life is a particular concern, because periodic replacement raises the effective cost of the system. It is uncertain how customers will react to the cost and hassle of periodically replacing the hybrid battery pack. Despite the concerns, lead-acid batteries will be used in some relatively low-power hybrid vehicle applications due to their low cost.

3.2.2.2 Nickel–Metal Hydride Battery NiMH batteries have higher power and energy density and a much longer life cycle compared to lead-acid batteries. They are also completely safe and their power output is not affected by the battery state of charge. The main concern with nickel–metal hydride batteries is that they are very expensive. Other concerns are high self-discharge rates, low-temperature operation, and higher cooling requirements. NiMH batteries are used in higher power hybrid designs, such as the Toyota Prius, the Honda Insight and Civic Hybrid, and the upcoming small sport utility hybrid vehicles from Ford and General Motors.

3.2.2.3 Lithium-Ion Battery Lithium-ion batteries have higher energy and power density compared to NiMH batteries. They also have better low-temperature performance and a low self-discharge rate. However, they need improvements in durability and cost before they will be commercially viable for hybrid electric vehicles.

3.2.3 Power Electronics

Direct current motors generally use three-phase motor windings. Three switching modules are needed to control current flow to the proper windings as the motor turns. These switches must be able to turn large amounts of current on and off rapidly and accurately, as well as to control the direction and phase of the current. This operation produces a substantial amount of heat and the power control unit must be cooled to keep the heat from affecting the electronics. The size and cost of the power electronics are substantial, due to the high power demands of the system and the need for cooling.

The power control unit includes a DC–DC converter to supply 12 V for the conventional electrical system. Improvements to the switching method are being incorporated to reduce switching losses and increase the efficiency over the output range of the converter.

3.2.4 Cooling System

Both the battery pack and the power electronics must be cooled. Charging and discharging a battery is a chemical reaction, which generates heat as a by-product. High battery temperatures degrade performance and shorten battery life. Power electronics generate a lot of heat and do not function properly if they get too hot. Efficient cooling systems are needed for both the battery pack and the power electronics. Due to the lower heat generated by newer power electronics, it is now possible to cool the power electronics with air discharged from the battery. Combining the battery pack and power electronics into a single unit with an integrated cooling system offers substantial reductions in overall size and weight.

3.3 Fuel Efficieny Benefits

3.3.1 Engine Downsizing

The greatest demands on horsepower and torque occur while accelerating and climbing grades. Minimal power is needed to maintain a vehicle's speed while cruising on a level road. By using an electric motor to provide a power boost to the engine when appropriate, a smaller engine can be used. Other

things being equal, a smaller engine is more efficient for a given load, because it has lower frictional losses, less heat loss to the engine block and cylinder head, and larger throttle openings, which reduce the energy lost in forcing intake air past the throttle (commonly referred to as pumping losses).

3.3.2 Regenerative Braking

A large amount of energy is lost in conventional vehicles when the vehicle slows down or brakes. This energy is primarily consumed by braking and engine friction, which includes mechanical friction and pumping losses. The electric motor can be used as a generator to capture this energy. When the vehicle is slowing down or braking, the electric motor captures—or regenerates—electrical energy from the vehicle's forward momentum that would otherwise be wasted. This energy is stored in the battery pack until it is reused during acceleration, when power from the gasoline engine is boosted by power generated by the electric motor. This process is referred to as "regenerative braking." If the electric motor is bolted to the engine, then some of the potential regenerative braking energy is lost to engine friction. Installation of a clutch between the engine and the electric motor eliminates engine friction during braking and increases the amount of energy available for recharging the battery. However, the clutch adds cost and increases packaging problems, especially for front-wheel-drive vehicles.

3.3.3 Idle Off

Hybrid systems allow the engine to be shut off at idle, eliminating idle fuel consumption and emissions. The powerful electric motor can restart the engine far more quickly than can a conventional starter motor. In most cases, the motor is powerful enough to spin the engine up to normal idle speed in only about 0.2 sec. This allows smooth, rapid restarts that are acceptable to the driver. In addition, fuel injection usually does not start until after the engine is already up to normal idle speed, which virtually eliminates unburned fuel and its attendant hydrocarbon emissions. The average vehicle sits and idles about 20% of the time, thus turning off the engine at idle can reduce fuel consumption by about 5–10%.

3.3.4 Engine Efficiency

Proper integration of the electric motor and the internal combustion engine can have a major impact on the efficiency of the engine during different driving conditions. Optimization of the overall system allows the engine to operate in more efficient modes. For example, the high torque of the electric motor allows the engine to turn more slowly, and hence more efficiently, during highway driving while still maintaining adequate acceleration. Another example is careful integration with transmission operation, to keep the engine running at higher efficiency speed and load points.

The boost from the electric motor can also facilitate use of innovative engine designs, such as the Atkinson cycle gasoline engine. The Atkinson cycle uses a longer expansion stroke to extract more energy from the combustion process and boost efficiency. The downside of the Atkinson cycle is that it generates much lower peak torque and horsepower. This is more acceptable on a hybrid vehicle because the electric motor can provide power boost during acceleration.

3.3.5 Electrical Accessories

The additional electric power can be used to improve the efficiency of engine accessories, such as the air conditioner, power steering pump, and water pump. Currently, most accessories are driven by mechanically connecting them to the engine with belts, gears, or chains. The mechanical efficiencies are often very low. Another problem is that the accessory speed goes up and down with the engine speed. This is very wasteful, because the accessory must be sized to provide adequate operation at low engine speed. Using electric motors to power accessories allows much more efficient designs, because they can be operated independently of the engine and only as needed.

On conventional vehicles, the use of electric accessories is limited by alternator load, which is already reaching the limits of a 12-V alternator. The vehicle electric demand has increased from a few hundred watts 10 years ago to over a kilowatt today. Inducing high currents in a 12-V system results in high efficiency losses and requires large, heavy, and expensive electrical cables. The high power and higher voltage available from the hybrid electric motor can enable use of electric air-conditioning compressors and electric pumps for power steering and circulation of fluids.

3.4 Design Considerations

Design of hybrid systems is an exercise in trade-offs. Larger electric motors and battery packs allow more acceleration assist and smaller engines, recapture more regenerative braking energy, and enable some all-electric operation to replace the most inefficient engine operation modes. Very large battery packs and

motors allow off-board recharging from the electric grid. However, battery packs are very expensive and their additional weight affects performance. Thus, it is highly desirable to limit the size of the battery pack. The type of battery also has a profound impact on durability, size, and cost. Finally, the design and sophistication of the hybrid system, including individual components and how they are integrated, can have a greater impact on fuel efficiency than the simple inclusion of individual hybrid features. Juggling the relative benefits, costs, and packaging of all the different options is difficult. Each manufacturer has a different idea of how to balance the factors. Ultimately, the market will determine what works and what does not.

Hybrid vehicles can incorporate different fuel-efficient techniques, such as idle off, regenerative braking, engine downsizing, motor-only operation, and off-board recharging. Except for idle-off, each of these features can be utilized to different levels.

3.4.1 Integrated Starter Generator

The most basic hybrid type is the integrated starter generator. These are conventional vehicles that have a higher power electric motor that can turn the engine off when the vehicle is stopped and restart it as needed. The starter generator motor can be mounted on the driveshaft or it can be connected with a belt. Most systems under development use a 42-V lead-acid battery pack, although 12-V systems are also being evaluated. These systems are relatively inexpensive and can provide some additional power for accessories. However, the fuel economy improvement is generally only 5–10%. An integrated starter generator system generally is not classified as a hybrid vehicle unless its electrical motor also provides propulsion assistance.

3.4.2 Full-Function Hybrid

The next step is to use a larger, full-function electric motor and a high-voltage battery pack to provide assist on all accelerations. This design significantly increases the cost of the system, but it greatly expands the amount of regenerative braking, enables engine downsizing, and allows integration of the engine and motor functions. Depending on the sophistication of the system and the amount of engine downsizing, fuel economy improvements can range from about 20 to 50%.

Trade-offs between fuel economy and performance are also possible. One interesting design concept is to add an electric motor without downsizing the engine. This can dramatically increase vehicle performance, especially at low speeds, when the electric motor has very high torque output, while still providing significant fuel savings from idle-off and regenerative braking. The result can be thought of as an environmental supercharger. Because performance is valued more highly than fuel economy by most U.S. new-vehicle purchasers, this may be a way to help public acceptance of hybrid vehicles.

3.4.3 Motor-Only Operation

The next step in design sophistication is to add motor-only operation. These systems use the electric motor to propel the vehicle at low speeds and loads, eliminating engine operation in its most inefficient modes. Unfortunately, the incremental efficiency gains from motor-only operation are modest. There is little efficiency benefit from motor-only operation above about 20 mph, because the engine can operate reasonably efficiently at higher speeds. Even below 20 mph the engine must be used for harder accelerations, so the primary efficiency benefits are in congested, low-speed urban driving, such as those encountered in Japan. In addition, large amounts of motor-only operation require larger battery packs, driving up cost and weight. Limited use of motor-only operation at low speeds can achieve some efficiency improvement at reasonable cost, but the overall impact under typical U.S. driving conditions is relatively small.

3.4.4 Grid-Connected Hybrids

The final step in hybridization extends the motor-only range by allowing the battery to be recharged from the electric grid. However, the value of electric-only operation has greatly diminished as ultraclean gasoline engines have become available. In any case, the very high cost and weight of the large battery packs needed for these systems generally preclude their use in production.

4. REAL-WORLD HYBRID EXAMPLES

In the real world, engineers can get creative about applying hybrid concepts to actual vehicles. Hybrid vehicles in production or being readied for production in the United States illustrate a number of innovative hybrid techniques.

4.1 Toyota Prius

The Prius was the first modern hybrid to be placed in production, starting sales in Japan in late 1997. A

version for the United States was introduced in 2000 (Fig. 6). Rather than choose between series and parallel systems, Toyota developed a new type of hybrid system that combines some of the advantages of both (Fig. 7). The Prius powertrain is based on the parallel type. However, it uses a planetary gear system and a separate generator to optimize the engine's operation point and to allow series-like operation. Toyota's hybrid system requires a power split device and an extra generator and inverter. However, it offers great flexibility in engine operation and calibration. The system also acts as a continuously variable transmission (CVT), eliminating the need for a conventional transmission.

The Prius used a relatively small NiMH battery pack, rated at 288 V and about 2 kWh of storage capacity. The electric motor is rated at 44 hp, allowing limited acceleration and cruise at light loads and low speeds on just the electric motor. The Prius also uses an Atkinson cycle engine for maximum efficiency. The overall design excels at low-speed fuel efficiency and the city rating of 52 mpg is much higher than the highway rating of 45 mpg. These values represent a fuel economy improvement of about 80% on the U.S. EPA city driving cycle, although less than 20% on the U.S. EPA highway cycle. On the Japanese driving cycle, involving relatively low speeds and many starts and stops, the Prius offers about a 100% improvement over a car of comparable size and performance.

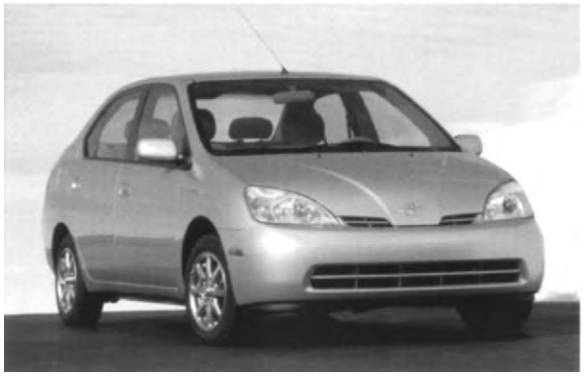


FIGURE 6 The Toyota Prius. Courtesy of Toyota Motor Company.

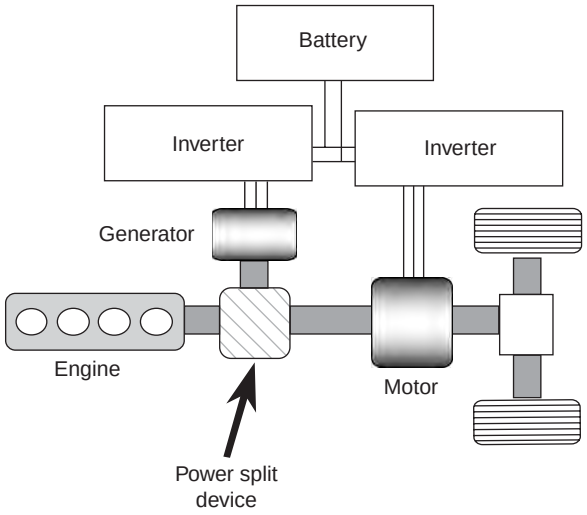


FIGURE 7 Hybrid system for the Toyota Prius, illustrating the planetary gear system and separate motor and generator. Courtesy of Toyota Motor Company.

4.2 Honda Insight and Civic Hybrid

Honda was the first company to market a modern hybrid vehicle in the United States, introducing the Honda Insight in late 1999 (Fig. 8). Honda developed an electric motor assist system for the Insight (Fig. 9), which integrates a high-torque, high-efficiency brushless DC motor between the engine and the transmission. This 10-kW (13-hp) motor is only 60 mm (2.4 in) thick and is connected directly to the engine's crankshaft. This is a straightforward method for packaging a parallel hybrid system and minimizes the weight increase.

The Insight's 144-V NiMH battery pack is rated at about 1 kWh of storage and weighs only about 22 kg (48 pounds). This helps to maintain in-use performance and efficiency while maintaining most of the hybrid system benefits. The Insight also incorporates a wide range of other efficiency technologies, such as an all-aluminum body, aerodynamic improvements, and a 1.0-liter three-cylinder engine with variable valve timing, reduced friction, and lean burn. With a manual transmission, the fuel economy ratings are 61 mpg in the city and 68 mph on the highway, the

FIGURE 8 The Honda Insight. Courtesy of Honda Motor Company.

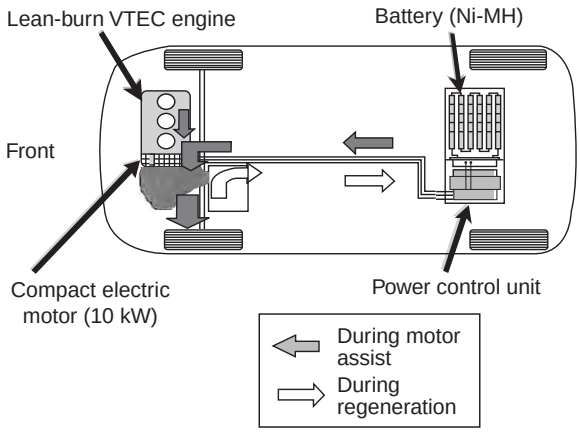


FIGURE 9 Hybrid system layout for the Honda Insight, with a variable valve-timing and lift electronic control (VTEC) engine and a nickel–metal hydride (NiMH) battery. Courtesy of Honda Motor Company.

best fuel economy ratings ever recorded by the EPA for a production vehicle.

In 2003, Honda expanded the integrated motor-assist (IMA) system to the Civic. This was the first time a hybrid system was offered as a powertrain option on a conventional vehicle. A number of improvements were made to increase motor torque, integrate many of the functions of the battery pack and power electronics, and improve battery pack efficiency. The Civic Hybrid also has a cylinder-idling system that deactivates valves on three of the four cylinders during deceleration and braking, cutting engine friction by 50% and allowing more power to be recaptured by the electric motor. The Civic hybrid with a manual transmission is rated at 46 mpg city and 51 mpg highway. With the automatic continuously variable transmission (CVT) option, fuel economy is 48 mpg city and 47 mpg highway. These values represent fuel economy improvements of about 55% (average of manual and automatic) on the U.S. EPA city driving cycle and about 30% on the U.S. EPA highway cycle.

4.3 Toyota Estima Minivan

Introduced in 2002 and sold only in Japan, the Toyota Estima small minivan is the first through-the-road parallel hybrid vehicle. The front powertrain combines a gasoline engine with a 216-V, 13-kW, belt-driven integrated starter generator. A second 216-V electric motor, rated at 18-kW, provides on-demand power to the rear wheels for temporary four-wheel-drive operation. The computer can choose any combination of the three power sources and either of the motor/generators can be used to recharge the NiMH battery pack.

4.4 Future Compact Sport Utility Vehicle Hybrids

Ford has announced plans for a hybrid versions of their small sport utility vehicle (SUV), the Ford Escape, in 2005. The hybrid system includes a 300-V NiMH battery pack, an efficient electric air-conditioning compressor that operates even when the engine is turned off at idle, and relatively large electric motors that provide motor-only operation at low speeds and loads. The drive motor is rated at 65 kW (87 hp), plus it has a 28-kW motor/generator. The Escape hybrid combines the electric motor, separate generator, power electronics, and planetary gear into a single unit that takes the place of a conventional transmission. This arrangement provides flexible power combinations and is similar in concept to the planetary gear system used in the Toyota Prius.

Toyota plans to produce the first through-the-road parallel hybrid system for the U.S. market based on its Lexus RX330 compact SUV in 2005. A V6 gasoline engine powers the front wheels while two electric motors, one in the front and one in the back, provide the vehicle with all-wheel-drive capability. This vehicle also illustrates a hybrid electric design optimized for performance, providing the vehicle with very high acceleration ability along with a significant boost in fuel economy. DaimlerChrysler also plans to offer a through-the-road hybrid system on future models, such as the Jeep Liberty and Dodge Durango.

4.5 Hybrids as Electric Generators

Both General Motors and DaimlerChrysler plan to offer similar 42-V hybrid systems on full-size pickup trucks in 2004. An AC induction motor/generator is integrated into the drivetrain and is used for idle-stop and limited acceleration assist. Both use 42-V lead-acid batteries, which are recharged with regenerative braking. The fuel economy improvement is relatively modest, about 10–15%, because the engines are not downsized, in order to preserve cargo-hauling and trailer-towing abilities. However, the fuel economy improvement is secondary to the on-board electrical-generating capability provided from the on-board engine and generator on these vehicles. The GM pickups provide continuous 20-amp power through

each of two 110-V outlets. The Dodge Ram can provide up to 20 kW of either 110- or 220-V power. These systems are primarily targeted at construction site work, but they may also appeal to campers and outdoor enthusiasts and to farmers and homeowners for backup power in case of a power failure.

4.6 Military Applications

Military actions have some unique characteristics that make hybrid vehicles very desirable. For example, fuel may need to be transported hundreds, or even thousands, of miles in overseas military action. Fuel is the single largest supply component in a military action and the price of delivering a gallon of fuel to a battlefield can easily exceed $30 per gallon. Thus, even a 10–15% improvement in fuel efficiency is extremely desirable to the military, as long as it does not impact performance and utility.

The military has to furnish its own electricity in a hostile country and the use of electronic controls and components for weapons, intelligence, and communication is accelerating. Currently, the only way to meet these rising electrical needs is to haul around generators, which could be eliminated by generating electricity directly from hybrid vehicles. Importantly, the military is also very interested in the stealth capability of hybrid powertrains. The battery pack can supply electricity to power on-board electronics for substantial periods, eliminating the noise and heat signature from a running engine. Military vehicles would also have the ability to travel limited distances on the electric motor alone, making it easier to move equipment without detection.

4.7 Urban Buses and Delivery Vehicles

Urban buses and delivery vehicles are ideal applications for hybrid vehicles. The electric motor is very good at launching a vehicle and accelerating at slow speeds, plus the frequent stops in urban operation offer much larger benefits from regenerative braking. In addition to larger fuel savings, brake maintenance is also reduced because the regenerative braking system helps to slow the vehicle down. Because brake repairs are the single largest maintenance item for urban buses, this is a significant cost saving.

As of early 2003, roughly 200 urban buses with hybrid electric powertrains had been placed in demonstration fleets. Results from the demonstration fleets have been positive, and Orion Bus Industries has a contract to deliver 325 production buses with a series hybrid system, diesel engine, and lead-acid battery pack to New York City Transit.

The first urban delivery hybrid application is FedEx's recent purchase from Eaton Corporation of 20 delivery trucks with parallel hybrid systems. The trucks use a smaller, four-cylinder diesel and fuel economy is expected to increase by 50%. Eaton is using lithium-ion batteries, which will be the first market experience with the durability of lithium-ion batteries in a hybrid vehicle. If these trucks meet project goals, the program has the potential to replace FedEx's 30,000 medium-duty trucks with hybrid vehicles over the next 10 years.

5. HYBRID ADVANTAGES AND DISADVANTAGES

5.1 Advantages

Hybrids have a number of positive features that are desired by customers, in addition to the obvious fuel savings:

- Hybrids use widely available fuel, thus there are no concerns about creating a new infrastructure to support fueling or recharging.
- The customer benefits from extended range and fewer trips to the gas station.
- The vehicle is quieter inside; both at idle when the engine is turned off and on the highway, where electric motor assist allows the engine to run at slower speeds.
- Hybrid systems have good synergy with other fuel-economy technologies. For example, not only could the hybrid system supply electrical power for fully adjustable electromechanical valves, but the motor and the valve timing could also be coordinated to optimize the benefits from the adjustable valves.
- Emissions can be reduced in a number of ways. Upstream emissions from fuel production and distribution are reduced in direct proportion to the reductions in fuel consumption. Increased driving range reduces evaporative emissions from refueling. Hybrid systems also reduce engine-out hydrocarbon and NO_x emissions. A conventional starter spins the engine up to only about 200–250 rpm and fuel must be combusted to propel the engine up to normal idle speeds. Combustion is very unstable at such low engine speeds and much of the fuel during the starting process goes out the tailpipe unburned, as hydrocarbon emissions. The greater power available from the hybrid electric motor can

quickly spin up the engine and fuel is not injected until after the engine has reached normal idle speed. This allows instantaneous combustion and reduced tailpipe hydrocarbon emissions. Hybrids also offer reductions in engine-out NO_x emissions. NO_x is formed when nitrogen and oxygen in the air are compressed at high temperatures. NO_x formation correlates closely with engine load, such that far more NO_x is generated during acceleration than it is during cruises. The hybrid system reduces the load on the engine during acceleration, with a corresponding reduction in NO_x formation. These reductions in engine-out emissions make it easier to achieve the ultralow levels that will be required starting in 2004.

- There is little impact on how the vehicle operates. The vehicle drives and operates similar to a conventional vehicle.
- Brake life is extended, because the electric motor captures some of the energy that is usually dissipated in the brakes.
- Part-time all-wheel drive can be added, if desired, by using a through-the-road hybrid system.
- If desired, the motor can be used to boost performance instead of reducing engine displacement.
- The motor can supply electricity for features desired by customers, such as off-board electricity generation, heated seats and windows, brake- and steer-by-wire systems, multimedia components, dynamic ride control, and radar and camera safety/warning systems. This feature is greatly underappreciated because electrical technology and accessories do not currently exist to take advantage of the high power from hybrid vehicles. A growth in customer demand for high-power features could be the "ace in the hole" for hybrid system popularity.

5.2 Potential Barriers

Traditionally, customer distrust of new technology has slowed market penetration. However, public reaction to introduction of hybrids in the United States has been surprisingly positive. Both the Insight and the Prius exceeded their original sales targets and public surveys indicate widespread knowledge and acceptance of hybrid technology. For example, a survey conduced by J. D. Power and Associates in 2002 found that 60% of new vehicle buyers would strongly consider a hybrid electric vehicle.

There are claims that development resources would be better spent on fuel cells. Certainly fuel cells are a promising long-term option. Hydrogen fuel cells have virtually no emissions and are extremely efficient. Large-scale production of hydrogen would use resources other than oil. However, fuel cells still have a lot of issues to resolve. Cost and size must be dramatically reduced, plus major breakthroughs are needed to enable practical on-board hydrogen storage. Durability must also be proved. Even after all these problems are solved, the problems and cost of creating an entirely new infrastructure must be addressed. Thus, fuel cell vehicles, at best, will need a long time for development and may yet turn out to be a dead end. Hybrids can help fill the gap until fuel cells and their related infrastructure are ready. In fact, hybrids might become the preferred technology until nonfossil sources of hydrogen are ready, especially because widespread adoption of hybrid vehicles will reduce the benefits of switching to fuel cell vehicles. Hybrids are also a useful hedge in case fuel cell development hits an insurmountable barrier. Finally, hybrid vehicles are a good path toward fuel cell vehicles. Many of the hybrid technologies can likely be incorporated into fuel cell systems.

Room must found for the electric motor, battery pack, and power electronics. This is not a problem on larger vehicles, but can pose challenges for smaller cars. A small reduction in interior space can be expected on smaller vehicles. The additional hybrid components also add some additional weight to the vehicle, which reduces performance. For example, the hybrid system adds about 100 lb to the Civic. However, this effect is relatively minor compared to the performance and efficiency boosts from the electric motor.

Many of the benefits of reducing fossil fuel use are shared by all of society, but the new vehicle purchaser must bear the burden of high first cost. Experience with vehicle emission standards has shown that, despite widespread support for pollution controls that raise the cost for everyone, few customers are willing to pay more individually for vehicles with lower emissions. New vehicle purchasers do respond rationally to the cost of putting fuel in their vehicles. This means that they severely discount expected lifetime fuel savings, both due to the time value of money and due to expected sale of the vehicle in a few years. It also means that most new vehicle purchasers are relatively indifferent to technologies that enhance fuel economy. The cost of driving a mile today is only about half what it was in the early 1970s before the first oil crisis, because fuel prices, adjusted for inflation, are the lowest in history and new vehicles are far more efficient than they were 30 years ago. Given the multitude of other factors affecting purchase decisions, it is no

surprise that efficiency ranks very low in the car-buying process.

Increased costs due to long-term maintenance and repair of hybrid components are another potential barrier to acceptance of hybrid electric vehicles. Although the electric motor and power electronics are very reliable and durable, there is a potential concern with battery packs. Batteries in hybrid vehicles should be far more durable compared to the batteries in electric vehicles. This is because most battery deterioration occurs at the lowest 20% and highest 20% of the battery state-of-charge. These conditions are unavoidable with electric vehicles, which already suffer from limited range. However, the batteries on hybrid vehicles are carefully controlled to always stay in the middle 60% range, allowing much longer battery life. Still, hybrid batteries are not expected to last the life of the vehicle and are expensive to replace. NiMH batteries last a lot longer than do lead-acid batteries, but cost a lot more when they do finally need to be replaced.

The bottom line is that there is really only one factor holding back hybrid systems—cost, including battery replacement costs. Unfortunately, hybrid systems are not cheap. Manufacturers are understandably reluctant to discuss the cost of their hybrid systems, so it is difficult to determine a realistic cost. The Electric Power Research Institute (EPRI), in a study in 2002 on hybrid costs, estimated that the incremental lifetime cost for a high-volume, full-function hybrid system, including in-use battery replacement, would be about $2500 to $3600 for a compact car, $4000 to $5500 for a midsize SUV, and $4500 to $6300 for a large SUV. To put this estimated hybrid cost into context requires looking at what customers might be willing to pay in exchange for the fuel savings, both in the United States and worldwide. To do this necessitates making a few assumptions. The most critical is customer discounting of fuel savings.

In the March 2002 issue of *The Power Report*, J. D. Power reported some findings from their research on consumer interest in hybrid vehicles. One was that about one-third of potential hybrid buyers would still buy a hybrid even if the savings from reduced fuel costs during their ownership period would be less than the extra cost of purchasing the hybrid option. This is encouraging, in the sense that there are some customers who are willing to pay extra for hybrids. However, if this statement is turned around, it implies that two-thirds would not buy a hybrid vehicle if the extra cost of purchasing it were more than the fuel savings just during their ownership period. This is roughly equivalent to assuming that most customers value the fuel savings only for about the first 50,000 miles of driving.

The following formula, using miles per gallon (mpg), is used to calculate the fuel savings from hybrid vehicles:

$$\left[\frac{50,000\ \text{miles}}{\text{baseline mpg}}\right] - \left[\frac{50,000\ \text{miles}}{\text{baseline mpg} \times (1 + \%\,\text{mpg increase})}\right] \times \text{fuel price}.$$

Estimates using this formula were made for three different vehicle sizes: small cars, midsize cars, and large trucks. Estimates were also made for two different fuel prices: $1.50 per gallon to represent the United States and $4.00 to represent Japan and many countries in Europe. A reasonable factor for the fuel economy improvement from just the hybrid system and corresponding engine size reduction is probably about 30–40% over EPA's combined cycles. Sensitivity cases of 20% (for very mild hybrids) and 80% (for advanced hybrid systems) are also shown in Table I. The results in Table I are supported by the J. D. Power survey on hybrids, which also found that consumers would be willing to pay about $1000 to $1500 more for a hybrid powertrain.

From a societal view, the fuel savings over the full life of the vehicle (which are about three times the values in Table I) are likely to justify the additional cost of hybrid systems. However, the typical new vehicle customer does not value the fuel savings enough to pay for the incremental cost. To address the cost differential, hybrid system costs need to come down or the price of gasoline must increase significantly.

Some customers who drive a lot or value the benefits and features more highly are willing to pay

TABLE I

Customer Value of Hybrid Fuel Savings[a]

Hybrid FE increase[b]	Fuel cost	Small car (34 mpg baseline)	Midsize car (23 mpg baseline)	Large truck (14 mpg baseline)
+20%	$1.50/gal	$368	$543	$893
	$4.00/gal	$980	$1449	$2381
+40%	$1.50/gal	$630	$932	$1531
	$4.00/gal	$1681	$2484	$4082
+80%	$1.50/gal	$980	$1449	$2381
	$4.00/gal	$2614	$3865	$6349

[a] Savings for the first 50,000 miles.
[b] FE, fuel economy.

the premium for a hybrid vehicle. Hybrid sales will increase as the costs come down. However, hybrids will probably not break into the mainstream market in the United States until the cost of hybrid systems comes down to less than $1500 and/or gasoline prices increase to over $3.00 per gallon. In the long term, increased demand for high-power features could increase the price customers are willing to pay for hybrid vehicles and make it easier to achieve the necessary cost reductions.

5.3 Practical Hybrid Designs

Because battery costs dominate customer acceptance of hybrid vehicles, it is not practical to design a hybrid system with a large battery pack for cars and light trucks. All light-duty hybrid vehicles will continue to be powered primarily by the internal combustion engine for the foreseeable future. Although manufacturers are introducing a wide variety of hybrid systems, they all share the basic design of a parallel system, with electric motors used primarily to assist the internal combustion engine and recapture regenerative braking energy. Some manufacturers choose to add motor-only operation at low-load, low-speed conditions and series hybrid operation under some conditions, but the efficiency gains are limited by the need to minimize the size and cost of the battery pack.

6. FUTURE TECHNOLOGY DEVELOPMENT

6.1 Electric Motor

The permanent magnet is one of the primary costs of a brushless DC electric motor. Because magnet production is a long-established and optimized technology, only modest cost reductions can be expected on magnets. Indeed, the electric motor in general is a mature technology. However, integration into a hybrid electric vehicle poses some unique packaging challenges, so the first hybrid electric vehicles are paying some additional premiums to cover development and tooling costs. These can be expected to come down in the future as production volumes increase. In the long run, switched reluctance motors may offer advantages over the brushless DC motor, because they provide good efficiency and reliability with reasonable cost and weight. They may offer a modest decrease in cost in the long term.

6.2 Energy Storage

Energy storage costs are the single largest long-term obstacle to widespread market acceptance of hybrid electric vehicles. Although lead-acid batteries are likely to be used initially in many hybrid vehicles due to their relatively low cost, their limited power density and short cycle life will force manufacturers to turn to advanced batteries in the future.

Nickel–metal hydride batteries are the near-term choice, due to their higher power density, longer cycle life, and better response to high-power pulses. Unfortunately, nickel–metal hydride batteries are expensive. Although some cost reduction is expected in the future, costs are likely to remain at high levels because of inherently high material costs. Nickel–metal hydride batteries also have high cooling requirements, due to high heat generation.

Lithium-ion batteries need improvements in cycle life and cost before they will be commercially viable for hybrid electric vehicles. However, lithium-ion battery use is rapidly growing in consumer markets, such as in laptops, and their cycle life and cost are expected to improve. Eventually, lithium-ion batteries should provide a lighter and slightly lower cost alternative to nickel–metal hydride, as well as better performance at low temperatures. Although development of lithium-ion batteries will improve the situation, their costs will likely still be a barrier to widespread acceptance of hybrid systems. A breakthrough in energy storage is needed if hybrid electric vehicles are to become universally accepted.

One possible long-term solution might be to combine a small battery pack with an ultracapacitor. Because of their relatively low power density, battery packs are sized to handle peak power demands during acceleration and regenerative braking. This leads to more energy storage than is necessary to support efficient hybrid operation. Ultracapacitors have very high power density, but can store only a small amount of energy. In theory, combining a battery pack with an ultracapacitor would allow the ultracapacitor to provide peak power for short acceleration and regenerative braking events, while the much smaller battery pack could recharge the ultracapacitor after acceleration and adsorb excess energy after regenerative braking. In essence, the ultracapacitor would act as a buffer for the battery pack, reducing the battery size requirements and energy storage system costs. This system is still in early stages of development and the savings in battery cost may prove to be less than the additional complexity and cost of the ultracapacitor.

Another possible technology is the solid-state battery, such as lithium-polymer batteries. Because they do not rely on liquids as the storage medium, solid-state batteries have the potential to be much smaller than conventional batteries. However, these batteries are still in their infancy and currently have lower energy density, shorter cycle life, and higher manufacturing cost compared to lithium-ion batteries.

6.3 Power Electronics

As with all electronics, advances in power electronics should continue steadily into the future, continuously improving performance and reducing cost. Higher temperature components are being developed that require less cooling, further improving performance and reducing cooling costs and packaging problems. The different components and modules are also being integrated to reduce size and complexity. Hybrid systems can also enable secondary improvements, by providing the electrical power needed to replace mechanical devices and pumps with superior electronic equivalents. This offers significant opportunities to reduce costs and provide features desired by customers.

7. FUTURE HYBRID MARKETS

Hybrid vehicles will continue to expand in niche applications. They offer a number of desirable features, in addition to fuel savings, and will appeal to a limited number of car and light-truck purchasers who value these features or the fuel savings enough to pay the increased price. Another niche market is through-the-road all-wheel-drive hybrid systems, which will appeal to some drivers in snowy climates. On-board electric generating will appeal to a small segment of contractors, farmers, and recreational users. Performance hybrids, in which the electric motor is used to boost acceleration instead of downsizing the engine, have not been tested in the market yet, but may appeal to another niche market of performance enthusiasts. A hybrid designed with both high-performance and part-time AWD might be especially appealing to SUV and luxury vehicle customers, because such a design could allow customers to have all the features they want, while still improving fuel economy by about 20%. Finally, although they do not represent huge markets, military, urban bus, and urban delivery applications are very promising and hybrids may come to dominate these applications in about 10 years.

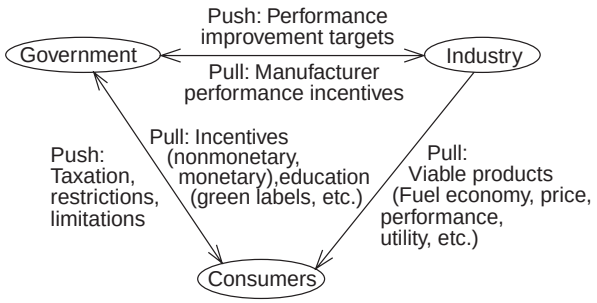


FIGURE 10 Schematic emphasizing the importance of interaction between consumers, government, and industry.

Hybrid system costs will come down as the market expands. The lower costs will, in turn, draw in more customers. J. D. Power predicts that hybrid vehicles will account for about 1%t of the market by 2005, 3% by 2009, and nearly 5% by 2013. However, in the long term, it is not clear if hybrid vehicles will stagnate at 5–10% of the market or if costs will eventually drop enough for hybrid vehicles to break into the mass market. An unknown factor is how much customer demand will increase for high-power features in the long term. Strong demand for high-power features would greatly help market acceptance of hybrid vehicles.

The industry can provide a market "pull" by providing products desired by the consumer. But customers cannot be "pushed" into buying vehicles they do not want. Government programs to stimulate demand, provide incentives, and educate the customer could dramatically affect acceptance of hybrids and market penetration (Fig. 10).

SEE ALSO THE FOLLOWING ARTICLES

Alternative Transportation Fuels: Contemporary Case Studies • Batteries, Transportation Applications • Biodiesel Fuels • Biomass for Renewable Energy and Fuels • Energy Efficiency, Taxonomic Overview • Ethanol Fuel • Fuel Cell Vehicles • Fuel Cycle Analysis of Conventional and Alternative Fuel Vehicles • Fuel Economy Initiatives: International Comparisons • Internal Combustion (Gasoline and Diesel) Engines • Transportation Fuel Alternatives for Highway Vehicles

Further Reading

An, F., Stodolsky, F., and Santini, D. (1999). "Hybrid Options for Light-Duty Vehicles, SAE Paper No 1999-01-2929." Society of Automotive Engineers, Warrendale, Pennsylvania.

Aoki, K., Kenji, N., Kajiwara, S., Sato, H.,Yamamoto, Y. (2000). "Development of Integrated Motor Assist Hybrid System," SAE Paper 2000-01-2059. Society of Automotive Engineers, Warrendale, Pennsylvania.

DeCicco, J. M. (2000). Hybrid Vehicles in Perspective: Opportunities, Obstacles, and Outlook. American Council for an Energy Efficient Economy, Washington, D.C.

Electric Power Research Institute. (2002). "Comparing the Benefits and Impacts of Hybrid Electric Vehicle Options for Compact Sedan and Sport Utility Vehicles," 1006892. EPRI, Palo Alto, California.

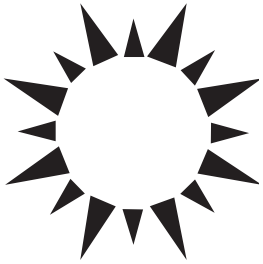
German, J. M. (2001). Hybrid vehicles go to market. *TR News* **213**(March–April), 15–23.

Itazaki, H. (1999). "The Prius that Shook the World." Nikkan Kogyo Shimbun, Ltd., Tokyo, Japan.

Wakefield, E. H. (1998). "History of the Electric Automobile Hybrid Electric Vehicles." Society of Automotive Engineers, Warrendale, Pennsylvania.

Hybrid Energy Systems

J. F. MANWELL
University of Massachusetts
Amherst, Massachusetts, United States



1. **Introduction to Hybrid Energy Systems**
2. **Characteristics of Hybrid Energy Systems**
3. **Technology Used in Hybrid Energy Systems**
4. **Energy Loads**
5. **Renewable Energy Resource Characteristics**
6. **Design Considerations**
7. **Economics**
8. **Trends in Hybrid Energy Systems**

Glossary

biomass Any organic material that is used as a fuel.
deferrable load An electrical load that must be supplied a certain amount of energy over a specified time period in an isolated electrical network, but that has some flexibility as to exactly when.
diesel generator A device for producing electricity through the use of a diesel engine.
dump load A device used to dissipate power in order to maintain stability in an isolated hybrid energy system.
fuel cell A device for continuously producing electricity through a reaction between hydrogen and oxygen.
hybrid energy system A combination of two or more energy conversion devices (e.g., electricity generators or storage devices), or two or more fuels for the same device, that, when integrated, overcome limitations that may be inherent in either.
load (1) An energy consuming device; (2) the cumulative power requirement due to a number of devices connected to an electrical network.
maximum power point tracker A DC-DC converter whose function is to maximize the power produced by a variable voltage DC generator, such as PV panels or a wind turbine, connected to a fixed voltage bus.
optional load An electrical load which can be supplied by excess power in an isolated electrical network but need not be supplied if no excess power is available.
penetration The instantaneous power from the renewable generator divided by the total electrical load being served in an isolated electrical network.
photovoltaic panel (PV) A semiconductor-based device that converts sunlight to electricity.
renewable energy source An energy source that is derived ultimately from the sun, moon, or earth's internal heat.
supervisory control A central control system whose function is to ensure proper operation of all the devices within a hybrid energy system.
weak grid An electrical network whose operational characteristics can be affected by the presence of a generator or load connected to it.
wind turbine A device that converts the power in the wind to electrical or mechanical power.

The term hybrid energy system refers to those applications in which multiple energy conversion devices are used together to supply an energy requirement. These systems are often used in isolated applications and normally include at least one renewable energy source in the configuration. Hybrid energy systems are used an alternative to more conventional systems, which typically are based on a single fossil fuel source. Hybrid energy systems may also be used as part of distributed generation application in conventional electricity grid. The most general definition is the following: "Hybrid energy systems are combinations of two or more energy conversion devices (e.g., electricity generators or storage devices), or two or more fuels for the same device, that when integrated, overcome limitations that may be inherent in either." This definition is useful because it includes a wide range of possibilities and the essential feature of the multiplicity of energy conversion. This article focuses on stationary power systems, where at least one of the energy conversion devices is powered by a renewable energy source (which, in the context of this article, is one based ultimately on the sun).

1. INTRODUCTION TO HYBRID ENERGY SYSTEMS

A considerable interest has emerged in combined or 'hybrid' energy systems. In the context used here,

 © 2004 Elsevier Inc. All rights reserved.

that refers to an application in which multiple energy conversion devices are used together to supply an energy requirement. These systems are often used in isolated applications and normally include at least one renewable energy source in the configuration. Hybrid systems are used as an alternative to more conventional systems, which typically are based on a single fossil fuel source.

1.1 Definitions

Hybrid energy systems have been defined in a number of ways. The most general, and probably most useful, is the following:

"Hybrid energy systems are combinations of two or more energy conversion devices (e.g., electricity generators or storage devices), or two or more fuels for the same device, that when integrated, overcome limitations that may be inherent in either."

This definition is useful because it includes a wide range of possibilities and the essential feature of the multiplicity of energy conversion. Note that this broad definition does not necessarily include a renewable energy based device and allows for transportation energy systems. In the present discussion, however, we focus on stationary power systems, where at least one of the energy conversion devices is powered by a renewable energy source (which is one based ultimately on the sun, moon, or earth's internal heat).

For the purpose of comparison, it is useful to consider briefly the nature of conventional energy systems that are normally used where hybrid system might be used instead. There are basically three types of conventional systems of interest: (1) large utility networks, (2) isolated networks, and (3) small electrical load with dedicated generator.

Large utility networks consist of power plants, transmission lines, distribution lines and electrical consumers (loads). These networks are based on alternating current (AC) with constant frequency. Such networks are frequently assumed to have an infinite bus. This means that the voltage and frequency are unaffected by the presence of additional generators or loads.

Isolated electrical networks are found on many islands or other remote locations. They are similar in many ways to large networks, but they are normally supplied by one or more diesel generators. Generally, they do not have a transmission system distinct from the distribution system. Isolated networks do not behave as an infinite bus and may be affected by additional generators or loads.

For many small applications, it is common to supply an electrical load with a dedicated generator. This is the case, for example, at construction sites, highway signs, and vacation cabins. These systems are also normally AC but have no distribution system.

1.2 Applications for Hybrid Energy Systems

There are numerous possible applications for hybrid power systems. The most common examples are (1) remote AC network, (2) distributed generation applications in a conventional utility network, and (3) isolated or special purpose electrical loads.

The classic example of the hybrid energy system is the remote, diesel-powered AC network. The basic goal is to decrease the amount of fuel consumed by diesel generators and to decrease the number of hours that they operate. The first addition to "hybridize" the system is to add another type of generator, normally using a renewable source. These renewable generators are most commonly wind turbines or photovoltaic panels. Experience has shown, however, that simply adding another generator is not sufficient to produce the desired results. Accordingly, most hybrid systems also include one or more of the following: supervisory control system, short-term energy storage, and load management. Each of these will be described in more detail. An example of a typical hybrid energy system, in this case a wind/diesel system, is illustrated in Fig. 1.

During the 1990s, management of many large electrical networks changed so that it is now possible

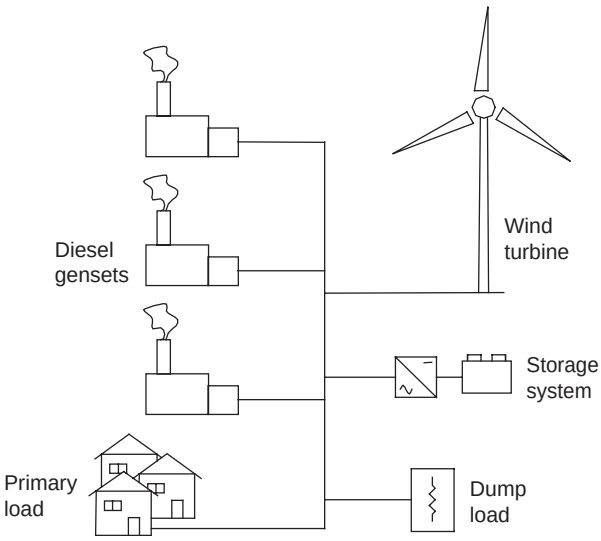


FIGURE 1 Schematic of wind diesel system with storage.

for individuals or businesses to add generation in the distribution system of the utility. This is known as distributed generation (DG). Distributed generation can have a variety of purposes, but in any case, it is sometimes desirable to combine a number of energy conversion devices together with the distributed generator. This results in a hybrid energy system in the distributed generation application.

Hybrid systems can be of particular value in conjunction with an isolated or special-purpose application. One example is that of using photovoltaic panels, together with some battery storage and power electronic converters to supply a small amount of energy to a load in a remote location. Other examples include water pumping or water desalination. In these applications, electrical loads may take a range of forms. They may be conventional AC, DC, or even variable voltage and variable frequency.

1.3 Impetus for Hybrid Energy Systems

There are a variety of reasons why a hybrid system might be used. An important reason is the reduction of fossil fuel use. Fossil fuel can be costly, especially in remote locations where the cost of transporting the fuel to the site must be considered. In many remote locations fuel oil must be stored, often for an entire winter. Reducing the amount of fuel use can reduce storage costs.

An isolated hybrid power system can be an alternative to power line construction or power line upgrade. Hybrid power systems, because they include so many components, tend to be relatively expensive. When they can be used as an alternative to power line construction, the savings can help compensate for the cost of the hybrid system.

Under some conditions a power line already exists, but is not capable of carrying the desired current. This could occur, for example, when a new load is added. In this case, the hybrid system would be a type of distributed generation. The hybrid system would only need to provide the difference between the total load and what the existing lines could carry. The presence of the existing line could simplify the design of the hybrid system, in that it need not necessarily be required to set the power line frequency and might not have either storage or dump loads (q.v.).

Distributed generation has advantages in a number of situations, and, in some cases, the overall benefit is enhanced by using a hybrid system. One common application for distributed generation is combined heat and power (CHP). In this case, waste heat associated with fossil fuel combustion is used for space or process heating, thereby increasing the overall energetic efficiency of the fuel use.

Use of hybrid systems can result in local environmental benefits. In particular, diesel generators (in isolated applications) should run less often and for shorter periods. This will reduce locally produced air pollution and noise, as well as reducing risks associated with fuel transport.

2. CHARACTERISTICS OF HYBRID ENERGY SYSTEMS

The characteristics and components of a hybrid system depend greatly on the application. The most important consideration is whether the system is isolated or connected to a central utility grid.

2.1 Central Grid Connected Hybrid Systems

If the hybrid system is connected to a central utility grid, as in a DG application, then the design is simplified to a certain degree and the number of components may be reduced. This is because the voltage and frequency are set by the utility system and need not be controlled by the hybrid system. In addition, the grid normally provides the reactive power. When more energy is required than supplied by the hybrid system the deficit can be in general be provided by the utility. Similarly, any excess produced by the hybrid system can be absorbed by the utility. In some cases, the grid does not act as an infinite bus, however. It is then said to be "weak." Additional components and control may need to be added. The grid connected hybrid system will then come to more closely resemble an isolated one.

2.2 Isolated Grid Hybrid Systems

Isolated grid hybrid systems differ in many ways from most of those connected to a central grid. First, they must be able to provide for all the energy that is required at any time on the grid or find a graceful way to shed load when they cannot. They must be able to set the grid frequency and control the voltage. The latter requirement implies that they must be able to provide reactive power as needed. Under certain conditions, renewable generators may produce energy in excess of what is needed. This energy must be

dissipated in some way so as not to introduce instabilities into the system.

There are basically two types of isolated grid hybrid systems which include a renewable energy generator among their components. These are known as low penetration or high penetration. In this context, "penetration" is defined as the instantaneous power from the renewable generator divided by the total electrical load being served. Low penetration, which is on the order of 20% or less, signifies that the impact of the renewable generator on the grid is minor, and little or no special equipment or control is required. High penetration, which is typically over 50% and may exceed 100%, signifies that the impact of the renewable generator on the grid is significant and special equipment or control is almost certainly required. High-penetration systems may incorporate supervisory control, so-called dump loads, short-term storage, and load management systems.

Two important considerations in an isolated system are whether the system can at times run totally on the renewable source (without any diesel generator on) and whether the renewable source can run in parallel with (i.e., at the same time as) the diesel generator. It is most common for one or the other to be possible (and normal). It is less common that both modes of operation are possible. This latter system offers the greatest fuel savings but is more complicated.

2.3 Isolated or Special Purpose Hybrid Systems

Some hybrid systems are used for a dedicated purpose, without use of real distribution network. These special purposes could include water pumping, aerating, heating, desalination, or running grinders or other machinery. Design of these systems is usually such that system frequency and voltage control are not major issues, nor is excess power production. In those cases where energy may be required even when renewable source be temporarily unavailable, a more conventional generator may be provided. Renewable generators in small isolated systems typically do not run in parallel with a fossil fuel generator.

3. TECHNOLOGY USED IN HYBRID ENERGY SYSTEMS

A wide range of technology may be used in a hybrid energy system. This section describes some of them in more detail than was done in previous sections. Devices to be discussed include energy consuming devices (loads), rotating electrical machinery, renewable energy converters, fossil fuel generators (often, but not always, diesels), energy storage devices, power converters, control systems, and load management devices. Some of the various possible devices and arrangements that may be found in hybrid energy system are illustrated in Fig. 2. The primary focus here is on isolated network hybrid systems, but much of the technology applies to other types of hybrid systems as well.

3.1 Energy Consuming Devices

Hybrid energy systems typically use the same types of energy consuming devices that are found in conventional systems. These include lights, heaters, motors, and electronic devices. The combined energy requirement of all the devices is know as the total load, or just load. The load will typically vary significantly over the day and over the year. An example of a load varying over a year on an isolated island is shown in Fig. 3.

3.2 Rotating Electrical Machinery

Rotating electrical machinery is found in many places in a hybrid energy system. Most such machines can function as either motors or generators, depending on the application. This section focuses on the generating function.

3.2.1 Induction Generators

The induction generator has been the most common type of generator used in wind turbines. They are also occasionally used with other prime movers, such as hydro turbines or landfill gas fueled internal combustion engines. More information on induction machines is provided elsewhere in this encyclopedia, so they will not be discussed in detail here. As far as hybrid systems, however, there are two important considerations. First of all, induction machines require a significant amount of reactive power. This is not a fatal problem, but it does affect the design of the system in a number of ways.

The second consideration is starting. When an induction machine is brought on line from stand still, it requires much higher current than when operating normally. Provision must be made to ensure that the hybrid system has the capability of starting any induction motors or generators that it connected to it.

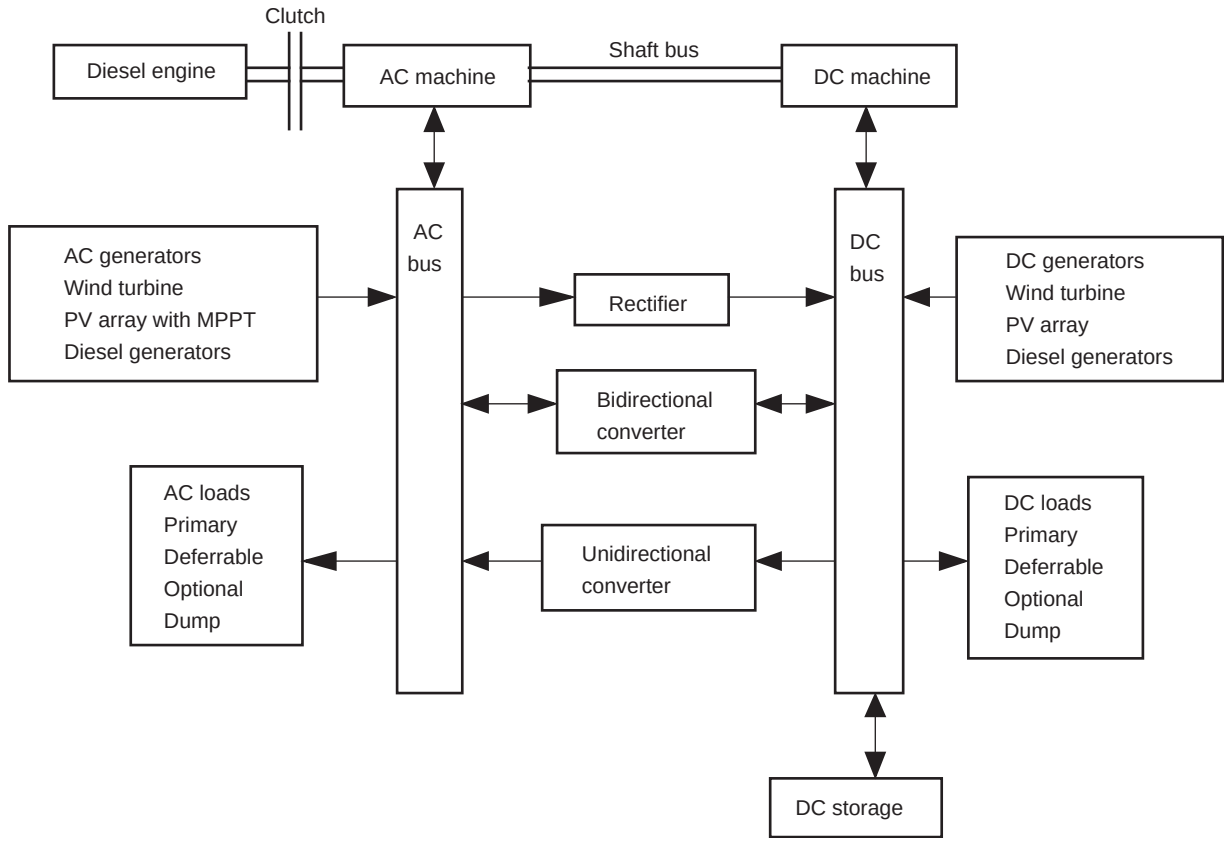


FIGURE 2 Devices and arrangements in hybrid energy systems.

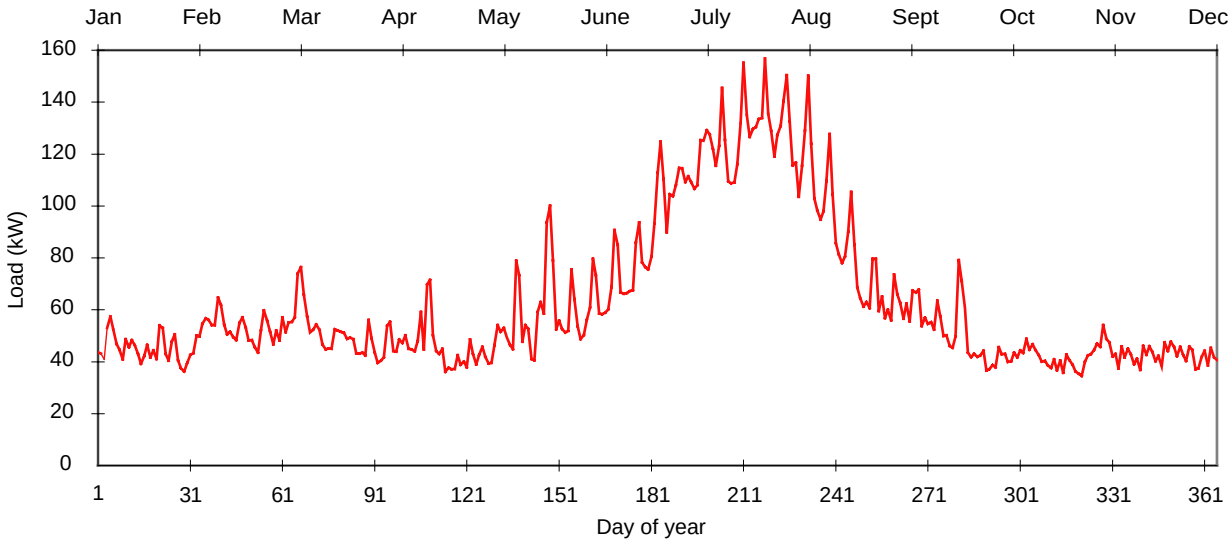


FIGURE 3 Daily electrical loads on Cuttyhunk Island, Massachusetts.

3.2.2 Synchronous Generators

Synchronous generators (SG) may also be used with a variety of prime movers. These generators are discussed in detail elsewhere in this encyclopedia. In hybrid system applications SGs are of one of two types: those with electromagnetic fields and those using permanent magnets. The first type, operating with a voltage regulator, can maintain the voltage on

the electrical network and provide reactive power required by other devices in the system. Permanent magnet SGs are often used in small wind turbines. Such generators cannot maintain voltage and so are normally used in conjunction with power electronic converters.

A synchronous machine can be brought on line so that it will run in parallel with other generators. When this is done, particular attention must be taken so that the various generators are in phase with each other.

3.3 Renewable Energy Generators

Renewable energy generators are devices that convert energy from its original form in the renewable energy source into electricity. Renewable energy generators that are most likely to be found in hybrid energy systems include wind turbines and photovoltaic panels. Some hybrid energy systems use hydroelectric generators, biomass fueled generators, or fuel cells. It should be noted that many renewable energy generators include rotating electrical machines acting in the generating mode, which is also called a generator. It should be clear from the context what is meant.

3.3.1 Wind Turbines

Wind turbines are devices that convert the energy in wind into electricity. A typical wind turbine is shown in Fig. 4. The main parts of a wind turbine are the rotor, the drive train (including the generator), main frame, tower, foundation, and control system. The rotor consists of the blades and a hub. The blades serve to convert the force of the wind to a torque that ultimately drives the generator.

The two most important features of a wind turbine as far as a hybrid energy system is concerned are the type of generator and the nature of the rotor control. Most wind turbines use induction generators, although some use synchronous generators. In either case the generator may be connected directly to the electrical network or it may be connected indirectly through a power electronic converter.

There are two main forms of rotor control on wind turbines: stall control and pitch control. The most important function of rotor control is to protect the wind turbine from high winds. In hybrid systems, rotor control is also important because it affects on how energy flows are regulated within the hybrid system. Under some conditions (with pitch control), the rotor can be controlled to facilitate start up or to reduce production when full output is not required.

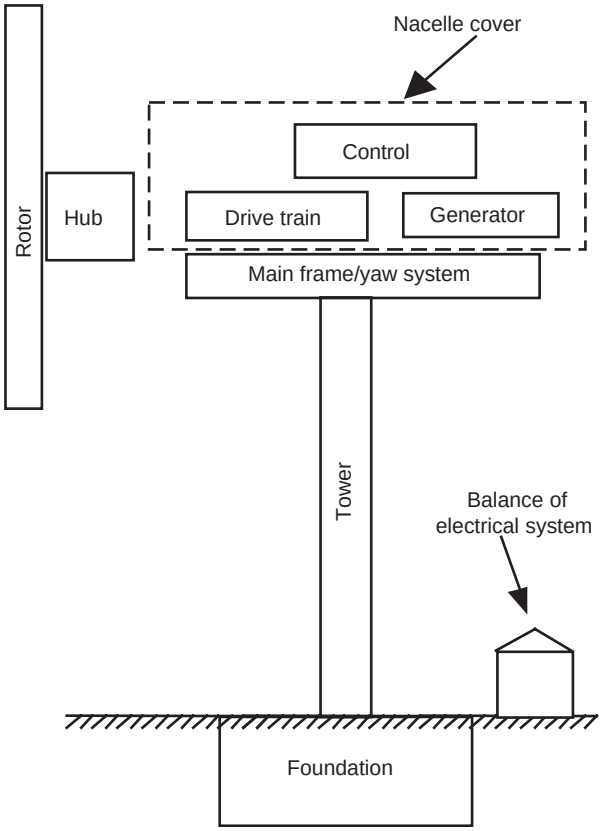


FIGURE 4 Components of typical wind turbine.

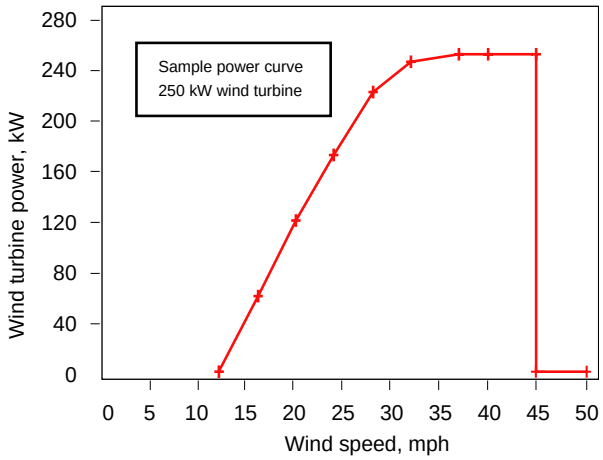


FIGURE 5 Typical wind turbine power curve.

Details on rotor control may be found elsewhere in this encyclopedia or in Manwell *et al.*

The power generated by a wind turbine varies in relation to the wind speed. That relation is summarized in a power curve, a typical example of which is illustrated in Fig. 5.

3.3.2 Photovoltaic Panels

Photovoltaic (PV) panels are used to produce electricity directly from sunlight. PV panels consist of a number of individual cells connected together to produce electricity of a desired voltage. Photovoltaic panels are inherently DC devices. To produce AC, they must be used together with an inverter.

Most PV cells are made from crystalline silicon. PV cells produce current in proportion to the solar radiation level (up to a certain voltage). The current/voltage relation of a typical silicon cell at a fixed level of solar radiation is shown in Fig. 6. Since power is proportional to the product of current and voltage, the power from a PV cell will continue to increase until the current begins to drop.

Because the maximum voltage from individual cells is less than 1 V, multiple cells are connected together in series on a PV panel.

The actual radiation level at any given time at a particular spot on the earth's surface will vary significantly over the year and over the day. See Section 5 for more details. More information on PV panels themselves is provided elsewhere in this encyclopedia.

3.3.3 Hydro Turbines

Hydropower is one of the oldest forms of electricity production and is used in large electricity networks as well as small ones. Small hydroelectric plants are used in isolated systems in many parts of the world. Hybrid systems, which include hydroelectric plants with other forms of generation, are relatively rare, but they do occur. The most common of these are in locations where the resource varies significantly over the year, and there is not enough water in certain seasons. A diesel generator is then used instead of the hydroelectric generator during those seasons. The primary requirements of a hydroelectric plant are a continuous source of water and a change in elevation. More information on hydroelectric generators is provided elsewhere in this encyclopedia.

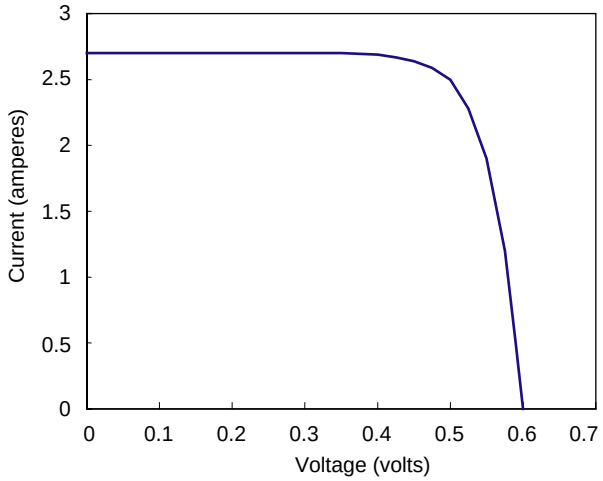


FIGURE 6 Photovoltaic cell current versus voltage.

3.3.4 Biomass Fueled Generators

Biomass fueled generators are occasionally used in hybrid energy, though most often when they are employed in isolated networks they are the only power plants used. Biomass fueled generators are quite similar to conventional coal or oil fired generating plants, except for the combustors themselves and the fuel handling equipment.

Biomass is any organic material which is used as a fuel. For purposes of power generation, the most common sources of biomass include wastes from wood products or the sugar cane industry. Other agricultural wastes are sometimes used as well.

Biomass generators are sometimes fueled with landfill gas. Such generators most commonly employ an internal combustion engine as a prime mover. Occasionally fuel cells have been used.

3.3.5 Fuel Cells

Fuel cells can be used in hybrid power systems, and are expected to become more common as their costs drop. Fuel cells ultimately run on hydrogen, but in many cases the input fuel is some other gas, such as natural gas or landfill gas.

Fuel cells can be thought of as a continuous battery. The fundamental reaction is between hydrogen and oxygen, the product of which is water and electric current. The reaction takes place in the vicinity of a membrane, which serves to separate the various components of the electrolyte and the electrodes. The electrodes are the terminals of the fuel cell and carry the current into an external circuit. Like batteries, fuel cells are inherently DC devices. Fuel cells can be used in an AC network if their output is converted to AC via an inverter.

When natural gas or landfill gas is used as the fuel, it must first be used to produce hydrogen. This is done in a reformer.

There are a number of different designs for fuel cells. They include proton exchange membrane (PEM), solid oxide, and molten carbonate. More information on fuel cells is presented elsewhere in this encyclopedia.

3.4 Fossil Fuel Generators

Fossil fuel generators are commonly used in hybrid energy systems. In fact, most isolated power systems at the present time are based on fossil fuel, using internal combustion engines as prime movers. Most medium-sized and larger isolated systems use diesel engine/generators. The smallest systems sometimes use gasoline. Some very large isolated power systems sometimes use conventional oil-fired steam power plants. They will not be discussed here, however.

3.4.1 Diesel Engine/Generators

Diesel engine generators are discussed elsewhere in this encyclopedia, so they will only be summarized here. Emphasis is given to those aspects of relevance to hybrid energy systems.

Diesel generators typically consist of three main functional units: a diesel engine, a synchronous generator with voltage regulator, and a governor.

The fuel injection system is an important part of the diesel engine. Its function is to inject the proper amount of fuel into the proper cylinder at the appropriate time in the cycle. The timing of the injection is determined by the design of the engine itself; the amount of fuel is determined by the governor.

The diesel engine is normally connected directly to a synchronous generator. A voltage regulator ensures the proper voltage is produced. The frequency of the AC power is directly proportional to the engine speed, which in turn is controlled by the governor.

Historically diesel generators have employed "droop" type governors. A set of flyballs is driven by the engine, so the speed varies in proportion to the engine speed. Through a set of linkages, the amount of fuel that can be injected at any time is varied according to how far the operating speed differs from (or "droops") from nominal. As the electrical load on the generator increases, the droop increases and more fuel is injected. Diesel generators are now more likely to have electronic governors without droop, but the function is the same.

An important consideration regarding diesel generators is their fuel consumption, both at full load and part load. Diesel fuel consumption is frequently described in terms of electricity produced per gallon of fuel consumed. Full load values ranging from a low of 8 kWh/gallon to 14 kWh/gallon have been reported.

Diesel engine generators are often called on to follow the load. That means that their output must be equal to the system load (or to the system load less the production of any other generators that might be on; this is called the net load). As the load may go up and down, so must the electricity generated. This is known as part load operation. Generally, the conversion efficiency is less at part load than at full load. Fuel consumption over the full range of operation is summarized in fuel curves. In these curves, fuel consumption is graphed against power. Figure 7 shows a typical example.

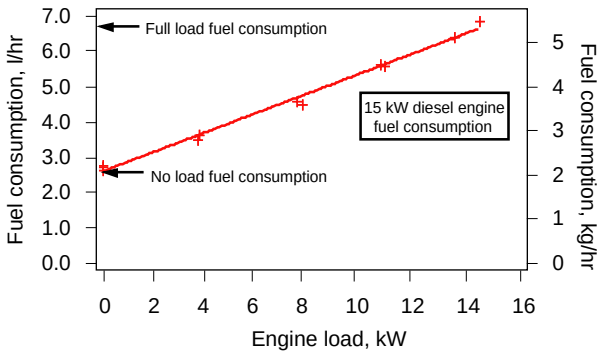


FIGURE 7 Typical diesel engine generator fuel curve.

Regardless of efficiency considerations, manufacturers normally recommend that diesel generators not be run below some specified minimum power level, known as the minimum load. Typically, the minimum recommended load is between 25% and 50% of rated. Engines run for long periods at levels below the minimum recommended can experience a number of problems.

3.4.2 Gasoline Generators

Gasoline generators are sometimes used for very small hybrid energy systems applications. These generators have the advantage that they are readily available throughout the world and are relatively expensive. They are similar in many ways to diesel generators, except that use spark ignition, and have lower compression ratios. They also typically use carburetors rather than fuel injection. The main disadvantage of gasoline generators is that they are less efficient than diesel.

3.5 Energy Storage

Energy storage is often useful in hybrid energy systems. Energy storage can have two main functions. First of all, it can be used to adapt to a mismatch between the electrical load and the renewable energy resource. Second, it can be used to facilitate the control and operation of the overall system. There are basically two types of energy

storage, convertible and end use. Convertible storage is that which can readily be converted back to electricity. End-use storage can be applied to a particular end-use requirement but may not readily be converted back to electricity.

3.5.1 Convertible Storage

There are a number of convertible storage media, although only a few of them have been used frequently in hybrid energy systems. The most commonly used form of convertible storage is the battery. Less commonly used, but frequently discussed, forms include pumped hydroelectric, flywheels, compressed air, and hydrogen.

3.5.1.1 Batteries Batteries are the most commonly used form of convertible storage for hybrid energy systems. They have been used both for short term (less than 1 hour) and long term (more than 1 day) storage. A number of types of batteries have been developed. The most common type of storage battery for hybrid applications is the lead acid battery. Nickel cadmium has also been used occasionally. Batteries are discussed elsewhere in this encyclopedia, so only those aspects most relevant to hybrid energy systems are summarized here.

As far as hybrid energy systems are concerned, there are five important performance characteristics of batteries: (1) voltage, (2) energy storage capacity, (3) charge/discharge rates, (4) efficiency, and (5) battery lifetime.

Batteries by their nature are DC. Individual batteries are made up by a number of cells in series, with each cell nominally two volts. Complete batteries are typically 2, 6, 12, or 24 Volts. The actual terminal voltage will depend on three factors: (1) state of charge, (2) whether the battery is being charged or discharged, and (3) the rate of charge or discharge.

The energy storage capacity is primarily a function of battery voltage and the amount of charge it can hold and then return. Charge is measured in units of current times time (Ampere-hours). The amount of charge that is stored in a battery at any particular time is often described with reference to its full state by the term "state of charge" (SOC). Discharging and charging back to a given level (normally fully charged) is referred to a cycle. The total amount of charge that a battery can hold is primarily a function of the amount of material used in the construction.

Battery capacity is normally specified with reference to a specific discharge rate. This is because the apparent capacity of batteries actually differs with charge and discharge rate. Higher rates result in smaller apparent capacities.

As energy storage media, batteries are not 100% efficient. That is, more energy is expended in charging than can be recovered. Overall efficiencies are typically in range of 50 to 80%.

An important characteristic of batteries is their useful lifetime. Experience has shown that the process of using batteries actually decreases their storage capacity until eventually the battery is no longer useful. The important factors in battery life are the number of cycles and the depth of discharge in the cycles. Depending on the type of battery, the number of deep cycles to which a battery can be subjected ranges from a few thousand down to hundreds or even tens of cycles. The cycle life of a typical battery is illustrated in Fig. 8.

3.5.1.2 Pumped Hydro One form of convertible storage that has been applied in some hybrid energy systems is pumped storage. In this case, water is pumped from one reservoir at a low elevation up to one at a higher elevation. The amount of energy that can be stored is a function of the size of the reservoir and the difference in elevation. The overall efficiency of the storage is a function of the efficiency of the pumps and turbines (which may be the same devices) and the hydraulic losses in the pipes connecting the two reservoirs. The use of pumped storage is limited by the lack of sites where such facilities can be installed at a reasonable price.

3.5.1.3 Flywheels Flywheels can be used to store energy in a hybrid system. A flywheel energy storage system consists of the following components: (1) the flywheel itself, (2) an enclosure, usually evacuated to minimize frictional losses, (3) a variable speed motor/generate to accelerate and decelerate the

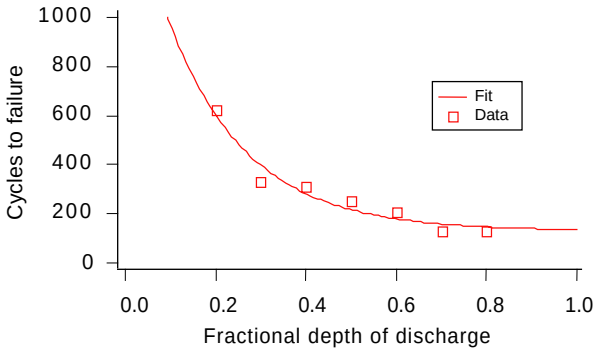


FIGURE 8 Battery cycle life.

wheel, and (4) a power electronic converter. Energy is absorbed when the wheel is accelerated up to its maximum speed and is released when the wheel is decelerated to some lower speed. A power electronic converter accompanies the flywheel and motor/generator because the input and output power to and from the motor/generator is typically variable voltage and variable frequency AC. Flywheels typically store relatively small amounts of energy, but they can absorb or release the energy at high rates. Thus in hybrid power systems they can be used to smooth short term fluctuations in power (on the order of seconds or minutes) and they can facilitate system control.

3.5.1.4 Compressed Air Compressed air can be used for storage in hybrid systems, and some experimental prototypes have been built. Efficiency is relatively low, however, and this method of storage has not been widely used.

3.5.1.5 Hydrogen Hydrogen can be produced by the electrolysis of water, using a renewable energy source for the electricity. The hydrogen can be stored indefinitely and then used in an internal combustion engine/generator or a fuel cell to generate electricity again. This method of storage appears to have a lot of potential, but it is still quite expensive.

3.5.2 End-Use Storage

End-use storage, as opposed to convertible storage, refers to the situation in which some product is created through the use of electricity when it is available. The product is then stored and made available when it is needed. Load management schemes often incorporate some end-use storage. Three examples are given next.

3.5.2.1 Thermal Energy A common form of end-use storage is thermal energy, most often in the form of hot water. Hot water can be used for space heating applications, domestic hot water, swimming pools, and so on. Hot water can be stored relatively inexpensively in insulated water tanks. Depending on the location of the water tanks relative to the end use, the efficiency of the storage can be quite high.

3.5.2.2 Pumped Water Pumped water is another form of end-use storage. This form of storage has a long history of use in conjunction with windmills. In this application, water for any plausible purpose is pumped into a reservoir or storage tank, usually elevated, from which it can be released as needed. When the storage is elevated, then the water can flow by gravity to the point of use, so no further input of externally produced energy is required.

3.5.2.3 Pure Water Another, less common form of end-use storage is the production and then storage of pure water from salty or brackish water. Depending on the quality of water at the input of the process, ultrafiltration, reverse osmosis, or vapor compression evaporation may be used to produce the pure water. All of these are energy intensive, so the implicit energy density of the storage is high.

3.6 Power Converters

For any hybrid energy system to function properly, it is common that one or more power converter be incorporated into the system. These are either electromechanical or electronic devices.

3.6.1 Electromechanical Power Converters

There are at least two types of electromechanical power converters that have been used in hybrid energy systems. They include the rotary converter and the synchronous condenser.

3.6.1.1 Rotary Converter A rotary converter is an electromechanical device that converts AC to DC or vice versa. When it is converting AC to DC it is a rectifier. When operating the other way it is an inverter. The rotary converter consists of two rotating electrical machines that are directly connected together. One of them is a DC machine; the other is an AC machine. Either of them can run as a motor or generator depending on the intended direction of power flow. The AC machine can be either an induction machine or a synchronous machine. Which is used will depend on the requirements of the system.

Rotary converters have the advantage that they are a rugged and well-understood technology that has been around for many years. Their disadvantage is that their costs are high and their efficiencies are lower than are electronic devices that can serve the same purpose.

3.6.1.2 Synchronous Condenser Synchronous condenser is the name given to a synchronous machine that is connected into an electrical network to help in maintaining the system voltage. The synchronous machine in this case is essentially a motor to which no load is connected. A voltage regulator is connected to the synchronous machine and functions in the same way as described

previously. Synchronous condensers are used in hybrid energy systems when there are, for at least some of the time, no other synchronous machines connected. This is often the case in those systems that include diesel generators but that are intended to allow all the diesels generators to be turned off under some circumstances. For example, a synchronous condenser could be used in a wind/diesel system in which the wind turbines used induction generators and could at times supply all of the electrical load. The diesel could then be turned off, but the synchronous machine would be needed to supply the required reactive power to the induction generators and maintain system voltage in the process.

3.6.2 Electronic Power Converters

A wide range of electronic devices has been developed and adapted for use in hybrid energy systems. Many of them serve similar functions to the electromechanical devices described previously, but have a number of advantages, such as lower cost, higher efficiencies, and greater controllability. The devices of greatest interest include rectifiers, inverters, dump loads, and maximum power point trackers.

3.6.2.1 Rectifiers A rectifier is a device that converts AC to DC. The primary functional elements are diodes, which only let current pass one way. By suitable layout of the diodes AC in a single or three-phase circuit is converted to a rippling, but single direction, current. Capacitors may be used to smooth the resulting current.

3.6.2.2 Inverter An inverter is a device that converts DC to AC. The primary switching elements are either silicon controlled rectifiers (SCRs) or power transistors (IGBTs). They are arranged in a bridge circuit and switched on (and off, in the case of transistors) in such a way that an oscillating waveform results. Some inverters operate in conjunction with other devices that set the system frequency. These are referred to as line commutated. Other inverters have the capability to set frequency themselves. These are called self-commutated.

3.6.3 Maximum Power Point Trackers

Another electronic device that may be used in hybrid energy systems, particularly ones with photovoltaic panels is the maximum power tracker. This device is DC-DC converter than can be partially thought of as a DC transformer. Its function to provide a particular desired output voltage to the rest of the system, while adjusting the voltage at the input to allow the maximum power production by the generator that is connected to it. In the case of photovoltaic panels, the voltage at the input will be the maximum power point voltage corresponding to the incident solar radiation level. Maximum power point trackers have also been used with small wind turbines.

3.7 Dump Loads

A dump load is device that is used to dissipate power in order to maintain stability in an isolated hybrid energy system. Dump loads are used primarily to maintain power balances. They may also be used to control frequency. The most common use of a dump load is in isolated wind/diesel systems where the penetration of the wind energy generation is such that instantaneous power levels sometimes exceed the system load less than minimum allowed diesel power level. The dump load control can sense the excess power and dissipate whatever is required to ensure that the total generated power is exactly equal to the actual system load plus that which is dissipated. The dump load is virtually a requirement in high penetration wind/diesel systems using stall controlled wind turbines, because adjusting the power is impossible. When pitch controlled turbines are used it may be possible to dispense with the dump load, since power can be limited by pitching the blades.

3.8 Supervisory Controller

Many hybrid systems, especially the more complex ones, have a supervisory controller to ensure proper operation of all the devices within the system. The possible functions of a supervisory controller are illustrated in Fig. 9. The controller itself consists of three main functional units: (1) sensors, (2) logical unit, and (3) control commands.

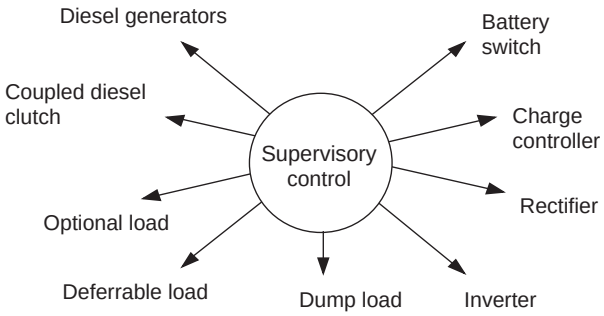


FIGURE 9 Supervisory controller functions.

Sensors are distributed throughout the system. They provide information on power levels, operating conditions, and so on. The information gathered from the sensors is directed to the logical unit. The logical unit is based on a computer or microprocessor. It will make decisions based on an internal algorithm and the data from the sensors. Algorithms can be fairly simple or quite complex.

The decisions made by the logical unit are referred to as dispatch decisions, since their function relates to dispatching of the various devices in the system. Dispatch in this sense refers to turning a device on or off, or in some cases to setting its power level. The dispatch instructions from the supervisory controller are sent to controllers of the various devices in the system.

4. ENERGY LOADS

The energy supplied by a hybrid system can be categorized in a variety of ways. The first has to do whether the energy supplied is electricity or heat. Within the electrical category, electricity loads are often divided into primary and secondary loads. Primary loads are those that must be served immediately. Secondary loads are associated with load management, and they may be further divided in what are known as deferrable and optional loads. Deferrable loads have some flexibility in when they are served, while optional loads are those which are only served if there happens to be sufficient excess energy available to do so. Regardless of type, it may be noted that loads frequently vary significantly from one season to the next as well as over the week and over the day, as was discussed earlier.

5. RENEWABLE ENERGY RESOURCE CHARACTERISTICS

It is worth considering briefly the time varying nature of the various renewable resources that might be used in a hybrid system, because that nature will affect the design and operation of the system.

5.1 Wind

The wind resource is ultimately generated by the sun, but it tends to be very dependent on location. Over most of the earth, the average wind speed varies from one season to another. It is also likely to be affected by general weather patterns and the time of day. It is not uncommon for a site to experience a number of days of relatively high winds and for those days to be followed by others of lower winds. The daily and monthly average wind speed for an island off the coast of Massachusetts, illustrating these variations, is illustrated in Fig. 10. The wind also exhibits short-term variations in speed and direction. This is known as turbulence. Turbulent fluctuations take place over time periods of seconds to minutes.

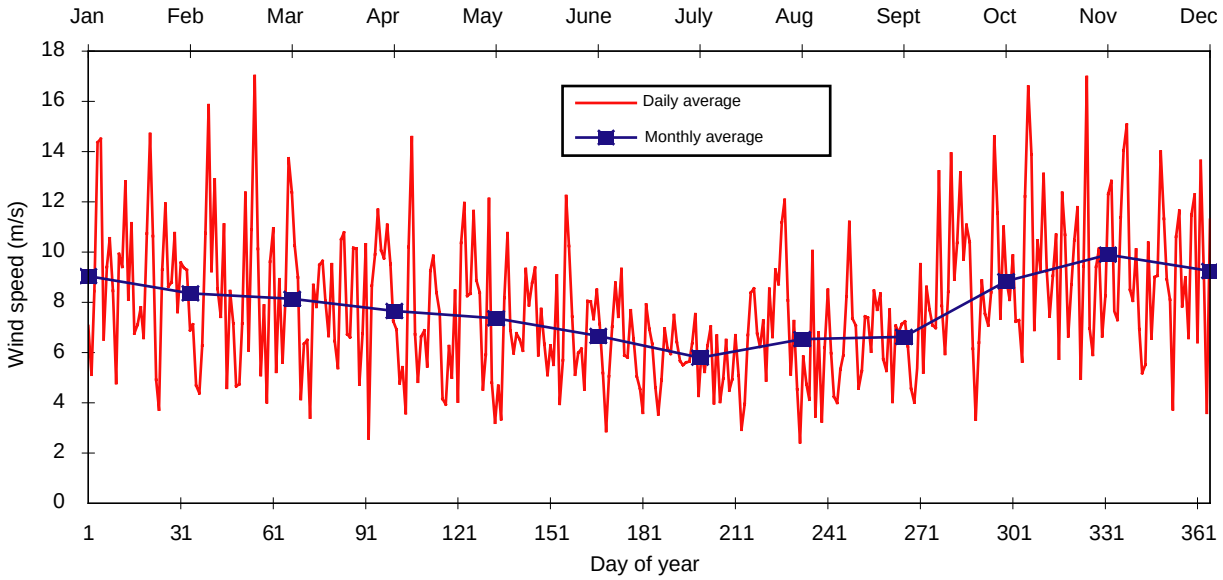


FIGURE 10 Daily and monthly wind speeds on Cuttyhunk Island, Massachusetts.

5.2 Solar Radiation

The solar radiation resource is fundamentally determined by the location on the earth's surface, the date, and the time of day. Those factors will determine the maximum level of radiation. Other factors, such as height above sea level, water vapor or pollutants in the atmosphere, and cloud cover, decrease the radiation level below the maximum possible. Solar radiation does not experience the same type of turbulence that wind does, but there can be variations over the short term. Most often, these are related to the passage of clouds. Figure 11 illustrates the solar radiation over a 5-day period in December in Boston, Massachusetts.

5.3 Hydropower

The hydropower resource at a site is a function of the amount of flowing water available (discharge) in a river or stream and the change in elevation (head). The head is usually relative constant (affected only by high water during storms), but the amount of water available can vary significantly over time. The average discharge is determined by rainfall and the drainage area upstream of the site on which the rain falls. Discharge will increase after storms and decrease during droughts. Soil conditions and nature of the terrain can also affect the discharge. Short-term fluctuations are normally insignificant.

5.4 Biomass

Sources of biomass are forest or agricultural products. The resource is ultimately a function of such factors as solar radiation, rainfall, soil conditions, temperatures, and the plant species that can be grown.

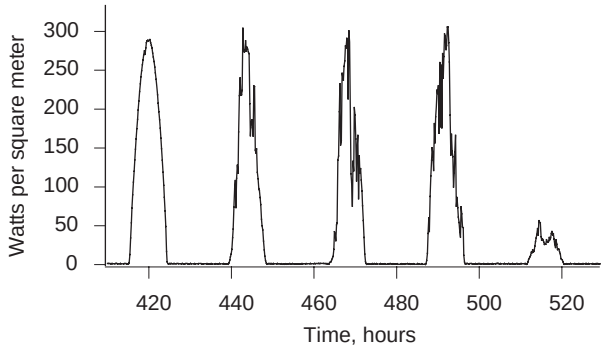


FIGURE 11 Hourly solar for 5 days radiation in Boston, Massachusetts.

6. *DESIGN CONSIDERATIONS*

The design of a hybrid energy system will depend on the type of application and the nature of the resources available. The primary consideration is whether the system will be isolated or grid connected. Other important considerations include (1) how the various generator types will be intended to operate with each other, (2) excess power dissipation, (3) use of storage, (4) frequency control, (5) voltage control, (6) possible dynamic interactions between components. Many of the possible technologies were discussed previously and will not be repeated here.

6.1 Load Matching

In the design of an isolated hybrid system with renewable energy source generators one matter of particular concern is known as load-matching. This refers to coincidence in time (or lack thereof) of the renewable resource and the load. The match between the two will affect the details of the design and will also determine how much of the energy that can be generated from the renewable energy source can actually be used. When too much is available, some of it may have to be dissipated. When too little is available, either storage or a conventional fueled generator will be needed.

Under ideal circumstances, the energy requirement will match the available resource. Often, however, that is not the case. In many islands of the world, for example, the largest loads are during the summer tourist season, but the wind resource is greatest in the winter. Figures 3 and 10 above illustrated a typical example of this for the load and wind, respectively.

When there is a mismatch between the resource and the load, the system must be designed to function properly under all conditions. For example, a wind/diesel system designed for the situation shown in Figs. 3 and 10 would most likely be designed so that wind energy could supply most of the load in the winter, keeping most or all of the diesel generators off for long periods, but in the summer, the diesels would continue to run and the wind turbines would serve as fuel savers. Figure 12 illustrates the possible energy flows for a 50-second period for a hypothetical wind/diesel system on an isolated island. The system illustrated has a primary load, an optional heating load, a dump load, a wind turbine, and multiple diesel engine/generators, at least one of which must always remain on.

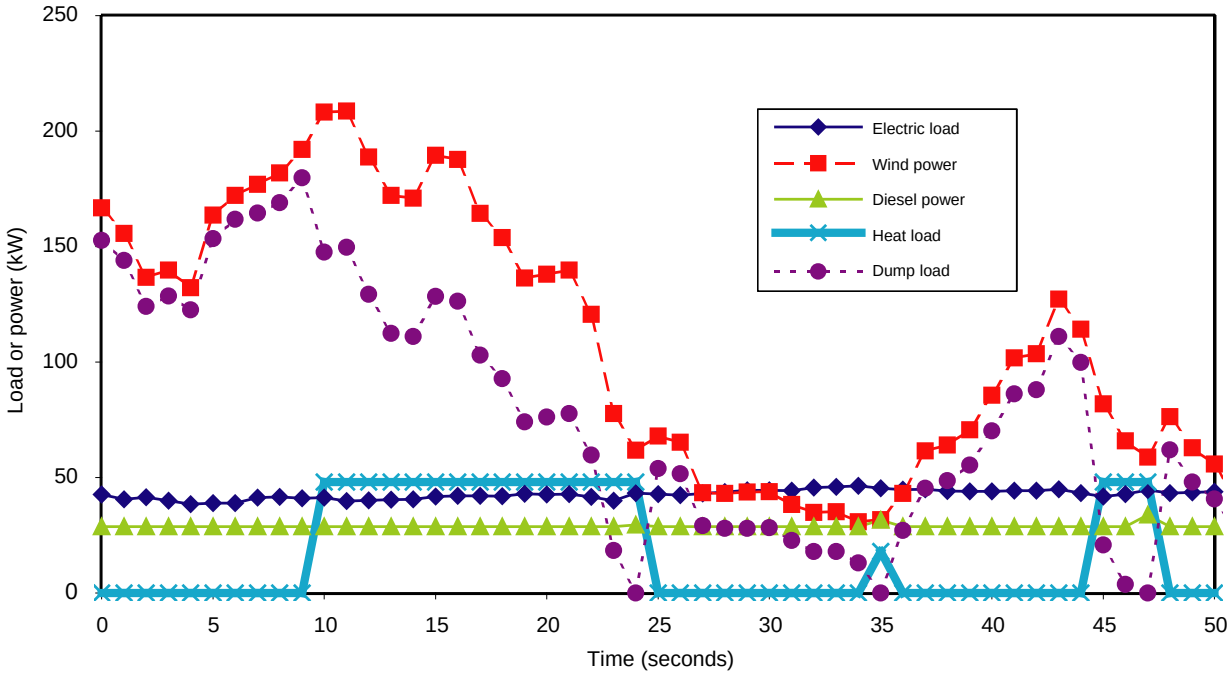


FIGURE 12 Power flows in hypothetical wind/diesel system.

7. ECONOMICS

Whether or not a hybrid energy system is actually used, and what it looks like, is strongly related to the economics of the system. In particular, will the cost of energy from the hybrid system be lower than that of a more conventional alternative? The cost of energy of a hybrid system is determined primarily by two factors: the cost of the system and the amount of useful energy that is produced. Other factors that are also important include the value of the energy, the cost of conventional energy, lifetime of the system, maintenance costs, and financial costs.

The cost of a hybrid energy system is first of all affected by the cost of the individual components that make up the system. Installation of the components and integrating them into a functioning unit will also contribute to the total cost. Generally, the renewable energy generators themselves are the most expensive items. For example, at the present time, wind turbines cost on the order of $1000 to $2000/kW, depending on the size. Photovoltaic panels cost on the order of $6000 to $8000/kW. Diesel generators are also costly, but considerably less so than wind turbines. Costs of $250 to $500/kW are typical. Lead acid batteries cost $100 to $200/kWh.

The energy that can be produced by a renewable energy generator will depend on the type of generator, its productivity at different levels of the resource, and the distribution of the occurrences of those levels. For example, a wind turbine will produce different amounts of power depending on the wind speed. If the average wind speed each hour over a year is known, this information can be combined with the data in the turbine power curve to predict the total energy that the turbine could produce. Similar curves describe the output of photovoltaic panels. These curves, combined with hourly data on the solar resource, can be used to predict the total annual energy from the panels.

When a renewable energy generator is operating in a hybrid energy system, predicting the useful energy is not as simple as it would appear, based on the previous discussion. First of all, because of a likely mismatch between the available resource and the load, it is may be that not all of the energy can be used when it is available, except in very low penetration systems. When not all of the load is supplied by the renewable generator, the rest is made up a conventional generator. Because of the difficulty in determining the amount of useful energy produced by the renewable energy generators in a hybrid system and predicting the actual fuel savings, detailed computer models are frequently used to facilitate the process.

The value of energy in a hybrid energy system is related to the nature of the energy produced, the cost of the alternatives, and the vantage point of the

operator of the system. For example, the operator of an island diesel power system, considering installing a wind turbine, would want to know that the cost of the turbine would be offset by the decrease in diesel fuel purchased.

Economics of hybrid energy systems are typically evaluated by a technique known as life cycle costing. This method takes account of the fact that hybrid energy systems have relatively high initial costs but long lifetimes and low operating costs. These factors, together with various financial parameters, are used to predict present value costs, which can then be compared with costs from the conventional alternatives.

8. TRENDS IN HYBRID ENERGY SYSTEMS

Hybrid energy systems are still an emerging technology. It is expected that technology will continue to evolve in the future, so that it will have wider applicability and lower costs. There will be more standardized designs, and it will be easier to select a system suited to particular applications. There will be increased communication between components. This will facilitate control, monitoring, and diagnosis. Finally, there will be increased use of power electronic converters. Power electronic devices are already used in many hybrid systems, and as costs go down and reliability improves, they are expected to be used more and more.

SEE ALSO THE FOLLOWING ARTICLES

Biomass for Renewable Energy and Fuels • Cogeneration • Economics of Energy Supply • Flywheels • Fuel Cells • Geothermal Power Generation • Hydrogen Production • Hydropower Technology • Photovoltaic Energy: Stand-Alone and Grid-Connected Systems • Solar Thermal Power Generation • Waste-to-Energy Technology • Wind Farms

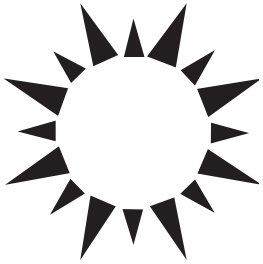
Further Reading

Hunter, R., and Elliot, G. (1994). "Wind-Diesel Systems." Cambridge University Press, Cambridge, United Kingdom.

Lundsager, P., Binder, H., Clausen, H-E., Frandsen, S. Hansen, L., and Hansen, J. (2001). "Isolated Systems with Wind Power." Riso National Laboratory Report, Riso-R-1256 (EN), Roskilde, Denmark, www.risoe.dk/rispubl/VEA/veapdf/ris-r-1256.pdf.

Manwell, J. F., Rogers, A. L., and McGowan, J. G. (2002). "Wind Energy Explained: Theory, Design and Application." John Wiley & Sons, Chichester, United Kingdom.

Hydrogen, End Uses and Economics



DAVID HART
Imperial College London
London, United Kingdom

1. **Introduction to Hydrogen Energy End Uses**
2. **Transport End Uses**
3. **Stationary Heat and Power End Uses**
4. **Other Hydrogen End Use Technologies**
5. **Economics of Hydrogen Systems**
6. **Conclusions**

Glossary

external costs of energy Costs, such as health damage from pollution, that can be directly attributed to energy use but are not included in its price.

fuel cell Device in which hydrogen can be electrochemically combusted to produce electricity and heat.

hydrogen embrittlement Reduction in material strength due to atomic hydrogen interaction with material.

hydrogen energy Hydrogen used for production of power, heat, or another energy service.

liquid hydrogen Hydrogen cooled to –253°C, where it changes from a gaseous state to a liquid state.

Hydrogen is an appropriate fuel for many energetic end uses, from the production of heat in buildings to the generation of power for transport. It can be burned in place of conventional fuels, but it can also be electrochemically oxidized rather easily in comparison with most hydrocarbons. This allows it to be used efficiently in fuel cells and broadens the available scope of end use applications. In addition, hydrogen provides a potentially valuable form of energy storage for intermittent renewable energy systems such as wind and solar power. Currently, hydrogen energy is used to any great extent only within the space industry, where its combustion with pure oxygen provides propulsion, typically for secondary rocket stages. However, chemical hydrogen is in widespread use in the oil industry, where it is used to clean, refine, and upgrade petroleum products for energy uses. Many small hydrogen energy systems are in development, including hydrogen vehicles using both internal combustion engines and fuel cells; stationary power and heat production plants, and small-scale electrical and electronic systems. None is yet commercially available outside very specific cases because the use of hydrogen is typically more costly than the use of conventional commercial fuels. The economics of hydrogen energy will improve as production, storage, and end use technology become more widespread and cheaper. If environmental benefits such as low emissions are valued, hydrogen should become even more competitive with conventional energy.

1. INTRODUCTION TO HYDROGEN ENERGY END USES

Hydrogen has been used as an industrial chemical since its discovery and naming during the late 18th century, but rarely for its energetic properties. The use of hydrogen as an energy carrier was proposed early on by Jules Verne, among others, but hydrogen is found only in compound form on earth, and so readily available conventional fuels, such as coal and oil, were preferred. Interest in hydrogen energy increased sharply toward the end of the 20th century. This increased interest was prompted by growing global concerns about (1) emissions of carbon dioxide (CO_2) and local pollutants such as NO_x and (2) potential reductions in the availability of sources of primary fossil energy. Hydrogen energy offers the possibility of solving these problems. CO_2 emissions can be reduced because no carbon atoms are present in the hydrogen, local pollutants can be reduced because hydrogen can be burned cleanly, and

 © 2004 Elsevier Inc. All rights reserved.

the wide range of possible hydrogen sources allows for flexibility in production. The extent to which these solutions are possible depends on all the elements of the hydrogen fuel chain: production, storage and transportation, and end use.

As with all fuels, the end use of hydrogen energy technologies is that of providing a service—illumination, warmth, and mobility. Various technologies may be used to achieve this, and various categorizations may be applied to the end uses.

Hydrogen end uses are frequently categorized in one of two ways: by type of energy output (e.g., heat, light, electricity) or by end use type (e.g., transport, stationary power generation). In many cases, these categorizations become blurred, for example, as the coproduction of heat and power becomes more widespread. The economics of each end use depend strongly on a wide variety of factors, including the source of the hydrogen but also the cost, lifetime, and efficiency of the end use technology. Hydrogen end use technologies are generally conventional heat and power production technologies with modifications to accommodate the various characteristics of the hydrogen fuel.

Importantly, hydrogen energy technologies are in their infancy, and little is published about the costs of many of them. The costs can be expected to change rapidly and probably over a large range.

Table I shows technologies that can be used with hydrogen in a variety of generic applications. The potential for providing transport, heat, stationary power, or power for small-scale electronics is indicated. A range of hydrogen end use technologies is shown, although other conversion devices, such as Stirling engines, may also be used.

1.1 Hydrogen End Use Technologies

As indicated in Table I, many hydrogen technologies can be applied to different end uses, either unchanged or with minor modifications. This is also common to other fuels.

The hydrogen used in transport can be used in various prime movers for most applications. For example, automotive companies are developing hydrogen vehicles based on both internal combustion engine and fuel cell technology. Aircraft will be powered primarily by gas turbines, although fuel cell aircraft are under consideration. Ships may use modified versions of diesel engines. The various prime movers offer different characteristics, as do those used with conventional fuels.

Similar technologies will be deployed in stationary power. Fuel cells are viewed as a promising technology with development work still to be done, and conventional machinery, such as turbines and gas engines, may be used effectively in stationary applications. Fuel cells are also the only hydrogen technology likely to be deployed in commercial electronics applications.

Hydrogen, especially when used in a very pure state, can embrittle certain materials and cause them to become weak. Its high flame speed may also cause problems. Therefore, in some instances, especially in the development of hydrogen turbines, new or advanced material developments have been required. Hydrogen in small-scale electronics would almost certainly be stored as a compound, probably in a metal hydride.

1.1.1 Spark Ignition, Diesel, and Wankel Engines

Hydrogen can be burned in a standard spark ignition engine with marginal modification, but only with a performance penalty. This comes about due to the large differences between the flame speed and flammability range of hydrogen and those of hydrocarbon fuels. The fuel flow and ratio must be optimized for hydrogen by using a pressure regulator (used for all gaseous fuels) rather than a liquid pump and by modifying the timing of the spark. Larger fuel-mixing equipment and fuel injectors are also necessary because hydrogen is less energy dense than a liquid fuel and a greater volume of it must be introduced for combustion. Then, there is less room for air in the cylinder, the expansion is reduced, and consequently so is the output power—by as much as 30% relative to the original gasoline engine.

If a dedicated hydrogen engine is used, its performance may be better than a comparable gasoline engine. The compression ratio of the cylinder can be increased, as is done with natural gas engines, by forcing more of the fuel mixture into the available space before combustion occurs or by using a turbocharger or supercharger to force more air into the cylinder. If the mixture is cooled, it is also possible to increase the amount introduced. Adopting one of these measures can allow the hydrogen engine to operate at a power up to 25% greater than the equivalent petroleum engine.

The wide flammability range of hydrogen (mixtures of between 4 and 75% in air) means that it is possible to operate the engine at a very lean mixture, increasing fuel efficiency and ensuring complete fuel combustion. Fuel efficiency is also improved because the mixing of gaseous hydrogen with air is better than that of a liquid fuel. The thermal efficiency of a hydrogen

engine can be on the order of 25% better than a gasoline equivalent, although the benefit in miles per gallon is not as simple to calculate because it also depends on the driving cycle being undertaken. The combustion temperature is lower with a lean mixture than with a rich one, reducing NO_x formation.

Hydrogen can also be used in modified diesel engines, given the introduction of a suitable ignition process, because hydrogen is not a compression ignition fuel. The Wankel rotary engine, operating on a slightly different principle from conventional piston engines, is particularly suitable for hydrogen fueling. The engine itself relies on combustion of the fuel/air mixture within an unusually shaped chamber known as an epitrochoid, which drives the rotary equivalent of a piston. The Wankel rotary engine has a separate intake and combustion chamber, allowing the hydrogen mixture to be taken in without backfire problems and enabling smoother running.

Backfiring has been an issue associated with the use of pure hydrogen in internal combustion engines in addition to knocking. The use of water injection reduces these tendencies, as does the injection of the hydrogen in liquid form into the chamber. Preignition of the hydrogen may occur if a mixture of hydrogen and air comes into contact with a hot surface. This can be avoided by mixing the fuel and air only within the cylinder and ensuring that this is done only just before the spark plug fires so as to leave no opportunity for preignition.

Use of hydrogen in internal combustion engines of any type in stationary applications is likely to be simpler than that in vehicles. Hydrogen storage is unlikely to be a critical parameter, and the system is more likely to run under steady-state conditions. Costs for engines are expected to be similar to those for conventional machines, but a premium of 10 to 20% may be attached to early examples.

1.1.2 Turbines

Hydrogen-burning turbines for vehicles are being investigated primarily for aircraft. Again, emissions from conventional kerosene fuels are a driver. Major development work is being undertaken to examine the potential for a passenger or cargo airliner, but work on hydrogen aircraft dates back to the 1950s when military projects were conducted.

The need to maintain range means that aircraft are nearly certain to use liquid hydrogen, and turbine designs are based around this expectation. In an aircraft engine, the liquid hydrogen would be passed through a heat exchanger to be vaporized before driving a turbine and a fan to compress incoming air to be driven through the engine. The hydrogen would then be burned in the afterburner to provide thrust.

Stationary hydrogen gas turbines are also under development, notably in Japan as part of a major government research project. Hydrogen's combustion properties offer advantages but have disadvantages as well. The fact that hydrogen burns cleanly is advantageous because no ash particles and residues that can produce sediment or corrosion on the turbine blades arise, unlike the burning of fossil fuels. In addition, the inlet temperature of the turbine can be raised much higher than would be possible with fossil fuels, producing higher thermal efficiencies under the constraints of the Carnot limit. However, very high temperatures and operation under pure hydrogen/oxygen conditions cause great stresses in materials, and so early turbine tests were conducted with nitrogen dilution. While reducing materials problems, this solution may result in production of high levels of NO_x, and that may be unacceptable.

Limited research is being conducted on hydrogen-burning turbines, although a 500-MW combined cycle turbine system burning liquid hydrogen and oxygen is under investigation in Japan. A thermal efficiency of 60% is predicted.

1.1.2.1 Economics of Hydrogen Engines The costs of dedicated and converted hydrogen combustion engines are not public because is available commercially. However, the minimal modification required, and the comparatively standard materials used, means that costs are expected to vary little from standard engine costs of between U.S. $30/kW for mass-produced vehicle engines and $500/kW for more specialized stationary machinery. Operating costs will depend not only on efficiency but also on the cost of fuel. Hydrogen supplied in large quantities can be competitive with gasoline, but until it becomes widespread, higher costs (perhaps double or triple) are likely.

Hydrogen turbines are even further from commercial costing, although capital costs of between U.S. $1000 and $2000/kW have been suggested.

1.1.3 Fuel Cells

The fuel cell works on the principle of electrochemical combustion and offers the opportunity for highly efficient conversion of hydrogen and an oxidant (usually air) to electrical energy. Polymer electrolyte membrane fuel cells (PEMFCs) are undergoing major development and deployment programs for automotive use, and the first demonstration fuel cell buses using hydrogen were put into service during the

1990s. Cars using fuel cell technology and hydrogen fuel were introduced in very limited numbers late in 2002, although fully commercial vehicles are not expected until around 2010.

A fuel cell-powered vehicle is an electric vehicle. The fuel cell produces electricity from the hydrogen reaction, and this is then channeled through a power conditioning system into either one electric motor that takes the place of the internal combustion engine or smaller motors mounted independently on each wheel. No transmission linkage or gearbox is required.

Fuel cell vehicle development has been driven particularly by emissions regulations such as the California zero emissions vehicle (ZEV) mandate, which has forced automotive manufacturers to develop vehicles that produce no regulated pollutants. Only battery and hydrogen fuel cell electric vehicles currently satisfy the criteria of the mandate, and it appears that fuel cell development may outpace that of batteries, meaning that fuel cell vehicles are more likely to come into widespread use in the future. However, suitable performance and cost reduction must be demonstrated for this to happen. In addition, a refueling infrastructure for hydrogen must be developed if that is the fuel chosen.

Fuel cells are equally applicable in stationary end uses, where they are likely to be used not only for electricity generation but also for heat. A range of fuel cells in addition to the PEMFC can also be used for stationary end uses, with each offering specific advantages and drawbacks. Low-temperature fuel cells are theoretically more efficient at producing electricity from hydrogen than are high-temperature fuel cells. However, they do not produce useful heat except perhaps for space heating applications. Many high-temperature fuel cells will operate directly using a hydrocarbon fuel source such as natural gas rather than running on hydrogen. Although they convert the natural gas to hydrogen on the anode before oxidation, they cannot strictly be termed "hydrogen technologies" given that only a proportion of fuel cells for stationary power and heat production are expected to operate purely on hydrogen.

1.1.3.1 Economics of Fuel Cells The economics of a fuel cell system are highly unclear. Capital costs are high, as the technology is still generally in laboratory-scale development, and are quoted in the wide range U.S. $500 to $10,000/kW over the range of technologies. Operating and maintenance costs, lifetime, and durability still remain to be proven.

1.1.4 Catalytic Burners

A catalytic burner uses a coating of a catalyst, such as platinum, to electrochemically combust a fuel over a wide range of fuel/air ratios. A pure catalytic burner uses only the electrochemical reaction between fuel and oxidant to generate heat, whereas a hybrid system uses a catalyst to enhance a conventional combustion reaction. Hydrogen is fed to the catalytic surface, where it reacts with oxygen in the air to produce heat and water vapor. A catalytic burner is self-starting; the reaction begins when hydrogen is fed to the catalyst, and it ceases again as the flow stops. The temperature can be controlled over a wide range to enable applications as diverse as catalytic cooking stoves and low-intensity heating panels that could be hung, uncovered, on the wall of a room. Emissions from catalytic combustion systems will typically be zero.

1.1.5 Gas Boilers

Hydrogen can be burned in a boiler in the same way as can any fuel gas. Hydrogen fuel and air are fed to the combustion zone, where they are ignited to produce heat.

Using hydrogen in a natural gas boiler is much the same as using natural gas. A number of small modifications to the boiler are required, the most important of which is a reduction of the orifice size to control the flow of the gas. The flow rate of the input gas must also be adjusted to take care of the fact that hydrogen is much less dense than natural gas but has a higher calorific value.

Emissions of NO_x may be slightly increased initially, but optimizing the flow rate and the introduction of boilers dedicated to hydrogen use should solve this problem, whereas carbon monoxide (CO) and hydrocarbon emissions both will be reduced.

1.1.6 Liquid and Gaseous Hydrogen and Hydrogen Mixtures

Apart from the technologies themselves, hydrogen will be used in various applications in different forms.

Aircraft apart, hydrogen can be stored on-board most vehicles in liquid or gaseous form or in metallic or other compounds. Vehicles using both gaseous and liquid systems have been demonstrated successfully and continue to be developed. Liquid hydrogen systems typically offer the advantage of greater energy density than does gaseous hydrogen, allowing greater range for less space occupied within the vehicle. However, very high-pressure storage systems of 10,000 psi/700 bar offer similar energy densities

to liquid hydrogen. But both combustion engines and fuel cells will use the hydrogen in a gaseous state.

In addition, hydrogen may be mixed with other fuels to improve combustion characteristics. This has been the subject of experiment since German engineer Rudolf Erren conducted early conversions of internal combustion engines during the 1930s. However, it is difficult to mix hydrogen with a liquid fuel effectively, and methane (or natural gas) has often been chosen. Mixtures of hydrogen in methane varying from 15 to 50% by volume have been examined. The effect of this mixing is to generally improve the combustion characteristics of the natural gas, which has a low flame speed. Increasing this speed assists in reducing hydrocarbon formation at the cylinder walls, thereby lowering production of CO, but also enables the engine to be run on a leaner mixture, cutting down NO_x emissions. CO_2 emissions from the combustion process will also be reduced because less carbon is contained in the fuel. Nevertheless, it appears likely that pure hydrogen will be preferred in automotive applications.

In stationary power systems, pure hydrogen will be used in remote areas where an integrated renewable power system can be put into place. The hydrogen will be used as an energy storage buffer to compensate for the intermittency of the system, but excess hydrogen could also be used for other purposes (e.g., as a transport fuel). Pure hydrogen may also be used in many other applications, but hydrogen/methane mixtures are being considered as a way of reducing CO_2 emissions from natural gas networks. Up to 20% hydrogen by volume can be introduced into the natural gas transmission system in many countries, without requiring end use technologies such as gas ovens to be adjusted and while remaining within gas standard specifications. CO_2 emissions from end use are reduced by the same amount as the hydrogen introduced into the system.

2. TRANSPORT END USES

Hydrogen is being explored for use in transport in a wide variety of applications. Motivated primarily by regulations to reduce emissions of CO_2 and regulated pollutants, all major automotive manufacturers have hydrogen research and development (R&D) under way for cars and some for buses. Smaller vehicles (e.g., scooters) and larger vehicles (e.g., ships, airplanes) are also being developed. Hydrogen's low density creates particular problems for transport applications because the range of vehicles using hydrogen fuel may be short, but it may be mixed with other fuels to give some of the advantages of each.

2.1 Hydrogen Vehicle Types

Hydrogen could be used to power vehicles in all sectors, although it is expected to be used first in smaller scale applications such as buses and cars and only later in replacing the fuel for very large engines such as marine diesels.

Geographical locations that are sensitive to emissions may be the first to adopt hydrogen vehicles, and these may be specific to the needs of the location. For example, many cities in Southeast Asia suffer from high levels of pollution generated by two-wheeled vehicles with two-stroke engines, and so hydrogen scooters, both internal combustion engine and fuel cell, are under development. Conversely, hydrogen buses are seen as a good early technology for European cities.

2.1.1 Niche Vehicles: Scooters, Forklifts, Battery Replacements

Scooters and small motorcycles can be run on either fuel cells or internal combustion engines, and developments of each are under way. Hydrogen for these applications is typically stored either in compressed gas form or in a metal hydride, providing a range on the order of 100 km and very low emissions.

Forklift trucks conventionally operate on batteries in warehouse and food industry environments where emissions must be zero to meet health and safety regulations. Hydrogen used in fuel cells can also meet these regulations without the long recharge periods required by many batteries.

Many other specialist vehicles, including aircraft tugs at airports, operate using batteries. Hydrogen used in fuel cells is under consideration as a replacement fuel for these vehicles.

2.1.2 Common Vehicles: Cars, Buses, Vans, Trucks

The largest potential for hydrogen vehicles may come in the main automotive markets, although the cost and technology targets for hydrogen cars are very challenging. However, the use of hydrogen in cars is viewed as one leading option to enable continued mobility while reducing some of its environmental impacts. Analysis carried out in Germany has suggested that hydrogen offers the best long-term combination of low emissions, renewable feedstock potential and vehicle performance, and should be seen as the main long-term strategic fuel for the country.

Significant debate is ongoing as to the best way in which to introduce hydrogen into the car market. Both buses and cars using liquid hydrogen burned in internal combustion engines are being demonstrated in North America, Japan, and Europe. The advantage over fuel cell vehicles is that the engine technology is close to maturity in terms of both cost and technical development. However, liquid hydrogen is bulky to carry and is subject to losses through "boil-off," where a small rise in temperature causes evaporation and a subsequent pressure buildup within the storage vessel and gas is then vented. In addition, the efficiency of the internal combustion engine is lower than that of a fuel cell, so both vehicle range and fuel cost can be worse than with a fuel cell vehicle.

Fuel cell vehicles have also been demonstrated using hydrogen stored as a liquid, as compressed gas, or in metal hydrides. The vast majority of these have used PEMFCs, although some early bus systems used phosphoric acid fuel cells (PAFCs). As of early 2003, some tens of buses and cars had been tested or were within demonstration projects. Fuel cell vehicles are entering into limited precommercial demonstration, but the technology must still be proven to be reliable under the same conditions as conventional engines, costs must be reduced, and fueling infrastructure must be provided. Fuel cell vehicles are not expected to be available commercially until around 2010.

Fuel cell and hydrogen internal combustion engine light duty vans are also undergoing tests, but limited work is ongoing in developing hydrogen fueling for heavy-duty vehicles such as long-distance trucks. The characteristics of these vehicles are (1) high efficiency brought about by the use of heavy diesel engines and numerous gears to allow the engine to operate at close to optimal efficiency at all times and (2) very long engine life (typically hundreds of thousands of hours). Fuel cell technology is not yet at a stage where it can compete with these characteristics, and hydrogen storage is not yet advanced enough to allow sufficient range for the use of hydrogen in internal combustion engines in trucks.

2.1.3 Large/Specialist Vehicles: Locomotives, Boats, Ships, Planes, Submarines, Aerospace

Large-scale vehicles using hydrogen have also been considered, and some have been put into practice. In submarines and spacecraft, the use of hydrogen is comparatively common. In the former, the use of hydrogen in fuel cell systems means that emissions-free propulsion can be provided with only a very low heat and noise signature. These characteristics are invaluable for a submarine operating in "stealth" mode, where it aims to remain undetected. Typically, the hydrogen is stored in metal hydrides, which are both chemically stable and heavy, providing ballast for the submarine. Oxygen is also stored on-board, and the exhaust of the fuel cell—pure water—can be vented directly into the sea water, where it produces no telltale bubbles. The fuel cell is used not as the primary drive but rather as a supplementary power source.

Spacecraft use liquid hydrogen and liquid oxygen as fuel for rocket boosters. The mix contains approximately 40% more power per unit mass than do other rocket fuels, enabling greater payloads to be transported. A typical specific impulse for a rocket engine using liquid hydrogen/liquid oxygen is approximately 420 s, as compared with a liquid oxygen/kerosene rating of perhaps 250 s. Hydrogen has also been used in fuel cells on-board spacecraft, from the Gemini mission to the Space Shuttle, to produce both electrical power for the equipment and drinking water for the astronauts.

Aircraft have been designed to use hydrogen, and several small-scale tests have been conducted, including a Russian experiment where one of the turbines of a Tupolev aircraft was converted and operated successfully on liquid hydrogen during a test flight, but no complete aircraft yet exists. Liquid hydrogen contains less energy per unit volume than does kerosene, meaning that larger fuel tanks will be required in hydrogen aircraft to give them the same range as in conventional aircraft. These tanks cannot be situated in the wings, unlike in conventional airplanes, because the shape of the tanks is not suitable for maintaining liquid hydrogen at very low temperatures and the fluid will not act as ballast in the same way as will kerosene. Instead, the use of large tanks in the top of the main body of the aircraft has been suggested. The volume of hydrogen required will be on the order of four times that of kerosene for equivalent energy content. However, hydrogen is very light, and this will enable hydrogen aircraft either to be lighter at takeoff than their conventional counterparts or to carry greater payloads.

The use of locomotives using hydrogen has been suggested, but limited investigation has been carried out. Erren conducted early work during the 1930s, operating a few locomotives on hydrogen mixtures. Similarly, little analysis has been conducted on ships, particularly on a large scale. Several small boats operating on rivers have been converted to hydrogen and fuel cell use, and the potential exists for further exploitation of this area.

3. STATIONARY HEAT AND POWER END USES

In addition to transport, hydrogen can be used effectively in producing heat, power, or a combination of the two. Hydrogen can be used in place of other fuels in a wide range of different technologies, as suggested earlier.

3.1 Hydrogen for Heat Production

Hydrogen's generally high cost means that it is unlikely to be used as a fuel simply for heating because other fuels are effective for this purpose. However, if hydrogen is available cheaply, boiler, catalytic heater, or internal combustion engine technologies may be used to produce heat. Hydrogen used in fuel cells will also produce heat, and even low-temperature fuel cell technologies such as PEMFCs can prove to be sufficient for space heating and hot water requirements.

3.2 Hydrogen for Power Production

Hydrogen is typically seen as a form of energy storage and clean fuel, and it is most frequently considered in the context of power production.

The most efficient and cleanest technology for producing power from hydrogen is the fuel cell, but high cost and limited availability mean that few have been introduced outside of demonstration markets. The first small backup PEMFC generator (1.2 kW) was made available late in 2002 at a price of U.S. $6000, including a hydrogen supply stored in a metal hydride within the package. Other fuel cell-based hydrogen power systems have typically been on the order of several or tens of kilowatts, although a pure hydrogen-using PAFC operates in the city of Hamburg, Germany, and produces 200 kW of power along with a similar amount of heat for a local district heating system. The largest fuel cell system run to date was an 11-MW PAFC system in Japan that operated on hydrogen produced from steam-reformed natural gas to produce electrical power and heat.

3.2.1 Remote Integrated Hydrogen Power Systems

Many islands and other geographically remote areas suffer both from high imported fuel costs and from pollution caused when that fuel is used in providing heat and power. A typical example might be the use of diesel oil imported to many small islands and used in old generator sets for power production. However, these islands frequently have high levels of indigenous renewable energy resources such as solar, wind, and wave power. Using renewable resources to produce electricity as required, with any excess electricity subsequently used in electrolysis to produce hydrogen, can provide an integrated power system with no polluting emissions.

A basic system would include renewable electricity production, electrolysis and hydrogen storage (typically as a compressed gas), and an end use technology such as a fuel cell or combustion engine. Sizing the system is a complex process that depends on the fluctuation between renewable power availability and local demand, both diurnally and annually, and the volume of hydrogen storage required may be large to meet all fluctuations between supply and demand. The system may also be used in conjunction with a conventional power supply to meet either baseload or peaking power production. The economics of the system will also depend significantly on local factors, and electricity produced in this way may be expensive, possibly on the order of U.S. $1.00/kWh. However, this is frequently competitive with imported diesel while avoiding the pollution associated with the use of hydrocarbon fuels.

Hydrogen systems may be equally integrated with biomass production of hydrogen, making them both flexible and appropriate for use in a wide range of countries and circumstances. Rural electrification in many areas, including developing countries, may be enhanced through the use of these technologies.

3.2.2 Large-Scale Hydrogen Power Production

As hydrogen energy becomes more widespread, it may also be used for large-scale power production, with end use technologies of tens of megawatts or even gigawatts. Large-scale hydrogen combustion turbines could potentially be sited closer to urban areas than can fossil fuel-powered turbines because they would produce no pollution and, hence, would require no smokestacks. A 500-MW liquid hydrogen-powered turbine is under development in Japan, with the expectation that it will be used in the future as part of that country's commitment to hydrogen energy.

Hydrogen could also be used in fuel cells for this purpose, with several smaller fuel cells (perhaps on the order of 1 MW) connected to provide a modular power production opportunity of greater size. However, it might be that the modular nature of the fuel cells, and the potentially distributed nature of hydrogen production resources, offers better opportunities for decentralized hydrogen power systems.

4. OTHER HYDROGEN END USE TECHNOLOGIES

A small number of technologies not directly associated with heat, power, or transport production may also operate on hydrogen. In general, these are expected to be technologies in which hydrogen plus a fuel cell replaces a battery. Modeling suggests that under optimal conditions, a hydrogen and fuel cell combination could provide both more power and more energy from the same weight and volume than do many advanced batteries. However, such small hydrogen fuel cell systems have not yet been optimized, and so batteries continue to offer the most convenient solution.

Possible end uses include military power packs for communications, laptop computer power, and battery replacement for mobile phones. However, in the smallest applications, such as mobile phones, methanol or another liquid fuel may be more suitable than hydrogen due to its higher volumetric energy density.

5. ECONOMICS OF HYDROGEN SYSTEMS

The fundamental economics of hydrogen energy systems are the same as those of conventional energy systems. However, hydrogen energy systems suffer from two peculiarities in comparison with the majority of conventional technologies. The first is that hydrogen is typically one of the more expensive fuels to transport, so that the economics become significantly worse with long-distance fuel movement. In conventional fossil fuel systems, fuel transport costs are generally a marginal factor in the overall economics. In addition, hydrogen technologies are still largely under development, and so costs may vary widely even in similar applications.

5.1 An Introduction to the Costs of Hydrogen Systems

The basic economics of hydrogen systems are dependent, as is any energy system, on three primary factors: (1) fuel (hydrogen) cost, (2) equipment capital cost, and (3) equipment performance (e.g., lifetime, maintenance, efficiency).

As has been discussed, the basic cost of hydrogen is higher than that of the raw materials from which it is produced. Losses are inherent in any conversion process, and the conversion equipment also has a cost associated with it. However, the cost of the end use service is what should be considered as far as possible—the cost per unit of luminosity, of mobility, or of warmth. Because of this, the efficiency and other characteristics of the end use technology have a critical role.

High fuel cost can be compensated, to a greater or lesser extent, by high end use efficiency. As a simplistic example, if fuel costs are twice as high for hydrogen but the efficiency of end use is doubled in comparison with a conventional fuel, the cost of the end use due to fuel will remain unchanged. Other costs, such as those for maintenance and capital, may still alter the final delivered service cost. For this reason, although fuel cells suffer very high capital costs in the near term due to their early development status, they can offer some expected economic benefits related to efficiency. In addition, maintenance costs are expected to be low in the long term, although this has yet to be demonstrated.

5.2 Environmental Costs

An increasingly important element to be considered in the economics of any energy system is that of environmental costs, both internal and external. A primary criticism of fossil energy systems is that they have severe environmental impacts associated with most stages of the fuel chain, from discovery and extraction of raw materials to emissions from processing and final conversion. A significant advantage of many hydrogen systems is the absence or reduction of these environmental impacts.

Emissions from end use are one of the effects most commonly considered. In conventional systems, these tend to have the greatest health impact, although emissions from refining and processing must also be taken into account. Ideally, to ensure valid comparisons, emissions from entire fuel chains should be evaluated.

These emissions can also be considered as part of the economic equation for evaluation of an energy system. To do so requires their representation as costs using principles of externality costing that are not trivial. However, this enables expensive fuels that produce little or no pollution to be compared, on a somewhat more equal basis, with cheap fuels that have a high health impact cost not included in the price of the fuel. Hydrogen can be considered one of the more expensive but less polluting fuels.

6. CONCLUSIONS

Hydrogen energy end use technologies are not yet mature, unlike conventional technologies. However, they offer potentially significant advantages in terms of low or zero emissions and flexibility in fuel sources. As immature technologies, costs are high and reliability and durability are not yet proven. Nevertheless, early demonstrations of hydrogen vehicles using internal combustion engines and fuel cells are under way and expected to become more widespread.

Similarly, hydrogen energy systems for power and heat production are operating under demonstration conditions in many areas. The possible integration of hydrogen energy storage with renewable energy sources offers the prospect of economically efficient remote power systems and reductions in the external costs of energy associated with many fossil fuels.

Over the long term, hydrogen energy is likely to be used in many everyday situations, and hydrogen energy technologies will replace many conventional technologies. The economics of such hydrogen technologies will improve as they enter widespread use, and the added value of low pollution should further enhance the value of such systems.

SEE ALSO THE FOLLOWING ARTICLES

Alternative Transportation Fuels: Contemporary Case Studies • *Arid Environments, Impacts of Energy Development in* • *Fuel Cells* • *Fuel Cell Vehicles* • *Hydrogen, History of* • *Hydrogen Production* • *Hydrogen Storage and Transportation* • *Internal Combustion (Gasoline and Diesel) Engines* • *Transportation Fuel Alternatives for Highway Vehicles*.

Further Reading

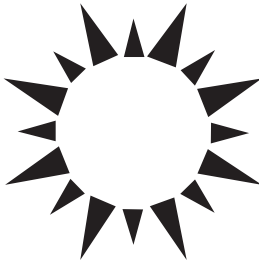
Hoffmann, P. (2001). "Tomorrow's Energy: Hydrogen, Fuel Cells, and the Prospects for a Cleaner Planet." MIT Press, Cambridge, MA.

Ogden, J. M. (1999). Prospects for building a hydrogen energy infrastructure. *Annu. Rev. Energy Environ.* **24,** 227–279.

Padró, C. E. G., Putsche, V. (1999). "Survey of the Economics of Hydrogen Technologies," NREL/TP-57027079. National Renewable Energy Laboratory, Golden, CO.

Hydrogen, History of

SETH DUNN
Worldwatch Institute
New Haven, Connecticut, United States



1. **Discovery of Hydrogen**
2. **Origins of Interest in Hydrogen Energy**
3. **Roots of the Hydrogen Economy Concept**
4. **Globalization of the Hydrogen Movement**
5. **Future of Hydrogen History**

Glossary

electrolysis The process of splitting water into its components, hydrogen and oxygen, by means of an electrical current.

energy carrier A form of matter that carries energy from one point to another, as distinct from an energy source, for which hydrogen is often mistaken.

fuel cell An electrochemical device that combines hydrogen and oxygen, producing an electrical current and water as a by-product.

hydrogen economy A term first introduced in the 1970s to describe the concept of an energy system based primarily on the use of hydrogen as an energy carrier.

proton-exchange membrane One type of fuel cell, featuring a membrane through which protons but not electrons can pass, forcing the latter to move along an electrode and generate a current.

The history of hydrogen as an energy carrier begins with its discovery and the comprehension that it is one of the basic building blocks of the universe. When the early Greek philosophers named water, air, fire, and earth to their list of elements, they did not realize that water in fact consisted of two elements. One of these elements is oxygen, the respiratory prerequisite for human life. The other is hydrogen, which modern science has revealed as the lightest and most abundant element in the universe.

1. DISCOVERY OF HYDROGEN

By the most recent estimates, hydrogen accounts for approximately 75% of the mass of the entire universe and accounts for more than 90% of all molecules. According to the Harvard astrophysicist Steven Weinberg, between 70 and 80% of the observable universe is composed of hydrogen, with the rest attributable to helium, the universe's second lightest element. This is why hydrogen is called *Wasserstoff*—the stuff of water—in German.

The first recorded production of hydrogen took place in the 15th century, when the Middle Age physician Theophrastus Paracelsus dissolved iron in the acid vitriol, producing hydrogen gas. Paracelsus did not note that hydrogen was flammable; this was left to Turquet de Mayeme and Nicolas Lemery, French scientists who, in the 17th century, mixed sulfuric acid with iron, yet did not think that the resulting gas could be an element unto itself, only a burnable form of sulfur.

The actual classification and formal description of hydrogen would have to wait until the 18th century. As the early Greeks would have no doubt appreciated, the modern discovery of hydrogen was tied to advances in identifying its fellow component in water, oxygen. Such advances would, of course, require first that scientists reconsider the conventional notion of air as a basic element. The first to question this long-standing characterization was Herman Boerhaave, a Dutch doctor and naturalist who believed that air contained an ingredient that made breathing and combustion possible. Wrote Boerhaave in 1732, "The chemist will find out what it actually is, how it functions, and what it does; it is still in the dark. Happy he who will discover it." Boerhaave had a modicum of intellectual support in the 17th-century writings of British scientists Robert Boyle and John Mayo, who suspected that some substances in the air were responsible for the process of combustion. But this ran counter to the phlogiston theory, another 17th-century idea. The phlogiston theory, one of the first attempts to explain the process of combustion, was first published in 1697. Its origins lay with a German scientist named Georg Ernst Stahl, who

 © 2004 Elsevier Inc. All rights reserved.

contended that a substance known as phlogiston gave all matter the ability to burn. Because phlogiston disappeared from material during the combustion process, and because it was considered impossible to reduce the substance to a pure state, this was not a theory easily disproved. Modern chemistry has since shown that Stahl had it precisely backwards: the process of burning arises from adding a substance—oxygen. The idea that combustion consisted of the release of phlogiston served as an impediment to the recognition of oxygen and hydrogen as gases.

Fortunately for scientific progress (but rather unfortunately for Stahl), the late 18th century would witness a scientific race of sorts to successfully isolate oxygen. Among those who discovered, but did not name, the element were Joseph Priestley, a British minister; Carl Wilhelm Scheele, a doctor of Swedish–German descent; and Antoine Laurent Lavoisier, by then France's leading chemist. Although Scheele was the first to produce pure oxygen, between 1771 and 1772, the other two beat him to the press, publishing their findings in 1774. Because the scientists studying air were also investigating water, the isolation and identification of hydrogen proceeded in parallel fashion. But most chemists, such as Boyle, still saw the hydrogen they produced as another type of air. That would change with the English nobleman Henry Cavendish, the first to discover and describe many of the important characteristics of hydrogen. Ironically, Cavendish did not name hydrogen because he was not fully free from the phlogiston theory—he thought he had discovered the substance in its pure state. But his key breakthrough was to identify two kinds of "factitious air": "fixed air," or carbon dioxide, and "inflammable air," or hydrogen. In a 1766 paper to the Royal Society of London, Cavendish provided exact measurements of the weights and densities of both gases, revealing the inherent lightness of hydrogen. (Recognition of this led the French physicist Jacques Alexandre Cesar Charles to fly a hydrogen-filled balloon to a height of close to 2 miles in 1783.) In his *Experiments on Air* treatise of the 1780s, Cavendish showed it was possible to mix hydrogen with air, ignite the mixture with a spark, and produce water.

Cavendish's findings attracted Lavoisier, who tried his experiments in reverse. Splitting water molecules into hydrogen and oxygen in an experiment considered to be a breakthrough, Lavoisier combined hydrogen and oxygen to produce 45 g of water (which are still preserved in the French Academy of Science). In February 1785, Lavoisier conducted, before a large group of scientists, definitive experiments proving that hydrogen and oxygen are the fundamental constituents of water. His greatest work, *The Method of Chemical Nomenclature*, labels the "life-sustaining air" oxygen and the "inflammable air" hydrogen.

The first large-scale production of hydrogen was helped along by France's historical circumstances: the storming of the Bastille in 1789 and the subsequent warfare of the French Revolution. Guyton de Norveau, chemist and member of the Comité de Salut Public (Committee for Public Salvation), proposed that the army use hydrogen-filled captive balloons as observation platforms, a proposal that was approved after he and Lavoisier repeated the latter's 1783 experiment. The scientist Jean Pierre Coutelle built the first hydrogen generator; this was a furnace with a cast iron tube, containing iron fillings, with steam piped in at one end and hydrogen emerging at the other. A standard generator was later developed and included temperature-control systems, precursors to the coal gas generators that would later be used for lighting and heating in the early 19th century; coal gas would often be misnamed "hydrogen gas."

2. ORIGINS OF INTEREST IN HYDROGEN ENERGY

In 1800, six years after Lavoisier went to the guillotine in the melee of the French Revolution, another important scientific discovery was made. This was electrolysis, the splitting of water into hydrogen and oxygen by the passage of an electric current through it. William Nicholson and Sir Anthony Carlisle, British scientists, had discovered electrolysis, only a few weeks after Alessandro Volta built the first electric cell. As the 19th century progressed, other proposals to put hydrogen to practical use emerged, and from intriguing sources—including the clergy. In 1820, Reverend W. Cecil, a Fellow of Magdalen College and of the Cambridge Philosophical Society, read "On the Application of Hydrogen Gas to Produce Moving Power in Machinery." His treatise described the space and time limitations of water-driven and steam engines, and the advantages of a hydrogen engine in providing motive force in any location and with little preparation needed. The transactions of the society provide considerable detail of the engine, but no evidence as to whether it was ever built by Cecil.

Hydrogen, and its unusual properties, became a topic of growing interest among not only scientists but also science fiction writers. One of the most

well-known fictional discussions of hydrogen is found in Jules Verne's *The Mysterious Island*, one of the writer's final works, published in 1874, almost precisely one century before modern hydrogen research took off. Set in the American Civil War, the book at one point depicts five characters speculating on the future of the Union, and the impact on commerce and industry were coal supplies to dwindle. Whereas four of the characters express fear about the loss of coal-powered machinery, railways, ships, manufacturing, and other forms of modern civilization, the other character, the engineer Cyrus Harding, asserts that they will burn "water decomposed into its primitive elements":

> *Yes, my friends, I believe that water will one day be employed as fuel, that hydrogen and oxygen which constitute it, either singly or together, will furnish an inexhaustible source of heat and light, of an intensity of which coal is not capable. Some day the coalrooms of steamers and the tenders of locomotives will, instead of coal, be stored with these two condensed gases, which will burn in the furnaces with enormous calorific power. There is, therefore, nothing to fear. As long as the earth is inhabited it will supply the wants of its inhabitants, and there will be no want of either light or heat as long as the productions of the vegetable, mineral, or animal kingdoms do not fail us. I believe, then, that when the deposits of coal are exhausted we shall heat and warm ourselves with water. Water will be the coal of the future.*

Verne may have skipped two energy transitions along the path from coal to hydrogen—oil and natural gas—and therefore he missed what energy expert Robert Hefner would in 2000 depict as the "age of energy gases" (Fig. 1). Nor did Verne mention what the primary source of the hydrogen might be. But it is a prescient passage, given the state of late-19th-century science.

Another adventure story, *The Iron Pirate*, authored by Max Pemberton and published in England in 1900, featured speedy battleships using hydrogen engines. The book influenced the British scientist W. Hastings Campbell, as he related to colleagues at Britain's Institute of Fuel in 1933. His reference was noted in the *Journal of the Institute of Fuel* as "another instance of the very annoying persistence with which art always seemed to anticipate discoveries."

Interest in hydrogen as a fuel had begun to blossom in the wake of the First World War, prompted in part by concern about access to energy supplies. At first, interest was strong in Germany and England, as well as in Canada. The early 20th century had seen the emergence of several new companies with a stake in hydrogen production, such as the Canadian Electrolyser Corporation Limited (now Stuart Energy Systems, one of the world's leading electrolytic hydrogen plant manufacturers). The firm arose from the interest of its founder, Alexander Stuart, in utilizing the nation's considerable excess hydroelectric capacity. The first commercial uses of its electrolyzers were not, however, to make hydrogen fuel, but rather to make hydrogen and oxygen for steel-cutting procedures. Other uses included fertilizer production, and in 1934 Ontario Hydro constructed a 400-kW electrolysis plant. But World War II and growing use of natural gas in Canada halted the project.

Back in Britain, a young Scottish scientist, J. B. S. Haldane, prophesied, in a lecture at Cambridge University in 1923, hydrogen as the fuel of the

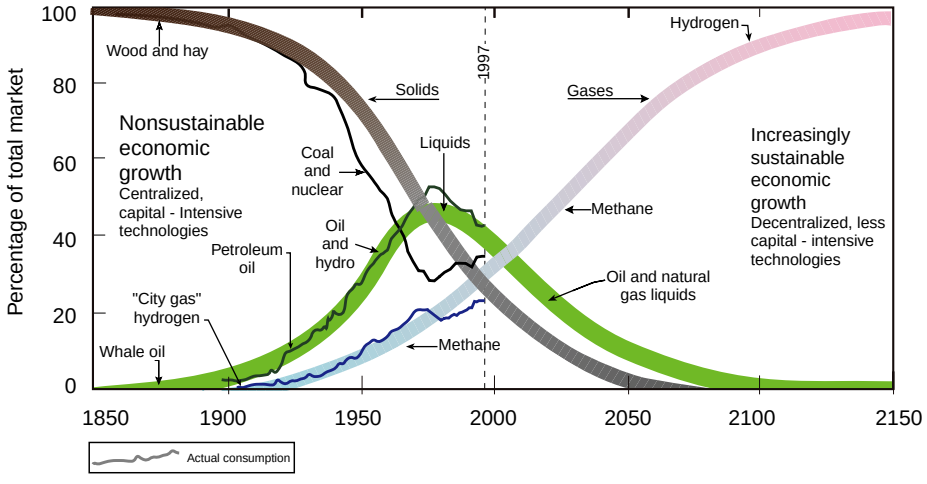


FIGURE 1 In the "age of energy gases," the transition in global energy systems. Reprinted from Dunn (2001), with permission.

future—derived from wind power, liquefied, and stored. Future British energy needs would be met by metallic windmills that decomposed water and stored the gases in vast underground reservoirs for later use. Britain's other early researchers and proponents of hydrogen as a fuel included Harry Ricardo, one of the developers of the internal combustion engine, and R. O. King, a member of the British Air Ministry Laboratory.

Germany counted numerous hydrogen supporters among its engineers and scientists, some of them influenced by Verne. Franz Lawaczeck, a turbine designer, sketched concepts for car, trains, and engines and is considered the first to propose the transport of hydrogen energy by pipeline. More whimsical was Hermann Honnef, who envisioned huge steel towers with windmills producing hydrogen. Use of liquid hydrogen in aircraft attracted attention in Italy, the United States, and Germany in the late 1930s. Lecturing before the American Institution of Electrical Engineers, Igor Sikorski cited the fuel's potential for improving the speed and performance of long-distance aircraft. But the most intriguing application was for the German zeppelins, in wide use in the 1920s and 1930s. As zeppelins lost weight through fuel consumption, they typically blew off hydrogen; engineers proposed burning the hydrogen as extra fuel, increasing output and saving energy. Although tests in the Mediterranean were promising, there is no evidence that the principles were applied to typical flights.

A prominent figure in early hydrogen advocacy was the German Rudolf Erren. Erren spent much of the 1930s in Britain, where he developed advanced combustion processes that would allow the use of hydrogen as a fuel or "booster" additive. In the mid-1930s, he and the engineer Kurt Weil proposed to the Nazi government that it convert from the internal combustion engines to these multifuel systems. At the time, Germany was interested in economic self-sufficiency and reducing dependence on imported liquid fuel. Eventually, between 1000 and 4000 cars and trucks were converted to the new system. Erren's efforts extended elsewhere. In Dresden, a hydrogen rail car ran for several years. In England, delivery vans and buses were tested; a prototype plane was readied and hydrogen-propelled submarines and torpedoes drew public interest in the early 1940s. But Erren's possessions were confiscated during World War II, after which he was repatriated to Germany. None of his engines survived the war years.

The war spurred direct interest in hydrogen in areas of the world where fuel supplies were in danger of being cut off. As Australia's wartime fuel needs grew and supplies from Borneo were lost to Japan, the government of the state of Queensland authorized construction of a hydrogen plant in Brisbane, using off-peak electricity. But the Allied victory in 1945 and the reversion to cheap oil and gasoline halted progress—as may be said about postwar hydrogen progress more generally.

3. ROOTS OF THE HYDROGEN ECONOMY CONCEPT

The early 1950s saw a resumption of interest in hydrogen. The British scientist Francis Bacon developed the first practical fuel cell. Though invented in 1836, the fuel cell had yet to see significant applications outside the laboratory, a situation that would soon change with the U.S. space program.

In the 1960s, the idea of hydrogen as a medium for energy storage spread. German physicist Eduard Justi proposed in 1962 the recombination of hydrogen and oxygen in fuel cells, and in 1965 the production of solar hydrogen along the Mediterranean, for later piping to Germany and other nations. John Bockris, an American electrochemist from Australia, proposed in 1962 the energizing of U.S. cities with solar energy from hydrogen, an idea he expanded on in his 1975 book, *Energy: The Solar-Hydrogen Alternative*. Bockris traces the coining of the term hydrogen economy to a 1970 discussion at the General Motors (GM) Technical Center in Warren, Michigan. As a consultant to GM, Bockris was discussing possible alternatives to gasoline in the wake of growing environmental awareness, and the group agreed that "hydrogen would be the fuel for all types of transports." GM began experimental work on hydrogen, but would later be eclipsed by overseas carmakers.

Various visions of the hydrogen economy began to emerge. The Italian scientist Cesare Marchetti, an influential advocate in Europe, promoted the large-scale production of hydrogen from the water and heat of nuclear reactors. Derek Gregory and Henry Linden of the Institute of Gas Technology, leaders in early U.S. hydrogen research and development, were motivated by the prospect of hydrogen as a substitute for natural gas. Space scientists and engineers continued to pursue liquid hydrogen potential. Despite the limited efforts of GM, Ford, and Chrysler, interest among automotive engineers grew. The Army, Navy, and Air Force pursued liquid

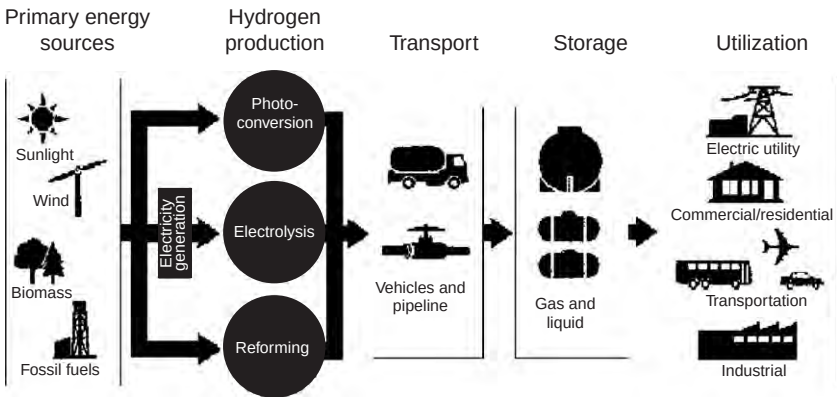


FIGURE 2 The components of a hydrogen energy system. Reprinted from Dunn (2001), with permission.

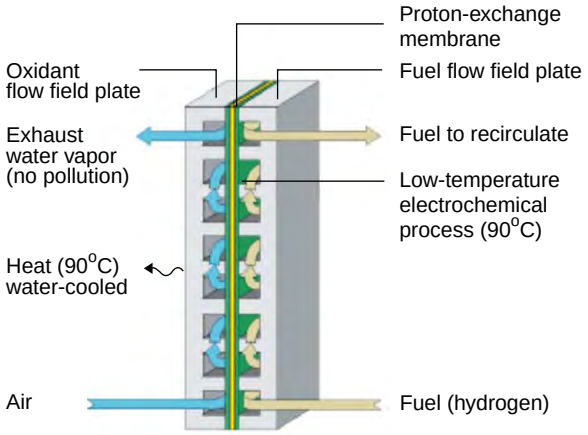


FIGURE 3 A proton-exchange membrane fuel cell. Reprinted from Dunn (2001), with permission.

hydrogen and oxygen applications, their experiments spilling over into the space program. During the 1970s, aided by the two oil shocks, both environmental concerns and energy security accelerated interest in hydrogen. Popular journals such as *Business Week*, *Fortune*, and *Time* ran stories on hydrogen development. Researchers formed groups such as the International Association for Hydrogen Energy (which currently publishes the *International Journal of Hydrogen Energy*).

In the 1980s and 1990s, governments and international organizations began to pay additional attention to the idea of a hydrogen system and its components of production, delivery, storage, and use (Fig. 2). The United States, West Germany, European Community, and Japan began to lay out funding for hydrogen research and development in the mid-1970s, but would cut back as the effects of the oil shock waned. Nevertheless, the institutionalization of hydrogen into energy policy and strategy was underway. The 1980s were a paradoxical period for the hydrogen prospect. Official government support by and large waned, particularly within the U.S. Administration, which significantly cut funding for alternative fuel budgets to make room for its expanding Cold War weapons buildup. One unintended consequence of this benign neglect was the decision of a creative but frustrated engineer named Geoffrey Ballard, then under the employ of the Department of Energy, to leave Washington, D.C. for the west coast of Canada. In Vancouver, British Columbia, Ballard and several like-minded scientists decided to focus on developing electric batteries—a focus that soon switched to the proton-exchange membrane (PEM) fuel cell (Fig. 3).

Although the fuel cell had, by the 1980s, enjoyed several decades of use in the U.S. space program, efforts to bring the technology down to earth had been hampered by issues of cost and efficiency. Most fuel cells relied on significant amounts of the expensive metal platinum and were large and bulky, minor considerations for space programs but major problems for terrestrial commercial applications. Through a combination of ingenuity, government seed money (from the Canadian defense agency), venture capital, and persistence, Ballard's company achieved a 20-fold reduction in platinum requirements for its PEM fuel cell, with commensurate cost reductions. In 1993, Ballard rolled out the first PEM fuel cell bus at its headquarters near Vancouver. That same year, it entered into a cooperative agreement with Daimler Benz (later to become DaimlerChrysler) to jointly develop fuel cells for cars and buses, a partnership that was joined by Ford in 1997, with an overall estimated investment of $300 million. Daimler announced a $1 billion, 10-year commitment to

fuel cell development; Daimler's founders were the first to bring the internal combustion engine car to market, and the company appears intent on doing the same for the fuel cell vehicle.

The late 1990s were characterized by an auto race of sorts, as companies paraded demonstration fuel cell vehicles at auto shows and the media buzzed over whether and when carmakers would unveil their first commercial vehicles. Executives also seemed to be staging another competition based on who could sound most bullish, even a bit visionary, about the hydrogen prospect. Echoing the GM engineers of three decades ago, Executive Director Robert Purcell announced in 2000, at the annual meeting of the National Petrochemical and Refiners Association, that "our long-term vision is of a hydrogen economy." As with the introduction of fuel-efficient vehicles in the 1980s, and hybrid electric vehicles in 2000–2002, the Japanese automakers quietly poured money into efforts that threatened to put them ahead of their U.S. and European competitors. In late 2002, Honda and Toyota announced they would be the first to introduce fuel cell cars to the U.S. market.

Energy suppliers were also giving more attention to hydrogen as the century came to a close. In 1999, Royal Dutch/Shell established a core business, Shell Hydrogen, to pursue business opportunities related to the hydrogen economy—an announcement that caused British Petroleum, ChevronTexaco, and ExxonMobil to follow with similar but smaller investments—and advocacy. Testifying before the U.S. Congress, Frank Ingriselli of Texaco (soon to become ChevronTexaco) asserted that "greenery, innovation, and market forces are propelling us inexorably towards hydrogen. Those who don't pursue it, will rue it." Shell Hydrogen would become involved in one of the most interesting hydrogen developments of the late 20th century: the announcement by Iceland that it planned to become the world's first hydrogen economy. The idea of using hydrogen in Iceland was first proposed in the 1970s by a chemistry graduate student named Bragi Arnason as a way to utilize the island's abundant geothermal and hydro resources to reduce oil import reliance. As oil price volatility resurfaced in the 1990s, it coincided with the nation's struggle to address greenhouse gas emissions and with growing global interest in fuel cells and hydrogen. Acting on the recommendations of a panel chaired by Arnason (who had by then earned the nickname "Professor Hydrogen"), the Icelandic government formed in 1999 a consortium, Icelandic New Energy, with Shell, DaimlerChrysler, and the Norwegian power company Norsk Hydro, to investigate the potential for transforming the nation's energy economy to hydrogen between 2030 and 2040. The consortium planned to convert gradually, from petroleum to hydrogen, the buses of the capital, Reykjavik, followed by the island's entire fleets of buses, cars, and fishing boats.

Other islands, likewise vulnerable to oil price volatility, soon followed Iceland's lead. Hawaii approved in early 2001 a jump-start grant to support a public/private partnership in hydrogen research and development, intended eventually to make Hawaii a major player in hydrogen exports. The South Pacific island of Vanuatu launched a feasibility study for transitioning to hydrogen, and plans to be completely renewable-energy-based by 2020. Cuba is also contemplating a move to hydrogen.

These developments shed useful light on the converging forces that are renewing interest in hydrogen. Technological advances and the advent of greater competition in the energy industry are part of the equation. But equally important motivations for exploring hydrogen are the energy-related problems of energy insecurity, air pollution, and climate change—problems that are collectively calling into question the fundamental sustainability of the current energy system. In the 21st century, it is likely that there will be five drivers for hydrogen—technology, competition, energy security, air pollution, and climate change. They are far more powerful drivers than those that existed during the 20th century, and they help to explain why geographically remote islands, stationed on the front lines of vulnerability to high oil prices and climate change, are in the vanguard of the hydrogen transition.

4. GLOBALIZATION OF THE HYDROGEN MOVEMENT

Iceland and other nations represent just the tip of the iceberg in terms of the changes that lie ahead in the energy world. The commercial implications of a transition to hydrogen as the world's major energy currency will be staggering, putting a $2 trillion energy industry through its greatest tumult since the early days of Edison, Ford, and Rockefeller. During 2002, over 100 companies were aiming to commercialize fuel cells for a broad range of applications, from cell phones, laptop computers, and soda machines to homes, offices, and factories, to vehicles of all kinds. Hydrogen was also being researched for

direct use in cars and planes. Particularly active were fuel and auto companies, who were spending between $500 million and $1 billion annually on hydrogen. Leading energy suppliers had created hydrogen divisions, and major carmakers continued to pour billions of dollars into a race to put the first fuel cell vehicles on the market between 2003 and 2005. Buses were being test-driven in North America, Europe, Asia, and Australia. In California, 23 auto, fuel, and fuel cell companies and seven government agencies had partnered to fuel and test drive 70 cars and buses. Hydrogen and fuel cell companies had captured the attention of venture capital firms and investment banks anxious to get into the hot new space known as or energy technology (ET).

The geopolitical implications of hydrogen are enormous as well. Coal fueled the 18th- and 19th-century rise of Great Britain and modern Germany; in the 20th century, oil laid the foundation for the unprecedented economic and military power of the United States, of which the fin-de-siècle superpower status, in turn, may be eventually eclipsed by countries that harness hydrogen as aggressively as the United States tapped oil a century ago. Countries that focus their efforts on producing oil until the resource is gone will be left behind in the rush for tomorrow's prize. As Don Huberts, Chief Executive Officer of Shell Hydrogen, has noted, "The Stone Age did not end because we ran out of stones, and the oil age will not end because we run out of oil." Access to geographically concentrated petroleum has also shaped world wars, the 1991 Gulf War, and relations between and among Western economies, the Middle East, and the developing world. Shifting to the plentiful, more dispersed hydrogen could alter the power balances among energy-producing and energy-consuming nations, possibly turning today's importers into tomorrow's exporters.

The most important consequence of a hydrogen economy may be the replacement of 20th-century "hydrocarbon society" with something far better. The 20th century saw a 16-fold increase in energy use—10 times more than the energy used by humans in the 1000 years preceding 1900. This increase was enabled primarily by fossil fuels, which account for 90% of energy worldwide. Global energy consumption is projected to rise by close to 60% over the next 20 years. Use of coal and oil is projected to increase by approximately 30 and 40%, respectively.

Most of the future growth in energy is expected to take place in transportation, where motorization continues to increase and where petroleum is the dominant fuel, accounting for 95% of the total. Failure to develop alternatives to oil would heighten growing reliance on oil imports, raising the risk of political and military conflict and economic disruption. In industrial nations, the share of imports in overall oil demand will rise from roughly 56% today to 72% by 2010. Coal, meanwhile, is projected to maintain its grip on more than half the world's power supply. Continued increases in coal and oil use will exacerbate urban air problems in megacities such as Delhi, Beijing, and Mexico City, which experience thousands of pollution-related deaths each year. And prolonging petroleum and coal reliance in transportation and electricity would increase global carbon emissions from 6.1 to 9.8 billion tons of carbon by 2020, accelerating climate change and the associated impacts of sea level rise, coastal flooding, and loss of small islands; extreme weather events; reduced agricultural productivity and water availability; and the loss of biodiversity.

Hydrogen cannot, on its own, entirely solve each of these complex problems, which are affected not only by fuel supply but also by factors such as population, over- and underconsumption, sprawl, congestion, and vehicle dependence. But hydrogen could make a major dent in addressing these issues. By enabling the spread of appliances, more decentralized "micropower" plants, and vehicles based on efficient fuel cells (of which the only by-product is water), hydrogen use would dramatically cut emissions of particulates, carbon monoxide, sulfur and nitrogen oxides, and other local air pollutants. By providing a secure and abundant domestic supply of fuel, hydrogen would significantly reduce oil import requirements, providing the independence and energy security that many nations crave.

Hydrogen has also been increasingly recognized by the scientific community as facilitating the transition from limited nonrenewable stocks of fossil fuels to unlimited flows of renewable sources, playing an essential role in the "decarbonization" of the global energy system needed to avoid the most severe effects of climate change (Fig. 4). The United Nations 2001 *World Energy Assessment* emphasized "the strategic importance of hydrogen as an energy carrier"; the accelerated replacement of oil and other fossil fuels with hydrogen could help achieve "deep reductions" in carbon emissions and avoid a doubling of preindustrial CO_2 concentrations in the atmosphere, a level at which scientists expect major, and potentially irreversible, ecological and economic disruptions. Hydrogen fuel cells could also help address global energy inequities, providing fuel and

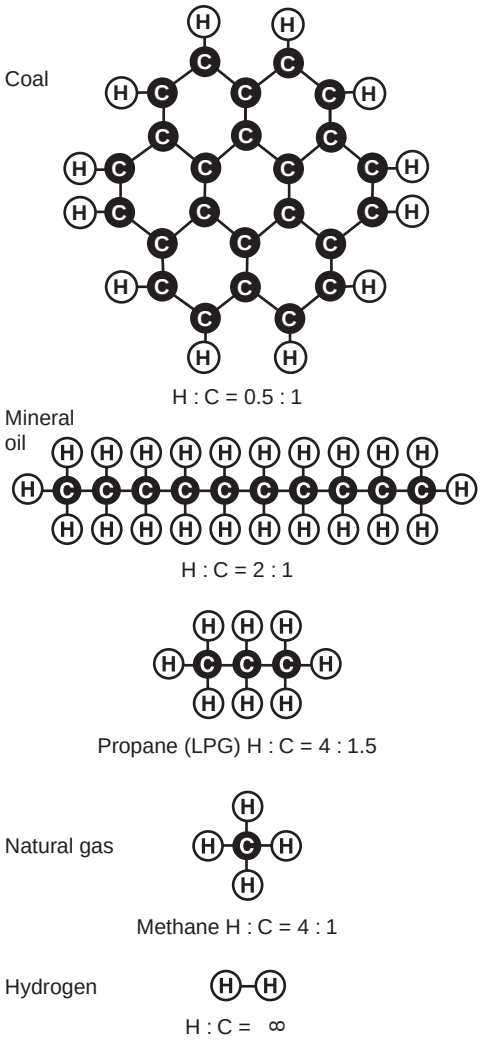


FIGURE 4 The atomic hydrogen:carbon ratios of selected energy sources. LPG, Liquid propane gas. Reprinted from Dunn (2001), with permission.

power and spurring employment and exports in the rural regions of the developing world, where nearly 2 billion people lack access to modern energy services.

Despite these potential benefits, and despite early movements toward a hydrogen economy, its full realization has faced an array of technical and economic obstacles. The feasibility of production, delivery, and use all awaits further improvement. Hydrogen is only beginning to be piped into the mainstream of the energy policies and strategies of governments and businesses, which still tend to aim at preserving the hydrocarbon-based status quo and expanding fossil fuel production. Market structures have been typically tilted toward expanding fossil fuel production. Subsidies to these energy sources, in the form of direct supports and the "external" costs of pollution, were estimated at $300 billion annually in 2001. The perverse signals in the energy market that lead to artificially low fossil fuel prices and encourage their production and use make it difficult for hydrogen and fuel cells (for which the production, delivery, and storage costs are improving but look high under such circumstances) to compete with the entrenched gasoline-run internal combustion engines and coal-fired power plants. This could push the broad availability of fuel cell vehicles and power plants a decade or more into the future. Unless the antiquated rules of the energy economy, aimed at keeping hydrocarbon production cheap by shifting the cost to consumers and the environment, are reformed, hydrogen will be slow to make major inroads.

One of the most significant obstacles to realizing the full promise of hydrogen is the prevailing perception that a full-fledged hydrogen infrastructure—the system for producing, storing, and delivering the gas—would immediately cost hundreds of billions of dollars to build, far more than a system based on liquids such as gasoline or methanol. This is tied to the "chicken-and-egg" dilemma that confronts any new infrastructure. Automakers have been hesitant to mass-produce fuel cell vehicles without assurance of a sufficient fueling network. Energy suppliers, meanwhile, have been resistant to ramping up hydrogen production without assurance of a sufficient number of vehicles to use the fuel. Because of this dilemma, auto and energy companies have been investing millions of dollars into the development of reformer and vehicle technologies that would derive and use hydrogen from gasoline and methanol, keeping the current petroleum-based infrastructure intact. To some analysts, this incremental path—continuing to rely on the dirtiest, least secure fossil fuels as a bridge to the new energy system—represents a costly wrong turn, both financially and environmentally. If manufacturers "lock in" to mass-producing inferior fuel cell vehicles just as a hydrogen infrastructure approaches viability, trillions of dollars of assets could be wasted. Furthermore, by perpetuating petroleum consumption and import dependence and the excess emission of air pollutants and greenhouse gases, this route would deprive society of numerous benefits. By the late 1990s, fossil fuels were the source of some 95% of the hydrogen being produced worldwide—approximately 400 billion cubic meters in 1999, primarily not for energy but for the refining of petroleum and the manufacture of resins, plastics, solvents, and other industrial commodities. Over the long run, critics argued, this proportion would need

to be shifted toward renewable sources, not maintained, for hydrogen production to be sustainable.

The "fuel choice" question heated up in the 1990s, complicated by competing life-cycle and "well-to-wheels" studies by government, industry, and nongovernmental organizations. A growing number of scientists in government and industry openly challenged the conventional wisdom of the incremental path. Their research suggested that the direct use of hydrogen would be the quickest and least costly route—to the consumer and the environment—toward a hydrogen infrastructure. Their studies pointed to an alternative pathway that would initially use the existing infrastructure for natural gas (the cleanest fossil fuel and the fastest growing in terms of use) and employ fuel cells in niche applications to buy down their costs to competitive levels, spurring added hydrogen infrastructure investment. As the costs of producing hydrogen from renewable energy fell, meanwhile, hydrogen would evolve into the major source of storage for the limitless but intermittent flows of the sun, wind, tides, and Earth's heat. The end result would be a clean, natural hydrogen cycle, with renewable energy used to split water into hydrogen, which would be used in fuel cells to produce electricity and water—which then would be available to repeat the process (Fig. 5). Some of the experts argued that there were no major technical obstacles to the alternative path to hydrogen. As one researcher put it, "If we really wanted to, we could have a hydrogen economy by 2010." But the political and institutional barriers appeared formidable. Both government and industry had devoted more resources toward the gasoline- and methanol-based route than to the direct hydrogen path, though they were open to the latter: remarked one industry executive, "everyone is placing bets on several horses."

Meanwhile, hydrogen still received a tiny fraction of the research funding that was allocated to coal, oil, nuclear, and other mature, commercial energy sources. Within energy companies, the hydrocarbon side of the business continued to argue that oil will be dominant for decades to come, even as other divisions prepare for its successor. Hydrogen also posed a management challenge, that of looking beyond continuous, incremental improvement of existing products and processes, and toward more radical, "disruptive" technological change. And little had been done to educate people about the properties and safety of hydrogen, even though public acceptance, or lack thereof, would in the end make or break the hydrogen future. The societal and environmental advantages of the cleaner, more secure path to hydrogen illustrated the essential—and underrecognized—role for government in the hydrogen transition. Indeed, without aggressive energy and environmental policies, the hydrogen economy is likely to emerge along the more incremental path, and at a pace that is inadequate for dealing with the range of challenges posed by the incumbent energy system. Neither market forces nor government fiat could, in isolation, move society down the more direct, more difficult route. The challenge appears to be for government to guide the transition, setting the rules of the game and working with industry and society toward the preferable hydrogen future.

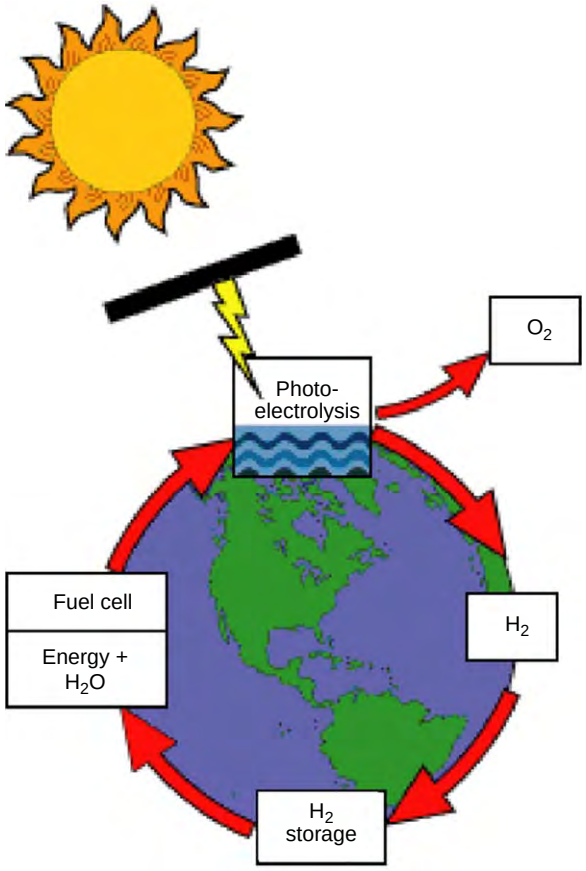


FIGURE 5 A renewable-hydrogen cycle. Reprinted from Dunn (2001), with permission.

This catalytic leadership role, some advocates have argued, would be analogous to that played by government in launching another infrastructure in the early years of the Cold War. Recognizing the strategic importance of having its networks of information more decentralized and less vulnerable to attack, the U.S. government undertook critical research, incentives, and public/private collaboration

into the development of what we now call the Internet. A similar case could be made for strategically laying the groundwork for a hydrogen energy infrastructure that best limits vulnerability to air pollution, energy insecurity, and climate change. Investments made at present would heavily influence how, and how fast, the hydrogen economy emerges in coming decades. As with the Internet, putting a man on the moon, and other great human endeavors, the dawning of the hydrogen age was unlikely to result from business as usual.

5. FUTURE OF HYDROGEN HISTORY

While its effects continue to ripple outward, the events of September 11, 2001 will become an important reference point in the history—written and unwritten—of hydrogen's development as an energy source. Interest in hydrogen surged in the aftermath of the 2001 terrorist attacks on the United States. Subsequent discussions of the appropriate military response revealed the extent to which dependence on Middle East oil had complicated U.S. foreign policy choices in that region. By chance, Shell had planned to unveil its latest long-term energy scenarios in New York City the following month. Their exposition of uncertainties, discontinuities, the "unthinkable," and the "unknowable" struck a strong chord at a moment when an unthinkable event had just happened—one that, to more business and political decision makers, made the vision of a hydrogen economy even more desirable than before.

Presenting the scenarios, Shell chairman Philip Watts presented two scenarios, "Dynamics as Usual" and "Spirit of the Coming Age," and three factors that would drive changes in the energy system: resource scarcity, new technology, and shifting social and personal priorities. Under the first scenario, there is an evolutionary progression or "carbon shift" from coal to gas to renewable energy. The second scenario involves "something more revolutionary: the potential for a truly hydrogen economy, growing out of new and exciting developments in fuel cells, advanced hydrocarbon technologies and carbon dioxide sequestration." The prominence of hydrogen in this scenario—not a prediction, but a tool for helping managers plan for different possible futures and make better strategic decisions—is one indication of the growing attention being given to hydrogen by the world's leading multinationals.

Hinting at hydrogen's ascendancy as well is *Tomorrow's Energy* by Peter Hoffmann, a journalist who had followed hydrogen and fuel cell developments since the 1970s and who edited the well-respected industry periodical *Hydrogen and Fuel Cell Letter*. His work originally began as an update of his 1981 book *The Forever Fuel*. But it soon became a new book, due to the progress in hydrogen developments that had since been made. Among the late-20th-century advances reported by Hoffmann was the latest discovery related to one of hydrogen's biggest public relations problems: the Hindenburg. The prospect of hydrogen as a fuel often raises safety questions, in part because of its association with the German airship that exploded in 1937, taking 36 lives. For years, it was widely believed that the cause of the explosion was the ignition of the hydrogen gas used for lifting the ship. Bain Addison, a retired National Aeronautics and Space Administration (NASA) scientist who doubted this explanation, carefully studied film footage and documents relating to the incident. In 1997, he publicized his surprising finding that the zeppelin's cover had been painted with iron oxides and aluminum, compounds used in rocket fuel. Addison also uncovered evidence that the Nazi government was aware of the design flaw but suppressed this information.

Beyond the Hindenburg, the safety of hydrogen can be compared with that of other fuels in use today. Hydrogen does have a wide range of limits for flammability and detonability, which means that a broad range of mixtures of hydrogen in air can lead to a flame or explosion. However, the lower limits are most relevant to transportation uses, and in this regard, hydrogen is comparable to, or better than, gasoline and natural gas. At this lower flammability limit, the ignition energy, i.e., the energy in a spark to ignite a fuel mixed in air, is about the same for hydrogen and methane. Hydrogen is also nontoxic, unlike methanol or gasoline in higher concentrations. And hydrogen is very buoyant, escaping quickly from leaks, whereas gasoline puddles, causing its fumes to build up. The prevention, detection, and management of hydrogen leaks comprise an important safety issue, requiring that areas where hydrogen is stored and dispensed be well ventilated.

Although the chemical industry routinely handles large quantities of hydrogen safely, the question is whether this safety record will be transferred to hydrogen vehicle and refueling systems. Several studies in the 1990s explored this question. A 1994 report by researchers at Sandia National Laboratories stated that "hydrogen can be handled safely, if

its unique properties—sometimes better, sometimes worse, and sometimes just different from other fuels—are respected." A 1997 study by Ford Motor Company concluded that, with proper engineering, the safety of a hydrogen fuel cell vehicle would potentially be better than that of a gasoline or propane vehicle. To ensure safe and standardized practices for using hydrogen, several national and international organizations have begun to develop codes and standards for hydrogen and fuel cells.

Another indicator of the increased attention given to hydrogen is the publication of *The Hydrogen Economy* by social critic Jeremy Rifkin. Rifkin, who has written extensively on the social and economic implications of scientific and technological trends, writes that hydrogen and the communications revolution will create a relationship that "could fundamentally reconfigure human relationships in the 21st and 22nd centuries. Since hydrogen is everywhere and is inexhaustible if properly harnessed, every human being on Earth could be 'empowered,' making hydrogen energy the first truly democratic energy regime in history." The resulting hydrogen energy web "will be the next great technological, commercial, and social revolution in history." Rifkin's utopian vision has engendered some justifiable skepticism from the corporate world. A review in *Business Week* opined that "the hydrogen hype…may be too much, too soon. While hydrogen's abundance and cleanliness make it the favorite as the fuel of the future, fossil fuels still have a cost advantage. And it's not obvious that even when hydrogen does come into widespread use, democracy will follow." The article argues that sharp cost declines in the cost of renewable energy, not a shortage of oil, will be "the factor that forces the transition to a hydrogen economy.…The hydrogen economy will dawn when systems based on renewables and hydrogen beat hydrocarbons in straight competition for consumers' dollars."

Indeed, hydrogen still faces considerable cost hurdles at all stages of the energy system, from production to transportation to storage to use. In addition to major advances in renewable energy, improvements in storage technologies such as carbon nanotubes and metal hydrides, development of hydrogen-compatible pipelines, and further strides with fuel cells are among the prerequisites to making hydrogen competitive with hydrocarbons. The most economic method of producing hydrogen at present, via the reformation of natural gas, has been mostly limited to feedstock production; extracting hydrogen from electrolysis from solar or wind power costs three to five times more.

At the same time, and as recognized to varying degrees by business and government, renewables and hydrogen do not face a level playing field today, making concerted public action essential to the hydrogen transition. Japan had set the pace in the global race to hydrogen with its 1993 announcement of a 10-year, $2 billion commitment to promote hydrogen internationally through its World Energy Network (WE-NET) Project. But the stakes had been raised by 2002. The European Commission (EC) unveiled a $2 billion, 4-year investment in hydrogen technology research. In announcing the initiative, EC President Romano Prodi argued that the program would be as important to Europe as the space program was for the United States in the 1960s. "It's like going to the moon in a series of steps.…We expect an [even] better technological fallout."

The U.S. government, often criticized for supporting hydrogen less aggressively compared to its European and Japanese counterparts, also moved forward in 2002, announcing a "Freedom Car" partnership with the "big three" automakers to develop fuel cell vehicles. The National Academy of Sciences commissioned a panel study to examine the technical potential of transitioning to a hydrogen economy. Soon after, the Department of Energy unveiled an ambitious hydrogen road map, a blueprint crafted by public and private experts to coordinate the long-term development of a hydrogen economy. In an introduction to the report, Secretary Spencer Abraham sounded almost reminiscent of Jules Verne: "To talk about the 'hydrogen economy' is to talk about a world that is fundamentally different from the one we now know. A hydrogen economy will mean a world where our pollution problems are solved and where our need for abundant and affordable energy is secure…and where concerns about dwindling resources are a thing of the past."

It has taken more than a century, but the idea of hydrogen as an energy source has moved off the pages of science fiction and into the speeches of business and political leaders. In other words, hydrogen energy has evolved from a matter of scientific and technological curiosity to a major strategic issue for political and corporate decision makers. Whether it continues to sustain such interest will determine whether a hydrogen economy truly emerges or whether hydrogen remains, as enthusiasts have called it for several decades, "tomorrow's energy" or "the fuel of the future."

The history of hydrogen as an energy source is in many respects only beginning, with its future path

and direction full of many possibilities. However it unfolds, the story of hydrogen in the 21st century promises to be as exciting as *The Prize*, Daniel Yergin's epic about oil in the 20th century. Whether, how, and how fast the hydrogen economy is realized hold enormous consequences for humanity's social, economic, political, and environmental future. That, in turn, will hinge on whether hydrogen becomes not only an interesting possibility, but also a compelling imperative. The hydrogen endeavor is a truly an historic one, bringing to mind President John F. Kennedy's impassioned appeal for the Apollo space program: "There are risks and costs to a program of action, but they are far less than the long-range risks and costs of comfortable inaction."

SEE ALSO THE FOLLOWING ARTICLES

Coal Industry, History of • Combustion and Thermochemistry • Electricity Use, History of • Hydrogen, End Uses and Economics • Hydrogen Production • Hydrogen Storage and Transportation • Manufactured Gas, History of • Natural Gas, History of • Nuclear Power, History of • Oil Industry, History of

Further Reading

Dunn, S. (2000). The hydrogen experiment. *World Watch* **13**(6), 14–25.

Dunn, S. (2001). "Hydrogen Futures: Toward a Sustainable Energy System." Worldwatch Pap. 157. Worldwatch Institute, Washington, D.C.

Hoffmann, P. (2001). "Tomorrow's Energy: Hydrogen, Fuel Cells, and the Prospects for a Cleaner Planet." MIT Press, Cambridge, Massachusetts.

Jensen, M. W., and Ross, M. (1999). The ultimate challenge: Building an infrastructure for fuel cell vehicles. *Environment* **42**(7), 10–22.

Koppel, T. (1999). "Powering the Future: The Ballard Fuel Cell and the Race to Change the World." John Wiley & Sons, New York.

McNeill, J. R. (2000). "Something New Under the Sun: An Environmental History of the Twentieth-Century World." W. W. Norton & Co., New York.

Ogden, J. M. (1999). Prospects for building a hydrogen infrastructure. *Annu. Rev. Energy Environ.* **24,** 227–279.

Rifkin, J. (2002). "The Hydrogen Economy." Penguin Putnam, New York.

Royal Dutch/Shell Group. (2001). "Energy Needs, Choices and Possibilities: Scenarios to 2050." Royal Dutch/Shell, Amsterdam.

Veziroglu, T. N. (2000). Quarter century of hydrogen movement, 1974–2000. *Int. J. Hydrogen Energy* **25,** 1143–1150.

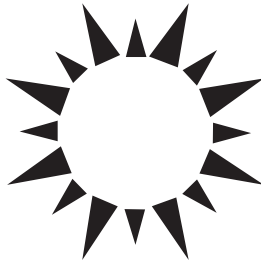
U.N. Development Programme, U.N. Department of Economic and Social Affairs, and World Energy Council. (2000). "World Energy Assessment Report." United Nations, New York.

U.S. Department of Energy. (2002). "National Hydrogen Energy Roadmap." Department of Energy, Washington, D.C.

Yergin, D. (1991). "The Prize: The Epic Quest for Oil, Money, and Power." Simon & Schuster, New York.

Hydrogen Production

GENE D. BERRY
Lawrence Livermore National Laboratory
Livermore, California, United States



1. **Introduction**
2. **Hydrogen Production Using Fossil Energy**
3. **Thermodynamics of Water Splitting**
4. **Multiple-Step Thermochemical Cycles for H_2O Decomposition**
5. **Electrolysis of Water and Steam**
6. **Overall Prospects for Hydrogen Production**

Glossary

Carnot efficiency The maximum efficiency with which thermodynamic work can be produced from thermal energy flowing across a temperature difference; the Carnot efficiency is given by $\eta_{Carnot} = 1 - \frac{T_{low}}{T_{high}}$.

Gibbs free energy The minimum thermodynamic work (at constant pressure) to drive a chemical reaction (or, if negative, the maximum work that can be done by the reaction); the standard Gibbs free energy necessary to decompose H_2O into H_2 and O_2 is 237 kJ/mol. The Gibbs free energy decreases with temperature, but increases with the pressure of hydrogen (and oxygen) produced.

lower heating value (LHV) Thermal energy released when a fuel is burned with air (oxygen) at ambient temperature and pressure, assuming that water in the exhaust remains as vapor, where 1 kg of hydrogen has an LHV of 120 MJ, nearly the same as a gallon (3.785 L) of gasoline; the higher heating value (HHV) of hydrogen is 142 MJ/kg, the thermal energy released when hydrogen reacts with air (oxygen) to produce liquid water exhaust.

Nernst equation The thermodynamic minimum voltage (i.e., in equilibrium) of an electrochemical reaction; it is given by the Gibbs free energy per mole of electrons involved in the reaction divided by Faraday's constant (F), defined as 96,500 Coulombs of charge per mole of electrons.

overvoltage The additional voltage, in excess of the Nernst voltage, required to overcome electrochemical reaction kinetics and resistance to the flow of ions and electrons necessary to drive electrochemical reactions at finite reaction rates; water electrolysis has a Nernst voltage of 1.23 V under ambient conditions.

steam reforming The dominant method of commercial hydrogen production; it is based on reacting methane with water to produce hydrogen and carbon dioxide according to the (endothermic) reaction: $CH_4 + 2H_2O \rightarrow 4H_2 + CO_2$.

thermochemical (hydrogen) cycle A multiple-step chemical reaction cycle that sums to the overall production of hydrogen (and oxygen) by water decomposition; the Carnot efficiency of the highest temperature step places a theoretical limit on the overall hydrogen production efficiency of such a cycle.

Composed of just one proton and an accompanying electron, and accounting for 90% of all matter in the universe, hydrogen (H_2) is by far the simplest, most abundant, and lightest element. H_2 offers distinct advantages as an alternative fuel. Its extreme buoyancy allows H_2 to disperse rapidly when accidentally released or spilled, ultimately escaping the earth's atmosphere or returning as water (H_2O) when combined with atmospheric oxygen (O_2). H_2 can be produced from H_2O using virtually any energy source(s), even electricity, allowing strategic economic and environmental flexibility, especially if generated by electrolysis. H_2 is ecologically ideal because it is carbon free, forming H_2O when burned (or otherwise reacted with O_2) to release chemical energy. Among alternative fuels, only H_2 is likely to be light enough for aircraft and, therefore, the only alternative transportation fuel capable of being universal.

1. INTRODUCTION

The widespread use of hydrogen (H_2) faces three fundamental obstacles. First, as the lightest known molecule, H_2 possesses extreme physical properties that make compressing, liquefying, and storing H_2 costly and energy intensive. Second, high-pressure

 © 2004 Elsevier Inc. All rights reserved.

gaseous H_2, or even liquid hydrogen (LH_2), has a very low energy density relative to polyatomic hydrocarbon fuels (e.g., gasoline [$\sim C_8H_{18}$], natural gas [CH_4], ethanol [CH_3CH_2OH], methanol [CH_3OH]), whose molecules contain more hydrogen atoms than does H_2. Finally, like electricity, H_2 is an energy carrier that must be generated from high-value nuclear, solar, or fossil energy (almost certainly by decomposing water [H_2O]). H_2 consequently entails an energetic and economic premium proportional to the energy consumed in H_2 production and storage. This premium will force H_2 to be more costly than fossil fuels (per unit energy) until nonfossil energy sources cost significantly less than fossil fuels.

Higher fuel economy can compensate for the high cost of H_2 in applications where heavier fossil fuels are at a disadvantage (e.g., spacecraft, aircraft). H_2 can also be used with the greatest efficiency of any fuel (1) in conventional combustion engines, (2) by flameless oxidation on catalytic surfaces, or (3) by electrochemical oxidation to directly generate electricity in fuel cells. Future high efficiency ($\sim$80 mpg, 34 km/L equivalent) H_2-fueled hybrid electric or fuel cell automobiles will travel 400 miles (640 km) on less than 5 kg of H_2 fuel (energy equivalent to 5 gallons of gasoline).

A full-fledged H_2 economy will likely require not just high fuel efficiency but also full realization of (and capitalization on) the unique benefits of H_2 as an alternative fuel. These include enduring ecological advantages when made from carbonless energy; a broad spectrum of H_2 production energy sources, methods, scales and related storage options and distribution infrastructures; and universal utility throughout transportation (land, sea, and air), energy storage, and industrial sectors.

H_2 fuel can be generated by decomposing H_2O using thermal, electrolytic, or chemical methods, perhaps in combination (Fig. 1). The various H_2 production approaches are characterized by extreme contrasts in technological maturity, energetics, ecological potential, and their implications for feasibility on a global scale.

It is the author's belief that the worldwide replacement of fossil fuels by H_2 will be driven chiefly by the potential scope, efficiency, and energetic costs of H_2 production technologies as they approach the thermodynamic limits of H_2O decomposition. Consequently, this article does not consider H_2 production methods that are ultimately inefficient pathways for harnessing solar energy or are likely to have limited applicability (e.g., biomass gasification; biological, photobiological, and photochemical production). Instead, this article surveys historical and current H_2 production using fossil energy, the thermodynamics of producing H_2 exclusively from H_2O, and the two chief technological options for carbonless H_2 production on a global scale: multiple-step thermochemical cycles and electrolysis.

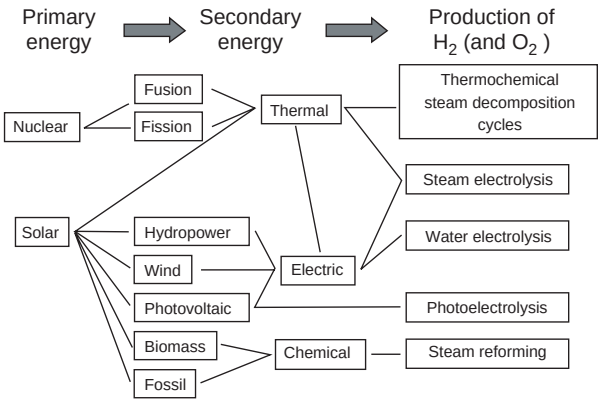


FIGURE 1 Flowchart of hydrogen production pathways and methods. Hydrogen is not an energy source but rather a carrier of energy from nuclear or solar primary energy sources (or indirectly through fossil, biomass, wind, and/or hydroelectric pathways). Energy can be stored in hydrogen by decomposing water using chemical, thermal, and/or electrical energy by steam reforming, electrolysis, photoelectrolysis, or thermochemical cycles.

2. HYDROGEN PRODUCTION USING FOSSIL ENERGY

Following Lavoisier's identification of hydrogen as an element of water in 1783, large amounts of H_2 gas were first produced in 1794 by reacting H_2O with iron (Fe) for suggested use in military observation balloons in France. H_2 was produced as a heating fuel during the late 19th century by reacting coal (carbon) with H_2O to generate mixtures of H_2 and carbon monoxide (CO)—known as town gas, blue gas, or blue water gas. Pipelined town gas is still used in parts of Europe, but these gases were ultimately displaced by natural gas (chiefly CH_4) in the United States following World War II, although the oil shocks of the 1970s revived interest in coal gasification to generate H_2 as a precursor to synthetic natural gas (SNG).

Natural gas prices have since fallen and economics now favor the reverse, that is, producing H_2 from natural gas (CH_4) by the process known as

steam-methane reforming, essentially harvesting H_2 from H_2O using natural gas by the following overall (endothermic) reaction:

$$2H_2O(l)+CH_4(g)+245kJ \rightarrow CO_2(g)+4H_2(g). \quad (1)$$

Steam methane reforming can be 70% efficient based on the lower heating value (LHV) of hydrogen (120 MJ or 33.3 kWh/kg), with substantially lower capital investment than other H_2 production methods. Purge gases, diverted feed methane, and even nonfossil sources (e.g., solar, nuclear) can provide the additional thermal energy to drive steam reforming at peak temperatures of approximately 1100 K. The product H_2 gas is typically separated by pressure swing adsorption (PSA) relying on the weak adsorption properties of H_2 relative to other larger molecules. CO is the principal contaminant in H_2 from this process. The vast majority of commercial H_2 is currently produced by steam reforming of natural gas at or near the point of use. This H_2 is used in refinery operations (upgrading crude oil) or chemical manufacture (e.g., NH_3, CH_3OH), with only approximately 3% transported as compressed gas or cryogenic liquid (LH_2) in the "merchant" H_2 market.

Unfortunately, steam reforming, or other fossil-fueled H_2 production methods, generate carbon dioxide (CO_2) and are generally not environmentally advantageous. The (high) energy intensity of producing and storing H_2 can actually increase net energy use, attendant pollution, and greenhouse gas emissions associated with "fossil" H_2 relative to direct use of CH_4 or other fossil fuels.

In general, the (high) energy intensity of H_2 production and storage can be justified in applications where H_2 can be used with far greater fuel economy than can hydrocarbon fuels and electricity is not a viable substitute (e.g., supersonic intercontinental aircraft). More recently, the potential ease of CO_2 sequestration has been advanced as a justification for fossil-fueled H_2 production because CO_2 emissions would generally be available to sequester without diluent nitrogen (N_2) from the atmosphere. CO_2 sequestration will be limited to large centralized H_2 production facilities, implying a substantial H_2 pipeline infrastructure and/or energy-intensive cryogenic LH_2 distribution. Sequestration is also likely to be relevant only for H_2 production using coal given that conventional natural gas and petroleum are projected to be minority contributors to global CO_2 emissions over the 21st century.

3. THERMODYNAMICS OF WATER SPLITTING

The desire to use nonfossil energy to efficiently produce H_2 underlies interest in decomposing H_2O into its elements: H_2 and O_2. So-called water-splitting approaches are broadly divisible into thermochemical and electrochemical methods, which can differ dramatically in terms of primary energy sources, scale, intermittent operability, and associated infrastructure implications. However, both approaches fundamentally share the same thermodynamic problem, that is, decomposing H_2O.

H_2 is an energetic and attractive fuel due to an extreme chemical affinity for oxygen (O_2). The earth's atmosphere is so O_2 rich that in thermodynamic equilibrium, even a single H_2 molecule would not remain free within it. The hydrogen–oxygen bond in H_2O is so strong that 1000 cubic meters (m^3) of pure steam (1 atm) in thermodynamic equilibrium will contain just 10 dissociated molecules on average.

Producing H_2 fuel from H_2O requires substantial energy to oppose these thermodynamic forces. The theoretical minimum energy input (work) is given by

$$\text{Minimum Work} = \Delta G_{std}(T) + RT\ln\left(\frac{[H_2][O_2]}{[H_2O]}\right)^{1/2} \quad (2)$$

where ΔG_{std} is the standard Gibbs free energy of water decomposition, R is the universal gas constant (8.314 J/mol Kelvin), T is absolute temperature, and $[H_2]$, $[O_2]$, and $[H_2O]$ are the partial pressures (in atmospheres) of hydrogen, oxygen, and water, respectively.

The minimum work to decompose saturated water vapor in equilibrium with liquid water and atmospheric air (0.2 bar O_2, 300 K) is 235 kJ/mol H_2O. Thermodynamic tables presume a higher O_2 concentration (1 bar, 300 K), raising the standard minimum work of H_2O decomposition to 237 kJ/mol, equivalent to 1.229 V, the well-known theoretical minimum voltage for water electrolysis.

Furthermore, H_2O decomposition generates entropy (i.e., positive ΔS) by creating three molecules (two H_2 and one O_2) from every two H_2O molecules and, therefore, is endothermic (i.e., absorbs heat), making H_2O decomposition thermodynamically favored by higher temperatures. This is reflected in the reduced work (ΔG) requirement but larger thermal energy ($T\Delta S$) demand at higher temperatures (Fig. 2). Overall, the thermodynamic requirements

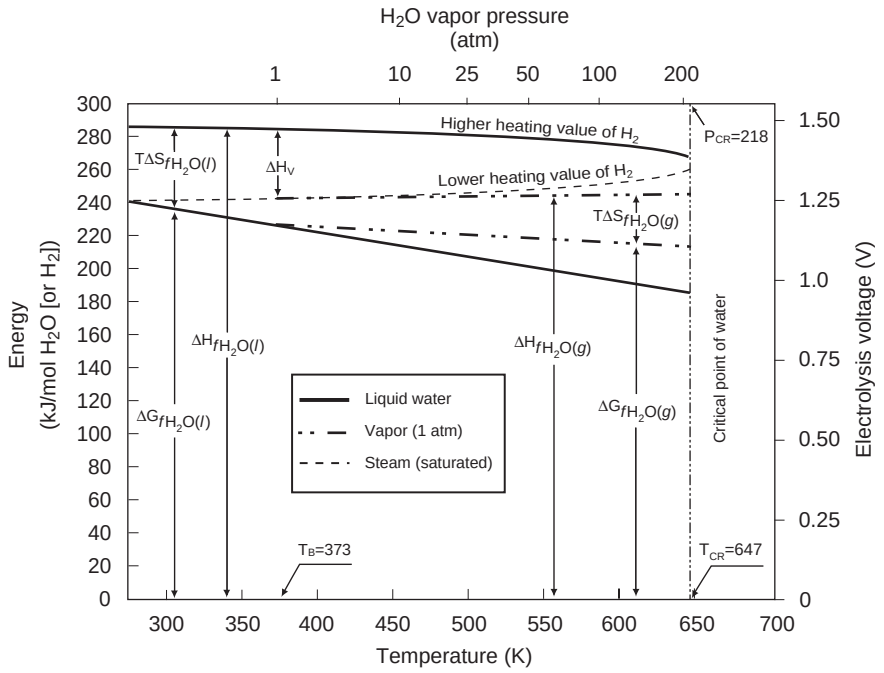


FIGURE 2 Theoretical energetics of water decomposition to produce hydrogen. Generating 1 atm of hydrogen (and oxygen) from liquid water (thick lines) requires a total energy input corresponding to the enthalpy of formation of water $(\Delta H_{f\mathrm{H_2O}(l)})$ equivalent to the higher heating value of hydrogen (286 kJ/mol). This energy can be supplied as work or as a combination of work $(\Delta G_{f\mathrm{H_2O}(l)})$ and thermal energy $(T\Delta S_{f\mathrm{H_2O}(l)})$. The total energy to electrolyze steam at 1 atm (dash double dot) $(\Delta H_{f\mathrm{H_2O}(g)})$ is considerably lower due to the thermal energy absorbed (ΔH_v) by the vaporization of water, representing the difference between the higher and lower heating values of hydrogen. This difference approaches zero as the thermodynamic properties of saturated steam (dashed line) and liquid water converge at the critical point (647 K, 218 atm). The work required to produce hydrogen from 1 atm steam $(\Delta G_{f\mathrm{H_2O}(g)})$ decreases with temperature due to a corresponding increase in absorbed thermal energy $(T\Delta S_{f\mathrm{H_2O}(g)})$. The work required to produce hydrogen from high-pressure (top-axis) liquid water $(\Delta G_{f\mathrm{H_2O}(l)})$ decreases more sharply with temperature, but the thermal energy requirement $(T\Delta S_{f\mathrm{H_2O}(l)})$ is higher as well. The right-hand axis indicates the energy equivalent voltage to generate hydrogen by electrolysis of water or steam.

for H_2 production from liquid H_2O are minimums of heat and work that sum to the decomposition enthalpy (ΔH) of 286 kJ/mol H_2O, equivalent to the thermoneutral electrolysis voltage of 1.48 V and the corresponding energy content of H_2 fuel as defined by its higher heating value (HHV).

4. MULTIPLE-STEP THERMOCHEMICAL CYCLES FOR H_2O DECOMPOSITION

At sufficiently high temperatures (> 3000 K) the thermodynamic work of steam decomposition (ΔG) will shrink to zero and H_2O will decompose spontaneously, theoretically permitting H_2 production using only thermal energy. The conceptual simplicity of thermal decomposition is potentially attractive because it circumvents the intermediate step of electricity generation prerequisite for H_2 production by water electrolysis. However, steam decomposes only slightly at temperatures up to 1200 K, forming approximately 5 ppm H_2 (and 2.5 ppm O_2). Steam decomposition increases exponentially at higher temperatures, reaching 4% H_2 at 2500 K (1 atm), but at these temperatures H_2 and O_2 molecules themselves begin to decompose (Fig. 3), forming hydroxyl (OH) molecules and even individual atoms (O and H), by reactions that generate even greater entropy than does steam decomposition.

This presents fundamental difficulties for direct thermal production of H_2 at very high temperatures (T > 2500 K). Not only are H_2 and O_2 molecules unstable at these temperatures, but most other materials are as well. The related practical issues of long-term material stability, bulk diffusion, reaction with very hot gases or monatomic species, and separation of product gases are generally thought to preclude H_2 production by single-step steam decomposition.

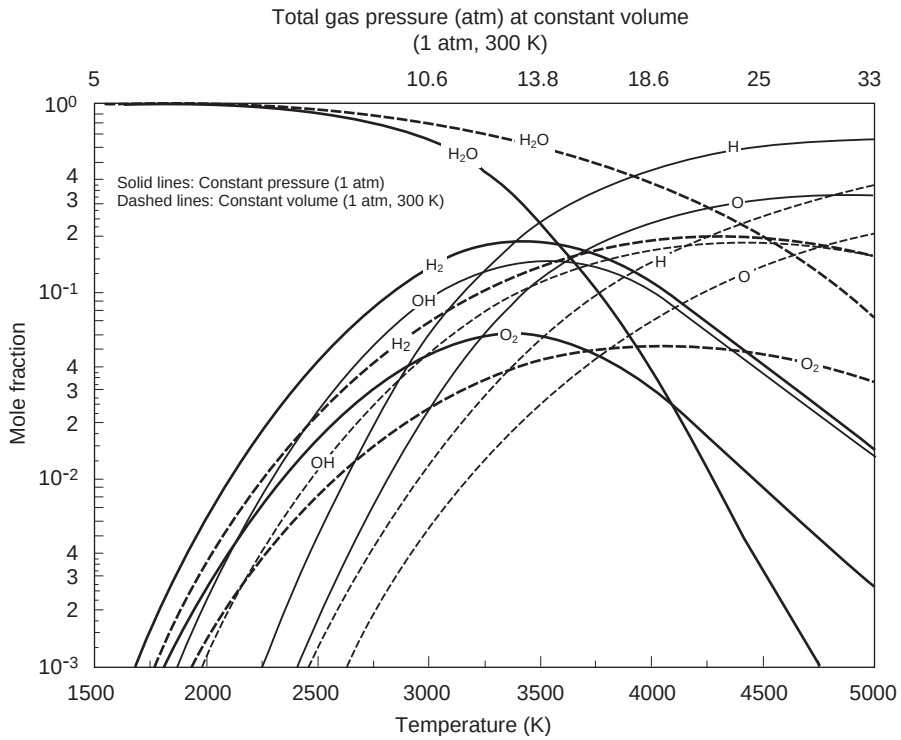


FIGURE 3 Equilibrium thermodynamics of water vapor decomposition. Water vapor decomposes when heated under conditions of constant pressure (solid lines) or volume (dashed lines). Hydrogen and oxygen are the major decomposition products at low temperatures, where less than 1% of water vapor decomposes. At higher temperatures, a greater fraction of water decomposes, but the thermodynamic equilibrium becomes complex and hydroxyl (OH) molecules are also generated in significant amounts. Above the very high temperatures ($\sim$2500 K) necessary for significant H_2O decomposition, H_2 and O_2 molecules themselves begin to decompose into individual atoms, complicating direct water decomposition schemes. The upper axis indicates the total gas pressure for the constant volume case. At constant volume, pressure increases as the temperature and number of molecules increase. The initial condition is 1 atm, 300 K.

First explored by Funk in 1966, interest in producing H_2 by multiple-step thermochemical H_2O decomposition cycles skyrocketed during the oil shocks of the 1970s. In theory, reduced reaction temperatures ($T < 1200$ K) compatible with advanced nuclear heat sources, or perhaps concentrated solar energy, can be achieved by dividing H_2O decomposition into multiple steps chosen for optimal properties and operated under varied conditions but that, overall, still sum to H_2O decomposition.

The simplest illustrative example is a notional two-step thermochemical decomposition cycle in which a conceptual oxide material AO absorbs heat, decomposing at a high temperature (e.g., 1173 K), releasing O_2, and later reacts with H_2O to liberate H_2, releasing heat near room temperature (373 K):

$$\text{Heat} + 2\text{AO} \rightarrow 2\text{A} + \text{O}_2 \ (1173\ \text{K}) \tag{3}$$

$$2\text{H}_2\text{O} + 2\text{A} \rightarrow 2\text{AO} + 2\text{H}_2 + \text{Heat} \ (373\ \text{K}). \tag{4}$$

Because both reactions sum to steam decomposition, the thermodynamic properties (i.e., entropy and enthalpy) of the individual reactions also sum to the properties of steam decomposition:

$$\Delta S_1 + \Delta S_2 = \Delta S_{H_2O} \tag{5}$$

$$\Delta H_1 + \Delta H_2 = \Delta H_{H_2O}. \tag{6}$$

These relationships, in combination with the temperature difference ($T_{high} - T_{low}$) between reaction steps, enable calculation of the theoretical thermodynamic requirements (i.e., $\Delta G = 0$ for each reaction) sufficient for both to be driven by heat alone:

$$\Delta G_1(T_{high}) = 0 = \Delta G_2(T_{low}) \tag{7}$$

$$\Delta H_{high} = T_{high}\Delta S_{high} \tag{8}$$

$$\Delta S_{high} = \frac{[\Delta G_{H_2O}(T_{low})]}{[T_{high} - T_{low}]}. \tag{9}$$

Assuming that the second reaction step is run at 373 K, the theoretical minimum reaction entropy needed for a high temperature step at 1373 K is 225 J/mol $H_2 \cdot$ K, a value that is five times as high as steam

decomposition itself (45 J/mol $H_2 \cdot K$). This must be compensated by a low-temperature step with a correspondingly large negative reaction entropy (-180 J/mol $H_2 \cdot K$). Unfortunately, it is doubtful whether chemical reactions based on practical (i.e., simple) substances that generate such large changes in entropy can be found. Candidate reactions for a theoretical two-step thermochemical cycle with a peak temperature below 1000 K have not been proposed to date. Instead, cycles with additional reaction steps have been examined.

Dividing steam decomposition into more than two steps relaxes the thermodynamic requirements on each reaction step, permitting lower peak reaction temperatures. Temperature limitations on nuclear heat sources stimulated theoretical consideration of more than 100 thermochemical cycles using three, four, and five reaction steps during the 1970s, but interest waned during the 1980s, with work carried out only in Japan. Multiple-step reaction cycles complicate thermochemical water splitting substantially, reducing practical efficiencies and increasing equipment cost. Recent work has narrowed to practical consideration of two quite different thermochemical reaction cycles envisioned for use with high-temperature ($\sim$1200 K) nuclear reactors, which can be considered prototypical: the sulfur–iodine (SI) and UT-3 cycles.

The SI cycle is a three-step thermochemical water decomposition cycle conducted entirely during the fluid phase (Fig. 4). The SI cycle relies principally on the high-temperature (and high-entropy) decomposition of the complex molecule hydrogen sulfate (H_2SO_4) into many simpler ones (H_2O, SO_2, O_2), from which O_2 is removed. A second lower temperature reaction step regenerates H_2SO_4 by adding H_2O and captures the liberated hydrogen as hydrogen iodide (HI), which is then decomposed in a third step, producing H_2 at pressures of up to 20 atm. The iodine (I_2) is recycled for use in the second reaction, closing the thermochemical cycle conceptually expressed by the following reaction sequence:

$$H_2SO_4 \rightarrow SO_2 + H_2O + 1/2\ O_2 \quad (1100\ K) \qquad (10)$$

$$SO_2 + 2H_2O + I_2 \rightarrow 2HI + H_2SO_4 \quad (400\ K) \qquad (11)$$

$$2HI \rightarrow I_2 + H_2 \quad (625\ K). \qquad (12)$$

Development of the SI cycle was pursued at General Atomics in San Diego during the 1970s, with bench scale experiments predicting H_2 production efficiencies as high as 44% (LHV). Potential improvement of the second reaction step appears possible by avoiding separation of HI from excess I_2. Interest in the cycle has recently been renewed at General Atomics.

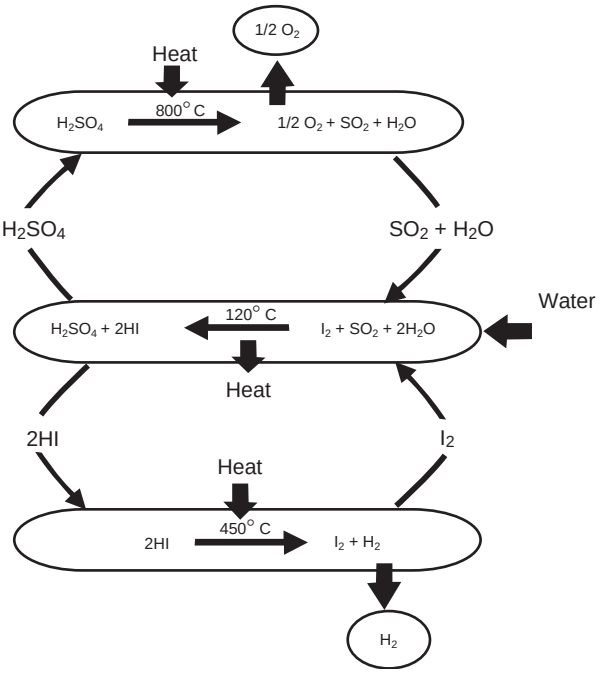


FIGURE 4 The sulfur–iodine cycle. The vertical schematic shows the leading thermochemical hydrogen production cycle using sulfuric acid and iodine. Sulfuric acid is decomposed in a high-temperature step, rejecting liberated oxygen. A second step in the cycle regenerates this sulfuric acid through the addition of water. The hydrogen from this additional water is removed as hydrogen iodide, which is decomposed in a third step liberating iodine for reuse and producing hydrogen for removal from the cycle. The thermal efficiency of this cycle is thought to be capable of approaching 50% on a higher heating value of hydrogen (142 MJ/kg) basis.

The high-temperature, gas-cooled nuclear reactors (HTGR) envisioned to power the SI (or other) thermochemical cycle(s) have been commercially demonstrated, producing electricity up to scales of 330 MWe. These reactors are projected to be capable of 48% efficiency, with capital costs of $1100/kWe and operating costs of $0.015/kWh for fuel, maintenance, and the like, resulting in busbar electricity costs of approximately $0.03/kWh.

Reactors designed to generate H_2 would rely on helium (He) coolant, transporting heat to the thermochemical cycle at approximately 1150 K, requiring additional capital investment for the thermochemical hydrogen plant (even with a credit for elimination of the electrical generation equipment). Assuming that 75 kWh of nuclear heat could produce 1 kg of H_2 (44% LHV efficiency) and that annual operations and plant maintenance costs

average 7% of capital investment ($315/kW$_{thermal}$), HTGR SI plants are projected to be capable of producing H_2 for approximately $1.50/kg, equivalent to $0.045/kWh$_{thermal}$ or $12.50/GJ H_2 (LHV). These costs would be lower than those of other nonfossil H_2 production approaches and would be competitive with costs of steam reforming if natural gas prices rise in the future.

An alternative H_2O decomposition cycle, known as the UT-3 cycle, has been under development for decades at the University of Tokyo. This cycle differs substantially from the General Atomics SI cycle in that it relies entirely on gas–solid reactions, with batched forward and reverse reactions occurring in stationary beds instead of recycling fluid reactants and products. The UT-3 cycle can be described thermodynamically by the following sequence of four reactions:

$$CaBr_2(s)+H_2O(g)\rightarrow CaO(s)+2HBr(g)\ \ (1030\,K) \tag{13}$$

$$3FeBr_2(s)+4H_2O(g)\rightarrow Fe_3O_4(s)+6HBr(g)+H_2(g)\ \ (835\,K) \tag{14}$$

$$Fe_3O_4(s)+8HBr(g)\rightarrow Br_2(g)+3FeBr_2(s)+4H_2O(g)\ \ (490\,K) \tag{15}$$

$$Br_2(g)+CaO(s)\rightarrow CaBr_2(s)+1/2\ O_2(g)\ \ (945\,K). \tag{16}$$

The chemistry of this cycle has been tested on the pilot plant scale. Operation consists of two sets of identical stationary beds. Two beds are calcium (Ca), cycled between the oxide (CaO) and bromide ($CaBr_2$). The other two beds cycle iron (Fe) between the oxide (Fe_3O_4) and bromide ($FeBr_2$). Hydrogen bromide (HBr) and H_2O are circulated throughout the beds, and subatmospheric oxygen and hydrogen are extracted from this vapor stream by membranes.

The first bed decomposes input H_2O (steam) over $CaBr_2$ at the peak reaction temperature of approximately 1030 K, yielding CaO and HBr. Similarly, the $FeBr_2$ bed is oxidized by H_2O, forming Fe_3O_4 and H_2/H_2O/HBr mixture, from which H_2 is extracted. The remaining H_2O and HBr regenerate both the $FeBr_2$ from Fe_3O_4 and the $CaBr_2$ from CaO, ultimately producing an O_2/H_2O mixture from which O_2 is extracted and additional H_2O is added to repeat the cycle.

The UT-3 cycle has a projected H_2 production efficiency of 34% (LHV), producing H_2 at subatmospheric pressures. The non-steady-state nature of the cycle could potentially reduce efficiency in long-term operation due to attrition of solids in the bed.

H_2 production by thermochemical cycles faces fundamental challenges beyond the details of any given cycle. To achieve commercial success, every step of the cycle must be accomplished with virtually perfect material recycling to avoid pollution and/or raw material cost issues given the enormous quantities of H_2 (billions of tonnes [metric tons] annually) needed in a mature hydrogen economy. The development of much simpler two-step cycles would facilitate material recycling greatly but necessarily require higher temperatures ($T > 1000\,K$) and the new development of high-temperature heat sources.

Higher temperatures would also raise the theoretical limit on thermochemical H_2 production given by the Carnot efficiency:

$$\eta_{thermal} = \left(1 - \frac{T_{low}}{T_{high}}\right) \cdot \frac{\Delta H_f}{\Delta G_f}, \tag{17}$$

where ΔH_f and ΔG_f are the enthalpy and Gibbs free energy of formation for liquid water under standard conditions, a ratio of 1.20.

5. ELECTROLYSIS OF WATER AND STEAM

Electrolysis is a conceptually simple and elegant method of decomposing water (and other exceptionally stable chemical compounds). Electrolysis was discovered by Nicholson and Carlisle shortly after the invention of the voltaic cell by Alessandro Volta in 1800. Early 19th-century scientists studied electrolysis for decades in advance of knowledge of the chemical elements, knowledge of thermodynamics, or the advent of the internal combustion engine. The ecological and global implications of producing H_2 by industrial scale electrolysis of water were first noted by Jules Verne in his 1874 novel *The Mysterious Island*. Following scientific groundwork by Faraday, Arrhenius, Gibbs, and Nernst, the first commercial electrolyzer was built in 1902.

The fundamental unit of an electrolyzer is an electrochemical cell consisting of three essential elements: an anode, a cathode, and an intervening electrolyte. Electrons flow from the anode (where oxidation occurs) to the cathode (where reduction occurs). The intervening electrolyte conducts ions between the anode and cathode (electrodes), maintaining overall charge neutrality of an electrochemical reaction. Water electrolysis technology is easily

classified into three approaches, with each approach conducting one of three ionic species across the electrolyte: liquid alkaline electrolytes, which transport hydroxide (OH^-) ions; proton exchange membranes, typically polymers that transport protons (H^+); and ceramic solid oxide membranes, which conduct oxygen (O^{2-}) ions (Fig. 5).

In electrolysis, the current is directly proportional to the H_2 production rate, which is accelerated by the voltage applied in excess of the thermodynamic minimum. Well-designed electrolyzers avoid parasitic side reactions and achieve Faradaic (current) efficiencies of virtually 100%. Consequently, the electricity consumption (and efficiency) of an electrolysis cell is a function only of the applied voltage necessary to achieve a satisfactory H_2 production rate (e.g., an electrolysis current density of 0.1–1.0 amp/cm^2).

Decomposing water requires a minimum voltage determined by the chemical potential of gases (or liquids) at each electrode, and these variables, in turn, are functions of temperature and relative gas concentrations. These variables influence the overall Gibbs free energy difference (ΔG) between products and reactants. When this is divided by the moles of electrons (n) transferred per mole of H_2 generated and Faraday's constant (F), it gives the electrolysis voltage (E) using the Nernst equation:

$$E = \frac{\Delta G}{nF}. \qquad (18)$$

Under 1 atm each of H_2 and O_2, the minimum work (voltage) necessary to decompose water at 300 K is 237 kJ/mol H_2O or 1.23 V, corresponding to 32.6 kWh/kg of H_2 produced. If liquid, H_2O absorbs substantial heat ($\sim$7 kWh/kg H_2) from the surroundings when electrolyzed into H_2 and O_2.

Thermodynamic irreversibilities (resistance to ion flow through the electrolyte, electron flow through electrodes and wires, and additional overvoltage necessary to accelerate electrode reactions) currently generate waste heat greater than 7 kWh/kg H_2, making water electrolysis exothermic in today's commercial electrolyzers. The transition from endothermic to exothermic occurs (Fig. 6) when the electrolysis voltage exceeds 1.48 V, the thermoneutral voltage corresponding to the HHV of hydrogen (142 MJ/kg H_2) or, equivalently, the enthalpy of water decomposition (286 kJ/mol H_2O). Water electrolysis efficiencies have been customarily quoted based on this thermoneutral (enthalpic) voltage (1.48 V) rather than on the lower Gibbs free energy

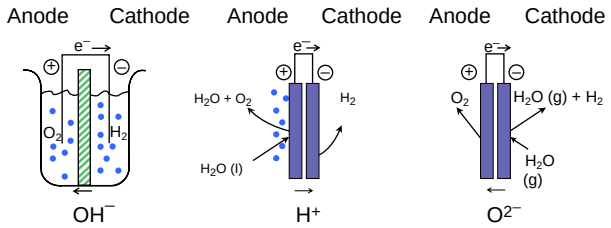


FIGURE 5 The three classes of electrolysis technology. The three classes of electrolysis technology are most easily differentiated by electrolyte. Electrons in a conventional alkaline electrolysis cell (left) generate hydrogen and hydroxide ions that diffuse through an aqueous potassium hydroxide (KOH) electrolyte, where they are converted to oxygen and water at the anode, giving up electrons to the external circuit. The more recently developed proton exchange membrane (PEM) (middle) replaces the corrosive KOH electrolyte with a polymeric membrane that, when hydrated, conducts protons (H^+) from the anode to the cathode, where they combine with electrons to generate hydrogen gas. Steam electrolysis (right) relies on solid oxide ceramics (usually heavily doped ZrO_2) to conduct oxygen ions extracted from steam at the cathode and release them as oxygen gas at the anode, returning electrons to the external circuit.

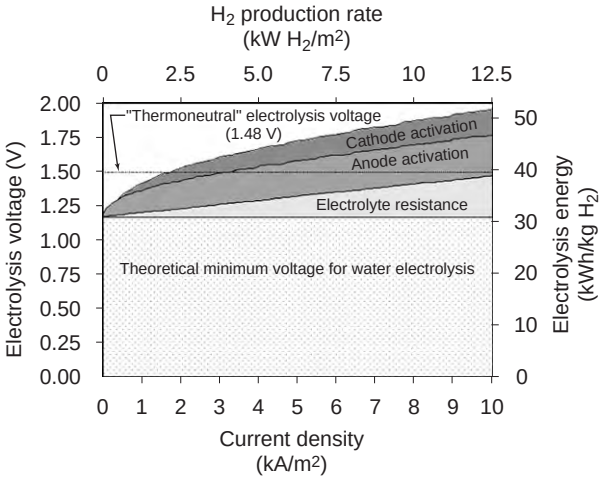


FIGURE 6 Electrolysis voltage and overvoltage mechanisms versus current density. The chart shows operating voltage versus current density for an advanced alkaline electrolyzer producing 1 atm of hydrogen and oxygen at 363 K (90°C), using aggressive technical assumptions. The thermoneutral voltage for water electrolysis is 1.48 V. If excess heat is available, the theoretical minimum (i.e., Nernst) voltage for water electrolysis can be much lower (1.18 V), corresponding to electricity energy consumption of only 31.3 kWh/kg H_2. Typically, excess heat is generated by overvoltage losses due to resistance to ion flow in the electrolyte, activation voltage to accelerate O_2 production at the anode (+), and activation voltage to accelerate H_2 evolution at the cathode (–). Electrolyte resistance is linear with electrolysis rate (current), whereas activation overvoltages rise more slowly and are shown as logarithmic. The chart indicates that the sum of the Nernst voltage and overvoltages (at 10,000 A/m^2) can rise to 1.9 V, corresponding to an electrolysis energy consumption of approximately 51 kWh/kg H_2 or 65% electrolysis efficiency (lower heating value). For reference, 1 gallon of gasoline contains approximately 33 kWh energy equivalent to 1 kg of hydrogen fuel.

voltage (1.23). However, this measure is obviously flawed given that, by the measure, the maximum "efficiency" of a perfect electrolyzer could be 120% (1.48/1.23). The correct definition of electrochemical efficiency uses the minimum reversible thermodynamic electrolysis voltage under standard conditions (1.23 V) divided by the actual voltage applied during electrolysis.

The thermodynamics of water decomposition discussed earlier (Fig. 2) apply equally well to thermochemical or electrolytic methods. H_2O decomposition has a positive reaction entropy, reducing the thermodynamic minimum electrolysis voltage at warmer temperatures and potentially allowing for greater use of the excess heat evolved in practical electrolyzers. Warmer temperatures sharply reduce the minimum electrolysis voltage for liquid water, declining from 1.23 V at 300 K to 1.17 V at 373 K. The effect is less dramatic for steam electrolysis (1 atm) above the boiling point of water (from 1.17 V at 373 K to 1.11 V at 573 K) because water vapor has a higher entropy than does liquid water.

Elevated temperatures also increase the effect of steam, hydrogen, and oxygen pressure(s) on minimum electrolysis voltages. High steam pressures reduce (and high hydrogen and oxygen pressures increase) electrolysis voltage logarithmically according to the following (approximate) equation:

$$E = \frac{\Delta G_{std}(T) + RT\ \ln\frac{[pH_2][pO_2]^{1/2}}{[pH_2O]}}{nF}. \quad (19)$$

A rigorous thermodynamic treatment recognizes the minimum electrolysis voltage as the difference in chemical potential of oxygen (or hydrogen) between the equilibrium gas concentrations at the anode and cathode of any electrolytic cell. This results in the minimum work of electrolysis given by

$$\begin{aligned} &\text{work}_{minimum}(\text{mole of } H_2O) \\ &\quad = RT\ \ln\left(pO_2^{0.5}\right) + \Delta H_{std} - T\Delta S_{std} \\ &\quad + RT\ln\left(\frac{pH_2}{pH_2O}\right) \end{aligned} \quad (20)$$

when oxygen ions are transported across the electrolyte.

Alternatively, for a proton-conducting electrolyte, the minimum work is given by

$$\begin{aligned} &\text{work}_{minimum}(\text{mole of } H_2O) = RT\ \ln(pH_2) \\ &+ \Delta H_{std} - T\Delta S_{std} + RT\ln\left(\frac{pO_2^{0.5}}{pH_2O}\right). \end{aligned} \quad (21)$$

The major distinction to be drawn from these two equations is that although both equations yield identical voltages under the same overall H_2O, H_2, and O_2 pressures, the ideal voltage (work) requirement of oxygen ion-based electrolysis depends on the absolute pressure (concentration) of O_2 at the cathode but only on the relative H_2/H_2O ratio at the anode. Conversely, ideal electrolysis voltages across proton-conducting membranes depend instead on the absolute H_2 anode pressure (concentration) and the relative O_2/H_2O ratio at the cathode.

At finite reaction rates (i.e., current density), two kinetic mechanisms (ionic transport across the electrolyte and activation of the rate limiting step at each electrode) raise electrolysis voltages above the ideal thermodynamic minimums, reducing efficiency (Fig. 6). First, ions flowing between electrodes and through the electrolyte encounter ohmic resistance that can be minimized only by using thinner electrolytes, more conductive electrolytes, or larger electrolyte cross-sectional areas (reducing current density). Second, electrons traveling from the external circuit to the cathode, and from the anode to the external circuit, also encounter ohmic resistance, which is typically lower than electrolyte resistance and is, to some extent, a design issue rather than fundamental. Overall, the resistance(s) to ion and electron transport during electrolysis is determined essentially by cell geometry and interconnections, varying with temperature as electrode and electrolyte resistivities change.

The ultimate limits to electrolysis efficiency are controlled by a more fundamental factor: the activation overpotential(s) necessary to accelerate the electrochemical reactions at the anode and cathode. Fortunately, electrolysis reactions accelerate exponentially with activation overpotential (η) according to the Butler–Volmer equation:

$$I = I_0(T) \cdot \left(\exp\left(\frac{(1-\beta)F\eta}{RT}\right) - \exp\left(\frac{-\beta F\eta}{RT}\right)\right). \quad (22)$$

The factors determining the activation overpotential (η in volts) necessary to achieve a given current density at an electrode are physical constants (ideal gas constant R and Faraday's constant F), environment (absolute temperature [T]), and electrochemical surface chemistry fundamentals (exchange current density I_o [T] and a reaction symmetry factor β that is usually ~ 0.5). The most important factor is the exchange current density I_o, which describes the speed with which adsorbed molecule, ion, and electron populations transform into one another on

the surface of each electrode in dynamic equilibrium (i.e., when no net electrolysis current is flowing).

The electrode kinetics that determine exchange current density (I_o) are also influenced by temperature, partial and absolute activities (pressures) of reactants and products, and the density, reactivity, and location of reaction sites. It is thought that electrochemical reactions on electrodes take place almost exclusively near the triple phase boundary, where electrons from the electrode, ions from the electrolyte, and surface-adsorbed gas molecules all exist in close proximity and can readily participate in reactions. If so, structure and geometry issues at the electrode/electrolyte interface can limit electrolysis efficiency. However, beyond issues of cell and electrode geometry, fundamental material, electronic, ionic, and defect chemistry properties, particularly at the surface, drive electrode kinetics and define the ultimate limits to electrolysis efficiency.

The three major electrolysis techniques are based on three classes of electrolytes (aqueous alkaline solution, proton exchange membrane, and solid oxide) at very different stages of commercialization and development. Because the electrolyte determines the operating temperature needed for sufficiently high ionic conductivity, choice of electrolyte is fundamental to an electrolyzer.

Aqueous alkaline electrolysis is the most mature electrolysis technology developed since the 19th century. Alkaline electrolyzers use an aqueous potassium hydroxide (KOH) electrolyte with metallic electrodes in solution. Proton exchange membrane (PEM) electrolyzers, which entered space and defense applications beginning in the 1970s and are now in commercial use as laboratory hydrogen generators, rely instead on a polymer electrolyte membrane with high protonic conductivity. Steam electrolysis at very high temperatures ($>1000\,K$) was researched alongside solid oxide fuel cells (SOFCs) beginning in the 1960s, was developed in Germany during the 1980s, and is still under development today. Virtually all steam electrolysis concepts rely on solid oxide ceramic materials that conduct oxygen ions at very high temperatures ($>1000\,K$).

Alkaline water electrolysis has developed slowly since commercialization during the early 20th century. A few very large plants ($\sim 100\,MW\,H_2$ output power or $3000\,kg\,H_2/h$) were built through the mid-20th century near large hydroelectric sources to produce H_2 for NH_3 synthesis. Developments have included electrolysis at elevated pressures to minimize electrolyte resistance from gas bubbles and a "zero-gap" design in which the electrolyte is essentially absorbed in an asbestos diaphragm sandwiched between the electrodes. Alternative diaphragm materials more resistant to corrosion than asbestos have been researched, but none has yet appeared to simultaneously satisfy the economic and technical criteria as well as asbestos. Nearly all alkaline electrolyzer manufacturers use the so-called filter-press electrolyzer design linking electrolysis cells in series to construct compact, higher pressure electrolyzers. Lurgi Corporation manufactures electrolysis units capable of electrolyzing water at pressures of 3 MPa (~ 500 psi). Only Electrolyzer Corporation in Canada uses a unipolar electrolyzer design linking electrolysis cells in parallel, reducing the consequences of any individual cell failure in exchange for higher cell resistance and lower H_2 production pressures.

The fundamental reactions at the electrodes of an alkaline electrolyzer are as follows:

$$\text{Cathode:}\quad 4H_2O + 4e^- = 4H_2 + 4OH^- \qquad (23)$$

$$\text{Anode:}\quad 4OH^- = O_2 + 2H_2O + 4e^-. \qquad (24)$$

The high concentration of hydroxyl (OH^-) ions in an alkaline electrolyte such as potassium hydroxide (KOH) promotes rapid reaction kinetics and, therefore, smaller activation overvoltages, especially for oxygen evolution at the anode. Higher temperatures would improve electrode kinetics and reduce electrolyte resistance. But corrosion issues restrict alkaline electrolysis to moderate temperatures ($<90^\circ C$). All commercial alkaline electrolyzers use aqueous KOH electrolyte to minimize corrosion.

PEM electrolyzers eliminate the corrosion issues of alkaline electrolysis by using very pure water and replacing the KOH electrolyte with a proton-conducting polymer membrane. Electrochemical devices using PEMs began during the 1950s. DuPont's introduction of Nafion, a perfluorinated ionomer, improved the durability of electrolyte membranes, enabling long-life applications. The basic process of PEM electrolysis is described by the two electrode half-reactions:

$$\text{Cathode:}\quad \left(4H^+ + 4e^- = 2H_2\right) \qquad (25)$$

$$\text{Anode:}\quad \left(2H_2O = 4H^+ + 4e^- + O_2\right). \qquad (26)$$

PEM electrolysis differs fundamentally from alkaline electrolysis in that it is acidic rather than basic. Hydrated protons, rather than hydroxide ions, are conducted across the polymer electrolyte. Proton mobility is very high in perfluorosulfonic acid polymers because protons jump between sulfonate

groups attached in high concentrations to a backbone polymer chain. Proton conductivity is controlled by the concentrations of sulfonic groups but increases relatively slowly with temperature, allowing operation near room temperature. PEM electrolysis requires very active noble metal catalysts to achieve high exchange current densities at room temperature and, therefore, to achieve rapid reaction rates.

The development of commercial PEM electrolysis has its roots in military and aerospace applications, principally for O_2 generation rather than H_2 generation. During the 1980s, PEM membranes also entered industrial use for the production of chlorine by electrolysis of brine. The principal advantages of PEM electrolysis are the use of pure water and high-pressure operation, potentially higher than 200 atm ($\sim$3000 psi) without mechanical compression. PEM electrolyzers for H_2 generation have begun to be commercialized only recently, in the wake of much larger development efforts devoted to closely related stationary and automotive PEM fuel cell technology.

Similarly, steam electrolysis development using doped ZrO_2 electrolytes is an offshoot of larger efforts in ceramic SOFC development. The basic principle of steam electrolysis using solid oxide electrolytes is given by the following reactions:

$$\text{Cathode: } H_2O(g)+2e^- = H_2(g)+O^{2-} \quad (27)$$

$$\text{Anode: } O^{2-} = 1/2\ O_2(g)+2e^-. \quad (28)$$

These reactions occur at very elevated temperatures (1000–1200 K), necessary for adequate conduction of oxygen ions (O^{2-}) through thin layers of ZrO_2 doped with Y_2O_3 to produce large concentrations of oxygen ion vacancies. High temperatures narrow the range of materials that can be used in steam electrolyzer construction, complicated by gas-tight sealing of brittle ceramic materials. But these very high temperatures lower theoretical electrolysis voltages dramatically, accelerate electrode reactions, and raise ionic mobility in the electrolyte, resulting in substantially lower electricity use for steam electrolysis than for any other electrolysis technology.

Steam electrolysis differs fundamentally from other approaches in that the electrode reactions occur between gas phase reactants and crystalline ionic solids rather than aqueous solutions or saturated polymers. A key issue is maximizing the electrode/electrolyte interfacial area in contact with reactant and/or product gases because it is thought that virtually all electrolysis occurs in this region. Contact resistance to ion flow across this interface is also thought to limit the ultimate high current density performance of high-temperature steam electrolysis.

Steam electrolysis was pursued in Germany throughout the 1980s, resulting in pilot plant designs based on ceramic tubes of steam electrolysis cells connected in series. This led to predicted electrolysis requirements of approximately 35 kWh/kg H_2, far lower than the alkaline or polymer water electrolysis requirements of 50 to 60 kWh/kg H_2. However, high-temperature steam electrolysis requires substantial additional thermal energy to generate steam for operation at hot temperatures. Recent solid oxide electrolysis research in the United States has grown out of SOFC development. Technology Management in Cleveland, Ohio, is pursuing steam electrolysis using excess thermal energy from an SOFC.

A long-term possibility for steam electrolysis is the replacement of oxygen ion-conducting ceramic electrolyte materials with proton-conducting ceramics. In principle, this would have the advantage of permitting the production of pure H_2 rather than an H_2/H_2O mixture on the anode side of a steam electrolyzer. The electrode reaction kinetics may differ as well, providing potential opportunities.

In general, high-temperature steam electrolysis appears to offer significant potential to improve electrolysis efficiency but is currently in embryonic development relative to alkaline and PEM electrolysis technologies. Both of the "advanced" electrolysis technologies (PEM and steam electrolysis) would likely develop far more rapidly if PEM fuel cells and/or SOFCs achieve mass commercialization, implying superb underlying electrochemical performance for either or both of these technologies.

6. OVERALL PROSPECTS FOR HYDROGEN PRODUCTION

Currently, the most economic method of H_2 production ($1.00/kg H_2) is steam reforming of natural gas, but future cost reductions are unlikely. In addition, the economic and energetic effectiveness of producing H_2 fuel by consuming natural gas must be compared with direct use of natural gas instead. Because the energy efficiency of direct use may be greater, using natural gas directly may both cost less and produce lower overall greenhouse gas emissions than does using H_2 produced by steam reforming.

Coal gasification to produce H_2 may become competitive if natural gas prices rise, but coal gasification has a higher capital, energy, and carbon intensity than does steam reforming of natural gas. If

CO_2 proves to be sequesterable on a global scale at low cost, both coal and natural gas could power H_2 production well into the 21st century, but at a higher cost than today's approximately \$1.00/kg H_2.

It is likely that thermochemical cycles powered by future fission or (eventually) fusion heat sources can produce low-cost H_2 (e.g., \$1.50/kg) at moderate pressures, continuously, and on large scales (>200 MW). Very large scales (>2000 MW) may be required in the case of fusion-driven thermochemical H_2 production. The characteristics of thermochemical cycles could have advantages for H_2 production integrated with NH_3 synthesis or LH_2 production for aircraft.

Beyond these two applications, the economic and energetic effectiveness of dedicating dispatchable (i.e., nonintermittent) nonfossil energy sources (e.g., nuclear) to H_2 generation is unclear. Generating electricity instead of H_2 from these sources is generally easier, more efficient, and likely to be more profitable. Therefore, thermochemical H_2 production schemes based on nuclear energy face a strategic question: are high-temperature nonfossil heat sources better applied, in economic and/or environmental terms, to dedicated thermochemical H_2 production or electricity generation, from which any surplus (i.e., low-value) electricity could generate H_2 by electrolysis? Answers to this question depend in part on electric generation mixes and pricing structures. More fundamentally, they depend on the efficiency of electrolysis technology.

Production of H_2 by electrolysis eliminates the need to dedicate primary energy sources to generating only hydrogen or electricity. For moderate capital costs (~\$500/kW LHV H_2), electrolysis permits flexible use of primary energy sources to coproduce both electricity and H_2, potentially in response to real-time supply/demand fluctuations and corresponding price signals. Electrolytic H_2 production (either for storage or for transportation fuel) is likely to be the highest value use of surplus electricity from intermittent renewables (e.g., solar, wind), and widespread electrolysis is consequently necessary for these nonfossil energy sources to become dominant.

Electrolysis would also permit a broad range of H_2 refueling infrastructures and onboard storage technologies. H_2 vehicles equipped with metal hydride onboard storage could be refueled overnight in a garage using low-pressure electrolysis. High-pressure electrolysis is the most efficient method of H_2 compression and would allow refueling of vehicles at H_2 pressures of at least 200 atm (~3000 psi). Refueling pressures could be boosted, if necessary, by mechanical compression to more than 350 atm (~5000 psi).

However, the strategic and infrastructure advantages of electrolysis are counterbalanced by higher H_2 production cost due nearly entirely to the price of electricity. If future nonfossil electricity prices drop as low as \$0.05/kWh, conventional electrolytic H_2 will cost \$2.50 to \$3.00/kg (energy equivalent to \$2.50–\$3.00/gallon of gasoline) for the electricity alone.

Although these prices can be easily tolerated by 80-mpg equivalent fuel cell automobiles, which need only approximately 150 kg H_2 annually (\$450/year), they are not economically competitive for H_2-fueled freight trucks and aircraft, which will consume an increasing share of transportation energy in the future. The economic viability of H_2 for these applications will require lower electricity prices and/or lower electrolysis voltages. High-temperature steam electrolysis, whether based on oxygen ion- or proton-conducting ceramic electrolytes, has the greatest prospects for achieving very low electrolysis voltages (~1.0 V) but might not easily withstand intermittent operation, requiring either a dispatchable high-temperature heat source (e.g., nuclear) or thermal storage. Steam electrolysis may be best suited for cost-sensitive continuous operations such as production of NH_3 or LH_2 fuel for aircraft. If the electricity to produce H_2 is not continuously available, such as when wind and solar power are widespread or when electricity transmission and distribution are strained during peak hours, the flexibility of alkaline or PEM technology may offset the necessarily higher electrolysis voltages. Assuming that surplus electricity is available in the future for \$0.05/kWh, improving electrolysis voltages from the current 1.9 to 2.0 V to near the thermoneutral voltage (~1.5 V) could reduce the electricity cost of H_2 to approximately \$2.00/kg.

SEE ALSO THE FOLLOWING ARTICLES

Alternative Transportation Fuels: Contemporary Case Studies • Hydrogen, End Uses and Economics • Hydrogen, History of • Hydrogen Storage and Transportation • Natural Gas Processing and Products • Oil Refining and Products

Further Reading

Berry, G., and Lamont, A. (2002). Carbonless transportation and energy storage in future energy systems. *In* "Innovative Energy Strategies for CO_2 Stabilization" (R. Watts, Ed.), pp. 181–210. Cambridge University Press, Cambridge, UK.

Bockris, J. O. (1976). "Energy: The Solar–Hydrogen Alternative." Halsted Press, New York.
Casper, M. S. (1978). "Hydrogen Manufacture by Electrolysis, Thermal Decomposition, and Unusual Techniques." Noyes Data Corporation, Park Ridge, NJ.
Cox, K., and Williamson, K. (1977). "Hydrogen: Its Technology and Implications." vols. 1–5. CRC Press, Cleveland, OH.
Hoffmann, P. (2001). "Tomorrow's Energy." MIT Press, Cambridge, MA.
Ogden, J., and Williams, R. (1989). "Solar Hydrogen: Moving beyond Fossil Fuels." World Resources Institute, Washington, DC.
Winter, C. J., and Nitsch, J. (1988). "Hydrogen as an Energy Carrier." Springer-Verlag, New York.

Hydrogen Storage and Transportation

GENE D. BERRY, JOEL MARTINEZ-FRIAS, FRANCISCO ESPINOSA-LOZA, and SALVADOR M. ACEVES
Lawrence Livermore National Laboratory
Livermore, California, United States

1. **Introduction**
2. **Hydrogen in the Context of Fuels and Energy Storage Modes**
3. **Historical Approaches to Hydrogen Storage and Transportation**
4. **Liquid Hydrogen**
5. **Compressed Hydrogen**
6. **Compressed Cryogenic Hydrogen**
7. **Hydrogen Storage by Chemical Absorption in Metal Hydrides**
8. **Hydrogen Storage by Physical Adsorption on Carbon**
9. **Hydrogen Storage in Chemical Carriers**
10. **Future Prospects for Hydrogen in Transportation**

Glossary

adiabatic A process (e.g., expansion of a gas) in which no heat is transferred to (or from) the external environment; in a reversible adiabatic expansion, as a gas cools, its internal energy is reduced by the amount of work done by the gas on the environment.

critical point The highest temperature and pressure at which the liquid and gas phases of a substance coexist in equilibrium; the critical point of hydrogen is 33 K and 13 atm.

dormancy The duration for which a cryogenic (liquid) hydrogen vessel can absorb heat from the surroundings without venting hydrogen vapor, thereby preventing the vessel internal pressure from exceeding the maximum allowable working pressure.

lower heating value (LHV) Thermal energy released when a fuel is burned with air at ambient temperature and pressure, assuming that the water exhaust remains as vapor, where 1 kg of hydrogen has an LHV of 120 MJ, nearly equivalent to the LHV of 1 gallon (3.785 L) of gasoline.

multiple-layer vacuum insulation Thermal insulation consisting of multiple layers of highly reflective metallized plastic located between the walls of an evacuated vessel.

steam reforming The dominant method of commercial production of hydrogen from methane and water according to the following reaction: $CH_4(g) + 2H_2O(l) + 245kJ \rightarrow 4H_2(g) + CO_2(g)$.

Composed of a single proton and an accompanying electron, hydrogen is the simplest and lightest element. As a diatomic molecule, hydrogen gas (H_2) is the lightest substance known, nearly 15 times lighter than air and able to escape the earth's atmosphere. Terrestrially, hydrogen is found chemically bound only within heavier substances. In addition to its abundance in the world's oceans and biological matter, millions of years of photosynthetic decomposition of water (H_2O) and atmospheric carbon dioxide (CO_2) is generally believed to be responsible for the creation of the world's fossil fuels, sequestering sizable reserves of hydrogen in hydrocarbon form. This "fossil hydrogen" accounts for more than 30% of the fuel energy in oil and more than 50% of that in natural gas (chiefly CH_4). Today, H_2 is used principally in fertilizer manufacture and as a rocket fuel and chemical. Nearly all commercial H_2 is essentially harvested from CH_4 by reaction with H_2O in a process known as steam methane reforming, which produces H_2 at low cost (if low-cost natural gas is available). Unfortunately, it also generates the by-product CO_2, a greenhouse gas.

1. INTRODUCTION

Recently both industry interest and public awareness of H_2 as a future automobile fuel has increased

 © 2004 Elsevier Inc. All rights reserved.

dramatically. Hydrogen (H_2) differs from other fuels in that it can be produced without carbon dioxide (CO_2) by simple decomposition of water (H_2O) using electricity and/or heat from solar, wind, fission, or fusion power sources. Therefore, H_2 is a versatile and universal carbonless energy carrier, a necessary element for future energy systems aimed at being free of air pollution, CO_2, and other greenhouse gases. If generated from renewable energy, H_2 becomes the crucial link in an inexhaustible global fuel cycle based on the cleanest, most abundant, natural, and elementary substances: H_2, oxygen (O_2), and H_2O.

The physical and chemical properties of hydrogen make its use superior to fossil fuels. H_2 is a simple non-toxic molecule that generates power cleanly and efficiently, even silently and without combustion if desired. However, widespread use of H_2 has been challenging due to its low energy density relative to conventional (hydrocarbon) fuels. Energy density fundamentally drives the feasibility of H_2 fuel by determining the capital, materials, volume, and energy needed for onboard storage.

The economic and physical consequences of energy density, in turn, become strategic, placing limits on the scale, efficiency, cost, range, and shape of H_2 vehicles and refueling infrastructure(s) linking the production of H_2 with its use onboard vehicles. Consequently, improvements in the technology, economics, and energetics of onboard hydrogen storage will likely do the most to speed (or slow) and shape the transition to H_2 as a universal transportation fuel.

2. *HYDROGEN IN THE CONTEXT OF FUELS AND ENERGY STORAGE MODES*

Hydrogen's position in the energy density hierarchy of fuels and energy storage modes is dictated by physical and chemical fundamentals. Gravitational energy storage is relatively weak and at the base of the energy density hierarchy (Fig. 1). The pneumatic energy of air compressed to 700 atm ($\sim$10,000 psi) is roughly 100 times greater than the gravitational energy of an equal volume of water elevated 370 m. Energy density 10 times greater still is present in chemical bonds. Mechanical energy storage (e.g., kinetic energy of a flywheel) can be on the order of chemical bond energies but can also be constrained by geometry. Batteries generate electricity directly (if slowly) from the energy in chemical bonds but are relatively heavy because they must retain their reaction products to be recharged.

Fuels quickly (e.g., combustion) release the energy in chemical bonds but sacrifice efficiency. Fuels also have very high energy densities when they can rely on the atmosphere to provide reactants (O_2) and absorb exhaust products (H_2O and/or CO_2). On a lower heating value (LHV) basis, 3.75 gallons (1kg) of liquid hydrogen (LH_2) is energy equivalent (33.3 kWh, 120 MJ) to 1 gallon of gasoline (2.6 kg), making H_2 by far the lightest, but also the least compact, fuel due chiefly to the extreme simplicity of the H_2 molecule.

Unlike polyatomic hydrocarbon fuels such as gasoline (C_8H_{18}), H_2 consists of just two monovalent atoms. This unique structure gives H_2 extreme combustion characteristics (e.g., low ignition energy, low emissivity, wide flammability limits, high flame speed) and physical properties (e.g., high diffusivity, thermal conductivity, buoyancy, incompressibility). An extremely low boiling point (20.3 K), second only to helium, gives LH_2 a very large coefficient of thermal expansion, expanding nearly 40% between 20 and 30 K.

Particularly striking is the very low atomic density of H_2 both as a compressed gas and as a cryogenic liquid. There are fewer hydrogen atoms by volume in LH_2 than in metal hydrides, most fuels, and common molecular liquids such as octane (C_8H_{18}), ethanol (CH_3CH_2OH), methanol (CH_3OH), cryogenic liquid methane (LCH_4), ammonia (NH_3), hydrazine (N_2H_4), hydrogen peroxide (H_2O_2), and even H_2O (Fig. 2).

The low intrinsic energy density of H_2 can be counterbalanced by its efficient utilization. Although automotive scale fossil-fueled heat engines are limited by variance in output and the balance between thermal and frictional losses to efficiencies of approximately 40%, hydrogen-powered fuel cells can circumvent the Carnot cycle and directly generate electricity at approximately 60% efficiency (or higher at partial load). This improves the comparison between H_2 and other fuels on a deliverable energy basis (Fig. 1). The elemental simplicity of H_2 also allows the inverse process (i.e., H_2 generation by electrolysis) to be elegant and efficient, even on a small scale. Thermochemical production of H_2 is simpler, and therefore more efficient, than production of synthetic hydrocarbon fuels (e.g., CH_3OH), which typically involve H_2 as an intermediate.

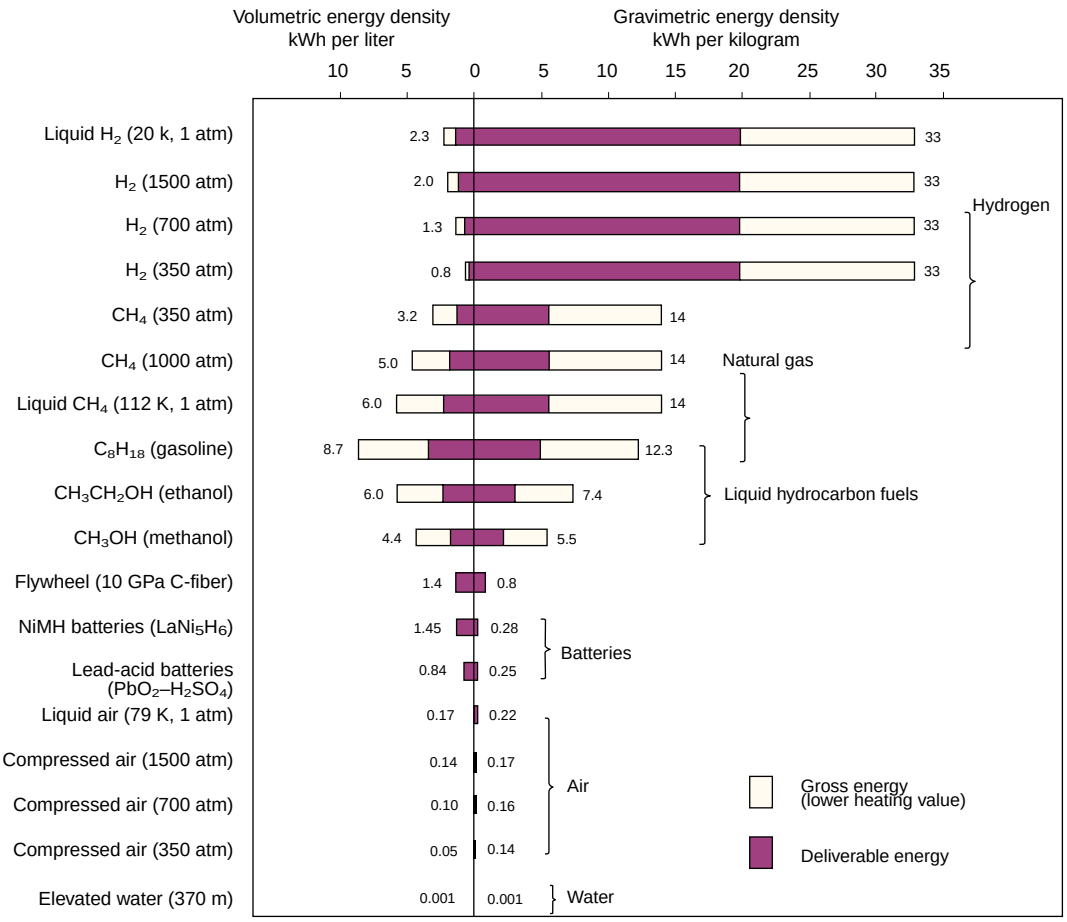


FIGURE 1 Theoretical intrinsic gravimetric and volumetric energy densities of fuels (on an LHV basis) and energy storage technologies. Compressed air and gaseous fuels are shown at 300 K and various pressures, and liquids are shown at 300 K or (if cryogenic) their boiling point. The mass and volume of containment vessels are excluded. Energy densities for flywheels and batteries include only active mass. The values given for compressed air represent maximum possible system work. Numerical values indicate gross energy densities. Maximum deliverable energy densities are indicated graphically for assumed efficiencies of 100% for all energy storage systems, 60% for hydrogen-powered fuel cells, and 40% for hydrocarbon fuels used in automotive scale combustion engines.

3. HISTORICAL APPROACHES TO HYDROGEN STORAGE AND TRANSPORTATION

During the first half of the 20th century, hydrogen was stored infrequently due in part to its low density. At the time, the most important use of H_2 was NH_3 synthesis. Convenience and economics favored producing H_2 by electrolysis, where inexpensive hydroelectric power was available, or else by fossil-fueled thermochemical methods.

H_2 was first used as a fuel in combination with carbon monoxide (CO), distributed by pipeline, and sometimes stored in underground caverns. These H_2/CO mixtures, known as "town gas," were produced by reaction of H_2O with coal and are still distributed by pipeline in Europe today. Pure H_2 pipelines have been operated for decades in the chemical industry, and it should be possible to distribute H_2 through existing natural gas (CH_4) pipeline systems, albeit with reduced energy flow rates and higher compression energies due to the low energy density of H_2. Consideration must also be given to H_2-induced embrittlement of carbon steel pipeline walls. Pipelines and underground storage will be the most economical method of distribution and storage of gaseous H_2 if and when the demand for H_2 becomes large and stable.

When H_2 demand is small, variable, or localized, on-site production by electrolysis or truck transport of compressed hydrogen is viable. Because the high capital and transportation costs of heavy metallic cylinders are magnified by the low energy density of

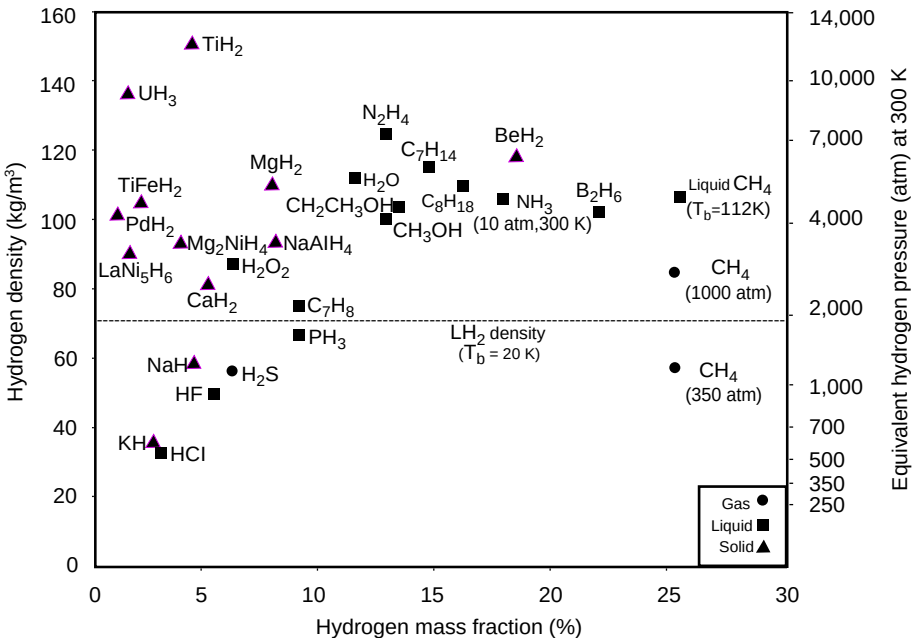


FIGURE 2 Density and mass fraction of hydrogen contained in substances spanning a spectrum of ionic, covalent, and metallic (both high- and low-temperature) hydrides. Hydrocarbon fuels are also shown, including compressed methane (CH_4) at 300 K. The right-hand scale (nonlinear) shows the pressure needed to achieve an equivalent density of compressed hydrogen gas at 300 K. The dashed line indicates that the density (70 kg/m^3) of liquid hydrogen (LH_2) at its boiling point (20 K) is substantially lower than the atomic hydrogen density of many compounds.

gaseous H_2, even moderately sized gaseous H_2 end uses (e.g., fuel stations) are more economically met by lightweight LH_2 transport by truck, on-site steam reforming, or electrolysis.

H_2 was first liquefied by James Dewar in 1898, but liquefaction did not begin on an industrial scale until the 1950s. LH_2 capacity grew during the 1960s to supply the U.S. space program, resulting in large commercial LH_2 plants (60,000 kg/day). Smaller LH_2 plants in Europe (5,000–20,000 kg/day) are less energy efficient. U.S. LH_2 capacity ($\sim$250,000 kg/day) is substantially greater than in other parts of the world, and LH_2 is distributed exclusively by tanker trucks with capacities of up to 4000 kg.

4. LIQUID HYDROGEN

LH_2 use has become routine for launching spacecraft, but liquefying H_2 remains very energy intensive despite decades of industrial experience. Hydrogen liquefaction is complex and energy intensive relative to the other bulk gases (CH_4, N_2, and O_2). Even the noble gases (except helium) are more easily liquefied than is H_2. One complication is that unlike nearly all gases, H_2 has a negative Joule–Thomson coefficient at room temperature, warming when expanded by throttling. H_2 must instead be cooled through work on the environment done by expansion of the gas (e.g., through a turbine). Theoretically, H_2 could be liquefied in a single step if expanded adiabatically from approximately 20,000 atm (to 1 atm). In practice, H_2 is precooled by a closed-loop LN_2 refrigerator, cyclically compressed to moderate pressures ($<$100 atm), and expanded before liquefying by throttling through a Joule–Thomson expansion valve.

An additional complexity is that H_2 molecules, whether liquid or gaseous, exist with two different nuclear spin arrangements. The higher energy form (spins aligned) is orthohydrogen. The lower energy form (spins opposite) is parahydrogen. The relative amounts of each form are in dynamic equilibrium, determined solely by temperature. The equilibrium fraction of parahydrogen is only 25% at 300 K but grows to 99% at 20 K. Spontaneous conversion of the higher energy orthohydrogen to the more stable parahydrogen form at cryogenic temperatures releases enough thermal energy ($\sim$500 kJ/kg H_2) to vaporize 10% of a given quantity of LH_2 within 10 h and 50% within 1 week.

The conversion of orthohydrogen to parahydrogen is intentionally accelerated in LH_2 plants, generating additional heat that must be removed.

This increases the theoretical work of H_2 liquefaction from 3.25 to 3.92 kWh/kg LH_2. In industrial practice, producing 1 kg of pure para-LH_2 requires 10 to 14 kWh, equal to 33 to 42% of the fuel energy (LHV) in LH_2. Magnetocaloric refrigeration, employing cyclic magnetization of tailored materials in very high magnetic fields, may ultimately be capable of reducing liquefaction energy requirements to perhaps 7 kWh/kg LH_2.

Despite the energy intensity of cryogenics, LH_2 is the favored method of large-scale hydrogen storage (e.g., ~270,000 kg LH_2 at Cape Kennedy, energy equivalent to 9 million kWh). LH_2 is particularly suited for aircraft, where the higher gravimetric energy density of hydrogen (120 MJ/kg for LH_2 vs 45 MJ/kg for jet fuel) counterbalances its roughly four times greater volume. The first successful test of LH_2 was in one engine of a modified B-57 aircraft during the late 1950s. In 1988, the Soviet Union reported successful testing of hydrogen in one engine of a modified civilian airliner.

Several design studies by Lockheed examined LH_2 airliners during the 1970s, evaluating lightweight fuel tanks that could store two to three times their own weight in fuel (25,000 kg LH_2). These studies indicate that the most favorable locations for LH_2 tanks are at both ends of the fuselage rather than within or supported from the wings. LH_2 aircraft would likely be 10% longer than fossil-fueled aircraft but would compare favorably in terms of fuel economy, wingspan, engine size, and takeoff noise, weight, and distance.

An LH_2 airliner designed to fly 400 passengers 10,000 km would weigh 75% as much as a conventional aircraft at takeoff and use 90% of the fuel energy. If designed for an 18,000-km range, an LH_2 aircraft could weigh only 50% as much as a fossil-fueled airliner at takeoff and use 65% of the fuel energy. It is generally agreed that LH_2 aircraft will compare even more favorably to conventional aircraft for supersonic air travel. LH_2 aircraft may also be able to use the cooling capacity of LH_2 to chill the wings and other surfaces so as to maintain laminar airflow and reduce aerodynamic drag.

The rationale for LH_2 storage onboard automobiles is similar to the case for aircraft but is not as clear-cut. The advantages of LH_2 originate from the fact that it is relatively compact (70 kg H_2/m^3), minimizing the capital investment necessary to centrally produce, liquefy, distribute, and store LH_2 throughout the refueling infrastructure and onboard automobiles. Automotive LH_2 issues have been researched for more than two decades, principally by BMW. Through five generations of prototype vehicles, BMW has developed engines and onboard LH_2 storage sufficient to drive a 2100-kg sedan 580 km (360 miles) using 13.5 kg (50 gallons) of LH_2 (a gasoline equivalent fuel economy of 28 mpg or 11.3 km/L). BMW has built and tested a fleet of 15 dual-fuel LH_2 sedans with reduced fuel economy and H_2 range (320 km or 200 miles). Manual and automated LH_2 stations can now refuel vehicles in less than 3 min. The world's first public LH_2 filling station opened in Munich, Germany, in 1998, but automotive LH_2 storage issues remain.

Automotive LH_2 tanks, storing only 5 to 15 kg LH_2, have surface/volume ratios 200 to 400 times larger than those for aircraft (25,000 kg LH_2), sharply curtailing capital cost and volumetric efficiency advantages. Typical automotive LH_2 tanks occupy a volume 1.3 to 1.5 times greater, and weigh 5 to 10 times more, than the LH_2 they contain. One fundamental difficulty is the extreme thermal expansion coefficient of LH_2 at temperatures and pressures below the critical point (33 K, 13 atm) where LH_2 and H_2 vapor coexist. In some cases, as LH_2 warms inside a tank, H_2 vapor actually condenses to accommodate LH_2 thermal expansion. Therefore, LH_2 tanks are fueled only 85 to 90% full to prevent LH_2 spills. This intrinsic volume penalty makes heat transfer reduction without increasing vessel volume critical. LH_2 vessels are typically double-walled aluminum or steel, sandwiching approximately 2 cm (70 layers) of multiple-layer insulation *in vacuo*.

Multiple-layer insulation can reduce heat flow into automotive LH_2 vessels to less than 1 W, a significant fraction of which is conducted through plumbing and supports suspending the inner vessel to avoid crushing the insulation. LH_2 tanks are designed to vent H_2 when warmed to 27 to 28 K (where vapor pressure is 6–8 atm). Most materials have very small heat capacities at such low temperatures. The heat capacity of aluminum is 50 times lower at 30 K than at 300 K. Below 30 K, the thermal inertia of an LH_2 tank is due principally to the LH_2 itself. Warming 140 L (5 kg) of LH_2 from 21 to 31 K requires nearly the same thermal energy as does melting 2 L of ice. A continuous 1-W heat leak will cause a 5-kg LH_2 vessel to "boil off" H_2 vapor at a rate equivalent to approximately 300 W of lost fuel energy after only 5 days. This is probably the greatest challenge facing onboard LH_2 storage for automobiles.

Dormancy times and boil-off rates of LH_2 tanks can significantly degrade effective fuel economy unless vehicles are driven far enough, on average, to consume H_2 faster than it evaporates. To avoid

venting H_2 vapor, an 80-mpg (34-km/L) equivalent LH_2 vehicle fueled with 5 kg LH_2 must be driven an average of approximately 15 miles (25 km) daily to offset 1 W of heat transfer (Fig. 3). If parked, the vehicle could remain dormant (i.e., no venting of H_2 vapor) for only 5 days. If future LH_2 tanks absorbed heat at only approximately 0.25 W (representing "perfect" insulation, magnetic vessel suspension, or perhaps active refrigeration), 80-mpg equivalent vehicles driven an average of only 4 miles (6.5 km) daily would never vent H_2, and 5-kg LH_2 tanks could remain dormant for 20 days.

However, without such advances, high(er) fuel economy vehicles will likely lose some fuel to evaporation unless they are routinely driven more than approximately 15 miles daily, reducing effective fuel economy. This could create an incentive for drivers to refuel often so as to keep their fuel tanks cold, or it could encourage excess driving. Vehicles venting H_2 can raise safety issues in unventilated areas (e.g., garages), and LH_2 vehicles parked for long periods of time (e.g., at airports) might not retain enough fuel to reach an LH_2 station.

In the final analysis, LH_2 storage is the most energy-intensive method of onboard H_2 storage, but it is also the lowest capital cost, and likely the most compact and perhaps safest, method. LH_2 storage may be most appropriate for vehicles storing more than 5 kg of H_2 fuel. The low-capital but energy-intensive cost structure of LH_2 storage is best suited for vehicles that are refueled infrequently but are occasionally driven long distances. For vehicles driven and refueled frequently (e.g., 12,000 miles/year for a 60-mpg equivalent H_2 vehicle with 5 kg LH_2 onboard), the capital cost advantage (perhaps ~$1000 saved for each vehicle) of LH_2 storage may be offset by the higher (energy) cost (~$0.50/kg) of LH_2 versus compressed H_2.

Therefore, to be economically compelling, it appears that LH_2 tank technology will need to be capable of about 2 weeks of unrefueled moderate (20 miles/day on average) driving in high-fuel economy H_2 vehicles. Improved LH_2 tanks with heat transfer rates of 0.5 W or lower would be energetically efficient (venting no H_2 vapor) at this level of vehicle usage and would provide dormancy of 10 days for 5-kg LH_2 tanks.

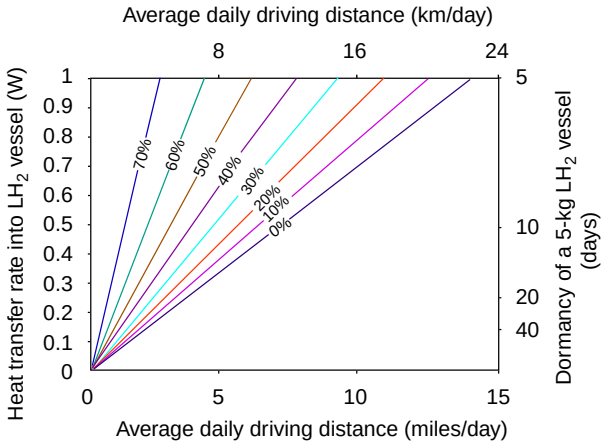


FIGURE 3 Cumulative LH_2 fuel lost to venting (boil off) of H_2 vapor due to continuous heat transfer (left axis) into the LH_2 tank of an 80-mpg equivalent H_2 vehicle as a function of its average daily driving distance (horizontal axes). The vehicle is assumed to refuel 84% full with 21 K LH_2 and vent hydrogen as necessary to keep the internal H_2 vapor pressure at or below 6 atm. The right axis indicates the dormancy (days before venting any H_2 vapor) of a 5-kg LH_2 tank if the vehicle is parked immediately on refueling.

5. COMPRESSED HYDROGEN

Currently, the principal alternative to LH_2 is to compress gaseous hydrogen (GH_2). Use of gaseous fuels onboard vehicles is fairly well established. There are approximately 250,000 compressed natural gas (CNG) vehicle cylinders currently used in the United States. There is also relatively broad experience with GH_2 in many demonstration vehicles, especially buses, which can accommodate large storage vessels. Under ambient conditions (1 atm, 300 K), H_2 has a density of only 0.0818 kg/m^3. Compression of H_2 up to pressures of 250 atm, a standard pressure for CNG vehicles, is thought to be insufficient for GH_2 vehicles. GH_2 at 250 atm is four times bulkier than LH_2 and 15 times bulkier than the energy equivalent volume of gasoline.

Because the theoretical (isothermal) work of compressing hydrogen rises only logarithmically with pressure, ambient temperature compression is the most energy-efficient method for densifying H_2 in both theoretical (Fig. 4) and practical terms. Compressing GH_2 to 250–1000 atm requires 1.5–2.0 kWh/kg theoretically and 2.5–4.0 kWh/kg in practice, substantially less energy than is required for conventional H_2 liquefaction (10–14 kWh/kg). Electrolysis at high pressure can compress GH_2 at near theoretical efficiency.

Compressed H_2 storage has historically been limited by the strength, and especially the weight, of pressure vessel materials. Current high-strength (1.5–3.0 GPa ultimate strength) composites permit lightweight vessels to approach the fundamental limits of GH_2 storage (Fig. 5) imposed by the

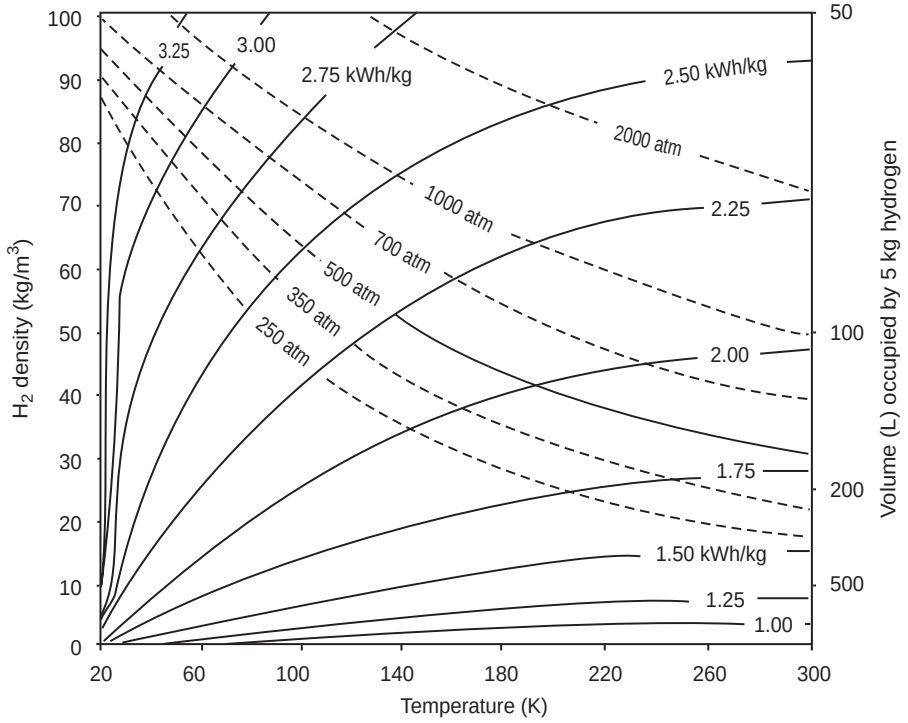


FIGURE 4 Theoretical minimum work to densify hydrogen from the reference gas density (0.0818 kg/m^3) at 300 K and 1 atm pressure to the densities (left axis) and temperatures (bottom axis) shown. The corresponding volume occupied by 5 kg of hydrogen gas, energy equivalent to 5 gallons (19 L) of gasoline, is shown on the right-hand axis. Dashed lines indicate the pressure needed to achieve a given hydrogen density as a function of temperature. The density that can be achieved for a given amount of work (solid lines) declines sharply at cryogenic temperatures. The work necessary to convert orthohydrogen to parahydrogen is not included but can be important for hydrogen stored more than a few days at temperatures below 60 K. The ideal work of converting normal (75% ortho) LH_2 to pure para-LH_2 is 0.7 kWh/kg at the boiling point (20 K), where LH_2 has a density of 70 kgH_2/m^3.

nonideal gas behavior of H_2. Even at 300 K, H_2 is not an ideal gas. It is less compressible and becomes more so at higher pressures, reaching only approximately 80% of ideal gas density at 350 atm ($\sim$5000 psi) and 60% at 1000 atm. Even for idealized infinitely thin pressure vessels, this incompressibility limits the hydrogen storage density achievable by high-pressure GH_2 storage.

For a given storage pressure, the strength of practical wall materials places a lower limit on vessel wall thickness. Thicker vessel walls occupy volume otherwise available for H_2 gas storage. Theoretical minimum wall thicknesses are more than doubled in practice to provide a burst pressure safety factor of 2.25 times the actual H_2 storage pressure. Future pressure vessel designs, using onboard sensors, may be allowed to use burst pressure safety factors of 1.8.

The net impact of H_2 gas behavior, burst pressure safety factor, and vessel wall strength is best shown by considering the amount of H_2 gas that can be stored in a vessel of fixed external volume (Fig. 6). The volume between the rear seat and the trunk of an automobile is approximately 250 L (66 gallons). This volume could accommodate H_2 storage in a 54-cm diameter cylindrical pressure vessel 120 cm in length with ellipsoidal end caps. Although actual vessel configurations will vary, H_2 storage density depends little on vessel aspect ratio or size. However, the density of H_2 gas itself depends nonlinearly on pressure. A 250-L volume (300 K) contains 5.9 kg H_2 at 350 atm or 10 kg H_2 at 700 atm but only 12.3 kg H_2 at 1000 atm, representing the ideal maximum H_2 storage capacities of (infinitely thin) pressure vessels with 100% volumetric efficiency.

Ambient temperature pressure vessels made of high-strength materials (1.5–3.0 GPa or 220,000–440,000 psi ultimate strength) achieve 90 to 95% volumetric efficiency (hydrogen volume/vessel volume) at pressures up to 350 atm. Above 350 atm, as thicker walls become necessary, wall strength and safety factor are significant influences on volumetric efficiency and, therefore, on storage capacity. For

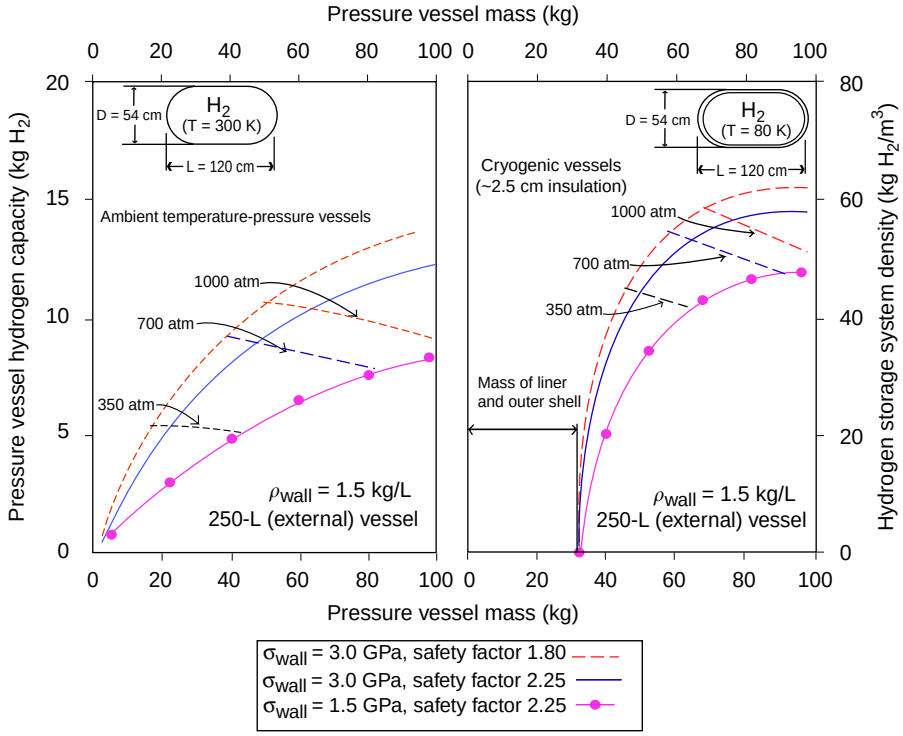


FIGURE 5 Mass of three idealized carbon composite pressure vessels as a function of H2 storage capacity at 300 K (left) and 80 K (right). All vessels shown are cylindrical with ellipsoidal end caps and identical external volumes (250 L), and dimensions (120 cm long, 54 cm diameter). Vessel wall thickness calculations include thick-wall effects. The heaviest vessels are designed with a burst pressure safety factor of 2.25 and wall stress of 1.5 GPa. The intermediate-weight vessels in both panels have twice the wall stress (3.0 GPa). The lightest vessels have identical wall stress (3.0 GPa) but are designed to a (lower) burst pressure safety factor of 1.8. The cryogenic (80 K) pressure vessels include an additional mass of approximately 30 kg for 3 mm aluminum liner, 2 mm outer shell, and 2.5 cm multiple-layer cryogenic insulation. Overall hydrogen storage system density is shown on the right-hand axis.

example, a 700-atm (~10,000-psi), 250-L (external volume) pressure vessel made of 1.5-GPa material with a safety factor of 2.25 can store only 7.6 kg H_2 instead of the theoretical 10 kg H_2. The same 250-L vessel made of materials with twice the strength (3.0 GPa) would have thinner walls and could store 8.6 kg H_2 at 700 atm.

The disadvantage of low-strength wall materials is compounded by the nonlinear relationship between the density of H_2 gas and pressure. Reducing the wall strength of a 250-L vessel storing this same 8.6 kg H_2 from 3.0 to 1.5 GPa increases wall thickness, displacing available H_2 storage volume requiring a 40% higher pressure (1000 atm) to overcome. At higher pressures (>1000 atm), an ultimate H_2 storage density is reached as hydrogen gas becomes more incompressible and vessel walls must be ever thicker. These two factors ultimately limit maximum H_2 storage density to pressures of approximately 10% of vessel wall material strength and 60% volumetric efficiency, whereas maximum economic GH_2 pressures are likely to occur at approximately 5% of vessel wall material strength.

The attractiveness of automotive compressed H_2 storage will grow with the strength of the materials available for pressure vessel construction. Pressure vessels for H_2 vehicles have been certified up to 10,000 psi (~700 atm), enabling a 250-L pressure vessel to store 8 kg H_2, adequate for a 650-mile (1100-km) range in an 80-mpg equivalent H_2 vehicle. Vessels made of very strong materials (3.0 GPa) could store more than 10 kg H_2 using higher pressures (>1000 atm). Future vessel materials stronger than 6.0 GPa (900,000 psi) and vessel pressures exceeding 1500 atm would be needed for GH_2 storage densities (at 300 K) comparable to LH_2 tanks (13 kg H_2 in 250 L).

One complication of all gaseous fuel storage, even at moderate pressures, is refueling overpressure. When a pressure vessel is quickly filled with gas, the temperature rises due to the compression of gas within the vessel by subsequent gas molecules.

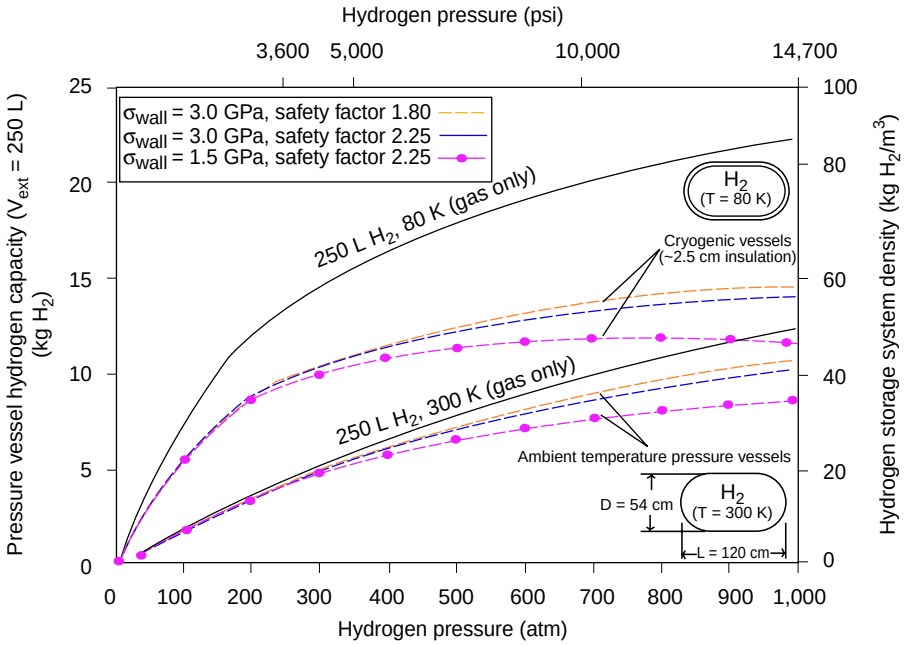


FIGURE 6 Hydrogen storage capacity versus pressure for idealized pressure vessels of varying wall stress and burst pressure safety factor at ambient (300 K) and cryogenic (80 K) temperatures. All vessels shown are cylindrical with ellipsoidal end caps and identical external volumes (250 L) and dimensions (120 cm long, 54 cm diameter). Vessel wall thickness calculations include thick-wall effects. Three vessels are shown for each temperature case: 3.0 GPa wall stress with a safety factor of 1.80, 3.0 GPa wall stress with burst pressure safety factor of 2.25, and 1.5 GPa wall stress with burst pressure safety factor of 2.25. Ambient vessels assume negligible (i.e., metallized polymer) liner thickness. Cryogenic vessels have 2.5 cm multiple-layer insulation, 3 mm internal metal liner, and 2 mm outer shell. Corresponding hydrogen storage system densities are shown on the right-hand axis. The quantities of H2 gas contained in a 250-L volume as a function of pressure at 80 and 300 K are also shown for reference.

Quickly filling an empty vessel to 350 atm with H_2 will theoretically warm it from 300 to 440 K, reducing storage capacity by 30% and requiring a 50% overpressure (525 atm) to overcome. Fortunately, in practice, pressure vessels absorb some of this thermal energy, and so warming to only approximately 350 K is observed. This reduces storage capacity by only 12%, and the storage capacity can be restored with 15% overpressure (400 atm). Overpressure effects vary little with filling pressure between 350 and 1000 atm and are lessened because H_2 is easier to compress as it warms. CH_4 exhibits opposite physical properties, becoming less compressible as it warms. The storage capacity reduction from warming CH_4 by 25 K is equivalent to H_2 warmed by 50 K. Therefore, although hydrogen vehicles will have warmer fast fill temperatures, fast fill overpressures may be comparable for H_2 and CNG vehicles.

Safety, both real and perceived, is an often raised criticism of compressed H_2 storage. However, the safety risks of storing compressed gases are not a simple function of pressure. The overall safety of pressure vessels can be counterintuitive, for although vessel wall strength and impact resistance increase directly with storage pressure, the maximum mechanical energy released by sudden expansion (e.g., in a vessel rupture) of the stored gas (hydrogen) does not. That is, 1 kg H_2 stored at 70 atm (300 K) will release a theoretical maximum mechanical energy of 0.55 kWh if suddenly expanded to atmospheric pressure (cooling substantially in the process). This maximum energy release increases only slightly if H_2 is stored at much higher pressures (Fig. 7). Raising pressure from 70 to 1000 atm (a 1400% increase) increases the maximum (theoretical) mechanical energy release by only 10% while shrinking vessel volume 83% and strengthening the vessel wall many times over. Over the likely range of onboard GH_2 storage pressures (350–1000 atm), the maximum mechanical energy release is nearly constant at 0.6 kWh/kg GH_2. Therefore, compact, higher pressure vessels with thick walls are likely safer than larger, lower pressure vessels with thinner walls.

In addition to pressure, the crash-worthiness of compressed H_2 storage depends on vessel shape,

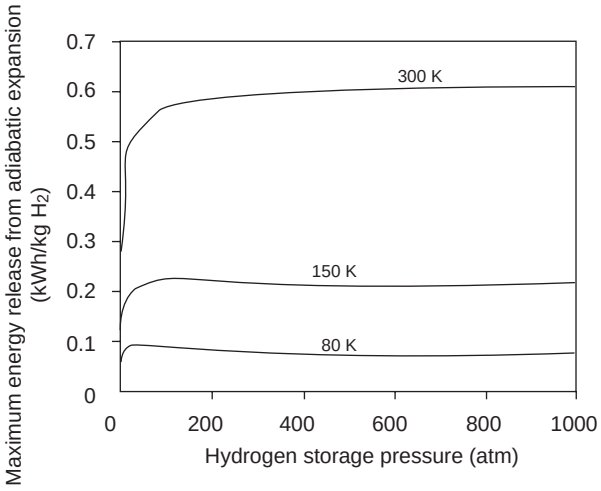


FIGURE 7 Maximum mechanical energy (per kilogram of hydrogen) released on instantaneous expansion of H2 gas (e.g., from a pressure vessel) as a function of initial storage pressure at 80, 150, and 300 K. This mechanical energy is the theoretical maximum available work based on reversible adiabatic expansion from the pressure shown to 1 atm, calculated from internal energy differences of H2 gas before and after isentropic expansion. For comparison, the chemical energy of hydrogen fuel is 33.3 kWh/kg.

orientation, volume, location, and likely accident scenarios. Vehicles with GH_2 stored onboard have been dropped from cranes, simulating a 30-mph rear-end collision, without releasing any hydrogen gas. It is generally thought that the best location to store H_2 is probably underneath/behind the rear seat of a vehicle. This sets an upper bound on vessel volume, aspect ratio, and thereby automotive onboard H_2 storage capacity given the limits of pressure and material properties (Fig. 6). Current carbon composite pressure vessels are designed to withstand pressures 2.25 times higher than their operating pressure and to leak before bursting. They are also subjected to a variety of safety tests, most notably bonfire, gunfire, and drop tests.

6. COMPRESSED CRYOGENIC HYDROGEN

Compressed H_2 is much denser at cryogenic temperatures (Fig. 4), potentially reducing storage volume dramatically, although cryogenic vessels incur volume inefficiencies similar to LH_2 tanks due to an inner liner, multiple-layer cryogenic insulation *in vacuo,* and a metallic outer shell. Built to withstand higher pressures than do LH_2 tanks, cryogenic pressure vessels could reduce H_2 vapor boil-off rates and extend dormancy when filled with LH_2. Alternatively, cryogenic pressure vessels can lower energy intensity if fueled with cryogenic GH_2 rather than LH_2. Cooling GH_2 to 80 K requires less energy than does liquefying H_2 at 20 K (Fig. 4). Additional energy could be saved by avoiding conversion of orthohydrogen to parahydrogen given that the equilibrium fraction of parahydrogen (25% at 300 K) is only 50% at 80 K. Flexibly fueled cryogenic vessels could reduce energy intensity and fuel cost further if ambient (300 K) GH_2 were used for urban driving and cryogenic (80 K) GH_2 or LH_2 were used only for occasional long trips.

Flexibly refueled cryogenic vessels will need to store H_2 at 200 to 400 atm to compare favorably with conventional ambient temperature pressure vessels. An insulated pressure vessel of moderate strength (1.5 GPa) material with an external volume of 250 L could contain 2.5 kg H_2 at 200 atm and 300 K, sufficient for a 200-mile (320-km) range in an 80-mpg equivalent H_2 vehicle. If filled with 80 K GH_2, the same 250-L, 200-atm vessel would hold 8.5 kg H_2 (680-mile or 1100-km range). Pressures of 1000 atm would be needed to store the same 8.5 kg H_2 at 300 K in a conventional pressure vessel of the same strength wall material (1.5 GPa) (Fig. 6). This dramatic relative advantage of flexibly fueled cryogenic vessels declines sharply in excess of 200 atm. Doubling H_2 pressure (to 400 atm) nearly doubles ambient H_2 capacity of a cryogenic vessel (to 5 kg GH_2 sufficient for a 400-mile or 660-km driving range). However, cryogenic (80 K) H_2 density over the same pressure range improves only 30% (Fig. 4). A conventional 250-L GH_2 vessel could store 5.5 kg H_2 at 400 atm (Fig. 6), sufficient for a 440-mile (725-km) range without the cost and complexity of cryogenics.

Cryogenic pressure vessels could also be designed to operate exclusively on LH_2, accepting the full energy intensity of liquefying H_2 but improving thermal endurance over low-pressure LH_2 tanks, enabling very compact storage for high-fuel economy vehicles requiring small amounts of fuel (kg LH_2). Pressure vessels can be filled more completely than can LH_2 tanks because the former can withstand the high pressures generated by LH_2 thermal expansion. Whereas conventional LH_2 tanks fueled 85% full must vent H_2 when warmed to 27 to 28 K, a 100-atm, LH_2-fueled pressure vessel could remain dormant until warmed to 33 K (45 K if 85% full). When fueled with para-LH_2, cryogenic vessels could be cooled substantially if catalysis of the endothermic conversion of cold (40–80 K) parahydrogen vapor to orthohydrogen were feasible.

Carbon composite pressure vessels lined with approximately 3 mm of aluminum have passed

high-pressure ambient and cryogenic cycling tests representative of a 240,000-km vehicle life without any degradation in burst pressure. They have also passed a number of certification tests similar to those for ambient pressure vessels (e.g., gunfire, bonfire).

A potentially significant safety distinction between cryogenic and ambient tanks is that cryogenic gases release much lower amounts of mechanical energy when suddenly expanded from high pressure. The theoretical maximum mechanical energy that can be released by sudden (adiabatic) expansion of hydrogen from 350 atm is nearly eight times smaller at 80 K than at 300 K (Fig. 7).

Although cryogenic pressure vessels may ultimately offer safety and other technical advantages over ambient pressure vessels, and they may offer energy intensity or dormancy advantages over LH_2 tanks, cryogenic vessels are a far less mature technology. To date, preliminary evaluations of cryogenic pressure vessels versus ambient high-temperature tanks indicate comparable or lower safety risks in the former, but insulation performance, the use of cryogenic pressure vessels on vehicles, and flexible refueling have yet to be demonstrated.

7. HYDROGEN STORAGE BY CHEMICAL ABSORPTION IN METAL HYDRIDES

Although H_2 vehicles can feasibly use LH_2, GH_2, or cryogenic GH_2 storage, H_2 has higher storage costs (both capital and energy) and occupies greater volume in each of these forms than do competing alternative fuels, especially CH_4 (natural gas), even when H_2 is stored at higher pressures and/or lower temperatures. H_2 can also be stored by unique methods, circumventing the need for high pressures and/or cryogenics.

The most mature of these approaches is the reversible chemical reaction of metals with H_2 to form what are known as metal hydrides. The roots of this approach date back to Graham's 1866 discovery of H_2 absorption in palladium. Many metals react with and absorb hydrogen, releasing heat. Later, this hydrogen can be recovered on moderate heating. Metal hydrides have the advantage of storing H_2 at high density but are typically very heavy, storing only 1 to 2 wt% H_2. Most metal hydrides and other compounds have atomic hydrogen densities of 90 to 110 kg H_2/m^3 (Fig. 2)—30 to 60% denser than LH_2. However, hydride storage systems usually achieve only approximately 50% volumetric efficiency because the hydrides are in powder form and expand when chemically absorbing H_2.

Chemisorption of H_2 by virtually every element has been studied without finding an ideal candidate for H_2 fuel storage. Most are too heavy, bond too strongly to hydrogen, or are insufficiently abundant. Magnesium is sufficiently light to store 7.6 wt% H_2, forming MgH_2, the elemental hydride closest to automotive viability. But the strong magnesium–hydrogen bond requires significant energy (75 kJ/mol H_2 at temperatures near 550 K) for H_2 release, making MgH_2 (and related alloys) unattractive for current fuel cell vehicles with exhaust temperatures of only approximately 380 K.

The field of candidate H_2 storage materials was expanded from elements to compounds with the discovery of the reversible hydride $ZrNiH_3$ by Libowitz in 1958. Entire classes of intermetallic hydrides were subsequently developed. These are known as the A_2B, AB, AB_2, and AB_5 hydrides, in which "A" denotes elements that strongly absorbs hydrogen (e.g., Mg, Ti, La) and "B" denotes elements that do not, usually lighter transition metals in the first row of the periodic table (e.g., V, Cr, Fe, Co, Ni, Cu). During the late 1960s, two prototype practical hydrides, $LaNi_5H_6$ and $TiFeH_2$, were developed to adsorb H_2 interstitially, with low enough bond energies ($\sim$30 kJ/mol H_2) to release H_2 between 273 and 373 K and to adsorb H_2 quickly at low pressures (1–10 atm). Many variations on these materials have been studied, with some finding application in nickel metal hydride batteries, but all known low-temperature hydrides store at most 2 wt% H_2. Practical automotive hydride systems will need additional mass, reducing effective storage densities to 1.1 to 1.5 wt% H_2, corresponding to a 330- to 450-kg system mass for 5 kg of onboard H_2 storage capacity (energy equivalent to 5 gallons of gasoline). Daimler–Benz built and tested a fleet of 10 hydride vans and station wagons during the 1980s. They were successfully operated over 240,000 km in West Berlin over 4 years, proving automotive hydride storage to be feasible, if perhaps too heavy. The most significant issue for hydride storage is probably refueling speed. H_2 absorption during refueling generates heat that, if not removed quickly, warms the hydride, requiring higher pressures to drive further hydrogen absorption, undercutting an advantage (low pressure) of metal hydrides.

More recent hydride research has ventured beyond the classical crystalline intermetallic hydride compounds, focusing on solid solution alloys as well

as disordered, amorphous, and nanocrystalline materials. However, such hydrides may be metastable or susceptible to degradation over many absorption/desorption and thermal cycles.

The greatest research progress has arguably been made with hydride complexes, in which hydrogen and a metal atom form a complex ion, stabilized by another element. Magnesium nickel hydride (Mg_2NiH_4) is a prototypical example, with Mg stabilizing an $(NiH_4)^{-2}$ ion. Complex hydrides can be formed (e.g., Mg_2FeH_6) from metals that do not themselves form intermetallic compounds, further expanding the range of potential hydride storage materials. Complex hydrides made of light metals (e.g., Li, Be, B, Na, Al) contain 5 to 14 wt% H_2 but typically do not absorb H_2 directly. The hydride with the highest known ratio of hydrogen to metal atoms, storing 2.7 wt% H_2 at twice the density of LH_2, is the complex hydride $BaReH_9$, but it is irreversible and is synthesized by aqueous techniques.

A breakthrough in hydride research occurred during the mid-1990s when Bogdanovic reported that sodium alanate ($NaAlH_4$), a lightweight complex hydride of low-cost materials, could reversibly store and release H_2 at temperatures of 400 K if catalyzed by titanium and/or zirconium. $NaAlH_4$ is more complex than the intermetallic hydrides. Hydrogen atoms reside on lattice sites in $NaAlH_4$ and are released through multiple decomposition steps involving metal atom diffusion and, therefore, warmer temperatures.

When $NaAlH_4$ is warmed, it initially releases half of its contained H_2, decomposing into aluminum (Al) and an aluminum-poor complex, Na_3AlH_6, which contains strongly bound hydrogen. On further warming, Na_3AlH_6 releases additional H_2 and decomposes into Al and sodium hydride (NaH), retaining a quarter of its original hydrogen atoms.

Based on these reactions $NaAlH_4$ contains 7.46 wt% H_2, of which 5.60 wt% is theoretically recoverable (3.73 wt% H_2 if only the low-temperature decomposition is used). The decomposition of Na_3AlH_6 is endothermic (47 kJ/mol H_2), releasing 1 atm H_2 at 380 K. When the cycle is reversed for refueling, H_2 is absorbed by Al and NaH, generating temperatures in excess of 455 K (the melting point of $NaAlH_4$). These high temperatures can drive refueling pressures as high as 100 atm but are needed to accelerate the sluggish refueling kinetics of $NaAlH_4$.

The tightly bound hydrogen in NaH reduces the effective maximum H_2 storage density of $NaAlH_4$ to approximately 70 kg H_2/m^3 (comparable to the density of LH_2). This may be offset by potential hydride bed engineering advantages. Volumetric efficiency of 75% would be needed for $NaAlH_4$ systems to achieve H_2 storage density comparable to that of other hydride systems. Future automotive (5 kg H_2) $NaAlH_4$ systems could perhaps occupy 100 L and weigh 170 kg.

8. HYDROGEN STORAGE BY PHYSICAL ADSORPTION ON CARBON

A second approach is to physically adsorb and store H_2 molecules on the surface of an adsorbent (e.g., carbon) rather than to dissociate H_2 molecules on a metal surface and absorb hydrogen atoms throughout the body of a metal, forming a metal hydride. Physical adsorption of H_2 molecules onto high-surface area adsorbents has been studied for decades. Initial research began during the 1960s and continued through the early 1990s, examining cryogenic H_2 adsorption onto high-surface area carbon materials. However, the advantage of cryogenic adsorbents was found to decline with pressure because the volume of the adsorbent itself displaces ever larger quantities of H_2. At pressures exceeding approximately 200 atm, removing the adsorbent usually increases cryogenic H_2 storage density.

Recent research has focused more directly on storage, pursuing H_2 adsorption at low pressures and ambient temperatures, where the relative advantage of adsorbents is greater. Ambient temperature H_2 adsorption has required engineered carbon materials, with atomic scale porosity and high surface areas. Graphite nanofibers are a class of engineered carbon materials that received significant attention, with experimental claims of extraordinary gravimetric H_2 storage densities (>50 wt% H_2). Subsequent work has been unable to duplicate these claims but also has produced widely varying results.

Carbon nanotubes are a subsequent class of engineered adsorbents. They are atomic scale tubes with graphite structure walls and internal diameters of 20 to 100Å. Intercalating lithium or potassium in multiple-wall carbon nanotubes has been reported to enable dry adsorption of approximately 2 wt%w H_2.

Single-wall carbon nanotubes are a further refinement, capable of higher purity ($\sim 50\%$) fabrication, with diameters of approximately 10 Å, nearly the 2.9-Å kinetic diameter of H_2. Single-wall nanotubes form compact hexagonal bundles. Early experiments on single-wall nanotubes measured an adsorption enthalpy of approximately 20 kJ/mol H_2, substantially

larger than the 4 kJ/mol H_2 seen with graphite adsorbents. Experiments on higher purity nanotubes measured desorption of H_2 from two classes of adsorption sites: one active near 300 K and another active only at higher temperatures. Current evidence indicates that 4 to 8 wt%H_2 adsorption is possible at 300 K, which for a bulk nanotube density of approximately 1 g/cm^3 would equate to intrinsic maximum storage densities of 40 to 80 kg H_2/m^3.

Current understanding of H_2 adsorption within carbon nanotubes (or other engineered adsorbents) is embryonic. Theoretical calculations and simulation results vary depending on methodology and assumptions. The reality of H_2 adsorption within a nanotube array may be very complex, depending on tube diameter distributions, interstitial sites, nanotube electronic and structural properties, defects, and changes caused by adsorbed H_2 molecules. The ideal adsorbent would cause H_2 molecules to adsorb, order, and condense at room temperature, akin to a phase change, representing the ultimate H_2 storage technology. Whether this ideal is possible within the theoretical limits of H_2 adsorption or the capabilities of economically fabricated carbon nanotube adsorbents needs to be known before their ultimate H_2 storage potential can be evaluated.

9. HYDROGEN STORAGE IN CHEMICAL CARRIERS

In addition to hydride and adsorbent storage, H_2 can be stored by chemical absorption in liquids. Aromatic molecules with carbon–carbon double bonds are the leading candidates for liquid chemical hydrogen storage. For example, methylcyclohexane (C_7H_{14}) releases three H_2 molecules when heated and becomes toluene (C_7H_8). Using this reaction, a theoretical 6.16 wt% H_2 can be stored at a density of 49 kg H_2/m^3. However, significant thermal energy (a theoretical minimum of $\sim$70 kJ/mol H_2) at temperatures of 650 K or higher is needed to release this H_2. Therefore, H_2 storage in aromatic liquids is probably best suited to compete with LH_2 for stationary long-term storage of H_2 in bulk, where the energy and capital costs of H_2 release are offset by the very low capital costs of storing liquid chemicals.

Methanol (CH_3OH) is a liquid chemical H_2 carrier containing 12.5 wt% H_2 at high density (100 kg H_2/m^3). H_2 accounts for 75% of the fuel value (19.92 MJ/kg) of CH_3OH, which is produced on a large scale with efficiencies as high as 75% from H_2/CO/CO_2 mixtures derived from natural gas. CH_3OH can be decomposed to H_2 and either CO or CO_2 (with H_2O). Both reaction pathways are very endothermic (65 and 40 kJ/mol H_2, respectively), making H_2 release from CH_3OH energy intensive unless waste heat is available to vaporize the methanol and/or steam. The CO_2 reaction product, toxicity, and high solubility in (ground) water cloud the ecological attractiveness of CH_3OH as an H_2 carrier. Methanol synthesis could be CO_2 neutral if atmospheric CO_2 or perhaps surplus biomass were used as the necessary carbonaceous feedstock.

NH_3 is the most attractive carbonless liquid chemical H_2 carrier. NH_3 boils at 240 K but can be condensed at 300 K under moderate pressure (10 atm) containing 17.6 wt% H_2 with an atomic hydrogen density of 105 kg H_2/m^3—150% denser than LH_2. Decomposing NH_3 into H_2 (and nitrogen [N_2]) is thermodynamically easier (30 kJ/mol H_2) than other liquid chemical H_2 carriers but is kinetically complicated by the N_2 triple bond. Rapid NH_3 synthesis is currently practical only at elevated temperatures (and pressures) in large plants (1000 tonnes [metric tons] NH_3/day) achieving efficiencies of 75 to 80% from CH_4.

Higher efficiencies are possible using atmospheric N_2 and nonfossil H_2 sources because NH_3 formation is exothermic. NH_3 has been proposed as an alternative fuel in its own right but has two major drawbacks: chemical reactivity with certain metals and toxicity. Although NH_3 can be lethal in concentrations of 1000 to 5000 ppm, it is easily detectable by smell at less than 20 ppm and is lighter than air. NH_3-based H_2 storage onboard vehicles would likely require avoiding any NH_3 emissions through nearly perfect decomposition to N_2 and H_2. NH_3 may hold greater immediate promise for H_2 distribution and transportation to refueling stations, avoiding the lower energy density, higher capital cost, and higher energy intensity of LH_2 transport, distribution, and storage.

10. FUTURE PROSPECTS FOR HYDROGEN IN TRANSPORTATION

Unlike other fuels, hydrogen possesses a broad array of possible storage approaches and attendant vehicle and refueling infrastructure implications. Each differs dramatically in capital and energy intensity, maturity, and perceived or potential business and safety risks. For some applications, a given H_2 storage approach is clearly best. For example, LH_2 storage is clearly the only option for H_2 aircraft,

especially those designed for intercontinental and perhaps future hypersonic air travel. Economies of scale and infrequent refueling argue for the use of LH_2 in ships and trains as well.

Since H_2 fuel began receiving serious interest in the 1970s, technological advances have weakened the case for LH_2 in light-duty vehicles and perhaps even in large trucks. Hybrid electric vehicles have been commercially introduced, and at the time of this writing, it is quite possible that fuel cell vehicles will be introduced before 2005. Either technology allows the production of H_2 automobiles that achieve fuel economies equivalent to 60 to 100 mpg (25–42 km/L). This leap in fuel economy permits H_2 vehicles using as little as 5 kg H_2 onboard storage to achieve a 300- to 500-mile driving range. However, these fuel economy improvements would be insufficient to weaken the case for LH_2 without the tripling in strength of lightweight composite materials since the 1970s. If future higher strength materials ($\sim$ 7 GPa or 1 million psi) become routinely available, compressed H_2 may rival the density of LH_2 storage.

Flexibly refueled cryogenic pressure vessels would permit light-duty vehicles to compromise between LH_2 and GH_2 and be compatible with a multiple-mode refueling infrastructure: centralized LH_2 production and delivery to stations alongside a spectrum of on-site GH_2 production technologies at stations and perhaps garage electrolysis.

Decentralized H_2 production, and particularly slow (e.g., overnight) refueling, would be ideal for metal hydride storage, easing refueling times and reducing electrolysis pressures. The lighter higher temperature hydrides (e.g., $NaAlH_4$) are not yet sufficiently developed for vehicles, but the heavier ambient temperature hydrides (e.g., $TiFeH_2$) will be increasingly feasible as H_2 vehicle fuel economy increases. In addition, electrolytic garage refueling of onboard hydride storage, or perhaps moderate-pressure GH_2 storage, would conceivably permit early H_2 light-duty vehicles to circumvent the greatest strategic barrier facing any alternative fuel vehicle: the initial lack of refueling infrastructure.

The ultimate shape of the refueling infrastructure and hydrogen energy economy may depend on whether LH_2 or GH_2 is the dominant fuel for future large commercial trucks. If gaseous hydrogen and/or hydride storage turns out to be sufficient for light-duty vehicles, the choice of hydrogen storage for trucks could determine whether LH_2 will be an everyday road vehicle fuel or principally used only in aircraft. The use of LH_2 would favor a centralized H_2 infrastructure for efficient liquefaction, whereas a GH_2 infrastructure would very likely have lower ultimate costs due to reduced energy intensity. A GH_2 infrastructure could also be more easily decentralized, relying on electrolysis at filling stations, in garages, or ultimately onboard passenger vehicles.

The factors determining the onboard H_2 storage technology best suited for commercial heavy trucks are very different from those for light-duty vehicles. A future H_2-fueled light-duty vehicle will refuel with 4 to 5 kg H_2 approximately 40 times annually. A liquefaction energy cost premium of \$0.50/kg for LH_2 would amount to only \$80 to \$100/year for such vehicles.

H_2-powered commercial heavy trucks will consume 100 to 200 times as much fuel as will light-duty vehicles, with the former refueling with 75 to 100 kg of H_2 daily and being much more sensitive to fuel costs. Potential GH_2 fuel cost savings of as little as \$0.25/kg relative to LH_2 would probably justify large investments (\$30,000–\$40,000) in onboard GH_2 storage for trucks. However, large amounts (100 kg) of high-pressure GH_2 storage on heavy trucks would benefit little from economies of scale and could raise overall safety and volume concerns. Composite vessels pressurized with 700 to 1000 atm GH_2 would be 6.0 to 7.5 times larger than equivalent energy 100-gallon (380-L) gasoline tanks.

For heavy trucks, cryogenic pressure vessels may best balance hydrogen storage volume, capital costs of onboard storage, and energy intensity-related fuel costs. Safety issues of cryogenic vessels could be eased by the substantially reduced maximum mechanical energy release at cryogenic temperatures. Dormancy, boil-off, and insulation requirements for cryogenic vessels would be minimal for heavy trucks. Cryogenic vessels store H_2 at approximately twice the density of ambient vessels at the same pressure, or they can achieve comparable H_2 storage density with far lower pressures, vessel wall strength, and likely capital costs. Finally, although the energy intensity of an 80 K H_2 infrastructure has yet to be demonstrated, its theoretical energy intensity is significantly lower than that of LH_2 (Fig. 4) and likely easier to approach than is the case with current H_2 liquefaction technology (10–14 kWh/kg LH_2).

If developed successfully, $NaAlH_4$ would allow hydride storage for heavy trucks with moderate desorption temperatures. However, a 100-kg H_2 storage system for a heavy truck would still weigh

approximately 3000 kg (six times heavier than two full 80-gallon diesel tanks currently used on heavy trucks).

Whichever blend of hydrogen storage technologies ultimately develops for both light-duty and heavier vehicles, it is likely that the most critical characteristic over the long term will be the flexibility, versatility, and efficiency of the hydrogen production and utilization infrastructure each storage technology permits. Hydrogen storage technologies with greater flexibility (cryogenic pressure vessels, hydrides, and perhaps engineered adsorbents) offer benefits that may be significant over the long run by permitting an H_2 production, storage, and utilization infrastructure that easily adapts.

A versatile hydrogen infrastructure that can benefit from or even encourage technological change (e.g., scale efficiencies, vehicle fuel economies), dynamically balance energetic and capital costs, and offer intangible benefits (e.g., energy infrastructure security, vehicle safety, sustainability) may ultimately be of greater value than the sum of its individual H_2 production and storage components.

SEE ALSO THE FOLLOWING ARTICLES

Coal Storage and Transportation • Hydrogen, End Uses and Economics • Hydrogen, History of • Hydrogen Production • Natural Gas Transportation and Storage • Storage of Energy, Overview.

Further Reading

Berry, G., and Lamont, A. (2002). Carbonless transportation and energy storage in future energy systems. *In* "Innovative Energy Strategies for CO_2 Stabilization" (R. Watts, Ed.), pp. 181–210. Cambridge University Press, Cambridge UK.

Cox, K., and Williamson, K. (1977). "Hydrogen: Its Technology and Implications Vols. 1–5." CRC Press, Cleveland, OH.

Hoffmann, P. (2001). "Tomorrow's Energy." MIT Press, Cambridge, MA.

Ogden, J., and Williams, R. (1989). "Solar Hydrogen: Moving Beyond Fossil Fuels." World Resources Institute, Washington, DC.

Peschka, W. (1992). "Liquid Hydrogen." Springer-Verlag, New York.

Sandrock, G., Suda, S., and Schlapbach, L. (1992). Applications. *In* "Hydrogen in Intermetallic Compounds II" (L. Schlapbach, Ed.), pp. 197–246. Springer-Verlag, New York.

Winter, C. J., and Nitsch, J. (1988). "Hydrogen as an Energy Carrier." Springer-Verlag, New York.

Hydropower Economics

BRIAN K. EDWARDS
Brian K. Edwards Associates
Downers Grove, Illinois, United States

1. Introduction
2. Development and Regulation of Hydroelectric Power in the United States
3. Policy Issues Relating to Dams and Hydroelectric Power
4. Generating Hydroelectric Power
5. Optimal Provision of Hydroelectric Power
6. Summary and Conclusions

Glossary

dam head The distance between the elevation of a reservoir and the top of the penstock.
elevation The number of feet the surface of a reservoir is above sea level.
forebay Where water is stored behind a dam prior to release for generation.
hydroelectric power The electric power generated by the flow of water.
hydro-shifting The practice of storing water behind a dam during off-peak hours for release during on-peak hours.
penstock A tube through which water flows, from the afterbay of a dam to the turbines, for generation.

With few exceptions, the academic research relating to hydroelectric power has come from the power systems (engineering) literature. Some published works have addressed hydroelectric power from an economics perspective, but many of these articles have addressed environmental issues associated with water use; these have included valuing damages to ecosystems and valuing the recreational benefits of water resources, and some have estimated the benefits of changes in downstream river flows sometimes attributable to changes in how dams are operated. In short, issues relating to dam operations have largely been relegated to the periphery of discussions of other, though certainly not unimportant, topics. Thus, the focus of this article will be on the use of water to generate hydroelectric power, by presenting a simple model of how a dam operator decides to use the water resource to generate that power.

1. INTRODUCTION

One of the difficulties associated with any economic analysis of hydroelectric power is how widely the functions of dams vary, not only within individual countries, but between countries. Studies by the International Commission on Large Dams and the World Commission on Dams reveal a great diversity in dam use. For example, one-third of all countries depend on hydropower for over one-half of their electricity. One-third of European dams are used to generate hydroelectric power and another 20% are used for irrigation. Nearly two-thirds of Asian dams are used for irrigation, but only 7% are used to generate hydroelectric power. In North and Central America, dam use is almost equally divided between irrigation, hydroelectric power, water supply, flood control, and recreation functions. In the United States, hydroelectric generation accounts for approximately 10% of total electric generation. In South America, nearly one-fourth of the dams are used for hydroelectric generation. Nearly three-fourths of African dams are used for irrigation and water supply. In Australasia, nearly 50% of the dams are used for water supply and another 20% are used to generate hydroelectric power.

These statistics suggest that the author of any article on dams and hydroelectric power must take care to avoid making too general an analysis of dams, or at least must provide adequate warning to readers expecting a complete and self-contained presentation on dams and hydroelectric power, covering every possible area of interest. Thus, the primary focus here is on the use of water to generate hydroelectric power; a simple model of a dam operator is presented in that regard, and other issues are ignored, i.e., those

 © 2004 Elsevier Inc. All rights reserved.

relating to operational restrictions that limit the rate at which water can be released, hydrothermal generation (i.e., the joint provision of hydroelectric power with other kinds of electrical power), hydraulically coupled dams (i.e., dams connected by a common river), and the recreational impacts of dam operations. However, some of the analysis presented here will relate to these other uses of water, but only marginally. For example, the operating restrictions often imposed on dam operators are motivated by environmental, recreational, and water use considerations. However, there is no attempt to integrate these other uses of water explicitly into this analysis. Nevertheless, the model presented here could be extended to consider alternative water uses explicitly, and these extensions would indeed represent fruitful extensions of the work presented herein. For example, the model could be extended to include explicit agricultural or recreational uses of the water that supplement the treatment of the water resource as an input in electricity generation. There could also be an attempt to integrate the benefits and costs of hydroelectric power generation into a broader analysis that also considers the recreational benefits of alternative flow regimes explicitly. In this case, trade-offs between using water to generate hydroelectric power and the benefits of imposing restrictions on changes in water release rates to improve the value of downstream recreational activities could be examined.

A brief legal history of hydroelectric power with particular emphasis on the United States is presented in Section 2. An outline of some of the energy, environmental, and economic development issues surrounding dams and hydroelectric power is presented in Section 3. In Section 4, a simple economic model of hydroelectric power generation is presented. In this model, a decision maker will decide on the hourly pattern of water releases that maximizes net receipts subject to an equation of motion that describes the dynamics of water supply.

2. *DEVELOPMENT AND REGULATION OF HYDROELECTRIC POWER IN THE UNITED STATES*

Although the development and regulation of hydroelectric power has a long history of federal oversight, the demand for and supply of such power have largely been regulated at the state and local levels. Unless a dam has been constructed specifically for federal purposes such as flood control or navigation, hydroelectric power generation has to comply with state proprietary water laws. Under the Federal Pollution Control Act, as amended by the Clean Water Act Amendments of 1977 (Clean Water Act), unless state and local environmental laws conflict with federal law, hydroelectric projects must also comply with state and local environmental laws.

In the beginning of the 20th century, most hydroelectric power development was private, developed primarily through holding companies, so that by 1916, most of the 87 hydroelectric projects in the United States were in private hands. During this period, the federal government increased its role in hydroelectric power development primarily through federal reclamation projects by the United States Bureau of Reclamation (BOR) and water resource projects by the United States Army Corps of Engineers (Corps). Over the years, the Corps has been the largest single producer of hydroelectric power in the United States, and currently operates 75 dams with an installed capacity of over 21,000 MW.

The BOR was created by the Reclamation Act of 1902. This act authorized the Secretary of the Interior to develop irrigation and hydropower projects in 17 western states. The increased involvement by the BOR was a natural consequence of the need to manage scarce water supplies in the western United States. Despite much of the emphasis on hydroelectric power generation, the development of dams resulted from the need to manage water supplies effectively in a predominantly arid region. Moreover, hydroelectric power was often used during the construction phases of these projects to provide electrical power for the processing of materials and the running of sawmills, concrete plants, and construction equipment, and allowed for operation of construction activities during the night. After construction was finished, hydroelectricity was used to power drainage pumps that moved water to higher elevations than was possible with gravity-flow canals. Many of these hydroelectric power facilities found that they had surplus power, thus it was necessary to establish the terms under which such surplus power could be sold. To that end, the Towns Sites and Development Act (1906) gave the Secretary of the Interior authority to lease surplus power or power privileges. As a result, surplus power was sold to existing power distribution systems, benefiting local industries, towns, and farming communities.

Indeed, the Bureau of Reclamation's first hydroelectric facility was built to support construction of the Theodore Roosevelt Dam, located on the Salt River northwest of Phoenix, Arizona. Although the

primary purpose of the hydroelectric power was to power construction equipment, the surplus power generated by this plant was sold to the local community. Public support for additional power led to an expansion of the capacity of this dam, so that in 1909 five generators were placed in operation, providing power for irrigated agriculture and the Phoenix area. This expansion of hydroelectric power encouraged the economic growth of the Phoenix area, including expanding electrically irrigated agriculture to cover more than 10,000 acres, and providing all of the residential and commercial power requirements of the area.

In 1920, the Federal Power Act established the Federal Power Commission (FPC), giving them authority to issue licenses for developing hydroelectric projects in public lands and on navigable waters. In 1928, the Boulder Canyon Project Act authorized construction of the Hoover Dam. Even though initiated before the New Deal, the construction of the Hoover dam has come to symbolize many of the large-scale public works projects that came to characterize the New Deal. Until the 1930s, the FPC took a fairly *laissez-faire* approach to requests for licenses for power projects, granting most of these requests with little federal review or oversight. It was not until the administration of President Franklin Roosevelt that the role that the federal government currently plays would be established. The Public Utility Act of 1935 expanded the FPC's jurisdiction, which hitherto had been limited to licensing hydroelectric generation, to include setting wholesale electricity rates. Under the Public Utility Act, however, regulation of consumer rates remained with state public utility commissions. The administration of Franklin Roosevelt also ushered in the "big dam" period, which, in addition to the Hoover Dam, saw the construction of many other large multipurpose hydroelectric projects, including the Grand Coulee Dam on the Columbia River in Washington state (authorized by the Rivers and Harbors Act of 1935) and the Central Valley Project in California. The Tennessee Valley Authority (TVA) was created in 1933 and the Bonneville Project Act of 1937 created the Bonneville Power Administration (BPA). The construction of the dams had much to do with the migration to the west that occurred during the 1930s. During World War II, cheap hydroelectric power encouraged firms involved in the defense effort to locate in the west.

From the legal developments, two federal policies emerged: (1) primary federal regulation of the construction and use of private hydroelectric projects and (2) federal construction and control of the largest such projects. With the exception of increased use of federal subsidies and tax incentives designed to encourage small hydroelectric power development at both existing and new facilities, these two policies remain in effect.

Even in the 1930s, the environmental impacts of hydroelectric power generation, including the lowering of water tables brought on by the rapid growth of irrigated agriculture, began to be recognized and would eventually lead to legislation that addressed these environmental impacts. Most of the early legislation laid the groundwork for the construction and operation of hydroelectricity projects, but it would not be until 1968, with the passage of the Wild and Scenic Rivers Act, that these environmental concerns would begin to be addressed. That Act protected rivers in their natural state by excluding them from consideration as hydroelectric power generation sites. One year later, the National Environmental Policy Act (NEPA) required federal agencies to take into account environmental considerations in the construction and operation of hydroelectricity generating facilities. In 1973, the Endangered Species Act listed endangered species and their critical habitats, many of which are directly affected by the construction and operation of dams. The Fish and Wildlife Coordination Act was first amended in 1946, to ensure that fish and wildlife would receive equal consideration for protection by federal agencies. Sections 401 and 404 of the Clean Water Act ensured that federal and nonfederal entities would comply with state water quality standards by requiring state certification as a condition to Federal Energy Regulatory Commission (FERC) license approval.

Other regulations addressed the financing and administration of dams. The Federal Water Power Act of 1920 regulated hydroelectric development of navigable waterways. The Reclamation Project Act of 1939 extended the contract term to 40 years for sale of power or lease of power privileges, giving preference to public entities. The Flood Control Act of 1944 gave the Secretary of the Interior authority to market power from Army Corps of Engineers projects and authorized the Pick–Sloan Missouri Basin Program. The Federal Columbia River Transmission Act of 1974 authorized the Bonneville Power Administration to issue revenue bonds.

In 1977, the United States Department of Energy was created by the Department of Energy Organization Act. This Act transferred the existing Power Marketing Administrations to the newly created

Department of Energy and created the Western Area Power Administration. The Public Utility Regulatory Policies Act (PURPA) of 1978 encouraged small-scale power production facilities; exempted certain hydroelectric projects from federal licensing requirements, and required utilities to purchase—at "avoided cost" rates—power from small production facilities that use renewable resources.

3. POLICY ISSUES RELATING TO DAMS AND HYDROELECTRIC POWER

The past few years have seen many changes in the energy industries in the United States and other countries. The British energy industry has already undergone deregulation. Deregulation in the United States, at least at the retail level, has largely proceeded on a state-by-state basis, and wholesale bulk power markets are in an advanced stage of deregulation. The hopes of deregulation are to increase competition and eventually lower energy prices to consumers, but also to allow pricing signals to direct energy resources to their most efficient use and provide incentives to users to conserve energy.

However, there is much more at stake than economic efficiency. Because energy and air pollution are often joint products, socially unacceptable levels of air pollution can result regardless of whether markets are used to ration electricity. Proponents of hydroelectric power point out that hydro generation produces no emissions of acid rain or ozone precursors, and dams have also not created the sort of not-in-my-backyard (NIMBY) concerns that have stigmatized nuclear power facilities and proposed nuclear waste storage facilities.

Despite the potential for displacing emissions, however, hydroelectric facility construction is virtually irreversible. Although some smaller dams have been torn down, it is difficult to fathom how the Hoover Dam could be decommissioned and taken apart without causing energy and economic disruptions in the American West. Moreover, dam construction alone can change the environment dramatically, even before any hydroelectric power is generated, the Xiaolangdi and Three Gorges dams in China being two recent cases in point.

Once built, dams can adjust their rate of generation quickly, but this very flexibility causes rapid changes in the downstream flow of rivers, which can erode downstream sandbanks, change the deposition of downstream river sediment, thereby altering the ecological makeup of the downstream environment, and degrade the quality of downstream recreational activities such as fishing. Unfortunately, reducing these fluctuations can cause other environmental problems by mitigating the positive effects of annual spring floods. These floods can replenish downstream sandbanks eroded by the river flows over the course of the year. The United States Bureau of Reclamation has been experimenting with periodic sustained high releases from Glen Canyon Dam to mimic the spring flooding that used to occur there naturally. The argument that hydroelectric power displaces air pollution precursors has been challenged. Storage dams can become silted with vegetation that can rot, emitting carbon dioxide and methane, two greenhouse gases. These increased greenhouse gas emissions could, in some cases, offset any fossil fuel emission reductions from thermal generation.

The salmon populations that Lewis and Clark observed in the Columbia River in abundance during the early years of the 19th century have fallen dramatically since then, perhaps a result of the current system of dams on that river. Hydroelectric power threatens fish species throughout the world by disrupting the movement of many fish species to their spawning sites. In 2000, more than 1500 fishermen and farmers occupied the Pak Mun hydroelectric dam in the rural northeast of Thailand, demanding that the floodgates of these dams be opened and the dams be demolished. Stocks of sturgeon, striped bass, Atlantic salmon, and alewives have increased, water quality has improved, and more animals, plants, and birds now occupy the banks of the Kennebec River in Maine since the bulldozing of the Edwards Dam in 1999. In the United States, over 40 dams have either been removed, or slated for removal, and dam-removal advocates have recently pointed their attention at the Snake River, a main tributary of the Columbia River, where four dams have virtually stopped the flow of salmon.

Dams can also have unforeseen impacts even when they accomplish their intended purposes. Flood control dams might encourage people to live closer to rivers, believing that the risks of flooding are mitigated by the dam. This can be catastrophic, as demonstrated in 1993, when floods in Mississippi overwhelmed the dam system, causing billions of dollars in damages to communities along that river. Presumably, these damages would have been less severe had people chosen to locate further from the river. To the extent that irrigation dams are successful, they can also encourage the wasteful use

of water, as evidenced by farmers switching from drip and other less water-using irrigation methods to more wasteful spray methods of irrigation. To the extent that hydroelectric power is sold at below-market rates, as is done in many parts of the United States, electrical irrigation methods are encouraged, possibly inducing farmers to specialize in more water-intensive crops, and possibly leading to more rapid depletion of aquifers. What else can account for the high share of American cotton grown in Arizona, or, for that matter, the preponderance of golf courses in the Phoenix area?

Hydroelectric power is often touted as a reliable source of electricity, but its availability often depends on the whims of nature. Years of low rainfall in Brazil, a country that depends on hydroelectric power for more than 90% of its electricity needs, have lead Brazil to approve the licensing of more than 20 new gas-fired generating plants and to consider the construction of additional nuclear capacity. In the meantime, Brazil faces the possibility of rolling blackouts to bring electricity consumption in line with their generating capacity. Shortages of water in the western United States have revealed weaknesses in how the state of California deregulated its power industry. Sufficient water might have allowed enough power to be generated to prevent, or at least forestall, the energy problems the state has faced.

4. GENERATING HYDROELECTRIC POWER

Water is stored behind a dam in what is called a forebay. The quantity of water in the forebay can be expressed in terms of acre-feet of volume, but can also be measured in terms of elevation, which refers to the number of feet the surface of the reservoir is above sea level. The elevation of the reservoir will determine an important attribute of a dam that has a direct bearing on the amount of hydroelectric power than can be generated, namely, the head of the dam, which refers to the distance between the elevation of the reservoir and the top of the penstock, which is a tube through which water flows to the generator. The higher the head, the faster the water will fall through the penstocks, and hence the greater the amount of kinetic energy that the generator will eventually convert into electrical energy. The technical relationship between water release, dam head, and generation can be expressed as $q = (\text{CFS} \times \text{head})/\alpha$, where q is hydroelectric output, CFS is the water release rate (measured in cubic feet or cubic meters per second), and head is the head of the dam (usually measured in feet or meters). The coefficient α is a factor that converts the product of water releases and dam head to hydroelectric generation. For example, if we use a figure of $\alpha = 8.8$ and release water at a rate of 5000 CFS from a dam with a 400-foot head, we generate 22,773 hp of electricity. Dividing this figure by 746 and converting (from watts to megawatts) yields 170 MW of hydroelectric generation.

To facilitate integrating the technical relationship between these factors into hydroelectric generation for use in the models that follow, we combine the relationship between elevation, content, and head into the following simpler equation that we will use as our production function for hydroelectric generation:

$$q_t^h = q_t^h(r_t,\ W_t). \tag{1}$$

In this equation, hydroelectric power, q_t^h, is generated in period t by releasing water at a rate given by r_t. The argument for reservoir content in the production function, W_t, will capture the combined effects that changes in content have on elevation, which in turn influences dam head, which finally influences the productivity of water releases. If we were to take a more formal approach, we would define the hydroelectric generation production function as $q_t^h = \tilde{q}_t^h\{r_t,\ h_t[e_t(W_t)]\}$. In this equation, we define head as a function of elevation, $h_t = h_t(e_t)$, and elevation as a function of content, $e_t = e_t(W_t)$. We derive Eq. (1) by successively substituting the head and elevation equations. Nevertheless, we assume that Eq. (1) is continuously differentiable in both arguments, that both first derivatives are positive, and that both second derivatives are negative. Accordingly, hydroelectric generation will be higher, for a given dam content, the higher the rate of water release. We also assume that hydroelectric generation will be higher, the higher dam content, for a given rate of water release.

5. OPTIMAL PROVISION OF HYDROELECTRIC POWER

Our simple model of hydroelectric generation assumes that the hydroelectric power dispatcher generates power and sells it in a large market, with the hydroelectric power sold in a market at a price given by the inverse demand function $p_t(Q_t)$, where Q_t represents the quantity of electricity sold during period t. We express industry output as $Q_t = \tilde{Q}_t + q_t^h(r_t, W_t)$, where, $\tilde{Q}_t$ represents power sold by the other generators or energy marketers serving this particular market. Water release costs are given by

$c_t^h(r_t)$which we assume to be continuously differentiable with positive first and second derivatives. Accordingly, profits in each period are given by the following equation:

$$\pi_t = p_t(Q_t)q_t^h(r_t,\ W_t) - c_t^h(r_t). \tag{2}$$

The only constraints that the hydroelectric power operator faces in our simple model is an equation of motion, $W_t = W_{t-1} - r_{t-1} + f_{t-1}$, which has reservoir content determined by water releases and inflows into the reservoir (f_{t-1}), and a requirement that the dispatcher operate within the operating characteristics of the generator. For now, we will assume that the dispatcher satisfies this constraint implicitly.

The dispatcher will choose rates of water release to maximize the discounted present value of profits subject to the equation of motion over t periods $(t = 1, \ldots T)$. We use a finite-time horizon because dam operations are often subject to operating plans promulgated by regulatory agencies that specify water release and other requirements over annual, monthly, and other finite time horizons. Nevertheless, the resulting discrete-time Hamiltonian function is given by

$$\begin{aligned} H = \sum_{t=1}^{T-1} & [p_t(Q_t)q_t^h(r_t, W_t) - c_t^h(r_t) \\ & + \gamma_{t+1}(W_{t+1} - W_t + r_t - f_t)] \\ & + P_T(Q_T)q_T^h(r_T, W_T) - c_T^h(r_T) \\ & + \gamma_T(W_T + r_T - f_T). \end{aligned} \tag{3}$$

In Eq. (3), we define γ_{t+1}as the discrete-time costate associated with the equation of motion and define γ_T as the final period costate on the equation of motion.

The first-order conditions on the control variable are given by

$$\frac{\partial H}{\partial r_t} = p_t\frac{\partial q_t^h}{\partial r_t} + \frac{\partial p_t}{\partial q_t^h}\frac{\partial q_t^h}{\partial r_t}q_t^h - \frac{\partial c_t^h}{\partial r_t} - \gamma_{t+1} \leq 0 \tag{4}$$

$$r_t\frac{\partial H}{\partial r_t} = 0. \tag{5}$$

Equations (4) and (5) indicate the conditions that govern hourly water release rates. Equation (4) can be rearranged to form the following equation:

$$\frac{\partial H}{\partial r_t} = p_t\frac{\partial q_t^h}{\partial r_t} + \frac{\partial p_t}{\partial q_t^h}\frac{\partial q_t^h}{\partial r_t}q_t^h \leq \frac{\partial c_t^h}{\partial r_t} + \gamma_{t+1}. \tag{6}$$

The first part of the right-hand side of Eq. (6) is the marginal benefit of releasing water in period t $(p_t\partial q_t^h/\partial r_t + (\partial p_t/\partial q_t^h)(\partial q_t^h/\partial r_t)q_t^h)$. The second part of the right-hand side of Eq. (6) is the marginal cost of releasing water in period t $(\partial c_t^h/\partial r_t + \gamma_{t+1})$ and includes two components. The first part $(\partial c_t^h/\partial r_t)$ is interpreted as the marginal cost of releasing water. The second part (γ_{t+1}) is the shadow price of water and captures the impact that releasing water now has on future net benefits. We include it in the marginal cost component of Eqs. (4)–(6) because it reflects the opportunity cost of using the water resource to generate hydroelectric power in the current period.

Nevertheless, Eqs. (4) and (5) are the Kuhn–Tucker necessary conditions for an optimum and can be interpreted as follows. If Eq. (4) is strictly negative, and the marginal benefits of releasing water in period t are less than the marginal costs, Eq. (5) will then hold with equality and water releases in that period will be zero. Otherwise, water is released at a rate up to the point where Eq. (4) holds with equality, and will still satisfy Eq. (5). We can take the first part of Eq. (4) and rearrange it into terms more familiar in microeconomics:

$$P_t\frac{\partial q_t^h}{\partial r_t} + \frac{\partial p_t}{\partial q_t^h}\frac{\partial q_t^h}{\partial r_t}q_t^h = p_t\frac{\partial q_t^h}{\partial r_t}\left(1 + \frac{1}{\eta_t}\right), \tag{7}$$

where η_t is the elasticity of demand for electricity. This allows us to express the right-hand side of Eq. (7) as $MR_t(1 + 1/\eta_t)$. On making this substitution, and assuming an interior solution, our first-order condition becomes

$$p_t\frac{\partial q_t^h}{\partial r_t}\left(1 + \frac{1}{\eta_t}\right) = \frac{\partial c_t^h}{\partial r_t} + \gamma_{t+1}. \tag{8}$$

According to this equation, the operator releases water at that rate which equates the marginal benefit of releasing water to the marginal cost of releasing water.

In this model, water has two values. The first, which we call the flow value, captures the role that plays a direct input in hydroelectric power generation. The second, which we call the stock value, reflects the effect that stored water has on the generating efficiency of the dam in future periods through effects that storing additional water have on the head of the dam. *Ceteris paribus,* higher release rates in the current period increase hydroelectric generation in the current period, but reduce the potential future reservoir elevation, thereby reducing the generating efficiency of future releases.

For a given reservoir size, releasing water in the current period increases revenue by $p_t\{\partial q_t^h(r_t, W_t)/\partial r_t\}$ and costs the hydroelectric power dispatcher $\partial c_t^h/\partial r_t$. The net impact of releasing water in the current period is $p_t\{\partial q_t^h(r_t, W_t)/\partial r_t\} - \partial c_t^h/\partial r_t$.

However, direct water release costs are invariant with respect to reservoir elevations, so additional water in the reservoir provides additional generation without incurring higher operating expenses. If we differentiate Eq. (3) with respect to the reservoir content, we obtain an equation that describes the dynamic conditions governing the change in the marginal cost of water.

$$\frac{\partial H}{\partial W_t} = p_t \frac{\partial q_t^h}{\partial W_t} + \frac{\partial p_t}{\partial q_t^h}\frac{\partial q_t^h}{\partial W_t} q_t^h + \gamma_{t+1} - \gamma_t = 0. \quad (9)$$

We can use the previous result involving the elasticity of demand to restate Eq. (9) as

$$p_t \frac{\partial q_t^h}{\partial W_t}\left(1 + \frac{1}{\eta}\right) = \gamma_{t+1} - \gamma_t = 0. \quad (10)$$

In Eq. (10), the $p_t(\partial q_t^h/\partial W_t)(1 + 1/\eta)$ term represents the marginal benefit of additional stored water measured in terms of the value of additional hydroelectric power generation for a given rate of water release. The term defining the change in the shadow price of the water, $\gamma_{t+1} - \gamma_t$, reflects the net contribution of the water in future periods. Rearranging Eq. (10) as a difference equation (for $\Delta\gamma = \gamma_{t+1} - \gamma_t$) and solving it yields Eq. (11), which expresses the current-period shadow value of water in terms of the value that the stored water has on remaining future period revenues, less thecost of the additional hydroelectric generation in these periods.

$$\tilde{y}_t = \sum_{t=1}^{T-1} p_t \frac{\partial q_t^h}{\partial W_t}\left(1 + \frac{1}{\eta_t}\right) + p_T \frac{\partial q_T^h}{\partial W_T}\left(1 + \frac{1}{\eta_T}\right). \quad (11)$$

According to this equation, the benefits of storing water will decrease over time. In the last period, water has value only for its generating value in that period. We also see that γ_t is the sum of these benefits for the remaining periods, which decline as we approach the end of the time horizon.

We now present a simulation example of a simple version of the above model. In order to implement the simulation, we will make a number of simplifying assumptions. First, we will run the simulation for 24 periods. We will also assume that spot electricity prices are determined exogenously and vary over the course of the simulation. We can easily think of the 24 periods as hours, and so we can easily imagine that our "day" is characterized by periods of lower demand (and hence lower spot prices) and periods of higher demand (and hence higher spot prices). The prices will be expressed in terms of dollars per megawatt and will follow the time path given in Fig. 1. According to Fig. 1, prices are low during the early and late hours of the day. We can easily think of these as off-peak hours. As the hours progress, however, prices rise until they reach their highest point sometime during the middle of the day. We easily think of these as on-peak hours. We will also assume that generating hydroelectric power costs a constant \$15/megawatt to generate.

We assume that the reservoir begins with 120,000 acre-feet of content and that hourly inflows fluctuate randomly at a rate between 200 and 800 CFS. Generation will occur according to $q_t^h = 0.0005 r_t h_t$ (where h_t is the head of the dam in period t). We also assume that dam head depends on content, according to $h_t = 0.002 W_t$. We assume finally that the operator must release a total of 2000 acre-feet of water over the 24 periods. The operator will choose rates of water release in each of the 24 periods to maximize the sum of net receipts subject to the equation of motion. For our simulation, we will use a discount rate of 5% (0.05). Finally, we will assume that generator capacity limits hourly release rates to 1500 CFS (a small dam!).

In Fig. 2, we graph hydroelectric generation (which is measured along the left-vertical axis) in each of the 24 periods. Along the right-vertical axis, we measure the rate of water release in each of the 24

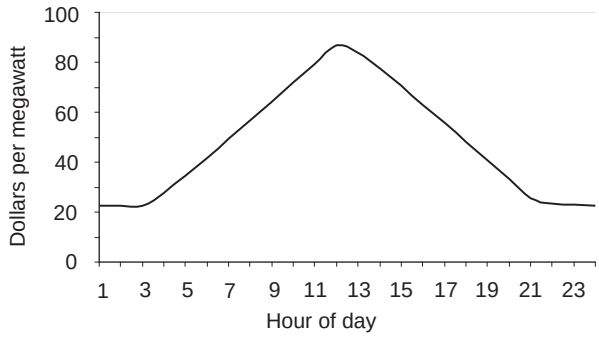


FIGURE 1 Hypothetical spot electricity prices for simulation.

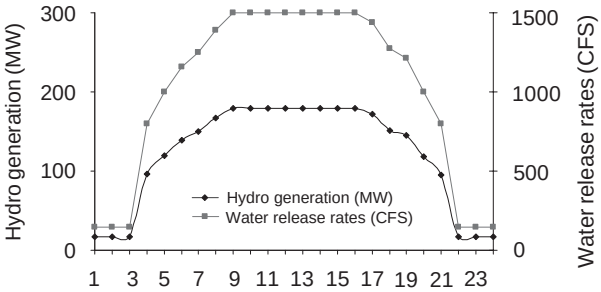


FIGURE 2 Water releases and hydroelectric generation. CFS, cubic feet/second.

periods. Because generation and transmission costs are constant throughout the day, but spot prices rise during the on-peak hours, it should come as no surprise that release rates (and generation) follow the pattern displayed by prices. From the perspective of power generation, this behavior typifies load following, whereby generation follows demand (as reflected in the higher on-peak period prices). In terms of hourly water release patterns, this illustrates what we term hydro-shifting, which refers to storing water behind the dam during off-peak hours and releasing the water for generation during on-peak hours.

6. SUMMARY AND CONCLUSIONS

The preceding discussion has presented a brief introduction to the economics of dams and hydroelectric power. Although the models and their results are simple, such a simplified approach represents a foundation for subsequent research and analysis of more complex issues involving dams and hydroelectric power. These models also couch the entire discussion of hydroelectric power firmly in the natural resource economics, and more specifically, in the water management domain, one that reflects the view of the author that operating dams and generating hydroelectric power require decisions involving how best to manage water resources.

SEE ALSO THE FOLLOWING ARTICLES

Electricity, Environmental Impacts of • *Electric Power Generation: Fossil Fuel* • *Electric Power: Transmission and Generation Reliability and Adequacy* • *Hydropower, Environmental Impact of* • *Hydropower, History and Technology of* • *Hydropower Resettlement Projects, Socioeconomic Impacts of* • *Hydropower Resources* • *Hydropower Technology*

Further Reading

Disheroon, F. R. (1993). Hydroelectric power. *In* "Environmental Law: From Resources to Recovery" (C. B. Campbell-Mohn and J. W. Futrell, Eds.), pp. 527–557. West Publishing Co., St. Paul, Minnesota.

Edwards, B. K. (2003). "The Economics of Hydroelectric Power." Edward Elgar, Cheltenham, UK.

Edwards, B. K., Flaim, S. J., and Ancrile, J. D. (1992). Maximizing the value of thermally integrated hydroelectric generating facilities. *In* "Modeling and Simulation, Proceedings of the Twenty-Third Annual Pittsburgh Conference on Modeling and Simulation (1992)" (W. G. Vogt and M. H. Mickle, Eds.), pp. 463–470.

Edwards, B. K., Howitt, R. E., and Flaim, S. J. (1996). Fuel, crop, and water substitution in irrigated agriculture. *Resource Energy Econ.* **18**(3), 311–331.

Edwards, B. K., Flaim, S. J., and Howitt, R. E. (1999). Optimal provision of hydroelectric power under environmental constraints. *Land Econ.* **75**(2), 267–283.

Ferguson, C. E. (1972). "Microeconomic Theory." Richard D. Irwin, Inc., Homewood, Illinois.

Francfort, J. (1997). "Hydropower's Contribution to Carbon Dioxide Emission Reduction," Idaho National Engineering and Environmental Laboratory (November). U.S. Department of Energy, Idaho Falls, Idaho.

International Commission on Large Dams (ICOLD). (1998). "World Register of Dams, 1998." ICOLD, Paris, France.

United States Department of the Interior, Bureau of Reclamation. (2001). "The History of Hydropower Development in the United States." Available on the Internet at http://www.usbr.gov.

Varian, H. R. (1992). "Microeconomic Analysis." W. W. Norton, New York.

World Commission on Dams. (2000). "Dams and Development: A New Framework for Decision-Making." Earthscan Publ. Ltd., London and Sterling, Virginia.

Hydropower, Environmental Impact of

GLENN CADA and MICHAEL SALE
Oak Ridge National Laboratory
Oak Ridge, Tennessee, United States
DENNIS DAUBLE
Pacific Northwest National Laboratory
Richland, Washington, United States

1. **Introduction**
2. **Water Quality Effects**
3. **Changes in Sediment and Nutrient Balance**
4. **Alteration of Terrestrial Habitats**
5. **Alteration of Aquatic Habitats**
6. **Interference with Movements of Aquatic Organisms**
7. **Greenhouse Gas Emissions**

Glossary

anadromous fish Fish that spend most of their lives in the sea and then migrate to fresh water to breed (e.g., salmon).
catadromous fish Fish that spend most of their lives in fresh water and then migrate to the sea to breed (e.g., eels).
entrainment The process by which an aquatic organism is drawn into the intake of a power plant along with the inflowing water.
epilimnion The upper, wind-mixed layer of a thermally stratified lake or reservoir; because of its exposure and turbulent mixing, dissolved gases are freely exchanged with the atmosphere.
eutrophication The process by which a lake or reservoir receives a large (often excessive) supply of nutrients, resulting in a high production of organic matter by plants and animals.
fishway A constructed device that allows fish to migrate upstream past dams, waterfalls, and rapids.
flushing rate The average length of time that water resides in a lake or reservoir, ranging from several days in small impoundments to many years in large lakes. The flushing rate is calculated by dividing the volume of water passing through the lake per unit time by the lake volume.
gas bubble trauma A condition in which fish tissues become supersaturated with dissolved gases that subsequently come out of solution and form bubbles in the gills, blood vessels, and other tissues.
hypolimnion The bottom-most layer of a thermally stratified lake or reservoir; it is isolated from wind mixing, often too deep for sunlight penetration and plant photosynthesis, and typically cooler and less oxygenated than the epilimnion.
metalimnion The middle or transitional layer of water in a reservoir, between the epilimnion and hypolimnion, in which the temperature exhibits the greatest difference in a vertical direction.
plankton Floating or weakly swimming organisms in the open water of a reservoir or river.
supersaturation of gases A state in which gases, such as nitrogen and oxygen, are held in solution in water in quantities exceeding the saturation level of the water (i.e., greater than 100% saturation) for a given temperature and pressure.
thermal stratification The temporary separation of a lake or reservoir into distinct vertical layers due to temperature differences.
trap efficiency The ability of a reservoir to trap and retain sediments and nutrients; it is usually expressed as a percentage of the inflowing sediments or nutrients that are retained in the reservoir.

Although free of many of the harmful emissions associated with other energy sources, hydroelectric power plants can affect the environment by impounding water, flooding terrestrial habitats, and creating a barrier to the movements of sediments, nutrients, and aquatic organisms. The ability of hydropower projects to alter the flow of water can affect both aquatic and terrestrial habitats downstream from dams.

 © 2004 Elsevier Inc. All rights reserved.

1. INTRODUCTION

Hydroelectric power production is largely free of several major classes of environmental impacts associated with nonrenewable energy sources (Table I). Unlike fossil-fueled power plants, hydroelectric generation does not emit toxic contaminants (e.g., mercury) or sulfur and nitrogen oxides that can cause acidic precipitation. Although construction of the hydropower project may result in temporary emissions from dust and equipment, these impacts are at least two orders of magnitude less than the continuing emissions from a coal-fired power plant. Inundation and decomposition of vegetation by the reservoir can cause the release of greenhouse gases (e.g., methane, carbon dioxide), but the significance of this source depends greatly on the age, size, and geographic location of the reservoir and on the amount of vegetation and soil carbon flooded. Hydroelectric power provides an opportunity to reduce the production of air pollutants and greenhouse gases when compared with power produced by the combustion of fossil fuels or biofuels.

Unlike most other forms of energy production, hydroelectric power plants generate few solid wastes. Land may be required for the disposal of material dredged from reservoirs or waterborne debris, but the amounts of land needed are very small compared with those needed for the continuing disposal of coal ash and slag. Similarly, hydropower production does not create hazardous or radioactive wastes that require safe long-term storage facilities. Many other environmental impacts associated with the overall fuel cycles of other energy sources are minor or nonexistent for hydroelectric power. These externalities include impacts associated with resource extraction (e.g., coal mining, oil drilling), fuel preparation (e.g., refining), and transportation (e.g., oil spills, other accidents).

Although hydroelectric power plants have many advantages over other energy sources, they also have potential environmental impacts. Most of the adverse impacts of dams are caused by habitat alterations. Reservoirs associated with large dams can inundate large areas of terrestrial and river habitats and displace people from their homes. Diverting water from the stream channel or curtailing reservoir releases to store water for future electrical generation can dry out streamside (riparian) vegetation. Insufficient water releases degrade habitat for fish and other aquatic organisms in the river below the dam. Water in a reservoir is stagnant compared with that in a free-flowing river. Consequently, waterborne sediments and nutrients can be trapped, resulting in the undesirable proliferation of algae and aquatic weeds (eutrophication). In some cases, water spilled from high dams may become supersaturated with nitrogen gas, resulting in gas bubble disease in aquatic organisms inhabiting the tailwaters.

Hydropower projects can also affect aquatic organisms directly. The dam can block upstream movements of fish, and this can have severe consequences for anadromous fish (e.g., salmon,

TABLE I

Potential Environmental Benefits and Adverse Impacts of Hydroelectric Power Production

Benefits	Adverse impacts
No emission of sulfur and nitrogen oxides	Inundation of wetlands and terrestrial vegetation
Few solid wastes	Emissions of greenhouse gases from flooded vegetation at some sites
Minimal impacts from resource extraction, preparation, and transportation	Conversion of a free-flowing river to a reservoir
Flood control	Replacement of riverine aquatic communities with reservoir communities
Water supply for drinking, irrigation, and industry	Displacement of people and terrestrial wildlife
Reservoir-based recreation	Alteration of river flow patterns below the dam
Reservoir-based fisheries	Loss of river-based recreation and fisheries
Enhanced tailwater fisheries	Desiccation of streamside vegetation below the dam
Improved navigation on inland waterways below the dam	Retention of sediments and nutrients in the reservoir
	Development of aquatic weeds and eutrophication
	Alteration of water quality and temperature
	Interference with upstream and downstream passage of aquatic organisms

steelhead, American shad), catadromous fish (e.g., American eels), and riverine fish that make seasonal migrations to spawn (e.g., sturgeon, paddlefish). Fish moving downstream may be entrained (i.e., drawn into the power plant intake flow). Entrained fish are exposed to physical stresses (e.g., pressure changes, shear, turbulence, strike) as they pass through the turbine, and these may cause disorientation, physiological stress, injury, and/or mortality.

This article describes the potential environmental impacts associated with hydropower production as well as measures that are employed to mitigate or eliminate these impacts.

2. WATER QUALITY EFFECTS

Construction of hydropower facilities (dam, powerhouse, roads, and transmission lines) will result in the disturbance and movement of soil. Water quality would be degraded if soil erosion increases the amounts of suspended material in the water (turbidity) or blankets the bottom of the river (sedimentation). Oils, greases, and chemical wastes may accidentally be spilled into the water during project construction. Well-known engineering practices, including soil stabilization techniques and storm-water retention dikes, can reduce impacts to water quality during construction. In most cases, the continuing effects of operation of the hydropower project are of greater concern. The remainder of this article considers the impacts of hydropower operation.

2.1 Temperature

The most common water quality issues associated with hydroelectric power production are changes in temperature and dissolved oxygen concentrations in the water released from the dam. Reservoirs in temperate regions often become thermally stratified, especially if they are deep or sheltered from the wind. When reservoir water begins warming during the spring, a warm surface layer floats over the colder bottom waters. In time, three distinct layers of water form: a warm upper layer (epilimnion), a narrow zone of rapidly decreasing water temperature (metalimnion), and a cool bottom layer (hypolimnion) (Fig. 1). This stratified condition continues throughout the summer until the surface waters cool during the autumn and top-to-bottom mixing occurs. Although thermal stratification is common in temperate regions, not all natural lakes and artificial reservoirs stratify. In general, reservoirs that are most likely to stratify are deep, have a low surface area/volume ratio and flushing rate, and are protected from mixing by their orientation to wind and the surrounding topography.

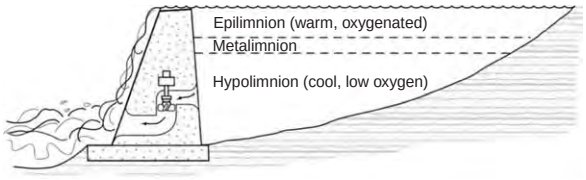


FIGURE 1 Thermal stratification of a hydropower reservoir.

Many hydroelectric turbine intakes withdraw water from the depths of the reservoir, and summer water temperatures in the discharged water are lower than would be the case in the free-flowing river. The opposite situation may occur during the winter. Because the large volume of water in the reservoir cools more slowly than does an unimpounded river, winter discharges from a hydropower plant may be warmer than natural. Altered water temperatures in the tailwaters have profound effects on aquatic organisms. Cold-water discharges during the summer can slow the growth rates and reduce productivity of fish and aquatic invertebrates, but they can also allow the establishment of cold-water fisheries in areas where the natural rivers are too warm. Conversely, warm-water discharges during the winter can speed the metabolic rate of aquatic insects and fish eggs, so that they develop and emerge before the appropriate season.

Mitigation of temperature effects is not difficult but can be expensive. Devices such as propellers have been used to break up stratification in small reservoirs. For large reservoirs, multiple-level intakes allow water to be withdrawn and mixed from different depths, so that water of the appropriate temperature can be discharged into the tailwaters.

2.2 Dissolved Oxygen

Thermal stratification can also decrease the concentrations of dissolved oxygen in the water discharged from the reservoir. The relatively stagnant hypolimnion is isolated from the processes of atmospheric diffusion and wind mixing that replenish dissolved gases in the epilimnion. Plant and animal respiration, bacterial decomposition of organic matter, and chemical oxidation all can act to progressively remove dissolved oxygen from bottom waters. This situation may be exacerbated by the input of high levels of oxygen-consuming organic materials that enter the reservoir from the watershed upstream. In

temperate regions, the decline in dissolved oxygen concentrations begins at the onset of stratification during the spring and continues until either all of the oxygen is used or reoxygenation occurs during the autumn turnover of the water body. Turbines that withdraw from the hypolimnion may discharge water that is low in dissolved oxygen, affecting the aesthetic qualities (taste, odor, and appearance), communities of aquatic organisms, and waste assimilation capacity of the tailwaters.

Low dissolved oxygen concentrations create chemically reducing conditions, which in turn may affect other water quality parameters. For example, decomposition of sulfur and nitrogen compounds in the absence of dissolved oxygen may result in the buildup of toxic hydrogen sulfide and ammonia in the hypolimnion. Chemically reducing conditions also increase the solubility of iron, manganese, and some heavy metals. These elements may be mobilized from the sediments and suspended particulate material and may be discharged into the tailwaters. Hypolimnetic discharges from hydropower facilities that have low dissolved oxygen concentrations and elevated levels of iron and other metals, manganese, ammonia, and sulfides may have adverse effects on downstream aquatic organisms and other water users. Adequate levels of dissolved oxygen (≥ 5 mg/L in most waters) are necessary not only for aesthetic qualities (taste and odor) but also to support a balanced community of aquatic organisms. Hydrogen sulfide creates odor problems and may cause downstream fish kills. High levels of iron and manganese increase the cost of domestic water treatment because they impart objectionable tastes to drinking water and can stain laundry and plumbing fixtures.

There are a number of structural measures that can be employed to increase dissolved oxygen concentrations in the discharge water (e.g., aerating turbines). Dissolved oxygen levels can also be increased through modifications in dam operations, including fluctuating the timing and duration of flow releases, spilling surface water from the top of the dam, and flow mixing with multiple-level intakes.

2.3 Gas Supersaturation

Dissolved gas concentrations may be raised as well as lowered by the presence of a hydroelectric dam. Spilling water over the crest of the dam (Fig. 2) often increases the concentration of total dissolved gases because entrained air bubbles dissolve under pressure when the water plunges into deep tailwater pools. As this water surfaces in the shallow river channel downstream, it becomes supersaturated ($>100\%$ saturation) with gases relative to the lower water pressures at the surface. If the subsequent rate of equilibration is slow, gas supersaturation may persist in flowing waters for long time periods and distances from the dam.

FIGURE 2 Spilling water from the Bonneville Dam on the Columbia River in the United States.

Dissolved gas saturations of 110% or greater are considered potentially damaging to aquatic organisms. Water that has become supersaturated with dissolved nitrogen can cause gas bubble trauma in fish, a potentially lethal condition similar to the "bends" experienced by deep divers who return to the surface too quickly. Gas supersaturation leading to gas bubble trauma has been most commonly reported at large dams in the Columbia River basin in the United States. Historically, gas supersaturations in the range of 115 to 143% have occurred in the Columbia and Snake rivers during periods of high spilling, but the problem has declined during recent years as a result of the implementation of mitigation measures.

The necessary conditions for nitrogen gas supersaturation at hydropower projects include a deep plunge pool (so that air bubbles will dissolve under elevated pressures) and a high dam (so that spilled water has sufficient velocity to reach the bottom of the pool). One successful method for preventing nitrogen gas supersaturation is to install flip lips, that is, structures at the base of the spillway that direct the spilled water horizontally downstream instead of deep into a plunge pool. By keeping spilled water (with entrained air bubbles) near the surface, there is less chance for excess nitrogen gas to become dissolved.

2.4 Eutrophication

When a free-flowing river enters a reservoir, the water slows down, spreads out, and warms up, often

creating excellent conditions for the growth of plants. Submerged rooted vegetation and emergent vegetation (e.g., cattails) proliferate in shallow water areas. The relatively low-flow conditions also favor the growth throughout the reservoir of microscopic planktonic plants (algae). In tropical areas, they also favor the growth of floating plants such as water hyacinth and water lettuce.

Depending on the circumstances, the proliferation of aquatic vegetation in the reservoir can be either beneficial or damaging. Rooted vegetation provides a valuable habitat for aquatic invertebrates and fish, and algae provide an essential food source for small planktonic animals and, in turn, plankton-eating fish. Often, the production of these organisms in a reservoir may far exceed that in a free-flowing river. On the other hand, the excessive undesirable growth of plants, known as eutrophication, can be a serious problem. Although many species of planktonic algae are food sources, some species that thrive in warm nutrient-rich waters release chemicals that are toxic to animals. Large plants can interfere with the operation of a hydropower project by clogging the intake structure, increasing the evapotranspiration loss of water from the reservoir, reducing the effective storage volume of the reservoir, and trapping more sediments and nutrients (and thereby decreasing the useful life of the reservoir). The decomposition of excessive growths of algae and large aquatic plants can degrade water quality in the reservoir and in the tailwaters downstream, impairing alternative water uses (e.g., fishing, swimming, drinking).

Surface water released from an eutrophic reservoir often has high concentrations of algae and low concentrations of nutrients compared with the inflowing river. This encourages the production of plankton-feeding organisms in the tailwaters. On the other hand, bottom water may have relatively high levels of nutrients due to the decomposition of organic material in the reservoir sediments. Discharges of nutrient-rich bottom water may stimulate excessive production of attached plants in the river channel below the dam, and this often has adverse impacts on habitats of fish and other aquatic organisms.

Mitigation of alterations in the nutrient balance of the river and reservoir is possible, but it is often costly and complicated. Excess growth of large aquatic plants can be controlled by mechanical harvesting or the introduction of herbivorous fish. Microscopic planktonic algae are more difficult to control. It is often easier to reduce the input of nutrients from the watershed or flush nutrients from the reservoir to limit algal production.

3. CHANGES IN SEDIMENT AND NUTRIENT BALANCE

Reservoirs may act as traps for sediments carried by the inflowing river. A stream entering a reservoir spreads out, reducing the water velocities and sediment-carrying capacity. A delta forms at the entrance to the reservoir, with the size and location of the delta depending primarily on the size of the sediment particles in the river and the velocity of flow through the reservoir. Coarse-grained heavy sediments carried by the stream are deposited immediately. Fine sediments remain in suspension and can be deposited farther into the reservoir or carried out of the reservoir in the outflow.

The ability of a reservoir to trap and retain sediments is known as the trap efficiency, expressed as the percentage of the total inflowing sediment that is deposited. Sediment trapping can be beneficial in cases where the reservoir retains soil that eroded from the watershed due to improper land use practices. However, reservoirs with high trap efficiencies will also tend to have the shortest useful life spans. The deposited sediment fills up the reservoir and, unless removed by dredging, reduces the storage capacity needed for hydropower production. Sediment trapping can also have biological effects. Compared with a turbid river, sedimentation will increase the clarity of water, benefiting aquatic plants and sight-feeding animals. Gravel deposits at the entrance to a reservoir will often provide an excellent fish-spawning habitat. On the other hand, fine sediments deposited elsewhere in the reservoir may blanket habitat and smother clams, fish eggs, and other nonmobile organisms.

Reservoirs also trap inflowing nutrients by two processes: (1) sedimentation of nutrient-bearing particulate matter and (2) transformation of dissolved nutrients into particulate material, which is then subject to sedimentation. Dissolved nutrients may be converted to particulate forms by physicochemical processes (adsorption and precipitation) or through uptake by plants and bacteria. The impacts of nutrient trapping may be beneficial or harmful, depending on the site. Nutrients trapped in the reservoir can support increased production of aquatic plants and fish. However, eutrophication caused by increased nutrient availability often diminishes water quality.

Removal of sediments and nutrients by the reservoir generally results in the tailwaters being "starved" of these materials, relative to the stream above the hydropower project. Clear water discharged from the reservoir has erosive power to pick up available sediment particles from the riverbed and stream banks below the dam. The river channel below the dam may be eroded, and scour may damage riverside structures (e.g., bridges, culverts, road embankments). The eroded sediments will be deposited farther downstream. Similarly, nutrients may be trapped in the reservoir sediments, leaving the discharged water nutrient poor compared with the water flowing into the reservoir. This can be beneficial in rivers where excessive nutrients are a form of water pollution. However, particularly in large rivers, reductions in nutrient flow may drastically lower the productivity of important fisheries.

The simplest way in which to mitigate for adverse sediment and nutrient trapping is to dredge the reservoir as needed. Numerous mechanical and hydraulic dredging techniques are available. Sediments in some reservoirs can be flushed out through pipes or notches in the dam. Large reservoirs impound enough water, so that sediments can be flushed out at any time, whereas in smaller reservoirs sediments might be able to be flushed only during floods and other high stream flow events.

4. ALTERATION OF TERRESTRIAL HABITATS

Creation of a reservoir behind a dam will flood terrestrial vegetation and displace associated wildlife and human populations. The significance of this impact depends on the size and location of the reservoir. A small reservoir will not cause major losses of terrestrial habitats and, in fact, may actually have benefits. For example, a hydropower reservoir typically creates a larger water surface than the original river channel that it flooded. Consequently, the reservoir can provide more habitat for waterfowl and, in arid regions, a more permanent source of drinking water for wildlife. Human populations may benefit from additional uses for hydropower reservoirs: flood control, recreation, fisheries, and a reliable source of water for drinking, industry, and agriculture. On the other hand, a large hydropower reservoir may flood villages and towns, displace large numbers of people, displace or drown terrestrial wildlife, and eliminate important floodplain vegetation. The decomposition of flooded vegetation may degrade water quality and increase the emissions of greenhouse gases.

Road construction and site clearance for the hydropower dam, powerhouse, and associated facilities may eliminate wildlife habitat, and construction noise and activities could interfere with the nesting of birds or migratory movements of terrestrial mammals.

Downstream from the reservoir, stream flow changes during construction and operation of the hydropower project can affect streamside (riparian) vegetation and associated animals. As with the reservoir, these effects may be positive or negative, depending on the site, design of the project, and mode of operation. For example, relatively constant releases from the reservoir associated with hydropower production could provide reliable flows in rivers that naturally experience wide variations in stream flow. This could encourage the establishment of riparian vegetation. On the other hand, some riparian communities require the seasonal flooding that is associated with natural unimpounded rivers, and they may decline under more uniform controlled flow releases. Some hydropower projects are operated in a "pulsing" mode, with water released from powerhouses only during portions of the day when electricity is most needed or valuable and with discharges from powerhouses being halted for the remainder of the day. By alternating flooding and drying within a single day, pulsing projects create an unnatural flow regime that can have serious impacts on both aquatic organisms and riparian communities.

5. ALTERATION OF AQUATIC HABITATS

Many, but not all, hydropower projects are associated with dams and reservoirs that are constructed either to store water or to increase the height, or hydraulic head, of water above the power plants. The water stored behind hydropower dams is used to shift river flows from times of surplus water to other times when water use is more valuable, for example, from the spring season when river flows are high to the summer season when they are low.

When dams and reservoirs are constructed on rivers that were previously free flowing, the aquatic habitats can be changed substantially, both in the reservoirs and downstream from the dams. The degree of habitat change depends on the reservoir operation (how and when the active storage is used to detain flows) and the physical dimensions of the reservoir. Some hydropower projects divert water out

of the natural river channel but do not have large impoundments. Such a diversion project usually has a bypassed reach of the river that has a substantially reduced water flow and modified aquatic habitat as well as a pipeline or canal that routes the diverted water to a powerhouse downstream.

5.1 Habitat Change in the Reservoir

In addition to inundation of terrestrial habitat, riverine habitats are altered by the creation of reservoirs. Natural riverine habitats are generally characterized by relatively shallow water with fast currents and a substrate of larger sand, gravel, boulders, and/or exposed bedrock. Reservoirs impounded behind dams are usually deeper and slower moving than free-flowing rivers. Because water velocities in these impoundments are greatly reduced, finer sediments that would have been transported downstream tend to accumulate. Silt covers the gravel and boulders that formed the bottom substrates in the former stream channel. Silt and sand habitats often support less varied invertebrate and fish communities. Conversely, the quiet waters and muddy bottoms of reservoirs may allow the development of beds of aquatic plants that are very productive habitats for some aquatic animals and waterfowl. Because of these habitat changes, animals and plants that are characteristic of river ecosystems are replaced by animals and plants that are adapted to lake-like conditions.

If the reservoir is deep and stagnant enough and the climatic conditions are conducive, the deep colder water may become isolated from the warmer surface water, a condition known as thermal stratification. The deep-water zone (hypolimnion) provides a very different habitat for aquatic organisms than does the unimpounded river. Depending on the transparency of the water and the depth of the reservoir, light might no longer reach the bottom, preventing aquatic plants from growing there. Storage and release of water for power production alternately floods and dries out the shoreline areas, further limiting the growth of aquatic plants along the margins of the reservoir.

5.2 Altered Flows below the Reservoir

A hydropower storage reservoir can retain enough water that the river below the dam dries up. Similarly, a diversion project withdraws water from the river, passes it through a canal or pipeline, and returns it to the river downstream from the powerhouse. In both types of hydroelectric projects, stream flows and aquatic habitats in a portion of the river are greatly or entirely reduced. This may eliminate fish and invertebrates that were residing in the affected reach and may constrain the movements of migratory animals.

Releasing a predetermined amount of water down the river channel is often required to sustain the other instream uses of water, including maintenance of fish and wildlife communities, streamside vegetation, recreation, aesthetics, water quality, and navigation. Providing flows downstream from a storage reservoir or hydroelectric diversion is simple; water can be spilled from the dam instead of diverted to a pipeline or stored in a reservoir. Because releasing water to support instream uses below the dam frequently makes that water unavailable for generation of electricity, hydropower operators are interested in providing sufficient, but not excessive, releases. Methods have been developed to ascertain the instream flow requirements of many of these uses, but the needs of biological resources are often difficult to assess. A variety of instream flow assessment methodologies are available to help determine how much water needs to be released to maintain aquatic and riparian communities.

6. INTERFERENCE WITH MOVEMENTS OF AQUATIC ORGANISMS

Hydropower projects can affect aquatic organisms directly by impeding their migration, blocking their movement upstream and/or downstream, and indirectly by altering the flow patterns and water quality conditions to which they respond. Fish are more likely to be affected than are other aquatic organisms because fish have more complex behaviors. Probably the most important change to physical habitats that influences fish movements is reduced water velocity upstream of dams. Flowing streams are changed to reservoirs, and the average depth and cross section are increased markedly. Higher water temperatures may also result from the increased surface area and reduced water velocities in reservoirs. The rate of downstream transport of aquatic invertebrates and other organisms can be slowed when water velocities upstream of hydroelectric dams are reduced substantially or when a large proportion of the stream is diverted for hydroelectric generation.

As discussed earlier, project operations such as spilling surface water can result in increased concentrations of dissolved gas, whereas storage practices

may cause lower concentrations of dissolved oxygen to be present than is the case under riverine conditions. Water temperature may also be altered due to storage practices. For example, in rivers with extensive hydroelectric development, there may be a phase shift where the timing of seasonal maximum or minimum water temperature is altered. Because aquatic organisms require a certain range of conditions to survive, grow, and reproduce, any major changes in water quality can affect their relative abundance and distribution. The following discussion focuses on the effects of hydropower projects on the upstream and downstream movements of migratory fish.

6.1 Upstream Passage

Most hydropower dams pose a physical barrier to upstream-migrating fish. Fish migrate upstream for a variety of reasons, including to return to natal areas to spawn (e.g., adult salmon, steelhead, American shad), to complete their freshwater rearing periods (e.g., juvenile American eel), and to feed (e.g., sturgeon, sucker). Many freshwater fish move seasonally between large rivers and their tributary streams. These migration routes may be blocked by dams, eliminating large areas from production of fish.

Many hydroelectric projects have ways in which to assist the upstream movement of fish, including fish ladders, trap-and-haul operations, and fish elevators. In all cases, fish are slowed while they search for and decide whether to use the upstream passage facilities. Delay of adult salmon and steelhead at Columbia River dams is one factor contributing to their mortality during migration from the Pacific Ocean to spawning areas. However, adult salmon generally migrate faster through reservoirs than they do through nonimpounded or natural sections of rivers that have higher water velocities against which the fish must swim.

Critical features of fishway design that affect passage success of fish include water velocity, depth, slope, and availability of resting areas. The entrance design is also an important consideration. Some fish species have a difficult time in locating passageway openings or auxiliary flows provided to attract those seeking upstream passage.

The majority of fishways in North America have been designed for efficient upstream passage of salmonids (Fig. 3). However, other species of fish might not respond similarly to these fishways due to their unique behaviors. For example, adult white sturgeons are often too large to navigate certain designs. The upstream passage of adult lampreys is also poor relative to the number of fish that approach hydropower projects.

6.2 Downstream Passage

Fish that migrate downstream past a hydropower project have three primary routes of passage. They may (1) be drawn into the power plant intake flow (entrainment) and pass through the turbine, (2) be diverted via bypass screens into a gatewell and to a collection facility or the tailrace, or (3) pass over the

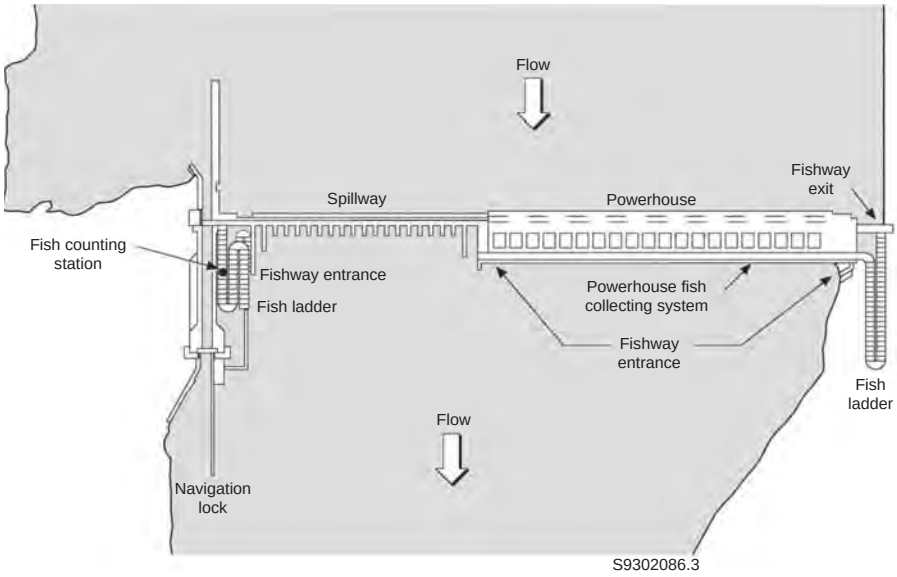


FIGURE 3 Generic design of a hydroelectric facility in the Columbia River system showing the location of fishways.

dam in spilled water. Recent modifications made to dams to decrease numbers of turbine-passed fish include guiding migrating fish toward spillbays and using surface bypass systems (Fig. 4) and behavioral guidance walls. Ice and trash sluiceways have also been modified to provide a surface passage route for migrating fish.

Entrained fish are exposed to physical stresses (e.g., rapid change in pressure, shear, turbulence, blade strike) that may be injurious. In the best existing turbines, up to 5% of turbine-passed fish may be injured or killed, and mortalities in some turbines may be 30% or more. Several new design concepts are under development with a goal of reducing mortality of turbine-passed fish to 2% or less.

Nonturbine passage routes pose some risk to fish as well. Weak swimmers, such as juvenile lampreys and small resident fish (e.g., minnows, sunfish), may be impinged or injured on contact with bypass screens. Design features that influence the ability of a diversion screen to guide fish safely include screen type (e.g., rotating, fixed), screen material, size of openings, length relative to intake opening size, angle, and approach velocity. The design and location of outfalls from fish bypasses are also critical to minimizing exposures of bypassed fish to predatory fish in the dam tailrace. Fish that pass via the spill of a high-head dam are subjected to extremely high and variable water velocities, may be abraded by contact with the dam face, and may collide with submersed structures below the dam, including those designed to dissipate the high energy of spilled water. However, project spill may reduce the residence time of migrating fish immediately above the dam during which they are vulnerable to predators in the reservoir. Thus, design of the entire project, from forebay entrance to tailrace exit, affects the safe and efficient passage of fish.

FIGURE 4 Surface view (looking downstream toward the project) of the prototype surface collector at the Bonneville I powerhouse on the Columbia River. This structure was designed to provide a surface passage route for juvenile salmon and steelhead that would otherwise pass through turbines.

6.3 Effects of Water Quality Alterations on Movements

Adverse water quality conditions resulting from operation of hydroelectric projects may influence the movement of aquatic organisms. For instance, dissolved gases in excess of 100% saturation can occur under conditions of high project spill, which in turn can lead to gas bubble trauma. Some storage reservoirs may experience low dissolved oxygen concentrations and/or thermal stratification. Fish could avoid or be attracted to these conditions, depending on their life history requirements, relative sensitivity of their sensory systems, and their physiological sensitivity. The movement and distribution of aquatic organisms may be influenced by water temperatures and dissolved gas changes caused by hydropower development. Hydropower operators can modify project structures or operations to mitigate for high total dissolved gas levels and improve passage conditions for fish.

7. GREENHOUSE GAS EMISSIONS

Concern about the climatic effects of increased emissions of greenhouse gases, primarily carbon dioxide and methane, is a relatively recent issue for electric power production, and the contribution of hydroelectric plants is still being clarified. It had been assumed that hydropower projects emit few, if any, greenhouse gases compared with those emitted by fossil-fueled power plants. However, decomposition of inundated vegetation and other organic matter within the reservoir can result in greenhouse gas emissions that may continue for decades after the initial flooding. The amounts of greenhouse gases released from a hydropower reservoir vary widely, depending on geography, altitude, latitude, water temperature, reservoir size and depth, depth of turbine intakes, hydropower operation, carbon input from the river basin, and reservoir construction (e.g., whether vegetation was cleared from the reservoir before the dam was closed). Recent studies suggest that greenhouse gas emissions from reservoirs in higher latitudes (e.g., Canada, Scandinavia) are low but that they may be a greater concern in tropical areas. For example, studies of 30 reservoirs in Brazil

found a 500-fold difference in greenhouse gas emissions among dams. The lowest emissions from Brazilian reservoirs were similar to those from Canadian lakes and reservoirs, whereas the highest emissions were within the range of gross emissions from fossil-fueled power plants. Thus, hydropower plants cannot be automatically assumed to release less greenhouse gases than do fossil-fueled sources of electricity. Emissions must be determined on a case-by-case basis.

SEE ALSO THE FOLLOWING ARTICLES

Electricity, Environmental Impacts of • *Entrainment and Impingement of Organisms in Power Plant Cooling* • *Hydropower Economics* • *Hydropower, History and Technology of* • *Hydropower Resettlement Projects, Socioeconomic Impacts of* • *Hydropower Resources* • *Hydropower Technology* • *Wind Energy Technology, Environmental Impacts of*

Further Reading

Baxter, R. M. (1977). Environmental effects of dams and impoundments. *Annu. Rev. Ecol. Systematics* **8,** 255–283.

Mattice, J. S. (1991). Ecological effects of hydropower facilities. *In* "Hydropower Engineering Handbook" (J. S. Gulliver and R. E. A. Arndt, Eds.), pp. 8.1–8.57. McGraw–Hill, New York.

Petts, G. E. (1984). "Impounded Rivers: Perspectives for Ecological Management." John Wiley, New York.

World Commission on Dams. (2000). "Dams and Development: A New Framework for Decision-Making," report of the World Commission on Dams. Earthscan Publications, London. www.dams.org.

Hydropower, History and Technology of

JOHN S. GULLIVER and ROGER E. A. ARNDT
University of Minnesota
Minneapolis, Minnesota, United States

1. **Introduction**
2. **Hydropower Potential**
3. **What Nature Gives Us**
4. **How Hydropower is Captured**
5. **Hydroturbines**
6. **Ecological Considerations**
7. **Advantages and Disadvantages of Hydropower Development**

Glossary

axial-flow turbine A collective term for turbines with axial flow through the runner blades axially to the turbine shaft; both propeller turbines and Kaplan turbines are axial-flow turbines.

base load Typically, the minimum load over a given period of time.

capacity The greatest load that a piece of equipment can safely serve.

dam A massive wall or structure built across a valley or river for storing water.

design head The head at which the turbine is designed to operate at maximum efficiency.

draft tube The diffuser that regains the residual velocity energy of the water leaving the turbine runner.

energy The power of doing work for a given period, usually measured in kilowatt-hours.

forebay The upstream part of the bay-like extension of the river for the location of a powerhouse.

Francis turbine A radial-inflow reaction turbine where the flow through the runner is radial to the shaft.

generator A machine powered by a turbine that produces electric current.

gross head The difference between headwater level and tailwater level at the powerhouse.

headrace The portion of the power canal that extends from the intake works to the powerhouse.

headwater The water upstream from the powerhouse, or generally the water upstream from any hydraulic structure creating a head.

headwater elevation (or headwater level) The height of the headwater in the reservoir.

hydraulic efficiency An efficiency component of the turbine, expressing exclusively the power decrement due to hydraulic losses (e.g., friction, separation, impact), including the losses in the scroll case and the draft tube.

hydroelectric power The electric current produced from water power.

hydroelectric power plant A building in which turbines are operated, to drive generators, by the energy of natural or artificial waterfalls.

hydropower plant (or hydropower development) The comprehensive term for all structures (one powerhouse and pertaining installations) necessary for using a selected power site.

hydropower station A term equivalent to the powerhouse and sometimes including the structures situated nearby.

hydropower system Two or more power plants (and therefore two or more powerhouses) that are cooperating electrically through a common network.

intake (or intake works or headworks) A hydraulic structure built at the upstream end of the diversion canal (or tunnel) for controlling the discharge and preventing silt, debris, and ice from entering the diversion.

Kaplan turbine An axial-flow reaction turbine with adjustable runner blades and adjustable guide vanes.

kinetic energy Energy that a moving body has because of its motion, dependent on its mass and the rate at which it is moving.

load The amount of electric energy delivered at a given point.

load demand A sudden electrical load on the generating units, inducing the rapid opening of the turbines.

load factor The ratio of the annually produced kilowatt-hours and of the energy theoretically producible at installed capacity during the whole year.

load rejection A sudden cessation of electrical load on the generating units, inducing the rapid closure of the turbines.

main shaft The rotating element that transmits torque developed by the turbine runner to the generator rotor

 © 2004 Elsevier Inc. All rights reserved.

or that transmits torque developed by the motor to the pump impeller.

mechanical efficiency An efficiency component of the turbine, expressing exclusively the power losses of the revolving parts due to mechanical friction.

needle valve A streamlined regulating body moving like a piston in the enlarged housing of the valve.

net head The part of the gross head that is directly available for the turbines.

nozzle (or jet nozzle) A curved steel pipe supplied with a discharge-regulating device to direct the jet onto the buckets in impulse runners.

peak load The greatest amount of power given out or taken in by a machine or power distribution system during a given time period.

Pelton turbine The main type of turbine used under high heads.

penstock (or pressure pipe) A pressurized pipeline conveying the water in high-head developments from the headpond or the surge tank to the powerhouse.

plant discharge (or plant discharge capacity) The maximum discharge that can be used by all turbines of the power plant with full gateage (i.e., the entire discharging capacity of the turbines).

pondage That rate of storage in run-of-river developments that can cover daily peaks only.

power The rate at which work is done by an electric current or mechanical force, generally measured in watts or horsepower.

powerhouse The main structure of a water power plant, housing the generating units and the pertaining installations.

propeller-type turbine The collective term for axial-flow reaction turbines; in this terminology, it denotes two types: fixed-blade propeller turbines and adjustable-blade propeller turbines (i.e., Kaplan turbines).

pumped-storage development A combined pumping and generating plant; it is not a primary producer of electrical power, but by means of a dual conversion, it stores the superfluous power of the network and returns it during peak load periods, as would a battery.

reaction turbine A collective term for turbines in which the water jet enters the runner under a pressure exceeding the atmospheric value. The water flowing to the runner still has potential energy, in the form of pressure, that is converted into mechanical power along the runner blades.

reservoir An artificial lake into which water flows and is stored for future use.

run-of-river plant A development with little or no pondage regulation such that the power output varies with the fluctuations in the stream flow.

runner The rotating element of the turbine that converts hydraulic energy into mechanical energy; for reversible pump-turbines, the element is called an impeller and converts mechanical energy into hydraulic energy for the pump mode.

scroll case (or spiral case) A spiral-shaped steel intake guiding the flow into the wicket gates of the reaction turbine.

semi-scroll case (or spiral case) A concrete intake directing flow to the upstream portion of the turbine with a spiral case surrounding the downstream portion of the turbine to provide uniform water distribution.

setting The vertical distance between the tailwater level and the center of a turbine runner.

specific speed A universal number that indicates the machine design: impulse, Francis, or axial.

tailrace The portion of the power canal that extends from the powerhouse to the recipient watercourse.

tailwater The water downstream from the powerhouse, generally the water downstream from any hydraulic structure creating a head.

tidal power plant (or tidal power station) A power station that uses the potential hydraulic power originating from the tidal cycles of the sea.

Turbine A device that produces power by diverting water through blades of a rotating wheel that turns a shaft to drive generators.

turbine discharge capacity The maximum flow that can be discharged by a single turbine at full gateage.

turbine efficiency The entire efficiency of the turbine (i.e., the product of hydraulic mechanical and volumetric efficiencies).

water power A general term used for characterizing both power (kilowatts) and energy (kilowatt-hours) of watercourses, lakes, reservoirs, and seas.

Hydropower has a long and rich history, being important to the development of modern industrial society and essential to the electrification of the world. This article reviews some of this history and the hydropower potential that remains in the world today. Next, the water cycle and how hydropower is extracted from this cycle will be discussed. Finally, some of the ecological considerations and the advantages and disadvantages of hydropower development will be addressed.

1. INTRODUCTION

Falling or flowing water has been used to perform work for thousands of years, with the particular uses varying with the social and political conditions of the times. Although the Greeks and Romans knew of waterwheels since the 3rd century BC, these labor-saving devices were not used extensively until the 14th century. Early tasks included grinding grain, sawing wood, powering textile mills, and (later) operating manufacturing plants. Mills or factories

were located at the hydropower sites to directly use the available energy. By the end of the 18th century, there were approximately 10,000 waterwheels in New England alone. The power output of these early plants, usually limited to about 100 kW, is compared with other power sources in Fig. 1.

During the 19th century, hydropower became a source of electrical energy, although some form of hydroelectric turbine development can be traced back as far as 1750. Benoit Fourneyron is credited with developing the first modern hydroelectric turbine in 1833. The first hydroelectric plant is documented as coming online September 30, 1882, in Appleton, Wisconsin, and is still functioning. However, there is some dispute over this; Merritt cited the Minneapolis Brush Electric Company as beginning operation of a hydroelectric plant some 25 days earlier. The generation of electricity from falling water expanded the need for larger hydroelectric plants because the energy did not need to be used on-site. The transmission of power over long distances became economical in 1901 when George Westinghouse installed alternating current equipment at Niagara Falls in New York, further expanding the potential uses of hydropower.

As Fig. 1 indicates, the power capabilities of water turbines became larger as the need grew. During the 1930s, large dams and ever-increasing turbine capacities became the norm. The power capacity of steam turbines was also increasing rapidly, and the relative cost of electricity continued to fall. Finally, during the period between 1940 and 1970, the cost of operating and maintaining older, smaller hydroelectric plants became greater than the income they could produce, and many were retired. This is seen in Fig. 2, where small hydropower capacity decreased as overall hydropower capacity climbed rapidly in the United States. A similar trend occurred in European countries. Hydropower development in other parts of the world was insignificant before 1930, as indicated by world hydropower production in Fig. 3.

The largest hydropower facility in the world is currently the Itaipu Dam on the Parana River, located between Brazil and Paraguay. This hydroplant's 12,600-MW capacity is the equivalent of 12 large thermal powerplants. It is upstream from the incomparable Iguasu Falls. China is building an even larger project at Three Gorges. The world's largest capacity hydroelectric plants in 2000 are listed in Table I.

There are two other basic sources of hydropower besides that which is extracted from the world's rivers: tidal power and wave power. Small tidal mills to provide mechanical power existed hundreds of years ago. However, the first hydroelectric tidal plant

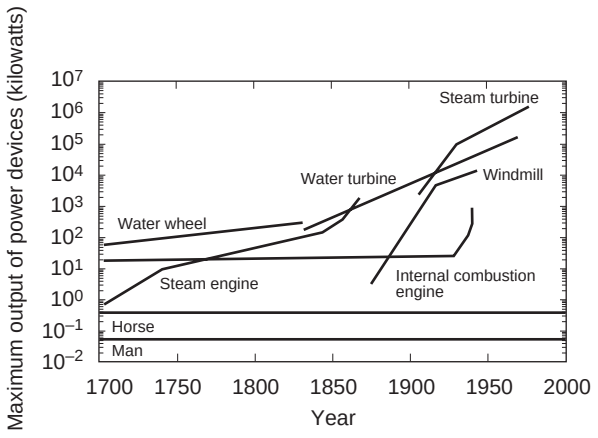


FIGURE 1 The maximum power output of selected power devices over the period from 1700 to 1970. From Chappel, 1984.

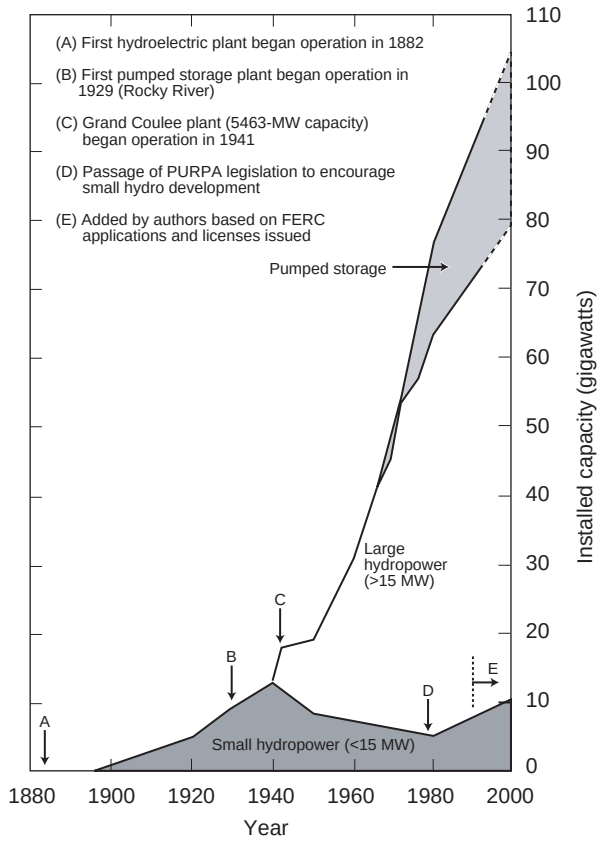


FIGURE 2 Installed hydroelectric capacity in the United States for the period from 1882 to 2000. FERC, Federal Energy Regulatory Commission. Assembled from Federal Energy Regulatory Commission, 1992, and International Waterpower and Dam Construction annual yearbooks.

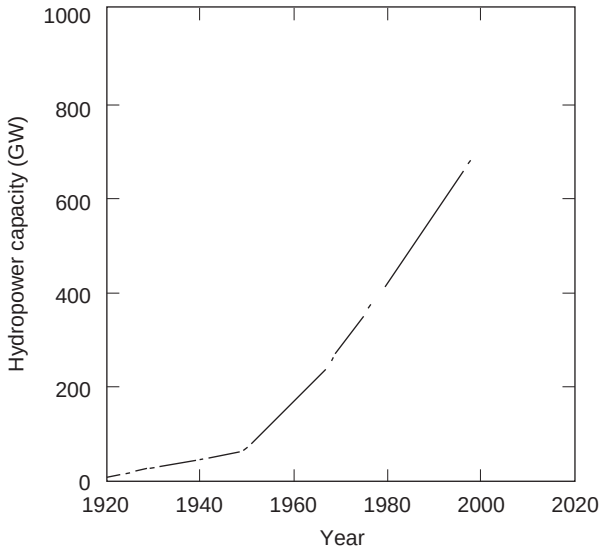


FIGURE 3 World hydropower production. Assembled from International Water Power and Dam Construction annual yearbooks.

was developed during the 1960s on the La Rance estuary in northern France. Producing up to 240 MW of power, this tidal plant uses a dam across a cove mouth to form a pond. Sluice gates open to let water flow in during the rising tide and then close with the returning tide as water is directed through a standard hydroturbine.

The feasibility of tidal power depends on the range of tide experience and on finding a location where an inordinately long dam does not need to be built. Thus, only a handful of tidal plants have been developed since the La Rance plant: a 10-MW plant in 1986 and a number of smaller plants in China; an 18-MW plant in 1984 at an existing flood structure at Annapolis Royal in Nova Scotia, Canada; and a 400-kW plant in the Soviet Union. The turbines at Annapolis Royal are of interest because they can turn in both directions, capturing energy from both the incoming and outgoing tides.

The United Kingdom has studied the feasibility of building a dam across the Severn estuary, generating up to 7000 MW of power with 192 hydroturbines. Canada has conducted a similar study on the Bay of Fundy.

All tidal power projects are low head, up to 36 feet (11 m). Gulliver and Dotan found that the cost per kilowatt produced is approximately proportional to $H^{-0.53}$, where H is the hydraulic head across the dam. Obviously, any low-head hydropower project must be planned carefully.

2. HYDROPOWER POTENTIAL

On a worldwide basis, hydropower represents roughly one-quarter of the total electrical energy generated. There is approximately 3800 GW of technically feasible hydropower potential existing in the world that could produce 14,370,000 GWh per year. Of this amount, 2100 GW (8,082,000 GWh/year) is feasible under current economic conditions and 700 GW (2,645,000 GWh/year) was installed in 2001. The technically feasible potential

TABLE I

World's Largest Capacity Hydroelectric Plants in 2000

Name of dam	Location	Rated capacity (MW)		Year of initial operation
		Present	Ultimate	
Itaipu	Brazil/Paraguay	12,600	14,000	1983
Guri	Venezuela	10,000	10,000	1986
Grand Coulee	Washington State (United States)	6494	6494	1942
Sayano-Shushensk	Russia	6400	6400	1989
Krasnoyarsk	Russia	6000	6000	1968
Churchill Falls	Canada	5428	6528	1971
La Grande 2	Canada	5328	5328	1979
Bratsk	Russia	4500	4500	1961
Moxoto	Brazil	4328	4328	n.a.
Ust-Ilim	Russia	4320	4320	1977
Tucurui	Brazil	4245	8370	1984

Source: Mermel, T. W. (2000). The world's major dams and hydro plants. "International Water Power and Dam Construction," 2000 Handbook. Wilmington, Kent, UK.

is that considered developable based on physical considerations without considering current economics or other environmental issues. It represents 39% of the total energy in the world's rivers. The available and developed power in 2001 is given by continent in Fig. 4. The tremendous potential in Asia, Africa, and South America is apparent, amounting to more than half of the world's total hydropower potential. On the other hand, hydropower developed or under construction in Europe and North American is at 73 and 70% of the economically feasible potential, respectively. The future hydropower development in these two continents is likely to be in the form of upgrades in power and pumped storage.

3. WHAT NATURE GIVES US

Hydroelectric power, similar to many of humanity's activities, is produced by insertion into the hydrologic cycle, where water evaporates from the oceans, seas, lakes, and the like into the atmosphere, forms clouds of condensed water drops, falls as precipitation (e.g., rain, sleet, snow), and flows into rivers and eventually into the ocean. This cycle has significant variability given that floods and droughts are common. A dam placed in a river forms a reservoir with a given hydraulic head or a difference in elevation between the water behind the reservoir and the water below it. This hydraulic head is used to create power according to the relation

$$P = \eta\gamma QH, \tag{1}$$

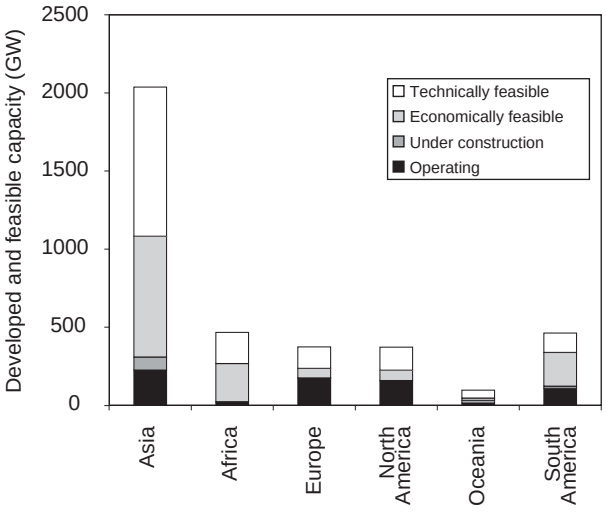


FIGURE 4 World hydropower resources by continent. Assembled from "International Hydropower and Dams," 2001 Handbook.

where P is the power produced (in watts), H is in meters, Q is the discharge (flow) routed through the hydroelectric turbine (in cubic meters/second), γ is the specific weight of water (in Newtons/cubic meter), and η is the overall efficiency of the hydroelectric facility (fraction between 0 and 1).

The parameter that is highly variable in Eq. (1) is the discharge, Q. At a typical dam and reservoir on a river, the discharge over the dam during floods can be 1000 times greater than that during droughts. Predicting this discharge for a typical year represents much of the risk in developing a hydroelectric facility because it typically requires 30 years of records to reduce the uncertainty in mean discharge to $\pm 15\%$.

A watershed or drainage basin is the region contributing flow to a given location such as a prospective hydropower site on a stream. Any watershed is composed of a wide variety of foliage, soils, geological formations, streams, and the like and, therefore, is unique. Because no two watersheds are the same, comparisons of runoff from two separate watersheds and generalized relationships for watershed runoff are not accurate. The data obtained from stream gauges in the watersheds of interest should be used whenever possible.

The ratio of stream runoff to precipitation within a given watershed depends primarily on seven factors that are illustrated in Fig. 5 and described as follows:

• *Interception.* Interception is the precipitation stored on vegetative cover and then evaporated. Interception accounts for between 10 and 20% of annual precipitation in a well-developed forest. The interception of crops varies greatly. However, some approximate values for a 25-mm storm are as follows: cotton, 33%; small grains, 16%; tobacco, 7%; corn, 3%; alfalfa, 3%; meadow grass, 3%.

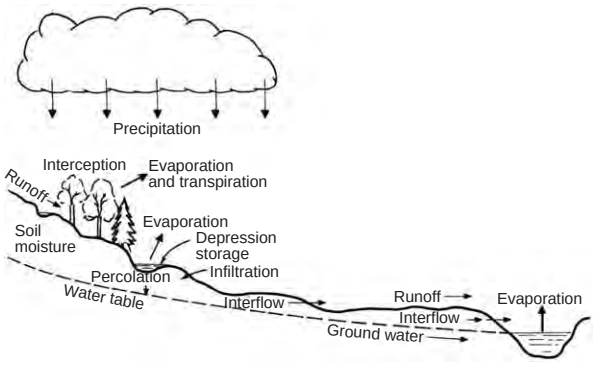


FIGURE 5 Processes affecting the relationship between precipitation and runoff.

• *Depression storage.* Depression storage is the rainwater retained in puddles, ditches, and other depressions in the surface. It occurs when the rainfall intensity exceeds infiltration capacity. At the end of the storm, water held in depression storage is either evaporated or infiltrated into the soil. The depression storage capacity of most drainage basins is between 13 and 50 mm of precipitation. A specific type of depression storage that is handled separately is a blind or self-enclosed drainage basin, that is, a portion of the drainage basin that does not drain into the stream network but rather is self-enclosed, usually with a lake, marsh, or bog at the center. Blind drainage basins are normally excluded from the hydrologic analysis of a hydropower site.

• *Surface runoff.* Surface runoff is the precipitation that moves downslope along the soil surface until it reaches a stream or lake. It is primarily associated with flood events, although in larger watershed basins the effects of surface runoff can be felt for up to a month. A small portion of any precipitation event will fall directly into the channel or stream. This immediately becomes runoff. It is grouped together with runoff from impervious areas, such as parking lots, streets, and buildings, because this runoff occurs rather quickly after precipitation.

• *Infiltration.* Infiltration is the passage of water through the soil surface. It usually implies percolation, which is the movement of water through unsaturated soil. Infiltration capacity depends on soil porosity, moisture content, and vegetative cover. Sandy or highly organic soils will have a greater infiltration due largely to increased porosity. A wet soil will have a lower infiltration capacity. Vegetation cover increases infiltration by retarding surface flow, increasing soil porosity with the root system, and reducing rain packing of the soil surface.

• *Soil moisture.* The precipitation that infiltrates into the soil will first be used to replenish soil moisture. Over time, the soil moisture is taken by plant root systems and eventually is transpired from plant foliage as part of the photosynthetic process. Water used to replenish soil moisture will not appear as stream flow. Soils with a high percentage of decayed plant material have a large capacity to retain moisture.

• *Interflow.* Interflow is water that infiltrates through the soil and moves laterally in the upper soil layers until it reemerges as surface runoff. A thin soil surface covering rock, hardpan, or plow bed will usually have large quantities of interflow. Interflow often emerges as a spring in riverbanks. It will not usually affect flood peaks but will increase stream flow at a steady rate for some time after the peak.

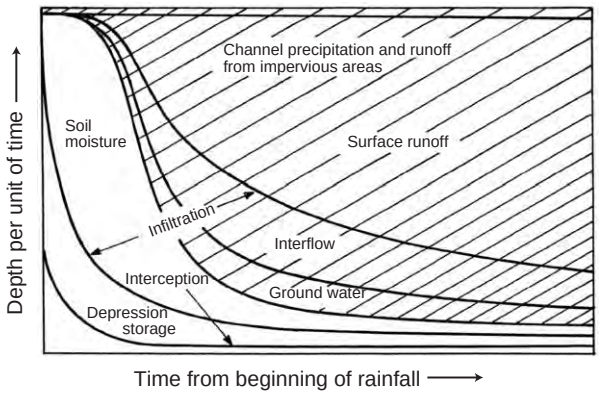


FIGURE 6 Schematic diagram of the segmentation of rainfall for an extensive storm in a relatively dry basin. The shaded area indicates precipitation that eventually results in a runoff.

• *Groundwater flow.* If the infiltrated water percolates downward until it reaches the water table (i.e., zone saturated with water), it will eventually reach stream as groundwater flow, which is the primary source of base flow for streams. Groundwater flow influences stream flow on a seasonal, rather than a weekly, time scale.

A schematic diagram of the segmentation of rainfall for an extensive storm in a relatively dry basin is given in Fig. 6. The shaded area indicates the quantity of rainfall that will eventually become stream flow. The general order in which the various types of flow reach the stream is as follows: channel precipitation, surface runoff, interflow, ground water.

The seven parameters that influence the runoff process, which depends on foliage, soil type, geological formation, and watershed geomorphology, can vary greatly within a given watershed and will certainly vary between two distinct watersheds. Therefore, it should be no surprise to discover that two adjacent watersheds of similar drainage areas have entirely different runoff characteristics.

4. HOW HYDROPOWER IS CAPTURED

The broad variety of natural conditions at hydropower sites has resulted in development of many different types of hydropower schemes. This is a disadvantage so far as engineering costs are concerned, but it provides a great deal of stimulus to those

involved in hydropower development. A longitudinal section of a typical scheme is provided in Fig. 7.

In the run-of-river development, a short penstock or dam directs the water through the turbines. The powerhouse is often an integral part of the dam, as shown in Fig. 8. The natural flow of the river remains relatively unaltered. A more complex development occurs at diversion or canal projects, where the water is diverted from the natural channel into a canal or long penstock, as shown in Fig. 9. This results in a significant change in the flow of water in a given reach of the river, sometimes for a considerable distance.

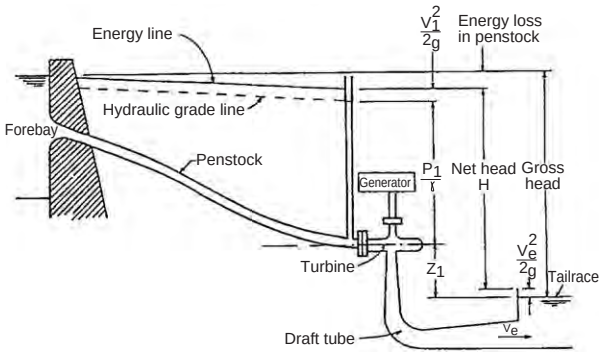


FIGURE 7 Schematic of a hydropower installation. $V_1^2/2g$ is head lost before the turbine, P_1/γ is the positive head extracted by the turbine, Z_1 is the suction head on the turbine, and $V_e^2/2g$ is the velocity head remaining at the draft tube outlet.

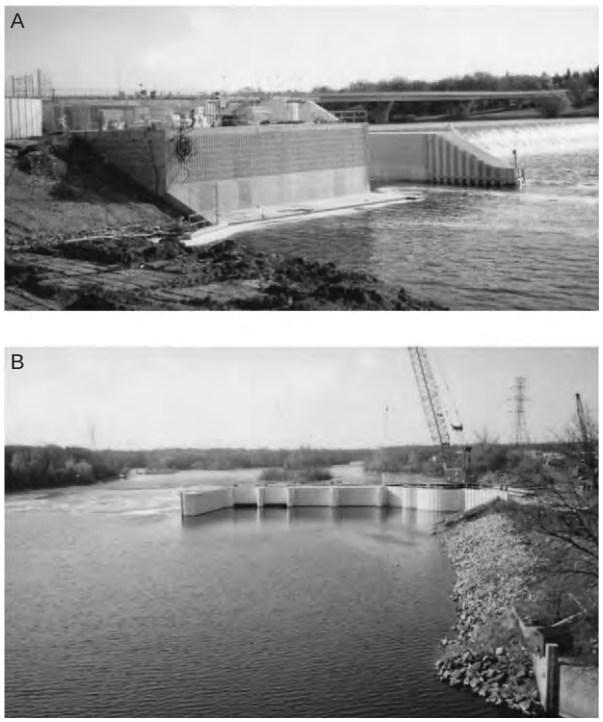


FIGURE 8 The 8.4-MW St. Cloud hydropower facility is a run-of-river development with the powerhouse composing a portion of the dam. Net head is 5.2 m. (A) Upstream view of the spillway and intakes into gates and powerhouse. (B) Downstream view of the powerhouse and spillway. Courtesy of M. A. Mortenson Company.

Storage regulation developments are defined as those in which an extensive impoundment at the power plant, or at the reservoir upstream of the power plant, allows for regulation of the flow downstream through storage. Water is stored during high-flow periods and is used to augment the flow during low-flow periods. This allows for a relatively constant supply of energy over the course of the year.

Pump storage facilities are normally large developments in which water is pumped from a lower reservoir during off-peak hours when the cost of energy is low. The pumps are run in reverse as turbines during the hours of peak demand to augment the power supplied from other sources. Thus, the pumped storage facility serves the same function as a large battery. It is used to allow large base load facilities to operate at continuous power output in cases where it is uneconomical to allow the power output of a large plant to fluctuate.

Tidal power plants have also been developed or considered. These are located in areas where there

FIGURE 9 The Mayfield project on the Cowlitz River in Washington. It is composed of an arch dam and spillway (background), a power tunnel leading to surge tanks (left), and penstocks leading to a 122-MW powerhouse (center front). Notice how the layout maximizes efficient use of the terrain. Courtesy of Harza Engineering.

are large tidal fluctuations. Low-head turbines are used to harness the energy in the tidal cycle, and an entire bay or estuary is enclosed by a low dam.

Hydropower schemes can either be single purpose, having as their only purpose the production of electricity, or multipurpose, having hydropower production as just one aspect of the total use of the facility. Multipurpose facilities include those in which hydropower is developed in conjunction with irrigation, flood control, navigation, and water supply. Hydropower plants are also categorized by the type of use. For example, a base load plant is one in which the power is used to meet all or part of a sustained and constant portion of the electrical load. Energy from these plants that is available at all times is referred to as "firm power." The need for power varies during the course of the day, and power requirements over and above the base load requirement are met by peak load facilities. These are plants in which the electrical production capacity is relatively high and the volume of water discharged through the units can be changed readily to meet peak demands. Storage or pondage of the water supply is necessary for these load demands because such plants can be started and stopped more rapidly and economically than can fossil fuel and nuclear power plants.

A typical hydropower facility consists of the following components:

- The dam or diversion structure with the associated reservoir
- The powerhouse structure and its foundation
- Hydraulic conveyance facilities, including the head race, headworks, penstock, gates and valves, and tailrace
- The turbine-generator unit, including the guide vanes or wicket gates, turbine, draft tube, speed increaser, generator, and speed-regulating governor
- Station electrical equipment, including the transformer, switch gear, automatic controls, conduit, and grounding and lightning systems
- Ventilation, fire protection, communication, and bearing cooling water equipment
- Transmission lines.

The selection of the proper turbine for a given site depends on the head and flow available as well as the use that is to be made of the facility. Because there is such a broad variation in available conditions in the field, many different turbine configurations have been developed and are available. It is essential that a hydropower engineer understand in detail the rationale for selection of hydroturbines. The other items listed, although not unique to hydropower production, are also crucial to a properly operating facility.

5. HYDROTURBINES

The water turbine has a rich and varied history. It has been developed as a result of a natural evolutionary process from the waterwheel. Originally used for direct drive of machinery, the use of the water turbine for the generation of electricity is a comparatively recent activity. Much of its development occurred in France, which did not have the cheap and plentiful sources of coal that England had and that sparked the industrial revolution during the 18th century. During the 19th century, France found itself with its most abundant energy resource being water. To this day, *Houille Blanche* (literally "white coal") is the French term for water power.

In 1926, the Société d'Encouragement pour l'Industrie Nationale offered a prize of 6000 francs to anyone who would "succeed in applying at large scale, in a satisfactory manner, in mills and factories, the hydraulic turbines or wheels with curved blades of Bélidor." Bélidor was an 18th-century hydraulic and military engineer who, between 1737 and 1753, authored a monumental four-volume work, *Architecture Hydraulique,* a descriptive compilation of hydraulic engineering information of every sort. The waterwheels described by Bélidor departed from convention by having a vertical axis of rotation and being enclosed in a long cylindrical chamber approximately 1 m in diameter. Large quantities of water were supplied from a tapered sluice at a tangent to the chamber. The water entered with considerable rotational velocity. This preswirl, combined with the weight of water above the wheel, was the driving force. The original tub wheel had an efficiency of only 15 to 20%.

Water turbine development proceeded on several fronts from 1750 to 1850. The classical horizontal axis waterwheel was improved by engineers such as John Smeaton (1724–1792) of England and J. V. Poncelet of France (1788–1867). This resulted in waterwheels having efficiencies in the range of 60 to 70%. At the same time, reaction turbines (somewhat akin to modern lawn sprinklers) were being considered by several workers. The great Swiss mathematician Leohard Euler (1707–1783) investigated the theory of operation of these devices. A practical application of the concept was introduced in France in 1807 by Monnoury de Ectot (1777–1822). His machines were, in effect, radially outward-flow

machines. The theoretical analyses of Burdin (1790–1893), a French professor of mining engineering who introduced the word "turbine" into engineering terminology, contributed much to our understanding of the principles of turbine operation and underscored the principal requirements of shock-free entry and exit with minimum velocity as the basic requirements for high efficiency. A student of Burdin, Benoit Fourneyron (1802–1867), was responsible for putting his teacher's theory to practical use. Fourneyron's work led to the development of high-speed outward-flow turbines with efficiencies on the order of 80%. The early work of Fourneyron resulted in several practical applications and the winning of the coveted 6000-franc prize in 1833. After nearly a century of development, Bélidor's tub wheel had been officially improved. Fourneyron spent the remaining years of his life developing some 100 turbines in France and Europe. Some turbines even found their way to the United States, the first around 1843.

As successful as the Fourneyron turbines were, they lacked flexibility and were efficient over only a narrow range of operating conditions. This problem was addressed by S. Howd and U. A. Boyden (1804–1879). Their work evolved into the concept of an inward-flow motor as a result of the work of James B. Francis (1815–1892). The modern Francis turbine, shown in Fig. 10, is the result of this line of development. At the same time, European engineers addressed the idea of axial-flow machines, which today are represented by "propeller" turbines of both fixed-pitch and Kaplan types, as shown in Fig. 11.

Just as the vertical-axis tub wheels of Bélidor evolved into modern reaction turbines of the Francis and Kaplan types, development of the classical horizontal-axis waterwheel reached its peak with the introduction of the impulse turbine. The seeds of development were sown by Poncelet in 1826 with his description of the criteria for an efficient waterwheel. These ideas were cultivated by a group of California engineers during the late 19th century, one of whom was Lester A. Pelton (1829–1908), whose name is given to the Pelton wheel (shown in Fig. 12), which consists of a jet or jets of water impinging on an array of specially shaped buckets closely around the periphery of a wheel.

Modern impulse units are generally of the Pelton type and are restricted to relatively high-head applications (Fig. 3). One or more jets of water impinge on a wheel containing many curved buckets. The jet stream is directed inward, sideways, and outward, thereby producing a force on the bucket, which in turn results in a torque on the shaft. All kinetic energy leaving the runner is "lost." A draft tube is generally not used because the runner operates under approximately atmospheric pressure and the head represented by the elevation of the unit above tailwater cannot be used. Because this is a high-head device, the loss in available head is relatively unimportant. The Pelton wheel is a low-specific speed device. Specific speed can be increased by the addition of extra nozzles, with the specific speed increasing by

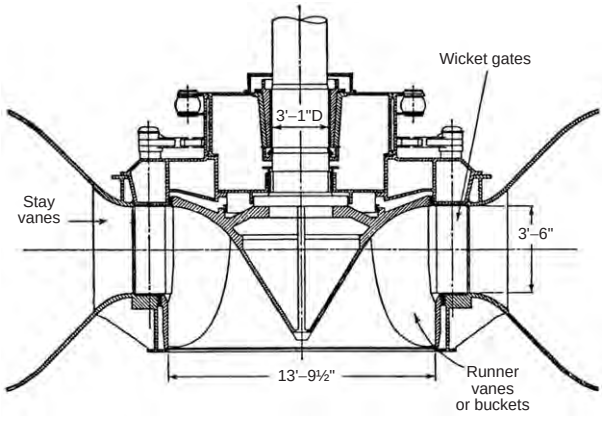


FIGURE 10 Francis turbine. From Daily (1950).

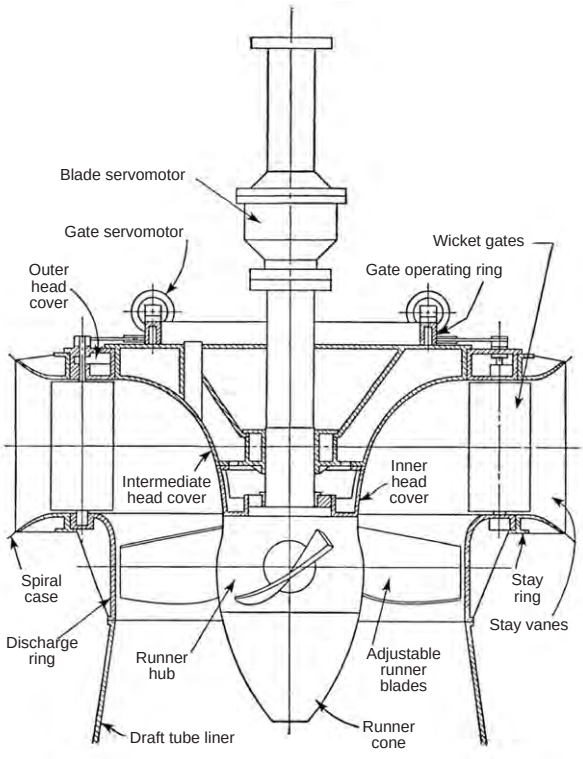


FIGURE 11 Smith–Kaplan axial-flow turbine with adjustable-pitch runner blades. From Daily (1950).

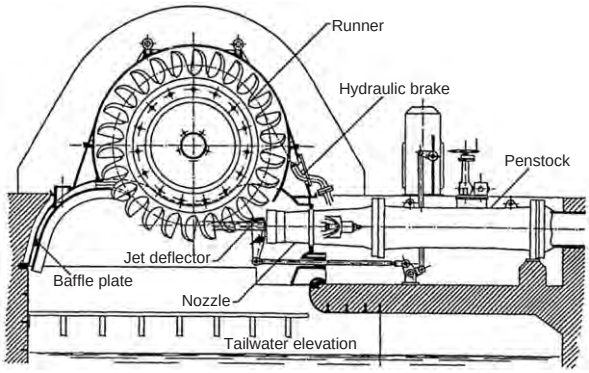


FIGURE 12 Cross section of a single-wheel, single-jet Pelton turbine. This is the third highest head Pelton turbine in the world ($H = 1447$ m, $n = 500$ rpm, $P = 35.2$ MW, $N_s \sim 0.038$). Courtesy of Vevey Charmilles Engineering Works. Adapted from Raabe, J. (1985), "Hydro Power: The Design, Use, and Function of Hydromechanical, Hydraulic, and Electrical Equipment." VDI-Verlag, Dusseldorf, Germany.

the square root of the number of nozzles. Specific speed can also be increased by a change in the manner of inflow and outflow. Special designs such as the turbo and crossflow turbines are examples of relatively high-specific speed impulse units.

Over more than 250 years of development, many ideas were tried. Some were rejected, whereas others were retained and incorporated into the design of the hydraulic turbine as we know it today. This development has resulted in highly efficient devices, with efficiencies as high as 96% in the larger sizes. In terms of design concept, these fall into roughly three categories: reaction turbines of the Francis design, reaction turbines of the propeller design, and impulse wheels of the Pelton type.

Most Pelton wheels are mounted on a horizontal axis, although newer vertical-axis units have been developed. Because of physical constraints on orderly outflow from the unit, the maximum number of nozzles is generally limited to six or less. Although the power of a reaction turbine is controlled by the wicket gates, the power of the Pelton wheel is controlled by varying the nozzle discharge by means of an automatically adjusted needle, as illustrated in Fig. 12. Jet deflectors, or auxiliary nozzles, are provided for emergency unloading of the wheel. Additional power can be obtained by connecting two wheels to a single generator or by using multiple nozzles. Because the needle valve can throttle the flow while maintaining essentially constant jet velocity, the relative velocities at entrance and exit remain unchanged, producing nearly constant efficiency over a wide range of power output.

5.1 Performance Comparison

The physical characteristics of various runner configurations are summarized in Fig. 13. It is obvious that the configurations change with speed and head. Impulse turbines are efficient over a relatively narrow range of specific speed, whereas Francis and propeller turbines have a wider useful range. An important consideration is whether or not a turbine is required to operate over a wide range of load. Pelton wheels tend to operate efficiently over a wide range of power loading due to their nozzle design. In the case of reaction machines that have fixed geometry, such as Francis and propeller turbines, efficiency can vary widely with load. However, Kaplan turbines can maintain high efficiency over a wide range of operation conditions. The decision of whether to select a simple configuration with a relatively "peaky" efficiency curve or take on the added expense of installing a more complex machine with a broad efficiency curve will depend on the expected operation of the plant and other economic factors.

Note in Fig. 13 that there is an overlap in the range of application of various types of equipment. This means that either type of unit can be designed for good efficiency in this range but that other factors, such as generator speed cavitation and cost, may dictate the final selection.

5.2 Speed Regulation

The speed regulation of a turbine is an important and complicated problem. The magnitude of the problem varies with size, type of machine and installation, type of electrical load, and whether or not the plant is tied into an electrical grid. It should also be kept in mind that runaway or "no load" speed can be higher than the design speed by a factor as high as 2.6. This is an important design consideration for all rotating parts, including the generator.

The speed of a turbine has to be controlled to a value that matches the generator characteristics and the grid frequency,

$$n = \frac{120f}{N_p}, \tag{2}$$

where n is the turbine speed in revolutions per minute (rpm), f is the required grid frequency in hertz, and N_p is the number of poles in the generator. Typically, N_p is in multiples of 4. There is a tendency to select higher speed generators to minimize weight and cost. However, consideration has to be given to speed regulation.

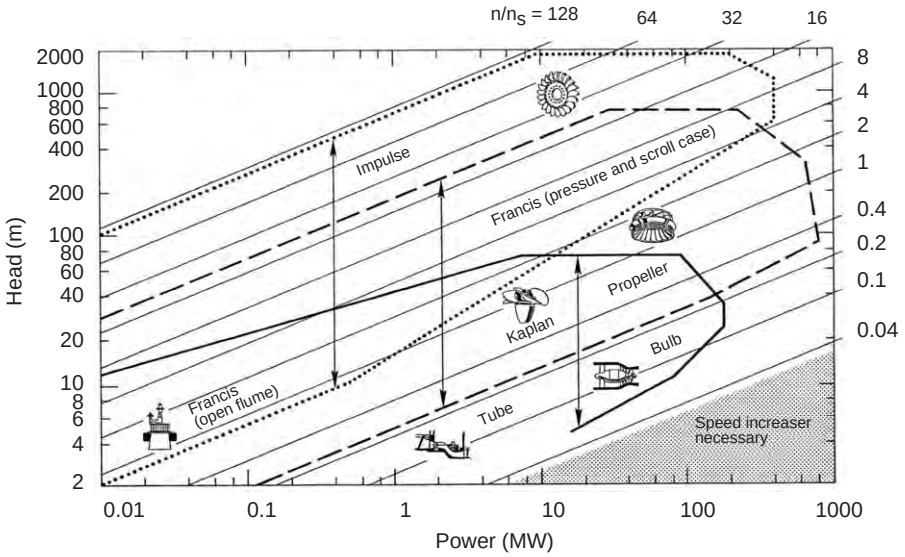


FIGURE 13 Application chart for various turbine types. n/n_s ratio of turbine speed in revolutions per minute n to specific speed defined in the system; $n_s = nP^{1/2}/H^{3/4}$ (with P in kilowatts). From Arndt, R. E. A. [1991]. Hydraulic turbines. *In* "Hydropower Engineering Handbook" [J. S. Gulliver and R. E. A. Arndt, Eds.]. McGraw–Hill, New York.

It is beyond the scope of this subsection to discuss the question of speed regulation in detail. Regulation of speed is normally accomplished through flow control. Adequate control requires sufficient rotational inertia of the rotating parts. When load is rejected, power is absorbed, and this accelerates the flywheel; when load is applied, some additional power is available from deceleration of the flywheel. Response time of the governor must be carefully selected because rapid closing time can lead to excessive pressures in the penstock.

A Francis turbine is controlled by the opening and closing of the wicket gates, which vary the flow of water according to the load. The actuator components of a governor are required to overcome the hydraulic and frictional forces and to maintain the wicket gates in a fixed position under steady load. For this reason, most governors have hydraulic actuators. On the other hand, impulse turbines are more easily controlled. This is due to the fact that the jet can be deflected or an auxiliary jet can bypass flow from the power-producing jet without changing the flow rate in the penstock. This permits long delay times for adjusting the flow rate to the new power conditions. The spear or needle valve controlling the flow rate can close quite slowly, say in 30 to 60 s, thereby minimizing any pressure rise in the penstock.

Several types of governors are available that vary with the work capacity desired and/or the degree of sophistication of control. These vary from pure mechanical, to mechanical–hydraulic, to electrohydraulic. Electrohydraulic units are sophisticated pieces of equipment and would not be suitable for remote regions. The precision of governing required will depend on whether the electrical generator is synchronous or asynchronous (i.e., induction type). There are advantages to the induction type of generator. It is less complex and, therefore, less expensive; however, it typically has slightly lower efficiency. Its frequency is controlled by the frequency of the grid it is feeding into, thereby eliminating the need for an expensive conventional governor. It cannot operate independently and can only feed into a network, and it does so with a lagging power factor that may or may not be a disadvantage, depending on the nature of the load. For example, long transmission lines have a high capacitance, so in this case the lagging power factor may be an advantage.

Speed regulation is a function of the flywheel effect of the rotating components and the inertia of the water column of the system. The start-up of the rotating system, t_s, is given by

$$t_s = \frac{I\omega^2}{P} \tag{3}$$

where I is the moment of inertia of the generator and turbine (in kilogram-meters), ω is the rotational speed of the unit (in radians/second), and P is the power generated (in watts).

The starting-up time of the water column, t_p, is given by

$$t_p = \frac{\Sigma LV}{gH} \tag{4}$$

where L is the length of water column segment, V is the velocity in each segment of the water column, and g is the acceleration of gravity.

For good speed regulation, it is desirable to keep $t_s/t_p > 4$. Lower values can also be used, although special precautions are necessary in the control equipment. It can readily be seen that higher ratios of t_s/t_p can be obtained by increasing I or decreasing t_p. Increasing I implies a larger generator, which also results in high costs. The start-up time of the water column can be reduced by reducing the length of the flow system, by using lower velocities, or by adding surge tanks, which essentially reduce the effective length of the conduit. A detailed analysis should be made for each installation because, for a given length, head, and discharge, the flow area must be increased to reduce t_p, leading to associated higher construction costs.

5.3 Cavitation and Turbine Setting

Another factor that must be considered prior to equipment selection is the evaluation of the turbine with respect to tailwater elevations. Hydraulic turbines are subject to pitting due to cavitation. For a given head, a smaller, lower cost, high-speed turbine runner must be set lower (i.e., closer to tailwater or even below tailwater) than does a larger, higher cost, low-speed turbine runner. Also, atmospheric pressure or plant elevation above sea level is a factor, as are tailwater elevation variations and operating requirements. This is a complex subject that can be accurately resolved only through model tests. Every runner design will have different cavitation characteristics. Therefore, the anticipated turbine location or setting with respect to tailwater elevations is an important consideration in turbine selection.

Cavitation is not normally a problem with impulse wheels. However, by the very nature of their operation, cavitation is an important factor in reaction turbine installations. The susceptibility to cavitation is a function of the installation and the turbine design. This can be expressed conveniently in terms of Thoma's sigma, defined as

$$\sigma_T = \frac{H_a - H_v - z}{H}, \tag{5}$$

where H_a is the atmospheric pressure head, H_v is the vapor pressure head (generally negligible), and z is the elevation of a turbine reference plane above the tailwater (Fig. 7). Draft tube losses and the exit velocity head have been neglected.

The parameter σ_T must be above a certain value to avoid cavitation problems. The critical value of σ_T is a function of specific speed. The Bureau of Reclamation suggested that cavitation problems can be avoided when

$$\sigma_T > 0.26 N_S^{1.64} \tag{6}$$

where N_S is the specific speed of the turbine or $N_S = n(\text{rpm})\ P(\text{kW})^{1/2}/H^{5/4}$ (feet). Equation (6) does not guarantee total elimination of cavitation; it guarantees only that cavitation is within acceptable limits. Cavitation can be totally avoided only if the value σ_T at an installation is much greater than the limiting value given in Eq. (6). The value of σ_T for a given installation is known as the plant sigma σ_P. Equation (6) should only be considered as a guide in selecting σ_P, which is normally determined by a model test in the manufacturer's laboratory. For a turbine operating under a given head, the only variable controlling σ_P is the turbine setting z. The required value of σ_P then controls the allowable setting above tailwater:

$$z_{allow} = H_a - H_v - \sigma_p H \tag{7}$$

It must be borne in mind that H_a varies with elevation. As a rule of thumb, from the sea level value of 10.3 m, H_a decreases by 1.1 m for every 1000 m above sea level.

6. ECOLOGICAL CONSIDERATIONS

The ecological impacts of hydroelectric power production that are perceived depend on the scale at which attention is focused. At the global environmental scale, hydropower is still seen as a clean form of energy. It is a replacement of fossil fuel power production, which is an inherently dirty source of energy. Terrestrial ecology is disturbed by mining activities, transportation of the fuel, and plants that burn the fuel. Water from rivers and lakes that is used to cool power generation systems returns hotter than normal, altering the ecology of these water bodies. Finally, atmospheric impacts of fossil fuel power production are substantial, including the production of nitrous oxides, sulfur compounds, heavy metals, and carbon dioxide. The future state of the earth and its atmosphere is affected by our power production through burning fossil fuels.

Although hydroelectric power production on the global scale is a relatively clean alternative to power produced through fossil fuels, on the local scale (i.e., the river system) the ecological impacts of hydropower can be significant. The dam and reservoir,

which most hydroelectric power facilities require, significantly alter the aquatic habitat and the species present. A dam will typically hinder migrating fish species in their spawning activities and cause water quality problems such as dissolved gas bubble disease. The reservoir used for storage can alter flows, such that floods and droughts are less frequent, and will catch most of the sediment. This will alter the river morphology in the downstream reach because the flood flows are important to forming river morphology. In tropical climates, parasite transmission can increase below a dam (e.g., below the Aswan High Dam). With extensive development of dams, a river will become a series of reservoirs and the river habitat will have disappeared entirely, such that many clams or mussel species will no longer exist and the dominant fish species will have been altered.

There are only selected locations on a river that are appropriate for dams and reservoirs. These locations are also the best for recreational activities such as canoeing, rafting, and fishing. Therefore, there is significant public interest in any proposed dam/hydropower facility that is proposed. As the prospective locations for dams and hydropower facilities are developed, there are fewer remaining locations for the recreational activities that need to compete for these sites. From Fig. 4, it can be seen that between 70 and 73% of the economic hydropower potential of Europe and North American has been developed. The recreational and environmental interests of the two continents need to be considered, and it is unlikely that much of the remaining hydropower potential of these two continents will be developed. However, in Asia, Africa, and South America, there is substantial hydropower remaining, and the power needs of the 21st century will be significant. The building of new dams and hydropower facilities in these regions is more likely.

7. ADVANTAGES AND DISADVANTAGES OF HYDROPOWER DEVELOPMENT

It is obvious that hydropower development is a renewable energy source, and this is a significant advantage. When hydropower development occurs at existing dam sites, the environmental impact is often minimal. However, significant changes can occur when an existing dam site is developed for power generation. For example, water that would normally flow over a spillway, where a large amount of aeration would take place, is now funneled through turbines, where little or no aeration occurs. The resulting substantial difference in the rate of aeration at a given point in the river can have a notable influence on the dissolved oxygen content of a considerable reach of river. In certain areas, especially those near large municipalities, the environmental impact can be significant.

Although the fuel costs of hydropower plants are negligible, their construction and capital equipment costs are usually substantially greater per unit of installed capacity than those of thermal power plants. This can be seen as either a benefit or a detriment. The economic feasibility of hydropower development is very sensitive to the difference between a discount rate, used to bring future income and costs to present value, and an assumed escalation rate, used to predict the future cost of electricity. The discount rate is taken as the load interest rate, which is usually fixed. The escalation rate is an estimate, which can vary greatly. For example, during a period of escalating inflation, a hydroelectric project that is already in place looks like a very good investment.

When considering hydropower development within a given region, it is also important to look at the overall economics of that region. For example, it should be noted that hydropower development means that a substantially larger percentage of the investment capital can stay within a given region because much of the developmental work can be done by local engineers and contractors. The more sophisticated coal-fired and nuclear power plants are designed and built by specialized contractors, and this often means that large amounts of capital leave the local economy. In many instances, the same is true for the amount of capital necessary for fuel for thermal power plants. This substantial drain on the economy can be very significant. In addition, hydropower facilities require minimal maintenance and do not have the same requirements for skilled personnel as do the more sophisticated thermal power plants.

There are other advantages of hydropower, and especially of small hydropower suitable as appropriate technology in less developed economic regions. Many future possibilities in small hydropower development will depend on the economic climate. There is a definite market for small turbine technology. If this market is developed, one could expect that significant improvements, both in operational characteristics of turbines and in reduced cost, can lead to small hydropower facilities being more cost-effective. There are many other aspects of small hydropower that have not yet been explored. In many cases in developing nations, small hydropower stations becomes the catalysts for the development of

small manufacturing facilities. One case in point is electrolytic manufacturing of fertilizer. It should also be noted that the current miniprocessor technology has developed to the point where it is feasible to operate a system of small hydropower plants on a completely automated basis, with only a traveling crew of workers needed for maintenance. In some cases, the number of small sites developed could supply the same amount of power as could one large nuclear power plant, without the safety or security hazards normally involved with a large-scale development. If the total amount of power is distributed over several small plants, overall reliability can increase because it is very unlikely that all plants would suffer outages at the same time.

However, there are also several disadvantages to small hydropower development. The most obvious is the fact that economy of scale does not prevail. This results in high initial cost for a relatively low installed capacity. In many cases, these plants are run-of-river; that is, their capability for generating power fluctuates wildly with the seasons, and this prevents a system of small power plants from acting as an equivalent base load plant. In many areas of the world, peak power is available during the late spring, whereas peak demand occurs during the midsummer or midwinter. This mismatch of power need and availability can be quite serious.

In addition to the lost economy of scale, there are other disadvantages that relate to the head available. For example, low-head facilities are those in which the available head is less than approximately 20 m. Because available power is proportional to the product of flow and head, larger amounts of flow must be handled to generate a given power level at lower head. Thus, the size of the machine increases, producing a disproportionate increase in cost for the amount of power developed.

SEE ALSO THE FOLLOWING ARTICLES

Coal Industry, History of • *Electricity Use, History of* • *Hydropower Economics* • *Hydropower, Environmental Impact of* • *Hydropower Resettlement Projects, Socioeconomic Impacts of* • *Hydropower Resources* • *Manufactured Gas, History of* • *Natural Gas, History of* • *Nuclear Power, History of* • *Oil Industry, History of* • *Wind Energy, History of* • *Wood Energy, History of*

Further Reading

Armstrong, E. L. (1985). The global outlook for additional hydropower use. In "Waterpower '85." American Society of Civil Engineers, New York.

Arndt, R. E. A. (1981). Cavitation in fluid machinery and hydraulic structures. *Annu. Rev. Fluid Mech.* **13,** 273–328.

Chapman, D. W. (1986). Salmon and steelhead abundance in the Columbia River in the nineteenth century. *Trans. Am. Fish. Soc.* **155,** 662–670.

Chappel, J. R. (1984). "The Future of Hydropower." EG&G Idaho, Idaho Falls, ID.

Daily, J. W. (1950). Hydraulic machinery. *In* "Engineering Hydraulics" (H. Rouse, Ed.). John Wiley, New York.

Deudney, D. (1981). "Rivers of Energy: The Hydropower Potential, Worldwatch Paper 44." Worldwatch Institute, Washington, DC.

Federal Energy Regulatory Commission. (1992). "Hydroelectric Power Resources of the United States, Developed and Undeveloped," FERC-0070. FERC, Washington, DC.

Gulliver, J. S., and Arndt, R. E. A. (1991). "Hydropower Engineering Handbook." McGraw-Hill, New York.

Gulliver, J. S., and Dotan, A. (1984). Cost estimates for hydropower at existing dams *J. Energy Eng.* **110,** 204–214.

International Electrotechnical Commission. (1963). "International Code for the Field Acceptance Tests of Hydraulic Turbines," Publication 41.

Isom, B. G. (1969). The mussel resource of the Tennessee River. *Malacologia* **7,** 397–425.

Merritt, R. H. (1979). "Creativity, Conflict, and Controversy: A history of the St. Paul District," U.S. Army Corps of Engineers, 008-022-00138-7. U.S. Government Printing Office, Washington, DC.

Rouse, H., and Ince, S. (1963). "History of Hydraulics." Dover, New York.

Shea, C. P. (1988). "Renewable Energy: Today's Contribution, Tomorrow's Promise," Worldwatch Paper 81. Worldwatch Institute, Washington, DC.

Smith, N. (1980 January). The origins of the water turbine. *Sci. Amer.* **242**(1).

Van der Schalie, H. (1972). World Health Organization Project Egypt 10: A case history of a schistosomiasis control project. *In* "The Careless Technology: Ecology and International Development" (J. T. Farver and J. P. Milton, Eds.). National History Press, Garden City, NY.

Wirth, B. D. (1997). Reviewing the success of international flooding of the Grand Canyon. *Hydro Rev.* **16**(2), 10–16.

Hydropower Resettlement Projects, Socioeconomic Impacts of

ADRIAN C. SLEIGH.
National Centre for Epidemiology and Population Health,
Australian National University
Canberra, Australian Capital Territory, Australia

SUKHAN JACKSON
School of Economics, University of Queensland
Brisbane, Queensland, Australia

1. **Introduction**
2. **Hydropower for Energy: Past and Present**
3. **Purposes of Hydrodams**
4. **The Height and Might of Dams**
5. **Opposition to Dams**
6. **Proliferation of Large Dams around the World**
7. **Winners and Losers in Hydropower Projects**
8. **Displaced, Host, Upstream and Downstream Communities**
9. **Construction Workers**
10. **Indigenous People and Gender Effects**
11. **Planning Blights and Intergenerational Differences**
12. **Health Impacts**
13. **Health Impact Assessment**
14. **Conclusion**

Glossary

anadromous Fish born in fresh water that descend to live in the ocean before ascending the river again to spawn and reproduce.
eutrophic Nutrient-enriched (water); for reservoirs this is usually due to the large biomass of vegetation trapped within the reservoir when it filled.
head For a hydrodam, the vertical distance between the water surface in the reservoir and the exit of water from the power-generating turbines.
oustees A word used by social scientists to describe people forced to leave their habitats and livelihoods by the reservoirs and related works of dams or other large projects.
reparation Compensation for a loss. For involuntary resettlers, this should cover in money or kind (or both) what they have lost and what they need to start again.
salinization Accumulation of salt in soil or water.
silt Sediment deposited by rivers and composed of fine soil particles that are usually rich in nutrients needed for agriculture.
waterlogging Saturation of soil with water.

In November 2000 the World Commission on Dams (WCD) released a landmark report, the first rigorous and inclusive global review of large dams. Worldwide, up to a third of these dams are for hydropower and another third are multipurpose. This article focuses on the social impact of such dams, especially the associated involuntary resettlement that has affected millions of people and will affect millions more.

1. INTRODUCTION

The World Commission on Dams (WCD) was formed in 1998 by a joint initiative of the World Conservation Union (IUCN) and the World Bank (WB) after a historic meeting of leading dam proponents and opponents in Switzerland, with proceedings reported by Dorcey and others. The WCD was asked to discover the truth about the cost, effect, and benefit of large dams and it functioned

 © 2004 Elsevier Inc. All rights reserved.

independently for 2 years at a cost of $10 million. The money was well spent, and at the end of that period the WCD produced a comprehensive report and numerous support documents, and then dissolved itself. The report and support documents remain on the Internet available to all, and they have changed forever the debate about dams. Unfounded assertions have been replaced with independent observations that constitute the largest multifaceted knowledge base on dams. The balanced analysis, development-oriented core values, and suggested guidelines are already influencing decisions about dams. Social impacts are now prominent elements in rights-and-risks assessments suggested for decision making about all future dams. Some governments have rejected this approach, but the overall views about dams and development have surely changed.

At the end of this article is a list of books, articles, reports, and other documents that support and expand on all the issues raised. The most useful general source of additional information is the "Dams and Development" report of the World Commission on Dams and many other documents available on its Web site. Both McCully and Smith give extensive accounts of the history of dams. Several references listed address aspects of India's Sardar Sarovar and other Narmada river dams, as well as China's Three Gorges dam; these are especially informative case studies for the political economy and social impacts of very large dams in the countries most committed to dam-based development. The works of Cernea, Scudder, McCully, Brody, Dorcey, Goodland, and Jackson have broader application for social impacts of large dams. Health impacts are covered by works cited for Hunter, Jobin, McCully, Sleigh, and the World Health Organization.

2. HYDROPOWER FOR ENERGY: PAST AND PRESENT

Water manipulation dates from the dawn of civilization. Mesopotamians built dams and irrigation canals at least 8000 years ago and town water supply systems operated from 3000 BC in Jordan. Ancient water systems and dams have been found in Asia and South America. Humans began to capture energy from flowing water using wheels with buckets to lift water in ancient Egypt and Sumeria. The Romans used watermills to grind corn 2100 years ago. A thousand years ago, England already had thousands of watermills. Nineteenth-century Britain led the charge to develop large dams and built nearly 200 above 15 m, usually to store water for the growing industrial cities. In 1832, Benoit Fourneyron had developed a water turbine to capture the potential energy of falling water, far more useful than flow-dependent waterwheels. Engineers later used turbines to generate electricity, and the first hydro plant began producing in Wisconsin in 1882. Hydropower integrates four technologies—dam building, water turbines, electric generators, and electricity transmission to other locations—enabling human society to produce energy from water at one location and use it at another. Such a useful invention soon spread and the size of dams and power stations increased from 30-m heads (in 1900) to 200-m heads (1930s). Modern dams began to spread to the developing world. In 1902, the British built the Low Aswan Dam to regulate the Nile and irrigate cotton fields and eventually India and China began to build many large dams. After 1950 these two countries invested huge sums in dams, foregoing other potential developments. From 1950 to 1980, the Chinese built about 600 such dams per year, but quality was poor. Many dams burst, some catastrophically. For example, in 1975 more than 250,000 people were killed in Henan when dams burst after an extraordinary rainfall. The early dam failure rate in China was much higher than elsewhere.

The World Bank, multilateral development banks, the Food and Agriculture Organization (FAO) and UN Development Program (UNDP) played a big role in promoting large dams and irrigation schemes in developing countries. Bilateral funds through the U.S. Agency for International Development (USAID) and the British Overseas Development Administration (ODA) also helped to plan and fund many dams. Since the 1970s, aid has supported construction companies in rich countries once the work at home had declined. Every year about $20 billion is spent on large dams, and 100 to 200 are commissioned.

3. PURPOSES OF HYDRODAMS

The water levels of most of the world's rivers fluctuate with seasonal rains, melting snows, or droughts, and in the arid zones there are flash floods. It makes sense to capture and store the abundant waters and later release it during the low flow periods or use it for power. So dams have been built for the purposes of irrigation, municipal water supply, and hydroelectricity generation. Flood control and transportation benefits are also important.

Cities are growing all over the developing world, and they require water supplies and reservoirs. Since the beginning of the 20th century, hydropower has steadily increased and it now supplies 19% of worldwide electricity.

Many large dams are multipurpose. For example, the Hoover Dam in the United States provides water for irrigation in California's Imperial Valley, piped water for the residents of Los Angeles and central Arizona, hydroelectricity to consumers in the region, and it controls flood in the lower Colorado River Valley.

4. THE HEIGHT AND MIGHT OF DAMS

Dams are described as major, large, or small. The size is usually based on height, the reservoir area, the generation capacity, or the area under irrigation. The International Commission on Large Dams (ICOLD) defines a large dam as one that is more than 15 m from foundation to crest (taller than a four-story building). But tallness is not always an indication of the dam's impact on ecology and on the number of displaced people. A dam is defined as major (or a super dam) on the basis of either (1) a height of at least 150 m, volume at least 15 million cubic meters, and reservoir storage at least 25 cubic kilometers or (2) a generation capacity of at least 1000 megawatts. It is difficult to calculate the number of large and major dams throughout the world partly because of unreliable information from China and the former USSR. Pearce estimated in 1992 that there were then more than 100 super dams with a height of more than 150 m, of which three quarters were built in the previous 35 years. Since then, others have been built and many more are planned or under way.

5. OPPOSITION TO DAMS

Since the 1970s, natural scientists, hydrologists, economists, and even engineers have become ever more concerned about the negative impacts of large and major dams. The Hoover Dam and the Grand Coulee Dam may have been regarded as successes, but thousands of other large dams performed poorly and were abject failures as instruments of development. First the economists began to criticize the cost overuns and schedule delays of many dams (1960s). Then environmentalists began to report on adverse effects on biodiversity and riparian ecology (1970s). Ecologists now point to the natural advantages of silt-depositing annual floods that recharge soils, groundwater, wells, and wetlands. The reservoirs created by large dams extinguish wildlife, fish, herbs, and fruits, natural bounties that sustain the rural poor and indigenous populations. The environmental focus is now on sustainability of power generation systems, including hydropower.

By the 1980s social scientists began to publicize the negative impacts on oustees, communities hosting those resettled, construction workers attracted to the project, and downstream fishers and farmers. Michael Cernea was especially important for bringing these social concerns into the World Bank, and he has written widely on the topic. For several decades, health researchers have been reporting a great variety of adverse health impacts, especially infectious diseases associated with tropical dams. These reports are well summarized by the works of Hunter and others, Jobin, and the World Health Organization. Finally the project-affected people have globally united their dissenting voices (1990s) culminating in their Declaration of Curitiba. They demand that dam builders put an end to violent expulsions, pay reparations for livelihood losses, and grant negotiating rights to all those affected.

Impact assessments are now obligatory for virtually all dam planners but are often subject to manipulation, censorship or delay until finance, design, or construction decisions become irreversible. Some well-documented examples include the grievously flawed impact assessments for India's Narmada dams and China's Three Gorges Dam. But many politicians and dam builders remain so confident of their product that few can even imagine that dams could do more harm than good. Yet for many large dams, the harm inflicted on people and nature is irreversible.

6. PROLIFERATION OF LARGE DAMS AROUND THE WORLD

After World War II, many rich nations began to export their 20th century model of economic development to the poor nations of the world, including the emphasis on building large dams. Now at least 45,000 such dams exist, nearly half constructed in China since 1949. Nearly 40,000 of all large dams were commissioned after 1950. The WCD estimated that the total cost to governments exceeded $2 trillion, and most dams had substantial cost overruns and have performed below functional expectations. Despite such enormous investments,

systematic reviews of large dam performance were never undertaken and even crude economic outcomes were rarely evaluated. The failure to report negative outcomes, and the apparent (but usually unmeasured) benefits of some large hydropower, flood control, or multipurpose projects in North America and Europe, contributed to the post–World War II determination of many Asian, South American, and African leaders to build large dams for national development, with activity peaking in the decades after 1970. The 20th century leaders of dam-based development, in order of number of dams, were the United States, Spain, Canada, the United Kingdom, Italy, Norway, and Germany. The followers included China (now with about half the world total), India, Japan, South Korea, Turkey, Brazil, South Africa, and Mexico.

The WCD reported that expected benefits were invariably exaggerated by dam proponents, and this had helped create an unstoppable development momentum. About 90% of the global inventory of large dams is located in the 15 countries listed, but 125 other nations have at least one such dam and many more are planned.

Brody, in a theme paper prepared for the WCD, notes that pro-dam sentiment in the United States in the 1930s, 1940s, and 1950s influenced world opinion profoundly. This enthusiasm is captured perfectly by the American folk musician Woody Guthrie, who wrote "The Grand Coulee Dam" (for the Bonneville Power Authority in the Pacific Northwest) and set it to the tune of a famous lament (Lead Belly's immortal song "Irene Goodnight").

Uncle Sam took the challenge in the year of 'thirty-three,
For the farmer and the factory and for all of you and me,
He said 'Roll along, Columbia, you can ramble to the sea,
But river, while you're rambling, you can do some work for me.'

Now in Washington and Oregon you can hear the factories hum,
Making chrome and making manganese and light aluminum,
And there roars the flying fortress now to fight for Uncle Sam,
Spawned upon the King Columbia by the big Grand Coulee Dam. ©

Now that we know the adverse impact of large dams on many ecosystems, and millions of poor people worldwide, Guthrie's words seem rather naïve. But they must be read in the context of his bitter experience of the great 1930s drought, the resulting dustbowl in rural Oklahoma and neighboring states, and the forced exodus of millions of impoverished farmers after the bank foreclosures that resulted. The dams he supported on the Columbia river system were sources of hope to many Americans, and they did stimulate the north-west economy.

By the 1970s, the main locus of large dam building had moved to the developing world. Since then, the World Bank has provided the core of the massive finance needed and has led the development of guidelines to assess, monitor, and mitigate environmental and social impacts. Over that period, the social impacts became more noteworthy because dam building affected the poorer and more densely populated parts of the world at a time of increasing interest in human rights and environmental protection. Global reporting and access to information also improved. The adverse social and environmental impacts of large dams were widely (but haphazardly) reported and such information fed an ever-growing, and increasingly vocal, global opposition. Now most such dams are being built in developing countries. Meanwhile, in the United States, decommissioning dams has become more frequent than building them, a trend expected to spread to most of the developed world in the early 21st century. So the social impact of large dams is now largely an issue for the developing world, and the pro-dam lobbies in China, India, Iran, Turkey, Brazil, and many other countries remain very powerful. Anti-dam dissent is poorly tolerated in most of those countries, and intimidation and violence are frequent.

Now the World Bank is far less enthusiastic than before. Since the 1980s, it was stung by strident criticism of its support for dams, which included reference to its failure to follow its own guidelines for mitigating social impacts and even an accusation of a coverup of a state-supported massacre of nearly 400 indigenous people who refused to move for a WB-supported dam in Guatemala. McCully gives a full account of these criticisms in his book *Silenced Rivers,* published in 2001. In the 1990s, the World Bank began to withdraw from or avoid the most troublesome projects, such as India's Sardar Sarovar dam or China's Three Gorges dam. These dams are proceeding without World Bank finance. Even private and government international dam-finance agencies are now subject to increased scrutiny, and some have withdrawn from controversial projects.

By the late 1990s, the dam debate had reached an impasse, and only a substantial and independent global inquiry into large dams could bridge the gulf between those who opposed such dams and those who financed and built them. The resulting World

Commission on Dams included engineers, environmentalists, anti-dam activists, civil servants responsible for water resources, lawyers, scientists, and others drawn from all parts of the world. Most observers expect that fewer dams will be built in the future, planning and cost-benefit estimations will be more honest and comprehensive, and the development emphasis will focus on the rights, risks, and livelihoods for affected people. As well, technological advances in power generation, especially gas turbines (at first), then wind and solar power (next 10 years), and finally fuel cell (hydrogen-based and renewable) should make future hydropower projects less frequent. Hydroelectric power and central electrification will become progressively more expensive relative to other options, despite the fact that hydrodams performed closer to their estimated economic and production targets than did most irrigation dams in the WCD assessment.

Asia has so far exploited only about 11% of its hydropower capacity and Africa about 3%, so the hydropower supply momentum in China and India, and perhaps Africa too, will persist for some time. China, India, Turkey, and Russia expressed grave doubts about the negative WCD findings and pro-dam sentiment remains powerful in other countries as well. The need to improve domestic and industrial water supplies, to electrify cities and rural areas, and to irrigate land to produce food for the growing world population are expected to be the key demand drivers of most future large dams.

7. WINNERS AND LOSERS IN HYDROPOWER PROJECTS

The winners in hydropower and other large dam projects are those who gain electricity, domestic water supplies, sanitation, drought relief, sustainable irrigation, flood control, improved water transport, industrial and employment opportunities, food security, better roads, telecommunications, and other benefits. As well, other economic multipliers may arise and whole regions may flourish. This is an impressive list and is the very basis of modernization and liberation from the drudgery of traditional low-technology rural life. Electricity, radio, and domestic water supplies completely transform the lives of those who gain these goods, and better health and education are the most obvious outcomes.

But the WCD (and many other reports before and after) found that the benefits of large dams were rarely so comprehensive, and the common river-basin disbenefits included waterlogging and salination of irrigated land, unseasonal reversal of downstream flows, dangerous river flows at unpredictable times, loss of vital silt deposits, encroachment of seawater in river deltas, and obstruction of migrating anadromous fish. Subtle changes included the switch to cash cropping, frequently with negative consequences for the nutrition of vulnerable women, children, and the elderly, and concentration of land ownership.

More direct and even more negative effects occur for those forced to leave their homes and land, for communities that are broken up, and for other communities that host the oustees. Migrants arrive to construct the dam or to take advantage of ecological changes they are better equipped to exploit, such as lake fishermen after reservoirs form. Funds distributed in compensation for losses are often embezzled or subject to nepotism. All these changes produce social friction and fights are common. Many dam-affected groups become sick due to the water impoundment or its ecological or socioeconomic effects. In many well documented cases (and others not documented), the evictions of oustees have been violent and sometimes lethal. Details of such ill-treatment are given in the publications of the WCD and McCully and in many other sources. All these groups described previously are the losers. They are not beneficiaries of development dams. Sometimes winners and losers coexist in a single community or even within one family.

In his paper on social impacts of large dams, prepared for WCD, Brody listed several important principles to consider, including the following:

1. Dams take one set of resources, a river and the lands along its banks, and transform them into another set of resources, a reservoir, hydropower, and irrigation. The means of production and livelihood for one group of people are changed into benefits for people living elsewhere.
2. Powerful economic and political interests attach to such undertakings. The larger the dam, the more costly the project, the greater its construction momentum.
3. The larger the project and the greater the momentum, the greater the risks to the social well-being and human rights of those directly affected.
4. For consultation and participation to be authentic and effective, they must take place in a way and at a time when they can influence

decisions about the project and mitigation of its impacts.

8. DISPLACED, HOST, UPSTREAM AND DOWNSTREAM COMMUNITIES

The adverse human outcomes of blocking the world's rivers have been even more dramatic than the ecological ones. But social impacts vary enormously within and among these groups. Upstream communities are either displaced directly, often with inadequate or no reparation, or become hosts for some of the oustees. The latter situation causes frequent conflict, especially (as is often the case) if land is scarce and resettlers arrive with more available capital than the hosts. The same tensions often arise for downstream communities that are forced to host oustees.

Upstream communities displaced are often the worst affected. In most dams studied by the WCD, and according to numerous other reports, compensation for oustees is almost always inadequate and sometimes is not offered at all. Dams have flooded 40 million to 80 million people off their land, with 16 million to 38 million displaced in India (differentially indigenous and always poor) and more than 10 million peasants displaced in China (just for the Yangtze basin). It seems probable (but usually unstudied) that many people never recovered from the ordeal, either economically or psychologically, and this trauma has been passed to the next generation. Displaced communities were usually fragmented and destroyed as cultural groups. Many more people lost their fisheries, the irrigation of annual floods, and the usufruct of flooded lands. Lost archeological and cultural sites could not be replaced.

Dam resettlement authorities often state they aim to offer land-for-land to people displaced. In fact, this is often not possible, but even if it is done the problems for those resettled are not over. The replacement land is likely to be less valuable, less fertile, less watered, and quite different to the land they occupied before. Essential infrastructure is often lagging or not provided at all. As the new land may be higher up, and sloping, water will be the first problem and may be an insurmountable obstacle to successful agricultural activity. The next obstacle will be to choose suitable crops, and these are likely to be unfamiliar to the farmer. Without scientific agricultural support, results will be poor. Even with that support, the first few crops may be disastrous. So resettled farmers need to be supported and monitored, and this must continue for several years. It rarely does. But unless resettlers livelihoods are established successfully and return to the level they had before or better, they will be long-term losers and may never recover.

Upstream communities not displaced must learn to live with a changed environment. Travel across the landscape is obstructed by the reservoir and new and unfamiliar industries and occupations arise. Inflation may affect the area due to the influx of money to build the dam. For very large reservoirs, the climate may change with relative humidity and temperature increasing and fog may become more frequent. There are other positive impacts as well. Transport may improve and journeys that once took days (by road or foot paths) may now take only 1 to 2 hours across the water. If the dam authority develops a boat service or creates a new township at the top of the reservoir to provide municipal services, people are better off.

Downstream communities are affected by large dams for hundreds of kilometers. The riparian impacts are greatest but the effects of trade, transport, and food supplies may be felt far away from the river basin. The benefits of electricity and domestic water supplies are often downstream. Some downstream communities are winners and others are losers. Farmers and fishers are at high risk of adverse impacts due to retention of silt by the dam, unseasonal water flows, loss of water to distant irrigators, and destruction of fisheries. Irrigation can be a mixed blessing; if not done well, it leads to salination and water logging, noted by the WCD to affect 20% of such areas.

9. CONSTRUCTION WORKERS

Large dams attract construction workers in large numbers. A temporary workforce of 10,000 to 30,000 is not uncommon and the peak of work may last 4 to 8 years. Sometimes whole families are employed, but often marginalized men, or women, are the workers recruited. Conflict is common and injuries and accidental death are frequent. Illnesses due to introduced or endemic infectious diseases, including sexual infections and AIDS, are expected. Health services are often well developed for this important group whose welfare is crucial to completion of the dam. Infrequently these services are made available to the local population or become the basis for boosting

regional health services after the dam is completed. If the latter occurs, the whole community benefits.

10. INDIGENOUS PEOPLE AND GENDER EFFECTS

The World Bank has expressed great concern for indigenous people affected by dams. These groups are so dependent on their local environments they cannot recover once they are removed. In India and South America, indigenous people have borne disproportionate shares of the displacement burdens. Many instances of violence and social catastrophes have been recorded. It is now thought that if indigenous persons are going to be affected, dams should not proceed unless equivalent lands can be found. Such substitute lands are quite improbable.

The gender effects of dam displacement have rarely been studied and were a special focus of the WCD report. The commission noted that in any situation where females were already disadvantaged dams often make things worse. Compensation, if paid, is usually given to husbands or sons, even if the group is matrilineal. This reflects the male gender bias of the state. Common lands lost are not compensated even if they were the location of food and herb gathering, small game hunting, and other social activities conducted exclusively by women. Women also suffer when violence erupts, which is more likely among the displaced. Once a family is impoverished, unequal females may become even worse off and may turn to prostitution and crime to survive. Gender-sensitive planning has been very rare for large dams and other infrastructure projects.

11. PLANNING BLIGHTS AND INTERGENERATIONAL DIFFERENCES

For many years and sometimes many decades before a large dam is built, the project is widely discussed. During this long period, all investment in the area is substantially curtailed by both the state and the local residents. Consequently, the people are poorer than they would have been if no dam was contemplated, and public services and infrastructures are run down or not provided at all. This planning blight should be considered when calculating compensation for oustees, but almost invariably it is ignored.

The most vulnerable groups for social impacts of large dams are the aged, children, and the infirm. Within families, a young couple might see the cash compensation and resettlement in a new house elsewhere, or even a chance to move to a city, as a golden opportunity to escape a dreary rural life. But their parents could be devastated by the same choices, become disoriented and depressed, and lose all hope. They would sink into poverty while their children rose to become relatively wealthy; in such circumstances, family harmony breaks down or parents are even abandoned. Other groups may be selectively disadvantaged, such as those who have no official residence in the place they live and must leave. This arises for up to ten per cent of displaced populations in China, and they receive nothing at all even if they lose a long-established business such as a restaurant or noodle bar. Others may be forced into the cash economy and quickly stripped of their life savings by the predatory behavior of swindlers they encounter in their new town. For many displaced persons—uncounted millions since the 1950s—the final result of their large dam has been loss of community, culture, burial sites, identity, habitat, and livelihoods. This leads to poverty and urban or rural squatting without adequate housing, water, electricity, food, or employment. They have no voice and are forgotten.

12. HEALTH IMPACTS

Dams built to stimulate development should not produce disease. But they do. Most information on the health impacts of large dams comes from studies conducted in Africa and the Caribbean. Vector-borne diseases have been the most noticeable disease effect, especially epidemics of parasitic and viral infections around large (or small) dams and irrigation systems. These begin soon after reservoirs form once the dams are closed. First among these is bilharzia, also known as snail fever, water belly, or schistosomiasis. The disease is caused by one or more pairs of blood flukes (schistosomes) whose eggs pass out in human urine or faeces, infect snails, multiply, and produce free swimming larvae that reinfect human hosts. Several species of schistosome infect humans, including one in China that also infects many domestic mammals. The disease is present in 73 countries including most of Africa, parts of the Middle East, some Caribbean islands, Brazil and several other South American countries, China, Laos, Cambodia, the Philippines, and Indonesia. There are numerous examples of entire communities becoming infected due to large dams; average life spans are shortened and many

people are debilitated and impoverished as a result. Attempts to prevent snail breeding in reservoirs and canals by engineering modifications or chemicals have met with limited success and are described in detail in Jobin's book *Dams and Disease*.

The next most reported dam-induced infection is malaria and this is spread even more widely than schistosomiasis. Other mosquito-borne infections include the nematode worm that causes elephantiasis, also known as filariasis, and several serious viral diseases, including Rift Valley fever, Japanese encephalitis, and dengue. Black fly populations can breed on dam spillways in parts of Africa and Latin America, and transmit the nematode worm that causes river blindness, also called onchocerciasis.

Much less reported, but perhaps of even greater importance, are the diseases of poverty and poor sanitation because they are made worse by inadequate resettlement. These problems include maternal anaemia, death in childbirth, childhood malnutrition, and infections such as diarrhoea, pneumonia, and tuberculosis. Because they were present before the dam was built and are relatively common, they are not often noted as problems exacerbated by large dams.

Even less well recognized and studied are the mental health problems produced by involuntary loss of habitat and livelihood. There is no doubt that they exist, but little is known about their consequences. Some indigenous people are so deeply affected that they die soon after displacement. Other oustees become depressed or suicidal. Violence and social breakdown are also serious problems and are inevitable if communities are displaced and not reformed with adequate support.

Poisoning and injury are well documented consequences of large dams. Methyl-mercury is formed in the reservoir and ascends the food chain, reaching humans via the fish. Other heavy metals may be collected in the runoff after flooding certain industrial plants. Organic solvents and other poisons can pollute reservoirs also. Biological toxins pollute the water if algae bloom or cyanobacteria proliferate, which is very likely in eutrophic waters typical of tropical reservoirs soon after they fill. Injury is a constant risk for construction workers and many are injured or killed by falling or after being struck by flying objects at the dam site or in the powerhouse or other related works. Drowning is also a risk after the reservoir forms, especially in tropical areas where people may start recreational swimming with no safety measures in place and little experience of deep water. Alcohol makes this risk higher. Inebriated revelers may swim out into reservoirs and be struck by boats around busy jetties or further out, or they may simply drown.

Finally there is a risk of a catastrophe inherent for all large dams. They can collapse and when overtopped or breached due to a geological fault, flood, or a landslide-induced giant wave, the results have sometimes been disastrous. Landslides may be made more likely by the earthquakes induced by the weight of water stored behind high dams. Accounts of such catastrophes, and seismic activity induced by impounded water, are given in the WCD report and in McCully's book *Silenced Rivers*. The most spectacular landslide example occurred in the Italian Alps in 1963. It caused a huge wave that rose 110 m above the 261-m wall of Vaiont Dam, overtopping the dam and killing 2600 people in villages located downstream. The most lethal example is the multiple dam collapse in Henan, China, in 1975, which killed at least 250,000 people downstream.

13. HEALTH IMPACT ASSESSMENT

Health impact assessment is not often requested by dam planners and builders. It involves a baseline (pre-dam) study of existing risks and an estimation of new risks that will or could arise. The at-risk community should then be followed and steps taken to reduce or prevent the identified health risks. In some instances, the risks are too high and cannot be controlled and the dam should not proceed. This almost never happens because such risks are not identified in time. Usually, at best, a health assessment representative is placed on an environmental or social panel to monitor the outcomes and is not resourced or supported with the needed array of public health expertise (epidemiology, nutrition, infectious diseases, entomology, malacology, toxicology). It is not possible to have much effect in these circumstances. In some cases the dam builder has investigated the risks well but not integrated the process with other preventive activity. This was done for China's Three Gorges Dam, but the geographers led by Chen and others made assessments without being connected to the health system, and the health personnel in the area were not involved. No follow-up activity resulted, and no mitigation of risks occurred. More than 1 million people are being displaced and many millions more affected along the 600-km reservoir or in the densely populated middle reaches of the river stretching hundreds of kilometers below the dam.

14. CONCLUSION

Many social benefits flow from well-planned, economically justified, and equitably executed large dams. But very few dams have been completed that way. In future, the rights-and-risks approach recommended by the World Commission on Dams, with full participation of project-affected people, is likely to be a requirement before attracting international finance for a new dam. Until then we have to address the enormous adverse social and health impacts created by many of the existing 45,000 large dams, a legacy of 20th-century enthusiasm that was often misplaced. Some of these dams have harmed a great many people, and some were successful as agents of development. It would be unfortunate if future dams did not take all this experience into account, and most people would agree that large dams should not be built if they make people sick.

Acknowledgments

Lyrics to "The Grand Coulee Dam" (words and music by Woody Guthrie) are used by permission. TRO—© Copyright 1958 (renewed), 1963 (renewed), 1976 Ludlow Music, Inc., New York, NY.

SEE ALSO THE FOLLOWING ARTICLES

Electricity, Environmental Impacts of • Hydropower Economics • Hydropower, Environmental Impact of • Hydropower, History and Technology of • Hydropower Resources • Hydropower Technology • Oil-Led Development: Social, Political, and Economic Consequences

Further Reading

Alvares, C., and Billorey, R. (1988). "Damming the Narmada." Third World Network/Appen, Penang, Malaysia.

Barber, M., and Ryder, G. (eds.). (1993). "Damming the Three Gorges. What Dam Builders Don't Want You to Know." Earthscan, London.

Brody, H. (2000). Social impacts of large dams equity and distributional issues (*Thematic Rev.,* World Commission on Dams), www. dams. org/docs/kbase/contrib/soc192.pdf (accessed September 24, 2003).

Cernea, M. M. (1988). Involuntary resettlement in development projects: Policy guidelines in World Bank-financed projects. *World Bank Technical Paper,* 80. World Bank, Washington, DC.

Cernea, M. M. (1999). "The Economics of Involuntary Resettlement: Questions and Challenges." World Bank, Washington, DC.

Cernea, M. M., and McDowell, C. (eds.). (2000). "Risks and Reconstruction: Experiences of Resettlers and Refugees." World Bank, Washington, DC.

Chen, Y., Shi, M., Zhao, M., Chen, M., Huang, X., Luo, X., Lie, B., Wu, Y., Yang, L., Yan, J., Wang, J., Zhong, D., Yao, S., Fan, F., Wang, D., and Han, A. (eds.). (1990). "Atlas of the Ecology and Environment in the Three Gorges Area of the Changjiang River." Science Press, Beijing.

Declaration of Curitiba (1997). "Affirming the Right to Life and Livelihood of People Affected by Dams." Approved at the First International Meeting of People Affected by Dams. Curitiba, Brazil, March 14, 1997, www. irn. org/programs/curitiba. html (accessed September 24, 2003).

Dorcey, T., Steiner, A., Acreman, M., and Orlando, B. (eds.). (1997). "Large Dams. Learning from the Past, Looking at the Future." Gland, Switzerland: World Conservation Union (IUCN) and World Bank Workshop Proceedings, April 11–12.

Fearnside, P. M. (1994). The Canadian feasibility study of the Three Gorges Dam proposed for China's Yangtzi River: A grave embarrassment to the impact assessment profession. *Impact Assessment* **12,** 21–53.

Goodland, R. (1997). Environmental sustainability in the hydro industry. *In* "Large Dams. Learning from the Past, Looking at the Future." (T. Dorcey, A. Steiner, M. Acreman, and B. Orlando, Eds.), pp. 69–102. Gland, Switzerland: World Conservation Union (IUCN) and World Bank Workshop Proceedings, April 11–12.

Hunter, J. M., Rey, L., Chu, K. Y., Adekolu-John, E. O., and Mott, K. E. (1993). "Parasitic Diseases and Water Resource Development." World Health Organization, Geneva.

Jackson, S., and Sleigh, A. (2000). Resettlement for China's Three Gorges Dam: Socio-economic impact and institutional tensions. *Communist and Post-Communist Studies* **33,** 223–241.

Jackson, S., and Sleigh, A. C. (2001). Political economy and socio-economic impact of China's Three Gorges Dam. *Asian Studies Review* **25,** 57–72.

Jobin, W. (1999). "Dams and Disease: Ecological Design and Health Impacts of Large Dams, Canals and Irrigation Systems." E & FN Spoon, New York.

McCully, P. (2001). "Silenced Rivers. The Ecology and Politics of Large Dams." Zed Books, London.

Morse, B., and Berger, T. (1992). "Sardar Sarovar: Report of the Independent Review." Resource Futures International, Ottawa.

Pearce, F. (1992). "The Dammed: Rivers, Dams and the Coming World Water Crisis." The Bodley Head, London.

Scudder, T. (1997). Social impacts of large dam projects. *In* "Large Dams. Learning from the Past, Looking at the Future" (T. Dorcey, A. Steiner, M. Acreman, and B. Orlando, eds.), 41–68. Gland, Switzerland: World Conservation Union (IUCN) and World Bank Workshop Proceedings, April 11–12.

Sleigh, A., and Jackson, S. (1998). Public health and public choice: Dammed off at China's Three Gorges? *The Lancet* **351,** 1449–1450.

Sleigh, A. C., and Jackson, S. (2001). Dams, development and health: A missed opportunity. *Lancet* **357,** 570–571.

Smith, N. A. F. (1971). "A History of Dams." Peter Davis, London.

World Commission on Dams (2000). "Dams and Development. A New Framework for Decision Making." Earthscan, London, www.dams.org (accessed September 24, 2003).

World Health Organization (1999). "Human Health and Dams." World Health Organization, Geneva (submission to the World Commission on Dams), www.dams.org/docs/kbase/working/health.pdf (accessed September 24, 2003).

Hydropower Resources

GAROLD L. SOMMERS
Idaho National Engineering and Environmental Laboratory
Idaho Falls, Idaho, United States

1. **The Resource**
2. **How Hydropower Resources are Assessed**
3. **Global Distribution**
4. **Current Resources**
5. **Potential of Hydropower Resources**

Glossary

benefit/cost ratio (B/C) Ratio of the present value of the economical benefit to the present value of the project cost, computed for comparable price level assumptions.

capability The maximum load that a generating unit, generating station, or other electrical apparatus can carry under specified conditions for a given period of time without exceeding approved limits of temperature and stress.

capacity The maximum power output or load for which a turbine generator, station, or system is rated.

dam A wall or structure built across a valley or river for storing water.

dependable capacity The load-carrying ability of a hydropower plant under adverse hydrologic conditions for a specified time interval and period.

distribution system The portion of an electric system dedicated to delivering electric energy to an end user; the distribution system "steps down" power from high-voltage transmission lines to a level that can be used in homes and businesses.

diversion dam A dam used in conjunction with long tunnels, canals, or pipelines to divert water to a powerhouse located a distance from the dam.

drawdown The distance that the water surface of a reservoir drops from a given elevation as the result of withdrawal of water.

energy The capacity for performing work; the electrical energy term generally used is kilowatt-hours and represents power (kilowatts) operating for some time period (hours).

forebay The water intake area for a canal, penstock, or turbine designed to reduce water velocity and turbulence so as to settle suspended material and keep it from entering the system.

generator A machine that converts mechanical energy into electric energy.

gigawatt (GW) Unit of electric power equal to 1 million kW.

head The difference in elevation between the headwater surface above and the tailwater surface below a hydroelectric power plant under specified conditions.

horsepower A unit of the rate of doing work equal to 33,000 foot pounds per minute or 745.8 W (Britain), 746 W (United States), or 736 W (Europe).

hydroelectric plant or hydropower plant An electric power plant in which the turbine generators are driven by falling water.

hydroelectric power Electric current produced from waterpower.

hydrology The scientific study of the properties, distribution, and effects of water on the earth's surface, in the soil and underlying rocks, and in the atmosphere.

hydrologic cycle Water constantly moving through a vast global cycle in which it evaporates from lakes and oceans, forms clouds, precipitates as rain or snow, and then flows back to the ocean; the energy of this water cycle, which is driven by the sun, is tapped most efficiently with hydropower.

kilowatt (kW) Unit of electric power equal to 1000 W or approximately 1.34 horsepower; for example, it is the amount of electric energy required to light 10 100-W light bulbs.

kilowatt-hour (kWh) The unit of electrical energy commonly used in marketing electric power; it is the energy produced by 1 kW acting for 1 h.

kinetic energy Energy that a moving body has due to its motion, dependent on its mass and the rate at which it is moving.

megawatt (MW) A unit of power equal to 1 million W; for example, it is the amount of electric energy required to light 10,000 100-W bulbs.

megawatt-hour (MWh) Unit of electric power equal to 1000 kWh.

microhydropower A hydroelectric plant with a rated capacity of 100 kW or less.

peaking capacity That part of a system's capacity that is operated during the hours of highest power demand.

penstock A closed conduit or pipe for conducting water to a powerhouse.

plant factor The ratio of the average output to the installed capacity of the plant, expressed as an annual percentage.

power (electric) The rate of generation or use of electric energy, usually measured in kilowatts.

pumped-storage hydroelectric plant A plant that usually generates electric energy during peak-load periods by using water previously pumped into an elevated storage reservoir during off-peak periods when excess generating capacity is available to do so; when additional generating capacity is needed, the water can be released from the reservoir through a conduit to turbine generators located in a power plant at a lower level.

reservoir An artificial lake into which water flows and is stored for future use.

turbine A machine for generating rotary mechanical power from the energy of a stream of fluid (e.g., water, steam, hot gas); turbines convert the kinetic energy of fluids to mechanical energy through the principles of impulse and reaction or a mixture of the two.

watt (W) The unit used to measure production/usage rate of all types of energy; the unit for power.

watt-hour (Wh) The unit of energy equal to the work done by 1 W in 1 h.

This article provides a brief overview of the hydropower resource—how this resource is assessed, current resources worldwide, and future potential development.

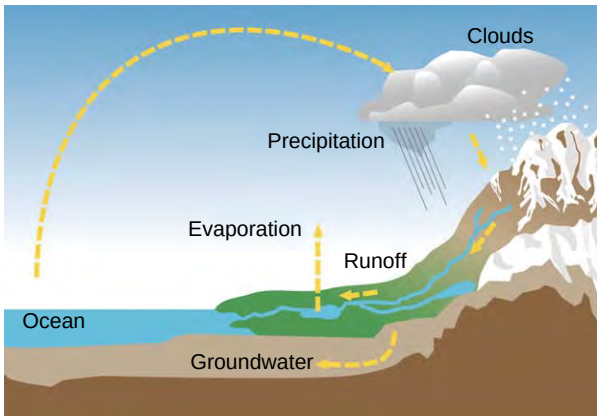


FIGURE 1 The hydrologic cycle.

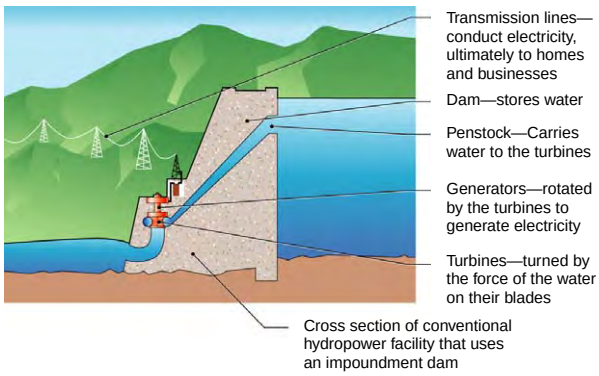


FIGURE 2 Impoundment hydropower project.

1. THE RESOURCE

Hydroelectric power comes from water at work. To generate electricity, water must be in motion. This kinetic energy turns the blades of a water turbine, which changes the kinetic energy to mechanical (machine) energy. The turbine shaft turns a generator, which then converts this mechanical energy into electricity. Because water is the initial source of electrical energy, this technology is called hydroelectric power or "hydropower" for short.

Water constantly moves through the hydrologic cycle (Fig. 1), which is driven by solar energy. In the hydrologic cycle, atmospheric water reaches the earth's surface as precipitation. Some of this water evaporates, but much of it either percolates into the soil or becomes surface runoff. Water from rain and melting snow eventually reaches ponds, lakes, reservoirs, or oceans where evaporation is constantly occurring. The hydrologic cycle ensures that water is a renewable resource.

Hydropower plants can be located on rivers, streams, and canals, but dams are needed for a reliable water supply. Dams store water for later release for purposes such as irrigation, domestic and industrial use, and power generation. The reservoir acts much like a battery, storing water to be released as needed to generate power. The dam also creates a head, that is, a height from which water flows. A pipe (penstock) carries the water from the reservoir to the turbine.

Typical conventional hydropower projects are of three types: impoundment, diversion, and run-of-river. Impoundment projects (Fig. 2) use a dam to store water. Water is released to meet water use demands, including generation of electricity. Diversion projects (Fig. 3) channel a portion of the river through a canal or penstock to produce electrical energy. A small diversion dam may be required to channel a portion of the water from the river. Run-of-river projects (Fig. 4) use the flow of water within the natural range of the river. They may require a small impoundment to develop head. Run-of-river projects generally blend in with the natural river system.

FIGURE 3 Diversion project.

FIGURE 4 Run-of-river project.

Unconventional hydropower projects include pumped storage. These projects pump water from a lower reservoir to an upper reservoir during times when demand for electricity is low. During periods of high electrical demand, the water is released back to the lower reservoir to generate electricity.

2. HOW HYDROPOWER RESOURCES ARE ASSESSED

The first step in assessing a potential water resource for hydropower is to calculate how much power can be produced. The estimated power is determined by the volume of water available (flow) and the vertical distance the water falls (head). A given amount of water falling a given distance will produce a certain amount of energy. In the English system of measurement, flow is commonly measured in cubic feet per second (cfs). Head is measured in feet.

The capacity of hydropower plants is measured in kilowatts or megawatts. Potential capacity can be calculated as follows:

$$kW = Q \times H/11.8,$$

where

- kW = electrical power;
- Q = flow rate in cubic feet per second;
- H = head in feet; and
- 11.8 = a constant (conversion of foot pounds to horsepower and then horsepower to kilowatts).

Note: the equation does not include efficiency losses in the turbine, generator, penstock, and so on.

Power potential can be converted to potential annual energy production, which is expressed in kilowatt-hours as follows:

$$kWh = kW \times 8760 \times PF,$$

where

- kWh = annual energy;
- kW = project capacity;
- 8760 = 24 h per day × 365 days per year; and
- PF = plant factor.

Annual energy production is based on installed capacity and available water for power production. In most cases, water flow varies throughout the year. During low-flow periods, energy production is reduced. Therefore, when calculating annual energy production, a plant factor must be applied. If the actual plant factor cannot be determined based on flow conditions, an average can be used to estimate energy production. The average plant factor for hydropower projects in the United States is approximately 50% (0.5).

The potential installed capacity for a specific project is used to evaluate development potential. In general, a river with a large flow (Q) or potential drop in elevation (H) will be desirable for producing hydropower. However, specific project evaluations include issues such as project economics and accommodating the necessary legal, institutional, and environmental constraints. The sum of the installed capacity of all the projects in a region or country represents that region's or country's potential resource.

Specific issues can be applied to each project, further reducing its potential resource. Some projects cannot be developed because they are located in excluded areas. Excluded areas include wild and scenic rivers, national wilderness areas, and national

parks, monuments, preserves, refuges, and historic areas. Other projects may not be developed, or developed to full potential, due to environmental, legal, and/or institutional issues. Environmental concerns may include releasing water for nonpower production to improve water quality or fish passage. Other issues could be cultural, historical, geological, recreational, scenic, loss of productive land, and displacement of people.

Undeveloped hydropower resources are generally reported by potential capacity. This capacity can be evaluated based on the various conditions and constraints that require specific site information and characteristics.

3. GLOBAL DISTRIBUTION

In 1998, hydroelectricity provided approximately 21.6% of the worldwide electricity capacity and 18.8% of the worldwide generation of electricity. Figure 5 illustrates the top hydroelectric generating countries in the world. Table I shows the installed hydropower capacity and energy production by major geographical regions worldwide.

In 1999, the developed hydropower capacity in the United States was approximately 79,700 MW. In 2000, the electrical energy production was approximately 269,000 GWh/year. This is approximately 7.1% of the total electricity in the United States.

4. CURRENT RESOURCES

Water is one of the most valuable resources, and hydropower makes further use of this renewable treasure. The United States has invested more than $150 billion (in 1993 dollars) in hydropower facilities. This investment does not include the cost to build most of the large federal dams in the western part of the country, but it does include the hydropower portion of these projects. The U.S. hydropower capacity is more than 94,000 MW (includes pumped storage), enough to supply the electrical power needs of approximately 28 million households.

Hydroelectric facilities have many characteristics that favor their use and encourage upgrading existing plants and developing new projects. Hydroelectric facilities offer several benefits, including the following:

- use a renewable resource to generate power;
- are highly reliable and have low operating costs; and
- can start up quickly and have the capability to adjust their output (load following) capability and peak capacity.

As an added benefit, reservoirs have scenic and recreational value for campers, fishermen, and water sports enthusiasts. Water is the home for fish and wildlife as well. Dams add to domestic water supplies, control water quality, provide irrigation for agriculture, and avert flooding. Dams can actually improve the quality of downstream conditions by allowing mud and other debris to settle out.

Just as there is a wide variety of natural settings for hydropower sites, there is a variety of developmental schemes that have various benefits and impacts. The most common distinction is between large and small dams. Dams are often referred to as high-head and low-head dams, reflecting the

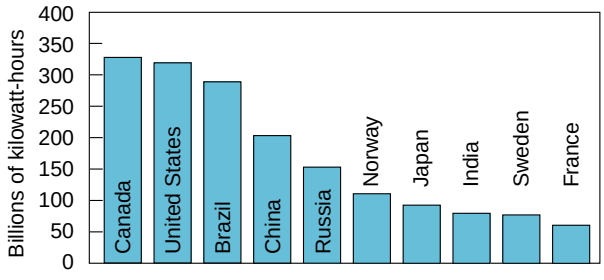


FIGURE 5 Top hydroelectric-generating countries in 1998. From the Energy Information Administration. (2000). "Annual Energy Review 1999," Table 11.15. EIA, Washington, DC.

TABLE I

Worldwide Hydropower Capacity and Energy Production

Geographic region	Installed capacity (~MW)	Energy production (~GWh/year)
Africa	20,651	80,575
Asia (including Russia and Turkey)	241,624	793,045
Australia/Oceania	13,271	42,000
Europe	175,625	593,377
North and Central America	158,000	700,000
South America	111,459	531,168
World totals	720,630	2,740,165

Source. Aqua~Media International Westmead House. (2002). "The International Journal on Hydropower and Dams: World Atlas and Industry Guide" (pp. 13–15). Aqua~Media International Westmead House, Sutton, UK.

concept that hydraulic head (the height from which water drops before reaching the turbine) is a key aspect of electrical generation. As a general rule of thumb, the electrical generating capacity of large hydropower projects ranges from approximately 25 to more than 10,000 MW. Small hydropower projects range from 1 to 25 MW. Smaller yet are the minihydropower projects (less than 1 MW) and microhydropower projects (less than 100 kW). Mini- and microhydropower projects are often located on very small streams to provide decentralized electrical power in remote locations. Many of the most controversial aspects of hydroelectric development, such as the social impacts of human resettlement or inundation of terrestrial habitat, arise from the creation of large reservoirs associated with large dams. However, large dams are far less numerous than small dams. For example, the International Commission on Large Dams estimates that there are approximately 40,000 large dams (dams with a height of 15 m or more or shorter dams with high discharges) and 800,000 small dams worldwide.

Existing power plants can be upgraded, or new power plants can be added at current dam sites, without significant additional effect on the environment. Only approximately 2400 of the nation's 80,000 dams are currently used to generate power (Fig. 6). In addition, new facilities can be constructed with consideration of the environment. For instance, dams can be built at remote locations; power plants can be placed underground; and selective withdrawal systems can be used to control the water temperature released from dams. Facilities can incorporate features that aid fish and wildlife such as salmon runs and resting places for migratory birds. For a history of hydroelectricity's progression in the United States, see Table II.

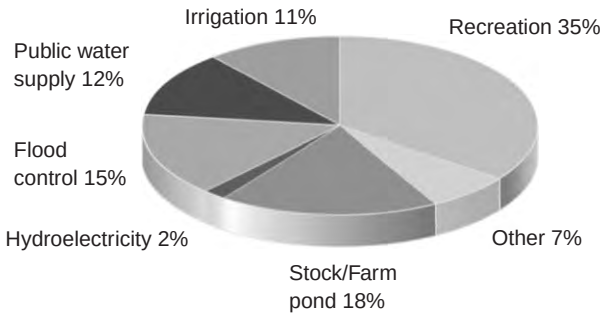


FIGURE 6 Primary benefits of U.S. dams. From the U.S. Army Corps of Engineers, National Inventory of Dams.

5. POTENTIAL OF HYDROPOWER RESOURCES

The potential of hydropower resources is generally evaluated in two steps. The first looks at the resources that are technically feasible. The site physical characteristics are evaluated and generally include site access, water flow data, potential development schemes, potential head, potential installed capacity and energy production, and special site issues. Technically feasible projects require a water resource that can be developed using known engineering and construction techniques. The second step looks at other conditions that apply to developing the resource, including economic factors, environmental issues, legal issues, and institutional concerns. In most cases, economic screening is the second step in the evaluation. The economic factors include investment costs (total cost of the project), interest rate (cost of borrowing money), debt service life (duration to

TABLE II

History of Hydroelectricity's Progression in the United States

- July 1880: Michigan's Grand Rapids Electric Light and Power Company generates electricity by a dynamo belted to a water turbine at the Wolverine Chair Factory; it lights 16 brush-arc lamps.
- 1881: City of Niagara Falls street lamps powered by hydropower.
- 1886: Approximately 45 water-powered electric plants exist in the United States and Canada.
- 1887: San Bernardino, California, is first hydroelectric plant in the western states.
- 1889: Approximately 200 electric plants in the United States use waterpower for some or all generation of electricity.
- 1901: First federal Water Power Act is passed.
- 1907: Approximately 15% of electric generating capacity in the United States is provided by hydropower.
- 1920: Approximately 25% of U.S. electrical generation is hydropower.
- 1920: Federal Power Act establishes Federal Power Commission authority to issue licenses for hydro development on public lands.
- 1935: Federal Power Commission's authority is extended to all hydroelectric projects built by utilities engaged in interstate commerce.
- 1938: Bonneville Dam is first federal dam on the Columbia River.
- 1940: Approximately 40% of electrical generation is hydropower.
- Conventional capacity in the United States triples between 1921 and 1940 and nearly triples again between 1940 and 1980.
- Currently, approximately 7% of U.S. electricity comes from hydropower; approximately 80,000 MW is conventional capacity and 14,000 MW is pumped storage.

TABLE III

Worldwide Hydropower Potential

Geographical region	Technically feasible potential (~GWh/year)	Economically feasible potential (~GWh/year)
Africa	1,750,000	1,100,000
Countries with major resources:		
Angola	90,000	65,000
Cameroon	115,000	103,000
Congo, Democratic Republic of	774,000	419,000
Ethiopia		260,000
Madagascar	180,000	49,000
Asia (including Russia and Turkey)	6,800,000	3,600,000
Countries with major resources:		
China, People's Republic of	2,200,000	1,270,000
India	660,000	
Japan	135,000	114,000
Russian Federation	1,670,000	852,000
Turkey	215,000	123,000
Vietnam	100,000	80,000
Australia/Oceania	270,000	107,000
Countries with major resources:		
Australia		30,000
New Zealand	77,000	40,000
Europe	1,035,000	791,000
Countries with major resources:		
Austria	54,000	50,000
France	72,000	70,000
Iceland	64,000	44,000
Italy	69,000	54,000
Norway	200,000	187,000
Spain	70,000	41,000
Sweden	200,000	130,000
North and Central America	1,663,000	1,000,000
Countries with major resources:		
Canada	981,000	536,000
Mexico	49,000	32,000
United States	528,000	376,000
South America	2,700,000	1,600,000
Countries with major resources:		
Argentina	130,000	
Bolivia	126,000	50,000
Brazil	1,300,000	764,000
Chile	162,000	
Colombia	200,000	140,000
Ecuador	133,000	106,000
Venezuela	261,000	100,000
World Totals	14,218,000	8,198,000

Source. Aqua~Media International Westmead House. (2002). "The International Journal on Hydropower and Dams: World Atlas and Industry Guide" (pp. 13–15). Aqua~Media International Westemead House, Sutton, UK.

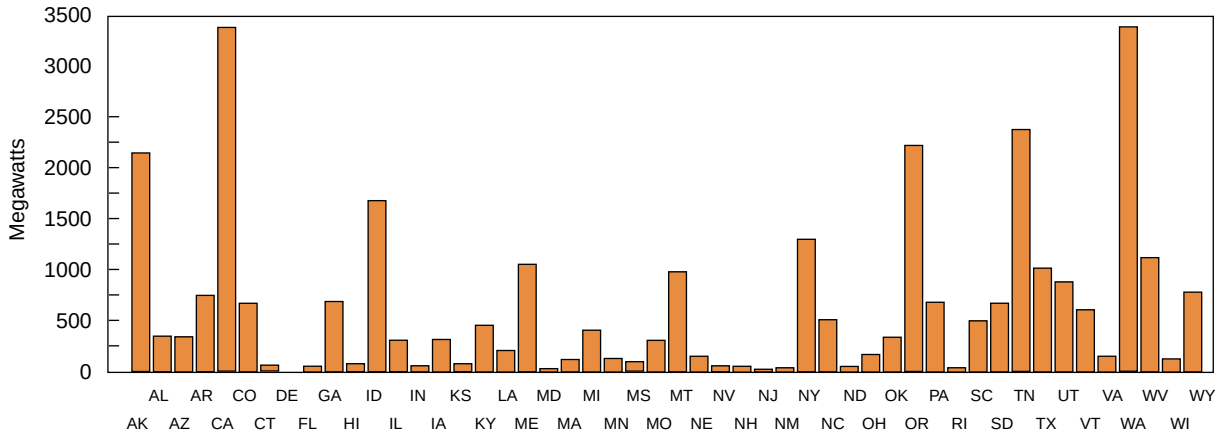


FIGURE 7 Undeveloped hydropower potential by state. From the Hydroelectric Power Resources Assessment database (FERC) and Hydropower Evaluation Software (INEEL). DOE has modeled the undeveloped conventional hydropower potential in the United States. This does not include developed capacity. Various state agencies have reviewed the modeled results and provided input. The 50-state undeveloped conventional hydropower potential is approximately 30,000 MW. The model includes environmental, legal, and institutional constraints to development.

repay the investment with interest), energy value (value or income from the energy produced), operations and maintenance cost (annual cost to operate and maintain the plant), and escalation rates (future value of energy and future operation/maintenance costs). These engineering economic factors can be used to determine the cash flow and benefit/cost ratio. If the resource appears feasible after this step, additional screening for other conditions can be applied to further evaluate the resource. Conditions are unique to each project, and evaluation requires site-specific information. The additional conditions may also affect economics, so a new economic evaluation might be appropriate after the other conditions have been determined.

In some cases, the other screening issues can be applied after the technically feasible resources have been identified. For example, if the environmental issues prevent a resource from being developed, no further evaluation is needed.

Table III shows the additional hydropower potential worldwide by major geographical area, based on technically and economically feasible resources.

The Federal Energy Regulatory Commission (FERC) has provided estimates of the undeveloped hydropower resources in the United States in its Hydropower Resource Assessment (HPRA) database. Other agencies, such as the Army Corps of Engineers, the Bureau of Reclamation, and the Power Marketing Administration (Alaska Power Administration, Bonneville Power Administration, West Area Power Administration, Southwestern Power Administration, and Southeastern Power Administration), have conducted regional HPRA studies.

In June 1989, the U.S. Department of Energy (DOE) initiated development of a National Energy Strategy to identify the energy resources available to support the expanding demand for energy in the United States. Public hearings conducted as part of the strategy development process indicate that the undeveloped hydropower resources were not well defined. One of the issues was that some assessments did not estimate the undeveloped hydropower capacity based on site characteristics, stream flow data, and available hydraulic heads.

Therefore, the DOE developed uniform criteria to perform HPRA studies. They used the FERC HPRA database as the basis for reevaluations. Information from other databases was factored into the assessments. The information was also reviewed by various state agencies. After the review process, the specific site information was computer modeled to further screen development potential base on environmental, legal, and institutional constraints.

The DOE HPRA studies involved 5677 sites and initially determined a capacity of approximately 70,000 MW. This computer modeling of these undeveloped hydropower resources based on environmental, legal, and institutional constraints resulted in an estimated total undeveloped capacity of approximately 30,000 MW. Figure 7 illustrates the undeveloped hydropower potential by state.

The DOE continues to conduct assessments of hydropower resources in the United States.

SEE ALSO THE FOLLOWING ARTICLES

Biomass Resource Assessment • Hydropower Economics • Hydropower, Environmental Impact of • Hydropower, History and Technology of • Hydropower Resettlement Projects, Socioeconomic Impacts of • Hydropower Technology • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Wind Resource Base

Further Reading

Aqua ~ Media International Westmead House. (2002). "The International Journal on Hydropower and Dams: 2002 World Atlas and Industry Guide." Aqua ~ Media International Westmead House, Sutton, UK.

Conner, A.M., et al., (1998). "U.S. Hydropower Resource Assessment Final Report." (DOE/ID–10430.2). U.S. Department of Energy, Idaho Operations Office, Idaho Falls, ID.

Gulliver, J. S., and Arndt, R. E. A. (eds.). (1991). "Hydropower Engineering Handbook." McGraw–Hill, New York.

Hall, D. G., *et al.* (2002). "Low Head/Low Power Hydropower Resource Assessment of the Arkansas White Red Hydrologic Region." U.S. Department of Energy, Idaho Operations Office, Idaho Falls, ID.

Hall, D. G., *et al.* (2002). "Low Head/Low Power Hydropower Resource Assessment of the Pacific Northwest Hydrologic Region." U.S. Department of Energy, Idaho Operations Office, Idaho Falls, ID.

Hydrologic Engineering Center and Institute for Water Resources. (1979). "Feasibility Studies for Small Scale Hydropower Additions: A Guide Manual." U.S. Army Corps of Engineers, Hydrologic Engineering Center, Davis, CA.

U.S. Bureau of Reclamation, Power Resources Office. (2001). Hydroelectric power. www.usbr.gov/power/index.html.

U.S. Department of Energy. (1997). "Hydropower: America's Leading Renewable Energy Resource" [brochure]. U.S. Department of Energy, Idaho Operations Office, Idaho Falls, ID.

U.S. Department of Energy. (2001). "Hydropower: Partnership with the Environment" [brochure]. U.S. Department of Energy, Idaho Operations Office, Idaho Falls, ID.

Hydropower Technology

PEGGY BROOKSHIER
U.S. Department of Energy
Idaho Falls, Idaho, United States

1. **Brief History**
2. **Hydrologic Cycle**
3. **Hydropower Plant Components and Functions**
4. **General Description of the Types of Hydraulic Turbines**
5. **Past Research and Development Efforts**
6. **Current R&D**
7. **Future of R&D**

Glossary

alternating current (**AC**) Electric current that reverses direction many times per second.
ancillary services Operations provided by hydroelectric plants that ensure stable electricity delivery and optimize transmission system efficiency.
cavitation Noise or vibration causing damage to the turbine blades as a result of bubbles that form in the water as it goes through the turbine, causing a capacity loss, head loss, and efficiency loss; the bubbles collapse when they pass into higher regions of pressure.
direct current (**DC**) Electric current that flows in one direction.
draft tube A water conduit that can be straight or curved, depending on the turbine installation, and that maintains a column of water from the turbine outlet and the downstream water level.
efficiency A percentage obtained by dividing the actual power or energy by the theoretical power or energy; it represents how well the hydropower plant converts the energy of the water into electrical energy.
flow Volume of water, expressed as cubic feet or cubic meters per second, passing a point in a given amount of time.
head Vertical change in elevation, expressed in either feet or meters, between the headwater level and the tailwater level.
headwater The water level above the powerhouse.
low head Head of 66 feet or less (as defined by the U.S. Department of Energy).
penstock A closed conduit or pipe for conducting water to the powerhouse.
runner The rotating part of the turbine that converts the energy of falling water into mechanical energy.
scroll case A spiral-shaped steel intake guiding the flow into the wicket gates located just prior to the turbine.
small hydro Projects that produce 30 MW or less (as defined by the U.S. Department of Energy).
tailrace The channel that carries water away from a dam.
tailwater The water downstream of the powerhouse.
ultra low head Head of 10 feet or less (as defined by the U.S. Department of Energy).
wicket gates Adjustable elements that control the flow of water to the turbine passage.

Hydropower, also known as hydroelectric power, is the use of water to produce power. Harnessing water to perform work has been going on for thousands of years. The Greeks used waterwheels for grinding wheat into flour more than 2000 years ago. Besides grinding flour, the power of the water was used to saw wood and to power textile mills and manufacturing plants. This article looks briefly at how hydropower began, why it is considered renewable, the parts of a hydropower plant, types of turbines and when they are used, and what research and development is occurring.

1. BRIEF HISTORY

The evolution of the modern turbine began during the mid-1700s when a French hydraulic and military engineer, Bernard Forest de Bélidor, wrote *Architecture Hydraulique*. In this four-volume work, he described using a vertical axis versus a horizontal axis machine. During the 1700s and 1800s, water turbine development continued. In 1880, a brush arc light dynamo driven by a water turbine was used to provide theater and storefront lighting in Grand Rapids, Michigan, and in 1881, a brush dynamo connected to a turbine in a flour mill provided street

lighting in Niagara Falls, New York. These two used direct current (DC) technology. Alternating current (AC) is used today. That breakthrough came when the electric generator was coupled to the turbine, resulting in the first hydroelectric plant in the United States—and in the world—located in Appleton, Wisconsin.

2. HYDROLOGIC CYCLE

The hydrologic cycle, shown in Fig. 1, depicts the continuous water cycle from evaporation, to precipitation, and back to the water source. This is why hydropower is considered a renewable energy.

A rough rule of thumb for determining potential power is as follows:

$$P(\text{kW}) = \underset{\text{English}}{[e\text{H(ft)Q(ft}^3\text{/s)}]/11.82}$$

$$= \underset{\text{Metric}}{[e\text{H(m)Q(m}^3\text{/s)}] \times 9.81},$$

where P = generator output (power) in kilowatts (kW), e = overall plant efficiency, H = head (foot [ft] or meter [m] depending on which equation is used), and Q = flow (cubic feet or cubic meters per second depending on which equation is used. This equation assumes a constant water density of 62.4 lb/ft^3 or 9806 N/m^3.

3. HYDROPOWER PLANT COMPONENTS AND FUNCTIONS

There are many reasons why dams are built. Figure 2 shows the primary purpose of dams in the United States. Many dams were built for other purposes, and hydropower was added later.

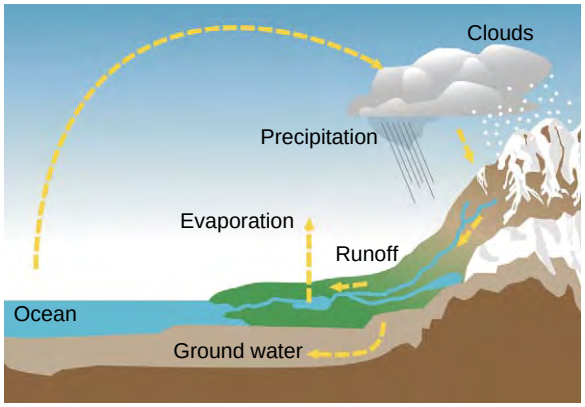


FIGURE 1 Hydrologic cycle.

Worldwide, there are approximately 840,000 dams. However, only a small percentage of them produce power. For example, the United States has approximately 80,000 dams, but only 2400 produce power. The rest are for recreation, stock/farm ponds, flood control, water supply, and/or irrigation.

In general, the purpose of dams is to control water by impounding it, raising its level, or diverting it. Some dams are built to contain sediments that are carried in the water (e.g., mining wastes).

There are three major categories of dams:

- *Embankment:* constructed from the natural materials soil and rock (two types of embankment dams: earthfill and rockfill)
- *Concrete:* constructed from concrete (four types of concrete dams: arch, gravity, combination arch dam with gravity tangents, and buttress)
- *Composite:* generally consisting of concrete gravity or buttress sections in combination with earthfill or rockfill sections

Determining which category and type of dam to build is dependent on many factors, including topography, geological conditions, availability of materials, spillway size and location, and environmental considerations.

There are three types of hydropower facilities:

- *Impoundment.* This type requires a dam to store water. Water may be released either to meet changing electricity needs or to maintain a constant reservoir level.
- *Diversion.* This type channels a portion of the river through a canal or penstock, which may or may not require a dam.
- *Pumped storage.* This type pumps water, when demand for electricity is low, from a lower reservoir to an upper reservoir. During periods of high demand, the water is released back to the lower reservoir to generate electricity.

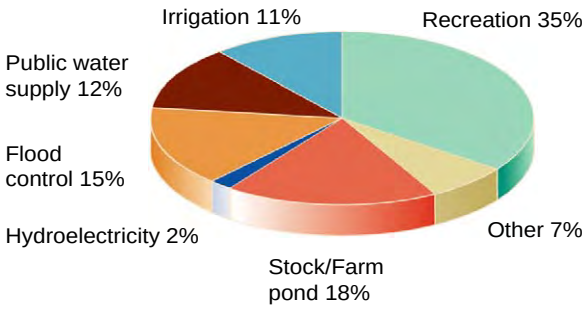


FIGURE 2 Primary purposes or benefits of U.S. dams. From the U.S. Army Corps of Engineers, National Inventory of Dams.

Figure 3 illustrates the components of a typical hydropower facility that requires a dam. Figure 4 illustrates a facility that does not use a dam. Figure 5 depicts a pump storage facility. Figure 6 depicts a typical reversible pump.

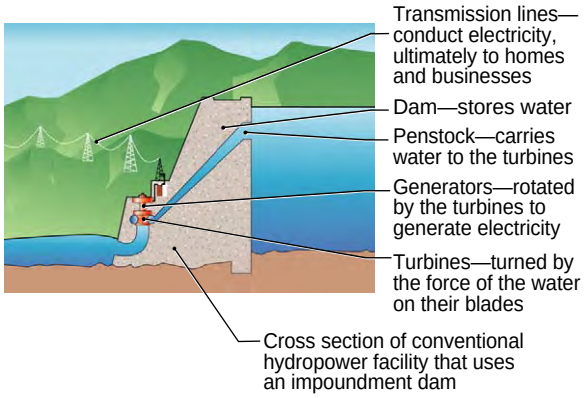


FIGURE 3 Components of a typical hydropower facility.

FIGURE 4 Tazimina Hydropower Project—diversion.

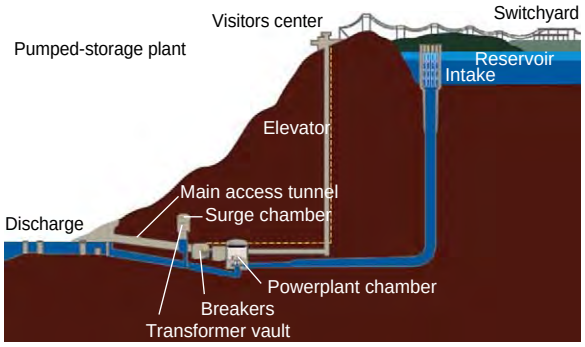


FIGURE 5 Cutaway diagram of a typical pump storage facility. From the Tennessee Valley Authority.

4. GENERAL DESCRIPTION OF THE TYPES OF HYDRAULIC TURBINES

Hydraulic turbines have two main classifications: impulse and reaction. The impulse turbine generally uses the velocity of the water to move the runner and discharges to atmospheric pressure. The water stream hits each bucket on the runner. There is no suction on the down side of the turbine, and the water flows out the bottom of the turbine housing after hitting the runner. An impulse turbine is generally suitable for high-head, low-flow applications.

A reaction turbine develops power from the combined action of pressure and moving water. The runner is placed directly in the water stream flowing over the blades rather than striking each one individually. Reaction turbines are generally used for lower head, higher flow applications in comparison with impulse turbines. However, as shown in Fig. 7, there is a lot of overlap.

4.1 Impulse Turbine

4.1.1 Pelton

A pelton wheel has one or more free jets discharging water into an aerated space and impinging on the buckets of a runner. Draft tubes are not required for an impulse turbine because the runner must be located above the maximum tailwater to permit operation at atmospheric pressure. Figure 8 shows a cutaway of a Pelton.

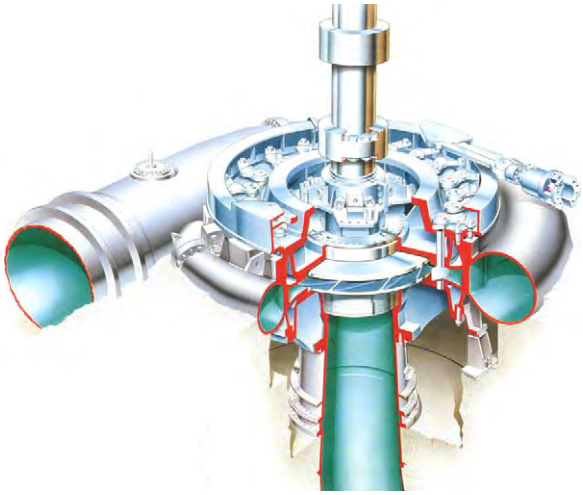


FIGURE 6 Cutaway diagram of a reversible pump used at pump storage facilities. It pumps to an upper reservoir and acts as a turbine when water flows to the lower reservoir. From GE Hydro.

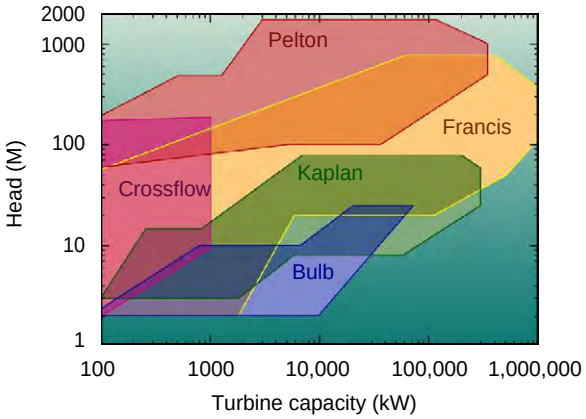


FIGURE 7 Head and flow ranges for various types of turbines.

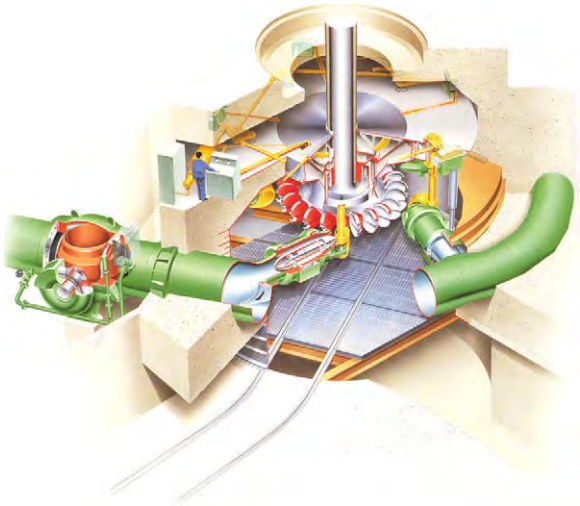


FIGURE 8 Cutaway of a typical Pelton turbine. From GE Hydro.

A turgo wheel is a variation on the Pelton and is made exclusively by Gilkes in England. The Turgo runner is a cast wheel whose shape generally resembles a fan blade that is closed on the outer edges. The water stream is applied on one side, goes across the blades, and exits on the other side.

4.1.2 Cross-Flow

A cross-flow turbine is drum shaped and uses an elongated, rectangular section nozzle directed against curved vanes on a cylindrically shaped runner. It resembles a "squirrel cage" blower. The cross-flow turbine allows the water to flow through the blades twice. The first pass is when the water flows from the outside of the blades to the inside, and the second pass is from the inside back out. A guide vane at the entrance to the turbine directs the flow to a limited portion of the runner. The cross-flow was developed

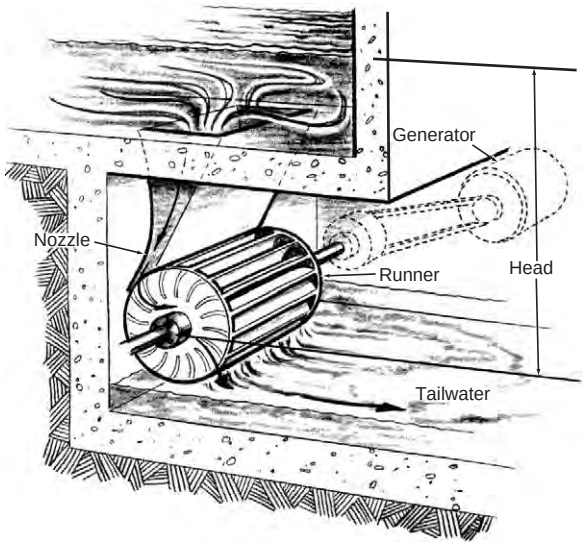


FIGURE 9 Cutaway of a typical cross-flow turbine.

to accommodate larger water flows and lower heads than is the case with the Pelton. Figure 9 shows a cutaway of a cross-flow turbine.

4.2 Reaction Turbine

4.2.1 Propeller

A propeller turbine generally has a runner with three to six blades in which the water contacts all of the blades constantly. One may picture a boat propeller running in a pipe. Through the pipe, the pressure is constant; if it were not constant, the runner would be out of balance. The pitch of the blades may be fixed or adjustable. The major components, besides the runner, are a scroll case, wicket gates, and a draft tube. There are several different types of propeller turbines:

- *Bulb turbine*. The turbine and generator are a sealed unit placed directly in the water stream. Figure 10 shows a cutaway of a bulb turbine.
- *Straflo*. The generator is attached directly to the perimeter of the turbine.
- *Tube turbine*. The penstock bends just before or after the runner, allowing a straight line connection to the generator.
- *Kaplan*. Both the blades and the wicket gates are adjustable, allowing for a wider range of operation. Figure 11 shows a cutaway of a Kaplan turbine.

4.2.2 Francis

A Francis turbine has a runner with a number of fixed buckets (vanes)—usually nine or more. Water is introduced just above the runner and all around it

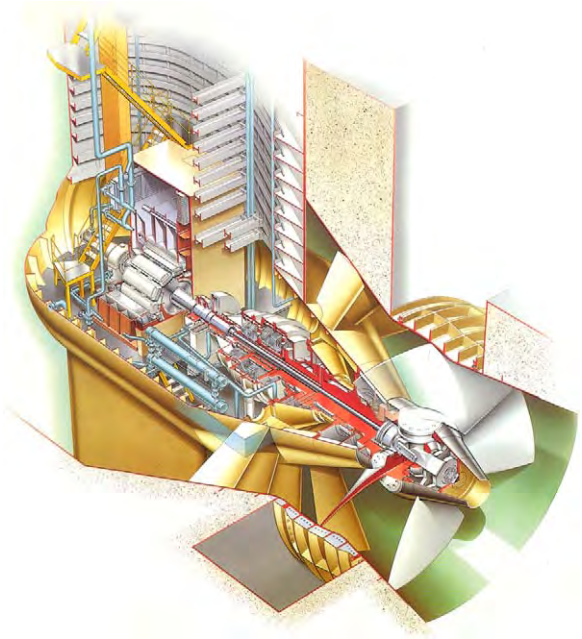


FIGURE 10 Cutaway of a typical bulb turbine. From GE Hydro.

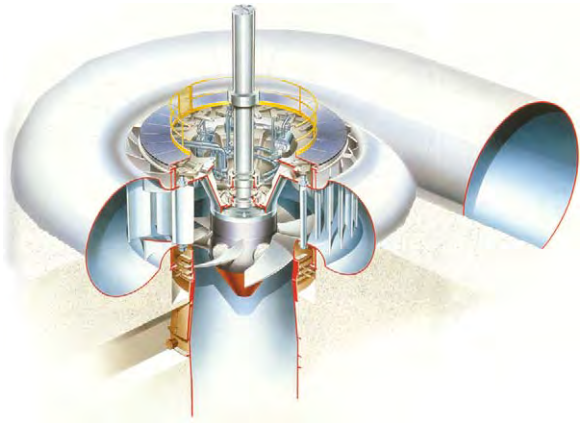


FIGURE 11 Cutaway of a typical Kaplan turbine. From GE Hydro.

and then falls through, causing it to spin. Besides the runner, the other major components are the scroll case, wicket gates, and draft tube. Figure 12 shows a cutaway of a Francis turbine.

4.2.3 Kinetic

Kinetic energy turbines, also called free-flow turbines, generate electricity from the kinetic energy present in flowing water rather than from the potential energy from the head. The systems may operate in rivers, manmade channels, tidal waters, or ocean currents. Kinetic systems use the water

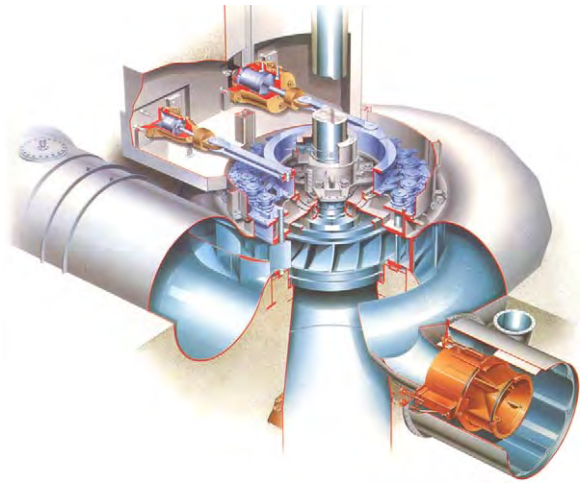


FIGURE 12 Cutaway of a typical Francis turbine. From GE Hydro.

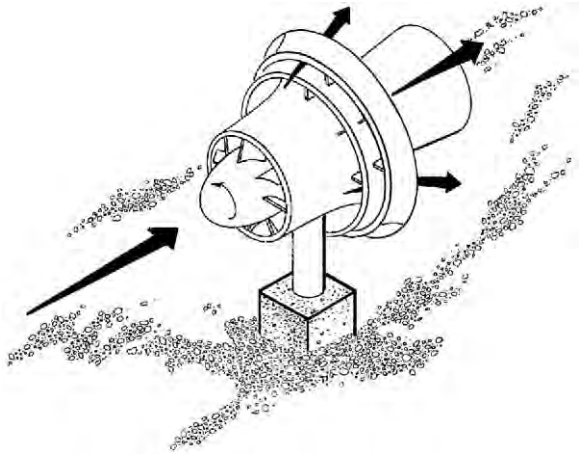


FIGURE 13 Ducted free-flow turbine.

stream's natural pathway. They do not require the diversion of water through manmade channels, riverbeds, or pipes, although they might have applications in such conduits. Kinetic systems do not require large civil works; however, they can use existing structures such as bridges, tailraces, and channels. Figure 13 shows a ducted free-flow turbine.

Turbine selection is first based on head and flow. As can be seen in Fig. 7, many different types of turbines overlap. Other deciding factors include how deep the turbine must be set, efficiency, and cost.

5. PAST RESEARCH AND DEVELOPMENT EFFORTS

Up to around 1980, research and development (R&D) efforts focused mainly on improving turbine

efficiency, reducing cavitations, and increasing generation. Whereas older units had efficiency ratings as low as 60%, the new units have efficiency ratings of approximately 90%. The U.S. Department of Energy (DOE) also looked at new technologies for developing low-head, ultra low-head, and small hydro projects. These technologies included the following:

- *Pumps used as turbines.* This technology uses either actual off-the-shelf pumps or pumps slightly modified by the pump manufacturer as a turbine. Water is now flowing down through the pump rather than up. In some cases, the efficiency is as high as 85%.
- *Float-in powerhouses.* This concept is where the powerhouse is constructed in modules at a shipyard and then floated, acting as a barge, to an existing dam, where it would be sunk into place and the voids would be filled with concrete. This is very similar in concept to modular homes.
- *Marine thruster.* This technology uses thrusters, normally used to maneuver ships while docking, to generate power. Initial tests showed an efficiency of approximately 60%.
- *Free-flow turbines.* This technology uses the flow of the water only to generate power. Therefore, no dam or reservoir is needed. This type of turbine can be anchored to the bottom of a river or hung from bridges. This concept has now been applied to ocean currents.
- *Hydraulic air compressor.* This technology uses entrainment of air bubbles in a penstock full of water, which is allowed to flow from the headwater to a deep chamber, where the then compressed air is collected for use in a gas turbine. The method of using the compressed air determines the overall plant efficiency, which is approximately 85%. The extraction of the work from the compressed air by direct expansion through a gas turbine will produce shaft power that may be used to generate electricity. In addition, the exhaust can be very cold and may be used for cooling.
- *Inexpensive cross-flow turbines.* This concept looked at designing and building an inexpensive cross-flow turbine by using cheaper materials and building in a conventional machine shop.
- *Siphon penstock.* This technology is a penstock that goes over a dam rather than through the dam. Thus, the integrity of the dam is not jeopardized, and the cost of installation is less than what it would be going through the dam. This would be useful at an existing dam with no hydropower. However, it has a lift limitation of 32 feet at sea level.

6. CURRENT R&D

Although hydropower turbine manufacturers have incrementally improved turbine technology to improve efficiencies, the basic design concepts have not changed for decades. These late 19th- and early 20th-century designs did not consider environmental effects because little was known about environmental effects of hydropower at the time.

During the 1980s, the environmental concerns in the United States became more important in hydropower projects, both existing and planned. This trend has been slowly spreading across the globe. In 1996, DOE looked at the hydropower R&D being conducted by federal agencies as well as the private sector. Approximately $63.5 million was spent on hydro R&D, with approximately 84% of that going toward fish passage, behavior, and response-type R&D.

Hydropower has many competitors for its source of power, that is, water. Water is used for drinking, irrigation, recreation, fish and wildlife, navigation, power, cooling, and the like. As shown earlier in Fig. 2, dams also have multiple purposes. With so many competing uses for both the water and the dams, it is understandable why an existing hydropower project, let alone a planned project, would face many barriers and issues.

The hydropower industry recognizes that hydropower plants have an effect on the environment. The industry also recognizes that there is a great need to bring turbine designs into the 21st century. The industry visualizes innovative turbines designed from a new perspective. This perspective would look at the "turbine system" (which could include everything except the dam and powerhouse) intended to balance environmental, technical, and economic considerations.

6.1 Environmental Challenges

Although hydroelectric power plants have many advantages over other energy sources, the potential environmental impacts are also well known. Most of the adverse impacts of dams are caused by habitat alterations. Reservoirs associated with large dams can inundate large amounts of terrestrial and river habitats and displace human populations. Diverting water out of the stream channel (or curtailing reservoir releases so as to store water for future electrical generation) can desiccate riparian (streamside) vegetation. Insufficient stream flow releases degrade habitat for fish and other aquatic organisms in the affected river reach below the dam. Water in

the reservoir is stagnant compared with that in a free-flowing river; consequently, waterborne sediments and nutrients can be trapped, resulting in the undesirable proliferation of algae and aquatic weeds. In some cases, water spilled from high dams may become supersaturated with nitrogen gas and cause gas bubble disease in aquatic organisms inhabiting the tailwaters.

A hydropower project can also affect aquatic organisms directly. The dam can block upstream movements of migratory fish (e.g., salmon, steelhead, American shad, sturgeon, paddlefish, eels). Downstream-moving fish may be drawn into the power plant intake flow and pass through the turbine. These fish are exposed to physical stresses (e.g., pressure changes, shear, turbulence, strike) that may cause disorientation, physiological stress, injury, and/or mortality.

R&D is currently under way to help fishery biologists and turbine designers better understand what is happening in the turbine passage. Biological tests are being conducted that will quantify the physical stresses that cause injury or mortality to fish. In addition to these tests, tools are being developed to help both engineers and biologists. These tools include developing a "sensor fish," which is similar to a crash dummy fish. It will be able to measure the physical stresses in a turbine passage and be used instead of live fish to gather information. Another tool is the development of a computational fluid dynamics program that models potential fish behavior in the turbine passage. The test results and tools will help turbine manufacturers to design a more environmentally friendly turbine that will reduce the physical stresses to which fish are exposed. New products such as greaseless bearings eliminate the possibility of petroleum products being released into the water.

6.2 Turbine

During the mid-1990s, DOE began the Advanced Hydropower Turbine System (AHTS) program as part of the Hydropower Program. The goal of the AHTS program is to develop environmentally friendly turbines. DOE funded the conceptual designs of four turbine types: three by Voith–Siemens and one by the team of Alden Research Laboratory (ARL) and Northern Research and Engineering Corporation (NREC). Voith–Siemens looked at redesigning a Kaplan and Francis turbine and developing a dissolved oxygen-enhancing turbine. ARL redesigned a pump impeller, used in the food processing industry to pump tomatoes and fish, to be used as a turbine. Figure 14 shows what the ARL–NREC design would look like.

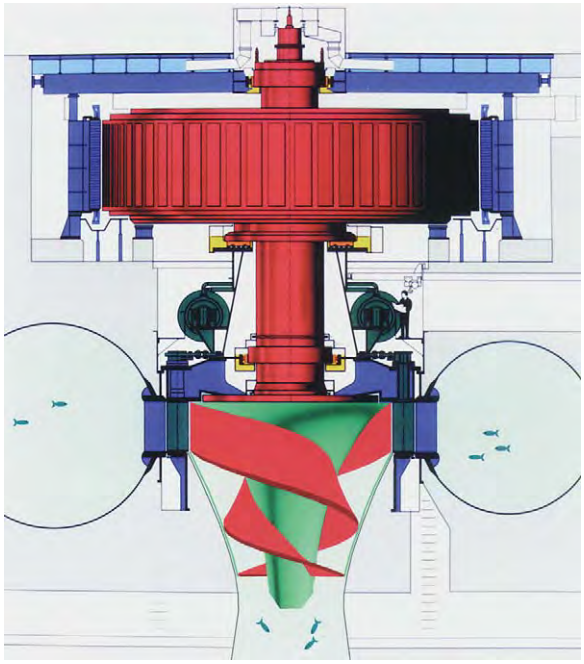


FIGURE 14 Cutaway of an ARL/NREC fishfriendly turbine. From the Alden Research Laboratory.

The ARL–NREC design proceeded into the proof-of-concept stage, where the biological and engineering tests were finalized. Part of the proof of concept was to verify the biological design assumptions and the issue of whether the results of biological testing of a smaller model can be scaled upward to a full-sized turbine.

Many turbine manufacturers have begun designing environmentally friendly turbines, based on a potential market not only in the United States but also worldwide. In 2002, DOE selected three projects where environmentally friendly turbines will be installed and tested. Two projects involve improving downstream fish passage, and one involves increasing dissolved oxygen in discharged water.

6.3 Other

Other federal agencies and utilities are conducting research looking at how hydro plants can be operated to improve fish habitat.

7. FUTURE OF R&D

In 1992, and again in 2001, HCI Publications conducted the Hydro R&D Forum. The purpose

of the forum was to develop an agenda for hydro industry R&D activities worthy of pursuit during the years ahead. The forum brought together representatives from various parts of the hydro industry.

R&D needs were identified and prioritized by the following topic areas:

- Performance and production
- Environmental issues
- Innovative operations and maintenance (O&M)
- Multiple-purpose values
- Human resource issues

The top issues in each area were then prioritized. Following are the top 10 R&D needs identified in 2001:

1. *Quantify and communicate hydro's value compared with the power mix*. Needs include developing objective criteria for measuring and valuing hydropower's nonpower attributes. Criteria should be based on internationally recognized standards, permit comparisons across generation sources, recognize emissions as well as effects on aquatic and land resources, and take into account the entire life cycles of facilities associated with electricity production.

2. *Continue AHTS program development and deployment*. Achieving progress in the AHTS program, which aims to reduce mortality of fish passing through turbines, is a goal. Recommended new areas of emphasis include improving water quality without compromising energy generation efficiency.

3. *Develop business model for O&M decision support based on condition monitoring*. Condition monitoring equipment and systems are increasingly being used. However, the availability of new information poses the conundrum of how to use this information optimally for making both major and minor business decisions in support of facilities operation and maintenance.

4. *Improve methods to quantify costs and benefits from ancillary services*. Hydro plant owners and operators recognize that the ancillary services often provided by their facilities are valuable. However, market structures often fail to appropriately recognize these values. As a consequence, returns to owners and projects are often less than they would be if these values were acknowledged appropriately.

5. *Improve hydro education and outreach*. There is a need to ensure that constituencies are suitably informed about hydro issues. In this area, there are needs to identify effective educational methods and content, research the best ways in which to transfer knowledge, update and maintain factual content and methods, and develop communication networks to inform policymakers, nongovernmental organizations, and the public.

6/7 (tie). *Support green power acceptance for hydro*. In some regions, electricity is marketed as "green" (e.g., from sources that are environmentally beneficial compared with the alternatives). Projects in this area include determining how to gain acceptance for hydro within green marketing programs.

6/7 (tie). *Improve hydropower competency management program*. Workforce issues are increasing in priority, especially in light of loss and projected loss of personnel due to retirements. Projects in this area include focusing on inventorying the skills and competencies that are needed for hydro facilities operation and developing techniques and strategies for addressing identified needs.

8. *Improve hydro-related R&D technology transfer*. Much information is available that is, or may be, applicable to problems and areas of need. However, better tools and processes are needed for communicating and sharing information, both from research activities and from operational experience.

9. *Determine operating life effects due to more severe load operations*. During recent years, many hydro projects have begun to operate in ways that were not envisioned during their original designs. Work is needed to determine how equipment is affected as well as to identify mitigative actions and tools for predicting and preventing failures.

10. *Establish protocols for measuring mitigation effectiveness*. During recent years, hydro project owners have engaged in many instances of implementing environmental mitigation measures. However, it is generally not known whether the measures have been effective. That is, did they accomplish the desired results? Knowledge gained from research in this area could help in designing future mitigative measures and in avoiding ineffective mitigation.

SEE ALSO THE FOLLOWING ARTICLES

Hydropower Economics • Hydropower, Environmental Impact of • Hydropower, History and Technology of • Hydropower Resettlement Projects, Socioeconomic Impacts of • Hydropower Resources

Further Reading

Alward, R., *et al.* (1979). "Micro-Hydro Power: Reviewing an Old Concept." National Center for Appropriate Technologies, Butte, MT.

Armstrong, E. (1977). Selection of the type of dam. *In* "Handbook of Dam Engineering" (A. Golze, Ed.). Litton Educational Publishing, New York.

Chappell, J. (1984a). "Hydropower Hardware Descriptions." EG&G Idaho, Idaho Falls, ID.

Chappell, J. (1984b). "New Technology for Small Hydropower Installations." EG&G Idaho, Idaho Falls, ID.

Gulliver, J. and Arndt, R. (1991) *Hydropower Engineering Handbook*, McGraw Hill, Inc.

Idaho National Engineering Laboratory. (1997). "Hydropower Research and Development." Idaho National Engineering Laboratory, Idaho Falls, ID.

McKinney, J., *et al.* (1983). "Microhydropower Handbook." EG&G Idaho, Idaho Falls, ID.

Indoor Air Quality in Developing Nations

MAJID EZZATI
Harvard School of Public Health
Boston, Massachusetts, United States

1. **Household Fuel Use Patterns in Developing Countries**
2. **Solid Fuel Smoke and Indoor Air Pollution**
3. **Human Exposure to Indoor Smoke from Solid Fuels**
4. **Health Effects of Exposure to Indoor Smoke**
5. **Health Benefits of Household Energy Transitions**
6. **Knowledge Gaps and Research Needs for Reducing Indoor Air Pollution**
7. **Emerging Indoor Air Quality Issues in Developing Countries**

Glossary

acute lower respiratory infections (ALRI) Infections of the lower part of the respiratory tract. Although the division between lower and upper parts of the respiratory tract sometimes takes place at the level of the larynx, many infections affect multiple parts of the respiratory tract, especially where the affected areas are smaller. Further, infections of the bronchi (brochitis) and of the lungs (pneumonia) are often considerably more severe than those in other parts of the respiratory tract and these infections have more specific symptoms. For these reasons, many public health and medical protocols use ALRI to refer to bronchitis, pneumonia, and broncho-pneumonia and combine the infections of other parts of the respiratory tract in the category of acute upper respiratory infections (AURI).

acute respiratory infections (ARI) Infections of the respiratory tract, which consists of the lungs, the bronchi, the trachea, the larynx, the pharynx, the tonsillar glands, the eustachian tube, the nasal cavities, and the sinuses. Otitis media (middle ear infection) is also sometimes included in the definition of ARI.

PM_{10} Particulate matter less than 10 μm in aerodynamic diameter.

Indoor air quality (IAQ) research deals with the presence, levels, health effects, and control of physical, chemical, and biological factors in indoor environments, including homes, workplaces, and vehicles. IAQ research in industrialized countries has examined hundreds of specific factors (e.g., temperature, various chemicals, and mold), sources of pollution (e.g., environmental tobacco smoke, occupational factors, consumer cleaning products, and moisture), and control technologies (e.g., ventilation). In developing countries, however, the great majority of IAQ research, which is considerably less extensive compared to research in industrialized nations, has focused on the role of household energy use and indoor smoke from cooking and heating because this source of indoor air pollution has the largest public health effects.

1. HOUSEHOLD FUEL USE PATTERNS IN DEVELOPING COUNTRIES

Globally, almost 3 billion people rely on biomass (wood, charcoal, crop residues, and dung) and coal as their primary source of domestic energy. Biomass accounts for more than one-half of household energy in many developing countries and for as much as 95% in some lower income regions (Fig. 1). There is also evidence that in some countries the declining trend of household dependence on biomass has slowed, or even reversed, especially among poorer households. Slow economic growth, especially in sub-Saharan Africa, lack of energy infrastructure in remote rural areas for delivery of alternative energy sources, and uncertainty about the price of alternative fuels are among the likely causes of the persistence of biomass fuels.

 © 2004 Elsevier Inc. All rights reserved.

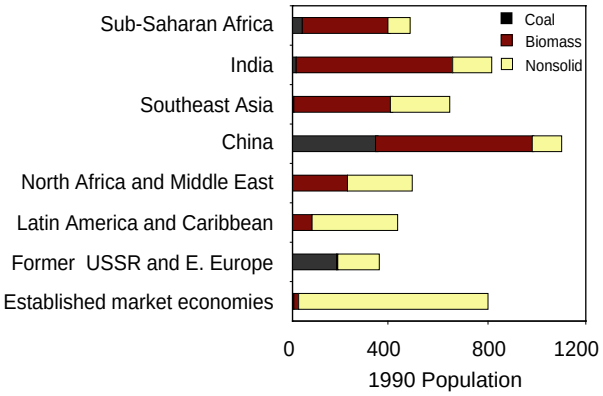


FIGURE 1 Sources of household energy by region. Data from the World Health Organization (1997).

FIGURE 2 The traditional "three-stone" open fire is used in many developing country households. Photograph by M. Ezzati.

2. SOLID FUEL SMOKE AND INDOOR AIR POLLUTION

Smoke from biomass and coal combustion contains a large number of pollutants with known health hazards, including particulate matter, carbon monoxide, nitrogen dioxide, sulfur oxides (mainly from coal), formaldehyde, and polycyclic organic compounds (e.g., carcinogens such as benzo[*a*]pyrene). The concentrations of each of these pollutants vary among the different forms of solid fuels, with animal dung and crop residues having some of the highest level emissions of particulate matter, one of the important indicator pollutants for health effects.

Pollution levels are especially high when burning is done using simple technologies such as open "three-stone" fires, which result in incomplete and inefficient combustion (Fig. 2). Monitoring of pollutants and personal exposures in biomass-burning households has shown concentrations many times higher than

FIGURE 3 Wood smoke rising from the roof of a house in Central Kenya on a typical day. Photograph by M. Ezzati.

those observed in industrialized countries (Fig. 3). The latest National Ambient Air Quality Standards of the U.S. Environmental Protection Agency, for instance, required that the daily average concentration of particles below 10 μm in diameter (PM_{10}) be less than 150 $\mu g/m^3$ (annual average below 50 $\mu g/m^3$). In contrast, typical 24-hour average concentrations of PM_{10} in homes using biofuels may range from 200 to 5000 $\mu g/m^3$ or more throughout the year, depending on the exact type of fuel, stove, and housing. Levels of carbon monoxide and other pollutants also often exceed international guidelines.

3. HUMAN EXPOSURE TO INDOOR SMOKE FROM SOLID FUELS

Exposure to air pollutants is very high in indoor environments in developing countries. Smith has estimated that at the aggregate level (i.e., without accounting for particle size, chemical composition, and source), approximately 80% of total global exposure to airborne particulate matter occurs indoors in developing nations. Details of exposure for various household members, and the roles of both pollution and behavior (e.g., location with respect to stove and activities), have been studied and evaluated using new tools and technology.

Biomass smoke is very episodic, with peaks occurring when fuel is added or moved, the stove is lit, cooking pots are placed on or removed from the fire, or food is stirred. These are also the times that those household members who cook are closest to the fire (Fig. 4). When these patterns of exposure, which incorporate both pollution and behavior, are accounted for, it can be seen that those household members who take part in cooking or are near the

FIGURE 4 Household members involved in cooking are exposed to episodes of high levels of pollution when they work directly above the fire. Photograph by M. Edwards. From Ezzati and Kammen (2002), *Environmental Health Perspectives.*

cooking area (mostly women and children) are exposed to pollution levels that are tens of times higher than those in even the most polluted cities in the world (Fig. 5). Further, a substantial portion of this exposure occurs during short peak periods when emissions are highest and the cook is closest to the fire (referred to as high-intensity exposure in Fig. 5).

4. HEALTH EFFECTS OF EXPOSURE TO INDOOR SMOKE

Exposure to indoor air pollution from the combustion of solid fuels has been implicated, with varying degrees of evidence, as a causal agent of several diseases in developing countries. In a review of the epidemiological evidence for the health effects of indoor smoke from solid fuels, Bruce *et al.* concluded

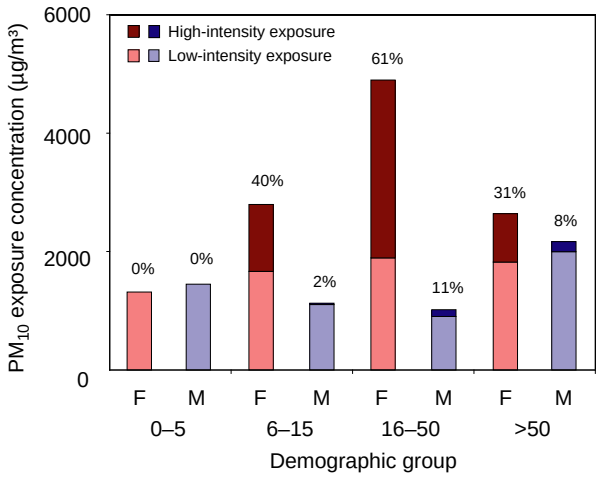


FIGURE 5 Exposure to indoor air pollution from biomass fuels in rural households of central Kenya. For each demographic subgroup, the total height of the column is the group average exposure concentration divided into the average for high-intensity (darker shade) and low-intensity (lighter shade) components. The percentages indicate the share of total exposure from high-intensity exposure. The high-intensity component of exposure occurs in less than 1 hour, emphasizing the intensity of exposure in these episodes. From Ezzati, Saleh, and Kammen (2000), *Environmental Health Perspectives.*

that, despite some methodological limitations, epidemiological studies provide compelling evidence of causality for acute respiratory infections (ARI) and chronic obstructive pulmonary disease (COPD). This is reinforced by experimental data, studies of pathogenesis, and indirect evidence from environmental tobacco smoke (ETS) and ambient air pollution studies. The relationship between coal smoke (but not biomass smoke) and lung cancer has also been consistently established in a number of epidemiological studies. A relationship between biomass smoke and lung cancer may also exist, because biomass smoke contains carcinogens, although at lower concentrations compared to coal smoke. Due to smaller risks and greater difficulties in exposure quantification, epidemiological studies on biomass smoke and lung cancer have so far been inconclusive. For other diseases, including asthma, upper aerodigestive tract cancer, interstitial lung disease, low birth weight and perinatal conditions, tuberculosis, and eye diseases, Bruce *et al.* classified the evidence as tentative.

Conservative estimates of global mortality due to indoor air pollution from solid fuels show that, in 2000, more than 1.6 million deaths were attributed to this risk factor. This accounts for approximately 4–5% of total mortality worldwide. Approximately 1 million of the deaths were due to childhood acute

TABLE I

Global Mortality Attributable to Indoor Smoke from Solid Fuels in 2000[a]

Country grouping	Number of deaths in children under 5 years	Number of deaths in adults[b]
High-mortality developing countries	808,000	232,000 (74%)
Lower mortality developing countries	89,000	468,000 (76%)
Demographically and economically developed countries	13,000	9,000 (79%)

[a] Data from World Health Organization (2002).
[b] Numbers in brackets show the fraction of adult deaths that are among women.

respiratory infections, with the remainder due to other causes, dominated by COPD and then lung cancer among adult women (Table I). The estimated health effects, currently limited to those diseases with strong epidemiological evidence, would of course be even larger if causality and quantitative relationships were established for other diseases. At the same time, even the diseases quantified so far show that the public health effects of indoor smoke in developing countries are enormous. As a result, coupled with the attention to details of exposure, health research has shifted to the "exposure–response" relationship for indoor air pollution and disease. Unlike early studies that divided the population into exposed and nonexposed, the exposure–response relationship, which quantifies the health effects along a continuum of exposure levels (or a larger number of categories), allows considering the benefits of partial exposure reduction for designing new interventions (Fig. 6).

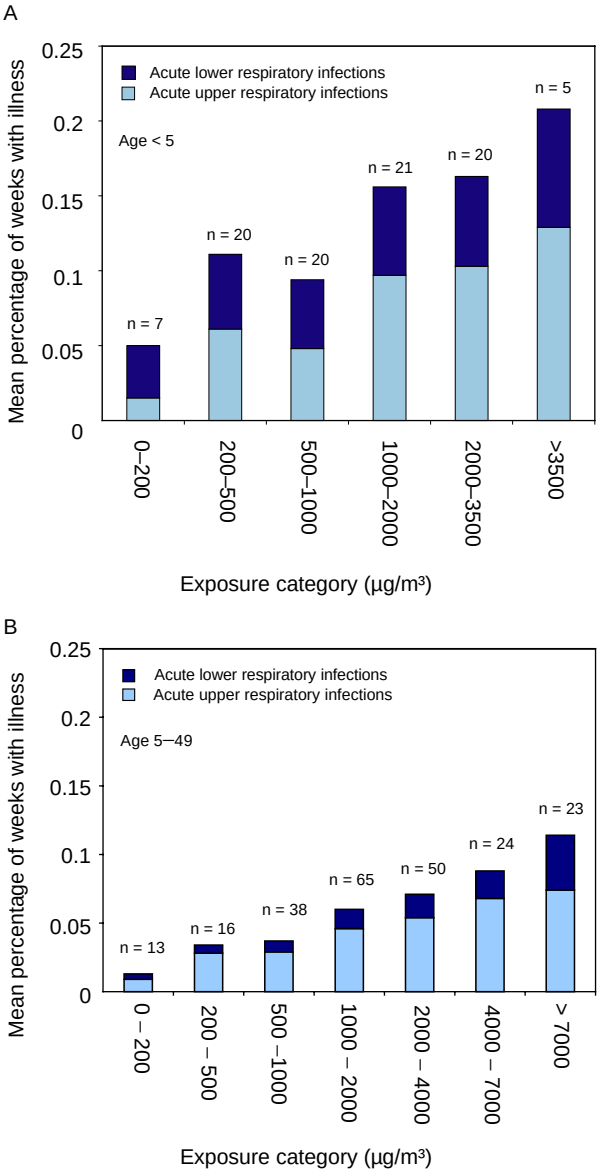


FIGURE 6 The relationship between exposure to indoor particulate matter and acute lower (more severe) and upper respiratory infections, for (A) children under the age of 5 years and for (B) young and middle-aged adults from a study in Central Kenya. As seen, disease increases with exposure level but the relationship is not linear; *n* is the number of people in each exposure group. Reproduced from Ezzati and Kammen (2001), with permission.

5. HEALTH BENEFITS OF HOUSEHOLD ENERGY TRANSITIONS

Reducing exposure to indoor air pollution from household energy use can be achieved through interventions in of the following areas:

- Emissions source and energy technology (fuel–stove combination).
- Housing design and ventilation.
- Behavior and time–activity budget.

Interventions involving emissions sources and fuel and stove combinations (Fig. 7) include using cleaner fuels (e.g., nonsolid fuels such as kerosene, natural gas, and electricity, or even charcoal, which has lower emissions compared to wood), stoves with better combustion conditions and lower emissions, and enclosed stoves with a chimney that removes

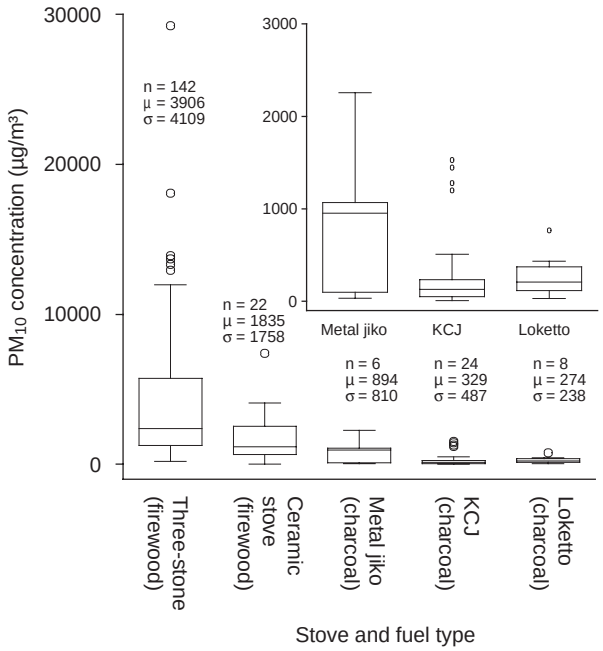


FIGURE 7 Daylong average of PM_{10} concentration for various stove–fuel combinations in Central Kenya (calculated over burning period). The insert in the upper right-hand corner is a more detailed version of the plot for the last three stoves; *n* is the number of days of measurement, µ is the sample mean, and σ is the standard deviation. The box-plots used here show a summary of the distribution of the variable. The lower and upper sides of each rectangle show the 25th and 75th percentiles and therefore enclose the middle half of the distribution. The middle line, which divides each rectangle into two, is the median. The large variability for different stove–fuel combinations shows that how a stove is used is as important in determining pollution as the technology. KCJ, Kenya ceramic jiko. Reproduced from Ezzati and Kammen (2002), *Environmental Health Perspectives*.

much of the emissions from indoor environments. Choices in housing and ventilation include houses with a separate cooking area or more windows. Modifications in behavior and in time–activity budgets include drying wood before use, having shorter burning periods, improving fire-handling practices to reduce the number and frequency of very smoky smoldering events, and keeping children away from the cooking area. The different intervention options are likely to have differential benefits for various household members. For example, cleaner fuels with lower emissions would benefit all household members regardless of age and gender. On the other hand, it has been estimated that relocating the stove to an outside cooking location without reducing its emissions provides larger relative benefits for male household members and children who do not cook.

Most current research in developing countries focuses on household energy technology (fuel–stove combinations) as the primary means for exposure reduction. Although it is fairly well established that transitions to cleaner fuels can provide large reductions in most pollutant levels, recent emphasis has been on "improved" (high-efficiency and/or low-emission) stoves and fuels, which may provide more affordable options in the near future compared to a complete shift to nonsolid fuels.

Attempts to reduce indoor air pollution through the design and introduction of new household energy technology should acknowledge the inherent interaction of technology and socioeconomic factors and include this interface in the research process. Initial research and development efforts to improve stove designs were often marked by a lack of detailed data on stove performances. Efficiencies and emissions, for example, were often measured in controlled environments, with technical experts using the stoves under conditions very dissimilar to those in the field. More recently, the attention of the research community has shifted from such ideal operating conditions to monitoring stove performance under actual conditions of use, taking into account the various social and physical factors that would limit the use of these stoves altogether or would result in "suboptimal" performance. These include factors such as food preferences or the ability to use the technology, the cost of the stove and fuel, the durability and ease of repair or replacement of the stove, uncertainties in fuel costs and access due to seasonality or political factors. Despite these complexities, a number of efforts have provided evidence that behavioral changes and the introduction of new energy technology can provide health benefits that are comparable in magnitude to other interventions (such as hospital-based case management for childhood ARI).

6. KNOWLEDGE GAPS AND RESEARCH NEEDS FOR REDUCING INDOOR AIR POLLUTION IN DEVELOPING COUNTRIES

Important research over the past few decades has illustrated that indoor smoke from solid fuel poses important health risks in developing countries. Unfortunately, affordable and effective interventions for reducing these risks are limited. This may be because in designing new interventions, the

complexities of household energy use and exposure have been often overlooked, and there is a lack of infrastructure to support technological innovations, marketing and dissemination, and maintenance. Even less is known about combinations of technologies that may be used in a household and the factors that motivate the households to adopt them.

In broad terms, design and implementation of appropriate interventions and policies that will positively impact health effects of exposure to indoor smoke require answers to five research questions:

1. What factors determine the details of exposure and what are the relative contributions of each factor to personal exposure? As described earlier, these factors include the energy technology (stove–fuel combination), the housing characteristics and ventilation (such as the size of the house and the building materials, the number of windows, and arrangement of rooms), and behavioral factors (such as the amount of time spent indoors or near the cooking area). It is important to be aware that these factors often interact, requiring a more integrated approach to research. Attention to details of individual exposure has been a subject of more recent research.
2. What is the quantitative relationship between exposure to indoor air pollution and disease (i.e., the exposure–response relationship) along a continuum of exposure levels?
3. Which determinants of human exposure will be influenced, and to what extent, through any given intervention or group of interventions?
4. What are the impacts of any intervention on human exposure and on health outcomes, and how would these impacts persist or change over time?
5. What are the broader environmental effects of any intervention, its costs, and the social and economic institutions and infrastructure required for its success?

7. EMERGING INDOOR AIR QUALITY ISSUES IN DEVELOPING COUNTRIES

In terms of aggregate health effects, household solid fuel use is currently the most important source of indoor air pollution in developing countries. Although relatively equitable economic and rural development has contributed to reducing the use of solid fuels in some settings (e.g., in some newly industrialized Asian countries), indoor air pollution is likely to remain an important health risk in poorer developing countries in the absence of successful intervention programs. More broadly, indoor air quality issues in developing countries are dynamic phenomena that require dynamic research and policy responses.

One of the emerging issues in indoor air quality relates to the demographic, economic, and ecological changes of the past few decades. Some of the poorest households have been forced to use inferior sources of fuel (e.g., dung or crop residues instead of wood and charcoal). This is partially due to the rising price of fuels and because real income has remained unchanged or has even declined in some developing countries, especially in sub-Saharan Africa. The interactions of poverty, population, environment, and health, however, have complex characteristics that are dependent on local socioeconomic, geographical, and demographic factors. For example, if a stand of mature trees is cleared to open space for additional cultivation or grazing area, or for use in the lumber industry, the fallen trees are often burned *in situ* or processed into charcoal for sale in a distant town. Households that formerly relied on fallen limbs and dead wood from that stand of trees find they must travel farther to meet their fuelwood needs. If no mature stands with a sufficient stock of deadwood remain within a reasonable distance, poor households, unable to afford commercial fuels, may begin to cut smaller trees, which leads to a further loss of tree cover. If smaller trees are not an option, or prove insufficient to meet demand, then some households may turn to agricultural residues or animal manure. This shift has consequences that extend beyond the use of a lower quality and more polluting fuel. Crop residues are often used as fodder, and using traditional fodder as fuel can lower the value of a family's livestock or lower the quality of the animals' manure. When they are not used as fodder, crop residues are often left in the field as ground cover to protect topsoil between growing seasons. They may be ploughed back into the soil or burned on top of it before the next crop is planted; both activities return nutrients to the soil. Using these residues for fuel can leave topsoil unprotected between the harvest of one crop and the sowing of the next, leading to soil erosion and a loss of nutrients. Similarly, using animal manure as fuel takes away a valuable fertilizer, leading to lower yields or forcing the family to rely on expensive inorganic fertilizers. When coupled with increased malnutrition and exposure to other risks associated with poverty in these same households, the health

effects of indoor air pollution would be greatly magnified. Therefore, while much of the literature has considered improvements in indoor air pollution as a result of economic development, the energy and health inequalities as a result of income inequality need particular attention.

Another emerging issue relates to urbanization and other forms of sociodemographic change, to changes in housing and neighborhood design, and to increased use of manufactured chemicals. The implications are that indoor air quality for some developing country households will be increasingly determined by pollution from sources other than their own energy use (nearby households, industries, transportation, insecticides, and chemical cleaners). Outdoor/indoor air quality interactions have been subject to research in industrialized countries. In developing countries, these issues should begin to appear on research and policy agendas, because many involve decisions that cannot be easily reversed.

Finally, smoking has been on the rise in developing countries over the past few decades. Environmental tobacco smoke, which has been subject to regulation in a number of industrialized countries, will increasingly affect the health of the many sectors of society in developing nations (e.g., workers in buildings where workplace smoking is allowed) and should receive early attention.

SEE ALSO THE FOLLOWING ARTICLES

Air Pollution from Energy Production and Use • Air Pollution, Health Effects of • Clean Air Markets • Climate Change and Public Health: Emerging Infectious Diseases • Development and Energy, Overview • Hazardous Waste from Fossil Fuels • Indoor Air Quality in Industrial Nations • Technology Innovation and Energy • Transitions in Energy Use • Women and Energy: Issues in Developing Nations • Wood in Household Energy Use

Further Reading

Agarwal, B. (1983). Diffusion of rural innovations: Some analytical issues and the case of wood-burning stoves. *World Dev.* **11,** 359–376.

Balakrishnan, K., Sankar, S., Parikh, J., Padmavathi, R., Srividya, K., Venugopal, V., Prasad, S., and Pandey, V. L. (2002). Daily average exposures to respirable particulate matter from combustion of biomass fuels in rural households of southern India. *Environ. Health Perspect.* **110,** 1069–1075.

Bradley, P. N., and Campbell, B. M. (1998). Who plugged the gap? Re-examining the woodfuel crisis in Zimbabwe. *Energy Environ.* 9, 235–255.

Bruce, N., Perez-Padilla, R., and Albalak, R. (2000). Indoor air pollution in developing countries: A major environmental and public health challenge. *Bull. World Health Org.* **78,** 1078–1092.

Ezzati, M., and Kammen, D. M. (2001). Indoor air pollution from biomass combustion as a risk factor for acute respiratory infections in Kenya: An exposure-response study. *Lancet* **358,** 619–624.

Ezzati, M., and Kammen, D. M. (2002). Evaluating the health benefits of transitions in household energy technology in Kenya. *Energy Policy* **30,** 815–826.

Ezzati, M., and Kammen, D. M. (2002). The health impacts of exposure to indoor air pollution from solid fuels in developing countries: Knowledge, gaps, and data needs. *Environ. Health Perspect.* **110,** 1057–1068.

Ezzati, M., Mbinda, B. M., and Kammen, D. M. (2000). Comparison of emissions and residential exposure from traditional and improved biofuel stoves in rural Kenya. *Environ. Sci. Technol.* **34,** 578–583.

Ezzati, M., Saleh, H., and Kammen, D. M. (2000). The contributions of emissions and spatial microenvironments to exposure to indoor air pollution from biomass combustion in Kenya. *Environ. Health Perspect.* **108,** 833–839.

Kammen, D. M. (1995). Cookstoves for the developing world. *Sci. Am.* **273,** 63–67.

Leach, G., and Mearns, R. (1988). "Beyond the Woodfuel Crisis: People, Land, and Trees in Africa." Earthscan, London.

Manibog, F. R. (1984). Improved cooking stoves in developing countries: Problems and opportunities. *Annu. Rev. Energy* **9,** 199–227.

Reddy, A. K. N., Williams, R. H., and Johansson, T. B. (eds.). (1996). "Energy after Rio: Prospects and Challenges." United Nations Publications, New York.

Smith, K. R. (1987). "Biofuels, Air Pollution, and Health: A Global Review." Plenum Press, New York.

Smith, K. R. (1988). Air pollution: Assessing total exposure in developing countries. *Environment* **30,** 16–34.

Smith, K. R. (1993). Fuel combustion, air pollution exposure, and health: Situation in developing countries. *Annu. Rev. Energy Environ.* **18,** 529–566.

Smith, K. R., Metha, S., and Feuz, M. (2002). The global burden of disease from indoor air pollution: Results from comparative risk assessment. *In* "Indoor Air 2002: Proceedings of the 9th International Conference on Indoor Air Quality and Climate," pp. 10–19. Indoor Air 2002, Monterey, California.

Smith, K. R., Samet, J. M., Romieu, I., and Bruce, N. (2000). Indoor air pollution in developing countries and acute lower respiratory infections in children. *Thorax* **55,** 518–532.

Smith, K. R., Shuhua, G., Kun, H., and Daxiong, Q. (1993). One hundred million improved cookstoves in China: How was it done? *World Dev.* **21,** 941–961.

Spengler, J. D., and Chen, Q. (2000). Indoor air quality factors in designing a healthy building. *Annu. Rev. Energy Environ.* **25,** 567–600.

Spengler, J. D., Samet, J. M., and McCarthy, J. F. (eds.). (2001). "Indoor Air Quality Handbook." McGraw-Hill, New York.

Spengler, J. D., and Sexton, K. (1983). Indoor air pollution: A public health perspective. *Science* **221,** 9–17.

Westhoff, B., and Germann, D. (1995). "Stove Images: A Documentation of Improved and Traditional Stoves in Africa, Asia, and Latin America." Brandes & Aspel Verlag, Frankfurt, Germany.

World Health Organization (WHO). (1997). "Health and Environment in Sustainable Development." World Health Organization, Geneva.

World Health Organization (WHO). (2002). "World Health Report 2002: Reducing Risks, Promoting Healthy Life." World Health Organization, Geneva.

World Resources Institute (with UNEP/UNDP and World Bank). (1999). "World Resources 1998–1999: A Guide to the Global Environment." Oxford University Press, New York.

Indoor Air Quality in Industrial Nations

ADRIAN F. R. WATSON
Manchester Metropolitan University
Manchester, United Kingdom

1. **Introduction**
2. **Sick Building Syndrome**
3. **Contribution of the Physical Environment to Indoor Air Quality**
4. **Indoor Chemical Pollutants**
5. **Indoor Biopollutants**
6. **Specific Microenvironments**
7. **Control of Indoor Air Pollution**
8. **Conclusions**

Glossary

biopollutants A collective term to describe airborne microorganisms, including fungi, bacteria, viruses, inflammatory agents (endotoxins), and indoor allergens such as house dust mites, which are found commonly within all indoor environments.

building-related illness A discrete, identifiable disease or illness that can be traced to a specific pollutant or source within a building (Legionnaires' Disease, for example).

environmental tobacco smoke (ETS) The mixture of smoke from the burning end of a cigarette and the smoke exhaled by the smoker (also called passive smoking).

radon (Rn) A radioactive gas formed in the decay of uranium. The radon decay products (also called radon daughters) can be inhaled into the lungs, where they continue to release radiation as they decay further.

sick building syndrome (SBS) A collection of common symptoms that affect the health and productivity of a substantial number of workers within a building; these symptoms may diminish or go away during periods when occupants leave the building. The symptoms cannot be traced to specific pollutants or sources within the structure.

ventilation The replacement of stale air within a space by clean air through dilution of existing indoor air with either cleaned and filtered outdoor air or recirculated indoor air; in combination with the removal of indoor pollutants using exhaust fans and filtration.

volatile organic compounds (VOCs) The collective term for all vapor-phase chemicals that contain organic carbon. There are numerous indoor sources of VOCs, including cooking and heating activities, common household products, and building materials.

Good indoor air quality is important for occupant health, contributing to a favorable working or home environment and a sense of comfort, health, and well being; when indoor air pollutants are present, however, indoor air quality deteriorates. To date, much of the attention relating to air pollution has focused on the risk and health effects of outdoor air quality, despite the fact that indoor levels of air pollutants are often 2–5 (and occasionally more than 100) times higher than outdoor concentrations. Recently, though, comparative risk studies such as those performed by the U.S. Environmental Protection Agency (EPA) have consistently ranked indoor air pollution among the top five environmental risks to public health. Furthermore, in industrialized countries, people spend the majority of time indoors, thereby increasing their exposure to indoor air pollutants relative to outdoor air pollutants. A typical adult spends approximately 60% of their time at home; 25% is spent in the workplace; 5%, at public places, shops, leisure pursuits, etc.; 5%, within transport vehicles; and 5–10%, outdoors. Some populations, often those most susceptible to health effects, such as the elderly, infants, the sick, and their caregivers, spend even more time indoors, usually in the home. However, within the general public, there is a notable lack of appreciation of this risk; the level of education regarding indoor air quality issues is low and the indoor environment is considered a place of refuge and safety. Furthermore, unlike outdoor air quality, indoor air problems are often subtle and do not always produce easily recognized impacts.

 © 2004 Elsevier Inc. All rights reserved.

1. INTRODUCTION

The main factors influencing indoor air quality include building design, construction materials, ventilation rates, and building maintenance; outdoor air quality; seepage of pollutants through the ground; occupant activities and preferences; and the occupants themselves. Figure 1 shows typical sources of indoor air problems in the home. The indoor air pollutants posing the greatest hazard include carbon monoxide, which is responsible for many accidental deaths every year; and radon and environmental tobacco smoke, both of which are implicated in the incidence of cancer. Other pollutants of concern are volatile organic compounds (VOCs), nitrogen dioxide, house dust mite allergens (all of which are implicated in the incidence of respiratory disease, especially in children), fungi, bacteria, and asbestos. A summary of the key indoor pollutants, sources, and effects is shown in Table I.

2. SICK BUILDING SYNDROME

Of all indoor air issues, sick building syndrome (SBS) has probably been the subject of most interest and study. It is a poorly defined syndrome that is described as involving a "building in which complaints of ill health (usually >20% of occupants) are more common than may be expected" and has been formally recognized as a disease by the World Health Organization (WHO) since 1986. SBS concerns are largely workplace based and relate to the health of individual workers and the productivity of the organizations employing them. A clear distinction can be made with building-related illnesses (BRIs), which are clinically diagnosed diseases such as humidifier fever or Legionnaires' disease, and are not considered to be SBS.

SBS is a relatively recent phenomenon; its incidence grew in the 1970s with the advent of more

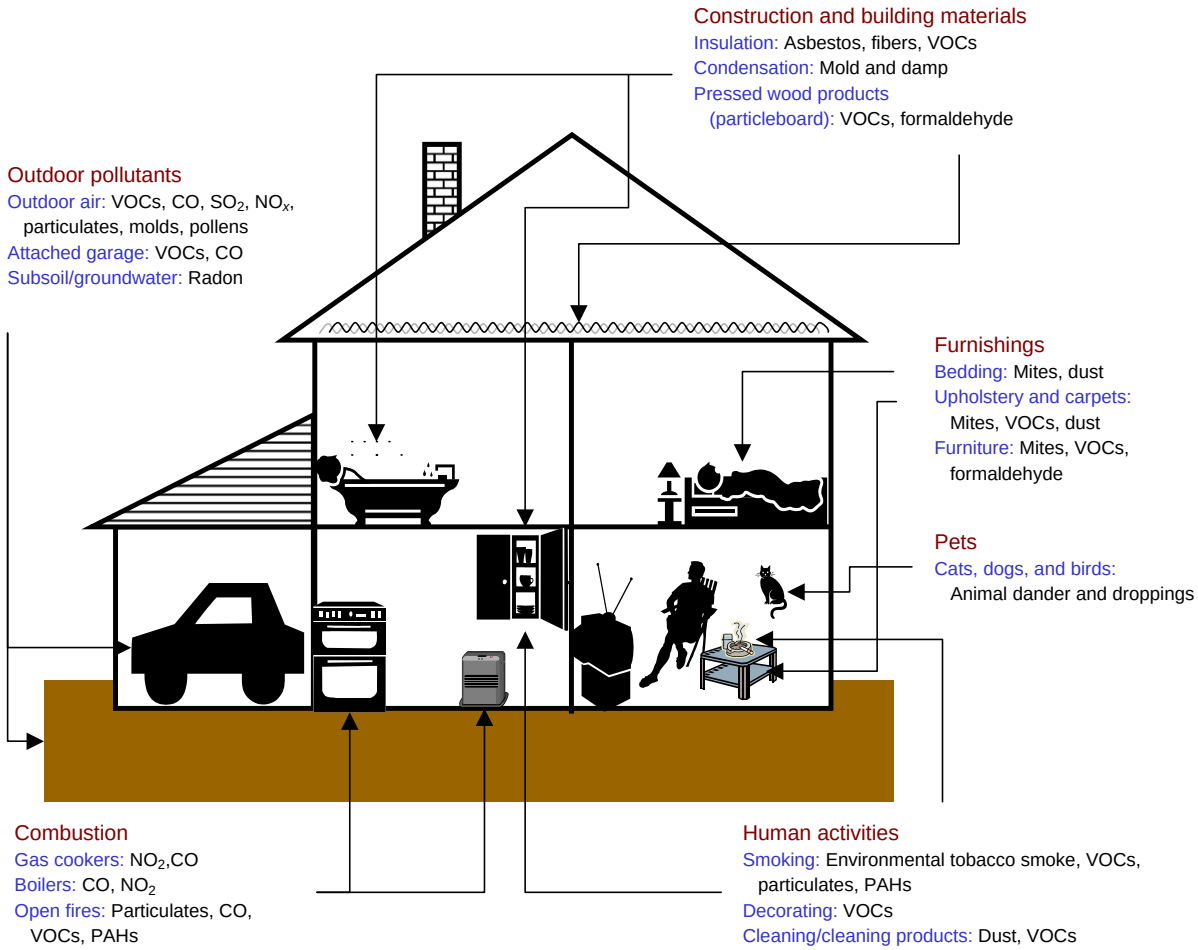


FIGURE 1 Sources of indoor air pollutants in the home. VOCs, Volatile organic compounds; PAHs, polycyclic aromatic hydrocarbons.

TABLE I

Indoor Air Pollutants, Sources, and Health Effects[a]

Indoor pollutant	Sources	Health effects
Carbon monoxide (CO)	Combustion appliances, environmental tobacco smoke, infiltrated exhaust from attached garages	Reduces capacity of blood to carry oxygen; increases frequency and severity of angina; headaches, decreased alertness, flulike symptoms in healthy adults; exacerbation of cardiopulmonary dysfunction in compromised patients; asphyxiation leading to death by suffocation
Radon	Rocks and soils containing naturally occurring radium and radon, well water, some building materials	Lung cancer
Environmental tobacco smoke (ETS)	Tobacco smoking (cigarettes, cigars, second-hand smoke)	Lung cancer, headaches and nausea, irritation to mucous membranes, chronic and acute pulmonary effects, cardiovascular effects
Biological contaminants (viruses, bacteria, fungi, molds, insect and arachnid excreta, endotoxins, pollen, animal and human dander)	Outdoors, humans, animals, damp building areas, improperly maintained heating ventilation and cooling systems, carpeting	Infectious diseases, allergic reactions (e.g., acute cases of asthma and skin irritations), toxic effects such as Legionnaires' disease
Volatile organic compounds (VOCs)	Paints, stains, adhesives, solvents, cleaners, pesticides, building construction materials, office equipment, upholstery fabrics, environmental tobacco smoke	Eye and respiratory irritation, neurotoxic effects, hepatotoxic effects, cancer
Formaldehyde (also a VOC)	Environmental tobacco smoke, foam insulation, resin in particle board, plywood paneling furnishings, carpets, upholstery fabrics	Headaches, dizziness, upper respiratory irritation, allergic responses (e.g., skin rashes), cancer
Polycyclic aromatic hydrocarbons (PAHs)	Environmental tobacco smoke, kerosene heaters, wood stoves	Cancer, cardiovascular effects, decreased immune function, atherosclerosis etiology
Pesticides	Pesticide application indoors and outdoors	Neurotoxicity, hepatoxicity, reproductive effects
Asbestos	Asbestos cement, ceiling and floor tiles, insulation, other building materials	Easily inhaled to cause lung damage, asbestosis, mesothelioma, cancer
Nitrogen dioxide (NO_2)	Combustion appliances, environmental tobacco smoke	Eye and respiratory tract irritations, decreased pulmonary function (particularly in asthmatics and children)
Particulate matter	Combustion appliances, environmental tobacco smoke	Cancer (soot, polycyclic aromatic hydrocarbons adsorbed to particles), irritation of respiratory tissues and eyes, decreased lung function
Dust, sprays, cooking aerosols	Personal activity	Unknown (can range from irritation to cancer)

[a] Data from Goodish (2001).

thermally efficient buildings, brought about by the 1973–1974 energy crisis. It is most common in new and recently refurbished buildings that contain complex heating, ventilation, and air-conditioning (HVAC) units, although it can be found in naturally ventilated buildings. The common features of SBS buildings are that they are airtight, warm (isothermal), energy-efficient buildings with forced ventilation and internally there are often a lot of textile coverings and soft furnishings. However, the symptoms of SBS cannot usually be traced to specific pollutants or sources within the building.

SBS symptoms are common among the general population but are experienced more frequently in the workplace. Typical symptoms include dryness or irritation of the eye, nose, or throat; lethargy, headaches, and dizziness; dry skin, itching, or rashes; airway infections, coughs, and wheezing; and unpleasant odor and taste sensations. Occasionally, more serious symptoms, such as high blood pressure,

have been reported. The symptoms are relieved when the worker leaves the workplace and recur following return to the affected building. SBS is poorly understood; because the etiology, symptoms, and cures often appear vague, it has been suggested that nonphysical issues within the workplace, such as organizational structure and work intensity, may contribute to an outbreak of SBS.

3. CONTRIBUTION OF THE PHYSICAL ENVIRONMENT TO INDOOR AIR QUALITY

The physical environment, characterized by temperature, humidity, and ventilation properties (and lighting), is the major determinant of comfort in the workplace, and departures from satisfactory conditions can have harmful health effects. Although workplace physical conditions tend to be controlled by local health and safety laws, there are no specific standards on the physical environment. However, two major professional groups, the Chartered Institute of Building Services Engineers (CIBSE) and the American Society of Heating, Refrigerating, and Air-Conditioning Engineers (ASHRAE) have been responsible for developing guidelines for workplace temperature, humidity, and ventilation.

3.1 Temperature and Air Movement

The temperature and air movement in the workplace are often the result of local workplace negotiation and Approved Codes of Practice (ACoP), rather than being enshrined in law, because the wide range of building functions requires a flexible approach. Cold temperatures can affect dexterity and mobility, and may increase physical and visual strain, with added problems for people with muscular pain, arthritis, and heart conditions. The design criteria of the CIBSE suggest a minimum temperature of 16°C in winter with a normal range of 21–23°C dry resultant temperature. In summer, the normal dry resultant temperature range is 22–24°C. The World Health Organization (WHO) recommends 24°C as the maximum temperature for working in comfort. Exceeding this temperature can lead to dehydration, heat stress, and ultimately heat stroke.

Air movement and drafts are also controlled in the workplace; health and safety laws merely stipulate that every enclosed workplace must have effective and suitable ventilation to provide a sufficient quantity of fresh or purified air, without specifying specific ventilation requirements. To assess fresh air, the CIBSE and ASHRAE usually identify carbon dioxide levels as a suitable indicator of human bioeffluents (odors) and a diagnostic tool in determining whether human comfort (odor) criteria are satisfied. The CIBSE quotes air exchange levels of 8 liters/sec per person (lsp) in no-smoking environments although design consultants increasingly use higher values (12–16 lsp). In the home, typical air movements are approximately 7 lsp. In many homes and workplaces, windows or other openings provide sufficient ventilation, although mechanical ventilation should be used if the openings are not suitable.

3.2 Humidity

Humidity is the amount of moisture in the air, and at saturation point, the relative humidity (RH) is 100%. To ensure occupant comfort, RH should be maintained between 40 and 70%. Health effects have been observed at the low levels of humidity (<30% RH), often seen during cold, dry winter months; desiccating effects seen in individuals can result in flu-like symptoms, such as stuffy or runny nose or sore eyes. However, indoor humidity should be maintained below 70% because high humidities can support the development of molds and can promote the growth of pathogenic or allergenic organisms such as *Legionella*.

4. INDOOR CHEMICAL POLLUTANTS

Indoor air quality is often compromised by a wide range of different chemical pollutants from a broad range of sources; the majority are gaseous (e.g., carbon monoxide), but some have particulate components such as environmental tobacco smoke and some are fibrous. The associated health risks should not be exaggerated because concentrations of indoor pollutants within homes and workplaces only usually affect morbidity through mild irritation, but occasionally some indoor chemical pollutants can lead to more extreme effects, such as lung cancers or accidental death. The indoor chemical pollutants that are of most current concern and interest are carbon monoxide, nitrogen oxides, volatile organic compounds, environmental tobacco smoke, radon, and fibers.

4.1 Carbon Monoxide

Carbon monoxide (CO) is a colorless, odorless gas produced by the incomplete combustion of carbon-containing substances. It is poisonous at high concentrations and is responsible for many accidental deaths worldwide (e.g., 600 per year in the United States). Outdoor levels can be determinants of indoor levels, particularly when a home has an attached garage or is in close proximity to heavily trafficked roads; however, the major source of indoor CO is incomplete fuel combustion (e.g., gas-fueled cooking and heating appliances, unvented paraffin space heaters, leaking chimneys and furnaces, and fireplaces), particularly where poor installation or ventilation or malfunctioning equipment reduces the oxygen concentration necessary for complete fuel combustion. Typical indoor concentrations range from 0.9 to 2.7 $mg\,m^{-3}$, although 15-minute peak concentrations have been measured at 180 $mg\,m^{-3}$ in homes using gas cooker grills.

CO preferentially combines with hemoglobin in the blood to form carboxyhemoglobin (COHb), interfering with the distribution of oxygen throughout the body. At high concentrations, COHb can cause unconsciousness and death, whereas lower concentrations cause a range of flulike symptoms such as headaches, dizziness, weakness, nausea, fatigue even in healthy people, and episodes of increased chest pain in people with chronic heart disease. Fetuses, infants, elderly people, and individuals with anemia or with a history of heart (angina) or respiratory disease can be especially sensitive to CO exposures. Because CO pollution is undetectable by human sensory perceptions of smell or taste, when accidental CO poisoning occurs, the victim tends to lapse slowly into unconsciousness and death with no realization about what is occurring. For this reason, CO is often given emotive terms such as "the silent killer."

Control of CO is achieved through regular professional maintenance and careful use of fuel appliances, ensuring that heating appliances are vented outdoors. Vehicles should not be idled inside a garage that is attached to the house.

4.2 Nitrogen Oxides

Like CO, the main indoor sources of nitrogen oxides (NO_x) are poorly ventilated combustion appliances, although in the absence of indoor sources, outdoor concentrations are often higher. The most commonly observed nitrogen oxides are nitrous oxide (N_2O), nitrogen oxide (NO), and nitrogen dioxide (NO_2), which is of most interest from a health perspective.

In the United Kingdom, 2-week average NO_2 concentrations range from 25 to 70 $\mu g\ m^{-3}$ with a gas cooker in the home and from 13 to 40 $\mu g\,m^{-3}$ without a gas cooker, although 1-hour averages of over 1000 $\mu g\,m^{-3}$ have been recorded in kitchens, which breaches most of the outdoor guideline levels (WHO guideline, 200 $\mu g\,m^{-3}$) for NO_2. Evidence from health studies of indoor NO_2 is inconclusive, but there is some evidence of an association with respiratory symptoms in children.

4.3 Volatile Organic Compounds

In indoor air quality studies, the term "volatile organic compounds" is typically used to describe all vapor-phase organic compounds. However, because VOCs exhibit a wide range of boiling points, they can be further subdivided into very volatile organic compounds (VVOCs; gaseous) with a boiling point ranging from <0 to 50–100°C, volatile organic compounds (50–100 to 240–260°C boiling point), and semivolatile organic compounds (SVOC; 240–260 to 380–400°C boiling point).

The VOC class of indoor air pollutants consists of more compounds than any other class, and hundreds of individual compounds have been identified at concentrations of 1 $\mu g\,m^{-3}$ or more. Total VOC concentrations typically range from 50 to 1000 $\mu g\,m^{-3}$ over long periods. However, some newly built offices and houses have higher source emissions, and over 130 individual compounds have been measured at a total VOC concentration of 13 $mg\,m^{-3}$ in these buildings. VOC sources are ubiquitous in the office and home environment, and include evaporative surfaces (such as vinyl flooring, carpets, or soft furnishings), building products, pressed wood products, adhesives, sealants, cleaning agents, paint, ventilation systems, combustion sources, cooking and space heating activities, tobacco smoke, or the infiltration of outdoor air.

Emissions of VOCs are either continuous (long-term, constant-strength-release sources dependent on temperature, relative humidity, and air velocity) or discontinuous (time-dependent, variable-strength, short-term emissions). Consequently, the variety of sources and emission characteristics confers on VOCs the most dynamic behavior of all indoor pollutants. Table II summarizes the common indoor VOCs and their sources.

One of the most researched individual VOCs is formaldehyde, a low-molecular-weight aldehyde found at variable concentrations indoors, ranging

TABLE II

Sources of Volatile Organic Compounds

Compound	Source materials
Aromatic hydrocarbons (benzene, toluene, xylenes, napthalenes) and aliphatic hydrocarbons	Paints, adhesives, sealants, gasoline, combustion products, damp-proof membranes, wallpaper, carpets, vinyl floor covering, creosote-impregnated timbers, tobacco smoke
Styrene	Insulation foam, jointing, textiles, disinfectants, plastics, paints, carpets, rubber floor tiles
Hexane	Floor covering, wallpaper, insulation foam, tobacco smoke
Chloroform	Chlorinated water
Terpenes (limonene, α-pinene)	Scented deodorizers, polishes, fabrics, fabric softeners, cigarettes, food, beverages
Polycyclic aromatic hydrocarbons and polychlorinated biphenyls	Combustion products (tobacco smoke, burning wood, kerosene heaters, burning paint)
Acrylic acid esters, alcohols	Polymers, aerosols, window cleaners, paints, paint thinners, cosmetics, adhesives, vinyl flooring
Ketones (e.g., acetone)	Lacquers, varnishes, polish removers, adhesives
Ethers	Resin, paints, varnishes, lacquers, dyes, soaps, cosmetics
Esters	Plastics, resins, plasticizers, lacquer solvents, flavors, perfumes

from 0.01 to 1.2 mg m^{-3} within homes (by comparison, outdoor levels in the United Kingdom range from 0.002 to 0.030 mg m^{-3}). Formaldehyde is commonly used in urea-formaldehyde- or phenol-formaldehyde-based resins, in foam insulation, as a binder in particleboard and wood paneling, and in many other building and household materials. Formaldehyde emitted from these materials into the building behaves as a common odorant at low concentrations (0.06–1.2 mg m^{-3}) and as an eye and airway irritant at higher concentrations (>0.5 mg m^{-3}). Smoking also contributes significantly to personal exposure, and as much as 25% of indoor formaldehyde can be accounted for by environmental tobacco smoke.

Industrial and occupational control of VOCs is achieved through health and safety legislation, but in the indoor environment no standards exist, although guidelines have been suggested in many industrialized countries (e.g., the WHO has recommended a guideline concentration of formaldehyde of 0.1 mg m^{-3} as a 30-minute average). Control is most often achieved through good ventilation practice and by the regulated prohibition or restricted use of products that emit VOCs, such as insulation foams or particleboards.

4.4 Environmental Tobacco Smoke

Environmental tobacco smoke (ETS), sometimes called secondhand smoke, has been recognized in many countries, including the United States, Australia, and the United Kingdom, as a leading cause of acute and chronic health effects in both smokers and nonsmokers, contributing to lung cancer, ischemic heart disease, serious respiratory illness, and asthma. Active cigarette smoking is considered the major preventable cause of morbidity and mortality in the majority of industrialized countries.

Environmental tobacco smoke is a combination of the sidestream smoke released from the burning end of the cigarette and mainstream smoke exhaled by the active smoker. It is a complex mixture of over 3800 gaseous and particulate (the visible smoke) chemical species, with burning tobacco the sole source. Smoking indoors raises the concentrations of indoor pollutants such as respirable particulates, NO_2, CO, acrolein, nicotine, and other VOCs at a rate determined by the number of smokers, the rate of smoking activity, room size, and air exchange rates within the room. Because of the lower temperatures associated with the burning cone of the cigarette, most partial-pyrolysis products are enriched in sidestream smoke rather than in mainstream smoke, and consequently sidestream smoke is often considered to contain higher concentrations of toxic and carcinogenic compounds.

Many health effects of involuntary smoking have been demonstrated in fetuses and children. Perinatal health effects on the baby include growth effects such as decreased birth weight, prematurity, and congenital malformations. Postnatal health effects on the

developing child may lead to increased risk of Sudden Infant Death Syndrome, reduced physical and cognitive development, and possibly longer term effects such as increased risk to childhood cancers (e.g., leukemia). During childhood, involuntary exposure to ETS has been linked to lower respiratory tract illnesses such as bronchitis and pneumonia, increased occurrence of asthma, and acute, persistent, or recurrent middle-ear disease.

The weight of scientific argument has led the International Agency for Research on Cancer (IARC) and many other regulatory authorities to consider ETS as a known (class A) carcinogen in adults. Other effects of involuntary ETS exposure in adults include irritation to eyes, nose, and upper and lower airways. There have been reports of increased coronary heart disease morbidity and mortality in active smokers, and acute respiratory morbidity in both active and passive smokers.

Control of ETS involuntary exposure has proved difficult; ventilation has not been shown to be fully effective and regulations specifying segregation of smokers into smoking areas or prohibition of smoking in locations such as the workplace and public buildings are often an emotional issue among the public.

4.5 Radon-222

Unlike other indoor chemical hazards, radon arises as a result of natural processes, not public or industrial activities. Radon is an inert, odorless, colorless, and tasteless radioactive gas, produced as part of the nuclear decay chain of naturally occurring ores bearing uranium-238 , such as granite. Because the radioactive half-life of radon-222 is 3.8 days, it has sufficient time to diffuse through the soil, into buildings, where it can reach potentially dangerous concentrations, or into the outdoor atmosphere, where dilution and dispersion render radon safe.

Radon enters buildings though three main routes: migration through cracks in foundations, basements, and lower floors of houses; dissolved in groundwater, where it can enter wells and then homes; or via radon-contaminated building materials used for home construction. Radon decays with the emission of an α particle, a form of high-energy, high-mass ionizing radiation that is very damaging to biological tissue, and a series of daughter decay products. The radioactively active decay products include materials such as plutonium-218 and lead-214, which can form molecular clusters or attach to aerosols and then be subsequently deposited on the lung epithelium, where further radioactive decay (with the release of both α and β particles) will damage the surrounding tissue, leading to lung cancer. Worldwide, there has been concern regarding the environmental health risk of radon in areas with the appropriate geology. The U.S. EPA considers radon to be the second greatest risk factor for lung cancer after smoking, and the cause of an estimated 5–20,000 deaths annually in the United States. Evidence also suggests that the risk of lung cancer increases 10–20 times if there is a combination of exposure to radon and ETS, as opposed to each pollutant individually.

Unlike most indoor air pollutants, radon is subject to many international regulations. Regional environmental or radiological protection agencies are required to test for radon in areas where concentrations in dwellings approach the action level (e.g., in Japan and United States, $\sim$150 becquerels (Bq) m^{-3}; United Kingdom, 200 Bq m^{-3}; Sweden, 400 Bq m^{-3}). If levels are considered actionable, control is relatively easy, usually by the identification and sealing of entry points, and/or increasing the level of ventilation in the home to increase the air exchange rate, thus reducing exposure.

4.6 Fibers (Asbestos)

Fibrous products are commonplace in modern buildings and are often employed for their thermal and acoustic insulation properties. Commonly used fibers are asbestos, a generic group term for several natural fibrous silicate minerals that were used widely until the 1980s, and vitreous and cellulose fibers, which are now commonly used as asbestos replacements.

Asbestos-containing materials (ACMs) have been used widely in building construction chiefly because of the fire prevention, chemical resistance, and insulation properties they offer relatively inexpensively. Example uses include roofing and pipe cladding; floor tiles, spray-applied fireproofing surfacing, and cement products. The low cost of ACMs has led to their widespread use in construction, particularly in public buildings such as schools, hospitals, and government buildings. Asbestos and asbestos-containing materials are among the most strongly regulated materials in the industrialized world and are classified as highly dangerous by most regulatory agencies, despite research that often questions this perceived level of risk, which is actually lower than that attributed to radon exposure. Many countries now require its removal from public buildings and have banned its use in current construction.

Asbestos is a particular hazard because the crystalline structure that underlies the larger fibers can cause fracture of asbestos-containing materials into thin, biologically active fibers, $>5\,\mu m$ long and $<3\,\mu m$ in diameter, which can become airborne. The small size of the fibers facilitates their easy passage into the alveolar region of the lungs, where they can cause great damage. The three principal important diseases linked with asbestos exposure are mesothelioma, asbestosis, and asbestos-related lung cancer. Mesothelioma, a malignant tumor that arises from the outer lining of the lung, tends cause mortality by compressing lung tissue. It is a rare condition associated with a long latency period (20–40 years) from first exposure and is almost always fatal despite treatment. Asbestosis, a disabling irreversible generalized scarification of the lung causing loss of pulmonary elasticity, can be fatal in severe cases. Asbestos-related lung cancer is a malignancy that arises within the lung when asbestosis is present. The development of these diseases is dose-related and the doses needed to induce them are substantial, although even very brief exposures to high levels of asbestos have been linked to asbestosis. Because the dose thresholds for all three diseases are orders of magnitude greater than exposures normally experienced by people who do not work with ACM, the risks are considered to be primarily through occupational exposure.

Vitreous and cellulose fibers are both irritants, but are considered safer than asbestos. Vitreous fibers, manufactured from sand or mineral wool, fracture into relatively larger fibers with a morphology different from that of asbestos, tending to settle out of the air more quickly, reducing exposure. Cellulose fiber is a recycled product made from materials such as wood dust or newsprint, with the addition of a fire retardant. It can act as a growth medium for microbiological growth in the presence of moisture, with potential respiratory tract irritation effects.

5. INDOOR BIOPOLLUTANTS

Despite being implicated in 35–50% of recent indoor air quality cases, indoor biopollutants, in contrast to chemical pollutants, have not been as widely studied. The wide range of biopollutants includes many airborne microorganisms; fungi, bacteria, and viruses, of which the fungi and bacteria are the most well studied; and inflammatory agents (endotoxins) and allergens from house dust mites, cockroaches, pets, and vermin. Like chemical pollutants, biopollutants can have indoor and outdoor origins. Pollens, a major outdoor group of biopollutants, are found indoors at highly variable levels, determined by outdoor levels and indoor/outdoor air exchange rates, but they are still believed to have a significant effect on morbidity. Biopollutants believed to be of most concern indoors are fungi and bacteria, especially the bacterium, *Legionella*, and allergens.

5.1 Fungi and Bacteria

Although many outdoor sources exist, particularly for fungi, many bacteria and fungi are found indoors associated with various forms of organic matter (e.g., surface coatings of walls, wood, and foodstuffs). Commonly seen bacteria include *Bacillus*, *Staphylococcus*, and *Micrococcus*, and common fungal species include yeasts, *Cladosporium*, and *Aspergillus*. Some species may be opportunistically pathogenic to humans, including the fungus *Aspergillus fumigatus* and the bacterium *Legionella*. The diversity and biomass of indoor fungi and bacteria vary with nutrient source, water availability, and temperature.

Mold and dampness are particularly associated with poor ventilation, and improved ventilation can be expected to reduce mold and fungi growth. Both bacterial and particularly fungal counts are influenced by season, with higher levels observed in the summer and autumn than in spring and winter. Several indoor infections have been attributed to the indoor fungi and bacteria, and a number of studies have indicated a relationship between home dampness (although not specific bacteria or fungi) and respiratory morbidity in children and adults. However, a lack of standardized sampling and analysis protocols for fungi mean that clear dose–response mechanisms in damp buildings have still to be established. A summary of reported health effects associated with exposure to bacteria and fungi can be seen in Table III.

5.2 Allergens (Endotoxins and House Dust Mites)

Endotoxins, proinflammatory substances present found in the outer membrane of gram-negative bacteria, have been associated with sick building syndrome. Levels in house dust above a median level of $1\,ng\,mg^{-1}$ have been shown to increase the severity of asthma, particularly in the presence of dust mite allergen. Sources of endotoxins are poorly defined but include contaminated humidifiers and areas that have suffered from severe water damage.

TABLE III

Diseases and Disease Symptoms Associated with Exposure to Bacteria and Fungi[a]

Disease syndrome	Examples of causal organisms
Asthma	Various *Aspergillus* and *Penicillium* spp., *Alternaria*, *Cladosporium*, mycelia sterilia, *Wallemia sebi*, *Stachybotrys*, *Serpula* (dry rot)
Atopic dermatitis	*Alternaria*, *Aspergillius*, *Cladosporium*
Extrinsic allergic alveolitis	*Cladosporium*, *Sporobolomyces*, *Aureobasidum*, *Acremonium*, *Rhodotorula*, *Trichosporon*, *Serpula* (dry rot), *Penicillium*, *Bacillus*
Humidifier fever	Gram-negative bacteria and their lipopolysaccharide endotoxins; actinomycetes and fungi
Rhinitis (and other upper respiratory symptoms)	*Alternaria*, *Cladosporium*, *Epicoccum*

[a]Data from Humfrey *et al*. (1996).

Other allergens include house dust mites (HDMs), cockroach allergens, and pet allergens. House dust mites are microscopic insects that grow rapidly in warm, damp conditions, thriving in carpets and bedding and feeding on dust, animal dander, and flakes of skin. HDM feces contain allergens that can trigger asthmalike reactions. Cockroach allergens are often found in kitchens and schools, and the common pet allergens from cats, dogs, and birds are found ubiquitously in urban indoor environments, either through pet ownership of because of transport of the allergen on contaminated pet owners to other locations.

5.3 Legionella (Legionnaires' Disease)

Legionnaires' disease is an uncommon form of pneumonia, first identified in 1976 in the United States after a large outbreak among a group of retired American service personnel who were attending a legion convention in a hotel in Philadelphia, Pennsylvania. Typical symptoms, including coughing, upper respiratory tract symptoms, and high fever, mainly affect susceptible populations (for example, the elderly and heavy smokers), but outbreaks can be fatal and the relatively high mortality rate means that the disease is associated with considerable public anxiety.

Infection is through the lungs, i.e., the route is airborne, and occurs as people breathe air containing fine water aerosols contaminated with *Legionella*. *Legionella* is a common bacterium found naturally in environmental water sources such as rivers, lakes, and reservoirs, usually in low numbers. However, if sufficiently high numbers of the bacteria get into water systems in buildings, they can present a risk to humans. Notable building sources of *Legionella* are air-conditioning equipment (cooling water systems, condensers), hot/cold water systems, humidifiers, nebulizers and medical humidifiers, showers, spray taps/mixer taps, water-based coolants (lathes, diamond wheels, and other rotating machines that generate aerosols), and whirlpool baths/spas. Wet cooling towers are considered a particular hazard, and have been identified as the cause of several outbreaks, because they generate large amounts of fine droplets, usually at rooftop level where there is potential for air currents to carry the droplets over a relatively large area.

Prevention and control of *Legionella* is through the careful design, maintenance, and operation of cooling towers (ideally, dry cooling towers) and water systems, removal of the conditions for bacterial growth, i.e., careful control of temperature of water systems, marinating water flow through pipes to avoid stagnation, and cleaning and disinfection (chlorination) of water systems to kill bacteria. There is no evidence that water systems in domestic homes present any risk from *Legionella*, but a significant proportion of public buildings, such as hotels, hospitals, offices, and factories, are known to be contaminated.

6. SPECIFIC MICROENVIRONMENTS

Although the majority of indoor air quality issues are general concerns relating to either the home or the office environment, there are certain locations where the indoor air quality has raised particular anxieties, either because there may be specific localized pollutant sources, or because of health and amenity issues requiring specialized control measures.

6.1 Hospitals

Usually the most important factor to consider with air pollution health effects is that a fraction of the population, usually the very young, the elderly, and the ill, will be affected more than the rest. The hospitalized population is largely drawn from just this fraction, and in most cases they spend 100% of their time exposed to an indoor environment. Within the

hospital environment, microorganisms such as viruses, bacteria, and fungi are all potential killers, patients have much lower immunity levels due to their poor health and because prescribed drugs may further suppress their immune system. Unfortunately, hospitals tend to be environments in which pathogens are concentrated; airborne microorganisms found in hospitals are generated within the building by the staff, patients, and visitors, with only a minority of the microorganisms, typically fungal spores, originating outdoors. Generally, the higher the occupancy level, the greater the microbial burden in the air; also, general hospital tasks can also contribute to the abundance of microorganisms, which means that microbial concentrations are often quite variable.

Examples of airborne infection sources include violent sneezing and coughing, which release and disperse droplets; activities such as bed making and areas such as cleaning, laundry, and waste storage rooms, all of which can cause microorganisms to become airborne. Airborne transmission is known to be the route of infection for diseases such as tuberculosis and has also been implicated in some hospital outbreaks of diseases such as methicillin-resistant *Staphylococcus aureus* (MRSA) and Legionnaires' disease. Nevertheless, the general role that airborne pathogen transmission has to play in hospital-acquired infection is still poorly understood, but is likely to be a much greater factor than is currently recognized.

Control of bioaerosols to prevent the spread of infection is best achieved through use of appropriate hospital ventilation systems and isolation rooms. This is technically difficult and costly, because many of the airborne pathogens are extremely small (e.g., viruses can be as small as 0.001 μm, although they infect via transport on larger particles), therefore air requires cleaning using high-efficiency particulate air (HEPA) filtration, which can achieve 99.997% removal of particulates. However, the combination of high cost and hospital policy often limits mechanical ventilation to the principal medical treatment areas, such as operating theaters and associated rooms, and patient wards are typically naturally ventilated, even though it is common practice for treatment to occur on these wards.

6.2 Schools

Although sick building syndrome is widely known to affect office workers and their performance, there is comparatively little awareness of the effects of the classroom environment on children's educational performance. This is despite the facts that children are particularly sensitive to air quality; because they breathe higher volumes of air relative to their body weights and are actively growing, therefore they have greater susceptibility to environmental pollutants compared to adults, and they spend more time in the school environment than in any environment other outside the home.

Schools, relative to other kinds of buildings, are seen as particularly likely to have environmental deficiencies that could lead to poor indoor air quality; furthermore, chronic shortages of funding in schools contribute to inadequate operation and maintenance of facilities. In the United States, approximately 25% of public schools have reported that ventilation was unsatisfactory, and indoor air quality has been reported to be unsatisfactory in about 20% of schools. Classrooms also have a high occupancy density in comparison to offices (United Kingdom occupancy density values: schools, 1.8–2.4 m^2 per person; offices, $\sim$10 m^2 per person), which can be a particular problem for primary school children, who spend prolonged periods in one classroom.

Current evidence linking pollutant levels in classrooms and student performance obtained from studies worldwide is limited but persuasive. Poor indoor air quality can adversely affect the comfort and health of students and staff, causing short-term health problems such as fatigue and nausea, as well as long-term problems such as asthma, all of which can affect concentration, attendance, student performance, and absenteeism levels with potentially important immediate and lifelong effects for the student.

6.3 Museums

The main function of museums, art galleries, libraries, and archives is the storage and display of important artifacts over long periods of time. Unfortunately, indoor pollutants, such as VOCs (particularly formaldehyde and acetic acid), water vapor, ozone, and sulfur dioxide, are all known for their ability to damage cultural property. Although deterioration is usually slow and progressive, prolonged exposure of archival material to some pollutants can cause severe damage. The pollutants can have outdoor sources but can also be released from the archive objects or from archive enclosures (particularly wooden enclosures). Therefore, to assist the safeguarding of collections, it is important to assess the air quality characteristics of the objects' microenvironments, galleries, and building as a whole. The maintenance of a stable humidity, to avoid the destructive effects of natural humidity variation, is seen

as being very important. Many collections are stored at "dry" humidity with typical values of 20–30% RH being maintained for corrodible collections, and 50–60% RH for organic/wooden collections. The use of and exposure of artifacts to high-VOC emission materials such as paints, adhesives, and cleaning products should also be tightly controlled.

Examples of archive deterioration include the action of sulfur oxides on photographic materials and collections of papers, causing brittleness and yellowing, and the emission of corrosive formic and acetic acids from sealed oak wood display cabinets, causing tarnishing or corrosion of metals (e.g., bronze or lead) and calcareous materials (e.g., egg shells or terra cotta).

7. CONTROL OF INDOOR AIR POLLUTION

In comparison to outdoor air quality, control measures for indoor pollutants are often considered poor. Currently, the main control of indoor air quality is usually assumed to be suitably maintained and operated ventilation and air-cleaning equipment; unfortunately, although this is effective for some parameters, these systems are not designed to maintain all aspects of air quality. Improvements, particularly in the short and medium term, can be gained through application and development of policy initiatives to control the sources of pollutants, such as smoking cessation in public places, restricting the sale of polluting products, imposing limits to activities that cause pollution, or developing standards for indoor pollutants. Longer term improvements may be likely if public awareness of indoor air quality is improved, such that eventually public and political pressure is applied to policymakers. This can be achieved using public health initiatives, providing better consumer product information, and improving the awareness of consumers' legal rights and the liabilities of landlords and manufacturers. The following sections focus on ventilation, using heating, ventilation, and air-conditioning systems, as the best current practice for indoor air quality control, and on legislation as the best hope for future short- and medium-term improvement.

7.1 Heating, Ventilation, and Air-Conditioning Systems

Although the functions and architecture of buildings are often taken for granted, they are highly complex dynamic systems, and the internal air quality within buildings depends on many interrelated factors. The principal engineering solution to providing good indoor air quality and thermal comfort is through ventilation and air-cleaning strategies. Ventilation is the replacement of stale air from a space by clean air through a combination of dilution of the existing indoor air, with either cleaned and filtered outdoor air or recirculated indoor air; and removal of indoor pollutants, using exhaust fans and filtration. To achieve this, a comprehensive heating, ventilation, and air-conditioning (HVAC) system is required, which must also maintain temperature and humidity at comfort levels for the building inhabitants. The main ventilation strategies to improve air quality are shown in Fig. 2.

Ventilation has a major impact on the energy use in buildings: it is estimated that more than 40% of primary energy in Organization for Economic Co-operation and Development (OECD) and European countries is consumed in buildings and half of that is dissipated through ventilation and air infiltration. The energy crisis of 1973–1974 acted as an impetus to an improvement in ventilation and building design, with an aim of conserving energy while creating an environment with good internal air quality. Therefore, ventilation systems that recirculate a large proportion of the indoor air have become more popular, owing to the large energy savings hat can be made by retaining heat within climate-controlled buildings. However, the high increase in reported indoor air quality problems, such as sick building syndrome, started in the 1980s, and the rise in symptoms and complaints appeared to correlate with energy conservation procedures. Consequently,

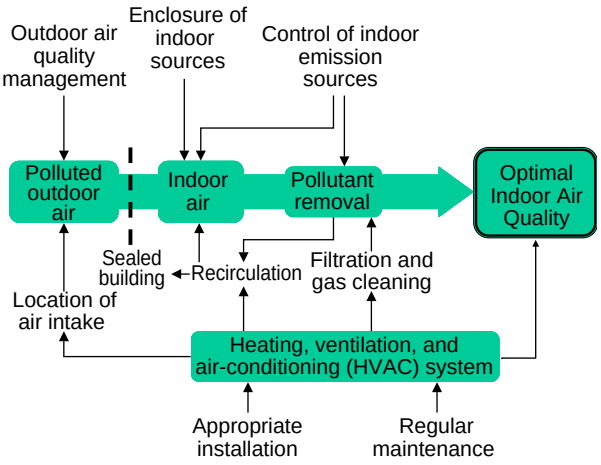


FIGURE 2 Ventilation strategies to achieve optimum indoor air quality.

careful design of HVAC systems is still required to provide a comfortable indoor environment for building inhabitants. To help achieve this, ventilation standards, such as standard 62 of the American Society of Heating, Refrigerating, and Air-Conditioning Engineers and guidelines of prepared by the Chartered Institute of Building Engineers have been developed to assist ventilation engineers.

The effectiveness of the HVAC system is dependent on a number of variables: initially, success depends on the proper design of the system relative to the building, then on installation and adherence to appropriate maintenance and operating procedures, the activity of people in the building, and the external air temperature, humidity, and air quality. Good outdoor quality air is a prerequisite for effective ventilation, because it provides the diluent for the internal stale air; however, in urbanized locations, constraints on outdoor air pollution such as traffic and industry mean that it is increasingly difficult to avoid outdoor pollutant entry across the building envelope. For this reason the location of the air intake duct is important, and close proximity to pollutant sources should be avoided.

Filtration is almost universally used in conjunction with mechanical ventilation and can be used to deal with both internal (recirculatory air filters) and external air pollutants. Gases are controlled using gas adsorption towers, and particle filters are available to trap particles of a wide range of different sizes. For example, in environments in which reduction of fine particle fractions of $<2.5\,\mu m$ is necessary, (e.g., hospitals), high-efficiency particle-arresting filters are available, although they are expensive to use because of the engineering and design constraints associated with such a high degree of filtration. Filtration is expensive and cannot be used in buildings that are naturally ventilated, or excessively leaky, because air will bypass the filtration units. Hence, for good ventilation, the building also has to be airtight, the tightness further contributing as a barrier to outdoor pollutants. Many energy-recovering ventilation systems cannot function unless the building is completely airtight.

Another feature of good ventilation design is the ability to contain polluting processes to specific locations and not disperse the pollutant around the building. This is achieved by negative-pressure regimes to keep pollutants in a space, and/or mechanical extraction to trap and remove the pollutants. Although HVAC systems can significantly improve air quality, if correct operation and maintenance procedures are not followed carefully indoor air quality problems can actually be created; for example, high humidities facilitate the growth of molds, and low humidity may be a source of discomfort.

7.2 Legislation

With the exception of radon and occupational environments, for which exposures are often covered by health and safety legislation, there are few indoor air pollution legislative measures or standards in industrialized countries. However, in 2000, the WHO issued statements and guidelines supporting the rights of individuals to healthy indoor air based on the basic principles of human rights, biomedical ethics, and ecological sustainability. The guidelines outlined the responsibility of both individual occupants and decision makers to create suitable indoor environments through public education programs, and other associated reduction or elimination strategies.

It is recognized that the creation of indoor air guidelines would be very difficult to implement for a number of reasons; for example, would guidelines be health or comfort based, should they be equivalent to outdoor standards for the same pollutants, would they apply to homes and offices, and who would be responsible for the compliance and monitoring of policies and guidelines? Despite these difficulties, it is apparent that there are benefits to establishing indoor air guidelines. Several countries, such as Germany, Norway, and Poland, have described target levels for pollutants, Australia uses indicators of good air quality rather than quantitative limits, and in some areas of the United States (California), VOC guidelines have been established for new building manufacturers. Other countries, such as the United Kingdom and South Africa, are also working to develop guidelines using typically observed concentrations and sources emission rates, which are both meaningful and contribute to public health.

8. CONCLUSIONS

As an environmental parameter, indoor air quality has the potential to affect health, comfort, and well being to an equal or greater extent than more well-known environmental issues such as outdoor air pollution; yet indoor air quality is still one of the more poorly understood environmental risks, by researchers, legislators, and the general public alike. Although the risks should not be overstated, it seems clear that there is a need for greater research into the indoor environment; for example, there is a need to

study the roles of copollutants (such as a combined exposure to biological allergens and chemical irritants, which may exacerbate health responses) and to develop improvements in ventilation practices, to engineer improved air quality levels alongside energy efficiency requirements. Finally, as health-based standards exist for outdoor air pollutants, policymakers and scientists should collaborate to develop and apply guidelines or standards to the indoor environment as well, particularly because indoor air pollutant concentrations are usually higher.

SEE ALSO THE FOLLOWING ARTICLES

Air Pollution from Energy Production and Use • Air Pollution, Health Effects of • Climate Change and Public Health: Emerging Infectious Diseases • Hazardous Waste from Fossil Fuels • Indoor Air Quality in Developing Nations • Refrigeration and Air-Conditioning

Further Reading

Godish, T. (2001). "Indoor Environmental Quality." Lewis, Boca Raton, Florida.

Green, E., and Short, S. (1998). "IEH Assessment on Indoor Air Quality in the Home (2): Carbon Monoxide." Institute for Environment and Health, University of Leicester, UK.

Humfrey, C., Shuker, L., and Harrison, P. (1996). "IEH Assessment on Indoor Air Quality in the Home." Institute for Environment and Health, University of Leicester, UK.

Spengler, J. D., McCarthy, J. F., and Samet, J. M. (eds.). (2001). "Indoor Air Quality Handbook." McGraw-Hill, New York.

Industrial Agriculture, Energy Flows in

DAVID PIMENTEL
Cornell University
Ithaca, New York, United States

1. Energy Resources
2. Energy Inputs in Corn Production
3. Trends in Energy Use in U.S. Corn production
4. Conclusion

Glossary

british thermal unit (BTU) The thermal energy required to raise the temperature of 1 lb (avoirdupois) of water 1°F at or near 39.2° F, the temperature at which water is at maximum density; equal to about 0.252 calories.

calorie The amount of heat required at a pressure of 1 atm to raise the temperature of 1 g of water 1°C.

kilowatt-hour A unit of work or energy equal to that expended in 1 hour at a steady rate of 1 kW.

Energy is important in the development of land, water, biological, and human resources in crop production. In addition to human energy, sunlight and fossil energy are the primary energy resources utilized in agricultural production. Because all technologies employed in crop production require energy resources, the comparative measure of energy flow in crop production provides a sound indicator of the technological changes that have taken place in agriculture. A major advantage in assessing technological changes in crop production lies in the fact that energy values (kilocalories, kilowatt-hours, joules, and British thermal units) for various resources and activities remain constant, in contrast to economic values, which are continually changing depending on the relative supply and demand of the resources and services. Another advantage of using energy as a measure of change in crop production technology is that it can help assess the impact of substitution of different forms of energy for various practices, as well as the impact of substitution of land, water, biological, and labor resources for fossil energy. The objective here is to assess the changes that have occurred in intensive agricultural production in the United States during the past 55 years. To the agricultural ecologist, a historical analysis of the technological changes that have occurred in crop production over the past decades can provide valuable insight for future agricultural policies.

1. ENERGY RESOURCES

The use of fossil fuels in crop production is a relatively recent event; the use of tools and techniques requiring fossil fuels did not grow rapidly until after 1940 (Tables I–III). As recently as 1850, the primary energy resources for society were fuel wood and human labor. Today, the United States consumes nearly 100 quads (1 quad $= 10^{15}$ Btu) of energy annually. On a per capita basis, this is the equivalent of about 8000 liters of oil per person per year. Approximately 17% of total U.S. energy is used in the food system, which represents about 1500 liters of fuel per person annually. This 17% includes 6% for agricultural crop and livestock production, 6% for processing and packaging of foods, and 5% for transport and preparation of food.

Since the oil crisis in 1973, the quantity of energy used in U.S. crop production has generally continued to grow. At the same time, the on-farm workforce has continued to decline. Some agriculturists proudly point to the fact that only about 2% of the U.S. workforce is involved in farming and thus are feeding the rest of the U.S. population. However, this is a misleading statistic because farmers hardly feed themselves entirely. They go to the same supermarket that everyone does to purchase food. Furthermore,

 © 2004 Elsevier Inc. All rights reserved.

TABLE I

Quantities of Various Inputs to Produce Corn Using Mechanization, from 1945 to 1985[a]

	Year				
Input	1945	1954	1964	1975	1985
Labor (hours/ha)	57	42	27	17	10
Machinery (kg/ha)	22	35	49	50	55
Fuel (liters/ha)					
Gasoline	120	150	125	60	40
Diesel	20	30	65	70	75
Manure	3000	1000	1000	1000	1000
Nitrogen (kg/ha)	8	30	55	111	152
Phosphorus (kg/ha)	8	13	36	56	58
Potassium(k/ha)	6	20	28	62	75
Lime (kg/ha)	145	124	203	220	426
Seeds (kg/ha)	11	17	21	21	21
Insecticides (kg/ha)	0	0.2	0.4	0.5	0.6
Herbicides (kg/ha)	0	0.1	0.4	3	3.5
Irrigation (%/ha)	11	2	5	16	18
Drying (kg/ha)	43	77	725	2290	3800
Electricity ($\times 10^3$ kcal)	8	24	60	90	100
Transport (kg/ha)	170	262	325	298	322
Yield (kg/ha)	2132	2572	4265	5143	7400

[a] From Pimentel, Dazhong, and Giampietro (1990).

they depend on Detroit for their tractors and other farm machinery, depend on the oil industry to fuel their tractors and trucks, and depend on the petrochemical industry for fertilizers and pesticides. After crops and livestock are harvested on a farm, the farmer relies on the transport and food-processing sectors to move and process the foods. Eventually, the foods are shipped to wholesalers and are handled by food retailers. Thus, for every farm worker, it estimated there are three to four farm-support workers. If the total workforce in the food system is considered, then about 20% of the U.S. workforce is involved in supplying food. Interestingly, about 20% of the workforce was involved in the food system in the United States in the 1920s. The primary difference today is that only a small portion in the work force is located on the farm.

2. ENERGY INPUTS IN CORN PRODUCTION

Corn yields in the United States remained relatively static, at about 1880 kg/ha, from 1909 to 1940 (Fig. 1). It was not until about 1945 that corn yields started to increase. At about this same time, hybrid corn varieties were introduced, commercial fertilizers and pesticides were applied, irrigation was used, and other new energy-intensive technologies in crop production were employed. From 1945 to 2000, corn yields increased about fourfold (Fig. 1). Concurrently, total energy input has increased also about fourfold (Tables I–III).

2.1 Labor Input

To raise corn by manpower alone requires about 1200 hours of labor per hectare during the growing season. Perhaps a few early settlers from Europe raised corn by hand, but by the time the United States gained independence, most corn culture involved the use of draft animal power, including oxen and horses. The use of 200 hours of oxen power to produce a hectare of corn reduces the manpower input, from 1200 hours for hand-produced corn, to about 400 hours. Thus, each hour of oxen input replaced 4 hours of manpower input. The estimate is that 120 hours of horsepower reduces the total labor input for corn production to about 120 hours, or 1 hour of horsepower replaces about 10 hours of manpower. Horsepower dominated U.S. corn production between 1910 and 1920.

Today, with the heavy mechanization of agriculture, the labor input in corn production has been reduced to about 6.2 hours per hectare (Table III). However, note that because of the technological changes, human labor is no longer required for direct power delivery. Thus, the labor input today is only 1/194th of the input required to produce corn by hand; however, this does not take into account, as mentioned, all the indirect labor input that goes into agricultural production, which includes the production of tractors and fuels. If the indirect labor input is taken into consideration, then current U.S. corn production uses about 1/40th of the labor input for hand-grown corn. This is still a dramatic reduction in the total amount of human activity required to produce corn over the past few decades.

2.2 Machinery and Power

Engine power has made a tremendous difference in U.S. society, as well as elsewhere in the world. This is illustrated by analyzing the manpower equivalent present in 3.78 liters (1 gallon) of fuel. This much fuel fed to a small gasoline engine will provide 20% of the heat energy produced in the form of power. Thus,

TABLE II

Energy Input for Various Items Used in Corn Production from 1945 to 1985[a]

	Year				
Item	**1945**	**1954**	**1964**	**1975**	**1985**
Labor[b]	31	23	15	10	6
Machinery[c]	407	648	907	925	1018
Fuel[d]					
Gasoline	1200	1500	1250	600	400
Diesel	228	342	741	912	878
Nitrogen[e]	168	630	1555	2331	3192
Phosphorus[f]	50	82	227	353	365
Potassium[g]	15	50	70	155	187
Lime[b]	46	39	64	69	134
Seeds[i]	161	421	520	520	520
Insecticides[j]	ND	13	27	50	60
Herbicides[k]	ND	7	40	300	350
Irrigation[l]	125	250	625	2000	2250
Drying[m]	9	15	145	458	760
Electricity[n]	8	24	60	90	100
Transport[o]	44	67	89	82	89
Total	2492	4111	5935	8855	10,309
Ratio	3.4	2.5	2.9	2.3	2.9
Yield	8528	10,288	17,060	20,575	29,600

[a] In thousands of kilocalories/hectare. ND, No data exist, although the value is >0. From Pimentel, Dazhong, and Giampietro (1990).

[b] Food energy consumed per laborer per day was assumed to be 3110 kcal from 1700 to 1970, 3300 kcal in 1975, and 3500 kcal from 1980 to 1983.

[c] The energy input per kilogram of steel in tools and other machinery was 18,500 kcal.

[d] A liter of gasoline and a liter of diesel fuel were calculated to contain 10,000 and 11,400 kcal, respectively. These values include the energy input for mining and refining.

[e] Nitrogen = 21,000 kcal/kg.

[f] Phosphorus = 6300 kcal/kg.

[g] Potassium = 2500 kcal/kg.

[h] Limestone = 315 kcal/kg.

[i] When hybrid seed was used from 1945 to 1985, the cost was 24,750 kcal/kg.

[j] Chlorinated insecticides dominated use from 1945 to 1964, and the energy input was calculated to be 67,000 kcal; from 1970 to 1985, carbamate and phosphate dominated use and the energy input for these was calculated to be 100,000 kcal/kg.

[k] Phenoxyl herbicides dominated use from 1945 to 1959, and the energy input was calculated to be 67,000 kcal/kg; 1964 to 1985, other types of herbicides dominated use, and energy input for these are calculated to be 100,000 kcal/kg.

[l] Water used per irrigated hectare was assumed to be 37.5 cm from 1945 to 1970 and 45 cm from 1974 to 1983. Irrigation was calculated to be 300,000 kcal/cm.

[m] The quantity of corn per hectare that required drying is shown. The energy required per kilogram dried was 200 kcal.

[n] Includes energy input required to produce the electricity.

[o] For the goods transported to the farm, an input of 275 kcal/kg ws included.

from the 31,000 kcal in the 3.78 liters of fuel, about 7.2 kWh of power can be produced, which is the equivalent of 10 hp/hour or 100 manpower hours. Therefore, the 3.78 liters of gasoline provides about 2.5 weeks of human work equivalents. This is part of the reason for the dramatic reduction in labor inputs, from 1200 hours/ha to produce corn by hand, to only 6.2 hours/ha employing mechanization (Table III).

Interestingly, the total amount of fuel consumed has changed only slightly, from about 880 liters/ha in 1975 to about 850 liters/ha today (Tables II and III). The reasons for the improvements include more efficient engines, a change from gasoline to diesel fuel, and larger, more efficient farm machinery. The weight of the machinery per labor hour has grown from 2.2 kg/hour in 1945 to about 9 kg/hour in 2000.

TABLE III

Energy Inputs and Costs of Corn Production per Hectare in the United States in 2000

Inputs	Quantity	Energy (kcal × 1000)	Costs
Labor	6.2 hr[a]	250[b]	62.00[c]
Machinery	55 kg[d]	1414[e]	103.21[f]
Diesel	90 liters[g]	900[e]	23.40[b]
Gasoline	56 liters[g]	553[e]	14.60[b]
Nitrogen	148 kg[i]	2738[j]	81.40[k]
Phosphorus	53 kg[i]	219[l]	12.72[k]
Potassium	57 kg[i]	186[l]	17.67[k]
Lime	699 kg[i]	220[e]	14.00[m]
Seeds	21 kg[d]	520[e]	74.00[i]
Irrigation	8.1 cm[n]	941[o]	81.00[p]
Herbicides	2.1 kg[i]	210[e]	21.00[q]
Insecticides	0.15 kg[i]	15[e]	6.00[r]
Electricity	13.2 kWh[g]	34[e]	2.38[s]
Transportation	222 kg[t]	268[e]	66.60[u]
Total	—	8468	579.98
Yield, 8590 kg[v]	—	30,924	—
		Input: Output (kcal) = 1:3.80	

[a] National Agricultural Statistics Service (1999), Available on the Internet at http//usda.mannlib.cornell.edu. Accessed 3/8/2002.
[b] It is assumed that a person works 2000 hours/year and utilizes an average of 8100 liters of oil equivalents per year.
[c] It is assumed that farm labor is paid $10/hour.
[d] Pimentel and Pimentel (1996).
[e] Pimentel (1980).
[f] Hoffman *et al.* (1994).
[g] U.S. Department of Agriculture (1991), "Corn-State. Costs of Production." Stock #94018. USDA, Economic Research Service, Economics and Statistics System, Washington, D.C.
[h] Diesel and gasoline assumed to cost 26.5¢/liter.
[i] U.S. Department of Agriculture, (1997), "Census of Agriculture." Available on the Internet at http://www.ncfap.org. Accessed 8/28/2002.
[j] An average of energy inputs for production, packaging, and shipping per kilogram of nitrogen fertilizer; from Food and Agriculture Organization, United Nations (1999), "Agricultural Statistics," available on the Internet at http://apps.fao.org, accessed 11/22/1999; and Duffy, M. (2001), "Prices on the Rise: How will Higher Energy Costs Impact Farmers?" available at http://www.ag.iastate.edu, accessed 9/3/2002.
[k] Manitoba Agriculture and Food (2002), "Soil Fertility Guide." Available on the Internet at http://www.gov.mb.ca. Accessed 9/2/2002.
[l] Food and Agriculture Organization, United Nations (1999), "Agricultural Statistics." Available on the Internet at http://apps.fao.org. Accessed 11/22/1999.
[m] Assumed to be 2¢/kg; see Clary, G.M., and Haby, V.A. (2002), "Potential for Profits from Alfalfa in East Texas." Available on the Internet at http://ruralbusiness.tamu.edu. Accessed 9/2/2002.
[n] U.S. Department of Agriculture (1997), "Farm and Ranch Irrigation Survey (1998). 1997 Census of Agriculture," Volume 3, Special Studies, Part 1. USDA, Washington, D.C.
[o] Batty and Keller (1980).
[p] Irrigation for 100 cm of water per hectare costs $1000; see Larsen, K., Thompson, D., and Harn, A. (2002), "Limited and Full Irrigation Comparison for Corn and Grain Sorghum." Available on the Internet at http://www.Colostate.edu. Accessed 9/2/2002.
[q] It is assumed that herbicide prices are $10/kg.
[r] It is assumed that insecticide prices are $40/kg.
[s] Price of electricity is 7¢/kWh (USBC, 2002).
[t] Goods transported include machinery, fuels, and seeds that were shipped an estimated 1000 km.
[u] Transport was estimated to cost 30¢/kg.
[v] U.S. Department of Agriculture (2001), "Agricultural Statistics: USDA" I–1–XV–34p. USDA, Washington, D.C.

Hence, where liquid fuel inputs have declined, fossil energy inputs for machinery have risen. In corn production during the early years of mechanization, the principal fuel use in tractors was 120 liters of gasoline and only 20 liters of diesel fuel. Over time, there was a shift toward diesel fuel; today, at least

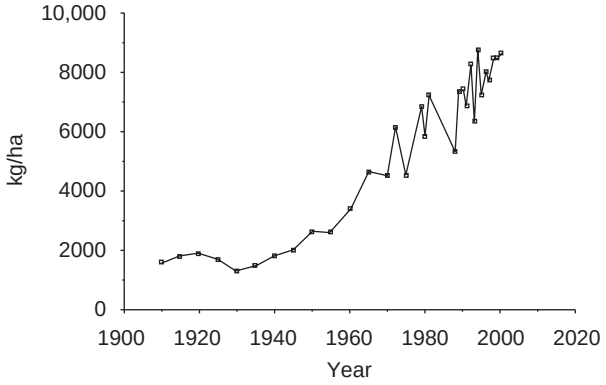


FIGURE 1 Corn grain yields from 1910 to 1999. From U.S. Department of Agriculture (1910–2000). "Agricultural Statistics." USDA, Washington, DC.

90 liters of tractor fuel is diesel and only 56 liters is gasoline (Tables I–III). Diesel fuel has advantages because it can provide 20–25% more power, or greater efficiency, per unit fuel energy, compared with gasoline. This is part of the reason for the improved corn yield per liter of fuel. Another reason for the improved fuel consumption per hectare has been the use of large farm equipment, which allows more operations to be performed during less time. The result is less fuel consumed per hectare of corn production.

2.3 Fertilizers

Early hand-produced corn and other crops depended on the nutrients that accumulated in the soil and derived from wild vegetation growing on uncultivated land during at least a 20-year fallow period. For example, in early slash/burn agriculture, vegetation was cut and burned to release the nutrients to the soil for crop production. Usually the land had to lay fallow for about 20 years so that nutrients would accumulate in the soil and vegetation. The land could then be tilled and planted for crops for approximately 2 years.

Early U.S. corn production was primarily organically based; that is, nutrients for corn production were provided mostly by livestock manure and green manure (legumes). In most cases, the farming system for corn required 2 ha of land to produce 1 ha of corn. For instance, 1 ha would be planted to a legume, such as clover or vetch, and the following year the legume would be plowed under and planted to corn. The legume would provide from 60 to 100 kg of nitrogen per hectare. As corn and other crops were harvested and removed from the land, the soils were being depleted of phosphorus, potassium, and calcium.

In 1945, when commercial fertilizers started to be used, only 8 kg/ha of nitrogen and phosphorus and 6 kg/ha of potassium were applied (Table I). By 2000, nitrogen application rates reached a high of 148 kg/ha, a nearly 20-fold increase. Note that the energy inputs for nitrogen alone in 1985 were greater than the total energy inputs for all items used in corn production in 1945 (Tables I and II). During the 38-year period, there was reported to be a 30% improvement in the efficiency of producing nitrogen fertilizer. Today the energy input for producing nitrogen for the farm, including production, packaging, and transport, is reported to be 18,590 kcal/kg.

Although the amounts of phosphorus and potassium applied per hectare rose significantly from 1945 to date, their use clearly did not grow as rapidly as occurred with nitrogen; both phosphorus and potassium require significantly less energy per kilogram to produce compared with nitrogen. The quantities of lime applied to land for growing corn also rose about threefold from 1945 to date, but lime is the least energy costly of the fertilizers that are used in corn production (Tables I–III).

2.4 Pesticides

Little or no pesticide was applied to U.S. corn crops in 1945. The quantity of insecticide used in corn production rose from 0.1 kg/ha in 1950 to 1.5 kg/ha in 2000. The insecticides applied to corn in the early 1950s and 1960s were primarily chlorinated insecticides such as dichlorodiphenyl trichloroethane (DDT) and dieldrin. Starting with the ban of DDT and other chlorinated insecticides in the early 1970s, there was a gradual shift from chlorinated insecticides to carbamate and phosphate insecticides. Although the quantity of insecticide applied per hectare increased only slightly with the newer insecticides, the energy required to produce a kilogram of insecticide increased and the toxicity increased 10- to 20-fold.

Changes in herbicide use also occurred in corn production starting in 1950. The first herbicide used in corn production was 2,4-dichlorophenoxyacetic acid (2,4-D), a phenoxy herbicide that was relatively efficient to produce in terms of energy input per kilogram. The newer triazines, such as atrazine, were applied during the 1960s to date and others were added later. The newer herbicides, although used at somewhat lower dosages, required more energy to produce compared with 2,4-D. The total energy

input for chemical weed control increased about 30-fold from 1954 to date (Tables I–III).

2.5 Irrigation Water

Corn requires enormous amounts of water for production. For instance, to produce about 8000 kg/ha requires about 5 million liters of water for photosynthesis and other physiological activities in corn production. To pump this much water from a depth of only 100 m and apply it to corn requires 12 million kcal. Irrigation of corn acreage has increased from less than 1% in 1945 to 16% today. The energy required for irrigation correspondingly increased from 125,000 kcal/ha to 941,000 kcal/ha (Tables II and III). The energy input for irrigation water during the period from 1945 to 2000 rose more than 7.5 fold.

2.6 Corn Drying

Another major change in corn production technology during the past 55 years has been the way corn is harvested and stored. Initially, corn was harvested as corn on the cob. Gradually, more corn has been harvested as grain directly in the field, and most corn is currently harvested in this manner. Corn harvested directly as grain in the field contains 25–30% moisture. The corn must not contain more than 13–15% moisture before being placed in storage. About 200 kcal of energy is required to dry 1 kg of corn. Harvesting corn on the cob and then drying the corn in corn cribs has been calculated to use 33% less fossil energy compared to harvesting the corn as grain and then drying it using fossil energy. The corn on the cob in the crib is dried by wind or solar energy.

2.7 Electricity

Electricity is used for a variety of purposes on farms, including moving goods in and out of storage, running fans, and operating many pieces of equipment. The energy inputs for electricity given in the calculations in Tables I–III include the primary energy required to produce the electricity. About 3 kcal of coal, for example, is required to produce 1 kcal of electricity. Note that the electricity energy tally does not include electricity used by the farm family.

2.8 Transport

Transportation costs in Tables I–III are relatively small because only machinery, fuel, and nitrogen fertilizer transport costs are calculated; the transport costs for the other items were already included. Note that overall, energy inputs for transportation increased sixfold from 1945 to 2000, confirming the intensiveness of agricultural management over this period.

3. TRENDS IN ENERGY USE IN U.S. CORN PRODUCTION

Energy inputs in corn production have continued to increase from hand and draft animal power systems to highly mechanized systems of today. Only an estimated 716,000 kcal was used to produce 1 ha of corn by hand in the U.S. in 1700, whereas more than 8 million kcal is employed in U.S. corn production today. This represents more than a 10-fold increase in energy inputs (Tables I–III). Corn yields during this period grew slightly more than 4.5-fold.

Several improvements in resource conservation and in technology occurred starting in 1975 and continue to evolve to date. Most of these changes have occurred as the result of higher prices for inputs, such as fertilizers and pesticides, and for fuels (especially natural gas prices), and lower prices for corn. There was a move away from gasoline use, toward diesel fuel, with important energy conservation consequences. Diesel fuel improves efficiency in tractors and other vehicles because diesel fuel has more kilocalories per gallon compared to gasoline (8179 kcal/liter for gasoline versus 9235 kcal/liter for diesel fuel). At same time, machinery was getting larger and more appropriate for the larger agricultural fields that were managed. Farmers were also maintaining their farm machinery more efficiently and keeping the engines well tuned.

Fertilizers have also been utilized more effectively than previously. For instance, more nitrogen fertilizer is being applied today during the growing season, when the corn crop can use the nitrogen, rather than over the frozen ground during the winter months, as was the previous practice. During the late 1960s and early 1970s, nitrogen was often applied during the winter months when labor and machinery were relatively abundant. Some of the nitrogen applied during the winter months was wasted because it leached into groundwater or washed into streams and rivers. The nitrogen pollution problem continues today; nitrogen has become a serious pollutant in groundwater, rivers, and lakes.

How do solar energy inputs relate to the use of fossil energy use in corn production? The solar energy reaching 1 ha during the year averages about

14×10^9 kcal. During the 4-month growing season in the temperate region, 7×10^9 kcal reaches 1 ha. With a corn grain yield of 8590 kg/ha plus another 8590 kg/ha in stover (fodder) biomass, the total energy in the corn represents about 0.4% of the solar energy captured. For the corn grain, the percentage is only 0.2%, or an input of nearly 50 kcal of solar energy per kilocalorie of corn grain produced. During the winter months, there is no growth of corn or other vegetation in the temperate region.

4. CONCLUSION

During the past 55 years, there have been a great many changes in U.S. corn production technology. A major change has been the increased use of fossil energy for labor, cropland, irrigation, nutrients, and pest control, all of which has contributed to increased environmental degradation. Fossil energy use has increased more than 3.4-fold during the past 55 years; during the same period, corn yields have increased 4-fold. The increase in the intensity of energy use, especially concerning oil-based sources of energy, makes U.S. agriculture very vulnerable to fluctuations in world oil prices and supplies. An understanding of energy resource use can aid in the search for conservation measures and improved agricultural technologies.

SEE ALSO THE FOLLOWING ARTICLES

Aquaculture and Energy Use • Diet, Energy, and Greenhouse Gas Emissions • Environmental Change and Energy • Fisheries and Energy Use • Food System, Energy Use in • Hunting and Gathering Societies, Energy Flows in • Industrial Ecology • Lifestyles and Energy • Livestock Production and Energy Use • Transitions in Energy Use

Further Reading

Batty, J. C., and Keller, J. (1980). Energy requirements for irrigation. *In* "Handbook of Energy Utilization in Agriculture" (D. Pimentel, Ed.), pp. 35–44. CRC Press, Boca Raton.

Ferguson, A. (2003). Implications of the USDA 2002 update on ethanol from corn. *Optimum Population Trust* (April 2003), pp. 11–17.

Giampietro, M., Ulgiati, S., and Pimentel, D. (1997). Feasibility of large-scale biofuel production. *BioScience* **47**(9), 587–600.

Hodge, C. (2002). Ethanol use in US gasoline should be banned, not expanded. *Oil Gas J.* (September 9), 20–30.

Hoffman, T. R., Warnock, W. D., and Hinman, H. R. (1994). "Crop Enterprise Budgets, Timothy-Legume and Alfalfa Hay, Sudan Grass, Sweet Corn and Spring Wheat under Rill Irrigation, Kittitas County, Washington. Farm Business Reports EB #1173." Washington State University, Pullman.

Lieberman, B. (2002). "The Ethanol Mistake: One Bad Mandate Replaced by Another." Competitive Enterprise Institute, September 17, 2002. Available on the Internet at http://www.nationalreview.com.

National Academy of Sciences NAS. (2003). "Frontiers in Agricultural Research: Food Health, Environment Communities." National Academy of Sciences, Washington, D.C.

Pimentel, D. (1980). "Handbook of Energy Utilization in Agriculture." CRC Press, Boca Raton.

Pimentel, D. (2001). The limitations of biomass energy. *In* "Encyclopedia of Physical Science and Technology," pp. 159–171. Academic Press, San Diego, CA.

Pimentel, D. (2003). Soil Erosion: A major environmental threat. *In*: "Making Development Work: A New Role for Science" (G. Leclerc, and C. A. S. Hall, Eds.). University of New Mexico Press, Albuquerque (in press).

Pimentel, D., and Pimentel, M. (1996). "Food, Energy and Society." Colorado University Press, Boulder.

Pimentel, D., Dazhong, W., and Giampietro, M. (1990). Technological changes in energy use in U.S. agricultural production. *In* "Agroecology" (S. R. Gleissman, Ed.), pp. 305–321. Springer-Verlag, New York.

Pimentel, D., Doughty, R., Carothers, C., Lamberson, S., Bora, N., and Lee, K. (2002). Energy inputs in crop production: Comparison of developed and developing countries. *In* "Food Security & Environmental Quality in the Developing World" (R. Lal, D. Hansen, N. Uphoff, and S. Slack, Eds.), pp. 129–151. CRC Press, Boca Raton.

Pimentel, D., Houser, J., Preiss, E., White, O., Fang, H., Mesnick, L., Barsky, T., Tariche, S., Schreck, J., and Alpert, S. (1997). Water resources: Agriculture, the environment, and society. *BioScience* **47**(2), 97–106.

Pimentel, D., Bailey, O., Kim, P., Mullaney, E., Calabrese, J., Walman, L., Nelson, F., and Yao, X. (1999). Will the limits of the Earth's resources control human populations? *Environ. Dev. Sustain.* **1**, 19–39.

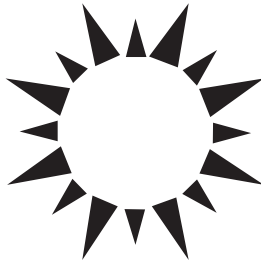
United States Bureau of the Census (USBC). (2002). "Statistical Abstract of the United States 2002." USBC, U.S. Govt. Printing Office, Washington, D.C.

Wereko-Brobby, C., and Hagan, E. B. (1996). "Biomass Conversion and Technology." John Wiley & Sons, New York.

Youngquist, W. (1997). "GeoDestinies: The Inevitable Control of Earth Resources Over Nations and Individuals." National Book Company, Portland, Oregon.

Industrial Ecology

AMIT KAPUR and THOMAS E. GRAEDEL
School of Forestry and Environmental Studies, Yale University
New Haven, Connecticut, United States



1. **Introduction to Industrial Ecology**
2. **Methods and Tools of Industrial Ecology**
3. **Industrial Ecology and Energy**
4. **Conclusion**

Glossary

design for environment An engineering perspective in which environmentally related characteristics of a product, process, or facility design are optimized.

ecoefficiency A business strategy to produce goods with lower use of materials and energy, to realize economic benefits of environmental improvements.

industrial ecology An approach to the design of industrial products and processes that evaluates such activities through the dual perspectives of product competitiveness and environmental interactions.

industrial metabolism A concept to emulate flows of material and energy in industrial activities from a biological systems perspective.

industrial symbiosis A relationship within which at least two willing industrial facilities exchange materials, energy, or information in a mutually beneficial manner.

life cycle assessment A concept and a methodology to evaluate the environmental effects of a product or activity holistically, by analyzing the entire life cycle of a particular material, process, product, technology, service, or activity. The life cycle assessment consists of three complementary components: (1) goal and scope definition, (2) inventory analysis, and (3) impact analysis, together with an integrative procedure known as improvement analysis.

material flow analysis An analysis of flow of materials within and across the boundaries of a particular geographical region.

pollution prevention The design or operation of a process or item of equipment so as to minimize environmental impacts.

recycling The reclamation and reuse of output or discard material streams for application in products.

remanufacture The process of bringing large amounts of similar products together for purposes of disassembly, evaluation, renovation, and reuse.

In industrial ecology, the approach to understand industry–environment interactions is to move from contemporaneous thinking, or thinking about past mistakes, to forward thinking. The objective is to minimize or eliminate environmental impacts at the source rather than to rely on traditional end-of-pipe measures in a command and control regime. If properly implemented, industrial ecology promotes business competitiveness and product innovation. In addition, industrial ecology looks beyond the actions of single firms to those of groups of firms or to society as a whole.

1. INTRODUCTION TO INDUSTRIAL ECOLOGY

Industrial ecology is a nascent and challenging discipline for scientists, engineers, and policymakers. Often termed the "science of sustainability," the contemporary origins of industrial ecology are associated with an article titled "Strategies for Manufacturing," written by Frosch and Gallopoulos and published in 1989 in *Scientific American*. However, historically, indirect references to the concept of industrial ecology date back to the early 1970s. The multidisciplinary nature of industrial ecology makes it difficult to provide a consistent and universally accepted definition, but the following statement captures the essence of the topic:

> *Industrial ecology is the means by which humanity can deliberately and rationally approach and maintain sustainability, given continued economic, economic, cultural, and technological evolution. The concept requires that an industrial ecosystem be viewed not in isolation from its surrounding system, but in concert with them. It is a*

 © 2004 Elsevier Inc. All rights reserved.

> *systems view in which one seeks to optimize the total materials cycle from virgin material, to finished material, to component, to product, to obsolete product, and to ultimate disposal. Factors to be optimized are resources, energy and capital.*
>
> *—Graedel and Allenby, 2002*

Several core elements characterize the discipline: the biological analogy, the use of systems perspectives, the role of technological change, the role of companies, ecoefficiency and dematerialization, and forward-looking research and practice. Each of the themes offers a plethora of methods and tools for analysis. In following section, some of the most important aspects and tolls of the core elements, especially those particularly relevant to energy, are discussed.

2. METHODS AND TOOLS OF INDUSTRIAL ECOLOGY

Industrial ecology offers a realm of methods and tools to analyze environmental challenges at various levels—process, product, facility, national, and global—and then to come up with responses to facilitate better understanding and provide suitable remedies. Some of the important components in the industrial ecology toolbox are discussed in the following sections.

2.1 Life Cycle Assessment

A central tenet of industrial ecology is that of life cycle assessment (LCA). The essence of LCA is the examination, identification, and evaluation of the relevant environmental implications of a material, process, product, or system across its life span, from creation to disposal or, preferably, to recreation in the same or another useful form. The formal structure of LCA, contains three stages: goal and scope definition, inventory analysis, and impact analysis, each stage being followed by interpretation of results. The concept is illustrated in Fig. 1. First, the goal and scope of the LCA are defined. An inventory analysis and an impact analysis are then performed. The interpretation of results at each stage guides an analysis of potential improvements (which may feed back to influence any of the stages, so that the entire process is iterative). There is perhaps no more critical step in beginning an LCA evaluation than to define as precisely as possible the evaluation's scope: What materials, processes, or products are to be considered, and how broadly will alternatives be defined? To optimize utilization of resources in an LCA exercise, the depth of analysis should be keyed to the degree of freedom available to make meaningful choices among options, and to the importance of the environmental or technological issues leading to the evaluation.

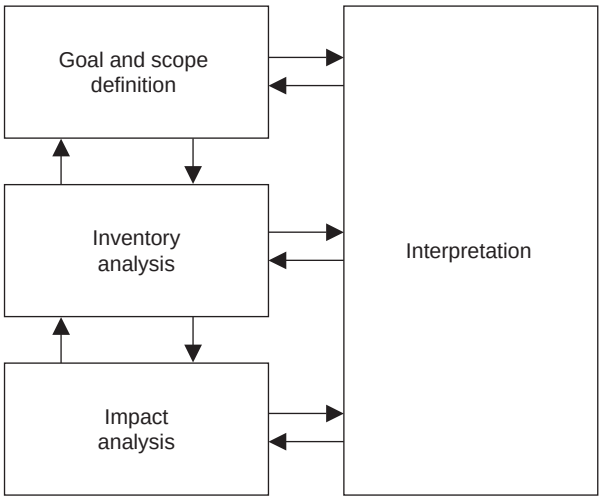


FIGURE 1 Stages in the life cycle assessment of a technological activity. The arrows indicate the basic flow of information. Reprinted from *Guidelines for Life-Cycle Assessment: A Code of Practice*, Society of Environmental Toxicology and Chemistry (1993), with permission.

The inventory analysis is by far the most well-developed component of LCA. It uses quantitative data to establish levels and types of energy and materials used in an industrial system and the environmental releases that result. The impact analysis involves relating the outputs of the system to the impact on the external world into which outputs flow, or, at least to the burdens being placed on the external world. The interpretation-of-results phase occurs when the findings from one or more of the three stages are used to draw conclusions and recommendations. The output from this activity is often the explication of needs and opportunities for reducing environmental impacts as a result of industrial activities being performed or contemplated.

A comprehensive LCA can be expensive and time consuming. As a consequence, more efficient approaches (streamlined LCAs, or SLCAs) have been developed with the intention of retaining the useful broad-scope analysis of the LCA while making the activity more tractable. In the case of either LCA or SLCA, the effort helps the analyst think beyond the boundaries of a particular facility or process to encompass the full measure of associated environmental implications.

2.2 Design for Environment

Product design engineers are always faced with the challenge of optimizing the multitude of attributes that determine the success or failure of the product. The paradigm for such design considerations, Design for X (DfX), states that X may be any of a number of attributes, such as assembly, compliance, disassembly, environment, manufacturability, reliability, safety, and serviceability. Design for Environment (DfE) is the DfX-related focus of industrial ecologists. The core theme of DfE philosophy is that it should improve the environmentally related attributes of a product while not comprising other design attributes such as performance, reliability, aesthetics, maintainability, cost, and time to market. DfE approaches systematically evaluate environmental concerns during the product life cycle stages of premanufacture, manufacture, delivery and packaging, use, and end of life, and accordingly set targets for continual improvements. The choice of materials during pre-manufacture and their efficiency of utilization during product manufacture, energy use during manufacturing and product use, and the environmentally friendly disposal or reincarnation of products at end of life are some of the prime considerations in DfE. DfE is also a "win-win" proposition in that it provides a corporation with a competitive edge in an ever-tightening regulatory environment, and promotes ongoing product innovation.

2.3 Industrial Symbiosis

The industrial ecologist views the economy as a closed system, similar to a natural system, in which the "residues" from one system are the "nutrients" for another. The concept known as industrial symbiosis is a current topic of research for industrial ecologists and environmentalists in identifying strategies to enable businesses to "close the loop." The objective is to create or encourage the formation of industrial production systems that function similarly to biological food chains. In either natural or industrial systems, symbiosis occurs when two or more organisms form an intimate association, either for the benefit of one of them (parasitic symbiosis) or for both or all of them (mutualistic symbiosis), such that there is high degree of synergy between input and output flows of resources. The most well-known industrial symbiosis example is the Kalundborg industrial system in Denmark. The qualitative material flows at Kalundborg are shown in

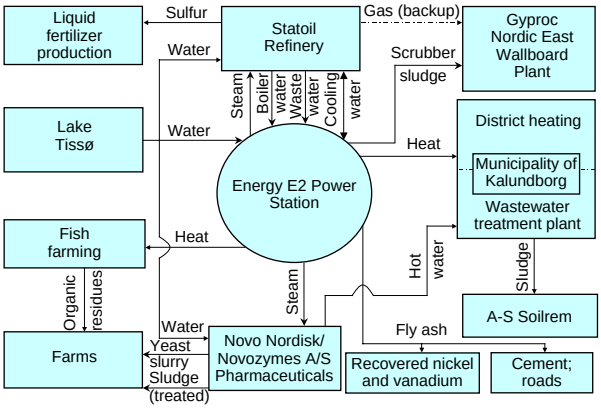


FIGURE 2 Material flows in the Kalundborg, Denmark industrial ecosystem. Reprinted from the research work of M. R. Chertow, Yale University, with permission.

Fig. 2. The heart of this industrial ecosystem is Denmark's largest power plant, which exchanges various residues (gypsum, steam, fly ash, etc.) with neighboring entities, i.e., the Statoil refinery, the Nova Nordisk pharmaceutical unit, the Gyproc plasterboard facility, and the town of Kalundborg.

Industrial symbiosis may occur opportunistically or can be planned. Coordinating industrial symbiosis appears to offer the promise of developing industrial ecosystems that are superior environmentally. Such a system would need to involve a broad sectoral and spatial distribution of participants, and to be flexible and innovative. The formation of ecologically balanced industrial systems results in numerous environmental and economic benefits. Economic benefits, which are shared by participating businesses, governments, and communities, are the primary driving force for setting up such industrial configurations. Entrepreneurs can gain appreciable cost savings from reduced waste management, reduced infrastructure costs, and improved process and product efficiency. There are opportunities for other cooperative ventures such as joint purchasing, combined waste recovery and treatment, employee training, environmental monitoring, and disaster response. The tangible environmental benefits include the reduction of greenhouse gas emissions and toxic air emissions, improving efficiency and conservation in the use of energy, materials and water, improving land use planning and green space development within the industrial complexes, and promotion of pollution prevention and recycling approaches.

2.4 Ecoefficiency, Dematerialization, and Decarbonization

The anthropogenic activities of production and consumption, both being integral parts of a robust economic world, are considered to be major factors in the sustainability debate. The concept of ecoefficiency evolved in the early 1990s, prior to the Earth Summit. Simply stated, ecoefficiency means doing more with less. The World Business Council for Sustainable Development defines ecoefficiency as being "attained by the delivery of competitively priced goods and services that satisfy human needs and bring quality of life, while progressively reducing ecological impacts and resource intensity throughout the life cycle, to a level at least in line with the earth's estimated carrying capacity." Businesses worldwide have begun to embrace ecoefficiency as a mutually beneficial strategy. Ideologically, ecoefficiency and dematerialization are synonymous. Dematerialization can be achieved by making products either smaller or lighter (e.g., personal computers today as compared to a decade ago), by replacing a product with an immaterial substitute (e.g., large-scale use of electronic mail rather than regular postal mail), or by reducing the use of material-intensive systems. Intensity of use of materials, defined as resource use per gross domestic product, is a commonly used metric to evaluate dematerialization trends. The key determinants that influence patterns of intensity of use are changes in material composition of products (material substitution and technological change) and product composition of income (structural changes in the economy and intrasectoral shifts). The variation of material intensity of use over the past century for some of the materials used extensively in the United States economy is shown in Fig. 3. There is an appreciable decline in the use of metals except for aluminum, whereas the use of paper and plastics continues to grow. These trends also indicate that the composition of materials in the United States economy changed over this period from dense to less dense, i.e., from iron and steel to light metals, plastics, and composites. At the product level, similar trends can also be observed; for example, the change in the weight composition of an average automobile in the United States is shown in Fig. 4. Plastics, composites, aluminum, and other specialty materials account for the decline in the use of conventional steel in automobiles over the 1978–2000 period.

In energy systems analysis, dematerialization is analogous to decarbonization, which refers either to moving away from carbon-intensive energy sources to completely carbon-free sources such as hydrogen or solar energy. Although evidence for any comprehensive dematerialization trend remains ambiguous, the situation is clearer with respect to energy. Historically, trends in fuel market share of global primary energy use indicate a transformation from carbon-intensive fossil fuels such as wood and coal to less intensive ones such as oil and natural gas, as shown in Fig. 5. The decline in carbon intensity has been very slow, approximately 0.3% per year over the 1850–1994 time period.

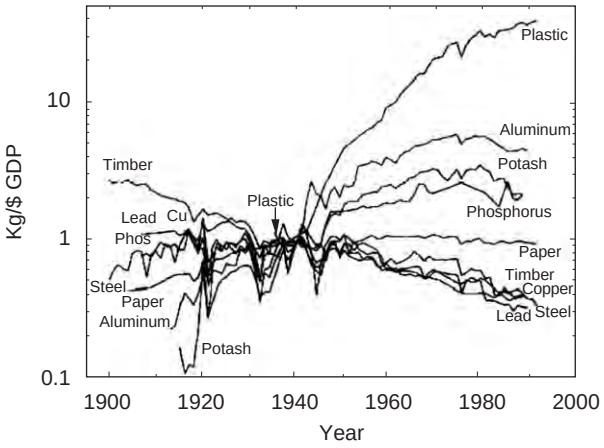


FIGURE 3 The intensity of use of materials in the United States, 1900–1990. The annual consumption data are divided by gross domestic product (GDP) in constant 1987 dollars and normalized to unity in the year 1940. Reprinted from Wernick (1996), with permission.

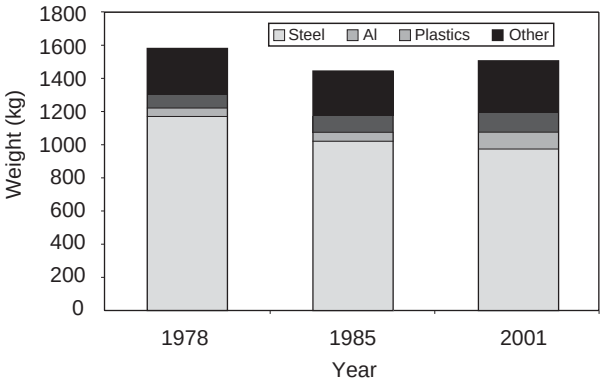


FIGURE 4 The average material composition of a domestic automobile in United States. Data from the U.S. Department of Energy (2001).

2.5 Industrial Metabolism

The extraction, production, and waste treatment of materials cause environmental problems that call for

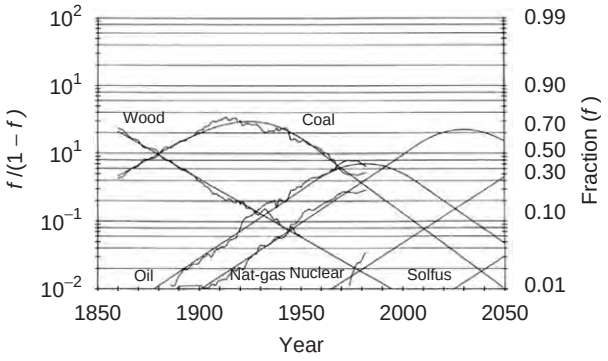


FIGURE 5 Global primary energy substitution, 1860–1980, and projections to 2050 (expressed in fractional market shares, *f*. Smooth lines represent model calculations and jagged lines are historical data. "Solfus" is a term employed to describe a major new energy technology (for example, solar or fusion). Reprinted from Nakićenović (1997), with permission.

an intervention by all the stakeholders: governments, corporations, and individuals. The approach to link society's management of materials to sustainable development and environmental quality is now well defined and understood. The concept of "industrial metabolism," first introduced by Robert Ayres, establishes an analogy between economy and environment on a material level. Societies mobilize material from Earth's crust to create "technomass," in analogy to nature's biomass. The ecosystems in the biosphere try to close the loop by cycling resources and wastes repeatedly. In contrast, modern society tends to use materials once and then discard them. Industrial ecologists have begun to assess the physical economy through the lens driven by biosphere guiding principles. The objective of such analytical studies from a systems perspective is to determine the anthropogenic contribution to natural material flows, the causal factors, and the spatial and temporal distribution of environmental problems.

Material balances, based on the universal law of mass conservation, have historically been used as an instrument to describe material flows through the economy. Within that framework, material flow analysis (MFA) and substance flow analysis (SFA) are two important approaches to assess the current and future state of resource flows and accumulation in the economy and environment. MFA usually tracks material intensities of national economies or sectors thereof, concentrating on bulk flows, whereas SFA is intended for flows of specific substances to identify specific environmental problems and propose remedial/prevention strategies.

An example of an SFA, a contemporary copper cycle for Germany, is shown in Fig. 6. The cycle for Germany represents a highly industrialized country without significant virgin copper resources. Unable to

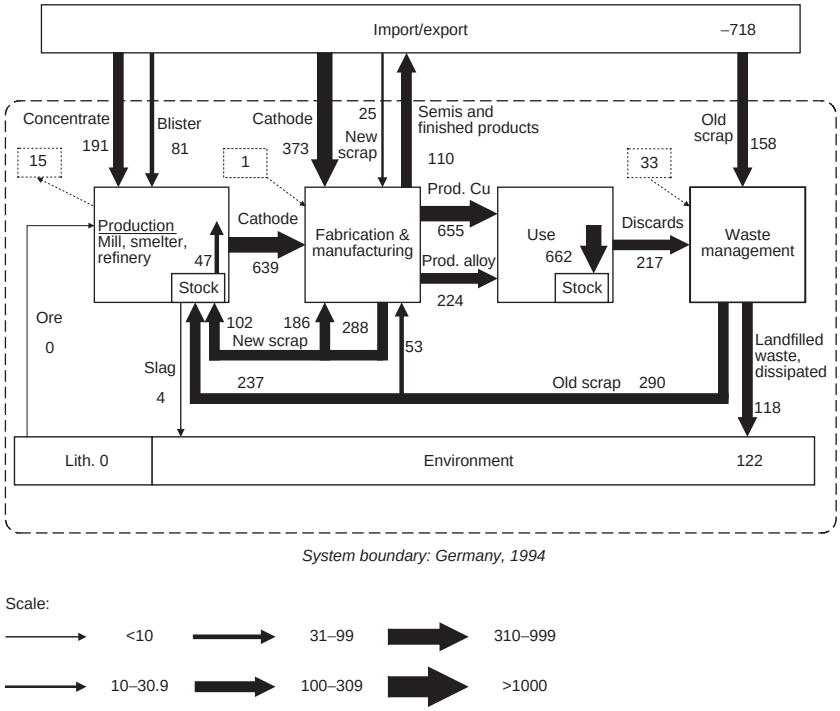


FIGURE 6 The contemporary copper cycle for Germany in 1994. All values are in Gg of Cu per year.

rely on an internal supply, Germany imports copper in different forms at every industrial life stage (production, fabrication and manufacturing, and waste management). Its reuse of discarded and imported scrap is so extensive that nonvirgin copper supplies represent about 45% of all inflows to its industrial facilities. Nonetheless, Germany landfills about 120 Gg Cu/year, more than any country in Europe. This substance diagram can be converted to an energy diagram if the flows are multiplied by the energy needed to effect a transfer of ore unit of copper from ore reservoir to another. The results make it obvious that the extraction of copper from ore is the most energy-intensive step in the entire cycle.

2.6 IPAT Equation

The environmental impact of materials use has often been conceptually formed by the "IPAT" equation:

$$I = P \times A \times T \quad (1)$$

in which the overall environmental impact I is expressed as the product of population P, the affluence A (expressed, for example, as gross domestic product [GDP] per person), and technology T (expressed, for example, as environmental impact per unit of per capita GDP). Historically proposed by Commoner, Ehrlich, and Holdren, the IPAT equation is now the "master equation" of industrial ecology. If the technology factor is expanded somewhat, the equation can be rewritten as

$$I = P \times A \times M \times D \times H \quad (2)$$

where M is the materials intensity, D is the dissipation factor, and H is the hazard factor, which depends on the chemical form of the material lost and the susceptibility of the receiving ecosystem. In words, Eq. (2) becomes

$$\begin{aligned} &\text{Environmental impact} \\ &= (\text{Population}) \times (\text{GDP/person}) \\ &\quad \times (\text{units of material/GDP}) \\ &\quad \times (\text{units of pollution/unit of material}) \\ &\quad \times (\text{impact/unit of pollution}) \end{aligned} \quad (3)$$

In the energy sector, the equivalent of the IPAT equation is the Kaya's identity, a mathematical expression for energy-related carbon emissions that can be written as

$$C = (P) \times (GDP/P) \times (E/GDP) \times (C/E) \quad (4)$$

Where the total energy-related carbon emissions C is expressed as a product of population P, GDP per capita, energy intensity E/GDP, and carbon intensity of energy use C/E. Although the master equation and Kaya's identity should be viewed as conceptual rather than mathematically rigorous, they can be used to suggest goals for technology and society. The technology-related terms M, E, and C offer the greatest hope for a transition to sustainable development, especially in the short term, and it is modifying these terms that is among the central tenets of industrial ecology.

3. INDUSTRIAL ECOLOGY AND ENERGY

Industrial ecology and energy use are inextricably linked because mobilization and utilization of materials in our present technological society are indispensable without the utilization of energy. However, global commons are threatened by environmental emissions from energy use. Simultaneous improvement in material and energy intensities of use is one of the desirable goals to preempt ecological damage across different scales and levels. In the following sections, the discussion focuses on energy considerations during different product life cycle stages.

3.1 Energy Considerations in Material Choice

Embodied energy, or "embedded energy," is a concept that includes the energy required to extract raw materials from nature, plus the energy utilized in the manufacturing activities. Inevitably, all products and goods have inherent embodied energy. The closer a material is to its natural state at the time of use, the lower is its embodied energy. Sand and gravel, for example, have lower embodied energy as compared to copper wire. It is necessary to include both renewable and nonrenewable sources of energy in an embodied energy analysis. The energy requirements for acquisition in usable form from virgin stocks of a number of common materials are shown in Fig. 7. From an industrial ecology perspective, a manufacturing sequence that uses both virgin and consumer-recycled material is usually less energy intensive than is primary production. The energy requirements for primary and secondary production of various metals are shown in Fig. 8. For one of the most commonly used industrial materials, aluminum, the energy requirement for secondary production of aluminum is approximately 90% less than the

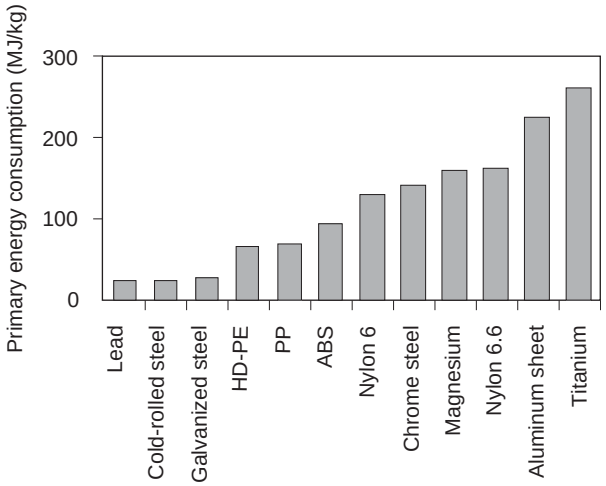


FIGURE 7 The primary energy consumption required to produce 1 kg of various materials. HD-PE, High-density polyethylene; PP, polypropylene; ABS, acrylonitrile butadiene styrene. Adapted from Schuckert *et al.* (1997).

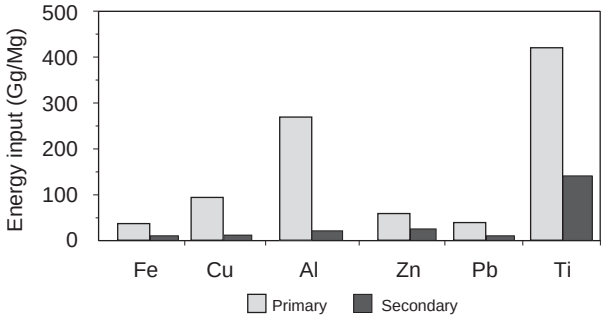


FIGURE 8 Energy requirements for primary and secondary production of metals. Data from Chapman and Roberts (1983).

requirement for primary production using virgin resource. Therefore, an efficient recycling operation can lead to potential savings in energy consumption and associated environmental damage.

3.2 Energy Considerations in Product Manufacture

Although the materials extraction and processing sectors have the highest energy intensity, these industries are suppliers to the intermediate processing industries, so there cannot be a plan to decrease industrial energy use solely by eliminating the extraction industries (which would not be possible in any case). The consumption of energy in selected manufacturing industries is shown in Fig. 9. Production petroleum and coal accounts for the largest energy use. Most of this energy use is attributable to

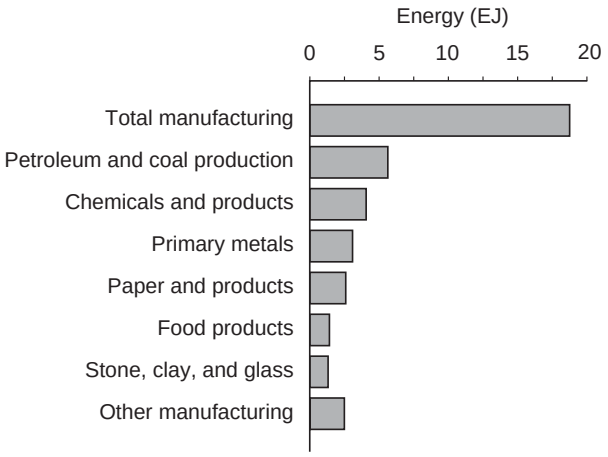


FIGURE 9 Consumption of energy in selected manufacturing industries. Data from the U.S. Department of Energy (1990).

the mining and processing of coal and the refining of petroleum. The trend toward desulfurization of crude oil and the production of high-octane gasoline without the use of metal-containing additives place ever-increasing energy demands on the refining operations. Refinery operations are generally subject to careful supervision and continuing engineering effort to improve efficiencies, but increased attention to cogeneration, heat exchange, and leak prevention is likely to offer opportunities for further improvements. Chemicals and chemical products rank among the industries depicted in Fig. 9 as well, although about a third of the amount shown represents petroleum and natural gas as feedstocks for products, rather than the fuel that is consumed to produce energy. Of the remaining two-thirds, a substantial amount is used in the generation or removal of process heat as a result of temperature differences between process streams and the heating and cooling streams. The production of compressed gases is another energy-intensive operation in this sector. Primary metals are the third industry listed in Fig. 9. Although the extraction of ore from the ground and its shipment are quite energy intensive, the bulk of the energy use is in crushing rock and recovering the target ore, and in generating the large amounts of process heat needed to extract metal from ore and to produce ingots and other purified products.

To maximize the efficiency with which water and energy are used in manufacturing and to minimize overall water and energy loss rates are major goals of pollution prevention. These actions are seen as complementary to those directed toward routine emissions to air, water, and soil, or to leaks or

accidents. Energy audits for the different process operations and for the overall facility are always helpful to indicate opportunities to reduce energy use. A particularly successful energy conservation program to date was initiated by the Louisiana Division of the Dow Chemical Company in 1982. Many of the improvements embodied techniques useful industry-wide, such as installing insulation on pipes carrying hot fluids, cleaning heat exchanger surfaces frequently to improve heat transfer efficiency, and employing point-of-use fluid heaters when storage or long pipelines create the potential for heat loss. The Dow energy contest results are summarized in Table I. It is clear that energy conservation, either through good housekeeping, improved technology, or process change or modification, is always beneficial.

3.3 Energy Considerations during Product Use

Designing innovative products that provide maximum benefit and service to customers and that are also simultaneously environmentally responsible is a challenge and an opportunity for product designers. DfE guidelines, evolved through feedback from customer expectations and regulatory policies, offer a framework to identify, prioritize, and implement product design improvements.

On a life cycle basis, the "in-use" phase is often dominant in terms of energy use. For a modern jet engine, for example, an improvement of 0.1% in fuel-burning efficiency is a more important design change than would be difficult material choices, improved manufacturing efficiency, or end-of-life design, so far as the environment is concerned. Energy-efficient designs may sometimes involve new approaches; the result may not only lower cost operation but may also improve product positioning (from a sales standpoint), particularly in areas of the world that are energy poor.

TABLE I

Energy Conservation Projects at Dow Chemical Company, Louisiana Division

Project	1982	1984	1986	1988	1990	1992
Number	27	38	60	94	115	109
Average return on investment (%)	173	208	106	182	122	305

[a]Data from Nelson (1994).

3.4 Energy Considerations in Remanufacturing and Recycling

When products reach the end-of-life stage, they can revert into input materials if they are recovered and recycled. Recycling operations are usually viable if the quantities recovered are large enough and the discard stream is homogeneous and concentrated. Even with extensive recycling, however, a growing economy still has a need for virgin resources, and processing of any kind requires energy.

In remanufacturing, products are either refurbished or reconditioned to the same quality control as new products, and are subsequently returned to market. Remanufacturing is more energy efficient as compared to recycling, because no new materials have to be processed. Remanufacutring also saves on landfill space and costs. From a DfX perspective, design for disassembly is important in remanufacturing because modular design of products will facilitate disassembly with less effort and time. In the United States, a remanufacturing tax credit has been included in the Clean Energy Incentives Act to encourage businesses. However, the lack of product "take back" laws to make producers responsible for their products at the end of the product life span, and the typical consumer's psychological preference for new rather than refurbished products, provide barriers against an extensive and well-integrated remanufacturing industry.

A perspective for product designers, remanufacturers, and recyclers is shown in Fig. 10, and details are provided in Table II. Different types of products are plotted on the basis of their product lifetime and technology cycle. There are four product categories,

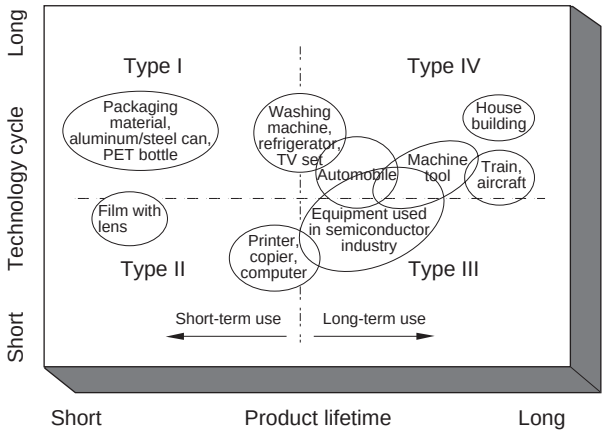


FIGURE 10 End-of-life plot for different product categories. PET, Polyethylene terephthalate. Reprinted from Masui (2002), with permission.

TABLE II

Guidelines for Product Designers and Recycling Technology Developers

Product category	End-of-life scenario	Product designer goal	Recycling technology developer goal
Type I	Material recovery	Ease separation of components for recycling high quality material	Develop separation technologies accounting for different physical properties of materials that cannot be sorted
Type II	Remanufacturing	Enhance reusability by using common parts and modular components in product family	Develop efficient cleaning and inspection technologies to reduce remanufacturing cost
Type III	Lengthen product life by upgrading	Extend product life by modular design of key devices that define value of product	Develop nondestructive techniques for removal of key components
Type IV	Lengthen product life by maintenance	Enhance ease of product disassembly for facilitating maintenance	Develop diagnostic technologies for maintenance

each of which offers a suite of environmental and nonenvironmental challenges to designers and recyclers. The complexity of the challenge increases in anticlockwise direction, moving from Type I products to Type IV in Fig. 10. Generally speaking, the energy consumption of these products also increases from Type I to Type IV. This characterization focuses on only two attributes and does not include functional and operational efficiency aspects, compatibility, number of parts, etc., but it does provide a framework for development of strategies to deal with products at their end of life.

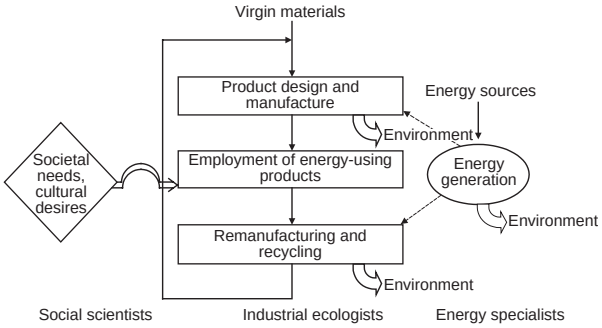


FIGURE 11 Satisfying the energy-related needs and desires of societies and cultures. The principal actors in each column of the diagram are shown at the bottom.

4. CONCLUSION

To an extent not generally appreciated, the study of industrial ecology is simultaneously the study of energy—its methods of generation, its employment in driving the engines of industry, and its use by individuals everywhere as they employ the technology of the modern world. Consider the diagram in Fig. 11. The driver for all energy use, shown on the left, is the contribution of societal needs and cultural desires related to energy. These define the type and number of energy-using products required to satisfy those needs and desires. The requisite energy is produced by the energy industry, indicated on the right. In the center are the design, use, and recycling of products. It is these steps that are the province of industrial ecology, which aims to satisfy the needs and desires while minimizing the use of energy in product design and manufacture, the use of energy in product operation, the use of energy-intensive virgin materials, and the magnitude of energy-related environmental impacts. The evolution of our society in the direction of increased sustainability will require not only that energy be widely available, but that it is used with maximum efficiency, and in such ways that its use produces minimal environmental impact. A close partnership between the energy industry and industrial ecologists will be required to make that vision a reality.

SEE ALSO THE FOLLOWING ARTICLES

Commercial Sector and Energy Use • Ecological Footprints and Energy • Ecological Risk Assessment Applied to Energy Development • Goods and Services: Energy Costs • Industrial Agriculture, Energy Flows in • Industrial Energy Efficiency •

Industrial Energy Use, Status and Trends • Industrial Symbiosis • Life Cycle Assessment and Energy Systems

Further Reading

Ayres, R. U. (1989). Industrial metabolism. *In* "Technology and Environment" (J. H. Ausubel and H. E. Sladovich, Eds.), pp. 23–49. National Academy Press, Washington, D.C.

Ayres, R. U., Norberg-Bohm, V., Prince, J., Stigliani, W. M., and Yanowitz, J. (1989). "Industrial Metabolism, the Environment and Application of Material Balance Principles for Selected Chemicals." International Institute for Applied Systems Analysis (IIASA) Report RR-89-11. IIASA, Laxenburg, Austria.

Ayres, R. U., and Simonis, U. E. (1994). "Industrial Metabolism, Restructuring for Sustainable Development." United Nations Press, Tokyo.

Ayres, R. U., and Ayres, L. W. (1996). "Industrial Ecology: Towards Closing the Materials Cycle." Edward Elgar Publishers, Cheltenham, United Kingdom.

Chapman, P. F., and Roberts, F. (1983). "Metal Resources and Energy." Butterworths, London.

Erkman, S. (2002). The recent history of industrial ecology. *In* "Handbook of Industrial Ecology" (R. U. Ayres and L. W. Ayres, Eds.), pp. 27–35. Edward Elgar, Northampton, Massachusetts.

Frosch, R. A., and Gallopoulos, N. E. (1989). Strategies for manufacturing. *Sci. Am.* **261**(3), 94–102.

Graedel, T. E. (1996). "Streamlined Life-Cycle Assessment." Prentice Hall, Upper Saddle River, New Jersey.

Graedel, T. E. (2000). Evolution of industrial ecology. *Environ. Sci. Technol.* **34**(1), 28A–31A.

Graedel, T. E., and Allenby, B. R. (2002). "Industrial Ecology," 2nd Ed. Prentice Hall, Upper Saddle River, New Jersey.

Graedel, T. E., van Beers, D., Bertram, M. B., Fuse, K., Gordon, R. B., Gritsinin, A., Kapur, A., Klee, R. J., Lifset, R. L., Memon, L. A., Rechberger, H., Spatari, S., and Vexler, D. (2004). The multilevel cycle of anthropogenic copper. Accepted for publication, *Environ. Sci. Technol.*

Grübler, A., and Nakićenović, N. (1996). Decarbonizing the global energy system. *Technol. Forecast. Social Change* **23,** 97–110.

Kaya, Y. (1990). "Impact of Carbon Dioxide Emission Control on GNP Growth: Interpretation of Proposed Scenarios." Paper presented to the International Panel on Climate Change (IPCC) Energy and Industry Subgroup, Response Strategies Working Group. IPCC, Paris, France [mimeo].

Kneese, A. V., Ayres, R. U., and d'Arge, R. C. (1970). "Economics and the Environment." Resources for the Future, Washington, D.C.

Lifset, R. L., and Graedel, T. E. (2002). Industrial ecology: Goals and definitions *In* "Handbook of Industrial Ecology" (R. U. Ayres and L. W. Ayres, Eds.), pp. 3–15. Edward Elgar, Northampton, Massachusetts.

Masui, K. (2002). "Life Cycle Strategies." Available on the Web site of the Agency of Industrial Science and Technology, Mechanical Engineering Laboratory at http://www.mel.go.jp (accessed on 11/16/02).

Nakićenović, N. (1997). Decarbonization as long-term energy strategy. *In* "Environment, Energy and Economy—Strategies for Sustainability" (Y. Kaya and K. Yokobori, Eds). United Nations University Press, Tokyo.

Nelson, K. E. (1994). Finding and implementing projects that reduce waste. *In* "Industrial Ecology and Global Change" (R. Socolow, C. Andrews, F. Berkhot, and V. Thomas, Eds.), pp. 371–382. Cambridge University Press, Cambridge, UK.

Schuckert, M., Beddies, H., Florin, H., Gediga, J., and Eyerey, P. (1997). Quality requirements for LCA of total automobiles and its effects on inventory analysis. *In* "Proceedings of the Third International Conference on Ecomaterials," pp. 325–329. Society of Non-Traditional Technology, Tokyo.

Society of Environmental Toxicology and Chemistry (SETAC). (1993). "Guidelines for Life-Cycle Assessment: A Code of Practice." SETAC, Pensacola.

United Nations Environment Program (UNEP). (2002). "The Industrial Symbiosis in Kalundborg, Denmark." Available on the United Nations Web site at http://www.uneptie.org (accessed on 11/16/02).

U.S. Department of Energy (DOE). (1990). "Manufacturing Energy Consumption Survey: Changes in Energy Efficiency 1980–1985." DOA/EIA-05169(85). Energy Information Administration, U.S. DOE, Washington, D.C.

U.S. Department of Energy (DOE). (2001). "Transportation Energy Data Book: Edition 21." Oak Ridge National Laboratory, U.S. DOE, Oak Ridge, Tennessee.

Wernick, I. K. (1996). Consuming materials: The American way. *Technol. Forecast. Social Change* **23,** 111–122.

Industrial Energy Efficiency

WOLFGANG EICHHAMMER
Fraunhofer Institute for Systems and Innovation Research
Karlsruhe, Germany

1. **Definitions of Industrial Energy Efficiency**
2. **Industrial Energy Efficiency Trends**
3. **Energy Efficiency Benchmarking**
4. **Potentials for Industrial Energy Efficiency Improvement**
5. **Obstacles to Industrial Energy Efficiency Improvement and Transaction Costs**
6. **Long-Term Potential for Industrial Energy Efficiency Improvement**

Glossary

barriers/obstacles to energy efficiency Factor that explains why energy efficiency measures are not carried out in the industrial sector despite the fact that they are economic according to company criteria.
energy efficiency benchmarking Comparison of the energetic performance of a company with the performance of other companies working in the same field.
energy efficiency improvement Decreasing the use of energy per unit service or activity without substantially affecting the level of these services.
energy intensity The amount of energy used per unit of activity or service.
exergy Ability of an energy form to perform work. Electricity has 100% exergy, while low-temperature heat has only a few percentage points.
transaction costs Costs occurring to change from one situation to another, such as from the current state to a more energy efficient state.

The industrial sector still represents in many countries a large fraction of energy consumption of typically 30 to 40% of the overall final energy consumption of a country. Although this sector has made continuous progress in energy efficiency over time (the article explains how to measure energy efficiency in the industrial sector), many examples from daily practice show that the economic energy efficiency potential in the industrial sector is far from being exhausted. Energy efficiency has many facets. Section 1 of this article introduces various concepts for energy efficiency in the industrial sector and explains why misunderstandings might arise from the nonexplicit use of different concepts. Section 2 uses the macroeconomic efficiency concept to describe the development of industrial energy efficiency, in particular the decoupling of economic growth and energy consumption, but also the pitfalls of this type of energy efficiency indicators. Section 3 shows how, at the microlevel, companies can use the benchmarking concept to improve their energy efficiency. Section 4 deals with the potentials for energy efficiency in the industrial sector (at a time horizon of 10 to 20 years) and how to realize them, while Section 5 deals with obstacles to industrial energy efficiency and transaction costs. Section 6 takes a look beyond the next two decades and points to the role of industrial energy efficiency research and development.

1. DEFINITIONS OF INDUSTRIAL ENERGY EFFICIENCY

When energy efficiency improvements are discussed for the industrial sector, quite often very different concepts are used for defining energy efficiency, which convey different messages that can even be contradicting. This is illustrated by the following tale.

In a public hearing in the year 2015, the (hypothetical) president of the European Commission was asked how much the energy efficiency of the iron and steel industry had improved between 2005 and 2015 in the framework of voluntary agreements on the reduction of CO_2 emissions. Naturally he did not know, but he had an expert to ask. The expert, after a few seconds of inquiry, came up with the following answer: "If you measure energy efficiency in the iron/steel industry in terms of energy consumed per value added, then efficiency has improved by 10%; if you measure energy efficiency in terms of energy consumed per ton of crude steel, then energy efficiency

 © 2004 Elsevier Inc. All rights reserved.

has got worse by 10%." The president was puzzled but decided to communicate only the first half of the sentence to the press.

The following analysis tries to provide an explanation to the contradictory answer of the expert to the hypothetical president of the European Commission. The figures used in the example are fictitious and chosen in such a way that the example is particularly striking: in the case of one indicator, the efficiency improves (monetary units); in the case of another, it deteriorates (physical units). The development of real indicators appears somewhat less spectacular because both types of indicators improve, but to a different degree (the energy efficiency improves more when measured in monetary units). A detailed analysis of the European iron/steel industry between 2005 and 2015 might show that the production of crude steel remained fairly stable as well as the energy consumed to produce 1 ton of crude steel. But what actually happened was that the iron/steel industry, in order to secure its future, discovered a new field of products (for example, the production of mobile phones), the production of which added, let's say, about 10% to the energy consumed for the production of the crude steel alone. Calculating the ratio between energy consumption and the constant crude steel production shows an increase of 10% of the specific energy consumption. At the same time, the value added created in the iron/steel industry increased due to the new activity by 20%. The energy intensity that links the 10% increased energy consumption to the 20% increased value added shows a decrease of roughly 10%.

This still does not answer the question whether the iron/steel industry has become more or less energy efficient. The contradiction is to be found in the different use of the term "energy efficiency" in the two parts of the sentence that the expert communicated to the president. When the term "energy efficiency" is used in conjunction with value added, it includes all kinds of structural changes within the branch considered—that is, energy efficiency is interpreted in a more economic sense: the structure is becoming more efficient; less energy is necessary to produce one unit of value. The correct answer to the president after a more refined analysis would have been: The energy efficiency of the core business has neither increased nor decreased. It has remained stable. The observed increase or decrease in "efficiency" has come about through the development of a new business with low energy-intensity (i.e., the sector has a different delimitation, which is not necessarily fully comparable with the previous state when looking at the overall branch). A further decomposition of the sector using physical units for the old crude steel production and value added for the new mobile phone business would reveal the correct answer that energy efficiency has not changed (assuming that the energy intensity for the production of the mobile phones was also fairly stable over time).

This example illustrates two important issues. First of all, when talking about improvements in industrial energy efficiency, it is important to mention which kind of energy efficiency definition is used (see Table I) and how it is operationalized in the form of energy efficiency indicators. The efficiency definitions in Table I start from very technical efficiency definitions at the level of processes: for example, efficiency in terms of energy output related to energy inputs in the case of a steam boiler. A second way would be by taking into account the different value or possibility to carry out work

TABLE I

Definitions for Industrial Energy Efficiency

Type of efficiency	Basic concept
First law efficiency	The energy transfer of a desired kind by a device or system divided by the energy supplied to the device or system to operate it.
Second law efficiency	The theoretical minimum available work needed for a given task divided by the actual available work used to perform the task.
Microeconomic I: Pareto efficiency	There are no feasible distributions of utility in which everyone is at least as well off, and at least one person is better off, than under the current distribution.
Microeconomic II: Cost-effectiveness	The rate of return on the investment equals or exceeds the hurdle rate of the investor.
Macroeconomic Energy efficiency I	The total consumption of energy over a relevant economic unit divided by the economic output of that unit.
Macroeconomic Energy efficiency II	The total consumption of energy over a relevant social unit divided by the current level of social welfare or level of sustainable income.
Institutional efficiency	Institutions are structured to minimize transaction costs.
Pareto efficiency with priced externalities	Standard microeconomic efficiency with the addition of monetary values assigned to externalities.
Distributional efficiency	Distributional objectives defined on moral and social grounds; prices set to achieve objectives cost-effectively.

(exergy) of different energy forms. For example, the exergetic efficiency of the same boiler would be much lower than its energetic efficiency. This can be understood in the following way: if the boiler could convert money coins, and would convert three Euro coins to three cent coins, the energetic efficiency would say three coins in, three coins out, so the efficiency is 100%, while the exergetic efficiency would consider what is written on the coins and would say three Euro in, three cents out (i.e., the efficiency is 3%). The third definition of efficiency is an economic definition at the microlevel stating that the world is always organized in such a way that an economic optimum at the microlevel is achieved. Other definitions occur at the level of macroeconomic units (e.g., whole countries) and relate energy inputs to the macroeconomic output of the units. The definitions continue to more general concepts, including externalities or institutional inefficiencies.

The second lesson is that in communicating results on energy efficiency improvements care has to be taken to avoid confusing messages by mixing up many different concepts. If one leaves the expert arena and goes to the political arena, where the indicators and concepts are to be applied, confusion is rapidly generated as the differences between the concepts are not recognized.

2. INDUSTRIAL ENERGY EFFICIENCY TRENDS

The objective of energy efficiency improvements in the industrial sector is the decoupling of the energy consumption and the corresponding environmental burdens (e.g., in the form of CO_2 emissions) from the economic growth of the sector. Figure 1 shows that for the European Union this aim was reached in the 1990s: while the industrial value added increased by 26% in that period of time, energy consumption remained stable at the level of 1990. CO_2 emissions from industrial activities decreased even more strongly. In other regions of the world, such as in developing countries or the United States, this decoupling is still far from being achieved. Also, electricity use in the industrial sector is still growing, even in the European Union, at a level close to the growth of the value added, but in some countries decoupling effects for electricity due to saturation start to become visible.

The most commonly indicator used to measure progress of energy efficiency in the industrial sector is the energy consumption per value added. Nevertheless, this indicator has the disadvantage that it does not separate changes in energy intensities due to changes in the industrial sector structure from changes in energy efficiency in a more technical sense. Figure 2 illustrates this with the example of the countries forming the European Union. Not only the energy intensity decrease during the 1990s differs widely between the EU countries, but also the level of industrial energy intensity. This is mainly due to differences in the industrial structure of the countries (i.e., the respective share of more or less energy-intensive branches in total industrial production). These differences should be taken into account when comparisons of industrial energy intensities among countries are carried out. If the energy intensities are adjusted to an average European structure, the differences are considerably less pronounced than in the case of actual intensity (compare the dark arrow in Fig. 2 with the light gray arrow).

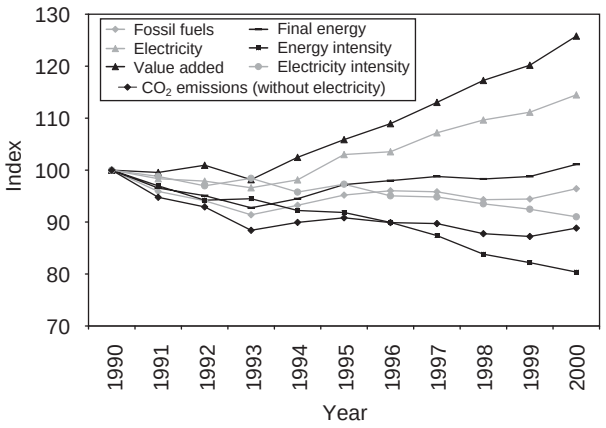


FIGURE 1 The decoupling of economic growth, energy consumption, and environmental burden: energy consumption, value added, energy intensity, and CO_2 emissions in EU industry, 1990–2000.

Why should one wish to separate different factors that influence the energy intensities? At the end it might not matter whether a reduction in energy consumption comes from changes in industrial structures or from energy efficiency measures. But, in fact, the point is that one might wish to influence energy efficient technologies by an adequate policy, while industrial structures might be dictated by considerations outside the energy field. Separation therefore allows to more adequately monitor factors of influence that one really wants to target by policy action.

As shown in Fig. 3 with the example of the European Union, the energy intensity of industry in a country or region tends to decrease more when the industrial value added increases strongly and to

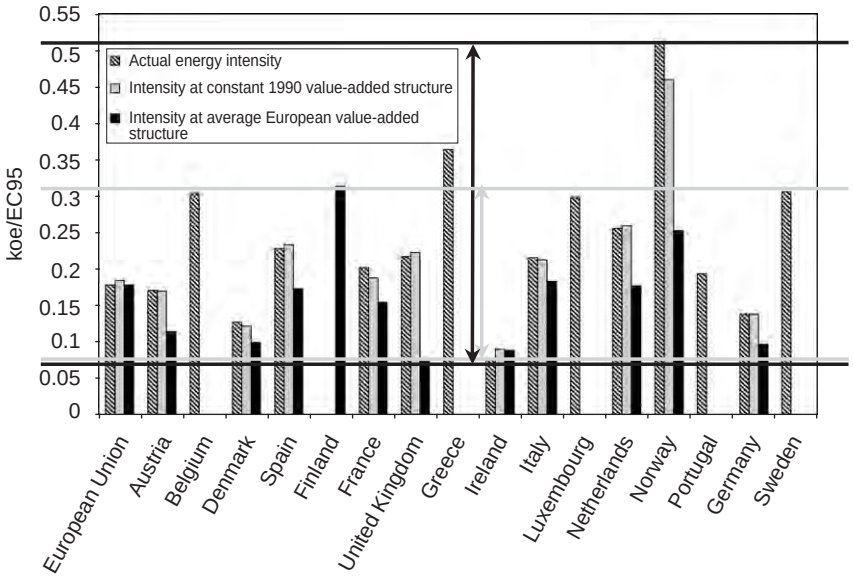


FIGURE 2 Manufacturing energy intensities in the EU countries 1999: actual, at constant 1990 value-added structure, at average European value-added structure.

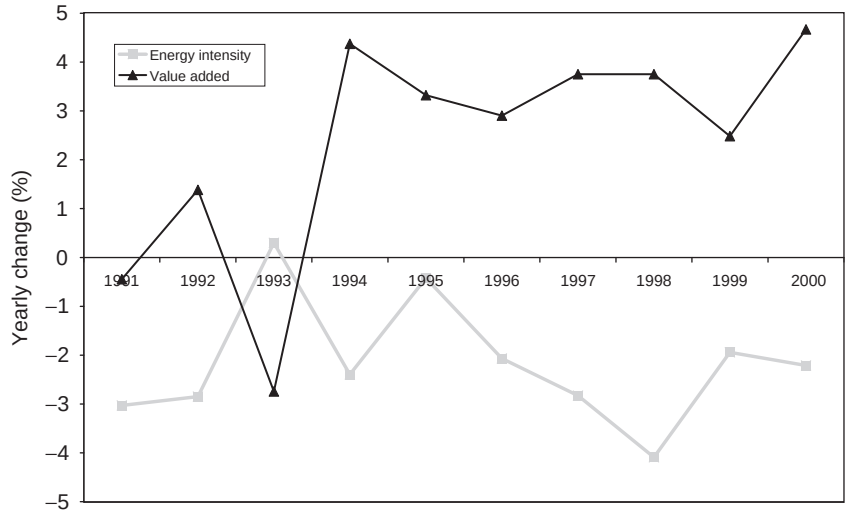


FIGURE 3 Annual growth rates of value added and energy intensity in EU industry, 1990–2000: The magnitude of the energy intensity variation is highly correlated to the industrial growth (the business cycles).

decrease less (or even increase) in the reverse situation (influence of the business cycle). The explanation is that there is one part of energy consumption that is rather independent from the production level (e.g., heating and lighting of premises) and another part that is more or less proportional to the production level (e.g., machinery, furnaces). In case of a production decrease, only the second part of consumption decreases, whereas the first one remains constant. As a result, the energy consumed per unit of production tends to increase.

3. ENERGY EFFICIENCY BENCHMARKING

At the level of companies, it appears also to be very useful to compare the own energy efficiency performance with other companies in the same field by using benchmarks. Benchmarks are widely used in many different areas of the society as well as in industrial practice to improve performances through competition and comparison with others. Energy efficiency benchmarking consists in the establishment

of energy indicators that are related to the activity of an industrial company and allow in principle to compare the energetic performance of a single company with other companies in the same field of activity. Examples in practice have shown that differences in energy efficiency practice from simple to double are possible for the same type of activity (see Fig. 4).

Benchmarking, which can be found in many different areas outside the energy efficiency field, is an ongoing process of continuous improvement: finding out how the "best" companies meet these standards; setting targets for business activities according to the "best practice" that can be found; adapting and applying lessons learned from these approaches and ideas to meet and exceed the standards; identification of areas where improvement would make the greatest difference to the bottom line, to key areas of the business or to customer relationships, establishing what makes a difference in customers' perceptions between a run-of-the-mill and an excellent supplier. Why introduce energy efficiency benchmarking in a company? It significantly reduces waste, rework, and duplication. Benchmarking increases awareness of what you do and how well you are doing it. A better process understanding leads to a more effective management in both energy efficiency terms and in overall production efficiency terms. Benchmarking helps set credible energy efficiency targets and identifies what to change and why. It removes blinkers and attitudes.

Energy efficiency benchmarking is hampered in practice by a variety of different factors of influence such as production volume (see Fig. 5), use of production capacity over time, changes in the process mix of the company, differences in product type and quality, energy efficiency measures, non-production-elated energy use, and storage of energy carriers (solid and liquid fuels).

In a carefully designed benchmark process, it is possible to take into account a variety of these factors. Nevertheless, benchmarking appears as limited, especially when no detailed knowledge on energy consumption is available at the process level in the company. In this context, benchmarking can be seen as a prescreening step toward an energy audit. It gives a quick result and can be performed on a large number of companies. Benchmarking and energy audit are therefore complementary exercises. From an engineering point of view, very often critics is therefore manifested toward energy efficiency benchmarks by stating that "all companies are individuals and they cannot be compared with each other." On this basis, benchmarking is considered as useless. This argument can be debated, first of all, on more general grounds:

- We are well using benchmarks in many different occasions of daily life, starting from the grades received at school without taking care of the individuality of each person.
- We are using benchmarks in companies well in other fields, for example, to compare their economic performance without paying attention either to the individuality of companies.
- We underestimate the dynamics of improvement triggered by the comparison with others, which leads us to detailed questions about why we are different with respect to the energetic performance elsewhere observed and possible measures for improvement.

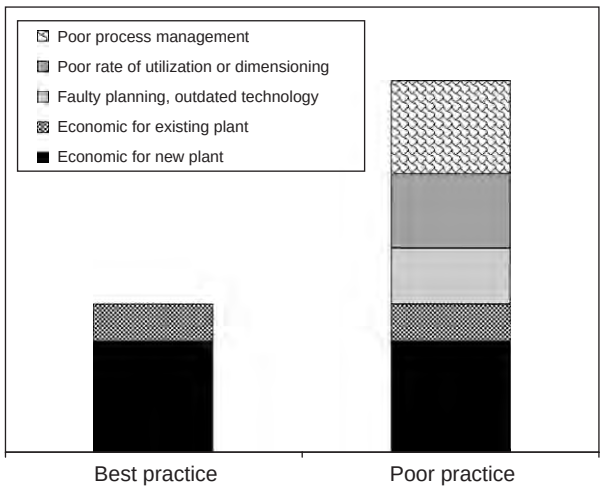


FIGURE 4 Why there are differences from simple to double in energy consumption for producing the same product.

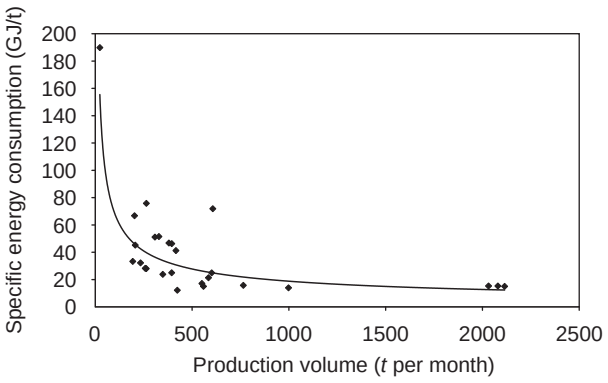


FIGURE 5 Scale effects of specific energy consumption (energy consumed per ton of product): yarn-producing companies in Thailand of different size.

In addition to these more general arguments in favor of benchmarking, there are increasingly also methodologies in development to design individually tailored benchmarks for each company. In fact, for this purpose the really observed energy consumption within the company is compared to a hypothetical company that has the same process mix but good practice processes. The advantage of this concept is that it is not necessary in particular to have detailed knowledge of the real energy consumption within the company at the process level. The disadvantage is, however, that extended knowledge has first to be provided to define and evaluate quantitatively process chains and basic processes for the comparison. This requires a lot of experience with benchmarking and a lot of knowledge accumulated with individual industrial processes and might be an obstacle at an early level of the introduction of a benchmarking process. Nevertheless, the so developed tailored benchmarks are possibly the most convincing ones from an engineering perspective.

4. POTENTIALS FOR INDUSTRIAL ENERGY EFFICIENCY IMPROVEMENT

Potentials for industrial energy efficiency improvement have been investigated on many occasions for different regions of the world. Results from investigations on European industries in the context of the contribution of energy efficiency to the reduction of global and local emissions provide the following wider conclusions, which are representative for many other studies and show the importance of energy efficiency for the general field of environment (see Table II):

- Cost-effective and currently available energy efficiency measures, which are technically proven, still offer the potential to make a significant contribution to reduction of pollution in the EU: an economic potential of 12 to 14% primary energy savings across the whole range of installations subject to the so-called Integrated Pollution Prevention and Control (IPPC) has been identified providing emissions saving of between 8 and 11%, mainly combustion related (CO_2, NO_x, SO_x).
- The calculate net cost benefit to the EU is 14 billion Euro at a total capital cost of 35 billion Euro for manufacturing industry.

TABLE II

Primary Energy Savings Calculated for European Union Industries in 2000 Covered by Integrated Pollution Prevention and Control IPPC (This Can Be Associated Roughly to Companies Entitled to Participate in Emission Trading Schemes)

Sector	IPPC coverage	Savings potential (PJ)	%
Food	31%	140–180	9–12
		(100–120)	(6–8)
Textiles	34%	70–80	12–14
Paper	98%	380–440	21–24
Chemicals	74%	460–790	16–27
		240–350	(8–12)
Nonmetallic minerals	86%	152	13
Steel	100%	387	15
Nonferrous metals	97%	147	21
Refineries	100%	770–980	25–32
Livestock	2%	2	10
Large combustion plants	98%	1830	6
Waste management	not quantified	40–90	

Note. It is assumed in this study that CHP replaces the average mix of thermal generation in the electricity supply industry. For some countries, it is more likely that CHP would replace gas turbines as these are the marginal technology. If this occurs, the primary energy savings would be much lower. The corresponding figures are shown in brackets in the table.

- Energy efficiency measures potentially save industry money, whereas end-of-pipe technologies to abate local pollution generally cost money.
- Low-cost measures such as energy management still offer significant scope for savings (around 8% of the total savings identified).
- Combined heat and power generation (CHP) offers a significant potential for primary energy savings ($\sim$36% of the identified savings) but involves relatively high capital costs. It is therefore likely that this potential does not represent a realistic short-term target for energy or emissions reductions.
- Energy savings from cost-effective improvements in cross-sector devices such as motors and drives, boilers, and compressed air plants are also available and are estimates to contribute 5% energy reduction.
- Process-specific energy savings represent a large proportion of the savings identified in many sectors (half of the total savings) and also benefit from other cost-savings factors such as throughput increases, quality, and labor savings.

This latter point illustrates that in many occasions energy efficiency measures do not only contribute to save energy in a company but while investigating these measures, often additional improvements for the production process are discovered at the same time, which improve considerably payback periods. Thus, energy efficiency, in addition to protecting the environment, contributes to render industrial processes more efficient and competitive (energy efficiency = production efficiency; see Fig. 6).

Energy efficiency improvements in the industrial sector may constitute large business opportunities. To illustrate this case, electricity saving measures in Western European industries on a time horizon 2015 are classified according to the following three categories:

- *Category I: Technical Add-on Improvements.* This category assembles measures that achieve electricity use reductions by modifying existing technologies, for example, with new or improved control devices or through improved process design.
- *Category II: Industrial Services.* This includes measures that help to improve the performance of existing technologies—that is, mainly by proper maintenance and operation. A major current trend is to outsource "housekeeping functions" that are not relevant for the core business of the company.
- *Category III: New Production Facilities.* This category represents the design and construction of new plants and production facilities (not necessarily new technologies).

Efficiency markets resulting from the implementation of efficiency measures that achieve electricity use reductions by modifying existing technologies (category I) or that help to improve the performance of existing technologies can be directly linked to the reduced energy costs. In total, the implementation of category I measures corresponds to a cumulated investment volume of 19,381 Mill € in 18 years for the Western European industrial sector (see Table III). The efficiency potential of category II measures results in an efficiency market volume of 7013 Mill. € for industry. All measures together result in a yearly efficiency market of 1466 Mill. € for industry in Western Europe.

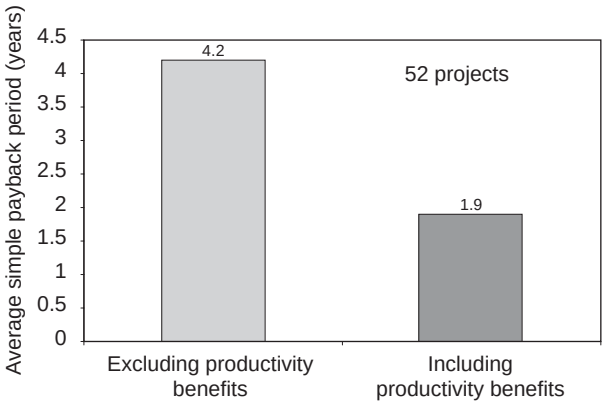


FIGURE 6 Energy efficiency = production efficiency.

Much higher market volumes arise from investments due to stock renewals and production growths of the analyzed technologies (category III measures). These investments bear a large potential for the uptake of energy efficient technologies. Especially the stock renewal of motors and motor-driven systems and the new construction of large office buildings bear large investment potentials.

5. OBSTACLES TO INDUSTRIAL ENERGY EFFICIENCY IMPROVEMENT AND TRANSACTION COSTS

There is a gap between the opportunities of the efficiency markets and the actual implementation of energy efficient measures in the industrial sector in practice. The reasons for this gap is commonly seen in a wide range of obstacles and market imperfections, which impede the widespread use of energy-efficient technologies. In theory, given all the benefits of energy efficiency at the microeconomic and macroeconomic levels, a perfect market would invest in, and allocate the rewards from, new energy-efficient technologies and strategies. But in practice, many obstacles and market imperfections prevent profitable energy efficiency from being fully realized:

- Improved energy efficiency is brought about by new technology, organizational changes, and minor changes in a known product. This implies that investors and energy users are able to get to know and understand the perceived benefits of the technical efficiency improvement. It also implies that investors and users have to be prepared to realize the improvement and to take time to absorb the new information and evaluate the innovation. But many users, especially small and medium-sized companies, do not have enough knowledge, technical skills, and

TABLE III

Efficiency Markets in Western Europe for Electricity Savings in the Industrial Sector

			Efficiency markets Western Europe			
			Cat. I[a]	Cat. II[b]	Cats. I and II	Markets per year
	Electricity use, 1997 (in TWh)	Economic savings	Mill. €	Mill. €	Mill. €	Mill. €/a
Industrial sector	966.8					
EAF steel	42.7	28%	187	152	339	
Primary aluminium	49.5	23%	454		454	
Chlorine	27.4	18%	1,017		1,017	
Air separation	13.3	6%	110		110	
Motors and drives	648.3	10%	11,663		11,663	
Compressed air	*76.0*	34%	706	3786	4492	
Pumps and fans	*236.3*	15%	3,515	3076	6,590	
Furnaces	61.8	28%	1,728		1,728	
Sum analysis	**842.9**		**19,381**	**7,013**	**26,394**	**1,466**

[a] Category I: Technical add-on improvements.
[b] Category II: Industrial services.

market information about possibilities for energy savings, particularly if energy converting off-sites are concerned (and not the production itself).

• The same energy consumers, even if they gain knowledge, often have trouble raising funds for energy-efficiency investments. Their capital may be limited, and additional credit may be expensive. Especially when interest rates are high, small firms tend to prefer to accept higher current costs and the risk of rising energy prices instead of taking a postponed energy credit.

• Large firms typically establish internal hurdle rates for energy efficiency investments that are higher than the cost of capital to the firm (e.g., by imposing payback requirements of 2 or 3 years for energy converting off-sites with technical life times of 15 years). This fact reflects the low priority that top managers place on increasing profits by raising energy productivity.

• In many companies, particularly small and medium-sized companies, all investments except for buildings and infrastructure are decided according to payback periods instead of internal interest rate calculations (see Table IV). If the lifetime of energy-saving investments is longer than that of existing production plants and machinery and if the payback period is expected to be even for both investments, entrepreneurs implicitly expect higher profits from energy-saving investments.

• Many firms (especially with the current shift toward lean firms) suffer from a shortage of trained technical staff because most personnel are busy maintaining production. Especially for small and medium-sized enterprises, installing new energy-efficient equipment is far more difficult than simply paying for.

• Insufficient maintenance of energy-converting systems and related control equipment causes substantial energy losses. Outsiders are not always welcome (external consultants, utilities), especially if proprietary processes are involved. Many companies cannot evaluate the risks connected with new equipment or control techniques in terms of their possible effects on product quality, process reliability, maintenance needs, or performance. Thus, firms are less likely to invest in new, commercially unproven technology.

• Cogeneration of heat and power has a considerable potential in a number of industrial branches, in particular the chemical industry, the pulp and paper industry, and the food sector. Yet the monopolistic structure of the electricity sector in many countries has led to high prices for maintenance and peak power and rather low buyback rates, as well as to dumping prices in the case of planning new cogeneration capacity. As a result, many auto producers restrict the capacity of the cogeneration plant to their minimum electricity and

TABLE IV

Internal Rate of Return in Percentage per Year for Industrial Energy Saving Measures (Continuous Energy Saving Is Assumed over the Whole Useful Life of the Plant)

	Useful lifetime of a plant (in years)							
Payback time (years)	3	4	5	6	7	10	12	15
2	24%	35%	41%	45%	47%	49%	49.5%	50%
3	0%	13%	20%	25%	27%	31%	32%	33%
4		0%	8%	13%	17%	22%	23%	24%
5			0%	6%	10%	16%	17%	18.5%
6		Unprofitable		0%	4%	10.5%	12.5%	14.5%
8						4.5%	7%	9%

Profitable investment possibilities (gray shaded area) eliminated by a 3-year payback time requirement.

heat needs, although they may wish to produce more heat by cogeneration. This obstacle is diminishing with increasing liberalization of the electricity markets opening up new opportunities for cogeneration and energy service companies.

It becomes obvious that manufacturers of energy-converting technologies and off-sites can surmount most of these obstacles by offering energy services in the form of contracting. In essence, many obstacles that are faced by small and medium-sized companies are business opportunities for large technology supplying companies. These opportunities are supported by the trend that companies in the industrial sector try to outsource their capital-intensive off-sites such as boilers, generation of technical gases, cooling, and logistics.

In the climate change debate, it is controversial to what extent measures should be undertaken to limit CO_2 emissions. Among the contentious aspects are the costs of these measures and, more specifically, the question whether there is a potential to reduce energy consumption at negative net costs, (i.e., a net benefit). According to numerous technoeconomic studies (i.e., studies that base their economic evaluations on detailed technology data), the investment costs of such measures will be recovered by the energy cost savings and other benefits they generate—even without counting their climate benefits. These so-called no-regret potentials are estimated for the year 2020 at 10 to 20% of global emissions of greenhouse gases. However, studies on the importance of this potential also affirm that multiple obstacles related to market imperfections or to juridical and administrative barriers impede their exploitation.

Other analyses with a more conventional economic background reject the possibility that profitable energy-saving measures may exist that are not undertaken autonomously by market forces. The fact that these measures are not implemented is interpreted as a rational response to real but hidden costs, which are supposedly not included in the evaluations of the no-regret potentials. In this context, the studies often refer to transaction costs as well as prohibitively high policy costs for the elimination of obstacles.

6. LONG-TERM POTENTIAL FOR INDUSTRIAL ENERGY EFFICIENCY IMPROVEMENT

Often the technical potentials of efficient energy use have a time horizon of only 15 to 20 years. Except for a few individual technologies, very little is known about the specific energy demand that industrial plant and machinery may have in, say, 40 or 60 years' time. This generally short time horizon of the (known) energy efficiency potentials gives rise to differing and inconsistent perceptions of the possibilities for technical and research-policy action on the energy supply and demand sides and therefore causes preference to be given to research on and development and application of new technologies in the energy conversion sector while devoting little funds to industrial energy-efficiency research and development (see Fig. 7).

Seen from a purely technological, natural science viewpoint, one can visualize potentials for eliminating more than 80% of present-day primary energy consumption in the industrialized countries.

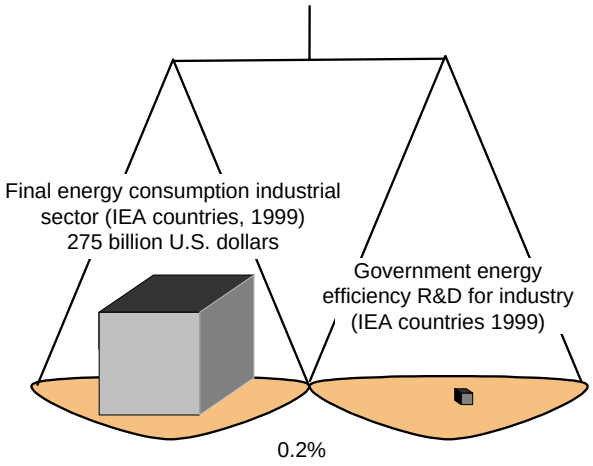


FIGURE 7 David against Goliath: research and devlopment spending for industrial energy efficiency in IEA countries versus the monetary value of industrial energy consumption, 1999.

Essentially, two basic principles can be adduced in support of this view: the better use of exergy of fossil fuels, heat, and electricity as well as the reduction of specific useful energy demand.

6.1 Better Use of Exergy of Fossil Fuels, Electricity, and Heat

The capability (or availability) of energy use (exergy) possessed by a given form of energy is often disregarded today when technical visions involving energy efficiency are made. The overall exergetic efficiency (i.e., the relation between useful energy and the availability of primary energy) is estimated for industrialized countries to be in the order of 10%. If this low efficiency is to be improved, it will be necessary to use energy with small losses of availability or only applying heat at small temperature differences, for example:

- Employ better use of available energy of burning fossil fuels via gas turbines, combustion engines, or fuel cells. In all these cases, electricity (of high exergetic value) is generated before the remaining energy in the form of heat is used.
- Try to avoid heat losses by using heat exchangers or vapor compressors. With increasing fuel prices in the long-term future, further technical progress, and additional economy-of-scale effects, heat recovery and energy cascading seem to have a bright future and an important role to improve energy efficiency in industry, the commercial sector, and the residential sector (as far as useful energy demand cannot be drastically reduced, discussed later).
- Try to increase temperature levels of waste heat in order to use the "upgraded" heat for useful purposes via heat pumps and heat transformers.

These technical possibilities open up a considerable energy-saving potential, in industry—for example, up to a theoretical maximum of 50%.

6.2 More Efficiency by Reduced Useful Energy Demand

The other major possibility for energy conservation is the reduction of specific useful energy demand, which is not a constant of nature but depends rather on the present level of technical know-how and energy price levels on the decision-making process and habits of consumption:

- The quality and quantity of thermal insulation used, for instance, reduce the heat demand in cooling equipment and production machinery. High-quality insulation material and vacuum solutions will be developed for heat transport, cooling equipment, and production machinery.
- Energy-intensive industrial processes may be substituted by less energy intensive processes or process chains (thin strip casting or even spraying of metals, of steel and aluminum in particular, introduction of the carbon technology in pig iron production, autothermal reforming in the production of ammonia and methanol, selective steam cradling in petrochemistry, fluidized bed cement production, direct route from salt to monomers of vinylchloride, etc.).
- Catalysts, enzymes biotechnology, and control techniques will render possible process substitutions and improvements in energy-intensive branches of industry to a very remarkable extent. High-energy demand of activation of chemical reactions, high pressure, and temperatures of processes may be rendered unnecessary by new catalysts or biotechnological processes. Traditional low- to medium-temperature processes may be substituted by biotechnological processes (e.g., dying of textiles or separation of intermediate products in the chemical or food industry).
- New applications of separation processes such as membrane processes, crystallization, extraction and absorption will only use a small percentage of the useful energy that is needed today in thermal separation processes in the basic products industry and the food industry.
- New materials (special plastics, for example, and foamed nonferrous metals) will reduce the

TABLE V

Long-Term Options to Improve Industrial Energy Efficiency by Levels of up to 80% beyond the Next 20 to 30 Years

Improve the use of the availability of energy (exergetic efficiency is only about 10% today) by	Reduce specific useful energy (it is not a constant of nature) by
Burning fuels: first via gas turbines, combustion engines, and fuel cells before heat is used	Insulation and air tightness
Using heat exchangers, vapor compressors, heat transformers, and heat pumps	Substitution of energy-intensive processes (catalysts, enzymes, membranes)
	New materials, lasers, inductive electric processes
	Recycling of energy-intensive materials
	Resubstitution by wood, natural fibres, new materials
	New concepts of vehicles and transportation systems
Theoretical energy saving potential: >50%	Theoretical energy saving potential: >60%
Total theoretical saving potential: >80%	

weight of moving parts and vehicles and, hence, reduce specific energy demand.

- Increasing recycling of energy-intensive materials (steel, aluminum, copper and other nonferrous metals, plastics, glass, paper/cardboard) can be anticipated, due to the necessity to reduce solid wastes that do not get societal acceptance for additional land fills and incineration plants.
- Substitution of energy-intensive materials by natural resources such as wood, natural fibers, and natural raw materials for chemicals is also a technical option, which will gain more attention in the long term because of its sustainability aspects.

The avoidance of heat and power losses, process substitution, and the closure of energy-intensive material cycles might have a theoretical technical energy-saving potential of at least 60% in industrialized countries. This estimate takes full account of the expectation that specific consumption of energy-intensive materials could be cut back by improved material properties and design changes.

Last, the useful service life of products made from energy-intensive materials could be lengthened by better corrosion protection and design for easier maintenance. This would reduce the cycling rate of energy-intensive materials, as well as the specific energy consumption of the respective energy services.

Adding together the above mentioned potentials, the theoretical long-term potential of efficient energy use undoubtedly lies above 80%. However, determined energy efficiency research and environmental protection policies are called for to maintain a high rate of efficiency increase over such a long period (see Table IV and V).

SEE ALSO THE FOLLOWING ARTICLES

Cogeneration • Commercial Sector and Energy Use • Discount Rates and Energy Efficiency Gap • Economics of Energy Efficiency • Energy Efficiency, Taxonomic Overview • Exergy Analysis of Energy Systems • Goods and Services: Energy Costs • Industrial Agriculture, Energy Flows in • Industrial Ecology • Industrial Energy Use, Status and Trends • Industrial Symbiosis

Further Reading

Bosseboeuf, D., Lapillonne, B., Eichhammer, W., and Landwehr, M. (1999). "Energy Efficiency Indicators: The European Experience." ADEME, Paris.

Golove, W. (1994). "A Preliminary Taxonomy of Efficiency Notions. International Comparisons of Energy Efficiency." Workshop Proceedings, Lawrence Berkeley Laboratory, Berkeley, CA.

Jochem, E., Mannsbart, W., Radgen, P., Schmid, Ch., Criqui, P., Worrell, E., and Price, L. (2001). "World-wide Electricity Demand and Regional Electricity Efficiency Markets with Particular Technological Focus on Western Europe." Research Report FhG-ISI, Karlsruhe.

Lapillonne, B., Schlomann, B., and Eichhammer, W. (2002). "Energy Efficiency in the European Union 1990–2000. SAVE-ODYSSEE." Project on Energy Efficiency Indicators (www.odyssee-indicators.org).

Ostertag, K. (2002). "'No-Regret' Potentials in Energy Conservation: An Analysis of Their Relevance, Size and Determinants." Physica Verlag, Heidelberg.

Industrial Energy Use, Status and Trends

ERNST WORRELL
Lawrence Berkeley National Laboratory
Berkeley, California, United States

1. **Introduction**
2. **Global Manufacturing Energy Use**
3. **Energy Services and Energy Efficiency**
4. **Energy-Intensive Industries**
5. **Cross-Cutting Energy Services**
6. **Energy Intensity Trends**
7. **Potential for Energy Efficiency Improvement**

Glossary

cross-cutting Energy technologies found throughout the industrial sectors and processes.
energy efficiency improvement Decreasing the use of energy per unit activity or service without substantially affecting the level of this activity or service.
energy intensity The amount of energy used per unit of activity or service.
energy-intensive industries Industries for which energy costs represent a large part (e.g., more than 10%) of total production costs.
energy service The activity or service supplied by the use of energy (e.g., lighting a room).
specific energy consumption The amount of energy used per unit of activity, expressed in physical terms (e.g., energy use per ton of steel).

Industry is a large energy user in nearly all countries. Approximately half of all industrial energy use is used in specific processes in the energy-intensive industries. On the other hand, various general energy conversion technologies and end uses can also be distinguished (e.g., steam production, motive power, lighting). Opportunities and potentials exist for energy savings through energy efficiency improvement in all sectors and countries. Technology development, as well as policies aimed at dissemination and implementation of these technologies, can help to realize the potential benefits. Technologies do not now, nor will they in the foreseeable future, provide a limitation on continuing energy efficiency improvements.

1. INTRODUCTION

Industrial production is the backbone of economic output in nearly all countries. Over the past decades, manufacturing industrial production has been growing in most economies. Industrial energy use can be broken down into that of the energy-intensive industries (e.g., primary metals, pulp and paper, primary chemicals, oil refining, building materials) and the non-energy-intensive industries (e.g., electronics, food). Energy use in the industrial sector is dominated by the production of a few major energy-intensive commodities such as steel, paper, cement, and chemicals. In any given country or region, production of these basic commodities follows the general development of the overall economy. Rapidly industrializing countries will have higher demands for infrastructure materials, and more mature markets will have declining or stable consumption levels. The regional differences in consumption patterns (expressed as consumption per capita) will fuel further growth of consumption in developing countries. In these "heavy" industries, energy is a very important production cost factor in addition to labor costs and raw material costs, driving a change toward higher energy efficiency (Table I).

Markets in the industrialized countries show a shift toward more service-oriented activities and, hence, non-energy-intensive industries. Still, energy-intensive industries will remain the largest energy consumers during the coming decades. Because of the great difference in energy intensity between energy-intensive industries and all others, changes in output

 © 2004 Elsevier Inc. All rights reserved.

TABLE I

Energy Intensities and Energy Purchases in Selected U.S. Industries

Sector	1973		1985		1994	
	Energy intensity (primary energy)	Energy costs (percentage share of production costs)	Energy intensity (primary energy)	Energy costs (percentage share of production costs)	Energy intensity (primary energy)	Energy costs (percentage share of production costs)
Iron and steel (GJ/ton)	30.5	7	27.8	11	25.4	8
Pulp and paper (GJ/ton)	43.1	6	42.7	6	32.8	6
Cement (GJ/ton)	7.3	40	5.2	36	5.4	33
Primary aluminum (MWh/ton)	N/A	14	17.6	19	16.2	13
Petroleum refining (GJ/ton)	6.2	4	4.3	3	4.5	3

Source. Lawrence Berkeley National Laboratory.

Note. Energy intensity is expressed in primary energy, where the efficiency of electricity generation is assumed to be 33%. Energy intensity of primary aluminum production is given in MWh (1000 kWh).

shares of these industries can have a major impact on total industrial energy use. Many commodities (e.g., food, steel) are traded globally, and regional differences in supply and demand will influence total industrial energy use. Production trends also depend on regional availability of resources (e.g., scrap) and capital. Manufacturing energy use will also depend on the energy efficiency with which the economic activities are done.

This article discusses energy use patterns in industry and then assesses trends in industrial energy use and energy intensities. This is followed by a discussion of the potential for energy efficiency improvement and the future trends in industrial energy use.

2. GLOBAL MANUFACTURING ENERGY USE

In 1995, manufacturing industry accounted for 41% (131 exajoules [EJ], where $EJ = 10^{18}$ J) of global energy use. Between 1971 and 1995, industrial energy use grew at a rate of 1.7% per year, slightly less than the world total energy demand growth of 2.1% per year. This growth rate has slowed during recent years and was virtually flat between 1990 and 1995, primarily because of declines in industrial output in the transitional economies in Eastern Europe and the former Soviet Union. Energy use in the industrial sector is dominated by the industrialized countries, which accounted for 43% of world industrial energy use in 1995. Industrial energy consumption in these countries increased at an average rate of 0.6% per year between 1971 and 1990, from 49 to 54 EJ. The share of industrial sector energy consumption within the industrialized countries declined from 40% in 1971 to 33% in 1995. The decline partly reflects the transition toward a less energy–intensive manufacturing base, a shift toward a more service-oriented economy, and the continued growth in transportation demand resulting in large part from the rising importance of personal mobility in passenger transport use.

The industrial sector dominates in the economies in transition, accounting for more than 50% of total primary energy demand, the result of the emphasis on materials production, a long-term policy promoted under years of central planning. Average annual growth in industrial energy use in this region was 2.0% between 1971 and 1990 (from 26 to 38 EJ) but dropped by an average of 7.3% per year between 1990 and 1995. Even so, in 1995 industry accounted for 51% of primary energy use in these countries.

In the Asian developing countries, industrial energy use grew rapidly between 1971 and 1995, with an annual average growth rate of 5.9%, jumping from 9 to 35 EJ. It also accounted for the greatest share of primary energy consumption (i.e.,

58% in 1995). The fastest growth in this sector was seen in China and in other rapidly developing Asian countries. Growth in other developing countries was slightly lower.

The nature and evolution of the industrial sector varies considerably among developing countries. Some economies that are experiencing continued expansion in energy–intensive industry, such as those in China and India, show relatively unchanging shares of industrial energy use. In other countries, such as Thailand and Mexico, the share and/or growth of the transportation sector dominate. Many smaller countries have remained primarily agrarian societies with modest manufacturing infrastructures.

Future trends in industrial energy use will be affected by many different challenges. Globalizing trade patterns will affect industrial competition as well as technology transfer. Global environmental challenges such as climate change will lead to an increased focus on energy efficiency and changing energy consumption patterns and fuel choice. Regional and local environmental problems will also drive changes in industrial energy use in both industrialized and developing countries.

3. ENERGY SERVICES AND ENERGY EFFICIENCY

Energy is used to provide a service (e.g., producing a ton of steel, lighting a specified area). These services are called energy services. Energy consumers (including industry) are not interested in using energy; rather, they are interested in supplying the energy service in the most economic way. Energy efficiency improvement entails the provision of these services using less energy and without substantially affecting the quality of the service provided.

Approximately half of all industrial energy use is for specific processes in the energy-intensive industries. On the other hand, various general energy conversion technologies and end uses can also be distinguished (e.g., steam production, motive power, lighting). Hence, energy use in manufacturing industry can be broken down into various uses to provide a variety of services. A common breakdown distinguishes energy use for processes (called process specific) and energy use for buildings, utilities, and boilers (called cross-cutting). Note that boilers provide steam and hot water for processes as well as for cross-cutting energy services, as do motors.

Because of the wide variety in industrial processes, the discussion in the next section is limited to four energy-intensive sectors that are responsible for a large share of industrial energy use: iron and steel, pulp and paper, cement, and chemicals. This is followed in the subsequent section by a discussion of cross-cutting energy uses (i.e., boilers and motors).

4. ENERGY-INTENSIVE INDUSTRIES

4.1 Iron and Steel Industry

The first record of the use of iron goes back to 2500 to 2000 BC, and the first deliberate production of iron began around 1300 BC. Small furnaces using charcoal were used. High-temperature processes started to be introduced in Germany around 1300 AD. The design of these furnaces is essentially the same of that of modern blast furnaces. The furnaces still used charcoal, and the first reported use of coke was in 1718 in the United Kingdom. The higher strength of coke allowed larger furnaces to be built with increasing energy efficiency. By 1790, coke ironmaking contributed to 90% of the British iron production. The development of the modern blast furnace after World War II resulted in an annual reduction of energy intensity of 3 to 4%/year due to the use of improved raw materials, ore agglomeration, larger blast furnaces, and higher air temperature. Today, the blast furnace is the main process to make iron and provides the largest raw material stream in steelmaking.

Steel is produced by reducing the carbon content in the iron to levels below 2%. This reduces the brittleness of the material and makes it easier to shape. The first steelmaking process was invented in 1855 by Bessemer. In the Bessemer converter, air was blown through the hot iron, a process that oxidizes the carbon. This principle is still followed in modern steelmaking processes. In the United States, the last Bessemer converter was retired during the 1960s. During the late 19th century, the Siemens–Martin or open hearth furnace (OHF) was invented. The OHF process uses preheated air to oxidize the carbon and melt the steel. This process is currently found only in developing countries and in Eastern Europe. The United States was one of the industrialized countries that phased out the OHF at a very late stage. During the 1980s, the dominant process became the basic oxygen furnace (BOF), which uses pure oxygen instead of air. The BOF process was developed in Austria during the 1950s. The productivity of this process is much higher, as is the energy efficiency. An

alternative process is the electric arc furnace (EAF). The EAF process is used mainly to melt scrap. Performance of EAFs has improved tremendously, starting to use fuel and oxygen in addition to electricity. It is expected that in the future, the BOF and EAF processes will follow similar developmental paths. Liquid steel is cast into ingot or slabs and is shaped in rolling mills to the final product. Although most energy use is concentrated in the iron- and steelmaking, reduced material losses and productivity gains in casting and shaping (e.g., continuous casting, thin slab casting) have contributed to dramatic increases in the energy efficiency of steelmaking.

Today, the global iron and steel industry produces approximately 840 million tons of steel. The industry is made up of integrated steel mills, which produce pig iron from raw materials (e.g., iron ore, coke) using a blast furnace and produce steel using a BOF, and secondary steel mills, which produce steel from scrap steel, pig iron, or direct reduced iron (DRI) using an EAF. The majority of steel produced is from integrated steel mills (62% in 2001), although the share of secondary steel mills (or "mini-mills") is increasing (35% in 2000).

Global energy consumption for steelmaking is estimated at 20 EJ. The worldwide average energy intensity of steelmaking is estimated at 24 gigajoules/ton (GJ/ton), although large variations occur between countries and plants. Today, the most energy-efficient process would use 19 GJ/ton for integrated steelmaking and 7 GJ/ton for making steel out of scrap. Analyses have shown that many technologies exist that could improve energy efficiency further. For example, in the United States, the potential for energy efficiency improvement is estimated at 18% using proven and cost-effective practices and technologies with a payback period of 3 years or less. New technologies that could considerably lower the energy intensity of steelmaking are under development. Smelt reduction in ironmaking would integrate the production of coke and agglomerated ore with that of ironmaking, leading to reductions in production costs and energy use. The development of direct casting techniques that abandon rolling would increase productivity while reducing energy use further. Combined, these technologies could reduce the energy intensity of primary steelmaking to 12.5 GJ/ton and could reduce that of secondary steelmaking to 3.5 GJ/ton, reductions of 34 and 50%, respectively. In the highly competitive and globalizing steel industry, manufacturers must continuously look for ways in which to lower their energy intensity and costs.

4.2 Pulp and Paper Industry

Paper consists of aligned cellulosic fibers. The fibers may be from wood or other crops or from recycled waste paper. Starting with wood fibers, the fibers need to be separated from the wood, a process that is done by pulping the wood. The separation can be done by chemicals, by heat, or by mechanical means. In the chemical pulping process, chemicals and hot water are used to separate the cellulosis from the ligno-cellulosis. The amount of pulp produced is about half the amount of wood used. Chemical pulping results in high-quality paper. In mechanical the wood is ground under pressure, separating the fibers from each other. In mechanical pulping, the ligno-cellulosis is not removed, resulting in a lower quality paper (e.g., paper used for newsprint) but also in a higher recovery ($\sim$90% of the used wood). In chemical pulping, a lot of steam is used to heat the water and concentrate the chemical by-products. Recovery of the by-products to be recycled in the process can actually produce sufficient steam for the whole paper mill. The mechanical process uses large quantities of electricity, whereas some processes can recover steam from the grinding process. Waste paper is pulped by mixing with water, after which ink is removed and the pulp is refined. Paper recycling reduces the energy needs of the pulping process. Waste paper use in the production of paper varies widely due to the different structures of the industry in different countries.

Although energy efficiency improvement options do exist in the pulping step, greater opportunities exist in the chemical recovery step. The most common pulping process in the United States is the Kraft pulping process. Black liquor is produced as a by-product. The chemicals are recovered in a recovery boiler, combusting the ligno-cellulosis. Because of the high water content, the recovery boiler is not very efficient, and the steam is used to generate electricity in a steam turbine and steam for the processes. Gasification of black liquor would allow the use of the generated gas at a higher level of efficiency. This would make a Kraft pulp mill an electricity exporter.

In papermaking, the pulp is diluted with water at about 1:100. This pulp is screened and refined. The solution with the refined fibers (or stock) is fed to the paper machine, where the water is removed. In the paper machine, the paper is formed into a sheet and water is removed by dispersing over a wire screen. At the end of the forming section, 80% of the water is removed. The rest of the water is removed in

the pressing and drying section. Although only a small amount of water is removed in the drying section, most energy is used in the drying section. Hence, energy efficiency opportunities try to reduce the water content by increasing the water removal by pressing. In a long nip press, the pressing area is enlarged. The larger pressing area results in extra water removal. New technologies aiming to increase the drying efficiency considerably are under development. One technology, impulse drying, uses a heated roll, pressing out most of the water in the sheet that may reduce the steam consumption of the paper machine by 60 to 80%.

The pulp and paper industry uses approximately 6 to 8 EJ globally. Because energy consumption and intensity depend on the amount of wood pulped, the type of pulp produced, and the paper grades produced, there is a great range of energy intensities among the industrialized countries of the world. In Europe, energy use for papermaking varied between 16 and 30 GJ/ton of paper during the early 1990s. The Netherlands used the least energy per ton of paper, largely because most of the pulp was imported. Countries such as Sweden and the United States have higher energy intensities due to the larger amount of pulp produced. Sweden and other net exporters of pulp also tend to show higher energy intensities. Energy intensity is also influenced by the efficiency of the processes used. Many studies have shown considerable potentials for energy efficiency improvement with current technologies such as heat recovery and improved pressing technologies.

4.3 Cement Industry

Cement is an inorganic, nonmetallic substance with hydraulic binding properties. Mixed with water, it forms a paste that hardens due to formation of hydrates. After hardening, the cement retains its strength. There are numerous cement types due to the use of different sources of calcium and different additives to regulate properties. The exact composition of cement determines its properties.

In 1999, global cement production was estimated to be 1600 million tons. Because of the importance of cement as a construction material and the geographic abundance of the main raw materials, cement is produced in virtually all countries. The widespread production is also due to the relatively low price and high density of cement, and this in turn limits ground transportation due to high transport costs.

The most common raw materials used for cement production are limestone, chalk, and clay. The collected raw materials are selected, crushed, and ground so that the resulting mixture has the desired fineness and chemical composition for delivery to the pyro-processing systems. The grinding process differs depending on the pyro-processing process used. The feed to the kiln is called "raw meal."

Clinker is produced by pyro-processing the raw meal. The raw meal is burned at high temperatures, first by calcination of the materials and then by clinkerization to produce clinker. Various kiln types have been used historically or are used around the world. Besides the rotary kiln, the vertical shaft kiln is used mainly in developing countries. In industrialized countries, the ground raw materials are processed predominantly in rotary kilns. In processing without precalcination, the decomposition (calcination) of $CaCO_3$ to CaO and CO_2 takes place in the kiln. The clinker is cooled. The cooling air serves as combustion air. The largest part of the energy contained in the clinker is returned to the kiln in this way.

Grinding of cement clinker together with additives to control the properties of the cement is done in ball mills, roller mills, or roller presses. Coarse material is separated in a classifier to be returned for additional grinding. Power consumption for grinding depends strongly on the fineness required for the final product and the use of additives.

Cement production is a highly energy-intensive process. Cement making consists of three major process steps: raw material preparation, clinker making in the kiln, and cement making. Raw material preparation and cement making are the main electricity-consuming processes, whereas the clinker kiln uses nearly all of the fuel in a typical cement plant. Clinker production is the most energy-intensive production step, responsible for approximately 70 to 80% of the total energy consumed. Raw material preparation and finish grinding are electricity-intensive production steps. Energy consumption by the cement industry is estimated at 6 to 7 EJ or 2% of global primary energy consumption.

The theoretical energy consumption to produce cement can be calculated based on the enthalpy of formation of 1 kg of Portland cement clinker (which is $\sim$1.76 MJ). In practice, energy consumption is higher. The kiln is the major energy user in the cement-making process. Energy use in the kiln depends basically on the moisture content of the raw meal. Most electricity is consumed in the grinding of the raw materials and finished cement. Power consumption for a rotary kiln is comparatively low.

4.4 Chemical Industry

The chemical industry produces many intermediate compounds that are used as the basis for many chemical products. The chemical industry produces more than 50,000 chemicals and formulations. For example, ethylene, one of the most important bulk chemicals from an energy point of view, is used to produce products varying from solvents to plastics. Also, many processes in the chemical industry produce different coproducts. Chemical industries consume fuels and electricity as energy and feedstock. This makes energy analysis of the chemical industry more complicated compared than that of other industries.

A small number of bulk chemicals are responsible for the largest part of the energy consumption in the chemical industry. These are the so-called basic chemicals that are used as building blocks for many chemicals down the production chain. The most important basic chemicals are the family of petrochemicals (ethylene, propylene, butadiene, and benzene) from the organic chemical industry as well as ammonia and chlorine/caustic soda from the inorganic chemical industry.

Ethylene and its derivatives are feedstocks for many plastics and resins as well as for fibers and detergents. Global ethylene production is estimated at more than 80 million tons and growing. The United States is the world's largest ethylene producer, accounting for less than 30% of the world capacity. Since 1974, ethylene production has grown by 3% annually, while propylene has grown by more than 4% annually. Propylene has grown more rapidly—5% per year—during the past decade or so.

Ethylene and other coproducts are produced through cracking of hydrocarbon feedstocks. In the presence of steam, hydrocarbons are cracked into a mixture of shorter unsaturated compounds. In the cracking process, hydrocarbon feedstocks are preheated to 650°C (using fuel gas and waste heat), mixed with steam, and cracked at a temperature of approximately 850°C. A series of separation steps produce fractions consisting of ethylene, propylene, a C_4 fraction (e.g., butadiene), and pyrolysis gasoline (containing benzene, toluene, and xylenes). The gas mixture is rapidly cooled to 400°C (or quenched) to stop the reaction, during which process high-pressure steam is produced. Injection of water further decreases the temperature to approximately 40 to 50°C, and a condensate that is rich in aromatics is formed. The liquid fraction is extracted, while the gaseous fraction is fed to a series of low-temperature, high-pressure distillation columns. Feedstocks used are ethane, LPG, naphtha, gas oils (GOs), and (sometimes) coal-derived feedstocks. Many of the installations used today can handle different (if not all) types of feedstock.

The single most energy-consuming step in the petrochemical industry is the steam cracking of hydrocarbon feedstocks into a mixture of shorter unsaturated compounds. Recent estimates of global energy consumption for the production of ethylene and coproducts are not available. Global energy use is estimated at more than 1 EJ. Energy consumption for ethylene production can be separated in feedstock and energy use. Feedstock energy use is generally equivalent to the heating value of the product, that is, approximately 42 GJ/ton (lower heating value [LHV]). Specific energy consumption for the energy use of the process varies depending on the feedstock, technology, age of plant, capacity, and operating conditions. In general, it varies between 14 (ethane feedstock) and 30 GJ/ton of ethylene (gas oil feedstock), whereas older plants can use more energy.

The production of ammonia, a key component in the manufacture of nitrogenous fertilizers, is a highly energy-intensive process. Roughly 80% of ammonia production is used as fertilizer feedstock. Ammonia is produced through the high-pressure synthesis of gases. Ammonia production in 1999 was estimated at 130 million tons. Growth is found mainly in developing countries, where the production expansion is also concentrated.

Ammonia is produced by the reaction of nitrogen and hydrogen, the so-called Haber–Bosch process. The main hydrogen production processes used in the United States are steam reforming of natural gas and partial oxidation of oil residues. Hydrogen is produced by reforming the hydrocarbon feedstock and producing synthesis gas containing a mixture of carbon monoxide and hydrogen. The carbon monoxide is then reacted with steam in the water–gas–shift reaction to produce carbon dioxide and hydrogen. The carbon dioxide is removed from the main gas stream. The carbon dioxide is recovered for urea production or exported as a coproduct or vented. The hydrogen then reacts with nitrogen in the final synthesis loop to form ammonia. The ammonia is often processed to different fertilizers, including urea and ammonium nitrate, whereas ammonia is also used as an input to make other chemicals.

Energy intensity of ammonia making depends on the feedstock used and the age of the process. Natural gas is the preferred feedstock and is used for more than 80% of global ammonia production. Coal is still used in countries with limited natural gas supplies (e.g., China). The use of natural gas provides not only

higher production efficiency but also lower capital costs. Part of the energy input is used as feedstock; that is, the hydrogen in natural gas ends up in the ammonia. This is generally equivalent to the thermodynamic mininum energy consumption for Haber–Bosch synthesis, corrected for losses. Hence, feedstock use is estimated at 19 to 21 GJ/ton of NH_3 (LHV). Energy use varies between 9 and 18 GJ/ton. The most efficient plants in the world consume approximately 28 GJ/ton of NH_3 (feedstock and energy use). In the United States, specific energy consumption is equal to approximately 37 GJ/ton (LHV). An analysis of energy intensities of ammonia manufacture in Europe showed that the average 1989 specific energy consumption (SEC) in the European Union was estimated at 35.5 GJ/ton (LHV).

Chlorine production capacity is estimated at 50 million tons at 650 sites. Approximately half of the chlorine is produced in Asia. The production of chlorine gas is an energy-intensive chemical process. In the process, a brine solution is converted into two coproducts, chlorine gas and sodium hydroxide (caustic soda), through electrolysis. The three main electrolysis cell types that are used to separate and produce the chlorine gas and caustic soda are the mercury flow, diaphragm, and ion-selective membrane. In the diaphragm and membrane cells, the caustic soda requires an additional step of concentrating the solution so that it can meet market specifications for most products.

Chlorine production is a main electricity-consuming process in the chemical industry, as is oxygen and nitrogen production. Energy use for chlorine production depends strongly on the cell type used for the electrolysis. Typically, power consumption varies between 2800 (membrane process) and 4300 kWh/ton (mercury process). In addition, steam may be used to concentrate the caustic soda in the membrane process. Membrane cells require the least amount of energy to operate. The membrane process is considered the state-of-the-art process. Countries in Asia and Europe have a higher share of membrane processes than do the United States and Canada. In the United States, chorine production uses 48 TWh, with an average intensity of 4380 kWh/ton.

5. CROSS-CUTTING ENERGY SERVICES

5.1 Steam Production and Use

Besides the energy-intensive industries, many smaller and less energy-intensive, or light, industries exist. Light industries can include food processing, metal engineering, and electronics industries. In light industries, energy is generally a small portion of the total production costs. There is a wide variety of processes used within these industries. In general, a large fraction of energy is used in space heating and cooling, motors (e.g., fans, compressed air), and boilers. Industrial boilers are used to produce steam or heat water for space and process heating and for the generation of mechanical power and electricity. In some cases, these boilers have a dual function such as the cogeneration of steam and electricity. The largest uses of industrial boilers by capacity are in paper, chemical, food production, and petroleum industry processes. Steam generated in the boiler may be used throughout a plant or site. Total installed boiler capacity (not for cogeneration) in the United States is estimated at nearly 880 million MW. Total energy consumption for boilers in the United States is estimated at 9.9 EJ.

A systems approach may substantially reduce the steam needs, reduce emissions of air pollutants and greenhouse gases, and reduce operating costs of the facility. A systems approach that assesses options throughout the steam system and incorporates a variety of measures and technologies can help to find low-cost options. Improved efficiency of steam use reduces steam needs and may reduce the capital layout for expansion, reducing emissions and permitting procedures at the same time. In specific cases, the steam boiler can be replaced nearly totally by a heat pump (or mechanical vapor recompression) to generate low-pressure steam. This replaces the fuel use for steam generation by electricity. Emission reductions will depend on the type and efficiency of power generation.

Another option to reduce energy use for the steam system is cogeneration of heat and power. Based on gas turbine technology, this is a way in which to substantially reduce the primary energy needs for steam making. Low- and medium-pressure steam can be generated in a waste heat boiler using the flue gases of a gas turbine. Classic cogeneration systems are based on the use of a steam boiler and a back pressure turbine. These systems have relatively low efficiency compared with a gas turbine system. Steam turbine systems generally have a power-to-heat ratio between 0.15 (40 kWh/GJ) and 0.23 (60 kWh/GJ). The power-to-heat ratio depends on the specific energy balance of the plant as well as on energy costs. A cogeneration plant is most often optimized to the steam load of the plant, exporting excess electricity to the grid. The costs of installing a steam

turbine system depend strongly on the capacity of the installation. Gas turbine-based cogeneration plants are relatively cheap. In many countries (e.g., the Netherlands, Scandinavia), gas turbine cogeneration systems are standard in industry. The power-to-heat ratio is generally higher than that for steam turbine systems. Aero-derivative gas turbines may have a power-to-heat ratio of 70 to 80 kWh/GJ. Aeroderivative turbines are available at low capacities, but specific costs of gas turbines decrease sharply with larger capacities.

5.2 Motors and Motor Systems

Motor systems are used throughout any industrial operation. A motor system generally consists of an appliance, the drive train, and the motor. Motor systems consume more than 60% of industrial electricity use in the United States and nearly 70% in Europe. The share of motor electricity use varies by sector. For example, in the cement industry, as much as 90% of electricity is used in motors to grind raw materials and final products and to drive the kiln. Motors themselves are efficient, and well-designed and -maintained motors have conversion efficiencies of more than 90%. However, older and inefficient motors may have much lower efficiencies, and the efficiency losses are even higher in the total motor system.

Motors range from less than 1 kW to a few megawatts in size. The largest motors are found in kiln and mill drives in the cement industry or as compressors in the chemical industry. Smaller motors may be found in many locations and are often installed as a whole system (e.g., a packaged pump). Small motors account for more than half of the motors installed in U.S. industry but represent a much smaller share of electricity use because these motors are generally used for fewer hours, whereas the large motors may run continuously. The most important uses of motors are for pumps, compressed air, and fans. In the United States, nearly 25% of industrial motor electricity is used for pumping systems, 14% for fans, 16% for compressed air systems, and 23% for material processing. Other uses (e.g., material handling, refrigeration) represent more than 23% of motor electricity use.

There are many opportunities to improve efficiency in motor systems. Overall, the cost-effective potential for motor system efficiency measures is estimated at 15 to 25% of overall motor electricity use. The most important element of any motor efficiency effort is a systems approach, that is, to optimize the whole system and not focus on a single element (the motor). The most important opportunities are found in optimizing the system driven by the motor, managing the motor system through adjustable speed drives (ASDs), and improving the efficiency of the motor. Compressed air systems use approximately 16% of industrial motor electricity. However, the overall efficiency of compressed air systems is often less than 10%. There are large opportunities for reduction of energy use by reducing leaks, improving air intake, and abolishing inappropriate use of compressed air as an energy source (and replacing, e.g., by direct drives). ASDs have revolutionized motor systems by allowing for affordable and reliable speed control using rugged conventional induction motors. ASDs work by varying the frequency of the electricity supplied to the motor, thereby changing the motor's speed relative to its normal supply frequency. This is accomplished by rectifying supplied alternating current to direct current and then synthesizing an alternating current at another frequency. This is accomplished with an inverter, which is a solid-state device in modern ASDs.

6. ENERGY INTENSITY TRENDS

In aggregate terms, studies have shown that technical efficiency improvement of 1 to 2% per year has been observed in the industrial sector in the past. During and after the years of the oil shock, U.S. industrial energy intensity declined by 3.5% per year (between 1975 and 1983). Between 1984 and 1994, industrial energy intensity declined by less than 1% on average. Figure 1 gives an overview of energy intensity trends in the industrial sector in industrialized countries.

The trends demonstrate the capability of industry to improve energy efficiency when it has the incentive to do so. Energy requirements can be cut by improved energy management and new process development. In addition, the amount of raw materials demanded by a society tends to decline as the society reaches certain stages of industrial development, leading to a decrease in industrial energy use. At the same time, the mix of fuels used by industry changes over time, affecting energy intensity as well as emissions. Overall changes in the manufacturing fuel mix are driven by price changes, availability and reliability of supply, efficiency gains, and sector structure. For example, in the U.S. manufacturing industry, a general trend toward a reduced use of coal and oil since 1958 is obvious from Fig. 2. Although coal consumption remained fairly constant after the oil price shock of the early

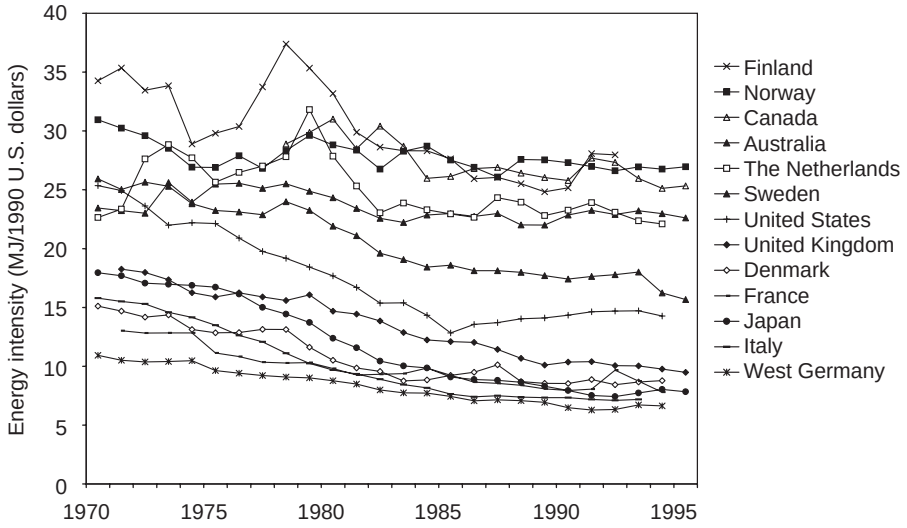


FIGURE 1 Industrial sector economic energy intensity trends in selected industrialized countries, 1970–1995. From the Lawrence Berkeley National Laboratory.

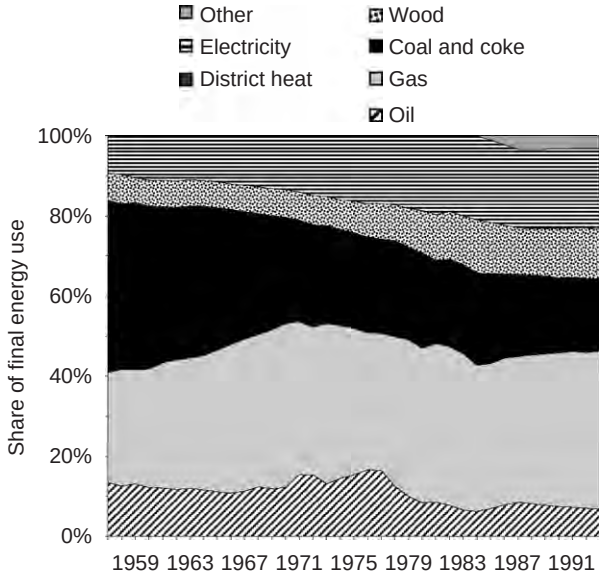


FIGURE 2 Relative fuel mix for U.S. manufacturing industry between 1958 and 1994. Energy is expressed by fuel as the share of the total final energy mix. From the Lawrence Berkeley National Laboratory.

1970s, it has declined further since. This change is partially a reflection of the decline of the energy-intensive industries in the U.S. economy. Natural gas, an easier and more efficient fuel to use, shows a dramatic increase in its share of the fuel mix over time. Also important is the increased use of electricity throughout industry, reflecting an overall trend in society of increased penetration of electric devices and equipment. Light industries tend to consume relatively more electricity than do energy-intensive industries.

The accounting of trends in structural shift, material intensity, and technical energy efficiency, as well as their interactions, can be extremely difficult. To understand trends in energy intensity, it is important to analyze the structure of the industrial sector. Reduction of energy intensity is linked closely to the definition of structure, structural change, and efficiency improvement. Decomposition analysis is used to distinguish the effects of structural change and efficiency improvement. Structural change can be broken down into intrasectoral (e.g., a shift toward more recycled steel) and intersectoral (e.g., a shift from steel to aluminum within the basic metals industry). A wide body of literature describes decomposition analyses and explains the trends in energy intensities and efficiency improvement. Decomposition analyses of the aggregate manufacturing sector exist mainly for industrialized countries but also for China, Taiwan, and selected countries such as those in Eastern Europe. The results show that different patterns exist for various countries, possibly due to specific conditions as well as differences in driving forces such as energy prices and other policies in these countries. More detailed analyses on the subsector level are needed to understand these trends better. Changes in energy intensities can also be disaggregated into structural changes and efficiency improvements at the subsector level. In the iron and steel industry, energy intensity is influenced by the

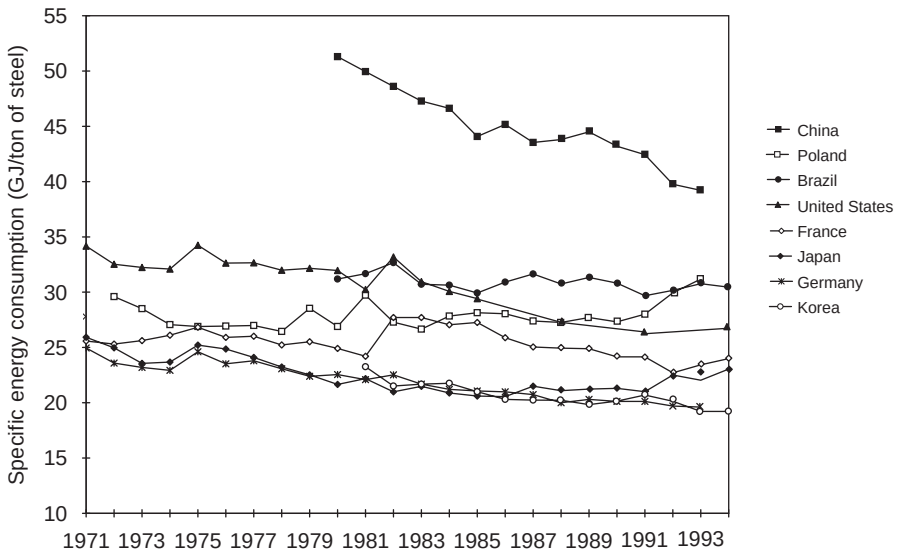


FIGURE 3 Trends in physical energy intensity (GJ/ton of crude steel) in seven countries between 1971 and 1994. From Worrell *et al.* (1997).

raw materials used (e.g., iron ore, scrap) and the products produced (e.g., slabs, thin rolled sheets). A study on the iron and steel industry examined trends in seven countries that together produced nearly half of the world's steel. Figure 3 shows the trends in physical energy intensity in these countries, expressed as primary energy used per ton of crude steel. The large differences in intensity among the countries are shown, as are the trends toward reduced intensity in most countries. Actual rates of energy efficiency improvement varied between 0 and 1.8% per year, whereas the energy intensity increased in the case of the restructuring economy of Poland.

7. POTENTIAL FOR ENERGY EFFICIENCY IMPROVEMENT

Much of the potential for improvement in technical energy efficiencies in industrial processes depends on how closely such processes have approached their thermodynamic limit. There are two types of energy efficiency measures: (1) more efficient use of existing equipment through improved operation, maintenance, or retrofit of equipment; and (2) use of more efficient new equipment through the introduction of more efficient processes and systems at the point of capital turnover or expansion of production. More efficient practices and (new) technologies exist for all industrial sectors. Table II outlines some examples of energy efficiency improvement techniques and practices.

A large number of energy-efficient technologies are available (Table II) in the steel industry, including continuous casting, energy recovery, and increased recycling. Large technical potentials exist in most countries, ranging from 25 to 50% even in industrialized countries. New technologies that are under development (e.g., smelt reduction, near net shape casting) will reduce energy consumption as well as environmental pollution and capital costs. In the chemical industry, potentials for energy savings in ammonia making are estimated at up to 35% in Europe and between 20 and 30% in Southeast Asia. Energy savings in petroleum refining are possible through improved process integration, cogeneration, energy recovery, and improved catalysts. Compared with state-of-the-art technology, the savings in industrialized countries are estimated at 15 to 20% and are higher for developing countries. Large potentials for energy savings exist in nearly all process stages of pulp and paper production (e.g., improved dewatering technologies, energy and waste heat recovery, new pulping technologies). Technical potentials are estimated at up to 40%, with higher long-term potentials. Energy savings in cement production are possible through increased use of additives (replacing the energy-intensive clinker), use of the dry process, and a large number of energy efficiency measures (e.g., reducing heat losses, using waste as fuel). Energy savings potentials of up to 50% exist in the cement industry in many countries through efficiency improvement and the use of wastes, such as blast furnace slags and fly ash, in cement making.

TABLE II

Examples of Energy Efficiency Improvement Measures in Energy-Intensive Industry

Iron and steel
Heat recovery for steam generation, preheating combustion air, high-efficiency burners
Adjustable speed drives, heat recovery coke oven gases, dry coke quenching
Efficient hot blast stove operation, waste heat recovery for hot blast stove, top gas power recovery turbines, direct coal injection
Recovery BOF–gas, heat recovery of sensible heat BOF–gas, closed BOF–gas system, optimized oxygen production, increased scrap use, efficient tundish preheating
UHP process, oxy-fuel injection for EAF plants, and scrap preheating
Heat recovery (steam generation), recovery of inert gases, efficient ladle preheating
Use of continuous casting, "hot connection" or direct rolling, recuperative burners
Heat recovery, efficient burners annealing and pickling line, continuous annealing operation
Chemicals
Process management and thermal integration (e.g., optimization of steam networks, heat cascading, low and high-temperature heat recovery, heat transformers), mechanical vapor recompression
New compressor types
New catalysts
Adjustable speed drives
Selective steam cracking, membranes
High-temperature cogeneration and heat pumps
Autothermal reforming
Petroleum refining
Reflux overhead vapor recompression, staged crude preheat, mechanical vacuum pumps
Fluid coking to gasification, turbine power recovery train at the fluid catalytic cracker, hydraulic turbine power recovery, membrane hydrogen purification, unit-to-hydrocracker recycle loop
Improved catalysts (reforming), hydraulic turbine power recovery
Process management and integration
Pulp and paper
Continuous digester, displacement heating/batch digesters, chemi–mechanical pulping
Black liquor gasification/gas turbine cogeneration
Oxygen predelignification, oxygen bleaching, displacement bleaching
Tampella recovery system, falling film black liquid evaporation, lime kiln modifications
Long nip press, impulse drying, other advanced paper machines
Improved boiler design/operation (cogeneration), distributed control systems
Cement
Improved grinding media and linings, roller mills, high-efficiency classifiers, wet process slurry
Dewatering with filter presses
Multistage preheating, precalciners, kiln combustion system improvements, enhancement of internal heat transfer in kiln, kiln shell loss reduction, optimize heat transfer in clinker cooler, use of waste fuels
Blended cements, cogeneration
Modified ball mill configuration, particle size distribution control, improved grinding media and linings, high-pressure roller press for clinker pregrinding, high-efficiency classifiers, roller mills

Source. Worrell, Levine, Price, *et al*. (1997).

In the United States, various studies have assessed the potential for energy efficiency improvement in industry. One study assessed the technologies for various sectors and found economic energy savings of 7 to 13% over the business-as-usual trends between 1990 and 2010. Technologies such as the ones described previously (Table II) are important in achieving these potentials.

However, barriers may partially block the uptake of those technologies. Barriers to efficiency improvement can include unwillingness to invest, lack of available and accessible information, economic disincentives,

and organizational barriers. The degree to which a barrier limits efficiency improvement is strongly dependent on the situation of the actors (e.g., small companies, large industries). A range of policy instruments is available, and innovative approaches or combinations have been tried in some countries. Successful policy can contain regulation (e.g., product standards) and guidelines, economic instruments and incentives, voluntary agreements and actions, information, education and training, and research, development, and demonstration policies. Successful policies with proven track records in several sectors include technology development and utility/government programs and partnerships. Improved international cooperation to develop policy instruments and technologies to meet developing country needs will be necessary, especially in light of the large anticipated growth of the manufacturing industry in this region.

SEE ALSO THE FOLLOWING ARTICLES

Commercial Sector and Energy Use • Decomposition Analysis Applied to Energy • Global Energy Use: Status and Trends • Goods and Services: Energy Costs • Industrial Agriculture, Energy Flows in • Industrial Ecology • Industrial Energy Efficiency • Industrial Symbiosis • Information Technology and Energy Use • International Comparisons of Energy End Use: Benefits and Risks

Further Reading

Ang, B. W. (1995). Decomposition methodology in industrial energy demand analysis. *Energy* **20,** 1081–1096.

Brown, M. A., Levine, M. D., Romm, J. P., Rosenfeld, A. H., and Koomey, J. G. (1998). Engineering–economic studies of energy technologies to reduce greenhouse gas emissions: Opportunities and challenges. *Annu. Rev. Energy Envir.* **23,** 287–385.

De Beer, J., Worrell, E., and Blok, K. (1998). Future technologies for energy efficient iron and steel making. *Annu. Rev. Energy Envir.* **23,** 123–205.

Howarth, R. B., Schipper, L., Duerr, P. A., and Strom, S. (1991). Manufacturing energy use in eight OECD countries: Decomposing the impacts of changes in output, industry structure, and energy intensity. *Energy Econ.* **13,** 135–142.

International Energy Agency. (1997). "Indicators of Energy Use and Efficiency: Understanding the Link between Energy and Human Activity." IEA/OECD, Paris.

Nadel, S., Elliott, N., Shepard, M., Greenberg, S., Katz, G., and de Almeida, A. T. (2002). "Energy-Efficient Motor Systems: A Handbook on Technology, Program, and Policy Opportunities," 2nd Ed. American Council for an Energy Efficient Economy, Washington, DC.

Nilsson, L. J., Larson, E. D., Gilbreath, K. R., and Gupta, A. (1995). "Energy Efficiency and the Pulp and Paper Industry." American Council for an Energy Efficient Economy, Washington, DC.

Price, L., Michaelis, L., Worrell, E., and Khrushch, M. (1998). Sectoral trends and driving forces of global energy use and greenhouse gas emissions. *Mitigation Adaptation Strategies Global Change* **3,** 263–319.

Schipper, L., and Meyers, S. (1992). "Energy Efficiency and Human Activity: Past Trends, Future Prospects." Cambridge University Press, New York.

World Energy Council. (1995). "Energy Efficiency Utilizing High Technology: An Assessment of Energy Use in Industry and Buildings" (prepared by M. D. Levine, E. Worrell, N. Martin, and L. Price). WEC, London.

Worrell, E., Cuelenaere, R. F. A., Blok, K., and Turkenburg, W. C. (1994). Energy consumption by industrial processes in the European Union. *Energy* **19,** 1113–1129.

Worrell, E., Levine, M. D., Price, L. K., Martin, N. C., van den Broek, R., and Blok, K. (1997). "Potential and Policy Implications of Energy and Material Efficiency Improvement." UN Commission for Sustainable Development, New York.

Worrell, E., Martin, N., and Price, L. (1999). "Energy Efficiency and Carbon Emission Reduction Opportunities in the U.S. Iron and Steel Industry." Lawrence Berkeley National Laboratory, Berkeley, CA.

Worrell, E., Price, L., Martin, N., Farla, J., and Schaeffer, R. (1997). Energy intensity in the iron and steel industry: A comparison of physical and economic indicators. *Energy Policy* **25,** 727–744.

Industrial Symbiosis

MARIAN R. CHERTOW
Yale University
New Haven, Connecticut, United States

1. **Introduction**
2. **Definition of Industrial Symbiosis and Related Terms**
3. **The Kalundborg Model**
4. **Elements and Tools of Industrial Symbiosis**
5. **Spatial Scale of Industrial Symbiosis**
6. **Technical, Regulatory, and Business Issues**
7. **Future Directions**

Glossary

by-product exchange An organized means of sharing discarded products and materials among firms.
cascading of water and energy Repeated use in different applications of a resource, such as water or energy, that degrades over successive uses.
cogeneration The simultaneous production of electricity and heat, usually in the form of steam, from the same fuel-burning process.
eco-industrial park or estate A community of manufacturing and service businesses located on a common property seeking enhanced environmental, economic, and social performance through collaboration in managing environmental and resource issues.
industrial ecology The study of the flows of materials and energy in industrial and consumer activities, of the effects of these flows on the environment, and of the influences of economic, political, regulatory, and social factors on the flow, use, and transformation of resources.
industrial ecosystem The place where industrial symbiosis occurs, ideally for optimizing the consumption of materials and energy where the effluents of one industrial process serve as the raw materials for another process.
integrated biosystem Exchange in which at least two biological subsystems are part of an integrated process to reduce emissions and reuse agricultural by-products.
materials tracking A means of identifying and quantifying all significant material inputs and outputs of each firm in a given industrial system.
sustainable development Development that meets the needs of the present without compromising the ability of future generations to meet their own needs.
symbiosis A biological relationship in which at least two otherwise unrelated species exchange materials, energy, or information.
utility sharing Cooperative effort among proximate firms and other organizations to pool and assign water and energy resources that might otherwise be generated or supplied individually or through a large centralized authority.

Industrial symbiosis is part of a new field called industrial ecology. Industrial ecology is principally concerned with the flow of materials and energy through systems at different scales, from products to factories and up to national and global levels. Industrial symbiosis focuses on these flows through networks of businesses and other organizations in local and regional economies as a means of approaching ecologically sustainable industrial development. Industrial symbiosis engages traditionally separate industries in a collective approach to competitive advantage involving physical exchange of materials, energy, water, and/or by-products. The keys to industrial symbiosis are collaboration and the synergistic possibilities offered by geographic proximity.

1. INTRODUCTION

The term industrial symbiosis was coined in the small municipality of Kalundborg, Denmark, where a well-developed network of dense firm interactions was encountered. The primary partners in Kalundborg, including an oil refinery, a power station, a gypsum board facility, and a pharmaceutical company, share ground water, surface water, wastewater, steam, and fuel, and they also exchange a variety of by-products that become feedstocks in other processes. High levels

 © 2004 Elsevier Inc. All rights reserved.

of environmental and economic efficiency have been achieved, leading to many other less tangible benefits involving personnel, equipment, and information sharing. Many other examples of industrial symbiosis exist around the world and illustrate how the concept is applied.

This article describes the nature of the industrial symbiosis in Kalundborg. Then, it discusses the many elements of industrial symbiosis such as energy and water cascading, cogeneration, and materials exchange. It also examines tools such as input/output matching, stakeholder processes, and materials tracking. Then, the article discusses how industrial symbiosis is a useful umbrella term because it can describe exchanges across entities regardless of whether they are colocated, located near one another but not contiguous, or located within a broader spatial area such as regionally. It also examines technical and regulatory considerations that have come into play in various locations and that can facilitate or inhibit industrial symbiosis. Finally, it considers future directions with regard to industrial symbiosis based on historical and current experience.

2. DEFINITION OF INDUSTRIAL SYMBIOSIS AND RELATED TERMS

The term symbiosis builds on the notion of biological symbiotic relationships in nature where at least two otherwise unrelated species exchange materials, energy, or information in a mutually beneficial manner, with the specific type of symbiosis known as mutualism. So, too, industrial symbiosis consists of place-based exchanges among different entities that yield a collective benefit greater than the sum of individual benefits that could be achieved by acting alone. Such collaboration can also foster social values among the participants that can extend to surrounding neighborhoods. As described in what follows, the symbioses need not occur within the strict boundaries of a park, despite the popular use of the term eco-industrial park to describe organizations engaging in exchanges.

At the same time interest began to develop in industrial symbiosis and eco-industrial parks, a number of other parallel tracks advanced that might be construed, broadly, as green development. These include residential, commercial, industrial, and community development as captured in terms such as sustainable architecture, green buildings, sustainable communities, and smart growth. Eco-industrial development or sustainable industrial development narrows down the possibilities to refer predominantly to industrial and commercial activities. Cooperating businesses that include a materials/water/energy exchange component qualify the activity as industrial symbiosis. The materials exchange component has also been referred to as a by-product exchange, by-product synergy, or waste exchange and may also be referred to as an industrial recycling network.

As with the term industrial park, the term eco-industrial park refers to eco-industrial development on a particular plot of real estate. An eco-industrial park may include many ecologically desirable goals, including mechanisms to reduce overall environmental impact, conserve materials and energy, and foster cooperative approaches to resource efficiency and environmental management. The terms industrial estate and eco-industrial estate are more commonly used in Asia and can include communities of workers who live in or near the group of businesses constituting the industrial estate. Gunther Pauli popularized the notion of zero-emissions parks to emphasize the drive toward sustainable industrial development. The agricultural community has found that while traditional agriculture incorporated cyclical reuse of by-products, industrial agriculture became much more linear, consuming materials and disposing of wastes. Therefore, many zero-emissions researchers describe integrated biosystems in which at least two biological subsystems are part of an integrated process to reduce emissions and reuse agricultural by-products.

Some writers refer to eco-industrial networks to capture a broad range of environmental and economic activities among businesses. Just as economic clusters have come to mean a group of businesses that are sectorally related by the products they make and use, such as the furniture cluster in central North Carolina, the term eco-industrial clusters is sometimes used to describe environmental interactions among firms in the same or related industries. Industrial complexes of sectorally related firms have been successful for the past several decades in overall pollution reduction in industries such as pulp and paper, sugarcane, textiles, and plastics.

Although materials and energy exchanges have been a significant part of industry for centuries, focus on environmental attributes is much more recent. In a foundational 1989 article on industrial ecology, Frosch and Gallopoulos described the underlying notion of an “industrial ecosystem” in which “the consumption of energy and materials is optimized and the effluents of one process … serve as the raw

materials for another process." Others have extended the ecosystem metaphor to see related industrial activities as a food web and to interpret the roles of various scrap and remanufacturing businesses as the scavengers and decomposers of these systems.

In any multidisciplinary field such as industrial ecology, there are strands from many disciplines and paths of research that are the antecedents of current understanding. Industrial symbiosis was not popularized until the first articles and analyses of the industrial region in Kalundborg were published during the early 1990s. One origin of industrial symbiosis is from the chemical industry, which embeds an intrinsic value chain of materials as they are degraded. Other terms discussed previously originate in ecology, agricultural studies, engineering, and/or business economics. Cogeneration and utility sharing have been justified on engineering, environmental, and economic grounds. Indeed, private industry sees cost efficiency as a driving factor, whereas city planners, economic development experts, and real estate developers also emphasize land use, social and environmental aspects, and the synergies that can arise from colocation. Van Berkel summarized three types of synergies that help to characterize trends emerging in different types of symbiosis network: (1) synergies across the supply chain, (2) synergies from shared use of utilities, and (3) synergies from local use of by-products. As new projects are formed and old projects evolve, more common definitions will emerge.

3. THE KALUNDBORG MODEL

The model of industrial symbiosis was first fully realized in the industrial district at Kalundborg. Although it is continually evolving, there are currently some 20 exchanges occurring among the symbiosis participants involving water, energy, and a wide variety of residue materials that become feedstocks in other processes. The waste exchanges alone amount to some 2.9 million tons of materials per year. Water consumption has been reduced by a collective 25%, and the power station has experienced a 60% reduction in water use through cycling. Some 4500 homes receive district heat, which replaced 3500 small oil-fired units. A key coordinating role is played by the Kalundborg Center for Industrial Symbiosis, formed by the symbiosis partners under the auspices of the Industrial Development Council of the Kalundborg region. In addition to several companies that participate as recipients of materials or energy, many of which are beyond the local region, the specific industrial ecosystem in 2003 consisted of six main business partners located in Kalundborg as follows:

- *Energy E2 Asnæs Power Station:* the largest plant producing electricity in Denmark, a 1037-MW facility fired by orimulsion and coal
- *Statoil Refinery of Kalundborg:* Denmark's oldest and largest refinery, with a production capacity of approximately 5.5 million tons of crude oil annually
- *Gyproc A/S:* part of BPB Gyproc, producing plasterboards and plasterboard-based systems for the construction industry
- *Novo Nordisk A/S:* an international, biotechnology, and pharmaceutical company offering a wide range of insulin products to world markets
- *Novozymes A/S:* the largest producer of enzymes in the world, with its largest single factory in Kalundborg
- *A/S Bioteknisk Jordrens Soilrem:* specialist in remediation of soil contaminated by oil, chemicals, and heavy metals that cleans approximately 500,000 tons of contaminated soil annually

The municipality of Kalundborg and the intermunicipal waste treatment company Noveren I/S also participate actively in the symbiosis.

Figure 1 shows the interrelationships of the symbiosis participants. Each exchange was developed as an economically attractive business arrangement between participating firms through bilateral contracts. It is significant to mention that this symbiosis was not based on a planning process and that it continually evolves. In 1961, low levels of ground water locally led the refinery to pipe in surface water from a nearby lake, which now also supplies the power station and pharmaceutical plant. The gypsum board plant began receiving excess butane gas from the refinery beginning in 1972, and this was recently discontinued. Since the 1980s, the number of exchanges has accelerated. Regulation has played an indirect role over the years; for example, the national ban on placing organic waste streams into landfills caused the pharmaceutical company to seek arrangements to apply its sludges on agricultural lands. Social cohesion is regularly cited as a key element of success in the Kalundborg symbiosis.

Rather than a static system of locked-in firms and technologies, individual participants in the symbiosis have changed significantly over time, and the

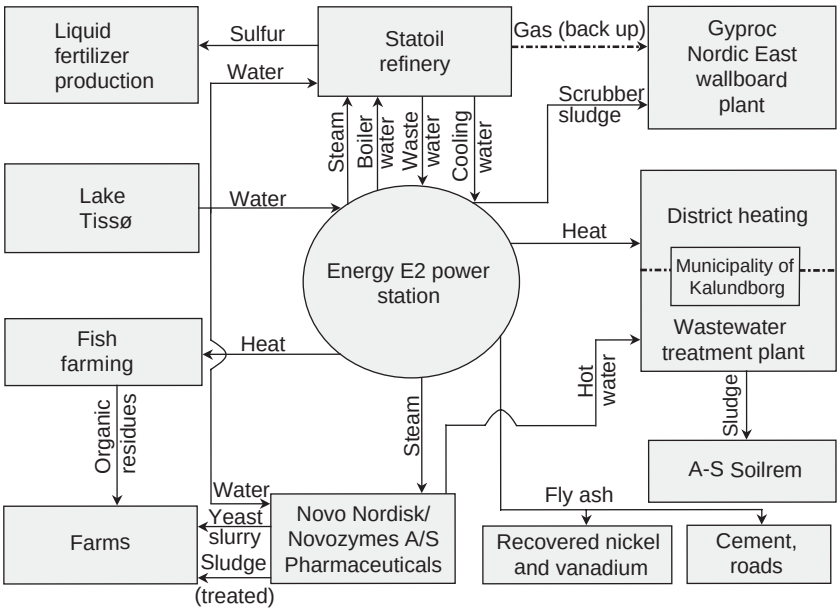


FIGURE 1 The industrial symbiosis at Kalundborg, Denmark.

ecosystem as a whole has adapted. Over the past several years, Kalundborg's Statoil Refinery doubled its capacity based on North Sea claims, the Asnæs Power Station switched from coal to orimulsion to comply with mandated carbon dioxide (CO_2) reduction, and the pharmaceutical plant split into two ventures, eliminated some product lines (including penicillin), and increased others. Although each individual business change alters the makeup of the industrial symbiosis, the changes collectively have not diminished the overall nature of the symbiosis. In fact, in the case of the power station's flue gas desulfurization program that creates calcium sulfate (or gypsum) used by the plasterboard plant, the change in fuel at the power station from coal to orimulsion led to the creation of even more gypsum, thereby increasing the benefits for the participants. Rather than tie themselves to a single supplier, the symbiosis participants try to insulate themselves from supplier interruptions by diversifying sources to reduce business risk, just as in traditional supplier–customer relationships.

4. ELEMENTS AND TOOLS OF INDUSTRIAL SYMBIOSIS

Drawing on industrial ecology, industrial symbiosis incorporates many elements that emphasize the cycling and reuse of materials in a broader systems perspective. A brief discussion of several elements is presented, followed by a discussion of useful analytical tools for industrial symbiosis. These elements include embedded energy and materials, life cycle perspective, cascading, loop closing, and tracking material flows.

4.1 Embedded Energy and Materials

To create a product, resources are used for extraction of materials, transportation, primary and secondary manufacturing, and distribution. The total energy and materials used is the amount embedded in that product. Reusing by-products in an industrial symbiosis preserves the embedded materials and energy for a longer period as part of the cycling emphasized in industrial ecology. Cogeneration is a specific means of cycling embedded energy by reusing waste heat to produce electricity or by using steam from electric power generation as a source of heat.

4.2 Life Cycle Perspective

As described by Graedel and Allenby in the first textbook in the field, industrial ecology is "a systems view in which one seeks to optimize the total materials cycle from virgin material, to finished material, to component, to product, to obsolete product, and to ultimate disposal." A life cycle perspective ensures a breadth of focus that is not limited to what happens within one facility or factory but rather considers the entire set of environmental

impacts that occur at each stage of industrial development and use across entities. With respect to industrial symbiosis, a life cycle perspective is helpful in assessing symbiotic opportunities—the junctures in the product life cycle at which the by-product of concern can be considered for another use.

4.3 Cascading

Cascading occurs when a resource, such as water or energy, is used repeatedly in different applications. In successive uses, the resource is of lower quality, a lower level of refinement, and/or lower value. A cascade must include at least one use beyond the virgin use of the resource and is generally conceptualized as a downward step diagram. The cascade terminates when either a considerable amount of energy must be put in to recover value from the resource or the resource is discarded. Cascading is a common strategy for industrial symbiosis because the firm producing the used resource can save on treatment or disposal and possibly can earn compensation in exchange for the value of the resource. The environmental benefits of cascading are numerous, including the reduced use of virgin resources, the avoided impact of resource extraction, and the reduced deposition of waste into the environment.

4.4 Loop Closing

If cascading is conceptually stepwise, loop closing is more circular. A general name for many different variations of reuse and recycling of resources, loop closing occurs when a resource has a cyclical flow embedded in the industrial ecosystem and the resource, rather than becoming entirely degraded, reappears akin to its original form. Thus, glass bottles may be washed out directly and returned to use, or glass bottles may be collected and crushed into cullet before the glass cullet is melted to make new glass containers. Both methods close the loop and return the glass to a form similar to its previous form. The economic and environmental benefits of loop closing are similar to those of cascading.

4.5 Tracking Material Flows

Also key to industrial symbiosis studies is the tracking of material, water, and energy flows. This form of accounting captures instances of loop closing, cascading, and unidirectional flows. Over time, many applications of materials tracking at different scales have yielded specific tools, such as material flow analysis (MFA) and substance flow analysis (SFA), that formalize tracking practices. Materials tracking for symbiosis identifies and quantifies all significant material inputs and outputs of each firm in the subject industrial system. The results are analyzed to suggest opportunities for exchange of materials among firms as well as opportunities for more efficient resource use in the industrial ecosystem.

To the extent that new symbioses can be planned or existing exchanges can be augmented, several tools have proven to be useful in industrial symbiosis analysis. These tools are industrial inventories, input/output matching, stakeholder processes, and materials budgeting.

4.6 Industrial Inventories

Once a district has been identified as a candidate for industrial symbiosis, it is useful to begin with an inventory of local businesses and other resources, including utilities and relevant institutions. Because confidentiality is a critical aspect of dealing with private companies, data can be collected generically concerning the inputs and outputs of relevant industrial processes to achieve a base analysis from which to assess further goals.

4.7 Input/Output Matching

Key to symbiosis is the matching of inputs and outputs to make links across industries. There are various ways in which to collect these data in a systematic fashion, including written and oral surveys and literature review. The U.S. Environmental Protection Agency (EPA) commissioned three pieces of input/output matching software during the late 1990s to be used as a planning tool to allow communities to investigate what mix of specific types of industries might support industrial symbiosis. FaST (Facility Synergy Tool) was a database of industry profiles describing typical inputs and outputs of specific types of facilities. It had a data input screen and a search mechanism to identify possible input/output matches among facilities. DIET (Designing Industrial Ecosystems Tool) allowed scenario analysis of various combinations of facilities. It included a linear programming optimization model to enable planners to optimize environmental, economic, and/or employment objectives and to change the relative weights of each. REaLiTy (Regulatory, Economic, and Logistics Tool) assisted

with sorting out regulatory hurdles likely to be confronted depending on the actual materials chosen for exchange.

A caveat on these and related models is that they can overemphasize idealized what-if scenarios with too little recognition of the time-consuming processes involved in attracting any business, let alone the optimal suite of symbiotic partners. It is also appropriate to add the cautionary note that most industrial products generally have not been designed for reuse, and so great care must be taken to see that the proposed match across systems is sound with respect to regulation and integration of best practices.

4.8 Stakeholder Processes

To the extent that industrial symbiosis involves different layers of unconnected participants, a broad array of community involvement techniques is warranted. Both the Londonderry Eco-Industrial Park in New Hampshire and the Sustainable Industry Park in Cape Charles, Virginia, assembled many diverse stakeholders and then conducted design charettes to seek input on what an eco-industrial project should look like in the local context. Whether and how to pursue specific covenants or conditions as a type of deed restriction is one topic of stakeholder meetings. Applied Sustainability gathered experience convening stakeholders from business and government in its efforts to create by-product synergy in Tampico, Mexico, and Alberta, Canada. Openness among participating companies and continued coordination by a stakeholder group such as an advisory council is important both to establish and to maintain the momentum of a symbiosis.

4.9 Materials Budgeting

Materials budgeting, a type of materials tracking discussed previously, can be used to map energy and material flows through a chosen system. Formally, in industrial ecology, materials budgeting embraces three concepts: (1) reservoirs, where a material is stored; (2) flux, which is the amount of material entering or leaving a reservoir per unit time; and (3) sources and sinks, which are rates of input and loss of specific materials entering or leaving a system. Because it helps to identify both stocks and flows, materials budgeting can be a basic building block of an industrial symbiosis analysis. By tracking material flows, Schwarz and Steininger determined the existence of a symbiotic network in Styria, Austria, larger than that at Kalundborg.

5. SPATIAL SCALE OF INDUSTRIAL SYMBIOSIS

In general, industrial symbiosis occurs locally or regionally across participating companies. Increasing the distance among firms lessens the breadth of exchange opportunities because it is not cost-effective to transport water and steam beyond regional boundaries, whereas by-products can often travel much farther. Observing numerous instances of industrial symbiosis, Chertow devised a taxonomy of materials exchange types to consider spatial and organizational elements. These include through waste exchanges (type 1); within a facility, firm, or organization (type 2); among firms colocated in a defined eco-industrial park (type 3); among local firms that are not colocated (type 4); and among firms organized "virtually" across a broader region (type 5).

5.1 Type 1: Through Waste Exchanges

Most often focused at the end-of-life stage of a product or process, examples of these exchanges would include contributions of used clothing for charity and collection of scrap metal or paper by scrap dealers or municipal recycling programs. Waste exchanges formalize trading opportunities by creating hard-copy or online lists of materials that one organization would like to dispose of and another organization might need. The scale of trades can be local, regional, national, or global. The exchanges accomplish various input/output savings on a trade-by-trade basis rather than continuously. They feature exchange of materials rather than of water or energy.

5.2 Type 2: Within a Facility, Firm, or Organization

Some kinds of materials exchange can occur primarily inside the boundaries of one organization rather than with a collection of outside parties. Large organizations often behave as if they are separate entities and may approximate a multifirm approach to industrial symbiosis. Significant gains can be made within one organization by considering the entire life cycle of products, processes, and services, including upstream operations such as purchasing and product design. In the context of state-owned enterprises popular in Asia, material exchanges can extend along the supply chain of a product still under single ownership. The Guitang Group, a state-owned

enterprise in China, expanded from sugar refining to include alcohol and paper production as a means of deriving income from the use of its own by-products, specifically molasses and bagasse. The group has recently extended its exchange network into a community-wide one to receive by-products of other sugar producers.

5.3 Type 3: Among Firms Colocated in a Defined Eco-industrial Park

In this approach, businesses and other organizations that are contiguously located can exchange energy, water, and materials and can go further to share information and services such as permitting, transportation, and marketing. Type 3 exchanges occur primarily within the defined area of an industrial park or industrial estate, but it is also common to involve other partners over the fence. The areas can be new developments or retrofits of existing ones. The Londonderry Eco-Industrial Park was established in a green field adjacent to an industrial zone. The Industrial Estate Authority of Thailand has a plan to adopt sustainable practices at 28 industrial estates at various stages. In Canada, Burnside Industrial Park in Dartmouth, Nova Scotia, is an example of retrofitting an existing industrial park to improve environmental performance. The park is spread over 2500 acres, with more than 1200 businesses employing some 18,000 people. Côté reported that researchers from Dalhousie University have been devising principles and strategies to work within the park to encourage its transformation into an industrial ecosystem.

5.4 Type 4: Among Local Firms That Are Not Colocated

Partners in this type of exchange need not be sited adjacent to one another but rather are located within a small geographic area, as in Kalundborg, where the primary partners are within roughly a 2-mile radius of each other. Type 4 exchanges draw together existing businesses that can take advantage of already generated material, water, and energy streams and also provide the opportunity to fill in new businesses based on common service requirements and input/output matching. In the Kwinana industrial area south of Perth, Australia, heavy process industries were established beginning in the 1950s and now employ some 3600 people. Through their industrial council, the companies have resolved to develop regional synergies and have doubled the number of companies involved in exchanges over the past decade or so.

5.5 Type 5: Among Firms Organized Virtually across a Broader Region

Given the high cost of moving and other critical variables that enter into decisions about corporate location, very few businesses will relocate solely to be part of an industrial symbiosis. Type 5 exchanges depend on virtual linkages rather than colocation. Although still place-based enterprises, type 5 exchanges encompass a regional economic community in which the potential for the identification of by-product exchanges is greatly increased by the larger number of firms that can participate. An additional attractive feature is the potential to include small outlying agricultural and other businesses. Self-organized groups, such as the network of scrap metal dealers, agglomerators, and dismantlers that feed particular mills and subsystems (e.g., auto recycling), could be considered as type 5 systems. Specific projects organized in the Triangle J region of North Carolina, in Tampico, and in Alberta sought to identify relevant material and energy inputs and outputs and to match companies within their regions to each other when doing so is economically and environmentally efficient.

6. TECHNICAL, REGULATORY, AND BUSINESS ISSUES

Because industrial symbiosis requires interorganizational cooperation, including knowledge of physical flows, this creates both barriers and opportunities beyond those of more conventional development projects. These include technical issues, regulatory issues, and business issues.

6.1 Technical Issues

In general, symbiotic industrial facilities need to be in close proximity to avoid large transportation costs and energy degradation during transit. High-value by-products, such as pure sulfur from sour gas treatment, are exceptions. Industrial symbioses of types 3 and 4 usually incorporate at least one anchor tenant with a large, continuous by-product stream such as a power plant or refinery. Wastes that are largely organic in nature (e.g., the effluent from

fermentation of pharmaceuticals or brewing), as well as raw agricultural or forestry by-products, can also be attractive. Use of organic streams from fermentation as feed or fertilizer requires assurance that toxic components or organisms are absent. Materials production, such as the manufacture of wallboard, is technically challenging and requires close matching of compositions. Supply security is important to the users of by-product streams, as would be the reliability of more conventional materials suppliers located farther away. The problem of achieving sufficient scale when aggregating by-products is significant because even if discarded materials are collected from numerous facilities, the total volume can fall far short of the raw materials necessary to support a new operation.

In Japan, scarcity of land for waste disposal has prompted significant cooperation and technological change in affected industries. The Japanese cement industry now consumes approximately 6% of the country's waste, including half of the fly ash produced by power plants. As a result, the Japanese cement industry reports the smallest energy consumption per ton of cement produced among developed countries. Taiheiyo Cement has been developing a new technology to convert plastic waste (including PVC) and other industrial wastes into raw materials and fuel. Another technology decomposes dioxin from fly ash through heating at high temperatures inside cement kilns. Heavy metals are extracted by wet-type refining technology to be recycled for the nonferrous metal industry.

6.2 Regulatory Issues

Industrial symbiosis is often at odds with environmental regulatory requirements, which may preclude by-product exchanges or at least serve as a very strong disincentive. For example, in the United States, the Resource Conservation and Recovery Act (RCRA) regulates the treatment, storage, and disposal of numerous wastes as a means of averting risks stemming from the improper management of hazardous waste. The law requires that by-products be matched to specific mandatory protocols through a very extensive set of rules that leaves little room for innovative schemes for by-product reuse as feedstocks elsewhere. This inflexibility is based in large part on a deep-seated fear of sham recycling, an undertaking where the generator of a waste product makes a show of reusing that by-product merely to escape treatment requirements. Industrial symbiosis must be distinguished from such efforts if it is to develop within the current regulatory system.

In many developing countries, governments lead the process for establishing industrial parks and estates. Lowe found that this often slows the development process. In addition, these projects are likely to rely on funding from development banks for which industrial symbiosis projects might seem unconventional and, thus, fall outside of development bank guidelines.

In some instances, regulatory actions encourage industrial symbiosis. Landfill bans in key European countries have driven symbiotic practices such as the reuse of organic wastes prohibited from land disposal in Denmark and The Netherlands. Very high tipping fees for waste disposal in Canada and climate change levies in the United Kingdom have been cited as stimulating innovation and action in by-product reuse.

6.3 Business Issues

Although private actors need not be the initiators, they clearly must be committed to the implementation of industrial symbiosis because, in most instances, the industrial symbiosis flows either belong to private actors or will be shared with them in the case of municipal wastewater linkages. Whether key private actors can appropriate sufficient benefit from environmental gains is a challenge to industrial symbiosis, especially given that the level of business benefit to various partners is not uniform. Businesses generally want to address nonproduct problems at the lowest cost and with the least use of resources. Such objectives might not include the time or inclination to work with others, especially concerning low-value wastes.

As a practical matter, all significant development projects take a long time, require substantial capital investment, and must constantly meet the test of whether investment in them will create sufficient returns. These issues are compounded with eco-industrial projects by the need for multiparty planning and coordination and the attendant transaction costs, including the risk that a proposed partner will relocate.

Industrial symbiosis raises the question of whether the desire to reuse waste streams comes at the expense of adhering to pollution prevention principles calling for the elimination of waste at the front end of the process. Some suspect that industrial symbiosis projects favor older dying industries and keep them going rather than fostering a new generation of clean technology. Overall, industrial

symbiosis could potentially discourage companies from updating their systems, plant, and equipment, instead substituting the veil of interdependence.

At the first level of analysis, it is reasonable to assume that companies will do what is in their economic interest. If, through incremental improvements or through broader scale process redesign, a company can eliminate waste in a cost-effective manner, rational actors will do so. In this sense, pollution prevention comes first. However, it is plausible that the opportunity for symbiosis might make the proposed process improvement fall lower in priority in a company's capital outlay scheme, in which case the company's own economic decision making might favor the symbiosis over pollution prevention.

7. FUTURE DIRECTIONS

An interesting debate in the industrial symbiosis literature is the extent to which industrial ecosystems are, like their natural counterparts, self-organizing and, if so, the extent to which additional policy and planning would advance eco-industrial development more quickly. Coordinating activities through representative councils, however, has been an important step in advancing symbiosis. Conceivably, the private sector could grab hold of industrial symbiosis as a logical extension of resource productivity by groups such as the World Business Council for Sustainable Development and its affiliates. Governments could latch onto eco-industrial parks as another way in which to redevelop brownfields and structure new industrial developments. On the one hand, a public already nervous about genetic engineering may become more restive about material exchanges that bring one industry's residues to an altogether different industry. On the other hand, region-specific water shortages and the scarcity of some materials accelerate the trend toward cycling and reuse.

Variety and experimentation in industrial symbiosis developments around the world will shed light on what is and is not successful, what the largest risks are, how regulatory hurdles can be overcome, what can be financed, and what is most environmentally beneficial and technologically desirable. The form could splinter back toward specific industry clusters around key materials such as plastics or around wastes such as in recycling parks. Technology may embed its own reuse such as the cycling of steam back into the industrial process in combined cycle power plants. It is also possible that much broader visions, which combine industrial symbiosis with other types of urban and regional planning, will prevail. The Japanese Central Government and the Ministry of Trade and Industry have designated several eco-towns to promote environmentally friendly practices such as zero emissions and material exchanges. The quest to find the most beneficial forms of resource sharing is part of the journey toward sustainable industrial development.

SEE ALSO THE FOLLOWING ARTICLES

Cogeneration • Earth's Energy Balance • Ecological Footprints and Energy • Ecological Risk Assessment Applied to Energy Development • Industrial Agriculture, Energy Flows in • Industrial Ecology • Industrial Energy Efficiency • Industrial Energy Use, Status and Trends • Sustainable Development: Basic Concepts and Application to Energy

Further Reading

Chertow, M. (2000). Industrial symbiosis: Literature and taxonomy. *Annu. Rev. Energy Environ.* **25,** 313–337.

Cohen-Rosenthal, E., and Muskinow, J. (eds.). (2003). "Eco-Industrial Strategies: Unleashing Synergy between Economic Development and the Environment." Greenleaf Publishing, Sheffield, UK.

Côté, R. P. (2000). "A Primer on Industrial Ecosystems: A Strategy for Sustainable Industrial Development." www.mgmt.dal.ca/sres/pdfs/primer.pdf.

Ehrenfeld, J., and Chertow, M. (2002). Industrial symbiosis: The legacy of Kalundborg. *In* "Handbook of Industrial Ecology" (R. Ayres and L. Ayres, Eds.). Edward Elgar, Cheltenham, UK.

Frosch, R., and Gallopoulos, N. (1989). Strategies for manufacturing. *Sci. Am.* **261**(3), 144–152.

Gertler, N., and Ehrenfeld, J. R. (1996). A down-to-earth approach to clean production. *Technol. Rev.* **99**(2), 48–54.

Graedel, T., and Allenby, B. (2003). "Industrial Ecology." Prentice Hall, Englewood Cliffs, NJ.

Lambert, A., and Boons, F. (2002). Eco-industrial parks: Stimulating sustainable development in mixed industrial parks. *Technovation* **22,** 471–484.

Lowe, E. (2001). "Eco-industrial Park Handbook for Asian Developing Countries," report to Asian Development Bank. RPP International, Emeryville, CA.

Inflation and Energy Prices

PAULO S. ESTEVES
Banco de Portugal
Lisbon, Portugal

PEDRO D. NEVES
Banco de Portugal and Universidade Católica Portuguesa
Lisbon, Portugal

1. **Evolution of Oil Prices and Inflation since the Early 1970*s***
2. **Pathway from Oil Prices to Inflation**
3. **Impact of Oil Price Changes on Inflation**
4. **Oil Price Volatility**

Glossary

energy consumption The expenditures made by households on energy products (fuels, electricity, gas, and heatenergy), which represent, in general, slightly less than 10% of total household expenditures.
energy intensity Energy consumption per unit of output.
first-round effects The impact of energy price changes on overall consumer prices, reflecting the reaction to energy prices (direct effect) and to energy content costs of other consumer goods and services (indirect effect).
inflation A generalized and sustainable increase in prices, usually measured using the rate of growth of the Consumer Price Index or the gross domestic product deflator.
oil price shock A high and abrupt increase (decrease) in oil prices.
second-round effects The feedback effects on consumer prices produced by other variables (typically wages) that were primarily influenced by the first-round effects of energy prices on inflation.

The remarkable energy price stability that characterized the period following World War II was interrupted in the early 1970s; the sharp increase in oil prices that occurred at that time followed the creation of the Organization of Petroleum Exporting Countries in 1973. Oil prices are, by far, the key factor underlying energy prices, even today, despite the growing role of alternative energy sources. Changes in energy prices, in turn, have an effect on inflation. The channels through which changes in oil prices affect inflation are extremely complex and are very likely to change over time. Using a study by the Organization for Economic Cooperation and Development, however, estimates of the impact of oil price increases on inflation can be made (despite the considerable uncertainty surrounding the estimates, they constitute very useful rules of thumb). A distinction can be made between first- and second-round effects. The expenditure share on oil and the taxation of energy products are the main factors influencing first-round effects. The credibility of monetary policy plays a key role in the magnitude of the second-round effects on prices. One of the most notable characteristics of oil price behavior is its extreme volatility. In particular, oil price volatility is an important source of changes in inflation; oil price futures, despite the very nice empirical properties that they exhibit, such as typically reverting to a mean price of $20–25 (U.S. dollars) per barrel, tend to produce considerable deviations in oil prices.

1. EVOLUTION OF OIL PRICES AND INFLATION SINCE THE EARLY 1970s

In the post-World War II period, until the beginning of the 1970s, oil price fluctuations were very small. From 1949 to 1970, average annual fluctuations of oil prices in U.S. dollars, as measured by the absolute value of year-to-year price changes, were of the order of 1%. Therefore, the real price (i.e., inflation adjusted) slightly declined throughout this period. This so-called Golden Age period was characterized by a remarkable price stability and very strong gross domestic product (GDP) growth in the main industrialized economies. The stability of oil prices was an

 © 2004 Elsevier Inc. All rights reserved.

important element behind the low inflation and strong economic growth.

In the 1970s, the price of oil increased dramatically as a result of the deliberate action of the Organization of Petroleum Exporting Countries (OPEC), by controlling the quantity supplied to the world market. In 1974, the oil price reached $11.50 per barrel, more than triple the price from the previous year and almost five times above the 1972 price (unless otherwise specified, the prices given herein are U.S. dollars). Some years later, in 1980 and 1981, oil prices reached values above $30 per barrel, duplicating levels achieved in the previous years.

The direct link between oil prices and inflation is illustrated by examining inflation rates; in Organization for Economic Cooperation and Development (OECD) countries, inflation reached maximum levels of about 15% in 1974 and 13% in 1980 (Fig. 1), as a result of rising oil prices. It is now widely accepted that monetary policy did not deal adequately with the first oil price shock in 1973, contributing to a situation that led to the so-called Great Inflation period. Indeed, ambivalent monetary policy, conflicted over price stability and short-term growth objectives, failed to provide an anchor to inflation expectations. It is also worth mentioning additional circumstances that were threatening price stability at the time: latent inflation pressures had emerged in OECD countries at the beginning of the 1970s, reflecting capacity constraints; powerful trade unions and the overall acceptance of wage indexation favored the existence of second-round effects; and the Bretton Woods institutional framework collapsed and floating exchange rates were adopted.

The rise in worldwide inflation after the oil price increase of 1979–1981 was much more short-lived than the one in the mid-1970s. The reaction of monetary policy to the oil price shock was clearly different, mainly orientated toward the objective of maintaining price stability. In this context, it is worth mentioning that a new strategic framework for monetary policy was adopted in some countries; the intermediate monetary targeting adopted by the Bundesbank deserves special reference.

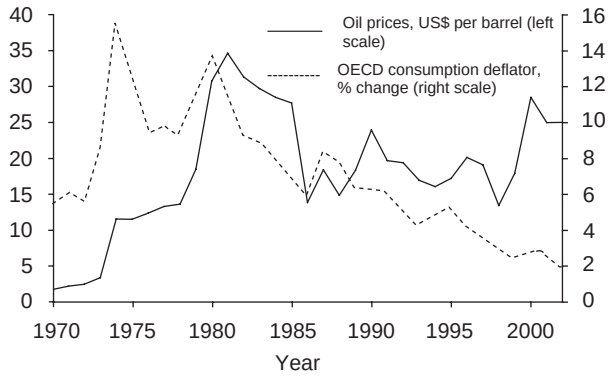


FIGURE 1 Oil prices and inflation, 1971–2002. Data from the OECD and Datastream.

In the first half of the 1980s, the oil cartel was not able to enforce the production quotas set for its members. The supply of oil increased when some countries produced more than their quotas and new producers entered the market, and oil prices started to decline. The price collapse of 1986, following serious disagreements within OPEC, marked a turning point from the previous upward trend in oil prices. Oil prices started to fluctuate within a narrower band (Fig. 1), while inflation recorded a noticeable downward trend from double-digit figures to figures close to 2%. However, oil price volatility remained a central feature of the world oil market. Oil prices decreased to around $10 per barrel in late 1998 and early 1999. After seriously misjudging the oil market during that period and even contributing to the collapse of prices, OPEC successfully pushed prices upward, overshooting its goal, and prices recovered vigorously, reaching slightly more than $30 per barrel in 2000. In 2003, oil prices were affected by the war in Iraq, remaining close to $30 per barrel in the first half of the year.

Despite all of these fluctuations in oil prices, inflation remained close to 2–3% in the main advanced economies, at least from the second half of the 1990s onward. Within the main advanced economies, the sensitivity to oil price fluctuations has declined over the past two decades. Moreover, during this period, an increase in the credibility of monetary policy decisively diminished the size of the second-round effects associated with oil price fluctuations. Central banks clearly signaled that they were focused on future inflation, rather than on current inflation, and therefore were not reacting to changes in headline-driven inflation created by transitory factors, unless if these were likely to induce changes in inflation expectations.

2. PATHWAY FROM OIL PRICES TO INFLATION

The channels through which oil price changes affect consumer prices are distinguished by first-round effects (including both direct and indirect effects)

and second-round effects. First-round effects reflect the fact that oil, as well as goods and services dependent on oil (transport, for instance), is included in the Consumer Price Index (CPI). The impacts on the energy components of the index are the so-called direct effects, and the impacts on the components of the index that have a content of energy are the so-called indirect effects. The size of first-round effects is affected by a large number of factors, including the share of energy included in total consumption expenditures, the energy intensity of production, the rate of taxation of energy, profit margins, and exchange rates.

The larger the share of energy (fuels, electricity, gas, and heat energy) in consumption expenditures, the larger the direct impact on the CPI of a given energy price change. The impact of a change in oil prices in the energy component of the price index is not instantaneous, however. Consumer fuel prices react almost immediately, but price adjustments in gas or heat energy can take somewhat longer to occur in normal circumstances. The magnitude of direct effects is affected by the importance and structure of tax rates, because oil is heavily taxed. If the tax rate has an *ad valorem* structure (i.e., the tax corresponds to a fixed proportion of the final price), an increase in the cost of energy of 1% will be transmitted to the consumer also as a 1% increase. However, if the tax corresponds to a specific tax (i.e., x cents by unit of energy), an increase in the cost of energy of 1% will be transmitted to the consumer as a somewhat smaller than 1% increase, at least before any adjustment of the tax, which is not likely to coincide in time with the usual fluctuations in oil prices in the international markets. Usually, tax systems are more complex than these very simple illustrations, reflecting the existence of both *ad valorem* and specific components. Therefore, the direct impact of oil price changes on the energy component of the consumer price index is affected by taxes. Figure 2 shows the sizable discrepancies of the tax component in unleaded gasoline prices for the G7 countries (G7 refers to the Group of Seven leading industrial nations that meet annually for economic and political discussions).

As previously stated, changes in energy prices will cause changes in the prices of goods and services that have been produced using energy resources. These changes are indirect effects. The most obvious case is the transport sector (air and surface), which is based on components that are highly energy intensive, but a large proportion of the price index is also likely to be affected. The magnitude of the effects on prices of goods and services depends, naturally, on the importance of energy as an input. The intensity of oil dependence can be measured by the consumption of oil per unit of GDP, as illustrated in Fig. 3. Given the strong decline in the use of oil per unit of GDP, the impact on prices of a given oil price shock is presently smaller than it was in the 1970s. This decline in oil intensity reflects the adjustment of advanced economies to a new era of higher and more unpredictable oil prices, through the use of alternative sources of energy and more energy-efficient technologies. Worldwide, the primary energy intensity has declined by more than 25% since 1970, slightly less so in the United States. The greatest decline in primary energy intensity has occurred in Japan and in the European Union countries, where the amount of oil used per unit of GDP has approximately fallen by half. The reduced sensitivity of the world economy to oil price changes resulting

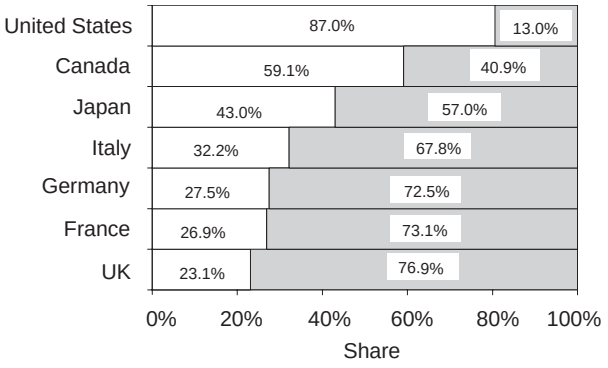


FIGURE 2 Gasoline prices and taxes (percent shares for unleaded gasoline, fourth quarter of 2002). Unshaded bars, before taxes; shaded bars, tax component. Data from the OECD.

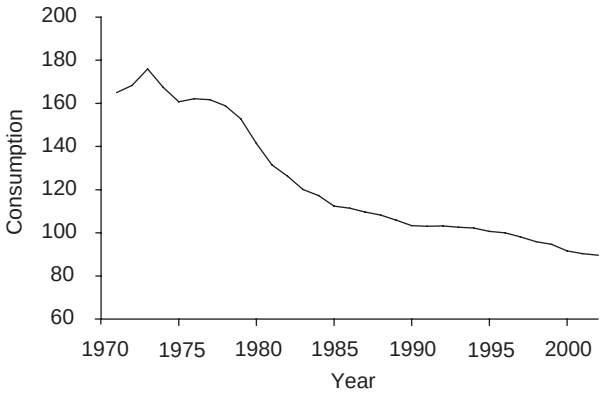


FIGURE 3 OECD consumption of oil per unit of GDP (in real terms; 1995 = 100). Data from the OECD.

from lower oil intensity has been reinforced by the decline of the oil prices in real terms (Fig. 4). Using the U.S. GDP deflator as the relevant price indicator, the relative oil price stood throughout the 1990s at levels that correspond to about half of its maximum levels reached in the early 1980s.

The size of the first-round effects is affected by additional factors, namely, the exchange rate and the profit margins. Oil prices are set in dollars in the international markets. Therefore, developments in oil prices denominated in an alternative currency (the euro, for example) reflect also the behavior of the exchange rate. Concerning profit margins, retailers always have the option to accommodate (partially or totally) the increases in a given input, such as energy, by contracting the profit component of the final price. The behavior of profit margins reflects a wide range of factors, including the cyclical position of the economy or the financial position of the retailers. Table I shows for the G7 countries the very strong correlation between annual rates of change of oil prices, based on national currencies, and the consumer energy components of the national CPIs. Table I also shows that the energy component of the CPI tends to react by slightly more than a ratio of 1:10 to the change in oil price. Combining this sensitivity with the present weight of energy in final household consumption, which tends to be slightly below 10%, it can be seen that inflation reacts by approximately 1:100 to the change in oil price. This rule of thumb gives a rough approximation of the magnitude of the direct first-round effects of an oil price change on inflation.

In addition, price developments are also affected by the so-called second-round effects. As shown in Table II, changes in consumer energy prices tend to produce contemporaneous but also lagging effects on the remaining components of CPI. This is the main reason why inflation measures that exclude the evolution of energy prices—usually referred to as core inflation indicators or measures of underlying inflation—tend to be lagging rather than leading indicators of price developments. Typically, second-round effects are associated with a circular wage–price causality. If workers manage to increase their nominal wages in line with the rise of consumer prices, rather than accepting lower real wages, additional inflation pressures arise, through a wage–price spiral. Therefore, the magnitude of the second-round effects clearly depends on the credibility of monetary policy, which is key to the

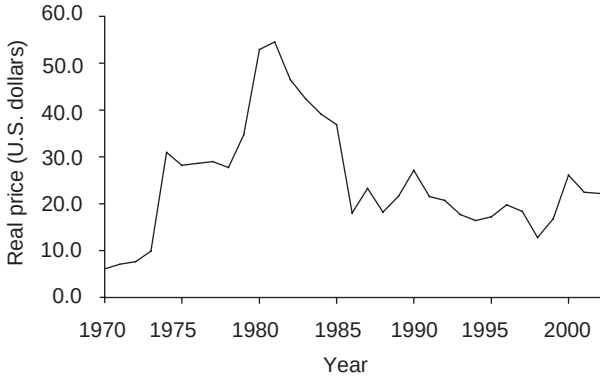


FIGURE 4 Oil prices in real terms (using the U.S. GDP deflator; 1995 = 1). Data from the OECD and Datastream.

TABLE I

Correlations between Oil Prices and Energy Consumer Prices[a]

	Consumer energy prices[b]	
Country	Correlation coefficients	Elasticity
United States	0.756	0.141
Japan	0.765	0.127
Germany	0.661	0.097
France	0.811	0.137
Italy	0.797	0.154
United Kingdom	0.563	0.098
Canada	0.418	0.065
Average	0.682	0.117

[a] Annual rates of growth, 1971–2001. Data from the OECD and Eurostat.
[b] Prices based on national currencies.

TABLE II

Correlation between Energy Prices and Overall Consumer Prices Excluding Food and Energy in period $t+i$[ab]

Country	$t-1$	t	$t+1$	$t+2$
United States	0.343	0.694	0.734	0.503
Japan	0.491	0.707	0.479	0.333
Germany	0.343	0.694	0.734	0.503
France	0.550	0.676	0.708	0.591
Italy	0.506	0.672	0.607	0.493
United Kingdom	0.642	0.877	0.747	0.379
Canada	0.642	0.717	0.658	0.452
Average	0.503	0.720	0.667	0.465

[a] Standard deviation of annual rates of growth, 1971–2001. Data from the OECD.
[b] $i = -1, 0, 1, 2$

formation of inflation expectations. For instance, second-round effects associated with the first oil price shock were larger than the ones associated with the second oil price shock, and this can to a large extent be explained by the different reactions of the monetary authorities to each of the shocks.

Considering the oil price shock of 1973–1974, there is strong evidence that substantial second-round effects took place in many industrial economies. This outcome reflects the combined effects of latent inflationary pressures (already present at the beginning of the 1970s), the large magnitude of the oil price shock, the relatively low flexibility of labor markets, and, last but not least, the accommodative stance of monetary policies in the more advanced economies. It is commonly accepted by economists that a tighter monetary policy was needed, as suggested by the fact that real interest rates became negative in 1974 and remained so until 1978, in a large number of countries. The lack of antiinflationary credibility contributed to unsustainable levels of real wages and thus to marked increases in the unemployment rate.

The experience with the second oil price shock was different. Despite the fact that inflation remained high in the period 1981–1982 (at double-digit figures), it fell significantly in subsequent years. A decisive factor for that evolution was the decline in real wages, in clear contrast to what had happened in the first oil price shock. In addition, the response of monetary policy, characterized by increases in the nominal interest rates, assured that real interest rates were positive and clearly above the levels observed in 1973–1974. Monetary policy was relatively successful in the second oil price shock, as measured by the subsequent reduction of inflation and the moderate evolution of inflation expectations.

3. IMPACT OF OIL PRICE CHANGES ON INFLATION

The focus of this section is on the quantitative assessment of the impact of oil price fluctuations on inflation. For that purpose, the empirical evidence provided by an OECD study is briefly described. The study considered a permanent oil price increase of about $10, vis-à-vis a baseline where oil prices were projected to remain at the $20–25 level (this shock corresponds to a price increase of about 40–50%). The OECD study presents estimates of the impact on GDP, world trade growth, and inflation, for the G7 countries. The estimates were obtained through the simulation of the OECD macroeconometric model Interlink, assuming fixed nominal exchange rates, real interest rates kept at baseline levels, and unchanged real government expenditures. The main results of this exercise, with respect to the impact on inflation, are reported in Table III.

Considering the average results for G7 countries, the impact of oil prices in the first year is slightly above the 1:100 rule of thumb, corresponding basically to the direct first-round effects, considering also some initial indirect effects. Effects in subsequent years are, naturally, smaller, corresponding to some lagging indirect first-round effects and to the second-round effects. As mentioned previously, the magnitude of the second-round effects depends on the assumptions on the reaction functions of the monetary authorities (constant real interest rates in this simulation).

Discrepancies in the estimates for different countries, Italy and Canada in particular, illustrate the difficulties associated with this type of exercise, owing to several factors. In first place, the transmission channels underlying the effects are extremely complex and are very likely to change over time. A macroeconometric model constitutes, in general, a very simplified representation of the real world and therefore does not capture precisely the functioning of the economy. Second, econometric simulations tend to replicate the average behavior of the economy since the early 1970s, despite the marked changes that took place in key variables, such as the oil intensity of the G7 economies or the weight of indirect taxation in final consumer prices. Finally, the estimates of the second-round effects associated with the change in oil price will depend on the different assumptions on the reaction of the monetary

TABLE III

Oil Price Effect on Inflation[a]

	Years after the shock		
Country	1	2	3
United States	0.5	0.2	0.1
Japan	0.4	0.0	0.0
Germany	0.4	−0.1	−0.2
France	0.6	0.1	0.1
Italy	1.0	0.2	0.0
United Kingdom	0.4	0.3	0.4
Canada	0.1	−0.1	0.2
Average	0.5	0.1	0.1

[a] A 50% rise. Data from Dalsgaard *et al.* (2001).

authorities, i.e., the credibility of monetary policy, and on how economic activity is affected. There is no doubt that there is an inverse relationship between oil prices and aggregate economic activity. Among the several channels through which oil price movements affect economic activity, the supply-side effect is probably the most common one. Rising oil prices, signaling reduced availability of a basic input to production, reduce potential output and the natural rate of unemployment, simultaneously explaining a negative relationship between oil price and economic activity and a positive relationship between oil price and inflation. However, a lot of uncertainty remains about the quantification of these effects. For instance, recent research indicates that there is a nonlinear relationship between oil price and real activity. In particular, the evidence indicates that oil price increases tend to produce stronger effects than do oil price decreases, and that increases after a period of stable prices tend to produce larger effects, as compared to increases that correct previous declines in prices. This type of result is also likely to hold when assessing the relationship between oil price changes and inflation.

4. OIL PRICE VOLATILITY

The volatility in oil prices since the early 1970s is a remarkable feature of energy economics. Annual fluctuations in the oil price level, as measured by the absolute value of year-to-year price changes, averaged only 1% in the 1949–1970 period; from 1970 to date, these fluctuations increased dramatically, reaching an order of magnitude of 30% per year. Even in the relatively stable period from 1986 to 1997, oil prices were more volatile than other primary commodities.

There is a very large set of reasons behind oil price volatility, including seasonal changes in demand; shifting demand conditions, in line with world GDP fluctuations; shocks in supply conditions that routinely affect, on a regular but rotating basis, some of the main oil producers (Iraq and Venezuela, in the recent past, for instance); and difficulties within OPEC in exercising its market power. The problems experienced by OPEC are due both to limitations of the knowledge of all the relevant aspects in the functioning of the market and to difficulty in maintaining cooperation between cartel members; these underlying problems have sometimes contributed to undesirable swings in the price of oil, as occurred in 1998–2000.

4.1 Oil Price Volatility as an Important Source of Inflation Changes

The contribution of energy prices to the volatility of the CPI is considerable. This point is illustrated in Table IV for the G7 countries. Despite the previously mentioned decline in oil prices in real terms, the average increase in consumer energy prices was slightly more than the overall increase in consumer prices. This evolution reflects an increase in taxation of energy products, which has been commonly justified for ecological reasons. At the same time, and this is the important aspect to stress at this stage, energy prices exhibited a volatility that was more than double the observed volatility in total consumer prices. Therefore, oil price changes are extremely important for the inflation level, because energy consumption exerts a weight slightly below 10% on total consumption, and even more important for fluctuations of inflation.

4.2 How to Forecast Oil Prices?

The evolution of oil prices is typically subject to a very high degree of uncertainty, given the extremely volatile nature of conditions that affect prices. Information on the future of oil prices is, however, extremely important for market operators as well as for central banks. Central banks, in particular, have a forward-looking perspective, attaching a very important role to inflation forecasts. This uncertainty tends to reinforce the importance of the relationship

TABLE IV

Oil Prices and Inflation[a]

	Average		Standard deviations	
Country	Total consumer prices	Consumer energy prices	Total consumer prices	Consumer energy prices
United States	5.1	5.8	3.1	9.4
Japan	3.8	3.3	5.0	9.3
Germany	3.3	4.5	2.0	7.4
France	5.8	6.7	4.3	9.5
Italy	8.9	9.0	6.0	11.1
United Kingdom	7.6	8.0	5.7	9.2
Canada	5.3	7.1	3.5	7.6
Average	5.7	6.3	4.2	9.1

[a] Using annual rates of growth, 1971–2001. Based on authors' computations using OECD data.

between oil prices and inflation, independently of estimates of the effects of a permanent change in oil prices on inflation. Two alternative technical assumptions on oil prices are worth mentioning:

1. The carryover assumption assumes that oil prices will remain constant; as in other markets (the exchange market provides a clear illustration), this hypothesis is justified by the fact that short-term changes in oil prices are statistically close to a pure mean zero random variable.
2. The futures market assumption assumes that the path of oil prices will follow the prices implicit in the agreements for future oil transactions.

Unfortunately, both techniques have produced considerable errors in past assessments, contributing to wrong evaluations of future inflation developments. On one hand, the carryover assumption reflects the idea of ignorance. If it is not possible to know if oil prices will increase or decrease, this assumption corresponds to putting the speculation somewhere in the middle, assuming that oil prices will remain constant. Obviously, this assumption will not account for the short-term volatility of oil prices and for the mean reversion process that has characterized the evolution of oil prices since the mid 1980s. On the other hand, futures markets tend to anticipate this mean reversion property, in the sense that they tend to converge, in a 1-year horizon, toward a range of $20–25 per barrel. However, the track record of futures on oil prices from 1995 onward has been rather poor (Fig. 5). In the short run, futures prices quite often missed the direction of change in oil prices. Additionally, the speed of adjustment and the mean level underlying the mean reversion process characterizing oil price behavior do not seem to be constant.

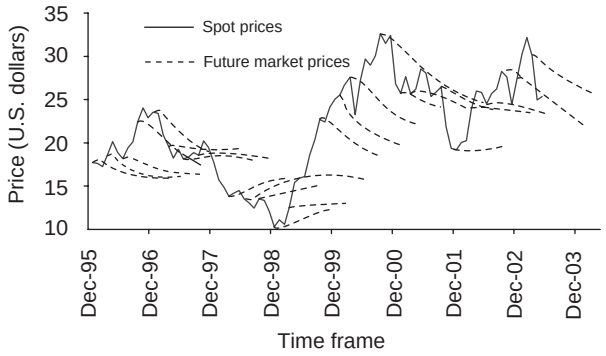


FIGURE 5 Spot prices versus future market prices (monthly figures, December 1995 through March 2003). Data from the International Petroleum Exchange.

An adequate measure of the degree of uncertainty attached to the evolution of oil prices in the near future is provided by the implicit volatility of the distribution of contracts. In that respect, it may also be interesting to analyze the role of speculators in energy futures markets. This has received a great deal of scrutiny in recent years, even though recent research suggests that the speculators, an important group among the commodity-fund managers, do not seem to be the driving force behind the volatility often observed in the crude oil market.

SEE ALSO THE FOLLOWING ARTICLES

Business Cycles and Energy Prices • Economics of Energy Demand • Economics of Energy Supply • Energy Futures and Options • Innovation and Energy Prices • Oil Price Volatility • OPEC Market Behavior, 1973–2003 • Prices of Energy, History of • Stock Markets and Energy Prices

Further Reading

Bank of England. (2000). What do the recent movements in oil prices imply for world inflation? *Q. Bull.* (May 2000), 147–149.

Brown, S. P. A., and Yücel, M. K. (2002). Energy prices and aggregate economic activity: An interpretative survey. *Q. Rev. Econ. Finance* **42,** 193–208.

Bruno, M., and Sachs, J. (1985). "Economics of Worldwide Stagflation." Harvard University Press, Cambridge, Massachusetts.

Dalsgaard, T., André, C., and Richardson, P. (2001). "Standard Shocks in the OECD Interlink Model." Organization for Economic Cooperation and Development (OECD) Economics Department Working Papers, No. 306. OECD, Paris.

European Central Bank (ECB). (2000). *ECB Econ. Bull.* (November 2000), 21–26.

Hamilton, J. D. (2002). What is an oil shock? *J. Econometr.* **133,** 363–398.

Hooker, M. (1986). What happened to the oil price–macroeconomy relationship? *J. Monetary Econ.* **38,** 195–213.

Kohl, W. L. (2002). OPEC behavior, 1998–2001. *Q. Rev. Econ. Finance* **42,** 209–233.

Marques, C. R., Neves, P. D., and Silva, A. (2002). Why should central banks avoid the use of the underlying inflation indicator? *Econ. Lett.* **75,** 17–23.

Smith, J. L. (2002). Oil and the economy: Introduction. *Q. Rev. Econ. Finance* **42,** 163–168.

Stuber, G. (2001). The changing effects of energy-price shocks on economic activity and inflation. *Bank Can. Rev.* (Summer 2001), 3–14.

Weiner, R. J. (2002). Sheep in wolves' clothing? Speculators and price volatility in petroleum futures. *Q. Rev. Econ. Finance* **42,** 391–400.

Information Technology and Energy Use

KURT W. ROTH
TIAX LLC
Cambridge, Massachusetts, United States

1. **The Direct Impact of Information Technology on Energy Consumption**
2. **Indirect Energy Consumption**
3. **Conclusions**

Glossary

active power mode Usage mode when equipment is performing its primary function (e.g., copier printing).
computer network equipment Equipment central to the operation of computer networks, including server computers, uninterruptible power supplies (UPSs), local area network (LAN) switches, routers, hubs, and wide area network (WAN) switches.
indirect impacts Energy consumption influenced by, but not directly attributed to, information technology (IT) equipment (e.g., the energy required to manufacture a copier).
office equipment Equipment typically deployed in offices to enhance productivity (e.g., personal computers, monitors, copiers, printers, facsimile [fax] machines).
off mode Usage mode when equipment is turned off but still plugged in (e.g., a copier manually turned off).
power management-enabled (PM-enabled) rate The percentage of information technology (IT) equipment that enters lower power draw modes after periods of nonuse.
stand-by mode Usage mode when the equipment is ready to perform, but is not actually performing, its primary function (e.g., copier fuser roll warm and ready to print).
suspend mode Usage mode when the equipment has powered down and cannot directly transition to active mode (e.g., copier transitioned to low-power mode with cold fuser roll).
telecommunications network equipment Equipment central to the operation of telecommunications networks, including: wireless base stations, fiber-optic terminals, analogue phone network equipment, private branch exchanges (PBXs), and wireless phones.

Information technology (IT) equipment includes devices that produce and process information (e.g., office equipment) and that transmit information electronically (e.g., computer and telecommunications network equipment). IT equipment directly consumes energy by drawing power. Analyses estimate that IT equipment directly accounts for approximately 3% of annual U.S. electricity consumption or a bit more than 1% of annual U.S. primary energy consumption. To clarify, primary energy includes the energy consumed to generate, transmit, and distribute electricity to end-use sites. For example, the average kilowatt-hours (kWh) of site electricity in the United States in 2000, equal to 3413 Btu of energy, required 10,775 Btu of source energy. (This Energy Information Administration estimate takes into account the electricity production mix for the entire country.) The nation's hundreds of millions of devices also have several potential indirect impacts on energy consumption. IT equipment can influence the national energy intensity (i.e., energy consumption per dollar of gross domestic product [GDP]) due to productivity changes, changes in work patterns (e.g., telework and telecommuting, e-commerce), and shifts in the national economy. On a local level, IT equipment affects peak electricity demand, whereas on the building level, it alters the energy consumption of heating and cooling systems. The manufacture of IT equipment also consumes energy, as does the manufacture of the paper used by printers, copiers, and facsimile (fax) machines.

1. THE DIRECT IMPACT OF INFORMATION TECHNOLOGY ON ENERGY CONSUMPTION

Information technology (IT) equipment (Table I) annual energy consumption (AEC) depends on three

 © 2004 Elsevier Inc. All rights reserved.

TABLE I

Major Energy-Consuming Information Technology Equipment

Office equipment	Network equipment
Copy machines	Analogue phone network
Facsimile machines	Fiber optic terminal
Monitors and displays	Local area network switches
Personal computers	Server computers
Printers (laser and inkjet)	Routers
	Server computers
	Uninterruptible power supplies
	Wireless telephony base stations

factors: equipment operating patterns, power draw by mode, and the installed base (Fig. 1). As a result, most IT equipment AEC estimates are bottom-up estimates that quantify each of the three factors. The operating pattern represents the number of hours spent per year in various modes of operation, each with a different power draw level. Typically, office equipment has at least three modes, as illustrated in Table II.

Unit energy consumption (UEC) for a given device equals the sum of the product of hours in each operating mode and the average power draw for each mode. Because the active power mode typically has much higher power draw levels than the other modes, the number of hours spent in active mode has a major impact on the UEC. Two factors, the night status and the power management-enabled (PM-enabled) rate, determine the number of hours spent in active mode for most kinds of office equipment. Weekend and night hours account for a sizable majority of potential operating hours, so night audits of commercial buildings are performed to determine the approximate operating patterns of office equipment during these periods. Many types of office equipment have PM capability, by which the equipment enters the low-power suspend mode after a prescribed period of inactivity. Clearly, the PM-enabled rate influences the daytime and night/weekend operating patterns of various equipment and is best derived from audits of commercial buildings. More accurate UEC estimates need to consider the actual (i.e., measured) power draw in each mode

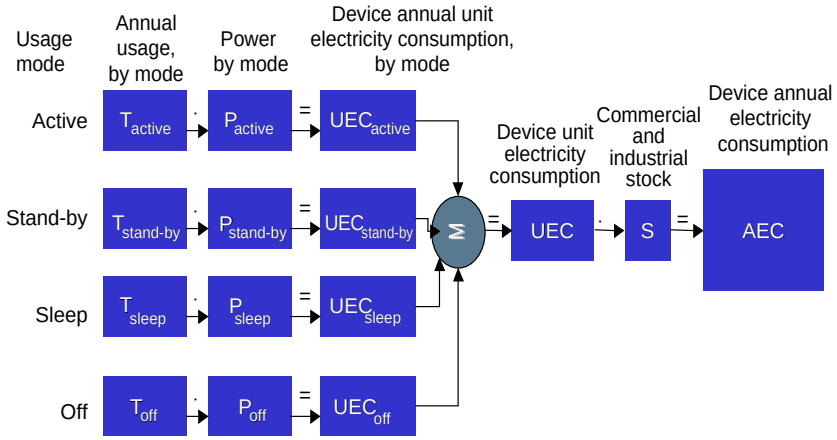


FIGURE 1 Bottom-up annual electricity consumption calculation methodology. From Roth *et al.* (2002).

TABLE II

Office Equipment Usage Modes

Mode type	Description	Examples
Active	Device carrying out intended operation	Monitor displays image; copier printing
Stand-by	Device ready to out intended operation (but not doing so)	Monitor displays screen saver; copier ready to print
Suspend	Device not ready to carry out intended operation (but on)	Monitor powered down but on; copier powered down but on
Off	Device not turned on (but plugged in)	Monitor off, plugged in; copier off, plugged in

instead of the nameplate power rating given that the active mode power draw for most IT devices is typically less than 50% of the nameplate value. The installed base of various types of IT equipment comes from either industry studies of equipment, market saturation data, or estimates based on sums of annual device sales over equipment lifetimes.

Circa 2000, nonresidential IT equipment energy consumption was roughly six times greater than residential IT equipment energy consumption. Consequently, this section focuses on nonresidential IT equipment and energy consumption.

1.1 Commercial and Industrial IT Equipment

In 2000, nonresidential IT equipment consumed just under 100 TWh Table III. Personal computers and their monitors represented just over 40% of the total AEC (~42 TWh site or 0.46 quads). The equipment forming the backbone of the Internet (server computers, computer networks, telephone networks, and uninterruptible power supplies [UPSs]) consumed approximately 30% of all nonresidential IT equipment electricity (~30 TWh site or 0.33 quads primary). The following subsections discuss the energy consumption of the major equipment types in further detail.

TABLE III

U.S. Nonresidential IT Equipment Annual Electricity Consumption Summary

Device type	AEC (TWH)	Percentage of AEC
Communications networks	30	31
Server computers[a]	11.6	13
Telephone network equipment[b]	6.6	7
Computer network equipment[c]	6.4	7
UPSs	5.8	6
Monitors and displays	22	23
Monitors	18.8	19
General displays	3.4	4
PCs[d]	20	20
Imaging devices	15	16
Copiers	9.7	10
Printers	5.7	6
Other	10	10
Total	97	100

Source. Roth *et al.* (2002).

[a] Includes data storage.

[b] Includes cell site equipment, transmission (fiber-optic), PSTN, PBXS, wireless phones.

[c] Includes LAN switches, routers, hubs, WAN switches, Modems/Remote Access Server (RAS), cable modem termination systems (CMTS).

[d] Includes desktop personal computers, workstations, and laptop/notebook personal computers.

1.1.1 Personal Computers and Workstations

The installed base of approximately 71 million personal computers and 2.5 million workstations in 2000 consumed 17.4 and 1.8 TWh, respectively. In most desktop PCs, the chipset, central processing unit, memory, and AC/DC power supply draw the most power in active mode, with an average 2000 vintage desktop PC drawing 55 to 60 W in active mode. More recent PCs can draw much more power in active mode, sometimes in excess of 100 W, particularly when configured with powerful graphics cards. On the other hand, the steadily growing notebook (laptop) computer segment typically draws 15 to 25 W total (including the 14- or 15-inch LCD monitor) and employs more sophisticated PM schemes in a quest to increase "unplugged" running times. Large quantities of PCs left on at night (~50% based on limited night audits) and low PM-enabled rates (~25%) are responsible for the majority of PC annual energy consumption. The past incompatibilities between common PM schemes and operating systems that caused the systems to crash have been resolved, creating the potential for higher future PM-enabled rates. On the other hand, PC network cards remain a problem in that entering suspend mode can cause the PC to drop the network connection. To avoid the inconvenience of disconnection from the network, some users disable PM features.

Future PC energy consumption will depend on increases in microprocessor and graphics card active power draw, changes in PM-enabled rates, and transitions to notebook (and smaller) computing devices.

1.1.2 Monitors and Displays

Most of the nation's approximately 61 million monitors are associated with desktop PCs and workstations, whereas some 13 million displays rely on video input (as seen in airports). Together, they consumed approximately 22 TWh of electricity in 2000. That same year, cathode ray tube (CRT) monitors accounted for more than 95% of the display market due to their low cost, with liquid crystal display (LCD) devices accounting for most of the remaining market. A CRT converts electrical

signals to the visual display seen on the screen, using an electron gun to emit a beam of electrons and project them onto a screen. Anodes accelerate the electrons, which are then steered by a varying electromagnetic field onto different parts of the screen, where they interact with a coating of phosphor compounds that convert the electron beam into an image. In essence, the CRT "paints" an image on the phosphor layer. In contrast to monochrome displays, which use a single electron gun, color monitors use three separate electron guns (red, blue, and green) to create color images.

Most CRT monitors (17 inches is the most common size) draw from 70 to 90 W in active mode, whereas similarly sized LCDs draw approximately 60% less power. Monitors have significantly higher PM-enabled rates than do PCs (~60%), resulting in monitors spending fewer hours in active mode. Recent growth in the market share of LCD monitors (to ~20% worldwide in 2002) and laptop computers (~25% of global PC sales), as well as increases in PM-enabled rates, will tend to decrease net monitor energy consumption, whereas further increases in screen size will increase energy consumption.

1.1.3 Copy Machines

Commercial buildings house approximately 9 million copy machines that consumed nearly 10 TWh of electricity in 2000. Analogue and newer digital copy machines all function in the same basic manner. The machine captures the image, either via a light source (analogue) or via a scanner (digital), and transfers it to the hot toner drum, which then transfers it in the form of toner to paper sheets. Copy machines come in a wide range of bands (speeds), ranging from several copies per minute (cpm) for desktop units to approximately 100 cpm for high-performance units.

Copy machines have high UEC values due primarily to the high power draw during copying (typically in excess of 1000 W) and the high stand-by power levels (hundreds of watts) needed to maintain fuser rolls in a hot ready state. In fact, the stand-by (hot fuser roll) mode typically accounts for most copy machine energy consumption, whereas active copying mode accounts for only a fraction of the consumption. Copiers wait substantial time periods before powering down into suspend mode because reheating a cooled fuser roll can take several minutes, which is often an unacceptably long delay in the fast-paced office environment. Advances in copier fuser systems, including toner materials with lower melting temperatures, could substantially decrease copier energy consumption in active and stand-by modes by decreasing the fusing temperature, which would reduce the amount of energy needed to keep the fuser rolls hot. Lower fuser roll temperatures also would decrease the warm-up time, perhaps to only a few seconds, enabling copiers to spend more time in the low-power suspend mode. The gradual integration of copier and printer function into a single device will also affect future copier and printer energy consumption.

1.1.4 Printers

Three printer types—laser, inkjet, and impact—constitute most of the printer market and consume nearly 6 TWh of electricity per year, with laser printers accounting for approximately 80% of total printer AEC. Typically, a laser printer is a shared resource among several users in a computer network, whereas inkjet printers may serve as personal printers or the sole printers in smaller businesses. Impact printers are generally used in the "back office" as part of order fulfillment systems where it is necessary to print information onto multiple copies of forms via carbon copies.

Although inkjet printers dominate the residential market, laser printers remain the printer of choice for offices due to their high print quality and printing rates. Laser printers consume much more energy than do inkjet devices primarily because the fuser rolls must remain at high temperatures to bond the toner to the paper quickly and completely. During printing, the laser printer actively supplies resistance heat to ensure effective bonding. In addition, laser printers in stand-by mode require heat to avoid delays in response to a print request. Most laser printers have PM capabilities, but surveys suggest that only perhaps half of laser printers are PM enabled.

If color laser printers become faster and their prices continue to decrease, their presence in the office could increase significantly in the future. This would challenge the inkjet market to find another niche to fill. On the other hand, the development of higher speed and quality inkjet printers could supplant much of the laser printer market and lead to an overall reduction in printer energy consumption.

1.1.5 Server Computers

In today's usage, a server generally refers to a computer that is not directly associated with a specific human user. Instead, servers provide common functions to a group of users or perform back-end processing invoked on a scheduled basis or by other computers. Servers come in a wide range of

sizes. At the upper end are the traditional mainframes, repositioned as high-end servers. The market segment formerly called minicomputers is now classified as midrange or workhorse servers. Server functions run the gamut of computing tasks, with the exception of those involving direct human interaction. High-end servers often carry out traditional batch processing applications (e.g., billing) or fulfill high-volume transaction processing applications (e.g., banking, airline reservations). Midrange servers typically handle database applications and can be the main back-end computers for medium-sized businesses. These larger systems feature superior reliability relative to PC-derived systems and often require large storage and input/output (I/O) handling capacity. Low-end servers typically take on tasks that require smaller storage and/or lower criticality (e.g., local area network [LAN] file and print management). Many perform as Web servers at large hosting centers (also known as server farms, data centers, or colocation facilities) operated by Internet service providers (ISPs). Other common roles for low-end servers include e-mail or lower volume specialized applications.

In general, CPUs and hard drives, as well as power supplies, account for the bulk of server power draw, with actual power draw depending significantly on the server class (Table IV) and configuration. Additional peripherals, primarily data I/O devices, play a key role in the functioning of server computers, particularly in high-end machines. There are two types of data storage: optical and disk systems. Optical storage is used for archiving data and involves only the writing of data to the optical storage disks. Disk storage typically interacts (reading and writing data) more frequently with outside systems and must be constantly accessible. Optical and disk data storage devices consume approximately 0.9 and 0.7 TWh per year, respectively.

1.1.6 Uninterruptible Power Supplies

Although UPSs are not IT equipment per se, they play an increasingly important role in ensuring its reliability. UPSs ensure the continuous flow of high-quality power to critical IT equipment. They provide power quality to eliminate transient spikes or sags in power that could adversely affect the performance or could damage electricity-consuming equipment. In case of power service failure, they also provide sufficient backup power to last through the outage, shut down the electronic device, and/or bring back up power generation online.

Battery chargers/rectifiers, stand-by batteries (usually lead acid), and inverters are the primary components of most UPSs. UPS applications relevant to IT equipment range from small (1 kVA) stand-by devices for individual PCs or workstations to very large (>100 kVA) online UPSs for data center applications. Overall, IT-related UPSs represent approximately 70% of UPS sales and consume nearly 6 TWh annually.

There are three major types of UPSs: stand-by, online, and line interactive. Stand-by and line-interactive systems have similar operating principles; relative to stand-by systems, line-interactive architectures typically offer superior PM because they have a live inverter. In standard operational mode, both architectures allow power to pass through the UPS and into the electronic device while providing varying degrees of power conditioning, typically surge suppression to counter voltage spikes and a filter to reduce unwanted harmonics. If the UPS battery has run down, the UPS battery charger will charge it. When power ceases flowing to the device, the UPS detects the lack of electric power and the transfer rapidly establishes (clamps) an electrical connection between the device and the battery power source to enable continued operation. Line-interactive and stand-by systems typically consume only a

TABLE IV

U.S. Server Computer Characteristics circa 2000

	Server class			
Attribute	Low end $<$\$25,000	Workhorse \$25,000–\$100,000	Mid-range \$100,000–\$1 million	High end $>$\$1 million
Installed base (thousands)	4,050	575	185	16.5
Typical active power draw (W)	125	650	1225	2500
AEC (TWh)	4.5	3.3	2.0	0.4

few percent of the electronic device load because in default operational mode they allow power to pass through to the load with minimal PM. The ferroresonant UPS is a notable exception in that it generates significant quantities of heat due to inefficiencies in the transformer.

In contrast, online UPSs accept AC power, which they condition and rectify (convert from AC to DC). Subsequently, the DC output charges the batteries and then passes into an inverter, which converts the DC power to well-conditioned AC power. If the input power fails, the battery continues to supply power to the inverter and provide high-quality power to the load without interruption because it has no transfer switch. Online systems offer high reliability but cost more than other UPS systems and are relatively inefficient, typically dissipating 10 to 15% of the power flow in heat when operating near the rated capacity. A new UPS architecture enables the input power to largely bypass the AC–DC–AC conversion process. Instead, the UPS primarily supplies the difference between the input and load powers, resulting in significantly higher efficiencies than is the case with double conversion systems (typically dissipating less than 5% of power flow near the rated capacity).

UPS efficiency also depends on the actual load seen by the UPS relative to the rated load. UPSs have a basic power demand (for fans, power supplies, etc.) that establishes a "base" load. However, the electric components are sized for the full UPS power rating and dissipate less heat at smaller loads. The net result is that the overall UPS efficiency curve is relatively flat when the load falls between 50 and 100% of the off rated load but often drops off below 50% and falls off precipitously under 20%. In general, UPSs are oversized, often designed for 70 to 80% load on larger systems and typically operating at 50% or less of design load. Several reasons for this practice exist, including general design conservatism, allowances for future expansion, greater redundancy, longer load runtimes after power failure, and parallel redundant systems for maintenance purposes. Circa 2002, many data centers had low occupancies, causing UPS loads to fall well short of expected loads. As a consequence, many UPS systems are likely see less than 25% of their rated loads and operate at low efficiencies.

1.1.7 Computer Network Equipment

Computer networks use a range of equipment that collectively connects the global Internet with millions of desktop and server computers. This equipment is generally divisible into two categories: LAN (within a building or campus) and wide area network (WAN, beyond the campus). LAN gear interconnects desktops and servers using private bandwidth, whereas WAN gear generally uses common carrier facilities to provide external access, either within a company (intranet), between companies (extranet), or to the global Internet.

In general, computer network equipment power draws vary little with data throughput rates, as devices always remain "hot" to respond to surges in traffic. In spite of round-the-clock operation, the total electricity consumed per year by all computer network equipment is only approximately 6 TWh per year (or ~6% of nonresidential IT energy consumption). LAN switches account for perhaps half of the total. Specialized processors and AC/DC power supplies are the major energy-consuming components of all computer network equipment except passive hubs.

1.1.7.1 LAN Gear: Hubs and Lan Switches In buildings, desktop computers are generally connected via short wiring to some kind of hub. Whether this is a simple hub or a LAN switch, it is often a rack-mounted device supporting anywhere from as few as 4 to a few hundred ports in a single rack. Functionally, hubs provide a physical connection between the local network and other devices, such as computers and printers, with most hubs providing signal conditioning in addition to connectivity (in contrast to the passive hubs of the past). LAN switches are more sophisticated devices that provide multiple paths between inputs and outputs, so that traffic between two nodes (e.g., one desktop and one server) does not affect traffic between other nodes. These higher function switches provide additional security features typically performed by routers, but they are scaled down to be cost-effective on the LAN. As the price approaches levels once occupied by simple hubs, the market is moving toward greater acceptance of higher layer switching.

1.1.7.2 Routers Routers have a basic set of primary functions, including examining, filtering, and routing incoming data packets. Routers examine each incoming packet and determine where to send it based on its address. Large high-capacity routers with high-speed interfaces (e.g., OC-48 or greater) serve major backbone routes, whereas subscriber connections tend to use large numbers of lower speed interfaces (e.g., DS-1 to DS-3). At subscriber locations, a small edge router usually provides the

connectivity. Corporate users and smaller ISPs typically have mid-sized routers.

1.1.7.3 WAN Switches Like routers, WAN switches help to manage the flow of large amounts of data between locations. They differ in their use of connection-oriented telecommunications protocols, with some hybrid switch/router products providing both functions. WAN switches tend to be large devices used by phone service carriers, ISPs, and/or high-end enterprise customers.

1.1.8 Telecommunications Network Equipment

Telephone networks differ from computer networks in that they were designed primarily to carry voice information instead of data. Four primary categories of telephone switching equipment exist. The transmission network includes long-distance fiber-optic connections among major cities and locations handling long-distance telephone calls. The mobile telephone network includes the base stations (in towers) that connect mobile telephones to the larger phone network. The public switched telephone network (PSTN) denotes the established copper wire connections between telephone companies' central offices and residences and buildings. Lastly, the private branch exchange (PBX) is a private in-house phone system, typically found in larger office buildings or campus settings.

Although telephone systems were originally designed to carry voice traffic, the distinction between voice and data traffic has become increasingly blurred since the advent of the Internet. For example, local public telephone networks originating from central offices now provide Internet access via dial-up modems and digital subscriber lines (DSL), and the long-distance fiber-optic transmission networks originally installed to handle voice traffic now carry Internet data. Mobile phone networks have begun to offer basic Web access, notably in Japan and Europe.

In 2000, telephone network equipment consumed approximately 6.6 TWh of electricity (or just under 7% of all electricity consumed by nonresidential IT equipment).

1.1.8.1 Mobile Telephone Networks Wireless (also known as cellular) phones became commonplace during the 1990s. Cell sites contain base station radio transmission gear (analogue and/or digital) that transmits and receives signals to and from cellular telephones, with the number of transmitters and their power dependent on the cell size, location, and chosen radio technology. The transmission gear is located as close as possible to the antenna (e.g., in a hut at the foot of a tower). Cell sites are characterized by their size (i.e., broadcasting power and effective receiving and transmission radius), which ranges from large macro cells to small pico cells. Transmission gear power draws vary significantly with the amount of traffic passing through the site, although the equipment does require a base load to keep it hot (i.e., ready to transmit information). Peak power draws depend greatly on the size of the installation, ranging from close to 10 kW for macro cells to as little as a few hundred watts for pico sites.

The approximately 100,000 cell sites in the United States in 2000 consumed more than 2 TWh, representing the largest component of telephone network energy consumption (this does not include additional energy used to cool the sites). The rollout of third-generation wireless networks with higher power draw and the expected uptick in geographical coverage will likely increase the average power draw at the base station as well as the number of base stations, making it nearly certain that base station energy consumption will increase.

Battery-powered, compact, and designed to maximize time of operation without recharging, cellular phones consume energy at low levels (on the order of a watt). Roughly 28 million nonresidential cellular phones, representing approximately 30% of the total stock of cellular phones, consumed close to 0.5 TWh in 2000.

1.1.8.2 Fiber-Optic Terminals The original long-distance transmission networks used copper wire, with microwave towers entering long-distance service during the 1950s. Fiber conversion began during the 1980s, buoyed by the enhanced signal quality and much higher bandwidth of fiber service compared with microwave service. Most long-distance traffic now passes through fiber. The current U.S. fiber network consists of numerous trunk routes between major hubs in each state, with fiber beginning to supplement and replace copper wire to major facilities, such as office buildings, to take advantage of its higher voice and data bandwidth.

Fiber-optic communication systems transmit data by converting electronic signals, such as voice signals and data packets, into laser-generated light pulses. The light pulses pass through a glass fiber wrapped in plastic (multiple fibers combined together make a fiber-optic cable) to a receiver, where they are translated back into electronic signals. Light signals that pass through long distances require an amplifier to maintain signal quality and intensity. As the

quality of fiber-optic devices has improved, the distance required between amplifiers has grown (to ~80 km in 2000).

Fiber terminals, which send, receive, and multiplex data (e.g., to a T1 connection), dominate the energy consumption of fiber-optic networks. Drawing roughly 200 W per terminal, the approximately 1 million terminals in 2000 consumed nearly 2 TWh of electricity. Continuing increases in network traffic, coupled with growth in fiber connectivity (to commercial buildings and potentially to homes), will likely increase fiber terminal energy consumption in the future.

1.1.8.3 Public Switched Telephone Network The PSTN refers to the established copper wire connections between telephone companies' central offices and residences and buildings. Analogue phone lines to residences and businesses dominate energy consumption of the PSTN, accounting for roughly 165 million of the more than 190 million phone lines installed in the United States in 2000. Central office switches typically power analogue phone lines, but the analogue line components that account for much of the switch power draw only draw power on establishing a connection between two phones (i.e., when a call is completed). Other switch elements that continuously draw power include common equipment (one administrative module per switch) and line equipment (e.g., concentration modules, trunk ports). The sheer variety of switching equipment in vintage, scale, and complexity makes it difficult to characterize the installed base. Furthermore, there is a dearth of publicly available power draw data. Together, these two factors complicate development of an accurate estimate of total energy consumption. Nonetheless, because each phone draws an average of approximately 3 W per active line, the energy consumed by analogue phone lines appears to be quite small (on the order of 1 TWh per year).

1.1.8.4 Private Branch Exchanges PBXs are private in-house phone systems, typically found in larger office buildings or campus settings. Commercial phone systems, such as those made by Lucent and Nortel, drive the energy consumption of PBXs, which operate round-the-clock and consume approximately 1 TWh per year.

1.1.9 Other Equipment Types

Other IT equipment accounts for approximately 10% of nonresidential IT energy consumption. Facsimile (fax) machines (3 TWh) and point-of-service terminals (1.5 TWh) consume the most energy. Other equipment types consume smaller quantities of energy: automated teller machines, desktop calculators (adding machines), dictation equipment, scanners, small handheld devices (e.g., personal digital assistants), supercomputers, typewriters, very small aperture terminals (VSATs), and voice mail systems.

1.2 Residential IT Equipment

As noted earlier, residential IT equipment energy consumption appears to be roughly six times less than nonresidential IT energy consumption, with desktop PCs and monitors accounting for the majority. The actual AEC of residential IT equipment has a very high uncertainty, reflecting poorly understood usage patterns. Nonetheless, there are several reasons to believe that residential IT energy consumption is significantly less than nonresidential IT energy consumption. Most residential IT equipment almost certainly has significantly lower usage (in active mode) due to the length of the business day relative to residential usage patterns. Furthermore, residences have much smaller installed bases of large energy-consuming devices, notably server computers, telephone network equipment, copiers, many types of computer network equipment, and laser printers. Finally, the more energy-intensive (and typically larger) IT equipment tends to be deployed in nonresidential settings, for example, faster copy machines and laser printers, larger UPSs, core routers, and LAN switches.

The potential exists for residential IT equipment energy consumption to grow rapidly in the future. PCs (and associated monitors and printers) have yet to achieve residential saturation equal to the level of the commercial sector, meaning that there is room for growth. Similarly, as the percentage of households connected to the Internet continues to increase from just over 50% circa 2001, increases in residential IT usage will follow. The fact that only approximately 10% of Internet subscribers in 2001 had broadband (primarily cable modems and DSL), which requires modems and tends to increase PC usage (to support "always-on" operation), represents another growth opportunity. Finally, the long-awaited arrival of widespread home networking and its associated infrastructure could increase future residential IT energy consumption.

2. INDIRECT ENERGY CONSUMPTION

2.1 Overview: Effect on National Energy Intensity

IT equipment indirectly affects energy consumption in several ways. For example, IT-enabled process controls enhance production efficiency, whereas IT-enabled telework and telecommuting (T&T) changes building usage and personal transportation patterns. The demand for IT equipment results in additional energy consumption to manufacture IT equipment. Together, these and many other facets of IT affect the macroeconomic energy intensity of the national economy (energy consumed per dollar of gross domestic product (GDP). Because of the number of factors at work and the sheer complexity of economic systems, quantifying the indirect energy impact of IT is significantly more difficult than understanding the direct energy impact. Nonetheless, evidence strongly suggests that the magnitude of the indirect impacts is comparable to—and likely exceeds—the magnitude of the direct impact.

Some of the most compelling evidence for this claim comes from energy intensity data from the late 1990s, when the Internet was experiencing widespread commercialization. During this period, the energy intensity of the U.S. economy began decreasing at a sharper rate, approaching decline rates characteristic of the early and mid-1980s, a time of high energy prices (Fig. 2). Notably, the sharp decrease in energy intensity, which coincided with a significant increase in economic productivity, occurred despite historically low (in real terms) energy prices, leading many to ascribe much of the sharp decrease in energy intensity to IT. Because the steeper energy intensity decrease began only recently, it is premature to draw a solid conclusion. If IT-based innovations do accelerate the long-term decrease in energy intensity, the impact of IT equipment on national energy consumption could be significantly larger than the 1% of energy directly consumed by such equipment.

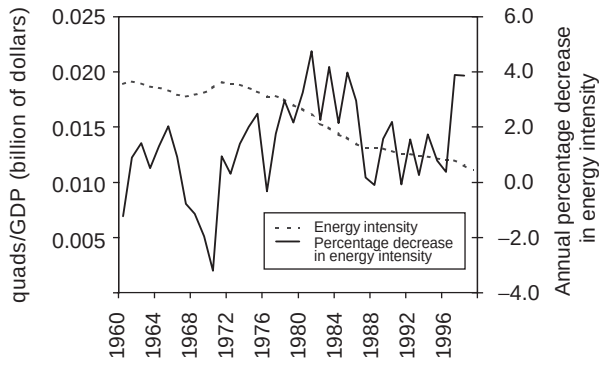


FIGURE 2 U.S. energy intensity and energy intensity decrease, 1960 to 1998. From Roth *et al.* (2002).

IT equipment affects national energy consumption in at least three general ways: (1) it causes structural changes to the economy, (2) it changes the mechanics of commerce and work (e-commerce), and (3) it consumes energy during manufacture.

2.2 Structural Changes in the Economy

Rapid growth in the low-energy intensity IT sector may decrease the energy intensity of the entire economy by shifting a larger portion of the nation's economic output to lower energy intensity activity. Overall, the IT-related sectors of the economy directly consume less energy per dollar of GDP produced than does the economy as a whole. Consequently, if IT sectors grow faster than the economy as a whole (i.e., if they become a larger portion of the economy), the energy intensity of the entire economy will be decreased. From 1990 to 1997, IT-related sectors grew about five times faster (in dollar terms) than the U.S. economy as a whole, contributing more than one-third of real GDP growth over that period. This may have accounted for a portion of the decrease in U.S. energy intensity during the late 1990s.

2.3 E-Commerce

E-commerce, short for electronic commerce, denotes the myriad of ways in which electronic transmission of content can substitute for the exchange of physical goods (dematerialization). Online shopping and banking, electronic music files and books, automated business-to-business exchanges, auctions, and electronic purchasing systems all are manifestations of e-commerce. E-commerce has long existed in certain forms (banking and stock trading), but its use increased when the public began embracing the Internet on a larger scale during the mid-1990s. Telecommunications and computer networks, along with servers and data storage devices, form the backbone of e-commerce. Other devices, including computers, monitors, and printers, enable people to directly interact and exchange information with other people on these networks.

Although e-commerce may result in net national energy savings, many of its applications have an

ambiguous net energy impact after considering all of the ramifications. For example, at first glance, it appears that a person ordering three books online reduces the energy consumed to acquire the books because the shipment to the front door via ubiquitous U.S. Post Office shipping (or by a private delivery company) replaces an automobile trip to a bookstore. However, many goods are shipped via air freight, which consumes nearly as much energy as does an auto trip. Furthermore, even if the books were bought online, the purchaser might have made the same automobile trip to buy something else at the same location as the bookstore or may have substituted a different trip for the trip to the bookstore. In either of these scenarios, the online order would show a net negative impact on energy consumption in the transportation sector. Online purchasing also requires significantly more packaging material, which consumes energy to manufacture and ship. Indeed, the net energy outcome is very sensitive to the specifics of the purchase: the transportation mode used and distance traveled for the store trip, the shipment mode(s) and distances traveled for e-commerce, the number of items purchased on the trip, the efficiency of the conventional book distribution network, the degree to which e-commerce reduces overproduction of books, and even the shipment packaging. "Rebound" effects (discussed in a subsequent subsection), such as any additional energy-consuming activities undertaken during the time saved (if any) by online ordering, also warrant consideration. All of these factors complicate this (or any) assessment of the net energy implication of e-commerce.

Many types of e-commerce require significant changes in personal and organizational behavior to achieve significant market presence, and these changes are typically difficult to realize and occur over long time periods. For example, shopping is not simply about the purchase; for many, it is a pleasurable activity that cannot be easily replaced by an online alternative. Similarly, the degree to which employees will adopt T&T is unclear. In some cases, T&T may be a favorable alternative for workers who spend a lot of time out of the office (e.g., sales personnel). On the other hand, a workplace fills important social needs and provides crucial opportunities for informal communication among workers. Security concerns about sharing personal information may also impede the development of e-commerce.

Ultimately, cost savings—not energy savings—drive companies to adopt e-commerce strategies, with the greatest savings expected to come from reduced procurement costs. In spite of optimistic projections for the growth of e-commerce during the current decade, e-commerce accounted for only a small fraction of U.S. and international commerce as recently as 2002. The net and potential impacts of e-commerce on national energy consumption will not become clear until businesses adopt e-commerce on a large scale.

The various manifestations of e-commerce have the potential to reduce energy consumption in the industrial (production), transportation, and building sectors.

2.3.1 Industrial Sector (Production)

Fundamentally, IT facilitates the acquisition and processing of data, increasing the flow of information about industrial processes and production. The information, intelligence, and control provided by IT can improve manufacturing and office productivity, reducing error rates and the time and energy expended to deal with errors. For example, in the back-office realm, a major manufacturer switching from a conventional ordering system to an online system could dramatically reduce its order rework rate. In one instance, such an improvement reduced order errors from 25 to 2%, saving the company $500 million while decreasing energy intensity.

2.3.2 Transportation Sector

Although not an aspect of e-commerce per se, IT equipment can also enhance the efficiency of industrial processes and product designs. IT equipment integrated with sensors could realize significant national savings through device improvement (e.g., an automobile) or enhanced system/process efficiency. For instance, a brewery installed a computer-based factory monitoring system integrated with a plant-wide intranet. The system improved process monitoring and production scheduling, resulting in a decrease in the level of waste bottles and cans from 5 to 0.1% of production.

IT can also enhance the effectiveness of product design by shortening the design cycle and improving the energy efficiency of design. To cite one example, aircraft companies routinely perform computational fluid dynamics (CFD) analyses on computers to model the flow around airplanes and calculate the drag. Relative to conventional wind-tunnel testing, CFD can be used to cost-effectively explore a wider range of design possibilities in a shorter period of time. This leads to improved airplane fuel efficiency, a key consideration in the purchase of commercial airliners.

E-commerce has the potential to reduce transportation energy consumption by reducing the number of trips and/or improving their efficiency. For example, roughly one-half of all trucks on the road carry no goods and are simply traveling to their next destination. An IT-enabled transportation exchange enables shippers to bid for unfilled space in truck fleets, increasing profits while also saving energy. Electronic distribution of music files and books precludes the need for transport of physical objects to the customer as well as the need for a trip to a store. In the case of T&T, electronic transfer of information (data and voice) makes trips to the office unnecessary. Similarly, video teleconferencing can replace long-distance travel.

2.3.3 Building Sector

E-commerce could reduce building energy consumption by reducing warehouse and retail building floor space. Improved supply chain management facilitated by IT should reduce inventories stored in warehouses, decreasing the square footage of floor space required and the energy consumed to condition and light that floor space. "Virtual storefronts" could replace a portion of the brick-and-mortar establishments. This could, in turn, greatly decrease (by more than 10-fold) the amount of energy required to operate businesses by reducing retail floor space (substituting low-energy warehouse-based operations for higher energy retail space) and increasing storage densities in warehouses relative to retail space. In both cases, reductions in warehouse and retail floor space would also displace a portion of projected new building construction, avoiding the energy consumed in the actual building construction and embodied in the building materials.

T&T can directly save energy by removing the need for office space. For example, assume that a major telecommunications company hopes to reduce office floor space per employee by approximately one-third by 2005 (relative to 2000) through vigorous promotion of telecommuting, which would lead to similar savings in energy consumption. In turn, reduced floor space demand would decrease the energy consumed in the construction sector, both directly (energy used to erect a building) and indirectly (energy consumed throughout the construction materials supply chains).

2.3.4 E-Commerce and the Rebound Effect

All purchasing decisions have ramifications well beyond the direct impact of a single purchase, making consideration of the energy impacts integrated over the entire economy very important. In energy terms, some rebound (the amount of additional energy consumed by other activities that occur due to the sale) is inevitable. Predicting the precise form of rebound experienced, as well as its magnitude, is often difficult. The rebound effect can operate on several levels relative to IT and may be intersector, meaning that an approach that reduces energy consumption in one sector tends to increase energy consumption in another.

Online bookstores clearly illustrate the importance of considering rebound effects. Lower costs for online book sales could force down the price of books, increasing the economy-wide demand for them as well as the national energy consumption to transport them. Net book production energy consumption would depend on the extent to which increased sales balance any decrease in book remainder rates. Moreover, if online sales do not actually displace physical bookstores, the energy used to support the infrastructure of online booksellers would not replace but rather supplement the energy already consumed in the retail sector. Rebound effects could also bring the savings of T&T into question. People who elect T&T may choose to live farther from the office, increasing automobile mileage for many trips and specifically during days when the employees do go to the office. If people move to less densely populated areas, the length of their other trips could increase. People working at home may also take advantage of their freedom to run more errands by car. On a somewhat larger scale, if T&T results in a net shift of the population to more remote locations, construction of the infrastructure (e.g., roads, electricity, sewage) to support the population shift will consume additional energy.

2.4 Manufacture of IT Equipment

2.4.1 IT Equipment

The manufacture of IT equipment consumes an appreciable amount of energy. A commodity I/O model of the U.S. economy based on 1997 U.S. Department of Commerce models and data enables life cycle analysis of the impact of the production of various equipment and materials, including IT equipment. It is important to note that the model includes not only the energy and resources directly consumed in device manufacture but also the resources consumed throughout the entire supply chain. The model indicates that the energy consumed

throughout the supply chain to produce IT equipment in the United States in 1997 is on the same order as the direct energy consumption of all IT equipment in a single year. Presumably, manufacturing efficiencies have improved since 1997, while device shipment levels have grown in both unit and dollar terms over the same period. Consequently, the conclusion that the energy consumed throughout the supply chain to produce IT equipment is on the same order as direct IT equipment energy consumption still holds. When considering the energy impact of IT equipment manufacturing in a given country, however, it is important to consider that a portion of the equipment is manufactured in a different than where it directly consumes energy.

2.4.2 Paper

Copy machines, printers, and fax machines all consume paper. The direct energy consumption estimates for these devices take into account the energy consumed to print an image on the paper but do not reflect the energy used to manufacture the paper. A study during the late 1990s estimated that manufacturing a piece of paper from wood requires about 17 Wh and that manufacturing it from recycled paper requires 12 Wh, whereas electrographically producing an image consumes an average of roughly 1 Wh. As a result, the roughly 900 billion sheets of paper consumed in the United States in 2000 by inkjet and non-roll-fed laser printers, copiers, and fax machines required approximately 15 TWh to manufacture. In addition, very high-volume roll-fed laser printers consumed another 600 billion sheets. Assuming similar energy consumption values per sheet of paper, these 600 billion sheets would have consumed approximately 10 TWh to manufacture. In sum, the energy consumed to manufacture the paper consumed by copiers and printers in 1 year likely exceeds the total electrical energy consumed to operate all of the commercial copiers and printers for 1 year ($\sim$16 TWh).

2.5 Building HVAC Energy Consumption

IT equipment located in conditioned areas of commercial buildings acts as an internal heat source, directly influencing heating, ventilation, and air-conditioning (HVAC) energy consumption. During cooling season, the electricity consumed by the equipment is equivalent to electric resistance heating, increasing the air-conditioning loads. The increase in cooling load also augments fan ventilation energy consumption by requiring the HVAC system to introduce additional cooled air into the conditioned space and remove it, and this in turn necessitates more cooling energy to mitigate the additional heat dissipated by the fans in the conditioned space. Conversely, during heating season, IT equipment displaces a portion of the heat that the building's heating system(s) would typically provide, generally 1 Btu of heating load per Btu dissipated by equipment. That is, the heat dissipated by the equipment reduces heating loads, supplanting the existing heating with inefficient resistant heating. Overall, IT equipment effectively lowers the outdoor balance temperature, above which the building requires air conditioning and below which the heating system operates. The precise impacts of the equipment on building loads vary greatly with the local climate as well as with the type of building in question, thereby complicating quantification of its impact. For example, the extremely high density of IT equipment in data centers requires year-round cooling in most cases, whereas IT equipment in office buildings displaces a portion of the heating load. The efficiency of the HVAC system also has a major impact on the energy expended to meet the modified loads. Taking all factors into account, the net national (U.S.) effect is likely to be a moderate increase in commercial building HVAC energy consumption.

HVAC system designers typically take IT equipment into account, modeling it as a heat source of from 0.4 to 1.1 W/ft^2, with an average value of approximately 0.8 W/ft^2. Naturally, some spaces, such as computer rooms, can generate much higher local loads, as do the portions of data centers that are densely packed with servers and computer and telecommunications network equipment (typically 20–60 W/ft^2, with levels projected to rise rapidly in the future).

2.6 Peak Electricity Demand Impact

The power draw of IT equipment during peak periods (typically mid- to late afternoon on hot summer days) makes three distinct contributions to demand. The first two, the direct power draw of the equipment and the power draw of the additional air conditioning required to cool the heat dissipated by the equipment, were described previously. The third, additional power demand resulting from equipment power factors, increases power demand at the power plant and building levels. Nonresidential IT equipment dominates the contributions to peak electricity

demand due to operating patterns and the relative magnitude of its energy consumption.

Most nonresidential IT equipment will be in use or ready for use (in active or stand-by mode) during this period. This results in a ratio of peak to (annual) average power draw of approximately 1.2 to 1.5, which is less "peaky" than office buildings as a whole (ratio of $\sim$1.8) due to dramatic increases in HVAC power draw during peak demand periods.

Power factors do not directly increase energy consumption per se but do increase power plant output. A study during the late 1990s found that most office equipment has power factors of approximately 0.6, primarily due to low-quality power supplies. In practical terms, this means that the power source must generate 1.67 times more power than is demanded by the equipment, and this increases the effective electricity demand as well as electricity transmission and distribution losses. On the other hand, server computers and network equipment tend to have higher power factors ($>$0.9). Taking into account all of these factors, on average, IT equipment appears to increase the total peak power demand in any given part of the United States by approximately 3%. The actual peak load impact for a given building depends on the actual density of IT equipment. For example, office buildings will face much larger increases in peak electricity demand from IT equipment than will retail establishments.

3. CONCLUSIONS

IT equipment directly consumed approximately 100 TWh of electricity in the United States in 2000, equal to just under 3% of all electricity produced, with monitors and desktop personal computers ranking as the two largest end uses ($\sim$40 TWh or 40% of the total). The infrastructure that powers the Internet—server computers, computer and telecommunications network equipment, and UPSs—consumed a total of approximately 30 TWh ($\sim$30%).

Indirectly, IT equipment influences national energy consumption in several ways that appear to be on the same order of magnitude as the direct impact, including the manufacture of IT equipment and several aspects of e-commerce: dematerialization, improved production efficiency, reduced inventories, and T&T.

National energy intensity (i.e., energy consumption per dollar of GDP) data reveal that the decrease in energy intensity sharpened during the late 1990s despite the low energy prices of that period. Because this coincided with the rise of the Internet in the United States, it suggests that IT may indeed reduce energy consumption relative to a business-as-usual forecast, slowing the growth of national energy consumption if not the absolute quantity of energy consumed. Closer consideration of e-commerce shows that careful consideration of changes in business paradigms and their interactions with multiple sectors of the economy, including rebound effects, is essential to developing a more accurate estimate of the magnitude and sign of the net energy impact of IT.

3.1 Future IT Energy Consumption

IT direct energy consumption will likely increase in the future due to growing equipment stocks to support the Internet infrastructure (server computers, telecommunications, and computer network equipment) and continued growth in the installed base and use of residential IT equipment. Both trends are driven by greater integration of society with IT. The future technology pathways of IT equipment will also play a major role, including the degree of substitution of portable computing for desktop computing, replacements for CRT monitors (e.g., with LCDs), and changes in PM-enabled rates. The deployment of ubiquitous sensors and controls in conjunction with communications could also increase direct energy consumption such as in home networks.

On a national scale, IT equipment has the potential to realize substantial reductions in future energy consumption relative to business as usual as e-commerce and other IT-related practices that dematerialize the economy and improve the efficiency of energy-consuming systems and equipment become more prevalent. The degree to which rebound effects will erode IT-enabled energy savings remains a critical issue warranting continued study.

SEE ALSO THE FOLLOWING ARTICLES

Electric Power Measurements and Variables • *Industrial Energy Use, Status and Trends* • *Microtechnology, Energy Applications of* • *Rebound Effect of Energy Conservation* • *Technology Innovation and Energy*

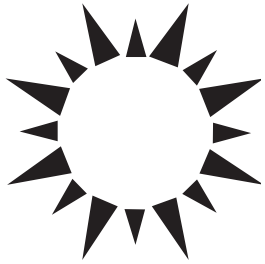
Further Reading

Matthews, H. S. (2001). "The Environmental Implications of the Growth of the Information and Communications Technology Sector." Organization for Economic Cooperation and Development, Paris.

Plepys, A. (2002). The grey side of ict. *Environ. Impact Assess. Rev.* **22**, 509–523.

Roth, K. W., Goldstein, F., and Kleinman, J. (2002). "Energy Consumption by Office and Telecommunications Equipment in Commercial Buildings—Vol. 1: Energy Consumption Baseline," final report by Arthur D. Little to the U.S. Department of Energy, Office of Building Technology, State and Community Programs, NTIS number PB2002-107657. www.tiax.biz/aboutus/pdfs/officeequipvol1.pdf.

Information Theory and Energy



SVEN ERIK JØRGENSEN
Danish University of Pharmaceutical Sciences
Copenhagen, Denmark

1. **Introduction**
2. **Information and Entropy**
3. **Shannon's Information Theory and Kullbach's Measure of Information**
4. **Exergy and Information**
5. **Order, Information, Genes, and Proteins**
6. **Closing Remarks**

Glossary

Boltzmann's constant A value (k) relating the average energy of a molecule to its absolute temperature; $k = 1.3803 \times 10^{-23}$ joules/molecules degree Kelvin.

detritus Dead organic matter; has an average free energy content of 18.7 kJ/g.

exergy The amount of work a system can perform when it is brought into thermodynamic equilibrium with its environment; entropy-free energy.

negentropy The difference between the entropy at thermodynamic equilibrium and the entropy of the present state.

The science of information theory has its root in the work of the theoretical physicists Ludwig Boltzmann and J. Willard Gibbs. At the beginning of the 20th century, Boltzmann and Gibbs related probability to entropy and free energy and thereby established the basis for the later development of information theory. The more modern formulation of information theory is attributed to Claude Shannon and Warren Weaver. Their approach was rooted in the problems of coding and decoding information during World War II. Shannon formulated an index of information known as Shannon's Information Index. This is sometimes referred to as "Shannon entropy," which has contributed to much confusion because of the unfortunate mixing of the two concepts, i.e., entropy and information. In the 1980s, biologists Daniel Brooks and E. O. Wiley, in their book *Evolution as Entropy*, added to this confusion. In addition, entropy has been applied by Ilya Prigogine and G. Nicolis to express the amount of information embodied in biological systems; the concept of negative entropy (negentropy) was defined by Erwin Schrödinger in his famous book *What is Life?*, and later also by Karl Popper. Jeffery Wicken should also be mentioned in this context, because he criticized the Brooks–Wiley attempt and introduced a more stringent interpretation of the relationship between entropy and information. Many researchers, including Myron Tribus, Werner Ebeling, Mikhail Volkenstein, and Rainer Feistel, have contributed to the clarification of the relationships between entropy and information, and although classical information theory is primarily discussed herein using Shannon's and Weaver's work and the later refinement by Leon Brillouin, the interpretations of the relationship between information and entropy of the many other contributors have been included.

1. INTRODUCTION

Thermodynamic information has found its widest application in biology due to the extremely high complexity of biological systems. This application of information theory will be presented here through an introduction to the concept of exergy. This thermodynamic concept was introduced in the 1950s, but was not applied in the ecological/biological contexts until the 1970s. These very complex concepts, involving the relationship between order and information as applied to biological systems, are presented here in the context of the genome and the proteom.

2. INFORMATION AND ENTROPY

In statistical mechanics, entropy is related to probability. A system can be characterized by averaging

 © 2004 Elsevier Inc. All rights reserved.

ensembles of microscopic states to yield the macrostate. If W is the number of microstates that will yield one particular macrostate, the probability P that this particular macrostate will occur as opposed to all other possible macrostates is proportional to W. It can further be shown that

$$S = k \ln W, \tag{1}$$

where k is Boltzmann's constant (1.3803×10^{-23} J/molecules degree Kelvin). Entropy is a logarithmic function of W and thus measures the total number of ways that a particular macrostate can be constituted microscopically. The entity S may be called the uncertainty or the required thermodynamic information, meaning the amount of information needed to describe the system, which must not be interpreted as the information actually available. The more microstates there are and the more disordered they are, the more information is required and the more difficult it will be to describe the system.

If an ecosystem is in thermodynamic equilibrium, the entropy, S_{eq}., is higher than it is in an ecosystem in nonequilibrium. The excess entropy may be denoted as the thermodynamic information and is also defined as the negentropy, NE:

$$I = S_{eq} - S = \text{NE}. \tag{2}$$

In other words, a decrease in entropy, S, will imply an increase in information, I, and erosion or loss of information implies an increase in entropy. When $S_{eq} = S$ or $I = 0$, the system cannot be distinguished from its environment and the information about the system is thereby lost. Further, the principle of the second law of thermodynamics corresponds to a progressive decrease of the information content. An isolated system can evolve only by degrading its information.

Schrödinger formulated Boltzmann's Eq. (1) as follows:

$$S = k \ln D, \tag{3}$$

where S is the entropy, k is Boltzmann's constant, and D is the quantitative measure of the atomic disorder. The value D partly covers the heat motion and partly the random mixing of atoms and molecules. Furthermore, $1/D$ may be defined as order, Or. The equation may therefore be reformulated:

$$-S = k \ln(1/D) = k \ln(\text{Or}). \tag{4}$$

In other words, negative entropy is a measure of order.

That entropy is increased by dispersion, which creates more disorder, more possible microstates, and more randomness, can be shown by a simple model consisting of two chambers of equal volume, connected with a valve. One chamber contains 1 mol of a pure ideal gas (i.e., pressure and volume $pv = RT$, the ideal gas constants) and the second chamber is empty. If the valve between the two chambers is opened, and it is assumed that the change in energy $\Delta U = 0$ and time T is constant, an increase in entropy will be observed:

$$\Delta S = \int \partial Q/T = Q/T = W = R \ln V_2/V_1 = R \ln 2, \tag{5}$$

where ΔS is the increase in entropy, V_2 is the volume of gas in the chamber after the valve is opened, and V_1 is the volume of gas before the valve is opened. Thus, paradoxically, the greater the attempt to maintain order, the greater the requirement for energy and the greater the environmental stress (entropy), because all energy transformations from one form to another imply production of waste heat and entropy according to the second law of thermodynamics.

3. SHANNON'S INFORMATION THEORY AND KULLBACH'S MEASURE OF INFORMATION

Entropy of information is frequently applied to social and biological problems. Although entropy of information is analogous to thermodynamic entropy, it is not the same thing, because thermodynamic entropy is derived from physical laws only. Shannon and Weaver introduced a measure of information that is widely used as a diversity index by ecologists. Shannon's index states that

$$H = -K' \sum_{i=1}^{n} p_i \log_2(p_i), \tag{6}$$

where p_i is the probability distribution of a species and K' is a constant ($K' = 1$ when the equation is used to express biodiversity, for instance). Unfortunately, the term "Shannon's entropy" was introduced as an alternative name for Shannon's index. Several authors have attributed this to be due to John von Neumann, who is said to have used the argument that because nobody knows what entropy is, Shannon would have an advantage and would probably win any discussion. The symbol H is thus used here to avoid confusion and to stress that H and S to a certain extent are two different concepts. Both

S and H increase with an increasing number of possible (micro)states, but it is common to equate information with knowledge and entropy with lack of knowledge. The higher the Shannon's index is, the more knowledge is needed to cover, for instance, the entire representation of species in an ecosystem. By increasing diversity, H increases. If the information about this diversity is obtainable, which is assumed (for instance, the number of organisms for each of the species in an ecosystem must be counted to be able to discuss what the biodiversity is), the knowledge of the system has increased correspondingly. In contrast, S considers the uncertainty or the lack of knowledge that increases by increasing diversity (complexity). Entropy implicitly assumes that this knowledge is not available but is necessary to be able to describe the system fully. If the system is at thermodynamic equilibrium, whereby all of the components (molecules) are randomly distributed, the number of microstates is equal to the number of molecules. If, on the other hand, the system has a certain order, it will be possible to group some of the components (molecules) and the knowledge necessary to describe the system is correspondingly less. Notice also that Shannon's index uses $\log_2$ whereas entropy uses the natural logarithmic (ln; the number of microstates) measures, according to Eq. (1), for the complexity of the system.

In 1962, Brillouin formulated a rigid science of information theory, based on probability theory. His definition of information, I, is based on a comparison of the information content in two situations: (1) the *a posteriori* situation in which p_0 is the probability of an event or the presence of a component and the information is $I_0 = 0'$, and (2) the *a priori* situation in which the probability of the same event is p_1 and the information $I_1 > 0$. This can be found from Eq. (7), which is known as Kullbach's measure:

$$I_1 = kp_0 \ln p_0/p_1, \tag{7}$$

where k is Boltzmann's constant. Equation (7) means that I expresses the amount of information that is gained as a result of the observations. In observing the system used for the computation of the entropy production by opening a valve between two connected chambers, it is expected that the molecules will be equally distributed in the two chambers, i.e., $p_0 = p_1$ is equal to 1/2. If, on the other hand, it is observed that all the molecules are in one chamber, $p_1 = 1$ and $p_0 = 0$. As can be seen, the entropy obtained by application of Kullbach's measure is the same as that obtained by the calculation of entropy,

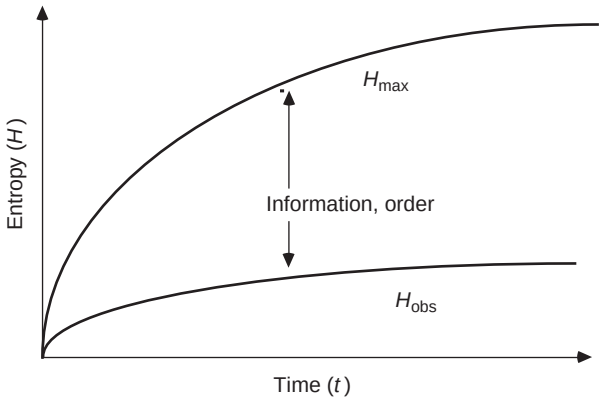


FIGURE 1 H_{max} and H_{obs} are plotted versus time. The difference between H_{max} and H_{obs} represents I, which increases with time, t. Data from Brooks and Wiley (1986).

because $R = kA$, where A is Avogadro's number, and there is proportionality to the number of molecules.

Figure 1 shows the relationship, as presented by Brooks and Wiley, in which H_{max} corresponds to the entropy of the ecosystem if it were at thermodynamic equilibrium, whereas H_{obs} is the actual entropy level of the system. The difference between H_{max} and H_{obs} covers the information or order. It means that

$$H_{max} = \log W, \tag{8}$$

where W is the total number of microstates available to the system. The value H_{obs} is defined according to the Eq. (9):

$$H_{obs} = -\sum_{i=1}^{n} p_i \ln(p_i). \tag{9}$$

Brooks and Wiley have interpreted this development of entropy in a variety of ways:

1. H_{obs} is interpreted as complexity: the higher the complexity, the more energy is needed for maintenance and therefore more is wasted as heat. The information in this case becomes the macroscopic information.
2. H_{obs} is translated to realization, whereas H_{max} becomes the total information capacity. Information may in this case be called constraints. Notice, however, that the strict thermodynamic interpretation of H_{max} is H at thermodynamic equilibrium, which does not develop (change) for an ecosystem on Earth.
3. H_{obs} represents the observed distribution of genotypes and H_{max} is any genotype equally likely to be found. The information becomes the organization of organisms over genotypes.

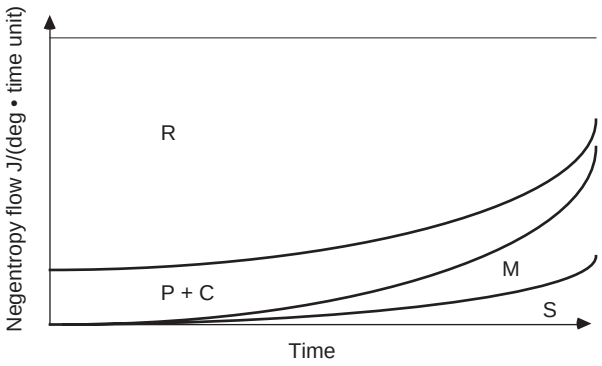


FIGURE 2 Due to biological evolution, there is tentative development in the application of the negentropy flow (solar radiation) on Earth versus time. Four cases are shown: the negentropy is not used, but is reflected (R); or is used by physical (P) and chemical (C) processes on Earth, P + C; or is used for maintenance (M) of the biological structure; or is used to construct biological structure (S).

Brooks and Wiley's theory seems inconsistent with the general perception of ecological development: order increases and entropy therefore decreases at the cost of entropy production of the environment. The misinterpretation probably lies in the translation of order to information. By increasing order, the amount of information needed decreases. Note that entropy in this context covers the amount of information needed.

Figure 2 attempts to give a different and tentative picture of the development as applied to the negentropy flow (i.e., the solar radiation) on Earth due to evolution. The negentropy flow is considered approximately constant, although the solar radiation has shown some (relatively minor) changes. Four different applications of negentropy are considered: unused negentropy (reflection of the radiation), negentropy used for physical and chemical processes, negentropy used for maintenance of life (respiration), and negentropy used for construction of biological structures. The latter two increase at the expense of reflection. It has been widely discussed whether ecosystems attempt to maximize H_{obs} (i.e., negentropy = information = $H_{max} - H_{obs}$). The difference between biological and thermodynamic entropy is also still the subject of discussion.

4. EXERGY AND INFORMATION

What is exergy? Definitions from a number of sources can be applied. Exergy is defined as the amount of work (entropy-free energy) a system can perform when it is brought into thermodynamic equilibrium with its environment. Figure 3 illustrates the definition. The system considered is characterized by the extensive state variables, S, U, V, N_1, N_2, N_3, ..., N_x, where S is the entropy, U is the energy, V is the volume, and N_1, N_2, N_3, ..., N_x are moles of various chemical compounds, and by the intensive state variables, T, p, μ_{c1}, μ_{c2}, μ_{c3}, ..., μ_{cx}. The system is coupled to a reservoir, a reference state, by a shaft. The system and the reservoir form a closed system. The reservoir (the environment) is characterized by the intensive state variables, T_0, p_0, μ_{0c1}, μ_{0c2}, μ_{0c3}, ..., μ_{0cx}, and because the system is small compared with the reservoir, the intensive state variables of the reservoir will not be changed by interactions between the system and the reservoir. The system develops toward thermodynamic equilibrium with the reservoir and is simultaneously able to release entropy-free energy to the reservoir. During this process, the volume of the system is constant because the entropy-free energy must be transferred through the shaft only. The entropy is also constant because the process is an entropy-free energy transfer from the system to the reservoir, but the intensive state variables of the system become equal to the values for the reservoir. The total transfer of entropy-free energy in this case is the exergy of the system. It is seen from this definition that exergy is dependent on the state of the total system (= system + reservoir) and is not entirely dependent on the state of the system. Exergy is therefore not a state variable. In accordance with the first law of thermodynamics, the increase of energy in the reservoir, ΔU, is

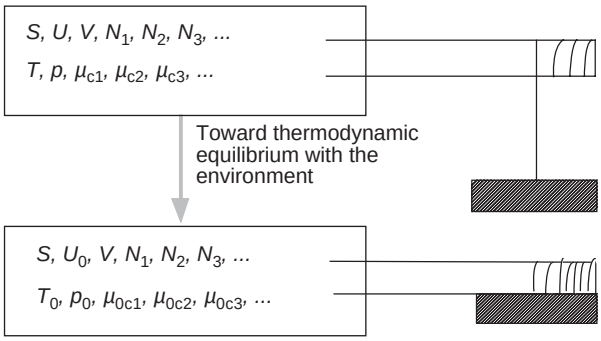


FIGURE 3 The definition of exergy.

$$\Delta U = U - U_0, \tag{10}$$

where U_0 is the energy content of the system after the transfer of work to the reservoir has taken place.

According to the definition of exergy (Ex),

$$\mathrm{Ex} = \Delta U = U - U_0. \tag{11}$$

Because

$$U = TS - pV + \sum_{\mathrm{c}} \mu_{\mathrm{c}} N_i \tag{12}$$

(see any textbook on thermodynamics), and

$$U_0 = T_0 S - p_0 V + \sum_{\mathrm{c}} \mu_{0\mathrm{c}} N_i, \tag{13}$$

we get the following expression for exergy:

$$\mathrm{Ex} = S(T - T_0) - V(p - p_0) + \sum_{\mathrm{c}} (\mu_{\mathrm{c}} - \mu_{0\mathrm{c}}) N_i. \tag{14}$$

Because the reservoir equals the reference state, it is possible, for instance, to select the same system but at thermodynamic equilibrium, i.e., all components are inorganic and at the highest oxidation state, if sufficient oxygen is present (nitrogen as nitrate, sulfur as sulfate, and so on). The reference state will in this case correspond to the ecosystem without life forms and with all chemical energy utilized or as an "inorganic soup." Usually, this implies considering that $T = T_0$ and $p = p_0$, which means that the exergy becomes equal to the difference between the Gibbs free energy of the system and the same system at thermodynamic equilibrium, or the chemical energy content includes the thermodynamic information of the system. Notice that Eq. (14) also emphasizes that exergy is dependent on the state of the environment (the reservoir = the reference state), because the exergy of the system is dependent on the intensive state variables of the reservoir.

Note that exergy is not conserved only if entropy-free energy is transferred, which implies that the transfer is reversible. All processes in reality are, however, irreversible, which means that exergy is lost (and entropy is produced). Loss of exergy and production of entropy are two different descriptions of the same reality, namely, that all processes are irreversible, and unfortunately there is always some loss of energy forms that can do work to energy forms that cannot do work (heat at the temperature of the environment). So, the formulation of the second law of thermodynamic by use of exergy is as follows: All real processes are irreversible, which implies that exergy inevitably is lost. Exergy is not conserved, whereas energy, of course, is conserved by all processes, according to the first law of thermodynamics. It is therefore wrong to discuss, as already mentioned briefly, the energy efficiency of an energy transfer, because it will always be 100%, whereas the exergy efficiency is of interest, because it will express the ratio of useful energy to total energy, which always is less than 100% for real processes, according to the second law of thermodynamics.

It is therefore of interest to set up for all environmental systems, in addition to an energy balance, also an exergy balance. The concern is loss of exergy, because it means that "first-class energy," which can do work, is lost as "second-class energy" (heat at the temperature of the environment), which cannot do work. So, the particular properties of heat and the fact that temperature is a measure of the movement of molecules create limitations in the possibilities to utilize all forms of energy to do work. Due to these limitations, a distinction must be made between exergy, which can do work, and anergy, which cannot do work, and thus all real processes imply inevitably a loss of exergy as anergy.

Exergy seems more useful than entropy to describe the irreversibility of real processes, because it has the same units as energy and is a form of energy, whereas entropy by definition is more difficult to relate to concepts associated with typical conceptions of reality. In addition, entropy is not clearly defined for "systems far from thermodynamic equilibrium," particularly for living systems. Moreover, it should be mentioned that the self-organizing abilities of systems are strongly dependent on temperature. Exergy takes temperature into consideration, but entropy does not. By definition, exergy at 0 K is 0 and is at a minimum. Negative entropy, as previously discussed, does not express the ability of a system to do work. Exergy therefore becomes, in contrast to entropy, a good measure of "creativity," which increases proportional to temperature. Furthermore, exergy facilitates the differentiation between low-entropy energy and high-entropy energy, because exergy is entropy-free energy. Finally, note that information contains exergy. Boltzmann showed that the free energy of the information at hand (in contrast to the information needed to describe the system) is $kT \ln I$, where I is the information known about the state of the system [for instance, that the configuration is 1 out of W possible (i.e., that $W = I$) and k is Boltzmann's constant ($= 1.3803 \times 10^{-23}$ J/molecules K)]. This implies that one bit of information has the exergy equal to $kT \ln 2$. Transference of information from one system to another is often almost an entropy-free energy transfer. If the two systems have different temperatures, the entropy lost by one system is not equal to the entropy gained by the other system, whereas the exergy lost by the first system is equal to the exergy transferred and is equal

to the exergy gained by the other system, provided that the transference is not accompanied by any loss of exergy. In this case, it is obviously more convenient to apply exergy than entropy.

The exergy of a system measures contrast—it is the difference in free energy if there is no difference in pressure and temperature, as may be assumed for an ecosystem or an environmental system and its surroundings against the surrounding environment. If a system is in equilibrium with the surrounding environment, the exergy is of course zero. Because the only way to move systems away from equilibrium is to perform work on them, and because the available work in a system is a measure of ability, a distinction must be made between the system and its environment, or thermodynamic equilibrium alias (for instance, an inorganic soup). Therefore, it is reasonable to use the available work, i.e., the exergy, as a measure of the distance from thermodynamic equilibrium.

Because it is known that ecosystems, due to the through-flow of energy, have the tendency to move away from thermodynamic equilibrium, losing entropy or gaining exergy and information, the following proposal for the relevance for systems can be put forward: systems, receiving a through-flow of energy, attempt to move toward a higher level of exergy. This proposal for exergy development in ecosystems makes it pertinent to assess the exergy of ecosystems. It is not possible to measure exergy directly, but it is possible to compute it using Eq. (14). Assuming a reference environment that represents the system (ecosystem) at thermodynamic equilibrium, which means that all the components are inorganic, are at the highest possible oxidation states (if sufficient oxygen is present and as much free energy as possible is utilized to do work), and are homogeneously distributed in the system (no gradients), the situation illustrated in Fig. 4 is valid. The chemical energy embodied in the organic components and in the biological structures contributes by far the most to the exergy content of the system, thus there seems to be no reason to assume a (minor) temperature and pressure difference between the system and the reference environment. Under these circumstances, it is possible to calculate the exergy content of the system as coming entirely from the chemical energy: [$T = T_0$ and $p = p_0$ in Eq. (14)]:

$$\mathrm{Ex} = \sum_{\mathrm{c}} (\mu_{\mathrm{c}} - \mu_{0\mathrm{c}}) N_i. \tag{15}$$

Equation (15) represents the nonflowing chemical exergy, which is determined by the difference in chemical potential ($\mu_{\mathrm{c}} - \mu_{0\mathrm{c}}$) between the ecosystem and the same system at thermodynamic equilibrium. This difference is determined by the concentrations of the components in the system and in the reference state (thermodynamic equilibrium), as is the case for all chemical processes. The concentrations in the ecosystem can be measured, but the concentrations in the reference state (thermodynamic equilibrium) can be based on the usual chemical equilibrium constants. Given the process

$$\text{Component A} \longleftrightarrow \text{inorganic decomposition products}, \tag{16}$$

there is a chemical equilibrium constant, K:

$$K = \frac{[\text{inorganic decomposition products}]}{[\text{component A}]}. \tag{17}$$

The concentration of component A at thermodynamic equilibrium is difficult to find, but can be based on the composition of A; the concentration of component A is found at thermodynamic equilibrium from the probability of forming A from the inorganic components. By these calculations, the exergy of the system, compared with the same system at the same temperature and pressure, but in the form of an inorganic soup without any life, biological structure, information, or organic molecules, is found. Because ($\mu_{\mathrm{c}} - \mu_{0\mathrm{c}}$) can be found from the definition of the chemical potential, replacing activities by concentrations, the following expression for the exergy

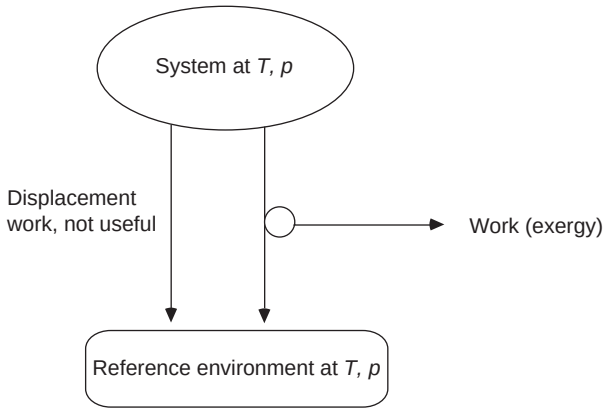


FIGURE 4 The exergy content of the system is calculated for the system relative to a reference environment of the same system at the same temperature (T) and pressure (p), but as an inorganic soup with no life, biological structure, information, or organic molecules.

is obtained:

$$\mathrm{Ex} = RT \sum_{i=0}^{i=n} C_i \ln C_i/C_{i,0}, \qquad (18)$$

where R is the gas constant, T is the temperature of the environment (and the system; see Fig. 4), and C_i is the concentration of the ith component expressed in a suitable unit (e.g., for phytoplankton in a lake, C_i could be expressed as milligrams/liter or as milligrams/liter of a local nutrient). Also, $C_{i,0}$ is the concentration of the ith component at thermodynamic equilibrium and n is the number of components. Of course, $C_{i,0}$ represents a very small concentration (except for $i = 0$, which is considered to cover the inorganic compounds), corresponding to a very low probability of forming complex organic compounds spontaneously in an inorganic soup at thermodynamic equilibrium. For the various organisms, $C_{i,0}$ is even lower, because the probability of forming the organisms is very low, due to their embodied information, which implies that the genetic code should be correct. By using this particular exergy, based on the same system at thermodynamic equilibrium as a reference, the exergy becomes dependent only on the chemical potential of the numerous biochemical components that are characteristic of life. This is consistent with Boltzmann's statement that life is struggle for free energy.

The total exergy of an ecosystem cannot be calculated exactly, because it is not possible to measure the concentrations of all components or to determine all possible contributions to exergy in an ecosystem. If the exergy of a fox is calculated, for instance, the calculations will reveal only the contributions coming from the biomass and from the information embodied in the genes of the fox, but not the contributions from blood pressure, sex hormones, and so on. These properties are at least partially covered by the genes, but is that the entire story? The contributions from the dominant components can be calculated (for instance, by the use of a model or measurements), which covers the most essential components for a focal problem. The difference in exergy by comparison of two different structures (species composition) is decisive here. Moreover, exergy computations always yield only relative values, because the exergy is calculated relative to the reference system.

It is possible to distinguish between the exergy of information and of biomass. The value p_i, defined as c_i/A, where

$$A = \sum_{i=1}^{n} c_i \qquad (19)$$

is the total amount of matter in the system, is introduced as new variable

$$\mathrm{Ex} = ART \sum_{i=1}^{n} p_i \ln p_i/p_{i0} + A \ln = A/A_0. \qquad (20)$$

As $A \approx A_0$, exergy becomes a product of the total biomass A (multiplied by RT) and the Kullback measure:

$$K = \sum_{i=1}^{n} p_i \ln(p_i/p_{i0}), \qquad (21)$$

where p_i and p_{i0} are probability distributions, *a posteriori* and *a priori* to an observation of the molecular detail of the system. Specific exergy is exergy relative to the biomass and for the ith component: Sp. $\mathrm{Ex}_i = \mathrm{Ex}_i/c_i$. This implies that the total specific exergy per unit of area or per unit of volume of the ecosystem is equal to RTK.

It is interesting in this context to draw a parallel with regard to development of entropy for the entire universe. Classical thermodynamic interpretations of the second law of thermodynamics predict that the universe will tend to "heat death," whereby the entire universe has the same temperature, no changes ever take place, and a final overall thermodynamic equilibrium results. This prediction is based on the steady increase of entropy according to the second law of thermodynamics: thermodynamic equilibrium is the attractor. It can, however, be shown that, due to expansion of the universe, the system is moving away from thermodynamic equilibrium at a high rate of speed. Due to incoming solar radiation energy, an ecosystem is able to move away from thermodynamic equilibrium (i.e., the system evolves, obtains more information, and undergoes organization). The ecosystem must produce entropy for maintenance, but the low-entropy energy flowing through the system may be able to more than compensate for this production of disorder, resulting in increased ecosystem order or information.

5. ORDER, INFORMATION, GENES, AND PROTEINS

Information theory has found widespread use in the analysis of biological macromolecules. Schrödinger, in his order principle, proposed that information

about the organization of biological systems was laid down in the system. This was confirmed when Watson and Crick elucidated the structure of DNA, and a platform was provided that made it possible to apply information theory more comprehensively to biological systems.

Again, the expression for exergy per unit of volume is

$$\mathrm{Ex} = RT \sum_{i=0}^{i=n} C_i \ln C_i / C_{i,0}, \tag{22}$$

where R is the gas constant, T is the temperature of the environment, and C_i is the concentration of the ith component expressed in a suitable unit (e.g., for phytoplankton in a lake, C_i could be expressed as milligrams/liter or as milligrams/liter of a local nutrient. The concentration of the ith component at thermodynamic equilibrium is $C_{i,0}$ and n is the number of components. This is, of course, a very small concentration, except for $i = 0$, which is considered to cover the inorganic compounds; however, $C_{i,0}$ is not zero, but corresponds to a very low probability of forming complex organic compounds spontaneously in an inorganic soup at thermodynamic equilibrium.

It has been suggested that the exergy of structurally complicated material can be estimated on the basis of the elementary composition of the material. This, however, has the disadvantage that higher organisms and microorganisms with the same elementary composition will have the same exergy, which is in complete disagreement with the lower probability of forming more complex organisms, i.e., the lower concentration of $C_{i,0}$ in Eq. (22). The problem related to the assessment of $C_{i,0}$ has been discussed, and a possible solution was proposed by Jørgensen. For dead organic matter (detritus), which is given the index 1, the solution can be found from classical thermodynamics. Detritus has a free energy (chemical energy) of about 18.7 kJ/g. For the biological components, 2, 3, 4, ..., N, the probability, p_{i0}, consists at least of the probability of producing the organic matter (detritus), i.e., $p_{1,0}$, and the probability, $p_{i,a}$, of finding the correct composition of the enzymes (number and sequence, a, of amino acids) that determine the biochemical processes in the organisms. In living organisms, 20 different amino acids comprise the enzymes, and each gene determines, on average, the sequence of about 700 amino acids. For *Homo sapiens*, it has been found that as many as 38,000 amino acids are determined by one gene. The value for $p_{i,a}$ can be found from the number of permutations that represent the evolutionary selection of a particular organism based on its characteristic amino acid sequence. This means that

$$p_{i,a} = a^{-Ng_i}, \tag{23}$$

where a is the number of possible amino acids ($= 20$), N is the number of amino acids determined by one gene ($= 700$), and g_i is the number of non-nonsense genes. The following equations are available to calculate p_i:

$$p_{i,0} = p_{1,0}\, p_{i,a} = p_{1,0}\, a^{-Ng} \approx p_{1,0} \cdot 20^{-700g} \tag{24}$$

and the exergy contribution of the ith component (per unit of volume) can be found by combining these equations:

$$\begin{aligned} \mathrm{Ex} &= RTc_i \ln c_i / (p_{1,0} a^{-Ng_{c_{0,0}}}) \\ &= (\mu_1 - \mu_{1,0}) c_i - c_i \ln p_{i,a} \\ &= (\mu_1 - \mu_{1,0}) c_i - c_i \ln(a^{-Ng_i}) \\ &= 18.7 c_i + 700(\ln 20) c_i g_i. \end{aligned} \tag{25}$$

The total exergy can be found by summing up the contributions originated from all components. The contribution by inorganic matter can be neglected because the contributions by detritus and (to a greater extent) by biological materials are much higher due to an extremely low concentration of these components in the reference system. The contribution by detritus (dead organic matter) is 18.7 kJ/g times the concentration (in grams/unit of volume), as previously indicated, and the exergy of living organisms consists of $\mathrm{Ex}_{1,\mathrm{chem}} = 18.7$ kJ/g times the concentration c_i (grams/unit of volume) and

$$\begin{aligned} \mathrm{Ex}_{i,\mathrm{bio}} &= RT(700 \ln 20) c_i g_i \\ &= RT(2100) g_i c_i, \end{aligned} \tag{26}$$

where $R = 8.34$ J/mol; if an average molecular weight of 10^5 is assumed for the enzymes (proteins), the following equation for $\mathrm{Ex}_{i,\mathrm{bio}}$ at 300 K is obtained: $\mathrm{Ex}_{i,\mathrm{bio}} = 0.0529 g_i c_i$, where the concentration now is expressed in grams/unit of volume and the exergy is in kilojoules/unit of volume. The exergy (per unit of volume) for the entire system, $\mathrm{Ex}_{\mathrm{total}}$, can be found:

$$\mathrm{Ex}_{\mathrm{total}} = 18.7 \sum_{i=1}^{N} c_i + 0.0529 \sum_{i=1}^{N} c_i g_i, \tag{27}$$

where g for detritus ($i = 1$) of course is 0. Table I illustrates how each contribution depends on selected biological systems as well as on the detritus. A

TABLE I

Energy of Living Organisms

Organism	g_i	$\beta = (Ex_{i,chem} + Ex_{i,bio})/Ex_{i,chem}$	Exergy (kJ/g)[a]
Detritus	0	1	18.7
Minimal cell	470	2.3	43.8
Bacteria	600	2.7	50.5
Algae	850	3.4	64.2
Yeast	2000	5.8	108.5
Fungi	3000	9.5	178
Sponges	9000	26.7	499
Mold	9500	28.0	524
Plants, trees	10,000–30,000	29.6–86.8	554–1623
Worms	10,500	30.0	561
Insects	10,000–15,000	29.6–43.9	554–821
Jellyfish	10,000	29.6	554
Zooplankton	10,000–15,000	29.6–43.9	554–821
Fish	100,000–120,000	287–344	5367–6433
Birds	120,000	344	6433
Amphibians	120,000	344	6433
Reptiles	130,000	370	6919
Mammals	140,000	402	7517
Humans	250,000[b]	716	13389

[a] Based on the energy contained in detritus. On average, the exergy of 1 g of detritus is 18.7 kJ.

[b] See text for discussion.

weighting factor ß is introduced to be able to cover the exergy for various organisms in the unit detritus equivalent or chemical exergy equivalent:

$$Ex_{total} = \sum_{i=1}^{N} \beta_i c_i \quad (\text{as detritus equivalent}). \qquad (28)$$

Through the use of Eq. (28), the calculation of exergy accounts for the chemical energy in organic matter as well as for the (minimum) information embodied in living organisms. The latter contribution is measured as the extremely small probability of forming living components (for instance, algae, zooplankton, fish, mammals, and so on) spontaneously from inorganic matter. Weighting factors, defined as the exergy content relative to detritus (see Table I), may be considered quality factors that reflect how developed the various groups are and to what extent they contribute to the exergy due to their content of information, which is reflected in the computation. This is, according to Boltzmann, who gave the following relationship for the work, W, embodied in the thermodynamic information:

$$W = RT \ln N, \qquad (29)$$

where N is the number of possible states, among which the information has been selected. For species, N is the inverse of the probability of obtaining the valid amino acid sequence spontaneously. The Kullback measure of information covers the gain in information, when the distribution is changed from p_{i0} to p_i. Note that K is a specific measure (per unit of matter). When multiplied by the total concentration, K yields the exergy [see Eqs (20) and (21)].

Exergy calculated by use of the equations given herein has some clear shortcomings:

1. Although minor, some approximations have been made.
2. The non-nonsense genes for all organisms are not known.
3. The exergy embodied in the proteins (enzymes) is calculated only in principle, and there are other components of importance for the life processes. These components contribute less to the exergy compared to the enzymes, and the information embodied in the enzymes controls the formation of these other components (for instance, hormones), but it cannot be excluded that these components contribute to the total exergy of the system.
4. The exergy of the ecological network is not included. If the exergy of models is calculated, the network will always be relatively simple and the contribution coming from the information content of the network is negligible.
5. A simplification (for instance, models or diagrams) of the ecosystem is always used. This implies that only the exergy contributions of the components included in the simplified image of the ecosystem are calculated. The real ecosystem will inevitably contain more components, which are not included in the calculations.

It is therefore proposed to consider the exergy found by these calculations as a relative minimum exergy index to indicate that there are other contributions to the total exergy of an ecosystem, although they may be of minor importance. In most cases, however, a relative index is sufficient to understand the reactions of ecosystems, because the absolute exergy content is irrelevant for the reactions. It is, in most cases, the change in exergy that is of importance to understand the ecological reactions.

The weighting factors presented in Table I have been applied successfully in several ecological models. The relatively good results obtained, despite the uncertainty of their assessment, seem to be explicable only by the robustness of the application of the factors in modeling and other quantifications. The

differences between the factors (for microorganism, vertebrates, and invertebrates) are so obvious that it does not matter if the uncertainty of the factors is very high—the results are not influenced. On the other hand, it would be an important progress to get better weighting factors from a theoretical point of view, because it would enable modeling the competition between species that are closely related.

There is no doubt that the correct estimation of ß values should be based on the number of proteins that control the processes in the cells of various organisms. Information about all human genomes is available, and it has been found that the number of non-nonsense genes is not 250,000, as indicated in Table I, but rather is 40,000. On the other hand, it is also clear that the number of amino acids controlled by one gene is more than 700 for *Homo sapiens*. for some genes, it may be as high as 38,000. The weighting factor in Table I may therefore be approximately correct. The importance of proteins in biology and physiology is reflected in the intensive and enormous amount of analytical work that has been invested for many decades to find the composition of the human proteins. Current terminology for genetically determined proteins, "proteoms," emphasizes that these biological components are of particular importance. The great interest in proteoms is due to their roles in control of life processes. Many of these proteins may be the medicines of the future. The use of enzymes in industrial production is in its infancy, and there is enormous potential for their application in many more industrial processes.

Although the key to finding better β values is the proteom, knowledge about the number of proteoms in various organisms is very limited; more is known about the number of non-nonsense genes. It may be possible, however, to compare knowledge about non-nonsense genes, the overall DNA content, and the limited knowledge about the number of proteoms and evolutionary data, and see some pattern emerge. This could be used to give better (although still very approximate, at this stage) β values. For *Homo sapiens*, it is presumed that 200,000 different proteins are produced by the cells and that they contain about 15,000 amino acids on average. This would yield a biological exergy of $RT(\ln 20) \times 3 \times 10^9\ c_i$; with a molecular weight of 100,000, the β value would be 12,275, or considerably higher than the value in Table I. The other values in the table may be similarly different. Current knowledge about the genome and the proteom is unfortunately very limited.

A live frog of 20 g will have an exergy content of $20 \times 6433\,\text{kJ} \approx 0.1296\,\text{GJ}$, whereas a dead frog will have an exergy content of only 374 kJ, although both frogs have the same chemical composition (at least within a few seconds after death). The difference is rooted in the information, or rather the difference in the useful information. The dead frog has the information a few seconds after its death (the amino acid composition has not yet been decomposed), but the difference between a live frog and a dead frog is the ability to utilize the enormous information stored in the genes and the proteom of the frog. The amount of information stored in a frog is surprisingly high. The number of amino acids required to exist in the right sequence is 84,000,000, which equates, for the correct selection of 1 of 20 possible amino acids, to $84{,}000{,}000^{20}$ possible microstates. This amount information is able to ensure reproduction and is transferred from generation to generation, which ensures that evolution can continue because what is already a favorable combination of properties is conserved through the genes. Because of the very high number of amino acids (84,000,000), it is not surprising that there will always be minor differences in the amino acid sequences from frog to frog. These may be the result of mutations or of minor mistakes in the copying process. Such variation is important because it provides possibilities to "test" which amino acid sequence gives the best result with respect to survival and growth. The best result, representing the most favorable combination of properties, will offer the highest probability of survival and will support the highest reproductive rate, and the corresponding genes will therefore prevail. Survival and growth mean more exergy and a bigger distance to thermodynamic equilibrium. Exergy can therefore be used as a thermodynamic function to quantify Darwin's theory. It is interesting in this context that exergy also represents the amount of energy needed to tear down a system. This means that the more exergy the system possesses, the more difficult it becomes to kill the system and the higher is the probability of survival.

6. CLOSING REMARKS

Information carries energy and also "negative" entropy in the sense that information about organization reduces randomness. Total randomness corresponds to maximum entropy, i.e., no information. All information contains energy, even first-class energy, which can do work, i.e., exergy.

A complete description of a dynamic system requires that all exchanges of mass (all elements and chemical compounds), energy, and information are known. Mass, energy, and information are the basic building blocks of all systems. Although information carries energy (exergy), it is necessary to specify all the information transfer processes, because not all energy is associated with information. For instance, heat energy has no or very little information content. Mass and energy are also convertible according to Einstein's famous equation: energy $= mc^2$. For complete information, it is necessary to know the transfer processes of both mass and energy. In this context, it is notable that information has a relatively low cost of energy [see; for instance; Eq. (29)]. Mass on Earth is limited, and the amount of energy is also limited to the diurnal inflow of solar radiation. Anthropogenic use of previously stored solar radiation in the form of fossil fuels is also limited due to environmental impacts. It is therefore not surprising that humans recognize the need to continue development by focusing on generation of information and a wider and better use of information. Success of the "information age" is based on the reality that information has (almost) no limitation, except, of course, for its low energy cost. The amount of information on Earth cannot, of course, increase by more than the inflow of solar radiation, unless, over a limited time, previously stored energy (e.g., fossil fuel) is used. The solar energy received on $1\,m^2$ of Earth is about 2 GJ/year. In accordance with Eq. (29), this would correspond to information on the order of about 10^w, where $w \approx 10^{27}$, provided, of course, that 100% of the energy could be transformed into information. So, there is a long way to go before solar radiation is exhausted by conversion to information. The exergy content of an average human being (75 kg) would, according to Eq. (28) and Table I, be around 750 GJ, or the solar radiation on $375\,m^2$/year, but it has taken 4 billion years to reach the concentration of information found in a human being. These figures are maybe the best way to get an idea about the relationships between information and energy.

SEE ALSO THE FOLLOWING ARTICLES

Complex Systems and Energy • Entropy and the Economic Process • Exergoeconomics • Exergy Analysis of Energy Systems • Exergy: Reference States and Balance Conditions • Thermodynamics, Laws of • Value Theory and Energy

Further Reading

Boltzmann, L. (1905). "The Second Law of Thermodynamics. Populare Schriften, Essay No. 3" (address to the Imperial Academy of Science in 1886). [Reprinted in English in "Theoretical Physics and Philosophical Problems, Selected Writings of L. Boltzmann." D. Reidel, Dordrecht.]

Brillouin, L. (1949). Thermodynamics and cybernetics. *Am. Sci.* **37**, 554–568.

Brooks, D. R., and Wiley, E. O. (1986). "Evolution as Entropy." University of Chicago Press, Chicago.

Jørgensen, S. E. (2002). "Integration of Ecosystem Theories: A Pattern." 3rd Ed. Kluwer Academic Publ., Dordrecht, Boston, and London.

Jørgensen, S. E., and Meyer, H. F. (1977). Ecological buffer capacity. *Ecol. Modelling* **3**, 39–61.

Jørgensen, S. E., and Mejer, H. F. (1979). A holistic approach to ecological modelling. *Ecol. Modelling* **7**, 169–189.

Jørgensen, S. E., Patten, B. C., and Straskraba, M. (2000). Ecosystem emerging IV: Growth. *Ecol. Modelling* **126**, 249–284.

Jørgensen, S. E., and Bendoricchio, G. (2001). "Fundamentals of Ecological Modelling." 3rd Ed. Elsevier, Oxford.

Jørgensen, S. E., and Marques, J. C. (2001). Thermodynamics and ecosystem theory, case studies from hydrobiology. *Hydrobiologia* **445**, 1–10.

Kay, J., and Schneider, E. D. (1992). Thermodynamics and measures of ecological integrity. *In* "Ecological Indicators." pp. 159–182. Elsevier, Amsterdam.

Prigogine, I. (1980). "From Being to Becoming: Time and Complexity in the Physical Sciences." Freeman, San Francisco.

Shannon, C., and Weaver, W. (1949). "The Mathematical Theory of Communication." University of Illinois Press, Chicago.

Ulanowicz, R. E. (1986). "Growth and Development. Ecosystem Phenomenology." Springer-Verlag, Berlin and New York.

Ulanowicz, R. E. (1997). "Ecology, the Ascendent Perspective." Columbia University Press, New York.

Wicken, J. S. (1987). "Evolution, Thermodynamics and Information. Extending the Darwinian Program." Oxford University Press, Oxford.

Innovation and Energy Prices

DAVID POPP
Syracuse University
Syracuse, New York, United States

1. **The Concept of Technological Change**
2. **Macroeconomic Trends**
3. **Invention and Innovation**
4. **Diffusion**

Glossary

diffusion The final step of technological change; the process during which new innovations are gradually adopted by new users.
factor substitution A shift in the proportion of inputs (or factors) of production used that occurs in reaction to a change in the relative prices of these inputs.
induced innovation The theory that the direction and magnitude of innovative activity is shaped by market forces such as prices.
innovation The second step of technological change, where a new concept is developed into a commercially viable product.
invention The first step of technological change, marking the initial creation of a new product.
research and development (R&D) Spending by individual actors on the creation and development of new inventions and innovations.
technological change The process by which new technologies are developed and brought into use; includes three steps: invention, innovation, and diffusion.
total factor productivity (TFP) Output per unit of total input, where increases in the level of output produced for a given level of inputs suggest that the productivity of these inputs has increased.

Technological change has potentially broad impact in the field of energy. Improvements in energy efficiency enable production to proceed with less energy consumption per unit over time. The potential development of alternative energy sources, such as fuel cells and solar energy, offers prospects for environmental improvements and helps to address concerns over exhaustibility of current energy resources. As such, it is important to understand the process by which such technologies evolve. This article focuses on the means by which incentives, through either prices or policy, influence technological change in the energy sector.

1. THE CONCEPT OF TECHNOLOGICAL CHANGE

It is useful to understand how economists decompose the reaction of individual actors to a price change. Consider, for example, the reaction of a firm to higher energy prices. Economists use a production function to represent the relationship between inputs and outputs. Typical notation is

$$Y = f(L, K, E, M, t),$$

where Y is output, L is labor, K is capital, E is energy, M is materials, and t is technology. A production function can be represented by an isoquant that illustrates all possible combinations of labor and energy that can produce a given level of output. Figure 1 provides an example of a production process using two inputs: labor and energy. Note that along any one isoquant, technology is held constant; that is, each isoquant represents production possibilities for a given technology. The straight line P_0P_0 in the figure is an isocost line. The isocost line represents all combinations of labor and energy that cost the same. Costs increase for isocost lines moving away from the origin. The firm chooses how to produce a given level of output by finding where these two lines are tangent. At that point, the costs of production are lowest. Graphically, this is the lowest possible isocost line that touches the isoquant.

When energy prices rise, we expect the firm to use fewer energy inputs. In Fig. 2, an increase in energy prices makes the isocost line steeper, as shown by the line segment $P'_0P'_0$. Intuitively, for the same level of expenditure, the firm now purchases fewer units of

 © 2004 Elsevier Inc. All rights reserved.

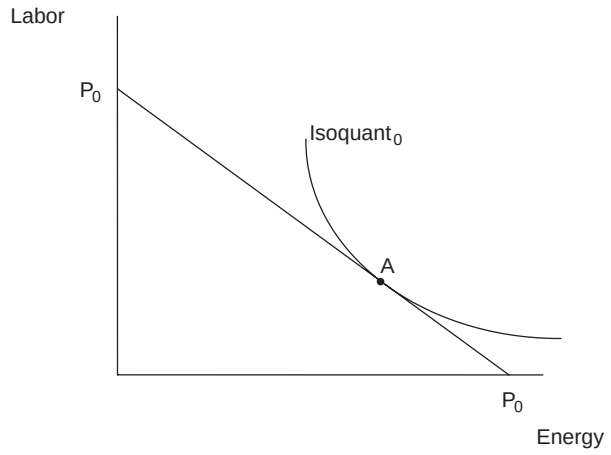


FIGURE 1 Diagram illustrating the use of an isoquant and isocost line to find the cost-minimizing choice of inputs. At time *t*, the production function is represented by isoquant$_0$. Along the isoquant, total output is constant. Initial prices are given by the isocost line P_0P_0. (Note: In this diagram, P does not represent actual prices but is simply a label for the relative price lines.) Initial production is at point A, where the two lines are tangent. This is the lowest possible isocost line that touches the isoquant. The interpretation is that this is the cheapest way possible to produce the level of output represented by the isoquant.

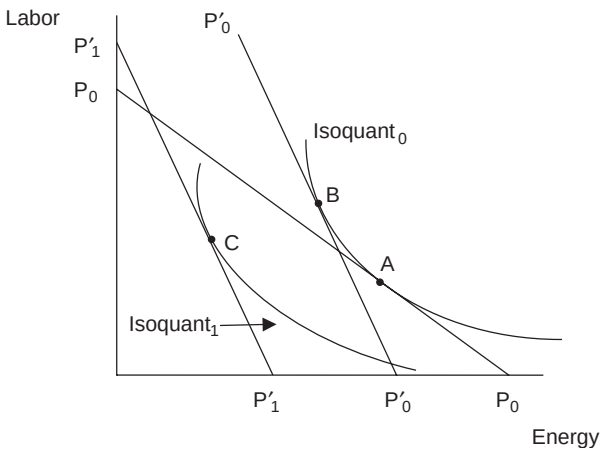


FIGURE 2 Diagram illustrating the distinction between factor substitution and technological change. At time *t*, the production function is represented by isoquant$_0$. Initial prices are given by the isocost line P_0P_0. Initial production is at point A. As energy prices rise, the isocost curve becomes steeper, as shown by the line $P'_0\,P'_0$. As a result of the price increase, the choice of inputs shifts to point B; that is, more labor and less energy are used. Because technology is constant along the isoquant, the shift from A to B represents factor substitution. During the next period, technological advances shift the isoquant to isoquant$_1$. If relative prices remain the same, the cost-minimizing combination of labor and energy is now at point C. The move from B to C is an example of technological change.

energy. As a result of the price increase, the cost-minimizing combination of labor and energy now involves more labor and less energy. An example of this would be a firm making less use of energy-intensive machinery and replacing the machines' capabilities with more manual labor. Such a change in reaction to a price change, illustrated by the shift from point A to point B on Fig. 2, is known as factor substitution. Note that the same production technology is being used but that different amounts of inputs are used in the process.

Of course, factor substitution is not the only reaction that a firm may have to higher energy prices. Over time, we would expect new machines that use energy more efficiently to be produced. New technology is represented by a new production function:

$$Y = f(L, K, E, M, t').$$

Because technology is constant along an isoquant, technological change results in an inward shift of the isoquant; the same level of output can now be produced using fewer inputs. In Fig. 2, technology shifts the isoquant from isoquant$_0$ to isoquant$_1$. Assuming that the ratio of prices for labor and energy remains the same, the choice of inputs moves from point B to point C. The movement from point B to point C on the figure is an example of technological change.

This process of technological change proceeds in stages. Writing in 1942, Schumpeter referred to the process as "creative destruction." First, an idea must be born. This stage is known as invention. New ideas are then developed into commercially viable products. This stage is referred to as innovation. Often, these two stages of technological change are studied together under the rubric of research and development (R&D). Finally, to have an effect on the economy, individuals must choose to make use of the new innovation. This adoption process is known as diffusion. At each stage, incentives, in the form of either prices or regulations, will affect the development and adoption of new technologies. This article begins with an overview of results pertaining to innovation and diffusion. A discussion of diffusion follows.

2. MACROECONOMIC TRENDS

To provide an introduction to the links between energy prices and innovation, it is helpful to look at trends in macroeconomic data. Figure 3 presents data on innovation, energy efficiency, and energy prices in the United States from 1972 to 1999.

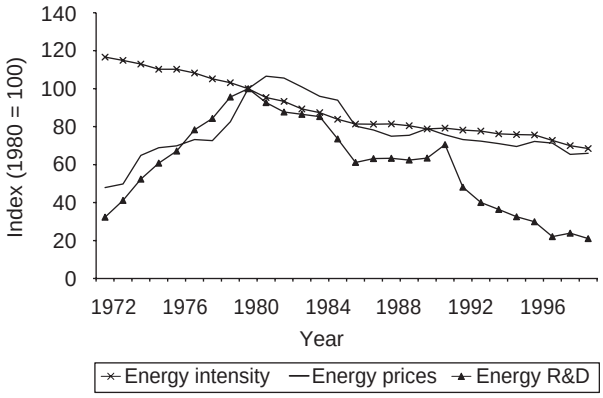


FIGURE 3 Diagram illustrating trends in energy prices, energy R&D, and energy intensity. All values are indexed so that 1980 = 100. As one would expect, there is a positive relationship between energy R&D and energy prices but an inverse relationship between both prices and R&D and energy intensity. Note that energy R&D begins to fall earlier than do energy prices, providing potential evidence of diminishing returns to energy R&D. Also note that energy intensity continues to fall even as energy prices and energy R&D both decline, suggesting continued diffusion of new energy technologies.

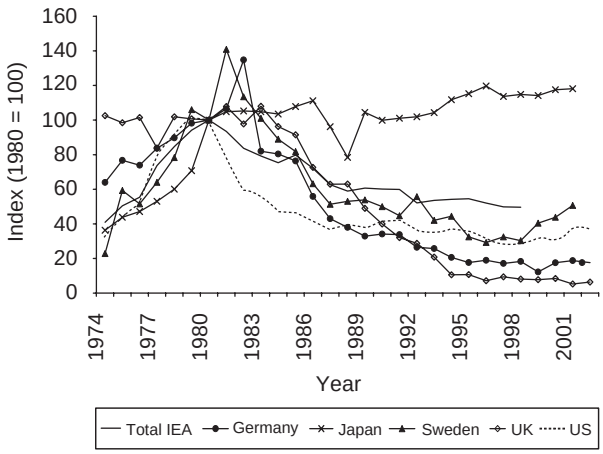


FIGURE 4 Diagram illustrating international trends in energy R&D spending. Values are indexed so that 1980 = 100. To create the index, all R&D figures have been converted to 2002 U.S. dollars. Total IEA spending includes totals for 23 IEA member countries (data for Luxembourg were unavailable). (Source: author's own calculations, based on data from the IEA Energy Technology R&D Statistics Service.)

Efficiency is represented by energy intensity—that is, the ratio of total energy consumption to real gross domestic product (GDP). The energy price index combines prices on all energy sources, weighted by consumption. Finally, energy R&D data represent all energy R&D performed by industry, whether financed by industry or by government. In each case, the data are indexed so that 1978 values equal 100.

The data provide a few interesting trends. First, note the strong relationship between energy prices and energy R&D. Such trends are found not just in the United States but also in other industrialized countries. Figure 4 shows the same energy R&D data for selected International Energy Agency (IEA) member countries as well as for a total of 23 member countries. Only Japan has maintained high energy R&D levels since energy prices have fallen.

Second, looking at the relationship between energy intensity and energy prices in Fig. 3, note that energy intensity fell as energy prices were rising. Price-induced substitution away from energy certainly played a role in the decline of energy intensity, especially immediately after the first energy crisis. However, simple substitution does not tell the whole story. Even when energy prices returned to pre-crisis levels, energy intensity remained low. If the technologies being used had not changed during the 1970s, there would be no reason for firms to continue using fewer energy inputs as energy prices returned to low levels. Note also that energy intensity did not level off until several years after energy prices and energy R&D had fallen. This suggests that technology played a role and that the diffusion of these new energy-efficient technologies took time.

Empirical analysis of these trends supports the interpretation that technological change played a role in improving energy intensity. A 2001 study by Popp, looking at energy consumption and innovation in 13 energy-intensive industries, found that approximately two-thirds of the decrease in energy consumption in response to higher energy prices comes from factor substitution, with the remaining one-third being due to the adoption of new technologies resulting from innovation induced by the price shock. At the macro level, studies of the effect of technological change on energy consumption typically regress energy consumption on the prices of energy and other inputs and include a time trend to capture the effects of technological progress. Early studies of industrial energy consumption that focused on data from before the first energy crisis of 1973 found that technological change was energy using; that is, energy use per unit output increased over time. Of course, using data from this period would ignore energy-saving innovations developed after the energy crises of the 1970s. Indeed, more recent work that included post-1973 data has found technological progress since the energy crisis to be energy saving.

3. INVENTION AND INNOVATION

3.1 Incentives and Innovation

To better understand the links between energy prices and innovation, it is important to look beyond the macroeconomic trends to understand how individual firms and consumers react to changes in energy prices. This section begins by looking at the process of innovation. To consider the links between energy prices and innovation, we need to begin with a more general question: what factors influence inventive activity? In general, economists focus on the ways in which individual actors respond to incentives. For a firm deciding whether or not to undertake a potential research project, potential incentives to consider include the likely cost of performing the research, the likelihood that the project will end with a successful result, and the potential payoffs from a successful project.

The hypothesis that higher energy prices lead to the development of more energy-saving technologies is derived from demand-pull theories of innovation. Demand-pull theories note the relationship between market forces and potential returns to an innovation. For example, higher energy prices increase the potential payoffs to energy research. Higher energy prices make energy-efficient inventions more valuable, either because the dollar value of potential energy savings is larger or because the market for energy-efficient inventions will be larger.

Other theories of innovative activity focus on the role that existing scientific knowledge plays in determining the pace and direction of technological change. Such theories, referring to the supply of knowledge available, are known as technology-push theories. Research is an uncertain activity. Even when energy prices are high, rational investors will not invest research dollars into energy-efficient research if the prospects of success are not good. Technology-push theories of technological change focus on the importance of the existing base of scientific knowledge to new inventions. The base of knowledge on which inventors can build helps to determine the likelihood of success. Technological advances make new inventions possible, and it is these breakthroughs that influence the direction of technological change. For example, the invention of microcomputer chips led to the development of a generation of electronic equipment. Thus, a thorough treatment of incentives and innovation must look at both demand- and supply-side influences.

The idea that the creation of ideas through invention and innovation will be influenced by economic incentives is not new. The concept of induced innovation was first introduced in 1932 by John Hicks, who noted that changes in factor prices would lead to innovation to economize on use of the more expensive factor. In this framework, one would expect higher energy prices to lead to greater research efforts to develop new energy sources and to improve energy efficiency.

As measures of innovative activity, such as patents and R&D, have become more readily available, empirical economists have begun to estimate the effects that prices and environmental policies have on energy and environmentally friendly innovation. Such studies take one of two approaches. First, authors have compared measures of innovative activity to environmental regulatory stringency. Typically, data on pollution abatement spending are used as a proxy for regulatory stringency in such studies. These studies find correlations between innovation and regulation both across nations and across time. Other studies use energy prices and related regulations as the mechanism that induces innovation. Although the observed price changes might not be policy related, the results can also be applied to situations where policy affects prices, such as a carbon tax. In one recent study, Newell, Jaffe, and Stavins examined the effect of both energy prices and energy efficiency regulations on technological advances in energy efficiency for air conditioners and natural gas water heaters. They found that both lead to more innovation and that energy prices have the largest inducement effect.

Figure 5 presents an example of typical findings in this literature. The figure shows how patenting

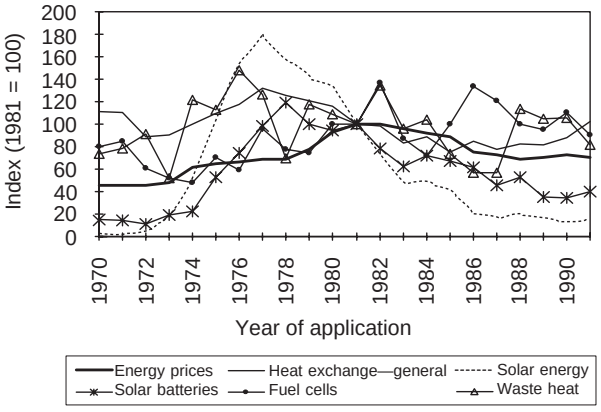


FIGURE 5 An index of patenting activity in five fields related to energy as well as an index of energy prices. Patenting activity is measured as successful U.S. patent applications per year. Only patents granted to U.S. inventors are included. From Popp (2002). Copyright 2002 American Economic Association; reproduced with permission of the *American Economic Review*.

activity for various energy technologies responded to changes in energy prices. Patents provide a detailed record of each invention. Classifications found on the front page of each patent can be used to identify its technological field. Economists have found that patents, sorted by their date of application, provide a good indicator of R&D activity. As a result, patent counts not only serve as a measure of innovative output but also are indicative of the level of innovative activity itself. Note that technologies such as solar energy experienced large jumps in patenting activity immediately following the first energy crisis. For example, there were just 10 solar energy patents in 1972. This figure jumped to 36 in 1973, 104 in 1974, and 218 in 1975. Thus, prices (or other regulations that increase the cost of using fossil fuels) can be expected to stimulate new research quickly.

Figure 5 also provides evidence of the importance of considering technology-push influences. Not only did energy R&D respond quickly to energy prices, but it also dropped off more quickly than did energy prices. Energy prices did not reach a peak until 1981. Nonetheless, patenting activity in most technologies peaked during the late 1970s. These data highlight the importance of the potential returns to R&D. Inventors "stand on the shoulders" of their predecessors. As a result, the quality of the knowledge stock available to an inventor is an important positive contributor to the level of innovative activity at any time. As innovation proceeds, it becomes more and more difficult to improve on the existing technology. In this fashion, doing research can be seen as analogous to drilling for oil. As more and more knowledge is drained from the "well," it becomes more difficult to achieve successful advances. The combination of standing on the shoulders of previous inventors and diminishing returns to research over time suggests that the level of induced R&D will fall over time.

3.2 The Returns to Research

As noted previously, in a market economy, researchers will not undertake a research project unless the expected returns justify the cost. Evidence from the returns to R&D comes from many sources such as case studies of successful innovation and econometric studies of firm productivity. Consistently, economists studying the returns to research have found that the returns to society as a whole are quite high. Most studies find rates of return of between 20 and 50%. Such rates are much higher than the rates of return on investment that firms typically receive. These high rates of return would suggest that further investments in R&D are warranted. Given this, why aren't R&D investments higher?

Two main characteristics about R&D investment help to explain this phenomenon. One is that research is an uncertain activity. The returns to research are highly skewed. Figure 6 illustrates the distribution of returns to individual research projects typically found in the literature. Although many projects yield little value, successful projects can yield returns of billions of dollars. High returns for successful projects help to compensate firms for the risk of undertaking projects that yield little or no returns. Second, the knowledge created by R&D is a public good. Firms capture the fruits of their research through sales of new products or via cost savings in production. However, these are not the only benefits that new knowledge provides to society. Once new knowledge is publicly available, it can be applied repeatedly without decay and can serve as the building block for other improvements. Moreover, without public policy such as intellectual property rights, others cannot be excluded from making use of the new knowledge. As a result, the firm is unable to capture all of the benefits of new knowledge to society, so that the social returns to R&D are higher than the private returns to R&D. Typical findings are that the social rates of return are approximately four times higher than private rates of return.

The high social returns of research have two important implications. First is that government policy can play an important role in the creation of new knowledge. Government policies help either by enabling firms to capture more of the benefits of their research (e.g., by providing monopolies via patent protection) or by financing research that the private sector is unwilling to conduct on its own. The latter

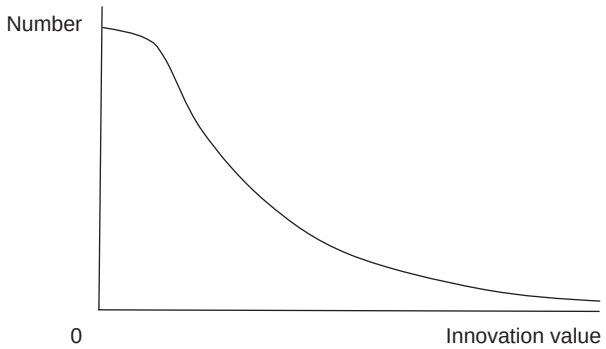


FIGURE 6 Diagram showing a typical distribution of the returns to R&D. The bulk of innovations have little or no commercial value. However, the right tail of the distribution includes innovations that are worth billions of dollars.

is particularly important for basic research that broadens general knowledge but does not have specific commercial applications. For example, basic advances in chemistry and physics are essential to developing new solar energy technologies. Because basic research serves as the building block for many other research efforts, and because the commercial benefits of basic research might not be readily apparent, firms are particularly reluctant to take on basic research projects. Indeed, basic research carried out by the government often provides the impetus for future research projects carried out by the private sector.

Second, high social returns suggest that the potential opportunity costs of doing R&D must be considered when looking at energy R&D policy. High social returns exist for all types of research, not just energy research. Given this, consider a simple economy in which output, Q_t, is devoted to consumption, C_t, investment in physical capital, I_t, or two types of research—energy, $R_{E,t}$, and other, $R_{O,t}$—so that

$$Q_t = C_t + I_t + R_{E,t} + R_{O,t}.$$

When energy R&D increases by \$1, \$1 less is available for the other activities. The opportunity cost of the first two is simply valued at \$1. However, if the social rate of return is four times higher for R&D than for other investments, giving up \$1 of other R&D activity has the same effect as giving up \$4 of other investment.

This is important because empirical work suggests that at least some increases in energy R&D will come at the cost of other forms of R&D. Research activities are carried out by highly trained scientists and engineers. Because years of training are needed to enter the field, the supply of scientists and engineers available at any one time is inelastic; that is, it cannot increase quickly when new research needs arise. For example, a 1998 study by Goolsbee found that one of the chief beneficiaries of R&D tax subsidies are scientists and engineers, who receive larger wages when subsidies are increased.

Applying this to energy R&D, a simple regression of energy and total R&D spending in the United States suggests that approximately one-half of the energy R&D spending that took place during the 1970s and 1980s came at the expense of other R&D. Thus, at least part of the increased energy R&D of the 1970s and early 1980s came at the expense of other potentially productive R&D projects. At the same time, productivity of the economy as a whole fell during the energy crisis. Researchers are still trying to grapple with the macroeconomic effects of such a dramatic shift in research efforts. Much empirical work has focused on potential causes of decreased productivity during the 1970s, and the general conclusion is that increased energy prices are not enough to explain the entire decrease. Nonetheless, increased energy prices and the resulting shift in innovation toward energy efficiency played at least a partial role.

4. DIFFUSION

Of course, induced invention and innovation is only half the story. Technological advances are of little use unless society makes use of the innovation. Thus, diffusion is also important. Diffusion is likely to play a particularly important role for problems dealing with long-term consequences such as climate change. Most innovation takes place in highly industrialized countries. In 1998, 85% of all R&D in the Organization for Economic Cooperation and Development (OECD) countries was conducted in just seven countries. In fact, roughly 44% of all OECD R&D was done in the United States alone. Thus, as policymakers pay increased attention to potential increases in fossil fuel consumption for fast-growing countries such as China, it is important to pay attention to not only how policy will help to induce the development of new technologies but also how policy can help to encourage the adoption of these technologies in developing countries.

4.1 Diffusion within a Country

Studies of the diffusion of individual technologies consistently find that diffusion is a gradual process. Typically, the rate of diffusion can be represented by a sigmoid, or "S-shaped," curve over time; that is, the rate of adoption rises slowly at first, speeds up, and then levels off as market saturation approaches. Figure 7 illustrates such a curve, with market penetration plotted on the y axis and the passage of time plotted on the x axis.

Traditionally, researchers have used one of two approaches to modeling diffusion of a new technology. Each offers a slightly different explanation for the pattern observed in Fig. 7. The epidemic model of diffusion proposes that information is the primary factor limiting diffusion. Adoption is slow at first because few people (or firms) know about the technology. However, as more people adopt the technology, knowledge of the technology spreads

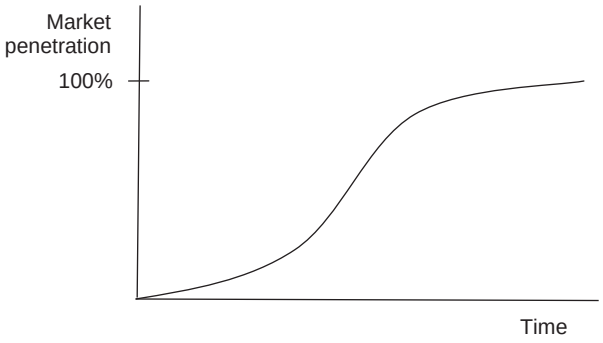


FIGURE 7 Diagram showing how the adoption of new technologies varies over time. Initially, only a few early adopters will choose to use a new innovation. Eventually, mass adoption takes place until only a few stragglers remain. As these stragglers eventually adopt the technology, adoption approaches 100% of the potential market.

quickly, leading to a period of rapid adoption. Economists often use the analogy of a contagious disease to describe this period of adoption; the more people who are "infected" by the technology, the more likely that others will also become "infected." Eventually, few potential adopters remain, as nearly everyone has adopted the technology, so that the rate of adoption levels off again. In a classic 1957 study using this framework, Griliches noted that the rate of diffusion is at least partially determined by economic factors such as the expected rate of return for adoption. Other work using the epidemic model typically focuses on firm characteristics such as firm size to explain variations in the rate of diffusion. These studies find that larger firms adopt new innovations first. Such firms are more likely to have access to capital necessary for new investments, and their size enables them to better bear the risks that come with the adoption of any new technology.

The second approach to studying diffusion focuses is the probit model. The probit model focuses on heterogeneity among firms. In this model, firm heterogeneity leads to a distribution of expected returns from adopting the new technology. Because adoption is costly, only firms above a threshold great enough to justify the costs of adoption will choose to adopt the technology at any given time. Over time, the technology gets cheaper and its quality improves, so that more firms cross the adoption threshold. In this framework, environmental policy and energy prices enter by increasing the expected returns from adopting new energy technologies, thus, increasing the likelihood that firms will cross the threshold. As would be expected, studies using this framework find that higher energy prices encourage the adoption of more energy-efficient technologies. Similarly, the presence of stronger environmental regulations also increases the probability of adopting such technologies. Nonetheless, the response still varies by firm characteristics, and again larger firms tend to adopt first.

Note that the two models suggest different interpretations of the gradual rate of diffusion. The epidemic model suggests that adoption generates positive externalities and, thus, is slower than would be optimal. In this framework, policies to encourage early adoption of new technologies, such as alternative energy sources, would be desired to ensure rapid deployment of new technologies. Conversely, the probit model suggests that gradual diffusion is optimal because differences in adoption decisions simply result from profit-maximizing decisions made by heterogeneous firms. Recent work combining these two models suggests that variations across firms are the main explanation for differences in adoption, as suggested by the probit model.

Turning to adoption studies that focus on energy-efficient technology, a common theme emerges: such technologies are underused. In many cases, individual firms fail to adopt technologies for which the present discounted value of energy savings is greater than the costs of the new technologies. Research suggests several possible explanations for this "energy efficiency paradox." One is that consumers lack information about the potential energy savings that new technologies offer. A second is that consumers place more importance on the upfront installation costs of new technologies than on potential future savings. This may be because they are uncertain about future savings (the upfront installation costs are certain, but the potential energy savings are uncertain given that energy prices may change), because they face credit constraints, and/or because they truly value future consumption at a low rate.

4.2 International Diffusion

The studies mentioned previously focus on the adoption of a single technology across individual actors within a single country. At the international level, there is a large body of research on technological diffusion generally but little work focusing on energy technologies specifically. One exception is a 1996 study by Lanjouw and Mody that looked at patenting activity for various environmental and energy-efficient technologies. This study found that the majority of environmental technology patents in developing countries came from foreign countries.

Moreover, it found that policies in one nation may affect innovation of technologies in another nation. For example, the authors noted that the majority of vehicle air emissions patents granted in the United States were from foreign nations, even though the United States was the first country to adopt strict emissions standards.

Given the importance of diffusion for bringing new energy technologies to developing countries, understanding the general lessons from empirical work on international diffusion is important. There are two potential avenues through which foreign knowledge can have an influence in the domestic economy. First, foreign inventions may be adopted directly by domestic firms. Second, rather than being adopted directly by domestic users, foreign knowledge may affect the productivity of domestic R&D. The "blueprints" represented by foreign patents may serve to inspire additional innovation by domestic inventors. Such productivity increases are knowledge spillovers in that the knowledge represented by the foreign patent creates a positive externality—knowledge from abroad that is borrowed by domestic inventors.

Because most R&D takes place in just a few countries, both types of diffusion are important. For direct adoption, many of the same lessons from diffusion studies within a country apply here as well. In particular, local conditions are an important determinant of the rate of diffusion across countries. Regarding knowledge spillovers, researchers have followed several strategies. At a macro level, several published articles estimate the effects of domestic and foreign R&D on total factor productivity (TFP) growth. Typically, accumulated levels of R&D are combined to create both domestic and foreign stocks of R&D effort. The foreign R&D stocks are weighted by international trade flows, so that these articles attempt to measure the effect of international trade on knowledge flows. These articles find a positive effect for foreign R&D, although this effect may vary by country. In general, the relative contribution of foreign knowledge is inversely related to economic size and level of development. At a more micro level, studies looking at the effect of foreign direct investment (FDI) and foreign knowledge spillovers have found mixed results. Whether the effect of FDI on growth is positive or negative depends on individual industry and country characteristics, with inward FDI being more successful for developed countries. These studies suggest that the ability of a country to successfully absorb new knowledge, such as having a well-educated workforce, is important for successful technology transfer.

SEE ALSO THE FOLLOWING ARTICLES

Business Cycles and Energy Prices • Depletion and Valuation of Energy Resources • Energy Futures and Options • Inflation and Energy Prices • Prices of Energy, History of • Stock Markets and Energy Prices • Technology Innovation and Energy • Transitions in Energy Use

Further Reading

Binswanger, H. P., and Ruttan, V. W. (eds.). (1978). "Induced Innovation: Technology, Institutions, and Development." Johns Hopkins University Press, Baltimore, MD.

Goolsbee, A. (1998). Does government R&D policy mainly benefit scientists and engineers? *Am. Econ. Rev.* **88,** 298–302.

Griliches, Z. (1957). Hybrid corn: An exploration of the economics of technological change. *Econometrica* **25,** 501–522.

Jaffe, A. B., Newell, R. G., and Stavins, R. N. (2003). Technological change and the environment. *In* "Handbook of Environmental Economics." (K. G. Mäler and J. Vincent, Eds.), Vol. 1, pp. 461–516. North-Holland, Amsterdam, Netherlands.

Lanjouw, J. O., and Mody, A. (1996). Innovation and the international diffusion of environmentally responsive technology. *Res. Policy* **25,** 549–571.

Newell, R. G., Jaffe, A. B., and Stavins, R. N. (1999). The induced innovation hypothesis and energy-saving technological change. *Q. J. Econ.* **114,** 941–975.

Popp, D. (2001). The effect of new technology on energy consumption. *Resource Energy Econ.* **23,** 215–239.

Popp, D. (2002). Induced innovation and energy prices. *Am. Econ. Rev.* **92,** 160–180.

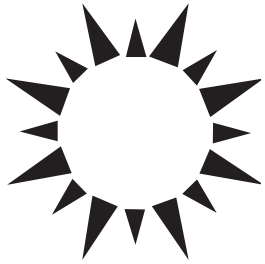
Rosenberg, N. (1982). "Inside the Black Box: Technology and Economics." Cambridge University Press, New York.

Schumpeter, J. (1942). "Capitalism, Socialism, and Democracy." Harper, New York.

Stoneman, P. (ed.). (1995). "Handbook of the Economics of Innovation and Technological Change." Blackwell, Cambridge, MA.

Input–Output Analysis

STEPHEN D. CASLER
Allegheny College
Meadville, Pennsylvania, United States



1. **Introduction**
2. **The Input-Output Model**
3. **Extensions of the Basic Model**
4. **Secondary Products**
5. **The Energy-Based Input-Output Model**
6. **Structural Decomposition Analysis: Evaluating Changes over Time**
7. **Conclusions**

Glossary

energy intensity A measure of the total Btu's of energy needed both directly and indirectly to produce a dollar's worth of a product for sale to final demand.

fixed coefficients In production theory, a situation in which the required quantity of each input used is a constant multiple of the level of output.

intermediate product Output from an industry that is used as an input by another producing sector, as opposed to output sold to final demand.

inverse coefficient A measure of the total value of output needed both directly and indirectly to produce a dollar's worth of a product for sale to final demand.

open model In input-output analysis, a formulation of the model in which all elements of final demand for goods and services are assumed to be exogenously determined.

producer prices Prices that exclude the transportation costs and wholesale and retail trade margins that are included in the actual prices paid by purchasers.

secondary product A good or service produced by an establishment that differs in classification from the establishment's principal or primary product.

structural decomposition analysis The evaluation and measurement of causes of change in input-output relations through time.

technological/structural coefficient The ratio of the amount of an input used by a sector to the sector's total output, assumed to be constant under the fixed-coefficients assumption.

The input-output approach to analysis of economic activity provides a concise means of representing interrelationships between producing, final demand, and value-added components of the economy at any point in time. As its name implies, the range of inputs used by industries is accounted for, as is the sale of outputs by industries to other industries and to final demand. Because these supply and demand relationships are considered simultaneously for the full range of producing sectors, the input-output model provides a general equilibrium framework useful for analyzing all transactions within the economy.

1. INTRODUCTION

The input-output model's historical origins lie with the French Physiocrat Francois Quesnay, whose *Tableau Economique* portrayed the nature of exchange between farmers, landlords, and manufacturers in the mid-1700s. Fundamental theoretical and empirical work on the modern model was conducted by Wassily Leontief, who was awarded a Nobel Prize in economics for his contributions. The model lends itself to applications ranging from studies of local to national economies. Relative to other forms of analysis, the input-output model's main attributes are its consideration of the full range of production activities, the number of sectors it can accommodate, its ability to measure the mutual interdependence among sectors of the economy in meeting final demand requirements, and its reliance on data from only one time period. The model is also very flexible. With slight modification, it enables the study of issues related to energy use, employment, pollution emissions, and much more.

1.1 The National Income and Product Accounts and the Input-Output Model

The relations that underlie the input-output model are organized around the various measures of income

 © 2004 Elsevier Inc. All rights reserved.

and output found in the national income and product accounts. This consistency is seen in the equivalent measures of gross national product as the sum of final expenditure on goods and services, the sum of the factor payments, and the sum of the values added. The data used in the input-output model show inter-industry sales of intermediate products among the industries or sectors of the economy. Intermediate products are produced in one sector of the economy and then used as inputs by other sectors in the production of their output. The model also accounts for sales to final demand—goods and services purchased for personal consumption, investment, government activities, and net exports (exports minus imports). Finally, value added, the wages, interest, profit, and rent paid to factors of production during the production process by each industry, is included.

To illustrate how inter-industry sales, final demand, and value added are connected in the input-output framework, consider an economy composed of three producing sectors: agriculture, manufacturing, and services. Transactions among these sectors are shown in the first three rows and columns of Table I. The arrows pointing to the right indicate the sales of the goods and services produced by the sectors listed in the rows. For example, in row 1, agricultural output is shown as being sold for use as an input to the agriculture sector itself and to manufacturing and services. Its output is also sold to the various components of final demand, consumption, C, investment, I, government, G, and net exports, X. The sum of the sales of agricultural output across this row equals the total value of output produced by the agriculture sector. A similar interpretation holds for the rows of the table showing sales of output by manufacturing and services.

Within columns of the table, for any sector, the value of each of the inputs used by that sector is measured. Intermediate inputs are measured in the first three rows of each column. Further down the table, within the value-added row, are the total factor payments to labor, capital, entrepreneurs, and land for each producing sector. The sum of these factor payments equals total value added for each sector. The sum of the values added generated by all sectors equals gross national product, as does the sum of the values of products sold to final demand. Finally, the value of total output produced by any sector—sales of its output to producing sectors plus output sold to final demand—is equal to the sum of the values of intermediate inputs used by that sector plus that sector's value added. Corresponding row and column sums are equal in value.

In the United States, the input-output accounts are constructed by the U.S. Department of Commerce, Bureau of Economic Analysis. Benchmark accounts are based primarily on economic census data, which are collected by the Bureau of the Census at 5-year intervals. The 1992 benchmark accounts measure the inputs used and outputs sold by 498 industries. The inputs and outputs of the model are generally measured in dollars, but they can also be measured in physical units. The accounts published by the Bureau of Economic Analysis are dollar based, and goods and services are valued using producer prices.

2. THE INPUT-OUTPUT MODEL

There are two ways in which the output produced by any industry in the economy can be used: as an input by other producers and to satisfy final demand. For example, some of the corn produced in the agriculture sector is used in that sector to produce more corn, some is sold as feed to cattle producers, some is used in the production of corn-based sweeteners by manufacturers, and some is sold directly to consumers as part of final demand. Such transactions by all producing sectors are the foundation of the input-output model. For applications and manipulations of the model, the range of sales to producers and final demand is represented using linear equations. Using the standard notation and the convention that when two subscripts are present the first refers to input and the second output, X_{ij} represents the total amount of input i used to produce output j, X_i represents the total output of sector i, and Y_i represents the total final demand for output i.

2.1 The Fundamental Balance Equation

In equation form, the relation between the total production of output i and its various uses across the economy is represented as

$$X_i = X_{i1} + X_{i2} + \ldots + X_{in} + Y_i. \qquad (1)$$

That is, the total output of good i on the left side of the equation equals the sum of good i sold as an input to all n producing sectors of the economy, the X_{ij}'s, plus the final demand for good i. Equation (1) shows the fundamental balance between the output produced by a sector and all its potential uses.

TABLE I

Interindustry Transactions, Final Demand, and Value Added[a]

	Intermediate sales			Final demand	
	Agriculture	**Manufacturing**	**Services**	$C+I+G+X$	**Total output**
Agriculture	→ Agriculture input to agriculture	→ Agriculture input to manufacturing	→ Agriculture input to services	→ Agriculture sales to final demand	Total agriculture output
Manufacturing	→ Manufacturing input to agriculture	→ Manufacturing input to manufacturing	→ Manufacturing input to services	→ Manufacturing sales to final demand	Total manufacturing output
Services	→ Services input to agriculture	→ Services input to manufacturing	→ Services input to services	→ Services sales to final demand	Total services output
Value added	Value added by agriculture	Value added by manufacturing	Value added by services	Total value added = Total final demand = GNP	
Total output	The value of inputs to agriculture plus its value added equals the value of agriculture output.	The value of inputs to manufacturing plus its value added equals the value of manufacturing output	The value of inputs to services plus its value added equals the value of services output		

[a] The input-output model is organized around the sales of products used as inputs by industries and consumed by final demand. There is a balance between the value of a sector's total output and the value of its inputs plus value added.

2.2 Technological Coefficients

A key assumption of the input-output model is that output is related to input use through fixed coefficients. This assumption implies that to double the amount of output produced requires that all inputs be exactly doubled, which in turn implies constant returns to scale (but not vice versa). The fixed-coefficients assumption means that a sector's use of any input is a constant multiple of its output. In equation form, this relationship is expressed as

$$X_{ij} = a_{ij}X_j, \tag{2}$$

where a_{ij} represents the ij^{th} technological or structural coefficient, the units of input i required to produce a single unit of output j. For example, suppose sector i produces eggs and sector j produces cakes. If it takes two eggs to produce one cake, then $a_{ij} = 2$; if it is desired to produce 500 cakes, then $X_{ij} = 2 \times 500$, or 1000 eggs. If input-output relations are measured as values rather than physical units, physical inputs and outputs are weighted by their respective prices in forming technological coefficients. Thus, if the price of an egg is 50 cents and a cake costs \$3, the technological coefficient corresponding to two eggs per cake is 0.33, 33 cents of egg input is required per dollar of cake output. Production of \$1500 of cake output requires \$500 worth of egg input. Incorporating the definition shown in Eq. (2) into Eq. (1), the balance equation between sector i's total output and its various uses across the economy becomes

$$X_i = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{in}X_n + Y_i. \tag{3}$$

2.3 Solving for Total Outputs

When all the sectors of the economy are considered, a system of simultaneous linear equations results. The endogenous variables are the outputs of each sector, the X_i's; the fixed technological coefficients or a_{ij}'s are parameters, and the Y_i's or final demands are exogenous variables. This system of equations for an economy of n sectors appears as

$$\begin{aligned} X_1 &= a_{11}X_1 + a_{12}X_2 + \ldots + a_{1n}X_n + Y_1 \\ X_2 &= a_{21}X_1 + a_{22}X_2 + \ldots + a_{2n}X_n + Y_2 \\ &\vdots \\ X_n &= a_{n1}X_1 + a_{n2}X_2 + \ldots + a_{nn}X_n + Y_n. \end{aligned} \tag{4}$$

On the right-hand side of every equation, output levels X_1 through X_n are present, and each is multiplied by a unique technological coefficient, a_{ij}. All goods are not necessarily required as inputs to particular sectors; some of the technological coefficients can be zero.

Using matrix algebra, the various elements of the model are conveniently organized in terms of outputs, technological coefficients, and final demand. The equivalent matrix-algebra representation of the n equations shown in Eq. (4) is

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ & \vdots & & \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} + \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}. \quad (5)$$

With the letter X representing the column vector of total outputs, A the $(n \times n)$ matrix of technological coefficients, and Y the column vector of final demands, the entire system of equations is expressed simply as

$$X = AX + Y. \quad (6)$$

The outputs contained in the vector X are the unknowns in the model, and they can be solved for as follows. First, rearranging Eq. (6),

$$X - AX = Y$$

or

$$(I - A)X = Y.$$

Here, the letter I is the identity matrix, which serves many of the same functions in matrix algebra as the number 1 does in standard algebra. It consists of one's along its main diagonal with zeros elsewhere. These equations imply that final demand is equal to what remains when goods used as inputs by producing sectors are subtracted from total output. However, final demand is not a residual—the purpose of production is to satisfy a given level of final demand. The solution for the vector of total outputs that will do so is

$$X = (I - A)^{-1}Y. \quad (7)$$

The $n \times n$ matrix $(I-A)^{-1}$ is known as the Leontief inverse. Solving the system of equations in terms of the inverse, $(I-A)^{-1}$, is simply a more efficient means of finding a solution compared with solving the system through substitution. Its use makes it especially easy to solve for the new output levels that arise when there is a change in the final demand vector—rather than re-solving the entire system by substitution, it is only necessary to post multiply the existing Leontief inverse by the new final demand vector to find the new levels of outputs.

Elements of the Leontief inverse account for the full range of interrelationships between producing sectors. For example, a \$1 rise in the final demand for a product simultaneously results in increased demand for all inputs used by the sector that produces it. But producing this first set of inputs implies an increased demand for the inputs required by all sectors that supply the initial set, and so on. Thus, inverse coefficients measure the direct and indirect requirements of a good that are necessary to produce \$1 of some output for final demand.

2.4 Finding the Inverse and Total Output: Numerical Example

To illustrate the interpretation of the Leontief inverse and the mathematics involved in solving for the vector of sectoral outputs, consider a hypothetical economy consisting of two goods, coal and steel. It is assumed that the output of each good is measured in dollars. Production of either good is also assumed to require both coal and steel. Thus, to produce \$1 of coal requires 10 cents of coal and 40 cents of steel. To produce \$1 of steel requires 55 cents of coal and 20 cents of steel. In the notation of the input-output model, with sector 1 designated as coal and sector 2 as steel,

$$a_{11} = 0.10,\ a_{21} = 0.40,\ a_{12} = 0.55,\ a_{22} = 0.20.$$

2.4.1 The Technological Matrix

Using the hypothetical numerical values, the technological or structural A matrix from Eqs. (6) and (7) is

$$A = \begin{bmatrix} 0.10 & 0.55 \\ 0.40 & 0.20 \end{bmatrix}.$$

The rows of the technological matrix show how much of each particular input is required to produce \$1 of output in the two producing sectors of this economy. Columns reveal the input requirements of each sector. For sector 1, coal, a total of 50 cents of intermediate inputs is required to produce \$1 worth of coal output, 10 cents of coal and 40 cents of steel. The sum of technological coefficients in each column must be less than 1 or else the production process is not economically feasible—producing \$1 of output should not cost more than a dollar. Aside from intermediate inputs, in producing a dollar's worth of

output, the remaining value comes from value added. Hence, the remaining 50 cents of the \$1's worth of coal output is comprised of coal's value added. Similarly, based on the second column of the technological matrix, which shows the steel sector's input requirements, a total of 75 cents of intermediate inputs is required to produce \$1 of steel; therefore, to make up the rest of the dollar's worth of steel output, 25 cents of each dollar of steel is comprised of value added.

2.4.2 Solving for the Leontief Inverse

The solution for the economy's vector of outputs requires formation of the Leontief inverse. The foundation for this inverse is the matrix $(I - A)$, calculated as

$$(I - A) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 0.10 & 0.55 \\ 0.40 & 0.20 \end{bmatrix} = \begin{bmatrix} 0.90 & -0.55 \\ -0.40 & 0.80 \end{bmatrix}.$$

Based on the procedure used to find the inverse of a matrix,

$$(I - A)^{-1} = \begin{bmatrix} 1.6 & 1.1 \\ 0.8 & 1.8 \end{bmatrix}.$$

2.4.2.1 Interpreting the Inverse Coefficients As previously noted, the values within the inverse represent the direct and indirect output necessary to supply \$1 of output for final demand. Thus, to supply \$1 of coal for final demand requires a total value of coal production of \$1.60. One dollar of this amount represents the dollar's worth of coal that is required to satisfy final demand. The remaining 60 cents represents the value of the coal used to produce coal and well as the value of coal used to produce the steel that is necessary for coal production, and so on.

As another example, the direct and indirect coal required to produce \$1 of steel equals \$1.10. The interpretation and implicit derivation of this measure of direct and indirect coal use can be understood by considering the various *linkages* between the coal and steel sectors. From the technological matrix, A, when steel output increases by \$1, the steel sector requires 55 cents of coal directly ($a_{12} = 0.55$). However, since it takes 10 cents of coal to produce \$1 of coal ($a_{11} = 0.1$), an additional 5.5 cents of coal is needed *indirectly* to produce the 55 cents of coal used by the steel industry ($0.1 \times 0.55 = \$0.55$). In addition, production of a \$1 of coal requires 40 cents of steel ($a_{21} = 0.4$). Since it takes 55 cents of coal to produce \$1 of steel, to produce 40 cents of steel it takes 22 cents of coal ($0.55 \times 0.4 = \$0.22$). It also takes steel to produce steel ($a_{22} = 0.2$). But the production of the 40 cents worth of steel used by coal requires the use of 8 cents of steel ($0.2 \times 0.4 = \$0.08$), which again implies the use of more coal to produce it. When the direct requirement for coal by the steel sector is added to the infinite chain of all such indirect requirements, the inverse coefficient 1.1 is derived.

2.4.2.2 Output Multipliers The sums of column entries within the Leontief inverse are known as output multipliers. Given a \$1 increase in final demand for a particular good, the corresponding output multiplier provides a measure of the resulting change in total output in the economy. Output multipliers reveal where a dollar of spending on some component of final demand will have the largest impact on total output. For example, the output multiplier for the coal sector is 2.4—a \$1 rise in final demand for coal will lead to a rise in overall production of \$2.40, with \$1.60 of output from the coal sector and \$0.80 of output from the steel sector. Alternatively, if the final demand for steel rises by \$1, the output multiplier is 2.9.

2.4.3 Finding Total Outputs

In addition to the measurement of the direct and indirect requirements per dollar of output, one of the most important uses of the input-output model is the determination of the total output necessary to satisfy some level of final demand. This determination involves a direct application of Eq. (7). For example, suppose final demand for coal and steel equals \$100 and \$200, respectively ($Y_1 = \$100$ and $Y_2 = \$200$). From Eq. (7), total output is calculated as

$$X = (I - A)^{-1} Y$$

or

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 1.6 & 1.1 \\ 0.8 & 1.8 \end{bmatrix} \begin{bmatrix} 100 \\ 200 \end{bmatrix}.$$

The equivalent algebraic expression is

$$X_1 = (1.6 \times 100) + (1.1 \times 200) = \$380$$
$$X_2 = (0.8 \times 100) + (1.8 \times 200) = \$440.$$

These equations imply that to satisfy a \$100 final demand for coal and a \$200 final demand for steel requires total coal production of \$380 and steel production of \$440. Alternatively, if final demand is *predicted* to rise to \$120 for coal and \$210 for steel,

estimated coal and steel output equals \$423 and \$474, respectively.

2.5 The Model's Accuracy over Time

Economists understand that the technological coefficients used to form the Leontief inverse are only representative of input requirements during the time period for which data are collected, because relative price changes and technological change lead to changes in input requirements per unit of output. However, there is general agreement that the model accurately approximates the impact on output of changes in final demand for short periods of time around the date for which the table was formulated. For example, the Leontief model based on data for 1939 was able to correctly predict steel requirements at the end of World War II. Therefore, even though benchmark tables are not available on a yearly basis in the United States and are generally published several years after the benchmark date, the data they contain are still useful for analysis of the state of the economy in nearby years.

3. EXTENSIONS OF THE BASIC MODEL

A number of important extensions and applications of the input-output model are seen in empirical investigations. The most widely used include the price model, which makes it possible to estimate the prices of outputs that are consistent with physical technological coefficients and value added; the labor model, which enables the estimation of direct and indirect labor requirements and the employment levels necessary to meet a given level of final demand; the closed model, in which elements from final demand and value added are made endogenous; and the pollution-based model, which provides a means of linking production activities to the emission of pollutants.

3.1 The Leontief Price Model

The technological coefficient a_{ij} can be measured as the physical units of input i per unit of output j or the dollars of input i required to produce \$1 of output j. In the Leontief price model, the a_{ij}'s are measured in physical units, for example, the tons of coal it takes to produce a ton of steel. Multiplication of the ij^{th} technological coefficient by the i^{th} price, $P_i a_{ij}$, results in a measure of the dollars of input i required per physical unit of output j (dollars per unit of input i times units of i per unit of j equals dollars of input i per unit of output j). For example, if 2 tons of coal are required to produce 1 ton of steel and if the price of coal is \$4 per ton, then \$8 of coal are required to produce a ton of steel. The total value of all intermediate inputs required to produce one physical unit of output j is therefore equal to the sum of each of sector j's price-weighted input coefficients.

In addition to the per-unit value of all the intermediate inputs used in its production, the price of a good is comprised of the value added per unit of output that is generated through the use of labor and other factors of production. For example, if 40 cents of coal and \$1.60 of steel are required as intermediate inputs to produce a ton of coal, and value added per ton of coal is \$2, then the price of coal is \$4. When all sectors of the economy are considered, the following system of price equations results:

$$\begin{aligned} P_1 &= P_1 a_{11} + P_2 a_{21} + \ldots + P_n a_{n1} + va_1 \\ P_2 &= P_1 a_{12} + P_2 a_{22} + \ldots + P_n a_{n2} + va_2 \\ &\vdots \\ P_n &= P_1 a_{1n} + P_2 a_{2n} + \ldots + P_n a_{nn} + va_n. \end{aligned}$$

Each price on the left-hand-side of the equation is seen to equal the sum of the values of the given sector's intermediate inputs per unit of output plus the sector's value added per unit of output. In matrix algebra, this system of products and sums is

$$P = PA + va. \tag{8}$$

where P is a $1 \times n$ vector of prices and A is the same matrix of technological coefficients found in the standard model, with coefficients measured in physical units of input per physical unit of output. Based on Eq. (8), the vector of prices is solved for as

$$P = va(I - A)^{-1}. \tag{9}$$

Equation (9) shows that the row vector of prices is equal to the value added row vector postmultiplied by the Leontief inverse.

3.1.1 The Price Model: Numerical Example

To illustrate the calculation of prices, consider the two-sector economy comprised of coal and steel production. As in the previous example, coal is sector 1 and steel is sector 2. In physical units, it takes $a_{11} = 0.1$ tons of coal to produce a ton of coal, $a_{12} = 1.375$ tons of coal to produce a ton of steel, $a_{21} = 0.16$ tons of steel to produce a ton of coal, and $a_{22} = 0.2$ tons of steel to produce a ton of steel. The economy's matrix

of physical-unit technological coefficients is therefore

$$A = \begin{bmatrix} 0.100 & 1.375 \\ 0.160 & 0.200 \end{bmatrix}.$$

The coefficients for coal's use of coal and steel's use of steel are exactly the same as in the value-based example—since the same price of coal applies to both the coal used as an input and the total amount of coal produced, the prices in the numerator and denominator cancel, with a similar effect for steel's use of steel. Subtracting the matrix of technological coefficients from the identity matrix and solving for the inverse yields

$$(I-A)^{-1} = \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 0.100 & 1.375 \\ 0.160 & 0.200 \end{bmatrix} \right)^{-1} = \begin{bmatrix} 1.60 & 2.75 \\ 0.32 & 1.80 \end{bmatrix}.$$

Suppose value added per ton of coal is \$2 and value added per ton of steel is \$2.50. Based on Eq. (9), the solution for the economy's prices is

$$P = va(I-A)^{-1} = [2.0 \quad 2.5] \begin{bmatrix} 1.60 & 2.75 \\ 0.32 & 1.80 \end{bmatrix} = [4 \quad 10].$$

The price of coal is \$4 per ton and the price of steel is \$10 per ton.

3.1.2 Prices and the Dollar-Based Model

Prices can also be calculated using the dollar-based technological matrix and the vector of dollar-based values added:

$$P = va(I-A)^{-1} = [0.5 \quad 0.25] \begin{bmatrix} 1.6 & 1.1 \\ 0.8 & 1.8 \end{bmatrix} = [1 \quad 1].$$

Prices estimated using the dollar-based model always equal 1. This result follows from the fact that such prices represent the dollar value of output per *dollar* of output (versus dollars per *unit* of output for a standard price). The dollar value of output per dollar of output equals 1 by definition. For each good, such prices can be thought of as price indices, which equal 1 in the base year or, in this case, the year in which the input-output model is formulated. Using a dynamic version of the price model, the effects on prices of changes in elements of the technological matrix or the vector of values added can be analyzed relative to this base. Such changes are discussed in the section on structural decomposition analysis.

3.2 Direct and Indirect Labor Requirements

Another important and widely used application of the input-output model is the estimation of direct and indirect labor requirements—measures of the total labor required given a \$1 increase in final demand for an output. These measures can also be used to estimate the economy's total employment given any level of final demand. Calculation of such labor intensities requires data on the direct labor input requirements per dollar of output for each sector of the economy. For example, to produce \$1000 of coal might require five workers, while production of a \$1000 of steel might require four workers. Estimation of the row vector of direct and indirect labor requirements, λ, simply involves postmultiplication of the $(1 \times n)$ vector of direct labor requirements per dollar of output, L, by the value-based Leontief inverse:

$$\lambda = L(I-A)^{-1}. \tag{10}$$

3.2.1 Direct and Indirect Labor Requirements: Numerical Example

Using the value-based Leontief inverse for the two-sector economy comprised of coal and steel, if it takes five workers to produce \$1000 of coal and four workers to produce \$1000 of steel, the direct labor coefficients per dollar of output are 0.005 workers per dollar of coal and 0.004 workers per dollar of steel. The economy's direct and indirect labor requirements are then calculated as

$$\lambda = L(I-A)^{-1} = [0.005 \quad 0.004] \begin{bmatrix} 1.6 & 1.1 \\ 0.8 & 1.8 \end{bmatrix} = [0.0112 \quad 0.0127].$$

To produce \$1000 of coal requires a direct and indirect labor input of 11.2 workers, while production of \$1000 of steel requires 12.7 workers from all sectors of the economy.

The logic underlying these measures can be seen for the coal sector as follows. Since it takes 0.005 workers to produce \$1 of coal and \$1.60 of coal directly and indirectly to produce \$1 of coal for final demand, (0.005×1.6) or 0.008 workers are required everywhere in the economy to produce the \$1.60 of direct and indirect coal output. Coal production also requires steel, and \$1 of steel production requires 0.004 workers. Therefore, production of the \$0.80 of steel required directly and indirectly to produce \$1 of

coal takes 0.0032 workers. The sum of these values equals the 0.0112 direct and indirect workers necessary to produce \$1 of coal for final demand.

3.2.1.1 Measuring Total Labor Requirements Finding the economy's total labor requirements to satisfy a given level of final demand simply involves postmultiplication of Eq. (10) by the final demand vector. If final demand for coal is \$100 and final demand for steel is \$200, total employment, E, is 3.66 workers:

$$E = L(I-A)^{-1}Y = [0.0112 \quad 0.0127]\begin{bmatrix}100\\200\end{bmatrix} = 3.66.$$

This result follows directly from the fact that $(I-A)^{-1}Y$ measures total output by sector, which for the simple economy used in the value-based example led to production of \$380 of coal and \$440 of steel. If it takes five workers to produce \$1000 of coal, it takes 1.9 workers to produce \$380; if four workers are needed to produce \$1000 of steel, 1.76 workers are required to produce \$440 of steel. Total employment in the economy is therefore, $(1.90 + 1.76)$, 3.66 workers.

3.3 Closing the Model with Respect to Labor

Rather than using direct employment coefficients to calculate direct and indirect labor requirements, it is possible to form similar measures by closing the input-output model with respect to households and the labor services they provide. In general, closing an input-output model refers to incorporating elements from final demand and value added directly into the matrix of technological coefficients. In the case of labor, some or all of the consumption goods used by households within final demand are assumed to be the inputs required to produce labor services (e.g., food is a necessary input for physical human effort). This consumption component is stripped out of final demand and included as an additional column of the inter-industry transactions table. Division of consumption expenditure for each good by total consumption expenditure results in the technological coefficients for the production of labor services. Similarly, payments for labor services by industries are stripped out of each sector's value added and included as an additional row of the transactions table. Division by the value of each sector's output results in technological coefficients that measure the value of labor input per dollar of output for each industry. Given these adjustments, the expanded technological matrix is used to find the Leontief inverse, which now includes dollar-based direct and indirect labor requirements.

3.4 Direct and Indirect Pollution Emissions

The estimation of direct and indirect labor requirements makes use of direct labor coefficients, measures of the direct labor input required to produce \$1 of a sector's output. Such coefficients can be defined for any variable not included in the inter-industry transactions table and are generally defined as direct impact coefficients. For example, if there is an interest in measuring direct and indirect water use, the direct impact coefficients are gallons of water used per dollar of output for each sector, and direct and indirect water requirements can be solved for in the same way that direct and indirect labor requirements are found.

An important use of such coefficients is the estimation of direct and indirect pollution emissions. For example, suppose production of each \$1000 of coal leads to 10 tons of carbon dioxide emissions, while production of a \$1000 of steel results in 20 tons of emissions. The direct-impact coefficients are then 0.01 and 0.02. Analogous to calculation of direct and indirect labor requirements, direct and indirect carbon dioxide emissions are solved for as

$$\pi = E(I-A)^{-1} = [0.01 \quad 0.02]\begin{bmatrix}1.6 & 1.1\\0.8 & 1.8\end{bmatrix} = [0.032 \quad 0.047].$$

Here, E represents the row vector of carbon dioxide's direct emissions coefficients for each sector. The Leontief inverse, $(I-A)^{-1}$, is estimated using the value-based model. The measures that result from this calculation show that each \$1000 of coal produced results in 32 tons of direct and indirect carbon dioxide emissions, while the total emissions everywhere in the economy from the production of \$1000 of steel equals 47 tons.

Total carbon dioxide emissions for a given level of final demand are found by postmultiplying the row vector of direct and indirect emissions by the column vector of final demand. Based on the previous hypothetical values, total emissions are

$$ETOT = \pi Y = [0.032 \quad 0.047]\begin{bmatrix}100\\200\end{bmatrix} = 12.6.$$

Since each dollar of coal production results in direct and indirect carbon dioxide emissions of 0.032 ton, production of $100 of coal for final demand leads to 3.2 tons. Added to the 9.4 tons of emissions induced by the final demand for steel, $(0.047 \times \$200)$, satisfaction of the economy's total final demand leads to 12.6 tons of carbon dioxide emissions.

4. SECONDARY PRODUCTS

Complicating the estimation of input-output coefficients for real-world applications are secondary products, products that do not fall within the classification of goods that are primary to the industry that produced them. For example, steel is the primary product of the steel industry, but steel producers may produce aluminum as well. If secondary products are not accounted for, sales of output from the steel industry to other sectors will give misleading estimates of the amount of steel used in particular production processes—some buyers of output from the steel industry might really be using aluminum to produce their output.

To deal with the secondary product problem, a distinction is made between a commodity and an industry. A commodity is any well-defined good or service, consistent with the government's classification of products at a specified level of aggregation. An establishment can produce many different commodities but is classified into a particular industry based on the primary commodity produced—the commodity that makes up the bulk of the establishment's output.

One of two assumptions is generally used to reallocate industrial inputs to each distinct commodity that is produced. First is the industry-technology assumption, which treats secondary products as if their production required the same inputs as those used to produce total output in the industry where they are produced. For example, if the steel sector produces aluminum as a secondary product, and aluminum constitutes 20% of the steel industry's output, then 20% of each of the steel industry's inputs is assigned to the production of aluminum as a commodity. The other assumption used in estimating commodity input requirements is the commodity technology assumption, under which the composition of inputs needed to produce a unit of commodity output is the same regardless of the industry in which the commodity is produced. The industry technology assumption is most appropriate in cases where the mix of inputs used by primary producers is fundamentally the same whether secondary products are produced or not. Examples include chemicals that are secondary products of refining operations and the extraction of coal as a secondary product in metal ore-mining operations. Alternatively, the commodity technology assumption is most appropriate when only one input recipe is associated with the production of an output.

4.1 The Use and Make Tables

Under either the industry technology or commodity technology assumption, reallocation of inputs to the production of commodities involves two matrices: the use and make tables. The rows of the use table represent commodities, while its columns represent industries. Each cell in the use table shows the amount of the commodity listed on the row that is used as an input by the industry listed in the column. For example, the amount of the commodity coal used as an input by the steel industry is found at the intersection of the coal row and steel column. The sum of each row in the use table, which measures all intermediate commodity sales to industries, plus sales to final demand equals total commodity output.

As its name implies, the make table shows all the commodities that are produced by particular industries. Industries are listed along the rows, and the commodities they produce are designated in the columns. The main diagonal of the make table shows the values of commodity outputs produced by the industries to which they are primary. For example, the intersection of the coal row and column shows the total amount of the commodity coal produced by the coal industry. If some steel output is a secondary product of the coal industry, the value of this steel production is found at the intersection of the coal row and the steel column. Similarly, the intersection of the steel row and the coal column shows the coal produced as a secondary product by steel. The sum of values across any row of the make table equals total industry output for the industry listed on that row. The sum of values down any column of the make table accounts for all producers of a commodity and equals the total output of that commodity.

4.2 Estimating Commodity Output under the Industry Technology Assumption

Since it is used most frequently in estimating input-output relations when secondary products are present, the means by which the industry technology

assumption leads to a commodity-based transactions table and inverse is illustrated. To produce their primary and secondary products \$13.42 of the commodity coal is used by the coal industry, \$266.58 of coal is used by the steel industry, \$171.81 of the commodity steel is used by the coal industry, and \$68.19 of the commodity steel is used by the steel industry. In the use matrix, these commodity sales to industries appear as

$$U = \begin{bmatrix} 13.42 & 266.58 \\ 171.81 & 68.19 \end{bmatrix}.$$

The first row of this table shows the use of the commodity coal by the coal and steel industries, respectively, while the second row shows the use of steel by these industries. If final demand for coal is \$100 and final demand for steel is \$200, the total commodity output of coal is equal to \$380, the sum of elements in the first row of the use table plus final demand for coal. Similarly, the total commodity output of steel is \$440.

Next, suppose that \$338 of the commodity coal is produced by the coal industry—its primary product, along with \$62 of steel—its secondary product. The steel industry produces \$42 of coal as a secondary product and \$378 of steel. The economy's make matrix is

$$V = \begin{bmatrix} 338 & 62 \\ 42 & 378 \end{bmatrix}.$$

Adding elements across rows of the make table equals total industry output. Therefore, the total value of output in the coal industry, comprised of \$338 of coal and \$62 of steel, is \$400. Similarly, the total value of output produced by the steel industry is \$420. Adding entries in the columns of the make matrix yields total commodity output, \$380 of coal and \$440 of steel, which are the same values found by adding across rows of the use matrix, plus final demand.

Through the industry technology assumption, commodity inputs to industries in the use table are allocated to the production of commodities in proportion to the fraction of industrial output they comprise. These proportions are calculated by premultiplying the make matrix by the matrix $\hat{g}^{-1}$, which is composed of the reciprocals of industry outputs along its main diagonal, with zeros elsewhere:

$$\hat{g}^{-1}V = \begin{bmatrix} 1/400 & 0 \\ 0 & 1/420 \end{bmatrix} \begin{bmatrix} 338 & 62 \\ 42 & 378 \end{bmatrix} = \begin{bmatrix} 0.845 & 0.155 \\ 0.100 & 0.900 \end{bmatrix}.$$

The first row of the resulting matrix product shows that 84.5% of the output of the coal industry is coal, and 15.5% is steel. Similarly, 10% of the steel industry's output is coal and 90% is steel. Under the industry technology assumption, 84.5% of the commodity inputs to the coal industry found in the use matrix is used in the production of the commodity coal. This fraction of the coal industry's inputs is added to 10% of the inputs used by the steel industry to yield the total commodity input requirements of the commodity coal. Similarly, 15.5% of the coal industry's commodity inputs is assumed to be used in the production of the commodity steel, along with 90% of the inputs to the steel industry. The transformation of the use table into a table of commodity-by-commodity requirements is seen in the matrix product

$$U(\hat{g}^{-1}V) = \begin{bmatrix} 13.42 & 266.58 \\ 171.81 & 68.19 \end{bmatrix} \begin{bmatrix} 0.845 & 0.155 \\ 0.100 & 0.900 \end{bmatrix} = \begin{bmatrix} 38 & 242 \\ 152 & 88 \end{bmatrix}.$$

The coal industry used \$13.42 of coal to produce both coal and steel, but its coal output constituted 84.5% of total industry production. Hence, 84.5% of \$13.42 (\$11.34) from the coal industry is allocated to production of the commodity coal as an input requirement. The steel industry used \$266.58 of coal to produce both steel and coal. Since 10% of steel's output is coal, 10% of \$266.58 (\$26.66) is assumed to be used for coal production. The total amount of the commodity coal used to produce coal by all industries in the economy is therefore \$38. A similar explanation holds for the other elements in the final matrix product.

To form commodity-based technological coefficients, commodity inputs are divided by their respective total commodity outputs. This transformation is accomplished using the matrix $\hat{q}^{-1}$, which consists of the reciprocals of commodity outputs along its main diagonal, with zeros elsewhere. The resulting matrix is defined as BD. Using matrix algebra, the complete transformation of the use table into a commodity-based matrix of technological

coefficients is seen as

$$BD = (U\hat{g}^{-1})(V\hat{q}^{-1}) = (U\hat{g}^{-1}V)\hat{q}^{-1}$$
$$= \begin{bmatrix} 38 & 242 \\ 152 & 88 \end{bmatrix} \begin{bmatrix} 1/380 & 0 \\ 0 & 1/440 \end{bmatrix}$$
$$= \begin{bmatrix} 0.10 & 0.55 \\ 0.40 & 0.20 \end{bmatrix}.$$

Finally, industry technology commodity-by-commodity total requirements are estimated as

$$(I - BD)^{-1} = \begin{bmatrix} 1.6 & 1.1 \\ 0.8 & 1.8 \end{bmatrix}.$$

Elements within this inverse are interpreted in the same way as elements within the usual Leontief inverse, $(I - A)^{-1}$, except that they relate to direct and indirect commodity input requirements (versus industry requirements) needed to produce \$1 of commodity output for final demand. Postmultiplication of the inverse by the final demand vector yields the vector of total commodity outputs.

5. THE ENERGY-BASED INPUT-OUTPUT MODEL

The energy-based input-output model uses a mixed or hybrid-unit modification of the standard approach to measure direct and indirect energy input use. It traces its development to the work of Hannon, Herendeen, and Bullard. In the model, energy flows are measured in physical units, usually British thermal units or Btu's, while the output of nonenergy sectors is measured in dollars. This formulation enables analysts to determine which sectors of the economy are the most energy-intensive in terms of direct energy inputs, the energy it takes to produce energy, and in the use of energy to produce nonenergy inputs. Relative to the dollar-based model, it provides an accurate means of estimating the economy's physical energy requirements for a given level of final demand and of estimating changes in energy requirements given final demand changes. Measurement in physical units provides useful information regarding the depletion of finite energy products such as crude oil, natural gas, and coal. The model can be easily modified to measure the emissions of pollution that result from the combustion of fossil fuels. Finally, the energy-based model provides a foundation for evaluating direct and indirect changes in energy efficiency and understanding the factors responsible for such changes.

5.1 Elements of the Energy-Based Model

In the standard input-output model, the balance between the total output produced by some sector and its use as an input across sectors and final demand is

$$X_i = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{in}X_n + Y_i,$$

where a_{ij} represents the dollars of input i used to produce \$1 of output j, X_j represents the total value of output j, and Y_i is the value of final demand for good i. In the energy input-output model, if a sector produces one of the energy products coal, crude oil and natural gas, refined petroleum products, gas utilities, or electricity, its energy use and sales are measured in Btu's. The matrix that shows the use of energy inputs by energy producers, nonenergy producers, and final demand is called the direct energy transactions matrix. For the energy sectors of the economy, it supplants the dollar-based flows of the standard input-output table.

With energy use and total energy output measured in Btu's and nonenergy input use and total output measured in dollars, there are four ways in which the units of the energy input-output model's technological and corresponding inverse coefficients appear. The technological coefficient a_{ij} measures Btu's of input per dollar of output when input i is an energy product and output j is a nonenergy product. It measures Btu's of input per Btu of output if a technological coefficient for energy use by an energy-producing sector is considered. For nonenergy inputs to energy sectors, the units of a_{ij} are dollars of input per Btu of output, since the outputs of energy producing sectors are measured in Btu's. Finally, if sales of nonenergy inputs to nonenergy producers are considered, the units of a_{ij} are the standard dollars of input per dollar of output.

In constructing an energy-based model, it is customary to rearrange the industrial classification system so that the energy inputs to producing sectors are shown in the top rows of the table, with nonenergy producing sectors below. With input-output sectors so arranged, Table II shows the units of technological and inverse coefficients for the various dollar and energy-based measures.

Based on the mixed-unit technological matrix, energy intensities are calculated using the equation

$$\varepsilon = e(I - A)^{-1}. \quad (11)$$

TABLE II

Units of Technological and Inverse Coefficients in the Energy-Based Model[a]

	Energy producers	Non-energy producers
Energy inputs	Btu/Btu	Btu/$
Nonenergy inputs	$/Btu	$/$

[a]Using the mixed-unit formulation of energy-based input-output model, technological and inverse coefficients are measured in four possible ways.

Equation (11) shows that the economy's energy intensities are a simple function of the mixed-unit Leontief inverse. Specifically, the matrix e selects the energy rows from the inverse. In the general case of m energy products, ε is an $m \times n$ matrix of energy intensities and e is an $m \times n$ matrix composed of ones and zeroes, with a 1 appearing in each row in the location that corresponds to the energy-producing sector.

5.1.1 Primary Energy Intensities

A summary measure of the Btu's of all energy types used by a sector is known as a primary energy intensity—a weighted sum of energy intensities for each fuel, where, to avoid double counting, weights are based on the fraction of total output comprised of newly extracted energy produced by fuel sectors. Examples of primary energy include oil from oil wells, energy obtained from geothermal sources and water power, and energy released from nuclear reactors. In calculating the primary energy intensity for each sector, the weight for the coal and crude oil and natural gas intensities is 1, because 100% of the energy output of these sectors comes from the earth. Alternatively, the weight for refined petroleum products is about 0.05, because only 5% of the energy sold by that industry originates as a refined petroleum commodity. The rest comes from sales of crude oil input to refining operations.

5.1.2 The Energy Input-Output Model: Numerical Example

Using the two-sector model composed of coal and steel production, suppose the coal sector requires 950 thousand Btu's of coal and steel requires 6050 thousand Btu's of coal to produce their total output. Final demand for coal is 2500 thousand Btu's. Hence, the total output of coal is 9500 thousand Btu's. To produce coal, $152 of steel is required by the coal sector, and $88 of steel is required by the steel sector. Final demand for steel is $200. Total output of steel therefore equals $440. Dividing the inputs to each sector by total sectoral output leads to the formation of the hybrid-units technological matrix, A.

$$A = \begin{bmatrix} 950/9500 & 6050/440 \\ 152/9500 & 88/440 \end{bmatrix} = \begin{bmatrix} 0.10 & 13.75 \\ 0.016 & 0.20 \end{bmatrix}.$$

Each Btu of coal production directly requires 0.1 Btu's of coal (the 1000s of Btu's in the numerator and denominator cancel). Production of 1000 Btu's of coal requires 1.6 cents of steel input. Each dollar of steel output requires 13.75 thousand Btu's of direct coal input and 20 cents of direct steel input.

Using Eq. (11), the solution for the economy's vector of energy intensities is

$$\varepsilon = e(I - A)^{-1} = [1 \quad 0] \begin{bmatrix} 1.600 & 27.5 \\ 0.032 & 1.8 \end{bmatrix} = [1.6 \quad 27.5].$$

To produce 1 Btu of coal for delivery to final demand requires a total of 1.6 Btu's of coal everywhere in the economy (i.e., the coal required to produce coal, and the coal required to produce the steel used by coal as an input, and so on). To produce $1 of steel output requires 27.5 thousand Btu's of coal, directly and indirectly.

Consistent with the energy conservation condition, the product of the vector of energy intensities and the mixed-unit final demand vector equals the Btu equivalent of total coal production:

$$\varepsilon Y = [1.6 \quad 27.5] \begin{bmatrix} 2500 \\ 200 \end{bmatrix} = 4,000 + 5,500 = 9,500.$$

Note that this relationship corresponds in principle to the identity in the all-dollar model that relates the economy's vector of total outputs, X, to final demand, Y, $X = (I - A)^{-1}Y$.

5.1.3 Energy Input-Output Models: The Accuracy of Alternative Techniques

Using the dollar-based table and direct energy coefficients (ratios of Btu's of direct energy use to the value of total output) to calculate energy intensities in a manner analogous to calculating labor intensities will lead to inaccurate results when the prices paid for energy products differ across using sectors, which is generally the case. Such price differences result from variations in the mix of fuels

that make up energy input aggregates (bituminous versus lignite under the classification coal, for example) or price differences brought about by the volume of fuel purchased. The problem with energy intensities calculated using direct energy coefficients and the dollar-based inverse lies with the correct estimation of total physical energy use given changes in final demand. For example, if the price of coal used by the coal sector differs from the price of coal paid by steel, the direct and indirect dollars of coal used by coal and steel will be associated with different quantities of coal, depending on whether the coal use occurs in the coal or steel sector. The hybrid-units approach does not suffer from this problem—it will correctly estimate total energy requirements for any level of final demand.

6. STRUCTURAL DECOMPOSITION ANALYSIS: EVALUATING CHANGES OVER TIME

Applications of the input-output model can be extended to measure and explain sources of change in the economy through time. The evaluation of such changes is known as structural decomposition analysis, which traces its roots to Leontief, Anne P. Carter, and Hollis B. Chenery. The modern classification of studies falling under this heading is attributable to Rose and Miernyk. As a form of comparative static analysis, the object of most structural decompositions is to isolate the separate sources of changes in the variables that comprise the input-output model and to measure their importance. The foundation for such analysis begins with the fundamental input-output identity that relates the economy's vector of total output to final demand:

$$X = (I - A)^{-1}Y.$$

From one time period to the next, assuming that changes are infinitesimal, the change in the vector of outputs is

$$\Delta X = \Delta(I - A)^{-1}Y + (I - A)^{-1}\Delta Y. \qquad (12)$$

Based on Eq. (12), there are two major sources of change. First, for a given final demand, changes in the vector of outputs are seen to result from changes in the Leontief inverse, $\Delta(I - A)^{-1}$, which are brought on by a change in the technological matrix, A. In turn, such change stems from variations in any of the n^2 structural coefficients within the A matrix. Second, changes in output will occur given a change in the final demand vector, ΔY, for a given Leontief inverse. These changes are caused by the overall growth of final demand and by variations in the demand for particular products within and across final demand categories. Underlying changes in both structural coefficients and final demand are the full range of economic forces, including relative price changes, economies of scale, and technological change.

6.1 Decomposing Changes in Final Demand

The input-output model makes it possible to estimate new output requirements given changes in final demand by simply postmultiplying the Leontief inverse by the new or predicted final demand; however, it is possible to extend the range of measured effects. For example, the final demand vector can be decomposed so that energy and nonenergy sectors are considered separately. The inverse matrix is assumed to be constant from one period to the next. Given n sectors in the economy, Y_t^e is an $n \times 1$ vector that shows the final demand in period t for energy products, with zeros elsewhere, and Y_t^{ne} is an $n \times 1$ vector showing final demand for nonenergy products in period t, with zeros in the energy locations. The sum of these vectors equals the final demand vector in period t:

$$Y_t = Y_t^e + Y_t^{ne}.$$

Given a change in final demand from period t to $t + 1$, the overall change in final demand is $\Delta Y = Y_{t+1} - Y_t$, which can be expressed in terms of energy and nonenergy inputs as

$$\Delta Y = \Delta Y^e + \Delta Y^{ne}.$$

Using these distinct changes in final demand, the overall change in output for all sectors of the economy can be broken down into that caused by changes in final demand for energy and that caused by changes in final demand for nonenergy products:

$$\Delta X = (I - A)^{-1}\Delta Y^e + (I - A)^{-1}\Delta Y^{ne}.$$

6.1.1 Structural Decomposition of Final Demand: Numerical Example

Using the two-sector dollar-based model, suppose that final demand for coal rises from \$100 to \$120, while final demand for steel rises from \$200 to \$240.

The overall change in final demand is

$$\Delta Y = \begin{bmatrix} 20 \\ 40 \end{bmatrix}.$$

Based on the standard output identity, the overall changes in coal and steel outputs are \$76 and \$70, respectively.

$$\begin{bmatrix} \Delta X_1 \\ \Delta X_2 \end{bmatrix} = (I - A)^{-1}\Delta Y = \begin{bmatrix} 1.6 & 1.1 \\ 0.8 & 1.8 \end{bmatrix}\begin{bmatrix} 20 \\ 40 \end{bmatrix} = \begin{bmatrix} 76 \\ 88 \end{bmatrix}.$$

In terms of energy and nonenergy effects, final demand change is decomposed as

$$\Delta Y = \Delta Y^e + \Delta Y^{ne} = \begin{bmatrix} 20 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 40 \end{bmatrix}.$$

The estimated changes in both coal and steel output caused solely by the change in final demand for the energy product coal are \$32 and \$16, respectively:

$$\begin{bmatrix} \Delta X_1^e \\ \Delta X_2^e \end{bmatrix} = (I - A)^{-1}\Delta Y^e = \begin{bmatrix} 1.6 & 1.1 \\ 0.8 & 1.8 \end{bmatrix}\begin{bmatrix} 20 \\ 0 \end{bmatrix} = \begin{bmatrix} 32 \\ 16 \end{bmatrix}.$$

The changes in output for both coal and steel due to the change in final demand for the nonenergy product steel equal \$44 and \$72, respectively:

$$\begin{bmatrix} \Delta X_1^{ne} \\ \Delta X_2^{ne} \end{bmatrix} = (I - A)^{-1}\Delta Y^{ne} = \begin{bmatrix} 1.6 & 1.1 \\ 0.8 & 1.8 \end{bmatrix}\begin{bmatrix} 0 \\ 40 \end{bmatrix} = \begin{bmatrix} 44 \\ 72 \end{bmatrix}.$$

Aside from breaking final demand into mutually exclusive groups of rows that include changes for particular sectors in the economy (energy, mining, manufacturing, etc.), it is also possible to consider the effects of changes in the separate categories of final demand—consumption, investment, government spending, and net exports. In addition, it is possible to isolate the effects of changes in the mix and level of final demand on the level of output.

6.2 Decomposing Changes in the Inverse

Changes in the Leontief inverse are brought about through changes in the elements of the technological matrix, A—that is, the a_{ij}'s. One approach to separating and measuring the effects of such change begins with the definition of inverse change and leads to an additive set of partial inverse-change effects, whose sum will equal the true change in the inverse, when changes in technological coefficients are infinitesimal. To understand this approach to inverse decomposition, let $\gamma = (I - A)^{-1}$. It follows that

$$\gamma(I - A) = I \text{ or } \gamma - \gamma A = I.$$

Through the product rule of calculus, the overall change in the inverse is

$$\Delta\gamma - \Delta\gamma A - \gamma\Delta A = 0$$

or

$$\Delta\gamma = \gamma\Delta A\gamma. \tag{13}$$

From Eq. (13), the change in the inverse over two time periods is equal to the change in the structural matrix, pre- and postmultiplied by the base-period inverse. Using this equation, changes in the inverse induced by changes in individual technological coefficients or groups of coefficients can be evaluated. Possible groupings include input coefficients for each sector (the columns of the technological matrix) or the use of particular inputs by all sectors of the economy (the rows of the technological matrix). Using the two-sector model in which coal energy requirements are shown in row 1 and requirements for the nonenergy product steel are shown in row 2, the overall change in the technological matrix can be written

$$\Delta A = \Delta A^e + \Delta A^{ne} = \begin{bmatrix} \Delta a_{11} & \Delta a_{12} \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \Delta a_{21} & \Delta a_{22} \end{bmatrix}.$$

Using this definition of the change in the structural matrix in Eq. (13) yields the distinct effects of inverse change brought about by changes in energy and nonenergy input requirements:

$$\Delta\gamma = \gamma(\Delta A^e + \Delta A^{ne})\gamma = \gamma(\Delta A^e)\gamma + \gamma(\Delta A^{ne})\gamma.$$

Postmultiplication by the final demand vector results in measures of the separate effects on total output brought on by the changes in the energy and nonenergy inputs.

6.2.1 Structural Decomposition of the Inverse: Numerical Example

With sector 1 producing coal and sector 2 producing steel, suppose the technological matrix and inverse in

period t are

$$A_t = \begin{bmatrix} 0.10 & 0.55 \\ 0.40 & 0.20 \end{bmatrix} \text{ and}$$

$$\gamma = (I - A_t)^{-1} = \begin{bmatrix} 1.6 & 1.1 \\ 0.8 & 1.8 \end{bmatrix}.$$

In period $t+1$, suppose the technological coefficients change such that

$$A_{t+1} = \begin{bmatrix} 0.12 & 0.60 \\ 0.48 & 0.24 \end{bmatrix}.$$

The overall change in the technological matrix and the distinct effects of changes in coal and steel input requirements are

$$\Delta A = \begin{bmatrix} 0.02 & 0.05 \\ 0.08 & 0.04 \end{bmatrix} = \begin{bmatrix} 0.02 & 0.05 \\ 0.00 & 0.00 \end{bmatrix} + \begin{bmatrix} 0.00 & 0.00 \\ 0.08 & 0.04 \end{bmatrix}.$$

Given a \$100 final demand for coal and a \$200 final demand for steel, an estimate of the change in output for both the coal and steel sectors caused solely by the changes in coal energy requirements is

$$\begin{bmatrix} \Delta X_1^e \\ \Delta X_2^e \end{bmatrix} = (\gamma \Delta A^e \gamma) Y$$

$$= \left(\begin{bmatrix} 1.6 & 1.1 \\ 0.8 & 1.8 \end{bmatrix} \begin{bmatrix} 0.02 & 0.05 \\ 0.00 & 0.00 \end{bmatrix} \begin{bmatrix} 1.6 & 1.1 \\ 0.8 & 1.8 \end{bmatrix} \right) \begin{bmatrix} 100 \\ 200 \end{bmatrix}$$

$$= \begin{bmatrix} 47.36 \\ 23.68 \end{bmatrix}.$$

Similarly, the estimated change in outputs caused solely by changes in technological coefficients associated with the use of the nonenergy-input steel is

$$\begin{bmatrix} \Delta X_1^{ne} \\ \Delta X_2^{ne} \end{bmatrix} = (\gamma \Delta A^{ne} \gamma) Y$$

$$= \left(\begin{bmatrix} 1.6 & 1.1 \\ 0.8 & 1.8 \end{bmatrix} \begin{bmatrix} 0.00 & 0.00 \\ 0.08 & 0.04 \end{bmatrix} \begin{bmatrix} 1.6 & 1.1 \\ 0.8 & 1.8 \end{bmatrix} \right) \begin{bmatrix} 100 \\ 200 \end{bmatrix}$$

$$= \begin{bmatrix} 52.80 \\ 86.40 \end{bmatrix}.$$

6.2.2 Fields of Influence and Inverse Important Coefficients

The process of inverse decomposition can be used to illustrate the concept of fields of influence and inverse important coefficients. Fields of influence refer to the degree and range of inverse coefficient change in response to a change in a single technological coefficient. Inverse important coefficients are those technological coefficients that are associated with the strongest fields of influence. They play the largest role in bringing about changes in output through their effect on the inverse. Knowing which coefficients are inverse important can be useful in identifying where technological improvements will have their greatest impact and in updating key coefficients to achieve the greatest level of accuracy when complete data for an input-output table in a new time period are unavailable.

6.3 Structural Decomposition: Problems in the Discrete Time Framework

The measurement of structural changes caused by inverse and final demand change is complicated in real-world applications where variations in technological coefficients and final demand occur simultaneously and in discrete rather than infinitesimal increments. In this environment, the change in total output is

$$\Delta X = \Delta(I - A)^{-1} Y + (I - A)^{-1} \Delta Y + \Delta(I - A)^{-1} \Delta Y.$$

The primary sources of output change caused by inverse and final demand change are still present, but there is an added effect. The final term to the right of the equality, $\Delta(I - A)^{-1} \Delta Y$, is an interaction term that involves simultaneous changes in both the inverse matrix and final demand vector. Its presence makes it difficult to isolate the effects of pure changes in the inverse and final demand.

In empirical applications, interaction terms are sometimes ignored, sometimes split in equal proportions and assigned to primary effects, and sometimes reported as separate effects along with the primary sources of change. While there is no generally agreed upon solution to the problem, one internally consistent means of separating interaction effects involves the use of weights related to the magnitudes of primary sources of change.

7. CONCLUSIONS

As this short overview shows, the input-output model provides a simple and convenient way to

portray transactions among the various sectors of the economy. At any point in time, the model provides a snapshot of the inputs required by specific industries and the requirements of all industries for specific goods and services. Because these transactions are considered simultaneously, through the calculation of inverse coefficients, the model makes it possible to show how production in one sector of the economy affects production in all others. This overview barely scratches the surface of the wide range of input-output applications. It is an especially important tool in regional analysis, where the mutual interdependence of multiregional economic activity is investigated. The model has been used to study the interrelationship between the economy and the environment, where, for example, pollution output from industrial activity becomes an input to various ecological processes. Aside from these applications, there is also continuing interest in improving its accuracy. Because input-output tables are published infrequently, research continues to focus on the stability of coefficients through time, as well as on finding ways to create accurate updates to tables when complete data for new time periods are unavailable. Similarly, since benchmark tables are not available for all regions, there is ongoing interest in developing accurate regionalized versions of the national tables.

SEE ALSO THE FOLLOWING ARTICLES

Complex Systems and Energy • Depletion and Valuation of Energy Resources • Economics of Energy Demand • Economics of Energy Supply • Energy Futures and Options • Goods and Services: Energy Costs • Modeling Energy Supply and Demand: A Comparison of Approaches • Multicriteria Analysis of Energy

Further Reading

Bullard, C., and Herendeen, R. (1975). The energy cost of goods and services. *Energy Policy* **1,** 268–277.

Carter, A. P. (1970). "Structural Change in the American Economy." Harvard University Press, Cambridge, MA.

Casler, S. D. (2001). Interaction terms and structural decomposition: An application to the defense cost of oil. *In* "Input-Output Analysis: Frontiers and Extensions" (M. Lahr and E. Dietzenbacher, Eds.), pp. 143–160. Palgrave, New York.

Hannon, B., Blazcek, T., Kennedy, D., and Illyes, R. (1984). A comparison of energy intensities: 1963, 1967, and 1972. *Resources and Energy* 5(1), 83–102.

Leontief, W. W. (1951). "The Structure of the American Economy." Oxford University Press, New York.

Miernyk, W. H. (1965). "The Elements of Input-Output Analysis." Random House, New York.

Miller, R. E., and Blair, P. D. (1985). "Input-Output Analysis: Foundations and Extensions." Prentice Hall, Englewood Cliffs, NJ.

Rose, A., and Casler, S. D. (1996). Input-output structural decomposition analysis: A critical appraisal. *Economic Systems Research* 8(1), 33–62.

Rose, A., and Chen, C. Y. (1991). Sources of change in energy use in the U.S. Economy: A structural decomposition analysis. *Resources and Energy* **13,** 1–21.

Rose, A., and Miernyk, W. (1989). Input-output analysis: The first fifty years. *Economic Systems Research* **1,** 229–271.

Sonis, M., and Hewings, G. J. D. (1989). Error and sensitivity input-ouput analysis: A new approach. *In* "Frontiers of Input-Output Analysis" (A. Rose, K. Polenske, and R. Miller, Eds.), pp. 244–262. Oxford University Press, New York.

U.S. Department of Commerce, Bureau of Economic Analysis (1997). Benchmark Input-Output Accounts for the U.S. Economy, 1992. *In* "The Survey of Current Business," December.

Integration of Motor Vehicle and Distributed Energy Systems

TIMOTHY E. LIPMAN
University of California, Berkeley and Davis
Berkeley and Davis, California, United States

1. **Introduction**
2. **Distributed Power Generation and Distributed Energy Resources: An Emerging Paradigm in Power Generation**
3. **Electric-Drive Vehicles: Battery, Hybrid, and Fuel Cell Vehicles to Improve Efficiency and Reduce Emissions**
4. **Vehicle-to-Grid (V2G) Power: Using Motor Vehicles to Complement Utility Grids**
5. **EVs for Remote and Backup Power**
6. **EVs for Power Generation and Ancillary Services**
7. **Hydrogen Energy Stations and Other Motor Vehicle and Distributed Generation Integration Schemes**
8. **Key Technical and Regulatory Issues for EVs Used as Distributed Generation**
9. **Conclusions**

Glossary

battery electric vehicle An electric-drive vehicle that derives the power for its drive motor(s) from a battery pack.

distributed energy resources Power generation, storage, and metering/control systems that allow power to be used and managed in a distributed and small-scale manner, thereby siting generation close to load in order to minimize electricity transmission and maximize waste heat utilization.

fuel cell vehicle An electric-drive vehicle that derives the power for its drive motor(s) from a fuel cell system; hybrid fuel cell vehicles would also derive drive motor power from a supplemental battery or ultracapacitor.

hybrid electric vehicle An electric-drive vehicle that derives part of its propulsion power from an internal-combustion engine and part of its propulsion power from an electric motor, or that uses an internal combustion engine to power a generator to charge a battery that in turn powers one or more electrical drive motors.

vehicle-to-grid (V2G) power Electrical power that is produced or absorbed by a motor vehicle to provide real power, reactive power, or electrical grid ancillary services to local electrical loads and/or utility grids.

A new generation of advanced, small-scale power technologies is rapidly emerging based on the concept of distributed generation (DG) and, more broadly, distributed energy resources (DER). DER systems have the potential to allow for decentralized electricity production with reduced needs for transmission and distribution of power, thereby increasing the efficiency of end-user energy use, particularly when coupled with the utilization of waste heat for local heating or cooling needs (known as combined heat and power [CHP] or "cogeneration"). At the same time, DER can help to reduce the need for siting large power plants and transmission lines. Furthermore, DER systems can also offer environmental benefits through installation of clean technologies such as photovoltaic (PV), wind, and fuel cell systems, and by replacing or displacing the construction of relatively dirty peaker power plants.

1. INTRODUCTION

The motor vehicle industry is undergoing dramatic change with regard to the vehicle technologies that are being introduced and those that are being researched and demonstrated. After years of experimentation with different types of alternative-fuel vehicles (AFVs), the automobile industry is now commercializing hybrid electric vehicles (HEVs) with hopes that fuel cell vehicles (FCVs), now in a precommercial prototype development phase, can be commercialized by the second decade of the 21st century. Toyota Motor Company has led the way in HEV development, based on the successful launch of

 © 2004 Elsevier Inc. All rights reserved.

the Prius hybrid vehicle in the United States and Japan and the hybrid Estima minivan in Japan. Toyota has now sold more than 100,000 HEVs globally, and Honda has also been successful with the hybrid Civic, introduced in 2002, after limited success with the small Insight hybrid.

These electric-drive vehicles (EVs)—including battery electric vehicles (BEVs) as well as HEVs and FCVs—could have a variety of interactions with the stationary power sector and, with distributed energy resources (DERs) in particular, in ways that could provide value and reduce the costs of vehicle ownership. The only conventional connection would be recharging BEVs as simple load for the electric utilities (or DER systems), or similarly recharging plug-in HEVs or refueling hydrogen FCVs with hydrogen made from water using electricity to operate an electrolyzer. Of these options, recharging BEVs is occurring and will continue, although these vehicles have met with limited market success where they have been introduced. It is unclear if plug-in HEVs will be commercialized, as they have the advantage of being able to operate in a zero-tailpipe emission mode but will tend to be more expensive than other types of HEVs. Finally, refueling FCVs with grid-power electrolysis is not very efficient or cost-effective, and will likely only be done in the early years of vehicle introduction until better hydrogen refueling arrangements for FCVs are made.

However, all of these types of EVs could interact with utility grids, especially DER systems, in far more complex and interesting ways than simply as consumers of electricity. First, vehicles equipped with electrical power generating or storage capacity have the ability to supply needed ancillary services to utility grids. These services include grid voltage support, frequency regulation, spinning reserves to provide adequate power reserve margins, volt-amperes reactive (VARs) for power factor correction, and emergency backup power, among others. Furthermore, with regard to the ability to store or produce power, these vehicle types can act as buffer systems for the utility grid, taking up power off-peak when it is inexpensive (as many BEVs already do) and then delivering it on-peak when it is often much more highly valued. This could prevent the need to construct additional peaker power plants that are only needed during times of peak demand and that tend to be less tightly controlled for emissions than constantly-operating baseload plants. One need only consider the fact that an electrified vehicle fleet in the United States would have approximately 14 times as much power generating capacity as all of the stationary power plants in the country to realize that there is an enormous amount of power embodied in our fleets of motor vehicles!

This article reviews the potential linkages and synergies between emerging distributed power systems and EVs, particularly with regard to the potential for vehicle-to-grid (V2G) power from various vehicle types. The chapter is organized as follows. First, DG and DER systems are briefly described, along with a discussion of their advantages relative to other options. Second, electric-drive vehicle introduction is also briefly summarized. Third, various applications in which DER systems and EVs can potentially interact are discussed. Finally, emerging concepts and needs for future research are summarized, along with key points from the previous sections.

2. DISTRIBUTED POWER GENERATION AND DISTRIBUTED ENERGY RESOURCES: AN EMERGING PARADIGM IN POWER GENERATION

While DG is generally considered to include only power generation technologies—such as fuel cells, microturbines, small generator sets, PV systems, and so on—DER also includes energy storage and demand-side technologies and systems (for example, smart metering and thermostats, used in conjunction with real-time electricity pricing). These definitions vary, as does the upper-end definition of DG size above which power generation is considered central and not distributed—this can be 1 megawatt (MW), 10 MW, or even 100 MW. Regardless of the exact definition, however, the possibility of DER adoption on a significant scale offers interesting possibilities for improving the environmental quality of energy production while at the same time potentially working hand-in-hand with electricity market changes such as real-time pricing of electricity (to which DER could respond by turning on and off or ramping up or down) to complement the needs of the utility grid.

Table I shows some of the emerging DER technologies that are being developed and deployed, along with a few important characteristics. As the costs of these technologies fall through better design and economies of scale, and enabled by constraints on powerplant and transmission line siting, they are

TABLE I

Example Types of DER Technologies and Systems

System type	Primary use	Likely fuel/energy input	Commercialization status
Absorption cooling	CHP[a]	Waste heat	Early commercial
Energy storage systems	Energy storage	Electrical energy	Commercial
Fuel cells	Power generation	Natural gas, landfill gas, biomass, wind, others	Early commercial
Industrial turbines	Power generation	Natural gas	Commercial
Microturbines	Power generation	Natural gas, landfill gas, biomass, others	Commercial
Photovoltaics (PV)	Power generation	Solar radiation	Commercial
Real-time pricing systems	Load curtailment	Minimal electrical energy	Demonstration
Reciprocating engines	Power generation	Natural gas, landfill gas, biomass, others	Commercial
Waste heat recovery	CHP	Waste heat	Commercial
Wind turbines	Power generation	Wind energy	Commercial

[a]CHP is combined heat and power.

expected to compete more and more favorably with the traditional option of central power generation.

With regard to the challenge of developing and deploying DER systems, it seems clear that (1) DER penetration into electricity/natural gas markets is not only revolutionary technically but also controversial politically due to its potential economic impacts and (2) in order to meet growing energy needs while considering the environmental consequences of power generation, growth in the use of renewable energy, CHP, and other very clean energy production will be very important to meeting air quality and climate stabilization goals. With regard to this second point, in 1998 the former U.S. secretary of energy developed a goal for the United States, which is still official Department of Energy policy, of doubling the contribution of CHP to the nation's power supply by 2010 (46 gigawatts [GW] in 1998 to 92 GW). Furthermore, in California, the state assembly approved a renewable portfolio standard (RPS) that required 20% of California's electricity to be renewably generated by 2017, and the New York legislature passed a 25% RPS standard, following on the success of other RPS standards in Texas and other states. Programs such as these help to advance the renewables-based advanced power-generation programs in these states and also provide a general boost for DER technologies and efforts to work on grid interconnection issues.

The overall vision for the future that some electric industry practitioners and observers share—of a network of DER, wireless, and other information technologies that become integrated into the electricity system to facilitate real-time pricing of electricity and a significant amount of local power generation for local use—will surely take many years to become a reality. However, various states in the United States and countries around the world are well on their way toward pursuing this vision by developing the underlying technologies and complete systems and working to also develop the needed utility interconnection standards, address market and rate design issues, and redesign administrative protocols.

These efforts can be contentious with many different stakeholder groups involved, and they are also intimately connected with other broader issues that are also often controversial such as the overall concept of utility deregulation. The various roles played by state legislatures, public utility commissions, the Federal Electricity Regulatory Commission, electrical grid operators such as the California Independent System Operator, governors' offices, and interest groups are complicated and overlapping to some extent, and even making clear which groups have authority in certain areas has been far from straightforward. Nevertheless, there is growing interest in DER systems among a variety of industrial, commercial, governmental, and nongovernmental entities, and continued development and deployment of these systems will occur as the regulatory and administrative issues are slowly resolved.

While the politics involved make it difficult to predict the timing or exact nature of the manner in which DER systems will ultimately be integrated into existing utility grids (greenfield-type developments being a somewhat different matter), it is clear that the prospects of DER adoption also open many interesting side possibilities. These include several

involving transportation sector linkages, such as (1) V2G power and grid ancillary services; (2) artful integration of DER in conjunction with such concepts as hydrogen energy stations (H_2E stations) that would co-produce electricity for local loads or the utility grid along with high-purity hydrogen to refuel fuel cell vehicles (FCVs), and (3) combination with information technologies such real-time pricing of electricity and intelligent transportation systems.

DER can also potentially interface with many other sectors, including the obvious potential benefits to end-user groups (commercial, industrial, ranches/farms, etc.), such as maritime, recreational, and military applications. Furthermore, technological innovation related to DER development for markets in developed nations may ultimately spill over into village power and remote/rural applications throughout the developing world, just as cell phone technology obviated the need for telephone cables and wires, thereby potentially affecting the lives of billions of people in the future.

3. ELECTRIC-DRIVE VEHICLES: BATTERY, HYBRID, AND FUEL CELL VEHICLES TO IMPROVE EFFICIENCY AND REDUCE EMISSIONS

Within the automotive industry, the development of electric-drive vehicles is about as revolutionary as the concept of DER relative to traditional power generation. After nearly 100 years of dominance by steel-bodied, gasoline-powered, internal combustion engine vehicles, we now are seeing introduction of HEVs and much research and development for even more advanced vehicles such as FCVs that would operate on hydrogen, or possibly methanol or other fuel. Toyota and Honda have already commercialized HEVs in the United States and Japan, and General Motors, Ford, and Nissan, among others are planning to follow in the next few years with HEV market introductions of their own.

BEVs, HEVs, and FCVs share many similarities, while being rather distinct in other respects. First, all of these vehicles are likely to incorporate battery systems for energy storage, but for different reasons. For BEVs, batteries are the sole energy storage device for the vehicle and provide all of the power to the electric drive motor (or motors). For HEVs, batteries power the electric motor (or motors), but additional power comes from the combustion engine. The main purpose of the batteries and motor is to allow the combustion engine to run better and to turn off when not needed, as well as to allow braking energy to be recaptured when the vehicle is decelerating as in BEVs (i.e., regenerative braking). For FCVs, regenerative braking is also a rationale for adding batteries, but additional considerations are the need for batteries to help start up the fuel cell system and the prospect of reducing costs by using battery power to reduce the size (and thus cost) of the fuel cell system.

Thus, in addition to battery and regenerative braking systems, these vehicle types all employ an electric motor to add propulsion power to the vehicle. For BEVs and FCVs, one or more electric motors provide all of the vehicle propulsion power. Most types of HEVs would also use a combustion engine to provide power to the wheels, but one type known as a series HEV would simply use the combustion engine to run a generator to help charge the battery and not use the engine to power the wheels.

The electric motors used in these vehicle types are of two different forms: (1) alternating current (AC) induction motors and (2) brushless permanent magnet (BPM) motors. To convert the direct current (DC) produced by the battery or fuel cell system to the proper current (and voltage) needed to operate these types of electric motors, a power inverter is required. These power inverters use insulated gate bipolar transistors (IGBTs) as the high-power switching devices to convert DC power to AC (or a similar synchronous output to operate BPM motors). Since these types of electric vehicles will all necessarily include a power inverter device that is capable of producing a high-quality AC waveform and power of at least 10 kilowatts (kW) and up to 100 kW (for typical BEV, HEV, and FCV designs), it is then a relatively minor matter to design in an interface to allow these vehicles to interact with utility grids so that they can both absorb and discharge real power and VARs, as well as acting as spinning reserves and backup power units.

One rationale within the industry for developing electric vehicles (particularly HEVs), other than the potential environmental and social benefits that they can provide, is that requirements for onboard power in motor vehicles are growing and beginning to stretch the limits of traditional 12-volt automotive electrical power systems. New motor vehicles are requiring this additional power to operate new types of accessories, such as vehicle navigation and auto

alert safety systems; onboard televisions and DVD players, electrically actuated and controlled engine, exhaust, and auxiliary systems; and additional electric motors to control seats and window blinds. Other electrical gadgetry soon to be incorporated into some vehicle models includes electronic tire pressure monitoring/control, bumper-mounted side-view and rear-view cameras, and 120-volt AC accessory outlets. To meet these growing needs for onboard electrical power, a new 42-volt standard is being incorporated into future automobile designs, and this can easily be accommodated by BEVs, HEVs, and FCVs by using a DC-to-DC converter to step the full system voltage (typically 150 to 500 volts) down to the 42-volt level needed for the auxiliary systems.

4. VEHICLE-TO-GRID (V2G) POWER: USING MOTOR VEHICLES TO COMPLEMENT UTILITY GRIDS

As noted previously, EVs of all types can interact with utility grids in much more complex ways than simply as load for electric utilities. With the partial deregulation of the utility industry in several U.S. states, opportunities have arisen for the various sections of the electricity supply value chain to be supplied by different groups in competition with one another (for generation, operation of the grid, billing and customer services, etc.) rather than by a single vertically integrated utility as has historically been the case in most places. Within this context, of a competitive market for a range of services related to the provision of electricity, EVs can play interesting roles, either similar to or actually in conjunction with those of DG and DER systems.

There are several potential possibilities for using EVs as distributed generating resources. Various types of EVs could be used to do the following:

- Produce power to meet the demands of local loads
- Provide additional power to the grid in a net-metered or electricity buy-back scenario, helping to meet demands in times of capacity constraint
- Provide emergency backup power to residences, offices, hospitals, and municipal facilities
- Provide peak shaving for commercial sites, simultaneously reducing electricity energy charges ($/kWh) and demand charges ($/kW-peak per month)
- Provide ancillary services to the grid, such as spinning reserves, voltage and power quality support, and VARs
- Provide buffering and additional power for grid-independent systems that rely on intermittent renewable energy systems, or for microgrids that can benefit from additional generation or storage capacity

Although only some of these possibilities have been explored, research has shown that EVs can be used to provide a range of important services to utility electrical grids. In fact, the potential for producing electrical power from vehicles is enormous—the generating capacity of an electrified U.S. motor vehicle fleet would be approximately 14 times the entire capacity of all of the stationary power plants in the country. For example, in California a fleet of 100,000 FCVs could produce about 1.5 GW of power for the grid, assuming 30 kW net fuel cell output power per vehicle and 50% vehicle availability. This 1.5 GW of power represents approximately 5% of California's peak electrical power demand on a typical day. The following sections describe various potential uses of electric vehicles as DER systems, as well as other schemes for integrating motor vehicles into these small-scale energy systems.

5. EVs FOR REMOTE AND BACKUP POWER

One obvious use for EVs in a power generation role is the notion of using the vehicles to produce power for remote uses (such as construction sites, recreational uses such as camping and fishing, etc.). This is similar to the idea of using EVs for backup power in the event of a power loss to a residential or commercial building, except that under the U.S. national electrical code the latter is not permissible (due to concerns about line-worker safety) while the former is allowed. However, with adequate safeguards it should soon be possible to allow EVs and other grid-tied DG systems to be used in an islanded mode with adequate safeguards to provide emergency backup power, with minor changes in existing codes and regulations.

One prototype vehicle has already been built with this remote/backup power capability in mind. The Dodge contractor special is a HEV with the capability of producing approximately 2 kW of electrical power for sustained periods of time.

In fact, an emerging standard that may eventually be incorporated into nearly all future vehicles is based on an electrical system for vehicle auxiliary systems that operates at 42 volts versus the 14-volt standard for conventional systems. This standard has become necessary due to increasing desire to equip vehicles with electronic systems and the need to provide more power for these systems without increasing wire diameter and raising costs. New or planned electronic gadgetry on vehicles include navigation systems, extensive onboard communications, voice-actuated controls, exterior AC power supplies, computer controlled power-assisted active suspension, collision-avoidance systems, electric air conditioner compressors, drive-by-wire steering, side- and rear-view bumper cameras, electronic tire pressure control, and generally greater computer power for increasing control of various vehicle systems. The power demands from these new electrical systems are expected to push the power demands onboard vehicles from several hundred watts to a few thousand or perhaps even several thousand watts in the future.

At the same time that the 42-volt standard can support the addition of electrical devices in conventional vehicles, the increased power levels available from these systems in even conventional combustion-engine vehicles could enable the provision of a significant level of emergency or backup power—perhaps on the order of 3 to 5 kW—although local regulations may prevent backup power from being provided with the vehicle engine on. The duration of the power supply capability would then be limited to a few hours, by the capacity of the 42-volt battery system onboard the vehicle.

6. EVs FOR POWER GENERATION AND ANCILLARY SERVICES

With regard to using EVs to generate power and support utility grids, research has shown that EVs can be used to provide a range of important services in this regard. When not in use, BEVs could be used for emergency backup power or to buffer the utility grid by charging off-peak, when electricity is plentiful, and supplying it back during times of peak demand and capacity constraint. HEVs and FCVs could act as generators, producing electricity from a liquid or gaseous fuel to meet the demands of connected local loads or other nearby power needs. EVs could also provide certain types of support services to utility grids, such as spinning reserves and grid frequency regulation. While contracting for these services would likely require a service aggregator to bid the reserve or regulation function from EVs to local system operators, BEVs in particular appear to be well suited to these uses. For example, Alec Brooks of AC Propulsion has estimated that the frequency regulation of the California grid (a service that costs the California Independent System Operator millions of dollars per year) could be provided by 50,000 to 100,000 EVs, and these vehicles could provide this service as they were recharging their battery packs. AC Propulsion has demonstrated this capability with one of their prototype vehicles, and Fig. 1 shows the scheme by which they envision EVs receiving, by radio signal, the control error in the frequency of the utility grid and then responding as needed.

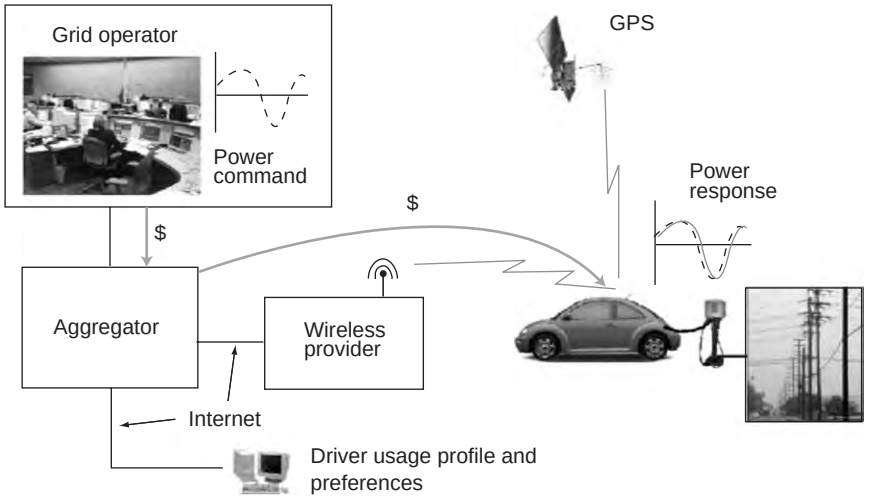


FIGURE 1 One concept of V2G Operation. BEVs acting as grid frequency regulation service providers. Data from Brooks (2002).

As noted earlier, the gross power generating capability of vehicles is enormous, and vehicles with significant onboard electrical generating capability allow that power to be generated efficiently for various uses, many of which have economic value that can translate into reduced costs of ownership for these new vehicle types. The most comprehensive published analysis of the potential for EVs and FCVs to act as DG resources was conducted by Kempton *et al*. This report was prepared for the California Air Resources Board by researchers at the University of Delaware, Green Mountain College, and the University of California. The analysis also involved collaboration with EV drivetrain manufacturer AC Propulsion, whose Generation 2 electric motor controller unit allows for bidirectional grid interface that can support V2G connections. The Kempton *et al*. effort examined the economic potential of using various types of EVs to produce power for buildings and the grid, as well as to provide grid ancillary services such as spinning reserves, nonspinning reserves, and grid frequency regulation. Table II presents some of the key findings of the Kempton *et al*. report, in terms of the range of annual values that might be expected for different EV types and for three different power-generation or ancillary services.

The particular case of using FCVs to produce power has drawn particular attention due to the potential for FCVs to become commercialized as mass replacements for conventional vehicles during the 2008 to 2012 timeframe. The significant power generation capability of these vehicles—from perhaps 20 kW for small hybrid FCVs (with additional battery power available) to 100 kW or more from larger FCVs with minimal battery or ultracapacitor— suggests that FCVs have the potential to become significant power generation resources in the future.

In the study discussed earlier, Kempton *et al*. concluded that FCVs could compete in the peak power market, with generating costs on the order of $0.18/kWh, but they could not compete with base-load power. Also, an earlier analysis of the potential for FCVs to provide power for buildings was conducted by Kissock. This analysis assumed that FCVs equipped with 25 kW nominal (37.5 kW peak) proton exchange membrane (PEM) fuel cell systems would be used to produce electricity, or electricity and heating/cooling in a cogeneration mode, for New Jersey and Texas residences, for a New Jersey hospital, and for a Texas office building. The analysis concluded that annual savings of up to $2500 per FCV docking station could be realized with the residential setting (with some cases showing negligible benefits), that annual savings of $1200 to $8800 were possible for each docking station at the hospital, and that annual savings of $2300 to $2900 were possible for the office building cases.

In a somewhat more detailed analysis, Lipman *et al*. have shown that while the costs of producing power from stationary and motor vehicle fuel cell systems are rather different in terms of the costs involved—for stationary systems much of the cost is capital cost, while for FCVs much of the cost is for periodically refurbishing the vehicle's fuel cell stack—the economics of producing power from these systems can be comparable, and potentially competitive with grid power in the future, if the systems are operated with similar overall efficiencies. However, results are highly sensitive to the compounding effects of variations in several key input variables such as natural gas and electricity prices, and fuel cell system costs, durability

TABLE II

Vehicle Owner's Annual Net Profit from V2G

	Peak power	Spinning reserves	Regulation services
Battery, full function	$267	$720	$3162
	(510–243)	(775–55)	(4479–1317)
Battery, city car	$75	$311	$2573
	(230–155)	(349–38)	(4479–1906)
Fuel cell, on-board H_2	$-50 (loss) to $1226	$2430 to $2685	$-2984 (loss) to $811
	(2200–974 to 2250)	(3342–657 to 912)	(2567–1756 to 5551)
Hybrid, gasoline	$322	$1581	$-759 (loss)
	(1500–1178)	(2279–698)	(2567–3326)

The figures represent $net and (revenue-cost). These are representative midrange figures extracted from full analysis in the report.
Source. Kempton *et al*. (2001).

levels, and fuel cell stack refurbishment costs. Where FCVs are used to produce power to meet very low load levels, such as at individual residences (e.g., and operate at very high turndown ratios), their efficiencies suffer and they become much less economically attractive at producing power. This suggests that net-metering policies—which would allow FCVs to be operated at higher power and efficiency levels and for excess power to be supplied to the grid for a credit—may be critical to improving the economics of FCVs used to produce power in this way.

Figures 2 and 3 present results from the Lipman *et al.* analysis, comparing the costs of producing power from FCVs and stationary PEM fuel cells at both California office building and residential locations, and with three different sets of economic assumptions. These results show that the costs of power production at residences from FCVs can be highly variable, but that the economics of producing power from FCVs at commercial locations are better due to lower commercial natural gas prices, better fuel cell system efficiencies in meeting higher loads, excess FCV power availability during the day when it is highly valued, and other factors.

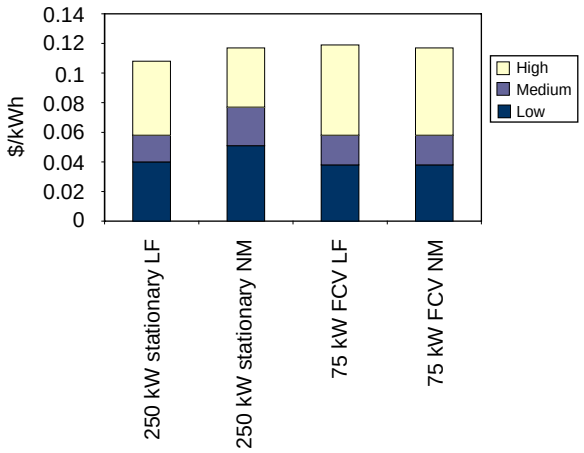


FIGURE 2 FCVs used as DG versus stationary PEM fuel cells at medium California office building setting. LF, load following; NM, net metering. Data from Lipman *et al.* (2002b).

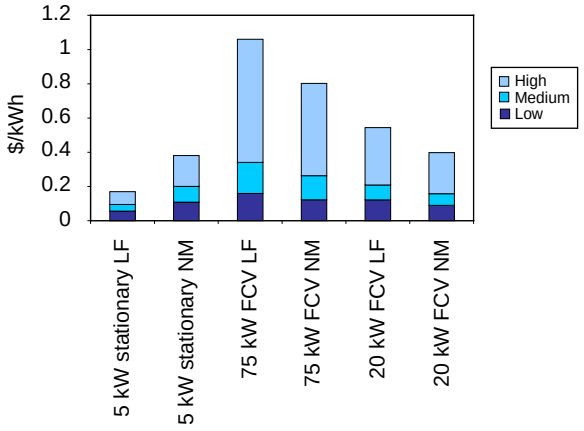


FIGURE 3 FCVs used as DG versus stationary PEM fuel cells at California residential setting. LF, load following; NM, net metering. Data from Lipman *et al.* (2002b).

7. HYDROGEN ENERGY STATIONS AND OTHER MOTOR VEHICLE AND DISTRIBUTED GENERATION INTEGRATION SCHEMES

In addition to using various types of vehicles with electricity-generation capability for standalone, backup, and power/ancillary services provision, there also are several other ways in which vehicles can interact with distributed energy systems. These includes opportunities for fueling motor vehicles in ways that are complementary to the development of distributed energy systems, and additional opportunities for EVs to act in conjunction with emerging schemes and plans for electrical power generation.

One interesting option relates to the need for developing hydrogen-refueling infrastructure for early FCVs and is the concept of the hydrogen energy station (or H_2E station). These H_2E stations would be either dedicated refueling facilities, or a key component of the energy production, use, and management portion of a commercial or industrial facility. The energy station component would consist of a natural gas reformer or other hydrogen generating appliance, a stationary fuel cell integrated into the building with the potential capability for combined heat and power (CHP) production, and a hydrogen compression, storage, and dispensing facility. In essence, H_2E stations would seek to capture synergies between producing hydrogen for a stationary fuel cell electricity generator that provides part or all of the power for the local building load (as well as the capability to supply excess electricity to the grid) and refueling FCVs with additional high-purity hydrogen that is produced through the same hydrogen-generation system.

In principle, many different H_2E station concepts and designs are possible, including the following:

- Service station–type designs that are primarily intended to produce hydrogen for FCV refueling

but that also supply electricity to the service station
- Office building–based designs that primarily provide electricity and waste heat to the building but also include a small off-shoot for FCV refueling
- Distributed generation facilities that are primarily intended to supply excess electricity to the power grid but that also include some provision for FCV refueling

In addition, FCVs parked near the H_2E station for any sizable length of time could in principle supply electricity to the building or grid, since they would have access to a fuel supply.

Analysis of the economics of several potential H_2E station designs shows that for the California setting examined, where prevailing commercial and residential electricity rates are high, H_2E stations appear to be relatively attractive ways of supplying small numbers of FCVs with hydrogen. In general, H_2E stations tend to be more attractive economically than dedicated hydrogen refueling stations, especially for low numbers of vehicles supported per day, but the economics depend importantly on several variables including natural gas and electricity prices, captial equipment costs, the hydrogen sales price, and fuel cell maintenance and stack refurbishment costs. Figure 4 shows that for this set of the service station cases examined the H_2E station designs are more attractive economically than dedicated hydrogen refueling stations. None of the dedicated or $H_2$2E stations are economically profitable with so few vehicles supported, but the H_2E stations offer a lower cost pathway to ultimately supporting greater numbers of vehicles profitably.

Figure 5 shows how after a transitional period with low numbers of vehicles supported, H_2E stations could begin to turn a profit with greater numbers of vehicles refueled per day and with reasonable hydrogen selling prices of around $20 per GJ. This figure shows that a 10% simple ROI target can be met with the service station type of H_2E station, but only with relatively high hydrogen sales prices of about $20 per GJ and only with about 50 or more vehicles per day refueled. At lower hydrogen sales prices of $10–15/GJ, the economics of this type of station do not look attractive, even with significant numbers of vehicles refueled. However, with lower natural gas prices than the $6/GJ assumed in this future high cost case, this picture would change somewhat, with hydrogen sales prices of $15/GJ potentially becoming profitable for large numbers of vehicles per day supported.

Additional options for integrating motor vehicles with distributed energy systems include incorporating EVs into designs for microgrids. Microgrids are small clusters of DER technologies that act together to supply power and heating/cooling to one or more adjacent buildings and that would connect with the

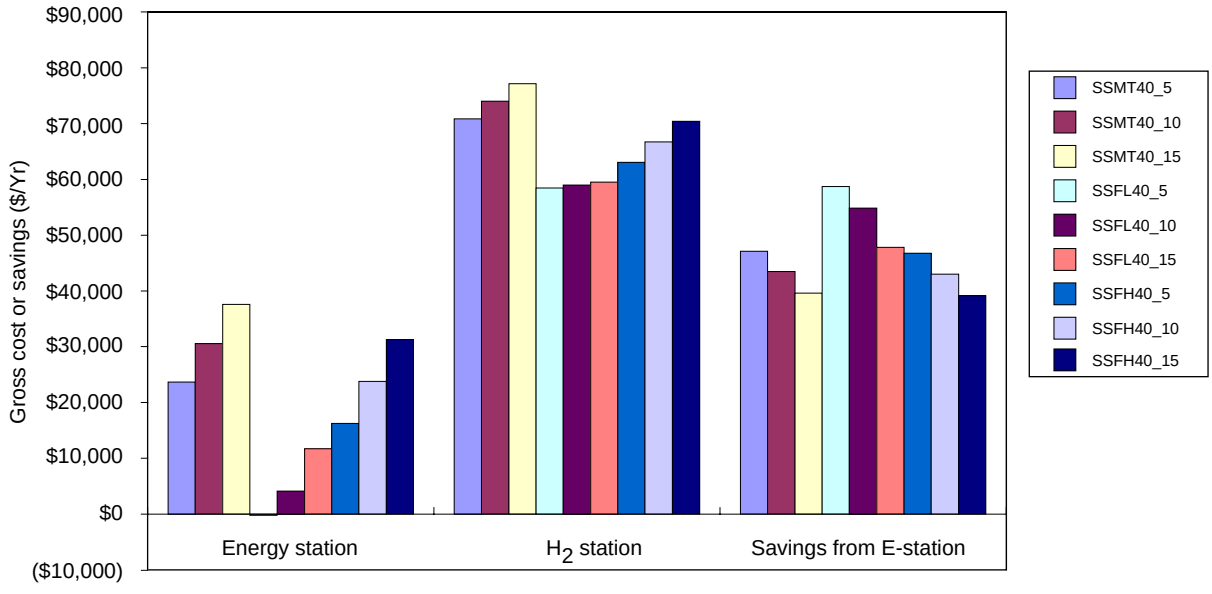


FIGURE 4 Estimated future annual costs of H_2E stations with 25 and 40 kW fuel cell and 5 to 15 FCVs refueled per day, compared with costs of dedicated H_2 stations (H_2 price of $10/GJ). FL, future low cost case; FH, future high cost case; MT, medium-term case; SS, service station; X_Y, fuel cell size in kW_# of FCVs per day refueled. Data from Lipman *et al.* (2002a).

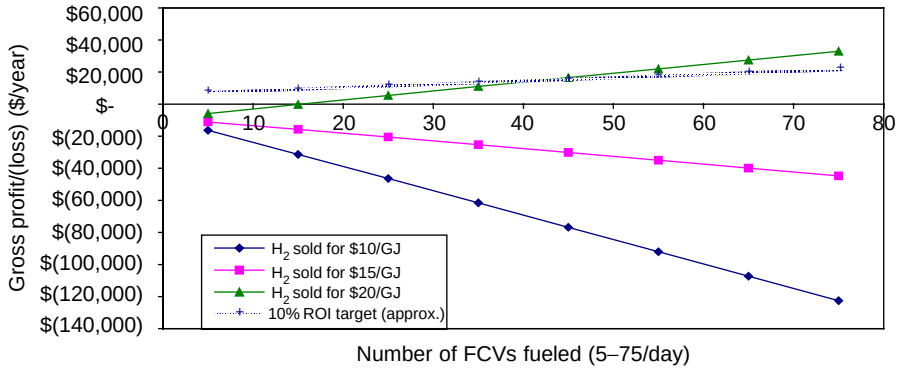


FIGURE 5 Estimated profit/loss from H_2E-service station with 40 kW fuel cell, 5 to 75 FCVs refueled per day, and future high costs (w/approx. 10% ROI target). Data from Lipman *et al.* (2002a).

main grid at one interface point. Microgrids could act as model citizens for the grid by supporting it in various ways, and they could benefit themselves from supplying local generation with grid power at various times (depending on electricity rate schedules and other variables). EVs could play interesting roles in microgrids by supplying electrical storage or power-generation capability, with the availability of the EV resources coinciding with the arrival of people at various commercial and industrial microgrid locations and their associated energy needs.

Similarly, EVs could help to provide additional power and buffering for small standalone renewable energy systems based on PV and wind power, along with a battery storage system. Small renewable systems that are off-grid because of their remote location are becoming increasingly common, but often are of relatively low power to keep costs low. Again, these systems could be bolstered by the use of EVs to provide additional power to operate high-power equipment for short periods of time, thereby allowing the size of the remote renewables system to be minimized.

Finally, another interesting opportunity for combining motor vehicles and distributed energy systems involves efforts to reduce emissions from heavy-duty diesel trucks. Approximately 30% of the fuel use and emissions associated with long-haul diesel trucks is due to the idling that these trucks do when they are stopped, primarily to support electronic devices in the truck cabin (refrigerator, microwave, television, climate control, etc.) by operating a generator with the main diesel engine. This is highly inefficient and polluting, and efforts are under way to examine opportunities for reducing these emissions. One such option would be to equip trucks with fuel cell auxiliary power units (APUs), running on pure hydrogen or possibly diesel fuel that is converted into a hydrogen-rich gas stream with an onboard fuel reformer. If operated on hydrogen, these heavy-duty truck APU systems could be refueled with hydrogen produced at truck stop hydrogen stations, possibly of a type of H_2E station design as discussed earlier, where hydrogen is used both for truck APU and FCV refueling, as well as to produce electrical power and heat for the truck stop using a stationary fuel cell. An alternative to using fuel cells as APUs onboard the trucks would simply be to have the trucks plug in to electrical outlets when near truck stops, and the electricity could be supplied either by the main utility grid or by a DG system at the truck stop. This would only reduce idling near truck stops, but would perhaps be easier to implement that the fuel cell option. Also note that each of these types of installations would be good candidate for integration of renewable PV or (if remote sites) wind power to further reduce environmental impacts.

8. KEY TECHNICAL AND REGULATORY ISSUES FOR EVs USED AS DISTRIBUTED GENERATION

Despite the potential benefits of using EVs to produce power and other grid services when the vehicles are parked, there are several key issues confronting the use of EVs in this manner. Some of these issues, and some potential solutions, are as follows.

First, with regard to grid interconnection issues for electric vehicles, reverse flow of electricity from EVs in event of loss of power in the grid (e.g., for emergency backup applications) is not permitted in

the United States under the national electrical code. Vehicles could in principle interconnect with utility grids in a manner such that they were supplying net power, assuming they were to meet relevant DG interconnection standards, but this would only be allowed by a few states under existing net metering laws. Also, from a technical perspective, reverse flow of electricity into local distribution systems eventually will reach a limit due to the generally unidirectional character of the existing high-voltage transmission system. Also, utility on/off control of grid-connected EVs may be essential for lineworker safety, requiring a complex control system. Solutions to these issues include revising the national electrical code to allow reverse-flow from vehicles for backup generation, with appropriate safeguards. Public utility commissions and other utility industry groups should consider analysis of retrofitting substations to ensure that transformer tap changers and line-drop compensators are compatible with reverse flow into high voltage transmissions systems, as well as investigating utility-to-vehicle wireless communication technology to provide emergency shut-down, real-time load, and electricity price information for utility control centers and vehicle owners.

Second, with regard to fuel cell operation when using FCVs to generate power, nonhybrid fuel cell systems with 75 to 100 kW peak power (a typical power level for full-sized vehicles) will likely be power limited to 30 or 40 kW for continuous operation while the vehicle is at a standstill due to thermal management issues. Fuel cell systems connected to loads in grid-independent operation (e.g., where they cannot lean on the grid) may be subject to transient demands that would require a hydrogen fuel supply buffer or battery support system for adequate performance. Also, operating fuel cell systems in residential settings and meeting local loads only, which are often as low as 1 kW to 2 kW and rarely exceed 4 kW to 5 kW, will likely produce low fuel cell system operating efficiencies, particularly for stacks designed to operate at high pressure. Potential solutions to these problems are to operate fuel cell systems at 4 kW to 30 kW and sell excess power to grid through net metering or other buyback scenarios. Alternately, industry may develop off-board auxiliary blower air supply systems for fuel cell stacks that allow low pressure, higher efficiency operation at low load levels, or power-on-demand systems where only a part of the vehicle's fuel cell stack is used in a distributed power mode, similar to displacement-on-demand systems that are beginning to be incorporated into light-duty trucks.

Third, with regard to fuel supply for FCVs, vehicles designed to operate on pure hydrogen may not be able to use reformate (a mix of hydrogen and other gases produced from a liquid or gaseous hydrocarbon fuel) due to carbon monoxide and sulfur catalyst poisoning issues. To operate for longer than a few hours by using some of the hydrogen in the vehicle's tank, vehicles will require some sort of hydrogen production support system based on steam methane reforming, electrolysis, partial oxidation reforming, or auto thermal reforming. These issues can be addressed by potentially employing vehicle fuel cell stacks with platinum-ruthenium catalysts that can operate on either neat hydrogen or reformate (e.g., the Ballard Mark 900), or by producing purified hydrogen at small scales. In this regard, there are important research and development requirements for developing low-cost and small-scale hydrogen reformers, and particularly interesting would be the development of multifuel reformers for vehicles that can operate on pipeline natural gas as well as the vehicle's primary fuel, thus eliminating the need for an off-board fuel reformer for V2G power from FCVs.

9. CONCLUSIONS

The emergence of DER solutions for distributed power production and the commercialization of various type of EVs in the motor vehicle industry are in some ways similar, with both representing potential paradigm shifts in their respective sectors. As these evolutionary and perhaps ultimately revolutionary changes occur, whether slowly or quickly, a range of interesting opportunities will emerge for interaction between the two sectors and synergistic development of infrastructure and services. Some of the possibilities for integration of these types of systems include the following:

- Using various types of EVs for emergency backup or remote power
- Using various types of EVs for peak power or grid ancillary services such as grid frequency regulation and spinning reserves
- Developing H_2E stations that would coproduce electricity/heat/cooling for local loads and the grid as well as pure hydrogen to refuel FCVs
- Combining efforts to reduce diesel truck idling emissions with distributed energy systems for truck stops

Among these options and the various possibilities within them, particularly promising appear to be the

prospects of using HEVs and FCVs for utility grid spinning reserves, FCVs for electricity peak shaving at commercial buildings, BEVs for utility grid frequency regulation, H_2E stations as a potential way to reduce the transition costs of developing hydrogen infrastructure for FCVs, and opportunities for reducing diesel truck idling through truck stop electrification or hydrogen refueling for fuel cell APUs. Initial investigations have suggested that all of these ideas are promising, but additional details such as how vehicles would bid these services into power markets and be compensated for the services they provide have yet to be fully considered. However, advances in information technologies such as wireless communications and real-time metering suggest that these logistical issues associated with V2G are unlikely to be major obstacles.

In conclusion, the tremendous amount of power available in the motor vehicle fleet and the prospects for electrifying the fleet in the future suggest a potentially large power resource that could be tapped to assist in the provision of future electrical power and electricity market services. With some careful planning and engineering, and in the context of the overall market development of distributed energy systems, the ideas discussed here may be just the beginning of the possibilities for integrating the motor vehicle and stationary power systems in creative and mutually beneficial ways.

SEE ALSO THE FOLLOWING ARTICLES

Alternative Transportation Fuels: Contemporary Case Studies • Batteries, Transportation Applications • Biodiesel Fuels • Ethanol Fuel • Fuel Cell Vehicles • Fuel Cycle Analysis of Conventional and Alternative Fuel Vehicles • Fuel Economy Initiatives: International Comparisons • Hybrid Electric Vehicles • Hydrogen, End Uses and Economics • Internal Combustion Engine Vehicles • Transportation Fuel Alternatives for Highway Vehicles • Vehicles and Their Powerplants: Energy Use and Efficiency

Further Reading

Brodrick, C-J., Lipman, T. E., Farschi, M., Lutsey, N., Dwyer, H. A., Sperling, D., Gouse, S. W., Harris, D. B., and King, F. G. (2002). Evaluation of fuel cell auxiliary power units for heavy-duty diesel trucks. *Transportation Research-D* **7**(4), 303–315.

Brooks, A. N. (2001). "Electric Drive Vehicles: A Huge New Distributed Energy Resource." EVAA Electric Transportation Industry Conference, Sacramento, CA.

Brooks, A. N. (2002). "Vehicle-to-Grid Demonstration Project: Grid Regulation Ancillary Service with a Battery Electric Vehicle." Prepared for the California Air Resources Board, Contract number 01–313, December 10.

Kempton, W., and Letendre, S. (1997). Electric vehicles as a new power source for electric utilities. *Transportation Res.-D.* **2**(3), 157–175.

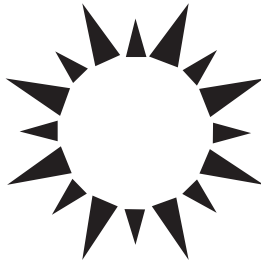
Kempton, W., Tomic, J., Letendre, S., Brooks, A., and Lipman, T. (2001). "Vehicle-to-Grid Power: Battery, Hybrid, and Fuel Cell Vehicles as Resources for Distributed Electric Power in California," Inst. of Transportation Studies, Davis, UCD-ITS-RR-01–03, Prepared for California Air Resources Board (under contract #ARB00–612) and Los Angeles Department of Water and Power, June.

Kissock, J. K. (1998). "Combined Heat and Power for Buildings Using Fuel-Cell Cars." ASME International Solar Energy Conference, Albuquerque, NM.

Lipman, T. E., Edwards, J. L., and Kammen, D. M. (2002a). "Economic Analysis of Hydrogen Energy Station Concepts: Are 'H2E-Stations' a Key Link to a Hydrogen Fuel Cell Vehicle Infrastructure?" Energy Development and Technology Working Paper Series, EDT-003, University of California Energy Institute (UCEI), November.

Lipman, T. E., Edwards, J. L., and Kammen, D. M. (2002b). "Economic Implications of Net Metering for Stationary and Motor Vehicle Fuel Cell Systems in California." Program on Workable Energy Regulation (POWER) Paper Series, PWP-092, University of California Energy Institute (UCEI), February.

Intelligent Transportation Systems



SUSAN A. SHAHEEN
University of California, Berkeley and Davis
Berkeley and Davis, California, United States

RACHEL FINSON
University of California, Berkeley
Berkeley, California, United States

1. **Introduction**
2. **Definition of ITS**
3. **ITS and Energy Impact Evaluation Tools**
4. **Predicted and Early Findings**
5. **System Integration and Conclusion**

Glossary

bus rapid transit (BRT) Seeks to improve bus service by reducing travel time and providing enhanced rider information; exclusive rights-of-way, bus lanes, adjusting stop locations, wider doors, preboarding payment, and supportive land use policies contribute to BRT improvements.

commercial vehicle operations (CVO) The application of electronic and wireless intelligent transportation systems technologies to address a range of trucking industry concerns; CVO approaches include border crossing clearance and safety via electronic clearance and manifesting, automatic vehicle location, vehicle-to-fleet management center communications, on-board safety diagnostics to enable more effective roadside safety inspections, and hazardous materials incident response.

electronic toll collection (ETC) Enables the instant payment of highway tolls when a vehicle passes through a toll station via an electronic roadside antenna (or reader) and a pocket-sized tag containing a radio transponder (typically placed inside a vehicle's windshield); the toll tag transponder transmits radio frequencies to the toll reader, and the appropriate fare is deducted automatically.

incident management Consists of three key areas: traffic surveillance (incident detection and verification), clearance, and traveler information; also covered by this area are emergency management services, which coordinate local and regional incident response to traffic accidents, security threats, and hazardous material spills.

intelligent transportation systems (ITS) A wide range of technologies—including electronics, information processing, wireless communications, and controls—aimed at improving safety, efficiency, and convenience of the overall surface transportation network.

intelligent transportation systems (ITS) user services User services that categorize a wide range of ITS technologies from the user's perspective. User service areas currently address high-level problems and needs such as route guidance, electronic payment, and parking services.

ramp metering Traffic signals employed at freeway on-ramps to control the rate of vehicles entering the freeway; metering rates are set to optimize freeway flow and to minimize congestion.

traffic management and surveillance Includes incident management, ramp metering, traffic signal control, traveler information, and traffic surveillance; traffic surveillance tools provide the data needed to manage the roadways.

traffic signal control Can integrate freeway and surface street systems to improve traffic flow and vehicular and nonmotorized traveler safety and also provide priority services for transit or high-occupancy vehicles; traffic control technologies include traffic surveillance, ramp metering, lane control systems, and traffic signals.

transit management Consists of four key areas: (1) transit vehicle tracking, (2) fare payment, (3) traveler information, and (4) personalized public transportation alternatives; transit vehicle tracking includes communication between vehicles and transit centers.

transportation management center (TMC) The hub where transportation, operations, and control data are collected, combined, and distributed to manage the transportation network (including emergencies and incidents) and to generate traveler information; the TMC relies on various intelligent transportation systems tools to collect its data, including electronic toll collection, radar, closed circuit video equipment, and loop detectors.

traveler information Provides the traveling public with information regarding available modes, optimal routes,

 © 2004 Elsevier Inc. All rights reserved.

and costs in real time either pre-trip or en route via in-vehicle information and changeable message signs along roadsides or at transit stations.

vehicle control technologies Intelligent transportation systems technologies that can help to avoid collisions, prevent or lessen injuries when crashes do occur, and ultimately lead to full vehicle automation; some existing vehicle control technologies include adaptive cruise control, antilock brakes, and electronic system malfunction indicators.

Although world energy consumption has increased only marginally since 1998 ($\sim$0.5%), greater efficiencies are still required in transportation. Indeed, the Texas Transportation Institute estimated total congestion costs (delay plus wasted fuel) at $67.5 billion for 75 U.S. urban areas in the year 2000. This represents 3.6 billion in delay hours and 5.7 billion gallons of excess fuel consumed due to congestion-related delays. The goal of intelligent transportation systems (ITS) is to improve surface transportation safety, efficiency, and convenience.

1. INTRODUCTION

Intelligent transportation systems (ITS) technologies include state-of-the-art wireless, electronic, and automated technologies. Collectively, these technologies have the potential to integrate vehicles (transit, trucks, and personal vehicles), system users, and infrastructure (roads and transit). Automated and in-vehicle technologies include precision docking for buses, automated guideways, and collision avoidance systems. Many ITS technologies can help to optimize trips (route guidance), diminish unnecessary miles traveled, increase other mode use, reduce time spent in congestion, reduce dependence on foreign oil, and improve air quality. Furthermore, when ITS technologies are applied to system management (transit and highways) and vehicle design, they can reduce fuel consumption by

- facilitating optimal route planning and timing;
- smoothing accelerations/decelerations and stop-and-go driving;
- reducing congestion;
- enabling pricing and demand management strategies;
- increasing the attractiveness of public transportation mode use;
- adjusting vehicle transmission for varying road conditions and terrain;
- facilitating small platoons of closely spaced vehicles (i.e., safer vehicles could enable weight reduction without compromising occupant safety).

Although ITS technologies are still in the early phase of deployment, many have shown the potential to reduce energy use. During the past 10 years or so, fuel consumption impacts of the following ITS technologies have been studied: (1) traffic signal control, (2) traffic management and surveillance (e.g., ramp metering), (3) incident management, (4) electronic toll collection (ETC), (5) traveler information, (6) transit management, (7) commercial vehicle operations (CVO), and (8) vehicle control technologies. They are the focus of this article. Nevertheless, ITS impacts, including benefits, unintended consequences, and aggregate effects, are still not well understood. The field of intelligent transportation systems, energy consumption impacts, current measurement tools, and early findings are discussed in this article.

2. DEFINITION OF ITS

In 1991, the concept of ITS emerged when transportation professionals recognized that electronic technologies could begin to play a significant role in optimizing surface transportation, and the U.S. Congress legislated the national ITS program. Since then, computer, communication, and sensor technologies have improved dramatically, and ITS technologies have emerged in highway and transit jurisdictions worldwide.

ITS deployment can been categorized into three stages:

- Stage 1: Test and implement early ITS technologies (or building blocks).
- Stage 2: Link early ITS technologies.
- Stage 3: Develop an integrated system of ITS technologies.

The public sector has been the dominant driver of stage 1 ITS technology research and development. During this phase, early ITS technologies were applied to improve traditional operations. Key accomplishments include the deployment of

- traffic management centers in urban areas to monitor freeway traffic and early incident notification;
- traffic signal control and ramp metering to improve traffic flow and safety;
- improved traveler information;

- commercial vehicle screening and ETC;
- satellite-based dispatching systems in transit operations;
- ride-matching services via the Internet;
- in-vehicle navigation systems in private vehicles.

Whereas stage 1 of ITS will continue for years, stage 2—linking early ITS technologies—is also progressing, as is evident by efforts to mainstream ITS with conventional capital improvement projects (e.g., connecting traffic management centers with advanced traffic signal coordination, ramp metering adjusted in real time, traveler information systems coordinated with in-vehicle devices). During this phase, industry has already initiated work to understand the fundamental nature of driving and of driver/operator behavior in transportation models. Furthermore, the process has started to move from a heavy focus on stage 1 public sector investment toward customer-oriented operations. Examples of next steps—included in stage 2—are (1) public transportation system improvements, (2) coordination among various freight modes (e.g., trucks, rail), (3) improved system management through real-time data and performance tools, and (4) in-vehicle crash avoidance systems.

Ultimately, further integration is needed for the longer term goal of widespread ITS deployment to be realized—culminating in stage 3. An integrated system requires a network of technologies working together along critical corridors and urban centers. Nevertheless, the challenge still remains in getting from "here" (individual building blocks) to "there" (a comprehensive network of technologies managing and disseminating information).

2.1 ITS Categories and Deployment Planning

To guide the development and deployment of ITS technologies, the U.S. Department of Transportation released its National ITS Architecture (designed in conjunction with its external advisory committee—ITS America—and stakeholder input) during the mid-1990s. The National ITS Architecture reflects a mature product, developed and updated over a 9-year period. It provides a common framework and language for planning and implementing ITS so that systems could be integrated functionally and geographically and be interoperable from one location to another. The National ITS Architecture is composed of 32 ITS user services bundled into eight categories: (1) travel and transportation management, (2) public transportation options, (3) electronic payment, (4) CVO, (5) emergency management, (6) advanced vehicle control and safety systems, (7) information management, and (8) maintenance and construction. For an overview of the 32 ITS user services, see Table I.

The National ITS Architecture provides an implementation roadmap that starts with ITS building blocks (or early market packages). The building blocks are characterized as low risk, reflecting early market opportunities and tangible benefits. Early market packages include (1) surface street control, (2) freeway control, (3) ETC, (4) traveler information, (5) transit vehicle tracking, (6) transit operations, and (7) electronic clearance (e.g., tolling, commercial vehicles).

The National ITS Architecture is tied to a 10-year plan (released in 2002) that addresses goals, objectives, user service requirements, and expected benefits. The five main goals outlined in the 10-year plan are (1) safety, (2) security, (3) efficiency/economy, (4) mobility/access, and (5) energy/environment. Furthermore, the 10-year plan develops a series of programmatic and enabling themes to describe the opportunities, benefits, and challenges of future transportation systems. There are four programmatic themes: (1) an integrated network of transportation information, (2) advanced crash avoidance technologies (including in-vehicle electronics), (3) automatic incident detection, notification, and response, and (4) advanced transportation management (including traffic and transit).

Two of the four ITS programmatic themes (the second and fourth themes just listed) could result in significant energy consumption benefits in the future. The advanced crash avoidance technologies programmatic theme focuses primarily on reducing the number of vehicle crashes. Adaptive in-vehicle electronics, integral to crash avoidance, is forecasted to reduce fuel consumption by (1) smoothing accelerations and decelerations (particularly for commercial and transit vehicles), (2) responding automatically to stop-and-go driving, (3) anticipating and adjusting the throttle and transmission for varying road conditions and terrain, and (4) enabling the safe movement of platoons of tightly spaced trucks and transit and other vehicles. In the longer term, such safety devices could also permit the introduction of lighter weight vehicles for greater fuel economy. In addition, route guidance products that help drivers to plan optional routes—in case of an incident—may significantly reduce miles driven, saving fuel and helping to mitigate congestion.

TABLE I
ITS User Services

User services bundle	User services
Travel and transportation management	Pre-trip travel information
	En route driver information
	Route guidance
	Ride matching and reservation
	Traveler services information
	Traffic control
	Incident management
	Travel demand management
	Emissions testing and mitigation
	Highway rail intersections
Public transportation operations	Public transportation management
	En route transit information
	Personalized public transit
	Public travel security
Electronic payment	Electronic payment services
Commercial vehicle operations	Commercial vehicle electronic clearance
	Automated roadside safety inspection
	On-board safety monitoring
	Commercial vehicle administration processes
	Hazardous materials incident response
	Commercial fleet management
Emergency management	Emergency notification and personal security
	Emergency vehicle management
Advanced vehicle control and safety systems	Longitudinal collision avoidance
	Lateral collision avoidance
	Intersection collision avoidance
	Vision enhancement for crash avoidance
	Safety readiness
	Pre-crash restraint deployment
	Automated vehicle operation
Information management	Archived data function
Maintenance and construction management	Maintenance and construction operations

Source. From Iteris Inc. and Lockheed Martin (2002). "Executive Summary National ITS Architecture Version 4.0." U.S. Department of Transportation, Washington, DC.

The advanced transportation management programmatic theme is also predicted to have a considerable effect on energy use in the future. Tools included in this area aim to intelligently and adaptively manage vehicle flows within the physical infrastructure and often across multiple jurisdictions and modes. Advanced transportation management systems rely on area-wide surveillance and detection, rapid acquisition and evaluation of traffic flow data, and predictive capabilities.

ITS America forecasts that these programmatic areas—crash avoidance and transportation management systems—will be critical to achieving the 10-year plan's energy goal: saving a minimum of 1 billion gallons of gasoline per year. The remainder of this article focuses on tools (models, field operational tests, and deployments) used to evaluate ITS impacts on energy consumption and early findings.

3. ITS AND ENERGY IMPACT EVALUATION TOOLS

In 1998, the U.S. Environmental Protection Agency (EPA) released a technical report that examined methodologies and research efforts aimed at evaluating the energy and environmental impacts of ITS. The report concluded that developing an ITS fuel consumption and emission impact assessment is an exceptionally challenging exercise due to the complex relationship among ITS, travel behavior, and transportation system management. Traffic simulation and travel demand models can aid in this research; however, more work is needed.

A key theme of ITS is the integrated deployment of information networks to support travel, increase transportation infrastructure use, and better manage demand. However, this integration requires models that can simulate the effects of information on traffic flow at corridor levels and travel behavior at regional levels. Independently, traditional tools (microsimulation, regional travel demand, and emission and fuel consumption models) cannot adequately capture critical linkages among ITS technologies and various feedback loops.

For example, travel demand models are typically unresponsive to information-related improvements (e.g., road conditions) and intersection- and corridor-level changes. Such impacts are usually evaluated using traditional microsimulation tools, which lack the behavioral assumptions of travel behavior models. Thus, new approaches are needed to capture complex ITS interactions and impacts. These tools must be flexible enough to (1) reflect different scenarios (e.g., market penetration and behavioral assumptions), (2) capture ITS impacts on dynamic

mode and route choice, and (3) simulate individual vehicle driving patterns.

Given the need for more complex tools, several advanced microsimulation models (e.g., INTEGRATION, DYNASMART) were developed to simulate the dynamic interactions among ITS technologies—including energy impacts. For example, INTEGRATION was designed with the purpose of simulating integrated networks, composed of freeways and arterial roads, for various ITS scenarios. INTEGRATION provides effectiveness measure estimates for individual vehicles, links, origin–destination trip pairs, and complete networks—including travel time, fuel consumption, and vehicle emissions. In addition, Traf–Nestim (another modeling tool) can be used to enhance the modal fuel consumption algorithms of INTEGRATION.

DYNASMART was developed to evaluate the dynamic effects of advanced traveler information and management systems on networks. This tool employs simple fuel consumption and emission models that reflect the relationship between uniform acceleration rates and speed. Consequently, this model cannot make adjustments for ambient temperatures, cold starts, and second-by-second driving behavior.

TRANSIMS, another ITS model initiative, stands for Transportation Analysis and Simulation System. This tool integrates a set of new travel forecasting and analysis procedures. TRANSIMS was developed to meet the travel forecasting and air quality analysis needs of metropolitan and state transportation and planning agencies. It is composed of six integrated modules: (1) population synthesis, (2) activity-based travel demand generation, (3) intermodal trip planning, (4) traffic microsimulation, (5) environmental analysis, and (6) feedback. The environmental module estimates air quality, energy consumption, and carbon dioxide emissions based on travel behavior outputs. As of 2003, TRANSIMS had been tested using data from the city of Dallas, Texas, and the city of Portland, Oregon. Developers anticipated that the fuel consumption module would be available for use in 2005.

Another common tool for evaluating ITS technologies is field operational tests (FOTs). Researchers and practitioners frequently conduct FOTs to demonstrate and evaluate ITS technologies in real-world settings. In some cases, there is an overlap between microsimulation tools and FOTs. Modeling is frequently used to simulate and predict real-world ITS deployment impacts. Furthermore, FOT data are often employed in modeling studies to simulate ITS impacts at different levels (e.g., corridor and higher market penetration). Early results, reported in the next section, reflect modeling, FOT, and deployment findings.

4. PREDICTED AND EARLY FINDINGS

Understanding of ITS fuel consumption impacts is still limited. A summary of predicted impacts for eight ITS categories in which fuel consumption has been studied over the past decade or so (at various levels) is presented in Table II. The eight categories are (1) traffic signal control, (2) traffic management and surveillance (e.g., ramp metering), (3) incident management, (4) electronic toll collection, (5) traveler information, (6) transit management, (7) commercial vehicle operations, and (8) vehicle control technologies. Energy impacts listed are both direct and indirect and are unsubstantiated by modeling or actual field test results.

Over the past decade or so, ITS technologies have been modeled, tested in limited FOTs and evaluated in full-scale deployment (e.g., electronic tolling, ramp metering, bus automatic vehicle location). Knowledge about ITS impacts is growing, as is an understanding of how to maximize benefits. Although fuel consumption is not the primary motivator of ITS (congestion relief and traffic and transit management are the key drivers), understanding of ITS energy impacts is increasing. Indeed, by reducing stop-and-go traffic and improving route guidance, ITS can have a positive impact on fuel consumption.

In the following subsections, early modeling, FOT, and deployment energy consumption results are quantified (when applicable). Although this analysis spans a wide range of ITS technologies, it does not represent an exhaustive review of ITS and fuel consumption impacts. Finally, there is a brief discussion regarding the potential effects and complexities of understanding latent or new demand that may result from ITS capacity enhancement strategies.

4.1 Traffic Signal Control

Traffic signals serve a variety of functions. They can be used to manage traffic speeds, vehicle merging and corridor crossings, and interactions between vehicles and low-speed or nonmotorized modes (e.g., bicycles, pedestrians, wheelchairs) at intersections. ITS

TABLE II

Predicted ITS Technology Impacts

ITS technology	Predicted impacts
Traffic signal control Traffic control technologies include traffic surveillance, ramp metering, lane control systems, and traffic signals	Reduced stop/idle delay times for vehicles traveling on main lanes and at intersections Better response to incidents and special events Some mode shifts to/from transit Reduced bus round-trip travel time, implying increased speed
Traffic management and surveillance ITS-based traffic management approaches include incident management, ramp metering, traffic signal control, traveler information, and traffic surveillance	Reduced freeway travel times due to ramp metering Increased ramp travel times Changes in acceleration (increased emissions) Dynamic changes in freeway exit points due to changeable message signs, thereby increasing freeway speeds Increases in high-occupancy vehicle travel resulting from high-occupancy vehicle bypass to ramp metering Increased travel times on arterials due to spillage of vehicles waiting to get on freeways Better use of existing freeway capacity as a result of lane control improvements
Incident management Incident management consist of three key areas: (1) traffic surveillance (incident detection and verification), (2) clearance, and (3) traveler information; in addition, this area includes emergency management services	Reduced incident response times Increased travel reliability Reduced delays, especially those due to incidents on highways and freeways (e.g., lane closures) Fewer secondary incidents caused by initial incidents Dynamic changes in destination, mode, and/or route choice
Electronic toll collection ETC technologies enable the instant payment of highway tolls when a vehicle passes through a toll station via an electronic roadside antenna (or reader) and a pocket-sized tag containing a radio transponder	Increase in toll lane capacity Fuel savings and a decrease in mobile emissions by reducing or eliminating waiting times
Traveler information Traveler information systems serve two primary functions, providing (1) information about transportation services and performance of the system to users and (2) an information linkage among various ITS components	Increased shifts in mode, destination, time of trip, and route choice over time
Transit management Transit management consists of four key areas: (1) transit vehicle tracking, (2) fare payment, (3) traveler information, and (4) personalized public transportation alternatives	Decreased boarding times Increased transit system reliability Improved transit fleet use Improved coordination between transit services, such as bus and rail transfers, promoting overall transit use Increased opportunity to provide more demand-responsive transit services (e.g., mode shifts)
Commercial vehicle operations CVO approaches include border crossing clearance and safety via electronic clearance and manifesting, automatic vehicle location, vehicle-to-fleet management center communications, on-board safety diagnostics to enable more effective roadside safety inspections, and hazardous materials incident response	Electronic clearance, safety, credentialing, and administrative processing save time by automating traditionally manual screening and inspection procedures Because of vehicle location technologies, emergency response teams can more easily locate and respond to HazMat accidents Reduced waiting times at international borders due to electronic clearance (inspectors can focus on noncompliant vehicles) On-board safety monitoring automatically alerts drivers of deficiencies in their vehicles' performance, ensuring greater roadway maintenance and safety

continues

Table II continued

ITS technology	Predicted impacts
Vehicle control technologies	
Advanced sensing, communication, and computing technologies represent the range of ITS technologies that can help to avoid collisions, prevent or lessen injuries when crashes do occur, and ultimately lead to full vehicle automation; some existing vehicle control technologies include adaptive cruise control, antilock brakes, and electronic system malfunction indicators	Reduce accidents resulting from unsafe headway, driver inattention, and errors in recognition and judgment Reduce antisocial driving behavior such as road rage Increased freeway capacity due to more closely spaced vehicle platoons Reduced fuel consumption and emissions due to traffic flow smoothing and vehicle platooning

microsimulation models enable researchers to assess the impacts of various traffic signal controls on emissions, travel times, and fuel consumption. Traffic signal controls can be set to optimize one or more desired goals (e.g., time savings, energy reduction). To maximize fuel efficiency, traffic signal controls can be fixed to reduce vehicle accelerations, decelerations, and idle times, all of which contribute to increased fuel consumption. Numerous—and sometime conflicting—goals must be balanced in traffic management. For instance, emission and fuel efficiency goals are typically not achieved using the same traffic control strategies employed to reduce travel times.

Using traffic signal controls to minimize energy consumption is not new, and studies to quantify energy impacts in this area predate ITS. For example, during the two energy crises of the 1970s (1973–1974 and 1979), researchers found that synchronizing traffic signal green lights to minimize stop-and-go traffic reduced energy consumption.

More recently, traffic signal control has become a predominant component of ITS, and its impacts have been evaluated in simulation and real-world settings. Overall, traffic signal control studies reveal fuel efficiency benefits ranging between 1.6 to 50.0%, with most results at less than 20%. Variability in estimated savings is a result of the following factors: quality of existing timing plans, network configuration, traffic patterns, and signal equipment.

4.2 Traffic Management and Surveillance

Ramp metering is one of several ITS technologies designed to manage traffic flow. The goal of ramp metering is to safely space vehicles merging onto a highway while minimizing speed disruptions to existing flows. Considerations include (1) public misunderstanding and system dislikes, (2) overflow of cars onto surface streets while waiting to enter ramps, and (3) driver use of arterial streets to avoid ramp meters. The most significant benefit of ramp metering is passenger time savings. Emission and fuel consumption impacts are mixed.

Ramp metering causes vehicles on ramps to stop and go, and this behavior consumes more fuel than does free-flow driving. Ramp metering also results in smoother vehicle flow on freeways because vehicles enter in a staggered and controlled manner, reducing bottlenecks that would otherwise impede traffic. This results in reduced fuel consumption. These two factors (increased stop-and-go traffic on on-ramps and decreased traffic flow disruption on highways) appear to negate each other. More detailed studies are necessary to understand how ramp metering effects interact and how they affect fuel consumption.

Whereas the larger category of traffic management and surveillance also includes incident management, traffic signal control, and traveler information, this subsection focuses on ramp metering primarily because each of the other categories is examined separately. Other traffic management tools include improved surveillance using loop detectors, closed circuit television, radar, lasers, video image processing, and vehicles equipped with toll tags or global positioning systems to determine travel times. Their impacts on fuel consumption relate directly to their use in incident management and traveler information.

4.3 Incident Management

ITS contributions to incident management include improved surveillance, verification, and dispatch to manage an incident. The use of changeable message signs and personal communications devices, such as cell phones and personal digital assistants (PDAs), can assist with early notification for upstream

drivers, resulting in reduced incident-related congestion because drivers have more time to select an alternative route. Improved incident management can decrease fuel consumption by reducing the delay and congestion associated with the blocked traffic. Although incident delay reductions are limited, model calculations for a Maryland initiative (called CHART) have shown fuel savings of 4.1 million gallons per year.

4.4 Electronic Toll Collection

ETC allows for electronic payment of highway and bridge tolls as vehicles pass through a toll station. Vehicle-to-roadside communication technologies include electronic roadside antennas (or readers) and pocket-sized tags containing radio transponders (typically placed inside a vehicle's windshield).

Studies show that ETC saves time and reduces energy consumption by reducing the stop-and-go traffic associated with vehicle queues approaching toll plazas, stopping to pay a toll, and accelerating to rejoin regular traffic flow. Analyses also support that fuel savings offset operating costs. ETC systems have the capacity to move five times as many vehicles as do conventional toll lanes. One study along the New Jersey Turnpike found savings of 1.2 million gallons of fuel per year due to reduced delays at toll plazas employing ETC. Approximately three-fourths of the reported savings accrued to passenger cars and one-fourth to commercial vehicles.

4.5 Traveler Information

Effective traveler information requires the accurate collection and dissemination of real-time travel information to transportation managers and the public to aid them in making informed decisions about travel time, mode, and route. A wide array of ITS technologies assist with traveler information, including in-vehicle guidance, Web sites, cell phones, PDAs, and changeable message signs to distribute user information.

The actual impact of traveler information on fuel consumption depends on a number of factors. For example, if ITS technologies assist drivers with route selection and guidance, fuel consumption benefits will likely be greater the less familiar a driver is with an area. The timeliness and delivery medium of information will also influence the degree to which travelers use it and subsequent energy impacts. Fuel consumption benefits might include mode shifts (e.g., from a single-occupancy vehicle to transit or bicycle) and energy savings proportional to travel time reductions achieved by taking alternate routes.

4.6 Transit Management

Transit managers are already implementing ITS technologies to improve service, including automatic vehicle location, real-time bus arrival signage, traffic signal priority, and automated information announcements. The degree to which ITS technologies improve transit services and attract new riders determines the significance of fuel consumption impacts.

Bus rapid transit (BRT) encompasses the use of a series of ITS technologies, route planning, exclusive right-of-ways, and management to improve service—each of which can reduce travel times. Increases in bus ridership due to BRT implementation have been reported in the United States, Australia, and Europe. BRT improvements along certain routes attract new riders and encourage existing riders to use the service more frequently. If a mode shift occurs from single-occupancy vehicles to BRT, there is an efficiency benefit. If the previous mode was nonmotorized, such as walking or cycling, the impact on fuel efficiency is negative. If additional riders are attracted from another bus route, the impact on fuel efficiency is neutral. An estimated 30 to 50% of additional riders on a new BRT route are new transit riders. Furthermore, faster journey times and reduced acceleration, deceleration, and idle times—resulting from fewer stops and signal priority—have been shown to reduce fuel consumption. Indeed, signal priority modeling results indicate a 5% reduction in fuel consumption.

4.7 Commercial Vehicle Operations

ITS applications in commercial vehicle management primarily include automatic vehicle identification and weigh-in-motion. The purpose of automatic identification and weigh-in-motion technologies in CVO is to enable the weighing and cataloging of trucks without causing vehicles to stop and queue in line.

Simulation modeling and on-road testing reveal increased fuel efficiency due to weigh-in-motion technologies. Measured against static scales, high-speed weigh-in-motion systems demonstrate the greatest fuel benefits. Weigh-in-motion ramps that require trucks to slow—but not stop—also result in fuel savings, although not as significant as the savings with high-speed weigh-in-motion stations.

The purpose of automatic vehicle identification is to identify trucks, drivers, and loads as a companion function to weigh-in-motion.

In addition, CVO can result in fewer trucks being forced to bypass weigh stations due to full queues at static scales. Although commercial ITS applications demonstrate a clear fuel benefit, this value depends on the number and nature of stations passed.

4.8 Vehicle Control Technologies

ITS technologies that automate vehicle control systems aim to improve vehicle safety, efficiency, and comfort. These technologies include intelligent cruise control, antilock brakes, electronic system malfunction indicators, and automated highway systems (e.g., platooned vehicles). Simulation research indicates that some automated vehicle control technologies could have a positive impact on fuel consumption.

Intelligent cruise control refers to technologies that can identify the distance in front of a vehicle on a highway and correspondingly modify a car's controlled speed to accommodate lane merging and changes in the speed of vehicles ahead. This results in reduced fluctuations in the speed of controlled cars, resulting in a positive fuel efficiency effect. One simulation study showed fuel savings ranging from 8.5 to 28.5% when 10% of vehicles in a lane are equipped with intelligent cruise control.

Another group of vehicle control technologies is being tested for automated highway systems. The concept behind automated highways is to employ technologies that facilitate vehicle-to-vehicle and vehicle-to-roadside communication to improve safety and system efficiency. In this way, vehicles can operate in very close proximity to each other. Simulations indicate a 5 to 15% reduction in fuel consumption due to aerodynamic drafting effects.

4.9 Human Factors

As already demonstrated, many ITS technologies can have a positive impact on fuel consumption. Some of the energy benefits accrue from reduced congestion and stop-and-go driving, resulting in smoother traffic flows. In effect, ITS technologies can increase existing roadway capacity without increasing infrastructure. Latent (or additionally generated) demand is often at the forefront of infrastructure enhancement discussions. In regions with the greatest congestion, travel demand may be suppressed due to travel time costs. Under such conditions, adding or managing infrastructure to ease congestion may result in additional travel demand until delay costs rise enough to suppress travel again. Studies show that increased roadway infrastructure can also result in modal shifts from transit ridership to personal vehicles.

If ITS technologies are successful at reducing congestion and travel delay, roadway capacity will be increased without adding new infrastructure. An outstanding question regarding ITS capacity enhancement is whether or not added capacity will result in latent demand for highway use—similar to some demonstrated infrastructure expansion impacts (e.g., mode shifts from transit to auto use). If improved traffic throughput due to ITS results in increased highway demand and modal shifts away from transit, some of the systemwide fuel efficiency benefits may be negated. Because evidence to support latent demand is still not definitive, and final stage 3 ITS implementation has not yet been achieved, no conclusions can be drawn regarding the causal relationship between ITS capacity improvements and latent demand.

Other ITS technologies (e.g., real-time road pricing) may be used to offset increased travel demand to send signals to the traveling public regarding travel costs. By employing sophisticated surveillance technologies, vehicle identification, and electronic payment systems, ITS can assist in administering congestion pricing within incremental time blocks, thereby reducing congestion and energy use.

5. SYSTEM INTEGRATION AND CONCLUSION

At the current stage of ITS development and deployment, several early predictions regarding fuel efficiency impacts are being realized (e.g., electronic tolling, signal priority for buses). However, modeling and early ITS deployments are still occurring largely in discrete applications rather than across integrated regionwide networks. To understand the full effects of ITS on transportation systems (traffic and transit), technologies must be deployed in a comprehensive manner. Similarly, the full energy impacts of ITS cannot be known until technologies are integrated and complex dynamics, including human factors, are modeled and tested. Given the complex interaction of ITS technologies and human factors, development

and use of suitable tools to measure environmental consequences, including fuel consumption, will remain important. Interrelationships among various ITS elements will determine the ultimate direction and degree of impacts. In addition, impacts will vary among regions with different traffic patterns and system use.

Overall, integration of individual ITS components can be expected to multiply benefits by providing the traveling public with a wider array of choices and real-time information. In the EPA's 1998 technical report on ITS, the agency identified four key dimensions to comprehensive ITS, requiring integration of

- information across key ITS elements such as transit management, freeway management, emergency and incident response, and traffic signal control;
- information across regions such as coordination between jurisdictions to ensure smoother traffic flow;
- transit management systems across regions to ensure integrated user services;
- incident management information across regions to provide the fastest possible response.

Ultimately, the integrated ITS vision should provide operational efficiencies and interjurisdictional coordination benefits that result in comprehensive and improved system management. The magnitude of these benefits will depend on factors such as market acceptance of available technologies to deliver information, user-perceived accuracy of the information provided, and level of personalized traveler information. Not surprisingly, the greatest travel time and energy benefits will come from traveler information persuading users to take public transportation or postpone their trips until congestion has cleared.

SEE ALSO THE FOLLOWING ARTICLES

Energy Efficiency, Taxonomic Overview • Passenger Demand for Travel and Energy Use • Transportation and Energy, Overview • Transportation and Energy Policy • Vehicles and Their Powerplants: Energy Use and Efficiency

Further Reading

Barth, M. (2000). An emission and energy comparison between a simulated automated highway system and current traffic conditions. *In* "2000 IEEE Intelligent Transportation Systems Conference Proceedings." Dearborn, MI.

Bose, A., and Ioannou, P. (2001). "Analysis of Traffic Flow with Mixed Manual and Intelligent Cruise Control Vehicles: Theory and Experiments" (California PATH Research Report UCB-ITS-PRR-2001-13). California PATH, Richmond, CA.

BP. (2002). BP statistical review of world energy 2002. www.bp.com/centres/energy2002/usenergy.asp.

Intelligent Transportation Society of America. (2002). Delivering the future of transportation: The National Intelligent Transportation Systems program plan—A ten-year vision. www.itsa.org.

Intelligent Transportation Systems Joint Program Office, U.S. Department of Transportation. (2003). ITS benefits and unit costs database. www.benefitscost.its.dot.gov.

New Technology and Research Program. (1996). "Advanced Transportation Systems Program Plan." California Department of Transportation, Sacramento.

Schrank, D., and Lomax, T. (2002). The 2002 urban mobility report. www.mobility.tamu.edu.

Shaheen, S. A., Young, T. M., Sperling, D., Jordan, D., and Horan, T. (1998). "Identification and Prioritization of Environmentally Beneficial Intelligent Transportation Technologies" (Research Report UCD-ITS-RR-98-1). Institute of Transportation Studies, University of California, Davis.

U.S. Department of Transportation (2002). The National ITS Architecture version 4.0. itsarch.iteris.com/itsarch.

U.S. Environmental Protection Agency. (1998). "Assessing the Emissions and Fuel Consumption Impacts of Intelligent Transportation Systems (ITS)" (EPA 231-R-98-007). EPA, Washington, DC.

Internal Combustion Engine Vehicles

K. G. DULEEP
Energy and Environmental Analysis, Inc.
Arlington, Virginia, United States

1. Introduction
2. Vehicle Energy Efficiency
3. Improving Efficiency by Reducing Tractive Energy Required
4. Improvement of Engine Efficiency
5. Increasing the Efficiency of Spark Ignition Engines
6. Increasing the Efficiency of Compression Ignition (Diesel) Engines
7. Intake Charge Boosting
8. Alternative Heat Engines

Glossary

engine efficiency Amount of energy produced by the engine per unit of fuel energy consumed.
fuel economy Vehicle distance traveled per unit volume of fuel consumed.
internal combustion engine vehicle Vehicle where primary motive power is derived from an engine that converts fuel energy to work using the air-fuel mixture as the working fluid.
light duty on highway vehicles Cars and light trucks with a fully loaded weight below 6000 k (13,200 lbs).
off-highway vehicle Vehicles designed to operate primarily on unpaved surfaces.

The vast majority of vehicles used in the world are powered by internal combustion engines (ICE). Other forms of propulsion such as electric motors or external combustion steam engines are used in specialized applications that account for a small fraction of the total vehicle fleet. Most vehicles are now powered by reciprocating piston engines that use the Otto cycle (also called the spark-ignition engine) or the diesel cycle (also called the compression-ignition engine). Gas turbines are used primarily in marine vessels and aircraft and are not discussed here. A small number of vehicles using the Wankel engine have also been sold. Internal combustion engine–powered vehicles typically account for one-quarter to one-third of total energy consumption in most countries, and their fuel consumption and fuel efficiency are issues of major concern.

1. INTRODUCTION

The on-highway fleet of vehicles accounts for over 95% of all vehicles in operation worldwide (which is in excess of a billion vehicles). The remainder is composed of off-highway vehicles, equipment such as forklifts or bulldozers, and motorcycles. Annual sales of on-highway vehicles exceeded 57.6 millions units worldwide in 2002 with 39.5 millions unit classified as passenger cars and 18.1 million units classified as trucks. The distinction between cars and trucks is not always clear (especially for light trucks) but trucks are usually used for cargo hauling or for carrying more than six passengers. Cars span the gross vehicle weight (GVW) range from 1 to 3 tons, while trucks typically span the GVW range of 2 to 40 tons. Off-road and specialized vehicles can be much heavier.

The majority of cars and light trucks (under five tons GVW) are powered by spark ignition engines, while most trucks that weigh more than 5 tons GVW are powered by diesel engines. Since the early 1990s, diesel engines have become more popular for cars and light trucks in the European Union (EU). The diesel engines' share in the new car market was over 50% in 2002 in countries such as France and Austria. In contrast, few diesel engine–powered cars and light trucks are sold in North America.

 © 2004 Elsevier Inc. All rights reserved.

2. VEHICLE ENERGY EFFICIENCY

Vehicle energy efficiency is generally defined in terms of fuel economy measured in miles per gallon (mpg) or kilometers per liter of fuel. It can also be measured in terms of fuel consumption, which is the inverse of fuel economy, in units of liters per 100 km or gallons per 100 miles. The fuel economy of a vehicle is strongly dependent on the vehicle's overall weight, but is also dependent on the efficiency of the engine, as well as the matching of the engine characteristics to the vehicle's operational requirements. The fuel economy of a particular vehicle is dependent on the load carried, the driving cycle, the ambient temperature, and the characteristics of the road such as its gradient and surface roughness. Hence, the fuel economy of a specific vehicle can vary widely depending on how and where it is used.

In most developed countries, on-road light vehicles (cars and light trucks) are certified for emissions and fuel economy by the government. The fuel efficiency rating is measured in a laboratory-controlled environment and on a specified driving cycle. In the United States and Canada, for example, light vehicle fuel economy is measured on a "city cycle" with an average speed of about 20 mph and a "highway cycle" with an average speed of about 50 mph. All aspects of the fuel economy test, ranging from the ambient temperature to the specification of the fuel used, are tightly controlled and this results in a fuel economy measurement that is repeatable to within $\pm 2\%$, typically. While this measured fuel economy may differ significantly from the fuel economy for the same vehicle in any specific use, the measured value provides a comparative benchmark for vehicle fuel economy that is useful from a vehicle buyer's perspective and from an engineering perspective.

The sales-weighted average test fuel economy of all new vehicles sold in the United States is about 28 mpg for cars and 21 mpg for light trucks. Much of this difference between car and light-truck fuel economy is attributable to the fact that light trucks are larger and heavier than cars, but some of the difference is also attributable to the fact that cars utilize higher levels of efficiency enhancing technology. Fuel economy levels in Australia and Canada are similar to the U.S. levels, but cars in the EU have about 25% higher fuel economy, on average. The higher fuel economy in the EU is partly due to the smaller size and weight of cars sold and partly due to the higher penetration of diesel engines, which are more efficient than spark ignition engines.

Studies conducted by technical agencies have concluded that vehicle fuel economy can be increased substantially from average values without any reduction of attributes such as interior space of cargo carrying ability. The sources of energy loss and the technology available to reduce these losses are described later.

A simple model of energy consumption in conventional automobiles provides insight into the sources and nature of energy losses. In brief, the engine converts fuel energy to shaft work. This shaft work is used to overcome the tractive energy required by the vehicle to move forward, as well as to overcome driveline losses and supply accessory drive energy requirements. The tractive energy can be separated into the energy required to overcome aerodynamic drag force, rolling resistance and inertia force. It is useful to consider energy consumption on the U.S. city and highway test cycles, which are reference cycles for comparing fuel economy.

Denoting the average engine brake specific fuel consumption over the test cycle as bsfc, we have fuel consumption, FC, given by

$$\mathrm{FC} = \frac{\mathrm{bsfc}}{\eta_\mathrm{d}} [\mathrm{E_R} + \mathrm{E_A} + \mathrm{E_K}] + \mathrm{bsfc}\ \mathrm{E_{AC}} + \mathrm{G_i}(\mathrm{t_i} + \mathrm{t_b})$$

where η_d is the drive train efficiency, E_R is the energy to overcome rolling resistance, E_A is the energy to overcome aerodynamic drag, E_K is the energy to overcome inertia force, E_{AC} is the accessory energy consumption, G_i is idle fuel consumption per unit time, and t_i, t_b are the time spent at idle and braking.

The first term in the above equation represents the fuel consumed to overcome tractive forces. Since the Federal Test Procedure (FTP) specifies the city and highway test cycle in terms of speed versus time, E_R, E_A, and E_k can be readily calculated as function of the vehicle weight, the tire rolling resistance, body aerodynamic drag coefficient, and vehicle frontal area. Weight reduction reduces both inertia force and rolling resistance.

It should be noted that not all of the inertia force is lost to the brakes, as a vehicle will slow down without the use of brakes, at zero input power due to aerodynamic drag and rolling resistance. Braking energy *loss* is approximately 35% in the city cycle and 7% on highway cycle. The fuel energy is used not only to supply tractive energy requirements but also to overcome transmission losses, accounting for the transmission efficiency that is in the first term.

The second term in the equation is for the fuel consumed to run the accessories. Accessory power requirements are required to run the radiator cooling

fan, alternator, water pump, oil pump, and power steering pump. Air conditioners also absorb power but are not reflected in official fuel economy estimates since they are not turned on during the FTP. Idle and braking fuel consumption are largely a function of engine size and idle RPM, while transmission losses are function of transmission type (manual or automatic) and design. The engine produces no power during idle and braking but consumes fuel, so that factor is accounted for by the third term. Table I shows the energy consumption (as a percentage) by all of these factors for a typical U.S. midsize car of mid-1990s vintage, with a 3-liter displacement s.i. engine, four-speed automatic transmission with lock-up, and power steering.

The values in Table I can be utilized to derive sensitivity coefficients for the reduction of various loads. For example, reducing the weight by 10% will reduce both rolling resistance and inertia weight forces, so that tractive energy is reduced by $(30.35 + 40.22) \times 0.1$ or 7.06% on the composite cycle. Fuel consumption will be reduced by $7.06\% \times 0.6544$, which is the fraction of fuel used by tractive energy, or 4.6%. This matches the common wisdom that reducing weight by 10% reduces fuel consumption by 4 to 5%. However, if the engine is also downsized by 10% to account for the weight loss, fuel consumption will be reduced by 5.8% since idle and braking fuel consumption will be reduced in proportion to engine size. In addition, there will be some reduction (0.5%) in transmission and drivetrain loss.

Fuel economy can be improved by two primary methods: (1) by reducing the power required to propel the vehicle and (2) by increasing the engine efficiency. To estimate the effects of different technology improvements that affect engine power required or the efficiency of the engine, it is useful to keep certain vehicle attributes constant. Vehicle attributes of interest to consumers are passenger room, cargo space or payload capability, acceleration performance, and vehicle comfort/convenience features. The impact of technology on fuel economy is typically measured while keeping these attributes constant.

Reducing the power required to propel the vehicle reduces engine load and can be accomplished by reduction of weight, aerodynamic drag, rolling resistance, or accessory loads. Engine efficiency increases can be accomplished not only by engine technologies but also by improved drivetrain technologies that improve the match between engine operating point and vehicle power requirements. Spark ignition engines convert only about 20 to 25% of fuel energy to useful work during typical driving so that a doubling of engine efficiency is theoretically possible without changing the basic Otto cycle.

3. IMPROVING EFFICIENCY BY REDUCING TRACTIVE ENERGY REQUIRED

Since vehicle weight is one of the most important variables determining fuel economy, weight reduction is an important method of improving fuel economy. The vehicle's weight is distributed between the body structure, the drivetrain, the vehicle's

TABLE I

Energy Consumption as a Percentage of Total Energy Requirements for a Typical Midsize Car[a]

	City	Highway	Composite[b]
Percentage of total tractive energy			
Rolling resistance	27.7	35.2	30.35
Aerodynamic drag	18.0	50.4	29.43
Inertia (Weight) force	54.3	14.4	40.22
Total	100	100	100
Percentage of total fuel consumed			
Tractive energy	57.5	80.0	65.44
Accessory energy	10.0	6.5	8.76
Idle + Braking consumption	15.0	2.0	10.41
Transmission + Driveline loss	17.5	11.5	15.39

[a] Midsize car of inertia weight = 1588 kg, $C_D = 0.33$, $A = 2.1\,m^2$, $C_R = 0.011$, 3L OHV V-6, power steering, four-speed automatic transmission with lock-up, air conditioning.

[b] Highway fuel economy is 1.5 times city fuel economy, and composite figures are based on the U.S. EPA 55% city/45% highway fuel consumption weighting.

interior, and vehicle suspension/tires. The first two component groups account for over 75% of a vehicle's weight. Weight can be reduced in all four component groups by improved structural design as well as by the use of alternative materials.

Improved structural design and packaging has been made possible through advanced computer simulations of structural strength, so that material use and shape can be optimized for the loads encountered. Most modern cars feature unibody designs where the body panels carry the structural loads, but several older models as well as many light trucks continue to use a separate chassis to carry structural loads. Heavy trucks, however, almost always utilize a separate chassis on which body components are mounted. A new architecture called space frame designs have emerged where structural loads are carried on skeletal frame from which body panels are hung. Improved packaging by optimization of component placement, body layout, and drivetrain location can also yield weight benefits. The placement of the engine transversely between the front wheels and driving the front wheels provides significant packaging benefits over front engine, rear-wheel-drive packages for light-duty vehicles.

The use of alternative materials such as ultra-high-strength steel, aluminum, and plastic composites is another way to reduce weight. Because of its low cost, steel, and cast iron continue to be the material of choice for body structures. Aluminum is already widely used for engine blocks and cylinder heads, and it is also used in critical suspension components. Some luxury cars now feature all-aluminum bodies, which weigh 30 to 35% less than their steel counterparts. Plastic composites are also widely used in body closures such as fenders, hood, and decklid with weight savings of 20 to 25% relative to steel parts. Such composites also see wide usage for lightweight interiors in the vehicle dashboard, seats, and door panels. Specially constructed prototypes maximizing the use of lightweight alternative materials have shown that weight reduction of 25 to 30% (relative to a conventional average steel vehicle) is possible, although with higher cost and with manufacturability constraints. The use of alternative materials in heavy trucks may not reduce loaded weight but will permit a larger payload to be carried.

Aerodynamic drag can be reduced by styling the vehicle's exterior shape and guiding the vehicle's interior airflow. At the speeds experienced by a typical vehicle, low drag shapes are a result of careful attention to airflow at the front of the vehicle, rear wheel wells and outside mirrors, and at the end of the roof. A measure of the drag is the aerodynamic drag coefficient, C_D, which is defined as

$$C_D = \frac{\text{Drag force}}{\frac{1}{2}\rho v^2 A},$$

where ρ is the density of air, v is the velocity of airflow, and A is the vehicle frontal area.

In the early 1980s, cars had drag coefficients of 0.45 to 0.5. By 2000, the most aerodynamic cars had drag coefficient of 0.25 to 0.28. Trucks typically have higher drag coefficients because of their boxy shape and increased ground clearance, relative to cars. Prototype cars with drag coefficients as low as 0.15 have been built, but such designs typically involve reduction of vehicle attributes such as reduced rear passenger headroom, reduced rear visibility, or reduced cargo space. Nevertheless, drag reduction still offers opportunities to reduce fuel consumption.

The tires' rolling resistance is the third major contributor to overall load. The tire rolling resistance coefficient (C_R) is a measure of tire energy loss, and is defined as

$$C_R = \frac{T}{L \cdot R},$$

where T is the torque required at any speed, R the tire radius, and L the vertical load on the tire. Typically, most modern tires have a C_R in the range of 0.009 to 0.012.

The tire rolling resistance results from a combination of tire-to-road friction and hysteresis. As the tire deforms, heat is dissipated in the tire's sidewall and tread due to the visco-elastic nature of rubber. In comparison, a steel wheel riding on steel rails has about one-tenth the rolling resistance of a rubber tire. Tire rolling resistance can be reduced by improved design of the tire tread, shoulders, and belts. In addition, the tire material formulation can significantly reduce hysteresis loss. The use of silica compounds mixed with rubber has been found to reduce rolling resistance, without affecting other desirable properties such as braking and wet traction. Design improvements and changes in belt material are also capable of reducing C_R with limited or no reduction of desirable attributes. It appears possible to reduce C_R by 15 to 25% in the short term and by up to 40% over the long term ($\sim$25 years).

4. IMPROVEMENT OF ENGINE EFFICIENCY

Engine efficiency on the driving cycle is the most significant determinant of vehicle fuel economy for a

vehicle of a specific weight. Heat engine efficiency can be stated in several ways. One intuitively appealing method is to express the useful energy produced by an engine as a percentage of the total heat energy that is theoretically liberated by combusting the fuel. This is sometimes referred to as "the first law" efficiency, implying that its basis is the first law of thermodynamics, the law of conservation of energy. Another potential, but less widely used, measure is based on the second law of thermodynamics, which governs how much of that heat can be converted to work. Given a maximum combustion temperature (usually limited by engine material considerations and by emission considerations), the second law postulates a maximum efficiency based on an idealized heat engine cycle called the Carnot cycle. The ratio of the first law efficiency to the Carnot cycle efficiency can be utilized as a measure of how efficiently a particular engine is operating with reference to the theoretical maximum based on the second law of thermodynamics. However, the most common measure of efficiency used by automotive engineers is termed brake specific fuel consumption (bsfc), which is the amount of fuel consumed per unit time per unit of power. In the United States, the bsfc of engines is usually stated in pounds of fuel per brake horsepower hour, whereas the more common metric system measurement unit is in grams per kilowatt-hour (g/kwh). The term brake here refers to a common method historically used to measure engine shaft power output. Of course, all three measures of efficiency are related to each other.

The efficiency of Otto and diesel cycle engines is not constant but depends on the operating point of the engine as specified by its torque output and shaft speed (revolutions per minute or RPM). Engine design considerations, frictional losses, and heat losses result in a single operating point where efficiency is highest. This maximum efficiency for an s.i. engine usually occurs at relatively high torque and at low to mid-RPM within the operating RPM range of the engine. At idle, the efficiency is zero since the engine is consuming fuel but not producing any useful work. When considering efficiency in a vehicle, the maximum efficiency need not, by itself, be an indicator of the average efficiency under normal driving conditions, since engine speed and torque vary widely under normal driving. The maximum efficiency of an engine is of interest to automotive engineers, but a more practical measure of efficiency is its average efficiency during "normal" driving or during the official city and highway fuel economy test.

4.1 Theoretical Maximum Engine Efficiency

The characteristic features common to all piston internal combustion engines are as follows:

1. Intake and compression of the air or air-fuel mixture
2. Raising the temperature (and hence, the pressure) of the compressed air by combustion of fuel
3. The extraction of work from the high-pressure products of combustion by expansion
4. Exhaust of the products of combustion

Combustion of the homogenous air-fuel mixture in a spark ignition engine takes place very quickly relative to piston motion and is represented in idealized calculations as an event occurring at constant volume. According to classical thermodynamic theory, the thermal efficiency, η, of an idealized Otto cycle, starting with intake air-fuel mixture drawn in at atmospheric pressure, is given by

$$\eta = 1 - 1/r^{n-1}, \tag{1}$$

where r is the compression (and expansion) ratio and n is the ratio of specific heat at constant pressure to that at constant volume for the mixture. The equation shows that efficiency increases with increasing compression ratio.

Using an n value of 1.4 for air, the equation predicts an efficiency of 58.47% at a compression ratio of 9:1. A value of $n = 1.26$ is more correct for products of combustion of a stoichiometric mixture of air and gasoline. A stoichiometric mixture corresponds to an air-fuel ratio of 14.7:1, and this air-fuel ratio is typical for most spark ignition engines sold in the United States. At this air-fuel ratio, calculated efficiency is about 43.5%. Actual engines yield still lower efficiencies even in the absence of mechanical friction, due to heat transfer to the wall of the cylinder and the inaccuracy associated with assuming combustion to be instantaneous. Figure 1 shows the pressure-volume cycle of a typical spark ignition engine and its departure from the ideal relationship.

Compression ratios are limited by the octane number of gasoline, which is a measure of its resistance to preignition or knock. At high compression ratios, the heat of compression of the air-fuel mixture becomes high enough to induce spontaneous combustion of small pockets of the mixture, usually those in contact with the hottest parts of the combustion chamber. These spontaneous combustion events are like small explosions that can damage the engine and

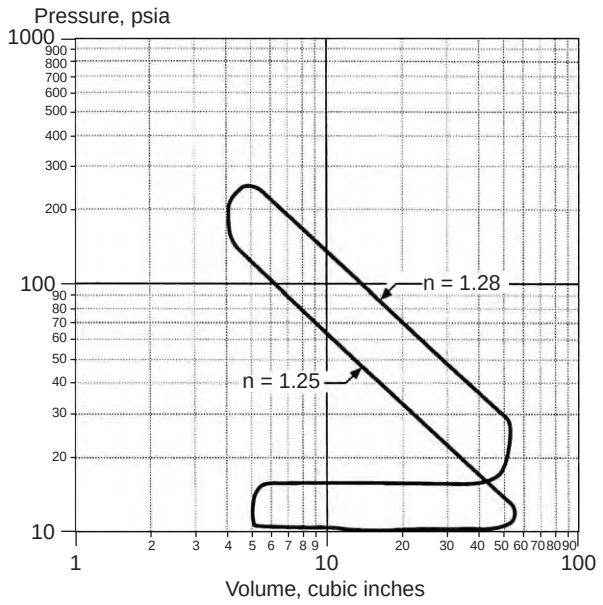


FIGURE 1 Pressure-volume diagram for a gasoline engine. Compression ratio = 8.7–11.

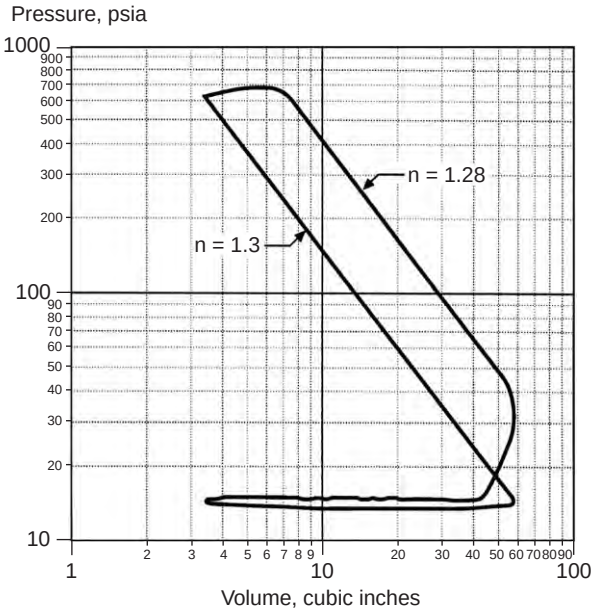


FIGURE 2 Pressure-volume diagram for a diesel engine. Compression ratio = 17–13.

can reduce efficiency depending on when they occur during the cycle. Higher octane number gasoline prevents these events, but also costs more and requires greater energy expenditure for manufacture at the refinery. The octane number is measured using two different procedures resulting in two different ratings for a given fuel, called motor octane and research octane number. Octane numbers displayed at the pump are an average of research and motor octane numbers, and most engines sold in the United States require regular gasoline with a pump octane number of 87. Typical compression rates for engines running on regular gasoline are in the range of 9:1 to 10:1.

The diesel, or compression ignition, engine differs from the spark ignition engine in that only air, rather than the air-fuel mixture, is compressed. The diesel fuel is sprayed into the combustion chamber at the end of compression in a fine mist of droplets and the diesel fuel ignites spontaneously upon contact with the compressed air due to the heat of compression. The sequence of processes (i.e., intake, compression, combustion, expansion, and exhaust) is similar to that of an Otto cycle engine. However, the combustion process occurs over a relatively long period and is represented in idealized calculations as an event occurring at constant pressure (i.e., combustion occurs as the piston moves downward to increase volume and decrease pressure at a rate offsetting the pressure rise due to heat release). Figure 2 shows the pressure-volume cycles for a typical diesel engine and its relationship to the ideal diesel cycle. If the ratio of volume at the end of the combustion period to the volume at the beginning of the period is r_c, or the cutoff-ratio, the thermodynamic efficiency of the idealized constant-pressure combustion cycle is given by

$$\eta = \frac{1}{r^{n-1}}\left[\frac{r_c^n - 1}{n(r_c - 1)}\right]. \qquad (2)$$

It can be seen that for $r_c = 1$, the combustion occurs at constant volume and the efficiency of the diesel and Otto cycle are equivalent.

The term r_c also measures the interval during which fuel is injected, and it increases as the power output is increased. The efficiency equation shows that as r_c is increased, efficiency falls so that the idealized diesel cycle is less efficient at high loads. The combustion process also is responsible for a major difference between diesel and Otto cycle engines. In an Otto cycle engine, intake is air throttled to control power while maintaining a near constant air-fuel ratio; in a diesel engine, power control is achieved by varying the amount of fuel injected while keeping the air mass inducted per cycle at near constant levels. In most operating modes, combustion occurs with considerable excess air in a c.i. engine, while combustion occurs at or near stoichiometric air-fuel ratios in a modern s.i. engine.

At the same compression ratio, the Otto cycle has the higher efficiency. However, diesel cycle engines normally operate at much higher compression ratios, since there are no octane limitations associated with this cycle. In fact, spontaneous combustion of the fuel is required in such engines, and the ease of spontaneous combustion is measured by a fuel property called cetane number. Most c.i. engines require diesel fuels with a cetane number over 40.

In practice, there are two kinds of c.i. engines, the direct injection type (DI) and the indirect injection type (IDI). The DI type utilizes a system where fuel is sprayed directly into the combustion chamber. The fuel spray is premixed and partially combusted with air in a prechamber in the IDI engine, before the complete burning of the fuel in the main combustion chamber occurs. DI engines generally operate at compression ratios of 15 to 20:1, while IDI engines operate at 18 to 23: 1. The theoretical efficiency of a c.i. engine with a compression ratio of 20:1, operating at a cutoff ratio of 2, is about 54% (for combustion with excess air, n is approximately 1.3). In practice, these high efficiencies are not attained, for reasons similar to those outlined for s.i. engines.

4.2 Actual versus Theoretical Efficiency

Four major factors affect the efficiency of s.i. and c.i. engines. First, the ideal cycle cannot be replicated due to thermodynamic and kinetic limitations of the combustion process, and the heat transfer that occurs from the cylinder walls and combustion chamber. Second, mechanical friction associated with the motion of the piston, crankshaft, and valves consume a significant fraction of total power. Since friction is a stronger function of engine speed rather than torque, efficiency is degraded considerably at light load and high RPM conditions. Third, aerodynamic frictional losses associated with airflow through the air cleaner, intake manifold and valves, exhaust manifold, silencer, and catalyst are significant, especially at high airflow rates through the engine. Fourth, pumping losses associated with throttling the airflow to achieve part-load conditions in spark ignition engines are very high at light loads. Note that c.i. engines do not usually have throttling loss, and their part load efficiencies are superior to those of s.i. engines. Efficiency varies with both speed and load for both engine types.

Hence, production spark ignition or compression ignition engines do not attain the theoretical values of efficiency, even at their most efficient operating point. In general, for both types of engines, the maximum efficiency point occurs at an RPM that is intermediate to idle and maximum RPM and at a level that is 60 to 75% of maximum torque. On-road average efficiencies of engines used in cars and light trucks are much lower than peak efficiency, since the engines generally operate at very light loads during city driving and steady-state cruise on the highway. High power is utilized only during strong accelerations, at very high speeds, or when climbing steep gradients. The high load conditions are relatively infrequent, and the engine operates at light loads much of the time during normal driving.

During normal driving, the heat of fuel combustion is lost to a variety of sources and only a small fraction is converted to useful output, resulting in the low values for on-road efficiency. Figure 3 provides an example of the heat balance for a typical modern small car with a spark ignition engine under a low speed (25 mph or 40 mph) and a high-speed (62 mph or 100 mph) condition. At very low driving speeds typical of city driving, most of the heat energy is lost to the engine coolant. Losses associated with other waste heat include radiant and convection losses from the hot engine block and heat losses to the engine oil. A similar heat loss diagram for a diesel c.i. would indicate lower heat loss to the exhaust and

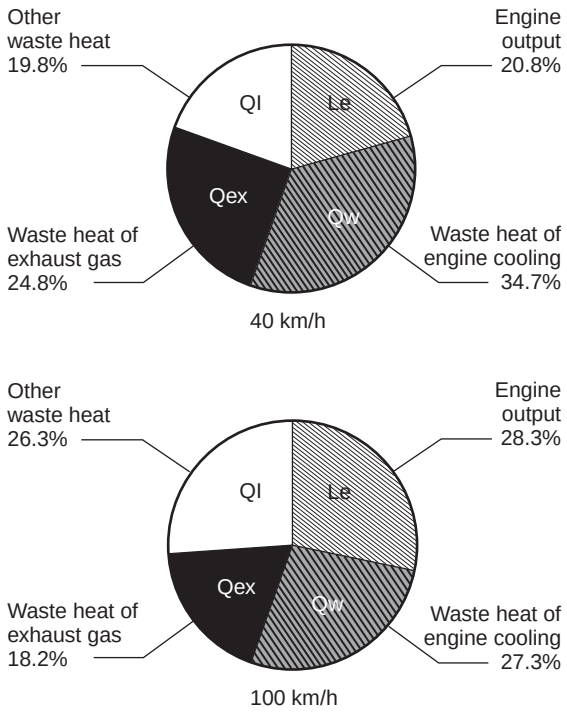


FIGURE 3 Heat balance of a passenger car equipped with 1500cc engine 17.

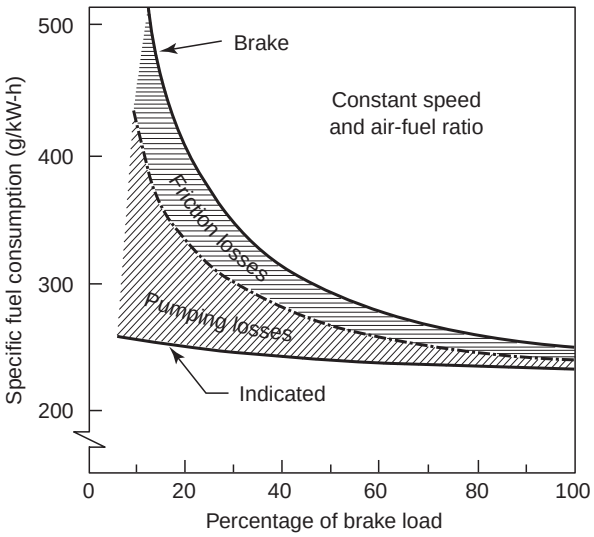


FIGURE 4 Specific fuel consumption versus engine load 18.

coolant and an increased fraction of heat converted to work, especially at the low speed condition. During stop-and-go driving conditions typical of city driving, efficiencies are even lower than those indicated in Fig. 3 because of the time spent at idle where efficiency is zero. Under the prescribed U.S. city cycle conditions, typical modern spark ignition engines have an efficiency of about 18%, modern IDI c.i. engines have an efficiency of about 21%, and modern DI diesel have an efficiency of about 23%.

Another method of examining the energy losses is by allocating the power losses starting from the power developed within the cylinder. The useful work corresponds to the area that falls between the compression and expansion curve depicted in Figs 1 and 2. The pumping work that is subtracted from this useful work, referred to as indicated work, is a function of how widely the throttle is open and, to a lesser extent, the speed of the engine. Figure 4 shows the dependence of specific fuel consumption (or fuel consumption per unit of work) with load, at constant (low) engine RPM. Pumping work represents only 5% of indicated work at full load, low RPM conditions, but increases to over 50% at light loads of less than two-tenths of maximum power.

Mechanical friction and accessory drive power, on the other hand, increase nonlinearly with engine speed but do not change much with the throttle setting. Figure 5 shows the contribution of the various engine components as well as the alternator, water pump and oil pump to total friction, expressed in terms of mean effective pressure, as a function of RPM. The brake mean effective pressure is

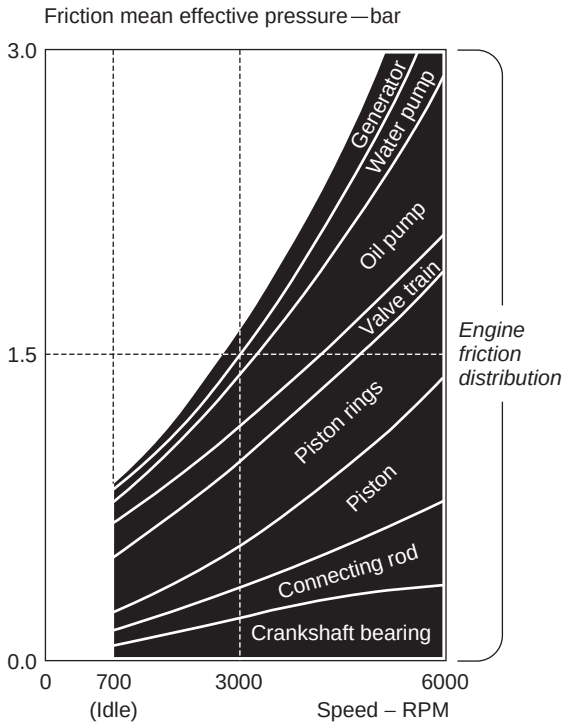


FIGURE 5 Friction distribution as a function of engine speed 19.

a measure of specific torque, or torque per unit of engine displacement; typical engine brake mean effective pressure (bmep) of spark ignition engines that are not supercharged range from 8.5 to 10 bar. Hence, friction accounts for about 25% of total indicated power at high RPM (~6000) but only for about 10% of indicated power at low RPM (~2000) in spark ignition engines. Friction in a c.i. engine is higher because of the need to maintain an effective pressure seal at high compression ratios, and the friction mean effective pressure is 30 to 40% higher than that for a dimensionally similar s.i. engine at the same RPM. Since the brake mean effective pressure of a diesel is also lower than that of a gasoline engine, friction accounts for 15 to 16% of indicated maximum power even at 2000 RPM. Typical bmep values for a naturally aspirated c.i. engine range from 6.5 to 7.5 bar.

5. INCREASING THE EFFICIENCY OF SPARK IGNITION ENGINES

5.1 Design Parameters

Engine valvetrain design is a widely used method to classify spark ignition engines. The first spark

ignition engines were of the side-valve type, but such engines have not been used in automobiles for several decades, although some engines used in off-highway applications, such as lawn mowers or forklifts, continue to use this design. The overhead valve (OHV) design supplanted the side-valve engine by the early 1950s, and continues to be used in many U.S. engines in much improved form. The overhead cam engine (OHC) is the dominant design used in the rest of the developed world. The placement of the camshaft in the cylinder heads allows the use of simple, lighter valvetrain, and valves can be opened and closed more quickly as a result of the reduced inertia. This permits better flow of intake and exhaust gases, especially at high RPM, with the result that an OHC design typically can produce greater power at high RPM than an OHV design of the same displacement.

A more sophisticated version of the OHC engine is the double overhead cam (DOHC) engine where two separate camshafts are used to activate the intake and exhaust valves, respectively. The DOHC design permits a very light valvetrain as the camshaft can actuate the valves directly without any intervening mechanical linkages. The DOHC design also allows some layout simplification, especially in engines that feature two intake valves and two exhaust valves (4-valve). The 4-valve engine has become popular since the mid-1980s and Japanese manufacturers, in particular, have embraced the DOHC 4-valve design. The DOHC design permits higher specific output than an OHC design, with the 4-valve DOHC design achieving the highest specific output, in excess of 70 BHP/liter of displacement.

5.2 Thermodynamic Efficiency

Increases in thermodynamic efficiency within the limitations of the Otto cycle are obviously possible by increasing the compression ratio. However, compression ratio is also fuel octane limited, and increases in compression ratio depend on how the characteristics of the combustion chamber and the timing of the spark can be tailored to prevent knock while maximizing efficiency.

Spark timing is associated with the delay in initiating and propagating combustion of the air-fuel mixture. To complete combustion before the piston starts its expansion stroke, the spark must be initiated a few crank angle degrees (advance) before the piston reaches top dead center. For a particular combustion chamber, compression ratio, and air-fuel mixture, there is an optimum level of spark advance for maximizing combustion chamber pressure and, hence, fuel efficiency. This level of spark advance is called MBT for maximum for best torque. However, MBT spark advance can result in knock if fuel octane is insufficient to resist preignition at the high pressures achieved with this timing. Hence, there is an interplay between spark timing and compression ratio in determining the onset of knock. Retarding timing from MBT reduces the tendency to knock but decreases fuel efficiency. Emissions of hydrocarbons and oxides of nitrogen (NO_x) are also dependent on spark timing and compression ratio, so that emission constrained engines require careful analysis of the knock, fuel efficiency, and emission trade-offs before the appropriate value of compression ratio and spark advance can be selected.

Electronic control of spark timing has made it possible to set spark timing closer to MBT relative to engines with mechanical controls. Due to production variability and inherent timing errors in a mechanical ignition timing system, the average value of timing in mechanically controlled engines had to be retarded significantly from the MBT timing. This protects the fraction of engines with higher than average advance due to production variability from knock. The use of electronic controls coupled with magnetic or optical sensors of crankshaft position has reduced the variability of timing between production engines and also allowed better control during transient engine operation. Engines have been equipped with knock sensors, which are essentially vibration sensors tuned to the frequency of knock. These sensors allow for advancing ignition timing to the point where trace knock occurs, so that timing is optimal for each engine produced regardless of production variability.

High-swirl, fast-burn combustion chambers have been developed to reduce the time taken for the air-fuel mixture to be fully combusted. The shorter the burn time, the more closely the cycle approximates the theoretical Otto cycle with constant volume combustion and the greater the thermodynamic efficiency. Reduction in burn time can be achieved by having a turbulent vortex within the combustion chamber that promotes flame propagation and mixing. The circular motion of the air-fuel mixture is known as swirl, and turbulence is also enhanced by shaping the piston so that gases near the cylinder wall are pushed rapidly towards the center in a motion known as squish. Improvements in flow visualization and computational fluid dynamics have allowed the optimization of intake valve, inlet port, and combustion chamber geometry to achieve desired flow characteristics. Typically, these designs

have resulted in a 2 to 3% improvement in thermodynamic efficiency and fuel economy. The high-swirl chambers also allow higher compression ratios and reduced spark advance at the same fuel octane number. The use of these types of combustion chambers has allowed compression ratio from about 8:1 in the early 1980s to 10:1 in the early 2000s, and further improvements are likely. In newer engines of the 4-valve DOHC type, the sparkplug is placed at the center of the combustion chamber, and the chamber can be made very compact by having a nearly hemispherical shape. Engines incorporating these designs have compression ratios of 10:1 while still allowing the use of regular 87 octane gasoline. Increases beyond 10:1 are expected to have diminishing benefits in efficiency and fuel economy and compression ratios beyond 12:1 are not likely to be beneficial unless fuel octane is raised simultaneously.

5.3 Reduction in Mechanical Friction

Mechanical friction losses are being reduced by converting sliding metal contacts to rolling contacts, reducing the weight of moving parts, reducing production tolerances to improve the fit between pistons and bore, and improving the lubrication between sliding or rolling parts.

Friction reduction has focused on the valvetrain, pistons, rings, crankshaft, crankpin bearings, and the oil pump. Valvetrain friction accounts for a larger fraction of total friction losses at low engine RPM than at high RPM. The sliding contract between the cam that activates the valve mechanism through a pushrod in an OHV design, or a rocker arm in an OHV design, can be substituted with a rolling contact by means of a roller cam follower. Roller cam followers have been found to reduce fuel consumption by 2 to 4% during city driving and 1 to 2% in highway driving. The use of lightweight valves made of ceramics or titanium is another possibility for the future. The lightweight valves reduce valve train inertia and also permit the use of lighter springs with lower tension. Titanium alloys are also being considered for valve springs which operate under heavy loads. There alloys have only half the shear modulus of steel and fewer coils are needed to obtain the same spring constant. A secondary benefit associated with lighter valves and springs is that the erratic valve motion at high RPM is reduced, allowing increased engine RPM range and power output.

The pistons and rings contribute to approximately half of total friction. The primary function of the rings is to minimize leakage of the air-fuel mixture from the combustion chamber to the crankcase, and oil leakage from the crankcase to the combustion chamber. The ring pack for most engines is composed of two compression rings and an oil ring. The rings have been shown to operate hydrodynamically over the cycle, but metal-to-metal contact occurs often at the top and bottom of the stroke. The outward radial force of the rings are as a result of installed ring tension and contribute to effective sealing as well as friction. Various low-tension ring designs were introduced in the 1980s, especially since the need to conform to axial diameter variations or bore distortions have been reduced by improved cylinder manufacturing techniques. Reduced tension rings have yielded friction reduction in the range of 5 to 10%, with fuel economy improvements of 1 to 2%. Elimination of one of the two compression rings has also been tried on some engines, and two-ring pistons may be the low friction concept for the future.

Pistons have also been redesigned to decrease friction. Prior to the 1980s, piston had large "skirts" to absorb side forces associated with side-to-side piston motion due to engine manufacturing inaccuracies. Pistons with reduced skirts diminish friction by having lower surface area in contact with the cylinder wall, but this effect is quite small. A larger effect is obtained for the mass reduction of a piston with smaller skirts and piston skirt size has seen continuous reduction since the 1980s. Reducing the reciprocating mass reduces the piston-to-bore loading. Secondary benefits include reduced engine weight and reduced vibration. Use of advanced materials also result in piston weight reduction. Lightweight pistons use hypereutectic aluminum alloys, while future pistons could use composite materials such as fiber-reinforced plastics. Advanced materials can also reduce the weight of the connecting rod, which also contributes to the side force on a piston.

The crankshaft bearings include the main bearings that support the crankshaft and the crankpin bearings and are of the journal bearing type. These bearings contribute to about 25% of total friction, while supporting the stresses transferred from the piston. The bearings run on a film of oil and detailed studies of lubrication requirements has led to optimization of bearing width and clearances to minimize engine friction. Studies on the use of roller bearings rather than journal bearings in this application has shown further reduction in friction is possible. Crankshaft roller bearings are used only in some two-stroke engines such as outboard motors for boat propulsion, but their durability in automotive applications has not been established.

Coatings of the piston and ring surfaces with materials to reduce wear also contribute to friction reduction. The top ring, for example, is normally coated with molybdenum, and new proprietary coating materials with lower friction are being introduced. Piston coatings of advanced high temperature plastics or resin are used widely and are claimed to reduce friction by 5% and fuel consumption by 1%.

The oil pumps generally used in most engines are of the gear pump type. Optimization of oil flow rates and reduction of the tolerances for the axial georotor clearance has led to improved efficiency, which translates to reduced drive power. Friction can be reduced by 2 to 3% with improved oil pump designs for a gain in fuel economy of about half of a percent.

Improvements to lubricants used in the engine also contribute to reduced friction and improved fuel economy. There is a relationship between oil viscosity, oil volatility, and engine oil consumption. Reduced viscosity oils traditionally resulted in increased oil consumption, but the development of viscosity index (VI) improvers had made it possible to tailor the viscosity with temperatures to formulate multigrade oils such as 10W-40 (these numbers refer to the range of viscosity covered by a multigrade oil). These multigrade oils act like low-viscosity oils during cold starts of the engine, reducing fuel consumption, but retain the lubricating properties of higher viscosity oils after the engine warms up to normal operating temperature. The development of 5W-20 oils and 5W-40 oils can contribute to a fuel economy improvement by further viscosity reduction. Friction modifiers containing molybdenum compounds have also reduced friction without affecting wear or oil consumption. Future synthetic oils combining reduced viscosity and friction modifiers can offer good wear protection, low oil consumption, and extended drain capability along with small improvements to fuel economy in the range of 1 to 3% over current oils.

5.4 Reduction in Pumping Loss

Reductions in flow pressure loss can be achieved by reducing the pressure drop that occurs in the flow of air (air-fuel mixture) into the cylinder and the combusted mixture through the exhaust system. However, the largest part of pumping loss during normal driving is due to throttling, and strategies to reduce throttling loss have included variable valve timing and lean-burn systems.

The pressure losses associated with the intake system and exhaust system have been typically defined in terms of volumetric efficiency, which is a ratio of the actual airflow through an engine to the airflow associated with filling the cylinder completely. The volumetric efficiency can be improved by making the intake airflow path as free of flow restrictions as possible through the air filters, intake manifolds, and valve ports. The shaping of valve ports to increase swirl in the combustion chamber can lead to reduced volumetric efficiency, leading to a trade-off between combustion and volumetric efficiency.

More important, the intake and exhaust processes are transient in nature as they occur only over approximately half a revolution of the crankshaft. The momentum effects of these flow oscillations can be exploited by keeping the valves open for durations greater than half a crankshaft revolution. During the intake stroke, the intake valve can be kept open beyond the end of the intake stroke, since the momentum of the intake flow results in a dynamic pressure that sustains the intake flow even when the piston begins the compression stroke. A similar effect is observed in the exhaust process, and the exhaust valve can be held open during the initial part of the intake stroke. These flow momentum effects depend on the velocity of the flow which is directly proportional to engine RPM. Increasing the valve opening duration helps volumetric efficiency at high RPM but hurts it at low RPM. Valve timing and overlap are selected to optimize the trade-off between high and low RPM performance characteristics.

Efficiency improvements can be realized by changing the valve overlap period to provide less overlap at idle and low engine speeds and greater overlap at high RPM. In DOHC engines, where separate crankshafts actuate the intake and exhaust valves, the valve overlap period can be changed by rotating the camshafts relative to each other. Such mechanisms have been commercialized engines show low RPM torque improvements of 7 to 10% with no sacrifice in maximum horsepower attained in the 5500 to 6000 RPM range. Variable valve overlap period is just one aspect of a more comprehensive variable valve timing system.

The oscillatory intake and exhaust flows can allow volumetric efficiency to be increased by exploiting resonance effects associated with pressure waves similar to those in organ pipes. The intake manifolds can be designed with pipe lengths that resonate, so that a high-pressure wave is generated at the intake valve as it is about to close, to cause a supercharging effect. Exhaust manifolds can be designed to resonate to achieve the opposite pressure effect to purge exhaust gases from this cylinder. For a given pipe

length, resonance occurs only at a certain specific frequency and its integer multiples so that, historically, tuned intake and exhaust manifolds could help performance only in certain narrow RPM ranges. The incorporation of a resonance tanks using the Helmholtz resonator principle in addition to tuned length intake pipes has led to improved intake manifold design that provide benefits over broader RPM ranges. Variable resonance systems have been introduced, where the intake tube lengths are changed at different RPM by opening and closing switching valves to realize smooth and high torque across virtually the entire engine speed range. Typically, the volumetric efficiency improvement is in the range of 4 to 5% over fixed resonance systems.

Another method to increase efficiency is by increasing valve area. A 2-valve design is limited in valve size by the need to accommodate the valves and sparkplugs in the circle defined by the cylinder base. The active flow area is defined by the product of valve circumference and lift. Increasing the number of valves is an obvious way to increase total valve area and flow area, and the 4-valve system, which increases flow area by 25 to 30% over 2-valve layouts, has gained broad acceptance. The valves can be arranged around the cylinder bore and the sparkplug placed in the center of the bore to improve combustion. Analysis of additional valve layout designs that take into account the minimum required clearance between valve seats and the sparkplug location suggest that fivevalve designs (3 intake, 2 exhaust) can provide an additional 20% increase in flow area, at the expense of increased valvetrain complexity. Additional valves do not provide further increases in flow area either due to noncentral plug locations or valve-to-valve interference.

Under most normal driving conditions, the throttling loss is the single largest contributor to reduction in engine efficiency. In s.i. engines, the air is throttled ahead of the intake manifold by means of a butterfly valve that is connected to the accelerator pedal. The vehicle's driver demands a power level by depressing or releasing the accelerator pedal, which in turn opens or closes the butterfly valve. The presence of the butterfly valve in the intake air stream creates a vacuum in the intake manifold at part throttle conditions, and the intake stroke draws in air at reduced pressure, resulting in pumping losses. These losses are proportional to the intake vacuum and disappear at wide open throttle.

Measures to reduce throttling loss are varied. The horsepower demand by the driver can be satisfied by any combination of torque and RPM since

$$\text{Power} = \text{Torque} \times \text{RPM}.$$

The higher the torque, the lower the RPM to satisfy a given power demand. Higher torque implies less throttling, and the lower RPM also reduces friction loss so that the optimum theoretical fuel efficiency at a given level of horsepower demand occurs at the highest torque level the engine is capable of. In practice, the highest level is never chosen because of the need to maintain a large reserve of torque for immediate acceleration and also because engine vibrations are a problem at low RPM, especially near or below engine speeds referred to as lugging RPM. Nevertheless, this simple concept can be exploited to the maximum by using a small displacement high specific output engine in combination with a multispeed transmission with five or more forward gears. The larger number of gears allows selection of the highest torque/lowest RPM combination for fuel economy at any speed and load, while maintaining sufficient reserve torque for instantaneous changes in power demand. A specific torque increase of 10% can be utilized to provide a fuel economy benefit of 3 to 3.5% if the engine is downsized by 8 to 10%. In light vehicles, the number of forward gears has been increasing from 3 to 5, and six-speed transmissions are likely to be standard in the future. The continuously variable transmission is other development that allow continuous change of gear ratios over a specific range.

Lean-burn is another method to reduce pumping loss. Rather than throttling the air, the fuel flow is reduced so that the air-fuel ratio increases, or becomes leaner. (In this context, the c.i. engine is a lean-burn engine.) Most s.i. engines, however, do not run well at air: fuel ratios leaner than 18:1, as the combustion quality deteriorates under lean con;ditions. Engines constructed with high swirl and turbulence in the intake charge can run well at air: fuel ratios up to 21:1. In a vehicle, lean-burn engines are calibrated lean only at light loads to reduce throttling loss, but run at stoichiometric or rich air: fuel ratios at high loads to maximize power. The excess air combustion at light loads has the added advantage of having a favorable effect on the polytropic coefficient, n, in the efficiency equation. Modern lean-burn engines do not eliminate throttling loss, but the reduction is sufficient to improve vehicle fuel economy by 8 to 10%. The disadvantage of lean burn is that such engines cannot yet use catalytic controls to reduce emissions of oxide of nitrogen (NO_x), and the in-cylinder NO_x emission

control from running lean is sometimes insufficient to meet stringent NO_x emissions standards. However, there are developments in lean NO_x catalysts that could allow lean-burn engines to meet the most stringent NO_x standards proposed in the future.

Another type of lean-burn s.i. engine is the stratified charge engine. Research is focused on direct-injection stratified charge (DISC) engines where the fuel is sprayed into the combustion chamber, rather than into or ahead of the intake valve. Typically, this enables the air: fuel ratio to vary axially or radially in the cylinder, with the richest air: fuel ratios present near the sparkplug or at the top of the cylinder. Stratification requires very careful design of the combustion chamber shape and intake swirl, as well as of the fuel injection system. Advanced direct injection systems have been able to maintain stable combustion at total air: fuel ratios as high as 40:1. Such engines have been commercialized in 2000 in Europe and Japan.

Variable valve timing is another method to reduce throttling loss. By closing the intake valve early, the intake process occurs over a smaller fraction of the cycle, resulting in a lower vacuum in the intake manifold. It is possible to completely eliminate the butterfly valve that throttles air and achieve all part load settings by varying the intake valve opening duration. However, at very light load, the intake valve is open for a very short duration, and this leads to weaker in-cylinder gas motion and reduced combustion stability. At high RPM, the throttling loss benefits are not realized fully. Throttling occurs at the valve when the valve closing time increases relative to the intake stroke duration at high speeds, due to the valvetrain inertia. Hence, throttling losses can be decreased by 80% at light load, low RPM conditions, but by only 40 to 50% at high RPM, even with fully variable valve timing.

Variable valve timing can also provide a number of other benefits, such as reduced valve overlap at light loads/low speeds (discussed earlier) and maximized output over the entire range of engine RPM. Fully variable valve timing can result in engine output levels of up to 100 BHP/liter at high RPM with little or no effect on low-speed torque. In comparison to an engine with fixed valve timing that offers equal performance, fuel efficiency improvements of 7 to 10% are possible. The principal drawback has historically been the lack of a durable and low-cost mechanism to implement valve timing changes. A number of new systems have been introduced that are ingenious mechanisms with the required durability.

6. INCREASING THE EFFICIENCY OF COMPRESSION IGNITION (DIESEL) ENGINES

Compression ignition engines, commonly referred to as diesel engines, are in widespread use. Most c.i. engines in light-duty vehicle applications are of the indirect injection type (IDI), while most c.i. engines in heavy-duty vehicles are of the direct injection type. In comparison to s.i. engines, c.i. engines operate at much lower brake mean effective pressures of (typically) about 7 to 8 bar at full load. Maximum power output of a c.i. engine is limited by the rate of mixing between the injected fuel spray and hot air. At high fueling levels, inadequate mixing leads to high black smoke, and the maximum horsepower is usually smoke limited for most c.i. engines. Naturally aspirated diesel engines for light-duty vehicle use have specific power outputs of 25 to 35 BHP per liter, which is about half the specific output of a modern s.i. engine. However, fuel consumption is significantly better, and c.i. engines are preferred over s.i. engines where fuel economy is important.

Due to the combustion process, as well as the high internal friction of a c.i. engine, maximum speed is typically limited to less than 4500 RPM, which partially explains the lower specific output of c.i. engine. In light-duty vehicle use, an IDI engine can display between 20 to 40% better fuel economy depending on whether the comparison is based on engines of equal displacement or of equal power output in the same RPM range. The improvement is largely due to the superior part load efficiency of the c.i. engine, as there is no throttling loss. At high vehicle speeds (>120 km/hr), the higher internal friction of the c.i. engine offsets the reduced throttling loss, and the fuel efficiency difference between s.i. and c.i. engine narrows considerably.

Most of the evolutionary improvements for compression ignition engines in friction and pumping loss reduction are conceptually similar to those described for s.i. engines, and this section focuses on the unique aspects of c.i. engine improvements.

6.1 Design Parameters

Note that c.i. engines have also adopted some of the same valvetrain designs as those found in s.i. engines. While most c.i. engines were of the OHV type, European c.i. engines for passenger car use are of the OHC type. The c.i. engine is not normally run at high RPM, so that the difference in specific output

between an OHV and an OHC design is small. The OHC design does permit a simpler and lighter cylinder block casting, which is beneficial for overcoming some of the inherent weight liabilities. OHC designs also permits the camshaft to directly activate the fuel injector in unit injector designs, which are capable of high injection pressure and fine atomization of the fuel spray. Four-valve OHC or DOHC designs allow central placement of the fuel injector in the cylinder, which enhances uniform mixing of air and fuel.

6.2 Thermodynamic Efficiency

The peak efficiency of an IDI engine is comparable to or only slightly better than the peak efficiency of an s.i. engine, based on average values for engines in production. The contrast between theoretical and actual efficiency is notable, and part of the reason is that the prechamber in the IDI diesel is a source of energy loss. The design of the prechamber is optimized to promote swirl and mixing of the fuel spray with air, but the prechamber increases total combustion time. Its extra surface area also results in more heat transfer into the cylinder head.

Direct injection (DI) systems avoid the heat and flow losses from the prechamber by injecting the fuel into the combustion chamber. The combustion process in DI diesels consists of two phases. The first phase consists of an ignition delay period followed by spontaneous ignition of the fuel droplets. The second phase is characterized by diffusion burning of the droplets. The fuel injection system must be capable of injecting very little fuel during the first phase and provide highly atomized fuel and promote intensive mixing during the second. Historically, the mixing process has been aided by creating high swirl in the combustion chamber to promote turbulence. However, high swirl and turbulence also lead to flow losses and heat losses, thus reducing efficiency. The newest concept is the quiescent chamber where all of the mixing is achieved by injecting fuel at very high pressures to promote fine atomization and complete penetration of the air in the combustion chamber. New fuel injection systems using unit injectors can achieve pressures in excess of 1500 bar, twice as high as injection pressures utilized previously. Quiescent combustion chamber designs with high-pressure fuel injection systems have provided to be very fuel efficient and are coming into widespread use in heavy-duty truck engines. These systems have the added advantage of reducing particulate and smoke emissions.

DI engines have entered the light-duty vehicle market, but these engines still utilize swirl type combustion chambers. In combination with turbocharging (see Section 6), the new DI engines have attained peak efficiencies of over 41%. Fuel economy improvements in the composite cycle relative to IDI engines are in the 12 to 15% range, and are up to 40% higher than naturally aspirated s.i. engines with similar torque characteristics. It is not clear if quiescent combustion chambers will be ever used in DI engines for cars, since the size of the chamber is quite small and fuel impingement on cylinder walls is a concern.

Although the efficiency equation shows that increasing compression ratio has a positive effect on efficiency, practical limitations preclude any significant efficiency gain through this method. At high compression ratios, the size of the combustion chamber is reduced, and the regions of dead air trapped between the cylinder and piston edges and crevices became relatively large, leading to poor air utilization, reduced specific output, and, potentially, higher smoke. Moreover, the stresses on the engine increase with increasing compression ratio, making the engine heavy and bulky. The compression ratios are already somewhat higher than optimal to provide enough heat of compression so that a cold start at low ambient temperature is possible.

6.3 Friction and Pumping Loss

Most of the friction reducing technologies that can be adopted in s.i. engines are conceptually similar to those that can be adopted for diesels. There are limitations to the extent of reduction of ring tension and piston size due to the high compression ratio of c.i. engines, but roller cam followers, optimized crankshaft bearings, and multigrade lubricants have also been adopted for c.i. engine use. Since friction is a larger fraction of total loss, a 10% reduction in fraction in a c.i. engine can lead to a 3 to 4% improvement in fuel economy.

Pumping losses are not as significant a contributor to overall energy loss in a c.i. engine, but tuned intake manifolds and improved valve port shapes and valve designs have also improved volumetric efficiency of modern c.i. engines. Four-valve designs, in widespread use in the heavy truck market, have appeared in passenger cars, but their benefits are smaller in c.i. engine use due to the low maximum RPM relative to s.i. engines. Nevertheless, the 4-valve head with a centrally mounted injector is particularly useful in DI engines since it allows for

symmetry in the fuel spray with resultant good air utilization.

Variable valve timing or any form of valve control holds little benefit for c.i. engines due to the lack of throttling loss, and lack of high RPM performance. Valve timing can be varied to reduce the effective compression ratio, so that a very high ratio can be used for cold starts, but a lower, more optimal, ratio for fully warmed up operation.

7. INTAKE CHARGE BOOSTING

Most c.i. and s.i. engines for light vehicle use intake air at atmospheric pressure. One method to increase maximum power at wide open throttle is to increase the density of air supplied to the intake by precompression. This permits a smaller displacement engine to be substituted without loss of power and acceleration performance. The use of a smaller displacement engine reduces pumping loss at part load and friction loss. However, intake charge compression has its own drawbacks. Its effect is similar to raising the compression ratio in terms of peak cylinder pressure, but maximum cylinder pressure is limited in s.i. engines by the fuel octane. Charge boosted engines generally require premium gasolines with higher octane number if the charge boost levels are more than 0.2 to 0.3 bar over atmospheric in vehicles for street use. Racing cars use boost levels up to 1.5 bar in conjunction with a very high octane fuel such as methanol. This limitation is not present in a c.i. engine, and charge boosting is much more common in c.i. engine applications. Most c.i. engines in heavy-duty truck applications use charge boosting.

Intake charge boosting is normally achieved by the use of turbochargers or superchargers. Turbochargers recover the wasted heat and pressure in the exhaust through a turbine, which in turn drives a compressor to boost intake pressure. Superchargers are generally driven by the engine itself and are theoretically less efficient than a turbocharger. Many engines that use either device also utilize an aftercooler that cools the compressed air as it exits from the supercharger or turbocharger before it enters the s.i. engine. The aftercooler increases engine specific power output by providing the engine with a denser intake charge, and the lower temperature also helps in preventing detonation, or knock. Charge boosting is useful only under wide open throttle conditions in s.i. engines, which occur rarely in normal driving, so that such devices are usually used in high-performance vehicles. In c.i. engines, charge boosting is effective at all speeds and levels.

7.1 Turbochargers

Turbochargers in automotive applications are of the radial flow turbine type. The turbine extracts pressure energy from the exhaust stream and drives a compressor that increases the pressure of the intake air. A number of issues affect the performance of turbomachinery, some of which are a result of natural laws governing the interrelationship between pressure, airflow, and turbocharger speed. Turbochargers do not function at light load because there is very little energy in the exhaust stream. At high load, the turbocharger's ability to provide boost is a nonlinear function of exhaust flow. At low engine speed and high load, the turbocharger provides little boost, but boost increases rapidly beyond a certain flow rate that is dependent on the turbocharger size. The turbocharger also has a maximum flow rate, and the matching of a turbocharger's flow characteristics to a piston engine's flow requirements involves a number of trade-offs. If the turbocharger is sized to provide adequate charge boost at moderate engine speeds, high RPM boost is limited and there is a sacrifice in maximum power. A larger turbocharger capable of maximizing power at high RPM sacrifices the ability to provide boost at normal driving conditions. At very low RPM (for example, when accelerating from a stopped condition), no practical design provides boost immediately. Moreover, the addition of turbocharger requires the engine compression ratio to be decreased by 1.5 to 2 points (or 1 to 1.5 with an aftercooler) to prevent detonation. The net result is that turbocharged engines have *lower* brake specific fuel efficiencies than engines of equal size, but can provide some efficiency benefit when compared to engines of equal mid range or top end power. During sudden acceleration, the turbocharger does not provide boost instantaneously due to its inertia, and turbocharged vehicles can have noticeably different acceleration characteristics than naturally aspirated vehicles. New variable geometry turbochargers have improved response and better boost characteristics over the operating range.

Turbochargers are much better suited to c.i. engines since these engines are unthrottled and the combustion process is not knock limited. Airflow at a given engine load/speed setting is always higher for a c.i. engine relative to an s.i. engine, and this provides a less restricted operating regime for the turbocharger. The lack of a knock limit also allows increased

boost and removes the need to cap boost pressure under most operating conditions. Turbocharged c.i. engines offer up to 50% higher specific power and torque, and about 10% better fuel economy than naturally aspirated c.i. engines of approximately equal torque capability.

7.2 Superchargers

Most s.i. engine superchargers are driven off the crankshaft and are of the Roots blower or positive displacement pump type. In comparison to turbochargers, these superchargers are bulky and weigh considerably more. In addition, the superchargers are driven off the crankshaft, absorbing 3 to 5% of the engine power output depending on pressure boost and engine speed.

The supercharger, however, does not have the low RPM boost problems associated with turbochargers and can be designed to nearly eliminate any time lag in delivering the full boost level. As a result, superchargers are more acceptable to consumers from a driveability viewpoint. The need to reduce engine compression ratio and the supercharger's drive power requirement detract from overall efficiency. In automotive applications, a supercharged engine can replace a naturally aspirated engine that is 30 to 35% larger in displacement, with a net pumping loss reduction. Overall, fuel economy improves by about 8% or less, if the added weight effects are included.

Superchargers are less efficient in combination with c.i. engines, since these engines run lean even at full load, and the power required for compressing air is proportionally greater. Supercharged c.i. engines are not yet commercially available, since the turbocharger appears far more suitable in these application.

8. ALTERNATIVE HEAT ENGINES

A number of alternative engines types have been researched for use in passenger cars but have not yet proved successful in the market place. A brief discussion of the suitability of four engines for automotive power plants is provided next.

The Wankel engine is the most successful of the four engines in that it has been in commercial production in limited volume since the 1970s. The thermodynamic cycle is identical to that of a four-stroke engine, but the engine does not use a reciprocating piston in a cylinder. Rather a triangular rotor spins eccentrically inside a Fig. 8–shaped casing. The volume trapped between the two rotor edges and the casing varies with rotor position, so that the intake, comparison, expansion, and exhaust stroke occur as the rotor spins through one revolution. The engine is very compact relative to a piston s.i. engine of equal power, and the lack of reciprocating parts provides smooth operation. However, the friction associated with the rotor seals is high, and the engine also suffers from more heat losses than an s.i. engine. For these reasons, the Wankel engine's efficiency has always been below that of a modern s.i. piston engine.

The two-stroke engine is widely used in small motorcycles but was thought to be too inefficient and polluting for use in passenger cars. One development is the use of direct injection stratified charge (DISC) combustion with this type of engine. One of the major problems with the two-stroke engine is that the intake stroke overlaps with the exhaust stroke resulting in some intake mixture passing uncombusted into the exhaust. The use of a DISC design avoids this problem since only air is inducted during intake. Advanced fuel injection systems have been developed to provide a finely atomized mist of fuel just prior to spark initiation and to sustain combustion at light loads. The two-stroke engines of this type are thermodynamically less efficient than four-stroke DISC engines, but the internal friction loss and weight of two-stroke engine is much lower than a four-stroke engine of equal power. As a result, the engine may provide fuel economy equal or superior to that of a DISC (four-stroke) engine when installed in a vehicle. Experimental prototypes have achieved good results, but the durability and emissions performance of advanced two-stroke engines is still not established.

Gas turbine engines are widely used to power aircraft, and considerable research has been completed to assess its use in automobiles. Such engines use continuous combustion of fuel, which holds the potential for low emissions and multifuel capability. The efficiency of the engine is directly proportional to the combustion temperature of the fuel, which has been constrained to 1200°C by the metals used to fabricate turbine blades. The use of high-temperature ceramic materials for turbine blades coupled with the use of regenerative exhaust waste heat recovery were expected to increase the efficiency of gas turbine engines to levels significantly higher than the efficiency of s.i. engines.

In reality, such goals have not yet been attained partly because the gas turbine components become less aerodynamically efficient at the small engine

sizes suitable for passenger car use. Part load efficiency is a major problem for gas turbines due to the nonlinear efficiency changes with airflow rates in turbomachinery. In addition, the inertia of the gas turbine makes it poorly suited to vehicle applications, where speed and load fluctuations are rapid in city driving. As a result, there is little optimism that the gas turbine powered vehicle will be a reality in the foreseeable future.

Stirling engines have held a particular fascination for researchers since the cycle closely approximates the ideal Carnot cycle, which extracts the maximum amount of work theoretically possible from a heat source. This engine is also a continuous combustion engine like the gas turbine engine. While the engine uses a piston to convert heat energy to work, the working fluid is enclosed and heat is conducted in and out of the working fluid by heat exchangers. To maximize efficiency, the working fluid is a gas of low molecular weight like hydrogen or helium. Prototype designs of the Stirling engine have not yet attained efficiency goals and have had other problems, such as the containment of the working fluid. The Stirling engine is, like the gas turbine, not well suited to applications where the load and speed change rapidly, and much of the interest in this engine has faded.

SEE ALSO THE FOLLOWING ARTICLES

Alternative Transportation Fuels: Contemporary Case Studies • Combustion and Thermochemistry • Fuel Cycle Analysis of Conventional and Alternative Fuel Vehicles • Fuel Economy Initiatives: International Comparisons • Hybrid Electric Vehicles • Internal Combustion (Gasoline and Diesel) Engines • Transportation Fuel Alternatives for Highway Vehicles • Vehicles and Their Powerplants: Energy Use and Efficiency

Further Reading

Amman, C. (1989). "The Automotive Engine—A Future Perspective." GM Research Publication GMR-6653.

Lichty, C. (1967). "Combustion Engine Processes." John Wiley & Sons.

National Academy of Sciences (2001). "Effectiveness and Impact of Corporate Average Fuel Economy Standards." National Academy Press.

Office of Technology Assessment (1995). "Advanced Automotive Technology—Visions of a Super—Efficient Family Car." Report to the U.S. Congress, OTA-ETI-638.

Weiss, M.A.(Ed.) (2000). "On the Road in 2020." Massachusetts Institute of Technology Report MIT-EC00-003.

Internal Combustion (Gasoline and Diesel) Engines

ROBERT N. BRADY
HiTech Consulting Ltd
Burnaby, British Columbia, Canada

1. **Introduction to Basic Force/Motion Laws**
2. **Constant Volume versus Constant Pressure Cycles**
3. **Pressure–Volume Curve**
4. **Two-Stroke Cycle Gasoline Engine: Basic Operation**
5. **Four-Stroke Cycle Gasoline Engine: Basic Operation**
6. **Four-Stroke Cycle Polar Valve Timing**
7. **Two-Stroke Cycle Diesel Engine: Basic Operation**
8. **Two-Stroke Cycle Diesel Polar Valve Timing**
9. **Four-Stroke Cycle Diesel Engine: Basic Operation**
10. **Direct versus Indirect Diesel Engine Injection Concepts**
11. **Review of Two-Stroke versus Four-Stroke Cycles**
12. **Comparison of Two-Cycle versus Four-Cycle Engines**
13. **Horsepower Description**
14. **Torque Description**
15. **Diesel/Gasoline Engine Comparison**
16. **Summary: Gasoline versus Diesel Power**

Glossary

blower A gear driven air pump usually consisting of two, two-lobe, or three-lobe rotors rotating within an aluminum housing to supply air flow to the engine cylinders at low pressure; it is typically used on two-stroke cycle heavy-duty diesel engines.

compression ratio A term used to describe the cylinder and combustion chamber swept volume when the piston is at bottom dead center (BDC) versus that when the piston is at top dead center (TDC); although the term compression ratio is used, in reality it is a volumetric ratio and not a pressure ratio.

naturally aspirated An non-supercharged internal combustion engine that relies solely on valve timing and piston movement to inhale or breathe the air/fuel mixture into the cylinder on a gasoline engine or the atmospheric air only on a diesel engine.

supercharged An engine that, through a combination of a gear-driven blower/pump and valve timing, contains a cylinder pressure greater than atmospheric at the start of the compression stroke; the degree of pressure retained determines the level of supercharging.

turbocharger An exhaust gas-driven device consisting of a shaft with two vaned fan-type wheels at each end; at the hot end, the turbine fan/wheel is driven by the hot pressurized exhaust gases, whereas at the opposite end, the compressor fan/wheel pressurizes the ambient air supply into the engine intake manifold.

In dealing with internal combustion (IC), reciprocating (back and forward motion) gasoline and diesel engine operating cycle theory, one needs to consider the basic laws of energy because the forces developed within the engine cylinder permit work to be developed. Both gasoline and diesel engines operate on the principle of IC, where chemical energy releases heat from the burning of gasoline in the former and diesel fuel in the latter. The heat is then converted into mechanical work.

1. INTRODUCTION TO BASIC FORCE/MOTION LAWS

In dealing with internal combustion (IC), reciprocating (back and forward motion) gasoline and diesel engine operating cycle theory, one needs to consider the basic laws of energy because the forces developed within the engine cylinder permit work to be developed. During the years 1665 and 1666, Isaac Newton developed three laws that describe all of the states of motion, rest, constant motion, and accelerated motion. These three laws explain how forces cause all of the states of motion:

1. An object at rest will remain at rest, and an object in motion will remain in motion at

 © 2004 Elsevier Inc. All rights reserved.

constant velocity, unless acted on by an unbalanced force.
2. Newton's second law of motion shows how force, mass, and acceleration are related, where Force = Mass × Acceleration.
3. The third law of motion states that for every action, there is an equal and opposite reaction.

The gasoline engine has been in use for well over a century now and was patented by the German technician Nikolaus August Otto in 1876; hence, the term that is sometimes applied to this type of engine is the "Otto cycle." The sequence of events is based on four individual piston strokes, two up and two down, consisting of intake, compression, power, and exhaust, with one power stroke occurring every 720° or two full engine crankshaft rotations.

The diesel engine concept was named after the French-born German engineer Rudolf Christian Karl Diesel. Diesel's concept was also designed to operate on the four-stroke cycle principle, and a patent was granted to him in 1892. However, it was not until 1895, on his third design attempt, that he successfully proved his theory of compression ignition operation.

Both gasoline and diesel engines operate on the principle of IC, where chemical energy releases heat from the burning of gasoline in the former and diesel fuel in the latter. The heat is then converted into mechanical work. The combustion chamber is formed between the crown (top) of the piston and the underside of the cylinder head. The term "thermodynamic cycle" is often attached to the operation of the gas and diesel engine because the word "thermodynamic" describes the science concerned with the conversion of heat to mechanical energy, the control of temperature, and the like. "Thermo" refers to heat, whereas "dynamic" refers to motion or being full of force and energy. Because gas and diesel engine pistons reciprocate, the combustion heat is transferred from the piston, through a connecting rod attached to the piston at one end and the crankshaft of the engine at the opposite end, to create mechanical rotary motion. Figure 1 illustrates the major components of a four-stroke cycle, heavy-duty, electronically controlled, overhead camshaft diesel engine. A four-stroke cycle, electronically controlled, gasoline fuel-injected engine is very similar in its arrangement other than the fact that a spark plug is used. To create power, all IC chamber engines require the following:

- Air (oxygen) for combustion
- Fuel (something to burn)
- Ignition (a means of starting the fire)

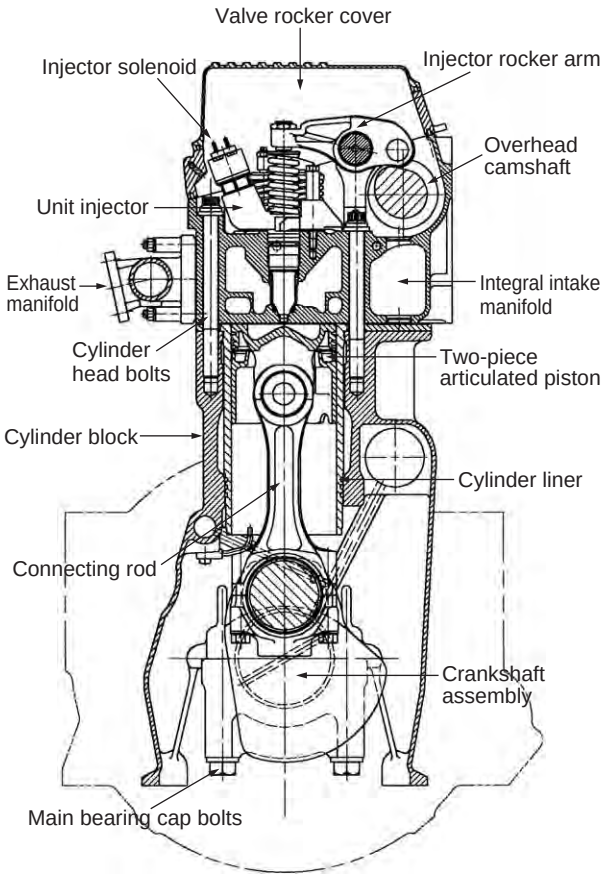


FIGURE 1 Design features of a 3406E Caterpillar heavy-duty truck engine that employs a turbocharger, air-to-air charge cooling (AACC), an overhead camshaft, and electronically controlled unit injectors and governing. Courtesy of Caterpillar Inc.

2. CONSTANT VOLUME VERSUS CONSTANT PRESSURE CYCLES

There are two major differences between a gasoline engine and a diesel engine. First, the gasoline engine operates on a theoretical air standard cycle commonly referred to as being a "constant volume cycle," where periodic combustion and generation of work take place. This term more fully indicates that combustion of the air/fuel charge is completed at the instant the piston is at top dead center (TDC) on its compression stroke, where pressure increases but the volume above the piston crown remains constant. Combustion requires a high-tension spark from an electrical/electronic ignition system distributed to a spark plug to initiate a flame front. Combustion time is primarily a function of the speed with which the flame propagates, typically in the range of 20 to 40 ms. In

most engines, the air/fuel mixture is formed outside the combustion chamber. Therefore, it forms a largely homogenous (composed of parts of the same kind) charge at the time combustion is initiated. In gasoline direct injection (GDI) engines, the air/fuel mixture formation is introduced directly into the combustion chamber and is referred to as being heterogeneous (composed of parts of different kinds).

Today's automotive gasoline engines operate in the "stoichiometric" air/fuel ratio range. Simply put, it is the ratio (usually of mass) between air and flammable gas or vapor at which complete combustion or chemical combination takes place. Approximately 14.5 kg of air is required for complete combustion of 1 kg of gasoline. Therefore, one can say that the air/fuel mixture is approximately 14.5:1.0. This ratio permits the three-way catalytic converter to treat raw exhaust emissions with maximum effectiveness as per the U.S. Environmental Protection Agency (EPA) standards. This is achieved by the use of electronic engine controls/sensors in harmony with one or more exhaust gas oxygen feedback sensors to constantly advise the engine electronic control module (ECM) as to how far from stoichiometric after combustion the air/fuel ratio is at a given time. The ECM then varies the injector solenoid on/off time (duty cycle) either to lean out or to enrich the air/fuel mixture so as to maintain the desired stoichiometric setting.

Second, unlike the gasoline engine, the diesel engine does not operate on a stoichiometric air/fuel ratio. In a diesel, the air/fuel ratio is extremely lean at an idle speed (minimum fuel delivery) because the air entering the cylinders is unthrottled (not restricted as it is on a typical gasoline engine). At higher loads/speeds, the diesel air/fuel ratio will run richer because more fuel is injected to produce more power. Therefore, at an idle speed, the diesel engine can exhibit air/fuel ratios as high as 90:1 or 100:1, with some engines running even leaner. At full-load/high-speed operation, the air/fuel ratio can drop down as low as 25:1 or 30:1. In addition, a diesel engine operates on what is referred to as a theoretical "constant pressure cycle," where fuel is presumed to be supplied as the piston moves down the cylinder on its power stroke at a rate such that the cylinder pressure remains constant during the combustion process. The diesel is also said to operate with a heterogeneous charge of compressed air produced during the piston's upward-moving compression stroke, supported by a finely atomized spray of high-pressure liquid fuel injected prior to the piston reaching TDC. The generated heat created in the trapped air from the upward-moving piston compression stroke (hot high-pressure air only) causes the injected diesel fuel to vaporize. After a short time (ignition) delay, where the air and injected fuel mix, the self-ignition properties of the air/fuel mixture initiate combustion to establish a flame front within the combustion chamber. Therefore, additional fuel that is being injected has no ignition delay but burns instantly. The higher compression ratio used with current design, high-speed, heavy-duty diesel engines can develop peak cylinder pressures between 1800 and 2300 psi (12,411–15,858 kPa) and peak temperatures of 3500°F (1927°C). These high-pressure waves and their speed of propagation throughout the combustion chamber represent one of the items that creates the inherent noise attributable to a diesel engine when running.

(In reality, no IC engine, whether gasoline or diesel, operates on either the constant pressure or the constant volume combustion phase. Each requires a few degrees of crankshaft rotation for completion of combustion, and there is also a rise in cylinder pressure during the combustion process. Therefore, a dual cycle of the two that falls somewhere between the Otto and Diesel curves would more closely represent the theoretical curve for consideration with both the gasoline and diesel IC engine cycles. Because IC engines do not operate according to ideal cycles but rather operate with real gas during combustion, they are characterized by flow and pumping losses, thermodynamic losses, and mechanical losses due to friction.)

3. PRESSURE–VOLUME CURVE

To clarify the constant volume and constant pressure cycles described in the previous section, Fig. 2 illustrates what actually transpires during the two most important strokes within a four-stroke cycle engine, namely during the compression and the power strokes. A four-stroke cycle diesel has been chosen for this description, but the curve would be somewhat similar for a gasoline engine. This pressure–volume (PV) diagram represents the piston from a position corresponding to 90° before top dead center (BTDC) as it moves up the cylinder on its compression stroke to 90° after top dead center (ATDC) on its power stroke.

4. TWO-STROKE CYCLE GASOLINE ENGINE: BASIC OPERATION

The term "two-stroke cycle" is based on the fact that the piston moves through two individual strokes (one

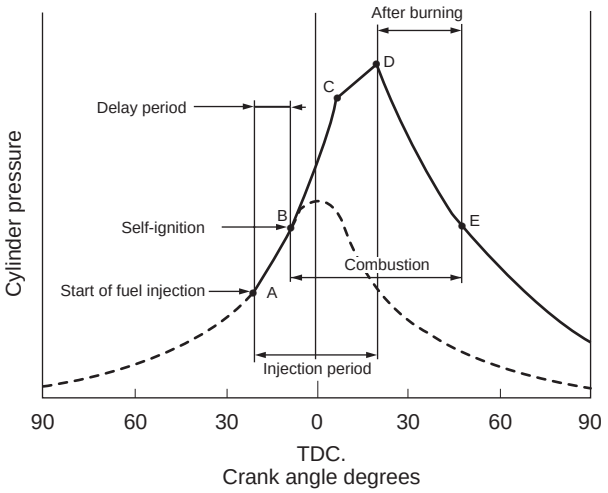


FIGURE 2 Graph illustrating the pressure–volume curve diesel engine combustion operating principle. Courtesy of ZEXEL USA, Technology and Information Division.

up and one down) to provide one power stroke for every 360° of crankshaft rotation. Two-cycle gasoline engines have been in use for many years in motorcycles, lawnmowers, outboard motors, and small industrial applications due to their simplicity of design, construction, lightness, and inexpensive mass production techniques. Although many models are air cooled, there are some that can be liquid cooled. However, these engines tend to emit a burning smell from the fuel/oil mix, are noisy, are fuel thirsty, emit higher exhaust emissions, and are more prone to failure compared with their four-stroke cycle counterparts.

The actions of both intake and exhaust occur between the compression and power strokes. In Fig. 3A, the power stroke occurs once the spark plug fires, the air/fuel mixture is ignited, and an immediate flame front is created. The heat and pressure created from the chemical energy of the burning air/fuel mixture forces the piston down the cylinder. The chemical energy is converted into mechanical energy by both the piston and the connecting rod to rotate the engine crankshaft.

In Fig. 3B, as the piston is driven down the cylinder, the specially shaped piston crown uncovers the exhaust port shown on the left-hand side and the hot burned gases start to escape from the combustion chamber. On the right-hand side of the diagram, as the piston continues down, the intake port is opened to permit a new air/fuel mixture charge from the engine crankcase to enter the cylinder. In Fig. 3C, a small reed valve opens to permit the air/fuel mixture to enter the crankcase. This occurs on the piston upstroke (compression) due to a slight vacuum created in the crankcase. Atmospheric pressure through the carburetion system causes the air/fuel mixture to flow through the reed valve and enter the crankcase. Also shown in Fig. 3C, as the piston begins its upstroke and both cylinder ports are covered, the air/fuel mixture from the sequence in Fig. 3B is compressed into the combustion chamber. When the piston reaches TDC, the spark plug ignites the mixture and the power stroke is repeated.

5. FOUR-STROKE CYCLE GASOLINE ENGINE: BASIC OPERATION

Four-stroke cycle gasoline and diesel engines predominate as the global prime mover of choice. In gasoline engines, the introduction of electronic controls and GDI systems, in conjunction with dual overhead camshafts, distributorless ignition systems, variable geometry turbochargers, intercoolers, four-valve cylinder head designs, three-way exhaust catalytic converters, and the like, has greatly improved not only their thermal efficiency (fuel economy) but also the control of exhaust gas emissions.

In the four-stroke cycle, the piston moves through four individual strokes (two up and two down). The two upward-moving piston strokes are the compression and exhaust strokes, whereas the downward-moving piston strokes are the intake and power strokes within each cylinder. The individual strokes of the piston shown in Fig. 4 include the following:

- *The intake stroke where the piston moves down the cylinder.* The intake valve(s) is held open by a rotating camshaft to permit an air/fuel mixture to enter the cylinder. This stroke can typically range between 240 and 260° obtained by opening the intake valve BTDC and then closing the valve ABDC.
- The compression stroke where the piston moves up the cylinder to pressurize the air/fuel mixture while the intake and exhaust valves are held closed by valve springs.
- *The power stroke where the air/fuel mixture is ignited by a high-tension spark from the spark plug.* The length or duration of the power stroke is controlled by opening the exhaust valve(s) before bottom dead center (BBDC). Typical duration of the power stroke averages approximately 140° of crankshaft rotation. The number of engine cylinders will

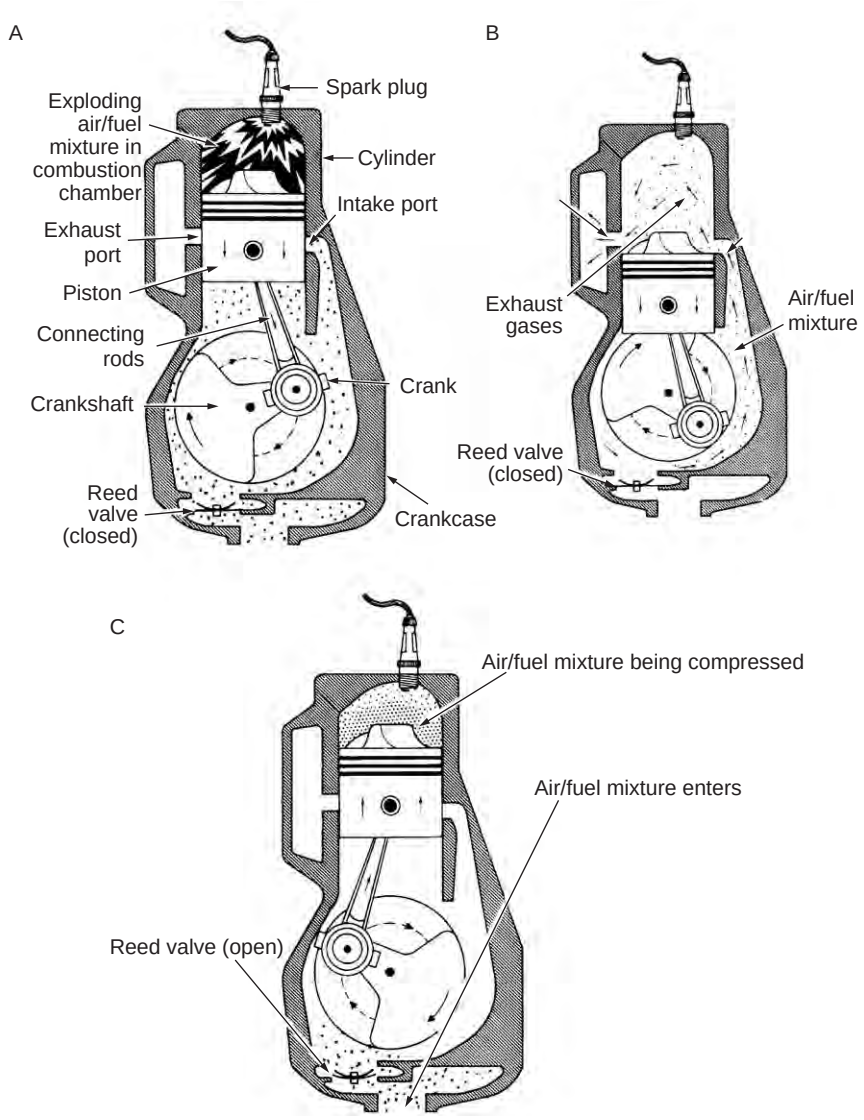


FIGURE 3 Two-stroke cycle gasoline engine operating cycles with crankcase scavenging. (A) Power or downstroke. (B) Exhaust and intake. (C) Compression or upstroke. Courtesy of Prentice Hall.

determine how many power strokes are delivered within the 720° of crankshaft rotation in the four-cycle engine. For example, in a four-cylinder engine (720° divided by four), one power stroke will be delivered for every 180° of crankshaft rotation, whereas in a six-cylinder engine (720° divided by 6), one power stroke will be delivered for every 120° of crankshaft rotation.

• The exhaust stroke, which begins when the exhaust valve(s) is opened before the piston reaches bottom dead center (BDC), typically between 30 and 40° of crankshaft rotation BBDC. The exhaust valves may remain open until 20° of crankshaft rotation after top dead center (ATDC) for a total duration of between 230 and 240° of crankshaft rotation.

6. FOUR-STROKE CYCLE POLAR VALVE TIMING

The actual valve opening and closing in a four-cycle gasoline or diesel engine can be considered similar. Figure 5 illustrates in graphical form the piston position and valve timing duration of the individual strokes.

Figure 6 illustrates one example of the opening and closing degrees for a four-stroke cycle,

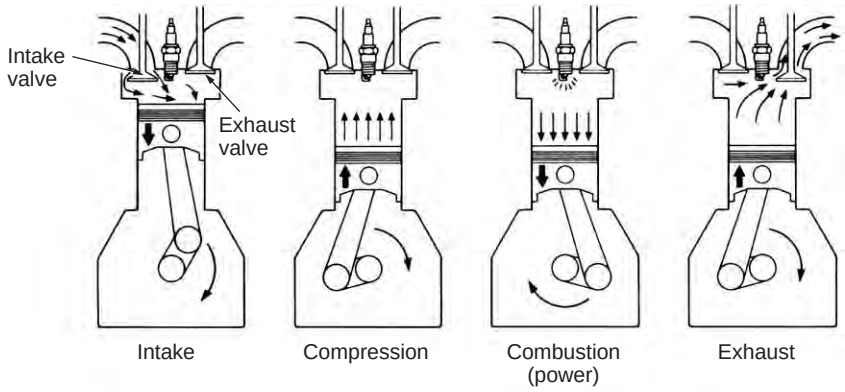


FIGURE 4 Schematic operation of a four-stroke cycle gasoline engine. Courtesy of Prentice Hall.

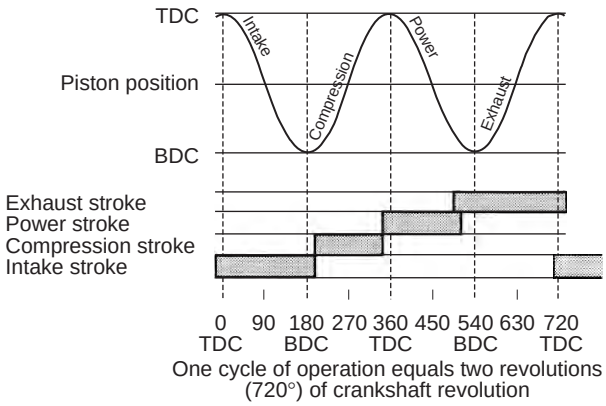


FIGURE 5 Graphical representation of the piston position and individual stroke duration for a four-stroke cycle gasoline engine. Courtesy of Prentice Hall.

high-speed diesel engine. This diagram is commonly referred to as a "polar valve timing diagram" because it shows the piston TDC at the north pole position and the piston BDC at the south pole position. For simplicity, the figure shows the four piston movements throughout the 720° of crankshaft rotation superimposed in a graphical 360° format. On the right-hand side of the diagram, the figure shows the piston downward movement that occurs during both the intake and power strokes, whereas on the left-hand side, it shows the upward-moving piston during both the compression and exhaust strokes. Whereas the individual strokes are illustrated in graphical form in Fig. 5 through a 720° rotation, Fig. 6 shows the number of degrees of valve opening in crankshaft degrees as being either BTDC, ATDC, BBDC, ABDC, TDC, or BDC.

In Fig. 6, the following sequence of events occurs:

1. *Intake stroke:* occurs when intake valves open at 16° BTDC and close at 36° ABDC (total duration is 232° of crankshaft rotation)
2. *Compression stroke:* occurs when the intake valves close at 36° ABDC until TDC (total duration is 144° of crankshaft rotation)
3. *Power stroke:* starts at TDC and continues until the exhaust valves open at 60° BBDC (total duration is 120° of crankshaft rotation)
4. *Exhaust stroke:* occurs when exhaust valves open at 60° BBDC and close at 16° ATDC (total duration is 256° of crankshaft rotation)

In this polar valve timing diagram, note that both the intake and exhaust valves are open for a period of 16° BTDC and for 16° ATDC. This design feature provides what is commonly referred to as a positive valve overlap condition, which in this example amounts to 32° of crankshaft rotation. Opening the intake valve(s) toward the end of the exhaust stroke (16° BTDC) permits the incoming fresh air to help fully scavenge any remaining exhaust gases from the cylinder through the still open exhaust valve(s) that is closed at 16° ATDC.

7. TWO-STROKE CYCLE DIESEL ENGINE: BASIC OPERATION

In the high-speed, heavy-duty, two-stroke cycle diesel engine design shown in Fig. 7, only camshaft-operated poppet exhaust valves are used. The intake of air is obtained through a series of ports machined around the circumference of the cylinder liner at its midsection. A gear-driven roots-type blower is employed to supply the air necessary for engine operation. Typically, the blower supplies air at a pressure between 4 and 7 psi (27.6–48.2 kPa) above atmospheric. High-speed, higher power output two-cycle engines can also employ an exhaust-driven turbocharger in series with the blower to supply a

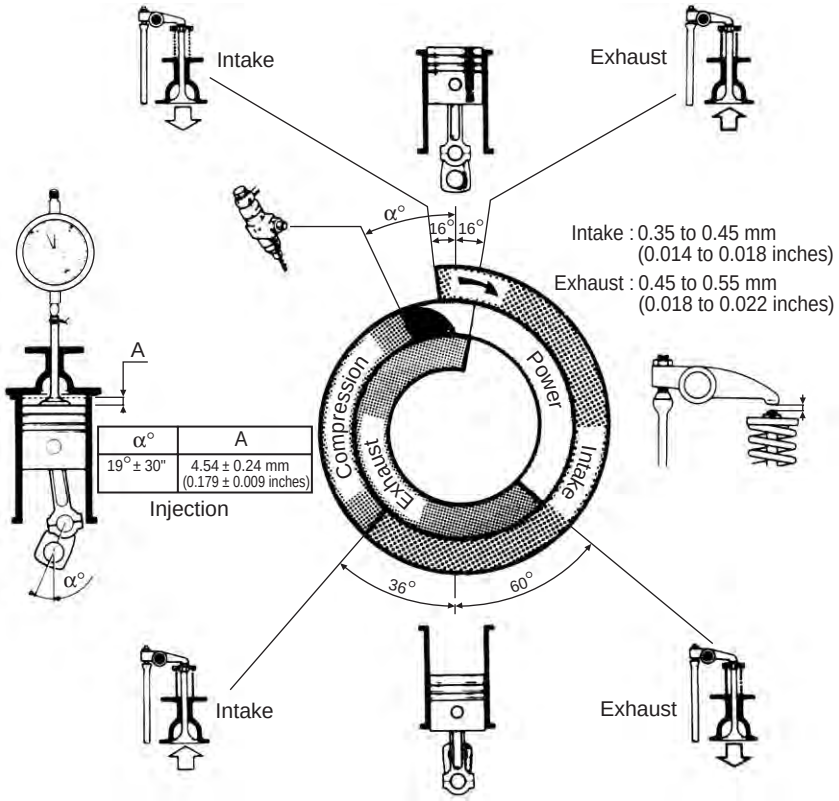


FIGURE 6 Polar valve timing diagram showing the individual stroke degrees for a four-cycle diesel engine. Courtesy of Volvo/Mack/RVI Inc.

higher boost air supply. This results in a boost under full-load rated speed conditions of between 50 and 60 inches Hg (24.5–29.5 psi or 169–203 kPa. In the two-cycle diesel, every upstroke of the piston is compression and every downstroke creates power. Therefore, only 360° of crankshaft rotation is necessary to create one power stroke versus the 720° of rotation required with the four-stroke cycle engine.

In the four-stroke cycle engine design, the pistons basically act as an air pump to permit air to enter the cylinder and to expel exhaust gases from the cylinder. In the two-stroke cycle design, the individual intake and exhaust strokes are basically eliminated. The pressurized air from the blower/turbo is stored in an air box that completely surrounds the ports of every cylinder liner. The air box acts somewhat like a plenum or storage chamber. Anytime the cylinder ports are uncovered, pressurized air can flow into and through the liner. If the exhaust valves are closed, air is trapped within the cylinder. If the exhaust valves are open, scavenging of exhaust gases will occur. The gear-driven blower, or the blower in conjunction with a direct mounted or upstream mounted exhaust-driven turbocharger, supplies the air flow necessary for several functions:

- Scavenging of exhaust gases from the cylinder
- Cooling of internal components such as the cylinder liner, the piston, and exhaust valves (approximately 30% of the engine cooling is achieved by air flow)
- Air required for combustion purposes
- Crankcase ventilation by controlled leakage of air past the oil control rings when the piston is at TDC

The four operational events that occur in the two-cycle engine are as follows:

1. When the piston is at TDC on its power stroke, it is pushed down the cylinder by the force of the high-pressure expanding gases. At approximately 98° ATDC, the exhaust valves are opened, permitting the pressurized exhaust gases to start leaving the cylinder via the opening exhaust valves.
2. At approximately 60° BBDC, the downward-moving piston begins to uncover the cylinder liner

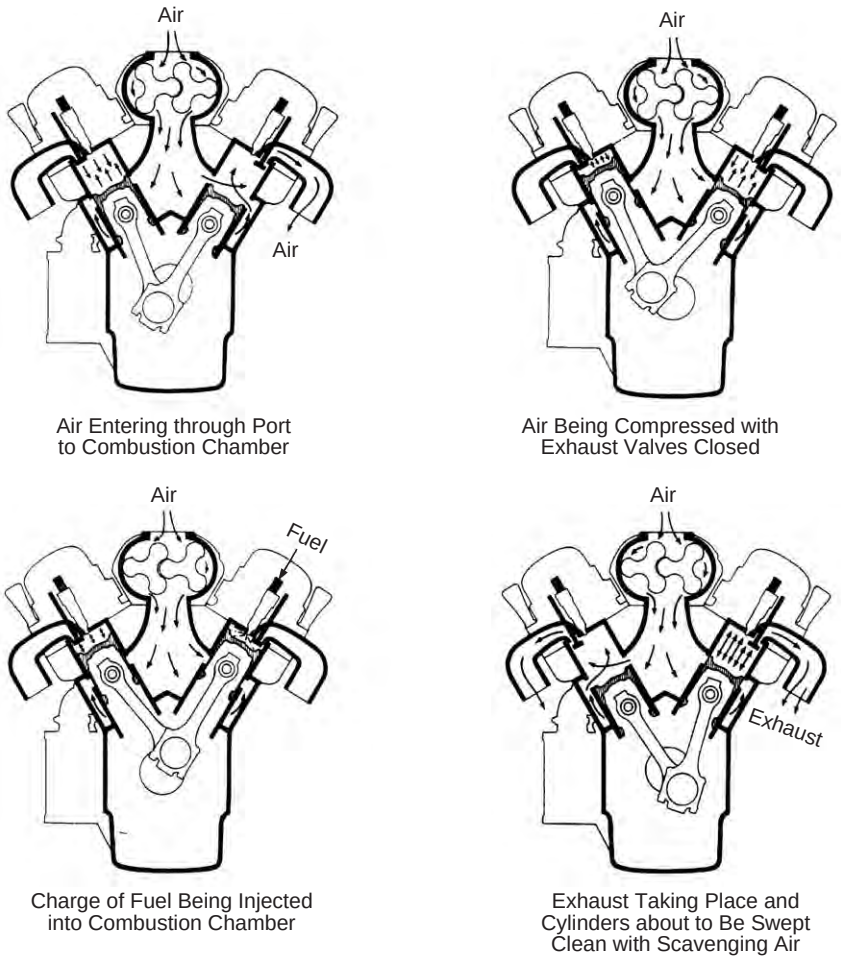


FIGURE 7 Two-stroke, heavy-duty, high-speed diesel principle of operation. Courtesy of Detroit Diesel Corporation, Daimler–Chrysler.

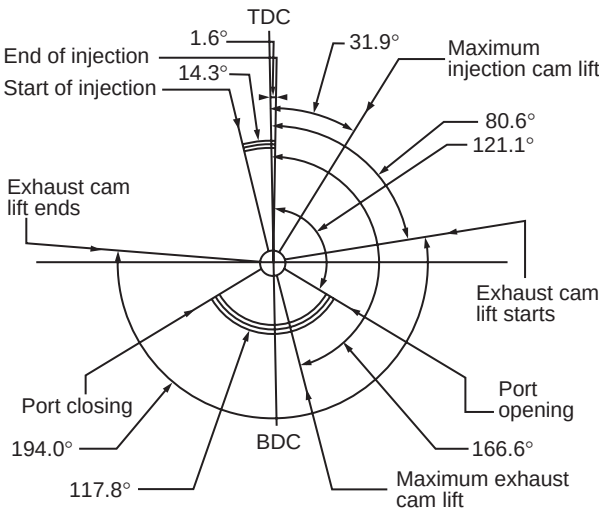
ports. At this point, the air box pressure is higher than the cylinder exhaust gas pressure, permitting this air to start entering the cylinder to assist in scavenging the burned exhaust gases from the cylinder. The liner ports are fully open when the piston reaches BDC for maximum air flow.

3. At approximately 60° ABDC, the upward-moving piston closes the liner ports. During the duration of 60° BBDC and 60° ABDC (120°), the air box pressure functions to provide exhaust scavenging, cooling, and fresh air for combustion purposes. It is during this 120° of liner port opening that exhaust and intake occurs.

4. The camshaft is timed to permit the exhaust valves to remain open for several degrees after the liner ports are covered by the upward-moving piston. This ensures that all burned exhaust gases have been fully scavenged from the cylinder and that a full charge of fresh air is retained in the cylinder. When the exhaust valves are closed, the compression stroke begins and the trapped air pressure and temperature increases. Based on engine speed and load, high-pressure finely atomized fuel is injected into the combustion chamber BTDC. The burning fuel reaches maximum pressure at TDC to force the piston back down the cylinder on its power stroke, and the sequence is repeated.

8. TWO-STROKE CYCLE DIESEL POLAR VALVE TIMING

Figure 8 illustrates one example of the actual degrees of valve timing used for a two-stroke cycle turbocharger and blower scavenged engine model.



	0.014 hot lash	0.016 cold lash
Exhaust valve lift starts-	98.2° ATDC	98.9° ATDC
Exhaust valve lift ends-	246.0° ATDC	243.0° ATDC

Maximum injection cam lift- 0.2755
Maximum exhaust cam lift- 0.3270

FIGURE 8 Polar valve timing diagram for a two-stroke cycle, heavy-duty, high-speed diesel engine. Courtesy of Detroit Diesel Corporation, Daimler–Chrysler.

9. FOUR-STROKE CYCLE DIESEL ENGINE: BASIC OPERATION

The operation of the four-stroke cycle diesel engine is illustrated in Fig. 9. On the intake stroke, only air (not an air/fuel mixture) is supplied into the cylinder, whether the engine is naturally aspirated or turbocharged. Air flow into the diesel engine cylinder is nonrestricted, meaning that the operator does not control a butterfly valve within the intake manifold as he or she does on a gasoline engine; rather, the throttle pedal or hand throttle is directly connected to the fuel injection pump control rack in mechanically governed fuel systems so as to directly control the fuel delivery rate. In an electronically controlled diesel engine, the position of the hand or foot throttle determines the voltage input signal to an ECM. The ECM then activates the electric solenoid on the injector (Fig. 1) to allow fuel delivery into the combustion chamber. Therefore, the start, duration, and end of fuel delivery (injection) is based on the duty cycle or injector solenoid activation period in crankshaft degrees.

Because of the much higher compression ratio used in the diesel, the heat of the compressed air toward the end of the compression stroke is sufficiently high to cause the injected liquid atomized diesel fuel droplets to convert to a vapor after a small ignition delay. Once the fuel vaporizes, ignition occurs to establish a flame front and the remaining injected diesel fuel burns immediately. By the time the piston reaches the end of the compression stroke, the burning fuel causes a further increase in the pressure above the piston.

As more fuel is injected and burns, the internal gases get hotter and the pressure increases further, pushing the piston and connecting rod downward to rotate the crankshaft during the power stroke. Once the exhaust valve(s) is opened, the pressurized gases force their way out of the cylinder and the exhaust stroke begins. When the piston starts its upward stroke again, it pushes the remaining exhaust gases from the cylinder. With a positive valve overlap condition, the intake valve(s) in the cylinder will be opened before the piston reaches TDC. Typical valve timing in a four-stroke cycle diesel is similar to that in a four-stroke cycle gasoline engine.

10. DIRECT VERSUS INDIRECT DIESEL ENGINE INJECTION CONCEPTS

Diesel engines can use either a direct injection (DI), an indirect injection (IDI), or what is referred to as a swirl chamber (SC) design. DI concepts are the design of choice in most diesel engines today due to their superior fuel consumption qualities, higher available injection pressures, and cleaner exhaust emissions. Figures 1, 7, and 9 show a DI design where the high-pressure atomized diesel fuel is injected directly into the combustion chamber that is formed by the shape of the piston crown and the underside of the cylinder head. Injection pressures ranging between 25,500 and 30,000 psi (176–207 Mpa) are obtained in today's electronically controlled, high-speed, heavy-duty engines. Various piston crown designs are in use with DI engines. The injector spray tip uses multiple holes/orifices to deliver the injected fuel throughout the compressed air mass of the combustion chamber, with the number, size, and spray-in angle of the orifices being determined by the piston crown shape.

The IDI design is also referred to as a precombustion chamber (PC) type, shown in Fig. 10A, where an electric glow plug facilitates ease of combustion, particularly in cold weather operation. In the PC or

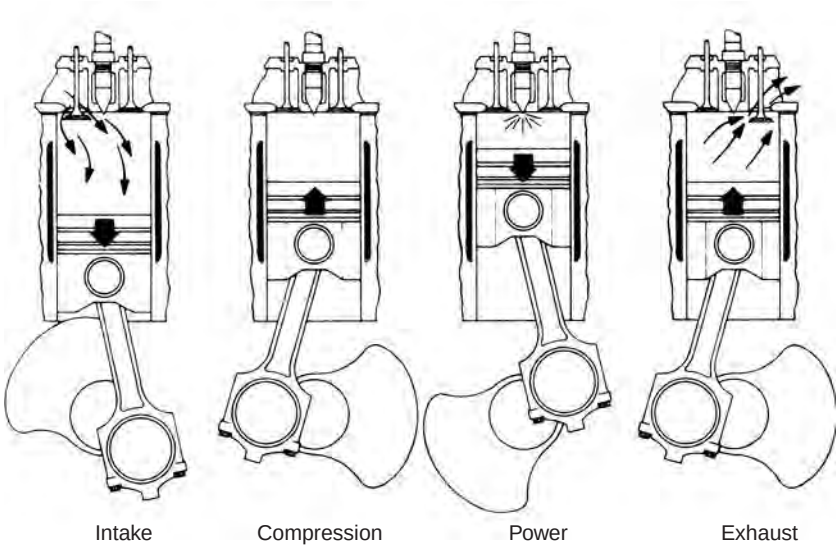


FIGURE 9 Sequence of individual piston and valve events for a four-stroke cycle diesel engine. Courtesy of Detroit Diesel Corporation, Daimler–Chrysler.

IDI design, a small auxiliary chamber, generally centrally located with respect to the main combustion chamber, supports the glow plug and the fuel injector/nozzle. The nozzle spray tip used generally consists of a single hole/orifice to deliver fuel. The small auxiliary chamber typically contains between 25 and 35% of the clearance volume with the piston at TDC. Once ignition occurs, the expanding high-pressure gases escape into the main chamber above the piston to force it down. Design variations exist between the PC and IDI concepts. Some engines also employ an SC design with a glow plug, shown in Fig. 10B, where the combustion chamber is usually located off to one side of the cylinder head. The high-pressure gases enter through a throat or passage onto the top of the piston to drive it down the cylinder on the power stroke. An SC design, although similar to the PC system, contains approximately 50 to 60% of the total clearance (compression) volume when the piston is at TDC. Both the SC and PC/IDI designs can be characterized as divided chamber engines where they are widely used in high-speed light industrial and automotive passenger cars. However, later engine models in these applications are now using more DI designs.

11. REVIEW OF TWO-STROKE VERSUS FOUR-STROKE CYCLES

These two cycles can be summarized by considering that the piston operation is divided into what are commonly referred to as "closed" and "open" periods. The closed period occurs during the power stroke, and the open period occurs during the time when the inlet and exhaust strokes are occurring. Consider the following sequence for the two-stroke cycle:

- Closed period:
  - a–b: compression of trapped air
  - b–c: heat created by the combustion process
  - c–d: expansion or power stroke
- Open period:
  - b–e: blow-down or escape of pressurized exhaust gases
  - e–f: scavenging of exhaust gases by the blower and/or blower–turbocharger combination
  - f–g: air supply for the next compression stroke.

All of the preceding events in a two-cycle diesel engine occur within 360° or one complete turn of the engine crankshaft/flywheel. Now consider the following sequence for the four-stroke cycle:

- Closed period:
  - a–b: compression of trapped air
  - b–c: heat created by the combustion process
  - c–d: expansion or power stroke
- Open period:
  - d–e: blow-down or escape of pressurized exhaust gases
  - e–f: exhaust stroke
  - f–g: inlet and exhaust valve overlap
  - g–h: induction stroke
  - h–i: compression stroke

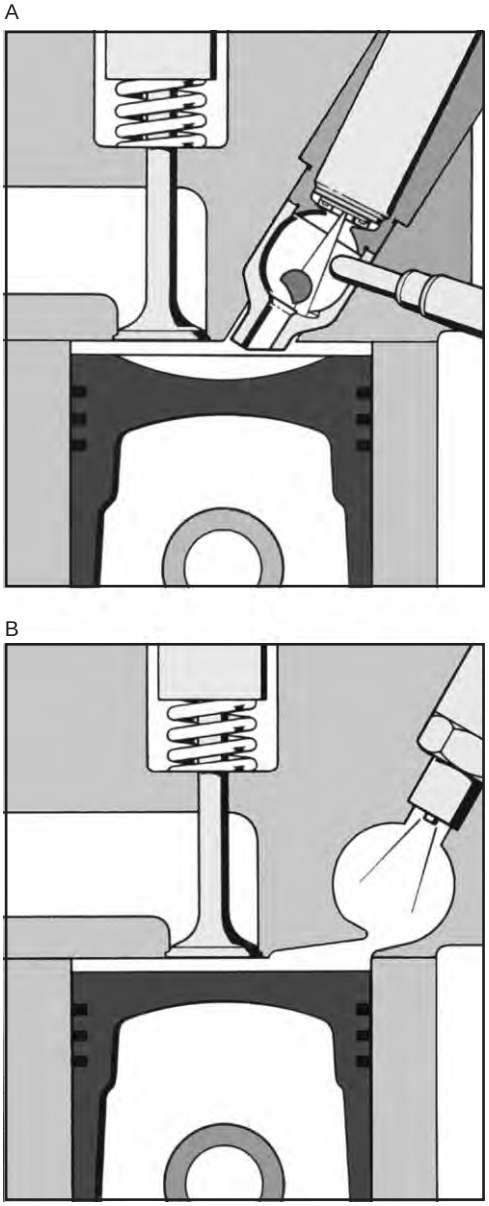


FIGURE 10 (A) Precombustion chamber design concept. (B) Swirl chamber design concept. Courtesy of Robert Bosch GmbH.

All of the preceding events require 720° of crankshaft/flywheel rotation, in contrast to the 360° required in the two-cycle engine model.

12. COMPARISON OF TWO-CYCLE VERSUS FOUR-CYCLE ENGINES

Consider the following specific operating differences between the two-stroke and four-stroke cycle engines, both gas and diesel. In the four-stroke cycle engine, there is a longer period available for the scavenging of exhaust gases and the separation of the exhaust and inlet strokes. In addition, with a shorter valve overlap period versus the port valve concept in the two-stroke engine, there tends to be a purer air charge at the start of the compression stroke in the four-cycle model than there is in a conventional blower air-supplied two-cycle model.

The four-stroke cycle engine pumping losses occur during the intake and the exhaust strokes when the pistons basically act as air pumps. The thermal (heat) loading on the piston, valves, cylinder head, and cylinder liner tend to be lower on a four-cycle engine because the power stroke occurs only once every two crankshaft revolutions versus once every revolution in the two-cycle engine. It is also easier to lubricate the pistons, rings, and liners in a four-stroke engine due to the absence of ports that are required around the midpoint of the two-stroke liner. Adding a turbocharger and the blower can increase the air flow substantially in the two-stroke engine, closer to the characteristics of the four-stroke model. Both four-stroke and two-stroke engine designs have pumping losses. In the two-cycle engine, the power losses required to drive the gear-driven blower reduce the net engine power output. In addition, the two-stroke engine requires a much larger air flow capacity to operate than does the equivalent displacement four-cycle engine.

Compare the pumping losses mentioned previously between the two-stroke and four-stroke cycle engines. In general, on a non-turbocharged two-cycle engine, the power required to drive the blower is less than the four-cycle pumping losses when the engines are operating at less than 50% of their rated speed/load. However, from 50% up to rated speed, the four-cycle engine's pumping losses tend to be roughly two-thirds those of the two-cycle model. On two-cycle models that also employ an exhaust gas-driven turbocharger along with a bypass blower feature, these pumping losses are reduced substantially. The two-cycle engine tends to have a slightly higher fuel consumption curve due to its double power stroke principle through the 720° for a four-stroke cycle model plus the higher overall pumping losses experienced in the two-cycle model above 50% of rated speed. The brake mean effective pressure (BMEP), which is the average pressure exerted on the piston crown during the power stroke, is generally lower on a two-cycle diesel. If one compares equivalent displacement and horsepower-rated engines at 450 to 475 hp (336–354 kW), the BMEP in the two-stroke engine model would be

approximately 115 to 128 psi (793–883 kPa). On the four-cycle model, the BMEP would be approximately double the BMEP figures cited for the two-cycle model.

The brake-specific fuel consumption (BSFC) of a two-stroke engine tends to be higher than that for a comparably rated four-cycle engine. BSFC is simply the ratio of fuel burned to the actual horsepower produced.

To summarize the advantages and disadvantages of these two different cycles, we can cite the following:

Advantages—four cycle:

1. Better volumetric efficiency (VE) over the engine speed range (VE is a measure of the "breathability" of the engine or the extent to which the cylinder of an engine is completely filled by the incoming charge following an exhaust stroke. It is also important to understand that the VE is a ratio of masses, not of volumes.)
2. Low sensitivity to pressure losses in the exhaust system
3. Better control of the cylinder charging efficiency through control of valve timing and matching of the intake manifold system

Disadvantages—four cycle:

1. Valve control potentially more complex
2. Engine power density lower due to only one piston power stroke every 720° of crankshaft rotation

Advantages—two cycle:

1. Simple engine design, low weight, and low manufacturing costs
2. More favorable torsional forces (shorter number of crankshaft degrees between power strokes)

Disadvantages—two cycle:

1. Higher fuel consumption
2. Higher hydrocarbon exhaust emissions (lower scavenging efficiency
3. Lower BMEP due to lower VE
4. Higher thermal load because every upstroke is compression and every downstroke is power (no separate intake and exhaust strokes)
5. Can exhibit poorer idle quality due to the higher percentage of residual gases retained in the cylinder (particularly in gasoline engines).

13. HORSEPOWER DESCRIPTION

To clarify and appreciate the work produced in an IC engine, it is appropriate to understand the two main terms "horsepower/kilowatts (kW)" and "torque." Failure to understand these differences will result in confusion when discussing the operation of any IC engine.

The term "horsepower" is used in the nonmetric vernacular to describe how much work an IC engine can produce, with work being equal to Force × Distance. Power produced is a measure of the rate at which work is done during a given time period, with 1 hp equivalent to a rate of work equal to 33,000 foot-pounds/min. In the metric system of measurement, 1 hp is equal to 0.746 kW; therefore, 1 kW is equal to 1.341 hp. In an IC engine, the force of the hot high-pressure expanding gases acting on the piston crown during the power stroke forces the piston down the cylinder and transfers the reciprocating (back and forward) motion via a connecting rod to convert it into rotary motion at the engine crankshaft. Therefore, the rate of work performed within the engine cylinder will increase as the engine crankshaft revolutions per minute (rpm) rotates faster (greater number of power strokes delivered during a given time period). The product of the BMEP in the cylinder and the piston displacement produces the piston work, and the actual number of working cycles (power strokes) per unit of time is what determines the indicated horsepower (ihp), which is nonusable power at the crankshaft or internal power produced in the cylinder without taking frictional losses into consideration. For example, an engine specification that indicates the engine can produce 500 bhp (brake horsepower or usable power at the flywheel) at 5000 rpm will not produce this same power at 2000 rpm because the rate of doing work is much less at the lower speed. Also, one should bear in mind that horsepower cannot be multiplied.

14. TORQUE DESCRIPTION

The word "torque" means a twisting or turning force applied to a fulcrum point that tends to cause torsion or rotation. Torque is expressed in pound-feet or pound-inches in the English system of measurement, whereas it is expressed as Newton-meters in the metric system. In an IC engine, the power that drives the piston down the cylinder is transferred to torque through the connecting rod that is bolted to the rotating crankshaft. The connecting rod functions as a mechanical reaction lever; therefore, its length and the distance between the center line of the connecting rod and the center line of the rotating crankshaft will

TABLE I

Comparison of Pressures and Temperatures Created within a Gasoline Engine versus a Diesel Engine

	Gasoline	Diesel
Intake	Air and fuel	Air only
Compression	8.0:1 to 10.5:1 ratio	15:1 to 23:1 ratio
Typical average values	130–175 psi pressure (896–1206 kPa) 540–650°F temperature (282–343°C)	425–600 psi pressure (2930–4137 kPa) 1000–1200°F temperature (538–649°C)
Combustion	Spark ignition	Compression ignition
Power	Maximum pressure 450–600 psi (31103–41137 kPa)	Maximum pressure 1800–2300 psi (12411–15858 kPa)
Exhaust	Temperature of 1300–1800°F (704–982°C)	Temperature of 700–1100°F (371–593°C)
Thermal efficiency[a]	22–28% average	32–44% average

Note. At the end of the compression stroke in some gasoline engine models with a homogenous air/fuel charge, the pressure can be as high as 290–580 psi (20–40 bar), with temmperatures ranging between 752 and 1112°F (400–600°C).

[a] Percentage of return in work value for investment in fuel or percentage of return for every dollar of fuel consumed in the engine. Current fuel injected and electronically controlled gasoline engines can return thermal efficiency values in the 30% range. Waste heat recovery stationary diesel power plants can return total thermal efficiency values in the middle to high 50% range.

have a direct influence on the amount of torque generated at a given engine speed. The torque is actually what keeps the crankshaft rotating; therefore, it is responsible for the ability of the engine to handle a load and for the engine's capacity to do work. Maximum torque is usually generated at approximately 60 to 65% of the engine's rated revolutions per minute because it is a combination of force and time acting on the piston crown plus volumetric efficiency. Unlike horsepower, engine torque can be multiplied by directing its energy through clutches and/or torque converters, then through a series of gear ratios within the transmission, and then through the additional gearing in the axle. Therefore, if one increases the engine speed, torque will decrease but horsepower will increase, and vice versa.

15. DIESEL/GASOLINE ENGINE COMPARISON

Table I highlights the basic differences between a gasoline engine and a diesel engine during the four strokes of intake, compression, power and exhaust.

16. SUMMARY: GASOLINE VERSUS DIESEL POWER

The heavier construction, higher compression ratio, nonthrottled air supply, and greater heat value per pound of diesel fuel, as well as the ability to produce rated horsepower (kW) at a lower speed, develop substantially higher torque curves, return superior fuel economy (better thermal efficiency), and offer "life-to-overhaul" typically three to four times longer than that of a gasoline engine, are major attributes of the diesel engine. Today's heavy-duty truck electronically controlled diesel engines offer warranties of 500,000 miles (805,000 km) plus in heavy-truck applications, with accumulated mileage of between 800,000 and 1,000,000 miles (1,287,000–1,609,000 km) between major overhauls. Factored over the life of the engine/equipment, a substantial savings amounting to thousands of dollars per year in operating costs is realized by an owner/operator. In fleet use with hundreds or thousands of pieces of rolling stock, these features and advantages can amount to literally tens of millions of dollars.

SEE ALSO THE FOLLOWING ARTICLES

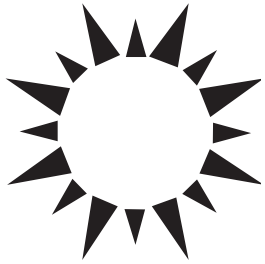
Biodiesel Fuels • Combustion and Thermochemistry • Internal Combustion Engine Vehicles • Mechanical Energy • Rocket Engines • Transportation Fuel Alternatives for Highway Vehicles • Vehicles and Their Powerplants: Energy Use and Efficiency

Further Reading

Brady, R. N. (1996). "Modern Diesel Technology." Prentice Hall, Englewood Cliffs, NJ.

Brady, R. N. (1997). "Heavy-Duty Trucks, Powertrains, Systems, and Service." Prentice Hall, Upper Saddle River, NJ.
Brady, R. N. (2001). "Automotive Electronics and Computer Systems." Prentice Hall, Upper Saddle River, NJ.
Brady, R. N., and Dagel, J. F. (2002). "Diesel Engine and Fuel System Repair." Prentice Hall, Upper Saddle River, NJ.
Lilly, L. R. C. (1984). "Diesel Engine Reference Book." Butterworth, London.
Robert Bosch Gmb, H. (1999). Bosch Automotive Handbook, 5th ed. Robert Bosch GmbH, Stuttgart, Germany. (Distributed by Society of Automotive Engineers, 400 Commonwealth Drive, Warrendale, PA 15096-0001.)

International Comparisons of Energy End Use: Benefits and Risks



LEE SCHIPPER
World Resources Institute
Washington, D.C., United States

1. **What Are International Comparisons of Energy Use?**
2. **Approach and Data Sources**
3. **Sectoral Examples**
4. **Overall Comparisons of Energy Use**
5. **Lessons from International Comparisons**

Glossary

activity (or output) The basic unit of accounting for which energy is used at the sectoral level.
degree day The difference between the average indoor and outdoor temperatures during a day, integrated over the entire heating season.
delivered energy (or final consumption) Energy supplied, for example, to a building, factory, or fuel tank and ultimately converted to heat, light, motion, or another energy service; transformation and distribution losses up to the final user are not included.
energy efficiency The ratio of energy converted into a desired form to energy put into the conversion; more commonly, it is a measure of energy converted to useful output divided by the energy "consumed" to produce this output.
energy intensity Energy "consumed" per unit of activity or output; it is usually presented at the most disaggregated subsectoral or end use level, but it can also be presented as energy use for an entire sector per unit of sectoral activity or output.
energy services Implies actual services for which energy is used (e.g., heating a given amount of space to a standard temperature for a period of time).
primary energy Delivered energy plus losses incurred in converting energy resources into purchased heat and electricity.
structure or subsectoral activity (or subsectoral output) The shares or modal mix (e.g., trucks, rail, and domestic shipping for freight; cars, two-wheelers, buses, rail, ship, and air for travel), ownership of using equipment by households (e.g., area per capita heated, refrigerators per capita), and manufacturing output by branch expressed as a share of total output.
useful energy Delivered energy minus losses assumed to occur in boilers, furnaces, water heaters, and other equipment in buildings; it is used for estimates of heat provided in space and water heating and cooking.

International comparisons of energy use entail detailed examination of the patterns of energy use over two or more countries. Comparisons are carried out in a disaggregated, bottom-up manner that permits comparing between or among countries measures of different economic and human activity and the energy uses associated with each activity.

1. WHAT ARE INTERNATIONAL COMPARISONS OF ENERGY USE?

International comparisons strive to understand how energy uses differ among countries of similar or varying geography, resource endowments, energy policies, incomes, and other differentiating characteristics. When undertaken carefully, international comparisons both reveal how energy uses differ and point to why these differences arise. Energy use patterns are complex, and comparisons among countries (or regions) at an aggregate level rarely reveal the nature of differences, the reasons for these differences, or the underlying policies that may contribute to these differences. The risks of international comparisons lie in the temptation to draw conclusions at too shallow a level of analysis. The benefits, which outweigh those risks, reveal

 © 2004 Elsevier Inc. All rights reserved.

themselves in insights as well as in pointers as to where to probe even deeper.

1.1 Why International Comparisons?

The most important reason for international comparisons lies in the need to understand how energy and development are linked. As economies grow worldwide, energy use increases, albeit at a slower rate. Analyses of how past development of wealthier countries affected energy use yield clues as to what is in store for developing countries as they grow. Still, over a time span of many decades, this growth in demand is worrisome due to pressures on world oil supplies, the environment, and global releases of carbon dioxide (CO_2) and other greenhouse gases (GHGs) associated with energy use. Energy use and its link to CO_2 emissions through fossil fuels plays a key role in negotiations and agreements related to climate. Understanding how wealthier countries use—and abuse—energy provides keys to less wealthy countries on how economic growth might diverge increasingly from energy use growth through careful management and clever policies.

Another reason for undertaking cross-country comparisons of energy use is to identify energy-efficient technologies. This is because politicians are wary of advocating energy use strategies that rely on more than just technologies. Unfortunately, a lack of sufficient data does not allow this to be done fully. One reason is that determining the true technological component of a particular energy intensity requires detailed study. For example, consider the case of space heating. The intensity of space heating in Sweden, measured as heat per unit of floor area and average winter outdoor temperature, is considerably lower than that in most other European countries. This suggests—but does not by itself prove—higher efficiency for heating in Sweden, that is, better insulation as well as efficient heating systems, heat distribution, and controls. New Zealand and Japan also have very low intensities of space heating. But how efficient are the space heaters, and what are the real thermal transfer properties of walls and windows, in New Zealand or Japanese homes vis-à-vis those in Sweden? Perhaps New Zealanders or Japanese simply do not heat very much, even in the cold regions. Available data suggest that this is the case, although adequate information for other countries is difficult to obtain at an aggregate level. Only a very detailed study of actual energy use and internal temperatures in hundreds of homes, as was done twice in Sweden, can verify that the real levels of thermal comfort in Sweden are very high. The conclusion is indeed that space heating in Sweden, with high indoor temperatures but low intensities, is very efficient. International comparisons alone point to an interesting possibility, but further work is needed to verify whether the superficial differences are really related to efficiency.

Comparisons guide policy formulation. Understanding how another country arrived at a particular structure may provide insights into how to steer one's own course. For example, careful study of the efficient state of housing in Sweden could lead to many ideas for promoting efficiency in the United States. Comparison of the effects of New Zealand's pricing of liquid petroleum gas (LPG), compressed natural gas (CNG), and diesel road fuels (whatever the reasons for the various levels of prices) with those for other countries illustrates the impacts of price differentials on fuel choices for new vehicles. Such international comparative studies can highlight important technologies that save energy as well as key policies that might promote or hinder enhanced energy efficiency.

In the end, international comparisons of energy use promote understanding on deep analytical levels, as well as on political levels, as to how and why energy use patterns differ. These differences explain much of national differences in energy and environmental strategies, particularly those aimed at restraining GHG emissions. Without a good understanding of each other's basic patterns, it is hard for leaders to agree on how to mold policies and measures to restrain carbon emissions. With that understanding, policymakers still need to look before they leap. Fortunately, international comparisons of energy use help to map out both known and unknown facets of energy use, enlightening policy formulation and execution for many purposes.

1.2 Aggregate Comparisons: Why They Tell Us So Little

Figure 1, from a series of national studies compiled by the International Energy Agency (IEA), shows a comparison of energy use and gross domestic product (GDP) among a number of countries and groups of countries. The division of energy use by GDP (or, alternatively, by population) is useful because the denominator normalizes energy use by an extensive quantity that is known to be a zero or even first-order driver of energy use. Without such a normalization, comparing any of the four Nordic

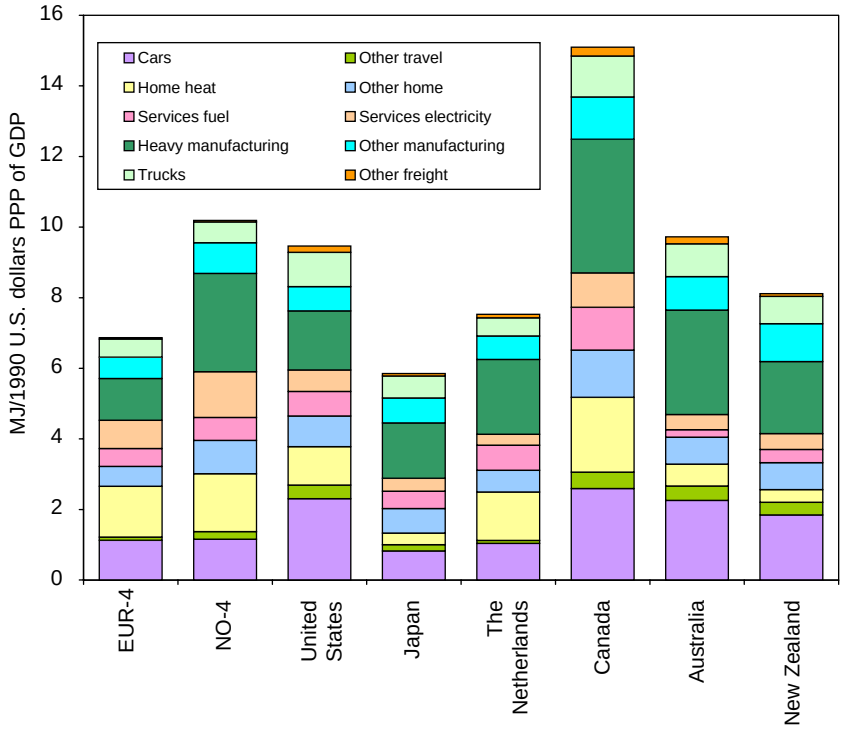


FIGURE 1 Energy use from 10 sectors/end uses relative to GDP. The EUR-4 consists of France, West Germany (or unified Germany from 1991 onward), Italy, and the United Kingdom. The NO-4 consists of Denmark, Finland, Norway, and Sweden. GDP is measured using purchasing power parity (PPP) to convert real local 1990 currency to U.S. dollars.

countries (NO-4) with the United States, which dwarfs them in geographical size, population, and total GDP, seems meaningless. Even comparing Canada, the United States, and Australia, three of the largest and wealthiest countries in the world, is hard to justify because their climates, industrial output mix, and housing structures differ considerably. Disaggregating energy uses into the sectors and activities shown, but keeping GDP as shown, is insightful but still cloaks most of the differences in how energy is used and, therefore, hides many of the reasons why energy use differs.

Instead of comparing energy use per sector against GDP, comparisons must start at a disaggregated level. Decomposition of changes in energy use can be summarized by the relation known as ASI. Put simply, it relates energy use E to three multiplicative terms on the right-hand side of an equation, with the sums over both indexes:

$$E = A_k * S_{k,j} * I_{k,j}.$$

In this decomposition, A represents overall sectoral activity (e.g., GDP in manufacturing). S represents sectoral structure (shares of output by two- or three-digit ISIC [International Standard Industrial Classification], branch of manufacturing or other industry, modal shares in transportation, and per capita equipment ownership in households). The k runs over the main sectors (here households, the service sector, manufacturing, other industries, travel, and freight). The j represents the subsectors, branches, or activities within each sector. The $I_{k,j}$ represents the energy intensities of each subsector j of activity k. The structure or relative output of each sector is represented by $S_{k,,j}$. The exact units of each of these variables are defined in what follows. The product $A_k * S_{k,j}$ in sector k gives output or activity in absolute values (e.g., total steel production, total distance traveled). This is also known as energy services. Multiplying by the energy intensity $I_{k,j}$ gives total energy use for subsector j of sector k. A fourth term could be added to the right-hand side to differentiate among energy carriers given that this is important for judging the importance of electricity or calculating carbon emissions for a given energy use.

Divide both sides by GDP (G), suppressing the subsectors j for notational ease:

$$E/G = [(A * S_k)/G] * I_k.$$

Clearly, E/G can differ because the sectoral A_k/G differ, because the elements of S differ, or because the I differ. Differences in individual I values are related

to, but not identical to, differences in energy efficiency. Differences in *S* values represent, within each sector in the structure, differences in energy use that can be related to technology, climate, industrial policy, natural resources endowment, land use, urban structure, and the like. Differences in *A* values, which aggregate sectoral activity into a few key indicators, also indicate broad differences in structure, output, and geography. The overall differences by sector (in *A*) or by subsector (in *S*) represent information that is as important to explaining differences in energy use/GDP as is the *I*. Because the *I* values are themselves only related to energy efficiencies, the overall ratio *E*/*G* cannot give an accurate picture of differences in energy efficiencies. Forming the ratio *E*/GDP lumps all of the differences in each country's structure into one parameter and then lumps all of the energy uses into a second parameter. These two aggregate numbers then represent neither the end uses of energy nor the structural features that drive many of these differences. This mixing leaves nothing to compare for meaningful international comparisons of energy use. Needless to say, the quotient of *E*/GDP itself cannot represent energy efficiency given that all of the activities and energy uses are aggregated first.

A particular problem is with GDP itself. Private vehicle uses, household activities related to energy consumption, and use of trucking fleets by firms for their own haulage account for 40% or more of national energy use but do not appear in GDP beyond the associated expenditures for energy because there are no transactions involved. The "outputs" of these activities do not show up as transactions anywhere.

This also means that Fig. 1 cannot tell us whether energy uses in any country or group of countries shown are efficient or inefficient. The figure only reveals how much fuel is used for cars relative to GDP, not relative to kilometers driven, which is the intensity related to fuel efficiency. However, energy use per kilometer only partly explains the obvious differences revealed in the figure. Total kilometers driven, relative to GDP, turns out to be just as important as fuel use per kilometer in explaining the total level of car fuel use/GDP shown in the figure. Knowing both of these vital elements—the *S* and the *I*—is crucial to making fair comparisons. Knowing enough about the characteristics of vehicles that shape the *I*, as well as the individual technological components, adds insights but becomes harder and harder unless only new vehicles (or homes, appliances, etc.) are compared due to enormous data availability problems. Hence, the minimum requirement for an international comparison is to be able to compare basic elements of *A, S,* and *I* between or among two or more countries.

Furthermore, even the use of the word efficient has to be scrutinized. It is dangerous to say that any country is energy efficient, even based on the disaggregated uses of energy shown in Fig. 1. First, countries themselves do not use energy; rather, individuals and public and private entities use energy. This is not a trivial point. For example, even in the centrally planned economy of the former Soviet Union, energy uses were affected by firms' and individuals' choices as to both how and how well to use energy. Although the ratio of energy use to GDP in the Soviet Union was far higher than in most other countries, the inefficiencies of energy uses explained only perhaps half of the differences. The rest was explained by very high production of energy-intensive raw materials, huge shipping distances, and cold climate. The size and cold climate of the Soviet Union certainly do not make that country inefficient, just as the lack of major space heating needs in Australia does not make that country efficient. Offsetting these factors were very small floor space for homes and commercial buildings, few household appliances, and very few cars (driven not very far) compared with Western countries. These latter factors did not make the Soviet Union any more efficient than the former ones made that country inefficient. Although choices of factory, transportation, and housing technologies may have been largely the results of central planning, these choices lead to explanations for only part of the high ratio of energy use to GDP. Those choices may be labeled inefficient, but the country per se may not be. For market-oriented nations of the Organization for Economic Cooperation and Development (OECD), it is even less obvious how to define national energy efficiency. Instead, international comparisons must be undertaken to explain how ratios of energy to GDP differ and then explain why the results occur.

Moreover, scholars attach energy efficiency to a particular energy-using process. When that process is taken by itself, such as the conversion of fuel to motion in a car motor (in reality, a collection of processes), its efficiency can be compared with that in other countries. But even at this level, automobiles differ in size and features. Thus, it turns out that automobiles in the United States actually use less energy per kilometer and per unit of weight than do those in nearly every other country. By that measure, U.S. automobiles are efficient. But they also use more

energy per kilometer because they weigh more than cars elsewhere. Thus, one can say that the U.S. car fleet is more fuel intensive than fleets in any other advanced country. The reason is that American behavior "favors" large cars, largely due to low car prices and low fuel prices. Behavior and cultural choices, inhibited or reinforced by policies, affect how the efficiencies of technologies are transformed into the actual energy intensities of energy uses one can measure or infer. Thus, analysis should be very mindful of what is being labeled as efficient.

The example of space heating noted previously is also instructive. Are the Japanese or New Zealanders, who heat to lower indoor temperatures than do Americans and most Europeans, more efficient at space heating or simply frugal? Here, as with the example of cars, differences in revealed energy intensities may represent what is seen as economically efficient for each country's households. Because that implies a different meaning of efficiency than one often used by engineers, one must use the word efficient carefully. Indeed, it might be better to avoid misunderstandings—to reduce the risks of international comparisons—by avoiding the use of the word efficiency entirely.

What can we learn about how different societies use energy? Energy uses and activity levels for approximately three dozen energy uses or activities are known in more than a dozen IEA member countries, in most cases over a period of time from years to decades. Similar estimates have been made for former Soviet bloc countries and at least for a year for most of the largest developing countries. These comparisons are reasonably accurate portrayals of the structure of energy use and its corresponding energy intensities. But the comparisons rarely measure energy efficiencies. Measuring how well energy is used on an economy-wide basis takes far more data than are normally available. With this distinction firmly in mind, the benefits of comparisons are maximized and the risks of misunderstandings are minimized.

1.3 Early Work in International Comparisons

Much of the forgoing emerges from studying literature on energy uses that has emerged since the 1960s, when data illustrating disaggregated national energy use patterns became available. At the same time, the energy and economics literature is replete with analyses and comparisons of energy consumption and GDP. However, such aggregate comparisons should not be mistaken for comparisons of energy use. These comparisons flourished due to the lack of data. Although the manufacturing sectors of most countries were nominally covered by energy balances whose detail matched national accounts or other measures of output, few countries separated energy consumption into the residential, commercial/services, and agricultural sectors from a common aggregate, and very few countries attempted to split the transportation sectors into modes (e.g., automobiles, trucks). Understandably, then, most scholars turned to the energy/GDP comparisons, squeezing information like water from a stone.

Work during the 1970s by Strout and colleagues at the Massachusetts Institute of Technology (MIT), as well as work undertaken at the World Bank, began to relate aggregate energy use to important "structural" variables measuring what happens in an economy such as production of raw materials, vehicle ownership, and even winter climate. Although it lacked real energy use data for these variables, the work did show that, on an aggregate energy use level, structural variables explained part of intercountry differences or changes in use in a country over time. High production of energy-intensive raw materials (e.g., steel), cold winter temperatures, and large shipping distances help to explain high energy use relative to GDP; conversely, low values of these parameters yield lower values of E/GDP.

Adams and Miovic made an important early disaggregated comparison of energy use in the manufacturing sector. They found the reduction in energy use per unit manufacturing during the 1950s and 1960s hard to explain because the price of energy fell during that period. However, they discovered that increases in energy efficiency and shifts of fuel mix toward oil and gas (with fewer combustion losses than coal), all occurring as the scale of manufacturing facilities increased, naturally occurred even as energy became cheaper; that is, technology outpaced energy prices. This work was the first to reveal the strength of international comparisons at a disaggregated level.

Much of the lack of detail in comparisons beyond the manufacturing sector made during the 1960s and 1970s may have been due to the problems of obtaining data. A massive ad hoc data collection for more than a dozen countries, using 1972 as a base year and building on the ASI relationship shown previously, was done for the World Alternative Energy Strategies study at MIT. Systematic data began to appear only during the late 1970s,

supported by national energy authorities in North America, much of Europe, and Japan. Experts reconstructed energy use patterns, often back to the 1960s, again with the ASI approach. The goal was nearly always to build a firm understanding of the past as a way of projecting future options in a disaggregated manner.

An early prominent binary comparison of energy uses was carried out by SRI International, comparing West Germany and the United States using data from 1972–1973. A more detailed binary comparison of energy uses, comparing the United States and Sweden, was published in 1976.

Both studies found that structural differences, which measured different levels of activity relative to GDP, accounted for significant differences in energy use relative to GDP. For example, Sweden and West Germany produced more steel, relative to GDP, than did the United States, raising energy use in the former two countries. But auto ownership and use in West Germany was lower than that in the United States relative to GDP, lowering energy use in the former country, whereas in Sweden this structural feature was close to that of the United States. On the other hand, both studies also identified differences in energy intensities, particularly automobile fuel use per kilometer, energy use per year for each major appliance, and energy use per tonne (metric ton) of various raw materials, suggesting that U.S. energy intensities were higher than those in the two European countries. For the first time, analysts could pinpoint key elements of the differences among sets of countries.

The first full comparison of a number of countries was carried out by Resources for the Future (RFF) in 1977. In "How Industrial Societies Use Energy," RFF used available data to quantify key differences in the structure and intensity of energy uses among the United States, Japan, Canada, and several European countries (the United Kingdom, France, Sweden, Italy, and The Netherlands). This epic investigation found that differences in both energy intensities and the structure of energy use contributed to differences in the overall ratio of energy use to GDP.

A detailed series of sectoral analyses of individual countries that included comparisons with up to 11 other countries was carried out at Lawrence Berkeley National Laboratory (LBNL) during the 1980s and 1990s and then at the IEA during the late 1990s. Drawing on 20 years of data (from 1973 to the mid-1990s) not available to the RFF study, and adding important countries such as Australia (with nearly no heating demand but a very energy-intensive industry structure) and Finland (the coldest of the IEA countries) to those studied, these analyses reinforced some of the key RFF conclusions but also quantified how the relative importance of structural and intensity differences evolved over that period. For one thing, structural difference among countries narrowed. The importance of manufacturing of heavy materials dropped in many countries, particularly Japan. Automobile ownership and use increased in Japan and Europe, closing the gap somewhat with the United States. Ownership of central heating in Europe and ownership of appliances in Europe and Japan gained on levels in the United States, closing that gap somewhat.

The IEA in 1997 developed energy indicators for more than a dozen countries based on these comparisons. (The United States, Canada, Japan, the four Nordic countries, West Germany, France, Italy, The Netherlands, the United Kingdom. Australia, and New Zealand were added in subsequent works.) The objective of this work was to describe the majority of IEA countries at the aforementioned level of detail, with the ultimate goal of introducing such data into the IEA databases for ongoing analysis and comparisons. An effort was made to standardize the approach and encourage other countries to develop data. Much of the effort was directed toward EUROSTAT as a way in which to ensure the highest level of quality of, and access to, data and indicators.

2. APPROACH AND DATA SOURCES

The landmark RFF comparison was the most disaggregated comparison of several countries ever undertaken through the late 1970s. Since the time of that work's base year (1972), many industrialized countries began to systematically improve their data. The IEA work, which has been the most exhaustive published work to date, described a number of key steps to define an approach to international comparisons of energy use, either across entire economies or in certain sectors. The same steps also enable disaggregated time-series analysis of energy uses in a given country. The key step is to be able to attach an activity measure to each energy use, that is, to make a disaggregated normalization.

- Disaggregation of each sector's energy uses for a given year (or over time) into subsectors (e.g., a limited number of branches of manufacturing), modes (e.g., passenger and freight transport), or

end uses (e.g., households)—in all, approximately 32 to 36 energy uses or sectors (certain compromises may permit lowering the number of subsectors, as is noted in the chapter reviewing each sector)
- Identification and quantification of structural parameters—output or activity—for each of the subsectors or branches of energy use
- Quantification of energy intensities for each output or activity
- Comparison of structural parameters and energy intensities
- Synthesis of results

If time-series analysis is required, this can be done use by use:

- Comparison of time trends in energy uses, energy intensities, and structural measures
- Brief comparison of trends with changes in measures of income and, in some cases, energy prices
- Decomposition of total sectoral energy use trends into underlying components and comparisons of these results with similar analyses for other countries

Using index techniques, individual sectoral trends in structure or intensity can be reaggregated and compared among sectors within a country over time. If many countries are involved, these aggregates can be compressed further to look at how overall trends in activity, structure, or intensity affect total energy use or energy use relative to GDP.

A hybrid approach is the safest. A comparison across individual years is acceptable. However, comparing time series of individual parameters or indicators shows whether the values for any particular country are robust and stable over time or waver due to errors, uncertainties, or unusual events. An aviation strike in Australia caused an unusual drop in actual flying in 1990 as planes flew empty after the strike. Thus, the intensity of aviation for that year was not representative of any long-term trend. For this reason, many of the comparisons presented here will be shown over time.

The hybrid approach can also be expanded in two dimensions. For example, plots of a structural variable (e.g., GDP) against another variable (e.g., floor area, automobile use) can be shown. Both variables evolve over time, and this is implicit in the portrayal. But by coupling two variables together and eliminating time as an axis, one can portray an approximate causal relationship.

2.1 Sectors and Definitions

The sectors commonly studied are listed in Table I. These typically cover 85% of all the final or primary energy use consumed in most IEA countries. The end uses or subsectors account for all energy use in each sector except in transport. There, approximately 10% of the fuels normally consumed are ignored (e.g., off-road vehicles, military, private aviation). Some gaps in data commonly arise in other sectors; for example, parts of agriculture, mining, and construction, as well as street lighting and public utilities (e.g., pumping of water or district heat), are either lumped in with energy sector losses or aggregated with the commercial/services sector. Finally, bunkers of international shipping or aviation and, in some countries, international trucking can be estimated but are usually not included in comparisons due to lack of activity data.

The approach in most studies uses final energy to measure the amount of fuels, electricity, or heat consumed within a branch of manufacturing, by a certain kind of vehicle, or within a house for a specific end use. Trends in primary energy use are referred to occasionally. Primary energy assigns to electricity (and heat in some countries) the average ratio of primary energy resources consumed to produce electricity (or heat) made available to the economy during the year in question. There is a large literature on how to count primary electricity from hydro, geothermal, and nuclear sources, so important that the official Swedish energy balances give aggregate primary energy using three different accounting schemes. If carbon emissions are of interest, standard coefficients can be used to infer the carbon emissions from combustion of fossil fuels.

Good comparisons are carried out in a transparent way. Data sources are clear, and data are available to allow analysts to adopt different conventions for counting primary energy or for measuring key structural variables such as floor area and vehicle ownership and use. Some countries, notably the United States, make the original survey data available at nominal or no cost, whereas others keep the disaggregated data "secret." Although some kinds of raw data may indeed be commercially valuable or may need to remain confidential to protect information from individual consumers, an open debate about energy use policies requires open access to data on how energy is used. Unfortunately, some major countries still do not conduct surveys regularly to determine how energy is used, in part due to the expense of such surveys and in part due to the burden

TABLE I

Measures of Activity, Sectoral Structure, and Energy Intensities for International Comparisons

Sector (k)	Subsector (j)	Activity (A_k)	Structure ($S_{k,j}$)	Intensity ($I_{k,j} = E_{k,j}/S_{k,j}$)
Residential/household		Population		
	Space heat		Floor area per capita	Heat[a]/floor area
	Water heat		Person/household	Energy per capita[b]
	Cooking		Person/household	Energy per capita[b]
	Lighting		Floor area per capita	Electricity/floor area
	Appliances (5–7)		Ownership[c] per capita	Energy/appliance[c]
	Other household electricity uses		—	—
Passenger transport		Passenger-kilometers		
	Cars		Share of total passenger-kilometers	Energy/passenger-kilometer
	Bus		"	"
	Rail		"	"
	Inland air		"	"
Freight transport		Ton-kilometers		
	Trucks		Share of total ton-kilometers	Energy/ton-kilometer
	Rail		"	"
	Coastal shipping		"	"
Commercial/services			"	"
	Services total	Floor area or GDP	(Not defined)	Energy/floor area or GDP
Manufacturing (with ISIC code)		Value added		
341	Paper and pulp		Share total value added	Energy/value added
351 352	Chemicals			
36	Nonmetallic minerals			
372	Iron and steel			
372	Nonferrous metals			
31	Food and beverages			
All other three-digit codes	Other manufacturing			
Other industry		Value added		
	Agriculture and fishing		Share total value added	Energy/value added
	Mining			
	Construction			

[a] Adjusted for annual winter climate variations and for changes in the share of homes with central heating systems.
[b] Adjusted for home occupancy (number of persons per household).
[c] Includes ownership and electricity use for six major appliance types.

of response. This lack of information also clouds the energy debate.

3. SECTORAL EXAMPLES

Table I gives a breakdown of six main energy use sectors used in international comparisons. This section reviews three of these in some detail and makes shorter remarks on two others.

3.1 The Residential/Household Sector

Population is the basic measure of activity in the residential/household sector. The residential/household sector "produces" space and water heat, cooking, cool temperatures for refrigeration and air conditioning, light, heat and mechanical action for washing and drying, and a number of other energy services. Lacking real measures of "output" (i.e., liters of water heated to a given temperature), some

of the measures of structure are based on per capita ownership of appliances or characteristics of dwellings themselves such as home area per capita. Changes in these values cause the "structural changes" in this sector.

The residential/household sector consists of energy consumed in dwellings. Calculation of this can be complicated due to secondary dwellings, homes in nonresidential buildings, and the like. Surveys of private expenditures, which follow people and households rather than dwellings, offer a reliable way of counting households or families that can be related to the number of occupied dwellings in one-, two-, or multifamily buildings. Usually, detached (one-family), semidetached (two-family), and row houses and mobile homes are counted with one- or two-family dwellings. Homes in dormitories and larger industrial or commercial premises are usually included with multifamily dwellings. Because both household and housing surveys have been carried out for decades in IEA countries, all of these results are generally compatible among countries.

Energy consumption as reported by energy suppliers cannot be easily separated into the residential/household and commercial building sectors because some fuels, notably oil, are often provided to large buildings where the supplier cannot say to which sector the customer belongs. Even if a gas or electric subscription is labeled residential, there is no guarantee that the label is correct given that so many utilities classify customers by size of overall consumption, not by the nature of the consumers. Because most energy suppliers belong to worldwide trade associations, there is good agreement across countries over the definitions of the residential/household sector and general recognition of where the definitions and data are the weakest.

Estimates of total energy use in the residential/housing sector vary in accuracy. Typically, the estimates for electricity and gas are more accurate than those for oil, coal, and wood. Only a "bottom-up" household energy use or energy expenditure survey, as carried out regularly in a handful of IEA countries (e.g., Japan, United States, The Netherlands, Sweden, France), gives an accurate measure of how much energy is used in this sector. Remaining countries make estimates based on utility deliveries, estimates of distribution of fuels supplied to large buildings for central heating, and so on. Overall, counting of gas and electricity from suppliers is the most accurate method, whereas counting of liquid and solid fuels is somewhat less accurate, with errors as large as 10% occurring due to these problems of misidentification.

3.1.1 Key Trends and Indicators for Comparison

The overall position of the residential/household sector in the mid-1990s in a number of countries and regions is shown in Fig. 1. Final energy use in the household sector had grown relatively slowly in most IEA countries since the 1980s as the main energy uses of space heating and electric appliances became saturated. Water heating and cooking represent the other two major categories of energy use. These have grown slowly during the past two decades or so, and their relative importance in overall energy uses continues to decline. The share of electricity has been rising both because electricity has gradually captured space and water heating (and cooking) from fuels and due to the increased importance of appliances, particularly those for information and communication. The uneven penetration of electricity for these is cause for caution in making comparisons. With much fewer losses in conversion within the home, the use of electricity (or district heat) for a given purpose appears to be lower than that of fossil fuels, which suffer from combustion losses in furnaces, boilers, and stoves. In contrast, counting the fuel primary energy requirements of electricity may overstate energy use relative to fossil fuels. In extreme cases of high electricity penetration (e.g., Norway, Sweden, Canada) or high use of district heating (e.g., Sweden, Denmark, Finland), it is advantageous to try to separate electricity, district heat, and nonelectricity, and this requires more data and provides an even more disaggregated form of the ASI equation.

Space cooling, largely from electricity, is usually contained in electric appliances. It is significant (5% of final energy use) only in Japan and the United States, but it is recorded (if under 2%) in France, Italy, Spain, Greece, Australia, and New Zealand. In warm developing countries such as India, Thailand, China, and Brazil, space cooling is rising in importance among household energy uses. Because many of these countries have little heating supplied by fossil fuels, electricity tends to be the most important energy form supplied to the residential/household sector.

3.1.2 Space Heating

Energy use for space heating ranges from 4 to 5 GJ per capita in warmer countries to as much as 30 to 40 GJ per capita in the United States, Canada, and countries in Northern Europe. Space heating is driven by the number of homes, the indoor temperature relative to the outdoor temperature, and the number of hours per year the homes are actually heated. Although population and the average area per home are growing slowly in IEA countries, the

number of homes continues to increase with the number of households at an annual rate as much as 1% faster than the population. Hence, per capita home area is increasing due to both factors. Higher GDP per capita both permits larger homes and allows more people to be able to afford to live apart. Thus, per capita area per unit of per capita GDP is a structural variable driving heating, and this is shown in Fig. 2. What can be seen is a strikingly large spread of per capita area at a given GDP.

To normalize for differences in climate, the number of degree-days is included. A degree-day represents the difference between the average indoor and outdoor temperatures during a day added up for the entire heating season; values (compared with a nominal 18°C indoor temperature) vary from as low as 900 degree-days in Australia to more than 4600 degree-days in Canada and Finland. These figures are averaged for the approximate distribution of population. An important conclusion is that both home area (relative to GDP) and winter climate differ enormously among the developed countries, contributing to significant differences in energy use.

The most useful intensity variable is formed by dividing total energy use for heating by total floor area and dividing further by degree-days. To portray the approximate differences between fuels, with combustion losses, and electricity or district heating, with no combustion losses in the home, "useful" energy is introduced. A nominal 66% efficiency for liquid and gaseous fuels and a 55% efficiency for solid fuels is taken, and the totals are added to the delivered values of heat and electricity. This avoids some of the problems of differences in fuel mix noted previously. The result gives a meaningful comparison of space heating intensity, as shown in Fig. 3.

What is striking about Fig. 3 is that, after normalizing for these important structural variables, large differences in space heating intensity still remain. The low values for Sweden, Denmark, and Norway reflect very good levels of insulation, as is somewhat the case for the United States. The low value for Japan reflects lower indoor temperatures and only partial heating of homes or rooms.

3.1.3 Electric Appliances

There are seven major appliances: white goods (refrigerator, refrigerator–freezer, and freezer), wet goods (dish and clothes washers, dryers), and air conditioners. The ownership (saturation or penetration, i.e., number/100 households) of major goods is known from private household surveys, censuses and household expenditure surveys, and household energy surveys. Important differences in size and characteristics are difficult to capture in a stock description but are known approximately. Such characteristics include freezer compartment size, whether washers use hot or cold intake water, and the like. Although these cannot be translated directly into energy use, they give qualitative explanations for differences in energy use.

Energy uses are more complicated to estimate. Only a few experiments have actually metered uses over enough households to make statistically valid estimates of annual energy use. Household ownership and total electricity consumption data can be analyzed using regression techniques to yield rough estimates of consumption per device. Indeed, in most

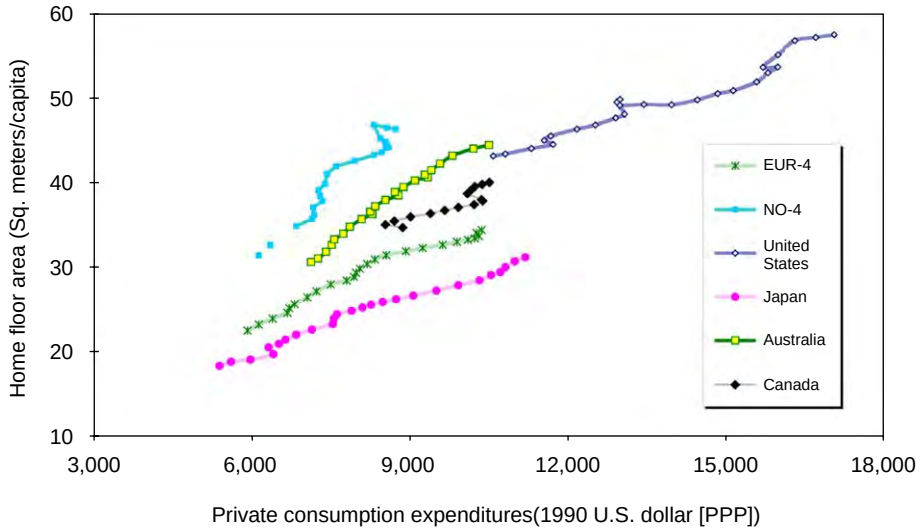


FIGURE 2 Home floor area and private consumption expenditures. See Fig. 1 for abbreviations.

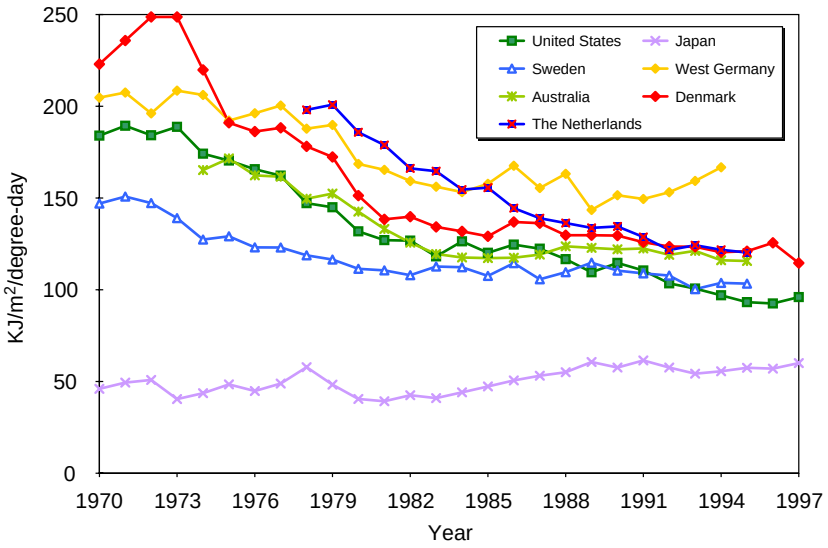


FIGURE 3 Space heating intensity: useful energy.

countries, utilities and their associations, researchers, and statistical bureaus often build models of ownership of each appliance by its age. Tests of new devices are more accurate measures of consumption or at least of given cycles (for wet goods).

Knowing the approximate energy use of new appliances of each type and year made, the model can estimate average energy for each appliance by integrating over the presumed age distribution of appliances. Survey or anecdotal information is used to estimate use of wet appliances (e.g., number of loads of washing or drying). The results can be compared with information on how residential electricity use varies for the entire sector by season and by hour of the day, and this also gives clues as to how much power each kind of appliance uses. The final results, in kilowatt-hours/appliance/year, give a rough indication of how much electricity is used for each kind of appliance. Although appliances in the United States have fallen in energy intensity since the imposition of efficiency standards during the late 1980s, they tend to use more energy than those in Europe and Japan due to differences in size/features and efficiency. But taken together, ownership, size, and features are still a greater source of differences in electricity use for appliances than is efficiency today.

3.1.4 Cooking and Water Heating

These end uses have become progressively less important in some countries but still account for 3 to 10 GJ per capita of delivered energy. By the 1980s, hot water penetrated into virtually every household in upper income countries. Some water is heated in dishwashers and clothes washers by electric elements and so should be transferred to "water heating."

Cooking has become a minor energy use. Typically, a gas stove might account for only 1 GJ per capita of home energy use, and an electric stove might account for barely half of that. Fewer and simpler meals eaten at home, more meals cooked with microwaves or other electric specialty appliances, and smaller household size all have reduced final cooking energy significantly from its values during the 1960s.

There is no obvious structural variable for either water heating or cooking. Hot water consumed (e.g., baths, loads of clothes) is not known to any precision, nor is the number (or kind) of meals cooked. Experience suggests that the amount of energy used for either purpose grows with the square root of household size. In comparing energy use for either purpose across countries, differences in household size can account for important differences in energy use per household. However, on a per capita basis, the effect is reversed; large households share energy uses across more members and, hence, tend to have lower per capita energy use. Thus, as household size shrinks, energy use per capita climbs. Either way, cooking and water heat use and habits explain the greater part of differences in per capita use.

There are modest differences in the efficiency of cooking appliances and water heaters. Imposition of electronic ignition for gas appliances, use of microwave or halogen cooking, and greater water tank insulation for all kinds of tanks explain small differences among countries.

3.1.5 Summary of Residential/Household Energy Uses

A comparison aggregating residential/household energy use by end use that accounts for key structural differences among countries can be devised. This permits separation of structural factors that are income driven, such as home area, from those that are endogenous, such as climate and household size. Such comparisons strip away differences not directly due to policies or technologies, thereby making comparisons more fruitful.

For example, Fig. 4 summarizes residential/household energy uses across a number of countries. (In this figure and subsequent graphs, Europe denotes the average of France, Italy, West Germany, and the United Kingdom.) Electricity is shown separately from fuels and district heat to illustrate its penetration. For comparison across a wide range of winter climates, space heating is scaled to the average number of degree-days for IEA countries ($\sim$2700 degree-days to base 18°C). This scaling alone shows how household energy use in Japan and Australia would increase while falling in Denmark and especially Canada.

The high use in the United States after normalization is explained by the largest homes and largest appliances as well as by somewhat higher than average intensities for heating and appliances. The low use in Japan, even after adjustment for winter climate, is explained by the small size of homes and low indoor temperatures (i.e., frugality). The normalization shows that Denmark's use is raised relative to that of somewhat warmer European countries, whereas that of Australia is lowered considerably by a very mild winter climate. Relatively low energy prices in Australia, the United States, and Canada boost energy use in those countries relative to the four European countries and especially Denmark and Japan.

3.2 The Commercial/Services Sector

As Fig. 1 shows, the commercial/services sector accounts for 10 to 15% of final energy use. The basic energy services include those of the household sector as well as ventilation. Power for computers, information, motors, and air conditioning in this sector is proportionally more important than that in homes, whereas energy use for water and space heating is less important. The basic measure of output is GDP arising in ISIC sectors 6 to 9. The basic measure of structure that is readily used is the total area within buildings. Some further detail of how that area is distributed among building types is known, and there are occasional surveys of energy use by fuel and building types in a handful of countries.

This sector is very heterogeneous and contains buildings that are not used primarily for homes,

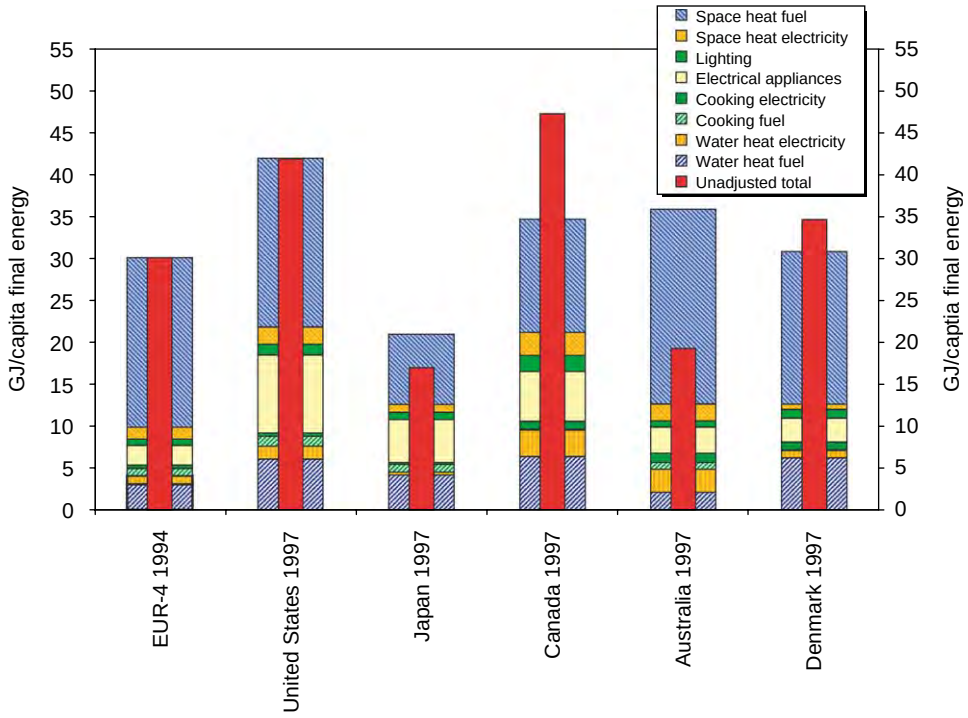


FIGURE 4 Energy use in homes, with space heating normalized to an average winter climate. See Fig. 1 for abbreviations.

industry, or farms. The economic sectors include warehousing and storage, retail and wholesale, transportation, private services (e.g., banking and insurance, hotels/restaurants), and public services (e.g., medical, educational, cultural, and sports facilities). The physical sectors by which buildings are surveyed directly might include "offices," "medical," "educational," "restaurants," "stadiums," and/or even "parking." Economic data are reported by building owners or operators, whereas physical data have to be gathered by careful surveys of large samples of buildings. Both are useful for understanding how the sector uses energy so as to facilitate comparisons. Unfortunately, neither kind of definition overlaps well with the other; much cooking occurs in buildings other than restaurants, offices are found in all kinds of buildings, and so on. Hence, comparisons of individual building types and their energy uses must be viewed with caution, although such comparisons are probably fair at the national level (e.g., all hospitals, all schools).

Unlike the residential/household sector, less is known about end uses of energy in the commercial/services sector. The reason is the enormous heterogeneity of the sector. As implied previously, laundry (and sanitary hot water, sterilization, etc.) is important in hospitals but is relatively unimportant in offices and schools, and space heating is important in some buildings but is increasingly provided indirectly in buildings, such as offices and stores, by heat from other end uses, particularly electronic machines. Comparisons of energy use based on metering studies and engineering estimates do exist, but they are very difficult to compare except among similar building types. Although a great deal is known about the efficiencies of individual motors, space heaters, lighting systems, and the like, it is virtually impossible to aggregate these into any meaningful figure of overall efficiency. Comparisons of intensities of fuel and electricity use, and of space heating use, are justified. However, beyond that, only very detailed examination of buildings through surveys yields enough information to say how efficient energy use is in any building type, not to mention across the entire sector or between countries.

3.2.1 Key Trends and Indicators for Comparison

The commercial sector's extent in square meters, and hence its energy use, has been driven by growth in the services component of GDP. Figure 5 shows this coupling clearly. Countries with large areas relative to sectoral GDP have a great deal of space (e.g., United States, Australia) or are relatively cold (e.g., Canada, Sweden, Denmark). Japan has little space because it is crowded, whereas area in Italy appears to be underreported. The large differences in floor use at a given GDP implied in the figure are clues to differences in overall sectoral energy use as suggested in Fig. 1.

Energy use in the services sector, in all but the warmest countries, was traditionally driven by the need for space heating, which was provided by fossil fuels and (in a few countries) by district heating supplied from central heat or heat-and-power stations. The same fuels generally provided water heating as well. The importance of heating is

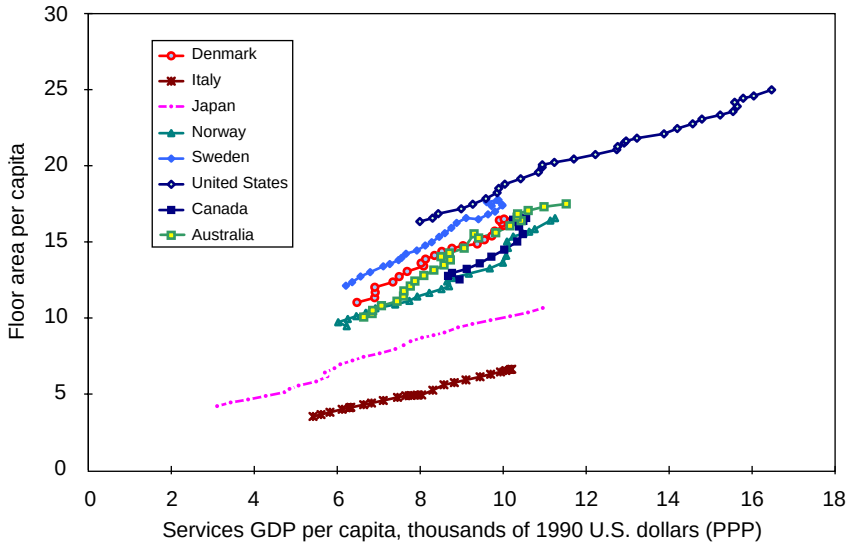


FIGURE 5 Commercial sector: floor area and GDP. PPP, purchasing power parity.

apparent from the high values of fuel/district heating per square meter for Finland and Canada from the 1970s and 1980s. The decline reflects both heat saving and switching to electric heating. In contrast, the very low value for Norway reflects the very high penetration of electricity for space heating even during the early 1970s as well as tight building shells. The waves of energy saving during the 1980s, and some substitution toward electricity, left fuel and heat use much lower, relative to area, as Fig. 6 shows. Although this indicator is inexact as a measure of space heating, detailed heating surveys by fuel and heating type that cover France, Denmark, the United States, and Sweden give approximately the same picture for each main heating fuel over time.

The decline in heating needs, coupled with growth in the demand for electricity services (e.g., lighting, information systems, ventilation, motors), has left electricity as the most important energy source, typically providing at least 40% of final energy (and much more if electric space heating is common). Figure 7 shows an interesting portrayal of electricity use in services as a function of service sector GDP. The high uses at a given GDP in Norway, Sweden, and Canada are largely a result of high space heating penetration. The sudden drop in Norway's value at

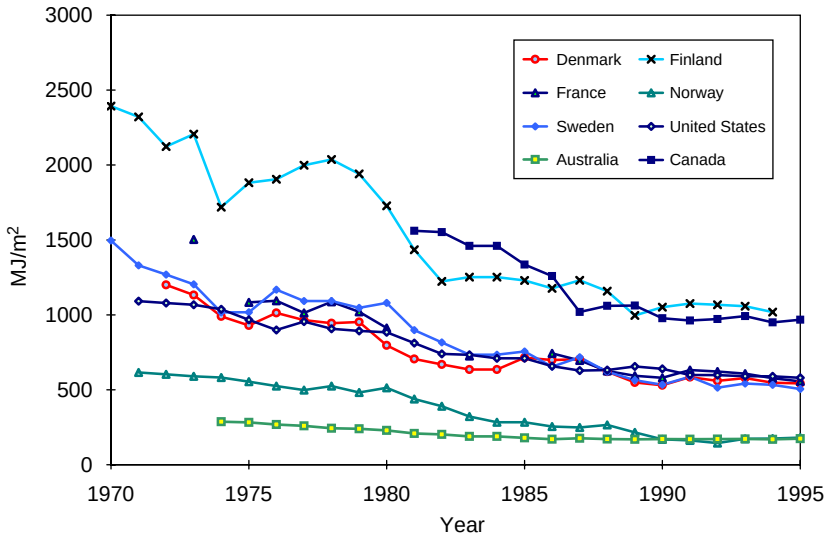


FIGURE 6 Fuel/district heat intensity in the services sector.

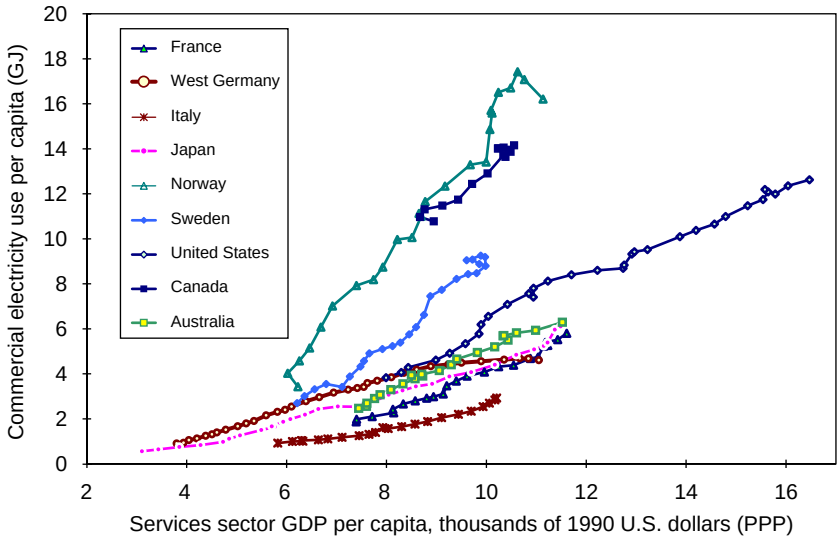


FIGURE 7 Services sector: electricity and GDP.

approximately U.S. $10,500 per capita (which occurred during the mid-1990s) arose when electricity use for heating was cut during a mild winter after electricity prices rose. No country appears to show a bulge of growth during recent years that might signal more electricity use for information and communication.

Although it is very difficult to estimate various end uses in this sector, results for space heating shown in Fig. 8 do at least suggest how space heating needs increase with increasing winter severity. The slope of this relationship should increase with colder temperatures. During mild periods, most of the heating needs of a large building can be met from waste heat from other uses. But during the coldest periods (or in the coldest countries), a disproportionate amount of heat comes from the heating systems. The unusually high position of Italy and Japan may be due to a significant undercounting of total heated area, also suggested in Fig. 5. Indeed, this is an example where international comparisons themselves point to fundamental problems with data, giving cause to reject results from some countries.

3.3 Manufacturing

Manufacturing accounts for 15 to 35% of energy use in IEA countries. Although it was traditionally the most important energy consumer before 1973, manufacturing's role in overall energy use has steadily eroded as the sector's overall output growth has slowed, energy has been saved, and other sectors have surged ahead.

3.3.1 Introduction

Manufacturing (ISIC 3) is a subset of the industry sector, which also includes agriculture (ISIC 1), mining (ISIC 2), and construction (ISIC 4) in most countries. Manufacturing typically accounts for 90% of energy use in industry unless a country has a large energy sector. Largely because these sectors are so poorly documented, most analysis is limited to manufacturing.

3.3.2 Methodology and Data

Manufacturing energy use is commonly disaggregated into six separate subsectors and a subsector that contains all remaining subsectors. In the formulation, output is measured in value added or GDP arising in each sector. The point where value is added corresponds approximately to the point where energy consumption is measured. The six subsectors are food and kindred products (ISIC 31), paper and pulp (ISIC 341), chemicals (ISICs 351 and 352), nonmetallic minerals (ISIC 36), iron and steel (ISIC 371), and nonferrous metals (ISIC 372). In a few countries with very specialized heavy industry (e.g., Norway, Sweden, The Netherlands, Australia, Finland), disaggregating chemicals, nonferrous metals, and/or paper and pulp gives additional insights into the very high energy intensities of those branches when disaggregated only this far. On the other hand, a more common aggregation— two digits at the ISIC

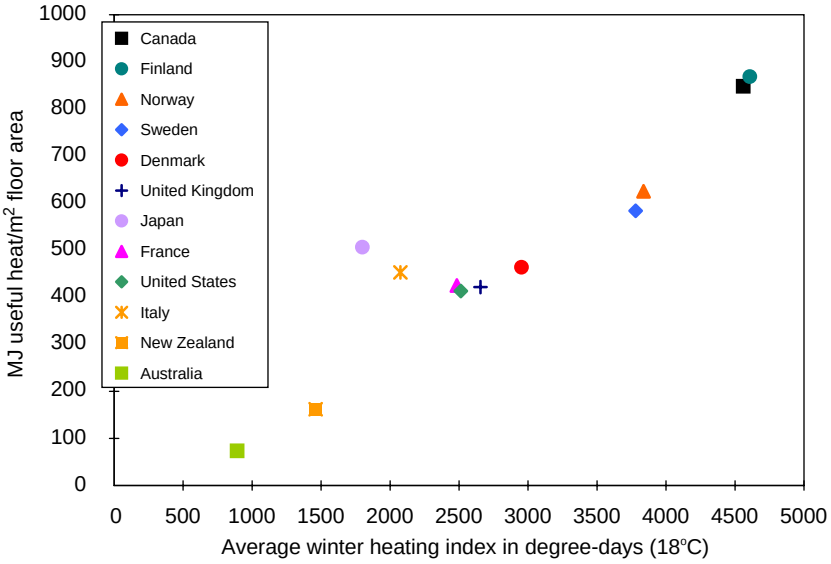


FIGURE 8 Estimated services sector space heating energy use and winter climate.

level—that combines energy-intensive branches, such as paper/pulp and heavy chemicals, with their respective less energy-intensive relatives, such as printing and rubber/plastics, obscures important features of manufacturing.

Alternatively, Phylipsen and colleagues at Utrecht University in The Netherlands and Price and colleagues at Lawrence Berkeley National Laboratory in the United States have carried out extensive comparisons and time-series analyses using data on physical output of materials and energy use for manufacturing those materials. Data on production and energy use typically cover paper, pulp, basic steel, primary and secondary aluminum, other nonferrous metals, cement, some basic chemicals, and refined petroleum. But these data are not available for all countries or for long time periods except in a few countries and industries. The economic approach is useful for explaining overall changes in national energy use, whereas the physical approach is more appropriate for understanding the energy efficiencies of basic processes. But ultimately, only the economic approach covers virtually all of manufacturing, and it is the one adopted in the long series of IEA studies.

3.3.3 Manufacturing Output and Structure

Manufacturing output, measured as aggregate value added or GDP arising, has generally risen slowly in most IEA countries, rising rapidly during economic boom times but declining sharply with recessions. As a share of total GDP across the IEA, manufacturing has fallen slightly due to the continued rise of services or, as in the United Kingdom and Norway, growth in the energy sector. However, manufacturing has not lost a significant share of GDP to services; instead, it is other sectors of industry that have suffered.

Within the manufacturing sector, important structural changes have taken place. These are measured by the share of output (as GDP or value added) in each branch, weighted by the energy intensity of that branch in a base year (rolling weights may also be chosen). In Japan, and to some extent West Germany and the United States, the energy-intensive industries have lost significant share of total output over the past three decades or so. In countries with particular niches for raw materials or cheap energy, such as Australia, Norway, and The Netherlands, the reverse has occurred. Not surprisingly, there is a wide variation of the share of output concentrated in the aforementioned energy-intensive industries, with Denmark at less than 20% and Finland, The Netherlands, Australia, and Norway at well above 30%.

These important movements and differences alone suggest that total manufacturing value added or GDP alone is not a good measure with which to compare countries. At a minimum, the shares in the several subsectors must be used for comparison. Overall differences in output per capita explain part of the difference in energy use per capita or GDP over all sectors. Differences within manufacturing—the share of energy-intensive raw materials in overall output—are nearly as important in explaining differences in total manufacturing or total energy use per capita.

3.3.4 Energy Use and Intensities

Figure 9 shows the energy intensities of the key sectors (also known as branches) of manufacturing in a number of IEA countries during the mid-1990s. The wide differences among sectors result from the intrinsic differences in the processes employed in each sector. But the large differences in intensities among countries within a sector arise for two reasons. Differences in energy efficiencies of processes do cause differences in sectoral energy intensities. But further differences in the composition of even a three-digit ISIC sector also lead to differences among countries. For example, aluminum production in Norway, Australia, and Canada is dominated by primary aluminum production, whereas in West Germany and Japan most is from recycled scrap and/or imported raw aluminum. Paper production in West Germany and Japan is fed largely by imported pulp, whereas "paper" production in Sweden, Norway, and Canada consists of a large surplus of energy-intensive, low value-added pulp that is exported. Similar differences arise in the chemicals and iron/steel sectors. Therefore, the results presented in Fig. 9 cannot be used to make widespread conclusions about energy efficiencies of processes. However, these intensities can be used to explain differences in aggregate energy use relative to output or changes over time. Again, within limits, differences in individual branch intensities can be very useful in explaining national differences in energy use for manufacturing.

3.3.5 Physical Intensities

With some care, physical intensities of production of raw materials can be derived from industry-wide data on production, energy use, and energy expenditures as well as from information on individual facilities. For example, Lawrence Berkeley National Laboratory and the University of Utrecht have carried out a series of comparisons of production and energy for steel, cement, paper and pulp, and

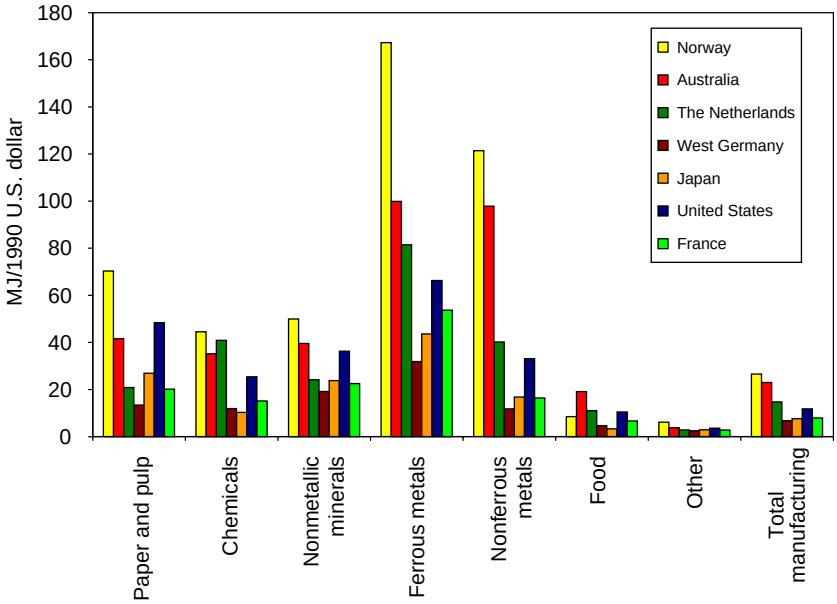


FIGURE 9 Energy intensities in manufacturing: 1995.

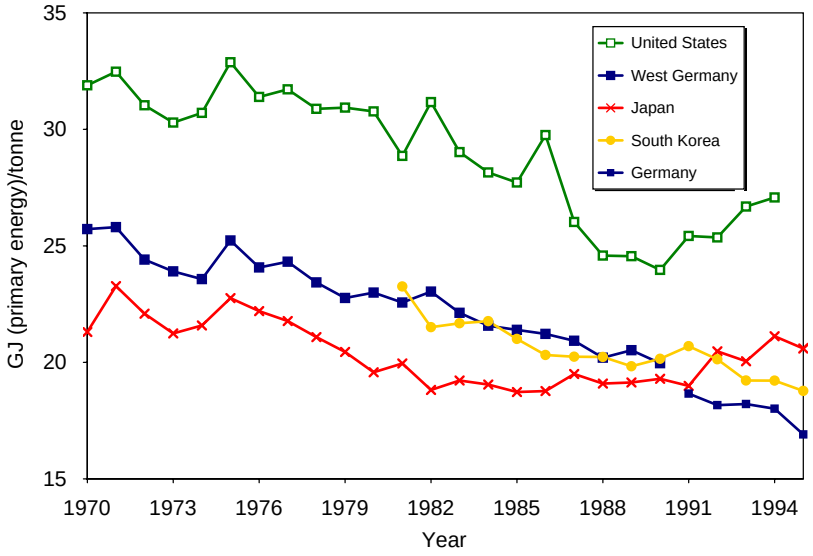


FIGURE 10 Physical intensities of steel production. Data from the Lawrence Berkeley National Laboratory and the University of Utrecht.

other important materials. The results for steel production are shown in Fig. 10.

Differences in national figures are explained by the actual efficiency of production, by the amount of recycled material used, by the nature of the process (e.g., basic oxygen vs blast furnace for steel, different pulping or cement-making processes), and by the actual nature of the final product (e.g., form of steel, type of paper). Efficiency depends on the age and size of the plant; newer and larger facilities are nearly always more energy efficient than older ones. And the level of housekeeping and vigilance in each factory also determines how much energy is used for each basic product. The differences over time are explained both by energy savings occurring within a certain kind of plant or process and by shifts toward processes or plants that are intrinsically less energy intensive and, to a great extent, more energy efficient.

An important caveat must be stated here. The fact that processes are physically more efficient in one country does not mean that they are the appropriate processes for another country, nor does it mean that one level of energy efficiency is "best." Factors such as plant size and vintage depend on many conditions, including energy prices. And, of course, energy prices also influence the level of housekeeping undertaken to save energy.

3.3.6 "Mine–Yours": A Summary Comparison

Summarizing overall differences in manufacturing energy use is tricky. Each part of the ASI equation exhibits significant differences among countries, and within a country the *I* values vary enormously from branch to branch of manufacturing. A comparison that has been developed to summarize these components is used to compare aggregate energy intensity of manufacturing (it was noted previously that output per capita itself varies by more than two to one among IEA countries). To account for differences in the mix of output, one calculates the aggregate manufacturing energy intensity that would have occurred for each country if all countries had the same shares of branch output that make up the average for 13 IEA countries but if each maintained its own branch energy intensity. The resulting aggregate intensity is shown in Fig. 11 ("IEA-13 structure and national intensity") next to the actual aggregate manufacturing intensity in 1994. Where the intensity increased in the alternative calculation (e.g., United States, United Kingdom, Italy, France), the country in question had an own-output structure that was less energy intensive than the average for all countries; where a decline occurred, the own-output structure was more energy intensive. For Australia, The Netherlands, Finland, Norway, and Canada, the difference is large, indicating the importance of the high shares of energy-intensive products. Unquestionably, the mix of branch output accounts for a significant variation in aggregate manufacturing energy intensity.

Differences in individual branch intensities can be summarized in a similar way. In Fig. 11, "IEA-13 intensity and national structure" displays the energy intensity that would have occurred in each country if it had the average energy intensity for the 13 countries in every manufacturing sector but its own sectoral output mix. In this case, an increasing energy intensity (e.g., France, West Germany, Japan, Italy, United Kingdom) compared with the actual value means that a country's energy intensity was lower than the average for all countries; a decline indicates that the country's intensity was higher than average. This is especially the case for the big producers of raw materials (e.g., Australia, Canada, The Netherlands, Norway, Finland). Note that the United States is the only country with higher energy intensity than average and, at the same time, with a (slightly) less energy-intensive structure than average within this group of countries. As with mix, differences in the

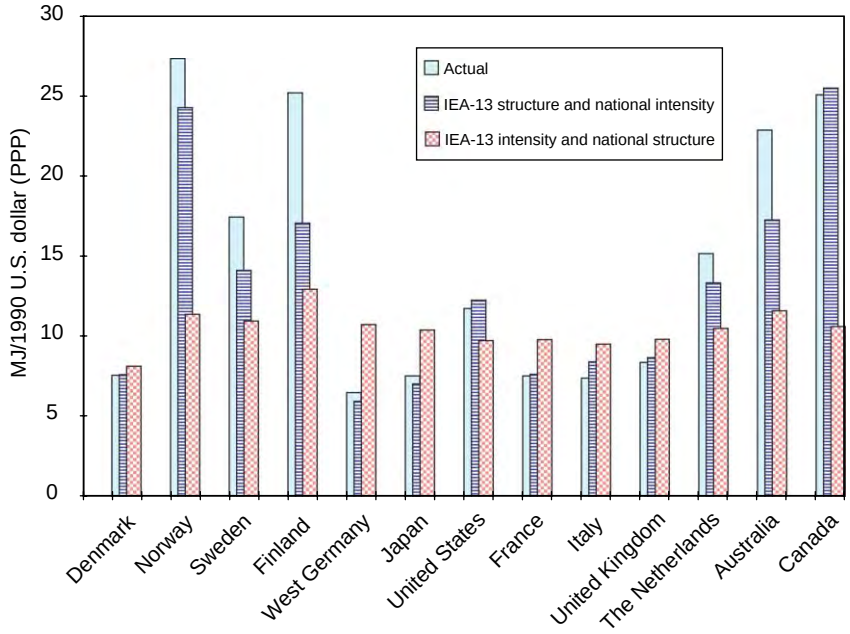


FIGURE 11 "Mine–Yours" comparison of manufacturing energy use. PPP, purchasing power parity.

energy intensities of branches also account for much of the differences in aggregate manufacturing energy intensities.

3.4 Passenger Transport

Passenger transport, or travel, measures human mobility. The most common measure is passenger-kilometers. These are measured by the distances that vehicles move multiplied by the number of people in the vehicles (e.g., 1.5 people/vehicle in private cars, 10–200 in buses, more in trains and aircraft). Roughly 80% of passenger transport is provided by people with their own cars, cycles, or feet. The rest is provided by transport for hire, most of which is air travel in the United States or intercity and commuter rail (including metros) and buses in Europe. The distribution of travel by mode measures the structure of the sector. Except for electricity used for rail and trolley buses and a small amount of CNG, liquid fuels dominate energy use in this sector.

Passenger cars dominate energy use in passenger transport. Their ownership and use grow with income, as Fig. 12 shows. Car use per car varies very little over time, so portraying car use captures mainly the effects of greater ownership. The growth in Japan at the highest income occurred because car sales spurred by tax changes continued to boom during the early 1990s even as economic growth slowed down, pushing up car use. Similarly, car use in Canada grew even when GDP per capita fell during the early 1990s. These twists illustrate the importance of looking at more than 1 year of data in an international comparison. They also show how the portrayal of a key variable against a driving factor such as income suggests important relationships that can be probed by formal statistical tests.

Portraying car use versus income shows the importance of income as a driving factor, but saturation (i.e., decreased growth in ownership or use) is clear for the countries in the upper right hand part of Fig. 12. Note that for every country shown, car use during the mid-1990s was higher than it was during the early 1970s, similar to the upward march of per capita house area shown in Fig. 2.

The wide variation at a given income between the United States, Canada, and Sweden, on the one hand, and Europe and Japan, on the other, is explained in part by fuel prices, incomes and car ownership policies (e.g., new car taxes, annual fees), and geography. Surprisingly, Sweden had relatively low fuel prices (for Europe) until the late 1980s and had very low taxation of company cars, which once represented nearly half of all new cars sold. This kind of practice probably accounts for Sweden reaching the level of car use/GDP seen by the United States, Canada, and Australia, that is, countries with even larger areas and lower fuel prices. This illustrates how a comparison points to, but does not prove, how a policy may have important impacts on energy use. A similar implication arises from the large areas of homes relative to private expenditures shown in Fig. 2.

When data on driving and the number of people in a car are combined, travel in passenger-kilometers can be developed. Together with estimates of travel from other commercial modes (including data from national travel surveys), the overall pattern of mobility can be established, as is shown in Fig. 13.

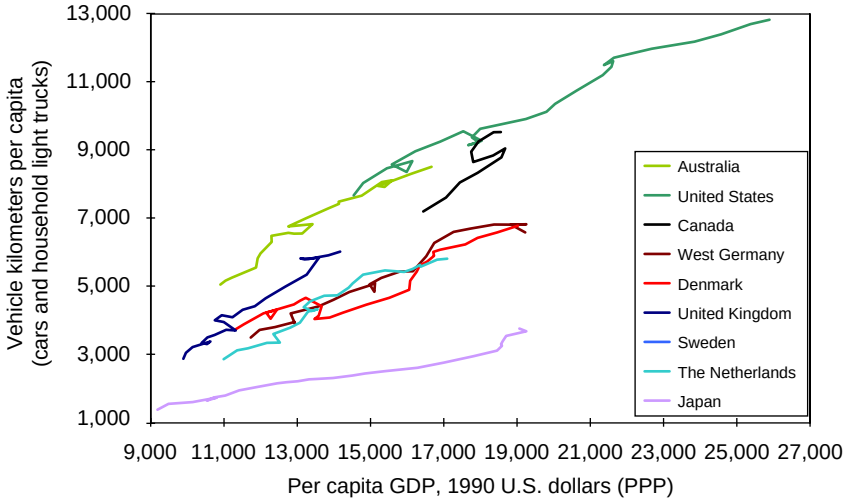


FIGURE 12 Car driving and per capita GDP: 1970 to 1998. PPP, purchasing power parity.

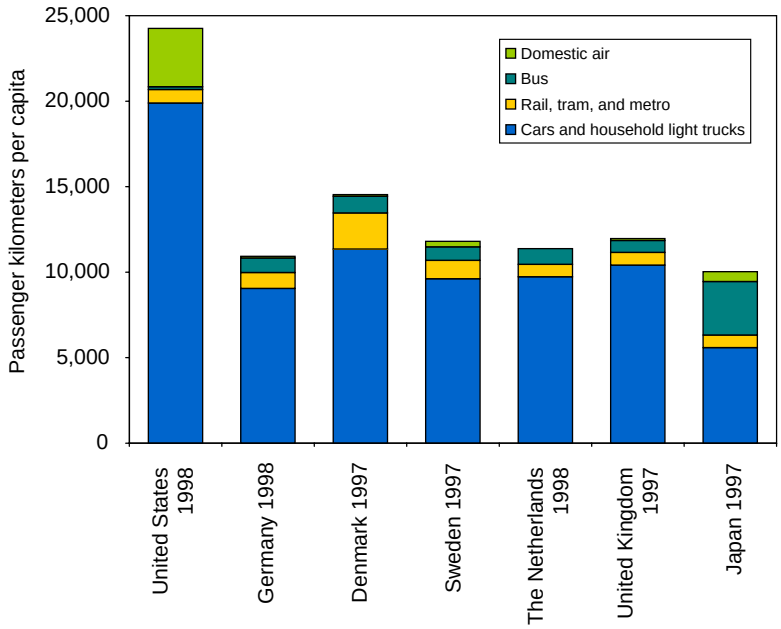


FIGURE 13 Per capita passenger transport by mode.

The high values for the United States follow both from higher car ownership relative to GDP and from higher annual driving distances. For every country shown, total travel was higher during the 1990s than during the 1970s and was increasingly dominated by cars. Indeed, the growth shown in Fig. 12, as well as some increases in air travel, accounted for virtually all of the increases in travel from the 1970s to the 1990s.

The most important energy intensity in the sector is automotive fuel economy or fuel consumed per unit of distance. Figure 14 shows that the United States has the most fuel-intensive cars, followed closely by Australia, Canada, and Japan. The latter might seem surprising, but poor traffic in Japan raises fuel intensities.

When the average loads are included, the energy intensities of car travel can be tabulated in megajoules per passenger-kilometer. These can be compared with the (lower) values for rail and bus as well as with the values for air travel, which are higher than those for car travel in Europe but lower than those for car travel in the United States. Multiplying each modal intensity by travel per capita in that mode gives energy use per capita by mode, which is shown in Fig. 15.

Although each mode serves a different purpose and has different characteristics, the difference in intensities among the modes, combined with the shifts in modes toward car and air travel, explains much of the change—usually an increase—in overall average energy use per passenger-kilometer of travel. Combined with higher per capita travel, this meant that energy use increased over the past three decades. Note that the only country in which per capita energy fell over the period shown was the United States, a result of the strong decline in the intensities of both automobile and air travel. Canada, which was affected by the same trends as the United States, had similar declines in intensities but a greater increase in car ownership and use; hence, per capita energy for travel was slightly higher in 1995 than in the earlier period. Note that by decomposing total change in energy into activity, structural, and intensity components, the analysis is able to show which caused energy use to increase and which caused it to decrease. This decomposition gives credit for policies, technologies, or efforts that led to saving energy but whose effects may have been obscured by the impact of other forces on overall energy use.

Using the data on modal structure and modal intensities, it is possible to compare transportation energy use among countries in an innovative "mine/yours" way that brings out the relative importance of these three components in much the same way as the same approach was used to explain differences in aggregate manufacturing energy intensities. However, this time the activity term is included explicitly, so the variable is per capita energy use and the results of imputing other countries' activity, modal structure, and modal intensities are compared with actual per capita energy use. Figure 16 shows first actual per capita use for a number of countries. Next,

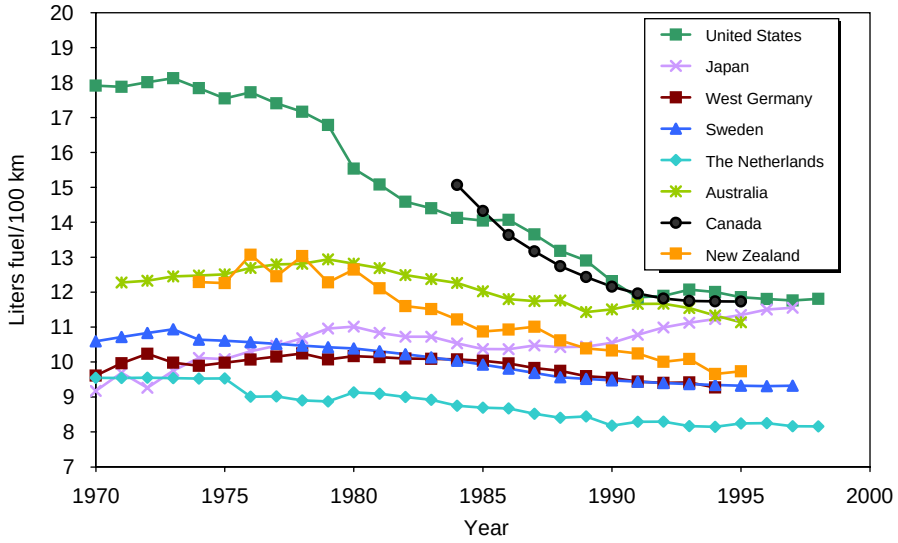


FIGURE 14 Automobile/personal light-truck fuel intensities: 1970s to 1990s.

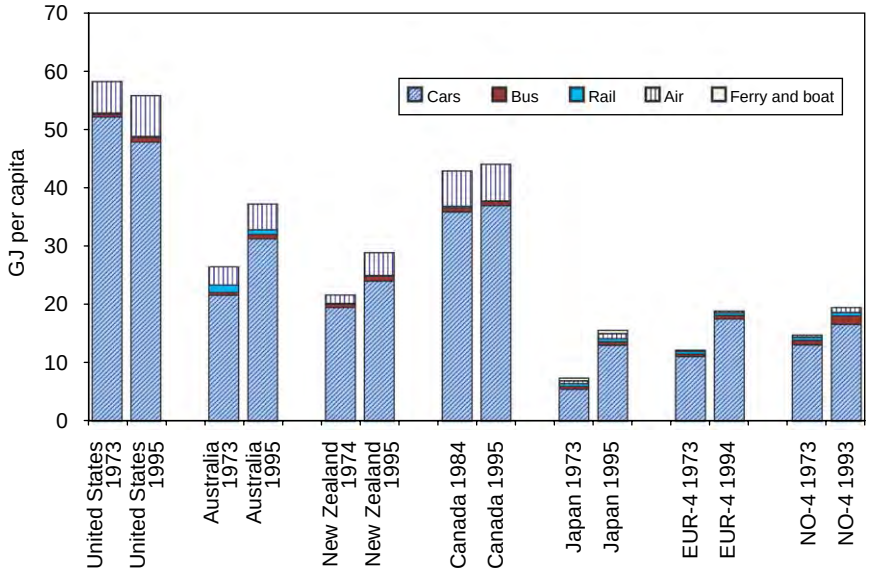


FIGURE 15 Per capita energy use for travel by mode. See Fig. 1 for abbreviations.

energy for each country is calculated as if total travel corresponded to the average value for the 13 countries considered. Then, the same calculation is carried out using the average modal mix and activity. Finally, the calculation is done for actual activity and modal mix but with the average intensities.

When a country's "total" is reduced by substitution of one set of average parameters, that is an indication that the country's own parameters boost energy use. The comparison shows that total travel and modal intensities explain most of the differences among countries. Thus, one can see for the United States, Australia, and Canada that activity, modal structure, and intensities all lead to higher energy use than the average. Conversely, giving Japan or Europe higher activity or a modal split more weighted toward cars would raise energy use in those countries. Note that with the exception of Japan, the impact of changing modal mix alone is small. This is because in every country except Japan, travel is dominated by cars (Fig. 14).

3.5 Domestic Freight Transport

Freight accounts for roughly 15% of a nation's energy use (Fig. 1), and its importance has been

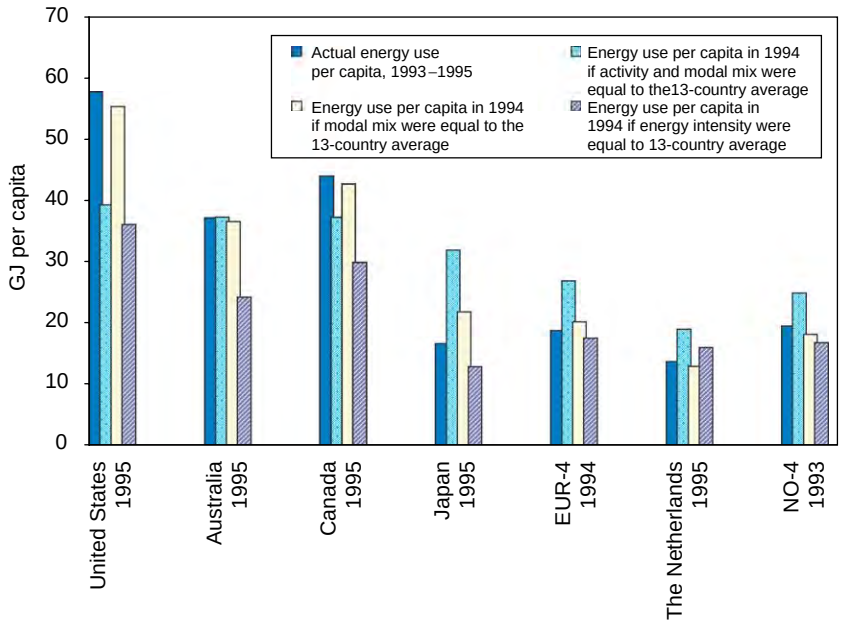


FIGURE 16 "Mine–Yours" comparison of energy use for passenger transport (travel). See Fig. 1 for abbreviations.

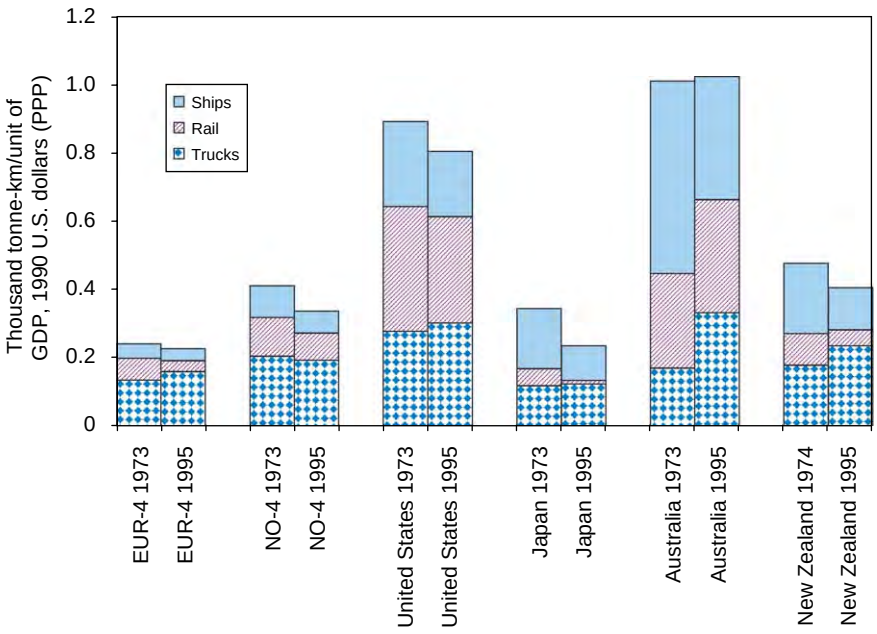


FIGURE 17 Domestic freight activity relative to GDP. See Fig. 1 for abbreviations.

rising. Output is measured in ton-kilometers. A modal mix of overall ton-kilometers—trucking, rail, and domestic shipping (coastal and river activity)—makes up structure. Freight hauled by pipeline, air freight, and freight sent overseas are excluded. The key driver for freight is GDP, which makes a convenient normalization, in addition to population, for aggregate comparison. Most energy used is for trucking, which also dominates total freight hauled in Japan and Europe.

Figure 17 gives a glimpse of domestic freight hauled in 1995 by mode relative to GDP. That normalization removes the most gross differences among countries. Clearly, the high levels of freight relative to GDP in the largest countries (e.g., United States, Australia) are related to their size. It is also

clear from further analysis that these countries produce more bulk raw materials than do the smaller countries portrayed. These larger countries also haul a much larger share of their freight on rail or domestic shipping (river barge and ship, coastal shipping). These differences explain much of the overall differences in the ratio of freight energy use to GDP implicit in Fig. 1.

Energy use for trucking has to be estimated carefully from surveys to separate it from energy use for cars and other road vehicles. Although diesel fuel dominates trucking, diesel is also important in a number of European countries for cars, so it cannot simply be assigned to trucks. On the other hand, many light trucks used exclusively for freight or commercial purposes run on gasoline, so this must also be added to energy use. For rail, it is difficult to separate fuels and electricity into that used for freight and that used for travel, although in principle national railroads could do this. Energy use for domestic shipping is well identified because marine fuels are so identified and because fuels used for foreign or overseas shipping (bunkers) are counted separately for tax purposes. Overall energy use is so dominated by trucking that if this mode is portrayed accurately, comparisons are relatively straightforward.

Trucking is the most energy-intensive mode after air freight, which still accounts for less than 3% of all freight. Trucking energy intensity depends mostly on the type and size of vehicle, use of each vehicle's capacity, and traffic conditions. In contrast, actual efficiencies of trucking (in terms of fuel used per unit of horsepower or load pulled) vary little from vehicle to vehicle due to widespread competition to sell trucks. This means that comparisons of trucking energy use either per kilometer driven or per ton-kilometer hauled are not representative of the efficiencies of individual types of vehicles. Unfortunately, few data exist on energy use per kilometer of trucks of a given size, so little can really be said about truck efficiency on the basis of international comparisons. Thus, the energy intensity of trucking is analogous to that of a manufacturing branch or space heating—related to, but not identical with, the inverse of energy efficiency.

Over time, the share of trucking freight haulage has increased, but trucking has become somewhat less energy intensive. Because trucking is so much more energy intensive than other modes, this shift raised energy use nearly enough to offset the impact of lower energy intensities. Still, trucking dominates energy use in all of the countries studied, as Fig. 18 shows.

Notice that the differences in energy use/GDP for freight among countries are smaller than the differences in levels of freight/GDP. This is because the most freight-intensive countries rely the least on energy-intensive trucking and because trucking tends to be less energy intensive along wide-open roadways than on the congested highways of the smaller countries in Europe and Japan. Thus, averaging in the 65% of U.S. domestic freight that is hauled by ship or rail to a relatively non-energy-intensive

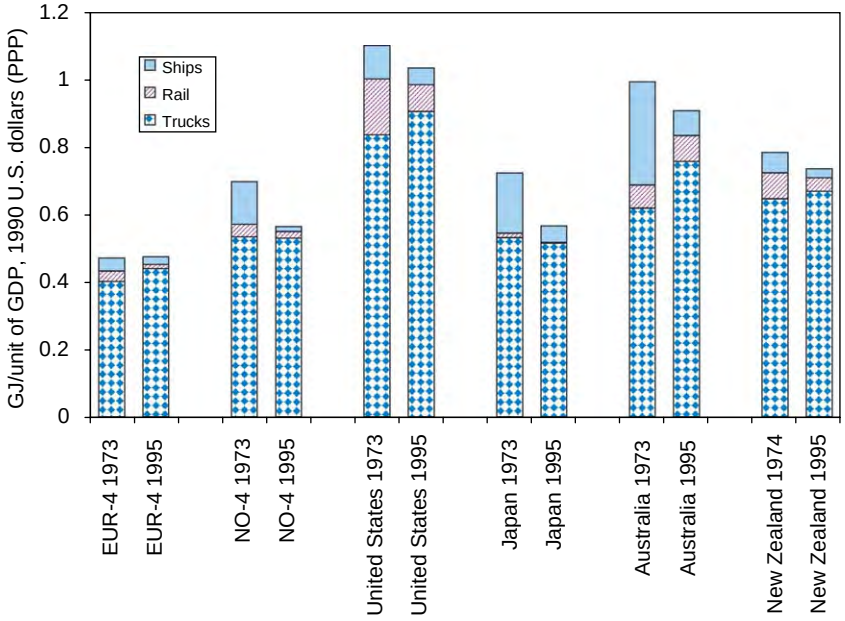


FIGURE 18 Energy use for domestic freight relative to GDP. See Fig. 1 for abbreviations.

trucking sector gives the United States a relatively low aggregate freight energy intensity and overall freight energy use/GDP that is only $2\frac{1}{2}$ times that in Europe. This is an important example often found in international comparisons: a feature that raises energy use (in this case, country size) is associated with lower energy use due to modal structure (share of trucking) and energy intensity.

4. OVERALL COMPARISONS OF ENERGY USE

It was argued early in this article that the ratio of energy use to GDP is a misleading way in which to compare overall energy use among countries. Still, it is important to tie together all of the information presented here in a meaningful way. The results can and should be related to GDP as well because wealthier nations generally do use more energy. This section summarizes a method based on the mine–yours formulation shown for travel and manufacturing that decomposes overall differences in energy/GDP ratios into structural and intensity components.

Figure 19 shows the results expressed in terms of *E*/GDP. The first column shows the actual ratio of final energy to GDP in 1995. The second column shows what that ratio would be if the average set of structural parameters for the 13 countries were used in calculating *E*/GDP. The final column uses the average energy intensities throughout the economy to repeat the calculation. (It is worth noting that in this calculation, each country is held out of the calculation of "others." Otherwise, the United States and, to some extent, Japan show little variation because these two countries dominate the averages.)

The results are not startling given the sectoral differences shown previously. Overall energy use in the United States would be significantly lower if it had the structure of other countries' energy use, due mainly to transportation, households, and services. This is important because the public perception—that manufacturing contributes to the United States' higher energy use—is actually the reverse. Both the size and structure of manufacturing restrain U.S. energy use relative to the average of the other countries. Sweden's energy use would also be lower for similar reasons because average structure would mean warmer winters and less energy-intensive manufacturing. Australia's energy use would be higher because the impact of its very energy-intensive manufacturing sector is more than offset by the lack of demand for space heating. The energy use of Japan and France would also rise due to imposing colder winters, greater car use, larger homes, and greater service sector areas. Thus, the comparisons show that structural differences are significant in explaining differences in energy use among countries.

Energy intensities are nearly as important. For Canada, the imposition of other countries' intensities would have a very large downward impact, but this is also true, to a lesser extent, for the United States,

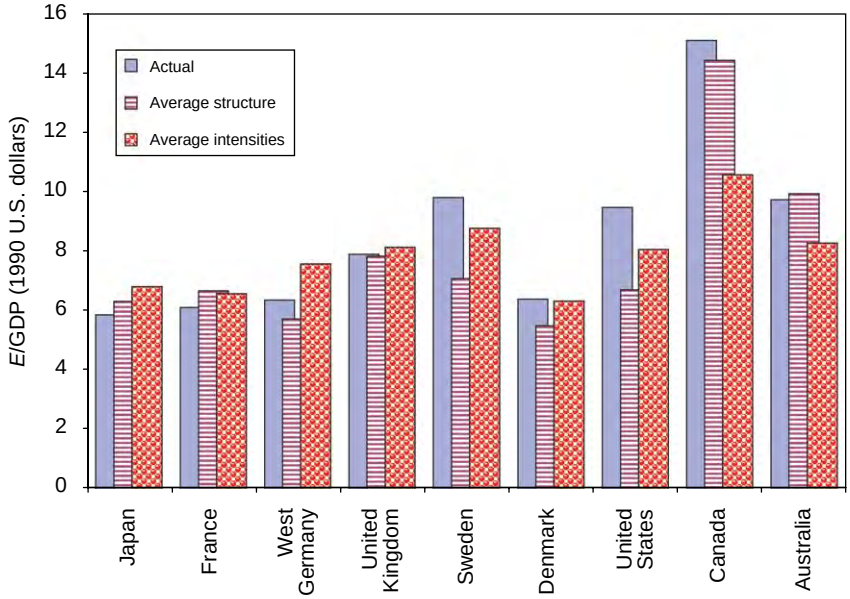


FIGURE 19 Summary comparison across all sectors and uses: mine–yours.

Australia, and Sweden. Thus, we can say that these countries have the highest energy intensities. Japan, France, and West Germany have the lowest energy intensities.

Overall, this approach illustrates how differences in the structure of energy use, as well as differences in individual energy intensities, explain differences in the ratio of energy use to GDP. Interestingly, the United Kingdom emerges as the "average" country that is affected the least by this procedure for aggregating differences.

5. LESSONS FROM INTERNATIONAL COMPARISONS

The most important lesson from international comparisons is that structural differences explain about as much of the differences in energy use relative to GDP as do intensity differences. This is not surprising given the huge variation in key structural parameters listed in Table II. Factors that drive these differences—incomes, housing and vehicle taxation policies, natural resource endowments, subsidies to one industry or another—may in fact be as important as energy policies per se and energy prices in explaining the differences in energy uses.

The second lesson is that energy intensities make acceptable, if imperfect, indicators of energy efficiencies. Considerable work is required before the efficiency of processes and devices can be evaluated. It was noted that the fuel intensity of cars is related both to characteristics consumers choose (e.g., weight, power) and to the real technological efficiency of cars (e.g., energy/weight or energy/power plus kilometers driven). Similarly, the energy intensity of space heating depends both on indoor heating habits and on the insulation of the structure or efficiency of the heating systems. And the intensity of appliances depends on their size and features, their frequency of use, and the efficiency of energy conversion to heat, cooling, motor power, and the like. Usually, but not always, intensities make good indicators of efficiencies. Thus, the comparisons of efficiencies implied by intensities must be made with caution.

Can these comparisons be extended to developing and transitional countries? The answer is a cautious yes. Data for Mexico, India, Brazil, China, and a number of transitional countries (e.g., Poland) have been presented for analysis in a number of forums. For example, Fig. 20 shows such data for Mexico. The data for the household sector come from surveys, whereas those for the productive sectors except transport come from detailed energy balances. For these sectors, structural data are also available. The difficult sectors are travel and freight. The problem for most developing counties is the lack of accurate data on the number of vehicles actually in circulation and the distances they are driven as well as a breakdown of how gasoline and diesel (and in a few cases LPG or CNG) are divided among two- or three-wheelers, cars, light trucks, heavy trucks, buses, and other vehicles.

The final lesson is the most difficult one. Collecting and processing data for these comparisons is not expensive, but it is not free. The United States, which pioneered surveys in the stationary sectors (e.g., households, manufacturing, services), has reduced its effort in both coverage and time and has abandoned most of its effort to accurately measure both activity

TABLE II

Key Differences in Economic Structure and Energy Prices: Late 1990s

- Heating degree-days (using a base indoor winter temperature of 18°C) vary from 900 in Australia and 1800 in Japan to more than 4500 in Canada and Finland. The population-weighted average for 14 countries is 2800
- Automobile use per capita varies from a high of 13,500 km/year in the United States to a low of 3900 km/year in Japan. Trucking ton-kilometers per capita vary by a factor of two, and the ratio of freight hauled/GDP varies by more than a factor of two, due to geography and other considerations.
- Home size varies from more than 155 m^2 per capita (or roughly 60 m^2 per capita in the United States, at the high end to less than 90 m^2 and closer to 30 m^2 per capita in Japan at the low end.
- Per capita manufacturing output varies by two to one. The manufacturing share of GDP varies from less than 20% in Australia and Denmark to nearly 30% in other countries, and the share of the energy-intensive raw materials sectors varies from more than 30% in Australia, The Netherlands, Finland, and Sweden to only 15% in Denmark.
- Similar considerations in the service sector reveal nearly a two-to-one ratio of built-up area to population or to service sector GDP.
- Real road fuel prices vary by a factor of three when national purchasing power is taken into account. Home heating fuel prices vary by a factor of two, and electricity prices in any sector vary by as much as a factor of three. GDP per capita varies by 1.75 to 1 across the countries considered, again when compared using purchasing power parity. Some countries permit full deduction of mortgage interest from incomes (e.g., United States), whereas others permit almost none at all. New care taxation varies from very low (e.g., United States) to effectively 200% (e.g., Denmark). Company car schemes accounted for as much as 50% of sales in Sweden and Great Britain during many periods between 1970 and 1995.

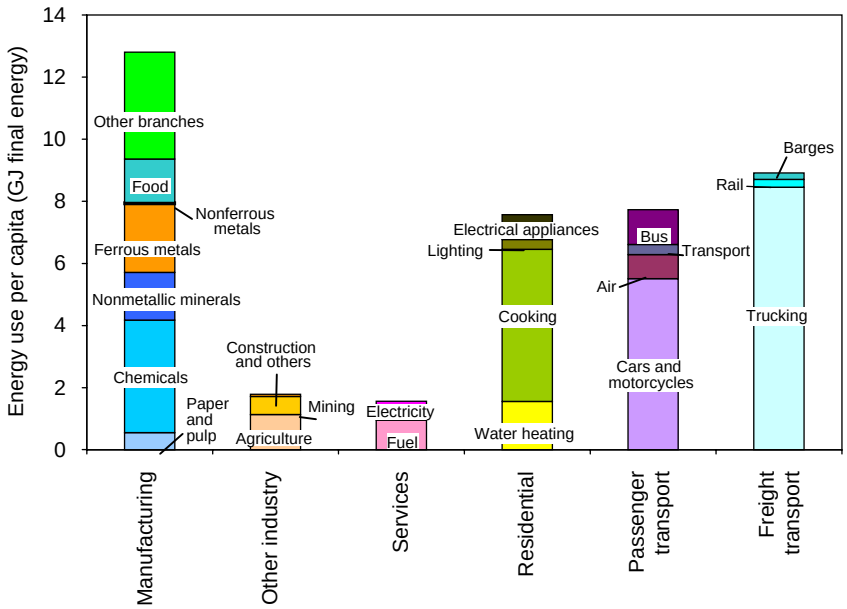


FIGURE 20 Energy uses in Mexico: 1994. From Claudia Sheinbaum, Universidad Autonoma de Mexico.

and fuel use in highway vehicles. Yet fairly accurate measures of energy use for other sectors and uses portrayed in this article are still available and even used to produce detailed inventories of carbon emissions. Other countries, notably Canada and Australia, have stepped up their efforts to institutionalize both data collection and data analysis. The IEA efforts noted previously have also stimulated a number of IEA member governments to take careful inventory of their own energy data and how they are analyzed and published.

These improvements should not stop and, in fact, should be extended to developing countries. The reason was given early in this article: a fair deal on restraining carbon emissions depends on all parties understanding the basis of each other's emissions, which for most countries are dominated by carbon arising in the combustion of fossil fuels.

Thus, the benefits of international comparisons of energy use are a better understanding of one's own consumption patterns as well as those of one's neighbors, partners, and even adversaries. The costs of those comparisons, in terms of data-gathering efforts, are relatively small given that many of the structural data are collected anyway. Even with sectoral surveys every few years, the price tag is less than one-thousandth of a country's monetary energy budget. The risks of misuse of information, particularly the confusion of intensities with real efficiencies and the assignment of "targets" for energy use based on other countries' experience, are real. But the risks of not understanding even one's own energy use patterns, and hence of missing opportunities for applying more efficient uses or for applying (or removing) policies that lead to inefficiencies or waste, may be even larger.

SEE ALSO THE FOLLOWING ARTICLES

Commercial Sector and Energy Use • Conservation Measures for Energy, History of • Development and Energy, Overview • Economic Geography of Energy • Energy Ladder in Developing Nations • Fuel Economy Initiatives: International Comparisons • Global Energy Use: Status and Trends • Industrial Ecology • Industrial Energy Efficiency • Industrial Energy Use, Status and Trends • Sustainable Development: Basic Concepts and Application to Energy

Further Reading

Adams, F. G., and Miovic, P. (1968). On relative fuel efficiency and the output elasticity of energy consumption in Western Europe. *J. Industr. Econ.* **17**, 41–56.

Darmstadter, J., Dunkerley, J., and Alterman, J. (1978). "How Industrial Societies Use Energy." Resources for the Future, Washington, DC.

Krackeler, T., Schipper, L., and Sezgen, O. (1998). Carbon dioxide emissions in OECD service sectors: The critical role of electricity use. *Energy Policy* **26**, 1137–1152.

Milstein, J., and White, R. (1975). "Comparison of Energy Use in the U.S. and West Germany." Stanford Research Institute International, Menlo Park, CA.

Phylipsen, G. J. M., Blok, C., and Worrell, E. (1998). "Handbook on International Comparisons of Energy Efficiency in the Manufacturing Industry." Department of Science, Technology, & Society, Utrecht University, Netherlands.

Price, L., Phylipsen, D., and Worrell, E. (2001). "Energy Use and Carbon Dioxide Emissions in the Steel Sector in Key Developing Countries." Lawrence Berkeley National Laboratory, Oakland, CA.

Schipper, L. (1997). "Indicators of Energy Use and Efficiency." International Energy Agency, Paris.

Schipper, L. (1997). "The Link between Energy and Human Activity." International Energy Agency, Paris.

Schipper, L., Johnson, F., Howarth, R., Andersson, B., Andersson, B. G., and Price, L. (1993). "Energy Use in Sweden: An International Perspective." Lawrence Berkeley National Laboratory, Oakland, CA.

Schipper, L., and Lichtenberg, A. (1976). Energy use and well being: The Swedish example. *Science* **194,** 1001–1013.

Schipper, L., and Marie-Lilliu, C. (1999). "Carbon-Dioxide Emissions from Transport in IEA Countries: Recent Lessons and Long-Term Challenges." Swedish Board of Transportation and Communications Research, Stockholm.

Schipper, L., Murtishaw, S., and Unander, F. (2001). International comparisons of sectoral carbon dioxide emissions using a cross-country decomposition technique. *Energy J.* **22**(2), 35–75.

Schipper, L. J., Ting, M., Khrushch, M., and Golove, W. (1997). The evolution of carbon dioxide emissions from energy use in industrialized countries: An end-use analysis. *Energy Policy* **25,** 651–672.

Schipper, L., Unander, F., Marie-Lilliu, C., and Walker, I. (2000). "Energy Use in New Zealand in an International Perspective: Comparison of Trends through the Mid-1990s," prepared for the Energy Efficiency and Conservation Authority of New Zealand. International Energy Agency, Paris.

Schipper, L., Unander, F., Marie-Lilliu, C., Walker, I., and Gorham, R. (2000). "Energy Use in Australia in an International Perspective: Comparison of Trends through the Mid-1990s." International Energy Agency, Paris.

Schipper, L., Unander, F., and Ting, M. (1996). "Energy Use and Efficiency in Denmark in an International Perspective." Energistyrelsen, Copenhagen, Denmark.

Unander, F., Karbuz, S., Schipper, L., Khrushch, M., and Ting, M. (1999). Manufacturing energy use in OECD countries: Decomposition of long-term trends. *Energy Policy* **27,** 769–778.

Unander, F., and Schipper, L. (2000). "Trends in Norwegian Stationary Energy Use: An International Perspective." International Energy Agency, Paris.

International Energy Law and Policy

THOMAS W. WÄLDE
Center for Energy, Petroleum and Mineral Law and Policy, University of Dundee
Dundee, United Kingdom

1. **Introduction**
2. **Organization of Petroleum Exporting Countries**
3. **International Energy Agency**
4. **Energy Charter Conference and Secretariat**
5. **International Atomic Energy Agency**
6. **United Nations**
7. **OECD Nuclear Energy Agency**
8. **European Union**
9. **Conclusions**

Glossary

Energy Charter Conference and Secretariat The mission of the Energy Charter is to promote open, efficient, sustainable, and secure energy markets and a constructive climate conducive to energy interdependence on the basis of trust between nations.

International Atomic Energy Agency (IAEA) An independent intergovernmental science- and technology-based organization, in the United Nations family, that serves as the global focal point for nuclear cooperation.

International Energy Agency (IEA) The energy forum for 26 countries whose member governments are committed to taking joint measures to meet oil supply emergencies. The IEA collects and publishes a wide array of energy and environmental data and research reports.

Nuclear Energy Agency (NEA) A specialized agency within the Organization for Economic Cooperation and Development, an intergovernmental organization of industrialized countries. The mission of the NEA is to assist its member countries in maintaining and further developing, through international cooperation, the scientific, technological, and legal bases required for the safe, environmentally friendly, and economical use of nuclear energy for peaceful purposes.

Organization of Petroleum Exporting Countries (OPEC) Eleven-nation organization, the members of which supply a major share of global oil; members collectively agree on global oil production and/or price targets. Members include Algeria, Indonesia, Iran, Iraq, Kuwait, Libya, Nigeria, Qatar, Saudi Arabia, United Arab Emirates, and Venezuela.

World energy laws and policies are the responsibility of numerous international agencies, including the Organization of Petroleum Exporting Countries, the International Energy Agency, the Energy Charter Conference and Secretariat, the International Atomic Energy Agency, the United Nations and its agencies, the Organization of Economic Cooperation and Development's Nuclear Energy Agency, and the European Union. Understanding the development and operation of these organizations is integral to understanding the role and effectiveness of each organization in promoting sustainable development.

1. INTRODUCTION

The concept of sustainable development now dominates the natural resources, energy, and environmental discourse with its accommodating notion of developing resources now, without compromising the future. Sustainable development suffers from an imbalance: rhetoric overwhelms action, affirmation of moral values overwhelms implementation, and good intentions prevail over getting good results. Results can be achieved only if the "game" is properly understood: this requires understanding the context of an action, including the economic, social, and cultural elements of the relevant environment; the concepts, which are usually culturally formed in some relationship with the often historical and possibly obsolete economic and technical

 © 2004 T. Wälde. Published by Elsevier Inc. All rights reserved.

contexts; and the roles, parameters of action, interests, and dominant perceptions of the main players.

Although sustainable development can perhaps be described as ideological input, activities such as technical and financial assistance and work with rules (formulating, implementing, revising) are the output of international agencies. This output is one of the factors that determines if, how, and to what extent and in what shape and form sustainable development leaves the realm of morality and ideology and enters the realm of socioeconomic and organizational reality. Very little practical research seems to have been done on how international agencies participate in the emergence of rule systems in the field of energy. The analyses here are meant to sketch out observations on the role of selected key international agencies in the field of energy law (policies and regulation). The outcome should help to understand somewhat better than before how sustainable development can be achieved more effectively through the activities of international agencies in the energy sector. The emphasis is not only on the formal mandates, a popular approach in the examination of international organizations by public international lawyers, but also and perhaps with greater emphasis on the "real-life" operations and strategies pursued by such agencies.

International organizations play a central role in the emergence of international energy law. They represent a response to the globalization of the regulatory challenges. As national regulation loses its grip, international agencies emerge to mirror the global scope of action of relevant actors (multinational companies, banks, nonstate actors, and criminal groups), but also the global scope of problems, i.e., transboundary environmental impact, and the need to create more level playing fields for competition in global trade, development, and administration of a system of global legal disciplines to ensure that the wealth-creating machine of the global economy can function properly, including its social safeguard. International agencies in this area are therefore more a prototype of true international regulation; they help member states to regulate better by providing model guidelines, collective and better regulatory intelligence, and a forum of dialogue and collegiality. As we shall see, they are primarily intergovernmental collegial networks. This, in turn, helps national regulation to be more aligned. National administrations usually establish formal and informal channels for collaboration as part of their participation in international agency activities (conferences, workshops, elaboration of model codes).

There is an explicit, visible role of international agencies in the regulatory process—servicing the negotiation and administration of treaties, formal elaboration of standards and guidelines, and, though rarely, direct regulatory power. Elaborating technical standards, usually in collaboration with experts in governments, companies, and industry associations, is a latent, but key, direct regulatory role. Such standards may be directly binding, but in most cases they have an indirect legal effect: as standards of due diligence for defining civil responsibility, by explicit incorporation into contracts or by reference in national or international (e.g., General Agreements on Tariffs and Trade/World Trade Organization-based) law. They also work by providing national governments, domestic experts, and social forces with a ready-made set of usually well-researched, tested, and often elsewhere-applied rules.

But international organizations also have an implicit, informal, and less visible role: Serving as a forum for discussion, i.e., an actor (often with some autonomy and independent dynamics) with agenda-setting power, international agencies also provide legitimacy and a benchmark for national regulatory action. There is a feedback process between the national level, the nongovernmental organization (NGO) and business level, the authoritative media (foremost the *Financial Times*), and the international level. Innovations and new concepts jump from the domestic level to the international agency level. They are here processed, refined, and negotiated into a more universally accepted format. They then leapfrog by way of an agency endorsement to the level of national regulation. At times, international agencies acquire true international regulatory competencies. The European Union (EU) is the best example. The new EU energy law emerging out of the 1996/1998 electricity and gas directives and its current amendment proposals illustrate the balance between true EU Commission powers and national regulatory powers to be exercised within the EU energy law framework. The more technical the subject and the more homogeneous the membership, the more it is likely that international agencies obtain true and direct regulatory powers, often exercised jointly with national enforcement authorities. The International Maritime Organisation (IMO), the World Intellectual Property Organization (WIPO), and other specialized agencies illustrate this angle. Here, there are superior technical and efficiency reasons for transferring direct regulatory powers, and less

"sovereignty" reasons related to more politicized issues for keeping such power at home.

International agencies serve as vehicle by transnational coalitions (e.g., national ministries, international business associations, and international agency officials and their expert networks, or, with a similar structure, national environment agencies, international NGOs, and the environmental agency network) to impose or influence by persuasion of national regulation, usually to counter the stronger influence of domestic pressure groups (protectionist business or labor lobbies; ideological pressure groups). The agency, with its experts, funds, and machinery for consultancy work, meetings, and publications, serves as forum for the elaboration of common rules by such specialized communities. The national representatives then return to their domestic regulatory process with the added legitimacy and technical and informational underpinning gained in the international agency process. The international agency, so to speak, is the nest to which and from where "birds of the same feather" fly into their domestic political and regulatory context. This leads to another nonconventional observation.

One generally ignored quality of international agencies is their scapegoat role for domestic politics. There is nothing more convenient than putting the blame and political responsibility for necessary action, which is strongly contested by vociferous interest groups, on international agencies and international obligations. The EU is the perfect example of national civil servants negotiating a regulatory agenda with the EU Commission staff. National politicians then blame an external and therefore intrinsically unpopular organization for the result they have all designed jointly. The current politicization of the most visible international agencies [World Bank, International Monetary Fund (IMF), World Trade Organization (WTO), Organization for Economic Cooperation and Development (OECD)], resulting in agitation by the inchoate and disparate forces of the antiglobalization movements, is a consequence of governments redirecting protest to faceless international agencies.

The primary and traditional function of international agencies is to provide the secretariat services for conferences, out of which emerge regulatory instruments such as multilateral treaties. The WTO services the General Agreements on Tariffs and Trade (GATT)-related treaties (and subsidiary instruments such as protocols and understandings), the Energy Charter Secretariat, the protocols and decisions of the Energy Charter Conference, and the IMO for various offshore pollution and dumping protocols. International agencies always proclaim that they are mere "servants" of the governments meeting in conference. This is an essential element of the official self-presentation. They will use such defense in particular when attacked by outside forces (the press, NGOs) for disliked treaties, treaty administration, and their mandate-based activities. This is formally correct: Intergovernmental organizations consist of a "secretariat" providing faceless "clerical" services to their governmental masters: preparing studies, conference services, support of negotiations, and reports on implementation. Formal decision powers, treaty-making, and the making of subsidiary law under treaties rest, as a rule, with the conferences or governing bodies composed of government representatives—economic ministry delegates in the case of the OECD, finance ministry delegates in the case of the World Bank (here acting through the World Bank executive directors), environmental ministry delegates in the United Nations Environment Program (UNEP), and so on.

This view of the role of the organization is usually described in the charters and other legislative documents setting up the organization. But such defensive escape into anonymity, though formally correct, also intentionally ignores the reality. An international organization, while professing to be only the servant of its governmental masters, develops within a short while a life, interests, external constituencies, and an organizational dynamic of its own. Pure formal descriptions focusing on founding instruments and organizational self-representation are therefore unable to capture the reality of an international agency and its role in the evolution of energy law. International agencies foremost develop a corporate self-interest in perpetuating themselves, even when their original purpose has evaporated. Because they are usually not set up with explicit "sunset" clauses, i.e., a finite end, like a task force, they will continue to exist even when the original purpose no longer exists nor requires the organization's continued existence. International agencies will therefore work hard at pretending that the original purposes still exist and are valid and will justify the considerable expense, usually by membership fees, to maintain them. They will, in addition, adapt to respond to issues they see as suitable for their survival. The fact that international agencies leap like lemmings, usually all together, after the most recent fad, reflects this survival instinct. A study of the self-justifications of agencies would lead to the identification of a "life cycle" of fashionable terms,

with innovation, reluctance, gradual acceptance, enthusiastic pursuit, widespread use, and finally obsolescence and quiet disappearance.

International agencies will develop constituencies (in more fashionable terms, "stakeholders," e.g., environmental ministry delegates, academic experts, consultants, or environmental NGOs) that will support the continuous existence of the agencies and the need to fund them. Because few if any agencies are subject to an independent assessment to gauge the continuing need for them, they are very hard to eradicate once established. Most informed assessment is partisan, and a nonpartisan assessment is usually uninformed, because it is without access and resources for an in-depth assessment. To develop a proper understanding of the role of international agencies is therefore quite difficult, if not impossible. It is understandable why most of the international law literature does little more than regurgitate statutory instruments and organizational propaganda rather than attempting an independent examination with the aim of at least coming closer to an image of reality.

The work of international agencies is closely related to the functioning of efficient regional and global energy markets. Only the market mechanism can unleash the human forces of innovation and efficiency. The market mechanism has, if left to its own devices, shortcomings (externalities, failure to provide desirable public goods, excessive volatility). The task of regulation, and, therefore, for regional and global markets, of the protoregulatory functions of international agencies is to deal with these deficiencies, i.e., to strengthen the effective functioning of international energy markets, to internalize externalities, to take care of truly "public goods," and to smooth out, to the extent feasible, excessive volatility. There is an ever-present risk of overregulation: "state failure" is the other side of market failure. Overregulation imposes excessive transaction costs, prevents innovation, and thus damages the capacity of society to create both economic and technological resources to deal with the challenges of sustainable development. In international organization, the risk of looking to state-led solutions is greater than in a national regulator: the international agency is twice removed from democratic legitimation, almost fully removed from the pressure of competitive markets, and interacts mainly with like-minded constituencies, i.e., governmental delegates, NGOs, and agency-selected experts with a natural proagency and usually proregulatory bias. This inherent weakness of international organizations is to some extent mitigated by the reduced effectiveness of their action, in particularly of the mainly "talk-shop" organizations.

In the remainder of this discussion, the relevant international agencies are reviewed, with comments on their roles and functions. Due to space constraints, only the international agencies of most significance for energy are discussed.

2. ORGANIZATION OF PETROLEUM EXPORTING COUNTRIES

In attempting to deal with truly "international" energy law, the Organization of Petroleum Exporting Countries (OPEC) is the international organization with the greatest impact on the oil sector; OPEC also has the greatest influence on energy and energy-related environmental questions, and not only on production and trade, but also on investment. OPEC countries currently control about 75% of the world's oil reserves and 40% of oil production. What is more, most low-cost oil is produced by OPEC countries. This means that a prolonged slump in prices will tend to enhance these market shares. [Low oil prices will also tend to increase the efficiency in high-cost areas (e.g., the North Sea). Significant results in terms of lowering production cost have been achieved by the Cost Reduction Initiative in the New Era (CRINE) incentive in the United Kingdom, for example. This incentive for efficiency is not present in the Middle Eastern low-cost producers. But the difference between, say, North Sea and Saudi production cost is still very large.] High-cost producers also tend to deplete their reserves much more rapidly. This means that the lower the price, the greater the subsequent enhancement of the future OPEC's market share.

The suppression of the existence of OPEC found in the Western international law literature reflects a Euro- or U.S.-centric perspective. OPEC is an actor that does not fit well into a traditional perspective given that it has been seen, since the 1970s, as at least coresponsible for high oil prices, nor into the modern "green" perspective, because it pushes, naturally, for hydrocarbon consumption. The fact that a high oil price has as much an energy-efficiency impact as most other measures undertaken by Western government does not help—it is a high oil price imposed by the interest of producing countries and not on, as the EU states, large gasoline excise taxes, for, at least ostensibly, environmental reasons. It does also not fit into the mind-set of Western NGOs: OPEC does not conform with the dominant view of developing countries as victims and passive beneficiaries in need

of high-minded Western support and guidance, but it is rather an independent actor outside Western control. It is also barely, if at all, the object of NGO attention or exposed to NGO pressure. NGOs play virtually no role in the political systems of almost all OPEC countries.

OPEC was founded in the 1960s through a Venezuelan–Saudi initiative to reduce dependence on the international oil companies. At the root of OPEC's formation were pricing disputes with the companies and United States import restrictions. The main objective was to establish a common front vis-à-vis the oil companies with respect to several fiscal issues ("expensing royalties" in lieu of crediting them against tax; calculation of the proper reference price for oil taxes and royalties). Its objectives evolved in the 1960s in response to U.S. government and oil company action moving gradually from a focus on tax matters to oil pricing. First put forward as a developing country producers' association to develop links, information, and discussion, it supported the efforts of major producing countries in the 1960s and 1970s (Iran, Gulf countries, and Venezuela) to take over foreign-owned operations and to set prices. OPEC Resolution XVI.90 of 1968 formulates a common agenda and the legal position of producing countries. This crucial, now quite dated, resolution also expresses explicitly the otherwise implicit OPEC policies restricting acreage access by the international oil companies; the principle is still alive in the Middle East, but has been disregarded in Indonesia, Venezuela, and Nigeria. In terms of impact on international law, that was perhaps the period of the major impact: OPEC's position was fully aligned with the then-prevailing "Third World" claim for a "new international economic order (NIEO)" with "permanent sovereignty over natural resources" and exclusive jurisdiction over foreign investment, including the right to nationalize. OPEC's success then fed directly into the formulation of the NIEO resolutions of the United Nations General Assembly, and those in turn provided and amplified legitimacy to OPEC actions, without much attention paid to the fact that the increase in the oil prices then affected developing countries most.

With most of the upstream oil and gas reserves and production in the hands of producing countries' state enterprises by 1980, OPEC became the forum for dialogue, and sometimes collective action, by its members. The very early days of OPEC focused on royalty/tax questions, but by the 1970s, the focus was price setting. This was later replaced by the use of production quotas for each country. The intention then was that through the change of the production quota in response to market conditions, prices within desired price bands could be achieved. With oil revenues dramatically increased after 1973 and again 1981, the OPEC Special Fund for development aid was set up.

OPEC is an international organization headquartered in Vienna; there is a staff of about 120 (with 25 senior professionals from member countries). The secretariat does not, unlike other international agencies, seem to have developed a major autonomous role. It is engaged in research, monitoring of market trends, and technical preparation of the meetings of ministers. Attempts to acquire a larger role of its own seem to be regularly rebuffed by members. The International Energy Agency (IEA), for example, born as a reaction to the formation of OPEC, has a much larger, much more serious, and even more autonomous role and outreach. It has built up a reputation for high-level competence, probably due to the autonomy given by its members and the emphasis on professional competence. An important part of OPEC's mandate is presenting OPEC's view to the public, and, in particular, developing informed response to the criticism of OPEC, mainly in Western countries—a function of bundling the expertise in and for OPEC countries and providing an "OPEC view" to the world. But even so, it seems that the major countries (in particular, Saudi Arabia, able to rely on ARAMCO, the world's largest oil company) do not fully trust the OPEC secretariat, either with respect to sharing intelligence with it or using its work, but instead operate their own oil market intelligence machinery. The use of OPEC by its member states is quite limited. Perhaps given the great differences among them (with nothing uniting them other than oil dependence), OPEC member states seem to use OPEC much less for jointly building strategic expertise and negotiating capacity and for developing a professional network for intelligence, standard-setting, and policy formulation than, for example, the much more homogeneous industrialized countries grouped in the IEA. The studies that are presently being carried out by the World Bank on petroleum revenue management, or work on petroleum law, contracts, taxation, safety regulation, and environmental protection, would constitute the natural mandate for an organization of oil producers. But member states seem to prefer to limit their solidarity to the essential core, i.e., production and pricing policies, rather than build up OPEC as a common producing countries' research, think-tank, and standard-setting

organization, as is the function of international agencies such as the OECD, IEA, or Nuclear Energy Agency (NEA). It seems that the organization, including its system of decision-making, its mandate, its staffing, and its budget level, is much in need of modernization to adapt to post-2000 globalization. This is a world that is very different from the confrontation with the "Seven Sisters" in 1959, which was at the root of OPEC's foundation. OPEC (or better, its member states) would benefit from carrying systematic benchmarking of the structure, organization, recruitment, and funding mechanisms in place in organizations such as the IEA.

The posts of the Secretary-General (mainly 3 years) and of the Chair of the board of governors (1 year, on alphabetic rotation) are filled in relatively short-term rotation. The principal organ is the Ministerial Conference; management is under the authority of the board of governors. Special issues (in particular, monitoring of the oil price and production quotas) is handled by specialized ministerial committees. The Economic Commission reviews reports and is responsible for technical recommendations for the ministerial council. Its budget (1995) was about $15 million (U.S. dollars). From all accounts, the organization seems to have little influence on the oil policy per se. There is little evidence that the secretariat brokers compromise in the typically very fractious dialogue. OPEC is also hampered by the fact that more significant secretariat positions can be filled only by nationals from OPEC member states. Because OPEC has, in comparison to the IEA, for example, a much smaller number of member states, this is one constraint (there are others) for its staff work. The pool of qualified professionals is therefore much smaller, because often in international agencies, member-state influences play a large role in recruitment. Member-state control over OPEC activities (including recruitment) reportedly goes much further than in other international organizations. This may be due, *inter alia*, to the small number of member states and the sensitivity of oil production to those states, but also to the prevailing culture of caution in the most influential member states. The fact that OPEC members are, apart from being major oil producers, much less culturally homogeneous than, say, the OECD and IEA members, may also be a hindrance to the full development of the inherent potential of OPEC's secretariat. It seems that OPEC does not grant to its staff the status of independent civil servants. Though this principle is, in the other international organizations established, not always taken fully seriously in practice, it does help to establish a professional culture. Long-term, contractually tenured, international civil servants tend to develop an "esprit de corps" and identify with the organization rather than with their country of origin. Civil servants, on the other hand, who are appointed on relatively short- and fixed-term contracts on the recommendation of their sponsoring home governments, will naturally have their prime loyalty to the home government.

In spite of such constraints, OPEC has developed considerable expertise in the operation of petroleum markets, probably much through links with the very large state oil companies in its member countries. It publishes regularly the *OPEC Bulletin*, the *OPEC Review*, and the *Monthly Oil Market Report*; these bring together know-how, with very much a focus on authors from its member states. OPEC, as an international organization, provides and services a forum for discussion and determination of concerted action by the member states. Such action consists, in the main, as of 2002, of production quotas per member state plus their adjustment in response to oil prices. Currently, OPEC pursues a policy of increasing production levels when prices over a period exceed a specified price band, and lowering production levels when prices decline below the specified price band (the farewell open letter by the then outgoing Secretary General Ali Rodriguez Araque, available from the OPEC Web site, is of interest in this regard).

OPEC's current role in the evolution of international energy law is marked by two key issues. First, the organization was founded on the basis of the producing countries' natural interest to increase and stabilize revenue, i.e., "mineral rent" from its control over oil and gas resources. That is still its raison d'ętre. But there are tensions between short-term maximization through price versus long-term strategies centered on market share for OPEC oil and gas as against non-OPEC competitors and nonhydrocarbon alternatives. Here, there is a conflict between the high excise taxes of Western (in particular, EU) governments and OPEC policy. [OPEC estimates that the Group of Seven (G7) nations in 1996 obtained oil tax incomes totaling U.S. $270 billion, while OPEC petroleum export revenues were U.S. $160 billion.] High excise taxes (up to four times or more the price in gasoline in the United Kingdom, for example) have an environmental justification. They internalize external costs to the environment, and by road traffic. But they are also a convenient cover for large tax income to compensate for the more visible lowering of income tax rates. The OPEC-consuming

country conflict is not about a higher price for petroleum-based energy, but rather about who gets most of it. The EU and the Unite States have tried to deflect political blame for high gasoline prices onto OPEC. They have, however, not been ready to accept the OPEC interest in stabilized oil prices and a "fair" OPEC share of the mineral rent on a formal negotiating agenda. Second, OPEC is naturally disinclined to view with favor the use of Western (very much currently EU) government policy to use heavy pressure to move away from hydrocarbons in favor of renewable energy sources, because this would devalue the OPEC reserves. But such policies could go hand in hand with a price- and production-based supply restriction.

Climate change is not a mid-term threat to OPEC countries if production (based on investment) is kept in balance with demand. The OPEC policy of implicitly keeping controls on investment and explicitly on production is quite compatible with the more extreme antihydrocarbon positions (exiting from hydrocarbons by restricting supply) taken by NGOs such as Greenpeace. OPEC policy can be interpreted not only as a price stabilization (increase) policy, but also as a conservation policy, in the sense of Article XX of GATT. But current Western thinking is not favorable to the use of regulatory instruments (trade, investment, and pricing rules) to smooth pricing volatility. The history of the largely failed commodity stabilization instruments of the 1970s and 1980s does not encourage new tinkering with similar instruments. Country-based income stabilization is another matter. In a volatile industry, it makes sense to skim off surplus in rich years and add to invested funds to add income in lean years. Many, if not all, OPEC and Western producing countries (Alaska, Norway; Kuwait; Abu Dhabi, and Venezuela) have developed different types of oil income funds. These are now proposed or established for new developing-country producers. In essence, income is stored away and made more difficult to access except in cases of emergency or severe budget pressures due to historically low oil prices. Such income stabilization may not make oil prices less volatile, but would make low price periods easier to suffer.

The role of OPEC is also likely to come under scrutiny from WTO law; several OPEC members are now in the process of accession. GATT/WTO obligations do not apply to OPEC, which is not a WTO member, but to its member states. Production quotas such as the ones currently used are "export quotas" under GATT Article XI. One justification may be under Article XI(2)(b), i.e., measures "necessary for international marketing of commodities"; the next defense could be found in GATT Articles XX and XXI. The question is justifiability under Article XX, mainly XX(g), conservation of exhaustible natural resources, or XX(h), pursuance of obligations under any intergovernmental commodity agreement that conforms to criteria submitted to the WTO parties and not disapproved by them. Acceptance as a legitimate measure under an international commodity agreement might be one way, but it is unlikely that governments, unless a comprehensive deal is struck with OPEC, would approve such arrangements at present. OPEC quotas are intended primarily to maintain and increase price levels. Do they have a conservation function? It is at present not a primary rationale for these measures, but it can be seen as a secondary justification. Conservation is certainly the effect of a higher price and government-induced limitation on production. But the condition is that such conservation measures must be applicable equally to domestic production. That this is done currently is questionable, but needs more in-depth investigation. OPEC countries would also rely on the national security exception (GATT Article XXI); acceptance of this is far from certain, but the concept has been interpreted by major trading countries (United States; EU) very widely. The dependency of OPEC countries on oil production, not comparable to the role of oil in other countries, would be an argument. GATT does not include any formal reference to "permanent sovereignty over natural resources" (UN General Assembly Resolution 1801 of 1962) or "energy sovereignty" [Energy Charter Treaty (ECT) Article 18]. But this principle could be seen as controlling or at least influencing the interpretation of the national security and conservation exceptions (GATT Article XX), either directly as customary international law or indirectly as a result of GATT interpretation for maximum compatibility with customary international law. Accession negotiations and conditions could carve out an exception for participation in OPEC export quota schemes (UAE, Nigeria, Qatar, Venezuela, Indonesia, and Kuwait are WTO members; Algeria and Saudi Arabia are in accession negotiations; Iran and Libya's applications for accession are explicitly being blocked by the United States; Iraq is not a member and is not involved in accession discussions). However, accession to the WTO is now subject to increasingly restrictive conditions; getting early into the club means having to live with less such constraints. The WTO may

have been, in 1948, primarily about access to manufacturing goods, with little interest in energy security. But this has changed for the influential blocks in the WTO. There is and will be more and more of an effort to extract concessions favoring U.S. and EU energy security concerns from the resource-owning countries requesting membership. For example, in the case of Russia, the dual energy price (i.e., higher export price, lower domestic prices for both energy exports, and pipeline tariffs) is currently a stumbling block [as is the prohibition on the WTO Trade Related Investment Measures (TRIMs) rules]. Nevertheless, the legal instruments for a "deal" are available, but political will and some creativity are required to identify the contours of a deal that will improve the situation of both sides.

Possible future membership of the OPEC countries in the Energy Charter Treaty would raise the same questions as raised for the WTO, because the ECT provides for non-GATT members the application of GATT provisions, with some qualifications. Different from the GATT, however, the ECT, in Article 18, explicitly recognizes "Energy Sovereignty" and "the optimization of (resource) recovery and the rate at which they may be depleted or otherwise exploited" [Article 18(3)]. Arguably, ECT membership therefore poses fewer problems for OPEC countries than does the GATT (though the GATT/WTO include several OPEC members and others in accession discussions, whereas the ECT does not). There may, however, be soft-law disciplines under the ECT for "export taxes" on oil [Article 29(4): states shall "endeavor" not to increase export levies], and under Article 6 (competition law) as well as Article 5, prohibition on TRIMs. Some of these issues could be solved through "understandings" negotiated by countries requesting accession (or, in the case of Russia, before ratification). Such understandings could include a limitation or long transition process for the TRIMs obligation (Article 5), a recognition of OPEC production control (limiting any argument about the competition law, Article 5), or an understanding that the ECT does not affect issues that are controversial between the EU and energy exporters (e.g., Russia and Algeria) such as destination clauses and initial access to new infrastructure such as pipelines.

To sum up, OPEC, as an international organization and forum that facilitates concertation among the major oil- and gas-producing countries, is now increasingly pulled into the institutional structure of the global economy. There will have to be a give-and-take on both sides to conclude such integration successfully. OPEC fulfills, different from more hostile sentiments in the 1970s, silently important functions for both domestic producers and international oil companies by being the organization most keen (and most potent) to help stabilize prices by helping producers to manage production. Whether this ability, which was not evident in the 1980s and 1990s, is maintained or, as happens with most cartels, will fade again is outside our ability to forecast. Sustainable development requires greater application of energy efficiency, minimization of emissions that are harmful for the global (and localized) climate, and possibly restrictions on the supply and use of hydrocarbons. Such policies, eagerly pursued by NGOs and the EU, for example, are unlikely to succeed if proper account is not taken of OPEC, the major international agency of the major oil-producing countries. This analysis suggests that there may be more compatibility than meets the eye or that is intuitively implicit in the conventional reference to the "OPEC cartel." An overall deal is possible, but requires a more active and creative effort at identifying communities of interest and much stronger leadership in pursuing and negotiating them on both sides. An arrangement could require some concessions by OPEC in terms of managing the oil price as a contribution to a stabilizing world monetary policy (e.g., lower prices in a recession, higher in a boom). It would require better guarantees of security of supply to concerned parties (e.g., the United States, EU, and China). Oil prices could also be linked to import prices for the producing countries. In exchange, there could be some examination of the very high excise taxes on gasoline and some other developed-country policies affecting the producer states. A higher price for oil together with a discipline on supply could be in the interest of both the OPEC countries and the environmentalist community, and in the long-term interest of consumer countries in a stable and secure oil supply. An unfettered global oil market is probably not in the interest of anybody—contrary to recurrent allegations, in particular in the United States. An unfettered oil market without political influence never existed. It would drive down oil prices to very low levels, would close down most non-OPEC production (including in the United States), countercurrent to Kyoto and energy efficiency objectives, would discourage development of renewable energy, and would very likely result in extreme (and therefore, bad for the global economy) swings of the oil price.

3. INTERNATIONAL ENERGY AGENCY

The IEA in Paris is the Western response to OPEC, though to mention this reverse-mirror role explicitly seems to be taboo. The IEA is the main international organization dealing with energy, though its mandate, membership, and operations are very limited; essentially, it is an intergovernmental energy policy institute for Western countries and also manages an emergency sharing system. Its importance also derives from the fact that there is no global intergovernmental energy agency. Quite likely, the division between the Western (IEA) and producer country world (OPEC) is the reason that there has not been enough political interest and effort to create a World Energy Agency (WEA). Existing international agencies with an energy mandate (UN, UN Conference on Trade and Development (UNCTAD), UNEP, UN International Development Organization (UNIDO), and World Bank) would also see their turf threatened if a World Energy Agency were to be created and would seek to take on such mandates.

The IEA was founded in 1974 on a suggestion by Henry Kissinger made in 1973, i.e., at the height of the first oil crisis, when there were sudden increases in oil prices and takeovers of foreign-owned oil production by the producing countries. OPEC, whose existence was disregarded in the 1960s, came to be seen as the main instigator and as a powerful cartel threatening the oil supply of western countries–in particular, the supply of the United States. The IEA was, and is, explicitly limited to western (that is, most OECD) countries. [Members, at present, include the EU countries, plus Hungary and Czech Republic, Turkey, the United States, Japan, Canada, Australia, New Zealand, and South Korea. The EU participates, but is not a member (though it could accede). Norway is an associate of the IEA, but not of the emergency system, based on a special agreement with the IEA.] South Korea, Hungary, and the Czech Republic have recently joined the IEA. There have been discussions with Mexico (which originally applied for membership), but in the meantime, Mexico has withdrawn its application. The IEA was not meant to be or to become a universal energy organization. Like with the EU, there have been intermittent calls for a UN Energy Organization, but these have so far come to nothing. The basic purpose of the IEA was to develop a system of collective energy security mirroring the collective producer power embodied, and then at its height, in OPEC. Such collective energy security operates through the continuing emergency-sharing system administered by the IEA. This system was never put into operation, though that may have been close at the time of the first Gulf war in 1991. [At the time of writing this article, a second Gulf war looked likely. The IEA was again preparing plans to ready its emergency-sharing system. But ultimately, the IEA desisted from implementing the emergency-sharing system because Saudi Arabia committed to use its massive production reserve to increase production. The Saudi "emergency production potential" is clearly superior to the IEA emergency sharing, though its availability in the future rests on the solidity of the U.S.–Saudi relationship, which, at present (summer 2003), is in jeopardy.] The most significant supply disruptions in the EU came about not because of OPEC and Middle Eastern conflict, but because of resistance to another round in gasoline tax increases by the British and other EU governments in October 2000.

The IEA's raison d'être has therefore diversified away from its earlier focus—Western solidarity in the face of OPEC-faced threats—toward energy-focused research and market intelligence, i.e., a type of work that is as well carried out by academic institutions, consultancy, and investment firms. The IEA advantage or difference is that it is publicly and internationally funded, and therefore with much more historical continuity, financial stability, and less dependence on markets, clients, and national budgeting processes, as compared to private or nationally based public energy research institutes. Though still in charge of the OECD countries' energy-sharing mechanism, it now fulfills a function of centralized research and intelligence, quite similar to the role of national energy institutes or the pooled research and intelligence function of international business associations. Focused on something that is still seen as strategic, and insulated from the questioning of national agencies of that type by its character as an international organization, it may have to justify its continuing usefulness by defining more closely the "public good" it delivers and its distinctive comparative cost and quality advantage over private, more market-driven organizations.

The International Energy Program (IEP) of 1992 is the treaty constituting the IEA. Membership is limited to OECD countries, but there is no automatic membership of OECD countries. Accession therefore requires a double hurdle—to the OECD and then to the IEA. The close-to-membership association agreement with Norway might, though, be a relatively simple way of protomembership for

non-IEA countries (if this was considered desirable). An IEA path toward a universal energy agency is therefore in all likelihood blocked; given the common economic and political orientation, a link-up with the Energy Charter Conference could be advocated, if the United States were to join (*vide infra*), but a full merger seems not to be in the cards as long as Russia and the other former Soviet countries are not OECD members. The IEA is open to membership by the European Community (Article 72 of the IEP), but this has not happened as yet. The EU Commission participates in all meetings without voting right. There seems to be a low-profile competition with the EU Commission. Decision-making by its governing board is mainly by majority voting, not by consensus as in the OECD. This should make the IEA in theory able to act more rapidly; presumably, in view of energy disruptions, that was the intention of a voting arrangement different from the OECD arrangement. Votes are weighted by pro-rata oil consumption, highlighting the economic weight and role in emergency sharing.

The IEA is in a complex relationship with the OECD as an "autonomous agency," with its own Executive Director and its own budget (though formally integrated into the OECD budget) and governance process, with some integration into and use of OECD administrative services. Funding is by membership contribution based on the OECD scale, plus voluntary, project-related contributions. The IEA has a separate division on "nonmember countries," which is used for an energy policy dialogue, economic studies, and joint research projects. A modest "global energy window" is thus open to the IEA via its nonmember activities and structures. Different from national or international agencies (e.g., World Bank, UN, UNIDO, UNCTAD, IMF, and IMO), the IEA seems not to have carried out technical assistance (policy advice; legislative, tax, and institutional reform; training; assessment of energy projects and programs; privatization; investment promotion) in member or nonmember countries. That is regrettable, given that the IEA has a large core of expertise, a comparative edge in its specialty field, and an energy policy link with all the IEA and many non-IEA countries. Energy assistance, when provided by other agencies (including the EU Commission), is never the central focus of expertise and interest and therefore is often provided with rather uneven know-how. Member states must have preferred energy policy assistance to come, rather, through their bilateral aid agencies, perhaps with the idea to gain a competitive advantage for their own energy industries and consulting firms, rather than provide regular program funding for technical assistance to the IEA. Also, the IEA may have intentionally avoided vigorously seeking voluntary contributions that would have placed it into competition with most other international agencies. The rough-and-ready tumble of international aid competition does usually not sit easily with well-funded international organizations staffed by Western civil servants on secondment. The dual character—rarefied and rare emergency sharing here, intergovernmental research institute there—may also not sit easily with the competitive vigor required to develop a sizable role in policy advice.

The traditional core of the IEA is its emergency oil-sharing program, now rather described better with the wider term as "emergency response measures." This consists, first, of measures by member states to reduce demand and to maintain oil stocks at 90 days of net imports. If emergency situations for the whole IEA group occur (two levels of group shortage: a 7% and a 12% shortfall, to be determined by the IEA Executive Director), a rationing plan is triggered that also requires surplus countries to provide for imports into deficit countries. IEA-based oil companies may have to be directed by member states to reorder supplies. There is a regular coordination mechanism between the IEA and such oil companies. The trigger is an actual shortage, rather than a sudden price explosion, so that the mechanism is rather more akin to a wartime international rationing plan than to a market intervention mechanism. The distinction between the two is sometimes questioned. Price spikes can both indicate and cause a disruption of physical supplies, and market intervention can help to manage scarcity situations. The United States, for example, has been more interventionist in deploying their strategic reserve in high-price situations, e.g., in 2000/2001. With two more recent governing board decisions in 1984 and 1995, the flexibility of the IEA and the member states to respond in a coordinated fashion to preemergency situations was increased, basically by a consultation procedure leading to the use ("drawdown") of the oil stocks. None of these mechanisms has ever been activated, though the beginning of the Gulf war in 1991 led to calls to start at least with some use of oil stocks to calm markets.

Otherwise, and in view of the largely dormant character of the emergency program, the IEA has become very much a collective study organization for

its members. There is regular reporting to enhance the transparency of oil markets; in-depth review of country energy policies (with recommendations and preparation of "outlooks" that provide national and commercial actors with some idea about the future of energy demand and supply), and a formulation of "shared goals," which at present comprise liberalization of markets and environmental issues. The IEA is unlikely to take a position on implementation of the Kyoto Protocol to the UN Framework Convention on Climate Change as long as there is a strong divergence between the United States (and most of its companies) and the EU (and its companies). Different from other more visible international agencies (e.g., the OECD, with its failed effort to negotiate a multilateral agreement on investment; the WTO, World Bank, and IMF, with their exposure to the anti-globalization movements), the IEA has so far escaped from much public attention. This relative immunity, though, is also likely to make it less aware of nongovernmental views and demands and less able to engage in a dialogue with them. Finally, the IEA provides an institutional vehicle for research in energy technology through, so far, over 40 collaboration agreements it sponsors, including with nonmember countries. There should be a potential here to engage not only scientific and technological, but also more applied and policy-oriented, technical assistance.

The IEA is in a curious situation. A child of the oil agitation of the 1970s, it may find itself in quest for justification of its existence. With the realization that the oil-exporting states are much more dependent on export than are the importing states, the fact that the emergency program has so far never been activated can be more easily explained. With oil declining in the energy mix, substituted in part by gas, and now, under the signs of Kyoto change, the new push toward renewable energy resources, the oil focus of the IEA risks obsolescence. Energy security is now no longer exclusively a matter of oil supply, but also of gas, coal, uranium, and electricity supply. Energy security for the IEA member countries, in particular the EU, means a favorable investment situation in producing countries and favorable legal and institutional conditions for transport and transit of energy resources (and electricity) and physical infrastructure (plus a regulatory framework maximizing its use) such as pipelines, interconnectors, and storage facilities. Though the IEA has dealt with such issues in various studies, it has no operational, or policy, or policy advisory role. Also, its character as a strictly Western, OECD-type of organization may be in question as globalization and the forces now triggered call rather for universal organizations, with an ability to conduct a global dialogue with all relevant stakeholders, conduct globally focused research, and prepare globally relevant policy studies corresponding in coverage with the globalization of energy markets. This was brought home starkly to the OECD when it tried to negotiate (in its club atmosphere) a multilateral investment code that was mainly relevant and intended to be ultimately applied to non-OECD countries.

There does seem to be a need for a truly universal energy agency, because energy continues to be the mainstay of the global economy. One way would be to maintain the emergency-sharing system of oil, but to expand the organization's focus on all energy sources, expand membership (perhaps in associate form) to all countries wishing to join, and put more emphasis on developing energy-related technical assistance. The IEA's nonmember countries and its research contract areas would seem to provide a nucleus for such expansion. An IEA consisting of two components—an emergency-sharing, OECD-based side and a universal, all-energy-based side—might be envisaged. This could be developed gradually, and might emerge over time anyway, but it would also make sense to consider amending the IEA agreement to provide for formal associate membership to non-OECD countries. In terms of influence on international energy law, the IEA's contribution has been mainly the emergency-sharing system as a free-standing element. IEA studies, generally of high quality and often prepared by seconded government and industry staff, will have an influence on policymaking by accentuating policy shifts (as presently toward liberalization and environmental mechanisms), but there is little direct influence on national energy law reform, as, for example, through bilateral or World Bank and UN technical assistance, nor has the IEA so far played a visible role in the evolving WTO-, North American Free Trade Agreement (NAFTA)-, or EU-based trade law of energy nor in the many environmental treaties, protocols, and guidelines now emerging.

With or without the IEA, the policy recommendation of the author is for a truly global World Energy Agency (WEA). The energy industries are coalescing into a truly globalized industry. This means there is a need for the research institute-type work of the IEA, the market-stabilizing influence of OPEC, and the protoregulatory work of the OECD, but involving, on an equal level, all stakeholders,

i.e., governments, companies (for example, in the way they are involved in the International Labor Organization procedures), other international agencies (IEA, OPEC, OECD, World Bank, UN agencies), and nonstate actors such as industry, professional associations, and NGOs. As the EU Commission recently noted, energy in developing countries is "an orphan without a parent international organization."

4. ENERGY CHARTER CONFERENCE AND SECRETARIAT

The Energy Charter Conference, served by its Secretariat, is the most recent addition to specialized, energy-focused international organizations. It is based on the Energy Charter Treaty (with a headquarters agreement with Belgium), and has the formal status of an international organization. The Secretariat services the negotiations for secondary and follow-up instruments (protocols) and supports treaty implementation. The ECT is an energy-focused treaty with all European countries, the states of the former Soviet Union, plus Australia and Japan as members. (The United States and Canada have signed the 1991 precursor European Energy Charter, but not the 1994 Energy Charter Treaty. China, Nigeria, Algeria, Saudi Arabia, and other countries have acquired observer status; several Mediterranean countries are considering the possibility of accession. Russia and Norway have signed, but have not as yet ratified. They are obliged to apply the Treaty provisionally, though.) The treaty deals mainly with investment protection (in the style of modern bilateral investment treaties) and with trade (adopting WTO rules for energy trade between states where at least one state is not a WTO member (now mainly Russia, Ukraine, and the Asian countries of the former Soviet Union). The ECT also deals with transit in a novel way.

The main activities of the Conference/Secretariat are at this time the elaboration of a transit protocol providing more specifics to the more general Article 7 of the ECT. A supplementary treaty to provide legally binding and specific rules for access for foreign investors ("preinvestment") was supposed to be negotiated when the ECT was signed in 1994. Negotiations took place between 1995 and 1998. Their final conclusion depends on the resolution of a number of outstanding political issues. The political interest in the supplementary treaty seems to have expired. There was also resistance, reportedly mainly from France and the EU. The interest in expanding investment arbitration (available under Article 26 of the ECT) may also have waned as the full potential of investment arbitration, in particular under the NAFTA, became clear.

The trade part of the Treaty is losing significance, because most East European countries are about to join the WTO. It is, for example, still relevant for the issue of energy import restrictions now being imposed by EU countries (in particular, Germany and Austria) on nuclear-based electricity, to the extent that such restrictions would affect non-WTO members (e.g., Russia or Ukraine).

The ECT, different from the quite similar (content-wise) Multilateral Agreement on Investment (MAI; *vide supra*), has not been politically very visible. As a result, the treaty and the ECT organization have been spared NGO onslaught. As in the IEA case, that has also compelled them, less so than the large international organizations, to engage in a dialogue with nongovernmental voices. The ECT's future lies mainly in securing its current membership through ratification by Russia (which seems possible), extending its membership to other Asian and Mediterranean countries, and serving at least as a model for regional energy relations in Asia (Asia–Pacific Economic Cooperation, or APEC), Latin America, and Africa. The strength of the treaty and the justification for a permanent Secretariat will also depend on the respect for the treaty's investment protection, and here mainly in the higher risk former Soviet Union countries. But the role of the Secretariat in getting such respect is limited, though the Secretary-General has a role under Article 7, in particular, for appointing a conciliator with provisional decision-making powers in case of transit disputes.

The staff of the ECT is quite small (under 30) and it has at this time no substantial budget for extracurricular activities, e.g., studies and country reports, like the IEA style, or technical assistance, like the UN and World Bank style. EU technical assistance has, though, at times supported ECT implementation projects. The Energy Charter Secretariat sits not too comfortably between the EU Commission (both are located in Brussels) and the IEA in Paris. Formally, the EU Commission has no particular role in the EC Secretariat; the EC is one of the 52 members. But politically and financially, the EU in its entirety is the main sponsor of the ECT. The absence of the United States reinforces this role.

The ECT is one of several privileged dialogue facilities between Europe and the former Soviet

Union, though for East European countries, the ECT will decline in significance because accession to the EU, and thereby full adoption of the energy "acquis communautaire," has acquired priority. But for the EU and the Commission in particular, it does not seem clear whether they wish to use the ECT machinery or its internal instruments for creating a legal and institutional framework facilitating trade and investment with its major energy partners. Several EC initiatives (e.g., the Inogate energy transit project) seem to overlap, duplicate, or compete with the ECT transit protocol effort. Similarly, instruments of EU external energy policy, such as association, partnership, and cooperation agreements and the Lome/now Cotonou agreements, could substitute for work through the ECT.

In terms of influencing non-OECD countries (transitional and developing economies) toward market-economy models for organizing the energy sector, the EC Secretariat/Conference overlaps with much larger organizations such as the EU and the World Bank, and the nonmember activities of the IEA. If the United States were part of the ECT, an argument could be made for much more collaboration, up to merger, between the EC Secretariat/Conference and the IEA. If, on the other hand, Russia were to ratify the ECT formally and other countries were to accede, the ECT could become either a major channel of EU dialogue with energy producers around the world (including OPEC countries) or, alternatively, a jointly "owned" and therefore more equal organization for regulating their economic relationship, rather than the more one-sided reliance on EC agreements.

The ECT and its organizational contribution to international energy law is currently mainly through the service of the ECT, one of the very few multilateral treaties exclusively devoted to energy, and thus a key element of international energy law. By support of negotiations for an energy transit protocol, the Energy Charter organization is also directly involved in the emergence of new and very relevant international energy law. As energy markets integrate regionally and globally, the role of transport, transit, and interconnectors becomes much more important than it was in a period when most energy industries were segregated into national areas. By the end of 2003, the negotiations for a transit protocol came to a halt, due to so far irreconcilable differences mainly between Russia (strongly influenced by Gazprom) and the EU.

5. INTERNATIONAL ATOMIC ENERGY AGENCY

The International Atomic Energy Agency (IAEA) is the one universal agency dealing with the peaceful use of nuclear energy. It has played a vital role in international nuclear security and a minor role, if at all, in the development and application of nuclear power. [The IAEA has, over the past 15 years, been relatively starved for cash, part of the general policy of Western countries to curtail funding to the universal international agencies. This has, arguably, not been a very wise policy, because the IAEA's core functions—nuclear safety, controls over trade in nuclear materials, and disciplines on nonproliferation—have become, even before September 11, 2001, more acute than ever, in the context of the collapse of the former Soviet Union nuclear industries, acquisition of nuclear power (for peace and for war) by developing countries, and the threat of nuclear terrorism. It is assumed that the Al-Qaeda attack on New York in 2001 will reverse that trend.] The IAEA was established by treaty, effective in 1957, one of the few instances of a successful East–West collaboration during the Cold War. Seated in Vienna, with a staff of over 2000, a budget of over U.S. $300 million, and the usual setup (general conference, board of governors, and headed by a Director-General), it is not much known outside the specialized nuclear community, and has, as yet, not been exposed to the antinuclear movement in any significant way. [This is arguably so, because the security functions of the Agency (setting of security standards, control of nuclear materials trade, and nonproliferation) have been largely supported by the antinuclear movement. The IAEA has also never taken a strong position in favor of establishing more nuclear power capacity.] As a universal organization (related to the UN as such, but not a specialized UN agency in the narrow sense), it is used by Western countries for purposes of controlling nuclear risk in the developing, and now former Communist, countries, but it has to coexist (with some underlying competitiveness) with the specialized nuclear agencies of the OECD countries (OECD Nuclear Energy Agency; *vide infra*) and Euratom, the EU's special nuclear agency. (The Euratom treaty, essentially governing nuclear materials transfers within and into the EU, was concluded in 1957.)

Among other things, the IAEA original mandate included (1) research on the peaceful use of nuclear energy, including scientific and technical information exchange and training, (2) a safeguard system to ensure that nuclear materials are not diverted to

military purposes, and (3) setting of safety standards. Over the years, the main functions have been standard-setting and the safeguards system. Its main task now is to safeguard the use of nuclear materials and facilities in member countries under the Non-Proliferation Treaty (NPT; Article III deals with IAEA verification), but also in non-NPT countries (e.g., India, Pakistan, and Israel, although most of their facilities are kept outside the reach of IAEA safeguards).The nuclear powers emerging in the Cold War encouraged nonnuclear powers into the IAEA and NPT treaty system. Based on special bilateral agreements with the state, the IAEA carries out monitoring of facilities to ensure safety in terms of operational standards, but also in terms of non-proliferation. The nonproliferation issue has become relevant not only in relation to states (e.g., the Iraqi efforts to build secretly a nuclear arms industry), but also in relation to illegal trade in nuclear materials and weapons from state to state and possibly to terrorist groups. In 2002, with the security threat by terrorists to nuclear installations seen in a much more acute light, the future high-priority IAEA activities are most likely to focus on nuclear security, both with respect to defense against attacks on nuclear installations and with respect to terrorist threats to build and use small-scale nuclear explosives against civilian targets.

Apart from nonproliferation, which now has a significance for "rogue states" and smaller states keen on nuclear weaponry (i.e., not countries that are too big for international pressuring, such as China and India), the IAEA is to develop and help apply technical standards and guidelines for nuclear plant safety, waste disposal, and decommissioning of nuclear plants. The IAEA has here, like most specialized agencies, no direct regulatory powers, but has a major, if not the most dominant, influence on regulation by national agencies with direct regulatory powers. For those agencies, in particular from smaller countries, it would be inefficient to replicate the amount of effort at information-gathering, consultation, and best-practices definition that goes into IAEA standard-setting. In a material, though not formal, way, the IAEA therefore acts as a global nuclear regulatory agency. The standard-setting activities are the most relevant for nuclear operators. Standard-setting is preceded by extensive information-gathering, including the International Nuclear Information System (INIS) for scientific literature. Reactor safety protection and radiation protection are implemented through Operational Safety Review Teams assessing specific nuclear power plants; the Operational Safety Indicators Program, providing for plant-specific safety indicator; and the Incident Reporting System, which reports on incidents in nuclear power plants.

The safeguarding activities are the third major, and currently very relevant, function. The IAEA carries out a large number of inspections of individual nuclear facilities and materials (over 2000 usually per annum, depending on budgetary constraints). Several treaties (the nonproliferation treaty of 1968; Tlatelolco, for Latin America; and Rarotonga of 1985, for the South Pacific) oblige member states to submit to IAEA inspections. Bilateral safeguards agreements, now with most states, govern the details of the safeguarding inspections. It is hard to deny in these situations that the IAEA has acquired not only material, but also formal, regulatory powers in such situations by delegation via treaty. Following the revelations of Iraqi efforts to build an atom bomb, the safeguards system was enhanced. Under the Additional Protocol (INFCIRC/540) to the NPT safeguards agreement (INFCIRC/153), the IAEA acquired greater and more immediate access to sites in NPT countries, and state parties were obliged to disclose more information about their nuclear activities and installations. Ratification, though, has been relatively slow, much due to the U.S. government's depreciation of multilateral arms control. With nuclear terrorism now threatening the United States, this is likely to change. As other developing countries embark on major nuclear power programs (e.g., Iran, China, and Pakistan), they will be encouraged to accede to the Additional Protocol as a sign of goodwill and to avoid "blacklisting" and other forms of sanctions. The IAEA has also an important role in nuclear trade control. Under the Nuclear Suppliers Guidelines, to which nearly all supplier countries subscribe, nuclear materials exported must be placed under permanent IAEA safeguards.

The IAEA provides technical assistance in matters of nuclear safety (out of voluntary contributions), but it is here in some competition with other agencies, e.g., the European Bank for Reconstruction and Development (EBRD) program on nuclear safety in Eastern Europe or the OECD/NEA. The EBRD is involved only in the Ukraine (Chernobyl Shelter Fund), Lithuania (Ignalina nuclear power plant closure), Bulgaria (Kozloduy plant upgrade/closure), and Russia (Rovno Unit 4 and Khmelnitsky Unit 2 plants) under its Nuclear Safety Account. The OECD/NEA is involved only in some advisory activities on emergency testing, technical aspects of power plants, and a project in Russia (the TASPLAV

project, concerning a Russian experimental facility where the reactor core material can be melted).

The IAEA has been a significant contributor to international nuclear law, in particular in the area of nonproliferation and safety of materials and installations, by administering relevant treaties, servicing the negotiation of new multilateral treaties and protocols, and designing the periodically reformed technical standards. The IAEA also publishes regularly, in its *Legal Series*, updated information and analysis on nuclear law (the *Nuclear Law Bulletin*, on the other hand, is published by the OECD/NEA). The IAEA was the moving force in creating the Vienna Convention on Civil Liability, and also the Brussels Convention on the Liability of Operators of Nuclear Ships, the Convention on Civil Liability for Maritime Carriage of Nuclear Materials, the Convention on Physical Protection of Nuclear Materials, the Convention on Early Notification of Nuclear Accidents, the Convention on Assistance in Case of Nuclear Accidents, and the revised Vienna Convention on Civil Liability for Nuclear Damage. The Vienna Convention served as a model for the subsequently negotiated Paris Convention. Other agreements sponsored by the IAEA relate to radioactive waste, nuclear safety, radiation, emergency planning, safeguards, and the nuclear liability agreements mentioned (with some involvement of the OECD/NEA) in the Paris and Brussels nuclear liability conventions.

6. *UNITED NATIONS*

The UN system consists of the United Nations proper with its various departments and other units and specialized agencies. Some of those other groups, such as the World Bank or the IMO, are for practical purposes completely independent. The survey here cannot do justice to the panoply of activities by the UN system, its main and secondary organs, and many secretariat groups and specialized agencies. What follows is therefore a selection with comments, rather than a systematic survey.

6.1 Climate Change Secretariat

For the energy industries, in particular the oil and gas and coal industry, the one UN activity with most relevance is the UN Framework Convention on Climate Change, with its Secretariat in Bonn The fate of the Kyoto Protocol is not clear, with its specified caps on CO_2 and other relevant greenhouse gas emissions in industrialized (including post-Soviet) countries, unspecific goodwill obligations on developing countries, introduction of emission trading and other emission reduction measures [Clean Development Mechanism (CDM) and Joint Implementation (JI)], and absence of the United States, the largest CO_2 emitter. (The target for post-Soviet Union socialist countries is currently largely irrelevant because it was based on the much higher CO_2 output during Communist times. This was the political price to get these countries on board the Kyoto protocol, but it also means that they do not have to undertake any meaningful efforts at present.) But the negotiations held by the Committee of the Parties (COP) for implementing the Kyoto Protocol are likely to put pressure on governments, in particular within the EU and the EU accession countries, to favor renewable (and possibly later nuclear) electricity generation and reduce coal, and possibly later oil and then gas-based power generation and consumption in transportation. With U.S. absence, there will also be a trade issue to the extent that the implementation of the Kyoto mechanisms is likely to develop intracorporate, national, and international trade in emission rights, but also financing of joint implementation and clean development mechanisms; it is hard to see how U.S. companies, operating outside the Kyoto membership, can be full beneficiaries of the emerging trade in emission rights and equipment/services for climate-change management when there is a tension between free trade under the WTO agreements and restricted trade among the Kyoto member states. [It is necessary to examine here the implications of GATT Article XXI(g) and analogies to the role of regional economic integration organizations and the noninclusion of international organizations in the WTO membership and obligation system.]

6.2 Compensation Commission

The political arm of the UN has had an involvement in oil and gas affairs through the UN Compensation Commission (UNCC), instituted after the Gulf War to administer Iraqi liability for war damage, in particular large-scale environmental and other damage to the oil production facilities in Kuwait by Security Council resolutions 687 and 705 (1991) and 986 (1995). The UNCC was created in 1991 as a subsidiary organ of the UN Security Council. Its mandate is to process claims and pay compensation for losses and damage suffered as a direct result of Iraq's unlawful invasion and occupation of Kuwait. A specified percentage of the revenue (in 2001, 25%) from authorized Iraqi oil exports is earmarked for the compensation of damages resulting directly

from the invasion of Kuwait by Iraq. Such damages are to include, according to the Security Council resolution, commercial losses, "environmental damage," and depletion of natural resources. [There is no doubt that Iraq by far exceeded the right of extraction of a belligerent occupant. It seems to have carried out a large-scale destruction of the Kuwaiti oil industry installations, according to the UN Compensation Commission Governing Council (Document S/AC.26/2001/16 of 22 June 2001, in particular p. 65, re Kuwait). The issue is, if the belligerent occupant, with the right of usufruct, can continue to extract the "normal" amount of (technically depletable) hydrocarbon (or water) resources.- With the occupation of Iraq in 2003, the priority allocation of Iraqi oil revenues to compensation under the UNCC will change, for the benefit of Iraqi occupation purposes and to the detriment of 1991 Gulf War compensation.]

The UNCC is in form and name not a tribunal, but an administrative process set up to expedite claims. (See the explicit statement on the UNCC Web site on "claims processing": "The Commission is thus neither a court nor a tribunal with an elaborate adversarial process. Rather, the Commission was created as a claim resolution facility that could make determinations on a large number of claims in a reasonable time. As such, the Commission operates more in an administrative manner than in a litigation format. The Commission's claims processing procedures were prescribed by the Security Council and were further elaborated by the Governing Council.") The Commission has received over 2.3 million claims. Practical and expedient justice is therefore the primary aim, different from the U.S.–Iran Claims Tribunal, wherein claims are still being litigated in depth. Most claims have been settled as of 2002. There have been decisions awarding very large amounts to Kuwait for damage to its oil installations, for oil extracted, and for environmental damage caused by oil spills. These have been subject to criticism of overvaluation, an issue that may be more problematic in view of the legal and financial resources available to Iraq to put its own position effectively. In practice, however, two things have transpired that, in the opinion of many, may vitiate this concern. First, at the Panels' direction, the Commission staff has taken an aggressive role in verifying the claims. As a result, the Commission sought and received significant budget increases to permit the legal and valuation staff to conduct thorough investigations of the circumstances of many of the commercial and environmental claims. Some claimants objected that the process was more intrusive than they would have experienced under most normal adversarial processes. Budget figures and the success rates for category E claims bear this out. Second, Iraq has had far more access than the Commission's designers anticipated. Iraq receives all claims submitted to the E1 (oil sector) Panel and is permitted to make its own submissions. In the larger claims, the E1 Panel has also held oral proceedings and has permitted Iraq to appear through counsel and argue the issues raised for decision. The panels only issue reports and recommendations—the final decision (not always the same) is made by the Governing Council. Reportedly, various governments, including Russia and France, used experts to review the E1 (oil-related) awards. On the other hand, participants of the process, both inside and outside the UNCC, have also communicated to this author that the Governing Council's decisions were intensely political, with only marginal adjustment of the procedures to afford to Iraq more than a mere formal opportunity to argue against, in particular, the valuations proposed by the UNCC's consultants.

The UN Compensation Commission practice should lead to international precedent for valuation of damage to the oil industry and oil-related environmental damages. The activity of the UNCC in the field of oil industry-related liabilities would merit deeper examination. It is regrettable that this significant precedent impact is somewhat weakened by the absence of full "due process" to the de facto defendant, Iraq. The panels of the UNCC have reportedly, in response of such criticism, made considerable efforts to stretch the existing rules to provide as much of a hearing to Iraq as possible. Future litigants will cite the actions as examples of legitimate ways to address problems such as valuation of oil and gas losses. This is particularly so because there is not much direct precedent. The remaining work of the UNCC will undoubtedly be overshadowed by the reconstruction of Iraq, which, by the end of 2003, was still not very clear. The U.S.-led effort to obtain a general and far-reaching debt reduction from all relevant debtor countries may therefore also impact on the UNCC, which, after all, manages Iraqi debt toward, *inter alia* but primarily for Kuwait.

6.3 United Nations Development Program

The UN Development Program (UNDP) is the main UN development funding program, fed by voluntary

contributions. Though the main funding is allocated to each country according to a population/poverty factor and spent according to national priorities, UNDP also runs several energy-related programs relating to small-scale energy development, development of renewable energies [mainly through the Global Environmental Facility (GEF), jointly with the World Bank], and implementation of the Montreal Protocol by assisting developing countries to eliminate activities that contribute toward depletion of the ozone layer. It is difficult to discern any appreciable effect on energy law in these activities; there is occasional funding by UNDP, if it fits into country priorities, of technical assistance for legislative reform in the oil and gas or energy sector.

The UN/UNDP in technical and financial assistance in the energy sector is, apart from the inevitable special attention for Kyoto and Montreal Protocol issues (renewable energy; ozone layer), not in any particular way focused on energy and certainly not on energy law. The significant developments in the energy industry over the past decade—privatization, liberalization, and postprivatization economic regulation—seem to have passed the UN system almost unnoticed. This is probably because the World Bank (for transition economies in some competition with EBRD) has taken leadership and "ownership" of these issues, but also because the UN system has had trouble modernizing its internal culture and outlook (and staff), all mired like old generals in the philosophy of the NIEO, with little substantial contact with industry, banks, or modern market-oriented thinking. Multilateral development funding has also been declining, and UNDP projects have shifted from the "harder" topics of energy to the topics that are more fashionable in the UN discourse (that is, poverty eradication, human rights, advancement of women, and sustainable development). In competition with the World Bank, the UN system has not been able to capitalize on its competitive advantage—greater sympathy for and trust by developing countries and (somewhat) greater independence from the United States—to develop concepts that are both in tune with the modernization of formerly state-oriented economies and less ideological than the philosophies that the Western-dominated institutions (World Bank, EBRD, and OECD) have imposed with rigor and purity, but also with less realism, practicality, and critical judgment, on developing and post-Soviet Union countries in the 1990s. It is both easy to speculate on the role of the UN in the energy sector and difficult to prove anything, because there is no independent assessment. As with all other international agencies, information and competence for critical assessment are divorced and self-assessment generally amounts to a mixture of propaganda and paraphrase of formal remits and mostly fallacious, if not even sometimes fraudulent, reports on own successes. From the accounts available, the UN activity appears diluted, ad hoc, not subject to systematic independent assessment in terms of cost-effectiveness, and with little, if any, lasting impact. This may be due partly to the organization's mode of operation and heavy bureaucratic processes, but also due to the fact, relevant for any provider of technical assistance, that aid may not be the least effective method to upgrade economic and energy competencies in countries that are seriously underdeveloped, in institutional, structural, governance, and cultural terms. Aid, to put it directly, does in principle not work where there is no absorptive capacity and culture.

6.4 United Nations Environment Program

The UN Environment Program in Nairobi is mandated to develop a global approach to environmental issues of sustainable development. As do all UN organizations, UNEP organizes training workshops for disseminating state-of-the-art know-how to developing countries, conferences to identify key issues and develop policy recommendations, and technical assistance to help developing countries to adopt modern policies. There is some focus on incorporating environmental considerations into energy planning. More of interest for international energy law is UNEP's work providing administrative support, including for the negotiation for subsequent protocols, to international treaties, such as the Vienna Convention for the protection of the ozone layer and its 1987 Montreal Protocol, the Basel Convention on hazardous waste transport, and the United Nations Framework Convention on Climate Change (UNFCCC). These, though not directly "energy laws," have at times a tangential impact on, in particular, the oil industry.

UNEP, though, has never been fully accepted as the lead agency on global environmental challenges and the global community's policy response. As part of the UN system, it suffers from the lack of political and technical credibility throughout Western countries and international companies; major activities in its field are carried out by the Global Environment Facility rather than by UNEP, or the UNFCCC Secretariat in Bonn. The "greening" of the World

Bank under its president, James Wolfensohn, has simultaneously reduced the need of Western countries for a UN institution in the environment field. Different from some of the accepted specialized agencies, seen in practical terms as fully independent and professionally very competent, UNEP has not achieved such status. It is rather covered by the negative view from which most of the standard UN activities (conferences, constant organizational restructuring, and making of ever more pious resolutions) suffer. A significant contribution to energy law, though not international or national environmental law, has, so far, been absent. That may differ, though, if UNEP's current efforts to promote guidelines on best practices with respect to environmental disclosure in the oil industry develop. Here, UNEP has stepped into the middle of nonconventional international energy law evolution by authoritative international and internal corporate guidelines.

In summary, energy is not treated well in the UN system. Its political appeal has been overshadowed by the great popularity of environmental and now human rights activities, which offer the chance to develop mutual benefits for both NGOs and the UN system to help each other to more political legitimacy. Energy is not only one of the most significant nuts-and-bolts issues of economic development, intercountry trade, but also a core element of sustainable development and climate change. With energy, after liberalization and privatization, developing from a mostly mere country issue to a great opportunity for mutual benefits from transnational trade, developing countries seem to gain less from this potential, because their institutional framework and intercountry politics are in most cases discouraging energy trade.

There is a case for a specialized UN agency dealing with energy matters, both in terms of monitoring world developments, linking with other issues (environment, transport, nuclear, shipping, and climate change), developing policy proposals for a global negotiating agenda (where really needed), relating with industry, and providing technical assistance in technical, institutional, and regulatory areas. Such an agency would work satisfactorily only if it were not part of the UN system as such, but rather a professionally competent specialized agency. It should be organized with a considerable input from both industry and competent NGOs and professional associations, i.e., in the way that the IMO, IAEA, IEA, and WTO are set up, rather than as a general UN department, with its inevitable slack and wastage.

7. OECD NUCLEAR ENERGY AGENCY

The NEA is a specialized agency integrated (much more closely than the IEA) into the OECD, though with some internal operational autonomy and directly under the Nuclear Energy Steering Committee, a subcommittee of the OECD Council. The 28 members comprise the OECD countries, with some exceptions (Poland and New Zealand); they account for 85% of the world's installed nuclear energy capacity. The main functions of the NEA relate to research, data collection, and the information exchange relating to the peaceful use of nuclear power, in particular, safety of operations, transport of nuclear materials, workers' protection, and waste management. In the past, the NEA managed nuclear operations directly. It works in the areas of nuclear safety and regulation; nuclear energy development; radioactive waste management; radiological protection and public health, nuclear law, and liability; nuclear science; and data collection related to the nuclear industry.

The NEA also has a protoregulatory role by present decisions of the OECD to its member states in the area of operational safety of nuclear plants, severe nuclear accidents, and radioactive waste disposal. Its own guidelines and standards are recommendatory, i.e., not legally binding. It works here through the Committee on Nuclear Regulatory Activities, the Committee on the Safety of Nuclear Installations, and the Radioactive Waste Management Committee; these bring together the national nuclear authorities. Different from the IAEA, it does not carry out safeguard inspections. Its relations are mainly with the nuclear authorities in member countries. As in all other international agencies, there is a connection between its primary network function and the identification of "best practices" arising out of the technical dialogue and regulatory comparison. The NEA also functions as a channel to transpose radiation protection norms developed by the International Commission on Radiological Protection into OECD decisions. A process of revision of the 1990s radiological protection recommendations is currently underway. In 1998, there was an evaluation of the NEA. The principal recommendations have been to incorporate sustainable development into its conceptual framework, to integrate better with the broader energy policy perspectives of the IEA, to develop a better collaboration with the IAEA based on complementarity and through an agreement, and to accept new members

(in particular, with major nuclear operations), but also to avoid duplication (in particular, in the area of technical assistance). It is not certain that the inherent organizational logic and self-interest of the NEA will allow such cooperative strategies to be implemented (interagency cooperation is usually recommended by external advisers and agencies pay lip-service to it); the reality is usually that there is interagency competition for interesting projects, funding, public profile, and organizational mandates, with often large-scale duplication.

One of the NEA functions is to develop and disseminate information on nuclear law. The objective is greater harmonization. It has carried out technical assistance on nuclear law reform (including nuclear liability in case of accidents) in Eastern Europe and Asia, usually in collaboration with the EU and the IAEA. It publishes the authoritative *Nuclear Law Bulletin* and has compiled several analytical/comparative studies on nuclear law in its member countries and Eastern Europe. It also runs a professional training program on nuclear law with the University of Montpellier. Finally, the NEA services the Paris Convention on third-party liability in the field of nuclear energy and the Brussels Convention complementing it. (The NEA Steering Committee may recommend the exclusion of nuclear installations or materials from the operation of the Paris Convention.) The NEA also collaborates with nonmember countries (particularly in the former Soviet Union) in the area of nuclear law.

Different from the IAEA (which has real regulatory powers), or the EU (which has real money for technical assistance), the NEA should be considered rather in the club model of the OECD: a forum for exchange and therefore dissemination and improvement of best nuclear regulatory practices. From the antinuclear perspective, the NEA is not necessarily an adversary: Its emphasis on nuclear safety, including its work on decommissioning of nuclear plants, fits as well with an antinuclear perspective. There is no record that "civil society" so far has included the NEA in its group of top evil forces driving globalization, such as the World Bank, the IMF, and the WTO. The NEA has in the past not taken much of a position in the debate over the continued justification of nuclear power. But with its most recent study on the link between greenhouse gases, climate change, and nuclear power, it has identified in detail the contribution that nuclear power makes, and can make, to achieving the Kyoto targets. It demonstrates that nuclear power is responsible for virtually zero CO_2 emissions, quite different from the large to very large amounts of CO_2 emissions from coal, oil, and gas consumption in power plants. Given the current debate about nuclear power, with opposition from traditionally antinuclear NGOs, and support based on its pro-Kyoto impact, the NEA should be expected to continue to be a significant, if not vital, function in the very specialized field of nuclear law and regulation.

8. EUROPEAN UNION

The EU is not an international agency, but rather, under international law, a persona that is situated between a supranational organization and a federation of states. Each of the European Communities (EC; Euratom) has its own legal personality. The European Union combines the European Communities plus, as established by the Treaty of EU (the Amsterdam Treaty), the foreign and security policy and justice and home affairs "pillars." The Council can authorize the Presidency to negotiate agreements with third parties, binding the EU. The EU's formal legal status is, therefore, as with many EU matters, unclear; perhaps it could be qualified as a "partial and budding international legal person." The "energy law" it produces is therefore both part of international energy law and the internal, domestic energy law of both the Union and (through direct effect and implementation by national law) the member states. The EU has, among all international organizations, been the most active producer of energy law over the past 15 years, primarily in the design and implementation of its target, an integrated energy market. This development is now in full motion, including implementation in member states, but far from completion. The EU is also a most interesting case to watch: its energy law is a pilot exercise for creating integrated energy markets in other regions (e.g., North America, South America and the Americas, Asia-APEC, and around Russia). It is also the dominant model in spheres of intensified economic cooperation of the EU [accession countries, Eastern Europe and Commonwealth of Independent States (CIS), and Mediterranean] where there is now a legal obligation and a de facto pressure to adopt the single-market instruments. Finally, the energy integration methods and experience of the EU provide an example of energy (and wider economic) integration in the global economy. It is the world's laboratory for ways to create integration benefit out of opening up national, hitherto largely segregated, markets for

cross-border investment and trade; for identification of the obstacles that have been overcome (and that have not yet been overcome, including the current cultural, institutional, and political limits to integration); and for the new challenges (in particular, environment and climate change) and ways to deal with them. The EU is therefore at this time the most relevant precedent case, not necessarily for instant copying, but for identifying challenges, issues, and policy instruments and their likely impact (including resistance to them). The importance of the EU as the global economy's laboratory for modern, postprivatization energy law as an instrument of economic and environmental regulation in emerging integrated energy markets cannot be overestimated.

In terms of organizational structure, the European Commission is both the conventional Secretariat servicing the treaties [Treaty of Rome and Treaty of Amsterdam, Euratom, and European Coal and Steel Community (ECSC)] and also an independent actor with colegislation, regulatory, and enforcement powers. In terms of political weight, it is the driver of integration policies, setting the agenda, organizing the process of information, consultation, and coalition building. Energy competence is located mainly in the Directorate General Transport and Energy (DG TREN, formerly DG XVII; handling the single energy market, Synergie), but energy-related competencies are also exercised by the DG Competition and the DG Environment, and the directorates [in particular DG Relex (ex-DG I), handling international assistance and foreign affairs—e.g., the TACIS and PHARE programs]. The Council represents the member states and the intergovernmental facet of the EU.

The European Court of Justice (ECJ) is the most influential and active international court. It has been instrumental in developing key notions of EU law going beyond the intergovernmental character of international treaties and developing the EC/EU into something between a supranational organization and a federal state, mainly through the concepts of supremacy and direct effect of EU law, and by a mostly integration-oriented and policy-based, rather than letter-based, interpretation of EU law. It has, though, been reticent to decide on matters that would require a large-scale industrial restructuring and establishment of a regulatory system, as in the 1997 case involving various EU member states' energy import/export monopoly. In these cases, the Court has observed "regulatory restraint" and has waited for the negotiations for new energy directives led by the Commission to achieve their result.

The new EU energy law cannot be compared to conventional international public law, which primarily deals with, first, division of proprietary and regulatory jurisdiction, and with, second, the existing quite general, unspecific, and indirect impact of many, often not yet ratified, international conventions, on the energy industry. EU energy law has a quite different goal: it is aimed primarily at restructuring the institutional and legal foundation of national energy industries so that a truly EU-wide energy industry can gradually emerge, while at the same time introducing on the Community-level and member state-level a more even playing field, with respect to environmental regulation and initiatives toward reducing greenhouse gas emissions. It is part of the global paradigm of privatization and liberalization. This entails dismantling of existing monopolies and barriers to cross-border trade, providing the legal basis for competitive markets, and, finally, nudging the industries toward real competition. It is the legal form of a proactive, rather than reactive, industrial restructuring in the EU energy industries.

EU energy law consists primarily of the EU treaty and its key general provisions (Articles 28–31, 49, 50, 81, 82, 86, and 87) for freedom of movement and controls on anticompetitive conduct as applied specifically to energy situations, on one hand, and, on the other, on a series of directives, chief among them being the 1996 Electricity and the 1998 Gas Directives (and their amendments in 2003/2004). There is no separate chapter or policy in the EU Treaty on energy (though it has often been advocated), but only some references [Articles 3(u), 154, and 174; to natural resources and environment, 175) of marginal significance. Because energy is one of the most strategic industries, it can be argued that an energy chapter is desirable, but also that energy is automatically covered and included as a key component in any reference to industry, commercial, and economic affairs. The energy directives are being implemented by the member states (and by the East European accession states based on "Europe agreements"), with perhaps still a too large diversity. Primary law could have been used to dismantle the export and import monopolies and provide third-party nondiscriminatory access to the "essential facilities" of electricity and gas transport, storage, and distributions systems owned mainly through monopoly, but there was not enough political will and power in the Commission, too much resistance from member states and the energy monopolies, and no green light from the European Court of Justice. In its October 1997 judgment on export/import

monopolies, the ECJ essentially told the Commission it had to restructure the existing regime by negotiated and agreed-upon specific directives, and not by ad hoc actions focusing on specific issues in the much more complex sequence of energy operations endorsed by the Court. The Court in essence declined political responsibility for such restructuring and mandated the Commission to seek a negotiated (and thereby accepted and easier to implement) solution.

Liberalization initiatives have been accompanied by emerging measures for compliance with the EU's Kyoto targets for greenhouse gas emissions by promotion of energy efficiency and renewable energy sources (RESs). These will, in turn, require compatibility with the EU's rules on state aids (which are responsive to renewable energy and other truly environmental goals), transparent and nondiscriminatory procurement, and EU-wide trade. Other significant measures have been the obligation of state and private energy utilities and oil and gas licensees to procure in a transparent and nondiscriminatory way [i.e., to abstain from formal or informal protectionism (Utilities Directive)] a duty on member states to issue oil and gas exploration and development licenses in a similar transparent and nondiscriminatory way [i.e., exclude preference for domestic companies or companies with a desired domestic procurement record (Licensing Directive)].

The EU, however, not only produces "internal" energy law, but it also participates actively, though in a still inchoate form, in "general" international energy law. The gradually coalescing "federal" character, with its tension between necessary, unitary negotiating and deal-making competence toward the outside world, and the maintenance of such elements of national sovereignty by member states make the EU a particularly awkward, indecisive, and contradictory international actor. It is now recognized, based on several authoritative ECJ decisions, that the EU has exclusive competence in trade matters (e.g., WTO negotiations), but only "mixed" competence in "investment matters" [e.g., General Agreement on Trade in Services (GATS), Trade Related Aspects of General Property Rights (TRIPS), ECT, and MAI]. There is no explicit competence for the EC to enter into treaties relating to energy matters in the EC Treaty, thus trade, investment, and other powers are relied on, and usually in the form of joint EC/member states accession. The result is that the EU is very inflexible and slow to react and is therefore at a disadvantage in international negotiations, which require decisiveness and clarity. The EU is very dependent on import of primary energy sources, in particular oil (apart from the United Kingdom), gas (apart from The Netherlands, Denmark, and the United Kingdom). This important dependence is coupled with the fact that most energy sources are in volatile regions close to the EU (Russia, Central Asia, Caucasus, Algeria, and other Mediterranean countries; Middle East/Gulf countries; Angola and Nigeria), with political insecurity in these countries spilling over to the EU in terms of security of supply and domestic political disruptions (terrorism, volatile situations involving ethnic minorities from these countries, and the unwilling but inevitable implication of the EU in the U.S./Israel and Arab conflict). But a unitary and focused EU action is here impeded by numerous special interests of its member states: the United Kingdom and French relationships with former colonies; the German sensitivity to anything endangering peaceful relations with its Eastern neighbors; the Norwegian special status as non-EU member, but more or less subject to EU energy law via the European Economic Area (EEA) agreement (and a major, and stable, oil and gas supplier); the United Kingdom's "special relationship" with the United States; and the French "special" competitiveness with the United States in political and cultural affairs. EU actions are also influenced by the opaque relationship with the United States. The EU needs the United States as senior partner (in particular, for security measures, because the EU usually cannot act decisively nor employ effective security forces), but there is also an underlying tension, out of economic competition, some resentment (in particular, France) at U.S. hegemony, and much greater linkage of the EU to various countries and social, political, and religious forces in its greater neighborhood.

The most visible success of the EU is the Energy Charter Treaty; here, the EU has managed its probably most visible tangible success in being essential in moving a +52-country EC investment, trade, and transit treaty to legal effectiveness and implementation. The ECT is not legally "owned" by the EU, but the EU is its major financial and political sponsor. The treaty's transit Article 7 is now being developed into a specific energy transit protocol that reflects the EU's interest in facilitating diversified supply of energy (oil, gas, and perhaps electricity), in particular from the prospective oil and gas countries now making up the former Soviet Union. The Treaty, though, has not been followed, as was the original negotiating mandate, by a "supplementary agreement" dealing with

privatization and access for investors (though a full-text draft exists). The Treaty is now of increasing interest to countries outside its original sphere (East and West Europe); for OPEC countries, it would be a multilateral treaty whereby they would not have to face U.S. obstruction.

Other noted instruments have been the "Europe agreements" with the East European accession states; these oblige the accession states to transpose EU law (including energy law and the new directives), within a time span, into their national system. There is a focus on developing "trans-European energy networks" (i.e., pipelines, transmission grids, and interconnectors) and other current priorities of the EU in the energy sector: promotion of renewable energy, compliance with Kyoto obligations, and nuclear safety of the problematic safety standards of nuclear reactors in Eastern Europe. The accession countries are under considerable pressure, as weaker, EU-entry-seeking parties, to adopt the EU "acquis communautaire," i.e. the current state of EU energy law. On the other hand, though they are being pressured to adopt this "acquis," the making of which they have had no influence, they are not likely to benefit from the liberalization of the EU if powerful domestic interests of domestic member states are affected. There is, for example, considerable reluctance (and search for legitimate reasons, mainly based on environmental pretext) to provide for free energy imports into the EU from Eastern Europe. But this is the nature of the relative bargaining power—seekers to join a club have to accept its rules, without being certain that their presence is universally welcome.

The major direction of current EU involvement in international energy law is through the Kyoto process. The EU has become the main promoter and sponsor of the Kyoto Protocol, setting targets on greenhouse emissions, and this role has increased with the exit of the United States. It is not dissimilar from the way the Energy Charter Treaty has moved to completion, with the EU pushing and the United States exiting. The strong position taken in the Kyoto process reflects the EU's comparatively greater domestic interest in environmental matters, as compared to the United States. The reasoning is that, first, the European Parliament has a sizable representation of "green party" members and that, second, the EU Commission, as an international agency searching for ways of making up for its low level of political legitimacy, is more responsive to pressures from environmental NGOs. Parallel to Kyoto, the EU has also been promoting tighter measures against marine pollution, mainly by using port state leverage, influencing the Oslo/Paris Atlantic (OSPAR) marine environment convention committee toward full prohibition of offshore decommissioning of oil platforms, participating in all relevant international conventions with an at least tangential impact on energy industry operations, issuing a number of directives mandating environmental impact assessment for power plants, and developing systems of ecolabel measures (e.g., energy efficiency of appliances).

8.1 Euratom

Euratom, a separate treaty, but fully integrated into the Community institutions and managed by the Commission, was the EC's early response to the need to develop and monitor safety systems for nuclear energy materials. It was originally premised on the 1950s enthusiasm for the "peaceful use of nuclear energy," fears over security of supply of oil, and a supply-sharing system for uranium. None of these factors present at the origin of Euratom is present today, except to some extent the security-of-supply aspect. Euratom's original objectives were frozen during the early decades by conflicts with France, which wanted to see Euratom deployed to develop an exclusively European (i.e., not U.S.-linked, and largely French-influenced) nuclear industry.

The nuclear industry in the EU (particularly outside France and the southern member states) is an embattled industry. Austria has closed down its nuclear industry and Sweden and Germany are engaged in a slow process toward closing and decommissioning. The future of these processes can, however, not be predicted with any certainty because the enthusiasm for the Kyoto Protocol and the opposition to nuclear industry are difficult to square. If nuclear industry is on the wane, then the future task of Euratom will be decommissioning and safe storage of used nuclear materials. If nuclear industry reemerges, then safety rules, rehabilitation of East European nuclear power plants, and management of nuclear waste will be priorities. The new threat from terrorism will pervade as well in European nuclear law. There will be a tightening of rules of nuclear safety (installations, materials). The implications of a successful or almost successful terrorist attack on a European nuclear installation are hard to predict; the consequence could be an acceleration of the trend toward closure or a much greater investment in security.

Euratom develops safety norms, supports research, cooperates closely with industry, scientists,

and other international organizations (in particular, the IAEA and the OECD/NEA), and can invest in projects, including the right to raise loans for that purpose. Nuclear research constitutes one of the major components of Commission-funded research. The current nuclear-related EU Commission research budget is over 1.2 billion euros, from 2002 to 2006; loans available to fund rehabilitation of East European nuclear reactors should exceed 2 billion euros. Euratom produces, through the Community institutions, directives, regulations, and administrative decisions. It is given the right to own nuclear materials within the EU. It runs a system of safeguarding for nuclear materials (excepting those for defense purposes) within the EU. Like its sister organizations (the IAEA and the OECD/NEA), Euratom has managed to keep largely out of the limelight of public opposition to nuclear power development. Its specialized, technical character, its generally endorsed focus on nuclear safety, and its absence from public debate over nuclear energy may have helped. Such positioning may have also helped the nuclear activities of the European Commission to have a more comfortable life, but it is not certain that retreat from challenge and debate has served these organizations or the issue of nuclear energy. As organizations with special knowledge of nuclear industry and the most extensive networks, the Euratom service of the Commission (as the IAEA and OECD/NEA) should have developed a position to be able to make informed arguments about nuclear energy and highlight in the public eye both the risks and the particularly newly emerging benefits in terms of climate change. At present, the Commission's nuclear services have taken a more active role in the debate on nuclear industry: They stress the climate change/Kyoto target and security-of-supply advantages of nuclear energy; they currently propose a much deeper engagement with the safety issues of problematic nuclear reactors in Eastern Europe and improved safety standards and safeguard procedures.

The function of Euratom in safeguarding nuclear materials will only increase as international terrorism develops increasing sophistication and inevitably tries to utilize nuclear materials. Here, though, the main risk of access to nuclear materials, equipment, and expertise is likely to be in countries with run-down nuclear establishments and weak systems of governance. The Euratom service of the Commission as well as the IAEA and OECD/NEA are likely to gain more prominence because they will have to serve as international instruments to deny access to nuclear power to terrorist organizations.

In summary, the EU is currently, in spite of its many and deep institutional weaknesses, the one supranational organization with the most visible impact on national energy laws (mainly in accession, transition, and many developing countries), the main laboratory for economic regulation of cross-border energy trade and investment under the current paradigm of environment-friendly postprivatization liberalization, and the main sponsor of major international energy initiatives, mainly the Kyoto Protocol and the Energy Charter Treaty. International energy law can no longer be seriously studied without understanding both the internal and the external dimensions of EU energy policy. EU energy policy is likely to progress at a tortoise pace, pushed away from conventional energy (coal, oil, and gas) by the new emphasis on renewable energy sources, but, on the other hand still, and for all the foreseeable future, dependent on strategic oil and gas imports from highly insecure producing countries. It has not yet been able to grapple with its major energy dilemma: A large part of the EU's energy demands is covered by nuclear energy. Nuclear electricity is currently and for the foreseeable future, together with large-scale hydropower, the only substantial energy source that is not responsible for greenhouse gas emissions. The EU's 2000 Security of Supply Green Paper and subsequent reports make this clear. Simultaneously, there is no political leeway for the EU to promote maintenance and expansion of nuclear power until the green movements in the EU have made a choice between their traditional opposition to nuclear energy, based on the risk of accidents and disposal of nuclear waste, and the indubitable fact that nuclear energy, among the currently relevant energy sources, is the most Kyoto-friendly choice. Should the reorientation of subsidy and support lead to a true upsurge of nonnuclear renewable energies, then this dilemma might fade, but if it does not, then there seems to be no choice—either embracing, vigorously and expansively, nuclear again or forgetting about control of greenhouse gases.

9. CONCLUSIONS

All international agencies have at their core a "cross-border network" function: They bring "colleagues" together, typically from a ministry with a similar regulatory and operational task. There is a Secretariat, which acts to organize the network and its typically very formalized gatherings and provides

permanent expertise (mainly of a comparative nature). This function is not often perceived so clearly, but it may be the vital function of the organization. The network now increasingly includes, often relegated to a lesser rank in communication priority, nonstate actors; these are nongovernment, commercial, and professional organizations and consultants. If the agency does not have direct regulatory powers (they rarely do), the "network" nevertheless exercises considerable influence over national and international energy law: it identifies best or at least prevailing and current practices. These provide a persuasive and practical blueprint for national regulation. Experiences are discussed and models are taken home or brought to the meetings. Copying other people's work and working off legal precedent has always constituted the practical core of the legal profession.

The much more formal and directly "legal" focus of international agency work revolves around their respective treaties. The Secretariats service the treaties and the governmental delegates negotiate (with low-profile Secretariat influence) such treaties and then "implement" them at home. It is only in the EU, a hybrid between an international organization and a federated country, that the international authorities (Commission, Council, and Court) have some direct regulatory powers.

Different from human beings, who grow largely along the lines of a genetic program, international organizations mutate: they come with a fixed mandate, the organizational constitution, which seems at the beginning to be written in stone. But all agencies here surveyed have developed quite differently; some parts of their original mandate have become obsolete and dead text, other parts have been developed and new mandates have been effectively acquired, often based on bold reinterpretations of the original terms of reference. The OECD and its predecessor started out to manage the Marshal Fund; it is now the major intergovernmental policy think tank on any subjects that move governments. The World Bank started off as a financing agency for reconstruction in the countries that were U.S. allies in Europe after World War II; the World Bank sees itself now as the "premier development institution." The IAEA started off to encourage nuclear power; it is now mainly an institution to develop and enforce disciplines related to nuclear risk (by accident), mismanagement, or terrorism. What the IAEA lacks is a natural death. It is very hard to make an obsolete international agency disappear; the agency and its leaders, staff, clients, and constituency will cling to life, sometimes with (self-) deception over its continued usefulness, often with a desperate attempt to latch on to the currently fashionable paradigm so that criticism against the agency can be presented as criticism of the current high-ground moral values. International agencies seem to adapt more easily in form than in substance. They rebaptise readily what they have always done and will continue to do in the fashions of the day; they are much better at redesigning their public relations than what they do and how they do it. Probably every organization examined is in some way or other seriously out of date, mainly with respect to its internal structure and organization. All should be modernized toward more professional competence, but also toward inclusion of nonstate actors in their formal decision processes, although that is unlikely given the highly conservative and anti-change bias built into their organizational structure and culture. Their funding, almost exclusively by government contribution (e.g., national taxes), is also in need of review. More dependence on the value of services rendered in a more competitive situation would make them more modern, responsive, and efficient. Some of them (e.g., OPEC or the IEA) were created in response to a particular crisis that no longer exists in the same form; here, adaptation has been particularly difficult and modernization particularly pressing. But as intergovernmental organizations, they are mainly controlled by diplomats, who are removed from elections, politics, competition, and markets, so that the need for modernization is delayed by a double wall of insulation.

There is a considerable difference between two types of international agencies. The general ones (primarily the UN, also to some extent the OECD) are mainly "talking-shops." They do not have a very specific mandate and certainly very little or no regulatory or operational focus, but they serve to accommodate the public concerns and themes of the day, to process them into organizational language of some (though in reality much less than is claimed) authority. This is an important function because they conduct a public dialogue, albeit in very stilted form, about global politics. The weaker the nation state, and the greater pressure for public participation, now expressed on the international level mainly by the NGOs, the more is there a need for a "parliament" in the original sense, i.e., a "talking-shop." Not surprisingly, general international agencies have found it least problematic to accommodate the pressure of NGOs for involvement. Both actors look for legitimacy, and reciprocal recognition of legiti-

macy helps both. They operate politically in symbiosis. The problematic legitimacy of both is less of a problem because the operational impact is minimal; even international treaties emerging are typically neither legally effective (because not ratified) nor (even if ratified) specific enough to make a difference.

Both the international agencies and the manifold actors within "civil society" need to open up to each other, in a professional way. There is a good reason for "civil society" to be present in the talk-shop forums. But there is also the temptation that NGOs and such agencies continuously run like lemmings into the same, regularly changing, fashionable direction, scoring successes that do not count. There is a good reason for the specialized, and practically influential, international organizations to open up, not just formally, and not only by being more transparent, but by effectively allowing formal representation of nonstate actors on their governing boards. Multinational companies sometimes now have representatives from public-interest organizations and academia on their boards. Intergovernmental organizations are a child of the club of nation states. Such adjustment to a world in which governments do not play exclusively and in which many nonstate actors are now significant actors is difficult for international organizations and their nation-state governors. Even if nonstate actors now compete with governmental actors on the global scene, it is difficult for the places where governments still call the tune to yield a part of their formal power. Such an opening of the decision-making structures would only accommodate the 1940-type of intergovernmental structure to the changing reality of global society where non-state actors are now operating next to the conventional sovereign states. How to do it is much more difficult. NGOs, for example, are mainly self-appointed, with weak governance, transparency, and accountability. Commercial companies, and their associations and professional groupings, on the other hand, are better structured and in a formal sense more easily identifiable. In principle, international organizations should gradually co-opt those who represent power (commercial, financial, political, and public opinion) into their governing structures. Similarly, treaty negotiations should, as the OECD debacle with the MAI has shown, incorporate in a much more active sense those who have power and a voice in the field. There are inchoate precedents: the ILO tripartite system of decision-making (itself a child of the 1950s corporatist worldview) and the meek OECD Business Industry Advisory Council (BIAC) and Trade Union Advisory Council (TUAC) (largely, it seems, lunch opportunities for retired functionaries). But these need to be developed, experimented with, and, in the end, formalized. There is no reason why those with a powerful voice (International Union for the Conservation of Nature [IUCN], Shell, the International Chamber of Commerce, or the International Bar Association) should not be able to be part of the directorium of international organizations and part of the treaty negotiators whose consensus is necessary. For me, the unresolved issue is if the not fully responsible "activist" NGOS, such as Greenpeace *et al.*, should be in the governing boards of international organizations. As in domestic law for political parties, there should probably be a set of quality standards reflecting transparency, governance, and accountability. From "civil society" this would require a sea change in attitude: from a merely critical, destructive, and political-campaigning attitude to a position whereby constructive alternatives have to be designed and bargained for, and whereby positive responsibility and accountability for results (not just for criticism) have to be accepted.

Finally, this survey of the role of international agencies, with a focus on energy law and policy, also posits that there is an ongoing process of formalizing power relationships over weakly governed, underdeveloped countries reminiscent of (though certainly not identical with) the colonialism of the 19th century. Then, the power of the state, the wealth of the investors, and the values of the missionary movements propelled European states to control a large part of the world. Today, the formal trappings of statehood and sovereignty are everywhere. But there is exercise of "soft" and "structural" power by the rich societies through their NGOs (reborn 19th-century missionaries), through their multinational companies, through their controlled and funded international agencies, and through the legal, financial, educational, and cultural tools of exercising influence and co-opting elites, rather than outright formal ownership, as in the past. The treaties and their subsidiary tools now being used to impose "good governance" on less civilized societies all tend to be formal, on their face, but asymmetric in substance and actual practice. They are all directed toward telling underdeveloped countries and their governments what to do, but rarely, in practice, do they target the rich countries.

Acknowledgments

This article is based on a more extensive and in-depth study in OGEL 4 (2003), available from http://www.gasandoil.com/ogel;

also from Wälde, T. (2003). The role of selected international agencies in the formation of international energy law and policy towards sustainable development. *In* "Energy Law and Sustainable Development" (A. Bradbrook and R. Ottinger, Eds.), pp. 171–203. IUCN Environmental Law Programme, Paper No. 47.

SEE ALSO THE FOLLOWING ARTICLES

European Union Energy Policy • Fuel Economy Initiatives: International Comparisons • Global Energy Use: Status and Trends • International Comparisons of Energy End Use: Benefits and Risks • Labels and Standards for Energy • Nongovernmental Organizations (NGOs) and Energy • OPEC, History of • Renewable Energy Policies and Barriers • Sustainable Development: Basic Concepts and Application to Energy • United Nations Energy Agreements • World Environment Summits: The Role of Energy

Further Reading

Bunter, M. (1998). "Promotion and Licensing of Petroleum Prospective Acreage." Kluwer, Amsterdam.

Bradbrook, A., and Ottinger, R. (2003). "Energy Law and Sustainable Development." IUCN Environmental Law Programme, Paper No. 47. International Union for Conservation of Nature and Natural Resources, Bonn, Germany.

CEPMLP Internet Journal (available on the Internet at http://www.cepmlp.org/journal).

International Energy Law and Taxation Review (published by Sweet & Maxwell).

Journal for Energy and Natural Resources (published by the International Bar Association).

Maniruzzaman, A. (2002). Towards regional energy cooperation in the Asia Pacific, some lessons from the energy charter treaty. *J. World Invest.* **3,** 1061.

Oil, Gas Energy Law Intelligence (available on the Internet at www.gasandoil.com).

Redgwell, C. (2001/2002). International regulation of energy activities. *In* "EU Energy Law," (M. Roggenkamp, Ed.), pp. 13–96. Oxford Univ. Press, Oxford.

Smith, E., *et al.* (2001). "International Petroleum Transactions." 2nd ed. Rocky Mountain Mineral Law Foundation, Denver.

Tomain, J., Hickey, J., and Hollis, S. (1989). "Energy Law and Policy." Anderson, Cincinnati.

Waelde, T. W. (1996). "The Energy Charter Treaty, Gateway for East–West Investment Trade." Kluwer, Amsterdam.

Zedalis, R. J. (2000). "International Energy Law—Rules Governing Future Exploration, Exploitation and Use of Renewable Resources." Ashgate, Aldershot.

Investment in Fossil Fuels Industries

MARK E. FISCHER
Fischer–Seitz Capital Partners, LLC
Lyons, Colorado, United States

1. **Breadth of Investment Opportunities in Fossil Fuels Industries**
2. **Crude Oil and Natural Gas Exploration, Development, and Production**
3. **Coal Development and Production**
4. **Crude Oil Refining and Marketing**
5. **Natural Gas Processing, Transmission, and Distribution**
6. **Electricity Generation, Transmission, and Distribution**

Glossary

barrel of oil-equivalent A unit measure of oil and natural gas that assumes that 6000 (United States) or 10,000 cubic feet (Canada) of natural gas is equivalent to one barrel of oil; the United States focuses on average energy equivalency, whereas Canada works off average historical market value.

discretionary cash flow Usually defined as the sum of net income, depletion, depreciation, amortization, exploration expense (or dry hole costs), and deferred income taxes.

finding, development, and acquisition costs (FD&A) The capital costs of exploring for, developing, and acquiring reserves, usually expressed in a currency per proved barrel of oil-equivalent added during the period of spending.

hedging The practice by producers of avoiding spot pricing for some of their oil or natural gas volumes through the use of forward, future, swap, and option contracts.

price realization The unit price of oil or natural gas realized by a producer for certain volumes; individual price realizations typically differ from benchmark pricing due to differences in energy content, impurities, distance from major markets, and fixed-price contracts.

proved reserves The estimated quantities of crude oil, natural gas, and natural gas liquids that geological and engineering data demonstrate with reasonable certainty to be recoverable during future years from known reservoirs under existing economic and operating conditions; the standards that reserves must meet to be defined as "proved" are set by the Securities and Exchange Commission in the United States and by similar regulatory authorities in other countries.

standardized measure of discounted future net cash flows (SEC PV10) A measure of the net present value of a set of proved reserves established by the U.S. Securities and Exchange Commission; the measure projects the future net cash flows that would be generated from the liquidation of a set of reserves under certain pricing and cost assumptions and then uses a standard 10% discount rate to calculate the net present value of the projected stream.

successful efforts accounting An optional accounting practice whereby only capital spending that results in the booking of proved reserves is capitalized as property, plant, and equipment and the rest is written off immediately as exploration expense; the alternative practice is full-cost accounting, whereby all capital spending for exploration and development is immediately capitalized regardless of the results.

While the fossil fuels industry is mature, it offers multiple investment opportunities driven by technological innovation, its cyclicality, an ever-changing competitive landscape, increasing environmental requirements, and a declining resource base. Successful investment in the fossil fuels industry requires a solid understanding of the supply and demand dynamics of the highly cyclical oil and natural gas markets as well as of the unique assets of hydrocarbon producers and how they are accounted for financially and accurately valued. Investment in the regulated subsectors of the fossil fuels industry further necessitates a firm grounding in government regulation of natural monopolies. This article covers the key issues of which an investor in the fossil fuels industry should be aware.

 © 2004 Elsevier Inc. All rights reserved.

1. BREADTH OF INVESTMENT OPPORTUNITIES IN FOSSIL FUELS INDUSTRIES

Although fossil fuels industries are typically considered mature, they actually offer a wide variety of investment opportunities, from aggressive high-tech growth to ultra-safe income. Growth in global consumption of fossil fuels is slower than overall economic growth because energy efficiency continues to rise, but above-average growth opportunities are offered by companies able to increase their production of low-cost energy and, thus, steal market share from companies producing high-cost energy. Service companies that develop new technologies to reduce fossil fuels production costs also offer growth opportunities. As protection of the environment becomes increasingly important, the definition of "cost" expands from consumption of capital and labor resources to include the creation of pollution. Thus, today's energy companies with above-average growth prospects are not only oil and gas exploration and production operations that can find and produce hydrocarbons at below-average costs but also companies developing cleaner coal-burning technologies and companies trying to commercialize innovative fuel cell technologies.

Although prospects for industry-wide above-average unit growth are dim, profit growth could accelerate if technology begins to lose the battle against the ever-declining resource base and real energy prices rise. Until now, real prices for fossil fuels, as for nearly all natural resources, have been flat or down over time, as improved extraction technologies have offset ever scarcer resources to exploit. However, it is conceivable that the decline in the resource base could overwhelm technological advances as global demand continues to climb and that real fossil fuels prices would have to rise to cover the higher production costs. Although rising costs would fully offset the benefits of higher prices in the very long term, in the nearer term producers with low sunk capital costs would be net beneficiaries. Growth opportunities would also arise for producers of renewable energy as well as for oilfield services and equipment firms, which would experience greater demand for their products and services per unit of energy produced. Energy industry analysts hotly debate how close the world is to experiencing a secular increase in real fossil fuels prices. Less highlighted is the possibility that an unforeseen technological breakthrough will result in lower real fossil fuels prices in the future, perhaps because technological breakthroughs are by definition difficult to foresee.

The lack of above-average growth opportunities in fossil fuels industries may limit the long-term appreciation potential of many energy investments, but it in no way restricts the cyclicality and volatility of most investments in the unregulated, capital-intensive sectors of the industry. These sectors typically experience large swings in capital investment that result in alternating shortages and surpluses. When energy is short and prices rise, energy investments perform well, but when overinvestment leads to surpluses, energy investments perform poorly.

The exception to this rule is investments in energy industries with government-regulated returns. For example, in the transmission and especially the distribution of energy, in most cases it would be inefficient to foster the construction of two or more competing transportation systems, so the government allows a single company to maintain a monopoly on transportation of gas or electricity in a given market. However, the government sets transportation rates for the company on a cost-plus basis targeting a "fair" return for the company's shareholders. Profitability of these heavily regulated companies typically depends on management's ability to keep costs in line with, or lower than, those projected for the regulatory agency's return calculation. Profitability can shift sharply if costs drift widely from projected levels and the regulatory agency suddenly adjusts pricing up or down to restore returns to "fair" levels.

2. CRUDE OIL AND NATURAL GAS EXPLORATION, DEVELOPMENT, AND PRODUCTION

With crude oil and natural gas providing approximately 62% of the world's energy supply, the oil and gas exploration and production industry offers significant investment opportunities. That said, many of the world's oil and gas reserves are owned by the governments of the countries in which they are located and may not be invested in by outsiders. The oil and gas reserves available for investment typically are owned by large so-called integrated oil companies or smaller so-called independent oil and gas producers. Integrated oil companies ("integrateds") not only explore for, develop, and produce oil and

gas but also at least refine the oil, and most of these companies transport it and market it as well. In addition, some own and operate petrochemical plants. Investments in integrateds offer exposure to the exploration and production business, but they are not so-called pure plays. They also provide exposure to the refining, transportation, and marketing businesses.

In contrast, independent oil and gas producers ("independents") mainly explore for, develop, and produce oil and gas, and they rarely engage in other businesses. Although nearly all integrateds are public companies and have equity market capitalizations of $20 billion to $200 billion U.S., many independents are small private companies and some have capitalizations of less than $1 million U.S. The public independents typically have equity market capitalizations of $10 million to $20 billion U.S. Independents offer a pure play on the oil and gas exploration and production business.

Exposure to the oil and gas exploration and production business is also available through investments in so-called oilfield services and equipment companies ("oil service companies"), which are firms that supply specialized services and equipment to integrateds and independents to help them find, develop, and produce oil and gas. For example, oil service companies plan, perform, and analyze seismic surveys, rent onshore and offshore drilling rigs with crews, run sensors down wells to detect oil and gas (so-called wireline logging), fracture reservoirs with acid or water pressure to maximize production, and cement casing and tubing in wells to prepare them for production. The upper end of the range of equity market capitalizations of oil service firms is a little higher than that of independents.

2.1 Key Drivers of Value Addition for Oil and Gas Producers

For the oil and gas exploration and production operations of integrateds and independents, the key drivers of value addition are unit growth, pricing, and costs. As in any business, companies that add value for shareholders are those whose return on capital employed exceeds their weighted cost of capital. However, because the lead time between the discovery of new reserves and first production is often years, sophisticated analysts often include in their calculation of annual returns not only net income but also the estimated increase in the net present value (NPV) of a company's reserves.

2.1.1 Unit Growth

Oil and gas investors monitor both production and reserves growth. Typically, to grow profits, an oil and gas producer must increase production. To do this, oil and gas companies drill new production wells, boost volumes from existing producing wells, or purchase producing wells from other companies. New well-stimulation and reservoir-fracturing technologies continue to increase the rates at which oil and gas fields can be drained. To experience net growth, companies must add more production than they lose to the natural declines on their existing producing wells. Depending on the rate at which hydrocarbons flow out of a reservoir and underground pressure on the hydrocarbons drops, production from a producing well can drop 5 to 70% per year. Typically, the faster a reservoir produces, the faster its production declines. The worldwide average annual natural decline rate for crude oil production is estimated to be 10 to 15%, but for North American natural gas production the estimate is 20 to 30%. Although companies are eager to monetize their oil and gas reserves as quickly as possible to maximize their return on capital, from a practical standpoint, public companies find it difficult to manage high decline rates. The timing of reinvestment opportunities does not always match a high-production, high-decline rate company's strong cash inflows, which can lead to volatile production rates and profitability. Just as public market investors accord premium valuations to companies with high stable growth rates, they impose discount valuations on companies with volatile returns, even if the long-term growth rate is above average. A private oil and gas producer with a limited need for external capital is freer to pursue high-return, high-decline rate investments because the associated profit volatility affects its cost of capital much less.

To be able to drill new production wells, an oil and gas company must discover new oil and gas reservoirs through exploration activities or it must purchase so-called undeveloped acreage that contains discovered hydrocarbons that have not yet been fully brought into production or even brought into production at all. Larger, well-capitalized oil and gas companies that can afford the often long lead times between discoveries and first production tend to favor exploration to replenish their development opportunities, whereas most smaller firms adopt a so-called acquisition-and-exploitation strategy. They mostly buy into existing discoveries that are not fully developed and try to complete the development and produce the reserves with a more favorable cost

structure than those of the larger companies typically selling the properties. Nearly all oil and gas producers maintain a portfolio of assets ranging from so-called unproven acreage, where no discoveries have yet been made, to mature producing properties on the verge of generating negative cash flow. The weightings of higher risk, higher potential early-stage assets and lower risk, lower potential mature properties in a company's mix define its risk/reward profile.

For production growth to be sustainable, an oil and gas company must replenish not only the production it is losing to natural declines but also the underlying reserves that it is producing. Theoretically, an oil and gas company could engage exclusively in exploration and/or reserves trading activities, leaving production from the reservoirs to those companies to which it sells reserves. In reality, the handful of independents that have come closest to implementing this strategy have encountered cash flow problems and financing difficulties. As with production, to grow reserves, a company must add more reserves than it produces. Unlike production, reserves are more difficult to define. The accounting systems in the countries where stocks of companies with hydrocarbon assets trade all define reserves differently. However, there are three basic categories of reserves: possible, probable, and proved. In addition, there are three subcategories of proved reserves: undeveloped, proved developed nonproducing, and proved developed producing.

2.1.1.1 Possible These reserves are also called "unrisked potential reserves." Reservoir engineers estimate this quantity by assuming that all of the potential reservoirs visible from seismic surveys of a given set of acreage are full of hydrocarbons and that the distribution of reservoirs across the acreage where seismic has not been acquired is similar to that found across the acreage where seismic has been acquired. The financial statements of oil and gas companies rarely refer to possible reserves because they are such an approximation.

2.1.1.2 Probable These are reserves that analysis of drilling, geological, geophysical, and engineering data does not demonstrate to be proved under current technology and existing economic conditions but where such analysis suggests the likelihood of their existence and future recovery. Although the U.S. Securities and Exchange Commission (SEC) does not require oil and gas companies listed on U.S. exchanges to report estimates of probable reserves, these are reported by Canadian and U.K. oil and gas companies. As the name suggests, probable reserves are more likely to be there than not, but odds of an upward or downward revision of the estimate are nearly even. Supposedly, the estimate has been risked for a betting person, and theoretically, this is the number that investors should care about most. If an investor can find a set of probable reserves that are undervalued, he or she should purchase them because an upward revision is slightly more likely than a downward revision and a return to fair value is likely over time. In reality, many investors fear that probable reserves are estimated too optimistically.

2.1.1.3 Proved These are the estimated quantities of crude oil, natural gas, and natural gas liquids that geological and engineering data demonstrate with reasonable certainty to be recoverable during future years from known reservoirs under existing economic and operating conditions. Most investors focus on proved reserves as defined by the SEC. Nearly all major oil and gas companies in the world now have their stocks trading on U.S. exchanges, either as ordinary shares or as so-called American depositary receipts (ADRs) or shares (ADSs), so U.S. oil and gas accounting has become the de facto global standard. All companies with oil and gas reserves whose stocks trade on U.S. exchanges must report their year-end reserves according to U.S. generally accepted accounting principles (GAAP) in a Form 10-K (for ordinary shares) or 20-F (for ADRs or ADSs).

Unlike probable reserves, proved reserves are defined conservatively to make downward revisions much less likely than upward revisions. In other words, on average, oil and gas companies have more reserves than their reported proved reserves. This partially explains why oil and gas companies typically trade at premiums to reasonable estimates of the liquidation value of their proved reserves. Investors correctly assume that the companies' actual reserves are probably larger than the reported proved amounts. The other explanation for the typical premium is growth potential. Unless an oil and gas company is a so-called royalty trust or is poorly managed, investors assume that it is not a static entity in liquidation but rather a profitable growing enterprise.

There are three types of proved reserves: undeveloped, proved developed nonproducing, and proved developed producing:

- *Undeveloped.* These are reserves that are expected to be recovered from new wells on developed acreage where the subject reserves cannot be

recovered without drilling additional wells. These reserves have been "proved" by a successful exploration well, but they have not been readied for production. No proved reserves of any kind are typically booked according to seismic interpretation alone. After the first successful exploration well, incremental proved reserves are booked as additional successful wells are drilled and define the size of the discovery. Most oil and gas companies are reluctant to book any proved reserves until a development plan for the discovery has been approved by the partners in the field as well as by the host government. Given the high-cost and time-consuming nature of capital-intensive development work (e.g., drilling wells, building production platforms and facilities), proved undeveloped reserves are usually worth only 15 to 30% of proved developed producing reserves. In other words, after reserves are discovered, 70 to 85% of the capital needed to bring them into production still must be spent. Given the often multiple-year lead times between discovery and initial production, the time value of money has a much more negative impact on the NPV of proved undeveloped reserves than on that of proved developed producing reserves.

- *Developed nonproducing*. These are proved reserves that can be expected to be recovered from existing wells with existing equipment and operating methods. These proved reserves have been readied for production, but production has not commenced. Oil and gas wells often encounter reservoirs at various depths, and for technical reasons it does not always make sense to commingle production from several different reservoirs at various depths in the same well. The reserves in the reservoirs that are "in line" awaiting their turn to be produced are proved developed nonproducing.
- *Developed producing*. These reserves have the highest NPV of all the categories because they are the only reserves generating cash flow. In addition, they are reserves for which all of the exploration and development capital already has been spent.

Understanding growth in proved reserves is much more complicated than understanding production growth. For simplicity's sake, many investors focus on growth in total proved reserves, but this can be misleading. For example, a company could add significant value by spending capital to transfer proved reserves to the developed category from the undeveloped category without any growth in total proved reserves. Conversely, another company could boost proved undeveloped reserves significantly with a large discovery in a frontier area but add limited value due to a long lead time until first production and heavy capital requirements to develop the discovery. Based on growth of total proved reserves, the latter company would appear to be the faster grower, but it could easily be adding value more slowly than is the former company. In other words, the NPV of proved reserves per unit varies widely depending on their stage of development. In addition, NPV per unit is affected by the quality of the oil or gas (which determines price received), the distance from markets, and the remaining unit development costs.

The key unit of measure for crude oil is the barrel, which is equal to 42 gallons. For natural gas, it is the Mcf or unit of 1000 cubic feet. Although different grades of oil and gas have different levels of energy content, on average, 1 barrel of oil contains the same amount of energy as does 6 Mcf of natural gas or approximately 6 million British thermal units (BTUs). In the United States, investors use this approximate energy equivalency to calculate so-called barrels of oil-equivalent or cubic feet of natural gas-equivalent. For example, a company with 1000 barrels (MB) of oil reserves and 6 million cubic feet (MMcf) of natural gas reserves is said to have total reserves of 2000 barrels of oil-equivalent (MBoe) or 12 million cubic feet of natural gas-equivalent (MMcfe). Somewhat differently, in Canada, an equivalency of 10 Mcf of natural gas per barrel of oil is standard. The Canadian ratio reflects the long-term historical pricing relationship between the two forms of hydrocarbon that was driven by the premium price per BTU accorded to oil because it can be used as a transportation fuel and not only as a heating fuel. However, during recent years, the price ratio has been narrowing for global crude oil and North American natural gas because North American natural gas production is maturing faster than global crude oil production (i.e., unit costs are rising faster for gas), and increasingly stringent environmental requirements are generating a new price premium for more environmentally friendly natural gas.

2.1.2 Pricing

Oil and gas are commodities, and in a perfectly competitive market all of the producers would be price takers. Although this model accurately describes the North American natural gas market, it is too simple for the global crude oil market. Approximately 41% of global crude oil production is controlled by the nations that belong to the Organization of Petroleum Exporting Countries (OPEC), whose goals are to maximize the global

price of crude oil without losing market share and to stabilize the price of oil. OPEC controls much of the world's lowest-cost oil production, and in a free market its market share would be expected to rise. However, OPEC fears that any increase in its market share would be more than offset by a decline in prices. Indeed, if OPEC increased low-cost production faster than total oil consumption, the world would need less oil from the high-cost end of the curve and the cost of the marginal high-cost production that sets the price would fall. OPEC could boost prices by actually relinquishing market share and forcing consumers up the supply curve to even higher cost marginal high-cost production, but as OPEC discovered during the early 1980s, such a strategy is unsustainable. In 1986, Saudi Arabia lost interest in artificially maintaining the price of oil above $25 per barrel by cutting back its production because it realized that its production would soon need to approach zero to achieve its goal.

OPEC also tries to stabilize crude oil pricing by cutting back its production temporarily during times of weak demand and by moving its production up toward maximum capacity when demand is cyclically strong. Fluctuating levels of cohesion among OPEC members and economic uncertainty make the process highly imperfect.

Although OPEC has an undeniable impact on the global crude oil market, it has not been able to eliminate the cyclicality of the industry. Like the freer North American natural gas market, the oil market is characterized by alternating periods of under- and overinvestment driven by pricing. Low prices cause underinvestment, which leads to shortage, which in turn raises prices and results in ensuing overinvestment, which makes prices fall again. Investors as a group typically overreact to pricing of the moment, making oil and gas investments more attractive when hydrocarbon prices are low and less appealing when oil and gas prices are high.

Because oil and gas prices are highly volatile and have such a dramatic impact on the ability of oil and gas producers to add value, investors make tremendous efforts to forecast fossil fuels prices. However, because of the large number of factors affecting oil and gas prices, they are extremely difficult to forecast with any accuracy over any period of time. In the past, the best assumption for crude oil for the long term has been reversion to mean historical real pricing, but as was highlighted earlier, some analysts believe that a secular upward move in real crude oil prices is under way, perhaps dating back to the early 1970s. For the 100 years through 1970, the real price of crude oil (in 2001 U.S. dollars) fluctuated mostly between $10 and $20 per barrel, but since 1970 the typical range has increased to $20 to $30 per barrel. Real North American natural gas prices may have also already entered a secular upward trend based on diminishing returns to drilling activity, but for now this trend may be muted by a loss of gas demand to cheaper competing fuels such as coal and residual fuel oil.

Although investors and the press often refer to benchmark prices, such as the near-month futures contracts for West Texas Intermediate (WTI) crude oil, Brent U.K. North Sea crude oil, and Henry Hub Louisiana natural gas, the actual prices realized by each oil and gas producer from each field in each region can vary widely from the benchmarks. So-called price realizations are affected by the energy content of the oil or gas, the quantity of impurities to remove (e.g., sulfur, carbon dioxide), the quantity of valuable by-products to extract (e.g., natural gas liquids), the need for refining, and the transportation costs to major markets. In addition, oil and gas producers are increasingly using futures and options to hedge the price realizations for a portion of their future production (so-called hedging). Their main goal is to add stability to their operating cash flows so as to be able to maintain more consistent capital spending programs in the face of oil and gas price volatility. Some companies also try to lock in pricing that is above historical averages when they can. One of the results of more widespread hedging is that the financial results of oil and gas operations no longer necessarily reflect spot market pricing of oil and gas during a given period.

2.1.3 Costs

Perhaps the aspect of an oil and gas operation that is most important to investors is its ability to maintain low costs and further reduce them. Essentially, even volume growth is a function of low costs. Companies that can add reserves and production at low unit costs are able to add more units with less capital. The price umbrella provided by the marginal high-cost producers provides strong returns to the industry's low-cost producers.

Accountants divide costs into several categories: lease operating expense, production taxes, exploration expense, depletion, depreciation and amortization, general and administrative expense, interest expense, income taxes, and dividends.

2.1.3.1 Lease Operating Expense These are the costs incurred to maintain and run the facilities that

produce the oil and gas. Low lease operating expenses are typically driven by high flow rates per well and widespread automation. These costs usually range from less than $1 per barrel of oil-equivalent at prolific world-class oil and gas fields up to more than $8 at high-cost marginal operations. Average lease operating costs are $3 to $4 per barrel of oil-equivalent for diversified public companies.

2.1.3.2 Production Taxes These are severance, ad valorem, and excise taxes on the revenues generated from sales of oil and gas produced. They vary widely from one fiscal regime to another. Investors typically analyze them as a percentage of revenues. The percentage usually ranges from 0 to 15%.

2.1.3.3 Exploration Expense Most oil and gas companies use so-called successful efforts accounting as opposed to full-cost accounting. In successful efforts accounting, all expenditures for exploration activities are immediately expensed during the period in which they are recognized to have been unsuccessful. Only exploration spending that results in booked proved reserves is capitalized as property, plant, and equipment. In contrast, in full-cost accounting, all exploration spending is capitalized regardless of the outcome. Exploration expense is the periodic charge off of the exploration spending that has been unsuccessful at "successful efforts" companies. For oil and gas companies with significant exploration programs, the number can be large at times and typically fluctuates widely. In contrast, companies adopting an acquisition-and-exploitation strategy generally have low and stable exploration expense. For "full-cost" companies, these expenditures pass through the income statement as depletion of capitalized reserve value over time and occasionally in a single large charge called a ceiling test writedown. Every quarter, full-cost companies are required by the SEC to compare the book value of their oil and gas reserves to an estimate of the NPV of their reserves assuming quarter-end spot oil and gas prices. When oil and gas prices dip, full-cost companies with a lot of capitalized unsuccessful exploration spending usually must take ceiling test writedowns.

2.1.3.4 Depletion When oil and gas companies spend money to explore for and develop reserves, most of it is capitalized as property, plant, and equipment. As the reserves are produced, companies take a noncash charge for every barrel of oil and Mcf of gas produced. This allows the companies to match the revenues from selling production during a given period with the costs incurred to generate that production, even if some of those costs were incurred many years earlier. Although there are different accounting methods, essentially the per unit charge is calculated by dividing the net capitalized costs associated with the proved developed reserves by the number of reserves. Unit depletion swings in a range very similar to lease operating expense: from less than $1 to more than $9 (U.S.) per barrel of oil-equivalent. The average for public companies is $4.50 to $5.50.

Over time, unit depletion should be similar to so-called unit finding, development, and acquisition costs. For oil and gas companies, investors monitor the annual cost of adding proved reserves. They divide the total annual spending for exploration, development, and acquisitions by the number of proved reserves added. Although this is an imprecise calculation on its own without reference to the NPV of the reserves added and not accounting for the number of proved reserves moved to the developed category from the undeveloped category, investors look for unit finding, development, and acquisition costs of diversified oil and gas companies to run less than $6 per barrel of oil-equivalent over 3- to 5-year periods. Note that for successful efforts companies and full-cost companies that have taken significant ceiling test writedowns, unit depletion is typically less than unit finding, development, and acquisition costs because a portion of these costs has been expensed through exploration expense or ceiling test writedowns.

2.1.3.5 Depreciation and Amortization With most assets of oil and gas operations in depleting reserves and not in depreciating factories or amortizing goodwill, this is not usually a significant expense.

2.1.3.6 General and Administrative Expenses Although these are not generally large for capital-intensive oil and gas companies, low unit general and administrative expenses can be a sign of corporate efficiency. However, it is usually not possible to calculate unit general and administrative expenses for integrateds because general and administrative expenses are not broken out for the exploration and production division but instead are reported for the whole company.

2.1.3.7 Interest Expense This is the interest on a company's debt as well as fees to maintain credit lines. It is important to know how much debt is at

fixed interest rates and how much is at variable rates. Variable rate debt is usually less expensive but subject to the risk of interest rate increases. For oil and gas companies, investors are typically comfortable with ratios of debt to total book capitalization of 45% or less and ratios of debt to operating cash flow of two times or less. Of course, operating cash flow fluctuates with oil and gas prices, so investors' comfort with a given debt level may change with the oil and gas price outlook.

2.1.3.8 Income Taxes Like production taxes, these vary from one fiscal regime to another. In general, the lower the operating costs of extracting hydrocarbons, the higher the taxes imposed by the government with jurisdiction over the reserves. Essentially, governments try to "levelize" the risk-adjusted returns available to oil and gas companies in the global marketplace for reserves. If a country is fortunate enough to hold low-cost oil and gas reserves, its government tries to keep the excess return for itself, allowing the oil and gas company extracting the reserves a return only marginally high enough to be competitive. That said, actual after-tax returns vary widely because the marketplace demands upside potential in exchange for assuming high risk, and when high-risk projects work out the payoff can still be large.

2.1.3.9 Dividends Some oil and gas companies include preferred stock in their capital structure and pay cash dividends on the preferred stock. These dividends, unlike interest on debt, are typically not deductible from income taxes. In general, the larger an oil and gas company, the more likely it pays a cash dividend to common shareholders and the larger the dividend itself. Larger oil and gas companies usually cannot find good reinvestment opportunities for all of their operating cash flow and so return some of it to common shareholders in the form of cash dividends. Although cash dividends are income tax-inefficient because common shareholders are taxed twice on the company's profits (once as taxable income to the corporation and then again as dividend income to the shareholder), they continue to be paid largely as a way of signaling stability to investors. An oil and gas company that can afford to pay a high dividend is viewed as more stable than a firm that does not pay a dividend. Greater stability results in a higher valuation, all other things being equal, and this may reduce a company's cost of capital enough to offset the higher tax costs of paying a dividend.

2.2 Valuing Oil and Gas Producers

The inventories of reserves held by oil and gas producers make valuing these operations more complex, but also potentially more precise, than valuation of other industrial concerns. Essentially, all investments are worth the NPV of future expected cash flows. But because the further in the future, the more uncertain the cash flows, investors typically focus on expected cash flow for the next 12 months, which is difficult enough to forecast. Once this has been projected, an estimate of maintenance capital spending (usually just depletion, depreciation, and amortization) is subtracted to project free cash flow (for which accounting net income or earnings is a proxy). A target multiple is then assigned to this free cash flow (or net earnings) projection based on the investor's assessment of the accuracy of the projection as well as growth prospects and risks beyond the next 12 months.

An investor could value oil and gas assets in this way, but it would be difficult to quantify the impact on growth prospects and risks of companies' differing inventories of reserves. For example, if two companies both hold a 50% interest in the same producing oilfield but one also has a nonproducing interest in another oilfield (to be brought into production in 3 years), both would be expected to generate identical free cash flow over the next 12 months, but the growth prospects of the company with the second nonproducing interest would be superior to those of the company with only the producing interest. The question is how much higher a multiple of next 12 months projected free cash flow should be assigned to the company with both the producing and nonproducing interests.

2.2.1 Asset Values: Vice to Virtue

The fact is that oil and gas producers' large inventories of proved reserves make their future cash flows more certain than those of many other types of commercial and industrial operations; thus, they can be turned into a valuation advantage. Making key assumptions about natural production decline rates, future oil and gas prices, future development costs, and the growth in future operating costs and taxes, investors can calculate a relatively precise estimate of the liquidation value of a set of oil and gas reserves for a given discount rate. The SEC recognized this fact when it decided to require all U.S.-traded companies with oil and gas assets to calculate annually its so-called Standardized Measure of Discounted Future Net Cash Flows, referred to as

the "SEC PV10" by investors because it estimates the NPV of the net cash flows expected using a standardized 10% discount rate. For the other key assumptions, companies go to their reservoir engineers and accountants, although the SEC requires that all companies use year-end oil and gas price realizations flat into the future. Companies that want to add credibility to their SEC PV10 have it audited, along with their proved reserves, by an independent reservoir engineering firm. The SEC PV10 is published is each company's 10-K or 20-F.

The SEC PV10 contains valuable information that is underused by many investors. The key analytical problem with the SEC PV10 is that each year a different set of price assumptions is used according to year-end pricing. For example, if an investor believes that the real price of oil will be approximately $20 per barrel in the future, but because the year-end spot price dipped to $14 per barrel the SEC PV10 shows him or her the liquidation value of a given set of reserves at $14 per barrel, the investor might not find the SEC PV10 to be a useful valuation tool. Unfortunately, this investor might end up ignoring valuable information about decline rates and future costs embedded in the SEC PV10 because of an unpalatable pricing assumption. However, the SEC PV10 offers the sophisticated investor enough information to broadly replicate the underlying discounted cash flow analysis with the pricing assumptions used, allowing him or her to solve for estimated decline rates and future costs. The sophisticated investor can then replace the year-end pricing assumptions mandated by the U.S. SEC with those to his or her liking and calculate a valuable estimate of the NPV of the expected future cash flows from a given set of reserves. To get from here to a value for a company's equity, the investor need only subtract its net debt.

Although sophisticated investors can squeeze much incremental value out of the SEC PV10 with in-depth analysis, even just as reported, the SEC PV10 can be helpful. In a given year, SEC PV10s for different companies have the virtue of having been calculated using the same year-end oil and gas price assumptions. Investors trying to understand the true value of a company's proved reserves can divide its SEC PV10 by its number of proved reserves to calculate its unit SEC PV10 (expressed in dollars as SEC PV10 per barrel of oil-equivalent). In comparing this value to the SEC PV10s per barrel of oil-equivalent of other oil and gas assets, investors can discover by how much the NPV of the company's reserves is at a premium or discount to the industry average. Investors can then compare the premium or discount of the company's market value (market equity plus net debt) per barrel of oil-equivalent relative to the industry average to its unit SEC PV10 premium or discount to assess the accuracy of its market valuation relative to that of peers. Note that this valuation analysis is less effective at the end of years in which the relationship of the price of oil to that of natural gas is atypical, and particularly in analyzing oil and gas companies with an oil and gas mix very different from the industry average. The closer the mix to the industry average, the more effective the analysis.

In practice, unit SEC PV10s range from $*X*/Boe to about 3*X*, where *X* depends on pricing assumptions. As highlighted previously, reserves are more valuable, the more developed, the faster the production rate, the higher the quality of the oil and gas, the closer to major markets, the lower the development and operating costs, and the more benign the fiscal regime where they are located. Although some investors assign a fixed asset value per barrel of oil and per Mcf of natural gas to all reserves, analysis of unit SEC PV10s demonstrates how needlessly imprecise this form of analysis is.

2.2.2 Cash Flow: Simple but Potentially Misleading

Many investors value oil and gas assets based on operating cash flow or earnings before interest, taxes, depletion, depreciation, amortization, and exploration expense (EBITDAX). Like the free cash flow (or net earnings) analysis mentioned previously, this is not a formal discounted cash flow analysis but rather an analysis of next 12 months expected operating cash flow or EBITDAX and the appropriate multiple to apply to it. Investors typically focus on expected operating cash flow before volatile and difficult to forecast working capital changes, which is essentially the sum of net income, depletion, depreciation, amortization, exploration expense, and deferred income taxes. This is also called discretionary cash flow because it is cash flow that management has the discretion to use in multiple ways (i.e., exploration, development, acquisitions, debt reduction, stock repurchases, dividend increases).

When it comes to oil and gas assets, investors do not like free cash flow or net earnings because the volatility in oil and gas prices makes these residual metrics even more volatile. Multiples of free cash flow and net earnings are virtually unanalyzable much of the time. In contrast, operating cash flow and EBITDAX have the virtue of being less

"residual" and, thus, less volatile, but they have the defect of offering less useful information. If an earnings-like valuation metric does not account for maintenance capital spending (as operating cash flow and EBITDAX do not), it has little useful information about earning power. In fact, a company that has analyzable operating cash flow, but whose maintenance capital spending is higher than its operating cash flow, is effectively in liquidation. There is no way in which to gauge this from operating cash flow or EBITDAX alone.

Operating cash flow and EBITDAX are actually more proxies of asset value than of free cash flow. They can help investors to compare oil and gas operations with each other so long as the operations have similar ratios of reserves to production. However, when one operation has a significantly higher ratio of reserves to production than does another, the analysis breaks down. Essentially, using operating cash flow or EBITDAX as a proxy for asset value means assuming a fixed relationship between cash flow and asset value, and this does not hold when one company has much more asset value (or reserves) per dollar of cash flow (or barrel of oil-equivalent of production) than does another. Even when reserves-to-production (R/P) ratios are similar, asset value per dollar of cash flow can vary for the same reasons that unit SEC PV10s differ from company to company, but varying R/P ratios show the pitfalls of cash flow valuation even more clearly.

Although reported operating cash flow and EBITDAX are less volatile than free cash flow and net earnings, they still fluctuate too widely with oil and gas prices to be solid anchors of value. Sophisticated investors often recalculate historical and projected cash flow using so-called normalized pricing, just as they calculate asset values using projections of average long-term real pricing for oil and gas. Normalized pricing makes operating cash flow and EBITDAX better proxies for asset value, although it does not solve the other problems highlighted previously.

Using multiples of market value (defined as market equity plus net debt) to EBITDAX to value oil and gas assets carries all the problems listed previously, but multiples of market equity alone to operating cash flow have yet another defect. Operating cash flow moves with financial leverage. When an oil and gas producer purchases oil and gas properties using incremental debt, its operating cash flow nearly always rises even if there is no increase in its net asset value. In fact, operating cash flow typically increases even if net asset value falls. This occurs because without reinvestment, operating cash flow from oil and gas properties falls with natural production declines while the interest on the incremental debt is constant (assuming fixed-rate debt). Even if the NPV of the expected stream of operating cash flows from the oil and gas properties is below the NPV of the stream of expected interest payments (i.e., shareholder value has been destroyed in the purchase), the first 12 months of operating cash flow nearly always exceeds the first 12 months of interest, resulting in a net increase in operating cash flow for the acquiring oil and gas producer. (Of course, this does not occur if the properties acquired are relatively underdeveloped and production is low relative to reserves.)

The analysis is slightly different, but the conclusion is identical, if one looks at the acquired properties as a going concern rather than as a simple liquidating entity. If they are a going concern, a large portion of the operating cash flow needs to be reinvested to maintain and grow production. The expected future net cash flows from the properties become the flat or rising operating cash flow minus the capital required to maintain or grow the production. Even if the NPV of this expected cash flow stream were lower than the NPV of the interest on the incremental debt used to fund the acquisition (i.e., shareholder value is being destroyed), expected operating cash flow for the next 12 months for the acquiring company would rise because the acquired properties' operating cash flow would exceed the interest on the debt assumed to fund the purchase.

The valuation problem arises when investors decide that the equity of a certain oil and gas producer is worth a given multiple of its operating cash flow and that producer increases its operating cash flow through a debt-financed acquisition that adds no net asset value. These investors might mistakenly conclude that the value of the equity has risen with the operating cash flow. Instead, the value of the equity in this example is unchanged, and the appropriate multiple of the market equity to operating cash flow has actually declined. In fact, high-debt oil and gas independents typically see their equity trade at steep discounts to the peer average relative to operating cash flow, and this is not only because of concerns over the risks of high leverage. These same high-debt companies usually see their total market value (market equity plus net debt) trade at much smaller discounts, if any, to the peer average relative to EBITDAX. This unleveraged multiple corrects for the distortions created by the timing differences between oil and gas property operating cash flow and interest payments.

2.3 Oilfield Services and Equipment Companies: Less Would Be More

Investors who believe that rising hydrocarbon demand and the declining hydrocarbon resource base are about to overwhelm the pace of technological improvements like oilfield services and equipment ("oil service") investments. The less oil and gas the industry gets out of every square kilometer shot with seismic and every meter drilled into the earth, the more demand for oilfield services and equipment. The oil service industry has leverage to diminishing returns to extraction activity, and unlike the producing industry, it should experience a minimal cost squeeze. Although oil and gas producers must purchase many industry-specific services and equipment, oil service companies purchase many industry-nonspecific inputs (e.g., steel, electronic components, construction services) that they render oil and gas industry specific.

Of course, leverage cuts both ways. When oil and gas prices decline, producer capital spending, which is tied to producer cash flows, typically falls more, and producer capital spending is oil service companies' revenues. The equities of many oil service companies are among the worst performers during oil and gas industry downturns. And if an unexpected technological breakthrough were to result in lower real oil and gas prices and less demand for oilfield services and equipment on a secular basis, oil service companies would probably be the worst performers in the oil and gas industry.

Fortunately, the financial analysis of oil service companies is less complex than that of oil and gas producers. Oil service companies are industrial concerns that purchase capital and labor inputs and transform them into services and equipment that are sold to oil and gas producers. They do not have large inventories of hydrocarbon reserves at various stages of development to value. However, operationally, they can be more complex to analyze. Some produce proprietary patented products and services, for which they can extract monopolistic or oligopolistic pricing until competitors develop even better products or services. Investors need to understand the market for each key product or service that a given oil service company offers. For each product or service, is the market growing or contracting, and at what rate? Is this company gaining market share with superior offerings in terms of quality or price, or is it losing share? Is pricing in each market rising due to high demand or underinvestment, or is it falling due to low demand or overinvestment? Are this company's costs increasing or decreasing, and what is the trend for competitors' costs? These all are questions that investors in oil service companies must address, and they are similar to the issues facing investors in any industrial concern.

Although financial theory would dictate that oil service companies be valued based on free cash flow or its accounting proxy, net earnings, many investors also focus on operating cash flow and earnings before interest, taxes, depreciation, and amortization (EBITDA). As with oil and gas producers, investors find reported net earnings difficult to analyze because of the broad cyclicality of the industry. Reported net earnings of oil service companies are sometimes unanalyzable for multiple quarters during lengthy industry downturns. Some investors use estimates of so-called normalized earnings, but others fear that these types of estimates might not be sufficiently grounded in reality. So, many investors fall back on operating cash flow and EBITDA because they are less volatile and analyzable quantities are reported quarterly. Investors analyze historical cycles in the oil service industry and observe the multiples of operating cash flow and EBITDA at cyclical peaks and troughs as well as at mid-cycle. They then compare the current multiples with the historical ones to value the companies. Larger, more diversified oil service companies operating in the less cyclical sectors of the industry typically trade at higher multiples, as do smaller companies operating in fast-growing niche sectors.

3. COAL DEVELOPMENT AND PRODUCTION

Although coal is broadly criticized as an energy source because it creates the most pollution, it remains a very important fuel, supplying approximately 25% of the world's energy needs. Investing in coal-producing operations raises many of the same issues as does investing in oil and gas operations, with a few key differences. Unlike oil and gas, for which there is active exploration, coal is already so abundant that few companies are searching aggressively for more reserves. Instead, coal-producing operations focus their efforts on reducing extraction and transportation costs.

Like oil and gas, coal is found with different characteristics in different deposits. The value of coal rises with its energy content and vicinity to markets, and it falls with the quantity of associated sulfur and other pollutants.

Most governments are trying to reduce their countries' use of coal because of the associated pollution. However, global energy needs may continue to force up the use of coal. At the same time, the power generation industry is trying to develop technologies that will allow cleaner burning of coal so that this low-cost abundant fuel may continue to be used.

With coal so abundant and yet so unpopular, investors typically do not expect any secular increase in real pricing anytime soon. The key to value creation for coal-producing operations is to reduce extraction and transportation costs and increase their production of low-cost, high-quality coal, crowding out of the market higher cost, lower quality competing coal.

Although there are several public companies whose principal business is coal production, many coal-producing operations belong to private companies and publicly traded conglomerates. There are a limited number of publicly traded pure plays on coal production.

4. CRUDE OIL REFINING AND MARKETING

The crude oil refining business is well known to investors for its difficulty in generating attractive returns. Although the business is cyclical, profit margins typically swing from poor to mediocre and then back to poor without ever getting good. Essentially, the refining industry suffers from chronic overinvestment. Ironically, few new refineries have been built in developed countries during recent decades. However, refining capacity has continued to grow as engineers upgrade and de-bottleneck existing plants and effectively increase their capacity. Increasing environmental requirements have forced companies to invest in their refining assets to reduce emissions, and this type of work often boosts capacity with limited incremental investment. In developing countries, refineries continue to be built despite weak returns because the plants create employment and allow crude oil-producing countries to export more finished product.

Although the ability to transport refined products (e.g., gasoline, jet and diesel fuels, heating oil, residual fuel oil) around the world inexpensively creates a global market for these products, marketing profitability is highly regionalized. In truly competitive markets, it is difficult to generate a competitive advantage in marketing refined products. However, for gasoline and diesel fuel, the geographic location of service stations can be a competitive advantage. Many buyers focus on convenience more than on price, so highly trafficked locations can be advantageous. For most other refined products, as well as for gasoline and diesel fuel, it takes a monopolistic or oligopolistic market for marketers to be able to generate above-average returns. A further problem facing gasoline and diesel fuel marketers has been the emergence of so-called hypermarkets that sell transportation fuels at very thin margins to attract motorists to the other retail outlets in their shopping centers. Although the hypermarkets have the potential to more than offset their thin margins in gasoline and diesel fuel marketing with higher profits from other stores, nondiversified marketers in the same market must suffer below-average profitability to remain price competitive.

Like coal-producing assets, refining and marketing operations typically are part of conglomerates, mostly part of so-called integrateds. As we noted previously, integrateds are usually very large companies with oil and gas exploration and production operations, oil refining and marketing assets, and (often) petrochemical operations. However, there are a small number of publicly traded companies that are pure plays on refining and marketing. Investors in refining and marketing assets usually try to take advantage of the cyclicality in the industry by buying assets when margins are depressed and trading out of them when margins recover. Some investors are expecting growing global demand for refined products to outstrip supply capacity increases, which may be constrained by more and more stringent environmental requirements, but so far this has not occurred.

5. NATURAL GAS PROCESSING, TRANSMISSION, AND DISTRIBUTION

5.1 Processing

Once natural gas has begun to flow from the wellhead, it must be treated and transported to end users. The degree to which it must be processed depends on both its natural gas liquids (NGLs) content and its impurities. Natural gas often contains NGLs, which comprise up to 15% of the volume. These include ethane, propane, butanes, and pentanes. Approximately 60% of NGLs produced go into the refinery and petrochemical markets, whereas the other 40% are used for space heating. Because

NGLs usually have approximately 65% of the energy content of crude oil, they normally trade at a composite 65% of crude prices. When they occasionally trade below that level, the lighter NGLs (ethanes and some propanes) are kept in the gas stream and sold for their heating value. The impurities include brine, silt, and water (so-called BS&W), and these are filtered out with the help of a separator. In several producing basins, gas is produced with high proportions of sulfur (hydrogen sulfide) and carbon dioxide. The hydrogen sulfide and carbon dioxide are removed before the gas enters the pipelines that take it to end users.

Until the energy industry consolidation of the 1990s, there were public companies for which natural gas processing was a dominant business. However, today nearly all natural gas processing operations are part of much larger energy conglomerates, so there are no more pure plays on natural gas processing. That said, the key profit driver of natural gas processing operations is the spread (or gross margin) between composite NGL prices and the "wet" gas feedstock, which is known as the fractionation (or "frac") spread. The typical NGL barrel contains about 37% ethanes, 33% butanes, 10% normal butanes, and 13% pentanes (or natural gasoline). The frac spread in the United States has averaged approximately $6/barrel (or $0.14/gallon) for the past decade or so, but it has been volatile due to the volatility of both crude oil and natural gas prices.

Another determinant of processing profitability is the type of gas purchase contract used by the NGL processors. There are three types of contracts: keep whole, percentage of proceeds (POPS), and fee based. Keep-whole contracts were the norm historically but have become obsolete due to the increased volatility of gas prices. With this type of contract, the NGL processor buys so-called wet gas at the plant gate, strips out the NGLs, and resells them and the now "dry" gas (just methane) back into the market. The processor takes all of the price risk. Today, approximately 35% of NGL volumes are still processed under keep-whole contracts. In contrast, processors with POPS contracts agree to pay their wet gas suppliers a fixed percentage of their proceeds from reselling the NGLs stripped out and the residue dry gas, so their margins are less volatile. POPS contracts cover approximately 40% of the NGL market. The remaining 25% of NGL volumes are processed under fee-based contracts. With this type of contract, the NGL processor simply strips out the NGLs for a fee. These contracts are most popular with Rocky Mountain processors because gas price volatility is so high in this region of North America.

5.2 Transmission

Natural gas transmission refers to the transportation of natural gas from a producing area to a consuming area, whereas distribution indicates the transportation of natural gas from so-called city gates on the transmission lines to the very locations where it is burned (e.g., residences, office buildings, factories, power plants). For distances of up to 4000 to 5000 miles across land or shallow water, natural gas is typically transported via pipelines. For longer hauls, and especially across deep bodies of water, natural gas is usually liquefied at low temperatures (–260° Fahrenheit) and transported to markets in seafaring liquefied natural gas (LNG) tankers. The methane is then warmed up and re-gasified at LNG terminals near end-user markets. Historically, the high capital and operating costs associated with the construction and operation of natural gas liquefaction facilities and LNG tankers have made LNG most attractive to energy-importing countries more interested in diversification and security of energy supply than in lowest-cost energy. Liquefaction facilities cost more than $1 billion, LNG tankers cost approximately $160 million, and re-gasification plants cost approximately $300 million. Japan, South Korea, and Taiwan, which rely on imports to meet most of their energy needs, have chosen to purchase LNG from Southeast Asia to reduce their reliance on oil from the Middle East.

Because natural gas pipelines are costly to build, it is often most economical for there to be only one or two lines from a producing region to a given consuming one. This market structure would give the pipeline owners a monopoly, so governments typically intervene to protect consumers from monopoly pricing of transmission services. Governments set the prices to be charged so that the owners of a pipeline can earn a "fair" return on their investment. In the United States, these returns are normally calculated as a 12% return on stockholders' equity, assuming a 40 to 60% debt-to-equity capital structure. Periodically, U.S. pipeline owners have the right to file for new tariffs via a so-called general rate case. However, as many as 15 of the 30 transmission companies in the United States have not filed rate cases during the past decade or so. Because of the way in which rates are designed and the lack of any rate case requirement, U.S. interstate pipelines typically offer among the steadiest streams of earnings in the energy industry.

Until April 8, 1992, U.S. natural gas pipeline companies were gas merchants, buying the commodity from producers and reselling it to local distribution companies (LDCs), power plants, and industrial users. The Federal Energy Regulatory Commission's Order 636 reregulated the pipelines as so-called common carriers. Pipelines now sell space to owners of natural gas that want it transported. Owners may include marketing arms of diversified energy companies as well as natural gas producers or consumers. Pipeline customers can choose so-called firm transportation, which guarantees them space in a pipeline over a set time period (often 5–10 years), or they can buy often-discounted "interruptible transportation," which provides them with transportation services on a space-available basis. Typically, industrial customers and power generators that can switch between natural gas and residual fuel oil use interruptible transportation to minimize costs. In contrast, LDCs that have commitments to provide gas to heat residences in winter must use more expensive firm transportation.

Natural gas pipelines are owned mostly by publicly traded so-called diversified energy companies, which may also own unregulated oil and gas exploration and production assets, gas processing plants, and natural gas trading and marketing operations as well as other regulated utilities such as local gas distribution companies and even power generation, transmission, and distribution assets. The profits of regulated natural gas pipeline companies are driven by their ability to find new capital projects that regulators will approve and to maintain as wide a spread as possible between their regulated revenues and their costs. Earnings are also affected by regulators' changing definitions of a "fair" return on assets. Regulators' assessments of fair returns can depend on interest rates and perceived equity risk premiums. Natural gas pipeline companies in North America typically have limited expansion opportunities, generate significant free cash flow, and pay common dividends above stock market averages.

A corollary business to natural gas transmission is gas storage. In the colder regions of the world, a large portion of natural gas consumption is for space heating, which creates wide seasonal fluctuations in demand. In these countries, gas storage is critical to meeting peak seasonal demand requirements without having to build up excess productive capacity that would remain idle much of the year. Because gas is expensive to transport over large distances, unlike crude oil, it is not economical to move gas hemisphere to hemisphere to satisfy seasonal fluctuations in consumption. Technology has been developed that allows natural gas to be reinjected in emptied oil and gas reservoirs. Gas storage caverns are also owned mostly by diversified energy companies, which rent space to electric and gas utilities, wholesale merchants, and other commodity players.

5.3 Distribution

LDCs in North America are both natural gas merchants and common carriers, but some gas marketing companies would like to see all of them become common carriers so that they could sell gas directly to residential and commercial consumers. But so far, most so-called retail unbundling programs have not been successful. Supporters of current regulation question the benefit to consumers of being allowed to choose from which company to buy a commodity such as natural gas, especially given that in many instances LDCs do not charge a premium over their cost of gas and just pass it through. Whether local distribution companies remain merchants or become common carriers like transmission companies, their monopolies on gas distribution in a given region will almost certainly continue to be regulated to protect consumers from monopoly pricing.

Like transmission companies, LDCs charge prices that are set by regulators to give the companies a "fair" return on their investment. Profitability is driven by factors similar to those that affect the earnings of regulated transmission companies. LDCs are typically either publicly traded or owned by publicly traded diversified energy companies. LDCs usually have limited expansion opportunities, generate significant free cash flow, and pay common dividends above stock market averages.

6. ELECTRICITY GENERATION, TRANSMISSION, AND DISTRIBUTION

A lot of the elements reviewed in natural gas transmission and distribution are applicable to the electricity market. Historically, electricity generation, transmission, and distribution assets have been owned by regulated integrated electric utility companies. Like natural gas transportation and distribution companies, electric utilities have held monopolies in their markets and had their prices set by government regulators to protect consumers from monopoly pricing and to offer the utilities a "fair" return on investment. However, during recent

years, there has been significant political pressure to deregulate (or "unbundle") the electricity generation and transmission industry. The movement started in Chile and the United Kingdom, where each part of the business (distribution, transmission, and generation) was separated. As with the natural gas industry, transmission and distribution remain a natural monopoly and, thus, will continue to be regulated. Generation is clearly the best candidate for competitive forces.

The genesis of competition in North America and the United Kingdom stems from the development of an unregulated generation market. Early on, so-called independent power producers (IPPs) and, in some cases, so-called cogenerators (which produce steam/heat and electricity) introduced competitive power and superior technology in markets where inefficiencies and old technology dominated. Although electricity is generated by hydropower, nuclear power, and the burning of oil, natural gas, and coal, most of the innovation in generation has been associated with natural gas, which is appealing because it is environmentally friendly and plant construction costs are relatively low. The emergence of highly efficient, so-called combined-cycle gas turbines (CCGTs) has been an important driver.

There are fundamentally three different turbine technologies in the market today. The oldest is the so-called steam turbine system, in which a plant generates power by boiling water using a fuel (e.g., coal, oil, gas, nuclear fission, waste) to produce steam. The pressure from the steam turns a turbine, which in turn creates electricity. The steam is recycled back into water in the boiler, creating a loop. The second turbine technology is actually based on a technology as old as civilization in that it involves wind, hydropower, and other generation technologies that use a physical force to turn a turbine as a substitute for burning fuel. The third is the gas turbine technology, which uses the equivalent of a big jet engine to produce power; the turbine is directly injected with natural gas, which burns and turns the turbine in the process. This is also called a "simple cycle" gas turbine. A CCGT plant combines the first and third technologies. Natural gas is fed into a gas turbine, which produces both electricity and heat. The heat is then applied to a boiler, which produces steam to turn a second turbine, thereby creating two outputs or cycles.

With the emergence of IPPs and cogenerators and their new, cleaner CCGT plants, many power markets seemed ready for competition during the 1990s. However, so far it has turned out that the level of capacity use in a market has been a key determinant of the success or failure of deregulation. One of the main reasons why deregulation in the United Kingdom has worked so well is that there is substantial overcapacity in its generation market. Where supply exceeds demand by a comfortable margin, price is driven by marginal high-cost production. However, in underserved markets, price can temporarily far exceed marginal high cost as scarce power is rationed by price to the end users most desperate for it. In natural gas markets, storage and competing residual fuel oil serve as natural ceilings for prices, but electricity is difficult to store and substitute. As a result, unregulated power prices can be very volatile. In fact, over the past few years, North America has experienced periods of electricity shortage when prices, normally about $30 per megawatt-hour (MWh), traded higher than $7000/MWh in some regions.

Beyond IPPs and cogenerators, deregulation has created another new entrant on the power scene: the unregulated trading and marketing company, also known as a merchant company. Merchant companies, which have appeared mostly in North America, purchase power from generators and resell it to distribution companies and industrial customers. Merchant players try to profit from arbitrage opportunities between regions, closing out these inefficiencies. When a merchant's parent company owns power plants as well, the parent may even temporarily shut down a plant if it can make more money by selling the fuel originally committed to the plant to another user for a higher price through its merchant arm.

Although the goal of deregulation has been to increase efficiency and reduce costs to consumers, it recently resulted in apparent market manipulation in California, and there is now political pressure to reregulate electricity generation in some states of the United States. Power marketing companies continue to argue that consumers would benefit if they were permitted to choose from which company they would like to purchase electricity. The U.S. Federal Energy Regulatory Commission has proposed a so-called standard market design, which would open up the power transmission grid to the wholesale marketplace. Currently, only about 30% of the power generation market is unbundled. All power lines would become common carriers, from which marketing companies and others could purchase space on transmission lines.

Many traditional electric utilities and new independent power producers are publicly traded or are owned by publicly traded companies. Like natural

gas transportation companies, electric utilities typically have limited expansion opportunities, generate significant free cash flow, and pay common dividends above stock market averages. During the 1990s, looming shortages of power generation capacity in North America made growth companies of some independent power producers and merchant participants with plans to build new power plants. However, the slower economic growth of recent years has resulted in many of these projects being cancelled or deferred.

SEE ALSO THE FOLLOWING ARTICLES

Electric Power Generation: Fossil Fuel • Energy Futures and Options • Hazardous Waste from Fossil Fuels • Natural Gas Resources, Global Distribution of • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Petroleum Property Valuation • Resource Curse and Investment in Energy Industries • Trade in Energy and Energy Services

Further Reading

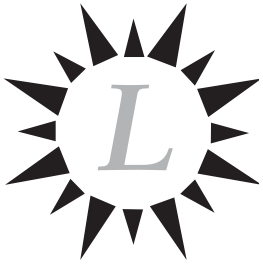
Duarte, J. (2002). "Successful Energy Sector Investing: Every Investor's Complete Guide." Prima Publishing, Roseville, CA.

Jansen, R. A. (1998). "Profits from Natural Resources: How to Make Big Money Investing in Metals, Food, and Energy." John Wiley, New York.

McCormack, J., Smith, A. L., and Fan, A. C. (2001). One way to close the oil industry's credibility gap: Hedge PDP production and buy back stock. *EVAluation* **3**(6).

McCormack, J.L.,Vytheeswaran, J. (1998). How to use EVA in the oil and gas industry. *J. Appl. Corp. Finance* **11**(3).

Labels and Standards for Energy



LLOYD HARRINGTON
Energy Efficient Strategies
Warragul, Australia

PAUL WAIDE
PW Consulting
Manchester, United Kingdom

1. **Why Labels and Standards?**
2. **Energy: The Invisible Attribute**
3. **When to Label and When to Use MEPS**
4. **Types of Labels**
5. **Regrading of Labels**
6. **Setting MEPS Levels**
7. **Review and Update of MEPS Levels**
8. **Efficiency Targets**
9. **Related Program Measures**
10. **Enforcement and Implementation Issues**
11. **Test Procedures**
12. **Tested Energy Consumption versus In-Use Energy**
13. **Performance versus Energy Consumption**
14. **Evaluation**
15. **Impacts of Efficiency Standards and Labeling Programs**

Glossary

categorical labels Where products are allocated into predetermined energy efficiency categories or classes, usually defined as a function of capacity and energy consumption; categories remain fixed, irrespective of the range of products on the market at any particular time.

check test A part or full test to verify the energy and performance claims or compliance claims, minimum energy performance standards (MEPS), or other performance requirements.

comparative labels Allow consumers to form a comparative judgment about the energy efficiency (or energy consumption) and relative ranking of all products that carry a label; comparative labels usually carry information such as size and/or capacity, energy consumption, and an energy efficiency rating scale.

continuous scale labels Where the energy consumption of the best and worst products on the market are shown, together with the relative energy consumption of the product that is labeled, allowing the consumer to assess the position of the product under consideration in comparison with the best and worst products; continuous scale labels can effectively compare products only of a similar type and size.

endorsement labels Indicate that products belongs to the "most energy efficient" class of products or that they meet a predetermined standard or eligibility criteria; these labels are essentially "seals of approval" that the products meet these requirements.

minimum energy performance standards (MEPS) A policy tool used to force low-efficiency products out of the market; MEPS either ban the sale of products that fail to meet specified minimum energy efficiency levels or restrict the availability of low-efficiency products.

Energy labeling is a policy tool that informs consumers about the energy performance of appliances and thereby encourages them to purchase appliances that provide the services they need with less energy consumption. In providing information to consumers about equipment energy consumption and operating costs that would otherwise be invisible or unavailable, energy labeling enables consumers to make more balanced and rational purchasing decisions. Energy labels can also help consumers to identify the most efficient products on the market. In effect, energy labeling attempts to provide a market "pull" for more energy-efficient products while simultaneously presenting information that might discourage the purchase of less efficient products. Energy labeling programs can be voluntary or mandatory and can vary considerably in style and type. Minimum energy performance standards (MEPS), also called minimum energy efficiency standards, are a policy tool used to force low-efficiency products out of the market. MEPS either

 © 2004 Elsevier Inc. All rights reserved.

ban the sale of products that fail to meet specified minimum energy efficiency levels or restrict the availability of low-efficiency products. MEPS are used to "push" the energy efficiency of the market. To be effective, most MEPS programs need to be mandatory, although there are examples of voluntary standards that have been effective, usually in the form of efficiency targets.

1. WHY LABELS AND STANDARDS?

Energy consumed by equipment and appliances in the industrial, commercial, and residential sectors of the developed world is a major source of global greenhouse gas emissions as well as of other pollutants such as NO_x and SO_x. Energy efficiency programs such as energy labeling and minimum energy performance standards (MEPS), where legislation and regulation are used to improve product energy efficiency, are among the most cost-effective and widely used measures employed to reduce these emissions (Fig. 1). Improved energy efficiency also has many economic benefits. Because of this, labels and MEPS are now a couple of the most common energy policy tools in use today. Nearly 60 countries (with the expansion of the European Union) now have labeling and/or MEPS programs in operation, and together these cover roughly 40 product types. Most Organization for Economic Cooperation and Development (OECD) countries now have labeling and MEPS programs in force for a range of products.

2. ENERGY: THE INVISIBLE ATTRIBUTE

Many people argue that consumers are not interested in energy per se but rather are interested in the services that energy can provide such as cold beer, hot showers, conditioned air, lighting, and operation of machinery and computers. Although this is undoubtedly true (what can one do with a raw kilowatt-hour or megajoule?), energy consumption is the key determinant of operating cost of most appliances and equipment and, therefore, is of significant concern to consumers. However, in places where energy prices are low (or where users are not charged directly for energy), energy consumption and efficiency are of less interest.

Some product attributes are readily observable on inspection by consumers or are readily available in product literature, whereas other attributes are not at all obvious. For example, in the case of a refrigerator, the volume available for the storage of food is clearly visible to consumers when the unit is inspected in a showroom. Product information also gives storage volumes and external dimensions (this information is essential for installation). Conversely, inspection of an appliance such as a refrigerator does not reveal its energy consumption, energy efficiency, or standard of performance. In fact, the determination of these performance attributes requires careful testing and the use of sophisticated equipment, which is not readily accessible to consumers.

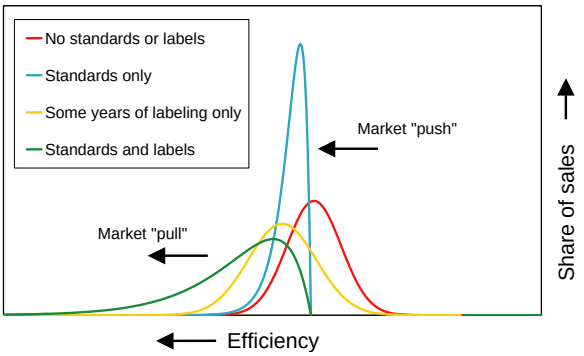


FIGURE 1 Conceptual impact of energy labeling and MEPS.

Information on nonobservable attributes can be obtained by consumers only where the manufacturer chooses to provide such information or where there is a requirement for the information to be disclosed such as through an energy labeling program. Where such nonobservable attributes are declared (and hence become visible to consumers), and where even a small market segment responds to these attributes (so-called informed consumers), manufacturers tend to concentrate on improving the rating and performance of those observable and visible attributes at the expense of other nonobservable attributes.

3. WHEN TO LABEL AND WHEN TO USE MEPS

It is important to remember that labeling and MEPS programs aim to influence the selection of products by consumers at the point of sale by making higher efficiency units more attractive (through labeling) or by making less efficient ones unavailable (through MEPS). Labeling or MEPS cannot be expected to have any significant ongoing influence on consumers' use of products once they have been purchased and installed.

Many of the necessary prerequisites for a successful labeling or MEPS program are common to both; that is, these programs are most applicable

- to products that use a significant amount of energy;
- when there is a large market (in terms of sales volume) for the product or where rapid growth is forecast;
- when there is already a range of energy efficiencies on that market or where significant improvements in energy efficiency are technically possible and economically feasible.

The main point to consider when selecting either energy labeling or MEPS is the purchasing process and payment of energy bills. Energy labeling is likely to be most effective where consumers inspect products in a retail outlet, select the products, and are also responsible for the payment of the resulting energy bills. This would typically apply to the purchase of major household appliances. Energy labeling is likely to be less effective in cases where

- a third party (e.g., a contractor/tradesperson) is responsible for the selection of the product (e.g., a water heater may be selected and installed by a plumber);
- products are not generally on retail display to allow direct comparisons between products (e.g., products are generally selected from catalogs or lists rather than on display in a showroom);
- the purchaser of the product is not responsible for the payment of the energy bills (e.g., a landlord purchasing an appliance for a tenant, capital purchases being separated from operational budgets such as energy bills in a large company or department)—the so-called "split incentives" market failure.

In these cases, purchasers are likely to be "label resistant" in that they either do not see the label or have no economic interest in selecting a more efficient product at the time of purchase to reduce operating costs. Therefore, MEPS is often a more effective policy option in these cases; removal of the least efficient models means that purchasers have no choice but to acquire a more efficient unit.

MEPS is invisible to purchasers, whereas labeling attempts to engage consumers to consider energy efficiency and use this as a selection criterion in their purchasing decisions. Labeling is an attempt to improve the operation of the market by bringing to the attention of consumers the operating cost at the time of purchase. From this perspective, even those who oppose regulatory energy policy intervention are usually in favor of energy labeling because this is seen as an essential part of smooth rational market operation. MEPS, on the other hand, is an attempt to overcome market failures such as split incentives.

An energy label works

- by providing consumers with data on which to base informed choices (to select the most efficient and suitable product available);
- by encouraging manufacturers to improve the energy performance of their models;
- by encouraging distributors and retailers to stock and display efficient products (i.e., those with good energy ratings).

In many ways, the most influential of these points is likely to be the second. If a manufacturer withdraws an inefficient product from the market and replaces it with a more efficient one, consumers are no longer able to purchase the less efficient product. For MEPS programs in isolation, the first point is not relevant, whereas manufacturers and retailers are forced (rather than encouraged) to stock more efficient products.

Often, products are most suited to either labeling or MEPS. However, some products, particularly those that are ubiquitous and that use substantial amounts of energy, may justify simultaneous use of both energy labeling and MEPS. A common product that is subject to both energy labeling and MEPS is household refrigerators. Energy labeling may provide an effective pull at the more efficient end of the market, whereas MEPS can redress those lower efficiency elements of the market that are "label resistant." Another strategy that is often employed is to introduce energy labeling initially and to assess the market impact over a period of, say, 5 years. If the rate of progress in terms of energy efficiency is judged to be too slow, MEPS may be introduced in parallel to accelerate progress. The advantage of this approach is that labeling can be used to gather a good deal of data in terms of products on the market and their energy efficiency, making subsequent setting of MEPS levels and implementation more straightforward.

4. TYPES OF LABELS

Energy labels fall broadly into two main types: comparison labels and endorsement labels.

4.1 Comparison Labels

Comparison labels are where key information on the product energy consumption and/or performance compared with that of products of an equivalent type is shown for comparative purposes. Additional measures of energy efficiency (e.g., a star, numerical, or other efficiency rating) may also be shown. This type of system works best when it is mandatory for all products to carry a label (so that poor performers can be more readily identified and avoided by consumers). Experience has shown that where labeling is not mandatory or where mandatory provisions are not enforced, energy labels on appliances with lower ratings are actively removed by retailers to improve their chances of selling these products (retailers and manufacturers actively dislike having products with poor energy ratings on display). Comparison labels tend to be sponsored by governments, although there are exceptions.

4.2 Endorsement Labels

Endorsement labels indicate that the products carrying the labels belong to the "most energy-efficient" class of products or meet a predetermined standard or eligibility criteria. These labels generally consist of a logo or mark that indicates the products have met the standard and generally contains little or no comparative energy efficiency information (although this may be available through lists of endorsed products). These labels merely inform consumers that the products meet the required standard. Criteria for some schemes are updated on a regular basis; therefore, these labels may carry the year of qualification. Endorsement labeling programs are mostly of a voluntary nature. An endorsement label may be specifically for energy efficiency or may be a broader "eco" label. Eco label programs endorse products that have low impact across a wide range of environmental factors; even though energy consumption levels often have a high priority within the rating system. It is a system that works best when only a limited proportion of the market is eligible for endorsement, typically the top 10 to 40% of performers. Endorsement labels can be sponsored by governments, but sponsorship by utilities, industry, and environmental groups is also very common.

Endorsement and comparative labels can coexist and do so in many countries and for many products. They tend to be complementary rather than competitive.

The most commonly used comparison labels employ a scale with absolutely defined energy efficiency categories (so-called categorical labels). These labels allow consumers to easily assess the efficiency of a product in relation to an absolute scale by means of a simple numerical or ranking system. The concept is that it is much easier for consumers to remember and compare a simple efficiency ranking scale (e.g., 1, 2, 3; 1 star, 2 star, 3 star; A, B, C) than to remember and compare energy consumption values.

The other type of comparison label design is continuous scale labels, where the best and worst energy consumption of products on the market is shown together with the relative energy consumption of the products that are labeled, allowing consumers to assess the position of the products under consideration in comparison with the best and worst products available.

In general, the visual designs of comparison labels in use around the world (Fig. 2) can be grouped into three basic types:

- *Dial label.* This type of label has a "dial" or gauge, with greater efficiency linked to advancement along the gauge (more efficient represented by a clockwise arc). This type of label is used in Australia, Thailand, and Korea. The number of stars or the "grading" numeral on the scale depends on the highest preset threshold for energy performance that the model is able to meet.
- *Bar label.* This type of label uses a bar chart with a grading from best to worst. All grade bars are visible on every label, with a marker next to the appropriate bar indicating the grade of the model. This approach is used in Europe, Russia, and Mexico as well as in some South American, African, and Middle Eastern countries.
- *Linear label.* This label has a continuous linear scale indicating the highest and lowest energy uses of models on the market, locating the specific model within that scale. Because energy (rather than efficiency) is used as the comparison indicator, it is necessary to group models into similar size categories for comparison. This model is used in the United States and Canada.

There are also many other comparative energy labels that have no graphic concept to support the indication of energy efficiency. These generally rely on text to explain the efficiency or some numeric indicator of efficiency (e.g., EER for air conditioners). Some labels show only energy consumption.

The majority of endorsement labeling schemes use a wide range of criteria to determine eligibility. Although energy often plays a prominent role in

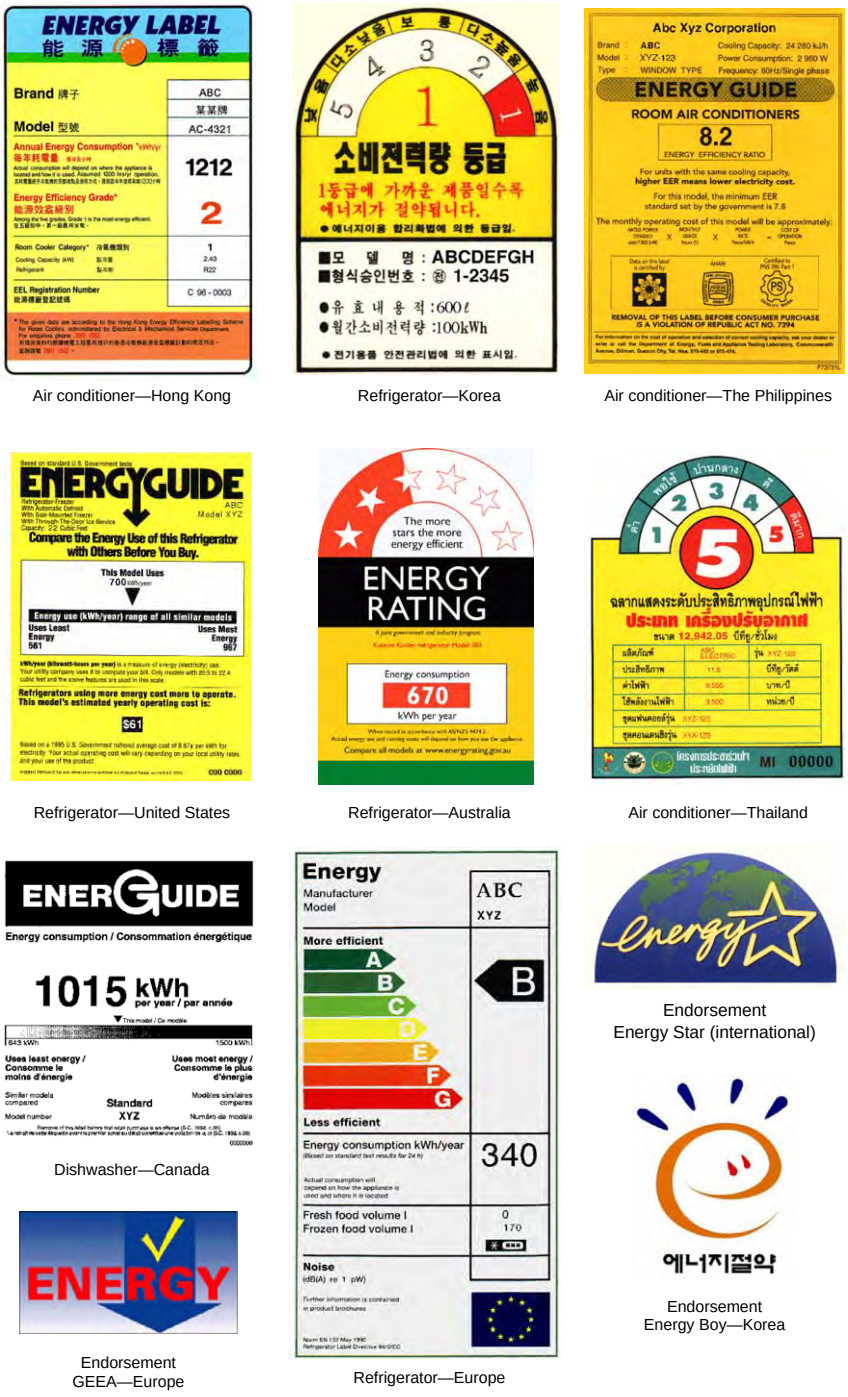


FIGURE 2 Examples of energy labels from around the world. Labels are presented for illustrative purposes only, and are not to scale or the correct color.

many of these programs, it is by no means the predominant criteria in most schemes. However, there are a few endorsement labeling schemes in place that focus exclusively on energy and energy efficiency. These are the U.S. Environmental Protection Agency (EPA) Energy Star label (used as an international endorsement label for information technology equipment and consumer electronics, with the emphasis on standby power consumption for these products, and domestically in North America for a wider range of products), the Korean Energy Boy label (essentially covering international

products and criteria covered by Energy Star), and the European Group for Efficient Appliances label for consumer electronics (emphasizing standby power consumption for these products).

Labeling programs are not necessarily restricted within country borders. For example, the European Union (EU) label not only extends to EU member states but also has been adopted by several other European and non-European countries. The U.S. Energy Star program for office equipment has been adopted in many countries around the world. In the South and Latin American region, attempts have been under way to design and implement a unified energy labeling program.

The purpose of an energy label is to convey key information to consumers to assist them in a purchase decision. Therefore, it stands to reason that there are likely to be key differences in the type of information that it is necessary to convey to consumers, depending on their socioeconomic, physical, and cultural contexts. Differences in language and, to a lesser extent, literacy are also key factors that will influence the design and presentation of an energy label within a particular region and culture. If an energy label is to be effective, it has to be understood, liked, and used by consumers.

5. REGRADING OF LABELS

If an energy label is doing its task, the efficiency of products available on the market will improve gradually over time. This means that adjustments will need to be made to energy labels to keep them relevant.

In the case of continuous scale labels, where the best and worst products currently on the market are shown, it is necessary to update the end points of the labels on a regular basis, typically annually. Although this keeps the labels relevant, it creates administrative difficulties in that new end points need to be regularly defined, notified, and distributed. Sometimes, labels with older and newer end points can end up in the same store and there is nothing to identify these differences. This system creates problems if a new product that lies outside the current range comes onto the market between updates (this is difficult to convey to consumers in a nonambiguous fashion). The other issue is that continually changing scale end points acts as a disincentive for manufacturers to improve the efficiency of their products in that they have no idea where their products will lie relative to end points that are continuously changing. There is also evidence that continuous scale labels are more difficult for consumers to use in an effective manner.

For categorical labels, the efficiency category definitions usually remain fixed for a considerable time, typically 5 but up to 10 years (or longer if the energy efficiency progress in the market is slow). From manufacturers' perspective, this tends to encourage the development of more efficient products because they know during the development phase how their products are likely to rate on the categorical efficiency scale for some time to come and will make some effort to reach the next highest rating category if this can be done in a cost-effective manner. As products improve in their efficiency over time, there tends to be efficiency category creep; that is, the proportion of high-efficiency ratings increases and that of low-efficiency ratings decreases (Fig. 1). The disadvantage of the categorical system is that when a rating system has to be changed or regraded (and it will have to be revised eventually), there is likely to be some disruption to the market during the transition. Model efficiency will appear to fall as categories are regraded, and the appearance of labels with two different rating scales in retailers at the same time both need to be handled carefully to avoid consumer confusion. The other element of difficulty is that manufacturers tend to be strongly resistant to a downgrading of the current efficiency ratings for their products (they dislike supplying products that have very low efficiency ratings; however, there is no point in having a categorical label scale where the bottom two or three scales are empty, with all models bunched into the higher efficiency categories). So, an orderly and well-planned transition is essential.

6. SETTING MEPS LEVELS

MEPS can be set at a range of levels, from gentle to aggressive. At the gentle end of the scale, MEPS levels would eliminate only a modest part of the current market, would eliminate only the least efficient models currently offered for sale, or would even set as a backstop that does not eliminate any current models but would stop any lower efficiency models from entering the market. At the most aggressive level, MEPS can be set at levels that eliminate all current models from the market and/or mandate the adoption of new technology and designs that are at the current leading edge to meet the prescribed efficiency levels. Of course, there is a continuum between these levels, and the precise level

selected will depend on a range of factors such as capability of local manufacturers (and whether there are any), manufacturing costs, local energy prices, and incomes. For modest MEPS levels, the implementation period can be relatively short (1 year or less), but for more stringent levels, the implementation period tends to be longer (up to 3 years in more extreme cases). The balance of local manufacture and imports for a particular product may also influence the levels and timing of MEPS.

There are two traditional methods for setting MEPS levels: the so-called statistical approach and the engineering approach. In the statistical approach, efficiency levels are set so that the market impact is manageable in terms of the number of models eliminated (typically 20–50% of models would be eliminated from the market, with at least some models left in all major categories and types of products). In this approach, the MEPS levels are defined in terms of the range and characteristics of products currently on the market. This approach is practical only where there is some significant variation in the efficiency of products already on the market.

The engineering approach is where a technical analysis of product efficiency is used to assess a range of MEPS levels. Typically, a range of measures and their incremental cost are assessed against a base case market product. The cost-effectiveness of each additional measure is assessed to determine the optimum efficiency level. Other factors that are typically considered when setting MEPS using an engineering approach are the capability and financial position of local product manufacturers, energy prices, and likely impact on retail product prices after MEPS.

An alternative hybrid approach to setting MEPS levels is to review the levels already in use by a country's major trading partners and to adopt one of these as a local MEPS level. This approach has the merit of avoiding questions regarding the technical feasibility of the proposed MEPS level and also ensures that suitable products would be available in the market after the implementation of MEPS. Harmonization of MEPS levels, although of secondary importance, would also reduce restrictions to free trade of products. This approach is feasible only where test procedures and product types in each of the relevant market are similar or related.

MEPS as a policy option has an advantage over energy labeling in that faster, larger, and more certain energy savings can usually be achieved through implementation of aggressive MEPS levels. Although this may be justified in terms of cost-effectiveness, implementation of aggressive MEPS levels can create problems and difficulties among manufacturers, especially where they are forced to invest large amounts of capital to meet MEPS requirements. In this regard, enforcement of MEPS, especially where the levels are aggressive, is critical to protect the investments of those manufacturers that comply with the requirements.

7. REVIEW AND UPDATE OF MEPS LEVELS

As with energy labels, improvement of efficiency levels within the market may warrant the reassessment of MEPS levels from time to time. The approaches used to review MEPS levels are the same as setting initial levels and depend on the market structure, mix of products, and level of stringency proposed. Where overseas levels are adopted locally, a periodical review of changes in overseas MEPS levels may be warranted.

Strategically, there is merit in setting initial MEPS levels at a moderate level (even with a short implementation time frame) to allow the regulatory system to be initiated and proved in a framework where the stakes for the main players are not too severe. Once the system operation is settled and manufacturers are accustomed to the concept and processes involved, the scene is set for implementation of more stringent or aggressive MEPS levels.

One issue to bear in mind is that when more stringent MEPS levels are implemented for products that also have energy labels, it may be necessary to regrade the energy label categories to take account of the new MEPS levels because all low-efficiency categories may be eliminated from the market. In this respect, close coordination between MEPS revisions and energy labeling revisions need to be maintained. In addition, the implementation of very aggressive MEPS levels may reduce or even eliminate energy efficiency differences among products, obviating the benefits of energy labeling, at least in the short term. The viability of energy labeling in conjunction with very aggressive MEPS levels (especially where these are upgraded regularly) would need to be reviewed from time to time.

8. EFFICIENCY TARGETS

As an alternative to regulated mandatory MEPS programs, it is possible to have program variations

that achieve comparable outcomes yet are implemented on a voluntary basis. These usually take the form of voluntary agreements between governments and industry to achieve prescribed efficiency targets. Efficiency targets may be sales-weighted averages within each manufacturer's sales or product range or may be the voluntary elimination of products that do not meet prescribed minimum efficiency levels. These arrangements work best where strong industry associations (whose members dominate the supply of products to the market) are a party to the arrangements. Examples of these types of arrangements can be found in Europe for products such as televisions, videocassette recorders, clothes washers, dishwashers, water heaters, and external power supplies. A related approach is used in Japan (the "Top Runner" program), where the government sets voluntary efficiency targets that are based on the most efficient products on the market today. Manufacturers are asked to ensure that their sales-weighted average efficiencies meet these targets by the prescribed date (typically 5 years). Although there is no regulatory framework to enforce the voluntary targets, manufacturers clearly understand that those who fail to meet these will be publicly named and humiliated by the government, making the system quasi-mandatory in effect.

9. RELATED PROGRAM MEASURES

As an adjunct to energy labeling and MEPS, there are a number of related measures that can complement these programs. These include retailer support, integrated promotion, efficiency awards, and lists of current models on the market.

9.1 Retailer Support

Retailers play a key role in the effectiveness (or lack thereof) of an energy labeling program. A hostile retailer can neutralize the impact of an energy label by discrediting the information or the label's validity or message. Retailer education and supply of point-of-sale information can help to build the effectiveness of the program.

9.2 Integrated Promotion

Where a number of products are included in a regulatory program, there is an opportunity to market the program in a coordinated fashion and to deliver to the marketplace complementary messages that reinforce the importance of energy efficiency. This can extend to other products, such as homes, cars, and windows, that have a similar rating system.

9.3 Efficiency Awards

An awards scheme for the most efficient products will encourage manufacturers to seek these awards and will highlight the most efficient products to consumers. A range of options are possible, including annual or ongoing awards for products that meet specified criteria and/or the development of an endorsement label that is used in conjunction with a comparison label. Lists of the most efficient products (e.g., those eligible for awards), either at the point of sale or on the Internet, will make these products more attractive and accessible.

9.4 Lists of Current Models on the Market

This is a complementary tool that is primarily for products that carry energy labels but that can also be useful for products covered by MEPS or efficiency targets. A complete lists of products can be made available in the form of brochures or, more commonly now, through product lists on the Internet. These are usually sorted so that the most efficient products are at the top to encourage consumers to consider these products first. Of course, interactive product lists on the Internet can be sorted by whatever criteria are desired by users.

10. ENFORCEMENT AND IMPLEMENTATION ISSUES

A range of issues and options need to be considered when implementing an energy labeling or MEPS program. Most of these relate to regulatory options.

10.1 Registration of Products

Some programs make registration of all regulated products mandatory, whereas others require manufacturers to hold details only of the products that are submitted to market. Registration certainly makes compliance and enforcement more straightforward, but this involves some organization and additional work for all parties, especially where there are a large number of products on the market. It is very difficult

to keep track of products on the market without some sort of registration system.

10.2 Supply of Test Reports

Manufacturers may be required to supply test reports with their registrations to confirm energy labeling or MEPS claims. Alternatively, they may be required to retain such reports for a specified period.

10.3 Self-Testing

In some cases, manufacturers are allowed to test their own products in their own laboratories, whereas other programs require independent third-party testing.

10.4 Accreditation

Tests may or may not have to be undertaken in an accredited or approved laboratory.

10.5 Certification

Manufacturers may be allowed to certify that their own products conform to the requirements/claims made. Alternatively, they may be required to use an independent third-party certification system.

10.6 Check Tests

It is critical that claims of energy efficiency be verified through a program of targeted compliance checks. These might not be random; in fact, they are most cost-effective if all available market intelligence is used to identify products that are most likely to not comply (e.g., excessive or unlikely claims of efficiency, brands with poorer compliance records, third-party referrals from competitors, products with high model turnovers). This requires a system of credible response to third-party claims of noncompliance and expedient enforcement of cases where noncompliance is proved.

10.7 Correct Display of Labels

An important aspect of a mandatory labeling program is checking retailers to ensure that labels are correctly displayed on appliances at the point of sale. An expedient way of dealing with minor infringements (e.g., spot fines) may be the most effective way in which to ensure high levels of compliance.

11. TEST PROCEDURES

Test procedures (often also called "test standards") are the method of testing used to determine appliance performance, energy consumption, and hence energy efficiency. They are critical in that they allow the comparison of products on a fair basis.

The test procedures used to determine the energy consumption (and, where relevant, the performance) of an appliance or a product can also have a large influence on the measured energy consumption. Factors such as ambient temperatures (for refrigerators and air conditioners), minimum wash temperatures (for dishwashers and clothes washers), and initial moisture content (for clothes dryers) all are critical. Although some of these parameters are specified in international standards, these do not always suit regional or national requirements for energy and performance testing. Often, a national standard will contain test conditions that are specific to and reflective of local climate and/or consumer patterns of use.

For appliances and equipment, the energy test procedure is the foundation for MEPS, energy labels, and other related energy programs. It provides manufacturers, regulatory authorities, and consumers a way in which to consistently evaluate energy use and savings across various appliance models. A well-designed test procedure services the needs of its users economically and with an acceptable level of accuracy and correspondence to real conditions of use. On the other hand, a poorly designed energy test procedure can undermine the effectiveness of everything built upon it.

Therefore, energy test procedures are a critical element underpinning all energy programs that seek to measure and improve the energy efficiency of appliances and equipment.

Test procedures cover all aspects of testing of the product such as the following:

- Ambient temperature
- Water quality and temperature
- Test loads
- Instrumentation and equipment
- Special materials and methods
- Duty cycles and/or loading patterns

Ideally, a good test procedure should have the following characteristics:

- Repeatability (producing the same result each time the same model is tested in the same test laboratory)

- Reproducibility (producing the same result each time the same model is tested in different test laboratories)
- Reflective of consumer use
- Simple but effective
- Covering existing products as well as new and forthcoming technologies
- Able to represent cultural and environmental influences and user patterns

Unfortunately, it is rarely possible for a test procedure to meet all of these requirements simultaneously. Clearly, the requirements for repeatability and reproducibility are paramount in any test procedure that is used to regulate products or judge manufacturer claims. These are often regarded as key and fundamental requirements for any test procedure and tend to take precedence over other aspects.

The ability to be reflective of real consumer use is also a key requirement. The problem is that consumers usually behave and act in different ways, so more often than not, there is a distribution of patterns of use rather than a single pattern. Variation in the frequency of use (e.g., of a clothes washer) is reasonably easy to determine, and it may even be possible to reflect this to some degree in an energy label or advisory brochure. However, consumer selection of various wash programs is more difficult to represent, as is variation in, say, washing temperatures. Being able to represent (or simulate) climate and temperature impacts is also an important consideration.

The issue of new and emerging technologies is a critical one. Increasingly, there appear product configurations and types that were never envisaged by the people who developed the original test procedure for a product. The presence of load-sensing devices (in products such as clothes washers), automatic dryness sensors (in clothes dryers), adaptive and smart controllers, and fuzzy logic (adjusting and optimizing operation and performance on the basis of the previous day/week/month/year of use) makes it increasingly difficult to test some product types, especially with outmoded procedures. Many of these devices and controls will actually save energy in actual use. However, some savings will be fake; that is, the savings will appear only in the test procedures. The key issue is how to test products without being tricked into thinking that they are more or less efficient than they are likely to be in actual use.

The alignment of international test procedures will

- facilitate international trade;
- decrease testing and approval costs for manufacturers;
- allow the free movement of the most efficient products (note that products with low energy efficiencies may still be barred if they do not meet local MEPS levels);
- facilitate international comparisons;
- assist in the diffusion of advanced energy-saving technologies.

12. TESTED ENERGY CONSUMPTION VERSUS IN-USE ENERGY

A key (but by no means obvious) difference between energy labeling programs is the assumptions that lie behind the calculation of the energy consumption and related performance data shown. Many energy labels have built into them assumptions about the frequency and duration of use for the calculation of energy consumption and related parameters. For example, in the case of clothes washers, the data shown on some energy labels is based on an assumed number of washing loads per year. Such estimates are usually based on averages from surveys and data collected from the country or region where the energy label is to be used. However, it is important to remember that no consumer or user is typical; there is always likely to be a wide distribution of patterns of use and consumer related behavior for all products.

Some analysts argue that the energy label will, or at least should, be of most economic value and of most interest to customers who use their appliances most intensively, so it may be appropriate for values on the label to be calculated for a higher than average frequency of use (or perhaps median use). In terms of values shown on an energy label, large "annual" figures for energy consumption and cost will have a greater influence on consumer decisions than will small numbers such as energy per usage cycle or per day. For example, 10-year running cost data (a proxy for expected minimum appliance life for many products) shown on energy labeling brochures would often be comparable to the appliance purchase cost. In this respect, there is a potential contradiction, albeit a minor one, between energy labels as a consumer information program and as a policy measure for increasing energy efficiency.

The economics of energy efficiency is a key area of consumer interest with respect to energy labeling. Most consumers express interest in the cost of energy

used to operate an appliance. However, conveying this information through an energy label has many problems, including variations in energy tariffs within a country and by time of day. In addition, consumers can easily confuse cost information shown on energy labels. In fact, it has been found to be unclear to some consumers whether the cost figures shown on a label relate to the cost of energy used to operate the appliance or to the savings in energy operating cost. Some consumers also mistakenly believe that energy cost values on an energy label are related to the appliance purchase price.

Climatic considerations are critical for some products (especially air conditioners and refrigerators and, to some extent, water heaters), and historically this has been poorly handled in international and regional test procedures for these product types. These products, which typically have widely varying temperature performance coefficients for different models, are usually tested under a single static temperature condition, which often neither is representative of nor facilitates the estimation of performance under other conditions (including real use). Of course, "real use" and a "single representative test point" can never be reconciled given that many countries have climate zones ranging from cool temperate to humid tropical; a single test condition can never be representative of such a range. For such products, a complex conversion algorithm (i.e., a computer simulation model) is probably the only feasible long-term option by which in-use energy consumption can be estimated more accurately from results recorded under standard test conditions.

13. PERFORMANCE VERSUS ENERGY CONSUMPTION

The measurement of appliance performance in conjunction with energy consumption is a challenging issue. Within a particular appliance type, the energy consumption and performance are related (e.g., making food colder in a particular refrigerator requires more energy, increasing the wash temperature of a clothes washer increases both the wash performance and the energy consumption). Measurement of the performance of products is widely accepted as part of most product test procedures. Things such as internal temperature control of a refrigerator, cooling or heating capacity of an air conditioner, shaft power output of an electric motor, and volume of hot water delivered by a water heater all are measures of performance. In fact, consumers and businesses are not generally as interested in the energy that is consumed by a product as they are in the energy service that the product performs for them. Nearly all test procedures measure the performance in some way.

Performance is critical as the energy service and energy consumption combine to provide a measure of the energy efficiency of the product, that is, the "energy service delivered per unit of energy consumption." In general, statements of energy efficiency are meaningless unless the level of performance is either specified or declared.

If there is no performance measurement in the test procedure (or no realistic task for the product to perform), it is often necessary to prescribe attributes such as capacity and configuration. Some advanced products are able to trick a nonperformance test procedure. For example, dishwashers or clothes washers that have soil sensors could terminate the washing operation prematurely if a clean load were used. Such machines would operate only with very short program cycles and would achieve inordinately high energy ratings under the test procedure, whereas in actual use the energy would be much higher. Hence, comparative advice based on such tests would mislead consumers.

14. EVALUATION

It is critical that any energy program be subjected to ongoing evaluation and assessments. This will ensure that the program is operating smoothly, that problems are identified and resolved, and that an accurate estimate of program impacts is obtained. There are two main types of evaluation of labeling and MEPS programs: process evaluation and impact evaluation.

14.1 Process Evaluation

Process evaluation is usually qualitative and is a measure of the operation of the program. Unfortunately, this is sometimes seen as relatively less important by policymakers. In reality, process elements are especially critical to the successful implementation of a labeling program (in contrast to a MEPS program) because they cover aspects such as the following:

- consumer priorities in purchasing an appliance;
- consumer awareness levels of labels;
- correct display of labels in retailers;

- correct indication of energy performance on labels;
- the operation of the regulatory system such as ease of product registration.

Elements such as check testing and evaluation of administrative efficiency (e.g., registration times and processes) may be relevant to both labeling and MEPS programs.

14.2 Impact Evaluation

Impact evaluation is an attempt to quantify the energy and related environmental (especially greenhouse gas) effects arising from the implementation of a labeling or MEPS program. Impact elements include tracking of product efficiency trends and quantification of energy, greenhouse gas, and demand savings resulting from the implementation of energy labeling or MEPS. Impacts can be very difficult to determine accurately in some cases, especially where it is problematic to benchmark the efficiency of the market prior to program implementation (e.g., due to lack of data). Another problem is that once an energy program has been in place for a period of time, it becomes increasingly difficult to determine a "base case" against which to compare the program impact. (There is only "one reality"; for example, after 20 years of energy labeling, a "no label" scenario can hardly be considered as a base case for comparison.) Even where the overall impact can be determined with reasonable accuracy, it can be difficult to accurately allocate the share of the efficiency improvement where a number of influences have been working in combination, for example, where labeling, MEPS, and the introduction of new technology (which may be unrelated to labeling or MEPS) all have had an influence on the energy efficiency of products over a defined period. Advanced techniques can be used to partly separate the impact of these various influences, but this is not always precise.

It is critical that planning and data collection for evaluation be incorporated into the program design and implementation. Collection of relevant data needed to establish a baseline after the program has been implemented is usually difficult or impossible.

15. IMPACTS OF EFFICIENCY STANDARDS AND LABELING PROGRAMS

Notwithstanding the difficulties of evaluating the impacts of appliance energy efficiency programs, various efforts have been made to quantify their impact to date as well as their potential impact if they were to be strengthened. A publication from the International Energy Agency (IEA) in 2003 pulled together much of the available information among OECD countries and produced a new synthesis of the probable impacts of the MEPS and labeling programs that are currently in place. At the time, residential appliances and equipment used 30% of all electricity generated in OECD countries and accounted for 12% of all energy-related carbon dioxide (CO_2) emissions. They were the second largest consumer of electricity and the third largest emitter of greenhouse gases in the OECD. Since 1973, primary energy demand in the OECD residential sector has grown faster than that in all other sectors (other than transport), and in terms of electricity demand growth, it has outstripped all but primary energy demand in the commercial buildings sector over this same period.

Figure 3 shows aggregate (projected) residential electricity demand in OECD countries from 1990 to 2030 for three scenarios:

- The "no policies" scenario that projects what would have happened if no MEPS, targets, or labeling programs had been implemented in OECD countries
- The "current policies" scenario that projects electricity demand with the existing set of policies

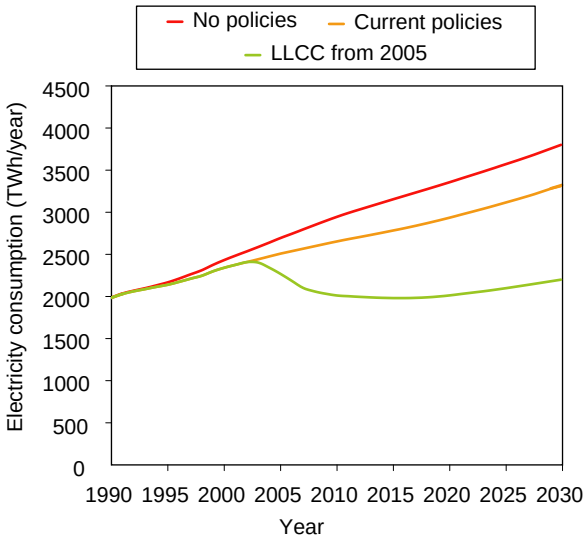


FIGURE 3 Aggregate residential electricity demand in the OECD for three appliance efficiency scenarios. LLCC, least life cycle cost.

- The "LLCC from 2005" scenario that projects electricity demand if the efficiency of residential electrical equipment sold from 2005 onward were to correspond to the least life cycle cost efficiency level from consumers' perspective, that is, if the efficiency of the products sold from 2005 onward were to be at a level by which the sum of the purchase prices and the discounted costs of operating the appliances over their lifetimes are minimized

The IEA results show that the appliance policies enacted since 1990 reduced OECD residential electricity consumption by 3.8% in 2000 and will go on to reduce it by 12.5% in 2020 compared with what would have happened if the policies had not been introduced. These savings resulted in a reduction of some 37 million tonnes (Mt) of CO_2 emissions in 2000 and avoided the need for at least 20 gas-fired power stations. Even without further strengthening, these same policies will go on to reduce emissions by 126 Mt CO_2/year by 2010 as more efficient equipment replaces less efficient equipment in the stock.

However, the IEA analysis also shows that a very significant cost-effective potential still exists to strengthen the impact of appliance efficiency programs. By using efficiency policy to target the most cost-effective level of efficiency (the least life cycle cost) for appliances from 2005 onward, IEA member countries could save more than 642 TWh electricity/year, or some 322 Mt CO_2/year, by 2010 compared with what they will save under existing policy settings. In terms of greenhouse gas emissions, this would be equivalent to taking more than 100 million cars off IEA roads or doing without nearly 200 gas-fired power stations. The IEA estimated that targeting the least life cycle cost for residential appliances could achieve up to 30% of OECD member countries' targets under the Kyoto Protocol on climate change. By 2030, a policy of targeting the least life cycle cost for residential appliances (from 2005) would avoid more than 1110 TWh/year of final electricity demand, or 572 Mt CO_2/year, equivalent to taking more than 200 million cars off OECD roads.

Because the efficiency level associated with the least life cycle cost necessarily corresponds to the lowest net equipment purchase and operating costs, the pursuit of these policies is highly cost-effective even if climate change targets are not a concern. The IEA estimated that each tonne of CO_2 saved through 2010 by current policies in OECD member countries in North America will be attained at a net cost saving of U.S. $78; that is, the net cost of CO_2 abatement is U.S. −$78/tonne CO_2. For OECD member countries in Europe, the equivalent value is Euro 241/tonne CO_2. Under the least life cycle cost from 2005 scenario, the net cost of CO_2 abatement is U.S. $65.5/tonne CO_2 in OECD member countries in North America and Euro 168.9/tonne CO_2 in OECD member countries in Europe by 2020. These are slightly less cost-effective per tonne saved than are the existing policies, but they result in much greater total cost savings.

SEE ALSO THE FOLLOWING ARTICLES

Consumption, Energy, and the Environment • *Economics of Energy Efficiency* • *Energy Efficiency, Taxonomic Overview* • *Industrial Energy Efficiency* • *International Energy Law and Policy* • *Obstacles to Energy Efficiency* • *Vehicles and Their Powerplants: Energy Use and Efficiency*

Further Reading

Bertoldi, P., Ricci, A., and de Almeida, A. (Eds.) (2001). "Energy Efficiency in Household Appliances and Lighting," proceedings of the Second International Conference on Energy Efficiency in Household Appliances and Lighting. Springer-Verlag, Berlin.

Bertoldi, P., Ricci, A., and Wajer, B. (Eds.) (1998). "Energy Efficiency in Household Appliances," Proceedings of the First International Conference on Energy Efficiency in Household Appliances and Lighting. Springer-Verlag, Berlin.

Egan, K., and du Pont, P. (1998). "Asia's New Standard for Success: Energy Efficiency Standards and Labeling Programs in 12 Asian Countries." International Institute for Energy Conservation, Washington, DC.

Energy Efficiency Strategies, *et al.* (1999). "Review of Energy Efficiency Test Standards and Regulations in APEC Member Economies," Report APEC 99-RE-01.5, prepared by Energy Efficiency Strategies and others for Asia-Pacific Economic Cooperation (APEC) secretariat, Singapore.

Harrington, L., and Damnics, M. (2001). "Energy Labelling and Standards Programs throughout the World," report commissioned and published by the National Appliance and Equipment Energy Efficiency Committee, Australia. www.energyrating.gov.au.

International Energy Agency. (2000). "Energy Labels and Standards." IEA, Paris. www.iea.org.

International Energy Agency. (2003). "Cool Appliances: Policy Strategies for Energy Efficient Homes, April 2003." IEA, Paris. www.iea.org.

Wiel, S., and McMahon, J. (eds.). (2000). "Energy Efficiency Labels and Standards: A Guidebook for Appliances, Equipment, and Lighting." Collaborative Labelling and Appliance Standards Program (CLASP), Washington, DC. www.clasponline.org.

Land Requirements of Energy Systems

VACLAV SMIL
University of Manitoba
Winnipeg, Manitoba, Canada

1. **Biomass Energies**
2. **Direct Solar Radiation**
3. **Wind and Water Energy**
4. **Fossil Fuels and Thermal Electricity Generation**
5. **Aggregate Demand and Energy Consumption**
6. **Energy Transitions**

Glossary

insolation The average share of solar (electromagnetic) energy absorbed by the earth's surfaces that amounts to only about 12% of the solar constant.

phytomass Living or dead organic matter (biomass) of plants; in energy literature, it refers particularly to that of trees and crop residues used for fuel.

power density Average long-term power flux per unit of land area, usually expressed in watts per squared meter (W/m^2).

solar constant Total flux of solar (electromagnetic) energy that falls on 1 square meter of the extraterrestrial area above the earth's atmosphere at a vertical angle; its value is approximately 1.37 kW/m^2.

thermal electricity generation Generation where either the combustion of one of the fossil fuels or nuclear fission is used to produce steam to drive turbogenerators.

Land requirements of energy production depend primarily on the nature of harnessed resources. The two sets of activities that define the modern high-energy civilization—extraction of fossil fuels and thermal generation of electricity—require a relatively limited amount of land. In contrast, all forms of renewable energy conversions are highly land intensive, but the differences among individual techniques range over three orders of magnitude. Production of phytomass fuels, the principal source of heat in preindustrial economies, makes the largest relative claims on land. In contrast, direct conversions of solar radiation, whose contribution to the world's supply of commercial energy remains minuscule, need one or even two orders of magnitude less land than does either the natural growth of forests or the cultivation of phytomass for biofuels. Hydroelectric generation can be as land demanding as phytomass production, and although it provides less than 3% of the world's commercial primary energy supply, the aggregate area of reservoirs makes up approximately 60% of all land occupied by the world's energy infrastructures. Socioeconomic and environmental implications of this pronounced spatial dichotomy will become increasingly important during the 21st century as the world's energy supply, now dominated by fossil fuels and thermal electricity, shifts gradually to a system that will be much more dependent on conversions of renewable energies. By far, the most revealing way of comparing land requirements of various energy systems is to calculate their power densities, that is, their energy flux per unit area of the earth's surface. The standard unit of power density is expressed in watts per square meter (W/m^2), and this fundamental measure is applicable to all energy flows, be they parts of production or consumption systems, natural or anthropogenic processes, or renewable or fossil fuel conversions. There are immutable natural limits to all commercial power densities and movable technical barriers that can raise the rates through higher efficiencies. Natural limits for fossil fuel-based techniques are given by the accumulations of coals and hydrocarbons, and unless we resort to space-based capture of solar energy, all renewable energy conversions will be limited to a fraction of solar radiation that reaches the earth's surface.

1. BIOMASS ENERGIES

Because of the inherently low efficiency of the photosynthetic process, no form of energy supply

 © 2004 Elsevier Inc. All rights reserved.

has such low power densities, and hence such high land demands, as does the production of phytomass. Recent estimates of the global terrestrial net primary productivity (NPP) average approximately 55 billion metric tons (tonnes) (Gt) of carbon per year, that is, approximately 120 Gt of dry biomass that contains (assuming the mean energy density of 15 billion joules per metric ton [GJ/tonne]) some 1800×10^{18} joules (EJ). This productivity prorates to less than 0.5 watts per square meter (W/m^2) of ice-free land. Rates for forests are naturally higher, but they do not surpass 1.1 W/m^2 even in the richest tropical ecosystems. In all natural ecosystems, a large part of the NPP is consumed by heterotrophs (ranging in size from bacteria to megaherbivores); hence, the phytomass that is actually available for energy conversion is only a fraction of the originally produced biomass.

These inherently low power densities mattered little so long as the woody phytomass was required in relatively small quantities, but they led to considerable logistic problems for traditional societies whose growing cities and expanding industries were entirely dependent on phytomass for their thermal energy needs. Clearing of a primeval forest in preindustrial Europe could produce several hundred tonnes of wood per hectare, but this was obviously a historical singularity. Once that rich growth was gone, the wood was commonly cut in 10- to 20-year rotations from coppicing hardwoods whose annual increment was between 5 and 10 metric tons per hectare (tonnes/ha) or 0.25 to 0.5 W/m^2. Converting wood to more energy-dense and less-polluting charcoal greatly increased the primary demand because primitive methods of charcoaling needed as much as 10 kg (and commonly at least 5 kg) of wood per kilogram of charcoal and resulted in production power densities of no more than 0.1 W/m^2. Spatial implications of these low power densities became burdensome for all large preindustrial cities and even more so for the highly energy-intensive iron smelting.

For example, in 1810, when the United States produced less than 50,000 tonnes of pig iron, approximately 2600 km^2 of forest, an area of roughly 50×50 km, had to be cut down every year to produce the requisite charcoal. A century later, the annual output of 25 million tonnes per year would have required, even with much lower charging rates, approximately 170,000 km^2 of forest every year, a square with the side equivalent to the distance between Philadelphia and Boston. This single example makes it obvious that only the substitution of charcoal by coal-derived coke made it possible to expand the modern pig iron production to more than a half-billion tonnes per year.

Although modern intensively cultivated field crops and fast-growing trees are grown with power densities that match or surpass the photosynthetic rates in the world's most productive tropical ecosystems, land requirements of biofuel production remain well below 1 W/m^2 (Fig. 1). Only with fast-growing trees (most commonly poplars, willows, pines, and eucalypti) that are intensively cultivated as short-rotation crops and produce exceptionally high harvests of 20 tonnes/ha of dry matter, will the above-ground phytomass yield more than 1 W/m^2. Even the best possible conversions—with efficiencies greater than 50% for cogeneration of electricity and heat and 55 to 60% for conversion to methanol or ethanol—result in power densities of just 0.5 to 0.6 W/m^2 for the final energy supply. These rates are one to two orders of magnitude below the power densities of electricity generation using water, wind, or solar radiation. These low power densities will limit the share of the total primary energy that could be derived from biofuels.

In fact, any large-scale schemes of biofuel production would be questionable even if they were to have much better net energy returns than has been the case so far. This is because we are already appropriating approximately 40% of all terrestrial NPP through our field and forest harvests, animal grazing, land clearing, and forest and grassland fires. Any future large-scale reliance on cultivated biofuels would have to increase this appropriation by a large margin, a shift that not only would preclude food production and eliminate wood harvests for timber and paper from areas devoted to biofuels but also would destroy many irreplaceable environmental services that are now provided by mature natural ecosystems or by only partially disturbed forests.

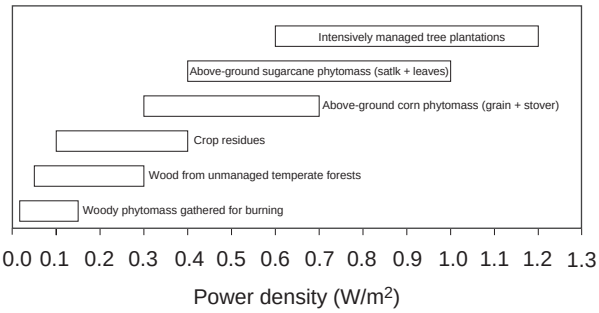


FIGURE 1 Power densities of various forms of biomass production.

2. DIRECT SOLAR RADIATION

Global mean of the total insolation—the solar radiation absorbed by the earth's surfaces—averages 168 W/m^2, only approximately 12% of the extraterrestrial solar constant. Annual maxima surpass 250 W/m^2 in the cloudless, unpolluted, high-pressure belts of subtropical deserts, where the daily peaks are approximately 1.2 kW/m^2. Except for passively heated structures (every house, no matter how poorly built and unsuitably oriented, is a solar house), preindustrial societies had no technical means to convert solar radiation into useful energy. Moreover, traditional architectures were much better for devising solutions for passive cooling in warm climates than for capturing solar radiation in cold locations. Modern conversions of solar radiation have followed two distinct routes: central solar power techniques, which concentrate the insolation by means of troughs, mirrors, or dishes, and photovoltaics (PV), which converts radiation directly into electricity.

Peak solar-to-electric conversion efficiencies are just over 20% for troughs, 23% for power towers, and 29% for dishes, but annual rates are much lower, that is, 10 to 18% for troughs, 8 to 19% for towers, and 16 to 28% for dishes. Thus, troughs and towers located at the sunniest sites and operating with the best efficiencies could have peak power densities of approximately 60 W/m^2 of collecting surface, and in less sunny locations the densities would go down to less than 40 W/m^2. Spaces between collectors, service roads, and buildings as well as exclusion zones will reduce those rates by at least 30%.

Maximum theoretical efficiency of PV cells is limited by the range of photon energies, whereas the peak practical rates are reduced above all through reflection from cell surfaces and the leakage of currents. Large differences remain when comparing theoretical, laboratory, and field performances as well as when contrasting the best performances of high-purity single crystals, polycrystalline cells, and thin films. Theoretical single-crystal efficiencies are 25 to 30%, and lenses and reflectors can be used to focus direct sunlight onto a small area of cells and boost conversion efficiencies to more than 30%.

Stacking cells sensitive to different parts of the spectrum could push the theoretical peak to 50% and to 30 to 35% in laboratories. Actual efficiencies of commercial single-crystal modules are now between 12 and 14%, and after a time in the field they may drop to below 10%. Thin-film cells can convert 11 to 17% in laboratories, but modules in the field convert as little as 3 to 7% after several months in operation. Multijunction amorphous Si cells do better, converting at least 8% and as much as 11% in large-area modules. The latter performance translates to no more than 15 to 18 W/m^2 for most locations, whereas with the eventual 17% field efficiency PV cells would be averaging approximately 30 W/m^2.

3. WIND AND WATER ENERGY

Wind energy is distributed very unevenly in both space and time. Many sites around the world have mean annual wind speeds of 7.0 to 7.5 m/s, enough to produce power densities of 400 to 500 W/m^2 of vertical area swept by rotating blades 50 m above ground. Translating these rates into horizontal power densities requires a number of sequential assumptions. Sufficient spacing between the units is needed to eliminate upstream wakes as well as to replenish kinetic energy. Spacing equal to 5 rotor diameters is enough to avoid excessive wake interference, but at least twice that distance is needed for wind energy replenishment in large installations. In addition, no more than 16/27 (59.3%) of wind's kinetic energy can be extracted by a rotating horizontal axis generator (the rate known as the Betz limit), and the actual capture will be still lower.

Consequently, machines with 50-m hub height and with 10×5 diameters spacing would have to be 500 m apart, and in locations with wind power density averaging 450 W/m^2 (the mean value common in the Dakotas, northern Texas, western Oklahoma, and coastal Oregon), they would intercept approximately 7 W/m^2. However, the average 25% conversion efficiency and 25% power loss caused by wakes and blade soiling would reduce the actual power output to approximately 1.3 W/m^2. Power density of 700 W/m^2, turbine efficiency of 35%, and power loss of 10% would more than double that rate to approximately 3.5 W/m^2. Actual rates are highly site specific. Rated performance of California's early wind projects was almost exactly 2 W/m^2, and the most densely packed wind farms rate up to 15 W/m^2. In contrast, more spread-out installations have power densities mostly between 5 and 7 W/m^2.

Unlike in the case of phytomass, the low power density of wind is not in itself an insurmountable obstacle to harnessing this large resource because all but a small share of areas occupied by wind facilities (commonly less than 5% of the total) can still be used for farming or ranching. Moreover, in large countries,

windy land is abundant. A recent U.S. study that used a variety of land exclusion scenarios (due to many environmental and land use reasons) concluded that the country's windy land adds up to approximately 460,000 km^2, or about 6% of all land in the contiguous United States, and that it could support generating capacity of approximately 500 GW.

Hydroelectricity generation operates with generally low power densities, and reservoirs have inundated large areas of natural ecosystems, including many unique and highly biodiverse forests and wetlands. Reservoirs behind the world's large dams (those higher than 30 m) now cover nearly 600,000 km^2 (an area nearly twice as large as Italy), and those used solely or largely for electricity generation add up to approximately 175,000 km^2. This prorates to approximately 4 W/m^2 in terms of installed capacity and to 1.7 W/m^2 in terms of actual generation, but power densities of individual projects span three orders of magnitude. Dams on the lower courses of large rivers impound huge volumes of water in relatively shallow reservoirs. The combined surface of the world's seven largest reservoirs is as large as the total area of the Netherlands, and the top two impoundments—Ghana's Akosombo on the Volta River (8730 km^2) and Russia's Kuybyshev on the Volga River (6500 km^2)—approach the size of small countries such as Lebanon and Cyprus. Power densities of these projects are less than 1 W/m^2 (Fig. 2).

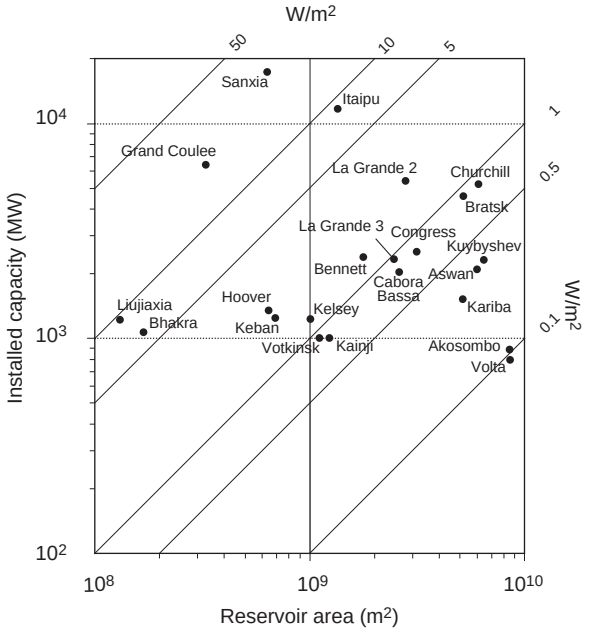


FIGURE 2 Power densities of large reservoirs used for hydroelectricity generation.

In contrast, dams on the middle and upper courses of rivers with steeper gradients have much smaller and deeper reservoirs, and their power densities are on the order of 10^1 W/m^2. Itaipu, the world's largest operating hydro project so far (12.6 GW), has a power density of 9.3 W/m^2. Grand Coulee (6.48 GW) rates nearly 20 W/m^2. China's Three Gorges (Sanxia, which is to become the world's largest hydro project, with 17.68 GW of installed capacity, when finished in 2008) will have approximately 28 W/m^2. In addition, some Alpine stations surpass 100 W/m^2 (Fig. 2). The world's record holder will be the planned Nepali Arun project, whose 42-ha reservoir and 210-MW installed capacity translate to approximately 500 W/m^2. Because load factors of most hydro stations are typically less than 50%, and because many projects are built to supply electricity during the hours of peak demand and hence have even shorter operating hours, effective power densities are only 25 to 50% of the just-cited theoretical rates.

4. FOSSIL FUELS AND THERMAL ELECTRICITY GENERATION

Energy density of recoverable coal deposits varies greatly with the thickness of seams and heat content of the fuel. Typical energy densities of large coal basins range from no more than 10 GJ/m^2 for poor lignites in thin seams to about 50 GJ/m^2 for good bituminous coals in multiple seams. The best sites have densities one or nearly two orders of magnitude higher: Arizona's Black Mesa ($\sim$200 GJ/m^2); Montana's Ashland (415 GJ/m^2); Fortuna, West Germany's largest brown coal mine in Rheinland (430 GJ/m^2); parts of Victoria's Latrobe Valley, with brown coal seams up to 100 m ($>$800 GJ/m^2); and Queensland's Blair Athol, with 30 m of bituminous coal ($\sim$900 GJ/m^2).

Power densities of coal extraction are the lowest (less than 500 W/m^2) in poorly designed surface mines producing low-quality lignites or sub-bituminous coals (with the best example being that of large brown coal mines in the former East Germany), whereas open cast mines working thick, high-quality seams of bituminous coal can extract several kilowatts per square meter (Australia has a number of such operations). Most coal mines, be they surface (with proper land reclamation) or underground (including all above-ground facilities), produce the fuel with power densities between 1 and 10 kW/m^2 (Fig. 3).

Prorating the most likely global estimate of recoverable hydrocarbons, an equivalent of nearly 750 Gt of crude oil according to the latest U.S.

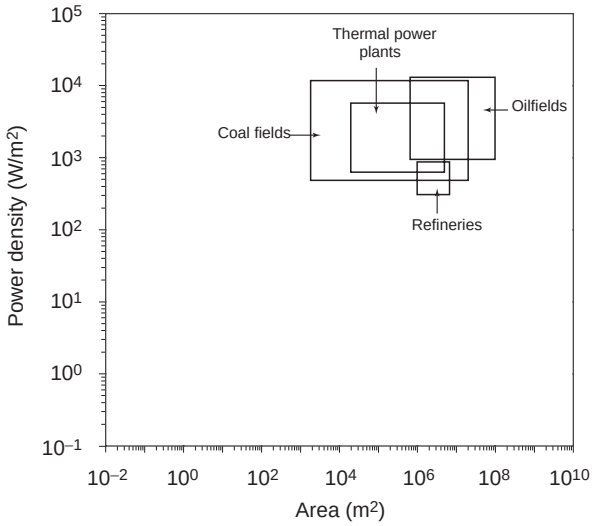


FIGURE 3 Typical power densities of fossil fuel extraction and processing.

Geological Survey study, more than roughly 70 million km^2 of known and prospective sedimentary basins gives a median energy density of approximately 450 million joules (MJ)/m^2. But the richest basins contain 1 to 10 GJ/m^2, and many giant hydrocarbon fields go well above that. Alaskan Prudhoe Bay rates about 25 GJ/m^2; Saudi al-Ghawar, the world's largest oilfield with roughly one-seventh of the world's oil reserves, contains close to 100 GJ/m^2; and Kuwaiti al-Burqan, with the world's second largest oil reserves, contains approximately 1000 GJ/m^2. Recovery of hydrocarbons has always required much less land per unit of extracted energy than does coal mining, and recent advances in directional and horizontal drilling mean that even extreme reaches of an underground oil or gas reservoir can be accessed from a single borehole, further minimizing any surface disruption.

Because usually less than 1% of the area overlying an oilfield is actually taken up by surface structures, even unremarkable hydrocarbon fields, such as the old U.S. ones dominated by stripper wells, can produce oil with power densities surpassing 2 kW/m^2 of land affected by the extraction. This includes not just exploratory and production wells but also mud and water ponds, storage yards, buildings and sheds, gathering pipelines, and access roads (Fig. 3). Rates for the world's richest Middle Eastern oilfields are commonly between 10 and 20 kW/m^2. Typical power densities of natural gas extraction are as high as 10 to 15 kW/m^2, but this rate is reduced by up to an order of magnitude once the rights-of-way for gathering pipelines and field facilities for gas processing are added to well enclosures.

The following generalizations summarize the land requirements of fossil fuel production and processing. Coals and hydrocarbons are now produced with power densities unmatched by any other fuel or source of primary electricity; 400 to 4000 W/m^2 might be a good conservative range, including the majority of coal mining and oil and gas extraction rates. In comparison with their extraction, processing of fossil fuels requires relatively little space. Coal preparation and cleaning have throughput densities between 8 and 10 kW/m^2. Crude oil refining requires more space due to safety precautions in siting high-pressure operations and storage facilities and in preventing spills, and its throughput power densities are generally lower (3–4 kW/m^2).

Pipelines usually need 25 to 30 m for construction, after which only access strips of up to 10 m wide may be necessary. Compressor stations take up to 20,000 m^2 at 80- to 120-km intervals, and pumping stations take up to 10 times as much every 130 to 160 km. Tanker terminals and liquefied natural gas facilities need relatively small areas but require extensive security zones. Railroads that were built exclusively to move coal by using unit trains over distances of 500 to 1500 km to large power plants have annual throughput densities between 100 and 400 W/m^2.

Only a small part of land required by modern thermal electricity generating plants is devoted to structures housing boilers and turbogenerators. Combustion of fossil fuels in large boilers proceeds at power densities often surpassing 5 million watts (MW)/m^2, and modern turbogenerators have outputs higher than 1 MW/m^2. In contrast, auxiliary structures (including switchyards, storage sheds, and office buildings), water-cooling facilities, and air pollution controls take up much more space, and even more land is needed when substantial on-site fuel storage is necessary (Fig. 4). Once-through cooling (returning warmed water into streams or bays) has minimal spatial claims, but spray ponds require approximately 400 m^2/MW and ordinary cooling ponds require between 4500 and 5000 m^2/MW. Cooling towers cut this demand to between 30 and 60 m^2/MW for both wet and dry natural draft structures, whereas mechanical draft units need as little as 10 m^2/MW.

Electrostatic precipitators removing more than 99% of all fly ash are now the norm on coal-fired stations, but because they are required to handle huge volumes of hot gas, they take up about as much space as do boilers. Desulfurization units can be more compact, but the disposal of captured fly ash

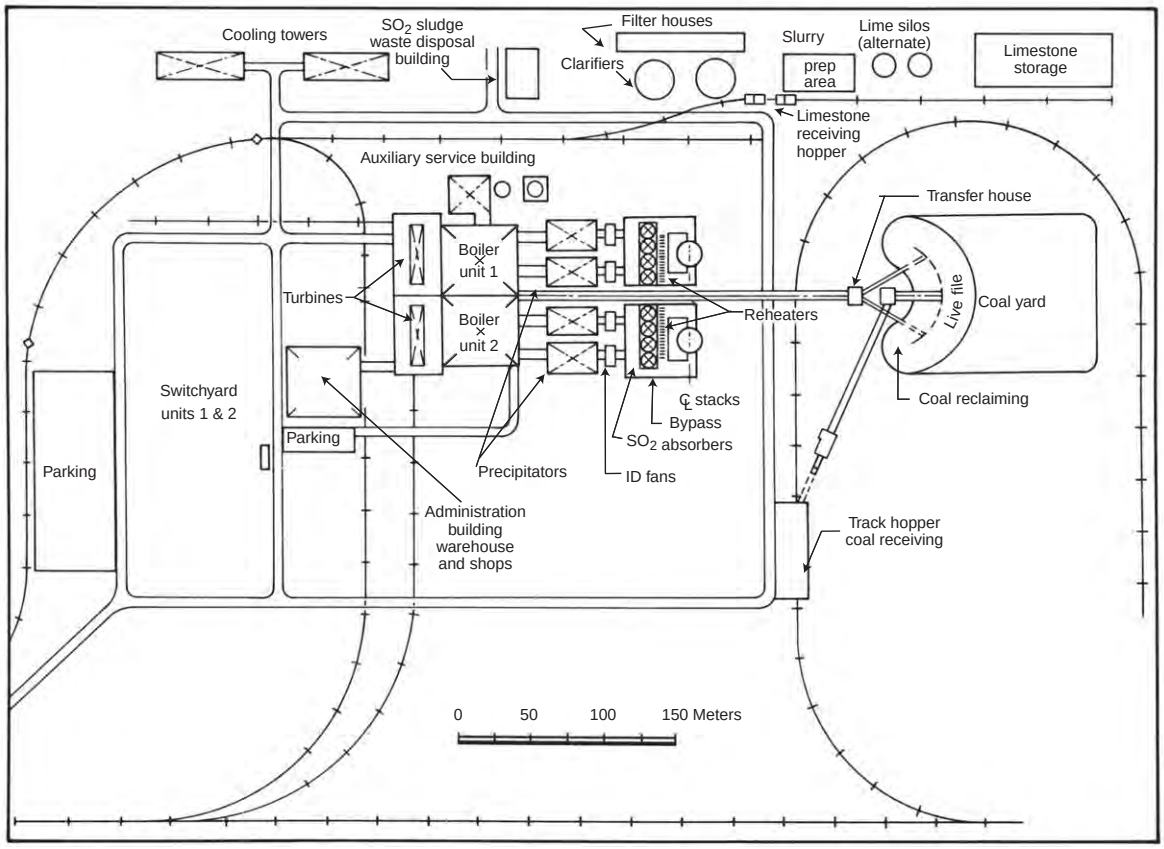


FIGURE 4 Land requirements of a coal-fired electricity-generating plant with two units rated at 600 MW_e each.

and ponding of the sulfate sludge require relatively large areas of adjacent land or costly arrangements for trucking or railway transport to dump sites. Long-term storage of fly ash captured by electrostatic precipitators needs 20 to 40 m^2/MW, and during a plant's lifetime of 35 to 50 years between 700 and 2000 m^2 of ash disposal space will be needed per installed million watts (assuming that coal contains about 10% of ash). Flue gas desulfurization facilities claim between 400 and 600 m^2/MW, and the lifetime disposal of sludge requires between 200 and 600 m^2/MW of installed capacity.

Hydrocarbon-fueled plants supplied by pipelines will have only small storage requirements, but coal piles needed for 60 to 90 days of operation (in case the railway deliveries were interrupted by a strike) are a different matter; they have storage densities of 25 to 100 thousand watts (kW)/m^2, similar to those of switchyards. Buffer zones between plants and the nearest settlements and land kept in reserve for future expansion may easily double or even triple the total land claim.

Given the variety of fuels, water-cooling and air pollution control arrangements, and fuel storage needs, it would be misleading to cite a single typical figure, but most coal-fired power plants will have overall power densities (including all land within a protective fence) of less than 1 kW/m^2, and the power density of the complete fuel cycle (extraction, transportation, and generation) will be less than 300 W/m^2. Analogical rates, for the plant alone and for the total fuel cycle, are approximately 2 kW/m^2 and less than 500 W/m^2 for oil-fired stations and are 3 and 1 kW/m^2 for the natural gas-fired ones. Land claims of nuclear plants that require no extensive fuel storage and no air pollution control facilities are dominated by exclusion and buffer zones, and with rates commonly approximating 1.5 kW/m^2, they are less than those of oil-fired stations but more than those of coal-fired ones.

5. AGGREGATE DEMAND AND ENERGY CONSUMPTION

Any global figure of the total area of land claimed and affected by infrastructures devoted to the

extraction and processing of fossil fuels and to generation of electricity must be only a useful approximation rather than an accurate summation. The author calculated that in the year 2000 the aggregate land claim of the global energy infrastructure reached approximately 290,000 km^2, equal to an area slightly smaller than Italy. Although the absolute figure is certainly large, it is perhaps more revealing to see the total in a relevant global perspective and to compare it with other anthropogenic conversions of land. The area of 290,000 km^2 amounts to less than 2% of the total area of natural ecosystems (mostly forests and grasslands) that has been converted to crop, pasture, industrial, transportation, and urban land during the past 250 or so years.

Less than 2% of the nearly 300,000 km^2 was claimed by the production and processing of coals and hydrocarbons, implying an average power density of approximately 1.75 kW/m^2 for the world's fossil fuel extraction (Fig. 5). Rights-of-way of hydrocarbon pipelines claimed roughly one-tenth of the total, and those of high-voltage transmission lines slightly claimed more than a quarter of the total. Naturally, claims by pipelines and transmission lines are a form of land use that is not directly comparable to the first category (extraction and generation) given that rights-of-way, particularly underneath high-voltage lines, can be used for crop farming, animal grazing, or shrub or small tree plantations. By far the largest share, approximately three-fifths of the total, was occupied by reservoirs storing water for the generation of hydroelectricity.

A more accurate estimate for the United States shows only approximately 1500 km^2 occupied by fossil fuel extraction and processing and by thermal

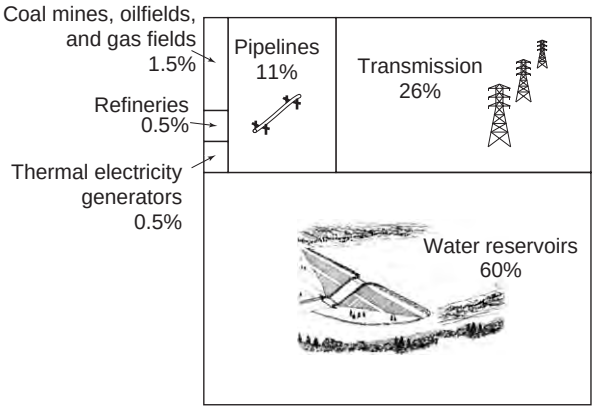


FIGURE 5 Approximate distribution of land claimed by the world's energy infrastructures during the late 1990s.

electricity generation but nearly 20,000 km^2 under water reservoirs. A similar amount of land has been devoted to rights-of-way by hydrocarbon pipelines and high-voltage transmission lines. In addition, there is more than 10,000 km^2 of land that was disturbed long ago by coal mining and that has yet to be properly reclaimed. All of these claims added up to approximately 50,000 km^2 during the mid-1990s, implying average power densities of approximately 1.9 kW/m^2 for the fossil fuel extraction and 45 W/m^2 for the entire U.S. energy infrastructure. These averages are just slightly higher than the global means of 1.75 kW/m^2 and 40 W/m^2, respectively.

But the actual effects of energy infrastructures on land use and land cover cannot be measured simply by adding up the areas whose use has been affected. Many energy facilities require land near large cities and, hence, often take over highly productive alluvial soils, whereas others are located in extraordinarily biodiverse and relatively fragile environments. Good examples include oil and gas production in the deltas of the Niger and Mississippi rivers and on Alaska's North Slope. Moreover, transportation and transmission corridors needed to develop energy resources may contribute to the impoverishment of surviving ecosystems because they increase their fragmentation, a process that is well known to reduce biodiversity of many species.

Although power densities of energy production range over five orders of magnitude, final energy uses in modern high-energy societies are confined mostly to between 10^1 and 10^2 W/m^2. Depending on the climate, construction, and the kind of converters inside the building, typical rates for single-family houses in temperate zones range from 20 to 100 W/m^2 of a foundation. Similar rates apply to low-rise office buildings and to structures that house industrial production of low-energy intensity, particularly assembly line manufacturing. Supermarkets and medium-height multiple-story office buildings use 200 to 400 W/m^2.

However, energy-efficient design and state-of-the-art converters can reduce these prevailing rates by large margins. For example, the average energy requirement of U.S. office buildings was approximately 33 W/m^2 of floor space during the late 1990s, whereas the Centex building in Dallas (the most efficient structure in its category as of 2002) used only 10 W/m^2 or only 90 W/m^2 of its foundations. Energy-intensive industrial enterprises (steel mills, smelters of nonferrous metals, refineries, and many chemical syntheses are the most prominent examples

in this category) require 300 to 900 W/m^2, and high-rise buildings require up to 3 kW/m^2.

Comparison of production and consumption power densities shows clearly that our overwhelmingly fossil-fueled civilization extracts fuels and generates thermal electricity with power densities that are commonly one to two (and sometimes even three) orders of magnitude higher than the common power densities of final energy use in buildings, factories, and cities (Fig. 6). These concentrated energy flows are then diffused through pipelines, railways, and high-voltage transmission lines, and that is why space taken up by extraction and conversion of fuels is relatively small in comparison with rights-of-way required to deliver fuels and electricity to consumers. For example, in the United States, the total area occupied by extraction of fossil fuels is on the order of a mere 0.01% of the country's area.

6. ENERGY TRANSITIONS

If a future overwhelmingly solar society were to inherit the existing urban and industrial infrastructure, it would have to rely on an energy-harnessing strategy that is an exact opposite of today's arrangements, that is, on concentration of diffuse energy flows that it would convert to useful forms of energy at rates ranging mostly between 0.3 and 30 W/m^2 (Fig. 7). Consequently, it would need several hundred, or even several thousand, times more space than does today's fossil fuel extraction to produce the same amount of energy. But it would also need extensive transmission networks to deliver electricity from regions endowed with high intensities of solar radiation or with strong winds.

A gradual increase of average conversion efficiencies (higher power densities of renewable generation) and reduced energy demand (lower power density), most notably for heating and lighting energy-efficient houses, should result in much more closer matches of power densities between production and use and, hence, lead to a much wider adoption of the distributed generation of electricity by individual households or small settlements. But to energize its cities and industrial areas, any society would have to concentrate diffuse flows to bridge power density gaps of several orders of magnitude.

By the year 2030, approximately 60% of the world's population will live in cities, and nearly a quarter of that total will live in "megacities" (with more than 10 million people) that will be dominated by clustered high-rises; parts of today's Hong Kong are a good preview of that future. Supplying these megacities by locally generated renewable energies will be impossible because the power density mismatch between the production and demand is simply too large. The same is true about manufacturing and even more so about often highly energy-intensive processing, especially smelting and chemical syntheses that are being increasingly concentrated in industrial parks.

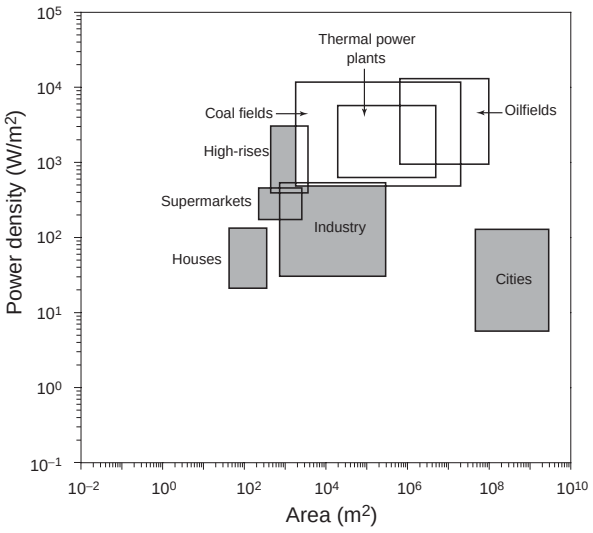


FIGURE 6 Comparison of typical power densities of fossil fuel production and modern urban consumption. The illustration shows how our current arrangements are based on the diffusion of concentrated primary energy supplies.

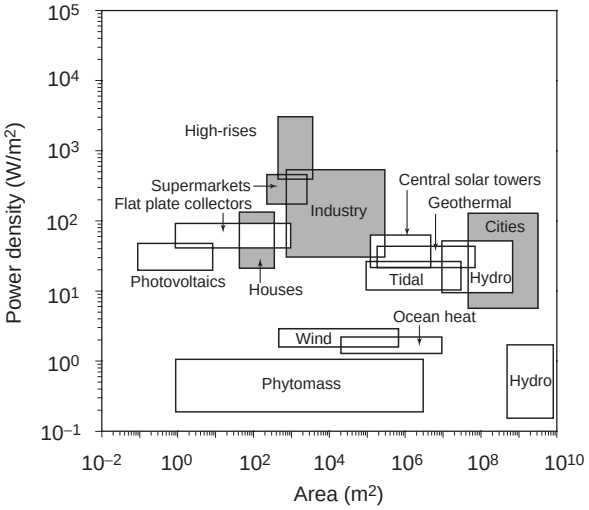


FIGURE 7 Comparison of typical power densities of renewable energy conversions production and modern urban consumption. The illustration shows the opposite process, that is, the need to concentrate diffuse sources of direct and indirect solar and geothermal flows.

In addition, a high-energy society relying on renewable flows would lose most of the flexibility now available in locating electricity-generating stations that can be sited either near large load centers (to minimize transmission distances) or far away from densely populated areas (to minimize air pollution levels). In contrast to these flexible arrangements, high-capacity concentrations of wind turbines and PV cells (or geothermal stations or land-based sea wave-converting machines) must coincide with the highest power densities of the requisite renewable flows. Unfortunately, nearly every contemplated instance of such a development would call for new high-capacity long-distance transmission links as well as for extensive upgrading of the existing lines.

For example, the United States now has only a rudimentary capacity to move electricity from the Southwest, the region of the most advantageous central solar and PV locations, to the Northeast. Similarly, there are no high-capacity transmission links from the windy Northern Plains to California. And the country is poorly prepared for even shorter electricity transfers. Although the U.S. transmission capacity nearly doubled between 1975 and 2000, when it is divided by peak load, the resulting ratio, which represents the distance to which a given plant can expect to sell energy during the conditions of peak demand, has been falling since 1984, and the trade area of a given generator is now only approximately 70% of the mean during the mid-1980s.

Complete absence or inadequate capacities of transmission links between regions endowed with high power densities of direct and indirect solar flows and densely populated load centers can be found in every large country that is contemplating higher reliance of renewable conversions. For example, China's windiest and sunniest areas are the northern grasslands in Inner Mongolia and the arid unpopulated western interior of Xinjiang and Tibet, respectively, whereas its largest load centers are along the eastern and southern coast. Linking the two would require spanning a distance of more than 3000 km and crossing some of Asia's most forbidding terrain.

Stochasticity of all solar energy flows would further increase the spatial impact of renewable conversion due to the need for either extensive energy storage or backup peak demand systems. Unfortunately, building pumped storages and using off-peak electricity generated by nuclear or fossil-fueled power plants to lift water into nearby reservoirs remains the only practical, but obviously a highly land-intensive, way of storing large quantities of almost instantly deployable energy. In spite of more than a century of diligent efforts to develop other effective forms of storage, all other options remain either inefficient or inadequate.

Even greater challenges would arise with any large-scale use of biofuels. For example, replacing the world's coal consumption (nearly 90 EJ in 2000) by woody biomass would require the cultivation of high-yielding trees on approximately 330 million ha, an area larger than all of the forests in the European Union and United States combined. And running the entire fleet of U.S. vehicles registered in the year 2000 on corn-derived ethanol would require planting the crop (with grain yields averaging 7.5 tonnes/ha and ethanol yields of 0.3 kg of alcohol/kg of grain) on an area 20% larger than the country's entire cultivated cropland. Such schemes are obviously impossible, but even their partial realization would have enormous environmental impacts.

These realities do not make large-scale reliance on renewables impossible, but they make it considerably more expensive and challenging to accomplish. Mismatch between low power densities of renewable energy flows and relatively high power densities of many modern final energy uses means that any large-scale adoption of renewable energy conversions by societies dominated by megacities and concentrated industrial production will require a profound spatial restructuring of the existing energy infrastructure, a process that will have many major environmental and socioeconomic consequences.

SEE ALSO THE FOLLOWING ARTICLES

Biomass Resource Assessment • City Planning and Energy Use • Ecological Footprints and Energy • Electric Power Generation: Fossil Fuel • Global Energy Use: Status and Trends • Global Material Cycles and Energy • Heat Islands and Energy • Hydropower Resources • Solar Thermal Power Generation • Transitions in Energy Use • Wind Resource Base

Further Reading

Dracker, R., and De Laquill, P. (1996). Progress commercializing solar-electric power systems. *Annu. Rev. Energy Envir.* **21,** 371–402.

Goetzberger, A., Knobloch, J., and Voss, B. (1998). "Crystalline Silicon Solar Cells." John Wiley, Chichester, UK.

Markvart, T. (ed.). (2000). "Solar Electricity." John Wiley, New York.
McGowan, J. G., and Connors, S. R. (2000). Windpower: A turn of the century review. *Annu. Rev. Energy Envir.* **25,** 147–197.
Perlin, J. (1999). "From Space to Earth: The Story of Solar Electricity." Aatec, Ann Arbor, MI.
Rojstaczer, S., Sterling, S. M., and Moore, N. J. (2001). Human appropriation of photosynthesis products. *Science* **294,** 2549–2551.
Smil, V. (1991). "General Energetics: Energy in the Biosphere and Civilization." John Wiley, New York.
Smil, V. (1994). "Energy in World History." Westview, Boulder, CO.
Smil, V. (2002). "The Earth's Biosphere: Evolution, Dynamics, and Change." MIT Press, Cambridge, MA.
Smil, V. (2003). "Energy at the Crossroads: Global Perspectives and Uncertainties." MIT Press, Cambridge, MA.

Leisure, Energy Costs of

SUSANNE BECKEN
Lincoln University
Lincoln, Canterbury, New Zealand

1. Background
2. Concept of Leisure, Recreation, and Tourism
3. Upstream/Downstream
4. Recreational Activities
5. Tourist Accommodation
6. Tourist Transport
7. Conclusion

Glossary

ecolabel (tourism) Certification recognizing good practice that enhances the environment or minimizes environmental impacts and aims to influence the purchasing behavior of tourists.
ecotourism Ecologically sustainable tourism with a primary focus on experiencing natural areas that fosters environmental and cultural understanding, appreciation, and conservation.
leisure Freely chosen activities pursued in free time that are often intrinsically satisfactory.
recreation Leisure activities occurring close to home.
sustainable tourism Tourism that meets the needs of current tourists and host regions while protecting and enhancing opportunities for the future.
tourism Leisure that involves traveling away from home for more than 1 night but less than 1 year.

Leisure constitutes an integral part of Western societies, which have increasing free time, higher disposable incomes, and an increasing demand for leisure activities. These trends are fortified by demographic trends such as an increasing number of senior citizens who engage in recreational activities or travel. Little is known about leisure in developing countries, and this may be explained to some extent by the (possibly perceived) limited economic significance of leisure and an absence of statistics on leisure behavior and trends. It seems highly likely that leisure activities in developing countries are less energy intensive than those in developed countries. However, on a global scale, it may be that particularly large countries, such as China and India, contribute significantly to energy consumption from leisure. Already, China is seen as one of the most promising markets for tourism development and also as an important source of tourists in the future.

1. BACKGROUND

Given the lack of knowledge on developing countries, this article focuses on Western countries, particularly those where leisure participation is increasingly frequent and diverse. In countries such as the United States, Canada, Japan, Germany, Australia, and New Zealand, the leisure industry continuously develops new products, provides an increasing variety of leisure experiences, and seeks to turn citizens of all ages into leisure consumers. This manifests in the increasing "commercialization" of leisure that encourages providers to compete in the market by providing more and more "thrills" and excitement. Examples of these developments include amusement and theme parks, indoor skiing centers, and entertainment complexes. Many of these offer light amusement that does not require active involvement by the leisure participant but involves considerable inputs of energy to sustain the amusement.

At the same time, in these societies there is a trend toward outdoor activities that provide opportunities for connection with the natural environment. Much of this outdoor recreation is energy intensive, not only through the activity itself (e.g., skiing, diving, boating) but also as a result of transport used to get to the natural amenity where the activity takes place.

Growing mobility manifests not only in travel to recreational activity sites but also in popular travel to foreign places either within the country of residence or abroad. This form of leisure, namely tourism, is a

 © 2004 Elsevier Inc. All rights reserved.

widely acknowledged and researched phenomenon of modern society now constituting the third largest industry globally (after the automobile and petroleum industries). Tourism is an important economic driver contributing between 10 and 12% to the world's gross domestic product.

Although nearly 80% of the world's tourism activity occurs within national borders (domestic tourism), international tourism is the faster growing segment. International visitor arrivals reached 699 million worldwide in 2000. This number is forecast to grow to 1.56 billion tourists by 2020. The top five countries in terms of visitor arrivals are France, Spain, the United States, Italy, and China; however, the market is becoming increasingly diverse, with many new destinations entering the scene. As a result of technological development, a highly competitive market with decreasing airfares, and increasing travel experience and expectations on the part of tourists, long-haul air travel is growing continuously. Tourists travel more frequently for shorter holidays to ever further destinations.

In many developing countries, tourism evolved as a promising tool to stimulate economic development while at the same time offering the potential to support efforts to protect natural and cultural heritage. Sustainable tourism, and more specifically ecotourism, has been discussed widely in this respect.

The trends just outlined result in an increasing demand for energy, mainly for transport but also for energy-intensive infrastructure and facilities. It has been recognized that recreation and tourism contribute considerably to resource use and environmental impacts such as climate change. For this reason, the UN Division for Sustainable Development identifies the indicator "time spent on leisure, paid and unpaid work, and traveling" as 1 of 17 indicators for the provisional core set of "indicators for changing consumption and production patterns."

2. CONCEPT OF LEISURE, RECREATION, AND TOURISM

Leisure has been defined from various perspectives. Most often, leisure is understood as time free of obligations to work, family, or community or to meeting other essential personal needs. Leisure further includes experiences, which are a source of intrinsic satisfaction and which are commonly associated with self-chosen behaviors—passive or active.

Recreation and tourism are leisure activities that have to be seen within the context of leisure. Recreation typically occurs spatially and temporally within reach of home, whereas tourism is leisure away from home. It can be argued that the same factors that encourage tourism (e.g., cheaper travel) also encourage people to extend their spatial and temporal expectations with regard to recreational activities. For example, a kayaking club might not be satisfied with local river conditions and may organize trips away from home to experience more demanding conditions. In this sense, recreation, travel, and tourism may be combined.

However, global statistics draw a line between tourists and recreationists by defining a tourist as any person being away for more than 24 h but less than 12 months from his or her usual environment. By this definition, tourists engage in travel and require accommodation; therefore, they rely on a broader range of services and products to meet their needs than do recreationists.

Visitors who travel for business or to visit friends or relatives are usually included in tourism statistics because they do not differ from "leisure tourists" in their infrastructure needs despite their differing travel motivations. Research on the emerging segment of MICE (meetings, incentives, conventions, and exhibitions) tourists has shown that these travelers engage in recreational activities to a substantial degree during their trips.

Much leisure takes place in the home. Watching television, reading, and listening to music are generally the most frequently undertaken activities. A compilation of studies of leisure participation in Western countries by Cushman, Veal, and Zuzanek provides country-specific information on leisure behavior. For example, Australians walk for pleasure and go swimming, in addition to engaging in home-based entertainment and visiting friends or relatives, in their free time. The most popular activities in Japan are eating out, driving, domestic sightseeing, karaoke, watching videotapes, and listening to music. For the American market, it was found that 94.5% of all U.S. citizens take part in outdoor recreation. Walking (66.7%), visiting a beach or waterside (62.1%), and family gatherings (61.8%) were the most popular activities. More energy-intensive forms of outdoor recreation, such as boating (30.0%) and off-road driving (13.9%), were also undertaken by large numbers of U.S. citizens.

The range of activities just listed indicates varying degrees of energy use associated with both recreation and tourism. Energy is demanded by leisure participants (e.g., consumption of electricity or petroleum) or by the industry. The leisure industry includes those

involved in providing goods, services, and facilities to people in their leisure time. Conceptually, it is useful to distinguish energy demand for upstream activities (e.g., construction or production of goods) and downstream activities (e.g., waste disposal) from energy use resulting from the operation of a recreational facility or activity.

2.1 The Leisure Industry

This article discusses the energy costs of leisure, starting with recreational activities and continuing with accommodation and transport. Recreational facilities or locations (this includes home) are used and visited by both recreationists and tourists. In the context of tourism, these are often called tourist attractions (built structures or natural attractions such as beaches and national parks). Tourist attractions constitute the core of the tourism product because they are the very reason why tourists travel. However, to support leisure activities, a much larger infrastructure and industry are required. Leisure participants get to their activities through various means of transport (walking and cycling are considered transport modes as well). Furthermore, they rely on (often activity-specific) infrastructure and equipment provided by ancillary industries. Tourists require transport and accommodation, and sometimes they also need the services of tour operators. International tourists make further arrangements for travel to their destinations. The various levels of leisure activities and associated services demanded by recreationists and both domestic and international tourists are illustrated in Fig. 1.

Ancillary services and tour operators are not included in the discussion. Although these industries are an important part of the whole leisure industry and are also likely to exert a substantial degree of influence through their purchasing behavior (cf. the tour operators initiative by the UN Environmental Program [UNEP]), they are generally small contributors in terms of energy use. The food and beverage industry is another important industry that provides services to leisure participants. However, little information is available, and this industry is not discussed separately.

3. UPSTREAM/DOWNSTREAM

Upstream and downstream activities constitute the indirect component of energy use such as the construction of recreational facilities and infrastructure and the production of activity-specific equipment. For example, the theater industry uses direct energy for heating and lighting and uses indirect energy via cleaning services, shops that sell the cleaning agents, transport operators who deliver the cleaning agents, and so forth. Energy involved with these processes is called "embodied" or indirect energy. The increasing commodification of leisure has contributed to an increased direct and indirect input of energy and material.

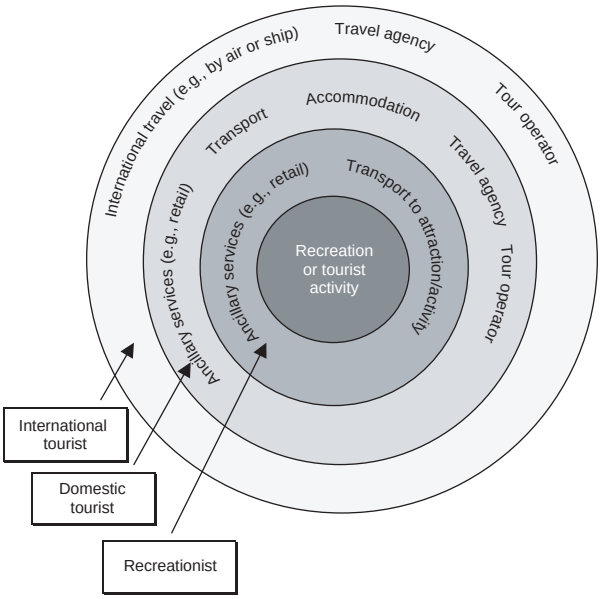


FIGURE 1 Overview of leisure components for recreationists, domestic tourists, and international tourists.

For some activities, energy use associated with upstream activities is considerable such as in motor sports and air activities (e.g., [para]gliding, skydiving). Skiing and snowboarding are other popular activities that require large energy inputs outside of the actual activity. In a recent study in Switzerland, Stettler found that the energy input for infrastructure and equipment amounts to 200 MJ per skier-day compared with 55 MJ for tennis and 15 MJ for soccer. In these examples, the indirect energy use is more than twice the direct energy use. In the case of snow sports, considerable energy costs result from artificial snowmaking, which can be considered an upstream activity because it is often undertaken to enable ski fields to open early in the season. A study on environmental impacts resulting from leisure in Japan found that the indirect energy use from leisure is about seven times larger than the direct energy use, whereas a study undertaken in New Zealand indicated that indirect energy costs of tourism are about the same as direct costs. In both cases, the

consideration of flow-on effects adds a substantial amount of energy to the direct energy costs.

Little specific information is available on downstream activities such as the disposal or recycling of equipment. However, it is believed that considerable energy use is involved with the disposal of sometimes highly technical sports equipment. Similarly, waste management associated with tourist accommodation is likely to consume considerable amounts of energy such as for transportation but also for waste (both solid and liquid) treatment.

Information on indirect energy use of leisure is scarce; therefore, this article focuses on direct energy use of recreational and tourist activities. It is acknowledged that total (both direct and indirect) energy use exceeds the direct energy use discussed in the remainder of this article.

4. RECREATIONAL ACTIVITIES

The discussion of recreational activities in this section focuses on active recreation away from home (often tourism related) because it is believed that energy use of such activities is larger than that of passive entertainment at home. It is acknowledged that home-based activities consume some energy, both directly (e.g., electricity) and indirectly (e.g., embodied energy in equipment or appliances). Recreational activities away from home include both visits to attractions as a permanent establishment or point of interest and "activities" in the true sense of the word. Activities include walking, swimming, biking, taking photographs, water sports, mountaineering, and taking guided tours.

4.1 Operation of Attractions and Activities

Many people engage in leisure on an individual basis (e.g., going for a walk, reading a book). Many of these activities require little direct energy input. However, an increasing number of activities are offered through institutions or businesses that consume energy for their buildings, operations, and transport.

The energy use of various recreational attractions or activities is strongly dependent on visitor volumes. The largest amusement park worldwide is Disneyland Tokyo, attracting more than 16 million visitors per year. On the other hand, there are many small (tourism) businesses operating for an exclusive market of several hundred visitors per season.

In a topical report on tourism, the European Environment Agency stated that recreation facilities can constitute a big source of energy demand in the local context. Ski fields in France consume between 2.1 and 2.6 PJ for 4000 ski lifts. Similarly, leisure facilities in urban environments are often main consumers of energy. The opera theater in Zurich, Switzerland, consumes approximately 1190 GJ per year, and the ice skating stadium in Berne, Germany, consumes approximately 5950 GJ per year. The very popular Japanese activity of playing Pachinko (similar to pinball) consumed a total of 49 PJ in 1995, just behind the energy consumption of eating and drinking places (59 PJ) and well ahead of other leisure activities (e.g., sports facilities, gardens, and amusement parks [9.3 PJ]).

In New Zealand, tourist attractions (e.g., museums, parks, historic sites, geothermal sites) consume on average 400 GJ per year, and entertainment facilities (e.g., shopping malls, theaters, amusement parks) consume approximately 1600 GJ per year. In contrast, businesses that offer recreational activities (e.g., scenic boat cruises, guided walks, adventure tour operators) are comparatively small energy users, using approximately 500 GJ per business per year. When comparing these attractions and activities on a per capita basis, the attractions become energy efficient, with an energy use of approximately 20 MJ per visit, compared with activities that require energy inputs of at least 50 MJ and up to approximately 2000 MJ per visit (e.g., scenic flights). Activity packages are energy intensive because they often offer an individual style and service orientation that requires energy at various stages as compared with (mass tourism) attractions. Furthermore, leisure activities often build on the use of motorized vehicles, either to get to the attraction or activity or for the activity itself (e.g., snowmobiling). In general, the input of petroleum fuels is characteristic of "activities," whereas "attractions" require electricity for their built environment.

An example of a typical tourist activity package in New Zealand is guided (alpine) walks (Fig. 2). Energy is consumed in operating the building where customers book the tour and receive their equipment. A shuttle bus takes customers to the glacier, where they start the guided glacier walk or go to the helicopter pad in the case of heli-hikes. After the activity, tourists are taken back to the main building, where they are served hot drinks. The total energy use per tourist is 88 MJ on average. Energy use per visit or per participation in an activity is a useful indicator for comparing various leisure activities (Table I).

The degree to which an activity depends on motorized vehicles largely determines its overall energy intensity. Sports traffic is an interesting example of the use of transport for recreation. For example, the average travel distance to sports activities in Switzerland was estimated to be 25 km return, resulting in an energy use of 60 MJ per sportsperson (considering specific transport mixes). Travel distances differ considerably for sports activities; the average one-way travel distance for auto sports is approximately 710 km, for skiing 87 km, for golf 32 km, and for jogging (the lowest) 4.6 km. Most sports travel is based on private motorized vehicles (78%), whereas only 18% are by public transport and 4% are by bike or on foot. Overall, sports traffic makes up more than 10% of primary energy demand in Switzerland; in 1995, 24 PJ was used by sports participants, 1.8 PJ by trainers and officials, and another 3.7 PJ by spectators.

The relatively new, yet very popular, "ecotourism" and "ecotourism activities" seem to be responsible and nature-based forms of tourism that seek to contribute to nature conservation, yet they are rarely discussed in the light of energy use. However, many ecotourism operators provide "ecotours" that take visitors to natural assets with various types of motorized vehicles, often four-wheel drives or boats. The environmental impact resulting from the energy use embodied in these tours has often been neglected, with researchers focusing on local environmental impacts. Clearly, against this background, there is a strong contradiction within the concept of ecotourism.

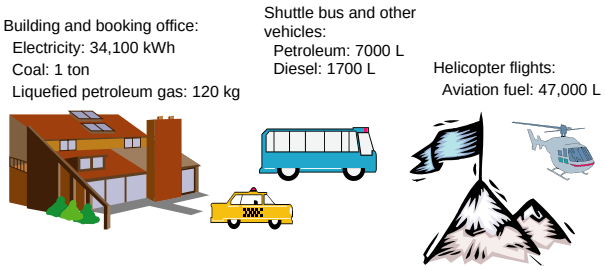


FIGURE 2 Direct energy use per year of a guided alpine walk operator in New Zealand.

4.2 Best Practice

Many initiatives have been reported to minimize the energy use of leisure activities. For example, the Australian National Maritime Museum in Sydney seeks to increase energy efficiency by improving the saltwater-cooled air-conditioning system, reducing fan speed, installing lighting sensors, and using energy-efficient office equipment. These initiatives resulted in considerable cost savings. Another example is SeeQuest Adventures (Canada), which has renounced motorized vehicles and has clients walk or paddle to campgrounds on its guided tours. The replacement of gasoline motors with electric ones by Costa Rica Expeditions helped to decrease the use of

TABLE I

Examples of Energy Intensities for Various Leisure Activities

Activity	Direct energy use per visit (MJ)	Source	Country of reference
Watching television (3.8 h/day)	1.8	Müller	Switzerland
Experience centers	29	Becken and Simmons	New Zealand
Museums, art galleries	10	Becken and Simmons	New Zealand
Gondola ride to Pilatus (Luzern)	8.5[a]	Geisel	Switzerland
Restaurant meal	18.0	Müller	Switzerland
Swimming in public pool	46.8	Müller	Switzerland
Swimming in public pool	119[a]	Motiva	Finland
Indoor ice skating	28.8	Müller	Switzerland
Skiing	90[a]	Motiva	Finland
Golf	12	Becken and Simmons	New Zealand
Rafting	36	Becken and Simmons	New Zealand
Heli-skiing	1300	Becken and Simmons	New Zealand
Scenic boat cruises	165	Becken and Simmons	New Zealand

[a] Primary energy, otherwise secondary energy (measured as heat content).

fossil fuels, whereas Odyssey Safaris reduces fuel consumption by training drivers and maintaining vehicles (e.g., tire pressure). Tasmania's West Coast Charters (Australia) achieved energy savings worth $6000 (Australian) per year by using wind-powered yachts instead of motorized boats.

Australia has implemented the first Nature and Ecotourism Accreditation Programme to ensure that an accredited company is committed to best environmental practice and provides quality experience. Energy efficiency is one key criterion in this program, specified as minimizing energy use in natural tourism and ecotourism facilities (including a biannual energy audit) and minimizing energy use for transport.

It can be concluded that the energy requirements of recreational activities range from minimal to extremely large. When attractions or activities are operated by businesses, the determining factors for varying energy efficiency include the visitation levels, management style and commitment, service level (e.g., shuttle transport), technical equipment, fuel mix, and vehicle use (including type of vehicles, occupancies, and travel distances).

5. TOURIST ACCOMMODATION

5.1 Description of the Accommodation Sector

The accommodation sector is extremely diverse. Not only are there very different types of accommodation within specific regions, but there is also a large diversity worldwide as a result of different cultural backgrounds. "Indigenous tourism" often involves staying at traditional housing forms such as long houses in Indonesia, maraes in Fiji, and yurts in Mongolia. In Western countries, commercial accommodation categories include mostly hotels (often rated through quality systems such as stars), resorts, motels or motor lodges, apartments, youth hostels and other budget accommodation, camping grounds (often with cabins or caravans), and partially commercial forms such as homestays, farmstays, and bed & breakfasts.

The energy use while staying in the private homes of friends or relatives is another important aspect of tourism accommodation. For example, U.S. domestic travelers spend 46% of their nights during travel at the homes of friends or relatives. In New Zealand, the share is even bigger, with domestic tourists staying more than half of all visitor nights at private homes.

Commercial accommodation offers various services in addition to a sleeping facility. These include washing and cooking facilities, eating and drinking places, entertainment services such as television lounges and casinos, and other amenities such as swimming pools, tennis courts, and fitness studios. At the very top end of services provided are "all-inclusive" resorts, where all tourist activities are provided in a packaged form through the accommodation business. This form of accommodation is widely established in the Caribbean and is now developing in places such as Mexico, Gambia, Malaysia, and the Mediterranean. In the case of all-inclusive holidays, the accommodation business often becomes a tourist destination in its own right, making the geographic location largely irrelevant. This high service level clearly has implications for the total energy use given the high energy inputs required to sustain such megaresorts. However, parts of this may be outweighed by the decreased transport needs of all-inclusive tourists.

5.2 Energy Use by Various Accommodation Categories

In strongly developed tourism destinations, the accommodation industry may contribute substantially to total energy demand. For this reason, a "roundtable" on ecoefficiency in the tourism sector was held in 1997 in the Caribbean. A large portion of electricity used on the Caribbean islands is to satisfy the large energy demand of the numerous tourist resorts (consumption of 6.1 PJ in 1994). This energy consumption carries a large "rucksack" given that electricity is usually generated from oil shipped to the islands. Total energy use for accommodation amounts to 29.2 PJ per year in Japan (0.1% of total national energy demand of $\sim$25,000 PJ), whereas accommodation in New Zealand consumes approximately 4 PJ per year ($\sim$1% of total national energy demand of 450 PJ).

For a single accommodation business, the energy use depends on a number of factors such as climate, size and type of building, technological standard, staff awareness, and visitation levels. For example, in the case of hotels, operational energy use ranges between 2000 GJ per year (e.g., New Zealand) and 3000 GJ per year (e.g., North America). Most energy savings could probably be achieved through the design, construction, or refurbishment of places by considering aspects of location and building design.

An important factor for energy use is the level of service provided. Accommodation can be split into

two broad categories: "comfort oriented" and "purpose oriented." Hotels, luxury lodges, and resorts clearly fall into the first category, whereas youth hostels, huts, and camping facilities are purpose oriented. This distinction makes it easier to establish benchmarks or codes of practice for accommodation businesses, which otherwise would ignore very different consumer needs (e.g., a business traveler vs a rucksack tourist) and, hence, different energy requirements to meet these needs.

The energy use per visitor night differs considerably between comfort-oriented and basic forms (Table II). Hotels and bed & breakfasts consume about three to five times the energy of motels, backpacker hostels, and campgrounds.

Energy costs constitute approximately 3 to 10% of business costs, which is the third highest cost after staff and food. Space heating or cooling, water heating, and refrigeration account for most of the energy use of accommodation businesses (Fig. 3). In warmer climates, space heating is redundant, but hotels use air conditioning to keep temperatures comfortable. In the Caribbean, air conditioning accounts for 44% of the energy use in a large hotel (more than 100 rooms).

The energy mix determines energy costs and environmental impacts. Most accommodation businesses rely mainly on electricity. For example, in New Zealand, 94% of the energy use for motels is supplied by electricity, compared with 74% for hotels and 44% for campgrounds. When generated from fossil fuels, electricity has a large amount of embodied energy. On the other hand, when based on renewable sources, particularly at remote locations (e.g., ecotourism lodges), equipment based on electricity (e.g., for water heating) has the potential to be sustainable.

5.3 Best Practice

Examples of best practice include building designs allowing for maximum use of sunlight, cross-ventilation, and weather protection; solar panels; a quantum hot water system; low-watt fluorescent lights and other energy-efficient equipment; and the installation of guest room occupancy sensors for heating, cooling, and lighting.

There are various environmental practice codes that support businesses in decreasing energy use. *Case Studies on Environmental Good Practice in Hotels* by UNEP and publications by the International Hotels Environment Initiative are two examples. There are also location-specific codes such as the energy code for energy-efficient buildings in Hawaiian hotels developed specifically for the unique conditions of Hawaii.

Single businesses engage in best practice to save costs, to decrease environmental impacts, and to gain a competitive advantage. For example, Turtle Island, Fiji, uses solar power for hot water and wind

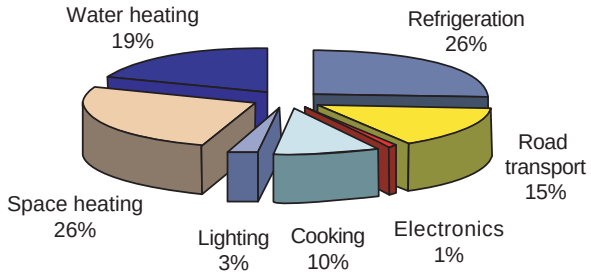


FIGURE 3 Energy end use in the New Zealand hotel sector. (Source: Energy Efficiency and Conservation Authority. (1996). Hotel sector energy use: Highlights. *Energy-Wise Monitoring Q.*, No. 4 [Wellington, New Zealand]).

TABLE II

Examples of Energy Intensities for Various Accommodation Categories

Accommodation type	Direct energy use per visitor-night (MJ)	Source	Country of reference
Hotel	151	Müller	Switzerland
Hotel	200	Brunotte	Germany
Hotel	circa 260	Gössling	Zansibar
Commercial accomodation	94	Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit	Germany
Hotel	155	Becken *et al.*	New Zealand
Bed & breakfast	110	Becken *et al.*	New Zealand
Motel	32	Becken *et al.*	New Zealand
Backpacker	39	Becken *et al.*	New Zealand
Campground	25	Becken *et al.*	New Zealand

generators for other electricity use. The Binna Burra Mountain Lodge in Queensland, Australia, is a Green Globe 21-certified company that reduced electricity use by 20% through energy-efficient lights. Total energy use was reduced by 15,000 kWh (189 metric tons CO_2) per year as a result of weatherproofing buildings, replacing diesel with gas for water heating, and using gravity for water supply. Transport energy use can also be reduced by transporting customers and/or staff in mini-buses to the accommodation businesses (e.g., Freycinet Experience, Tasmania, Australia), by replacing four-wheel drives with two-wheel drives (e.g., Australis, Australia), and by encouraging cycling. The need for travel can also be reduced by increasing the length of stay at a single location (e.g., Crystal Creek Rainforest Retreat, Queensland, Australia).

6. TOURIST TRANSPORT

6.1 Importance of Transport

Travel for leisure purposes is different from other travel in that it is chosen largely in terms of destination, length of travel, and transport mode. It has been estimated that leisure travel constitutes more than half of all travel in Western countries. Transport for recreation has already been discussed in the context of energy use associated with recreational activities or attractions. In this section, transport is examined in a broader tourism-related context.

Mobility is an essential prerequisite for tourism. Research on tourism systems has long emphasized that transportation connects the tourist generating and the destination region via transit routes and that transportation also ensures mobility within a destination, either to visit tourist attractions or to travel along recreational routes. Often, transport is more than a means to get from one place to another; it also constitutes an attraction in itself, as in the case of cruising trips, scenic flights, and railway journeys.

In the case of single-destination trips (often resort based), transport is used mainly to travel to the destination or to undertake multiple-day trips from the base location. An increasing number of trips are to more than one destination; this includes "touring tourists," who stay in different places within one country, and tourists who visit different countries on one trip such as round-the-world travelers. New Zealand is a typical site for both types of multiple-destination travel. Approximately 75% of all international visitors travel to more than one place within New Zealand, and 36% visit another country on their way to or from New Zealand.

The development of various forms of transport, such as private vehicles, high-speed trains, and airplanes, makes it possible for a large number of tourists to engage in travel. This manifests in mass tourism phenomena such as organized coach tours and chartered flights to resort holiday destinations. Today, one in five German air travelers is going on a package holiday by charter flight.

6.2 Transport Modes Used by Tourists

Leisure travel is based largely on private vehicles and airplanes. A study on travel in the United States in 1995 showed that a large majority of trips (81%) were made in personal-use vehicles and that these trips accounted for more than half of all person-miles. Air travel, although undertaken by only 15% of all travelers, accounted for most of the other person-miles. Other transport modes played a minor role for U.S. citizens (Table III). The situation is similar in New Zealand, where 77% of all domestic tourist trips are car based, with another 13% being airline flights. In the European Union, the share of other transport modes is higher. Although most travel is still by road (61%) and by (21%), train (15%) and ship (3%) are also frequently used modes of travel. The use of public transport for leisure is generally low because it is often perceived unreliable, expensive, time-consuming, and inconvenient.

TABLE III

Travel (>100 Miles) in the United States

Transport mode	Person-trips (thousands)[a]	Percentage	Person-miles (millions)	Percentage
Personal-use vehicle[b]	813,858	81.3	451,590	54.6
Commercial airplane	155,936	15.6	347,934	42.1
Intercity bus	3,244	0.3	2,723	0.3
Chartered tour bus	14,247	1.4	9,363	1.1
Train	4,994	0.5	4,356	0.5
Ship, boat, or ferry	614	0.1	1,834	0.2

Source. 1995 American Travel Survey.

[a] Of these trips, 23% are business trips, 33% are for visiting friends or relatives, and 30% are leisure trips.

[b] Private car, truck, or camper van.

The distance traveled per trip is considerable. For example, in the European Union, tourists travel on average 1800 km per trip, whereas in New Zealand, domestic tourists travel 623 km per trip (with 4.4 trips taken per person per year) and international tourists travel 1950 km during their stays in New Zealand. In 1995, the average trip length of Americans in a personal-use vehicle was 590 km and in the case of air travel was 1732 km.

6.3 Energy Use of Transport

In most countries, transport is an important source of energy demand. Worldwide transportation contributes 20 to 25% to all fossil fuel consumption. Given the dominant role of leisure in transport demand, there is a clear need to integrate leisure traffic into transport management plans and strategies to reduce energy use.

In Japan, total energy use associated with leisure transport amounted to 260 PJ in 1995, with the largest single contributor being rental car and taxi services (98 PJ). Tourism transport alone (excluding recreation) makes up 4% of New Zealand's national energy consumption. A recent report on tourism and the environment by the German government states that inbound holiday travel (excluding business and visiting friends or relatives) makes up 7% of all passenger travel's greenhouse gas emissions.

The most energy-intensive forms of travel are by air and sea. On a per passenger-kilometer (pkm) basis, air travel consumes between 2 MJ for international flights (e.g., 1.9 MJ/pkm for Lufthansa, 2.2 MJ/pkm for British Airways, 2.5 MJ/pkm for Scandinavian Airlines) and 3 MJ for domestic flights (e.g., 2.75 MJ/pkm for Air New Zealand). Ferry travel is comparably energy intensive (e.g., 3.5 MJ/pkm for Sydney Harbor Ferries, 2.0 MJ/pkm for Brisbane Ferries). In comparison, private cars consume between 3 and 4 MJ per vehicle kilometer; however, the per capita energy use can be decreased substantially by high load factors. The energy intensity of leisure-specific transport modes in the New Zealand context is presented in Table IV.

Vehicles that are designed particularly for recreational purposes, such as jet boats, motorboats, snowmobiles, and off-road vehicles, are often comparatively energy intensive. This may be explained by the fact that power and speed, rather than efficiency and economy, are features of the leisure experience.

On the other hand, a range of nonmotorized recreational transport options exist. These include cycling, walking, horseback riding, kayaking, sailing, and cross-country skiing. It is likely that these nonmotorized recreational transport options appeal to only a small sector of recreational consumers. Cycling has the largest potential to be accepted as an alternative transport mode, whereas the other examples are likely to be considered as mere recreational activities. Many countries recognize the integration of cycling into transport planning as an important tool for achieving sustainable transport. The Netherlands, Denmark, Germany, and Great Britain already have extensive cycle routes for recreational biking and networks that support long-distance cycle tourism. The conversion of former rail tracks into multipurpose user trails for bikers and walkers in North America has proven to be very popular.

TABLE IV

Energy Intensities of Transport Modes Used by Leisure Travelers in New Zealand

Transport mode	Direct energy use per passenger-kilometer (MJ/pkm)
Domestic air	2.75
Private car	1.03
Rental car/Taxi	0.94
Organized coach travel	1.01
Scheduled (intercity) bus	0.75
Backpacker bus	0.58
Camper van	2.06
Train	1.44
Cook Strait ferries	2.40
Yacht/Boat	1.75
Other ferries	3.53
Motorcycle	0.87
Helicopter	4.68

Source. Becken, S. (2002). "Tourism and Transport in New Zealand: Implications for Energy Use" (TRREC report 54). Lincoln University, New Zealand.

6.4 International Travel

International travel, which includes both air and sea travel, plays a special role because it is responsible for the largest share of energy use. Furthermore, it is not accounted for in national energy or greenhouse gas accounts. Because air travel in particular is growing consistently at rates of approximately 9% per annum, paralleling tourism growth rates, energy use by international travel is becoming increasingly important.

Aviation contributes approximately 2 to 3% to global energy demand, and at least half of this energy

use is due to tourist travel. An increasing number of countries, often developing or remote ones, rely on international tourism as a major economic activity and foreign exchange earner. However, associated with this is the dependency of these countries on fossil fuels. External costs of aviation are not accounted for at this stage in international agreements and airfares, but once they are implemented (e.g., through the Kyoto Protocol during the second commitment period), they could severely affect these destinations.

In the New Zealand case, the total energy use involved in transporting tourists to the country and back home by air amounts to 55.6 PJ per year. If this were added to the national energy account, the total would increase by more than 12% and render compliance with international agreements (e.g., the Kyoto Protocol) difficult.

Similar considerations apply to cruise ship tourism, although the total tourist volumes are much smaller than those for air travel. In 1994, there were approximately 5 million cruise passengers worldwide who undertook some 37 million cruising-passenger days. Many of these tourists buy "fly and cruise packages." Energy use by these tourists has rarely been analyzed for several reasons. First, cruising companies are large international firms that are reluctant to reveal commercially sensitive energy consumption figures. Second, fuel consumption varies considerably among different types of cruising ships. Third, the fuel is consumed for transport and electricity generation, so it includes on-board hospitality functions. Finally, cruise ship tourists consume additional energy on their often frequent land trips, and this renders accounting for the total energy requirements of cruise tourists even more complex. On the basis of shipping companies advertising on the World Wide Web, it is estimated that a cruising ship consumes between 1500 and 4000 MJ per passenger-day, and this is large compared with other transport or accommodation options.

6.5 Best Practice

A large number of initiatives have been made by the tourism industry to make transportation more sustainable. These are represented in business actions and initiatives by tourist destinations.

Business initiatives exist at different levels, ranging from small and locally operating companies to large international organizations. The largest reductions in energy use are expected from large organizations. For example, Budget Rent-a-Car rents electric cars in France and operates an electric shuttle bus at the Los Angeles airport (LAX). It has established a national bike hire service and distributes fuel-saving tips to customers. Similarly, Avis increased the percentage of small cars in its fleet and offers renting schemes based on mileage instead of days to provide an incentive to travel less.

British Airways is well known for its environmental actions, for example, by reducing the energy use of land-based operations, by providing mileage allowances for business travel by bicycle, and by encouraging the use of public transport. Both KLM Royal Dutch Airlines and Lufthansa developed "Air–Rail Projects" where connecting flights are substituted by rail travel. This initiative is supported by the German Touristik Union International (TUI). Furthermore, Lufthansa seeks to continuously modernize its fleet, improve air traffic management (navigation), and decrease ground energy use by employing cogeneration to heat water, leasing bicycles to staff, optimizing vehicle travel, and using flight simulators. The latter alone saves 450,000 tons of fuel per year. The East Japan Railway Company aims to decrease energy consumption by reducing the weight of train cars and by using regenerative brakes. Electric power plant generators will be replaced by combined steam and gas turbine cycle generators.

Tourist destinations have developed a variety of strategies to decrease transport, partly because the attractiveness of many tourist resorts is reduced by noise, pollution, and congestion. The Gemeinschaft Autofreier Schweizerischer Tourismuorte (GAST) is one example, where nine communities joined in 1988 to promote car-free tourist resorts. Although the use of motorized vehicles in these nine resorts is prohibited, most tourists (71% during summer, 69% during winter) still travel by car to and from the resorts. Another program in the European Alps, Sanfte Mobilitaet, explored the potential of cards for travel on local public transport (Bodenmais, Germany), cycle routes (Oberstdorf, Germany), and a "Talerbus" that connects several neighboring valleys (Lungau, Austria). The Talerbus proved to be very successful by attracting a large number of visitors. However, problems arose when the short-term funding provided by the European Union expired and the future existence of the bus system became uncertain.

The numerous actions by operators and destinations to decrease traffic are promising. However, it is the tourists or recreationists themselves who make travel decisions and who exert substantial influence through their consumer behavior. In this sense, it is critical to incorporate travel behavior into current transport management.

7. CONCLUSION

Tourism and recreation, especially outdoor recreation, depend on natural resources and a healthy environment. For this reason, tourism is often seen as a potential ally of nature conservation. However, the large energy consumption linked to leisure makes this association tenuous. The total leisure-related energy use in Switzerland was estimated to make up approximately 43% of total national energy use; 16% of total energy is consumed in the living environment (home), for example, 11% in leisure transport, and 6% each in recreational activities and tourism In Japan, the total energy use for leisure contributes about 12% to national energy demand. Tourism alone makes up approximately 6% of energy use in New Zealand, not including energy use of international travel to and from New Zealand.

Recreational facilities (e.g., parks, sports grounds) and activities (e.g., walking, other physical activities) often constitute the core of leisure. However, energy use and resulting impacts are mostly local and relatively small compared with other components of the leisure industry. The accommodation industry and other ancillary services are more important in terms of energy consumption. However, the dominant energy-consuming sector is transportation. Energy use for leisure transport is considerable and is expected to continue growing, particularly due to the heavy use of private vehicles and the increasing popularity of air travel. Transport is also the dominant contributor to an individual traveler's total energy use. International tourists in New Zealand consume approximately 4000 MJ per trip within New Zealand, of which 69% is for transport. If the energy use of international return flights by overseas visitors were included, transport would constitute 97% of a tourist's total energy use. Similar relations were found for holiday travel within Germany (excluding business and visiting friends or relatives); fully 63% of total energy use was for transport, 28% for accommodation, and 9% for recreational activities. Clearly, tourism is a highly energy-dependent industry that would be severely affected by economic measures to decrease energy use such as fuel or emissions taxes.

The leisure, recreation, and tourism industries have been proactive in reducing resource use and mitigating environmental impacts. Numerous business actions illustrate a shift in practice, and the introduction of ecolabels for the tourism industry is another step in this direction. There are approximately 100 different certification programs, ranging from regional to global schemes. Green Globe 21 is an example of a global certification program that was launched by the World Travel and Tourism Council but is now an independent company. Green Globe 21 seeks to increase awareness among tourism businesses and provides tools to implement environmentally sound practices (e.g., benchmarking). Reducing energy costs is a key element of the Green Globe 21 system.

To decrease energy use associated with leisure, and more specifically energy use associated with tourism, in an effective manner, concerted action by industry and consumers will be necessary. For example, the utility of ecolabels is uncertain because it is not yet clear whether business efforts and possibly increased costs in the short term are compensated by gaining a competitive advantage in the market. It seems particularly difficult to change the transport behavior of leisure participants who expect a high level of comfort and promptness of transport during leisure time. The balance between tourism as an important industry in many countries (and as a mechanism for development in some others) and as a significant contributor to greenhouse gas emissions (mainly as a result of long-distance flights) remains.

Acknowledgments

I am grateful for the useful comments on an earlier draft by Bob Gidlow and Chris Frampton.

Further Reading

Becken, S. (2002). Analysing international tourist flows to estimate energy use associated with air travel. *J. Sustainable Tourism* **10,** 114–131.

Becken, S., Frampton, C., and Simmons, D. (2001). Energy consumption patterns in the accommodation sector: The New Zealand case. *Ecol. Econ.* **39,** 371–386.

Becken, S., and Simmons, D. (2002). Understanding energy consumption patterns of tourist attractions and activities in New Zealand. *Tourism Mgmt.* **23,** 343–354.

Brunotte, M. (1993). "Energiekennzahlen für den Kleinverbrauch: Studie im Auftrag des Öko-Instituts." Freiburg, Germany.

Bundesministerium für Umwelt, Naturschutz, und Reaktorsicherheit. (2002). "Tourismus und Umwelt." www.bmu.de/fset800.php.

Cushman, G., Veal, A. J., and Zuzanek, J. (1996). "World Leisure Participation: Free Time in the Global Village." CAB International, Wallingford, UK.

Ecotourism Association of Australia. (2002). "NEAP: Nature and Ecotourism Accreditation Programme," 2nd ed. www.ecotourism.org.au/neap.pdf.

Font, X., Buckley, R. C. (eds.) (2001). "Tourism Ecolabelling: Certification and Promotion of Sustainable Management." CAB International, Wallingford, UK.

Geisel, J. (1997). "Ökologische Aspekte zum Reisebustourismus in Luzern und Ausflugstourismus zum Pilatus: Diplomarbeit im Studiengang Geooekologie." http://cobra.hta-bi.bfh.ch/home/gsj/downloads/uni_ka/da_goek.pdf.

Gielen, D. J., Kurihara, R., and Moriguchi, Y. (2001). The environmental impacts of Japanese leisure and tourism: A preliminary analysis. www.resourcemodels.org/page8.html.

Gössling, S. (2000). Sustainable tourism development in developing countries: Some aspects of energy use. *J. Sustainable Tourism* **8,** 410–425.

Holding, D. M (2001). The Sanfte Mobilitaet project: Achieving reduced car-dependence in European resort areas. *Tourism Mgmt*. **22,** 411–417.

Høyer, K. G. (2000). Sustainable tourism or sustainable mobility? The Norwegian case. *J. Sustainable Tourism* **8,** 147–160.

Motiva. (eds.) (1999). "Enduser's Energy Guidebook for Schools." www. motiva.fi.

Müller, H. R. (1999). "ETH-Pilotprojekt "2000 Watt Gesellschaft": Arbeitsgruppe "Freizeit und Energie." Bern, Switzerland.

Life Cycle Analysis of Power Generation Systems

JOULE BERGERSON and LESTER LAVE
Carnegie Mellon Electricity Industry Center
Pittsburgh, Pennsylvania, United States

1. **Introduction**
2. **Methods for Life Cycle Analysis**
3. **Brief Historical Review**
4. **Coal**
5. **Natural Gas**
6. **Hydro**
7. **Oil**
8. **Nuclear**
9. **Biomass**
10. **Wind**
11. **Solar**
12. **Conclusions**

Glossary

discharges Discharge of liquid effluent or chemical emissions into the air from a facility through designated venting mechanisms.

economic input-output life cycle analysis (EIOLCA) A tool used to assess the environmental impact of a particular product or service by linking all the various economic transactions, resource requirements, and environmental emissions required in its manufacture.

SETAC/EPA life cycle analysis Holistically analyzing the cradle-to-grave environmental impact of products, packages, processes, and activities, from raw material acquisition to final disposition.

To make informed decisions about electricity generation options, companies, concerned citizens, and government officials need good information about the environmental implications of each fuel and generation technology. To be most helpful, an analysis must examine the life cycle of each fuel/technology, from extraction of the materials to disposal of residuals. In this article, we review studies examining the life cycle environmental and health implications of each fuel and technology. We begin with an examination of the methods of life cycle analysis, then present a brief historical overview of the research studies in this area. Finally, we review and critique the alternative methods used for life cycle analysis. Our focus is on the recent studies of the health and environmental implications of each tehnology.

1. INTRODUCTION

Increases in electricity demand and the retirement of old generating plants necessitate investment in new generation. Increasingly stringent environmental regulations, together with other regulatory requirements and uncertainty over future fuel prices, complicate the choice of appropriate fuels and technologies.

Electricity generation, a major source of CO_2, SO_2, NO_x, and particulate matter, also produces large quantities of solid waste and contributes to water pollution. To make informed decisions about refurbishing old plants or investing in new ones, companies, concerned citizens, and government officials need good information about the environmental implications of each fuel and generation technology. New issues have surfaced recently, such as discharges of mercury and total greenhouse gas emissions. Since other potential issues loom (e.g. other heavy metals), an environmental analysis must examine the life cycle of each fuel/technology, from extraction of the materials to disposal of residuals.

This article reviews studies examining the life cycle environmental implications of each fuel and technology. It focuses on the coal fuel cycle since (1) it accounts for more than half of the electricity generated in the United States; (2) historically, the coal fuel cycle has been highly damaging to the

 © 2004 Elsevier Inc. All rights reserved.

environment and to health; (3) there are huge coal reserves in the United States, China, and Russia; and (4) the fuel is inexpensive to mine and likely to be used in large quantities in the future. The article begins with an examination of the methods of life cycle analysis. Recent advances in life cycle analysis offer large improvements over the methods of three decades ago and should help in choosing among fuels and technologies as well as modifying designs and practices to lower the health and environmental costs. Then a brief historical overview of the research studies in this area is presented. Finally, the alternative methods used for life cycle analysis are reviewed and critiqued. The article focuses on recent studies of the health and environmental implications of each technology. The studies agree that coal mining, transport, and combustion pose the greatest health and environmental costs. Among fossil fuel fired generators, natural gas power turbines are the most benign technology. Light water nuclear reactors received a great deal of attention in the early literature but are neglected in recent U.S. studies. The earlier studies found that the health and environmental costs of light water reactors were low, at least for the portions of the fuel cycle that were evaluated. The studies did not evaluate the disposal of spent fuel and thus are incomplete.

2. METHODS FOR LIFE CYCLE ANALYSIS

Life cycle analysis (LCA) is needed for informed decisions about alternative fuels and technologies. Modern LCA is divided into (1) scoping, (2) discharge inventory, (3) impacts, and (4) improvement. Since a comprehensive analysis is impossible, each analyst must decide, explicitly or implicitly what will be considered in the analysis.

The Environmental Protection Agency (EPA) and the Society of Environmental Toxicologists and Chemists (SETAC) developed and formalized methods for conducting LCAs in the 1990s. The basic steps are conducting mass and energy balances of each relevant process. Thus, the analyses tend to be time consuming and expensive; the first step, scoping, refers to reducing the number of processes to be studied to a feasible level. An LCA can be quicker if the results from previous analyses are used for each process. However, using old data lowers accuracy. Unfortunately, the quantitative estimates are generally uncertain and often controversial. For example, the "best" technology can change as the scope of the study changes. Holdren's 1978 criticism that electricity life cycle analyses exclude important aspects and take insufficient care still applies.

A new approach to LCA was developed using the Department of Commerce's 485×485 commodity input-output model of the United States economy and publicly available environmental data by Lave and Hendrickson. This approach is quick and inexpensive. The disadvantage is that it is at an aggregate level.

3. BRIEF HISTORICAL REVIEW

In the early 1970s, Lave and Freeburg as well as Sagan performed the first comprehensive analyses on the effects of power plants. They found that coal posed significant environmental risks, from mining, transport, and generation. Both studies found that oil and natural gas have much smaller environmental and health costs. Finally, both found that light water reactors have an even lower health burden, although neither could assess the environmental and health burdens of dealing with spent fuel, decommissioning old reactors, or of potentially catastrophic events.

Numerous additional studies were conducted throughout the 1970s and early 1980s that also primarily focused on coal and also found major problems. More recent papers have examined new technologies and newer data. In 2002, Pacca and Horvath as well as Meier conducted studies using a new life cycle analysis tool. Some recent studies focus on environmental burdens and impacts. Some compare several fuel cycles using CO_2 payback times (the time required to "payback" the CO_2 emitted from constructing the power plant), while others evaluate the global warming potential (GWP) of a fuel/technology. Pacca and Horvath proposed a metric, global warming effect (GWE) to compare technologies based on global impact by accounting for the time dependency of GWPs. Table I shows Pacca and Horvath's estimates of the total GWE for different fuels/technologies (coal, hydro, PV, wind farm, and natural gas) after 20 years of operation.

Renewable technologies have received increasing attention in recent studies. Many studies have concluded that renewable technologies, such as wind, hydro, solar-thermal and photovoltaic, are less attractive environmentally when evaluated using an LCA.

In 2002, Gagnon *et al.* reviewed previous LCAs. This paper focused on hydropower (run-of-river and

TABLE I

Global Warming Effect of Five Fuels/Technologies[a]

	Hydroelectric	Photovoltaic	Wind farm	Coal	Natural gas
Output (TWh)	5.55	5.55	5.55	5.55	5.55
			Emissions (MT CO_2 equiv.)		
CO_2 ($\times 10^6$)	5	1	0.8	90	50
CH_4 ($\times 10^6$)	0.01	0.008	0.0005	0.4	0.5
N_2O ($\times 10^6$)	0.09	0.09	0.007	2	4
GWE ($\times 10^6$)	5	1	0.8	90	50

[a] From Pacca and Horvath (2002).

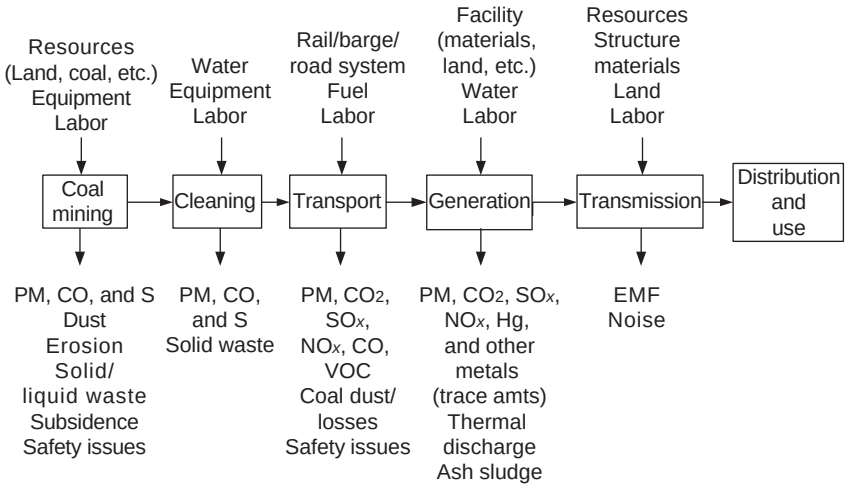


FIGURE 1 Example of life cycle of electricity (from coal).

with reservoir), nuclear energy, and wind power. They concluded that, although many studies have demonstrated technological innovation that promises to reduce emissions in one phase of the life cycle, it often increases emissions in other phases. They also suggest that a fair comparison requires including the backup power needed to achieve the same level of reliability. The environmental issues associated with management and storage of spent nuclear fuel as well as the potential for catastrophic events is only mentioned briefly while a considerable portion of the paper discusses the excellent performance of nuclear energy. This paper also suggests that estimates of land use should include the land that is damaged due to climate change and acid rain.

4. COAL

Figure 1 shows the fuel cycle of coal. We focus on comprehensive studies conducted by research teams at the Oak Ridge National Laboratory-Resources for the Future (ORNL-RFF), Argonne National Laboratory (ANL), and National Renewable Energy Laboratory (NREL). The ORNL-RFF study focused on valuing the externalities, NREL focused on a complete inventory of discharges of the main processes (including the construction and decommissioning of the plant), and ANL focused on the design of the plant and the quantification of impact on the environment.

The 1994 ORNL-RFF study, the most detailed of the three, was part of a series that also looked at natural gas, hydro, biomass, oil, and nuclear. Each study examined one plant in the southeast United States and one in the southwest United States. The plants were selected on the basis of easily available data, but are not representative of most plants in the United States. The 1999 NREL study assessed the environmental impacts of three pulverized coal boiler systems: a currently operating plant, a plant that meets the New Source Performance Standards

(NSPS), and a low-emission boiler system (LEBS) based on the design of a potential future plant. The 2001 ANL study focused on an advanced technology, an integrated gasification combined-cycle (IGCC) plant design based on the Shell entrained-flow gasifier; this plant was used to produce either electricity or both electricity and hydrogen. The assumptions made for the three studies are summarized in Table II.

Environmental standards for coal combustion have tightened considerably since 1970. For example, the current New Source Performance Standards (NSPS) are 0.60, 0.60, and 0.03 lb per million BTU of energy from coal for NO_x, SO_2, and particulate matter, respectively. A low-emissions boiler has emissions standards one-sixth of the NSPS for NO_x and SO_2 and one-third of the NSPS for particulate matter. Similarly, standards for underground mine safety, acid mine drainage, and restoration of strip mined land have become more stringent over time.

4.1 Mining

Problems from coal mining include injuries and chronic lung disease in miners, acid mine drainage, unrestored mining sites, dumping hilltops into neighboring valleys, air pollution, erosion, mining waste, subsidence, and disruption in underground water flows and storage. The environmental aspects of mining have received little analysis.

U.S. coals vary in moisture content (2 to 40%), sulfur content (0.2 to 8%) and ash content (5 to 40%). The energy content varies from lignite to subbituminous to bituminous coal. The ORNL-RFF study looks at two levels of sulfur content (0.7% and 2.1%). The NREL and ANL studies focused on high-sulfur coal (Illinois #6), providing no information about the range of coals currently used in the United States.

The ORNL-RFF study assumed that the coal came from surface mining. The NREL study examined the impacts of underground (longwall) and surface (strip) mining but concluded that the results were not significantly different. The ANL study assumed underground mining but did not conduct a full analysis of the impacts of mining.

4.2 Transportation

Coal is transported by rail, barge, truck, and pipeline. The environmental impacts and injuries vary considerably across modes. The ORNL-RFF estimates for rail injuries and deaths are shown in Table III. Transporting coal causes nearly 400 deaths annually in the United States, almost all to the general public.

4.3 Electricity Generation

Resources required to build the power plant were considered, but some studies gave little detail.

The ORNL-RFF study examined steel, concrete, land, and water. The coal feed requirements were taken into account in order to estimate the mining impacts (e.g., accidents), transportation impacts (e.g., road damage), and generation impacts (e.g., NO_x emissions). The study did not consider the resources or environmental issues associated with opening a new mine or constructing a new transport system, since they assumed that coal came from existing mines and transport systems. NREL evaluated concrete, steel, aluminum and iron, analyzing the resources used in the mining and transportation phases of the fuel cycle, including the transportation vehicles, land reclamation activities, and mining equipment.

4.4 Transmission of Electricity

Transmission has been neglected in most studies, based on the argument that all power plants require transmission. However, energy resources are located in different places, requiring different amounts of transmission. A few studies evaluate the environmental impacts of transmitting electricity. The impacts appear to be small, unless there are important health consequences from exposure to 60 hertz electromagnetic fields. In 1996, Knoepful developed and tested a framework for comparing the environmental impacts associated with various methods of shipping energy in Europe. This study concluded that for coal, generating electricity early in the fuel cycle and shipping the energy through high-voltage transmission lines can lead to significant impact reductions when compared to coal transport by barge and train. The results for oil and gas were not as clearly beneficial but have potential for environmental improvements.

4.5 Environmental Discharges and Impacts

The studies considered a wide range of discharges and impacts, from greenhouse gas discharges to water pollution discharges, and from human health impacts to recreational fishing impacts.

TABLE II

Summary of Assumptions for Three Coal Studies

	ORNL-REF SE and SW Ref	NREL			ANL	
		National average	NSPS	LEBS	Base	Co-product
Date of study	1994		1999		2001	
Plant size (MWe)	500	360	425	404	413	76/423
Technology	PC	PC	PC	PC	IGCC	H2/CC
Efficiency	35%	32%	35%	42%		
Capacity	75%	60%	60%	60%		
Type of plant	Based on two reference sites	National average	Average attaining NSPS	LEBS	Design simulation	Design simulation
Other plant details	Wet lime/limestone scrubber (90% efficient), electrostatic precipitator (99.2% efficient), Low nox burners, can meet NSPS	Baghouse filter, FGC system, heat recovery steam generator, steam turbine	Same as national average except higher FGC efficiencies	Low NO_x system with advanced burners, air staging, wet ash slagging system, enhanced FGC (CuO)	Conventional IGCC plant releasing CO_2 by combustion of the synthesis gas in a gas turbine	Production of electricity and H2 as energy carriers, 90% of CO_2 is recovered for disposal in geological storage
Coal type	Kentucky no. 9 and Navajo	Illinois no. 6	Illinois no. 6	Illinois no. 6	Illinois no. 6	Illinois no. 6
Extraction and processing	Surface mining only (strip and contour)	Raw material extraction, equipment manufacture, coal mining (surface-strip and underground mining-longwall), surface coal mining reclamation requirements, coal preparation and cleaning–jig washing (electricity and water required, refuse landfilled)			Underground mining, coarse cleaning at mine mouth (assumed refuse returned to mine)	
Transportation	Rail and truck	Transport and chemicals/materials to the mine site and power plant as well as the transport of coal—railcar, railcar, and barge, mine mouth			By rail only—coal losses from train considered, diesel fuel use, open rail cars loaded with crushed coal, did not include manufacturing diesel fuel or manufacture and maintenance of railcars	
Generation	Operation only	Operation, construction, and demolition, ash treated and landfilled or alternate use			Operation of power plant as well as construction and demolition of power plant, CO_2 and hydrogen pipelines	
Transmission	Not considered	Not considered			Not considered	
Data sources	Coal Technology—DOE 1988 coal technology, water and solid emissions—Meridian 1989, EPRI 1989, DOE 1989 material requirement emissions—Meridian 1989	Power generation—TEAM database, FETC (Utility Data Institute, 1993, 1993, and 1996; Keeth *et al*., 1983; U.S. EPA., 1985; Schultx, and Kitto, 1992; Combustion Engineering, 1981; Ladino *et al*., 1982; Walas, 1988; Collins, 1994; Larinoff, 1994; and Darguzas *et al*., 1997); plant details—average and NSPS plant—(Utility Data Institute, 1996), TEAM LEBS Plant—Ruth, 1997, Surface and Underground Mining—Bureau of Mines—Sidney *et al*., 1976 (surface), and Duda and Hemingway, 1976 (underground)			LCA analysis; LCAdvantage, process design; ASPEN simulation	

NSPS, New Source Performance Standards; LEBS, low emission boilers; PC, pulverized coal; IGCC, integrated coal gasification combined cycle; H_2/CC, electricity and hydrogen as energy carriers with carbon capture. From ANL (2001); NREL (1999); and ORNL-RFF (1994).

TABLE III

Number of Deaths/Injuries from Southeast Reference Site/Year by Rail[a]

			Occupational				
		Public	Trans.	Maint.	Other	Total	Total
Injuries	Low	0.7	2.1	3.7	0.37	6.2	6.9
	Mid	0.83	2.5	4.2	0.37	7	7.8
	High	0.97	2.8	4.6	0.43	7.8	8.7
Fatalities	Low	0.28	0.0052	0.0047	0.0013	0.011	0.29
	Mid	0.34	0.0062	0.0053	0.00155	0.013	0.35
	High	0.39	0.0071	0.0058	0.0017	0.015	0.41

[a] From ORNL-RFF (1994).

TABLE IV

Comparison of Emissions from Coal Studies[a]

	ORNL-RFF		NREL			ANL		
Emissions (ton/GWh)	Southeast ref site	Southwest ref site	Average	NSPS	LEBS	Base case	Coproduct case	Pacca and Horvath
CO_2	1100	1200	1100	1000	820	940	120	900
SO_2	1.8	0.87	7.4	2.8	0.79			
NO_x	3.0	2.3	3.7	2.6	0.60			
Particulate matter (PM)	1.6	1.6	10	11	0.12			
CO	0.27	0.27	0.23	0.28	0.21			
HC	0.099	0.13	0.23	0.22	0.21			
Trace metals								
As ($\times 10^{-4}$)	2.0	2.0	0.54					
Cd ($\times 10^{-6}$)	3.0	3.0	4.5					
Mn ($\times 10^{-4}$)	1.3	1.3	0.47					
Pb ($\times 10^{-5}$)	9.0	9.0	3.3					
Se ($\times 10^{-4}$)	0.50	0.50	4.5					

[a] Adapted from ANL (2001); NREL (1999); ORNL-RFF (1994); Pacca and Horvath (2002).

The ORNL-RFF study estimated monetary values for some impacts and characterized rather than measured or calculated additional impacts. NREL distinguished between human health and ecological health when discussing the impacts associated with the production of electricity. There was no attempt to estimate the dollar loss or magnitude of these impacts. ANL considered (1) natural environment impacts, such as acidification, eutrophication, smog, global climate changes, and ecotoxicological impacts (aquatic and terrestrial toxicity); (2) human health impacts, such as toxicological impacts, PM_{10} inhalation effects, and carcinogenic impacts; and (3) natural resources impacts (the depletion of fuels and consumption of water).

ANL made an attempt to calculate the relative impact of the plant designs studied. Table IV shows that modern technology can lower the adverse discharges from a coal-fired generation plant, due both to greater efficiency and better processes. SO_2 and NO_x emissions can be lowered by almost a factor of ten and particulate matter emissions by a factor of 100. Since little or no attention has been given to CO and HC, the new technologies have little effect.

4.6 Other Considerations

4.6.1 Abandoned Coal Mine Problems

Abandoned coal mines are ubiquitous, as shown in the Office of Surface Mining's (OSM) map (Fig. 2).

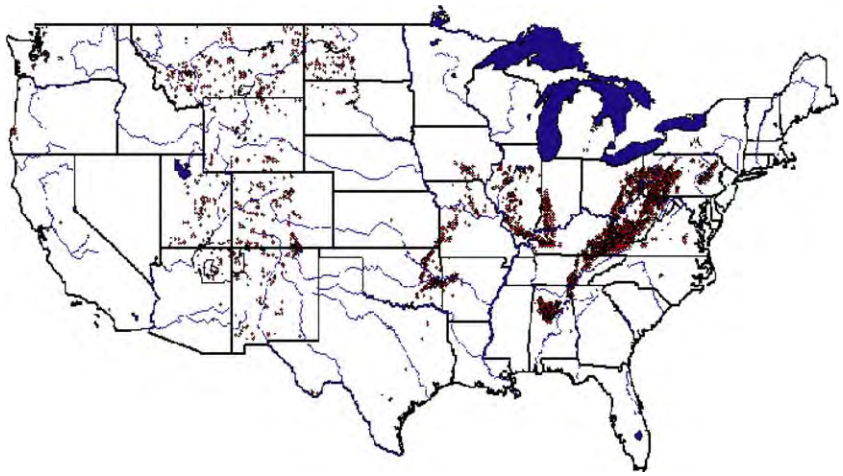


FIGURE 2 Locations of abandoned mine land problems eligible for funding from the U.S. Office of Surface Mining (OSM). From http://www.osmre.gov/aml/intro/zintro2.htm.

TABLE V

Office of Surface Mining—Unreclaimed Public Health and Safety Coal Related Problems by Problem Type[a]

Problem description	Measured as	Units unreclaimed	Cost of reclaiming problems (000$)
Clogged streams	Miles	1,700	48,000
Clogged stream lands	Acres	25,000	190,000
Dangerous highwalls	Feet	4,200,000	600,000
Dangerous impoundments	Count	760	14,000
Dangerous piles and embankments	Acres	8,800	250,000
Dangerous slides	Acres	4,700	72,000
Gases: Hazardous/explosive	Count	94	2,800
Hazardous equipment and facilities	Count	2,600	26,000
Hazardous water body	Count	970	54,000
Industrial/residential waste	Acres	400	10,000
Portals	Count	5,600	21,000
Polluted water: Agriculture and industry	Count	540	100,000
Polluted water: Human consumption	Count	4,200	3,70,000
Subsidence	Acres	8,500	480,000
Surface burning	Acres	440	17,000
Underground mine fire	Acres	4,200	860,000
Vertical opening	Count	2,400	37,000
Total			6,500,000

[a] From OSM—Abandoned Mine Land Inventory—Current Database.

Funding for site restoration is small. For example, Pennsylvania is estimated to need $15 billion worth of restoration work, while the tax collected from coal mining operations available to Pennsylvania is only about $21 million per year. The "high priority" problems monitored by the OSM are listed in Table V.

4.6.2 Acid Mine Drainage

Acid mine drainage is the main cause of polluted water in the United States with devastating effects on biological activity in many streams. In 1995, 2400 of 54,000 miles of streams in Pennsylvania were polluted by acid mine drainage from old mining

operations. The summary data from OSM show that there are still a total of 4688 miles of waterways that are affected by acid mine drainage; the cost of reclaiming those waterways is approximately $3.8 billion. Other organizations estimate that the total costs to reclaim waterways from acid mine drainage are much higher ($5 billion in Pennsylvania alone). While measures have been put in place to minimize the effects of acid mine drainage, it still occurs in abandoned mines as well a small percentage of new mines.

4.6.3 Coal Mine Fires

Coal mines contain hazardous and explosive gases, and there is a potential for long-lasting fires. The OSM estimates that there are currently 4163 acres burning, including 94 sites where hazardous or explosive gas is being emitted from underground mine fires, which can have an effect on humans in the vicinity of the site. The estimated cost of extinguishing these fires is $860 million. The most extreme case in the United States is in Centralia, Pennsylvania, where an underground fire has been burning for longer than 30 years. Attempts to extinguish it have failed, leading the government to buy all the property at a cost of $42 million as well as costs associated with the attempts to fight the fire.

5. NATURAL GAS

Nearly all of the new generation in the United States in the past 5 years has been fueled by natural gas. NREL estimates that 22% of energy consumed in the United States is natural gas; the Department of Energy predicts that by 2020, 33% of the electricity will be generated from natural gas. The life cycle stages of natural gas are construction and decommissioning of the power plant, construction of the natural gas pipeline, natural gas production and distribution, ammonia production and distribution, NO_x removal, and power plant operation.

The ORNL-RFF study concluded that the major sources of damage from the gas fuel cycle are emissions of particulate matter and ozone. The gas fuel cycle has lower net emissions of CO_2 than other fossil fuel cycles but still has greater discharges than renewable energy sources. Other important environmental consequences of this fuel cycle are the consumption of water, land impacts from drilling and exploration, the potential for pipeline fires and explosions.

TABLE VI

Comparison of Three Natural Gas Studies

	Natural gas		
	ORNL-RFF	NREL	Pacca and Horvath
Technology	CCGT	NGCC	CCGT
Capacity (MW)	500	505	1000
Efficiency	45-47%		
Output	3.2 TWh		5.5 TWh
	Emissions (ton/GWh)		
CO_2	640	440	500
SO_2	neg.	0.32	
NO_x	0.50	0.57	
Particulate matter (PM)	9.9	0.13	

From NREL (1999); ORNL-RFF (1994); and Pacca and Horvath (2002).

A summary of the results obtained from LCAs of natural gas are shown in Table VI. There is good agreement concerning CO_2 emissions. The studies don't make common assumptions about the plants.

6. HYDRO

Until recently, hydroelectric power was considered the most environmentally benign form of electricity. In recent years, however, many people have concluded that hydro may be one of the worst fuel cycles in terms of environmental damage. Most major waterways that have the potential to be used as hydroelectric generators in the United States have already been developed. However, projects involving retrofitting current dams as well as smaller scale diversion structures are possible. The ORNL-RFF study states that more than 77,000 dams in the United States have the potential for hydroelectric power development. The discharges from the hydroelectric fuel cycle are shown in Table VII. These values represent the discharges experienced in producing the materials to construct the hydroelectric power plant.

7. OIL

Since the mid-1970s, the share of oil has declined. In 1999, petroleum generated only 3.3% of U.S.

TABLE VII

Comparison of Emissions from Two Hydro Studies

Emissions (ton/GWh)	Hydroelectric	
	ORNL-RFF	Pacca and Horvath
	2° Emissions from manufacture	Upgrade of existing dam (5.5 TWh/yr)
CO_2	8.7	5
SO_2	0.027	
NO_x	0.074	
PM	0.0052	

From ORNL-RFF (1994) and Pacca and Horvath (2002).

electricity. The ORNL-RFF study investigated a hypothetical plant to be built in 1990, assuming effective pollution abatement technologies.

8. NUCLEAR

In 1987, 107 nuclear power plant facilities were operating in the United States. No new permits to construct nuclear power plants have been issued in three decades, and no applications are imminent in the United States. A major advantage to nuclear power is that generation does not release the pollutants that are a problem with fossil fuels. However, there are major concerns about the treatment and risks associated with the generation and storage of radioactive wastes and the possibility of a catastrophic release of radioactive material, as occurred at Chernobyl. Since the 1970s, little assessment has been done in the United States, in contrast to studies in other nations. An ORNL-RFF study investigates two hypothetical plants using pressurized water reactors. This is not state-of-the-art technology but reflects typical plants in the United States today. However, the study was conducted as if a new plant were being built with this technology.

9. BIOMASS

Biomass is a renewable fuel that could substitute for much of the coal currently used. In 1999, Rafaschiari *et al.* compared an integrated gasification combined cycle plant fired by dedicated energy crops (poplar short rotation forestry) to a conventional power plant. They found that biomass had less environmental impact than coal in almost all of the ecoindicators and normalized effects considered in this study. The most significant environmental effects from this fuel cycle are caused by the use of chemicals and fertilizers.

Several recent studies have concluded that the use of crops to generate electricity is preferred to their use as transport fuels from both an ecological and socioeconomical criteria. However, financial incentives are required to make these crops competitive fuels for electricity generation.

Some studies focus on the greenhouse gas emissions of this fuel cycle while others insist that considering ecological and socioeconomic sustainability of biomass crops is essential to gaining a clear perspective of this fuel cycle. A comparison of co-combustion between different biofuels and hard coal for electricity production from hard coal alone has also been conducted.

The ORNL-RFF study investigated two hypothetical plants. This study concluded that there are significant differences in damages, and thus externalities, among different sites (for example, benefits from erosion reduction differ by a factor of three) and for different biomass technologies. The use of advanced biomass conversion technologies could reduce NO_x emissions significantly compared to conventional wood burners. This biomass fuel cycle has near-zero emissions of CO_2. This study concluded that a biomass fuel cycle has less impact on global climate change than the fossil fuel cycles.

10. WIND

Wind power has been used by humans for thousands of years. It has been used to generate electricity (on a small scale) since the early to mid 1900s. By 1995, it was estimated that there were 17,000 commercial wind turbines in the United States. The main advantage of wind is that the generation phase does not emit environmentally harmful pollutants. However, there are several major issues to consider. In evaluating wind, it is important to account for the environmental impacts associated with the manufacture of the wind turbines as well as the land used for the wind turbine. The amount of energy that can be extracted from wind goes up with the cube of wind speed, making location important. Good wind locations often are located far from electricity demand centers and therefore require additional transmission to deliver the electricity. Finally, due to the intermittency of this technology, backup

power is required to supply electricity when the wind is not blowing.

In 2002, Lenzen and Munksgaard reviewed studies concerning the environmental impact of wind turbines. They suggest using an input-output based hybrid technique in order to minimize the uncertainties as well as using a standardized method of assessment. In 2000, McGowan *et al.* conducted a general review of the technology, design, trends, and their subsequent environmental impact.

11. SOLAR

The sun is the earth's greatest source of energy and the source of most renewable energy. The sun radiates energy (approx. 2.1×10^{15} kWh per day on earth) in the form of electromagnetic radiation. Although biomass, hydro, and waves are indirect forms of solar energy, solar thermal or photovoltaic are the direct ways of using solar energy to generate electricity. Solar thermal technology uses the radiation directly to heat water. Photovoltaic technology converts the sun's rays directly to electrical energy. One of the advantages of solar radiation is that the conversion of electromagnetic radiation to electricity occurs without environmentally harmful discharges. However, other stages of the fuel cycle do contribute to environmental damage. One of the major environmental issues with this fuel cycle is the manufacture and disposal of solar cells and other equipment required to capture the radiation before it is transformed into electricity.

The renewable technologies, except for wind, are not used widely anywhere in the world because of their cost. Because the technology will not be disseminated widely until its costs are competitive, we assess promising prototypes. Many studies evaluate the environmental implications of fuel cycles in terms of their contribution to global warming. This is only one aspect of the life cycle and may mislead readers.

In 1996, Mirasgedis *et al.* estimated the level of atmospheric pollutants emitted during the manufacturing process of solar water heating systems. They found that the LCA gaseous pollutant emissions from the production of solar water heating systems are much less than that resulting from generating electricity through conventional means in Greece.

In 2001, Greijer *et al.* evaluated the environmental life cycle implications of a nanocrystalline dye sensitized solar cell and compared this to a natural gas combined cycle power plant. This evaluation focused on CO_2 and SO_2 emissions per kWh. They found that the gas power plant emitted about 10 times the CO_2 emissions of the solar cell. The largest impact from the solar cell was the process energy required to produce it.

In 2000, Koner *et al.* looked at a photovoltaic generator and used the life cycle energy cost analysis to compare it to fuel generators (kerosene and diesel). They found that, at current market prices, the photovoltaic generators were comparable or less expensive than the fuel generators.

The toxic and flammable/explosive gases of concern in photovoltaic power systems are silane, phosphine, and germane as well as cadmium. Recycling the cell materials is possible, but the environmental consequences must be considered. Depletion of rare materials is also a concern. Energy use in the manufacturing stage is the largest contributor to emissions. An LCA of solar systems should consider the system integration aspects such as energy storage and the treatment of imports and exports.

12. CONCLUSIONS

A substantial amount of research has explored the life cycle implications of generating electricity using a range of fuels and technologies. This work has developed the framework and life cycle method as well as the implications of each fuel/technology. Most of the U.S. research has focused on coal, since it is the fuel for more than half of the electricity that is generated. The early technologies for generating electricity from coal produced many deaths and injuries from mining and transport as well as highly polluted air and water due to acid mine drainage and burning the coal. Increasingly stringent regulatory pressure has lowered both the injuries and environmental pollution from the life cycle of using coal to generate electricity. Large remaining problems are underground mining, transport of the coal, and CO_2 emissions from burning the coal. Although the technology exists to solve the remaining environmental problems, little of that technology has been implemented. Additional incentives will be needed to solve these problems.

Nuclear-powered turbines are perhaps the most benign fuel/technology, with only relatively small amounts of injury and environmental discharges. Concerns about disposing of radioactive waste and protecting plants against mishaps or terrorist attacks are not fully resolved. The next most benign is likely

to be natural gas. Biomass offers a solution to the CO_2 emissions problem, but this fuel is more expensive and may not be less polluting. Petroleum has been phased out of the U.S. electricity fuel market over the past quarter century; it is no longer important. On a life cycle basis, the renewable fuels have much higher environmental costs than might be suspected from an examination of a single part of the fuel cycle.

LCA has a major contribution to make in choosing among fuels and generating technologies, as well as in finding the parts of the fuel cycle of each that are most important to fix.

SEE ALSO THE FOLLOWING ARTICLES

Clean Coal Technology • Coal Mine Reclamation and Remediation • Electricity, Environmental Impacts of • Electric Power Generation: Valuation of Environmental Costs • Greenhouse Gas Emissions from Energy Systems, Comparison and Overview • Hydropower, Environmental Impact of • Life Cycle Assessment and Energy Systems • Nuclear Waste • Turbines, Gas • Wind Energy Technology, Environmental Impacts of

Further Reading

Argonne National Laboratory (ANL). (2001). "Life-Cycle Analysis of a Shell Gasification-Based Multi-Product Systems with CO_2 Recovery." The First National Conference on Carbon Sequestration. May 15–17. Washington, DC.

Carnegie Mellon Green Design Initiative, Economic Input-Output Life Cycle Assessment. (2003). http://www.eiolca.net/.

Gagnon, L., Belanger, C., and Ychiyama, Y. (2002). Life-cycle assessment of electricity generation options: The status of research in year 2001. *Energy Policy* **30**(14), 1267–1278.

Hendrickson, C., Horvath, A., Joshi, S., and Lave, L. B. (1998). Economic input-output models for environmental life cycle analysis. *Environ. Sci. Technol.* **32**(7), 184A–191A.

Holdren, J. P. (1978). Coal in context: Its role in the national energy future. *Houston Law Review* **15**(5), 1089–1109.

Knoepfel, I. H. (1996). A framework for environmental impact assessment of long-distance energy transport systems. *Energy* **21**(7/8), 693–702.

Lave, L. B., and Freeburg, L. C. (1973). Health effects of electricity generation form coal, oil and nuclear fuel. *Nuclear Safety* **14,** 409–428.

Lenzen, M., and Munksgaard J. (2002). Energy and CO_2 life-cycle analyses of wind turbines–Review and applications. *Renewable Energy* **26**(3), 339–362.

Oak Ridge National Laboratory and Resources for the Future (ORNL-RFF). (1992–1996). "Estimating Externalities of Fuel Cycles." Utility Data Institute and the Integrated Resource Planning Report. United States.

Office of Surface Mining. Abandoned Mine Land Reclamation. http://www.osmre.gov/osmaml.htm.

Pacca, S., and Horvath, A. (2002). Greenhouse gas emissions from building and operating electric power plants. *Environ. Sci. Technol. ACS* **36,** 15.

Rafaschieri, A., Rapaccini, M., and Manfrida, G. (1999). Life cycle assessment of electricity production from poplar energy crops compared with conventional fossil fuels. *Energy Conversion Manage.* **40,** 1477–1493.

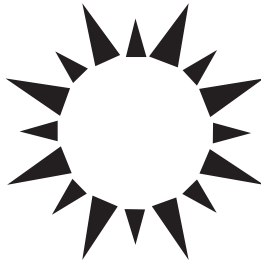
Sagan, L. A. (1974). Health costs associated with the mining, transport, and combustion of coal in the steam-electric industry. *Nature* **250,** 107–111.

Society of Environmental Toxicology and Chemistry (SETAC). http://www.setac.org/lca.html.

Spath, P. L., Mann, M. K., and Kerr, D. R. (1999). "Life Cycle Assessment of Coal-Fired Power Production." National Renewable Energy Laboratory. NREL/TP-570–27715. Golden, CO.

U. S. Environmental Protection Agency. Life Cycle Assessment–LCAccess. http://www.epa.gov/ORD/NRMRL/lcaccess/index.htm.

Life Cycle Assessment and Energy Systems



EVERT NIEUWLAAR
Copernicus Institute, Utrecht University
Utrecht, The Netherlands

1. **Definition and Role of Life Cycle Assessment**
2. **Goal Definition and Scoping**
3. **Inventory Analysis**
4. **Impact Assessment**
5. **Interpretation**
6. **Energy Issues in Life Cycle Assessment**
7. **Life Cycle Assessment of Energy Systems**

Glossary

functional unit Quantified performance of a product system for use as a reference unit in a life cycle assessment study.

life cycle Consecutive and interlinked stages of a product system, from raw material acquisition or generation of natural resources to the final disposal.

life cycle assessment (LCA) Compilation and evaluation of the inputs, outputs, and potential environmental impacts of a product system throughout its life cycle.

life cycle impact assessment (LCIA) Phase of life cycle assessment aimed at understanding and evaluating the magnitude and significance of the potential environmental impacts of a product system.

life cycle interpretation Phase of life cycle assessment in which the findings of the inventory analysis, the findings of the impact assessment, or both are combined, consistent with the defined goal and scope, to reach conclusions and recommendations.

life cycle inventory (LCI) analysis Phase of life cycle assessment involving the compilation and quantification of inputs and outputs for a given product system throughout its life cycle.

The environmental impact of energy systems has a diverse origin. The combustion of fuels leads directly to emissions and potential environmental harm, whereas the use of electricity does not lead directly to environmental impacts. However, the production of electricity from fuels leads to emissions of, for example, carbon dioxide and nitrogen oxides, which can lead to climate change and acidification, respectively. If wind turbines or photovoltaic cells produce the electricity, these emissions are avoided. However, the material-intensive production of wind turbines and photovoltaic cells is associated with environmental releases. Therefore, assessing the environmental impacts of energy carriers involves not only the process of using the energy carrier but also all related processes: from extracting primary energy from nature, its conversion into secondary energy carriers, its use, and any processing of waste flows at the end. Environmental life cycle assessment is the "cradle-to-grave" approach for assessing the environmental impacts of products such as electricity and petroleum. The cradle-to-grave approach involves all steps between extracting materials and fuels from the environment until the point where all materials are returned to the environment. The methodology of life cycle assessment and its application to energy systems are the subject of this article.

1. DEFINITION AND ROLE OF LIFE CYCLE ASSESSMENT

1.1 Definition and Characteristics

A life cycle refers to the life span of a product, from resource extraction, to manufacture, to use, to final disposal. The cycle aspect reflects that materials extracted from the environment are followed until they are ultimately returned to the environment.

Life cycle assessment (LCA) refers to the analysis and evaluation of product life cycles. In general, the assessment is limited to environmental issues only,

 © 2004 Elsevier Inc. All rights reserved.

although LCA could also imply the assessment of other issues (e.g., social, economic). Therefore, LCA is shorthand for environmental life cycle assessment. Economic tools such as life cycle costing do exist but are not considered in this context as part of LCA.

LCA addresses environmental aspects of products—actually product systems, as is explained later. This is a major distinction of LCA from other environmental tools such as environmental impact assessment (EIA) and substance flow analysis (SFA). The product-oriented approach in LCA means that environmental impacts related to a product are examined. Products can be physical items such as pencils, potatoes, and cars. The term "product" also includes services such as transport and health care. Actually, it is not the physical product or service that has the focus in an LCA; it is the function that is fulfilled by the product that is central. Products are provided by product systems that fulfill a certain function. For example, milk cartons and milk bottles are provided by packaging systems that provide the packaging function. For energy systems, the products are delivered energy (e.g., in the form of petroleum, electricity, or heat) delivered by energy supply systems from primary energy carriers. Ultimately, the delivered energy is used to provide energy functions (e.g., lighting, mechanical work, production of materials).

Another important characteristic of LCA is that it addresses complete chains. In principle, all processes from resource extraction through production, consumption, and ultimately waste management are included in the assessment (cradle-to-grave analysis). In most cases, this means that the activities in the life cycle are not controlled by a single company or consumer and that companies involving themselves with LCA must face the fact that the environmental profile of their products also depends on the environmental performance of other companies.

LCA is not only comprehensive with respect to the number of processes included in the assessment; it is also comprehensive with respect to the number of environmental impacts. In principle, all environmental impacts related to the product system are addressed. The environmental impacts assessed include, but are not limited to, the environmental issues most often related to energy conversion systems: acidification, climate change, depletion of primary energy carriers, human toxicity, and ecotoxicity.

Where possible, quantitative results are presented in an LCA. These can range from listing all emissions amounts, aggregated values within environmental themes, to single-value indicators such as ecopoints or ecoindicators. However, single-value indicators reflecting all environmental impacts are value laden because this would involve weighting of different types of environmental damage.

Because of its nature, LCA is not location specific or time specific; that is, it cannot directly tell what the actual environmental impact of the product will be at a certain time or place. All phases of the life cycle (e.g., manufacture, use, waste management) are usually not concentrated on one location, and the time between manufacture of the product and waste management after discarding the product can be several years.

1.2 Energy and LCA

In many LCAs, energy plays an important role. The use of energy by itself is not always considered to be an environmental issue. The depletion of primary energy carriers and the strong link between energy use and certain other environmental issues (most notably acidification and climate change) have often led to the use of energy as an environmental indicator. Energy supply systems are also a major contributor to environmental themes.

1.3 Methodology Outline

LCA methodology was developed during the 1990s and is still under further development. These developments have led to standardization by the International Organization for Standardization (ISO) in the ISO-14040 series. In these standards, the following steps are distinguished (Fig. 1).

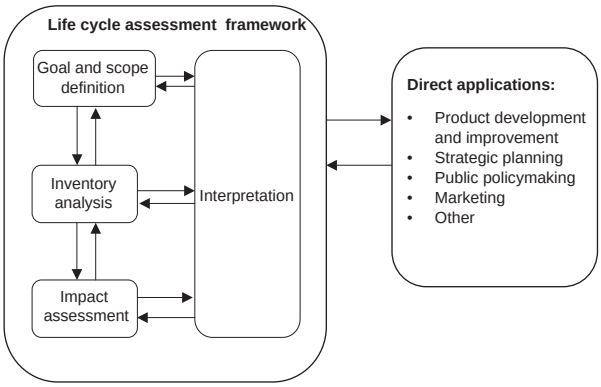


FIGURE 1 Phases of a life cycle assessment. From ISO (1997). Reproduced with the permission of the International Organization for Standardization, ISO. These standards can be obtained from any ISO member and from www.iso.org. Copyright remains with ISO.

1. *Goal definition and scoping*. The application and type of LCA are described, the product systems to be evaluated are defined, and the geographical and temporal scope are defined as well. This step also includes the definition of the functional unit, which will act as the reference for the subsequent steps.
2. *Inventory analysis*. The environmental releases (e.g., emissions, resource extractions) for the product systems and the functional unit are determined.
3. *Impact assessment*. The (potential) environmental impacts caused by the environmental releases analyzed in the previous step are determined.
4. *Interpretation*. The results of the inventory analysis and impact assessment are discussed, conclusions are drawn, and recommendations are made.

These four steps in an LCA are further discussed in subsequent sections of this article.

2. GOAL DEFINITION AND SCOPING

In the goal definition and scoping phase of LCA, the purpose and method of including environmental information in decision making are defined. The primary goal of an LCA is the selection of the best product alternative in terms of having the least effect on human health and the environment. This almost always means that comparisons are made: the product is compared with a competing product, or improved products are compared with the original or a design target. Also, the LCA can be used to identify life cycle stages where opportunities for reduction in resource use and emissions might be found. LCA can also play a role in guiding the development of new products. More specific goals can be defined; for example, how can the product system be changed such that a specific environmental impact (e.g., acid rain) can be reduced without the introduction of adverse effects in other environmental impact categories?

The goal definition and scoping phase determines the amount, type, and quality of the data requirements and, therefore, also the time and resources needed for the LCA study. For instance, during early stages of product development, many design choices exist. Limited amounts of rough data and a focus on the most important steps may be considered sufficient in such cases, especially when the results are used internally. In the case of comparing competing products, the justification of environmental claims may require comprehensive data acquisition and, therefore, more time and resources for the LCA study.

Depending on the goal, the LCA practitioner may decide not to include all life cycle phases in the scope of the research. This may be justified in cases where such phases are known to have negligible impact or where such phases are equal for all product alternatives considered.

LCA data and results will be expressed in terms of a functional unit that describes the function of the product system studied and also gives a quantification of this function. For electric power production, the function can be defined as "to deliver electric power to the electric grid." The functional unit is then 1 kilowatt-hour (kWh) electric power delivered to the electric grid. Comparisons between product alternatives (e.g., between coal-fired power stations and gas-fired power stations) must be made based on the same functional unit for all product alternatives. This means that 1 kWh of electricity from a coal-fired power station must be functionally equivalent to 1 kWh of electricity from a gas-fired power station.

3. INVENTORY ANALYSIS

3.1 Definition

In the life cycle inventory (LCI) analysis, the life cycle is drawn up and all energy and material requirements; emissions to air, water, and soil; and other environmental releases are quantified. The steps involved in performing an LCI are the development of a flow diagram, data collection, multi-output processes, and reporting.

3.2 Development of a Flow Diagram

The flow diagram reflects all of the processes that make up the product system and the inputs and outputs to each of the processes within the product system. From the goal and scope definition, it follows which processes will be included within the system boundary. Figure 2 gives an example of a product system for LCI analysis. It shows the processes that are included in the system boundary as well as product flows to or from other systems that are not included. Two types of processes are modeled here in a more generic way: transports and energy supply. This represents the fact that in nearly all cases some form of transport is needed between processes and the fact that energy is needed in many operations. All processes are connected with flows of (intermediate)

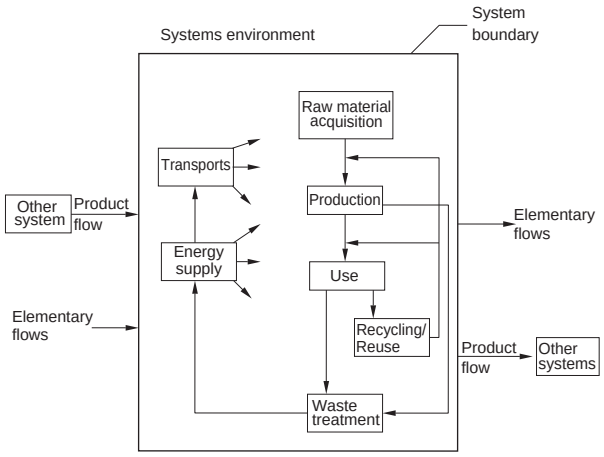


FIGURE 2 Example of a product system for life cycle inventory analysis. From ISO (1998). Reproduced with the permission of the International Organization for Standardization, ISO. These standards can be obtained from any ISO member and from www.iso.org. Copyright remains with ISO.

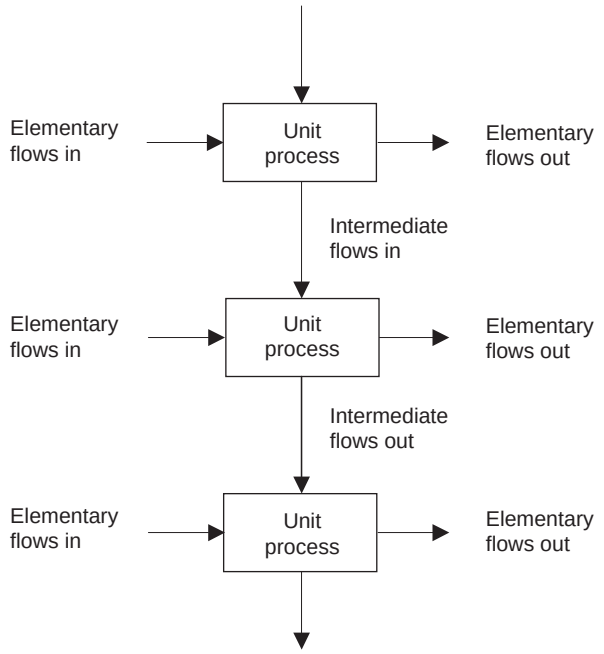


FIGURE 3 Example of a set of unit processes within a product system. From ISO (1998). Reproduced with the permission of the International Organization for Standardization, ISO. These standards can be obtained from any ISO member and from www.iso.org. Copyright remains with ISO.

products. Furthermore, elementary flows to and from the environment are shown. Elementary flows are either material or energy flows entering or leaving the system that have been drawn from the environment or discarded to the environment without previous or subsequent human transformation, respectively. In other words, elementary flows originate directly from the environment (e.g., energy and material resources, land use) or are discarded directly to the environment (e.g., emissions, heat, radiation, sound). Elementary flows originate from the processes within the system boundary, as is shown in Fig. 3.

3.3 Data Collection

Data collection is probably the most time-consuming step. From the goal definition and scoping step, it must be clear which requirements for data accuracy must be fulfilled. Most often, this will be a mix of site-specific data (obtained through measurements and calculations from measurements) and more generic, non-site-specific data that can be obtained from available databases or expert estimates.

3.4 Multi-Output Processes

In many cases, processes are encountered that do not have a single useful output. In such multi-output processes, often only one of the outputs is used in the life cycle. Other outputs are used in other life cycles that are not under investigation. The material and energy inputs and the elementary flows of these processes have to be divided (allocated) among the useful outputs in a "proper" way. Various possibilities exist for the allocation in multi-output processes. A well-known multi-output process is the combined production of heat and power (CHP). The fuel input in CHP plants is to be divided between heat and electricity. One possibility is system extension by introducing a reference technology for one of the products. In the case of heat production, the fuel input for heat is calculated as the fuel required for generating the heat in a reference boiler. The fuel input for electricity from the CHP plant is then obtained by subtracting the fuel input for heat production from the CHP plant fuel requirement. Of course, a reference electricity plant could be introduced instead of a reference boiler. Another approach is the use of a physical property of the products (e.g., mass or energy content) as an allocation factor. For CHP, the use of mass as an allocation factor is meaningless. Energy content (either on a heat/enthalpy basis or on an exergy basis) is more meaningful. A third approach is the use of economic value of the individual products as the proper allocation factor. All three approaches have been used in LCAs.

3.5 Reporting

The outcome of the LCI analysis is a list of all elementary flows to or from the environment for each product system investigated, resulting from the functional unit provided by the product systems. For the purpose of analyzing these results, they can be organized according to life cycle stage and/or media (air, water, or land). It is possible to use these results for further analysis such as comparison of product alternatives and identification of life cycle stages where significant environmental releases take place. However, these releases do not express the potential environmental impacts that may be caused. The translation of environmental releases to potential environmental impacts is the subject of the next section.

4. IMPACT ASSESSMENT

4.1 Definition

Life cycle impact assessment (LCIA) is the phase of an LCA where the evaluation takes place of the potential environmental impacts stemming from the elementary flows (environmental resources and releases) obtained in the LCI. LCIA consists of the following steps:

1. Selecting the relevant impact categories
2. *Classification:* assigning the elementary flows to the impact categories
3. *Characterization:* modeling potential impacts using conversion factors obtaining an indicator for the impact category
4. *Normalization (optional):* expressing potential impacts relative to a reference
5. *Grouping:* sorting or ranking the impact indicators
6. *Weighting:* relative weighting of impact categories; and evaluation and reporting

4.2 Impact Categories

Environmental impacts are the consequences caused by the elementary flows on human health, plants, and animals (ecological health) or by the future availability of natural resources (resource depletion). A number of impact categories are listed in the first column of Table I.

In the classification step, the LCI results are organized and combined into the impact categories selected in the previous step. An example is formed by the greenhouse gases that are classified in the global warming category (third column of Table I).

In the characterization step, characterization factors are used (where possible) in each impact category to combine all of the elementary flows to one indicator for the impact category. For greenhouse gases, the global warming potential for each greenhouse gas is used to combine all greenhouse gases into one indicator (global warming equivalent), expressed in carbon dioxide equivalents. The final two columns of Table I show the characterization factors and a description of the procedure. For some impact categories (e.g., global warming, ozone depletion), there is a consensus on acceptable characterization factors. For other impact categories, methods have been developed but further consensus regarding their application has to be established. Furthermore, characterization factors for regional environmental issues (e.g., acidification) may sometimes be defined such that they are relevant only for the region for which they were derived. Use of these factors in other regions has to be investigated.

In the (optional) normalization step, the indicator scores obtained in the characterization step are normalized by dividing by a reference value. Reference values can be the total emissions or resource uses for a given area (global, regional, or local) or the total emissions/resource uses for a given area per capita. The results do not imply a certain weighting among impact categories; they merely give an indication how much the environmental profile of the product contributes to the total environmental profile of, for example, the region.

4.3 Grouping and Weighting

In the grouping step, the indicators are sorted by characteristics such as emissions (e.g., air, water) and location (e.g., local, regional, global). Other forms of grouping, such as using a ranking system (high, low, or medium priority), also exist. In the last case, value choices regarding what is to be considered a high or low score play an important role, and transparency regarding the procedure used is essential.

The weighting (or valuation) step assigns relative values to the various impact categories based on perceived importance. Although scientific procedures may be used (e.g., multiple-criteria analysis), the weighting factors obtained are not strictly scientific. At best, they correctly reflect the preferences of the stakeholders considered in the (scientific) procedure. Because such preferences can change in time and among stakeholders, it can also be considered to use only the characterization (or normalization) results.

TABLE I

Life Cycle Impact Categories

Impact category	Scale	Relevant LCI data (i.e., classification)	Characterization factor	Description of characterization factor
Climate change	Global	Carbon dioxide (CO_2) Dinitrogenmonoxide (N_2O) Methane (CH_4) Chlorofluorocarbons (CFCs) Hydrochlorofluorocarbons (HCFCs)	Global warming potential (GWP)	Converts LCI data to carbon dioxide (CO_2) equivalents (Note: global warming potentials can be 50-, 100- or 500-year potentials)
Stratospheric ozone depletion	Global	Chlorofluorocarbons (CFCs) Hydrochlorofluorocarbons (HCFCs) Halons Methyl bromide (CH_3Br)	Ozone depletion potential (ODP)	Converts LCI data to trichlorofluoromethane (CFC-11) equivalents
Acidification	Regional local	Sulfur oxides (SO_x) Nitrogen oxides (NO_x) Hydrochloric acid (HCl) Hydrofluoric acid (HF) Ammonia (NH_3)	Acidification potential (AP)	Converts LCI data to sulfur dioxide (SO_2) equivalents
Eutrophication	Local	Phosphate (PO_4) Nitrogen monoxide (NO) Nitrogen dioxide (NO_2) Nitrates (NO_3) Ammonia (NH_3)	Eutrophication potential (EP)	Converts LCI data to phosphate (PO_4) equivalents
Photochemical smog	Local	Non-methane volatile organic compounds (NMVOC)	Photochemical ozone creation Potential (POCP)	Converts LCI data to ethylene (C_2H_4) equivalents
Ecotoxicity	Global continental	Releases to air, water, and soil	Fresh water aquatic ecotoxicity potential (FAETP) Marine aquatic ecotoxicity potential (MAETP) Fresh water sediment ecotoxicity potential (FSETP) Marine sediment ecotoxicity, potential (MSETP) Terrestrial sediment ecotoxicity potential (TETP)	Converts LCI data to 1,4-dichlorobenzene equivalent; separate factors for fresh water aquatic/sediment Marine aquatic/sediment, and terrestrial ecotoxicity (Note: different factors for global/continental scale and infinite or 20-, 100-, or 500-, year time horizon)
Human toxicity	Global continental	Releases to air, water, and soil	Human toxicity potential (HTP)	Converts LCI data to 1,4-dichlorobenzene equivalents (Note: different factors for global/continental scale and infinite or 20-, 100-, or 500-year time horizon.)
Resource depletion, abiotic	Global	Quantities of minerals/fossil fuels used	Abiotic depletion factor (ADP)	Converts LCI data to antimony equivalents
Land use	Global regional local	Land occupation	Increase of land competition	Converts LCI data to cumulative land use expressed in square meters/year

Sources. U.S. Environmental Protection Agency (2001), with corrections; Guinée (2002). See Guinée (2002) for a more complete listing, guidelines, and values for the characterization factors.

In fact, the ISO 14040 standard does not allow the use of weighting methods in comparative assertions that are to be disclosed to the public.

5. INTERPRETATION

Life cycle interpretation is the procedure during which the results of the LCI and LCIA are identified, checked, and evaluated. Life cycle interpretation consists of the following steps:

1. Identify significant issues.
2. Evaluate the completeness, sensitivity, and consistency of the data.
3. Draw conclusions and make recommendations.

Significant issues are the data elements that contribute the most to the results of the LCI and LCIA. These can be certain inventory parameters (e.g., energy use), impact categories (e.g., acidification), or certain life cycle stages (e.g., manufacturing).

In the second step, checks are performed regarding completeness, sensitivity of the significant data elements, and consistency (with regard to the goal definition and scope of the study). Such checks are required to reach consistent and reliable conclusions. In most cases, a clear answer cannot be given to the question of which product alternative is better from an environmental point of view. This can be caused by the uncertainty in the results or by differing outcomes in the impact categories that were considered relevant. The LCA process can then still give useful results in terms of better understanding of the environmental impacts of each product alternative (e.g., which impacts, where these impacts occur).

When LCA results are to be used to support local decision making, additional analysis addressing local aspects is required. For this, various forms of environmental impact assessment and risk analysis can be used.

6. ENERGY ISSUES IN LIFE CYCLE ASSESSMENT

Because energy is used in nearly every life cycle step, energy production often plays a key role in the LCA of any product system. Therefore, a consistent and accurate treatment of energy in LCA is required.

In a strict sense, energy is not a separate issue in LCA. The inputs and outputs of energy carriers are recorded individually in the inventory phase, but aggregate input/output is not. Similarly, the use of energy is not seen as an impact category during the impact assessment phase. The use of energy carriers leads to depletion of biotic and abiotic resources and can also lead to the emissions of various pollutants. The impacts in the related categories (e.g., [a]biotic depletion, climate change, acidification) are of concern in environmental assessment. On the other hand, an account of primary energy use in a product life cycle can be made relatively easy when the LCA data are available. Such data can be used to show overall energy performance and may also be used as a limited indicator of environmental performance.

The use of electricity in an LCA asks for careful treatment. Electricity is generated in various power plants within a certain country or region, each having its own environmental profile. It can be considered to take average values of energy use and emissions during a specific year by averaging over all power plants in the region. However, this might not reflect what is actually happening. For instance, if incandescent lamps are replaced by compact fluorescent lamps, a marginal approach is more adequate than an average approach. In the short-term marginal approach, the electricity savings caused by the substitution of lamps do not stem from the average of all plants (e.g., based mainly on coal) but rather stem from the plant that will not produce the electricity anymore. This could be an efficient natural gas-fired power plant or an older, less efficient power plant that can now be taken out of operation. For product choices that reduce electricity use in the future, present average and short-term marginal values may be inadequate. In such cases, long-term marginal values are needed. In many regions of the world, the dominant technology for new electric power plants is the natural gas-fired combined cycle power plant. Environmental data from this power plant are more adequate than present or future averages.

7. LIFE CYCLE ASSESSMENT OF ENERGY SYSTEMS

LCA methods can be used to analyze energy systems. In principle, the scope for such analyses can range from the analysis of individual energy production technologies to, ultimately, the analysis of complete national energy systems.

A convenient functional unit for energy supply systems is an amount of energy produced. For electric power systems, the kilowatt-hour of electricity produced is an example. However, an account

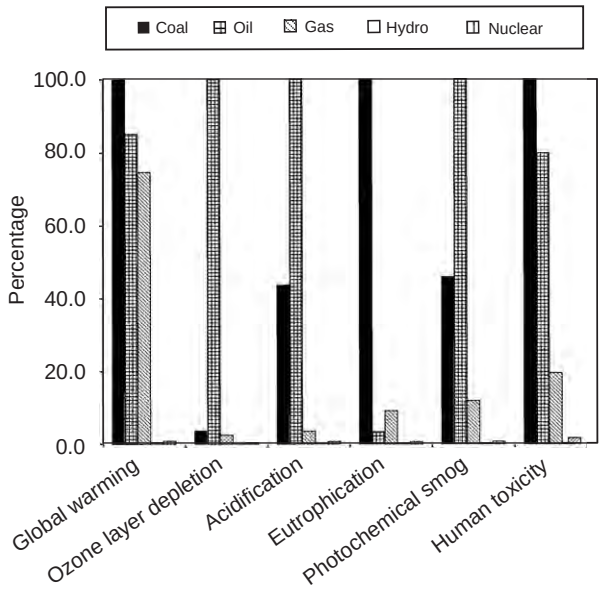


FIGURE 4 Comparison of various electricity generation systems with respect to various impact categories. For every impact class, the most polluting plant type is set to 100% (no normalization is applied here). From inventory data from Frischknecht [1996]; impact assessment characterization factors from Guinée [2002].

must be made of various factors such as operation times (load factor), time patterns in the supply of energy (especially for solar and wind energy), and capacities. These factors make that kilowatt-hour of electricity produced from one source (e.g., coal, natural gas, wind, solar, nuclear) not fully comparable to a kilowatt-hour produced from another source. So long as the total power generation mix satisfies certain demands regarding, for example, reliability, comparisons on a kilowatt-hour basis can be made. When high levels of intermittent supply sources (e.g., solar, wind) require the installation of storage options, these should be included in the LCA.

For fossil fuel-based energy systems, important environmental impacts will result from the energy conversion processes (mostly combustion) and the waste and flue gas treatment directly related to these conversion processes. In cases such as nuclear power, solar power, and wind power, the environmental profile of the materials used will largely determine the total environmental profile.

As an example, Fig. 4 shows the results for typical electricity-generating systems in Europe for six impact categories. It shows that coal is most polluting in the impact categories of global warming (climate change), eutrophication, and human toxicity and that oil is most polluting with respect to ozone layer depletion, acidification, and photochemical smog. Results are sensitive to the quality of the fuels and the specific cleaning technologies used.

SEE ALSO THE FOLLOWING ARTICLES

Climate Protection and Energy Policy • Ecological Risk Assessment Applied to Energy Development • Electric Power Generation: Valuation of Environmental Costs • Green Accounting and Energy • Industrial Ecology • Life Cycle Analysis of Power Generation Systems • Multicriteria Analysis of Energy • Net Energy Analysis: Concepts and Methods

Further Reading

Frischknecht, R. (1996). "Okoinventare für Energiesysteme." Bundesamt für Energiewirtschaft, Zurich, Switzerland.

Guinée, J. B. (ed.). (2002). Handbook on Life Cycle Assessment: Operational Guide to the ISO Standards. Kluwer Academic, Dordrecht, Netherlands.

International Organization for Standardization. (1997). "Environmental Management: Life Cycle Assessment—Principles and Framework" (ISO 14040). ISO, Geneva, Switzerland.

International Organization for Standardization. (1998). "Environmental Management: Life Cycle Assessment—Goal and Scope Definition and Inventory Analysis" (ISO 14041). ISO, Geneva, Switzerland.

International Organization for Standardization. (2000). "Environmental Management: Life Cycle Assessment—Life Cycle Impact Assessment" (ISO 14042). ISO, Geneva, Switzerland.

International Organization for Standardization. (2000). "Environmental Management: Life Cycle Assessment—Life Cycle Interpretation" (ISO 14043). ISO, Geneva, Switzerland.

U.S. Environmental Protection Agency and Science Applications International Corporation. (2001). "LCAccess: LCA 101." www.epa.gov/ord/nrmrl/lcaccess/lca101.htm.

Lifestyles and Energy

KORNELIS BLOK
Utrecht University
Utrecht, The Netherlands

1. Introduction
2. What Are Lifestyles?
3. The Structure of Household Energy Requirement
4. Trends in Household Energy Requirement
5. What Can a Consumer Do?
6. How Can Consumption Patterns Be Influenced through Policies?

Glossary

consumption patterns The total of household consumption defined as the package of goods and services that a household consumes (generally measured in terms of household purchases).

direct energy requirement The primary energy required to deliver energy carriers (e.g., electricity, natural gas, and gasoline) to households.

energy intensity of household consumption The primary energy requirement for a consumption item divided by the expenditure on this item. The energy intensity can be determined for individual goods and services but also for categories or for the complete consumption of households in a country.

indirect energy requirement The primary energy required to deliver goods and services other than energy carriers to households.

lifestyles A set of basic attitudes, values, and patterns of behavior that are common to a social group. In this article, lifestyles are operationalized in terms of consumption patterns.

primary energy requirement of a product The amount of primary energy needed to deliver a product to a consumer, including the energy needed for the production, use, and disposal of the product.

Through a range of studies, a good overview exists of how consumption patterns determine the primary energy requirements of households. Income is an important determinant. Household consumption is likely to increase in monetary terms and so may the primary energy requirement because there is not much tendency toward less energy-intensive consumption patterns. Changes in consumption patterns that substantially reduce household energy requirement are possible. To date, no clear policies can be defined that could actually lead to such changes.

1. INTRODUCTION

The concept of lifestyles is not a very well defined one. In all kinds of discussions about how to solve environmental problems, it is often suggested that people especially in industrialized countries should "change their lifestyles." Before we can discuss the relation between lifestyles and energy, we must better define the concept. This article first discusses how consumption patterns of households determine energy use. It distinguishes direct energy requirement (associated with, e.g., in the form of electricity or gasoline) and indirect energy requirement (needed to produce goods and services consumed by households). Then, attention is paid to the ways households can change their demand on primary energy carriers by changing their consumption patterns. On this basis, conclusions can be drawn on the extent to which energy use can be influenced by changing our lifestyles.

2. WHAT ARE LIFESTYLES?

In many policy documents and also in the scientific literature, lifestyles are often quoted. In one of the assessment reports of the Intergovernmental Panel on Climate Change, lifestyle is defined following the sociologist Max Weber as "a set of basic attitudes, values, and patterns of behavior that are common to a social group, including patterns of consumption and anticonsumption." As a concrete example of a lifestyle change, the report mentions the change of people's preference from car to bicycle transport.

 © 2004 Elsevier Inc. All rights reserved.

The concept is sometimes used in social science, but there are also scholars that doubt the usefulness of the concept. Lifestyles are associated with social status groups. According to many scientists, in modern societies stable status groups no longer exist. The limited use of the lifestyle concept is also demonstrated by the fact that lifestyle research has not led to a uniform and widely used classification. Nevertheless, the lifestyle concept is often used in marketing circles, although classifications vary widely.

This article focuses on people in households. Households carry out a wide range of activities, and most of them are associated with the purchase of goods and services. Some of these goods are energy carriers (electricity, heating fuels, and gasoline); however, all other goods and services also require input of energy over the life cycle to produce the good or service. We therefore make the concept of lifestyles operational in terms of consumption patterns: The lifestyle is defined by the package of goods and services that households consume. The energy impact of a consumption pattern is the total amount of energy associated with the production, use, and disposal of these goods and services.

3. THE STRUCTURE OF HOUSEHOLD ENERGY REQUIREMENT

A detailed study to establish the relation between household consumption patterns and associated energy use was carried out for The Netherlands and in somewhat less detail for the main European Union member states (Fig. 1A). In another study, the induced CO_2 emissions were established (Fig. 1C). It shows that the well-known categories, such as space heating and transport, are fairly important. However, food, an often neglected category in energy studies, contributes substantially. The direct energy use of households (heating fuels, electricity, and gasoline) comprises approximately half of the total energy use. Also, data are available for India (Fig. 1B). Here, the category food is even more important. Note that total energy requirement per capita in India is approximately a factor of 10 smaller than in the European Union countries.

The main determinant for the energy impact of household consumption is income. In the 1970s, it was established through input–output analysis that the income elasticity of household energy requirement is approximately 0.7 or 0.8 (i.e., with every

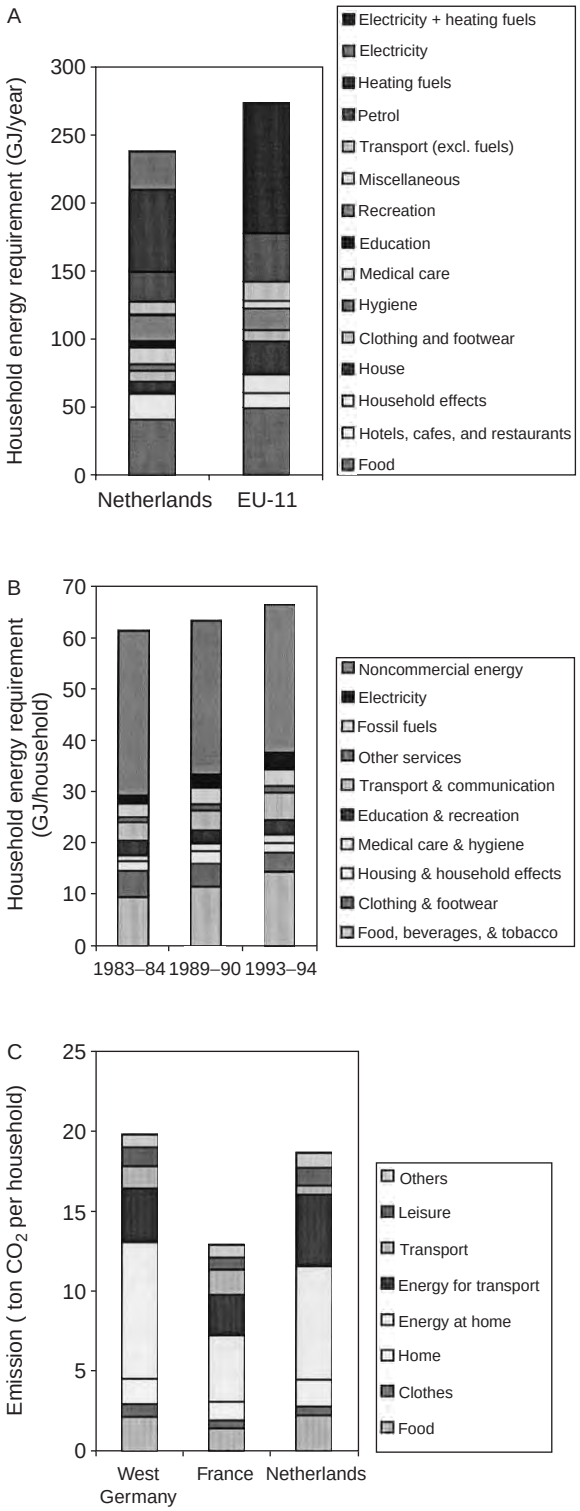


FIGURE 1 The breakdown of total energy requirement for households (A) in The Netherlands (from Vringer and Blok, 1995) and in 11 European Union member states (from Reinders *et al.*, 2003), (B) in India for various years (from Pachauri and Spreng, 2002), and (C) total induced CO_2 emissions in some countries (from Weber and Perrels, 2000).

percent of growth of household income, the household energy requirement increases by 0.7 or 0.8%). Recent studies have derived lower values (e.g., a value of 0.63 was found for The Netherlands and 0.67 was found for India, the latter actually indicated expenditure elasticity). The strong relation between income and household energy requirement is not only found within a country but also in a comparison between European countries (Fig. 2).

The composition of the consumption package changes with income (Fig. 3). Some categories only increase slowly with income (e.g., food and space heating fuels), whereas others are relatively unimportant at low incomes but increase strongly with higher incomes (e.g., electricity, transportation, and recreation).

Other factors, such as the size of the household, the age of the oldest household member, the life cycle phase of the household, the degree of urbanization, and the level of education, are less important in explaining household energy requirement. The degree to which they explain differences in household energy requirement varies between studies, but income, or expenditure, is always the most important variable. However, there are strong correlations between household income and some of the other factors (e.g., the number of persons per household and education level).

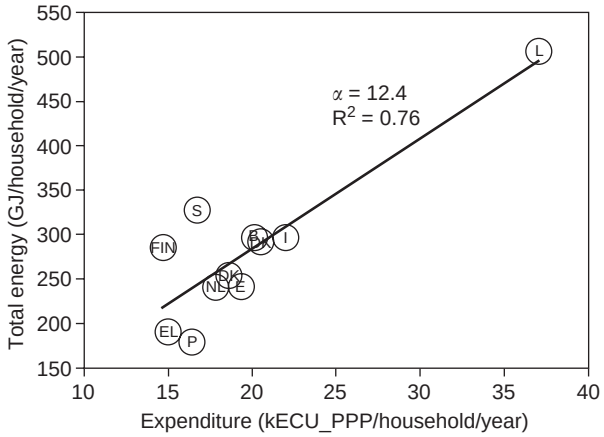


FIGURE 2 The relation between average household expenditure and average household energy requirement for European Union member states. B, Belgium; DK, Denmark; E, Spain; EL, Greece; FIN, Finland; I, Italy; L, Luxembourg; NL, Netherlands; P, Portugal; S, Sweden; UK, United Kingdom. From Reinders *et al.* (2003).

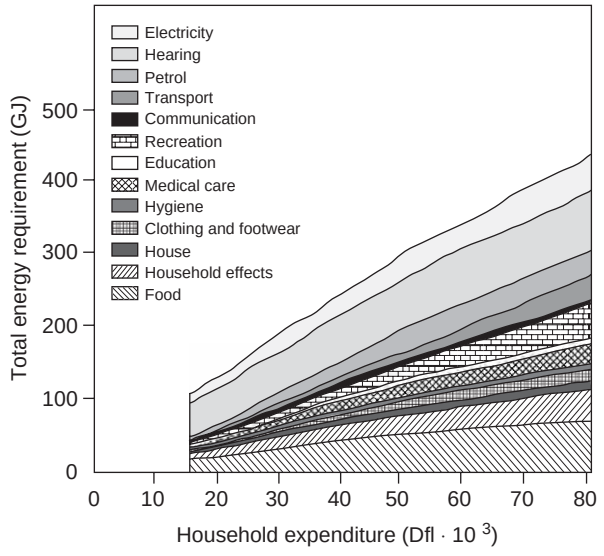


FIGURE 3 The development of total energy requirement for the various consumption categories with income. From Vringer and Blok (1995).

4. TRENDS IN HOUSEHOLD ENERGY REQUIREMENT

An important question regards how household energy requirement develops in time. We distinguish again direct and indirect energy requirement. Most information is available on direct energy requirement. Table I gives an overview of growth rates of direct parts of energy use over time. In Organization for Economic Cooperation and Development (OECD) countries, energy use for transportation and electricity services is growing at a rate comparable to that of gross domestic product, whereas

TABLE I

Overview of Growth Rates of Energy Use by Households Compared to Other Parameters[a]

	Average change during the period 1971–1995 (% per year)	Average change during the period 1985–1995 (% per year)
Household fuel use for space heating	0.0	0.8
Household electricity use	3.7	2.7
Fuel use for transportation	2.4	2.6
For comparison		
Population	1.5	1.6
GDP (ppp corrected)	0.9	0.8
Total energy use	2.7	2.4

[a] *Source*. "IEA/OECD Energy Balances." International Energy Agency, Paris, France.

energy use for space heating is lagging. There are several relevant factors.

For space heating, the development is the result of several factors: an increase in population (0.8% per year) and a decrease in the number of persons per dwelling (0.8–1.2% per year). These factors combined lead to an increase in the number of dwellings. In addition, an increase in the average size of the dwelling can be observed in most countries (0–0.8% per year).

For electric appliances, the increasing number of households plays a role. In addition, there is an increase in appliance ownership per household (typically by 1–3% per year during the past few decades). Only for a few large appliances is there a saturation of ownership rate. Apart from the ownership rate, other aspects change. For instance, there is a tendency to purchase larger refrigerators and to wash clothes more often.

For passenger cars, an increase in car ownership has occurred in all industrialized countries in the past decades; relatedly, total car mileage has increased. There is also an increase in car size and performance. Other transportation modes only contribute modestly to total passenger transportation performance. There are no substantial shifts between the various modes, except for the strong growth of the most energy-intensive mode—air transport.

All these factors can be considered as lifestyle factors, although some are strongly related to growth of population and economy. Most of the trends are likely to continue. All lead to an increase in primary energy requirement.

Adoption of improved technology (which can also be considered lifestyle related but is not discussed here in detail) partly or completely offsets the other tendencies. In the case of space heating, in most OECD countries the substantial effort to insulate buildings and improve boilers has led to a net decrease in energy use for this application. For the other areas, the energy efficiency efforts definitely had effects but were too dispersed to offset the growth drivers completely.

Also, in the case of the indirect energy requirement, there is an ongoing increase. For The Netherlands, an increase of 2.2% per year was observed during the period 1948–1988. Growth was smallest in the categories of food and clothes and highest in the categories of leisure, the house, and household effects. India, had a comparable growth rate, but especially food and transport and communication are growing.

We may ask whether the growth in household energy requirement is higher or lower than the growth in household expenditures. To answer this question, the concept of energy intensity of household consumption is introduced. This is defined by the energy requirement for a consumption item divided by the expenditure on this item. There are substantial differences in energy intensity of various consumer goods. A typical value for the average energy intensity of consumption is 7 MJ/Euro if direct energy use is excluded; if direct energy use is included, this increases to 12 MJ/Euro. Some items that have high energy intensities (20–30 MJ/Euro) are frozen food, eggs, sugar, refuse bags, toilet paper, aluminum foil, and flowers grown in greenhouses (the energy intensity of natural gas and electricity is more than 100 MJ/Euro; for gasoline, this value is lower if taxes are taken into account). Categories that have low energy intensities (<3 MJ/Euro) are music lessons, dance classes, maintenance and repair of electric appliances and central heating systems, use of taxis, telephone use, and expenditures on beauty parlors. The question now is did the average energy intensity change due to changes in household consumption patterns? Hardly. During a period of 50 years, the average decrease of energy intensity through changes in consumption patterns was only approximately 0.1% per year (Fig. 4). Note that the effect of energy efficiency improvements in the supplying sectors is likely to be much higher.

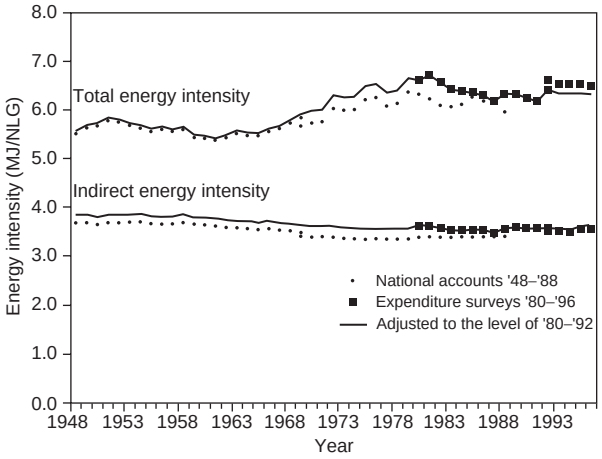


FIGURE 4 The development of energy intensity of household consumption in The Netherlands. The energy intensity of individual consumption items is kept constant at the 1990 level, so this figure only reflects the effect of changes in consumption patterns. The change in the average energy intensity of the indirect energy requirement is small. The total energy intensity shows changes caused by changes in the share of direct energy use: first an increase, mainly caused by the expansion of natural gas-based central heating systems, and later a decrease, mainly caused by insulation programs and improved boilers. 1 NLG = 0.45 Euro. From Vringer and Blok (2000).

For the indirect energy requirement, changes in consumer preferences in no way offset the increase in energy requirement related to increased income and total expenditures.

What are the mechanisms for the increase in primary energy requirement? As discussed by Røpke (1999), an important one is the increase in industrial productivity (forced by competition) that makes it possible to produce more goods, given the amount of hours worked by the total population. A variable is the total amount of time worked per person. Although there are tendencies to limit the number of working hours per week, there is only limited exchange of income for leisure time. Elements that play a role are the "hours-invariant" component of labor costs and the tendency to relate career making to full-time employment. Note that more leisure time could also lead to more energy use if the extra time is used for energy-intensive activities (e.g., those involving a lot of driving); however, the limitation of production and consumption that is ultimately related to shorter working time should lead to a limitation of energy use.

Psychological mechanisms on the consumer side are important. Although traditionally the role of producers in generating consumption (e.g., through advertising) is stressed, it is also important to recognize that the possession of goods is a way for people to position themselves in social relations.

One might further ask why there is not a stronger shift from material consumption to services consumption. Often, the idea is expressed that when basic material needs are fulfilled, consumption patterns will shift to nonmaterial items. One of the explanations is that productivity increases are larger for the production of material goods than for the production of services. This leads to lower relative prices for material goods than for services, and this limits the shift to buying more services.

These mechanisms together may explain the increasing consumption and the limited shift of consumption in a less energy-intensive direction. These are all strong mechanisms and they are not easily changed.

5. WHAT CAN A CONSUMER DO?

A consumer can change the energy impact of his or her behavior by making changes in direct energy consumption. Concrete actions include insulating the house, purchasing efficient appliances, and purchasing efficient cars. All these actions involve improving the energy efficiency of energy use within the household; households generally can save substantially on the direct energy requirement (typically 20–50%).

Although most energy can be saved with the choice of specific equipment, some savings are also achievable in the use of equipment. Consumers can save energy by switching off a light when a room is not in use, driving a car conservatively, lowering the thermostat; etc. Typical savings through good housekeeping range from 5 to 20%.

In this section, we focus on changes in the packaging of goods and services that are purchased by households with the aim of limiting the energy requirement of households. We can distinguish the following categories of relevant changes:

A. Households may choose an energy-saving variant within a product category. The functionality of the consumption essentially does not change. These variants sometimes require more effort or cost more money. Examples include the following:

- Increase the lifetime of products: This can be realized through the purchase of products with improved quality (e.g., shoes, furniture, and tools) or by having products repaired more often (e.g., household equipment and shoes).
- Product or material substitution: There are several examples in the area of food (e.g., the replacement of animal-based food by vegetable food, selection of food of the season, or selection of food produced locally) but also, for example, in the material choice for clothing and floor coverings.
- Share the use of products with other households (e.g., cars, tools, and newspapers).

B. Households may choose a more luxury variant within a product group. The functionality of the product remains the same in physical terms but is much more attractive for the household. The price per functional unit is substantially higher, which leads to a lower energy intensity. Examples include the following:

- The purchase of products with individualized design: Examples are more high-end or tailor-made clothes and furniture.
- Contract out certain activities, such as household work, clothes washing, or maintenance of the house.
- Eat outdoors or purchase luxury food (including ecological food).

C. Households may refrain from certain activities and replace them by other less energy-intensive ones. An example is a household that does not go on a

long-distance vacation but chooses to buy a piano and music lessons with the money saved.

A quantitative assessment of all these options identified a potential shift in consumption patterns in a period of approximately two decades that can lead to a decrease in indirect energy requirement by 10% and a reduction in energy intensity by 30% (Table II). This means that in principle it is possible for household expenditures to keep increasing while at the same time there is a decrease in indirect energy requirement.

Following such analysis, an interesting experiment was carried out in The Netherlands. Fourteen households were selected to test whether it is feasible to combine higher "future" financial budgets with lower energy budgets. All households received a bonus of 20% of their net salary. Before the experiment started, the total (direct + indirect) energy requirement amounted to 260 GJ per household. They were given a maximum energy requirement target of 207 GJ, which they had to achieve in 2 years. They achieved this target and at the end of the experiment used 182 GJ. They reduced direct energy requirement from 60 to 50 GJ (the selected households already had an energy-efficient house), and indirect energy requirement from 200 to 132 GJ (all data are averages for the 14 households). Households reduced expenditures on energy-extensive products such as meat and flowers. They only slightly reduced their vacation expenditures but managed to reduce energy use by half through shorter distances and less use of air transport. In the area of transport, energy requirement was reduced by more than 20%. However, expenditures decreased less due to the increased use of taxis and first-class public transport.

Most households were satisfied with the low-energy lifestyle they had to develop. In the course of 2 years, the households developed an appetite for low-energy expenditures, including quality products, delivery services, delicious or ecological food, exclusive clothes, antique furniture, and fancy parties (i.e., something that could be considered a "light" lifestyle). The most problematic aspects were the constraints perceived on car transport and vacation destinations.

The experiment showed that a low-energy lifestyle can be developed. However, it must be noted that the participating households were very well informed, had computer-aided support, were coached on an individual basis, and were motivated by the income bonus. Nevertheless, this shows that the estimated potential is more than just theory.

A more systematic study of how consumers prefer to change their consumption patterns is presented in Table III. Technical measures (improving the energy efficiency of the home and equipment) are preferred over behavioral measures, especially shifts in consumption. Furthermore, home energy-saving measures were more acceptable than transport energy-saving measures.

TABLE II

Effect on the Energy Intensity of Household Consumption through Various Measures[a]

Category of measures[b]	Effect on indirect energy requirement (%)	Effect on expenditures (%)	Effect on energy intensity of goods and services (excluding direct energy use) (%)
A	−17	−4	−14
B	+3	+23	−17
C	+3	+8	−6
Total	−10	+26	−29

[a] *Source*. Vringer, K., Potting, J., Blok, K., and Kok, R. (1995). A reduction of the indirect energy requirement of households. *In* "Proceedings of the 1995 Summer Study: Sustainability and the Reinvention of the Government—A Challenge for Energy Efficiency" (A. Persson, Ed.). European Council for an Energy Efficient Economy, Stockholm.

[b] See description in text.

6. HOW CAN CONSUMPTION PATTERNS BE INFLUENCED THROUGH POLICIES?

The Intergovernmental Panel on Climate Change (IPCC) presents the following list of policy options:

1. Environmental education: Although public awareness is already fairly high, environmental education could be more effective if it is recognized that it is linked to lifestyle and self-awareness.

2. Decreasing marginal satisfaction with increasing private material consumption: People might prefer other forms of satisfaction than ever-increasing consumption of purchased commodities. A suitable index of welfare and the stimulation of public dialogue on the goal of economic action could raise awareness of this issue.

3. New emphasis on immaterial and common goods: In general, these goods have lower energy and environmental impacts.

TABLE III

Acceptability of 23 Energy-Saving Measures and Their Characteristics[a]

Energy-saving measure	Domain	Strategy	Amount	Mean[b]	SD[b]
Switching off lights in unused rooms	H	2	+	4.6	0.68
Appliances not on stand-by	H	2	+	4.4	0.87
Energy-efficient heating system	H	1	++	4.3	0.94
Walking or cycling short distances (<2.5 km)	T	2	++	4.3	1.00
House insulation	H	1	++	4.3	0.89
Line drying of laundry	H	2	++	4.0	1.13
Compact fluorescent light bulbs	H	1	+	4.0	0.94
Applying radiator insulation (foil)	H	1	+	4.0	0.99
Energy-efficient refrigerator	H	1	+	3.9	1.01
Shorter showers	H	2	++	3.6	1.10
Walking or cycling short distances (<5 km)	T	2	++	3.6	1.18
Econometer in car	T	1	+	3.6	1.16
Energy-efficient car	T	1	++	3.4	1.15
Car–pooling	T	2	+	3.4	1.33
Drive at most 100 km/h on the highway	T	2	+	3.3	1.35
Using public transport	T	2	++	3.2	1.42
No greenhouse vegetables	H	3	+	3.1	1.14
Thermostat maximally 18°C	H	2	++	3.1	1.27
Energy-extensive presents (no flowers)	H	3	+	2.9	1.19
Rinsing the dishes with cold water	H	2	+	2.9	1.31
Vacation by train	T	3	+	2.8	1.32
Altering food pattern	H	3	+	2.7	1.26
Hiring a housekeeper	H	3	++	2.7	1.31

[a] Abbreviations used: H; home measures; T; transport measures; 1; increasing technical energy efficiency; 2; a different use of products; 3; shifts in consumption; +; small energy saving; ++; large energy saving.

[b] Mean and standard deviation (SD) on a scale from 1("unacceptable") to 5 ("very acceptable"). Data obtained from a survey of 455 households in The Netherlands. From Poortinga *et al.* (2003).

4. New deals in collective action: By improving communication before people have to decide in social dilemmas, people might be more willing to cooperate in the common interest.

5. Environmental legislation: Governments could have people pay incrementally higher prices for environmentally more benign goods.

6. Creative democracy: For example, the local implementation of Agenda 21 could be considered; these actions are directed not only at concrete action but also at promoting understanding.

The IPCC concludes that lifestyle research is neglected compared to technology research.

To what extent will these policy proposals actually lead to a change in behavior and an actual impact on the development of energy use and greenhouse gas emissions? In policy analysis, three basic mechanisms to influence behavior are distinguished: communication, regulation, and economic incentives.

The list of options put forward by the IPCC mentioned previously relies heavily on various forms of communication. However, communication as a stand-alone instrument is weak; only a part (generally <10%) of the population changes its behavior. This portion often declines to 1% when substantial costs are involved; these costs can either have a monetary character or be a burden.

Regulation (command-and-control) makes certain behavior mandatory or prohibited. Given the wide variety of options, regulation would only have a substantial effect if an extended web of rules were implemented. Moreover, most options represent incremental changes in behavior (e.g., eating less meat), which are not very suitable for regulation. Nevertheless, for specific situations (e.g., forbidding car traffic in certain areas, such as inner cities and nature reserves) regulation might be an option.

Economic incentives include both subsidies, which stimulate desired behavior, and taxes, which

discourage undesired behavior. Taxes and subsidies for individual options have the same problem as regulation: There are simply too many options to design specific taxes or subsidies for each of them. Generic energy or carbon taxes will discourage all options with a high energy or carbon dioxide emissions impact. However, the effect on behavior is probably minor due to dilution of the taxation in the product chain. For example, a carbon tax of 50 Euro/ton of CO_2 can be considered high (only present in some countries and only for small consumers). Such a tax will lead to a price increase of approximately 10% for a product with high energy intensity of 25 MJ/Euro. Such a price increase will probably have only a limited effect on the consumption of these products. Nevertheless, taxes or subsidies may be useful for a limited number of specific products with a high energy impact (e.g., meat consumption and air travel).

The discussion of the three basic mechanisms focuses on the adoption of individual options. It would be much more effective if complete lifestyle changes could be accomplished. However, how can this be stimulated? An intriguing idea is the rationing of individual carbon quota. Everyone receives a specified amount of carbon points, which can be traded. For every purchase, not only is the monetary price paid but also the carbon points are paid from one's carbon credit card. The practicality of such a system needs to be proven.

In conclusion, there is no golden key to the stimulation of consumption patterns with low energy impacts and greenhouse gas emissions.

SEE ALSO THE FOLLOWING ARTICLES

Consumption, Energy, and the Environment • Diet, Energy, and Greenhouse Gas Emissions • Global Energy Use: Status and Trends • Goods and Services: Energy Costs • International Comparisons of Energy End Use: Benefits and Risks • Leisure, Energy Costs of • Suburbanization and Energy • Transportation and Energy Policy • Urbanization and Energy

Further Reading

Anonymous (1997), "Indicators of Energy Use and Energy Efficiency." International Energy Agency, Paris.

Carlsson-Kanyama, A., Ekström, M. P., and Shanahan, H. (2003). Food and life cycle energy inputs: Consequences of diet and ways to increase efficiency. *Ecol. Econ.* **44,** 293–307.

Cohen, C. A. M. J., Schaeffer, R., and Lenzen, M. (2004). Energy requirements of households in Brazil. *Energy Policy*, in press.

Lenzen, M. (1998). Energy and greenhouse gas cost of living for Australia during 1993/94. *Energy* **23,** 497–516.

Noorman, K. J., and Schoot Uiterkamp, A. J. M. (Eds.) (1997). "Green Households? Domestic Consumers, Environment and Sustainability." Earthscan, London.

Pachauri, S. (2004). An analysis of cross-sectional variations in total household energy requirements in India using micro survey data. *Energy Policy*, in press.

Pachauri, S., and Spreng, D. (2002). Direct and indirect energy requirement of households in India. *Energy Policy* **30,** 511–523.

Poortinga, W., Steg, L., Vlek, C., and Wiersma, G. (2003). Household preferences for energy-saving measures: A conjoint analysis. *Econ. Psychol.* **24,** 49–64.

Reinders, A. H. M. E., Vringer, K., and Blok, K. (2003). The direct and indirect energy requirement of households in the European Union. *Energy Policy* **31,** 139–154.

Røpke, I. (1999). The dynamics of willingness to consume. *Ecol. Econ.* **28,** 399–420.

Schipper, L. (1996). Life-styles and the environment—the case of energy. *Daedalus* **25,** 113–138.

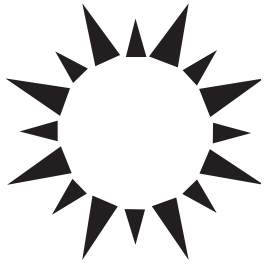
Toth, F., and Mwandosya, M. (convening lead authors) (2001). Decision making frameworks. *In* "Climate Change 2001—Mitigation" (B. Metz, O. Davidson, R. Swart, and J. Pan, Eds.), Report of Working Group III of the Intergovernmental Panel on Climate Change's Third Assessment Report. Cambridge Univ. Press, Cambridge, UK.

Vringer, K., and Blok, K. (1995). The direct and indirect energy requirement of households in The Netherlands. *Energy Policy* **23,** 893–910.

Vringer, K., and Blok, K. (2000). Long-term trends in direct and indirect household energy intensities: A factor in dematerialisation? *Energy Policy* **28,** 713–727.

Weber, C., and Perrels, A. (2000). Modelling lifestyle effects on energy demand and related emissions. *Energy Policy* **28,** 549–566.

Lithosphere, Energy Flows in



SETH STEIN
Northwestern University
Evanston, Illinois, United States

CAROL A. STEIN
University of Illinois at Chicago
Chicago, Illinois, United States

1. Plate Tectonics and Mantle Convection
2. Heat Flow
3. Heat Flow in Oceanic Lithosphere
4. Heat Flow in Continental Lithosphere
5. Global Heat Flow
6. Thermal Evolution of the Earth
7. Societal Implications

Glossary

asthenosphere The weaker shell of material underlying the lithosphere.
differentiation The process by which material within planets became compositionally segregated during their evolution.
geotherm The temperature as a function of depth in the earth.
heat flow The outward flow of heat through a portion of the earth's surface.
heat loss The outward flow of heat integrated over the earth's surface.
lithosphere The strong outer shell of a planet, approximately 100 km thick and composed of rigid plates on Earth.
plate Portion of Earth's lithosphere that moves as a coherent entity.
spreading center Divergent plate boundary in the oceans where hot material upwells to form new portions of the two plates.
subduction zone Convergent plate boundary where cooling oceanic lithosphere descends into the mantle.
thermal conduction Transfer of heat by molecular collisions.
thermal convection Transfer of heat by material motion due to temperature-induced density variations.
thermal history The temperature as a function of position and time on the earth or another planet.

It is said that heat is the geological lifeblood of planets. Planets are great heat engines whose nature and history govern their thermal, mechanical, and chemical evolution. On Earth, the heat engine is manifested by the plate tectonic cycle, whereby hot material upwells at spreading centers, also known as mid-ocean ridges, and then cools. Because the strength of rock decreases with temperature, the cooling material forms strong plates of the thick outer shell known as the lithosphere. The cooling oceanic lithosphere moves away from the ridges and eventually reaches subduction zones or trenches, where it descends in downgoing slabs, reheating as it goes. Thus, Earth's surface topography (e.g., mountain chains) and geological activity (e.g., volcanoes, geysers, earthquakes) reflect the planetary heat engine. As a result, measurements of heat flow at the earth's surface provide crucial constraints on how the heat engine works today and has evolved through time.

1. PLATE TECTONICS AND MANTLE CONVECTION

Plate tectonics is conceptually simple. It treats Earth's outer shell as made up of approximately 15 rigid plates, about 100 km thick, that move relative to each other at speeds of a few centimeters per year (roughly the speed at which fingernails grow). The plates are rigid in the sense that little deformation (and ideally no deformation) occurs within them, so they move as coherent entities and deformation occurs at their boundaries, causing earthquakes, mountain building, volcanism, and other spectacular phenomena. The direction of the relative motion between two plates at a point on their common boundary determines the nature of the boundary.

 © 2004 Elsevier Inc. All rights reserved.

At spreading centers, such as the Mid-Atlantic Ridge, both plates move away from the boundary, whereas at subduction zones, such as the Aleutian Trench, the subducting plate moves toward the boundary. At the third boundary type, transform faults such as the San Andreas Fault, relative plate motion is parallel to the boundary (Fig. 1).

These strong plates form Earth's lithosphere and move over the weaker asthenosphere below. The lithosphere and asthenosphere are mechanical units defined by their strength and the way in which they deform. The lithosphere includes two chemically distinct layers: the crust and part of the underlying upper mantle.

Evidence from seismic waves show that Earth is made up of a series of concentric shells with varying properties. From the outside, these are the crust (which extends to about 6 km depth beneath the oceans and 35–50 km beneath the continents), the underlying rocky mantle (which extends to 2900 km depth), the fluid iron outer core (which extends to 5150 km depth), and the solid iron inner core. This layered structure provides the primary evidence for the process of differentiation by which material within planets became compositionally segregated during their evolution. The differentiation process is thought to be largely controlled by Earth's thermal history—how temperature within the planet evolved with time.

The key to Earth's thermal evolution is the thermal convection system involving the mantle and core that removes heat from Earth's interior. Rocks within the mantle are sufficiently hot that they flow slowly over geological time, causing heat to be transported by this motion (i.e., thermal convection). Although much remains to be learned about this convective system, especially in the lower mantle and core, there is general agreement that at shallow depths the warm, and hence less dense, material rising below spreading centers forms upwelling limbs, whereas the relatively cold, and hence dense, subducting slabs form downwelling limbs. The lithosphere is a very thin layer compared with the rest of the mantle (100 km is ~3.4% of the mantle's radius), but it is where the greatest temperature change occurs, from more than 1400°C at a depth of 100 km to approximately 0°C at the surface. For this reason, the lithosphere is called a thermal boundary layer. Because of this temperature change, the lithosphere is much stronger than the underlying rock and so is also a mechanical boundary layer. This strong boundary layer is thought to be a primary reason why plate tectonics is much more complicated than simple models of convecting fluid. Moreover, the lithosphere, which contains the crust, is also a chemical boundary layer distinct from the remainder of the mantle. An important point is that the continental and oceanic lithospheres are distinct, although individual plates can contain both. The former is made of granitic rock, which is less dense than the basaltic rock composing the latter and so does not subduct. The oceanic lithosphere is continuously subducted and reformed at ridges and so never gets older than approximately 200 million years, whereas some continental lithosphere can be up to billions of years old.

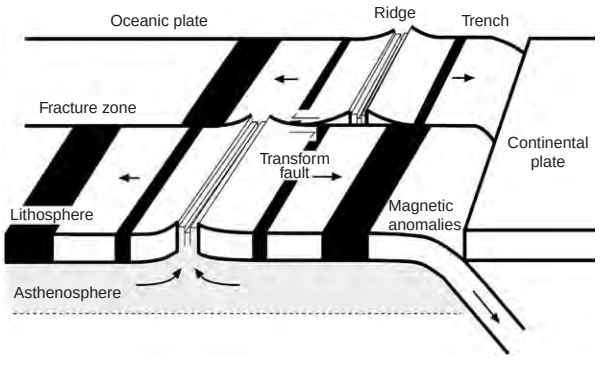


FIGURE 1 Plate tectonics at its simplest. Oceanic lithosphere is formed at ridges and subducted at trenches. At transform faults, plate motion is parallel to the boundaries.

2. HEAT FLOW

The challenge in understanding Earth's heat engine is that nearly all of our observations are made currently at the surface, whereas we seek to understand how it operates at depth over time. Our primary observation is the flow of heat at the surface today. Mathematically, heat flow q at the surface ($z = 0$) is the product of the vertical derivative of the temperature T with respect to depth z times the thermal conductivity:

$$q = k \frac{dT}{dz} \text{ at } z = 0.$$

(Normally, this equation requires a minus sign because heat flows from hot objects to cold ones. Without this sign, hot objects would get hotter. There is none here due to customary but inconsistent geophysical definitions; heat flow is measured upward, whereas depth is measured downward.) Hence, measuring the temperature at several depths

and the thermal conductivity gives the heat flow. The heat flow is the energy per unit time (power) per unit area flowing out from the earth and is reported in milliwatts per meter^2 or in "heat flow units" (HFU), where 1 HFU is approximately 42 mW m^{-2}.

Such measurements began in 1939 on land and in 1952 at sea. By now, heat flow has been measured at more than 30,000 sites worldwide. This is done at sea by using a probe that penetrates into the soft sediment on the sea floor and on land by using bore holes that have been drilled for oil or other purposes.

To relate surface heat flow to the temperature at depth as a function of time, we use the heat equation that describes the temperature $T(z, t)$ as a function of depth z and time t. A simple form, assuming that heat is transferred only by conduction, is

$$\rho C \frac{\partial T}{\partial t} = k \frac{\partial^2 T(z,t)}{\partial z^2} + \rho H,$$

where C is the specific heat per unit mass that shows how much material will heat up for a given temperature change, ρ is the density (mass/unit volume), and H is the rate at which heat is produced (by radioactivity or chemical reactions) per unit mass.

3. HEAT FLOW IN OCEANIC LITHOSPHERE

Oceanic heat flow data (Fig. 2, center right) show a clear variation with the age of the lithosphere. Values greater than 100 mW m^{-2} occur near the ridges and decrease smoothly to approximately 50 mW m^{-2} in the oldest oceanic lithosphere. Similarly, ocean depth is approximately 2500 m at the ridges and increases to 5600 m for the oldest sea floor.

These variations are the primary data used to constrain models of the thermal evolution of oceanic lithosphere and, hence, of the forces driving plate tectonics. They can be described using a simple but powerful model for the formation of the lithosphere by hot material at the ridge, which cools as the plate moves away. In this model, material at the ridge at a mantle temperature T_m (1300–1400°C) is brought to the ocean floor, which has a temperature T_s. The material then moves away while its upper surface remains at T_s. Because the plate moves away from the ridge faster than heat is conducted horizontally, we can consider only vertical heat conduction. Mathematically, this is the same as the cooling of a half-space originally at temperature T_m, whose surface is suddenly cooled to T_s at time $t = 0$.

The solution of the heat equation for this case can be viewed in terms of isotherms, or lines of constant temperature, in the plate. The depth to a given temperature increases as the square root of lithospheric age (Fig. 2, upper left). Such square root of time behavior occurs for any process described by a diffusion equation (e.g., the heat equation). For example, after a lava flow erupts, it cools as the square root of time.

The cooling of the lithosphere predicts both the depth and heat flow variations. As the rock cools, it gets denser and contracts, so ocean depth should vary with the square root of age, as observed. Similarly, the heat flow should decrease as the square root of age, as observed. Because ocean depth seems to "flatten" at approximately 70 million years, we often use a modification called a plate model, which assumes that the lithosphere evolves toward a finite plate thickness with a fixed basal temperature. The flattening reflects the fact that heat is being added from below, so the predicted sea floor depth and heat flow behave for young ages like in the half-space model but evolve asymptotically toward constant values for old ages.

Comparison with depth, heat flow, and geoid (gravity) data shows that the plate thermal model is a good, but not perfect, fit to the average data because processes other than this simple cooling also occur. For example, water flow in the crust transports some of the heat for ages less than approximately 50 million years, as shown by the spectacular hot springs at mid-ocean ridges. Thus, the observed heat flow is lower than the model's predictions, which assume that all heat is transferred by conduction. Some topographic effects, including the spectacular volcanic oceanic plateaus, result from crustal thickness variations. Because these and other effects vary from place to place, the data vary about their average values for a given age.

The cooling of the oceanic lithosphere as it moves away from ridges is part of the plate tectonic cycle, as is the cooled lithosphere heating up again as it subducts. Both portions give rise to thermal buoyancy forces, which drive plate motions, due to the density contrast resulting from the temperature difference between the plate and its surroundings. Often, the first is called ridge push, which is a confusing term because it is zero at the ridge and increases linearly with plate age. Similarly, the force due to cold, dense, subducting slab is called slab pull. Although it is useful to think of the two forces separately, both are parts of the net buoyancy force due to mantle convection.

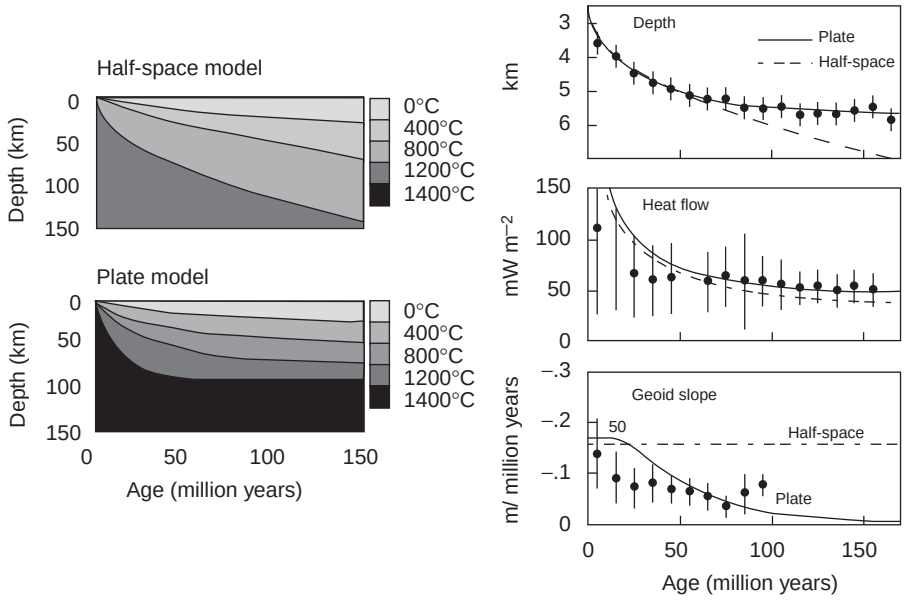


FIGURE 2 Models and data for thermal evolution of the oceanic lithosphere. Isotherms for thermal models are on the left-hand side. The lithosphere continues cooling for all ages in a half-space model and equilibrates for 70 million years lithosphere in a plate model with a 95-km thick thermal lithosphere. The plate model shown has a higher basal temperature than the half-space model. Comparisons of thermal model predictions to various data are on the right-hand side. All show a lithospheric cooling signal and are better (but far from perfectly) fit by a plate than by a half-space model. Reprinted from Stein and Wysession (2003), with permission.

4. HEAT FLOW IN CONTINENTAL LITHOSPHERE

Continental heat flow differs from oceanic heat flow because continental crust is not subducted. However, continental heat flow also varies with geological age, from approximately $80\,\mathrm{mW\,m^{-2}}$ in the youngest rocks to $40\,\mathrm{mW\,m^{-2}}$ in the oldest ones (Fig. 3).

Using the heat flow to infer the temperature with depth beneath continents is more complicated than doing so with depth beneath the oceans because the rocks of the continental crust contain much more of the radioactive heat-producing elements uranium, thorium, and potassium than does the oceanic crust. Hence, the geotherm, or temperature as a function of depth, is found near the earth's surface by making the simplifying assumption that the temperature is not changing with time. In this case, the heat equation can be solved to give

$$T(z) = T_s + \frac{q}{k}z - \frac{\rho H}{2k}z^2.$$

This solution (Fig. 4) has two interesting consequences. First, because of the heat production, the temperature at depth is less than it would be without heat production. Moreover, because the geotherm is less steep at depth, the heat flow at depth is less than

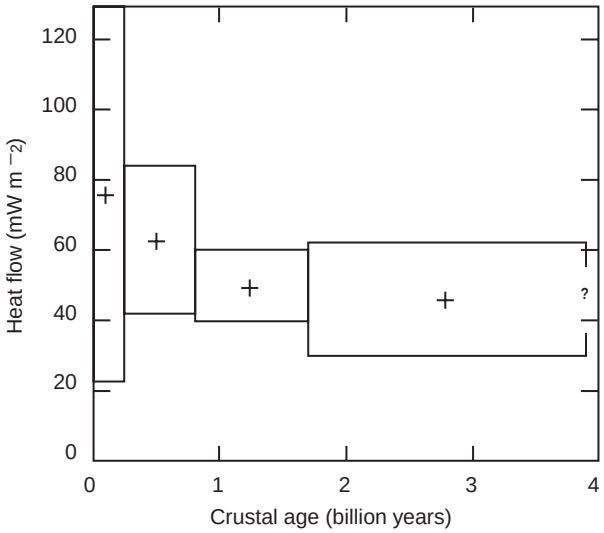


FIGURE 3 Heat flow versus age for continental lithosphere. Data from Sclater *et al.* (1980).

that at the surface. Physically, this is because some of the heat is generated in the crust.

A second consequence is that at relatively shallow depth, or approximately 100 km, the geotherm intersects the solidus, where rock begins to melt, and then the liquidus, where all of the rock has melted. This cannot be the case because seismic shear

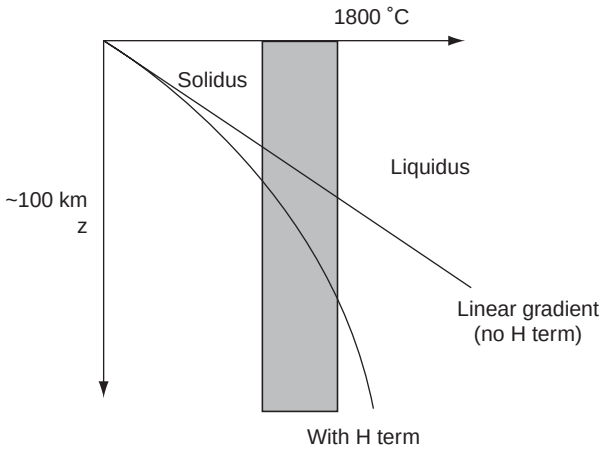


FIGURE 4 Schematic comparison of continental geotherms with and without crustal heat production (H term). In reality, the liquidus and solidus temperatures increase with depth, and convective effects need to be included.

waves travel through the entire mantle, but molten rock does not transmit them. The explanation for this paradox is that this geotherm is computed assuming that heat is transferred entirely by conduction, whereas at these high temperatures rock flows, and so heat, are transferred by convection. Hence, the lithosphere is the strong outer layer of the mantle convection system, within which heat transfer occurs by conduction, and is underlain by weaker rock, within which heat is transferred by convection that is much more efficient.

Intriguingly, the heat flow in old continental and old oceanic lithospheres are approximately equal. This is surprising given that the former includes the effect of crustal heat production, whereas the latter reflects only mantle heat flow. Whether this equality is simply a coincidence or reflects underlying physics is unclear.

5. GLOBAL HEAT FLOW

For many applications, we use the global average or total heat flow. To find these, we take the mean oceanic and continental heat flows of approximately 101 and 65 $\mathrm{mW\,m^{-2}}$ and weight them by their areas. This gives a global mean of about 87 $\mathrm{mW\,m^{-2}}$ or a total global heat loss of 44×10^{12} W. Of this loss, 70% is in the oceans and 30% is in the continents, reflecting both the higher heat flow and larger area of the oceans. The term "loss" indicates that the solid earth is losing heat. Heat loss is measured in units of energy per unit time (power).

6. THERMAL EVOLUTION OF THE EARTH

Heat flow provides a crucial constraint on the thermal evolution of the earth. A common approach is to consider conservation of energy for the whole earth. In this case, the change in the average temperature T as a function of time t is given by the balance between heat produced and that lost at the surface:

$$MC\frac{\partial T}{\partial t} = MH - Aq,$$

where M is the mass of the earth, A is its surface area, C is specific heat, q is the average heat flow, and H is the average rate of radioactive heat production. This equation can be applied to both the core and mantle or to the mantle alone. The heat flow used is an average of that coming from the mantle, or approximately 72 $\mathrm{mW\,m^{-2}}$, which is estimated by removing the 17% or so thought to be produced by radioactivity in the continental crust.

The balance between surface heat flow and heat production is given by the Urey ratio

$$Ur = \frac{MH}{Aq},$$

so the temperature evolves with time according to

$$\frac{\partial T}{\partial t} = \frac{Aq}{MC}(Ur - 1).$$

The Urey ratio is thought to be between 0.5 and 0.8, so heat is lost faster than it is produced and the earth has been cooling with time. Thermal history models are derived by assuming an initial temperature and distribution of radioactive elements, both of which decay with time. The temperature is thought to have been high, owing to the release of gravitational potential energy as the planet accreted, gravitational potential energy that was released as iron sank to form the earth's core, and radioactivity. Analytic parameterized models are used to estimate how convection transferred heat from the interior to the surface over time. An important aspect of these models is the strong dependence of mantle viscosity on temperature. Early in the earth's history, the planet was hotter, making mantle viscosity lower and heat transfer by convection more efficient. With time (Fig. 5), the earth cooled and the viscosity increased. This idea seems plausible because rocks formed roughly 3 billion years ago solidified at temperatures approximately 300°C higher than those of comparable rocks today.

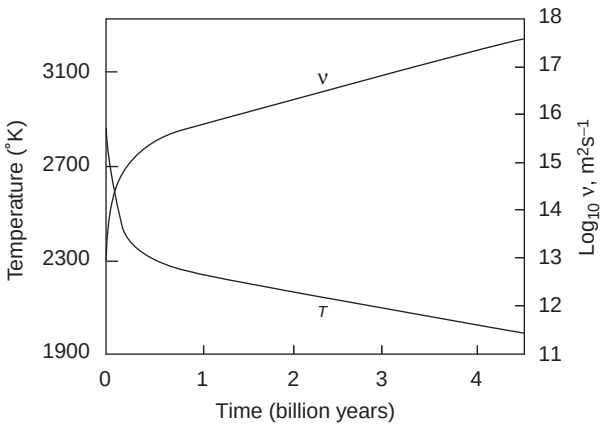


FIGURE 5 Average mantle temperature and viscosity history for a parameterized convection model of Earth's evolution. Reprinted from Schubert *et al.* (1980).

Hence, the plate tectonic cycle is evidence that the earth's cooling continues today as heat is transferred from the planet's interior. As cooling continues, we expect the strong lithosphere to thicken with time. Similarly, cooling causes the earth's solid inner core to grow with time at the expense of the liquid outer core. Because convection involves heat transfer by the motion of material, the thermal evolution of the earth gives rise to the chemical differentiation that produced the earth's current structure.

This approach gives insight into the differences among the terrestrial or inner planets of the solar system. What we know about the earth and our more limited knowledge of the moon and other planets suggests that although there are differences among the inner planets that reflect their initial compositions, there are also similarities in their evolution. As shown in Fig. 6, planets may follow a similar life cycle with phases including their formation, early convection and core formation, plate tectonics, terminal volcanism, and quiescence. This evolution is driven by the available energy sources and reflects the planets' cooling with time. Thus, even though the planets formed at about the same time, they are at different stages in their life cycles (as are a human and a dog born on the same date).

The earth is in its middle age, characterized by a relatively thin lithosphere and active plate tectonics. Conversely, it appears that the moon now has a thick lithosphere and is tectonically inactive. Thus, it seems to have lost much of its heat, presumably due to its small size, which favors rapid heat loss. In general, we would expect the heat available from the gravitational energy of accretion and radioactivity to increase with the planet's volume, whereas the rate of heat loss through the surface should depend on its surface area. Hence, the remaining heat should vary as

$$\text{remaining heat} \rightarrow = \frac{\text{available}}{\text{loss}} = \frac{(4/3)\pi r^3}{4\pi r^2} = \frac{r}{3},$$

so larger planets would retain more heat and be more active.

From such arguments, we might expect Mercury and Mars, which are larger than the moon but smaller than the earth, to have also reached their old age with little further active tectonics. Venus, which is comparable in size to the earth, might still be active but with episodic, rather than continuous, plate tectonics.

7. SOCIETAL IMPLICATIONS

One might not expect the flow of heat from the earth to be of great significance for society. This flow typically seems small because the ground is not hot when we touch it, except in geothermal areas. The average heat flow is much less than the solar constant, which measures the average power delivered per unit area at the top of the atmosphere, given that their ratio is $87\,\text{mW}\,\text{m}^{-2}/1372\,\text{W}\,\text{m}^{-2} = 6 \times 10^{-5}$. Even so, heat from the earth can have important consequences, both good and bad. For example, volcanic eruptions can be very destructive. However, islands such as Hawaii result. Moreover, the steady supply of volcanic gases are crucial to maintaining the atmosphere. Earthquakes are a destructive consequence of tectonic processes that shape topography that we enjoy and use.

Energy from the earth is also crucial for society's energy needs. Direct use of geothermal energy, although only a minor contributor to global energy supply, is important in a few areas, such as Iceland and New Zealand, where high temperature rock is close to the surface. Other energy sources ultimately require energy from the earth. Hydroelectric power depends on topography, raised by tectonic processes, to provide potential energy. Fossil fuels, such as coal, oil, and natural gas, result because organic matter has been buried to the depths required for the earth's heat to convert it to useful fuels. Nuclear power relies on the same radioactive elements that supply much of the earth's heat.

Finally, at a fundamental level, humanity evolved and survives in large part because Earth loses its heat by plate tectonics. Plate tectonics is crucial for the evolution of Earth's ocean and atmosphere because it

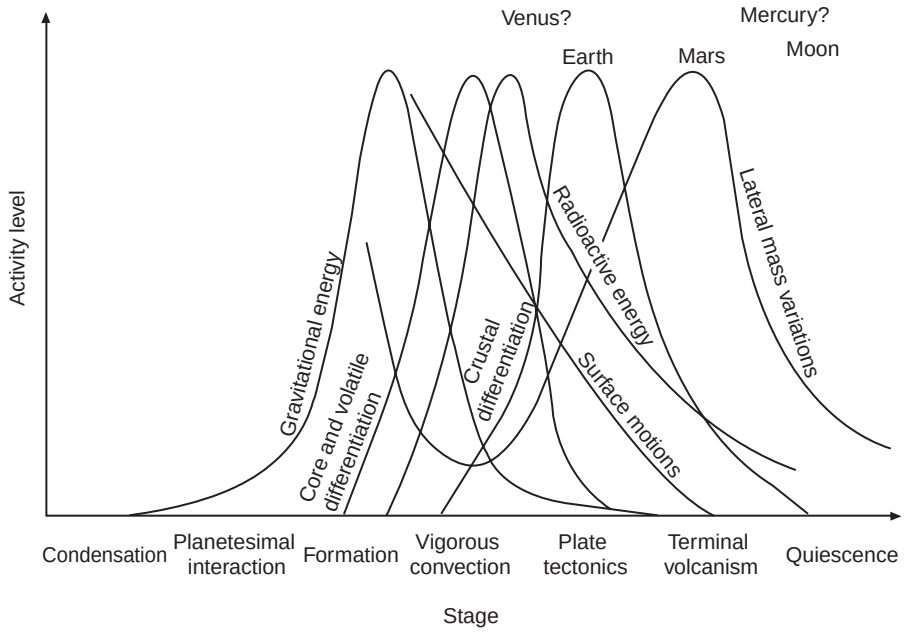


FIGURE 6 A model for the evolution of terrestrial planets showing the energy sources at each stage. Presented in text and verse by Kaula (1975).

involves many of the primary means (including volcanism, hydrothermal circulation through cooling oceanic lithosphere, and the cycle of uplift and erosion) by which the solid earth interacts with the ocean and atmosphere. The chemistry of the oceans and atmosphere depend in large part on plate tectonic processes, and many long-term features of climate are influenced by mountains that are uplifted by plate convergence and the positions of continents that control ocean circulation. Plate tectonics may explain how life evolved on Earth (at mid-ocean ridge hot springs) and is crucial for its survival (the atmosphere is maintained by plate boundary volcanism, and plate tectonics raises the continents above sea level).

SEE ALSO THE FOLLOWING ARTICLES

Earth's Energy Balance • Geothermal Direct Use • Lunar–Solar Power System • Ocean, Energy Flows in • Ocean Thermal Energy • Sun, Energy from • Thermal Energy Storage

Further Reading

Kaula, W. M. (1975). The seven ages of a planet. *Icarus* **26,** 1–15.

Pollack, H. N., Hurter, S. J., and Johnston, J. R. (1993). Heat loss from the earth's interior: Analysis of the global data set. *Rev. Geophys.* **31,** 267–280.

Schubert, G., Stevenson, D., and Cassen, P. (1980). Whole planet cooling and the radiogenic heat source contents of the earth and moon. *J. Geophys. Res.* **85,** 2511–2518.

Schubert, G., Turcotte, D., and Olson, P. (2001). "Mantle Convection in the Earth and Planets." Cambridge University Press, Cambridge, UK.

Sclater, J. G., Jaupart, C., and Galson, D. (1980). The heat flow through oceanic and continental crust and the heat loss of the earth. *Rev. Geophys. Space Phys.* **18,** 269–311.

Sleep, N. H. (1979). Thermal history and degassing of the earth: Some simple calculations. *J. Geol.* **87,** 671–686.

Stacey, F. D. (1992). "Physics of the Earth," 3rd ed. John Wiley, New York.

Stein, C. A., and Stein, S. (1992). A model for the global variation in oceanic depth and heat flow with lithospheric age. *Nature* **359,** 123–129.

Stein, C. A., and Stein, S. (1994). Constraints on hydrothermal heat flux through the oceanic lithosphere from global heat flow. *J. Geophys. Res.* **99,** 3081–3095.

Stein, S., and Wysession, M. (2003). "Introduction to Seismology, Earthquakes, and Earth Structure." Blackwell, Oxford, UK.

Verhoogen, J. (1980). "Energetics of the Earth." National Academy of Sciences, Washington, DC.

Livestock Production and Energy Use

DAVID PIMENTEL
Cornell University
Ithaca, New York, United States

1. **Animal Products Consumed in the U.S. Diet**
2. **Energy Inputs in Animal Product Production**
3. **Land Resources**
4. **Water Resources**
5. **World Food Needs**
6. **Conclusion**

Glossary

cellulose Any of several fibrous substances constituting the chief part of the cell walls of plants.
erosion The slow breakdown of rock or the movement and transport of soil from one location to another. Soil erosion in crop and livestock production is considered serious worldwide.
forage Vegetable food (as hay, grain) for domestic animals.
fossil energy Energy from sources extracted from the earth such as oil, coal, and natural gas.
hectare (ha) A metric unit of area equal to 10,000 square meters or 2.47 acres.
monoculture The cultivation of a single crop or product to the exclusion of other possible uses of the land.
per capita By or for each person.
salinization The process by which salts accumulate in soil.
waterlogging The saturation of soil with water causing the water table to rise high enough to expel normal soil gases and interfere with plant growth or cultivation.

Worldwide an estimated 2 billion people live primarily on a meat-based diet, while an estimated 4 billion people live primarily on a plant-based diet. The shortages of cropland, freshwater, and energy resources require most of the 4 billion people to live primarily on a plant-based diet; however, there are serious food shortages worldwide. For instance, the World Health Organization recently reported that more than 3 billion people are malnourished in the world. This is the largest number and proportion of malnourished people ever recorded in history. In large measure, the food shortage and malnourishment problem are primarily related to rapid population growth in the world plus the declining per capita availability of land, water, and energy resources required for food production. Meat, milk, and eggs contribute valuable nutrients to the human diet in the United States and the rest of the world. To produce animal protein successfully requires the expenditure of human and fossil energy to supply livestock forage and grain. The land, devoted to grain or forage for livestock production, is exposed to soil erosion, which slowly diminishes the fertility of the soil and its productivity. Additionally, animal production requires large inputs of water for the grain and forage crops and, to a lesser extent, directly for animal consumption. All of these factors interact to determine the ultimate success of animal production systems. This article includes analyses of the quantities of animal products produced; energy, land, and water resource inputs in livestock production; and meat, milk, and egg production.

1. ANIMAL PRODUCTS CONSUMED IN THE U.S. DIET

In the United States, more than 8 billion livestock are maintained to supply the animal protein consumed annually. In addition to the large amount of cultivated forage, the livestock population consumes about seven times as much grain as is consumed directly by the entire U.S. population.

From the livestock population of more than 8 billion, approximately 7.5 million tons (metric) of animal protein is produced each year (Table I). If distributed equally, it would be sufficient to supply

 © 2004 Elsevier Inc. All rights reserved.

TABLE I

Number of Livestock in the United States

Livestock	Number $\times 10^6$
Sheep	7
Dairy	13
Swine	60
Beef cattle	74
Turkeys	273
Broilers	8000
Eggs	77,000

about 75 g of animal protein daily per American. With the addition of 37 g of available plant protein, a total of 112 g of protein is available per capita. In contrast, the recommended daily allowance (RDA) per adult per day is 56 g of protein for a mixed diet for an adult. Therefore, based on these data, each American is consuming about twice the RDA for protein per day. About 144 kg of meat, including fish, are eaten per American per year. In addition, 271 kg of milk and eggs are consumed per capita in the United States per year.

2. ENERGY INPUTS IN ANIMAL PRODUCT PRODUCTION

Each year an estimated 45 million tons of plant protein are fed to U.S. livestock to produce approximately 7.5 million tons of animal protein for human consumption. To produce this animal protein, about 28 million tons of plant protein from grain and 17 million tons of plant protein from forage are fed to the animals (Table II). Thus, for every kilogram of high-quality animal protein, livestock are fed nearly 6 kg of plant protein. In the conversion of plant protein into animal protein, there are two principal costs: (1) the direct costs of production of the harvested animal including the grain and forage and (2) the indirect costs for maintaining the breeding animals (mother and father).

The major fossil energy inputs for grain and forage production include fertilizers, farm machinery, fuel, irrigation, and pesticides. The energy inputs vary according to the particular crop and forage being grown. When these inputs are balanced against their energy and protein content, grains and some legumes like soybeans are produced more efficiently in terms of energy inputs than fruits, vegetables, and animal products. In the United States, the average

TABLE II

Grain and Forage Inputs per Kilogram of Animal Product Produced, and Fossil Energy Inputs (kcal) Required to Produce, 1 kcal of Animal Protein

Livestock	Grain (kg)[a]	Forage (kg)[b,c]	kcal input/kcal protein
Lamb	21	30	57:1
Beef cattle	13	30	40:1
Eggs	11	—	39:1
Beef cattle	—	200	20:1
Swine	5.9	—	14:1
Dairy (milk)	0.7	1	14:1
Turkeys	3.8	—	10:1
Broilers	2.3	—	4:1

From [a]USDA (2001). "Agricultural Statistics." U.S. Department of Agriculture, Washington, DC; [b]Heischmidt, R. K., Short, R. E., and Grings, E. E. (1996). Ecosystems, sustainability and animal agriculture. *Animal Sci.* **74**(6), 1395–1405; [c]Morrison, F. B. (1956). "Feeds and Feeling." Morrison Publishing Co., Ithaca, NY.

protein yield of the five major grains (plus soybeans) fed to livestock is about 700 kg/ha. To produce a kilogram of plant protein requires about 10 kcal of fossil energy.

Forage can be fed to ruminant animals, like cattle and sheep, because they can convert forage cellulose into usable nutrients through microbial fermentation. The total plant protein produced on good U.S. pasture and fed to ruminants is 60% of the amount produced by grains (Table II). The current yield of beef protein from productive pastures is about 66 kg/ha, while the energy input per kilogram of animal protein produced is 3500 kcal. Therefore, animal protein production on good pastures is less expensive in terms of fossil energy inputs than grain protein production (Table II).

Of the livestock systems evaluated in this investigation, chicken-broiler production is the most efficient with an input of 4 kcal of fossil energy per 1 kcal of broiler protein produced (Table II). Broilers are a grain-only system. Turkey production, also a grain only system, is next in efficiency with a ratio of 10:1. Milk production based on a mixture of grain and forage also is relatively efficient with a ratio of 14:1 (Table II). Nearly all the feed protein consumed by broilers is grain, whereas for milk production about two-thirds is grain (Table II). Of course, 100% of milk production could be produced on forage. Both pork and egg production also depend on grain (see Table IV). Pork has a 14:1 ratio, whereas egg production is relatively more costly in terms of feed energy requiring a 39:1 ratio (Table II).

TABLE III

Calorie, Water, and Protein Availability per Kilogram of Animal Product

Livestock	Energy (kcal)	Water (%)	Protein (g)
Lamb	2521	47	220
Beef	2565	49	186
Turkey	1193	55	123
Egg	1469	74	116
Pork	2342	57	134
Dairy	647	87	34
Broiler	1357	71	238

From Pimentel, D. (1997). "Livestock Production: Energy Inputs and the Environment." Canadian Society of Animal Science, Proceedings. Canadian Society of Animal Science, Montreal, Quebec.

TABLE IV

Estimated Liters of Water Required to Produce 1 kg of Food and Forage Crops

Crop	Liters/kg
Potatoes	500
Wheat	900
Alfalfa	900
Sorghum	1100
Corn	1400
Rice	1900
Soybeans	2000
Broiler	3500
Beef	43,000

From Pimentel, D., Houser, J., Preiss, E., White, O., Fang, H., Mesnick, L., Barsky, T., Tariche, S., Schreck, J., and Alpert, J. (1997). Water resources: Agriculture, the environment, and society. *BioScience* **47**(2), 97–106.

The two livestock systems depending most heavily on forage, but still using significant amounts of grain, are the beef and lamb production systems The lamb system with a ratio of 57:1 and the beef system with a ratio of 40:1 are the two highest (Table II). If these animals were fed only on good-quality forage, the energy inputs could be reduced by about half depending on the conditions of the pasture forage as well as the management practices. Note that beef fed 200 kg of forage and no grain had an energy input per kcal protein output ratio of 20:1 (Table II). Rainfall is critical for all productive pasture systems.

Per kilogram of animal product foods, broiler chicken flesh has the largest percentage of protein and milk the lowest (Table III). Beef has the highest calorie content because of its high fat content and relatively low water content. Of the all the animal products, milk has the highest water content with 87%.

The average fossil energy input for all animal protein production systems studied is about 25 kcal of fossil energy input per kcal of animal protein produced (Tale II). This energy input is more than 10 times greater than the average input to output ratio for grain protein production, which was about 2.5 cal per kcal of protein produced. As a food for humans, however, animal protein has about 1.4 times the biological value as a food compared with grain protein.

3. LAND RESOURCES

Livestock production requires a large number of hectares to supply the grains, forages, and pastures for animal feeds. In fact, nearly 300 million hectares of land are devoted to producing the feed for the U.S. livestock population. Of this, 262 million hectares are pasture and about 30 million hectares are for cultivated grains. In addition to the large amount of forages and grass that are unsuitable for human consumption and are fed to animals, about 323 million tons of grains—or about 816 kg per American in the United States are fed to livestock to provide meat, milk, and eggs.

More than 99.2% of U.S. food is produced on the land, while less than 0.8% comes from oceans and other aquatic ecosystems. The continued use and productivity of the land is a growing concern because of the rapid rate of soil erosion and degradation that is taking place throughout the United States and indeed throughout the world. Each year about 90% of U.S. cropland is losing soil at an average rate 13 times above the sustainable rate of 1 t/ha/year. On croplands where most grain is produced, soil loss averages more than 13 t/ha/year from the combined effects of water and wind erosion. Also, our rangelands are losing soil an average of 13 t/ha/year. About 60% of United States rangeland is being overgrazed and is subject to accelerated erosion.

The concern about high rates of soil erosion in the United States and the rest of the world is evident when it is understood that it takes approximately 500 years to replace 25 mm (1 in.) of lost soil. Clearly a farmer cannot wait for the replacement of 25 mm of soil. Commercial fertilizers can replace some nutrient loss resulting from soil erosion, but this requires large inputs of fossil energy.

The future of all agricultural production that requires land, including that targeted for livestock, will feel the effects of land degradation, particularly when fossil fuel supplies decline and prices increase. Soil erosion losses, compounded by salinization and waterlogging, are causing the abandonment of nearly 1 million hectares of U.S. agricultural land per year. Some of the abandoned, degraded cropland may find use as either pasture or forest.

The costs of soil erosion are well illustrated by the loss of rich U.S. soils. Iowa, which has some of the best soils in the world, has lost more than one-half of its topsoil after only 150 years of farming. Iowa continues to lose topsoil at an alarming rate of about 30 t/ha/year, which is about 30 times faster than the rate of soil formation. The rich Palouse soils of the northwestern United States have similarly lost about 40% of their topsoil in the past century.

Despite the efforts of the USDA Soil Conservation Service, erosion rates in the United States have decreased only slightly during the past 50 years. This is the result of major changes in agricultural production, such as an emphasis on commodity price-support programs; widespread planting of crop monocultures; crop specialization; abandonment of crop rotations; the removal of tree shelter-belts; leaving the soil without protective biomass cover; and the use of heavy farm machinery. Concurrently these changes have been accompanied by the creation of fewer and larger farms where increased mechanization is a necessity.

Although modern farming practices are contributing to the soil erosion problem, the failure of farmers and governments to recognize and address the soil erosion problem is equally important if soil depletion is to be halted. Erosion often goes unnoticed by some farmers because soil loss is difficult to measure visually. For instance, one night's wind- or rainstorm could erode 15 t of soil per hectare as a sheet, which would be only 1 mm of soil; the next morning, the farmer might not even notice. This soil loss continues slowly, quietly, year after year, until the land is no longer productive. In addition, governments tend to ignore erosion because of its insidious nature and because it does not seem to be a major environmental crisis such as floods or tornadoes.

4. WATER RESOURCES

Agricultural production, including livestock production, consumes more fresh water than any other human activity. Western U.S. agriculture accounts for about 81% of the fresh water consumed after being withdrawn. Growing plants render all water nonrecoverable through evaporation and transpiration. In the United States, about 62% of the water used in agricultural irrigation comes from surface sources and 38% ground water sources.

The transfer of water to the atmosphere from the terrestrial environment by transpiration through vegetation is estimated to range between 38% and 65% of the rainfall depending on the terrestrial ecosystem. The vital photosynthetic processes and temperature control necessitates that the plants consume enormous amounts of water.

The water required to produce various food and forage crops range from 500 liters to 2000 liters of water per kilogram of plant biomass produced (Table IV). For example, a hectare of U.S. corn producing about 8000 kg per year transpires about 5 million liters of water during the growing season. Approximately 1000 mm (10 million liters per hectare) of rainfall or other sources of water are needed during the growing season for corn production. Even with 800 to 1000 mm of annual rainfall in the corn-belt region, corn usually suffers from some lack of water during the summer growing season. Producing 1 kg of beef requires about 43 times more water than it takes to produce 1 kilogram of grain. Livestock directly use only 1.3% of the total water used in agriculture. However, when the water required for forage and grain production is included, this dramatically increases the water requirement for livestock production. Producing 1 kg of fresh beef requires about 13 kg of grain and 30 kg of forage (Table II). This much grain and forage requires a total of 43,000 liters of water. On rangeland where the animal consumes about 200 kg of forage to produce 1 kg of beef, about 200,000 liters of water are need to produce the 1 kg of beef. With forage and some cereal crops, livestock can be produced in areas with low rainfall ranging from 150 to 200 mm per year. However, crop production and yields are low under such conditions.

Animals vary in the amounts of water required for their production. In contrast to beef, 1 kg of broiler chicken can be produced with about 2.6 kg of grain requiring approximately 3500 liters of water (Table IV).

Water shortages are already severe in the western and southern United States. The situation grows worse as the U.S. population and its requirements for water, including for agriculture, rapidly increase.

5. WORLD FOOD NEEDS

Worldwide human food needs are rising and will continue to rise with the world population. Currently, more than 3 billion people are malnourished based on calories, protein, vital minerals, and vitamin shortages in their diets. There are currently 6.2 billion people on Earth, and it is projected that the world population will double to more than 12 billion in less than 50 years, based on the current growth rate. The U.S. population is also increasing rapidly. The U.S. population is currently at 285 million and is expected to double to 570 million in about 70 years. Food security becomes at risk as more and more people need food, while the required resources of land, water, and energy decline per person.

Food consumption patterns in the United States and most other developed nations include generous amounts of animal products. More than half of U.S. grain and nearly 40% of world grain is being fed to livestock rather than being consumed directly by humans. Grains provide 80% of the world's food supply. Although grain production is increasing in total, the per capita supply has been decreasing for nearly two decades. Clearly, there is reason for concern in the future.

If all the 323 million tons of grain currently being fed to livestock was consumed directly by people, the number of people who could be fed would be approximately 1 billion. Also, if this much grain were exported, it would provide approximately $80 billion each year in income—this is sufficient income to pay for our current oil bill of $75 billion per year. Of course, exporting all the grain currently fed to livestock would reduce the average protein consumption of Americans from 112 gram per day to approximately 73 grams per day. Yet this intake would still be greater than the 56 gram of protein suggested by the RDA.

Exporting all U.S. grain that is now fed to livestock assumes that livestock production would change to a grass-fed livestock production system. Animal protein in the diet would then decrease from the current level of 75 grams to 36 grams per day, or about one-half. Again, the diet for the average American would be more than adequate in terms of protein consumption, provided that there was no change in the current level of plant protein consumed. In fact, consuming less meat, milk, and eggs and eating more grains and vegetables would improve the diet of the average American.

6. CONCLUSION

Meat, milk, and egg production in the United States relies on significant quantities of fossil energy, land, and water resources. Grain-fed livestock systems use large quantities of energy because grain crops are cultivated; in contrast, cattle grazed on pastures use considerably less energy than grain-fed cattle. An average of 25 kcal of fossil energy is required to produce 1 kcal of animal protein and requires approximately 10 times the energy expended to produce 1 kcal of plant protein. However, it should be noted that animal protein is 1.4 times more nutritious for humans than plant protein.

Nearly one-third of the U.S. land area is devoted to livestock production. Of this, about 10% is devoted to grain production and the remainder is used for forage and rangeland production. The pastureland and rangeland is marginal in terms of productivity because there is too little rainfall for crop production.

Livestock production is also a major consumer of water because grains and forage consumed by livestock require significant amounts of water for growth and production. To produce 1 kg of grain requires about 1000 liters of water. Based on grain and forage consumption, about 43,000 liters of water are required to produce 1 kg of beef. In regions where water is already in short supply and where aquifers are currently being mined faster than they can be recharged, major decisions will have to be made concerning all agricultural production, including grain and forage crops for livestock.

As human food needs escalate along with population numbers, serious consideration must be given to the conservation of fossil energy, land, and water resources. The careful stewardship of these resources is vital if livestock production, and indeed agriculture, will be sustainable for future generations. In the end, population growth must be reduced, in the United States and in throughout the world, if we are to achieve a quality life for ourselves and for our grandchildren.

SEE ALSO THE FOLLOWING ARTICLES

Aquaculture and Energy Use • *Conversion of Energy: People and Animals* • *Diet, Energy, and Greenhouse Gas Emissions* • *Fisheries and Energy Use* • *Food System, Energy Use in* • *Hunting and Gathering Societies, Energy Flows in* • *Industrial Agriculture,*

Energy Flows in • Land Requirements of Energy Systems • Lifestyles and Energy

Further Reading

FAO (1998). Food Balance Sheet. http://apps.fao.org (November 16, 2000).

Pimentel, D. (2004). Soil erosion: A major environmental threat. *In* "Making Development Work: A New Role for Science" (G. Leclerc and C. A. S. Hall, Eds.). University of New Mexico Press, Albuquerque, New Mexico (in press).

Pimentel, D., Doughty, R., Carothers, C., Lamberson, S., Bora, N., and Lee, K. (2002). Energy inputs in crop production: comparison of developed and developing countries. *In* "Food Security & Environmental Quality in the Developing World" (L. Lal, D. Hansen, N. Uphoff, and S. Slack, Eds.), pp. 129–151. CRC Press, Boca Raton, FL.

Pimentel, D., and Pimentel, M. (1996). "Food, Energy and Society." Colorado University Press, Niwot, CO.

Pimentel, D., and Pimentel, M. (2002). Sustainability of meat-based and plant-based diets and the environment. *Am. J. Clinical Nutrition* **78** (Suppl.), 660–663.

Thomas, G. W. (1987). Water: Critical and evasive resource on semiarid lands. *In* "Water and Water Policy in World Food Supplies" (W. R. Jordan, Ed.), pp. 83–90. College Texas A & M, University Press, Station, TX.

Troeh, F. R., Hobbs, J. A., and Donahue, R. L. (1999). "Soil and Water Conservation." Prentice Hall, Englewood Cliffs, NJ.

USBC (2002). "Statistical Abstract of the United States 2002." Bureau of the Census, U. S. Government Printing Office, Washington, DC.

WHO (2000). Malnutrition Worldwide. http://www.who.int (July 27, 2000).

Lunar–Solar Power System

DAVID R. CRISWELL
University of Houston
Houston, Texas, United States

1. **Isolated Spaceship Earth and the Human Race to Sustainability**
2. **The Immediate Global Energy Challenge**
3. **Lunar–Solar Power System**
4. **Lunar–Solar Power Demonstration Base**
5. **Scale and Cost for 20 TWe Lunar–Solar Power System**
6. **Conclusion**

Glossary

astronomical unit (AU) The radius of a Keplerian circular orbit of a point mass having an orbital period of $2 \times \pi/k$ days (where k is the Gaussian gravitational constant) or 149,597,870.691 km.

bootstrapping Use of local materials to manufacture a significant fraction of the tools to make additional manufacturing facilities.

energy The capacity to do work, E, which is formerly determined as the scalar product of the vector quantities force and distance ($E = \mathbf{F} \cdot \mathbf{s}$); it is measured scientifically in joules (J), and large quantities are often quoted in pentajoules (1×10^{15} J) and exajoules (1×10^{18} J).

power Work per unit time and formerly measured in watts ($1\,\mathrm{W} = 1\,\mathrm{J/s}$), kilowatts ($1\,\mathrm{kW} = 10^3\,\mathrm{W}$) characteristic of the household level of use, megawatts ($1\,\mathrm{MW} = 10^6\,\mathrm{W}$) for a community or small industrial activity, gigawatts ($1\,\mathrm{GW} = 10^9\,\mathrm{W}$) for use by a large city, terawatts ($1\,\mathrm{TW} = 10^{12}\,\mathrm{W}$) for use at the national or world level, and so on.

rectennas Many specialized radio antennas and diodes that are mechanically and electrically linked together into a large plane of rectifying antennas that receive microwave power and convert the incident microwave power into electrical power.

stand-alone power system Able to deliver a predictable and dependable flow of power without major ancillary equipment or alternative fuels or power sources.

Earth's inhabitants need, within 50 years, a new source of sustainable commercial power that provides at least 2 kilowatts of electricity per person or 20 terawatts globally. The new power system must increase Earth's resources and be independent of the biosphere, and the use of the power must not change the biosphere. Facilities built on the moon from lunar materials can capture a fraction of the 13,000 terawatts of solar power incident on the moon. This lunar–solar power system can deliver more than 20 terawatts of affordable electric power to Earth.

1. ISOLATED SPACESHIP EARTH AND THE HUMAN RACE TO SUSTAINABILITY

It is nearly impossible to comprehend the physical isolation of our tiny spaceship Earth and its cargo of humans and the other life within Earth's thin biosphere. To gain some sense of it, consider the Voyager 1 spacecraft, launched from Earth in September 1977. It is now traveling in a northward direction away from our solar system. Each year, it is 3.6 astronomical units (AU) farther away. On February 14, 1990, when Voyager 1 was approximately 11 AU away from our sun, it turned a long-range camera toward us and photographed our Earth, Venus, and our sun (Fig. 1A). The pale blue dot of Earth filled only 10% of one pixel of the camera (Fig. 1B).

Every living entity on Earth is totally dependent on sources of meager flows of geothermal power from within Earth—less than 30 terawatts of thermal power (TWt) and a tiny fraction of the approximately 175,000 terawatts of solar power (TWs) that intersects Earth. Life on Earth is totally subject to unpredictable major violent events. These include volcanic eruptions, sudden changes in regional and global climates, and impacts by asteroids and comets. Major volcanic eruptions can devastate large

 © 2004 Elsevier Inc. All rights reserved.

portions of a continent and quickly affect the atmospheric greenhouse. Earth's biosphere is affected over millions of years by changes in the orbit of Earth that change the solar power Earth receives. Its orbit around our sun is affected by the changing gravitational forces exerted on Earth by the sun, Jupiter, Saturn, and the other smaller planets. Earth was formed 4.6 billion years ago by the in-falling gas, dust, comets, asteroids, and primitive moons. The infall continues. Several metric tonnes of meteors are vaporized each day in Earth's upper atmosphere. Approximately once a year, a larger meteor releases the energy of a small nuclear bomb. It is projected that, on average, regional-scale disasters due to impacts of multi-hundred-meter-diameter objects will occur every 10,000 years, civilization-ending impacts by multi-kilometer-diameter objects will occur every 1 million years, and global cataclysms will occur every 100 million years. Such a collision 64 million years ago ended the age of dinosaurs and enabled small mammals to proliferate. One lineage evolved into humans.

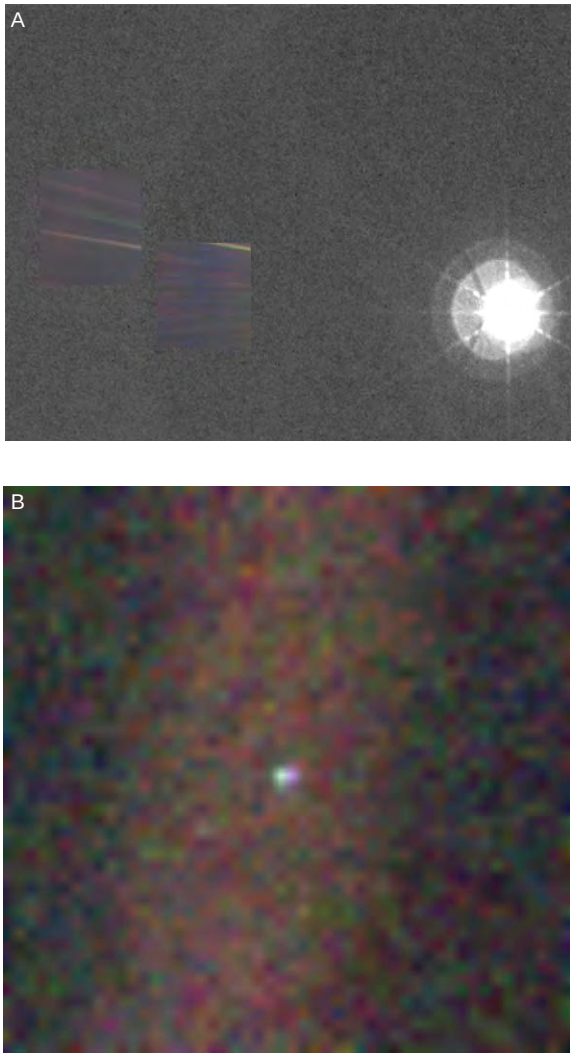


FIGURE 1 (A) Composite photograph of Earth (artifact-lined insert on left), Venus (middle artifact-lined insert), and our sun by the Voyager 1 spacecraft on February 14, 1990, at a distance of approximately 11 AU from the sun. (B) Image of Earth magnified to the greatest possible extent. It is a single pixel located at approximately the middle of the third "artifact" line from the top in the left-hand insert in panel A. Photographs extracted from the Voyager spacecraft file PIA00452 of the Jet Propulsion Laboratory.

The lifeless moon is our closest neighbor (Fig. 2). Surface temperatures near the equator of the airless moon vary from a daytime high of approximately 380°Kelvin (K) (107°C) to a predawn low of 120°K. A few craters near the lunar poles contain cold surfaces ($\sim$80°K) that have never seen the sun. Earth appears just above the horizon of the moon. Unlike the moon, Earth was large enough to gravitationally retain the water and gases that form its oceans and atmosphere. Earth's atmosphere provides a protective blanket, like that of a greenhouse, that recirculates about Earth a fraction of the solar power that would otherwise be reradiated quickly and directly back to space, as occurs from the lunar surface. This recirculation of solar heat provides Earth with a global mean temperature of approximately 295°K (22°C) within which water is liquid and water-based life can exist and evolve. Earth's atmosphere and

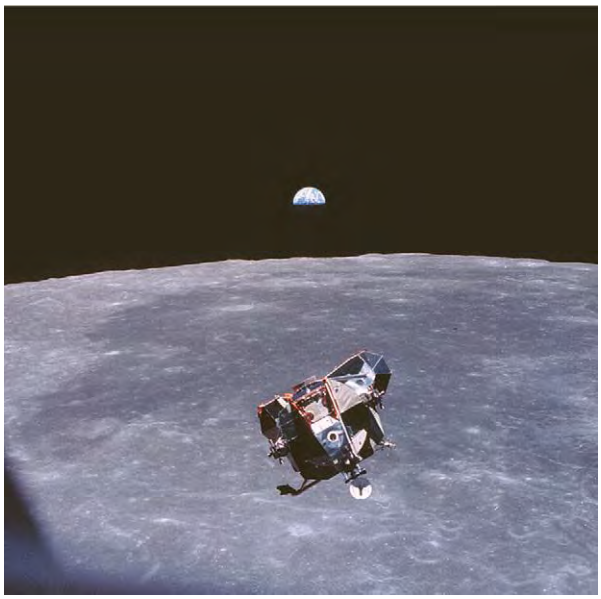


FIGURE 2 Photograph of Earth, moon, and Apollo 11 lunar module returning from the moon to the command service module in 1969. From NASA, ID AS11-44-6642.

oceans operate as a heat engine, driven by the sun, to transport this recirculated solar heat about the planet. This process evens out the temperature differences between night and day and among the equator, mid-latitude deserts (extreme of $\leq 331^{\circ}$K), and poles (extreme of $\geq 184^{\circ}$K). Days- to weeks- to years-long variations of the recirculating atmosphere at any one point on Earth are termed "weather." Decades- and centuries-long patterns in the recirculating atmosphere and oceans are termed "climate."

Carbon dioxide is the minor noncondensing and relatively long-lived component of the atmosphere that controls the greenhouse over centuries. The effectiveness of Earth's greenhouse varies over shorter time scales due to the concentrations of water vapor, methane, and other minor constituents that vary quickly in time, latitude, longitude, and altitude. Over the life of Earth, its greenhouse has adjusted to approximately a 30% increase in the output of solar power and changes in Earth's orbit. However, the greenhouse is not perfect. Major and "minor" ice ages occur. The external and internal forces that drive climate change are not fully understood. Interactions between the drivers are even less understood. However, one thing is clear: human industry is now changing the face of the planet. Human commercial power systems now output the equivalent of approximately 14 TWt of thermal power and are rapidly consuming the relatively small sources of fossil energy accumulated over a period of more than 100 million years from the ancient biosphere. Also, human power systems and industry are degrading, in many ways, the ability of the living biosphere to support us. In particular, our commercial power systems are increasing the concentration of greenhouse gases in the atmosphere at a faster rate than have any previous natural processes. The inhabitants of the uncontrolled spaceship Earth (note Figs. 1B and 2) need a new source of sustainable abundant power that is independent of the biosphere and whose use does not fundamentally change the biosphere. This new source of power must be started, at industrial scale, within the next two decades and come to maturity within 50 years.

2. THE IMMEDIATE GLOBAL ENERGY CHALLENGE

The World Energy Council Statement 2000 issued this energy challenge:

> *Slightly more than 1 billion people in the industrialized countries (about 20% of the world's population) consume nearly 60% of the total energy supply, whereas just under 5 billion people in developing countries consume the other 40% of the total energy supply. … The 2 billion poorest people ($1000 annual income per capita or less), a small but growing share of whom live in shantytowns with most still scattered in rural areas, use only 0.2 toe of energy [tonnes of oil-equivalent of thermal energy] per capita annually, whereas the billion richest people ($22,000 annual income per capita or more) use nearly 25 times more at 5 toe per capita annually [5 toe/person-year ~6.7 kilowatts of thermal power [kWt]/person].*

In 2002 the UN World Summit on Sustainable Development challenged the world to "diversify energy supply and substantially increase the global share of renewable energy sources in order to increase its contribution to total energy supply."

Over the past century, the economic output of a unit of commercial energy has increased at approximately 1% per year. If this trend continues, the current 6- to 7-kWt/person power usage by Western Europeans and Japanese might decline by approximately 50% for the same level of economic output. Thus, many argue that per capita global energy use will decline significantly over the 21st century. Unfortunately, this is not a realistic prospect for a sustainable global economy. Both the developed and developing economies obtain significant quantities of energy embedded as fresh air and water and energy-rich biomass. As world population increases, the per capita quantities of these biospheric resources decline. The current economies obtain necessary chemicals from mines in which economically important chemicals have been concentrated by ancient geological processes. In 1976, Goeller and Weinberg estimated the per capita power necessary, assuming the use of non-fossil fuels systems such as breeder reactors, to indefinitely sustain the 1960s U.S. and global economies. This advanced sustainable economy will obtain its fresh water from desalination. All of the agricultural, process, and industrial chemicals will be obtained at their average abundance in the crust of the earth. Goeller and Weinberg projected that a sustainable 1968-style U.S. economy will require approximately 15 kWt/person and that a 1960s global economy will require 7.5 kWt/person. Global power consumption for 10 billion people of 75 to 150 TWt is implied. Ayres and colleagues demonstrated that economic growth of the U.S. economy, from 1900 to 1998, was proportional to the total "useful work" accomplished. Useful work, approximately a product of total input exergy (i.e., available energy) and effectiveness of its use, is now the dominant factor in U.S. economic growth. National economies are only now beginning to face

the challenges of providing new levels of useful work to provide new goods and services that are essential to sustainable development. For the following discussion, it is assumed that ingenuity and economic adaptability will enable humankind to provide sustainable existence using only 6 to 7 kWt/person. However, it is likely that human needs and desires will require a greater level of per capita power.

The new power must be clean and significantly less expensive than it is now. The developed countries expend approximately 10% of their gross domestic product (GDP) on all phases of commercial energy. They now have a per capita GDP of approximately $22,000/person-year. By analogy, if the developing countries spend approximately 10% of their GDP on all phases of providing 6.7 kWt/person, the new source of power must initially supply thermal energy for less than 0.4¢/kWt-h. The new primary energy source and the new power system must be adequate for centuries and should have significant capacity for growth. Finally, it is practical to assume that the new commercial power system will use physical resources and technologies that are relatively well understood at this time.

Together, the developed and developing countries now consume the equivalent of approximately 14 TWt of thermal energy. To provide 10 billion people with 5 toe/person-year implies the global production of 67 TWt of commercial thermal power by approximately 2050. This corresponds to consuming the equivalent of approximately 900 billion barrels of oil per day or 11 times current global production. An energy-rich 10 billion people will consume 6700 TWt-years of energy per century. Commercial power is used to provide goods (e.g., steel, concrete) and services (e.g., fuel for transportation or heating). In 1880, only thermal power was available. However, by 1900, commercial electric power became available. By 2000, electricity powered 25% of the final production of goods and delivery of services. Electricity is projected to rise to approximately 100% of final power by 2050. This is because electricity, where it can be applied, is approximately three times more productive than thermal power. By 2050, approximately 20 to 30 terawatts of commercial electric power (TWe) will likely be as economically productive as 60 to 67 TWt of commercial thermal power.

The first column of Table I lists 25 options for generating global commercial power. The options are divided into five major categories: traditional fuels from the living and fossil biosphere, nuclear fuels, geothermal power, solar-derived power on Earth, and solar power supplied from facilities in space. The second column lists the equivalent total thermal energy, measured in terawatt-years, that can be provided by each option. The third column lists the estimated maximum electrical power, in terawatts, that can be provided by each option by 2050. The fourth column lists the anticipated trend in pollution production to 2050 for each power option. The final column lists the anticipated cost of electric energy in 2050 versus 1¢/kWe-h. Criswell provided considerable details on the analyses of these options and references.

Power flow through the living biosphere (row 1 in Table I) is simply too small to support a 20 TWe or 60 TWt commercial power system. Peat (row 2) and oil and gas (row 4) do not provide sufficient fuels. Coal (row 3) can potentially provide the order of 4500 TWt-year of fuel. However, a prosperous global economy would consume that fuel in approximately 70 years. It is estimated that natural gas hydrates (row 5), buried in the ocean silt along the margins of the continents, may contain the order of 10,000 TWt-year of thermal energy. However, the hydrate deposits are thinly spread over vast areas. Mining them likely will be expensive and disrupt the sea floor.

Conventional nuclear reactors that use uranium mined from the continents (row 6 in Table I) are projected to supply less than 430 TWt-year of commercial thermal energy. Breeder reactors can increase the total energy output by approximately a factor of 60. However, fission breeder reactors (rows 7 and 8) appear to be unacceptable politically and very expensive. Continuous net energy output from controlled nuclear fusion of deuterium (D) and tritium (T) is not yet demonstrated. It will be decades before the engineering and commercial characteristics of this "least challenging" fusion technology can be specified (rows 9 and 10). Fusion burning of helium-3 (^{3}He) with D (row 11) requires much higher temperatures and pressures. Again, it likely will be decades before commercial D–^{3}He fusion power plants can be engineered. In addition, the lunar soils are the only potential source of ^{3}He near Earth. The total quantity of ^{3}He that is potentially available is not known, and the economic efficiency with which it can be mined will require extensive demonstrations on the moon. Huge areas of the lunar surface would be totally disrupted by the mining operations.

Geothermal power (row 12 in Table I) is driven primarily by the slow radioactive decay of uranium (U) and thorium (T) within Earth. The original

TABLE I

Global Power System Options for 2050

Power system option	Fuel resource TWt-year thermal	2050 output electric power (TWe)	2050 pollution vs now	2050 cost vs 1¢/kWe-h
Biosphere and fossil				
1. Bioresources	<230	<0.2	More	$>$
2. Peat	<60	~ 0	More	$>$
3. Coal	$<4{,}500$	<4	Large	$>$
4. Oil and gas	$<1{,}300$	<8	Large	$>$
5. Natural gas hydrates	TBD $>10{,}000$	TBD	Large	Likely $>$
Nuclear				
6. Fission (no breeder, high-grade uranium ores)	<430	<1.5	Large	$>$
7. Fission (breed 238U/T)	$<33{,}000$	in #6	Large	$>$
8. Fission (breed ocean U)	$\sim 6{,}000{,}000$	in #6	Large	$>$
9. Fusion (D–T/U-Th)	$<6\times10^9$	in #6	Large	Likely $>$
10. Fusion (D–T)	$\gg 1\times10^9$	0 likely	More	Likely $>$
11. Fusion (D–^{3}He lunar)	? ~ 100 to 10^5	0 likely	More	Likely $>$
Geothermal				
12. Nuclear decay and accretional energy (0–7 km depth)	$\sim 9{,}000{,}000$	<0.5	Low	$>$
Solar within the biosphere				
13. Hydroelectric	<14	<1.6	Low	$>$
14. Salinity gradient to sea	$\sim 1{,}700$	<0.3	TBD	Likely $\gg$
15. Salinity gradient to brine	$\sim 24{,}000$	<0.3	TBD	Likely $\gg$
16. Tides	0	<0.02	Low	$>$
17. Ocean thermal	$\sim 200{,}000$	<0.1	Large	$\gg$
18. Wind (not stand-alone)	0	<6	Low	$>$
19. Terrestrial thermal solar	0	<3	TBD	$\gg$
20. Terrestrial solar photovoltaics	0	<3	TBD	$\gg$
Mixed conventional system				
21. (1, 3, 4, 6, 12, 13, 16, 18, 19, 20)	$\leq 3{,}200$	11($\sim$33 TWt)	Large	$>$
Solar power from space				
22. GEO solar power satellites (deployed from Earth)	0	<1	Low	$\gg$
23. LEO solar power satellites (deployed from Earth)	0	<0.1	Low	$\gg$
24. SPS beyond GEO (made of nonterrestrial materials)	0	<1	Reduce	Likely $\geq$
25. Lunar–solar power system (made on the moon)	0	≥ 20	Reduce	Likely $\leq$

Note. t, thermal; e, electric; 3 Wt $\sim$ 1 We in utility; TBD, to be determined; GEO, geosynchronous earth orbit; LEO, low earth orbit; SPS, solar power satellite.

materials that formed Earth also contributed a portion of their gravitational potential energy. The entire Earth releases only approximately 30 TWt of thermal energy. Most is released from the inaccessible sea floor and along the spreading boundaries between plates. Only a fraction of a terawatt is released as high-quality steam at continental sites such as the geysers in California. Large-scale geothermal power systems must mine the heat content of rocks between the surface of Earth and to a depth of less than 7 km. This shell of continental rocks contains less than 9 million TWt-years of thermal energy. However, it will be expensive to mine this relatively low-quality heat. Maintaining large-scale circulation of water through mined rocks is not demonstrated. Power installations on the surface will release considerable waste heat into the environment. The geological consequences of long-term mining of more than 600 TWt from this outer shell of Earth have not been considered.

Our sun delivers approximately 175,000 TWs of solar power to the disk of Earth. Unfortunately, the terrestrial renewable systems listed in Table I, such as wind (row 18) and solar (rows 19 and 20), are not stand-alone and must be supplemented by other power systems of similar capacity. This drives up the cost of the delivered energy significantly. Also, very large-scale renewable systems can directly affect the biosphere by changing the natural flows of power, water (row 13), and wind (row 18) and by modifying the reflective and emissive properties of large areas of Earth.

Expansion of Earth's existing commercial mixed power system in row 21 of Table I (a combination of coal, oil, natural gas, biomass, nuclear, hydroelectric, and miscellaneous renewables) can likely provide less than 3200 TWt-years of commercial energy due to economic and environmental constraints and a maximum output by 2050 of less than 11 TWe. In this mixed system, coal (row 3) and conventional fission (row 6) systems would exhaust their estimated fuels a few decades into the 22nd century. Even assuming increasing efficiency in economic output of a given unit of energy, the expensive mixed power system would not enable global prosperity by 2050. It must be noted that, under the modeling conducted by the International Institute for Applied Systems Analysis, decreasing the use of fossil fuels and increasing the use of renewable power systems resulted in decreased economic growth and less available per capita power.

Sustainable global energy prosperity requires low-cost, dependable, and direct access to solar power. Our sun, in Fig. 1A, is the dominant source of power and energy for Earth. It outputs 3.9×10^{14} TWs of solar power. It is powered by the nuclear fusion of hydrogen into helium and heavier elements, and it now converts approximately 1.4×10^{14} metric tonnes/year of hydrogen and helium into energy. Over the next 5 billion years or so of stable operation, the sun will release at least 2×10^{24} TWs-year of energy.

Commercial-scale solar power satellites, deployed from Earth (rows 22 and 23 in Table I) or the moon (row 24), will likely not be demonstrated until well into the 21st century. A huge fleet of extremely large satellites in orbit around Earth will be unacceptable due to concerns over safety and disruption of the night sky. Orders of magnitude improvements are required in materials, systems performance, assembly and maintenance, and transportation to space before such systems could provide commercially competitive electric energy to Earth. It appears that 24 of the power options cannot provide sustainable, pollution-free, and affordable global power by 2050.

The answer to supplying adequate, sustainable, and affordable commercial power is the moon. Earth's moon (Fig. 2) is dependably illuminated by 13,000 TWs of solar power. It is argued in the following sections that a reasonable fraction of the solar power incident on the moon can be sustainably delivered to Earth as commercial electric power at significantly less than 1¢/kWe-h.

3. LUNAR–SOLAR POWER SYSTEM

Figure 3 illustrates the essential features of the lunar–solar power (LSP) system: the sun, the moon and power bases (the bright spot on the left side and the dim spot on the right side of the moon), microwave power beams from each power base on the moon, relay satellites in orbit around Earth (not shown but discussed later), and a microwave receiver (i.e., rectenna) on Earth. Figure 4 shows the moon after the construction of an LSP system scaled to supply more than 20 TWe to Earth. There are 10 pairs of power bases on opposing limbs of the moon. The power bases receive sunlight, convert it to electricity, and then convert the electric power into microwave power beams. Each base transmits multiple microwave power beams directly to the rectennas on Earth when the rectennas can view the moon. Power beams between Earth and the moon are not esoteric. The Arecibo Radio Telescope in Puerto Rico has routinely beamed microwaves from Earth to the moon. Its beam intensity is approximately 20 W/m^2 going upward through the atmosphere. This is 10% of the maximum intensity ($\leq$230 W/m^2) proposed for transmission of commercial power.

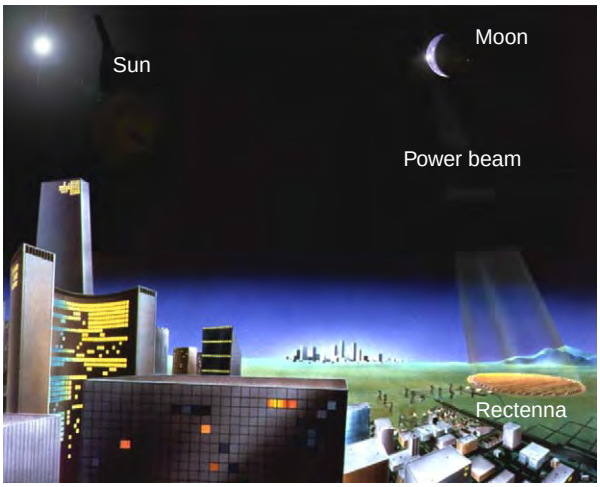


FIGURE 3 Sun, moon, beam, and rectenna.

FIGURE 4 Ten pairs of LSP bases.

Bases on the earthward side of the moon can be augmented by fields of photoconverters just across the far side of the moon from Earth. Power lines connect each earthward base and the extra arrays of photoconverters on the far side of the moon. One or the other of the two bases in a pair will receive sunlight over the course of a lunar month. Thus, a pair of augmented power bases can beam power toward Earth over the entire cycle of the lunar day and night.

The rectennas on Earth that receive the power are simply specialized types of centimeter-scale television antennas and electric rectifiers. A rectenna is illustrated in the lower right of Fig. 3. A rectenna converts the microwave beam it receives at more than 85% efficiency into electricity and outputs the pollution-free power to local electric distribution systems and regional grids. Most long-distance power lines will not be necessary. Rectennas are the major cost element of the reference version of the LSP system.

Microwave power beams, approximately 12 cm wavelength or 2.45 GHz, pass from the moon through Earth's atmosphere, clouds, fog, snow, smog, smoke, and normal rain with 1% or less of attenuation. An extremely heavy ($\geq$25 cm/h) rain will attenuate the beam by less than 30%. This level of rain is very rare in most regions and lasts for only a few tens of minutes. Beam intensity can be increased to maintain the electric output of the rectenna. Other frequencies can be used but will not be as efficient. A rectenna that receives a load-following power beam of less than 230 W/m^2 will, over the course of a year, output approximately 200 W/m^2 of electric power. The rectenna will output this power whether it is located in a desert, the tropics, or a polar region. In contrast, stand-alone solar arrays on Earth output much less average power per unit of area. A stand-alone solar array is one that feeds power directly to a grid and also to a storage system so that power can be provided during the night or when the sky is obscured. A stand-alone solar array on Earth will have an annually averaged output of less than 3 W/m^2 if it uses 1980s technology and of less than 20 W/m^2 if it uses advanced technologies. Solar arrays on Earth are captives of the biosphere, season, and weather, whereas the power output of the rectenna is independent of these limitations.

Rectennas on Earth can view the moon and receive power only approximately 8 h each day. Earth-orbiting satellites can redirect beams to rectennas that cannot view the moon. The redirectors enable load-following power to rectennas located anywhere on Earth. Rectennas on Earth and the lunar transmitters can be sized to permit the use of Earth-orbiting redirectors that are 200 to 1000 m in diameter. Redirector satellites can be reflectors. Alternatively, a relay satellite can receive a power beam from the moon and then retransmit several new beams to different rectennas on Earth. Unmanned and manned spacecraft have demonstrated the transmission of beams with commercial-level intensity in low Earth orbit. Demonstration-scale reflectors and retransmission technologies have been, and are now, operating in space.

Power beams that will be 1 to 20 times more intense than the maximum that is recommended for continuous exposure by the general population are proposed. Each tightly focused beam will be directed to rectennas that are industrially zoned to exclude the general population. Microwave intensity under the rectennas will be reduced to far less than is permitted for continuous exposure of the general population. Secondary electrical shielding can reduce the microwave intensity to a negligible level. The power beams do not pose a hazard to insects or birds. Aircraft can simply fly around the beams. In addition, passengers will be shielded by the metal skin of their aircraft. Beams can be turned off in a few seconds or decreased in intensity to accommodate unusual conditions.

Earth can be supplied with 20 TWe by several thousand rectennas whose individual areas total approximately 100,000 km^2. Rectennas placed over agricultural land and industrially zoned facilities enable multiple uses of the same land. Rectennas can be as large in area as is necessary to produce larger electric power output. Conventional electric power systems use far larger total areas. In the cases of strip-mined land or power line rights-of-way, energy production degrades the land for several years to decades and/or precludes multiple uses. Rectennas could be placed over such land and can be located relatively close to major power users and, thereby, minimize the need for long-distance power transmission lines. A rectenna as small as approximately 0.5 km in diameter, with an output of 40 MWe, is feasible.

Unlike conventional power systems, the LSP system supplies net new power to Earth. Its net new electricity can be used to power the production and recycling of goods and the provision of services and to restore the biosphere. LSP energy replaces the tens of billions of tonnes of oil, coal, and natural gas that are now burned each year and the hundreds of billions of tonnes of wind and flowing fresh water that would otherwise be diverted from their natural courses to produce a much smaller output of electricity. The LSP system requires very little continuous processing of mass on Earth to receive and distribute the electric power. The fundamental waste product of lunar–solar power on Earth is waste heat that is eventually converted into useful work and then into infrared photons and radiated back to space.

Figure 4 depicts 20 power bases, 10 on each limb or visible edge of this full moon. They are adequate, assuming 1980s technology, to provide 20 TWe of electric power to Earth. More advanced technology will enable bases that are so small they will be unobservable to the naked eye. The moon receives 13,000 TWs, so the LSP has considerable growth potential.

Approximately 14 days after the view in Fig. 4, the opposite or far side of the moon will be fully illuminated by the sun. Additional fields of solar converters located approximately 500 km across the limb of the moon from each base in Fig. 4 will receive full sunlight. These cross-limb bases of solar converters can be connected by power lines to the bases on the earthward side of the moon. Approximately once a year, the moon will be totally eclipsed by Earth for up to 3 h. It is reasonable to precisely plan for the amount of additional LSP energy that must be provided to Earth for use during a lunar eclipse. Continuous power can be provided on Earth by storing solar-derived energy on the moon and transmitting it to Earth during the eclipse. Alternatively, predictable amounts of energy can also be stored on Earth and released during the short lunar eclipse. Eclipse power can also be produced on Earth from conventional systems such as natural gas-fired turbines. Finally, mirrors, similar to proposed solar sails, could be placed in orbits high above the moon and reflect solar power to the bases during the eclipse. Such mirrors eliminate the need for expensive power storage.

4. LUNAR–SOLAR POWER DEMONSTRATION BASE

Figure 5 illustrates a demonstration power base. A power base is a fully segmented, multibeam, phased-array radar powered by solar energy. This demonstration power base consists of tens to

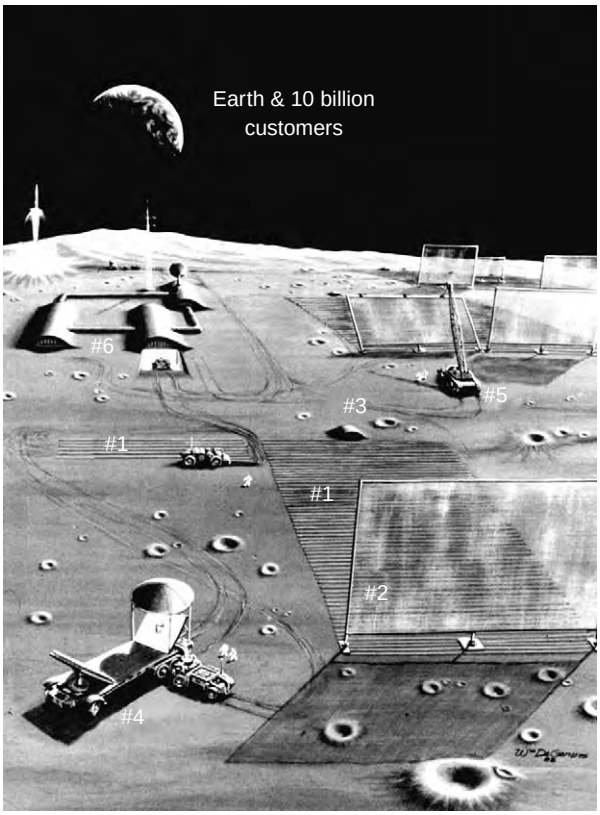


FIGURE 5 LSP demonstration base: multiple power plots [arrays of solar converters (♯1), microwave reflector (♯2), and microwave transmitter (♯3)], set of mobile factory (♯4) and assembly units (♯5), and habitat/manufacturing facility (♯6).

hundreds of thousands of independent power plots. A power plot is depicted in the middle to lower right portion of Fig. 5. Each power plot emits multiple sub-beams.

A power plot consists of four elements. There are arrays of solar converters (#1 in Fig. 5), shown here as north/south-aligned rows of photovoltaics. Solar electric power is collected by a buried network of wires and is delivered to the microwave transmitters. Power plots can use many different types of solar converters and many different types of electric-to-microwave converters. In this example, the microwave transmitters (#3) are buried under the mound of lunar soil at the earthward end of the power plot. Each transmitter illuminates the microwave reflector (#2) located at the anti-earthward end of its power plot.

Earth is fixed in the sky above the power base. The many small reflectors (#2 in Fig. 5) can be arranged in an area on the limb of the moon so that, when viewed from Earth, they appear to form a single large aperture as depicted by one of the power bases in Fig. 4. This enables a given power base to form very narrow and well-defined beams directed toward Earth. The moon has no atmosphere and is mechanically stable. There are no moonquakes. Thus, it is reasonable to construct the large lens from many small units. Individually controllable sub-beams send their microwave sub-beams toward each small reflector. The sub-beams are correlated to combine coherently on their way toward Earth to form one power beam. A mature power base can have hundreds to a few thousand sets of correlated microwave transmitters illuminating each reflector. The multiple reflectors will likely include additional sub-reflectors or lenses in front of each main reflector.

To achieve low unit cost of energy, the lunar portions of the LSP system are made primarily of lunar-derived components. Both the fixed (#6 in Fig. 5) and mobile (#4 and #5) factories are transported from Earth to the moon. Their high output of LSP components made on the moon from lunar materials greatly reduces the impact of the high cost of transporting the factories from Earth to the moon. Lunar factories will produce hundreds to thousands of times their own mass in LSP components.

Table II characterizes an LSP demonstration base. A period of 10 years is required to plan a demonstration base and establish it on the moon. Three sizes of base are modeled. Absolute costs are less important than the trend of cost versus the total power put in place after 10 years of operations. The machinery of the smallest base (column 1) installs 1 GWe of power that is delivered to Earth (1×10^9 We). A total cost of $60 billion is predicted, as measured in 1990 U.S. dollars. Cost of the LSP production equipment is $12 billion (1.a + 1.b). This base is estimated to have a mass of 2300 tonnes and requires 30 people on the moon. Electricity sold on Earth at $0.1/kWe-h generates $4.4 billion of revenue. The largest base (column 3) is scaled to install 100 GWe received on Earth. Total cost increases by a factor of 4. However, the sales of electric energy increase by a factor of 100. The largest base pays for itself. Cost of LSP production equipment, the mobile units in Fig. 4, increases by only a factor of 2. The production process will continue after the 10-year demonstration. Total power capacity (energy delivered to Earth) will grow, and net profit will continue to grow. Energy cost will drop steadily. The demonstration base can be established in

TABLE II

Cost (1990 Dollars) of LSP Demonstration Base

	1 (smallest)	2	3 (largest)
GWe installed after 10 years on the moon	1	10	100
GWe-years of energy to Earth over 10 years	5	50	500
Gross revenue (billions of dollars) ($0.1/kWe-h)	4.4	44	438
Net revenue (billions of dollars)	−56	−47	195
Total costs (billons of dollars) (sum 1 + 2 + 3)	60	91	243
1. Research and development (sum a + b + c + d)	42	51	86
a. LPS hardware (billions of dollars)	11	11	11
b. Construction system (billions of dollars)	1	3	11
c. Facilities and equipment (billions of dollars)	5	10	30
d. Transport (billions of dollars)	26	27	35
2. Space & Operations (billions of dollars)	17	34	103
3. 1970s reference rectennas on Earth (billions of dollars)	0.6	6	55
Dollars/kWe-h	1.4	0.2	0.06
Moon (tonnes)	2300	6200	22,000
Space (tonnes)	970	2700	9700
People (moon, LLO, and LEO)	30	85	300

LLO, low lunar orbit; LEO, low earth orbit.

less than 10 years. The size of the base and the production process can be steadily increased in size to enable 20 TWe delivered to Earth by 2050.

5. SCALE AND COST FOR 20 TWE LUNAR–SOLAR POWER SYSTEM

A full-scale LSP system, based on 1980s technology, would occupy approximately 15% of the surface area of the moon. This is illustrated in Fig. 4. The total area of all the bases in Fig. 4 appears to be small compared with this disk of the moon. However, this is an illusion. The area of each base curves around the spherical surface of the moon near the lunar limb as seen from Earth. Each base is six times longer in the east–west direction than it is wide in the north–south direction. As illustrated in Fig. 5, approximately 80% of the area would be the empty ground between the solar arrays. Using solar collectors on the far side of the moon and eliminating the solar mirrors in orbit around the moon increases the area of the 1980s-style bases to approximately 25% of the lunar surface. Technologies anticipated to be available in 2020, all demonstrated by advanced systems of today, and close packing of the solar cells on the surface of the moon reduce the area occupied by an advanced LSP system to less than 0.2% of the lunar surface.

Table III presents estimates of the tonnage of equipment that would be deployed from Earth over 70 years to deliver 20 TWe of electric power on Earth and the number of people required in space and on the moon. The estimates assume that a period of 10 years is required to place the production facilities on the moon. Next, there is 30 years of full-scale production that requires 53,000 tonnes/year of components and supplies launched to the moon. The model includes the additional equipment and materials needed to maintain the bases and rebuild 50% of the power collectors and transmitters between 2050 and 2070. "LSP REF" assumes that all production equipment, people, and supplies are taken from Earth to the moon. "10 × LSP REF" assumes that the equipment taken to the moon and used to make the power plots are 10 times more massive than, or one-tenth as productive as, the equipment in the LSP REF column. The "BOOT 90%" column assumes that 90% of the mass of the production equipment is made on the moon of lunar materials.

In the LSP REF model (column 3 in Table III), over the 70-year life cycle, the micromanufacturing

TABLE III

Life Cycle Cost for Heavy LSP, Reference LSP, and Bootstrapped LSP

	10 × LSP REF	LSP REF	BOOT 90%
Tonnes of lunar equipment			
Micro manufacturing	3,205,323	250,605	24,361
Hot forming	1,031,271	103,127	10,313
Beneficiation	321,229	32,123	3212
Habitats, shops, and mobile units	75,057	23,341	22,085
Chemical refining	246,945	24,941	2469
Gather and eject to orbit	43,185	4383	438
Excavation	83,115	831	80
Cold assembly	2776	278	28
Total (tonnes)	5,012,039	442,630	62,915
Number of people			
Moon	55,915	4717	436
Lunar orbit	4986	468	59
Earth orbit	5010	443	63
Cost of equipment and people (1977 dollars) for 20 TWe and 1000 TWe-year to Earth	$19.5 trillion	$1.75 trillion	$0.26 trillion
Energy engineering cost (dollars/kWe-h)			
Reference rectenna ($7.9 trillion)	0.0106	0.0037	0.0037
Reflective rectenna ($0.82 trillion)	0.0078	0.001	0.0004

of solar converters and similar products requires approximately 57% of the equipment mass or 250,000 tonnes. Approximately 23,000 tonnes of habitats, shops, and other support facilities equipment is required to support 4700 people needed during full production of 0.6 TWe/year of new power capacity. The intense and variable solar and cosmic radiation makes it impossible for people to work on the surface of the moon for long periods. A covering of approximately 3 m of lunar soil is required to protect people. Most surface operations are automated and/or controlled by people in habitats, as shown on the left side of Fig. 5.

The engineering cost in Table III is presented in 1977 U.S. dollars and does not include financing cost. The engineering cost of $1.75 trillion includes all lunar- and space-related activities and launch and manufacturing on Earth. The National Aeronautics and Space Administration (NASA) and U.S. Department of Energy (DOE) studies conducted during the 1970s were used to calculate the engineering cost of 20 TWe of rectennas at $7.9 trillion. Rectenna construction and operation is five times greater than the cost of all the lunar-related operations. A 20-TWe LSP system that delivers 1000 TWe-year of energy over a life cycle of 70 years is projected to achieve an average cost of $0.0037/kWe-h (0.37¢/kWe-h). Assuming a factor of three for inflation from 1977 to today, LSP REF electricity would be competitive with the lowest cost existing conventional power systems that provide electricity at a wholesale or busbar cost of $0.02 to $0.04/kWe-h. Power bases can be maintained indefinitely at far lower cost than is required to develop and construct them. Rectennas on Earth that use reflective concentrators will reduce the number of expensive rectifiers and also be less subject to mechanical loads from wind, rain, and ice. The projected cost of 20 TWe of reflector rectennas is approximately $0.8 trillion. The engineering cost of energy is reduced to $0.001/kWe-h. Over time, lower costs are possible. Thus, the LSP REF system can potentially provide electric energy at 10% or less of the cost of existing systems.

Table III assumes that the cost of launch from Earth to orbit is $470/kg (in 1977 U.S. dollars). Increasing the cost of Earth-to-orbit transportation to $5000/kg increases the upfront cost significantly, and life cycle engineering cost increases to approximately $5.2 trillion. However, the life cycle engineering cost of electricity from the LSP REF system increases by only approximately 20%. The system model assumes the extensive use of lunar-derived propellants. Solar electric propulsion is used for the transport of cargo from orbit around Earth to orbit around the moon.

The "10 × LSP REF" in Table III assumes that all of the production machinery must be 10 times more massive than that for LSP REF to achieve the same output of LSP components. Less productive manufacturing equipment (column 2), with lower output, will increase the size and cost of the lunar and space operations. Total tonnage of production machinery on the moon increases by a factor of 11. Engineering life cycle cost increases to approximately $20 trillion over the 70 years. However, the engineering cost of the electric power on Earth increases by only a factor of 3, to $0.0106/kWe-h, for the reference rectenna and to $0.0078/kWe-h for the reflective rectenna. The "10 × LSP REF" is still competitive with electricity from existing suppliers.

P. Glaser originated the concept of using large satellites in geosynchronous orbit around Earth to capture sunlight, change it into microwaves, and beam the power to Earth. During the 1970s and 1990s, NASA and DOE studied the deployment of large solar power satellites (SPS) from Earth to geosynchronous orbit. For delivery of energy to Earth, the LSP REF is projected to be at least 1000 times more efficient in its use of mass deployed from Earth than is deployment of an equal mass as an SPS.

On Earth, we make all of our tools for manufacturing and production from materials on Earth. This is called "local manufacturing" or "bootstrapping." The industrial materials obtained from lunar soils (e.g., glasses, ceramics, iron, aluminum, titanium, sulfur, silicon) can also be used on the moon to manufacture, as shown in Fig. 5, a large fraction of the mobile factories (#4 and #5), shops (#6), and habitats (#6). Of course, the production machinery in Fig. 5 must be designed for optimal use of parts manufactured on the moon. The bootstrapping equipment and operations will be fully refined during the demonstration phase.

The final column of Table III assumes that 90% of the mass of production equipment (#4, #5, and #6 in Fig. 5) is "booted" from lunar materials. The model is adjusted for the additional people needed to conduct this extra level of manufacturing. The model is also adjusted for a higher level of support from Earth in the form of remote monitoring and teleoperation. Only 63,000 tonnes of equipment is shipped to the moon over the period of 70 years. If components and supplies can be made entirely of lunar materials, then BOOT 90% enables the delivery of 1000 TWe-year to Earth at a specific energy of 63 tonnes/TWe-year. Engineering cost (in

1977 U.S. dollars) of electric energy could decrease to $0.0004/kWe-h (0.04¢/kWe-h). The production and logistics systems are assumed to make extensive use of tele-operation from Earth. The mass of facilities in orbit around Earth and the moon and the number of people are decreased significantly.

Major funding is required to build the LSP system, but it is reasonable when compared with global expenditures on energy and space to date. Today, the world expends approximately $4 trillion/year on all aspects of the 14 TWt global power system. Petroleum companies expend approximately $130 billion/year exploring and bringing to production new oil and gas fields. This continuing investment provides less than 6 TWt of commercial thermal power derived from petroleum. Global oil and natural gas production will likely peak within 10 years. The United States now expends more than $20 billion/year on civilian and military space activities. The current infrastructure of NASA, which can be adapted to enable the demonstration LSP, results from a total investment more than $600 billion since 1958.

Bootstrapping enables the exponential growth of the LSP power bases on the moon. Bootstrapping facilities are brought to the moon until sufficient manufacturing capacity is established to provide 20 TWe by 2050. Once full-scale production of power is achieved, many of the bootstrapping facilities can be directed to providing a wider range of new lunar goods and to enabling new services.

The Apollo- and Soviet-era lunar programs and the worldwide post-Apollo lunar research used well over $1 billion to provide the essential knowledge of lunar resources and the lunar environment. Post-Apollo satellites and deep space missions have mapped the major mineral and chemical composition of most of the moon.

The International Space Station program provides the international level of cooperation, technology, and operating systems that enables a quick return to the moon. A permanent international lunar base, established to conduct industrial research and development, can accelerate economic development of the moon. This can be done within a modest expansion of existing cash flow of the major civilian space programs. Automation and remote operations are a normal component of terrestrial industry. Major offshore oil and gas platforms use advanced capabilities, such as underwater robotics and complex operations and logistics, that can be employed on the moon and in space to build and maintain the lunar–solar power bases. Much work can be done by tele-operation and supervisory control from Earth.

6. CONCLUSION

The conventional energy resources of our isolated spaceship Earth (Figs. 1B and 2) are extremely limited and will be quickly consumed by a prosperous human race. The sun (Fig. 1A) is the only reasonable power source for a prosperous world. It contains adequate fuel, is a functioning fusion reactor that does not have to be built and operated, and retains its own ashes. The challenge is to build the transducers that can extract this power and deliver it to consumers on Earth at a reasonable cost. The moon receives 13,000 TWs of dependable solar power. The LSP system, built on the moon from lunar soils and rocks, is the transducer. The LSP system can deliver net new power to consumers on Earth that is independent of the biosphere and clean. Developing countries can afford LSP electricity. LSP electricity can accelerate the economic growth of all nations. Net new LSP electric power enables all nations to produce and recycle their goods and consumables independent of the biosphere. Transportation and services can be powered without consuming or affecting the biosphere. The challenges of the World Energy Council and the United Nations can be achieved by the year 2050.

SEE ALSO THE FOLLOWING ARTICLES

Earth's Energy Balance • Global Energy Use: Status and Trends • Lithosphere, Energy Flows in • Sun, Energy from • Sustainable Development: Basic Concepts and Application to Energy

Further Reading

Ayres, R. U., Ayres, L. W., and Ware, B. (2003). Exergy, power, and work in the U.S. economy, 1900–1998. *ENERGY: The Int. J.* **28**, 219–273.

Criswell, D. R. (1985). Solar system industrialization: Implications for interstellar migration. *In* "Interstellar Migration and the Human Experience" (B. R. Finney and E. M. Jones, Eds.), pp. 50–87. University of California Press, Berkeley.

Criswell, D. R. (2002). Energy prosperity within the 21st century and beyond: Options and the unique roles of the sun and the moon. *In* "Innovative Energy Strategies for CO_2 Stabilization" (R. Watts, Ed.), pp. 345–410. Cambridge University Press, Cambridge, UK.

Criswell, D. R. (2002). Solar power via the moon. *The Industrial Physicist,* April/May, pp. 12–15. (See letters to the editor and reprise in June/July, August/September, and October/November issues or at www.tipmagazine.com.)

Criswell, D. R., and Waldron, R. D. (1993). International lunar base and lunar-based power system to supply Earth with electric power. *Acta Astronautica* **29,** 469–480.

Clarke, A., and Trinnaman, J. (eds.). (1998). "Survey of Energy Resources 1998." World Energy Council, London.

Goeller, H. E., and Weinberg, A. M. (1976). The age of substitutability. *Science* **191,** 683–689.

Heiken, G. H., Vaniman, D. T., and French, B. M. (eds.). (1995). "LUNAR Sourcebook: A User's Guide to the Moon." Cambridge University Press, Cambridge, UK.

Murdin, P. (Ed.-in-Chief) (2001). "Encyclopedia of Astronomy and Astrophysics" Vols. 1–4. Nature Publishing Group, New York.

Nakicenovic, N., Grubler, A., and McDonald, A. (eds.). (1998). "Global Energy Perspectives." Cambridge University Press, Cambridge, UK.

Shirley, J. H., and Fairbridge, R. W. (eds.). (1997). "Encyclopedia of Planetary Sciences" (1st ed.). Chapman & Hall, London.

Weissman, P. J., McFadden, L-A., and Johnson, T. V. (eds.). (1999). "Encyclopedia of the Solar System." Academic Press, San Diego.

World Energy Council. (2000). "Energy for Tomorrow's World: Acting Now!" Atalink Projects, London.

Magnetic Levitation

DONALD M. ROTE
Argonne National Laboratory
Argonne, Illinois, United States

1. **Introduction**
2. **Characteristics of Attractive-Force-Based Suspension Systems**
3. **Characteristics of Repulsive-Force-Based Suspension Systems**
4. **Maglev Development History**
5. **Advantages of Maglev Technology for High-Speed Transportation**

Glossary

critical temperature The phase transition temperature of a superconductor.

ferromagnetic materials Technically, materials having very high magnetic permeability that depends on the magnetizing force. "Ferro" refers to iron, typical of the group. Physically, these are materials strongly attracted to magnets and are used in permanent magnets, including nickel and cobalt, and various ceramic compounds, notably those made of neodymium, iron, and boron.

magnetic levitation The noncontact support of an object with magnetic forces. In the context of transportation applications, Howard Coffey coined the term "Maglev."

magnetic permeability The ratio of change in magnetic flux density in a material to the change in magnetizing force applied to it.

reluctance The resistance of a material or combination of materials to magnetization.

This description of magnetic levitation focuses on applications of this technology to both low- and high-speed mass transportation systems. Methods of providing the three main functions of levitation, guidance, and propulsion are described in detail for the two basic types of magnetic suspension systems, namely, the attractive-force and the repulsive-force systems. Various magnetic field source design options, including the use of conventional electromagnets, permanent magnets, and superconducting magnets to accomplish these three functions, are discussed and compared. Alternative choices for propulsion system configurations, including the use of long- versus short-stator linear induction and synchronous motors, are also examined. This is followed by a brief history and update of the development of the technology and a discussion of the advantages and cost of maglev versus conventional high-speed trains.

1. INTRODUCTION

The term magnetic levitation has come to be used in a wide variety of different contexts ranging from suspending a small laboratory-scale stationary object so that it is isolated from vibrations of its surroundings (an isolation platform) to large-scale mobile applications such as maglev vehicles capable of carrying people and materials up to speeds of several hundred miles per hour or the proposed assisting in the launch of space vehicles. Depending on the nature of the application, some degree of physical contact may be required. However, if physical contact is to be completely eliminated, as in the case of very high-speed vehicles, then in addition to suspension, the functions of lateral guidance, propulsion, braking, energy transfer, and system control must be provided by noncontact means alone. Faced with a wide variety of options, maglev system designers must decide how the system should be configured and which components should be placed on board the vehicle and which should be mounted on the guideway. The decision depends on a variety of both technical and economic issues.

For maglev vehicle applications, two basic maglev suspension system designs have been widely used. The first employs the attractive force between magnets and ferromagnetic metals and is referred

 © 2004 Elsevier Inc. All rights reserved.

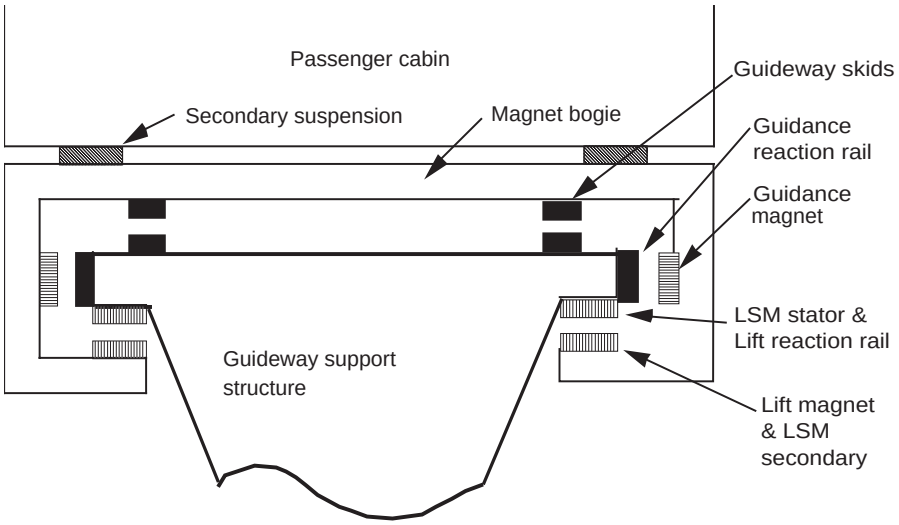


FIGURE 1 Schematic diagram of a long-stator EMS system. The magnet bogie, which supports the lift and guidance magnets, is connected to the passenger cabin via a secondary suspension system.

to as an electromagnetic suspension (EMS) system. The other uses the repulsive force generated by magnets moving relative to electrical conductors and is referred to as an electrodynamic suspension (EDS) system. Examples are illustrated schematically in Figs. 1 and 2.

2. CHARACTERISTICS OF ATTRACTIVE-FORCE-BASED SUSPENSION SYSTEMS

The task of trying to suspend a ferromagnetic object below a permanent or electromagnet magnet with an air gap between the magnet and the object is thwarted by an intrinsic instability. An air gap length exists where the force of magnetic attraction exactly balances the force of gravity. However, if the air gap length decreases slightly, the attractive force increases and the object is clamped onto the magnet. If the air gap increases slightly, the force is weakened and the object falls down. This inherent instability can be overcome by constantly changing the current supplied to an electromagnet's windings in response to signals being fed back to the magnet's power supply from a sensor that continuously monitors the air gap. More complex feedback control systems use additional control parameters, such as the absolute position of the object, its velocity, and its acceleration, to achieve certain suspension characteristics.

All EMS-based systems employ some type of feedback control mechanism to maintain stable levitation. (An interesting exception to the above rule, which has been demonstrated in the laboratory, consists of a permanent magnet suspended below a block of high-temperature superconducting material held below its critical temperature. Stable suspension results from a property of the superconducting state that immobilizes magnetic field lines in the superconducting material.)

A feature of EMS systems that is particularly attractive for maglev vehicles is that the suspension force is essentially independent of speed. Consequently, stable magnetic suspension can be achieved at rest as well as at any other speeds within the system's design limits. Typically, for vehicle applications, in order to compensate for the force of gravity with an attractive magnetic force, the vehicle and guideway components must be configured so that the onboard magnets are drawn upward toward the undersides of the guideway rails, as illustrated in Fig. 1.

The design of the lateral guidance and propulsion systems must, of course, be compatible with the lift system design. A variety of methods for producing lateral guidance in EMS systems have been developed. The German Transrapid high-speed EMS system utilizes separate magnets interacting with guideway-mounted side rails, as illustrated in Fig. 1. In that system, the guidance magnets are excited with feedback-controlled DC power supplies in essentially the same manner as the lift magnets. Several low-speed EMS systems, including the Japanese high-speed surface transport (HSST) and systems under development in Korea and the United States, use a suspension system in which the same magnets are

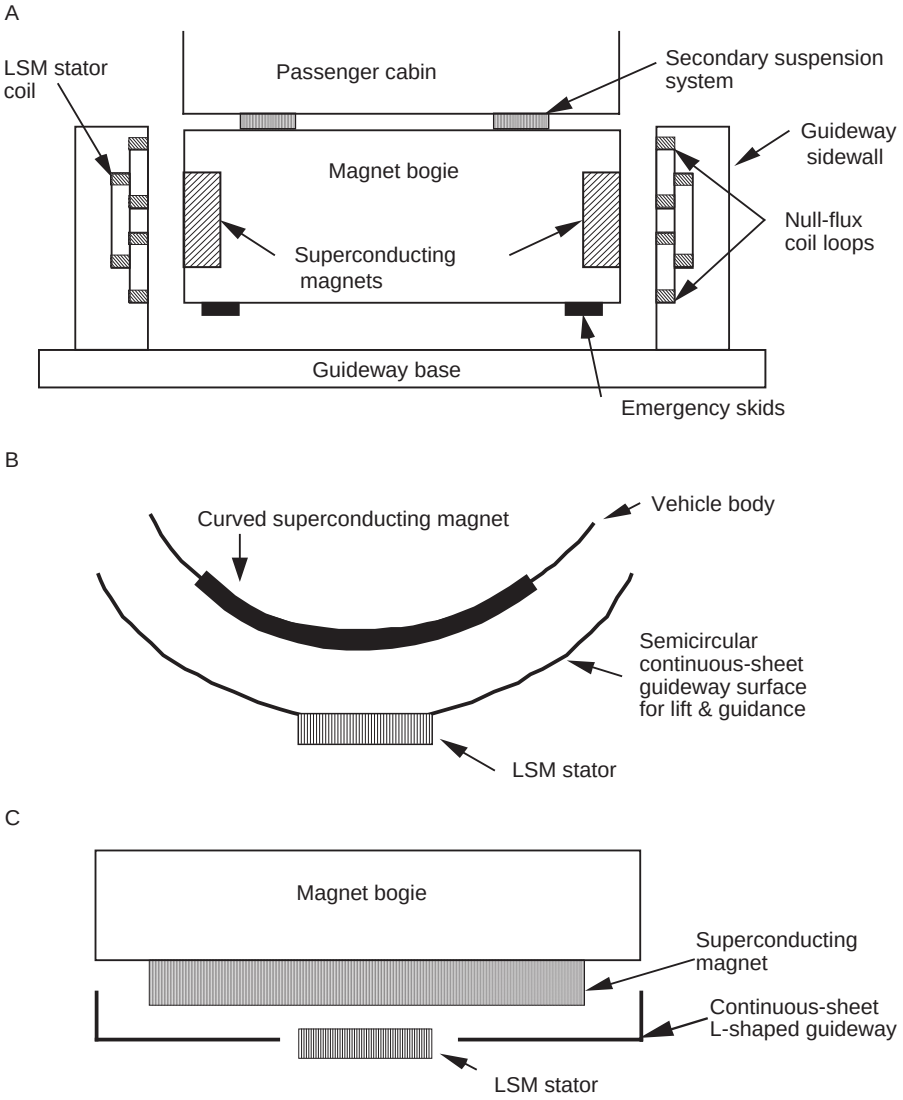


FIGURE 2 Examples of EDS systems using superconducting magnets. (A) Null-flux lift and guidance with air-core LSMs in a "U"-shaped guideway. (B) Continuous-sheet, semicircular guideway. (C) Continuous-sheet, "L"-shaped guideway.

used for both lift and lateral guidance, as illustrated in Fig. 3. The latter method utilizes the natural tendency of a ferrous metal object to orient itself relative to magnetic poles so as to minimize the reluctance of magnetic flux path to magnetization. In Fig. 3, the minimum reluctance position of the ferrous rail is centered between the two magnet poles. To increase lateral stiffness, the vehicle-borne magnets are alternately offset from the centerline. Departures from the lateral equilibrium position result in a strong centering force to be exerted on the rail. Another method proposed for combining lift and guidance functions in EMS systems is to place the reaction rails and magnets at an angle so that the magnetic attraction force has both a horizontal component and a vertical component.

In principle, any type of noncontact propulsion system could be used to propel a maglev vehicle. However, in practice, the choice depends on a number of important and often conflicting considerations, including compatibility with the lift and guidance system, power requirements, thrust requirements, operating speed, weight penalties, cost, and environmental constraints. With the exception of some special-purpose applications, linear induction motors (LIMs) and linear synchronous motors (LSMs) have generally been the propulsion means of choice. Both consist of a primary part that generates a traveling

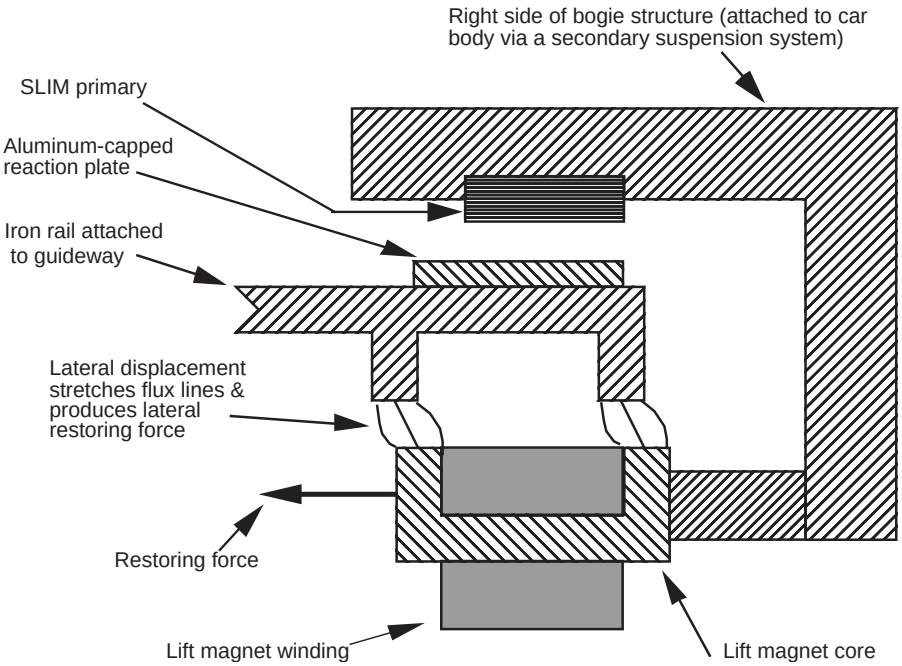


FIGURE 3 An EMS short-stator system used in several low-speed maglev designs.

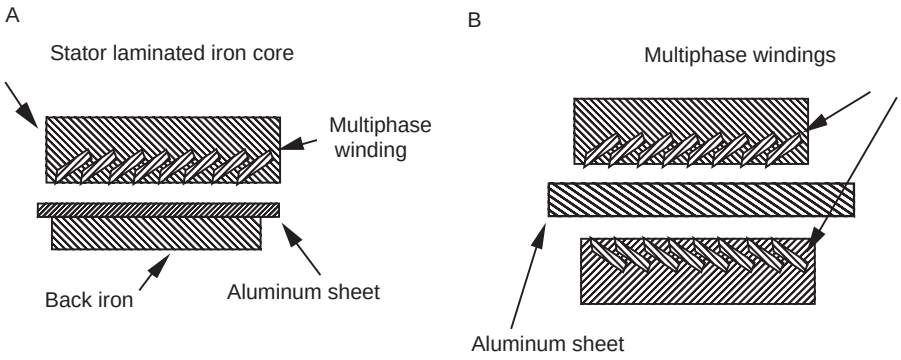


FIGURE 4 (A) Single-sided LIM. (B) Double-sided LIM.

magnetic wave, which in turn interacts with a secondary motor part to produce thrust. Air gaps for both types of linear motors are generally approximately 0.5 in. or less in EMS systems. A third type of linear motor uses electronic switching technology to turn electric currents on and off at precisely the right times to create thrust between the primary and secondary motor parts. The latter type, called sequentially excited linear motors, includes locally commutated linear synchronous motors and pulsed linear induction motors.

LIMs consist of an energized primary part, the stator, and a passive secondary part, the reaction rail (Fig. 4). The stator generally consists of a laminated iron core with transverse slots spaced along its length. Windings are placed in the slots to form a series of coils. In the most common form, three separate windings are placed alternately in the core slots so that when excited with three-phase electric current, a traveling magnetic wave is created. The speed of the traveling wave, called the synchronous speed, is given by the relation, $v_5 = 2pf$, where p is the distance between stator magnetic poles (the pole pitch) and f is the frequency of excitation in Hertz. The primary power supply is called a variable voltage variable frequency (VVVF) power supply.

When the reaction rail is placed adjacent to the stator, the traveling magnetic wave exposes the rail to time-varying magnetic fields, which in turn cause eddy currents to flow in the rail. Those

eddy currents react with the traveling wave and produce a force that has two components: one in the direction of motion, i.e., a thrust force, and another that acts normal to that direction, i.e., a repulsive force. As the speed between the traveling wave and rail (called the slip speed) is increased, the thrust initially increases, reaches a peak value, and then diminishes. The repulsive normal force increases with slip speed.

Depending on the motor configuration and its application, the reaction rail may be composed of ferrous or nonferrous metal or a combination of both. Single-sided LIMs (Fig. 4A) generally use iron or aluminum-capped iron rails for low-speed applications. Examples include the Japanese HSST and the maglev systems under development in the United States and Korea. The iron provides a return pathway for the magnetic flux emanating from the stator pole faces. This concentrates the magnetic flux in the air gap and improves motor performance in some cases. However, with iron reaction rails, the ferromagnetic effect produces a strong attractive force that acts normal to the direction of motion. As the slip speed increases, this attractive normal force is partially compensated for by the repulsive force between the stator magnetic pole faces and eddy currents induced in the rail as mentioned above. Since the iron is a relatively poor conductor of electricity, the repulsive normal force component tends to be small. When an aluminum cap is placed over the iron rail, the induced eddy currents are larger and the net normal force can be adjusted from attractive at low slip speeds (dominated by ferromagnetic attraction) to repulsive at high slip speeds (dominated by eddy-current repulsion). For high-speed applications, so-called end effects, particularly at the entry end of the motor, become very important and can considerably degrade motor performance. Consequently, to achieve satisfactory performance at high speed, they must be compensated for. Several techniques have been proposed, including the use of compensating stator windings and wound secondaries. Pulsed linear induction motors that eliminate the end effects have also been designed.

For LIM applications, a crucial question is whether to place the stator on board the vehicle and the reaction rail on the guideway (called the short stator option) or vice versa. The short-stator option adds substantially to the vehicle's weight, which puts an extra burden on the suspension system. It also requires that the propulsion power be generated on board or transferred to the vehicle from the guideway. On-board power generation requires the storage and conversion of chemical fuel to electricity, which produces pollutants and adds more weight to the vehicle. Use of fuel cells could reduce such impacts. Transfer of propulsion power to the vehicle from the guideway using sliding contacts or brushes has proven practical for low-speed systems only. The principle advantage of the short-stator system is low cost. With most of the technology carried on board, the guideway is relatively simple, and since the cost per mile of guideway must be multiplied by the guideway length, the total cost savings can be very substantial. However, several factors have limited the short-stator option to low-speed applications. First, the difficulty of transferring power to the vehicle increases with speed. Second, greater speed requires larger and heavier motors and accompanying power-conditioning equipment, which, in turn, places an ever-increasing burden on the suspension system, which, in turn, becomes larger and heavier and requires more power. A third issue is the degradation of performance of LIMs at high speed due to motor end effects. As noted earlier, the end effects must be either compensated for or eliminated by special LIM designs.

The long stator option, on the other hand, places the primary on the guideway and the reaction rails on the vehicle, thus reducing vehicle weight and eliminating the need to transfer propulsion power to the vehicle, which results in a lighter, simpler vehicle, but a more complex and expensive guideway. In principle, one could also achieve a better performing system, with greater acceleration and maximum speed. However, continuous generation of eddy currents in the on-board reaction rails is accompanied by ohmic heating, which poses large cooling requirements on board the vehicle.

Whereas low-speed maglev systems have generally used single-sided LIMs, double-sided LIMs using aluminum or copper sheets or shorted-turn coils (coils whose ends are connected together) have been used in applications requiring high acceleration, such as mass drivers and rail guns. Since this type of motor configuration has energized magnetic poles on both sides of the reaction rail, the net normal force is near zero and a good magnetic flux path is provided, leading to a high thrust capability.

LSMs also have a primary part that is energized by a VVVF power supply so as to produce a traveling magnetic wave. However, the secondary part differs in that it consists of a separate set of discrete magnetic poles (Fig. 5). The latter may be a series of permanent or electromagnets or ferrous

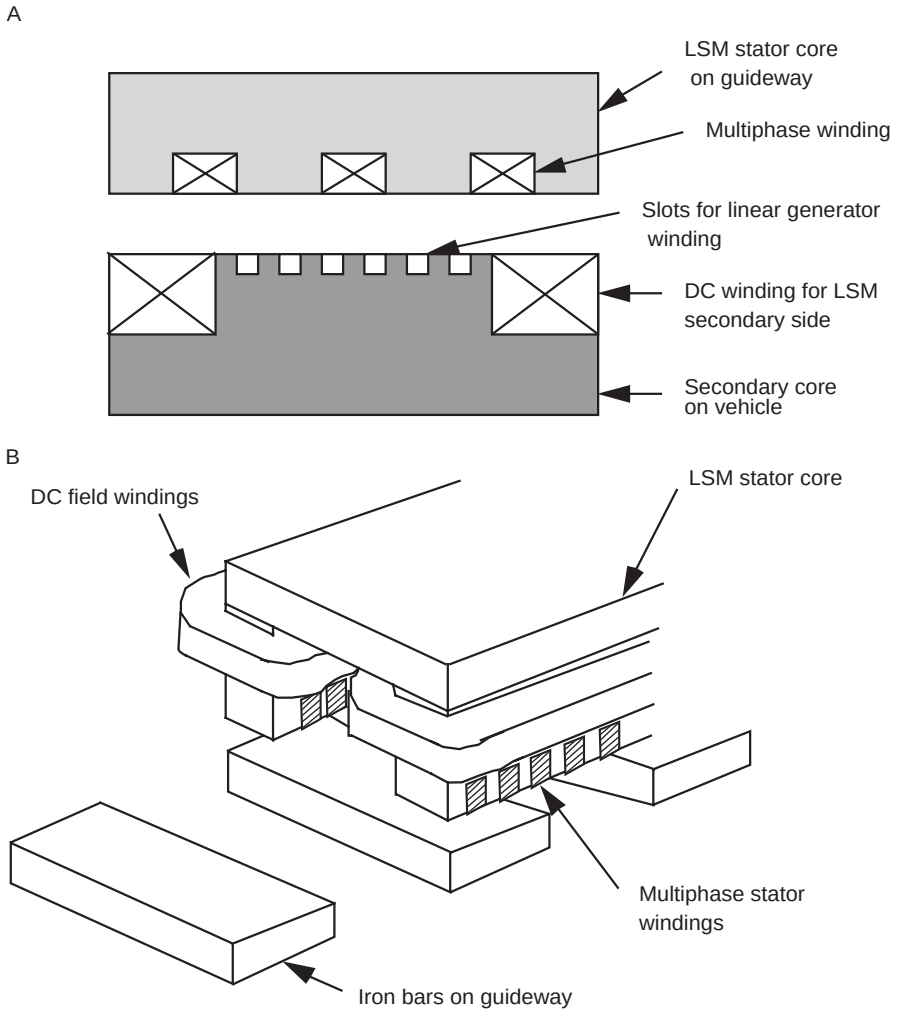


FIGURE 5 (A) Longitudinal cross section of a conventional iron-core LSM. (B) Isometric sketch of a homopolar LSM in which both the AC and DC windings are on the stator core.

metal bars that are magnetized by separate DC energized windings placed on the stator core. Thrust is produced by the interaction between the traveling magnetic wave and the secondary magnetic poles. In contrast to the LIM, the LSM secondary part always moves at the synchronous speed. Position sensors continuously monitor the position of the secondary poles relative to the poles of the traveling wave and their signals are used to maintain synchronization of the secondary part with the traveling wave. The speed is controlled by the frequency of the primary excitation and the thrust by the applied voltage.

The LSM configuration of choice for both EMS and EDS high-speed applications places the primary side on the guideway and the secondary on board the vehicle. This long-stator option has several advantages. It enables the same on-board magnets to be used for lift and as the secondary side of the LSMs. In the case of the German Transrapid system, the laminated iron core of the stator also serves as the reaction rail for the on-board lift magnets. Furthermore, it avoids the problem of transferring propulsion power from the guideway to the vehicle at high speed. It also reduces the weight burden on the vehicle suspension system and places vehicle control at the wayside, eliminating the need for an onboard engineer or driver. To improve efficiency, the primary side of the LSM is normally divided into blocks ranging in length from a few tenths of a mile to several miles. The wayside power control system monitors the position of each vehicle in the system and turns on a block as the vehicle approaches and turns it off after

it leaves. The system is designed so that only one vehicle (which may consist of one or more cars linked together as in a train) can occupy a block at a time. This allows short headways while ensuring positive vehicle separation.

3. CHARACTERISTICS OF REPULSIVE-FORCE-BASED SUSPENSION SYSTEMS

In contrast to EMS systems, EDS systems can provide stable suspension without a feedback system. Imagine a magnet held above a moving flat sheet of aluminum or copper. Following Ampere's law of induction, an electric field is induced in the moving metal surface that causes eddy currents to flow in closed loops near its surface. The eddy currents in turn set up their own magnetic field, whose polarity, in accordance with Lenz's law, opposes that of the magnet's field. Consequently, the magnet is repelled from the moving metal surface, countering the force of gravity. If the magnet is pushed down toward the moving metal surface, the induced currents and the resultant repulsive force increase, restoring the equilibrium position automatically. Conversely, if the magnet is moved upward, the levitating force decreases. Consequently, the system is said to be inherently stable. (Of course, the magnet must be held in position because the induced eddy currents also cause an electromagnetic drag force that tends to pull the magnet in the longitudinal direction along with the moving metal sheet.)

In practice, the stability of a magnetically suspended object is more complicated because any disturbance from the equilibrium position in the repulsive force case results in an oscillation about the equilibrium position. The oscillation may decrease or increase with time depending on whether the net damping force present is positive or negative, respectively. Both active and passive damping mechanisms may be employed to provide net positive damping and ensure stable suspension. In the attractive force case, the feedback system mentioned above provides an effective positive damping force that attenuates any oscillations.

The configuration of choice for most EDS systems places the magnets on board and the electrical conductors on the guideway. One disadvantage of the EDS system is that, because the lift force depends on the speed, mechanical support must be used at rest and at low speeds until the lift-off speed where the magnetic suspension force exceeds the force of gravity is reached. Pneumatic tires mounted on retractable landing gears can provide suitable mechanical support. Support at rest or at low speed can also be achieved with repulsive magnetic forces. This can be accomplished using AC-excited coils interacting with eddy currents induced in reaction rails or coils. However, this method tends to be rather energy-intensive.

A variety of combinations of vehicle-borne magnets and guideway-mounted electrical conductor configurations have been devised for EDS systems. These include the use of continuous-sheet and discrete-coil guideways. Three configurations are illustrated in Fig. 2. In all cases, the guideway-mounted conductors are completely passive. The null-flux coil configurations have high lift efficiency and can produce forces in two opposing directions (up and down or left and right).

Both permanent and superconducting magnets (SCMs) can be used as the magnetic field sources. Designs utilizing permanent magnets have been investigated by both American Maglev Technology, Inc. and by Lawrence Livermore National Laboratory. The latter group has incorporated permanent magnets in Halbach arrays, which provide greater field strength at the pole faces per unit weight and substantially reduced stray fields behind the arrays.

The principal advantage of SCMs is greater force per unit weight. This is due to the strong magnetic fields that can be produced over large volumes of space and because the SCMs do not use iron cores. Air gaps on the order of 10 to 20 cm are possible with SCMs. Such large air gaps help ensure against contacting the guideway surfaces when guideway irregularities or disturbances including wind gusts and seismic activity are encountered. However, in practical designs, component parts tend to take up at least some of that space, reducing the actual clearance between the outer vehicle and guideway surfaces to 6–8 cm. The major disadvantage is that the superconducting windings must be kept below their critical temperatures to maintain their superconducting states. This requires the use of specially designed cryostats and cryogenic refrigeration systems. Superconducting magnets used in transportation applications are of the "low-temperature" type, which are operated at liquid-helium temperatures. The Japanese have successfully developed efficient cryogenic systems for use with SCMs on their high-speed "linear motor cars." The newer "high-temperature" superconductors, which can be cooled with liquid nitrogen, are not yet ready for such applications.

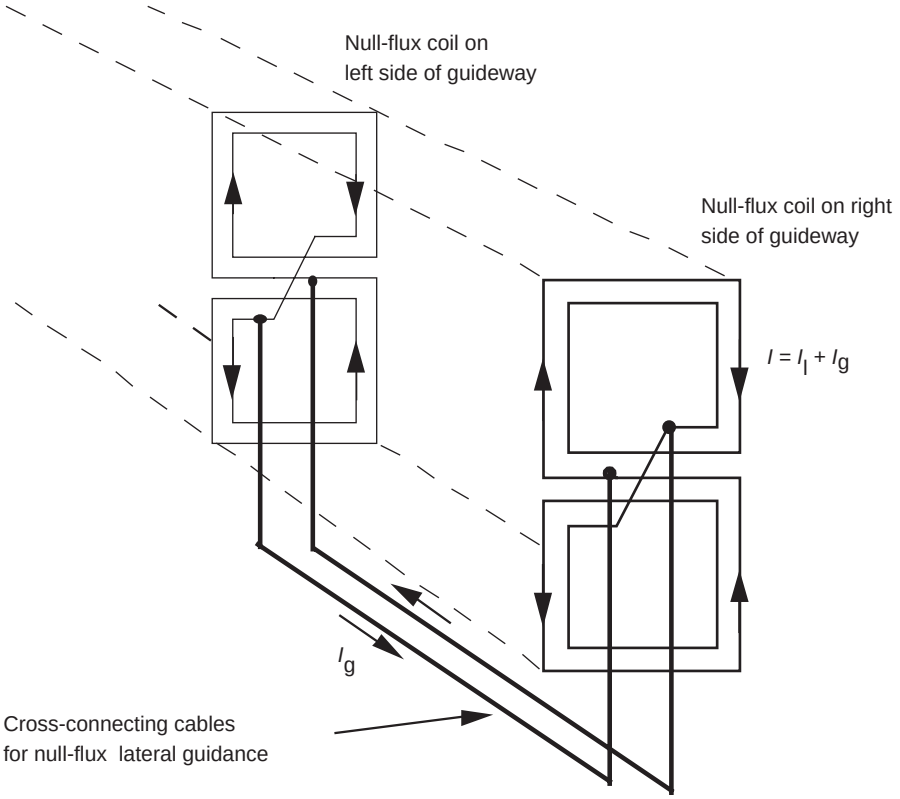


FIGURE 6 Connections of null-flux coil system for null-flux lift and guidance.

EDS systems use eddy current repulsion for guidance as well as lift. Depending on the type of lift and propulsion system being used, and the guideway configuration, either separate magnets or the same magnets can be used for lift and guidance. In the Japanese linear motor car system, the same onboard SCMs provide the magnetic field for lift and lateral guidance and serve as the field sources for the secondary side of the LSMs. Eddy currents induced in the null-flux coils mounted on the guideway sidewalls in this system interact with the magnetic fields of the SCMs to produce both lift and guidance forces. The guidance forces are enhanced by cross-connecting the null-flux coils on opposite sides of the guideway (Fig. 6). Null-flux lift and guidance rely on the same operating principle. That is, the two loops of a null-flux coil are wound in opposite directions so that when a magnet is in the neutral position, an equal amount of magnetic flux links both loops so that no net current flows in the two connected loops and hence, no force is generated. When the magnet is displaced relative to the neutral position, then one loop links more flux than the other, a net current flows in both loops, and a restoring force is produced.

4. MAGLEV DEVELOPMENT HISTORY

One of the earliest known mentions of the concept of high-speed ground transportation using maglev technology appeared in a fictional story written by Robert Goddard, father of the rocket, in 1907. By 1911, two patents were awarded to Graeminger for attractive force suspension systems. He proposed several methods, involving mechanical contact between the electromagnets and the reaction rail, for overcoming the inherent instability in attractive force systems. In the October 18, 1912 edition of *The Engineer*, the article "Foucault and Eddy Currents Put to Service" began with the line: "There was demonstrated some months ago at Mount Vernon, which is really a suburb of New York City, a levitating transmitting apparatus invented by Mr. Emile Bachelet." Bachelet actually conducted a series of experiments in which he levitated plates of various metals above a coil excited with 60 Hz single-phase current. The eddy currents induced in the plates by the time-varying magnetic field reacted against that field to produce the resultant lift force. He designed a test vehicle consisting of aluminum reaction plates

for lift and iron framing pieces that served as the reaction rails for the sequentially excited coils of a simple type of LIM. The energy requirements, though large, were minimized by the clever use of limit switches that turned the levitating and propelling coils on and off as the vehicle passed. Little further development was reported until the 1930s, when Hermann Kemper, a German inventor, demonstrated the first noncontact method for stably suspending an object with attractive magnetic force. His inventions, for which he was awarded patents in 1934 and 1935, were the basis for modern attractive-force-based systems.

Following a long hiatus, widespread interest in maglev technology suitable for intercity and urban applications was renewed in the late 1960s and early 1970s. After a decade of experimentation and analysis in Germany, it was decided in 1978 to use EMS technology with separate lift and guidance electromagnets and long stator LSMs with iron cores for propulsion (see Fig. 1). By 1985, the first 20 km section of the Emsland test track, which included a closed loop at the north end, was completed and the first high-speed prototype commercial vehicle (the TR06) reached a speed of 355 km/h. A second 10 km long closed loop was added to the south end, permitting continuous running in 1987. Continued technical refinements resulted in the TR07 200-passenger vehicle prototype, which reached speeds in excess of 450 km/h in the early 1990s. Visitors could purchase rides, for 20 Deutsch Marks, on the TR07, which reaches speeds near 450 km/h. Construction of the first commercial route is completed in China and is expected to begin revenue service in early 2004.

At the same time that the Germans were developing the high-speed EMS maglev system, several countries including Germany, Japan, and the United Kingdom undertook the development of EMS technology for low-speed applications. Japan Airlines initiated a long-term program in the early 1970s to develop and commercialize an EMS maglev system using the same "U"-shaped magnets and reaction rails for both lift and guidance and short-stator LIMs for propulsion (see Fig. 3). Commercialization was taken over by the Chubu HSST Development Corporation in 1991. Various versions of the HSST family of vehicles have been demonstrated in public venues since the early 1980s, but thus far no systems have been put in revenue service. Developments continued in the 1990s on a 1.6 km test track in Nagoya. The M-Bahn system, which used permanent magnets stabilized by a mechanical contact system and propelled by a LSM, was developed in Germany in the early 1980s and placed in demonstration service in Berlin in the late 1980s but was never operated in revenue service. The first commercial EMS maglev system in the world was installed at the Birmingham Airport in the United Kingdom in 1984. Based on experiments at the British Rail Research and Development Division, lift and guidance were provided by the same controlled DC electromagnets reacting against laminated iron rails in the longitudinal flux configuration and a LIM provided the propulsion. The system shuttled passengers from the air terminal to an adjacent railway station until it was retired in 1995. Short-stator EMS systems are also under development in other countries including Korea and the United States. A consortium of organizations, including American Maglev Technology Corporation, is building a people-mover system at the Old Dominion University in Virginia. Additional efforts at urban maglev technology development, under Federal Transit Administration sponsorship, are in progress at Sandia National Laboratory and at General Atomics Corporation.

Development of the EDS system technology commenced with U.S. scientists Powell and Danby, who received a patent in 1969 for the design of the first SCM-based suspension and stabilization system. It utilized the same set of vehicle-borne SCMs for null-flux lift and guidance. The use of SCMs greatly reduced the energy requirements that stymied further development of repulsive-force, eddy-current-based suspension concepts first demonstrated by Bachelet in 1912. They also allowed large air gaps on the order of 10 to 20 cm. In the early 1970s, researchers around the world, including those in the United States, Canada, the United Kingdom, Germany, and Japan, subsequently picked up on some of Powell and Danby's ideas and began developing their own versions of EDS systems using SCMs. North American participants included researchers at Massachusetts Institute of Technology, Stanford Research Institute, Ford Motor Co., and the Canadian Institute of Guided Ground Transport at Queen's University. Government support for maglev research in the United States stopped in the mid-1970s but work continued in other countries, including Canada, the United Kingdom, and, most notably, in Japan, where development has proceeded to the stage of commercial readiness on the high-speed Japanese linear motor car.

Following several years of basic development and testing of their superconducting maglev concept at the Japanese Railway Technical Research Institute,

construction of the 7.1 km long test track at Miyazaki on Kyushu Island was initiated in 1977. The guideway was of the inverted "T" type with levitation coils (called ground coils) placed horizontally on the floor and LSM stator windings mounted vertically on either side of the stem of the inverted "T." In 1979, the unmanned MLU-500 test vehicle incorporating SCMs reached a speed of 517 km/h. Shortly thereafter, the superstructure of the guideway was converted to a "U" shape and a long program of developmental and testing of the MLU family of vehicles commenced. In the late 1980s, the ground coils were replaced with null-flux lift coils mounted on the vertical sidewalls of the guideway. Up to that period, separate coils were used on the guideway for lift and combined propulsion and null-flux guidance. In the late 1980s and early 1990s, sections of the guideway coils were replaced with a new set of coils that served all three functions of lift, guidance, and propulsion. Meanwhile, the construction of a new 42.8 km test track began in the Yamanashi Prefecture approximately 70 km west of Tokyo. Originally planned so that it could be extended to form a commercial route from Tokyo to Osaka, the first phase consisted of an 18.4 km "priority section" that was completed in 1995. Due to the mountainous terrain, approximately 87% of its length is in tunnels. A 12.8 km portion of the priority section was double-tracked to permit tests of passing vehicles. The first test vehicle delivered to the test track in July 1995 consisted of a full-scale prototype commercial three-car train set referred to as the MLX01. Following preliminary trials, the test speed was gradually increased to 550 km/h in March 1998. After a second prototype vehicle arrived, tests were conducted with trains passing each other at relative speeds as high as 1000 km/h. A new system under development in Switzerland is the SwissMetro. It is a high-speed EDS system designed for operation in partially evacuated tunnels. Such a system is motivated by the mountainous terrain and the desire to reduce the propulsion power required to overcome the large aerodynamic drag force present at high speed in tunnels.

In the United States, the National Maglev Deployment Program, under the auspices of the Transportation Equity Act for the 21st Century, is in its second phase. Two consortia are competing for final approval of their proposed high-speed maglev projects. One connects Baltimore to Washington, DC and the other connects the Pittsburgh Airport to the city and its eastern suburbs. Another project, under the sponsorship of the state of Florida, and the direction of Powell and Danby, will link the Titusville Airport to the Kennedy Space Center and the cruise terminals at Port Canaveral. In addition, the development of low-speed maglev systems for urban mass transit applications is being sponsored by the Federal Transit Administration.

5. ADVANTAGES OF MAGLEV TECHNOLOGY FOR HIGH-SPEED TRANSPORTATION

High-speed maglev technology offers four main advantages: noncontact operation; low-mass vehicles (on a per seat basis, maglev vehicles weigh approximately one-third to three-quarters as much as high-speed trains); high speed; and wayside system control and self-propelled vehicles (each car contains its own secondary part of a LSM).

Noncontact operation means that vehicle traction does not depend on adhesion between contact surfaces, e.g., wheels and rails. Combined with lower mass, vehicles can accelerate and brake faster and can climb steeper grades independent of whether the surfaces are wet or dry. Noncontact also means lower noise levels than for conventional trains in urbanized areas, permitting higher urban and suburban speeds or, alternatively, reduced or eliminated noise abatement costs. In addition, maintenance costs associated with routine wheel, rail, and bearing wear are eliminated. (However, if the specific maglev system design requires a landing gear for low-speed operation, some maintenance costs will be incurred for landing gear wear.) The combination of noncontact operation and reduced vehicle weight translates into reduced stress on guideway support structures and construction costs. At grade, guideway costs have been estimated to be more for maglev systems than for conventional trains, but approximately the same as or even less than that for elevated guideways, depending on the specific technology. A technology assessment conducted by the U.S. Army Corps of Engineers in the mid-1990s estimated that the technology cost (omitting profits and land acquisition costs) for elevated maglev system concepts would range from $18 to $28 million/double-track mile and at grade systems from approximately $11 to $25 million/double-track mile (all in terms of 1993 U.S. dollars). Estimates place the total cost for the elevated German Transrapid system at approximately $35 million/double-track mile. Total costs for existing high-speed rail systems range widely

depending on country, terrain, and percentage of elevated track and tunnels used. In terms of 1993 U.S. dollars, costs range from $7.3 to $21.5 million/double-track mile in France and from $32 to $50 million/double-track mile in Germany.

High speed translates into reduced trip times. Existing maglev systems have demonstrated operating speeds of passenger-carrying vehicles in the range of 450 to 550 km/h. Higher speeds are possible in open air but become increasingly energy-intensive due to the rapidly increasing aerodynamic drag force with speed. However, higher speeds are practical in partially evacuated tunnels. Figure 7 shows a comparison of line haul times versus trip length for commercial jet aircraft and high-speed ground transportation (HSGT) modes. Line haul time refers to time spent in the vehicle from terminal to terminal and does not include terminal access or egress times or times spent in the terminals. Time spent on the ground by aircraft in major airports is clearly a disadvantage that is only compensated for by higher speeds at flight altitude at flight distances greater than approximately 400 to 500 miles. Figure 8 shows a more realistic comparison of the total trip times, including access and egress time and times spent in the origin and destination terminals for maglev and jet aircraft. The maglev makes two additional urban stops to pick up and discharge passengers (reducing access and egress times); the aircraft makes no additional stops. The maglev is the clear winner even for distances greater than 900 miles.

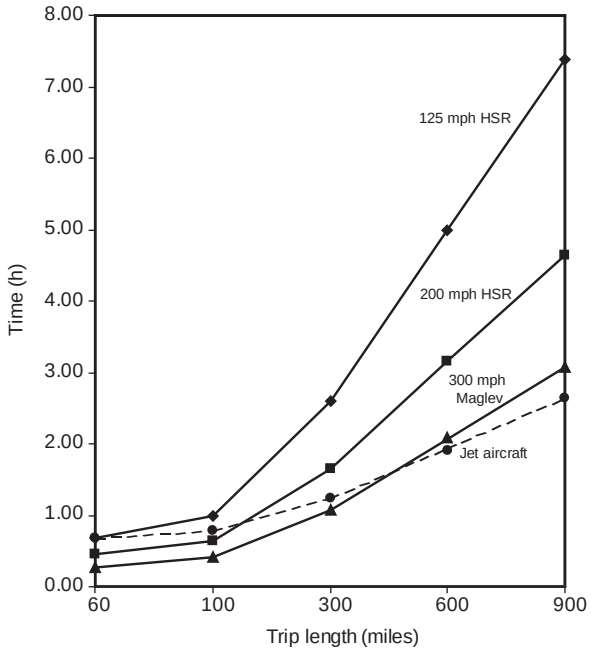


FIGURE 7 Line-haul trip times versus trip length for various high-speed modes.

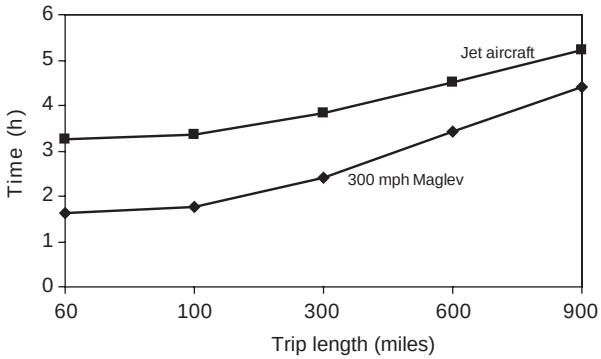


FIGURE 8 Total trip times versus trip length. Maglev makes two urban stops. Jet aircraft makes no stops.

Since maglev cars are self-propelled, they can be operated as individual cars (each carrying 50 to 100 passengers) or as linked cars, depending on travel demand. Furthermore, since each car or set of linked cars is controlled by the wayside control system, headways (times between vehicles) can be reduced to less than 30 s, if necessary, to accommodate travel demand. In contrast, most high-speed trains are operated as standard train sets consisting of one or two locomotives and 8 or 10 coaches. Consequently, the technology provides considerable flexibility to meet a wide range of passenger demand and still maintain a very high-frequency service.

Two public benefits of maglev vehicle systems are reduced dependence on petroleum-based fuels and reduced emissions of air pollutants. The energy used by a maglev system comes from the electric utility grid. Consequently, the emissions produced and energy consumed are strongly dependent on the mix of electricity-generating technologies that supply the grid. To the extent that renewable energy sources, such as hydroelectric, wind, solar, and nuclear technologies, are used, the emissions can be greatly reduced. Furthermore, new fossil-fuel-generating technologies, including combined-cycle gas turbines and fuel cells, that are substantially more efficient than traditional technologies are coming into the mix. In addition, nationally, less than 4% of the electricity is generated by combustion of petroleum-based fuels. Hence, to the extent that trips are diverted from conventional modes to maglev vehicles, the United States moves closer to energy independence and a cleaner environment. The reduction in emissions and energy use comes about for a variety of reasons including the high efficiency

of the maglev technology itself, the efficiency of the generating plants, and the manner in which the maglev vehicles are operated. The efficiencies are somewhat design-specific, e.g., systems incorporating SCM technology are more efficient than those that use conventional electromagnets. Generally, SCMs are operated in the "persistent current" mode, whereby once energized, the winding is disconnected from the power source and the current in the winding continues to flow without loss. Power is then required only to keep the SCM below its critical temperature. Propulsion energy is used only when the vehicles are moving (there is no need for idling). Maglev vehicle systems have the added advantage that they utilize three-phase power, which provides a more balanced load to the utility grid than catenary-based systems, which can use only single-phase power. In addition, the use of smaller coaches with shorter headways tends to distribute the load to the utility system more evenly over time. Finally, since maglev guideways are connected together electrically, there is a greater opportunity to convert the power derived from regenerative braking of some vehicles into propulsion power of other vehicles.

Figure 9 shows a comparison of energy used per seat · mile for jet aircraft and HSGT modes as a function of trip length. Aircraft consume the greatest amount of energy except for distances greater than approximately 700 miles. Figure 10 shows a similar comparison for which the HSGT modes make two additional urban stops and one in-route stop, whereas the aircraft makes only one in-route stop. In this case, the aircraft uses more energy for distances further than 900 miles.

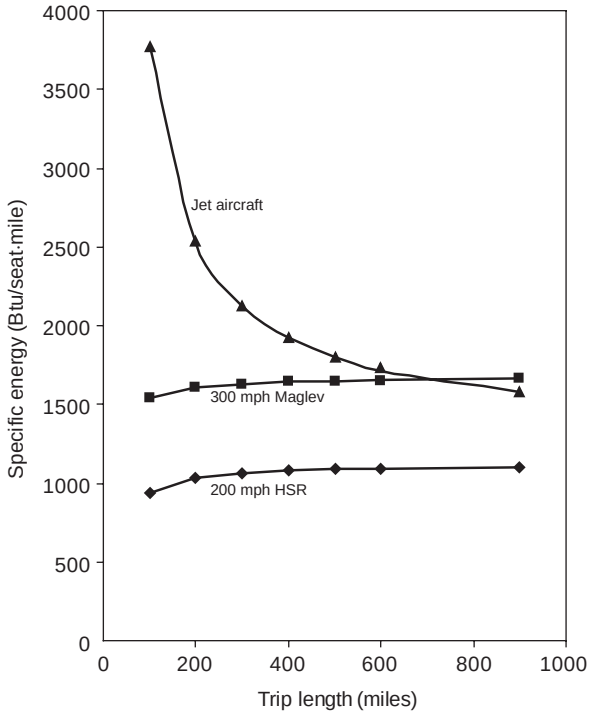


FIGURE 9 Energy per seat · mile used by HSGT and short-haul jet aircraft (no stops).

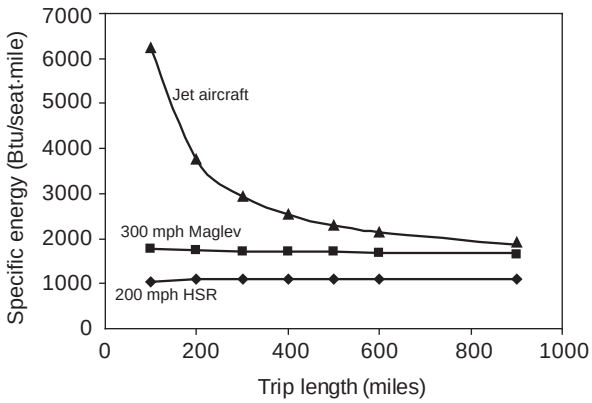


FIGURE 10 Same as for Fig. 9, but includes energy for two urban in-route stops and one in-route stop for HSGT and one in-route stop for jet aircraft.

SEE ALSO THE FOLLOWING ARTICLES

Intelligent Transportation Systems • Magnetohydrodynamics • Transportation and Energy, Overview • Transportation and Energy Policy • Ultralight Rail and Energy Use

Further Reading

German Transrapid World Wide Web Site. Available at http://www.transrapid.de/en/index.html.

He, J. L., Rote, D. M., and Coffey, H. T. (1992). "Survey of Foreign Maglev Systems," Argonne National Laboratory Report No. ANL/ESD-17. Argonne National Laboratory, Argonne, IL.

Japanese Maglev World Wide Web Site. Available at http://www.rtri.or.jp/index.html.

Laithwaite, E. R. (1987). "A History of Linear Electric Motors." Macmillan Education Ltd., London, UK.

Lever, J. H. (ed.). (1998). "Technical Assessment of Maglev System Concepts," Final Report by the Government Maglev System Assessment Team, U.S. Army Corps of Engineers, Cold Regions Research and Engineering Laboratory, Special Report 98-12. National Technical Information Service, Washington, DC.

Moon, F. C. (1994). "Superconducting Levitation." Wiley-Interscience, New York.

NMI Office. (1993). "Final Report on The National Maglev Initiative." FRA, U.S. Army Corps of Engineers, U.S. Department of Energy, Washington, DC.

Rhodes, R. G., and Mulhall, B. E. (1981). "Magnetic Levitation for Rail Transport." Clarendon, Oxford, UK.

Rote, D. M. (2001). "Guidelines for Estimating Trip Times, Energy Use, and Emissions for HSGT Technologies," Argonne National Laboratory Report No. ANL/ES/RP106522. Argonne National Laboratory, Argonne, IL.

Magnetohydrodynamics

PAUL H. ROBERTS
University of California, Los Angeles
Los Angeles, California, United States

PATRICK H. DIAMOND
University of California, San Diego
San Diego, California, United States

1. **Introduction**
2. **Pre-Maxwell Approximation**
3. **Ideal Magnetohydrodynamics and Magnetostatics**
4. **Nonideal Magnetohydrodynamics**
5. **Stability of Magnetically Confined Plasmas**

Glossary

Alfvén wave A wave that propagates along magnetic lines of force because of the tension of the field lines; also called a hydromagnetic wave.

Ampère's law The law that, in pre-Maxwell theory, determines the magnetic field created when electric current flows.

beta The name given to the ratio of the gas pressure to the magnetic pressure in a conducting fluid; plasma β is one of the figures of merit awarded to a magnetostatic equilibrium.

dynamo The process whereby a magnetofluid self-excites and amplifies a magnetic field.

electromotive force The electric field created by a changing magnetic field. Motive refers to its ability to move charge in a conductor (i.e., generate current).

Faraday's law The law that determines the electromotive force created by a changing magnetic field.

field line A curve that is everywhere parallel to the direction of the prevailing magnetic field; also called a line of (magnetic) force, it is part of a family of such lines, one of which passes through each point of space.

field line tension Part of the stress that a magnetic field **B** exerts on a medium and interpreted as a tension in a flux tube of B^2/μ_0, per unit cross-sectional area of the tube; μ_0 is the permeability of the medium.

figure of merit A number that quantifies the susceptibility of a plasma equilibrium to a certain type of magnetohydrodynamics instability.

flux tube See magnetic flux surface.

force-free field The field in a particular type of magnetostatic equilibrium in which the Lorentz force vanishes and in which field and current are parallel.

gradient A term used to indicate how rapidly a scalar or vector field changes with position. In the case of a scalar field, Φ, it is used to describe the vector field that is parallel to the direction in which Φ increases most rapidly; it is equal to the spatial rate of change of the scalar in that direction and is denoted by $\nabla\Phi$.

interchange-ballooning mode An instability driven by the combined effects of pressure gradient and average, or locally, unfavorable curvature of the field lines.

International Thermonuclear Experimental Reactor (ITER) a large-scale experiment, run by an international consortium, designed to study the physics of burning plasma.

kink safety factor A figure of merit quantifying the susceptibility of a plasma equilibrium to kink instability.

kink-tearing mode An instability driven by current gradients.

line of force A curve that is everywhere parallel to the direction of the prevailing magnetic field; also called a (magnetic) field line, it is part of a family of such lines, one of which passes through each point of space.

Lorentz force The force per unit volume on a fluid conductor that carries an electric current (density **J**) and lying in a magnetic field **B**; mathematically, it is the vector product of **J** and **B** and can also be represented by magnetic stresses.

magnetic helicity A property of a magnetic field that is preserved in the motion of a fluid that is a perfect electrical conductor.

magnetic pressure Part of the stress that a magnetic field **B** exerts on a medium and interpreted as an isotropic pressure $B^2/2\mu_0$, where μ_0 is the permeability of the medium.

magnetic reconnection The process by which a magnetofluid forms thin current layers that change the field line topology and dissipate magnetic energy.

magnetic relaxation The process by which a magnetofluid relaxes to its minimum energy state subject to certain constraints.

 © 2004 Elsevier Inc. All rights reserved.

magnetic shear The rate at which the direction of a magnetic field changes with position **x**; a special case is used in plasma physics as a figure of merit.

magnetic stresses Stresses that are equivalent in their dynamical effect to the Lorentz force.

magnetic surface, also called a flux tube or a magnetic flux surface A surface composed of field lines so that the magnetic field, **B**, is everywhere tangential to it. The flux of **B** contained within the surface is the same everywhere along its length.

Ohm's law The law that states that the electric current density, **J**, produced in a motionless conducting medium by an electric field **E** is proportional to **E**. The constant of proportionality is the electrical conductivity of the medium.

pre-Maxwell theory The approximate form of electromagnetic theory that existed prior to Maxwell's discovery of displacement currents.

scalar product A convenient mathematical abbreviation used to describe a quantity derived from two vectors, **F** and **G**, and in magnitude is equal to the product of their magnitudes, |**F**| and |**G**|, and the cosine of the angle between them; it is written as $\mathbf{F} \cdot \mathbf{G}$.

Taylor relaxation The process proposed by J. B. Taylor whereby a magnetofluid relaxes to its minimum energy state subject to the constraint of constant global magnetic helicity.

toroidal Alfvén eigenmode (**TAE**) A particular type of Alfvén wave appropriate to toroidal geometry.

vector product A convenient mathematical abbreviation used to describe a vector that is perpendicular to two other vectors, **F** and **G**, and that in magnitude is equal to the product of their magnitudes, |**F**| and |**G**|, and the sine of the angle between them; it is written $\mathbf{F} \times \mathbf{G}$, where **F**, **G**, and $\mathbf{F} \times \mathbf{G}$ form a right-handed triad.

Magnetohydrodynamics (*MHD*) is the study of the movement of electrically conducting fluids in the presence of magnetic fields. The magnetic field influences the fluid motion through the Lorentz force, which is proportional and perpendicular to the magnetic field and the electric current flowing through the conductor: The magnetic field is affected by the electric current created by an electromotive force which is proportional and perpendicular to the magnetic field and the fluid velocity. It is this duality between magnetic field and fluid flow that defines the subject of MHD and explains much of its fascination (and complexity).

1. INTRODUCTION

Magnetohydrodynamics (MHD) is the marriage of hydrodynamics to electromagnetism. Its most famous offspring is the Alfvén wave, a phenomenon absent from the two subjects separately. In fact, many consider the discovery of this wave by Alfvén in 1942 to mark the birth of MHD. Initially, MHD was often known as hydromagnetics, but this term has largely fallen into disuse. Like MHD, it conveys the unfortunate impression that the working fluid is water. In reality, the electrical conductivity of water is so small that MHD effects are essentially absent. Moreover, many fluids used in MHD experiments are antipathetical to water. Even as fluid mechanics is now more widely employed than hydrodynamics the terms magnetofluid mechanics or magnetofluid dynamics, which are already sometimes employed, may ultimately displace MHD. Magnetofluid is already widely used in MHD contexts.

Since electric and magnetic fields are on an equal footing in electromagnetism (EM), it may seem strange that the acronym EMHD was not preferred over MHD. There are two reasons for this. First, to invoke EM theory in its full, unapproximated form would, in most contexts, add complexity without compensating enlightenment. It usually suffices to apply the form of EM theory that existed in the 19th century before Maxwell, by introducing displacement currents, cast the theory into its present-day form. In this pre-Maxwell theory of EM, there are no displacement currents and the electric and magnetic fields are not on an equal footing; the magnetic field is the master and the electric field the slave. Consequently, MHD is an appropriate acronym, but EMHD is not. Situations in which this is untrue and in which full EM theory is needed involve relativistically moving fluids and are too seldom encountered to be described here; in this article, the pre-Maxwell approximation is used throughout. The second reason why EMHD cannot be used is that the acronym is generally understood to mean electron MHD, which studies the MHD of the electron fluid in a two-fluid description of a plasma, the ions forming the other fluid. This topic is also outside the scope of this article.

A significant branch of MHD is the study of Magnetostatic equilibria (MSE). This subject is the MHD analogue of hydrostatics, the branch of fluid mechanics that deals with fluids at rest, with the pressure gradient in the fluid balancing external forces such as gravity. Similarly, in MSE the fluid is motionless and the pressure gradient balances the Lorentz force (and any other forces present).

The two main applications of MHD are technological—to liquid metals and to plasmas. There is little doubt that the former has had the greater

impact on society. It includes the casting and stirring of liquid metals, levitation melting, vacuum-arc remelting, induction furnaces, electromagnetic valves, and aluminum reduction cells. Another application, the flow of a liquid metal in the blanket surrounding a thermonuclear reaction chamber, touches on the other main area: plasma MHD. The reactor contains a rarefied plasma of deuterium/tritium (DT) that is raised to a high enough temperature for these nuclei to fuse and release energy. The economic promise of such a device in generating magnetic fusion energy (MFE) has provided a powerful incentive for studying plasma MHD and has led to significant new insights, particularly into the structure and stability of MSE. In addition to these practical applications, the elucidation of a wide variety of magnetic phenomena in nature depends on an understanding of MHD. Astrophysics and geophysics provide abundant examples, including the magnetism of the earth, planets, and satellites, that of the sun and other stars, and that of galaxies.

This article describes the simplest form of MHD theory, in which the working fluid is a homogeneous, continuous medium. This may or may not (depending on context) be a satisfactory description of a plasma. In general, a plasma and its dynamics are described by kinetic equations that govern distribution functions for electrons, ions, and (if the plasma is incompletely ionized) neutral atoms. The distribution function, $f(\mathbf{x}, \mathbf{v}, t)$, of any of these species is proportional to the probablility that a particle of that species at position $\mathbf{x}$ is moving with velocity $\mathbf{v}$ at time t. In certain cases, such a description can be greatly simplified by replacing the kinetic equations for the functions f by a closed set of fluid equations. This happens when the dependence of each f on $\mathbf{v}$ is close to what it would be if the plasma were in local thermodynamic equilbrium, as seen by an observer moving with the mean velocity, $\mathbf{V}(\mathbf{x}, t)$, of the species at $\mathbf{x}$ and t. Such a distribution is termed Maxwellian. The distribution functions are then replaced by ion, electron, and neutral densities, n_i, n_e, and n_n, respectively, that give the total numbers of each species per unit volume at $\mathbf{x}$ and t, with each fluid having its own (mean) velocity. Although this may seem (and is) very complicated, it is far less forbidding than the alternative (i.e., solving for the evolution of the distribution functions).

The method of closure described previously necessarily involves an approximation of restricted validity. Single-fluid MHD also assumes that the characteristic length scale, $\mathscr{L}$, is large compared with the Debye length, $\mathscr{L}_{\text{debye}}$. This length quantifies the distance over which n_e and n_i can differ substantially from one another. When $\mathscr{L} \gg \mathscr{L}_{\text{debye}}$, these densities are closely equal and the plasma is very nearly electrically neutral. A second requirement for simple one-fluid MHD to hold is that $\mathscr{L}$ be large compared with the mean free path between particle collisions in the plasma and in comparison with the gyroradii. (A charged particle in a magnetic field would, in the absence of collisions with other particles, orbit about a field line in a spiral having a certain gyroradius that depends on its mass and charge and on the field strength.) These demands are more easily satisfied in many astrophysical contexts than in a thermonuclear reactor chamber. Nevertheless, one-fluid MHD can provide useful qualitative insights into the structure and stability of plasmas relevant to MFE.

To simplify the following discussion, the abbreviations shown in Table I will usually be employed. In particular, velocity means fluid velocity, density stands for the mass density of the fluid, field means magnetic field, current is short for electric current density, conductor means conductor of electricity, and potential means electric potential. Script letters—$\mathscr{L}$, $\mathscr{T}$, $\mathscr{V}$, $\mathscr{B}$, $\mathscr{J}$, $\mathscr{E}$, etc.—are used to indicate typical magnitudes of length, time, velocity, field, current, electric field, etc.

The aim of this article is to provide as simple an account of MHD as possible—one that is slanted toward the MFE field but one that involves a minimum of mathematics. Some familiarity with vector fields is necessary, however.

TABLE I

Abbreviations

Quantity	Symbol	Magnitude	SI unit
Time	t	$\mathscr{T}$	Seconds (s)
Position	$\mathbf{x}$	$\mathscr{L}$	Meters (m)
Velocity	$\mathbf{V}$	$\mathscr{V}$	m/s
Density	ρ	—	Kilogram/m^3 (kg/m^3)
Pressure	p	—	Newton/m^2 (N/m^2)
Temperature	T	—	Kelvin (K)
Field	$\mathbf{B}$	$\mathscr{B}$	Tesla (T)
Current	$\mathbf{J}$	$\mathscr{J}$	Amp/m^2 (A/m^2)
Charge density	ν	—	Coulomb/m^3 (C/m^3)
Potential	Φ	—	Volt (V)
Electric field	$\mathbf{E}$	$\mathscr{E}$	V/m
Conductivity	σ	—	Siemens/m (S/m)

2. PRE-MAXWELL APPROXIMATION

Some readers may value a brief review of the relevant concepts in EM theory that underpin MHD; others may prefer to skim this section and move ahead to Section 3.

Under discussion are systems having characteristic length, time, and velocity scales, $\mathscr{L}$, $\mathscr{T}$, and $\mathscr{V}$, respectively, for which $\mathscr{L} \ll c\mathscr{T}$ and $\mathscr{V} \ll c$, where c is the speed of light. In these circumstances, EM phenomena are well described by an approximation that was in use before Maxwell introduced displacement currents and his famous equations into EM theory. This approximation, which is also called the nonrelativistic approximation (since it implies that c is infinite), is summarized.

Electrostatics is the study of static electricity. Each positive (negative) electric charge experiences an attraction toward every negative (positive) charge and a repulsion from every other positive (negative) charge. The sum, **f**, of all these forces on a charge e is proportional to e and is conveniently represented by an electric field, $\mathbf{E} = \mathbf{f}/e$. Suppose that, in steady conditions, a test charge is carried around a closed curve Γ, where by test charge is meant a charge so small that it does not have any effect on the other charges. As the charge describes Γ, the agency that moves it will sometimes do work against the force $\mathbf{f} = e\mathbf{E}$ and will sometimes receive energy from it but, when the circuit is complete the net gain or loss of energy is zero. Such forces are termed conservative and this fact is expressed mathematically by the statement that $\mathbf{E} = -\nabla\Phi$, where Φ is the electric potential. (Here, $\nabla\Phi$ denotes the gradient of Φ, a vector that has Cartesian components $\partial\Phi/\partial x$, $\partial\Phi/\partial y$, $\partial\Phi/\partial z$.) If positive, $\Phi_1 - \Phi_2$ is the energy received from the electric field when a unit positive charge moves from a location where the potential is Φ_1 to one where the potential is Φ_2; if negative, it is the work that must be done to move the charge from the first location to the second. The force $\mathbf{f} = -e\nabla\Phi$ acts to move a positive charge in the direction of **E** from a region of higher potential to one of lower potential.

If a conductor carries a net charge, the mutual repulsion between charges quickly drives them to the surface of the conductor, where they arrange themselves with a surface charge density such that the surface is at a uniform potential Φ. The electric field is then normal to the surface and cannot move charges along the surface. The charge distribution is therefore static. Beneath the surface charge layer, **E** is identically zero, as is the volumetric charge density ϑ. If different areas of the surface are held at different potentials, as at the two ends of a straight wire, **E** is no longer zero in the conductor and a charge e experiences the force $\mathbf{f} = e\mathbf{E}$. If $e > 0$ (< 0) this moves it from (to) the area at the higher potential to (from) the area at the lower potential. The resulting flow of charge is called an electric current. Its density, **J**, is proportional to the force **f** and, according to Ohm's law, it is given by $\mathbf{J} = \sigma\mathbf{E}$, where σ is the electrical conductivity of the conductor in siemens/meter (S/m). To relate this to the more familiar form of Ohm's law, consider again the example of the straight wire. If its length is L meters, a potential difference of Φ volts between its ends creates an electric field of strength Φ/L (V/m) and therefore a current of density $J = \sigma\Phi/L$ A/m^2. If the cross-sectional area of the wire is A, the total current is $I = JA = \sigma A\Phi/L = \Phi/R$, where $R = L/\sigma A$ is the electrical resistance of the wire in ohms (Ω). This result, $I = \Phi/R$, is the form of Ohm's law encountered in elementary texts on electricity and magnetism.

In pre-Maxwell theory, the flow of charge resembles the flow of a constant density fluid, for which mass conservation requires that the net flow of mass into any closed volume is equal to the net flow out of that volume. In the same way, charge conservation requires that the inward flow of electric current into the volume balances its outward flow.

Whenever current flows, a magnetic field, **B**, attends it. This is shown in Fig. 1, again for a straight wire C having a circular cross section of radius a. In this case, **B** is in the θ direction, where θ is the azimuthal angle around the wire. Some lines of (magnetic) force are shown. A line of force, also called a field line, is actually not a line but a continuous curve whose tangent is everywhere

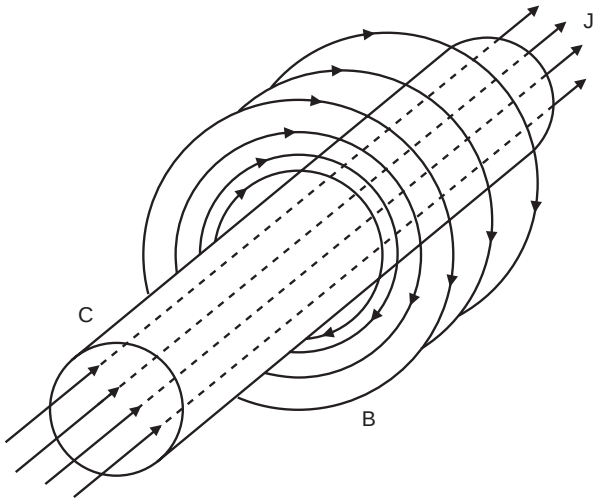


FIGURE 1 Lines of force created by a current-carrying wire C.

parallel to **B**. Although lines of force pass through every point in space, sketches such as Fig. 1 can only show a finite number. The density of lines indicates the strength of the field. Where lines bunch together (spread apart) the field is stronger (weaker). Magnetic charges, analogous to electric charges, do not exist. This means that a line of force may close on itself or cover a magnetic surface ergodically (i.e., without closing; Fig. 2), or it may continue to infinity, but it cannot terminate. It also means that (as for **J**) the net flow of **B** into any closed volume is equal to the net flow of **B** emerging from the volume. A bundle of lines of force contained in a magnetic surface form a flux tube (Fig. 3). The net flow of **B** into one end of a volume of the tube bounded by two of its cross sections is the same as the net flow that emerges from the other end. This is called the strength of the tube.

To determine **B** from **J** in a situation such as that sketched in Fig. 1, Ampère's law is required. This states that the component of **B** along any closed curve Γ is, when integrated around that curve, proportional to the net flux, I, of current through Γ. In a uniform medium, the constant of proportionality is its permeability, which in this article is always assumed to be the permeability of free space, $\mu_0 = 4\pi \times 10^{-7}$ H/m. (The fact that this value is exact is a quirk of the SI system of units, which makes use of one more quantity than is strictly necessary.) Taking the closed curve to be the (circular) line of force of radius r surrounding the wire, the required integral is $2\pi r B_\theta$, which by Ampère's law is also $\mu_0 I$, so that $B_\theta = \mu_0 I/2\pi r$; the field decreases with distance r from the wire as $1/r$. The field outside the wire does not depend on whether the current is uniform across the cross section of the wire ($J = I/\pi a^2$) or whether, as in a case considered in Section 3, it flows only along the surface of the wire with density $I/2\pi a$.

Faraday's law provides the final link between **B** and the electric field. This law applies to any closed curve Γ that is the periphery of a surface S. It states that the component of **E** along Γ, when integrated around Γ, is equal to the rate of decrease of the flux of **B** through S. This statement depends on a sign convention that can be understood from Fig. 4. If **E** is integrated around Γ in the direction of the arrows, then the flux of **B** through S is taken into the plane of the paper. (This right-handed rule also applies to Ampère's law; see Fig. 1.) Faraday's law shows that in unsteady EM states, **E** is not conservative (i.e., it cannot be expressed as $-\nabla\Phi$).

Faraday's law applies even when Γ is moving and changing shape, but in such cases **E** must be reinterpreted. To see why this is necessary, consider again Fig. 4. Suppose that **E** is zero and that **B** is uniform, constant in time and directed out of the plane of the paper. Now suppose that the sides *ad* and *bc* start to lengthen at speed U so that the area

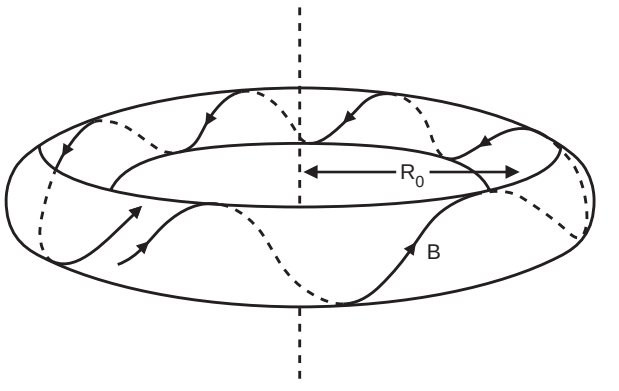


FIGURE 2 An axisymmetric magnetic surface filled ergodically by one line of force that does not close. The dashed line is the symmetry axis of the "donut"; it is also called the magnetic axis.

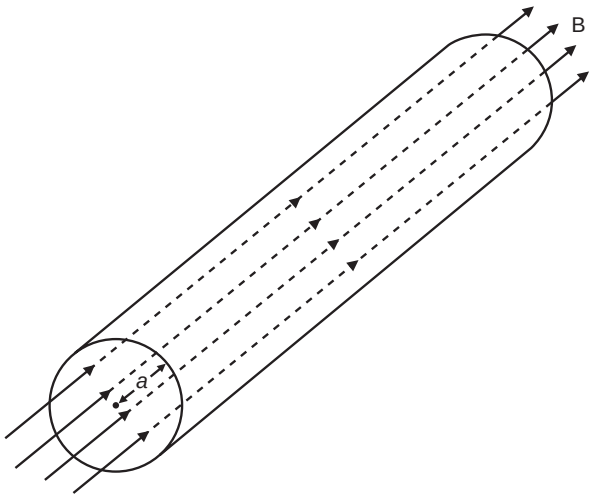


FIGURE 3 A straight flux tube. Its surface is composed of field lines.

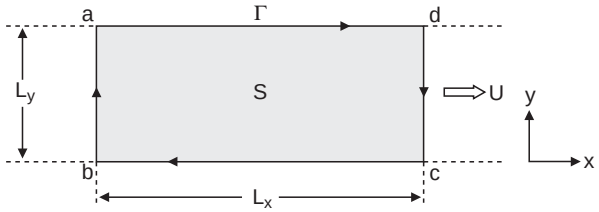


FIGURE 4 Application of Faraday's law to a growing rectangle S with periphery Γ. The sides *da, ab,* and *bc* are fixed. The side *cd* moves with speed U in the x direction. The sides *ad* and *bc* lengthen at the same rate.

S of the rectangle increases at the rate UL_y. The flux of $\mathbf{B}$ out of the plane of the paper increases at the rate BUL_y; in the positive sense defined previously, it decreases at the rate BUL_y. By Faraday's law, this must also be the integral of the component of the electric field along the sides of the integral. However, $\mathbf{E}$ is clearly zero on all sides of Γ, particularly on the stationary sides da, ab, and bc. One must conclude that when integrating the electric field around Γ, one is not integrating E_y on the moving side cd but some other electric field, E_y'. The required integral is then $E_y'L_y$ and Faraday's law gives $E_y' = -UB$.

This argument has uncovered a very significant fact: Although $\mathbf{B}$ (and $\mathbf{J}$) are the same in all reference frames, the electric field is not. In the current example, $E_y' = -UB$ in the reference frame moving with speed U in the x direction. It is convenient here to make use of a shorthand notation: $\mathbf{U} \times \mathbf{B}$ is called the vector product of $\mathbf{U}$ and $\mathbf{B}$. It is proportional and perpendicular to $\mathbf{U}$ and $\mathbf{B}$; its Cartesian components are $U_yB_z - U_zB_y$, $U_zB_x - U_xB_z$, and $U_xB_y - U_yB_x$. In this notation, the electric field in the frame moving with the side cd is $\mathbf{E}' = \mathbf{U} \times \mathbf{B}$. More generally, when the electric field in one reference frame $\mathscr{F}$ is $\mathbf{E}$, then it is $\mathbf{E}' = \mathbf{E} + \mathbf{U} \times \mathbf{B}$ in the reference frame $\mathscr{F}'$ moving with velocity $\mathbf{U}$ relative to $\mathscr{F}$. Returning to Faraday's law in the case in which Γ moves or changes shape, it is $\mathbf{E}'$ and not $\mathbf{E}$ that is integrated around Γ, where $\mathbf{E}'$ is evaluated at each point P of Γ using the velocity $\mathbf{U}$ at P.

The force on a charge e that is stationary in $\mathscr{F}'$ is $\mathbf{f} = e\mathbf{E}'$, which in $\mathscr{F}$ is more conveniently written as $\mathbf{f} = e(\mathbf{E} + \mathbf{U} \times \mathbf{B})$. Analogously, Ohm's law in a solid conductor moving with velocity $\mathbf{U}$ is $\mathbf{J} = \sigma(\mathbf{E} + \mathbf{U} \times \mathbf{B})$. To illustrate this modification of Ohm's law, suppose that the rectangle sketched in Fig. 4 is electrically conducting, with ad and bc being parallel rails along which cd slides, the electrical contact between cd and the rails being perfect. In general, a current I will flow around Γ in the direction indicated by the arrows. This current is driven through $cb + ba + ad$ by a potential difference Φ created by charges that accumulate at the sliding contacts c and d. At one extreme, when the resistance R_1 of $cb + ba + ad$ is large, $I = \Phi/R_1$ is small. By charge conservation, I is also the current flowing in cd and it is small because the electric field $E_y = (\Phi_c - \Phi_d)/L_y$ in cd created by the charges at c and d cancels out almost completely the electric field $-UB$ induced by the motion of cd; i.e., $|E_y'|$ is small. At the other extreme, when R_1 is small, the path $cd \to ba \to ad$ is almost a short-circuit, the contact charges at c and d are small, and these points are at nearly the same potential; E_y is small, $E_y' \approx -UB$, and the current $I = UBL_y/R_2$ flows around Γ in the direction indicated, where R_2 is the resistance of cd. In the general case between these two extremes, $I = UBL_y/R$, where $R = R_1 + R_2$ is the total resistance of Γ.

The modified Ohm's law also applies when the conductor is a fluid, but now $\mathbf{U}$ is replaced by the fluid velocity $\mathbf{V}$, which is usually not the same everywhere. At a point at which the fluid velocity is $\mathbf{V}$, the current density is $\mathbf{J} = \sigma(\mathbf{E} + \mathbf{V} \times \mathbf{B})$. By Ampère's law, this current will influence $\mathbf{B}$. This fact exposes one half of the relationship between $\mathbf{V}$ and $\mathbf{B}$, alluded to in Section 1, that characterizes the subject of MHD: Flow alters field in a fluid conductor. The other half of the relationship is a consequence of a force that a conductor experiences when $\mathbf{J}$ and $\mathbf{B}$ are nonzero. The force is proportional to, and perpendicular to, both $\mathbf{J}$ and $\mathbf{B}$. This is the Lorentz force, $\mathbf{J} \times \mathbf{B}$ per unit volume. Through this force, field alters the flow of a fluid conductor.

To illustrate the action of the Lorentz force, consider the conducting rectangle Γ sketched in Fig. 4. The sides ab and cd experience equal and opposite forces in the x direction, IBL_y and $-IBL_y$ respectively. These oppose the continual expansion of the rectangle (an example of Lenz's law). To maintain the expansion, forces must be applied to the sides ab and cd, and these do work at the rate $UIBL_y = (UBL_y)^2/R = I^2R$, which is precisely the rate at which electrical energy is converted into heat in the circuit Γ by its electrical resistance R. In this way, mechanical energy is converted into electrical energy and then "ohmically" into heat.

To state this result more generally, it is convenient to introduce a second shorthand notation: $\mathbf{F} \cdot \mathbf{G}$ is the scalar product of two vectors $\mathbf{F}$ and $\mathbf{G}$. In terms of their Cartesian components, it is $F_xG_x + F_yG_y + F_zG_z$. If $-\mathbf{V} \cdot (\mathbf{J} \times \mathbf{B})$ is positive at a point P, it quantifies the rate at which fluid loses its kinetic energy at P. This energy goes partially to increasing the magnetic energy density at P but some is radiated away from P through a "Poynting" energy flux $\mathbf{E} \times \mathbf{B}/\mu_0$. The remainder offsets the ohmic losses, J^2/σ per unit volume, at P. When $-\mathbf{V} \cdot (\mathbf{J} \times \mathbf{B}) > 0$ at P, the loss of kinetic energy provides a brake on the motion, which would cease even in the absence of viscous friction, unless maintained by some other force. When $-\mathbf{V} \cdot (\mathbf{J} \times \mathbf{B}) < 0$ at P, the reverse process occurs; field passes energy to motion, as in an electric motor.

The volumetric charge density ϑ in a moving fluid conductor is generally nonzero so that the fluid experiences a body force $\vartheta\mathbf{E}$ per unit volume. In the pre-Maxwell approximation, this is negligible in comparison with $\mathbf{J} \times \mathbf{B}$. These forces can be

reinterpreted in terms of stresses on the fluid and, not surprisingly, the electric stresses are negligible in comparison with the magnetic stresses. It is also found that the energy density of the electric field is insignificant compared with the energy density, $B^2/2\mu_0$, of the magnetic field. This highlights the unimportance of **E** relative to **B** in pre-Maxwell theory.

The magnetic stresses alluded to previously have two parts. One is an isotropic magnetic pressure, $p_M = B^2/2\mu_0$; the other is nonisotropic and can be interpreted as a field line tension of B^2/μ_0 per unit area, or AB^2/μ_0 for a flux tube of cross-sectional area A. Both parts increase quadratically with B. The magnetic pressure, which is already approximately 4 atm for a field of 1 T, is 10,000 times greater for $B = 100$ T.

The action of the field on the flow is recognized by including the Lorentz force $\mathbf{J} \times \mathbf{B}$ in the equation of motion that governs the fluid velocity **V**. The remaining forces affecting the motion include the gradient of the (kinetic) pressure p and a term, proportional to the (kinematic) viscosity ν, that describes frictional effects. When the Lorentz force is equivalently represented by magnetic stresses, the magnetic pressure p_M joins p in a "total" pressure $P = p + p_M$. The ratio $\beta = p/p_M = 2\mu_0 p/B^2$ is usually called beta. It is a significant quantity in plasma physics, where it is a figure of merit—one of several used to quantify the excellence or otherwise of a plasma containment device.

3. IDEAL MAGNETOHYDRODYNAMICS AND MAGNETOSTATICS

A fluid is termed ideal if it has zero viscosity and zero thermal conductivity but infinite electrical conductivity. Since **J** must be finite even though $\sigma = \infty$, Ohm's law $\mathbf{J} = \sigma(\mathbf{E} + \mathbf{V} \times \mathbf{B})$ implies that $\mathbf{E} = -\mathbf{V} \times \mathbf{B}$. Consequently, Alfvén's theorem holds: Lines of force move with the fluid as though frozen to it.

To establish this important result, let Γ be the periphery of a surface element S that moves with the fluid. Since $\mathbf{E}'$ vanishes everywhere on Γ, Faraday's law shows that the flux of **B** through S is unchanging. Suppose S is part of the curved surface C of a flux tube (Fig. 5A). The flux of **B** through S is zero because **B** is tangential to it; by Faraday's law, it remains zero as the contents of the flux tube are carried to a new location by the fluid motion **V**. Since this applies to every such S drawn on C, **B** remains tangential to C everywhere. Thus, it is still a flux tube at its new location. On shrinking the cross section of the tube to a single line of force, it is seen that this line of force moves with the fluid, as though frozen to it. A stronger statement can be made. Let Γ encircle the tube so that S becomes a cross section of the tube (Fig. 5B). The flux of **B** through S is the strength of the tube, as defined in Section 2. By Faraday's law, it is the same at its new location. This applies to every cross section. The fluid motion therefore preserves the strength of flux tubes and the integrity of the magnetic surfaces that contain them.

Alfvén's theorem provides a useful way of envisioning MHD processes in highly conducting fluids. For example, consider the eruption of a flux tube from the relatively dense region below the solar photosphere and into the more tenuous solar atmosphere (Fig. 6). Gravity drains the material in the ejected flux loop back to the solar surface, leaving the flux loop at low density and pressure, where it is in approximate equilibrium with its new surroundings. Such processes are responsible for expelling from the sun the field created within it.

Next, consider the way in which fluid motions can exchange energy with the field. Imagine a straight flux tube, initially of cross-sectional area A_0, lying in

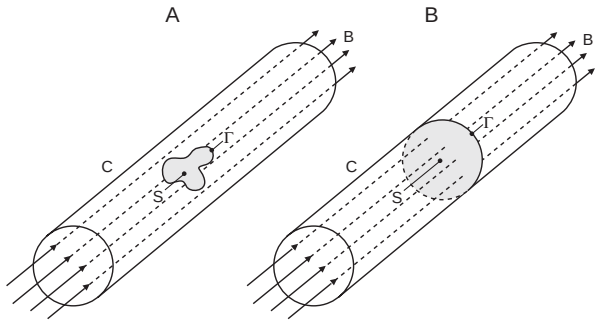


FIGURE 5 (A) An element S of surface area lying on the surface of a flux tube with periphery Γ. (B) A cross-section S of the flux tube with periphery Γ.

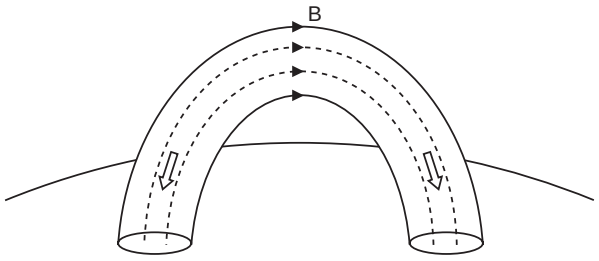


FIGURE 6 Fluid pulled by gravity down a flux tube that has erupted from the surface of the sun. Field directions are indicated by single arrows and fluid motion by arrowheads.

a compressible fluid. Suppose a motion compresses the tube uniformly so that, while remaining straight, its cross-sectional area is reduced to $A(<A_0)$. The density of the fluid it contains increases from its initial value of ρ_0 to $\rho = \rho_0(A_0/A)$. Since (by assumption) the fluid does not conduct heat, the tube compresses adiabatically, and its internal energy (i.e., the energy of the random molecular motions within it) increases so that its temperature rises. Because the strength of the tube cannot change, the field B_0 initially within the tube increases to $B = B_0A_0/A$ (i.e., by the same factor as ρ_0); thus, B/ρ remains at its initial value, B_0/ρ_0. The magnetic energy per unit length contained by the tube, however, increases from $A_0(B_0^2/2\mu_0)$ to $A(B^2/2\mu_0) = [A_0(B_0^2/2\mu_0)](A_0/A)$. In short, the kinetic energy of the compressing motion has been transformed into internal energy and magnetic energy. The reverse happens if $A > A_0$; the flux trapped in an intense tube tends to expand into surroundings where the field is weaker. For these reasons, sound traveling across a field moves faster (and much faster when β is small) than it would in the absence of the field; the magnetic pressure created by the field intensifies the restoring force in the compressions and rarefactions of the wave.

The transformation of kinetic energy into magnetic energy also occurs in an incompressible fluid when motions stretch field lines against their tension. The process may be likened to the storing of elastic energy in a stretched rubber band or to the energy transmitted to a violin string by plucking it. If the flux tube of cross-sectional area A_0 containing field B_0 considered previously is lengthened from L_0 to L, its cross section will decrease by the same factor ($A = A_0L_0/L$) and the field within it will increase by the same amount ($B = B_0L/L_0$). The magnetic energy it contains, which is proportional to B^2, is enhanced by a factor of $(L/L_0)^2$ from $(B_0^2/2\mu_0)L_0A_0$ to $(B^2/2\mu_0)LA = (B^2/2\mu_0)L_0A_0 = [(B_0^2/2\mu_0)L_0A_0](L/L_0)^2$. If $L = L_0 + \delta$, where $\delta \ll L_0$, the increase in magnetic energy is $(B_0^2/\mu_0)A_0\delta$. This is the work that the applied force had to do against the magnetic tension $(B_0^2/\mu_0)A_0$ of the field lines in stretching the tube by δ.

Tension has other important effects. Consider again the system shown in Fig. 1, but suppose now that the cylinder C is a compressible fluid or plasma surrounded by a vacuum. Suppose that the current I flows along the surface of C so that $I/2\pi a$ is the surface current density. The current I produces a field of strength $B_\theta = \mu_0 I/2\pi a$ on C. Because of their curvature, the field lines exert a "hoop stress" on C of $B_\theta^2/2\mu_0$ per unit area and directed radially inward. This compresses the fluid in C. Since $\mathbf{B}$ is zero in C, only (kinetic) pressure p can oppose the hoop stress. If $p = B_\theta^2/2\mu_0$, the forces are in balance, and the configuration is in *MSE*.

One may imagine that this MSE is brought about in the following way. Initially, A is large and p is small. After the current I and its associated field B have been set up, the forces on C are not in balance and the hoop stresses exerted by the field lines encircling C pinch the plasma column in much the same way as a stretched elastic band squeezes what it encircles. As these lines of force shorten and A becomes smaller, magnetic energy is converted into internal energy. The interior of C becomes hotter and its pressure p increases to oppose the further contraction of the field lines. Plausibly, the final result is an MSE of the kind envisaged previously, in which a plasma column has been strongly heated and prevented from expanding by the encircling field. Hot plasma is then confined within a magnetic bottle and away from solid walls. The possibility of this kind of magnetic confinement is of considerable interest in the MFE field. A hot reacting plasma must be prevented from leaching impurities from the solid walls of the reactor because these would greatly enhance the radiative losses, cooling the plasma and quenching the reactions.

In an MSE, the Lorentz force $\mathbf{J} \times \mathbf{B}$ balances the pressure gradient ∇p. This means that $\mathbf{B}$, being perpendicular to ∇p, is tangential to the constant–p surfaces, which are therefore magnetic surfaces, as defined in Section 2, and can be labeled by their value of p. Current lines (curves drawn parallel to $\mathbf{J}$) also lie on these surfaces and, like the field lines, they may close or traverse the surfaces ergodically. The total pressure P is continuous across the special magnetic surface that is the boundary of a plasma; if the exterior is a vacuum, P must balance the magnetic pressure, $B^2/2\mu_0$, in the vacuum.

If a region exists in which $\mathbf{J} \times \mathbf{B} = 0$, the field within it is termed force-free. Since $\nabla p = 0$, the pressure p is constant in the region, and p is no longer available to label surfaces on which the field and current lines are constrained to lie. Since $\mathbf{J}$ and $\mathbf{B}$ are parallel in a force-free field, $\mathbf{J} = \lambda\mathbf{B}$, where λ is constant on each field line, this being the component $J_\parallel = \mathbf{J} \cdot \mathbf{B}/B$ of $\mathbf{J}$ parallel to $\mathbf{B}$ on that line. The parameter λ may act as a surrogate for p in labeling the magnetic surfaces and constraining the field and current lines. If, however, λ is also constant, the field and current lines become unconstrained. A single field line and a single current line may fill the region. These are termed stochastic lines. They magnetically connect every part of the region to

every other part. Particular cases of this are presented in Sections 4.6 and 5.3.2.

The MSE just described is called the Z pinch because the current flows parallel to the axis of C, a direction usually labeled by z. The Z pinch is shown in Fig. 7A, together with two other MSEs. In Fig. 7B, the θ pinch, the current flows around the plasma column, a direction often denoted by θ. Between these extremes is the screw pinch (Fig. 7C). Here, the current flows in helices around C. The pitch $q_J = J_z/J_\theta$ of the helices depends on r, the radius of the cylindrical magnetic surface on which they lie. The field lines are also helices with a different pitch, q. The way that q varies with r is a significant factor for the stability of the equilibrium, and dq/dr is another figure of merit, called magnetic shear (see Section 5.3). It quantifies how rapidly the direction of the magnetic field changes as a function of r.

As applied to plasma devices, an obvious shortcoming of the linear or one-dimensional pinches shown in Fig. 7 is that, in practice, they must have ends, and these are sources of contamination. Greater interest therefore centers on toroidal or two-dimensional pinches. One may visualize a toroidal MSE as a linear pinch whose ends are joined together by bending it around on itself to form a donut of radius R_0, similar to that sketched in Fig. 2; R_0 is called the major radius and a the minor radius of the torus. The axis of symmetry is called the magnetic axis and distance from it is denoted by R. The z coordinate of the linear pinch is replaced by the angle ϕ around the magnetic axis, and increasing ϕ is called the toroidal direction. The coordinates r and θ of the linear pinch are approximately the same for both the linear and toroidal systems, r now being the distance from the axis running through the center of the donut and θ being the angle around that axis. Increasing θ is now called the poloidal direction. There is no toroidal MSE analogous to the θ pinch; in the absence of current flow in the toroidal direction, the Lorentz force cannot be balanced by a pressure gradient. Main interest centers on the toroidal screw pinch. The quantity $q = rB_\phi/R_0B_\theta$ is another figure of merit called the kink safety factor (see Section 5.3.2).

In addition to the one- and two-dimensional MSEs, there are also interesting three-dimensional configurations, so-called because they have no symmetry. Of fundamental concern in all these MSEs is the stability of the equilibria. This topic is discussed in Section 5.

Alfvén's theorem leads to a new phenomenon, the Alfvén wave. Consider again the straight flux tube of cross-sectional area A. Suppose first that the fluid is incompressible, and imagine that a transverse displacement bends the tube, carrying its contents with it, in obedience to the theorem (Fig. 8). The tension $\tau = AB^2/\mu_0$ of the tube acts, as in a stretched string, to shorten the tube. A wave results that moves in each direction along the tube with speed $V_A = \sqrt{(\tau/\Sigma)}$, where $\Sigma = \rho A$ is the mass per unit length of the string, with ρ being the fluid density. The wave velocity is therefore $\mathbf{V}_A = \mathbf{B}/(\mu_0\rho)^{1/2}$, which is usually called the Alfvén velocity. It is also the velocity with which energy can be transmitted along the field lines. In a compressible fluid, the situation is more complex because the presence of the field and its associated Lorentz force make sound propagation anisotropic. For example, the speed s of sound waves traveling parallel or antiparallel to the field $\mathbf{B}$ is unaffected by it: $s = \sqrt{\gamma p/\rho}$, where γ is the ratio of specific heats. For sound traveling perpendicular to the field, $s = \sqrt{(\gamma p + B^2/\mu_0)/\rho}$. For small β this is approximately V_A.

The MHD or Alfvénic timescale, $\tau_A = \mathscr{L}/V_A$, is very significant in many MHD contexts. It provides an estimate of how quickly a system responds to changes in its state. Such changes generate acoustic and Alfvén waves, the former crossing the system in a time of order $\tau_s = \mathscr{L}/s$, which for $\beta \ll 1$ is indistinguishable from τ_A. Thus, τ_A is the time required for the initial change to permeate the system. In particular, it is the dynamic timescale on which an MSE responds to a perturbation, and it is usually the timescale on which that equilibrium will, if unstable, be disrupted. It should be noted, however, that even if an equilibrium is stable according to ideal MHD, it

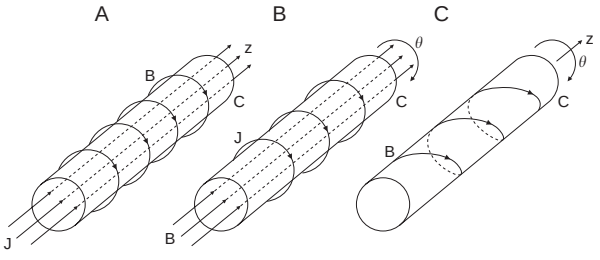


FIGURE 7 (A) Z pinch; (B) θ pinch; (C) screw pinch.

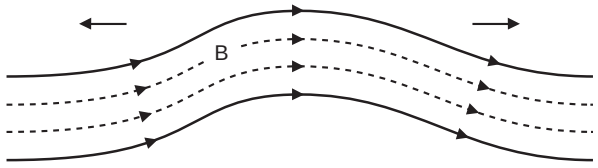


FIGURE 8 Alfvén wave propagation down a straight flux tube. The arrows indicate the directions of wave propagation.

may be unstable through diffusive effects ignored in ideal MHD. Such resistive instabilities are discussed in Section 5.

4. NONIDEAL MAGNETOHYDRODYNAMICS

4.1 Quantifying Deviations from the Frozen in Law

Alfvén's theorem poses a paradox: If field and fluid are inseparable, how did the fluid acquire its field in the first place? The answer is by diffusion, by which is meant, by diffusion of magnetic field. This happens because every real fluid has a finite conductivity.

In the reference frame $\mathscr{F}'$ moving with a particular point P of the fluid, Ohm's law is $\mathscr{J} = \sigma\mathscr{E}'$, in order of magnitude. Ampère's law shows that $\mathscr{J} = \mathscr{B}/\mu_0\mathscr{L}$, whereas Faraday's law implies that $\mathscr{E}' = \mathscr{B}\mathscr{L}/\mathscr{T}$. The timescale $\mathscr{T}$ on which **B** at P can change as P moves around is therefore the magnetic diffusion time, defined as $\tau_\eta = \mu_0\sigma\mathscr{L}^2 = \mathscr{L}^2/\eta$, where $\eta = 1/\mu_0\sigma$ is the magnetic diffusivity. On timescales $\mathscr{T}$ short compared with τ_η, the field is frozen to the fluid; for $\mathscr{T} \gg \tau_\eta$, it is not. Unless a field of scale $\mathscr{L}$ at P is maintained in some way, its lifetime is τ_η; also, a field of this scale cannot be implanted at P in a lesser time. One may consider these phenomenon as a type of EM memory that sets a minimum timescale for change, in the frame moving with the conductor.

Consider again the Z pinch described previously. After the potential difference has been applied between the ends of C at time $t = 0$, the current it creates initially flows only over the surface of C because, over short times, the plasma behaves as an ideal conductor in which **B** in the bulk of C cannot change from its initial zero value. Thereafter, the current sheet penetrates into C by diffusion, its thickness being of order $\delta_\eta = \sqrt{(\eta t)}$ at time t. After a magnetic diffusion time of $\tau_\eta = a^2/\eta$, the current will be distributed almost uniformly across the cross section of the column. As a digression, in practice the process may take rather longer since the diffusion of field into C is hampered by the increasing conductivity of the plasma as its temperature T rises through its adiabatic compression or heating. Similarly, the $z-$directed field initially trapped in a screw pinch grows as the column C collapses, whereas B_z in the surrounding vacuum does not. The tendency for the flux in C to diffuse out of the column is hampered by the diminishment of η that accompanies the compressional increase in T.

As another example, suppose that a plasma column C, such as the screw pinch shown in Fig. 7C, is contained within, but is not in contact with, a surrounding metallic wall. The field created outside C after the pinch has been initiated at time $t = 0$ cannot at first penetrate the walls because, over short times, they behave as perfect conductors in which **B** cannot change. The field is excluded from the walls by currents that flow in a thin skin on their surfaces, and these create their own field, which (when the walls are planar) is the mirror image of that of the plasma column (Fig. 9). The plasma column is repelled by its image (another example of Lenz's law), thus, preventing it from striking the wall. This stabilizing effect weakens in time because the thickness $\delta_\eta = \sqrt{(\eta t)}$ of the skin increases with t until ultimately the field completely penetrates the walls.

In a moving fluid, the advective timescale, $\tau_v = \mathscr{L}/\mathscr{V}$, quantifies the time taken (according to Alfvén's theorem) for a field of scale $\mathscr{L}$ to be carried over that distance. However, electrical resistance diffuses the field on a timescale of τ_η. If the magnetic Reynolds number, $Rm = \tau_\eta/\tau_v = \mathscr{L}\mathscr{V}/\eta$, is large, advection is more rapid than diffusion and Alfvén's theorem is useful in visualizing MHD processes. When $Rm \ll 1$, as is usually the case in technological applications of MHD involving liquid metals, the concept of frozen fields is not very useful. The prevailing field then provides an anisotropic friction that attempts to damp out motions perpendicular to itself on a timescale of $\tau_d = \rho/\sigma V_A^2 = \eta/V_A^2$, often called the magnetic damping time or the Joule damping time.

Energy can also be transmitted in a fluid by waves, and particularly significant for MHD systems is

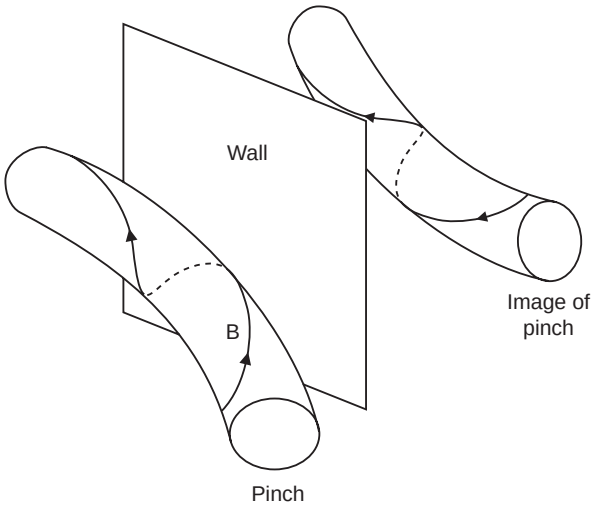


FIGURE 9 Image of a screw pinch in a conducting wall.

Alfvén radiation. This introduces a second way of quantifying diffusive effects: the Lundquist number, Lu. Alfvén waves are damped both by electrical conduction and by the viscosity ν of the fluid. The timescale over which this occurs is a combination $\tau_c = \mathscr{L}^2/(\eta + \nu)$ of τ_η and the viscous diffusion time, $\tau_\nu = \mathscr{L}^2/\nu$. For a liquid metal ($\nu \ll \eta$), $\tau_c \approx \tau_\eta$, and we shall now use τ_η in place of τ_c. During a time τ_η, the Alfvén waves travel a distance, $V_A\tau_\eta$. If the Lundquist number, $Lu = V_A\tau_\eta/\mathscr{L} = V_A\mathscr{L}/\eta$, is large, this distance greatly exceeds the characteristic length scale $\mathscr{L}$, and Alfvén radiation is significant. If $Lu \ll 1$, this is not the case since the time $\tau_d (= \eta/V_A^2)$ taken for the wave to damp out is small compared with the time $\tau_A (= \mathscr{L}/V_A)$ that the Alfvén wave requires to traverse the distance $\mathscr{L}$. The Lundquist number can be thought of as a magnetic Reynolds number, with the wave speed V_A replacing the advection speed $\mathscr{V}$.

4.2 Magnetic Reconnection

According to Alfvén's theorem, the topology of field lines cannot change in ideal MHD, but in a finitely conducting fluid ($\eta > 0$), field lines can sever and reconnect. This can happen fastest where scales are smallest (i.e., where the gradients of **B** are greatest). Magnetic reconnection is the process that forms thin current layers that change field topology and dissipate magnetic energy, even though the resistivity is modest and smoothly varying. Indeed, it is very important to distinguish magnetic reconnection, a process in which very thin (nearly singular) current sheets form in otherwise smoothly varying systems, from simple diffusive dissipation of magnetic energy when sharp profiles are built in *ab initio*. In magnetic reconnection, resistive diffusion and the magnetofluid dynamics combine to release energy stored in the magnetic fields and to reconfigure the magnetic topology of the system. Reconnection phenomena are usually subdivided into two classes, namely driven magnetic reconnection (discussed here) and spontaneous magnetic reconnection (described in Section 5 in the context of the tearing instability).

Perhaps the simplest and most fundamental model of magnetic reconnection is that of Sweet and Parker, hereafter referred to as the Sweet–Parker reconnection (SPR) model. This is a two-dimensional machine that steadily reconnects field lines. A cross section is sketched at successive times in Fig. 10. Resistivity acts everywhere but is most effective near a segment of the x axis. To simplify the argument, it is supposed that $\eta = 0$ everywhere except in reconnection region (shown

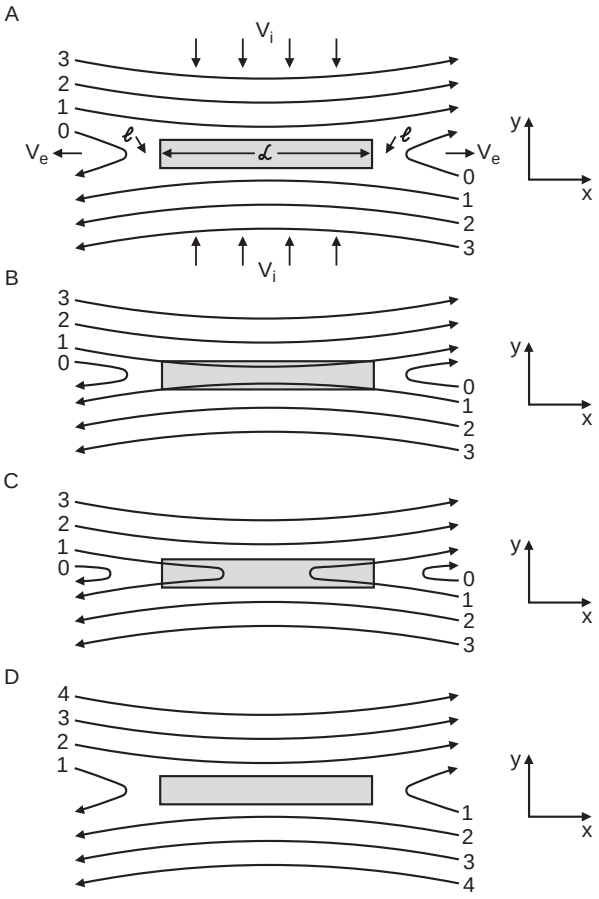


FIGURE 10 The Sweet–Parker reconnection mechanism. (A) Reconnection occurs when fluid motion brings slabs of oppositely directed magnetic fields together. The slabs have length $\mathscr{L}$ and velocity V_i. Pushing the slabs of field together creates a current sheet of thickness ℓ and ejects fluid at velocity V_e. The field lines labeled 1–3 approach the current sheet and field line 0 leaves it. (B) Field lines 1 enter the current sheet. (C) Field lines 1 are reconnected in the current sheet. (D) Field lines 1 leave the current sheet and move away in the $\pm x$ directions with velocity V_e. The configuration of field lines is now identical to that in A, but field lines n take the place of field lines n–1. Reconnection has changed the field line topology and dissipated magnetic energy.

shaded) of x width $\mathscr{L}$ and y width ℓ; it is assumed that ℓ is so much smaller than $\mathscr{L}$ that it is almost a current sheet. A steady inflow, V_i, in the $\pm y$ directions forces together two slabs of oppositely directed field $\pm B$ (Fig. 10A). The fluid is incompressible so that fluid must be ejected, with velocity V_e in the $\pm x$ directions. Figure 10B shows the field lines, labeled 1, moving into the reconnection region, and Fig. 10C shows them moving out after they have been reconnected. Their curvature implies a tension that assists the ejection of fluid in the $\pm x$ directions.

The object now is to determine V_i, V_e, and ℓ, assuming that the configuration is in a steady state

and that the Lundquist number $Lu = V_A\mathscr{L}/\eta$ is large. The argument rests on three simple balance laws:

1. Balance of fluid mass flowing into the reconnection layer with that flowing out: This gives $V\mathscr{L} = V_e\ell$.
2. Balance of fluid momentum in the reconnection layer. Since the flow in the reconnection layer is approximately one dimensional, it is governed by Bernoulli's law, which states that the sum of the total pressure P and the dynamic pressure $\frac{1}{2}\rho\ V^2$ is constant. Furthermore, the total pressure P is, in turn, the sum of the thermal pressure p and the magnetic pressure $p_M = \mathbf{B}^2/2\mu_0$. The sum, $S \equiv p + \mathbf{B}^2/2\mu_0 + \frac{1}{2}\rho\ V^2$, is therefore constant throughout the layer. By symmetry, the center of the configuration is a stagnation point ($\mathbf{V} = 0$), and at the midpoints on the long sides of the reconnection layer $\mathbf{V}$ is negligibly small so that $S \approx p + B^2/2\mu_0$. Outside the layer (i.e., for $|x| \gg \mathscr{L}$), the magnetic field is small so that at the ends of the layer where fluid is ejected in the $\pm x$ directions, $S \approx p + \frac{1}{2}\rho\ V_e^2$. Equating these two expressions for S shows that the ejection velocity is simply the Alfvén velocity: $V_e = V_A$.
3. Balance of magnetic energy: In the steady state, the influx of magnetic energy to the layer must equal the rate of ohmic dissipation in the layer. The influx rate is simply $(2V\mathscr{L})B^2/2\mu_0$ since the efflux from the narrow side edges of the reconnection region is negligible. The rate of ohmic dissipation is $(\mathscr{L}\ell)J^2/\sigma$, where by Ampère's law, $J = B/\mu_0\ell$. Since $\eta = 1/\mu_0\sigma$, the inflow velocity is related to the layer thickness by $V = \eta/\ell$.

Taken together, these three demands imply that

1. Since $Lu \gg 1$, the reconnection velocity, $V = V_A/\sqrt{Lu}$, is much smaller than V_A. It is proportional to a fractional power of the resistivity of the system; the growth rate of the tearing mode (discussed in Section 5) shows a similar fractional dependence.
2. The aspect ratio, $\ell/\mathscr{L}$, of the reconnection region is small, of order $Lu^{-1/2} \ll 1$. The direction of the merging fields imposes this anisotropy. It is consistent with the idea that magnetic energy is dissipated in a thin current sheet. Of course, the simultaneous smallness of V/V_A and the aspect ratio $\ell/\mathscr{L}$ is a consequence of their equality, by the balance law, and of the assumption $Lu \gg 1$.

Although conceptually simple and plausible, SPR tells a rather pessimistic story, namely that reconnection is a relatively slow process (i.e., $V = V_A/\sqrt{Lu} \ll V_A$, where $Lu \gg 1$ in cases of greatest interest). Many physical phenomena appear to require, for their explanation, fast reconnection (i.e., reconnection on ideal timescales or, equivalently, at Alfvénic velocities $V \sim V_A$). This is especially true of many solar and astrophysical phenomena, such as solar flares. The physics of fast reconnection is an active area of research in MHD and plasma physics.

Two main routes to fast reconnection have been proposed. Within the framework of MHD, one originally proposed by Petschek suggests that shock waves carry energy away from the reconnection layer. This rapid process of energy removal is thought to increase both the aspect ratio of the reconnection region and the reconnection rate. The Petschek mechanism predicts a reconnection velocity of order V_A, with only a weak logarithmic dependence on Lu [i.e., $V \sim V_A/\ln(Lu)$]. The alternate route to fast reconnection appeals to small-scale, non ideal, or kinetic processes that cannot be studied using the one-fluid model of MHD. All models of fast reconnection are controversial and they will evolve considerably in the years ahead.

4.3 MHD Turbulence

Many fluid systems are turbulent. This fact presents formidable obstacles to theorists and is likely to embarrass them for the foreseeable future. The motions and fields of a turbulent fluid have many length and timescales. In the classical picture of turbulence, these range from macroscales, $\mathscr{L}_{\text{macro}}$ and $\mathscr{T}_{\text{macro}}$, corresponding to scales on which the fluid is stirred, to dissipation scales, $\mathscr{L}_{\text{diss}}$ and $\mathscr{T}_{\text{diss}}$, which are dominated by viscosity. It is argued that $\mathscr{L}_{\text{macro}}/\mathscr{L}_{\text{diss}} \sim (Re)^{3/4}$, where $Re = \mathscr{V}_{\text{macro}}\mathscr{L}_{\text{macro}}/\nu$ is the Reynolds number. This is typically greater than 10^4, so the range of scales spans many decades and cannot be adequately resolved even by modern, high-speed computers.

In classic turbulence, interscale interaction occurs via a process of cascade in which larger eddies (i.e., fluctuations that have the appearance of vortical whorls) fragment into smaller eddies, thus producing a range of fluctuations on scales from $\mathscr{L}_{\text{macro}}$ to $\mathscr{L}_{\text{diss}}$. This process is encapsulated by a well-known parody by Louis Fry Richardson of a verse of Jonathan Swift:

> *Big whorls have little whorls, that feed on their velocity.*
> *And little whorls have lesser whorls, and so on to viscosity.*

In contrast to ordinary fluid turbulence, which can be thought of as a soup of eddies, MHD turbulence consists of a mixture of eddies and small-scale Alfvén

waves and is controlled by viscosity and resistivity. Thus, although the idea of a cascade remains useful for describing MHD turbulence, the interaction processes in MHD are considerably more complex than in nonmagnetic fluids. Perhaps the most important difference is that, since MHD turbulence involves Alfvén waves, it necessarily retains an element of memory or reversibility. Here, Alfvénic memory refers to the tendency of even small-scale magnetic fields (which are always present when Rm is large) to convert the energy in eddies (which cascades irreversibly) to reversible nondiffusive Alfvén wave motion. This is a consequence of the fact that wave motion is necessarily periodic and involves a restoring force.

Despite the difficulties discussed previously, there is considerable effort in the research community directed at numerical simulation of MHD turbulence. The large scales can be resolved numerically in so-called large eddy simulations (LES), but only if the effects of the unresolvable or subgrid scales (SGS) on the large scales are parameterized in some way. One popular, although controversial, expedient originated from an idea by Osborne Reynolds in the 19th century. He drew an analogy between the diffusive effects of the small-scale eddies and collisional diffusion processes at the molecular level. This suggested that the SGS similarly spread out the large-scale motions and fields diffusively, but on timescales $\tau_{\nu/\mathrm{turb}} = \mathscr{L}^2/\nu_{\mathrm{turb}}$ and $\tau_{\eta/\mathrm{turb}} = \mathscr{L}^2/\eta_{\mathrm{turb}}$ determined by turbulent diffusivities, ν_{turb} and η_{turb}, of order $\mathscr{L}_{\mathrm{eddy}}\mathscr{V}_{\mathrm{eddy}}$ that might be very large compared with the corresponding molecular values, ν_{mol} and η_{mol}.

Reynolds's idea is not totally satisfactory but is qualitatively useful. For example, the fact that $\eta_{\mathrm{turb}} \gg \eta_{\mathrm{mol}}$ in the solar convection zone explains, in a rough and ready way, why the timescale $\mathscr{T}$ of magnetic activity on the sun, as estimated by the period of the solar cycle, is on the order of a decade, even though $\tau_{\eta/\mathrm{mol}}$ exceeds the age of the sun: $\mathscr{T}$ is determined by η_{turb} and not η_{mol} so that $\mathscr{T} \approx \tau_{\eta/\mathrm{turb}}$.

Reynolds's idea also correctly implies that turbulence enhances viscous and ohmic dissipation. Although the magnetic Reynolds number Rm may be large when defined from $\mathscr{L}_{\mathrm{macro}}$ and $\mathscr{V}_{\mathrm{macro}}$, it may be small when $\mathscr{L}_{\mathrm{eddy}}$ and $\mathscr{V}_{\mathrm{eddy}}$ are used instead. Even if the characteristic macroscale field, $\mathscr{B}_{\mathrm{macro}}$, is large compared with the typical eddy field, $\mathscr{B}_{\mathrm{eddy}}$, so that the magnetic energy resides mostly in the macroscales, the magnetic energy may be dissipated mainly by the eddies. This is because $\mathscr{L}_{\mathrm{eddy}} \ll \mathscr{L}_{\mathrm{macro}}$ so that $\mathscr{J}_{\mathrm{macro}} = \mathrm{O}(\mathscr{B}_{\mathrm{macro}}/\mu_0\mathscr{L}_{\mathrm{macro}})$ may be smaller than $\mathscr{J}_{\mathrm{eddy}} = \mathrm{O}(\mathscr{B}_{\mathrm{eddy}}/\mu_0\mathscr{L}_{\mathrm{eddy}})$, even though $\mathscr{B}_{\mathrm{eddy}} \ll \mathscr{B}_{\mathrm{macro}}$. In many situations however, $\mathscr{B}_{\mathrm{eddy}} \gg \mathscr{B}_{\mathrm{macro}}$ so that the eddies contain most of the magnetic energy and are mainly responsible for its dissipation. Similarly, they may contain the majority of the kinetic energy and be the main cause of its viscous dissipation. The reader is cautioned, however, that the expressions for ν_{turb} and η_{turb} given here are estimates only. In particular, it has been appreciated that the Alfvénic memory intrinsic to MHD turbulence may cause significant reductions in ν_{turb} and η_{turb} in comparison with the estimates given here. The self-consistent calculations of ν_{turb} and η_{turb} are active topics of current research in MHD.

MHD propulsion and drag reduction are topics of considerable interest and usually concern systems for which $Rm \gg 1$ and in which MHD turbulence occurs. An extensive discussion of these topics is beyond the scope of this article.

4.4 Buoyancy in MHD

A real (nonideal) fluid has a nonzero thermal conductivity, K. The appropriate measure of the resulting conduction of heat is the thermal diffusivity, $\kappa = K/\rho C_p$, where C_p is the specific heat at constant pressure; the thermal diffusion time is $\tau_\kappa = \mathscr{L}^2/\kappa$. The finiteness of κ gives rise to the magnetic buoyancy of flux tubes. Such tubes are created by, for example, the turbulent motions in a stellar convection zone, such as that of the sun. These cause the prevailing magnetic field to become intermittent, with regions of strong field (i.e., flux ropes) being surrounded by regions of comparatively weak field. During a time $t \ll \tau_\eta$ after its formation, flux loss from a tube can be ignored. Radiative transport of heat, however, is very effective in a star and $\tau_\kappa \ll \tau_\eta$. Any difference in temperature between the rope and its surroundings diffuses away in a time of order τ_κ. It is reasonable to suppose that in the time interval $\tau_\kappa \ll t \ll \tau_\eta$, the temperature of the rope is the same as its surroundings but that the field B in the interior of the rope differs from the field $B_0 (< B)$ outside it. Magnetostatic balance requires that the total pressure exerted by the rope on its surroundings is approximately the same as the total pressure the surroundings exert on it: $p + B^2/2\mu_0 = p_0 + B_0^2/2\mu_0$. This means that the kinetic pressure p of the gas in the tube is less than that of its surroundings, p_0. However, since its temperature is the same, its density must be less. The tube is therefore buoyant and rises toward the stellar surface.

When the tube breaks through the stellar surface, dark regions occur that are called starspots (or sunspots in the case of the sun). That these spots are dark is another illustration of the constraining effects of the field. Near the stellar surface, heat is carried outward mainly by convective overturning, which within a tube is coupled to Alfvén waves that radiate energy and therefore tend to suppress the convective motions. The emerging heat therefore has to find a route that goes around the spots rather than through them. The spots are thus cooler and darker than their surroundings.

4.5 Amplification of Magnetic Fields in MHD

A magnetic field threading a sphere of solid iron of radius 1 m would disappear in approximately 1 s unless maintained by external current-carrying coils. Since $\tau_\eta \propto \mathscr{L}^2$, larger bodies can retain their fields longer, but many celestial bodies (planets, stars, galaxies, etc.) are magnetic and are believed to have been magnetic for times long compared with their magnetic diffusion times. The age of the geomagnetic field, for instance, exceeds 3×10^9 years, even though τ_η is less than 10^5 years. A process must exist that replenishes these fields as they diffuse from the conductor. This process is the same as that operating in commercial power stations: self-excited dynamo action. From a field **B**, fluid motion creates the electromotive force (emf) $\mathbf{V} \times \mathbf{B}$ that generates the electric currents **J** necessary to create, by Ampère's law, **B**.

Although it was stated previously that in turbulent flow eddies may be responsible for greatly enhancing diffusion of the large-scale part, $\mathbf{B}_{\text{macro}}$, of **B**, these small-scale motions can, if they lack mirror symmetry, also act to replace the lost flux. The simplest manifestations of such symmetry breaking are helical motions. Helicity is defined as $\mathbf{V} \cdot \boldsymbol{\omega}$ where $\boldsymbol{\omega}$ is the vorticity of the flow, a vector that has Cartesian components $\partial V_z/\partial y - \partial V_y/\partial z$, $\partial V_x/\partial z - \partial V_z/\partial x$, $\partial V_y/\partial x - \partial V_x/\partial y$ and that is often written as curl**V** or $\nabla \times \mathbf{V}$. Helicity is a measure of how mirror symmetric the flow is; like a common carpenter's screw, a helical flow does not look the same when viewed in a mirror. Helicity is created naturally by the rotation of large systems such as the solar convection zone, where the rising and falling motions created by buoyancy acquire vertical vorticity through the action of the Coriolis force. Small-scale helical motions create a large-scale emf that, in the simplest case, is proportional to $\mathbf{B}_{\text{macro}}$ and is conventionally written $\alpha \mathbf{B}_{\text{macro}}$. This α effect may suffice to maintain $\mathbf{B}_{\text{macro}}$ and hence $\mathbf{B}_{\text{turb}}$ as well, thus creating a turbulent dynamo. The α effect is the cornerstone of a subject now called mean field electrodynamics. (Confusingly, this is often known by the acronym MFE, which we reserved as an abbreviation for magnetic fusion energy.)

The α effect illustrates the danger, mentioned previously, of adopting the Reynolds ansatz too uncritically; through their anisotropy and lack of mirror symmetry, small-scale motions can create large-scale emfs that are not recognized when η_{mol} is merely replaced by η_{turb}. It should also be mentioned that the issue of Alfvénic memory appears once again in the context of the dynamo and the α effect. Recent research has suggested that because of memory effects and constraints on the rate of magnetic reconnection, the α effect at large Rm may be much weaker than it would be were the dynamical effects of the small-scale eddies ignored, as was done in the early days of mean field theory. Further developments in this important research area may be expected in the near future.

The interested reader can get an intuitive, hands-on perspective on a particular type of dynamo process by the following home demonstration, which also illustrates several of the processes described in this section and the last. Obtain an elastic (rubber) band. This loosely corresponds to the flux tube in Fig. 11A, its tension being the sum of the tensions in the magnetic field lines it contains (i.e., it is proportional to the energy of the field within it).

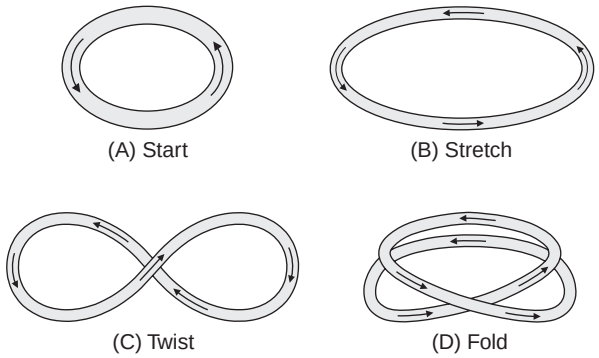


FIGURE 11 The stretch–twist–fold dynamo. (A) The initial undeformed flux tube; (B) the result of stretching it to twice its length; (C) the stretched tube is twisted into the shape of an infinity sign ∞; and (D) the result of folding one loop of the ∞ over onto the other. Reconnection in D can produce two loops of the same size as the initial tube.

The tension in the band may be systematically increased by a three-step, stretch–twist–fold process:

1. Stretch the rubber band to double its length.
2. Twist it into the shape of an infinity sign, ∞.
3. Fold the loops on top of one another to form two linked loops of the same size as the original band.

These steps are illustrated in Fig. 11B–D. In a similar way to the rubber band, the energy within and the tension of the flux tube are increased by a factor of 4. This is the essence of a dynamo, the creation of magnetic energy by helical motion.

In principle, these steps can be repeated over and over again, with the sense of twist in step 2 being always the same. This gives the band the sense of handedness of a screw motion and the broken reflection symmetry and helical motion that are crucial to the success of the dynamo. The tension in the band (analogous to the magnetic tension) increases progressively as the steps are repeated, and at some stage the reader's hands will tire, making it impossible to increase the tension further. This illustrates the end of the kinematic regime (in which fluid motion is prescribed *a priori* and the field grows) and the beginning of the dynamic regime (in which the field quenches its own further growth). If the reader lets go of the band at any stage, it will immediately relax back to its initial state. This illustrates the crucial importance of magnetic reconnection as a means of locking in the amplified magnetic field. In other words, magnetic reconnection provides the crucial element of irreversibility in the dynamo. The analogue of magnetic reconnection in the case of the rubber band could be achieved by fusing the two loops together after the stretch–twist–fold process.

In recent years, much effort has been expended on the study of fast dynamos, which are defined as processes of field amplification that operates on timescales independent of τ_η in systems with $Rm \gg 1$. It should be apparent to the reader that a fast dynamo necessarily requires fast reconnection in order to lock in the dynamo-generated field. Thus, the problems of fast reconnection and those of the fast dynamo are intimately linked.

4.6 Magnetic Helicity and Relaxation in MHD

No doubt the reader has noticed that the past three sections, which discuss magnetic reconnection, turbulent transport of magnetic fields, and the dynamo process, all concern the following broader question: Given certain initial conditions, what magnetic configuration does the magnetofluid ultimately adopt? This question seeks to identify the final state when some very complex dynamical processes, involving the dissipative phenomena mentioned previously (as well others, some unknown), are at work.

Apart from a very few simple, exactly soluble cases (that nearly always have an unrealistically high degree of symmetry built in), there are two main ways of finding the final state.

The first is to integrate the MHD equations directly using a computer. This approach forces the researcher to confront all the thorny issues of MHD turbulence, SGS modeling, etc. described previously. Moreover, such brute force tactics are often expensive and inefficient. One is naturally motivated to seek other approaches that are bolder but simpler. Foremost among these is constrained relaxation. This variational method seeks to identify the final state as one that minimizes the magnetic energy, subject to certain constraints. The central issue then becomes the identification and inclusion of the most important constraint or constraints.

A natural candidate for consideration as a constraint is the magnetic helicity. This is the analogue of the (kinetic) helicity, which was previously defined as $\mathbf{V} \cdot \boldsymbol{\omega}$, where $\boldsymbol{\omega} = \mathrm{curl}\mathbf{V}$. Conversely, $\mathbf{V}$ can be derived from $\boldsymbol{\omega}$ by integration in an uncurling operation. Likewise, a vector $\mathbf{A}$ can be obtained from $\mathbf{B}$ by a similar uncurling operation. This defines a vector potential, $\mathbf{A}$. The scalar product $K = \mathbf{A} \cdot \mathbf{B}$ is the magnetic helicity. It is conserved in the motion of an ideal fluid, and this is why it is significant in MHD. Physically, magnetic helicity quantifies the self-linkage or knottedness of the magnetic field lines and is thus a measure of the topological complexity of the field. Indeed, the invariance of field line topology in ideal MHD is the origin of the conservation of magnetic helicity.

If an MSE is analyzed using ideal MHD, it is found that magnetic helicity is conserved in detail in the sense that if we regard the MSE as an assembly of infinitesimally thin flux tubes of volume Δv and labeled by parameters α and β, their individual magnetic helicities, $K_{\alpha,\beta} = (\mathbf{A} \cdot \mathbf{B})_{\alpha,\beta} \Delta v$, are conserved, although they are not the same for each tube. If instead nonideal MHD is used in analyzing the MSE, the configuration will relax toward its final state on a timescale determined by the resistivity and other slow diffusion mechanisms. In this process,

individual flux tubes lose their identity through magnetic diffusion and turbulent cascade. Thus, the local magnetic helicity, $K_{\alpha,\beta}$, associated with a given thin flux tube, is *not* conserved. However, J. B. Taylor made the insightful conjecture that the global magnetic helicity (i.e., the sum of $K_{\alpha,\beta}$ over the system) would be approximately conserved on the timescale of the relaxation.

Global helicity is the integral of K over the entire system. It is, in some sense, a coarse-grained topological invariant that is also related to the self-inductance of the plasma. Taylor relaxation theory proposes to determine relaxed states by minimizing magnetic energy subject to the constraint of constant global magnetic helicity. The resulting Taylor states have the property that the current is linearly proportional to the field (i.e., $\mathbf{J} = \lambda \mathbf{B}$, where λ is a constant). This, is a force-free field of the special type described in Section 3 for which the component, $J_{\|}$, of $\mathbf{J}$ parallel to $\mathbf{B}$ is everywhere the same. In Section 5.3.2, we discuss kink-tearing instabilities of toroidal MSE. These instabilities are current driven (i.e., they exist only because $\nabla J_{\|}$ is nonzero). Since $\nabla J_{\|} = 0$ in the Taylor relaxed state, it is stable to kink-tearing modes.

Taylor relaxation theory has been very successful in explaining the relaxed mean field profiles of low β confined plasmas, especially the reversed field pinch (RFP) and spheromak. Of particular note is Taylor's successful prediction of the F–Θ curve for the RFP. The F–Θ curve relates the degree of field reversal to the poloidal field at the edge of the RFP and is an experimentally derived signature of the final RFP state. In relation to Section 4.4, it is worth noting that the relaxed state is maintained by a cyclic dynamo process in which dissipative excursions from the Taylor state trigger instabilities that drive the plasma back to relaxed state.

The main mystery of Taylor relaxation theory is why it works so well. In other words, why is the global helicity the most important constraint on turbulent relaxation? There are at least three plausible answers:

1. *Enhanced dissipation:* Turbulence and dissipation drive magnetic reconnection, which destroys domains of local magnetic helicity on all but the largest scale. Thus, as the system evolves, the global magnetic helicity is the only surviving topological invariant.
2. *Field line stochasticity:* In the initial state of the configuration, the field lines lie on magnetic surfaces, each labeled by its hydrostatic pressure p. However, as the configuration relaxes to the force-free state of constant p, the identity of the surfaces evanesces, and field lines become increasingly stochastic. Ultimately, as the turbulent relaxation proceeds, the entire MSE is threaded by a single field line. Since magnetic helicity is calculated by volume integration over a region enclosed by a magnetic surface, and all other magnetic surfaces are destroyed, only global magnetic helicity is relevant to a stochastic state.
3. *Selective decay:* In three-dimensional MHD turbulence, energy cascades to small scales. It can be shown, however, that magnetic helicity cascades inversely to large scales. As a result, magnetic helicity accumulates on the largest scales of the system, with minimal coupling to dissipation. Thus, magnetic helicity is dynamically rugged, whereas energy decays.

It should be noted that these are only plausibility arguments, and that a rigorous justification of the Taylor hypothesis remains an elusive goal of current research.

5. STABILITY OF MAGNETICALLY CONFINED PLASMAS

5.1 Generic Issues

Configurations of MHD fluids are analyzed using the concepts of equilibrium and stability. These are used to classify a variety of possible configurations according to their intrinsic interest, potential utility, importance for science and technology, and so forth. They are vital in the context of magnetic plasma confinement for the success of controlled fusion (hereafter referred to as MFE). This is the only application considered here, but there are many others in astrophysics, geophysics, space plasma dynamics, and MHD power generation.

MHD equilibria are configurations in which the plasma pressure gradient, ∇p, balances the $\mathbf{J} \times \mathbf{B}$ force (Lorentz force) and any other body forces acting on the magnetofluid. An equilibrium in which the balance is dominated by the pressure gradient and the Lorentz force is referred to as a *MSE*. All equilibria of interest in the context of MFE are MSEs. In general, the magnetic fields in MSEs are produced in part by external coils and in part by plasma currents. They are characterized by the pressure and current profiles $p(r, \theta, \phi)$ and $\mathbf{J}(r, \theta, \phi)$. In toroidal MSEs, these are independent of ϕ, the angle around the torus in the toroidal

direction, but still depend on θ, the angle that increases in the so-called poloidal direction.

Interesting MSEs almost always have closed magnetic flux surfaces (i.e., surfaces that enclose a specified amount of magnetic flux). These are labeled by an effective radius r, and because they are so significant, we generally suppress the dependence of p and $\mathbf{J}$ on θ and ϕ. Charged particles in a plasma cannot readily cross a flux surface and are therefore confined. For this reason, an MSE is sometimes called a magnetic bottle.

MSEs have a high degree of symmetry, corresponding to closed toroidal, spheroidal, or helical configurations. Toroidal MSEs of interest in the context of MFE include the tokamak (based on a Russian acronym for maximum current) and the RFP. Together, these two configurations are referred to as toroidal pinches since plasma is confined by a toroidal current that produces a magnetic field, which in turn acts to pinch the plasma column. A tokamak has a strong, externally applied toroidal magnetic field, whereas a RFP plasma generates its own toroidal field by a process of constrained relaxation, whereby some of the toroidal current is dynamically twisted into the poloidal direction. This relaxation process is closely related to that of the dynamo described previously. RFPs are intrinsically dynamic (i.e., not quiescent) in nature. Spheroidal MSEs include the spheromak and the spherical torus, whereas helical MSEs include the heliac, a type of stellarator. Stellarator MSEs do not rely on inductively driven currents as do toroidal pinches, but they sacrifice the virtues of toroidal symmetry. In this section, the discussion is limited for reasons of brevity and clarity to toroidal pinches in general, and tokamaks in particular. The MSE of a tokamak is particularly simple in that the pressure and toroidal current profiles, $p = p(R)$ and $J_\phi = J_\phi(R)$, can be parameterized by a single coordinate, R, that may be thought of as a (generalized) distance from the symmetry axis.

Once the existence of a closed MSE is established, the next question is whether it is magnetohydrodynamically stable—that is, whether initial, small perturbations that break the symmetry of the equilibrium tend to reinforce themselves and grow. This growth is due to the release of potential energy stored in the initial MSE and is accompanied by a relaxation of the pressure and/or current profiles $p(r)$ and $J_\phi(r)$, which tends to lower the initial potential energy and degrade the quality of the nascent MSE. Instability is a linear concept and is relevant for only a few growth times (at most), until the dynamics enter the nonlinear regime, characterized by finite (i.e., not small) amplitude perturbations. Instabilities saturate (i.e., cease growing) either when the plasma arrives at a new secondary equilibrium, which is a MSE with lower potential energy than the initial equilibrium, in which the symmetry is usually broken as well, or when the kinetic energy of the plasma motion is lost through dissipation (e.g., through viscosity). The latter often occurs as a consequence of the cascade to small scales characteristic of MHD turbulence.

MHD instabilities are usually classified as either pressure driven [when they are associated with relaxation of $p(r)$] or current driven [when associated with relaxation of $J_\phi(r)$]. They are either ideal or resistive. Ideal instabilities occur without a violation of Alfvén's frozen flux constraint, whereas resistive instabilities require some decoupling of field and fluid, usually by collisional resistivity, to trigger the energy-release process. MHD instabilities are of great interest in the context of MFE, because they limit the class of viable MSEs and thus severely constrain the performance of magnetic fusion devices.

Entire books have been devoted to the MHD stability of confinement devices for MFE. There is space here to describe only a few of the basic concepts and their application to MFE. One way of investigating dynamical stability makes use of the fact that ideal MHD conserves the total energy E of the system. This is the sum of the kinetic energy E_K, the magnetic energy E_M, and the internal energy of the plasma, E_I. Before perturbation, $\mathbf{V} = 0$ and therefore $E_K = 0$ so that $E = E_M + E_I$. The perturbation creates a small initial E_K and this may grow at the expense of $E_M + E_I$, indicating that the initial state is unstable; if the initial state is linearly stable, E_K remains small.

The situation may be described conceptually by analogy with a ball rolling under gravity on a surface S. Then E_K is the kinetic energy of the ball and its height z above some reference level plays the role of $E_M + E_I$. A magnetostatic equilibrium is analogous to the ball being at rest at some point P where the surface is level (i.e., where the tangent plane at P to S is horizontal, as at the points indicated in Fig. 12). In Fig. 12A, the ball rests at a global minimum of z, and the system is not only linearly stable but also globally stable (i.e., stable to all perturbations no matter how large they are or in which horizontal direction they occur). In Fig. 12B, the ball is at an unstable equilibrium that is destroyed if perturbed in any horizontal direction. Figure 12C also illustrates an unstable state; although it is true that some

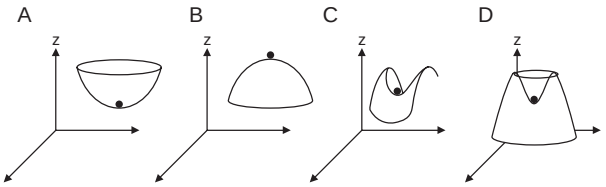


FIGURE 12 Equilibrium points (•) of a ball rolling on a surface. (A) global stability; (B) instability at a maximum of z; (C) instability at a saddle point of z; (D) stability for small perturbations but not large.

perturbations increase z (suggesting stability), other directions of perturbation exist that will cause the ball to descend permanently from the saddle point. In Fig. 12D, a sufficiently large perturbation will take the ball over an adjacent hill, after which it descends permanently to a state remote from its starting point (i.e., the initial state is linearly stable but unstable to finite amplitude perturbations).

Similarly, if $E_M + E_I$ increases for every small perturbation of a MSE, it is linearly stable, as in Figs 12A and 12D. The MSE is unstable if any perturbation exists that reduces $E_M + E_I$, as in Figs 12B and 12C. Figure 12C is the better analogue for a MSE since many directions of perturbation do not threaten a MSE; perturbations that bend field lines increase E_M and those that compress the plasma increase both E_M and E_I. Other directions of perturbation, however, may exist that reduce $E_M + E_I$, and these are the perturbations that should be understood and, as far as possible, eliminated. These directions are usually associated with perturbations that relax the current or pressure gradients.

5.2 Magnetically Confined Plasmas and MFE

The hope of controlling the release of thermonuclear energy by the fusion of light nuclei provides a very powerful incentive for the study of plasma physics and MHD. To make light nuclei react and release energy, they must be forced together against their mutual electrostatic repulsion. This happens naturally at the enormous pressures characteristic of the interiors of stars, such as the sun. The aim of MFE, of course, is to create high plasma pressure MSEs using magnetic fields. The efficiency of the MSE is determined (in part) by the dimensionless parameter β, (which is the ratio of the plasma pressure p to the magnetic pressure $B^2/2\mu_0$. For fusion to occur, high pressure must be sustained long enough for the ignited plasma to replace its energy, via the fusion burn, faster than it loses it through leakage of energy across the confining fields. Thus, both high plasma pressure and a sufficiently long plasma energy confinement time, τ_E, are necessary for the success of MFE. The Lawson criteria for ignition of a fusion burn are that the ion temperature T_i exceeds $4.5 \times 10^7 K$ and that the product of density and energy confinement time $n\tau_E \geq 10^{20}\,\mathrm{m}^{-3}\,\mathrm{s}$. These two conditions can be combined into a single criterion often referred to as the Lawson triple product. This requires that $nT\tau_E > 4.5 \times 10^{27}\,\mathrm{m}^{-3}\,\mathrm{Ks}$. This product is proportional to $\beta B^2 \tau_E$, indicating that high beta and large fields are desirable for fusion, assuming that τ_E does not degrade rapidly with either. In addition, high β is intrinsically desirable for economic reasons.

The required Lawson number is lower for a DT plasma than for a D plasma. Tritium, which is virtually nonexistent in nature, may be bred in a lithium blanket surrounding the ignited plasma in which alpha particles escaping from the burning DT plasma react to create T, which can subsequently be extracted to fuel the fusion reactions. Of course, the commercial viability of MFE demands even better performance. The fusion reactions occurring during confinement must recoup the cost of creating and heating the plasma and all other overheads of the system, such as collecting T from the lithium blanket and extracting the spent fuel from the reaction chamber. This necessarily raises the required Lawson triple product in comparison with the value needed for ignition. In practice, Lawson triple products in excess of $2.4 \times 10^{28}\,\mathrm{m}^{-3}\,\mathrm{K\,s}$ are thought to be required for a fusion power plant producing 1 GW of power.

The performance of a magnetic confinement device is determined by

1. *The magnetic configuration*: the external magnetic fields and plasma currents that together define the geometry of the MSE and the strength of the magnetic bottle. Today's tokamaks rely heavily on optimization of the magnetic geometry by shaping of the cross section to achieve peak performance.
2. *Heating power and method*: the means by which the plasma is heated. These include ohmic heating, neutral beam injection, radiofrequency heating, and self-heating by the slowing down of fusion-generated alpha particles. Most scenarios for achieving ignition require auxiliary (i.e., non-ohmic) heating, usually via injection of radiofrequency waves that resonate with a characteristic frequency (known as the cyclotron frequency) of a class of plasma particles.

3. *Fueling*: the means by which fuel is injected. These include gas-puffing, pellet injection, and beam fueling.

4. *Boundary control*: the means by which the deleterious effects of impurity accumulation and other undesirable consequences of plasma–wall interaction are minimized by control of the plasma boundary. This is usually accomplished by a special magnetic configuration called a divertor. This diverts plasma outside the confinement region to a separate chamber, where interaction with walls occurs.

5. *Momentum input*: the means, usually via neutral beam injection, whereby plasma rotation is generated. Toroidal rotation is desirable for optimizing energy confinement and controlling certain types of MHD instabilities.

The principal limitations on fusion plasma performance are

1. *Collisional transport*: There is an irreducible minimum to the cross-field losses, that is, for a given magnetic geometry, *some* losses through diffusion are due to collisions between charged particles (analogous to, but dynamically different from, molecular diffusion in fluids) and these are unavoidable.

2. *Radiative losses*, especially due to *impurities* in the plasma.

3. *Macroscopic MHD instabilities*: MHD instabilities are large-scale perturbations that constrain possible pressure and current profiles. Robust ideal instabilities set hard limits on $p(r)$, $\mathbf{J}(r)$, β, etc. that place an upper bound on the possible Lawson number. Resistive instabilities may set soft limits (i.e., bounds that can be exceeded but only at unacceptably high cost in, for example, heating power). Both ideal and resistive instabilities can lead to disruptions, which are catastrophic events that result in termination of the plasma discharge and may also damage the confinement vessel.

4. *Microinstabilities*: These are small-scale instabilities, driven by local temperature and density gradients, leading to small-scale turbulence that degrades energy and particle confinement through turbulent transport. Microinstabilities typically require a description in terms of two (or more) fluids or one using kinetic equations. Discussion of these is beyond the scope of this article. Microinstabilities typically produce cross-field leakage of energy and particles that far exceeds that due to collisions. They essentially control τ_E, which enters the Lawson triple product.

Instabilities are the key players that limit achievable Lawson numbers. MHD phenomena typically determine β and the plasma pressure p, whereas microinstabilities determine τ_E. Thus, the interplay of macro- and microinstabilities is a recurring theme in the design of MFE devices.

MHD instabilities have played a prominent role in the history of MFE, especially in elucidating the behavior of tokamaks. The early predictions of rapid progress in MFE relied on naive expectations based only on considerations of collisional and radiative losses. Fusion researchers soon learned, however, that an evil genie named instabilities lurked in the magnetic bottle and was only too anxious to escape. The instabilities it created were difficult to control. Thus, the hopes of the MFE pioneers were dashed by the struggle against unexpected, premature termination of discharges by disruptions created by MHD instabilities. The genie had to be persuaded to remain in the bottle. The instabilities had to be understood and minimized or eliminated.

Much, indeed most, research on tokamaks in the late 1960s and 1970s focused on current-driven instabilities called kink-tearing modes, which can disrupt the current profile and thus the discharge. The output of this phase of fusion research was a greatly improved understanding of the parameter space of viable tokamak operations, particularly with respect to current profile and magnetic field configurations. The emphasis of tokamak research in the 1980s and early to mid-1990s shifted to microinstabilities and the turbulent transport associated with them. The aim here, simply stated, was to understand and predict the parameter scaling of τ_E. Great progress was made.

A watershed in this line of research was reached with the discovery of spontaneously formed transport barriers at the edge of, and within, the plasma. A transport barrier is a localized region in which turbulence due to microinstabilities is greatly reduced or extinguished so that cross-field leakage declines to very low, collisional levels. This, in turn, results in the formation of regions of steep pressure gradient. Transport barriers usually form via a spontaneous transition to a state of strongly sheared poloidal and/or toroidal flow, which tears apart the eddies driven by microinstabilites before they can cause significant leakage. Present-day transport barriers can sustain temperature gradients that seem almost incredibly large, in excess of 1 million degrees per inch. Profiles with and without transport barriers are shown in Fig. 13.

The discovery and exploitation of transport barriers stimulated a renewed interest in MHD stability in the late 1990s, a trend that continues to

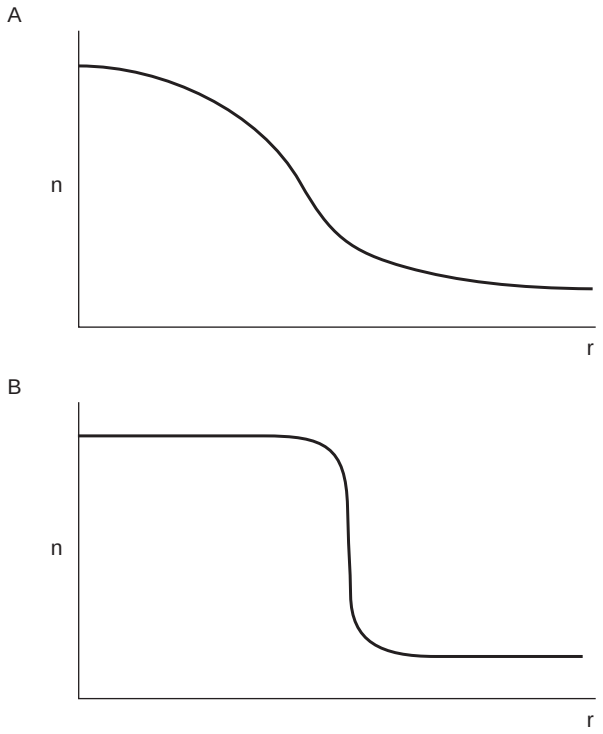


FIGURE 13 Profiles with and without transport barrier. (A) Normal profile; low confinement. (B) Profile with transport barrier; high confinement. Particle density is denoted by n.

this day. In confining the plasma better, a transport barrier also creates a larger pressure gradient in the plasma column that brings it closer to the verge of ideal MHD instability. In the jargon of MFE, enhanced confinement is necessary to reach the β limit. Thus, much of the current research on MFE is devoted to avoiding or mitigating the effects of pressure-driven ideal MHD instabilities via cross-section shaping, active feedback techniques, and profile control. Paradoxically, after 40 years of research on how to exterminate microinstabilities, effort is now being expended on finding ways to stimulate them at opportune times. The aim is to reap the full benefits of transport barriers by avoiding disruptions while at the same time controlling impurity accumulation.

5.3 Examples of MHD Instabilities in MFE Plasmas

As mentioned previously, the MHD instability of MFE plasmas is an exceptionally complex topic but is the subject of several excellent monographs. Here, we give only an introduction to the major pressure gradient and current gradient instability

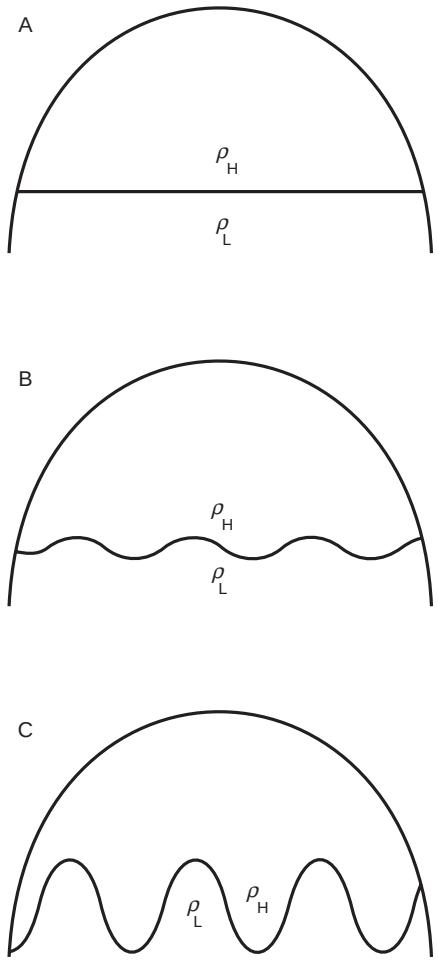


FIGURE 14 Development of Rayleigh–Taylor instability. (A) Initial equilibrium, with heavy fluid (density ρ_H) on top of light fluid (density ρ_L). (B) Ripples form on the interface. (C) Ripples are amplified by Rayleigh–Taylor instability.

mechanisms; interchange-ballooning modes and kink-tearing modes, respectively.

5.3.1 Interchange-Ballooning Modes

These modes, like all ideal MHD instabilities driven by the pressure gradient, are related to Rayleigh–Taylor instability, familiar from classical (nonmagnetic) fluid dynamics. Rayleigh–Taylor instability occurs when ripples are excited on the interface between a heavy fluid (e.g., water) sitting atop a lighter fluid (e.g., air) in a gravitational field, as shown in Fig. 14. Everyone knows what happens: The water falls out of the glass. What is not so apparent is that the initial configuration of the water and air is in equilibrium, but that the equilibrium is unstable to the growth of ripple perturbations. The experimentally inclined reader (with inexpensive carpeting) may easily convince himself or herself of

this by filling a glass to the brim, placing a strong piece of cardboard (preferably two-ply) over the surface of the water and in contact with the rim of the glass, and then inverting the glass. Because the presence of the cardboard prevents the formation of surface ripples, the water will not fall, so long as the integrity of the cardboard is maintained. Of course, the free energy source for this instability is simply the gravitational potential energy stored in the initial configuration (i.e., the elevated heavy fluid); the gradient that relaxes is simply the density gradient.

The analogue of the Rayleigh–Taylor instability in the MHD of magnetically confined plasmas is the interchange instability. Interchange instability relaxes the gradient in density, or more generally in the pressure gradient, by lowering the effective gravitational potential energy of the system. It does this by interchanging a tube of high-pressure fluid with a tube of low-pressure fluid, as shown in Fig. 15. In a magnetically confined plasma, the role of gravity is played by the centrifugal force exerted on the plasma particles as they traverse curved field lines. This results in a net body force that resembles a gravitational force. Thus, if field lines curve or sag away from regions of higher pressure, the system is interchange unstable, whereas if they curve or sag toward regions of higher pressure, the system is interchange stable (Fig. 16). Equivalently, an interchange stable system is said to have favorable curvature, whereas one that is unstable is said to have unfavorable curvature. Also, it is important to realize that the conceptual image of an interchange of two plasma tubes is motivated by the fact that potential energy release will be maximal when the perturbation does not spend any portion of its energy budget on bending field lines. In other words, a pure interchange instability does not couple to Alfvén waves. In reality, the physical appearance of an interchange instability resembles that of a convection roll (Fig. 17).

In magnetic confinement devices of practical interest (including tokamaks), the magnetic curvature is not constant but, rather, varies along the field lines. Interchange stability is then determined by an average of the magnetic curvature over the extent of the field line. One key virtue of the tokamak is that its configuration has favorable average curvature and thus is interchange stable. However, as one follows a field line around the tokamak, one traverses regions of locally unfavorable curvature on the outboard side of the torus and locally favorable curvature on the inboard side. Thus, if perturbations are larger and stronger in regions of locally unfavorable curvature

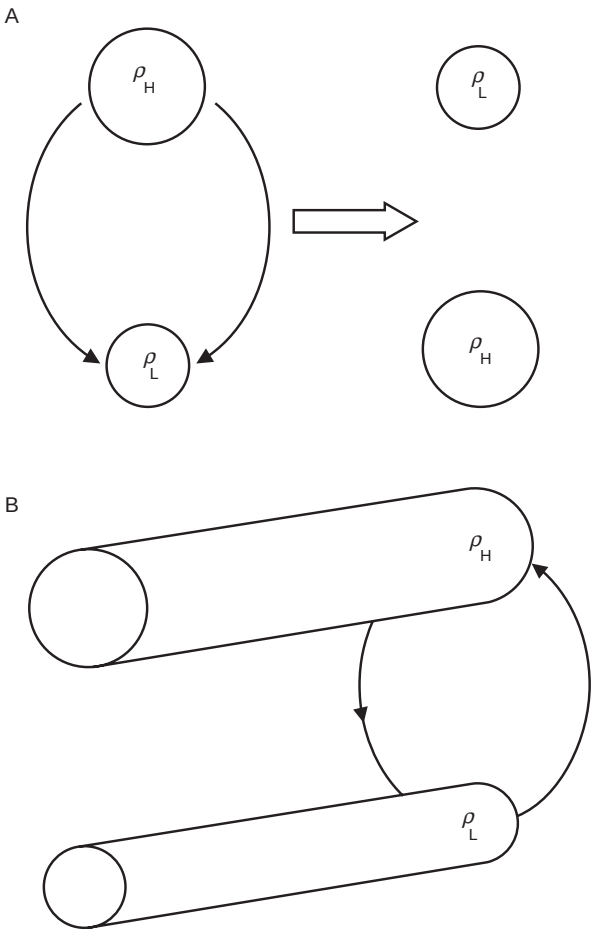


FIGURE 15 Interchange instability. (A) Interchange instability switches tubes of heavy and light fluid (densities ρ_H and ρ_L, respectively. (B) Oblique view of tube interchange; note that tubes, which are aligned with field lines, are not bent.

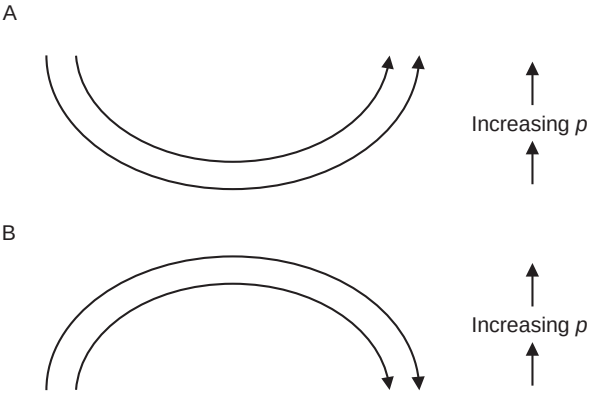


FIGURE 16 Favorable and unfavorable curvature. (A) Unfavorable curvature: Field lines sag away from the region of high pressure p. (B) Favorable curvature: Field lines curve toward the high-pressure region.

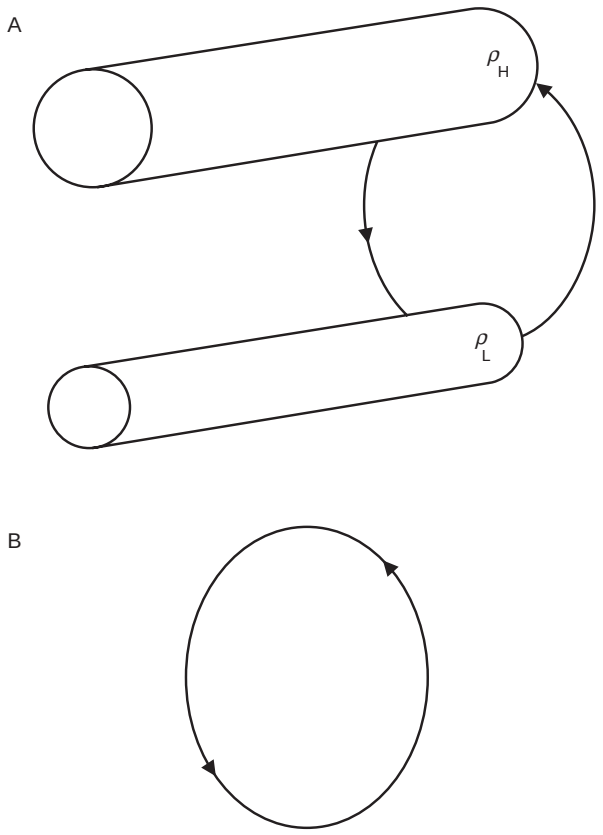


FIGURE 17 Depiction of an interchange convection roll; see legend to Fig. 15.

than they are in the locally favorable regions, instability is possible. These are called ballooning instabilities because the perturbation in the plasma pressure balloons the plasma outward in regions of unfavorable curvature. A physical picture of a ballooning instability may be gained by thinking of what happens when one overinflates a bicycle tire. Of course, the tire eventually ruptures. If there is no flaw in the tire material, the rupture will be likely to occur on the outboard portion of the wheel because the inboard portion is supported by the structural frame of the wheel and is thus mechanically more resilient. Indeed, sometimes an aneurysm will form in the tire. Finite-amplitude ballooning may be thought of as a kind of aneurysm in the magnetic surface that is driven by high plasma pressure interior to that surface. Finite-amplitude ballooning-kink instabilities are shown in Figs 18 and 19. Figure 18 shows development of the instability in a poloidal cross-section. Notice how fingers form on the outboard side of the torus. Figure 19 shows the magnetic flux surfaces. Notice how the ballooning instability produces crenelations in these.

Although ballooning instabilities are closely related to interchange instabilities, they differ crucially in that the perturbations vary along the field lines and therefore must bend the field lines. Thus, ballooning instabilities couple localized interchange motions to Alfvén waves. For ballooning instability to occur, the energy released by interchange motions in the unfavorable regions must exceed the energy expended in bending the magnetic field lines. Thus, there is a minimal or critical pressure gradient required for ballooning instability to occur. This is in sharp contrast to the interchange mode (in a system with a curvature that is on average unfavorable), which, in ideal MHD, can occur for any pressure gradient. The critical pressure gradient for ballooning instability plays a central role in the ultimate determination of the maximum achievable β for a specific configuration (i.e., its beta limit). In practice, the magnetic geometry and magnetic field strength determine the critical pressure gradient for ballooning instability. One important feature of magnetic geometry is the magnetic shear, which parameterizes how rapidly the direction of field lines changes as a function of radius r. Magnetic shear plays an important and subtle role in the detailed dynamics of ballooning instabilities. A complete discussion of this is beyond the scope of this article.

5.3.2 Kink-Tearing Modes

Kink-tearing modes are all driven by the current gradient, although the symptom of instability is often related to the pitch of the magnetic field lines. The physical nature of the modes is best illustrated by a sequential discussion of sausage instabilities, kink instabilities, and tearing instabilities.

Consider the Z pinch (Fig. 7A). Figure 20 illustrates a type of perturbation, sometimes called a sausage mode, which is axisymmetric around the axis of the pinch. The existence of the sausage mode of instability can be explained rather simply. As discussed in Section 2, the field B_θ outside the plasma column, created by a current I flowing down it, is azimuthal in direction and of strength $\mu_0 I/2\pi r$. On the surface of the column, the field strength is therefore $\mu_0 I/2\pi a$, where a is the local radius of the column, which varies when the tube is pinched. The pinched column then resembles a tube of sausage meat that has been compressed to create a chain of links. The surface field and the associated field line tension are greater where the column necks and smaller where it bulges. This is the field that confines the plasma column, and since I is constant, the tension it exerts on the necks of the chain further

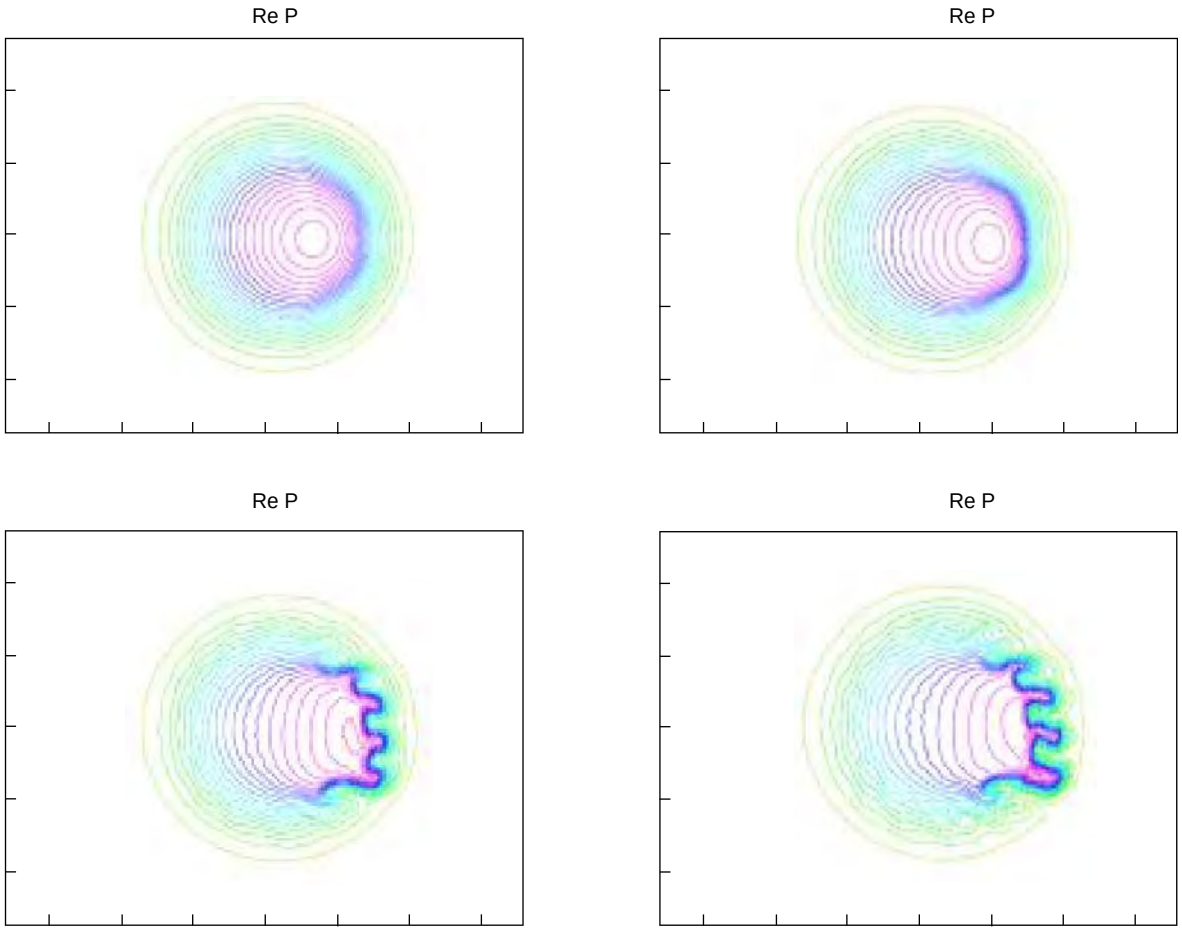


FIGURE 18 Development of a ballooning mode as seen in a poloidal cross section of the torus. The magnetic axis is on the left of each panel. The sequence of panels is as follows: top left, top right, bottom left, and bottom right. Note that perturbations are larger on the outboard midplane regions (on the right). Also note that as the perturbation grows, it evolves to form fingers or spikes. Courtesy of S. C. Jardin and the Center for Extended MHD Modeling, Princeton Plasma Physics Laboratory.

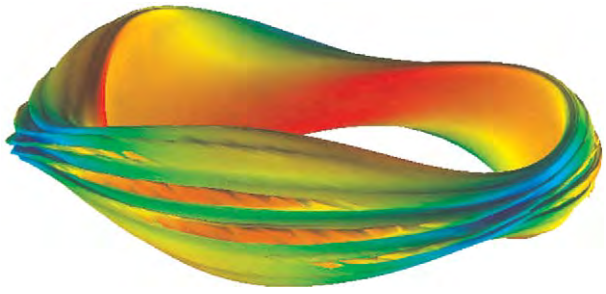


FIGURE 19 Development of a ballooning-kink mode in a torus. Note how a side view of the finger formation reveals the tendency of the ballooning mode to crenelate the plasma and break it into ribbons. Courtesy of S. C. Jardin and the Center for Extended MHD Modeling, Princeton Plasma Physics Laboratory.

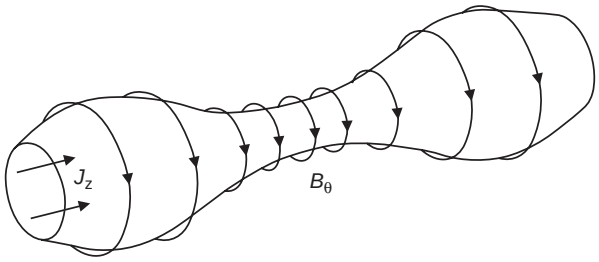


FIGURE 20 The sausage mode of instability of a Z pinch; lines of force encircling a current-carrying plasma column are shown. J_z indicates the cross section through which the current enters.

increases the initial deformation of the column. This self-reinforcing, positive feedback loop is indicative of an instability process that tends to break up the column into a chain of plasmoids (i.e., plasma droplets). For obvious reasons, the instability is called the sausage instability. The mechanism of sausage instability is closely related to that causing a thin stream of water from a faucet to break up into a

line of spheroidal droplets. In this case, surface tension plays the role of field line tension.

The sausage instability may be stabilized by an axial field B_z of sufficient strength within the column, which provides a tension that opposes the formation of necks. The critical strength of B_z necessary for stabilization can be estimated by comparing the field energy released by necking of the column with the energetic costs of deforming the axial field B_z. Because necking motions must conserve axial magnetic flux, the change, $\Delta(B_z^2/2\mu_0)$, in energy density of the axial field is $-B_z^2\Delta a/\mu_0 a$, where $\Delta a/a<0$. The corresponding change, $\Delta(B_\theta^2/2\mu_0)$, in the azimuthal field energy density is $B_\theta^2\Delta a/2\mu_0 a$. It follows that the net change in magnetic energy is positive if $B_z^2>\frac{1}{2}B_\theta^2$. This gives the required axial field strength necessary to stabilize the instability, as depicted in Fig. 21. Note that when the plasma tube is bent into a torus, one arrives at a simple model of the tokamak configuration.

The dual presence of both a poloidal field and a strong, vacuum toroidal field greatly improves the MHD stability of a toroidal pinch by eliminating both the sausage instability and the related bending instability shown in Fig. 22 and by introducing magnetic shear, which is useful in controlling pressure-driven instabilities. Indeed, the tokamak

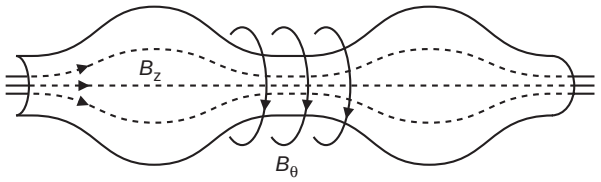


FIGURE 21 Stiffening of the *Z* pinch against sausage modes of instability by the addition of a field B_z along the plasma column (creating a screw pinch).

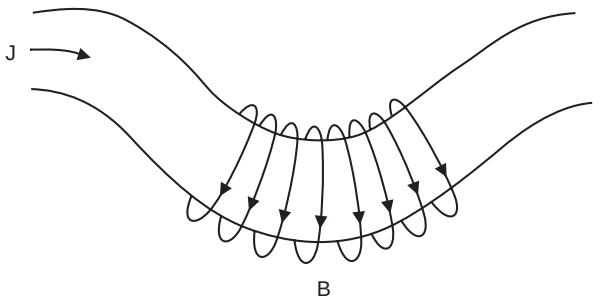


FIGURE 22 The bending mode of instability of the *Z* pinch. Note how the lines of force are crowded together on the inside of the bend and are moved apart on the outside of the bend; the concomitant difference in magnetic pressure acts to reenforce the displacement.

magnetic configuration is de facto defined by the pitch of its magnetic field lines, which is given by $q(r)=rB_\phi/R_0B_\theta$. With the (possible) exception of a finite region around the magnetic axis, the $q(r)$ of standard tokamaks exceeds 1 and increases with r so that the magnetic shear is positive. The maximum $q(r)$ in most standard tokamaks is between 2 and 4. The intrinsic dual periodicity of toroidal configurations means that perturbations also have their own effective pitch. This is a rational number, m/n, where m and n are positive integers, with m being the number of times the perturbation circles the torus in the poloidal direction while completing n turns in the toroidal direction. Thus, it is possible for the field line pitch $q(r)$ to equal (or resonate with) the perturbation pitch m/n at certain radii, $r_{m,n}$, where $q(r_{m,n})=m/n$. These special radii define magnetic surfaces that are called rational or resonant surfaces.

Resonant surfaces, and more generally field line pitch, are crucial to an important class of instabilities called kink-tearing instabilities. The onset of these instabilities is determined by $q(r)$ and by the radial gradient of the current profile. Since $q(r)$ is determined primarily by $B_\theta(r)$, kink-tearing instabilities are also conveniently described as current-driven instabilities. When the perturbation pitch m/n exceeds $q(a)$ for $r<a$, the instability does not involve magnetic reconnection and is called an external kink. It may then be described by ideal MHD. The effects of boundaries, particularly conducting walls, are vitally important to external kink dynamics and stability (see Section 4 and Fig. 9). When the perturbation is resonant [i.e., when $m/n=q(r)$ for $r<a$], reconnection is usually involved and the instability is called a tearing mode since magnetic field lines tear and reconnect. Tearing modes involve the formation of current filaments. In certain cases, the instability dynamics of a resonant kink (usually with $m=1$) can be described, at least in its initial phases, by ideal MHD. Such instabilities are called ideal internal kinks.

Ideal kink instabilities are current driven, but it is really the pitch of the magnetic field lines that signals the onset of the instability. Kink instabilities have long wavelength in the toroidal direction and cause the plasma column to snake helically around its equilibrium position in the manner shown in Fig. 23. Before the column is perturbed, the current flows parallel to the magnetic field. After perturbation, it follows the helical path shown and produces a magnetic field resembling that shown in Fig. 23. This is only part of the total field; to obtain the total field, one must imagine augmenting the **B** shown in the

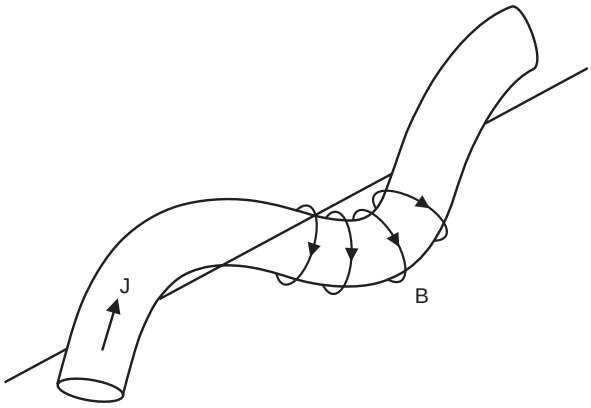


FIGURE 23 Kink instability of a toroidal equilibrium. In the initial state, the current flows along field lines. In the perturbed state (shown here), the current follows a helical path and creates a field, **B**, of the type shown. The remainder of the field (not shown) is along the helically distorted column. The net result is a field that has a component perpendicular to **J** that creates a Lorentz force enhancing the perturbation.

figure by the field along the column. The net result is a field, **B**, that has a component orthogonal to **J**. This creates a $\mathbf{J} \times \mathbf{B}$ force that reenforces the initial perturbation, thus promoting instability.

Kink instabilities are similar to the helical deformations of an elastic tube or band that occur when it is twisted beyond a critical value. An example of such kinks is sometimes seen on a wire connecting a telephone handset to its cradle. The twist of the band is clearly analogous to the initial pitch (or twist) of field lines, as given by $q(r)$. Kink instabilities reflect the tendency of the plasma to lower its energy by assuming a secondary, helical equilibrium that replaces the initial configuration with one of helical symmetry. The ultimate radial extent of the helical deformation is determined by several factors, including the amount of plasma current, the toroidal field strength, and the size of the gap between the plasma boundary and the conducting walls of the containment vessel. Since the plasma current channel expands during kink instabilities, these constitute a simple but very direct route to an undesirable disruption of the discharge, should the current column come into contact with a wall of the containment vessel.

In practice, external kink instabilities can be avoided by not operating the discharge with $q(a)$ close to a low-order rational number, such as, $\frac{3}{2}$, 2, or 3, and/or by having a conducting wall sufficiently close to the plasma column. Kink instabilities are also predicted to occur when $q(r) < 1$. For this reason, internal kinks are ubiquitous in tokamaks and are thought to be related to the sawtooth phenomenon. This refers to quasi-periodic relaxation oscillations of the central temperature of the tokamak that, in diagnostic recordings, produce a jagged trace reminiscent of the teeth on a saw.

Since internal kink perturbations usually do not pierce the plasma boundary or bring the current channel into contact with the vessel wall, they are usually not directly associated with disruptions. Internal kinks do, however, affect the performance of tokamaks by limiting the central plasma pressure that can be achieved.

The dynamics of energy release in resonant kinks or tearing modes are determined by nonideal effects (i.e., by effects that break the Alfvén frozen flux constraint). Typical of such nonideal effects is plasma resistivity η. However, it is important to keep in mind that the free energy source for tearing modes is the same as that for kinks, namely the current gradient. Nonideal effects, such as resistivity, are only the triggers of the instability and are significant only in a narrow diffusion layer, surrounding the resonant surface, in which reconnection occurs. Just as kinks tend to form helical secondary equilibria, so do tearing modes. However, since tearing modes are resonant, reconnection occurs, resulting in the formation of magnetic islands in the region near the resonant surface, as shown in Fig. 24. These islands may be thought of as consequences of current filamentation and concomitant reconnection. Magnetic islands are MHD analogues of Kelvin cat's eyes or nonlinear critical layers, which form as a consequence of certain hydrodynamic shear flow instabilities. When magnetic islands grow to a size at which they touch the wall, or when magnetic islands at neighboring resonant surfaces overlap (resulting in stochastic field lines in a region that pierces the surface of the plasma), MHD turbulence is generated, resulting in turbulent transport, rapid heat loss, and a sudden expansion of the current channel. When the current strikes the wall, a large impurity influx commences, and this raises the effective

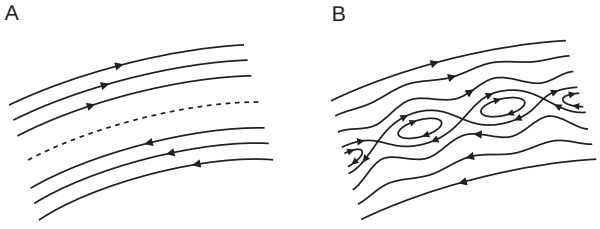


FIGURE 24 The formation of magnetic islands. (A) Initial state; the neutral line is shown dashed. (B) Field configuration after magnetic islands have formed around the neutral line.

plasma resistivity, which in turn quenches the current, creating a disruption. For this reason, tearing mode dynamics and stability are of great interest to MFE.

A final comment on tearing instabilities concerns their rate of growth. Ideal MHD instabilities grow exponentially, on Alfvénic (e.g., kink) or sonic (e.g., ballooning) timescales. Tearing modes initially grow at a rate proportional to $(\tau_A^{-2}\tau_\eta^{-3})^{1/5}$, where τ_A and τ_η are the Alfvén and (global) resistive diffusion times defined in Sections 3 and 4. Since $\tau_A \ll \tau_\eta$, tearing modes grow more slowly than ideal MHD modes. However, linear growth rates are of little relevance to the dynamics of the tearing mode since magnetic islands rapidly enter a nonlinear phase of algebraic growth. This occurs when the width of the magnetic island exceeds the width of the (narrow) reconnection layer of the linear theory. A major current research topic in tokamak MHD concerns the nonlinear growth and saturation of magnetic islands.

5.4 Interplay of MHD Instabilities and Tokamak Operation

Considerations of turbulent transport and MHD stability define the possible viable operating space of present-day tokamaks. The MFE program has sought to exploit understanding of MHD stability limits, etc. in order to expand and optimize the tokamak operating space. The tactics and tricks used include

1. *Shaping the cross section* to increase β limits: The goal here is to design a shape that increases the relative fraction of a magnetic field line that lies in the region of favorable curvature.
2. *Programming current and heating evolution* in order to avoid conditions that are prone to the onset of disruptions (i.e., in order to improve ideal stability limits).
3. *Controlling current profiles* by driving currents locally through the injection of radiofrequency waves.
4. *Using active feedback*, including current drive, to mitigate or retard the growth of undesirable perturbations.
5. *Driving plasma rotation* by neutral beam injection in order to stabilize certain MHD instabilities.

A recurrent theme in fusion research is the often rather diabolical ways in which various instabilities conspire to limit the operating space and tokamak performance. As mentioned previously, eliminating microinstabilities using transport barriers enhances confinement times but opens the door to pressure-driven instabilities (i.e., interchange ballooning modes). Operating at high current is good for energy confinement and helps limit pressure-driven ballooning modes, but it increases the possibility of disruption due to kink-tearing modes. Employing low currents improves stability but often degrades confinement and fails to achieve maximal machine performance. The list is long and continues to grow. These numerous and accumulating questions and concerns ensure that both the theoretical and experimental MHD of tokamaks and other MFE devices will remain vital and vibrant research areas well into the 21st century.

5.5 The Future of MHD in MFE

There is little doubt that in the near future the focus of the world MFE program will be the International Thermonuclear Experimental Reactor (ITER), a tokamak shown schematically in Fig. 25. The ITER program is an international project involving the collaboration of many countries. Its primary mission is to study plasma dynamics under conditions of a fusion burn, in which the main source of plasma heating is from collisional friction between the alpha particles produced by the thermonuclear reactions and the electrons.

A burning plasma is different from the standard electron–ion plasma discussed so far in (at least) two very important respects. First, a burning plasma contains a very substantial population of high-energy alpha particles. This situation is inconsistent with the foundations of MHD, which treats both ions and electrons together as a single fluid. The absolutely minimal viable theory of the dynamics of a burning plasma is a two-component description involving an MHD-like model of the bulk plasma and a kinetic equation for the energetic alphas; these have a velocity distribution that is far from Maxwellian and cannot therefore be adequately modeled by fluid equations. The two components of the model couple to each other electrodynamically.

The second important distinction between a burning plasma and a standard plasma is that high-energy alpha particles can resonate with plasma waves of higher frequency than can bulk particles, thus encouraging a new class of instabilities. Of particular importance among these are toroidal Alfvén eigenmodes (TAEs), which can be destabilized by resonance with alphas and so relax the alpha particle density and temperature gradients. TAEs are

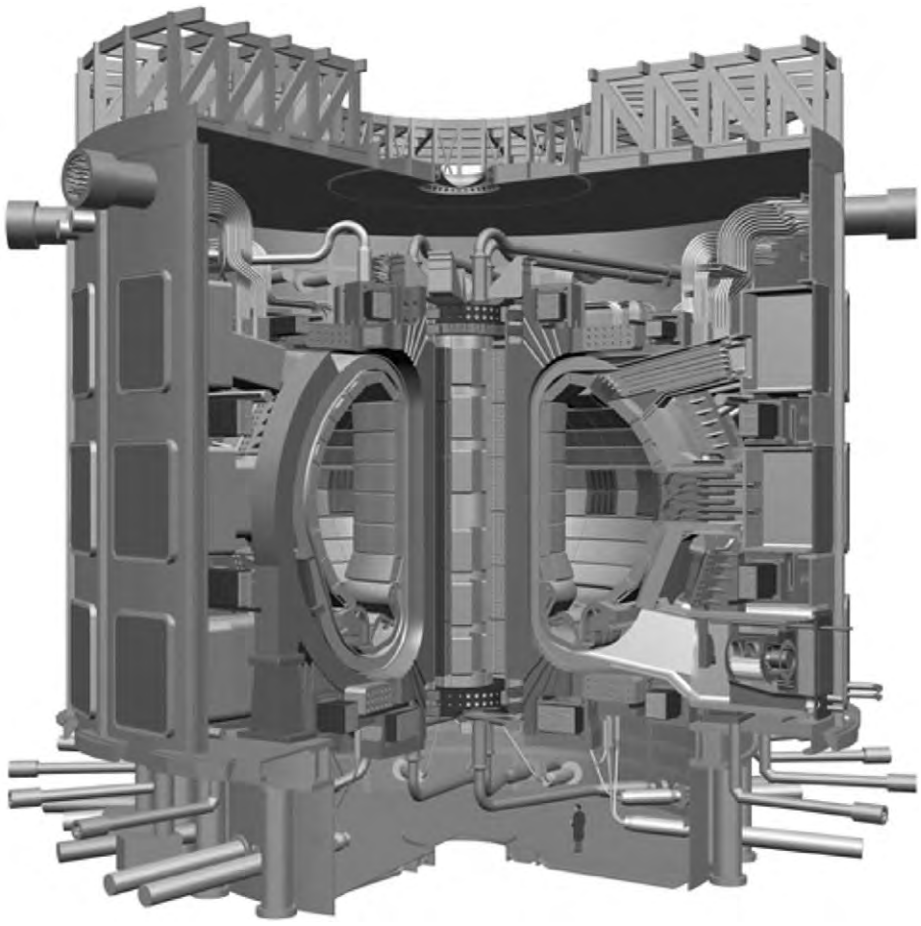


FIGURE 25 The tokamak ITER. To appreciate the scale of the device, see the figure of a person near the bottom.

critical to MFE since excessive instability-induced alpha losses can prevent, or prematurely quench, ignition. TAEs and other energetic particle-driven instabilities have been a topic of great interest in theoretical, computational, and experimental MFE research, with beams of externally injected high-energy particles used to simulate the effects of the alphas. Clearly, research in alpha particle-driven instabilities will be a major part of the ITER program.

Traditional MHD problems will be major foci of the ITER program as well. A key topic will certainly be feedback stabilization of instabilities, with the aim of avoiding or mitigating disruptions. Note that the disruption of a burning plasma could cause very significant damage to the ITER confinement vessel. Also, any successful fusion power plant must be effectively disruption free. Because the ITER plasma will be very hot, with concomitantly infrequent collisions between particles in the plasma, the traditional MHD model will need extension in order to include dynamical effects relevant to the ITER plasma. Such effects include possible nonideal, kinetic modifications to MHD as well as transport coefficients that describe dissipation due to micro-turbulence rather than collisions. Theoretical and computational research on such extended MHD models is being pursued vigorously throughout the world today. In addition, it is anticipated that subgrid scale models of turbulent transport will be an integral component of extended MHD models in the future.

SEE ALSO THE FOLLOWING ARTICLES

Electrical Energy and Power • Electricity Use, History of • Electric Motors • Electromagnetism • Magnetic Levitation

Further Reading

Bateman, G. (1978). "MHD Instabilities." MIT Press, Cambridge, MA.

Biskamp, D. (2003). "Magnetohydrodynamic Turbulence." Cambridge Univ. Press, Cambridge, UK.

Biskamp, D. (1997). "Nonlinear Magnetohydrodynamics." Cambridge Univ. Press, Cambridge, UK.

Chandrasekhar, S. (1961). "Hydrodynamic and Hydromagnetic Stability." Cambridge Univ. Press, Cambridge, UK.

Childress, S., and Gilbert, A. D. (1995). "Stretch, Twist, Fold: The Fast Dynamo." Springer, Heidelberg.

Davidson, P. A. (2001). "An Introduction to Magnetohydrodynamics." Cambridge Univ. Press, Cambridge, UK.

Drazin, P. G., and Reid, W. (1981). "Hydrodynamic Stability." Cambridge Univ. Press, Cambridge, UK.

Freidberg, J. P. (1987). "Ideal Magnetohydrodynamics." Plenum, New York.

Itoh, K., Itoh, S.-I., and Fukuyama, A. (1999). "Transport and Structure Formation in Plasmas." Institute of Physics Press, Bristol, UK.

Kadomtsev, B. B. (1992). "Tokamak Plasma: A Complex Physical System." Institute of Physics Press, Bristol, UK.

Krause, F., and Rädler, K.-H. (1980). "Mean-Field Magnetohydrodynamics and Dynamo Theory." Pergamon, Oxford, UK.

Mestel, L. (1999). "Stellar Magnetism." Oxford Univ. Press, Oxford, UK.

Parker, E. N. (1979). "Cosmic Magnetic Fields." Oxford Univ. Press, Oxford, UK.

Taylor, J. B. (1986). Relaxation and magnetic reconnection in plasmas. *Rev. Modern Phys.* **58**, 741–763.

Tillack, M.S., Morley, N.B. (1999). Magnetohydrodynamics. *In* "14th Standard Handbook for Electrical Engineer" (D. G. Fink and H. W. Beaty, Eds.), pp. 11-109–11-144. McGraw-Hill, New York.

Wesson, J. (1997). "Tokamaks." 2nd Ed. Oxford Univ. Press, Oxford, UK.

White, R. B. (1989). "Theory of Tokamak Plasmas." North-Holland, Amsterdam.

Manufactured Gas, History of

JOEL A. TARR
Carnegie Mellon University
Pittsburgh, Pennsylvania, United States

1. Origins and Characteristics of the Industry
2. The Period of Coal Gas Domination, 1816–1875
3. Innovations and Challenges from Competing Fuels, 1875–1927
4. The Transition to Natural Gas, 1927–1954
5. Environmental Impacts of the Manufactured Gas Industry
6. Conclusions

Glossary

carbureted water gas Gas produced by spraying blue gas (water gas) with liquid hydrocarbons and thermally cracking it to form carbureted water gas and tars.
coal gas A method of heating bituminous coal in a retort to destructively distill volatiles from the coal in order to produce fuel gas for illuminating and heating purposes.
water gas A fuel gas composed primarily of hydrogen and carbon monoxide produced through the interaction of steam with incandescent carbon, usually from anthracite coal or coke.
Welsbach mantle A mantle made of cotton fabric impregnated with rare earth that becomes incandescent when exposed to a gas flame.

This article examines the development of the manufactured gas industry in the United States from its origins early in the 19th century to its demise in the post-World War II period. It discusses the production technologies of coal gas, water gas, and oil gas and their methods of distribution. It also discusses industry innovations, such as the development of carbureted water gas and of the Welsbach mantle, and their effect on regulatory standards. It explores the shift from the use of manufactured gas to natural gas in the period from the 1920s through the 1950s, with the consequent demise of the industry. Finally, the article examines the environmental effects of the industry and the issues of the cleanup of former MGP sites.

1. ORIGINS AND CHARACTERISTICS OF THE INDUSTRY

Manufactured gas was one of the most critical energy sources and fuels that provided American cities, as well as other cities throughout the world, with light and energy during much of the century from 1850 to 1950. Manufactured gas is produced by heating coal or other organic substances such as wood, pitch, or petroleum, driving off a flammable gas made primarily of carbon monoxide and hydrogen as well as other gases. The discovery that flammable gas could be produced in this manner was first made in Great Britain by John Clayton in the late seventeenth century followed by the work of Frenchmen Jean Pierre Minckelers and Philippe Lebon in the late 18th and early 19th centuries. At the turn of the century, the British engineer William Murdock developed methods of gas lighting that were used in cotton mills.

Demonstrations and applications of gas lighting began in the United States in 1802, based on European developments. Gasworks were installed to provide better illumination for cotton mills and lighthouses during the first decade of the century. In 1817, the Baltimore municipality granted the Gas Light Company a franchise to lay gas pipe in the city streets and to light them. Others cities quickly followed: New York, 1825; Boston, 1829; New Orleans, 1835; Philadelphia, the first municipally owned plant, 1836; and San Francisco, the first western plant, 1854. The U.S. Census reported 30 manufactured gas plants in 1850, 221 in 1859, 390 in 1869, 742 in 1889, 877 in 1904, and 1296 in 1909, but the actual number was probably much larger.

The manufactured gas utilities that emerged ranged in size from very small plants in small towns (cf. 5000 population) that served a limited number of consumers (200–300) to very large plants with extensive distribution systems that served thousands

 © 2004 Elsevier Inc. All rights reserved.

of customers with gas for home lighting, cooking, and heating, served municipalities by illuminating city streets, and provided gas for a number of industries. Urban gas utilities were systems that generated fuel for their customers from a central plant (or plants in the case of the very large utilities) and distributed it through a piped system. By far, it was most widely used for lighting purposes. Individual gas production units were also manufactured and were used for isolated factories and homes (spirit gas lamps). From the late 19th century onward, various industries in need of low-Btu (British thermal unit) gas frequently installed coal-using producer gas units in their factories.

Municipal and state governments regarded gas manufacturing plants as public utilities and required that they acquire franchises to use the city streets for their distribution lines. In order to profitably survive, gas manufacturing utilities required long rather than short franchises—the exact term, however, produced many heated political battles. Franchises were originally local but frequently became the concern of state authorities because the natural monopoly characteristics of the utilities were resistant to market forces. Battles over issues such as price, luminosity and energy ratings, and technology, as well as price and duration, frequently took place between utility firms and municipal and state authorities. Although some municipally owned gas works existed, the industry remained largely private.

2. THE PERIOD OF COAL GAS DOMINATION, 1816–1875

In the period from 1816 to 1875, the number of centralized gas works grew to between 300 and 400, primarily using technology developed in Great Britain. The most important innovations during these years concerned fuel, distillation technology, distribution, and metering technology.

Although most firms eventually used coal as their feedstock for gas production, other substances were also initially used. The Baltimore Gas-Light Company, for instance, used pine tar as the feedstock for its initial gas production, the New York Gas Works used rosin, the Brooklyn Gas Works consumed British coals, and the Wilmington, North Carolina Gas Works used wood from the Carolina Pine Barrens. Several small eastern cities used distillate of turpentine to manufacture rosin gas and Philadelphia experimented with wood as late as the 1850s. Utility managers originally avoided coal because of its impurities, but in the period from 1815 to 1820, British gas manufacturers developed methods to remove such impurities. In 1822, Baltimore became the first American city to consume coal to produce gas, using bituminous coal from Virginia mines for its feedstock.

Since bituminous coal mines were not easily accessible for most cities, the cost of the fuel kept prices relatively high or forced the use of organic materials such as wood. Expensive fuel, therefore, provided a reverse salient (a concept originated by historian of technology Thomas Hughes), in terms of extending the size of the market for improved lighting. Whereas bituminous coal, considered "the best in the world for gas light," existed in western Pennsylvania and was used in Pittsburgh for gas production and for home heating, transportation costs over the Allegheny Mountains were extremely high, limiting its use in eastern markets.

The major transportation breakthroughs came in the late 1840s and 1850s with the slack watering of the Pittsburgh area rivers and the completion of railroad connections from Philadelphia to Pittsburgh in 1852. These transport improvements facilitated the shipping of high-quality bituminous gas coal from western Pennsylvania to gas works in eastern cities at greatly reduced transportation costs. By the late 1850s, for instance, the Westmoreland Coal Company, a large western Pennsylvania producer of gas coal, was supplying 58 gas works.

Responding to the lower fuel costs and the anticipation of a larger market for cheaper light, a number of gas works were founded in the 1850s and 1860s. Between 1850 and 1860, the U.S. Census reported that the number of manufactured gas establishments jumped from 30 to 221, a rise of over 600%—an increase much greater than that of urban population growth. By 1870, the total number of manufactured gas plants reported was 390, another large jump.

The manufacture of coal gas consisted of three operations: distillation, condensation, and purification. Although there were many different processes patented for the production of coal gas, by the 1850s those most widely used operated by heating bituminous coal in a retort to destructively distill volatiles from the coal. The leading types of coal-carbonization retorts were the horizontal, inclined, and vertical. Horizontal retorts were most commonly used; until approximately 1850, they were made of cast iron, when superior clay refractories were substituted. A typical horizontal bench consisted of six

retorts and a producer gas furnace that used coke or coal for heating the retorts. Coal tar condensed from the gas was also used as a retort fuel. Wet and dry lime processes were used in purification methods, as was iron oxide, beginning in 1870.

Proper modes of metering, gas storage, and distribution were critical to the successful operation of a gas firm. Gas holders, or gasometers, as they were called, were essentially inverted iron bells in a tub of water. The early gas holders were usually made of masonry and often set in the ground, with iron plates for the bell. A number of firms initially used wooden mains for distribution, but in ~1820, cast iron pipes manufactured in British foundries were introduced into the United States; however, some wooden mains of tamarack logs lasted into the late 19th century. By the 1830s, American foundries were producing their own cast iron pipe, reducing dependence on the British. From ~1820 through 1850, pipe was made in horizontally cast sections (4–5 ft in length). These were subject to considerable core sagging and slag accumulations but in 1850 foundries shifted to casting in vertical molds, producing a much stronger and longer (12 ft) pipe product. Nineteenth century cast iron gas mains often lasted 50 to 75 years.

The distribution system was another major cost for the gas companies. Firms occasionally used smaller diameter mains than was ideal, causing problems and poor service as the system expanded. In Philadelphia, for instance, the city-owned utility used pumps to force gas through mains too small to accommodate the demand, often resulting in customers receiving gas under excess pressure. Though the gas utilities seldom extended conduits to undeveloped streets, they quickly installed pipe on streets in growing areas in order to benefit from rises in real estate prices.

Cities initially used gas primarily for street illumination but cost reductions plus innovations such as improved burners and fixtures and metering led to a broader market. Gas companies originally charged flat rates per burner, leading to expansion limits for street lights and domestic sales. Meters were essential for price reduction for both municipalities and domestic consumers but were fought in many city councils. The original meters were "wet" meters, using water, but in the 1850s firms began replacing them with more accurate dry meters.

The manufacture of coal gas provided the potential for firms to recover and sell by-products such as coke, tar, light oil, and ammonia. However, unlike Great Britain, there was no market for such products in the United States (aside from the coke) before the late 19th century. The United States lacked a coal-based chemical industry and was almost completely dependent on imports for dyestuffs, drugs, and explosives and other products of coal tar and light oil origin. Ammonia was regarded as a nuisance until after the Civil War, when ammonium sulfate was used for fertilizer. Gasworks themselves burned by-product coke under their retorts and sold some on the open market, but it had limited appeal.

By the 1870s, over 400 manufactured gas utilities existed and the industry had spread throughout the country. Over half of the firms, with over 50% of the labor and capital, were located in just four states: New York (71), Massachusetts (52), Pennsylvania (43), and Ohio (32). Price, however, still limited home gas service to a relatively affluent segment of the market. Gas companies had originally hoped to make large profits from street lights, but municipal franchises applied stringent regulations as to the numbers of public street lights and rates paid, and many firms defaulted on this part of the contract or were continually petitioning city councils for rate increases. Because costs were high and margins low, gas companies were reluctant to expand to lower income markets, putting the industry at risk from lower priced fuels. In the 1850s, coal oil lamps provided much of the lighting needs of those who could not afford gas, the fuel supplied from refineries concentrated in Pittsburgh and Ohio. These lamps furnished up to 14 cp and were quite flexible to use. In the post-Civil War period, however, two new competitors for manufactured gas emerged to challenge the industry for the lighting market—kerosene and electricity.

3. INNOVATIONS AND CHALLENGES FROM COMPETING FUELS, 1875–1927

Substantial competition for urban lighting markets developed in the half century from 1875 to 1927. Municipalities and states also subjected the industry to increased regulation. The first competitor for the lighting market was the kerosene lamp, an outcome of the discovery of oil in Pennsylvania in 1859, which largely captured the low end of the illumination market. Improvements in petroleum refining and transportation techniques and the discovery of new sources of supply resulted in reduced oil prices. By the 1870s and 1880s, most of the urban working

class, as well as a substantial part of the middle class, used kerosene lamps for domestic illumination.

Electric arc lamps began competing with manufactured gas for the street lighting market in the 1880s. Since gas companies found these lighting contracts unprofitable, it was a limited loss. By the 1880s, domestic lighting accounted for 90% of gas revenues and arc lamps posed no threat in this market because their intensity made them unusable in the home. In 1881, however, Thomas Edison, the inventor of the incandescent lamp, developed a central power station in New York City that supplied power for home lighting that directly competed with interior gas light. Edison had wisely based his distribution system on that of the gas industry, constructing an "electrical analogue to the gas system."

The manufactured gas industry responded to this challenge with the development of carbureted water gas. The concept behind water gas, which was first observed in 1780 and patented in France in 1834, involved the action of steam on incandescent carbon, usually anthracite coal or coke. When the steam, composed of oxygen and hydrogen, came into contact with the incandescent carbon, it produced a fuel gas composed of hydrogen and carbon monoxide, as well as other products. The gas had a heating value of approximately 300 Btu/ft^3, burned with a blue flame, and had little illuminating power. The last characteristic diminished the utility of water gas for commercial use, because the market for manufactured gas in most of the 19th century was almost entirely in lighting.

In 1875, however, Thaddeous S. C. Lowe, a Pennsylvania inventor, patented a process to enrich the blue gas by spraying it with liquid hydrocarbons and thermally cracking the gas to form carbureted water gas and tars. The Lowe process was acquired by the United Gas Improvement Company (UGI) of Philadelphia, formed in 1882. Before this time, utilities had been slow to adopt the water gas process and the states of Massachusetts and New Jersey actually outlawed water gas because of gas poisoning incidents. UGI, however, aggressively pushed the technology and the rapid growth of the petroleum industry made oil available at a relatively cheap price. The gas industry proceeded to shift a substantial amount of its manufacturing processes from the production of coal gas to carbureted water gas. Originally carbureted water gas manufacturers used naphtha as the process oil because it was relatively inexpensive, but after 1895, as rising demand caused the price of naphtha to rise, gas manufacturers switched to the petroleum fraction known as gas oil. The invention and adoption of the automobile increased the demand for the higher petroleum fractions even further and in ~1930 the carbureted water gas industry switched to a yet cruder grade of petroleum.

The use of oil rather than coal as the feedstock for gas production was another major step by the manufactured gas industry. Oil gas was made by the thermal decomposition or cracking of petroleum. Whale oil had been used in several gas plants early in the 19th century, but it was not until 1889 that a modern refractory process to use petroleum was invented. L. P. Lowe, the son of the discoverer of carbureted water gas, was the inventor. The first oil gas plant was constructed in California and in 1902 an oil gas plant was erected in Oakland, California for lighting purposes. The so-called Pacific Coast oil gas process was used mostly in Pacific coast cities—an area that was poor in coal but rich in petroleum.

The number of carbureted water gas plants in the United States grew steadily. In 1890, the census reported returns from 742 gas manufacturing firms, probably approximately 75% of the total. In this census, coal gas represented 46.7% of the total product, whereas water gas made up 38% of the total. Between 1890 and 1904, coal gas production dropped precipitously to 11.2% of total gas produced, straight water gas rose to 48.6%, and a new category, mixed coal and water gas, absorbed 36.4% of the market. Between 1904 and 1914, coal gas declined even further, from 11.2 to 5.2% of total gas production; carbureted water gas also declined from 48.6 to 44.2% and mixed gas advanced from 36.4 to 42.4%. Because Great Britain (the first water gas plant was built in 1890) and other European nations did not have access to inexpensive petroleum, they did not adopt carbureted water gas to the same extent as did the United States. The U.S. gas industry, therefore, became increasingly dependent on oil, whereas the European industry remained largely based on coal.

Carbureted water gas had a number of advantages over coal gas. Its production required less space, less capital, and less labor, as well as a smaller storage capacity, because plants could be started up in a relatively short time. Many coal gas plants installed auxiliary carbureted water gas machines to increase their flexibility in regard to load, mixing the two gases. Without the auxiliary water gas equipment, firms would have had to add additional storage or construct additional retorts, which would have been idle for many hours in the day. Carbureted water gas

also had a higher illuminating value and a flexible heating value; a plant could produce gas varying in heating value from 300 to 800 Btu/ft^3. Utility operators could produce carbureted water gas with widely varying Btu contents by changing the amount of oil cracked into the blue gas. Higher Btu carbureted water gas could be mixed with gas with a lower Btu grade from blue gas or coal gas retorts, thereby increasing the available gas volume without the necessity of increasing the number of water gas sets.

The addition of petroleum hydrocarbons to water gas provided a gas that had illumination values equal to those of coal gas but at a lower cost. Most municipal franchises to gas companies required illuminating power of between 15 and 17 cp. The manufactured gas burners most commonly used before 1886 were the bat-wing, the fish-tail, and the Argand burners. These burners depended on the luminosity of the flame in order to fulfill legal candlepower requirements. Therefore, when carbureted water gas was burned as a fuel, the burners produced an equal or better flame at a lower cost. The lamps, however, still depended on the incandescence of the carbon particles in the flame itself for their illuminating power.

In 1885, Carl Auer von Welsbach of Germany, working in the laboratories of Robert Wilhelm von Bunsen, produced another major innovation in the consumption of manufactured gas. Welsbach patented a gas mantle made of cotton fabric impregnated with rare earth. When the mantle was placed near a gas flame from a Bunsen burner, it "became incandescent," producing a steady white light of 30–50 cp. This increased the light that could be obtained from a given quantity of gas approximately six times, utilizing the principle of incandescence rather than the direct illuminating properties of the gas. Thus, it mirrored the innovation of the gas industry's chief competitor, the incandescent electrical lamp. The Welsbach mantle provided an inexpensive and superior incandescent lamp to meet the competition of the electric light and helped modernize an industry whose existence was under severe competitive threat.

Although carbureted water gas was necessary for the older burners in order to meet existing legal candlepower and calorific requirements, it did not have to be used with the Welsbach mantle, which produced an equally brilliant light with cheaper blue water gas. Yet, legal candlepower requirements for manufactured gas required that higher priced carbureted water gas rather than blue gas had to be used with the Welsbach mantle. At the time of the mantle's introduction, municipal franchises usually stipulated a candlepower standard for gas ranging from approximately 14 to 16 for coal gas and 18 to 36 for water gas. The illuminating power of the Welsbach Mantle, however, which was not dependent on a higher candlepower gas rating, made these standards obsolete.

Statutory change to accommodate the new technology occurred slowly and some utilities actually opposed shifting to a different standard because they feared a possible profit decline. Beginning with the Wisconsin Public Utilities Commission in 1908, however, states and municipalities began requiring that gas be measured in Btu or heat units per cubic foot of gas (as was done in Great Britain) rather than lighting units. Of the 1284 gas firms reporting in the 1914 Manufacturing Census, 1055 reported candlepower standards and 811 reported standards expressed in Btu heating values. By 1919, there had been a substantial change. Of 1022 establishments reporting, only 389 reported candlepower standards, whereas 737 reported Btu heating value standards.

The use of gas solely for lighting resulted in poor load factors for gasworks. Although they needed generating, storage, and distribution equipment capable of handling demand during peak lighting periods, their capital equipment was often utilized for only 10–12 h per day. The use of larger gasholders (gasometers) was one way utilities could reduce investment in gas-producing equipment. Smaller plants frequently ran their equipment during part of the day and relied on their gasholders to supply gas at night. Larger companies could not risk supply cutoffs and had to run separate units of gas producers, even though they too used gas holders for reserve. From the 1870s through the 1890s, several important improvements were made in gasholders, including the use of iron and steel tanks and the installation of the first three-lift holder tank. Because of limited capacity, however, gas holders were not a satisfactory solution to the load factor.

Competition from electricity, despite the success of the Welsbach mantle, consistently ate into the gas industry's market share. Utilities could attempt to spread their loads as a way to meet competition and to use their technology more efficiently. The gas industry moved increasingly into markets outside of lighting, such as cooking stoves, water heating, and house heating, although firm aggressiveness varied from place to place. In both Baltimore and Philadelphia, for instance, both large markets, the gas companies only gradually moved away from lighting as their only product. In other cities, however, such as Denver and Los Angeles, utilities introduced gas-consuming

appliances more rapidly. Two technical innovations aided in the expansion of these markets: the Bunsen burner (1885), which premixed gas with air, and thermostatic controls, initially applied to water heating in 1899. This movement away from light and into household appliances reflected the recognition that manufactured gas was a fuel or energy source. It was also a "technical mimicry" of electricity and an attempt by an older technology to survive by copying a more recent technology, in this case electricity.

Even though the gas industry developed a substantial market share in the home utility domain, it still faced excess productive capacity. In ~1900, gas firms began to try to capture industrial customers to absorb this capacity. Efforts were made to market gas engines as well as to sell gas for heating purposes. Potential industrial customers, however, often found it more economical to use low-Btu gas to meet their energy requirements rather than high-Btu carbureted water gas. Frequently they used producer gas, a low-grade but inexpensive gas made on the firm's site from coal. Gas companies also found it difficult to enter industrial markets because of state requirements for a fixed number of Btu per cubic foot (usually 450–550 Btu) in their gas, which raised gas prices. This Btu rating was higher than many industries desired or would pay for, limiting the number of industrial customers. The gas industry, therefore, began to push for the sale of gas with Btu as a standard and sold by the therm (100,000 Btu), as was done in Great Britain, rather than by the cubic foot. Industries could then obtain less costly low-Btu gas to supply the heat they needed in their processes. Changes in this regard were reported in the 1920 Census of Manufacturers. The census noted that the prevailing Btu standard was lower in 1919 than it had been in 1914, as the number of gas firms providing high-Btu gas dropped sharply. Between 1919 and 1927, the industrial use of manufactured gas increased from 70.4 billion to 136.4 billion ft^3.

Gas utilities could also compete with electricity by investing in electrical generating equipment or by acquiring electrical companies. In 1887, approximately 40 gas companies supplied electric lighting, usually arc lights; by 1889, the number was 266, or approximately 25% of U.S. gas companies, and in 1899, 362 or nearly 40% of gas companies provided electric lighting. A counter trend was for electrical utilities to acquire gas utilities, and large electric firms or holding companies made acquisitions of small gas companies increasingly in the 20th century.

Water gas possessed another disadvantage compared to coal gas in terms of potential profits—the character of its by-products. Markets had developed in the late 19th century for coal gas by-products such as tar for creosote in railway tie preservation, pitch for roofing and waterproofing, and refined tar for road surfacing. The carbureting process produced gaseous, liquid, and solid by-products, but they were not as salable as those from coal gas. The liquid was "water-gas tar," which was lighter and contained less valuable constituents than coal tar. The solid was coke, the sale of which could substantially affect the profit position of plants. Water gas coke, however, was lower in thermal efficiency than coal gas coke. In addition, water gas firms, rather than entering the coke markets, consumed the coke they produced as water gas fuel. Finally, unlike coal gas, the water gas process did not produce ammonia, cyanides, or phenolic compounds as by-products. The ammonia was especially valuable and could constitute 6–10% of the cost of the coal used. Before the invention of the Haber process for the synthetic production of ammonia, coal carbonization was the main source of fixed nitrogen, and coal carbonization processes found increasing favor in the 1920s among some gas producers.

Between 1914 and 1929, the amount of manufactured gas distributed expanded from approximately 204 billion to over 450 billion ft^3. A little more than 50% of the gas produced in 1929 was water gas, approximately 40% was classified as coal gas, although the sum also included gas from by-product ovens, and oil gas was 6.2%. The most significant new addition to the census categories was gas from by-product coke operations.

Between approximately 1916 and 1929, a revolution occurred in the technology used by the iron and steel industry to make coke. From 1850 through 1910, most of the coke manufactured for use in blast furnaces was carbonized in beehive coke ovens. These ovens produced a high-quality coke from bituminous coal, but wasted all of the coal by-products. The by-product coke oven had originally been developed in Europe, and beginning in the 1890s several by-product coke plants were constructed in the United States. The real expansion of the United States industry, however, occurred just before World War I and in the following decade. The by-product coke oven captured the valuable coal chemicals for use and also produced a gas surplus. That is, by-product ovens could be heated by 35 to 40% of the coal gas produced in the oven, leaving approximately 60% to be marketed. Coke oven gas had a heating value of approximately 560 Btu and could be easily sold to other customers. By 1932, by-product coke firms

produced 25% of the manufactured gas distributed by utilities to domestic, commercial, and industrial customers in the United States.

During the 1920s, the manufactured gas utilities themselves moved into the by-product industry. In order to meet peak period requirements, utilities with carbureted water gas technology had to manufacture large amounts of coke. By utilizing by-product ovens rather than retorts, utilities could combine gas manufacture with coke manufacture to meet peak load requirements and to produce a higher quality of coke for sale than that produced in retorts. During the 1920s, construction firms such as Koppers Company and Semet-Solvay built coke oven batteries for a number of large utilities, including Brooklyn Union Gas, Consolidated Gas Company of New York, New Jersey Public Service Gas & Electric, Rochester Gas & Electric Company, and People's Gas & Coke Company of Chicago. In addition, entrepreneurs built a number of merchant plants that had as their chief products foundry coke, domestic coke, and coke for making water gas, as well as producing gas to sell under contract to public utilities. These plants had the advantage of flexibility in meeting demand peaks for either gas or for coke as well as lowering the cost of gas generation. By the end of the decade, a clear trend existed for manufactured gas utilities to adopt carbonization equipment that produced both high-quality gas and high-quality coke, and by 1932, 18.7% of the manufactured gas produced by gas utilities was coke oven gas. Thus, as a 1926 text on fuels noted, "the gas and the coke industry gradually are merging as far as equipment requirements are concerned, and…the tendency is toward the same type of plant for both branches."

4. THE TRANSITION TO NATURAL GAS, 1927–1954

Natural gas is a substance that occurs in nature under three conditions: in porous domes that form a gas field (nonassociated gas), in association with petroleum deposits (casinghead or associated gas), and mixed with the oil so that it must be trapped out of it. Chemically, there is no difference between natural gas and manufactured gas—each is composed of hydrocarbons—but they differ in heating value. Manufactured gas usually averages between 550 and 600 Btu, whereas natural gas is approximately 1000 Btu. Natural gas is not distributed evenly throughout the environment. In the late 19th and early 20th centuries, there were four regions where natural gas was piped in for use in cities and industries: western Pennsylvania and West Virginia, northern and central Indiana, locations around Los Angeles including the San Joaquin Valley, and eastern Kansas. Towns and cities close to these fields utilized the natural gas until it was exhausted.

In the first decades of the 20th century, additional natural gas discoveries were made in the Southwest, but gas from these fields was initially wasted because of the absence of nearby markets and because the technology to produce pipelines long enough to reach markets was nonexistent. In the 1920s, however, significant breakthroughs occurred in the making and laying of welded pipe technology. In addition, the mathematical understanding of the role of tensile steel strength, pipe thickness and diameter, pipeline capacity, gas pressure, and compressor station spacing to pipeline capacity greatly increased. Pipeline construction passed beyond the craft stage into the stage where decisions were made by mathematical analysis.

From 1925 to 1935, pipeline companies and utilities constructed a number of pipelines ranging from 200 to over 1000 miles in length, connecting southwestern and California gas fields to urban markets. Cities such as Houston and Beaumont (1925), Wichita (1927), Denver (1928), St. Louis (1928), Atlanta (1929), San Francisco (1929), Minneapolis (1930), Chicago (1931) and Washington, DC (1931) were supplied with natural gas, driving out manufactured gas. The depression and the war slowed pipeline construction, but in 1944 the Tennessee Gas Transmission Company completed a 1265-mile pipeline from the Texas Gulf Coast to West Virginia as part of a wartime push to replace diminished Appalachian supplies.

The last major regions to be linked by pipeline with natural gas fields were the Middle Atlantic States and New England. New York City and Philadelphia were considered especially prized markets, since New York City alone consumed 40% of the nation's manufactured gas. These cities were supplied with natural gas in 1947 through the conversion of Big Inch and Little Inch, wartime government-constructed petroleum pipelines, to natural gas and the sale of these pipelines to private enterprise. New England was the last major region to receive natural gas, with Boston being the most distant urban market to draw from the southwestern gas fields. The postwar period also saw a boom in the construction of pipelines serving California and Florida, as well as new major pipelines to the midwest.

In these new markets for natural gas, the logistics of conversion of manufactured gas distribution systems into natural gas distribution systems was a major undertaking. Manufactured gas utilities did not always view conversion with favor, even though there seemed to be many advantages to the change. Some utility managers worried that natural gas supplies would prove insufficient to maintain their service commitments. Corporations were also reluctant to abandon their substantial investments in gas manufacturing plants and equipment.

In the conversion to natural gas as a fuel, the gas could be circulated in the same distribution system and consumed by the same appliances used by manufactured gas. However, because of the different Btu ratings of the gases, conversion required gas main and appliance adjustments. In addition, engineers and managers had to decide whether to use natural gas or natural gas mixed with manufactured gas after conversion. Occasionally regulatory commissions overturned the decisions made by utility managers.

The 1946 Federal Natural Gas Investigating Committee recommended that natural gas be used primarily for enrichment purposes. This strategy would expand the production capacity of existing plants, provide backup in case of supply interruptions, and furnish peak load shaving abilities. A number of firms followed the recommendation for mixed gas. In 1945, for instance, 301.4 million therms of manufactured gas and 988.5 millions therms of natural gas were used to produce mixed gas; by 1954, this total had reached 936 million therms of manufactured gas and 2012.7 million therms of natural gas. Other utilities, especially after World War II, went directly to natural gas, usually to avoid the costly necessity of adjusting gas appliances twice. In many cases, utilities kept their gas-making equipment for emergencies and for peak shaving purposes.

Once the decision concerning the type of gas to be used had been made, utilities faced major organizational tasks of conversion and adjustment, regardless of whether mixed or natural gas had been chosen. In 1931, an American Gas Association Subcommittee issued a report on "The Study of Factors Involved in Change-over from Manufactured to Natural Gas." The committee made four major recommendations: (1) that cities be divided into districts for conversion purposes and that each district be converted to natural gas separately; (2) that those handling the conversion task be thoroughly trained and that the work force be large enough so that the largest district could be covered in 1 week; (3) that customers be kept fully informed at all times of the conversion process; and (4) that a universal adapter be used to adjust orifices. The committee also warned of the drying-out effects of the natural gas on joints and valves in a system that had formerly only carried wet manufactured gas, with resultant leaks. The committee concluded, however, that compared to manufactured gas, natural gas presented fewer hazards. In approaching the conversion process, utilities communicated extensively with one another regarding the pitfalls to be avoided.

The largest single task facing a utility converting to natural gas was adjusting domestic and commercial appliances to the new gas. All gas-using appliances had to be adjusted at the orifice because of natural gas' higher Btu rating and its slower rate of flame propagation compared to 550–600 Btu for manufactured gas. Management of conversion could involve thousands of specially trained workers. In order to carry out the conversion tasks, the larger utilities hired engineering management firms. Some utilities, however, managed their own transitions rather than hiring outside contractors, forming schools and training men specifically for the job and relying on experience from other utilities.

With the transformation in the gas industry, the fuel flow now originated from a distant source rather than traveling a relatively short distance from gasworks to customer. From a regulatory point of view, this shifted responsibility to the federal government because pipelines often crossed state boundaries. Pipeline companies became subject, after the Natural Gas Act of 1938, to federal control, which had a constraining and shaping effect on the industry.

5. ENVIRONMENTAL IMPACTS OF THE MANUFACTURED GAS INDUSTRY

The various by-products produced by the manufactured gas industry, if not captured and disposed of in a safe manner or sold, could become environmental externalities, damaging surface and groundwater quality, poisoning soils with toxic substances, and producing odors and smoke. British gas engineers were aware during the industry's early years of the potential for environmental pollution. In London, for instance, in 1821, effluents from gasworks killed fish and eels in the Thames. In 1841, in his *Practical*

Treatise on the Manufacture of Gas, published in London, engineer Thomas Peckston warned of the possible escape of "tar and ammonical liquor" from the "tar cistern" because it could "percolate through the ground" and render water in springs and wells "unfit for use."

Americans also became aware of the injurious environmental effects of manufactured gas effluents on the environment as the industry expanded. New York State, for instance, passed a statute in 1845 restricting the disposal of gas house wastes in streams in the New York City area; Pittsburgh, Pennsylvania passed a similar statute in 1869 and New Jersey approved legislation limiting the disposal of gas house wastes in streams in 1903 and 1911. Plaintiffs filed numerous nuisance suits against manufactured gas companies in the late 19th and early 20th centuries for offenses such as odors, pollution of wells, damage to trees, and the killing of finfish and shellfish. Various city councils and state bodies undertook investigations of the environmental effects of manufactured gas wastes as well as their pricing policies beginning in the middle of the 19th century. Warnings about the pollution potential of gas house wastes appeared with increasing frequency in the professional literature during the first decades of the 20th century.

Because of the potential costs of nuisance suits as well as the damage to the industry's reputation, in 1919 the American Gas Association established a committee on the "Disposal of Waste from Gas Plants." The 1920 report of the committee noted that gas plant wastes could drive "away fish and damage oyster beds," create "damage to paint on pleasure boats," cause "objectionable odors," result in "pollution of wells [by] contaminating ground water," create "deposits in sewerage systems," and cause "pollution of drinking water supplies where water is chlorinated." The committee provided numerous recommendations to help manufactured gas plant operators to deal with pollution disposal questions and even volunteered to visit sites with problems. The committee continued to meet throughout the 1920s and to explore these issues.

The gradual phasing out of the manufactured gas industry in the 1950s and 1960s because of the conversion to natural gas resulted in the elimination of many plants and a gradual loss of institutional memory concerning their operations. The renewed interest in land and groundwater pollution that developed after Love Canal in 1978 and the subsequent passage of the Superfund Acts, however, focused renewed attention on the sites of former manufactured gas plants (MGPs), many of which had been stripped of gas-making technology and structures. Investigation of these sites by state and federal environmental officials identified a number of toxic by-products remaining in the soil and in groundwater from past MGP operations. Some of the MGPs had operated for a century or so and most sites contained accumulated wastes from the gas-making processes. Thus, an industry largely of the past that had operated approximately 2000 plants in towns and cities throughout the nation and provided it with fuel for light and heat for over a century was discovered to have left a heavy environmental burden for the future.

6. CONCLUSIONS

This article has focused on the evolution of the manufactured gas industry in the United States and the transition to natural gas as well as its environmental legacy. Manufactured gas systems were urban public utilities that, because they utilized city streets for their distribution systems, required first municipal and later state approval for their operations and rates.

Manufactured gas systems shared a number of characteristics with other urban networks, such as water supply, sewerage, and district heating. These were "grid-based systems" that required a special physical network devoted to supplying users. They were also centralized rather than decentralized and had natural monopoly characteristics, a factor that affected their rate-making and consumer relations and, at times, resulted in their becoming deeply involved in politics.

The substitution of natural gas for manufactured gas resulted in the gradual death of the manufactured gas industry as an entity that produced a product rather than merely distributing it. However, much of the original industry's grid remained in the form of inner-city distribution systems. The substitution of natural gas for manufactured gas did not require network reconfiguration but only adjustments because the delivery system and appliances from the first system were retained for use with the natural fuel.

The industry had large environmental effects, both during its period of operation and after, because of wastes discarded during its processing of coal and oil for purposes of producing a fuel for illumination and energy. As an industry it furnished both illumination and energy to American cities for many decades, but its significance today is measured primarily in terms of the environmental burdens it has left to contemporary society.

SEE ALSO THE FOLLOWING ARTICLES

Coal, Chemical and Physical Properties • Coal, Fuel and Non-Fuel Uses • Coal Industry, History of • Electricity Use, History of • Hydrogen, History of • Hydropower, History and Technology of • Natural Gas, History of • Nuclear Power, History of • Oil Industry, History of • Solar Energy, History of • Transitions in Energy Use

Further Reading

American Gas Association. (1956). "Historical Statistics of the Gas Industry." American Gas Association, New York.

Anderson, N., Jr., and DeLawyer, M. W. (1995). "Chemicals, Metals and Men: Gas, Chemicals and Coke: A Bird's-Eye View of the Materials That Make the World Go Around." Vantage Press, New York.

Binder, F. M. (1974). "Coal Age Empire: Pennsylvania Coal and Its Utilization to 1860." Pennsylvania Historical and Museum Commission, Harrisburg, PA.

Blake, A. E. (1922). Water gas. *In* "American Fuels" (R. F. Bacon and W. A. Hamor, Eds.), Vol. 2, pp. 995–1015. McGraw-Hill, New York.

Block, E. B. (1966). "Above the Civil War: The Story of Thaddeus Lowe, Balloonist, Inventor, Railway Builder." Howell-Nort Books, Berkeley, CA.

Castaneda, C. J. (1993). "Regulated Enterprise—Natural Gas Pipelines and Northeastern Markets, 1938–1954." Ohio State University Press, Columbus, OH.

Castaneda, C. J. (1999). "Invisible Fuel: Manufactured and Natural Gas in America, 1800–2000." Twayne Publishers, New York.

Castaneda, C. J., and Pratt, J. A. (1993). "From Texas to the East: A Strategic History of Texas Eastern Corporation." Texas A&M University Press, College Station, TX.

Collins, F. L. (1934). "Consolidated Gas Company of New York." The Consolidated Gas Company of New York, New York.

Davis, R. E. (1935). Natural gas pipe line development during past ten years. *Natural Gas* **16,** 3–4.

Eavenson, H. N. (1942). "The First Century and a Quarter of American Coal Industry." Privately Printed, Pittsburgh, PA.

Elton, A. (1958). Gas for light and heat. *In* "A History of Technology." (C. Singer, *et al.*, Eds.), Vol. IV, pp. 258–276. Oxford University Press, New York.

Harkins, S. M., *et al.* (1984). "U. S. Production of Manufactured Gases: Assessment of Past Disposal Practices." Environmental Protection Agency, Research Triangle Park, NC.

Hasiam, R. T., and Russell, R. P. (1926). "Fuels and Their Combustion." McGraw-Hill, New York.

Hughes, T. P. (1983). "Networks of Power: Electrification in Western Society, 1880–1930." Johns Hopkins University Press, Baltimore, MD.

Hunt, A. L. (1900). Gas, manufactured. *In* "U. S. Bureau of the Census. Twelfth Census of Manufacturers (1900), Part IV: Reports on Selected Industries," p. 19. U.S. Government Printing Office, Washington, DC.

Hyldtoft, O. (1995). Making gas: The establishment of the Nordic gas systems, 1800–1870. *In* "Nordic Energy Systems, Historical Perspectives and Current Issues" (A. Kaijser and M. Hedin, Eds.), pp. 75–100. Science History Publications, Canton, OH.

Jacobson, C. D., and Tarr, J. A. (1998). No single path: Ownership and financing of infrastructure in the 19th and 20th centuries. *In* "Infrastructure Delivery: Private Initiative and the Public Good" (A. Mody, Ed.), pp. 1–36. The World Bank, Washington, DC.

Kaijser, A. (1990). City lights: The establishment of the first Swedish gasworks. *FLUX* **1,** 77–84.

Kaijser, A. (1993). Fighting for lighting and cooking: Competing energy systems in Sweden, 1880–1960. *In* "Technological Competitiveness: Contemporary and Historical Perspectives on the Electrical, Electronics, and Computer Industries" (W. Aspray, Ed.), pp. 195–207. IEEE Press, New York.

Kaijser, A. (1999). Striking bonanza: The establishment of a natural gas regime in the Netherlands. *In* "The Governance of Large Technical Systems" (Olivier Coutard, Ed.), pp. 38–57. Routledge, London, UK.

King, T. (1950). "Consolidated of Baltimore 1816–1950." The Company, Baltimore, MD.

Leinroth, J. P. (1928). Industrial gas in the United States—Growth and future trends. *In* "Transactions of the Fuel Conference, Vol. II, The Carbonisation Industry Utilization of Fuels," pp. 1211–1217. Percy Lund, Humphries & Co., London, UK.

Matthews, D. (1987). The technical transformation of the late nineteenth-century gas industry. *J. Econ. History* **47,** 970–980.

Passer, H. C. (1953). "The Electrical Manufacturers 1875–1900." Harvard University Press, Cambridge, MA.

Passer, H. C. (1967). The electric light and the gas light: Innovation and continuity in economic history. *In* "Explorations in Enterprise" (H. G. J. Aitken, Ed.). Harvard University Press, Cambridge, MA.

Platt, H. L. (1991). "The Electric City: Energy and the Growth of the Chicago Area, 1880–1930." University of Chicago Press, Chicago, IL.

Rose, M. H. (1995). "Cities of Light and Heat: Domesticating Gas and Electricity in Urban America." Pennsylvania State University Press, State College, PA.

Sanders, M. E. (1981). "The Regulation of Natural Gas Policy and Politics, 1938–1978." Temple University Press, Philadelphia, PA.

Schivelbusch, W. (1988). "Disenchanted Night: The Industrialization of Light in the Nineteenth Century." University of California Press, Berkeley, CA.

Stotz, L., and Jamieson, A. (1938). "History of the Gas Industry." Stettiner Bros, New York.

Tarr, J. A. (1998). Transforming an energy system: The evolution of the manufactured gas industry and the transition to natural gas in the United States (1807–1954). *In* "The Governance of Large Technical Systems" (O. Coutard, Ed.), pp. 19–37. Routledge, London, UK.

Tarr, J. A., and Dupuy, G. (eds.). (1988). "Technology and the Rise of the Networked City in Europe and America." Temple University Press, Philadelphia, PA.

Thoenen, E. D. (1964). "History of the Oil and Gas Industry in West Virginia." Education Foundation, Charleston, WV.

Thorsheim, P. (2002). The paradox of smokeless fuels: Gas, coke and the environment in Britain, 1813–1949. *Environ History* **8,** 381–401.

Troesken, W. (1996). "Why Regulate Utilities: The New Institutional Economics and the Chicago Gas Industry, 1849–1924." University of Michigan Press, Ann Arbor, MI.

Tussing, A. R., and Barlow, C. C. (1984). "The Natural Gas Industry: Evolution, Structure, and Economics." Ballinger, Cambridge, UK.

Weber, F. C. (1922). The future of the artificial gas industry. *In* "American Fuels" (R. F. Bacon and W. A. Hamor, Eds.), Vol. 1, pp. 1095–1106. McGraw-Hill, New York.

Williams, T. I. (1981). "A History of the British Gas Industry." Oxford University Press, Oxford, UK.

Williamson, H. F., and Daum, A. R. (1959). "The American Petroleum Industry: The Age of Illumination 1859–1899." Northwestern University Press, Evanston, IL.

Marine Transportation and Energy Use

JAMES J. CORBETT
University of Delaware
Newark, Delaware, United States

1. Introduction
2. Overview of Marine Transportation
3. Energy Choice and Marine Propulsion Development
4. Trends, Opportunities, and Challenges

Glossary

bunker Marine fuel, typically for international shipping. The term is adapted from the days of coal-fired steam boilers; fuel was stored aboard ship in coal bunkers and was shoveled into the boiler; when oil replaced coal, the bunkers became fuel tanks, but the name stuck.

deadweight ton (dwt) The weight of cargo, fuel, and supplies needed to make a vessel fully loaded; in other words, it is the weight difference between a completely empty and a fully loaded vessel.

gross registered ton (grt) The capacity on a ship, in cubic feet, of all spaces within the hull (body of the ship) and the enclosed spaces above the deck available for passengers, crew, and cargo, divided by 100; simply, 100 cubic feet of capacity is equal to 1 gross registered ton; the equivalent term, gross ton (gt), is also used.

marine distillate oil (MDO) A group of distillate oils that can be burned in auto-ignited engines (typically called diesel fuels). MDO can be specified with a range of properties such as density, flashpoint, and maximum sulfur content. Sulfur contents may not exceed 2% by weight. This fuel requires very little pretreatment for proper combustion in marine engines.

marine gas oil (MGO) A group of distillate oils that are typically clean and clear, used for emergency diesel engines, lifeboat diesel engines, etc. Marine gas oil is comparable to No. 2 fuel grade (onroad diesel fuel) used commercially ashore. MGO can be specified with a range of properties, such as density, flashpoint, and maximum sulfur content.

residual fuel (heavy fuel oil) (HFO) A group of residual fuels that can be burned in auto-ignited engines (typically called diesel fuels). HFO can be specified with a range of properties such as density, flashpoint, and maximum sulfur content. This fuel typically requires heating and filtering for proper combustion in marine engines. The maximum sulfur content allowed is 5% by weight, but the world average is much lower ($\sim$2.7% sulfur).

tonne-kilometer (tonne-km) The transport of 1 tonne of cargo over a distance of 1 km.

More than any other transportation mode, most modern shipping activity is designed to minimize fuel consumption because it helps minimize operating costs. As a result, waterborne commerce generally is considered an energy-conserving mode of transportation. On average, ships and barge tows accomplish the work of moving cargo on oceans and inland rivers at much less energy per tonne-kilometer, compared to trucking, rail, or air modes. This is especially true for cargoes shipped in bulk, such as oil or grain (see Table I). Published data suggest that there are important exceptions to this effort: (1) naval ships are designed for power and performance for military readiness and warfare and (2) some passenger ships (e.g., fast ferries) need to meet high-speed service requirements as a condition of high-value passenger service. The focus in this article is on the characteristics of the marine transportation system, the role of energy conservation in marine propulsion development, and trends in energy use for marine transportation as ships meet increased demand for cargo trade worldwide.

1. INTRODUCTION

Conservation of energy in marine transportation has been important since before the classical times of the Greeks and Romans. Shipping has been an important

 © 2004 Elsevier Inc. All rights reserved.

TABLE I

Comparison of Modal Energy Intensities[a]

Mode	National averages of energy intensity[b] (MJ/tonne-km)	Modeled ranges of energy intensity[c] (MJ/tonne-km)
Average road freight	1.8–4.5	
Heavy-duty trucks	0.60–1.0	
Rail (freight trains)	0.40–1.0	0.25–0.50
Marine freight (average)	0.10–0.40	
General cargo		0.20–0.50
Container ship		0.20–0.30
Bulk carrier		0.10–0.25
Oil tanker		0.10–0.20
Air freight	7–15	

[a] As the data show, marine transportation is less energy intensive, on average, compared to other transport modes. Specific energy intensities depend strongly on individual vehicle load factors and patterns of use.

[b] Data from United Nations Framework Convention on Climate Change Working Paper No. 1, Appendix A, Table 21.

[c] Data from International Maritime Study of Greenhouse Gas Emissions from Ships, Appendix A4, Figure 4.9.

human activity throughout history, particularly where prosperity depended primarily on commerce with colonies and interregional trade with other nations. From ancient times, when ships sailed mostly in sight of land using wind-driven sails or slave-driven oars, to today's oceangoing fleet of steel ships powered by the world's largest internal combustion engines, the history of shipping includes many examples of technical innovations designed to increase ship performance while conserving energy, capital, and labor. Whether the primary energy for ship propulsion has been wind and sail or engines and petroleum, all major innovations in marine transportation throughout history have involved balancing these factors. Since the 1800s, when industrialization made combustion power available to shipping, marine propulsion technology development has focused on ways to move a ship using less fuel and smaller crews (See Table I).

2. OVERVIEW OF MARINE TRANSPORTATION

The role of marine transportation in today's commerce and society must be considered to understand energy use practices and trends. Marine transportation in the 21st century is an integral, if sometimes less publicly visible, part of the global economy. The marine transportation system is a network of specialized vessels, the ports they visit, and transportation infrastructure from factories to terminals to distribution centers to markets. On a worldwide basis, some 44,000 oceangoing vessels move cargo more than 33 billion tonne-km annually (see Table II). In the European Union, marine transportation moves more than 70% (by volume) of all cargo traded with the rest of the world; in the United States, more than 95% of imports and exports is carried by ships. This work is accomplished by ships using 2 to 4% of the world's fossil fuels.

Maritime transportation is a necessary complement to and occasional substitute for other modes of freight transportation (see Fig. 1). For many commodities and trade routes, there is no direct substitute for waterborne commerce. (Air transportation has replaced most ocean liner passenger transportation, but carries only a small fraction of the highest value and lightest cargoes.) On other routes, such as some coastwise or short-sea shipping or within-inland river systems, marine transportation may provide a substitute for roads and rail, depending on cost, time, and infrastructure constraints. Other important marine transportation activities include passenger transportation (ferries and cruise ships), national defense (naval vessels), fishing and resource extraction, and navigational service (vessel-assist tugs, harbor maintenance vessels, etc.).

2.1 Marine Transportation System Complexity

The marine transportation system is a complex element of global trade. From an energy consumption perspective, the most important elements are the vessels; complexity of the marine transportation may be seen most easily by examining the complex mix of ships and boats that comprise the system's vehicles.

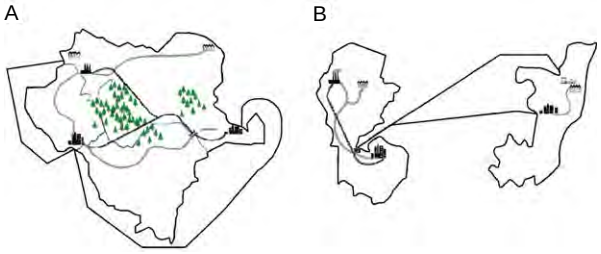


FIGURE 1 The marine transportation system as (A) a substitute for and (B) a complement to land modes.

TABLE II

Profile of World Commercial Fleet, Number of Main Engines, and Main Engine Power

Ship type	Number of ships	Percent of fleet	Number of main engines	Percent of main engines	Installed power (MW)	Percent of total power	Percent of energy demand[a]
Cargo fleet							
Container vessels	2662	2	2755	2	43,764	10	13
General cargo vessels	23,739	22	31,331	21	72,314	16	22
Tankers	9098	8	10,258	7	48,386	11	15
Bulk/combined carriers	8353	8	8781	6	51,251	11	16
Noncargo fleet							
Passenger	8370	8	15,646	10	19,523	4	6
Fishing vessels	23,371	22	24,009	16	18,474	4	6
Tugboats	9348	9	16,000	11	16,116	4	5
Other (research, supply)	3719	3	7500	5	10,265	2	3
Registered fleet total	88,660	82	116,280	77	280,093	62	86
Military vessels[b]	19,646	18	34,633	23	172,478	38	14
World fleet total	108,306	100	150,913	100	452,571	100	100

[a] Percent of energy demand is not directly proportional to installed power because military vessels typically use much less than their installed power, except during battle. The average military deployment rate is 50 underway time per year; studies indicate that when underway, naval vessels operate below 50% power for 90% of the time. Therefore, energy demand was adjusted in this table to reflect these facts.

[b] The data on which military vessel power was estimated specified the number of engines aboard ship associated with propulsion systems.

Ships, barges, ferries, towboats, and tugboats are among the most recognized and romanticized parts of the marine transportation system, perhaps second only to the water. Unlike automobiles, of which many thousands of each model are built and sold each year and a popular car design may remain largely the same for years, each ship design is unique to only one or a few vessels.

Because of different service requirements, ship vary greatly in terms of hull designs, power plants, cargo loading/discharge equipment, and other features. Inland-river and coastal towboats range in power from 0.5 to 4 megawatts (MW), or 600 to ~5000 horsepower (hp), and can push between 1 and more than 30 barges in a tow. Passenger ferries often have capacity to carry 150 to 300 passengers and range in power from 2 to 4 MW (~2600 to ~5000 hp), depending on speed requirements; some ferries, such as the Staten Island ferries in New York, have capacity for 6000 passengers. Roll-on, roll-off ships (or RoRos, so called because vehicles and other rolling cargo are driven onto the vessel rather than loaded by crane) can carry between 200 and 600 automobiles and have installed power between 15 and 25 MW (~20,000 and ~33,000 hp). Tankers and dry bulk carriers can carry more than 250,000 tons of cargo, with installed power often in the range of 25 to 35 MW (~33,000 to ~46,000 hp). Although container ships are not often as large as tankers, they have much larger power plants to accommodate greater vessel speeds. Average container ships carrying between 1750 and 2000 standard shipping containers have installed power of 20 to 25 MW (~26,000 to ~33,000 hp); larger container ships with capacity for more than 4000 containers can have installed power of 35 to 45 MW (46,000 to 60,000 hp), and the largest container ships can carry more than 6000 standard containers with engines rated as high as 65 MW (86,000 hp).

To accommodate complexity in geography, cargoes, and vessels, ports and terminals also vary greatly. The primary common goals at ports and terminals are to achieve rapid vessel loading and unloading and to maximize the rate of cargo transport

through the port facilities. Many ports are designed to facilitate container movements between ship, truck, and rail. Transportation planners and engineers have been working to speed this port transfer with on-dock rail, larger cranes, and improved queuing schedules for trucks. Some ports are megahubs that attract much of the world trade (by value and volume), whereas other ports are networked ports that handle much of the niche cargoes or serve as feeders for megaports. Each of the world's ports is unique by location (proximity to market), waterway configuration, and ecosystem, making general characterizations difficult. (Although each of these developments affects the overall energy consumption required to move cargo, the land-side energy demands of industrial port terminals are not addressed herein.)

In terms of the work performed, measured in tonne-kilometers or ton-miles, the marine transportation system rivals the other modes. The relative share of cargo moved by water, compared with truck and rail modes, varies from year to year, but U.S. waterborne commerce represents between 22 and 24% of the total tonne-kilometer of cargo movements in recent years. Truck and rail modes in the United States each account for about 25–29% of annual cargo tonne-kilometer. In Europe, the European Commission reports that waterborne commerce between nations (called shortsea shipping) accounts for nearly 41% of the goods transport market, compared to 44% for trucking and only 8% for rail and 4% for inland rivers.

One of the reasons that marine transportation moves so many tonne-kilometers of cargo is cost efficiency. Choice of mode is a function of cost (shipping is usually cheaper), time (shipping is often slower), and other quality-of-service factors. Only pipelines move goods at lower average cost than shipping, because of their fixed capital and ability to move fluids (liquids and gases), at very low energy requirements. Bulk shipment of liquid and dry goods such as oil and grain costs about two to three times as much per tonne-kilometer as pipelines, averaging between 2.5 and 3.5¢/tonne-km according to U.S. national transportation statistics. These average costs are generally less than but similar to rail, whereas marine transportation of bulk goods can cost much less than trucking. Except for high-value containerized goods that are shipped intermodally, the cost per tonne-kilometer for trucking can be more than an order of magnitude greater than the cost of marine transportation.

In energy conservation terms, shipping is not only among the least costly modes of transportation but efforts to keep down costs focus operators directly on fuel savings. Fuel costs can be the first or second largest cost of operating a vessel (not considering vessel capitalization); labor is generally the other largest cost and may vary with crew nationality. Fuel costs can range between 20 and 60% of operating costs, depending on vessel design and route factors. In order to be competitive, shipping companies try to reduce costs and/or add value to their service. These economic motivators have been very strong in the shipping industry and motivated much of the progress in energy conservation on ships.

2.2 International Fleet Propulsion Profile and Vessel Energy Use

The world's ships are primarily powered by diesel engine systems, consisting of one or more diesel engines in one of three typical propulsion configurations. Diesel direct-drive engines are typically large, slow-speed diesel engines directly connected to the propeller; they rotate at exactly the same revolutions per minute (rpm) as the propeller (0–130 rpm), except when connected to controllable-pitch propellers. Diesel gear engines drive the propeller indirectly through a set of gears, allowing the engine speed to be selected independently of the propeller. Diesel electric engines drive generator systems that turn the propeller; these engines typically operate at a constant speed, with electric power providing variable power to the propeller shaft. Diesel electric and diesel gear engines can operate at speeds greater than propeller revolutions per minute; in the international fleet, these are generally medium-speed engine designs operating between 130 and 2000 rpm. A fourth major category of engine, steam turbine systems, relies on conventional steam generation to drive a turbine geared to the propeller shaft; these marine systems typically have thermal and fuel efficiencies lower than modern marine diesel efficiencies and are rarely installed on new vessels. Other propulsion systems, not in wide use today, include combustion turbines (gas turbines) and older or nontraditional designs such as wind-powered vessels, reciprocating steam pistons, and nuclear power (for some military vessels). The evolution of engine technologies is discussed further in Section 3.

Most of the energy required by marine transportation is used on cargo or passenger transport vessels (ships that move cargo or passengers from one place to another in trade). A profile of the internationally registered fleet of ships greater than 100 gross tons is

shown in Table II. Transport vessels account for almost 60% of the ships and nearly 80% of the energy demand of the internationally registered fleet (not including military ships); considered along with military ships, cargo ships account for 40% of the world fleet and 66% of world fleet fuel use. Cargo ships are analogous to on-road vehicles because they generally navigate well-defined (if unmarked) trade routes, similar to a highway network. Other vessels are primarily engaged in extraction of resources (e.g., fishing, oil, or other minerals), or are primarily engaged as support vessels (vessel-assist tugs, supply vessels). Fishing vessels are the largest category of nontransport vessels and account for more than one-quarter of the total fleet. Fishing vessels and other nontransport ships are more analogous to off-road vehicles, in that they do not generally operate along the waterway network of trade routes. Rather, they sail to fishing regions and operate within that region, often at low power, to extract the ocean resources. As a result, fishing vessels require much less energy.

The exception to energy conservation designs in shipping would be the nearly 20,000 military vessels worldwide, which are sized for power and performance by their military mission, not for plant fuel economy. The naval ship is designed for sustained speeds in excess of their endurance (cruise) speed; as a result, the installed power in the world Navy is nearly 40% of the total installed power of the entire commercial fleet. However, the world's Navy ships use less energy than their installed power would suggest, because most of the time they operate in peaceful conditions at endurance speeds, partly to achieve fuel/cost savings. Average engine power characteristics for the U.S. Navy show that military ships operate below 50% power for 90% of the time that they are underway. Moreover, military vessels spend greater time in port than do modern commercial ships. Whereas commercial ships cannot effectively earn profits unless underway, Navy studies claim that military ships typically spend as much as 60–70% of their time in port; this is confirmed by public records of Naval ship activity, which show that typical deployment rates range between 40 and 55% of the fleet. This means that, in practice, military ships would demand some 14% of total energy required by the world fleet (commercial plus military fleets), much less than their installed power implies (Table II).

Fuel consumption is generally proportional to installed power on commercial vessels, and this power is concentrated in the larger commercial ships. About two-thirds of commercial ships are powered by four-stroke compression-ignition engines (operating on the compression-ignition, or diesel, cycle, and therefore referred to as diesel engines). More than one-quarter are powered by two-stroke diesel engines. Recent fleet statistics report that 6% of the ships have "unknown" diesel engines (i.e., either two- or four-stroke engines) and only 1% are turbine-driven. Most turbine-driven vessels (80%) are steam turbines with oil-fired boilers; the number of aero-derivative gas turbine engines in the commercial fleet is very low.

When estimating fuel consumption from commercial ships, the installed engine power rather than the number of engines or vessel size provides the most direct indication of energy requirements. The registered fleet has approximately 84,000 four-stroke engines with total installed power of 109,000 MW and some 27,000 two-stroke engines with total installed power of 164,000 MW. Engines with "unknown" cycle types and "turbines" together make up only about 2.5% of total installed power for main engines. This suggests that 27,000 two-stroke marine prime movers account for almost 60% of the commercial fleet's total energy output and fuel consumption, and more than half of the energy demanded by the world oceangoing fleet (including military ships). A majority of these engines are large-bore, low-speed diesel engines with above 10 MW rated output. Two-stroke engines are the main consumers of bunker fuel and therefore are the major sources of oceangoing ship emissions, followed by four-stroke engines.

The fuel types used in marine transportation are different from most transportation fuels. Marine fuels, or bunkers, used in today's fleet, can be generally classified into two categories: residual fuels and other fuels. Residual fuels, also known as heavy fuel oil (HFO), are a blend of various oils obtained from the highly viscous residue of distillation or cracking after the lighter (and more valuable) hydrocarbon fractions have been removed. Since the 1973 fuel crisis, refineries have used secondary refining technologies (known as thermal cracking) to extract the maximum quantity of refined products (distillates) from crude oil. As a consequence, the concentration of contaminants such as sulfur, ash, asphaltenes, and metals has increased in residual fuels. Residual fuel has been described by industry professionals as "black, thick as molasses, and has an odor that's pungent, to put it politely." By contrast, "other" bunkers used in international shipping include marine distillate oils of higher grade; however, these fuels often consist of a blend of heavier

fractions (residuals) with distillate oils and are not to be confused with the more highly refined fuels used in motor vehicles. Domestically, ferry vessels and harbor craft may use a more familiar diesel fuel that currently meets less stringent standards for maximum contaminant concentrations; however, in the United States, at least much of the marine fuel used by inland towboats, workboats, and ferries is rebranded on-road diesel. Few commercial ships, but many recreational vessels (not addressed here), use spark-ignited engines and marine gasoline oils (MGOs).

To reduce fuel costs, marine engines have been designed to burn the least costly of petroleum products. Residual fuels are preferred if ship engines can accommodate its poorer quality, unless there are other reasons (such as environmental compliance) to use more expensive fuels. Of the two-stroke, low-speed engines, 95% use heavy fuel oil and 5% are powered by marine distillate oil (MDO). Fuel consumed by 70% of the four-stroke, medium-speed engines is HFO, with the remainder burning either MDO or marine gas oil. Four-stroke, high-speed engines all operate on MDO or MGO. The remaining engine types are small, high-speed diesel engines all operating on MDO or MGO, steam turbines powered by boilers fueled by HFO, or gas turbines powered by MGO.

2.3 National and International Energy Statistics for Ships

The world's fleet operates internationally, but individual ships are registered under national authorities. Understanding which nations have maritime and trade interests is important to understanding which nations provide fuel for these ships. In terms of fuel sales, the nations providing the most fuel for ships are typically nations with strong interests in the cargoes or services those ships provide. Table III presents a summary of the top nations selling international marine fuels during the 1990s, according to the World Energy Database. The United States provides most of the world's marine fuels by far, and together the top 20 nations selling international marine fuels (shown in Table III) account for more than 80% of total marine fuel sales.

According to the Energy Information Administration's World Energy Database, some 140 million tonnes of fuel has been recorded as sold for international marine fuel consumption in recent years. At least 70% of these fuel sales has been identified as residual fuel, although recent work by Jos Olivier and co-workers at the National Institute for Public Health and the Environment in The Netherlands [Rijksinstituut voor Volksgezondheid & Milieu (RIVM)] suggests that residual fuel probably accounts for more than 80% or more of oceangoing fleet energy. (Typically, statistics for international "other" bunkers include international fuel sold to aircraft, which are excluded from the quantities discussed here.) This lack of clarity in the definitions makes the statistic somewhat uncertain, but the general trends are well documented.

Organization for Economic Cooperation and Development (OECD) nations account for roughly half of these fuel sales and provide a good illustration of historical consumption trends in the overall fleet, according to the International Energy Agency. Detailed fuel statistics for OECD nations over the past several decades (Fig. 2) show that marine fuel consumption has increased gradually, along with a growth in global trade and transportation. Over the entire period shown, the increase in marine fuel sales

TABLE III

International Marine Fuel Sales by Nation

Percent of total sales (1990–1999, inclusive)	Nations selling residual bunkers	Nations selling other bunkers
>15%	United States (18%)	United States (22%)
6–15%	Singapore (9%), Russia (9%), Netherlands (8%), United Arab Emirates (8%)	Saudi Arabia (12%)
2–5%	Japan, Saudi Arabia, Belgium, South Korea, Spain, South Africa, Greece	Hong Kong, Singapore, Netherlands, South Korea, United Kingdom, Spain, Russia
~1%	France, Taiwan, China, Italy, Egypt, Netherlands Antilles, Hong Kong, United Kingdom, Germany	Thailand, Greece, India, Belgium, Italy, Brazil, Indonesia, Denmark, Egypt, Venezuela, Germany
<1%	47 other countries	92 other countries

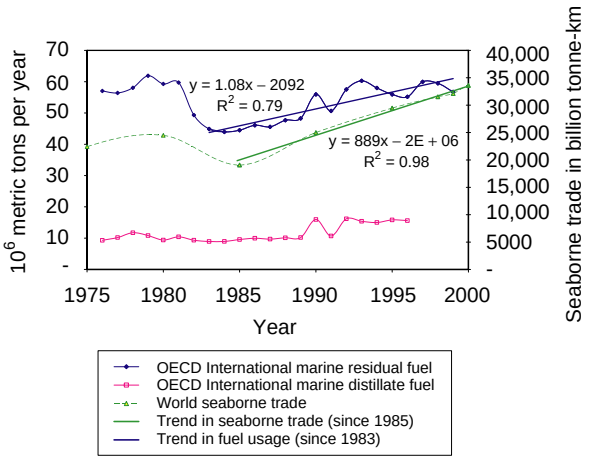


FIGURE 2 International marine fuel sales for Organization of Economic Development and Cooperation (OECD) nations and world seaborne trade, according to the International Energy Agency, Energy Information Agency, and United Nations data. The regression equations represent a simple linear statistical annual growth relationship. For residual fuels, the slope (1.08) suggests that, on average, an additional 1.08 million metric tons of residual marine fuel is sold each year by OECD nations; this represents an annual growth rate of about 2% for all international marine residual fuel sold. Note that there was significant variability in residual fuel sales during the 1990s and that the trend for international marine distillate furl is much flatter. For seaborne trade, the slope suggests that each year, ships move cargo an additional ~900 tonne-km; this represents an annual growth rate of about 3.5% in seaborne trade.

averages only about 0.2%/year, but since 1983 (i.e., after the recession in the early 1980s), these data show about a 2% annual average growth in residual fuel use. This growth in fuel consumption is about the same as recent annual growth rates in the fleet of ships, according to United Nations statistics. In other words, the fuel economy per ship may be roughly constant. However, over the long term, the fleet may be performing more useful work (i.e., transporting more cargo over longer distances) with a less-than-proportionate increase in energy demand. During a similar period (1985–2000), as shown in Fig. 2, seaborne trade increased slightly faster (~3.5%/year on average) than fuel sales; this could indicate that economies of scale and/or improved cargo logistics may work to reduce the energy consumed per cargo tonne-kilometer of work performed by the fleet.

An important caveat to consider is that these international fuel statistics do not represent the total fuel consumed by the international fleet of ships. Even making simplistic and conservative assumptions about the number of hours that vessels operate main engines and applying general fuel consumption rates for marine diesel engines, the international fleet of ships presented in Table II would require more fuel than reported as international marine bunkers. Recent estimates based on ship activity and installed engine power suggest that the world fleet of ships (including cargo, noncargo, and military vessels) consumes some 284 million metric tonnes of fuel per year, with more than 200 million metric tonnes required for cargo ships alone. This represents nearly 4% of all fossil fuel use.

Clearly, internationally registered vessels must be consuming domestic supplies of marine fuels. However, fuel statistics are generally well understood, especially for the OECD nations that sell most of the world's marine fuels. There are two reasons for the discrepancy. First, not all statistical sources for marine fuels have defined international marine fuels the same way. Original International Energy Agency (IEA) definitions have been reworded to be more consistent with reporting guidance under the *Revised 1996 IPCC Guidelines for National Greenhouse Gases*. The IEA defines "international marine bunkers (fuel) [to] cover those quantities delivered to seagoing ships of all flags, including warships. Consumption by ships engaged in transport in inland and coastal waters is not included." The IEA defines national navigation to be "internal and coastal navigation (including small craft and coastal vessels not purchasing their bunker requirements under international marine bunker contracts). Fuel used for ocean, coastal and inland fishing should be included in agriculture." Second, domestic fuel consumption inventories include fuel provided to oceangoing ships, at least during voyages that do not meet the definition criteria for international marine fuels.

Reconciling fleet fuel consumption between international and domestic fuel supplies is not simple. To get agreement with international fuel statistics and with the definitions, about 31% of the fuel consumed by transport vessels must occur while operating in shortsea or coastwise service using domestic fuels, and all of the nontransport ships (fishing, research, tugs, etc.) operate on fuel included in domestic inventories. This appears to be the case. The U.S. Energy Information Administration publishes state-by-state consumption statistics by fuel type and by end-use sector for the United States, the world's largest provider of marine fuels over the past 10 years at least (accounting for some 21% of marine residual sold by OECD nations). These Combined State Energy Data System (CSEDS) statistics identify residual fuel consumption in the transportation

sector in the United States domestic inventory, which represents marine residual fuels primarily because no other mobile source category consumes residual fuel.

A comparison of marine residual fuels on a state-by-state basis, alongside state-by-state waterborne commerce statistics, shows good agreement. First, total residual fuel consumption reported in the CSEDS agrees exactly with domestic residual consumption statistics in the World Energy Database, providing a consistency check. Second, the 10-year average residual consumption by marine transportation in the United States equals more than 40% of total residual consumption, more than enough to accommodate the difference between fleet energy requirements and international residual fuel statistics. Third, the top states for waterborne commerce (measured by total cargo volume) are also the top states for marine fuel sales. This suggests that most of domestic marine fuel consumption is going to cargo transport ships, and to vessels that directly support commerce. In terms of energy conservation, it is also worth noting that, although domestic land-based consumption of residual fuel has decreased significantly over the past two decades, marine transportation fuel consumption has not changed much, either internationally (Fig. 3) or domestically (Fig. 4).

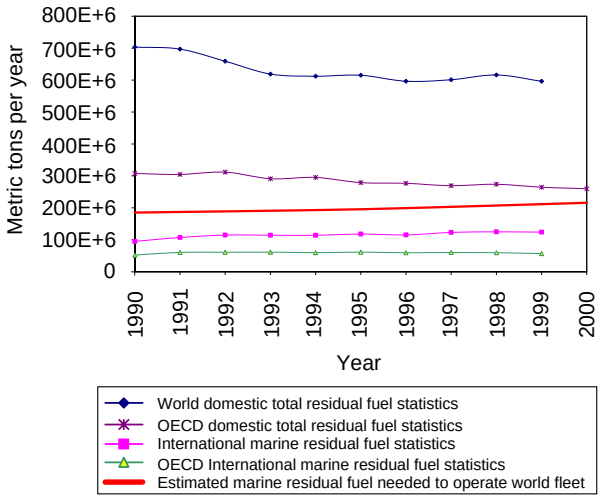


FIGURE 3 Comparison of domestic residual fuel consumption and international marine residual fuels worldwide and by Organization for Economic Cooperation and Development (OECD) members.

3. ENERGY CHOICE AND MARINE PROPULSION DEVELOPMENT

Aside from the shift of human labor (oars) to wind-driven sails, the first modern energy conversion in marine transportation was the shift from sails to combustion. Two primary motivators for energy

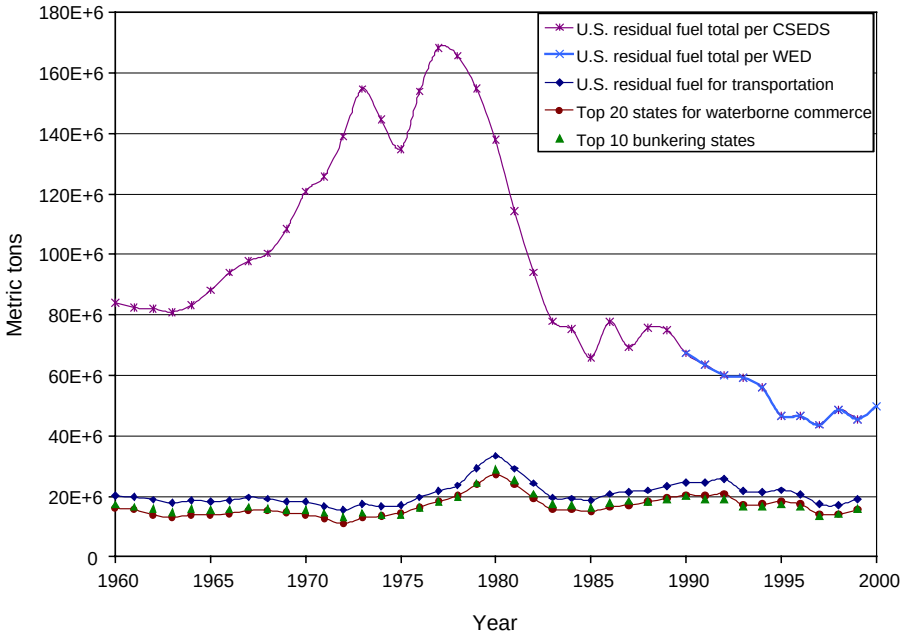


FIGURE 4 United States domestic residual fuel statistics, including residual fuel for transportation. Nearly 100% of residual fuel used in transportation is used by oceangoing marine vessels; nearly all inland vessels, harbor craft, and small commercial workboats use diesel fuel. CSEDS, Combined State Energy Data System; WED, World Energy Database.

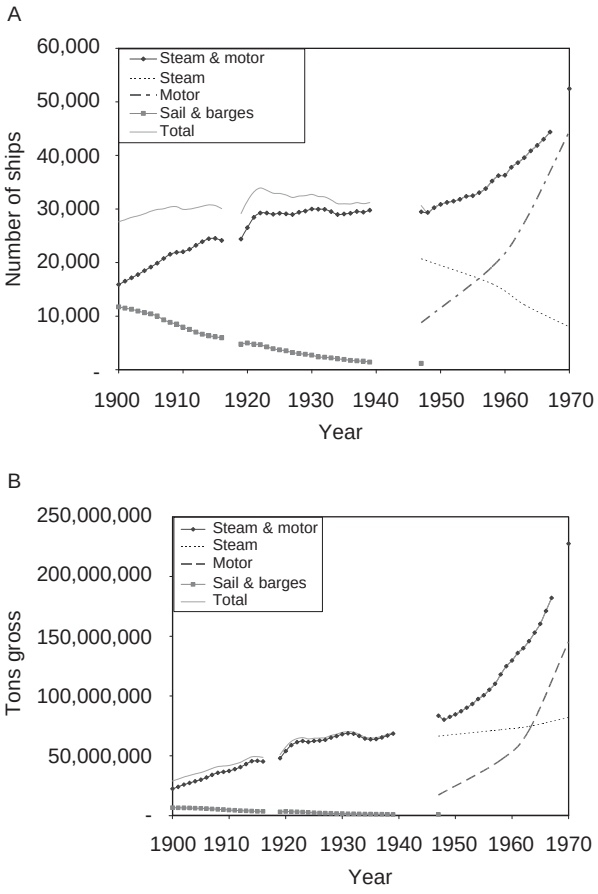


FIGURE 5 Conversion of ship propulsion, from sail to combustion propulsion, for internationally registered ships greater than 100 gross tons. (A) The conversion from sail as measured by number of ships took approximately 50 years to complete. (B) The effective fleet conversion of vessel tonnage to higher performing combustion systems took less than 20 years. During the shift from steam to motor, after 1948, the number of steamships decreased steadily. Economies of scale meant that ships with steam plants were still being built (mostly tankers) and that they were among the largest ships in the fleet. Note that ship registry data were not published during major world wars. Data from Lloyd's Register of Shipping Statistical Tables for various years.

technology innovation—greater performance at lower cost—caused this conversion. Figure 5 illustrates how this shift was completed during the first half of the 20th century, using data from Lloyd's Register Merchant Shipping Return for various years. Essentially, newer and larger ships adopted combustion technologies as part of an economy of scale. These technologies enabled trade routes to emerge, regardless of the latitudes, without consistent winds (referred to as the doldrums) supporting both international industrialization and modern political superpower expansion. As shown in Fig. 5, the conversion of fleet

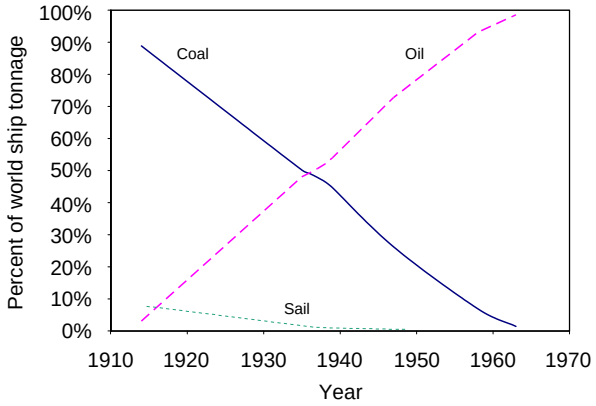


FIGURE 6 First major shift of world shipping to alternative fuel (coal to oil). Note that due to sparse data, particularly during world wars, these curves may misrepresent how smooth the transition was; the switch to oil was likely accelerated during periods of heavy shipbuilding (e.g., during the war periods and times of rapid economic growth). Data from Lloyd's Register of Shipping Statistical Tables for 1914, 1935–1939, 1947, 1958, and 1963.

tonnage to the preferred technology was achieved much more rapidly than was the phaseout of smaller ships using the outdated technology; this lead in conversion by tonnage was because the new technology was installed on the larger and newer vessels. Initially, these ships were powered by coal-fired boilers that provided steam first to reciprocating steam engines and later to high-speed steam turbines that drove the propellers. Later, the introduction of the industry's first alternative fuel, petroleum oil, enabled the introduction of modern marine engines. This pattern is repeated in many technology changes for marine transportation: some ship operators continue to use long-lived vessels purchased on the secondhand market while industry leaders replace their fleets to achieve new markets or realize economies of scale.

The switch from coal to oil was clearly motivated by a desire to conserve fuel. Switching fuels was both a direct means of conservation and a technology enabler during the introduction of internal combustion engines that ultimately replaced steam turbines and boilers. Of course, the reasons were economic: According to British Admiral Fisher's remarks to Winston Churchill in 1911 (quoted in Yergin's 1991 book, *The Prize*, page 155), a cargo steamer could "save 78 percent in fuel and gain 30 percent in cargo space by the adoption of the internal combustion propulsion and practically get rid of stokers and engineers." Essentially, the commercial sector was converting to oil-fired boilers powering marine turbines and introducing

oil-fueled internal-combustion, compression-ignition engines, commonly known as the diesel engine.

Figure 6 illustrates how international shipping switched from coal to oil over five decades. It is worth emphasizing that energy conservation was not the primary purpose for this conversion to "alternative fuel" in the early 1900s. Oil-powered commercial ships required fewer crew members and enjoyed a greater range of operations between fueling. This was not only of commercial interest; military vessels appreciated these advantages and the fact that refueling at sea could be accomplished more quickly and easily. Oil-powered ships also accelerated more quickly than did coal-powered systems, and could achieve higher speeds. When Churchill converted the British Navy to oil, despite the fact that this conversion made the British fleet dependent on a fuel not found within the British islands, the industry conversion was irreversible. After the fuel conversion was implemented, the next big shift was to more fuel-efficient marine diesel engines. A historical overview by Joe Evangelista in the American Bureau of Shipping's *Surveyor* describes how marine diesels came to dominate marine propulsion:

> *The first oceangoing diesel ship, the 370-ft Selandia, was delivered to the East Asiatic Company in 1912 by Burmeister & Wain (B&W) shipyard, Copenhagen, Denmark. In the 20 years since the first commercial diesel motor was proven, through the collaboration of Rudolph Diesel and Maschinenfabrik Augsburg-Nürnberg (MAN), the new type engine had been the subject of intense development and speculation. It promised industry a revolutionary new power source, and in so doing threatened the world's coal producers. So Selandia's sailing was one of the biggest news items of the time. Winston Churchill, Britain's First Lord of the Admiralty, visited the vessel and hailed it as an achievement of world proportion.*
>
> *Selandia was powered by twin B&W eight cylinder, four-stroke diesels, each developing 1,250 hp. Within the year, competitors like MAN, Sulzer, Vickers and Krupp were working to develop two-stroke engines with an output near 2,000 hp per cylinder—beginning a power competition that has not paused since. Observing these early efforts, Rudolph Diesel commented, "If, as seems probable, these tests give satisfactory results, the era of very large diesel engines has come."*
>
> *The king of propulsion for the biggest commercial vessels—tankers, bulkers, and containerships—is the large, slow-speed, two-stroke diesel engine. It became a market force during the late 1920s, as growing ship size and power requirements began slipping beyond the capabilities of four-stroke engines. By making the intake and exhaust phases occupy a fraction of a stroke, instead of a full stroke each, the two-stroke conferred several advantages. It doubled the power output per unit size over four-strokes, and made higher power outputs accessible through bigger piston sizes and longer strokes. In addition, the low speed two-stroke allowed direct coupling to the propeller. Over time, two-stroke manufacturers would also come to claim—as they do today—higher efficiency and greater overall reliability for these engines.*
>
> *The two-stroke diesel's 50-year battle against steam for supremacy of the seas was settled amid the skyrocketing fuel prices of the 1970s. In 1975, for example, of 614 engines delivered for the biggest commercial ships, 78 percent (482) were low-speed diesels, and the remaining 123 were steam turbines. In 2000, of the 1,059 main engines delivered to vessels over 2,000 dwt, 60 percent were low-speed two-strokes. Most of the remainder was medium-speed four-strokes.*
>
> *Four-stroke marine diesels are applied everywhere from racing boats and motor yachts to cruise ships, {freight and passenger} vessels and shuttle tankers (and in a large number of stationary power plants). Their higher speed requires some intermediary between engine and propeller (reduction gearing or diesel-electric drive), a setup offering advantages of quieter operation, lower vibration, and highly flexible engine room layout and location. Competition for this variety of business is divided among a large field of manufacturers, led by Wärtsilä (51 percent of the market in 2000) and MAN B&W Diesel.*
>
> *In their competition for higher performance [and fuel economy], diesel designers focused most of their efforts on raising power while controlling engine weight. To gage their success, consider just the power increase between two 420 mm-bore, 12-cylinder two-stroke engines, separated by 30 years. In 1968, B&W's 1242-VTBF-90 delivered a total output of 6,600 bhp on a stroke of 900 mm at a speed of 220 rpm. In 2001, MAN B&W's 12S42-MC delivered 17,640 bhp on a stroke of 1,764 mm at a speed of 136 rpm—a power increase of about 2.5 times within nearly the same engine silhouette. Such power increases were realized across the diesel spectrum. Top engine power jumped in that time from some 40,000 to 100,000 bhp. Average piston speeds increased from about 6.6 m/s to about 8.5 m/s.*
>
> *The tremendous increase in internal forces and pressures that accompanied this power rise was handled not by doubling the physical size of the engine, but through advanced materials, forging, and structural technologies. There have also been great gains in thermal efficiency, the measure of an engine's ability to get mechanical work out of the energy potential of the fuel. Efficiency has been raised from roughly 40 percent in 1975 to 50 percent today, with hybrid systems pushing 55 percent. In the orderly, high-tech workshops of today's diesel designers, the massive components of even the biggest engines—like piston crowns nearly a meter in diameter and fuel injectors 600 mm long—are hand-machined to watchworks precision in order to coax every fraction of power from the combustion process. New methods of fuel treatment, computerized injection and cylinder controls, and new exhaust handling technologies optimize power while lowering the pollutant content of engine emissions. And now, the last bastion of steam propulsion, the LNG carrier, is under assault from compact, attractively-priced diesels that burn a combination of oil and natural gas.*
>
> *—Joe Evangelista, in the American Bureau of Shipping's Surveyor (Spring 2002), pp. 14–20*

Although the introduction of motor ships dominated new construction, the phaseout of existing sail and steam vessels still took many decades. For example, in 1948, steam ships accounted for 68% of the ships in the fleet and 79% of the fleet tonnage, whereas motor ships accounted for 29% of ships and only 20% of the tonnage; sail still powered 4% of vessels but only 1% of registered ship tonnage. By 1959, motor ships accounted for 52% of vessels and 39% of registered tonnage in the fleet, and in 1963, motor ships represented 69% of vessels and 49% of registered tonnage. By 1970, motor ships dominated the fleet both in terms of ships and cargo tonnage, with 85 and 64%, respectively.

For maritime nations, this rapid modernization of the world fleet was anything but shared. In fact, a nation's fleet modernizes according to economic and political motivators within each country. This means that some large merchant fleets simply age while some of the latest technologies go to emerging maritime nations. For example, after World War II, the United States had the second largest fleet by number and the largest fleet by tonnage capacity; its fleet was populated by ships built during the war to carry supplies. These ships were nearly all steam turbine powered, and they would last at least two decades. As a result, the U.S. merchant fleet of ships did not invest significantly in newer motor-driven vessels; moreover, despite stated intentions to promote a modern and competitive fleet, policymakers continued a practice of industry subsidy and protection that discouraged domestic shipbuilding and subsidized older, costlier ships that burned more fuel and required larger crews. Even as late as the 1990s, when the world merchant fleet was more than 97% diesel powered, the U.S. oceangoing fleet was more than 50% steam-powered.

4. TRENDS, OPPORTUNITIES, AND CHALLENGES

The marine transportation system has clearly considered the importance of energy conservation in its technological and economic development. However, whether these achievements continue will depend on how marine energy systems respond to growth in trade (increased demand to perform work), emerging environmental regulations (competing constraints on power system design), and increased attention to CO_2 reductions to mitigate climate change (higher expectations for energy conservation improvements). Consideration of each of these in terms of the global fleet of ships that will continue to provide a primary transportation and trade service worldwide is the focus of this final section.

4.1 Increasing Trade Volumes

Global trade has increased dramatically over the past century, and projections continue to forecast a doubling of trade volume for major nations, and as much as tripling in cargo volumes to major world ports. Carriage of these increasing trade volumes will by necessity include water transportation, and energy conservation will be a factor. There exist three possible ways for the marine transportation system to accommodate these forecasted cargoes. First, the existing fleet might simply carry more through better logistics and more fully loaded ships to maximize efficiency. Second, additional ships may be added to carry the additional cargoes, resulting in an increase in ship traffic overall. Third, the fleet might transition to larger vessels without increasing the overall number of ships in service in order to provide greater capacity within the same network of routes. In practice, each of these alternatives may be applied in some combination.

First, on certain routes and for certain cargoes, it may be possible to carry more cargo with the current fleet. However, there is an upper bound to this alternative. Most shipping routes are practically unidirectional, meaning that most of the cargo flows in one direction. For example, oil is primarily imported into the United States, and grain is primarily exported. Bulk carriers typically transit empty (or under ballast) in one direction and return with a load of cargo; this results in a practical upper bound of 50% capacity overall on routes for bulk carriers (e.g., averaging 100% one way and 0% the other, or more realistically, averaging 80–90% one way and 10–20% on the return). For container and general cargoes, some routes are able to secure a backhaul; however, there is limited ability for ships to realize this idealized goal. Where a strong backhaul market can be established, the backhaul freight rates usually are much lower and may not directly cover the costs of the voyage; here, as for bulk carriers, the freight rate for "inbound" cargoes must cover most of the cost of the round-trip voyage. For inland-river and shortsea shipping, backhauls are more practical, because vessel capacities are smaller. In fact, the industry appears to behave as though there is an upper bound. Most ships carry loads that average 50–65% capacity or less. When cargo capacities exceed 70%, it can be an indication that

too few ships are available for the route; under these conditions, ships are added to the route. Where possible, this may help reduce growth in energy consumption by the fleet.

Second, adding ships to increase route capacity is easier than it might be for on-road vehicles in the current transportation system. This is primarily because ocean routes are "unconstrained" highways. For most route distances, few modifications will be needed to enable more vessel traffic. However, in and near port regions, this may not hold. The number of berths at terminals and cargo handling rates provide at least two important limits on how many vessels a port can accommodate in a given period. No consistent metric to evaluate port capacity exists, but clearly, the capacity of the marine transportation system will be limited if ships have to wait for a berth at a port. This may involve increased energy demand for waterborne transportation, and unless these added ships divert cargoes from land-based modes, the overall energy consumption will likely increase.

The third alternative is to continue increasing economies of scale for shipping with larger ships, better cargo stowage technologies, and improved cargo logistics planning. Examples of this include the 6000-tonne equivalent unit (TEU) containerized "megaships" and supertankers. Very large ships are being employed on some routes where the cargo volumes are strong, and this practice is projected to grow. However, trade logistics may limit the cost-effectiveness of applying this option across the fleet. There will continue to be important niche routes and specialty cargoes that require vessels smaller than the megaships. This alternative may allow for even greater growth in trade while minimizing increases in energy consumption.

In general, the challenge of decoupling energy requirements from a growing global trade will be significant. The cost motivation of carriers to reduce fuel consumption will continue to promote conservation research and innovation; however, cheaper fossil-based fuels may limit the conservation progress in shipping. This is an important consideration, because many of the alternative energy sources are much more expensive than residual fuels and will require matching engine technology to the fuel.

4.2 Environmental Constraints and Continuing Conservation Efforts

Environmental impacts from shipping have been the focus of increasing attention over the past decades, primarily focused on waterborne discharges and spills of oil, chemical, and sewage pollution. A number of international treaties and national laws have been adopted, along with industry best practices, to prevent these pollution releases through accident or substandard operation. However, the shipping industry is seen as an important hybrid between land-based transportation and large stationary power systems. On the one hand, these are non-point-source vehicles operating on a transportation network of waterways; on the other hand, these vehicles operate ship systems that compare to small power-generating plants or factories. As industry leaders and policymakers at state, federal, and international levels recognize these facts, environmental and energy performance is being measured and regulated. Recent efforts to mitigate environmental impacts from shipping (for example, invasive species in ballast water, toxic hull coatings, and air emissions) are relatively new for the industry and will take decades to address.

Environmental regulation of the shipping industry typically lags behind regulation of similar industrial processes (engines, etc.) on land. At the international level, the International Maritime Organization has effectively developed a set of treaties addressing much of these ship pollution issues. Under the International Convention for the Prevention of Pollution from Ships, these regulations (MARPOL 73/78, a combination of two treaties from 1973 and 1978) apply global standards to environmental protection practices aboard ship. However, national and state jurisdictions can apply stricter regulations to certain vessel operations and to the ports that receive these ships. For example, the U.S. Environmental Protection Agency recently regulated shipping under mandated authority in the Clean Air Act. Another example includes efforts by port states in Sweden and Norway, and by individual ports in California, to apply market-based or voluntary regulations to reduce air pollution from ships entering their ports.

Environmental regulation may make some energy conservation efforts more difficult, by introducing environmental trade-offs. For example, hull coatings are specialized for marine conditions to prevent hull fouling; some of these coatings have been shown to affect the marine environment, particularly in ports and harbors, by releasing toxic chemicals or metals. These compounds slowly "leach" into the seawater, killing barnacles and other marine life that have attached to the ship. But the studies have shown that these compounds persist in the water, killing sea life,

harming the environment and possibly entering the food chain. Toxic hull coatings are being banned by national and international regulations, and this may threaten to increase energy consumption by shipping. Replacing these coatings with less toxic alternatives may result in increased rates of hull fouling and increased fuel requirements.

4.3 Future Alternative Fuels and Energy Conservation

Future energy conservation efforts in shipping are increasingly considering factors other than cost. Certainly, energy independence and fossil fuel resource availability continue to play a factor. Traditional pollution mitigation also plays a role. And most recently, climate change concerns are motivating reductions in CO_2 emissions that are directly proportional to fossil fuel combustion. Marine fuels, like all petroleum products, are about 86% carbon. If achieving greater reductions in CO_2 from marine transportation is to be achieved through means other than cost-driven fuel economy, then alternative fuels will likely be introduced.

It is worth noting that alternative fuels for energy conservation were the focus of significant attention during the energy crisis in the 1970s. In fact, the National Research Council conducted a large study, Alternative Fuels for Maritime Use, motivated explicitly by President Carter's national energy policy. That report proposed a 20-year plan to conduct research and development toward conversion of fleet to alternative energy sources; however, very little marine transportation research was pursued because energy independence descended from the top of the public policy agenda after the crisis. To a large extent, this work is being updated by current policy efforts to achieve CO_2 reductions to mitigate climate change impacts. Whether the political attention will sustain such a long-term research and development effort remains to be seen; after all, it took nearly five decades to switch the industry from coal to oil even with the performance advantages.

The current focus on alternative fuels in maritime applications involves various types of fuel, including biofuels such as biodiesel, clean diesel products, natural gas, and hydrogen. Solar, wind, and other noncombustion sources (e.g., nuclear) are sometimes identified but do not appear to be particularly suited for modern ships or global seaborne trade. Hydrogen is attracting increasing attention as a long-term goal among alternative fuels; for that reason it will be the main example for this discussion, although the issues discussed generally apply to other fuels as well.

There are challenges to the successful introduction of alternative fuel in any transportation mode. Five factors that may favor marine transportation as a cost-minimizing mode for hydrogen (or other alternative fuel) introduction have been discussed by Farrell et al. in a 2003 *Energy Policy* article. In general, converting to hydrogen fuel in freight modes first may offer certain cost-saving advantages over passenger-only modes of transportation. Some factors may apply equally or better to nonmarine modes as well, but here they are discussed only in context with marine transportation.

First, vehicle design and performance factors are important constraints on less dense fuels such as hydrogen. When the fuel-storage:payload ratio is high, switching to hydrogen will involve greater cost trade-offs. For smaller ferries, this may mean carrying significantly fewer passengers; for large cargo ships, the loss of cargo capacity may be less. Also, if the alternative fuel (hydrogen) has less energy density, then there may be impacts on acceleration, which would matter less to slower, oceangoing vessels than to high-speed harbor craft.

Second, ship operations and management may be well suited to adopting alternative fuels. Ship crews are generally well trained and licensed, and requirements for shipboard qualifications are becoming stricter. This means that commercial vessels with professional crews may be able to adopt innovative technologies more safely and at lower cost than the general public can accomplish. Moreover, if there is an increased risk of accident from the alternative fuel (which may not be strictly true for hydrogen versus petroleum fuels), then introducing a new fuel into marine cargo transportation may be accomplished through crew training that minimizes exposure to the general public. Also, fueling operations for marine vessels occur at commercial docks or by fuel barges in a commercial port harbor. Furthermore, because most commercial vessels are operated around the clock and throughout the year, investments to convert to alternative fuel may be spread across a greater period of operation than, say, is the case for passenger vehicles, which are typically driven only a few hours each day.

Third, one of the most important factors in alternative fuels is the cost of energy infrastructure. Larger, more centralized refueling locations tend to reduce the cost of delivering the alternative fuel. In the case of marine transportation, about 30 major

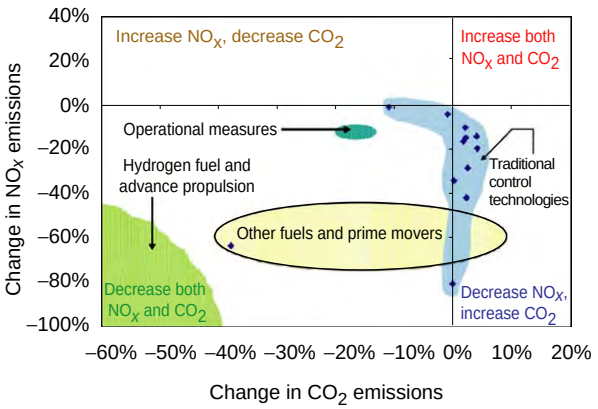


FIGURE 7 Potential trade-off between technologies to achieve reductions in traditional pollutants (NO_x) and fuel economy (in terms of CO_2).

fuel providers (refiners) in the United States supply most of the marine fuel, and most of that fuel is delivered in the top 20 ports for waterborne commerce.

Fourth, reduced air emissions are one of the benefits offered by many alternative fuels, and improved environmental performance may be greater in less regulated industries such as international and domestic shipping, compared to passenger cars, for which emissions standards have been applied increasingly for more than three decades. Converting ships to hydrogen will reduce emissions of traditional pollutants significantly, depending on the prime mover used. Figure 7 illustrates the potential trade-offs between different technologies and fuels aimed at reducing ship air emissions versus achieving greater energy conservation. It appears likely that any of the trade-offs will not be motivated by a reduction in vessel operating cost.

Fifth, shipbuilding and marine engine production rates will also be important to any fuel conversion. On the one hand, the facts that ships use tailored designs with generally similar power systems and that they are built one at a time make the process of innovation more effective than for massed-produced vehicles such as passenger cars. On the other hand, the few number of ships that are built annually and their long operating lifetimes make economies of scale for new propulsion systems difficult to achieve. However, marine engines are currently available that operate on gaseous fuels, and the long tradition of engine designs that can handle a wide range of fuel properties may contribute to successful demonstration efforts.

SEE ALSO THE FOLLOWING ARTICLES

Aircraft and Energy Use • Fisheries and Energy Use • Internal Combustion (Gasoline and Diesel) Engines • Tanker Transportation • Transportation and Energy, Overview • Transportation and Energy Policy • Transportation Fuel Alternatives for Highway Vehicles

Further Reading

Corbett, J. J., and Fischbeck, P. S. (1997). Emissions from ships. *Science* **278**(5339), 823–824.

Corbett, J. J., and Fischbeck, P. S. (2000). Emissions from waterborne commerce in United States continental and inland waters. *Environ. Sci. Technol.* **34**(15), 3254–3260.

Corbett, J. J., and Koehler, H. (2003). Improving the accuracy of ship emissions inventories. *JGR Atmospheres* **28** (D20), 4650–4666.

Energy Information Administration (EIA). (2001). "World Energy Database and International Energy Annual 2001." EIA, Washington, D.C.

Evangelista, J. (2002). Big diesel balancing act [quarterly magazine of the American Bureau of Shipping]. *Surveyor* (Spring 2002), 14–20.

Farrell, A. E., Keith, D. W., *et al.* (2003). A strategy for introducing hydrogen into transportation. *Energy Policy* **31**(13), 1357–1367.

International Energy Agency (IEA). (1977–1997). "Quarterly Oil Statistics" [various issues]. Organization for Economic Cooperation and Development, Paris, France.

International Organization for Standardization (ISO). (1987). "International Standard, Petroleum Products—Fuels (Class F)—Specifications of Marine Fuels." ISO, Geneva, Switzerland.

Kendall, L. C., and Buckley, J. J. (1994). "The Business of Shipping." Cornell Maritime Press, Centreville, Maryland.

Lloyd's Maritime Information System (LMIS). (2002). "The Lloyds Maritime Database." Lloyd's Register, Fairplay Ltd., London.

Lloyd's Register of Shipping. (1970–1994). "Merchant Shipbuilding Return." Lloyd's Register of Shipping, London.

National Research Council (U.S.) Commission on Sociotechnical Systems. (MTRB). (1980). "Alternative Fuels for Maritime Use." National Academy of Sciences, Washington, D.C.

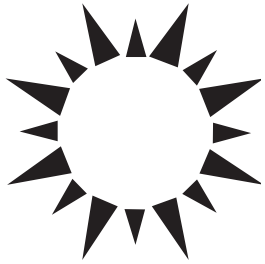
Skjølsvik, K. O., Andersen, A. B., *et al.* (2000). "Study of Greenhouse Gas Emissions from Ships (MEPC 45/8 Report to International Maritime Organization on the outcome of the IMO Study on Greenhouse Gas Emissions from Ships)." MARINTEK Sintef Group, Trondheim, Norway; Center for Economic Analysis, Carnegie Mellon University, Pittsburgh; and Det Norske Veritas, Oslo, Norway.

United Nations Conference on Trade and Development (UNCTAD). (2002). "Review of Maritime Transport 2002." United Nations, Geneva.

United States Army Corps of Engineers (USACE). (1999). "Navigation Data Center (NDC) Publications and U.S. Waterway Data." Available on compact disk from Water Resources Support Center, NDC, Alexandria, Virginia.

Yergin, D. (1991). "The Prize." Touchstone, New York.

Market-Based Instruments, Overview



STEPHEN FARBER
University of Pittsburgh
Pittsburgh, Pennsylvania, United States

1. Types of Instruments
2. Principles and Applications to Energy-Related Environmental Management

Glossary

command-and-control A regulatory method by which acceptable activities are tightly prescribed, such as designating the permissible technologies for pollution control.
double dividend The claim that environmental taxes not only benefit the environment, but the collected revenues can be beneficially used to reduce other tax distortions.
emissions allowances Certifiable permits to emit pollutants.
emissions reduction credits Certifiable credits received for reducing emissions below allowance levels or beyond required reductions.
externalities Inadvertent effects on persons, such as adverse health effects of pollution, from activities of production or consumption.
marginal control cost Additional costs of controlling emissions by one more unit (ton, pound, etc.).
market-based instruments Regulatory tools that use market-like forces to provide flexibility and incentives to meet regulatory objectives.
performance bonds A financial obligation to pay when performance criteria are not met.
pigovian taxes Levies that reflect the social costs of activities, such as pollution, that are not typically accounted for in traditional markets.
tradable permits Quantified permits to engage in an action, such as emissions, whose ownership can be transferred among parties.

Market-based instruments (MBI) are proposed as one of several types of regulatory tools for attaining environmental, ecological, and resource management objectives. MBI mimic processes found in traditional markets, such as price signals, trading, and information. They allow behavioral flexibility by entities acting in their own self-interest. MBI include the following specific instruments: prices, marketable rights, subsidies, liability, and information. They contrast with command-and-control (CAC) instruments, which stipulate with varying degrees of specificity what actors must do and how they must do it, for example, establishing a technology standard for pollution control or specifying allowable emissions. MBI are also referred to as incentive-based instruments, although the latter term is a bit misleading as CAC instruments with fines for violations create incentives. Each of these MBI create circumstances where some activities critical to regulatory objectives are not fully prescribed, leaving varying degrees of discretion open to the regulated parties. Interest in MBI reflects concerns for cost-effective management and the need to persuade regulated parties to buy into regulation. They provide more flexibility in the achievement of environmental goals than CAC. Whereas CAC instruments have been successful in moving toward these goals, their typical "one size fits all" rigidity and centralized specification of behavior standards have led to concerns of overly costly regulations, as actors are allowed little flexibility and have minimal incentives for developing new control technologies. Cost savings obtained by shifting from CAC to MBI may be a prerequisite to the attainment of higher, more costly, environmental standards.

1. TYPES OF INSTRUMENTS

1.1 Prices

The function of prices as a MBI is to force actors to explicitly consider the environmental, ecological, or

 © 2004 Elsevier Inc. All rights reserved.

resource implications of their activities. By placing a price on an activity, the actor can then consider the trade-offs involved in avoiding that activity. For example, a price placed on each ton of emissions should induce actors to consider the cost of reducing emissions by 1 ton compared to the price they must pay if they do not reduce. This is illustrated in Fig. 1. Emissions reductions become progressively more expensive, as the Marginal Control Cost illustrates. These control costs are attributable to any options the actor wishes to employ to avoid emissions, including end-of-pipe controls, internal process changes, input selection, and reduced product output. The actor would have complete flexibility in selecting options and would presumably select those that are least costly. If a price is levied for emissions, an actor seeking to keep emissions-related costs low will find it cheaper to reduce emissions up to level ER* than to pay the price of not reducing emissions. The total costs of reducing emissions by an amount ER* in Fig. 1 are represented by area B and the net cost savings by avoiding the price levy are represented by area A. The actor will not reduce emissions beyond ER*, as it costs more to do so than would be saved by avoiding the levy.

Prices may be levied directly on the activity of concern to the regulator, such as emissions, or on related activities, such as the sulfur content of fuels. Where the levy is placed determines its effectiveness. Administrative concerns may determine where and how the price is levied. For example, it is administratively simpler to place a price levy on permitted activities, such as emissions volumes, than on other less readily observable activities. Legal and political concerns may determine whether the levy can be a tax or a fee, where the latter is reserved for levies that compensate the regulator for services rendered, such as a permit fee for financing the permitting process.

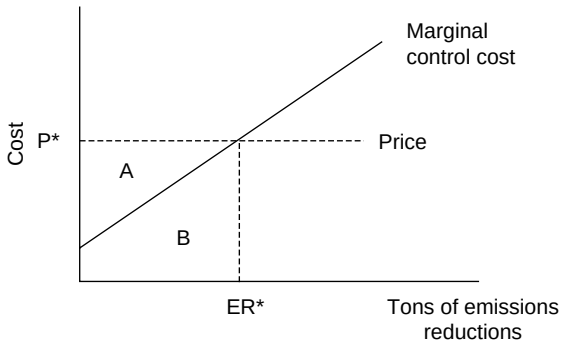


FIGURE 1 Emissions pricing and control.

The terms fines and penalties are reserved for charges based on violations of stipulated activities and are typically associated with CAC instruments.

Financial assurances, such as performance bonds, are a potentially useful MBI that place an explicit price on undesirable events. Typically the actor must post some financial assurance, such as a bond or letter of financial assurance, that can be drawn upon if some undesirable event, such as a spill or incomplete mining remediation, occurs. A price is placed on the event and the actor must pay it when the event occurs. The anticipation of this price obligation guides the actor's behavior. These assurances both manage undesirable activities through a price signal and ensure equity by holding the actor financially accountable, rather than leaving others with unwanted liabilities.

Pricing has many advantages from a social perspective. A single price sends the same signal to all actors, inducing each of them to control behavior to the same level of marginal control costs, which means that there are not some actors whose control costs far exceed the costs of others. It appears simple to administer, although it requires verification through monitoring, which may be more expensive than the cost savings from control efficiencies. Actors have incentives to consider new technologies that reduce their control costs, but only when the prices remain the same and are not lowered as all actors find cheaper methods of control. But actors may not be supportive of pricing, as they must pay for control plus the levies on remaining activities or emissions. Of course, levies may be placed on activities above a certain level to counteract potential opposition to this "double cost."

The benefits of pricing socially adverse environmental or ecological impacts of activities are reductions in welfare losses from overuse of the environment and overuse of resources in activities that harm the environment. In addition, some have claimed that using these environmental prices for public revenues allows the reduction in perverse or resource distorting taxes, such as income taxes, that diminish welfare. The benefit to welfare from environmental enhancement and reductions in otherwise distorting taxes is referred to as the "double dividend" or environmental tax reform. There is controversy over whether the second dividend is large or even positive. Environmental taxes may not raise significant revenues if the environmental activity can be easily reduced when taxes are imposed. The welfare costs of preexisting noncompetitive market structures may only be increased with such taxes. And revenue-neutral shifting of taxation to undesirable

environmental activities and away from desirable activities, such as work, would increase commodity prices to such an extent that real wages would fall, discouraging labor participation.

1.2 Marketable Rights

Endowing the actor with rights to some activity and allowing the actor to market those rights provide flexibility in achieving regulatory objectives. For example, a power plant can be allowed to emit a given volume of SO_2. But if it is allowed to buy or sell these permitted emissions allowances, they become marketable rights or tradable permits. The prudent plant will then buy or sell them based on the market prices of these allowances compared to the cost of emissions control. Plants that find it more costly to control emissions to levels permitted under their initial allowances than the market price of allowances will seek to buy allowances in the market and emit more than their initial allocation and vice versa. The result is that high-cost plants can avoid excessive control costs and low-cost plants can control to higher levels than their initial allowances. These marketable rights shift control responsibilities across sources so that a total emissions objective can be attained at lower costs than under the initial allocation of responsibility and they provide individual plants with complete flexibility in determining how much and how they will control emissions.

The rights can be for allowable activities, such as emissions allowances, or for reductions in activities, such as emissions reduction credits. The principle of trading is the same; responsibility is shifted from those who find it more costly to meet objectives to those who find it less expensive. For example, in Fig. 1, if an actor is required to reduce emissions more than ER* and the price of emissions reduction credits is P*, the actor will buy emissions reduction credits, thus allowing emissions to be reduced to a lesser degree. The actor would end up reducing emissions by ER*, as it would be less expensive to buy reduction credits than incur the greater costs of control. Similarly, an actor with an initially assigned responsibility of less than ER* will find it profitable to reduce emissions more, as more can be received for the sale of these excess reductions than it costs to create them.

The prices established through marketable rights trading guide the behavior of the actors, just as direct prices set by regulators. So there is a strong parallel between the two sets of instruments. They both have the potential to accomplish regulatory objectives at lower costs and with more flexibility than CAC instruments.

1.3 Subsidies

Subsidies are the opposite of the pricing instrument outlined in Section 2.1. Whereas a price is a payment extracted from actors enabling them to do something, a subsidy is a payment to an actor to do something. An example would be a payment to a power plant for each ton of SO_2 removed from the waste stream. In Fig. 1, if the subsidy is P* for each ton removed, the plant would remove ER*, as the reward for removal exceeds the cost of emissions reduction up to that point. So the subsidy and price instruments can achieve the same regulatory objectives and associated flexibility and cost-effectiveness. The major difference is in the wealth position of the actors under the two instruments. With a pricing system, the actor's wealth after regulation is the original wealth minus control costs (B); in a subsidy system, the actor's postregulatory wealth is the original wealth plus the net of subsidy payments over control costs (A). Clearly, the actor's postregulatory wealth is greater under the subsidy system.

The subsidy could be directly on the behavior the regulator wished to encourage or it could be indirectly related to it. For example, the subsidy could be directly on emissions reductions or a payment for installing a scrubber. The subsidy could be an explicit payment, or an indirect benefit, such as a special tax privilege.

A deposit–refund system is a hybrid price-subsidy system. For example, an electric battery purchase may require a deposit, which is a price. But if the battery is returned, the actor receives a refund, which is a subsidy for its return. The magnitude of the deposit and the differential between the deposit and the refund guide both the original purchase and the disposal of the battery.

1.4 Liability

Legal principles, such as common laws of trespass and negligence or statutorily designated liabilities, place implicit costs on actors for undesirable activities. Actors then face potential liabilities for undesirable and harmful activities. These are liabilities beyond regulatory obligations such as CAC fines and penalties. In a sense they are prices, but with a higher degree of uncertainty over magnitude and incidence than explicit regulatory pricing instruments. The potential obligation for an undesirable

activity is probabilistic, requiring a range of proofs from responsibility to magnitude of harm.

A variant of liability MBI is third-party insurance coverage. The purpose of insurance is to shift financial responsibility from an actor to the insurer. The insurer then has an incentive to establish premiums in accordance with risk and to monitor the behavior of the insured. Insurance coverage costs then reflect the insurer's assessment of risks, providing a signal of potential liabilities and the insurer becomes a quasi-regulator through monitoring. Insurance may not protect actors against liabilities due to negligence or illegal behavior.

1.5 Information

Information provided to the public about actors' behaviors may change environmental management activities. For example, public knowledge about a firm's emissions, or the "greenness" of its production processes, may galvanize public attitudes and induce firms to become good "corporate citizens." This moral suasion may extend to actual impacts on firms' profitabilities, as demands for their products are diminished or it becomes more difficult and costly to operate in the community.

2. PRINCIPLES AND APPLICATIONS TO ENERGY-RELATED ENVIRONMENTAL MANAGEMENT

The MBI are based on their ability to meet regulatory objectives in a manner that increases actor flexibility and reduces the costs of meeting objectives. The designs of these MBI are based on some principles of application. These principles suggest that some MBI are more appropriate for some circumstances and management objectives than others. These principles are discussed below, along with observations on strengths and weaknesses. Applications to energy-related issues are also presented for each MBI.

2.1 Prices

2.1.1 What to Price?

The regulator has control over actor behavior through the price. Prices may be placed on the behavior most immediately related to the regulatory objective. If the objective is to control emissions, the price can be placed directly on those emissions. The actor then has a host of activities that can be employed to avoid paying the price. However, the further the priced activity is from the regulatory objective, the less perfectly and the more costly it is for the regulator to manage for that objective. For example, SO_2 emissions from power plants may be regulated by placing a price on those discharges. Alternatively, levies based on the sulfur content of fuels would induce plants to shift to lower sulfur fuels but not induce them to change internal production processes, alter production levels, install scrubbers, etc. The pricing of fuel content does not induce the plant to engage in a whole host of activities downstream from the fuel purchase that could also reduce emissions, perhaps more cheaply.

Pricing the regulatory target may not be feasible, either economically or technologically. For example, nonpoint source run-off of nutrients from agriculture would be a candidate for targeted pricing if it were not so difficult to measure that run-off. Although it may not be impossible to measure, either directly or through nutrient transport modeling, it may be more feasible to price the nutrient inputs through charges on fertilizer. The total costs of meeting regulatory objectives must consider both the costs to agricultural production and the administrative costs. Pricing at points only indirectly related to the regulatory targets may be most cost-effective when considering both of these costs.

2.1.2 The Price Level

The level of the prices depends on the regulatory objective. A general economic principle is to set prices for activities at levels that reflect the cost to society of those activities. When traditional markets do not fully reflect these full costs, as would be the case for externalities such as pollution, a tax can be levied to correct for these price distortions. This tax is referred to as a "Pigovian Tax," named after the British economist who proposed taxing activities according to the damages they imposed on society but which were not accounted for in typical market decisions. This levy forces actors to "internalize" the adverse consequences of their actions. Such a levy would assure society that all benefits and costs of activities would become accounted for in market supply and demand conditions. For example, if it is known that a ton of emissions will change ambient concentrations by a known amount and that an estimable number of persons will become ill, one could set the emissions charge at a level reflecting the costs of those illnesses. For example, suppose that each ton of emissions harms society by an amount P*

in Fig. 1. A levy of P* would induce actors to consider those emissions and they would reduce emissions as long as the levy exceeded the costs of emissions reduction. In Fig. 1, they would control to point ER*.

But regulators are not likely to know these social costs with much exactness or to a level of certainty to avoid challenges of "arbitrary and capricious." The alternative is to establish charges related to the known costs of control. For example, in Fig. 1, if the regulator knows the marginal control costs over a wide range of emissions reductions, it can set the price to induce the actor to reduce emissions by ER*. In fact, setting the price and observing actor behavior will indicate what the marginal cost function is, if the actor engages in a rational comparison of control costs to prices.

When price levels are designed to reflect the social costs of activities, such as pollutant damages, non-uniformly mixed pollutants must bear different prices. Suppose the same emissions from two plants result in different health impacts, where 1 ton from plant A has half the impact as 1 ton from plant B. The levy on plant A should then be half that of plant B.

Pigovian taxation is designed to induce actors to behave in the best interests of society by guiding their activities and resources so that the full social costs are considered. When prices on activities do not reflect the potential social costs of those activities, the guidance advantages of MBI are lost. For example, some emissions pricing, such as permit fees that vary with permitted volumes, is designed to collect sufficient revenues to fund the regulatory program. Such prices may be greater or less than the damages they impose on society. They are not Pigovian taxes, but simply fees for services rendered.

2.1.3 Pricing with Uncertainty

In principle, when a price is placed on activities, the actor will engage in control of those activities to the point where the cost of control equals the price, as shown in Fig. 1. When control costs are not known, and a small change in price can result in large changes in control behavior, it may not be prudent to use price as a MBI when the adverse consequences of the activity can be severe. For example, polychlorinated biphenyls have such serious health impacts that it would be better to directly regulate them through quantitative restrictions or CAC regulations than to depend on pricing to ensure safe and acceptable levels. Clearly, there is a trade-off between the cost of being wrong in setting inappropriate prices and the cost of foregoing the incentive and flexibility benefits of price MBI. When the cost of being wrong can be too great, price MBI are not adequate.

2.1.4 Pricing Applications

Pricing has been used as a means of managing energy-related resource activities and environmental impacts in a variety of ways. Severance taxes and leasing payments, although they have the potential of a MBI, are typically levied as sources of revenue and payment for transfer of ownership from the public to private actors rather than a resource management instrument. Superfund taxes in the United States on inputs to chemical and refinery production, which were ended in 1995, would not qualify as MBI, as they are designed for revenue-raising and not to create incentives for behavioral changes. Fuel-use taxes are also primarily revenue-generating rather than behavior-inducing. Many U.S. states charge water effluent or air emissions permit fees, but the levels of these fees are based on the revenue needs of the regulatory programs. Such taxes and fees are not Pigovian taxes.

France introduced an industrial emissions levy on large power plants and incinerators in 1985. The charge was on sulfur oxide emissions, but the purpose was to raise revenues to finance air monitoring and subsidize installation of pollution control equipment. Japan has implemented an emissions charge for sulfur oxides also, although the charge and revenues are based on needs to compensate applicants for the health damages compensation fund. The Netherlands has implemented a manure charge system, whereby farmers pay for phosphate loadings on their land above what they are allowed to apply. Sweden has levied a charge on nitrogen oxide emissions from large furnaces. Several European countries find it more convenient to place levies on products rather than on environmental impacts, per se. For example, Norway and Sweden have imposed levies on pesticides and fertilizers. These countries and Finland have also levied carbon taxes on motor fuels, but the primary purpose is to raise revenues.

The state of Maine has a complex air-emissions permit fee system, where fees per ton of several pollutants are based on volumes, with increasing per-ton fees for larger volumes and more toxic emissions. In the South Coast Air Quality Management District of California, facilities exceeding allowable emissions must pay excess emissions fees. This district also levies fees on both criteria pollutants and air toxics, and toxics fees are considerably higher than criteria fees. Many U.S. states levy charges on the

generation and management of hazardous wastes. Taxes on landfill and incinerator disposal of hazardous waste have been high enough to reduce disposal by these methods, although it is unclear whether they have reduced the generation of such wastes or shifted disposal across state lines or to illegal disposal. The United States imposes gas guzzler taxes on automobiles, presumably to increase fuel efficiencies. The United States imposed taxes on ozone-depleting chemicals, based on their depletion factors, until such chemicals were completely phased out in the United States in 1996. Most U.S. states impose fertilizer fees, but they are based on revenues for fertilizer inspection. Texas introduced a clean fuel incentive surcharge on boiler fuel oil. Used oils and fuels derived from hazardous wastes are exempt. This surcharge is designed to shift fuel use rather than raise revenue.

2.2 Tradable Permits

The basis of tradable permits as a useful regulatory tool stems from the notion that all parties will benefit from free and voluntary trades. Markets are institutions for transfers of property rights. Ronald Coase, in his famous Coase Theorem, proposed that in the absence of transactions costs of trade, these property rights would be voluntarily redistributed so that the social value is maximized. It did not matter to whom the original property rights were assigned; the outcome would be the same and coincide with what society deemed best. By implication, if a fixed number of permits to do something were made available and trading was allowed, those permits would go to the entities most capable of profiting from them. For example, a fixed number of emissions permits, distributed across polluters, would be traded so that actors with low control costs would sell their permits to high-cost controllers; hence, low-cost controllers would do more control. The result was low social costs of attaining an emissions objective, a socially desirable goal.

2.2.1 What to Trade?

A wide variety of activities are amenable to the establishment of rights and trading. In fact, any activity that is identifiable and to which an enforceable property right can be feasibly established is ripe for tradable permitting. Emissions allowances establish a volume of permitted emissions, give an enterprise permits for that volume, and define the terms of trading. Enterprises may be given emissions reduction credits when they reduce emissions below allowable levels. They trade these reduction credits with others who will not reduce their emissions to allowable levels. Rules would be established that define the property rights implied by the permits, such as how long they are active, whether they can be banked and used later, whether some spatial trades are not allowed, and trading ratios across locations or types of activities. Tradable water rights, leaded gasoline, land development, and timber harvesting illustrate the broad range of tradable rights or permits. Some have even suggested biodiversity trading.

2.2.2 Problems with Trading

Permit and rights trading can have adverse consequences when the regulatory objective is to ensure an acceptable ambient concentration; for example, unrestricted trading of nonuniformly mixed pollutants may create "hot spots" where ambient conditions exceed objectives. This suggests trading rules in which restrictions or trading ratios are based on ambient concentration effects. A buyer of emission reduction credits may be restricted in how many they can purchase or trading among greenhouse gases may be based on CO_2 equivalents of different gases.

Since permits are required for activities, the level of permits one has will determine business viability. Actors may seek to hoard permits or employ the permitting process to enhance their market position. When tradable permitting results in market monopolization, the possible increased consumer costs must be weighed against any social gains from the tradable permit program.

A small number of trading partners limits the viability of the trading market in several ways. Transactions costs associated with finding and making trades may be higher than when the market has a large number of active participants. Also, manipulation of the market is more likely the smaller the number of traders. Any manipulation defeats the purpose of the trading and eliminates social gains. Monopolization of the trading market has the same disadvantages as monopolization of any market.

Limiting the time during which purchased permits or rights remain valid also reduces the viability of the trading market, as the property rights diminish over time. This is especially a problem when the regulator wants to reduce levels of the activity in the future. Allowing existing permits to depreciate just reduces the demand for them and reduces the potential gains from trading.

Although Coase's Theorem suggests that the initial assignment of rights or permits does not

matter to the final outcomes, this may not be true when the initial assignment affects the wealth of actors. A permit or right endows wealth. This wealth can be employed in a number of ways. Otherwise, failing enterprises may be able to obtain access to capital that allows them to endure or this wealth may be used to alter political processes in an actor's favor. These wealth effects suggest that the initial assignment does matter to the final outcome.

2.2.3 Applications

Tradable permits have been used in a variety of environmental regulatory applications. The first U.S. application was in air nonattainment areas in the early 1970s, where new sources had to obtain emissions reductions credits, or offsets, from existing sources. Those who generated reductions credits were allowed to bank them for future use. In the course of phasing out lead in gasoline, U.S. refineries were allowed to trade lead content. It is instructive that U.S. Environmental Protection Agency enforcement of trading relied on reports and random gasoline samples. There were fraudulent trades discovered, but only in less than 1% of the cases. The United States allowed chlorofluorocarbon producers to trade during the phase-out of ozone-depleting chemicals during the early 1990s.

U.S. power plants have participated since the mid-1990s in an active sulfur dioxide trading program under the Clean Air Act's Acid Rain Program. Emissions allowances are calculated based on allowable emissions per million British thermal units (Btu) times average Btu consumption. Plants with emissions less than allowables can trade with plants that exceed their allowances. The U.S. Environmental Protection Agency holds some allowances and offers them at annual auctions. This trading program has been estimated to save approximately $2.5 billion per year compared to costs without trading. Eight northeastern U.S. states also participate in a NO_x trading program.

The South Coast Air Quality Management District in southern California established a Regional Clean Air Incentives Market (RECLAIM) in the early 1990s. Emissions allowances for NO_x and SO_x were established and sources were allowed to trade emissions reduction credits, where trading ratios were set for trades between geographic areas. The district estimates that compliance costs were reduced by nearly 50% as a result of trading. Several other states, including Illinois, Michigan, New Jersey, and Texas, have trading programs similar to RECLAIM.

Trading within and between point and nonpoint sources has been applied in several cases in the United States. The state of Washington has a grass burning tradable permit system. Water effluent trading has had limited applications in the United States. Biological oxygen demand trading in the Fox River of Wisconsin had limited success because of limitations the state placed on trades and the fact that since the Clean Water Act does not explicitly allow trading, minimum technology-based standards placed on effluents limited source control flexibility. Phosphorous discharge trading in the Dillon Reservoir in Colorado has also had limited success. Trading between point (wastewater treatment plants) and nonpoint (development) was allowed. As treatment plants found cheaper means of source control, their demand for permits diminished. As a result, the market has not been active. The ability of point sources to reduce nutrient effluents inexpensively has limited the application of tradable nutrient permits in the Tar Pamlico Basin in North Carolina.

The Kyoto Protocol allows greenhouse gas trading through several provisions. The Joint Implementation provision allows Annex I countries, larger developed countries, to acquire emissions reduction units from other Annex I countries. These reduction units can be based on projects that either reduce emissions or increase sinks for greenhouse gases. The Clean Development Mechanism allows Annex I countries to invest in emissions reduction or sink projects in non-Annex I countries in order to obtain emissions reduction units. International Emissions Trading provisions then allow Annex I countries to trade with one another to meet greenhouse gas targets.

Trading programs designed to protect ecosystems and resources have also been established. Wetlands mitigation banking in the United States allows persons who engage in activities that destroy or impair wetlands to purchase credits from a wetlands bank, where persons who have created or improved wetlands place deposit credits. Rules of trading often restrict trades to geographic areas and employ trading ratios that presumably ensure no net loss of wetlands. Credits can be based on simple acreage or on quality-based measures. Transferable development rights in the United States have been used to shift residential and commercial developments to more appropriate urban areas in order to avoid urban sprawl. Landowners in protected areas are provided development rights to alternative areas in lieu of development on their property and can sell these rights to others. This market has had limited success in the United States, partially due to

municipalities' unwillingness or inability to restrict property rights to development and to the fact that effective sprawl-controlling trades may be between jurisdictions. Historic allocations of water rights in California to agriculture have created circumstances where water has not gone to its highest and best use, especially municipal and wetlands needs. The introduction of tradable permits has facilitated the movement of water rights from agriculture to these higher valued needs.

2.3 Subsidies

Subsidies are a means of rewarding actors for activities they may not have undertaken otherwise. Subsidies may be particularly useful when the actor is not legally obligated to undertake an activity or where restrictions on activities are not feasible for economic, political, or social reasons. For example, enforcement of nonpoint pollution restrictions in agriculture is costly and imperfect and farmers may be incapable, financially or politically, of controls. Pollution reduction may be achieved by subsidizing best management practices or the use of environmentally friendly fuels may be encouraged through subsidies to production or demand. Subsidies may be direct payments, such as price supports for crops used in ethanol production, or favored tax treatment, such as reductions in ethanol fuel taxes.

Although subsidies violate the "polluter pays" principle, by which social costs of activities are borne by the polluter, they may be the only means of achieving regulatory objectives. They face public resistance in cases where the undesirable activities are determined to be socially unacceptable and the public balks at paying someone for doing something they should have done anyway. Also, the perverse effect of subsidies may be to encourage the unacceptable activities in order to enhance the subsidy levels. For example, farmers may claim they would have drained more wetlands if they are subsidized for not farming wetlands.

The United States employs a variety of subsidy instruments for environmental and energy management. Favorable tax treatment in accelerated depreciation allowances for oil and gas drilling and production activities lowers energy costs. Reduced federal and state taxes on ethanol fuels, public procurement of recycled items, and favorable tax treatment of pollution control investments are further examples of subsidies impacting the demands for environmentally related goods and services. The RECLAIM program in southern California bought back old, high-polluting automobiles. An interesting, but short-lived, tax-based incentive program was the Louisiana Environmental Scorecard. Enterprises applying for industrial tax relief received exemptions based on their environmental performance, including violations, recent behavior, and Toxic Release Inventory (TRI) emissions per employee. A scoring system resulted in tax relief ranging from 0 to 100% of potential relief.

Of course, there are also many types of subsidies that harm the environment, such as energy development subsidies, below-cost timber sales, and grazing fees. Favorable tax treatment in timber, such as allowing enterprises to expense management and reforestation costs rather than capitalizing them, lowers timber costs and encourages timber development. Price support programs, such as those for sugar, have encouraged farming in sensitive ecosystems, such as the Florida Everglades. These agricultural price support programs also encourage the increased use of fertilizers.

2.4 Liabilities

The use of liabilities relies on the actor's willingness and ability to foresee the financial implications of its activities. Depending on the actor's foresight and time horizon, liabilities may be an effective means of attaining regulatory objectives. But there are many factors that may limit their effectiveness. The actor must foresee the implications of activities. There must be a high probability that adverse actions will eventually be penalized. But this depends on the effectiveness of the legal system and its enabling laws. Proofs must be made and penalties must ensue. Each element of the legal process is probabilistic. For example, can long-term health effects of a pollutant be proven? Can it be proven that a particular actor was responsible for the emissions? Will harmed parties pursue retribution? Will juries or judges convict? Will they impose punitive damages beyond proven damages? And to make liabilities an even less perfect enforcement tool, actors will use varying discount rates in assessing implications for current actions. They may accept bankruptcy as a better alternative to control.

Burdens of proof alter the liability landscape. It is one thing to have to prove that harm was done before halting an activity. It is another if actors must prove that no harm will be done before they can engage in an activity. Performance bonds may be one MBI that would hold society harmless while actors are allowed activities that may eventually prove

harmless. Actors would post them while engaging in the activity and bond values would be based on worst-case damages scenarios. Required bonding levels would be diminished as the actions are proven to be less harmful than the worst-case basis. The bonding makes the burden of proof of no harm fall on the actor. This would be appropriate when activities may have tragic consequences, such as loss of life or irreversible changes in ecosystem conditions. Bonding is a means of allowing potentially harmful activities when it is uncertain what their impacts will be. Actors have an incentive to discover their true implications.

Environmental performance bonds are used in the mining industry to ensure proper remediation of surface mines. Mining bonds in Pennsylvania have been set at values only one-tenth the average reclamation costs, leaving the state holding the financial responsibility for complete restoration. Bonding for oil and gas mining activities could similarly be applied to exploration, development, and production activities, especially in ecologically sensitive areas. Bonding of farmers for nutrient runoff, erosions control, and other adverse effects of agriculture could be used, although the ability of farmers to finance such bonds may be limited.

In the United States, the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA) applies strict liability and joint and several liability to hazardous substance releases. Strict liability means that guilty actors are liable regardless of whether the act is purposeful or a result of negligence. Joint and several liability means that any party to a release can be held liable for the entire cost, regardless of its identifiable responsibility. Damages to natural resources from oil spills are included in CERCLA and require compensation to the public for lost resources.

Personal injuries from toxic agents can be litigated under U.S. tort law. However, a successful suit requires considerable demonstration of causality by plaintiffs. The willingness of courts to accept statistical analyses is limited, making proof of causation more difficult.

2.5 Information

Information provided to the public regarding actors' activities can have dramatic repercussions on the market conditions of the actors. Boycotts related to the environment can have impacts on product demand. Community opposition can increase costs of operations, such as legal battles over zoning and property assessments, and actors may just feel some social responsibility to be good citizens for the environment. Firms may be interested in the production conditions of suppliers and better environmental records may enhance labor supply conditions. Capital markets may look favorably at firms with better environmental records, as a reflection of potential liabilities.

Information may be casual and less than perfect, based on perceptions rather than facts. But it may also be well-established and certifiable, such as appliance and vehicle energy efficiencies or toxic releases. Good examples of effective information systems in the United States are the "right-to-know" federal and state statutes. The federal Emergency Planning and Community Right-To-Know Act introduced the publicly available TRI reporting system. Over 600 toxic chemicals have been reported by larger manufacturing facilities. Generation, treatment, and disposal information on the chemicals is available to the public. The TRI program has earned wide acclaim for its effectiveness. One study has shown that firms whose stock prices fell after releasing the TRI data were more likely to reduce TRI emissions in the future. California enacted Proposition 65, which requires firms to post warnings to the public of potential exposures to listed carcinogenic or toxic chemicals. Warnings include product labels and placards in business establishments.

Product labels, such as the Green Seal program in the United States, inform consumers about the environmental friendliness of products and their inputs and production processes. Manufacturers pay the Green Seal organization for an evaluation. The U.S. Federal Trade Commission has been developing "Green Guides" for recyclable, degradable, and environmentally friendly products. Scientific Certification Systems is a U.S. for-profit business that scores timber companies for their harvesting and management practices. The company also prepares internal environmental audits for firms. Several states and nonprofit foundations sponsor environmental awards.

SEE ALSO THE FOLLOWING ARTICLES

Clean Air Markets • Economic Growth, Liberalization, and the Environment • Industrial Ecology • Market Failures in Energy Markets • Markets for Biofuels • Markets for Coal • Markets for Natural Gas • Markets for Petroleum • Modeling Energy

Markets and Climate Change Policy • Stock Markets and Energy Prices • Subsidies to Energy Industries

Further Reading

Andersen, M. S., and Sprenger, R. (eds.). (2000). "Market-Based Instruments for Environmental Management: Politics and Institutions." Edward Elgar, Cheltenham, UK.

Costanza, R., and Perrings, C. (1990). A flexible assurance bonding system for improved environmental management. *Ecol. Econ.* **2,** 57–77.

Field, B. C. (1997). "Environmental Economics: An Introduction." McGraw-Hill, New York.

Hanley, N., Shogren, J. F., and White, B. (1997). "Environmental Economics in Theory and Practice." Oxford University Press, Oxford, UK.

OECD (1993). "Taxation and the Environment: Complementary Policies." OECD, Paris, France.

O'Riordan, T. (ed.). (1997). "Ecotaxation." St. Martin's Press, New York.

Rosenzweig, R., Varilek, M., and Janssen, J. (2002). "The Emerging International Greenhouse Gas Market." Pew Center on Global Climate Change, Arlington, VA.

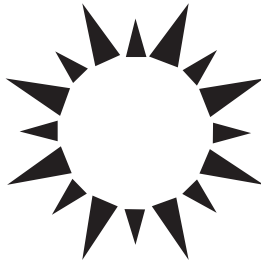
Shogren, J. F., Herriges, J. A., and Govindasamy, R. (1993). Limits to environmental bonds. *Ecol. Econ.* **8,** 109–133.

Stavins, R. N. (Ed.). (1991). "Project 88—Round II: Incentive for Action, Designing Market-Based Environmental Strategies," A Public Policy Study Sponsored by Senator Timothy E. Wirth, Colorado, and Senator John Heinz, Pennsylvania, Washington, DC.

Tietenberg, T. (2000). "Environmental and Natural Resource Economics." Addison-Wesley, New York.

Tietenberg, T., Button, K., and Nijkamp, P. (eds.). (1999). "Environmental Instruments and Institutions." Edward Elgar, Cheltenham, UK.

Market Failures in Energy Markets



STEPHEN J. DECANIO
University of California, Santa Barbara
Santa Barbara, California, United States

1. "Classical" Market Failures
2. Is Optimization the Correct Model of Behavior?
3. Additional Considerations

Glossary

consumers' surplus A measure of the benefit to consumers, net of the cost of the good, of their being able to purchase it at a particular price.
cost–benefit analysis The formal comparison of the economic costs and benefits of some particular action or policy.
general equilibrium model A model designed to represent the utility- or profit-maximizing behavior of all the agents in the economy, the solution of which is a set of prices and goods allocations that satisfy the maximization conditions of the agents.
Herfindahl index A measure of industry concentration equal to the sum of the squares of the market shares of the firms in the industry.
hurdle rate A rate of return that investment projects must achieve in order to be undertaken.
inelastic demand (or supply) A situation in which large changes in the price of a good lead to only small changes in the quantity demanded (or supplied).
marginal cost The additional per-unit cost of production as output increases.
price taker An economic agent that is not able to influence, by its own actions, the price at which any good is transacted.
producers' surplus A measure of the benefit to producers, net of the cost of production, of being able to sell their output at a particular price.
risk-adjusted cost of capital The price a firm must pay to borrow funds for an investment project of a particular risk.
traveling salesman problem Given a finite number of cities along with the cost of travel between each pair of them, find the cheapest way of visiting all the cities and returning to the starting point.

Market failure can be understood only as the opposite of some concept of market success. But what is market success? Defining it as a general improvement of the conditions of life is too vague; requiring achievement of the best of all possible worlds is too utopian. For economists, market success is ordinarily thought of as achieving Pareto optimality, a situation in which no one's material well-being can be improved without reducing the well-being of at least one other person. If the economic decisions of all the individuals in the society (that is, their activities as they participate in markets of all types) lead to a Pareto optimum, the criterion of market success has been met. Anything that falls short constitutes market failure. One of the achievements of modern economic theory has been to demonstrate that there is a set of ideal conditions under which decentralized decision-making leads to a Pareto-optimal social outcome. This remarkable result is embodied in the so-called First Welfare Theorem, which states that if every relevant good is traded in a market at publicly known prices (i.e., if there is a complete set of markets), and if households and firms act perfectly competitively (i.e., as price takers), then the market outcome without any government regulation or intervention will be Pareto-optimal. This result sets a scientifically meaningful standard—one with empirical content that is not a tautology—against which real-world situations can be compared. Policies cannot improve on the economic efficiency of a Pareto optimum, although it should be noted that many different Pareto optima having different income distributions are possible. Closely related to the notion of Pareto optimality is the concept of the Kaldor-Hicks test. According to the Kaldor-Hicks criterion (first introduced in the 1930s), situation B is socially preferable to situation A if those who would gain from moving from A to B

 © 2004 Elsevier Inc. All rights reserved.

could compensate the losers and still have something left over. This type of move will be referred to subsequently as one that is cost-beneficial, drawing on the terminology of cost–benefit analysis. Market failure also occurs if the economy exhibits opportunities for cost-beneficial actions (investments, policy implementations, etc.) that are not undertaken. It should be noted that to be cost-beneficial, it is only necessary that a policy or private action generate enough additional output to make it *possible* for the gainers to compensate the losers. In the real world, such compensation is rarely paid even when cost-beneficial moves are undertaken. This is a feature of the political system, however, that is not entailed in the definition of market success or failure.

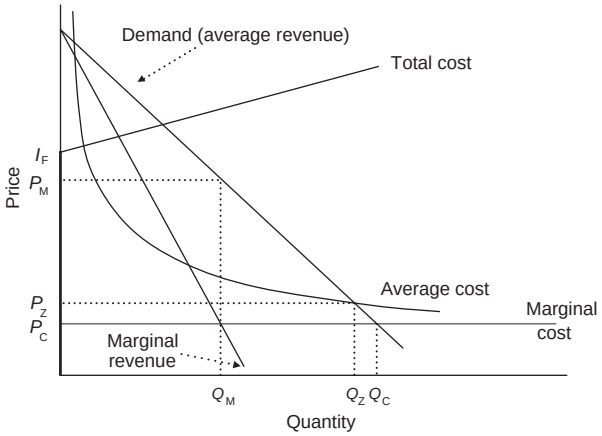


FIGURE 1 Prices and quantities under monopoly, zero profit, and perfect competition.

1. "CLASSICAL" MARKET FAILURES

In energy markets, the requirements of the First Welfare Theorem (FWT) are systematically violated, leading to a wide variety of market failures. Of course, the conditions for the FWT to hold are quite strong and are unlikely to be perfectly satisfied anywhere, but in the case of energy markets the failures are conspicuous.

1.1 Monopoly

Perhaps the most commonly recognized breakdown of the necessary conditions for market success is that energy markets are frequently subject to problems of monopoly. Some of the key agents do not act as price takers but would, in the absence of legal constraints, be able to control prices and quantities so as to improve their own position at the expense of general economic efficiency. The foremost source of such market power is the existence of economies of scale in production. Economies of scale imply that the unit cost of producing the output is declining as the scale of operations increases. If this decline in unit costs prevails over the entire range of market demand, the industry is said to constitute a natural monopoly. In energy markets, perhaps the clearest case of this is in electricity distribution, which historically has been characterized by very large fixed-cost investments in the grid.

Consider a simplified graphical representation of the electricity delivery market. Building the "backbone" grid requires a large initial investment, but once it is in place, the additional cost of connecting a new customer (that is, building a line from the grid to the customer) is small and roughly constant. In a stylized representation, this gives rise to Total Cost, Average Cost, and Marginal Cost curves, as depicted in Fig. 1. Total Cost consists of two parts: a fixed initial investment (I_F) that must be incurred before the first unit of output can be sold, plus an amount that increases linearly with the quantity of electricity delivered. This type of Total Cost curve implies an Average Cost curve that is falling over its entire range and a Marginal Cost curve that is horizontal. Demand is given by the Average Revenue curve, which slopes downward over the relevant range of potential output. The Marginal Revenue curve (defined as the additional unit of revenue per unit of additional output sold) is also a downward-sloping line with a slope steeper than the demand curve. Linearity of the demand and marginal cost curves is not essential to the argument, but simplifies Fig. 1 and the exposition.

The electricity distribution utility's economic profit (after accounting for payments to all factors of production including management and the return to capital) will be zero when average revenue equals average cost; in contrast, the perfectly competitive equilibrium will be the point at which price equals marginal cost, whereas an unregulated monopoly would maximize its profit by delivering an amount of product such that the marginal revenue equals marginal cost. The prices in these three cases would be P_M for the monopoly, P_Z for the point of zero economic profit, and P_C for the perfectly competitive equilibrium.

Market failure occurs because the unregulated monopolist charges too high a price for the electricity and delivers too little of it. If the fixed costs are high enough, no potential competitor could enter the market and make a profit. Even if the entrant's fixed

cost were low, no entrant would be able to make any positive profit so long as the entrant's average cost for whatever market share it could obtain were greater than the price prevailing in the market after the newcomer's entry. The monopolist collects "supernormal" profits, that is, profits over and above all the costs of production. Elimination of these monopoly profits requires that the price be no higher than average cost and the minimum output for which this is possible is (Q_Z, P_Z). This zero-profit point is often the target aimed for by utility regulators—the price the utility is allowed to charge is only enough to cover all costs including a "fair" return on capital.

Yet even at the zero-profit point (Q_Z, P_Z), the market failure is not entirely corrected. Although it is true that at this point there is no incentive for entry into the market and the existing firm is economically viable, the zero-profit point is not Pareto-optimal because it is possible to design taxes on consumers and subsidies to the utility that would induce the utility to expand its output (and lower its price) to (Q_C, P_C). At (Q_C, P_C) every member of society is potentially better off than at (Q_Z, P_Z). This can be seen by comparing the consumers' gain in moving from (Q_Z, P_Z) to (Q_C, P_C) to the subsidy that would be required to bring the utility back to the break-even point. Alternatively, it can be shown that the total surplus (producers' plus consumers') increases in moving from (Q_Z, P_Z) to (Q_C, P_C). Society gains on net in moving to (Q_C, P_C).

The example illustrated by Fig. 1 is a simple case in which economies of scale are the cause of the monopoly problem. In the real world, monopoly power is often more difficult to detect and may be very hard to control. Markets are rarely occupied by a monopolist producer, and although measures of concentration (such as the Herfindahl index) based on market shares are often used to provide a rough indication of potential market power, the absence of a general theory of oligopoly pricing makes unambiguous inferences of market power from data on market shares impossible. To complicate matters further, temporary market power can emerge in almost any dynamic setting. It often involves a time-sensitive component. Entering a market can take time—to build parallel infrastructure, establish a sales network, or obtain the required regulatory permissions, for example. Thus, it may be possible for a producer to enjoy a monopoly position in the short run that cannot be maintained over the long term. Much actual business practice involves seeking temporary monopoly advantages—being the first to offer a new product or service or developing a new process or technique that is protected by patent rights or as a trade secret. In such cases, the monopoly return may be viewed as a way to compensate firms for the expense of research and development and provides an incentive to innovate. The kind of temporary monopoly advantage that leads to technological progress can be an indicator of dynamism in the market economy. In the perfectly competitive equilibrium in which no one is earning any supernormal profits, there is no incentive for anyone's behavior to change. In the real world, the search for profit opportunities must have some positive probability of a payoff, or innovation would wither.

In many other circumstances, however, monopoly power not based on economies of scale has no social benefit. Taking advantage of temporary shortages, or collusion to create a shortage, is rightly condemned as a market failure. Energy markets are prone to problems of this type because of complex technical interdependencies or inelastic supply and demand curves (as in electricity markets). Controlling the opportunism that such conditions engender is one of the reasons firms exist and is also a legitimate role for government intervention. The debacle of "deregulated" electricity markets in California in 2000–2001 is an example of the kinds of things that can go wrong when such opportunities for the exercise of market power are not held in check. When electricity supply became tight in the summer of 2000, even producers with relatively small market shares could exercise market power because of the extreme inelasticity of short-term demand.

1.2 Externalities

A second source of market failure in energy markets stems from the existence of externalities. A negative externality arises when production or consumption decisions of some agents directly and adversely affect the happiness of other agents and the affected agents have no way of obtaining compensation for the harm done to them. (A similar inefficiency occurs if the externality is positive in that it provides benefits to some members of society that they do not have to pay for.) In the case of energy production and consumption, externalities are the rule rather than the exception. Examples include the local air pollution of coal-fired power plants (which produce harmful particulates and oxides of nitrogen and sulfur as a natural result of the combustion process), the congestion of streets and highways that results when individuals make uncoordinated personal

transportation decisions, and the climate-changing effect of the carbon dioxide that is emitted whenever fossil fuels are burned.

Essentially, the market failure from externalities comes about because some markets simply do not exist. There is no way that individuals can buy or trade "climate stability," for example. At the root of the externality is the absence of well-defined and enforceable property rights. Clear property rights of a suitable type would enable individuals to obtain compensation for the adverse effect on their well-being of the externality-producing activities of others. When such rights do not exist, people have no way of expressing their preferences through their willingness to exchange goods they own. In the case of monopoly, market failure occurs because not all participants in the market are price takers; with externalities, markets fail because they are not complete. The First Welfare Theorem does not hold because not every good is traded.

A number of strategies can eliminate the market failures caused by externalities, but all involve some kind of policy decision or collective action. Taxes (or subsidies in the case of positive externalities) can be imposed that make the private costs associated with the externality-producing activities (such as the burning of fossil fuels) correspond to what producers would have to pay those harmed by the externality if the appropriate property rights existed. Alternatively, the government can create new property rights and assign them to individuals so that the requisite market transactions can occur with low transactions costs. This enabling of new markets is exemplified by the creation of SO_2 allowances under the 1990 Clean Air Act (CAA) Amendments in the United States. Under the CAA Amendments, SO_2 emissions are curtailed in the aggregate (thereby creating a cost for SO_2 emissions and correcting the externality) and trading of allowances among producers makes it possible for the net aggregate reduction in SO_2 emissions to be achieved a minimum cost. Provided the participants in the newly created markets behave as price takers and not as monopolists or oligopolists, this kind of policy can eliminate the externality that arises from nonexistent markets.

It should be noted here that the distribution of wealth and income will be affected by how the property rights are assigned or tax revenues are collected and spent. Different assignments of rights will give rise to different Pareto-optimal equilibria and those different outcomes need to be judged according to a criterion of justice rather than economic efficiency (because all Pareto-optimal outcomes are efficient). According to the definition being used here, market failure pertains to the failure to achieve an efficient outcome, not to the intrinsic desirability or fairness of the resulting pattern of prices, incomes, and consumption.

1.3 Public Goods

A third source of market failure in energy markets is that energy production and consumption entail the production or diminishment of public goods such as macroeconomic stability or the scientific knowledge derived from basic research. A public good is a commodity or service such that if it is supplied to any of the members of society, it can be enjoyed by all at no extra cost. The consumers of the public good are not able to exclude others from its consumption. The classic examples are national defense and public order. A military establishment that is successfully able to deter potential aggressors and defend the nation's territory against attack necessarily provides this service for all citizens alike. This "nonexcludability" of public goods gives rise to the problem of free riders—if national defense protects citizens equally regardless of their contribution, individuals have no incentive to pay for the military. If left to the unregulated market, not enough public goods will be supplied because of this free rider problem. This is the fundamental reason why taxes that are used to purchase public goods must be compulsory.

It might be thought that fuels and energy services are commodities like any other. In their direct benefits this is true: the owner of an automobile can exclude others from driving it and a consumer's electricity bill reflects only his or her own consumption of electricity. However, because the use of energy is ubiquitous, and because the means of converting fuels into useful work requires complex and expensive capital equipment (power plants, motor vehicles, pipelines, etc.), the commodity-like nature of energy products gives rise to the problem of macroeconomic stability. Fuel prices can change much more rapidly than the capital stock can be adjusted and therefore the volatility of fuel prices creates situations in which capital that was profitable at previous fuel price levels becomes uneconomic at current fuel price levels. Uneconomic capital equipment is taken out of production and this in turn leads to job losses and unemployment. Thus, movements in energy prices can have a ripple effect throughout the economy, resulting in fluctuations in employment and output that affect the well-being of all. This is precisely what happened during the oil price shocks

of the 1970s. Unanticipated sudden increases in oil prices associated with market and political events had macroeconomic consequences around the world. Policy initiatives such as the Strategic Petroleum Reserve (SPR) are intended to stabilize the price of oil and thereby ameliorate macroeconomic disruptions caused by oil price volatility. The SPR provides a valuable buffer against the potential macroeconomic consequences of unexpected oil price movements just as discretionary monetary policy is an appropriate tool for responding to changes in aggregate demand.

Macroeconomic vulnerability to energy price shocks is sometimes characterized as a national security problem. Given that modern military vehicles require sophisticated power plants (whether fossil-fuel driven or nuclear), it is certainly true that the availability of suitable fuels is necessary for military operations. For some countries, such as the United States, domestic fuel sources are sufficient to satisfy military requirements in wartime so direct military needs are not part of the national security problem for those countries. Also, the relative proportions of foreign and domestic oil consumed are no indication of macroeconomic vulnerability. Because oil is a commodity traded in world markets, a supply disruption anywhere has an effect on the world oil price. So, for example, even if U.S. oil demand were supplied 100% domestically, a world oil price shock would have adverse macroeconomic effects in the United States as the domestic price followed the world price. Thus, the national security problem associated with energy has to do primarily with the flexibility of a nation's capital stock in the face of energy price fluctuations. In addition, foreign oil purchases may constitute a national security problem if the supply comes from a hostile state or states whose military capability is increased because of the wealth gained from their petroleum sales.

A second public goods aspect of energy markets is the role of research and development (R&D). Scientific knowledge is very much a public good, because once new knowledge is published it is available to all without the possibility of exclusion. The producer of the knowledge cannot hope to recover more than a fraction of the benefits resulting from the knowledge. Useful knowledge created through the research and development process has spillovers to firms or other organizations and individuals that did not carry out the research. As a result, there will be too little research and development in the aggregate unless the government intervenes. If the optimal level of research were being undertaken, the rate of return to R&D would be equal to the rate of return on other forms of investment. In fact, empirical studies consistently find social rates of return to R&D that substantially exceed the market return to capital. This discrepancy provides a rationale for government support of R&D, whether through universities, government research laboratories, tax incentives, or direct subsidies.

A related potential market failure comes about because of the tendency for unit costs to decline as producers and consumers gain experience with a product or technology—the phenomenon of "learning by doing." This dynamic pattern of cost decline has been observed across many technologies, including energy technologies such as wind power generators and solar voltaic panels. Unless private investors correctly anticipate the timing and extent of the cost reductions due to learning by doing, they may fail to invest sufficiently in promising emerging technologies. Consumers may be unwilling to pay the high price for new technologies in the early stages of their development, even though the technologies would be quite attractive if they could be utilized in sufficient volume to realize cost reductions from learning by doing and economies of scale. This is sometimes referred to as the "chicken and egg" problem. In some cases, market forces will lead to the emergence of the best technologies, but not always. A long history of the gradual improvement of an inferior technology (along with the growth of infrastructure associated with it) may lead to the lock-in of that technology and barriers to the adoption of superior alternatives.

2. IS OPTIMIZATION THE CORRECT MODEL OF BEHAVIOR?

The market failures described in Section 1 are commonplace staples of textbooks. There is another class of energy market failures about which there is less consensus. Prominent among these is the energy efficiency gap: the notion that even though sizable improvements in energy efficiency could be achieved by making investments that would be profitable at current prices, these investments are not carried out. Sometimes this phenomenon takes the form that the hurdle rates required for energy-efficiency investments are considerably higher than the cost of capital to the firm for projects of comparable risk. The existence of this type of market failure is controversial, because if profitable opportunities are available,

why do profit-maximizing firms fail to undertake them? (By definition, a profitable investment is cost-beneficial because the resources devoted to it create more wealth than if the investment were not made.) A very large body of empirical literature attests to the existence of these opportunities, yet the systematic failure to maximize profits conflicts with standard practice in economic modeling. It would seem that either the assumption of profit maximization does not hold or the studies documenting the possibility for energy efficiency improvements have missed essential elements of full cost accounting.

To see why the resolution of this issue has implications for market success or failure, consider the two situations depicted in Fig. 2. In both Fig. 2a and Fig. 2b, the axes represent desirable outputs from different activities that might be undertaken by the firm. The horizontal axis shows the reduction in pollution that would follow from enhancing energy efficiency (by suitable investments, changes in operating procedures, etc.) and the vertical axis shows the economic output from the other activities of the firm. The curves PP′ in the two diagrams represent the theoretical "production-possibilities frontier" of the firm. This frontier is the set of possible combinations of the two kinds of goods that the firm could produce from a given set of resources if it were fully optimized.

The empirical question is whether the firm is always operating on its production-possibilities frontier. Figures 2a and 2b represent alternate answers to this question. In both cases, the points A and A′ represent the initial situation of the firm and points B and B′ show the position of the firm after taking some action such as increasing its energy efficiency. If the firm is fully optimized, it is not possible to achieve reductions in pollution without sacrificing some other profitable activity. This case is shown in Fig. 2a. In contrast, Fig. 2b shows a situation in which it is possible for the firm to realize gains simultaneously from investments in energy efficiency and other activities. This case corresponds to the possibility that there are profitable energy-saving investments available to the firm. The increase in energy efficiency accounts for the pollution reduction along the horizontal axis and the additional profit earned from these investments allows the firm to expand its output in other directions. If the firm is not organized optimally to begin with, improvements are possible along multiple dimensions of its performance.

The situation displayed in Fig. 2b is commonly referred to as a market failure, even though it is perhaps more appropriate to describe it as an organizational failure or institutional failure. The point remains that in Fig. 2b it is possible to take actions that would be cost-beneficial. To assess how common or realistic is the situation shown in Fig. 2b compared to Fig. 2a, it is necessary to examine what kinds of factors might give rise to a firm's operating inside its production-possibilities frontier.

2.1 Nonunitary Organizations

Primary among the causes of organizational failure is the fact that firms are made up of distinct individuals having different interests. The firm has no mind or will of its own; it acts according to a set of decision-making procedures and there is no guarantee that the collective action of the individuals making up the firm will lead to fulfillment of the firm's formal objective (the maximization of profit). Even though

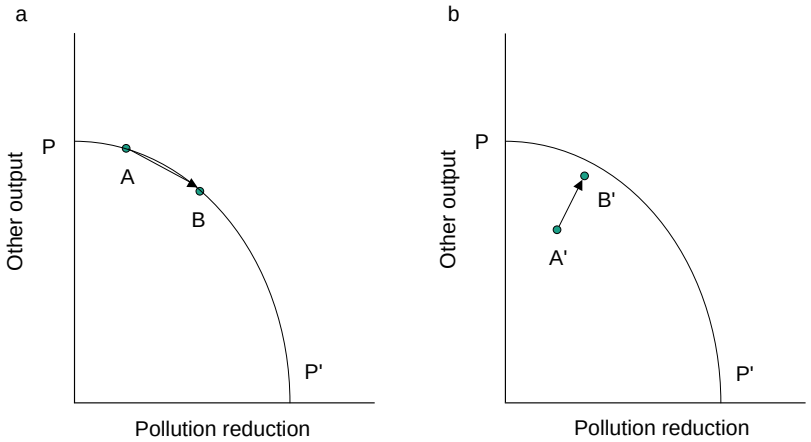


FIGURE 2 An energy efficiency gap? (a) Firm that is fully optimized. (b) Firm that is not fully optimized.

there are gains from cooperation (firms would not exist otherwise), the diverging objectives of the agents making up the firm are almost certain to result in a failure of complete optimization. The type of conflict arising from the nonunitary nature of firms and other productive organizations has historically been referred to as the principal–agent problem. The principal (in the case of firm, this would be the shareholders whose interest is that the firm maximize profits) cannot guarantee that the agents (the managers and employees of the firm) will act to further the principal's objective rather than improving the conditions of the workplace, creating perquisites of employment, or otherwise diverting profits into additional benefits for the agents. Because profit maximization by firms is required for Pareto optimality in a market economy, any deviation from that standard caused by managers' and employees' self-interested behavior is a cause of market failure.

The various manifestations of the principal–agent problem have been developed by scholars specializing in the theory of the firm. Adolph Berle and Gardiner Means in the 1930s identified how the separation of ownership and control in the modern corporation enables managers to advance their own interests at the expense of shareholders, but an awareness of the general problem can be traced back to Adam Smith and earlier. Managers typically have access to detailed inside information not available to the principals. To prevent managers from using this information to their benefit, principals sometimes have recourse to operating procedures that are second-best in terms of profit maximization, but that control the principal–agent conflict. This explanation has been proposed for the excessively high hurdle rates required for approval of investment projects in some firms. Principals cannot know the actual rates of return on all potential projects. If managers are interested in maximizing "organizational slack" (which includes managerial perquisites and inessential expenditures), the second-best solution for the principals involves setting a hurdle rate greater than the risk-adjusted cost of capital. The high hurdle rate serves as a substitute for expensive postproject auditing. The benefit to the principals comes at the cost of foregoing some profitable projects (for example, low-risk energy-saving investments in lighting retrofits or heating, ventilation, and air conditioning upgrades) having rates of return that are lower than the hurdle rate.

Managers also can exhibit myopia, risk aversion, and bias in project selection that make the managers' lives easier at the expense of the profit maximization objective of the owners of the firm. It is impossible for the owners to determine from the outside just what contribution each employee of the firm makes to its success; hence, the market rates of compensation for the agents cannot be known. Many of the incentive compensation plans that are observed in firms (stock options, performance bonuses, etc.) are compromises intended (on the part of the principals, at least) to bring the interests of owners and their agents into closer alignment. All of these sources of deviation from perfect profit maximization apply to the general operations of the firm, including the firm's energy use decisions.

2.2 Bounded Rationality and Agent-Based Models

The principal–agent problem undermines full optimization when all agents are entirely rational in the strong sense required by neoclassical economic theory. That is, the agents are assumed to have unlimited capacity to carry out whatever calculations are required to maximize their own utilities. This assumption of unlimited computational capacity is questionable, however. The alternative notion of "bounded rationality," first proposed by Herbert Simon, is based on the realization that the calculations necessary to achieve full optimization in economic activities are effectively beyond the reach of real people and firms.

Since the time of Simon's initial work, advances in mathematics and computer science have formalized the concept of computational complexity. It is widely believed (but not yet proven) that a wide class of practical problems such as the Traveling Salesman Problem require computation time that grows "faster than polynomially" with the size of the problem. John Rust and others have pointed out that many calculations in economics are subject to formal computational complexity constraints, and Michael Garey and David Johnson's long list of intractable problems (of the same class as the Traveling Salesman Problem) includes many standard management problems such as network design, storage and retrieval, database management, and sequencing and scheduling. Achievement of full maximization (including its energy efficiency component) involves calculations of this type in many cases. It is an empirical question how large the deviation from optimality might be in any particular instance, but the existence of

computational limits constitutes a formidable barrier to complete optimization by firms.

Aside from these mathematical foundations of bounded rationality, as a practical matter limitations on management time and attention mean that cost-saving energy efficiency investments not central to firms' strategic priorities are often given short shrift. Even this commonsense observation indicates a departure from Pareto optimality, because optimally efficient firms would take advantage of every profitable opportunity. Bounded rationality provides one justification for policies that simplify business decision-making by setting standards. In addition, the success of voluntary pollution prevention programs in the United States and other Organization for Economic Cooperation and Development countries testifies to the possibility of policy-driven cost-beneficial improvements if the government provides high-quality information and support services to boundedly rational firms. Michael E. Porter has even proposed the bold hypothesis that environmental regulation may, in some circumstances, stimulate firms to better performance than they would have achieved in the absence of regulation.

Some economists have begun to embody the concept of bounded rationality in models in which the entities (or agents) follow relatively simple rules of behavior, rather than being presumed to maximize an objective function. Because the rules are simple, it is possible to set up numerical simulations of systems containing large numbers of such agents and observe the development of such systems over time in a laboratory-like setting. Models of this type have the potential for reproducing features of the population of firms, such as the size distribution or the birth and death rates of companies, that do not emerge naturally in conventional maximization models. If the performance of the firms in agent-based models is compared to the ideal of full optimization, room for improvement will be seen. If the economic system is in fact made up of agents following relatively simple rules, then cost-beneficial gains are possible.

2.3 Behavioral Economics

Both principal–agent problems and bounded rationality take for granted that at least the agents in the economy are trying to maximize utility as it is conceived by neoclassical economic theory. Both within and outside of economics, however, there is recognition that people do not necessarily behave that way. Psychologists have noted that decisions about energy-saving investments frequently do not conform to the strict requirements of the economic paradigm. "Organizational Behavior" and "Management Science" are recognized as fields of inquiry distinct from economics. Indeed, it would be difficult to understand the considerable demand for management training and expertise (Master of Business Administration and executive education programs, for example) were it not for the fact that management activities, such as motivating employees, coordinating complex operations, and visualizing the uncertain future, require skills and talents of the highest order. The intricacies and subtleties of human and organizational behavior are far from being reducible to a simple economic calculus.

Some economists have begun to explore the consequences of behavior outside the conventional maximizing framework. A Special Section in the March 2000 edition of the *Journal of Economic Behavior and Organization* was devoted to "Psychological Aspects of Economic Behavior" and a session of the year 2000 annual meeting of the American Economic Association included papers titled "Emotions in Economic Theory and Economic Behavior" and "Thinking and Feeling." If irrational exuberance can account for stock market bubbles measured in trillions of dollars, it is perhaps naive to think that all decisions regarding energy technologies and investments are explicable as purely rational investment decisions. Market pressures surely push firms and individuals in the direction of rational choice, but social forces that individuals are barely conscious of, such as the imperatives of maintaining prestige and self-image and even considerations of style and fashion, are also powerful influences on behavior.

2.4 Evolutionary Economics

The inadequacy of the neoclassical model of the firm ("maximization of profits subject to the constraint of current technology and resources") has led some economists and other social scientists to explore a variety of alternative characterizations of production. Among the most promising is evolutionary economics—the idea that firms and the economy can be described and analyzed in terms analogous to biological populations and ecosystems. In this view, the market acts in the same way as natural selection in evolutionary biology to shape the characteristics of the population of firms. If a firm is not sufficiently fit (i.e., profitable), it dies (goes bankrupt or is acquired by a healthier firm), whereas the more profitable firms expand and propagate.

What are the implications of the evolutionary model for market failure as it has been defined here? First, it is important to realize that the biological analogy does not carry over perfectly. In biological evolution, species evolve over time through mutation and natural selection, whereas in the economy, firms are capable of consciously changing or adapting to new circumstances. Humans can learn, so the "inheritance of acquired characteristics" is to be expected in economic evolution, unlike in biology. Despite this difference, the crucial similarity between economic and biological evolution is that in both the evolutionary pressures of natural selection lead to improvements in performance over time. However, it is a common mistake to confuse market pressures that reward profitability with a mechanism that somehow leads to the *maximization* of profits (or full optimization by firms). According to such a view, market selection results in firms acting "as if" they maximize profits, regardless of what the managers think they are doing or limitations on their access to information or computational capabilities. It was pointed out by Armen Alchian in 1950 that though market competition favors firms that realize positive profits, this by no means implies that surviving firms maximize profits.

In neither economics nor biology can the evolutionary process be expected to achieve optimization of the fitness of all members of the population. Evolution takes existing populations as its raw material and works through incremental change. The consequence is that evolutionary processes are path dependent. Small differences in initial conditions can lead to different paths of technological progress or institutional development or, alternatively, to different ways of accomplishing the same function through different institutional arrangements or technologies. There is no reason to think that the particular path "selected" by the evolutionary process arrives at an optimal solution. Similarly, in the evolutionary context, optimality can be defined only in terms of the environment in which the population's members operate. If that environment is changing rapidly, the evolutionary process will lead to improvements that never entirely catch up to the changing conditions. In addition, some degree of diversity in the population (of organisms or firms) is likely to provide resilience against sudden shifts in the environment. This indicates that not all firms will be equally adept at dealing with any particular environment that exists at a moment in time. Finally, it is a fallacy to think that the population in existence today is optimally evolved, for if that were so, why was it not true also of the population that existed a year ago, a decade ago, or a century ago?

These departures from complete optimization apply to the behavior of firms and economic agents in general and are not specific to the energy sector or to energy markets. Nevertheless, the ubiquity of departures from optimality certainly indicates that market failure in energy markets (and in the energy decisions made by firms) will also be widespread. There is no reason to think that incomplete optimization is more severe with respect to energy decisions than it is with respect to other important decisions taken by firms, say with regard to employee benefit packages, advertising, or organizational structure. Nevertheless, the energy sector and the energy decisions of firms have been scrutinized extensively in an effort to discover empirical examples of failures to optimize. This is partly an historical accident because of the intense interest in energy issues that followed the oil price shocks of the 1970s. From a scientific standpoint, this focus on energy has been highly productive, for it led to the large literature documenting the energy efficiency gap and the noneconomic factors that influence energy decision-making. This empirical literature does not support the presumption of complete optimization.

3. ADDITIONAL CONSIDERATIONS

The discussion thus far has been couched almost entirely in economic terms. As such, it has given no consideration to the very real possibility that markets can fail even if Pareto optimality is achieved. A perfectly functioning market system may be a failure from a higher social or philosophical perspective. Ordinary economic theory ignores the question of where individuals' preferences come from and how they are shaped by the social context. In the same way, economic analysis typically takes the pattern of ownership of property rights as given, even though different assignments of those rights (and creation by the State of rights distinct from the set currently in existence) can have a major impact on the justice of social arrangements. Even if energy markets were not subject to market failure as outlined above, it might still be the case that the prices and consumption patterns emerging from them would not appropriately reflect the interests of future generations or that they might violate standards of fairness that transcend economics.

In addition, there is no guarantee even given a particular set of endowments of property rights that

the Pareto-optimal competitive equilibrium will be *unique*. Without any problems of monopoly, externality, etc., it is easy to set up simple models that exhibit multiple equilibria. Each of these equilibria is a Pareto optimum, but they can be quite different in their equilibrium prices, allocations, and distributions of income. There is no consensus on how society might go about choosing among such equilibria from a welfare perspective or about what dynamic processes might lead an economy to one or another of the equilibrium points (or from one to another if a particular equilibrium is disturbed). This lack of a theory of equilibrium choice is a gap in economic knowledge that would need to be filled before too much reliance is placed on Pareto optimality as a goal of policy.

It should also be noted that the Pareto criterion for market success makes operational sense only if individuals' utilities are not interdependent. If some individuals' utilities depend not only on their own consumption of goods but also on the consumption of others, then the First Welfare Theorem no longer holds. Striving after positional goods (that derive at least part of their value as a result of comparison to others rather than because of the goods' intrinsic characteristics) also leads to inefficiency. If organizations of one type (such as economics departments or universities) all seek to increase their academic rankings, a large part of their effort is bound to be wasted because not all of them can succeed in moving up. Market failures of this type are no less important than more conventional kinds of deviations from Pareto optimality nor should they be ignored in formulating policy.

At the same time, the ubiquity of market failures in the energy sector does not mean that markets should cavalierly be replaced by political decisions or central planning. Even where markets fail, there is no guarantee that their replacement by some other allocative process would perform better. The political process is also vulnerable to failure, and sometimes to more spectacular failure, than markets. Political systems may restrict competition, inhibit the generation and transmission of accurate information, and replace satisfaction of the wants and needs of the general population with implementation of the preferences of the decision-making elite. It is unrealistic to substitute the assumption that the government always acts in the best interests of society as a whole for the assumption that markets always conform to the requirements of the FWT. Neither assumption is justified. Nonetheless, an awareness of the possibilities of market failure should caution against embracing the market fundamentalism that rejects every attempt to make improvements in the functioning of the economic system and dismisses all forms of regulation as unwarranted government intrusion on the free functioning of the Invisible Hand. In energy markets, the balance between decentralization and regulation is a delicate one, but that only makes it more important to maintain.

SEE ALSO THE FOLLOWING ARTICLES

Economic Growth, Liberalization, and the Environment • Economics of Energy Demand • Economics of Energy Supply • Energy Efficiency, Taxonomic Overview • Market-Based Instruments, Overview • Markets for Biofuels • Markets for Coal • Markets for Natural Gas • Markets for Petroleum • OPEC Market Behavior, 1973–2003 • Stock Markets and Energy Prices

Further Reading

Antle, R., and Eppen, G. D. (1985). Capital rationing and organizational slack in capital budgeting. *Manage. Sci.* **31,** 163–174.

Borenstein, S. (2002). The trouble with electricity markets: Understanding California's restructuring disaster. *J. Econ. Perspect.* **16,** 191–211.

Brekke, K. A., and Howarth, R. B. (2002). "Status, Growth and the Environment." Edward Elgar, Cheltenham, UK.

Conlisk, J. (1996). Why bounded rationality? *J. Econ. Lit.* **34,** 669–700.

DeCanio, S. J. (1993). Barriers within firms to energy-efficient investments. *Energy Policy* **21,** 906–914.

DeCanio, S. J. (2003). "Economic Models of Climate Change: A Critique." Palgrave-Macmillan, Houndmills, UK.

Dennis, M. L., Soderstrom, E. J., Koncinski, W. S. Jr, and Cavanaugh, B. (1990). Effective dissemination of energy-related information. *Am. Psychologist* **45,** 1109–1117.

Epstein, J. M., and Axtell, R. L. (1996). "Growing Artificial Societies: Social Science from the Bottom Up." M.I.T. Press, Cambridge MA.

Howarth, R. B., and Sanstad, A. H. (1994). Normal markets, market imperfections and energy efficiency. *Energy Policy* **22,** 811–818.

Interlaboratory Working Group. (2000). "Scenarios for a Clean Energy Future." Interlaboratory Working Group on Energy-Efficient and Clean-Energy Technologies, Oak Ridge National Laboratory and Lawrence Berkeley National Laboratory. ORNL/CON-476 and LBNL-44029, Oak Ridge, TN and Berkeley, CA.

Kehoe, T. J. (1998). Uniqueness and stability. *In* "Elements of General Equilibrium Analysis" (A. Kirman, Ed.), pp. 38–87. Blackwell, Oxford, UK.

Laitner, J. A., DeCanio, S. J., and Peters, I. (2001). Incorporating behavioural, social, and organizational phenomena in the assessment of climate change mitigation options. *In* "Society, Behaviour, and Climate Change Mitigation" (E. Jochem,

J. Sathaye, and D. Bouille, Eds.), pp. 1–64. Kluwer Academic, Dordrecht, The Netherlands.
Loewenstein, G. (2000). Emotions in economic theory and economic behavior. *Am. Econ. Rev. Papers Proc.* **90,** 426–432.
Mas-Colell, A., Whinston, M. D., and Green, J. (1995). "Microeconomic Theory." Oxford University Press, New York.
Nelson, R. R., and Winter, S. G. (1982). "An Evolutionary Theory of Economic Change." Belknap Press of Harvard University Press, Cambridge MA.
Nelson, R. R. (1995). Recent evolutionary theorizing about economic change. *J. Econ. Lit.* **33,** 48–90.
Porter, M. E., and van der Linde, C. (1995). Breaking the stalemate *Harvard Bus. Rev.* **73,** 120–134.
Romer, P. M. (2000). Thinking and feeling. *Am. Econ. Rev. Papers Proc.* **90,** 439–443.
Rust, J. (1997). "Dealing with the Complexity of Economic Calculations." Economics Working Paper Archive at Washington University, St. Louis. Available at http://econwpa.wustl.edu:8089/eps/comp/papers/9610/9610002.pdf (accessed 5/26/03).
Sanstad, A. H., DeCanio, S. J., Boyd, G. A., and Koomey, J. G. (2001). Estimating bounds on the economy-wide effects of the CEF policy scenarios. *Energy Policy* **29,** 1299–1311.

Markets for Biofuels

BENGT HILLRING
Swedish University of Agricultural Sciences
Uppsala, Sweden

1. Introduction
2. Experiences from Sweden
3. Experiences from Finland
4. Experiences from Denmark
5. Experiences from Austria
6. European Experiences
7. Conclusion

Glossary

bioenergy Super ordinate to the terms "biofuel" and "wood fuel."
biofuels Fuels produced directly or indirectly from biomass.
biomass Material of biological origin, excluding material embedded in geological formations and transformed to fossil.
forest fuel Wood fuel that has not been used previously.
upgraded wood fuels Wood fuel briquette with a diameter equal to or less than 25 mm; wood fuel pellet with a diameter less than 25 mm; wood fuel powder with a main fraction less than 1 mm.
wood fuel All types of biofuels originating directly or indirectly from woody biomass.
woody biomass Biomass from trees, bushes, and shrubs.

The use of biofuels is widespread over various countries and regions all over the world. However, the reasons for use of biofuels are different. Historically, it has been a very important resource for what we today call the industrialized countries. There are several examples of successful industrial use in mining industries, ore industries, and the like. For example, charcoal still is used in ironworking industries in Brazil. Biofuels may be used for cooking, heating, industrial applications such as steam production, electricity production, and fuel for vehicles as well as in the ironworking industry. Modern use of biofuels, as described in this article, includes industrial production in district heating networks and production of electricity. This development has increased trade of biofuels and in some European regions has created markets for various qualities of biofuels.

1. INTRODUCTION

The use of bioenergy is a strategic resource used in the work to fulfill the Kyoto Protocol to replace fossil fuels on a large scale. Bioenergy and biomass generally could be used to mitigate greenhouse gas emissions and the global climate change. In some countries, the use of biofuels already holds a significant share of the energy supply to industrial processes, heat production, and (in some cases) electricity production. The Nordic countries are examples of this. Large-scale production and use of biofuels has to be done in an environmentally sound way. New technology and system solutions give biofuels possibilities to compete in new markets.

Traditionally, biofuels have been used in the same geographical regions in which they were produced. During more recent years, this pattern has been changed in Northern Europe by industrial and large-scale use of various forms of biomass for district heating such as a vast supply of recycled wood and forest residues. The trade situation has come about as a result of means of control on waste and energy. Sea shipments allow bulk transports of biofuels over long distances at low cost.

Climate change due to anthropogenic emissions of greenhouse gases is one of the greatest environmental challenges today, and greenhouse gas concentrations in the atmosphere will continue to rise unless there are major reductions in greenhouse gas emissions. Carbon dioxide (CO_2) is the most important greenhouse gas, and increasing the use of biomass for energy is an important option for reducing CO_2 emissions.

 © 2004 Elsevier Inc. All rights reserved.

The European Union (EU) has no common energy policy, but there are different common goals mainly given in the European Commission's (EC) "White Paper for a Community Strategy," which sets out a strategy to double the share of renewable energy in gross domestic energy use in the EU by 2010 (from the recent 6–12%, with some 85% of the renewables being bioenergy). The strategy includes a timetable of actions to achieve this objective in the form of double the share of renewable energy in gross domestic energy use in the EU by 2010, including a timetable of actions to achieve this objective in the form of an action plan. The EC's green paper in November 2000, "Towards a European Strategy for the Security of Energy Supply," introduced the objective of substituting 20% of traditional fuels by alternative fuels in the road transport sector by 2020. A proposal for a directive "on the promotion of the use of biofuels for transport," adopted by the EC in November 2001, requires that an increasing proportion of all diesel and gasoline sold in the member states be biofuels, starting with 2% in 2005 and increasing progressively so as to reach a minimum of 5.75% of fuels sold in 2010.

The EU member countries are jointly decreasing the emissions of six different greenhouse gases with by 8% on average during the period starting in 1990 and ending in 2008–2012 (depending on type of action and type of gas). An example of this policy is the EU-funded research programs, where the focus is on nonfossil fuels and nonfossil techniques to target the goal of reducing emissions of greenhouse gases.

The Scandinavian countries and Austria have been the pioneering countries in the modern use of bioenergy. Other countries, such as Germany, the United Kingdom, The Netherlands, and Spain, are expected to increase their use of biofuels significantly. There are variable forest resources and variable uses of biofuels in Europe, and these are the key factors behind the trade that has grown rather quickly during the past years. Because of the incentives to increase the use of biofuels, so the prediction for the short and medium term is an increase in trade biofuels in Europe.

2. EXPERIENCES FROM SWEDEN

Sweden is one of the leading European countries in use of biomass based mainly on the large forest resources. Biofuels are used in the forest products industry, in the district heating sector, and for traditional heating mainly of houses in the countryside. Industrial use is 184 PJ per year of mainly wood fuel and black liquor, with the majority of this being internal use. The residential sector uses 38 PJ per year, also mainly internal use, on farms or self-supply among house owners.

Swedish experience of a commercial biofuel market is based on the development of the district heating sector over a period of at least 20 years, from a low level during the 1970s to a substantial market reaching 96 PJ in 2000. Wood fuel counts for 55% of the biofuel supplied to the district heating sector, reaching 54 PJ in 2000. The average growth rate of supplied wood fuel between 1990 and 2000 was approximately 10% per year. Wood fuels compete on the market with other untaxed biofuels. The highly taxed fossil fuels are not competitively priced for heat production.

Prices for biofuels, peat, and the like are reported by the Swedish National Energy Administration.

2.1 Imports

During the early 1990s, the import of biofuels was established in the Swedish market.

Many different assortments and qualities of biofuels were traded. Various projects were carried out to adapt fuels to existing technology. In some areas, cofiring between various biofuels and biofuel and coal is established.

Later results from the imports indicate much large activity and import figures reaching as much as one-third of the supply of biofuels to the district heating sector. No reliable figures for later than 1997 were available at the time of this writing, but the estimated level of imports today is a stabilization of the levels from the mid-1990s. One reason is the weaker Swedish currency, which made it more expensive to buy from other countries. An example of this is the strong U.S. dollar, which made it too expensive to continue with the import of wood pellets from North America.

Swedish currency is stronger and prices have increased due to a larger market. This gives good ground conditions for an increased import (Table I).

Importers are the energy companies themselves or specialized fuel companies. The imported fuels show a spectrum of various assortments and qualities. Sometimes, imports are double counting, and this gives a range of the imports (Table II).

2.2 Exports

During recent years, the acceptance of the Kyoto Protocol has started activities in many European

TABLE I

Imports of Untaxed Fuels (Biofuels) to Sweden in 1992, 1995, and 1997 (in Pentajoules)

	1992	1995	1997
Biofuels	2–4	11–15	29–36
Approximate share of wood fuels	0.2 (estimated)	0.33–0.45	0.5–0.62

Sources. NUTEK (1993) and Vinterbäck and Hillring (2000).

TABLE II

Imports to Sweden of Untaxed Fuels in 1997: Combined Results from Telephone Survey to Fuel Users and Fuel Dealers

Assortment (quality)	Pentajoules	Share provided by importers (percentage)
Tall oil	8.73	59
Wood pellets	4.35	74
Recovered wood chips	3.65	80
Green wood chips	2.70	58
Peat	1.48	0
Rubber and tires	1.29	48
Railway sleepers	0.58	100
Round wood	0.21	100
Bark	0.19	100
Waste wood	0.17	100
Recycled paper	0.14	0
Sawdust	0.07	100
Total 1	23.56	(No double counting)
Total 2	14.70	(Only importers, i.e., maximum double counting)

Source. Vinterbäck and Hillring (2000).
Note. Figures in the survey did not represent the total population of possible users and importers. To get an estimation of the total, import figures were multiplied by 1.36.

countries, as mentioned earlier. Interest in exporting projects has started due to the fact that Sweden and the other Nordic countries, as well as the Russian Federation, have large forest resources. Importing countries in this future scenario are the densely populated European countries and some of the Mediterranean countries. No regular trade is carried out yet, but there have been projects in this field.

3. EXPERIENCES FROM FINLAND

In Finland, the annual use of solid wood fuels is approximately 127 PJ. During a recent 5-year period, more than 100 district heating plants and 500 MW_e of new additional capacity for electricity production from wood-based fuels were commissioned in Finland (total capacity 2000 MW_e). Each year, Finland exports approximately 1 PJ of wood pellets to Denmark, Sweden, and The Netherlands as well as peat and wood chips to Sweden. Each year, the Finnish forest industry imports approximately 13 million solid m^3 of wood for raw material, mostly from the Russian Karelia. Out of this, 11% is bark, which is used for energy production (11 PJ). This import of wood fuel as a separate fraction is equal to 1.3 PJ.

The Finnish national solid biofuel trade is carried out by various organizations. There are big diversified companies working on energy, wood processing, and biofuels business. These companies can exploit industrial wood residues from their own mills for production of wood chips and/or pellets. They can integrate harvesting of logging residues into timber or pulp wood harvesting and, in so doing, can keep fuel prices competitive. In small-scale heating of municipal buildings such as schools and small district heating plants, a special heating entrepreneurship model has been used in fuel procurement as well as in the operation and maintenance of the boiler plants. Usually, these entrepreneurs are cooperative, harvest small-sized wood of their own woodlots, or purchase industrial wood chips or cutter shavings from the local wood processing industry. The number of such entrepreneurs in Finland has increased to more than 130. In large-scale wood fuel procurement, various modern information technologies, such as mobile telephones and the Internet, are used in positioning fuel storage and logging sites, in ordering chippers, and so on.

In Finland and Austria, several electronic firewood marketplaces are established on the Internet (e-trade) for local, national, and even international consumers to select log and pellet suppliers.

4. EXPERIENCES FROM DENMARK

Denmark is a net importer of biofuels and imports trunks as well as wood chips and wood pellets. The import of wood chips covers approximately 25% of the annual Danish consumption of roughly 2 to 3 PJ. The Baltic states and Germany represent nearly 90%

of these imports, with the remainder coming from Poland. These imports consist mainly of wood qualities originating from clear-cuttings and trunks that are not chipped until after their arrival at a Danish harbor. Selling to the Danish market is competitive because it may be difficult to sell wood chips in Germany; furthermore, the prices in Denmark are a little higher than in Germany (Table I). In addition, an unknown amount of moist industrial by-products (e.g., bark, sawdust, chips) is imported.

The international market for wood fuels can be compared with the trading of straw for energy purposes. Straw is a local commodity, and in Denmark there is a geographic fluctuation in the prices due mainly to the bridge toll (Store Bælt) or ferry crossing, which make the straw transport very expensive. Long-distance transport of straw bales is expensive but is not an environmental problem given that a truck loaded with 12 tonnes of straw must drive 17,000 km with the straw before its emissions have compensated for the straw's CO_2-displacing effect in a heating plant.

5. EXPERIENCES FROM AUSTRIA

In Austria, biofuels have a 10.5% share of the annual consumption of approximately 105 PJ. Fossil fuels provide a significantly lower percentage of the primary energy in Austria compared with the European average. Austria has committed itself to an annual 13% reduction in greenhouse emissions. This is a highly ambitious goal that is all the more difficult and costly to achieve given that the per capita share of greenhouse gas emissions in Austria is already relatively low due to the fact that Austria meets a high percentage of its energy demand from hydropower, biomass, and natural gas.

The country is rich in both forest resources and agricultural land, giving Austria the opportunity to use bioenergy. Policy to promote the use of various biofuels in district heating is rather small scale compared with that in the Scandinavian countries.

Trade of wood fuels occurs with neighboring countries such as Germany, Slovenia, and Italy; however, such trade occurs at rather low levels.

6. EUROPEAN EXPERIENCES

Trade between countries has been established in various parts of Europe. In most countries, the customs statistics do not record trade in enough detail to identify international trade of different biomass types. Today, solid biofuels such as wood residues, pellets, and wood chips are already traded in Europe. In some countries, there is a growing interest in international trade because it can provide biofuels at lower prices, larger quantities, and better quality than can domestic alternatives.

6.1 European Trade

Solid biofuels such as wood residues, pellets, and wood chips are now traded in Europe and reached a level of nearly 50 PJ in 1999. Trade among European countries has resulted in a growing interest in biofuel trade because international trade can provide fuels at lower prices. In several cases, the national biomass market is not yet developed well enough for organized international trade. On the other hand, projects may benefit in countries with unexploited biomass resources when fuels are available on an international market. Although there may be cross-border trade of, for example, domestic firewood between neighboring countries, this trade is more or less occasional and beyond official statistics.

The largest volumes of biofuel are traded from the Baltic countries (Estonia, Latvia, and Lithuania) to the Nordic countries (especially Sweden and Denmark but also Finland). Some volumes are also traded from Finland to other Nordic countries as well as among neighboring countries in Central Europe, especially The Netherlands, Germany, Austria, Slovenia, and Italy. The traded biofuels most often include refined wood fuels (e.g., pellets, briquettes) and industrial by-products (e.g., sawdust, chips) as well as (in Central Europe) wood waste. The annual production of wood pellets in Europe is estimated to be 1.2 to 1.3 million tonnes.

Some biofuels are also traded intercontinentally. Sweden has imported biofuels from Canada, and Italy imports firewood from Northern Africa. In addition, Germany exports some firewood to the Middle East and Far East.

Scandinavian biofuel markets have increased, with national energy policies contributing strongly to this trend. Taxes on energy with a clear environmental profile were introduced in Scandinavian countries during the early 1990s. Fossil fuels are heavily taxed in heat production, whereas biofuels are untaxed. In electricity production, all fuels are untaxed, whereas consumers pay a tax. In Finland and Sweden, the investment supports called for growth in the

TABLE III

Minimum, Maximum, and Average Fuel Prices (Including Taxes) in 18 Selected European Countries: 1999 (in GJ)

Fuel	Minimum country	Maximum country	Average
Forest residues	1.02, Germany	8.33, Italy	3.42
By-products, forest products industry	0.58, Romania	9.07, Poland	2.38
Firewood	1.01, Slovakia	14.00, United Kingdom	5.26
Wood waste	−4.00, Ireland	3.31, Poland	0.97
Refined wood fuels	3.24, Latvia	18.22, Germany	8.37
Other biomass	0.83, Slovakia	12.00, Poland	4.68
Peat	2.10, Finland and Latvia	3.75, Ireland	2.83
Heavy fuel oil	1.40, Slovakia	12.00, Ireland	6.74
Light fuel oil	3.10, Slovakia	14.30, Denmark	6.74
Natural gas	1.10, Slovakia	16.21, Italy	5.80
Coal	1.19, Poland	12.78, Germany	4.53

Sources. Vesterinen and Alakangas (2001) and Hakkila *et al.* (2001).

TABLE IV

European Prices of Wood Fuels for Combustion Plants of the Size 1 to 5 MW_{th}: 1999 (in €/GJ, Including Energy Taxes and Excluding Value-Added Taxes)

Country	Bark, sawdust, and chips	Wood chips	Wood pellets
Denmark	4.2	4.5	5.0
Finland	1.6	3.0	7.5
Germany	3.1	3.7	6.1
Sweden	2.9	3.4	4.8
France	1.1	4.0	10.6
Latvia	0.8	1.6	3.3

Sources. Vesterinen and Alokongas (2001) and *STEM* (periodical).

capacities and also contributed to the demand for biofuels.

6.2 Fuel Prices

Prices of various fuels were collected in a 1999 survey. Since then, oil prices have risen significantly, and this also might have influenced the prices for other fuels (Table III).

6.3 European Price Level

The increase in trade among European countries has established international prices for wood fuels. Table IV gives examples of prices. Of course, there is a range in prices set by the market situation, production cost, and cost of competing fuels. By-products were traded at approximately 3 to 4 €/GJ, wood chips at 3.5 to 4.5 €/GJ, and wood pellets at 5 to 6 €/GJ.

7. CONCLUSION

There is a growing international trade in biofuels, and the largest volumes are traded from the Baltic countries (Estonia, Latvia, and Lithuania) to the Nordic countries (especially Sweden and Denmark but also Finland). Some volumes are also traded from Finland to other Nordic countries as well as among neighboring countries in Central Europe, especially The Netherlands, Germany, Austria, Slovenia, and Italy. The traded biofuels are most often refined wood fuels (e.g., pellets, briquettes) and industrial by-products (e.g., sawdust, chips) as well as (in Central Europe) wood waste.

There are international market prices established in Europe; however, volumes are still limited. Both the use and trade of wood fuels are expected to increase, at least in Europe, in the future.

SEE ALSO THE FOLLOWING ARTICLES

Biodiesel Fuels • Biomass for Renewable Energy and Fuels • Biomass Resource Assessment • European Union Energy Policy • Forest Products and Energy • Market-Based Instruments, Overview • Market Failures in Energy Markets • Markets for Coal • Markets for Natural Gas • Markets for Petroleum • Wood in Household Energy Use

Further Reading

Alakangas, E., Hillring, B., and Nikolaisen, L. S. (2002). Trade of solid biofuels and fuel prices in Europe. *In* "12th European Biomass Conference: Biomass for Energy, Industry, and Climate Protection" (W. Palz *et al.*, Eds.). Amsterdam, Netherlands.

Hakkila, P., Nousiainen, I., and Kalaja, H. (2001). "Use and Prices of Forest Chips in Finland in 1999." VTT, Jyväskylä, Finland.

Hillring, B. (2002). Incentives for co-firing in bio-fuelled industrial steam, heat, and power production. *Renewable Energy* **28,** 843–848.

Nikolaisen, L. S. (2001). Trade with solid biomass. *In* "Proceedings of the Conference, Bioenergy 2001," pp. 77–80. BioPress, Risskov, Denmark.

NUTEK. (1993). "Forecast for Biofuel Trade in Europe: The Swedish Market in 2000." Swedish National Board for Industrial and Technical Development, Stockholm.

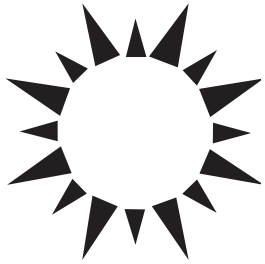
Swedish National Energy Administration. (2001). "Energy in Sweden: Facts and Figures 2001." SNEA, Stockholm.

Vesterinen, P., and Alakangas, E. (2001). "Export–Import Possibilities and Fuel Prices in 20 European Countries." VTT, Jyväskylä, Finland.

Vinterbäck, J., and Hillring, B. (2000). Aufbau eines Europäishen Handels mit Holz-Brennstoffen. *Holzforschung Holzverwertung* **6,** 98–102.

Markets for Coal

RICHARD L. GORDON
Pennsylvania State University, University Park
Pennsylvania, United States



1. **Patterns of Coal Consumption**
2. **End Uses**
3. **The Basis of Change: Coal in "Energy"**
4. **The Mythology of Energy Shortages**
5. **Innovations in Mining, Transportation, and Use**

Glossary

coal The solid fossil fuel ranging from hard anthracite to the softer subbituminous coals.
coking coal Coal converted into coke for pig iron manufacture and other uses.
energy Petroleum, natural gas, coal, nuclear power, waterpower, and alternatives such as geothermal and solar.
steam coal Coal used for heating in boilers.
synthetic fuels Liquid fuels and gases made from other sources such as coal.
unit trains A train dedicated entirely to hauling coal between a specific origin (almost always a coal mine) to a specific destination (either a consumer such as an electric power plant or a transfer terminal).

In many discussions of coal, its abundance is extolled with the implication that coal consumption can be and should be massively expanded. In practice, coal use became increasing specialized. Electricity generation is the major market. Although this is not true in every major consuming center, it is increasingly true for most. This article seeks to delineate and explain this pattern. Discussion begins by indicating patterns of coal use by country and end use. The situation in the early 21st century and differences from earlier years are examined.

1. PATTERNS OF COAL CONSUMPTION

Coal consumption is even more unevenly distributed around the world than total energy use. The amount of coal consumed is affected greatly by the level of economic activity in the country. The relationship, however, is not direct. Wide differences prevail among countries in "energy intensity" as measured by the amount of energy consumed per unit of gross domestic product (GDP) and in the relative role of coal in energy use. Were this not complicated enough, further issues arise about the proper ways to measure economic activity and energy use. Table I summarizes the critical data for the leading coal-consuming countries.

The most striking aspect of the tabulation is the existence of countries with high GDP and energy use and low coal use. France, Italy, Brazil, and Mexico are the main examples. In the French case, the critical influence was a decision to allow a slow death of the domestic coal industry and rely heavily on nuclear power. (Over 80% of French electricity has been nuclear generated since the middle 1980s.) The other three countries simply display a tendency to low coal use characteristic of countries lacking extensive domestic coal supplies.

Higher coal roles necessarily mean that a country will be more important in coal use than in total energy. Several important cases arise. Every country getting the majority of its energy from coal is among the leaders in coal tonnage consumed. China and India are the primary examples. Russia, the Ukraine, and Kazakhstan are notable for unusually high energy/GDP ratios.

Another consideration is that coal is a much less mobile fuel than oil. The basic problem is that solids are more expensive to move than liquids. Moreover, coal has 70% or less heat content per ton than oil.

The vast majority of coal is consumed in the country of production. Again, however, substantial international differences prevail (see Table II). In particular, the four leading coal-consuming countries are all substantial producers, and international trade in coal is less than 10% of each of these countries' output. In contrast, other major producers export

 © 2004 Elsevier Inc. All rights reserved.

TABLE I

Economic Activity, Energy Use, and Coal Use

	GDP 1998[a] (billion US$ at 1990 prices and PPP[b]s)	Total Energy/GDP[a] (tonnes of oil equivalent per thousand 1990 US$ PPP)	Total energy 2001[c] (million tons oil equivalent)	Coal 2001[c] (million tons oil equivalent)	Coal share[c]	Electricity share in coal[a]
World	33,653.523	0.284	9124.8	2255.1	24.7%	64.6%
United States	7,043.638	0.310	2237.3	555.7	24.8%	95.1%
China	4,154.508	0.248	839.7	520.6	62.0%	44.9%
India	1,487.012	0.320	314.7	173.5	55.1%	66.7%
Russia	664.945	0.875	643.0	114.6	17.8%	64.4%
Japan	2,581.473	0.198	514.5	103.0	20.0%	47.2%
Germany	1,483.293	0.232	335.2	84.4	25.2%	87.7%
South Africa	195.894	0.567	107.0	80.6	75.3%	59.2%
Other Asia–Pacific Islands			95.8	58.7	61.3%	15.7%
Poland	257.313	0.375	87.7	57.5	65.6%	62.7%
Australia	362.814	0.289	109.9	47.6	43.3%	85.1%
Korea	539.059	0.303	195.9	45.7	23.3%	66.1%
United Kingdom	1,077.407	0.216	224.0	40.3	18.0%	70.3%
Ukraine	75.978	1.881	131.1	39.0	29.7%	69.7%
Taiwan	398.142	0.193	85.4	30.9	36.2%	66.8%
Canada	610.413	0.384	274.6	28.9	10.5%	84.8%
Kazakhstan	52.116	0.749	43.1	24.7	57.3%	60.9%
Czech Republic	92.227	0.445	41.6	21.3	51.2%	66.7%
Turkey	461.224	0.157	70.2	20.4	29.1%	46.5%
Spain	557.744	0.202	134.6	19.5	14.5%	81.3%
Indonesia	665.131	0.185	97.9	16.7	17.1%	69.1%
Brazil	907.057	0.193	173.6	14.0	8.1%	15.5%
Italy	1,037.380	0.162	177.2	13.9	7.8%	53.7%
France	1,138.864	0.225	256.4	10.9	4.3%	53.4%
Other Europe			41.1	10.6	25.8%	
Greece	118.096	0.228	31.1	9.5	30.5%	88.7%
Thailand	360.047	0.192	63.0	8.8	14.0%	57.6%
Netherlands	305.165	0.244	88.6	8.4	9.5%	69.3%
Romania	64.645	0.613	38.2	7.7	20.2%	61.2%
Other Middle East			91.9	7.2	7.8%	100.0%
Belgium and Luxembourg	209.057	0.298	63.9	7.1	11.1%	43.0%
Other Africa			95.8	6.5	6.8%	
Mexico	757.984	0.195	127.7	6.3	4.9%	62.9%

[a] *Source. Energy Balances of OECD Countries* and *Energy Balances of Non-OECD Countries* (disk versions), The International Energy Agency.

[b] PPP, purchasing power parities, which are calculations that compare currencies in terms of relative domestic purchasing power over all commodities.

[c] *Source. BP Statistical Review of World Energy*, June 2002.

higher proportions of their output and many major consumers are heavily import dependent (see Table III).

The United States alone of the long-time major coal producers remains cost competitive with oil, natural gas, and coal from other countries. The

TABLE II

2001 Consumption and Production of Coal[a]

	Coal consumption	Coal production	Consumption as percentage of production
United States	555.7	590.7	94.1%
China	520.6	548.5	94.9%
India	173.5	161.1	107.7%
Russian Federation	114.6	120.8	94.9%
Japan	103.0	1.8	5722.2%
Germany	84.4	54.2	155.7%
South Africa	80.6	126.7	63.6%
Other Asia Pacific	58.7	68.1	86.2%
Poland	57.5	72.5	79.3%
Australia	47.6	168.1	28.3%
South Korea	45.7	1.7	2688.2%
United Kingdom	40.3	19.6	205.6%
Ukraine	39.0	43.6	89.4%
Taiwan	30.9		
Canada	28.9	37.6	76.9%
Kazakhstan	24.7	40.6	60.8%
Czech Republic	21.3	25.8	82.6%
Turkey	20.4	12.9	158.1%
Spain	19.5	8.0	243.8%
Indonesia	16.7	56.9	29.3%
Brazil	14.0	2.1	666.7%
Italy	13.9		
France	10.9	1.5	726.7%
Other Europe	10.6	12.1	87.6%
Greece	9.5	9.0	105.6%
Thailand	8.8		
Netherlands	8.4		
Romania	7.7	7.3	105.5%
Other Middle East	7.2		
Belgium and Luxembourg	7.1		

[a] *Source. BP Statistical Review of World Energy*, June 2002.

United Kingdom, Germany, Japan, and France are countries with substantially reduced coal production and shifts to imported coal. In every case, depletion has eliminated resources that can compete with alternative fuels or coal from elsewhere.

New more competitive sources of coal have emerged. Many of these—Australia, Indonesia, Colombia, and Venezuela—export well over half their coal; another case, South Africa, approximately 30%. Japan, France, and South Korea are particularly striking cases where domestic production is a small fraction of consumption. The role of domestic supplies is greater in the United Kingdom and Germany but contrasts sharply from the prior tendency to be a net exporter. Korea and Taiwan were never significant coal producers.

TABLE III

Coal Production and Exports 2001[a]

	Production	Exports	Exports as percentage of production
China	940,000	91	9.68%
United States	927,868	44	4.74%
Russia	330,000	36	10.91%
India	315,000		
Australia	257,288	194	75.40%
South Africa	221,273	69	31.18%
Poland	103,000	23	22.33%
Canada	32,450	30	92.45%

[a] *Source. Zahlen Zur Kohlenwirtschaft*, Statistik der Kohlenwirtschaft, E.V. (June 2002).

2. END USES

The history of coal use involves greatly increased specialization of use. In particular, consumption to generate electricity became the dominant use. The United States is the prototype for this development. Data are available on U.S. coal use since 1920 and show a massive shift in coal-use patterns (Table IV). In 1920, coal was used throughout the economy. Over one-half was employed as a general-purpose fuel for "industry" (which as reported apparently included residential and commercial use). Over one-quarter of use was to fuel the railroads. The mix of use persisted to 1945 (and the level of use was only 10% higher in 1945 than in 1920).

A precipitous decline of coal use prevailed from 1945 to 1961. The adoption of the diesel-electric engine eliminated the railroad consumption of coal. Sales for general industrial use and retail deliveries declined. Use in coking fluctuated without showing any persistent trend. However, growth in electric power use was maintained.

The post-1961 history involved continuation of both gains in sales to electric utilities and declines elsewhere. Indeed, coking joined other industry and retail deliveries as a declining sector. Since rail use

TABLE IV

U.S. Coal Use by End Use[a]

	Electric utilities	Railroads	Coking	Industry	Retail	Bunkers	Total
1920	30,099	135,414	76,191	255,959		10,486	508,148
1929	39,729	131,100	86,787	257,214		4,287	519,117
1945	71,603	125,120	95,349	148,198	119,297		559,567
1960	176,685	3,046	81,385	96,017	40,948		398,081
1970	320,182	298	96,481	90,156	16,114		523,231
1980	569,274		66,657	60,347	6,452		702,730
1990	773,549		38,877	76,330	6,724		902,893
2001	968,990		26,075	63,361	4,127		1,062,554
1920	5.9%	26.6%	15.0%	50.4%	0.0%	2.1%	100.0%
1929	7.7%	25.3%	16.7%	49.5%	0.0%	0.8%	100.0%
1945	12.8%	22.4%	17.0%	26.5%	21.3%		100.0%
1960	44.4%	0.8%	20.4%	24.1%	10.3%		100.0%
1970	61.2%	0.1%	18.4%	17.2%	3.1%		100.0%
1980	81.0%		9.5%	8.6%	0.9%		100.0%
1990	85.7%		4.3%	8.5%	0.7%		100.0%
2001	91.2%		2.5%	6.0%	0.4%		100.0%
2001/1960	448.4%	−100.0%	−68.0%	−34.0%	−89.9%		166.9%

[a] *Sources*. U.S. Bureau of Mines Reports (for years up to 1945) and U.S. Department of Energy, Energy Information Administration, *Annual Energy Review*, Table 7.3 (for later years).

had almost disappeared, further losses in transportation were minute. Electric power use gains, however, began to outweigh the losses and coal use began rising. These trends have persisted. The other sectors shrink while electric power use rises. By 2001, electric power absorbed 91% of coal use. General industrial use was 6% and coking use was 2%.

The tendency to increased importance of electric power as a coal consumer prevails in most of the main coal-consuming countries. The details again differ but most countries and the world as a whole use more than half of their coal for electricity production. The main exceptions are China and Japan. In both cases, electricity is becoming an increasing portion of coal use. The Chinese electricity use already exceeds that for any other end use. Japanese electricity use is nearing parity with coking use.

Germany uses approximately 88% of its coal in electricity. Here the pattern of steady electricity growth and declines elsewhere produced steadily declining total coal use.

The below average degree of electricity use in South Africa reflects its unique venture into synthetic fuels from coal. South Africa under white rule undertook an extensive effort to produce liquid fuels, gas, and chemicals from coal. This generated at most approximately one-third of the inputs for producing petroleum products. Oil procurement problems were not as dire as feared. Even at the height of the United Nations boycott, major European oil companies such as Shell, British Petroleum (BP), and AGIP openly operated in South Africa. The majority of coal use in Poland is for electricity. The use has declined sharply since the fall of communism.

3. THE BASIS OF CHANGE: COAL IN "ENERGY"

The central proposition about coal is that it is a problem-plagued source of "energy." To see this, the basic propositions about energy need review. Energy as usually treated is defined to include only inanimate energy sources and usually not all of them. Five types of sources—coals, oils, gases, uranium, and waterpower—are invariably covered. Depending on their importance in a country and the interests of the observer, such substitutes as wood, animal wastes, other sources of biological origin, wind, and solar energy also are treated. (Coverage of these sources

has expanded in large part reflecting interest in increasing the use of "alternative fuels." The data show that present utilization is heavily concentrated in poor countries. China and India consume the largest amounts of such fuels and together account for approximately 38% of reported world totals. However, in both cases, the majority of reported fuel use comes from the five usually reported fuels. Smaller, even more impoverished countries such as all sub-Saharan Africa other than South Africa obtain the vast majority of their fuel from alternative sources. In short, these fuels are widely considered the last resorts for those not integrated into the world economy.) Thus, energy consists of things that are burned to secure work or close substitutes for such material.

Moreover, all three fossil-fuel sources actually are secured from heterogeneous occurrences. They differ within as well as among deposits in such key characteristics as amount and form of the fuel contained, the extent of contamination by other material, and the ease of extraction. In particular, solid fuels range from the very hard (anthracite) to the boggy (peat). Although classifications differ among countries, coals extend at most to the softer, more ash-laden forms called lignites or brown coals.

Outside the United States, lignite is treated as being distinct from the harder coals—anthracite, bituminous, and subbituminous—and even subdivided into better and worse qualities. Thus, the Germans distinguish between braunkohle and pechkohle. In contrast, the U.S. reporting practice was established long before lignite became a significant part of output. U.S. data still combine lignite with harder coals. (Conversely, the procedures arose when the United States had substantial anthracite production. Anthracite was reported separately long after it ceased to be important. Under the new methodology, anthracite is reported with bituminous and lignite with inclusion of separate anthracite subtotals.)

Those hard coals used to make coke for ironmaking are termed metallurgical or coking coals. All others are termed steam coals. The coals are not necessarily different; often the only difference is the use made. Considerable evidence exists that producers shift between the two sectors as market conditions warrant. (In particular, Queensland long has reported the disposition of coal from each of its mines, and shifts between steam and coking sales can be observed.) This may or may not alter what is mined and how it is processed. The flexibility, in any case, links the steam coal and coking coal sectors of the market.

The work done by these fuels, in turn, involves primarily providing heating, lighting, and cooling to residential, commercial, and industrial buildings, powering industrial processes, and fueling automobiles and other forms of transportation. With the rise of electric power in the late 19th century, a tendency has arisen increasingly to consume fuels to produce electricity.

The physical characteristics of each fuel are a major influence on their production, transportation, processing, and utilization economics and on the information that is known about them. Coals are solid fuels that are more readily found than oil, gas, or uranium. Being solid, however, is otherwise an economic drawback. Solids by definition lack the fluidity that makes oil and gas so attractive and, as noted, have less heat content per ton than oil. As a result, extraction, transportation, storage, handling, and use are more difficult than is true for oil and gas. With transportation, storage, and handling, more material must be handled by less supple methods. Coal must be pushed through every stage. For example, a coal-using facility requires larger, more complex, more expensive boilers than those needed to obtain the same amount of heat from oil or gas.

In particular, although oil can be fairly cheaply transformed physically and chemically, coals are expensive to transform. The expense has been such that only an economy such as South Africa with a strong desire for self-sufficiency heavily engages in coal transformation.

Among the drawbacks of coal are the environmental impacts. The original problem was of the high waste content that emerged as ash, much of which was discharged during combustion. This difficulty was alleviated at a cost by the development and adoption of techniques to collect the ashes before they would have been discharged. The next major concern was the sulfur contained. This was, in fact, also a problem for natural gas and oil but removal or shifts to naturally less-sulfur-containing sources were more economic than with coal. The coal options are naturally lower sulfur coals such as in the U.S. Rocky Mountain States or adoption of postcombustion clean-up technologies. Further concerns arise from other contaminants in coal.

The newest concern is that coal emits more of the carbon dioxide believed to be a critical greenhouse gas in the global warming threat. Thus, adoption of vigorous measures to limit greenhouse gas emission could produce major new pressures to reduce coal use.

During the 20th century, coal changed from a dominant fuel to one often beleaguered by competition from other fuels. With the prevailing scarcity pattern, the role of coal is to fill niches in which its drawbacks are less. Coal is attractive when location and size of operations justify use of coal. Energy use on a larger scale produces economies. Just where they are exhausted is unclear. It is at a level much smaller than the electric power market but apparently much larger than that maintained by most manufacturing plants and commercial facilities. These scale economies narrow, but do not eliminate, the cost disadvantages of using coal. If large cheap-enough coal supplies exist, savings on fuel costs will justify endurance of higher other costs and lead to coal use by electric utilities. Large boilers predominantly in electricity generation are a prime and indeed steadily growing market.

Coal as the input for the coke used in pig iron manufacture is a niche that has shrunk. The remainder of coal use is scattered, predominantly in manufacturing plants.

Steel is an old, beleaguered industry that has undergone remarkable changes. Coal use has been squeezed by adverse developments in this market. The alterations involve the location of raw material suppliers and steel producers, the technologies used in different stages of steelmaking, and the product mix. These have been good for consumers but bad for coal. Historically, coke-based iron-making provided most of the inputs for steelmaking. In what is termed the hot-metal approach, pig iron is refined into steel. Some scrap also can be used. Steelmaking technology has changed so that less coal is needed.

First, steel output has fluctuated without much net gain. New technology lessens processing losses and thus the gross input of steel per ton of finished steel output. The hot-metal route, moreover, is losing ground to the alternative of producing steel entirely from scrap (using electricity). This occurred despite introduction of the oxygen furnace that reduced costs of the ore-based route. A secondary consideration is the ability to supplement coke with other fuels such as oil and uncoked coal. Thus, coke has many direct and indirect rivals.

More innovations that would further reduce coal use are under consideration. Certain products such as the thin sheets used for automobile bodies could not be made satisfactorily from steel from scrap, but technology has emerged to overcome this limit. For decades, steel companies have experimented with technologies to make iron without coke. Application still is limited to countries in which coking coal is expensive to attain, but hope of change persists.

The growing dependence of the coal industry on electric-power customers has many important technical and economic implications. Coal suppliers are dealing with large entities participating in an industry with far more flexibility in fuel choice than any other fuel users. This means substantial substitution possibilities and widespread knowledge of these possibilities.

The components of an electric power network include generating units of different fuel-using capabilities in many different locations. A unit consists of a generator and whatever causes it to operate. A plant has the usual meaning of a total facility. Although one-unit plants exist, it is common to have several units. These sometimes consist of several identical units but often differ from one another in various degrees. Some simply have units that differ only in age and size, but many cases arise of major differences among units. One U.S. plant has a coal unit and a nuclear unit. Many have both boiler-powered plants and combustion turbine units (see below). Often only some of the boilers are designed to burn coal.

Nuclear units using technologies that rigidly determine fuel choice are one pole of the spectrum of flexibility. At the other extreme are plants on the U.S. East Coast physically able to burn coal, oil, and natural gas and located so that they can be economically supplied by all three fuels. Much lies in between. A boiler designed for coal can be (and often has been) converted to oil and gas use. In the 1960s era of cheap oil, utilities located at the center of interfuel competition on the U.S. East Coast built triple-firing capability and shifted among coal, oil, and gas as prices changed. Such plants were readily served by tankers carrying fuel oil and close to coalfields. Gas use was attractive at least during the warm months in which home heating use ceased. When natural gas appeared likely to remain a low-cost fuel, utilities in gas-producing states (most notably Texas and California) built plants designed to burn only gas.

When cheap oil seemed likely in the late 1960s, plants that burned only oil were built. This occurred mainly on the East Coast. Prevailing public policy then freed East Coast power plants from import controls that discouraged oil use elsewhere. A few inland (Illinois and Michigan) plants also arose in the expectation that Canadian oil sources would be an attractive alternative.

Finally, the demand for electricity fluctuates, with some of the variation being unpredictable; equipment,

moreover, is not continuously available. Thus, the need exists to maintain spare capacity to meet demand surges given equipment availability. In these circumstances, great short-run flexibility exists. Output mixes can and do shift in response to changing fuel prices. Plants already using cheaper fuel sources are used more; plants that can easily shift to cheaper fuels do so and also are utilized more.

In the longer run, the mix of available plants can be altered to allow still greater use of cheaper fuels. Thus, the enthusiasm for nuclear power in the early 1970s inspired a flurry of orders that was (partially) implemented through the rest of the decade and throughout the 1980s.

Similarly, as gas became more expensive, gas-using utilities started building coal-fired units. This process started in West North Central and Mountain states and later spread to the large oil-producing states such as Texas, with California choosing for various reasons to rely on coal-generated electricity transmitted from plants in Arizona, New Mexico, Nevada, and Utah.

4. THE MYTHOLOGY OF ENERGY SHORTAGES

In contrast to the limited prospects just sketched, restoration of coal to a role of eminence often was advocated. Tendencies exist to exaggerate both the advantages and disadvantages of coal use. Coal is regularly proclaimed a highly abundant fuel. The implication often is that a great potential has not been realized and coal should be restored to its historical role as the leading source of energy for general use. These assertions, however, ignore or at least mute debate of the formidable barriers to such a massive return. The correct view is that the industry has done extremely well in a hostile environment.

At least immediately after the rapid oil price increases of 1973–1974, discussions of energy routinely argued that a massive return to coal is essential. (Those with longer memories know this idea often has risen and fallen since oil competition emerged in the middle of the 19th century and that the coal industry perennially reiterates the contention.)

The advocacy of a return to coal is based on the presumption that the economics will change radically within a few decades. Some add that since so much oil comes from a small group of politically unstable countries engaged in efforts to cartelize supplies, the supply problems will emerge sooner. Coal then will become competitive again.

This then inspires an unending debate in which both sides remain impervious. Both views have some truth on their side, albeit not to an equal extent. Nothing lasts forever, but oil and gas supplies are greater than those expecting an immediate demise realize. The disparity between expectation and outcome arose from misunderstanding of the basic economic principles affecting coal. Unfortunately, many people believe in the desirability of more expensive energy. Therefore, clearly premature predictions of vanishing oil and gas supplies frequently appear and error fails to discredit the advocates.

The available data suggest that the physical supply of coal is much greater than that of oil and gas. Since oil and gas, as noted, are harder to find than coal, the appearance may be incorrect. The data are biased because coal outcrops more than oil seeps. Thus, more can be and is readily learned about physical endowments of coal than about oil and gas. Indeed, coal resource data are based almost entirely on physical availability. In contrast, oil and gas reserves cover only the well-delineated resources in fields developed for production. Thus, whereas coal estimates change little, oil and gas resources are steadily augmented.

Ignorance remains about the true state of future energy supplies because acquiring the knowledge involves costs from which investors see inadequate benefit. Those who might use the knowledge to assist investment in developing energy supplies are aware that they can safely delay their decisions until possible use is nearer at hand.

Even if the assertions about physical abundance are true, relative physical availability does not dictate a massive swing to coal. Physical endowment is not economic supply and not even a good proxy for economic supply. Supply relates to the cost of extraction and use. The best is used first. The available physical stocks of oil and gas contain a larger portion of the cheapest resources to produce, process, and use than do the available coal resources. This is why oil and gas attained a larger energy role than coal.

The impacts of physical availability will occur so far in the future that research and development may have succeeded in making competitive energy sources that have even greater physical availability. It can readily be seen that at least another half-century's worth of extensive oil and gas supplies remains. Oil and gas prices may remain at 2002 levels for many more decades. The balance among coal, oil, and gas, therefore, will not change radically at least until the middle of the 21st century.

No guarantee exists that the coal industry is ensured an ultimate victory. The interim could permit the development of many alternative sources of energy. Instead of being the only option, coal may be one of several. Over such a time span, more will be learned about the availability of oil, gas, and coal and if profitable, even more ample energy supplies (e.g., uranium—given supply-augmenting techniques called breeding, nuclear fusion, and solar energy) will be developed as will new ways to use them.

For example, nuclear fission involves the release of energy by splitting atoms. Such fission occurs only with unstable materials such as uranium 235 and plutonium. However, as demonstration projects proved, it is possible to build reactors in which the splitting produces both energy and the conversion of uranium 238 into a fissionable material. Since more fissionable atoms are added than used, the process is termed breeding. Plans to build a larger demonstration project were terminated by the Carter administration. Although it stressed environmental concerns, another problem was that the effort was another example of reaction to premature warnings of resource exhaustion.

Indeed, in the years since 1973, coal has become increasingly déclassé. Resource pessimists are stressing a future in which conservation and alternative energy rather than coal (or even worse for them nuclear power) replace oil and natural gas.

Thus, many outcomes are possible. It would be a serious mistake to commit the world to any option for this distant future. The need to act has not arisen and it is still not known what the best approach will be when the time comes.

The enthusiasts of government assistance to promote rapid changes in energy utilization believe that oil and gas soon will become very expensive and that energy producers and consumers are not acting rapidly enough to prepare.

Two (questionable) reasons exist for arguing that public policy has a major role in ensuring a more efficient transition. The first relates to the externalities in research on energy possibilities. The core issue here is whether government assistance will be more efficient than forming consortiums of private companies. A second issue is the perennial debate about whether markets or governments take better care of the future. Governments possess greater resources and thus potentially are better able to hedge against risks. However, the need to win elections puts stress on preserving the status quo and concentrating on quick payoffs. Thus, controversy persists over how much of a role should be played by governments.

Contentions about the abundance of coal often lead to proposals for increasing coal use. The coal industry thought (probably incorrectly) that President Nixon was committed to doubling coal output; President Carter made a clear commitment to such a doubling. Neither president did much to promote the outcome.

In the Nixon case, optimism was killed by analysis. A massive effort was undertaken to appraise the feasibility of Nixon's call for attaining energy independence. Those studying coal-production prospects tentatively suggested that, with effort, coal output could be doubled. The work actually grossly understated the ability to expand coal output. It posited barriers that failed to matter.

However, this proved irrelevant. Those producing the computer analyses that synthesized all the studies into estimates of possible market clearing production–consumption patterns could not devise a plausible combination of market conditions that would lead to utilizing all the physical possibilities for expanding coal output and thus reflected economic realities that limit the ability to use coal.

Extensive evidence of the ability of energy markets to adapt to changing circumstances is available. Energy consumption growth slowed radically with the rises in oil prices of the 1970s. The pattern of energy use also changed. A substantial part of the alteration involved shifts by U.S. electric utilities to coal and nuclear power. U.S. refineries long produced a higher proportion of gasoline and less of fuels used in industry than other industrialized countries. The U.S. pattern reflected the availability of low-cost alternatives, such as natural gas and coal, that were formidable competition for oil products in industrial markets. Changing oil prices combined with rising natural gas, nuclear energy, and foreign coal supplies produced a European move to a similar pattern.

Coal possesses a niche that has expanded. Efforts to alter this must carefully appraise the consequences of reducing coal use.

As attention turned to the environmental side effects of energy production, processing, distribution, and use, it appeared that more serious effects were associated with coal than with oil and natural gas. This further reduced the competitiveness of coal.

Policy makers deal with these contradictions by professing to want more coal consumption while imposing barriers to its production and use. Whatever the realities, the practical implication has been that generally greater stress is placed on taming

the environmental impacts of coal than on promoting its use.

Contrary to statements often made by alarmists about energy, the existence of "special" uses is another irrelevance. All the proposed nonmarket criteria prove defective. Both economic theory and energy practice indicate deficiencies with fascination about critical uses. Basic supply–demand analysis shows that what is served depends on market conditions. The greater the supply, the more demands that can be met. The case for consumption control involves misunderstanding of these basic economic principles.

The proposed alternatives to market allocation have serious drawbacks. Each suggestion for administratively allocating energy has clear defects and conflicts prevail among different proposals. The proposed rules for overriding market decisions prove neither clear nor appropriate.

One of several widely used concepts of urgency is satisfying some end use such as transportation and chemical production. Some concentrate on favoring types of users—favoring households over commercial establishments, commercial establishments over factories, and factories over electric power plants.

The differences in urgency, however, can be greater within any broad category than between categories. The chemical uses of fuels extend from life-saving drugs to fancy wrappings. The marginal chemical uses of fuels, such as packaging, are less urgent than some of the applications simply to provide heat. An apparently superior use may not persist as with coke for iron-making. A new basis, moreover, may conflict with the old. The widespread political preference for direct household use over use in producing goods for households is inconsistent with a stress on chemicals.

More basically, favoring direct consumption is clearly harmful to such final consumers. Indirect uses are undertaken to provide goods for final consumers. Such uses can outbid households for energy because the industrial use has more value to consumers than direct consumption of energy.

However, the price system is preferable to political earmarking in determining preferable uses of energy resources. Markets decide on the basis of willingness to pay; this approach responds to the differences within broad categories ignored by policies giving blanket favoritism. No convincing evidence exists of market failures to anticipate future demands. In addition, the preferred ways of meeting a need are not even the only ones already known and new modes could emerge. Thus, even if best uses could be identified, saving fuels for them may not prove the best way to serve them.

The contentions perpetrate one of the oldest errors in economics (dating back at least to John Law, a notorious speculator of the early 18th century) over when it is efficient to extend use beyond the most vital ones. The debate is epitomized by the diamond–water paradox—why does a decorative good sell for so much more than something vital for life? When supply is sufficiently ample, it proves good economics to serve less valuable uses—water for car washes and gas for a barbeque.

The environmental impacts associated with extracting and using coal justify regulation. However, the controls should deal directly with the environmental problems. An indirect indicator, such as energy use, can be a poor proxy for pollution impacts.

Nonintervention avoids both tying up resources and unnecessarily acting before relevant information develops. Too rapidly anticipating demand can lead to choice of technologies that prove obsolete. Preserving reserves may be as onerous for this century as preserving a supply of horses sufficient to cover the return of animal power would have been to the last century.

Thus, whatever the truth about the availability of coal, it does not by itself determine the optimal pattern of coal-consumption evolution. Even if a shift back to coal were ultimately to occur, this does not necessarily require that oil and gas be hoarded and more critically indicate that special government policies are needed to promote that hoarding.

5. INNOVATIONS IN MINING, TRANSPORTATION, AND USE

Coal and its producers have a reputation for stodginess that is belied by the record. Over the years since 1960, a major building program has taken place. Substantial expansion occurred in the western United States. Other important developments included those in Australia, South Africa, western Canada, Colombia, Indonesia, and Venezuela.

This has involved many innovations in extraction and distribution technology. Mechanization of underground mining has increased considerably. Better machines have been developed to cut the coal and transfer it to the intramine transportation system. Methods to mechanize the installation of mine supports have been adopted. An earlier simpler step was to develop a technology for placing bolts to hold

up roofs. Traditionally, the ability to avoid expensive systems of supporting the roofs with elaborate props was an advantage of U.S. mining. Extensive European efforts to mechanize prop installation proved so attractive that they became widely adopted in U.S. mining. A technology that failed to reverse the unprofitability of deep, thin-seamed European mines even more greatly benefited shallower mines in thicker seams. Similarly, heavy equipment builders developed large machines called draglines to remove the masses of material that cover coal extracted by surface mining.

Other innovations have occurred in transportation. Railroads around the world run dedicated (unit) trains between mines and their customers (or a transfer point to water-borne shipment). Various methods have been devised to speed loading and unloading. One approach is to turn each car upside down and to empty it into a receiving pit; another is to empty the coal from the bottom of the cars as they slowly move over the receiving pit. Conveyers can move the coal to storage areas and then into ships or power plants. Bulk shipping terminals are used to transfer coal from trains to ships.

The most important changes in utilization were improvements in the efficiency of electric power plants. Such improvements clearly proceeded for much of the 20th century. Concerns are widespread, however, that the innovation process had hit serious barriers by at least the 1970s.

However, the intrinsic difficulties have thwarted the desire to widen use. Considerable attention has been devoted for many decades to a combination of transformation and utilization technologies that would facilitate coal use. Coal can be converted by long-known technologies into both gases and liquids of various qualities. Thus, the gas might be as good as naturally occurring gas (methane) or have a much lower energy content. The liquid might be gasoline or a thicker fuel suitable only for large boilers. Various alternative utilization technologies have been proposed. However, widespread commercial adoption of these technologies has been difficult to attain. The costs remain difficult to reduce to competitive levels.

A key element of coal promotion rhetoric is claiming that such new technologies will soon emerge. For example, in 1948, the Bituminous Coal Institute (an industry trade association that was subsumed into the National Mining Association) proclaimed in a pamphlet extolling the virtues of coal that "...coal and lignite comprise over 98% of the U.S. mineral fuel energy reserves...." (p. 12). The discussion suggested that coal use would expand through synthesis of liquids and gases from coal, greater use of electricity, and better methods of direct coal burning, including utilization in gas turbines (an adaptation of the jet engine for stationary use, called a combustion turbine). Only the electricity prediction came true. The gas turbine was developed but still cannot directly use coal. Attention shifted to coupling the turbine to a coal gasification plant.

Two main realms of research arise: improved combustion technologies and synthetic fuel approaches. In both realms, a range of options exists. At least three alternative combustion technologies have been explored. The range of synthetics extends from better solids to efforts to produce gasoline and methane (the principal component of natural gas).

Hybrids have been suggested. For example, the combined-cycle approach involves attaching a conventional boiler to a combustion turbine. The boiler utilizes the waste heat from the turbine. Efforts have been made to develop a technology to produce synthetic gas from coal with a heat content above that from 19th century gasification technologies but less than that of methane. Building such a gasifier next to a combined cycle could allow plant coal use. However, none of the technologies proposed has proved to be economically feasible.

These concepts have regularly been presented by various U.S. government agencies as new salvations for coal. U.S. government agencies have gone from the Office of Coal Research in the 1960s to the Energy Research and Development Administration of the 1970s to President Carter's synthetic fuels program to the Department of Energy's clean coal technologies in the 1990s. Throughout, it has been difficult to produce competitive new technologies. This necessarily engenders skepticism about current claims.

SEE ALSO THE FOLLOWING ARTICLES

Clean Coal Technology • *Coal, Chemical and Physical Properties* • *Coal Conversion* • *Coal, Fuel and Non-Fuel Uses* • *Coal Industry, Energy Policy in* • *Coal Industry, History of* • *Coal Mining, Design and Methods of* • *Coal Preparation* • *Coal Storage and Transportation* • *Market-Based Instruments, Overview* • *Market Failures in Energy Markets* • *Markets for Biofuels* • *Markets for Natural Gas* • *Markets for Petroleum*.

Further Reading

Adelman, M. A. (1995). "The Genie Out of the Bottle: World Oil since 1970." MIT Press, Cambridge, MA.

Adelman, M. A., Houghton, J. C., Kaufman, G., and Zimmerman, M. B. (1983). "Energy Resources in an Uncertain Future, Coal, Gas, Oil, and Uranium Supply Forecasting." Ballinger, Cambridge, MA.

Barnett, H. J., and Chandler, M. (1963). "Scarcity and Growth: The Economics of Natural Resource Availability." Johns Hopkins Press for Resources for the Future, Baltimore, MD.

Gordon, R. L. (1987). "World Coal: Economics, Policies and Prospects." Cambridge University Press, Cambridge, UK.

Simon, J. L. (1996). "The Ultimate Resource 2." Princeton University Press, Princeton, NJ.

Markets for Natural Gas

AAD F. CORRELJÉ
Delft University of Technology
Delft, The Netherlands

1. **Introduction**
2. **Historical Development of Markets for Natural Gas**
3. **The Liberalization of Gas Markets**
4. **The Future of Natural Gas: New Patterns of Supply and Demand?**

Glossary

distribution of gas Operations necessary to deliver natural gas to the end users, including low-pressure pipeline transportation, supply of natural gas, metering, and marketing activities vis-à-vis the several types of customers.

liberalization Reduction of public authorities' intervention to coordinate the behavior of actors in (gas) markets.

market for gas System of institutional, economic, and physical elements and determinants through which gas is produced, transported, and sold by suppliers to the consumers.

natural gas Mixture of methane and other gases, as it is found in gas fields in geological layers of the earth crust, that may be either associated or not with crude oil. Natural gas is not a fraction of distilled crude oil.

production Exploration, drilling, production, and the collection of gas from the fields' wellheads.

regulation of gas market Intervention by public authorities to coordinate gas businesses' and consumers' behavior through sets of rules, standards, and prescriptions.

transmission Long-distance, high-pressure pipeline transport of gas from the producers to the distribution systems in the consumer markets.

A continuous, steady growth of gas consumption in domestic households, industry, and power plants in North America, Europe, Asia, and Latin America has gradually turned natural gas into a major source of energy during the past century. Main drivers in this development are the technical and economic advantages of natural gas as a clean, versatile, and easily controllable fuel. A major challenge, however, has been the establishment and management of the markets for gas. This article examines how gas markets have been constituted in the past as a function of economic, technical, and institutional components and how they have been altered in recent times. In particular, it illustrates how the setup and later restructuring of these markets have been influenced by the characteristics of the complex production, transmission, and distribution systems through which the gas is transported from increasingly remote on- and offshore fields to a wide range of users.

1. INTRODUCTION

Natural gas supplies approximately one-fourth of the world's total commercially traded primary energy requirements. A continuous, steady growth in gas consumption in domestic households, industry, and power plants has gradually turned natural gas into a major source of energy. Main drivers in this development are the technical and economic advantages of natural gas. It is a clean, versatile, and easily controllable fuel for which no on-site storage is needed. Further growth in gas consumption is expected as a consequence of its relatively low carbon content compared to that of coal and oil products. On this basis, gas is often considered the form of energy that will be the "bridging fuel" to a sustainable energy system, sometime after 2050.

Unlike other main sources of energy, such as oil and coal, gas is not traded on an actual world market. This is because gas is made available to its consumers by means of complex production and transport systems, through which it is moved often from remote fields to its users. The geographical reach of these pipeline transport and distribution systems is essential for supply and demand to develop.

 © 2004 Elsevier Inc. All rights reserved.

Traditionally, the development and exploitation of these systems posed a great challenge because of the great risks and uncertainties involved. Huge investments must be committed to facilities that, once constructed, have only one purpose and destination and no alternatives. In addition, producers, transporters, and consumers are tied into a relationship of mutual dependency. In response to these characteristics and the specific local setting of systems, a variety of contractual relationships and organizational structures were established to reduce the risk involved and to establish the terms of trade for a longer time period so that the producers' as well as the consumers' investments would not be jeopardized.

This has resulted in the development of truly regional gas markets in different areas of the United States and later in continental Europe, the United Kingdom, Japan, the Soviet Union, and Latin America. Each market has its own market structure, characteristic institutional framework and role for governments and local authorities, and specific outcome in terms of the economics of supply and demand.

During the 1980s, a gradual shift in economic thought began to take shape in which the stabilizing role of the state and the need to control markets in general were questioned. It was argued that the state would never be able to coordinate the economy more efficiently than the market. The state could never acquire and process the necessary information to do so, whereas government failures would undermine the economy's efficiency. Moreover, in the process of planning, the government ran a serious risk of being affected by interest groups or by political deadlock. The arguments for restructuring were reinforced by a plea to integrate national and regional markets for goods because international trade theory argued that economic welfare would be enhanced by allowing production to take place in the most efficient location or country. Because countries differed widely in their energy resource endowments, national (energy) markets needed to integrate to the extent that the process of producing and trading energy would no longer be confined to national territories. To achieve this, national trade regimes had to remove existing barriers to trade, whereas the physical infrastructure to efficiently transport energy between and within countries, such as pipelines, ports, and railways, had to be developed.

Gradually, in the several gas-consuming regions processes of structural and regulatory change were undertaken. Again, the evolution of these processes was influenced by local, economic, and (geo)political circumstances. This is reflected in the timing of these processes, the speed with which they evolved, and the structural models chosen to reform the gas supply "systems" into actual gas "markets." This article provides some illustrations of the manner in which some of the main gas markets have been organized in the past and how they are currently being restructured. Due to space limitations, it does not provide a wide overview of such developments. Neither does it present a quantitative overview of the actual development of gas consumption and supply in the countries and areas discussed.

2. HISTORICAL DEVELOPMENT OF MARKETS FOR NATURAL GAS

The use of natural gas on a large scale is a relatively recent phenomenon, occurring in the 20th century. Much earlier, however, gas was used in local systems as so-called "town gas" produced in coke works, sometimes associated with steel mills, or in specific gas manufacturing plants. In part, however, it involved natural gas, from small-scale underground gas reservoirs, supplying nearby villages and cities. Thus, in the United States the first recorded major use of natural gas was in 1821 in Fredonia, New York. In Russia, the first use of natural gas was associated with the production of oil in the Baku in 1871.

It was not until the 1920s that developments in welding technology facilitated the construction of steel pipes capable of handling higher pressures. This allowed the development of much larger systems over extended areas, through which actual systems of wells could be connected with consumers attached to local distribution systems, and in which depleted fields could be replaced by newly found deposits farther away. In the United States, the evolution of long-distance gas began in 1925. Elsewhere, it would take until the end of the 1950s until these systems developed. In Europe, major gas discoveries in The Netherlands in 1959 marked the start of the development of a natural gas system during the 1960s, stretching to Germany, Belgium, France, Switzerland, and Italy. The United Kingdom followed in 1967 after the discovery of significant fields in the North Sea. In the Soviet Union, large-scale use of gas began in the 1950s.

The systems that have developed since then normally involve the following: the production of gas, pipeline transportation, distribution, trading, and supply. The production segment involves exploration, drilling, production, and the collection of

gas from the fields' wellheads to move it to the transmission pipelines. Gas transmission involves the long-distance, high-pressure pipeline transport of gas from the producers to the consumer markets. Trading refers to the resale of natural gas in the wholesale market and retail market. Natural gas distribution consists of the operations necessary to deliver natural gas to the end users, including low-pressure pipeline transportation, supply of natural gas, metering, and marketing activities vis-à-vis the several types of customers.

Traditionally, a major challenge has been the establishment and management of these complex production, transport, and distribution systems. There is great risk involved. Huge, specific investments have to be made into facilities that, once constructed, have only one purpose, namely to produce gas and transport that gas from a specific field to the consumers attached to the system. If either the consumers decide to turn to alternative fuels or the gas supplier stops producing, the system becomes useless and the capital invested worthless. Thus, the system ties producers and the consumers in a heavily dependent relationship. Each side has to face the risk that the other side will drop out for whatever reason or that prices will increase or decrease to unacceptable levels. For producers, the former implies a volume risk in the sense that sunk investments in exploration and production become worthless because they will not be able to sell the gas they produce; for consumers, it may mean that their gas-fired appliances and equipment will fail to perform. The price risk implies that because of the tight relationship between them, either producers or consumers are able to put pressure on the other side to sell the gas at either too high or too low prices.

In a growing market, a key challenge over time is to maximize capacity utilization of the system, as well as revenue, to cover high fixed costs, including an acceptable level of profit. The main instrument to achieve this goal is a supply portfolio that dynamically balances the sales to these several markets as a function of the relative prices that these markets will bear and the capacity available in the system. In the energy market, gas has to compete with other fuels. Depending on the sector, fuels such as coal, city gas, fuel oil, and heating oil are substitutes for natural gas.

Main sectors with specific characteristics are the domestic household sector, the commercial market (also involving public use in schools, hospitals, etc.), the large industry sector, and the electricity production sector. In each of these sectors, the use of gas for purposes of space or process heating, hot water supply, or as a feedstock involves particular patterns of demand with respect to the load factor and daily and seasonal swing factors. Gas for domestic and other space heating purposes is subject to an enormous seasonal load variance. The electricity sector is a dynamic market, which relatively easily switches from one source of energy to another. A cyclical pattern develops over time. The highly profitable market for domestic household hot water supply and cooking is normally associated with sales in the space heating segment, driven by the cold season. Sometimes short-term sales to industry may be used to fill the free capacity during the summer.

This cyclical pattern is very significant for gas marketing in systems in which normally a constant, close-to-supply capacity production of gas is required for geological and technological/economic reasons. It is also significant for the planning of transmission capacity, distribution, and storage systems used to transport the gas from the wellhead to areas of consumption and to individual consumers. The pattern imposes very stringent requirements for coordination of marketing and planning and investment in production, transport, and storage capacity. Depending on specific regional circumstances, these aspects may be more or less relevant, but in general they induce a great degree of risk and, thus, affect suppliers' and consumers' willingness to invest.

A crucial spatial element in the evolution of the supply systems during the second half of the 20th century was the phenomenon that, by and large, the distance over which gas was transported from the production location to the centers of consumption increased. As a principle, gas exploration, the development of reserves, and eventual production must take place as close to the location of gas consumption as possible. Transport is expensive and it reduces the flexibility of gas supply to meet the daily and seasonal variation in demand. Therefore, the development of gas resources should evolve in a pattern of concentric circles around the center(s) of gravity as regards consumption.

This pattern, of course, is not completely consistent with what is actually observed. Old fields maintain production because they are huge, because technological development facilitates a "second life," or because they are conserved for reasons of national resources policy. Nevertheless, most of the "greenfield" gas production will take place increasingly farther away, requiring longer pipelines and more compressors to move the gas through the pipelines. It

will also involve the pipelines passing through more countries. Moreover, in many places large volumes of gas are produced in joint production, dissolved in crude oil. The characteristics of the joint production of gas and oil traditionally resulted in major difficulties in coordinating the output and marketing of both types of output. Often, the commercialization of these volumes of gas was not economical because oil production took place far away from potential customers. The cost of transporting the gas by pipeline was prohibitive, so it was flared. However, expected shortages in gas in some markets such as the United States, the declining cost of liquefied natural gas (LNG) transport and handling, and environmental considerations may change this practice. Of course, if oceans are to be crossed, the whole concept of supply through pipelines must be abandoned and LNG is the only option. Thus, geographical patterns of gas supply may experience radical changes in the near future.

In response to these characteristics, a variety of contractual relationships were established between gas producers, transporters, and consumers. In these contracts, the volume and the price risk were reduced by fixing prices and establishing specific terms of trade for a longer time period so that the producers' as well as the consumers' investments would not be jeopardized. Often, it involved a role for public entities, taking ownership and management over parts of these systems. These factors have resulted in the development of truly regional gas markets, each with its own market structure, characteristic institutional framework and role for governments and local authorities, and specific outcome in terms of the economics of supply and demand.

Traditionally, in the U.S. system, strict state and federal regulation stabilized a market with predominantly private firms. In Europe, a variety of national arrangements were unified within an international system of trade and supply. This Europe-wide system was highly coordinated under the aegis of a few governments of gas-producing countries and a limited number of gas-producing oil companies. In the United Kingdom, a single state-owned distribution company, British Gas, purchased all gas from private and public offshore producers and supplied it to the consumers. In Japan, LNG was imported under long-term contracts between buyers consortiums of Japanese gas and power companies and foreign suppliers, often state oil and gas companies from producer countries or the international majors. In the Soviet Union, the planning bureaucracy stabilized the market. In Latin America, stabilizing regulatory arrangements were established in the countries in which gas developed as a significant source of energy, namely Argentine, Chile, and Venezuela. The following sections illustrate the specific patterns of market coordination in the United States, Europe, and the Asian Pacific area (Figs. 1 and 2).

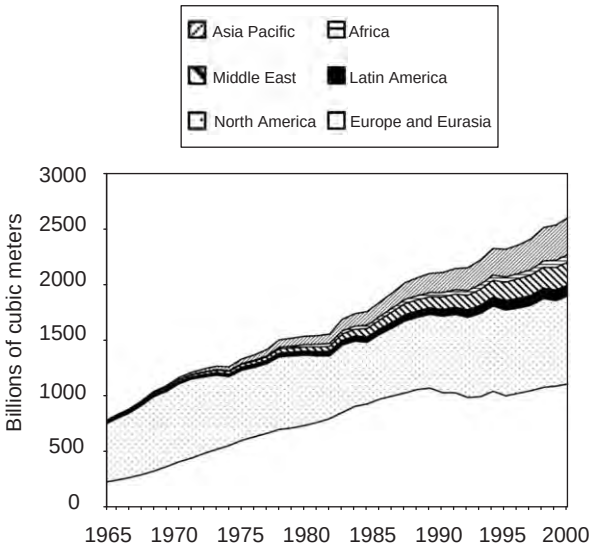


FIGURE 1 Natural gas consumption by region.

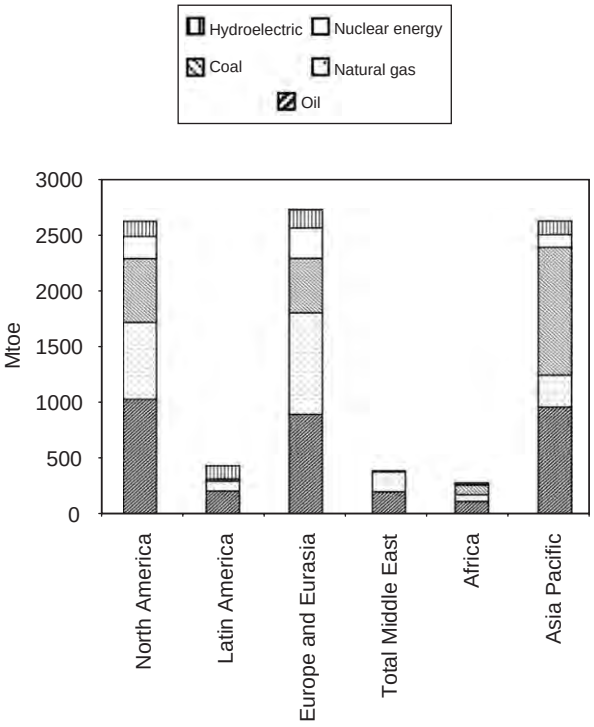


FIGURE 2 Primary energy consumption by source and region.

2.1 The United States

The U.S. natural gas industry is the largest gas industry in the world and it comprises a huge number of fields of various sizes. Initially, the development of gas use was haphazard and locally confined. After a period of limited oversight, with mainly municipality involvement, the Natural Gas Act of 1938 established a basis for the regulation of gas prices and the activities of companies. The following period saw a rapid expansion of the industry and the industry's regulation. Interstate transactions came under regulation by the Federal Energy Regulatory Commission (FERC). Intrastate transactions came to be regulated by the state public utility commissions. The industry was vertically separated into production, pipeline transportation, and distribution. However, all transactions were regulated and took place under long-term contracts. Thus, the industry was vertically integrated. Most of the gas produced was sold by the producers to pipeline companies under long-term, take-or-pay contracts. These pipeline companies supplied the distribution companies also under long-term contracts.

Attempts to regulate wellhead prices at a low level, however, discouraged gas exploration and production. Also, the transportation and distribution systems became private monopolies, imposing high costs on consumers. This delayed the growth of the industry considerably in a period in which enhanced indigenous energy supply was seen as crucial. The costly gas shortages in the 1970s prompted Congress to embark on a reform of the gas industry. In 1978, Congress adopted the Natural Gas Policy Act, through which the FERC was enabled to liberalize interstate natural gas markets.

2.2 Europe

Crucial to the development of the European system—and different from conditions elsewhere—was the ability to produce flexible load curves from the Groningen field in The Netherlands. Because of its geological structure and because of large investments in wellhead production capacity, the field had a capacity of approximately 100 billion m^3 annually. Daily capacity was sufficient to supply the required peak demand for the whole northwest European gas system in the early 1970s. On this basis, the European supply system was developed. Initially, The Netherlands supplied the lion's share of all requirements, to be sold to Germany, Belgium, and France.

However, during the 1970s, new foreign suppliers appeared, attracted by the high revenues from gas sales in Europe. The producers in these countries undertook the exploration and production activities and sold the gas produced to the state-owned wholesale buyer and exporters, such as Gasunie in The Netherlands, British Gas, GFU in Norway, Gazprom in Russia, and Sonatrach in Algeria. These firms coordinated the production and sales both nationally and abroad with other suppliers. They sold the gas at the borders to transmission companies and/or traders that transported the gas throughout Europe and supplied it to the national and local gas distribution companies and large industrial customers. With the exception of Germany, every country in Europe had one transmission company, operating the whole national high-pressure transport system. The local, municipal, or regional distribution companies operated and owned the local, low-pressure network, through which the gas was transported to small domestic users and businesses.

Trade between the several parties in the gas chain was arranged through medium- or long-term contracts. The producers delivered the gas via long-term contracts, (15–20 years) to transmission companies. These companies had medium-term contracts (1–5 years) with the local distribution companies and large-scale users. The small users had exclusive supply agreements with their local distribution companies. Throughout Europe, the transmission companies and local distribution companies had been given a local monopoly for pipeline operation and distribution.

By means of concessions, public ownership, indicative planning, and regulated prices, the producers and distributors were able to coordinate purchase and sales volumes of gas and prices. Through these contractual provisions, the risk involved in the funding of the expensive production and transport facilities over the longer term was reduced. This, of course, stimulated enormous investment in the system, as long as gas could be produced and transported at such a cost that the end-use price allowed for a reasonable profit. An important element of the contracts was the so-called take-or-pay clause, obliging the distributor or consumer to pay even when the whole volume of gas agreed upon had not been taken from the system, for example, because of lagging demand.

The only competition to gas emerged from the threat that consumers would switch to other energy sources, of which oil was the most relevant. However, the end-use gas prices were always slightly less than local prices for substitutable oil products. Importantly, the contracts contained destination clauses through which the volumes of gas were

committed to a specific area and could not be sold elsewhere at a higher profit or undercut official sales. The revenues to the producers were thus dependent on the retail prices for oil products to the several types of end users (the netback principle). The transporters and distributors were remunerated with a fixed fee, which was dependent on the cost of the pipeline and the negotiation position of its operator (normally having a monopoly position).

2.3 Asia and the Pacific

Unlike the United States and the European natural markets, the Asian Pacific gas markets are essentially markets for LNG. Japan and other main LNG users, such as Korea and Taiwan, import most of their LNG under long-term contracts. Most of the gas is used in power generation, and relatively smaller shares, compared to the United States and the European Union (EU), are used in industry directly. The shares of domestic household use are comparable.

Traditionally, pipelines have been a difficult issue in Asian markets as a consequence of geopolitical tensions between supplier, transit, and consumer countries, sometimes in combination with problems of scale with regard to the size of the markets and relatively low prices for other fuels. As a consequence, the whole region consists of a number of national submarkets with their own dynamics, economics, and institutional frameworks.

The national markets in Japan, India, China, Taiwan, and Korea are heavily regulated. In some countries, natural gas is supported and subsidized, whereas in others gas is priced at its full cost. Both situations may hamper further development: The former because it may impose an additional burden on the subsidizing states, and the latter because it may scare away the (potential) consumers that have to make a choice for natural gas. Regarding the Asian natural market, there is a huge potential for growth in gas consumption in most countries, as well as in gas supply from a few of them. There are, however, major problems specific to each country in bringing about the expansion of the gas systems, not to mention a market.

3. THE LIBERALIZATION OF GAS MARKETS

As it is generally stated, liberalization of the gas industry envisages to reduce the costs of supply, enhance quality, and enhance the overall efficiency of the gas supply system. By providing the consumers, or traders, with a choice with respect to their suppliers and the type of contract, the customer is in a position to select the supplier offering the most attractive conditions. It is assumed that suppliers will try to gain, or protect, their market share by improving supply and price conditions. The traditional wholesale traders and distribution companies will have to develop new strategies and adjust their organizations because they will lose both their secured market, while the integration and coordination between their several activities in the supply chain disappears. These firms will also adjust their structure and operations—through mergers and acquisitions—so that their size and activities fit the requirements of the newly emerging competitive market.

Competition in the gas market can be achieved by altering the structure of the market, as suggested by economic theory. Whereas the traditional perspective denied the feasibility of competition in the gas industry, structural change started from the hypothesis that the introduction of competition would be possible in some segments and that this would improve the performance of the whole system. Actually, only the pipeline transportation and distribution segments of the industry were accepted as being a natural monopoly because of economies of scale and scope, high fixed costs of pipeline construction, and relatively low variable costs. The other segments, production and wholesale and retail trade, were considered to be potentially competitive markets. Therefore, by changing the ownership structure of the sector and by dividing the competitive segments into a number of different firms, competition could be introduced into the sector. A number of models can be distinguished to break up the industry.

One model separates production from the rest of the industry and introduces competition among producers. Producers sell natural gas to a gas utility, which then resells it to the end users. The transactions between the producers and the utility lead to the development of a wholesale natural gas market, in which natural gas is traded for further resale. Regulation is needed to restrict the market power of the gas utility relative to both the consumers and the producers. End-user prices are regulated. The price of gas sold by producers to the utility is determined through competitive bidding for a supply contract with the gas utility.

A more ambitious model introduces open access in pipeline transportation to third parties that either sell

or buy gas from each other. The gas utility thus supplies gas to the captive, small consumers and it provides transportation services to large consumers that purchase gas independently in the wholesale market. A gas utility may also be further separated into a pipeline company and several distribution utilities, all providing access to their networks. Open access promotes competition in the wholesale and end-use market. Producers benefit because of the increase in the number of potential buyers, whereas distribution utilities or large end users benefit from a greater choice in gas suppliers and conditions. The high transaction costs involved in buying gas create room for natural gas traders, which aggregate demand and supply for a number of smaller market participants by purchasing natural gas and transportation services on their behalf. In this model, the end users have to be protected from the monopoly power of gas utilities. The price of transport and other services is another crucial factor in achieving competition in the wholesale market.

A third model separates natural gas supply from pipeline transportation and distribution and introduces full competition into these markets. Unbundling creates a level playing field for all participants in the natural gas market and creates a large number of supply companies that purchase natural gas in the wholesale market and resell it downstream using the transportation systems of pipeline and distribution companies. As competition among suppliers reduces excess margins and forces the savings on to the end users, the need for stringent price regulation at the wholesale and retail level will be reduced. In this model, new flexible short-term trading and contractual arrangements will be provided to balance supply and demand and give market participants the flexibility they need. Liquid spot markets are expected to emerge, yielding prices that continuously reflect the market value of natural gas at a specific location.

The advantages of a more efficient market may be offset by the transaction costs, high risks, and high volatility in volume and price. It increases the uncertainty of demand for gas and transportation services, exceeding or falling below available capacity. In part, this may be resolved by contracts that offer secure or interruptible transportation services, storage and conversion contracts, and other techniques to moderate changes in supply and demand. However, it is too early to effectively evaluate the outcome of processes of structural change. Only in a very few countries, such as the United Kingdom, has the third model been implemented to its full extent. Other European countries are contemplating this step in several forms, and the United States is also moving in this direction.

It has become clear that a structural change of the gas market to introduce competition requires complex organizations and administrative systems to carry out and control all these transactions. Strict and independent regulatory oversight is required to keep owners of distribution networks or other parties from abusing their dominant economic positions. Indeed, such transport systems remain natural monopolies, of which the owners and operators are in the position to jeopardize others' interests.

3.1 The United States

In 1978, Congress adopted the Natural Gas Policy Act authorizing FERC to liberalize interstate natural gas markets. Since the major producing and consuming regions in are in different states, this had a major impact on the operation and efficiency of the natural gas industry. During the 1980s and 1990s, FERC gradually established a framework for the actions of market forces in the natural gas industry.

By Order No. 436, FERC instituted an open-access regime for interstate pipeline transportation. This regime enabled local distribution utilities and large end users to bypass pipeline companies' gas sales and purchase natural gas directly from producers. Although the open-access regime was voluntary for pipeline companies, it was widely accepted because it enabled them to increase the utilization of pipelines. However, large-scale implementation occurred only after FERC resolved the issue of how the costs of the transition to open access were to be distributed. Because customers were allowed to exit long-term supply contracts, the pipeline companies were left with large take-or-pay obligations to producers. FERC Order No. 500 allowed them to pass a share of the transition costs to procurers, distribution utilities, and end users. Subsequently, the Wellhead Decontrol Act of 1989 deregulated the wholesale price of natural gas in all interstate transactions, promoting competition in the wholesale natural gas market.

Order No. 636 required pipeline companies to unbundle natural gas sale operations from pipeline transportation activities and set up separate transportation and trading affiliates. This attracted many new companies into marketing and promoted fierce competition among marketing firms. Order No. 636 also reformed the regulation of interstate pipeline transportation. It allowed resale of transportation

contracts by shippers, creating a secondary transportation market in which shippers can purchase pipeline capacity from other shippers that have temporary or permanent spare capacity. This enhances an efficient allocation of transportation contracts among shippers and high utilization of natural gas pipelines.

A subsequent series of measures was designed to promote transparency and flexibility trading practices. It facilitated short-term capacity resale, shippers' choice of delivery locations on interstate pipeline systems, and the standardization of contracts and pipeline system operation. Additional rules guided the development of a short-term transportation market in which capacity and interruptible contracts can be traded among pipeline companies and shippers.

However, by the end of the 20th century questions began to arise about the future of the U.S. gas and energy markets. Fear began to arise that the supply of new, additional volumes of gas, in response to the continuous increase in demand, would not materialize. Low prices in the context of specific market circumstances, fiscal regulations, disputes about pipeline construction, and difficulties in the establishment of storage facilities all contributed to an awkward situation with regard to gas supply. The difficulties in the operation of the market mechanism seem to be a consequence of the fact that excess production, transport, and storage capacities for gas and substitute fuels have been eroded, as illustrated by a number of local and regional energy crises. The workings of the market seem to have reduced the incentives for private industry to maintain sufficient spare capacity. It remains questionable whether this phenomenon is part of a learning experience, or whether the energy is structurally unable to establish long-term balanced behavior.

3.2 Europe

The situation in Europe has not reached that of the United States. There are no signs of immediate capacity shortages in the gas industry, although the situation in the United States is raising questions. The main issue, however, seems to be the progress of the liberalization of the EU gas market and the lagging implementation in several member countries. During the 1980s, the perception that the economies of the EU countries would be better off if linked, through integration and/or internationalization, was gaining widespread recognition as European countries were increasingly confronted with large internationally operating firms from the United States and Japan selling goods and services in their relatively small and often somewhat protected home markets. In 1985, Jacques Delors, supported by powerful European industrial leaders, launched the Single European Market initiative. The aim was to achieve a free market for the whole community by dismantling impediments to the free movement of goods, services, labor, and capital.

The EU gas directive (98/30/EC) mandated that member states of the EU had to adjust the organization and regulation of their national gas sectors in such a way that consumers would receive the right to contract with national or foreign gas suppliers of their choice. It had fundamental consequences for the structure of the existing gas market and the way in which the gas was produced and moved to the final consumer. First, the member states had to open up a first tranche of 20% of the national market and after 10 years 35% of the market had to be open for supplies by third parties.

To facilitate supplier and trader access to consumers, two main changes were necessary. First, the EU member states had to adjust their rules so that producers would no longer be forced to sell their production to one national wholesale gas buyer. The EU hydrocarbons directive stated that producers were to be given the right to sell their gas to anybody interested and acceptable to them, including large consumers, intermediate traders, and retail sellers. To achieve this, a second important requirement was that the existing transport and distribution network had to be made accessible to third parties. Producers, traders, and consumers should be given the right to contract transport capacity in these systems to transport the gas to the customer at the end-use location. Instead of selling the gas to the network owners and traders, the producers and independent traders had to acquire transport and distribution services for a fee.

By 2003, the gas directive has been implemented by the EU member countries only partially and to a varying extent. In part, this is a consequence of the fact that the EU countries are free to choose how to implement the provisions of the directive. It provides a framework, leaving the actual instruments of regulation, including the way in which regulatory oversight is created, to the discretion of the member states. This has created a situation in which some of the countries, such as the United Kingdom, The Netherlands, Spain, Austria, and Denmark, are fairly advanced in opening up their markets. Also, the trading instruments referred to in the discussion of

the U.S. situation are being developed in these countries.

However, other countries are delaying the implementation of the gas directive or the national adoption of an effective regulatory framework. Thus, access provisions to the several national EU markets are quite different and the market positions of the several companies involved are unequal. In addition, there are important differences between countries' perspectives on issues such as security of supply, in light of their dependence on a limited number of suppliers, and the recognition of the interests of gas-producing countries, such as The Netherlands and Norway, vis-à-vis the consumers.

Unless Europe radically changes its position on nuclear energy, increased dependence on gas imports will be unavoidable for the foreseeable future, with a possible share of gas in the EU primary energy portfolio of approximately 30%. In principle, there is enough gas in the area to meet Europe's additional demand for energy for the foreseeable future. However, doubts have been cast over this supply perspective because it is maintained that policy-makers in the consumer countries fail to accept the conditions necessary to bring this gas to the EU markets. The size of the investments required, their timing, and the quantities of gas involved in the new developments are such that producers and suppliers from outside the EU cannot undertake these ventures without a high degree of certainty that the gas will be taken by the market at the time it is produced. This implies that such supplies need to be carefully arranged with the contractual (and other) support of those buyers in the market that have the ability to evaluate and aggregate the many small parcels of demand in local markets.

During the past 40 years, the European gas industry obtained security of demand by means of long-term supply contracts with provisions on a take-or-pay basis, providing limited flexibility, and price indexation ensuring the competitiveness of gas in the specific markets. As noted previously, this industry structure is being dismantled and a new system is being introduced that focuses on short-term rather than long-term deals. This is causing significant uncertainty regarding the market and the business environment. Gas-producing countries, particularly Russia, Algeria, and Norway, and some of the large oil companies and financing groups involved are showing signs of confusion and discontent with these developments. For example, at conferences in Algiers in May and September 2002, with the exception of Norway and The Netherlands, it was made clear that the concepts that have emerged from the process of liberalization are not helpful in creating an environment for an expansion of the gas market.

4. THE FUTURE OF NATURAL GAS: NEW PATTERNS OF SUPPLY AND DEMAND?

This examination of the structuring of markets for natural gas has illustrated the importance of the way in which governments, alone or together, are involved in the organization of these markets. Although natural gas has a number of inherent advantages as a clean, versatile, and easily controllable fuel, its development also involves fundamental problems. In the past, the establishment and management of the newly developing markets for gas in the United States and Europe occurred essentially under a system of tight coordination of the economic, technical, and institutional components of these systems. The rationale for this was clear: Private investors simply could not afford to take the technical, financial, and market risks.

During the past 25 years, it has been argued that times have changed and that the state should withdraw from such ventures. However, as argued in this article, a number of experiences have posed new questions with respect to the feasibility of markets as a coordinating device for large-scale, technologically and (geo)politically complex systems such as international gas markets. It remains to be seen to what extent the development of increasingly advanced risk sharing and trading mechanisms, and regulatory instruments to mimic market behavior in parts of such systems, may be sufficient to overcome the need for public regulation. Although the pseudomarket may be a forceful means to slim down a sector, it may not necessarily be helpful in establishing and expanding the complex gas production, transmission, and distribution systems that will be required to support growth of gas use in North America, Europe, Asia, and Latin America.

SEE ALSO THE FOLLOWING ARTICLES

Market-Based Instruments, Overview • *Market Failures in Energy Markets* • *Markets for Biofuels* • *Markets for Coal* • *Markets for Petroleum* • *Natural Gas Industry, Energy Policy in* • *Natural Gas Processing and Products* • *Natural Gas Resources,*

Global Distribution of • Natural Gas Transportation and Storage • Oil and Natural Gas Drilling • Oil and Natural Gas: Economics of Exploration • Oil and Natural Gas Leasing • Oil and Natural Gas Resource Assessment: Classifications and Terminology

Further Reading

Adelman, M. A. (1962). The supply and price of natural gas. *J. Ind. Econ.* (Suppl.).

Arentsen, M. J., and Künneke, R. W. (2003). "National Reforms in European Gas." Elsevier, Amsterdam.

BP (2003). "BP Statistical Review of World Energy, June 2003: A Consistent and Objective Series of Historical Energy Data." BP, London.

Correljé, A., Vander Linde, C., and Westerwoudt, T. (2003). "Natural Gas in The Netherlands: From Cooperation to Competition?" Oranje Nassau Groep/Clingendael, Den Haag, The Netherlands.

Davis, J. D. (1984). "Blue Gold: The Political Economy of Natural Gas. World Industries Studies 3." Allen & Unwin, London.

Estrada, J., Moe, A., and Martinsen, K. D. (1995). "The Development of European Gas Markets: Environmental, Economic and Political Perspectives." Wiley, Chichester, UK.

Juris, A. (1998a). "The Emergence of Markets in the Natural Gas Industry, Policy Research Working Paper No. 1895." The World Bank, Private Sector Development Department, Private Participation in Infrastructure Group, Washington, DC.

Juris, A. (1998b). "Development of Natural Gas and Pipeline Capacity Markets in the United States, Policy Research Working Paper No. 1897." World Bank, Private Sector Development Department, Private Participation in Infrastructure Group, World Bank: Washington, DC.

Mabro, R., and Wybrew-Bond, I. (eds.). (1999). "Gas to Europe: The Strategies of the Four Major Suppliers." Oxford Univ. Press, Oxford, UK.

Newbery, D. M. (2001). "Privatization, Restructuring and Regulation of Network Utilities." MIT Press, Cambridge, MA.

Odell, P. R. (2001). "Oil and Gas: Crises and Controversies 1961–2000. Vol. 1: Global Issues." Multi Science, Brentwood, UK.

Odell, P. R. (2002). "Oil and Gas: Crises and Controversies 1961–2000. Vol. 2: Europe's Entanglement." Multi Science, Brentwood, UK.

Peebles, M. W. H. (1980). "Evolution of the Gas Industry." Macmillan, London.

Stern, J. P. (2001, November). Traditionalists versus the new economy: Competing agendas for European gas markets to 2020, Briefing Paper No. 26. Royal Institute of International Affairs, Energy and Environmental Program, London.

Wybrew-Bond, I., and Stern, J. (eds.). (2002). "Natural Gas in Asia: The Challenges of Growth in China, India, Japan and Korea." Oxford Univ. Press, Oxford, UK.

Markets for Petroleum

M. A. ADELMAN
Massachusetts Institute of Technology
Cambridge, Massachusetts, United States

MICHAEL C. LYNCH
Strategic Energy and Economic Research, Inc.
Winchester, Massachusetts, United States

1. Oil Markets
2. Natural Gas

Glossary

forward market A market in which physical crudes are traded for future delivery, usually only a few months into the future.
integration (vertical) The degree to which companies participate in upstream and downstream aspects of the industry (crude production, shipping, refining, and marketing).
liquefied natural gas (LNG) Methane that is cryogenically cooled to form a liquid for easier storage and transport.
marker crudes Large-volume crudes that are widely traded and referenced in other contracts.
middle distillates Light petroleum products that are heavier than gasoline, primarily used as diesel fuel and heating oil.
methyl tertiary butyl ether An octane enhancer used in gasoline, manufactured from methanol.
New York Mercantile Exchange (NYMEX) The largest futures market for petroleum.
resid Residual fuel oil; also called distillate fuel oil 6 or C (in Japan).
sour/sweet Degree of sulfur content, where high sulfur means more sour.
take-or-pay A contract requirement that the buyer accept a specified amount of the contracted volume or pay for it.
upgrade To improve the value of the crude or product by cleaning or lightening it through a variety of processes.
Very Large Crude Carriers Supertankers.
West Texas Intermediate (WTI) The U.S. marker crude.

Two major changes characterize energy markets of the early 21st century: globalization and growing competitiveness. Where markets used to be fragmented and opaque, in part due to heavy regulation, they have become increasingly connected to one another at the same time that transparency and competition have increased. This reflects a combination of falling transportation costs and policy changes as many governments pursue economic reform. In addition, as economic growth spreads around the world, there are many more companies large enough to operate on a global or at least regional basis.

1. OIL MARKETS

The world petroleum market remains the most important of the energy markets, in terms of both volume and global penetration. The ability of oil to provide energy in a wide variety of uses, combined with the high energy density that most petroleum products contain, makes it the most flexible and useful of fuels. Although prices are relatively high in 2003, oil has long been one of the cheaper fuels, especially after the discovery of the supergiant oil fields in the Middle East.

Until 1970, oil consisted of the U.S. domestic market, which had many small producers, refiners, and marketers, and the international market, which was dominated by a small number of multinational corporations, known as the Seven Sisters. Before World War II, and for nearly two decades afterward, those companies controlled almost all oil that was traded on the international market, producing it, shipping and refining it, and marketing it in the major consuming countries. Although some oil was sold on long-term contracts to independent companies or partly owned affiliates, most oil remained within the systems of these large companies.

The situation has completely changed in the past two decades, especially since the oil crises of the

 © 2004 Elsevier Inc. All rights reserved.

1970s. A large portion of the world's oil is now produced by national oil companies, especially in the Organization of Petroleum Exporting Countries (OPEC) nations. Some of that oil is sold on the world market, some is sold on longer-term contracts to individual companies, and some is sold to their former concessionaires. In addition, a host of smaller companies, some privately held and others controlled by governments—mostly in the Third World—produce and sell oil on a national, regional, and occasionally global basis.

1.1 Spot Markets

One of the major changes in the world oil market in the past two decades, compared to most of the 20th century, is the rise of the spot market. Although the United States has long had a very active internal trade in petroleum, given the abundance of small producers, on a global level the amount of crude traded on the spot market has been relatively small. Oil products usually are traded only regionally, because of the higher cost of delivery. Gasoline and middle distillates especially require clean tankers so that they do not become contaminated and this has relegated their trade to local and regional markets. Rotterdam, the Mediterranean, and to a lesser extent Singapore have long seen a relatively active spot trade in oil products.

However, because the number of small producers has increased so much, more oil has entered the world's spot market. Additionally, many of the larger oil companies no longer keep their production within their own systems. A major company might sell African production into the spot market and buy spot crude in Asia rather than ship their own production between those markets, depending on the respective prices of crude and shipping. Also, advances in refinery technology have made it much easier for refiners to handle different types of crude, so that they are much more able to utilize many types of crude.

As a result, where spot crude trade was 3 to 5% of world oil markets until the 1970s, it is now estimated to be 30 to 35% on a regular basis. This could change over time with the rise of the new players in the market, such as China National Petroleum Co. and Petrobras, in Brazil, which have different goals and behavior patterns from the U.S. and European privately held oil companies. Increasing integration on the part of OPEC countries might also eventually reduce their sales on the open market.

1.2 Futures Market

One new development (actually a revival of a 19th century market) is the use of financial futures to hedge and/or speculate on the price of oil. This resumed in 1978, with the trading of heating oil in the United States on the New York Mercantile Exchange (NYMEX). Most futures trading takes place on the NYMEX and consists primarily of contracts for sweet crude [West Texas Intermediate (WTI)], gasoline, and middle distillates. Although this is a very active market, it is much less developed than those for many other commodities. The amount of open contracts on a given day is only a few times daily physical trade, whereas for many agricultural commodities, the trade is several times higher. Additionally, trading beyond 3 to 6 months is very thin.

In part, this reflects the fact that this market is still, relatively speaking, a new one and many of those active in the physical market are very large and thus are reluctant to use futures to hedge their trading, for fear their volumes would overwhelm other traders and thus drive the market. Actors such as OPEC producers also have particular foresight into future oil prices, since their decisions directly affect prices, so that any trading they undertook would influence the market. If the Saudis were to go short, for example, traders would take that as evidence that they were planning to raise production and would sell along with them. This is different from that of many other commodities, where much of the uncertainty reflects factors that are beyond the control of the players in the market, such as weather and economic uncertainty, rather than policy choices by a cartel.

However, the futures market does provide substantial transparency, especially combined with the Internet. Virtually everyone in the world can have access to oil prices on an instantaneous basis. Compared to the 1970s, when world oil prices were rarely published, this is a huge difference. It has been argued that the existence of a futures market can, by itself, influence oil prices, as speculators will tend to behave in a herd and often move in illogical ways, that is, reacting to psychological events such as headlines rather than fundamental indicators such as oil inventories. There is some truth to this, but the impact tends to be exaggerated; foresight is still rewarded and mistakes penalized, on average. Deviations from the "true" price of oil tend to be relatively small and brief and it is usually the case instead that observers are attempting to rationalize oil price movements that they cannot explain otherwise.

1.3 Price Setting and Volatility

For most of the history of the world oil market, the price of oil has been set or at least strongly influenced by one or another body. This began in the late 19th century with Rockefeller's Standard Oil Trust, which effectively monopolized the market by setting the price at which it would buy crude. After the Trust was broken up by the U.S. government, prices became more volatile, until two particular steps were taken to stabilize them. In Texas in the 1930s, the state government gave control over production to the Texas Railroad Commission, which set quotas for individual producers based on its estimate of the market situation. Other producing states in the United States followed their lead and imposed similar quotas. Since the United States represented the bulk of the world oil market through the middle of the century, this had the effect of stabilizing the world crude oil price, as the export price equaled the domestic price.

This was especially true since the world oil trade was dominated by the Seven Sisters, or major oil companies. They established a price known as "Gulf plus," which meant they would pay the price of crude oil in the Gulf of Mexico plus the cost of transportation to the point of sale. Prior to World War II, this was economically logical as Gulf crude represented the marginal barrel. In order to protect this price, the majors agreed to limit their production to avoid undermining crude prices. Since most production growth after the 1930s came from the Middle East, where consortia consisted of interlocking major oil companies, it was very easy for them to cooperate in this manner. (Earlier, they had been inclined to poach on one another.) In fact, the consortia had the power to penalize any partner who wanted additional oil that would allow them to take market shares from a competitor. It was only in the 1960s, as independent oil companies began to break into new areas such as Libya, that the system began to erode.

In the 1950s, the world oil price was often referred to as the posted price; however, this was increasingly a tax reference price. Companies paid taxes to the governments based on the posted price, but the actual price was determined by their tax payment plus the cost of production. As the cost of production was declining, actual prices increasingly diverged from the posted price. But since most crude stayed within single-company systems and was not traded openly, there was very little transparency in the market.

By the middle of the 1970s, however, most OPEC nations had completely nationalized the producing operations in their own territories and thus the posted price no longer represented the tax on production, but rather the actual sales price set by the governments. At that time, the official sales price came into wide use. OPEC nations would hold meetings to agree on a standard price, which they would then offer to all buyers. Production then fluctuated according to changes in sales, but prices remained relatively steady. Thus, Fig. 1 shows that

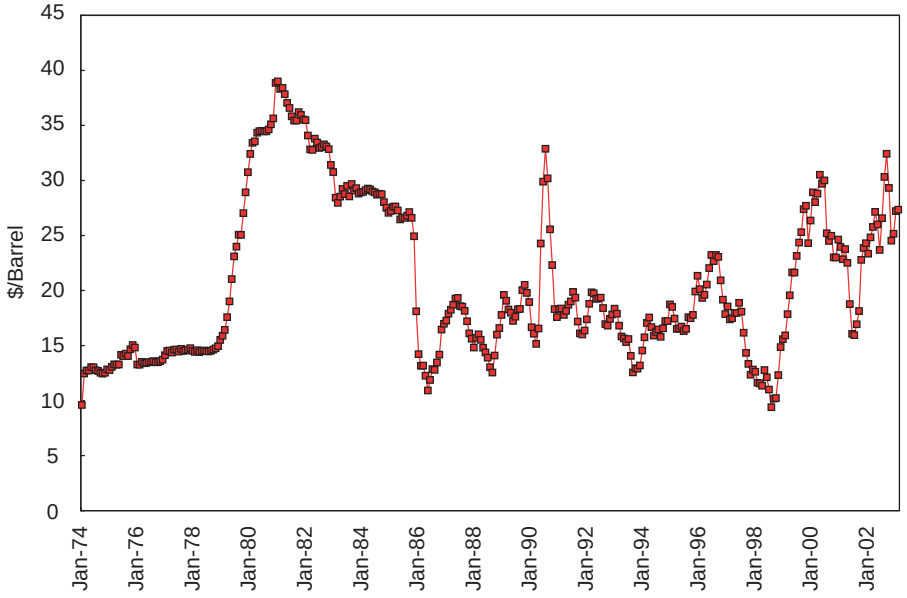


FIGURE 1 U.S. refiners' acquisition cost for imported crude (nominal dollars). From Monthly Energy Review, U.S. Department of Energy.

the world price moved very little from month to month during this period, except for the Iranian oil crisis in 1979.

The system broke down as a result of OPEC setting the price too high for too long a period after 1979, causing a substantial drop in the demand for OPEC oil, the impact of which was heaviest on Saudi Arabia, whose cash sales had practically vanished by 1985. During this period, the Saudis had been acting as the cartel's swing producer, meaning that they would absorb seasonal and short-term fluctuations in demand for OPEC oil. But they found themselves actually absorbing the long-term decline in demand for OPEC oil instead and had to initiate a price war to force the other members to both lower the price of crude oil in the long term and accept a greater share of the burden.

Since that time, prices have fluctuated much more on the world market, as Fig. 1 shows. The cartel is now allowing price to balance the market rather than their production. Since there is substantial uncertainty about global oil supply demand and inventories in the short term, the cartel is always setting a target and achieving it imperfectly.

In the past few years, prices have become even more volatile, reflecting changes in the overall market. The oil crises of the 1970s resulted in a huge amount of surplus capacity in OPEC, as well as in the global refining and shipping industries. Over time, this declined (Fig. 2), and there has been less surplus capacity to deal with short-term market tightness. Instead, prices have had to balance the market on several occasions, including 1996 and 2000. In both of those cases, most OPEC nations were producing at full capacity and the market did not balance. (Of course, the Saudis could have raised production further in both cases, but chose not to do so.)

1.4 Marker Crudes

Although crude petroleum is thought of as a fungible commodity, that is, one within which units are highly interchangeable, there are substantial quality differences in different crude types, which translates into differences in value. There are three primary aspects of quality: weight, indicating the amount of carbon relative to the amount of hydrogen; sulfur content, which is often referred to as sweetness (low sulfur) or sourness (high sulfur); and contaminants, particularly heavy metals, such as nickel and vanadium. Typically, a given field will contain one particular type of crude whose qualities are well known to buyers and even a region can produce sweet or sour crude, heavy or light crude, and so forth. Also, it is

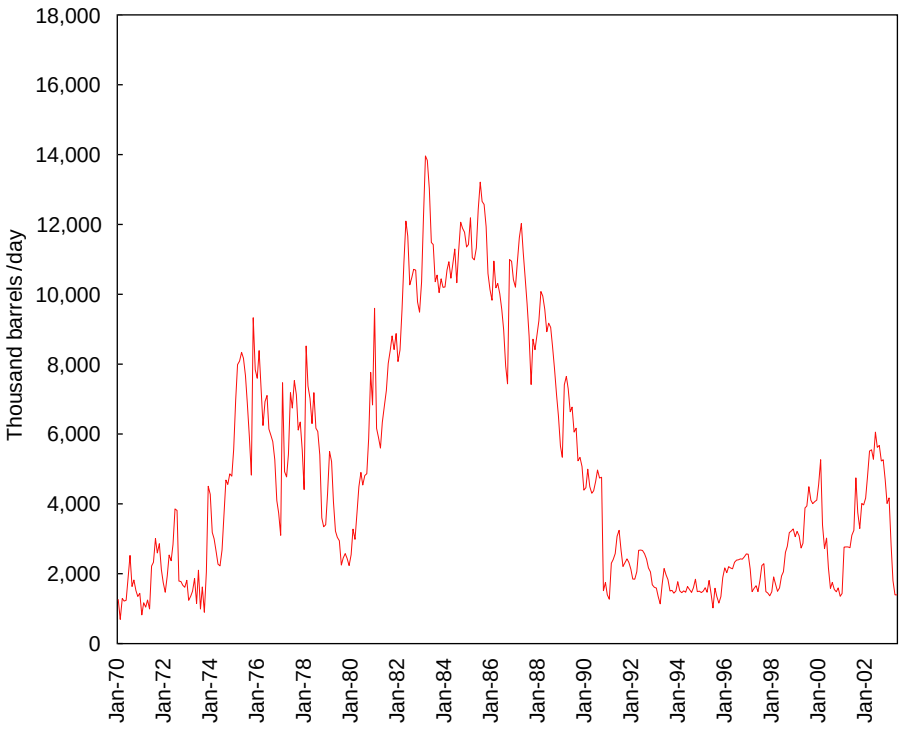


FIGURE 2 OPEC surplus capacity (thousand barrels/day). From U.S. Central Intelligence Agency and Energy Intelligence Group.

common for crudes produced in an area to be blended so that buyers do not have to differentiate between minor differences in small volumes.

Since it is difficult to track all of the many prices for crudes according to quality differences, many contracts rely on what is called a marker crude. In Europe, Brent crude is used, whereas in the United States, WTI is the primary marker crude and Asian buyers and sellers tend to rely on Dubai crude. All are traded in the open market and have well-publicized prices. Most contracts will fix prices relative to those crudes, adjusted according to differences in the quality of the crude being traded.

However, these crudes are increasingly suffering from problems of declining volumes, which leaves them vulnerable to squeezes that can cause their prices to diverge from the overall market. This has been a problem particularly for Brent crude produced in the British North Sea, which is actively traded in the forward market, and where the number of cargoes is approximately 24 per month. It is thus possible for a company to corner the market in a given month, leaving players in the forward market unable to find physical crude cargoes.

Some have suggested that the Saudis once again accept the role as a marker crude for its Arab Light, which is one of the largest volume crudes produced, but the Saudis have refused and do not allow their crudes to be traded in the spot market as a rule. They fear that Arab Light would become the target of manipulation, possibly causing their prices to rise relative to the market, thereby hurting sales, as well as making in their role in the price-setting more obvious and thus subject to political criticism.

1.5 Trade Flows

Over time, the trade in petroleum has changed significantly. In the 1970s, for example, the United States began importing increasing amounts of crude from the Arabian Gulf. However, rising production in the North Sea and the Atlantic Basin caused a sharp drop-off in that trade. Europe relies heavily on Russian and African production for its import needs and Asian imports are primarily from the Middle East. The United States still imports significant amounts of Middle Eastern crude, but most of its import needs are met by short-haul Atlantic basin producers: Canada, Mexico, Venezuela, and West Africa.

This primarily reflects the difference in transportation costs, where differences as little as pennies per barrel will shift trade between regions. Although countries such as Mexico and Venezuela may prefer to avoid being overly dependent on sales to the U.S. market, they receive much more for their production by not having to pay the higher transportation costs into more distant markets. At the same time, some oil-exporting nations continue to have an almost political concern about their market shares in certain areas, especially the U.S. market, where the Saudis sell approximately 1 million barrels of crude a day at a loss (compared to prices in Asia) to retain a foothold. It is hard to know how much this reflects a marketing strategy, such as diversification, versus a diplomatic tactic, maintaining the so-called "special relationship" with the United States.

At the same time, differences in demand and a variety of market fluctuations mean that the price in different markets can vary over time by as much as $2 per barrel for similar crudes. This then encourages balancing shifts in the trade flows to offset the differential. However, the value of crude in a given market is a reflection of a combination of factors, such as type of product demanded in the region, refinery sophistication, and openness of markets. However, isolated markets such as California are more likely to face higher prices on a regular basis.

Figure 3 shows the oil trade flows in the year 2002.

1.6 OPEC Integration

The tendency of OPEC national oil companies to move downstream is also an important recent change. Especially since the early 1950s, most crude oil was shipped to consuming nations and refined there. This was partly in response to the problems that resulted from the Iranian nationalization of British Petroleum holdings, including the Abadan refinery, in 1951, which led to a policy of "refining on the consumer's doorstep" in many industrialized nations. Also, the development of Very Large Crude Carriers made it cheap to ship very large quantities of crude.

However, as the OPEC nations gained control over their own operations, and especially as their income rose, they decided to refine their own oil where possible, partly as a corporate strategy and partly as an economic development strategy. Seeking to capture the larger share of the value added to their raw materials, OPEC countries first sought to build export refineries on their own territory and then to buy refineries overseas in consuming nations. As Fig. 4 shows, their own refinery capacity rose sharply in the late 1970s and early 1980s and then leveled off as their oil revenues weakened. Subsequently, they

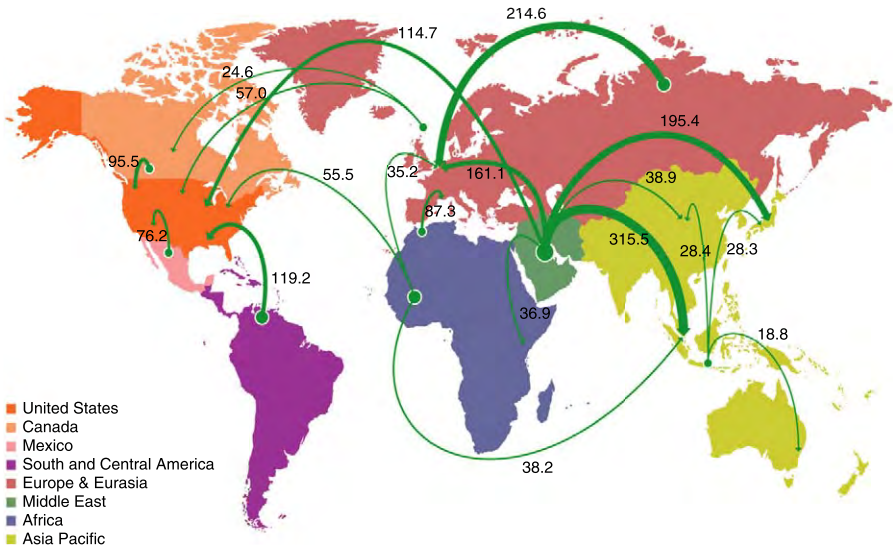


FIGURE 3 World oil trade flows (in millions of tons). From BP Statistical Review of World Energy Industry.

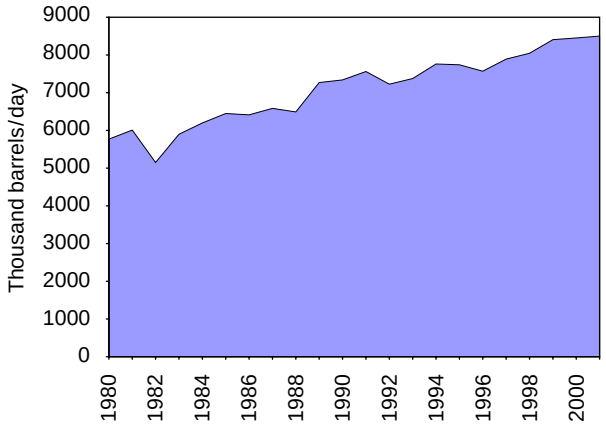


FIGURE 4 OPEC refinery capacity (thousand barrels/day). From OPEC.

realized that it was cheaper to buy into existing, underutilized assets in consuming nations, and Venezuela, Kuwait, and Saudi Arabia in particular have led in this regard, owning part or all of numerous refineries, especially in Europe and the United States.

1.7 Regional Differences

The wide variety of differences in energy usage around the world reflect a combination of economic circumstances and geography. Most important has been the abundance of commercial hydrocarbons in the United States, which meant that most stationary energy needs, such as for power generation and industrial boilers, were met by coal and later natural gas, so that petroleum was used more in the transportation sector than in many other countries. This trend continued when nuclear power became an important source of electricity beginning in the 1970s. Many other parts of the world switched from coal to oil for large users after World War II and have only just begun moving toward a more balanced mix of energy (Fig. 5).

The ready availability of cheap energy has also meant that U.S. industry and consumers have come to anticipate the ready availability of cheap energy. As Fig. 6 shows, this still remains the case. In most parts of the world, most types of energy are more expensive than they are in the United States, with a few exceptions.

Certainly, it is true that taxes play a role in raising energy prices in many parts of the world. Many countries tax gasoline at much higher levels than the United States, with gasoline taxes often higher than the final price of gasoline in the United States. But also, subsidies for coal long made it expensive in some parts of the world, such as Germany, although they are diminishing, and the high cost of natural gas transportation as well as past contracting practices have made that fuel expensive in Europe and much of Asia.

The higher rate of taxes on gasoline stems from a combination of reasons: energy security, fiscal policy, and simple protectionism. At the same time, diesel fuel is often taxed at much lower rates, which is why diesel automobiles are much more common in Europe than the United States. One side effect of this is that Europe has a surplus of gasoline produced, which it ships to the United States, whereas

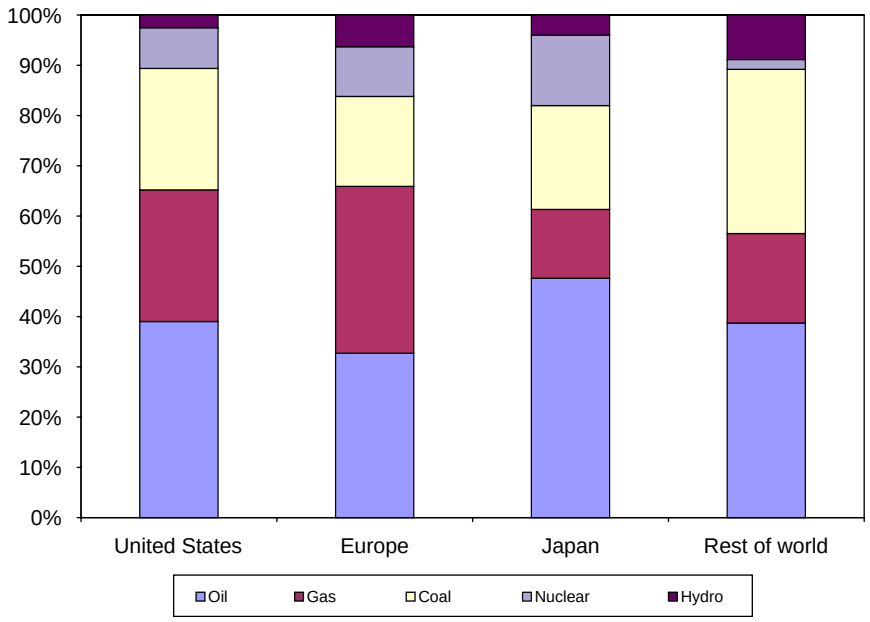


FIGURE 5 Primary energy share by region, in the year 2002. From BP Statistical Review of World Energy Industry.

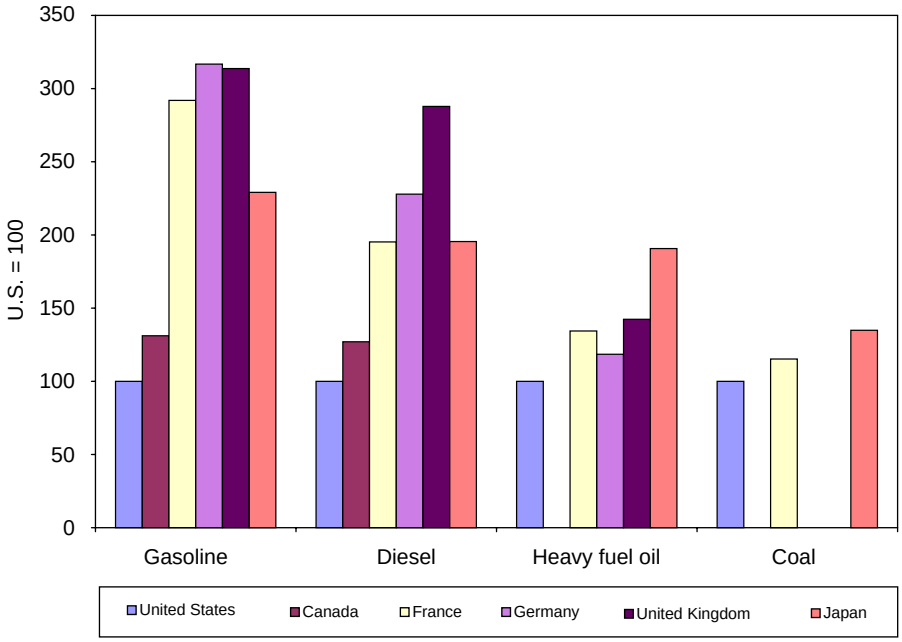


FIGURE 6 End-user prices in various countries (indexed to the United States). From International Energy Agency and Monthly Energy Review, U.S. Department of Energy.

the United States produces a surplus of middle distillates, which it exports to Europe.

1.8 Product Markets

There are a large number of products derived from crude petroleum, including many minor specialty chemicals that are not traded in quantity. But the three primary products, gasoline, middle distillates—including diesel fuel and heating oil—and residual fuel oil, make up the bulk of world oil traded, $\approx$75%. Each of these serves somewhat different markets, given their chemical nature and the competition they face.

Gasoline is used almost exclusively in transportation. Its volatility makes it explosive and unusable in many applications and its high price makes it less competitive with other fuels, such as coal and natural gas. Thus, it is rarely used in stationary applications, such as boiler fuel or to operate backup power generators, as diesel is. Because of its use in internal combustion engines, it is necessary for gasoline to remain very clean and uncontaminated and as a result it can be moved only in specialized vehicles, which means that bulk transport is somewhat expensive. Some gasoline moves internationally, but at volumes that are much smaller than those for crude oil.

Middle distillates play a dual role because of their nature. They are less volatile than gasoline and thus they can be used in diesel engines, which rely on high compression to ignite the fuel rather than a spark plug. But also, they can be used in small boilers, especially as a heating fuel for households. Typically, higher sulfur middle distillates are used as heating fuel, a practice that will increase as more advanced engines require lower and lower sulfur fuels.

Residual fuel oils, which are the heaviest of the major products, are named after their original origin, that is, the part of the crude that is left in the barrel after the higher value products have been refined. These fuels tend to be thicker, are harder to handle, and thus are not used by individual consumers but instead are used in factories, power plants, and large vessels, in which case they are referred to as bunker fuels. It is also the case that many of the contaminants in the original crude oil are left behind in the residual fuel oil, so that they are the least wanted of the fuels, although additional processing can improve their value by upgrading a portion to light fuels, or removing pollutants.

Prices for the different fuels are determined in large part by their competition, where gasoline faces virtually no competition in the transportation sector, automobiles being unable to switch fuels. And although diesel engines cannot readily switch fuels, some users are capable of switching, such as utilities running diesel-fired turbines. Some turbines can be switched to natural gas, should it be cheaper, but most utilities also have the option of running different power plants depending on relative cost, in which case a diesel turbine might even be competing with coal plants. Residual fuel oil, or resid, competes directly with natural gas, as many boilers are capable of using either fuel, as well as coal, typically the cheapest fuel available for power plants.

However, the use of both residual fuel oil and middle distillates is somewhat limited in the United States, as many power plants face pollution regulations that make it difficult for them to reduce usage of natural gas in favor of those fuels, which tend to give off higher levels of sulfur and other contaminants. Even so, in recent years the United States has seen as much as a half million barrels a day increase in the usage of middle distillates and residual fuel oil when natural gas prices were high.

Other major petroleum products include jet fuel, which is obviously used primarily in the airline industry and as such is very sensitive to the economic situation rather than prices of competing fuels. Liquefied petroleum gas (LPG) is another fuel used in many parts of the world for heating and cooking, especially when there is no distribution network for natural gas. Rural areas rely on this quite a bit in developed nations and less developed nations often use it even in urban areas as a convenient source of fuel for households.

Petroleum products are not completely interchangeable by any means but there is a degree of flexibility at the refinery level. By use of catalysts, cracking, blending, and other processes, refiners have the ability to alter their product streams, thereby accommodating changes in the market somewhat. For example, U.S. gasoline production increases during the summer and the production of heating oil rises in the fall and winter.

Similarly, some products can be altered when markets dictate. A variety of products can be used in petrochemical plants, including natural gas, LPG, and naphtha (which is an ingredient of gasoline), and refiners can moderate the amount of jet fuel or middle distillate they produce when conditions warrant.

1.8.1 Product Quality

The past decade has seen a major change in the impact of the environmental regulations on product formulas, especially for gasoline and diesel fuel. Historically, there have been tight regulations on the sulfur content of residual fuel oil in some parts of the industrialized world since the early 1970s, and lead in gasoline was phased out in the United States beginning at that time. These measures were in response to local pollution and had a relatively small impact on oil product trade inasmuch as low-sulfur crudes produced residual fuel oil with a low sulfur content, so that it was regularly available. As demand for low-sulfur crude began to outstrip availability, desulfurization units had to be installed at refineries. On the other hand, lead was an additive in gasoline

rather than an inherent part of the product and could be replaced with other octane enhancers.

However, regulations have become both stricter and more widespread in the past decade, covering many parts of Europe and increasingly the less developed countries, requiring both lower sulfur content and also specific formulas for gasoline and diesel fuel. In the United States, there are numerous formulas for gasoline depending on the location and season, that is to say, the particular environmental problem being addressed. Factors such as benzene content, Reid vapor pressure, and sulfur content face more rigorous regulations.

This has required refiners to make two types of changes. First, they must reduce the amount of sulfur in the products been produced, but in many cases, additives are required to enable them to meet both octane needs and specific formula requirements. Also, one of the most popular additives, methyl tertiary butyl ether, is being banned in many places as a water pollutant, requiring refiners to search for other additives, such as ethanol, to replace it, although it is more expensive and difficult to handle.

Aside from requiring substantial investments in upgrading capacity, the effect has been to segregate markets in many places, as some products are not clean enough to sell legally in some markets. Most Third World refineries are not capable of producing the kinds of products needed in some parts of the industrialized world and, depending on the specific product requirements, product trade cannot take place even within a country at certain times of the year.

As a result, there have been a number of localized problems with product supply, as any difficulties at the refinery level in an area cannot be offset by nearby supplies because they do not meet the specific environmental regulations in that location. California is in the worst situation because of the severe air pollution problems and the lack of connections to other markets—products must be shipped in via the Panama Canal or across the Pacific Ocean, if they can meet the environment rules. The problem overall has been exacerbated by the rise in capacity utilization in the downstream industry (Fig. 7), which has reduced the amount of capacity available to meet unexpected supply problems.

2. *NATURAL GAS*

Natural gas is unusual, a premium fuel with high value to customers but low prices at the wellhead. Because of its very low density, transportation costs are extremely high. The result is that natural gas markets developed much later than those for other fuels and regional markets remain relatively isolated from one another, although that is slowly changing.

Before the development of welding technologies that allowed high-pressure steel pipelines to be constructed, global natural gas markets consisted primarily of synthetic natural gas manufactured from coal, called town gas, and the exploitation of local deposits particularly for industrial purposes. Synthetic gas had a relatively low energy content and

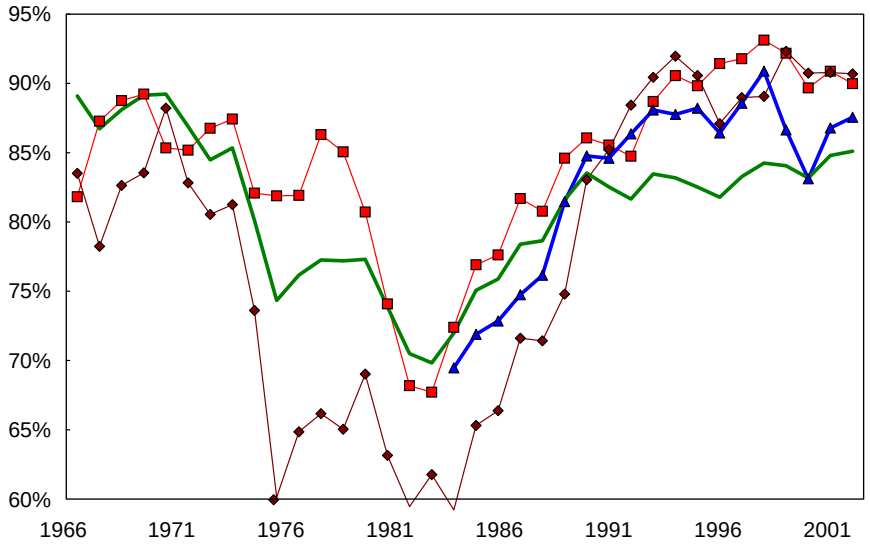


FIGURE 7 Refinery capacity utilization by region. Boxes, United States; diamonds, Europe; triangles, Asia/Australia/New Zealand; boldface line, world. From BP Statistical Review of World Energy Industry.

produced a large amount of waste products. It was used mainly in lighting in urban areas during the 19th century and began to be replaced by electricity in the early 20th century.

The early downstream petroleum industry—refining and petrochemicals—often relied on natural gas deposits as a source of cheap energy. Since long-distance transportation was impossible until the postwar era, most gas could be sold only locally and the lack of demand kept prices very low. Given this fact, it was usually not sought after intentionally, but found by those seeking oil, and often gas discoveries were abandoned unless there was a nearby facility that could use it. Some industrial centers developed specifically to take advantage of existing natural gas deposits that had no market, including the chemical industry in West Virginia and later Texas.

The U.S. natural gas market began to develop after World War II with the construction of numerous long-distance pipelines to bring the gas to consumers. European markets developed later, especially after the 1962 discovery of the supergiant Groningen field in the Netherlands, a low Btu gas that was extremely cheap to produce. Asian markets did not begin to develop until the 1970s, as high oil prices made liquefied natural gas (LNG) trade economically feasible, with the prices for gas and/or services set on a cost-plus basis. Fig. 8 shows the trade flows in 2002.

2.1 U.S. Gas Markets

The United States was the first to develop a large, well-articulated natural gas market because of the ready abundance of large amounts of gas resources and the early development of long-distance pipelines. The construction of oil pipelines from the Gulf of Mexico to the East Coast during World War II was a boon to the industry, as the two major pipelines were converted to natural gas after the war. But the ownership of large amounts of gas reserves by major oil companies and the nature of gas distribution as a natural monopoly raised concerns about possible abuse by reserve owners. In response, the U.S. government insisted that long-distance gas pipelines be owned by independent companies and that they be treated as regulated monopolies.

As a result, the U.S. gas market consisted of producers particularly in the Southwest who sold their production to pipeline companies that delivered and resold the gas to local distribution companies in the consuming areas. Since wellhead natural gas prices for the interstate market were regulated after 1954, this meant that the production, transportation, and distribution of natural gas were regulated by a combination of the federal and state governments.

However, since wellhead prices were allowed to rise only in response to inflation, the oil price boom of the 1970s made natural gas very competitive with oil in many areas and resulted in rising demand. At the same time, low prices discouraged drilling and supplies began to become scarce. The Natural Gas Policy Act of 1978 deregulated prices of some supplies and some pipeline companies signed grievously high-priced contracts for them, planning to roll the expensive gas in with their cheap, price-controlled gas. This proved disastrous, as weak demand meant that pipelines cut back on cheap gas in favor of high-

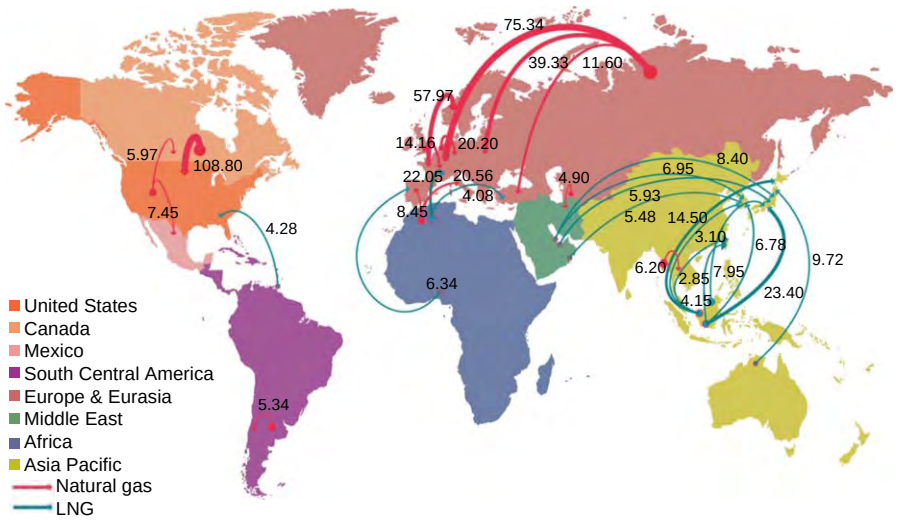


FIGURE 8 International trade flows of natural gas. From BP Statistical Review of World Energy Industry.

priced gas, for which they had guaranteed the volumes they would lift ("take-or-pay" clauses). The government deregulated all wellhead prices as a result, and the average price fell (Fig. 9).

By the 1980s, the long-distance pipeline business had also grown to the point where there were few areas that did not have multiple sources of supply, making it much more difficult to monopolize supply in an area. This led to the deregulation of the pipeline business and caused it to turn into a common carrier, rather than buyers and sellers of natural gas. It also led to the rise of natural gas traders, who located buyers and sellers of gas and then procured the necessary pipeline capacity to provide delivery.

At this point, the U.S. gas market split into the competitive, large end-user sector made up of power plants and other industrial facilities and the captive residential market, which remained regulated. The former could bid for natural gas on the open market and, having the option of switching to cheap residual fuel oil, could see its prices capped by that competing fuel. But residential users remained captive to the local distribution companies, having to pay their capital cost plus the cost of fuel.

The past few years have seen a sudden departure from the previous two decades of relatively low and stable prices, as rising demand for gas in the power sector and poor upstream performance combined with cold weather to yield extraordinarily high prices. Most analysts think that U.S. gas prices will remain well above those observed in the 1990s, but the ultimate outcome is unclear.

2.2 Canada and Mexico

The Canadian market resembles the U.S. gas market in some respects, including the presence of many gas-producing companies and the need to transport gas long distances to the major consuming areas. These include the major cities of eastern Canada, served by TransCanada Pipeline (TCPL), as well as the exports to the U.S. Midwest and Northeast, where Canada is a major source of supply.

Historically, TCPL was a regulated monopoly with partial government ownership that used its systemwide revenues to expand its pipeline further eastward, including projects that might not have been economical on their own. The province of Alberta contains the bulk of Canadian gas resources and production is gathered into a number of large pipelines and then sent south and east. There is also a major chemical industry in Alberta, which acts as a primary consumer, and the booming tar sands fields, which require heat to produce their heavy oil.

In Mexico, natural gas is produced and shipped by Pemex, the government's national oil company. Most of the gas is associated gas that is found with oil fields and only recently has Pemex been seeking to find and develop gas fields. The gas is primarily used to displace residual fuel oil in industrial and utility

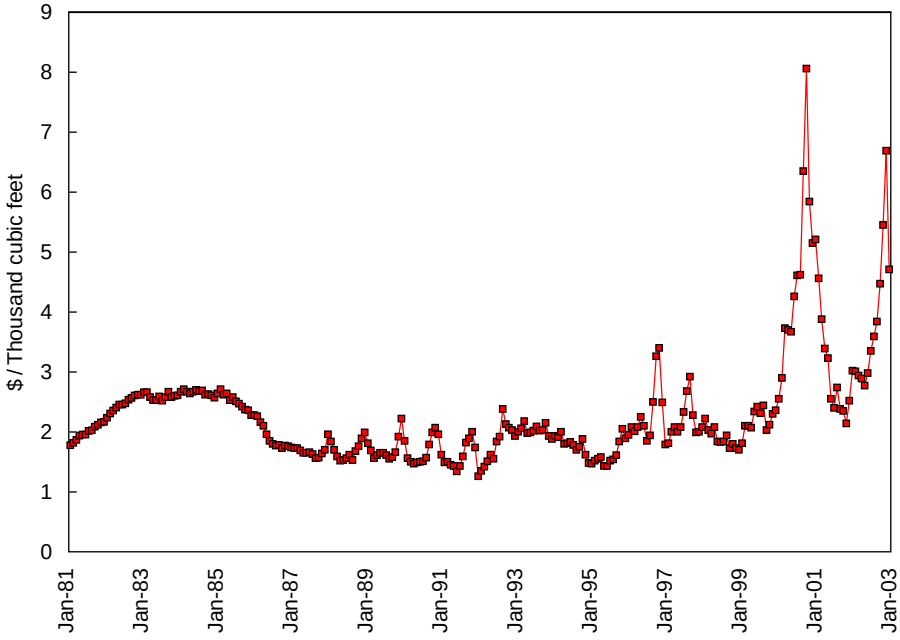


FIGURE 9 U.S. average wellhead natural price. From Monthly Energy Review, U.S. Department of Energy.

uses, which rely heavily on it. (The government monopoly on oil and gas has had the perverse effect of raising oil consumption by causing underinvestment in natural gas.)

2.3 European Gas Markets

In 1959, the discovery of the supergiant Groningen field in the Netherlands led to the beginning of the modern gas era in Western Europe. Before that time, countries such as Germany, France, and Italy had minor amounts of domestic gas production and many cities had town gas facilities. The Dutch began the large-scale international trade of natural gas in Europe and encouraged the building of national pipeline systems, by initially offering their gas at prices designed to be competitive with coal, which was the main competing fuel at the time. Their strategy was to build markets and encourage customers to make the necessary expenditures to put in the infrastructure needed for their gas to be exploited.

With the 1970s, Norwegian gas discoveries began to be marketed into Western Europe and the development of liquefied natural gas shipping also brought supplies from Algeria and Libya to the Continent. The Russians, who have the largest natural gas reserves in the world, developed long-distance pipelines from Siberia into Western Europe in order to gain hard currency. With the oil price shocks of the 1970s, these suppliers became more aggressive in seeking higher gas prices, with the Algerians going so far as to demand prices above those for crude oil, arguing that natural gas was a superior fuel. Many of these price demands did not stick, although exporters succeeded in raising prices to approximately those of crude oil equivalence, as Fig. 10 indicates.

The European gas market remains dominated by a few suppliers on one side and a few consumers on the other. In France, the buyer is Gaz de France, which delivers and resells to other customers. In Germany, Ruhrgas continues to control most of the national pipeline system, and in Italy, Snam is the main pipeline company. The heavy investment needed for the development of the gas infrastructure encouraged governments to maintain state-owned monopolies to run their gas pipeline systems and these companies have been particularly willing to accept long-term contracts with the suppliers paying prices indexed to oil prices. The European Union has been trying for some years to open up the closed markets.

The British market is somewhat unique, in that it was separated from the Continental gas market until recently, and it produced most of its own supplies in the North Sea for the past several decades. Initially, British Gas served as a monopsony buyer, purchasing gas produced by oil companies in the North Sea at fixed prices and transferring it through its pipeline system to companies and residential users. For a time, extremely low prices were paid to domestic gas

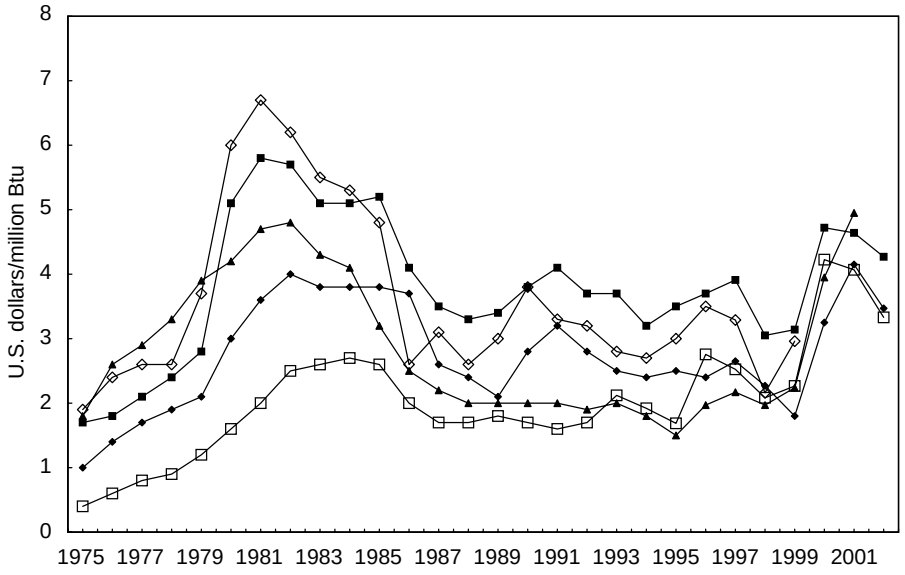


FIGURE 10 World gas and oil prices, $/million Btu (nominal). Closed squares, Japan LNG; diamonds, Western Europe gas imports; triangles, U.S. gas imports; open squares, U.S. wellhead gas; open diamonds, crude oil CIF. From BP Statistical Review of World Energy Industry and Monthly Energy Review, U.S. Department of Energy.

producers, which discouraged drilling and led to British Gas importing gas from Norway at much higher prices. The Thatcher government discontinued that practice, privatizing British Gas and allowing producers to sell their gas freely.

Norway has become a primary exporter to the European Continent, with a very mixed heritage. Although most of its natural gas has been discovered and produced by private companies, the nation relies on a national oil company, Statoil, as a primary participant and operator in many fields and used the Gas Fuel Committee to act as a monopoly marketer of its gas until recently. This reflected partly a desire to gain oligopoly rents for its gas by not having different producers bid the export price down to competitive levels and partly the fact that much of its gas exports went to areas that were dominated by monopsony buyers, particularly in France and Germany.

Norwegian sales did not begin until the 1970s, when the oil crises created a buyers' market for non-OPEC energy supplies. Thus, the Norwegians were able to achieve premium prices for their gas, selling the gas reserves in given fields and gearing their development plans according to the contracts they signed. However, after selling the gas from the supergiant Statfjord field at a significant premium in 1981, they found substantial buyer resistance in the weakening market, worsening as the price of oil collapsed in 1986. This led to a reversal of policy and an effort to price at competitive levels in order to lock in buyers.

The natural gas market is growing strongly, particularly in southern Europe, and new LNG trade is being arranged from West Africa and potentially the Middle East. At the same time, new pipelines from North Africa and Russia are being built and the Caspian producers are attempting to place their reserves into the lucrative European market.

2.4 Russian Gas

Russia contains the largest natural gas reserves in the world and is the world's largest producer and exporter. Natural gas provides the country with approximately one-fifth of its foreign currency earnings and accounts for a similar proportion of the federal government's budget. The domestic and export markets are dominated by the company Gazprom, a privatized firm that is little changed from the Soviet days (including continued partial government ownership). Gazprom produces approximately 80% of the gas in Russia and the relatively new marketing firm Itera accounts for most of the rest, with a small amount produced from oil deposits by the various oil companies. Itera also buys gas from Caspian producers and resells it to other customers, principally the Ukraine.

Gazprom continues to be the primary exporter to the West, accounting for nearly all sales outside the country, including to most of the former Soviet Bloc nations as well as the daughter republics of the Soviet Union. Sales are essentially carried out on a monopoly basis, primarily to large buyers such as Gaz de France and Ruhrgas, typically on a long-term basis with prices fixed relative to petroleum prices, usually on a moving average basis.

Domestic gas sales are a different matter, where many areas have their own local distribution companies and buy gas from either Gazprom or Itera. However, difficulties in obtaining payments from many customers have meant that many local companies have gone bankrupt and some of them have been bought out by Gazprom. The domestic price remains fixed at a fraction of the international price, although the government hopes to raise that over time. As of 2003, it is approximately 20% of the European Border price, although residential customers pay a flat monthly fee, with no metering of consumption.

Russia also holds huge reserves in the Far East, particularly in the Yakut area and near Lake Baikal, where British Petroleum holds a major interest in the supergiant Kovytka field. Hopes are that exports will be begin soon to China and other Far Eastern countries, which would require the construction of large-scale, long-distance pipelines. The primary delay relates to questions of pricing and ownership of the pipelines, although most of the major players appear to be very interested in beginning such trade, as it would have the effect of reducing coal consumption in China—a major source of pollution both locally and regionally—and reducing reliance on Middle Eastern oil imports in Korea and Japan.

2.5 Asian Gas Markets

The Asian gas market has been one of the last to develop, primarily because most gas supplies are at a great distance from the major consuming areas. Aside from minor amounts of consumption in countries such as India and Pakistan, which have exploited small domestic deposits for some time, the primary consumption of natural gas has been in the form of LNG, transported across oceans. Unlike North America and Europe, much of the gas has

been consumed by the electric utility sector rather than by households.

The LNG trade began in the 1970s when Japanese utilities, looking to reduce their emissions resulting from direct consumption of crude oil, began to arrange for small quantities of LNG to be shipped from Alaska, Brunei, and other areas. Initial prices were relatively low to be competitive with crude oil prices. As crude oil prices soared in the 1970s, and many LNG exporters became more aggressive in seeking higher prices, the Japanese buyers agreed to pay prices equivalent to delivered crude oil prices in Japan. This practice has remained for some time and still dominates the trade in Asia.

Because of the greater expense of transporting natural gas in the liquefied form, the market has been much slower to develop than pipeline markets in North America and Europe. Natural gas's market share has grown slowly in Asia but remains far behind that of other regions and still does not penetrate the residential market significantly. In part, this reflects warmer climates in some areas, but it also reflects the expense of building residential distribution systems. Large consumers, such as power companies, require minimal investment in pipelines, which makes them ideal customers for a new regasification terminal, as it maximizes consumption relative to initial outlays.

However, the fact that price is not set by the market (i.e., long-run marginal costs) but rather by the price of crude oil has left prices for LNG usually well above costs and made the business fairly attractive. As a result, there are a number of projects unable to find customers. In many markets, this would mean downward pressure on prices, but such has been slow to develop in Asia for two reasons. First, many of the sellers already have large volumes committed at the higher prices and they do not wish to sacrifice the revenues from existing sales, since these customers would almost certainly demand lower prices for their contracts. Additionally, the high price of transportation means that supplies from more distant regions are not particularly competitive.

The Asian gas market is slowly changing due to a combination of falling liquefaction costs and a slowly developing gas pipeline system in the region. The construction of pipelines from Russia to Japan, China, and Korea would change the marketing of natural gas. Also, the slow growth of domestic gas usage in India, Pakistan, and Indonesia as well as development of pipeline connections in Southeast Asia mean that LNG is no longer as dominant as it once was.

2.6 Other Markets

Gas consumption is sparse in other parts of the world, but some usage occurs in areas with large resources, such as the Middle East, as the cheap gas attracts energy-intensive industry, and in areas where there is some gas and at least a local market, as in parts of Africa and Latin America. Nigeria has large amounts of associated gas and in addition to exporting LNG, it is trying to develop a local pipeline network, as are other countries with nonassociated gas, such as Bolivia and Tanzania. Power plants dominate the customer list, but this should change as the systems mature.

2.7 LNG Contracting

The LNG business is somewhat unusual, in that it involves a basic commodity that is traded almost exclusively on long-term contracts. The particularly high capital costs necessary for building liquefaction plants, combined with the small number of importers with the necessary facilities for unloading LNG, have made producers very sensitive to demand-side risk. In most cases, they have not begun construction until they have located enough gas reserves to supply their plant for its lifetime and sold nearly all of its planned production on fixed contracts.

Typically, plants would line up contracts to sell 80 to 100% of their output before beginning construction and most of those contracts would have very strict take-or-pay clauses, requiring that buyers accept most or all of their contract amounts on a regular basis. Buyers were usually not allowed to resell any of their gas either, which tended to make them hesitant to be aggressive in signing up new volumes.

Because most LNG, especially in Asia, was sold into markets without directly competing fuels, there was no real market price for the natural gas. As a result, nearly all contracts called for prices to be set equivalent to the crude oil price level, converted on a Btu basis. This has reduced the price risk, at least to some degree since crude oil prices are also volatile and uncertain, but it has generally meant that prices are adequate to provide sufficient if not quite hefty profits to the producers and the exporting governments.

However, this has also meant that market penetration has been constrained by the combination of strict contracts and high prices. The high storage costs for LNG meant that buyers did not wish to hold large inventories and thus have been reluctant to sign up for volumes that might not be used in a short period of time, reducing sales on a spot basis or

for seasonal purposes. Also, where the gas competes in the industrial sector with fuel oil, usually sold at a discount to crude oil, it finds itself unable to gain a large market share.

SEE ALSO THE FOLLOWING ARTICLES

Clean Air Markets • Market-Based Instruments, Overview • Market Failures in Energy Markets • Markets for Biofuels • Markets for Coal • Markets for Natural Gas • Oil and Natural Gas: Economics of Exploration • Oil and Natural Gas Leasing • Oil and Natural Gas Resource Assessment: Classifications and Terminology • OPEC Market Behavior, 1973–2003 • Petroleum Property Valuation • Stock Markets and Energy Prices • Strategic Petroleum Reserves

Further Reading

Adelman, M. A. (1993). "The Economics of Petroleum Supply." MIT Press, Cambridge, MA.

Bradley, R. L., Jr. (1996). "Oil, Gas and Government: The U.S. Experience." The Cato Institute, Washington, DC.

British Petroleum (BP). (2003). "BP Statistical Review of the World Energy Industry." British Petroleum, London.

Hartshorn, J. E. (1993). "Oil Trade: Politics and Prospects." Cambridge University Press, Cambridge, UK.

Horsnell, P. (1997). "Oil in Asia: Markets, Trading, Refining and Deregulation." Oxford University Press, London.

Horsnell, P., and Mabro, R. (1993). "Oil Markets and Prices: The Brent Market and the Formation of World Oil Prices." Oxford University Press, London.

International Energy Agency. (2002). "World Energy Outlook." International Energy Agency, Paris.

Leffler, W. L. (2000). "Petroleum Refining in Nontechnical Language," 3rd ed. Pennwell, Tulsa, OK.

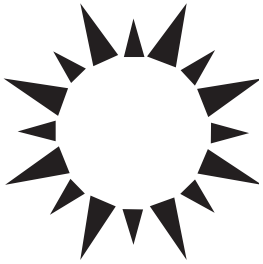
Lynch, M. C. (2002). Causes of oil price volatility. *J. Energy Dev.* **42**, 373–389.

Mitchell, J. V., Beck, P., and Grubb, M. (1996). "The New Geopolitics of Energy." Royal Institute of International Affairs, London.

Razavi, H., and Fesharaki, F. (1991). "Fundamentals of Petroleum Trading." Praeger, New York.

Tussing, A., and Tippee, R. (1995). "The Natural Gas Industry: Evolution, Structure, and Economics," 2nd ed. Pennwell, Tulsa, OK.

Marx, Energy, and Social Metabolism



JOAN MARTINEZ-ALIER
Universidad Autónoma de Barcelona
Barcelona, Spain

1. Social Metabolism
2. Podolinsky's Agricultural Energetics
3. Engels's Reaction to Podolinsky
4. Otto Neurath's Naturalrechnung

Glossary

Otto Neurath (1882–1945) A famous logical empiricist philosopher of the Vienna Circle, he was also an economist or economic historian, and a Marxist. Neurath introduced the idea of incommensurable values in the economy, a concept now embraced by some ecological economists.

Sergei Andreyevich Podolinsky (1850–1891) A Ukrainian physician, economist, public and political figure, and a contemporary of Karl Marx and Fredrick Engels, who performed early detailed studies of energy flows in agriculture.

social metabolism An idea adopted by Karl Marx and others in the 19th century that nature and society were linked by material floes. Marx largely ignored energy flows, while others embraced them as a key concept.

Marx (1818–1883) and Engels (1820–1895) were contemporaries of the physicists and physiologists who established the first and second laws of thermodynamics in the 1840s and early 1850s (J. R. Mayer, 1814–1878, J. P. Joule, 1818–1889, R. Clausius, 1822–1888, W. Thomson, later Lord Kelvin, 1824–1907). Marx and Engels followed with a few years' delay many of the remarkable scientific and technical novelties of their time including the developments in thermodynamics or energetics. Engels wrote in 1888 that the nineteenth century would be remembered as the century of Darwin, Mayer, Joule, and Clausius. It had been not only the century of the theory of evolution but also of the theory of the transformation of energy. However, Marx and Engels never expressed or translated their analytical concepts (productive forces, surplus value, exploitation) into the language of energy as they could have done in their mature work after the 1850s. They never said that the productivity of labor in agriculture and in industry would depend on the energy subsidy to the economic process. After all, they were historians and economists, not yet environmental historians, ecological economists, or industrial ecologists. Increases in productivity depended on the development of the productive forces (*Produktivkr…fte*), where "Kraft" (force) was not used with the physical meaning of energy. The productive forces could be furthered or could be fettered by the social relations of production. Capitalism had meant an enormous development of the productive forces but it provoked economic crises by its very nature—that is, the exploitation of wage labor. Malthusian crises of subsistances were not relevant. These were essentials points of their theory.

1. SOCIAL METABOLISM

The first volume of *Capital* was published in 1867. There were certainly precursors of a biophysical approach to the economy from the 1880s onward, and at least one of them was close to Marxism (such as S. A. Podolinsky, 1850–1891). These proto-ecological economists did not form a school. In the 1970s, the study of the human society and economy from a physical point of view (flows of materials and energy) started to be practiced by coherent research groups with little relation to Marxism. Histories of the use of energy in the economy are not connected to Marxism (even if their authors might sometimes have Marxist pasts). The notion of energy return on

 © 2004 Elsevier Inc. All rights reserved.

energy input (EROI) was applied in the 1970s to the economy by Charles Hall and other ecologists with no relations to Marxism. Explanations of economic growth model the use of energy in the economy (or rather physical work output as distinguished from energy [exergy]). They criticize neoclassical production functions and growth theories but draw no inspiration from Marxism.

Marx and Engels had a profound interest in the interactions between the human economy and the natural environment, particularly as regards capitalist agriculture. This was expressed in Marx's use in his own drafts after 1857–1858 and in *Capital,* of the notion of metabolism between humans and nature. Marxist use of "metabolism" (by Marx himself, also by Bukharin) is well known. Alfred Schmidt was the first author to insist in the role of the concept of *Stoffwechsel* in Marx's work on the development of capitalism, noting Moleschott's and Liebig's influence, noting also its substantive use for the cycling of plant nutrients. Marx became so keen on the concept of metabolism that in a letter to his wife, he nicely wrote that what made him feel like a man was his love for her, and not his love for Moleschott's metabolism or for the proletariat. There is a rediscovery of Marx's social metabolism. Authors working on industrial metabolism or in social metabolism look at the economy in terms of flows of energy and materials. They insist that the economy should be seen as a subsystem embedded in a larger physical system. Is this perspective to be found in Marx already?

Marx and Engels were one generation younger than the agricultural chemists who published from 1840 onward their researches on the cycles of plant nutrients (phosphorous, nitrogen, potassium), influenced by the debates on the threat of decreasing agricultural yields and the wholesale imports of guano, mainly from Peru, which became an essential bulk commodity for agricultural production. About 11 million tons of guano were exported from Peru in four decades. The analyses of the composition of imported guano, and also of other manures and fertilizers (bones, for instance) already well known to farmers, laid the foundations for agricultural chemistry. Liebig's name was associated, by his own wish, to a new future leading sector of the economy, the fertilizer industry. It may also be associated to an ecological vision. He is recognized as one of the founders of ecology, before the name itself was invented. Politically he developed an argument against latifundist agriculture and agricultural exports because the plant nutrients would not return to the soil. He was in favor of small-scale agriculture and dispersed settlements. Marx quoted this opinion very favorably on several occasions.

Before proceeding further, it must be emphasized that Marx and Engels were also contemporaries of the physiologist Jacob Moleschott (1822–1893). To the present day, Marxists call him a "vulgar materialist" (meaning a reductionist materialist, in opposition to historical materialists who favor dialectics and coevolution). However "vulgar," Moleshott was a main figure in the introduction of the concept of metabolism in physiology and in what now is called ecology. Moleshott had been dismissed after 1848 from Heidelberg, he taught in Zurich and in Rome, he became a well-known professor and editor of a scientific journal, and he had been an activist. Materialism was used against the doctrine of vitalism and against religion. His book on the theory of human nutrition of 1850 had the subtitle *Ffr das Volk*. This was certainly different in intent from the studies by other authors on how to feed cheaply and efficiently soldiers, industrial workers or slaves (in America). In 1851 he published *Physiology of Metabolism in Plants and Animals*, one year later, *Der Kreislauf des Lebens* (The Cycle, Circuit or Circle of Life). Engels, in a letter to Piotr Lavrov of 1875 discussing Darwinism, summarized the idea of the "cycle of life" in "the fact that the vegetable kingdom supplied oxygen and nutriment to the animal kingdom and conversely the animal kingdom supplied plants with carbonic acid and manure." This had been interpreted as cooperation in nature by that trinity of "vulgar materialists" (Vogt, B*f*chner, and Moleschott, who, incidentally, now were talking loudly about "struggle for existence") and had been stressed by Liebig.

Indeed, Marx's use of *Stoffwechsel* (metabolism) in *Capital* (vol. I) and, before that, in some drafts of the late 1850s was influenced by Liebig, probably by other authors, and certainly by Moleshott. Marx used "metabolism" in two senses. First, as a biological analogy or metaphor to describe the circulation of commodities. Second, more relevantly for the present article, he used the expression "metabolism between man and earth, or between society and nature" to refer specifically to the cycles of plant nutrients, quoting, and praising Liebig. Such metabolism was broken by an agriculture of spoliation.

Marx found Liebig supremely relevant because he talked on the natural conditions of agricultural fertility and on the possibility of development of the productive forces by the fertilizer industry. This was useful for the polemics against Malthus and for

the theory of land rent. Foster has analyzed in depth Marx's debt to Liebig, and he has wrongly dismissed Moleshott's influence on Marx's use of "metabolism." Foster does not quote Moleshott's books of 1851 and 1852 on the "circle of life" and on the physiology of metabolism in plants and animals.

In conclusion, Marx was a historian and an economist, he was also a student of agriculture and he read on agricultural chemistry and adopted the notion of "metabolism" between man and nature. The material flows between humans and nature were mobilized by human labor, except perhaps in very primitive societies that lived from gathering. The development of tools by humans was essential for the metabolism. Marx did not include the metabolic flow of energy in agriculture and in industry in his analysis, so he could not trace a fundamental distinction (as Lotka was to do 40 years later) between the use of energy in nutrition (endosomatic) and the use of energy by tools (exosomatic). This difference between biometabolism and technometabolism is crucial for the understanding of human ecology. We as a species have genetic instructions regarding our use of endosomatic energy but not our use of exosomatic energy, which must be explained by history, politics, economics, culture, and technology. This is not to be found in Marx. Not only he left energy flow aside, he did not count material flows in detail or even roughly, he was not a protoindustrial ecologist. But on the positive side, he did not see the economy as a closed system, or nature as an immutable purveyor of free gifts, as the neoclassical economists were to do.

Humans were part of nature, humans used nature's materials, we could increase its produce by the development of the so-called productive forces but we could also undermine the natural conditions of production. This was the case with capitalist agriculture. Marx wrote:

> *Capitalist production disturbs the metabolic interaction between man and the earth, i.e. it prevents the return to the soil of its constituent elements consumed by man in the form of food and clothing, hence it hinders the operation of the eternal natural conditions for the lasting fertility of the soil… Moreover, all progress in capitalist agriculture is a progress in the art, not only of robbing the workers, but of robbing the soil. (Capital, I)*

He added that the separation of town and country, caused by latifundist agriculture and by the concentration of sources of energy in cities, provokes an "irreparable rift" in the process of social metabolism. The result of this was a squandering of the soil, aggravated by trade, undermining the conditions of agricultural production. Engels wrote, "As the individual capitalists are engaged in production and exchange for the sake of the immediate profit, only the nearest, most immediate results must first be taken into account." In modern-day parlance, Engels was saying that individual capitalists discounted the future and did not include externalities in their accounts. He continued:

> *As long as the individual manufacturer or merchant sells a manufactured or purchased commodity with the usual coveted profit, he is satisfied and does not concern himself with what afterward becomes of the commodity and its purchasers. The same thing applies to the natural effects of the same actions. What cared the Spanish planters in Cuba, who burned down forests on the slopes of the mountains and obtained from the ashes sufficient fertilizer for one generation of very profitable coffee tress—what cared they that the heavy tropical rainfall afterward washed away the unprotected upper stratum of the soil, leaving only the bare rock! (Dialectics of Nature)*

Such quotations, which could be multiplied, should not be seen as anecdotal, they arose, as Foster has shown, from a theory of "metabolic rift." So Marx was "greener" than usually thought. However, what Greens can learn in Marx is not new to them, the knowledge of the cycles of plant nutrients and the damage to soil and forests.

"Metabolic rift" is not equivalent to "decreasing returns." Marx wrote to Engels on February 13, 1866, saying that Liebig's agricultural chemistry was more important for the discussion on decreasing returns than all the economists put together. Marx dismissed the notion of decreasing returns in agriculture, pointing out in the context of his praise for Liebig's agricultural chemistry and its promise of artificial fertilizers, that it did not make sense to assume in Britain that the produce of the land would increase in a diminishing ratio to the increase of the laborers employed, because in practice there was at the time both an increase in production and an absolute decrease (already) in the number of laborers. Marx was not worried about crisis of subsistances. Marxists attacked Malthus, which could be expected because Malthus had made the point that improving the situation of the poor was counterproductive because they would have more children. There were many debates on how many people could the earth feed around 1900. Some Marxists not only attacked Malthus, they also attacked the Neomalthusians of the late 19th century and early 20th century who were often political radicals and feminists (Paul Robin, Emma Goldman).

2. PODOLINSKY'S AGRICULTURAL ENERGETICS

The link between material metabolism (*Stoffwechsel*, literally exchanges of materials) and the flow of energy at the level of cells and organisms was made in the 1840s. It was then also understood that agriculture might be represented in terms of changes in the flow of energy and not only as an intervention in the cycling of plants nutrients (in 1845, Mayer used *Stoffwechsel* for energy flow). Metabolism was therefore used no only for materials, also for energy. Of course, materials could be cycled, energy could not. The theory of the direction of the flow of energy was developed after 1850 and the establishment of the second law.

Marx and Engels were interested in energy. For instance, Engels wrote to Marx on July 14, 1858, commenting on Joule's work on the conservation of energy (of 1840) as something of well known to them. Marx was of course interested in new sources of energy. One example will suffice: it was already discussed at the time whether hydrogen could be a net source of energy, depending on the energy requirement for electrolysis. Marx wrote to Engels on April 2, 1866, that a certain M. Rebour had found the means of separating the oxygen from hydrogen in water for very little expense. However, Marx did not refer to the flow of energy as a part of metabolism throughout his lifetime.

At a more general level, one interesting point arises from Engels's unwillingness to understand how the first and second laws could apply at the same time: the "dialectics of Nature" failed him there. As Engels became aware of Clausius's concept of entropy, he wrote to Marx on March 21, 1869:

> *In Germany the conversion of the natural forces, for instance, heat into mechanical energy, etc. has given rise to a very absurd theory—that the world is becoming steadily colder... and that, in the end, a moment will come when all life will be impossible... I am simply waiting for the moment when the clerics seize upon this theory.*

Indeed, not only the clerics but also W. Thomson (Lord Kelvin) brandished the second law in his religious tirades about the "heat death," although he had and could have no inkling of the source of energy in the sun in nuclear fusion. One understand Engels's dislike for the uses to which the second law was put. Josef Popper-Lynkeus (1838–1921), who with Ernst Mach became one of the sources of inspiration for the analytical, antimetaphysical philosophy of the Vienna Circle, complained since 1876 about W. Thomson's theological handling of Carnot's law. However, Engels's dislike of the second law was not only motivated only by its religious use, or abuse. He believed that ways would be found to reuse the heat irradiated into space.

Another interesting point is the reaction by Engels in 1882 (in letters to Marx) regarding Podolinsky's work on energy flows in agriculture. Podolinsky had studied, we might say, the entropy law and the economic process, and he tried unsuccessfully to convince Marx that this could be brought into the Marxist analysis. Podolinsky's work has been seen as a first effort to develop ecological economics. Politically he was not a Marxist, he was a Ukrainian federalist narodnik. He had criticized Marx's overpowering behavior at the congress of the International of 1872, praising the anarchist James Guillaume. However, Podolinsky himself saw his work on the energetics of agriculture as a contribution to Marxist theory, and he would not have objected to appear under "Marxism and Energy" in an encyclopedia. Writing to Marx on April 8,1880, he said, "With particular impatience I wait for your opinion on my attempt to bring surplus labor and the current physical theories into harmony."

In his article (published in long versions in Russian in 1880 and in German in 1883, and in short French and Italian versions in 1880 and 1881), Podolinsky started by explaining the laws of energetics, quoting from Clausius that although the energy of the universe was a constant, there was a tendency toward the dissipation of energy or, in Clausius's terminology, there was a tendency for entropy to reach a maximum. Podolinsky did not discuss explicitly the difference in thermodynamics between open, closed and isolated systems, although he stated explicitly, as the starting point of his analysis, that at the present time the earth was receiving enormous quantities of energy from the sun, and would do so for a very long time. All physical and biological phenomena were expressions of the transformations of energy. He did not enter into the controversies regarding the creation of the universe and its "heat-death," nor did he discuss the relations between thermodynamics and the theory of evolution. In March 1880 he published an article against social Darwinism. He certainly realized that the availability of energy was a crucial consideration for the increase (or decrease) of population. However, he thought that the distribution of production was explained by the relations between social classes: "in the countries where capitalism triumphs, a great part of work goes towards the production of luxury goods, that is to say, towards a

gratuitous dissipation of energy instead of towards increasing the availability of energy."

He explained that plants assimilated energy, and animals fed on plants and degraded energy. This formed the *Kreislauf des Lebens*:

> *We have in front of us two parallel processes which together form the so-called circle of life. Plants have the property of accumulating solar energy, but the animals, when they feed on vegetable substances, transform a part of this saved energy and dissipate this energy into space. If the quantity of energy accumulated by plants is greater than that dispersed by animals, then stocks of energy appear, for instance in the period when mineral coal was formed, during which vegetable life was preponderant over animal life. If, on the contrary, animal life were preponderant, the provision of energy would be quickly dispersed and animal life would have to go back to the limits determined by vegetable wealth. So a certain equilibrium would have to be built between the accumulation and the dissipation of energy.*

Not only plants, also human labor had the virtue of retarding the dissipation of energy. Human labor achieved this by agriculture, although the work of a tailor, a shoemaker, or a builder would also qualify, in Podolinsky's view, as productive work, since they afford "protection against the dissipation of energy into space." The energy available for humankind came mainly from the sun. Podolinsky gave figures for the solar constant. He explained how coal and oil, wind energy, and water power were transformations of solar energy. He mentioned tides as another possible source of energy. He then started his analysis of the energetics of agriculture, remarking that only a very small proportion of sun energy was assimilated by plants. Human work, and the work of animals directed by humans, were able to increase the availability of energy by agricultural activity. This he showed by comparing the productivity of different types of land use taking statistics from France (he was living at the time in Montpellier). Table I summarizes his data.

TABLE I

Annual Production and Energy Input (Only Work by Humans and Domestic Animals) per Hectare, Averages for France in 1870, According to Podolinsky

System	Production (kg)	Production (kcal) Energy input (kcal)
Forest	900 (dried wood)	2,295,000 Nil
Natural pastures	2,500 (hay)	6,375,000 Nil
Sown pastures	3,100 (hay, excluding seed)	7,905,000 37,450 (50 horse-hours and 80 human-hours)
Wheat	800 (wheat) and 2,000 (straw) (excluding seed)	8,100,000 77,500 (100 horse hours and 200 human-hours)

Podolinsky then compared wheat agriculture and sown pastures and forest, concluding that production was higher when there was an input of human and animal work. Thus, comparing wheat agriculture to natural pastures, each kcal put in contributed to an increase of 22 kcal of production. If forest were taken as the terms of comparison, the energy productivity of human and domestic animals work was even higher. Notice that Podolinsky was counting human and animal work—that is, not the food intake but the work done. He did not include solar radiation in the input of energy, because he was writing as an ecological economist. Solar radiation is indeed a free gift of nature (moreover, without an owner, so that there is no payment of rent).

Energy values of wood, ha,y and straw, 2550 kcal/kg, of wheat 3750 kcal/kg. Hours of work converted into kcal: 645 kcal/hour of horse-work, 65 kcal/hour of human-work.

The conclusion was that work could increase the "accumulation of energy on earth." Although Podolinsky mentioned guano, and although he must been keenly aware of the war then raging for Peruvian or Chilean saltpeter, he did not subtract from the output, or include in the input, the energy contents and cost of fertilizer. Nor did he consider the energy spent for steam engines for threshing. His methodology is nevertheless basically the same as that used later for establishing the energy balance of particular crops, or for small scale societies, or for the entire agricultural sector of particular countries.

Podolinsky then went on to explain the capacity of the human organism to do work. Otherwise "it would be difficult to explain the accumulation of energy on the surface of the earth under the influence of labour." Quoting from Hirn and Helmholtz, he concluded correctly that "man has the capacity to transform one-fifth of the energy gained from food into muscular work," giving to this ratio the name of "economic coefficient," remarking that man was a more efficient transformer of energy than a steam engine. He then used a steam engine metaphor to put across a general theoretical principle on the minimum natural conditions of human existence, from an energy point of view. He wrote that humanity was a "perfect machine" in Sadi Carnot's sense: "Humanity is a machine that not only turns heat and other physical

forces into work, but succeeds also in carrying out the inverse cycle, that is, it turns work into heat and other physical forces which are necessary to satisfy our needs, and, so to speak, with its own work turned into heat is able to heat its own boiler."

Now, for humanity to ensure its elementary conditions of existence, each calorie of human work must then have a productivity of 5 calories. Taking into account that not everybody is able to work (children, old people), and that there other energy needs apart from food energy, the necessary minimum productivity would be more like ten or more. If that minimum is not achieved, then, of course, "scarcity appears and, many times, a reduction of population." Podolinsky then established the basis for a view of the economy in terms of energy social metabolism, looking at the energy return to energy input in a framework of reproduction of the social system. He thought than he had reconciled the Physiocrats with the labor theory of value, although the Physiocrats (in the 18th century) could not have seen the economy in terms of energy flow.

Podolinsky also emphasized the difference between using the flow of solar energy and the stock of energy in coal. The task of labor was to increase the accumulation of solar energy on earth, rather than the simple transformation into work of energy already accumulated on earth, more so since work done with coal was inevitably accompanied by a great dissipation of heat-energy into space. The energy productivity of a coal miner was much larger than that a primitive farmer could obtain, but this energy surplus from coal was transitory, and moreover (Podolinsky added in a footnote) there was a theory that linked climatic changes to concentrations of carbon dioxide in the atmosphere, as Sterry Hunt explained at a meeting of the British Society for the Advancement of Science in the autumn of 1878. Notice here that the emphasis here was not on capital accumulation but on increasing the availability of energy and certainly also its dissipation. Podolinsky was not, however, at all pessimistic about the prospects for the economy, and he was hopeful for the direct use of solar energy for industrial purposes, referring to the "solar engine of M. Mouchot." One could envisage that one day solar energy would be used directly to make chemical syntheses of nutritive substances, by-passing agriculture. Thus, a proper discussion of the demographic question had to take into account "the relations between the general quantity of energy on earth and the quantity of energy who live on it," and this was a more relevant consideration, in his explicit view, than the Malthusian prognosis. He was also hoping (as he had written to Marx on March 30, 1880, sending his work to him) to develop applications of his energy accounts to different forms or modes of production. Marx died in 1883 and apparently never commented on Podolinsky's work, beyond a letter of acknowledgement at the beginning of April 1880. He asked Engels to write a comment, which Engels did not do until December 1882. Podolinsky himself became ill in 1881 or 1882 and never recovered until his death ten years later, back in Kiev.

3. ENGELS'S REACTION TO PODOLINSKY

Engels had some knowledge of the energetics of human physiology, and in a note of 1875 (later included in the *Dialectics of Nature*) he referred to Fick's and Wislicenus's experiment in climbing the Faulhorn in 1865, which became popularized under the name "a day of hunger for science." Fick (1829–1901) had written in 1857 and 1958 on the amount of kilocalories (2700) that a man would need and spend per day, when not working, and how different types of work implied different energy expenditures over that rate. An idea being circulated at the time (for instance, in an article in *Das Ausland* in 1877) was that the economic values of different types of work could be established in physical terms. Engels already explicitly rejected this notion in 1875.

On December 19 and 22, 1882, Engels wrote two letters to Marx (about 1000 words) explaining that Podolinsky had "discovered" the following facts, already well known. If the food intake of one person per day were equal, say, to 10,000 kcal, then the physical work done would be fraction of this energy. This physical work could become economic work if employed in fixing solar energy, for instance through agricultural activity. With a daily food intake of, say, 10,000 kcal, a man will work in agriculture. Podolinsky's discovery was that human labor had the power of detaining solar energy on the earth's surface and permitting its activity longer than would be the case without it. All the economic conclusions he drew from this were wrong. The economic labor, which a person working in agriculture performed by the employment of these 10,000 kcal, consisted in the fixation for a greater or less time of new energy radiated to that person from the sun. Whether, however, the *new* quantity of energy fixated by the application of the original 10,000 kcal of daily nourishment "reaches 5,000, 10,000, 20,000, or 1,000,000 kcal depends solely on the degree of

development attained by the means of production." In agriculture, "the energy value of the auxiliary materials, manures, etc. also enters into the calculation"—a perceptive comment that directly makes one think of the agricultural energy balances of the 1970s, which showed that the energy return to energy input of modern agriculture was on the decline although the economic productivity was increasing. "In industry," wrote Engels, all calculation comes to an end; in most cases the work added to the product (he wrote on December 19, 1892) "can no longer be expressed in kcal." Engels was right that in industrial production, there would be an economic added value although not a net gain of energy—on the contrary, there would be more available energy in the raw materials than in the final product. Actually, also in agriculture the energy in the product does not come from human work but it comes really from photosynthesis, human labor helping and guiding the process. In fact, Engels wrote on December 22, 1982:

> *Storage of energy through work really takes place only in agriculture; in cattle raising the energy accumulated in the plants is simply transferred as a whole to the animals, and one can only speak of a storage of energy in the sense that without cattle-raising, nutritious plants wither uselessly, whereas with it they are utilised. In all branches of industry, on the other hand, energy is only expended…. So that if one chooses one can translate into a physical language [ins Physikalische fbersetzen] the old economic fact that all industrial producers have to live from the products of agriculture, cattle raising, hunting, and fishing—but there is hardly much to be gained from doing so.*

Engels also wrongly wrote that "the energy value of a hammer, a screw or a needle calculated according to the costs of production is an impossible quantity," concluding, "In my opinion it is absolutely impossible to try and express economic relations in physical magnitudes." We are here as far away as possible from the "social metabolic" perspective.

Engels was most unfair when he wrote, "What Podolinsky has entirely forgotten is that man as a worker is not merely a fixer of *present* solar heat but a still greater squanderer of *past* solar heat. The stores of energy, coal, ores, forests, etc. we succeed in squandering you know better than I." Podolinsky had not forgotten this at all. Finally, Engels also wrote, "Podolinsky has strayed away from his very valuable discovery into mistaken paths because he was trying to find in natural science a new proof of the truth of socialism, and has therefore confused physics and economics."

Engels's negative reaction to Podolinsky's work, and Marx's 3-years silence from 1880 to 1883 (when he was still intellectually active, much involved in discussions on Russian peasant communes), may be seen as a missed change for an ecological Marxism. Actually Podolinsky's work is relevant not only in the Marxist context. His work and life has an entity of its own, apart from his encounters with Marx and Engels. He had a short life (he was severely ill the last 10 years of it), but he left a strong trace in Ukrainian federalist politics (as a friend of Drahomanov) and also in the Narodnik movements against the Russian autocracy (as a young colleague of Piotr Lavrov but also connected to the Narodnaya Volya group). He was trained as a medical doctor and physiologist. His work on energy and the economy received Vernadsky's approval. In a section of *La Geochimie* (1924) Vernadsky wrote about several authors (Felix Auerbach with his notion of *Ektropismus*, John Joly) who had explained that life was a process which reversed or slowed down the dissipation of energy. He then wrote a memorable phrase: Podolinsky had studied the energetics of life and tried to apply his findings to the study of the economy.

The link between the use of energy and human culture, in the form of "social energetics" (without historical statistical work), became well established and debated in European culture around 1900. Some Marxist authors adopted this outlook, and their work has been seen as an anticipation of Bertalanffy's systems theory, which grew out of the links between thermodynamics and biology. But, to repeat, there is no line of ecological Marxist history based on quantitative studies of material and energy social metabolism. Moreover, Lenin wrote a diatribe against Ostwald's social energetics in the context of his polemics against Bogdanov and his tirades against Mach's empiriocriticism. This was before the October Revolution of 1917. Afterward, Lenin's ill-considered remarks became sacred to the faithful. Thus, on the occasion of the publication of Engels's *Dialectics of Nature* in 1925 (out of drafts and notes he had left behind) and the 30th anniversary of his death, Otto Jenssen printed once again Engels's letters to Marx on Podolinsky (which were first published in 1919) and explained that Engels had anticipated a critique of Ostwald's social energetics even before Ostwald himself appeared on the scene.

4. OTTO NEURATH'S NATURALRECHNUNG

Marx wrote that wealth was undoubtedly created both by human labor and by nature, quoting William

Petty: "Labour is the Father, Nature is the Mother." Beyond facile jokes on whether fathers or mothers are more important, the interesting point is the distinction between value and wealth. Marx's objective was to explain in a capitalist wage economy the appropriation of surplus labor. Such exploitation was not so obvious as in a slave economy or a feudal (serf) economy, since wages were apparently freely established in a market. Surplus labor became surplus value.

An important contribution to the links between Marxism and a biophysical approach to the economy was made later, in the first half of the 20th century, by Otto Neurath (1882–1945). A famous logical empiricist philosopher of the Vienna Circle, he was also an economist or economic historian, and a Marxist in at least two senses. First, he defended a democratically planned economy based on physical accounting in energy and material terms (*Naturalrechnung*) influenced by Popper-Lynkeus's and Ballod-Atlanticus's quantitative, realistic "utopias," in the so-called Socialist Calculation debate of 1919 and following years. Neurath introduced the idea of incommensurable values in the economy, hence the interest in his work by ecological economists. Second, some years later, in the context of the Vienna Circle' project of the Encyclopedia of Unified Science of the 1930s and 1940s, he defended a dialectical view of history (although he did not like the word "dialectics") as the putting together the findings of the different sciences regarding concrete processes or events. The findings of one science, collected in the encyclopedia, with regard to one particular process or event, should not be contradicted by the findings of another science also present in the encyclopedia, and leave things at that. At attempt at reconciliation or removal of the contradiction should be attempted. "Consilience" (to use Edward Wilson's word) should be the rule of the encyclopedia.

Hayek's strong criticism against "social engineering" (1952) was directed, as John O'Neill has put it, "at the whole tradition which attempts to understand the ways in which economic institutions and relations are embedded within the physical world and have real physical preconditions, and which is consequently critical of economic choices founded upon purely monetary valuation." Translated into Foster's language: Hayek critized the view of the economy in terms of social metabolism. While Patrick Geddes, Wilhelm Ostwald, Frederick Soddy, and Lewis Mumford were all rudely dismissed by Hayek, it is the work of Otto Neurath that was the primary target of Hayek's criticism.

One is also reminded not only of the comments of Max Weber against Otto Neurath in *Economy and Society* but even more of his polemics against Wilhelm Ostwald in 1909. Ostwald (who was a well known chemist) was trying to interpret human history in terms of the use of energy. He proposed two simple laws, which are not untrue, and which might act or not in opposite directions. First, the growth of the economy implied the use of more energy and the substitution of human energy by other forms of energy. Second, this came together with a trend toward higher efficiency in the transformation of energy. Max Weber (1909) wrote a famous, ironic review of Ostwald's views, where he insisted on the separation between the sciences. Chemists should not write on the economy. The review was much praised by Hayek in the 1940s. Max Weber's basic point was that economic decisions by entrepreneurs on new industrial processes or new products were based on prices. It could so happen that a production process was less efficient in energy terms and nevertheless it would be adopted because it was cheaper. Energy accounting was irrelevant for the economy. Max Weber did not question energy prices as we would today when externalities such as the enhanced greenhouse effect or when the intergenerational allocation of exhaustible resources are taken into account.

Ostwald influenced many authors, among them Henry Adams (1838–1918) who proposed a "law of acceleration" of the use of energy: "the coal output of the world, speaking roughly, doubled every ten years between 1840 and 1900, in the form of utilized power, for the ton of coal yielded three or four times as much power in 1900 as in 1840." Since Marxism is in this article under critical review, it is worthwhile pointing out that from the opposite camp, Karl Popper, in *The Poverty of Historicism,* which is a polemic against Marxist laws of history, thought it worthwhile to attack also Henry Adams in a footnote, without even mentioning the word "energy" because he dared propose a historical law. Henry Adams was certainly not a Marxist; on the contrary he was a Boston aristocrat who thought the world would probably come to a bad end socially and technologically.

When the anthropologist Leslie White (1943) wrote on energy and the evolution of culture inspired by Morgan and Taylor as evolutionary anthropologists, he realized that Ostwald and Soddy were precursors of his views. He looked for stages of evolution in terms of harnessing energy in increasing magnitudes and in various forms, in a dialectical

framework between the technological level (defined by the availability and the efficiency in the transformation of energy), the social system (in terms of Marxist "relations of production"), and the cultural-symbolic system.

Marx's doubts on the benefits of economic growth, clearly expressed in his ecological critique of capitalist agriculture, was not forgotten within Marxism (think of Walter Benjamin) although the technological optimists, who believed in the development of the productive forces, predominated. One of most influential Marxist technological optimists of the 20th century was the historian of science J. D. Bernal. He was in the 1950s totally in favor of the "civil" use of nuclear energy, which Lewis Mumford was strongly criticizing already at the time. Mumford, described by Ramachandra Guha as the "forgotten American environmentalist" (forgotten in comparison to G. P. Marsh, John Muir, Gifford Pinchot, Aldo Leopold, and Rachel Carson), was a heir to Patrick Geddes, William Morris, and John Ruskin; he does not belong in the Marxist tradition.

Mumford's mentor, Patrick Geddes (1854–1932), a biologist and urban planner, early on attacked neoclassical economists such as Walras because they did not count flows of energy, materials, and waste, but he did not discuss Marx. Geddes proposed the construction of a sort of input-output table inspired by the *Tableau Economique* of the Physiocrat F. Quesnay. The first column would contain the sources of energy as well as the sources of materials that are used, not for their potential energy, but for their other properties. Energy and materials were transformed into products through three stages: extraction, manufacture, transport and exchange. Estimates were needed of the losses (dissipation and disintegration) at each stage. The quantity of the final product (or "net" product, in physiocratic terms) might seem surprisingly small in proportion to the gross quantity of potential product. Now, however, the losses at each stage were not accounted for in economic terms. The final product was not added value at all. It was the value remaining from the energy and materials available at the beginning once they had been through all three stages.

Geddes's scheme is relevant to the attempt by several authors to develop a theory of ecologically unequal exchange between the metropolitan centers and the world peripheries. In neoclassical economics, provided that markets are competitive and ruled by supply and demand, there cannot be unequal exchange. This could only arise from monopoly or monopsony conditions, or because of noninternalized externalities (or excessive discounting of the future). In an ecological-economics theory of unequal exchange, one could say that the more of the original exergy (available energy or "productive potential" in the exported raw materials) has been dissipated in producing the final products or services (in the metropolis), the higher the prices of these products or services. This was indeed implied by Geddes with different words. Thus, Hornborg concluded, "market prices are the means by which world system centers extract exergy from the peripheries," sometimes helped, one must say, by military power.

A social metabolic perspective shows that capital accumulation does not take place by itself, and it is not only based on the exploitation of labor. Capitalism must advance into commodity frontiers because it uses more materials and energy, therefore it produces more waste, and it undermines not only its own conditions of production but the conditions of livelihood and existence of peripheral peoples, who complain accordingly. Such ecological distribution conflicts are more and more visible. They cannot be subsumed under the conflict between capital and labor.

SEE ALSO THE FOLLOWING ARTICLES

Aggregation of Energy • Conversion of Energy: People and Animals • Ecological Footprints and Energy • Economic Thought, History of Energy in • Ecosystem Health: Energy Indicators • Ecosystems and Energy: History and Overview • Energy in the History and Philosophy of Science • Heterotrophic Energy Flows • Industrial Agriculture, Energy Flows in • Industrial Ecology • Photosynthesis and Autotrophic Energy Flows • Thermodynamics and Economics, Overview • Thermodynamic Sciences, History of

Further Reading

Ayres, R. U. (1989). Industrial metabolism. *In* "Technology and Environment," (J. Ausubel, Ed.). National Academy Press, Washington, DC.

Ayres, R. U., Ayres, L., and Warr. B. (2002). Exergy, power and work in the US economy, 1900–1998. CMER-INSEAD, working paper, 2002.

Cipolla, C. (1974). "The Economic History of World Population." 6th ed. Penguin, London.

Cleveland, C. J. (1987). Biophysical economics: Historical perspectives and current recent trends. *Ecol. Modelling.*

Cohen, J. (1995). "How Many People Can the Earth Support?" Norton, New York.

Debeir, J. C., Del/age, J. P., H/mery, D. (1986). Les servitudes de la puissance Une histoire de l'energie, Flammarion, Paris.
Engels, F. (1972). Dialektik der Natur (1925), MEW, vol. 20, Dietz Verlag, Berlin.
Fischer-Kowalski, M. (1998). Society's metabolism: The intellectual history of materials flow analysis, Part I: 1860–1970, Part II (with W. Huettler): 1970–98. *J. Industrial Ecol.* **2**(1) and **2**(4).
Foster, J. B. (2000). Marx's ecology. Materialism and nature. *Monthly Rev. Press*.
Fluck, R. C., and Baird, D. C. (1980). "Agricultural Energetics." Avi, Westport, CT.
Geddes, P. (1885). An analysis of the principles of economics. Proceedings of the Royal Society of Edinburgh, 17 March, 7 April, 16 June, 7 July 1884, repr. Williand and Norgate, London.
Gootenberg, P. (1993). "Imagining Development: economic ideas in peru's "Fictitious Prosperity" og guano, 1840–80." University of California Press, Berkeley, CA.
Haberl, H. (2001). The energetic metabolism of societies, Parts I and II. *J. Industrial Ecol.*
Hall, C. h., Cleveland, C. J., and Kaufman, R. (1986). "Energy and Resources Quality: The Ecology of the Economic Process." Wiley, New York.
Hornborg, A. (1986). Toward an ecological theory of unequal exchange: Articulating world system theory and ecological economics. *Ecol. Econ.* **25**(1), 127–136.
Jenssen, O. (ed.) (1925). Marxismus und Naturwissenschaft: Gedenschrifts zum 30. Todestage des Naturwissenschaftlers Friedrich Engels, mit Beitr…gen von F. Engels, Gustav Eckstein und Friedrich Adler, Verlagges. des Allgemeinen Deutschen Gewerkschaftsbundes, Berlin.
Kormondy, E. J. (1965). "Readings in Ecology." Prentice-Hall, Englewoods Cliffs, NJ.
Leach, G. (1975). "Energy and Food Production." IPC Science and Technology Press, Guildford.
Lenin, V. I. (1913). The working class and Neomalthusianism. *Pravda* **137**. (Reprinted in Collected Works, vol. 19.)
Martinez-Alier, J. (2002). "The Environmentalism of the Poor: A Study of Ecological Conflicts and Valuation." Edward Elgar, Cheltenham UK, Northampton, MA.
Martinez-Alier, J., Munda, G., and O'Neill, J. (1998). Weak comparability of values as a foundation for ecological economics. *Ecological Economics* **26,** 277–286.
Martinez-Alier, J., and Schlfpmann, K. (1987). "Ecological Economics: Energy, Environment and Society." Blackwell, Oxford.
Marx, K. (1969). Das Kapital, vol. I 1867, vol. III, 1894, Ullstein Verlag, Frankfurt-Vienna-Berlin.
Marx, K., and Engels, F. (1976). Lettres sur les sciences de la nature et les mathematiques, Mercure de France, Paris.
Masjuan, E. (2000). La ecolog a humana y el anarquismo ibrico: el urbanismo "orgnico" o ecol gico, el neo-malthusianismo y el naturismo social, Icaria, Barcelona.
Masjuan, E. (2003). Neomalthusianesimo e anarchismo in Italia: un capitulo della storia del pensiero ecologico dei poveri? *Meridiana* (Rome).
Mayer. J. R. (1893, 1911). Die organische Bewegung in ihrem Zusammenhang mit dem Stoffweschsel, Heilbronn, published also in Die Mechanik der W…rme: gesammelte Schriften, Stuttgart, 1893, and in W. Ostwald, Klassiker der exacten Naturwissensschaften, Akademische Verlag, Leipzig, 1911.
McNeill, J. R. (2000). "Something New under the Sun: An Environmental History of the Twentieth-Century World." Norton, New York.
Moleschott, J. (1850). "Lehre der Nahrungsmittel." F*f*r das Volk, Enke, Erlangen.
Moleschott, J. (1851). Physiologie des Stoffwechsels in Planzen und Thieren, Erlangen.
Moleshott, J. (1852). "Der Kreislauf des Lebens." Von Zabern, Mainz.
Mouchot, A. (1869, 1879). La chaleur solaire et ses applications industrielles. Gauthier-Villars, Paris, 1869; 2nd ed., 1879.
Naredo, J. M., and Valero, A. (1999). "Desarrollo econ mico y deterioro ecol gico." Argentaria-Visor, Madrid.
O'Neill, J. (1993). "Ecology, Policy and Politics." Routledge, London.
O'Neill, J. (2002). Socialist calculation and environmental valuation: Money, markets and ecology. *Science and Society*, spring.
Odum, H. T. (1971). "Environment, Power and Society." Wiley, New York.
Pfaundler, L. (1902). Die Weltwirschaft im Lichte der Physik, Deutsche Revue **22,** April–June, 29–38, 171–182.
Pimentel, D., *et al.* (1973). Food production and the energy crisis. *Science* **182,** 443–449.
Pimentel, D., M. (1979). Food, Energy and Society. Arnold, London.
Rappaport, R. (1967). "Pigs for the Ancestors: Ritual in the Ecology of a New Guinea People." Yale University Press.
Ronsin, F. (1980). La grve des ventres. Propagande no-malthusienne et baisse de la natalit/en France, 19–20 sicles, Aubier-Montagne, Paris.
Schmidt, A. (1978). "Der Begriff der Natur in der Lehre von Marx." 3rd ed. EVA, Frankfurt-Cologne.
Sieferle, R. P. (1982). Der unterirdische Wald. Energiekrise und industrielle Revolution, Beck, Munich (English trans., White Horse Press, Cambridge, United Kingdom, 2001).
Susiluoto, I. (1982). "The origins and development of systems thinking in the Soviet Union." Political and philosophical controversies from Bogdanov and Bukharin to present-day reevaluations, Suomalainen Tiedeakatemia, Helsinki.
Vernadsky, V. (1924). La G/ochimie. Alcan, Paris.
Weber, M. (1968). Energetische Kulturtheorien, *Archiv ffr Sozialwissenschaft und Sozialpolitik,* 29, 1909, repr. in Max Weber, Gessamelte Aufs…tze zur Wissenschatslehre, 3rd ed., J. C. B. Mohr (Paul Siebeck), T*f*bingen.
White, L. (1943). Energy and the evolution of culture. *American Anthropol.* **45**(3).
White, L. (1959). The energy theory of cultural development. *In* "Readings in Anthropology" (M. H. Fried, Ed.), vol. II, pp. 139–146. Thomas Y. Cromwell, New York.

Material Efficiency and Energy Use

EBERHARD JOCHEM
Swiss Federal Institute of Technology
Zurich, Switzerland
Fraunhofer Institute for Systems and Innovation Research (FH-ISI)
Karlsruhe, Germany

1. **The Conceptual Approach**
2. **Energy Saving Potentials by Reduced and Changed Demand for Materials**
3. **Joint Impact of Material Efficiency, Substitution, and Intensified Product Use and Plants**

Glossary

dematerialization The declining trend of material use per capita in highly industrialized countries. This is due to the facts that a large capital stock and infrastructure are already in place and that reinvestments tend to need less material for the same service (material efficiency) or the capital stock is used more efficiently (intensified use of durables and plants).

energy intensity The relation between primary energy use and the gross domestic product (GDP) of a country in a given year.

intensified use of durables and plants The same quantity of material services is able to be carried out with fewer durables (goods, vehicles, machinery, and plants) because of more intensive use. For this purpose, additional services (e.g., car sharing, solvent service, and harvesting service) and/or additional organization (e.g., contracting and just-in-time delivery) are needed.

material efficiency The same function of a product, a material, or a technical system is performed with less material due to more intelligent design or improved properties of the materials used. Material efficiency has the largest relevance for energy use with regard to vehicles and moving parts.

material intensity The relation between primary material demand and the GDP of a country in a given year. Sometimes, the expression is defined as the relation between primary material demand and the population of a country.

material substitution The substitution of a material by another material or a different system that delivers the same or similar functions to the application considered (e.g., e-mail instead of mailing letters or books). In this context, the energy use for producing the two optional materials and for their operation during their lifetimes is compared; a net energy gain is expected in the case of material substitution.

The more efficient use of materials by improved construction design and material properties has a long success story with regard to reducing the energy demand of energy-intensive materials, vehicles, and moving parts of machines and plants. However, this more efficient use of materials, which can reduce the specific material and energy demand by approximately 1% per year, is scarcely known or perceived as a challenging development. Until recently, it was not even considered to be a promising option for energy and climate change policy in many countries. In addition to the technological option of material efficiency, the substitution of energy-intensive materials by less energy-intensive materials and the intensification of product and plant use are other options that may play a more important role in reducing the quantity of materials produced and hence energy demand. Entrepreneurial innovations will support these technical options to achieve intensified uses of machinery and vehicles by pooling (leasing, renting, and car sharing). The impact on reduced primary energy demand may range from approximately 0.3% per year (autonomous technical progress in material efficiency without specific policies) to more than 0.6% per year if supported by various policies oriented toward sustainable development.

 © 2004 Elsevier Inc. All rights reserved.

1. THE CONCEPTUAL APPROACH

Considerations of future improvements in energy efficiency and energy demand often focus on energy-converting technologies and the distribution of grid-based energies where the energy losses amount to approximately 60% of primary energy in most economies. However, there are two additional areas for reducing future energy demand to which little attention is currently being paid (Fig. 1):

1. Energy losses at the level of useful energy (currently approximately 40% of the primary energy demand of industrialized countries) could be substantially reduced or even avoided through technologies such as low-energy buildings, membrane techniques or biotechnological processes instead of thermal processes, and lighter vehicles or reuse of waste heat.
2. The demand for energy-intensive materials could be reduced by recycling or substituting these materials with less energy-intensive ones; reusing products or product components; improving their design or material properties; and the intensified use of products, plants, and vehicles through pooling (e.g., car sharing and leasing or sharing of machines and energy-intensive plants).

This article focuses on some of the latter aspects (excluding material recycling and product reuse). Many energy demand analyzes and projections implicitly assume that the demand for energy services is a "given market demand" that has to be fueled by the various forms of useful and final energies. However, the demand for energy-intensive products and materials is not a fixed, noninfluenceable energy service; on the contrary, the demand reflects a certain status of consumer preferences, choices, and behavior; market traditions and policy boundary conditions; the division of labor in production; and technology, which can be positively influenced by many technical options and entrepreneurial innovations without reducing the services delivered by the (technically changed) products, machinery, or vehicles. Examples and their impact on energy demand are given in the following sections.

The concept of dematerialization, reducing energy demand by using less primary material, was developed more than 20 years ago. This concept integrates the different options of reducing energy services and is often measured by the indicator of material intensity, the ratio of material use in physical terms to gross domestic product (GDP) in deflated constant terms. Malenbaum was one of the first researchers to use this indicator at a global level and in 1978 made preliminary projections for the following two decades. Dematerialization is foremost a long-term concept, tied to long-term changes in the nature of technology, infrastructure, and the economy of a country. Care is required to distinguish it from short-term changes in material intensity resulting from business cycle effects. Empirical dematerialization analyses published by Bernadini and Galli in 1993 concluded two basic postulates:

1. The intensity of use of a given material follows the same bell-shaped pattern for all economies. At first, in its industrializing phase, it increases with per capita GDP, reaches a maximum, and eventually declines afterwards; the growth in material use declines to lower rates than GDP. The decline is

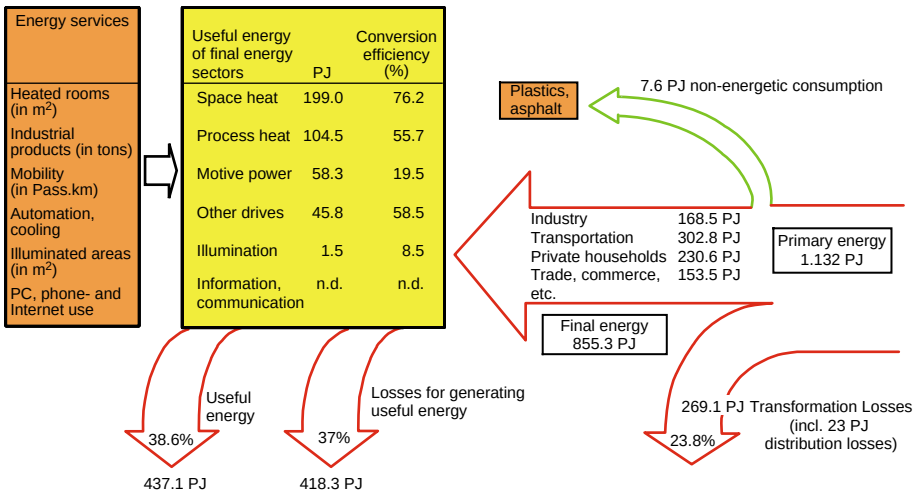


FIGURE 1 More efficient use of materials to reduce energy services demonstrated by the Swiss energy system from energy services to primary energy demand. Source: ISI, Karlsruhe Germany.

frequently accelerated by the growth of competing new materials with process, product quality, and cost advantages (e.g., aluminum or plastics instead of steel or glass).

2. The maximum intensity of material use declines the later it is attained by a given country, reflecting the technological learning and globalization of technical know-how over time.

These two postulates have been found to be valid for individual materials, and they also apply to primary energy intensity, which shows a very similar bell-shaped development. Integrating the substitution and life cycle concept of materials into a total material intensity, a further contribution to the decline of the total material intensity stems from saturation and the development of new markets based on more sophisticated intermediates and commodity products with a relatively higher value added per weight of material input.

For the total material use per capita (t/cap.) and year (including material demand for imported raw materials) of mature economies, it seems that its maximum level of approximately 80–90 t/cap. at an annual per capita GDP level of $25,000 has been reached. It is at this stage that per capita total material use may level off and even begin to decline, as Biondi and Galli found in 1992. This was confirmed by Cleveland and Ruth in 1999; see also declining U.S. figures from 100 t/cap. in the 1970s to 85 t/cap. in the 1990s published by Bringezu. This author and others, such as Rogers, stress that this trend of material use can be influenced by relevant product, recycling, and material-related policies of governments and companies in industry, trade, and other service sectors.

These changes in material intensity and their impact on energy demand were collectively treated in energy demand modeling and forecasts as intra- and interindustrial structural changes. Only recently have they been explicitly analyzed at technical levels of reduced energy and avoided greenhouse gas emissions in the national communications of the reporting process within the Framework Convention of Climate Change.

2. *ENERGY SAVING POTENTIALS BY REDUCED AND CHANGED DEMAND FOR MATERIALS*

The close relationship between the efficient use of energy-intensive materials and energy demand did not receive much attention from energy analysts and policy makers in the past, even though approximately one-third of total industrial energy demand is required for material production and it is obvious that moving parts and vehicles would need less energy during their lifetime if they weighed less. In most energy demand projections to date, more efficient use of materials, products, vehicles, and production plants or the substitution of materials were not seen as flexible and policy relevant but, rather, as a fixed level of activity. The following are the options to reduce energy demand by reduced material intensity discussed in this article.

1. More efficient use of materials by better design and construction; improved properties of materials, oils, or solvents, or even foamed plastics or metals; or reduced use of energy-intensive products such as nitrogen fertilizers. This strategy is particularly important in the case of moving parts or vehicles because lighter construction may contribute to significantly lower energy demand over the lifetime of the particular application (e.g., cars).

2. Substitution of highly energy-intensive materials by less energy-intensive materials or even by other technologies (e.g., substitution of steel and cement/concrete by wood; synthetic fibers by biobased fibers; and newspapers, journals, or books by electronic news and information).

3. Intensified use of products, machinery, and vehicles through short-term leasing, sharing, or renting, reducing the demand for materials by diminishing the nonused capital stock that may even include more efficiently used floor areas of residential and office buildings. This option of the strategy "renting instead of owning" seems to be a purely entrepreneurial and organizational issue at first, but it may also involve new supporting technologies (e.g., clear documentation by electronic devices on the use of the product or vehicle to identify responsibilities during its operation and to develop adequate rental tariffs).

All three elements contribute to structural changes within industrial production, mostly in the direction of lower energy intensity of total industrial production, and they offer numerous opportunities in the service sector to sell product services instead of products.

2.1 More Efficient Use of Materials

Energy-intensive materials as semifinished or final mass products make up a substantial part of

industrial energy demand; of particular interest are paper, glass, steel, aluminum, plastics, rubber products, solvents, and fertilizers. When used in mobile applications, they also contribute to high energy demand, particularly in transportation. Material efficiency means that less material per product function is needed in the future. The direct energy use for primary production of energy-intensive materials is more than 5% of the primary energy demand of an industrialized country, and the indirect energy use for transportation and moving parts (e.g., lifts, moving staircases, and machine components) amounts to more than 25–30% of primary energy demand. In total, approximately 35–40% of the global primary energy use is related to the production of materials, according to estimates of Hekkert and colleagues made in 2000. Examples of past and future increased material efficiency were published by the Enquête Commission of the German Parliament in 2002 and include the following:

- Lower weight of mass papers per square meter: This has been reduced by 15% during the past 30 years.
- Lower glass weight per bottle: This has been reduced by 30% during the past 30 years (ultralight glass bottles with PET films for safety reasons are entering the market).
- Lower weight steel products [e.g., cans (reduced by 50% during the past 30 years), vehicle steel sheets, machines, and appliances]: Competition between steel, aluminum, and plastics has resulted in revolutionary changes in steel shaping and finishing technologies. Powder metallurgy, thin casting, ion beam implantation, directional solidification, cold forging, cold automatic heating, and extruding have allowed savings up to 50% of steel input during the past 30 years;
- Lower specific weight of aluminum per application [e.g., packaging (reduced by 40% during the past 20 years), facades, window frames, vehicle sheets, and construction (airplanes, packing, and cans)].
- Lower specific weight of plastics through better quality plastics produced by improved catalysts and copolymers (e.g., the development of radial tires in the 1970s led to a 25% reduction in tire weight and doubled its lifetime; linear low-density polyethylene allows the use of thinner films in plastic products and savings up to 50%; in the future, the use of foamed plastics will reduce the weight of parts in vehicles, window frames, floor materials, and home and office textiles): Supercellular alloys of magnesium and aluminum and foamed nonferrous metals may contribute to less material input required for many applications.

Worrell and colleagues made the interesting observation in 1995 that the amount of nitrogen fertilizer used can be reduced by several measures applicable to many crops; the reduction potential is between 30 and 50% (e.g., for The Netherlands, from 230 to 130 kg/ha) within one decade.

2.1.1 Potentials and Drivers of Material Efficiency

Besides cost reduction potentials, material efficiency will also be enhanced by increasing the useful life, for example, by improving surface layers protecting buildings and infrastructure against ultraviolet radiation, acid rain, algae growth, and pollution. Surface protection is of particular interest in coastal regions with salty air. Other technological developments rely on the progress in microelectronics, information, and communication technology—miniaturization, design and quality control, and improved logistics. Computer-aided design and manufacturing has allowed the optimization of the number of components that can be produced by a given quantity of material, leading to a significant reduction of waste or to fewer unfinished or rejected goods.

As indicated by these examples, the strategy of reducing a product's material use while maintaining identical functionality is not a new development. In particular, much attention has been paid to post-consumer wastes by national policies or by the European Union regarding the reduction of packaging wastes. The first analytical work by Gielen in 1999 suggests that material efficiency has improved in the past decades by approximately 1% per year on average, which indicates a comparable rate of energy efficiency improvement. The question is whether these trends can be maintained or even accelerated. Competition among the materials and legal frameworks that strive for the internalization of the external costs of material production, waste disposal, and energy conversion will be a major driver for maintaining progress in material efficiency.

The future energy saving potential stemming from improved material efficiency has not been systematically analyzed. However, the methodological approach used to analyze the impact of material efficiency on energy demand (and energy-related emissions) has been developed and described by several authors, such as Worrell and colleagues in 1995 and Hekkert and colleagues in 2000. These

authors have also performed a number of case studies for several basic materials, products, and packaging materials. Tentative analyses suggest two major conclusions:

1. The energy savings during the lifetime of vehicles and moving components is approximately three times higher than the savings for reduced material demand.
2. The direct and indirect primary energy savings from the more efficient use of materials range between 0.1% per year (autonomous technical progress) and, according to the findings of the Enquête Commission, approximately 0.25% per year if technical progress is supported by relevant policies.

2.1.2 Obstacles and Issues of Research and Development

Many of the new materials and their specific properties and production processes are known today, but, as for energy efficiency potentials, various obstacles hinder the rapid diffusion of many new materials and the related benefit of reduced energy demand. Because many of the new material technologies are based on expensive, high-quality materials, price is often the most common barrier to their widespread implementation. For this reason, apart from research and development (R&D), Ayres stresses the fact that the promotion must also integrate favorable economic and policy conditions, including innovation and market introduction policies.

In some cases, there are conflicts of interest that have to be taken into account for future market introduction and acceptance. There may also be problems in the transition phase from today's products to light material products (e.g., safety aspects of heavy traditional cars and new light cars and road vehicles in cities).

In order to realize the high-end energy saving potential of material efficiency in the coming decades, some of the research involved is very basic applied research (e.g., foamed plastics or even metals and alloys of magnesium or aluminum). The R&D for the different applications seems to be very dispersed but is often very advanced applied research. In many cases, the higher quality of the improved materials and their applications require high-performance production machines and high production expertise; for example,

- R&D on new polymer catalysts, particularly for polyethylene, polypropylene, polystyrene, polyester, and interesting copolymers, to improve their properties and to substitute specialized, more energy-intensive plastics (polycondensates and polyadducts).
- R&D on improved surface protection to avoid corrosion and/or to increase the life span of the products, facades, or vehicles, including economic optimization over the product's and component's lifetime.
- R&D on foamed plastics and nonferrous metals for lightweight construction, particularly for moving parts and vehicles.

2.2 Substitution of Energy-Intensive Materials

The specific energy demand for producing materials can vary significantly. Replacing mass-produced, energy-intensive materials with less energy-intensive materials can generate a large energy saving potential. This depends on the specific substitution involved, where different product life spans may also have an influence. Substantial energy savings can be achieved by substitution of

- Heavy materials with lightweight materials in the transport sector. Conventional steel constructions, for example, may be replaced with lighter aluminum or magnesium constructions. Although the production of aluminum or magnesium components is more energy-intensive, this additional energy consumption can be compensated for during the components' lifetime. The average weight of U.S. cars declined approximately 20% between 1976 and 1990; most of this decline was accounted for by the use of aluminum, plastics, and composites as substitutes. Schön and colleagues concluded that this past achievement is likely to be repeated in the next 20 years by further substitution using nonferrous metals and plastics.
- Energy-intensive metals, paper, and glass with plastics; their advantages are often lighter weight, greater resistance to tear, relative inertness to moisture and organic environments, and less storage space requirements. Not only do plastics continue to substitute for metals, glass, and paper in traditional applications but also new polymers with improved properties (e.g., high strength-to-weight ratio, transparency, greater durability, and specific surface properties) are increasingly replacing older generation plastics widely in use today. These new polymers often have a lower manufacturing energy demand than the materials they displace when the comparison is made in terms of end-use qualities rather than on a unit weight basis.

• Petrochemically based plastics with biobased polymers or plastics or natural materials. Well-established substitution routes are oleochemical surfactants and lubricants from vegetable oils, as Patel and colleagues reported in 2000. Future examples are the increased use of wood as a construction material, the renaissance of flax and hemp as natural fibers instead of the use of chemical fibers, manufacturing and cooling oils, and bioplastics produced from starch. In the long term, Hüsing and colleagues expect major innovations such as polyhydroxyalkanoates from gene-manipulated plants, which have similar properties as those of the petrochemically based polypropylene.

There are many other small substitutions of materials, such as the substitution of copper in the telecommunication market with optical fibers with 30–40 times the transmission capacity of conventional wiring and by mobile telecommunications, substitution of metals with ceramics (e.g., spark plugs, insulation layers, titan oxide-based solar cells, superconductors, and high-temperature protection layers), and substitution of printed media with electronic information and communication.

A fundamental connotation of substituting old materials with new ones is the trend from commodities to specialty materials adapted for specific products, such as car frames or other car components and biodegradable packaging materials, specific functions of surfaces (e.g., zero corrosion and improved heat transfer), or specific mechanical properties per weight of material. These advanced materials (e.g., produced from nano powder) are becoming more important as they substitute for basic traditional metallic materials or plastics.

2.2.1 Potentials and Opportunities

The future energy saving potentials stemming from material substitution have not been systematically analyzed; the evidence for the energy saving effect of material substitution is restricted to a limited number of metals and plastics applied to a limited number of products. The impact of bioengineered basic natural materials, for example, is an unresolved issue but theoretically implies high energy saving potentials. In addition, the competition between different materials within the substitution process and the simultaneous process of material efficiency makes the analysis of energy saving potentials extremely complex. Initial tentative and partial analyses suggest results similar to those of the impact studies on material efficiency:

1. The energy savings due to substituted materials during the lifetime of vehicles and moving components are approximately three times higher than the energy savings due to material substitution in other fields of application.
2. The direct and indirect primary energy savings from material substitution are estimated to be less than 0.1% per year (autonomous substitution) and 0.2% per year or more if the technical progress by material substitution is supported by relevant policies. This potential may increase during the next decades due to new opportunities resulting from technologies such as nano- and biotechnology and genetic engineering.

2.2.2 Obstacles and Policy Issues

Traditional consumer wishes, the traditional material choice of producers, a certain "image" of a given material in its societal context, and sometimes the higher cost of the substituting and less energy-intensive material can adversely affect the realization of the energy saving potentials in various applications. Early R&D on biobased polymers and products, including applied genetic engineering to improve properties and yields (e.g., natural fibers, starch, and wood), may be important given the perspective of changing boundary conditions in the 2000 "World Energy Assessment," edited by Goldemberg and Johannson: increasing prices of oil and natural gas within the next two decades when global crude oil production is likely to pass its middepletion point. Reducing the production costs of the less energy-intensive materials by learning and economy-of-scale effects may also be an issue of public policy through limited financial incentives, procurement schemes, information and training, as well as by setting technical standards to be met by the new materials.

2.3 Intensified Use of Products, Machinery, and Vehicles

Durables and investment goods are generally purchased but could often be leased, rented, or hired, particularly in cases of small annual usage times (e.g., cars, 200–500 h per year; harvesting machines, <50 h; construction machines, <500 h; and many high-temperature industrial processes in small and medium-sized companies operating on one-shift schedules or <2500 h per year). Regarding material use (and the imbedded energy to produce durables and investment goods), Friend stated in the mid-1990s that the low annual usage time implies a huge idle capital and

material stock of an economy due to the concept of selling and owning products instead of selling and buying product-based services. In the case of high-temperature industrial processes, the energy losses for daily startup and shutdown may be substantial.

The strategy of product-based services of manufacturers or service companies is based on the concept of selling the total service (e.g., harvesting, partial outsourcing of energy-intensive production steps, using centralized voice mail instead of answering machines, and renting solvents from chemical suppliers) or pooling durables and investment goods. Mont describes many cases in great detail in which the intensified use of the durables and investment goods reduces their capital cost and compensates for the additional service cost for either the total service or contracting, operating, reservation, billing, controlling, and additional insurance. In many cases, ecological or environmental arguments and considerations contribute to the acceptance of these new services, which are labeled ecoefficient services by Popov and DeSimone. However, very seldom are these product-based services examined from the perspective of material efficiency and indirectly reduced energy use.

At first glance, using products, machinery, and vehicles more intensively through short-term leasing, sharing, or renting and its technical dimension of material efficiency (and, indirectly, of energy efficiency) seems to be a research area more related to the social sciences. In order to realize the strategy "renting instead of owning" at the commercial level, however, experience strongly suggests the need for supporting technologies in many cases. These technologies would offer the user clear documentation on the use of the product or vehicle, identify responsibilities during its operation, and develop adequate rental tariffs. These predominantly information and communication technologies require research focused on small, inexpensive, and reliable electronic systems.

2.3.1 Potentials and Opportunities

The realizable direct energy efficiency potential of these numerous options of intensified use of products, vehicles, and production plants has scarcely been analyzed and depends very much on the future acceptance and market diffusion of services such as car sharing (studied by Meijkamp in The Netherlands), pooling public (municipal) or company vehicle fleets or machinery (studied by Fleig), and moving elderly people into adequate but appropriately sized apartments (first analyzed by Zanger and colleagues in 1999). Assuming moderate market shares of two of these new services in 2050 (e.g., 10% of car-using households opt for car sharing, and 2% of elderly people move to smaller apartments with an average gain of $20\,m^2$ per move and apartment), the energy savings in industrialized countries may be on the order of 3% of current levels of primary energy demand. Increased renting of products and vehicles may indirectly affect energy use through, for example, the higher energy efficiency of "younger" capital stock or vehicle fleets, walking or cycling displacing the car for short trips, and a more sensible use of machines or vehicles due to the accompanying documentation system. The indirect effects are estimated to be approximately 10% of the direct effects.

The impact of these numerous possibilities of intensifying the use of products and production machinery and plants on energy demand is almost unknown because very few existing services have been assessed under energy aspects and many conceivable services to intensify the use of products, vehicles, or production facilities have not been offered on suitable markets. The estimates of known and offered services thus assume yearly primary energy savings of less than 0.02% per year (autonomous substitution), but in the longer term the Enquête Commission estimated that they may contribute to the annual energy savings of 0.1–0.2% of total primary energy demand if relevant incentives and policies are in place.

2.3.2 Obstacles and Policy Issues

To realize the full impact of intensifying the use of products, vehicles, and apartments, new entrepreneurial forms are also required that reduce the capital stock for those applications in which intensified use is being targeted.

Obviously, much research must be undertaken to develop a clearer understanding of the possible potentials of energy savings:

- Research on customer acceptance: In many cases, the various services will be innovative—for example, incentives for machinery and car sharing or improved awareness about the fixed and capital costs of owning machinery and vehicles with low annual operating hours.
- Research on the willingness of innovative companies to invest in pooling services: R&D on inexpensive accompanying information and technology-based monitoring and control techniques for easy and fair pricing, easy scheduling and access, and protection against theft.

New services may first require pilot projects with socioeconomic evaluations in order to design the

necessary policy and financial boundary conditions, clarify legal questions, and design professional training and educational programs.

3. JOINT IMPACT OF MATERIAL EFFICIENCY, SUBSTITUTION, AND INTENSIFIED PRODUCT USE AND PLANTS

Although the relation between these three elements of sustainable material use on energy demand has only been partially analyzed, the results of analytical technology-specific studies suggest substantial energy saving potentials. These potentials may materialize less quickly if policy action is not taken by governments, the scientific community, and the industrial players (Table I). It seems that the energy saving potential of the strategy of material efficiency may be most important during the next few decades but will have decreasing importance thereafter, whereas the potentials of material substitution and intensification of the use of products, vehicles, and production plants may require much additional research (e.g., on biomass-based plastics) and changes in social values and decision making (renting instead of owning) before they can substantially contribute to a reduced use of energy. However, the energy saving potentials of these two options may be quite high in the long term (i.e., over many decades during this century).

TABLE I

Estimates of Annual Primary Energy Saving Potential from Material Efficiency, Substitution, and Intensified Use of Products, Vehicles, and Production Facilities

Energy-related material aspect	Annual primary energy saving potential (%)		Comment
	Autonomous development	Policy-induced development	
More efficient use of materials	0.1	Up to 0.25	Relatively well analyzed, but still unknown potentials
Substitution of materials	<0.1	Up to 0.20	Open to major technical break throughs (biomass, nanotechnology)
Intensified use of durables, vehicles, etc.	<0.02	Up to 0.20	Unknown potential due to changes in societal values
Total effect	<0.2	Up to 0.5 or 0.6	

Preliminary estimates of the energy saving potentials for the next three decades have been made for an industrialized country such as Germany. The results showed approximately 450 PJ saved in the reference scenario, which is expected to occur without additional policy activities and is reflected in the assumptions on structural change in the economy (Table II). This energy saving potential corresponds to an annual decrease in the total primary energy demand of Germany of 0.12% per year. In the policy scenario, called the sustainability scenario, the energy saving potential was assumed to double. Approximately half the savings are estimated to come from improvements in road vehicles [material efficiency (light vehicles) and car sharing/truck leasing]. Other major savings are expected to result from the lighter construction of machinery and plants and the substitution of primary aluminum, which is a light but energy-intensive material.

There are indications that recycling seems to play a smaller role compared to material efficiency and material substitution mainly due to the savings in vehicles and moving parts of machinery and to the substitution of primary aluminum. The contribution of the intensified use of products, vehicles, and production plants is certainly underestimated, but analytical and empirical work still has to be done in many product areas to obtain reliable estimates of energy savings in this area of activity and entrepreneurial innovations. The results do not reflect changes in exports and imports of energy-intensive products, which may affect the saving potentials at the national level of an economy. Finally, the integration of the four options cited in Table II is highly dependent on socioeconomic conditions; hence, the joint effect on energy savings will depend on the constellation of the four options.

There is no doubt that the energy saving potentials will very much depend on future progress in the R&D of improved properties of existing materials, such as steel, nonferrous metals, cement, glass, plastics, and fibers, as well as the invention, market introduction, and diffusion of new materials (e.g., plastics or fibers based on biomass that may be genetically designed or modified). In addition, the intensified use of products, vehicles, and production plants will be dependent on changes in social values and entrepreneurial innovations that must

TABLE II

Impact of Material Recycling, Efficiency, Substitution, and Intensified Use of Products and Production Facilities on Primary Energy Demand: Reference and Sustainability Scenario for Germany, 2000–2030[a]

Material/group of product	Additional saving potential of primary energy until 2030, reference and sustainability scenario (PJ)									
	Intensified recycling of waste materials		Improved material efficiency		Material substitution		Intensified use of products, vehicles, etc.		Total	
	Ref.	Sustainability	Ref.	Sustainability	Ref.	Sustainability	Ref.	Sustainability	Ref.	Sustainability
Steel	30	60	5	30			n.a.	n.a.	35	60
Aluminum	23	43			100[b]	240[b]	n.a.	n.a.	83	163
Plastics	30	90			10	20	n.a.	n.a.	40	110
Cement/concrete, bricks			n.a.	n.a.	2	10	0	0	2	10
Bitumen, asphalt	5	11			n.a.	n.a.	0	0	5	11
Glass	5	10	5	25	5	10			5	45
Paper	11	40	11	55	10	50			32	145
Road vehicles										
Production	n.a.	n.a.	30	50	n.a.	n.a.	n.a.	n.a.	30	50
Use	n.a.	n.a.	250[b]	500[b]	n.a.	n.a.	45	110	170	360
Construction	n.a.	n.a.	80	120	n.a.	n.a.	n.a.	n.a.	80	120
Machinery/ plants, office and home appliances	n.a.	n.a.			n.a.	n.a.				
Dwellings	0	0	n.a.	n.a.	n.a.	n.a.	4	10	4	10
Total	104	254	256	530	77	210	49	120		
Weighted by 0.9	94	230	230	480	70	190	45	110	430	1020

[a] *Sources*. Enquête Commission (2002); and Schön *et al.* (2003). Ref, reference; n.a., not applicable.
[b] In total, only 50% of the value considered.

occur to realize the potentials expected in the coming decades.

The following conclusions can be made with regard to the preliminary findings of the relationship between the various options of sustainable material use and energy demand:

- The impact of material efficiency, material substitution, and the intensified use of durables has not been explicitly treated as an energy policy option in the past, nor has it been explicitly analyzed in energy demand projections. It has only been treated as an unexplained aggregate of structural change.
- If consciously adopted as a policy of sustainable development by interested companies in industry and the new service sectors, the impact of the three options for energy demand may be significant—an approximately 0.5% reduction in annual primary energy demand in the coming decades.
- The obstacles to these potentials are manifold, but the material and energy savings provide opportunities to recover the R&D expenses and additional investments; the value added in the new materials and systems may also be an incentive for companies as well as for governments.

The close relationship between material use and energy demand opens up an entirely new area of energy systems analysis and, more important, offers major entrepreneurial opportunities for a lighter, more efficient, and less polluting industrial society. It opens up new jobs in many service areas by substituting the use of natural resources by know-how-intensive products and services. These effects have been studied and calculated by Walz and colleagues for Germany, with the conclusion that more efficient use of material and capital stock is likely to contribute to a sustainable development in the long term.

SEE ALSO THE FOLLOWING ARTICLES

Cement and Energy • Consumption, Energy, and the Environment • Economics of Energy Efficiency • Energy Efficiency and Climate Change • Energy Efficiency, Taxonomic Overview • Industrial Energy Efficiency • Material Use in Automobiles • Obstacles to Energy Efficiency • Plastics Production and Energy • Recycling of Metals • Recycling of Paper • Steel Production and Energy

Further Reading

Ayres, R. U. (2001, November). The energy we overlook. *World Watch* **14**(6), 30–39.

Bernadini, O., and Galli, R. (1993). Dematerialization: Long-term trends in the intensity of use of materials and energy. *Futures* **25**(5), 431–448.

Biondi, L., and Galli, R. (1992). Technological trajectories. *Futures* **24**(7), 580ff.

Bringezu, St. (2000). "Use of Resources in Economic Regions [in German]." Springer, Berlin.

Enquête Commission (2002). Sustainable energy supply under the conditions of globalization and liberalization, Final report, No. BtgsDs. 14/0400, pp. 311–327. German Bundestag, Berlin.

Flaig, F. (ed.). (2000). "Sustainable Cycling Economy [in German]." Schäffer-Poeschel, Stuttgart.

Friend, G. (1994). The end of ownership? Leasing, licensing, and environmental quality. *New Bottom Line* **3**(11).

Gielen, D. J. Materialising dematerialisation: Integrated energy and materials system engineering for greenhouse gas emission mitigation. Thesis. Delft Technical University. The Netherlands

Hekkert, M. P., Joosten, L. A. J., Worrell, E., and Turkenburg, W. C. (2000). Reduction of CO_2 emissions by improved management of material and product use: the case of primary packaging. *Resources, Conservation and Recycling* **29** Issues 1–2, 33–64.

Hekkert, M. P., van den Reek, J., Worrell, E., and Turkenburg, W. C. (2002). The impact of material efficient end-use technologies on paper use and carbon emissions. *Resour. Conserv. Recycling* **36**(3), 241–266.

Hüsing, B., Angerer, G., Gaisser, S., and Marscheider-Weidemann, F. (2002). "Bio-technological Production of Basic Materials under Specific Aspects of Energy and Bio-polymers from Residues [in German]." Federal Environmental Protection Agency, Berlin.

Jochem, E., *et al.* (2000). End-use energy efficiency. *In* "World Energy Assessment" (J. Goldemberg and Th. Johannson, Eds.), pp. 173–217. UNDP/World Energy Council/UNDESA, New York.

Jochem, E., Favrat, D., Hungerbühler, K., Rudolph, V., Rohr, P., Spreng, D., Wokaun, A., and Zimmermann, M. (2002). "Steps Towards a 2000 Watt Society. Developing a White Paper on Research and Development of Energy-Efficient Technologies." CEPE/ETH, Zurich.

Maijkamp, R. (1998). Changing consumer behaviour through eco-efficient services: An empirical study on car sharing in The Netherlands. *Business Strategy Environ.* **7**, 234–244.

Malenbaum, W. (1978). "World Demand for Raw Materials in 1985 and 2000." McGraw-Hill, New York.

Mont, O. (2000). "Product-service systems. Shifting corporate focus from selling products to selling product-services: A new approach to sustainable development, AFR Report No. 288." Natur Vards Verket, Stockholm, Sweden.

Nakicenovic, N., Grübler, A., and MacDonald, A. (1998). "Global Energy Perspectives." Cambridge Univ. Press, Cambridge, UK.

Popov, F., and DeSimone, D. (1997). "Eco Efficiency—The Business Link to Sustainable Development." MIT Press, Cambridge, MA.

Rogers, W. M. (2000). "Third Millennium Capitalism: Convergence of Economic, Energy, and Environmental Forces." Quorum, Westport, CT.

Schön, M., *et al.* (2003). "System Analysis of Recycling Potentials of Energy-Intensive Materials and Their Impact on Efficient Energy Use [in German]." Fh-ISI, Karlsruhe, Germany.

Walz, R., *et al.* (2001). "The World of Labour in a Sustainable Economy—Analysis of the Impact of Environmental Strategies on the Economy and Labour Structures [in German]." Federal Environmental Protection Agency, Berlin.

Worrell, E., Meuleman, B., and Blok, K. (1995a). Energy savings by efficient application of fertilizer. *Resour. Conserv. Recycling* **13**(3/4), 233–250.

Worrell, E., Faaij, A. P. C., Phylipsen, G. J. M., and Blok, K. (1995b). An approach for analysing the potential for material efficiency improvement. *Resour. Conserv. Recycling* **13**(3/4), 215–232.

Zanger, C., Drengner, J., and Gaus, H. (1999). Consumer acceptance of increased life times of products and intensified use of products [in German]. *UWF* **1**, 92–96.

Materials for Solar Energy

CLAES G. GRANQVIST
Uppsala University
Uppsala, Sweden

1. **Solar Energy Materials**
2. **Fundamental Optical Properties of Materials**
3. **Transmitting and Reflecting Materials**
4. **Thin Films**
5. **Transparent Thermal Insulators**
6. **Solar Thermal Converters**
7. **Radiation Cooling**
8. **Solar Cells**
9. **Coatings for Glazings: Static Properties**
10. **Coatings for Glazings: Dynamic Properties**
11. **Solar Photocatalysis**

Glossary

cermet A mixture of ceramic and metal.
chromogenic Pertaining to the ability to change optical properties persistently and reversibly by the action of an external stimulus.
electrochromic Pertaining to the ability to change optical properties reversibly and persistently by the action of a voltage.
float glass Solidified glass floated on a bed of molten metal so that the glass surface is nearly atomically smooth.
infrared Referring to wavelengths of light longer than 0.7 μm.
iridescence A rainbowlike exhibition of color.
photocatalysis Chemical reactions on an "active" surface, stimulated by light.
photochromic Pertaining to the ability to change optical properties reversibly and persistently by the action of light irradiation.
Planck's law A relationship describing the spectral distribution of thermally emitted radiation.
smart window A window incorporating a chromogenic material.
spectrally selective Having distinctly different optical properties in different wavelength ranges.
thermochromic Pertaining to the ability to change optical properties reversibly and persistently by the action of temperature.
ultraviolet Referring to wavelengths of light shorter than 0.4 μm.

This article summarizes work on materials with properties tailored to meet the specific parameters of the spectral distribution, angle of incidence, and intensity of the electromagnetic radiation prevailing in our surroundings. Spectral and angular selectivity are introduced to specify many of the desired properties. Materials with optimized transmittance, reflectance, absorptance, and emittance are discussed. A general introduction of the properties of glasses, polymers, and metals is followed by an account of thin film technology, which is of importance for solar energy materials of many different kinds. The article treats materials for transparent thermal insulation, solar thermal converters, radiative cooling devices, solar cells, architectural windows with static as well as dynamic properties, and solar photo-catalysis. The aim is to introduce basic concepts and commonly used materials, and also to bring the exposition up to the work in today's scientific laboratories.

1. SOLAR ENERGY MATERIALS

Solar energy materials have properties tailored to meet the specific parameters of the spectral distribution, angle of incidence, and intensity of the electromagnetic radiation prevailing in our natural surroundings. Materials for thermal and electrical conversion in man-made collectors, as well as for energy-efficient passive design in architecture, are typical examples.

Figure 1 introduces the ambient electromagnetic radiation of Earth's environment in a unified manner. The most fundamental property of this radiation, which ensues from the fact that all matter emits electromagnetic radiation, is conveniently introduced

 © 2004 Elsevier Inc. All rights reserved.

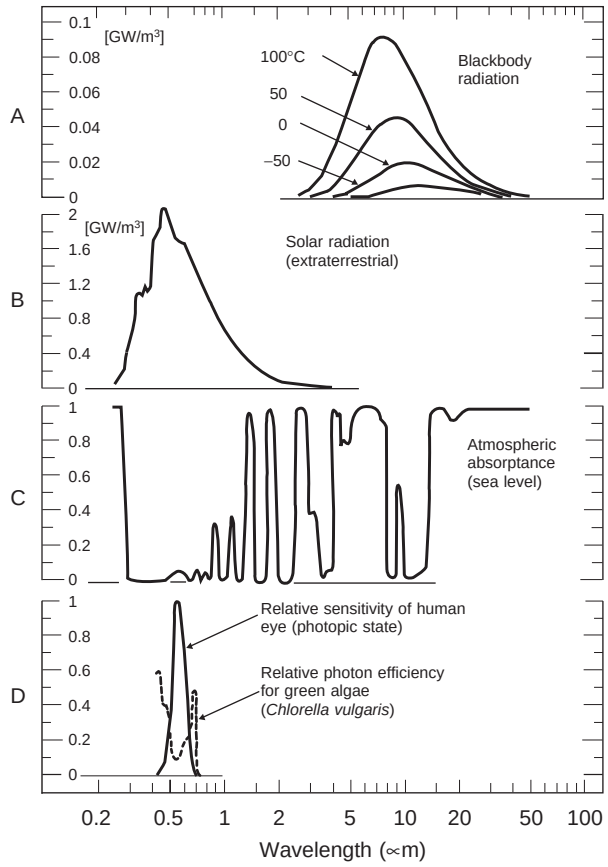


FIGURE 1 Spectra for (A) black-body radiation pertaining to four temperatures, (B) solar radiation outside Earth's atmosphere, (C) typical absorptance across the full atmospheric envelope, and (D) relative sensitivity of the human eye and relative photosynthetic efficiency for green algae.

by starting with the ideal blackbody. The emitted spectrum–known as the Planck spectrum–of blackbody radiation is uniquely defined if the temperature is known. Planck's law is a consequence of the quantum nature of the radiation. Such spectra are depicted in Fig. 1A for four temperatures. The vertical scale denotes power per unit area and wavelength increment (hence the unit GW/m^3). The spectra are bell-shaped and confined to the $2 < \lambda < 100$-μm wavelength range. The peak in the spectrum is displaced toward shorter wavelengths as the temperature increases. At room temperature, the peak lies at about 10 μm. Thermal radiation from a material is obtained by multiplying the Planck spectrum by a numerical factor (the emittance), which is less than unity. In general, the emittance is wavelength dependent.

A solar spectrum is depicted in Fig. 1B for radiation just outside Earth's atmosphere. The curve has a bell shape, defining the sun's surface temperature ($\sim 6000^\circ C$). It is important to observe that the solar spectrum is limited to $0.25 < \lambda < 3\,\mu m$, so that there is almost no overlap with the spectra for thermal radiation. Hence, it is possible to have surfaces with properties that are entirely different with regard to thermal and solar radiation. The integrated area under the curve gives the solar constant ($1353 \pm 21\ W/m^2$); this is the largest possible power density on a surface oriented perpendicular to the sun in the absence of atmospheric extinction.

Most solar energy conversion systems are located at ground level, and it is of obvious interest to consider to what extent atmospheric absorption influences solar irradiation and net thermal emission. Figure 1C illustrates a typical absorption spectrum vertically across the full atmospheric envelope under clear weather conditions. The spectrum is complicated with bands of high absorption–caused mainly by water vapor, carbon dioxide, and ozone–and bands of high transparency. It is evident that most of the solar energy can be transmitted down to ground level and only parts of the ultraviolet ($\lambda < 0.4\,\mu m$) and infrared ($\lambda > 0.7\,\mu m$) radiation are strongly damped. The maximum power density perpendicular to the sun is about $1000\ W/m^2$. Thermal radiation from a surface exposed to the clear sky is strongly absorbed except in the $8 < \lambda < 13$-μm range, where the transmittance can be large, provided that the humidity is moderately low.

Two biological conditions of relevance for solar-energy-related applications are illustrated in Fig. 1D. The solid curve shows the relative sensitivity of the human eye in its light-adapted state. A bell-shaped curve extends across the $0.4 < \lambda < 0.7$-μm interval, with its peak at 0.555 μm. Clearly, a large part of the solar energy comes as invisible infrared radiation. The dashed curve indicates that photosynthesis in plants operates at wavelengths in approximately the same range as those for the human eye, which is relevant for greenhouse applications.

Figure 1 depicts an important fact about ambient radiation: that it is spectrally selective, i.e., confined to specific and usually well-defined wavelength ranges. This property is of great importance for most types of solar energy materials. Another type of selectivity–known as angular selectivity–results from the fact that different angles may apply for different types of radiation; for example, solar radiation comes from a point far from the horizon during most of the day, whereas the visual contact between a person and his or her surroundings is often at near-horizontal lines-of-sight.

2. FUNDAMENTAL OPTICAL PROPERTIES OF MATERIALS

When electromagnetic radiation impinges on a material, one fraction is transmitted, a second fraction is reflected, and a third fraction is absorbed. Energy conservation yields, at each wavelength, that

$$T(\lambda) + R(\lambda) + A(\lambda) = 1, \quad (1)$$

where T, R, and A denote transmittance, reflectance, and absorptance, respectively.

Another fundamental relationship, also following from energy conservation and referred to as Kirchhoff's law, is

$$A(\lambda) = E(\lambda) \quad (2)$$

with E being the emittance, i.e., the fraction of blackbody radiation (cf. Fig. 1A) that is emitted at wavelength λ. It is evident that Eq. (2) is of relevance mainly for $\lambda > 3\,\mu m$.

It is frequently convenient to average the spectral data over the sensitivity of the eye, a solar spectrum (with or without accounting for atmospheric effects), or a blackbody emittance spectrum (for a certain temperature). Luminous, solar, or thermal values pertinent to the respective optical properties are then obtained; these are denoted X_{lum}, X_{sol}, and X_{therm}, with X being T, R, A, or E.

3. TRANSMITTING AND REFLECTING MATERIALS

3.1 Glasses

Glass can be used for protection against an unwanted environmental impact, for convection suppression, and as a substrate for surface coatings (normally referred to as thin films). Float glass, a highly standardized product used in a great majority of windows in buildings, is characterized by uniformity and flatness almost on the atomic scale. The latter feature is an outcome of the production process in which the molten glass is solidified on a surface of molten tin. The middle curve in Fig. 2 shows the spectral transmittance of 6-mm-thick standard float glass within the wavelength range relevant to solar radiation. A characteristic absorption feature at $\lambda \approx 1\,\mu m$, due to the presence of Fe oxide, limits T_{sol} to some extent. Special float glass is available with varying amounts of Fe oxide. Fig. 2 shows that low-Fe content leads to glass with very high T_{lum} and T_{sol} values, and also with substantial transmittance in the ultraviolet region.

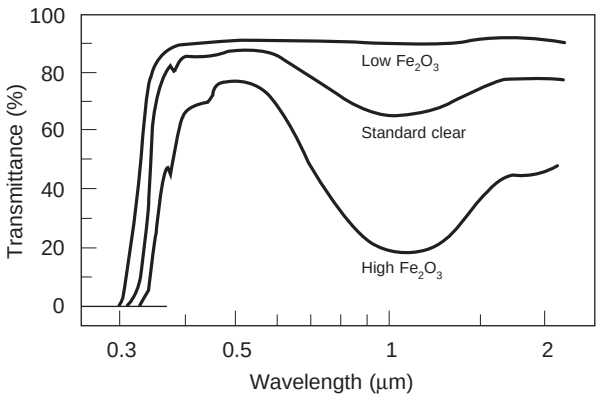


FIGURE 2 Spectral transmittance for float glass with three different amounts of Fe_2O_3.

Glass with a large Fe content limits T_{sol} whereas T_{lum} remains rather large; such glass has a distinctive greenish tint. The reflectance of each interface between glass and air is about 4% in the $0.4 < \lambda < 2$-μm range, implying that the maximum transmittance for a glass pane is 92%. Glass is strongly absorbing for $\lambda > 3\,\mu m$, and E_{therm} is as large as about 85%.

Laminated glass may be used for safety and other reasons. This glass comprises a layer of polyvinyl butyral sandwiched between two glass panes and bonded under heat and pressure. The laminate is an efficient absorber of ultraviolet light, and the transmittance at $\lambda < 0.38\,\mu m$ is almost zero for a layer thickness of the order of 1 mm. Photochromic glass is able to darken under irradiation of ultraviolet light from the sun and clears in the absence of such irradiation. Fatigue-free photosensitivity is normally accomplished by adding metal halides–particularly silver halide–to the vitreous matrix of the glass. A number of other additives are important as well. Figure 3 shows T_{lum} as a function of time for darkening and clearing of two photochromic glass plates. Darkening progresses rapidly and reaches about 80% after 1 minute. Clearing is much slower, though, and is still incomplete after 1 hour. The photosensitivity occurs almost entirely in the $0.4 < \lambda < 1.0$-μm interval, and hence the modulation of T_{lum} is much more pronounced compared to the modulation of T_{sol}. Photosensitivity exerts no effect on the reflectance or on E_{therm}.

3.2 Polymers

Many polymers can be transparent and can, in principle, replace glass. However, polymers degrade more easily than glass and hence have fewer

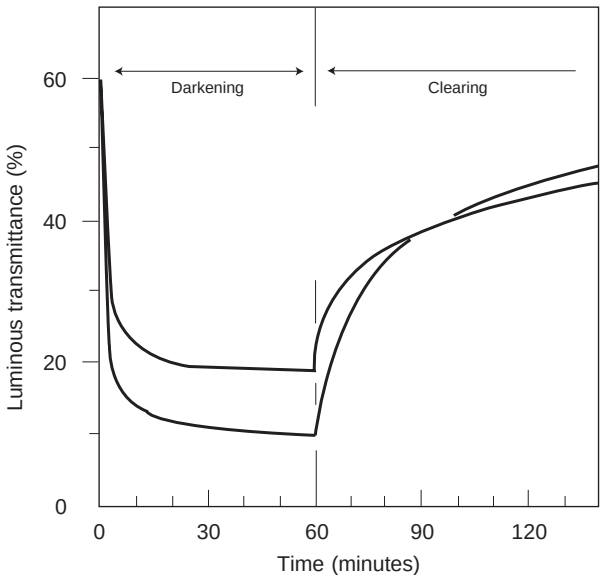


FIGURE 3 Luminous transmittance vs time for two photochromic glasses.

applications. Polyester foil deserves special mention because it can serve as a flexible and inexpensive substrate for thin films of many different kinds, transparent as well as reflecting. Another alternative is ethylene tetrafluoroethylene. Photochromism is known in a number of polymers, and detailed information is available for spirooxazine in a host of cellulose acetate butyrate. The coloration dynamics are similar to those for photochromic glass, but the bleaching is faster for the polymer and takes place in about 20 minutes. The photoeffects are limited to T_{lum}; they vanish at elevated temperatures.

3.3 Metals

Metal surfaces can be used to accomplish high R_{lum} and R_{sol} (for example, in mirrors of various types), or for providing low E_{therm}. Figure 4 shows spectral reflectance for five different metals. The highest values of R_{lum} and R_{sol} are found in silver (Ag) and aluminum (Al), and these metals are used in bulk form or as thin films supported by substrates in several solar-energy-related applications. Surface protection of such surfaces is normally needed to obtain long-term durability. More corrosion-resistant metals, such as platinum or rhodium, are less reflective. All of the metals can yield a low magnitude of E_{therm}.

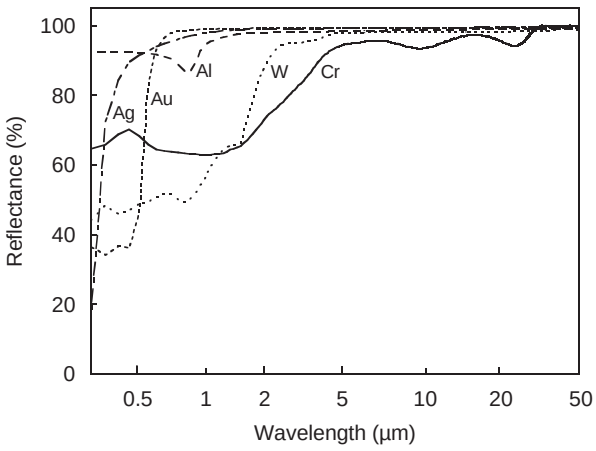


FIGURE 4 Spectral reflectance for some metal surfaces.

4. THIN FILMS

4.1 Thin Film Deposition

Surface treatments can be used to modify the optical properties of an underlying material and thus enhance its ability to adapt to the conditions of the ambient radiation (cf. Fig. 1). Many specific types of solar energy materials are used in this way. Thin film deposition, a particularly important application, represents a vast technological field, and only the most common methods are mentioned here.

Sputter deposition is widely used to make uniform coatings on glass, polyester, metal, etc. Typical film thicknesses are 0.1 to 1 μm. Essentially, a plasma of inert and/or reactive gases is set up under low pressure, and energetic ions in the plasma dislodge atoms from a solid plate of the raw material of the film (known as the "target"); these atoms are deposited as a uniform film on an adjacent surface (the substrate).

Evaporation can be an alternative to sputter deposition. Here, the raw material of the film is heated in vacuum so that a condensing vapor can transfer material to the substrate at a sufficient rate. Other technologies can be applied without recourse to vacuum: For example, sol-gel deposition involves dipping the substrate in a chemical solution, withdrawing it at a controlled rate, and subsequent annealing. Alternatively, the chemical solution can be applied by spray coating.

Chemical vapor deposition uses heat to decompose a vapor of a "precursor" chemical to produce a thin film of a desired composition. Electrochemical

techniques include cathodic electrodeposition from a chemical solution and anodic conversion of a metallic surface to form a porous oxide. Numerous alternative techniques exist.

4.2 Antireflection Treatment

Antireflection treatment (for example, to obtain high T_{lum} through glass) can be accomplished by the application of a layer with a refractive index that is close to the square root of the refractive index of the underlying glass (i.e., 1.23) and with a thickness corresponding to one-quarter of the wavelength of visible light (i.e., $\sim$100 nm). The condition imposed on the refractive index allows for good durability for only a few materials. Figure 5 shows an example of a glass slab of which two surfaces have been coated with a layer of porous silica obtained via sol-gel technology. It is seen that T_{lum} exceeds 99% in the middle of the luminous spectrum. Other antireflection treatments of practical interest include thin films of aluminum oxyfluoride or magnesium fluoride, and porous or microstructured surface layers produced by etching the glass surface in fluorosilicic acid or by application of a submicrometer-structured layer.

Simple antireflection treatments are able to lower the reflectance within a rather narrow wavelength range, but if a low magnitude of R_{sol} is required, a multilayer film with carefully adjusted refractive indices and thicknesses could be used. It should be noted that antireflection is efficient only for a specific range of incidence angles, especially in the case of multilayer films.

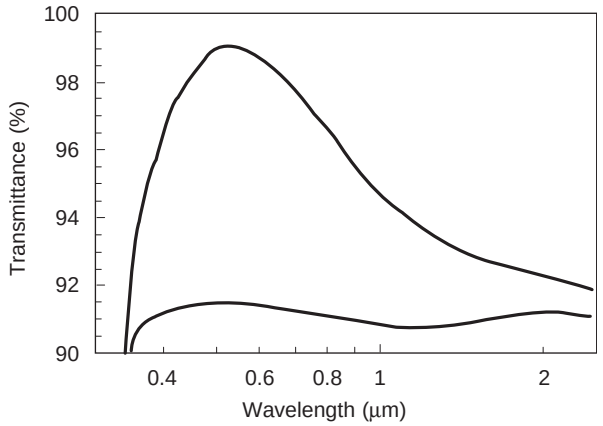


FIGURE 5 Spectral transmittance of a glass plate coated on both sides by antireflecting thin films of porous silica.

5. TRANSPARENT THERMAL INSULATORS

5.1 Principles

Heat transfer occurs via the additive effects of radiation, convection, and conduction. Cutting back this transfer is of concern for many different applications of solar energy. Radiation transfer can be diminished by having surfaces with low values of E_{therm}. Convection, which is of interest for gases, can be lowered by breaking up the gas into cells with dimensions of the order of 1 cm or less (or, possibly, by decreasing the gas pressure), and conduction heat transfer can be minimized by suitable constructions using low-conductance materials.

5.2 Solid Materials

Figure 6 shows four types of solid transparent insulation materials, including flexible polymer foils; polymer honeycomb materials; bubbles, foam, and fibers; and inorganic microporous materials such as silica aerogels. If the honeycomb cross-section is small compared to the cell length, a capillary structure exists. Foils and aerogels can be almost invisible to the eye, whereas honeycombs, bubbles, foam, and fibers cause reflection and scattering and hence a limited direct transmittance; thus, only the former group of materials is of interest when unperturbed vision is required–such as in many windows–and the other materials find applications in translucent wall claddings and solar collectors. The materials can be used in vertical, horizontal, or inclined positions, depending on the type of application.

Flexible polyester foils can be suspended to break up a thick gas slab into $\sim$1-cm-thick layers with

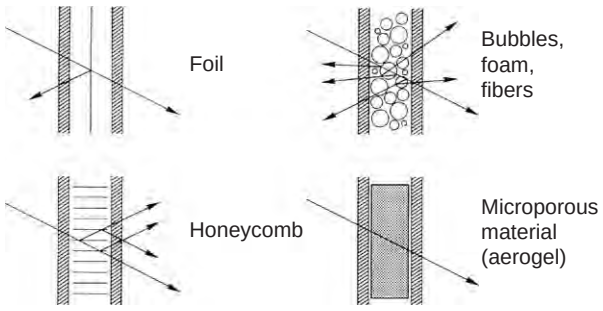


FIGURE 6 Principles underlying performance of four different types of solid transparent insulation materials placed between glass panes. Arrows signify light rays. Reflections at the glass surfaces are not shown.

diminished convection, and thin films can be applied to adjust the magnitudes of T_{lum} and T_{sol}, and to minimize E_{therm}. Similarly, the macroporous materials–typically made of polystyrol, polyamide, polyvinylchloride, or polycarbonate–produce gas-filled cells of a sufficiently small size to practically eliminate convection. A characteristic value of the heat transfer of the honeycombs, capillaries, bubbles, foam, and fibers is $1\,W/m^2\,K^{-1}$ for a material thickness of 10 cm.

Silica aerogel can be made by supercritical drying of colloidal silica gel under high temperature and high pressure. The ensuing material consists of $\sim$1-nm-diameter silica particles interconnected so that a loosely packed structure with pore sizes of 1 to 100 nm is formed. The relative density can be as low as a few percent of the bulk density. The material can be produced as translucent granules or as transparent tiles with a certain degree of haze. Figure 7 shows spectral transmittance through a 4-mm-thick tile. The low transmittance at short wavelengths is due to optical scattering. Heat transfer lower than $1\,W/m^2\,K^{-1}$ can be achieved in 1-cm-thick slabs of aerogel under favorable conditions (positioning between surfaces with low E_{therm} and at low gas pressure).

5.3 Gases

The heat transfer between two surfaces can be lowered by use of a gas other than air, as is common in modern fenestration technology involving hermetically sealed glazing units. The heat transfer across a typical 1.2-cm-wide gas slab is diminished on the $\sim$10% level if air is replaced by argon or CO_2, whereas larger effects are found with krypton or xenon. The relative role of the gas filling is enlarged (in practice, to the $\sim$20% level) if the radiation component to the heat transfer is decreased by the use of surfaces with a low E_{therm}. Infrared-absorbing gases can diminish the heat transfer across a given distance, but the limited availability of suitable gases makes this possibility less interesting. Among other options related to gases, some improvement of the thermal properties can be combined with sound insulation by the use of SF_6 gas. Finally, minimized heat transfer can be achieved by eliminating the gas, so that both the conduction and the convection components vanish. Obviously, the mechanical strength of the device must be sufficient to allow a vacuum. This necessitates inconspicuous spacers for "vacuum windows."

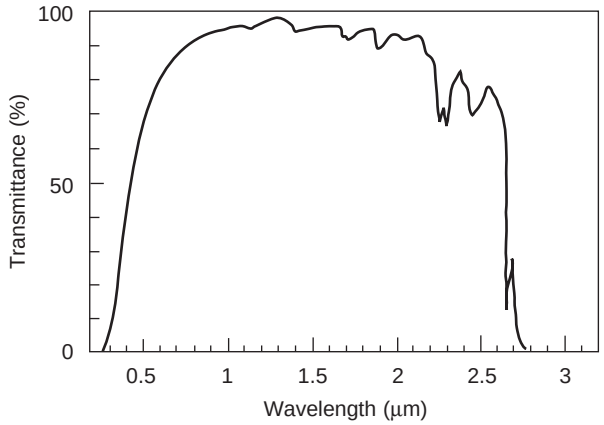


FIGURE 7 Spectral transmittance of a silica aerogel tile.

6. SOLAR THERMAL CONVERTERS

6.1 Principles

Photothermal conversion of solar energy, to produce hot fluid or hot air, takes place in solar collectors. Efficient conversion requires that a solar-absorbing surface is in thermal contact with the fluid or gas, and that thermal losses to the environment are minimized. The absorber should be positioned under a transparent cover, and a transparent thermal insulation material may be used as well. The back and sides of the solar collector should be well insulated using conventional techniques. The most critical part for the photothermal conversion is the actual solar absorber surface, which obviously should have maximum A_{sol}. Its radiation heat losses should be minimized, and this requires low E_{therm}. The absorber is nontransparent, and the conditions can be written in terms of an ideal spectral reflectance as

$$R(\lambda) = 0 \quad \text{for} \quad 0.3 < \lambda < \lambda_c\ \mu m \tag{3}$$

$$R(\lambda) = 1 \text{ for } \lambda_c < \lambda < 50\ \mu m, \tag{4}$$

where λ_c denotes a "critical" wavelength lying between 2 and 3 μm, depending on the temperature (cf. Fig. 1). Clearly, this is one example of how spectral selectivity is used to obtain optimized properties of a solar energy material.

A number of different design principles and physical mechanisms can be used in order to create a spectrally selective solar-absorbing surface. Six of these are shown schematically in Fig. 8. The most straightforward approach would be to use a material with suitable intrinsic optical properties. There are

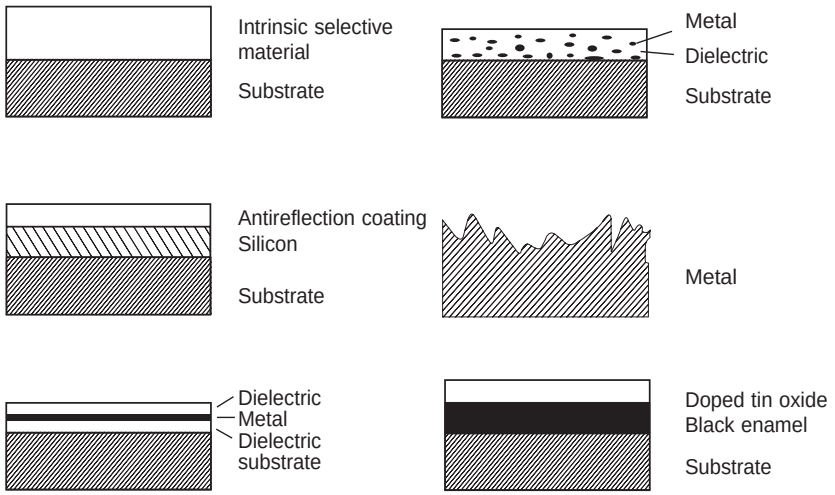


FIGURE 8 Cross-sectional schematic designs of six types of thin films and surface treatments for selective absorption of solar energy.

few such materials, though, and this approach has not been particularly fruitful. Semiconductor-metal tandems can give the desired spectral selectivity by absorbing short-wavelength radiation in a semiconductor with a bandgap corresponding to a wavelength of $\sim 2\,\mu m$ and having low E_{therm} as a result of the underlying metal. The pertinent semiconductors all have large refractive indices, which tend to give strong reflection losses and lower A_{sol}, implying that complex antireflection treatments must be invoked. Multilayer-metal tandems can be tailored so that the thin film stack becomes an effective selective absorber of solar energy. Metal/dielectric composite-metal tandems contain metal nanoparticles in a dielectric (often oxide) host. The composite is often referred to as a cermet. This design principle offers a large degree of flexibility, and the optimization of the spectral selectivity can be made with regard to the choice of constituent materials, film thickness, particle concentration, and grading, as well as the shape and orientation of the particles. The particles are much smaller than any relevant wavelengths of solar (or thermal) energy, and the composite behaves as an "effective medium" with properties intermediate between those of the metal and the dielectric. Textured metal surfaces can produce a high A_{sol} by multiple reflections against metal dendrites that are $\sim 2\,\mu m$ apart, but the low E_{therm} is rather unaffected by this treatment because the relevant wavelengths are much larger that the dendrite separation. Finally, selectively solar-transmitting films on blackbody-like absorbers can be used to impart spectral selectivity. The relevant thin films are discussed in the following section.

6.2 Spectrally Selective Thin Films

A number of thin films, coated onto metallic substrates by one of the thin film technologies discussed in the preceding section, have been developed in the past, and some of these have reached commercialization. These films normally exploit several of the design principles and physical mechanisms mentioned previously. Electrochemical techniques have been used for several decades to make spectrally selective surfaces (suitable for flat-plate solar collectors) based on electrodeposited films containing chrome or nickel (referred to as black chrome and black nickel, respectively) of complex compositions. Another technique employs anodic conversion of aluminum to make a porous surface layer, followed by electrolysis to precipitate out nickel particles inside the pores. The solid curve in Fig. 9 illustrates a spectral reflectance curve. $A_{sol} \approx 96\%$ and $E_{therm} \approx 15\%$ at 100°C are typical performance data. The durability of these films has been modeled in detail.

The electrochemical techniques involve the handling of large quantities of environmentally hazardous chemicals, and this has led to a current trend toward thin film techniques based on deposition in vacuum. The dotted curve in Fig. 9 shows a spectrum for an optimized sputter-deposited film based on nickel (Ni); the design includes an antireflection layer that boosts A_{sol}. The data correspond to $A_{sol} \approx 95\%$ and $E_{therm} \approx 10\%$ at 100°C. Other sputter-deposited and evaporated films of practical use include a number of cermets with stainless-steel and metal nitride, and compositionally graded films containing chrome oxide and chrome nitride. These films can be used

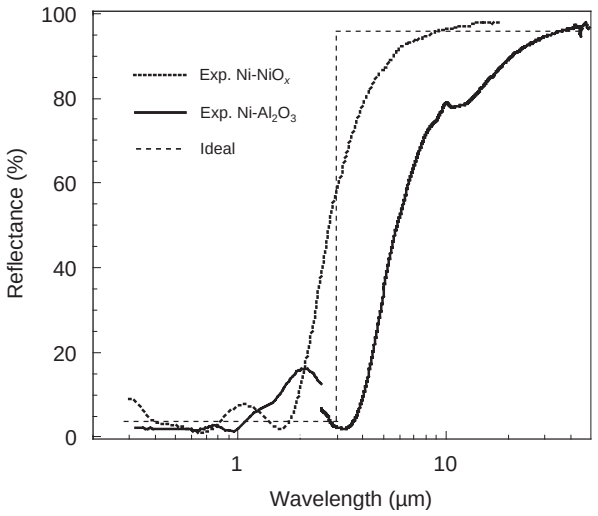


FIGURE 9 Experimental spectral reflectance results for two types of selectively solar-absorbing surfaces, compared to the ideal performance.

for different types of solar collectors, including concentrating ones. Advanced selectively solar-absorbing surfaces have been developed for solar thermal power stations with parabolic trough reflectors and tubular solar collectors. The absorber surfaces used in this case employ sputter deposition to manufacture a complex multilayer film stack incorporating graded molybdenum (Mo)/oxide films, antireflection treatment, and an emittance-suppressing Mo film on the stainless-steel tube serving as a substrate. These absorbers have $A_{sol} \approx 97\%$ and $E_{therm} \approx 17\%$ at 350°C.

6.3 Spectrally Selective Paints

Paints have an obvious advantage in that they can be applied by commonly available techniques. Typically, a strongly absorbing pigment (for example, an oxide of FeMnCu) is mixed in an infrared-transparent polymeric binder such as a silicone or a siloxane. The paint is applied in a layer 2 to 3 μm thick onto a metallic substrate with low E_{therm}. Characteristic data are $A_{sol} \approx 90\%$ and $E_{therm} \approx 30\%$; these data are not as good as those for the thin films. Possibilities to prepare thickness-insensitive spectrally selective paints by including metallic flakes in the binder may lead to interesting products in the future. Visibly colored solar-absorbing surfaces, which may be desirable for aesthetic reasons, can be obtained by relaxing the requirement on A_{sol} somewhat.

7. RADIATION COOLING

As was shown in Fig. 1, the atmosphere can be transparent to thermal radiation in the $8 < \lambda < 13$-μm "window" range. This effect can lead to radiation cooling of surfaces oriented toward the sky. The resource for this cooling has been evaluated through detailed calculations showing that the cooling power for a blackbody surface is 80 to 120 W/m^2, depending on the atmospheric conditions, when the surface is at ambient temperature. If nothing but radiation exchange is considered, the surface can drop to 14–26°C below the air temperature. In the presence of a nonradiative heat transfer equal to 1 W/m^2 K^{-1}, the maximum temperature drop is expected to be 10–20°C. Practical tests, using a multistage radiation cooler placed under an exceptionally dry atmosphere, have demonstrated that a temperature decrease as large as ~33°C can be achieved in practice by purely passive means.

The temperature drop can be enhanced if radiation exchange takes place only within the atmospheric "window." This calls for surfaces with infrared selective emittance, with an ideal reflectance of

$$R(\lambda) = 1 \quad \text{for} \quad 3 < \lambda < 8\ \mu\text{m} \quad \text{and} \quad 13 < \lambda < 50\ \mu\text{m} \tag{5}$$

$$R(\lambda) = 0 \quad \text{for} \quad 8 < \lambda < 13\ \mu\text{m}. \tag{6}$$

If cooling should take place also in the day, it is desired to have $R(\lambda) = 1$ for $0.3 < \lambda < 3$ μm. Calculations have shown that infrared-selective surfaces can have a cooling power between 60 and 95 W/m^2, depending on the atmosphere. Maximum temperature drops of 26–60°C with nothing but radiation exchange, and of 18–34°C when the nonradiative heat influx is 1 W/m^2 K^{-1}, have been predicted. The significance of spectral selectivity for reaching low temperatures is hence evident. Spectral selectivity approximately according to Eqs. (5) and (6) has been achieved with ~1-μm-thick silicon oxynitride films backed by aluminum and by some aluminized polymer foils, including Tedlar. Another possibility is to use confined gas slabs of NH_3, C_2H_4, C_2H_4O, or mixtures of these, with thicknesses of a few centimeters. In both cases, the low reflectance in the atmospheric "window" range is associated with molecular vibrations, and the high reflectance outside this range is caused by the aluminum.

Practical tests of the cooling ability of these materials have been carried out with devices in which the cooling surface, or the gas, is located under an infrared-transparent convection shield. High-density polyethylene foil has suitable optical

properties but requires stiffening in order not to cause convection heating due to mechanical movements caused by wind. Using a radiation-cooled gas, the coolant can be circulated and heat exchanged so that it provides a continuous source for cooling under the clear sky. Radiation cooling can be used for condensation of water from the atmosphere. Preliminary measurements have demonstrated that significant amounts can be collected even under unfavorable conditions (during drought months in a semidesert part of Tanzania). The material used in this case was a 0.4-mm-thick polymer foil pigmented by white TiO_2 and $BaSO_4$.

8. SOLAR CELLS

Solar cells for generating electricity depend on certain materials. In principle, all that is needed to generate electricity is a carrier in an excited state or states; a carrier can be excited from a ground state to an excited state by photon absorption, and to generate electricity requires some means of extracting or supplying carriers to or from these states. There are many options but, despite much research and development over decades, the attention has been focused almost entirely on a few materials.

Silicon (Si) is an excellent material for solar cells and is by far the most widely used one. It is nontoxic, very abundant, and has a well-established technological base due to its ubiquitous uses in microelectronics. Essentially, a slab or film of Si is n-doped and p-doped in the two film surface regions, and this structure is positioned between a transparent front electrode arrangement and an opaque metallic back electrode. (Elements that serve as dopants are either p-type acceptors or n-type donors; adjoining n and p regions form a junction that passes current from p to n.) The Si can be of three types: crystalline (cut from a single crystal ingot), polycrystalline (made from a multicrystalline ingot or prepared by "ribbon growth"), or amorphous and hydrogenated (typically a $\sim$0.3-μm-thick film made by glow discharge deposition of a silane gas). High photoelectric conversion efficiency demands low reflection losses, which can be achieved by an antireflecting film or–in the case of a single crystalline material–by using anisotropic etching to make a pyramid-type texturing.

Compound semiconductors of the III–V type (gallium, arsenic, aluminum, indium), based on GaAs, (Al,Ga)As, InP, or (In,Ga)P, can show high efficiency as well as good resistance to ionizing radiation; they are costly, though, and mainly used for space applications. Among the II–IV compounds (cadmium, tellurium), CdTe and Cd(S,Te) are known for their robust manufacturability and are of interest for ground-based solar cells, although environmental concerns regarding large-scale use of a technology including cadmium should not be overlooked. Among the I–III–IV_2 compounds (copper, indium, selenium), $CuInSe_2$, $Cu(In,Ga)Se_2$, and $Cu(In,Ga)\ (S,Se)_2$ are notable for their possibility to reach high photothermal conversion efficiencies in potentially low-cost thin film solar cells. Numerous other materials–including polymers–and approaches to solar cells are possible, and many other compound semiconductors may well turn out to have favorable properties. Some of those may find applications in thermophotoelectricity, i.e., cells operating with thermal irradiation. Others may be used in tandem cells, encompassing superimposed cells with optical bandgaps that are large on the exposed surface and gradually shrink toward the underlying substrate.

Alternatively, photoelectrochemical techniques can be employed to generate electricity, and studies of nanocrystalline dye cells are presently of interest. These cells absorb light in the dye molecules–normally containing a ruthenium-based compound–coated onto nanocrystals of TiO_2 made by colloidal technology or other means. Electrons are excited in the dye and can flow into the conduction band of the adjacent n-type TiO_2. The electron is transferred to a transparent front surface electrode, through the load, and to a counterelectrode having a thin layer of platinum. Here it reduces triiodine to iodide, which then diffuses to the dye molecules and reduces them back to their original state. Sealing and durability are issues that need to be resolved before widespread applicability is possible. Also, the fact that the fundamental principles of functionality remain obscure may hamper rapid progress.

Generally speaking, solar cells lose efficiency when they become warm. It is therefore of interest to devise cells and cell systems that minimize heating. One possibility is to use external surfaces that reflect the subbandgap wavelengths useful for photothermal conversion onto the solar cell while absorption prevails at longer wavelengths. Mirrors of SnO_2-coated aluminum have been developed for this purpose. One aspect of today's solar cell technology that requires particular attention, because it proliferates, is the global availability of the raw materials. Gallium, selenium, and ruthenium are all rare components of Earth's crust. Indium is also widely

believed to be rare, but recent assessments indicate that availability is not a serious issue.

9. COATINGS FOR GLAZINGS: STATIC PROPERTIES

9.1 Principles

Architectural windows and glass facades are problematic from an energy perspective. On the one hand, their primary function, to provide visual contact between indoors and outdoors, as well as day lighting, should not be compromised. On the other hand, there are often unwanted energy flows, with too much thermal energy leaving or entering the building via the window and a concomitant demand for space heating and cooling; furthermore excessive solar energy may be admitted, which puts additional demand on air conditioning. Present architectural trends tend to increase the window areas, so that the energy issue becomes even more pressing.

The radiation part of the heat transfer can be controlled by thin films with a low E_{therm}, which, in a multiply glazed window, should face one of the confined air (or gas) spaces. In this way, the heat transfer across a vertically mounted window can drop from ~ 3 to $\sim 1.5\,\text{W/m}^2\,\text{K}^{-1}$ for double glazing and from ~ 1.8 to $\sim 1.0\,\text{W/m}^2\,\text{K}^{-1}$ for triple glazing. Obviously, T_{lum} must be large for these films. The infrared part of the solar radiation, which transmits energy through the glazing but is not needed for vision, can be stopped by a thin film having low transmittance at $0.7 < \lambda < 3\,\mu\text{m}$. These demands have led to the development of two types of thin films, i.e., "low-emittance coatings" (low-E coatings), characterized by

$$T(\lambda) = 1 \quad \text{for} \quad 0.4 < \lambda < 0.7 \quad \mu\text{m} \tag{7}$$

$$R(\lambda) = 1 \quad \text{for} \quad 3 < \lambda < 50 \quad \mu\text{m} \tag{8}$$

and "solar control coatings," characterized by

$$T(\lambda) = 1 \quad \text{for} \quad 0.4 < \lambda < 0.7 \quad \mu\text{m} \tag{9}$$

$$R(\lambda) = 1 \quad \text{for} \quad 0.7 < \lambda < 50 \quad \mu\text{m}. \tag{10}$$

The spectral selectivity inherent in these relationships is of course idealized.

Another way to achieve energy efficiency is to use angular selective thin films on the glass. In this way, it is possible to take advantage of the fact that the indoor–outdoor contact across a window normally occurs along near-horizontal lines-of-sight, whereas solar irradiation comes from a small element of solid angle high up in the sky during most of the day. It is also possible to combine spectral and angular selectivity.

9.2 Spectrally Selective Thin Films

Very thin metal films can have properties resembling those defined in Eqs. (7)–(10). Calculations for extremely thin bulklike films of silver, gold, and copper predict that large values of T_{lum} and T_{sol} can be combined with low values of E_{therm}. However, such films cannot be made on glass or polyester. Rather, the thin film goes through a number of distinct growth phases, and large-scale coalescence into a metallic state with low E_{therm} requires thicknesses exceeding 10 nm, and then T_{lum} and T_{sol} are limited to about 50%. Most of the transmittance loss is due to reflectance at the film interfaces, implying that the transmittance can be boosted by embedding the metal film between dielectric layers, serving, essentially, for antireflection purposes.

Silver (Ag) has the best optical properties and is used in most applications. Gold (Au) and copper (Cu) tend to produce colored films that are undesirable for many architectural applications. Titanium nitride (TiN) can serve as a replacement for Au. The dielectric layer can be of several different kinds, such as Bi_2O_3, SnO_2, TiO_2, ZnO, and ZnS. The multilayer structure may also include corrosion-impeding layers, such as Al_2O_3. Figure 10 illustrates $T(\lambda)$ for three types of thin films based on TiO_2/Ag/TiO_2 and

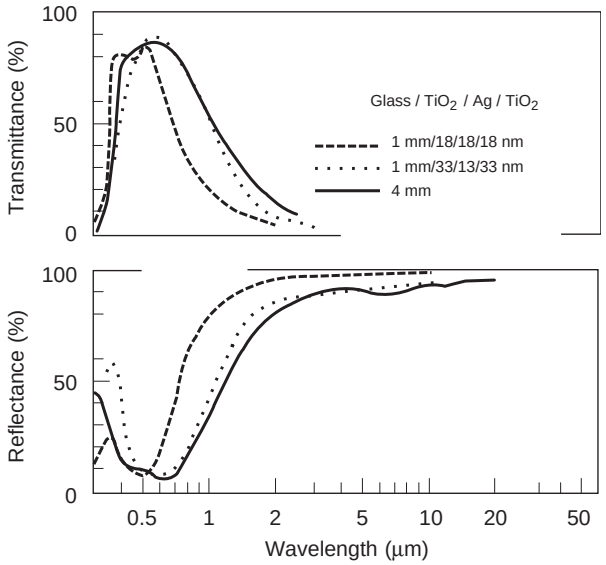


FIGURE 10 Spectral transmittance and reflectance results for glass having thin films based on TiO_2/Ag/TiO_2. Thicknesses are given for two of the film stacks.

deposited onto glass. Depending on the film thicknesses, it is possible to optimize T_{lum} or T_{sol}, whereas E_{therm} is invariably low. Thus, low-E coatings as well as solar control coatings can be made from metal-based thin films. Much of the glass that is used in modern windows and glass facades employs sputter-deposited films based on Ag. The films can be used only under protected conditions, such as inside sealed, double-glazed windows. The film thicknesses are small enough that iridescence, i.e., a rainbow-like exhibition of colors, is not a problem.

Doped oxide semiconductor coatings offer an alternative to the spectrally selective metal-based films just discussed. The practically useful materials include In_2O_3:Sn, SnO_2:F, SnO_2:Sb, and ZnO:Al. Many alternatives can be found among the ternary and quaternary oxides as well as among mixtures of the mentioned oxides. Nevertheless, the binary oxides have remained popular. The indium-based oxide has the lowest E_{therm}, the selenium-based oxides are most durable, and the zinc-based oxide is most popular for some applications (such as in many solar cells). The films are electrically conducting with a minimum resistivity of $\sim 10^{-4}\,\Omega\,\text{cm}$ for optimally prepared In_2O_3:Sn. Achieving such properties is notoriously difficult, though, and typically requires thermal treatment at $\sim 300^{\circ}\text{C}$ or above.

Figure 11 shows $T(\lambda)$ and $R(\lambda)$ in the full $0.3 < \lambda < 50$-μm range for a 0.2-μm-thick film of In_2O_3:Sn. The parameter denoted n_e is the density of free electrons, which is directly associated with the amount of tin (Sn) doping; in practice, this amount is of the order of a few percent. The data in Fig. 11 were computed from a fully quantitative theoretical model of heavily doped semiconductors and give a clear representation of the salient features developing on increased doping: an onset of high reflectance in the near-infrared (i.e., a lowering of E_{therm}), bandgap widening tending to enhance T_{sol}, and disappearance of absorption features in the thermal infrared (approximately at $20 < \lambda < 30\,\mu\text{m}$). The doped semiconductor oxides can be very hard and resistant to corrosion, and they can be applied on glass surfaces exposed to ambient conditions. The practically useful thicknesses are larger than $\sim 0.2\,\mu\text{m}$, implying that thickness variation, which may be difficult to eliminate in practical manufacturing, may lead to aesthetically unpleasing iridescence.

The refractive index of the doped oxide semiconductors is about 2 for luminous radiation, which tends to limit T_{lum} to $\sim 80\%$. Higher transmittance values can be obtained by antireflection treatment,

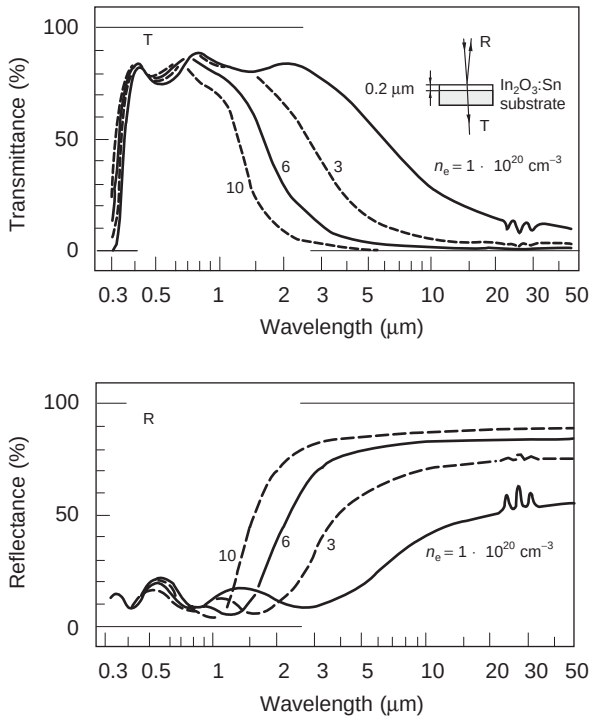


FIGURE 11 Spectral transmittance (T) and reflectance (R) computed from a fully quantitative model for the optical properties of In_2O_3:Sn, using the indicated values of film thickness and free-electron density (n_e).

and T_{lum} can then even exceed the value for uncoated glass. The antireflection layer lowers E_{therm}, but on an insignificant level of a few percent only.

9.3 Angular Selective Thin Films

Pronounced angular properties can be used to invoke energy efficiency for glazings, as mentioned previously. Figure 12 illustrates this feature as computed for a five-layer stack of thin films, specifically with two 12-nm-thick silver films embedded between three SiO_2 films with thicknesses of 120 or 170 nm. Strong angular effects prevail at the larger of these thicknesses, and T_{sol} is 23% for normally incident light and as large as 58% at an incidence angle of 60°. The data in Fig. 12 are symmetrical around the surface normal of the film. However, angular properties that are distinctly different at equal angles on the two sides of the normal can be obtained by the use of films with inclined columnar microstructures. Figure 13 shows a polar plot of T_{lum} recorded for a chromium-based film prepared by evaporation under conditions so that the incident flux arrives at the substrate at an almost glancing angle. The columns needed for the angular selectivity then

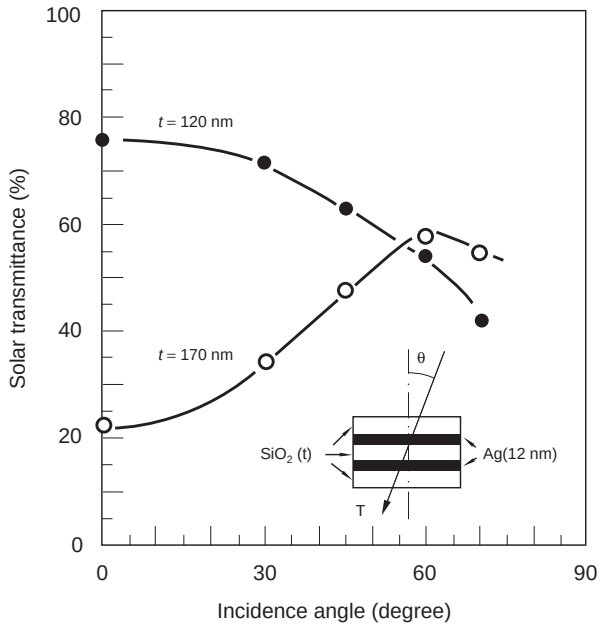


FIGURE 12 Solar transmittance for thin films of SiO_2/Ag/SiO_2/Ag/SiO_2 as a function of SiO_2 thickness (t) and incidence angle (θ).

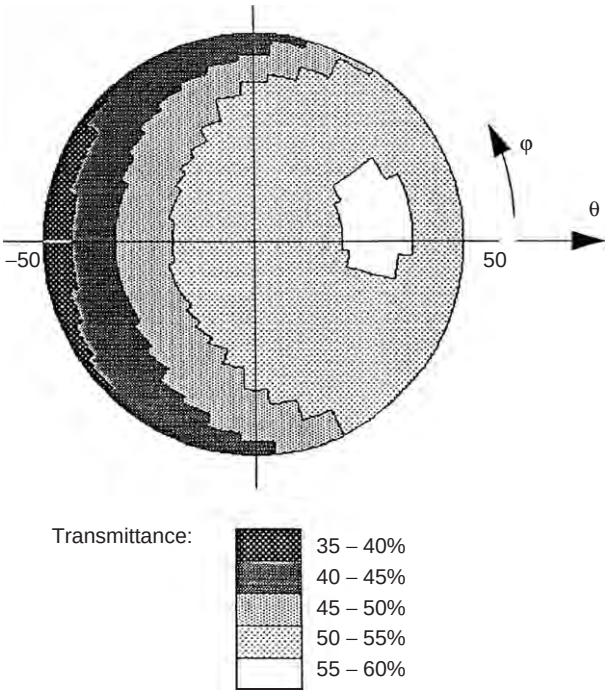


FIGURE 13 Polar plot of luminous transmittance through an obliquely evaporated Cr-based thin film. Data are given as a function of incidence angle (θ) and azimuthal angle (φ).

tend to form an array inclined in the direction toward the evaporated flux. Sputter deposition can also be used for making angular selective films.

10. COATINGS FOR GLAZINGS: DYNAMIC PROPERTIES

Windows and glass facades with variable properties have been the dream of architects for years. Such products are now becoming possible by exploiting "chromogenic" materials. Electrochromic devices are now being used in windows, albeit as yet on a small scale. A range of other options for dynamic fenestration also exists.

10.1 Electrochromic Devices

Electrochromic devices are able to vary their throughput of visible light and solar radiation on electrical charging and discharging, using a low voltage. Thus, they can regulate the amount of energy entering through a "smart window" so that the need for air conditioning in a cooled building becomes lessened. The energy efficiency inherent in this technology can be large, provided that the control strategy is adequate. Additionally, the transmittance regulation can impart glare control as well as, more generally, user control of the indoor environment. The absorptance, rather than the reflectance, is modulated so that the electrochromic devices tend to heat up in their low-transparent state, which has to be considered in a practical window construction. Modulation of E_{therm} is possible in principle, and development of variable-emittance surfaces for temperature control of space vehicles, as well as for other applications, is in progress.

Figure 14 shows an electrochromic five-layer prototype device that introduces basic design concepts and types of materials. A kinship, in principle, to an electrical battery is noteworthy. The central part of the construction is a purely ionic conductor (i.e., an electrolyte), either a thin film or a polymer laminate material; it must be a good conductor for small ions such as H^+ or Li^+. The electrolyte is in contact with an electrochromic layer and a counter-electrode. For a transparent device, the latter must remain nonabsorbing, irrespective of its ionic content, or it should exhibit electrochromism in a sense opposite to that of the base electrochromic film. This three-layer configuration is positioned between transparent electrical conductors, normally being doped oxide semiconductors of the kinds previously discussed; they are backed by glass or polyester foil. By applying a voltage of a few volts–conveniently obtained from solar cells–between the outer layers, ions can be shuttled into or out of the electrochromic

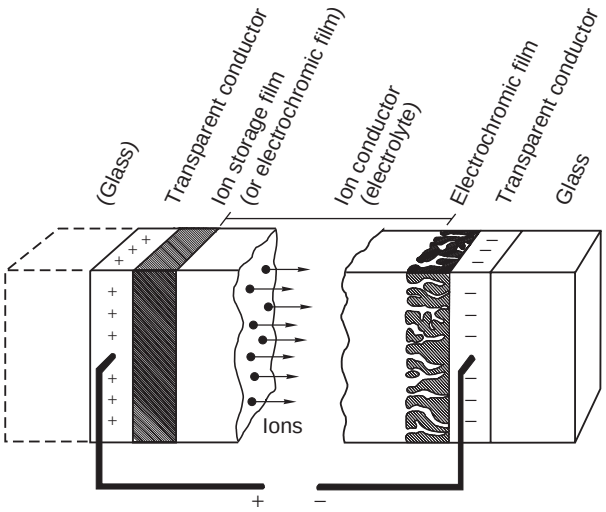


FIGURE 14 Basic design of an electrochromic device, indicating transport of positive ions under the action of an electric field.

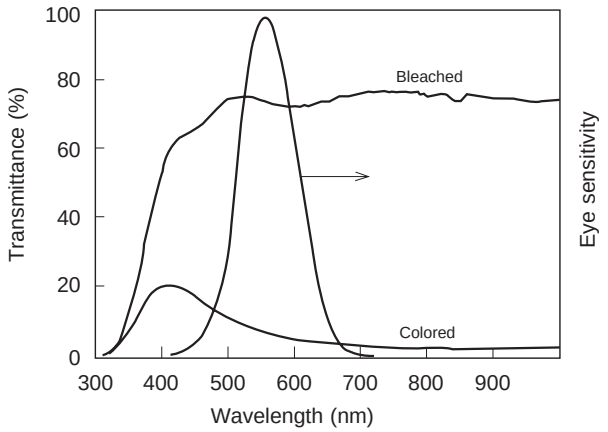


FIGURE 15 Spectral transmittance through an electrochromic foil including thin films of tungsten-based and nickel-based oxides. Data are shown for fully bleached and colored states.

film(s), the optical properties of which are thereby changed so that the overall optical performance is changed. The voltage needs to be applied only when the transmittance is altered, i.e., the electrochromic devices exhibit open-circuit memory. The time for going between bleached and colored states is largely governed by the electrical resistance of the transparent conducting films. Smart windows of the scale of a square meter require about ten minutes to change from a bleached state to a moderately colored state. Smaller devices can display much faster dynamics.

There are many materials and design issues for practical electrochromic smart windows. Electrochromism is found in numerous organic and inorganic materials, the latter mainly being transition metal oxides. Films of WO_3 are used as the base electrochromic layer in almost all devices studied until now. However, many options exist for the counterelectrode, with good results having been documented for oxides based on Ce, Co, Ir, Ni, and V, whereas other designs have used $KFe^{3+}Fe^{2+}(CN)_6$ (known as "prussian blue"). Additions of Al and Mg to oxides of Ir or Ni can enhance the bleached-state transmittance of the devices, which is important for architectural applications. In the case of Ir oxide, the additions lead to significant cost reduction. Many alternatives exist for the central electrolyte, which can be an adhesive polymer with high ionic conductivity or–in an all-solid-state approach with superimposed thin films–a hydrous layer of Ta_2O_5, ZrO_2, or SiO_2 serving as a conductor for H^+.

Figure 15 illustrates the optical performance of a laminated polyester-based electrochromic foil device comprising films based on WO_3 and NiO in the bleached state and after maximum coloration. It is seen that T_{lum} can be modulated between 74 and 7%. There is a trade-off between durability and maximum coloration, though, and a minimum T_{lum} of $\sim$30% may be more adequate for practical applications.

10.2 Some Alternative Chromogenic Technologies

"Gasochromic" thin films represent an alternative to the electrochromic devices, with specific pros and cons. Glass is coated with a WO_3-based thin film having an extremely thin surface layer of catalytically active Pt. The coated side of the glass is in contact with the gas confined in a carefully sealed double-glazed window. By changing the amount of H_2 in contact with the thin film, it is possible to insert a variable amount of H^+ and the transmittance is then modulated in analogy with the case of the electrochromic device. Among the many other possibilities to control T_{lum} and T_{sol} electrically, of particular note are suspended-particle devices and phase-dispersed liquid crystal devices. These have much swifter dynamics compared to electrochromic devices, but require higher voltages. Rapid technological progress, as well as the proprietary nature of much basic information, make a detailed comparison of the various techniques difficult.

Thermochromism represents another possibility to control T_{sol} in order to create energy efficiency in buildings. The most interesting material, a thin film based on VO_2, undergoes a structural transition at a certain "critical" temperature τ_c, below which the material is semiconducting and relatively nonabsorbing in the infrared. Above τ_c, it is metallic and

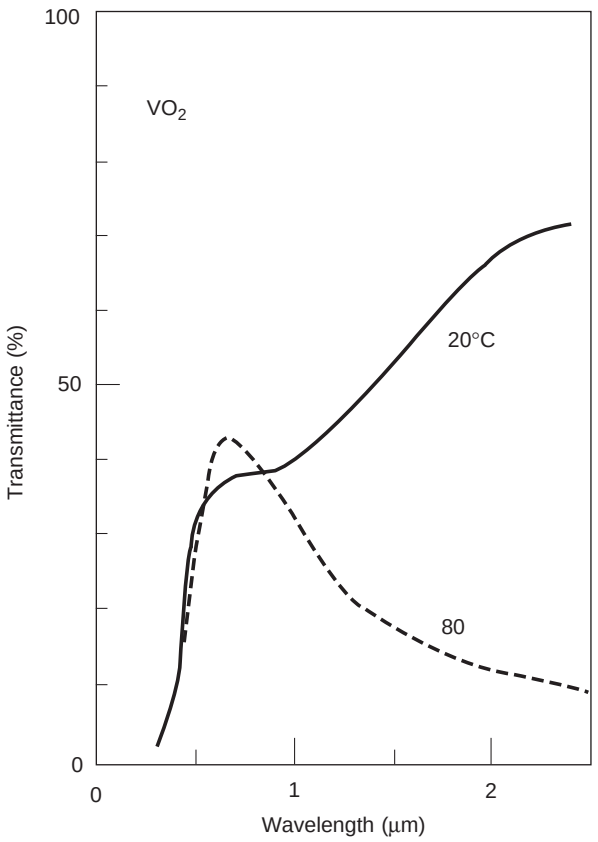


FIGURE 16 Spectral transmittance through a thermochromic VO_2 film at two temperatures.

infrared reflecting. Figure 16 shows typical data on $T(\lambda)$ above and below τ_c. Bulklike VO_2 has $\tau_c \approx 68^\circ C$, which obviously is too high for architectural applications in buildings, but τ_c can be decreased to room temperature by additions of Mo, Nb, or W, for example.

Still another possibility to vary T_{lum} as a result of temperature changes is found among the "cloud gels," which can be interposed between the panes in a double-glazed window. These are polymers for which "clouding"–i.e., transition to a diffusely scattering state–can set in above a certain τ_c due to a reversible thermochemical dissolution and a thermally induced modification in the lengths of the polymer molecules.

11. SOLAR PHOTOCATALYSIS

Under the action of solar irradiation, photocatalytically active surfaces can cause the breakdown of organic molecules as well as microorganisms adsorbed on these surfaces. Thus hazardous or otherwise unpleasant pollutants in air or water can be decomposed into tasteless, odorless, or at least less toxic compounds. Well-known photocatalysts are oxides of Cd, Ce, Fe, Sb, Sn, Ti, W, Zn, and Zr, with most developmental work having been devoted to TiO_2. All of these oxides are able to use ultraviolet light only. This is an advantage for some applications, and, for example, a double-glazed window may be transformed into a photocatalytic reactor with an outer ultraviolet-transparent pane (cf. Fig. 2) and an inner pane with a TiO_2 film. In other cases, it is also an advantage to employ visible light for inducing photocatalysis; transition metal additions to the oxides previously mentioned , as well as nitrogen incorporation, are able to widen the spectral range for photocatalysis.

Photocatalytic decomposition of water into hydrogen and oxygen has been investigated for many years, but it has turned out to be difficult to accomplish effectively using solar irradiation. Present research includes studies of noble-metal-loaded oxide catalysts of many kinds, sometimes covered with NaOH or Na_2CO_3. Ultrasonic irradiation of the catalyst can improve the efficiency of water splitting.

SEE ALSO THE FOLLOWING ARTICLES

Lunar–Solar Power System • Solar Cells • Solar Cookers • Solar Energy, History of • Solar Fuels and Materials • Solar Heat Pumps • Solar Ponds • Solar Thermal Energy, Industrial Heat Applications • Sun, Energy from

Further Reading

Gordon, J. (ed.). (2001). "Solar Energy: The State of the Art." James & James Science Publ., London.

Granqvist, C. G. (ed.). (1993). "Materials Science for Solar Energy Conversion Systems." Pergamon Press, Oxford, U.K.

Granqvist, C. G. (1995). "Handbook of Inorganic Electrochromic Materials." Elsevier Science, Amsterdam.

Kaneko, M., and Okura, I. (eds.). (2002). "Photocatalysis: Science and Technology." Springer-Verlag, Berlin and New York.

Lampert, C. M., Granqvist, C. G. (eds.). (1990). "Large-Area Chromogenics: Materials and Devices for Transmittance Control." SPIE—The International Society for Optical Engineering, Bellingham, Washington.

Mbise, G. W., Le Bellac, D., Niklasson, G. A., and Granqvist, C. G. (1997). Angular selective window coatings: Theory and experiment. *J. Phys. D: Appl. Phys.* **30,** 2103–2122.

Material Use in Automobiles

SUJIT DAS
Oak Ridge National Laboratory
Oak Ridge, Tennessee, United States

1. **Material Usage Trends in Automobiles**
2. **Factors Impeding Lightweight Materials Growth**
3. **Role of Lightweight Materials in Energy Consumption**
4. **Automotive Lightweight Materials Outlook**
5. **SAFETY Considerations in Automobile Lightweighting**

Glossary

automobile An average light-duty vehicle representing passenger cars, sport-utility vehicles, minivans, and light trucks.
composites A material composed of a reinforcing material (e.g., glass or carbon fibers) in a matrix material (e.g., polymers).
lightweighting Using lighter materials or efficient system design to reduce vehicle weight.
monocoque A type of motor vehicle design in which the body and frame are integrated into eggshell shape, resulting in a small number of relatively large parts.
recovery rate The percentage of usable recycled materials recovered from discarded vehicles.
recycling rate The percentage of vehicles from which materials are collected, separated, or processed and returned to the economic mainstream to be reused in the form of raw materials or finished goods.
thermoplastic A polymer in which the molecules are held together by weak secondary bonding forces that can be softened and melted by heat, then shaped or formed before being allowed to freeze again.
thermoset A polymer that solidifies irreversibly when heated due to a chemical reaction involving cross-linking between chains.
unibody A conventional body on frame motor vehicle body structure design composed of a relatively large number of hollow sheet metal components onto which body panels are welded by welding robots or in multispot welding units.

Material composition continues to play an important role in automobile engineering, particularly with the renewed importance of fuel economy standards and advanced vehicle designs such as hybrid and fuel cell vehicles. Of the numerous types of materials currently in use in automobiles, lightweight materials, particularly aluminum, polymer composites, and magnesium, have received increasing attention because of their potential to reduce the weight of vehicles. Most of these lightweight materials are more energy intensive to produce than conventional steel, the predominant material in current vehicles. A life cycle energy comparison that addresses the vehicle-use phase, which contributes the major share of a vehicle's life cycle energy use, is appropriate to examine the materials' energy implications. The effect of vehicle weight on its safety is an overriding sensitive issue to consumers, and it remains a hotly debated issue in vehicle lightweighting.

1. MATERIAL USAGE TRENDS IN AUTOMOBILES

Automobiles play an important role in the world economy with a share of about 50% of total transportation energy demand. The transportation sector accounted for about 27% of total U.S. energy consumption in 2000. Growth in U.S. energy consumption of the past decade averaged about 2% annually, with automobiles accounting for about 57% of total transportation energy in 2000. Between 1978 and 2001, the average U.S. automobile weight has decreased only by 7.3% (3570 lbs versus 3309 lbs), whereas sales-weighted fuel economy improvements in automobiles have been more than 40% (19.9 miles per gallon versus 28.6 miles per gallon). As the share of heavier SUVs, minivans, and light trucks continued to grow during the past decade, the percentage increase in average vehicle weight in the fleet is significantly higher. A combination of technology improvements and weight reduction brought

 © 2004 Elsevier Inc. All rights reserved.

about this fuel-economy gain. The technology improvements range from reduced internal engine friction to reduced aerodynamic drag. Major improvements were provided by the transition from rear- to front-wheel drive, which allowed interior volume to be maintained while weight and exterior size were reduced, and by conversion from carburetors to electronically controlled fuel injection. There were many other improvements as well: more efficient automatic transmissions, improved lubricants, lower rolling resistance tires, reduced accessory loads, material substitution, and so forth.

The technology choices in U.S. automobiles have been a concern since the environmental movement of the 1960s, the oil crisis of the 1970s, the Clean Air Act of 1970 mandating standards for emissions, and the Energy Policy and Conservation Act of 1975 mandating national standards for automotive fuel economy (popularly known as corporate average fuel economy). These events and policies prompted automakers to address fuel economy, formerly a minor factor in automotive design, as a vital engineering requirement. The 1994 Partnership for a New Generation of Vehicles (PNGV) program (lately superseded by Freedom Cooperative Automotive Research (FreedomCAR), which is unlike independent of any specific vehicle platform with a focus on fuel cell vehicles without any specific fuel economy goal), has a formidable goal of developing production-ready mid-sized vehicles that are 40% lighter than baseline vehicles and achieve 80 miles per gallon fuel economy. This program considered many new research areas, including advanced powertrain concepts with as much as 40 to 45% powertrain thermal efficiency, body and chassis weight reduction technologies, alternate design concepts, and more efficient accessories development. To achieve this goal, the focus was on hybrids and fuel cells for propulsion systems; aluminum, magnesium, and glass-fiber composites for body and chassis; and a range of improvements in accessories.

With improved vehicle performance and emission requirements, vehicle weight is becoming increasingly important to achieving better fuel economy. It is estimated that each 1% reduction in vehicle weight provides a 0.66% improvement in fuel economy, after a vehicle has been fully redesigned to account for the reduced power demands of a lower mass. The primary weight savings leads to secondary savings with lighter engine and powertrain configuration. Although more than 85% of fuel economy gain observed during the past three decades can be credited to technology improvements, the use of lightweight materials has played a key role in maintaining the vehicle weight, countering the weight added by the technology gains and passenger safety and comfort features. Furthermore, since 75% of a vehicle's energy consumption is directly related to factors associated with vehicle weight, the need to produce lightweight vehicles that also are safe and cost-effective is critical.

An average U.S. automobile weighs about 3240 lbs, of which body, chassis, and powertrain vehicle subsystems contribute 35%, 34%, and 27%, respectively, of its total weight. Compared to a U.S. automobile, European and Japanese automobiles are lighter in total vehicle weight due to a significant growth in the use of nonferrous metals and plastics/composites in the past. A European automobile has the least share of ferrous materials content (i.e., about 60%), and its current share of plastics/composites is estimated to be more than 10% of its vehicle weight. Despite the competition from other materials, ferrous metals continue to be the dominant material in an automobile, as shown for the United States in Fig. 1. Ferrous materials are flexible in use and relatively cheap, and their properties including its crash behavior and recyclability, are well understood and established. Although there has been a steady decline in conventional steel usage, the increased use of lighter, high-strength steel, with its improved formability in combination with efficient system design with lower weight, has helped to maintain steel's $>50\%$ share of total materials. Tailored blanks, hydroforming, and various metallurgical improvements have also kept the steel content of vehicles relatively constant. The decrease in iron content is due mainly to the switch to aluminum engine blocks and heads. The use of stainless steel is very limited to exhaust systems and engine gaskets applications today. The use of lightweight materials is largely responsible for the decline of 7.3% in total vehicle weight during the 1978–2001 period, as shown in Fig. 1.

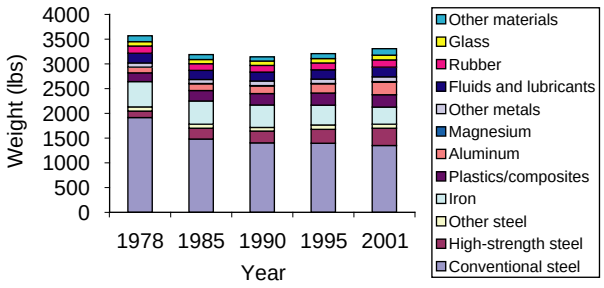


FIGURE 1 Material composition trend of a U.S. automobile.

Of the lightweight materials, aluminum, plastics/composites, and (to a far lesser extent) magnesium continue to have significant roles in automotive materials usage. The new Lincoln LS represents a current example of lightweight materials on a high-volume production vehicle. Aluminum, plastics, and magnesium total more than 20% of the vehicle's weight. Between 20 and 25% of the ferrous materials are high-strength steels selected to achieve weight reduction. Aluminum, composites, and magnesium have weight reduction potentials ranging from 25 to 75%, depending on the material replaced and automobile part application, as shown in Table I. Due to some inferior mechanical properties and manufacturing technology limitations, the mass reduction achieved using some of these lightweight materials in commercial applications is generally lower than the theoretical densities shown in this table. Consumer preferences—such as performance, style, comfort, economy, and the brand image—are relatively less important than many other hard-headed criteria—such as cost—which dictate the material choice selection by the vehicle manufacturer. Life cycle (e.g., recyclability, manufacturability, environmental issues, etc.) and engineering needs are also gaining momentum in material selection criteria.

Of the nonferrous lightweight metals, aluminum, having a density one-third that of steel, has seen the most use, with applications ranging from body closure panels to wheels and engines. Although aluminum parts are generally 1.5 times as thick as steel to achieve comparable stiffness, they still offer a significant weight savings. Aluminum use in automobiles has grown steadily from an average of 180 lbs in 1978 to 257 lbs in 2001, with roughly 80% of that in cast components, mostly engines and transmissions. This substitution of steel has been on a part-by-part basis, not the result of any radical design change. Aluminum is rapidly becoming the material of choice for vehicle's cylinder blocks.

Use of wrought aluminum sheet, however, is limited to air conditioning units and a few closure panels for the car body. Both panels and structures are still largely an unfilled dream for aluminum. The most popular body panel application is the hood (e.g., F-series pickup truck) due to weight balancing advantages. Liftgates are a rapidly growing application. Currently, only 5 to 10% of body panels are made of aluminum. A significant displacement of steel by aluminum would occur if and when steel as the primary material is being replaced in the chassis or the body of the car, since the latter contributes more than 30% of the average total vehicle weight. Aluminum panels on an extruded aluminum spaceframe were first used in the 1994 Audi A8—a low-volume production car. Price is a major problem presently, but the price (and its variability) is likely to be reduced as more aluminum scrap becomes available.

Composites have been the important peripheral contender in the material war between steel and aluminum with magnesium. Automotive composites consist of a reinforcing fiber in a polymer matrix, with polyester, vinyl ester, and epoxy resins the most often used matrixes. These polymer-based composites—having densities one-sixth to one-third that of steel, depending on the fiber type, fiber loading, and resin type—are the most common type of composites used in automotive applications. The fibers have densities ranging from 1.8–2.5 gm/cm^3, as shown in Table I; the commonly used polymer matrix materials have densities ranging from 1.1 to 1.4 gm/cm^3. About two-thirds of all composites currently use glass fiber and polyester or vinyl ester resins and provide a wide range of potential automotive applications, including body panels, suspension, steering, brakes, and other accessories. Demand varies widely by application. The composite materials have not changed significantly over the past 30 years, with glass fibers being the mainstay among reinforcement materials, while in some select applications carbon and aramid fiber materials are used. Today's average automobile is about 8% plastics and composites, combined, and their use has grown slowly, increasing by only 73 lbs during the 1978–2001 period, as shown in Fig. 1.

Polymer composites have been used for applications with low production volumes because they

TABLE I

Weight Savings and Relative Cost of Lightweight Materials

Lightweight material	Density (gm/cm^3)	Material replaced	Mass reduction (%)	Relative cost (per part)
High strength steel	7.8	Mild steel	10–24	1
Aluminum	2.7	Steel or cast iron	40–60	1.3–2
Magnesium	1.8	Steel or cast iron	60–75	1.5–2.5
Magnesium	1.8	Aluminum	25–35	1–1.5
Glass FRP	2.5 (fiber)	Aluminum	25–35	1–1.5
Carbon FRP	1.8 (fiber)	Mild steel	50–65	2–10+
Titanium	4.5	Alloy steel	40–55	1.5–10+
Stainless steel	7.9	Mild steel	25–40	1.2–1.7

require less lead times and lower investment costs relative to conventional steel fabrication. Important drivers of the growth of polymer composites have been the reduced weight and parts consolidation opportunities the material offers, as well as design flexibility, corrosion resistance, material anisotropy, and mechanical properties. Composites use is still predominantly in nonstructural elements of the vehicle because they are highly competitive for bolt-on exterior panels such as hoods, decklids, and fenders, especially in low- and mid-volume light-duty vehicles such as Corvette, Dodge Viper, and AP minivans. In addition, composite cross vehicle beams, radiator housings, fascia supports, bumper systems, and floor trays have contributed to weight savings. Corvette's body has been made largely of composites since the 1950s, and General Motor's Saturn and other vehicles use composite side panels. The latest growth in automotive composites has been mostly in exterior parts such as fenders, tailgates, and pickup beds.

Automotive manufacturers are becoming increasingly interested in carbon-fiber-based polymer composites because they offer high stiffness and low density for structural applications. They can also be made lighter than their glass-reinforced counterparts, providing a significantly higher weight-savings potential. Its most significant use to date has been in concept cars (e.g., about two decades ago in the Ford Granada and in General Motors' 1991 Ultralite). Carbon fiber composites became part of the government's research plans with the PNGV goal of achieving by 2011 a production-ready, 100-mpg, mid-sized vehicle, which is expected to contain at least 10 lbs of carbon fiber. In addition, there has been interest in fiber-reinforced thermoplastic composites that have the typical advantages of polymer matrix composites—such as weight savings, high strength, high stiffness, corrosion resistance, parts integration, and energy absorption—and have an indefinite shelf life, are recyclable, and are feasible for automated, high volume processing with a potential for rapid and low-cost fabrication. The most-used thermoplastics today are glass-filled thermoplastics developed for a variety of applications from intake manifolds to engine covers and, to a lesser extent, for body panels.

Magnesium is becoming a significant automotive material in varied applications such as engine blocks, seat frames, steering column components, transmission housings, wheels, and valve covers. Despite density that is 25% and 67% lower than aluminum and steel, respectively, magnesium use in automobiles has been limited to die castings because of its high cost and limited supply. The lower cost of producing die cast parts helps to offset its high cost. The market for automotive magnesium parts has grown rapidly, nearly 15% annually during the 1990s, achieving a level of about 8.5 lbs per vehicle in 2001. Some high-end models, such as the 2003 Dodge Viper, are slated to contain over 44 lbs of magnesium, mostly in the one-piece instrument panel that has been incorporated into the design. Magnesium use in wrought automotive products, such as seat frames, wheels, and body panels, has been very limited because of the lower process throughput and the metal's cost. Magnesium's creep resistance is inferior to aluminum's and limits its application in automotive powertrains today.

The "other metals" category in Fig. 1 mainly includes copper, brass, powder metal parts, and zinc die castings, where the latter represents only about 3% of this category. Zinc is also used for the coating of steel and being considered to protect automotive underbody parts. Nonferrous metals are important in powder metallurgy and metal matrix composites, both of which are being used increasingly in components such as camshafts, connecting rods, pistons, gears, and pulleys. The use of ferrous powder metallurgy parts in applications such as connecting rods has been trending upward primarily because it is less expensive than other methods. Metal matrix composites are gaining wide acceptance in niche applications such as brake discs, drums and calipers, pistons, prop shafts, and tire studs. Copper use in automobile radiators has been declining as manufacturers shift to aluminum, which is lighter and perceived to have a stable market price.

Another promising lightweight material that has only recently seen even limited use is titanium. It has potential application in suspension springs, exhaust systems, and a number of engine and chassis parts. Passenger car components that could benefit from titanium include engine valves, connecting rods, valve-spring retainers, and valve springs. Titanium has the benefits of being half the weight and twice the strength of steel (as shown in Table I), with great potential for a reduction in engine noise and vibration and improved durability due to reduced component loads, vehicle weight reduction, and improved powertrain efficiency. However, the cost (both raw material as well as component fabrication) is prohibitive, four times higher than even stainless steel. Titanium's high strength, low density, excellent corrosion resistance, and the relative worldwide abundance of its ore make it a

long-term viable lightweight alternative to steel in engine and body/chassis applications. Because of its high corrosion resistance at elevated temperatures, titanium is the only lightweight substitute for stainless steel in exhaust systems. The first use of titanium in a production automobile was for connecting rods in the Acura NSX V-6, and the use of titanium springs in the 2001 model year Volkswagen Lupo FSI has contributed to about 180 lbs reduction in weight.

2. FACTORS IMPEDING LIGHTWEIGHT MATERIALS GROWTH

Although lightweight materials have significantly penetrated the aerospace and defense industry, their cost has been the major barrier to their large-scale penetration in the automotive industry. A part-cost comparison (which includes both materials and manufacturing) of these materials indicates that relative part costs are anywhere from 1.3 to 10+ times higher than the materials they replace (see Table I). At the higher end of these cost differences are carbon fiber reinforced polymer composites, titanium, and magnesium, which are very little used in automobiles today. In most cases, the cost of the primary material is significantly higher than conventional steel's cost, and the manufacturing-process cost is only a small share of the overall part cost. Thus, any automotive system application with a greater potential for parts integration improves the economic viability of parts made of lightweight materials. These expensive lightweight materials have comparatively penetrated in Europe and Japan more than the United States due to lower fuel cost in the United States. In addition, there is a lack of economies of scale due to these materials' relatively smaller industry base. The current consumption of these materials is relatively small compared to steel, and in some cases (e.g., magnesium and rubber) the automotive industry accounts for a major share of the materials' total consumption, as shown in Table II. Most of the lightweight material industries are small, causing instability in material price that promotes substitution among the competing materials. Plastic in Table II does not include polymer composites, whose total consumption by the automotive industry is estimated to be 156 K tonnes in 2001.

Factors in addition to cost that have dampened the growth in lightweight materials use include relatively less pressure for improved fuel economy and from environmental pollution concerns, along with the availability of inexpensive gasoline and to some extent recent subsidies available to traditional materials such as steel. The existing well established manufacturing base is always a big positive factor for steel—there is always reluctance in the industry to avoid the enormous potential capitalization costs of large-scale conversion from steel to lightweight materials. All of the industries lack experience and play a small role in the automotive industry, situations that will only change with the materials' improved economic viability.

A widespread use of aluminum in automobiles would require a dramatic shift in wrought applications. Aluminum stamped parts are more expensive than extruded and cast parts because production-line rates are slower, there are difficulties in welding and assembly, aluminum sheet cost is higher as are die costs because they require special coatings to compensate for springback. The tendency of aluminum sheets to tear requires low stamping rates and extra hits for the stampings. Most aluminum used in outer-body panels is heat-treatable, 6000-series aluminum alloys, which are still relatively expensive to produce. The potential exists to produce 5000-series aluminum alloy for inner body panels at $1/lb using continuous casting and exploiting large economies of scale. Under existing fabrication conditions, only a drastically consolidated design such as a spaceframe is even slightly competitive with the conventional unibody design. A cost comparison indicates that aluminum body structures made of continuous-cast, coated aluminum at $1/lb (unlike ingot and stamped sheet rolling cost at $1.30/lb) cost 1.6 times more than body structures made of conventional steel. It is not only raw material cost difference, but also the rivet bonded assembly for aluminum structures that causes this large cost difference.

TABLE II

2001 Material Usage by the Automotive Industry

Material	U.S. total consumption (tonnes)	Automotive share (%)
Steel	90 M	14
Iron	9 M	31
Aluminum	11 K	32
Plastic	46 K	4
Magnesium	96 K	47
Zinc	1 M	23
Rubber	3 M	68

As with other lightweight materials, the price of resins and glass fiber reinforcements poses a major barrier to the growth of automotive composites. The long cure-cycle times required for automotive part applications make it impossible to maintain high automotive production rates. For composites to be cost competitive on a part-by-part substitution, improvements are necessary in cycle times and material utilization, which in some cases have been estimated to currently contribute 21% and 60%, respectively, of the total cost of carbon-fiber-reinforced thermoplastics. On a $/lb basis, the cost of polymer composites is about two to three times higher than steel, but a recent study comparing the monocoque designs indicates that the cost of glass-fiber reinforced thermosets and carbon-fiber-reinforced thermoplastics are about 62% and 76% higher than the conventional steel unibody. The material cost plays a key role in the economic viability of polymer composites, particularly at higher production volumes for glass-fiber-reinforced thermosets and for carbon fiber-reinforced thermoplastics composites. A 50% reduction in the carbon fiber cost (to prices in the $3 to $5/lb range) and smaller cost reductions in other thermoplastic materials are needed for the composites to be economically viable. Also impeding large-scale use of composites is a curious mixture of concerns about material issues such as crash energy absorption, recycling challenges, competitive cost pressures, the industry's general lack of experience and comfort with the material, and industry concerns about its own capabilities.

Technical and institutional drawbacks, in addition to cost, have limited magnesium use in vehicles, especially for wrought magnesium products. Because magnesium is very reactive, corrosive, hard to form at room temperature, and has a relatively high percentage of scraps and refuses, its part production cost is significantly higher than that of competing materials. In addition, magnesium has a small supply base, with annual production about 1400 times and 45 times less than steel and aluminum, respectively. Because it is a small industry and not traded on the London Metal Exchange, its price is more prone to swings as demand grows and absorbs available production. In addition, past supply expansions in magnesium production capacity have not been systematic enough to harness potential demand for low-cost material in the auto industry. Automakers' demand surpassed industry capacity, causing wild price swings, which caused parts manufacturers to switch from magnesium to other competing materials. Recent cheap imports, particularly from Russia and China, have caused a severe blow to the industry, forcing several magnesium production plants worldwide to close down, including one of the two remaining plants in the United States.

Several current governmental R&D programs (e.g., the Automotive Lightweighting Materials Program by the United States Department of Energy in collaboration with the automotive industry) focus on the viability of primary material production and manufacturing technologies for significantly reducing automotive vehicle body and chassis weight without compromising other attributes such as safety, performance, recyclability, and cost. Aluminum and glass-reinforced polymer composites have been considered as the near-term automotive materials, whereas, magnesium, titanium, and carbon-reinforced polymer composites are considered for the long term. Similarly, the Canadian Lightweight Materials Research Initiative (CLiMRI) coordinates research and development of materials and processes for weight reduction in transportation applications, besides the multidisciplinary auto-related research and development initiative "Auto21" to influence the vehicle in the 21st century. The Industrial and Materials Technologies program of the European Union indicates similar activities in this area of research. The resulting ongoing research is expected to address some of the aforementioned impediments to lightweight material use.

3. ROLE OF LIGHTWEIGHT MATERIALS IN ENERGY CONSUMPTION

Due to the increasing use of automotive lightweight materials, concerns have surfaced regarding the fuel savings, lower emissions, and weight savings, particularly with respect to aluminum and magnesium. Life cycle analysis has shown that high energy consumption during production of virgin aluminum and magnesium is offset by the reduced energy consumption in the use phase, bringing the breakeven point to within the lifetime of the car. The point at which a breakeven is reached depends on the amount of lightweight materials included, the amount of recycled material used for production of the part, and the system and conditions assumed in the study. Life cycle assessments with the consideration of "cradle to grave" aspects of a material used in automobiles have been instrumental in providing a

more holistic view of energy consumption, considering energy consumption beyond the materials production stage and demonstrating a significant potential for energy savings.

The major stages of "life" included in the life cycle analysis of automotive lightweight materials are material extraction, part production, use phase, and end-of-vehicle stage. The second and last life cycle stages consume considerably less energy than the first and third stages and have only sparse data about them available in the literature today. The use phase energy consumption has the biggest impact on the life cycle analysis, contributing 60 to 80% of the total energy consumption. The next two largest shares, in order, are material production and part manufacturing. The transportation necessary within each of the phases has often been included in the past studies. The life cycle energy benefit in switching to a lighter weight material is not just a function of weight saved; it is also dependent on material production energies and substitution factors and vehicle operational efficiency.

Table III shows the primary- and secondary-production energy of various materials used in automobiles today. There is considerable variation—as much as 40%—in reported values of material production energies, depending on the estimation methodology used and assumptions about production technology and its efficiency. The primary production energy is the sum of all energies required to extract raw materials from the earth (e.g., mine ore or pump oil) and to process (wash, concentrate, or refine) it into a usable form (ingot or rolled sheet). Much of this production energy is avoided by recycling, although getting a recycled material of a comparable purity to virgin stock is often not readily achieved due to the difficulty in removal of impurities. The recycled content is an important factor in the overall life cycle analysis, as less than 15% of the energy used for production of virgin material is needed for remelting in case of aluminum and magnesium, compared to 30% in the case of steel as shown in Table III. Recycling is gaining momentum with the End-of-Life directive in the European Union affecting the design phase and material decision making processes of automobile manufacturers as well as influencing other industrial activities such as automotive dismantlers and shredders. The directive's recycling targets are 80% and 85% by 2006 and 2015, respectively, with a considerable higher recovery rate of 85% and 95%, respectively, for the corresponding years. This is in contrast with 75% recycling rate today, mostly limited to ferrous materials.

TABLE III

Primary and Secondary Production Energy for Automotive Materials

	Energy (MJ/kg)	
Material	Primary	Secondary
Steel	30	9
Cast iron	34	24
Wrought aluminum	201	29
Cast aluminum	189	26
Cast magnesium	284	27
Reinforced plastic	56	37
Unreinforced plastic	79	14
Copper	140	35
Zinc	53	16

The life cycle energy consumption of a generic family U.S. sedan based on 120,000 lifetime mileage is estimated to be about 974 GJ, which corresponds to the consumption of 46,000 pounds of hydrocarbons (coal, oil, and natural gas). Weight reduction through the use of lightweight materials is shown to significantly reduce the life cycle energy of a vehicle. Several life cycle analyses on a single product basis indicate that although aluminum production is about ten times more energy intensive than steel production, aluminum in automotive applications is less energy intensive during its life cycle than conventional steel. Since vehicle operation dominates overall life-cycle energy use, "lightweighting" with aluminum increases fuel efficiency enough to reduce overall energy use.

For example, a radical use of carbon fiber composites and aluminum (i.e., replacement of all steel and iron with carbon fiber composites and aluminum, respectively, resulting in overall 40% vehicle weight savings) is predicted to yield a life cycle energy savings of 16% for a conventional mid-sized vehicle. The use of recycled materials improves life cycle energy consumption considerably—for example, a 7.5% life cycle energy reduction for a particular aluminum substitution example becomes 12% when the part is 50% recycled aluminum. If aluminum-intensive light-duty vehicles were commercialized on a mass scale by 2005, the reported national petroleum energy savings would be quite significant, about 4% and 5% by 2020 and 2030, respectively. The impacts of lightweight materials on non-conventional vehicles such as electric and hybrid are comparatively less due to the use of highly energy

efficient power-plants that reduce energy consumption in the vehicle use stage.

Life cycle energy comparisons on a single product basis are quite sensitive to the underlying assumptions made for major input parameters. For example, a comparative assessment of material substitution examining automatic transmission cases made from die-cast magnesium and aluminum finds aluminum use more favorable, despite magnesium's significantly higher weight reduction potential. This result is assumed to be due to the different recycling rates for the two materials. On the other hand, magnesium substitution ranging from 10 to 30% in automotive parts like gearbox housing and dashboards, indicated a breakeven point around 40,000 miles.

Recent studies argue that since new product designs are introduced progressively in a fleet, only a fleet level analysis can account for the temporal effects of use and the distribution of energy over a vehicle lifetime. Since the product manufacture and use are distributed over long periods of time that are not simple linear combinations of single product life cycles, it is all the products in use over a period of time, rather than a single product that are more appropriate for the life cycle analysis. This approach allows the prediction of time when benefits are realized. A recent study looking at the rolled products for body-in-white applications (in which aluminum provides the greatest weight saving potential compared to steel and ultra light steel autobody) found that it would take about 4 and 10 years, respectively, for aluminum vehicles to achieve a life cycle energy equivalence with steel and ultra light steel autobody on a single vehicle basis. It would take twice as long to realize the benefits of aluminum in the fleet.

4. AUTOMOTIVE LIGHTWEIGHT MATERIALS OUTLOOK

During the next two decades, a combination of customer expectations, environmental stewardship, and tightened regulations will lead to significant changes in the automobile technologies, where materials will be at the forefront. As technological improvements continue to address the cost and manufacturability issues of automotive lightweight materials, a shift from current high-end and niche applications to widespread applications on a part-by-part basis will be seen. The pressure to boost the light-truck corporate fuel economy with an anticipation of higher future sales in this vehicle category will provide impetus to the growth in the automotive lightweight materials during this decade. As long as the pressure of the 'hypercar" strategy based on a conceptual vehicle combining very low weight with a highly aerodynamic design and a electric drive system capable of achieving a fuel economy 80 to 100 mpg continues, a much greater role for composites, aluminum, magnesium, and possibly titanium in automotive use will be seen. These materials would enable less powerful engines to be used for equivalent performance, compounding the weight reduction throughout the vehicle. This higher usage of lightweight materials trend can be seen in the concept cars developed by the three automobile manufacturers (i.e., Precept by General Motors, ESX3 by DaimlerChrysler and Prodigy by Ford) during the end of the past decade to meet the three-times-greater fuel efficiency goal of the PNGV program, as shown in Table IV. There is a significant increase in aluminum use for all of these concept vehicles, whereas composites use increases significantly in the case of ESX3 only. Carbon fiber and titanium, now rarely used in vehicles, had a significant share in the material composition of these concept vehicles. Development of more efficient design and manufacturing initiatives, as well as parts consolidation, will be necessary for the widespread use of lightweight material

TABLE IV

Material Composition of PNGV Concept Vehicles by Three Automobile Manufacturers

Material	Baseline	Prodigy	ESX3	Precept
Wrought aluminum	47	462	330	304
Cast aluminum	159	271	120	820
Magnesium	6	86	122	7
Titanium	0	11	40	33
Platinum	0.0033	0.01	0.01	0.01
Ferrous	2168	490	528	487
Plastics	193	209	52	187
Resins (for composites)	28	40	428	86
Carbon fiber	0	8	24	22
Glass fiber	19	19	60	35
Lexan	0	30	20	0
Glass	97	36	70	57
Rubber	139	123	148	77
NiMH or LiIon batteries	0	138	88	138
Other	391	83	212	338
Total	**3248**	**2010**	**2250**	**2591**

applications. Lightweighting will become more important as the trend for comfort and safety related system gadgets continue to add weight to the vehicle. With the European directive establishing considerably higher recycling goals, recyclability will pose major barriers to these lightweight materials unless viable recycling technologies and infrastructure are solved and in place.

The development of high-strength steels and relevant design opportunities mean major weight savings can be made without resorting to other materials. Hence, steel is likely to be a substantial automotive material for at least the first third of the century. The $22 million ultra light steel auto body (ULSAB) project recently completed by American Iron and Steel Institute (AISI) demonstrated a 36% mass reduction using high-strength steels based on drastic design changes on benchmark vehicles using tailor-welded blanks, lighter material, and advanced forming processes. A similar ultra light steel auto closures (ULSAC) project demonstrated a number of weight-saving vehicle closure elements such as doors, hoods, decklids, and hatchbacks. Faster growth is predicted for high-strength steel than for aluminum in body panel applications. High-strength steels in optimized designs enable substantial weight reductions when made by innovative manufacturing and assembly technologies. In addition, an extensive use (i.e., about 80%) of advanced high-strength steels (AHSS)—a new category of steels that combine very high-strength with excellent formability—defines the latest ULSAB-Advanced Vehicle Concept (AVC) project by AISI.

Growth in the use of aluminum castings is anticipated to continue, particularly for engines and wheels. Another aluminum-casting growth area is anticipated to be suspension (e.g., suspension knuckles and control arms, as well as structural components, drivelines, and interiors). The possibility exists for other lightweight materials to capture some of these substitution opportunities (e.g., for intake manifolds to be made with reinforced nylon and other polymers and for transmission transfer cases made by magnesium casting). Panel and structures applications are still largely an unfilled dream for aluminum. Aluminum still has to overcome significant technological and economic hurdles before it can replace steel in the car body, and competing with recent advancements in high-strength steel will be difficult. The shift of aluminum producers from being merely material suppliers to being partners with the automakers will facilitate larger-scale use of aluminum. Developments in aluminum stamping, legislative pressures to improve of fuel economy, and recycling targets might make aluminum the primary material in the auto body.

Glass-fiber-based polymer composites using thermosets as the matrix material will likely hold their own, but usage will grow at a much lower rate than expected for reinforced thermoplastics, which are likely to capture nearly all the growth anticipated later this decade. These materials will play a significant role where part consolidation is great, at a system level rather than on part-by-part substitution. Low tooling cost and lightweighting potential will continue to be the driving force for increased penetration into the light truck area. Due to the cost and toughness advantages of thermoplastic composites, automotive underbody shield application areas look promising. Carbon-fiber-based polymer composites use will be limited to high-end and niche applications demanding substantially higher weight reduction; further growth will be impeded by the significant cost of carbon fiber.

There appears to exist a general commitment towards the use of magnesium by the automotive manufacturers as evidenced by the partnerships formed between them and major magnesium suppliers. These partnerships include Volkswagen and the Dead Sea Works, Ford and Australian Magnesium Company, and General Motors and Norsk-Hydro. Although magnesium remains a major competitor of aluminum, its growth will be limited mainly to niche automotive applications from the die casting industry until major breakthroughs are achieved in wrought automotive applications. The historical, worldwide growth rate for magnesium die casting has been 13 to 14%, a rate that appears to be sustainable with continued pressure to increase fuel economy and as more vehicles use magnesium dashboards, seats, transfer cases, knee bolsters, and so on. Since the largest current use of magnesium is for aluminum alloying, growth in use of these materials are complementary in the context of lightweighting the automotive fleet. If magnesium's wild price swings observed during the past decade—resulting from the mismatch between supply and demand and the more recent market oversupply with cheap imports from non-Western countries—continue, they will be detrimental to the industry's viability. It is anticipated that a slow expansion in the supply base and prices maintained at above $1.40/lb can lead to a more stable demand and supply growth—in the range of 20 lbs to 80 lbs of magnesium per vehicle by 2015 compared to less than 10 lbs today.

5. SAFETY CONSIDERATIONS IN AUTOMOBILE LIGHTWEIGHTING

It has been strongly argued that improving fuel economy by reducing vehicle weight reduces vehicle safety. Technological improvements that have played a key role in fuel economy improvements appear to be neutral or beneficial to safety. The relationship between vehicle weight and safety is important by itself. Vehicle safety concerns have been gaining momentum lately with the growth of light-duty trucks and sport utility vehicles (SUVs) and the pending legislation to boost the CAFÉ for light duty trucks by 1.5 mpg in model year 2007 vehicles. In 1988, light trucks constituted roughly 30% of the vehicle fleet; by 1994 they were more than 40%; and by 2000 they reached an estimated 45% of the fleet. As the number of car-light truck collisions increased, it raised the question whether the growth in the number and weight of light trucks is having an adverse impact on the safety of passenger car occupants and other road users, possibly exceeding any safety benefits of the vehicle-weight increases for the occupants of the trucks.

Safety issues surrounding a general downsizing are concerned with the details of how vehicle designs may change, differences in the performance of lighter weight materials, the precise distribution of changes in mass and size across the fleet, and interactions with other highway users. Of the driver, environment, and vehicle, the vehicle is the least important factor in highway fatalities. To isolate the effects of a less important factor from the effects of more important yet related factors is often not possible. Measure of vehicle exposure with which to control for confounding influences of drivers, environment, and other vehicle characteristics are almost always inadequate—resulting in biased inferences in the literature today.

In a collision between two vehicles of unequal weights, the occupants of the lighter vehicle are at a greater risk whereas occupants of the heavier vehicles benefit, and so there may not be a win-win situation from a societal perspective. In addition, because of the powerful heavier light-truck designs, lighter vehicles are at a serious safety disadvantage. A clear pattern has been observed: reducing a vehicle's weight increases net risk in collisions with substantially larger and heavier entities, reduces net risk in collisions with much smaller and more vulnerable entities (e.g., motorcyclists, bicyclists, and pedestrians) and has little effect on net risk in collisions with vehicles of about same size. Overall, it has been found that the variation of weight among cars results in a net increase of fatalities in collisions, as the reduction in the fatality risk for the driver of the heavier car is less than the increase of the fatality risk for the driver of the lighter car. This increase of fatalities in collisions is offset some by benefits to smaller, lighter highway users (pedestrians and cyclists) in collisions with lightweighted vehicles. The relative weight of the vehicles rather than their absolute values, as well as the distribution of vehicle weights, is what leads to the adverse consequences for the occupants of the lighter vehicle. In fact, some evidence exists, such as the 1997 National Highway Traffic safety Administration study, that proportionately reducing the mass of all vehicles would have a beneficial safety effect in vehicle collision. However, if cars are downweighted and downsized more than light trucks, the increased disparity in weights would increase fatalities.

There is also evidence that smaller, lighter cars have historically had a greater propensity to roll over, and single-vehicle rollover crashes are a major component of all traffic fatalities. The stability of vehicles depends on their dimensions, especially track width, and the height of their center of gravity. The correlation between vehicle weight and the stability has not been strong particularly for light trucks, given the historical tendency that reduced mass means narrower, shorter, less stable cars. It has been argued with some evidence that increase in fatalities due to downweighting and downsizing in single vehicle accidents is a matter of vehicle design rather than mass and that smaller and lighter vehicles have better crash avoidance capabilities. The effect of weight reductions on fatalities in passenger car rollovers might be smaller if weight could be reduced without changing track width.

A recent ACEEE study provides additional evidence to contradict conventional wisdom. This study was based on a measure of risk comparing driver fatalities to the number of vehicles on the road, associated with recent vehicle models sold between 1995 and 1999 with up-to-date safety designs and restraint technologies. The study found that many small cars have lower fatality rates among their drivers than do SUVs or pickups. SUVs have been the major concern, as they are the fastest growing segment of new vehicles, today comprising 21% of the new vehicle market, up from 6% just 13 years ago. This study further illustrated the importance of design in determining a vehicle's safety. Among all major vehicle groups, minivans have the lowest

fatality rates among their drivers, while pickups have the highest. Minivans generally are built on car rather than pickup truck platforms, which may reduce the risk to their drivers.

Overall, there does not appear to be any direct correlation between improved fuel economy and vehicle safety, perhaps largely because fuel economy improvement to date have been the result of technology advances rather than decreased weight. The relationships between vehicle weight and safety are complex and not measurable with any reasonable degree of certainty at present. The distribution of vehicle weights is an important safety issue. It is important to know how vehicle weight and size changes will affect the size/weight distribution of the fleet and, finally, how this distribution will affect different fatality rates (i.e., single-vehicle versus vehicle-to-vehicle rates). Vehicles of uniform weight would correct for the negative social externality that encourages individuals to transfer safety risks to others by buying ever larger and heavier vehicles. It appears that in certain kinds of accidents, reducing weight will increase safety risk, while in others it may reduce it. Reducing the weights of light-duty vehicles will neither benefit nor harm all highway users; there will be winners and losers. Lastly, there appears to be no definitive answer to the fuel-economy and safety question, but to the extent that fuel economy can be improved with advanced technology rather than size and weight reductions, adverse safety impacts, if any, can be minimized.

SEE ALSO THE FOLLOWING ARTICLES

Aluminum Production and Energy • Fuel Economy Initiatives: International Comparisons • Glass and Energy • Industrial Energy Use, Status and Trends • Internal Combustion Engine Vehicles • Material Efficiency and Energy Use • Motor Vehicle Use, Social Costs of • Plastics Production and Energy • Recycling of Metals • Steel Production and Energy • Transportation Fuel Alternatives for Highway Vehicles

Further Reading

Das, S. (2000). The life cycle impacts of aluminum body-in-white automotive material. *J. Metals* **52**(no. 8), 41–44.

Faller, K., and Froes, F. H. (2001). The use of titanium in family automobiles: Current trends. *J. Metals* **53**(7), 27–28.

Kahane, C. J. (1997). "Relationships between Vehicle Size and Fatality Risk in Model Year 1985–93 Passenger Cars and Light Trucks." NHTS Technical report, DOT HS 808 570. National Technical Information Services, Springfield, VA.

Kramer, D. A. (2002). "U.S. Geological Survey Minerals Yearbook – 2001." U.S. Geological Survey. Also available at http://minerals.er.usgs.gov/.

National Academy Press (2002). "Effectiveness and Impact of Corporate Average Fuel Economy (CAFÉ) Standards." U.S. Government Printing Office, Washington, DC.

Paxton, H. W., and DeArdo, A. J. (1997). What's driving the change in materials. *New Steel* **1**(2), 64–74.

Pinkham, M. (2001). Will the 21st century be the aluminum age? *Metal Center News* **41**(8), 32–35.

Powers, W. F. (2000). Automotive materials in the 21st century. *Adv. Mat. Proc.* **157**(5), 38–41.

Ross, M., Wenzel, T. (2001). "An Analysis of Traffic Deaths by Vehicle Type and Model," Report No. T021. American Council for an Energy Efficient Economy, Washington, DC.

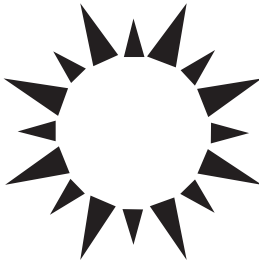
Schultz, R. A. (1999). "Aluminum for Lightweight Vehicles: An Objective Look At the Next 10 to 20 Years." Presented at the Metal Bulletin 14th International Aluminum Conference, Montreal, Canada, September 15.

Sullivan, J. L., and Hu, J. (1995). "Life Cycle Energy Analysis for Automobiles." SAE Paper no. 951829, Warrendale, PA.

Ward's Communications (2002). "Ward's Motor Vehicle Facts & Figures." Southfield, MI.

Mechanical Energy

JOSEPH PRIEST
Miami University
Oxford, Ohio, United States



1. Mechanical Energy
2. Work
3. Kinetic Energy
4. Work–Energy Principle
5. Conservative and Nonconservative Forces
6. Potential Energy
7. Conservation of Mechanical Energy
8. Energy Conversion
9. Applications of Mechanical Energy

Glossary

conservative force When the net work done by a force is zero for every path that ends up at the starting point.
energy The capacity or ability of an object to do work, with the joule (*J*) as the measuring unit.
Hooke's Law When an elastic object exerts a force proportional to the displacement of the object and in the opposite direction.
joule (J) The unit of work and energy; a force of 1 newton (N) acting over a distance of 1 m does 1 J of work.
kinetic energy The ability of an object to do work as a result of having mass and speed; in terms of mass (m) and speed (v), kinetic energy is $\frac{1}{2}mv^2$.
mechanical energy The sum of the kinetic energy and potential energy of an object.
nonconservative force When the net work done by a force is not zero in a path that ends up at the starting point.
potential energy The ability of an object to do work as a result of an advantageous position.
power The rate of doing work or converting energy; the watt (W) is the metric measuring unit, where 1 W is equal to a rate of 1 J per second.
restoring force A force on an object that tends to restore the object to its condition prior to application of the force.
watt (W) A rate of doing work or converting energy; the watt is the metric measuring unit, where a rate of doing work of 1 J per second is a watt (P = work/t).
work The result of a force acting on an object as the object moves from one position to another; in one-dimensional motion, work is the product of the force (F) and displacement (d).

The physical concepts of work, kinetic energy as energy due to motion, and potential energy as energy due to some favorable position are discussed in this article. Added together, kinetic energy and potential energy are called mechanical energy. In the absence of nonconservative forces (e.g., friction), the mechanical energy of an object does not change with time and is said to be conserved. The motion of a child on a swing is discussed in light of the conservation of energy. A wide variety of important types of kinetic energy are derived from the conversion of other forms of energy. Such is the case in a hydroelectric system that is discussed in some detail in the article.

1. MECHANICAL ENERGY

Energy has some meaning to everyone. A person often does not have energy following a bout with a cold. Concerns about depletion of our energy resources, solar energy, wind energy, and nuclear energy are common topics in newspapers and on television. Although energy has several meanings, in a physical sense it is considered "a capacity for doing work." An object can have a capacity for doing work due to its motion and by virtue of an advantageous position. The combination of these two types of energy is called mechanical energy. The implications of mechanical energy are tied to the meaning of work.

2. WORK

A person does physical work when cleaning a room. A student does mental work when preparing for an exam. Whether physical or mental, work involves an effort directed toward producing some outcome. In the physical sense, effort is associated with force (push or pull), and work is done when a force acts on an object as it moves through some distance. The

 © 2004 Elsevier Inc. All rights reserved.

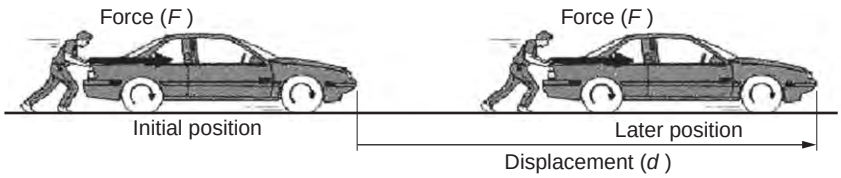


FIGURE 1 Illustration showing a car moving to the right that has been displaced by an amount d. During the displacement, the person exerts a force F to the right. The work (W) done by the person is the product of force and displacement ($W = Fd$).

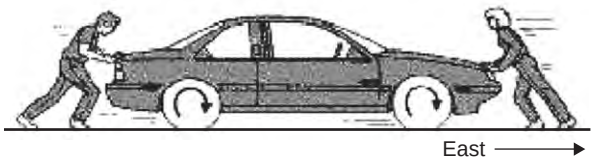


FIGURE 2 Illustration showing a car traveling east that is aided by a person pushing east and hindered by a person pushing west. The work done by the person pushing east is positive. The work done by the person pushing west is negative.

person cleaning a room does work on the broom by displacing it through some distance. In the simplest situation, the direction of the force is either the same as the direction of the displacement or opposite to the direction of the displacement. This being the case, the numerical value of the work (W) is the product of the force (F) and the displacement (d) (Fig. 1):

$$W = Fd \tag{1}$$

Measuring force in newtons (N) and displacement in meters (m) gives work the units of newton-meters. One newton-meter is called a joule (J).

The physical definition of work retains much of the popular notion with one important difference. No matter how much force (effort) is exerted, no work is done in the physics sense if the object does not move. A person may fret and sweat while pushing on a heavy box, but no work is done if the box does not move.

A force in the same direction as the direction of the movement produces an effect that is very different from that if the directions were opposite. A person pushing eastward on a car moving eastward tends to help the car along (Fig. 2). However, a person pushing westward on the same car tends to slow the car. We distinguish these two situations by labeling the work as "positive" when force and movement have the same direction and as "negative" when force and movement have opposite directions. The car is moving eastward under the action of one person pushing eastward and another person pushing westward. The person pushing eastward does positive work. The person pushing westward does negative work.

The net work on an object is the algebraic sum of the works done by each force acting on the object. Suppose that a sled moves 2 m eastward as a result of a girl pulling due east with a force of 100 N and a boy pulling due west with a force of 50 N. The girl does

$$+100\ \text{N} \times 2\ \text{m} = 200\ \text{J of work},$$

the boy does

$$-50\ \text{N} \times 2\ \text{m} = -100\ \text{J of work},$$

and the net amount of work is

$$+200\ \text{J} - 100\text{J} = +100\ \text{J}.$$

3. KINETIC ENERGY

A day does not pass without a person being involved in physical work. The forces that move one in walking and running perform work. Lifting food to one's mouth and chewing food involves forces doing work. On some days, a person works more easily and more efficiently than on other days. One's activities may be cast in terms of his or her ability to do work. Because physical work involves forces and movement, one might ask, "Under what conditions does something have a capacity for moving an object through some distance?" A car in motion has this capacity because if it slams into the rear of a car stopped at a red light, the struck car will surely move some distance. The energy associated with masses in motion is called kinetic energy. A big car colliding with a stopped car will "do more work" on the stopped car than will a smaller vehicle colliding with a stopped car. Similarly, a fast-moving car colliding with a stopped car will "do more work" on the stopped car than will a slow-moving car of the same type colliding with a stopped car. A numerical evaluation of kinetic energy should reflect these observations. Formally, the kinetic energy (K) of an object having mass m moving with speed v is defined as one-half the product of mass and square of speed:

$$K = \tfrac{1}{2}mv^2. \tag{2}$$

Both energy and work are measured in joules. Importantly, Eq. (2) shows that kinetic energy increases if the mass and/or the speed increase. Doubling the mass of an object while keeping its speed the same will double the kinetic energy. Doubling the speed of an object while keeping its mass the same will increase the kinetic energy by four times. The dependence on speed is especially significant.

4. WORK–ENERGY PRINCIPLE

At some time, an object such as a car may have speed v and kinetic energy $K = \frac{1}{2}mv^2$. Later, its speed may change to V so that its kinetic energy is $\frac{1}{2}mV^2$. The change in kinetic energy is the later value minus the earlier value, that is, $\frac{1}{2}mV^2 - \frac{1}{2}mv^2$. The work–energy principle states that the net amount of work on the object between the initial and later times is equal to the change in kinetic energy:

$$W_{net} = \tfrac{1}{2}mV^2 - \tfrac{1}{2}mv^2. \qquad (3)$$

When a hockey player hits a stationary 0.16-kg puck with a hockey stick and imparts to it a speed of 45 m/s (100 miles per hour [mph]), its kinetic energy changes from zero to $\frac{1}{2}(0.16\text{ kg})(45\text{ m/s})^2 = +162\text{ J}$. The change in kinetic energy is $+162\text{ J} - 0\text{ J} = +162\text{ J}$, and the net work is $+162\text{ J}$. The glove of a baseball catcher receiving a 0.15-kg baseball traveling 40 m/s (90 mph) reduces the speed of the baseball to zero. The change in kinetic energy of the ball is $0\text{ J} - \frac{1}{2}(0.15\text{ kg})(40\text{ m/s})^2 = -120\text{ J}$, and the net work is -120 J.

You see this principle in operation in many processes. The kinetic energy of water flowing over a dam increases as its speed increases. The increase in kinetic energy is the result of positive work done on the water by the gravitational force. When a car starting from rest is set into motion, its kinetic energy increases. This is due to work done by a force in the direction of motion of the car. Likewise, when the car slows down, its kinetic energy decreases. This is due to (negative) work by a force on the car acting in a direction opposite to the direction of motion.

5. CONSERVATIVE AND NONCONSERVATIVE FORCES

A box being lifted from the floor is acted on by an upward force provided by the lifter and a downward force due to gravity. Positive work is done by the lifter because the force and movement are in the same direction. The gravitational force does negative work because the force and movement are in opposite directions. If the person lowers the box back to the floor, the gravitational force does positive work because the force and movement now have the same direction. The negative work done by gravity in the upward movement is equal in magnitude, but of opposite algebraic sign, to the work done by the gravitational force in the downward movement. Accordingly, the net amount of work done by the gravitational force in the round trip of going up and back down is zero. When the net amount of work done by a force in any trip that ends up at the starting point is zero, the force is said to be conservative. If the net amount of work done by a force in a trip that ends up at the starting point is not zero, the force is said to be nonconservative. A friction force is a nonconservative force. A friction force acting on an object always opposes its movement so that the work due to friction is always negative. Consequently, the net work due to friction in a trip that ends up at the starting point is never zero.

6. POTENTIAL ENERGY

When an object such as a box is in an elevated position, it has "a capacity for doing work" because if it is dropped and hits something, it can exert a force on that something and push it through some distance. The box at an elevated height is an example of potential energy. Potential energy is associated with conservative forces. By definition, the change in potential energy when moving from one position to another is the negative of the work done by the conservative force acting during the change in position. Labeling U the potential energy and ΔU the change in potential energy, the definition may be expressed as

$$\Delta U = -W_{\text{conservative}}. \qquad (4)$$

6.1 Gravitational Potential Energy

The gravitational force on a box of mass m is mg, where g is the acceleration due to the gravitational force. If the box is raised a height h, the work done by gravity is $W = -mgh$. The change in potential energy of the box is then $\Delta U = -W_{\text{conservative}} = +mgh$. The positive sign means that the potential energy has increased. When the box falls from the height h, the work done by gravity is positive and the change in

potential energy becomes $\Delta U = -mgh$; the potential energy has decreased. The potential energy acquired by the box by placing it in an advantageous position can be recovered by letting it drop to its initial position on the floor. Water atop a dam in a hydroelectric plant has potential energy. When it falls toward the bottom of the dam, it loses potential energy but gains kinetic energy. At some point, the water does work by pushing on the blades of a water turbine, and the kinetic energy of the water is converted to rotational energy of the turbine.

6.2 Elastic Potential Energy

A material is said to be elastic if it returns to its original condition after being stretched or compressed. A spring, a rubber band, the bow of a bow and arrow, and a shock absorber on a car are examples of elastic objects. Somewhat like the earth pulling downward on a box that is being lifted upward, an elastic object pulls in an opposite direction to the force that is stretching or compressing it. The object has potential energy in the stretched or compressed condition due to its capacity to do work if it is released. For many elastic objects, the force exerted by the object, called the restoring force, is proportional to the extension or compression and in a direction opposite to the extension or compression. This being the case, the object is said to obey Hooke's Law. A force obeying Hooke's Law may be expressed as

$$F = -kx, \tag{5}$$

where F is the force exerted by the elastic object, x is the extension or compression, and k represents the strength of the force. For a linear spring, k is called the spring constant having units of newtons per meter. The stronger the spring, the larger the spring constant. The potential energy of a spring obeying Hooke's Law is given by

$$U = \tfrac{1}{2}kx^2. \tag{6}$$

Elastic potential energy is used in many ways. The kinetic energy acquired by an arrow has its origin in the elastic potential energy in the flexed bow. Pole vaulters acquire much of their vault from a bent pole. Toy guns expel projectiles by releasing the elastic potential energy of a compressed spring. Atoms in a molecule are held together by spring-like forces that lead to a form of potential energy. Release of that potential energy often leads to the emission of light.

7. CONSERVATION OF MECHANICAL ENERGY

Friction is always present in a mechanical system, but if it can be ignored, the mechanical energy (i.e., the sum of the kinetic energy and potential energy) does not change with time. To illustrate, a girl waiting to move down a playground slide has potential energy but no kinetic energy because she is at rest. Moving down the slide, she loses potential energy but gains kinetic energy. If friction between the girl and the slide can be ignored, the mechanical energy at any moment is unchanged. At the bottom of the slide, all of her initial potential energy would have been converted to kinetic energy. The constancy of mechanical energy is called the conservation of mechanical energy.

To the extent that friction can be ignored, the mechanical energy of a child on a swing is constant at any moment (Fig. 3). Held in an elevated position waiting for the swing to begin, the child has only potential energy. When released, the child gradually loses potential energy but gains kinetic energy. The sum of the two energies is unchanged at any moment. At the lowest portion of the swing, the potential energy is zero, making the kinetic energy a maximum. As the swing moves upward from the lowest position, the child gradually gains potential energy but loses kinetic energy. The child is momentarily at rest at the uppermost position of the swing, making the kinetic energy zero and the potential energy a maximum. If friction were absent, the back-and-forth motion would continue unabated and mechanical energy would be conserved at any moment. But as anyone who has taken a child to a playground knows, friction is always present and the motion gradually dies out unless the person pushes the swing to replace the energy lost to friction.

8. ENERGY CONVERSION

An object in an elevated position clearly has potential energy because if it is dropped and contacts something during its downward flight, it may do work on that something. Similarly, a compressed spring has potential energy because if it is released, it may strike something and do work on it. The general idea of potential energy as a capacity for doing work and the rearrangement of things when the potential energy is converted goes beyond these two mechanical cases. For example, gasoline has potential energy because if

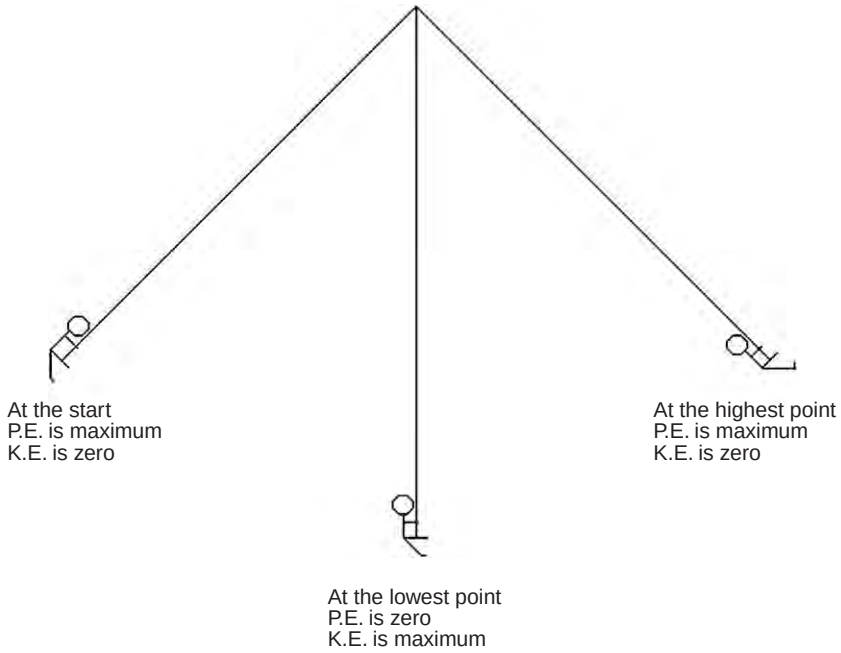


FIGURE 3 Illustration showing that, to the extent that friction can be ignored, the mechanical energy of a child in a swing is unchanged at any moment. P.E., potential energy; K.E., kinetic energy.

a gasoline vapor/air mixture is ignited in the cylinder of an internal combustion engine, the expanding gas pushes against a piston and does work on it. The root of the potential energy is found in the molecules from which the gasoline is formed. Energy is extracted from these molecules when their atoms are rearranged into different molecules during the combustion process. Similarly, the uranium fuel in a nuclear reactor has potential energy that is extracted from rearrangement of neutrons and protons through nuclear fission reactions. The energy produced from the nuclear reactions is then used to produce steam, which pushes against the blades of a turbine, producing rotational kinetic energy.

An industrial society finds myriad uses for kinetic energy. Pistons moving up and down in an internal combustion engine have kinetic energy. Wheels rotating on an automobile or a truck have kinetic energy. Water falling from atop a dam has kinetic energy. Around a home or in a factory, there are numerous motors providing rotational kinetic energy for a multitude of purposes. In all of these examples, the kinetic energy evolves from a conversion from some other form of energy. The force behind the movement of a piston comes from an expanding gas produced by the ignition of a gasoline/air mixture. The kinetic energy acquired by water rushing to the bottom of a dam comes from a conversion of gravitational potential energy. Motors convert electric energy to rotational kinetic energy. Whether large or small, and whether simple or complex, converters producing kinetic energy all subscribe to the principle of conservation of energy. Each one converts energy into some form of kinetic energy regarded as useful, and each one diverts energy that is not immediately useful and might never be useful. Because energy is diverted, the efficiency defined as

$$\text{efficiency} = \frac{\text{useful energy}}{\text{total energy converted}}$$

can never be 100%.

9. APPLICATIONS OF MECHANICAL ENERGY

9.1 Pile Driver

A pile is a large metal or wooden post driven into the ground. A pile driver (Fig. 4) lifts a rather massive object (hammer) above the pile and drops it. Each drop of the hammer drives the pile farther into the ground until the required depth is reached. Lifting the hammer requires mechanical work. In so doing, the hammer acquires potential energy. When released, the hammer gradually loses potential energy

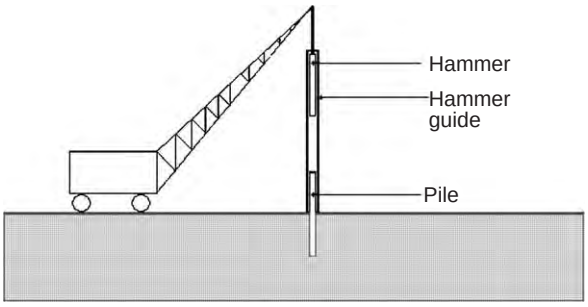


FIGURE 4 Basic components of a pile driver.

and gains kinetic energy. Work is done on the pile during contact, causing the kinetic energy to decline, eventually to zero.

9.2 Hydroelectric System

Whenever an object has gravitational potential energy, some agent had to do work on the object. In the case of a pile driver, there is a hoist of some sort. Water atop a dam has potential energy. Nature provides the means to fill the water reservoir through evaporation followed by condensation and rain. To see the energy transformations involved when the water falls from the top of the dam onto the blades of a turbine, it helps to visualize the motion of an object having boundaries. Accordingly, imagine the motion of a cube of water moving with the stream (Fig. 5). At the top of the dam, the cube of water of mass (m) and speed (V) has kinetic energy $\left(\frac{1}{2}mV^2\right)$, and potential energy ($mgh$) due to its position at the top of the dam. As the gravitational force pulls the cube from the top of the dam to the bottom, it loses potential energy but gains kinetic energy. At the bottom of the dam, all of the potential energy (mgh) has been converted to kinetic energy $\left(\frac{1}{2}mv^2\right)$. The total kinetic energy now includes its kinetic energy before falling plus the kinetic energy gained by falling:

$$\tfrac{1}{2}mv^2 = \tfrac{1}{2}mV^2 + mgh. \qquad (7)$$

The cube of water arrives at the paddle wheel with energy $\left(\frac{1}{2}mv^2\right)$. The force of the cube on the paddle wheel causes work to be done on the wheel. Accordingly, the paddle wheel rotates, acquiring energy (E) while the cube of water loses kinetic energy. The kinetic energy of the cube after it passes by the paddle wheel (K) equals the kinetic energy it had before striking the paddle wheel minus the energy acquired by the paddle wheel:

$$K = \tfrac{1}{2}mv^2 - E \qquad (8)$$

In a hydroelectric power plant, the turbine would drive an electric generator and the energy (E) would be used for generating electricity.

9.3 Pumped Storage Hydroelectric System

The demand for electric energy by a community varies with the time of day and with the time of year. An electric power utility must be prepared to meet these demands. This poses an engineering problem because there is no practical method of storing electric energy on a scale that will meet the demands of a large community. To meet short-term increases in demand, electric utilities employ generators that can be turned on and off on short notice. For example, they may use a gas turbine similar to a jet plane engine to drive a generator. Another scheme is to use a pumped storage hydroelectric system. Such a system does not rely on nature to replenish the water in a reservoir but rather uses electrically run pumps. Importantly, the system can generate electricity on short notice. A schematic diagram of the system is shown in Fig. 6. Water is forced to an elevated reservoir by a motor-driven turbine. The water in the reservoir has gravitational potential energy by virtue of the work done on it. When electricity is needed, the water is allowed to flow downward into the turbine that drives the motor, which now functions as an electric generator. The energy required to elevate the water is never completely recovered in the process. Nevertheless, the system is economical because the reservoir can be filled when electric energy demands and costs are low. It is also possible to have a system in which water flows from ground level to underground turboelectric generators. In this case, work has to be done to restore the water to ground level.

9.4 Warning

Steel is a very hard metal that behaves like a very stiff spring when compressed or stretched. Bolting two steel plates together compresses each plate to some extent. Although the compression may be small, the elastic potential energy $\left(\frac{1}{2}k\,x^2\right)$ in the plates can be large because the spring constant (k) is large. If the nut on the bolt holding the plates together is released gradually, the elastic potential energy declines gradually. But if for some reason the nut cannot be turned and the bolt is freed by chiseling the bolt in two, the elastic potential may be released suddenly, causing the nut to spring away. So violent is the separation that a person may be seriously injured if

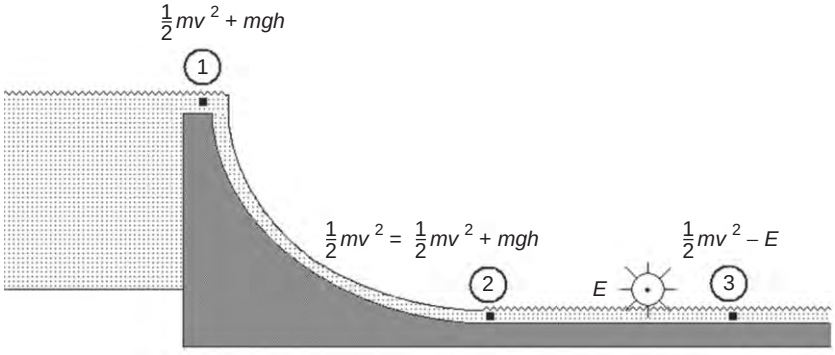


FIGURE 5 Energy transformations involved in water flowing over a dam. (Position 1) At the top of the dam, the cube of water has kinetic energy $(\frac{1}{2}mv^2)$ due to its motion and potential energy (mgh) due to its position above the bottom of the dam. (Position 2) At the bottom of the dam, all of the potential energy (mgh) has been converted to kinetic energy. The energy of the cube of water is kinetic $(\frac{1}{2}mv^2)$ and includes the kinetic energy it had at the top of the dam plus the kinetic energy acquired by falling over the dam. (Position 3) Passing by the paddle wheel, the cube of water transfers energy (E) to the wheel. Its energy after passing the wheel is still kinetic and is equal to its energy before impinging on the wheel $(\frac{1}{2}mv^2)$ minus the energy (E) imparted to the wheel.

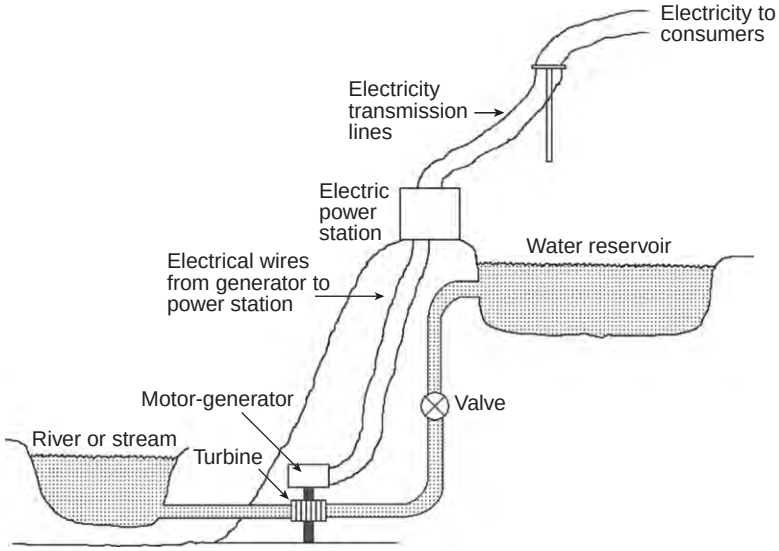


FIGURE 6 Principle of a pumped storage hydroelectric system. Water in an elevated reservoir has potential energy as a result of being pumped from a river or stream. Electricity is generated when the water flows through a turbine–generator combination on its way back to the river or stream.

he or she has the misfortune of encountering the nut. Such incidences have actually happened and have prompted warnings to workmen who may have to free a frozen nut.

SEE ALSO THE FOLLOWING ARTICLES

Conservation of Energy Concept, History of • Electrical Energy and Power • Energy in the History and Philosophy of Science • Forms and Measurement of Energy • Heat Transfer • Hydropower Technology • Storage of Energy, Overview • Thermodynamic Sciences, History of • Thermodynamics, Laws of • Work, Power, and Energy

Further Reading

Hobson, A. (2002). "Physics: Concepts and Connections," 3rd ed. Prentice Hall, Upper Saddle River, NJ.

Priest, J. (2000). "Energy: Principles, Problems, Alternatives," 5th ed. Kendall/Hunt, Dubuque, IA.

Serway, R. A., and Beichner, R. J. (2000). "Physics for Scientists and Engineers." Brooks/Cole, Pacific Grove, CA.

Serway, R. A., and Faughn, J. S. (1999). "College Physics." Brooks/Cole, Pacific Grove, CA.

Media Portrayals of Energy

JAMES SHANAHAN
Cornell University
Ithaca, New York, United States

1. Introduction
2. Media Coverage of Energy: History
3. Public Opinion about Energy
4. Media Effects on Public Opinion and Policy

Glossary

environmentalism A social movement of the 20th century focusing on the threats to human health posed by a variety of pollutants. Includes a broad swath of concerns, such as air and water pollution, climate change, wilderness protection, endangered species protection, and sustainability.

mass media The system of communication by which large, commercial organizations produce entertainment and news content for large, heterogeneous audiences. The term usually includes television, radio, newspapers/magazines, popular music, publishing, and films.

muckraker A name for a type of investigative newspaper or magazine journalist of the early 20th century. Muckrakers examined social problems, working conditions, environmental pollution, and other threats associated with the activity of large industry.

public opinion A collective indication or measurement of how the public feels about given issues. Usually measured using scientific random sampling techniques and opinion questionnaires or surveys. The term also refers to a more general, impressionistic perception of public sentiment on a given issue.

The mass media, especially newspapers, magazines, television, and movies, play a role in the portrayal of energy issues. Since the energy crisis of 1973, the attention of scholars has been turned to how media shape, frame, and influence audience perceptions of energy and the environment. During the period following the energy crisis, especially throughout the 1980s and early 1990s, issues of energy conservation were frequent topics of scholars' attention. Oil crises and the perceived dangers of nuclear power sensitized society to the potential importance of conserving energy. These issues were also frequently mentioned in the media. In these earlier days of the environmental movement, it was believed that mass media could be prominent tools to encourage conservation of energy, through adoption of energy-saving measures in the home, driving smaller and more fuel-efficient vehicles, and promoting alternative energy source use. The attitudes of this period highlighted the fact that media can play an important role in people's individual choices. Influences such as advertising, journalism, television programs, and other media need to be considered when determining how people make choices in energy use behavior.

1. INTRODUCTION

In general, the influence of media on energy use has been considered within the wider field of study of media effects on environmental concerns. This field of research has shown that media do have both direct and indirect impacts on how people develop conceptions of the environment. However, these effects are not always consistent; they do not always move in the same direction. Thus, it has been shown repeatedly that information derived from the media is an important factor in how people develop awareness about the environment. Because people directly experience only a small portion of the environment, knowledge about global environmental issues and problems must come from mediated sources. People who are more attentive to newspapers and other journalistic sources are more likely to be informed and knowledgeable about environmental issues and problems. Those who are particularly attentive to environment-specific media sources (such as environmental magazines or television programs) are better informed about environmental issues and are more concerned about them. On the

 © 2004 Elsevier Inc. All rights reserved.

other hand, general media attention (especially to entertainment media) is associated with less knowledge and concern about environmental issues. Frequent viewers of television programs, for instance, are less concerned about environmental issues in general. Further, they are more supportive of materialist ideologies that tend to promote higher energy consumption.

These countervailing examples show that it can be dangerous to lump all "media" together under one umbrella. Sometimes media outlets promote conservation of energy and concern for the environment; specific campaigns aimed at reducing energy consumption are often studied with this as the goal. Meanwhile, a more general attitude of promotion of consumption seems to guide the overall media atmosphere, which is built on advertising-supported ideologies of material growth. Given these tendencies, it is necessary to try to disentangle the various media effects, in order to get some idea about the key questions of interest. The main question that is the focus in this article concerns how the media influence public conceptions of energy within an environmental framework.

2. *MEDIA COVERAGE OF ENERGY: HISTORY*

Before discussing current issues in media coverage of energy, some of the main historical aspects of how media have treated the energy issue are summarized. This discussion will be contexted within the wider theme of how the media treat the environment.

It is a common conception that environmentalism in the public mind and in the media frame is a phenomenon of the latter 20th century. However, there are clear examples of how the mass media provided a stage for a discussion of environment and energy issues well back into the 19th century. Preservationists such as John Muir, for example, were using the national media in their attempts to set aside lands for national parks. Muir's articles in magazines such as *Century* attracted the attention of Eastern elite audiences to issues that were mainly situated in the West. These strategies helped develop national attention for environmental issues in the context of a nationally emerging mass media structure. Though a national mass media system had emerged as early as the 1830s, especially with the development of mass newspapers in major cities served by wire services, this structure was most firmly cemented after the Civil War. How did this affect coverage of energy issues?

2.1 Coal and Oil

Attitudes about sources of energy such as coal and oil are recoverable from publications of the 19th century, which followed the development of new energy sources with close attention. However, as with media coverage of any major issue, it is not possible to determine a single prevailing attitude toward an energy source such as coal or oil. Certainly, many of the major media sources adhered to a generally nationalistic ideology, oriented toward goals such as economic progress, westward expansion, and even imperial aspirations. Within this context, fossil fuel sources obviously played an important role in the development of America as a world power. Thus, one major current of the media mainstream was oriented toward a view of fossil fuels as expanding the possibilities for America, without worrying too much about potential negative consequences.

But worrisome themes also appeared in the press. The economic booms and busts of oil discoveries were linked to boom-and-bust media cycles: high-hoped optimism followed by cautionary tales was a not uncommon cycle in the press. Stories told of those who made and lost fortunes were as characteristic of news coverage in the oil years as they were in the days following the Internet bubble. Oil was often presented as something ridiculously easy to obtain and sell, violating all previous rules of the marketplace. After a boom had gone bust, the national dailies turned to tales of woe and ghost towns in their coverage.

As oil became an industry, it attracted a different kind of attention, that of the "muckrakers." Muckrakers were the forerunners of today's "investigative journalist." Their reformist spirit led them to attack industry on issues ranging from child labor to food safety to environmental issues. Ida Tarbell (see Fig. 1), writing in the pages of *McClure's,* a national magazine, attacked the Standard Oil Company for its monopolistic and predatory practices. Her family life in the Pennsylvania oil patch had been disrupted by Rockefeller's South Improvement scheme. Published serially from 1902 to 1904, Tarbell's "History of Standard Oil" is considered a founding piece of investigative journalism. But not all media controversies resulted in immediate change. Concerns about ethyl leaded gasoline were covered fairly extensively, especially in the New York City area, from 1924 to 1926. Worker deaths from lead

FIGURE 1 Ida Tarbell (1857–1944).

poisoning led crusading scientists to point out that lead additives could be harmful to the public at large, due to risks from lead deposition in exhaust. Although a committee was formed by the Surgeon General to look at the issue, no action was taken, even with significant attention from newspapers such as the *New York World* and *The New York Times*. It was not until the 1970s that the Environmental Protection Agency (EPA) would take action to remove lead from gasoline. Still other themes were developed earlier than one might expect. Recognizing American dependence on oil led to journalistic calls for conservation and development of renewable energy sources long before the onset of the environmental era in the 1970s. The concern was not environmental in nature, but stemmed from a worry about maintaining America's preeminence despite dependence on oil.

Typically, media attention to environmental issues is driven by both journalistic cycles and events. Events are spectacular disasters or phenomena with widespread impact. A "killer smog" in Donora, Pennsylvania in 1948 drew enough media attention to catalyze the development of smoke-abatement programs in cities around the United States. This was an early precursor to the more far-reaching programs of air pollution control that were eventually enacted in the 1970s. Indeed, air pollution themes had received sporadic attention throughout the 20th century, but the atmospheric impacts of fossil fuel consumption began receiving a closer look in the media in the 1960s. This peaked with the passage of national environmental legislation (the National Environmental Protection Act) in 1970.

Fossil fuels and American industry in general came under further media scrutiny in the 1960s. Rachel Carson's *Silent Spring* (1962), though not focused directly on fuels or energy sources, drew attention to the chemical industry and its promotion of dangerous chemicals such as dichlorodiphenyl trichloethane (DDT). *Silent Spring* sensitized Americans to risks both known and unknown in and about the general environment. Carson arguably inaugurated the age of environmentalism, in which the media played an important role in turning people's attention to a variety of environmental issues, with fossil fuel consumption at or near the top of the list of problems. Carson's book, serialized in the *New Yorker* magazine, received wide media attention, building on the earlier model for journalistic crusading that had been successfully tested by both John Muir and Ida Tarbell.

This touched off a frenzy of popular publishing activity focused on energy consumption and environmental issues. Books such as *The Waste Makers* (Vance Packard, 1960), *The Quiet Crisis* (Stewart Udall, 1963), *The Population Bomb* (Paul Ehrlich, 1968), *The Closing Circle* (Barry Commoner, 1971), *The Limits to Growth* (Donella Meadows *et al*., 1972), *The Poverty of Power* (Barry Commoner, 1975), *The Eco-Spasm Report* (Alvin Toffler, 1975), and *The Fate of the Earth* (Jonathan Schell, 1985) are examples of popular publications that made some impact on public sensibilities and opinions about energy consumption problems. These efforts were strengthened by press attention to the oil crises of the 1970s, which caused a great deal of concern about energy dependence. With energy issues bouncing around the press throughout the 1970s, the issue was often at the top of the public agenda, particularly when gasoline prices were rising. With a cast of villainous characters [the Organization of Petroleum Exporting Countries (OPEC) oil sheiks] and an issue striking at the heart of the American economy, foreign oil was the biggest news item in 1973, and to a lesser extent in 1979. However, specific events also masked cycles of press concern about energy issues. These cycles are generated not by specific external motivating factors, but by tendencies to adhere to narrative structure embedded within journalistic practice. For instance, in the late 1980s and early 1990s, journalists in both newspapers and magazines

and on television turned their attention to a broad range of environmental issues. The attention was initially brought about by an extremely hot and dry summer in 1988. Scientists' claims about global warming suddenly found a foothold in the media, when James Hansen of the National Aeronautics and Space Administration (NASA) claimed that global warming could be definitively tied to human agency. Bill McKibben published a widely read book, *The End of Nature* (1989), which argued that climate change meant that human influence was suddenly spread throughout the entire environment, and there was no more pristine "nature" to be found. The book was also serialized in the *New Yorker*.

Suddenly, the press was covering a wide range of issues, from climate change to acid rain to radon gas to ozone depletion. The years 1988–1991 represented an unprecedented period of public concern and attention to environmental issues. The Exxon *Valdez* accident (1991) was yet another event that encouraged even more environmental debate in the media. It was during these years that the public and policy debate began to give more credence to the idea that the media could be a positive force for social change, especially in the big, important areas such as climate change. Environmental concern was so strongly present in the mediated discourse of the United States that companies such as McDonald's and Burger King adopted new forms of packaging to counter charges that styrofoam hamburger containers consumed too much energy in production and were persistent in the waste stream. Even entertainment television began to manifest signs of a green consciousness. Hollywood celebrities increasingly focused on environmental issues such as rain forest conservation. Public relations agencies with the purpose of placing environmental references in television programs and feature movies came into being. Movies with explicitly and implicitly environmental themes were more common. *Dances With Wolves* (1990) was a prototypical environmentally themed movie of the period, focusing on the connection of the Sioux Indians to the land. Movies with more directly energy-related themes include *The China Syndrome* (1979), about a nuclear disaster uncovered by an inquiring reporter; the *Mad Max* series of movies (1989–2000) about a postapocalyptic Australian outback gripped in wars for the dwindling supply of oil; *Waterworld* (1999), a story about a post-climate-change world that has been inundated with water; and *Ferngully* (1992), a children's story about a destroyed forest that threatens a race of environmentally conscious sprites. On television, children's shows such as *Captain Planet* (1990–1996) often focused on themes of energy conservation.

After 1991, however, general media attention began to subside (see Fig. 2). Vice President Al Gore published *Earth in the Balance* in 1992, but he missed the peak of media and public excitement. The eruption of Mount Pinatubo in the Philippines led to cooler temperatures globally, which also seemed to cool press interest. Indeed, research has shown that news attention to climate change issues varies with the actual temperature, such that journalists are less likely to cover climate change during cooler periods. In any case, coverage for all environmental issues was declining throughout the mid-1990s. The pattern was following that predicted by social scientist Anthony Downs, who argued that press attention to environmental issues would always be cyclical due to the inherent nature of the issues. Because environmental issues are difficult to solve, public attention will fade when the costs of achieving environmental gains are calculated. Also, the issues are seen as neither interesting nor fascinating from a journalistic standpoint. Other theorists also pointed out that the news agenda has a carrying capacity. When one issue rises in salience, others decline. The first Gulf War (1991), even though it had an important energy motivation, turned attention away from the environmental aspects of energy issues.

By 1995, environmental issues were neither as much in the media nor in the scope of public attention. The celebration of "Earth Day," which in 1990 had received a lot of media attention, went barely noticed in 1995. A strain of antienvironmental

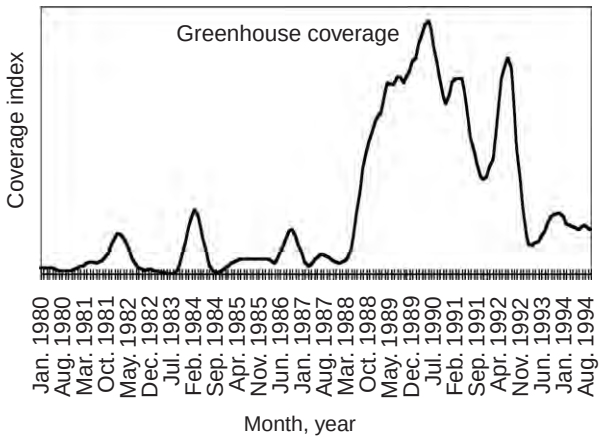


FIGURE 2 Coverage of climate change reported in *The New York Times*. From Shanahan and McComas (1999).

thinking was also becoming evident in the works of authors such as Julian Simon, who preached environmental optimism in works such as *The Ultimate Resource*. Readers of popular literature could see these ideas in books such as Greg Easterbrook's *A Moment on the Earth* (1995). Also, the energy industry was not at all inactive. Industry public relations groups such as the Western Fuels Association had engaged in active public communication campaigns to discredit the science associated with predictions about climate change. To some extent, these campaigns were successful, and public opinion eventually turned away from concern about climate change. Although the scientific community was not dissuaded in its opinions about climate change, lack of public resolve and media attention has made it difficult to develop stronger U.S. policies on climate change.

Since the mid-1990s, energy issues have struggled to receive major attention. Cheap oil has fueled the popularity of sport utility vehicles and trucks, which have been advertised heavily by car producers. Indeed, the reliance of the media on the automobile and energy-producing industries, through advertising revenues, is an important factor to consider when examining coverage of energy issues. Although the major media certainly did not shy away from criticism during disasters such as the Exxon *Valdez* incident, some critics have pointed out that mainstream news organizations are too reliant on funds from "Big Oil," or from Detroit.

Energy consumption, thus, is not currently viewed as a major issue on the American agenda, with the minor exception of concerns about dependence on oil and its relation to issues of susceptibility to terrorism. September 11, 2001 and the Iraq War (2003) activated certain sectors of society to begin thinking again about U.S. reliance on foreign oil. Slogans such as "No blood for oil" encapsulate the thinking of the social protest sectors that have sought to mobilize wider opposition to U.S. foreign policy in the Middle East. But the first Gulf War and the Iraq War were both very popular. The U.S. media sector assisted the war efforts in both cases by focusing on issues of patriotism and nationalism; their efforts were rewarded with strong opinion support. Energy issues were widely disregarded in both cases, except insofar as U.S. forces achieved successes in "protecting" energy resources such as the Iraqi oil fields. The media, having passed through a period of concern about environmental issues, revealed little of that concern in recent times.

2.2 Nuclear Energy

Nuclear energy, compared with traditional energy sources such as coal or oil, has received much different treatment in the media. Nuclear energy has always been portrayed in a more bipolar fashion, as offering either fantastic possibilities or horrifying risks. Coverage has tended to swing between these poles without resting much in the middle. As a consequence, public opinion has been more mistrustful.

After the detonation of atomic bombs at the end of World War II, there were attempts to promote peaceful uses of atomic energy. It would be fair to say that pro-atomic energy campaigns played a large role in shaping the media atmosphere. During the 1950s, the government encouraged a positive strain of thinking about nuclear energy that permeated the media. Dwight Eisenhower's "Atoms for Peace" speech of 1953 is one example of an attempt to create a rubric for public communication about nuclear energy that could escape from the dread perceptions engendered by its military history. But the fear brought about by the Cold War was made manifest in media coverage of various types. It was difficult for nuclear power to escape negative associations. Although many nuclear reactors were built and used, the attitude of the entertainment media in the 1960s toward nuclear power was focused either on science fiction or on doomsday scenarios. Movies such as *Failsafe* (1964: "It will have you sitting on the brink of eternity!") both informed and characterized the public fascination with and fear of nuclear weapons; also notable was Stanley Kubrick's *Dr. Strangelove* (1964), a satire of nuclear militarism based on *Failsafe*. Transfer of public fear from nuclear weapons to nuclear energy was probably inevitable. At a more quotidian level, nuclear energy had been portrayed as the source of horrific mutations and science-fiction accidents in more movies than can be recounted here [though they include *Attack of the Crab Monsters* (1957), *The Beast From 20,000 Fathoms* (1953), *The Creation of the Humanoids* (1962), *Godzilla, King of the Monsters!* (1956) and *Them!* (1954)]. And many movies (as well as television programs such as *The Twilight Zone* and *The Outer Limits*) used post-nuclear-holocaust scenarios as a standard setting or as a formulaic plot device.

Mediated public debate about nuclear energy as a power source emerged more strongly in the media in the 1970s. One major factor was the emergence of antinuclear activism. Media coverage of environmental activism can be a two-edged sword. Oftentimes,

specific stories about environmental issues present the environmental perspective as out of the mainstream. Activists and their causes are not at the core of public opinion concern. However, the salience derived from media coverage has been very beneficial to environmental causes. The nuclear activism issue is an excellent case in point.

Risk studies have shown that the risk from nuclear energy is perceived as an "imposed" risk. People are more likely to express concern about risks that they do not normally assume in the course of their daily lives. When such risks are covered heavily in the media, it may make little difference that actual risk ratios are relatively low. Such was certainly the case for nuclear power, which experienced negative attention from all sides, and never recovered from negative associations with the atomic bomb. The key media event that affected perceptions of nuclear power was Three Mile Island. Walter Cronkite ("the most trusted man in America") and other mainstream media were arguably quite alarmist in their coverage, making connections to the recently released movie *The China Syndrome*. Supporters of nuclear power have been extremely critical of the media model that was established with this incident; they blame it for the fact that citizens estimate risks from nuclear power as much higher than risks known in fact to be more dangerous, such as cigarette smoking.

With Chernobyl, a similar dynamic was introduced, although the media took a different tack given that the accident occurred within the Soviet Union. Studies showed that news agendas were dominated by the accident in the U.S. and especially in Europe. Though some coverage emphasized the relative safety of U.S. reactors compared to Soviet ones, this incident added yet another negative to the list of achievements of nuclear power. In the time between the two accidents at Three Mile Island and Chernobyl, the broadcast by the American Broadcasting Corporation (ABC) of *The Day After* (1983) was arguably the biggest entertainment media event of the 1980s that dealt with nuclear energy. It told the story of a U.S. city after a nuclear attack. It was watched by half of the U.S. population, and was seen as a feature film in many foreign countries. The Reagan White House was concerned about its antinuclear bias. The broadcast's effects on public opinion and knowledge about nuclear power were also studied, with inconclusive results, but undoubtedly few positive effects accrued for the nuclear energy industry, which was increasingly under fire.

3. PUBLIC OPINION ABOUT ENERGY

Public opinion polls have frequently asked citizens to comment on their thoughts, feelings, fears, and attitudes about energy sources. In terms of attitudes about energy in general, opinion and concern have varied. Through the 1970s and into the 1980s, in a variety of polls, about 80% of U.S. respondents said that the U.S. energy situation was "serious." This reflected general concern about energy prices and reliance on foreign oil. When assessing blame for the energy situation, respondents have been most likely to finger oil companies, although Arab states were often blamed as well. Oil companies, as is true of most energy institutions, receive very low levels of public support (nuclear power institutions receive the lowest ratings of all) in public opinion polls.

In the late 1980s and early 1990s, the U.S. moved beyond a debate on the energy supply, which was largely not an issue anymore. In the late 1980s, more consumers turned toward larger vehicles, and questions about greenhouse emissions and the role of fossil fuels in climate change moved to the front of the public agenda. These issues were foregrounded by the media, resulting in unprecedented levels of public concern about fossil fuels into the early 1990s. However, these concerns faded with the onset of the Gulf War and the economic downturn that followed. Coverage of climate change and other issues followed a classic cyclical pattern, as was noted above. How did this affect public opinion? One series of data showed that public concern about governmental spending on environmental issues was strongest during the time of peak media coverage. Concern was stronger among those who were heavier consumers of news media. In the late 1990s, concern about fossil fuels tended to drop, along with general concern about environmental issues. Although major international conferences on climate change, such as those held in Kyoto in 1997, did provide opportunity for new spates of energy-related coverage, these often focused on U.S. refusal to go along with world climate policy.

In general, as already noted, compared with fossil fuels, nuclear energy has sparked more negative feelings. In September of 1945, 93% of respondents to a National Opinion Research Center poll felt that there was a real danger of most people living on Earth being killed by nuclear bombs if a new world war broke out. But still, citizens generally strongly felt that it was acceptable to have used the bomb in 1945; they were also generally mixed on whether

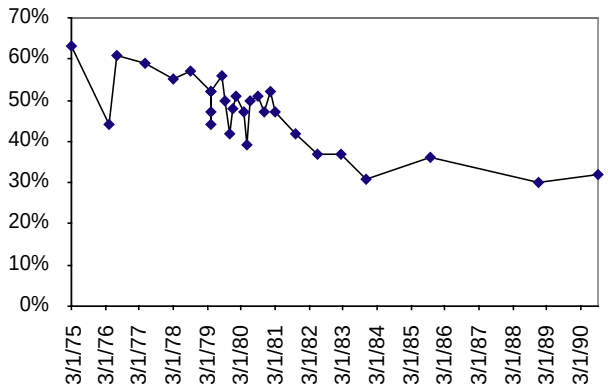


FIGURE 3 Percentage of poll respondents favoring building more nuclear plants in the United States (Harris poll data).

splitting the atom would turn out to be a net positive or negative for society.

In the 1970s, after the first oil shock, U.S. citizens felt that it was extremely important to develop alternative sources of energy. A 1976 Harris poll found that 80% of respondents favored developing nuclear energy sources, along with other sources such as solar energy and more oil exploration. But then, among other influences, news about nuclear power accidents in Three Mile Island and Chernobyl had effects on public opinion, though not initially as drastic as one might expect. Some researchers argued that nuclear power could "rebound" from the negative effects of such disasters. But over time, this proved not to be the case, as the public moved consistently toward negative views of nuclear power in the late 1980s and 1990s (Fig. 3). In addition to the perceived dangers evident in the disasters, other factors affecting support were antinuclear activism (often media based) and the perceived high expense associated with constructing nuclear plants. Also, the fictional portrayals mentioned above (*The China Syndrome* and *The Day After*) probably played a role.

4. MEDIA EFFECTS ON PUBLIC OPINION AND POLICY

To what degree do the media affect public attitudes about energy? Although many studies have examined relationships between media use and environmental attitudes, no clear finding has emerged. On a case-by-case basis, it is often difficult to show how specific media messages change public opinion, apart from extremely salient mass spectacles such as *The Day After.* For instance, studies of the effects of messages such as President Carter's "moral equivalent of war" speech on energy conservation showed minimal effects in terms of citizens actually conserving. Studies that attempted to tie proconservation media coverage or campaigns to specific energy consumption habits (such as lowering the thermostat) have also not shown much effect. At the micro level, it is clear that there are a myriad of forces that determine individual energy decisions, among which the media may play only a small role.

At a macro level, however, it is difficult to imagine a more powerful institution than the media for setting the social agenda and defining terms. Repeated exposure to messages favoring increased material consumption does play a role in actual consumption, as studies have shown. Viewers who spend a lot of time watching television ("heavy" viewers) are more likely to believe that the world is more affluent than it really is, they desire more material possessions, and they are less likely to be concerned about environmental issues. Such television viewers are even more likely to answer opinion questions about material wealth more quickly than "light" viewers; the implication drawn by psychologists is that heavy viewers work from a store of media-cultivated images that allow them to heuristically view the world as wealthier than it actually is. Media images also play a major role in determining people's risk perceptions and fears, particularly about nuclear energy. Thus, we can look at media effects at two levels. Environmental activists, energy producers, and other active participants in social policy focus most often on the day-to-day slate of messages that a citizen receives. These messages convey a welter of often conflicting impressions about the state of energy, what the polls refer to as the "energy situation." "Actors" with significant resources can play a major role in attempting to tip the balance of these conflicting messages. Issue-advertising models along the lines of those developed in the 1970s by Herbert Schmerz of Mobil Oil Company are one way that energy producers attempt to influence the policy discourse. As well, since the 1990s, most major energy producers have added environmental divisions to their corporate structure, in recognition of the increasing public relations importance of environmental issues. On the other hand, the incredible development and growth of environmental activist organizations have balanced out the claims and arguments of energy producers. Groups such as Greenpeace have been persistent and successful in developing media strategies. Groups with specific media missions include the Environmental Media Association, which strives to

make Hollywood productions more environmentally relevant.

Still, the fundamental tone of the media is oriented toward growth, toward what some sociologists called the "Dominant Social Paradigm." This paradigm values economic growth and encourages confidence in the ability of new technologies to fuel such growth. Scholars who have examined media from a broad cultural perspective have been most impressed with contributions of advertising to the creation and maintenance of this paradigm. Even though sociologists have detected the outlines of an emerging "New Environmental Paradigm," there is little doubt that the media system is still fundamentally structured to encourage growth in consumption.

Bill McKibben summed up these views in his *The Age of Missing Information* (1992), which dealt with a comparison of world views obtainable in the media versus those obtained in the real world. To McKibben, the power of television was not its specific messages, but rather the dizzying amount of content, most of which was focused on material consumption in one form or another. He complemented the views of scholars such as George Gerbner, who argued that the true effect of the media is the extent to which it cultivates perceptions of reality. The efforts of media researchers to document that the media do have an influence on perceptions of reality have not been in vain. In the energy sphere, consumption of media is tied to materialism, support for the dominant social paradigm, and less concern about environmental issues such as energy conservation.

Thus, the media have not changed much since the days of early coverage of oil booms. In the media, energy is the fuel of our technologically materialist culture. At the same time, media will sporadically play a watchdog role in suggesting environmental safeguards as far as our use of energy is concerned. Thus, even despite the various disasters, risks, problems, and accidents associated with fossil fuel consumption, the media do not fundamentally question our reliance on these fuels. Perhaps this is because everyone realizes we have no real alternative. On the other hand, with nuclear energy, we see the veto power of the media in action. In that nuclear power has not been an essential energy source, it has also been fair game for media criticism.

SEE ALSO THE FOLLOWING ARTICLES

Climate Change and Energy, Overview • Conservation Measures for Energy, History of • Consumption, Energy, and the Environment • Environmental Change and Energy • Geopolitics of Energy • Global Energy Use: Status and Trends • Lifestyles and Energy • Oil Crises, Historical Perspective • Public Reaction to Energy, Overview • Public Reaction to Nuclear Power Siting and Disposal • Public Reaction to Renewable Energy Sources and Systems

Further Reading

Allen, C., and Weber, J. (1983). How Presidential media use affects individuals' beliefs about conservation. *Journalism Q.* **68**(1), 98–110.

de Boer, C. (1977). The polls: Nuclear energy. *Public Opin. Q.* **41**(3), 402–411.

Downs, A. (1972). Up and down with ecology—The "issue attention cycle." *Public Interest* **28**, 38–50.

Erskine, H. (1963). The polls: Atomic weapons and nuclear energy. *Public Opin. Q.* **27**(2), 155–190.

Farhar, B. (1994). Trends: Public opinion about energy (in the polls). *Public Opin. Q.* **58**(4), 603–632.

McKibben, B. (1989). "The End of Nature." Random House, New York.

McKibben, B. (1992). "The Age of Missing Information." Random House, New York.

Neuzil, M., and Kovarik, W. (1996). Conflict management and scientific controversy. *In* "Mass Media and Environmental Conflict: America's Green Crusades," Chap. 6. Sage, Thousand Oaks, California. The importance of dramatic events. *Ibid*., Chap. 7.

Rosa, E., and Dunlap, R. (1994). Poll trends: Nuclear power: Three decades of public opinion. *Public Opin. Q.* **58**(2), 295–324.

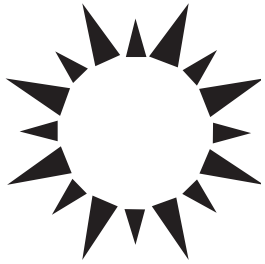
Shanahan, J., and McComas, K. (1997). Television's portrayal of the environment: 1991–1995. *Journalism Mass Commun. Q.* **74**(1), 147–159.

Shanahan, J., and McComas, K. (1999). "Nature Stories." Hampton Press, Cresskill, New Jersey.

Shanahan, J., Morgan, M., and Stenbjerre, M. (1997). Green or brown? Television's cultivation of environmental concern. *J. Broadcast. Electron. Media* **41**, 250–268.

Tarbell, I. (1902). The history of the Standard Oil Company. *McClure's Mag.* **20**(1), 3–17.

Microtechnology, Energy Applications of



RICHARD B. PETERSON
Oregon State University
Corvallis, Oregon, United States

1. **Introduction and Overview**
2. **Unit Operations**
3. **Systems**
4. **Materials, Fabrication, and Costs**

Glossary

coefficient of performance (COP) A figure of merit for cooling systems; it is defined here as the amount of thermal energy removed from a cooled space divided by the amount of work or heat supplied to the cooler to accomplish the heat removal.

fuel cell An electrochemical device for directly generating electricity by combining a fuel, such as hydrogen, with oxygen to form a reaction product, such as water; because theoretical performance is not tied to the thermodynamic Carnot efficiency, higher chemical-to-electrical energy conversion can occur.

fuel processor A chemical reactor specifically for converting a fuel from a complex mixture of hydrocarbons to a less complex, and often pure, form of usable fuel (e.g., hydrogen).

heat exchanger A device for transferring thermal energy from a hotter fluid to a colder one; heat exchangers come in a variety of configurations, including (but not limited to) parallel flow, counter-flow, cross-flow, compact, shell-and-tube, plate-and-frame, regenerative, recuperative, unmixed, and mixed streams.

logistics fuel Any number of liquid fuels widely used by the military for transportation and power generation; examples include JP-4, JP-6, diesel, gasoline, and kerosene.

microchannel array An array of channels with characteristic dimensions of less than 1 mm designed for conveying a heat and/or mass transfer fluid; when used as the basis for heat exchangers, boilers, and condensers, high rates of heat transfer result.

Micro Electro Mechanical Systems (MEMS) A class of devices typically made from silicon or employing it in the fabrication process; devices integrate electronics and mechanical devices onto a single substrate with typical feature sizes in the 1- to 50-μm range.

microreactor A miniaturized chemical reaction system characterized by small size (typically $<1\,cm^3$) and fabricated, at least partially, using methods of chemical or electrochemical etching, laser machining, electro discharge machining (EDM), or other microfabrication processes.

Microtechnology-Based Energy and Chemical Systems (MECS) A class of devices integrating heat and mass transfer components along with chemical reactor technology in a single integrated system; feature sizes are typically 50 μm to 1 cm.

nano-, micro-, and mesoscale Dimensional regimes for describing various levels of feature and/or device size; definitions vary, but a typical one would be nanoscale (10 nm to 1 μm), microscale (1–100 μm), and mesoscale (0.1–10.0 mm).

platelet architecture A fabrication scheme using a variety of micromachining techniques to place structural features in thin plates; several plates are then stacked together in a registered manner and bonded, producing a single part having an intricate array of embedded features.

process intensification The enhancement of heat and mass transfer rates as the characteristic length defining the process is decreased.

unit operation A process characterized by a single function; several unit operations can be combined together to produce, or generate, an end result.

The trend toward miniaturization has branched out into many fields of engineering. The area of Micro Electro Mechanical Systems (MEMS) has been established for well over a decade and is focused primarily on sensors and actuators, although other components such as gears, linkages, valves, and fluid mixers are common research and development topics. The energy area, including chemical processing, power generation, refrigeration, and heat

 © 2004 Elsevier Inc. All rights reserved.

pumping, is in the nascent stages of process and systems miniaturization. The developing area of Microtechnology-Based Energy and Chemical Systems (MECS) holds great potential for spawning new commercial sectors to the economy. The primary aim of the MECS area is to miniaturize and integrate the components necessary for advanced energy and chemical systems. Miniaturization will be especially important in areas where portability, compactness, weight, reliability, and point application are the driving considerations.

1. INTRODUCTION AND OVERVIEW

At the beginning of the 21st century, the trend toward miniaturization has branched out into many fields of engineering. The area of MEMS has been established for well over a decade and is focused primarily on sensors and actuators, although other components, such as gears, linkages, valves, and fluid mixers, are common research and development (R&D) topics. MEMS have feature sizes between 1 and 50 μm, with development in the area being heavily dependent on fabrication processes common in the electronics industry. The energy area, including chemical processing, power generation, refrigeration, and heat pumping, is in the nascent stages of process and systems miniaturization. This developing area is called by many terms—Micro Systems or Micro Systems Technology (MST) in Europe and Micro Chemical and Thermal Systems (MicroCATS) or MECS in the United States—and holds great potential for spawning new commercial sectors to the economy. The primary aim of the MECS area is to miniaturize and integrate the components necessary for advanced energy and chemical systems. Researchers in the field view this idea as a way of improving traditional energy systems while addressing the challenges of new applications. Miniaturization will be especially important in areas where portability, compactness, weight, reliability, and point application are the driving considerations.

Systems based on this technology rely on the extraordinary rates of heat and mass transfer associated with microcomponent architecture. That is, as a system's characteristic heat transfer or mass transfer path is reduced, rates of transfer increase significantly. Depending on the type of system considered and the performance metric chosen, the increase can scale as $1/L$ or $1/L^2$, where L is the characteristic length defining the transfer path. An example is the fabrication of heat exchangers from microchannel arrays where the thermal diffusion path is very small and the total surface area per unit volume is high. Early work on microchannel heat exchangers demonstrated heat transfer rates of 20 kW/cm^3, albeit under extreme flow conditions. This is the rate of heating needed by an average home on a cold day in the northern latitudes.

Other important effects are present with a reduction in scale. As with heat transfer, microscale devices have high mass diffusion rates. This leads to very fast and complete mixing in small volumes. Another feature present in microsystems is precise control over chemical reactions and biological processes by rapidly controlling system temperature. With flow-through microscale devices, temperature gradients can be very large (on the order of 100,000 K over a distance of a few microns or a time period of a few microseconds). Finally, by virtue of scale, the stress in miniaturized structural devices is lower for a given operating pressure. Therefore, it is practical to operate at higher pressures than in conventional processing situations. This can increase efficiency substantially or skew a chemical reaction to higher yields of a desirable product.

Enhancement of the primary transport processes by scale reduction is responsible for what researchers in the area call process intensification. This term captures the idea of using microscale structures in mesoscopic devices to enhance performance in small or miniaturized systems. However, size reduction cannot go on indefinitely. This is due to the increasing difficulty of transporting fluids and maintaining temperature differences within a miniaturized system. Excessive pressure drops, parasitic internal heat conduction, and mechanical friction are just a few of the existing problems when miniature energy systems are developed.

Integration is also a central issue in MECS research. To take advantage of mass production techniques leading to economical devices, it is desirable to have most if not all of the critical components of a system integrated together and fabricated simultaneously. Realistic implementation of this concept will probably rely on a number of additional steps, such as "pick and place" (common in the electronics industry), to complete the device. Aside from lowering production costs for single systems, additional capability and functionality can result from processing multiple devices at once. For

example, integrated arrays of components open up new approaches to solving challenging problems in the energy and chemical processing fields. In high-capacity applications such as chemical plants, production schemes based on massively paralleled arrays could be used. The principal advantage to this approach is the higher production efficiency associated with microscale processing and the inherent reliability of parallel architecture where the failure of one or even several individual devices would have a small effect on the overall production rate. If a few individual components required replacement, they can be isolated from the array, removed, and then replaced without shutting down the overall system.

A clear definition of an energy system must be given to elucidate the application areas for this new technology. In the context used here, the term "energy systems" generally involves energy generation, use, distribution, or conversion from one form to another. Furthermore, microtechnology-based energy systems involve the concept of microfabricated internal features providing process intensification over standard practices in common use today. Component and system size spans the mesoscale/microscale regime, as shown in Fig. 1. Nanoscale features may be important to this field during the coming decades but are beyond the scope of this discussion.

Intermediate size applications are the most likely to be commercially attractive in the short term, especially where MECS technology makes possible previously impractical activity. For instance, a MECS reformer to strip hydrogen atoms from hydrocarbon-based fuels will allow use of fuel cells in automobiles (with higher efficiencies and lower pollution). For this idea to work, high-temperature steam reformers and membrane separation are needed in small package configurations. Thermal management will also be critical to the development of practical systems. Heat loss from palm-sized reformers must be minimized through use of high-performance insulation, and heat exchangers are necessary to conserve thermal energy in the various internal gas streams. With the process intensification afforded by MECS technology, palm-sized fuel reformers for fuel cell-powered automobiles become feasible.

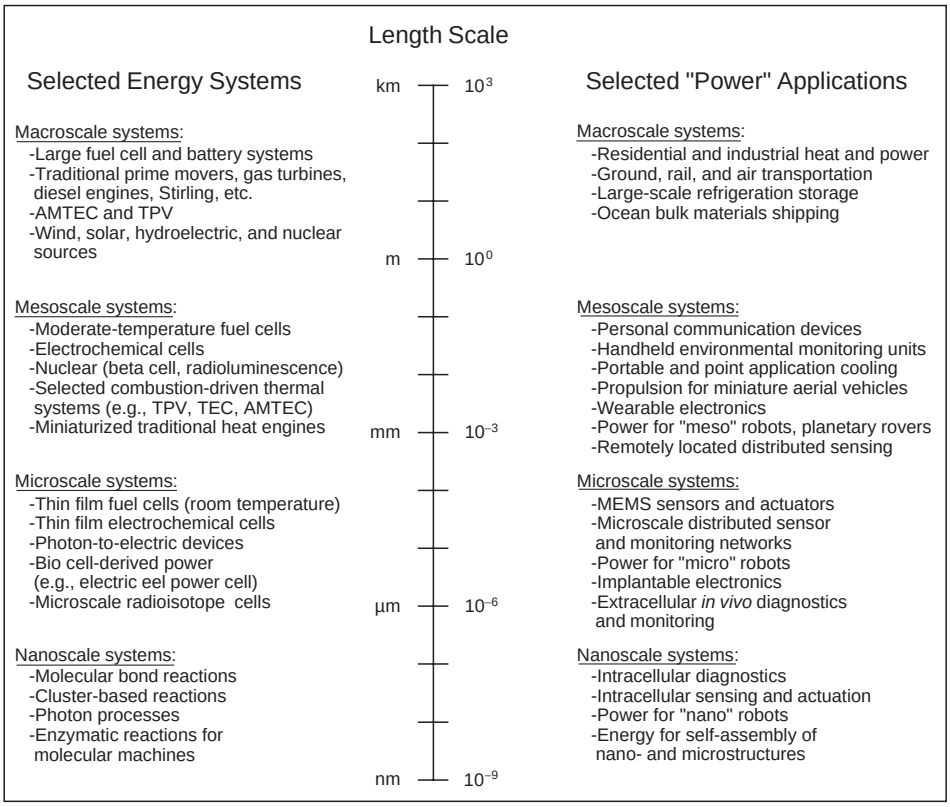


FIGURE 1 Length scales important to the energy area.

Two other areas where this technology can make an impact are small-scale power production and resource processing. As an example of the first application area, consider the chemical energy stored in a liquid hydrocarbon fuel. When burned with oxygen from the surrounding air, the heat generated can be two orders of magnitude higher than the energy contained in an electrochemical battery. If the heat can be converted to electricity with 20% or higher efficiency, stored energy densities for portable power generation can be 5 to 20 times those of current battery technology. This would require miniaturized and integrated components to achieve practical implementation in small-scale packages, but no fundamental limitations exist in developing a battery replacement operating from hydrocarbon fuels. Another application area would involve resource processing at remote locations such as on the surface of Mars. If human exploration of Mars is to take place, it makes sense to process native materials for fuel, oxygen, and water. This type of resource processing would be accomplished most reliably by large numbers of small devices working to generate the necessary power and performing the required chemical processing of the indigenous materials.

Other application areas are possible and include decentralization of heating, ventilation, and air-conditioning (HVAC) systems (super zoning), cooling of protective suits for hazardous environments, and micropropulsion of miniaturized air, land, and sea vehicles. Although the true value of MECS will become known only during the next few decades, the vision of bringing about both evolutionary and revolutionary changes to the energy area is one of the driving forces in this developing field.

2. *UNIT OPERATIONS*

A unit operation is a single process critical to the functioning of a larger integrated system. Several unit operations are typically joined together, either sequentially or in parallel, to accomplish the overall system function. For MECS, this overall function could include (but is not limited to) heat generation through combustion, heat-activated heat pumping, or power generation. The unit operations terminology derives from chemical engineering where filtration, evaporation, distillation, and batch reaction processing are relevant. In this work, a broadening of the term is implied to include other elementary operations important for the generation and conversion of energy.

2.1 Heat Exchangers, Evaporators, and Condensers

The most fundamental process from an energy systems standpoint is heat transfer to (or from) a working fluid. This unit operation typically occurs in small channels, the characteristic diameter of which is less than 1 mm. Other heat transfer configurations, such as fluid flowing through a network of posts or through a mesh, are also possible. Specific examples of this fundamental unit operation include phase change occurring in boilers, evaporators, and condensers. Single-phase systems where the fluid undergoes a temperature change only also falls into this category.

A slightly more complicated arrangement for transferring thermal energy from one fluid to another occurs in a heat exchanger. Many different configurations exist for this type of device. Classifications include parallel flow, counter-flow, and cross-flow for the primary arrangement of the flow as well as shell-and-tube, plate-and-frame, single pass, multiple pass, and other terminology specifying the physical arrangement. Regardless of the type and configuration, heat exchangers are designed to recover thermal energy from one fluid and deliver it to another to enhance the overall performance of a system (by way of conserving thermal energy).

Use of engineered microstructures in heat and mass transfer has the potential of enhancing transfer rates, thus leading to process intensification in thermal/mass diffusion limited situations. The typical microstructure used is the microchannel array. Characteristic dimensions of these structures are small enough so that laminar conditions exist throughout. Consequently, diffusional processes are responsible for thermal and species mixing and result in a time for process completion proportional to d^2/α, where d is the channel dimension (the diameter if circular) and α is the thermal diffusivity (for thermal mixing) of the fluid. Thus, smaller channels lead to higher rates of thermal and mass transfer, resulting in process intensification. Surface area per unit volume can also be increased in these devices. If special design accommodations are not made in microchannel systems, the penalty for enhanced transfer rates is often higher pressure drops across the device. One of these accommodations would be to array a larger number of parallel microchannels so that a shorter length of fluid channel results in a lower pressure drop, whereas the increased number of parallel paths maintains the required throughput. It is not uncommon for heat fluxes exceeding $100\,W/cm^2$ to be achieved with water flowing through microchannel arrays.

2.2 Valves

Microvalves have been developed for a variety of applications suitable for MECS. Valve classification can be divided into passive and active categories. The passive type is the most fundamental and is configured as a basic check valve where a thin flapper plate covers an orifice. Microvalves of this type have been fabricated using a variety of techniques, including lithography and etching, laser micromachining, and electro-discharge machining. Materials have included metal on metal, metal on polymers, and silicon with photoresists such as PMMA and polyimides. Soft lithography has also been used to create check valves. Work in this area has resulted in miniature peristaltic pumps driven by compressed air. Sizes of these microvalves vary from the 50-μm range (for orifice diameters) up to 1 mm. Larger sizes are typically classified as conventional macroscale devices.

Active valves have designs that rely on electric actuation in some specific form. Direct actuation by a piezoelectric element has been achieved where the sealing element in the valve is a simple piezoelectric bending element, or the so-called bimorph, in place of a passive flapper in the check valves discussed previously. Electromagnetic solenoid actuation has also been developed, but this activity borders on conventional solenoid valves found commercially. A new development in the electromagnetic actuation area is arrays of active valves with characteristic sizes of a few millimeters and integrated together on substrates containing manifolding and sensor placement. This level of sophistication is only beginning to emerge from laboratory work in the commercial sector.

Secondary electrical effects using heating of bimetallic strips or shape memory alloy elements are being used for actuating valves. Electrical current passing through the active element (a metal strip) itself, or through a nearby resistor for I^2R heating, is the most common way of actuating valves of this type. For the bimetallic valves, heating causes a differential thermal expansion of the two metals, leading to a bending or distorting in a suitably designed element. Shape memory elements can also be designed as a composite actuator to provide an active opening and closing of an orifice on electrical current flow. Thermopneumatic forces have also been harnessed for valve actuation. This method relies on the expansion of a material, often through phase change, on heating. A common problem with all valve designs is leakage in the adverse direction, that is, the direction opposite to the desired flow when closed. Due to scaling of the sealing area, which goes as the characteristic length of the valve squared, valves become progressively more leaky as the overall scale is reduced. This can be mitigated somewhat by design; however, effective low-leakage valves at the microscale remain problematic.

2.3 Pumps and Expanders

Many small-scale energy systems require effective miniaturized pumps and expanders. These devices are responsible for conveying a working fluid through various components to execute a cycle that may result in power production, heat pumping, or refrigeration. As an example, a pump is required for pressurizing the working fluid in Rankine power cycles. Miniaturized pumps could also cycle a heat transfer fluid through a cooling loop for heat rejection purposes. On the other hand, expanders are necessary for any cycle employing a working fluid to generate power. Specific examples include microturbines, miniaturized piston-based machines, and roots-type expanders.

Most candidate techniques for pumping and expanding are ineffective for MECS applications. However, there are a number of ongoing projects concerned with development of pumps, compressors, and expanders. Micro turbo machinery has been the focus of one such project. Its targeted application is to develop a gas turbine engine the size of a shirt button. Many spinoff applications will result if the project is successful, including compressors, pumps, and expanders—all based on high-speed rotating machinery. Note that scaling turbo machinery down to the several millimeter size regime requires careful attention to machining tolerances, heat transfer effects, and operating speed. It is anticipated that rotational rates of up to 2 million revolutions per minute (rpm) will be needed in such devices. Various piston-based expander and compressor configurations can also be miniaturized at least down to the several millimeter range, but sub-millimeter devices will be a challenge. Seals, friction, and heat transfer effects all have adverse scaling characteristics; hence, novel design and fabrication methods are needed.

Pumps in the true microscale regime can be fabricated using photolithographic techniques where flexible diaphragms are used and electrostatic or electrothermal mechanisms are used for driving the diaphragms. Simple check valves of the flapper kind can be used to control the flow. Current state-of-the-art pumps of this type have low pumping speeds and limited pressure increases. Other microscale pumps working with weak electro–fluid interaction, such as

electrohydrodynamic and even magnetohydrodynamic effects, have also been evaluated. Low pumping speeds and minimal pressure rises have been observed. Until effective designs are developed, mesoscale pumps, compressors, and expanders may have to be coupled with many parallel flow microscale components to be effective. Hybrid mesoscale/microscale devices would result from this. Recent developments in the actuator area could help to bring about practical mesoscale/microscale pumps and compressors. New polymer-based electro-active materials and magnetic shape memory actuators are now becoming available to MECS researchers.

2.4 Mixers

Mixing is a process where two or more liquids or gas phase components are brought together and combined to form a uniform composition. Both turbulent and diffusive processes are typically employed to achieve the mixing action. As the scale of the process is reduced, diffusion becomes dominant. Microtechnology mixers are designed to bring together individual components for mixing with the smallest diffusional scales possible. Mixing is an important unit operation in the chemical and energy industry and can be found in combustors, microreactors, and adsorbing processes.

Micromixing technology has been developed in several different directions, but the common theme with most concepts is to rapidly bring together the constituents where a very small diffusive length scale exists for the final mixing step. Mixing in microchannel geometries has been studied in "Tee" sections, opposed microjets, channel flow through arrays of posts, and various interdigited designs. Mixing in converging channels and jet-in-cross-flow geometries has also been studied. Practical micromixers have been developed and proven out using deep-etch photolithography (in silicon) where an interdigited arrangement yielded a length scale of 100 μm (distance between two streams). Note that the time for diffusional mixing is proportional to L^2/D, where L is the distance between two mixing streams and D is the diffusion coefficient. Hence, it is important to reduce the device size to achieve rapid and complete mixing.

2.5 Separation Units

Separation is a unit operation used in absorption heat pumps and chemical processing applications such as solvent extraction and product separations. A typical separation process in a heat pump application involves the desorption of ammonia from a water–ammonia solution. Although a number of configurations have been studied for this process at the macroscale, most are based on gravity and have relatively low rates of desorption. For developing small heat-activated absorption heat pumps, microtechnology can be applied through the use of thin "sheet" channels where desorption of ammonia is accompanied by heat addition. As the sheet film in the channel is made thinner ($<100\,\mu m$), diffusional processes become exceedingly fast and high rates of desorption can be achieved.

Practical implementation of the thin sheet geometry for desorption has taken the form of mechanical membrane gas–liquid contactors. This approach has been dictated by the consequences of capillary forces present at small-length scales. Surface tension and flow in the "lubrication" regime of fluid dynamics conspire to prevent unconstrained films from flattening out to the dimensions necessary for rapid desorption. With the use of a mechanically constrained liquid film, progress on desorber units has progressed to the point of making miniature heat-activated heat pumps feasible. To realize the potential of the concept, an integrated approach to the design of a thin film desorber must be used. Thus, combining a microchannel heat exchanger with a mechanically constrained thin film desorber has resulted in high-performance units ideal for small-scale systems.

2.6 Microreactors and Combustors

MECS will most likely employ combustion for driving processes such as vapor generation, endothermic chemical reactions, and (most notably) fuel reforming. Both fuel reformers and combustors will be of a miniature design relying on embedded catalysts for promoting chemical reactions at moderate temperatures (350–750°C). Many potential configurations exist depending on the application and constraints on the design. Microchannel arrays are a potential configuration; mesh and post architecture is another to achieve the desired surface area and small diffusional lengths necessary.

Small-scale fuel reforming is an important area of research in the microreactor area. Hydrogen production for fuel cells is the main driver for this activity. Fuels such as methanol, ammonia, and gaseous hydrocarbons have been tested in laboratory settings with reaction volumes on the order of $1\,mm^3$. These devices are constructed using several different

techniques, but etched silicon and platelet technology (described later) are two methods being explored. Note that thermal management of microreformers is exceedingly important given that heat loss can represent a significant fraction of the energy transfer rates present in the device. Consequently, small reformers will require both high-performance insulation and gas stream configurations that recover thermal energy in the exiting flows. A systems approach to design optimization incorporating thermal management, flow pressure drops, and component interfacing will be necessary for practical systems to emerge from the laboratory.

A heat source is required to drive endothermic reactions of fuel reforming and shift reactions. Evaporators and boilers also require a heat source. Of all the conventional sources of heat, combustion of liquid hydrocarbons has the highest stored energy densities. Compared with electrochemical cells, liquid fuels have energy densities 35 to 300 times greater than current battery technology. This assumes that the fuel has an energy density of 42 kJ/g (with air coming from the surroundings) compared with a zinc–air battery at 1.2 kJ/g or a lead–acid battery at 0.125 kJ/g. The development of moderate-temperature (450–1000°C depending on the type of fuel) combustors in a miniaturized form has led to a practical method of releasing this energy in a small overall component size. Microcombustors have been developed in a size range of a few cubic millimeters where a platinum catalyst has been used to promote reactions. Catalytic operation is a necessity because miniature and microscale devices have much larger surface/volume ratios, and hence a high degree of heterogeneous reactivity, compared with conventional macroscopic combustors. Also, true microscale operation typically takes place at length scales much smaller than the quench distance associated with the fuel being burned. However, high-temperature operation can mitigate the need for catalytic surfaces in some cases.

As the characteristic size of combustors is reduced from mesoscale to microscale, thermal management plays an increasingly important role in efficient combustor operation. Work currently taking place in miniaturized combustors includes development of small excess enthalpy burners fabricated using a three-dimensional printing process by stacking hundreds of individually patterned layers. A toroidal combustor has been developed in the form of a "Swiss roll" where the hot region of the burner is isolated from the surroundings by the inward spiral of reactants and the outward flow of products. Excellent heat recovery can be achieved with this design. There has also been work accomplished on the development of microscale combustors based on a bayonet-style counterflow heat exchanger where the outside surface of the combustor is insulated by either vacuum conditions or high-performance insulation. Hydrogen and propane have been catalytically burned in volumes of less than 0.5 mm^3 (Fig. 2) at temperatures in excess of 650°C. Other work has also taken place on small-scale diffusion flames and insulated burners. The current technology offers validation of the ability to construct small-scale heat sources for integration with other components to build miniaturized energy systems.

2.7 Balance of Plant

A critical issue with MECS development is the size and complexity of the "balance of plant." This includes subcomponents such as air movers, fuel delivery components, sensors, valves, and other associative concerns (e.g., power for start-up). These issues are being explored by a few researchers concerned with the development of total systems, but all too often this aspect of energy systems developed is relegated to a minor consideration when in fact it can be the deciding factor for practical systems. The balance of plant problem must be explored in detail and resolved for each specific case being examined. Over the coming years, solutions may begin to emerge through ongoing efforts in related microtechnology development.

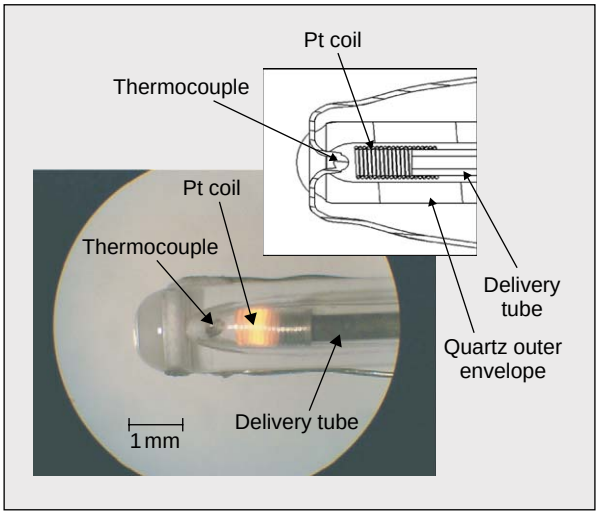


FIGURE 2 Example of catalytically promoted microcombustor. (Courtesy of Oregon State University.)

Specifically, microscale engineering and MEMS-type fabrication techniques may offer effective balance-of-plant solutions in an integrated packaging scheme.

3. SYSTEMS

Applications abound for microtechnology-based energy systems. Although not exhaustive, Fig. 1 conveys important developmental areas and provides a listing of technologies important in each size class. The current topic is concerned with the mesoscale/microscale size regime; overall systems of palm size or smaller are the main focus here. These systems are based on subcomponents that have microscale elements for process intensification but can fall into either the mesoscale or microscale regime. The focus here is concerned with power generation and cooling. These two applications will be the main areas for MECS during the coming years and could represent multi-billion-dollar industries by the end of the first decade of the new century.

3.1 Power Generation

Power generation can take the form of shaft work, electricity, or a propulsive effect (for applications involving flight or submersibles). Electrical power generation using fuel cells or microengines will become a direct competitor to batteries as the technology in MECS is developed into reliable systems. As mentioned earlier, stored liquid hydrocarbons have a large advantage over electrochemical storage in terms of energy density. However, batteries enjoy significant advantages in ease of use and reliability. Once MECS are developed to their full potential, they should provide portable applications capability with substantial benefits for both military and commercial uses.

3.1.1 Fuel Cells and Fuel Processing

Fuel cells are direct energy conversion devices that combine two reactants to produce electrical power. The reactants are typically a fuel such as hydrogen, or methanol, and oxygen from the air. Fuel cells require an electrolyte capable of passing an ionic charge carrier across an electronic conduction barrier where the ions are driven by a concentration gradient. Fuel cells also need a catalytic-based anode and cathode for reactant preparation. For mesoscale/microscale systems, fuel cells are best fabricated in thin film form. Depending on the desired power output of the system, the "footprint" may well be relatively large to supply the required power. Systems arranged in this manner are referred to as mixed scale systems on the basis that one critical dimension is small (the thickness), whereas the extent of the device (its footprint) can be the requisite size to satisfy a particular power application.

Thin film fuel cells operate across a broad range of temperatures. Proton exchange membrane (PEM) cells based on Nafion or similar material can operate at room conditions but provide better performance at elevated temperatures. The upper practical temperature limit for Nafion is approximately 100°C, although pressurized systems can go higher. The reason for this limitation is the requirement for keeping the membrane saturated with water to promote ion passage. Fuels for PEM cells include hydrogen and methanol as well as other fuels if reforming takes place. Direct methanol fuel cells have received much attention recently as a possible power source for portable electronics. Developments in PEM cells have resulted in new membrane materials operating near 200°C, where power densities can be higher and where catalysts on the fuel side (the anode) have less susceptibility to carbon monoxide poisoning. This is critical for cells consuming a reformer gas because carbon monoxide in low concentrations is usually present even after filtering.

Higher temperature systems are also a possible choice for mesoscale power systems. Solid oxide fuel cells (SOFCs) have traditionally been made with electrolytes of yittria-stabilized zirconia having a thickness greater than 100 μm. This has dictated operating temperatures approaching 1000°C due to low ion mobility through the electrolytes. Research on thin film SOFCs over the past decade or so has shown the possibility of operating at temperatures as low as 500°C. This makes them attractive for small-scale systems. Attributes of the thin film SOFCs at these lower temperatures include tolerance to many types of fuels (including carbon monoxide), no water management issues, and the possibility of operating with either internal reforming or direct fuel oxidation. If the power density of the thin film devices can be maintained at the lower temperatures, practical small-scale systems may result.

As mentioned earlier, practical PEM fuel cells require hydrogen to operate (with the exception being the work on direct methanol fuel cells). However, to carry sufficient quantities of hydrogen gas for cell operation, high-pressure canisters are required. This can be a safety hazard, especially if pressures greater than 100 atm are needed. An alternative is to extract hydrogen from a fuel rich

in this constituent (e.g., methanol, ammonia, hydrocarbons). Fuel reforming is the process by which this extraction occurs and has recently been the focus of much research. Note that other hydrogen sources have been considered, such as the so-called hydrogen-on-demand systems using borohydrides, but are not examined here.

Reforming takes place at temperatures from approximately 350°C for methanol to over 750°C for hydrocarbons and ammonia. The reactions are endothermic and require a source of heat to drive them to near completion. Small-scale reformers are configured so that fuel with water vapor passes over a catalyst within a microchannel array. The fuel molecules are "reformed" so that hydrogen, carbon dioxide, carbon monoxide, and excess water emerge from the device. Trace amounts of unreacted hydrocarbons can also contaminate the exit stream. The consequence of not having a pure hydrogen stream from the reformer is severe; if small amounts of carbon monoxide are present in the hydrogen supplied to a PEM fuel cell, poisoning of the anode catalyst results. Hence, reformers require a rigorous cleanup and filtering of the fuel stream. Nevertheless, mesoscale/microscale reformers with metallic membrane filters having internal volumes of less than a few cubic millimeters are being developed. As with all mesoscale/microscale devices operating at elevated temperatures, thermal management is important for conserving the energy of the process.

3.1.2 Miniature and Microscale Heat Engines

Many characteristics of traditional engines make them attractive for use in power generation and propulsion. They tend to be self-aspirating and rely on combustion, which at the macroscale is a very robust form of heat generation. Fuel is plentiful and inexpensive, with storage easily realized. The energy density of the fuel (or fuel plus container), when compared with electrochemical sources, is high. Along with these advantages come a number of drawbacks, especially where miniaturization is concerned. For example, an engine is thermodynamically restricted (by the Carnot efficiency) in its conversion of chemical energy to work due to the intermediate heat-generating step. This is in contrast to the direct energy conversion of fuel cells. However, conversion efficiency is respectable in macroscopic engines and can approach the 30 to 40% range at design speed and power output. Note that the overall conversion efficiency of fuel cells rarely exceeds 50% because of the cell's internal electrochemical irreversibilities and losses due to power conversion electronics, fuel use, and reforming. Thus, thermal engines provide a practical and reliable way of extracting chemical energy bound up in hydrocarbon fuels.

Most common thermal engines rely on compressing a cooler, condensed working fluid and expanding a hot, volume-expanded fluid. This is often accomplished through mechanisms that have rubbing or clearance seals, hot regions separated from cold structures, and bearings to allow differential motion between parts. All three of these defining characteristics suffer from adverse scaling effects. Although scaling laws are not discussed here, consider the case of maintaining the required temperature difference for the Brayton cycle (gas turbine). By its very nature, this engine exploits a difference in temperature to generate work from heat. This temperature difference is the driving force for engine operation, and as the temperature difference increases, better thermal efficiency results. However, as the size of the engine is reduced, its internal structure acts as a thermal pathway shunting a portion of the usable heat through two unproductive pathways. First, some of the heat is transferred to the surroundings without producing work. Second, heat is conducted through the connection between the turbine and the compressor, heating the inlet air stream. At smaller sizes, it becomes progressively harder to insulate the hot section of the engine to prevent heat flow through the two leakage paths. Simple calculations for palm-sized gas turbines having a 300 K inlet temperature and a turbine operating at 1800 K show the parasitic heat loss to be comparable to the overall power output. The parasitic effects become more pronounced as the engine is further miniaturized until no practical conventional design is possible.

Microthermal engines, as characterized by miniature Wankel engines, Stirling engines, and micro gas turbines, all are practical macroscopic engines but suffer from significant sealing problems, reduced subcomponent efficiency, friction, and thermal management issues when scaled to the mesoscopic size regime. To date, significant work has been invested in micro gas turbines and Wankel engines. The former is contemplated to have a rotor diameter of approximately 5 to 10 mm, a combustor temperature in excess of 1200°C, and rotation rates of up to 2 million rpm. Wankel engines and compressors have received considerable interest recently. The particular design features making this approach attractive are a simple overall design and no valves. Furthermore, a near two-dimensional layout of the engine (Fig. 3) would permit MEMS-type fabrication (in silicon or other suitable material) if workable designs emerge

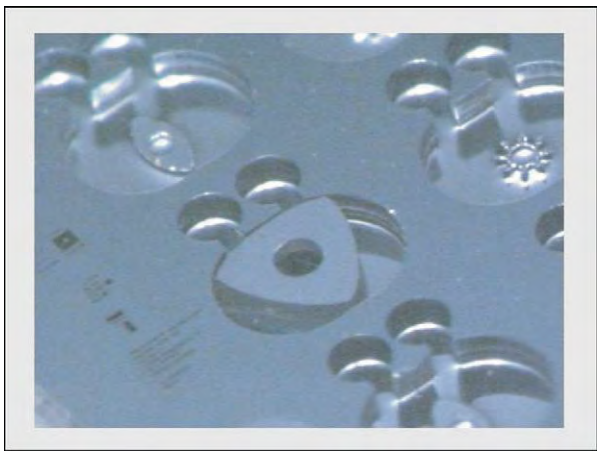


FIGURE 3 Lithographically deined 1-mm Wankel rotor in its housing. (Courtesy of University of California, Berkeley.)

from the laboratory. As with most miniature internal combustion engines, the Wankel presents serious questions regarding seals, internal heat transfer, and combustion in small spaces.

3.2 Cooling

An important technological application for MECS is small-scale systems cooling. As mentioned previously, an advantage will be realized in areas where portability, compactness, weight, and point application are the driving considerations. Many specific cases that meet these criteria exist. For example, cooling protective suits used in chemical and biological "hot spots" represent one important application. Super zoning in residential buildings for heating and cooling is another. At the smallest length scales, electronic chips would benefit from onboard and integrated cooling mechanisms. The following is a brief overview of some of the work being done in this area.

3.2.1 Refrigeration and Cryocooling

Microtechnology can be employed to produce miniaturized refrigeration and cryocooling systems. Although process intensification is typically the route used for miniaturization, microtechnology can also be used for reducing the size of mechanical components that are necessary for operation. For example, to create a small-scale vapor compression refrigerator, the heat transfer components comprising the condenser and evaporator can be made in a microchannel configuration for enhanced heat transfer rates. However, the mechanical compressor will often be the determining factor for overall size and weight. Advanced compressor designs can be developed by replacing existing components with micromachined ones, or entirely new types of compressors can be developed. In the former area, valve heads and drive mechanisms can be redesigned to take advantage of layered manufacturing techniques so that integrated components result. In the latter area, work has been under way focusing on electrostatically operated compressors micromachined out of silicon. Other approaches using thermopneumatic operation, electroactive polymers, and magnetic shape memory alloys have been pursued.

Cryocoolers can benefit from miniature components as well, but fundamental heat transfer issues must be considered. For example, counterflow heat exchangers or regenerators are necessary for cryocooler operation. This is due to thermal isolation requirements of the cold space from ambient temperature. The working fluid must pass relatively unimpeded (i.e., low pressure drop) through the heat exchanger while thermal energy is exchanged between the incoming and outgoing flows. To reduce the size of the overall cooler, each component (including the heat exchanger or regenerator) must be reduced in scale. However, this presents a heat transfer problem in the form of heat leakage into the cold space. As the size, and hence the length, is reduced, the temperature gradient along the heat exchanger increases, leading to enhanced heat transfer rates to the cold section. On further size reduction, a point is reached where the cooler load is entirely from this leakage and no useful heat lift takes place. Thus, there exists a limit, based on fundamental heat transfer principles, to which coolers can be reduced in size.

With these considerations, several of the techniques used at the macroscale for cryocooling have been investigated for miniaturization. For instance, both pulse tube and Stirling cycle devices have been studied for miniaturization. Conduction through the regenerators and heat exchangers limits the ultimate size of the system, but cryocoolers that are approximately a centimeter in length appear to be practical. Miniaturization of the cyclic compressors needed for cryocoolers is also an important need in this area. Typically, piston-based compressors are used; reducing their size involves all of the challenges as does the microengine area discussed earlier. Another possible approach to cryocooling uses a reverse Brayton cycle. Miniaturized turbo machinery could be effectively employed for this application. However, the performance of any Brayton cycle, whether

power producing or in cooling applications, is strongly influenced by the efficiency of the individual components. Thus, compressor and turbine elements in a reverse Brayton cycle machine must be relatively efficient for the cycle to be practical. Although this section has not covered solid-state coolers or refrigerators, much work has been under way for producing efficient small-scale devices (e.g., thermoelectric coolers). If high "figure-of-merit" materials become available for thermoelectric cooling, they will play an important role in miniaturized cooling applications.

3.2.2 Heat-Actuated Heat Pumps

Heat-actuated heat pumps are cooling systems requiring no electricity for operation. Rather, a heat source is used, avoiding the need for batteries. This consideration is critical for portable applications because on a weight basis, a stored liquid fuel has anywhere from 35 to 300 times the energy content of batteries, depending on the battery technology considered. Thus, although the simple solution of combining an electric motor, a battery package, and a vapor compression refrigerator may sound appealing, it actually suffers a severe volume and weight disadvantage when compared with a heat-actuated system. An efficient heat-actuated cooler, perhaps based on combustion of butane or a logistics fuel, can be especially applicable to cooling protective suits and vehicle interiors. In general, the coefficient of performance (COP) for a heat-actuated cooling system needs to be in the range of unity for effective systems to be fielded.

Two types of heat-activated cooling systems are considered viable for miniaturized systems. Both would rely on microtechnology for operation of critical components, and both would use combustion of a storable fuel to drive the process. The first is an absorption cycle heat pump with a working fluid of lithium bromide and water or ammonia and water. The key components of the cycle are shown schematically in Fig. 4. The feature of this device making it a possible choice for palm-sized cooling is the use of microtechnology in the absorber, desorber, and evaporator. As shown in the diagram, a small amount of electrical power is needed for pumping the two-component liquid to a higher pressure. However, more than 90% of the energy used in the device would come from combustion of the stored fuel.

Significant developmental challenges exist for miniature absorption coolers intended for protection suits. The absorber and desorber rely on thin,

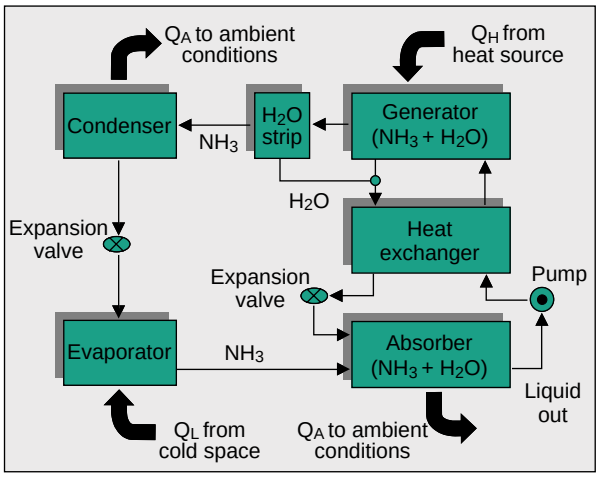


FIGURE 4 Schematic diagram of an ammonia/water absorption heat pump.

mechanically constrained membranes that mediate the separation of the two-component system. For practical use, orientation independence is required so that fieldable units can operate whether the wearer is standing upright or lying down. Finally, an integrated approach for the main components of the cooler, as well as for the balance of plant, must be developed so that economical fabrication is possible.

The second possible system for miniaturized heat pumps takes the form of an engine-driven vapor compression refrigerator. Microtechnology would be used throughout, but especially in the refrigerator components such as the compressor, evaporator, and condenser. The miniature engine for this type of heat pump has yet to be developed. Candidates are the aforementioned small-scale turbines and Wankel engines as well as other engine concepts using external combustion. High efficiency in a small package will be a key aspect to this heat pump concept. Furthermore, thermal management, including insulation around hot engine parts and energy recovery in the exhaust stream, will be critical if the engine-driven vapor compression cycle is to be realized.

4. MATERIALS, FABRICATION, AND COSTS

4.1 Criteria for Materials in the Energy Area

In fabricating microscale energy systems, a number of operating conditions must first be taken into account. Foremost are the operating temperatures throughout

the device. Often with mesoscale/microscale energy systems, two closely spaced parts of the same device have a large temperature difference between them; hence, a large gradient can be present. A primary constraint is to have the material withstand the maximum temperatures existing in the system, but the gradients and thermal expansion must also be tolerated. This is especially true if more than one material is used for construction where transitions must be made between material sets. Corrosion and oxidation are also important considerations when extended operating lifetime is required. Materials must be chosen for compatibility with the working fluid being used, especially at the temperature extremes of operation. The various materials comprising all wetted parts must not catalyze decomposition reactions at the fluid–surface interface—not even slightly—or else extended operation cannot be ensured. Furthermore, the materials must retain sufficient strength at the operating temperature and should have a low thermal conductivity. This latter criterion stems from the requirement of minimizing conductive heat loss throughout the device. Finally, cost, ease of machining, and material joining factors (e.g., brazing, diffusion bonding, soldering compatibility) must also be considered when selecting the materials of construction. With all of these criteria and constraints, the engineer must have a versatile material set to work with when designing and constructing MECS.

Other considerations come into play when selecting thermal insulation for isolating the elevated temperature (or cold) sections of the device from ambient conditions. Several potential material sets are available for this task, but operating temperature is one of the deciding factors in using a particular material or technique. For example, silica-based aerogel material has a very low effective thermal conductivity and can be used throughout the temperature range from cryogenic to elevated temperatures of up to 800°C. The aerogel can be either infiltrated with a gas or evacuated for very low thermal conductivity. However, radiation heat transfer in the infrared (IR) region must be effectively blocked at high and low temperatures. To accomplish this, carbon black is typically used as a dispersant. But because of oxidation of the carbon, this formulation is incompatible with air at temperatures above approximately 350°C. Other insulating systems, such as multifoil vacuum insulation and simple vacuum gaps, can be used. For long-term service, vacuum conditions must be maintained and might require gettering.

4.2 Types of Engineering Materials Available

Silicon is the material of choice for most MEMS due to (1) the feature size often being in the range of less than 20 μm requiring single crystalline material and (2) the dependence on fabrication techniques developed by the electronics industry for processing this material. However, it is a particularly poor choice for the larger thermally based devices that characterize MECS. The room temperature thermal conductivity of silicon is approximately 150 (W/mK). Use of this material leads to high rates of heat transfer across the short distances where temperature gradients occur. The thermal conductivity of silicon above 1000°C falls to below 25 W/mK, but the average value can still be high. Various metallic alloys and intermetallics, including common stainless steels, are potential fabrication materials for small-scale heat engines and coolers. Most stainless steels have a room temperature thermal conductivity between 11 and 15 W/mK. In contrast to silicon, the thermal conductivity of metals increases with increasing temperature. It is interesting to note that between approximately 1000 and 1200°C, the thermal conductivities of silicon and stainless steel are similar. However, heat loss through each material will depend on average thermal conductivities, and silicon will still have the higher average value. From a heat loss point of view, the best materials will probably turn out to be ceramics and glasses with amorphous structures. Fused silica is at the low end of the thermal conductivity range, with a room temperature value of 1.4 W/mK. This climbs to approximately 4 W/mK at 1200°C. It is also highly resistant to thermal shock due to its low thermal expansion coefficient. A class of material not known for elevated temperature use is polymers. However, polyimide (and other similar polymers) can be used up to approximately 400°C, and this could make these polymers attractive in certain applications.

In most MECS concepts, internal structures will be engineered to reduce internal thermal conduction, which implies thin sections. However, these same sections often contain a working fluid at elevated pressures and must be hermetic. Furthermore, they must support the stress due to the operation of the device and must not fail catastrophically. Single crystal silicon and other pure semiconductors (most notably silicon carbide) can be excellent structural material for microtechnology components. But they are brittle and unforgiving in a yield mode. The so-called engineering materials, including steels and

other metals, ceramics, and plastics, will probably be applied to the construction of larger microscale and mesoscale devices due to a wide selection of both properties and bonding techniques available. Material costs are reasonable for most of the engineering materials, and various forms such as thin sheets are commercially available. Another deciding factor, as discussed in the next subsection, is the availability of lower cost manufacturing techniques. In this respect, silicon requires a wide suite of expensive equipment for bulk and/or surface machining the material.

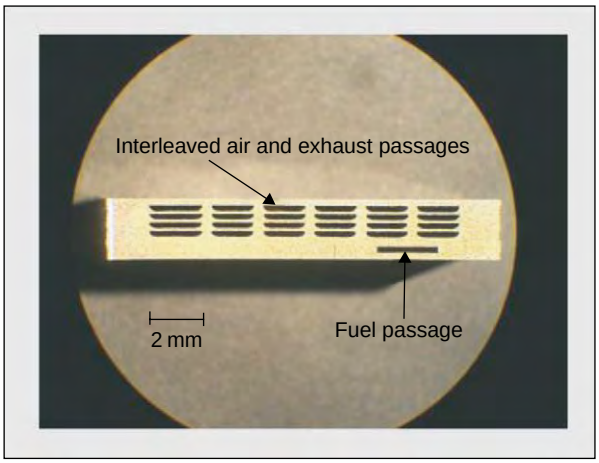


FIGURE 5 A three-stream counterflow heat exchanger fabricated using platelet technology. (Courtesy of Oregon State University.)

4.3 Microfabrication Techniques for Engineering Materials

Fabrication techniques developed for integrated circuit (IC) production have been refined to the extent of supporting a multi-billion-dollar industry. Chip manufacturing relies on silicon-based processing where micron-sized features are routinely used in production. MECS do not require the extremely small "line widths" needed for IC fabrication. Furthermore, in many energy applications, silicon is not the favored material, as discussed previously. Other fabrication techniques, such as LIGA, have been specifically developed for MEMS. Although many rely heavily on silicon processing, others can produce very small structures in metals electrodeposited on a surface or within a micromold. Again, for MECS applications, the feature size of these MEMS fabrication techniques is usually much smaller than what is needed.

Because MECS are fundamentally different from traditional ICs and MEMS, they require different materials and fabrication processes. One important fabrication method for microenergy applications is called microlamination or platelet technology (Fig. 5). Although its history dates back to the 1970s, when it was developed for liquid rocket engine injectors, it is now being pursued by a number of groups for fabricating MECS. The method is based on microlamination of metals, ceramics, and polymers. The process begins by surface machining, or cutting, a single lamina with a pattern containing the desired structure. The lamina is often a shim of a material having desirable mechanical and thermal properties important to the functioning of the final device. Once the pattern is cut, the shims are surface treated and precisely stacked in a prearranged order. The stack is then bonded together, forming a single block of material. For the platelet architecture to have utility, a machining method capable of fabricating structures in the laminating material is needed. The method must be versatile, easy to use, and capable of rapid machining (with through-cuts and surface texturing) in a wide variety of materials. One of the most general techniques is laser numerically controlled micromachining. It is most useful for prototype runs and can be used on metals and polymers. Other useful techniques specific to only one class of material can also be used. For example, chemical etching through a photographically defined mask (photolithography) can be used on most metals. The process is commercially available and can be employed for high-volume production runs. Another machining technique applicable to most metals is wire-based electrodischarge machining (wire EDM). Current machines on the market are numerically controlled and have wire diameters as small as 50 μm. The cutting of metal platelets with this technique is primarily a two-dimensional operation and can be as precise at 5 μm. Note that conventional high-speed milling, with end mills as small as 100 μm, can be achieved with most modern equipment and can be applied to both metals and plastics. Although wire EDM and conventional milling with small tools have been described in the context of platelet fabrication, their precision and speed can also be applied to a wide range of small mechanical components needed for MECS.

The properties of ceramics (and of materials such as quartz) are desirable from an energy systems point of view, but this class of material is difficult to machine and form. Relatively few techniques are available for cutting the requisite structures in ceramics. However, work is progressing on various

additive techniques for making small components out of both plastics and ceramics. Some of these techniques go by the name stereolithography and often employ a laser to selectively fuse a particle bed of the starting material into a desired shape.

SEE ALSO THE FOLLOWING ARTICLES

Fuel Cells • Heat Transfer • Technology Innovation and Energy

Further Reading

Ameel, T. A., Warrington, R. O., Wegeng, R. S., and Drost, M. K. (1997). Miniaturization technologies applied to energy systems. *Energy Conversion Mgmt.* **38,** 969–982.

Ehrfeld, W., Hessel, V., and Lowe, H. (2000). "Microreactors: New Technology for Modern Chemistry." Wiley–VCH, New York.

Epstein, A. H., and Senturia, S. D. (1997). Macro power from micro machinery. *Science* **276,** 1211.

Fernandez-Pello, A. C. (2002). Micro-power generation using combustion: Issues and approaches. "The 29th International Symposium on Combustion." The Combustion Institute, Supporo, Japan.

Goemans, P. A. F. M. (1994). Microsystems and energy: The role of energy. *In* "Microsystem Technology: Exploring Opportunities" (G. K. Lebbink, Ed.), pp. 50–64. Samsom BedrijfsInformatie, Aphen aan den Rijn, Netherlands.

Larminie, J., and Dicks, A. (2000). "Fuel Cell Systems Explained." John Wiley, New York.

Madou, M. (1997). "Fundamentals of Microfabrication." CRC Press, Boca Raton, FL.

Peterson, R. B. (2003). Miniature and microscale energy systems. *In* "Heat and Fluid Flow in Microscale and Nanoscale Structures" (M. Faghri and B. Sunden, Eds.). WIT Press, Southampton, UK.

Migration, Energy Costs of

CHARLES R. BLEM
Virginia Commonwealth University
Richmond, Virginia, United States

1. Importance of Migration
2. Migration Patterns and Ranges
3. Molecular Nature of Energy Stores
4. Magnitude of Energy Stores
5. Anatomical Storage Sites for Energy Reserves
6. Patterns of Energy Storage
7. Energetic Costs of Migration and Flight Ranges
8. Energetic Benefits of Migration

Glossary

β-oxidation Metabolic process whereby fatty acids are oxidized to produce energy for locomotor activity.
fatty acid Major energy source for many migrants; part of triglycerides.
glycogen Carbohydrate formed by the union of two glucose molecules.
homeotherm Animals that are insulated and regulate their body temperature, typically with the assistance of physiological heat production by shivering.
hyperlipogenesis Excessive production of lipid stores brought about by hyperphagia and increased enzyme activity.
Krebs cycle The tricarboxylic acid cycle; chemical process whereby metabolic substrate provides energy for utilization and storage by the organism.
metabolizable energy Energy obtained from food; the basic amount of energy available for all activities.
passerine Birds belonging to the order Passeriformes; relatively small perching birds, including sparrows, thrushes, and swallows.
poikilotherm Animals that do not regulate their body temperature with precision; typically, they get heat from external sources.
triglyceride Molecular form of storage lipid; a glycerine backbone to which three fatty acid molecules are attached.

Migration is defined as the periodic, regular (usually annual) movement of animals between breeding sites and wintering or feeding–maturation areas. Numerous species of animals migrate, but the majority making trips of significant distance include swimming (fish, sea turtles, sea snakes, cetaceans, and pinnipeds) or flying (insects, birds, and bats) species. The energetic cost of locomotion has been measured in numerous animals, large and small, as well as the energetic costs of locomotion (the latter in kJ/kg km).

1. IMPORTANCE OF MIGRATION

Because costs of transport of terrestrial migrants are high, and risks of predation severe, there are few species of small animals that migrate by walking or running, and all animals making terrestrial migrations of more than a few kilometers are relatively large mammals, such as wildebeest (*Connochaetes taurinus*), caribou (*Rangifer tarandus*), and bison (*Bos bison*).

In insects, migration occurs fairly frequently among species in the orders Lepidoptera (butterflies and moths) and Diptera (flies) and occasionally in Orthoptera (grasshoppers and locusts), Colleoptera (beetles), and Odonata (dragonflies). Many fish are migratory—so many so that migrants have been classified into three groups; anadromous, catadromous, and oceanic species. Anadromous fishes are those that are born in freshwater, mature in the ocean, and make spawning runs back in freshwater (salmon, herring, and lampreys). Catadromous fishes are those that are born in the ocean, then move up into freshwater habitats to mature, and finally make trips to spawning grounds back in the ocean. Oceanic species (sharks, tuna, and swordfish) migrate much like whales, following food, often with the assistance of currents such as the Gulf Stream. Like terrestrial mammals, migration is limited to larger fishes. Most small fish cannot swim rapidly and cannot store sufficient energy to travel long distances. Young eels are an exception. They are able to make much of their

 © 2004 Elsevier Inc. All rights reserved.

migration by drifting with the current. Among birds, many migrants are members of the order Passeriformes (perching birds) or belong to the various groups of waterfowl (ducks, swans, cranes, storks, grebes, and loons). However, just about every order within the class Aves includes at least a few migratory species. For example, nearly all temperate zone swifts, hummingbirds, and goatsuckers are migratory and some of the pelagic seabirds (shearwaters, albatrosses, and jaegers) are remarkable travelers that wander over huge expanses of ocean. Among mammals, cetaceans (whales), pinnipeds (seals), and bats include many migratory species.

Because the risks of migration are great, it is axiomatic that migrants must derive significant selective advantages from changing location. The focus of this article is on the balance of costs and benefits that affect the ultimate reproductive success of migrants. Although many of the advantages are potentially related to reduced competition and lower predation rates, some of the justification discovered by investigators of the phenomenon has involved energy acquisition and use. Without doubt, natural selection favors animals that minimize energy expenditure during migration, either to increase their potential range over barren habitat or to reduce the amount of foraging necessary after shorter trips.

Although we often think of poikilotherms (cold-blooded animals) as lethargic organisms less active than homeotherms (warm-blooded), it is significant that migration is fairly common in both groups. There is a relatively large amount of information regarding the physiology and ecology of migration of insects and fishes, but little attention has been given to similar aspects of terrestrial migrants. Furthermore, the information regarding the ecophysiology of flying migrants, primarily birds, far exceeds all other sources and this article reflects that bias.

2. MIGRATION PATTERNS AND RANGES

Migration, strictly defined, can be as restricted as a trip of a few hundred meters or as much as thousands of kilometers. Some salamanders (e.g., mole salamanders and newts) and many species of frogs (particularly tree frogs) emerge from the soil and travel to temporary ponds to breed. They then return to their subterranean hiding places, where they may remain for the next 6–11 months. The distance involved may be only a few meters.

The most extensive migration may be that of the Arctic tern (*Sterna paradisaea*), which literally spends its lifetime in migration, traveling 40,000 km or more, its trip interrupted only by an annual period of reproduction. Note, however, that this trip is not nonstop and the tern feeds along the way, can drink seawater, and rests on the surface of the ocean. The migration of the blackpoll warbler (*Dendroica striata*) may be viewed as even more remarkable because it may fly more than 80 h nonstop from New England to Bermuda or the Caribbean Islands with no food or rest. However, the monarch butterfly (*Danaus plexippus*), which travels more than 3000 km, or dragonflies, which apparently fly across the Gulf of Mexico, are no less striking in their abilities to migrate.

Patterns of migration vary among different groups of migrants. Migrants may travel exclusively during the day (diurnal), at night (nocturnal), or up and down mountainsides (altitudinal). Some insects and mammals (e.g., lemmings) said to be migratory actually are not because their movements are emigrations from population centers to food sources. Their return to their site of birth is not predictable and often does not happen at all. Most migrants travel in groups that either are loosely formed (insects and small birds) or are in distinct flocks (waterfowl and cranes). Onset of migration is usually controlled by weather conditions. For example, individuals of migratory fishes typically gather in schools (e.g., salmon, eels, and shad) before setting out. In some fish species, the number of migrants passing a point pulses in correlation with environmental conditions such as water temperature. Migratory insects may incorporate several generations in one annual circuit (e.g., monarch butterfly), during which time adults survive long enough to get to the next stopover and lay eggs, which develop into the adults covering the next leg of the migration. Birds, particularly small passerine species, are the icons of migration. They annually depart from wintering grounds in spring and from breeding grounds in autumn at fairly precise dates, take similar routes from year to year, and make extensive preparations for the trips. Migratory birds and some insects (e.g., dragonflies) must cross great expanses in which there is no suitable habitat to land or refuel. Birds crossing the Gulf of Mexico number in the millions, and in the spring their passage (together with numerous insects also in migration) may be readily seen on radar screens.

Birds crossing more hospitable terrain typically alternate one or more days of flight followed by

longer stopover periods, during which energy for the next flight period is accumulated. Short flights require relatively small energy reserves, whereas long flights depend on large fuel loads. The total amount of energy reserve is a compromise among the effects of loading on energetics, risks of the trip, and necessary duration of the flight. Furthermore, a key factor in acquiring the birds' optimal fuel reserves and for their resulting migration speed is the rate of fuel deposition at stopover sites. In small birds, metabolizable energy and intake rates are very high, particularly 1 day after reaching stopover (feeding) areas. This may indicate an increase in the capacity of the digestive tract after many hours of fasting during flight.

Some fish have similar adaptations for migration. Sockeye salmon (*Oncorhynchus nerka*) accumulate large fat reserves that are largely consumed during swimming over long migratory trips to upstream breeding sites (as much as 1200 km). In the course of their migration, they may use 91–96% of their lipid (possibly all of the lipid that can be mobilized) and more than 50% of the protein content of their flesh.

The general timing of migration is under the control of circannian (endogenous annual) physiological rhythms more finely tuned by changes in photoperiod. Increasing day length stimulates spring migration in a number of species, and decreasing day length stimulates fall migration. This is especially evident in birds. Day length is the most accurate environmental indicator of time of the year, but many species begin subtle preparation for migration before the photoperiod changes greatly. Timing also varies with age of the migrant, body size, distance of the migration, and rate of accumulation of energy stores.

Many migratory species routinely return to the site of their birth with precision. This is called philopatry or site fidelity. Philopatry has long been known for birds, but it has also been recognized in sea turtles, bats, cetaceans, and other species. Returning to specific locations, especially if breeding has been successful, is of obvious importance and may be related to the ability of the organism to accumulate sufficient energy in its previous breeding attempt. However, such precise migration requires special skills. The navigation of long-range migrant birds depends on their striking sense of location, including their ability to detect magnetic fields. Some species correct their direction of movement by integrating information from star patterns with these fields. Wintering site fidelity is less well-known, but there are examples of species (passerine birds) that appear to return to specific sites on the wintering grounds and this is especially true for nonpasserine birds, such as waterfowl and cranes, that travel in specific flyways and winter on the same refuges year after year.

No other single event in the life history of migratory organisms requires more energy than that expended in travel during this period. Indeed, within the span of a few days, some small passerine birds may spend a significant proportion of their total annual energy budget in spring and fall migration. The majority of this energy is expended in two activities: accumulation of energy stores for their journey and the locomotor costs of movement between starting points and ultimate destinations. Much of the evolution of migration centers on the amount and nature of the energy stores and on reduction of the costs of travel.

3. MOLECULAR NATURE OF ENERGY STORES

Several fuels have been shown to support the energetic demands of muscular activity of migrating animals. Among these are carbohydrates, lipids, and some amino acids. Not all animals use the same fuel and some may even use different substrates during different stages of a prolonged trip, particularly if feeding is not possible. Glycogen and trehalose, a disaccharide that hydrolyzes to glucose, may be used in metabolic pathways of some migrant Diptera (flies) and Lepidoptera (butterflies and moths). Other species typically store glycogen, which is transformed to glucose (a simple sugar) for use in the Krebs cycle. In birds, glucose and glycogen are used for muscular activity, but accumulation of lipid for migration has been observed widely; it has been demonstrated in at least 40 families of birds. Some insects and birds use large proportions of glycogen early in migration but begin to use lipid as glycogen stores are depleted. Protein can be used in cases of extreme exertion, but this usually is indicative of an animal near its limits of survival.

In many birds, some insects, and perhaps other animals, nonesterified fatty acids provide fuel for avian exercise, particularly as activity is prolonged. Declines in lipid reserves and downward shifts of respiratory quotients (carbon dioxide produced/oxygen consumed) during prolonged locomotion under controlled conditions in the laboratory support these observations. Premigratory fat deposits are generally

TABLE I

Energy Content of Common Fuels Used in Migration

Material	J/g
Carbohydrates	16.7–18.8
Proteins	~18.0
Lipids	37.7–39.7

composed of triglycerides containing a mixture of fatty acids of various chain lengths and degrees of saturation. The metabolic pathway in which these lipids are used as fuel in many insects, fishes, birds, and mammals is β-oxidation. In this process, fatty acids are broken down into two-carbon units, converted to acetyl-CoA, which then enters the Krebs cycle. A great deal of energy (38 ATP) may be generated for every two-carbon unit going through the biochemical pathway.

Lipid is superior to carbohydrate as an energy reserve because lipid contains more energy than carbohydrate (Table I), and storage of carbohydrate is typically accompanied by storage of water, which increases body mass but contributes no energy. The water may amount to approximately 3 g per gram of glycogen. In birds, the net result is that 1 g of lipid potentially provides more than 37 J, whereas 1 g of glycogen (including water) provides only approximately 5 J. It is obvious that lipid is superior to glucose or glycogen as a intrinsic fuel source for long-range migration, especially in flying animals.

Lipid depots are mostly in the form of triglycerides (triacylglycerols), which consist of three fatty acid molecules attached to a glycerol moiety. The triglyceride content of adipose tissue may exceed 80% in some species, particularly birds. The fatty acids can be released metabolically from the glycerol molecule either entirely or partially and are then carried to the mitochondria within muscle cells, where they are transformed to produce substrate for oxidative metabolism. They are converted to two-carbon (acetyl) fragments that are oxidized in the Krebs cycle, producing relatively large numbers of ATP molecules. The amount of energy produced depends on the carbon chain length of the fatty acid (Table II). Although fatty acids may vary from 10 to 24 carbon atoms (chain length), the most common fatty acids have chain lengths of 16 and 18 carbons. Triglyceride composition may vary seasonally, geographically, and with diet of the migrant. The most energy-rich forms of triglyceride are those with greatest chain length that are least saturated (more double bonds). Unfortunately, these have higher melting points and may be less easily mobilized for use. Migratory birds tend to have flight muscles that are largely "red" fibers with many mitochondria, a good blood supply, and the metabolic capacity for sustained exercise fueled by lipids.

TABLE II

Common Names, Melting Points, and Energy Content of Common Fatty Acids

Common name	c:d[a]	J/g	Melting point (°C)
Saturated fatty acids			
Caparic	10:0	35.3	32
Lauric	12:0	36.8	45
Myristic	14:0	38.0	54
Palmitic	16:0	38.9	63
Stearic	18:0	39.7	70
Arachidic	20:0	40.5	76
Behenic	22:0	41.0	80
Lignoceric	24:0	41.4	88
Unsaturated fatty acids			
Palmitoleic	16:1	38.5	33
Oleic	18:1	39.3	14
Linoleic	18:2	39.7	−5
Linolenic	18:3	40.0	−11
Arachidonic	20:4	—	−50
Erucic	22:1	—	33
Nervonic	24:1	—	43

[a] Ratio of carbon atoms (c) to double bonds (d).

4. MAGNITUDE OF ENERGY STORES

The amount of fuel stored for migratory trips is extensive in some species. For example, the blackpoll warbler, a bird that flies more than 80 h nonstop during fall migration, may add 10 g of lipid to its body mass of 11 g shortly before its fall migration. Hummingbirds that cross the Gulf of Mexico during migration may be even more obese. These birds arguably have the largest energy stores of any animal in the world, rivaled only by some mammals in their prehibernation period and some cetaceans and pinnipeds outside of migration periods. Insects and bats that migrate over fairly large distances also deposit substantial amounts of fat.

In general, the mass of energy stores varies with size of the migrant, the distance to be covered, and barriers to be crossed (Table III). For example, small

TABLE III

Lipid Reserves in Some Migrants

Species	Lipid (%)[a]
Short-range migrants	
Chorizagrotis auxiliaris (moth)	5–15
Aphis fabae (aphid)	31
Yellow-rumped warbler	55–84
Little brown bat	38
Long-range migrants	
Migratory locust	35–69
Monarch butterfly	43
Blackpoll warbler	342
Bobolink	273

[a] Percentage of total dry mass.

birds have relatively larger premigratory energy stores than large birds, everything else being equal. Transoceanic migrants must have large fuel supplies, especially if they are not able to land on the water or forage en route. Accumulating such reserves may be done by any or all of the following: hyperphagia (excessive eating) in the period prior to departure, increased efficiency of nutrient utilization, reduction of energy expenditure in activities not related to preparation for migration, and selection of food items high in energy. In most species, only hyperphagia seems to be important. Some migrants, particularly small birds, can fatten significantly in a short period of time, perhaps 1 or several days. Premigratory hyperphagia is best known in birds, but relatively little evidence of the phenomenon exists for insect, bats, or other forms. In birds, hyperphagia results in hyperlipogenesis and lipid deposition may rapidly occur. Some investigators have not been convinced that all of the fattening occurs as a result of hyperphagia, but there is little evidence of increased efficiency of assimilation, decreased activity, or decreases in standard metabolism in the premigratory period. In fact, small birds demonstrate intensive nocturnal activity under caged conditions (Zugunruhe) during the premigratory period that is not shown at any other time. This paradoxical behavior would be wasteful of energy but probably represents a stereotyped behavior that replaces actual migration. There have been several observations of animals making dietary shifts to items containing more energy during the period just prior to migration.

Accumulation of sufficient reserves in a short period of time has a number of advantages. First, the time of arrival of some birds on their breeding grounds may be affected by their ability to accumulate sufficient reserves on the wintering rounds (or v.v.). It appears that arrival dates are extremely important in gaining advantage on competitors, exploiting food resources that are only temporarily abundant, and avoiding inclement weather. Second, heavy fuel loads may increase wing loading in flying animals, thus making them vulnerable to predators, so rapid accumulation and use is adaptive.

5. ANATOMICAL STORAGE SITES FOR ENERGY RESERVES

Energy deposits are basically stored at two general locations within the migrant's body. Poikiolothermic animals tend to store lipids in discrete fat bodies that are located within the body cavity. (Fat bodies in the thorax and abdomen are the major sites of storage in insects.) This is also generally true for short-distance migrants, such as frogs and some salamanders, but may vary widely in other poikilotherms. For example, in migratory fish, lipid storage sites are in connective tissue, mesenteries of intestines, skeletal muscle, liver, or the head, skin, and tail. Homeotherms also deposit lipids within the body cavity but usually store large amounts of fat at subcutaneous sites. Whales are obvious examples. Their subcutaneous fat deposits were the foundation of the whaling industry. The difference between the spatial patterns of poikilotherms (fish, amphibians, and reptiles) and homeotherm (birds and mammals) may be due to heat conservation by the latter through the increased insulation provided by the subcutaneous fat. In birds, fat deposition may be done in a fairly discrete sequence in which the lipid is placed at fairly specific sites, beginning in the body cavity and ending at several subcutaneous sites. For example, the white-crowned sparrow (*Zonotrichia albicollis*) deposits lipid in 15 recognizable regions. In birds, subcutaneous layers of fat associated with feather tracts appear first. Subsequent fattening is manifested as greater amounts of subcutaneous fat plus some abdominal storage. In the final stages of fattening, intraabdominal deposits become extreme. When fat is deposited at both abdominal and subcutaneous sites, it is usually used in the opposite sequence as its deposition (i.e., abdominal fat is used first and subcutaneous fat last). It is logical that lipid nearest sites of utilization would be easiest to mobilize, but there is evidence that the molecular nature of the fatty acids

may differ with the location of the depot. Long-chain, saturated fatty acids have higher melting points than short-chain, unsaturated ones and are presumed to be more easily mobilized. Subcutaneous sites are cooler than abdominal sites, so long-chain, saturated fatty acids may be less easily used.

6. PATTERNS OF ENERGY STORAGE

The storage of energy in the depot fats that fuel most of the migratory trip may play a dominant role in the species' behavior for short periods. Some animals spend the majority of their lives in transit in a way that requires little energy storage for travel (e.g., tuna and seabirds). Storing huge reserves by cetaceans may reflect food availability on opposite ends of the their migratory pathway more than the energy demands of migration. Most other species have life histories that involve annual sequences of breeding and wintering, punctuated with two annual migrations. This pattern usually involves periods of energy accumulation and storage immediately prior to migration, the magnitude of which may be remarkable. In migrants that cross large areas in which food is unavailable, fuel storage is more pronounced. For example, transoceanic migrant passerines deposit large amounts of fat; intracontinental migrants stop periodically, never fattening extensively. Swimming migrants (fish and cetaceans) may likewise pass through areas in which food is scarce and thus must store large amounts of fuel (typically lipid). Terrestrial, short-range migrants either make very short trips, after which they feed for a few days, or feed as they go, stopping only during part of the day. In both instances, they have relatively small energy depots in comparison with long-range migrants.

In some cases, the temporal pattern of energy storage differs between spring and autumn migration. The difference may be a response to availability of food resources but also may be due to the benefits of precise timing of migrants traveling to breeding sites. In some small birds, fattening in spring occurs during a shorter period of time than does lipid acquisition in the fall, and nonmigrants show little seasonal variation in fattening, especially if they live in tropical habitats (Fig. 1). Precision in spring is beneficial because the earliest arrivals tend to obtain better territories and are more prone to reproduce successfully.

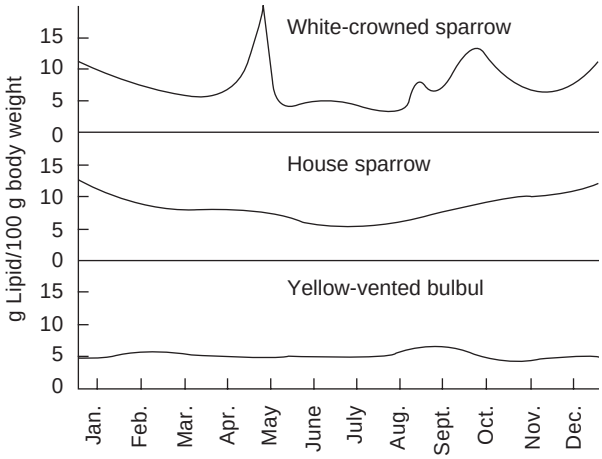


FIGURE 1 Annual cycles of lipid deposition in three passerine birds. The white-crowned sparrow (*Zonotrichia albicollis*) is a temperate zone migrant, the house sparrow (*Passer domesticus*) is a temperature zone nonmigrant, and the yellow-vented bulbul (*Pycnonotus goiavier*) is a tropical nonmigrant. From Blem (1990).

7. ENERGETIC COSTS OF MIGRATION AND FLIGHT RANGES

The energetic cost of migration, although striking in the short term, typically is not a great part of the annual budget of most animals. In long-range migrants such as some passerine birds, less than 1% of their annual energy is expended in migratory flight. In more leisurely migrants, this is similar, but the necessary energy can be collected as the migrant travels and thus the process is less demanding.

Long-range migrants typically have high rates of metabolism, great endurance, and a large metabolic scope (ratio of peak metabolic rate to standard or resting metabolic rate). Other factors affecting the ability of migrants to travel long distances include speed of travel and the mass of the migrant.

Migration is least costly for swimming organisms, more energetically expensive for flying organisms, and most costly for animals traveling terrestrially. Large animals have relatively low costs of transport and small animals have large costs (Table IV).

In flying migrants, flight range depends largely on the amount (and possibly chemical composition) of fuel stored prior to migration. The amount of fuel and the style of flight are a compromise among the costs of transportation of reserves, the risks of migration, and the nature of the migration (i.e., short-range intracontinental hops versus long-range transoceanic flights). However, energy demands of flight can be reduced by soaring, which is relatively inexpensive, and by flying with aid of favorable

TABLE IV

Costs of Transport of Some Species

Species	Mass (g)	Energy (J/kg km)
Mosquito	0.0082	0.033
Migratory locust	2.0	8.0
Hummingbird	4.0	16.0
Chaffinch	21.0	16.4
Laughing gull	310.0	6.1
Lemming[a]	61.0	164.0
Sockeye salmon	3.0	2.0
Sockeye salmon	1500.0	4.5

[a] Not really migratory.

winds. Insects fly low within the boundary layers of shelter, use updrafts, or fly with the wind. Birds sometimes use soaring, flying information, or fly with the wind to reduce costs of migration. The monarch butterfly, however, has been observed flying into winds up to 12 km/h. In fact, the volume of migration of insects and birds is often a function of weather patterns. Migrant birds in eastern North America are usually associated with southerly winds in spring and north/northwest winds in autumn and often follow fronts. In some species of birds, migration only occurs over part of the range. In such species, the migratory birds may have longer wings, apparently reducing increased wing loading brought about by fuel loads. Empirical studies indicate that avian metabolic rates increase less rapidly with body mass than does the ability to store and carry lipid reserves, thus providing longer ranges for larger birds.

In birds and insects that cross large distances where refueling is difficult (deserts) or impossible (oceans), mass of reserve is important in several ways. Theoretical and empirical studies both suggest that there is an optimal flight velocity with regard to fuel use. Low speeds (near hovering) are expensive, as are very great speeds (Fig. 2). Furthermore, large flying objects travel less expensively at higher velocities than small ones at lower speeds. This results in greater flight ranges for large birds, even though their energy reserves are not proportionately greater (Fig. 3). Additionally, as travel progresses and body mass decreases as fuel is utilized, the optimal speed becomes lower. Some researchers have observed that migrants travel more slowly as they reach their destination. However, untested assumptions about physiology and aerodynamics may greatly affect estimated flight ranges. In the earliest studies of

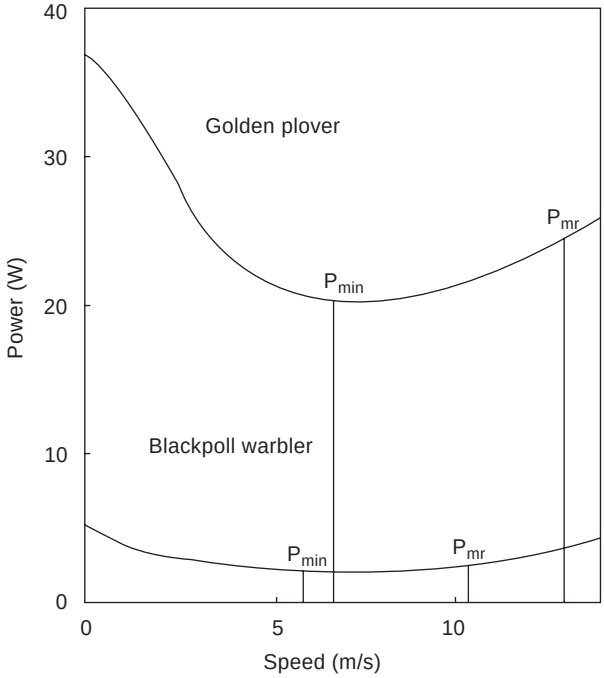


FIGURE 2 Theoretical power curves of two long-range migrants. Lines for minimum power required for flight extend to minimum energy costs (P_{min}); lines for power producing maximum ranges extend to velocities that produce these maximum migratory ranges (P_{mr}). From Blem (1980).

ruby-throated hummingbird migration (*Archilochus colubris*) across the Gulf of Mexico, some investigators doubted that potential flight ranges of hummingbirds were not sufficient to complete the trip. Subsequent measurements of hummingbird flight metabolism indicated that a 4.5-g hummingbird (containing 2 g of fat) could fly nonstop for 26 h (at a cost of approximately 3 J/h). At 40 km/h, its range was estimated at 1050 km, and it is now believed that a trans-Gulf flight is possible by these birds and the available field data support this observation.

8. ENERGETIC BENEFITS OF MIGRATION

Migration is a risky activity. For example, during migration there is a high potential for mortality from predation, and there are predators that appear to seek out migrant insects, birds, and fish. Small birds that are exclusively diurnal throughout their lives migrate at night, ostensibly to avoid predation. Inclement weather may take a toll. There are numerous records of mass mortality caused by storms during passage over large bodies of water.

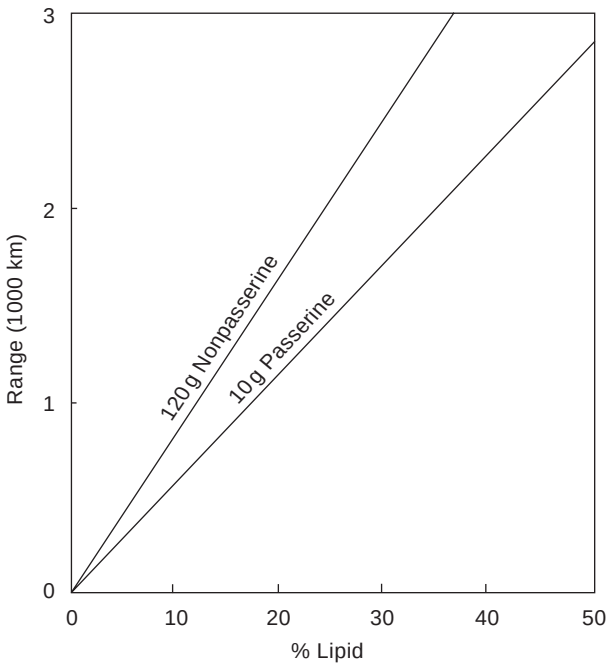


FIGURE 3 Nonstop flight ranges for two avian migrants as a function of their lipid reserves. These curves are often slightly to noticeably concave, which is not apparent here because of the narrow scale of the figure. Lipid content of the migrants is expressed as percentage of total wet body mass. From Blem (1980).

Rafts of dead birds, including kinglets, hummingbirds, and passenger pigeons (many years ago), and several species of insects have been found floating on seacoasts or along the shores of the Great Lakes following violent storms.

Starvation is a real possibility, particularly if the migrant flies or swims in opposing wind/water currents. At some locations, migrant birds are commonly seen in severely emaciated condition. They have little or no lipid, are greatly dehydrated, and are little more than "flying skeletons." They may rapidly perish if they do not quickly find food and water. Finally, simple accidents occur at high rates. In recent years, accidental death from striking towers, buildings, and other man-made structures has become a major source of mortality, estimated to amount to millions of birds each year in North America alone.

There are several possible benefits of migration that offset the risks that accompany extensive travel, including avoidance of predation and parasitism on the breeding grounds and reduced competition at more distant breeding sites. Many migrants appear to obtain energetic benefits from migration because periodic, heterogeneous food resources become available to them. In fact, some of these species literally never experience winter. They obtain the benefits of spring/summer throughout their lives. Some species (e.g., pinnipeds and cetaceans) may migrate only because of rich food sources at one or both ends of the journey. Longer day length on northern breeding grounds may provide energetic benefits for species whose foraging time would otherwise be limited. Tests of the hypothesis that birds or other migrants benefit energetically from the environment of the breeding area have provided both positive and neutral results. For example, the dickcissel (*Spiza americana*), a long-range neotropical migrant, attains a positive energy balance during the breeding season because of the longer photoperiod to which it is exposed in the north. American robins (*Turdus migratorius*) appear to obtain similar benefits. On the other hand, tests of field sparrows (*Spizella pusilla*) and tree sparrows (*Spizella arborea*) indicated no such benefits.

It is worth noting that reproductive output in some species increases with distance of migration. At distant breeding grounds, food may be more available as a result of the presence of fewer competitors. This not only provides obvious energetic benefits but also reduces risks of predation if the migrant needs to forage less during the breeding season. Foraging activity usually makes the animal more vulnerable to predation because its attention is directed elsewhere and it sometimes must move out of seclusion to find food sources. Long day length provides more time for foraging of diurnal animals (many insects and birds) and should provide energetic benefits, but these might be partially offset by relatively low environmental temperatures at northern breeding sites.

The energetic benefits of extended day length and higher ambient temperature at southern latitudes in winter may be a selective factor for fall migration in a variety of animals but are best understood in insects and birds. Furthermore, some species of birds appear to become distributed on their wintering grounds in a manner related to energetics. For example, in several species of small birds, sex ratios vary geographically, apparently because of differences in size and the metabolic requirements of the sexes. Males winter farther north and have larger body masses. Bergmann's rule, an ecogeographic principle used to describe ecotypic variation in vertebrates, suggests that body size increases with latitude in such a way that it provides energetic benefits. Larger individuals may have greater endurance of periods of fasting caused by inclement weather.

SEE ALSO THE FOLLOWING ARTICLES

Conversion of Energy: People and Animals • Food Capture, Energy Costs of • Heterotrophic Energy Flows • Human Energetics • Photosynthesis and Autotrophic Energy Flows • Reproduction, Energy Costs of

Further Reading

Alerstam, T. (1990). "Bird Migration." Cambridge Univ. Press, Cambridge, UK.

Berthold, P. (1996). "Control of Bird Migration." Chapman & Hall, London.

Blem, C. R. (1976). Patterns of lipid storage and utilization in birds. *Am. Zool.* **16,** 671–684.

Blem, C. R. (1980). The energetics of migration. *In* "Animal Migration, Orientation and Navigation" (S. A. Gauthreaux, Ed.), pp. 175–224. Academic Press, Orlando, FL.

Blem, C. R. (1990). Avian energy storage. *Curr. Ornithol.* **7,** 59–113.

Blem, C. R. (2000). Energy balance. *In* "Sturkie's Avian Physiology" (G. C. Whittow, Ed.), pp. 327–343. Academic Press, New York.

Gwinner, E. (1990). "Bird Migration: The Physiology and Ecophysiology." Springer, New York.

Klaassen, M. (1996). Metabolic constraints on long-distance migration in birds. *J. Exp. Biol.* **199,** 57–64.

Lindstrom, A., and Alerstam, T. (1992). Optimal fat loads in migrating birds: A test of the time-minimization hypothesis. *Am. Nat.* **140,** 477–491.

Pennycuick, C. J. (1975). Mechanics of flight. *Avian Biol.* **5,** l–75.

Modeling Energy Markets and Climate Change Policy

HILLARD G. HUNTINGTON and JOHN P. WEYANT
Stanford University
Stanford, California, United States

1. **Alternative Modeling Approaches**
2. **The Key Role of Energy Prices**
3. **Factors Influencing the Response to Climate Policies**
4. **Conclusions**

Glossary

autonomous energy efficiency improvement An exogenous technical change that is unrelated to energy prices or other variables predicted by the model.

foresight The ability of producers or consumers to anticipate future changes in market conditions. Perfect foresight means that they correctly anticipate the future (as predicted by the model); myopic foresight assumes that the conditions tomorrow will be like today's.

general equilibrium The joint solution to the market price for multiple energy and economic sectors.

induced technological change An endogenous technical change, usually resulting from price-induced behavior.

interindustry interactions The relationships between industries that reveal which sectors buy from and which sectors sell to each industry.

process analysis An engineering concept that refers to discrete technologies, each often requiring fixed input combinations.

production function A relationship that links inputs to output and specifies the rate at which each input can be substituted for each other input in response to shifts in input prices.

putty–clay capital stock malleability The assumption that original equipment cannot be modified once installed.

putty–putty capital stock malleability The assumption that both old and new capital can be reconfigured once installed, to fit the current price situation in each time period.

vintage Identifying equipment and its energy and other key characteristics by the year when it is introduced.

Various economic models of energy supply and demand have been used for global climate change policy analysis. Although these models are quite diverse in their structure and focus, all systems determine market-clearing equilibrium prices that balance production and consumption levels for different fuel types. On energy demand, these models (except for MARKAL-Macro) are "top-down" in that they seek to establish aggregate or sectoral relationships between consumption, prices, and economic activity, as opposed to the "bottom-up" assessments that focus on specific equipment and their energy use. This comparison focuses on the energy supply and demand adjustments that significantly influence the costs of various strategies for limiting carbon emissions, although the models also incorporate reduced-form representation of the climatic effects that determine the benefits of abatement.

1. ALTERNATIVE MODELING APPROACHES

A number of different economic models has been applied to the global climate change problem; the focus here is on 16 modeling systems by teams, most of whom have recently participated in model-comparison studies conducted by Stanford University's Energy Modeling Forum. The models are identified in Table I along with their principal investigators and sponsoring organization. Half of these teams are based in the United States and half outside of it. The model structures discussed herein appear to be most important for determining the effects of climate policies. Weyant has provided comprehensive discussions of the many major findings or insights from applying these models to this issue.

 © 2004 Elsevier Inc. All rights reserved.

TABLE I

Models Analyzing Post-Kyoto Energy Modeling Forum Scenarios

Model	Principal investigators
ABARE–GTEM (Australian Bureau of Agriculture and Resources Economics–Global Trade and Environment Model)	B. Fisher, V. Tulpule, D. Kennedy, and S. Brown (ABARE)
AIM (Asian–Pacific Integrated Model)	T. Morita, M. Kainuma (NIES, Japan), and Y. Matsuoka (Kyoto University)
CETA (Carbon Emissions Trajectory Assessment)	S. Peck (Electric Policy Research Institute) and T. Teisberg (Teisberg Assoc.)
FUND (Framework for Uncertainty, Negotiation, and Distribution)	R. Tol (Vrije Universiteit, Amsterdam)
G-Cubed (Global General Equilibrium Growth)	W. McKibben (Australian National University), P. Wilcoxen (University of Texas), and R. Shackleton (U.S. Office of Management and Budget)
GRAPE (Global Relationship Assessment to Protect the Environment)	A. Kurosawa (Institute of Applied Energy and Research Institute of Innovative Technology for Earth, University of Tokyo)
IGEM (Intertemporal General Equilibrium Model)	D. Jorgenson (Harvard University), P. Wilcoxen (University of Texas), and R. Goettle, M. Sing Ho, and D. Slesnick (Dale W. Jorgenson Associates)
MARKAL-Macro	S. Morris (Brookhaven National Laboratory), A. Manne (Stanford University), and P. Tseng (U.S. Department of Energy)
MERGE 3.0 (Model for Evaluating Regional and Global Effects)	A. Manne (Stanford University) and R. Richels (Economic Policy Research Institute)
MIT-EPPA (Massachusetts Institute of Technology Emissions Projection and Policy Analysis)	H. Jacoby/J. Reiner (MIT) and I. Sue Wing (MIT)
MS–MRT (Multi-Sector–Multi-Region Trade)	D. Montgomery/P. Bernstein (Charles River Assoc.) and T. Rutherford (University of Colorado)
NEMS (National Energy Modeling System)	R. Earley, S. Holte, M. Hutzler, A. Kydes, R. Eynon, *et al.* (U.S. Energy Information Agency)
Oxford Econometrics Model	Adrian Cooper and John Walker (Oxford Econometrics)
RICE (Regional Integrated Climate and Economy)	W. Nordhaus and J. Boyer (Yale University)
SGM (Second Generation Model)	J. Edmonds, H. Pitcher, and R. Sands (Pacific Northwest National Lab)
WorldScan	A. Gielen/H. Timmer (Central Planning Bureau, Netherlands) and J. Bollen (Rijksinstituut voor Volksgezondheid Milieuhygiene, Netherlands)

Although each model has unique characteristics and has proved to be extremely valuable for studying certain types of issues, the structures of the models can be described in terms of five basic categories shown in Table II. Many of the models now employ combinations of traditional modeling paradigms. One category of models focuses on carbon as a key input to the economy. These models do not distinguish between energy types and therefore are similar to models of aggregate energy use. Trends toward less carbon-intensive fuels are incorporated into their projections for carbon. They consider the cost of reducing carbon emissions from an unconstrained baseline by using an aggregate cost function in each country/region. This approach uses a simple vintaging structure to incorporate the time lags in reducing carbon intensity in response to increases in the price of carbon. In these models, all industries are aggregated together, and gross domestic product (GDP) is determined by an aggregate production function with capital, labor, and carbon inputs. These models generally omit interindustry interactions, include trade in carbon and carbon emissions rights but not in other goods and services, and assume full employment of capital and labor. The Regional Integrated Climate and Economy (RICE) and Framework for Uncertainty, Negotiation, and Distribution (FUND) models are examples of this category of models.

Another closely related category of models focuses heavily on the energy sector of the economy. These models consider the consumption and supplies of fossil fuels, renewable energy sources, and electric power generation technologies, as well as energy prices and transitions to future energy technologies. In general, they explicitly represent capital stock turnover and new technology introduction rate constraints in the energy industries, but take a more

TABLE II
Model Types

Economy model	Fuel supplies and demands by sector	Energy technology detail	Carbon coefficients
Aggregate production/cost function	—	CETA, MARKAL-Macro, MERGE 3.0, NEMS, GRAPE	FUND, RICE
Multisector general equilibrium	MIT-EPPA, WorldScan	ABARE–GTEM, AIM, MS–MRT, SGM	—
	G-Cubed, IGEM[a]	—	—
Multisector Macroeconometric	Oxford Econometrics Model	—	—

[a] Models combine multisector general equilibrium with multisector macroeconometric approaches.

aggregated approach in representing the rest of the economy. In these models, all industries are aggregated together, and GDP is determined by an aggregate production function with capital, labor, and energy inputs. These models generally omit interindustry interactions and assume full employment of capital and labor. The Model for Evaluating Regional and Global Effects (MERGE 3.0), Carbon Emissions Trajectory Assessment (CETA), MARKAL-Macro, National Energy Modeling System (NEMS), and Global Relationship Assessment to Protect the Environment (GRAPE) paradigms are examples of this category of models. MERGE 3.0 and CETA have the same basic structure, with nine and four regions, respectively. GRAPE includes a somewhat broader set of technology options, including especially carbon sequestration technologies.

A third category of models is those that include multiple economic sectors within a general equilibrium framework. They focus on the interactions of the firms and consumers in various sectors and industries, allowing for interindustry interactions and international trade in nonenergy goods. In these models, adjustments in energy use result from changes in the prices of energy fuels produced by the energy industries included in the interindustry structure of the model (e.g., coal, oil, gas, and electricity). Explicit energy sector capital stock dynamics are generally omitted. These multisector general equilibrium models tend to ignore unemployment and financial market effects. The Intertemporal General Equilibrium Model (IGEM), the Massachusetts Institute of Technology Emissions Projection and Policy Analysis (MIT-EPPA) model, and the WorldScan model are examples of this type of model. The Global General Equilibrium Growth (G-Cubed) model does consider some unemployment and financial effects and is, therefore, a hybrid general equilibrium/macroeconometric model. G-Cubed, MIT-EPPA, and WorldScan all include trade in nonenergy goods.

A fourth basic class of models combines elements of the first two categories. They are multisector, multiregion economic models with explicit energy sector detail on capital stock turnover, energy efficiency, and fuel-switching possibilities. Examples of this type of hybrid model are the Asian–Pacific Integrated Model (AIM), Australian Bureau of Agriculture and Resources Economics–Global Trade Environment Model (ABARE–GTEM), Second Generation Model (SGM), and Multi-Sector–Multi-Region Trade (MS–MRT) model. These models include trade in nonenergy goods, with AIM including energy end-use detail, GTEM and MS–MRT including some energy supply detail, and the SGM considering five separate supply subsectors to the electric power industry.

By including unemployment, financial markets, international capital flows, and monetary policy, the Oxford model is the only model included here that is fundamentally macroeconomic in orientation. However, as shown in Table II, the G-Cubed and IGEM models do consider some unemployment and financial effects, as well as international capital flows.

2. THE KEY ROLE OF ENERGY PRICES

Economic models assign an important role to the price of energy in determining the economy's adjustment to climate change policies. The models compute the price of carbon that would be required to keep emissions levels controlled at some predetermined level. Although it is easiest to think about these scenarios as carbon taxes or permits (which may or may not be tradable), the additional carbon

costs also reveal important information about any program that seeks to reduce carbon emissions.

Most economic models solve a set of mathematical equations to obtain the prices of goods and services and key inputs, including different forms of energy. The simultaneous solution of these equations represents an equilibrium in which supply equals demand among consumers and producers. In this sense, these models generally determine fuel prices endogenously by searching for those prices that will balance the supply and demand for each fuel. This approach is very valuable for policy analysis of control strategies that will affect the supply and consumption of different fuels. In this framework, an energy price increase can be either the motivation for, or the result of, greenhouse gas (GHG) emissions reductions. For example, governments may impose emissions taxes to motivate GHG reductions. Emissions taxes increase the costs of fuels directly, and economies will adjust to reduce the use of those higher cost fuels, substituting goods and services that result in fewer GHG emissions. On the other hand, governments may cap the total amount of emissions, distribute or sell emissions "allowances," and let the market determine the price and distribution of these allowances. Such a "cap and trade" system will induce changes in prices that are difficult to predict. Because a cap would essentially restrict the supply of carbon-based fuels, GHG consumers would bid up the price until demand for such fuels no longer exceeded supply. In this way, the higher prices reduce emissions, but also allocate the GHGs to their highest value uses.

The effects of higher fossil fuel prices would diffuse throughout the economy. Prices of secondary energy sources, such as electricity and oil products, would rise as the higher primary fuel costs are passed through into electricity rates, fuel oil, and gasoline prices. Higher fuel costs would also increase operating costs in transportation, agriculture, and especially industry. Although energy costs make up only 2.2% of the total costs in U.S. industries, they constitute up to 25% of the total costs in the most energy-intensive sectors (e.g., iron and steel, aluminum, papermaking, and chemicals). Each industry's ability to pass these cost increases along to customers through higher product prices would depend on the strength of the demand for its products and on the severity of international competition. Because many of the major trading partners of the United States would also be implementing similar climate policies, it is likely that the energy cost increase would result in higher prices for a broad range of consumer products. Households could also be affected through increased heating, transportation, and utility bills and, to a lesser degree, food bills and other costs of living.

A host of adjustments by producers and consumers in the economy would take place in parallel with the price increases, and, in fact, these substitutions would also serve to limit the extent of the price increases that would ultimately result. Higher energy costs would induce firms to accelerate the replacement of coal-based or obsolete plants with more energy-efficient or less carbon-intensive equipment. Utilities and their customers would seek alternatives to carbon-intensive coal-fired power plants, stimulating the market for hydro, nuclear, gas-fired, and renewable electricity sources. As coal prices rise relative to natural gas prices, modern gas-fired combined-cycle power plants would become even more competitive. Older, less efficient coal-fired plants would probably be retired from service or reserved for intermittent operations. Energy-intensive industries would also face a number of adjustment decisions: whether to retire obsolete facilities and concentrate production at more modern, low-cost facilities; whether to modify their furnaces to burn gas instead of coal; whether to generate their own electricity; whether to invest in a wide variety of energy-conserving process changes; whether to redesign products to save energy; and whether to alter their product mix. Ultimately, there would be an effective diminution in the value of the existing stock of plant and equipment because it is optimized for the set of input prices that prevailed when it was installed and would be suboptimal for the new price regime.

In the short run, consumers and producers would reduce their energy consumption by either consuming less energy services (for example, turning their thermostats down or driving their automobiles less) or producing less output. Consumers and producers may also, potentially, reduce energy use without reducing output by identifying energy efficiency measures previously believed to be uneconomic. In the intermediate time frame, there might be opportunities for fuel switching (or substitutions between other inputs) that would not involve substantial outlays for new equipment or infrastructure (for example, switching the fuel used in a multi-fuel-capable boiler from oil or coal to gas). In addition, consumers may be able to substitute goods that require less energy to produce (which would become relatively less expensive) for more energy-intensive ones (which would become relatively more expensive). In the long term, new technologies would be purchased that either use less GHG-intensive fuel or are more fuel efficient. In addition, new, less GHG-intensive technologies might become available over

time as a result of research and development expenditures or cumulative experience. The emergence of these new technologies might be related to the energy price increases, the base case trend of all other prices, or simply the passage of time.

Higher energy prices would lead to less energy use, and less energy use would decrease the productivity of capital and labor. These productivity changes would, in turn, generally result in a slowdown in the accumulation of capital equipment and infrastructure, and in lower wages for workers. Ultimately, even after all the adjustments have been made, consumers would have somewhat less income, which might cause them to adjust the amount of time they spend on work rather than leisure. The resulting adjustment in labor depends on two opposing effects. Workers would want to work more to make up for their loss in real income, but the lower wage would make working worth less relative to leisure. Replacing work with leisure would involve an additional change in welfare. Offsetting these welfare losses would be the benefits of reduced climate change and the benefit of making those responsible for GHG emissions pay for the damages they cause.

The complicated web of economic adjustments that would take place in response to rising prices of energy, or energy scarcity, makes the task of projecting the costs of GHG mitigation a challenge. Interpreting the results they produce is further complicated because different modeling systems emphasize different dimensions of the adjustment process. Also, different policymakers may be interested in different policy regimes, and in different impacts of climate change and climate change policies.

3. FACTORS INFLUENCING THE RESPONSE TO CLIMATE POLICIES

Baseline economic, technological, and political conditions, the opportunities to substitute away from fossil fuels, the nature of the capital stock turnover process, and the dynamics of technological progress are four very important factors that determine a model's response to climate change policies.

3.1 Baseline Conditions

One set of important issues is the base case emissions and climate impact scenarios, against which the costs and benefits of GHG mitigation policies are assessed. They are largely the product of assumptions that are external to the analysis. Each GHG mitigation cost analysis relies on input assumptions in three areas:

- Population and economic activity.
- Energy resource availability and prices.
- Technology availability and costs.

Most of the researchers projecting the cost of reducing carbon emissions have relied on worldwide population growth projections made by others (e.g., the World Bank or the United Nations). These external projections are generally based on results from very simple demographic models. There is less uncertainty about projections for the developed countries, where population is expected to peak very soon, than for the developing countries, where population is typically assumed to peak somewhere around the middle of this century. Very few of the researchers analyzing GHG emissions reductions make their own projections of economic growth. Most rely on economic growth projections made by others, or on external assumptions about labor force participation and productivity growth.

Another key set of assumptions concerns the price and/or availability of energy resources. The prices of fossil fuels (oil, natural gas, and coal) are important because producers and consumers generally need to substitute away from these fuels when carbon emissions are restricted. Optimistic assumptions about natural gas availability and/or substitutability can make carbon emissions reductions easier to achieve in the short run. Natural gas plays an important role because its carbon emissions are about 60% of those from coal, and 80% of those from oil, per unit of energy consumed. In addition, the amount of unconventional oil and gas production that will ultimately be technically and economically feasible is highly uncertain. It depends on future economic incentives for oil and gas exploration and production, which could (absent climate policies) retard the development of carbon-free renewable and higher efficiency end-use energy technologies. How oil exporters would react to a climate policy that would reduce the demand for oil imports is another key dimension of the energy supply picture.

A final set of key assumptions includes those made about the costs and efficiencies of current and future energy-supply and energy-using technologies. These factors tend to be critical determinants of energy use in both the base case and control scenarios. Most analysts use a combination of statistical analysis of historical data on the demand for individual fuels and process analysis of individual technologies in use or under development, in order to represent trends in

learning by doing (past experience with the technology) and its effect on reducing costs as well as other characteristics of energy technologies. Particularly important, but difficult, is projecting technological progress within the energy sector. Jorgenson and Wilcoxen have attempted to estimate systematically and empirically future trends in energy productivity at a national level, but such efforts are rare. Typically, analysts take one of two approaches: (1) the costs and efficiencies of energy-using and energy-producing technologies are projected based on process analysis, and the characteristics of these technologies are extrapolated into the future, or (2) the trend in energy demand per unit of economic activity, independent of future price increases, is assumed. Some recent analyses have attempted to blend the two approaches. At some point, these two approaches tend to converge, because the end-use process analyst usually runs out of new technologies to predict. It is then assumed that the efficiency of the most efficient technologies for which there is an actual proposed design will continue to improve as time goes on.

Projections of the benefits of reductions in GHG emissions are also highly dependent on the base case scenario employed. The greater the base case damages (i.e., the damages that would occur in the absence of any new climate policies), the greater the benefits of a specific emissions target. The magnitude of the benefits from emissions reductions depends not only on the base case level of impacts, but also on where they occur, and on what sectors are being considered. In fact, a number of additional socioeconomic inputs (e.g., income by economic class and region; infrastructure and institutional capability to adapt to changes) are required because they determine how well the affected populations can cope with any changes that occur. The task of projecting base case climate change impacts is particularly challenging because (1) most assessments project that serious impacts resulting from climate change will not begin for several decades and (2) most of the impacts are projected to occur in developing countries, where future conditions are highly uncertain. How well developing countries can cope with future climate change will depend largely on their rate of economic development.

3.2 Representation of Substitution Possibilities

As efforts are made to reduce GHG emissions, fossil fuel combustion and other GHG-generating activities become more expensive. Producers adjust to these price increases by substituting inputs (i.e., switching to inputs that generate fewer GHG emissions in manufacturing any particular product) and by changing their product mix (i.e., producing different products that require less GHG emissions to make). The extent to which inputs can be shifted depends on the availability and cost of appropriate technologies as well as on the turnover rate of capital equipment and infrastructure. These two factors, as well as consumer preferences, determine an industry's ability to produce and sell alternative mixes of products. Increases in the costs of fossil fuels and products that depend on fossil fuel combustion will reduce consumers' real incomes. Consumers will simultaneously decide (1) the extent to which they wish to adjust their mix of purchases toward less carbon-intensive products and (2) how to adjust their mix of work and leisure time to compensate for the reduction in their real income.

3.2.1 Short-Term vs. Long-Term Substitution

If businesses and households have several decades to complete the substitution process, the current stocks of energy equipment and associated infrastructure do not constrain the substitutions that they may make. Businesses and households are limited primarily by the available technologies and by their own preferences regarding how much of each available product they would buy at the prevailing prices.

If climate policy is long term, the transition to a lower carbon energy system can be relatively smooth and the costs relatively moderate. To reach such an outcome, economic incentives should be designed to motivate producers and consumers to invest in more energy-efficient and less carbon-intensive equipment when their existing equipment has reached the end of its useful life. Useful life is an economic concept that compares the costs of operating existing equipment with the costs of purchasing and operating new equipment. A new and better computer may be purchased after 3 years, even though the old computer could be "useful" for 10 years, because the new one has superior cost and performance characteristics. Or an old car may be kept running because the performance advantage of the new car is not worth the cost.

Over shorter time spans, however, existing plant and equipment can significantly constrain the behavior of firms and households, adding transition costs to the long-run costs of GHG control policies. Policies implemented on this time scale (i.e., within 10 years) will lead to reductions in energy services

(e.g., industrial process heat and home heating and cooling), some easy fuel switching, and an increase in the purchase and use of available energy-efficient products and services. They will also influence the rate of retirement and replacement of existing equipment. Energy-producing and energy-using goods are relatively expensive and long-lived. Thus, it will generally take a substantial increase in energy prices to induce those who own such equipment to replace any of it before the end of its useful life.

The importance of capital stock dynamics creates a formidable challenge for the analytical community. Some data on the characteristics of the energy-producing and energy-using capital stock are available. It would be ideal to have information on the costs of operating and maintaining every piece of equipment currently in use. This would enable analysts to calculate all the trade-offs between retiring equipment early and using other strategies to achieve the specified targets. Unfortunately, the data that are available are generally aggregated across large classes of consumers and generally include all existing capacity without regard to when it was installed. An important exception is power plant data, which are very disaggregated and include the age of the equipment. However, even these data are generally not sufficient to ascertain precisely the point at which the carbon price incentives will influence the rate of replacement of plant and equipment. Limitations on data require the analyst to make a number of assumptions regarding the aggregation and interpretation of the available data.

3.2.2 Two Approaches to Representing Substitution Possibilities

In many models, technologies are represented with "production functions" that specify what combinations of inputs are needed to produce particular outputs. The production function specifies the rate at which each input can be substituted for each other input in response to shifts in input prices. As new capital investment occurs and older capital is retired, the technology mix within the model will change.

Two basic types of production functions may be specified: aggregate production functions and technology-by-technology production functions, also known as process analysis. Some models (e.g., G-Cubed, SGM, and EPPA; see Table I for model identification) use smooth and continuous aggregate production functions that allow incremental input substitutions as prices change, even if the resulting input configuration does not correspond to a known technology. These models do not represent individual technologies. Such models often assume "nested" production functions. For example, at one level, substitutions are possible between energy, capital, and labor in producing final commodities; at a second level, substitutions are possible between electricity and fuel oil in producing energy; and, at a third level, substitutions are possible between coal and natural gas in producing electricity. Models employing continuous aggregate production functions do not account for individual technologies.

In contrast, other models (e.g., MARKAL-Macro and NEMS) draw from a menu of discrete technologies, each requiring fixed input combinations—i.e., each technology is essentially represented with its own production function. This approach is often referred to as "process analysis." These combinations correspond to those employed in actual, or anticipated, technologies that the modeler specifies. The technology-rich MARKAL-Macro model specifies over 200 separate technologies. For discrete technology models, different technologies become cost-effective as input prices change. Modelers then assume that these technologies are selected and used to produce outputs. Process analysis represents capital stock turnover on a technology-by-technology basis. The data and analysis requirements for this type of model can be substantial.

A number of systems use a process analysis approach within the energy sector and an aggregate production approach for the remainder of the economy (e.g., MERGE, MARKAL-Macro). When using either approach, it is important to be able to distinguish between the causes of changes in the selections the models make among the existing technologies. Sometimes the technology choice changes because prices change, and sometimes it changes because new technologies become available.

Some models represent both individual energy supply technologies and individual energy consumption technologies, and do not represent the remainder of the economy explicitly. With these models, however, the analyst must either (1) assume that "end-use" energy demands (such as the demand for home heating and automotive transport) do not respond to changes in the prices of those services or (2) employ a complex statistical estimation technique (which requires some historical data on the cost of end-use energy equipment) to estimate the price responsiveness. The choice of production function depends, in part, on the time frame under consideration and the level of technological disaggregation. Short-term models intended to shed light on precise technology choices specify production functions for

large numbers of separate technologies. In contrast, models concerned with longer term effects can safely characterize technological trends using aggregate production functions. Many models blend the two approaches. Although they allow for smooth input substitution in determining new capital investment, they fix input proportions for all equipment installed in a certain year (sometimes called a "vintage"). Similarly, a model may have smooth production functions for conventional fuels, yet stipulate discrete technologies for a particular noncarbon fuel (e.g., EPPA).

3.2.3 Capital Stock Turnover and Malleability

In modeling capital investment in plant and equipment and turnover, each system must use assumptions about the flexibility the investor has in choosing technologies and in changing their characteristics after installation. Data availability and computational considerations limit the choice of modeling assumptions that can be employed. Fortunately, there are some simple formulations that seem to give plausible results in most circumstances.

In almost all models, it is assumed that in making decisions about new capital investment, the decision maker (firm, individual, or government entity) has complete flexibility (particularly in the mix of capital and energy inputs required) in choosing among available technologies before their purchase. The models differ, however, in their assumptions about how much the characteristics of the capital equipment can be changed after it has been installed. These adjustments may be desirable if changes in input prices occur, but retrofitting to a certain set of characteristics is generally more expensive than installing equipment with the same characteristics initially. On the other hand, technological improvements may reduce the costs of the retrofitting over time.

Most models make one of two polar assumptions about this process. To describe these assumptions, the metaphor of soft putty and hardened clay has proved useful ("putty" representing a flexible scenario and "clay" representing a hardened or inflexible scenario). In a putty–clay or putty–putty formulation, the first term refers to the assumption about the degree of flexibility in original capital investment, and the second term refers to the assumption about the degree of flexibility in modifying that capital after it is installed.

- In a putty–clay formulation, it is assumed that the original equipment cannot be modified once installed. Putty–clay assumptions are more realistic in cases in which relative prices are changing rapidly. Here, new capital investments embody state-of-the-art technology and use input mixes that are appropriate for the price expectations that exist at the time of the investment. These characteristics then remain with that vintage until it is scrapped.
- In a putty–putty formulation, it is assumed that capital—old or new—can be reconfigured once installed to fit the current price situation in each time period. Under the so-called putty–putty assumption, the capital stock is a single entity that is neither broken down into separate vintages nor constrained to retain its initial technology and input mix. The term "putty–putty" is used to indicate that capital can be continuously reshaped both before and after investment has taken place. The inherited capital stock adjusts to changes in prices and technology as fully as brand new capital. In effect, the entire capital stock continually adapts itself to reflect current technologies and prices.

The precise details of the capital adjustment process differ from model to model. In some, there is a composite stock of old capital that reflects some average mix of inputs. In others, each vintage is identified and depreciated separately. In many models, the old capital stock cannot be altered. In others (e.g., NEMS), it can be retrofitted if doing so is more profitable than making brand new investments, or if it is required by regulation.

Modelers are just starting to experiment with various hybrids of the two formulations, i.e., putty–semiputty formulations, in which some retrofitting is allowed at some additional cost. One type of putty–semiputty specification allows plant and equipment to be retired before the end of its useful life if the operating cost of the old equipment is greater than the operating plus capital costs of replacement equipment. In this case, the remaining capital costs of the old equipment would have to be written off, so the changes in prices or new technologies would have to be quite significant for this to occur. Prices do rise to these levels in some models in Kyoto Protocol simulations in which the flexibility mechanisms are severely restricted.

3.3 Capital Stock Adjustment Process

Jacoby and Wing have described three characteristics of these models that are important in analyzing the time horizon for meeting the Kyoto targets: the time frame, the level of detail about capital stock and

production structure, and the specification of economic foresight. The first and most obvious attribute is the time interval over which a model solves its equations. If a model uses a 10-year time interval, the relatively long time period limits the model's ability to be used in analyzing phenomena occurring within a decade, such as the consequences of accepting a 2008–2012 Kyoto target after the year 2000. The results of such models may thus obscure important short-run dynamics of adjustment. The second important attribute of the models is the level of aggregation in the capital stock and the production structure. The level of aggregation affects how models represent the sources of rigidity in the production sectors of the economy. For example, the choice about whether to aggregate output and capital by sector or by technology determines the degree of substitution that is possible within the model's structure. Within a specific aggregate, substitutions are, by construction, assumed to be costless. Additional capital stock produces outputs using a combination of inputs that reflects (1) current and expected input prices and (2) the constraints and limits of existing technologies.

Models capture the aging of capital in different ways. In evaluating short-term adjustment to climate policies, the distinction between putty–putty and putty–clay specifications is critical. In the face of a stringent near-term policy, the putty–putty assumption may produce unrealistic results because this specification implies that large parts of the current capital stock can be transformed into more efficient and less carbon-intensive alternatives. However, for analysis of the long run, after fuel prices have settled at a new equilibrium level relative to other goods and services, the distinction is less important. In this postadjustment phase, the inherited capital stock will be increasingly fuel efficient and competitive under prevailing conditions, because those conditions will more closely match the conditions in place at the time the investments were made.

The third important characteristic of models of the capital stock turnover process is the way they treat foresight. Models may specify economic behavior as forward-looking or myopic. Forward-looking models assume that agents with perfect foresight find the path of emissions reductions that minimize discounted costs over the entire modeling horizon, choosing the timing and stringency of control measures so as to smooth optimally the costs of adjustment. In contrast, myopic models assume that economic agents seek to minimize the costs of policy on a period-by-period basis, and take little or no action in advance of the onset of carbon constraints. Model results can be very sensitive to assumptions about investor foresight. Models that assume perfect foresight allow emissions targets to be met at lower costs because investment decisions are made in the full certainty that emissions limits will be set and achieved. Models that assume some degree of myopia generate higher costs because investors must scramble to alter the capital stock as the target period approaches, prematurely scrapping existing capital (e.g., coal-fired power stations) and quickly investing in less carbon-intensive alternatives.

Of the models reviewed here, a great majority assume perfect foresight, whereas only one is constrained to be myopic (EPPA). Some models (e.g., G-Cubed) allow alternative assumptions under different runs and/or can set expectations differently for different sectors. The NEMS and SGM models can allow industrial or utility investors to give greater consideration to future conditions than individual consumers do.

In practice, investors do not have perfect foresight, nor do they suffer from complete myopia. Although there is inevitable uncertainty regarding future economic conditions, policymakers can reduce uncertainties by making credible commitments to meet targets or to initiate market-based policies. Model results clearly demonstrate that the more convinced investors are that emissions targets will become binding, the less costly the transition to lower carbon emissions.

3.4 Technological Change

How these opportunities will change with time and with people's experience with new technologies also have important effects. Technological change can be thought of as increasing the amount of a product that can be produced from a given amount of inputs, or as expanding the universe of opportunities for substitution of inputs and products that were described in the last section. Technological change is discussed separately from input and product substitution here because the underlying determinants are somewhat different, because technological change is less well understood, and because of the opportunities for synergy between public support and private investment in stimulating new technology development.

In 1942, Schumpeter identified three distinct types of technological change that take place continually in modern economies: (1) invention of completely new ways of satisfying human needs and wants, or the creation of new needs not previously identified or

satisfied, (2) innovation, which takes place through continual improvement and refinement of existing ways of doing things, and (3) diffusion of new technologies throughout and across economies. These processes are all important for climate policy. It often takes decades for innovation and invention to pay off. Even diffusion may be difficult to accelerate over a decade, though, because it takes time to distribute information, analysis, and experience from one user to another.

New technologies can allow firms to produce a particular product using a mix of inputs not previously available, including, for example, less energy. In addition, new technologies can lead to new products. These new products compete with existing products, with further implications for carbon emissions reduction policies. If these new technologies and new products produce less carbon, then carbon emissions will be lower, fewer emissions reductions will be needed, and/or emissions reductions will be less expensive. Projecting how technological change might progress over time, both with and without climate policies, is challenging. The processes by which technological change occurs are very complex and the data required to estimate how these changes have been made in the past are generally not available. However, there are several ways economic models represent technological change.

3.4.1 Induced Technological Change

Inventions of productive technologies or processes are, by their very nature, hard to predict. However, past experience has shown that they can be revolutionary enough to justify large expenditures in basic research in strategic areas. Innovations could be of great help in lowering the costs of reducing GHG emissions. Thus, it would be worthwhile to find an appropriate combination of government interventions and private sector incentives that encourage innovation. Thus far, however, most of the policy debate on the influence of technological change on climate change policy has focused not on technology policy options, but rather on how restrictions on GHG emissions reduce aggregate mitigation costs over time. This latter effect has been labeled "induced technological change" (or ITC for short). ITC has to do with price-induced behavior—i.e., what private firms will do in response to higher prices. It does not incorporate what firms will do anyway in trying to become more competitive through investing in research and development, or what they would do in response to government sponsorship of research and development or other direct government technology policies. There has been a good deal of discussion about the potential for induced technological change to substantially lower, and perhaps even eliminate, the costs of CO_2 abatement policies. These discussions have exposed very divergent views as to whether technological change can be induced at no cost, or at some cost.

Every ITC model must represent some incentive to induce technical change in one or more ways, such as the following examples:

1. The form of profits from innovations, as in the top-down models, which focus on the behavior of economic aggregates rather than the behavior of individual actors or the use of individual technologies.
2. At a more aggregate and abstract level, by means of cost-functions, research and development production functions, or empirical estimates. Similarly, the decision maker(s) considered may be either decentralized industries, representative firms, or a central planner.
3. By the inclusion of intrasectoral knowledge spillovers that are advances that individual firms within a sector cannot keep to themselves. For example, the level of investment may be determined by the rate of return the firm expects to earn on the research and development investment as compared with other available investment opportunities. However, the rate of innovation may far exceed that implied by the rate of return alone because other firms in the industry may be able to replicate the innovation.
4. By the dimension in which technological change is assumed to progress (i.e., new products or processes, substitution of inputs, or reorganization of production and distribution arrangements).

Some ITC models are based on empirical observations of past responses to energy price and policy changes. One advantage of this type of model is that different sectors may exhibit different rates of technological progress. However, only one model, IGEM, estimates all these parameters simultaneously because of the large amount of data necessary and the heavy computational burdens of such estimations. Another advantage is that this type of model implicitly takes into account real-world factors that are relevant to technological change and that are difficult to incorporate into conventional economic frameworks. Thus, this model relies on empirical observations of real events, not on a simplified representation of the phenomenon. All types and sources of short-term technical change are included. One disadvantage of this aggregation,

though, is that the approach may omit specific known technologies that are beginning to be introduced but that are not yet revealed in the available data. In addition, information about the underlying costs of research and development is lost. Also missing is explicit attention to how firms determine their research and development investments. Firms take into account both the cost of engaging in research and development and the expected benefits in terms of future profitability. Thus, models are unable to evaluate optimal policies with full consideration of the costs of research and development. Another disadvantage is that the model is as limited as the data set from which it is constructed. Only one historical path can be observed, and it is assumed that tomorrow's economy will respond to energy price changes in the same way as yesterday's economy. Thus, long-term technological change is beyond the feasible reach of this type of model. "Long-term" here refers to periods over which substantial technological development and major inventions may occur.

Nonetheless, empirical modeling of ITC may be valuable for short- to medium-term projections, or for estimating the short- to medium-term cost of policies on the economy. Empirical models may also be valuable in comparing or calibrating short-term projections from other types of ITC models. Also, the consideration of ITC helps clarify two key matters of debate: (1) whether prior studies (without ITC) have overstated the cost of achieving given emissions reduction targets and (2) the optimal size and timing of a carbon tax.

3.4.2 Autonomous Energy Efficiency Improvement

In contrast to the ITC models, many models include exogenous technical change. ("exogenous" can mean external to the model, or independent of price, or both.) A simple characterization of technological improvement, employed in many of the models, is a single scaling factor—the autonomous energy efficiency improvement (AEEI)—that makes aggregate energy use per unit of output decline over time, independent of any changes in energy prices. (Many modelers specify the AEEI as a percentage of gross domestic product growth, so that the value changes over time.) Although the definition of the AEEI varies from model to model, in all models it implicitly represents the effect of technological progress. In some models, it also represents one or both of two additional trends: (1) changes in the structure of the economy, resulting in a shift in the relative contribution of energy-intensive industry output to total economic output, and (2) an improvement in energy efficiency over time, reflecting the gradual removal of market barriers that prevent some energy consumers from choosing more efficient energy technologies.

Although the AEEI approach allows for energy improvements over time, it is limited in two respects. First, using the AEEI approach to represent technological change ignores price-induced technological progress (ITC). In reality, higher prices do spur greater innovation and more rapid diffusion of energy-saving technologies. Second, it is not clear what an appropriate rate for AEEI should be. This is important, especially for longer term projections, which are very sensitive to differences in assumed rates. More sophisticated specifications (often used in conjunction with an AEEI parameter) attempt to paint a more detailed picture of technological change by incorporating some degree of price sensitivity, distinguishing different sectors, and assessing changes to specific technologies.

3.4.3 Learning by Doing

In practice, much technological advance comes from learning by doing (LBD), which is the incremental improvement of processes through small modifications and adjustments. It is not until a technology is actually used that important lessons are learned that can be applied to its subsequent development. LBD is an integral part of the innovation process. Observations of past technological innovations show that initial installations are quite expensive, but that costs drop significantly the more the technology is used, and the more lessons are learned from using it. This type of learning may be the result of either exogenous or endogenous (induced) technological change.

The LBD approach does not reveal how learning occurs and whether the learning is associated with invention, innovation, or diffusion. Thus, it cannot evaluate which policies might be appropriate for increasing the learning associated with a technology. The approach also suffers from its inability to establish whether future cost reductions result from increased cumulative experience with the technology or whether they occur with the passage of time, which is closely associated with cumulative experience.

Although most models do not attempt to capture LBD, two models do mimic the process. MERGE assumes endogenous diffusion rates: the more investment there is in advanced technologies in the early years of the projection, the greater is the rate of

adoption in the later years. In the NEMS model, learning by doing is represented in the electricity generation sector, where the capital costs of particular types of new plants decline as more such plants are built.

4. CONCLUSIONS

As long as climate change policies are geared toward keeping a dynamic economy tied to some historical benchmark emissions level, projections of baseline economic and emissions conditions will dominate the costs of emissions control strategies. An equally important consideration is the adopted policy regime, such as the extent to which international emissions trading is permitted. In general, the more flexibility permitted as to where, when, and which GHG reductions may be used to satisfy a commitment, the smaller the economic impacts. In addition to these baseline and policy assumptions, the model structures also influence the cost estimates associated with climate change actions. Especially important are how each model's structure accounts for the rate and extent to which available inputs and products can be substituted for one another and the rate of improvement in the substitution possibilities over time (i.e., technological change). The representation of the substitution possibilities depends both on the available technologies and on how the retirement of existing equipment and the introduction of new technologies are represented. The more flexibility the model includes in the choice of technologies, retirement of old equipment, and introduction of new technologies, the lower the economic impacts of emissions reductions.

Technological change occurs when new technologies allow a particular good or service to be produced with fewer inputs, or when a new product is developed. Most models used to project GHG emissions and mitigation costs assume that technological change takes place steadily over time, but does not depend on changes in prices or the implementation of government policy options. Thus, different technologies are selected as prices change, but no new technologies are added to the menu. Analysts have recently started developing ways by which price-induced technological change and price-induced increases in the rate of diffusion of new technologies can be included.

The technological change that occurs over time, and that is included in most of the models, reduces the costs of mitigating carbon emissions because it decreases the base case trajectory of GHG emissions. However, it is probably unrealistic to assume that changes in energy prices will not alter the course of technological progress. In the short run, price increases should encourage the diffusion of new technologies. In the intermediate term, they should lead to a more rapid rate of increase in the rate of improvement of existing technologies and earlier remodeling or replacement of other facilities and equipment. In the long run, they should stimulate the development of brand new technologies. Both kinds of changes should reduce the average rates of GHG emissions per unit of output.

SEE ALSO THE FOLLOWING ARTICLES

Bottom-Up Energy Modeling • Business Cycles and Energy Prices • Carbon Taxes and Climate Change • Clean Air Markets • Climate Change and Energy, Overview • Climate Change: Impact on the Demand for Energy • Greenhouse Gas Abatement: Controversies in Cost Assessment • Market-Based Instruments, Overview • Modeling Energy Supply and Demand: A Comparison of Approaches

Further Reading

Jacoby, H. D., and Wing, I. S. (1999). Adjustment time, capital malleability and policy cost. *In* "The Costs of the Kyoto Protocol: A Multi-Model Evaluation" (J. P. Weyant, Ed.), pp. 73–92. International Association for Energy Economics, Cleveland, Ohio. [Special Issue of *The Energy Journal*.]

Jorgenson, D. W., and Wilcoxen, P. J. (1993). Energy, the environment and economic growth. *In* "Handbook of Natural Resources and Energy Economics" (A. Kneese and J. Sweeney, Eds.), pp. 1267–1349. North-Holland Publ., Amsterdam.

Manne, A. S., and Richels, R. G. (1992). "Buying Greenhouse Insurance: The Economic Costs of Carbon Dioxide Emission Limits." MIT Press, Cambridge, Massachusetts.

Nordhaus, W. D. (1994). "Managing the Global Commons: The Economics of Climate Change." MIT Press, Cambridge, Massachusetts.

Schumpeter, J. A. (1942). "Capitalism, Socialism, and Democracy." Harper & Brothers, New York and London.

Toth, F. L., and Mwandosya, M. (2001). Decision-making frameworks. *In* "Climate Change 2001: Mitigation" (B. Metz, O. Davidson, R. Swart, and J. Pan, eds.), pp. 601–688. Cambridge University Press, Cambridge, U.K.

Weyant, J. P. (ed.). (1996). Integrated assessment of climate change: An overview and comparison of approaches and results. *In* "Climate Change 1995—Volume 3: Economic and Social Dimensions of Climate Change" (J. P. Bruce, H.

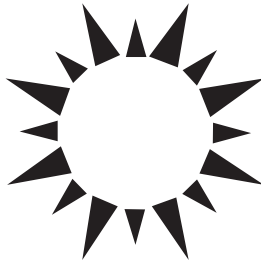
Lee, and E. F. Haites, eds), pp. 367–396. Cambridge University Press, Cambridge, U.K.

Weyant, J. P. (ed.). (1999). "The Costs of the Kyoto Protocol: A Multi-Model Evaluation." International Association for Energy Economics, Cleveland, Ohio. [Special Issue of *The Energy Journal*.]

Weyant, J. P. (2000). "An Introduction to the Economics of Climate Change Policy." Pew Center on Global Climate Change, Washington, D.C.

Weyant, J. P., and Olavson, T. (1999). Issues in modeling induced technological change in energy, environmental, and climate policy. *Environ. Model. Assess.* **4,** 67–85.

Modeling Energy Supply and Demand: A Comparison of Approaches



ALESSANDRO LANZA
Fondazione Eni Enrico Mattei, Milan, Italy
Eni S.p.A., Rome, Italy

FRANCESCO BOSELLO
Fondazione Eni Enrico Mattei
Milan, Italy

1. **An Introduction and a Short Historical Overview**
2. **Top-Down and Bottom-Up Models: Comparing Technical Features**
3. **Top-Down and Bottom-Up Models: Comparing the Methodologies**
4. **Top-Down and Bottom-Up Models: Possible Integration**
5. **Concluding Comments**

Glossary

constant elasticity of substitution (CES) technology Characterized by the fact that production factors are substitutable among each other at a given degree and that this substitution possibility is constant at any given production level.

Hicks–neutral technical change Refers to a technical improvement that uniformly reduces the input requirements associated with producing a given level of output.

Leontief technology Technology in which production is constrained by the input in the lowest availability; in this situation, increasing the availability of other production factors will not allow an increase in production.

panel data Composed of a repeated number of observations of the same group of individuals over time.

time-series data Composed of a repeated number of observations of a variable over time.

Walras's Law Often wrongly considered an equilibrium condition, a law stating simply that consumers fully expend their wealth.

After the first modeling efforts investigating the relationships between energy and economics emerged during the 1970s, two broad classes of modeling approaches appeared: the economic or top-down models and the technical/engineering or bottom-up models. The top-down models adopted a general perspective and described the economic linkages between energy demand and supply and the rest of the economic system, with the main goal of analyzing energy or wider economic policies. The bottom-up models adopted a focused view of the energy sectors and explored the various technological options, with the main goal of highlighting low-cost energy production opportunities.

1. AN INTRODUCTION AND A SHORT HISTORICAL OVERVIEW

The first modeling efforts investigating the relationships between energy and economics date back to the 1970s. Since the beginning, two broad classes of modeling approaches appeared: the economic or top-down models and the technical/engineering or bottom-up models. The first approach, adopting a general perspective, described the economic linkages between energy demand and supply and the rest of the economic system, with the main goal of analyzing energy or wider economic policies. The second approach, adopting a focused view of the energy sectors, explored the various technological options, with the main goal of highlighting low-cost energy production opportunities. During the same period, natural scientists became interested in energy modeling as well. To assess the role of anthropogenic

 © 2004 Elsevier Inc. All rights reserved.

greenhouse gas (GHG) emissions in determining global warming, they developed pioneering models consisting at that stage in simple extrapolations based on a restricted number of parameters where qualitative/quantitative experts' opinions were predominant with respect to direct estimation or calibration procedures. Simulation of energy and climate policies was well beyond the capacity of these first mathematical tools.

Interestingly, it was mainly environmental concerns that stimulated both bottom-up and top-down modeling efforts during the subsequent decade. In this respect, a particularly important event was the 1988 Toronto Conference on Climate that spurred a real boom in modelization. One of the outcomes of the conference was to establish a first preliminary limit for carbon dioxide (CO_2) emissions in the measure of a cut of 20% with respect to the 1988 level. This fostered immediate and great interest in the scientific and political communities to assess costs and benefits of such a target, its implications for the economic systems, and the feasible strategies for its accomplishment. Moreover, the need to analyse three different dimensions—energy, economics, and environment—encouraged a fruitful and unprecedented exchange of knowledge among disciplines.

Nowadays, bottom-up and top-down models are being used extensively to investigate climate change issues and the costs of the related policies. Moreover, it is a fact that even when these models are used to investigate energy policies and reforms not directly related to climate, they nevertheless take into account some environmental implications (e.g., GHG emissions are commonly considered a strategic variable to monitor anyway).

It could be said that the points in common between the two approaches end here. In fact, they have led to very different model properties and results. A striking example is again given by the case of determining the costs of climate change policies where top-down and bottom-up studies wound up producing opposite results. This induced two opposing views in the scientific community: one claiming the necessity to integrate the two methodologies so as to exploit the comparative advantages of both techniques and the other considering the two methodologies totally incompatible.

Recently, also due to the development of new flexible software packages and to the improvement in computers' computational capacity, the first view seemingly has prevailed, and a large number of "hybrid" models sharing top-down and bottom-up features is being developed. Still, the integration techniques are far from reaching an uncontroversial standard and securing full validation from the scientific community; bridging the gap between the two approaches developed for different purposes and permitting different analyses designed to answer different questions remain a difficult and complex task. Yet despite efforts to merge characteristics in the hybrid models, and despite convergence between model categories, the distinction remains an important one that is essential to understanding the policy conclusions of influential studies and to avoiding misinterpretation of model results.

In light of this, the aim of this article is to describe the main technical and theoretical features of bottom-up and top-down modeling approaches, presenting their results, highlighting their respective strengths and weaknesses, and discussing the main issues that have arisen due to a possible integration.

In what follows, this article presents the technical characteristics of the two modeling approaches. It then compares and comments on the theoretical rationales shaping the models' use and results. Subsequently, it describes possible ways in which the two approaches could be and have been reconciled. The final section concludes the article.

2. TOP-DOWN AND BOTTOM-UP MODELS: COMPARING TECHNICAL FEATURES

2.1 Top-Down Modeling: Technical Features

Top-down models are considered economic, general equilibrium, or aggregated models; they aim to giving a comprehensive picture of the functioning of an economic system, including the relationship between energy markets and the rest of the economy, based on the behavior of representative and rational economic agents maximizing an objective function. The information used to build top-down models "comes from the past"; usually, historical data describing energy–economic interactions are used to shape present and future market behavior. With respect to this, two broad parameterization techniques can be identified: calibration (originating models in which the key parameters forging agents' behavior are based on the information stemming from a given point in time, i.e., the base year) and estimation (originating econometrically estimated models in which the parameterization is obtained extracting information from longer time periods, i.e.,

through time-series or panel econometric techniques). In fact, it is common for both approaches to coexist within a model, especially due to constraints imposed by data availability.

Traditionally the economic modeling literature has divided between general equilibrium models (GEMs), whether static or dynamic (on their turn, classifiable in recursive dynamic and growth models), and macroeconometric models. The view of GEMs is "Walrasian"; that is, optimizing producers and households demand and supply goods and factors, perfectly flexible prices adjust excess demand and supply in all markets that clear, and profit maximization under perfect competition and free market entrance guarantee zero profits and the optimal distribution of resources. Macroeconometric models are "neo-Keynesian" in the sense that the economic system is demand driven; moreover, perfect competition is abandoned, especially in energy and labor markets where market power and bargaining processes determine price settings.

Notwithstanding the great heterogeneity within the top-down model family, the explicit aim of comprehensiveness, the focus on agents' decisions, and the idea that the past can describe the future consistently allow one to identify some important common features in the various modeling efforts.

2.1.1 Key Parameters and Variables

Usually, the most important parameters and variables constituting the main drivers of internal processes and results are the elasticity of energy demand in response to gross domestic product (GDP) or income changes; the elasticity of price substitution among capital, labor, and energy; the indicators of technical progress, usually assuming the form of non-price-induced decreases in energy use per unit of GDP (the so-called autonomous energy efficiency improvement [AEEI]) and/or of a more general autonomous factor-augmenting productivity (generally "Hicks-neutral" technical progress); the cost and availability of existing and future energy supplies; and the availability, conversion efficiencies, and costs of existing and future energy generation technologies, including "backstops" (discussed later). Together with the GDP growth rate, all of these elements determine future energy demand and the economic cost of a technological switch.

2.1.2 Disaggregation

It is a recognized fact that estimated econometric relationships among aggregated variables are generally more reliable than those among disaggregated variables; as a consequence, the behavior of the models is more stable using such variables. Thus, it is common to adopt a high level of aggregation for top-down models when they are applied to long time frames. An additional constraint on the level of disaggregation attainable is imposed by computational burdens. This is true in particular for dynamic top-down models in which the process of intertemporal optimization used to find the equilibrium path for the economic variables taking into account strategic behavior among various agents (countries or sectors) requires the solution of huge nonlinear systems. In practice, although a static top-down model can reach a good degree of sectoral disaggregation (e.g., the 57 sectors of the GTAP model), a dynamic top-down model usually presents fewer than 10 to 15 sectors (e.g., the G-cubed model) or even no sectoral disaggregation (e.g., the RICE '96 and RICE '99 models).

2.1.3 Technology

Being concerned with the feedback between energy and other sectors, and between macroeconomic impacts of given policies (e.g., energy or climate change policies), on the national or even global scale, top-down models tend to propose minimal details of the energy-consuming side of the economy. More specifically, technology is usually represented by the shares of the purchase of a given input in intermediate consumption in the production function, as well as in labor, capital, and other inputs, and by the previously mentioned AEEI and/or the factor accounting for Hicks-neutral technical change. Depending on the functional form of the production function, these parameters describe the degree of substitutability among inputs and the evolution of their productivity. In other words, they shape economically a technology that finally determines energy (and other input) demands for the various energy types. This reflects the idea that it is mainly the characteristics of the specific energy type, such as its price, that govern demand developments. It is important to highlight that shares are usually constant and that parameters of technological progress are usually exogenous, being represented by some kind of time trend. This implies, on the one hand, that the approach to technology is inherently static, somehow "freezing" substitutability relationships among various inputs to what has already been observed (in a given year or a given "past") and, on the other, that non-price-induced technological progress is not agents' decision. In other words, agents can decide where to place themselves along the

production frontier, but they cannot determine the shape and shift of that frontier. More recently, the issue of endogenizing technological change was tackled by several researchers. Again the perspective is intrinsically economic. Put simply, the point here is to endow agents with an additional decision variable, usually the amount of investment in research and development (R&D), competing with the others, the stock of which determines, according to some functional form, the productivity and energy efficiency of the input bundle. In this case, AEEI and technical progress do depend on agents' decisions.

2.1.4 The Supply Side

Another key aspect that top-down modeling studies have to consider is the availability of primary inputs, in particular of energy sources and their cost. The majority of top-down models distinguish between carbon-emitting and non-carbon-emitting energy supplies. Usually, fossil fuels (coal, oil, and natural gas) provide energy and produce GHGs as by-products, whereas electricity use is considered carbon free. Some models explicitly consider the possibility of accessing alternative forms of renewable energy supplies with low or zero carbon content. This is done mainly following the so-called "backstop approach," a rather common technique. Alternative energy sources are already technologically feasible and available, but they are not economically viable due to their high costs. As time passes, extracting energy from traditional energy sources becomes more costly due to reserve exhaustion. Accordingly, at a given point in time, new energy sources and technologies become competitive and enter into the market, determining a total or partial shift away from carbon-emitting sources. Thus, the assumed cost and availability of backstop technologies are key determinants of the long-term marginal cost of supply in top-down models.

2.2 Bottom-Up Modeling: Technical Features

Bottom-up models are considered engineering, partial equilibrium, or disaggregated models. Their original aim was to find least cost opportunities to meet an exogenous demand for specific energy services. More recently, following environmental concerns, they have been used to devise least cost opportunities to achieve a given energy efficiency or carbon emission reduction. Bottom-up models are based on technological and economic data that describe in great detail past, present, and future technological options that are used, or that can be used, for harnessing energy resources and convert them into energy services.

Bottom-up models can be grouped into two broad categories: spreadsheet models (which solve a simultaneous set of equations to describe the way in which a given set of technologies is or could be adopted) and simulation or optimization models (which simulate investment decisions endogenously). Again within the heterogeneity of bottom-up modeling, some common features can be devised.

2.2.1 Key Parameters and Variables

In a bottom-up model, the fundamental parameterization concerns the costs and initial values of installed capacities of technologies currently in use and of their alternatives, their residual lives, fuel and electricity costs, and the potential rates and limits of alternative technology penetration.

2.2.2 Disaggregation

As mentioned previously, a typical feature of bottom-up models is the high degree of detail or disaggregation in the representation of the energy sector. But note that disaggregation in bottom-up models has a different meaning with respect to top-down terminology. Most bottom-up models interconnect conversion and consumption of energy via energy carriers. Usually, energy carriers are disaggregated according to their involvement with primary supplies (e.g., mining, petroleum extraction), conversion and processing (e.g., power plants, refineries), and end use demand for energy services (e.g., boilers, automobiles, residential space conditioning). Demand for energy may, on its turn, be disaggregated by sector (e.g., residential, manufacturing, transportation) and by specific functions within a sector (e.g., air conditioning, heating, lighting).

2.2.3 Technology, Supply, and Demand

Bottom-up models capture technology in the engineering sense. This means, for instance, that technology A, with a given performance X determined by a whole set of parameters accessible at the direct cost C(A), on its turn determined by a whole set of parameters, is compared with technology B with a given performance Y accessible at the direct cost C(B). As an outcome, individual measures are ranked in order of increasing net cost to form "packets of options" represented as marginal cost or curves. These are called "conservation supply curves" (CSCs). More specifically, CSCs can be obtained through a three-step procedure: (1) energy-saving

potential and costs of all possible "packets" of options are evaluated, (2) the total "cost of conservation" curve is obtained as the lower convex bound of all conservation packets that may be justifiable at some energy price level, and (3) the CSC is obtained by plotting the slopes of each straight line segment of the total cost of conservation curve. Demand in these models is usually exogenous but can also be determined endogenously. The "typical" focus is on end use energy demand (for heating, lighting, ventilation etc.). This endorses the view that the evolution of energy demand is driven mainly by the various purposes for which energy is made use of. Household energy demands are typically specified by vintage models of a large number of end use technologies whose penetration rates follow a time profile with saturation levels. Penetration ratios could be specified to follow a logistic function, and in some cases parameters of these logistic functions are estimated for each type of appliance.

All of the descriptions in this section are summarized in Table I and Fig. 1.

3. TOP-DOWN AND BOTTOM-UP MODELS: COMPARING THE METHODOLOGIES

It should not be surprising that the different technical characteristics of the two modeling approaches, peculiar to two different disciplines adopting different perspectives, determine a difference both in the kind of questions the specific models can address and in the qualitative and quantitative results they can produce.

TABLE I

Main Features of Top-Down and Bottom-Up Modeling Approaches

Criterion	Top-down	Bottom-up
Level of disaggregation	• Usually low: 1–10 sectors or activities; can be high in some CGE static models	• High: a wide range of energy end uses represented
Behavior representation	• Comprehensive (general equilibrium approach = full economic feedback effects) but few energy-relevant details	• Detailed (at end use level) but not comprehensive (partial equilibrium approach, "no rebounds")
Representation of technologies	• Based on macro, input/output, or econometric analysis • Production functions determine substitution possibilities	• Based on engineering and cost data • Description of physical flows
Technological change	• Price and income effects • Usually exogenous technical progress (Hicks-neutral + AEEI); some models consider R&D and learning-by-doing-driven endogenous technical progress	• Assumptions on market shares or optimization • Projections of technological efficiency • Learning curves
Methodological approach	• Time-series/panel econometrics or calibration based on a single year • Economic growth estimated or exogenous	• Spreadsheet analysis (for descriptive reasons) • Simulation/optimization models
Efficiency gap	• No energy efficiency gap except in case of energy subsidies • Usually markets are fully competitive in CGE models; market power (oligopoly/monopolistic competition) in energy markets in Keynesian macroeconometric models	• Energy markets are not efficient (functioning inside the production frontier) • Potential for cost-effective energy savings
Assumptions about market barriers and hidden costs of new technologies	• Costs of adopting new technologies are reflected in observed behavior	• Significant market barriers prevent adoption of new technologies • Hidden costs tend to be low
Transaction costs of removing market barriers and imperfections	• High	• Low

Source. Adapted from International Energy Agency (1998).

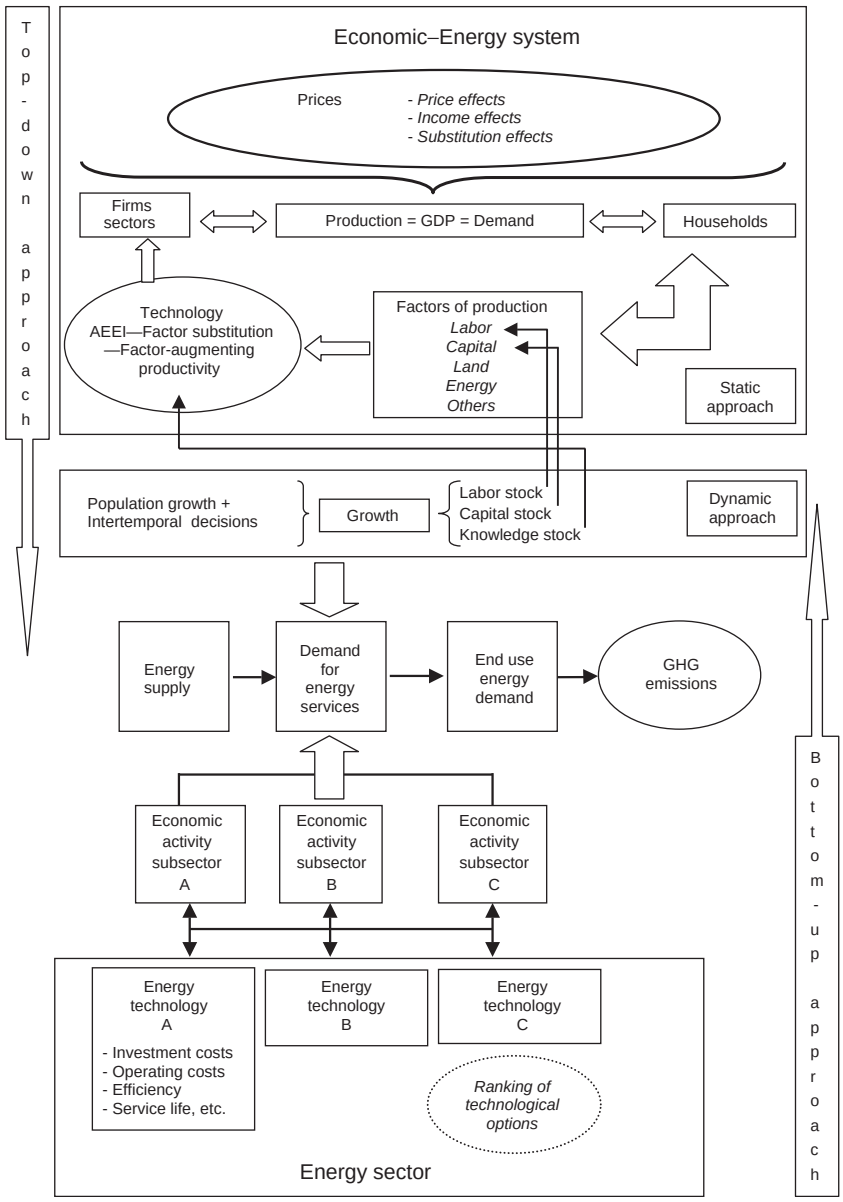


FIGURE 1 Schematic diagram illustrating the structural features of bottom-up and top-down modeling exercises. Top-down approaches are highlighted by the downward arrow, and bottom-up approaches are highlighted by the upward arrow.

A problem can arise when, using both tools to investigate the same issue with the explicit aim of providing policy advice, the two approaches end up with opposite answers. This is the specific case of climate change and climate change policies where both top-down and bottom-up models have been widely used to assess costs and benefits of various control options to decide how much, and according to which strategies (if any), to intervene. Unfortunately, top-down models generally conclude that mitigating global warming entails substantive costs for the economic systems, whereas bottom-up models demonstrate that potential costs are negligible.

Understanding the reason for this discrepancy offers the opportunity to present and compare the different methodologies. Basically, a top-down model can perform two kinds of investigations that are strongly market oriented: the so-called "if-then" or simulation analysis and the optimization analysis. In the first case, the question answered is of the following type: "what are the economic consequences (in terms of international/intersectoral

trade, GDP, welfare, etc.) of a specific perturbation (e.g., a tax levied on the production or consumption of a given good, a change in factor productivity and endowment) imposed on a system?" or "what is the tax or subsidy required to accomplish a given policy target?" In the second case, the question answered is of the following type: "what is the path of a given control variable (e.g., the tax, the investment, the abatement rate) allowing a given target minimizing cost or maximizing welfare to be reached?" The perspective of top-down models is "wide" or "general"; that is, in describing the effect of the perturbation imposed on the system, they try to describe and measure all of the (macro)economic implications and feedback originated within the system. The technological part is necessarily simple, embodied in a small number of parameters.

Conversely, bottom-up models focus on problem solving and are solution oriented. They identify least energy and/or least cost strategies to providing energy services or to accomplish given GHG emissions reduction targets. Basically, the user of a bottom-up model can specify a given set of constraints (e.g., on technology or on demand but also derived from policy variables as emissions limits or taxes), and the model will produce the least cost "technological" solution that meets the provided set of constraints. The outcome of the procedure allows various future technology configurations, the paths of their adoption, and the optimal allocations of their investment flows to be described and compared. Bottom-up models are focused on the energy sector that they describe in great detail but leave the general macroeconomic picture outside the model.

This illustrates two important differentiations. First, top-down models can consider new technological options only marginally. This implies that top-down models, unlike bottom-up models, observe that an important part of the emissions reduction potential of the energy sector is missing. Second, the global perspective of top-down models takes into account the so-called rebound effects that usually escape from bottom-up analyses. Said simply, under an engineering perspective, the possibility of accessing a new "clean" technology at low cost directly implies a cheap decrease in the polluting potential of an economic sector. In a top-down framework, lower prices imply savings and/or an increase in demand for that technology. Both of these elements—an increase in demand and additional savings that can be expended somewhere—usually lead to an increase in energy consumption or consumption tout court ("rebound") that, on its turn, may lead to an increase or a lower decrease in emissions.

Another basic difference is more "philosophical": top-down models adopt the perspective of markets led by rational agents. The direct consequence is that in a top-down world, new and low-cost technologies are hardly accessible because rational agents should have already exploited all of the best available options. Another consequence is that in top-down models, all physical and technological resources for supplying energy services are constrained to quantities available at the current time. A practical example is given by the previously mentioned treatment of backstop technologies. The possibility of obtaining energy from new sources is indeed accounted for, but notwithstanding technological progress, it is accounted for only at an increasing cost over time. Bottom-up models consider the future as "changeable"; they observe that technological progress can improve supply- and demand-side technologies, whose price can then decline. This is why bottom-up models usually highlight the potential for low-cost or even negative cost energy or emissions-saving opportunities.

In light of this, which approach is more reliable? Can the two approaches be reconciled or integrated? Regarding the first question, it must be recognized that both approaches present strengths and weaknesses. The strength of top-down models is their comprehensiveness; basically, these models point out that the energy sector cannot be considered in isolation from the whole economic system and, more important, that economic elements shaping and shaped by agents' behavior are very important in determining the technological quality and performance of a production system. The main limitation is top-down models' inherently "fixed" nature; they can offer valuable insights about the future so long as the relations obtained from the past remain unaltered and can be projected into the future. The main consequence is that a true analysis of impacts of technological innovation over time is precluded because these models implicitly neglect the fact that information, policy, or institutional changes can increase the availability of cost-effective technologies. Conversely, bottom-up models do highlight the possibility of low-cost options but suffer from their technology- and sector-specific focus. This neglects two important facts. On the one hand, even assuming that a technological option were available, it could be neither developed (due to, e.g., the huge investment necessary) nor diffused through the whole production system (due to, e.g., the existence of

transaction costs and other barriers imposed by investment planning). On the other hand, bottom-up models neglect rebounds or feedback that can counterbalance (sometimes dramatically) initial efficiency improvements.

These considerations could also shed some light on the issue of reconciling results. The basic point is to recognize that the two approaches are different. Additional confusion derives from the fact that sometimes the same terminology is used in the two fields but with different meanings. Top-down models are useful for describing macroeconomic (general equilibrium) effects and transmission channels related to taxes, various economic scenarios on energy and environment, and changes in the energy system. The subsequent quantification could be reliable so long as technical progress and cost-effective efficiency opportunities are limited and, more generally, the relationship among energy, economics, and the environment does not change over time. Conversely, bottom-up models are valuable tools for analyzing regulation and energy planning, restructuring of the energy supply sector, and introduction of technological standards and the technological development so as to quantify the aggregated development in energy efficiency, but they cannot be used generally, for example, to provide estimates of the costs of reducing GHGs on a global scale.

4. TOP-DOWN AND BOTTOM-UP MODELS: POSSIBLE INTEGRATION

As pointed out by some authors, it seems not only desirable but also natural to integrate the two approaches. Engineering data on technical potential and costs can offer valuable input to economic market investigations analyzing the effects of a given energy policy, whereas information about the economic intersectoral and international relationships described by top-down models can enrich the cost assessments of bottom-up research with the crucial dimensions of market effects.

Indeed, numerous attempts have been made to coherently and consistently exploit the benefits of both methodologies. One straightforward procedure is to use the results of bottom-up models as an input in macroeconomic models. This is the so-called soft link. An example derives from the MARKAL–MACRO model. MARKAL is a national system optimization model that determines the "best" energy technology solutions and provides marginal abatement costs, energy prices, and rates of technical change as input for MACRO, a simple production and utility function for the economy.

A "harder" link among bottom-up and top-down rationales is obtained when bottom-up and top-down modules can work interdependently. An example derives from the Hybris model, developed by Jacobsen, in which the macroeconomic top-down model ADAM for Denmark is linked to three different bottom-up energy modules: energy supply, demand for heat, and demand for electricity. Basically, the bottom-up modules aggregate or disaggregate variables to fit the specification of the macro model. The model can run different scenarios for bottom-up and top-down initiatives separately and then compare them to scenarios with "combination of reduction" initiatives to highlight the interdependencies among them.

Another notable example is given by the AIM Asian–Pacific model documented by Fujino and colleagues. A set of bottom-up modules that can reproduce detailed processes of energy consumption, industrial production, land use changes, and waste management, as well as technology development and social demand changes, is linked to top-down modules to compute GHG emissions and their general equilibrium relationship with international trade and via an aggregation computable general equilibrium (CGE) interface to assess country and global sustainable development paths considering the feedback among energy, economics, and the environment.

As claimed by several authors, a soft link methodology suffers from the major limitation of failing an effective reconciliation between the logic of the two approaches. In particular, the separation between technological choices at the bottom-up level and economic choices at the top-down level is not observed in the reality where the two decisions are indeed taken simultaneously. Moreover, the top-down part of the model usually is built following the "neoclassical" investment theory that does not allow distinguishing between investment decisions related to the "quality" of the capital stock (e.g., its energy-saving potential) and those related to the "quantity" of the capital stock (e.g., its productive capacity). The consequence is that energy-saving investment is necessarily ill modeled.

This has led to the development of so-called hard link models that attempt a full fusion between top-down and bottom-up features. One possibility is to add a qualitative dimension to the capital stock. An interesting example derives from the WARM model for the European economy, developed by Carraro and

Galeotti, in which economic agents decide not only the quantity but also the quality of the total capital stock. This is done assuming the existence of two kinds of capital: a polluting one and an environmentally friendly one (with the latter allowing for energy savings). Each year, a new capital vintage enters into use, and its quality is determined by the relative weight of the "dirty" and "clean" components endogenously chosen by agents. This composition determines a quality index of the total capital stock that is crucial in determining energy demand responses to environmental and/or energy policies.

Alternatively, Bohringer proposed a hybrid approach that can be applied to a static CGE framework in which energy sectors are represented by bottom-up activity analysis and the other production sectors are characterized by top-down production functions. What is interesting is that the model is solved as a unique equilibrium. Although the technicality of the argument is beyond the scope of this article, this is intuitively possible by observing that equilibrium conditions stemming from the bottom-up representation of a sector are compatible with the equilibrium conditions implied by the standard Arrow–Debreu framework constituting the basis for CGE static top-down models when they are stated in their most general form. The treatment of this kind of problem requires the use of the so-called complementary format proposed by Cottle and Pang. The model is formulated as a nonlinear system of inequalities satisfying the usual Arrow–Debreu general equilibrium conditions (non-negativity of supply minus demand for every commodity, zero profit condition, and Walras's Law). In top-down sectors, substitution is described by nested separable constant elasticity of substitution (CES) functions, whereas in bottom-up sectors, it is represented by a bundle of discrete Leontief technologies.

A possible integration procedure in a dynamic environment was instead proposed by Muller. The basic idea is to consider capital goods as heterogeneous, being characterized by both different production capacity and different energy efficiency. The representative firm chooses quality and quantity of new capital equipment simultaneously to minimize present and discounted future costs. Basically, this intertemporal optimization procedure is constrained by two laws of motion: one for capital stock and one for its quality. This last factor cannot be adjusted to the desired level immediately; rather, it can be improved only by replacing old inefficient vintages with new, more energy-efficient equipment. In this way, the model can take into account the double role of investment. One is the top-down motivation of giving up one unit of consumption today in exchange for increasing productive capacity and consumption tomorrow, and the other is the bottom-up motivation of sustaining higher costs today by buying an efficient capital stock in exchange for gaining energy savings due to this higher efficiency tomorrow.

5. CONCLUDING COMMENTS

A model can be defined as a simplified but sufficiently representative picture of reality. Thus, the very first task of a model is to describe reality—or some aspects of it—that should appear clearer and more understandable with the imposed simplification of the model than without it. In the case of energy modeling, it is unquestionable that both top-down and bottom-up approaches contributed greatly to improving the knowledge of the respective economic and technological dynamics related to the energy system. Nevertheless the different foci of the two methodologies—one economic the other technological—usually leads them to different forecasts and prescriptions. Moreover, it must be recognized that, notwithstanding numerous promising attempts to bridge the gap between the two views, this difference still remains. Under a scientific perspective, this is not dramatic. The process of gaining knowledge takes time, and accordingly, a difference in results can be expected, and the consequent debate can constitute a major driver for scientific improvements. For scientists, it is a strong push to develop finer bottom-up and top-down methodologies and more appropriate integration techniques. On the contrary, under a policy perspective, the difference is problematic. On the one hand, it may lower the trust of policymakers regarding both approaches; on the other hand, it may leave policymakers without a sure framework for their decision making. This issue is very concrete insofar as although a full recognized and accomplished scientific reconciliation is missing, the choice between top-down and bottom-up modeling has to be made on political grounds. What science can do, while working on improvements, is present the results of the two approaches and state clearly under which conditions, and according to which hypotheses, these results could be expected to hold. Symmetrically, policymakers should make clear which vision of the world has endorsed their strategies, in any case bearing in mind that tackling energy policies adopting a bottom-up or top-down angle in isolation is misleading.

SEE ALSO THE FOLLOWING ARTICLES

Bottom-Up Energy Modeling • Decomposition Analysis Applied to Energy • Depletion and Valuation of Energy Resources • Economics of Energy Demand • Economics of Energy Supply • Fuzzy Logic Modeling of Energy Systems • Input–Output Analysis • Modeling Energy Markets and Climate Change Policy • Multicriteria Analysis of Energy • National Energy Modeling Systems

Further Reading

Bohringer, C. (1998). The synthesis of bottom-up and top-down in energy policy modeling. *Energy Econ.* **20,** 233–248.

Carraro, C., and Galeotti, M. (1996). WARM: A European model for energy and environmental analysis. *Environ. Modeling Assess.* **1,** 171–189.

Fujino, J., Garg, A., *et al.* (2002). *In* "Climate Policy Assessment: Asia–Pacific Integrated Modeling" (M. Kainuma, Y. Matsuoka, and T. Morita, Eds.). Springer-Verlag, Tokyo.

Hertel, T. (1998). "Global Trade Analysis: Modeling and Applications." Cambridge University Press, Cambridge, MA.

International Energy Agency. (1998). "The Energy Future: Energy Modeling and Climate Change Policy, Energy, and the Environment." IEA, Paris.

Jacobsen, H. K. (1998). Integrating the bottom-up and top-down approach to energy–economy modeling: The case of Denmark. *Energy Econ.* **20,** 443–461.

Manne, A., and Wene, C. O. (1994). MARKAL–MACRO: A linked model for energy–economy analysis. *In* "Advances in System Analysis: Modeling Energy-Related Emissions on a National and Global Level" (J. Hake, M. Kleemann *et al.,* Eds.). Konferenzen des Forschungszentrums, Jülich, Germany.

McKibbin, W. J., and Wilcoxen, P. J. (1995). "The Theoretical and Empirical Structure of G-Cubed." Brookings Institution, Washignton, DC.

Muller, T. (2000). "Integrating Bottom-Up and Top-Down Models for Energy Policy Analysis: A Dynamic Framework." Centre Universitaire d'études, Université de Geneve, Switzerland.

Nordhaus, W., and Yang, Z. (1996). A regional dynamic general equilibrium model of alternative climate change strategies. *Am. Econ. Rev.* **86,** 726–741.

Popp, D. (2000). Induced innovation and energy prices. *Am. Econ. Rev.* **92,** 160–180.

Wilson, D., and Swisher, J. (1993). Top-down versus bottom-up analyses of the cost of mitigating global warming. *Energy Policy* **21,** 249–263.

Motor Vehicle Use, Social Costs of

MARK A. DELUCCHI
University of California, Davis
Davis, California, United States

1. **Background**
2. **Relevance of Analyses of the Social Cost of Motor Vehicle Use**
3. **A Conceptual Framework**
4. **Components of the Social Cost of Motor Vehicle Use**
5. **Results of an Analysis**
6. **Summary**

Glossary

annualized (amortized) cost An initial or up-front payment for an asset converted to an economically equivalent stream of regular payments over the life of an asset; conversion depends on the interest rate.

average cost The total cost of a given total quantity divided by the total quantity (e.g., the total cost of 4 billion barrels of oil divided by 4 billion barrels) (cf. marginal cost).

bundled cost A cost, such as the cost of parking, that is not explicitly and separately priced, but rather is included in the cost of other items that together are explicitly priced as a "bundle."

cost–benefit analysis A method of analysis in which the economic impacts of a proposed plan, policy, or project are identified, quantified, and valued in dollars for the purpose of comparing the total value of the negative impacts (costs) with the positive (benefits).

external cost The incidental cost of an economic transaction to persons who are not formally part of the transaction (see Section 4.5 for a more formal definition).

marginal cost The cost of an additional or incremental unit of consumption (e.g., the cost of the four-billionth barrel of oil) (cf. average cost). The marginal cost will exceed the average cost if the cost of producing each additional unit increases continually because of increasing scarcity of resources.

monetary cost A cost, such as the cost of gasoline, that is expressed in monetary terms (e.g., dollars) in market transactions. A nonmonetary cost is a cost, such as the cost of travel time, that is not expressed directly in money terms.

opportunity cost In economics, the opportunities or alternative resource uses given up as a result of a particular use of a resource.

optimal price A price on a good or service that incorporates all of the marginal costs to society of using the good or service.

private cost The opportunity cost of an action to a private individual, usually as part of a market transaction.

social cost The opportunity cost of an action to society as a whole; generally equal to the private cost plus the external cost.

The social cost of motor vehicle use—the all-inclusive economic cost to society of using motor vehicles—comprises explicitly priced private-sector costs, bundled private-sector costs, public-sector costs, external costs, and personal nonmonetary costs. Estimates of the social cost can be used for a variety of purposes, including analyses of efficient pricing of motor vehicle goods and services, cost–benefit analysis of motor vehicle projects and plans, and general policy analysis related to motor vehicle use. This article focuses on the classification, uses, and magnitude of estimates of the social cost of motor vehicle use.

1. BACKGROUND

Every year, American drivers spend hundreds of billions of dollars on highway transportation. They pay for vehicles, maintenance, repair, fuel, lubricants, tires, parts, insurance, parking, tolls, registration, fees, and other items. These expenditures buy Americans considerable personal mobility and economic productivity. But the use of motor vehicles costs society more than the hundreds of billions of dollars spent on explicitly priced transportation goods and services. There also are bundled costs, i.e., those goods and services that are not explicitly priced, but rather are bundled in the prices of

 © 2004 Elsevier Inc. All rights reserved.

nontransportation goods and services. For example, "free" parking at a shopping mall is unpriced, but it is not costless; its cost is included—bundled—in the price of goods and services sold at the mall.

In addition to these priced or bundled private-sector costs, there are public-sector costs, of tens of billions of dollars per year, to build and maintain roads and to provide a wide range of services that partly support motor vehicle use. These costs include those for police protection, the judicial and legal systems, corrections, fire protection, environmental regulation, energy research and regulation, military protection of oil supplies, and more. Finally, beyond these monetary public and private-sector costs, are the nonmonetary costs of motor vehicle use, which, by definition, are not valued in dollars in normal market transactions. A wide variety of nonmonetary costs exist, including the health effects of air pollution, pain and suffering due to accidents, and travel time. Some of these nonmonetary costs, such as air pollution, are externalities, whereas others, such as travel time in uncongested conditions, are what may be called "personal nonmonetary costs."

The social cost of motor vehicle use—the all-inclusive economic cost to society of using motor vehicles—is the sum of all of the costs just mentioned: explicitly priced private-sector costs, bundled private-sector costs, public-sector costs, external costs, and personal nonmonetary costs. These costs are listed in complete detail and classified more rigorously, in Table I. Over the past decade, a number of researchers have been doing detailed and comprehensive analyses of the social cost of motor vehicle use. In the following sections, the purpose of estimating the total social cost of motor vehicle use is explained, a conceptual framework and a cost classification are delineated, and recent cost estimates are presented and discussed.

2. *RELEVANCE OF ANALYSES OF THE SOCIAL COST OF MOTOR VEHICLE USE*

Researchers have performed social-cost analyses for a variety reasons and have used them in a variety of ways to support a wide range of policy positions. Some researchers have used social-cost analyses to argue that motor vehicles and gasoline are terrifically underpriced, whereas others have used them to downplay the need for drastic policy intervention in the transportation sector. In any case, social-cost analyses usually excite considerable interest, if only because nearly everyone uses motor vehicles. By itself, however, a total social-cost analysis does not determine whether motor vehicle use on balance is good or bad, or better or worse than some alternative, or whether it is wise to tax gasoline or restrict automobile use or encourage travel in trains. Rather, a social-cost analysis is but one of many pieces of information that might be useful to transportation policymakers. Specifically, a social-cost analysis can help in the following ways:

1. To establish efficient prices for transportation resources, such as roads or automobile emissions. A social-cost analysis can give some idea of the magnitude of the gap between current prices (which might be zero, as in the case of emissions) and theoretically optimal prices, and can inform discussions of the types of policies that might narrow the gaps and make people use transportation resources more efficiently. However, unless it is done with extraordinary specificity and explicitly with an eye to pricing, a social-cost analysis cannot tell us optimal prices for roads or emissions or anything else.
2. To evaluate all of the costs and benefits of alternative transportation investments in order to find the alternative that will provide the highest net benefit to society. A social-cost analysis, as is presented here, is conceptually one-half of the full social-cost-benefit analysis that must be performed in order to invest society's resources efficiently. At a minimum, a broad but detailed social-cost analysis can be a source of data and methods for evaluations of specific projects.
3. To prioritize efforts to reduce the costs of transportation. A detailed comparison of costs can help decide how to fund research and development to improve the performance and reduce the costs of transportation. For example, when considering funding research into the sources, effects, and mitigation of pollution, it might be useful to know that emissions of road dust might be an order of magnitude more costly than are emissions of ozone precursors, which in turn might be an order of magnitude more costly than are emissions of toxic air pollutants.

The analysis and estimates are presented here with these relatively modest informational purposes in mind, not to promote any policies regarding motor vehicle use. There is no declamation, for example, about the correct price of gasoline or the correct level of investment in the highway infrastructure.

3. A CONCEPTUAL FRAMEWORK

In speaking of the social cost of motor vehicle use, what is meant is the annualized social cost of motor vehicle use. The annualized cost of motor vehicle use is equal to the sum of periodic or operating costs (such as fuel, vehicle maintenance, highway maintenance, salaries of police officers, travel time, noise, injuries from accidents, and disease from air pollution) plus the value of all capital (such as highways, parking lots, and residential garages; items that provide a stream of services), converted (annualized) into an equivalent stream of annual costs over the life of the capital. This annualization approach essentially is an investment analysis, or project evaluation, in which the project, in this case, is the entire motor vehicle and use system. Of course, it is awkward to treat the entire motor vehicle system—every car, every gallon of gasoline, every mile of highway—as a "project" up for evaluation, but endeavoring to generate data and methods that will be generally useful in analyses of pricing, investment, and research will not avoid the scale and its awkwardness.

3.1 What Counts as a Cost of Motor Vehicle Use or Infrastructure?

In economic analysis, "cost" means "opportunity cost." The opportunity cost of action A is the opportunity that is foregone—what is given up, or used, or consumed—as a result of doing A. For some resource R to count as a cost of motor vehicle use, it must be true that a change in motor vehicle use will result in a change in use of R. Thus, gasoline is a cost of motor vehicle use because a change in motor vehicle use will result in a change in gasoline use, all else equal. But general spending on national health and education is not a cost of motor vehicle use because a change in motor vehicle use will not result in a change in the resources devoted to national health and education.

For the purposes of planning, evaluating, or pricing, it is of concern not only whether something is a cost of motor vehicle use, but, if it is a cost, exactly how it is related to motor vehicle use. For example, pollution is a direct, immediate cost of motor vehicle use: if motor vehicle use is changed a little, pollution is immediately changed a little. But defense expenditures in the Persian Gulf, if they are a cost of motor vehicle use at all, are an indirect, long-term, and tenuous cost (via congressional decisions about military needs, which might be informed by concerns about the security of oil supply for transportation). This sort of distinction is important because the more tenuously linked costs are harder to estimate, often delayed considerably with respect to the causal changes in motor vehicle use, and often highly dependent on the specific characteristics and amount of the change in motor vehicle use.

3.2 Costs versus Benefits

In this discussion, the dollar social cost but not the dollar social benefit of motor vehicle use is considered. Of course, it is not forgotten that there are benefits of motor vehicle use (a charge occasionally leveled against social-cost analysts) and there is certainly no presumption here that the benefits somehow are less important than the costs of motor vehicle use. Rather, because there is no credible way to estimate all of the benefits of motor vehicle use, there is no attempt to do so here. The emphasis here, however, is that not only does motor vehicle use provide enormous social benefit, but that this benefit, if it could be expressed in dollars, almost certainly would greatly exceed the full social cost.

Nevertheless, because this is a cost analysis only, it is not possible to say much about net dollar benefits or cost–benefit ratios, or whether a particular transportation system is worthwhile, or better or worse than another system. For example, this analysis indicates that motor vehicle use might cost more than is realized (i.e., that the total social cost appreciably exceeds the commonly recognized private costs, such as for fuel, vehicles, and maintenance and repair). But even if this is so, it does not mean that motor vehicle use costs more than it is worth, or that there should be a preference for any transportation option that might have near-zero external costs, or even any transportation option that might have lower total social costs. To make these evaluations, the dollar value of all the benefits as well as the dollar value of all the costs must be estimated.

3.3 Average Cost as a Proxy for Marginal Cost

To be relevant, a social-cost estimate must apply to a potential (realistic) context or policy. Given this, the question might asked: Is an estimate of the social cost of all motor vehicle use, which reveals what would be saved without a motor vehicle system and no motor vehicle use at all, of practical interest? Certainly an estimate of the total dollar cost of all motor vehicle use, by itself, is useful only for research purposes, not for evaluating any policy, because a policy requiring the elimination of motor vehicle use

TABLE I

Classification of the Social Costs of Motor Vehicle Use[a]

	Private sector monetary costs			External costs of motor vehicle use		
Personal non-monetary costs	**Goods and services produced and priced in the private sector (estimated net of producer surplus and taxes and fees)**	**Goods bundled in the private sector**	**Goods and services provided by government**	**Monetary externalities**	**Nonmonetary externalities**	**Nonmonetary impacts of the motor vehicle infrastructure**[b]
Travel time (excluding travel delay imposed by others) that displaces unpaid activities	Costs usually included in GNP-type accounts:	Annualized cost of nonresidential, offstreet parking included in the price of goods and services or offered as an employee benefit	Annualized cost of public highways (including on-street parking and off-street private investment)	Monetary costs of travel delay imposed by others (extra consumption of fuel, and foregone paid work)	Accidental pain, suffering, death, and lost nonmarket productivity, not accounted for by the economically responsible party	Land-use damage: habitat, species loss due to highways, MV infrastructure
Accidental pain, suffering, death, and lost nonmarket productivity inflicted on oneself	Annualized cost of the fleet (excluding vehicles replaced as a result of motor vehicle accidents)	Annualized cost of residential, offstreet parking included in the price of housing	Annualized cost of municipal and institutional off-street parking	Accident costs not accounted for by the economically responsible party (property damage, medical, productivity, and legal and administrative costs)	Travel delay (imposed by other drivers) that displaces unpaid activities	The socially divisive effect of roads as physical barriers in communities
Personal time spent working on MVs and garages, refueling MVs, and buying and disposing of MVs and parts	Cost of transactions for used cars	Annualized cost of roads provided or paid for by the private sector and recovered in the price of structures, goods, or services	Highway law enforcement and safety	Expected loss of GNP due to sudden changes in oil prices	Air pollution: effects on human health, crops, materials, and visibility[d]	Esthetics of highways and vehicle and service establishments
	Parts, supplies, maintenance, repair, cleaning, storage, renting, towing, etc. (excluding parts and services in the repair of vehicles damaged in accidents)		Regulation and control of MV air, water, and solid waste pollution	Price effect of using petroleum fuels for motor vehicles: increased payments to foreign countries	Global warming due to fuel cycle emissions of greenhouse gases (U.S. damages only)	
	Motor fuel and lubricating oil (excluding cost of fuel use attributable to delay)		MV and fuel technology R&D		Noise from motor vehicles	
	Motor vehicle insurance (administrative and management costs)				Water pollution: effects of leaking	

MV noise and air pollution inflicted on oneself	Priced private commercial and residential parking (excluding parking taxes) Costs not included in GNP-type accounts: Travel time (excluding travel delay imposed by others) that displaces paid work Overhead expenses of business and government fleets Private monetary costs of motor vehicle accidents (including user payments for cost of motor vehicle accidents inflicted on others, but excluding insurance administration costs)	for oil used in other sectors (not an external cost internationally) Monetary, non-public-sector costs of fires and net crimes[c] related to using or having MV goods, services, or infrastructure Police protection (excluding highway patrol); court and corrections system (net of cost of substitute crimes) Fire protection Motor-vehicle-related costs of other agencies Military expenditures related to the use of Persian Gulf oil by motor vehicles Annualized cost of the Strategic Petroleum Reserve	storage tanks, oil spills, urban runoff, road deicing Nonmonetary costs of fires and net crimes[c] related to using or having MV goods, services, or infrastructure Air pollution damages to ecosystems other than forests, costs of motor vehicle waste, vibration damages, fear of MVs and MV-related crime

[a] Abbreviations: MV, motor vehicle; GNP, gross national product; R&D, research and development.

[b] Although these are nonmonetary environmental and social costs of total motor vehicle use, they are not costs of marginal motor vehicle use, and hence technically are not externalities.

[c] These should be classified not as external costs, within an economic framework, but rather as costs of illegal or immoral behavior, within a framework that encompasses more than just economic criteria. However, regardless of how these are classified, they in fact are related to using or having motor vehicle goods, services, or infrastructure.

[d] The cost of crop loss, and some of the components of other costs of air pollution (e.g., the cost of medical treatment of sickness caused by motor vehicle air pollution), probably should be classified as monetary externalities.

is highly unlikely. However, an estimate of the cost of all motor vehicle use (i.e., as explained previously, an estimate of the annualized cost of the entire system) will be useful as a starting point to the extent that it accurately can be scaled down to an estimate of the cost of a more realistic "project size." That is, if the true cost of a proposal to increase the motor vehicle system and its use by 10% is approximately 10% of the cost of the entire motor vehicle system as estimated here, then the estimate of the cost of the entire system will be a useful starting point for evaluating the proposal.

Do social costs of motor vehicle use in fact scale linearly with output? In economic terms, the question can be stated as follows: Is the total cost function linear, so that marginal (or incremental) cost per unit is constant and equal to average cost per unit (where the average cost per unit is equal to the total cost divided by the total number of units)? In most cases, the answer is "not exactly," because most types of costs probably are not strictly a linear function of output. For example, it is known that the nonmarket costs of air pollution are a nonlinear function of motor vehicle pollution, and that congestion delay costs are a nonlinear function of motor vehicle travel. Still, even though most costs of motor vehicle use are not strictly a continuous linear function of motor vehicle use, down to the mile or gram or decibel or minute, in at least some scenarios of relatively large changes in motor vehicle use, the average cost ratio might be a serviceable approximation of the actual long-run marginal ratio of interest. At a minimum, some of the data and methods used in an analysis of total cost will be useful in an analysis of marginal cost.

3.4 Classification of Components of the Total Social Cost

In any social-cost analysis, the individual cost components, or cost items, should be identified and classified in consonance with how the cost estimates will be used. As discussed previously, estimates of the total social cost of motor vehicle use legitimately can be applied toward three ends: to ensure that motor vehicles are used efficiently, to evaluate transportation investments, and to prioritize efforts to reduce costs. Of these uses, only the first one, efficiency of use, comes with a set of widely accepted organizing principles. That is, when estimating costs in order to help policymakers improve the efficiency of the use of the transportation system, then costs should be organized into groups with similar characteristics with respect to this purpose of understanding the economic efficiency of motor vehicle use (for example, costs that are priced and efficiently allocated, unpriced but efficiently allocated, or unpriced and inefficiently allocated). Therefore, in Table I, the costs of motor vehicle use are grouped with respect to efficiency of use. However, there is an additional criterion that has nothing to do with efficiency per se: whether a cost is monetary, which means "valued in dollars in real markets." (For example, gasoline and parking are monetary costs, but air pollution is nonmonetary cost.) The distinction between monetary and nonmonetary costs is important methodologically, because it is much harder to estimate the nonmonetary costs. These distinctions result in the six categories of Table I. These are reviewed briefly next.

4. COMPONENTS OF THE SOCIAL COST OF MOTOR VEHICLE USE

The discussions that follow are based on the classification set forth in Table I.

4.1 Column 1: Personal Nonmonetary Costs

Personal nonmonetary costs are those unpriced costs of motor vehicle use that a person self-imposes as a result of the decision to travel. The largest personal costs of motor vehicle use are personal travel time in uncongested conditions and the risk of getting into an accident that involves nobody else. With respect to economic efficiency, the particular issue in the case of personal nonmonetary costs is whether drivers fully recognize the personal nonmarket costs that they face. If a person does not correctly assess these costs (that is, if the true value to the user does not equal the true cost to the user), then the person will drive more or less than would be the case if he or she were fully informed and rational. For example, people who, on account of ignorance or poor judgment, underestimate their risk of falling asleep at the wheel, once in a while will make trips for which the real but underestimated risk-cost exceeds the value, and which, consequently, in principle, should not be made.

4.2 Column 2: Priced Private-Sector Motor Vehicle Goods and Services, Net of Producer Surplus and Taxes and Fees

Priced private-sector costs are related to goods and services that are explicitly priced in private markets:

motor vehicles, motor fuel, maintenance and repair, insurance, and so on. Because these goods and services are bought and sold by private individuals, they are the most familiar components of the total social cost. Also, a portion of these goods and services constitute the "transportation" subaccounts of the gross national product (GNP) accounting.

The social cost of motor vehicle goods and services supplied in private markets can be calculated by starting with known prices and quantities sold, but is not equal to prices multiplied by quantities. Rather, the social (economic) cost is the area under what is called the private supply curve: the dollar value of the resources that a private market allocates to supplying vehicles, fuel, parts, insurance, and so on. To estimate this area, two items must be subtracted from total price-times-quantity revenues, i.e., producer surplus, and taxes and fees. (It is appropriate to start with revenues because prices and quantities can be observed, but not supply curves.) The producer surplus is deducted because it is defined as revenue ("profit," in lay terms) in excess of economic cost, and hence is a noncost wealth transfer from consumers to producers. Taxes and fees assessed on producers and consumers are deducted because they either are transfers from producers and consumers to government, or else are economically inefficient government charges for government services.

The deduction of producer surplus is not merely a theoretical aside: it bears directly on comparisons of alternatives. For example, in comparing the cost of motor fuel with the cost of alternative energy sources, it will not do to count the revenues received by oil producers as the cost, because a large portion of the revenues received by oil producers is extra "profit," far in excess of costs and "normal" profit (which is why major oil-producing countries are so wealthy). It also is important to note that the prices and quantities are realized in private markets rarely if ever are optimal in economic terms, not only because of taxes and fees, which cause economic distortions, but because of imperfect competition (e.g., monopoly), standards and regulations that affect production and consumption, externalities (discussed later), and poor information. Put another way, there is no simple dichotomous world in which private-sector prices are perfect and can be left alone and all other prices (or nonprices) need to be fixed. Rather, there are a variety of market imperfections in every sector of the economy, including the most competitive, unregulated private sectors, and a corresponding range of issues pertaining to pricing, taxation, regulation, and so on. In some cases, there may be as much concern about the market price of tires as there is about the best way to charge motor vehicle users for roads or about the complete lack of a price on motor vehicle emissions.

4.3 Column 3: Bundled Private-Sector Costs

Some very large costs of motor vehicle use are not explicitly priced as separate costs of motor vehicle use. Foremost among these are the cost of free, nonresidential parking, the cost of home garages, and the cost of local roads provided by private developers. However, all of these costs are included in the price of "packages" (such as houses and goods) that are explicitly priced. This bundling is not necessarily economically inefficient; in principle, a producer will bundle a cost and not price it separately if the administrative, operational, and customer (or employee) costs of collecting a separate price exceed the benefits. In a perfect market, it can be presumed that any observed bundling is economically efficient and that it would be economically undesirable to actually mandate unbundling. As concerns economic efficiency, then, the question is whether taxes or regulations (such as requirements that office buildings have a certain number of parking spaces, regardless of what the building owners think) and the like distort the decision to bundle and whether a supplier is correct in his assessment of the costs and benefits of bundling. To the extent that taxes and standards are distorting the market for parking, the ideal remedy is to eliminate the inefficient taxes and standards, not to force parking costs to be unbundled.

4.4 Column 4: Motor Vehicle Goods and Services Provided by the Public Sector

Government provides a wide range of infrastructure and services in support of motor vehicle use. The most costly of these is the capital of the highway infrastructure. Government costs are categorized separately because generally they either are not priced, or else are priced but not in the most efficient manner. Note that, whereas all government expenditures on highways and the highway patrol are a cost of motor vehicle use, only a portion of total government expenditures on local police, fire, corrections, jails, and so on is a cost of motor vehicle use. The portion of these expenditures that is a cost—a long-run or average cost, anyway—of motor

vehicle use has been estimated. This sort of allocation is valid for expenditures (such as for police protection) that arguably are economic costs of motor vehicle use, in the sense of "cost" as used previously. (For example, motor vehicle use results in motor vehicle-related crime, which consumes police-protection services, and motor vehicle fires, which consume fire-protection services.) Publicly provided goods and services that are not priced at all are arguably also monetary externalities, which are shown in column 5 of Table I and are discussed next. Those items that might be both a publicly provided good or service and a monetary externality are shown to straddle columns 4 and 5 in Table I.

4.5 Columns 5–7: Monetary and Nonmonetary Externalities

An externality is a cost or benefit that an action of person A imposes on person B, but that person A does not account for in his or her deliberations regarding the utility (usefulness) of the action. Environmental pollution, traffic delay, pain and suffering due to accidents, and the loss of GNP due to sudden changes in the price of oil are common examples of externalities. Some analysts distinguish "monetary externalities," which are valued in dollars in markets, from "nonmonetary externalities," which are not valued in any markets, even indirectly. Although monetary external costs are valued in markets, they are completely unpriced from the perspective of the responsible motor vehicle user and hence are externalities. The clearest example, shown in column 5 of Table I, is accident costs that are paid for in dollars by those who are not responsible for the accident. For example, vehicular repair costs inflicted by uninsured motorists are valued explicitly in dollars in private markets, but are unpriced from the perspective of the uninsured motorist responsible for the accident.

With respect to economic efficiency, the concern here, as with any external cost, is that the costs are not priced at all, and hence are associated with consumption that is larger than is socially desirable. The intuition behind this is straightforward: if people pay a price that is less than the full cost to society, they will consume more than they would if the price were raised to reflect the external costs. The largest monetary externalities (and hence the largest potential sources of economic efficiency) are those resulting from motor vehicle accidents and congestion on the highways.

Most environmental damages are nonmonetary externalities. Environmental costs include those related to air pollution, global warming, water pollution, and noise due to motor vehicles. To estimate these costs, complex physical processes and biological responses must be modeled and the dollar value of the responses must then be estimated. (The valuation step, which often is quite difficult, is required for nonmonetary but not for monetary externalities; hence the distinction between monetary and nonmonetary costs.) By far the largest environmental externality of motor vehicle use is the cost of particulate air pollution. Interestingly, a typically overlooked and completely unregulated emissions source, particulate matter kicked up from the road bed by passing vehicles, may be one of the largest sources of pollution damages—much larger than damage from ozone.

5. RESULTS OF AN ANALYSIS

The results of a comprehensive analysis of most of the costs of Table I are summarized by aggregate cost category in Table II. Note that the aggregated totals are shown here in order to provide a sense of magnitudes, not because such aggregated totals are inherently useful. Indeed, as discussed next, care must be taken to avoid misusing estimates of the total social cost of motor vehicle use. Table II also shows two subtotals of interest: all monetary costs (those that are valued in dollars in markets) and costs that normally are included in GNP-type accounts of the economic value of transportation. Costs normally included in GNP-type accounts are of interest because they are the explicitly priced private-sector costs and hence represent what most people readily identify as "costs of motor vehicle use." It can be seen that these are a small fraction of the total social costs of motor vehicle use.

An estimate of user payments for public highway and services is shown at the bottom of Table II. An important caveat regarding the use of this estimate is discussed in the next section.

5.1 Allocation of Costs to Individual Vehicle Categories

All of the costs shown in Table II pertain to all motor vehicles (autos, trucks, and buses). Although it can be interesting to estimate the cost of all motor vehicle use, it typically will be more useful to estimate the

TABLE II

Summary of the Social Costs of Motor Vehicle Use in the United States, 1990–1991

	Total cost ($10^9$$)		Percentage of total	
Cost item	Low	High	Low	High
Personal nonmonetary costs of motor vehicle use	527	968	32	29
Motor vehicle goods and services produced and priced in the private sector (estimated net of producer surplus, taxes, and fees)	827	980	49	30
Motor vehicle goods and services bundled in the private sector	76	279	5	8
Motor vehicle infrastructure and services provided by the public sector[a]	131	247	8	7
Monetary externalities of motor vehicle use	43	104	3	3
Nonmonetary externalities of motor vehicle use[b]	68	730	4	22
Grand total social cost of highway transportation	1673	3308	100	100
Subtotal: costs of usually included in GNP-type accounts	559	674	33	20
(2) + (3) + (5) subtotal: monetary cost only	1077	1610	64	49
Item: payments by motor vehicle users for the use of public infrastructure and services	114	206		

[a] Includes items in Table I that straddle columns 4 and 5.

[b] Includes motor vehicle noise and air pollution inflicted on oneself (from column 1 of Table I). Excludes nonmonetary impacts of the motor vehicle infrastructure (column 7 in Table I) and air pollution damages to ecosystems other than forests, costs of motor vehicle waste, vibration damages, and fear of motor vehicles and motor-vehicle-related crime (last item in column 6 of Table I).

cost of different classes of vehicles or of different fuel types, because analysts, policymakers, and regulators typically are interested in specific classes of vehicles and specific fuels, rather than all motor vehicles as a group. (For example, pollution regulations are set for individual classes of vehicles, not for all motor vehicles as a class.) For some cost items, such as the some of the costs of air pollution, analysts have estimated marginal costs as a function of vehicle type and other specific characteristics of motor vehicles and their use. For example, noise costs are estimated as a function of vehicle type, vehicle speed, distance from the road, the presence of sound barriers, housing values, and many other parameters. Air pollution costs are estimated as a function of vehicle characteristics, urban population, weather variables, and other parameters. Road maintenance and repair costs are estimated as a function of vehicle weight, traffic volume, and other factors.

If marginal cost functions are not available, total costs can be allocated to particular vehicle classes or uses on the basis of general cost-allocation factors. A cost-allocation factor shows the share of a particular vehicle class of some general measure of motor vehicle use. For example, it shows the share of light-duty gasoline autos of total vehicle miles of travel, or the share of heavy-duty diesel vehicles of total motor vehicle expenditures for maintenance and repair. The use of these allocation factors is straightforward. For example, the heavy-duty diesel vehicle fraction of total vehicle ton-miles per axle, multiplied by any total motor vehicle cost that is a function of vehicle ton-miles per axle, yields the amount of that cost that is assignable to heavy-duty diesel vehicles. Thus, if total expenditures for highway repair are known, and it is believed that highway repair costs are related to ton-miles of travel per axle, the ton-mile/axle allocation factors can be used to allocate the total expenditures to individual vehicle classes.

5.2 How the Results of a Social-Cost Analysis Should Not Be Used

The legitimate uses of estimates were discussed in previous section. Caution against several common misuses of these estimates is necessary, however. First, the temptation to add up all of the imperfectly priced costs and to express the total per gallon of gasoline, as if the optimal strategy for remedying every inefficiency of motor vehicle use was simply to raise the tax on gasoline, should be resisted. The economically optimal strategy for dealing with imperfect prices is considerably more complex than this. In the first place, some sources of economic

inefficiency, such as imperfect competition and a distortionary income tax policy, are not externalities, and hence are not properly addressed by taxation at all. Taxation can be an appropriate remedy for external costs, but it is a particular kind of taxation that is appropriate (one in which the tax is equal to the actual marginal external cost), and it turns out that there is not a single external cost, with the possible exception of CO_2 emissions from vehicles, that in principle is best addressed by a gasoline tax. For example, an optimal air pollution tax would be a function of the amounts and types of emissions, the ambient conditions, and the size of the exposed population; it would not be proportional to gasoline consumption.

Second, it might be misleading to compare the total social cost of motor vehicle use with the GNP of the United States, because the GNP accounting is quite different from and generally more restricted than social-cost accounting. Most importantly, the GNP does not include nonmarket items. Third, the considerable uncertainty in most social-cost estimates should be represented properly. With regard to the estimates of Table II, the uncertainty actually is greater than is implied by the low and high ranges, even where the high is much higher than the low, because the estimates shown do not include every conceivable component or effect of every cost, and do not always accommodate the entire span of data or opinions in the literature. (Also, the costs "not estimated" in column 6 of Table I should not be assumed to be trivial.) Fourth, although Table II shows highway user tax and fee payments and the government-provided motor vehicle goods and services to which the user payments nominally apply, it may not be economically meaningful to compare the user payments with the government expenditures. Most emphatically, it is not true that any difference between user payments and government expenditures is a source of economic inefficiency that must be eliminated in order to have efficient resource use. This is because economic efficiency does not require that the government collect from users revenues sufficient to cover costs; rather, efficiency results from following certain pricing rules, which when applied to goods such as highways need not result in user revenues covering costs. An efficient pricing scheme for government goods and services would look nothing like the present tax and fee system. Consequently, a comparison of current user tax and fee payments with current government expenditures reveals little, if anything, quantitative about economic efficiency. However, such comparisons may be of use in analyses of equity (questions of who pays for what), which is not addressed here.

Finally, estimates of *the total social* cost of motor vehicle use may be of use in the analysis of a particular policy or investment decision only if it is believed that the marginal costs associated with the particular policy or decision are reasonably close to average costs, which can be calculated easily from total costs. In this respect, average cost estimates derived from the results of Table II will be less and less applicable as times and places increasingly different from the United States in 1990 and 1991 (the basis of the estimates of Table II) are considered. However, even if total-cost estimates per se are irrelevant, the underlying data, methods, and concepts (which are presented in the sources listed in the reading list at the end of this article) might be useful in an analysis of specific pricing policies or investments.

6. SUMMARY

The social costs of motor vehicle use in the United States have been classified, discussed, and estimated. The analyses, which present some first-cut estimates of some of the costs, are meant to inform general decisions about pricing, investment, and research, providing a conceptual framework for analyzing social costs and developing analytical methods and data sources.

It should be clear that a social-cost analysis cannot describe precisely what should be done to improve the transportation system in the United States. There are several kinds of inefficiencies in the motor vehicle system, and hence several kinds of economically optimal measures. The magnitude of these inefficiencies is difficult to estimate accurately. Moreover, society cares at least as much about equity, opportunity, and justice as it does about efficiency. In sum, a total social-cost analysis contributes only modestly to one of several societal objectives regarding transportation.

SEE ALSO THE FOLLOWING ARTICLES

Bicycling • Cost–Benefit Analysis Applied to Energy • External Costs of Energy • Fuel Economy Initiatives: International Comparisons • Internal Combustion Engine Vehicles • Leisure, Energy Costs of •

Lifestyles and Energy • Material Use in Automobiles • Suburbanization and Energy • Transportation Fuel Alternatives for Highway Vehicles

Further Reading

Baumol, W. J., and Oates, W. E. (1988). "The Theory of Environmental Policy," 2nd ed. Cambridge University Press, New York.

Delucchi, M. A. (1998). "The Annualized Social Cost of Motor-Vehicle Use, 1990–1991: Summary of Theory, Methods, Data, and Results." UCD-ITS-RR-96-3 (1). Institute of Transportation Studies, University of California, Davis, California.

Friedrich, R., and Bickel, P. (eds.). (2001). "Environmental External Costs of Transport." Springer-Verlag, Stuttgart, Germany.

Greene, D. L., Jones, D. W., and Delucchi, M. A. (eds.). (1997). "Measuring the Full Social Costs and Benefits of Transportation." Springer-Verlag, Heidelberg, Germany.

Hohmeyer, O., Ottinger, R. L., and Rennings, K. (eds.). (1996). "Social Costs and Sustainability, Valuation and Implementation in the Energy and Transport Sector." Springer-Verlag, Berlin, Germany.

Hohmeyer, O., Ottinger, R. L., and Rennings, K. (eds.). (1996). "Social Costs and Sustainability, Valuation and Implementation in the Energy and Transport Sector." Springer-Verlag, Berlin, Germany.

Murphy, J. J., and Delucchi, M. A. (1998). A review of the literature on the social cost of motor-vehicle use in the united states. *J. Transport. Statistics* **1**(1), 15–42.

Multicriteria Analysis of Energy

R. RAMANATHAN
Sultan Qaboos University
Muscat, Sultanate of Oman

1. **Introduction**
2. **Why Is Multicriteria Analysis Necessary for Energy Issues?**
3. **MCDM: Terminologies, Classification, and Some General Characteristics**
4. **An Overview of Some MCDM Methods and Their Energy Applications**

Glossary

alternative One of several things or courses of action to be chosen by the decision maker. It is the most fundamental entity in a multicriteria decision-making model. Alternatives are also called solutions, especially when dealing with continuous variables, in the mathematical programming context.

attribute A surrogate measure of performance used to represent a criterion. Usually quantified, using some measurable unit, to identify the consequences arising from implementation of any particular decision alternative. Thus, whereas warmth is a criterion, temperature, measured in a suitable (say Celsius or Fahrenheit) scale, is an attribute.

criterion A tool allowing comparison of alternatives according to a particular significance axis or point of view.

decision maker (DM) The individual whose subjective opinions are considered in arriving at a solution for a decision problem.

decision problem A decision-making problem is characterized by the need to choose one or a few from among a number of alternatives.

group decision making (GDM) A situation in which the opinions of more than one decision maker need to be considered in arriving at a solution for a decision problem.

multiattribute decision making (MADM) Involves cases in which the set of alternatives is defined explicitly by a finite list from which one or a few alternatives should be chosen that reflect the DM's preference structure.

multicriteria decision making (MCDM) Making decisions in the face of multiple conflicting criteria. MCDM situations are also referred to herein as multicriteria analysis (MCA) decision situations.

multiobjective decision making (MODM) Involves cases in which the set of alternatives is defined implicitly by a mathematical programming structure with objective functions. Such alternatives are usually defined in terms of continuous variables, which results in an infinite number of alternatives. Multiobjective decision making is also referred to in the literature as multiobjective mathematical programming, multiobjective optimization, vector optimization, or simply multiobjective programming.

nondominated alternative An alternative that is clearly superior to others with respect to at least one criterion. The nondominated alternative is also called an efficient alternative, a noninferior alternative, or a Pareto-optimal alternative.

objective Usually represents the direction of improvement of the attributes; mainly used in mathematical programming problems. A maximizing objective refers to the case in which "more is better," whereas a minimizing objective refers to the case in which "less is better." For example, profit is an attribute, whereas maximizing profit is an objective.

preemptive priorities A situation in which it is assumed that a higher ranked objective is infinitely more important than any of the lower ranked objectives.

stakeholders People who affect and are affected by the decision under consideration.

The field of energy includes a number of issues that require analysis considering several criteria. For example, decisions on almost all energy projects require consideration of several, mostly conflicting, criteria related to economic development arising out of increased energy use, and environmental degradation arising out of increased energy production. Accordingly, multicriteria analysis has been applied to a number of energy-related problems.

 © 2004 Elsevier Inc. All rights reserved.

1. INTRODUCTION

Multicriteria analysis (MCA), or multicriteria decision making (MCDM) as it is more often called, is a subfield of operations research. It is a special case of the so-called decision-making problems. A decision-making problem is characterized by the need to choose one or a few from among a number of alternatives. The person who is to choose the alternatives is normally called the decision maker (DM). The preferences of the DM will have to be considered in choosing the right alternative(s). In MCDM, the decision maker chooses the most preferred alternative(s) on the basis of two or more criteria.

2. WHY IS MULTICRITERIA ANALYSIS NECESSARY FOR ENERGY ISSUES?

Actually, single-objective programming can be considered as a specific case of MCA. Though some problems have been considered with a single criterion (which is captured by the single-objective function of the linear programming models), many of the issues often have impact on more than one criterion. When it can be safely assumed that only one criterion is overridingly important, linear programming approaches that use the single-objective function should be employed. However, if impacts should be considered in terms of several criteria, the MCA methods should be employed. Further, note that MCA is much more general, involving not only multiobjective programming models, but also non-programming-based models.

2.1 Need for Multiple Criteria

The issues of the energy sector that can be tackled by quantitative methods (including single-criterion linear programming and MCA methods) abound, from the location of power plants, to transmission and distribution systems, to the choice of demand-side management programs, to the choice of options that mitigate global warming. Financial criteria are the most dominant considerations usually evaluated in single-criterion analysis of energy issues. However, the growing importance of environmental impacts of decisions in the field of energy has increased the importance assigned to environment-related criteria such as air pollution (emission of sulfur dioxide, nitrogen oxides, etc.), water pollution, human health, global warming, acid rain, oil spills, preservation of biodiversity, etc., increasing the need for multicriteria analysis of energy-related decision problems. Of course, other criteria, such as quality and reliability, are also relevant for specific problems.

Multicriteria analysis of energy problems encompasses technological, economic, social, environmental, risk, and financial criteria in making energy decisions. Multicriteria decisions require certain approaches that are not normally used in single-criterion decision-making situations. For example, it is necessary to carry out a trade-off analysis when distinctive characteristics in terms of cost, environmental emissions, human health, etc. are compared and weighed against each other. These comparisons equate to comparing apples and oranges, but are the necessary ingredient of any multicriteria analysis. The trade-off analysis is necessary only when the criteria employed are conflicting, i.e., when there is no alternative that performs well in terms of all the criteria. In some situations, such trade-offs may not be necessary. For example, in the power sector, there may be alternatives that have beneficial impacts on both environmental and economic objectives—most energy-efficient investments that are economically justifiable also bring about a reduction in emissions, and thus they score well in terms of both the criteria, i.e., environmental quality and economic efficiency. The two criteria are said to be not conflicting in this case. However, in most of the energy-related problems, alternatives are often evaluated in terms of conflicting criteria. For example, wind power plants score well in terms of environmental criteria but are more expensive than many other power generation options.

2.2 Group Decision Making

Another feature of multicriteria analysis is group decision making. Invariably, multiple criteria have to be considered to arrive at a decision, because different criteria are important to different people, who both affect and are affected by the decision under consideration. These people are normally referred to as the stakeholders. There may be different groups of stakeholders for a decision-making situation. For example, government and local authorities are normally stakeholders for most energy-related decisions because they would like to ensure that the resulting decision follows legal guidelines and is socially acceptable. A company, entrepreneurs, or potential investors are also stakeholders; here the greater interest is in the economic

and finance-related criteria. For a utility company, for example, the decision made about the problem of siting a power plant would have an economic impact. Environmentalists, public pressure groups (such as nongovernmental organizations and local media), and the local public, also stakeholders, may wish to ensure that the environment is not adversely affected. In some cases, opinions of experts and academia may also be important. Hence, involvement of all the stakeholders is important in arriving at a large-scale energy-related decision. And, the opinion of a stakeholder group should be synthesized from the opinions expressed by individual members of the group (such as individual local people, when the stakeholder group is the local public). MCA has been considered a valuable tool for facilitating early involvement of all stakeholders in the decision-making process, and enhances the fairness and transparency of the procedures for complex decision situations, including energy-related decision situations.

Ideally, the final decision of a stakeholder group should be arrived at by consensus. However, due to the conflicting nature of the criteria adopted by different people (or groups of people), especially in energy-related decision situations, it may not be possible to arrive at a consensus decision. In such cases, special procedures are required to estimate the most preferred solution for the group as a whole, from the opinions expressed by the group members. Most of the multicriteria methods provide such special procedures. Due to the need to consider the subjective opinions of a number of people, most of the MCA methods have been designed to be interactive.

3. MCDM: TERMINOLOGIES, CLASSIFICATION, AND SOME GENERAL CHARACTERISTICS

The field of multicriteria decision making has been succinctly defined as making decisions in the face of multiple conflicting criteria. Some special issues of journals have been devoted to the field of MCDM, including *Management Science* (Vol. 30, No. 1, 1984), *Interfaces* (Vol. 22, No. 6, 1992) (devoted to decision and risk analysis), and *Computers and Operations Research* (Vol. 19, No. 7, 1994). The *Journal of Multi-Criteria Decision Analysis*, since 1992, has published articles entirely devoted to MCDM. Special issues of the *Journal of the Operational Research Society* (April, 1982), and of *Interfaces* (November–December, 1991) provide a range of applications. Issue No. 2, Volume 133 (January 2001) of the *European Journal of Operational Research* is a special issue on goal programming.

3.1 MCDM Terminologies

Several terminologies are normally used when dealing with a decision problem that has multiple criteria. The terms "goals," "objectives," "criteria," and "attributes" are commonly found in the MCDM literature and can be used with interchangeable ease. The general meaning of these words is similar in most cases.

Alternatives are the most fundamental entities in a MCDM model. They are normally compared with each other in terms of criteria. Identification of criteria for a particular problem is subjective, i.e., varies for each problem. Criteria are normally developed in a hierarchical fashion, starting from the broadest sense (usually called the goal of the problem) and refined into more and more precise sub-goals and sub-sub-goals. In general, some rules should be followed in identifying criteria for any decision problem. They have to be mutually exclusive or independent, collectively exhaustive, and should have operational clarity of definition.

The criteria of a decision problem are usually very general, abstract, and often ambiguous, and it can be impossible to associate criteria directly with alternatives. Attributes are objective and measurable features of the alternatives, and are sometimes used to describe criteria. Objectives, used mainly in mathematical programming problems, represent directions of improvement of the attributes. The term "criterion" is a general term comprising the concepts of attributes and objectives. It can represent either attribute or objective, depending on the nature of the problem. Due to this, MCDM is considered to encompass two distinct fields, namely multiattribute decision making and multiobjective decision making. These fields are discussed in the next section.

3.2 A Classification of Different MCDM Approaches

The broad area of MCDM can be divided into two general categories, multiattribute decision making (MADM) and multiobjective decision making (MODM). MADM involves cases in which the set of alternatives is defined explicitly by a finite list

from which one or a few alternatives should be chosen that reflect the decision maker's preference structure. MODM involves cases in which the set of alternatives is defined implicitly by a mathematical programming structure with objective functions. Such alternatives are usually defined in terms of continuous variables, which results in an infinite number of alternatives.

3.3 Multiobjective Decision Making—Some Basic Concepts

In this section, a comparison of a multiobjective decision making problem with a single-objective decision-making problem is used to introduce certain key concepts frequently used in MCDM literature. Note that many of the concepts introduced in this section are also applicable to MADM problems.

A single-objective decision-making (SODM) problem is generally written in the following form:

$$\begin{aligned} &\text{maximize/minimize } Z = CX \\ &\text{subject to } AX = b,\ X \geq 0, \end{aligned} \tag{1}$$

where Z is the objective function. The set of equations, $AX = b,\ X \geq 0$, represents the constraint set of the problem. Methods for solving single-objective mathematical programming problems have been studied extensively for the past 40 years. However, almost every important real-world problem involves more than one objective. A general multiobjective decision-making (MODM) problem has the following form:

$$\begin{aligned} \text{maximize/minimize } Z_1 &= C_1 X \\ \text{maximize/minimize } Z_2 &= C_2 X \\ \text{maximize/minimize } Z_n &= C_n X \\ &\vdots \\ \text{subject to } AX &= b,\ X \geq 0, \end{aligned} \tag{2}$$

where the values Z_i, $i = 1, 2, \ldots, n$, represent the n objective functions. The set of equations, $AX = b$, $X \geq 0$, represents the constraint set of the MODM problem. Thus the MODM problem is similar to a SODM problem, except that it has a stack of objective functions instead of only one.

Let us consider a simplified example involving choice of a coal technology for electric power generation. Let the objectives be as follows:

$$\begin{aligned} \text{Minimize } Z_c &= \text{investment cost.} \\ \text{Maximize } Z_G &= \text{savings of emissions of global pollutants } (CO_2 \text{ here}) \\ \text{Maximize } Z_L &= \text{savings of emissions of local pollutants } (SO_2 \text{ here}) \\ \text{Subject to some constraints.} & \end{aligned} \tag{3}$$

Let us assume that we have some technologies (normally called alternatives of the decision problem), say pulverized fuel (PULV) technology, pressurized fluidized bed combustor (PFBC) technology, and integrated coal gasification combined cycle (IGCC) technology, that satisfy the constraints. This means that these are the three feasible technologies (or alternatives). The performance of the technologies in terms of the three criteria is shown in Table I. Note that the savings of emissions of CO_2 and SO_2 are lower and investment costs higher for PFBC technology compared to IGCC technology. Thus PFBC technology fares, on all the three objectives, worse than IGCC technology and hence it should not be considered any more. In MCDM terminology, PFBC technology is said to be inferior to (or dominated by) IGCC technology. Technologies PULV and IGCC are incomparable, as neither of them is at least as good as the other in terms of all the objectives. Although PULV is cheaper in terms of the investment cost compared to IGCC, it results in smaller savings of the environmental emissions. Hence, PULV and IGCC are said to be noninferior technologies and PFBC is called an inferior technology. (Please note that, in practice, comparisons of the

TABLE I

Some Characteristics of Coal Technologies for Power Generation in India[a]

Technology	CO_2 savings (g/kWh)	SO_2 savings (g/kWh)	Investment cost ($/kW)
Pulverized fuel (PULV)	220	3.30	1202
Pressurized fluidized bed combustor (PFBC)	240	3.32	1894
Integrated coal gasification combined cycle (IGCC)	590	3.75	1578

[a] Values estimated by the author from miscellaneous sources.

different coal technologies will not be so simple. For example, PFBC technology may fare better than other technologies if some other criteria are considered.) Thus a noninferior alternative is one that is clearly superior with respect to at least one criterion. The noninferior alternative is also called an efficient alternative, nondominated alternative, or Pareto-optimal alternative. In any MODM exercise, the first task is to identify the noninferior options. Note that there may be more than one noninferior option for any multiobjective programming (MOP) problem.

Once the noninferior alternatives are identified, it is necessary to evaluate them so as to arrive at the best alternative. Because none of the noninferior alternatives will optimize (i.e., maximize/minimize) all the objective functions, it is necessary to identify those alternatives that best satisfy all the objective functions (in the opinion of the decision-maker). This has to be contrasted with single-objective optimization problems that are concerned with arriving at one or more optimal solutions. Thus, although the underlying philosophy in SODM problems is in optimizing solutions, MODM aims at identifying one or more satisficing solutions. One way of reducing the number of Pareto-optimal alternatives is by providing performance targets for each of the objectives. For example, if the DM specifies no interest in those technologies that cost more than \$1500/kW, then IGCC technology is automatically excluded from the DM preference set, and PULV technology now remains the only choice. This is a satisficing solution because it satisfies all the performance targets, but obviously it need not be an optimal solution. This is the basic idea of goal programming (discussed later). In general, identification of alternatives (technologies here) that are considered best by the decision maker requires additional information on the preference structure among the objectives, which is not necessary for SODM problems.

The preference structure can be ordinal or cardinal. This essentially involves identifying the "importance parameters" or "weights" associated with the objectives. Suppose it is possible for the DM to provide weights (w) to the three objectives as follows: $w_C = 5$, $w_S = 2$, $w_I = 8$, where the subscripts C, S, and I represent CO_2 savings, SO_2 savings, and investment cost, respectively. These weights may be specified exogenously before attempting to analyze the MOP problem. They imply the trade-offs between the objectives that the DM is willing to consider. It is necessary to have a trade-off because the objectives are conflicting: given the nondominated options, it is not possible to maximize one objective without getting reduction in the performance of other objective. The information as to how much of reduction in one objective can be tolerated to improve performance in another objective is called the trade-off information. Note that this trade-off information is subjective and depends on the particular DM for whom the MODM problem is being analyzed. The given structure means, in the opinion of the DM, that the objective Z_S (SO_2 savings) is the least important objective, a reduction in investment cost by \$1/kW is considered by the DM to be equivalent to an increase of $8/2 = 4$ g/kWh in SO_2 savings, an increase in CO_2 savings by 1 g/kWh is considered by the DM to be equivalent to an increase of 2.5 g/kWh in SO_2 savings, etc.

Given such cardinal weights, the MODM problem reduces to the SODM problem, with the single objective of minimizing $Z = \sum w_i Z_i$ subject to the constraints. In this case, IGCC technology has an aggregate score of about 15,582, and is preferable to PULV (which has an aggregate score of about 10,723). Thus, the MODM can be reduced to a SODM problem if such cardinal weights can be unambiguously specified. However, in practice, many DMs find it difficult to specify a precise set of weights. Imprecision occurs also because the different objectives have different performance measures or they are measured in incommensurable units. In situations when specification of cardinal weights is difficult, it may be easier to provide ordinal rankings or provide simple ranks of the objectives. For example, given the weight structure, even if the weights cannot be specified exactly, the ranking pattern may remain the same, i.e., Z_I is the most important of all the three objectives, followed in importance by Z_C, and Z_S is the least important objective, or

$$Z_I \succ Z_C \succ Z_S. \tag{4}$$

Here, the symbol $\succ$ stands for "is preferred to." Thus, $Z_I \succ Z_C$ means that Z_I is preferred to Z_C.

When it is assumed that no further trade-off is possible, this may be considered as a case of preemptive priorities, i.e., the higher ranked objective is assumed to be infinitely more important than any of the lower ranked objectives, or $w_I \ggg w_C \ggg w_S$. Preemptive priorities are more commonly denoted by P, i.e., $P_I > P_C > P_S$. Under these preemptive priorities, PULV technology is preferable to IGCC technology. In fact, PULV technology will be preferred even if we have a hypothetical technology with CO_2 savings of 500 g/kWh, SO_2 savings of

9 g/kWh, and investment cost of $1203/kW. This is because, under the assumptions of the preceding preemptive priorities, a reduction of $1/kW is worth infinitely more than any amount of CO_2 savings or SO_2 savings. This is the basic principle behind a version of goal programming, called preemptive goal programming.

3.4 Multiattribute Decision-Making Methods—Some Common Characteristics

In general, MADM methods have some common characteristics:

1. The persons or organizations that are the stakeholders of the decision problem being considered have to be identified. These people will act as the decision makers whose opinions will be elicited by the MCDM methods. This stage is equally applicable to MODM methods, especially while eliciting the preference information for choosing one or a few from the set of all nondominated alternatives (solutions).
2. A model of the decision problem has to be constructed. Typically, the model will consider the main goal of the problem, relevant criteria, attributes, and alternatives. For example, if it is desired to evaluate some selected technologies, it is important to identify the key criteria that distinguish the characteristics of technologies, and all the relevant technologies to be considered as alternatives in the model. The model is called a hierarchical model in some MCDM methods, because it represents a hierarchy of decisions—the goal of the exercise (say, selection of the most important energy technologies), the criteria (economic significance, social significance, etc.), and alternatives (the technologies to be evaluated).
3. Ideally speaking, criteria have to be mutually exclusive and collectively exhaustive. Collective exhaustivity is critical in some methods, such as in the analytic hierarchy process, because final rankings could change for the addition or deletion of alternatives at a later stage.
4. The relative importance of criteria, which are implicit trade-off data (see previous section) used in identification of the best alternative(s), has to be assessed. Although assessment of the relative importance of criteria forms an important step in many MCDM methods, some methods do not use this information (e.g., data envelopment analysis). The procedure for this assessment differs for each method and forms generally the single most important distinguishing feature of each method. The procedures will be described in more detail in the next section for some important MCDM methods, along with some prominent applications of the methods in the field of energy. The assessment can be performed either by an individual, or by representatives of conflicting values acting separately, or by those representatives acting as a group. Again, some methods have special procedures to aggregate the opinions of individuals to form a group opinion, though in general the group members should be encouraged to arrive at a single unanimous assessment.
5. Assessment of alternatives with respect to different criteria.
6. Aggregation of performance of alternatives with respect to all criteria to provide the overall performance measures of alternatives. Simple additive or multiplicative aggregation is used in many methods, but can also be more sophisticated (e.g., data envelopment analysis uses linear programming). The overall performance measures could also be considered as the rankings of the alternatives.

Note that many of these characteristics are also valid for the MODM problems, especially when the preferences of the DM are elicited.

4. AN OVERVIEW OF SOME MCDM METHODS AND THEIR ENERGY APPLICATIONS

Several surveys are available that describe many of the MCDM methods suggested in the literature. In the following discussion, four selected MCDM methods that have been applied to a number of problems in the field of energy will be briefly described. Due to space limitations, additional methods cannot be discussed here, but are available in the references listed at the end of this article. Software packages are available for facilitating the implementation of most of the methods, but, again owing to space limitations, they will not be discussed here.

4.1 MADM Methods

4.1.1 Multiattribute Utility Theory

Utility measures the subjective "worth" of an outcome, even one that is not a monetary value. Traditionally, utility functions are defined for stochastic problems that involve uncertainty. In the case

of deterministic problems, the term "value functions" is more commonly used. The utility or value functions may be thought of as evaluative mechanisms that can be used to measure the value of a particular solution.

Utility functions are defined in terms of uncertainty and thus tie in the decision maker's preferences under uncertainty, revealing the DM's risk preference for an attribute. An uncertain situation faced by a decision maker can be considered similar to a lottery—the DM can earn $\$X$ with a probability p, and earn $\$Y$ with probability $(1-p)$. In these situations, a rational decision maker is expected to maximize *the expected utility*, given by $\$[pX+(1-p)Y]$. Utility functions are assessed by giving the DM a sequence of choice between alternatives or between alternative lotteries. The responses are used to generate functions.

Multiattribute utility theory (MAUT) consists of assessing and fitting utility functions and probabilities for each attribute, and then using the functions and probabilities to come up with rankings of alternatives. The utility functions for each attribute are aggregated to get the overall utility function. At least two methods of aggregation are used in MAUT, additive and multiplicative. The additive aggregation is

$$\begin{aligned} U(A) &= \textstyle\sum_i w_i u_i(a_i) \\ \textstyle\sum_i w_i &= 1,\ w_i \geq 0, \end{aligned} \tag{5}$$

where $0 \leq u_i(a_i) \leq 1$; $u(a_i)$ is the utility function describing preferences with respect to the attribute i, a_i represents the performance of the alternative A in terms of the attribute i, the scaling factors w_i define acceptable trade-offs between different attributes, and $U(A)$ represents the overall utility function of the alternative A when all the attributes are considered together. This form of additive aggregation is valid if and only if the decision maker's preferences satisfy the mutual preferential independence. Suppose that there is a set of attributes, X. Let Y be a subset of X, and let Z be its complement, i.e., $Z = X–Y$. The subset Y is said to be preferentially independent of Z, if preferences relating to the attributes contained in Y do not depend on the level of attributes in Z.

Under certain assumptions, the multiplicative aggregation can be written as $U(A) = \Pi_i\ u_i(a_i)$. This multiplicative utility function requires stronger assumptions such as the Utility independence. The utility functions may be used as objective functions for solving mathematical programming problems. Utility theory has been criticized because of its "strict" assumptions, which are usually not empirically valid. Because of the strict assumptions, practical applications of MAUT are relatively difficult, though there are several practical successful applications of MAUT. This has led to some simplifications of the MAUT concepts. For example, the multiattribute value theory is a simplification of MAUT without uncertainty and risk. The simple multiattribute rating technique is another simplification that makes weaker assumptions while eliciting utilities. These are described in the next few subsections.

4.1.1.1 Multiattribute Value Theory Multiattribute value theory (MAVT) is a simplification of MAUT: MAVT does not seek to model the decision maker's attitude to risk. As a result, MAVT is considered to be easier to implement. Value theory assumes that it is possible to represent a decision maker's preferences in a defined context by a value function, $V()$, such that if alternative A is preferred to alternative B, then $V(A) > V(B)$. For this representation to be possible, the decision maker's preferences should satisfy two properties: the transitivity property (if A is preferred to B, B is preferred to C, then A should be preferred to C) and comparability (given two alternatives A and B, the decision maker must be able to indicate whether A is preferred to B, or B is preferred to A, or there is no difference between the two).

Note that the value function is an ordinal function, i.e., it can be used only to rank alternatives. In contrast, utility function is cardinal, i.e., it can be used to measure the strength of preference among alternatives. Multiattribute value theory explicitly recognizes that the decision maker will use many attributes (criteria) when evaluating a set of alternatives. For each attribute i, a partial value function $v_i(a_i)$ describing preferences with respect to the attribute i is assessed by the decision maker, where a_i represents the performance of the alternative A in terms of the attribute i. Then, the overall value function $V(A)$ of the alternative when all the attributes are considered together is normally obtained using the additive form: $V(A) = \Sigma_i\ v_i(a_i)$. This is more generally expressed as follows:

$$\begin{aligned} V(A) &= \textstyle\sum_i w_i v_i(a_i) \\ \textstyle\sum_i w_i &= 1,\ w_i \geq 0, \end{aligned} \tag{6}$$

where $0 \leq v_i(a_i) \leq 1$. As mentioned above, the scaling factors w_i define acceptable trade-offs between

different attributes. Again, this additive value function is appropriate if and only if the decision maker's preferences satisfy the so-called mutual preferential independence discussed earlier.

4.1.1.2 Simple Multiattribute Rating Technique The simple multiattribute rating technique (SMART) follows the steps described in the previous section for modeling a decision problem. It uses the simple weighting technique for the assessment of importance of criteria, and for the assessment of alternatives with respect to criteria. To rate (i.e., assess the importance of) criteria, the DM is asked to start by assigning the least important criterion an importance of 10. Then, the DM has to consider the next-least-important criterion, and ask how much more important (if at all) is it compared to the least important criterion, and assign a number that reflects that ratio. This procedure is continued till all the criteria are assessed, checking each set of implied ratios as each new judgement is made. The DM will be given the opportunity to revise previous judgments to make them consistent. Once the numbers are assigned, the relative importance of criteria is obtained by summing the importance weights, and dividing each by the sum. Thus, the relative importance of the criterion $j(w_j)$ is the ratio of importance weight of this criterion to the sum. Note that $\sum_j w_j = 1$ by definition.

Alternatives are rated with respect to each criterion in a similar fashion. Though MAUT requires the development of complex utility functions for each criteria, SMART prefers to produce the rating using a more straightforward approach: the DM is asked to estimate the position of the alternative of a criterion on a scale of 0 to 100, where 0 is defined as the minimum plausible value and 100 is defined as the maximum plausible value. Once the two measures are available, the overall performance of an alternative i can be aggregated using the simple weighted average,

$$U_i = \sum_j w_i u_{ij}, \tag{7}$$

where U_i is the overall performance rating of alternative i, w_j is the relative importance of criterion j, and u_{ij} is the rating of the alternative i with respect to the criterion j. The alternative that has the maximum U_i is the most preferred alternative to achieve the goal of the decision problem. The values of U_i can be used to provide the overall rankings of the alternatives.

MAUT, or its simplified versions MAVT or SMART, has been used for several practical applications. Siting and licensing of nuclear power facilities is an interesting application of these methods in the field of energy. Siting of nuclear power facilities is an extremely complex task because there are many interest groups with their own set of multiple objectives. Five categories of objectives were considered in the study: environmental, human safety, consumer well being, economic, and national interest. Note that each category has one or more objectives. For example, the economic category comprises objectives such as maximize economic benefits to local community, maximize utility company profits, and improve balance of payments. Similarly, the consumer well-being category comprises two objectives: provide necessary power and minimize consumer power costs. There are many interest groups (stakeholders) for this problem. The power company is one of the interest groups. Objectives of interest to the company are maximizing company profits, satisfying consumer preferences for energy, minimizing the detrimental environmental impact of its facilities, and maximizing the net benefits of its facilities on the local communities. Hence, it is important to assess a utility function from the company that is a function of these four objectives. Eliciting the utility functions involves assessing the trade-off information among the four objectives. More detailed discussion of this application is available in the literature cited in the reading list at the end of this article.

Additive or multiplicative utility or value functions have been employed to handle multiple criteria in many applications in the field of energy: to integrate socioeconomic, environmental, and cost criteria in siting a power plant and routing its transmission lines; for the weighting of environmental and economic benefits in demand-side management programs; for the amalgamation of energy, controllability, flexibility, viability, and other related attributes of new resources in bidding procedures; and combining measures of air, water, land, visual, noise, and solid waste effects into an index of environmental impact for use in resource bidding and screening. Other energy-related applications of utility or value theory include air pollution control problems, transporting hazardous wastes, refueling a nuclear power plant, climate change analysis, the study of UK energy policy, selection of a portfolio for a solar energy project, selection of a technology for the disposition of surplus weapons-grade plutonium, performing "as low as reasonably practicable"

(ALARP) assessments in the nuclear industry, identifying appropriate technological alternatives to implement treatment of industrial solid residuals, and risk analysis in nuclear emergency management.

4.1.2 The Analytic Hierarchy Process

The analytic hierarchy process (AHP) is one of the most popular and widely employed multicriteria methods. In this technique, the processes of rating alternatives and aggregating to find the most relevant alternatives are integrated. The technique is employed for ranking a set of alternatives or for the selection of the best in a set of alternatives. The ranking/selection is done with respect to an overall goal, which is broken down into a set of criteria.

The application of the methodology consists of establishing the importance weights to be associated to the criteria in defining the overall goal. This is done by comparing the criteria pairwise. Let us consider two criteria, C_j and C_k. The DM is asked to express a graded comparative judgment about the pair in terms of the relative importance of C_j over C_k with respect to the goal. The comparative judgement is captured on a semantic scale (equally important/moderately more important/strongly important and so on) and is converted into a numerical integer value a_{jk}. The relative importance of C_k over C_j is defined as its reciprocal, i.e., $a_{kj} = 1/a_{jk}$. A reciprocal pairwise comparison matrix A is then formed using a_{jk}, for all j and k. Note that $a_{jj} = 1$. It has been generally agreed that the weights of criteria can be estimated by finding the principal eigenvector w of the matrix A:

$$AW = \lambda_{\max} w. \tag{8}$$

When the vector w is normalized, it becomes the vector of priorities of the criteria with respect to the goal; $\lambda_{\max}$ is the largest eigenvalue of the matrix A and the corresponding eigenvector w contains only positive entries. The methodology also incorporates established procedures for checking the consistency of the judgments provided by the decision maker.

Using similar procedures, the weights of alternatives with respect to each criterion are computed. Then, the overall weights of alternatives are computed using the weighted summation

$$\begin{pmatrix} \text{Overall weight} \\ \text{of alternative } i \end{pmatrix} = \sum_j \begin{pmatrix} \text{Weight of alternative } i \text{ with} \\ \text{respect to } C_j \times \text{weight of } C_j \\ \text{with respect to the goal} \end{pmatrix} \tag{9}$$

The popularity of AHP stems from its simplicity, flexibility, intuitive appeal, and ability to mix quantitative and qualitative criteria in the same decision framework. Despite its popularity, several shortcomings of AHP have been reported in the literature, which have limited its applicability. However, several modifications have been suggested to the original AHP, such as the multiplicative AHP (MAHP), to overcome these limitations. Some of the prominent shortcomings of AHP include the following limitations:

- Rank reversal: The ranking of alternatives determined by the original AHP may be altered by the addition of another alternative for consideration. For example, when AHP is used for a technology selection problem, it is possible that the rankings of the technologies get reversed when a new technology is added to the list of technologies. One way to overcome this problem is to include all possible technologies and criteria at the beginning of the AHP exercise, and not to add or remove technologies while or after completing the exercise. Some variants, such as MAHP, do not suffer from this type of rank reversal.
- Number of comparisons: AHP uses redundant judgments for checking consistency, and this can exponentially increase the number of judgments to be elicited from DMs. For example, to compare eight alternatives on the basis of one criterion, 28 judgments are needed. If there are n criteria, then the total number of judgments for comparing alternatives on the basis of all these criteria will be $28n$. This is often a tiresome exercise for the decision maker. Some methods have been developed to reduce the number of judgments needed.

AHP has been applied to a variety of decision problems in the literature. Several detailed annotated bibliographies of AHP applications are available. In an illustrative application of AHP for the selection of greenhouse gas mitigation options, there are three criteria—namely, cost-effectiveness, feasibility, and other national benefits—to rank five demand-side management options (variable speed drives, good housekeeping, energy-efficient motors, compact fluorescent lamps, and cogeneration). Some other energy-related applications of AHP include solar energy utilization, energy resource allocation, integrated resource planning, climate change negotiations, environmental impact assessment, energy choices for the United States and for members of the Organization of Petroleum Exporting Countries (OPEC), choice of an appropriate of energy mix for

the United States, and, options for high-level nuclear waste management. In some of the applications, AHP has been combined with other MCA methods, such as goal programming and compromise programming. In an application of AHP for global negotiations on climate change, the method has been used to compare the preferences of nations of the world to decide about the extent of reduction of greenhouse gases. In another application in utility planning, AHP has been applied to evaluate the best bids from private parties using criteria such as flexibility in starting date, fuel diversity, reputation of the bidder, and environmental impacts during generation and transmission.

4.2 MODM Methods

As discussed earlier, most of the MODM problems start with the following mathematical programming problem:

$$\begin{aligned} \text{maximize/minimize } Z_1 &= C_1X \\ \text{maximize/minimize } Z_2 &= C_2X \\ \text{maximize/minimize } Z_n &= C_nX \\ &\vdots \\ \text{subject to } AX &= b, \quad X \geq 0 \end{aligned} \tag{10}$$

where the values Z_i, $i = 1, 2, \ldots, n$, represent the n objective functions.

4.2.1 Goal programming

Goal programming (GP), likely the oldest school of MCDM approaches, was developed in the 1950s as an extension of linear programming. In its simplest form, the method of GP assigns so-called aspiration levels (also called targets or goals) for the achievement of different objective functions, and minimizes the deviations of actual achievement from the aspiration levels. It is important to stress here that that the term "goal" has the connotation of target when used in goal programming, whereas in the general MCDM context, the term "goal" is taken to represent a generalization of all criteria.

Consider Problem (10). Suppose T_i is the aspiration level for the objective function Z_i. This means that the decision maker expects to achieve around T_i for the objective function Z_i as given by the approximate equation, $Z_i \approx T_i$, which can be further written as $Z_i + d_i^- - d_I^+ = T_i$, where d_i^- and d_I^+ are deviational variables measuring deviations of actual achievement below and above the aspiration level. They are usually called underachievement and overachievement deviational variables, respectively. Because it is expected that the actual achievement will be as close to the aspiration levels as possible, the deviational variables are minimized in goal programming. Weights, reflecting the relative importance of the objective functions, can also be associated with the deviational variables during the minimization. Thus, one simple form of the goal programming objective function is given by the following relationship:

$$\min \sum_{i=1}^{n} \left(w_{in} d_i^- + w_{ip} d_i^+ \right), \tag{11}$$

where the values w_i are the weight given to minimizing the selected deviational variable from the aspiration level T_i. This objective has to be minimized subject to the original constraints of Problem (1).

Here are two major approaches in goal programming: minimizing the weighted function of goals (discussed previously) and preemptive goal programming that avoids the weighted summation in the goal programming objective function. In preemptive GP, the weights used in the objective function are preemptive weights as defined in Section 3.4. Other mathematical forms, such as minimizing the maximum deviation, fractional goal programming, and nonlinear GP, are also available in the literature. GP has been criticized by several authors. For example, it can, in some circumstances, choose the dominated solution. The criticisms can be overcome by careful applications of the method. In one application, GP has been used to provide energy resource allocation for the city of Madras in India. The minimum energy requirements for cooking, water pumping, lighting, and using electrical appliances form the main constraints of the model. The objective functions pertain mainly to the needs for minimizing cost, maximizing efficiency, minimizing consumption of petroleum products, maximizing employment generation, maximizing the use of locally available resources, minimizing emissions of oxides of carbon, sulfur and nitrogen, and maximizing convenience and safety. The trade-off information about the relative importance of these objectives has been obtained using the AHP.

4.2.2 Compromise Programming

The multiple objective functions Z_i in Problem (10) are often conflicting. Compromise among the multiple conflicting objective functions is defined as deviation or "distance" from the ideal solution. The compromise programming approach minimizes the

distance from the ideal solution. The distance is defined as

$$d_p = \left\{ \sum_{i=1}^{n} w_i^p \left[\frac{Z_i(x^*) - Z_i(x)}{Z_i(x^*)} \right]^p \right\}^{1/p}, \quad (12)$$

where x is the vector of decision variables in the optimization exercise, the values $Z_i(x)$, $i = 1, 2, \ldots, n$, are the n objective functions, $Z_i(x^*)$ represents the ideal solution [when the objective function $Z_i(x)$ is optimized separately without considering the other objective functions], w_i represents the weights to be provided to the objective function $Z_i(x)$, and the parameter p can range from 1 to ∞.

Note that the deviations of all of the objective functions, $Z_i(x^*) - Z_i(x)$, are divided by the respective ideal values. This is done because the objectives are measured in different units and are not comparable in absolute terms. In the literature, p has been interpreted as a degree of noncompensation among the objectives. The case $p = 1$ refers to when the objectives are perfectly compensatory, i.e., an increase in one can compensate for a decrease in the value of the other. The case of $p = \infty$, on the other hand, corresponds to absolute noncompensation, i.e., the objectives are of completely different natures, and improvement in one does not compensate for worsening of the others. Thus, when compromise programming is used, one will have to minimize d_p subject to the original constraints, $AX = b$, $X \geq 0$.

Compromise programming has received some energy applications, especially related to utility planning and environmental economics. In an application, the method has been used to aid the utility planners to select the best bids for generation capacity addition. Bids from independent power producers have been evaluated using multiple objectives—minimizing cost, minimizing emissions, and maximizing qualitative measures such as flexibility, fuel diversity, reputation, etc., subject to constraints representing demands, capacity limits for individual generator and transmission projects, budget constraints, hydro energy availability, fuel production and transportation capacity limits, and minimum utilization of generation capacity. Note that the first two objectives are quantifiable whereas the remaining objectives are not directly quantifiable. Although the compromise programming model has used the first two objectives directly, other qualitative objectives have been handled using AHP.

4.3 Other Methods

The field of MCDM features many more techniques for solving MCA problems. Only a few were described in the previous section. Other techniques include ELECTRE, PROMETHEE, an aspiration-level interactive method, TOPSIS, the reference point approach, multiobjective linear programming, and data envelopment analysis. A technique named preference ratios through intervals in multiattribute evaluation (PRIME) is a relatively recent method that deals with incomplete information about the preferences of the decision makers. An approach involving measuring attractiveness by a categorical-based evaluation technique (MACBETH) was developed in Europe in the 1990s. It is an interactive approach for cardinal measurement of judgments about the degrees of attractiveness in decision processes. Other methods based on an outranking approach (the approach initially used in the ELECTRE methods) include the QUALIFLEX method, the ORESTE method, and the TACTIC method. There are more methods, such as the utilitiés additives UTA) method, the ZAPROS method, and the NIMBUS method. Most of these methods are described in the literature mentioned in the list of suggestions for further reading.

SEE ALSO THE FOLLOWING ARTICLES

Bottom-Up Energy Modeling • Complex Systems and Energy • Depletion and Valuation of Energy Resources • Input–Output Analysis • Life Cycle Assessment and Energy Systems • Modeling Energy Markets and Climate Change Policy • Modeling Energy Supply and Demand: A Comparison of Approaches • Net Energy Analysis: Concepts and Methods

Further Reading

Dyer, J. S., Fishburn, P. C., Steuer, R. E., Wallenius, J., and Zionts, S. (1992). Multiple criteria decision making, multiattribute utility theory: The next ten years. *Manage. Sci.* **38**(5), 645–654.

Edwards, W. (1977). How to use multiattribute utility measurement for social decision-making? *IEEE Trans. Syst. Man Cybernet* **7/5**, 326–340.

Gal, T., Stewart, T. J., and Hanne, T. (eds.). (1999). "Multicriteria Decision Making: Advances in MCDM Models, Algorithms, Theory, and Applications." Kluwer Academic Publ., Boston.

Hobbs, B. F., and Meier, P. (2000). "Energy Decisions and the Environment: A Guide to the use of Multicriteria Methods." Kluwer Academic Publ., Boston.

Keeney, R. L., and Raiffa, H. (1976). "Decisions with Multiple Objectives: Preferences and Value Tradeoffs." Wiley, New York. [Another edition (1993) available Cambridge University Press.].

Roy, B. (1996). "Multicriteria Methodology for Decision Aiding." Kluwer Academic Publ., Dordrecht, The Netherlands.
Saaty, T. L. (1980). "The Analytic Hierarchy Process: Planning, Priority Setting and Resource Allocation." McGraw-Hill, New York.
Schniederjans, M. J. (1995). "Goal Programming: Methodology and Applications." Kluwer Academic Publ., Boston.
Shi, Y., and Zeleny, M. (eds.). (2000). "New Frontiers of Decision Making for the Information Technology Era." World Scientific Publ. Co., Singapore.
Zeleny, M. (1982). "Multiple Criteria Decision Making (McGraw-Hill Series in Quantitative Methods for Management)." McGraw-Hill, New York.

National Energy Modeling Systems

ANDY S. KYDES
U.S. Department of Energy
Washington, D.C., United States

AMIT KANUDIA
McGill University and KanORS Consulting Inc.
Montreal, Quebec, Canada

RICHARD LOULOU
McGill University and HALOA Inc.
Montreal, Quebec, Canada

1. **Introduction**
2. **Overview of NEMS**
3. **Analysis of a 20% Non-hydroelectric Portfolio Standard**
4. **Overview of MARKAL**
5. **Example Application of the Canadian MARKAL model**

Glossary

coefficient of performance The ratio of energy transferred out of a system (e.g., an air conditioner or heat pump transfers heat from inside a house to outside the house) to the energy input to the system.

competitive market Pertains to the degree that energy prices and supply are influenced or set by any particular group in the market; if energy prices and supplies are set by many market participants (no monopoly or monopsony power) and not by regulatory bodies, the market is said to be competitive.

corporate average fuel efficiency (CAFE) For new cars, a standard by which the average of all cars sold in the United States by a manufacturer must achieve 27.5 miles per gallon.

dynamic linear programming model A multiperiod linear programming model that is composed of multiple smaller linear program models (one for each time period) that are linked with intertemporal constraints; a dynamic linear programming model provides the optimal value of the objective over all time periods—decision making with perfect foresight.

efficiency The ratio of the energy output by an energy system to the energy content of the fuel input to the energy system.

exajoule (EJ) A unit of energy equal to 10^{18} joules; 1.0556 EJ is approximately 1 quadrillion Btu.

Gauss–Seidel convergence algorithm An iterative method for solving large linear or nonlinear models; in the National Energy Modeling System (NEMS), the algorithm finds where supply equals demand and the delivered prices for which this happens.

greenhouse gases Gases that have the property of absorbing or trapping energy in the earth's atmosphere; such gases include carbon dioxide, methane, and a number of manmade chemicals used in refrigeration and manufacturing, with the latter group being collectively known as "high global warming potential" gases.

intermittent technologies Pertain to technologies whose energy production is intermittent (e.g., wind, photovoltaics); for intermittent technologies, energy production depends on factors other than the availability of the technology (e.g., availability of wind and sunlight).

learning factor Pertains to the rate by which manufacturing costs decline for every doubling of capacity.

linear programming model A mathematical model with a linear objective function and linear constraints.

load shape Pertains to the time profile of energy consumption.

logit function A function form that often includes exponential functions in the denominator and numerator; it is typically used to determine market shares.

market diffusion Pertains to the adoption of a new technology or features into the market.

merit order dispatch Pertains to the order in which power plants are brought online and used to satisfy electricity demand; a merit order dispatch requires that plants be brought online in order of increasing operating plus fuel costs.

overnight capital cost Usually pertains to the costs quoted for electric power generation plants; overnight costs include all manufacturing costs, including owners' costs

 © 2004 Elsevier Inc. All rights reserved.

and project contingency but not interest during construction.

reference energy system A schematic representation of energy flows in an energy system, from energy production, to intermediate conversions and transformations, to delivery at end use.

renewable portfolio standard Pertains to a potential policy requiring that a minimum percentage of all generation be provided from renewable energy sources such as wood, wind, other biomass, and geothermal.

Residential Energy Consumption Survey (RECS) A survey that was developed and is published periodically by the Energy Information Administration.

technological change Pertains to changes in the development, adoption, and characteristics of technologies, including innovations, efficiency improvements, cost reductions, and the addition of new qualitative features such as variable speed fans.

technological optimism Pertains to cost and performance (mis)estimates for technologies that have not been commercialized. Prior to commercialization, manufacturing costs tend to be underestimated and performance tends to be overestimated.

unit energy consumption (UEC) A surrogate term used to measure efficiency in a process that combines structures and energy conversion systems.

The development and introduction of new technologies to the market, relative costs and performance of technologies, physical lifetimes of installed equipment (which influence the natural turnover rate), and consumer preferences are key factors that determine the rate of market diffusion of technologies and how quickly energy use patterns, energy efficiency, and energy-related environmental emissions in a society can change. Estimates of the costs and benefits of new energy policies, such as those that may be implemented to constrain carbon dioxide or other greenhouse gas emissions, will depend on estimates of how rapidly advanced technologies are accepted in the market and how quickly existing end use equipment reaches the end of its useful life (i.e., wears out). Consequently, the representation of technological change, market diffusion, and other economic flexibility embodied in any energy–economy modeling system will be major determinants of the projected energy–economy adjustments or responses to policy/price changes. This article provides an overview of the National Energy Modeling System (NEMS), which was developed and used by the Energy Information Administration (EIA) to address energy–economic–environmental policy questions for the U.S. government—typically Congress and the executive branch of government. A brief example of a recent NEMS application in assessing the potential impact of a 20%, nonhydroelectric, renewable portfolio standard (RPS) by 2020 on U.S. energy markets is also provided. The second national energy modeling system described in this article is the Canadian MARKAL model, one of the numerous variants of the market allocation (MARKAL) model, a dynamic linear programming formulation of the Canadian energy markets.

1. INTRODUCTION

The purpose of this article is to provide an overview of arguably two of the most prominent and influential national energy modeling systems used over the past decade or so: the National Energy Modeling System (NEMS) and the MARKAL model. NEMS was developed by the Energy Information Administration (EIA) during the early 1990s to analyze U.S. energy markets and is important because it is the primary analytical and projection tool used by the U.S. Department of Energy to analyze the energy, economic, and environmental impacts of proposed energy policies for the U.S. government. MARKAL is important because it has been used by more than 50 countries and 70 regional governments to analyze the impact of energy strategies and is supported by an international group of experts and the International Energy Agency (IEA).

The persistence of and refinements to these models over the past decade or so reflect both the increasing complexities of proposed energy–environmental policies and the recognition that energy policy has important impacts on an economy, even though energy's share of gross domestic product (GDP) may be less than 10%.

The next section of this article provides an overview of NEMS with more detailed descriptions of the residential and electricity generation markets to illustrate how two representative energy sectors are modeled in the system. This is followed by an example of the use of NEMS in policy analysis—a renewable portfolio study. The subsequent section provides an overview of MARKAL. The final section provides an example of policy analysis using the MARKAL model.

2. OVERVIEW OF NEMS

This section provides a brief overview of NEMS. A more extensive treatment of the NEMS model is

provided on the EIA's Web site (www.eia.doe.gov). NEMS is a large, technology-rich, regional, computer-based energy–economy model of U.S. energy markets for the midterm period through 2025.

In 1990, the secretary of the U.S. Department of Energy (DOE) directed that NEMS be developed. NEMS was designed and developed to support energy policy analysis and strategic planning, based on recommendations from the National Research Council (NRC) of the National Academy of Sciences.

Key features implemented in NEMS included (1) regional outputs of energy, economic, and environmental activity of the U.S. economy; (2) use of a modular modeling structure to facilitate and enable the model builders to work with particular aspects of the model independently; (3) integration of engineering and economic approaches to represent actual producer and consumer behavior; (4) use of a projection period spanning 20 to 25 years; and (5) incorporation of the broader energy analysis community and outside peer groups in the design and update of NEMS.

NEMS was completed at the end of 1993 and was used for the first time to develop the *Annual Energy Outlook 1994, with Projections to 2010*. More recently, NEMS has been extended to 2025 and further revised to address electricity restructuring and carbon mitigation issues and numerous other multipollutant policy studies.

NEMS is used to develop baseline projections that are published annually in the *Annual Energy Outlook* (*AEO*). NEMS is also used by analysts to prepare special studies that are requested by the U.S. Congress, the DOE Office of Policy, other DOE offices, and other government agencies. In accordance with the requirement that the EIA remain policy neutral, the *AEO* projections assume that all existing legislation, regulations, and policies remain unchanged. Furthermore, these projections depend on additional uncertain assumptions, including the estimated size of the economically recoverable resource base of fossil fuels, changes in world energy supply and demand, the rate at which new energy technologies are developed, and the rate and extent of their adoption and penetration. Consequently, the *AEO* projections are not meant to be predictions about the future. This article describes an updated version of NEMS used for the *AEO 2000*.

2.1 Purpose of NEMS

The primary purpose of NEMS is to analyze the energy-related consequences for the United States of alternative energy policies or pertinent economic or energy market influences. The policy questions of interest have determined the level of detail required within the structure of NEMS. For example, environmental issues relating to energy production and consumption have taken on new importance with the implementation of the Clean Air Act Amendments (CAAA) of 1990 and the proposed Kyoto Protocol on greenhouse gases (GHGs) in 1997. Accordingly, NEMS is designed to measure seven emissions (sulfur oxides [SO_x], nitrogen oxides [NO_x], carbon, carbon monoxide [CO], carbon dioxide [CO_2], volatile organic compounds, and mercury) released in the use of energy products to generate electricity and, in the case of carbon (or CO_2), to constrain national emissions using a pricing mechanism. The technology representation in NEMS is particularly helpful in the analysis of national carbon mitigation policies and utility sector SO_x, NO_x, and mercury mitigation policies due to its explicit representation of vintage (time-dependent) energy equipment and structures (e.g., building shells) and the careful tracking of vintage capital stock turnover rates. For similar reasons, NEMS contains sufficient detail in the transportation sector to project the use of alternative or reformulated fuels such as compressed natural gas, ethanol, and methanol. In addition to environmental concerns, NEMS is designed to account for existing and emerging government regulations (e.g., electricity restructuring), the potential for the development and use of new energy-related technologies, the increased use of renewable sources of energy (especially intermittent technologies), and the potential for demand-side management, conservation, and increases in the efficiency of energy use. These topics reflect the expected scope of current and future government policy.

The NEMS representation of energy markets focuses on four important interrelationships: (1) interactions among the energy supply, conversion, and consumption sectors; (2) interactions between the domestic energy system and the general domestic economy; (3) interactions between the U.S. energy system and world energy markets; and (4) interactions between current production and consumption decisions and expectations about the future.

2.1.1 Domestic Energy System/Economy Interactions

The general level of economic activity in sectoral and regional detail has traditionally been used as an explanatory variable or "driver" for projections of energy consumption and prices. In reality, energy prices and other energy system activities themselves

influence the level of economic activity. NEMS is designed to capture this "feedback" between the domestic economy and the energy system. Thus, energy price changes are capable of affecting macroeconomic variables such as GDP.

2.1.2 Domestic and World Oil Market Interactions

The world oil price (WOP) is a key variable in domestic energy supply and demand decision making. As a result, WOP assumptions have been a key starting point in the development of energy system projections. In fact, the U.S. energy system itself exerts a significant influence on world oil markets, and this in turn influences the WOP—another example of a feedback effect. World energy market supply and demand are first specified outside of NEMS. Given this, NEMS models the interactions between the U.S. and world oil markets, with changes in U.S. oil markets affecting world supply and demand. As a result, domestic energy system projections and the WOP are made internally consistent.

2.1.3 Economic Decision Making over Time

The production and consumption of energy products today are influenced by past decisions to develop energy resources and acquire energy using capital. Similarly, the production and consumption of energy at a future time are influenced by decisions made today and in the past. Current investment decisions depend on expectations about future market circumstances. For example, the propensity to invest now to develop alternative energy sources is greater when it is expected that future energy prices will be higher. NEMS allows the application of different kinds of foresight assumptions to be applied differentially to its individual submodules. This feature allows the consequences of different kinds of planning horizons to be incorporated into NEMS projections.

2.2 System Design of NEMS

Like its predecessor models (e.g., the Intermediate Future Forecasting System [IFFS]), NEMS incorporates a market-based approach to energy analysis. NEMS balances the supply of and demand for energy for each fuel and consuming sector, taking into account the economic competition between energy sources.

NEMS is partitioned into a modular system, which is solved by applying the Gauss–Seidel convergence method with successive overrelaxation. The modules of NEMS represent each of the fuel supply markets, conversion sectors, and end use consumption sectors and also include interactive macroeconomic and international modules. The primary flows between these modules are the delivered prices of energy and the energy quantities consumed by product, region, and sector but also include other information such as economic activity and technology characteristics. The delivered prices of fuels incorporate all of the activities necessary to produce, import, and transport fuels to the end users. Figure 1 provides a system overview of NEMS.

The integrating methodology controls the independent execution of the component modules. The modules are executed from the integrating module. To facilitate modularity, the components do not pass information to each other directly but instead communicate through a central data file. This modular design provides the capability of executing modules individually or to alternative modules, thereby allowing decentralized development of the system and independent substitute analysis and testing of individual modules. Furthermore, this modularity allows the flexibility of using the methodology and level of detail that is most appropriate to represent each energy sector.

A solution is achieved by equilibrating on the delivered prices and quantities of energy demanded, thereby ensuring an economic equilibrium of supply and demand in the consuming sectors. Each fuel supply, conversion, or end use demand module is called in sequence by the integrating module and solved, assuming that all other variables in the other energy markets are fixed. For example, when solving for the quantities of fuels demanded in the residential sector for an input set of energy product prices, all other sectors of the economy are held fixed. The modules are called iteratively until successive end use prices and quantities remain constant within a specified tolerance. This equilibration is achieved annually through the midterm period to 2025. Table I provides a summary of NEMS products and regional details. NEMS reflects market economics, industry structure, and energy policies and regulations that influence market behavior.

NEMS consists of four supply modules (oil and gas, natural gas transmission and distribution, coal, and renewable fuels), two conversion modules (electricity and petroleum refineries), four demand modules (residential, commercial, transportation, and industrial), one module to simulate energy–economy interactions (macroeconomic activity), one module to simulate world energy–domestic energy interactions (international energy activity), and one module to provide the mechanism for achieving a

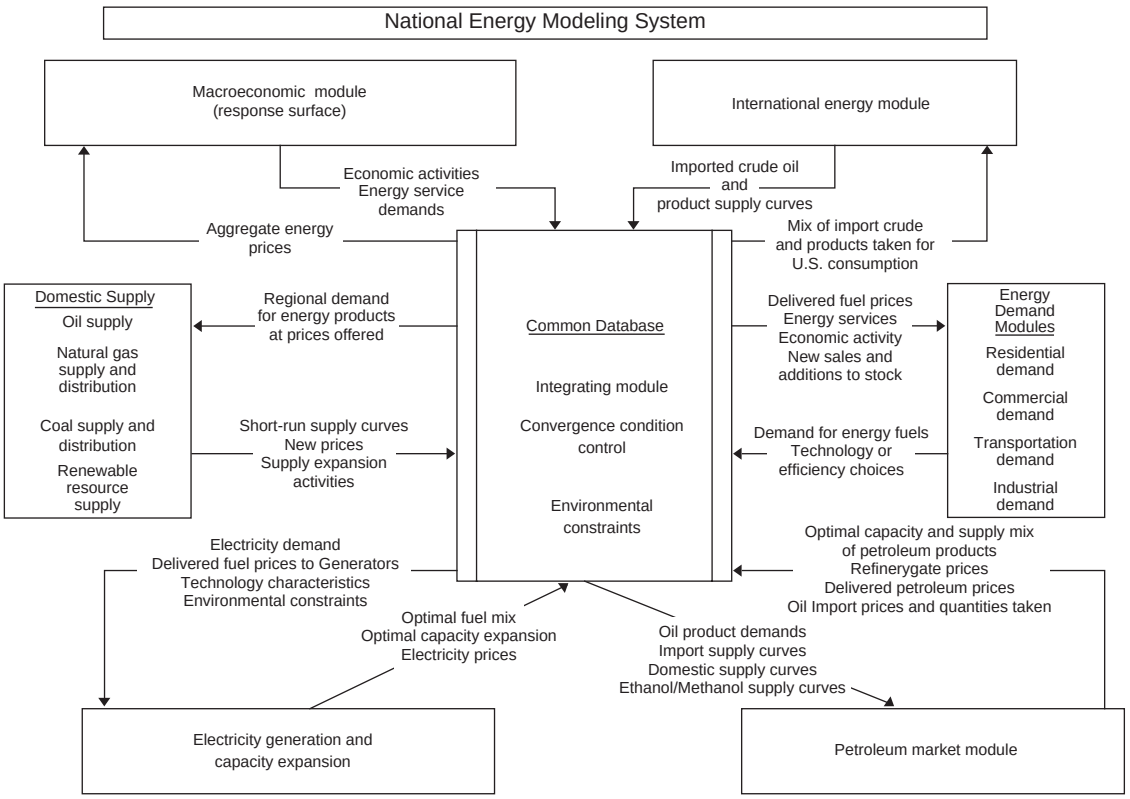


FIGURE 1 NEMS systems view.

general market equilibrium among all of the modules (the integrating module).

2.3 Foresight

All of the NEMS supply modules require assumptions of future energy prices and demands for energy so as to make capacity expansion decisions. Recognizing that there is valid evidence that decision making varies by sector, one implementation of foresight in NEMS allows each sector to use those foresight and behavioral assumptions deemed most appropriate for that sector. Two alternative decentralized options that have been implemented for foresight are myopic (i.e., assuming, within any forecast year, that the current prices will remain constant into the future for capacity expansion decisions) and extrapolative (i.e., assuming, within any forecast year, that expected prices, demands, and supplies are a function of historical trends and other critical assumptions regarding behavior). In the simplest case, the extrapolation may be a growth rate applied to the current forecast year.

Perfect foresight has been implemented using a recursive process where the previous run's prices are used as the expectations.

2.4 Emissions

Recognizing the importance of environmental issues associated with the use of energy, an environmental accounting capability has been incorporated within NEMS. Seven emissions are accounted for in the electricity generation sector: sulfur dioxide (SO_2), NO_x, CO, CO_2, carbon, volatile organic compounds, and (most recently) mercury emissions from power plants. A subset of these emissions is computed for energy production activities and fuel combustion. In addition, NEMS is designed to represent all current environmental regulations (e.g., CAAA of 1990) as well as other mandated costs for controlling toxic substances.

NEMS also incorporates the capability of constraining systemwide CO_2 emissions as well as SO_2 emissions, NO_x, and mercury emissions in the utility market—important features for policy analysis. One option for achieving a prescribed level of systemwide CO_2 emissions is accomplished by raising the cost of producing emissions until the total system emissions are reduced to the level of the constraint. Another approach used in the utility market for SO_2 is to incorporate the constraint directly into a linear program. This is

TABLE I

Summary of NEMS Modeling Details

Energy system activity	Categories	Regions
Primary supply		
Oil	Conventional	Six lower 48 onshore
	Enhanced	Three lower 48 offshore
	Heavy oil	Three Alaska
	Shale oil	
Gas	Conventional nonassociated	Six lower 48 onshore
	Conventional associated	Three lower 48 offshore
	Coalbed methane	Three Alaska
	Three other types	
Natural gas transmission and distribution	Residential, commercial, industrial, utility firm versus interruptible Peak versus offpeak	12: nine census divisions, with further subdivisions for key transportation issues
Coal	Four sulfur categories	16 supply regions
	Four thermal categories	
	Deep and surface mining types	
Renewables	Hydropower, wind, geothermal, solar thermal, photovoltaic, municipal solid waste, biomass	Natural Environment Research Council/ Refining/Census
Conversion		
Electricity (including uranium for nuclear)	Utilities, independent power cogeneration	13 supply regions Nine census demand regions
Refining	Five crude categories 19 products	Three petroleum allocation for defense districts
Energy demand		
Residential	Eight end use services	Nine census divisions
	Three housing types	
Commercial	Eight end use service	Nine census divisions
	11 building types	
Transportation	Six vehicle size categories	Nine census divisions
	10 vehicle age categories	
Industrial	35 industries	Four census regions shared to nine census divisions
	Nine primary industries	

the method used by the electric market module to constrain SO_2 emissions as mandated by the CAAA of 1990.

2.5 Policy Analysis Capability with NEMS

The following list illustrates the main analytical capabilities available with NEMS:

• *Federal tax policies/fees:* impacts of energy tax policies on the economy and energy system; impacts of Btu or carbon taxes; revenue recycling options (how portions of taxes are recycled through the economy: all to consumers, all to deficit reduction, or some other combination); federal reaction to rising unemployment (accommodating or not)

• *Emissions caps/restrictions:* carbon emissions cap (cap and trade impacts) from all sectors; national sulfur cap from electricity generation; NO_x cap from electricity generation; mercury emissions cap from electricity generation; other emissions caps

• *Effects of existing and proposed government laws and regulations related to energy production and use:* impacts of electricity restructuring on prices and treatment of stranded costs; impacts of renewable portfolio standards (RPSs); impacts of

increased use of renewable energy sources; potential savings from increased efficiency of energy use; changes in emission levels due to environmental policies; effects of appliance efficiency and building shell standards on energy consumption; impacts of fuel use restrictions (e.g., required use of oxygenated and reformulated gasoline, mandated use of alternative-fueled vehicles) on emissions, energy supply, prices, and economic growth; changes in natural gas prices and pipeline and import capacity in response to regulatory initiatives; impacts of automobile CAFE standards

- *Impacts of advanced/alternative technology menus/cost/performance:* potential impacts of new and advanced energy production, conversion, and consumption technologies; impact on technology learning of accelerated demonstration plants for utilities (government-funded/mandated capacity additions to accelerate learning and reduce capital and/or operating and maintenance [O&M] costs); impacts of new technologies on consumption and production patterns and emissions; impacts on the production of crude oil and natural gas resulting from improvements in exploration and production technologies
- *Oil market assessment:* responses of the energy and economic systems to changes in world oil market conditions as a result of changing levels of foreign production and demand in the developing countries.

2.6 NEMS Residential Sector

The residential demand module forecasts fuel consumption and energy equipment choices in the residential sector for 9 census divisions, 3 housing types (single-family, multifamily, and mobile homes), and 10 end uses of which 8 (space heating, space cooling, water heating, refrigerators, freezers, lighting, clothes dryers, and cooking) have equipment choices associated with them.

2.6.1 Important Interactions with NEMS

NEMS is initialized with information about the state of the U.S. energy system for a recent historical year. With respect to the residential demand module, this includes base year housing stock and retirement rates, appliance stocks and life expectancies, new appliances to be made available in current and future years with their costs and efficiencies, housing shell integrity (efficiency index), unit energy consumption per end use and household, and square footage per household and housing type. In addition, assumptions about population and associated demographics developed by the U.S. Department of Commerce are provided as inputs to the projection.

NEMS provides to the residential modules forecasts of residential energy product prices and new housing starts (Fig. 2). The housing stock submodule begins by adjusting the previous year's housing stock by adding new construction (new housing starts) and

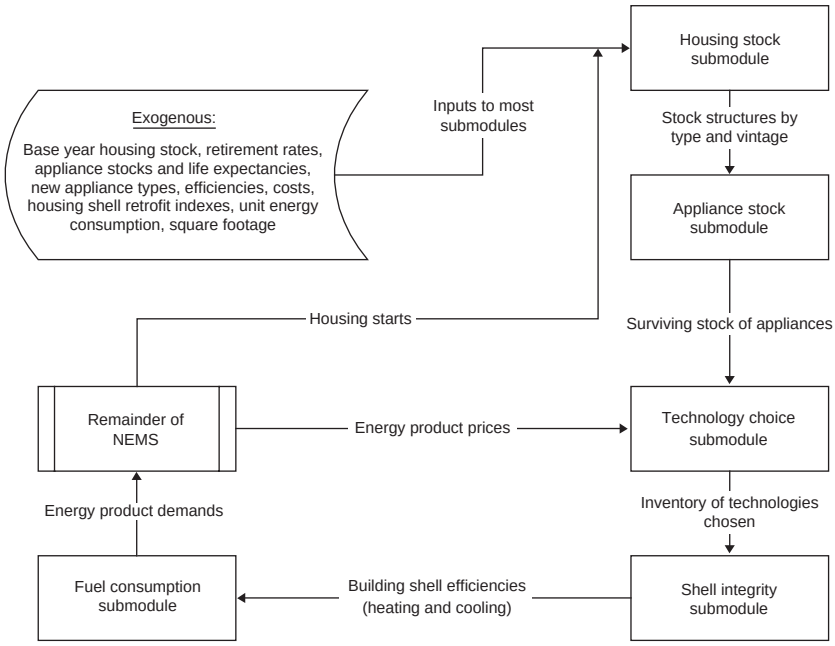


FIGURE 2 Residential demand module.

subtracting housing retirements. The new housing stock is allocated to each of the three housing types based on a combination of historical shares and income distribution and demographic changes. Control is then transferred to the appliance stock module, which (1) retires appliances in homes that were retired, (2) retires appliances that have reached the end of their useful lives, and (3) prepares the menu of technology choices with their associated cost and performance characteristics. Control is then transferred to the technology choice submodule. This submodule, using equipment cost and performance data with estimated consumer preference functions, selects appliances to replace retired appliances and appliances needed for new housing. Control is then passed to the shell integrity submodule, which adjusts the heating and cooling requirements of each building type based on an autonomous change to shell efficiency (improvements that are independent of price) and price-induced changes to building efficiency.

Once the appliances have been selected for each end use and end use energy intensities have been adjusted for changes to shell integrity and lifestyle changes such as increasing floor space per household, the energy consumption module computes energy consumption by fuel and census division, and this is passed back to NEMS for response. Figure 3 provides an overview of the residential demand module with an emphasis on the appliance and fuel choice components.

2.6.2 Equipment Characterizations and Choice Methodology

2.6.2.1 Technology Characterizations Residential sector energy technologies are characterized by (1) retail cost of equipment plus installation, (2) efficiency of the equipment, (3) dates available for purchase (efficiency standards may limit availability during future years), and (4) technological evolution, that is, either projected efficiency increases or projected cost declines. Cost reductions can be discrete

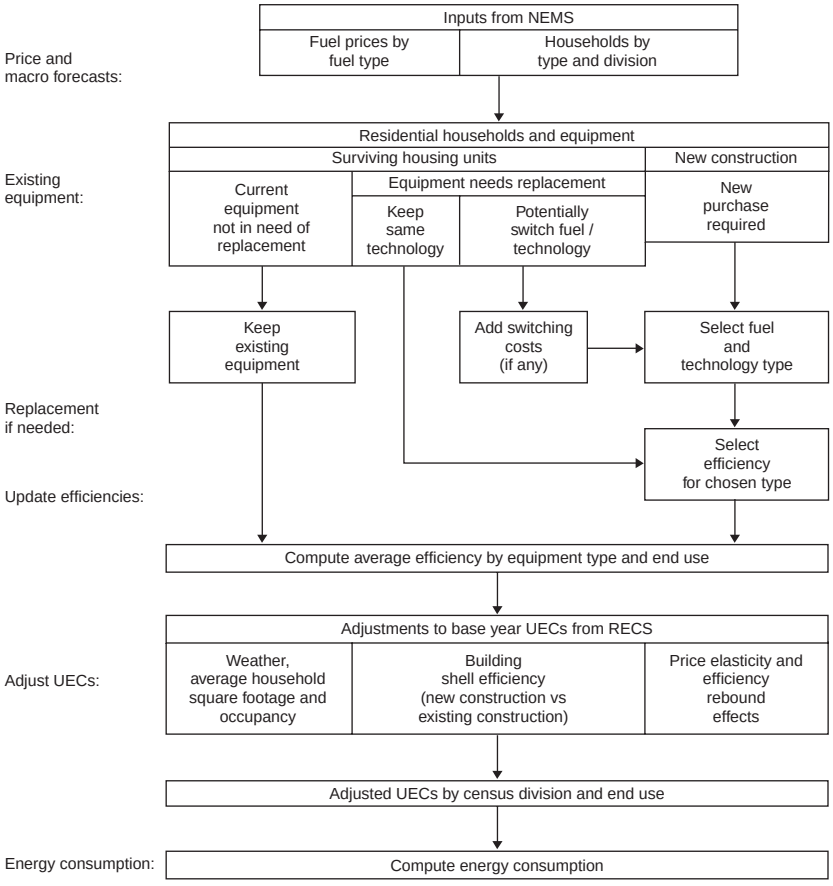


FIGURE 3 NEMS residential model overview. UECs, units energy consumption; RECS, Residential Energy Consumption Survey.

(e.g., a new characterization of a technology in a given year) or continuous (e.g., annually declining costs modeled for new heating technologies).

2.6.2.2 Market Penetration/Consumer Choice Algorithm The choice of equipment in the residential demand sector is made using a two-step selection process that incorporates the following important residential market attributes: (1) the inclination not to replace equipment until it fails (i.e., "if it isn't broken, don't fix it"), (2) the predisposition to stay with the same technology when equipment needs to be replaced, and (3) the incorporation of switching costs ("real and perceived") when changing to a different fuel is contemplated.

Technology choice/penetration is a two-step process in the NEMS residential energy model. Equipment choices are segmented by new and replacement equipment. For equipment decisions in new homes or for replacement decisions that consider fuel switching as an option, the first step is to decide on what fuel and general equipment type will be used for each end use (e.g., natural gas furnace for space heating). The second step is to select the efficiency of the equipment (e.g., the furnace). For replacement decisions that do not consider fuel switching or technology switching, the only decision made is to select the efficiency of the technology. NEMS uses a logit function to represent the decisions made in each of these two steps.

In the residential sector, technologies compete on the basis of relative importance of installed cost and operating cost for each technology. Equipment choices are segmented by new and replacement equipment for each of nine census divisions and each of three building types.

2.6.2.3 Fuel–Technology Class Choice The following calculation is performed as the first step in determining the fuel–technology class combination in new homes and when considering fuel switching when replacing worn equipment. When fuel switching is considered for replacement decisions, a matrix of switching costs from one equipment/fuel type to another is added to the installed capital costs before computing life cycle costs. Life cycle costs are based on the implicit hurdle rate for each end use and technology combination using 7 years as the time horizon over which discounted costs are calculated. These hurdle rates combine not only financial aspects (the cost of money) but also nonfinancial aspects (institutional and physical obstacles and perceived risks). The implicit discount rates range from 15 to more than 100%. The time horizon of 7 years reflects the fact that the average time between moves for homeowners in the United States is less than 7 years, and a longer time horizon for calculating discounted costs seems inappropriate for simulation purposes.

If LC_i is the life cycle cost of the representative technology for each fuel–technology class i, the share allocated to fuel–technology class i is given by the logit function

$$S_i = \frac{\exp(a_i + b_i LC_i)}{\sum \exp(a_j + b_j\, Lc_j)};\ \ j = 1,\ n.$$

2.6.2.4 Efficiency–Technology Decision Each efficiency choice function is calibrated to historical shipment efficiency data; that is, the model replicates efficiency choices for historical years. The efficiency–technology decision is based on a three-parameter logit function of the form

$$S_i = \frac{\exp(a_i + b_i\, FC_i + c_i\, O\&M_i)}{\sum \exp(a_j + b_j\, FC_j + c_j\, O\&M_j)};\ \ j = 1,\ k_i,$$

where FC_i is the first cost of technology i, $O\&M_i$ is the O&M cost of technology i, and k_i is the number of technologies that use fuel i.

2.7 Electricity Market Module

The electricity market module (EMM) represents the capacity planning, generation, transmission and distribution, and pricing of electricity subject to delivered prices for coal, petroleum products, and natural gas; O&M costs for existing and new generation equipment; the costs of renewable fuels for generation; the capital costs of power generation investments; and electricity load shapes and demand. The submodules consist of capacity planning, fuel dispatching, finance and pricing, and load and demand-side management (Fig. 4). In addition, nonutility supply and electricity trade are represented in the fuel dispatching and capacity planning submodules. Nonutility generation from cogenerators and other facilities whose primary business is not electricity generation is represented in the demand and fuel supply modules. All other nonutility generation is represented in EMM. The generation of electricity is accounted for in 13 supply regions, whereas consumption is satisfied in the 9 census divisions.

Operating (dispatch) decisions are provided by the cost-minimizing mix of fuel and variable O&M costs, subject to environmental constraints. Capacity expansion is determined by the least cost mix of all

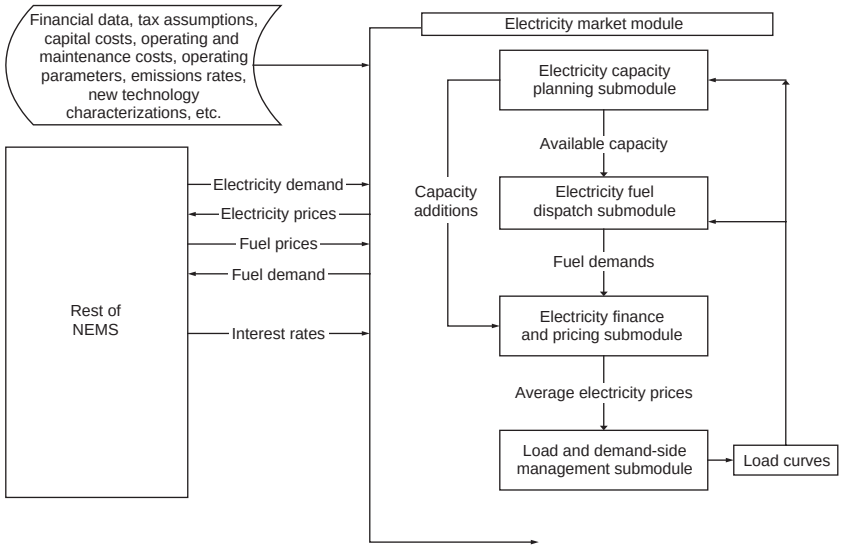


FIGURE 4 Electricity market module overview.

expected costs, including capital, O&M, and fuel, subject to meeting environmental restrictions and expected electricity demand. Construction of generating plants with long lead times is selected with planning horizons up to six periods into the future; the planning horizon can change with respect to the generating technology being considered. Electricity demand is represented by load curves, which vary by region, season, and time of day.

The solution to the submodules of EMM is simultaneous in that, directly or indirectly, the solution for each submodule depends on the solution to every other submodule. A solution sequence through the submodules can be viewed as follows:

- The load and demand-side management (LDSM) submodule processes electricity demand to construct load curves.
- The electricity capacity planning (ECP) submodule projects the construction of new utility and nonutility plants, the level of firm power trades, and the addition of scrubbers for environmental compliance.
- The electricity fuel dispatch (EFD) submodule dispatches the available generating units, both utility and nonutility, allowing surplus capacity in select regions to be dispatched for another region's needs (economy trade).
- The electricity finance and pricing (EFP) submodule calculates total revenue requirements for each operation and computes average and marginal cost-based electricity prices.

2.7.1 Load and Demand-Side Management Submodule

The LDSM submodule generates load curves representing the demand for electricity. The demand for electricity varies over the course of a day. Many different technologies and end uses, each requiring a different level of capacity for different lengths of time, are powered by electricity. For operational and planning analysis, an annual load duration curve, which represents the aggregated hourly demands, is constructed. Because demand varies by geographic area and time of year, the LDSM submodule generates load curves for each region and season.

2.7.2 Electricity Capacity Planning Submodule

The ECP submodule, a dynamic linear programming formulation of capacity expansion decision making, determines how best to meet expected growth in electricity demand given available resources, expected load shapes, expected demands and fuel prices, environmental constraints, and costs for utility and nonutility technologies. When new capacity is required to meet electricity demand, the timing of the demand increase, the expected use of the new capacity, the operating efficiencies, and the construction and operating costs of available technologies determine what technology is chosen.

The timing of the demand increase is important because the construction lead times of technologies differ. The ECP submodule looks up to six periods into the future when identifying new capacity needs. A multiperiod optimization is performed, whereby

capacity choices in each year are made by looking at several years in the future rather than at only a single year. Construction lead times also contribute to uncertainty about investment decisions. Technologies with long lead times are subject to greater financial risk. Compared with plants with shorter lead times, they are more sensitive to market changes in interest and inflation rates and are more vulnerable to uncertain demand projections that determine the need for new capacity. To capture these factors, the discount rate for each technology can be adjusted using risk premiums based on the construction lead time. The risk-adjusted discount rate results in the perception that a technology with a long lead time is less economically attractive than another technology with similar costs but a shorter lead time.

Uncertainty about investment costs for new technologies is captured in the ECP submodule using technological optimism and learning factors. The technological optimism factor reflects the inherent tendency to underestimate costs for new technologies.

The degree of technological optimism depends on the complexity of the engineering design and the stage of development. As development proceeds and more data become available, cost estimates become more accurate and the technological optimism factor declines. Learning factors represent reductions in capital costs due to "learning-by-doing." For new technologies, cost reductions due to learning also account for international experience in building generating capacity.

The decrease in overnight capital costs due to learning depends on the stage of technological development. The costs of a "revolutionary" technology are assumed to decrease much faster than the costs of mature technologies, ranging from 10% for every doubling of capacity for advanced technologies to 1% for every doubling of capacity for mature technologies.

Capital costs for all new electricity generating technologies (fossil, nuclear, and renewable) decrease in response to foreign and domestic experience. Foreign units of new technologies are assumed to contribute to reductions in capital costs for units that are installed in the United States to the extent that (1) the technology characteristics are similar to those used in U.S. markets, (2) the design and construction firms and key personnel compete in the U.S. market, (3) the owning and operating firm competes actively in the United States, and (4) there exists relatively complete information about the status of the associated facilities. If the new foreign units do not satisfy one or more of these requirements, they are given a reduced weight or are not included in the learning effects calculation.

Initially, investment decisions are determined in the ECP submodule using cost and performance characteristics that are represented as single-point estimates corresponding to the average (expected) cost. However, these parameters are also subject to uncertainty and are better represented by distributions. If the distributions of two or more options overlap, the option with the lowest average cost is not likely to capture the entire market. Therefore, ECP uses a market-sharing algorithm to adjust the initial solution and reallocate some of the capacity expansion decisions to technologies that are "competitive" but do not have the lowest average cost.

Fossil-fired steam plant retirements are calculated endogenously within the model. Fossil plants are retired if the market price of electricity is not sufficient to support continued operation. The expected revenues from these plants are compared with the annual going-forward costs, which are mainly fuel and O&M costs. A plant is retired if these costs exceed the revenues and the overall cost of electricity can be reduced by building replacement capacity.

Retirement decisions for nuclear capacity are also determined by the model. Four options for the operating license are considered. A unit (1) can be retired early (10 years prior to the end of the operation license), (2) can be retired when the license expires, or can be operated (3) an additional 10 years or (4) an additional 20 years by renewing the license. At each stage, the assumed aging-related expenditures due to capital additions, increased maintenance, and/or performance declines are compared with the cost of replacement capacity. A unit is retired if the aging costs, which are recovered over 10 years, exceed the cost of building new capacity.

The ECP submodule also determines whether to contract for unplanned firm power imports from Canada and from neighboring electricity supply regions. Imports from Canada are computed using supply curves developed from cost estimates for potential hydroelectric projects in Canada. Imports from neighboring electricity supply regions are computed in ECP based on the cost of the unit in the exporting region plus the additional cost of transmitting the power. Transmission costs are computed as a fraction of revenue. After building new capacity, the submodule passes total available capacity to the electricity fuel dispatch submodule and new capacity expenses to the electricity finance and pricing submodule.

2.7.3 Electricity Fuel Dispatch Submodule

Given available capacity, firm purchased power agreements, fuel prices, and load curves, the EFD submodule minimizes variable costs as it solves for generation facility use and economy power exchanges to satisfy demand during each time period and in each region. The submodule uses merit order dispatching; that is, utility, independent power producer, and small power producer plants are dispatched until demand is met in a sequence based on their operating costs, with least cost plants being operated first. Limits on emissions of SO_2 and NO_x (and mercury and CO_2 if such policies are requested) from generating units and the engineering characteristics of units serve as constraints. Coal-fired capacity can cofire with biomass to lower operating costs and/or emissions. During off-peak periods, the submodule institutes load following, which is the practice of running plants near their minimum operating levels rather than shutting them down and incurring shut-off and start-up costs. In addition, to account for scheduled and unscheduled maintenance, the capacity of each plant is derated (lowered) to the expected availability level. Finally, the operation of utility and nonutility plants for each region is simulated over six seasons to reflect the seasonal variation in electricity demand.

Interregional economy trade is also represented in the EFD submodule by allowing surplus generation in one region to satisfy electricity demand in an importing region, resulting in a cost savings. Economy trade with Canada is determined in a manner similar to that of interregional economy trade. Surplus Canadian energy is allowed to displace energy in an importing region if doing so results in a cost savings. After dispatching, fuel use is reported back to the fuel supply modules, and operating expenses and revenues from trade are reported to the EFP submodule.

2.7.4 Electricity Finance and Pricing Submodule

The costs of building capacity, buying power, and generating electricity are tallied in the EFP submodule, which simulates the cost-of-service method often used by state regulators to determine the price of electricity. Revenue requirements shared over sales by customer class yield the price of electricity for each class. Electricity prices are returned to the demand modules. In addition, the submodule generates detailed financial statements.

The EFP submodule also determines "competitive" prices for electricity generation. Unlike cost-of-service prices, which are based on average costs, competitive prices are based on marginal costs. Marginal costs are primarily the operating costs of the most expensive plant required to meet demand. The competitive price also includes a "reliability price adjustment," which represents the value consumers place on reliability of service when demands are high and available capacity is limited. Prices for transmission and distribution are assumed to remain regulated, so the delivered electricity price under competition is the sum of the marginal price of generation and the average price of transmission and distribution.

2.7.5 Emissions

The EMM tracks emission levels for SO_2, NO_x, and mercury. Facility development, retrofitting, and dispatch are constrained to comply with the pollution constraints of the CAAA of 1990 and other pollution constraints. An innovative feature of this legislation is a system of trading emissions allowances. The trading system allows a utility with a relatively low cost of compliance to sell its excess compliance (i.e., the degree to which its emissions per unit of power generated are below the maximum allowable levels) to utilities with a relatively high cost of compliance. The trading of emissions allowances does not change the national aggregate emissions level set by the CAAA, but it does tend to minimize the overall cost of compliance.

3. ANALYSIS OF A 20% NON-HYDROELECTRIC PORTFOLIO STANDARD

Concerns over the possibility that climate change may be caused by anthropogenic activities, particularly the combustion of fossil fuels, have raised interest in examining a series of policy options that may inhibit or reverse the growth of energy-related carbon emissions. Recently, a number of bills have been introduced by the U.S. Congress that would simultaneously reduce emissions of NO_x, SO_2, mercury, and CO_2 from power generators. Two of the more recent policy proposals include Senate bill S. 1766 and House bill H.R. 4. These analyses were developed using the *AEO 2002* version of the NEMS. Other related analyses have been performed at the request of the House Committee on Government Reform and the Senate Committee on Environment and Public Works. These may be viewed or downloaded from the EIA Web site.

The policy considered in this article is a 20% non-hydroelectric RPS. A typical RPS requires that a share of the power sold must come from qualifying renewable facilities. Companies that generate power from qualifying renewable facilities are issued credits that they can hold for their own use or sell to others. To meet the RPS requirement, each individual electricity seller must hold credits—issued to qualifying renewable facilities or purchased from others—equal to the share required each year. For example, a supplier of 10 TWh of retail electricity sales during a year with a 10% RPS requirement would have to hold 1 TWh of renewable credits. In a competitive market, the price of renewable credits would increase to the level needed to meet the RPS requirement. The RPS provides a subsidy to renewable generators (from nuclear, coal, natural gas, oil, and hydroelectric generators) to make them competitive with other resource options while allowing the market to determine the most economical renewable options to develop.

3.1 Assumptions of the RPS

In the 20% RPS case, we assumed the *AEO 2002* reference case assumptions in all areas except the following:

- A non-hydroelectric RPS of 20% by 2020 was assumed, growing linearly from approximately 2.5% in 2003 to 20% by 2020. A renewable credit trading system was assumed. Sellers of electricity must hold renewable credits that are equal to the minimum renewable portfolio fraction of generation assumed for each year.
- Because a renewable portfolio policy implies greater public acceptance and, as a consequence, lower legal and other preparation costs, the RPS incorporates the assumption that wind system costs will be lower, and the maximum allowable generation from intermittent technologies (e.g., wind, photovoltaics [PV]) was raised from 15 to 20% of total generation in the RPS case. (Many electricity market experts assert that even 15% intermittent power generation is too high without backup due to reliability. As a rough rule of thumb, intermittent power cannot exceed the reserve margin of the power control center.)
- All existing state and federal programs or mandates for renewable power generation are assumed to adhere to existing legislation and are not included in the costs attributed to federal RPS programs.

3.2 Results

In 2003, 12 states had RPSs, targets, or other incentive programs. Without the current federal and state wind programs, grid-connected wind systems or other intermittent renewable generation technologies are not expected to be economical in the United States over the midterm through 2020. Niche markets are another matter because electricity options are often quite limited.

Approximately 4.8 GW of capacity is expected to result from state RPS programs, and 2.2 GW is expected to result from other state-sponsored stimulus programs, over the 2001–2020 time frame. However, because of these capacity additions, technological learning is expected to result in cost reductions for wind; the projected lower wind costs become competitive with conventional fossil-fueled technologies in some regions of the United States during the 2015–2020 period. Renewable capacity built under these programs reduces the incremental quantity of renewable generation needed to comply with the federal RPS program.

3.3 Fossil Fuel Use and Electricity Market Impacts of the RPS

The penetration of renewable generation technologies reduces the construction of the more efficient gas combined cycle and coal gasification generation technologies that would have been built in the reference case, thereby reducing the overall stock efficiency of fossil-fueled electricity generation plants. In 2020, the RPS case is projected to build approximately 70 GW less combined cycle and 26.8 GW fewer new advanced coal units. Renewable generation capacity is projected to increase by approximately 166 GW, much of which is intermittent wind capacity ($\sim$99 GW above the reference case) (Fig. 5).

Fossil fuel consumption in 2020 is expected to be approximately 5.6% lower in the RPS case than in the *AEO 2002* reference case—down from 120.9 to 114.2 EJ. For electricity generation in 2020, fossil fuel consumption is expected to be approximately 19% lower than in the reference case—down from 37.7 to 30.4 EJ.

The RPS electricity production is expected to be slightly lower in 2020 than in the reference case ($\sim$30 TWh or 0.6%) due to the higher electricity prices that result from the RPS. However, electricity generation from non-hydroelectric renewables is expected to increase from approximately 160 TWh

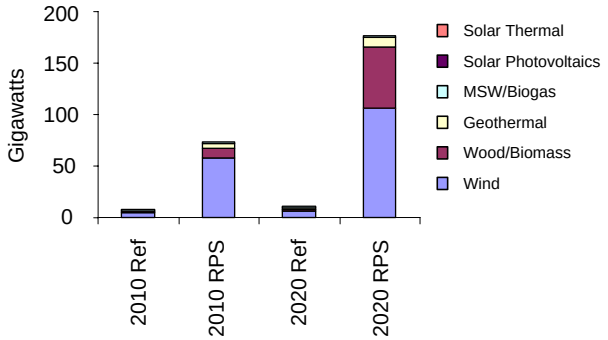


FIGURE 5 Renewable capacity additions: Reference versus RPS case. MSW, municipal solid waste; Ref, reference.

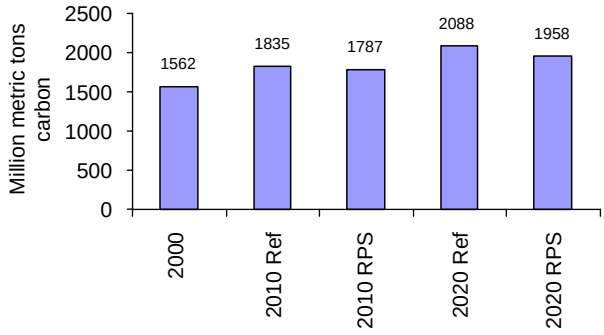


FIGURE 7 Carbon emissions: Reference versus RPS. Ref, reference.

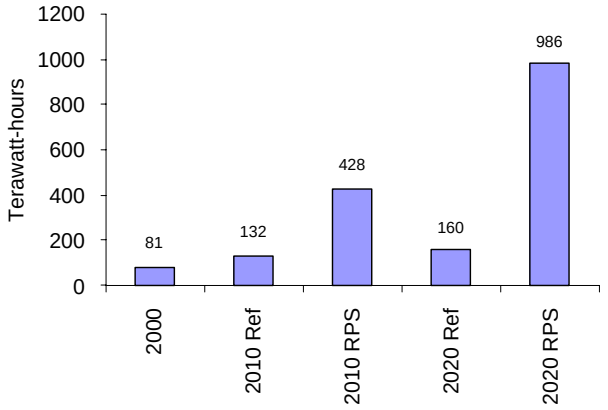


FIGURE 6 Non-hydroelectric renewable generation. Ref, reference.

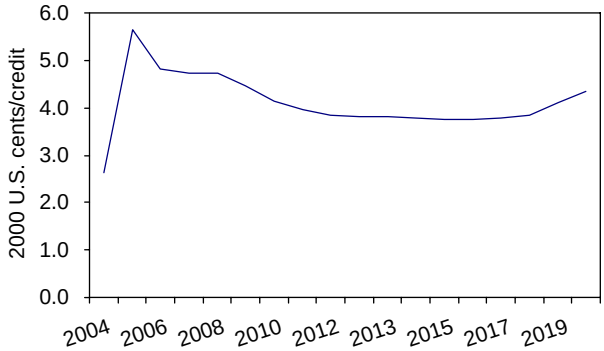


FIGURE 8 Renewable generation credit prices.

in the reference case in 2020 to approximately 986 TWh in the RPS case (Fig. 6). The primary contributors to non-hydroelectric renewable generation in the RPS case in 2020 are expected to be biomass generation from both dedicated plants and cofiring with coal (~476 TWh), wind systems (~335 TWh), and geothermal (~82.5 TWh).

3.4 GHG Emissions

NEMS projects only CO_2 emissions on a systemwide basis for the United States. In 2020, carbon emissions are projected to be approximately 6% lower in the RPS case than in the *AEO 2002* reference case (Fig. 7). However, carbon emissions from electricity generation are expected to be approximately 16.8% lower in the 20% RPS case than in the reference case despite meeting the 20% RPS target in 2020. Consequently, per capita carbon emissions are projected to decline by 0.4 metric tons (~5.5%).

3.5 Costs of the 20% Renewable Policy

Imposition of an RPS on the U.S. generation markets is projected to have a relatively mild adverse affect on delivered electricity prices but at a significant cost to the electricity industry. Electricity prices in 2020 in the RPS case are projected to be approximately 3% higher than in the reference case as a result of the costs that were added to electricity prices to conform to the RPS. The reduction in natural gas prices in the RPS case in 2020 relative to the reference case mitigates against the price increases that would otherwise have occurred due to the renewable credit prices paid and the additional capital invested in renewable generation technologies. However, the cost to the electricity industry over the next 18 years ranges between $38 billion and $59 billion (in 2000 U.S. dollars), using a discount rate ranging from 0 to 10%.

As the industry ratchets up its use of renewables to 20% by 2020, significant issues emerge regarding the ability of the U.S. renewable resources to continue to expand as evidenced by the rising renewable credit prices and the slowing expansion rate of wind and geothermal capacity (Fig. 8). For

example, geothermal resources are projected to have used more than 50% of all remotely competitive sites, although an additional 10 GW of capacity is possible at significantly inferior and costlier sites. Wind expansion in 2020 is expected to begin to be limited by the amounts of cost competitive wind resources remaining in regions that are sufficiently close to demand and transmission centers to allow further expansion. Although wind capacity is projected to reach 109 GW by 2020 under a 20% RPS, expansion beyond 150 GW is likely to be very expensive. The rapid growth in biomass generation eventually is likely to be limited by the competition between agricultural uses and the generation of an energy feedstock for biomass power generators. Furthermore, dedicated biomass gasification plants must be built within a 50-mile radius of their resources to be cost-effective; otherwise, transportation costs are likely to make biomass generation very expensive (by raising renewable credit prices).

3.6 Conclusions

A 20% renewable portfolio standard for the United States is expected to increase total consumer costs of electricity by approximately 3%. Although this does not appear to be significant on a national level, the regional distributional price effects can be quite significant. For example, producers in regions rich in coal- or gas-based generation are likely to experience much larger revenue reductions than will those that are rich in renewable resources. The RPS is likely to significantly increase the costs to the power generation industry, from $38 billion to $59 billion, for the period ending in 2020.

Whether the benefits of a 20% RPS outweigh the costs is a matter of considerable policy debate within the United States. The answer clearly depends on how costs and benefits are perceived and whether they can be measured at all.

4. OVERVIEW OF MARKAL

4.1 Purpose of MARKAL

MARKAL stands for market allocation. The model's original purpose was to evaluate energy technologies in a comprehensive and integrated framework. Since its initial implementation during the early 1980s, MARKAL has undergone several significant enhancements that have extended its scope to the computation of a competitive, dynamic, energy supply–demand equilibrium. The model represents all stages in the energy market—extraction, trade, conversion, and end use consumption—with each sector selecting from an array of technologies that produce and/or consume energy forms. The technologies, together with their input and output energy carriers, constitute a reference energy system (RES) of the regional energy system being modeled (local regions, countries, or provinces).

MARKAL's original capabilities as a tool for an economic assessment of technologies has been enhanced by the addition of new features to the model, including the ability to simulate endogenous technological learning, the inclusion of interregional trade variables, and the capability to represent major uncertainties.

MARKAL evaluates the economics of a technology in the context of the configuration of the entire energy system. This fact renders it difficult, if not impossible, to evaluate a single technology in isolation because its profitability depends on the presence or absence of competing technologies in all parts of the energy system.

MARKAL may be used in two distinct modes. In the first mode, a perfectly competitive energy market is simulated, where all economic agents (suppliers and producers) make their investment and operating decisions under perfect information and perfect foresight and minimize long-term energy system costs using a single discount rate. This mode is particularly useful for revealing the socially optimal set of technologies and for identifying those technologies that should be targeted to achieve the optimal technological mix over the long run. In the second mode, several market imperfections may be introduced to bring the model closer to a short- to medium-term forecasting tool. For instance, the assumptions of perfect information and of perfect foresight may be replaced by the introduction of uncertainty via stochastic programming or by using the model in a time-stepped fashion to simulate decision making with imperfect foresight. In addition, different discount rates may be assumed for different sectors (and even different technologies), thereby simulating the observed differences in behavior among economic agents. This forecasting mode is better suited to the short to medium term when simulating energy market agents under imperfect information.

4.2 MARKAL Model Structure

The MARKAL model minimizes energy system costs over a specific time horizon, subject to satisfying all of

the energy service demands and environmental or fuel use restrictions. This objective can also be viewed as minimizing the net social cost (or maximizing the social surplus) of the energy system, while satisfying a number of constraints, over the horizon.

We now proceed to flesh out several terms in the preceding description. To configure means to decide which technologies to invest in (and/or abandon) at different dates, which ones to operate (and at what level), how much of each energy form to produce during each period, how much of each energy form to buy and/or sell to other regions, and so on. In addition to time periods (which are usually 5 or 10 years long), MARKAL divides electricity demand into three seasons (winter, summer, and spring/fall) and two diurnal divisions (day and night). These time divisions result in six so-called time slices. These time slices are recognized only for technologies producing electricity or low-temperature heat, neither of which may be easily stored and so require a finer time disaggregation than do other energy forms. As a result, these two energy forms are disaggregated into time slices for each time period.

The energy system of a region includes energy supply sources such as mining, imports, and exports as well as processing, conversion of the various energy forms, and their transport (including interregional transport), distribution, and end use consumption by all sectors of the economy. These activities are represented in MARKAL by means of technologies, which consume and/or produce energy forms and/or materials (collectively termed commodities).

The end use technologies produce special kinds of commodities that are physical products or services for which demands are specified.

The set of technologies, demands, sources, and commodities determine the topology of the "reference energy system" of the region modeled. Figure 9 is a simplified representation of a typical MARKAL RES showing the five broad components usually recognized in each model instance: primary energy resources (SRC), energy conversion into electricity or low-temperature heat (CON), other energy processing (PRC), and energy end uses (DMD), with the right-most oval in the figure representing the demands (DM) for energy services and products.

Each "region" may be a country, a group of countries, or a province/state within a country. Multiregional MARKAL models have been created with a wide variety of geographical scopes, going from a relatively small community to the whole world divided into 15 regions. The regions are

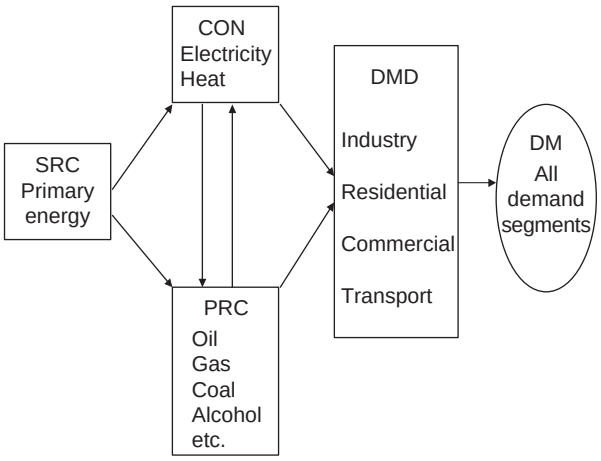


FIGURE 9 Simplified reference energy system. See text for abbreviations.

interconnected by technologies that transport energy forms (e.g., transmission lines, pipelines).

The "horizon" comprises at most nine periods, each having equal duration (usually 5 or 10 years, as chosen by the modeler). At least one of the initial periods is a historical period over which the model has no freedom and for which all of the quantities of interest are fixed to their historical values. This calibration to an initial period is one of the important tasks required for setting up a MARKAL model for a given region. The main variables that must be fixed are the capacity and operating levels of all technologies as well as extraction, exports, and imports for all energy forms and materials. Note that the initial period's calibration also influences the model's decisions over several future periods because the "profile of residual capacities" (i.e., capacities inherited from a historical period) is fully specified over the remaining lives of the technologies existing at the start of the model forecast horizon.

MARKAL minimizes the "total energy system cost," which includes the following elements: investments, fixed and variable annual O&M costs, commodity import costs, minus export revenues, and demand losses incurred from reduced product and service demands. To correctly account for the remaining value of all equipment that is still operating at the end of the horizon, a salvage value (i.e., residual value) of all such equipments is computed and subtracted from the total cost. This is an important feature; without salvage value, the investment decisions made by the model would be severely distorted, especially toward the end of

the planning horizon. Optionally, the user may also specify taxes and subsidies on some commodities (e.g., an emission tax on some pollutant). Because the various costs are incurred at different times over the horizon, each cost is first discounted to a fixed year before being added to the total cost. The same or different discount rates may be used for the various sectors and regions.

The constraints imposed by MARKAL are many. The main ones are as follows:

- *Satisfaction of demands*. A reference demand scenario is provided by the user, specifying the reference demand trajectories in all subsectors. These demands are set for the reference case but may later be modified by MARKAL if some alternate scenario is such that it alters the prices of end use demands (demands can be price elastic in MARKAL). Note also that prices are computed endogenously by MARKAL for all commodities and end use demands.
- *Conservation of investments*. If the model decides to invest in a piece of equipment at a certain period, the capacity is increased accordingly for the life of the equipment. At the end of that life, the capacity is decreased by the same amount (unless the model decides to extend the life of the equipment by investing in a life extension technology). While computing the available capacity at some time period, the model takes into account the capacity resulting from all surviving (unretired) investments up to that period. Some of those investments may have been made prior to the initial period and remain in operating condition (embodied by the residual capacity of the equipment), whereas other investments may have been made by the model at or after the initial period. However, the model is not forced to use all of the available capacity.
- *Use of capacity*. At each period, the model may use some or all of the available capacity in that period times the availability factor of the technology. In some cases, the model may decide to use less than the available capacity at certain time slices or even throughout the whole period. In other words, some capacity may be inactive during some time periods. Of course, this will occur only if such a decision contributes to minimizing the overall cost. Optionally, there is a provision for the modeler to force specific technologies to use their capacity.
- *Electricity balance*. At each period, during each time slice, and in each region, electricity produced plus electricity imported (from other regions) must be at least as much as electricity consumed plus electricity exported (to other regions) plus grid losses. A similar balance exists for low-temperature heat.
- *Balance for other energy forms*. This is the same as for electricity, but because time slice plays no role, there is only an annual balance equation.
- *Electricity peaking reserve constraint*. This constraint requires that at each time period and for each region, total available capacity exceeds the average load of the peaking time slice by a certain percentage. This percentage is called the peak reserve factor and is chosen to reflect the extra load at peak as well as uncertainties regarding electricity supply that may decrease capacity in an unpredictable way (e.g., water availability in a reservoir, unplanned equipment downtime). The peak reserve factor is typically set between 0.20 and 0.50. The peaking time slice is defined as the time slice when load is heaviest (it may be winter day in cold countries, summer day in warm countries, etc.).
- *Emissions constraint(s)*. The user may impose on the whole system upper limits on emissions of one or more pollutants. The limits may be set for each time period separately to simulate a particular emissions profile (also called emission target).
- *Base load (electricity generation only)*. The user may identify which technologies should be considered as base load technologies by MARKAL, that is, those whose operation must not fluctuate from day to night during a given season. The user may also specify what is the maximum fraction of night production that may be supplied from all base load technologies.
- *Seasonal availability factors (electricity sector only)*. The user may specify seasonal and even day/night limitations on the use of the installed capacity of some equipment. This is especially needed when the operation of the equipment depends on the availability of a resource that cannot be stored, such as wind or sun, or that can be partially stored, such as water in a reservoir.

The Canadian incarnation of MARKAL has the following specific features:

- It is composed of 14 linked MARKAL modules, one for each Canadian province and territory plus a U.S. model. All 14 modules are linked by a large number of energy trade variables as well as permit trading variables when required.
- Each provincial module contains full technological descriptions of energy extraction, electricity production, oil and gas processing, coal extraction, industrial processes of all energy-intensive industries, transportation, and residential and commercial end

uses. A typical provincial module contains more than 2000 technologies.

• The demand segments number more than 50 in each provincial module.

• There are more than 500 energy trade links among the 13 Canadian jurisdictions and the U.S. module, encompassing detailed representation of the oil and gas pipelines, electricity interties, and coal and refined petroleum products trading.

4.3 Representation of Energy Market Interactions

MARKAL maximizes the long-term net energy-related social surplus (i.e., the sum of energy-related producers and consumers surpluses) of a state, province, country, or group of countries. Equivalently, the model computes a least cost partial equilibrium on energy markets (i.e., the prices and quantities such that all quantities produced are consumed at the computed market clearing prices). In MARKAL, all energy forms (quantities by fuel type) and their prices are endogenous (computed by the model) and result from the optimization. The demands for products and services are also endogenous but in a different fashion; their price elasticities are chosen by the user, and these choices in turn determine the demand levels.

Mathematically, MARKAL optimizes a linear objective function subject to various constraints. The economic significance of the result of the optimization is twofold. First, the primal solution consists of the investment levels, capacities, and operating levels for all technologies as well as the amounts of each energy form extracted, traded, produced, and consumed by each technology. Second, the dual solution provides shadow prices for all constraints. Three types of shadow prices are of particular practical interest: those of the energy balance constraints (which provide prices for all energy forms), those of the emissions constraints (which provide the price of each pollutant whose emissions are capped), and those of the demand constraints (which provide a price for each demand for a good or service).

4.4 Technology Representation

Each technology is explicitly represented by its technical and economic parameters as follows:

- Nature and quantities of inputs and outputs per unit of technology
- Emission of each pollutant per unit of technology
- Annual availability factor of the technology
- Seasonal availability factors (electricity- and heat-producing technologies only)
- Technical life duration
- First period of availability
- Time profile of existing capacity (residual capacity at initial period)
- Investment cost per unit of capacity
- Fixed O&M cost per unit of capacity
- Variable O&M cost per unit of capacity
- Delivery cost per unit of each input into the technology
- Upper and/or lower bounds on capacity, on investment, and on operating level

Note that each parameter, except life duration and period of first availability, may be specified for each time period desired.

5. EXAMPLE APPLICATION OF THE CANADIAN MARKAL MODEL

In 1999, the Canadian government embarked on a systematic study of all aspects of climate change in Canada. This multistage process, named National Climate Change Implementation Process (NCCIP), directly involved a large number of stakeholders regrouped in 15 issue tables, charged with proposing a large array of measures and actions to reduce GHG emissions. The output from these issue tables was collected by the Analysis and Modeling Group (AMG), which was responsible for integrating the issue tables' proposals into a set of coherent strategies using several energy–economy models, including MARKAL. The MARKAL model was selected as one of two energy technology models to provide an integrated analysis of several paths to Kyoto, that is, to analyze the economic impacts on the Canadian economy of various ways of reducing GHG emissions. The results of the two energy technology models were then injected into the macroeconometric model TIM/RIM to complete the economic analysis with macroeconomic indicators.

The NCCIP was divided into two phases. During phase I, several paths to Kyoto were examined under fairly general economic instruments. During phase II, a restricted number of paths were examined more closely and the market instruments were made more precise.

5.1 Scenarios and Paths to Kyoto

5.1.1 Phase I of the AMG

Three scenarios were considered in addition to the baseline:

- *Canada acts alone (CA) scenario*. In this scenario, only Canada implements GHG reduction strategies, whereas the rest of the world does not. Therefore, the oil price remains the same as in the base case, and energy trade with the United States is negatively affected by Canada's GHG policies.
- *Kyoto–loose (KL) scenario*. Annex I countries (including the United States) implement Kyoto-like GHG reductions, and an international permit trading system is in place. The permit price is low ($25/metric ton [tonne] CO_2). Canadian electricity and gas trade is positively affected, but oil trade is negatively affected.
- *Kyoto–tight (KT) scenario*. The situation is the same as in the KL scenario except that the permit price is high ($50/tonne CO_2). Gas and electricity trade is positively affected, and oil trade is negatively affected but to different degrees than in the KL scenario.

A path describes the components of a GHG reduction strategy. Five paths were defined and simulated in the AMG phase I:

- *Path 0*. Force in all of the issue tables' measures without any imposed emissions target.
- *Path 1*. Impose a 1990−4.33% cap on each sector separately.
- *Path 2*. Impose a 1990−4.33% cap on the entire Canadian economy.
- *Path 3*. This is a variant of path 2 where the cap is imposed on approximately 90% of the Canadian emission sources, with the rest being treated via dedicated measures and actions.
- *Path 4*. This is a variant of path 3 where only approximately 40% of the GHG sources are covered by the cap, with the rest being treated via dedicated measures.

Path 0 is not really a path to Kyoto given that there is no guarantee that the desired target will be reached. The issue tables' measures are made available to the model but are not imposed in paths 1 to 4. In paths 1 to 4, only the sectors covered by permits could buy or sell permits from the international market. Path 2 is a 100% efficient path that achieves the target at minimum cost.

TABLE II

Cases Treated during Phase I

	Path 0	Path 1	Path 2	Path 3	Path 4
Scenario					
Canada acts alone	Y	Y	Y	Y	Y
Kyoto-tight			Y		Y
Kyoto-loose			Y		Y

Note. Y, yes.

The scenarios and paths potentially create 15 combinations. Of these, 9 were simulated during phase I plus the base case, as shown in Table II.

5.1.2 Phase II of the AMG

The CA scenario was dropped, and it was assumed that the Kyoto treaty was in effect and ratified by all industrialized countries except the United States.

Two carbon price scenarios were considered (high and low prices for permits). The paths considered (now called cases) were as follows:

- *Case 1*. This is similar to path 3 (maximum coverage of the economy). Permits were not allocated at all. Emitters must buy them at the prevailing carbon price.
- *Case 2a*. This is similar to path 4 (partial coverage of the economy). Permits were allocated gratis to covered industries in proportion to their emissions in 1990. The proportionality coefficient is calculated to achieve the Kyoto target.
- *Case 2*. This is similar to path 4 (partial coverage of the economy). Permits were allocated gratis but proportionally to the output of each industry in 2010.
- *Case 3*. This is similar to path 4 (partial coverage of the economy). Permits were allocated gratis according to a so-called triptych formula by sector and by province. The multiple-criteria formula was devised to favor development of some sectors and provinces while taking into account each sector/province's endowment in energy forms.

In addition to these, two sensitivity scenarios were simulated, both with case 2 only: one with an intermediate CO_2 price and the other with a supplementarity condition imposing a limit on the amounts of international permits that Canada was allowed to purchase. Table III shows the resulting 10 combinations that were simulated during phase II.

TABLE III

Cases Treated during Phase II

Carbon price	Case 1	Case 2	Case 2a	Case 3
High ($50)	Y	Y	Y	Y
Intermediate ($25)		Y		
Low ($10)	Y	Y	Y	Y
Supplementarity constraint		Y		

Note. Y, yes.

5.2 Some Results from the Analysis

5.2.1 Main Conclusions for Phase I

5.2.1.1 CA Paths

- Path 0 did not achieve the Kyoto target (fell short by more than 40%), and its total cost was quite large.
- Path 1 nearly achieved the Kyoto target (shortfall of 10%), with a total cost 2.5 times that of path 2.
- Path 2 (efficient path) achieved the Kyoto target at a net present value cost equal to $25 billion (the smallest of all paths).
- Path 3 (broad coverage) achieved the Kyoto target at a cost reasonably close to that of path 2. Because path 3 is more likely to be implementable in practice, it becomes a good candidate for further analysis and potential implementation.
- Path 4 (large final emitters coverage) achieved the Kyoto target at a cost closer to that of path 1 than to that of path 2. Therefore, narrow permit coverage is much less efficient than broad coverage.

5.2.1.2 KT and KL Paths When international permits are available, the overall Kyoto costs are significantly smaller than when Canada acts alone. In the KT path, Canada actually incurs negative costs when implementing the Kyoto GHG reductions. This is due to the increased energy trade with the United States (gas and electricity). In the KL path, overall costs for Canada are close to zero.

5.2.2 Main Conclusions for Phase II

- The overall cost for Canada of achieving the Kyoto target is very sensitive to the price of carbon rights, but it is much less sensitive to the case considered. For instance, for the high carbon price, the net present value varies from $+$12 billion to $+$18 billion, depending on the case analyzed. For the low carbon price, the net present value stays in the range of –$1 billion to –$2 billion.
- However, the sectoral and provincial detailed costs vary quite a bit from case to case. Cases 1 and 2a show heavy burdens on industries, whereas cases 2 and 3 succeed in smoothing the costs better across sectors. This is due to the output-based allocation of permits adopted in these two cases.
- The imposition of a 50% limit on permit acquisition proves to be quite costly for Canada. In the low carbon price (case 2), Canadian costs jump from –$2 billion to $+$5 billion.
- The triptych formula partially succeeds in alleviating the burden for the targeted provinces. It is inferred that some additional refinements of the formula would achieve a better repartition of the Kyoto burden than in case 2.

SEE ALSO THE FOLLOWING ARTICLES

Bottom-Up Energy Modeling • Computer Modeling of Renewable Power Systems • Decomposition Analysis Applied to Energy • Input–Output Analysis • Modeling Energy Markets and Climate Change Policy • Modeling Energy Supply and Demand: A Comparison of Approaches • Multicriteria Analysis of Energy • Net Energy Analysis: Concepts and Methods

Further Reading

Adams, D. M., Alig, R. J., Callaway, J. M., and McCarl, B. A. (1996). "The Forest and Agricultural Sector Optimization Model (FASOM): Model Structure and Policy Applications, USDA Forest Service Report PNW-RP-495." U.S. Department of Agriculture, Washington, DC.

Edmonds, J. A., Pitcher, H. M., Barns, D., Baron, R., and Wise, M. A. (1995). Modeling future greenhouse gas emissions: the second generation model description. *In* "Modeling Global Change" (L. R. Klein and F. C. Lo, Eds.), pp. 295–340. United Nations University Press, Tokyo.

Energy Information Administration (1994). "Annual Energy Outlook 1994, with Projections to 2010," DOE/EIA-0383. EIA, Washington, DC.

Energy Information Administration (1998). "Impacts of the Kyoto Protocol on U.S. Energy Markets and Economic Activity," SR/OIAF/98-03. EIA, Washington, DC.

Energy Information Administration (1999). "The Comprehensive Electricity Competition Act: A Comparison of Model Results," SR/OIAF/99-04. EIA, Washington, DC.

Energy Information Administration (2001). "Analysis of Strategies for Reducing Multiple Emissions from Electric Power Plants:

Sulfur Dioxide, Nitrogen Oxides, Carbon Dioxide, and Mercury and a Renewable Portfolio Standard," SR/OIAF/2001-03. EIA, Washington, DC. www.eia.doe.gov/oiaf/servicerpt/epp/index.html.

Energy Information Administration (2001). "Analysis of Strategies for Reducing Multiple Emissions from Power Plants with Advanced Technology Scenarios," SR/OIAF/2001-05. EIA, Washington, DC. www.eia.doe.gov/oiaf/servicerpt/epp/index.html.

Energy Information Administration (2001). "The National Energy Modeling System: An Overview," DOE/EIA-0581. U.S. Department of Energy, Washington, DC.

Gabriel, S., Kydes, A. S., and Whitman, P. (2001). The National Energy Modeling System: A large-scale energy–economic equilibrium model. *Operations Res.* **49**(1).

Hillier, F. S., and Lieberman, G. J. (1990). "Introduction to Operations Research." McGraw–Hill, New York.

Loulou, R., and Lavigne, D. (1996). MARKAL model with elastic demands: application to greenhouse emissions control. *In* "Operations Research and Environmental Management" (C. Carraro and A. Haurie, Eds.), pp. 201–220. Kluwer Academic, Dordrecht, Netherlands.

Manne, A., Mendelsohn, R., and Richels, R. (1995). MERGE: A model for evaluating regional and global effects of GHG reduction policies. *Energy Policy* **3**(1).

Murphy, F. H. (1983). An overview of the Intermediate Future Forecasting System. *In* "Energy Modeling and Simulation" (A. S. Kydes, *et al.*, Eds.), pp. 67–73. North-Holland, Amsterdam, Netherlands.

Murphy, F. H., Conti, J. J., Shaw, S. H., and Sanders, R. (1988). Modeling and forecasting energy markets with the intermediate future forecasting system. *Operations Res.* **36,** 406–420.

National Research Council (1992). "The National Energy Modeling System." National Academy Press, Washington, DC.

Nordhaus, W. D., and Yang, Z. (1996). A regional dynamic general equilibrium model of alternative climate change strategies. *Am. Econ. Rev.* **86,** 741–765.

National Energy Policy: Brazil

SERGIO V. BAJAY
State University of Campinas
Campinas, São Paulo, Brazil

1. **Organization of the Brazilian Energy Supply Industry**
2. **The National Council for Energy Policy**
3. **Fostering Energy Supply**
4. **Energy Efficiency and Research and Development Programs**
5. **Energy Prices and Social Issues**
6. **Energy and the Environment**
7. **An Integrated Approach**

Glossary

cogeneration The simultaneous production of power (either electrical or mechanical) and useful heat (e.g., process steam) using a single fuel source.

energy service company A business that implements energy conservation measures for its customers and is paid by them part of the corresponding cost savings.

firm power Continually available power, or power that is available for a large, prespecified, part of the time.

"free" electricity or gas consumer A consumer free to choose a supplier of electricity or gas, in contrast to a "captive" consumer.

independent power producer An electrical power producer that is not a generation utility, i.e., does not have a concession contract and is not regulated. Independent producers compete, at their own risk, with other producers and, sometimes, with generation utilities in a power supply market.

indicative forward planning Prospective studies, carried out or contracted by government bodies, indicating possible expansion paths and providing guidance about future investment needs to interested agents.

Some of the current national energy policies in Brazil were adopted during the two terms of President Fernando Henrique Cardoso, from 1995 to 2002. Other policies that were set up earlier have evolved over time to their present form. Since the middle of 2003, under the government of President Luis Inácio Lula da Silva, who took office in January 2003, there have been proposals to change some of these energy policies. Potential shifts that may impact the national energy policy of Brazil include institutional and managerial changes in the country's energy supply industry; the fostering of some types of energy supply development and demand-side management programs; the formulation of energy price regulations; tariff making, with cross-subsidies for some large consumers; the granting of subsidies to the poor for fuel and electricity purchases; increasing the cross-links between energy and environmental policies; and integrating the approach to energy policymaking and forward planning. In this latter matter, the roles of the Ministry of Mines and Energy and the National Council for Energy Policy are of paramount importance.

1. ORGANIZATION OF THE BRAZILIAN ENERGY SUPPLY INDUSTRY

From the 1940s and through the 1950s and 1960s, the federal government of Brazil, with the help of the state governments, undertook the charge of assuring, through state-owned companies, the supply of most of the electricity, oil, and gas consumed in the country. A state monopoly for the production, importation, processing (with the exception of private refineries existing at that time), and transportation of oil and gas was defined by the 1953 mandate, Law No. 2004, and was granted to Petrobras, a federal-government-owned company created for the purpose. Distribution and retail trade of oil products were kept out of the monopoly, instead being shared between BR, a subsidiary of Petrobras, and large transnational oil supply companies such as Shell, Exxon, and Texaco. Some Brazilian states formed state-owned companies to distribute and trade initially town gas and later natural gas.

 © 2004 Elsevier Inc. All rights reserved.

The Brazilian government created Eletrosul, Furnas, Chesf, and Eletronorte to generate and transmit electricity for, respectively, the southern, southeastern/midwestern, northeastern, and northern regions of the country, putting all of these entities under the control of a holding company called Eletrobras. Chesf also sells electricity for some very large consumers in the northeastern region, and Eletronorte, besides doing the same in the northern region, also provides distribution services in the capitals of the northern states. All of the state governments formed electrical energy distribution and retail supply companies; some of them, such as Cesp (in the state of São Paulo), Cemig (in the state of Minas Gerais), and Copel (in the state of Parana), also generate and transmit power. At a later stage, the federal government acquired control of two large electricity distribution utilities, Light and Escelsa, which supply areas located in the states of Rio de Janeiro and Espírito Santo, respectively.

Petrobras and Eletrobras, the latter with the cooperation of all of the large state-owned electricity supply utilities in the country, did the required forward (expansion) and operation planning exercises for the oil and gas and electrical power supply industries, respectively, and proposed the associated energy policies for the Ministry of Mines and Energy. This scheme, of having state-owned companies as the main operators in the Brazilian energy supply industry, involving both federal and state governments, succeeded up to the mid-1980s, when a number of policy positions cast shadows on the adequacy of the scheme for the future. The missteps included (1) the artificially low tariffs for electricity (mirroring most public service tariffs imposed by the federal government, in often vain efforts to control high inflation rates) and (2) the political misuse of electricity supply and gas distribution utilities (involving incompetent and often corrupt management and the initiation of construction of several plants, particularly electric power stations, primarily to reap political benefits to some politicians, but without the necessary funding to finish them on schedule), coupled with the desire of the federal government to have substantial and fast increases in domestic production of oil and gas. Discussions about what institutional changes should be made to correct the problems dragged along through several governments and lasted nearly a decade, up to the time when a deep financial crisis in the electricity supply industry required urgent action. President Fernando Henrique Cardoso, in the beginning of his first term in office, decided to sell all of the federally owned electrical power utilities to private investors and also to exert political and economic pressure on state governments to do the same. At the same time, two amendments to the Brazilian Constitution were passed in congress, ending the legal monopoly of Petrobras and allowing the state governments to grant concessions to investor-owned gas distribution utilities and not just to state-owned ones, as was the case before this change. The control of Petrobras continued to be in the hands of the federal government, but Petrobras was expected to compete with private companies in the production, importation, processing, and transportation of oil and gas, allowing, according to the government's wishes, lower prices in the market and substantial increases in the domestic production of these commodities.

At the end of President Cardoso's second term (December 2002), some competition was achieved in domestic exploration; four bidding rounds were carried out to grant exploration and production (EP) licenses, but no licenses have been granted for production of oil and gas (no large findings have occurred outside of Petrobras' EP areas) and few licenses were granted for importation of natural gas. No competition has occurred in the processing and transportation of oil and gas. Several transnational companies operating in this industry have preferred so far to set up joint ventures with Petrobras, rather than to challenge a competitor with such market power as Petrobras still has in Brazil.

Privatization in the electricity supply industry has occurred to a much more limited extent than was planned initially by the government. Around 70% of the distribution capacity was actually privatized but less than 30% of the generation capacity went to private ownership (the large-scale generation and transmission utilities Furnas, Chesf, Eletronorte, Cemig, and Copel continue to be state owned). This partial failure of President Cardoso's government plans was caused by strong political opposition to the privatization of these utilities, not just from opposition parties but also from the government's own rank and file, particularly after the electricity supply shortage of 2001. Rolling blackouts were avoided due to a power rationing program, in effect from June 2001 through March 2002; also, several short-construction-time generating plants were built, to provide reserve capacity, and some generation and transmission facilities were brought online ahead of schedule.

Big changes, however, were made in the Brazilian electrical power supply industry. A regulated third-party access system was mandated for both transmission and distribution networks. An independent

regulatory agency (ANEEL), a national system operator (ONS), and a wholesale market (MAE) were created; because of legal disputes among some utilities, the latter did not settle the short-term transactions (spot market) up to the end of 2002. The distribution and trade activities of the distribution utilities now have separate accounting systems, and for some utilities, the generation and transmission businesses were split into different companies during the privatization process. Some new agents in the electricity market were created, such as the independent power producers (IPPs), the "free" consumers (who, as opposed to the traditional "captive" consumers, can choose their electricity suppliers), and the pure traders (who do not own any generation, transmission, or distribution assets).

In the oil and gas supply industry, a negotiated third-party access scheme was defined by Law No. 9478, which detailed, in 1997, the new "rules of the game," to promote competition in the industry. The same law created the National Petroleum Agency (ANP), an independent agency that regulates the whole oil supply chain and the upstream activities of the natural gas supply chain.

The opening up of the Brazilian energy supply industry to private investors, in order to redirect public investments to other areas and to introduce competition in the industry, in line with what is happening in several other countries, was the main energy policy of President Fernando Henrique Cardoso. As a result of this policy, there are now both private and state-owned large companies in both main branches of the industry (oil/gas and electricity supply). Electricity tariffs, rising much faster than the country's inflation rates, represent a big problem facing the new federal administration. This has been exacerbated in the past few years by growing marginal costs (particularly for generation), by clauses in concession contracts linking the annual tariff updates to an inflation index (which has been overvaluing the strong devaluation of the local currency, the real, since 1999), and by the perception among private investors in the industry of a high-risk business environment, which, in the short term, either increases profit expectations or decreases investments.

In order to solve this problem, by making some changes in the current institutional model of the Brazilian electric power supply industry, the government of President Luis Inácio Lula da Silva intends to negotiate with the utilities the inflation index issue and to pursue programs to reduce the perceptions of high risk. The main changes proposed by the Ministry of Mines and Energy in July 2003 and, after discussions with the interested parties, reaffirmed with minor modifications in December 2003, are as follows:

1. The electricity supply market will be divided into two parts, one comprising the free consumers and the other comprising the captive consumers. Free consumers choose their suppliers among independent power producers, or traders, and buy from them their energy requirements, through freely negotiated bilateral contracts; the captive consumers deal with a distribution utility, through a pool managed by a new entity, the Administrator of Electricity Contracts (ACEE), which will replace the current electricity wholesale market (MAE) organization.
2. The tasks of ACEE will be management of long-term bilateral contracts among generators and distribution utilities and settlement of contractual differences for all market agents.
3. A new state-owned company, the energy research company (EPE) will be created to carry out the long-term (20 years ahead) and medium-term (10 years ahead) expansion planning exercises for the Ministry of Mines and Energy (MME); the resulting plans will be publicly discussed and eventually modified before final approval and implementation by the Ministry.
4. The plan for 10 years ahead will define the hydropower plant projects, the predefined energy and capacity generation blocks for thermal power plants, the regional constraints, and the transmission lines that should be auctioned by MME (no longer ANEEL), in addition to the required commissioning dates, to meet the forecasted demand of the pool consumers.
5. The bidding process referred to previously will allow proposals, by interested agents, of alternative projects to fulfill the energy supply or transmission needs as outlined in the plan. The proposal requiring the least revenue during the concession period will be the winning bid.
6. Preference will be given to public service generation utilities, instead of independent power producers, to supply the pool. Such utilities will sign a concession contract with MME and will have their firm power shared among all distribution utilities of the national interconnected grid, through compulsory long-term bilateral power purchase contracts.
7. Meeting the forecasted demand of the distribution utilities for the next 5 years should be fully assured through these long-term power purchase contracts. Special contractual arrangements are proposed for additional power purchases, in association with nonpredicted demand requirements.

8. EPE, ONS, ACEE, and ANEEL will help MME to oversee the supply conditions over the next 5 years, proposing corrective measures whenever necessary, which includes setting up generation reserve margins.

These changes were brought to the Brazilian Congress in December 2003, in the form of proposals of two new laws, to be enacted in the first half of 2004.

President Silva has emphasized that no further privatizations in the electricity supply industry will take place during his presidency. The partial deverticalization process carried out in this industry under the previous administration will continue in the new administration; vertically integrated utilities will form separate companies to carry out generation, distribution, and trade activities with free consumers, whereas separate accounting systems will suffice in the case of generation and transmission.

2. THE NATIONAL COUNCIL FOR ENERGY POLICY

The National Council for Energy Policy (CNPE) was created in 1997 by Law No. 9478, but was not actually installed until October 2000. According to a report issued in 2001 by a commission formed by the government to search for the causes of that year's electricity supply shortage, earlier activation of the council would have helped to avert or, at least, to minimize the effects of the shortage.

The National Council of Energy Policy consists of ten members, seven ministers of state bureaus (Mines and Energy, Planning, Economy, Environment, Industry and Trade, Civil House, and Science and Technology), one representative of the state governments, one representative of the universities, and one citizen expert on energy policy issues; the President of the Republic appoints the latter two members. The CNPE is headed by the Minister of Mines and Energy, who forwards proposals of energy policy resolutions to the President of the Republic; once approved, the proposals have the power of a presidential decree. Thus, CNPE is the most important forum for setting energy policies in the country. At the end of President Cardoso's second term, three technical committees were lending support for the council activities: one committee addressed the activities of the electrical power sector, another dealt with fuel supply chains, and the third focused on activities concerning required changes in the institutional model of the Brazilian electrical power supply industry.

3. FOSTERING ENERGY SUPPLY

Most of the national energy policies aiming to foster various forms of energy supply in Brazil were conceived in the 1970s. Policy development focused on medium- and large-scale hydroelectricity plants, coal-fired power plants, nuclear power stations, large-scale petroleum and gas production from offshore wells located in deep waters, and fuel alcohol production from sugarcane. Policies to boost the generation of electricity from gas-fired power plants, small hydropower stations, wind power, and biomass resulted from decisions made during President Cardoso's government. All of these policies are briefly reviewed in the following sections.

3.1 Hydroelectricity

Brazil has a large hydroelectric potential, i.e., 258,420 MW, of which just 23.9% corresponded to plants in operation in 2002 and 4.3% represented hydropower stations under construction at that time. Since the middle of the 20th century, particularly after the 1960s, the federal and state governments have made large-scale efforts to tap this valuable and comparatively cheap resource, building the plants themselves through state-owned utilities or, more recently, providing credit facilities for private investors, through the National Bank for Economic and Social Development (BNDES).

Of the total installed capacity of electrical power plants in Brazil as of December 2001, the share of hydropower stations was 82.25%; the corresponding figure for public supply plants at that time was 84.92%. During President Cardoso's government, there was a policy guideline establishing that investments in new hydroelectric power plants should be carried out preferentially by private entrepreneurs, with possible minority participation of state-owned utilities in the case of projects of strategic interest for the federal administration. Spokesmen from the new government have declared recently that state-owned utilities will have more opportunity to invest in hydroelectricity than they had in the previous administration.

3.2 Coal-Fired Power Plants

Brazil's recoverable coal reserves as of December 2001 were estimated at 32.4 billion tons, the largest coal reserves in Latin America; the mines are located in the southern states of Rio Grande do Sul, Santa

Catarina, and Parana. The coal's high content of ash and, in most of the mines, sulfur severely limits the use of Brazilian coal in the iron and steel industries, and the remoteness of the mines necessitates great transport distances; in the year 2001, 99.9% of the coal consumed in Brazil's iron and steel plants was imported. The lack of appropriate railway networks in the mining regions also adds a further difficulty to the transportation problem. The Brazilian coal-mining industry has therefore always depended on the construction of new coal-fired power plants to survive. However, these plants have never been competitive with hydropower stations in Brazil, and have thus required subsidies to be built and operated.

Under the old rules of the Brazilian electricity supply industry, state-owned utilities had been building coal-fired power stations in the southern states of the country, close to the mines, for strategic reasons (diversification of the fuel mix for power generation and, as a result of an industrial policy, aiming to increase the domestic production of components for such plants); a fund (CCC) created by an electricity surcharge was formed to subsidize the operation of these plants when required, i.e., during years and seasons of low inflows to the hydro plant reservoirs. A minimum capacity factor for the plants, however, has been fixed, because of minimum annual consumption levels specified in the coal supply contracts, required to keep the mines running.

The new rules of the game in the Brazilian electricity supply industry, aiming to foster industry competition, cast shadows on the future of the coal mining industry in the country; the CCC fund, for instance, will be downsized from 2003 to 2005 and eliminated in 2006, according to 1997 legislation (Law No. 9648). Law No. 10,438, passed by the Brazilian Congress on April 26, 2002, however, opened a new door for the coal producers; the resources of a new fund (CDE), created by this law for the electricity supply industry, can, among other uses, be employed to finance both old stations (for operating expenses, replacing the cash flows from the CCC fund) and new coal-fired power stations. The amount of the CDE fund to be made available for such purposes will be defined on a regular basis by the National Council for Energy Policy.

3.3 Nuclear Power Stations

President Cardoso's administration created Eletronuclear, a subsidiary of Eletrobras, to assume responsibility for the nuclear plants in Brazil. Brazil has two operational nuclear power plants, Angra-1 (675 MW) and Angra-2 (1.3 GW), both located at the town of Angra dos Reis, in the state of Rio de Janeiro. Angra-1 was bought from the U.S. company Westinghouse in 1969. The Angra-2 plant came online in 2000, 23 years and $10 billion (U.S. dollars) after construction began; it was the single outcome, in terms of power plant building, of a very ambitious nuclear power agreement signed between the Brazilian and German governments in 1975, which envisaged the construction, by Siemens of Germany, of eight nuclear power stations in Brazil; the agreement also specified the transfer of German technology related to fuel cycle activities (mining, processing, fuel enrichment, fuel element manufacturing, and reprocessing), and the joint development of a new uranium enrichment process.

The construction of a second nuclear plant (Angra-3, with an installed capacity of 1.3 GW), included in the Brazilian/German agreement, was started in 1981, involving foundation works and the acquisition of German equipment. Due to budget cuts and to some opposition from environmental groups, the construction was stalled and the equipment for the plant has been mothballed. However, the electricity supply crisis of 2001 bolstered interest in bringing the Angra-3 plant into service. Those in favor focus on the need of the country to diversify its sources of power generation and to take advantage of its substantial uranium reserves (the world sixth largest: 309,370 t of U_3O_8 as of December 2001); furthermore, there is the fact that about $750 million (U.S. dollars) has already been spent on the plant, including the purchase of about 60% of the required equipment. These resources will be lost if the project is abandoned and Eletronuclear will be unable to develop sufficient scale to become competitive. On the other hand, those against Angra-3 point out that the project will require around an additional $1.7 billion (U.S. dollars) and will take at least 5 years to be completed; it is also emphasized that the population still views nuclear energy with suspicion, because issues surrounding safety and the final disposal of the radioactive residues have not yet been resolved.

CNPE authorized Eletronuclear in 2001 to carry out the necessary economic and environmental feasibility studies (Resolution No. 05, approved in December 2001). In August 2002, the Council voted in favor of Eletronuclear resuming the construction of Angra-3 after the necessary environmental licenses have been granted, if the new government does not decide to halt the process (a CNPE meeting was

scheduled for May 2003 via Resolution No. 8, of September 17, 2002; major outlays in the project will occur only after that meeting). This decision was postponed for 2004 by the new federal administration, in part because of current surplus power supplies in MAE and in part because of urgent cash problems at Eletronuclear, requiring a tariff revision by ANEEL.

Apart from Angra-3, no other new nuclear plants are planned for Brazil. Such plants may be built in the future, but only if safer and less expensive new technologies, still at the research and development stage, succeed on an international scale.

3.4 Oil and Gas Production from Offshore Deepwater Fields

Brazil has the second largest proved oil reserves in South America (after Venezuela), at 1.35 billion m^3, or 8.48 billion barrels, as of December 2001; 88% of the total is in offshore basins and 80% is at depths above 400 m. The natural gas proved reserves as of December 2001 stood at 219.84 billion m^3, or 7.76 trillion ft^3, the fifth largest in South America behind Venezuela, Argentina, Bolivia, and Peru; 60% of the total is in offshore basins and 40% is at depths above 400 m.

The production of both petroleum and natural gas has been rising steadily in Brazil since the early 1990s, reaching, in 2001, 75.22 million m^3, or 1.3 barrels per day (bbl/d) of petroleum, which met 79.3% of the consumption in that year, and 14.04 billion m^3, or 1.36 billion ft^3 per day of gas. Imports of gas from Bolivia started in 1999 and imports from Argentina started in the next year, totaling 4.61 billion m^3, or 0.45 billion ft^3 per day in 2001. The offshore Campos Basin, north of Rio de Janeiro, is the country's most prolific production area for both oil and gas, containing around 80% of the national reserves. The Santos Basin also holds large gas fields. Brazil's oil imports come mostly from Venezuela and Argentina.

As was the case with many other national oil companies, Petrobras initially concentrated on building up its downstream infrastructure, particularly from 1965 to 1974. In the wake of the first oil price shock, in the middle 1970s, the Brazilian government ordered the management of Petrobras to implement three new policies aiming to decrease the effects of the oil price rises on the national balance of payments: (1) international expansion of the company in upstream activities, through a subsidiary, Petrobras International (Braspetro); (2) signature of service contracts, with a risk clause, with private oil companies in regions not yet under exploration; and (3) an increase in the national production of oil and gas through exploitation of offshore, mainly deepwater, fields, which make up most of the Brazilian reserves. The first two strategies failed, but the last one has been highly successful. Petrobras' accomplishments in deepwater production have been internationally acknowledged and the Campos Basin's success at proving giant reserves at great depths has attracted attention worldwide. Many companies have been encouraged to come to Brazil to participate in ANP's promoted bidding rounds, in order to develop upstream exploration, some of them without partnering with Petrobras.

In 1986, Petrobras began the first of the Procap programs (the Petrobras technological development program on deepwater production systems, or Procap 1000). The main objective of this program was to improve the company's expertise in oil and gas production in water as deep as 1000 m. It also consolidated Petrobras' production concept based on floating production systems. Petrobras' Procap 2000, launched in 1993, emphasized the development of technologies aimed at reducing investment and operational costs as well as improving efficiency and extending the working life of equipment at water depths of 1000–2000 m. Procap 3000, implemented in 2000, goes even further, seeking to develop technologies that will make oil and gas production in ultradeep waters, below 2000 m, technically and economically feasible. In 2000, the U.S. Geological Survey published new estimates of global oil reserves suggesting that Brazil might still have some 47 billion barrels of undiscovered oil, almost all in offshore fields, with about 35% in the Campos Basin.

Apart from the second half of the 1980s, when Petrobras' management became involved in a long battle with the Ministry of Economy, which decided to cut back the company's expenditures and investments, the rest of the time the federal administration has supported the company's effort to boost oil and gas production from offshore deepwater fields, striving in the medium term for self-sufficiency in oil production. This is likely to continue with the new administration.

3.5 Fuel Alcohol from Sugarcane

Since 1975, with the creation of the National Alcohol Program (Proalcool) by Federal Government Decree No. 76,593, Brazil has produced anhydrous alcohol from sugarcane; this alcohol is blended with

gasoline in Otto cycle car engines in proportions of up to 25%. With the second phase of Proalcool, which started in 1979 (Federal Government Decree No. 83,700), hydrated alcohol has also been produced for use in Otto cycle engines modified to run on 100% ethanol, or neat alcohol.

Currently, Brazil is the world's largest producer of sugarcane, with crops often yielding over 300 million tonnes of crushed cane per harvest season. Prior to Proalcool, the Brazilian share was less than 15% of worldwide production. During the 1970s, many ethanol distilleries were installed in the country, either as new plants or as distilleries annexed to existing sugar mills. The main alcohol-producing states are São Paulo (contributing over two-thirds of the total), Rio de Janeiro, Alagoas, and Pernambuco. Since Proalcool was created, two main products have been obtained from sugarcane: sugar and fuel ethanol. The former has been an important component of the basket of commodities exported by the country since the time when Brazil was a colony of Portugal; in contrast, exports of fuel ethanol have been sporadic and have faced many protectionist barriers abroad.

The production rate of fuel ethanol has varied according to the relative prices of both sugar, particularly in the export markets, and alcohol, which are the main factors affecting production, besides climatic and environmental variables. When sugar prices are high, the production of alcohol decreases, and vice versa. Up to the beginning of the current decade, the price of fuel ethanol was fixed by the government, tracking with the controlled price of gasoline; now both prices are determined by market forces, although they will eventually be subject to government pressures on Petrobras and on the alcohol producers when price increases are considered too high.

The main objective of Proalcool, rapid growth of the alcohol industry, in conjunction with subsidies that increased alcohol production capacity in the 1970s and early 1980s, has facilitated the building of a large alcohol industry. Within the industry, there is still considerable need for increasing energy efficiency and reducing production costs; government policies for this industry have so far failed to address the important issue of cost-effectiveness. The Proalcool program was discontinued in the early 1990s during President Collor de Mello's term in office. The federal government, however, continues to foster the production of fuel ethanol by maintaining the requirement of a mandatory blend of anhydrous alcohol with gasoline. The blend formulation set by the government is between 18 and 25% ethanol, depending on the annual production of ethanol, which, as already pointed out, is strongly affected by sugar prices abroad.

In support of the fuel ethanol program, there are discounts on federal taxes applied to alcohol-fueled cars (the IPI tax) and to fuel ethanol (the recently created CIDE tax). The share of alcohol-fueled cars in total sales of new cars dropped from 96% in 1985 to 1.15% in 2001, after a low of 0.07% in 1997. At the end of 2001, there was an aging fleet of neat alcohol-fueled cars, estimated at about 2.5 million vehicles. New policies have been considered by the federal government to boost the production of hydrated ethanol once more. The most important measures that have been envisaged are the compulsory addition of ethanol to diesel oil in buses and lorries and the establishment of government "green fleets" that will run on neat ethanol. None of these measures, however, has yet been adopted, and the outlook for hydrated ethanol production in Brazil is not promising. On the other hand, the prospects for future growth in the production of anhydrous alcohol to blend with gasoline are bright, not only because of the environmental benefits of such blends, in terms of reduction of air pollution, particularly in large cities, but also because of the good prospects for "flexible fuel" vehicles. These vehicles employ electronic fuel management technologies that allow use of any blend of anhydrous alcohol with ethanol. Some flexible fuel models are already available in the Brazilian market. Because the prices for these vehicles are higher than the prices for the low-ethanol-blend counterparts, the government is evaluating the adoption of financial incentives to help boost sales.

A rapid hydrolysis process to produce ethanol from sugarcane bagasse (crushed cane) is being developed in the state of São Paulo. A demonstration plant should be operating soon. If this technology proves economically feasible, it will allow an increase of around 30% in alcohol production with the use of 50% of the currently available sugarcane waste (tops and leaves, or "barbojo"), without any additional contribution from sugarcane plantations. There have been significant improvements in the productivity of both sugarcane agriculture and the ethanol-based industrial sector. These gains have been due to a combination of factors, including (1) introduction of new and improved sugarcane varieties, (2) better economies of scale from larger and more efficient new plants, and (3) technological improvements and energy conservation measures in old plants. However,

there is still room for further cost reductions. Finding better uses for the sugar and alcohol by-products, such as sugarcane bagasse, barbojo, and vinasse (organic wastewater), is certainly an excellent route to improving the economic performance of these plants. The prospects to increase the current generation of surplus electricity in cogeneration plants located in sugar mills and alcohol distilleries are also promising.

3.6 Gas-Fired Thermal Power Plants

Brazil has a large natural gas pipeline network to transport the gas produced in the Campos and Santos basins to the cities of Rio de Janeiro, São Paulo, and Belo Horizonte. There is also a long pipeline for collecting and transporting the output of gas fields located in the northeastern region of the country to the local capitals and industrial areas; there are plans to interconnect the two pipeline systems. There are also smaller transportation networks to receive the production from the offshore Espirito Santo Basin and from the onshore field of Urucu; the former delivers gas to Vitoria, the capital of the state of Espirito Santo, and to industrial areas in the northern part of that state, and should be connected soon to the Campos/Santos network. Commissioning of the onshore field of Urucu awaits completion of pipelines that will supply Manaus and Porto Velho, the capitals of the states of Amazonas and Rondonia, respectively. All of these gas pipelines are owned by Petrobras.

There are two international gas pipeline connections in the country. The first pipeline to connect Brazil to foreign gas sources was the Bolivia-to-Brazil pipeline, tapping Bolivia's Rio Grande sources and servicing the states of Mato Grosso do Sul, São Paulo, Parana, Santa Catarina, and Rio Grande do Sul; this pipeline came onstream in July 1999. In the Bolivian part of the pipeline, there is a diversion to supply a power plant and other consumers in Cuiaba, the capital of the state of Mato Grosso. Partners in the Brazilian section of the pipeline include Petrobras, which is the major shareholder, Enron, Shell, and BBPP Holdings. The second international pipeline links the city of Parana, in Argentina, to Uruguaiana, in the state of Rio Grande do Sul, Brazil, where it supplies gas to a 600-MW power plant. Transportadora de Gas del Mercosur is the pipeline's operator. Service began in July 2000. An extension of the pipeline, which will connect Uruguaiana to Porto Alegre, the capital of the State of Rio Grande do Sul, to service a new power plant in Porto Alegre, was planned but has been postponed. Additional Argentina–Brazil pipelines are in various stages of planning, although recent natural gas discoveries in Bolivia and Brazil could discourage the development of these projects. It is also possible that a second Bolivia–Brazil pipeline will be built.

The primary motivation behind the projects of most of the recently built or planned pipelines has been the hope for a fast buildup of natural gas demand in Brazil, in conjunction with expectations that there will be construction of a large number of gas-fired thermal power plants. The source of these expectations was the belief of President Cardoso's government that the private investors, under the new rules of the Brazilian electrical power supply industry, would prefer to build efficient, combined cycle, gas-fired power plants, as has been the case in many developed and developing countries, instead of new hydro plants, as has been the Brazilian practice in past decades. Some initial uncertainties among the potential investors caused the Brazilian government to step in; in September 1999, the Gas-Fired Thermal Power Plants Priority Plan, or simply PPT, was announced. The first version of the PPT identified 15 projects, totaling 12 GW, expected to be online by 2003. Specific regulations were established for these projects, such as a specific value for the upper limit pass-through that the electricity distribution companies are allowed to pass on to their "captive" consumers' tariffs. To reassure investors concerned about fluctuations in gas prices, which were in U.S. dollars and were indexed to a basket of fuel oils, the government, through Petrobras, set price ceilings on 20-year fuel supply contracts. There was just one gas price, revised quarterly, regardless of the power plants' location. In addition, the national development bank, BNDES, offered a special loan program.

Uncertainty among the investors remained, however, such that none of the 15 projects got underway. The variation of gas prices with fuel oil prices, the exchange rate, the quarterly price revisions, and the lack of synchronism between the revisions of electricity and gas prices generated investor anxiety. So Petrobras was required to offer an alternative solution with a blended gas price indexed to the U.S. Producer Price Index (All Commodities) (PPI), revised annually. In April 2000, the federal government issued a revised version of the PPT, with the new price option and, in response to political pressure from local politicians and state governors, increased the number of projects from 15 to 51, all over the country. Apart from the ambitious and unrealistic number of proposed plants, the rapid devaluation of the Brazilian currency, the real,

against the U.S. dollar created further difficulties, given that the gas price was set in dollars.

The electricity shortage of 2001 forced the government to launch its Emergency Thermal Power Plant Program, the last version of the PPT program, improving the conditions for all the project developers with gas already contracted or coming on-stream before June 2003 (this was later extended to December 2004), up to a maximum volume of 40 million m^3/day. For these plants, MME/MF Order No. 176, on June 1, 2001, set a new gas price formula, valid for 12 years. The timetable for tariff revisions was rescheduled to bring gas and electricity into line. Petrobras will assume the exchange rate risk for 1 year before passing it on to the power plants at the time of their tariff revision. The annual revision of gas prices considers the Brazilian inflation index IGPM, with a weighting of 20%, and the exchange rate plus the PPI, with a weighting of 80%. Prices will be renegotiated every 3 years and the gas supply contracts are transferable. A further benefit to the plants under the new program, established by Law No. 10,312 on November 27, 2001, was the elimination of PIS/PASEP and COFINS, two federal taxes on the gross revenue accruing from the gas sales to such plants. According to estimates made by the Ministry of Mines and Energy in October 2002, based on the regular follow-ups carried out by the Ministry and considering plants in operation, undergoing trial runs, and under construction, at several stages and contracting levels, 19 gas-fired thermal power plants are likely to come online by 2004, with a total installed capacity of 7157.6 MW, under the umbrella of the last version of the PPT.

When the gas supply contract for the Uruguaiana power plant was signed, during the early stages of the Brazilian electric power supply industry reform, new gas-fired thermal power plants were competitive with new hydropower stations, according to calculations made using the prevailing cheap gas price negotiated for the contract and the reference unit costs for both types of plants, in Brazilian reals. The sharp devaluation of the real since 1999 and high oil prices, however, changed this picture, against the gas-fired thermal power plants. The Brazilian government believed that, in the medium term, such plants would regain their competitiveness. This reasoning was based on (1) decreasing thermal power unit costs accruing from competition among gas suppliers and from the growth in industrial uses for the gas, creating the necessary conditions for the development of a secondary gas market, which, in turn, would allow more flexible "take or pay" and "ship or pay" contractual requirements, and (2) increasing hydropower unit costs arising from plants located farther from the main load centers. Thus the government of President Cardoso decided to subsidize the cost of transporting gas in the country. Law No. 10,604, enacted on December 17, 2002, allows such a subsidy up to R$500,000,000 per year, using the CIDE tax as a resource. With this subsidy, with the lower prices for the commodity expected to accrue from negotiations with the Bolivian government and producers, and with the substantially increased medium-term local production made possible by recent discoveries of large fields in the Santos and Campos basins, the government expects to reduce by $0.50 (U.S. dollars) per million British thermal units or more the price of the gas, which should make the gas-fired power stations competitive again in Brazil.

A fundamental issue is the fact that, in Brazil, in contrast to most other countries, the opportunity cost of natural gas for power generation in public supply plants is determined by hydro generation in new plants. Thus, indexing the price of such gas to the prices of a basket of fuel oils, as is traditional in the oil and gas industry, is meaningless, in economic terms, in Brazil. The new federal administration is less enthusiastic than the previous one about large expansion plans involving gas-fired thermal power stations. Regarding possible new energy policies to boost gas demand in the medium term, incentives may be given for other gas uses, particularly for cogeneration plants in the industrial and services sectors.

3.7 Generation of Electricity from Small Hydropower Plants, Biomass, and Wind Power

Power generation units employing renewable sources of energy (e.g., small hydropower plants, wind power, solar energy, and biomass) and cogeneration plants have received financial incentives in some countries. During the 1970s and part of the 1980s, the major reason was that they represented indigenous sources of energy, reducing the dependence on foreign sources. More recently, with globalization and the formation of economic blocks of countries, this argument has lost much of its early appeal. However, the potential of these generating units to create environmental benefits is being realized. The financial incentives can be orthodox, such as tax relief and attractive credit terms, or heterodox, such as (1) compulsory purchases by utilities of the power

generated by these plants at avoided costs; (2) purchase, for the public grid, of energy blocks through bidding restricted to some types of these plants; or (3) granting of purchase tariffs above the market rate for the energy generated in these plants (to be paid for by all consumers, or on a voluntary basis, depending on the willingness of consumers to pay more for "green" energy.

The sugar and alcohol, and paper and pulp sectors are the Brazilian industrial branches that rely most heavily on self-production to meet their electricity needs. They use mostly cogeneration plants burning industrial residues from biomass, e.g., sugarcane bagasse, firewood, and black liquor (a mixture of chemicals and dissolved wood materials). The new rules of the Brazilian electricity supply industry tend to encourage greater use of such industrial residues for process steam and power generation in cogeneration units, together with forestry residues and sugarcane waste (barbojo), since recent developments in harvesting machinery design and new collection practices are reducing the cost of the latter waste fuels for power generation, particularly if gasification is involved.

During the period 1996–1997, the Brazilian Minister of Mines and Energy discussed with interested parties a possible federal government decree that would oblige utilities to buy surplus power from cogenerators, up to a certain share of their market growth, more or less along the same lines of the American Public Utilities Regulatory Policies Act (PURPA) legislation, during its first phase. The project was badly designed and the proposed measures came up against the main directives dealing with the opening up of the Brazilian power sector. As a consequence, many utilities rallied against the project and succeeded in aborting it. Bearing in mind the American experience related to the application of the PURPA legislation, the board of directors of Brazil's regulatory agency for the electrical power industry, ANEEL, defined, through Resolution No. 021, on January 20, 2000, the minimum share of thermal energy production and the minimum total efficiency requirements that a cogeneration plant owner should meet to become a "qualified cogenerator." The creation of this qualification process aimed to set up procedures allowing the selection of eligible cogeneration units to receive incentives still to be defined. (Some of the requirements of ANEEL Resolution No. 021 should be revised, however, because they are either too slack regarding some combinations of technologies and fuels, or they are too strict regarding other combinations.) The first of the incentives was specified in December 2000 through MME Order No. 551, which included qualified cogeneration plants, using any kind of fuel available in the country, to be commissioned up to December 2003 (this was later extended to December 2004) in the PPT program, having rights to all program benefits.

For quite a while, Brazil's most important development bank, Banco Nacional de Desenvolvimento Econômico e Social (BNDES), has been offering some credit facilities for the building of cogeneration units and electrical power plants using nonconventional renewable energy sources. Although the conditions specified by BNDES are more favorable than what usually can be found on the Brazilian credit market, they are worse than those in the international market, particularly because of the long-enduring prevailing high interest rates in the country. Thus, this credit line was little used before 2001.

Before 2002, the owners of small hydropower stations (up to 30 MW) were the only renewable power producers to enjoy "heterodox" financial incentives in Brazil. In 1998, Law No. 9648 granted these producers access to any consumer with a contracted demand higher than 0.5 MW and relieved them from the payment of half the value of transmission grid use fees. The current minimum power demand limit, which defines "free" consumers, is 3 MW. Law No. 10,438, enacted on April 26, 2002, created the Incentive Program to Generate Electricity from Alternative Sources (Proinfa), comprising wind power, biomass, and small hydropower plants, to be implemented in two stages. Associated with this program, the law defined a new kind of agent in the Brazilian electrical power supply industry—the autonomous independent producer, whose business cannot be controlled or associated with any electricity generation, transmission, or distribution utility. Producers that do not meet this requirement can participate in the program, provided their share in the contracts does not exceed 25% (50% for wind power producers, in the first stage of the program), and no autonomous producer is precluded because of the requirements. Equipment manufacturers can be autonomous independent producers if at least 50% of the value of the equipment involved in the program is produced in Brazil.

Proinfa will hold public solicitations for each kind of power source. Priority will be given first to plants that have already obtained the Installation Environmental License (LI) and then to those holding a Preliminary Environmental License (LP). If more capacity is offered, satisfying the conditions above,

than the capacity scheduled to be contracted, the plants with the shortest remaining environmental license periods will be chosen. In the first stage of the program, 3300 MW, equally distributed among small hydropower plants, wind power stations, and biomass-fueled thermal power stations, will be installed up to the year 2006. Eletrobras will provide long-term contracts to purchase the energy produced by these plants, paying the so-called "economic value" associated with each technology, which should correspond at least to 80% of the average electricity tariff in the country. The cost of these acquisitions as well as the administrative cost of Eletrobras to manage this scheme will be shared among all categories of consumers in the National Interlinked System, proportional to measured individual consumption.

After completion of the first stage, a second stage will continue up to 2022, during which the generation from the plants should meet 15% of the annual load growth and, considering the results of the first stage, 10% of the electricity consumption in the country. Throughout 15-year-long contracts, Eletrobras will again buy the output of these plants, equally among the three technologies if there is enough supply. The purchase will, as before, be preceded by public calls and there will be the same selection criteria as used in Proinfa's first stage, but the price paid will be equal to the weighted average unit cost of new hydroelectric plants with an installed capacity above 30 MW and new gas-fired thermal power stations. The expenses associated with this purchase will again be shared among all consumers proportional to their measured consumption. The difference between the generation cost of each technology and the average unit cost will be paid straight to the producers, using the resources of a new fund (CDE) created by Law No. 10,438. ANEEL is responsible for overseeing the whole process, using, for this purpose, the Renewable Energy Certificates issued by the generators. The CDE fund consists of monies from the annual fees paid by the electrical power supply industry investors to the government for the right to use public goods, the revenues collected by ANEEL from the application of fines, and a new annual fee paid by all agents who sell electricity to consumers.

The creation of Proinfa by Law No. 10,438, inspired by successful legislation in Germany and Denmark, is a landmark approach to foster the generation of electricity from distributed renewable energy sources in Brazil. It has, however, some drawbacks, which should be addressed in future legislation. The first problem is the fact that the market share targets set for the generation of the renewable energy sources have no relationship to environmental policy targets, to the amount of these resources available in Brazil at reasonable costs, to the indigenous manufacturing capacity the government wishes to foster, or to supplementary power sources, e.g., new thermal power plants, required to complement the generation of random energy sources such as wind and hydropower. Also, there was no study of the impact of Proinfa targets on electricity tariffs, which is a major issue in a country with so many poor consumers. These targets should be reviewed in the future. Apart from eventually decreasing "economic values" for each technology, set by the Ministry of Mines and Energy, the program provides no further incentives to increase the cost-effectiveness of these sources; making the Renewable Energy Certificates tradable would be an important step forward in this direction.

In December 2003, the new federal administration put forward a proposal that would limit the annual addition of plants generating electricity from distributed renewable energy sources in the second stage of Proinfa, through specific auctions for such sources, to amounts which would not cause increases in the new pool prices (discussed in Section 1) above 0.5% in any single year and 5% on a cumulative basis. According to the same proposal, from January 2005 onward, the generators interested in participating in the auctions will have to prove that at least 60% of their equipment and services will be produced in Brazil; this share will increase to 90% in 2007.

4. ENERGY EFFICIENCY AND RESEARCH AND DEVELOPMENT PROGRAMS

Several energy efficiency programs have been sponsored by the federal government, as well as by the governments of some states (São Paulo, Bahia, Minas Gerais, and Rio Grande do Sul), since the 1970s. Of the national programs implemented in previous decades and still in operation, the most important ones are Procel, Conpet, and the mandatory energy efficiency programs run by the electricity distribution utilities and overseen by ANEEL.

The Ministry of Mines and Energy and the Ministry of Industry and Trade created, through the MME/MIC Order No. 1877, on December 30, 1985, the Program to Reduce the Waste of Electrical Energy (Procel), to be managed by Eletrobras. A

presidential decree of July 1991 gave the coordination of the program to the Ministry of Mines and Energy and increased its scope, which included electricity consumption labels for household appliances and electrical motors; electricity consumption audits in small and medium-size industrial and commercial premises; financing of research and development initiatives (particularly at the Eletrobras research center, Cepel), directed to the manufacturing of more efficient electrical appliances and motors; support for new legislation and regulation in the country concerning energy efficiency; support of projects directed to reduce losses in the generation, transmission, and distribution of electricity; setting up information systems and marketing activities on energy efficiency; and running educational and training programs aiming to bolster a culture of energy conservation. The Procel program has impacted households, the commercial and industrial sectors, public services such as illumination and water supply, and efficient management of electricity consumption in public buildings. Procel has gone through ups and downs. Since the mandatory energy efficiency programs run by the electricity distribution utilities were set up in the late 1990s, Procel's role has been downgraded; its activities have been redirected to support ANEEL in the evaluation of the utilities' programs.

A 1991 presidential decree created the National Program to Rationalize the Use of Oil Products and Natural Gas (Conpet), to be coordinated by the Ministry of Mines and Energy and to be operated by Petrobras. "Conpet in the School" is its main institutional project. In the transportation sector, Conpet has two successful projects, SIGA-BEM and ECONOMIZAR. SIGA-BEM is a partnership with BR, Petrobras' subsidiary company for the distribution and retail trade of oil products, directed to advise truck drivers in BR's filling stations about how to reduce the consumption of diesel oil in their vehicles. ECONOMIZAR is a partnership with the National Confederation of Transportation, which, through mobile units, provides assistance to garages and service shops of load/passenger transportation companies in the search of the optimal management of their diesel oil stocks, aiming to reduce specific fuel consumption and pollutant emissions of the serviced fleets. Most of Conpet's initiatives in the industrial sector have been directed to energy efficiency improvements in Petrobras' refineries, including a wider use of efficient cogeneration units. In the residential/commercial sectors, Conpet is responsible for pioneering the labeling of liquefied petroleum gas (LPG) consumption of stoves. So far, Conpet has been managing with less financial resources and a shorter scope of action compared to Procel, but has shown a much more stable performance in running its projects.

Since 1998, the concession contracts of the electricity distribution utilities have contained a clause that requires them to apply at least 1% of their annual income to energy efficiency and research and development programs, with at least 0.25% going to demand-side management programs and at least 0.1% going to research and development activities. ANEEL regulates these programs and oversees their results with the help of Procel's staff and experts from some state regulatory agencies that have contracts with ANEEL. ANEEL has set boundary conditions for the range of activities covered by these programs, which, in essence, has been similar to many of those developed earlier by Eletrobras in Procel. Law No. 9991, enacted in July 2000, rules that the electricity distribution utilities should apply annually at least 0.75% (0.5% up to December 2005) of their net operational income to research and development projects and at least 0.25% (0.5% up to December 2005) to energy efficiency programs on the demand side. Electricity generation utilities, independent power producers, and distribution utilities are also required by this law to spend at least 1% of their net operational income on research and development programs. Half of all these research and development resources will be managed by ANEEL, and the other half will be channeled to the Electricity R&D Fund, created by Law No. 9991, to be managed by the Ministry of Science and Technology. The new federal administration intends to redirect half of the resources managed by ANEEL to partially fund EPE's activities.

Petrobras has been sponsoring research and development activities related to the production chain of oil and natural gas since the early days of the company, particularly at its research center, Cenpes. The Petroleum National Agency has, since its installation in 1998, been regulating research and development programs in this field, with resources coming from the royalties paid for by oil and gas exploration and production concessions (Presidential Decree No. 2851, November 30, 1998, which created the National Fund for Scientific and Technological Development for the oil and gas industry) and by the concession contract of Petrobras (1% of the company's gross revenue accruing from production activities). The Ministry of Mines and Energy also funds several applied research projects in

various universities, involving alternative fuels and/or technologies, particularly in the Amazon region.

In terms of energy policy, the most important incentive in recent years to the search for a higher level of energy efficiency in Brazil was the enactment of Law No. 10.295, on October 17, 2001; this law allows the government to set up maximum levels of specific energy consumption, or minimum levels of energy efficiency, for energy-consuming equipment produced in Brazil or imported, after public hearings involving the interested parties. This law also mandates the government to promote energy efficiency measures in buildings. A permanent committee (CGIEE), with members from several ministries, was formed to set goals and to elaborate proposals for the public hearings (Decree No. 4059, December 19, 2001). The first type of equipment to fall under the mandate for minimum energy efficiency levels is the three-phase induction squirrel-cage rotor electrical motor (Decree No. 4508, December 11, 2002).

The mandatory energy efficiency programs run by the electricity distribution utilities, and the electricity supply shortage of 2001, boosted the market of the Energy Services Companies (ESCOs). The main barrier found for further development of this market is the financing of ESCOs, which has been addressed by the Ministry of Mines and Energy through proposals for the opening of new credit lines in state-owned banks and the setting up of a fund with resources provided by Eletrobras and/or BNDES. The certification of ESCOs and the technical qualification of energy efficiency projects are other measures being considered by the Ministry to push forward this market. The promotion of a greater use of high-efficiency electrical motors and household appliances, via credit facilities and tax reductions, and a gradual integration of the national programs directed to energy efficiency improvements are two other important policy measures that have been pursued recently by the MME.

5. ENERGY PRICES AND SOCIAL ISSUES

The price of oil products (gasoline, diesel oil, fuel oil, LPG, naphtha, aviation kerosene, and lubricating oil) in Brazil has been set by market conditions, without any regulation, since January 2002, as mandated by two laws (No. 9478/97 and No. 9990/00). The price of LPG was subsidized before January 2002 because a large number of poor people in the country use this fuel for cooking purposes. Resolution CNPE No. 4, on December 5, 2001, followed by Decree No. 4102 (January 24, 2002), and Law No. 10,453 (May 13, 2002), substituted the cross-subsidy then existing for all consumers of LPG with government grants just for the low-income consumers, registered in government aid-to-the-poor programs, using the resources of the CIDE tax. During the same meeting, on December 5, 2001, CNPE decided that the Ministry of Mines and Energy and the Ministry of Economy should keep regulating the prices of the natural gas produced in Brazil after December 2001, because, in practice, Petrobras had retained its monopolistic position. This control should remain in place until real competition in this market materializes. As defined at the end of the last government by the Ministry of Mines and Energy, this "real" competition will be obvious when there are at least three suppliers, none of which hold a market share larger than 75%. The state governments regulate the gas price for the consumers and there will be no competition in the downstream part of the gas chain in the short to medium term unless a general understanding is achieved among the federal and state governments or the Constitution is changed. There are currently far fewer subsidies available for the sugarcane producers and fuel alcohol manufacturers than in the past, but they still exist, using resources provided by the CIDE tax (Law No. 10,453, December 13, 2002), particularly for the poor northeastern region.

Frustrating the expectations of President Cardoso's government, so far few eligible "captive" electricity consumers have opted to become "free" consumers, i.e., to choose their supplier and negotiate the price to be paid for energy consumption. The main reason is that there are cross-subsidies for energy-intensive large-scale consumers, based on past regulations, discouraging the move from regulated to free arrangements. Recognizing this distortion, but concerned about likely losses in export revenues and job losses in the short term if large tariff increases are imposed on the energy-intensive industrial branches, the previous federal administration decided to spread these increases over 4 years, at 25% per year (Decree No. 4562, December 31, 2002). The current administration has thus decided to increase the transition period to 5 years, with just a 10% increase in 2003, provided the interested energy-intensive industrial consumer invests in generation expansion (Decree No. 4667, April 4, 2003). The government also aims to convince this type of industrial consumer to buy at favorable prices, through medium-term contracts (duration of 1 to 2 years), about 2000 MW, on average, out of a total

of 7500 MW, on average, available in the generation system in the first months of 2003, because of the decrease in demand caused by the current recession. The current recession was triggered by the electricity shortage of 2001, which also speeded up the construction and commissioning of several new power plants.

There have been cross-subsidies to low-income electricity consumers in Brazil for many years. Law No. 10,438 broadened the definition of the low-income consumer, particularly in the poor northern and northeastern regions of the country. In order to avoid sharp tariff increases for certain utilities, especially in those regions, the government also decided to create direct subsidies to low-income electricity consumers using resources provided by eventual extra income earned by the federally owned generation utilities. This extra income could come from trading in the public biddings, defined by Law No. 10,438; it was also decided, if necessary, during 2002 and 2003, to use resources from the RGR fund, managed by Eletrobras.

Expensive thermal energy, generated by engines fueled by diesel oil or, for the larger machines, fuel oil, supplies isolated networks, located mostly in the northern region of the country. This generation is subsidized by all consumers in the national integrated network through a fund known as CCC-Isol, which, according to Law No. 10,438, should last up to 2022.

Between 10 and 15% of the Brazilian population, comprising from 4 to 5 million households, mainly in remote, rural areas, has no access to electricity supply. To improve this situation, the Federal government set up two rural electrification programs in the 1990s, "Luz no Campo" and "Prodeem." Luz no Campo, managed by Eletrobras, has the goal of electrifying 1 million rural properties, basically through grid extensions, using resources from the RGR fund. Up to the beginning of 2003, 541,115 properties were electrified by the program. Prodeem, on the other hand, was conceived to meet the basic social demands, in terms of electricity supply, of isolated small rural communities through local generation, with preference to the use of renewable energy sources. The program, managed by the Ministry of Mines and Energy, has been employing photovoltaic panels as generating units in most cases. Law No. 10,438 determines that the payments for the use of public goods, and fines applied by ANEEL, which contribute to the CDE fund, should be applied with preference to rural electrification programs. The law also states that ANEEL should assign rural electrification targets to each distribution utility. The Agency board of directors hopes to have all Brazilian households electrified by 2016.

The new federal administration report on changes in the institutional model of the Brazilian electric power supply industry, issued in December 2003, proposes that the subsidies for power plants consuming distributed renewable energy sources should be borne by (small) tariff increases; thus the government hints that all of the CDE resources should be channeled with priority to rural electrification programs and for granting subsidies to poor electricity consumers.

6. ENERGY AND THE ENVIRONMENT

Brazil has an advanced body of legislation concerning the environment. Enforcement of this legislation, however, has had failures, mainly due to short budgets of the regulatory bodies, at both federal and state government levels; the situation has improved since the electricity shortage of 2001.

In terms of regulatory tools, the Brazilian environmental legislation uses "command and control" measures such as environmental licenses, pollutant emissions limits, and establishment of zones where certain activities are restricted due to potential environmental damages. The regulation leave little room for market-driven measures involving economic incentives and for negotiated agreements among the regulatory bodies and the agents they regulate, as is happening now in some countries. Environmental policies, planning, and regulation activities are decentralized in Brazil, involving not only federal and state government bodies, but also municipal ones. The same kind of decentralization was established by Law No. 9433, o August 1, 1997, for water resources. This law created a new agent, the Hydrographic Basin Committee, made up by representatives of municipalities, who are responsible for elaborating a Hydrographic Basin Plan and for defining the priorities of water usage in the basin; needless to say this committee is very important to the interests and activities of the energy supply industry.

There has been little connection so far between environmental and energy policies in Brazil. The energy supply shortage of 2001 brought together the work carried out by the Ministry of Mines and Energy and that of the Ministry of Environment, but essentially only on particular projects and mostly to speed up environmental licensing procedures. A joint

agenda for the electrical power sector set by the two ministries in 2002 is expected to enlarge the scope of joint activities, including formal exchanges between technical committees of CNPE and CONAMA, the National Council for the Environment, and a more proactive treatment of environment issues in the electricity supply industry's 10-year forward planning.

7. AN INTEGRATED APPROACH

Energy policies in Brazil have been formulated in the past mainly by the federal government. Separate policies have been developed for each energy supply industry branch (oil and gas, electricity, coal, nuclear, etc.), and these have had little or no relation to other public policies. This has been changing slowly in recent years. The installation of CNPE, which includes the seven most important ministers of state, was a big step forward toward the integration of energy policies with other public policies in Brazil. The development of long-term (20 years ahead) integrated prospective studies each year by the Ministry of Mines and Energy for CNPE, since 2001, for the energy sector as a whole, using alternative development scenarios that take into account the current and possible new economic, technological, and environmental policies, has provided a consistent technical background for such integration. There is, however, a big challenge still to be faced, which is to engage the state and municipal governments, similar to what has happened in the environmental and water resources areas, in order to decentralize to some extent the policymaking process, under the direction of CNPE and with the technical support and supervision of MME. Two important measures in this direction were taken in the second half of 2002 at MME, involving the integration of the National Energy Balance with the state balances, in terms of methodology and data basis employed, and the start of integrated resources planning studies in four Brazilian hydrographic basins.

SEE ALSO THE FOLLOWING ARTICLES

Alternative Transportation Fuels: Contemporary Case Studies • Development and Energy, Overview • European Union Energy Policy • Geopolitics of Energy • National Energy Policy: China • National Energy Policy: India • National Energy Policy: Japan • National Energy Policy: United States • Nationalism and Oil

Further Reading

Bajay, S. V., Carvalho, E. B., and Ferreira, A. L. (2000). Energy from biomass in Brazil. *In* "Industrial Uses of Biomass Energy—The Example of Brazil" (F. Rosillo-Calle, S. V. Bajay, and H. Rothman, Eds.), pp. 27–52. Taylor & Francis Inc., New York.

Comitê Coordenador de Planejamento da Expansão dos Sistemas Elétricos (2002). "Plano Decenal de Expansão 2003–2012—Sumário Executivo." Secretaria de Energia, Ministério de Minas e Energia, Brasília, DF.

Kelman, J., Ventura Filho, A., Bajay, S. V., Penna, J. C., and Haddad, C. L. S. (2001). "Relatório." Comissão de Análise do Sistema Hidrotérmico de Energia Elétrica (criada por Decreto do Presidente da República, em 22 de Maio de 2001), Brasília, DF.

Martins, A. R. S., Alveal, C., Santos, E. M., La Rovere, E. L., Haddad, J., Lisbôa, M. L. V., Correia, P. R. S., Schaeffer, R., Aguiar, S. C., and Bajay, S. V. (1999). "Eficiência Energética–Integrando Usos e Reduzindo Desperdícios." ANEEL/ANP, Brasília, DF.

Ministério de Minas e Energia. (2003). "Modelo Institucional do Setor Elétrico." Ministério de Minas e Energia, Brasília, DF.

Santos, E. M. (2001). "The Brazil Oil and Gas Sector—Outlook and Opportunities." CWC Publishing Ltd., London.

Secretary of Energy. (2003). "National Energy Balance." Ministry of Mines and Energy, Brasilia, DF.

National Energy Policy: China

MARK D. LEVINE and JONATHAN E. SINTON
Lawrence Berkeley National Laboratory
Berkeley, California, United States

1. Energy Policy under Central Planning, 1949–1979
2. Energy Policy under the Transition to a Market Economy, 1979–Present
3. Outlook for China's Energy Policy

Glossary

exajoule (**EJ**) A measure of energy; an exajoule is 10^{18} joules, equivalent to 0.9478 quads (10^{15} British thermal units).

Five-Year Plan (**FYP**) China's Five-Year Plans are developed by the State Development Planning Commission and are used to guide overall socioeconomic development policy.

State Development Reform Commission (**SDRC**) The comprehensive national agency in China that coordinates long-term planning of economic and social development; formerly the State Planning Commission, and then, until 2003, the State Development Planning Commission.

State Economic and Trade Commission (**SETC**) The agency in China that, until 2003, coordinated day-to-day government regulation and economic activities.

In China, the national government has a strong role in the energy sector. Under the central planning system, from the establishment of the People's Republic in 1949 to the initial economic reforms in 1979 championed by Deng Xiaoping, the government directly controlled extraction, generation, transport, and allocation of fossil fuels and electricity. During the transition to a market-oriented economy, the government has gradually, though not always consistently, withdrawn. It lifted controls on energy prices, created energy corporations, and assumed a regulatory role. Like China's other governmental structures, the institutional apparatus for energy policy has been frequently reorganized, with a single ministry responsible for energy in some periods and specialized agencies for each energy type in other periods. In the first three decades of the People's Republic, policy was concerned mainly with increasing supplies of coal, oil, natural gas, and electricity. In the early 1980s, after proposing ambitious economic development goals that would have outstripped conceivable growth in energy supply, China also adopted strong energy efficiency policies. There have been significant efforts to provide the two-thirds of the population that lives in rural areas with better energy services. In recent years, China has made greater efforts to integrate energy supply and efficiency policy with environmental protection and other social goals, in addition to bold economic development goals. The government faces energy security concerns as the dependence on imported oil and natural gas rises.

1. ENERGY POLICY UNDER CENTRAL PLANNING, 1949–1979

1.1 Energy for Industrialization

From 1949 until the 1970s, the Soviet system provided the model for industrial development in China. Investment in energy supply became the highest priority of national investment. As in the Soviet Union, there were large subsidies for extraction of natural resources. Subsidies for energy under the Chinese regime were intended to support expansion of heavy industry and to make energy available and affordable to all citizens. Energy and other natural resources were priced well below their cost of production. The result was very high growth in energy supply. China experienced rapid growth in commercial energy output, from a minuscule 0.7 EJ in 1949 to 18.9 EJ in 1979, an average annual growth rate of 11.8%. Large quantities of biomass fuels were used in rural areas, approximately 3 EJ in 1949 and currently about 6 EJ. The economy grew rapidly, also

 © 2004 Elsevier Inc. All rights reserved.

from a very small base. Energy supply grew even faster; energy elasticity (percent change in energy divided by percent change in GDP) from 1953 to 1979 was 1.26. Energy supply commanded a larger share of total capital investment than any other industrial sector. From 1970 to 1979, one-third of capital investment in industry was allocated to energy. From 1953 to 1979, 33% of energy investment went to coal, 24% went to oil and natural gas, and 43% went to electricity. No attention was paid to the environmental consequences of energy development. Large numbers of miners were killed in coal mines, and mines degraded surface water and groundwater and produced scars on the countryside. Emissions to the air from coal burning were uncontrolled, and solid waste was not treated before disposal.

Because of large subsidies provided to energy supply, lack of incentives to control its use, and absence of economic and environmental discipline, in 1978 Deng Xiaoping inherited a set of energy institutions that were deeply troubled. Most of them were state owned, and virtually all of them would have been bankrupt but for government subsidies. They were inefficient, overstaffed, lacking in modern technology, and without effective management structures and with few incentives to create them.

1.2 Energy Institutions

Under the Soviet-style bureaucratic structure, a web of constantly changing agencies managed planning, production, and distribution of energy. The State Planning Commission (SPC) included energy investment, production, and distribution in its Five-Year Plans (FYPs). The State Economic Commission coordinated and monitored implementation of annual and short-term plans. Day-to-day energy production was directed from a ministerial level. At the beginning of the 1950s, one body, the Ministry of Fuels and Power (MFP), managed production of all types of energy. In 1955, the State Council replaced the MFP with individual ministries for coal, petroleum, and electric power. During the 1960s, the Cultural Revolution reshaped China's institutional structures. Prevailing ideology favored institutional simplification, so in 1970 the Ministries of Coal, Petroleum, and Chemical Industry were merged into a Ministry of Fuels and Chemical Industries. Similarly, the Ministries of Electric Power and of Ministry of Water Resources Utilization were combined. Management of state-run enterprises in the energy sector was transferred to local governments. Poor performance of the energy sector eventually led the central government to resume control of larger enterprises and to reestablish separate ministries for coal and petroleum. Throughout the central planning period, ministries concentrated on short-term planning, and the SPC's FYPs provided only rough guides for long-term direction. Implementation relied on administrative decisions for production and allocation, and each year's targets were based on incremental changes from the previous year's targets.

2. ENERGY POLICY UNDER THE TRANSITION TO A MARKET ECONOMY, 1979–PRESENT

2.1 Energy Supply and Efficiency for Rapid Economic Growth

In 1979, Deng Xiaoping announced that China would quadruple its per capita gross domestic product between 1980 and 2000. This goal had significant implications for energy policy. If energy use continued to grow 26% faster than the economy, as it had over the previous 30 years, total energy use in 2000 would have reached 106 EJ. Actual consumption in 2000 was 38 EJ, while per capita GDP more than quadrupled (Fig. 1; Table I). For comparison, the United States in 2000 used about 117 EJ of energy. Had China continued its previous energy path after 1979, it would have suffered intolerable environmental insults. Hence, one of the

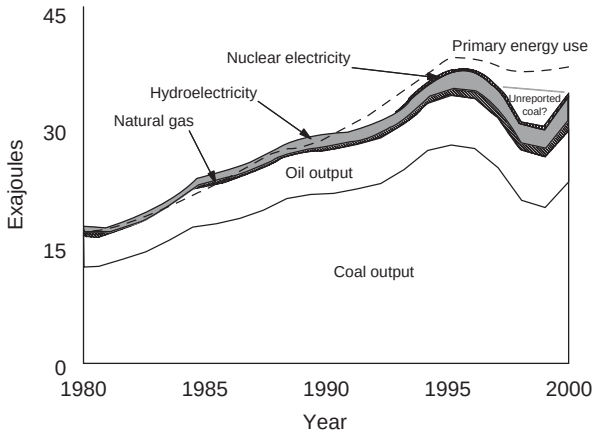


FIGURE 1 Primary energy output and use, 1980–2001. According to official statistics, China's energy output and consumption, mainly coal, rose from 1980 to the mid-1990s before falling temporarily. The difference between total energy output and use in most years represents net exports or imports of oil, except in the period from 1997 to 2000, when most of the difference is probably due to unreported output from small mines.

TABLE I

China's Energy Balances, 1949–2000[a]

	Exajoules used in year										
Energy balance	**1949**	**1955**	**1960**	**1965**	**1970**	**1975**	**1980**	**1985**	**1990**	**1995**	**2000**
Total primary energy supply	0.69	2.15	8.69	5.51	9.08	14.29	18.04	22.75	28.17	37.82	33.98
Coal	0.67	2.06	8.31	4.85	7.41	10.09	13.0	18.3	22.6	28.5	20.9
Oil	0.01	0.04	0.22	0.47	1.28	3.23	4.4	5.2	5.8	6.3	6.8
Natural gas	0.00	0.00	0.04	0.04	0.11	0.34	0.6	0.5	0.6	0.7	1.1
Hydroelectricity	0.01	0.05	0.12	0.15	0.28	0.63	0.7	1.1	1.5	2.1	2.6
Nuclear electricity										0.1	0.2
Net imports and stock changes							(0.64)	(2.33)	(2.28)	0.15	2.39
Total primary energy use		2.05	8.83	5.54	8.56	13.33	17.67	22.48	28.93	38.45	38.19
End use[b]											
Agriculture							1.02	1.19	1.42	1.61	1.70
Industry							11.50	14.46	18.89	26.62	24.95
Transport							0.85	1.09	1.33	1.72	2.91
Services							0.69	0.95	1.38	1.92	2.53
Households							2.81	3.90	4.63	4.61	4.37
Losses							0.81	0.91	1.29	2.03	1.84
Balance							0.38	0.27	(0.76)	(0.62)	(4.21)

[a] Data from the National Bureau of Statistics for various years; *China Statistical Yearbook* (China Statistics Press, Beijing, China).
[b] In Chinese energy accounts, electricity is converted according to the average amount of primary energy consumed in power plants to generate electricity.

most important issues to explore in China's energy policy post-1979 is how the country was able to cut its energy elasticity from 1.26 to a figure less than 0.5.

A second major issue regards the choice of energy forms most suitable to the country's economic development. This has meant a continuation of the emphasis on electricity development, which fostered production of the most valuable and versatile energy form, but limited investment capital to the coal, oil, and gas sectors. A third major issue was protecting the environment, because the greatest environmental impacts on air, water, and land come from energy supply and use. Although energy elasticity fell, energy supply and demand still grew rapidly, because China sustained a 20-year period of extraordinary economic growth (roughly 10% per year by official figures and about 8% according to some independent economic analysts).

Finally, the structure of energy markets was a critical issue in development of the energy system after 1979. It became clear that energy prices needed to reflect costs, which they did not under the central planning system. Energy shortages in the 1980s and nonproductive investments led the Chinese to strive for reforms that encouraged a stronger role for markets for the energy system and for other sectors.

2.2 Transformation of Energy Institutions

Reforms of China's political economy have touched all sectors. Responsibility for overall planning has remained the domain of the State Council, the State Development and Reform Commission (SDRC), and, until its disbanding in 2003, the State Economic and Trade Commission (SETC), but ministries have been repeatedly reshuffled. Authority for energy sector activities has been spun off to large state-owned corporations, which retain many of the same personnel but have more freedom to allocate investments and manage production. The state no longer has the resources to direct the economy through its agencies and now seeks to influence development through institutions that use other means of control. Institutions currently in place may have more of the gloss than the substance of a market-oriented system, but they reflect the irreversible shift to a market

economy. Moreover, local agencies have become larger players.

Since the early 1980s, four major reorganizations occurred: in 1981–1983, 1985–1988, 1993–1994, and 1998–2001. The first reorganization focused on the oil sector and split the Ministry of Petroleum Industry into the China National Petroleum Corporation (CNPC), the China National Petrochemical Corporation (Sinopec), which were responsible, respectively, for upstream exploration and production and downstream refining and marketing. The China National Offshore Oil Corporation (CNOOC) was created in 1979 to manage offshore oil development.

The second wave of reorganizations, in 1985–1988, encompassed all energy subsectors. In place of central ministries with direct responsibility for investment and production, large state-owned companies were formed. A Ministry of Energy (MOE) was established as a coordinating body under the SPC, although it was active only in electricity. A State Energy Investment Corporation (SEIC) was formed and given responsibility for the central government's major investments in the energy sector. In the 1980s, China also incorporated energy conservation into the institutional structure. Energy conservation offices were established at central and provincial levels in planning and production agencies, and over 200 energy conservation technology service centers were set up nationally.

In 1993, MOE was disbanded and once again replaced with separate ministries for coal and electricity. SETC was established and given responsibility for coordinating short-term activities in energy supply. In 1994, control over state investment was unified, and SEIC, along with other state investment companies, was subsumed by the new State Development Bank. The China Energy Conservation Investment Corporation, formed in 1988, was the only exception and remained independent.

A major reorganization in 1998 transformed most remaining industrial line ministries, including those for coal and electricity, into departments of the SETC. Simultaneously, staffs were cut by half. The result was even greater independence for energy corporations and there was concentration of authority over energy supply policy, with SDPC responsible for long-term guidance and SETC responsible for short-term direction. To introduce competition in the oil industry, wells and refinery assets were split geographically between Sinopec and CNPC, which thus became vertically integrated oil companies. Further reorganization of SETC in 2001 absorbed separate departments responsible for energy into a single department with authority for all industrial activity. In 2003, SETC was broken up and its functions spun off to the renamed State Development and Reform Commission and to other agencies. Possible future institutional reforms under consideration include establishment of a utilities commission to set rules for the electric power sector, and reestablishment of a Ministry of Energy.

Nongovernmental organizations (NGOs) have been absent in most areas of Chinese policymaking, but the China Energy Research Society (CERS) has been influential in charting the course from central planning of the energy system to reliance on markets.

2.3 Coal

Coal is China's major fuel, providing an average of 73% of total energy consumption from 1980 to 2000. For many reasons, including its abundance in China, coal received limited policy attention and less capital investment than did any other energy form besides natural gas. Figure 2 makes clear that, after 1984, investment in oil and gas outstripped coal and that electric power investment "took off." By 1997, investment in electricity was 50% more than for oil and gas, and 3.5 times that for coal. This policy

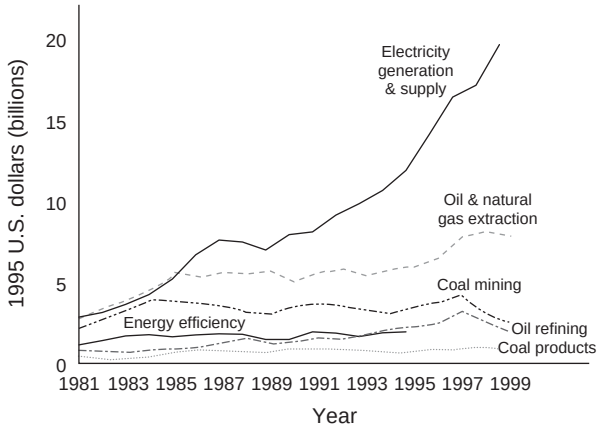


FIGURE 2 Energy supply and efficiency investment, 1981–1999. Investment in power generation and transmission has grown steadily since the early 1980s, and by the late 1990s made up three-fifths of all investment in the energy sector. Investment in other energy supply sectors remained flat until the late 1990s; during that period, investment in oil and natural gas extraction rose, while that in oil refining and coal mining fell, as capacities sufficed to meet demand. China has maintained a steady, even growing, level of investment in energy efficiency projects, in later years exceeding investment in coal mining. These data cover only state-owned and -invested enterprises, which account for the vast majority of energy investment in China.

essentially starved coal for capital, even more than previously. As a result, almost all expansion of coal output from 1980 through 1995 was from small collective and private mines, as shown in Fig. 3. This created a major new source of coal for China at virtually no investment from the central government. Unfortunately, it was an unsustainable policy. The government exerted little oversight over these small mines, which typically produce low-quality coal. Many of them are unsafe, contributing disproportionately to high accident and death rates. They cause serious environmental problems, polluting underground water sources, causing subsidence, and other harm. In most cases, the coal is used locally because it is not of high enough quality to be sold in regional markets, increasing local air pollution as well as pollution in homes (most rural and many urban households burn coal directly for cooking and heating).

Although the policy of encouraging the expansion of coal output from these small mines was clearly not desirable in the long term, there was a clear rationale in the 1980s. China faced a substantial shortage of coal; capital for energy was devoted to forms of energy that would best serve a modernizing economy,

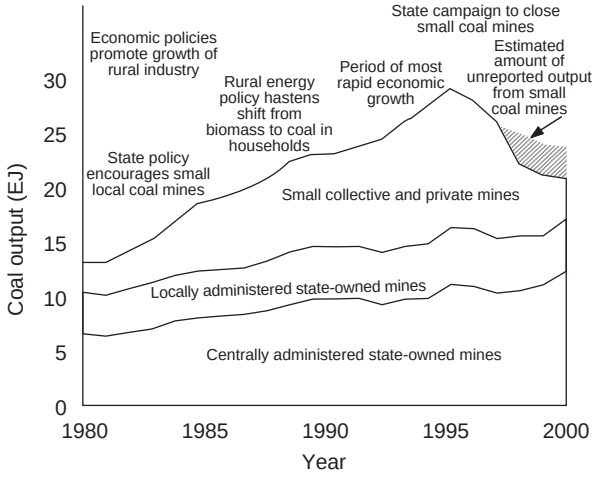


FIGURE 3 Coal production by type of mine, 1980–2001. National policy beginning in the early 1980s promoted small "township and village" mines. By 1995, these often tiny mines produced nearly half of China's coal, but at a tremendous cost in human life, wasted resources, and environmental pollution. Officially reported coal production peaked in 1996 at 1374 million metric tons. Small-mine output was already falling when, in 1998, the central government began a vigorous campaign to close small mines. Reported output from small mines fell by half within 2 years, but actual production from small mines may have been up to one-third higher than reported output, because many small mines continued to operate clandestinely.

and energy sectors were already consuming a large portion of total industrial investment, up from 37% in 1980 to 59% in 1997. Thus, China faced a serious dilemma in allocating investment. It followed a path that promoted economic development, but contributed to the increasingly serious environmental problems. By the late 1990s, China had not only overcome the energy shortages of the 1980s, but, because of declining economic growth, was in a position to reduce its coal consumption. This situation led to a campaign by the central government to close the small coal mines on a massive scale. As shown in Fig. 3, coal output from small collective and private mines declined from more than 600 Mt of coal (45% of all coal produced) to 200 Mt from 1996 to 2001, according to official figures. In fact, as shown by the upper right-hand shaded area in Fig. 3, it is unlikely that all of the private and collective coal mines were closed. These prohibited mines may have produced as much as 150 Mt of coal. Even so, the reduction of coal use from these mines was enormous.

China's coal industry is not appreciably stronger than it was some years ago. It is still undercapitalized, has poor environmental and safety records (about 6000 coal miners die annually in accidents), and has generally weak management, despite the country's heavy reliance on coal. Closing small mines has improved the environmental situation, but either alternatives to coal must be found or the coal industry must be modernized and pollution from both production and consumption reduced.

2.4 Oil and Natural Gas

In the late 1950s, the huge Daqing oil field was discovered in northeastern China, followed a few years later by the large Shengli field in Shandong. Development of these and other fields began in earnest in the early 1970s and oil production expanded rapidly. There was a widespread belief, in China and elsewhere, that the country would become the world's next major source of oil, with oil resources similar to those in the Middle East. The early promise of vast oil resources was not fulfilled. No large fields have been found in eastern China since the 1970s. Untapped oil resources in China are in harsh environments, i.e., in bitterly cold northwestern China and in offshore fields. Both sources are already being exploited. The need for advanced technology, experience, and capital to develop oil in such climates has meant that major projects are joint ventures with international oil companies, and are expensive.

The history of oil imports and exports tells an important story for China. As a result of the early oil discoveries and policy restrictions on nonindustrial oil use, China was an exporter for many years. The peak of its exports, at 1.6 EJ (36 Mt oil) occurred in 1985. Export earnings from energy that year were $7.1 billion (U.S. dollars), or 25% of export earnings for the entire economy. Net oil exports dropped to zero by 1993 and have since increased as dramatically as exports once did. By 2000, net imports were 3.3 EJ (76 Mt oil). With the automobile industry in its infancy in China, but growing very rapidly, dramatic increases in oil demand are expected. This oil will come from overseas, presenting China with a growing security issue.

China's interest in natural gas has grown, and a major exploration program has been initiated. Natural gas is still a very small part of total energy supply: 1.2 EJ (30 billion cubic meters) in 2001, or 3.4% of total energy production. However, production has been growing at an average of over 10% per year since the early 1990s. China is building a major pipeline from the northwest to Shanghai, to be completed in the middle of the first decade of the 2000s. It is not yet clear how the gas will be utilized, because the price of delivered gas will be very expensive. Natural gas has the clear advantage of over coal and oil of reducing pollutant emissions, but unless Shanghai decides to impose high externality costs on coal burning, or forbids the use of coal in various applications, development of the gas market may be constrained.

The major barrier to greater use of natural gas is its high cost in China. In the absence of large gas fields near the eastern population centers, gas will have to be brought in by pipeline from the far western region, imported from Russia or Central Asia, or imported as liquefied natural gas (LNG). These alternatives are costly and pose perceived geopolitical risks. Coal has been used in some industrial processes in highly inefficient ways (e.g., as a feedstock for fertilizer production) because of the lack of natural gas. Substitution of natural gas for coal would be a natural development. A second large potential market is as a substitute for electricity (e.g., in household water heaters), especially as direct coal burning is phased out. In small-scale applications, for which pollution control is prohibitively expensive, natural gas or oil is often the only viable alternative to coal. However, given China's ambitious plans for natural gas development, promoting gas for power generation is an obvious way to spur long-term growth of the fuel. Many markets will depend on China's policies to control emissions from coal combustion and the relative incremental costs of emissions controls and of natural gas.

Oil imports will continue to increase until an affordable substitute for liquid fuels from petroleum is found and widely disseminated. The drivers of future growth will be the rapidly expanding automotive and petrochemicals sectors. Without new large discoveries, China will become more dependent on imports, and has taken steps to mitigate dependency through investments in producing fields overseas. Predicting the future of natural gas is difficult. The key factors will be the rate of discovery of natural gas fields, their size and location relative to demand centers (thus determining the cost), and the development of suitable markets for the natural gas, which likely will remain expensive. After the current efforts to create a natural gas market, investment in continued exploration and development will depend on the assurance of markets for the gas, which will be closely related to environmental policy.

2.5 Electricity

From 1980 to 1998, electricity supply (and demand) grew at 8.3% per year, about the same rate as revised estimates of gross domestic product (GDP) growth. Over 80% of electricity is now generated by fossil-fired power plants, with the remainder coming mainly from hydropower, some from nuclear, and a tiny fraction from other sources. As already noted, the focus of China's energy policy has been electricity development. Investments in electrification have been highly beneficial to the nation in several ways. First, electric power is essential for a modernizing economy, because it provides the highest quality energy suitable for many uses that fuels cannot meet. Second, access to electricity in rural areas, including those suffering great poverty, provides a critical input into the economic and social development of the region. Moreover, it is easier to reduce pollutant emissions from large sources such as power plants, as compared to small ones such as household stoves, and a growing share of China's coal is used in power plants.

On the other hand, the policy of investing heavily in electricity has, as noted earlier, starved other supply sectors of funds. Two resulting problems are the poor quality of marketed coal and the lack of exploration for natural gas. Such problems could be overcome by making markets for international investments more open and transparent, a process that has proceeded slowly since the 1980s. Because China has long viewed energy as key to national

security, it has been cautious about relying on foreign energy supplies or encouraging "too much" foreign investment, although the definition of "too much" has never been made explicit and has apparently varied over time. Recently, the government has permitted companies outside of China to have a controlling interest in energy projects, subject to careful review.

China has the second largest hydropower resources after Russia, but the biggest dam sites are far inland, distant from coastal demand centers. The 1960s and 1970s saw preferential development of hydropower, but as electricity demand grew exponentially in the 1980s, China turned to coal. Large dams continued to be built, particularly with international development assistance, but more slowly than fossil plants, due to high costs of transmission, major problems encountered in earlier projects, and increasing environmental sensitivity. Since the late 1980s, most attention given to hydropower went to the giant Three Gorges project, but concerns about pollution from coal-fired generation and a desire to develop western provinces are helping to drive construction of other hydropower projects. As in most countries, large dams are planned as multipurpose projects, with water supply and flood-control objectives being particularly important. Pumped-storage facilities are also being built in coastal regions where power shortages during peak demand periods have sometimes been severe. Current policy aims to promote delivery of electricity from hydropower and coal from poorer western provinces to the wealthier coastal regions. This is intended to relieve projected shortfalls in generation capacity in the East, and to provide income to the West. Significant issues remain in funding investment in generation and transmission and in restructuring the regulatory system to allow large interregional transfers.

At least five factors affect the efficiency of China's fossil-fired power plants. Two factors—the high quality of large, new state-of-the-art power plants and the relatively high percentage of cogeneration (i.e., use of heat as well as power) in the power system—have contributed to improved efficiency. Two other factors—the small size and outmoded technology of many existing power plants and the relatively low quality of coal burned—have resulted in many inefficient plants. A final factor, the lack of natural gas as a supply source of electricity, has precluded the use of advanced combustion turbines that can produce electricity at high efficiency and with low levels of pollution (as well as favor cogeneration in many cases). In order to improve the technical and economic performance of power generation, the government has long had policies to close the least efficient power plants and to prohibit construction of new small power plants unless they are cogenerators. Over the past two decades, the average size of China's power plants has grown, and the average efficiency of fossil fuel power plants has improved.

The Chinese have long had the goal of developing nuclear power on a large scale, but growth has been slower than planned, largely because of high costs compared with coal. Nonetheless, as of 1993, China had installed eight nuclear generating units totaling 5980 MW, with another three units totaling 2510 MW under construction. China has sought to develop the capability to construct nuclear power on its own, but almost all construction so far has had active overseas involvement.

Hydropower has always been an important part of Chinese electricity supply. The 18,000-MW Three Gorges Dam will be the world's largest hydropower project, nine times the size of the Hoover Dam in the United States. It has been controversial from its inception because of environmental impacts, involving the relocation of 1.2 million people and the loss of historical sites. The reservoir is already being filled; the first turbine is expected to go online in 2004 and the project will be completed by 2009.

China has an active nonhydro renewable electricity program. There are large resources of geothermal energy, but because the high-temperature geothermal resources are in the West, far from end users, they contribute little to total electricity supply. Development of wind power is rapid, although from a low base. There are presently 28 wind farms totaling 403 MW. Two more, one near Shanghai and the other outside of Beijing, are committed for construction. There is discussion of policies to promote wind power, in the hopes that overcoming barriers will render wind a cost-competitive electricity source. Because China has many good-to-excellent wind regimes, wind could become competitive with coal-fired power, depending on environmental regulations.

The government is now restructuring the power system and planning to create regulatory commissions at the national and provincial levels, thus separating power generation from direct government oversight of performance and prices. Generation and transmission will also be separated. The government expects to introduce competition into the power generation market, although there may be only a

small number of companies owning large numbers of generators initially. The government will maintain authority over transmission lines and has ambitious plans for creating a national grid among its six regional and six provincial grids.

Some pressing issues regarding future electricity generation can be addressed by the following questions:

- Will China create meaningful environmental standards for new and existing power plants? When?
- Will China strive to replace coal in a significant portion of new power plants? What roles will natural gas, wind, and biomass play?
- Will the Three Gorges Dam work as planned, or will it run into significant operational and/or environmental problems?
- Will restructuring encourage improved performance of the electricity system?
- Will the bulk of domestic energy investment continue to be allocated to electricity? What will be the role of foreign investment?

2.6 Rural Energy

China's top priorities for rural energy development have been to meet basic fuel needs for cooking and heating, to provide electricity to as many households as possible, and to supply energy for rural industry. This last function has been accomplished mainly through allowing the development of local coal mines and grid extension. Most policy measures have been directed at fulfilling household energy needs. The vast majority of the rural population has at least minimal access to electricity. In remote areas in southwestern China, this has been accomplished through construction of small hydropower projects. Many of the "microhydro" projects of the 1960s and 1970s were poorly designed, but more recent projects have been more successful, and China has become a model for other developing countries. A growing number of remote communities are supplied by small wind generators. Still, as of 2000, about 16 million rural residents (i.e., slightly more people than the population of Florida) had no access to electricity. Most options for supplying those who remain without electricity are beyond local means, so the continued efforts to supply regions without electricity will depend on the central government's limited funds.

Throughout history, China's farmers have periodically faced serious fuel shortages. Various programs over the years have aimed to resolve those shortages. In the 1970s, for example, a campaign resulted in the construction of tens of thousands of small biogas digesters, intended to turn crop and livestock waste into fuel gas. Virtually all of these facilities failed, due to poor design, inadequate materials, and insufficient operating expertise. A new program, instituted in the late 1990s, emphasizes larger scales and better designs, and shows more promise. Direct combustion remains the major way biomass fuels are used. In the early 1980s, in response to rural fuel shortages, the government organized a large-scale National Improved Stove Program (NISP) to provide rural households with more efficient biomass stoves for cooking and heating, and later also improved coal stoves. By the early 1990s, the pressure on biomass supplies had eased in most areas, in part due to increased incomes and availability of coal and electricity. From the mid-1990s onward, support for the stove industry was replaced with extension services and certification systems to standardize stoves. The development and dissemination of improved stoves is now left to market actors. The government claims that by 1998, 185 million of China's 236 million rural households had improved stoves. Although many rural households now use fossil fuels, especially coal and some liquefied petroleum gas (LPG), and ownership of electrical appliances is increasing, nearly all rural households use at least some biomass. Solar water heaters and large plastic-covered greenhouses have become common features in rural areas. Passive solar design has been incorporated into some new housing, and solar cookers have been distributed in some western areas where fuel is short and sunlight is plentiful.

2.7 Energy Efficiency

The achievements of China's energy efficiency policy have been remarkable. As noted earlier, since 1980, energy has grown half as fast as GDP (Fig. 4). Virtually all other developing countries have seen energy demand grow at least as fast as GDP. In many ways, China serves as the "existence proof" that energy can grow at a substantially lower rate than that of the economy over a considerable period of time.

2.7.1 Energy Efficiency under the Planning System

China achieved results in the large-scale promotion of energy efficiency beginning with instruments of the planned economy. Energy efficiency became a national priority in 1980, after a confidential study by

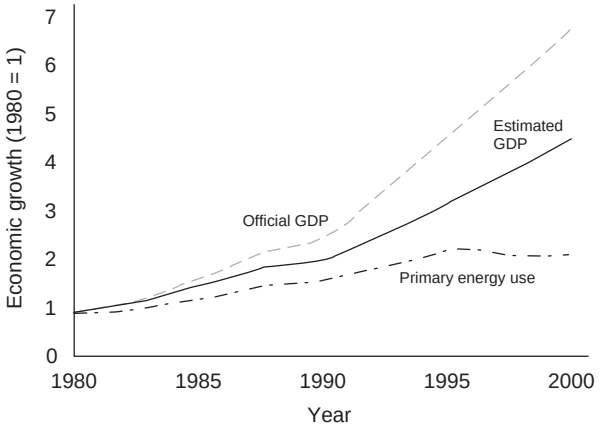


FIGURE 4 Official and estimated gross domestic product (GDP) and primary energy use, 1980–2001. Since 1980, according to official statistics, China's economy grew nearly sevenfold. Many economists believe official statistics exaggerate actual growth. Although experts differ on the correction to apply, a reduction of two percentage points in annual growth rates is a typical estimate. Even with revisions, China's economy in 2001 was 4.6 times larger than in 1980. Over the same period, primary energy use rose to 2.2 times its previous level. Unlike most developing countries, in which energy use rises faster than economic output, China has experienced growth in energy use only one-third to one-half as fast as economic growth.

a group of academics that proposed major changes in China's energy policy; among the proposals was establishment of a national efficiency initiative. In the Sixth FYP (1981–1985), China proposed and implemented a massive effort to promote energy efficiency. In 1981, investment in energy efficiency by the government was 9.5% of total energy investment (Fig. 2), or 1.5 billion yuan (about $830 million U.S. dollars in then-current exchange rates). Investment in energy efficiency, including cogeneration, grew to 3.2 billion yuan by 1985.

No other country in the world at that time, or since, has had a national program that allocated energy investment to energy efficiency, or one that was implemented so rapidly. China created a bureau of Comprehensive Energy Savings and Resource Utilization within the SPC, which was at that time China's top executive agency. This bureau undertook a wide range of reforms, some of which were carried out by a newly created (China) Energy Conservation Investment Corporation. Along with its regional offices, the bureau evaluated proposals from throughout the nation and chose the projects to be funded. The energy efficiency investments were only the most important of many measures put in place during the very innovative period of the early and middle 1980s. Other innovations that remained in place until the late 1990s concerned energy management, financial incentives, technical service centers, and research and development. Large factories were required to have energy managers to monitor energy use. Annual quotas for energy use were set and evolved into regulations relating to energy use of various types of equipment (e.g., motors and boilers) and manufacturing processes. Factories that performed better than the quota or exceeded the standards were given financial rewards. Many inefficient factories were shut down. Demonstration projects illustrated performance of efficient technologies.

Initially, energy conservation investments were financed by the national government. Later, the cost was shared with provincial governments. By the early 1990s, the investments had turned into low-interest loans. Other incentives included reduced taxes on energy-efficient products, subsidies to develop new energy-efficient technologies, and monetary rewards to energy-efficient enterprises. As mentioned previously, China created a network of over 200 energy conservation service centers that employed nearly 5000 staff at their peak; the staff performed energy efficiency feasibility studies, participated in investment projects, and trained energy managers and technicians. China instituted a national energy conservation week (in November), during which the national and local press spotlight energy efficiency and efficiency is included in school curricula.

China began an energy conservation research and development program in 1981 to complement its shorter term programs. The areas receiving the greatest attention were efficiency of coal combustion (e.g., improved briquettes and cookstoves), electricity end-use technologies (e.g., fans, pumps, and motor controls), boilers and furnaces for steel production, heat exchangers and waste heat recovery systems, transport technologies, fertilizer production processes, and industrial-process controls. Few of these programs have been formally evaluated. However, the reduction of the energy intensity of the Chinese economy from 1980 to the present is strongly suggestive of the overall success of the programs, which made more efficient equipment and practices widely known and available to the enterprises that were expanding production capacity in those years.

2.7.2 The Energy Conservation Law

In 1998, the national Energy Conservation Law came into force, codifying the country's approach to promoting energy efficiency under a more market-oriented economic system. Implementing provisions

are still being formulated to address some of the challenges:

- Creation of new fiscal and other incentives.
- Reassessment of existing energy price regulations.
- Establishment and enforcement of mandatory efficiency standards for common energy-using equipment and buildings.
- Integration of energy efficiency with environmental protection efforts.
- Improving efficiency of the rapidly growing rural industrial sector.
- Retaining and transforming the former system of efficiency centers.
- Institutional coordination in implementation, management, and supervision of energy conservation law.

2.8 Energy Pricing

Energy pricing remains heavily influenced by the central government, which has only slowly relinquished control. Under the planning system, state-dictated prices accompanied planned allocations to users. Controls on energy prices determined which energy enterprises and subsectors would make profits and suffer losses. Low coal prices, for example, consigned coal mines to a constant struggle to make ends meet, while price differentials between crude oil and oil products ensured that refineries became wealthy.

Energy-price reform began in the early 1980s and proceeded tentatively, because price policy was—and remains—politically sensitive. A portion of the energy products was allowed to be sold at prices higher than the in-plan prices, resulting in a multitrack pricing system. Today, most energy prices in China track international market levels. China's accession to the World Trade Organization in 2001 represents a step toward full integration of China's energy markets with international markets.

Price liberalization began with coal. In the early 1980s, the government allowed limited free markets for coal and permitted state-owned mines to sell small amounts at market prices, and then allowed a multitrack system. In 1993, the price of nearly all coal was liberalized and prices rose rapidly. State-owned enterprises, which had previously been required to purchase from state-owned mines, were able to buy lower cost (and lower quality) coal from small local mines. In the mid-1990s, coal prices leveled off and then fell as coal demand fell. Price controls remain in place for some users, particularly fertilizer plants that use coal feedstock.

Two distinct markets for coal have emerged, with two different pricing systems. Large mines and consumers, like power plants and steel mills, negotiate supply contracts at national semiannual coal-marketing conferences. The product is of relatively high quality, is shipped long distances, and is expensive. Local markets match small non-state-owned mines with consumers, often in the same area, and provide generally cheaper, lower quality coal. Prices vary significantly by region; in coastal areas, far from the major coal mines, domestic coal can be expensive by world standards, and small amounts of coal are imported.

Oil prices have followed a similar path of deregulation. Central control of wellhead prices contributed to slow growth in oil production during the 1980s, even though a two-tiered pricing system was introduced in 1982. In 1988 the government increased plan prices, but by a very small amount compared to market levels, and by 1990 the cost of production had surpassed the average in-plan wellhead price. Distortions in pricing of upstream and downstream products allowed refineries to profit at the expense of oil fields. Local governments built many small, inefficient refining facilities to capture rents created by this system. At the same time, oil exploration and extraction became much less attractive, resulting in a slowdown in oil field development.

In 1994, the two-tiered pricing system was replaced by a single set of central price controls on crude oil and oil products that were adjusted periodically with reference to international prices. Faced with fixed ex-refinery and retail prices up to twice the level of import prices, oil consumers sought alternatives to the high-priced oil. By 1998, smuggling had become rampant, and a new import and pricing regime—linking Chinese domestic prices for the first time to the Singapore market—was announced along with a crackdown on smuggling. Consequently, oil imports dropped in 1998 (before rising again; Fig. 5) and smuggling activity has declined. In 2001, the basis for the pricing regime was broadened to refer to Rotterdam and New York prices as well as Singapore prices. As yet, China has few domestic price-discovery mechanisms and continues to rely on references from international markets.

Natural gas prices, particularly the large portion of gas provided to fertilizer plants, remain tightly controlled by the government, which has tried to make investment in natural gas more attractive by adjusting domestic prices closer to international levels.

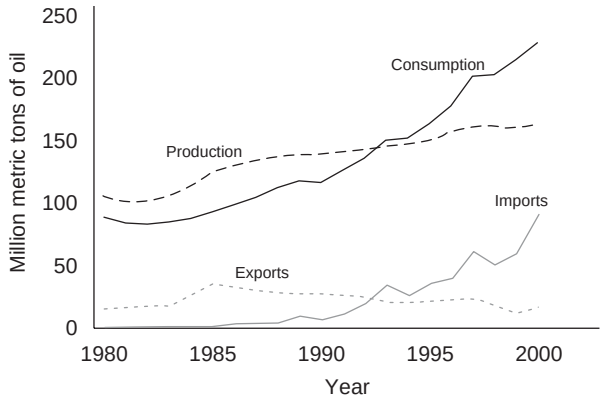


FIGURE 5 Oil production, consumption, and international trade, 1980–2000. In 1993, China went from being a net exporter to a net importer of oil.

However, gas pricing varies widely from city to city, and in the absence of a national pipeline network, a true market for natural gas does not yet exist.

Pricing policy in the electricity sector has been motivated by attempts to ease the serious electricity supply shortages that plagued much of China until the late 1980s. To encourage electricity production, the central government established a multitiered and diversified price system for the sector in 1985. Prices for enterprise-owned power plants (China's version of independent power producers) and small hydropower stations were set higher than plan prices for state-run utilities, and state-owned utilities were allowed to sell above-quota generation at higher prices. As shortages persisted in the late 1980s, the government allowed tariffs to be set on the basis of total costs plus a profit margin to encourage capacity expansion. In 1987, peak-load and seasonal pricing were introduced in some areas. In 1988, state-owned enterprises were levied an added electricity consumption tax for national electric power development, in addition to other national and local taxes on electricity, e.g., the national tax for financing the Three Gorges Dam project. Electricity price classifications, as well as rules and regulations for price settings, remain confusing, defeating any incentives for users to conserve. Prices vary greatly depending on region and class of user, as well as between urban and rural areas. The impact on prices of the current wave of power-sector reforms remains unclear.

2.9 Energy and the Environment

All across China, and particularly in the wealthy coastal provinces, cities and towns are becoming stricter in enforcing limits to pollutant emissions. When local administrations are supportive, environmental protection bureaus can levy significant emissions fees and fines, mandate process changes, and apply more drastic measures. Forceful application of environmental regulations could change industrial energy demand significantly. The regulation of sulfur dioxide emissions in China's legislatively defined "acid rain control zones" may, for instance, result in greater use of washed coal and installation of flue-gas desulfurization (FGD) equipment at power plants. Coal washing would provide a higher heat-content product that would burn more efficiently, reducing demand for coal, all else being equal. FGD, on the other hand, requires a great deal of a power plant's output, raising demand for coal inputs to power generation. Requiring urban factories to move or to replace equipment often results in the use of newer, generally larger, cleaner, and more efficient equipment. Sustained support for environmental policies is a factor in the continuing decline of coal use by households. In households, coal is being replaced by LPG, natural gas, town gas, and electricity.

3. OUTLOOK FOR CHINA'S ENERGY POLICY

China's demand for energy services will grow to meet the demands of its expanding economy and urbanizing population. Widely circulated baseline scenarios of primary energy demand in China to 2020 forecast total energy use to be between 70 and 90 EJ in 2020, compared to actual consumption in 2001 of 39 EJ (Fig. 6). The country faces numerous challenges in meeting this demand in a way that is economically, socially, and environmentally sound. It will be crucial for China to continue its progress in structuring markets, establishing institutions of corporate governance, revamping the finance sector, and finding other means to fulfill the social-welfare functions that formerly were provided by enterprises. These are broad tasks with multiple goals, but they will affect energy supply and use at least as much as policies aimed directly at energy.

Challenges for energy-supply policy abound. In the long term, commitment to developing renewable energy will have a tremendous impact on fuel structure, but for now conventional energy will be central. For coal, ensuring that large mines become financially viable remains a key task. How China finds ways to improve coal supply to meet environmental goals, such as reducing acid precipitation, will affect how coal is used. The debate, driven by

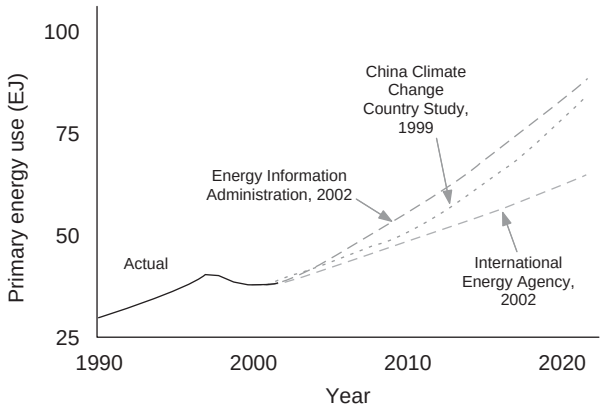


FIGURE 6 Major baseline scenarios of future fossil energy consumption in China to 2020. Widely circulated baseline scenarios of primary energy demand in China to 2020 forecast total energy use to be between 70 and 90 EJ in 2020, compared to actual consumption in 2001 of 39 EJ. The scenarios were adjusted to account for the discrepancies between forecast and actual energy use 2000; trends are the same as in original forecasts.

financial and security concerns, over whether China should rely more heavily on domestic coal resources instead of imported oil, also bears watching, though the potential for greater coal use will hinge again on coal quality, which, in turn, will depend on water availability in the arid coal-mining regions. Greater reliance on coal will signal higher growth in energy use and carbon emissions, because coal emits more carbon dioxide per unit of useful energy. In the long term, China's reliance on coal will be determined not by resources, but by climate change. If, as seems likely, significant reductions of carbon dioxide emissions from fossil fuel use in all countries are required, then coal use will be limited, unless an inexpensive way can be found to permanently sequester large amounts of carbon dioxide.

Because accession to the World Trade Organization has left China with fewer tools to restrict oil imports, policies that affect demand will be crucial. Transport policies that affect mode choices (e.g., road vs. rail; private vehicle vs. public transportation) and fuel-efficiency standards for vehicles, and policies affecting demand for petrochemicals, will influence China's oil imports. Natural gas is such a desirable fuel for so many reasons that consumption will likely be limited by supply. Policies affecting international trade in natural gas, support for construction of pipelines and distribution networks, pricing policies, and regulatory development all become relevant.

For electricity, the main question is how regulatory reform of utilities will proceed. Issues include how government will reduce its role in the management and operation of utilities, how generation will be separated from transmission and distribution, and how markets for electricity will be transformed and regulated. How China treats the activities of foreign participants in the sector will also be important, as well as national policy regarding development of nuclear power and nonhydro renewables. Changes in these areas will affect what types and scale of new generating units are built, system efficiency, environmental performance, and what kinds of demand-side electricity efficiency programs can be feasibly deployed in China. Energy efficiency policies are unlikely to bring about great change on their own, but they will help to create attractive opportunities for energy suppliers and users to raise efficiency—and slow growth in energy use—when the economic and institutional environments permit.

Within all these areas, there are substantial opportunities for international assistance and cooperation. On the supply side, efforts to help ease China's participation in international energy markets will be important to ensuring that China has access to adequate supplies of oil and natural gas. International experience is a valuable guide to establishment of China's national gas network. For the other oil-importing countries, a key challenge will be to find ways to accommodate China's needs for access to energy from the Middle East and Central Asia. Although China still has some claim to status as a developing nation, it is increasingly a heterogeneous country, with aspects that are highly developed and others that are much less developed. Consequently, the kind of multi- and bilateral assistance that would be most valuable is different from that usually provided to developing countries. As the country's financial system evolves, direct grants and loans will become less important than efforts to develop strong domestic financial institutions and commercial links to international capital markets. Promoting development and transfer of efficient technologies could be particularly valuable, including joint precommercial research and development. Even though most exchanges affecting how particular actors obtain and use energy will occur in the commercial realm, cooperation at the policy level remains important. If the guiding philosophy behind assistance is one of helping a partner build capabilities that will serve mutual interests, rather than one of providing a handout to a poor neighbor, then the chances will be much greater that strong, peaceful, trusting, collaborative relationships will develop, relationships that

will allow countries to work together to solve global challenges.

SEE ALSO THE FOLLOWING ARTICLES

Development and Energy, Overview • European Union Energy Policy • National Energy Policy: Brazil • National Energy Policy: India • National Energy Policy: Japan • National Energy Policy: United States • Rural Energy in China

Further Reading

Andrews-Speed, P. (2001). China's energy policy in transition: Pressures and constraints. *J. Energy Lit.* **VII**(2), 3–34.

Deng, K. Y., Gu, S. H., and Liu, W. Q. (1996). Rural energy development in China. *Energy Sustain. Dev.* **III**(3), 31–36.

Horii, N., and Gu, S. H. (eds.). (2001). "Transformation of China's Energy Industries in Market Transition and Its Prospects." Institute of Developing Economies, Japan External Trade Organization, Chiba, Japan.

International Energy Agency (IEA). (2000). "China's Worldwide Quest for Energy Security." Organization for Economic Cooperation and Development, Paris.

Lieberthal, K., and Oksenberg, M. (1988). "Policy Making in China: Leaders, Structures, and Processes." Princeton Univ. Press, Princeton, New Jersey.

Lu, Y. Z. (1993). "Fueling One Billion: An Insider's Story of Chinese Energy Policy Development." The Washington Institute, Washington, D.C.

McElroy, M. B., Nielsen, C. P., and Lydon, P. (eds.). (1998). "Energizing China: Reconciling Environmental Protection and Economic Growth." Harvard Univ. Press, Cambridge, Massachusetts.

National Bureau of Statistics. (2002). "China Statistical Yearbook 2002." China Statistics Press, Beijing, China.

Sinton, J. E., Levine, M. D., and Wang, Q. Y. (1998). Energy efficiency in China: Accomplishments and challenges. *Energy Policy* **26**(11), 813–829.

Smil, V. (1988). "Energy in China's Modernization: Advances and Limitations." M. E. Sharpe, Armonk, New York.

Thomson, E. (2002). "China's Coal Industry: An Economic History." Routledge Curzon Press, London.

World Bank. (1997). "Clear Water, Blue Skies: China's Environment in the New Century." The World Bank, Washington, D.C.

Yang, F. Q., *et al.* (1995). "A Review of China's Energy Policy." Report No. LBL-35336. Lawrence Berkeley National Laboratory, Berkeley, California.

Zhao, J. M. (2001). "Reform of China's Energy Institutions and Policies: Historical Evolution and Current Challenges." Energy Technology Innovation Project, John F. Kennedy School of Government, Harvard University. Harvard Univ. Press, Cambridge, Massachusetts.

National Energy Policy: India

R. K. PACHAURI and PREETY BHANDARI
The Energy and Resources Institute (TERI)
New Delhi, India

1. **Evolution of the National Energy Policy in India**
2. **Energy for All**
3. **Ensuring Security of Energy Supplies**
4. **Improving the Efficiency of the Energy System**
5. **Reducing the Negative Environmental Impacts**
6. **Imperatives for the National Energy Policy**
7. **Implications for National Energy Policy**

Glossary

biogas A gas composed of methane and carbon dioxide; produced from the anaerobic decomposition of organic material in landfills, biogas fuel provides a medium level of energy (British thermal units); also called biomass gas.

chulhas Traditional cooking devices that use biomass as a fuel.

coal bed methane (CBM) An environmentally friendly, clean fuel with properties similar to those of natural gas. Most of CBM is in an adsorbed state on the micropores on the surface of coal.

energy intensity The amount of energy required by an economy to produce one unit of national product.

fuel cell A type of cell capable of generating an electrical current by converting the chemical energy of a fuel directly into electrical energy. Fuel cells differ from conventional electrical cells in that the active materials, such as fuel and oxygen, are not contained within the cell, but are supplied from outside.

gas hydrates Solid, crystalline, waxlike substances composed of water, methane, and usually a small amount of other gases, with the gases being trapped in the interstices of a water and ice lattice. Gas hydrates form beneath the permafrost and on the ocean floor under conditions of moderately high pressure and at temperatures near the freezing point of water.

reserve replacement ratio The amount of natural resource added in a unit of time. It is calculated as the gross addition of reserves minus the production during the time frame, taken as a ratio of existing reserves.

solar thermal collector A device designed to receive solar radiation and convert it to thermal energy. Normally, a solar thermal collector includes a frame, glazing, and an absorber, together with appropriate insulation. The heat collected by the solar thermal collector may be used immediately or stored for later use. Solar thermal collectors are used for space heating, for domestic hot water heating, and for heating swimming pools, hot tubs, or spas.

In India, several committees appointed by the government to review energy policy from time to time have emphasized the need for a nationally coordinated approach to energy policy formulation. As a result, sporadic efforts have been made to bring together various departments or ministries dealing with energy for the purpose of arriving at a unified and integrated approach. However, this has not met with much success and initial efforts have been abandoned in every case. Currently, the National Development Council (NDC) functions as an umbrella organization to approve of each Five-Year Plan as prepared by the Planning Commission. The development of each successive Five-Year Plan involves a significant effort in consultation between the Planning Commission and several ministries; the plan receives the final stamp of approval from the NDC. The aggregation of component plans, however, is not an effective substitute for a properly integrated comprehensive national energy policy. For instance, several countries worldwide have a single energy ministry that serves the requirement of integration; with the tradition established in India, such an arrangement has neither been attempted explicitly nor has it presented great appeal to political decision makers. With rapid growth of the economy and increase in demand for energy, issues of energy security, efficiency of the entire energy system, and the effects of energy production, conversion, and consumption on the environment require policies that optimize some of these variables on an integrated basis, rather than by fuel or by specific source of energy. There is, therefore, a growing

 © 2004 Elsevier Inc. All rights reserved.

imperative for India to articulate a forward-looking integrated energy policy.

1. EVOLUTION OF THE NATIONAL ENERGY POLICY IN INDIA

The energy sector in India is administered at the apex level by four ministries and one department, representing the five supply sectors, namely, coal, oil and gas, power, renewable energy, and atomic energy. Coordination and integration are achieved within the Government of India (GoI) through the Cabinet and the Planning Commission. The respective ministries do coordination and integration with state governments, with the National Development Council providing a forum at the highest level.

This structure has evolved over time and has also led to a vacuum for coordinated vision and action on a sustained basis. However, this has not always been the position. A Fuel Policy Committee that had been set up in 1974 took an integrated view of the energy requirement of the country and made recommendations for the entire energy sector. With a rapidly changing international oil scenario, a Working Group on Energy Policy was set up in 1977. Its report, submitted in 1979 (after the beginning of the second oil shock), expressed that there could be no guarantee of steady oil supply to any importing country. It emphasized that although efforts were required to conserve resources of commercial fossil fuels through efficient production and utilization, it was essential to switch to renewable resources such as agricultural and animal wastes and to new and renewable sources such as energy from the sun and wind. An important recommendation of the Working Group was that energy planning and policy formulation should be done on a full-time basis and an energy ministry, or, alternatively, an energy commission, should be in place to deal with all these matters.

In 1982, the Ministry of Energy, comprising the Departments of Coal and Power, was expanded to include the Departments of Petroleum and Nonconventional Energy Sources. Thus, except for atomic energy, all other energy departments in the GoI had been brought under one ministry. In addition, in 1983, the GoI established an Advisory Board on Energy. The functions of the board included formulating an integrated energy policy covering commercial and noncommercial sources of energy. Its functions were advisory in nature and it was to submit its reports directly to the Prime Minister (Table I). Since then, this structure has become more decentralized, with all the energy subsectors falling under separate ministries or departments. It is only at the time of the formulation of Five-Year Plans (FYPs) that the sector is looked at as a whole. A brief review of the various pronouncements related to the energy sector in the FYPs is summarized in Table II.

TABLE I

Evolution of National Energy Policy

Major milestone	Impact/function
Fuel Policy Committee (1974)	Looked at energy sector as a whole
Working Group on Energy Policy (1977)	Report submitted in the aftermath of the second oil shock expressed apprehensions regarding security of oil supplies and thus recommended conservation of energy and encouragement of renewable energy
Expansion of Ministry of Energy to include Departments of Petroleum and Nonconventional Energy Sources (1982)	Except for atomic energy, all other forms of energy were brought under one ministry
Advisory Board on Energy (1983)	Set up to formulate an integrated energy policy covering commercial and noncommercial energy resources
Exploration rounds	The exploration rounds to attract the private sector were introduced in 1979, but not much headway was made till 1995
Committee on Integrated Coal Policy (1996)	Recommended adoption of coal conservation measures, inviting private capital in the sector, deregulating the coal and coal product prices, and setting up a regulatory body, among other things
Common Minimum National Action Plan for Power (1996)	Initiated the reforms in the power sector with trifurcation of state electricity boards, setting up of state and union level regulatory commissions, and rationalization of tariffs
New Exploration and Licensing Policy (NELP) (1997)	NELP was introduced in 1997 with modified terms and conditions; three rounds have been held with moderate success
Administered Pricing Mechanism (APM) dismantling (1997)	Phase-wise dismantling of APM started in 1997; APM was completely abolished early in 2003

Broadly speaking, the overall objective of a National Energy Policy (NEP) should be to meet the energy needs of all segments of population in the most efficient and cost-effective manner while ensuring long-term sustainability. This objective entails the following major tasks:

- Providing clean and affordable energy to all.
- Ensuring security of the energy supply.
- Improving the efficiency of the energy system.
- Reducing the adverse environmental impacts of energy use.

The following discussions review the Indian experience in light of these objectives.

TABLE II

Five-Year Plans[a]

Plan	Goal
First FYP (1951–1956)	Set out the objectives of national planning with immediate focus on agricultural and industrial development and attendant development of irrigation and power
Second FYP (1956–1961)	Emphasized the long-term perspective while planning for short-term horizons
Third FYP (1961–1966)	Elaborated the need for a comprehensive review of the extent of available information and data on India's natural resources
Sixth FYP (1980–1985)	Aimed at a decentralized approach to increasing availability of renewables to meet rural energy requirements; thrust was on management of oil demand, conservation, and renewable resources; contained the first oblique reference to the environmental impact of energy use
Seventh FYP (1985–1990)	Recognized that energy planning involves not only potential increase: in indigenous availability of energy, but also better utilization, because the trends in India's commercial energy consumption show high rates of growth and because of the growing share of oil dependency
Eighth FYP (1992/1993–1996/1997)	The environmental dimension was given due significance only in this plan, which emphasized the long-term need for promotion of technologies of production and consumption of energy that are environmentally benign and cost-efficient
Ninth (1997–2002) and Tenth FYP (2002–2007)	These plans placed special emphasis on the need to introduce reforms in the energy sector in order to improve efficiency in the sector and to enhance infrastructure development in view of the impending growth in the sector in the long term. Due importance has also been given to the pricing issues in the sector and the need to make prices indicative of the market. Concern was also raised regarding the direct and indirect subsidies present in the sector and the need to streamline them and make the process more focused. Energy conservation is to be promoted in the plan periods mainly through improvements in the industrial sector

[a]Data are from each plan, as specified by the Indian government.

2. *ENERGY FOR ALL*

2.1 Present Scenario in Energy Consumption

Total primary energy consumption in India has increased nearly fivefold in the three decades from 1970 to 2001. The total primary energy consumption stood at 437.7 million tonnes of oil equivalent (MTOE) in 2001/2002 (GoI 2002). The share of commercial energy in total primary energy consumed rose to more than 68% in 2001, up from 28% in 1950 and, as the economy develops, this share is expected to rise even further. However, the per capita energy consumption in India is still 479 kilograms of oil equivalent (kgOE), which indicates a tremendous scope for even higher energy consumption in the coming years. Additionally, the 5.4% annual growth rate of energy consumption achieved by India, over the period 1970–2001, is the 14th highest in the world.

Figure 1 shows the comparative position of India with respect to per capita income and per capita energy consumption. Countries with high per capita income are also characterized by high per

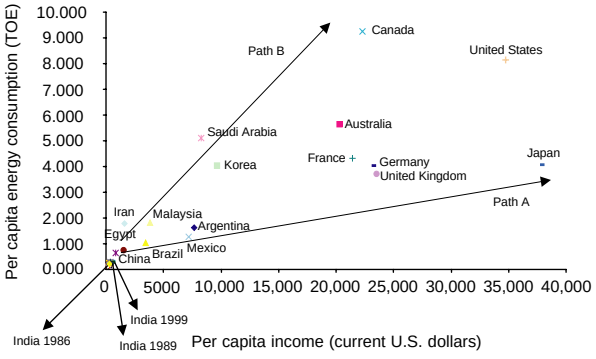


FIGURE 1 Comparison of per capita energy consumption and income in India and other countries. The gross domestic product and population data are from the World Bank; energy consumption data are from *BP Statistical Review of World Energy (2002)*.

capita energy consumption. This is not to suggest that India should be looking toward energy consumption levels of the West, as it marches on its path of development, because that would simply be unsustainable. In fact, it does indicate that there is tremendous scope for increase in energy consumption, admitting, though, that the link between energy and economy will gradually become weak as the economy matures and moves more toward the services sector. This is shown in Table III, which gives the trend in elasticity of consumption of various fuels with reference to the gross domestic product (GDP) over time. From Fig. 1, it is evident that Japan is more efficient than the United States because its energy consumption is much lower than that of the United States for the same level of per capita income. Accordingly, developing countries could aim to follow path A for development in order to build a sustainable economy. Development along path B will result in a high, potentially fatal dependence on energy for countries like India, which are dependent on imports for meeting their energy needs. Figure 2 shows the path that India has taken for the past two decades. As shown, since 1990, the per capita energy consumption in India has not risen commensurately with its per capita income, which indicates that it is possible to reduce energy intensity and/or improve energy efficiency in the economy while still enabling economic development.

The elasticity of consumption of various fuels with respect to the GDP has declined over the years, with current commercial energy elasticity lower than unity; this means that a 1% increase in the GDP will increase the commercial energy consumption by less than one. However, underprivileged sections of Indian society do not have access to efficient sources of energy. The World Bank estimates indicate that, of the 2 billion people who have no access to modern energy sources such as electricity and liquefied petroleum gas (LPG), about 700 million live in India. Within India, there is a big rural–urban divide in terms of access to modern energy sources (Figs. 3 and 4). Access to modern fuels can be improved by short-term government measures such as subsidies as was done by the introduction of Public Distribution Systems to make available kerosene to the underprivileged section. However, these measures have not proved to be sustainable in the long term. State interventions in the form of programs for poverty alleviation are important to enable continued use of modern resources even after subsidies are phased out.

TABLE III

Elasticity of Consumption of Commercial Energy with Respect to GDP[a]

Period	Coal	Oil	Electricity	Total commercial energy
1953/1954–1960/1961	1.10	2.14	3.02	1.37
1960/1961–1970/1971	0.53	2.31	3.04	1.16
1970/1971–1980/1981	0.98	1.83	2.06	1.37
1980/1981–1990/1991	0.31	1.12	1.57	0.82

[a]Data from Sengupta (1993).

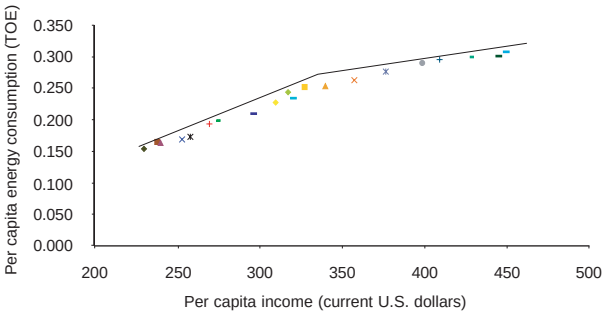


FIGURE 2 Trend in Indian per capita income and energy consumption (countries are represented by symbols as in Fig. 1). The gross domestic product data are from *Economic Survey 2001/2002*; population data are from *Provisional Census (2000/2001)*; energy consumption data are from *BP Statistical Review of World Energy (2002)*.

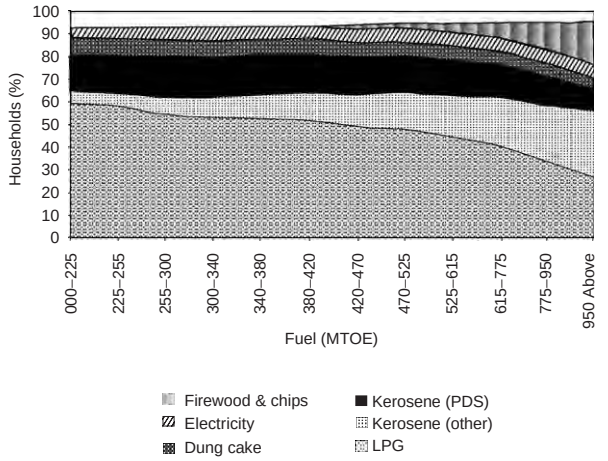


FIGURE 3 Fuels used by rural households (percentage). PDS kerosene is kerosene sold through the "fair price" shops under the Public Distribution System. Data are for 2001, from the National Sample Survey Organization.

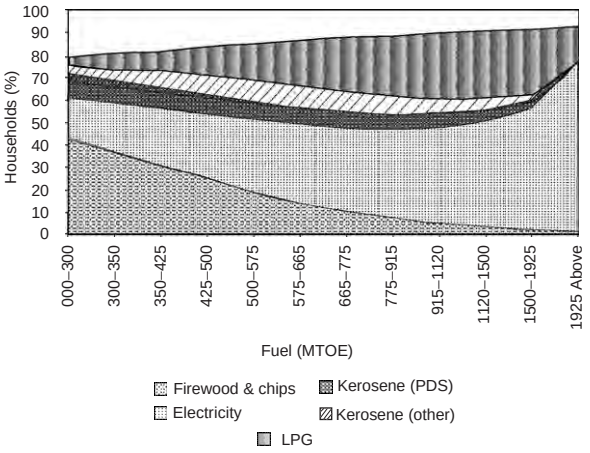


FIGURE 4 Fuels used by urban households (percentage). PDS kerosene is subsidized kerosene sold through the "fair price" shops under the Public Distribution System. Data are for 2001, from the National Sample Survey Organization.

Figs. 3 and 4 show that the percentage of electricity usage among all classes of urban households is quite high, as is the consumption of LPG. However, the dominant fuel among the rural households is firewood, which is both inefficient and leads to indoor air pollution. Thus, promoting fuels such as electricity and LPG would not only promote efficiency but would also help curb respiratory problems among rural women.

2.2 Energy–Economy Linkage

Energy is the key driver of many sectors of the economy, namely, agriculture, industry, commerce, services, and domestic. With advances in technology and the growing need for higher productivity, most of these sectors are also becoming more energy intensive. This is especially true in developing countries, which show a strong energy–economy linkage in the form of high-energy intensity/elasticity of the economy. The developed countries have been able to weaken this linkage due to development of more efficient processes that can do the same amount of work with less energy, or more work with the same amount of energy. However, developing countries—most of them highly industrialized—are still highly dependent on energy.

The picture is not different for India. Though India now boasts of a big services sector, accounting for 49% of its GDP, the energy intensity of Indian economy is still high. Figure 5, which shows the trends in the GDP and the primary energy consumption in India since 1970, indicates a strong linkage

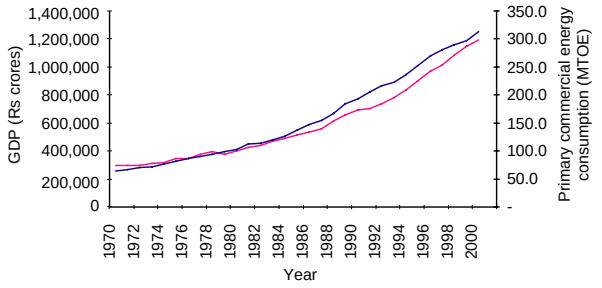


FIGURE 5 Comparative trends in gross domestic product (GDP) and total commercial energy consumption. The GDP is expressed in crores (=10 million) of rupees (Rs). The gross domestic product data are from *Economic Survey 2002/2003;* energy consumption data are from *BP Statistical Review of World Energy (2003).*

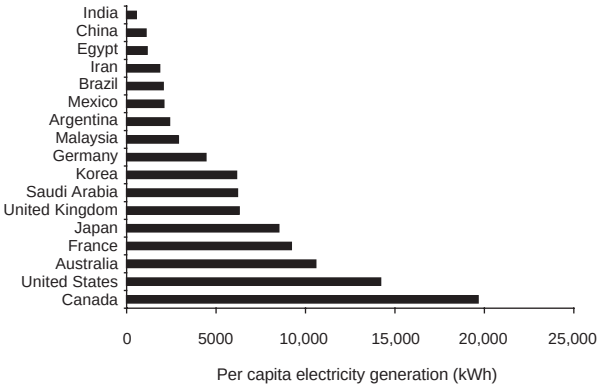


FIGURE 6 Comparison among countries with respect to per capita electricity generation (in kWh, kilowatt-hours). Electricity data are from the International Energy Agency; population figures are from the World Bank.

between energy and economy. However, the years since 1985 have witnessed a gradual distancing between the two curves that is indicative of a growing share of the services sector, which is less energy intensive as compared to the industrial sector. Per capita electricity generation and oil consumption levels in India are among the lowest in the world, lower even than those of its neighbors, as shown in Figs. 6 and 7.

3. ENSURING SECURITY OF ENERGY SUPPLIES

The issue of ensuring affordable supplies of energy in various required forms is central to India's socio-economic development. As already mentioned, the link between energy and economy in India is still strong and any disruption in energy supplies is bound

to have negative impact on the economy. The important issue is to assess both the vulnerability of the economy to an external supply shock and the measures that can be taken to cushion the economy against such a shock.

With the entry of the private sector, market forces in the energy sector have added a new dimension to the whole issue of energy security. Market forces will ensure that, during crises, demand and supply will always balance. Prices will indicate the relative shortage and surplus in the sector. Thus, the impact of external shock will be felt in prices rather than in demand and supply.

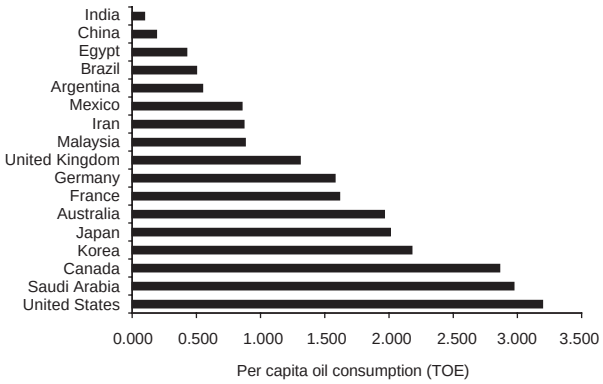


FIGURE 7 Comparison among countries with respect to per capita oil consumption (in TOE, tonnes of oil equivalent). Oil consumption data are from *BP Statistical Review of World Energy (2003)*; population figures are from the World Bank.

3.1 India's Fuel Mix

India's energy mix has not undergone much change in the past 30 years. Coal still remains the dominant fuel; 55% of the total primary energy is accounted for by coal, down from 58% in 1970, though the consumption of coal has grown by 5.21% annually. The share of oil has increased marginally from 30% in 1970 to 31% in 2001, though consumption has grown at 5.5% annually since 1970. Gas has witnessed major gain, increasing by a share of 8%, which is up from virtually nothing in 1970, with gas consumption growing at 13.73% annually. However, because of domestic discoveries, it is only since the 1980s that gas has really seen major improvement. The share of hydropower has declined, although it displayed an annual growth rate of 2.38% over the past 30 years. This is primarily due to high growth rates achieved by oil and gas (Table IV).

The high percentage of oil consumption in the economy, along with high oil elasticity, points toward increasing dependence of Indian economy on oil as a fuel. This dependence becomes more

TABLE IV

India's Fuel Mix[a]

	Mix per year (in MTOE)					
Fuel	**1953/1954**	**1960/1961**	**1970/1971**	**1980/1981**	**1990/1991**	**2001/2002**
Commercial primary energy						
Coal	23.62	35.64	36.48	56.96	94.68	133.89
Lignite	—	0.01	0.81	1.23	3.34	6.52
Crude oil	0.19	0.46	7.01	10.79	33.92	32.03
Natural gas	—	—	0.6	1.41	11.73	26.72
Hydropower	0.24	0.67	2.17	4	6.16	6.37
Nuclear power	—	—	0.63	0.78	1.6	5.15
Wind power	—	—	—	—	—	0.14
Total	24.05	36.78	47.7	75.17	151.43	210.82
Net imports (+)	2.2	6.04	12.66	24.63	31.69	87.85
Stock changes (−)	0.24	2.87	0.69	3.8	5.37	—
International bunkers (−)	0.53	0.5	0.24	0.21	0.14	—
Total commercial energy supply	25.48	39.45	59.43	95.79	177.61	298.67
Noncommercial primary energy supply	64.13	74.38	86.72	108.48	122.07	139.02
Total primary energy supply	89.61	113.83	146.15	204.27	299.68	437.69

[a] Data from the Tenth Five-Year Plan, developed in 2002 by the Planning Commission.

threatening if it is seen in conjunction with the meager quantity of domestic crude oil production and, consequently, the high level of dependence on imported oil. However, the economy's dependence on oil is much more than is suggested by the share of 31%. This is because of the heavy dependence of the transport sector on oil. The avenues for fuel substitution in the transport sector are limited, notwithstanding recent endeavors to use compressed natural gas (CNG) in lieu of gas oil and gasoline. Initiatives to increase use of solar-powered vehicles and fuel cell vehicles are still many years from impacting the share of oil use by transport sector. Certain reform strategies to improve energy efficiency, however, can be implemented with immediate effect in the transport sector:

- Measures to improve the efficiency of oil use in vehicles (this would also improve the already degraded urban environment).
- Taxes on fuel consumption to signal importance of fuel conservation.
- Strict adherence to vehicular fuel-efficiency parameters.
- Policy interventions to encourage faster turnover of old vehicles.
- Incentives to attract foreign players into the market to create competition, promoting the latest and most efficient technologies.
- Improvements in highways and internal city and town roads and promotion of foreign direct investment (FDI).
- Strengthening of public transport systems to reduce the use of personal vehicles.
- Policy options and improvements in railways to promote cargo movement and to reduce the pressure on roadways.
- Policy options and rapid work on the megaproject to link rivers across the country, to reduce the load on road transportation.

The issue of energy security assumes even greater importance when future energy demand and supply scenarios are examined.

3.2 Future Energy Demand

The preceding discussion highlighted the low per capita income and high per capita energy consumption that characterize the Indian economy. These trends indicate the possibility of even larger energy consumption in the coming years. The *International Energy Outlook 2003*, published by the U.S. Department of Energy, projects India's energy consumption

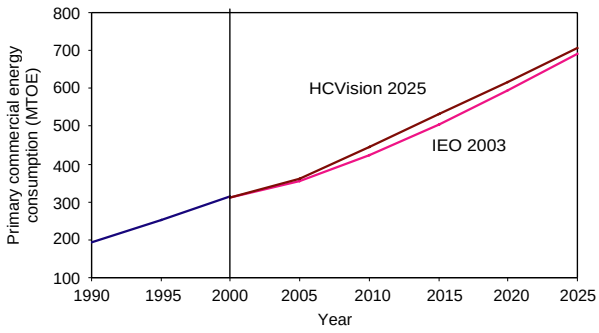


FIGURE 8 Past and projected energy consumption trend. MTOE, Million tonnes of oil equivalent. Data are from *Hydrocarbon Vision 2025* (HCVision) and *International Energy Outlook 2003* (IEO); additional data are from *TERI Energy Data Directory and Yearbook* for years 1999–2000.

at 690 MTOE by 2025, up from 322 MTOE in 2001, and an average annual percentage change of 3.2% during the period 2001–2025. Based on the population estimates given by the World Bank, the per capita energy consumption works out to be 529 kgOE, which is 10% higher than the current level. The energy intensity at this level of energy consumption, assuming a GDP growth rate of 6%, works out to be 13.67 kgOE per 1000 rupees, which indicates a decline of 43% over the 22-year period (2003–2025). This level of consumption will increase India's share in total world primary energy consumption to 4.27%, up from the current level of 3.1%.

Hydrocarbon Vision 2025, the long-term energy policy document published by the GoI, also projects energy demand and fuel mix for the period 2002–2025. It pegs India's total primary energy consumption at 616 MTOE in 2020, of which 238 MTOE will be derived from oil, representing 38% of the total; gas will contribute 16%; coal, 38%; hydro, 6%; and the rest will be met by nuclear power. These projections are not very different from the *International Energy Outlook 2003* projections, as shown in Fig. 8. The resultant change in fuel mix is shown in Fig. 9. The share of coal is projected to decrease, from 55% in 2001 to 38% in 2020 (Fig. 9), whereas that of natural gas is projected to double to 16%. The share of oil is projected to rise to 38%. This fuel mix indicates a gradual shift toward cleaner fuels (natural gas and oil) in the wake of rising environmental concerns. On the other hand, the projection also indicates a rise in dependence of the economy on imports, because domestic reserves of oil and gas cannot support such high consumption levels under the most optimistic scenario.

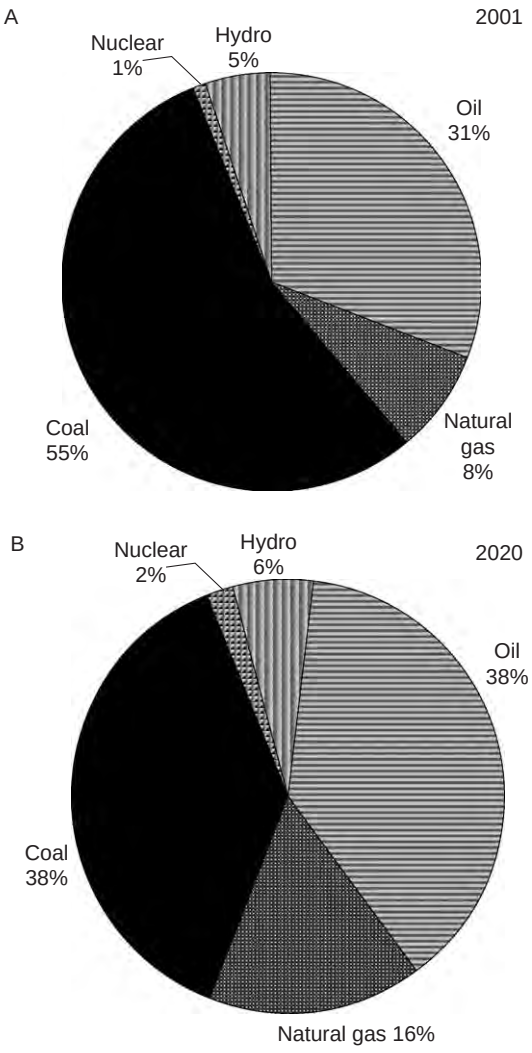


FIGURE 9 Current (A) and projected (B) commercial energy mix. Data are from *Hydrocarbon Vision 2025*.

3.3 Present Resource Availability and Future Outlook

Given that India is a huge consumer of energy, supply options are very limited. Already, India imports around 74% of its crude oil requirement and, though India has vast reserves of coal, coal imports are on the rise due to the low quality of native coal and the faulty taxation structure that makes coal imports cheaper for some consumers located near the coast. Development in the nuclear energy sector has increased in recent years, but it is unlikely that this sector will play a role in the development of Indian economy comparable to the role it played in France. India has a huge potential for hydropower, especially small/micro hydro/electric (hydel) power projects, but these opportunities are concentrated in a few states. This makes it imperative to improve the efficiency of the transmission system so that surplus power in one state can be reliably transferred to a deficit state. To achieve this, reforms in the power sector are essential. Table V shows the level of reserves in coal, oil, and natural gas in the country along with the reserve replacement ratio. The reserve replacement ratio for crude oil, though positive, is still low and the same goes for gas. India's position in coal seems to be comfortable but, as is discussed later, there are some environmental concerns.

TABLE V

Coal, Oil, and Natural Gas Reserves[a]

Fuel	Reserves	Current production	Reserve replacement ratio
Crude oil (million tonnes)	644.8	32.45	0.59
Natural gas (billion cubic meters)	647.5	26.57	0.33
Coal (million tonnes)	213,905.5	322.2	8.46

[a]Data from TEDDY 2002/03, see TERI (2003).

Efforts such as the New Exploration Licensing Policy (NELP) are being made to increase the indigenous production of crude oil, but the prospects for discovery of another oil field equivalent to the Bombay High field (discovered in the late 1970s) remain remote. Availability of gas in the form of liquefied natural gas (LNG) is slated to improve in the coming years, but there are many policy issues that remain to be sorted out before gas can be delivered cheaply in the country. Several pipeline projects are also planned, but political risks seem insurmountable. These issues are discussed in detail in the following sections. The locations of various coal, lignite, oil, and gas fields are shown in Fig. 10.

3.3.1 Crude Oil Supply Outlook

It has been almost two decades since Bombay High was discovered. Given the fact that consumption is around 100 million tonnes (MT) of petroleum products every year and production is only 33 MT, four more Bombay Highs would be required for India to become self-reliant in oil at present. As oil consumption rises in the future, discovering even more domestic reserves becomes imperative with

FIGURE 10 Crude oil, natural gas, and coal/lignite reserves in India. Data are from The Energy and Resources Institute.

respect to security in both oil supplies and prices. This appears far-fetched, given the lackluster response to the NELP by international oil companies, which were expected to bring in both technology and capital. Though the Oil and Natural Gas Corporation Ltd. (ONGCL) is planning to double its reserves in 20 years, the oil demand by then would also have more than doubled, leaving the import dependency unchanged. The Planning Commission has projected domestic production of only 45 MT by 2011/2012 against the projected demand for petroleum products ranging from 187 to 160 MT. This indicates high import dependency with all its repercussions on energy security and the foreign exchange reserves. The low prospect of discoveries in Indian sedimentary basins, due to poor exploration methodology, is an indication of the need for diversification of energy sources and application of effective and efficient exploration techniques to discover new oil and gas basins in India.

3.3.2 Electricity Supply Outlook

Given the high positive elasticity of electric power with respect to the GDP, reliable electricity supply is a necessary ingredient for economic development. The current installed capacity in India is 105,083 megawatts (MW). However, there is a 10.5% energy shortage and a 15.9% peak shortage. The high transmission and distribution (T/D) losses mean that even the peak demand of 78,841 MW, which is 26,242 MW less than the installed capacity, is not being met. To meet the electricity requirement, the National Thermal Power Corporation (NTPC) is planning to add 20,000 MW of additional capacity by 2020; but these plans are largely dependent on the availability of LNG at reasonable prices. The Central

Electricity Authority (CEA) projects an addition of 47,000 MW during the period 2002–2007, which seems highly unlikely, given that India has been able to add only about 20,000 MW per plan period previously. The private sector was expected to play an important role in improving the power situation in the country, but the slow pace of reforms has discouraged investors, and many independent power producers (IPPs) have thus abandoned their plans for new plants. As indicated previously, to realize the full potential of hydropower, substantial investments in T/D are required.

The government has targeted a nuclear-based capacity of 20,335 MW by 2020. Considering the fact that current nuclear-based capacity is only 2720 MW, this also seems a bit ambitious. This is not only because of high initial investment costs, but also due to many other issues concerning nuclear power, such as long development period, reactor safety, waste disposal, and the continuance of government support. Thus, in the present scenario, nuclear energy is not expected to play the same important role that it played in France. The important aspect that emerges is that interventions are required to expedite the pace of reforms, attract IPPs back into the energy market in India, and create investment opportunities in T/D of power grids to tap the high prospect of hydropower.

3.3.3 Gas Supply Outlook

Gas is slated to play an important role in India's energy mix. As per *Hydrocarbon Vision 2025*, the share of gas in total energy consumption is projected to be 16%, up from 5% at present. However, projections for indigenous gas production imply that much of this gas is likely to be imported, primarily in the form of LNG. The Planning Commission has projected domestic production of 105 million standard cubic meters per day (MSCMD) by 2011/2012; the demand projected by *Hydrocarbon Vision 2025* ranges from 216 to 313 MSCMD in the year 2011/2012.

The recent discovery of 10.5 trillion cubic feet (TCF) of natural gas by Reliance in the Krishna–Godavari Basin is expected to boost the current gas production by 50% in 2–3 years. This discovery, in the NELP 1 block, has led to an improved outlook on indigenous gas production, and many plans for importing natural gas are being reviewed. However, this gas is concentrated in southeastern India and the major demand centers are in northern and western India. Whether this gas will flow to these demand centers remains to be seen, but in the present scenario, the dependence on imported gas will continue. Several proposed LNG projects will serve to fill this gap, but progress is slow due to various issues concerning the reforms in the oil and gas sector in the country.

Several pipeline projects are also planned to reduce the demand–supply mismatch in the country, including the Indo–Iran pipeline, the Indo–Oman pipeline, and the Indo–Bangladesh pipeline. The security concerns surrounding the Indo–Iran pipeline are given great weight by the policymakers, rendering this project, too, a nonstarter. The Indo–Oman pipeline was officially declared dead when the feasibility report admitted the lack of technology for deepwater pipelines. Bangladesh has recently hinted at limited gas exports to India, but the issue is still under debate.

The GoI has also framed policies for the development of nonconventional gas resources such as coal-bed methane (CBM) and gas hydrates. India has 850 billion cubic meters (BCM) of assessed CBM resources and 6150 trillion cubic meters (TCM) of assessed gas hydrate reserves. Private sector players such as Reliance and Essar have evinced interest in CBM, and the National Gas Hydrates Program is underway to develop gas hydrates. All these initiatives to strengthen the nonconventional gas resources are steps in the right direction and will boost India's energy security.

4. IMPROVING THE EFFICIENCY OF THE ENERGY SYSTEM

Figure 11 shows energy intensity (kilograms of oil equivalent per 1000 rupees) and energy consumption

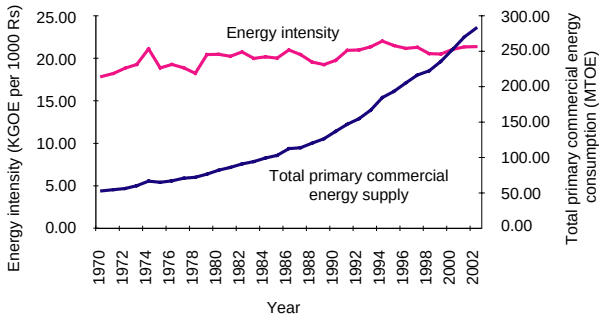


FIGURE 11 Trend in commercial energy intensity and total commercial energy consumption (in kgOE, kilograms of oil equivalent); Rs, rupees; MTOE, million tonnes of oil equivalent. Total commercial energy consumption data are from *BP Statistical Review of World Energy (2003)*; gross domestic product data are from *Economic Survey 2002/2003*.

over several decades. The rate of growth of commercial energy intensity of the economy has been slower than that of the total commercial energy consumption. This decline is due to the high growth in the services sector and/or improvements in the energy efficiency of the economy.

Sectoral trends in energy intensity also yield important insights into the country's energy efficiency. The energy intensity of the industrial sector has declined since the 1980s, primarily due to a decline in the oil intensity of the industry, although intensity of gas usage has increased. Presently, gas demand in the country is constrained due to deficiency in supply, even though gas is more efficient than oil in industrial applications. Energy intensity of the agriculture sector has risen, primarily due to a huge rise in electrical energy intensity of the sector. This trend is also expected to continue as Indian agriculture tries to rely more on man-made irrigation options than on monsoons. Also, new-age farming processes require measured irrigation at the right time to yield desirable results. This will further necessitate the use of pump sets in irrigation.

Energy intensity of transport (excluding railways) rose till 1989/1990, when it reached its peak, and declined thereafter. This is probably due to the introduction of more fuel-efficient vehicles following the liberalization of the economy in 1990/1991. Improving efficiency in the sector would entail introducing market-based reforms. Some progress has been made along these lines (a detailed analysis of the status of reforms to date is discussed later). According to estimates of the United Nations Development Program (Table VI), there is still room for improvement in energy efficiency in various sectors of the economy.

TABLE VI

Potential for Improvement in Energy Efficiency[a]

Sector	Energy efficiency (%)
Industry	
Iron and steel	15
Cement	17
Pulp and paper	20–25
Textile	23
Aluminium	15–20
Household	
Lighting	10–70
Refrigeration	25
Air-conditioning	10
Agriculture	
Pump sets	25–55
Transportation	
Cars	7.5–10.0
Trains (diesel)	5–10
Trains (electric)	5–10

[a] Data from United Nations Development Program (2000).

5. REDUCING THE NEGATIVE ENVIRONMENTAL IMPACTS

5.1 Effects of Conventional Energy on the Environment

Being a developing country, India does not have any commitments for reduction of CO_2 (carbon dioxide) emissions, though India has ratified the United Nations Framework Convention on Climate Change (UNFCCC) and is party to several international environmental treaties (e.g., the Montreal Protocol). However, the noncommitment on emissions should not lead India to a pursue development that is unsustainable and has adverse impacts on the environment. Through the concept of sustainable development, a long-term vision that can bind various policy objectives of each subsector of the economy should be developed. Reducing environmental distress should undoubtedly be among the main policy objectives of an integrated energy policy.

India's energy-related carbon emissions have grown ninefold over the past four decades. With 162 MT of carbon released from consumption and burning of fossil fuels in 2001, India ranked fourth in the world. Coal burned in electricity generation accounts for 64% of all carbon emissions from coal. Technology improvements and diversification of fuels for the power sector are an utmost priority. In a business-as-usual scenario, according to *International Energy Outlook 2003* projections, CO_2 emissions from the power sector will rise to 261 MT of carbon equivalent by 2020. The average annual percentage change in CO_2 emissions from 1990 to 2020 will be 2.2%. India should gear up now, because it may be called on to cut emissions in the second commitment period (2012–2018). Given the long lives of power plants, these issues have to be kept in mind while planning national electricity capacity. Due emphasis should be given to research and development to implement effective and efficient technology for future emissions reductions.

India's carbon intensity is also high, primarily due to low efficiency of coal-based generating plants. Taxation policies that make domestic coal competi-

tive with cleaner imported coal are also responsible for the high carbon intensity of the economy. An integrated energy plan would consider the effects of such energy policies on the already degraded environment of the country. In sum, analysis easily identifies the key areas—decentralized grass-roots initiatives and production and usage of renewable energy—for policy interventions for sustainable development and for reducing adverse environmental impacts.

5.2 Potential for Renewable Energy

The share of hydrocarbons in India's commercial energy mix is slated to rise to 54% in 2020 from the current level of 38% (see Fig. 9). Given the paucity of hydrocarbons in India, it is important to develop a renewable energy base that would reduce import dependence. Table VII shows the estimated potential for renewable energy in the country. Comparison of the potential with achievement paints a bleak picture. Although India ranks first in the world in utilization levels of solar cookers and biomass gasifiers, and can claim decent utilization levels in biogas plants and cookstoves, in solar photovoltaic energy, and in wind power, the potential to exploit these resources still remains unfulfilled. One major deterrent to the utilization of renewable energy is the high utilization cost of these technologies. However, these costs are slated to decline in the future, which will make the technologies economically competitive. Solar photovoltaic (PV) energy is an important alternative for power generation, and specific intervention is required to promote its use. The annual global market for PV energy is estimated to be around 200 MW and has grown at a rate of 15% over the past 15 years. In India, the annual PV production volumes are of the order of 10 MW, making India one of the largest markets and manufacturers of PV modules in the world after the United States, Japan, and the European Union.

Very few power-generating renewable energy sources are currently connected to the grid in India. As energy from renewables becomes more and more competitive, this aspect will automatically cease. Renewable energy can be used not only to generate power, but also for cooking, space heating, water desalination, etc. All these uses suggest that rural India represents a vast market for renewable energy; with provision of energy becoming more and more decentralized, such options will be explored more thoroughly. In fact, the Ministry of Nonconventional Energy Sources has targeted 5750 remote villages for electrification through the use of renewable energy by 2012.

TABLE VII

Potential and Achievement in Renewable Energy[a]

Energy source	Potential	Achievement[b]
Biogas plants	120 lakh[c]	32.75 lakh
Improved chulhas[d]	1200 lakh	338 lakh
Wind	45 000 MW	1507 MW
Small hydro	15 000 MW	1423 MW
Biomass power/ cogeneration	19 500 MW	358 MW
Biomass gasification		42.8 MW
Solar photovoltaic	20 MW per km^2	82 MW per km^2
Waste-to-energy fuels	1700 MW	17.1 MW
Solar water heating	1400-lakh m^2 collector area	6-lakh km^2 collector area

[a] Data from Ministry of Nonconventional Energy Sources (2002).

[b] As of 31 December 2001.

[c] Lakh is a term denoting 100,000; it is used here to indicate efficiency.

[d] A chulha is a fuel-burning stove; chulhas are made of many materials and have shapes and components that make them more or less efficient.

6. IMPERATIVES FOR THE NATIONAL ENERGY POLICY

Energy demand in India is projected to increase in the coming decades, as discussed earlier. Though efforts are being made to ensure reliable supply of energy at reasonable prices, entry of market forces due to the globalization and privatization process will play an important role in the process. Ensuring the reliable supply of energy, in required forms and at reasonable rates, has become one of the priorities of governments the world over, and this holds true especially for energy-deficient countries such as India. However, market forces are increasingly encroaching on this traditional domain of state influence. This essentially means that though demand and supply will always match, prices will reflect the shortage or surplus in the economy. Hence, energy shortages are likely to translate into higher prices. Because there is a strong energy–economy linkage, this will slow down the pace of development and depress the economy. Thus, the

National Energy Policy should aim first to weaken the energy–economy linkage, by improving the efficiency of energy use, and then to ensure that dependence on any one source is reduced effectively, thereby reducing the vulnerability of the economy to external shocks. The various policy issues that are critical to ensuring these aims are discussed in the following sections.

6.1 Role of Market Forces

Traditionally, energy has been one of the main domains of state influence in which the private sector had been excluded from effective participation. This is evident in developing countries, in which scarce resources are utilized in optimal ways in the interest of the entire nation. At the same time, the developed world has moved toward market forces. In the recent past, market forces have also begun to make their presence felt in increasing numbers of developing countries (India, Brazil, Mexico, and several East Asian countries). It is being gradually realized that the market is the best allocator of resources and that the government should at best be a facilitator and regulator that sends signals in the form of incentives to the market, to orient market objectives toward national aims. In India, entry of market forces in each sector of the economy was planned and allowed when reforms were launched in 1990/1991.

The entry of private players in the energy sector means that prices will become important tools for signaling shortage or surplus in the market. In the controlled era, quantitative restrictions played a major role in balancing demand and supply. This, however, created a lopsided consumption pattern and efficiency was put on the back burner. Industries, assured of fixed returns, paid no attention to the efficiency of the process, which increased the energy intensity of the economy. Wasteful spending by consumers who could afford it led to shortages in the market, which then forced the government to resort to quantitative restrictions in order to supply energy to those who could not afford it. Meanwhile, the import bill for the country continued rising, given the importance of oil in the energy mix.

With prices now becoming important tools, such tendencies are likely to be curbed. This is already showing in the improving energy intensity of the economy, which is becoming more and more competitive. For example, the prices of petroleum products have been decontrolled and are now set by oil companies fortnightly, based on international price of crude. This gives the consumer strong signals about the real value of these products and, hence, is likely to have an impact on the consumption pattern and fuel economy of vehicles. Such a rise of market forces, however, raises the question of regulation of the market by the government. Markets are nowhere near perfect; to prevent market abuse, regulation has to play an important role. Thus, regulating the energy market is one of the major tasks for the NEP.

Pricing policies adopted by the government so far have resulted in the largest state-run power-generating company in the country, the National Thermal Power Corporation, which holds the decision to use gas for power plants (Anta, Auraiya, Kawas, and Kayamkulam) in abeyance due to uncertainty over final delivered gas price. In the absence of an integrated LNG policy, the tax status of LNG remains uncertain and, hence, the cost of delivered LNG differs from state to state. On the other hand, an increasing number of studies are proving that gas will be competitive with coal, at least in the western region of the country. This uncertainty, in turn, affects the investment in the infrastructure sector of the country, which is crucial for development.

6.2 Status of Reforms

Reforms were launched in India almost a decade ago. So far, the scorecard has been mixed. In some sectors, reforms have seen greater consumer choice, availability of latest technologies, and reduction in prices due to increased competition. However, reforms in the energy sector have not been adequately effective, because this sector requires an integrated approach involving all the sectors that are directly or indirectly linked with it.

6.2.1 The Energy Conservation Act (2001)

The Energy Conservation Act (2001) was enacted in September 2001 to deal with all matters related to the efficient use of energy and energy conservation. Under the act, the Bureau of Energy Efficiency was set up to discharge the following functions:

- To look into the energy consumption norms for each energy-intensive industry and encourage proper labeling of energy consumption indicators on every electrical appliance.
- To provide guidelines for energy conservation building codes.

- To take measures to create awareness and disseminate information for efficient use and conservation of energy.
- To strengthen consultancy services in the field of energy conservation, develop testing and certification procedures, and promote testing facilities for certification and energy consumption labeling of equipment and appliances.
- To provide financing for certain measures taken by consumers to enhance efficiency in energy consumption.
- To maintain a list of accredited energy auditors to carry out energy audits of industries and recommend measures for improving efficiency.

The responsibility of the central and the state governments under the Energy Conservation Act is to ensure effective and efficient implementation of the suggestions given by the Bureau of Energy Efficiency, for which a system of penalties and incentives has been devised. On the whole, the enactment signifies the importance that the government has accorded to efficiency improvements in all sectors.

6.2.2 Oil and Gas Sector Reforms

In 1997, the government had established a timetable for abolishing the Administered Pricing Mechanism (APM) by 2002 and for deregulating the entire oil and gas sector. Though the APM has been abolished, it is not certain that the oil and gas sector has been freed of government diktats. Though the pricing of sensitive fuels (motor fuels, high-speed diesel fuel, LPG, and superior kerosene oil) is supposed to be market determined, prices are still set after consulting the government.

In order to prevent exploitation of consumers in the deregulated scenario, the Petroleum Regulatory Board Bill was introduced to ensure that each marketing company displays the maximum retail prices for the notified petroleum products and take steps in accordance with regulations to prevent profiteering by the entities. However, the bill has not been passed and has been referred back to the Parliamentary Standing Committee. Therefore, the regulator is still not in place and gas prices are not yet completely linked to international parity. The ONGCL still negotiates with the refineries for an internationally competitive crude price.

LNG importers have been demanding granting of infrastructure status for LNG that would make it eligible for a central sales tax of only 4%. Because the sales tax rates vary widely from state to state, LNG is likely to be costlier than coal, which jeopardizes investments being made by companies in building LNG terminals. To deal with this, the government has introduced the draft Integrated LNG Policy, which focuses on the various tax sops given to the project developers to make LNG viable in the country. It is expected that the delivered LNG will be 22% cheaper under the new policy and will bring it on par with 100% fuel oil parity price of the domestic gas. The government has also announced a new Petroleum Product Pipeline Policy on a common carrier principle. The new guidelines for grant of right of use (ROU) on land do not contemplate any restrictions or conditions for grant of ROU for crude oil pipelines. As per the guidelines, the investors will have complete freedom with respect to the pipelines originating from refineries or meant for captive use by companies for which ROU will be unconditional.

The delay in the sell off of the Hindustan Petroleum Corporation Ltd. and Bharat Petroleum Corporation Ltd. has not gone in the government's favor and has discouraged foreign investors. So far, NELP rounds have been only moderately successful in attracting small international oil companies, but have failed to generate interest among the very large-size oil companies. Whether this is due to low prospects of Indian sedimentary basins or due to faulty policy has to be examined. However, the recent discovery in the Krishna–Godavari basin should improve the prospect of Indian sedimentary basins in the international market. Subsequent to this discovery, the government decided to include more deepwater areas in the fourth round of NELP.

6.2.3 Power Sector Reforms

Mounting power shortages and critical financial condition of state electricity boards (SEBs) have rendered the boards unable to add significantly to the power generation capacity, prompting the national government, in 1991, to encourage private-sector participation, with an objective of mobilizing additional resources for the sector. The milestones in power sector reforms are briefly summarized in the Table VIII. However, power sector reforms are also plagued with multiple problems. Even though the government has provided incentives to the reforming state in the form of higher allocations for the sector, this has not yielded the expected result, i.e., restructuring of the SEBs. Trifurcation of the SEBs into generation, transmission, and distribution companies has been initiated in only a few states, with no visible benefits yet evident. In view of high electricity

TABLE VIII

Milestones in Power Sector Reforms[a]

The Ministry of Power formulated a Common Minimum National Action Plan (MNAP) in 1996. The MNAP introduced wide-ranging reforms in the power sector, touching every aspect from generation to transmission and distribution and the state of finances of state electricity boards. The agenda for reforms included the following decisions:

1. Each state/Union Territory shall set up an independent State Electricity Regulatory Commission (SERC), which shall have the tariff fixation, licensing, and planning powers. The Indian Electricity Act of 1920 and The Electricity [supply] Act of 1948 were amended to enable these changes
2. A Central Electricity Regulatory Commission (CERC) was set up to regulate the bulk tariffs for central generating power plant and transmission utilities. Licensing and planning will come under CERC when the Union government gives notice.
3. The action plan also aimed to rationalize retail tariffs, which were to be set by SERC. Norms for minimum tariffs and cross-subsidization were also to serve to guide SERC in the tariff fixation process
4. Private-sector participation was allowed in the distribution of electricity, with initial coverage limited to few areas
5. The action plan also envisaged greater autonomy for state electricity boards, which were to be restructured and corporatized and run on a commercial basis

[a] The progress on all these fronts has been slow. SERC has not been established in all states, and tariffs for many sectors are still not reasonable. Major issues, such as allowing Independent Power Producers (IPPs) to sell power directly to bulk consumers, are still unresolved and so is the sorry state of finances of the state electricity boards. With the status of LNG being uncertain, many many gas-based power plants have either been postponed or are being run on naphtha. The National Thermal Power Corporation is still uncertain about the status of its gas-based power plants at Anta, Auraiya, Kawas, Gandhar, and Kayamkulam. The transmission and distribution losses are still very high in a majority of states, which is inhibiting improvements of the financial positions of state electricity boards. The procedure for getting permission for setting up a generating station is still lengthy and opaque. All these factors have caused many foreign IPPs to either abondon or postpone their plans for setting up power stations in India.

generating requirements projected for the future, and to promote private-sector investments, the government recently passed the Electricity Bill (2003), which seeks to create a liberal framework of development for the power sector by distancing the government from regulation. It replaces three existing laws, the Indian Electricity Act of 1910, the Electricity (Supply) Act of 1948; and the Electricity Regulatory Commissions Act of 1998. The Electricity Bill seeks to promote competition in the electricity sector in India by decoupling generation, transmission, distribution, and supply of electricity. The bill also envisages preparation of a national electricity policy (including tariff) for the development of the power system based on optimal utilization of natural resources. In consonance with this policy, the Central Electricity Authority will prepare the National Electricity Plan once every 5 years. Perhaps the most important feature of the bill is the provision of open access to the transmission and distribution infrastructure in the country. Under the new bill, the generator and the consumer can individually negotiate the power purchase and use the common-access T/D system to meet the goals. The commissions aim to reduce cross-subsidization in the system and reward efficiency in performance. Thus, the Electricity Bill (2003) maintains the trend in electricity reforms witnessed the world over by exposing generation and the supply side of the market to competition, but placing T/D sections under an incentive regulation.

6.2.4 Coal Sector Reforms

A Committee on Integrated Coal Policy was initiated by the Planning Commission in 1996 to address problems in the coal sector. The proposals of this committee included the following major recommendations:

- Adoption of coal conservation measures.
- Augmentation of domestic coal production by inviting private capital.
- Integration of the exploratory efforts of coal and lignite.
- Deregulation of coal prices.
- Creation of a regulatory body.
- Establishment of more pithead thermal power plants.
- Augmentation of the railways and port infrastructure facilities.
- Acceleration of project clearance procedures.

The recent move by the government toward reforming the coal sector is exemplified in the constitution of Expenditure Reforms Commission (ERC). However, to date, deregulation of prices and the distribution of coal of all grades are the only recommendations that have been implemented by the government.

6.3 Subsidies

Subsidies form an issue of major concern, not only for the energy sector, but also for the entire economy. Problems due to subsidies arise when long-term gains are sacrificed for short-term gains, which are guided

by political will and have major influence on policymakers. It has been common experience that subsidies are not withdrawn until long after they have stopped serving the purpose for which they were instituted. Subsidies doled out by the government to one sector often have a lasting impact on some other sector.

The International Energy Agency (IEA) estimated in 1999 that energy savings from removal of subsidies amount to 7.2% of total primary energy supply, and that CO_2 emissions would be cut by 14%. The fiscal savings would amount to $8.6 billion. Eliminating all subsidies may not be feasible for the Indian political system, but rationalization of the plethora of subsidy schemes is definitely of prime importance and the government has taken important steps toward removing price controls on oil and coal and lowering of subsidies generally. Coal prices were decontrolled in 2000, and there are no longer any direct subsidies to coal production or consumption. Delivered coal prices, nonetheless, remain below market levels due to continuing subsidies on rail transportation. In April 2002, the government completed the dismantling of the APM for oil products and natural gas, and the removal of all subsidies, except for those for kerosene and LPG used by households. The Indian electricity sector is heavily subsidized. In 2000/2001, the average rate of subsidy expressed as a proportion of the estimated full cost of electricity supply was 93% for farmers and 58% for households. Industrial and commercial customers and the railways pay above-cost prices.

6.3.1 Effects of Subsidies on Sector Development

One prominent example of cross-subsidization impact, i.e., hurting one sector of the economy at the expense of another, is the imposition of a high electricity tariff on the Indian Railways. The Working Group on Energy Policy in 1980 had recommended electrification of 1000 km of route of railway track every year because of low costs and high efficiency of electrified engines. It is estimated that the cost of energy per 1000 gross tonne km for a diesel engine is 78 rupees, compared to 55 rupees for an electric engine. The cost of operation and maintenance per 1000 gross tonne km for a diesel-based engine is 26.5 rupees, whereas that for an electrified engine is 17.5 rupees. However, because the railways subsidize the residential and agricultural sectors and pay a high electricity tariff, only 550 km of track has been converted per year since 1980. This type of cross-subsidy, apart from exerting a toll on railway finances, undermines system efficiency and clogs up the railway network. More important, it makes the railways dependent on crude oil for operation, a situation not desirable from either the economic or the security point of view.

6.3.2 Timely Review of Subsidies Essential for Effectiveness

LPG subsidies were instituted in order to encourage consumers to use LPG as cooking fuel. The penetration of LPG in urban centers is now near saturation, yet subsidies are still in place. Efforts by the government to eliminate them have not been successful, even though the subsidy monies could be better used elsewhere. This serves as another example to highlight the impact that a policy framed in one sector may have on other sectors. The National Energy Policy should recognize linkages between sectors and account for them in the policy framework.

7. IMPLICATIONS FOR NATIONAL ENERGY POLICY

Energy consumption in India is bound to grow manifoldly as India aims for sustained economic growth; growth will primarily be driven by increasing commercial energy consumption, given the positive and greater-than-unity energy elasticity of the economy. Usage of commercial energy sources, particularly oil and gas and power, will remain crucial for development. Therefore, policy initiatives to ensure flow of investments in these sectors become imperative. Domestic energy supply is likely to fall short of indigenous demand and, hence, reliance on imports will be unavoidable. To meet the growing demand for energy in a sustainable manner, India needs to look at all possible ways of bridging the energy gap. To achieve this, India has to take a long-term view of economic imperatives and move away from crisis management to providing strategic direction for energy development.

Before embarking on an integrated energy policy, policymakers should study the effect of existing energy policies on the already degraded environment and initiate effective research and development for alternatives to the conventional form of energy use. A long-term integrated NEP should keep in view both macro- and microlevel implications. Various recent technological innovations are giving rise to a competitive fuel market. For example, in the new-age power plants, gas is more efficient than coal. There-

fore, with improved LNG supplies, prospects of using gas for power generation are also improving, which will lead to gas providing tough competition to coal. Hence, for balanced energy consumption, linkages between different ministries need to be strengthened; in view of the fuel choice facing the power sector today, pricing policies adopted by the government will play an important role in influencing the decisions made by power generators.

Environmental concerns can no longer be ignored and policies should take into account the likely impact on the environment, and every attempt should be made to reduce negative impacts. Reforms in the energy sector are also important, both to attract private sector investment and to improve the energy efficiency of the economy.

To sum up, achieving energy efficiency and sustainable development requires the following essential approaches:

- A cabinet committee under the chairmanship of the Prime Minister (for effective coordination and monitoring between ministries).
- Immediate and effective reforms in various energy sectors and subsectors, representing both the supply side and the demand side, with an integrated approach.
- Policy coordination across energy ministries and other ministries for an integrated energy policy.
- Effective linkages with foreign policy.
- Revival of the National Environmental Council and its promotion as the forum for discussion and debate on energy–environment policies for the country and the changes that need to be brought about from time to time.

Overall, it can be said that energy has to be seen as part of overall development strategy. Interventions in the energy sector provide only short-term benefits unless they are combined with matching interventions in other sectors that lead to the spread of economic opportunities among the poor.

Acknowledgments

The authors appreciate the assistance of Ms. Shobha Mishra.

SEE ALSO THE FOLLOWING ARTICLES

Development and Energy, Overview • European Union Energy Policy • National Energy Policy: Brazil • National Energy Policy: China • National Energy Policy: Japan • National Energy Policy: United States • Rural Energy in India

Further Reading

Energy Information Administration (EIA). (2003). "International Energy Outlook 2003." EIA, Washington, D.C.

International Energy Agency (IEA). (2002). "World Energy Outlook 2002." IEA, Paris.

Ministry of Law, Justice, and Company Affairs (Legislative Department). (2001). "The Energy Conservation Act, 2001." Ministry of Law, Justice and Company Affairs, Government of India, New Delhi.

Ministry of Nonconventional Energy Sources. (2002). "Annual Report 2001–2002." Ministry of Nonconventional Energy Sources, Government of India, New Delhi.

Ministry of Petroleum & Natural Gas. (2002). "Indian Petroleum and Natural Gas Statistics, 2000–2001." Ministry of Petroleum & Natural Gas, Government of India, New Delhi.

Ministry of Petroleum & Natural Gas. (2002). "The Petroleum Regulatory Board Bill, 2002." Ministry of Petroleum & Natural Gas, Government of India, New Delhi.

Ministry of Power. (2003). "The Electricity Act, 2003." Ministry of Power, Government of India, New Delhi.

Planning Commission. (1999). "Hydrocarbon Vision 2025." Planning Commission, Government of India, New Delhi.

Sengupta, R. (1993). "Energy Modelling for India; Towards a Policy for Commercial Energy." Planning Commission, Government of India, New Delhi.

The Energy and Resources Institute (TERI). (1998). "Green India 2047; Looking Back to Think Ahead." TERI, New Delhi.

The Energy and Resources Institute (TERI). (2001). "Green India 2047; Direction, Innovations, and Strategies for Harnessing Actions for Sustainable Development." TERI, New Delhi.

The Energy and Resources Institute (TERI). (2002). "Defining an Integrated Energy Strategy for India." TERI, New Delhi.

The Energy and Resources Institute (TERI). (2003). "TERI Energy Data Directory and Yearbook (TEDDY), 2002/03." TERI, New Delhi.

United Nations Development Program (UNDP). (2000). "World Energy Assessment." UNDP, New York.

World Bank (2000). "World Development Indicators, 2000." The World Bank, Washington, D.C.

World Bank Group. (2001). "Indoor Air Pollution. Energy and Health for the Poor 4." The World Bank, Washington, D.C.

National Energy Policy: Japan

PAUL J. SCALISE
Consultant
Tokyo, Japan

1. Energy Resources
2. Energy Deregulation
3. Social and Environmental Impacts
4. Conclusion

Glossary

deregulation The time period over which countries/states open up to full retail competition of a preferred energy supplier, down to the household level.

distribution network The low-voltage electricity network that runs from electricity substations to the end user. Distribution is effectively a monopoly activity.

economies of scale A cost function exhibiting "natural monopoly" properties; economies of scale are present if the marginal costs of production of a single-product firm are less than the average costs of production over the relevant range of output. Put differently, economies of scale are said to exist over the relevant range of output should unit costs decline with the volume of production (on a kilowatt-hour basis.)

independent power producer (IPP) A producer of electricity whose plant is not affiliated with a local utility company; independent plants operate in a competitive, unregulated environment.

kilowatt-hour (kWh) A standard unit of electric consumption corresponding to usage of 1000 W for 1 hour. A 100-W light bulb burning for 10 hours consumes 1 kWh.

tariff A public schedule detailing utility rates, rules, service territory, and terms of service; tariffs are filed for official approval with a regulatory agency.

total primary energy supply (TPES) This is made up of production + net exports − international marine bunkers ± stock changes.

The objectives of Japan's national energy policy today are defined by the Ministry of Economy, Trade, and Industry (METI) as the "four Es": energy security, environmental protection, economic efficiency, and economic growth. Although each goal is separately crucial to the long-term national good, implementation and compatibility of these four objectives have been in conflict for much of the time since World War II. The second largest economy in the world, Japan in the year 2000 represented 0.28% of the world's landmass, 2.1% of the world's population, and 15.3% of the world's gross domestic product. Such statistics form the backdrop to a densely populated industrial society that heavily depends on stable supplies of energy to survive. Yet, the country suffers from a lack of indigenous natural resources, especially fossil fuels such as oil, coal, and natural gas—resources that are crucial in the maintenance of infrastructure and rising aggregate consumption. As such, Japan imported nearly 84% of the total amount of its primary energy supply in fiscal year 2000. This rate is the highest of any major industrial country (save Italy.) The problems and prospects of Japan's changing economy and how this relates to its energy policy are discussed in this article.

1. ENERGY RESOURCES

Historically, geographical and commodity vulnerabilities have led Japan to a "dependence-based" energy strategy in which Middle Eastern oil-producing countries have been afforded special considerations. This strategy became critical during the oil shock in 1973, when the price of Brent crude oil rose substantially, thereby forcing Japan to take special measures not always compatible with the broader goals of its U.S. ally. Thus, reevaluation of Japan's domestic energy policy resulted in further energy diversification and, in particular, a major nuclear construction program.

With the end of the Cold War and advent of the economically stagnant 1990s, Japan's energy policy has moved away from one of pure "dependency" and

 © 2004 Elsevier Inc. All rights reserved.

"economic aid in exchange for oil" within highly regulated energy markets to a policy of gradual deregulation, market efficiency, and alignment with the United States to obtain regional security in the maintenance of its oil supply flow. Judging by the country's position in energy consumption among the Group of Seven (G7) countries (Table I), Japan ranks second in terms of total primary energy consumption as well as in oil and coal, third in nuclear and hydroelectric energy, and fifth in natural gas. Moreover, the country relies heavily on overseas sources for its total amount of primary energy supplied (84%) in fiscal year 2000. This percentage is the highest of any major industrial country. Consequently, Japan accounts for a very large proportion of the world's energy imports, ranking second after the United States in total imports of primary energy, first in coal and natural gas, and second in crude oil and oil products. This situation suggests that Japan is both a driver and a vulnerable target of material fluctuations in the price of fossil fuels on the world market.

1.1 Domestic Energy Supply and Demand: A History

The process of energy development in Japan's postwar economy can be divided into five stages: postwar reconstruction (1945–1955), the high-growth period (1955–1973) lasting until the 1973 oil crisis, the stable growth period (1973–1985) and its subsequent adjustments, followed by the incipient "bubble" growth period (1985–1991) that ultimately led to the country's current stage, the so-called restructuring period (1991–present). These five periods of development are different in their respective patterns and correspond to changes in Japan's energy supply and demand.

1.1.1 Postwar Reconstruction (1945–1955)

The postwar reconstruction period was characterized by its industrial reorganization, development, and modernization in the aftermath of Japan's defeat in World War II. In an attempt to meet the challenge of securing a stable supply of energy and saving foreign exchange, the Japanese government adopted "a priority production system" in 1946 that targeted the domestic extraction of coal and increased development of hydroelectric power. As shown in Fig. 1, coal continued to play a central role in Japan's primary energy supply until it was overtaken by oil in 1963. Concurrently, the shift toward greater use of electricity as an alternative to coal coincided with the establishment of nine regional, but privately owned, electric power companies in 1951.

1.1.2 Rapid Growth (1955–1973)

Throughout a stage of rapid growth from the mid-1950s to the early 1970s, Japan's real gross domestic product (GDP) growth averaged 9.2% per annum. One important factor in the development of this performance was the increasing supply of imported fossil fuels, such as inexpensive crude oil. In conjunction with this "oil revolution," the country structurally shifted its emphasis away from agricultural services to heavy and chemical industries; the

TABLE I

Energy Consumption, 2000: Japan's Rank among G7 Countries[a]

Rank	TPES (%)[b]	Oil (%)	Nuclear (%)	Hydro (%)	Coal (%)	Natural gas (%)
1st	United States (56.4)	United States (56)	United States (42.9)	Canada (43.1)	United States (66.9)	United States (58.2)
2nd	Japan (12.9)	Japan (13.9)	France (22.3)	United States (29.9)	Japan (11.6)	United Kingdom (9.3)
3rd	Germany (8.3)	Germany (6.9)	Japan (17.3)	Japan (10.5)	Germany (10.0)	Canada (8.0)
4th	France (6.3)	Canada (6.1)	Germany (9.1)	France (8.0)	United Kingdom (4.4)	Germany (7.7)
5th	Canada (6.2)	Italy (5.9)	United Kingdom (4.6)	Italy (5.3)	Canada (3.8)	Japan (6.9)
6th	United Kingdom (5.7)	United Kingdom (5.8)	Canada (3.9)	Germany (2.6)	France (1.8)	Italy (6.2)
7th	Italy (4.2)	France (5.4)	Italy (0)	United Kingdom (0.6)	Italy (1.6)	France (3.8)

[a] Data from International Energy Agency (2002); calculations by the author.
[b] TPES, Total primary energy supply.

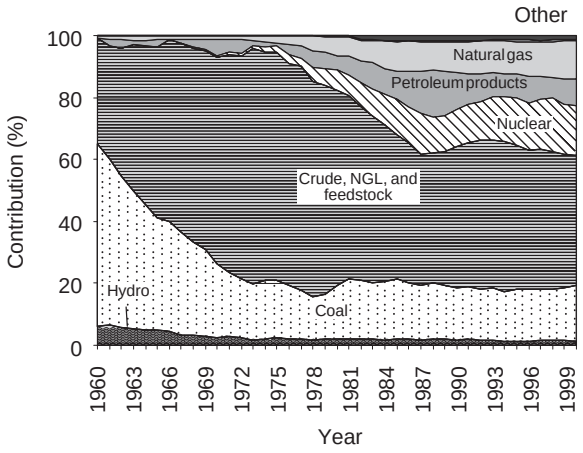


FIGURE 1 Long-term trends in primary energy supply by fuel. NGL, Natural gas liquids. Data from International Energy Agency (2002); calculations by the author.

government's primary objective was to support an "export-led" strategy of economic development in which oil consumption for domestic use increased materially. The share of oil as a percentage of the total primary energy supply increased from 17.6% in 1955 to 55.2% in 1965, while coal declined from 47.2% to 36.5% during the same period (Fig. 1). This trend continued as oil's share peaked at 76.4% in 1973, the year of the first world oil crisis. Oil replaced coal as the country's leading source of energy in 1963, and in that year thermal power replaced hydroelectric power as a percentage of total electric energy supplied.

1.1.3 The First Oil Shock and Stable Growth (1973–1986)

By 1973, Japan's energy self-sufficiency rate reached its nadir (10.6%). The country's dependence on imported oil (99.8%) suggested that any sudden fluctuation in world prices would cause serious discontinuities within the Japanese economy. The outbreak of the first oil crisis in 1973 dispelled the notion that cheap, readily available supplies of imported oil would always prevail. Crude oil prices increased fourfold, and Japan was compelled to rethink its national energy strategy. The second oil crisis in 1978/1979 reinforced this need. Industries began to adopt energy efficiency measures, and Japan's industrial framework shifted again from energy-intensive industries (e.g., basic materials) to less energy-demanding industries (e.g., assembly and processing). Consequently, energy consumption only grew 1.5% per annum from fiscal 1973 to 1979, and decelerated to 0.5% per annum from fiscal 1979 to 1985.

With the onset of the two oil crises, the energy policy of the government prioritized national energy security, and this was implemented in a few ways. First, the government sought to minimize its risk by increasing oil stockpiles and diversifying its imported oil sources. The second policy aim was to reduce dependence on oil by adopting a "diversified fuel-mix strategy." Alternative energy sources such as coal, liquefied natural gas (LNG), and nuclear energy were emphasized (Fig. 1).

1.1.4 Japan's "Bubble" Growth Era (1985–1991)

The Japanese economy witnessed a material turnaround in both performance and consumer activity during the so-called bubble years; this was a period characterized by rapidly rising real estate and stock market prices in highly speculative markets. In 1986, the material decline in crude oil prices, brought about by a reduction in the crude target prices of the Organization of Petroleum Exporting Countries (OPEC) for the first time since the 1973 oil crisis, led to a surplus on the world market. In turn, Japan's economy became a major beneficiary of the development as the compound annual growth of the primary energy supply from 1986 to 1991 increased by 3.9%, or 4.5% in industry, 4.3% in the household/service sector, and 5.3% in transportation. Japan, once again, was seeing an increase in energy consumption.

Concurrently, export-led companies attempted to compensate for the economic shocks imposed on the Japanese market of the previous decade, by rationalization efforts, cost-cutting, and increased exports. As the trade surplus with the United States increased, international pressure to shift Japan away from an "export-led growth strategy" to a "consumer-led growth strategy" intensified. In April 1986, the Advisory Group on Economic Structural Adjustment for International Harmony, a panel appointed by the prime minister and headed by former Bank of Japan Governor Maekawa Haruo, issued a set of recommendations otherwise known as the *Maekawa Report*. [The Japanese practice of putting family names of Japanese people first is followed in this article (for example, Ono Yoko, rather than Yoko Ono). Western names retain the standard form of first and last names (for example, Dwight Eisenhower).] This report set the stage for a series of proposals to restructure the Japanese economy with long-term ramifications for many domestic sectors, including energy. The concrete steps proposed

centered first on the expansion of domestic demand, focusing on areas such as industry restructuring, technology development, and changes to the quality of housing.

1.1.5 Structural Adjustment (1991–Present)

Since 1991, after the bursting of its bubble economy, Japan has faced unprecedented and difficult questions. Electricity demand growth, crippled by stagnant GDP elasticities, has fallen to barely flat levels, and incumbent energy industries have been forced to accept new regulatory enhancements to ensure greater market efficiencies under the banners of internationalization, globalization, and liberalization.

Throughout the decade, the government fundamentally shifted its policy from energy security to an overall reconsideration of its national agenda. Economic efficiency as an economic policy driver became another pillar to the METI's growing list of goals. Although deregulation incrementally lowered prices in energy industries, including oil, electricity, and gas, it also undermined profitability, forcing industrial consolidation and corporate efficiency measures.

1.2 Energy Consumption by Sector

Changes in demand can also offer valuable insights into the nature of energy growth. Since the 1973 oil shock, industrialized nations have made concerted efforts to conserve energy (Fig. 2). On an index with 1973 set at 100, Japan led the industrialized nations in energy conservation; between 1973 and 1987, the total consumption level as a percentage of GDP fell

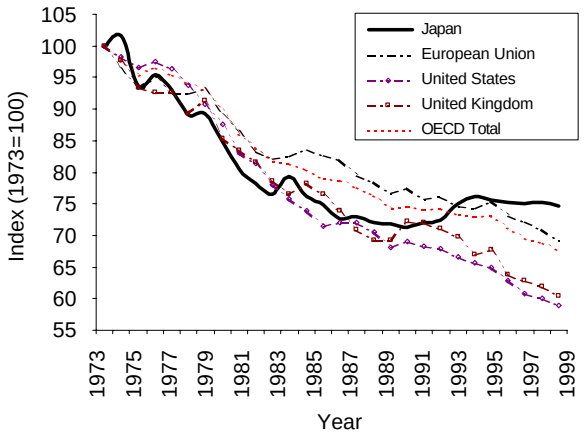


FIGURE 2 Index of total primary energy supply per unit of gross domestic product. OECD, Organization for Economic Cooperation and Development. Data from International Energy Agency (2002); calculations by the author.

(energy efficiency improved) at a compound annual rate of 2.3%, which was higher than that of any other Organization for Economic Cooperation and Development (OECD) country. As mentioned, such efficiency gains were due, in part, to a shift in orientation from heavy, energy-intensive industries, such as iron and steel, cement, nonferrous metals, and petrochemicals, to machine-based and high-technology industries, such as automobiles, appliances, consumer electronics, and industrial ceramics, and a further shift from secondary to tertiary industries with increasing emphasis on service trades. This shift can be expected to continue, but at a slower pace than in the past. Japan is under pressure to increase its domestic consumption to a level more in line with its trade competitors. This international pressure, in conjunction with growth in road transport, can be expected to increase energy intensity.

It is important to note, however, that the industrial sector has been the only contributor to major energy efficiency improvements. As Fig. 2 indicates, by the 1990s, the total primary energy supply per unit of GDP materially shifted; energy consumption levels within the household/service, commercial, freight transportation, and passenger transportation sectors markedly increased, thus outweighing previous efficiency gains in the industrial sector. An international comparison of per-household energy consumption reveals that, adjusting for heating figures related to atmospheric temperature differences among countries, Japanese data values are still quite low, suggesting that the current growth in energy consumption is the direct result of income elasticity among affluent economies. In 1973, for example, there was only one car for every two families in Japan. In 2000, the number of cars on the road reached over 52 million versus only 45 million households. Moreover, in 1970, fewer than 10 out of every 100 Japanese homes had such things as videocassette recorders, microwave ovens, and air conditioners, and only 43 out of every 100 homes had color televisions. By 2000, the number per 100 households had risen to 84 for videocassette recorders, 95 for microwave ovens, and 217 for air conditioners, and color televisions topped the list at 231.

2. ENERGY DEREGULATION

Deregulation of the Japanese economy began in the mid-1980s with the introduction of the so-called *Maekawa Report*. In the 1990s, deregulation ex-

tended into the energy industry; first materially affecting the petroleum industry and later extending into the electricity and gas industries. For many years, Japan's energy industry was a textbook example of a regulated industry, but the picture has changed. A string of events has taken place, altering the operational landscape, including partial liberalization of the electricity and gas retail markets and the lapse of the Refined Petroleum Import Law.

2.1 Electricity

The Japanese electricity market is not only one of the largest in the world, but also one of the most lucrative. In 2001, total market revenues (*ex* self-generation) were worth more than ¥15 trillion (=$115.3 billion U.S. dollars; $1=¥130). Tokyo Electric Power Corporation (TEPCO), the world's largest privately owned utility, boasted total assets of more than ¥141.3 trillion ($1087 billion), volumes of 275.5 TWh, and revenues of ¥5.2 trillion ($40.1 billion). The sheer mass of TEPCO translated into one of the highest net profit streams in Japan despite high interest payments on debt. The key feature to this arrangement has been the existence of vertically integrated regional monopolies that have been heavily regulated by the state. After a brief period of government wartime control, nine private power companies were reorganized in 1951 to come into control of most of the generation, transmission, and distribution businesses of electric power throughout the country. Their reclaimed monopoly status was the result of successful lobbying by utilities to return to a system of centralized private power, reminiscent of the early 1930s, not of the highly competitive era of the 1920s.

The greatest strength of the new "1951 system" was that it exploited economies of scale, maintained stable electricity prices for over 20 years, and kept dangerous dioxide emissions in check. Its greatest weakness was that it eventually disproved the theory; the 1951 regulatory structure failed to reconcile rising demand with mounting variable (fuel) and fixed (capital) costs. Over time, the weaknesses became more pronounced. As Fig. 3 indicates, from 1975 to 1985, average monopoly tariffs increased by almost 100% from 11 yen/kWh to 21 yen/kWh. Utility companies argued that the phenomenon was the result of exogenous variable shocks stemming from an overdependence on imported fossil fuels. Although true to an extent (99.8% of Japan's consumed oil is imported, 87% of which is imported

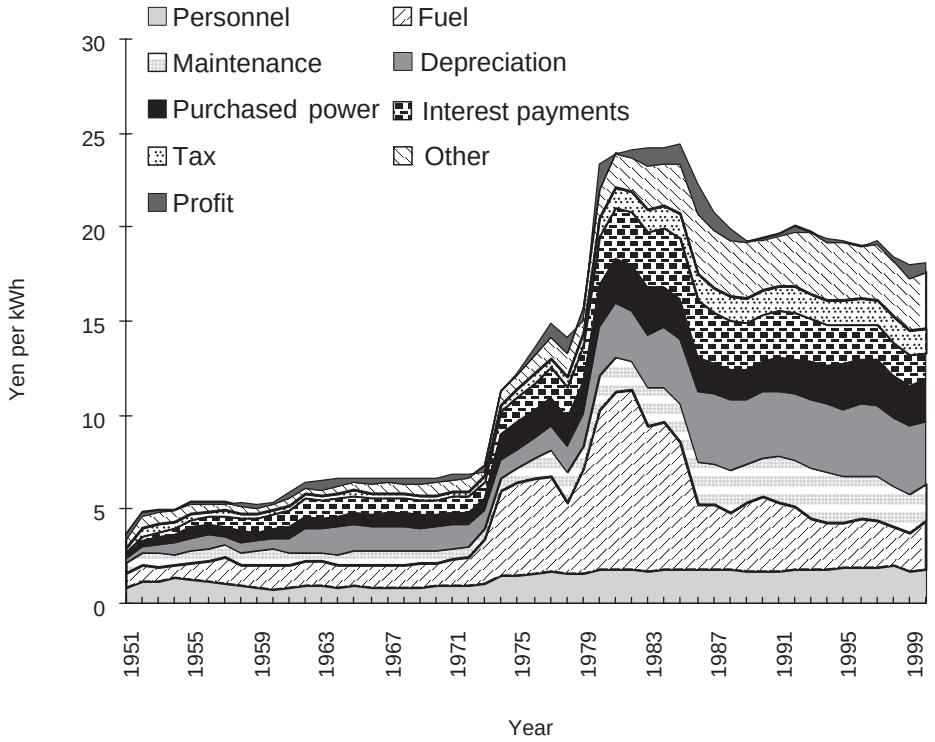


FIGURE 3 Average electricity tariff, fiscal years 1951–1999. Based on data from the 2002 annual report of the Tokyo Electric Power Company; calculations by the author.

TABLE II

Compound Annual Growth in Electricity by Industry Type[a]

Industry	Compound growth, fiscal years 1965–2000 (%)		Compound growth, fiscal years 1990–2000 (%)	
	Total volumes	Self-generation	Total volumes	Self-generation
Mining	−3.2	−1.5	−1.5	−5.3
Food	8.2	9.5	3.0	7.5
Textiles	0.3	11.9	−5.6	2.4
Paper	1.7	7.2	−1.2	2.0
Chemicals	0.5	5.9	−0.5	3.1
Oil	3.4	9.9	−4.8	8.1
Rubber	5.4	12.6	0.1	17.2
Ceramics	5.0	7.4	−2.3	4.5
Iron	2.4	5.5	−1.2	2.0
Nonferrous	3.2	1.3	1.4	0.6
Machinery	8.6	30.5	2.0	16.5
Other	7.1	8.3	1.7	9.3
Railroads	3.7	2.8	1.0	0.4
Total (manufacturing)	3.9	6.4	0.4	3.5
Grand total[b]	4.0	6.1	0.7	3.4

[a] *Source*. Energy Data and Modeling Center, *Handbook of Energy & Economics Statistics in Japan, 2002*; calculations by author.
[b] Grand total equals total manufacturing demand plus railroad demand and other.

from the Middle East), average tariff rates and the unit cost of fuel per kilowatt-hour continued to decouple after 1985. The lack of continuity became more pronounced as market rigidities and rising capital costs prevented a nominal readjustment to pre-1973 tariff levels. Nevertheless, the original presence of "natural monopoly" technologies continued to provide justification for a regulated power industry and accordingly to impose entry barriers, price controls, and obligations to serve. Unlike other industrialized nations, the postwar regulatory regime did not breakdown with the "energy crisis" of the 1970s. Rather, the then-Ministry of International Trade and Industry (MITI) overcame the crises through close cooperation with incumbent utilities, preserving the vertically integrated monopoly structure. The exploitation of economies of scale while controlling market power was not only considered the progressive "public interest" view of the country's utility regulation until the early-1990s, it was also a means of preserving the political status quo.

Industry restructuring began in earnest with the advent of the Hosokawa Morihiro coalition government in 1993. By that time, electricity prices from 1983 to 1993 were not only perceived to be high domestically, but also internationally. Pretax Japanese electricity tariffs were three times the average U.S. price in 1993. Broadly speaking, pressure to enact reforms was initiated by several actors, including the incumbent utilities concerned with the threat of self-generators (*jika hatsudensha*) capturing further market share (Table II) the export-led industries burdened with an appreciating currency and mounting operational costs; the potential new entrants seeking to duplicate their success in foreign deregulated markets; and the central government, which was looking not only to maintain its legal powers over the market, but also to placate foreign government pressures (*gaiatsu*).

In response to such growing pressure to lower electricity tariffs, the government decided to undergo regulatory reform. Revisions to the Electric Utility Industry Law were intended to produce a more competitive market, albeit incrementally. Phase one (1995) permitted the nine major power companies (*ex* Okinawa) to act as "single buyers" through a revision of the 31-year-old Electric Utility Industry Law. A bidding structure organized independently of government control was established whereby independent power producers (IPPs) and other self-generators could bid for new contracts for their supplementary power needs. Major players in phase one were predominantly steel makers such as Nippon Steel Ltd. and Kobe Steel Ltd. Phase two (2000)

expanded the scope of liberalization measures to include the following steps:

1. Retail choice for large-lot customers: Competition was opened to utilities servicing customers with contracts for 2 MW or greater and connected to a supply system of 2000 V or higher. Such customers included owners of large-scale office buildings, industrial factories, hotels, hospitals, and department stores. Unlike phase one, customers could now choose their preferred wholesale supplier, whether foreign or domestic.
2. Competitive bidding process for thermal capacity acquisition: The nine incumbent power companies (*ex* Okinawa) were required to place competitive bids when they purchased thermal capacity. Hydroelectric and nuclear power competition were (and are) beyond the scope of the regulatory changes.
3. Profit sharing: Incumbent utilities were no longer required to submit rate case applications for tariff reductions. This served as an incentive to reduce costs voluntarily because the incumbent utility could reserve part of the increased profit margin through reduced service costs.
4. Abolition of Article 21 of the Antimonopoly Law: Without this provision, discriminatory rights to railroad and utility companies (natural monopolies) were no longer upheld. Both the Japan Fair Trade Commission (JFTC) and the Ministry of Economy, Trade, and Industry enacted guidelines that legally prohibited any exclusionary activities inhibiting new entrants.

As of the date of writing of this article, a quantifiable results-oriented assessment of Japan's electricity deregulation suggests only tepid competition, average incumbent tariff declines, and minimal corporate efficiencies. New electricity market entrants account for only 0.85% of total installed capacity nationwide. Structural barriers to market entry, such as extensive backup, wheeling, and maintenance charges, preclude cost-competitive access to many incumbent networks (transmission and distribution). Furthermore, lack of available land for further capacity build and stringent environmental regulations in generation also present strong obstacles to market entry.

Incumbent electric power companies have begun to lower tariff rates in anticipation of future market competition, but, as Fig. 3 indicates, readjustments have been preemptive and voluntary, not forced through direct competition. On an index with 1990 set at 100, average electricity tariffs for Japan's three largest electric power companies (TEPCO, Kansai Electric Power Co., and Chubu Electric Power Co.) in 2002 fell to 93, suggesting only tepid market pressures to date.

2.2 Gas

In contrast to the electric power industry, the Japanese gas distribution industry is best characterized as vertically truncated, but within a framework of several regional monopolies. Non-Japanese firms have traditionally conducted the "upstream" activities of exploration and supply. The "downstream" activities of refinement and distribution have fallen within the purview of companies such as Tokyo Gas, Osaka Gas, Toho Gas, and Saibu Gas, to name only the four largest companies in terms of volumes, revenues, and assets. Historically, the Japanese gas industry was privately owned but publicly regulated in the postwar era under the Gas Utilities Industry Law. Policymaking and regulatory functions were the responsibility of the MITI (now METI). This situation, however, is beginning to show signs of change, due to the strategic importance of liquefied natural gas (LNG) to Japan's long-term energy needs.

The majority of LNG is used within the electric power sector. In 2000, power plants accounted for nearly 70% of total gas consumption. With the gradual liberalization of the electric power industry in the 1990s, liberalization and restructuring are currently underway in the gas industry as well. The first measures aimed at liberalizing Japan's gas market were adopted in June 1994 and went into effect in March 1995. Revisions were made to the Gas Utilities Industry Law that allowed gas utilities to sell gas on a retail basis to large-lot volume users, thereby circumventing incumbent gas distribution companies. These contestable customers were defined as those consuming more than 2 million cubic meters (MMcm) per year, representing approximately 720 gas users. The changes permitted non-gas companies (power, steel, oil, and other companies with access to gas) to supply large-lot industrial customers for the first time. The revisions also lifted tariff regulations so that parties in the liberalized segment of the market were free to determine price and other contract terms on a negotiated case-by-case basis.

Phase two of the Japan's gas liberalization was adopted in May 1999 and went into effect the following November. These revisions expanded the contestable market to eligible consumers of 1 MMcm per year, or approximately 993 large-lot users. Designated general gas companies were also ordered

to grant third-party access to gas pipelines on a nondiscriminatory basis. The designated companies are Tokyo Gas, Osaka Gas, Toho Gas, and Saibu Gas. The smaller gas distribution companies, such as Shizuoka Gas, Hiroshima Gas, Chubu Gas, and Hokkaido Gas, to name a few, have no obligation to provide access to their infrastructure.

In 2000, a subcommittee established by the METI recommended a pricing formula for third-party access to both the incumbent electric power transmission system and the incumbent gas distribution pipeline system. Known as the "wheeling charge," this forward-looking cost method allowed for historical comprehensive costs, including depreciation, taxes, and other operational costs, to be added to a stipulated rate in return for third-party access to the transmission network or pipeline. Based on these rules, the four gas companies issued rate schedules for transportation, backup, and storage in March 2000. Further transparency was also imposed regarding the accounting methods and public disclosure requirements to be used by the gas companies in allocating costs among various business segments.

At the time of writing this article, new entrants to the newly liberalized gas industry have been predominantly incumbent electric power companies. TEPCO, for example, as of January 2001, has supplied local city gas companies via its own low-pressure pipeline in Chiba Prefecture; since 2001, Chubu Electric Power Co. has had a joint venture with Iwatani International Corporation and Cosmo Oil to supply LNG via a lorry operation in Mie Prefecture, and Tohoku Electric Power operates a mass natural gas supply business through its subsidiary Tohoku Natural Gas using the Niigata–Sendai pipeline. METI is currently monitoring the outcome of these liberalization measures and subsequent moves by new entrants. The agency is also examining what measures could be taken next and recently impaneled a new committee to develop recommendations for further liberalization.

2.3 Petroleum

Unlike the vertically integrated electric power industry, but similar to the gas industry, Japan's privately owned petroleum industry has a long history of rigid regulations and failed government direction, resulting in a vertically truncated and horizontally fragmented industry. Incremental deregulation of the Japanese petroleum industry eventually illustrated that regulatory change could lead to greater consumer benefits and corporate consolidation, but not necessarily increased industrial structures or reduced fiscal burdens of the government. The reason for such a regulatory framework stems from the fact that crude oil, natural gas liquids (NGLs), and feedstock consumption meet 42% of Japan's annual energy needs. Japan accounts for 7% of total world oil consumption. Although the country's dependence on oil has decreased from 77% in 1973 thanks to a concerted effort to diversify the national energy supply, oil remains a key energy source (Fig. 1). At 99.8%, Japan's import dependence on oil is extreme and oil is therefore considered a strategic national resource, especially because 80% is sourced from the politically unstable Middle East.

Historically, the petroleum industry was bifurcated into the "upstream" activities, related to the exploration for and production of oil, and the "downstream" activities, related to oil refinement and distribution (wholesale and retail). These two activities coincided in the prewar era, but slowly gave way to strictly "downstream" activities owing to insufficient domestic supply, horizontal market fragmentation, and high fixed costs. Japanese oil firms became heavily dependent on the "majors" (large international oil companies) for upstream operations. It is within this framework that the 1962 Petroleum Industry Law was enacted. Essentially, the law was written to achieve a stable supply of oil by controlling downstream oil refining, effectively authorizing the separation of upstream and downstream activities through permits to establish refineries or to purchase equipment for them. Several laws were subsequently enacted to reinforce the 1962 framework; these include the Petroleum Supply and Demand Balance Law, the Emergency Temporary Law for Stabilizing the Nation's Livelihood, the Petroleum Stockpiling Law, and the Quality Assurance Law. As a by-product, these laws indirectly managed to protect the fragmented structure of the industry, thereby propping up many of the smaller firms that were dependent on existing regulations to stay in business.

For a brief period during the 1980s, foreign pressure (*gaiatsu*) from Western nations in conjunction with foreign companies caused a media stir in Japan. It was suggested that the majors and independent oil companies could circumvent these laws in order to achieve greater market share in Japan's abundant downstream activities. Although unsuccessful, what followed was a gradual easing of regulations and laws in order to bend to public pressures to reduce high gasoline, kerosene, and naphtha prices in a domestic market that was

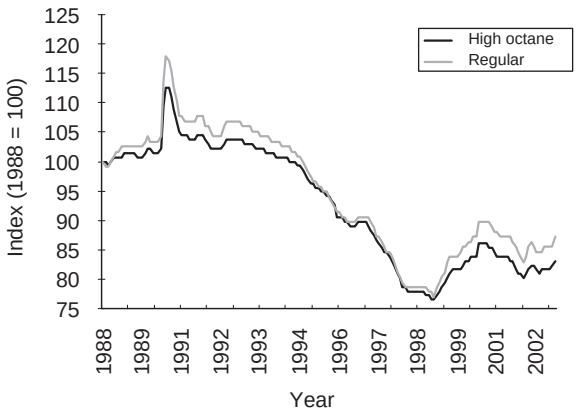


FIGURE 4 Gasoline per liter, fiscal years 1988–2002 (indexed). Based on data from the Oil Information Center; calculations by the author.

experiencing a clear slowdown in economic growth compared to the previous two decades. Technically, such liberalization of the petroleum industry began with the 1986 Refined Petroleum Import Law. The MITI eased its requirement for obtaining a license to import such refined products, answering calls from domestic and international actors to open markets. However, incumbent oil refiners successfully lobbied to limit the impact of the new legislation by requiring only those refineries in Japan with existing domestic production, stockpiling, and quality control to import oil products. The requirement effectively blocked all imports from nonincumbent players, making "petroleum liberalization" a paper tiger.

On March 31, 1996, the Refined Petroleum Import Law expired and the Gasoline Sales Law and Stockpiling Law were also revised. Emphasis shifted away from "energy stability" of imported petroleum to "energy efficiency," partly aided by advocates of reform in the newly established Hosokawa Morihiro coalition government. As shown in Fig. 4, extreme inefficiencies and ineffective legislation kept gasoline prices high in the 1980s, but real liberalization of the market led to greater market competition and increased consumer welfare gains despite yen depreciation and stable crude oil prices. Indexing average gasoline prices per liter at 100 in 1988, the price of regular and high-octane gasoline, respectively, fell steadily from their Gulf War high of 112 and 118 in 1990, to a historic low of 76.5 in 1999.

The key to such consumer welfare gains emanated from laws enacted to stir competition in the mid-1990s. With the Refined Petroleum Import Law removed, stringent restrictions on petroleum imports were also removed; import licensing was readjusted to require new capacity to decrease from 120 days' worth of product storage (a clear structural barrier to market entry) to only 60 days' worth of product. In turn, the abolition of the Specified District System, which prohibited new entrants from targeting fragmented districts in which there were already a number of gasoline stations, allowed for greater competition among retail players. Consequently, incumbent gasoline stations began to lower prices in an effort to build customer loyalty. Corporate profit margins, historically low by world standards, continued to decrease in the newly competitive operating environment. Consequently, consolidation, asset streamlining, and cost cutting were initiated in the industry. In 1999, Nippon Oil and Mitsubishi Oil, for example, merged to form the largest Japanese oil distributor by market capitalization (667 billion yen as of November 22, 2002), sales (estimated at 4 trillion yen in fiscal year 2002), and market share (23%).

3. SOCIAL AND ENVIRONMENTAL IMPACTS

Energy policy does not exist in a vacuum: like all governments, Japan's national energy policy must coexist with social and environmental issues. How these factors shaped Japan's changing energy policy landscape is discussed in the following sections. On the one hand, Japan's fossil fuel dependency did not add to national security. Alternative sources of energy, such as nuclear energy, were encouraged and developed to serve this end. On the other hand, Japan, as a leading nation, had environmental obligations such as those defined in the Kyoto Protocol in 1997. Environmental concerns were not always in conflict with Japan's wish to free itself from its fossil fuel dependency. However, the goals were not always in tandem.

3.1 Kyoto Protocol

In December 1997, delegates from 159 countries gathered in Kyoto, Japan, for the Third Conference of the Parties (COP3) to the United Nations Framework Convention on Climate Change (UNFCCC). The results of that conference, known as the Kyoto Protocol, addressed important commitments from developed countries to reduce emissions of carbon dioxide and other greenhouse gases between the years 2008 and 2012. Industrial economies were required to cut aggregate man-made emissions of six climate-affecting gases by an average 5.2% below

1990 levels by some point in the 2008–2012 time frame. The six gases, carbon dioxide (CO_2), nitrous oxide (NO_x), methane (CH_4), hydrofluorocarbons (HFCs), perfluorocarbons (PFCs), and sulfur hexafluoride (SF_6), were considered to be ozone-depleting chlorofluorocarbons. As such, developed countries agreed to incorporate into their efforts to implement the Kyoto Protocol measures that would promote, assist, and finance technology transfers in fighting occurrence of such emissions in developing countries.

Japan's emissions reduction target was agreed to be 6% and that for the European Union was 8%; the goal of the United States was to be 7%. As host nation, Japan stressed the importance of the Kyoto Protocol. Conservation and efficiency improvements yielded a double benefit. Conservation policies not only limited the country's overall dependency on imported fossil fuels, they also made export-led companies, such as the Japanese automakers, more competitive through fuel-efficient products.

One strategy to meet the stipulated goals of the Kyoto Protocol was to redouble efforts in the development of alternative sources of energy that reduce or eliminate pollution. If alternatives were palatable enough, significant reductions in the consumption of fossil fuels could be made. Solar, geothermal, fuel cell, and wind power sources were (and are) four such contenders in the implementation of such a strategy. These renewable energy sources currently hold only a small market share of the total primary energy supply. Wind power, for example, had an estimated 0.03% market share in 2000, but is projected to capture a 20% market share by 2020. The METI has been researching and developing practical development of these sources.

3.2 Nuclear Power

In complying with the goals of both the Kyoto Protocol and the METI's long-term energy strategy (energy security, environmental protection, economic efficiency, and economic growth), nuclear power continues to play an important, albeit controversial, role. Despite being the only country to suffer directly the fallout of nuclear weapons, Japan has adopted the peaceful use of nuclear technology to provide a material portion of its electricity generation. Today, nuclear energy accounts for some 30% of the country's total electricity production. However, nuclear accidents, policy contradictions, "not-in-my-backyard" protests, and regional funding issues present recurring obstacles to the government's plans to perfect a completely indigenous fuel source for the 21st century.

3.2.1 Nuclear Development and Policy: A Brief History

Following U.S. President Dwight Eisenhower's historic speech, "Atoms for Peace," at the United Nations in 1953, Japan began its nuclear research program. The Atomic Energy Basic Law was enacted in 1955 with the aim of ensuring the peaceful use of nuclear technology in Japan. Democratic methods, independent management, and transparency were (and are) the foundation of nuclear research activities, as well as the promotion of international cooperation. Several nuclear energy-related organizations were established in 1956 under this law to further promote development and utilization, including the Atomic Energy Commission; the Science and Technology Agency; the Japan Atomic Energy Research Institute (JAERI), and the Atomic Fuel Corporation (renamed the Power Reactor and Nuclear Fuel Development Corporation in 1967).

As Fig. 5 indicates, Japan has 52 reactors with a combined installed capacity of 45.7 GW, or approximately 30% of the nation's total installed capacity. The country imported its first commercial nuclear power reactor from the United Kingdom (Tokai-1) in 1966. This gas-cooled (Magnox) reactor built by General Electric Company (GEC) had a relatively small installed capacity of 160 MW; the reactor was finally decommissioned in March 1998. After this unit was completed, only light-water reactors (LWRs) using enriched uranium, either through pressurized water reactors (PWRs) or boiling water reactors (BWRs), have been constructed. Since 1970, 23 PWRs and 28 BWRs (including two advanced BWRs) have been brought online. Initially, Japanese electric power utilities purchased designs from U.S. vendors and built them with the cooperation of Japanese companies, who received licenses to then build similar plants in Japan. Companies such as Hitachi Co Ltd., Toshiba Co Ltd., and Mitsubishi Heavy Industry Co Ltd. developed the capacity to design and construct LWRs. An additional 11 reactors are in the planning stages or are currently under construction.

3.2.2 Reprocessing and Waste Disposal: National Policy

As already mentioned, one of the obvious goals of Japan's national energy policy is energy security. With the increase in nuclear power as a percentage of its total primary energy source, Japan hopes to

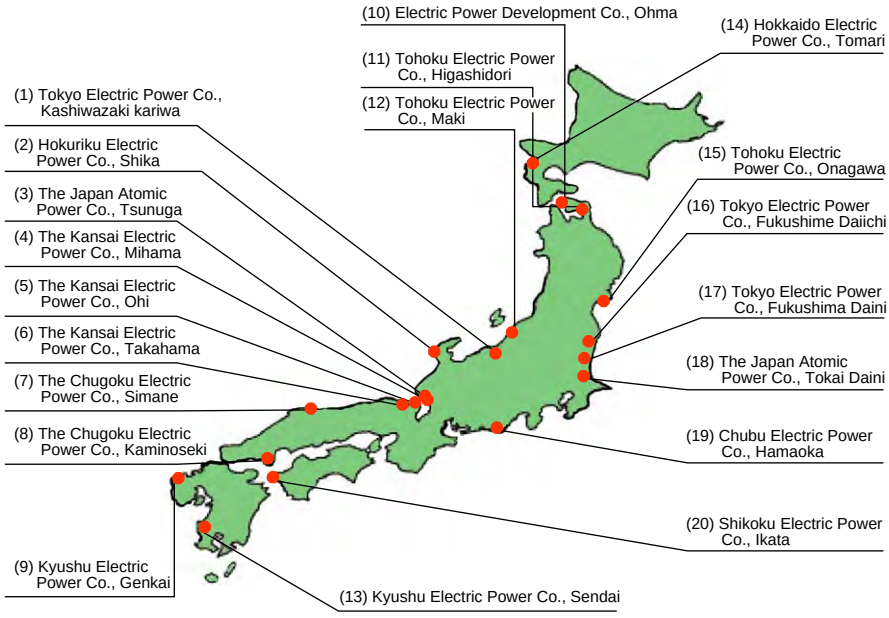


FIGURE 5 Nuclear power plants in Japan. As of 2003, 20 locations are power plant sites (or will be) for 63 units (52 operational, 3 under construction, and 8 in the planning stage), with an expected total output of 59,895 MW (45,742, 3838, and 10,315 MW, respectively). The locations on the map are numbered; the following key details the operational (OP), under-construction (UC), or planning-stage (PS) unit output at each location: (1) seven OP units, each >1000 MW; (2) one OP unit <1000 MW and one UC unit >1000 MW; (3) two OP units, one <500 and one >1000 MW, and two PS units, each >1000 MW; (4) three OP units, one <500 and two <1000 MW; (5) four OP units, each >1000 MW; (6) four OP units, each <1000 MW; (7) two OP units, one <500 and one <1000 MW, and one PS unit >1000 MW; (8) two PS units, each >1000 MW; (9) four OP units, two <1000 and two >1000 MW; (10) one PS unit >1000 MW; (11) one UC unit >1000 MW; (12) one PS unit <1000 MW; (13) two OP units, each <1000 MW; (14) two OP units, each <1000 MW, and one PS unit <1000 MW; (15) three OP units, each <1000 MW; (16) six OP units, one <500, four <1000, and one >1000 MW; (17) four OP units, each >1000 MW; (18) one OP unit >1000 MW; (19) four OP units, two <1000 and two >1000 MW, and one UC unit >1000 MW; (20) three OP units, each <1000 MW. The Japan Atomic Power Company Tokai plant closed in March of 1998. Map, plant locations, and output derived from the 2003 data of the Federation of Electric Power Companies.

reduce its import dependency on fossil fuels. However, the challenges facing the Japanese government are much more daunting with the introduction of nuclear power. Plutonium is essential as a major fuel for nuclear power generation. The more plutonium used by a country, the more likely the nation becomes influenced by international politics, especially because plutonium produced within the nuclear reactor can be used for nuclear weapons.

Recovering plutonium through a process known as the "fuel cycle" essentially reprocesses spent fuel. The theory is that once the nuclear fuel cycle is established domestically, nuclear power virtually becomes an indigenous energy. Until now, Japan relied on the reprocessing of spent fuel by European contracts through British Nuclear Fuels (BNFL) and Cogema, with vitrified high-level wastes being returned to Japan for disposal. However, this reprocessing has proved to be expensive and time consuming. In 2005, Japan Nuclear Fuel Ltd. (JNFL) will begin its first commercial enrichment plant at Rokkasho in northern Japan. Its planned capacity is to be 1.5 million separative work units (SWUs)/year (in the nuclear power industry, the separative work unit is a measurement of mass: 1 kg of separative work = 1 SWU). Spent fuel has been accumulating in Rokkasho since 1999 in anticipation of the full operation of the plant (shipments to Europe stopped in 1998).

The plutonium recovered by foreign reprocessing in the United Kingdom (BNFL) and France (Cogema) will be used in LWRs as mixed-oxide (MOX) fuel. MOX fuel was first intended to be used in the Takahama nuclear plant of Kansai Electric Power Company. However, local concerns surrounding the safety of MOX fuel in 2002–2003 created scheduling problems for the implementation of that program.

3.2.3 The Economics of Nuclear Power

The economics of nuclear power generation is largely controversial. Although the fuel cost of generation

is relatively inexpensive ($\geq$20% of total nuclear generation costs), its capital costs are increasingly expensive. In a deregulating market where the marginal cost of new entry is increasingly competitive vis-à-vis incumbent prices, the incumbent competitiveness of Japan's nuclear power remains unclear. Current nuclear projects depend not only on material cost-cutting within tight schedules, but also on operating such generators at reasonably high utilization rates over many years. By way of contrast, fossil fuel-fired power plants are relatively cheap and quick to build, but expensive to operate due to their cost of fuel (currently $\geq$70% of total thermal generation). Thus, the economics of nuclear power further inform the already contentious debate.

3.2.4 "Not-in-My-Backyard" Issues

The "not-in-my-backyard" (NIMBY) attitude is far from being a Japan-specific obstacle to national energy policy, but is one with significant long-term ramifications for the implementation of a successful nuclear power development program in Japan. Simply put, NIMBY refers to a grassroots movement to prevent the further construction and maintenance of nuclear power plants in local communities, where it is believed to represent critical environmental and safety hazards for the general public. A series of nuclear power accidents, mishaps, and scandals have further exacerbated the already volatile and emotional debate on the peaceful use of nuclear energy sources, thus eroding public support in Japan and reinforcing NIMBY sentiments. The accidents involved a sodium leak at the Monju fast breeder reactor (FBR), a fire at the Japan Nuclear Cycle Development Institute (JNC) waste facility connected with the Tokai reprocessing plant, and a 1999 criticality accident at Tokaimura. The latter accident, which claimed two lives and seriously injured three others, was the direct result of workers following an unauthorized procedures manual.

In 2002, an additional scandal—non-accident-related—erupted over an alleged cover-up of safety inspection procedures. Inspection of the shrouds and pumps around the nuclear reactor core, the responsibility of the electric power company, had been contracted out by the power company. In May 2002, questions emerged about data falsification and the significance of reactor shrouds and whether faults in the shrouds were reported to senior management. This incident further sullied the reputation of the incumbent electric power companies and the METI's designs for nuclear power development, in general.

4. CONCLUSION

Japan is a naturally resource-deficient island nation that relies heavily on imported fossil fuels for its energy needs. Stability in the international community continues to be a prerequisite for acquisition and maintenance of Japan's energy supplies. Especially in the case of oil, the fuel on which Japan is most dependent, securing access to stable supplies has been the mainstay of Tokyo's energy policy for over 30 years.

The two oil crises of the 1970s were set in motion by developments on the supply side, but if a third oil crisis should occur, the problem may be a strictly demand-side affair. World oil consumption, especially in Asian countries, continues to rise year after year, calling into question viable energy sources for future consumption. Japan has recognized the need to adopt a strategy that embraces both its imported energy dependency and its ability to offset such problems with indigenous natural forms of energy, such as nuclear, wind, solar, and geothermal power. In advancing the four Es—energy security, environmental protection, economic efficiency, and economic growth—the Japanese government endeavors to convince the Japanese public of the benefits of nuclear power, but with limited success. Nuclear accidents and scandals, both domestically and internationally, have tarnished nuclear power's image as a safe, viable alternative to imported fossil fuels. Moreover, NIMBY protests and the introduction of liberalization have added extra political and economic dimensions to this already contentious subject matter.

Deregulation of the energy industries (electricity, gas, and oil) will likely continue. Already, revisions to the basic laws have shown signs of material change. Prices in all three sectors have gradually fallen. Competition has led to partial consolidation in at least the petroleum industry, with prospects for the other sectors in the years to come. In conclusion, Japan's national energy policy is in a state of constant fluctuation and development. The foremost priority is stability of supply. Especially vital to this goal is maintaining an ample supply of oil and other imported fossil fuels needed to feed the world's second largest economy.

Acknowledgments

The author sincerely thanks the following people for reading an earlier draft of this article and/or for their many helpful suggestions throughout the course of this study: Dr. Chris Rowland and Dresdner Kleinwort Wasserstein (United Kingdom), Mr. Peter C. Evans (Massachusetts Institute of Technology, United States), and Dr. Yuki A. Honjo (JapanReview.net, Japan).

SEE ALSO THE FOLLOWING ARTICLES

European Union Energy Policy • National Energy Policy: Brazil • National Energy Policy: China • National Energy Policy: India • National Energy Policy: United States • National Security and Energy • Oil Price Volatility • World Environment Summits: The Role of Energy

Further Reading

Evans, P. (1997). "Japan's Deregulated Power Market: Taking Shape." A Cambridge Energy Research Associates (CERA) Global Power Forum Report. Cambridge Energy Research Associates, Cambridge.

Hein, L. E. (1990). "Fueling Growth: The Energy Revolution and Economic Policy in Postwar Japan." Harvard Univ. Press, Cambridge, Massachusetts.

International Energy Agency (IEA) (2002). "Energy Balances of OECD Countries." Organization for Economic Cooperation and Development, Paris. (CD-ROM)

Lesbirel, H. (1998). "NIMBY Politics in Japan: Energy Siting and the Management of Environmental Conflict." Cornell Univ. Press, Ithaca, New York.

Oyama, K. (1998). The policymaking process behind petroleum industry regulatory reform. *In* "Is Japan Really Changing Its Ways?: Regulatory Reform and the Japanese Economy" (L. Carlile and M. Tilton, Eds.), pp. 142–162. Brookings Institution Press, Washington, D.C.

Samuels, R. (1987). "The Business of the Japanese State: Energy Markets in Comparative and Historical Perspective." Cornell Univ. Press, Ithaca New York.

Scalise, P. (2001). "The Powers That Be: Japanese Electricity Deregulation." Dresdner Kleinwort Wasserstein, Tokyo.

National Energy Policy: United States

MIRANDA A. SCHREURS
University of Maryland
College Park, Maryland, United States

1. Energy Politics
2. George W. Bush's National Energy Policy

Glossary

Btu tax A tax on the heat or energy content of fuels. A British thermal unit (Btu) is defined as the amount of heat necessary to change the temperature of 1 lb of water at sea level by 1°F. The Broad-Based Energy Tax proposed in 1993 by the Clinton Administration would have placed a tax on energy based on its energy or Btu content. A gallon of diesel fuel, has more Btus than a gallon of liquefied natural gas. Thus, under the Clinton Administration proposal, fuels with a high energy content, which tend to be the dirty fossil fuels, would have been taxed at a higher rate than many alternative energies, which have lower average energy contents.

Corporate Average Fuel Economy (CAFE) Standard A miles per gallon standard established by law that manufacturers of cars and light trucks must obtain. In 2003, CAFE standards are set at a minimum of 27.5 miles per gallon averaged across a manufacturer's entire fleet. If a manufacturer does not meet the standard, it must pay a civil penalty of $5.00 for each 0.1 mile per gallon that the fleet does not obtain multiplied by the number of vehicles the manufacturer produces.

Energy Policy and Conservation Act (1975) This was one of the earliest laws enacted in the United States with the explicit purpose of regulating and reducing energy consumption. The Act also established a Strategic Petroleum Reserve in the Gulf of Mexico. During reauthorization, a national inventory of onshore energy sources was created and a home-heating oil reserve was established in New England.

Public Utilities Regulatory Policies Act (1978) This law was passed in the face of high energy prices in an effort to reduce dependence on foreign oil, to encourage energy efficiency, and to promote alternative energy sources. It required electric utilities to purchase power from independent companies that were able to produce power at a lower cost than what it would have cost the utility to generate the power. The Act is credited with the development of nonhydro renewable energies and has been especially favorable to the development of natural-gas-fired "cogeneration" plants that produce both electricity and steam.

Efforts to establish a national energy policy in the United States began under Jimmy Carter's presidency during the late 1970s, but then languished for over a decade. Since the early 1990s, there have been renewed attempts to establish a national energy policy. In 1992 a National Energy Strategy was formulated but U.S. dependence on imported energy continued to grow. Under the George W. Bush administration, there have been renewed efforts to establish a national energy policy, but sharp differences of opinion have prevented the passage of new energy legislation, at least for the time being.

1. ENERGY POLITICS

Understanding energy politics in the United States and efforts to develop a national energy policy requires an understanding of the different actors involved and their economic and political interests. There are widely divergent views regarding how to deal with America's large and growing appetite for energy and the pollution this produces.

The fossil fuel industries—the oil, coal, and natural gas producers—tend to support policies that favor the expansion of drilling and mining activities, including in protected lands and offshore. The automobile industry is typically opposed to legislation that mandates fuel efficiency improvements or raises gasoline taxes. The nuclear industry is eager to see a renewal in government support for nuclear

 © 2004 Elsevier Inc. All rights reserved.

energy. Environmental groups and the renewable energy industries tend to call for greater governmental support for energy conservation and renewable energy and legislation to control pollution from fossil fuel burning. Labor unions tend to be most concerned with the implications of governmental policies on jobs and consumer groups with the prices consumers will have to pay for energy. Each of these groups tends to frame energy issues in a different light, with some groups concerned more with economic security and quality of life concerns, others with national security interests, yet others with environmental protection, and some primarily with jobs.

Energy politics is also closely tied to the nature of the energy source, risk perceptions related to the use of the energy source, and the distribution of energy resources in the country.

Coal mining has been central to U.S. energy politics for well over a century and coal production is on the rise. However, although coal mining was a major employer in the past, the number of coal miners has steadily declined with time. Whereas in 1980 there were an estimated 220,000 coal miners in the country, in 2003 that number is down to approximately 100,000. Coal is a heavily regulated energy source because of the high safety risks for miners and the pollution emitted from coal burning as well as the damage that can be caused to lands from coal extraction.

Coal mining is heavily concentrated in Wyoming, West Virginia, and Kentucky, which together account for more than one-half of U.S. production of coal. Other major coal-producing states include Pennsylvania, Texas, Montana, and Illinois. Efforts to regulate coal production and coal burning in power plants for environmental reasons has resulted in considerable interstate politics, pitting producing states against downwind states suffering from acid rain.

Crude oil production is also heavily concentrated in a few states. Texas, Alaska, and California are the three largest producers, with Louisiana, Oklahoma, New Mexico, and Wyoming also being substantial producers.

Federal and Indian lands are of great interest to energy developers. Approximately 38% of total U.S. coal production was from Federal and Indian lands in 1999. The federal government owns the outer continental shelf, which is the territory that lies approximately 3 nautical miles from the shoreline and extends out 200 nautical miles. In 1998, the outer continental shelf (primarily in the Gulf of Mexico) was the source of close to 25% of domestic natural gas and 20% of crude oil production.

The politics surrounding nuclear energy are distinct from those regarding other energy sources. The future of nuclear energy remains highly uncertain and heavily debated. There are 103 licensed nuclear reactors operating in 65 plants in 31 states. In the 1960s and 1970s, the government provided the nuclear energy industry with substantial subsidies to offset the heavy initial investment required for new plant construction. The government also guaranteed the nuclear industry that it would develop a national nuclear waste depository. There are, however, many obstacles facing the nuclear energy industry. The accident at the Three Mile Island nuclear station in March 1979 that led to the permanent shutdown of reactor number 2 sent chills through the nation and intensified an already strong tendency for local communities and environmental groups to object to the building of new nuclear power plants. According to the report, "Nuclear Energy Policy," high construction costs are another serious problem for the industry; construction costs for reactors completed since the mid-1980s have been between $2 and $6 billion, or more than $3000 per kilowatt of electric-generating capacity (in 1997 dollars). Although no new nuclear power plants have been ordered since the Three Mile Island disaster and over 100 reactors have been canceled, 16 commercial reactors have received 20-year license extensions, with another 14 plants undergoing reviews for such extensions.

Over time, there has been some consensus established among stakeholders on the importance of energy efficiency improvements and energy conservation, but there is no real agreement on how the nation should deal with its large and growing appetite for energy. How much of a role should the government play in ensuring a stable energy supply? To what extent should government policy favor particular energy industries? Is deregulation of the electricity sector a good idea? What mix of energy sources should be achieved? To what extent should environmental considerations play a role in the nation's energy plans?

1.1 Energy Supply and Demand

The United States is the world's largest consumer of energy, accounting for 24% of total world energy consumption; it is a major importer of energy and is the world's largest source of greenhouse gas emissions. Approximately 86% of total fuel consumption is of fossil fuels. The United States is the world's second largest producer of coal after China and the second largest producer of natural gas after Russia. It also is a major producer of oil, representing 9% of

global production and nuclear energy, accounting for 31% of global nuclear electricity production.

Almost three decades after the 1973 oil embargo by the Organization of Petroleum Exporting Countries (OPEC), the United States remains highly dependent on energy imports. In 2001, the United States consumed 97 quadrillion British thermal units (Btu) of energy. Approximately 39% of this was oil, 24% natural gas, 23% coal, 8% nuclear, and 6% renewable energies (primarily hydro and biomass). In contrast, the United States produced approximately 72 quadrillion Btu of energy [33% coal, 28% natural gas, 21% oil, 11% nuclear, and 8% renewables (largely hydro and biomass)]. This means that approximately 30% of all energy consumed in the country is imported.

U.S. dependence on oil imports is especially large, at 53% of all oil consumed. Given that approximately 20% of oil imports are from the Persian Gulf and another 40% from OPEC, U.S. energy markets are very dependent on the politics of the Middle East, a highly volatile region. The United States has also become a net importer of natural gas.

Although there have been substantial improvements in energy efficiency over the past three decades, total energy consumption continues to rise as a result of a growing population and more energy-intensive lifestyles. Thus, although the U.S. economy has become approximately 60% more efficient in the past 30 years, per capita energy consumption levels in the United States are among the highest in the world. On average, an American uses 342 million Btu of energy per year, or almost twice as much as a Japanese (174 million Btu) or a German (172 million Btu). Moreover, as the U.S. population is expected to grow substantially in the next decades due to a combination of immigration and a relatively high birth rate compared with many other advanced industrialized states, energy demand is expected to continue to rise.

1.2 Energy Planning in U.S. History

Unlike many other advanced industrialized democracies, the United States does not produce regular multiyear national energy plans. Historically, the lack of interest in long-term energy planning in the United States stems both from a relative abundance of energy and from a fairly strong tendency toward neoliberal economics. Prior to the 1970s, energy policy change tended to be incremental and largely reactive.

Through the late 1950s, the United States produced at least as much energy as it consumed. In contrast with many countries in Europe, which nationalized energy industries in the first half of the 20th century, in the U.S. energy production remained in private hands and little consideration was given to long-term energy supply.

In the 1960s, energy consumption began to surpass energy production and by the 1970s, the United States had become a major importer of energy, and especially petroleum. Over the course of the 1970s, in response to the nation's severe environmental problems and sharp increases in energy prices, the government became increasingly involved in regulating the energy industry, promoting energy conservation, and providing incentive schemes for the targeted development of specific energy sources.

1.2.1 The Impact of the OPEC Oil Embargoes

The 1973 oil embargo by OPEC sent world oil prices soaring. In reaction to the embargo, in 1975 at Gerald Ford's urging, Congress passed the Energy Policy and Conservation Act, which established Corporate Average Fuel Economy (CAFE) standards for automobiles, extended domestic oil price controls, and created the Strategic Petroleum Reserve, an oil stockpile for emergency situations. Two years later, Congress created the Department of Energy (DOE).

Jimmy Carter was convinced that the nation's energy security demanded a comprehensive energy plan. Carter's 1978 National Energy Plan was the first attempt by a president to establish a national energy policy, one that called for both expanded production of coal and enhanced energy conservation. The plan included numerous measures for the promotion of renewable energies, provisions for energy conservation, and energy taxes. It also led to the establishment of the Public Utilities Regulatory Act, which required utilities to purchase energy from "qualifying facilities," effectively ending the electric utility monopoly on electricity production and helping to foster a market for renewable energy sources.

1.2.2 The Ronald Reagan Years

Many of the policy goals laid down in Carter's National Energy Plan, however, came under attack during the years of the Reagan administration. Reagan was eager to down-size government and remove many of the regulations he felt were burdening the energy industry. Although he failed in his effort to close the Department of Energy, he succeeded in winning Congressional support to rescind tax breaks for energy-saving devices and decreased government funding for research and development of renewable energy sources and energy conservation initiatives.

1.2.3 George H. W. Bush's National Energy Strategy and Energy Policy Act

Although George H. W. Bush had a political outlook similar to that of Ronald Reagan, like Carter he did see the need for the creation of a national energy plan. Thus, in 1989 he ordered the DOE to prepare a National Energy Strategy (NES). The NES was a response to growing concerns with global warming and rising energy prices in the late 1980s and early 1990s. The plan was criticized by environmentalists, however, because it included plans for oil drilling in the Arctic National Wildlife Refuge (ANWR) and did little to reduce U.S. carbon dioxide emissions. Indeed, a decade later these same issues remain two of the major points of contention between Republicans and Democrats in the formulation of a national energy plan.

The NES led to the adoption by Congress in 1992 of the Energy Policy Act, which began the opening of the electric utility industry to competition, established energy efficiency standards for federal facilities, mandated that the federal government replace some of its fleet with alternative fuel vehicles, established the (unmet) goal of having 10% of the nation's gasoline use be replaced by alternative fuels (as of the year 2000, only 3.6% of the nation's highway gasoline use had been replaced), called for voluntary reporting of greenhouse gas emissions and the establishment of a national inventory of greenhouse gases, and initiated measures to aid the nuclear energy industry (including a revision of the regulatory framework and environmental standards governing nuclear waste disposal at Yucca Mountain in Nevada).

Although the 1992 National Energy Policy Act was the most important energy legislation to be passed in over a decade, it failed to do much to move the nation toward greater energy independence or to do much to promote a more sustainable energy future.

1.2.4 William J. Clinton's Failed Btu Tax

Little progress in these directions was made under the Clinton administration, either. Clinton and Vice President Al Gore proposed a tax on the heat content of fuels (a Btu tax) in an effort to cut energy consumption (and thereby reduce dependence on imported oil) and to reduce greenhouse gas emissions. The Btu tax, however, was rejected by Congress in favor of a 4.3 cent per gallon increase in the federal gas tax, which, because of the historically low price of gasoline at the time, was politically palatable. The Senate also made known its lack of intention to ratify the Kyoto Protocol that the Clinton administration formally signed in 1998. Had the Kyoto Protocol been ratified by the Senate, major efforts at energy conservation and the establishment of a carbon emissions trading system would have been necessary.

2. GEORGE W. BUSH'S NATIONAL ENERGY POLICY

Soon after George W. Bush entered office, California was hit by rolling electricity blackouts and sharp increases in electricity prices. In reaction to the California "energy crisis" and fulfilling a campaign pledge, shortly after taking office Bush announced his intention to establish a national energy plan. Vice President Richard Cheney was given the task of leading an energy task force that was charged with drafting a plan. After 3 months of closed-door meetings and with considerable fanfare, in May 2001 the energy task force released its National Energy Policy (NEP) report, "Reliable, Affordable, and Environmentally Sound Energy for America's Future." The report and pursuant energy bills have been the subject of heated Congressional debate and a number of lawsuits.

The NEP report suggests that the nation is facing the most serious energy crisis since the oil embargoes of the 1970s. Moreover, in the coming decades, unless action is taken, "projected energy needs will far outstrip expected levels of production. This imbalance…if allowed to continue, will inevitably undermine our economy, our standard of living, and our national security." To deal with the crisis and future energy needs, the report calls for promoting energy conservation, repairing and modernizing the energy infrastructure, and increasing energy supplies.

Perhaps the strongest theme to come out of the NEP report is the need to increase the nation's energy supply. The NEP calls for oil drilling in ANWR, the promotion of clean-coal technology, nuclear energy development, and natural gas exploration. It also makes mention of the need to encourage renewable energy development, but suggests that renewable energies are unlikely to make a large dent in U.S. foreign energy dependence. A second theme of the report is the potential to make major gains in energy efficiency and new energy sources through technological developments. The report also urges a modernization and expansion of the nation's aging energy infrastructure (oil and natural gas pipelines, refinery capacity, and electricity transmission grids).

In total, the report included 105 recommendations, including many that called for cooperation

with foreign governments to improve the environment for energy investment and to improve the stability and security of supply.

The release of the NEP has had a mixed reception. Conservatives like it. The Heritage Foundation's Senior Policy Analyst for Energy and Environment calls the plan "a step in the right direction" because it calls for meeting the nation's energy needs through developing as yet untapped domestic energy sources, removing interstate transmission restrictions, modernizing energy delivery systems, promoting nuclear energy (which does not produce greenhouse gas emissions) and clean coal technology, and removing regulatory burdens that create market inefficiencies.

In contrast, environmentalists have been highly critical of the NEP. Greenpeace, for example, argues that the NEP is leading the nation "down the wrong road" because the plan fails to take steps to reduce greenhouse gas emissions, calls for 1300 new fossil fuel and nuclear power plants, favors oil extraction in ecologically sensitive areas (including ANWR and the Rocky Mountains), and will use taxpayer subsidies for the nuclear energy and fossil fuel industries.

Several lawsuits have been brought against the Vice President's office as well in relation to the closed-door process by which the NEP was drafted. The National Resources Defense Council (NRDC) sued the Department of Energy, a key member of the task force, under the Freedom of Information Act, for the release of thousands of pages of documents used by the task force in formulating the plan. Although the DOE provided some 13,000 pages that the NRDC has since made available on the Internet for public scrutiny, thousands of additional pages of documents were not released. In February 2002, the NRDC won a District of Columbia court motion requiring the DOE to expedite the release of the remaining documents.

Judicial Watch and the Sierra Club are suing the Vice President's office for the release of additional documents, task force minutes, and computer records that would shed light on who advised the task force and how this may have influenced the document's development. Their lawsuits are motivated by the concern that private industry representatives essentially functioned as members of the advisory group, and under the law, this would mean that the group's deliberations should be made open for public scrutiny. The collapse of Enron in early 2002 and the revelations that Enron had tried to manipulate California's electricity markets for profit have also played into these lawsuits since Cheney met with Enron's Chief Executive Officer, Kenneth L. Lay, several times while he was heading up the task force. Cheney's office has sought unsuccessfully to have the cases dismissed.

The terrorist attacks on the World Trade Center and the Pentagon on September 11, 2001, and the subsequent U.S. decision to go to war in Iraq, for a time shifted government and public attention to national security concerns.

Events in the Middle East did, however, feed into energy debates and renewed concern about U.S. oil dependence on the Middle East. They strengthened national sentiment on the need to reduce U.S. dependence on imported oil. But clear partisan differences remained on how greater energy independence should be achieved. The Republicans have tended to favor greater domestic production, including production of nuclear energy, whereas moderate Republicans have at times joined Democrats in calling for more energy conservation and support for nonnuclear renewable energies.

Thus, although recognizing the need for some kind of national energy policy plan, the 107th Congress failed to pass comprehensive national energy legislation even though both houses had passed energy bills. Major differences in the Democratic-controlled House and Republican-controlled Senate versions of energy legislation could not be bridged and, thus, Congress ended its session without passing energy legislation. The November 2002 elections returned control of the Senate to the Republicans and strengthened the position of Republicans in the House.

Despite the Republicans' strong showing in the 2002 election and high voter-approval ratings, the Bush administration has a poor environmental image. The administration has done little to win the confidence of environmentalists with its rejection of the Kyoto Protocol, its failure to regulate carbon dioxide emissions or to set more stringent fuel efficiency standards, and its efforts to open protected federal lands to mining, oil and natural gas drilling, logging, and recreational purposes. An August 23, 2003 public opinion poll found that of the 1011 adults surveyed by telephone 53% preferred the Democrats' approach to the environment and 29% preferred Bush's (the respective figures in January 2002 were 43% and 38%). Similarly, 42% of Americans said they preferred the Democrat's approach to energy compared with 33% who said Bush was doing a better job (the respective figures in January 2002 were reversed, with 33% favoring the Democrats and 46% favoring Bush's approach).

In a bid to improve its image on environmental and energy issues, the Bush administration has begun to focus the domestic energy debate on hydrogen fuel development. Hydrogen is being billed as a pollution-free fuel since the energy-generating chemical reaction between hydrogen and oxygen produces only water as a waste product. Hydrogen production, however, requires energy from other energy sources. Natural gas is the most affordable means of producing hydrogen although its cost is still four times as high as the cost of producing gasoline. The administration's hope is that, through technology development, it will be possible to produce hydrogen more affordably, including with the use of coal and nuclear power.

In his January 2003 State of the Union address, President Bush announced his intentions to fund hydrogen fuel technology research and development. The administration targeted $1.7 billion to be distributed over the subsequent 5 years for the Freedom CAR and Fuel Initiatives, public–private cooperative endeavors for the development of hydrogen fuel cells, hydrogen fuel cell-powered cars, and hydrogen infrastructure. The administration's stated goal is to have hydrogen-powered vehicles and a fuel distribution network in place by 2020.

The initiative has won praise, especially from Republicans and Democrats in states with automobile manufacturers, but has earned more mixed reviews from Democrats and environmentalists who have criticized the administration for placing so much emphasis on a still unproven energy source and providing relatively little investment for available renewable energy technologies. Concern has also been raised that in order to develop hydrogen it will be necessary to build more coal-burning fossil fuel plants and nuclear power facilities.

The 108th Congress took up the energy legislative debate that the 107th Congress failed to complete. The House of Representatives was the first to act. In April 2003, the House of Representatives passed H.R. 6 on a vote of 247 to 175. The bill includes approximately $19 billion in energy-related industry tax incentives over a 10-year period for alternative fuels, energy efficiency, electricity restructuring, and oil and gas production. The bill provides tax breaks to oil and gas companies to encourage production from marginal wells and offshore drilling, provides millions of dollars for clean coal technology, and permits utilities to more easily write off the cost of new transmission systems. The bill also grants tax credits for solar and wind power and encourages energy-efficiency improvements in homes.

On a vote of 228 to 197, the House killed a proposed amendment that would have removed the provision for oil drilling on 2000 acres of the Arctic National Wildlife Refuge. Moderate Republicans joined Democrats in calling the move ill-conceived, arguing that the Arctic National Wildlife Refuge should be protected as one of the nation's most important pristine areas and that a more effective approach would be to save oil through the introduction of higher automobile fuel efficiency standards. In an earlier vote, however, the House defeated a provision that would have required a 5% reduction in automotive fuel use by 2010 (or an average fuel efficiency standard of approximately 30 miles per gallon). The justification given by opponents of higher auto mileage standards was that it would be bad for the economy as it would make it harder for manufacturers to produce popular sports utility vehicles, would result in layoffs, and would pose safety hazards because the standards would require the production of smaller cars.

Democrats in the House also failed in their efforts to rewrite electricity rules to protect consumers from market manipulation that could lead to huge price swings and ban methyl tertiary butyl ether as a fuel additive despite its role as a potential source of groundwater pollution. The Natural Resources Defense Council has criticized the House bill for providing huge subsidies to the coal, natural gas, and nuclear industries but providing relatively few incentives for less polluting industries.

Attention then shifted to the Senate, which began voting on measures related to its national energy bill. Repeating events in the House of Representatives 3 months earlier, Senate Democrats failed in their effort to win support for a proposal for a 40-mile-per-gallon fuel-economy standard for passenger cars by 2015. Instead, the Republican majority voted to direct the Department of Transportation to work on a new standard, taking into consideration the impact it would have on jobs and consumer safety. The current CAFE standard of 27.5 miles per gallon has not been changed since the 1986 model year.

Deadlock in the Senate on a number of other energy issues was finally broken on July 31, 2003, when the Senate Republicans in a surprise move agreed to a suggestion by Senate Minority Leader Thomas A. Daschle that last year's energy bill (that was negotiated by a Democratic-controlled Senate) should be resurrected. The Republican majority agreed to this proposal, knowing that they would be able to write new provisions into it in the coming months. The Democrats extracted a promise that

separate votes would be taken in 2003 on climate change control and measures to give federal regulators greater authority to oversee utility mergers. In accepting the 2002 Senate Bill, the Republicans gave up on provisions for loan guarantees for new nuclear power plants.

The sense of urgency behind the need for a national energy policy was rekindled by the August 2003 electricity blackout that left New York City and large sections of the Northeast sitting in the dark. The primary cause of the blackout was determined to be an antiquated electricity distribution infrastructure.

In the following months in largely closed-door conferences, Republican leaders of the House and Senate met to try to iron out differences between the House and Senate versions of energy legislation. The controversial House decision to drill in ANWR was dropped from the bill by the Republican leadership after it became clear that the bill stood no chance of being passed in the Senate if it were included, but other controversial elements remained.

The bill proposed by the conferees included $23 billion in tax incentives primarily for coal mining, oil and exploration, the construction of new transmission lines and power plants, and the building of a natural gas pipeline from Alaska to the Midwest. It also included tax incentives for wind power and biodiesel fuel made from soybeans and a doubling of ethanol fuel mandates. While the House of Representatives passed the bill, it was blocked by a coalition of 32 Democrats, 7 Republicans (mostly from Northeastern states), and one Independent in the Senate. The bill was supported oddly enough by labour unions, energy companies, and renewable energy producers. It was opposed, however, by those concerned with its large price tag (estimated to be over $30 billion over 10 years). Senators from Northeastern states also objected to the incentives provided for Midwestern coal-fired utility plants because the pollution from these plants affects them. There was also bipartisan opposition to the provision that would have exempted manufacturers of methyl tertiary-butyl ether from product liability lawsuits.

Thus, although there was great demand for new energy legislation, strong differences between Democrats and Republicans, between Midwestern agricultural and mining states and Northeastern states, and between energy producers and environmentalists prevented passage of new energy legislation prior to the adjourning of the 108th Congress for the holiday break.

SEE ALSO THE FOLLOWING ARTICLES

Energy Development on Public Land in the United States • European Union Energy Policy • Fuel Economy Initiatives: International Comparisons • National Energy Policy: Brazil • National Energy Policy: China • National Energy Policy: India • National Energy Policy: Japan • National Security and Energy • Renewable Energy in the United States • Strategic Petroleum Reserves

Further Reading

Alliance to Save Energy. A nonprofit coalition of business, government, environmental, and consumer leaders: http://www.ase.org.

American Council for an Energy Efficient Economy. A nonprofit organization dedicated to energy efficiency: http://www.aceee.org. Chubb, J. E. (1983). "Interest Groups and the Bureaucracy: The Politics of Energy." Stanford University Press, Palo Alto, CA.

Davis, D. H. (1993). "Energy Politics." St. Martin's Press, New York.

Department of Energy (DOE). Created in 1978 and has as one of its missions the protection of national and economic security by providing a diverse and reliable supply of energy. Also responsible for nuclear energy: http://www.energy.gov.

Duffy, R. J. (1997). "Nuclear Politics in America: A History and Theory of Government Regulation." University of Kentucky Press, Lexington, KY.

Energy Information Administration. A statistical agency of the Department of Energy that provides energy data, forecasts, and analyses: http://www.eia.doe.gov.

House Committee on Energy and Commerce. The House of Representatives' committee that is responsible for the supply and delivery of energy, among many other issues. It also has jurisdiction over the Department of Energy: http://energycommerce.house.gov.

Ikenberry, G. J. (1988). "Reasons of State: Oil Politics and the Capacities of American Government." Cornell University Press, Ithaca, NY.

Jasper, J. M. (2000). "Nuclear Politics: Energy and the State in the United States, Sweden, and France." Princeton University Press, Princeton, NJ.

Kraft, M. E. (2003). "Environmental Policy and Politics." 3rd ed. Pearson Longman, New York.

Landsberg, H. H. (1993). "Making National Energy Policy." Resources for the Future, Washington, DC.

Natural Resources Defense Council (NRDC). An influential environmental action organization active on energy and environmental issues: http://www.nrdc.org/

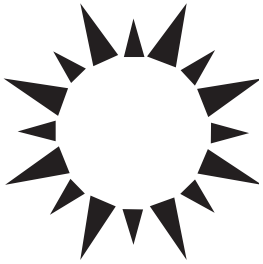
Senate Committee on Energy and Natural Resources. The Senate committee that deals, among other policy issues, with energy resources and development, including regulation, conservation, strategic petroleum reserves, and appliance standards; nuclear energy; surface mining; and federal coal, oil, and gas: http://energy.senate.gov/.

Stagliano, V. A. (2001). "A Policy of Discontent: The Making of a National Energy Strategy." Pennwell, Tulsa, OK.

Tugwell, F. (1988). "The Energy Crisis and the American Political Economy: Politics and Markets in the Management of Natural Resources." Stanford University Press, Palo Alto, CA.

Nationalism and Oil

VÍCTOR RODRÍGUEZ-PADILLA
National Autonomous University of Mexico
Mexico City, Mexico



1. Introduction
2. The Rise and Peak of Nationalism: 1920–1980
3. Pragmatism Substitutes Nationalism: 1981–2003

Glossary

economic rent What remains of revenue after deducting the production costs of a mineral deposit, including normal return (mean industrial return) on the required capital.

expropriation Obligatory transfer of the property of a particular party to state administration for the public interest, normally in exchange for an indemnity payment.

nationalization Authoritative transfer of private sector means of production to national collective ownership represented by the state for reasons of national security or public interest.

public monopoly Exclusive right of the state to undertake an economic activity such as oil or natural gas extraction.

sovereignty Abstract principle that designates the instance that detains legitimate authority, the only authority with the capacity to enact norms; the international community recognizes that sovereignty over natural resources in inherent to the nation and is exercised by its representative, the state.

state interventionism A growing and direct process of intervention by the state in the economy or a sector of the economy, acquiring property rights and undertaking business activities and productive activity.

Nationalism is a doctrine of a people under foreign domination who try to free themselves and form a sovereign nation. The oil industry evokes such doctrine—a historical phenomenon, a process, a value, a feeling, and a vision of a country. As a doctrine, nationalism claims the nation's ownership of its natural resources, the state's valuation of such resources, and the subordination of the management of reserves and production by development policies. It is a dogma fed on socialist, communist, anti-imperialist, and anti-capitalist ideas that have been expanded here and there, imitating the achievements made by other countries.

1. INTRODUCTION

The term "nationalism" is commonly used to describe state interventionism that permeated the oil industry since the beginning of the 20th century. It was born in Latin America with the first concessions and was generalized in all continents during the postwar period and culminated by "overturning" the Organization of Petroleum Exporting Countries (OPEC) during the 1970s. The nationalist intervention took on various forms. Taken to its ultimate consequences, it led to the nationalization of oil property and activities and finally to a state-owned monopoly.

In its character of patriotic value, nationalism means defense of sovereignty. Territorial sovereignty, one of the essential bases of nation-state, becomes extendible to the subsoil and, therefore, to the resources contained within. In certain countries, ownership and the state-owned monopolies over oil and natural gas have even been indispensable elements in the consolidation of the nation-states themselves.

Equally, nationalism translates a sentiment against an external action. It evokes the struggle for the liberalization of oil, of which greedy foreign companies have taken control. Just as the imperialist practices of the superpowers awoke the nationalism of oppressed populations, the predatory practices of the multinationals aroused oil nationalism. It was a reply to the plundering, the arrogance, and the arbitrariness of the concession system. And this sentiment has tended to prevail.

 © 2004 Elsevier Inc. All rights reserved.

However, nationalism also gave birth to independent development following a colonial past. To put oil at the service of development through ownership of the resources, a greater share of income and gradual control of related activities was the fundamental objective proposed by oil-producing countries. That nationalist vision contained a mix of values, interests, and aspirations that converted it into a vital challenge, an image that strongly took root in society.

Contrary to popular belief, nationalism does not exist per se. It emerges when the conditions are favorable, when the state cannot exercise sovereignty over its natural resources, and/or when foreign interests threaten that original, inalienable, and imprescriptible right. However, its achievements will be few if the context does not favor the state's intervention in the economy.

Because of these objectives, nationalism has been one of the political factors that have contributed the most to the structural transformations of the international oil industry. It called a halt to the concession system and performed a crucial role in recovering the ownership of natural resources and the administration of complete sectors of the oil industry by oil-producing countries. Following its peak during the 1970s, the nationalist fervor abated but did not disappear altogether. In a worldwide context characterized by the stagnation of oil prices, economic liberalism, and the end of the cold war, the international oil companies have returned to the oil-producing countries, but not on the same terms as before. Nationalism is currently lethargic but will reemerge if the circumstances change and become favorable once again.

2. THE RISE AND PEAK OF NATIONALISM: 1920–1980

History repeats itself from one country to another. A few years after oil exploitation began by multinational companies, a social discontentment arose and grew to become a national movement of rejection. And although the nationalist struggles took on different forms and accomplished diverse goals, they shared the same driving force: the reclaiming of the permanent sovereignty of the nation's natural resources. In 1917, Russia was nationalized without any indemnity whatsoever, but this related to a different logic: the assumption of control by the state of all means of production, not just the oil industry.

2.1 The Nationalist Movements

Nationalism was born in Latin America at the beginning of the 20th century. From there, it radiated all over the world. Historical reasons explain that course. On the one hand, it was the natural expansion zone for the North American oil industry; on the other hand, it was the zone that had achieved greater political consciousness and institutional maturity because it consisted of countries that had ceased to be colonies decades earlier. The savage capitalism of the concession system (Table I) and a heightened political consciousness conjugated to breathe life into the nationalist phenomenon.

Argentina created the first nationalized company in 1923, and Uruguay created the first state-owned refinery in 1931. The Bolivian government nationalized the assets of Standard Oil of New Jersey in 1937 after it was discovered that the company had played a dual role in the Chaco War of 1932–1935, a bloody conflict between Bolivia and Paraguay that was instigated and backed by Standard Oil and its rival, Royal Dutch Shell.

Lázaro Cárdenas expropriated in 1938 following failed attempts to control and put the oil industry at the service of internal development through laws and regulations. The 1917 constitution had established the state-owned ownership of oilfields but maintained the concession system. Standard Oil and Shell took advantage of this situation to divide up the territory, plunder the subsoil, and make Mexico the second-largest oil exporter in 1920 without a single tangible benefit for the country. The companies, although paid off, organized a blockade against Mexican crude and the national oil company (PEMEX).

At the end of the fierce dictatorship that turned the country into a vast petroleum field dominated by Shell, Esso, and Gulf, Venezuela changed the law in 1943 to recover a part of the oilfields and established a less unfair distribution of the oil benefits. But it was not until 1948 that Juan Pablo Pérez Alfonso managed to impose the "50/50" principle on the oil companies. A year later, such a principle was being applied in the Middle East oil-producing countries.

In Iran, the Anglo–Iranian Company rejected that distribution. In response, the Mossadegh government nationalized oil in 1951 and created the National Iranian Oil Company (NIOC). The International Court of Justice at The Hague, Netherlands, ruled in the Mossadegh government's favor, but the process was not completed due to a military coup d'êtat that ousted the government in 1952. The new authorities ratified state ownership of the oilfields and the

TABLE I

Inheritance from the Old Colonial Regime

The Concession System

Using skill, influence, and the support of its country of origin, an oil company obtained from the government of an underdeveloped country the concession for a vast territory to carry out oil exploration operations for 50 years or more. Frequently, it was the men close to power who obtained the concessions, but sooner or later, they were transferred to the oil companies. In the end, the majority of national territory was under concession.

The concession implied the exclusive right to explore, extract, and export the product so that no other investor could do the same in the zone indicated in the mining deed. The concessionaires organized and controlled the operations to suit their own criteria and interests. They decided the volume and nature of the investments, working areas, exploration plans, oilfields to be exploited, production capacity, extraction rates, and exportation volumes. The criteria used were linked to agreements between the oil companies and, in the last instance, to the evolution of a world demand broadly managed by the large multinationals. The role of the state was reduced to that of simple tax collector.

From there arose the dual economic structure. The oil industry was isolated from the rest of the economy. Its evolution depended on the behavior of world demand. From there, some oil areas suffered from overexploitation, and others suffered from stagnation. Also, the oil companies could monopolize very cheap oil to resell it at high prices. The main beneficiary was not the owner but rather the entity that extracted, exported, transformed, and sold the oil as products. The concessionaire ensured the highest benefits without the slightest obligation to satisfy the demands of economic and social development, neither of the country nor of the oil sector itself. In summary, the companies managed a geological heritage that was not theirs to their own convenience. They exercised sovereignty over oil resources of the country that had opened its doors to them. They usurped a right that was not theirs.

Another equally important problem was the manner in which the companies operated the concessions. With the complicity of weak or corrupt governments, they turned the oil areas into regions with their own laws, authorities, and police forces. They created a state within a state. Communities and farmers stripped of their land, workers in misery, accidents of gigantic proportions, destruction of oil fields, the accelerated wastage and exhaustion of reserves, enormous fortunes in the hands of a few, and interference in the internal affairs of the country were some of the concession system. It was not for nothing that this is historically considered as one of the most savage forms of capitalism. The unjust nature of such a system awoke nationalism, which caused its collapse.

Agreements between Large Companies

In 1928, Anglo–Iranian (BP), Royal Dutch Shell, Compagnie Française des Pétroles (Total), Standard Oil of New Jersey (Exxon), and Mobil Oil agreed, through the Red Line Agreement, to jointly penetrate the old frontiers of the Ottoman Empire (Turkey, Iraq, Syria, and the Arab peninsula) with the exception of Kuwait. Each company would make its own investments, independently of the others, in refining, distribution, and marketing. That same year, the three largest companies (Standard Anglo–Iranian, and Shell) signed the Achnacarry Agreement to conserve the market portions, regularize production, and determine world oil price, Mobil Gulf, Texaco, and Chevron joined the pact. These and other subsequent agreements allowed the "Seven Sisters Cartel" and Total to exercise total control over the extraction and exportation of Middle East oil. However, because the collusion among companies limited the aspirations of putting oil at the service of national development, the producing countries rebelled.

existence of the NIOC but maintained the concession system. That nationalization, aborted by the intervention of the International Oil Cartel and the large capitalist powers, had a double effect of dissuading other countries from following its example but confirming the legitimacy of the struggle. It propagated the idea that the state should ensure absolute control of oil operations through nationalization or negotiated agreements. The Iranian experience represented a real basis for the subsequent political changes in the region.

Nationalism gained strength during the 1960s. Certain Latin American countries created upstream monopolies but respected existing rights (e.g., Brazil, Argentina, Chile), whereas others nationalized (e.g., Cuba, Peru, Bolivia). The infection traveled as far as India and Syria. In parallel, numerous national companies were created in Venezuela, Kuwait, Saudi Arabia, Argelia, Iraq, and Libya, some of which were strengthened by technical assistance from the former Soviet Union and from companies on the margin of the cartel.

A decisive factor in the progress of nationalism was the creation of OPEC in 1960 by Iran, Iraq, Kuwait, Saudi Arabia, and Venezuela at a meeting in Baghdad, Iraq. These were later joined by Qatar in 1969, Indonesia and Libya in 1962, Abu Dhabi in 1967, Argelia in 1969, Nigeria in 1971, Ecuador in 1973, and Gabon in 1974. Born to combat the reduction in posted prices, unilaterally fixed by the oil companies, the cartel was not only dedicated to that. It encouraged the adoption of a single tax treatment in member countries, the recovery of ownership rights, and state control of production. In 1966, OPEC pronounced the principle of the countries' sovereignty over hydrocarbons. In 1968, it

pronounced the principle of national participation in the concessions. The organization's support turned out to be crucial when one of its members took action against the consortia. The example prospered. In 1965, the Latin American State Oil Reciprocal Assistance Organization was created. In 1968, the Organization of Arab Petroleum Exporting Countries (OAPEC) was formed.

Another fundamental factor in the rise of nationalism was the peak in the liberation and decolonization process. A total of 45 countries in Asia, Africa, and Oceania gained independence during the 1960s decade.

The movement of the nonaligned countries contributed to maintaining nationalist feelings on a high—on being constantly present against the political blockades and foreign domination, on the one hand, and for each country's right to liberty, independence, and autonomous development, on the other. However, the persistence of poverty was a reminder that access to independence was not enough to ensure progress. Economic independence must be conquered until a standard of living in accordance with the most elementary rules of human dignity is achieved. Developing countries wanted to make use of their wealth, organize their development, and cooperate on a basis of equality and reciprocal interest. From there arose, in particular, a movement ever more important in the rejection of the control of their natural resources by foreign companies and of reclaiming their national sovereignty. This discussion gained strength at international forums.

Two additional factors encouraged developing countries' aspirations of freeing themselves from the control of the large multinational companies. The first was the appearance and strengthening of new investors—the national companies of the industrialized nations such as France, Italy, Belgium, and Japan and the "independent" U.S. companies such as Occidental, Amoco, Conoco, Getty, and Arco—that were prepared to accept a share of the benefits more in favor of the producing countries with more restricted access to the oilfields as well as offering technical and commercial assistance to the national companies so as to gain certain permits and guarantee cheaper sources of supply. The second factor was the appearance of new production zones with high oil potential such as Libya and Nigeria. This led to fierce competition among companies, and the developing countries used this competition to their advantage to impose their conditions.

Nationalism reached its peak during the 1970s, not only as a consequence of a historical process but also due to a series of favorable circumstances. On the economic plane were tensions regarding the availability of oil as a result of transportation problems, increased global production costs associated with the development of high-cost oilfields (e.g., the North Sea, Alaska) necessary to respond to the rapid growth in demand, and the need for higher international process to correct imbalances in the U.S. supply system. On the political front were the rise to power or consolidation of nationalist governments and a new Arab–Israeli war. Taking advantage of the political climate, the OPEC countries rebelled against the companies. Advances in one country were reproduced in the others, dealing simultaneously with ownership rights, the distribution of benefits, and administration of oil operations.

In 1970, Libya managed to increase taxes and posted prices. Iran and Kuwait copied the measure. Argelia went further by unilaterally fixing such prices. These events set the criteria to begin negotiations between OPEC and the companies, resulting in the Tehran (1971), Tripoli (1971), and Geneva (1972–1973) agreements. The most important achievement was that the determination of prices ceased to be a decision made only by the oil companies. From that point onward, the OPEC countries intervened in such decisions.

Argelia nationalized the interests of six companies and 51% of the French concessions in 1970. The following year, Libya nationalized BP's assets. In 1972, Iraq nationalized the Iraq Petroleum Company, Ecuador nationalized its oil, and Libya acquired 50% of ENI's assets. In 1973, Saudi Arabia's share in the ARAMCO concessions took effect; from an initial value of 25%, it would reach 51% by 1982. The shah of Iran and the companies agreed to immediate nationalization of all the assets of the Anglo–Iranian Company in exchange for a supply contract guaranteed for a 20-year period. Libya nationalized 51% of the remaining concessions. Nigeria acquired a 35% share of BP–Shell assets.

On October 6, 1973, the Yon Kippur War broke out. Ten days later, OAPEC members reduced their exports and declared an embargo against the United States and the low countries for having assisted Israel against Egypt and Syria. Weeks later, the exporting countries started to fix sovereignty and posted prices without consulting the oil companies. By the end of the year, Iraq had nationalized Exxon's and Mobil's equity in the Basrah Petroleum Company. The following year saw the full nationalization of Iraqi oil.

In 1974, the General Assembly of the United Nations recognized the right to nationalize or

transfer to state the control and execution of mining projects. The Declaration and Action Program for the Establishment of a New International Economic Order stipulated the following:

> *In order to safeguard its resources, each state is entitled to exercise effective control over them and their exploitation with means suitable to its own situation, including the right to nationalize or transfer ownership of such resources to its nationals, this right being an expression of the full permanent sovereignty of the state. No state may be subjected to economic, political, or any type of coercion to prevent the free full exercise of this inalienable right.*

In 1974, Nigeria took a 55% share of all concessions. Kuwait, Qatar, the United Arab Emirates, and Saudi Arabia fixed their shares at 60%. The Saudis also increased the tax rate to 85% and fixed royalties at 20%. In 1975, Venezuela nationalized and Kuwait acquired all concessions, with the oil companies receiving compensation on the value of the assets and a long-term trading agreement with certain advantages in exchange for technological and transportation services. This formula served as an example to other countries.

2.2 From Nationalism to the Application of the Principle of Sovereignty

The "overturn of OPEC" encouraged nationalism in the Third World. The world's perception of ever more scarce and more expensive oil played in its favor. The rhythm and intensity with which each country imitated the organization depended on political will; the technical, human, and financial capacity of the public sector to develop oil industry activities; the geological potential; the level of production and importance of demand; the weight of the oil bill; external indebtedness; and the shortage of currency. These factors determined the goals and margin for maneuver of the state when faced by the multinationals.

Nationalist efforts were focused in several directions: the recovery of ownership rights, state administration of activities considered as strategic, the creation or strengthening of national companies, the substitution of concessions for the system of contracts, and tightening up on taxes imposed on companies with regard to access, operation, and distribution of benefits.

2.2.1 Recovery of Ownership Rights

Many countries substituted the royal prerogative system for the ownership system. In the first place, the mineral resources, before being discovered, belonged to nobody. The state used its royal prerogative power, formerly attributable to the king, to concede the right of use, determine the general conditions of the search and exploitation of the oilfields, and designate the right to mine or extract to the company of its choice. In the second place, under the ownership system, the oilfields are the property of the state, which may entrust their exploitation to a state-owned monopoly or grant a contract to a third party, establishing the most convenient terms and conditions on a case-by-case basis. If the resource is already being exploited, the state has the right to nationalize it. The third possibility, the occupation system, postulates that the owner of the soil or the surface is also the owner of the subsoil. This did not echo among the developing countries as being incompatible with history, politics, and the ideology dominant at that time.

2.2.2 Direct Control of Strategic Activities

Certain nations imposed state exclusivity on exploration, development, production, transportation, refining, processing, distribution, marketing, exportation, and importation of crude oil, natural gas, gasoline, and petrochemical products. Others assigned exclusive rights to exploration and production but left the other segments open. Still others limited exclusivity to crude oil imports. The remainder of the countries decided to promote competition in all of the links in the chain.

2.2.3 Incorporation or Strengthening of National Companies

One of the nationalist goals most appreciated by the producing countries was the state's direct intervention in oil production activities in two ways: (1) as an investor (through a share in the consortia or concessions) and (2) as the direct executioner of the projects. In this manner, the state became not only owner but also entrepreneur. That role led to the creation or strengthening of state-owned companies. These should be capable of taking control of nationalized assets, executing the projects, maintaining the rate of production, and (if necessary) negotiating agreements with the multinational companies to obtain capital, technology, and know-how. During the 1970s, 38 state-owned companies were incorporated and 19 already existing companies were strengthened.

2.2.4 Substitution of the Concession System for the Contract System

Abandoning the concept of concessions had become a question of principle. As a result, production sharing contracts and risk service contracts flourished

in their many forms. However, the title of "concession" did not disappear. It survived, adapting itself to the new circumstances, normally incorporating the state ownership of the oil and an obligatory share of the national company under association agreements. The concession system continued to be used by countries lacking known oil potential that had small, little explored territories and pragmatic governments. Countries that had interesting geological potential or that were already producers were inclined toward production sharing or association contracts. Nationalist countries, with experienced state-owned oil companies and great geological potential, were inclined toward risk service contracts. The companies quickly adapted to the new system because it did not exclude the possibility of limiting risks and making significant profits. Together with the World Bank, they promoted rate of return-based profit-sharing contracts.

2.2.5 Toughening the Conditions of Access, Operation, and Benefit Distribution

Nationalism meant searching for the maximum economic or industrial benefit that a country could obtain from the relationship with the companies. To achieve rapid, efficient, and complete prospecting of the sedimentary basins, as well as optimum extraction of hydrocarbons, the terms of the contracts were reduced, the investment obligations were increased, programmed periods were contracted, delivery of the contracted zones was accelerated, surface rents were increased, and the control and follow-up processes were improved.

To increase the share in the benefits, the tax burden was increased through greater premiums per contract signed and for commercial discovery (bonus), increases in royalties and taxes on benefits, reductions in cost oil (the part of production destined to recovering costs) and profit oil (the part that produces income for the companies), increases in amortization periods, greater state participation in the projects, and the disappearance in the depletion allowance for the reconstitution of oilfields. The tax collection mechanisms multiplied and were perfected with an emphasis on efficient collection of taxes such as the rent-skimming tax, cost stop, price cap, and windfall profit tax.

Rights, taxes, customs tariffs, obligations for national content, limitations for repatriating benefits abroad, commitments to create infrastructure, the use of local goods and services, the hiring of local labor, and the carrying out of social projects were established to prevent the flight of oil profits from national territory. With the intention of dominating the local oil industry, clauses were imposed relating to the transfer of technology as well as to training the national companies' employees. Finally, measures were demanded to protect the natural environment and the productive and social activities that may be affected by oil projects.

The most notable toughening up was seen in the large and medium oil-exporting countries that remained unnationalized, such as Egypt, Oman, and Qatar, and in the smaller producers that strongly embraced nationalism, such as Ivory Coast, Zaire, and Sri Lanka. Some countries, such as Argentina, India, and Turkey, were more pragmatic, following the trend but without frightening away the oil companies. Following the 1979 oil glut, a second wave of reclamations was observed, driven by the non-OPEC countries of Malaysia, Colombia, and Egypt as well as by other member countries, such as Ecuador, Indonesia, and Gabon, that previously had shown a degree of moderation.

3. PRAGMATISM SUBSTITUTES NATIONALISM: 1981–2003

The toughening of the conditions of access, operation, and distribution of benefits would not have been possible without favorable circumstances. The factors of the oil industry itself were particularly important. The dramatic growth in demand, the substantial increase in oil prices, and the general perception of scarcity caused savage competition among the oil companies to obtain mining rights. The Rome Club's catastrophic projections reflected the climate of that period. The perspective of ever more expensive and scarce oil forced companies to accept less advantageous conditions on the "invest now or go away" principle. Such circumstances did not last long. The prevailing conditions following the second oil glut were very different from those that had existed during the 1970s. The producing countries ceased to enjoy the correlation of forces in their favor due to changes both in the oil industry and in the economic and global political contexts that serve as a backdrop. Under those circumstances, thriving competitions developed among producing countries to attract foreign investment.

3.1 Nationalism Weakened by Lower Capital Availability

The industry structure changed as a result of the nationalist movements in the large exporting

countries. Vertical integration was broken, new markets were created, and new producers that did not join OPEC emerged. When the companies lost control of world reserves, the quantities produced, and the fixing of prices, relationships with the producing countries became highly dependent on the peaks and troughs of the market. From that point onward, the availability of risk capital for exploration and extraction projects depended directly on oil price levels, which were difficult to predict. In a tense market with perspectives of scarcity, the owners of the reserves had more leeway to impose their conditions. In a surplus market, the foreign operators could impose theirs.

Oil price quotations declined during the 1980s. The reduced budgets of the multinationals were not directed at developing countries as a priority but instead were directed at politically safe areas or at stock exchanges to absorb firms with proven reserves. Therefore, the companies began to pressure the producing countries for more favorable conditions. The World Bank and other credit institutions turned the thumbscrews by ceasing to finance public investment in oil. This situation changed little during the 1990s. Price increases during 1995–1996 and 1999–2002 were translated into greater availability of funds, but these were directed toward regions that were politically safe but had potential (e.g., the North Sea), at large producers that opened up their territories (e.g., Russia, Venezuela), at large producers that lowered their pretensions (e.g., Nigeria, Angola, Oman, Egypt), and at the frontier zones in industrialized countries (e.g., the Gulf of Mexico).

3.2 Economic Crisis, Neoliberal Paradigm, and a Unipolar World Close in on Nationalism

Changes in the general frame of reference also contributed to weakening the position of the producing countries. The solution to the delicate economic situation of the 1980s known as the "debt crisis," characterized by inflation, currency depreciation, and high levels of debt, was sought through structural adjustment programs agreed to with international financial organizations that invariably included measures to enable foreign investment. Credit restrictions of debt service obligations reduced the availability of public funds. This reduction directly affected the national oil companies, especially when such limitations combined with inadequate prices, poor efficiency, and the diversion of funds for purposes other than productive reinvestment to reduce their response capacity to face an ever growing demand due to industrialization processes, the gradual elimination of social backwardness, and population growth. And although the economic urgency was slowly disappearing, neoliberal politics continued to be applied under the pressure of the so-called "Washington consensus."

The frame of reference changed completely during the 1990s and the first few years of the new millennium. The North–South confrontation, which until then had formed the backdrop of the relationships between the producers and the companies, changed with the end of the cold war. With the collapse of the Soviet Union and the socialist regimes of Eastern Europe, the political, technical, economic, and human support that had been offered to several developing countries disappeared, leaving them with no other option but to turn to the multinationals.

On confirming U.S. dominance, pressure increased on those countries that had nationalized their oil industries. As a result of the Gulf War in 1990, Iraqi oil exports were administered under the UN "oil-for-food" program. Since 1996, Iran and Libya have been subject to U.S. economic sanctions that particularly affect the oil sector. Hugo Chavez's government, trying to regain governmental control over Petroleos de Venezuela (PDVSA) and restrain the opening up of foreign investment in Venezuela, has been subject to destabilizing pressures since 2001. Also, following the military intervention in Iraq in March 2003, Iraqi oil is under U.S. guardianship.

In parallel, the generalization of the neoliberal paradigm accelerated the opening of economies and the liberalization of activities as well as the flow of investment. The outline of the state decreased notably. The energy sector, one of final defenses of the state-entrepreneur, did not escape the change. The restructure included the opening of reserved areas, the segmentation of activities, the introduction of competition, and the sale of state-owned companies. The effervescence centered in the natural gas and electricity industries but also reached the oil industry. In a kind of historical revenge, the multinational companies and their governments took advantage of the new circumstances to reclaim positions and privileges that they had enjoyed in the past.

All countries ceded, although with varying scope and speed. The need to attract foreign capital was less urgent for those that possessed important oil potential and national companies with balanced trade and economies. For most, the adverse conditions limited their negotiating power. That notable

asymmetry had repercussions in OPEC. Ecuador left the organization in 1992, and Gabon left in 1995. The defense of sovereignty over hydrocarbons had united OPEC, but once ownership rights had been recovered, it was divided by pricing strategies.

3.3 Adjustments to Attract Risk Capital

The politics of openness are best understood by classifying countries based on their oil industry experience, geological potential, and economic health. In that sense, three main groups may be identified:

- *Countries that were completely dependent on external contributions of capital, experience, and technology.* Their negotiating power with the multinationals was very weak. They were highly sensitive in the economic context and to the international oil market situation. They had a modest or unknown oil potential, the production levels of some were low, and their economies generally were poorly developed. Countries in this category included New Guinea, Congo, Guatemala, The Philippines, and Paraguay.
- *Countries that partially or fully controlled exploration and production activities but did not achieve full control of their industry due to, among other factors, budget restrictions that weighed heavily on the national oil companies.* To maintain productions rates, they entered into association agreements and shared production agreements or risk service contracts. This category included medium-sized exporters such as Indonesia, Oman, Nigeria, Qatar, Ecuador, and Egypt as well as importers with solid state-owned companies such as Argentina, Brazil, and India.
- *Nationalized countries that enjoyed a generous geology and that had taken control of all sectors of the oil business.* They were able to undertake vast exploration campaigns, managing to obtain capital, know-how, and technology to carry out such projects. Their experience in commercial circles allowed them to optimize production in advantageous conditions and allowed them to undertake an internationalization process. In the beginning, their relationship with the companies was limited to service contracts, but later they agreed to negotiate agreements regarding access to production. This was the case with Iran, Iraq, Venezuela, Argelia, Kuwait, Saudi Arabia, and Mexico.

With regard to sovereign management of resources, the dynamics in which the countries in the first two groups evolved were similar; with some exceptions, they had always turned to the companies. Conversely, the countries in the third group, having nationalized and started an internationalization process, evolved in a different dynamic.

3.4 A Step Back for the Small and Medium Producers

The countries in the first group, small or medium producers or those with no production, introduced faster and more vigorous adaptations to attract the oil companies. For these, the principle of sovereignty was limited to ideological debate. They quickly concluded that encouraging exploration to prove the potential existence of hydrocarbons was the main objective in controlling the industry or maximizing the taxation of the eventual production of oil or natural gas.

On the other hand and with some exceptions, the countries in the second group reacted relatively slowly and moderately. The problems arising from the change of context did not seem to make a mark in their nationalist conviction. However, to the extent that the exploration budgets were reduced and the economic reforms deepened, the adjustments in oil contracting became ever more important.

Despite their differences, the small and medium producers were faced with a policy of adjustment whose content and essential components were comparable. They adopted the following measures (among others): assignment of a better cost oil and profit oil for the companies, reduction or elimination of the state's participation in projects, reductions in taxes, suppression of oil price controls and currency exchange, reduction or elimination of royalties, signing of seismic option contracts, elimination of local personnel training, and relaxation of the restrictions related to burning natural gas.

There were some countries that did not want to relax their nationalism and others that did just the opposite. They did not take just one step backward; they took many. For example, they accepted waiving the renegotiation of contracts in case of oil price increases, applying the same advantageous conditions that other companies obtained in negotiations with states, guaranteeing a substantial benefit to foreign companies, and exonerating the operators from the payment of taxes, the elimination of royalties, and so on.

In summary, the small and medium producers consented to sacrificing economic profit and desisted

from the objective of dominating the oil industry. However, they did not waive ownership rights, an eminently political claim.

3.5 Adjustments in the Large Producers

Because the countries in the third group had to wait, the large producers were the last to call the oil companies again. But they did so not to occupy the place they had before nationalization; instead, they did so to participate in the specific activities that the national companies had developed with difficultly due to lack of capital, technology, and experience. The important role that these factors play is seen with greater clarity during times of crisis—when oil prices collapse, debts increase, fiscal resources become stunted, and state company exploration and production budgets collapse. The insufficiency of these key factors explains the advance of pragmatism to the detriment of nationalism.

The first to call the oil companies again, in 1986, was Argelia, a country overwhelmed by huge foreign debt. Its government even proposed the sale of some of the largest oilfields. It was followed by Iran and Iraq. Faced with enormous needs to rebuild their economies after the war, they opened negotiations in 1989 to rebuild and expand production capacity with the help of the multinationals. In the case of Iraq, the granting of sharing production contracts to non-U.S. firms was not far removed from the strategy of obtaining powerful allies such as France and China, a strategy that pressured for the end of sanctions that the UN Security Council had imposed on Iraq following the Gulf War. Under the tutelage of the United States, the Iraqi oil industry would be privatized and the oilfields would be granted in concession.

In 1990, Venezuela began an opening up process through operating service agreements to reactivate inactive or abandoned fields as well as association agreements in liquid natural gas (LGN), heavy-duty oils, and orimulsion. In 1995, profit-sharing agreements were introduced. In 1999, the natural gas industry was opened. And in 2001, a 51% share was established for PDVSA in all exploration and production agreements.

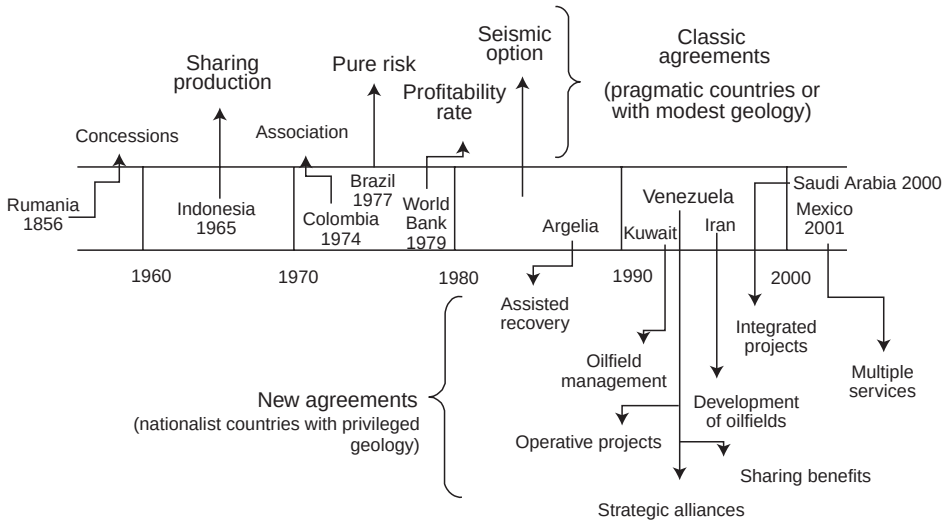
After the Gulf War, Kuwait called on the oil companies temporarily to rebuild its extraction installations. Government and Congress are face to face regarding the possibility of newly granting mining rights to the private sector. Saudi Arabia opened negotiations in 2000 to allow the multinationals to exploit nonassociated natural gas through agreements, including its use in the generation of electricity and the desalting of water. Finally, the Mexican government in 2001 announced its intention to enter into service agreements to share the risks and benefits of nonassociated gas and (later) oil, despite Congress's opposition.

What has been the nature of the changes in terms of the principle of sovereignty over natural resources? In the first place, no country has resigned ownership of oil and natural gas. In the second place, and leaving to one side Argelia and Venezuela, countries have not changed their oil legislation to grant mining deeds to the oil companies. What they have done is reinterpret the legal framework in such a way that the state companies may enter into contracts so that the international oil companies contribute capital, technology, and management capability, at the same time assuming part of the risk. As a consequence, new oil agreements have emerged that are unlike the concessions or contracts of the 1960s and 1970s (Fig. 1). Governments have tried to find politically acceptable ways of permitting the local oil industry to participate again, making a few visible institutional adjustments so as not to awake nationalist opposition, although not always successfully (as in the cases of Kuwait and Mexico).

3.6 The Opening of Economically Planned Countries

In the framework of Perestroika, a set of reforms aimed at modernizing a country internally and bringing it closer to the West as foreign policy, the Soviet Union began in 1985 by signing a series of technical assistance and technology transfer agreements, culminating in 1990 with the signing of exploration agreements with the Elf, Chevron, and Total companies. Since the fall of the Soviet Union in 1991, the 11 republics have been conceding various types of mining rights.

In the Eastern European countries, the return to a market economy and the changes in the Soviet oil industry, above all concerning the oil supply and the technical cooperation they received, arose mainly from the opening up of their territories to foreign capital. The first country to invite bids for the granting of exploration and production permits was Bulgaria in 1989. In mid-1991, all of the other countries (Lithuania, Estonia, Poland, Czechoslovakia, Hungary, Rumania, and Albania) had already signed agreements with the oil companies or were in negotiations to do so.



Other agreements: Technical assistence, transfer of technology, service contracts, joint ventures, etc.

FIGURE 1 Diversification of exploration and production agreements.

In the developing countries with planned economies, the opening up began before the fall of the Berlin Wall, such as in the cases of China (1979) and Vietnam (1986). Although the call to the Western companies is recorded in the framework of a broader economic opening, in both cases the need for capital and technology to develop offshore oilfields was observed. In the case of Cuba, the opening up was a logical consequence of the cooling off of relations with Moscow. In 1990, the Cuban authorities decided to turn to the European companies in the search for marine oilfields. Moreover, as a test of autonomy, in 1989, Mongolia created a state-owned company to negotiate association agreements with Western companies. Finally, Laos and Cambodia, which were practically unexplored, signed their first contracts in 1991.

3.7 Nationalism Weak but Not Extinct

The nationalist movements allowed the correction of a very unfavorable situation for underdeveloped countries. However, the international context changed rapidly, and the majority of countries did not have time to provide themselves with the means necessary to take economic control of their economic resources.

The following important conclusions may be drawn from the exercise of sovereignty over natural resources over the past three decades or so by developing countries:

- The management of the local oil industry, in terms of sovereignty, is a valid objective. Past experience has not demonstrated that such a thesis is erroneous.
- Although the state is the owner of the subsoil's resources, this does not guarantee economic control over hydrocarbons. The state must be able to directly execute such projects.
- The state directly manages the oil industry, but this does not automatically guarantee said economic control. Certainly, it is indispensable to have an operational sovereign instrument over hydrocarbons (i.e., to have a state-owned oil company), but that is not enough.
- The design of contracts and efficient fiscal mechanisms is a necessary but not sufficient condition to guarantee the correct exploitation of oilfields, the recovery of the greater parts of the profits, and the transfer of technology and know-how.

To achieve the economic control of natural resources required by developing countries, besides the aforementioned ownership rights, a state-owned company and an efficient taxation system are required to reach a high level of economic, social, political, and institutional development.

Also, such economic control is reached only when the product is delivered to the end user. Therefore, some countries, such as Kuwait and Iran, have allowed the state companies to invest and operate in the consuming countries in both the oil and energy sectors. This is an impossible task if the state

companies do not have sufficient room for maneuvering on the operative, administrative, financial, and strategic fronts. Paradoxically, state control over the oil business is refined with autonomy and internationalization such as in the case of Venezuela.

Nowadays, the balance of power is different from what it was during the 1970s. Total sovereignty over hydrocarbons, as it was conceived in the past, seems difficult to reach. Approximately two decades ago, steps were made in the opposite direction: transfer of assets to the private sector and downsizing or disappearance of state companies. This does not mean that nationalism is dead.

The persistence of a market dominated worldwide by purchasers, neoliberal economic policies, the conditions imposed by international financial organizations, and government pressure that supports and promotes the large oil companies are some of the factors that limit sovereignty over hydrocarbons. Military aggression only annihilates it, such as in the case of Iraq. In those cases, the existence of nationalism is ensured. It will exist insofar as the state's right to exercise permanent sovereignty over the nation's natural resources is not guaranteed or is threatened.

SEE ALSO THE FOLLOWING ARTICLES

Development and Energy, Overview • Economic Geography of Energy • Environmental Injustices of Energy Facilities • Geopolitics of Energy • National Security and Energy • Oil Industry, History of • OPEC, History of • War and Energy

Further Reading

Aïssaoui, A. (2001). "Algeria: The Political Economy of Oil and Gas." Oxford University Press, Oxford, UK.

Alnasrawi, A. (1991). "Arab Nationalism, Oil, and the Political Economy of Dependency." Greenwood, New York.

Angelier, J. P. (1976). "La rente pétrolière." Energie et Société, Paris.

Elm, M. (1992). "Oil, Power, and Principle: Iran's Oil Nationalization and Its Aftermath." Syracuse University Press, Syracuse, NY.

Ferrier, R. (1994). "The History of the British Petroleum Company I and II." Cambridge University Press, Cambridge, UK.

Ghadar, F. (1984). "The Petroleum Industry in Oil-Importing Countries." Lexington Books, Lexington, MA.

Linde, C. (2000). "The State and the International Oil Market: Competition and the Changing Ownership of Crude Oil Assets." Kluwer Academic, Boston.

Mommer, B. (2002). "Global Oil and the Nation-State." Oxford University Press, Oxford, UK.

Penrose, E. (1968). "The Large International Firms in Developing Countries: The International Petroleum Industry." Allen & Unwin, London.

Philip, G. (1982). "Oil and Politics in Latin America: Nationalist Movements and State Companies." Cambridge University Press, Cambridge, UK.

Sampson, A. (1976). "Les Sept sœurs: Les grandes compagnies pétrolières et le monde qu′elles ont créé." A. Moreau, Paris.

Tanzer, M. (1969). "The Political Economy of the International Oil and the Underdeveloped Countries." Beacon, Boston.

Taverne, B. (1999). "Petroleum, Industry, and Governments: An Introduction to Petroleum Regulation, Economics, and Government Policies." Kluwer Law International, The Hague, Netherlands.

Terzian, P. (1983). "L'étonnante histoire de l'OPEP." Editions Jeune-Africa, Paris.

Vandewalle, D. (1998). "Libya since Independence: Oil and State-Building." Cornell University Press, Ithaca, NY.

Yergin, D. (1991). "The Prize: The Epic Quest for Oil, Money, and Power." Simon & Schuster, New York.

National Security and Energy

WILFRID L. KOHL
Johns Hopkins University
Washington, DC, United States

1. Introduction
2. Oil: Background
3. Dimensions of Oil Security
4. Reliable Electricity Supply and Price Volatility
5. Natural Gas and Price Volatility
6. Terrorism and Energy Security

Glossary

Achnacarry agreement Second oligoplistic accord reached by major international oil companies at Achnacarry castle in Scotland to control downstream marketing of oil and divide it according to market shares existing in 1928.

futures market An organized market, such as the New York Mercantile Exchange, in which contracts are bought and sold to deliver a specified quantity of oil (or other commodities) at a specified future date at a price to be paid at the time of delivery. Futures contracts allow the holder to hedge risk.

Hubbert curve U.S. petroleum geologist King Hubbert predicted (correctly) in 1956 that U.S. oil production followed a bell-shaped curve and would peak in approximately 1970.

International Energy Agency Established in 1974 within the Organization for Economic Cooperation and Development to promote energy cooperation among consumer countries including oil crisis management; currently has 26 member countries.

Kyoto Protocol International agreement signed in 1997 that sets binding emissions reductions of greenhouse gases with an average 5.2% reduction below 1990 levels for industrial countries. As of 2002, the protocol had not yet entered into force. The European Union and Japan have ratified it, but the United States has declined to participate.

Organization of Petroleum Exporting Countries (OPEC) - Founded in 1960 in Baghdad; members are Algeria, Indonesia, Iran, Iraq, Kuwait, Libya, Nigeria, Qatar, Saudi Arabia, United Arab Emirates, and Venezuela.

Red Line Agreement Agreement in 1928 on oil production in most of the former Ottoman Empire by American and European major international oil companies; established structure of the Iraq Petroleum Company and was an important foundation of the international oil company cartel.

Seven Sisters A cartel of seven international oil companies that dominated the world oil market after 1928 until the rise of OPEC; five of the companies were American—Standard Oil of New Jersey (Exxon), Socony Vacuum (Mobil), Standard of California (Chevron), Texaco, and Gulf—two were European—Anglo-Iranian Oil, which became British Petroleum, and Royal Dutch/Shell.

Strategic Petroleum Reserve U.S. government oil reserve established under the Energy Policy and Conservation Act of 1975 to be used to mitigate effects of oil market disruptions; contained approximately 650 million barrels as of 2003 located in four underground salt caverns along the Texas and Louisiana Gulf Coast.

Texas Railroad Commission A state agency that in 1932 acquired authority to regulate oil production via a system of market demand prorationing within the state.

Traditionally, energy policy has sought security of supply, affordability, and limited impact on the environment. Until recently, energy security has been dominated by oil security, since oil has been the leading fuel and is subject to the influence of the OPEC cartel and the geopolitics of the world oil market. This article reviews the history of oil, how it became a strategic commodity, and the importance of oil today. It analyzes the multiple dimensions of oil security: the long-term outlook for oil supply, OPEC market power, oil and the economy and the costs of oil market disruptions, the role of oil imports and the balance of payments, oil crisis management, oil and environmental security, and the link between oil, foreign policy, and geopolitics. The article concludes by considering other energy security concerns: shortages and price volatility in electricity and natural gas, and potential threats to energy facilities by terroist attacks.

 © 2004 Elsevier Inc. All rights reserved.

1. INTRODUCTION

Energy policy is a subset of economic policy, foreign policy, and national and international security policy. Traditionally, energy policy has sought security of supply, affordability, and limited impact on the environment. According to the Bush administration's "National Energy Policy," energy should be "reliable, affordable, and environmentally sound." Until recently, energy policy was dominated by oil policy since oil has been the leading fuel in the United States and most industrial economies, and it has been subject to price volatility and the vagaries of the world oil market that is greatly influenced by the Organization of Petroleum Exporting Countries (OPEC) cartel. Most of this article therefore reviews the elements of oil security. It begins with a brief historical review of the oil market and how oil developed as a strategic commodity with an eye to its military uses. After World War II the focus shifted to its rapidly expanding role in the civilian economy and how to deal with oil supply interruptions and their macroeconomic effects as well as possible resource constraints.

Other types of energy security concerns have surfaced, including shortages and the effects on consumers of high electricity and natural gas prices as a result of problems with deregulation. In the future, there could be potential threats to energy facilities by terrorists in the wake of the September 11, 2001, attacks. These aspects are discussed toward the end of the article.

2. OIL: BACKGROUND

2.1 History

Early concerns about oil and security date back to World War I. In 1913, Winston Churchill, then First Lord of the Admiralty, decided to convert the British Navy from using coal to bunker oil, which was a cleaner and more efficient fuel. To ensure adequate oil supply, the British government took a majority share in the Anglo-Persian Oil Company (later renamed British Petroleum). This was one of the first major government interventions in the oil industry. Other navies soon followed the British example. The appearance of tanks and other motorized military vehicles highlighted oil's important new role as a strategic commodity. During World War I, allied nations established government controls over oil supply. Just after the war, the U.S. government established three Naval Petroleum Reserves.

Following World War I, as oil became more important in the civilian economy in transportation and industry, the U.S. Geological Survey predicted that the United States was about to run out of oil because of insufficient domestic reserves. The U.S. government in turn supported the efforts of American oil companies to gain access to concessions in the Middle East. In 1928, the Red Line and Achnacarry agreements provided mechanisms for coordination of supply and markets by a secret cartel of seven American and European international oil companies (the seven sisters), which dominated the international oil market until well after World War II. Following the discovery in 1930 of large oil fields in east Texas, which produced an oil surplus, state regulation of production began in 1933 under the mantra of conservation by the Texas Railroad Commission to help stabilize oil prices and the health of the domestic industry. (The Texas Commission later provided a model for OPEC.)

World War II demonstrated the growing importance of oil in modern warfare to fuel fighting ships, freighters, submarines, tanks, troop transports, and airplanes. Access to oil was a major reason behind Germany's invasion of the Soviet Union and Eastern Europe and Japan's advances on the Dutch East Indies and Southeast Asia, including the Japanese attack at Pearl Harbor. (The United States had placed a de facto oil embargo on Japan. U.S. naval forces could have interfered with Japanese efforts to secure oil supplies in Southeast Asia.) Meanwhile, the future importance of the Middle East, where major new oil fields had been discovered in the 1930s, was clearly recognized. In 1943, President Roosevelt announced that the defense of Saudi Arabia was vital to the United States and extended Lend-Lease aid. After the war, the U.S. government supported the efforts of four American oil companies (Exxon, Mobil, Texaco, and Chevron) to form a consortium to develop Saudi oil, the Arabian–American Oil Company (Aramco).

The first postwar oil crisis occurred in 1951 in Iran when a nationalist leader, Mohammed Mossadegh, seized power and decided to nationalize the Anglo-Iranian Oil Company (BP). Because of the strategic importance of Iran, the CIA helped to stage a countercoup that ousted Mossadegh 2 years later and restored the Shah. At the same time, the U.S. government took the lead in organizing a new Iranian oil consortium, which for the first time included participation by American oil companies.

During the Suez crisis in 1956 (the second postwar oil crisis), when Egypt's President Nasser nationalized the Suez Canal and Israel invaded the Sinai and began moving toward the Canal, Britain and France intervened and sent troops. Concerned about escalation of the crisis and Soviet involvement, the United States offered to provide oil to its European allies if they would agree to a cease-fire and withdraw, which they did. Excess U.S. oil production capacity enabled this diplomatic action. By the early 1970s, this excess capacity had disappeared.

After the postwar recovery, a boom period of U.S. and global economic growth stimulated increasing demand for crude oil and petroleum products. In the United States, this was driven by a rapid expansion of highway transportation and growing environmental concerns over coal-fired power plants. However, new oil supplies were also becoming available, especially from the Middle East. OPEC was founded in 1960 to put pressure on the international major oil companies not to reduce prices further in a period of increased competition in the oil market from independents and more than adequate supply. However, by the early 1970s, demand increases and rising U.S., European, and Japanese imports led to a much tighter market. This set the stage for stronger OPEC actions in the oil crises of the 1970s.

In October 1973, when the United States supported Israel in the Yom Kippur war, Arab nations responded with an oil embargo and production cutbacks. OPEC quadrupled oil prices, raising concerns about the security of oil supply to the Organization for Economic Cooperation and Development (OECD) nations and the long-term adequacy of oil resources. In response, the United States created a Strategic Petroleum Reserve (SPR) and took the lead in establishing the International Energy Agency (IEA) associated with the OECD. The principal purpose of the IEA was to establish an emergency oil-sharing system among member countries to assist in managing future supply interruptions.

The Iranian revolution caused an oil supply shortfall in 1979 and produced another oil crisis as oil prices doubled, led by the spot market and adopted by OPEC. The outbreak of the Iran–Iraq war in 1980 exerted further upward pressure on oil prices, which reached a high of $34/barrel for the Saudi benchmark crude in 1981. This action turned out to have negative consequences for OPEC because the high oil prices in the first half of the 1980s encouraged fuel switching, the development of more efficient technologies, and a strong increase in non-OPEC oil production, which, along with decreasing demand due to recession, reduced OPEC's market share from more than 30 million barrels/day (mbd) in the 1970s to as low as 16 mbd by the mid-1980s. Following Saudi Arabia's decision in 1985 to abandon its role as "swing producer," oil prices collapsed in 1986 to less than $10/barrel, although prices returned to the $15–20 range the next year after OPEC regrouped and reinstated production quotas.

By the time of the outbreak of the Persian Gulf War in 1990, the world oil market had become more competitive. Players included national oil companies, international majors, and larger numbers of independent oil companies. Oil was now traded like other commodities in forward and futures markets, providing a mechanism to hedge price risk. Market transparency was also stronger because of the expanding application of information technologies, which enabled faster adjustment to market trends. Also, several governments, led by the United States, had built up strategic oil stocks for use in emergencies.

The Persian Gulf crisis (1990–1991) demonstrated the traditional aspects of oil diplomacy and military force along with market-based mechanisms. After Iraq invaded Kuwait in October 1990, the United Nations (UN; urged by the United States) placed an embargo on oil shipments from Iraq, removing more than 4 mbd of crude from the world market. A U.S.-led UN expeditionary force was deployed to Saudi Arabia during the fall to protect Saudi Arabia and to prepare for military action to push Iraq out of Kuwait. Although some of the oil shortfall was made up by increased production from Saudi Arabia and other producers, oil prices escalated to nearly $40/barrel in part because of uncertainty regarding what would happen next. Although other motives may have played a role in the U.S. action, clearly the threat posed by Saddam Hussein to the security of the world oil market, and to the world economy, was a paramount concern. When air strikes against Iraq commenced in January 1991, the United States, Germany, and Japan initiated a drawdown of strategic oil stocks that was coordinated by the IEA. As a consequence of both actions, oil prices declined to normal levels closer to $20/barrel. During this oil crisis many buyers and sellers hedged oil transactions on the futures market.

At the end of the 1990s, another oil price collapse and then a price shock took place, which were much more the result of market forces. After a misjudgment of the oil market by OPEC at the end of 1997 when it expanded production as Asia was going into recession, oil prices plunged in 1998 and early 1999 to approximately $10 barrel, causing serious damage

to the economies of oil-producing countries. OPEC rallied, and acting with greater cohesion beginning in March 1999, it lowered production, causing prices to rebound and then increase to more than $30/barrel in 2000. In 2001, OPEC struggled to manage falling prices set off by a global recession made worse by the attacks of September 11, 2001, and the war on terrorism.

2.2 Importance of Oil Today

Although the importance of oil in the world economy has diminished somewhat since the 1970s, oil remains the largest source of primary energy in industrial countries at 40% (compared to 55% in 1980). The oil intensity [i.e., the amount of oil required to produce a unit of gross domestic product (GDP)] of industrial economies has fallen considerably due to fuel switching, structural change, and advances in energy efficiency. In 2000, oil imports represented only 4% of the total value of OECD imports, compared to 13% in 1981. However, oil is still very important, accounting for two-thirds of international energy trade. Oil reserves are less abundant than those of coal and natural gas and less evenly distributed. Most oil reserves are located in developing countries, which makes the oil market more amenable to cartel control than other commodity markets, hence the influence of OPEC.

For the United States, oil accounts for 39% of primary energy consumption, which in 2000 amounted to 19.5 mbd of petroleum products (Fig. 1). Of this amount, approximately 10 mbd was derived from crude oil and product net imports. Most oil (approximately 13 mbd) is consumed in the transportation sector. The United States accounts for approximately 25% of the world's oil consumption. The U.S. oil demand is expected to grow approximately 1.5% annually until 2020, led by growth in the transport sector. Net imports, which amounted to 53% of crude and product needs in 2000, are projected to increase to 62% in 2020. Imports come from a variety of countries, both OPEC and non-OPEC (Fig. 2). The U.S. domestic oil production was approximately 9.5 mbd in 2000. The U.S. onshore production is scheduled to decline, but production in the Gulf of Mexico is expected to increase, which will probably lead to stable or slightly expanded production in the next few years.

In peacetime, the U.S. military uses less than 5% of national consumption of crude oil and petroleum products. In case of military conflict, this requirement would increase. However, the military requirement represents a relatively small percentage of domestic oil consumption, and it is dwarfed by the needs of the civilian economy. Most of the incremental military demand in wartime would likely come from overseas in one or more war zones, and most of the needed supplies would be purchased abroad. In case of a large-scale conflict involving the domestic homeland, the military has mechanisms in place to procure oil and other fuels for defense needs. This would involve invoking the Defense Production Act and the Energy and Security Act. The Department of Defense maintains petroleum stocks in the United States and abroad.

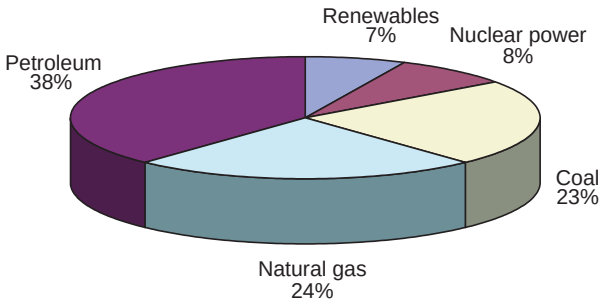


FIGURE 1 U.S. primary energy consumption, 2000. From the Energy Information Administration (2002).

3. DIMENSIONS OF OIL SECURITY

3.1 Long-Term Outlook for Oil Supply

The oil shocks of the 1970s and their associated uncertainties led to major questioning about future oil resource scarcity, as highlighted by the publication

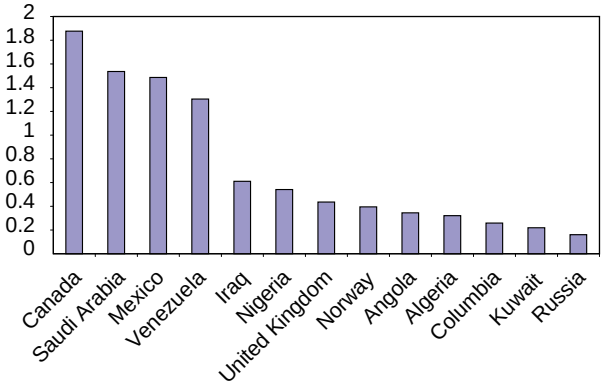


FIGURE 2 U.S. oil imports, source countries (million barrels/day, first 6 months of 2002). From *Monthly Energy Review*, Department of Energy, Washington, DC.

of the Club of Rome study, "The Limits to Growth." After the oil price collapse of 1986, this view was less prevalent. However, at the end of the 1990s debate had begun on the future of oil supplies, spurred by articles in prominent journals by geologists Colin Campbell and Jean Laherrere, who argued on the basis of applying the Hubbert curve to the world that conventional oil production would peak in the first decade after 2000 and set the stage for a coming oil crisis. (King Hubbert was a famous U.S. geologist who correctly predicted that U.S. oil production would peak in approximately 1970.) Their argument was based on the proposition that all large oil fields have already been found and that world reserve data are inaccurate, especially in the Middle East, where several OPEC member countries suddenly and mysteriously increased their reserve figures in the late 1980s. According to these pessimists, world recoverable oil reserves at the end of 1998 were estimated at 1800 billion barrels, and world production could peak by 2010 if not sooner. A similar perspective is presented by geologist Kenneth S. Deffeyes in his book, "Hubbert's Peak: The Impending World Oil Shortage."

Among the leading opponents of this view are economists M. A. Adelman and his associate Michael Lynch, who pointed out serious limitations of the Hubbert curve and emphasized the role of investment and new technology in expanding reserves. Oil reserves are better viewed as inventory, which is replenished by investment. Depletion is constantly delayed by new knowledge and advances in production technology. Oil supply forecasts have tended to be dominated by a pessimistic bias.

In 2000, the authoritative U.S. Geological Survey (USGS) published its latest estimates of world oil and gas reserves outside the United States that have the potential to be added during the period 1995–2025. The estimated volumes of undiscovered conventional oil are 20% greater than the 1994 estimate. The potential addition to reserves from reserve growth (e.g., due to applications of new technology) is also very large. When the new mean global estimates are combined with previous estimates for the United States, the USGS contends that worldwide ultimately recoverable reserves (URRs) of conventional oil total 3.021 trillion barrels, and natural gas liquids (frequently added to oil estimates) total an additional 324 billion barrels. URRs include cumulative production to date, identified remaining reserves, undiscovered recoverable resources, and estimates of "reserve growth" in existing fields. The 1994 USGS estimate listed URRs of conventional oil at approximately 2.3 trillion barrels.

The new USGS estimates have been adopted by two organizations that regularly publish widely respected energy market forecasts. Both the IEA in its 2001 "World Energy Outlook" and the U.S. Energy Information Administration (EIA) in its "International Energy Outlook 2002" take an optimistic view of future world oil supply and conclude that proven oil reserves are adequate to meet demand until 2020, with a world production peak to occur sometime thereafter. In addition, the world has very large reserves of unconventional oil, including Venezuelan heavy oil and Canadian tar sands, which will become economic to produce at higher oil prices. In short, the world appears to have sufficient oil reserves for the foreseeable future.

An important factor in the increased estimates for oil reserves, especially for oil reserve growth, is the recent advances in oil production technology, which have improved success rates and lowered costs of finding oil and production. These advances include three- and four-dimesional seismic for locating and evaluating underground deposits; directional drilling; floating production systems; deep-water platforms; and the general widespread application of computers and information systems by oil companies and service contractors to improve analysis, management, and communications. Direct production costs for the international oil companies are estimated to average $3–6/barrel worldwide.

3.2 OPEC Market Power

The world's conventional oil reserves are concentrated in a relatively few countries. Most of these countries are members of OPEC, established in 1960, which controls approximately two-thirds of global oil reserves and in 2002 slightly less than 40% of world oil production. OPEC has a formal organization, meets approximately four times per year, and attempts to act as a cartel to manage world oil supply and prices. Since 1982, it has set production quotas for its members. As an organization of 11 countries, not firms, OPEC has not always been successful at maintaining cohesion. Members have political as well as economic objectives and are known to cheat on their quotas. However, since the 1970s OPEC has been successful most of the time in using its market power generally to maintain the price of oil well above the costs of production (which are less than $2/barrel in the Persian Gulf or approximately $4/barrel including finding and development costs.)

It has also been successful at either causing or capitalizing on oil market disruptions to obtain very high prices. Basic to its success in this regard is the fact that oil demand and, to some extent, supply are inelastic and do not respond quickly to changes in price. There are few readily available substitutes for oil, especially in the transportation sector.

OPEC, however, has made mistakes. As previously noted, after locking in very high prices in 1981, which caused a recession and a precipitous decline in oil demand, along with an increase in non-OPEC production, OPEC lost market share and prices collapsed in 1986. However, OPEC regrouped and benefited from an oil price spike at the time of the Persian Gulf War in 1990, and Saudi Arabia increased its production following the UN embargo against Iraq. During the early and mid-1990s in a more competitive oil market, prices were more volatile and ranged from approximately $18 to $20/barrel. Following a bad decision by OPEC to increase production at the end of 1997, oil prices collapsed in 1998 and early 1999. However, OPEC—led by Saudi Arabia, the largest producer—reasserted cohesion in its March 1999 decision and succeeded in obtaining support from non-OPEC states Norway and Mexico, which resulted in lower production and a price recovery in 1999 and 2000—initially to more than $30/barrel before settling back into the mid-$20 range. Thus, OPEC has shown that it still has market power and is able to use it, even if it is not a perfect cartel. It has never been able to agree on an enforcement mechanism for its decisions. In economic parlance, OPEC is an imperfect monopolistic cartel of the von Stackelberg type. OPEC has a large enough market share to influence prices, but its influence is limited by the existence of other competitive suppliers.

Looking to the future, OPEC's market share is expected to grow, which will give it even more leverage on the world oil market. Both the EIA and the IEA predict that OPEC production will increase from 30 mbd in 2000 to approximately 60 mbd in 2020 to meet growing oil demand, which would represent a market share of more than 50%. Non-OPEC production is also projected to grow but less rapidly.

3.3 Oil and the Economy and the Cost of Oil Market Disruptions

Research has shown that sudden increases in oil prices are linked to inflation, increased unemployment, and higher interest rates. Eight of 10 post-World War II recessions in the United States (including several recent ones) were preceded by oil price shocks (Fig. 3). A study by the Stanford Modeling Forum concluded that the first oil shock of 1973–1974 reduced U.S. GNP in 1975 by 3–5.5%, and the second shock in 1979–1980 reduced U.S. GNP in 1981 by 2–4%. An OECD study reported the real income losses to all OECD countries due to the second oil shock at approximately 5% in 1980 and 8% in 1981. However, there is asymmetry in the oil–economy relationship. Whereas rising oil prices tend to retard economic growth, declining oil prices or an oil price collapse (such as occurred in 1986 and again in 1998) do not necessarily stimulate growth, apparently because there are more economic adjustment costs and coordination problems associated with rising prices.

David Greene, Donald Jones, and Paul Leiby of the Oak Ridge National Laboratory analyzed the macroeconomic impact of oil price increases on the U.S. economy. They contend that there are three types of economic losses:

Loss of potential GDP: When the price of oil is increased by monopoly power, oil becomes more scarce. The economy is able to produce less output with the same resources of capital, labor, materials, and land.

Macroeconomic adjustment costs: When prices increase rapidly, there are additional transitional costs because wages and prices are not able to adjust rapidly enough to the higher oil prices to permit the economy to operate at full employment. There is also the possibility of adjustment required by changes in monetary policy.

Wealth transfers from U.S. oil consumers to foreign oil exporters: These transfers are equal to the quantity of U.S. oil imports multiplied by the difference between the monopoly price and the competitive price of oil. These transfers go to OPEC and non-OPEC producers and are a cost to the U.S. economy.

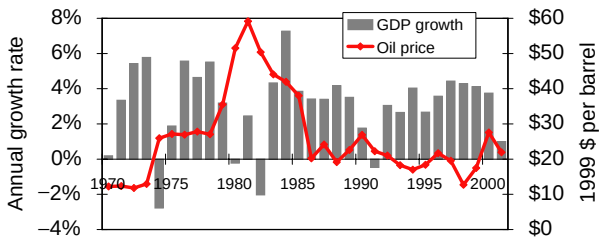


FIGURE 3 Oil price and economic growth, 1970–2001. From Greene and Tishchishyna (2000) and data updates, 2001.

In the most recent update of this analysis, David Greene and Nataliya Tishchishyna estimate the costs to the U.S. economy of oil dependence and oil monopoly power over 30 years at approximately $7 trillion 1998 dollars, which as they point out is approximately equal to the payments on the national debt during the same period. (This estimate is controversial since it assumes a competitive oil price of approximately $10/barrel. There are a range of views regarding what oil prices would be absent the OPEC cartel.) Whether or not this figure is correct, the costs of oil price shocks to the economy have clearly been substantial.

Greene and Tishchishyna define oil dependence as the product of a noncompetitive world oil market dominated by OPEC; high levels of U.S. oil imports; the importance of oil to the U.S. economy, where it is critical in the transportation sector; and the absence of readily available or affordable substitutes. The transportation sector is key because on-road vehicles account for two-thirds of U.S. oil consumption and there are no substitutes immediately available. In their view, the cost of oil dependence to the United States is a market failure because of the cartelization of the world oil market.

The United States can take some actions to reduce the costs of this dependence, including developing advanced technologies to increase efficiency in the transport sector, pursuing alternative energy technologies, and improving the technology of oil exploration and recovery to increase oil supply. Furthermore, the United States can diversify its sources of oil to more non-OPEC countries, a trend that is already under way.

It should be noted that the economic significance of oil consumption in the U.S. economy is much less today that it was in the past. In 1999–2000, the oil cost share of GDP was 1.4–2.0% of GDP. This compares with a 4–6% share in the early 1980s. Changing economic structure and improvements in efficiency of oil use have both contributed to this development. As discussed by Brown and Yûcel, the relationship between oil prices and economic activity also weakened by the end of the 1990s.

3.4 Role of Oil Imports and the Balance of Payments

In the media and political debates, oil security is frequently equated with the level of U.S. oil imports, which has been rising. (Net U.S. oil imports were 55% of oil consumption in 2001.) The transfer of wealth to pay for oil imports is a factor, as noted previously. However, equally if not more important is the level of overall oil consumption, which is related to the oil intensity of the U.S. economy.

In recent oil market disruptions, the increase in oil prices has been the most important factor. Any supply shortfalls have not lasted more than a few months. Since crude oil is traded on a world market, the amount of oil a country imports has little effect on the price of oil in the short term since that price is set on a world market. The vulnerability of the United States and other countries to oil shocks depends more on the level of oil consumption and the price of oil.

Arguments are often made that rising U.S. oil imports are harming the U.S. economy, and therefore national security, because of their contribution to the trade deficit. Although U.S. oil imports are increasing, they have amounted in recent years to 8–10% of the value of total U.S. merchandise imports. This is certainly acceptable for a complex economy the size of that of the United States. By comparison, several European countries and Japan import more oil (more than 90% of their oil needs) than does the United States, and this has not harmed their economies. (The situation can be different for a developing country with a simpler economy and reduced ability to pay for a large amount of oil imports.) The changing role of oil imports in the trade deficit is more a function of changes in the imports and exports of other commodities than it is in the role of oil imports. In short, the United States will need to get used to increasing levels of oil imports. Reducing import dependence would be very costly in the short to medium term. However, as noted previously and later, there are reasons to consider ways to reduce oil consumption overall in the future.

3.5 Oil Crisis Management

An international framework for oil crisis management is provided by the IEA, established in 1974 and located in Paris, where it is associated with the OECD. The 26 member countries are obligated to hold oil stocks equivalent to 90 days of imports. The heart of the agency is the formal oil-sharing mechanism, which is intended to mitigate future supply disruptions. The system is triggered by either a 7% or a 12% or greater supply shortfall in the daily rate of oil supplies to one or more member countries, whereupon a request can be directed to the secretariat that the oil-sharing mechanism be triggered. The 7% case would normally be dealt with by demand restraint/conservation measures, but the 12% plus case is more serious and might require

sharing of oil supplies. In any case, the governing board would need to be convened to decide on the appropriate response by a weighted majority vote. The system has never been triggered, although a request was considered in 1979.

With growing awareness that small market disruptions can yield sharp oil price spikes, IEA countries agreed in 1984 to a less formal system of international consultations aimed at release of government-controlled emergency stocks known as Coordinated Emergency Response Measures by member countries that possess such stocks (mainly the United States, Germany, and Japan, although other countries are considering the idea). This framework was first used in January 1991 for a coordinated stock release at the beginning of military action in the Persian Gulf War. In 1995, the governing board reached agreement that coordinated stock draw should be given priority in response to market disruptions regardless of size and before activation of the formal allocation mechanism.

The other purposes of the IEA are to promote the development of alternative energy sources and energy efficiency, to gather data and conduct studies on the world oil market and other energy markets, to encourage cooperation with nonmembers, and to assist in the integration of energy and environmental policies.

The European Union (EU) also has a requirement that member countries hold stocks of petroleum products covering at least 90 days of average daily consumption for the previous calendar year. Unlike the IEA, the EU lacks the necessary power to manage the stocks in a crisis. Member states own and control their own stocks. They are obliged to consult each other before releasing stocks from the strategic reserve. The EU is considering adding to its strategic oil reserves and managing them on a more centralized basis.

Stocks in IEA net importing countries have been declining during the past 15 years or so at the same time that OECD oil import dependence has been increasing. Meanwhile, the world oil market is growing, and much of the new oil demand will come from developing countries in Asia, which do not possess strategic oil stocks. Global oil security would be increased if Asian countries established minimum emergency stock requirements, as has been recommended by the Asia Pacific Energy Research Center.

The U.S. SPR was established in 1975 to help protect the nation against future oil market disruptions. Managed by the Department of Energy (DOE), the SPR oil was mainly acquired during the 1980s and is stored in underground salt domes along the coast of the Gulf of Mexico. At the end of 2003, the SPR contained 650 million barrels of oil. President George W. Bush has instructed the secretary of energy to proceed with further filling of the reserve up to its 700 million barrel capacity using principally royalty oil from federal offshore leases.

The president determines whether there is a severe supply interruption that warrants use of the SPR. If so, the secretary of energy announces an auction sale of a certain amount of oil at a price determined by the market. Bidders (oil companies) send their offers to the DOE and, after selection, submit a letter of credit or a cash deposit. The oil is then delivered through a pipeline or a marine terminal. The whole process can take 3 or 4 weeks to complete. The amount of import coverage afforded by the SPR depends on the level and rate of increase in imports. With U.S. net imports of approximately 11 mbd at the end of 2003, the current SPR offers approximately 53 days of import coverage if all imports were cut off, which is a very unlikely case given that U.S. oil imports come from a diversified group of nations.

There has been controversy about how and when to use the SPR. In the first Bush administration, the president refused to order an SPR drawdown despite a sharp escalation in oil prices in the fall of 1990 after Iraq invaded Kuwait (although the DOE did proceed with a SPR test run.) However, in January 1991, a drawdown of 33.75 million barrels of oil was authorized at the beginning of Desert Storm in coordination with stock draw by other allies under the IEA (later reduced to 17.3 million barrels of oil actually sold). That release, coupled with the success of the air war, helped to reduce oil prices and restore stability to the market. In September 2000, President Clinton authorized release of oil to bolster supplies at a time of high oil prices and low inventories of heating oil in New England before the start of winter. This was actually an oil swap, not a sale, because companies accepted obligations to take oil and return slightly more 1 year later. During the 2000 election campaign, candidate G. W. Bush criticized this action and argued for retaining the SPR for truly emergency uses.

In Germany, the government works closely with industry to administer a government-owned reserve, the Erdölbevorratungsverband, which stocks both crude oil and products to meet the 90-day IEA import requirement. Most other European countries rely on mandates to industry to hold private emergency stocks. In Japan, the Ministry of International Trade and Industry supervises government and private stocks.

3.6 Oil, Transportation, and Environmental Security

In the United States, transportation accounts for 27% of overall energy consumption and 68% of petroleum use. The petroleum share of transport energy in 1999 was 96%. Most transportation energy use (76%) comes from highway vehicles—light vehicles and trucks—with the rest coming from aircraft, boats and ships, pipelines, rail, and off-highway construction and farm vehicles. For European countries, transportation is also the most important and fastest growing sector for oil use, whereas the situation in Asia is more diversified, with more oil being used in industry and power generation. Oil use in transportation contributes importantly to air pollution and to greenhouse gases that are responsible for global climate change.

Transportation in the United States is a source of every major pollutant except sulfur dioxide. Of the criteria pollutants listed in the U.S. Environmental Protection Agency's (EPA) National Ambient Air Quality Standards, the oil/transport contribution (1999) is as follows: carbon monoxide, 77%; nitrogen oxides (NO_x), 55.4%; volatile organic compounds (VOCs), 47%; and very small amounts of particulate matter (PM-10) and sulfur dioxide. (Carbon dioxide is not yet regulated.) Although much progress has been made in reducing emissions from vehicles, and lead has been essentially phased out of gasoline, challenges remain. The most important air pollution challenge is low-level ozone, which is formed in the atmosphere by complex photochemical reactions involving VOCs and nitrogen oxides in the presence of sunlight. Exposure to high levels of ozone can cause coughing, eye irritation, sore throat, headache, and chest pain, especially for children and old people who have asthma conditions. Efforts to limit criteria pollutants by the EPA have been made under the Clean Air Act, originally passed in 1970 and amended several times.

There are still many U.S. cities that do not meet minimal standards for ozone, of which the precursors come from vehicle tailpipe emissions and evaporative emissions. The major problem seems to be NO_x emissions, which are increasing. (NO_x emissions also come from electric power plants.) Another problem is that although engines are much cleaner today, vehicle miles traveled continually increases as more cars are purchased. More efficient engines also tend to produce more NO_x. A third problem is that sport utility vehicles are being purchased in increasing numbers, and they have not been regulated as stringently as light-duty vehicles. They are less efficient and produce more emissions.

The Clean Air Act Amendments of 1990 provided for tier 1 standards on emissions from mobile sources covering nitrogen oxides in addition to carbon monoxide, hydrocarbons, and particulate matter for light-duty vehicles beginning with model year 1994. More stringent tier 2 standards will apply to all passenger vehicles starting in 2004, including a reduction in the sulfur content of gasoline in order to ensure the effectiveness of emission control technologies. Standards will also be tightened for heavy-duty vehicles (trucks) beginning in 2004, including a tighter combined NO_x and VOC standard. This will be followed in 2007 by new "ultra-low" sulfur content requirements for diesel trucks and specific NO_x emissions control technologies. Europe is not far behind the United States in strengthening its emissions requirements for mobile sources.

Oil use is also a major contributor to greenhouse gas emissions. According to the EIA, CO_2 emissions (the leading greenhouse gas) from industrialized countries accounted for 51% of the global total in 1999, followed by developing countries at 35% and transition economies in Eastern Europe/FSU at 13%. In the industrialized world, oil use contributed almost half (49%) of carbon dioxide emissions. The transportation sector is an important contributor. In the United States, transport contributed 33% of carbon dioxide emissions in 1999.

The Bush administration rejected participation in the Kyoto Protocol in March of 2001 and put forward an alternative voluntary plan to reduce the carbon intensity of the U.S. economy in spring 2002. However, it has not offered to regulate carbon dioxide emissions, which were absent from the administration's "Clear Skies Proposal" to strengthen limits on three other pollutants. In what may be a harbinger of the future, the state of California legislature passed a bill in July 2002, subsequently signed by the governor, calling for a plan by 2005 to reduce greenhouse gas emissions from cars and light trucks. Methods of achieving reductions are not specified but presumably might include fuel efficiency, hybrid engines, fuel cells, or other technologies. Cars and light trucks emit approximately 40% of greenhouse gas emissions in California. As has happened in the past, once again the state of California may be blazing a new trail in environmental regulations.

Currently, the future of the Kyoto Protocol is uncertain; it has been ratified by the EU and Japan but needs to be approved by the Russian parliament

before it can come into force. Several European countries have already been experimenting with carbon taxes.

3.7 Oil and Geopolitics, Foreign Policy, and International Security

Oil has implications for foreign policy, international security, and geopolitics. However, the implications are more international than national, except for the United States, for which the two are intertwined. The fact that two-thirds of the world's oil reserves are located in the Persian Gulf states makes them targets for aggression. The end of the Cold War and the disintegration of the Soviet Union removed the Soviet threat to the region. However, the Gulf remains an unstable place, subject to revolutions and regional wars, as the Gulf War of 1990–1991 demonstrated when Iraq attacked and annexed Kuwait and threatened Saudi Arabia. In that war, a UN military force was assembled, led by the United States, to protect Saudi Arabia and force Iraq to relinquish Kuwait. Although there were probably several reasons for the U.S. action, surely the protection of access to Gulf oil supplies was a major reason. It appears that the United States, as the world's remaining superpower, is the only country that can play this role.

A U.S. security role will continue to be needed in the Gulf, where there is no regional balance of power. That role is strengthened but has also become more complicated following the U.S. invasion and occupation of Iraq in 2003. Any threat to Gulf oil supplies would pose a threat to the world economy if supplies were cut off and oil prices escalated. The United States will oppose any serious threat to the world economy because it is increasingly interconnected with that economy, whatever the level of Persian Gulf imports into the United States (which is much lower today). Europe and Japan are much more dependent on imports of Gulf oil than is the United States.

As noted in "Energy Security," the DOE's 1987 report to the president, "Increased dependence on insecure oil supplies reduces flexibility in the conduct of U.S. foreign policy." A key example of this is U.S. relations with Saudi Arabia, for a long time considered a key ally in the Middle East presumably because of its leading role as an oil producer. However, Saudi Arabia is an autocratic regime, not a democracy, with weaknesses in the area of human rights and treatment of women, and it is the country that produced many of the terrorists who participated in the September 2001 attacks on the World Trade Center and the Pentagon. However, the "oil factor" reduces U.S. leverage and freedom of action vis-à-vis that country.

The globalization of the world oil market is widening the context of international oil security in the future. For example, the rapidly growing economies of Asia will import increasing amounts of oil from the Middle East in the years ahead. Oil tankers en route to Asia must pass through the narrow Strait of Malacca near Singapore. If the strait were closed by terrorists or military action, tankers would be required to add considerable distance and expense to deliver their oil, which would immediately raise freight rates worldwide. Will the United States sometime be asked to police this vital Strait?

Central Asia, as another example, is a region of increased interest for future oil production, assuming pipeline routes can be constructed to ship the oil to Western markets. However, the new republics around the Caspian Sea are not democracies and could become unstable during future transfers of power or revolutions or ethnic strife. A security framework for the region has yet to be devised.

Oil interacts with foreign policy in other ways. The United States maintains trade sanctions on many countries throughout the world, including Iraq (subject to UN sanctions after the Gulf War), Iran, and Libya. These three countries are major oil producers but they are off limits to U.S. companies for major oil development projects. Leaving Iraq aside as a special case, there is some question whether these sanctions are working effectively if they are not multilateral and accepted by other countries.

3.8 Strategies to Reduce Oil Dependence

The best long-term strategy to reduce the costs and risks of oil dependence lies in research and development of affordable alternatives to petroleum, especially in the transportation sector, in which there is a need for new technologies and fuels. Early efforts focused on developing vehicles that use compressed natural gas or liquefied petroleum gas or on alcohol fuel additives, such as ethanol or methanol, that can be blended with conventional gasoline. Electric vehicles have been tried, but they have been limited by the distances they can operate before batteries need to be recharged—and a breakthrough in battery technology has so far been illusive. Hybrid vehicles, which combine a small conventional internal combustion engine with an electric motor and offer increased efficiency and lower emissions, appear to be the best candidates for a transition period. Ultimately, fuel cell vehicles, which produce no

emissions except water and can run at high efficiency, are the best hope, although they would require a new infrastructure to provide hydrogen or a way to create it on board the vehicle. Fuel cells are still very expensive, but the second Bush administration has invested in a new government–industry partnership to promote their development.

It is unclear how much time this will take, but by 2010–2015 fuel cell vehicles could be available on the market, and by 2020 they could be expanding their market share. Such a development could have a major impact on reducing future world oil demand, although the transition to a new transportation technology will take a considerable amount of time.

4. RELIABLE ELECTRICITY SUPPLY AND PRICE VOLATILITY

Because electricity cannot be stored, security of supply is critical. Previously, this was ensured by monopoly utilities under government regulation. With the beginning of liberalized markets, there are new questions. Liberalization shortens contracts and threatens adequate investment. The California electricity crisis of 2000–2001 brought the problem sharply into focus.

The California crisis occurred 2 years after the state reformed its power market. In June 2000, in a period of exceptionally hot weather and in which there were some problems with grid operation, there were rolling blackouts in the San Francisco Bay area. Later in the summer, electricity prices tripled in the southern part of the state, and San Diego Gas and Electric asked the Federal Energy Regulatory Commission for price controls in wholesale markets. High wholesale prices led to questions about market power. The governor signed legislation placing rate caps on residential and small commercial users in southern California. In December, the price of electricity spiked to 30 ¢/Kwh. Pacific Gas and Electric (PG&E) and Southern California Edison, facing high pool prices and unable to raise retail prices, later defaulted on $12 billion of debt. PG&E eventually declared bankruptcy.

In early 2001, California experienced a series of short-duration rolling blackouts. Meanwhile, Governor Davis signed an emergency order authorizing California's Department of Water Resources to become a temporary buyer of power, allowing it to float revenue bonds to finance power purchases under long-term contracts. The state's experiment with deregulation and a power pool had apparently failed.

The details of what happened in California are still being investigated. There was a combination of bad planning and bad luck. A 3-year drought in the Northwest had reduced hydroelectric capacity normally available for import into California. Natural gas prices spiked during the same period because of insufficient pipeline capacity at the border and within the state. There was physical congestion in north–south transmission. When pool prices escalated, market power among wholesale producers (many of them out of state) became a problem. However, one of the most significant facts was long-term underinvestment in electric power plants in a state in which population, and therefore electric demand, was increasing. This was not an optimal condition on which to proceed with full-scale liberalization of the electric market.

As highlighted by a recent IEA study, "Security of Supply in Electricity Markets," during the past 20 or 30 years the electric systems of most OECD countries have generally maintained adequate reserve margins, even to the point of overinvestment, which ensured security of supply but at additional costs to consumers. In a liberalized market in which market players bear the costs and risks, the incentives to overinvest are removed. At the same time, there may be market imperfections that hinder the ability to achieve reliable supply, including limited demand-side response to market conditions, price distortions, policy barriers for certain technologies or the use of certain fuels, or cumbersome licensing procedures for new power plants.

The United States is midway in the complex process of restructuring the electric industry, previously regulated at both federal and state levels. Some states are further along in the process and have done better than California at making competition work (e.g., Pennsylvania and New York). Other states are holding back. The Federal Energy Regulatory Commission, which has played an important role in the process, is pressing utilities and states to form regional transmission organizations and has issued an important new proposed rule on standard market design. Meanwhile, Congress is considering new energy legislation, which may provide more guidance at the federal level.

In Europe, the United Kingdom and the Nordic countries have led the movement toward liberalized markets. Also, the EU has begun a process with its 1997 electricity directive to move the EU countries collectively toward more competition in the electric industry.

The issue of security of the U.S. electric supply was highlighted more broadly by a major power blackout on August 14, 2003, whcih lasted up to 2 days and affected some 50 million people in the Midwest and Northeast as well as in Ontario, Canada. The blackout reportedly began with downed power lines at a utility in Ohio. A U.S.–Canada Task Force is investigating the causes. The incident may lead to a stronger government role in enforcing reliability standards in the electric industry and in modernizing an aging power grid.

5. NATURAL GAS AND PRICE VOLATILITY

Natural gas is the fastest growing energy source worldwide, and its share of total energy consumption, according to the EIA, is projected to increase from 23% in 1999 to 28% in 2020. For the United States, the gas share is predicted to increase from 23.6% in 2000 to 26.5% in 2020. Gas is the cleanest of the fossil fuels, and it is the preferred choice for new electric power plants driven by gas turbines. Coal is likely to remain the lead fuel for power generation, but natural gas will expand its share considerably.

Gas is more difficult to transport than oil and requires pipelines or special liquefied natural gas (LNG) tankers. For this reason, natural gas tends to be traded in regional markets (e.g., North America, Europe, and Asia). However, although the amount of gas traded internationally (20%) is much less than the amount of oil (50%), international gas trade is also growing via international pipelines and expanded LNG trade.

There is no shortage of natural gas in the world. The USGS reports in its 2000 "World Petroleum Assessment" that only 10% of worldwide gas resources have been produced (compared to 25% for oil). Considerable volumes of gas remain to be discovered and developed. A major problem has been that much of the world's gas is located far from demand centers. More than half of the world's remaining reserves are in the former Soviet Union, Middle East, and North Africa. Two countries, Russia and Iran, hold approximately 45% of the world's reserves. Thus, there is some concentration of reserves.

Western Europe, which holds only approximately 2% of world gas reserves (mostly in the North Sea), imports approximately 40% of its requirements, mainly from Russia and Algeria. European gas demand is likely to expand rapidly during the next 20 years, which will require increased imports. The European gas industry is also in the process of deregulation following an EU directive in 1998 that set forth a staged plan of achieving more competitive gas markets.

In 1995, the IEA study on gas security concluded that most European countries could withstand gas supply interruptions, but it also noted that gas infrastructure is less flexible than that of oil. Could gas supply security become an issue in the future? If so, it might be desirable to establish a framework for gas security with obligations to hold strategic stocks and develop demand restraint programs. Since the power industry will be using increasingly more gas, reliable gas supply is crucial to the functioning of the electricity industry.

The United States has considerable gas supplies, but it imports approximately 15% of its gas needs from Canada. It also receives other gas imports via LNG. Although LNG shipments are currently small, they are projected to increase.

The U.S. gas industry has been deregulated and functions mostly on short-term contracts and spot sales. On the whole, the liberalized U.S. gas market has worked well. However, there have been some problems. In 2000, natural gas prices increased dramatically and were pushed even higher by very cold weather in November and December. Prices remained high through the first half of 2001. The mean price range during the period was $2.53–7.85 per million Btu, up from $1.98 per million Btu during 1995–1999. The spot price reached more than $10 per million Btu at the Henry Hub at the end of December 2000. What was troubling was the length of time that prices stayed at high levels. High-demand growth and cold weather, plus inadequate gas reserves in storage, explain part of the problem. So does the inelastic short-term supply response to price. Later in 2001, gas prices declined because of the slowdown in the economy and milder temperatures.

Although gas pipelines were used to capacity during the crisis period, there were apparently few infrastructure constraints except in and near California, where transmission capacity was not adequate to transport all the gas needed and natural gas prices spiked higher than elsewhere in the U.S. market (and contributed to the California electricity crisis).

In 2003, at the request of the Secretary of Energy, the National Petroleum Council published an

industry study, "Balancing Natural Gas Policy: Fueling the Demands of a Growing Economy." The study notes that traditional North American gas producing areas can only meet 75% of projected U.S. gas demand in 2025. This includes Canadian production which is reaching a plateau. The gap could be filled by encouraging increased production in the Rocky Mountains, by Arctic gas (which requires a new pipeline), and by expanded LNG imports. At the same time there will be a need to expand gas infrastructure and to deploy new technologies to increase the efficiency of gas use. The study notes that gas price volatility will likely continue, reflecting the variable nature of supply and demand in a free market.

6. TERRORISM AND ENERGY SECURITY

The September 11, 2001, attacks against the United States by the Al-Quaeda group have raised the specter that future terrorist assaults could be made against U.S. energy infrastructure. The subject is not new. Oil companies operating in Columbia have suffered attacks against pipelines and other facilities by guerillas operating in that strife-torn country. However, it is new for the United States. Although electrical/nuclear power plants might be considered especially vulnerable targets, certainly oil refineries and, to a lesser extent, oil and gas pipelines are potential targets as well. New efforts are being made to increase physical security around these facilities, but it is a daunting task. Cyber attacks against energy systems are also a possibility.

The analysis of energy infrastructure vulnerabilities has begun and will require extensive government–industry cooperation. A new Office of Energy Assurance was created soon after September 11, 2001, in the DOE, but it may be folded into the new Department of Homeland Security if endorsed by Congress.

Although physical security can be strengthened at large power plants and petroleum refineries, it will undoubtedly be impossible to protect power lines and pipeline systems from attack. Instead, strategies need to be developed to promote redundancies and reserves in our energy systems to be able to respond to interruptions and reinstate normal service as quickly as possible.

A study of this subject by the U.S. Energy Association, "National Energy Security Post 9/11," recommends some initial actions, which require the application of advanced energy technologies. First, detailed risk assessments need to be made of North American energy system vulnerabilities to terrorism, which might be called probabilistic vulnerability assessments, and corresponding countermeasures. Second, because many existing control systems are linked to the Internet and are therefore vulnerable to cyber attacks, steps should be taken to develop secure, private communications networks for energy systems, with adequately protected back-up systems. Third, with regard to the widely dispersed electricity grid, which is particularly vulnerable, a more flexible strategy of intelligent "islanding" is needed to minimize the impacts of an attack, along with self-healing mechanisms. For major energy facilities, there is a need to develop and install sensors to detect dangerous biological and chemical agents, which might be released in cooling towers, and to identify and deploy countermeasures. Contingency planning is needed for all kinds of attack scenarios against centralized and local facilities.

SEE ALSO THE FOLLOWING ARTICLES

Geopolitics of Energy • Gulf War, Environmental Impact of • Inflation and Energy Prices • National Energy Policy: United States • Nationalism and Oil • Oil Crises, Historical Perspective • Oil Industry, History of • Oil Price Volatility • OPEC, History of • OPEC Market Behavior, 1973–2003 • War and Energy

Further Reading

Adelman, M. A. (1995). "The Genie out of the Bottle: World Oil since 1970." MIT Press, Cambridge, MA.

Asia Pacific Energy Research Centre. (2003). Emergency oil stocks, and energy security in the APEC region, APEC No. 00-RE-01.2. Asia Pacific Energy Research Centre, Tokyo.

Campbell, C., and Laherrere, J. (1998, March). The end of cheap oil. *Sci. Am.* **278**(9), 78–83.

Deffeyes, K. S. (2001). "Hubbert's Peak: The Impending World Oil Shortage." Princeton Univ. Press, Princeton, NJ.

Energy Information Administration (2001). "U.S. Natural Gas Markets: Mid-term Prospects for Natural Gas Supply." Energy Information Administration, Washington, DC.

Energy Information Administration (2002). "International Energy Outlook, 2002." Energy Information Administration, Washington, DC.

Greene, D. L., and Tishchishyna, N. I. (2000). Costs of oil dependence: A 2000 update, ORNL/TM-2000/152. Oak Ridge National Laboratory, Oak Ridge, TN.

Greene, D. L., Jones, D. W., and Leiby, P. N. (1997.). The outlook for U.S. oil dependence. *Energy Policy* **26**(1), 55–69.

International Energy Agency (2001a). "World Energy Outlook, 2001 Insights." Organization for Economic Cooperation and Development, Paris.

International Energy Agency (2001b). "Oil Supply: The Emergency Response Potential of IEA Countries in 2000." Organization for Economic Cooperation and Development, Paris.

International Energy Agency (2002). "Security of Supply in Electricity Markets." Organization for Economic Cooperation and Development, Paris.

National Energy Policy Development Group (2001). "National Energy Policy." National Energy Policy Development Group, Washington, DC.

National Petroleum Council (2003). "Balancing Natural Gas Policy: Fueling the Demands of a Growing Economy," vol. **1**, Summary. Washington, DC.

Quarterly Review of Economics and Finance **42** (2002). [Special issue on the oil market]

U.S. Energy Association (2002, June). "National Energy Security Post 9/11." U.S. Energy Association, Washington, DC.

U.S. Geological Survey (2000). "World Petroleum Assessment." U.S. Geological Survey, Washington, DC.

Yergin, D. (1991). "The Prize: The Epic Quest for Oil, Money and Power." Simon & Schuster, New York.

Natural Gas, History of

CHRISTOPHER J. CASTANEDA
California State University
Sacramento, California, United States

1. **Origins**
2. **Natural Gas in Fredonia, New York**
3. **Early Commercial Utilization**
4. **Natural Gas in Pittsburgh**
5. **Natural Gas in the Southwest**
6. **Long-Distance Pipelines**
7. **Natural Gas in the Great Depression**
8. **Appalachian Gas and Federal War Planning**
9. **Gas in the Postwar Era**
10. **Deregulation**

Glossary

Appalachia The mountainous region stretching from northern Mississippi to southern New York and commonly characterized by rural communities and poverty.
British thermal unit (Btu) The amount of heat required to change the temperature of 1 lb of water 1°F at sea level.
Federal Energy Regulatory Commission (FERC) The successor to the Federal Power Commission; created in 1977.
Federal Power Commission (FPC) The federal regulatory agency responsible for regulating the interstate natural gas industry; created in 1920 and abolished in 1977.
hydrocarbons Organic compounds that are composed entirely of carbon and hydrogen. Petroleum products are composed of hydrocarbons, and methane, or natural gas (CH_4).
Insull, Samuel (1859–1938) A leader in the American public utility industry during the early 20th century. He was born in London and served as Thomas Edison's secretary as a youth.
manufactured coal gas A fuel gas for illuminating and heating purposes produced by heating coal in a retort and capturing the resulting vapors for distribution.
waste gas Natural gas, considered to be a nuisance in the production of oil.

Natural gas is a vital fuel for modern society. During the last 50 years of the 20th century, natural gas satisfied approximately 25% of U.S. energy demand. It has been used for electric power generation, industrial heat processes, domestic heating and cooking, and transportation fuel. Natural gas is composed primarily of methane, a hydrocarbon consisting of one carbon atom and four hydrogen atoms (CH_4). As a "fossil fuel," natural gas is rarely pure. It is commonly associated with petroleum and often contains other hydrocarbons, including butane, ethane, and propane. In the United States, substantial natural gas utilization did not begin until after the discovery of large quantities of both crude oil and natural gas in western Pennsylvania during 1859.

1. ORIGINS

Natural gas was observed and utilized in limited quantities during ancient times. References in literature to burning springs, burning bushes, or perpetual lights suggest that natural gas was used, albeit rarely, for heating. In ancient China, burning gas springs heated brine water in order to extract salt, and there were flaming gas springs in Greece and Rome. Recorded observations of burning springs in France, Italy, and Russia also exist. The philosopher Plutarch and theologian St. Augustine described lights that may have been produced by burning natural gas.

In colonial America, both George Washington and Thomas Jefferson observed natural gas springs. During the autumn of 1770, Washington participated in an expedition along the Ohio and Kanawha rivers in West Virginia and Ohio. Near the present-day town of Pomeroy, Ohio, Washington described a location "wch. the Indians say is always a fire." About perhaps the same site, Thomas Jefferson recorded his observations of "a hole in the earth…from which issues constantly a gaseous stream." Other visitors to these springs reported that hunters used them to cook food.

 © 2004 Elsevier Inc. All rights reserved.

Through the early 19th century, these "burning springs" had little practical or widespread use. Most importantly, there was no practical method of either capturing the gas emanating from the springs or storing and redirecting its flow through piping.

2. NATURAL GAS IN FREDONIA, NEW YORK

Residents of Fredonia, New York, were perhaps the first Americans to use natural gas for lighting on a regular basis. Gas springs in the vicinity of Fredonia had been observed in the early 1800s, but it was not until the mid-1820s that a local gunsmith named William Aaron Hart organized an apparently successful effort to utilize gas from the local gas spring to provide light for local homes and establishments.

Some accounts written much later state that citizens of Fredonia used natural gas to illuminate their town when the French military leader Marquis de Lafayette visited. Lafayette toured America during the years 1824–1825, and he traveled to New York in the summer of 1825. Of his visit to Fredonia, Lafayette's private secretary, A. Levasseur, recorded that they had observed a great many lights in the town. The local newspaper featured a story on the same events and noted lamps and chandeliers that provided illumination in the town during Lafayette's visit. Contemporary reports of Lafayette's visit do not mention gaslights at Fredonia; only the accounts of this event written much later mention gas lighting. Lafayette's secretary did note gas lighting at other locations. While in a Boston theater, Levasseur recorded observations of "gas blazing abundantly from numerous pipes, and throwing floods of dazzling light over the hall." These lights were fueled by manufactured coal gas, however, and not natural gas.

After Lafayette's visit, William Hart continued to develop his interest in natural gas in Fredonia. During 1827, he began work on a plan to supply natural gas to a lighthouse at nearby Barcelona Harbor. After the U.S. government granted him a contract for this service, he installed a primitive gas works. It consisted of a fish barrel placed over the gas spring located at Westfield along Lake Erie. The barrel served as a "gasometer," or gasholder. Hart sealed the gasometer and transported the gas for one-half mile through hollowed out pine logs to the lighthouse. Gas from the spring provided enough fuel to illuminate 144 burners and create a bright light.

3. EARLY COMMERCIAL UTILIZATION

It was not until Colonel Edwin Drake discovered oil in Titusville, Pennsylvania in 1859 that natural gas became a significant source of energy in the United States. Although Drake had been searching for oil, he found natural gas as well; oil and natural gas are often found in the same geologic structures. Natural gas discovered in eastern Pennsylvania was marketed to regional customers. Therefore, the Drake discovery heralded the beginning of both the modern U.S. oil and natural gas industries.

Prior to Drake's discoveries, there were few successful long-term attempts to utilize natural gas for either industrial or commercial purposes. By the mid-19th century, only those factories or towns located very near a natural gas well could utilize the fuel. The difficulty of containing a natural gas spring, storing the gas, and transporting it over long distances limited its utility. For example, significant natural gas discoveries such as the high-volume well discovered by William Tomkins in 1841, near Washington's burning spring on the Canadaway Creek, attracted some attention but little commercial interest. Alternatively, manufactured coal gas plants could be built and operated anywhere as long as coal, or the feedstock, was readily available. Thus, the manufactured coal gas industry developed much

TABLE I

Introduction of Manufactured Gas to Major Cities[a]

Year	City
1816	Baltimore
1825	New York City
1829	Boston
1832	Louisville
1835	New Orleans
1836	Philadelphia
1843	Cincinnati
1846	St. Louis
1849	Chicago
1854	San Francisco
1867	Kansas City
1867	Los Angeles
1871	Minneapolis
1873	Seattle

[a] The dates primarily reflect the year manufactured gas was first produced in the city for commercial use. In some cases, however, the date reflects when a city charter was granted, and charters were sometimes granted before and even just after gas service began.

more quickly than that of natural gas in the 19th century. By the mid-19th century, many towns and cities had a manufactured gas plant and a local distribution system that provided some coal gas for residential and business lighting. (see Table I).

The earliest recorded use of gas for industrial purposes in the United States occurred in 1840, near Centerville, Pennsylvania. The gas was used to distill salt from brine water. Gradually, in the 1860s and 1870s, local deposits of natural gas were utilized for a variety of industrial heating applications. Even in Fredonia, New York, where some residents and shop owners utilized natural gas for lighting beginning in the mid-1820s, a formal natural gas company was not organized for many years. In 1858, businessmen established the Fredonia Gas Light and Water Works Company to operate the local gas wells and discover new ones.

Natural gas was not used on a large scale until the 1880s, and gas wells were most likely to be abandoned when oil was not concurrently discovered. An example of early abandonment occurred in 1865 when a 480-foot drilling effort struck a natural gas reservoir near West Bloomfield, New York. The operators estimated the gas flow to be about 2000 cubic feet (mcf) per day; they directed the gas into a large balloon and attempted to measure the flow by calculating the time required to fill it. Because the investors were disappointed that oil was not discovered, they abandoned the project. Not everyone was disappointed that this well contained only natural gas. Several businessmen formed the Rochester Natural Gas Light Company and purchased the same gas well in 1870. The nearest town desiring natural gas was Rochester, about 25 miles away. The company constructed a pipeline system to connect the well with the town. They built a pipeline out of Canadian white pine. The 2- to 8-foot log segments were planed to a uniform 12.5-inch exterior diameter, and they were bored for an 8-inch interior diameter. Construction and maintenance of the wood pipeline system was particularly problematic, but the company began transporting natural gas during the winter of 1872. Consumers in Rochester discovered quickly that hotter burning natural gas was not easily interchangeable in their burners with manufactured coal gas. This situation resulted in lower than expected natural gas demand. Continuing problems with gas transportation facilities caused significant problems for the company; rotting and leaking wood pipelines simply prevented the adequate transportation of natural gas from well to consumer. Soon, the company stopped operations.

A more successful attempt to transport natural gas took place in 1872. New natural gas discoveries created a demand for specialized gas pipelines. In this case, a 2-inch wrought-iron line was constructed and used to connect a gas well 51 miles away. The line transported "waste gas" from nearby oil fields to Titusville. This pipe transported 4 million cubic feet (mmcf) per day to 250 customers, both residential and industrial.

The primary obstacle to the expansion of the natural gas industry in the mid-19th century was inadequate pipeline facilities and technology, not lack of supply. Hollow log pipelines leaked and disintegrated, but cast and wrought iron lines also suffered from significant intrinsic defects. Wrought iron lines in the period 1872–1890 were typically less than 8-inches in diameter, and the pipe segments were attached with couplings tightened with screws. Gas leaks were common problems. Most of the gas transported in pipelines during this period flowed under the natural pressure of the well without the aid of additional compression.

4. NATURAL GAS IN PITTSBURGH

Pittsburgh became the first major U.S. city in which industry utilized large volumes of natural gas for industrial heat processes. Abundant Pittsburgh area coal deposits and the importation of iron ore from the Juanita region in central Pennsylvania (and later from the Mesabi range) facilitated development of a substantial iron industry originally fueled by coal. Extensive coal burning for industrial heat created significant air pollution, and Pittsburgh became known as the "Smoky City." Contemporary newspaper articles noted the black smoke produced by burning coal. In 1884, *The New York Times* reported that natural gas would be used in Pittsburgh's industries to reduce coal smoke pollution.

The earliest recorded use of natural gas in a Pittsburgh iron works occurred in 1870–1871, but widespread natural gas utilization did not commence until the early 1880s, after the development of nearby gas wells. Entrepreneurs then organized new regionally based gas firms. One group of Pittsburgh area manufacturers established the Chartiers Valley Gas Company in 1883 to transport natural gas from local gas fields to their glass and steel plants. This company's first line extended from the Hickory gas field to Pittsburgh. The wrought-iron line was the first "telescoping" pipeline, meaning that a smaller diameter pipe installed at the well's origin led to a larger diameter pipe in the city. The telescoping line system

was useful for lowering the gas line pressure as gas flowed into the city. For pipe less than 12 inches in diameter, the typical connection was a screw coupling. Pipe segments were threaded on the outer edge of each length end that turned into a screw coupling.

As of 1886, the Chartiers firm also laid claim to operating the largest continuous pipe in the world. The company installed a 16-inch pipe extending from the Murrysville gas field to Pittsburgh. After 8 miles, the 16-inch line was fed into a 6-mile-long 20-inch pipe, and it in turn fed into a 5-mile section of 24-inch cast-iron tubing, tested at 300 pounds per square inch (psi). The National Tube Works constructed this line; J. P. Morgan controlled National Tube.

By the late 1880s, Pittsburgh had become the locus of the American steel and coal industry, and it was also the center of the natural gas industry. In 1886, there were 10 iron and steel mills using natural gas in their puddling furnaces, with many more planning to convert to gas. Six glass-making factories and reportedly every brewery in Pittsburgh used natural gas instead of coal. The Sampson Natural Gas Crematory also used the invisible fuel. Pittsburgh received its natural gas from the lines of six companies tied into 107 regional gas wells. Five hundred miles of pipeline transported natural gas from wells to the city, including 232 miles of line within the Pittsburgh city limits.

As natural gas utilization increased, local engineers addressed technological problems associated with its transportation. Solomon R. Dresser focused attention on drilling and oil-field-related technology. In 1880, he formed S. R. Dresser & Company and conducted pioneering work in pipe coupling. In 1887, Dresser received a patent for using a rubber ring in pipe joints to create a leakproof coupling. Although this method proved not entirely satisfactory, less than a year later Dresser designed a second coupling that was more effective. He developed a two-part mechanical device that pulled the pipe segments together. Between the tightened sections, an internal rubber ring created a seal. Dresser proved the leakproof qualities of this coupling method when he developed his own gas field near Malta, Ohio and used the couplings in a gas line that extended into town. The Malta Natural Gas Line established Dresser as a leader in the natural gas business, and his couplings attracted widespread favor; gas companies located throughout the country ordered them. As much as 90% of the gas pipeline industry used these couplings into the 1920s.

Improved couplings not only reduced leakage, they also lessened the possibility of explosions. In Pittsburgh, city ordinances prohibited gas lines from operating at pressures higher than 13 psi. This pressure limitation was intended to reduce the leaks prevalent in more highly pressurized lines. Leaks often resulted in accumulations of gas in cellars, leading to explosions and fires. Within the city, regulating valves further reduced the gas pressure. To prevent leaking gas from ending up in residential cellars, the Chartiers Valley Company used its patented "broken stone escape system." This system involved laying a pipe in a trench filled with dirt to the center of the pipeline. Workers then placed about 9 inches of broken stone on top of the line. A layer of tarpaper was then placed over the stone; dirt covered the tarpaper. The stone barrier was placed adjacent to every city lamppost so that escaping gas could vent through the stone. In addition, gas firms used "escape pipes," very small diameter lines leading from each pipe joint to a lamppost. Inspectors checked each escape pipe for possible leaks and identified the joint to which each escape line was connected. A system of 4-inch pipes distributed gas to individual residences. In these pipes, gas pressure was limited to about 5 psi. As the gas entered homes, an additional regulator/shutoff valve lowered gas pressure again to about 5 ounces per square inch, so that gas could be burned satisfactorily in gaslight fixtures.

George Westinghouse, inventor of the railroad air brake and a resident of Pittsburgh, also became involved in the expanding natural gas industry. He explored for natural gas close to home. He drilled for gas in his own backyard located in a fashionable Pittsburgh neighborhood. In late February 1884, a small volume of gas began flowing from the well. The workers continued drilling to a depth of about 1560 feet. At 3 a.m. one morning, Westinghouse awoke to the sound of a tremendous explosion and the loud sound of hissing gas from the well.

Westinghouse needed a company organization to proceed with his new plan of selling his natural gas to Pittsburgh-area customers. He purchased a moribund company, the Philadelphia Company, to produce the fuel. As President and Director of the company, Westinghouse watched the firm become one of the largest gas businesses in the Pittsburgh area. For additional supply, the company leased substantial gas production acreage in western Pennsylvania. By 1887, the Philadelphia Company supplied approximately 5000 residential and 470 industrial customers with gas from about 100 natural gas wells located on 54,000 acres.

Westinghouse's financial participation in the natural gas business brought his inventive mind in touch with

some of the major problems of this new industry. Between the years 1884 and 1885, he applied for 28 gas-related patents, and during his lifetime he applied for a total of 38 gas equipment patents. Some of Westinghouse's most important inventions for natural gas included a system for enclosing a main gas line in residential areas with a conducting pipe to contain gas leaks. Westinghouse also developed a method for "stepping down" the highly pressurized gas in main trunk lines to lower pressure in residential areas. To prevent accumulations of gas in homes and shops after gas service was shut down and then restarted, Westinghouse patented a pressure regulator and cutoff valve that automatically restricted gas flow when the pressure dropped below a particular point. Tragedies nonetheless occurred. On the morning of January 31, 1885, two major explosions at Thirty-fifth and Butler streets in Pittsburgh nearly leveled an entire city block, killing two and injuring 25 others, some severely. The first explosion occurred at George Hermansdorfer's butcher shop after an accumulation of gas in his cellar; two or three people were badly burned. People rushed to investigate the explosion when a second, larger explosion occurred nearby. Subsequent explosions caused substantial injury to life and property, damaging as many as 15 buildings. Local residents threatened a riot against the gas company, and a representative of the Fuel Gas Company made a stunning admission: the pipes had not been tested before the gas was turned on. Efforts to develop gas regulators and emergency shutoff valves were absolutely required to ensure that this fuel could be utilized safely.

Andrew Carnegie, Pittsburgh's foremost entrepreneur, understood that natural gas had superior heating characteristics. He wrote: "In the manufacture of iron, and especially in that of steel, the quality is also improved by the pure new fuel. In our steel rail mills we have not used a pound of coal for more than a year, nor in our iron mills for nearly the same period." The iron and steel maker also noted that natural gas burned much more cleanly compared to coal.

By 1885, 150 companies had charters to sell gas in Pennsylvania, but the future of natural gas was not certain. Gas fields tended to exhaust themselves within several years after discovery. Selwynn Taylor, a Pennsylvania mining engineer, believed that most regional natural gas fields would soon be exhausted, and the price of coal would rise to the levels existing prior to the discovery of regional gas fields. His beliefs were typical of the time: existing natural gas fields and current production, transportation, and distribution systems simply could not supply enough gas to satisfy the demand and natural gas was ultimately unreliable. Fears of short-lived wells aside, gas discoveries in other Appalachian states, first in Ohio and then West Virginia, made this fuel economically significant to the entire region. Industries located in cities such as Buffalo, Cleveland, Toledo, and Cincinnati all began using natural gas from nearby wells. Waste and poor planning, however, led to many failed ventures. In one episode, the Indiana Natural Gas & Oil Company had built the longest pipeline to date in 1891. The transmission system consisted of two parallel 8-inch lines extending from northern Indiana gas fields to Chicago, a distance of approximately 120 miles. These lines transported natural gas at 525 psi. The supply quickly declined and the lines were soon removed from service. Episodes such as this but on a smaller scale were repeated throughout the region.

Similar supply problems in Indiana continued. During the late 19th century, an area covering 7000 square miles included a large number of producing natural gas fields. Despite attempts to regulate the production and flow of natural gas, unrestrained gas demand soared in the state. By 1907, many of Indiana's once productive natural gas fields had expended their valuable fuel, and many natural gas customers had to return to manufactured gas utilization. Gas discoveries in Oklahoma and in the eastern and southern Kansas gas fields suffered similar stories of rapid development followed by depletion. Episodes such as these characterized the natural gas industry, as opposed to manufactured gas, as fairly undependable.

By the turn of the century, the natural gas industry was most developed in the Appalachian region. Productive gas wells in West Virginia, Pennsylvania, New York, Kentucky, Tennessee, and Ohio led to the establishment of regional gas firms that built pipelines to serve local markets in the entire region. Natural gas was used primarily for industrial purposes, but, where available, its higher heating content meant that it was a superior cooking and heating fuel, although appliances for these purposes were still not widely available until later in the 19th and early 20th centuries. Natural gas was a promising fuel, but its limited availability and dependability forced entrepreneurs to proceed cautiously with plans to develop fields and build pipelines.

5. NATURAL GAS IN THE SOUTHWEST

The discovery of massive southwestern natural gas fields and technological advancements in

long-distance pipeline construction dramatically altered the early 20th-century gas industry market structure. In 1918, drillers discovered a huge natural gas field that became known as the Panhandle Field, situated primarily in North Texas. In 1922, drillers discovered the Hugoton Field, located in the common Kansas, Oklahoma, and Texas border area (generally referred to as the midcontinent area). The combined Panhandle/Hugoton fields became the nation's largest gas-producing area, comprising more than 1.6 million acres. It contained as much as 117 trillion cubic feet (tcf) of natural gas and accounted for approximately 16% of total U.S. reserves in the 20th century.

As oil drillers had done earlier in Appalachia, they initially exploited the Panhandle Field for petroleum only while allowing an estimated 1 billion cubic feet/day (bcf/d) of natural gas to escape into the atmosphere. As new gas markets appeared, the commercial value of southwestern natural gas attracted entrepreneurial interest and bolstered the fortunes of existing firms. These discoveries led to the establishment of many new southern firms, including the Lone Star Gas Company, Arkansas Louisiana Gas Company, Kansas Natural Gas Company, United Gas Company, and others, some of which evolved into large natural gas companies.

The sheer volume of these southwestern gas fields and their distance from distant urban markets emphasized the need for advancements in pipeline transport technology. In particular, new welding technologies allowed pipeline builders in the 1920s to construct longer lines. In the early years of the decade, oxyacetylene torches were used for welding, and in 1923, electric arc welding was successfully used on thin-walled, high-tensile-strength, large-diameter pipelines necessary for long-distance compressed gas transmission. Improved welding techniques made pipe joints stronger than the pipe, and seamless pipe became available for gas pipelines beginning in 1925. Along with enhancements in pipeline construction materials and techniques, gas compressor and ditching machine technology improved as well. Long-distance pipelines became a significant segment of the gas industry beginning in the 1920s.

These new technologies made possible for the first time the transportation of southwestern natural gas to midwestern markets. Soon, the southwest supplanted Appalachia's position as the primary region for marketable gas production. Until the late 1920s, virtually all interstate natural gas transportation took place in the northeast, and it was based on Appalachian natural gas production. In 1921, natural gas produced in West Virginia accounted for approximately 65% of interstate gas transportation whereas only 2% of interstate gas originated in Texas. Most interstate gas flowed into western Pennsylvania and Ohio. Appalachian fields experienced serious depletion in the 1920s, however, and various state legislators attempted to prohibit out-of-state gas exportation. These attempts to corral natural gas for intrastate utilization were largely unsuccessful.

Between the mid-1920s and the mid-1930s, the combination of abundant and inexpensive southwestern natural gas production, improved pipeline technology, and increasing nationwide natural gas demand led to the creation of the new interstate gas pipeline industry. Metropolitan manufactured gas distribution companies, typically part of large holding companies, financed most of the pipelines built during this era. Despite the high cost of the long-distance lines, access to natural gas even for mixing with existing manufactured gas could be a profitable venture. Natural gas was so abundant it was often substantially less costly than coal gas.

In 1927, Cities Service built the first long-distance line originating in the Panhandle field. This 250-mile, 20-inch pipeline connected the Panhandle field with a Cities Service gas distributor in Wichita, Kansas. Standard Oil (New Jersey) also participated in several significant pipeline ventures during these years. The first of these was Colorado Interstate Gas Company. Organized in 1927 by Standard, Cities Service, and Prairie Oil & Gas, this firm built a 350-mile, 22-inch line originating at the Texas–New Mexico border and extending to Denver. In California, natural gas from the Buena Vista field in the San Joaquin Valley fueled industry and commercial establishments in Los Angeles, and in 1929, Pacific Gas & Electric (PG&E) constructed a 300-mile pipeline from the Kettleman field north of Los Angeles to bring natural gas to San Francisco. San Francisco was one of the first major urban areas to switch from manufactured gas to natural gas. Because the same volume of natural gas had nearly twice the heating content as coal gas, burners and airflow valves in stoves and water heaters had to be adjusted to accept the natural fuel. With near military precision, PG&E divided San Francisco into 11 districts that were successively converted to natural gas. Six hundred trained men divided into 35-member crews converted PG&E's service area within 5 months. The conversion of 1.75 million appliances cost $2 million, but natural gas was less costly for the utility to market compared to coal gas.

New long-distance gas lines and expensive conversion programs were necessary if gas utilities were

TABLE II

Estimated Waste of Natural Gas in the United States in Billions of Cubic Feet[a]

	Natural gas waste[b]		
Year	Total U.S.	Texas panhandle	Total U.S. natural gas consumption
1919	213	n/a	256
1920	238	n/a	286
1921	193	n/a	248
1922	233	n/a	254
1923	416	n/a	277
1924	343	n/a	285
1925	324	n/a	272
1926	417	220	289
1927	444	405	296
1928	412	351	321
1929	589	294	360
1930	553	252	376

[a] *Source*. Federal Trade Commission, "Report to the Senate on Public Utility Corporations," Senate Document No. 92, 70th Congress, 1st Session, Part 84-A, 1935, pp. 93 and 95.

[b] Waste means gas production that was flared or vented and otherwise not utilized. n/a, data not available.

going to meet consumer demand. The new holding companies marshaled tremendous amounts of capital to build pipelines, extend service, and promote gas utilization. They also became adept at advertising and marketing. Trained salesmen, company servicemen, and even co-opted plumbers touted gas. During the 1920s, utility companies offered for sale a wide variety of gas-powered appliances, including space heating units, water heaters, stoves, and even gas-powered refrigerators. By 1926, about 50,000 automatic water heaters had been installed in homes, but gas appliances were not inexpensive.

Another use for natural gas beginning in the late 19th century was carbon black production. Produced by burning natural gas, carbon black was used for coloring in paint and inks. It was also used as a reinforcing agent in rubber and automobile tires. Natural gas produced in fields not connected by pipelines to urban markets was a prime candidate for carbon black production. Even by the late 1930s, about two-thirds of the amount of marketable gas produced was either flared, vented, or used to make carbon black. (see Table II). But greater profits awaited entrepreneurs willing to finance pipelines connecting gas fields to urban and industrial gas markets. Urban natural gas utilization also brought forth efforts to develop a standardized odorant. Unlike coal gas, which typically has a distinct smell, natural gas is odorless. Thus, a leak or inadvertently opened valve might allow odorless gas to accumulate in an enclosed space and asphyxiate people, or explode. Experiments with odorants date to at least 1885, and in 1930, the Bureau of Mines conducted experiments with mercaptan, which later became the standardized gas odorizer.

6. LONG-DISTANCE PIPELINES

By the late 1920s, four public utility holding companies dominated the U.S. gas industry and sought to control interstate gas transportation as well. Two of the largest holding companies, Columbia Gas and Standard Oil (New Jersey), distributed more than half of the gas sold in the entire Appalachian region. Henry Doherty's Cities Service dominated the lower midwest. The largest public utility conglomerates included Middle West Utilities, Inc. and Insull Utility Investments, Inc., both controlled by Samuel Insull and headquartered in Chicago. By the late 1920s, Insull's empire included 248 gas, coal, and electric power firms serving 4741 communities in 30 states.

Planning for the first 1000-mile pipeline began in 1926 when Samuel Insull and associates discussed the possibility of building a natural gas pipeline connecting southern gas fields with Chicago area gas utilities. They sponsored engineering studies, considered a pipeline route, and examined potential gas acreage. In April, 1930 they first incorporated as the Continental Construction Corporation; a year later the company changed its name to the Natural Gas Pipeline Company of America (NGPL). NGPL's proposed 24-inch line would extend 980 miles from north Texas to Chicago. Commonly referred to as the "Chicago pipeline," this line would allow Insull to convert Peoples Gas Light & Coke Company's service area from dependence on manufactured coal gas to cleaner, hotter burning, and less expensive natural gas.

The NGPL venture was jointly planned, financed, and controlled by three utility holding companies and three other oil firms. The three holding companies were Samuel Insull's Insull & Sons, Henry Doherty's Cities Service, and Standard Oil of New Jersey. NGPL purchased its gas supply from gas fields controlled by the pipeline's owners. Standard Oil (NJ) agreed to furnish 25% of NGPL's requirements indirectly through the Canadian River Gas Company. Canadian River was a partnership of Cities Service and Prairie Oil & Gas, Standard's partners in

the Colorado Interstate line. The Texoma Natural Gas Company supplied the remaining 75% of NGPL's gas requirements.

Henry L. Doherty & Company contracted to build the NGPL line. Construction began in August, 1930, and the main line was completed 12 months later. A total of 418 million pounds of steel pipe buried 6 feet transported gas at 600 psi. Construction costs for the main line, nine compressor stations, and telephone lines totaled $35 million. Although NGPL's major market was Insull's Chicago area utilities, some gas was also sold to gas distributors in Kansas and other states. The first gas deliveries in Chicago commenced on October 16, 1931, and by January 1, 1932, the line was delivering 55 mmcf/d with an originally designed total capacity of 175 mmcf/d. With access to abundant volumes of natural gas, Chicago became the largest U.S. city to convert its utility distribution system to "mixed gas," and later to straight natural gas. Peoples Gas Light and Coke Company first began producing a mixed gas with a 800-Btu content. Mixed gas, a mixture of lower Btu coal gas and higher Btu natural gas provided a hotter burning flame than did coal gas alone, for both cooking and heating. Peoples Gas Light and Coke Company began charging for gas based on a price per "therm" (1 therm = 100,000 Btu) basis, rather than by volume; natural gas had nearly twice the Btu rating compared to an equal volume of manufactured gas.

Peoples Gas Light and Coke Company organized a massive campaign to merchandise gas house-heating equipment. The company placed full-page and three-quarter-page advertisements in newspapers serving Chicago and in 50 outlying communities; advertisements appeared on billboards, streetcars, and shop windows. In addition, the utility hired 270 company-trained salesmen, 60 heating engineers, and 14 sales directors to promote gas consumption. Within the first 10 weeks of the promotion, Peoples Gas Light and Coke Company installed about 10,000 conversion burners, and the company made 30,000 gas installations during the gas sales promotion. Servicemen adjusted existing residential furnaces to accept the higher Btu mixed gas. In order to convert appliances, gas mains required cleaning to remove oil residue and other impurities from the manufactured gas.

Also during this time, a consortium led by North American Light & Power Company, which owned gas and electric properties throughout the midwest, purchased from Odie R. Seagraves and William L. Moody III (Moody–Seagraves Interests) the beginnings of the pipeline these two men had planned to build from Seagraves's Hugoton gas field properties to Omaha, Nebraska. The North American Light & Power Company joined the Lone Star Gas Company and United Light & Power Company in a partnership to purchase the Moody–Seagraves project and rename it the Northern Natural Gas Company. North American financed the construction of Northern Natural, which was completed 1931. The 1110-mile, 24- and 26-inch line transported gas to various cities along its path to Minneapolis via Omaha.

During the 1930s, a third group of entrepreneurs formed a third pipeline to connect southwestern gas fields with midwestern customers. They incorporated the Panhandle Eastern Pipe Line Company. By 1936, it was transporting gas from the Texas Panhandle through an affiliated firm to Detroit, Michigan.

7. NATURAL GAS IN THE GREAT DEPRESSION

In the late 1920s and early 1930s, the most well-known public utility figure was Samuel Insull, a former personal secretary of Thomas Edison. Insull's public utility empire headquartered in Chicago did not fair well in the economic climate that followed the 1929 Wall Street stock market crash. His gas and electric power empire crumbled, and he fled the country. The collapse of the Insull empire symbolized the end of a long period of unrestrained and rapid growth in the U.S. public utility industry. In the meantime, the Federal Trade Commission (FTC) had launched a massive investigation of the nation's public utilities, and its work culminated in New Deal legislation that imposed federal regulation on the gas and electric industries. The Public Utility Holding Company Act (1935) broke apart the multitiered gas and electric power companies and the Federal Power Act (1935) and the Natural Gas Act (1938), respectively, authorized the Federal Power Commission (FPC) to regulate the interstate transmission and sale of electric power and natural gas.

During the Depression, the gas industry also suffered its worst tragedy in the 20th century. In 1937, at New London, Texas, an undetected natural gas leak at the Consolidated High School resulted in a tremendous explosion that virtually destroyed the Consolidated High School, 15 minutes before the end of the school day. Initial estimates of 500 dead were later revised to 294. Texas Governor Allred appointed a military court of inquiry that determined an

accumulation of odorless gas in the school's basement, possibly ignited by the spark of an electric light switch, created the explosion. This terrible tragedy was marked in irony. On top of the wreckage, a broken blackboard contained these words, apparently written before the explosion: "Oil and natural gas are East Texas' greatest mineral blessings. Without them this school would not be here, and none of us would be here learning our lessons." Although many gas firms already used odorants, the New London explosion resulted in the implementation of new natural gas odorization regulations in Texas.

8. APPALACHIAN GAS AND FEDERAL WAR PLANNING

During World War II, the Pittsburgh, Youngstown, and Wheeling areas contained hundreds of steel mills and metallurgical factories, as well as rubber and chemical plants that required large volumes of natural gas. Natural gas was vital to these factories because it burned at a constant specific temperature, providing high-quality product manufacture. Approximately 660 Appalachian area factories used a substantial amount of natural gas, and wartime energy demands put further pressure on Appalachian gas reserves. Appalachian natural gas production had peaked in 1917 at 552 bcf of natural gas, or about 63% of total U.S. gas production; this percentage declined to approximately 15% by the late 1930s. The decline resulted from diminishing Appalachian gas reserves as well as a proportionate increase in southwestern-produced gas. By 1943, Appalachian production alone was insufficient for meeting regional industrial, commercial, and residential demand. (see Table III).

TABLE III

Natural Gas Production by Region, 1912–1970[a]

Year	Region (%)[b]			Total marketed production (tcf)[c]
	Appalachia	Southwest	Other	
1912	74	22	2	0.56
1920	55	34	11	0.80
1922	46	37	17	0.76
1924	31	45	24	1.14
1926	26	50	24	1.31
1928	21	57	22	1.57
1930	17	61	22	1.94
1935	16	65	19	1.92
1940	15	68	17	2.66
1945	10	73	17	3.91
1950	6	80	14	6.28
1960	3	87	10	12.80
1970	2	90	8	21.90

[a] *Source.* U.S. Bureau of Mines, "Natural Gas Annuals and Minerals Yearbook (Government Printing Office, Washington, D.C.), various years; and Energy Information Administration, "Natural Gas Production and Consumption," Energy Data Reports, DOE/EIA-0131 (Government Printing Office, Washington, D.C., 1978). Also see, David Gilmer, "The History of Natural Gas Pipelines in the Southwest," *Texas Business Review* (May–June, 1981), p. 133.

[b] Appalachia includes Pennsylvania, Ohio, West Virginia, and Kentucky (and New York for 1920 only). Southwest includes Texas, Louisiana, Oklahoma, and Kansas.

[c] tcf, trillion cubic feet.

The intense drain on Appalachian reserves stimulated private entrepreneurial efforts to increase production and build new pipelines. At the same time, some industry executives were already looking forward to a burgeoning gas industry after the war. During one meeting held during 1942, J. French Robinson, a prominent gas utility executive, stated that "in the postwar sunshine of abundant materials for our use, we will be able to realize the potential values of natural gas to all this nation as never before." Patriotic fervor aside, the business of war stimulated both industrial production and entrepreneurial ambition.

To direct the federal government's wartime energy policy, Roosevelt chose Harold I. Ickes, who was then Secretary of the Interior. On May 28, 1941, Ickes assumed his new position as the first Petroleum Coordinator for National Defense; this agency was later renamed the Petroleum Administration for War (PAW). In this role, the new "oil czar" exercised special emergency powers over much of both the oil and gas industries. Despite initial industry fears, Ickes implemented a cooperative relationship with the energy industry during wartime. The PAW created a Natural Gas and Natural Gasoline Division to be responsible for the gas industry. E. Holley Poe, a former executive of the American Gas Association, headed the division. His charge was maintaining natural gas production and deliverability, particularly in the Appalachian region. Poe also attempted to marshal support for joint-industry cooperation while administering the wartime industry. The PAW's authority over natural gas was relatively modest compared to that of the Supply Priorities and Allocation Board (SPAB). The SPAB, which later merged into the War Production Board (WPB), had much broader powers over industry. Regarding

natural gas, the agency dictated specific gas sales allocation orders to gas pipelines.

During late 1941, representatives of the natural gas industry, military, PAW, WPB, and the American Gas Association met several times in different cities to discuss recommendations for restricting some classes of natural gas consumption and maintaining production levels during the war. J. A. Krug, Chief of the WPB Power Branch, was particularly concerned about potential shortages in Appalachia, southern California, and the midcontinent areas. He proposed a special "Limitation Order" for conserving natural gas. The order had two major goals: (1) to increase production and (2) to curtail nonessential consumption. Major General H. K. Rutherford wrote a letter of support and noted the critical situation faced by war industries dependent on natural gas.

In early February, 1942, the WPB issued Order L-31. This action called for voluntary compliance with pooling arrangements "to achieve practicable maximum output in the area or areas in which a shortage exists or is imminent." The order authorized the WPB to integrate natural gas systems, curtail gas sales when necessary, and reallocate existing gas sales. The WPB actively encouraged pipelines to transport gas at 100% load factor, to use gas storage fields whenever possible to free up pipeline capacity for gas transmission, and to develop curtailment schedules. Six months later, the WPB issued Order L-174, which imposed the same restrictions on the manufactured coal gas industry.

The PAW and WPB also addressed the Appalachian gas production problem. First, the PAW set guidelines for a new drilling program, M-68, for developing a nationwide oil and gas drilling program "consistent with the availability of material and equipment." This program limited drilling of gas wells to not more than 1 every 640 acres. Industry leaders objected to M-68, believing that it would stymie efforts to maintain current production levels. In response, the PAW issued new spacing provisions that permitted drilling one well on each 160 acres for specified deep horizons and one to each 40 acres for shallow wells.

The importance of Appalachian natural gas supply to the war effort was reflected in the disproportionate number of wells drilled there. Between 1942 and 1945, approximately 70% of all gas wells drilled in the country were drilled in Appalachia, even though overall production levels did not rise significantly. Wartime demand simply sped up the depletion of Appalachian gas fields. Government drilling and consumption regulations could not reverse this situation.

TABLE IV

Natural Gas Prices and Demand, 1945–1970[a]

Year	Marketed production	
	Trillions of cubic feet	Average wellhead price (cents/million cubic feet)
1945	4	4.9
1950	6	6.5
1955	9	10.4
1960	13	14.0
1965	16	15.6
1970	22	17.1

[a] *Source*. American Gas Association, Gas Facts. (Various years.)

9. GAS IN THE POSTWAR ERA

In the period following World War II, the natural gas industry expanded rapidly. A new round of long-distance pipeline construction made natural gas available throughout the nation. Natural gas fueled factories, electric power-generating facilities, and provided heat for homes and cooking. Demand for gas fuel rose dramatically as it became available. (see Table IV). In this postwar era, entrepreneurs organized several long-distance gas pipeline firms to connect southwestern gas supply with northeastern markets. New pipeline firms organized to sell natural gas to northeastern markets. One group of entrepreneurs purchased the so-called Big Inch and Little Big Inch pipelines from the United States government and converted them for natural gas transportation. The government had financed these lines during World War II to transport oil from the Texas Gulf Coast to the New York refinery area. Under new private ownership, the newly named Texas Eastern Transmission Corporation and affiliated lines delivered natural gas for the first time to Philadelphia, New York, and Boston. Two other new pipelines built either during or immediately after the war, the Tennessee Gas and Transmission Company and Transcontinental Gas Pipe Line Company, also began delivering southwestern-produced natural gas to northeastern customers in major urban areas.

Other pipelines extended from southwestern gas fields to growing urban markets on the West Coast and in the Southeast. California is a large producer of natural gas, but rapid population and infrastructure growth fueled the demand for more of it. El Paso Natural Gas became the first interstate pipeline to deliver natural gas to California, followed by

Transwestern Pipeline Company in the early 1960s. The Northwest Pipeline Company began transporting natural gas produced in the San Juan Basin in Colorado and New Mexico to customers in Seattle after 1956. In 1959, Florida Gas Transmission Company delivered the fuel to Floridians. By the mid-1950s, therefore, the beginnings of a national market for natural gas emerged. During the last half of the 20th century, natural gas consumption in the U.S. ranged from about 20 to 30% of total national energy utilization. However, the era of unrestricted natural gas abundance ended in the late 1960s.

The first overt sign of serious industry trouble emerged when natural gas shortages appeared in 1968–1969. Economists almost uniformly blamed the shortages on gas pricing regulations instituted by the so-called Phillips Decision of 1954. This law had extended the FPC's price-setting authority under the Natural Gas Act to the natural gas producers that sold gas to interstate pipelines for resale. The FPC's consumerist orientation meant, according to many economists, that it kept gas prices artificially low through federal regulation. Gas producers consequently lost their financial incentive to develop new gas supply for the interstate market, and shortage conditions developed.

10. DEREGULATION

The 1973 OPEC oil embargo exacerbated the growing shortage problem as factories switched boiler fuels from petroleum to natural gas. Cold winters further strained the nation's gas industry. The resulting energy crisis compelled consumer groups and politicians to call for changes in the regulatory system that had constricted gas production. In 1978, a new comprehensive federal gas policy dictated by the Natural Gas Policy Act (NGPA) created a new federal agency, the Federal Energy Regulatory Commission, to assume regulatory authority for the interstate gas industry.

The NGPA also included a complex system of natural gas price decontrols that sought to stimulate domestic natural gas production. These measures appeared to work almost too well and contributed to the creation of a nationwide gas supply "bubble" and lower prices. The lower prices wreaked additional havoc on the gas pipeline industry because most interstate lines were then purchasing gas from producers at high prices under long-term contracts. Some pipeline companies had also invested tremendous amounts of money in expensive supplemental gas projects such as coal gasification and liquid natural gas (LNG) importation. The long-term gas purchase contracts and heavy investments in supplemental projects contributed to the poor financial condition of many gas pipeline firms. Large gas purchasers, particularly utilities, also sought to circumvent their high-priced gas contracts with pipelines and purchase natural gas on the emerging spot market.

Once again, government was forced to act in order to bring market balance to the gas industry. Beginning in the mid-1980s, a number of FERC orders, culminating in Order 636 (and amendments), transformed interstate pipelines into virtual common carriers. This industry structural change allowed gas utilities and end-users to contract directly with producers for gas purchases. FERC continued to regulate the gas pipelines' transportation function, but pipelines ceased operating as gas merchants as they had for the previous 100 years. Restructuring of the natural gas industry continued into the early 21st century as once-independent gas pipeline firms merged into larger energy corporations.

Natural gas is a limited resource. Although it is the most clean burning of all fossil fuels, it exists in limited supply. Estimates of natural gas availability vary widely, from hundreds to thousands of years.

TABLE V

Natural Gas Production and Consumption in the United States[a]

Year	Total dry production	Total consumption
1970	21,014,292	21,139,386
1972	21,623,705	22,101,452
1974	20,713,032	21,223,133
1976	19,098,352	19,946,496
1978	19,121,903	19,627,478
1980	19,557,709	19,877,293
1982	17,964,874	18,001,055
1984	17,576,449	17,950,524
1986	16,172,219	16,221,296
1988	17,203,755	18,029,588
1990	17,932,480	18,715,090
1992	17,957,822	19,544,364
1994	18,931,851	20,707,717
1996	18,963,518	21,966,616
1998	19,125,739	21,277,205
2000	19,072,518	22,546,944

[a] *Source*. Energy Information Agency, "Supply and Disposition of Natural Gas in the United States, 1930–2000," Historical Natural Gas Annual, Government Printing Office, Washington, D.C.

In millions of cubic feet.

Such estimates are dependent on technology that must be developed in order to drill for gas in more difficult geographical conditions, and actually finding the gas where it is expected to be located. Methane can also be extracted from coal, peat, and oil shale, and if these sources can be successfully utilized for methane production, the world's methane supply will be extended another 500 or more years.

Since 1970, natural gas production and consumption levels in the United States have remained reasonably stable. During the 1980s, both consumption and production levels dropped about 10% from the 1970 levels, but by the later 1990s, production and consumption were both on the rise. (see Table V). In the absence of aggressive conservation programs, unexpected shortages, or superior alternative energy sources, natural gas consumption will continue to increase.

For the foreseeable future, natural gas will continue to be used primarily for residential and commercial heating, electric power generation, and industrial heat processes. The market for methane as a transportation fuel will undoubtedly grow, but improvements in electric vehicles may well dampen any dramatic increase in demand for engines powered by natural gas. The environmental characteristics of natural gas, however, should retain this fuel's position at the forefront of desirability of all fossil fuels, while supplies last.

SEE ALSO THE FOLLOWING ARTICLES

Coal Industry, History of • Electricity Use, History of • Manufactured Gas, History of • Natural Gas Industry, Energy Policy in • Natural Gas Processing and Products • Natural Gas Resources, Global Distribution of • Natural Gas Resources, Unconventional • Natural Gas Transportation and Storage • Nuclear Power, History of • Occupational Health Risks in Crude Oil and Natural Gas Extraction • Oil Industry, History of

Further Reading

Bragdon, E. D. (1962). "The Federal Power Commission and the Regulation of Natural Gas: A Study in Administrative and Judicial History." Ph.D. Dissertation. Indiana University.

Castaneda, C. J. (1999). "Invisible Fuel: Manufactured and Natural Gas in America, 1800–2000." Twayne Publishers, New York.

Castaneda, C. J., and Smith, C. M. (1996). "Gas Pipelines and the Emergence of America's Regulatory State: A History of Panhandle Eastern Corporation, 1928–1993." Cambridge University Press, New York.

DeVane, D. A. (1945). Highlights of the legislative history of the Federal Power Act of 1935 and the Natural Gas Act of 1938. *George Washington Law Rev.* XIV (Dec. 1945).

De Vany, A. S., and Walls, W. D. (1995). "The Emerging New Order in Natural Gas: Markets vs. Regulation." Quorum Books, Westport, Connecticut.

Frey, J. W., and Ide, H. C. (1946). "A History of the Petroleum Administration for War, 1941–1945." Government Printing Office, Washington, D.C.

Herbert, J. H. (1992). "Clean Cheap Heat: The Development of Residential Markets for Natural Gas in the United States." Praeger, New York.

MacAvoy, P. W. (2001). "The Natural Gas Market: Sixty Years of Regulation and Deregulation." Yale University Press, New Haven.

Peebles, M. W. H. (1980). "Evolution of the Gas Industry." New York University Press, New York.

Rose, M. H. (1995). "Cities of Light and Heat: Domesticating Gas and Electricity in Urban America." University of Pennsylvania Press, University Park.

Sanders, E. (1981). "The Regulation of Natural Gas: policy and politics, 1938–1978." Temple University Press, Philadelphia.

Stotz, L., and Jamison, A. (1938). "History of the Gas Industry." Stettiner Brothers, New York.

Tarr, J. A. (1998). Transforming an energy system: The evolution of the manufactured gas industry and the transition to natural gas in United States (1807–1954). *In* "The Governance of Large Technical Systems" (O. Coutard, Ed.), pp. 19–37. Routledge, London.

Tussing, A. R., and Barlow, C. C. (1984). "The Natural Gas Industry: Evolution, Structure, and Economics." Ballinger Publ., Cambridge.

Natural Gas Industry, Energy Policy in

MICHELLE MICHOT FOSS
Institute for Energy, Law and Enterprise, University of Houston
Houston, Texas, United States

1. Introduction
2. Examples of Natural Gas Policy

Glossary

associated/dissolved natural gas A type of natural gas that occurs in crude oil reservoirs either as free gas (associated) or as gas in solution with crude oil (dissolved gas).

dry natural gas The natural gas that remains (1) after the liquefiable hydrocarbon portion has been removed from the gas stream (i.e., gas after lease, field, and/or plant separation) or (2) after any volumes of nonhydrocarbon gases have been removed, when they occur in sufficient quantity to render the gas unmarketable. Dry natural gas is also known as consumer-grade natural gas. The parameters for measurement are cubic feet at 60°F and 14.73 pounds per square inch absolute.

methane A colorless, flammable, odorless hydrocarbon gas (CH_4); the major component of natural gas. It is also an important source of hydrogen in various industrial processes. Methane is a greenhouse gas.

natural gas A gaseous mixture of hydrocarbon compounds, the primary one being methane. The U.S. Energy Information Administration measures wet natural gas and its two sources of production (associated/dissolved natural gas and nonassociated natural gas) and dry natural gas (which is produced from wet natural gas).

nonassociated natural gas A form of natural gas that is not in contact with significant quantities of crude oil in the reservoir.

wet natural gas A mixture of hydrocarbon compounds and small quantities of various nonhydrocarbons existing in the gaseous phase or in solution with crude oil in porous rock formations at reservoir conditions. The principal hydrocarbons normally contained in the mixture are methane, ethane, propane, butane, and pentane. Typical nonhydrocarbon gases that may be present in reservoir natural gas are water vapor, carbon dioxide, hydrogen sulfide, nitrogen, and trace amounts of helium. Under reservoir conditions, natural gas and its associated liquefiable portions occur either in a single gaseous phase in the reservoir or in solution with crude oil and are not distinguishable at the time as separate substances. The Securities and Exchange Commission and the Financial Accounting Standards Board refer to this product as natural gas.

Natural gas has come a long way as a good with intrinsic value. This resource was once considered a mere by-product of oil and thus not worth the significant capital investment required to find, gather, treat, transport, and distribute it. The relative abundance, wide dispersion, and cleanliness of natural gas have propelled it to the forefront of the fossil fuels, so that natural gas today is poised to become a global commodity, a bridge fuel to the next energy future, and a source of molecular building blocks to new materials. The key policy challenges lie in differentiating the various markets associated with natural gas molecules and fashioning competitive supply, demand, and pricing mechanisms; designing appropriate policy approaches for components of the natural gas value chain that bear public service attributes and thus affect the public interest; and mobilizing capital investment while balancing efficiency and equity concerns.

1. INTRODUCTION

The term "natural" is used to distinguish gaseous and associated liquid hydrocarbons that are produced in the earth from those that are "manufactured," typically from coals (and often referred to as "town gas"). Methane (one carbon and four hydrogen atoms) is the most abundant molecular component of a natural gas stream, typically comprising as much

 © 2004 Elsevier Inc. All rights reserved.

as 85%, depending on whether the natural gas occurs with crude oil or separately. Other molecules in the natural gas stream, in varying proportions, include ethane, propane, butane, carbon dioxide, oxygen, and nitrogen, and impurities such as hydrogen sulfide and rare gases that must be removed before the natural gas can be used. Liquefied petroleum gas (LPG), a propane/butane mixture, is stripped out of the natural gas stream during field production; LPG can be shipped by pipelines or separately by truck or tanker for uses such as petrochemical manufacturing, charcoal starter fuel in backyard grills, or vehicle transport fuel. In many countries, particularly those with short heating seasons, LPG comprises the principal domestic fuel source for water and space heating and cooking. Natural gas liquids (NGLs)—ethane and larger molecules—also are stripped out as feedstocks for petrochemicals (methane can also be used instead of the heavier molecules for petrochemical applications, but much larger volumes of methane are required). The NGL molecules can also be transported by pipeline, truck, or tanker. Other than pipeline delivery, methane can be liquefied and thus transported economically over large distances in liquefied natural gas (LNG) ships or by truck; LNG cargoes may contain molecules other than methane if there is no processing at the point of liquefaction. Methane can also be compressed and transported shorter distances by truck as compressed natural gas (CNG); eventually, depending on emergence of viable technologies and cost structures, it may be possible to transport CNG on tankers. Compression and liquefaction are required for methane to be used as a vehicle fuel, whereas LPG can be used directly.

Worldwide, the natural gas industry has grown rapidly in recent years. Nations with rich natural gas resources have aggressively added new petrochemical capacity. These huge investments tend to be quite lumpy and the products are subject to intense global competition (with commensurate impacts on the feedstock molecules), leading to the well-known cyclicality in these businesses. Likewise, LNG investments are also sizable, lumpy, and subject to global forces, and are also fast growing. For nations that do not have large enough domestic demand relative to the size of their resource base, or that have not developed petrochemicals capacity for conversion of natural gas to other products, LNG is an important means of deriving value for their natural resource endowments through international trade. LPG as a domestic energy source has also grown, proving to be a relatively cheap (in terms of local infrastructure) and clean replacement for biomass. In particular, LPG has replaced wood (and is therefore preferable in places where deforestation has been rampant) and animal dung. LPG also represents an improvement in both cleanliness and safety over kerosene, also used as domestic fuel. Where LPG is in wide use, typical policy challenges include pricing and transparency in market organization. Because LPG is most often found as a domestic fuel in poorer countries, the tendency has been toward government control of final prices charged to customers and heavy subsidization of these prices. These policy approaches are usually neither fair nor effective; they are also costly to governments. In many countries, lack of transparent market organization for LPG contributes to widespread theft, increasing distortions and injuring the consumers least able to make alternative choices.

Above all, it is the growth in demand for piped methane as a clean convenient option for consumer energy markets that has garnered the greatest interest. Thus, from hereon "natural gas" refers to pipeline delivery of methane and the myriad operational and commercial aspects involved. Generally, worldwide, natural gas/methane infrastructure systems consist of upstream, midstream, and downstream operational elements. Upstream operations involve two components:

1. Exploration. This consists of activities that lead to discovery of new natural gas resources. Exploration risk is one of the strongest forms of risk.
2. Production. Extraction of discovered supplies from hydrocarbon fields either with crude oil (as associated natural gas) or separately (nonassociated natural gas). If natural gas is associated with crude oil, it must be separated.

Midstream operations are more extensive:

1. Gathering. Collection of natural gas production from multiple wells connected by small-diameter, low-pressure pipeline systems and delivery to a processing plant or long-distance pipeline.
2. Processing and treatment (if necessary). Processing is the separation of heavier molecules and unwanted substances such as water from a methane gas stream. If the gas stream contains impurities such as hydrogen sulfide, then treatment is required.
3. Storage. Containment of supplies, usually in depleted underground reservoirs or caverns such as those associated with salt domes. Storage can be located either near production or near demand.
4. Transportation. Delivery of gas from producing basins to local distribution networks and

high-volume users via large-diameter, high-volume pipelines. In countries that are federal republics, pipeline systems may be distinguished by whether they cross state or provincial boundaries (for example, in the United States, interstate pipelines as opposed to intrastate systems that operate within the state jurisdiction). Countries vary greatly with respect to allowable pipeline specifications for heat content, as measured by the British thermal unit (Btu) in the United States, Canada, and elsewhere, and as related to the presence of molecules other than methane in the piped stream. For example, pipelines transport and distribute methane or "dry" gas in Canada and the United States, which means that heavier molecules are removed before the natural gas stream enters the pipeline system. Exceptions do exist, such as the Alliance Pipeline, which transports "wet" gas from British Columbia to Chicago, where molecules other than methane are stripped out in processing for use in other markets. Pipeline standards generally are set for safety reasons.

5. Liquefaction, shipping, and regasification. Known collectively as the LNG value chain, this entails conversion of gas to a liquid form via refrigeration, resulting in a cryogenic fluid (temperature −256°F) for transportation from a producing country or region to a consuming country or region via ship. LNG is stored until it is returned to the gas phase (regasification, using vaporization) for pipeline transportation within the consuming region. In the United States, LNG is also used as a storage form of natural gas produced from domestic fields, until the gas is needed, primarily for peak use. Both storage and transport of LNG are done at nearly atmospheric pressure.

Finally, the infrastructure system consists of downstream operations:

1. Distribution. Retail sales and final delivery of gas via small-diameter, low-pressure local gas networks operated by local distribution companies (often termed gas utilities).
2. End use and conversion. Direct use or conversion for use in other forms (petrochemicals, electric power, or vehicle fuels).

The following commercial elements serve to bind the operating segments of the natural gas infrastructure system, linking suppliers, transporters, and distributors with their customers:

- Aggregation. Consolidation of supply obligations, purchase obligations, or both as a means of contractually (as opposed to physically) balancing supply and demand.
- Marketing. Purchase of gas supplies from multiple fields and resale to wholesale and retail markets. Retail marketing constitutes sales to final end users (typically residential, commercial, industrial, electric power, and public sectors).
- Capacity brokering. Trading of unused space on pipelines and in storage facilities.
- Information services. Creation, collection, processing, management, and distribution of data related to all the other industry functions listed here.
- Financing. Provision of capital funding for facility construction, market development, and operation start-up.
- Risk management, Balancing of supply, demand, and price risks.

Altogether, the upstream, midstream, downstream, and commercial elements constitute the natural gas value chain. The various segments are highly interdependent but, in an open, competitive market, they also can be highly competitive. The policy challenges associated with increased worldwide use are numerous. Frameworks for efficient discovery and optimal production are the first hurdle. Efficient and equitable mechanisms, often at odds, for pipeline transportation and local distribution are the second major hurdle. Methane is of little use in consumer energy markets without pipeline infrastructure. These large systems tend to be characterized by strong technical economies of scale and high barriers to entry. Particular problems also emerge with respect to the public interest/public service component of these facilities, mainly with respect to reliability and pricing of systems that are usually operated in monopoly, duopoly, or limited competitor regimes. In all cases, a specific policy quandary is how best to achieve prices for natural gas transportation and distribution through tariff designs that yield something close to what competitive markets might be able to achieve, with contestability (potential competition) usually providing a basis for market-based pricing in larger markets.

A third challenge is development of transparent markets for natural gas supply and consumption. If pipelines are an essential feature, a central question is whether molecules have intrinsic value or whether the combination of pipeline and molecule together must constitute the service. Increasingly, the trend has been to separate infrastructure and product (often termed "unbundling") and to search for ways

of providing competitive access to pipeline systems for multiple suppliers and users of natural gas (often termed "third-party access" or "open access"). In these cases, pipelines become like toll roads, priced through tariff design, whereas molecules become priced in discrete, competitive markets. When pipelines become subject to third-party access regimes, the commercial activities described earlier and associated with pipeline operations, i.e., the linking of suppliers to buyers, can be separated into competitive business activities. Once subjected to competitive markets, methane molecules become commoditized. With all market participants as price takers, new sources of risk and new policy challenges arise. In recent years, the rapid commoditization of methane in the United States and Canada triggered growth in marketing and trading (both of the physical product as well as financial derivatives) as separate businesses. With methane a commodity, and pipeline capacity close to being commoditized, balancing supply and demand across a market becomes both more efficient and more fragmented, as new combinations of activities and relationships across the value chain become established. The fuel's growing importance in the international economy, as natural gas becomes globalized via LNG shipments and disparate national and regional markets become linked, has meant new incentives for technical improvements in supply and demand management and balancing. The principal segments of a natural gas infrastructure system—exploration, production, transportation and distribution—share substantial capital requirements and comparable, albeit different, risks. The long lead times required for development of each sector's assets present both industry leaders and policymakers with the problem of adequately anticipating changes in supply and demand. These projections must be accurate and timely to attract the necessary long-term investment consistently and to minimize market disruptions and distortions. Inadequate projections create conflicts to the extent that they result in supply–demand imbalances, which neither industry nor government has the flexibility to correct in a timely manner.

Both the international trade linkages and the evolution of market-based policies for natural gas mean timely and accurate data and information on supply, demand, and prices, a fourth requirement. A fifth and increasingly complicated challenge is dealing with integration, with respect to industry organization and international trade. Industry organization can encompass both vertical (meaning up and down the value chain) and horizontal (meaning over some geographic or market extent) integration. Paradoxically, the forces for integration within a natural gas industry often occur in spite of policy objectives that seek to instill deintegration and competition as part of transitions to competitive markets. Integration of physical infrastructure across international boundaries has grown rapidly with increased demand for piped methane. As transportation and information technologies have improved, so have the opportunities for system linkages—first within a country, then among geographically contiguous nation states, and increasingly across the globe. With improved physical and commercial linkages comes an ever greater need for more complex, sophisticated, and coordinated policy solutions, posing new dilemmas in international trade.

Natural gas policy across nations is most easily differentiated by how the value chain is organized and operates with respect to the balance between markets and government—that is, sovereign ownership of or control over the critical segments of the value chain. The high degree of interdependency across the value chain segments, the propensity toward integration, the fact that, in most cases, large deposits of natural gas are associated with oil (a strategic commodity for many producing nations), the energy security aspects of natural gas supply and delivery, and the public interest/public service concepts embedded in pipeline infrastructure have all resulted in a strong pattern of government ownership, control, or, in the least, intervention in natural gas enterprises and markets. In most countries, the natural gas value chain has been developed through integrated, sovereign owned, or heavily controlled enterprises. The rare exceptions are the United States, Canada, and Australia, all of which have allowed the natural gas system infrastructure to emerge through the activities of private, investor-owned companies. For the United States and Canada, this experience extends roughly 100 years; for Australia, this has been the practice since the 1960s. The United States is even more unique in that private ownership of much of the resource base is allowed; indeed, it is a powerful tradition. Federal and/or state ownership of the natural resource is limited to certain onshore lands and offshore, and even in these cases, development of sovereign-owned and controlled resources has always been through competitive acquisition of leases in organized auctions and private investment in exploration and production (a tradition also maintained in Australia,

where states and territories play a comparable role, and in Canada, which has limited private ownership of natural gas resources in southern Alberta province; the vast majority of the resource base, both onshore and offshore Canada, is controlled by the provincial crown governments). Of great interest is the transition in many countries away from sovereign ownership and/or control of integrated natural gas enterprises, as well as the push for ever more competitive markets in many locations, including Canada and the United States.

To a large extent, these transitions have been driven by specific needs for increased efficiency and to introduce innovations, to solve fundamental problems in pricing and service, and to attract investment into the natural gas value chain. As with any industry, natural gas market development requires sufficient supply availability and enough demand to justify the infrastructure systems to connect buyers and sellers. To attract this investment, governments have experimented with policies designed to stabilize the investment environment by optimizing participant choice at predictable prices that reflect, or at least attempt to mimic, actual supply and demand conditions. During the past 15 years or so, the progression toward competitive markets has meant movement toward market determination of investment, and operation of assets subject to real-time supply-and-demand interactions. Under these conditions, actionable information must be timely, accurate, and transparent. For such information to be truly actionable, the decision maker also must have timely access to whatever system capacity the information prompts that decision maker (supplier, customer, or intermediary) to demand. Finally, competitive markets must comprise systems wherein this information and capacity cannot be dominated or manipulated by a few anticompetitive participants. These conditions are difficult to create, implement, and maintain, and imply new and changing roles for market participants and government overseers.

2. EXAMPLES OF NATURAL GAS POLICY

2.1 United States

2.1.1 Overview

Natural gas exploration and production on private lands, including environmental and safety controls, are regulated at the individual state level by conservation commissions. Exploration and production on state lands are controlled by separated state agencies charged with management of those lands. Exploration and production on federal lands, onshore or offshore, are managed by federal agencies. Natural gas exploration and production, gathering, and processing are all viewed to be "workably competitive" industries and are not regulated for prices. Tariffs for transportation within state boundaries on intrastate pipelines are regulated by state public utility commissions (PUCs). The PUCs also license new intrastate pipelines. Tariffs for transportation across state boundaries in interstate pipelines are regulated by the Federal Energy Regulatory Commission (FERC). The FERC also issues licenses (certificates of public need) for new interstate pipelines. Tariffs for natural gas distribution to final customers are regulated by PUCs. The most competitive gas service in the United States is for industrial customers. The least competitive service is to residential customers.

The FERC is governed by five appointed commissioners and operates as an independent authority. Enabling legislation for the FERC dates back to the 1930s (it was created as the Federal Power Commission and charged principally with development of water and hydroelectric facilities). The FERC's authority to regulate interstate natural gas commerce is embodied in the 1938 Natural Gas Act. The individual state PUCs were established at various times, generally between the late 1800s through the 1970s. Each state has a separate enabling legislation for the formation of its PUC. Notable exceptions are the state of Nebraska, which does not regulate natural gas, and Texas, where natural gas is regulated by the Texas Railroad Commission rather than the PUC. The PUCs also vary with regard to numbers of commissioners, whether commissioners are elected or appointed, and sizes of staffs and budgets. The FERC and state PUCs are funded through fees charged to regulated industries.

The style of regulation in the United States traditionally has been "cost of service" or "rate of return," which involves a determination of revenue requirements and rate structures based on costs provided by the regulated firms. A regulated company may be a local distribution company (gas utility), an intrastate pipeline, or an interstate pipeline. The regulated company's revenue requirements are the total funds that the company may collect from ratepayers (customers). Revenue requirements are calculated by multiplying the company's rate base by an allowed rate of return (ROR) and adding this

product to the company's operating costs (OCs), as shown in the following formula:

$$\text{Revenue requirement} = (\text{rate base} \times \text{ROR}) + \text{OC} - (\text{taxes} + \text{depreciation})$$

The rate base is the total value of the company's capital investments, which may include construction work in progress. The allowed rate of return constitutes a profit sufficient to pay interest on accumulated debt and to provide a "fair" return to investors. A fair return is determined through a comparable earnings test (whereby a company's earnings are measured against those of a firm facing comparable risks), a discounted cash flow approach (whereby a company's capital costs are estimated by analyzing conditions in the financial market), or some other method. Operating costs include expenses for purchased gas, labor, management, maintenance, and advertising. The costs of taxes and depreciation are also part of a company's revenue requirements. The regulatory process can be generally described as follows:

- A regulatory commission (a PUC or the FERC) first seeks to determine how much of an applicant's capital stock should be included in the rate base, then attempts to determine which elements of test-year costs and revenues should be allowed for regulatory purposes and whether to allow specific changes since the test year. The final step is to determine what the fair rate of return is for the company.
- States and the FERC have legal rules for deciding what should be included in the rate base, although the same is not necessarily true for the method of calculating allowed rate of return.
- States may vary from each other and from the FERC according to the particular sets of rules that are used (for example, to calculate rate base) and the impact of these rules on rate case decisions. However, over the course of the long history of natural gas regulation in the United States, the states and the FERC have generally shared practices fairly quickly.
- All regulators are constrained in their abilities to calculate cost of capital. This is due in part to general disagreement within the industry of how market cost of capital should be computed, and in part because commissions are not well equipped to deal with the complexities surrounding these issues. As a result, a critical component of a rate case proceeding is a commission's reliance on historical information, or precedent, as well as the testimony of interveners, parties with specific interests in the outcome of rate cases (principally large customers and consumer advocates representing small business and residential users; competing regulated firms may also intervene).
- All U.S. regulatory commissions hear rate cases, issue blanket rulings that set broad policy parameters, and act as judges and adjudicators on disputes.

With the implementation of unbundling (separation of pipeline transportation from natural gas sales, with nondiscriminatory open access or third-party access) in 1992, the FERC and many of the states now encourage market-based rates for transportation service. With respect to distribution, many states are experimenting with "incentive-based" regulation designed to encourage more efficient operation and capital cost decisions than has historically been achieved with cost-of-service regulation.

2.1.2 History

When comparing the United States to other countries, an important difference is that the U.S. natural gas system has always been characterized by the participation of private companies, with regulation as a substitute for competition to moderate private monopoly market power. For the most part, regulation has been directed toward pipeline transportation and local distribution. However, during periods of U.S. history, cost-of-service-style regulation was also applied to natural gas production at the wellhead with disastrous results. Table I illustrates the complex progression of regulatory policy eras in the United States.

In the early days of the U.S. natural gas industry, the construction and operation of natural gas distribution systems tended to be concentrated around local deposits of natural gas. Cities and towns that were near the early discoveries of natural gas in the late 1800s were often the centers of intense competitive activity as companies struggled to build competing systems. Because all of the early natural gas companies were private and because the intense competition reduced the returns to shareholders, state-level regulation of local distribution companies (LDCs) through public utility commissions evolved. The strategy was to stabilize investment returns to shareholders while attempting to mimic most of the benefits of competition to customers through regulation (competition by substitution, as it is often called). The form of regulation typically used was "cost of service," in which regulators granted a rate of return that was deemed to be reasonable to the LDCs. In exchange for a limited return, the LDCs enjoyed a monopoly franchise for service in a city or

TABLE I

Natural Gas Industry Regulation/Deregulation in the United States

Policy history	Impact
Competitive local distribution company industry	Emergence of natural gas industry as local distribution companies were established to provide town gas and later natural gas service
State public utility commissions, 1885–1927	Formation of Massachusetts Gas Commission in 1885 through regulation of intrastate pipelines (in all 48 states by 1927); followed the Supreme Court ruling (*Munn v. Illinois,* 1877) that established the basis for regulating monopolies (grain elevators and warehouses) as public utilities
Development of interstate transportation, 1930s	Technological advances (mechanized trenching and arc welding) allowed construction of long-distance pipelines to transport natural gas from large producing fields in the southwestern United States to key market in the Northeast and upper Midwest
Federal regulation of interstate transportation (Public Utility Holding Company Act and Federal Power Act of 1935; Natural Gas Act of 1938)	Interstate Commerce Act of 1887 provided the basis for federal intervention. A U.S. Supreme Court decision in 1934 (*Nebbia v. New York*, dealing with milk prices) expanded the basis for public utility regulation. Disputes centered on pricing natural gas in cross-state sales activities and market power of interstate public utility holding companies. The Federal Power Act established and authorized the Federal Power Commission; the natural gas industry was thus comprehensively regulated, from the burner tip, to intrastate transmission, to interstate transmission, by state and federal jurisdictions
Federal regulation of wellhead prices (U.S. Supreme Court *Phillips* Decision, 1954)	Dispute regarding pricing of natural gas produced in Oklahoma for delivery in Michigan led to cost-of-service regulation at the wellhead, with the FPC as agency with authority
Beginning of wellhead price decontrol (Natural Gas Policy Act, 1978; Public Utility Regulatory Policy Act of 1978; Powerplant and Industrial Fuel Use Act of 1978)	The FPC's inability to deal with the scope of wellhead regulation and provide sufficient adjustment to increase price ceilings and encourage production, as well as disparity in pricing natural gas sold in interstate markets, with gas sold in unregulated intrastate markets, and resulting curtailments, all led to decontrol. The Natural Gas Policy Act of 1978 extended wellhead price ceilings to the intrastate market, introducing the process of deregulation by loosening certification requirements to facilitate gas flows
First stage of open access for interstate pipelines (FERC Orders 436 and 500, 1985)	"Phased decontrol," with surplus conditions created by the Natural Gas Policy Act; the need for flexible pricing and transportation led to "special marketing programs" that released gas from long-term contracts into price-discounted supply pools. FERC Order 436 in 1985 created the open-access era and Order 500 provided some resolution for take-or-pay liabilities
Final restructuring rule for interstate pipelines (FERC Order 636, 1992)	FERC Order 636 continues separation of merchant and transportation functions of interstate pipelines
Regulation of interstate transportation and related services (FERC Order 627, 1999)	In Order 637, the FERC moved to improve competitive markets, mainly targeting capacity and operations for captive customers and to deal with pipeline affiliate issues

town but also had the obligation to serve all customers within that franchise.

The discovery of huge natural gas deposits in Texas and Oklahoma fundamentally changed the U.S. natural gas industries. Companies began to build long-distance pipelines to carry natural gas from the southwestern United States to the Northeast and Midwest, where gas was needed for winter heating. Almost immediately, disputes arose among individual states with regard regulatory jurisdiction over interstate sales of natural gas. By 1938, the U.S. government was prepared to step into the conflict. Passage of the Natural Gas Act (NGA) that year gave the Federal Power Commission (FPC) regulatory

authority over interstate natural gas commerce. This action was consistent with the philosophy of the times. Following the Great Depression, there was considerable mistrust of large businesses and greater faith in the ability of government to intervene and solve problems.

The NGA treated the interstate pipelines as natural monopolies. (It should be noted, however, that in debating the NGA, the U.S. Congress deliberated on contract carriage as an alternative approach.) The economics of early pipeline construction and operation and conditions in the early natural gas markets were considered to be such that it was unlikely for many companies to build competing facilities. As a result, the pipelines acted as "merchants," contracting with natural gas producers for supply and also with local distribution companies for deliveries. Disputes related to the price of natural gas in the interstate market did not end, however, and the federal government intervened again, this time through the Supreme Court. In the landmark *Phillips* decision in 1954, the Court concluded that the FPC should also have regulatory authority over the price of natural gas at the wellhead. By this time, thousands of natural gas wells had been drilled in Texas, Oklahoma, Louisiana, and other states, including the beginnings of the U.S. offshore industry. The task faced by the FPC was daunting, and the ability of federal regulators to perform efficiently was limited. Distortions began to show up immediately, most importantly in the difference between prices for natural gas in the regulated interstate market and prices in the unregulated intrastate market (meaning gas produced and sold within the boundaries of individual states). Demand for natural gas had grown and prices were rising in the intrastate market. As a consequence, producers shifted their strategies so as to sell more gas in that market. By the time of the oil embargoes and supply shocks in the early 1970s, insufficient amounts of natural gas were committed to the interstate market. During the severe winter in 1976, shortages and curtailments of natural gas supplies occurred all over the eastern United States. To make matters worse, because the interstate pipelines controlled all transactions, there was no way for natural gas producers to engage in sales directly with customers.

Broad dissatisfaction with how the natural gas sector was managed led to an unwinding of federal regulatory control. By the 1970s, public opinion regarding government management of economic activity, including energy, had begun to erode. The preference for market-based solutions was increasing. Already the United States was engaged in major transformations to reduce government intervention in other sectors, such as airline transportation, telecommunications, and banking. The first step was to remove regulatory control of natural gas at the wellhead, with the Natural Gas Policy Act (NGPA) of 1978, which also transformed the FPC into the Federal Energy Regulatory Commission. The strategy chosen by the U.S. government was flawed, with a tremendously complicated schedule for decontrol of natural gas from different formations, by year of discovery (vintage), and so on. The U.S. Congress created more than 200 different categories of natural gas. Overall, natural gas prices rose rapidly in response to demand. After a period of time, higher natural gas prices caused demand to fall as customers, especially large industrial users, shifted to cheaper fuels. Demand adjustments were further complicated by two additional laws that had been passed to deal with 1970s energy crises, the Public Utility Regulatory Policy Act (PURPA) and the Powerplant and Industrial Fuel Use Act (PIFUA), which, together with the NGPA, encompassed the National Energy Act. PURPA encouraged experimentation with "cogeneration" of electricity using natural gas at industrial facilities, which sold their electric power to electric utilities at a price that (presumably) reflected the costs utilities would avoid by not building new electricity generation capacity themselves. PIFUA, however, prohibited natural gas use in most industrial applications and by the electric utilities. The resulting confusion and fluctuations in prices created havoc on both sides of the interstate pipeline merchant contracts. Both pipelines and producers were left holding contracts with take-or-pay obligations that led to severe financial strain for many companies. Because interstate pipelines still acted as merchants, bottlenecks existed all over the natural gas system, preventing efficient transactions from taking place.

In 1983, the FERC began to put into place the policies that have led to the restructured natural gas market that we see in the United States today. Through a series of actions, the FERC began to dismantle the interstate pipeline merchant function. Pipelines came to be treated as common carriers, conduits through which any seller or buyer could ship gas. Natural gas came to be treated as a commodity, whereas previously oil companies had treated natural gas as a by-product with no intrinsic value. Pipeline construction technology had changed dramatically over the years, and many parts of the

United States were served by more than one interstate pipeline, reducing the natural monopoly advantage. The advent of sophisticated computer information systems allowed real-time trading of natural gas, and financial instruments (futures contracts and other mechanisms) enabled suppliers and customers to manage commodity price risk. Although conditions today are vastly different—natural gas in recent years has enjoyed a growing share of the U.S. energy mix, albeit with great price elasticity for certain kinds of demand—there is no doubt that policies in the past constrained market growth of the industry. Indeed, the position that Canada enjoys as a major exporter of natural gas to the United States (approximately 15% of U.S. consumption) is a direct outcome of the 1976 shortages.

2.1.3 Issues

Several issues remain following restructuring to restore and enhance competition in the U.S. gas system.

2.1.3.1 Market Disruptions The FERC's actions to implement open access on U.S. interstate pipelines created a "wholesale market" for natural gas, with competitive pricing, trading, and marketing activities; price risk management (the New York Mercantile Exchange established a natural gas futures contract in 1993); market mechanisms to facilitate trading of unused pipeline capacity; and a national standards board (the Gas Industry Standards Board) to facilitate commercial activity. Beginning in 2000, surging prices for natural gas and electric generation constraints as a result of extended drought in the Pacific Northwest resulted in collapse of the electric power market in California. Disparities in natural gas prices between the California internal market and other U.S. locations and related improprieties in natural gas trading, along with the bankruptcy of Enron Corporation, led to a general collapse in the "energy merchant" segment, including credit downgrades, additional bankruptcies, and severe losses in market capitalization. Energy merchant businesses deal in unregulated wholesale market activities, including construction and operation of competitive, unregulated infrastructure assets. Many energy merchants were affiliated with regulated natural gas interstate pipelines and utilities. Continued conflict around issues that emerged during these events heavily impacted natural gas markets in the United States and Canada. A subsequent natural gas spike in 2003, with ancillary concerns regarding natural gas field production trends and disputes regarding how natural gas price information is compiled and communicated in the marketplace, continue to retard further policy, regulatory, and commercial activity.

2.1.3.2 Lack of Competitive Service Lack of competitive service to small residential and commercial customers occurs at the end of distribution systems. Unbundling and open access to facilitate deliveries of competitive supply for smaller customers have not materialized as expected following FERC's restructuring rule. These initiatives generally are in the domain of state jurisdictions, but federal/state coordination, always an issue, is required in some instances.

2.1.3.3 Barriers to Entry of New Pipelines The certification process, already deemed to be too onerous given competitive market conditions, has been further complicated by the market disruptions of the 2000s and lack of financial capital available from distressed energy merchant businesses and parent pipeline companies.

2.1.3.4 Uncertainty with Regard to Electric Power Restructuring Following early success with natural gas, some states (notably California in 1994) and the FERC (through Orders 888 and 889), and with encouragement through the Energy Policy Act (EPAct) of 1992, proceeded to experiment with similar unbundling and open-access approaches for electricity grids. The 1992 EPAct supported creation of a bulk, wholesale market for electric power. The electric power market collapse in California and natural gas market disruptions have stymied further initiatives by the FERC for a segment of the energy sector that is considered to be crucial to growth of the natural gas industry.

2.2 Canada

The Canadian natural gas system parallels that of the United States, with an important exception. All natural gas resources in Canada are controlled by the provincial crown governments. Exploration and production activities are carried out by private, competing firms under the rules and regulations established by provincial energy ministries. This contrasts with the United States, where roughly two-thirds of natural gas resources and production are in the private domain (held either by companies or individuals).

Like the United States, Canada's transportation and distribution systems are owned and operated by private (investor-owned) companies regulated to

control for any monopoly power. Regulation of long-distance, interprovincial pipelines is carried out by the National Energy Board (NEB), which receives its authorization from the federal National Energy Board Act. The NEB, like the FERC, licenses new pipelines, sets tariffs for transportation, adjudicates disputes, and sets broad policy parameters with blanket rulings. Intraprovincial pipelines and local distribution systems are regulated by provincial energy utilities boards (EUBs). Each province has enabling legislation for its EUB. As with the state PUCs and the FERC in the United States, the EUBs and the NEB use similar methods for regulating their client industries, and are funded by these industries. The NEB and EUBs have also, traditionally, used cost-of-service ratemaking like the U.S. commissions do. Canada is a significant exporter of gas to the United States, supplying about 17% of U.S. (the lower 48 states) current demand. The NEB licenses and regulates all natural gas export activity from Canada.

Canada began to restructure its natural gas system ahead of the United States in the early 1970s with the Western Accord, which eliminated control of the natural gas supply by Canada's monopoly interprovincial pipeline, TransCanada. The Agreement on Natural Gas Markets and Prices in the 1980s was a statement in principle of support for a market-based natural gas system. The Open Access Order in 1986 unbundled the Canadian system and allowed contract carriage on Canada's pipelines. Since these steps were taken, the NEB has consistently encouraged market-based rates for transportation.

At the provincial level, EUBs followed the NEB with open access and market-based tariffs. Nearly every local distribution system in Canada offers some form of competitive supply to its core customers (residential and small commercial), with LDCs in Ontario moving toward full open-access systems for core customer service.

2.3 Britain

Like the United States and Canada, Britain's national gas grid evolved to transport "town gas" manufactured from coal by the state-owned British Gas (BG). However, unlike its North American counterparts, Britain's grid was not converted to natural gas until the 1960s. As the first gas came ashore from West Sole field in the North Sea in 1967, British Gas began the process of transforming its business. By comparison, discovery of natural gas in the United States at the turn of the 20th century, first in the Appalachians and later in the Southwest (Texas and Oklahoma), had already launched the U.S. natural gas industry. Britain has moved much more quickly than the United States to allow competition to residential and small commercial customers.

Restructuring of the British natural gas sector involved a series of steps:

1. Passage of the Oil and Gas (Enterprises) Act of 1982, which laid out the process for privatization.
2. Privatization of British Gas with the Gas Act of 1986 and creation of The Office of Gas (Ofgas) as regulator.
3. Initiation of competition in the contract market (large industrial and electric utility users who use more than 2500 therms/year) in 1989.
4. Accelerated competition in the contract market, with government targets and issuance of licenses to qualified customers taking more than 2500 therms/year at a single site in 1992.
5. Creation of the gas cost index in 1992 to limit cost passthroughs by British Gas to domestic customers who use 2500 therms/year or less.
6. Passage of the Competition and Service (Utilities) Act in 1992, which included natural gas service standards.
7. Passage of the Gas Act of 1995, which laid out targets for full competition in the domestic sector by 1998.
8. Creation of the Network Code in 1996 to ensure smooth operation of the liberalized United Kingdom gas industry.

The Oil and Gas (Enterprises) Act was passed in the year 1982 to set out the aims and objectives of the privatization of British Gas. The main issue was to break down the monopoly British Gas had in the gas supply market, so enabling the introduction of competition in gas supply. The year 1986 saw the implementation of the Gas Act, which brought about fundamental changes necessary for the privatization program to go ahead. First, Ofgas, the Office of Gas Supply, was established as the regulatory body over the gas industry. Its role is to monitor the privatization of the gas industry and to ensure customers' rights are recognized in every aspect. Second, British Gas was reorganized into a public limited company (Plc) as a monopolist with secure markets only for the medium term. BG Plc shares were released onto the stock market.

The mechanism for competition within the gas market had been implemented, with major new private sector companies created. Some were established by oil companies active in the North Sea;

others are ventures by the electricity companies and some are truly independent. The Gas Act emphasized the need for vigorous competition among the new independent suppliers with the objective of undercutting British Gas prices. The goal was to introduce competition in stages, with the industrial sector first, followed by commercial and then residential users.

2.3.1 Development of the Nonfranchise Market

In 1989, the first competition for large contract customers in industry and power generation appeared. The gas supply market did not really open up until 1992, when gas consumers buying over 2500 therms/year (million Btu, or mmBtu, per year) at a single site qualified to convert to independent suppliers of gas. The government's target was to have 60% of the contract market open to competition. By 1993, independents claimed about a 50% share of the contract market. In 1995, Ofgas reported that British Gas held a roughly 27% share. The complete breakdown of British Gas' supply monopoly occurred 1998, when every consumer of gas, including approximately 19 million domestic users, was given the option to choose their own supplier. Legislation approved in 1995 (the Gas Act of 1995) allowed liberalization of the residential market to begin in 1996.

Competition in the domestic gas supply market was introduced in stages. Under an amended Gas Act 1995, Ofgas became responsible for securing competition in gas supply to domestic customers, and issued licenses to competing companies for the transportation, shipping, and supply of gas. The Ofgas Technical Directorate enforced regulations on the quality and calorific value of gas, approved and stamped gas meters, and tested disputed meters. It also provided advice on a wide range of other technical issues. The principal job was to ensure British Gas did not take unfair advantage of its monopoly powers. This was done by limiting the prices the company could charge and setting standards for customer services. Two price controls were used (modifications have been made since restructuring was initiated). The first was a transportation and storage price control, which capped the prices TransCo, which is part of BG Plc and is the public gas transporter and operator of the pipeline system, can charge British Gas Trading and other users for transporting gas in its pipelines and storing gas. The pipeline system is operated on a contract carriage basis (suppliers arrange for transportation from beach to final customer), but with the recognition that pipeline transportation is monopolistic. The second control involved a supply price control that limited the prices British Gas Trading, now part of Centrica, could charge domestic consumers.

With the initiation of privatization in 1992, a gas cost index was used as part of the regulatory formula that limited BG's allowable revenue for the 18 million customers who use 2500 therms or less of gas per year. Ofgas introduced the index after reviewing the price control set by the government when BG was privatized in 1986. This had allowed BG to pass through to domestic customers the average actual cost of buying gas for all customers, which is about 40% of the total allowed price of gas. With introduction of the index, this part of the allowed price became based on the level of the index and was no longer linked to actual gas costs. In addition to price and service regulation, Ofgas oversaw conformance with the Network Code, established in 1996, which sets out the rights and responsibilities for TransCo and all gas shippers using the existing grid for transportation and storage. Under the code, shippers are liable for any imbalances in their daily shipments. With the network code fully operational by 1998, tariff formulas for gas suppliers will no longer apply. Gas prices will be left to market forces, including those charged to Britain's 18 million small commercial and residential customers.

2.3.2 Implementation of Domestic Competition

Key to final implementation in 1998 of domestic competition was the 1997 price control review for BG's domestic supply business. Proposals developed by Ofgas in the summer of 1996 allowed the average domestic customer to realize a reduction in gas bills, with continuing real price reductions for the following 2 years. In addition, TransCo put forth proposals for transportation/distribution tariffs. The main proposals were (1) a 3-year price control period, (2) continuation of the retail price index, (RPI) $-X$, price control on supply costs, with X set at 5% (currently 4%), (3) direct passthrough to customers of British Gas' gas purchase costs and transportation costs, (4) service standards to remain as now, but with increased compensation payments when specified standards are not met, and (5) scope for British Gas to make a profit margin of 1.5% on turnover built into prices to allow a reasonable return to shareholders, as British Gas faces the introduction of competition into the domestic market. The price of gas was made up of three components: the cost of purchasing gas from producers, the costs of transportation, and supply costs. The proposals focused on six areas in particular: the treatment of gas costs,

transportation charges, BG's operating costs in its domestic supply business, BG's profit margin, the structure of the price control, and the duration of the price control.

2.3.3 Issues

A number of primary issues were associated with Britain's gas restructuring. Take-or-pay (TOP) contract resolution was essential. As with U.S. and Canadian companies before restructuring, BG held contracts with North Sea gas producers for supply. With falling gas prices, as a consequence of competition, BG and its suppliers faced a situation in which the value of gas in the contract was higher than it was in the marketplace, rendering the contract unsustainable; new competitors in the United Kingdom market were able to contract for the cheapest gas supplies, thus renegotiation of the TOP contracts was necessary for implementation of restructuring, but the issues over reapportioning costs became politically difficult.

Setting the value of "*X*" in the price cap formula was difficult. Determination of this value relative to performance of gas companies is not simple, although it is less conflictual compared to establishment of the allowed rate base for rate of return (or cost of service) regulation traditionally used in the United States and Canada. Two other issues involved the low margins to gas suppliers and the slow implementation of domestic competition. With the onset of competition and falling gas prices, suppliers in the United Kingdom faced low and diminishing profit margins in their businesses. As in the U.S. and Canada, fear and uncertainty about reliability of gas service and the ultimate cost of service, particularly to residential customers, has delayed full implementation of domestic competition.

There was also the issue of coordination with other European initiatives. Britain leads Europe in creating a competitive natural gas market. In 1997, the European Union Council established a directive allowing large customers (above 25,000 cubic meters) to select from competitive suppliers with a target of the year 2000 for implementation by member states. Initially, each member state is to grant third-party access (TPA) to 20% of the market (reducing the customer threshold if necessary), with TPA reaching 28% after 5 years and 33% in 10 years. States have moved very slowing with this directive, with few countries establishing independent regulators or forcing incumbent monopolies to respond in spite of court actions.

TABLE II

Natural Gas Policy in Latin America

Country	Private participation in exploration and production	Private participation in pipelines	Independent regulator
Mexico	No	Yes	Yes
Colombia	Yes	Yes	Yes
Venezuela	Yes	No (pending)	No (pending)
Brazil	Yes	Yes	Yes
Bolivia	Yes	Yes	Yes
Peru	Yes	Yes	Yes

2.4 Latin American Experience

The Latin American experience with natural gas policy is summarized in Table II. During the past decade or so, experimentation in the region was active in tandem with other economic and political reforms. The disintegration of Argentina's economy and uncertainty surrounding other country regimes, as well as dissatisfaction with results of market reforms, have, as of this writing, posed a number of stumbling blocks to continued development. The region is rich in natural gas resources, but much of this supply is "stranded" as a result of inadequate domestic markets and the expense of exporting natural gas in the form of LNG (Trinidad and Tobago being a notable exception). With respect to exploration and production, Latin American countries have been characterized by the presence of sovereign, state-owned oil companies, such that privatization was an initial requirement in many cases (Mexico is unique in not pursuing this step). Argentina and Mexico represent distinct cases that reflect the range of issues in the region.

2.4.1 Argentina

Until its recent financial and economic travails, Argentina had made great strides to build a competitive market for natural gas. Most of that country's initiatives have survived the depression of the past few years, but market participants have suffered throughout the period of energy economic malaise.

In the late 1980s, the government of Argentina began to privatize state-owned energy companies as part of an economic reform drive to combat hyperinflation and a chronically underperforming domestic economy. These efforts included the

privatizations of Gas del Estado, which controlled the Argentine natural gas transportation and distribution grids, and Yacimientos Petroliferos Fiscales-Argentina (YPF-A), the national oil and gas company. Argentina now has two main transportation companies, Transportadora del Gas Norte (TGN) and Transportadora del Gas Sur (TGS), both owned and operated by consortia of Argentine, U.S., Canadian, and European companies. Eight investor-owned distribution systems are now in operation, also with combinations of Argentine and foreign direct investment. The transportation system and portions of the distribution systems operate under open-access conditions following the Canadian and U.S. systems. As in Canada and the U.S., physical, economic bypass is allowed.

The regulatory framework in Argentina is provided by Ente Nacional Regulador del Gas (ENARGAS), an independent national commission. ENARGAS maintains several objectives:

1. To protect consumers' rights.
2. To promote competition.
3. To guarantee long-term investment.
4. To regulate distribution and transport services.
5. To guarantee fair and nondiscriminatory tariffs.

For tariff regulation, ENARGAS uses a variation of the British RPI$-X$ methodology and incentive mechanisms to reward performance by the private operators. Prices to final customers are unbundled and the passthrough method is used to guarantee price transparency. The formula is specified as follows:

$$\text{Total price} = \text{gas price} + \text{transportation price} + \text{distribution margin}$$

Subsidized tariffs are allowed for certain customers but must be justified to the Presupuesto Nacional, the federal budget authority. No dumping of gas is permitted. Tariffs vary with distance and type of contract (firm or interruptible) as in the United States, Canada, and Britain. It is possible to adjust distribution tariffs as often as every 6 months in correlation with the U.S. price index and productivity factors. (Because the Argentine dollar was pegged to the U.S. dollar as an antiinflation strategy, U.S. price movements are used in a variety of Argentine economic sectors.) It is also possible to adjust distribution tariffs seasonally. The overall tariff structure is revised every 15 years.

ENARGAS has several roles (regulator, enforcer, judge) and functions principally to advise users; publish information on the industry's evolution and maintain a comprehensive library; approve transportation and distribution tariffs; authorize the building and expansion of the transportation and distribution networks; foster open access and nondiscriminatory treatment; normalize security and gas quality, measurement, and odorization processes; establish the rules for new permits; implement sanctions; perform regular inspections of sites; require periodical informs to the companies that participate in the industry; guarantee competitive business practices; and resolve controversies between companies and third interested parties through public consultations. The commission resolves disputes using public consultations and general resolutions. This counters the style of U.S. and Canadian commissions, which use open hearings and technical conferences to allow input into decision making and for dispute settlement. ENARGAS, an independent regulatory commission with no influence from Argentina's energy ministry, has three appointed commissioners. The commission uses both internal and external auditors to monitor its activities and reports to the national executive branch of the Argentine government.

Argentina's model is closest to the U.S. upstream regime. The potential exists in Argentina for a healthy gas (and power) commodity market with trading and marketing activities used by producers, marketers, and large customers (including local distributors) to hedge against natural gas price volatility. Yet, today YPF still controls around 65% of marketed natural gas production and about 75% of gas produced in fields. Thus, even the privatized YPF remains extremely powerful.

2.4.2 Mexico

Mexico has pursued a strategy of reserving upstream petroleum and gas exploration to Petroleos Mexicanos (Pemex), the national oil company. During the early days of Mexico's industry, oil and gas exploration was carried out by private foreign and Mexican companies. Disputes between the Mexican government and foreign operators, and political imperatives following Mexico's revolution, resulted in the 1938 nationalization of Mexico's oil industry. Article 27 of the regulatory law to Mexico's constitution stipulates that Pemex has sole control of the production of oil and gas and the products derived from the raw resources.

During the 1970s, hydrocarbon production did not keep pace with economic modernization, so that by 1973 Mexico found itself to be a net importer of crude oil. Critical discoveries restored Mexico's

stature as an oil producer and exporter. Investment in upstream activities continued until the early 1980s, when Mexico's external debt crisis, exacerbated by falling world crude prices, triggered a contraction in government spending. The collapse of crude prices in 1986 impacted upstream activity even more. Spending by Pemex on exploration and production dropped from its peak of approximately 86% of Pemex's total budget in 1982 to less than 60% in 1988. Historically, crude oil has been given priority because of its export potential and value. However, two factors contributed to an effort to increase natural gas production. One, in the late 1970s, was the desire to increase gas sales to the United States, which led to construction of the Cactus–Reynosa pipeline as a result of negotiations between Pemex and Border Gas, a consortium of U.S. companies. The 2-billion cubic feet (bcf)/day project was never realized because of disputes about pricing. The second factor was concern about inefficient utilization of energy. Energy use was, and is, highest in the energy sector. Expenditures were made to gather and transmit gas, especially from the huge Bay of Campeche fields, and to reduce wasteful gas flaring, which declined from 26% of production in 1970 to less than 3% by 1989. Domestic consumption of natural gas continued to grow in the 1980s, but with relatively little new investment in natural gas production and transmission made by Pemex. The result is Mexico's current situation of inadequate domestic production capacity to satisfy natural gas demand.

In the 1980s, in response to the critical economic situation Mexico faced after the oil market crash and currency devaluations, the Mexican government began to implement market reforms. Public opinion and political will for privatizing Pemex have historically been weak, but a series of major accidents, chronic shortages, and unreliable service forced the managers at Pemex to take action. The government gradually removed the obligations on Pemex to provide everything from roads to hospitals and schools as part of its social obligations to the state. Pemex began to reduce its huge employment from more than 250,000 to just over 133,000 today. In 1992, Pemex was reorganized into four functional subsidiaries for exploration and production, refining, natural gas and basic petrochemicals, and secondary petrochemicals. The government also changed Pemex's tax status by creating a corporate tax rather than controlling all of Pemex's revenues and returning some portion to the company for reinvestment. The corporate tax rate for Pemex remains high (more than 60%) and Pemex still does not have independence with respect to its budget planning.

In 1995, further, more dramatic steps were taken to reform Mexico's energy sector. The regulatory law to the constitution was changed to allow private investment in natural gas transportation, distribution, and storage, in recognition of the importance of this fuel to Mexico's economic development. A regulatory commission has been created (the Comisión Reguladora de Energía, or CRE), charged with the privatization of assets formerly controlled by Pemex. Rules have been established for pipeline tariffs and first-hand sales of imported gas from the United States (although Pemex is expected, at some point, to resume its bid to be a net exporter of gas). The CRE handles all auctions for state-owned assets (portions of the Pemex pipeline grid and local distribution networks) that are to be turned over to private operators, and uses a price cap formula similar to the Ofgas RPI$-X$ to regulate tariffs. The CRE has five appointed commissioners. A law passed by the Mexican Congress in 1995 established the CRE as an independent entity, but Mexico's energy ministry retains a great deal of influence. Like ENARGAS, the CRE handles conflicts and disputes with operators through private consultations rather than through the public meetings typical in the United States and Canada.

In spite of all of these reforms, deep problems exist in Mexico's oil and gas sector. Pemex has watched its market share of exported oil erode as other countries moved aggressively to lure private investment into their upstream businesses. The investment demands on Pemex for improvement and expansion projects are huge. Although the company has had some success with foreign placements of debt, many questions remain about Pemex's ability to finance capital improvements. Finally, although the effort to attract private investment into Mexico's natural gas pipeline and distribution segments continues, Pemex remains the sole supplier of natural gas, which will restrict growth of Mexico's natural gas market.

SEE ALSO THE FOLLOWING ARTICLES

Coal Industry, Energy Policy in • *Greenhouse Gas Emissions from Energy Systems, Comparison and Overview* • *Markets for Natural Gas* • *National*

Energy Policy: United States • Natural Gas, History of • Natural Gas Processing and Products • Natural Gas Resources, Global Distribution of • Natural Gas Resources, Unconventional • Natural Gas Transportation and Storage • Occupational Health Risks in Crude Oil and Natural Gas Extraction • Oil-Led Development: Social, Political, and Economic Consequences

Further Reading

Foss, M. M. (1995). "Natural Gas in the Twenty-First Century: Adjusting to the New Reality." Doctoral dissertation, University of Houston, Houston, Texas.

Foss, M. M. (1998). Worldwide transitions: Energy sector reform and market development. *Natur. Resourc. Environ.* (Spring 1998).

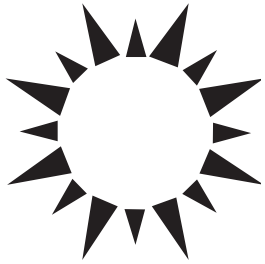
Foss, M. M. (2000). Perspectives on the international exploration business. *In* "International Oil and Gas Ventures: A Business Perspective" (G. Kronman and T. O'Connor, Eds.), pp. 11–39. American Association of Petroleum Geologists, Tulsa, Oklahoma.

Foss, M. M. (2002). Latin American gas and gas-fired power shows progress, risk. *In* "Natural Gas and Electric Industries Analysis" (R. Willett, Ed.), pp. 230–247. Financial Communications Company, Houston.

Tussing, A., and Barlow, C. (1984). "The Natural Gas Industry: Evolution, Structure, and Economics." Ballinger Publ. Co., Cambridge, Massachusetts.

U.S. Energy Information Administration (EIA). (1992–1998). "Natural Gas Issues and Trends." Available on the Internet at www.eia.doe.gov.

Natural Gas Processing and Products



RICHARD G. MALLINSON
University of Oklahoma
Norman, Oklahoma, United States

1. **Requirements for Gas Processing**
2. **Gas Dehydration**
3. **Gas Sweetening**
4. **Hydrocarbon Recovery and Fractionation**
5. **Conversion to Transportable Chemicals and Fuels**

Glossary

acid gas The acidic constituents of natural gas, H_2S, CO_2, mercaptans (RSH), CS_2, and COS.

dew point The temperature and pressure at which a gas mixture begins condensing to form a liquid phase; for a fixed composition, the point is reached by lowering the temperature or increasing the pressure except when the mixture exhibits retrograde condensation.

ethane, propane, butanes, liquefied petroleum gas (LPG), and natural gasoline Natural gas liquid products for which there are various purity grades; they are defined by a combination of maximum and minimum compositions, vapor pressures, and boiling points.

gas conditioning The process of removal of contaminants, usually acid gases, nitrogen, and water, from natural gas.

gas dehydration The process of removal of water from natural gas to achieve a low dew point that will avoid condensation of water during further processing, transportation, and use.

gas sweetening/treating The process of removal of acid gas components from natural gas.

gas to liquids (GTL) Processes for the chemical conversion of methane to other chemicals and fuels that are predominantly liquids; most frequently refers to Fisher–Tropsch-type processes that preferably produce hydrocarbons from C_5 to C_{30}.

liquefied petroleum gas (LPG) A natural gas liquid product composed primarily of a mixture of propane and butanes.

natural gas liquids (NGL) The higher hydrocarbons separated from natural gas that are usually produced and stored as liquids under pressure at ambient temperature.

natural gasoline A natural gas liquid product that consists primarily of pentanes and higher alkane hydrocarbons (C_5^+); it is a liquid at ambient temperature and atmospheric pressure and has a low octane number.

retrograde condensation The phenomenon where a hydrocarbon gas mixture of a particular composition will pass through the dew point and begin condensation of a liquid phase when the pressure is lowered.

sour gas Natural gas that has at least a small percentage of acid gases, specifically sulfur-containing components.

Gas processing is the preparation of raw natural gas as it is produced from the reservoir for transportation to markets for utilization. Traditionally, this has been primarily the removal of chemical constituents that are not desired due to safety reasons (e.g., H_2S, Hg), operability (e.g., water due to condensation and corrosion), and/or economics (e.g., CO_2 that lowers the heating value or those components that may be profitably separated and marketed as a separate product, natural gas liquids, C_2 to C_5^+ hydrocarbons). The final product for this has been compressed gas that is fed into pipeline networks. Increasingly, the processing required to transport the gas to markets for end use has included the production of liquefied natural gas by complex refrigeration processes as well as the chemical transformation of the methane into solid or liquid chemicals and fuels that are commonly produced from natural gas, are much more easily transported than the gas itself, and satisfy significant markets for the products. The most commonly produced are ammonia and its derivative fertilizers and methanol, but there is strong promise for production of liquid hydrocarbon fuels on much larger scales to satisfy increasing demand, especially for cleaner fuels.

 © 2004 Elsevier Inc. All rights reserved.

1. REQUIREMENTS FOR GAS PROCESSING

Feed gas to a processing plant can have extremely large variations in composition from one location to another. Produced gas, subject to removal of liquids and solids with separators only at the field locations, is frequently saturated with water due to intimate contact between gas and water in the reservoir formation. Gas associated with oil production may be saturated with hydrocarbons from ethane to C_{10}^{+}. In both cases, the amounts will vary with the pressure and temperature of the gas. In the case of water, this is commonly a proportion up to several percentage points. For hydrocarbons, this can be well above 10%. For the acid gases, primarily CO_2 and H_2S but including other sulfur-containing species such as mercaptans (RSH) and COS, reservoirs in which the majority of the gas volume is comprised of these species is possible. At some point, the high percentage of these contaminants will make the gas uneconomical for recovery, but dealing with several percentage points H_2S and CO_2 above 10% is common. The other contaminant commonly found in gas at these percentage levels and higher is nitrogen (and rarely helium). Lower levels of contaminants that must be removed are possible but less common, with mercury being one example (generally removed using solid adsorbents).

The requirements of gas processing are to remove the contaminants mentioned previously to appropriate levels. These levels are specified by contract and determined by the market for which the gas is destined as well as for safety and operability considerations. Typical "pipeline quality" gas specifications are discussed here.

Water must be removed to levels that will ensure that no condensation of the water will occur in any of the subsequent operations or distribution of the gas. This is both to minimize corrosion and to prevent damage to equipment due to two-phase flow where liquid water droplets can damage machines that compress or expand the gas and cause erosion by impingement on surfaces at high velocities. Dehydration also prevents the formation of methane hydrates, solid crystalline complexes of methane with water molecules, that can plug flow lines as well as damage equipment. To avoid condensation, a dew point for the gas is selected based on a specified temperature and pressure. This represents the combination of the minimum temperature and the maximum pressure to which the gas may be expected to be subjected, and at those conditions where no water will condense (e.g., 0°C and 6.7 MPa). Typical values of allowable water content at such conditions are in the range 50 to 110 mg/m^3 and vary by location due to climate and other factors.

For sulfur content, specifications for H_2S, mercaptans, and total sulfur are common, with H_2S and mercaptans each having limits in the range of 6 to 24 mg/m^3 and total sulfur having limits from 115 to 460 mg/m^3. These limits, although needed in part due to corrosiveness of the sulfur compounds, are due in significant part to their noxious and toxic properties.

Specifications for CO_2 and nitrogen are primarily set due to the fact that they are inert and have no heating value and, therefore, reduce the heating value of the product gas. There are typically limits of a few percentage points on each, but their amounts are also indirectly constrained by a heating value specification of the product gas.

The specifications for hydrocarbons are somewhat more complex. For methane, there is some minimum specification (e.g., 75%). For higher hydrocarbons, there may be individual maximum limits (e.g., 10% ethane, 5% propane, 2% butanes). These limits are based on ensuring against condensation, in the same fashion as for water but also for the combustion processes for which most pipeline gas is used. The presence of higher hydrocarbons changes the required air/fuel ratio for combustion and, when outside of design limits, can cause incomplete combustion with additional pollution and higher or lower temperatures, among other operational problems. From the standpoint of the gas processor, there are frequently economic incentives to recover the higher hydrocarbons from the gas. When present in sufficient quantities, the value of the hydrocarbons can exceed the cost of their recovery. Typical products of value include ethane, propane, butanes, and natural gasoline. These are collectively called natural gas liquids and, as mentioned previously, are frequently present in near saturation conditions (at their dew points) in gas associated with oil production. Each of these products has its own specifications that put minimum and maximum limits on the hydrocarbon constituents and may include heating value and boiling point ranges.

For liquefied natural gas (LNG), the product gas will have specifications with considerably lower contaminant levels to avoid contaminants that could condense or freeze at the very low temperatures involved in production of LNG (−161°C). For gas to chemicals and fuels processes, the feed gas may have similar specifications to pipeline gas except that sulfur (and possibly other trace contaminants) will be

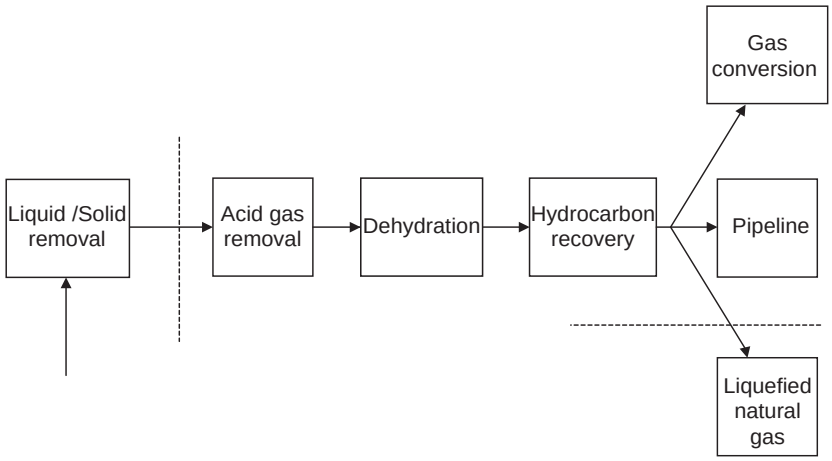


FIGURE 1 Block diagram of major gas-producing units.

further restricted due to its ability to deactivate catalysts used for the conversion reactions.

Figure 1 illustrates a block diagram of the processing units that may be found in a gas processing plant. The presence, order, and relative size of the process units will be dependent on the composition of the feed, the gas product specifications, and the specific processes selected for each unit. Gas conversion plants are considered separately given that these processes are a follow-on to the type of plant shown in Fig. 1.

2. GAS DEHYDRATION

Gas dehydration is accomplished by means of one of two primary processes (or occasionally both). Absorption of water vapor from the gas into a hygroscopic liquid is the predominant method, but adsorption of water onto solid desiccants is also used, particularly for achieving very low dew points.

The hygroscopic liquid is most frequently triethylene glycol (Fig. 2), but other glycols are sometimes used. Glycols are quite stable, but the presence of oxygen can make the solutions highly corrosive. Water is highly soluble in the glycol, and when a water-containing natural gas stream is placed in contact with a glycol solution, the water is absorbed from the gas phase into the liquid phase. The other important property is that the boiling point of the glycols, especially triethylene glycol, is much higher than that of water (288°C vs 100°C at atmospheric pressure). Therefore, the glycol–water mixture may be heated to revaporize the water at temperatures far below the boiling point of the glycol, resulting in

(1) $HO–CH_2CH_2–OH$

(2) $HO–CH_2CH_2–O–CH_2CH_2–OH$

(3) $HO–CH_2–CH_2–O–CH_2CH_2–O–CH_2CH_2–OH$

FIGURE 2 (1) Ethylene, (2) diethylene, and (3) triethylene glycol.

little covaporization of the glycol. In addition, aliphatic hydrocarbons are absorbed only in very small amounts, although aromatics are very soluble. Acid gases that may be present are also absorbed in small amounts, although their presence can affect he solubility of water in glycol during the absorption and can also result in absorption of aliphatic hydrocarbons.

Figure 3 presents a process flow diagram for a glycol dehydration system. The glycol contactor is a vertical cylindrical column with dimensions of from approximately 0.5 m up to several meters in diameter, with larger sizes required for larger gas flow rates being processed. Inside the column are usually 8 to 10 contacting trays placed above each other approximately 0.6 m apart. Although trays are common, packing materials that allow the liquid to trickle down over their surface to create a high gas–liquid contacting area are also used and can be preferred in a number of applications, particularly for increasing the capacity of existing trayed columns. The total height of the column is approximately 5 to 8 m and is not greatly affected by the flow rates. The lean (low water content) glycol solution is introduced at the top of the column and flows across each tray and down to the next. The water-containing natural gas is introduced into the

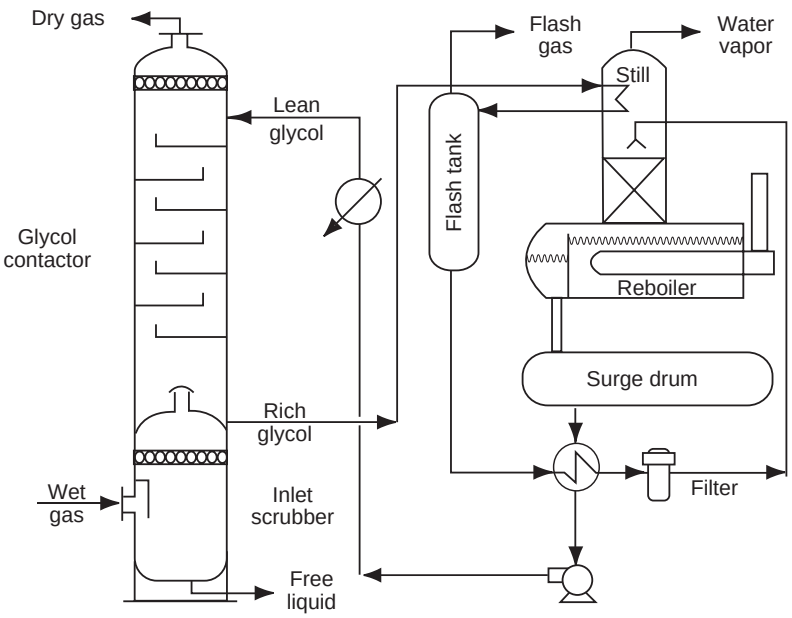


FIGURE 3 Typical glycol gas dehydration process diagram. Reprinted from Gas Processors and Suppliers Association. (1998). "GPSA Engineering Data Book," Vols. 1–2, 11th ed. GPSA, Tulsa, OK.

bottom of the column and flows upward through perforations or caps in the trays where the gas and liquid are highly mixed. This mixing provides good contact between the gas and the liquid to allow the transfer of the water from the gas to the liquid. This countercurrent flow pattern provides the best approach to equilibrium for the removal of water from the gas phase. The gas has less water remaining at each successive tray until it leaves the column at the top. The liquid increases its water content as it flows down through the column until it is discharged at the bottom. Typical applications require approximately 0.25 L of TEG/kg water removed and a lean (inlet) glycol concentration of approximately 99%. The contacting generally takes place at near ambient temperatures ($\sim 38^{\circ}$C) and at convenient pressures, with as little lowering of the gas pressure as possible (2–6 MPa).

The glycol must then be regenerated to return it to its lean concentration for reuse in this continuous process. The water-rich glycol flows from the bottom of the column to the still/reboiler, where it is heated to approximately 200°C and the water is boiled off at atmospheric pressure and vented. Any absorbed aromatics are also vaporized, and emissions handling equipment may be needed in this case. The lean glycol is then cooled and pumped back to the glycol contactor.

A number of variants of this process that improve performance, particularly under challenging circumstances such as where especially low dew points are required, are in commercial operation. In that case, a vacuum still can achieve higher lean glycol purity, say above 99%.

Solid desiccant dehydration makes use of the attraction between water in the natural gas and the surface of certain solid materials. In these adsorption processes, the surface properties are such that a weak bond will form between the solid surface and water when exposed to water during the gas phase. The bond strength is a function of temperature with high water-adsorbing capacity at ambient operating temperatures and lower capacity as the temperature increases. These materials can be fabricated into particles with diameters in the range of 1 to 5 mm and with high active surface areas and water adsorption capacities from 20% to more than 40% by weight. The most common materials are silica gels, activated aluminas, and crystalline molecular sieve zeolites.

Although it is a more expensive process than glycol dehydration, the solid desiccant process can achieve significantly lower dew points (as low as -60 to -100°C). Adsorption is also useful for simultaneous removal of acid gases.

The solid particles are placed as a bed in a vertical cylindrical column. The wet feed gas flows down through the column at a velocity designed to provide an appropriate residence time for the adsorption to take place at high rates, but it is limited by the loss of

pressure due to the friction of the gas flow through the void spaces between the particles that increases as the velocity increases. The adsorption takes place at a convenient gas pressure, say 2 to 6 MPa, at near ambient temperature.

The process operates continuously in a cyclic manner with two columns in parallel, as shown in Fig. 4. One column accepts the flow of natural gas for a period of time that allows the water-adsorbing capacity of the volume of particles in the column to be reached. A larger column of particles has a higher capacity and operates for a longer time but commonly for less than 12 h. During this time period, the other column is in the process of being regenerated, with the water removed by flowing a stream of hot ($\sim$315°C) regeneration gas over the particle bed. The regeneration gas is a portion of the dried gas that, after passing over the regenerating bed and taking up the "desorbed" water, is cooled back to ambient temperature. This stream passes below its dew point; thus, water condenses and is removed in the knockout drum. The gas, now saturated with water, is put back into the feed gas. It should be pointed out that at the high regeneration temperature, the regeneration gas can accept significantly more water than it can at ambient temperature; thus, only a relatively small portion must be reprocessed. When the adsorbing column is saturated, valves are switched to redirect the wet feed gas to the regenerated column and to begin regeneration of the "saturated" column.

Other processes for dehydration are available for commercial application; however, these are generally limited to specialty applications such as very small gas flow rates. Membrane permeation is a technology that continues to advance, although limitations on selectivity cause unacceptable losses of feed gas for dehydration. These alternative process choices are generally more costly for the most common gas dehydration applications.

3. GAS SWEETENING

A wide array of processes are available for the removal of the acid gases (primarily H_2S and CO_2 but including all sulfur compounds such as mercaptans, CS_2, and COS). The selection of the most appropriate one (both technically and economically) depends primarily on the feed gas composition, specifically which acid gases are present (or absent) and at what concentrations as well as the final product specifications. The major categories of sweetening processes are amine reaction processes, physical absorption processes, combination processes, and cyclic processes. Figure 5 shows a matrix that considers some of the criteria for process selection.

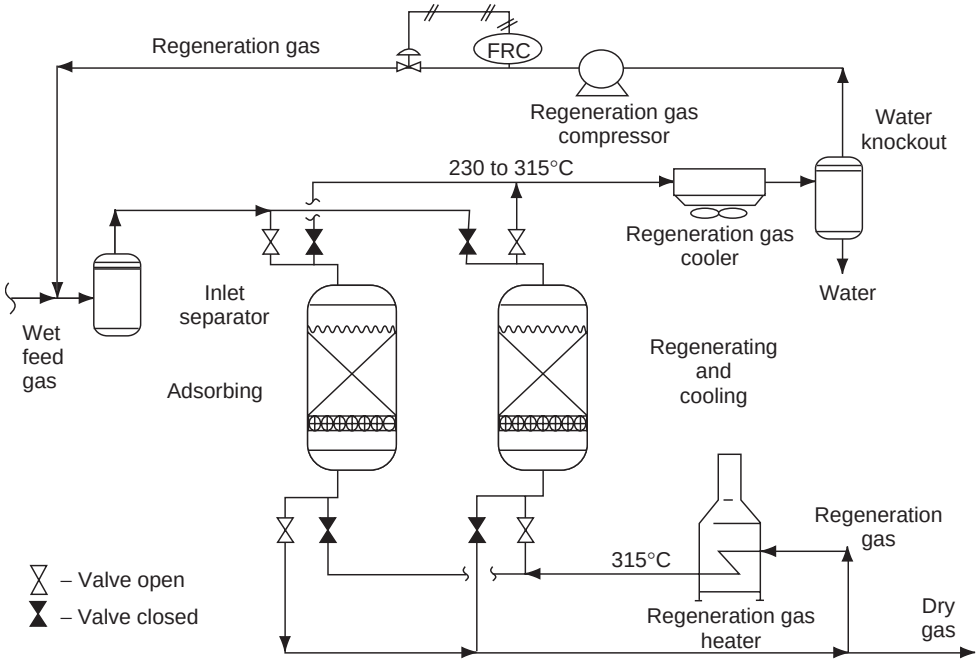


FIGURE 4 Solid desiccant dehydration process diagram. Reprinted from Gas Processors and Suppliers Association. (1998). "GPSA Engineering Data Book," Vols. 1–2, 11th ed. GPSA, Tulsa, OK.

	Normally capable of meeting 0.25 grain[a] H_2S	Removes mercaptans and COS	Selective H_2S removal	Solution degraded (by)
Primary amine	Yes	Partial	No	Yes (COS, CO_2, CS_2)
Secondary amine	Yes	Partial	No	Some (COS, CO_2, CS_2)
Tertiary amine	Yes	Partial	Yes[b]	No
Combination	Yes	Yes	Yes[b]	Some (CO_2, CS_2)
Physical solvent	Maybe[c]	Slight	Yes[b]	No
Solid bed	Yes	Yes	Yes[b]	No
Liquid redox	Yes	No	Yes	CO_2 at high concentrations
Sacrificial	Yes	Partial	Yes	No

[a] 0.25 grains H_2S/100 scf ≈ 5.7 mg/m^3 (4 ppmv).
[b] Some selectivity exhibited.
[c] Can make 0.25 grain under some conditions.

FIGURE 5 Matrix of sweetening process selection factors. Data from Gas Processors and Suppliers Association. (1998). "GPSA Engineering Data Book," Vols. 1–2, 11th ed. GPSA, Tulsa, OK.

The absorption of the acid gases into aqueous solutions is low due to low solubilities, but the placement of a basic amine into the solution allows the absorbed acid gas to react to form an acid–base complex. This creates a high capacity for absorption that depends on the concentration of the amine. Figure 6 shows the dominant amines used in these solutions. For a primary amine (RNH_2), one molecule of acid gas is complexed with each amine molecule. For a secondary amine (R_2NH), two molecules of acid gas are complexed, and for a tertiary amine (R_3N), three are complexed, except that CO_2 is not complexed with the tertiary amine, giving it a preferential selectivity for sulfur (Fig. 7). In practice, only approximately one-third of the theoretical complexation capacity (moles of acid gas complexed per mole of amine) is realized.

Monoethanolamine	$HOCH_2CH_2NH_2$
Diethanolamine	$(HOCH_2CH_2)_2NH$
Triethanolamine	$(HOCH_2CH_2)_3N$
Methyldiethanolamine	$(HOCH_2CH_2)_2NCH_3$
Diisopropanolamine	$(CH_3HOCHCH_2)_2NH$
Diglycolamine	$HOCH_2CH_2OCH_2CH_2NH_2$

FIGURE 6 Major amines used in gas sweetening.

$$RNH_2 + H_2S \longleftrightarrow RNH_3^+HS^-$$
$$RNH_2 + CO_2 \longleftrightarrow RNHCO_2^-H^+$$

FIGURE 7 Simplified amine acid gas chemistry.

Figure 8 shows a process flow diagram for an amine reaction-type gas sweetening process. The sour gas enters the column from the bottom and flows up and out the top while the lean (low acid gas content) amine solution flows down from the top to the bottom of the column. As with dehydration columns, the countercurrent flows of the two streams maximize the transfer of acid gas constituents to the liquid phase. The column internals are trays or packing in similar fashion to the dehydration column. The gas leaving the column may be passed through a water scrubber to recover more volatile amines (e.g., MEA, DEA). The gas at this point is essentially saturated with water and would then pass to a dehydration system. The column operates at near ambient temperature and a convenient gas pressure, say 38°C and 2 to 6 MPa. The acid gas-rich amine solution leaves the bottom of the column, is heated, and flows to the upper part of the stripping column. This also is a gas–liquid contactor that creates surface area between the gas and liquid phases that allow, at a higher temperature of approximately 110 to 130°C, the sour gases to be stripped into the gas phase. The bottom of the stripping column includes a reboiler where some of the heating of the amine solution occurs. The stripped or lean amine solution is then cooled and pumped back to the sweetening column. Both columns' sizes depend on the volumetric flow rates of the gas and liquid streams, with larger flow rates requiring larger diameters (in the range of 0.4 m to several meters), and on the amount of sour gas that is transferred. The amine concentration in the aqueous solution is typically in the range of 20 to 30%, and this defines the capacity of the solution and the amount that is needed to achieve the designed sweetening specification (e.g., liters of amine solution per cubed meter of gas treated). The acid gas leaving the stripper may be vented or burned if sulfur is absent or sufficiently low, but if significant amounts

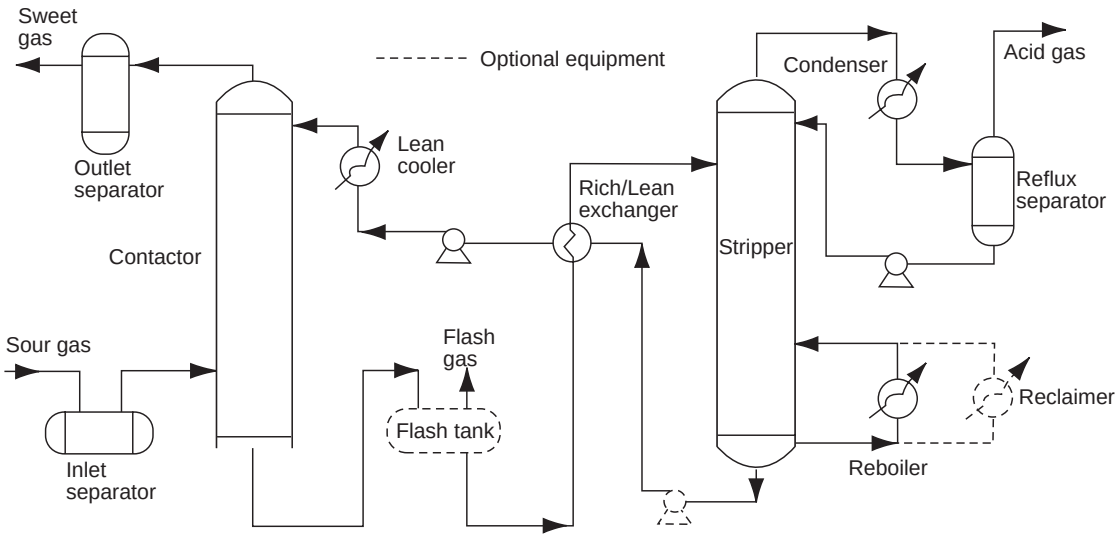


FIGURE 8 Example of an amine reaction-type sweetening process flow diagram. Reprinted from Gas Processors and Suppliers Association. (1998). "GPSA Engineering Data Book," Vols. 1–2, 11th ed. GPSA, Tulsa, OK.

are present, the sulfur-containing gas will have to be recovered.

Physical absorption processes operate in the same fashion as does glycol dehydration. A solvent that has a high capacity for acid gases is contacted with the sour natural gas in an absorption column (Fig. 3). In some cases, the solvent may be selective for sulfur compounds, may also absorb water so that the need for a separate dehydration step is eliminated, and/or may also absorb hydrocarbons advantageously. The regeneration of the solvent is generally different from the case with glycols in that it does not depend on boiling off the acid gases. Instead, pressure reductions allow the absorbed gases to vaporize, and a stripping column may be used to provide gas–liquid contacting to enhance the vaporization. Major processes using this method include the Selexol process using a polyethylene glycol derivative as the solvent, the Fluor solvent process using anhydrous propylene carbonate, and the Rectisol process using methanol.

Combined processes use both reacting and absorbing materials. Shell's Sulfinol process uses an aqueous amine along with Sulfolane, a proprietary physical solvent.

In general, cyclic processes operate by contacting with a solid adsorbent (e.g., iron sponge, zinc oxide, molecular sieve zeolites) that must be regenerated in similar fashion to the solid desiccant dehydration process or by contacting with a liquid solution (e.g., Chemsweet, Sulfa-check, Sulfatreat) that must be changed and regenerated at the end of the cycle.

Other technologies are available, including iron chelation redox-type processes and membrane permeation processes that may have applications in specific circumstances.

The major process for sulfur recovery is the Claus process with a number of possible variations. The sulfur-containing gas is first burned with air in a thermal reaction furnace to convert approximately one-third of the sulfurous gas to SO_2. At the high temperature ($\sim$1000°C), some elemental sulfur is formed. The gas is then cooled below the sulfur dew point to allow much of it to condense. The gas is then reheated to approximately 400°C and passed over an alumina-type catalyst that allows the reaction of H_2S and SO_2 to form sulfur and water. This reaction is equilibrium limited in that complete conversion is not thermodynamically feasible (lower temperature favors higher conversion) because the catalyst requires a high temperature ($\sim$400°C) to actively carry out the reaction. To achieve higher conversion to sulfur, the converted gas stream is then cooled below the dew point of sulfur that then condenses, and the remaining gas is once again heated and passed over a second catalytic reactor. This may be repeated for further conversion in a third catalytic reactor to achieve a high enough conversion to meet requirements for release of the remaining gas to the atmosphere, although further cleanup of this gas by combustion and (possibly) scrubbers may be necessary.

The primary use for the recovered sulfur is in the manufacture of sulfuric acid.

4. HYDROCARBON RECOVERY AND FRACTIONATION

The first objective of hydrocarbon recovery in natural gas processing is the reduction of the hydrocarbon dew point so that condensation cannot occur during subsequent processing, transportation, and use of the gas. Of course, the removal of these hydrocarbons causes a loss of volume of the product gas, called shrinkage, as well as a loss of heating value given that the higher hydrocarbons have a greater heat content per volume than does methane. If the higher hydrocarbons cannot be sold or used for fuel, this is a substantial loss of value that must be minimized. Historical practice frequently has been to burn those products to simply dispose of them. (Indeed, much the same could be said about natural gas found with oil, although this has changed substantially during the past decade or so.) The second objective is to maximize recovery of the higher hydrocarbons as various liquid products collectively referred to as natural gas liquids. This is now quite common, with the enhanced economic value more than offsetting the loss of value to the natural gas product.

Earlier in the article, it was mentioned that the required product dew point for water occurred at the combination of the highest pressure and the lowest temperature. For hydrocarbons, the situation is more complex in that gaseous hydrocarbon mixtures undergo the phenomenon of retrograde condensation. This is when, at a fixed temperature and composition, a lowering of the pressure may cause condensation. Thus, lowering the pressure moves the system toward the dew point rather than away from it (as is the case for most other systems, e.g., water vapor in natural gas). Thus, the operative criteria for higher hydrocarbon removal is their removal to levels that ensure that condensation cannot occur at the lowest expected temperature at any pressure. The point at which this occurs is referred to as the cricondentherm.

The removal of the higher hydrocarbons proceeds through cooling of the gas at a pressure of convenience to maintain as high a pressure for pipeline flow as possible. Figure 9 shows the normal boiling points of the major hydrocarbons found in natural gas. As can be seen, condensation and recovery of greater amounts of lighter hydrocarbons requires lower temperatures, and this is true at any pressure. Thus, the lowest temperature required will be dictated by the composition of the gas, the fractions of specific components to be recovered (or dew point to be achieved), and the operating pressure.

n-Pentane	36.07
i-Pentane	27.85
n-Butane	−0.5
i-Butane	−11.72
Propane	−42.04
Ethane	−88.60
Methane	−161.49

FIGURE 9 Boiling points of natural gas hydrocarbons at atmospheric pressure (in °C).

If the pressure is relatively low, a refrigeration process is used to chill the gas sufficiently to cause condensation of the higher hydrocarbons. These are typically propane and higher with small amounts of ethane. The condensate is then removed in a gas–liquid separator and flows to a stabilizer column that makes sure that any lighter hydrocarbons that are dissolved in the condensate, such as methane, ethane, and small amounts of higher hydrocarbons, can revaporize to an equilibrium composition and will not vaporize during later liquid processing, storage, and transportation. The stabilizer column is a vertical cylindrical column with internal trays or packing that promotes gas–liquid contact and mixing, as discussed earlier. The gas from the cold separator and the stabilizer column are then combined to form the product gas stream.

An alternative process is the use of a solvent to absorb the higher hydrocarbons, analogous to solvent dehydration. Lean oil absorption can be used as the process. The lean oil is simply another hydrocarbon stream of a sufficiently low vapor pressure that it remains in the liquid phase through the absorption and regeneration. Regeneration is carried out by heating the rich oil to allow the higher hydrocarbons absorbed from the natural gas stream to revaporize. Such plants are relatively uncommon today.

If the gas is available at sufficiently high pressures, and especially if greater recovery of natural gas liquids is desired for economic reasons, the cooling of the gas is achieved by allowing it to expand, losing some of its pressure. This is the mainstay of hydrocarbon recovery processing. There are two alternatives for this, with the simplest being the JT expansion making use of the Joule–Thomson effect where the expansion occurs as the gas flows through a restrictive valve. The design determines the degree of cooling required to achieve the amount of natural gas liquids condensation required, and additional

refrigeration can be provided if needed. This type of plant, although common, is relatively inefficient in that there can be no work recovered from the expansion.

Most newer plants, particularly those that wish to maximize recovery of gas liquids, allow the gas to expand through a turbine where the gas cooling is achieved and some work is recovered from the turbine. Usually, the turbine shaft drives a compressor also used in the process. The basic process flow diagram is shown in Fig. 10. The feed gas is first chilled by cooler gas streams in the plant, and any condensed liquids are removed in the low-temperature cooler before reaching the turbo expander. The pressure is reduced in the expander, from approximately 7 to 4 MPa in the example shown, while the temperature drops from 51 to −77°C. After the additional liquid condensate is removed in the very low-temperature separator (called the demethanizer), the cold gas is warmed by chilling the incoming feed gas. The product gas can now be partially repressurized by the compressor that is driven by the expander turbine. Additional recompression may or may not be required. There are many variations of this process to accommodate specific composition and recovery requirements, for example, the use of external refrigerants to achieve lower temperatures for greater ethane recovery. These processes can recover as much as 80 to 90% of the ethane (and virtually all of propane and higher hydrocarbons) in the natural gas.

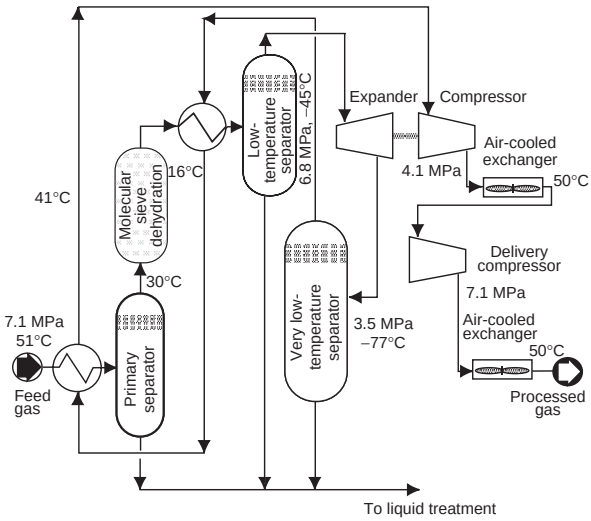


FIGURE 10 Turboexpander plant diagram. Reprinted from Rojey, A., Jaffret, C., Cornot-Gandolphe, S., Durand, B., Julian, S., and Valais, M. (1997). "Natural Gas Production Processing Transport." Editions Technip, Paris.

The processing of the gas to pipeline quality has now been completed. The liquid streams will require further separations at low temperature to produce an optimum product set. There are many possibilities for configurations that depend on both the composition of the liquids and the value of the various products. Products are defined by their composition, boiling range, and vapor pressure and may have several grades. Some major products include natural gasoline, commercial butane, commercial propane, and high-purity ethane. These products are produced by separation in one or more fractionation columns that contain trays or packing. The columns can appear in a wide range of sizes depending on the quantities of the components. Typical sizes are in the range of 0.5 to 3.0 m in diameter and 8 to 15 m in height. Higher purities require taller columns that give greater opportunity for the gas and liquid to mix. Temperatures are typically below ambient temperature, and pressures are typically in the range of 1 to 5 MPa. The conditions are selected to maximize the difference in volatilities of the components to be separated in each column. The volatility differences control whether a component appears primarily in the gas or in the liquid leaving a column. In a four-column fractionation, the raw natural gas liquids are fed to the first column, the deethanizer, with primarily ethane vaporizing and leaving from the top of the column and the remaining liquid flowing from the bottom of the column to the depropanizer, where primarily propane vaporizes and leaves from the top. The remaining liquid flows from the bottom of the column to the debutanizer, where the butanes leave from the top and flow to the fourth column that separates isobutane from normal butane. The liquid leaving the bottom of the debutanizer is the C_5^+ natural gasoline.

The product streams may flow by pipeline to subsequent processes or plants for use or to storage as liquids at ambient temperature and sufficient pressure to maintain them in the liquid state.

5. CONVERSION TO TRANSPORTABLE CHEMICALS AND FUELS

Natural gas has been a major feedstock for production of chemicals since early in the 20th century with the development of gas–solid catalytic chemistry. This is despite the intrinsic stability of the methane molecule that requires the conversion of methane to

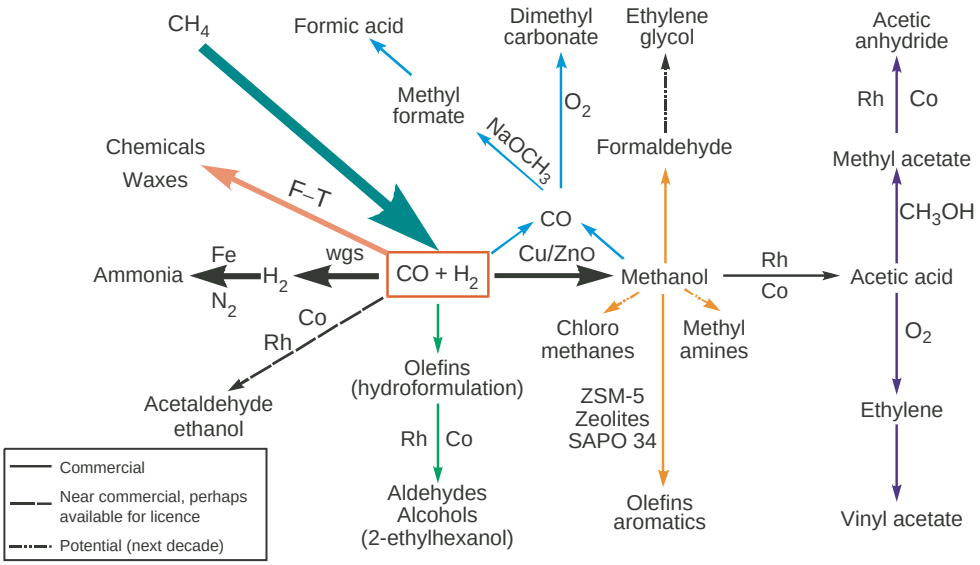


FIGURE 11 Chemical pathways for gas conversion.

carbon monoxide and hydrogen, known as synthesis gas, before further conversions to useful chemicals. Figure 11 illustrates the chemical pathways to form primary and secondary industrial chemicals from natural gas via the synthesis gas intermediate. Historically and through the current time, methanol and ammonia are among the most produced industrial chemicals. Ammonia represents the fixation of nitrogen and is the precursor for most nitrogen fertilizers. Methanol is used as a solvent as well as a feedstock for many other chemicals. Lastly, the production of liquid hydrocarbons via Fischer–Tropsch has a recent history of production from natural gas but a long history of production starting with coal. It appears that processes for production of liquid fuels from Fischer–Tropsch synthesis (FTS) are poised to undergo rapid expansion as markets for clean fuels and chemical feedstocks increase. All of these products satisfy one of the requirements for monetizing many more remote natural gas resources in that they are significantly more transportable than natural gas that is often located far from demand that can use pipelines and the relatively limited, but also increasing, markets for LNG.

The first step in the synthesis of ammonia, methanol, and liquid hydrocarbons is the production of synthesis gas. The primary means of carrying out this conversion has been methane steam reforming. This is shown in Fig. 12 as reaction 1. Steam reforming has the attribute of producing large amounts of hydrogen, one-third of which comes from the water. However, it is a very energy-intensive

(1) $CH_4 + H_2O \leftrightarrow CO + 3H_2$

(2) $CO + H_2O \leftrightarrow CO_2 + H_2$

(3) $CH_4 + 1/2\,O_2 \leftrightarrow CO + 2H_2$

(4) $3H_2 + N_2 \leftrightarrow 2NH_3$

(5) $CO_2 + 3H_2 \leftrightarrow CH_3OH + H_2O$

(6) $CO + 2H_2 \leftrightarrow -CH_2- + H_2O$

FIGURE 12 Natural gas conversion chemical pathways.

reaction. The reaction takes place on supported nickel metal catalyst pellets placed in tubes through which the natural gas and water feed flows. The thermodynamics require very high temperatures to achieve good conversions, and the reaction is conducted at a temperature of approximately 800°C. In addition, the reaction is highly endothermic, requiring large heat inputs. The combination of these two requirements leads to placing the reactor tubes inside a direct-fired combustion furnace where radiative heat transfer can provide the needed heat input. Conversion remains incomplete, and a secondary reformer with a different nickel catalyst is used at temperatures of 900 to 1100°C. Less heat input is required because the remaining conversion is smaller, and this may be supplied by feeding oxygen to cause the heat-releasing exothermic partial oxidation reaction, shown as reaction 3 in Fig. 12.

For ammonia production, the secondary reformer may use air for its oxygen supply because the accompanying nitrogen will be converted to ammonia. Also for ammonia production, the CO will be

unused, so it is converted to CO_2 by the water–gas shift reaction (reaction 2 in Fig. 12) that produces another mole of H_2 for each methane that was converted (for a total of four). Newer plants may make use of molecular sieve zeolite adsorbents to adsorb only H_2 for producing a pure hydrogen stream. In many current plants, the CO must be reduced to very low levels because it is a poison for the ammonia synthesis catalyst. This is done by converting the residual CO back to methane in a methanation reactor. The CO_2 is also removed using sweetening methods discussed earlier, leaving a nitrogen and hydrogen gas stream ready for the ammonia synthesis process.

Because of the energy intensity of steam reforming and its high H_2/CO ratio that is not needed for FTS (reaction 6 in Fig. 12), large-scale FTS plants are designed to use either partial oxidation or a combined reforming to produce synthesis gas with lower energy intensity. However, partial oxidation generally uses pure oxygen that adds to process cost and complexity. Combined reforming uses less oxygen by combining partial oxidation with steam reforming in several configurations. This balances the heat required by the endothermic steam reforming reaction against the heat released by the exothermic partial oxidation reaction, although high temperatures are still required.

Research is under way on membranes that separate the oxygen, particularly on certain ceramic oxygen transport membranes. This would eliminate the need for expensive cryogenic oxygen plants and make the reforming unit very compact.

Ammonia synthesis takes place on a supported iron catalyst. The overall reaction is shown as reaction 4 in Fig. 12. This is also an equilibrium limited reaction that is favored by low temperatures and very high pressures, taking place at approximately 400°C and 40 MPa. A process diagram for ammonia synthesis is presented in Fig. 13. The feed gas is compressed and preheated before flowing to the top of the converter. The reaction is exothermic and the gas temperature increases as the reaction proceeds, so the catalyst beds are separated into two or three distinct zones. Between each zone, cool hydrogen gas is injected to mix with and cool the reacting gas so as to allow a higher conversion. The conversion remains relatively low, and the product gas is cooled to below the ammonia dew point. The product gas then passes through the separator where liquid ammonia is recovered, and the remaining gas is recycled to the reactor feed.

Methanol synthesis takes place over a Cu/ZnO catalyst at a pressure of approximately 5 MPa and a temperature of 200 to 250°C. The reaction, shown as reaction 5 in Fig. 12, requires an H_2/CO ratio of 2, but this may be higher in practice. A process diagram is shown in Fig. 14. Feed gas is compressed and flows to a multiple-bed catalytic reactor in similar fashion to that in ammonia synthesis. This reaction is also equilibrium limited, and cool feed gas is added between catalyst zones to allow higher conversion. As with ammonia synthesis, methanol synthesis achieves a relatively low conversion of the synthesis gas feed, the product gas is cooled to condense the methanol and water reaction products along with

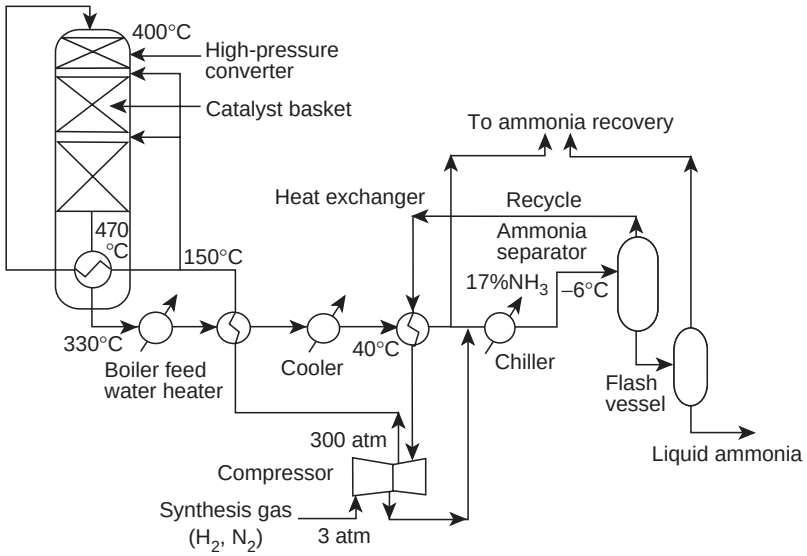


FIGURE 13 Ammonia synthesis process diagram.

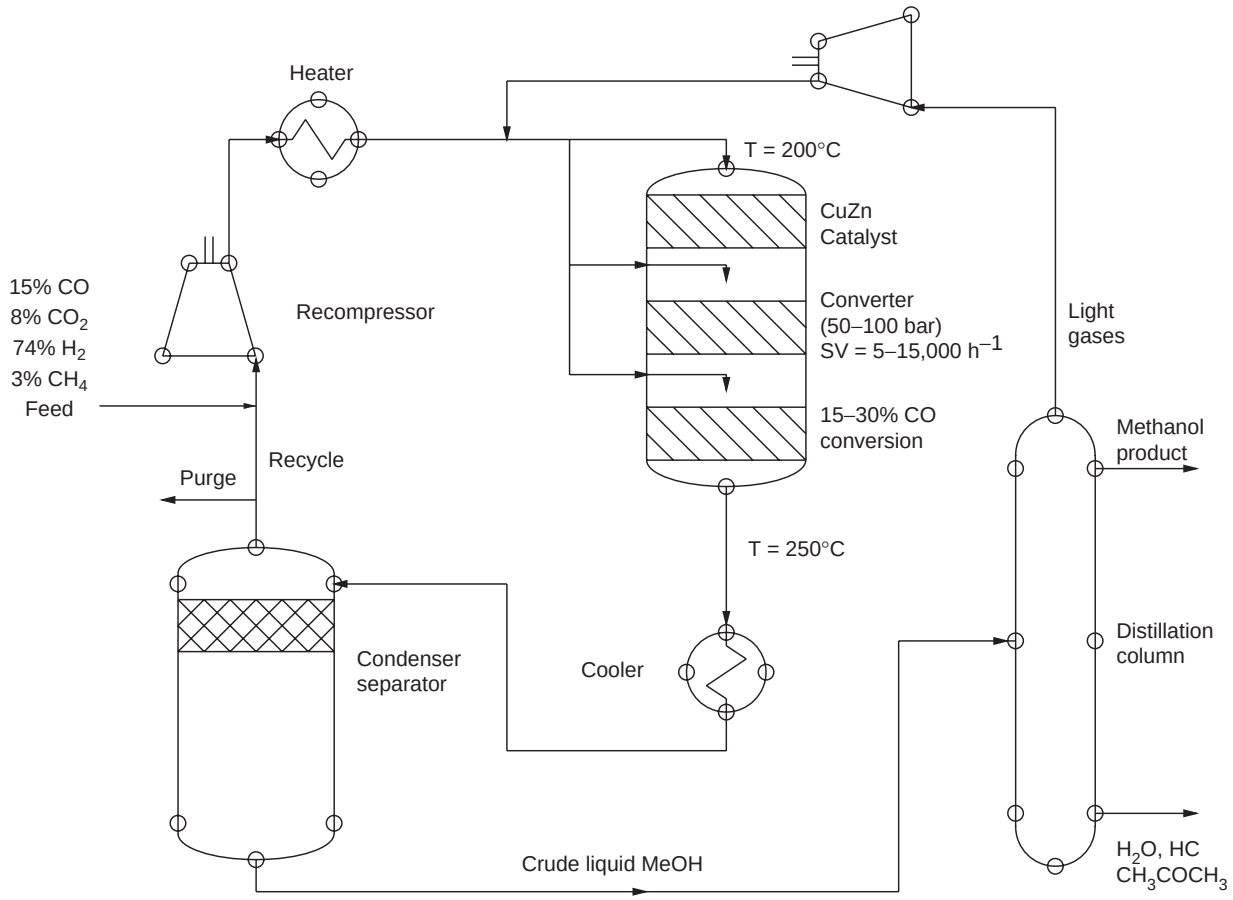


FIGURE 14 Methanol synthesis process diagram.

small amounts of by-products, and the gas is recycled to the feed. The crude product is distilled to produce various purity grades of methanol.

FTS has a complex chemistry that is simplified in reaction 6 in Fig. 12. The primary objective of large-scale Fischer–Tropsch gas-to-liquids (GTL) processes is to produce liquid hydrocarbons suitable for fuel use, and this is represented in reaction 6 by the production of the hydrocarbon-repeating unit $-CH_2-$. Under some conditions, such as high temperatures, oxygenates and olefins may be produced for specialty chemicals use. The reaction produces a wide range of molecular weight products that are predominantly straight chain alkanes. The product distribution produces some light gases (e.g., methane, ethane, propane) that are undesirable, some light liquid alkanes generally referred to as naphtha (a desirable feedstock for production of olefins), longer chain alkanes that are a suitable for diesel fuel, and even longer chain molecules that are solids at room temperature called waxes that may be converted into diesel fuel and other useful, but smaller market, products. The relative amounts of these may be adjusted by altering the catalyst and reaction conditions, but only within certain ranges. These are produced under so-called low-temperature synthesis conditions using supported cobalt- or iron-based catalysts. There are trade-offs between the two catalysts, with cobalt appearing to be more desirable despite its higher cost due to generally higher reaction rates and the fact that iron also catalyzes the water–gas shift reaction that undesirably consumes CO. A process diagram is shown in Fig. 15, which includes a complete process from natural gas to liquid hydrocarbons. The feed natural gas and oxygen both flow to the bottom of the fluidized bed synthesis gas (FBSG) reactor that uses combined reforming. The synthesis gas stream then flows to the hydrocarbon synthesis (HCS) reactor. The reactor operates with a catalyst suspended as a slurry in the liquid product with the gas feed bubbling up through the reactor. The reaction is exothermic, so heat is removed from the reactor with internal piping through which water is pumped and converted to

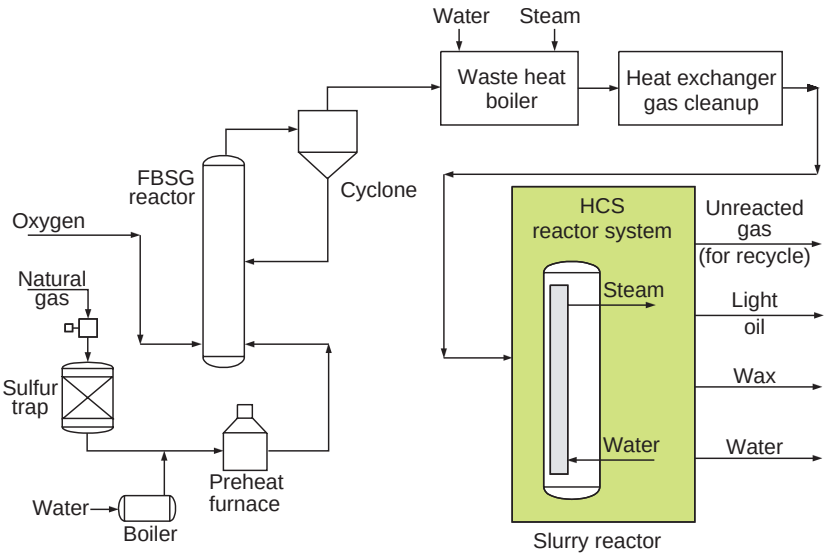


FIGURE 15 Example of a Fisher–Tropsch hydrocarbon synthesis process diagram. FBSG, fluidized bed synthesis gas reactor; HCS, hydrocarbon synthesis reactor.

steam. This maintains the desired reaction temperature of 200 to 250°C at the operating pressure of approximately 2.5 to 5.0 MPa. Relatively high conversions are achieved so that unreacted gas (and produced light hydrocarbons) is not recycled; rather, it is burned and its heat is recovered. The product liquid with catalyst flows out of the reactor, where the catalyst is separated and returned to the reactor suspended in a portion of the product hydrocarbon liquid. The water produced by the reaction is removed as a separate liquid phase, and the hydrocarbon liquid product is further refined to make final products ready for transportation by liquid pipelines or tankers.

SEE ALSO THE FOLLOWING ARTICLES

Markets for Natural Gas • Natural Gas, History of • Natural Gas Industry, Energy Policy in • Natural Gas Resources, Global Distribution of • Natural Gas Resources, Unconventional • Natural Gas Transportation and Storage • Occupational Health Risks in Crude Oil and Natural Gas Extraction • Oil and Natural Gas Drilling

Further Reading

Campbell, J. M. (1984). "Gas Conditioning and Processing," vols. 1–2, 6th ed. Campbell Petroleum Series, Norman, OK.

Farrauto, R. J., and Bartholomew, C. H. (1997). "Fundamentals of Industrial Catalytic Processes." Blackie Academic and Professional, London.

Gas Processors and Suppliers Association. (1998). "GPSA Engineering Data Book," vols. 1–2, 11th ed. GPSA, Tulsa, OK.

Kohl, A. L., and Riesenfeld, F. C. (1985). "Gas Purification.." 4th ed. Gulf Publishing, Houston, TX.

Rojey, A., Jaffret, C., Cornot-Gandolphe, S., Durand, B., Julian, S., and Valais, M. (1997). "Natural Gas Production Processing Transport." Editions Technip, Paris.

Satterfield, C. N. (1991). "Heterogeneous Catalysis in Industrial Practice." 2nd ed. McGraw–Hill, New York.

Twigg, M. V. (1989). "Catalyst Handbook." 2nd ed. Wolfe Publishing, London.

Wender, I. (1996). Reactions of synthesis gas. *Fuel Process. Technol.* **48**, 189.

Natural Gas Resources, Global Distribution of

RONALD R. CHARPENTIER
United States Geological Survey
Denver, Colorado, United States

1. Introduction
2. Worldwide Distribution
3. Impact of Distribution
4. Data Sources

Glossary

associated/dissolved gas The natural gas that is in or from oil fields, i.e., fields with a gas/oil volume ratio of less than 20,000 cubic feet per barrel. The natural gas may be in separate pools within the oil field, floating on the oil as a gas cap, or dissolved in the oil.

cumulative production The total volumes of natural gas that have been produced.

endowment The sum of cumulative production, reserves, reserve growth, and undiscovered resources.

future resources The sum of reserves, reserve growth, and undiscovered resources.

liquefied natural gas (LNG) The liquid form of gas; can be stored under low temperature and high pressure for transportation by ship.

nonassociated gas The natural gas that is in or from gas fields, i.e., fields with a gas/oil volume ratio of at least 20,000 cubic feet per barrel.

reserve (or field) growth The increases of estimated gas volume that commonly occur as oil and gas fields are developed and produced.

reserves The estimated quantities of natural gas, expected to be commercially recovered from known accumulations relative to a specified date, under prevailing economic conditions, operating practices, and government regulations. Reserves are part of the identified (discovered) resources and include only recoverable materials.

resource A concentration of naturally occurring natural gas in or on the earth's crust, some of which is currently or potentially economically extractable.

stranded gas Discovered natural gas that is partially or completely isolated from markets because of distances or insufficient transportation infrastructure, and thus is undeveloped or underdeveloped.

undiscovered resources Those resources postulated from geologic information and theory to exist outside of known oil and gas fields.

Natural gas is a mixture of hydrocarbon compounds that are in a gaseous state at standard conditions of temperature and pressure. Natural gas is very abundant worldwide. The greatest concentrations of conventional natural gas are in the Middle East and in the countries of the former Soviet Union. Because natural gas burns more cleanly and produces less greenhouse emissions compared to coal or oil, it is likely to provide a larger portion of the energy supply in the future.

1. INTRODUCTION

Some natural gas is found in fields that produce little or no volume of liquids. Such gas is called nonassociated gas; the fields that produce nonassociated gas have a gas/oil volume ratio of at least 20,000 cubic feet per barrel. Natural gas is also found in oil fields (fields with a gas/oil volume ratio of less than 20,000 cubic feet per barrel). In an oil field, natural gas can occur in separate pools or as a gas cap floating on the denser oil, in which cases it is termed "associated" gas. Oil field gas can also be dissolved in the oil, in which case it is termed "dissolved" gas.

1.1 Resource Classification

In the following discussion, conventional natural gas resources are classified in four categories: cumulative

production, reserves, reserve growth, and undiscovered resources. Cumulative production is that volume of natural gas that has already been produced. Reserves are those volumes remaining in already discovered oil and gas fields that are known to reasonable certainty. Reserve (or field) growth includes those volumes in already discovered oil and gas fields that are not known with certainty, but can be expected to be added to reserves in the future. Undiscovered resources are those volumes of natural gas in conventional oil and gas fields that have not yet been discovered. These volumes of gas are not static; as gas is discovered or produced, volumes shift from one category to another (e.g., from undiscovered resources to reserves, or from reserves to cumulative production). The sum of these four categories is called the endowment of natural gas. The future resources are the sum of the reserves, the reserve growth, and the undiscovered resources—that is, the endowment minus what has been already produced.

Many terms are used to identify categories and subcategories of resources. These terms are not always used consistently and must be approached with caution. The discussion here deals with conventional resources of natural gas, that is, those resources that are in discrete fields (usually with distinct gas/water or oil/water contacts). The bulk of historical natural gas production has been from conventional reservoirs. Unconventional natural gas resources, such as coal bed gas, tight gas sands, fractured shale gas, and gas hydrates, are commonly classified separately. They are treated here as unconventional gases.

1.2 Uncertainty of Estimates

Even more so than for oil, estimates of natural gas resources are subject to considerable uncertainty. This is true for all categories, including estimates of past production. Recorded gas volumes sometimes include total gas volume and at other times do not include that portion of natural gas that is noncombustible (nitrogen, CO_2, etc.). Generally, noncombustible components are less than 15% of gas volumes, but locally can be much greater. Throughout this article, total gas volumes are given (i.e., the noncombustible volume is included). Past production, as reported by the appropriate government agencies, may or may not include amounts of gas that were flared or reinjected into the reservoir to sustain pressure for additional oil recovery.

The amount of reserve growth in discovered fields is not well known, but is probably very significant. Especially in the case of stranded gas (discovered natural gas that is partially or completely isolated from markets because of distances or insufficient transportation infrastructure), there is little incentive to invest in the costs to evaluate fields fully. Although most studies of reserve growth have been based on more easily accessible data from the United States, studies of reserve growth outside the United States show that such growth occurs worldwide. Gas reserve growth outside the United States reported from 1977 to 1991 was at a rate greater than was growth of gas fields in the United States during that period.

2. WORLDWIDE DISTRIBUTION

A majority of the data used in this article on the worldwide distribution of conventional natural gas resources are from the U.S. Geological Survey (USGS) and the U.S. Energy Information Administration. The USGS provides probabilistic estimates of reserve growth and undiscovered resources, but only the mean estimates are cited here. The original USGS estimation of reserve growth was at the world level; for the analysis here, estimates of reserve growth by region were generated by allocating the USGS world reserve growth estimates proportional to the volumes of reserves in each region.

Table I and Fig. 1 show how conventional natural gas resources of the world are distributed into the eight regions used by the USGS. Note that 62% of the resource endowment is concentrated in just two regions—(1) the former Soviet Union and (2) the Middle East and North Africa. Also note that only in the North American region has a large proportion of the resource already been produced. The map in Fig. 2 shows how the resources that remain available are concentrated primarily in a part of the world that stretches from the Middle East, through the Caspian Sea, to West Siberia. Table II presents the natural gas production and consumption by region for the year 2000. The major trade movements of natural gas are shown in Fig. 3. These eight regions are discussed herein in decreasing order of their endowment; Tables I and II and Figs 1–3 should be referred to throughout the discussions.

The world distribution of unconventional gas resources is not well understood. Most of the exploration and production of unconventional resources has been in the United States. Geologic studies show that similar unconventional accumulations

TABLE I

World Conventional Natural Gas Resources by Region[a]

Region[b]	Trillion cubic feet						World cumulative production (%)	World reserves (%)	World reserve growth (%)	World undiscovered resources (%)	World future resources (%)	World endowment (%)
	Cumulative production	Reserves	Allocated reserve growth	Undiscovered resources	Future resources	Endowment						
1. Former Soviet Union	389.20	1682.95	1203.79	1611.26	4498.01	4887.20	22.21	35.12	32.89	31.01	32.95	31.73
2. Middle East and North Africa	73.49	1835.57	1312.96	1369.93	4518.47	4591.96	4.19	38.30	35.87	26.36	33.10	29.82
3. Asia Pacific	56.50	344.52	246.43	379.34	970.29	1026.79	3.22	7.19	6.73	7.30	7.11	6.67
4. Europe	204.30	279.71	200.08	312.37	792.15	996.46	11.66	5.84	5.47	6.01	5.80	6.47
5. North America	938.60	256.51	415.45	681.50	1353.46	2292.06	53.56	5.35	11.35	13.11	9.92	14.88
6. Central and South America	55.26	223.61	159.95	487.19	870.75	926.01	3.15	4.67	4.37	9.38	6.38	6.01
7. Sub-Saharan Africa and Antarctica	7.43	106.57	76.23	235.29	418.09	425.52	0.42	2.22	2.08	4.53	3.06	2.76
8. South Asia	27.59	63.07	45.11	119.61	227.79	255.38	1.57	1.32	1.23	2.30	1.67	1.66
Total	1752.37	4792.52	3660.00	5196.49	13649.01	15401.38	100.00	100.00	100.00	100.00	100.00	100.00

[a] All estimates reported here are mean estimates. Gas resource totals from U.S. Geological Survey. Allocated reserve growth from U.S. Geological Survey estimates, allocated (for non-U.S.) proportional to volumes of remaining reserves.

[b] Numbers correspond to maps in Figs. 2 and 3.

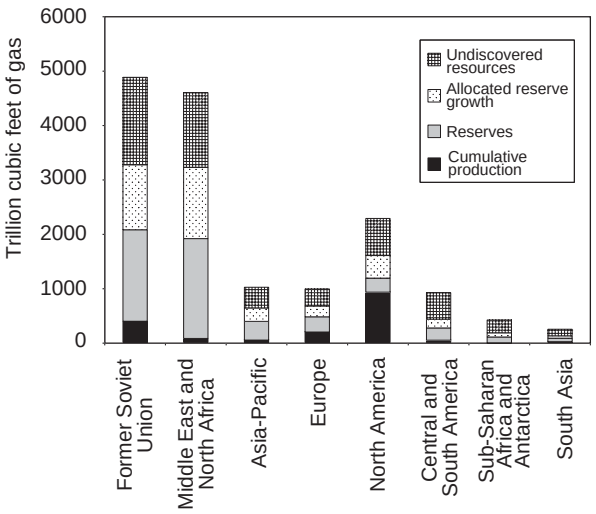


FIGURE 1 Conventional natural gas endowment (cumulative production, reserves, allocated reserve growth, and undiscovered resources) by region (see data in Table I).

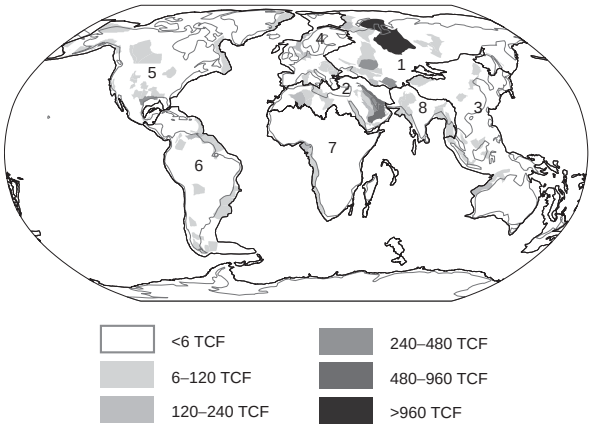


FIGURE 2 Natural gas of the world by geologic province as assessed in the *U.S. Geological Survey World Petroleum Assessment 2000*. The volumes are a sum of cumulative production, reserves, and undiscovered resources. Reserve growth was not included in the volumes. The shaded regions indicate conventional gas in trillions of cubic feet (TCF). Region key: (1) The former Soviet Union, (2) Middle East and North Africa, (3) Asia–Pacific, (4) Europe, (5) North America, (6) Central and South America, (7) sub-Saharan Africa and Antarctica, and (8) South Asia.

probably occur in many places worldwide, but quantitative estimates are not available.

2.1 Former Soviet Union

The former Soviet Union (FSU) has 33% of the world's future natural gas resources. In 2000, this region accounted for 29% of the world's natural gas production but had only 24% of the natural gas demand; thus, even though a major user of natural gas, the FSU is also a major exporter. Almost all of the export volume is to Europe through pipeline systems.

Of the 25.43 trillion cubic feet (10^{12} cubic feet, TCF) of production in 2000, 20.63 TCF (81%) was from Russia, mostly from the West Siberian Basin. The West Siberian Basin also has 63% of the FSU's natural gas reserves and 40% of the estimated undiscovered gas. The area of the Caspian Sea also contains very large estimated volumes of natural gas resources, accounting for 12% of the reserves and 21% of the estimated undiscovered gas. The Amu-Darya Basin of central Asia (Turkmenistan, Uzbekistan, Afghanistan, and Iran) has large reserves (9% of the FSU total) and estimated undiscovered gas (10% of the FSU total), but the gas is far from demand centers, so much of the gas is stranded. The Barents Sea has had few previous discoveries, but has significant estimated undiscovered gas (15% of the FSU total).

2.2 Middle East and North Africa

The Middle East and North Africa region also has 33% of the world's estimated future natural gas resources. This includes 38% of the world's reserves and 26% of the world's estimated undiscovered gas. In 2000, this region accounted for 13% of world production, but only 10% of the demand. The region is a major exporter to Europe through trans-Mediterranean pipelines and by transportation of liquefied natural gas (LNG). Much of the gas in the Arabian–Persian Gulf area is stranded, but some is exported to Japan and South Korea as LNG. The Arabian–Persian Gulf area accounts for 87% of the region's reserves and 90% of the estimated undiscovered gas. It includes the world's largest gas field, North Field (over 350 trillion cubic feet of natural gas), in Qatar. Algeria exports a significant volume of natural gas to Europe (69% of the region's exports) by both pipeline and as LNG.

2.3 North America

North America has 10% of the world's estimated future natural gas resources. Of all eight regions listed in Table II, North America is both the largest producer and the largest consumer of natural gas. Over half of the natural gas ever produced in the world has been from this region. It is the only region for which unconventional natural gas has been a significant portion of the production.

The United States is the largest producer (71% of North America's 2000 production) and the largest

TABLE II

World Natural Gas Production and Consumption in the Year 2000[a]

Region[b]	Trillion cubic feet		World production (%)	World consumption (%)	Surplus or deficit (%)
	Gas production	Gas consumption			
1. Former Soviet Union	25.43	20.54	28.89	23.51	+5.38
2. Middle East and North Africa	11.51	8.87	13.07	10.15	+2.92
3. Asia Pacific	7.29	8.20	8.28	9.38	−1.10
4. Europe	10.91	16.85	12.39	19.29	−6.90
5. North America	26.79	27.21	30.43	31.14	−0.72
6. Central and South America	3.43	3.30	3.89	3.78	+0.11
7. Sub-Saharan Africa and Antarctica	0.57	0.34	0.64	0.39	+0.25
8. South Asia	2.11	2.06	2.40	2.36	+0.04
Total	88.03	87.38	100.00	100.00	

[a] Data from Energy Information Administration (2001).
[b] Numbers correspond to maps in Figs. 2 and 3.

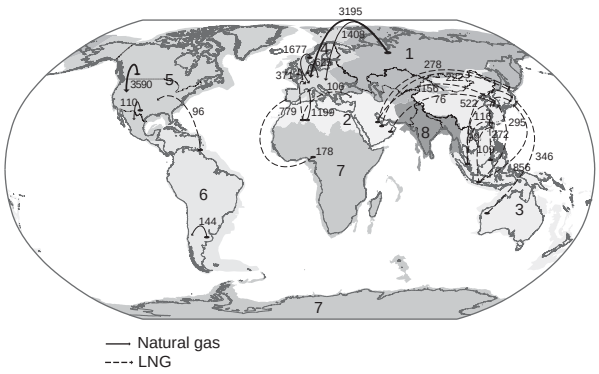


FIGURE 3 Major natural gas trade movements in billions of cubic feet (small numbers) for the year 2000. Region key (large numbers): (1) The former Soviet Union, (2) Middle East and North Africa, (3) Asia–Pacific, (4) Europe, (5) North America, (6) Central and South America, (7) sub-Saharan Africa and Antarctica, and (8) South Asia. Modified from British Petroleum Company (2002).

consumer (83% of North America's 2000 consumption). About 60% of U.S. production is from the onshore and offshore areas of the Gulf of Mexico. Almost all of the gas imported into the United States (94% of U.S. imports) comes from Canada by pipeline. Mexico imports only 3% of its demand, which is supplied by the United States.

2.4 Europe

Europe has 6% of the world's future natural gas resources. It is a net importer of natural gas, having consumed almost 6 trillion cubic feet of gas more than it produced in 2000 (35% of its consumption). Of the gas imported into Europe in 2000, 66% was from the former Soviet Union and 30% was from Algeria. Significant resources of European natural gas occur in the North Sea and adjoining areas. The largest producers in the North Sea area are the United Kingdom (35% of Europe's 2000 production), The Netherlands (24%), Norway (17%), and Germany (7%). Of these, Norway, The Netherlands, and the United Kingdom are net exporters. Almost all other European countries are net importers. Gas production in other parts of Europe is primarily from the Po Valley of northern Italy and Romania.

2.5 Asia–Pacific

The Asia–Pacific region has 7% of the world's future natural gas resources. The region accounts for 7% of world natural gas reserves, 7% of the estimated undiscovered gas, and 9% of the demand. Japan and South Korea, with little production, import LNG from Indonesia, Malaysia, Australia, and Brunei, and also from the Arabian–Persian Gulf. Areas of significant production are Malaysia and Indonesia, with 53% of the region's 2000 production. They have 42% of the reserves and 42% of the estimated undiscovered gas in the region. Northwest Australia also has significant resources—25% of the region's reserves and 29% of the region's estimated undiscovered gas—but some of the reserves are stranded. China has

significant resources also—13% of the region's reserves and 23% of the estimated undiscovered gas.

2.6 Central and South America

Central and South America have about 6% of the world's future natural gas resources. There is little interregional import or export, except for some LNG exported from Trinidad and Tobago to the United States. Argentina and Venezuela are the major natural gas producers of the region, accounting for 39 and 28%, respectively, of the region's production in 2000. Venezuela has 60% of the region's gas reserves and 21% of the estimated undiscovered gas. Argentina has 11% of the region's reserves and 8% of the estimated undiscovered gas. Other significant producers include Trinidad and Tobago, Brazil, Colombia, and Bolivia. Although Brazil currently has only 3% of the region's reserves, it has 40% of the estimated undiscovered gas, most of it in basins along the South Atlantic margin. Most of the region's natural gas is consumed in the country of production, but Argentina, Trinidad and Tobago, and Bolivia are net exporters. Chile and Brazil are net importers of natural gas.

2.7 Sub-Saharan Africa and Antarctica

Sub-Saharan Africa has about 3% of the world's estimated future natural gas resources. The Niger Delta of Nigeria dominates natural gas statistics for sub-Saharan Africa. Most (78%) of the natural gas already produced in sub-Saharan Africa was from the Niger Delta. The Niger Delta also has 82% of the reserves. Nigeria is the only major exporter of natural gas to other regions, that by transportation of natural gas liquids (NGLs). Of its exported gas, most (90%) is to Europe and 10% is to the United States.

In the past decade, deepwater oil exploration off the Atlantic coast of Africa has been very successful; several giant oil fields have been discovered. Improved geologic understanding about the South Atlantic coastal basins suggests that the undiscovered gas may be more widespread than was obtained from past production—52% of the estimated undiscovered gas is in Nigeria and 18% is in (mostly offshore) Angola. With an expansion of deepwater infrastructure and technology, gas production from these areas may become more economically viable, both from oil fields (associated/dissolved gas) and from gas fields (nonassociated gas).

No natural gas resources have yet been discovered in Antarctica. Even if exploration were allowed, any resources found would be very unlikely to be economic in the next few decades.

2.8 South Asia

South Asia has only about 2% of the world's future natural gas resources. South Asia is not well connected to other parts of the world by gas transportation routes and the gas production and consumption are roughly in balance. The region includes two of the world's largest deltas, the Indus and the Ganges-Brahmaputra, both with significant undiscovered resources of natural gas. The largest producers are Pakistan (about 40% of the region's production), India (38%), and Bangladesh (16%). Each of these countries consumes what it produces, with minimal imports or exports of natural gas. In contrast, Burma exports almost half of the gas that it produces.

3. IMPACT OF DISTRIBUTION

Although, natural gas resources are abundant and widespread, they have been underutilized compared to oil resources. Only about 11% of the recoverable conventional gas resource has been produced. Unconventional natural gas is even more abundant, but much of it is not likely to be developed in the next few decades. Tables III and IV list the major producing and consuming countries of the world. Russia and the United States dominate both lists. Russia is a net exporter of natural gas and the United States is a net importer. Some of the major consumers (Germany, Japan, Italy, and France) are only minor producers. Some major producers are only minor consumers (Algeria, Norway, Turkmenistan, and Malaysia).

The geographic distribution of supply is different from that of demand, making transportation critical. Areas such as northern Alaska, northwest Australia, and central Asia have large volumes of discovered natural gas that are poorly connected to areas of gas demand. Transportation to areas of high demand requires the construction of either natural gas pipelines or of facilities to use high pressure and low temperature to liquefy natural gas so that it can be transported by ship as LNG. The gas stranded in these areas could also become more relevant to world supply with improvements in gas-to-liquid technology.

TABLE III

Top 20 Producers of Natural Gas in Year 2000[a]

Country	Production (trillion cubic feet)
Russia	20.63
United States	18.99
Canada	6.47
United Kingdom	3.83
Algeria	2.94
Netherlands	2.57
Indonesia	2.36
Iran	2.13
Uzbekistan	1.99
Norway	1.81
Saudi Arabia	1.76
Turkmenistan	1.64
Malaysia	1.50
United Arab Emirates	1.41
Mexico	1.33
Argentina	1.32
Australia	1.12
Qatar	1.03
Venezuela	0.96
China	0.96

[a] Data from U.S. Energy Information Administration (2001).

TABLE IV

Top 20 Consumers of Natural Gas in Year 2000[a]

Country	Production (trillion cubic feet)
United States	22.55
Russia	14.13
United Kingdom	3.38
Canada	3.29
Germany	3.10
Ukraine	2.78
Japan	2.75
Italy	2.49
Iran	2.22
Saudi Arabia	1.76
Netherlands	1.72
Uzbekistan	1.51
France	1.43
Mexico	1.38
Argentina	1.17
Indonesia	1.08
United Arab Emirates	0.97
Venezuela	0.96
China	0.96
Pakistan	0.86

[a] Data from U.S. Energy Information Administration (2001).

4. DATA SOURCES

Information about natural gas resources is not static. Fortunately, several sources of information are easily accessible. The data about discovered resources used in this article came primarily from the U.S. Energy Information Administration. Their Web site (www.eia.doe.gov) provides extensive data on resources and energy usage for all energy sources (not just oil and natural gas) for the United States and the rest of the world. The data for reserve growth and undiscovered resources came primarily from the U.S. Geological Survey's *World Petroleum Assessment 2000*. The USGS has a Web site (energy.usgs.gov) where that assessment and additional extensive data for the United States and the rest of the world may be accessed.

Other sources of information can be found in several petroleum industry magazines (such as *Oil and Gas Journal* and *World Oil*) that publish annual statistical summaries. The British Petroleum Company (BP) also produces an annual statistical summary and has a Web site (www.bp.com) where extensive data is available.

SEE ALSO THE FOLLOWING ARTICLES

Depletion and Valuation of Energy Resources • Markets for Natural Gas • Natural Gas, History of • Natural Gas Industry, Energy Policy in • Natural Gas Processing and Products • Natural Gas Resources, Unconventional • Natural Gas Transportation and Storage • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Oil and Natural Gas Resource Assessment: Classifications and Terminology • Oil and Natural Gas Resource Assessment: Geological Methods • Oil and Natural Gas Resource Assessment: Production Growth Cycle Models

Further Reading

British Petroleum Company (BP). (2002). "BP Statistical Review of World Energy June 2002." BP, London. Also available on the Internet at http://www.bp.com.

DeGolyer and MacNaughton Consultants (2000). "Twentieth Century Petroleum Statistics," 56th ed. DeGolyer and MacNaughton, Dallas, Texas.

Energy Information Administration (EIA) (2001). "International Energy Outlook 2001." U.S. Department of Energy,

Washington, D.C. Also available on the Internet at http://www.eia.doe.gov.
International Energy Agency (IEA). (2000). "World Energy Outlook 2000." Organisation for Economic Co-Operation and Development, Paris.
PennWell Corp. (2000). "International Petroleum Encyclopedia." PennWell, Tulsa, Oklahoma.
U.S. Geological Survey (USGS) World Energy Assessment Team (2000). "U.S. Geological Survey World Petroleum Assessment 2000—Description and Results." U.S. Geological Survey Digital Data Series DDS-60 (four CD-ROMs). USGS, Reston, Virginia. Also available on the Internet at http://energy.usgs.gov/.

Natural Gas Resources, Unconventional

VELLO A. KUUSKRAA
Advanced Resources International, Inc.
Arlington, Virginia, United States

1. Background
2. Definitions for Unconventional Gas
3. Resource Estimates for Unconventional Gas
4. Unconventional Gas Development in the United States
5. Conclusion

Glossary

coalbed methane Natural gas stored within a coal seam (coal reservoir) requiring dewatering and/or lowering of pressure for its release and production.

gas hydrates Natural gas stored as a chlahrate within the molecular structure of water requiring the injection of heat or the lowering of pressure for dissociation and gas flow.

gas shales Natural gas stored within an organically rich shale interval (reservoir) requiring natural fractures and artificial well stimulation for acceptable rates of gas flow.

geopressured methane Natural gas stored as dissolved gas in highly pressured (geopressured) saline brines requiring the lowering of pressure for gas release and flow.

tight gas sands Natural gas stored within a low-permeability (tight) reservoir requiring natural and/or induced well stimulation for acceptable rates of gas flow.

unconventional gas Natural gas stored in a variety of geologically complex (unconventional) reservoirs requiring advanced exploration and production technology for their commercial development.

Unconventional natural gas has become a major source of energy supply in the United States. A large, favorable resource base, advances in knowledge, and steady progress in extraction technology have enabled significant portions of this resource to be converted to producible reserves. However, worldwide development and assessments of unconventional gas are still limited, concentrated primarily on coalbed methane and selected tight gas plays. This article provides a status report on our understanding of the size and nature of worldwide unconventional gas resources. The intent is that the information and insights in this article will be of value in accelerating the development of these large, complex natural gas resources.

1. BACKGROUND

For some time, geologists and resource analysts have known that large volumes of natural gas are locked in geologically complex, unconventional reservoirs, such as tight gas sands, gas shales, and coalbed methane (CBM). These unconventional gas reservoirs exist in the United States, Canada, and many other areas of the world. In the past, the natural gas resources stored in these geologically complex reservoirs have been judged as being too costly and technically difficult to recover to be a viable source for future natural gas supplies.

With advances in technology and understanding, the status and outlook for unconventional gas resources has become much more positive. Today, in the United States, unconventional gas provides 5.4 trillion ft^3 (Tcf) annually and accounts for 27% of domestic natural gas production, with each of the three unconventional gas sources—tight gas sands, coalbed methane, and gas shales—making a significant contribution, as shown in Fig. 1. The importance of unconventional gas is underscored by the recognition that 8 of the largest 12 U.S. natural gas fields exclusively produce unconventional gas (see Table I). Most significant, as set forth in the U.S. Department of Energy (DOE)/Energy Information Agency's Annual Energy Outlook 2003, unconventional gas is expected to become, by 2015, the largest single source of U.S. natural gas supply, surpassing

 © 2004 Elsevier Inc. All rights reserved.

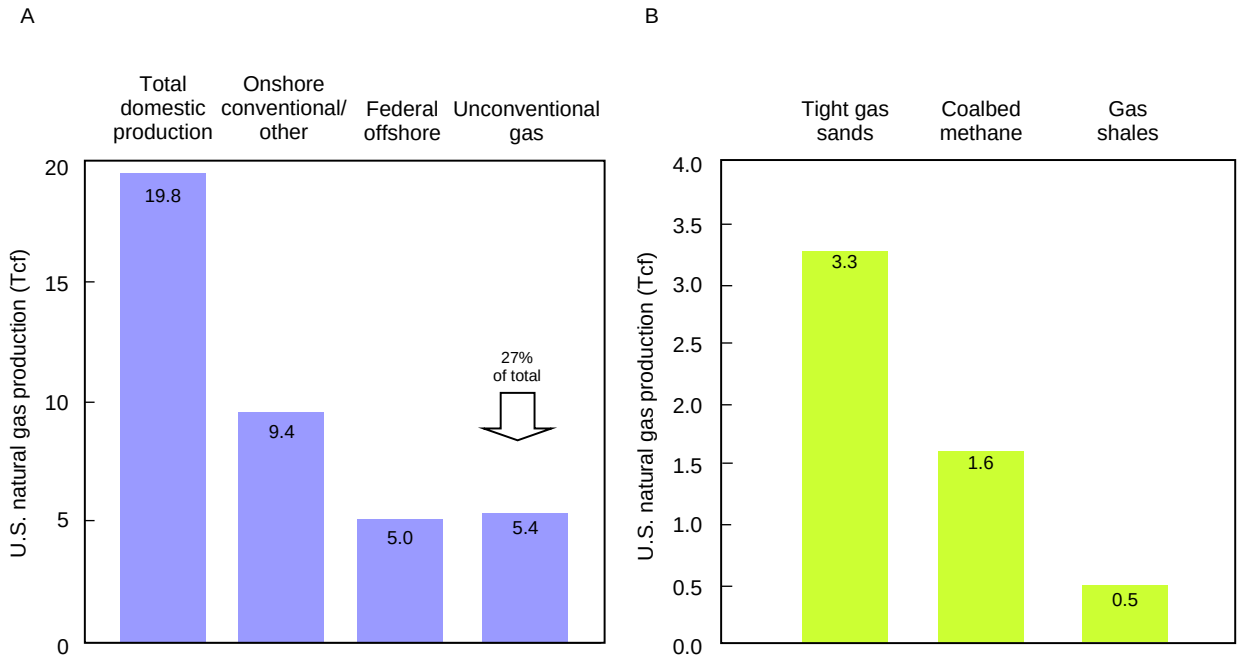


FIGURE 1 Sources of U.S. natural gas production in 2001 (Tcf). (A) From: Conventional/Offshore—Energy Information Agency Annual Reserve Report 2001. Unconventional—Advanced Resources International database. (B) From: Advanced Resources International Database.

TABLE I

Eight of the 12 Largest U.S. Natural Gas Fields Are Unconventional Gas Fields

Rank	Field name	Basin/State	Type of resource	Year 2001 production (billion cft^3)
1	Blanco/lgancio-Blanco	San Juan, NM/CO	CBM/Tight gas sands	812
2	Basin	San Juan, NW Powder River,	CBM/Tight gas sands	617
4	Wyodak/Big George	WY	CBM	224
5	Carthage	East Texas, TX	Tight gas sands	203
6	Jonah	GGRB, WY	Tight gas sands	176
8	Giddings	East Texas, TX	Tight gas sands/chalk	152
9	Wattenberg	Denver, CO	Tight gas sands	143
12	Newark East	Ft. Worth, TX	Gas shale	114

Source. Advanced Resources International, Energy Information Agency 2001 Annual Reserve Report.

production from onshore conventional gas fields and offshore sources (see Fig. 2).

Worldwide, the pursuit of unconventional gas is also growing in importance. Significant drilling for coalbed methane is under way in Canada's Western Canadian Sedimentary Basin; Australia is actively producing tight gas sands from the Cooper Basin and coalbed methane from the Bowen Basin, and China, India, and South Africa (among other countries) are actively exploring for new coalbed methane plays.

Still, little is known about the ultimate size, nature, and producibility of these vast natural gas resources. The purpose of this article is to expand this base of knowledge.

2. *DEFINITIONS FOR UNCONVENTIONAL GAS*

Unconventional natural gas is defined as conventional natural gas (primarily methane) trapped in

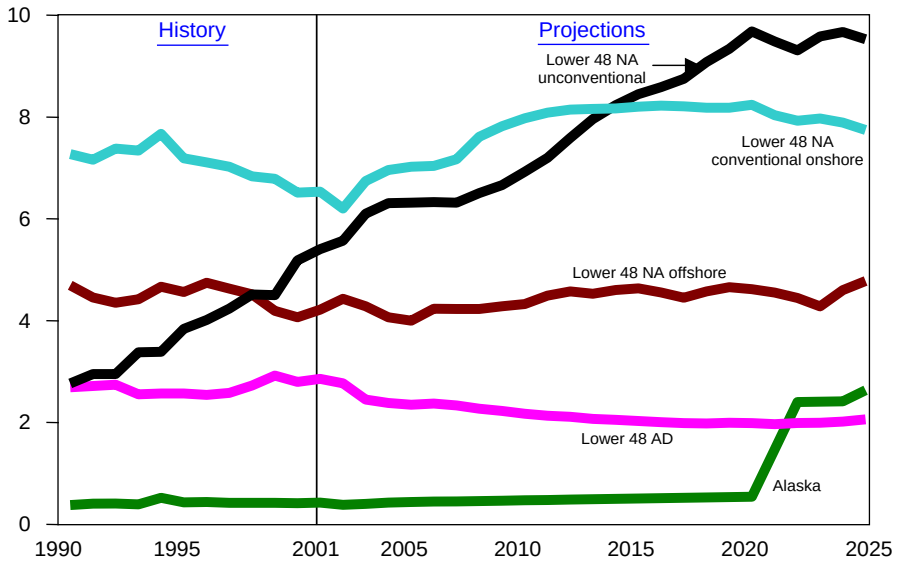


FIGURE 2 U.S. natural gas production, 1990–2025 (Tcf). *Source*: Advanced Resources International, Energy Information Agency (2003).

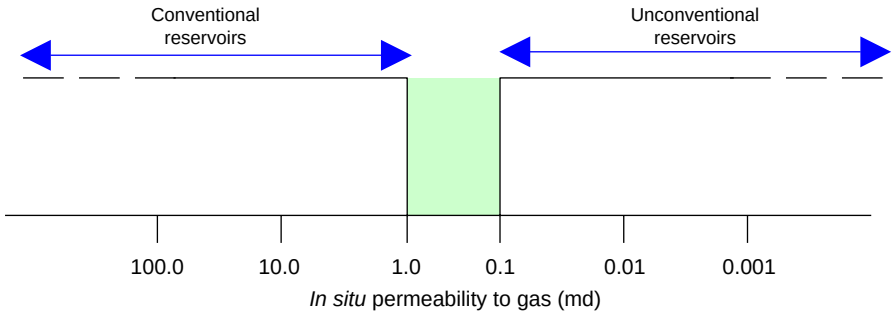


FIGURE 3 Permeability continuum for conventional through unconventional gas reservoirs.

geologically complex, nonconventional reservoirs, such as tight (low-permeability) sands, gas-bearing shales, and coalbeds. In the longer term, more speculative resources, such as gas hydrates, ultradeep gas, and geopressured methane, may be included in this resource category.

2.1 Tight Gas Sands

Tight gas sands are low-permeability (tight) gas-bearing reservoirs that have an *in situ* matrix permeability to gas of less than 0.1 md, exclusive of natural fracture permeability. Figure 3 illustrates the permeability continuum and criteria for conventional and unconventional gas reservoirs. In contrast to conventional gas reservoirs where the gas is held in structural or stratigraphic traps, tight gas reservoirs are areally extensive, often called continuous-type accumulations. These reservoirs occur in nearly all petroleum provinces, are usually abnormally pressured, and are often found in basin-center settings. Although they are commonly called tight gas sands, low-permeability reservoirs are found in a variety of rock types, including sandstone, siltstone, carbonate, limestone, dolomite, and chalk.

The porosity of lenticular and marginal-marine tight gas sand reservoirs is low, generally less than 12%, and the reservoir pores have high capillary pressure as a result of postdepositional cementation of conventional reservoir rocks. A second type of tight reservoir, such as a chalk, has high porosity of 15 to 35% but low permeability to gas because the rock is very fine-grained.

Artificial well stimulation, typically hydraulic fracturing, is required to attain commercial rates of gas flow from tight gas sands, unless the well

encounters an extensive cluster of natural fractures that enhances permeability. Identifying these naturally enhanced areas of permeability, the so-called "sweet spots" in an otherwise tight accumulation, is a major research and exploration objective for the industry.

2.2 Gas Shales

Gas shales are self-sourcing reservoirs where a major portion of the resource is adsorbed on and within the organic matter in the shale. To be productive and have adequate flow paths, the gas shale reservoir must be naturally fractured. Because of these conditions, gas shales are often called organic-rich gas shales or fractured gas shales.

Gas shale accumulations tend to be regionally pervasive, continuous-type accumulations and are abnormally pressured (generally underpressured). The quality of gas shales is judged by the total amount of organic matter (organic-rich shales are those containing more than 2% organic matter) and its degree of thermal maturity. Vitrinite reflectance values of over 1.5% are preferred for identifying areas with potential for thermogenic gas. Also important is the type of organic matter within the shales, the thickness and depth of the reservoir, and particularly the regional and local natural fracture intensity of the gas shale accumulation. These factors, in combination, determine the gas-producing potential of a gas shale reservoir.

Because the per well reserves of gas shales are often modest, low well costs, large-scale development (for economies of scale), induced fracture stimulation, and natural fracture (sweet spot) exploration are important for commercial, sustained production of natural gas from gas shales.

2.3 Coalbed Methane

Large volumes of methane, along with water and other gases, are generated in the process of converting organic matter to coal. This in-place generated gas, as well as deeper gas that may have migrated into the coal reservoir, is adsorbed (stored) upon and within the molecular structure of the coal. As such, coalbeds are both a source and a reservoir for natural gas. The amount of gas stored in the coal is a function of the coal's rank, pressure, temperature, and degree of gas saturation. High coal ranks and pressures, in general, lead to higher gas storage capacity. The storage capacity is commonly determined by using a sorption (or gas content) isotherm, as shown in Fig. 4.

The permeability of the coal matrix is nonexistent for all practical purposes. Gas moves through the angstrom-size pores in the coal by diffusion, an extremely slow flow process in most cases. As such, the primary flow paths for methane in coal seams are provided by the coal's cleat system, as well as its natural fracture overprint.

Dewatering and subsequent depressurization of the coal reservoirs are used to release the gas sorbed in the coal. Artificial well stimulation, such as hydraulic fracturing or cavitation, is used to create

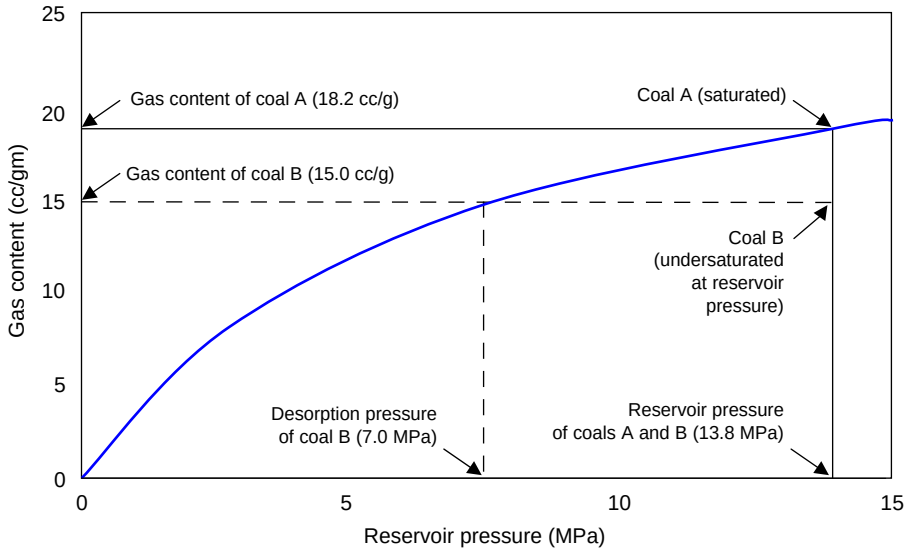


FIGURE 4 Idealized coalbed gas sorption isotherm showing relation between reservoir pressure and gas content for a saturated (coal A) and undersaturated (coal B) coal.

additional flowpaths to accelerate gas flow from the coal face and cleat systems to the wellbore.

As may be surmised from the above definitions for tight gas sands, gas shales, and coalbed methane, several features are common to unconventional gas resources:

1. First, the unconventional deposits may contain large, even massive volumes of gas in-place, often at low concentrations;
2. Second, despite the large volume of gas in-place, only limited areas will have sufficient richness and permeability for commercially developable reserves; and
3. Third, continued improvements in exploration and production technology, such as natural fracture detection, advanced reservoir diagnostics, and lower-cost well drilling and completion practices, are essential if these resources are to be effectively developed and converted to reserves.

Despite their widespread, global presence, the development of unconventional gas resources, particularly tight gas sands and gas shales, has been largely limited to the United States and much of the information in this article is based on this experience. The intent is that the information and insights in this article will be of value in developing the large, complex unconventional gas resources of the world.

3. RESOURCE ESTIMATES FOR UNCONVENTIONAL GAS

As more has been learned about the nature and location of unconventional gas, its resource estimates have increased and gained in confidence.

3.1 Initial Resource Estimates

The initial resource estimates for global unconventional gas were prepared by Kuuskraa and Meyers in the early 1980s, at a time when only limited information was available on the in-place and recoverable resource potential from tight gas formations, coalbed methane, fractured gas shales, or geopressured methane accumulations. Even less was known about methane hydrates or ultradeep gas.

As shown in Table II, the authors estimated that 570 Tcf of unconventional gas may be technically recoverable from the United States and Canada out of a gas in-place estimate of 2000 Tcf. The estimates for other areas of the world of 850 Tcf of technically recoverable resources were much more speculative and conservative and were based on analogues from the U.S. experience.

TABLE II

Technically Recoverable Resource Estimates for Unconventional Gas (Tcf)

	United States/Canada	Other areas	Total
Gas shales	50	250	300
Tight gas sands	460	400	860
CBM	60	200	260
Total	570	850	1420

Source. Kuuskraa and Meyers (1983).

3.2 Global Gas Resources Workshop

The second comprehensive set of resource estimates for unconventional gas was provided by the "Workshop on Global Gas Resources," convened in Vail, Colorado, in September 1994. The purpose of the workshop was to evaluate the worldwide gas resource base and the volume of resources that could be added with significant progress in technology. The workshop participants included presentations by 32 resource experts from 15 countries, representing international and domestic producers, state-owned energy companies, government organizations, universities, and research institutions.

The workshop participants concluded that advanced technology and unconventional resources could increase the 1993 U.S. Geological Survey (USGS) benchmark resource estimate for global natural gas by nearly 3000 Tcf, as shown in Table III and Fig. 5. Because unconventional resources in tight formations, gas shales, and coalbeds were not included in the 1993 USGS benchmark estimate (except for North America), the bulk of the 3000 Tcf of additional resource (producible with advanced technology to 2015) was estimated to be from unconventional gas.

Detailed estimates for unconventional gas were provided for the United States, Canada, and China by Fisher, Young and Drummond, and Miao, respectively. An overall review of global unconventional natural gas was provided by Kuuskraa. Of particular note were the closing remarks by Fisher on U.S. natural gas resources, setting forth long-term expectations for adding another 10,000 Tcf to the global gas resource base from unconventional gas and technology progress:

…the explicit incorporation of technology in resource estimates has doubled the U.S. estimate of ultimate recovery of natural gas…If the experience in the United States were extrapolated to the global resource base, the world reserves would exceed 20,000 Tcf, and they would be accessible at lower costs.

—Fisher, 1994

3.3 Resource Estimates

As the demand for natural gas has grown, considerably more attention has been placed on understanding the size, location, and producibility of unconventional gas.

TABLE III

Summary of Workshop Estimates for Natural Gas Resources (Tcf)

	USGS benchmark remaining resources,[a] 1993	Technology-enhanced resources, by 2015
North America	1394	2259
South America	715	987
Europe	617	783
Africa	813	1000
Middle East	2697	2799
Asia and Australasia	963	1733
Former Soviet Union	4361	4928
Total	11,560	14,489

[a] Total remaining resources are the sum of identified reserves and undiscovered gas (mean).

3.3.1 Rogner

Rogner, in 1997, published his latest assessment of global unconventional gas resources (see Table IV). Even after a considerable number of assessments, he noted that the data on unconventional gas is still limited, that estimates "are fraught with geological uncertainty," that "the technology implications for the eventual production of unconventional gas are poorly understood," and that "the data are speculative and should be read as such." Still, the information published by Rogner represented a significant step forward in quantifying both the gas in-place and the range of recoverable resources for unconventional gas. Rogner estimated the potential volume of unconventional gas in-place at over 35,000 Tcf globally.

Rogner estimated 10,000 Tcf of coalbed methane gas in-place, with over 3000 Tcf of gas in-place estimated for North and Central America. His regional distribution of coalbed methane was based on the geographical distribution of anthracite and bituminous coal deposits around the world, assuming average gas content values for these two coal types.

Rogner used the U.S. estimates for in-place natural gas from shale formations as an analogue to estimate 17,000 Tcf of global resources in gas shales, with

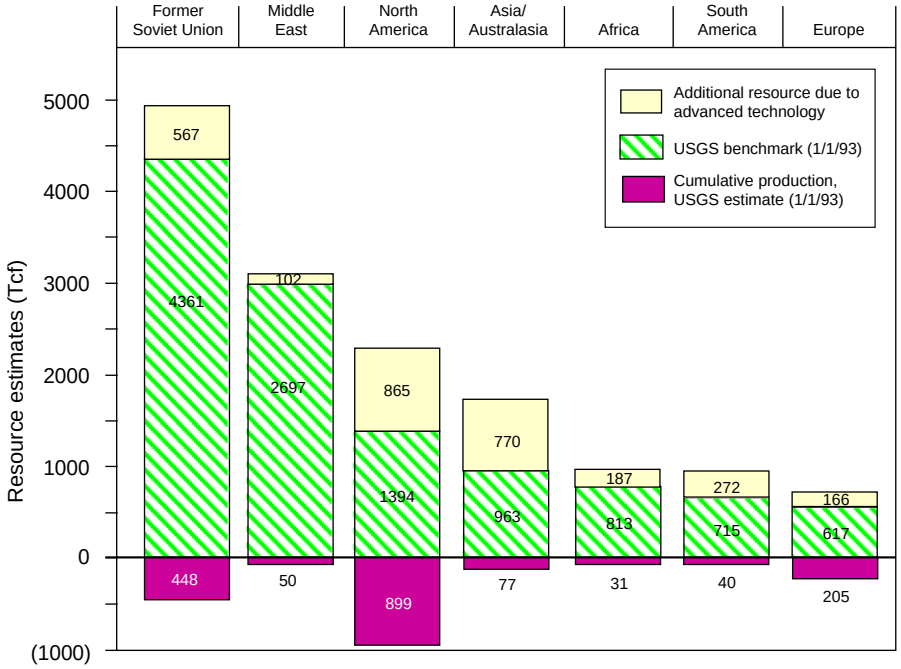


FIGURE 5 Global Gas Resources Workshop estimates for natural gas (Tcf).

TABLE IV

Estimates of Unconventional Natural Gas In-Place by Type (Tcf)

Region	Coalbed methane	Gas from fractured shales	Tight gas sands	Methane hydrates
North America	3270	4160	1480	260,000
Latin America and the Caribbean	40	2290	1400	190,000
Western Europe	170	550	380	30,000
Central and Eastern Europe	130	40	90	—
Former Soviet Union	4290	680	980	180,000
Middle East and North Africa	—	2750	890	10,000
Sub-Saharan Africa	40	300	850	20,000
Centrally planned Asia and China	1320	3820	380	20,000
Pacific OECD	510	2500	760	60,000
Other Pacific Asia	—	340	600	10,000
South Asia	40	—	210	20,000
World	9810	17,440	8020	800,000

over 4000 Tcf of gas in-place in North and Central America.

The 8000 Tcf global resource estimate and the regional distributions for tight gas sands were based on published data on conventional natural gas, with nearly 1500 Tcf estimated for North and Central America.

The global estimate for methane hydrate of 800,000 Tcf in-place was from an earlier estimate by MacDonald. A world map of locations of known and inferred gas hydrates in marine sediments and in continental permafrost was used to establish the regional distributions for this resource.

3.3.2 Nakicenovic

Nakicenovic, drawing on work by Rogner and others, published information on natural gas reserves and technically recoverable resources, including historical consumption of conventional and unconventional gas (see Table V). Reserves are assumed to be recoverable with available technologies at prevailing market conditions. Resources are technically recoverable volumes in addition to reserves, with less certain geological assurances, lacking economic feasibility, and requiring advances in technology for transforming these resources into reserves. The estimate of 33,200 Tcf of reserves and resources for unconventional gas was chosen to correspond to the highest plausible values from the literature. Nakicenovic postulated that potentially the resource base for unconventional gas could be twice as large as that for conventional gas and that the global natural gas resource base will be larger than that for oil.

3.3.3 Kuuskraa

In 1998, Kuuskraa published a country by country estimate for global coalbed methane, both for in-place and for technically recoverable resources (see Table VI). Kuuskraa estimated a CBM in-place value of 3000 to 8000 Tcf and a technically recoverable value of 510 Tcf. Kuuskraa's 1998 estimate for recoverable CBM resources is double his value published in the early 1980s. His most recent estimates were based on a combination of detailed CBM basin studies and information on the rank and depth of coal basins around the world. Future basins studies, particularly of the large CBM resources in Russia and the speculative CBM resources in Canada, the United States, and other countries, and advances in recovery technology would likely add considerably to these estimates of recoverable resources.

3.3.3.1 United States The United States has over 500 Tcf of high-quality coalbed methane resource (see Fig. 6). An additional 230 Tcf of CBM is in the deep coals of the Green River Basin and another 1000 Tcf of hypothetical CBM resource may exist in Alaska. Of this, 150 Tcf is estimated as being technically recoverable, including 29 Tcf of proved reserves plus past production. The estimates for technically recoverable CBM in the United States have steadily increased in the past 10 years and will likely continue to do so in future years. (Additional details on U.S. CBM development are provided in Section 4.)

TABLE V

Global Hydrocarbon Reserves and Resources

	Consumption		Remaining resource base		
	1860–1998	1998	Reserves	Resources	Total
Oil (billion barrels)					
Conventional	782	21	770	970	1940
Unconventional	47	2	970	2580	3550
Gas (Tcf)					
Conventional	2228	76	5700	10,400	16,100
Unconventional	28	N/A	8600	24,600	33,200

Sources. Nakicenovic *et al*. (1996, 1998); World Energy Council (1998); Masters *et al*. (1994); Rogner *et al*. (2000).

TABLE VI

International Coalbed Methane Resources (Tcf)

Country	CBM resource in-place (Tcf)	CBM recoverable resource (Tcf)
Russia	550–1550	N/A
China	350–1150	70
United States	500–1730	150
Australia	310–410	60
Canada	570–2280	140
Indonesia	210	30
Western Europe	120	N/A
Southern Africa[a]	100	20
India	90+	20
Poland/Czech Republic	70	N/A
Turkey[a]	50	10
Ukraine	50	N/A
Kazakhstan[a]	40	10
Total	3010–7840	510

[a] Assumes 20% recovery of gas in-place.

3.3.3.2 Australia Commercial-scale CBM projects exist in the Bowen and Gunnedah Basins, providing over 100 million ft^3 per day of production, with exploration under way in the Clarence-Moreton and Sydney Basins. The CBM resource in these four basins is estimated at 260 Tcf of gas in-place, plus 50 Tcf for the Galilee Basin, bringing the total to 310 Tcf. An additional 100 Tcf of gas-in-place may exist in the deep coals of the Cooper Basin. Approximately 60 Tcf of the CBM resource is estimated as recoverable, although considerable uncertainties exist on well productivities and economics. Given the substantial size of the resource, CBM production in Australia will likely be demand-limited. New gas pipelines to industrial gas users, such as to Mt. Isa, and the completion of a southeastern Australia gas grid are opening new markets, helping boost to both conventional gas and coalbed methane consumption.

3.3.3.3 Canada Canada has 570 Tcf of coalbed methane resources in-place, in formations that meet certain minimum reservoir parameters. This high-graded portion represents approximately one-quarter of the total Canadian CBM resource of 2280 Tcf. Using U.S. coal basins as the analogue, approximately 140 Tcf, in the higher quality reservoir settings, is judged to be potentially recoverable. The largest portion, 100 Tcf, is from the Cretaceous-age coals in the plains and foothills of Alberta. Approximately 35 Tcf of CBM is from British Columbia and 5 Tcf is potentially recoverable from Nova Scotia and Prince Edward Island.

3.3.3.4 China China has a large CBM resource base, estimated at 1150 Tcf, of which 350 Tcf is judged to be in areas geologically favorable for exploration. Of this, 150 Tcf is in eight main CBM exploration areas and perhaps 70 Tcf is technically recoverable. Much of the CBM resource, though located close to urban industrial markets in eastern China, occurs in structurally complex basins. The premier CBM area in China is the eastern margin of the Ordos Basin, where the structural setting is simpler and favorable permeability has been measured. Two new gas pipelines from the Ordos Basin to Beijing and Xian could accommodate 200 million cubic feet (MMcfd) of future CBM production. In addition, China is the world's largest coal producer and emits the highest volumes of mine-sourced coalbed methane into the atmosphere, an estimated

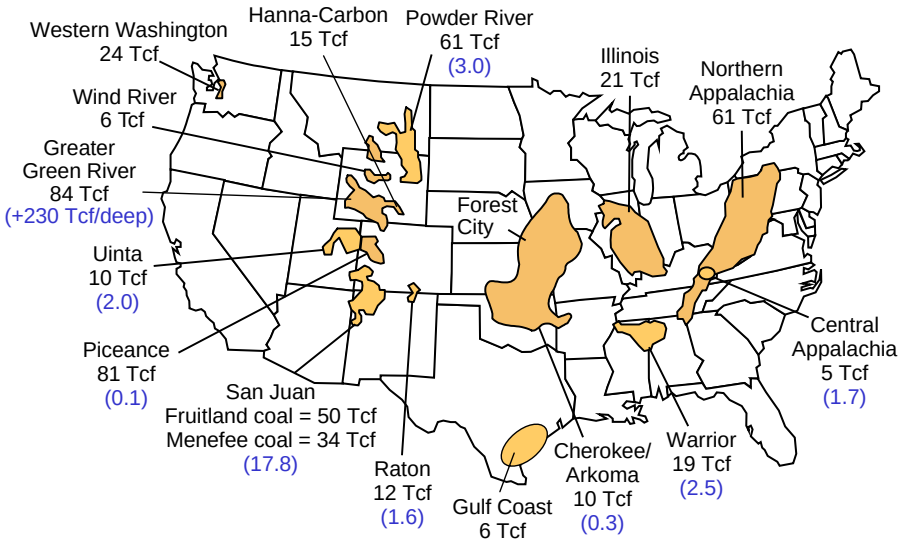


FIGURE 6 Coalbed methane resources of the United States. Values indicated (in Tcf) represent gas in-place; numbers in parentheses indicate resources that have been produced or are reserves (in Tcf) as of 2002; shaded areas indicate coal basins. *Source*: Advanced Resources Database.

2 billion cubic feet (Bcfd). Capture and use of this coal mine methane would provide a valuable source of clean energy and significantly reduce the emissions of a powerful greenhouse gas.

3.3.3.5 India India has 90 Tcf of coalbed methane resource in-place. Of this 44 Tcf is in six Damodar Valley coal fields, 40 Tcf is in the Cambay Basin of Gujarat, and 6 Tcf is in the Sohagpur and Satpura Basins. Additional CBM resources exist in other Gondwana and Tertiary coal basins of the Godavari and Mahanadi Valley that have yet to be appraised. An estimated 20 Tcf of CBM may be technically recoverable from the Damodar Valley and the Cambay Basin. Production of India's coalbed methane resource will be assisted by a strong natural gas market and plans for expanding its natural gas pipelines.

3.3.3.6 Indonesia A new entry on the world coalbed methane scene, Indonesia is estimated to have 210 Tcf of CBM resources, primarily in the Sumatra and East Kalimantan coal basins. Of this, approximately 30 Tcf is estimated as being in sufficiently high-quality settings to be potentially recoverable. Given its large conventional gas resources, uncertainties about markets, and lack of current activity, development of Indonesia's CBM may progress slowly.

3.3.3.7 Poland/Czech Republic The coalbed methane resources of Poland and the Czech Republic are estimated at 70 Tcf. This includes 47 Tcf in the Upper and Lower Silesian Basins of Poland, 20 Tcf in the Czech Republic, and 3 Tcf in the Lublin Coal Basin. Additional CBM resources may exist in the deep coals (below 1500 m) of Poland. Because of uncertainties with respect to permeability and gas saturation, no estimates of recoverable resources exist.

3.3.3.8 Russia Ultimately Russia may hold the largest volume of coalbed methane, though much of it is still poorly defined. The Kuznetsk Coal Basin (Kuzbass), on the southern slopes of the west-central Siberian plain, covers an area of 10,000 mi^2. The numerous coal seams in this basin extend to below 5000 ft in the basin center and hold an estimated 450 Tcf of CBM resource. The Pechora Basin, located on the Arctic Circle adjacent to the Barents Sea, holds an estimated 175 billion tons of high- to low-volatility bituminous coals and an estimated in-place CBM resource of 100 Tcf. Given the lack of data, no estimates are available on technically recoverable resources. The massive Tungusk Basin in the center of Siberian Russia that extends from the Arctic Circle to the Mongolian border may hold over 1000 Tcf and ultimately may become one of the largest CBM basins in the world.

3.3.3.9 Ukraine The Ukraine has 50 Tcf of coalbed methane in-place, primarily in the Donetsk Coal Basin. Given the high geologic complexity of the basin, the limited data, and the extensive past

mining in the basin, estimates of recoverable resources are still speculative.

3.3.3.10 Western Europe Western Europe is estimated to hold 120 Tcf of coalbed methane in place. The Carboniferous coal basins in Western Europe share many similarities: moderately gassy coals, concerns over natural degassing, numerous thin coal seams distributed over a large stratigraphic interval, and structurally complex settings. Coalbed methane has been tested in France, Germany, Great Britain, Belgium, the Netherlands, and Spain. Thus far, the results have been inconclusive, providing little basis for estimates of recoverable resources.

3.3.3.11 Other Counties Coalbed methane exists in numerous other countries, with still limited information on quality and producibility. Prominent are Turkey with 50 Tcf, Kazakstan with 40 Tcf, and Southern Africa with 90 Tcf of gas in-place. Approximately 20% of this in-place resource may be technically recoverable. South America (Brazil, Colombia, and Venezuela) has no publicly available assessments for CBM resources. Future exploration may add these and other countries to the world's productive CBM ledger.

3.4 Dynamic Resource Base

The review of unconventional gas resource estimates provides a valuable insight. As more is learned about the nature of these resources and the technologies by which to produce them, the resource estimates have changed dramatically.

- For geologically sound unconventional gas resources, such as tight gas sands, gas shales, and coalbed methane, the historical review shows that the recoverable resource and reserve estimates have steadily increased.
- For geologically uncertain resources, such as geopressured methane, little additional information has been obtained and the expectations for recoverable resources are low.
- For the very speculative unconventional gas resources, such as gas hydrates, much too little is known about their in-place gas concentrations and reservoir settings to provide scientifically reliable estimates on recoverable volumes.

An example of the expected growth in the U.S. unconventional gas resource base is shown in Fig. 7 and Table VII. Figure 7 shows the expected growth in tight gas sands, gas shales, and coalbed methane resource between the years 2000 and 2025, stemming from an assumed continually improving knowledge base and technology. Table VII provides an update to Fig. 7, using the latest information for the U.S. unconventional gas database maintained by Advanced Resources International.

Looking ahead, it is useful to recognize that the information and expectations for unconventional gas represent a "snapshot in time." As new plays are

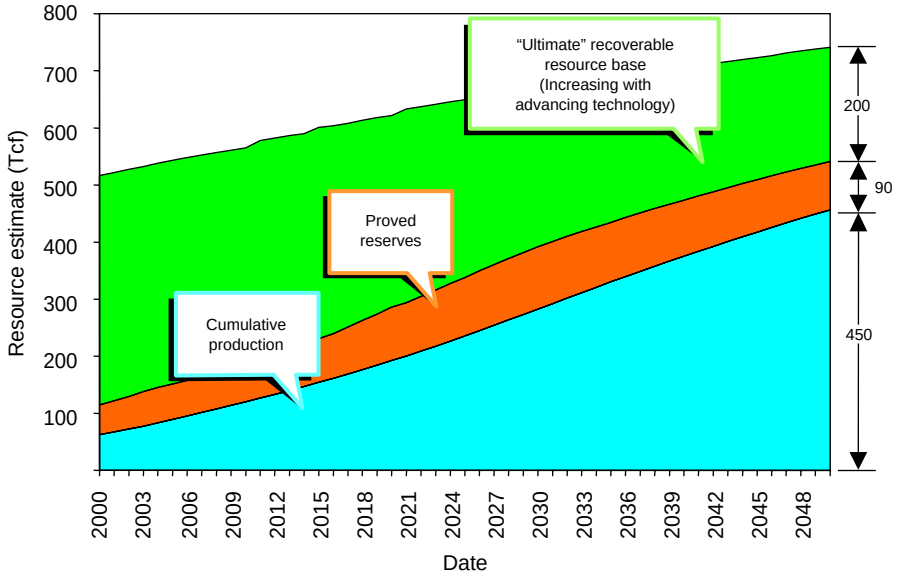


FIGURE 7 Projected growth of unconventional gas resources, years 2000 to 2050. *Source*: Modified by Advanced Resources from National Petroleum Council (1992).

TABLE VII

Estimates of Technically Recoverable U.S. Unconventional Gas Resources (Tcf)

	Year 2002 technology			Year 2025 technology:
	Proved reserves/production	Undeveloped resource	Ultimate resource	Ultimate resource
Tight gas sands	149	292	441	520
Gas shales	13	62	75	90
Coalbed methane	29	95	124	150
Total	191	449	640	760

tested, additional wells are drilled, and more efficient technologies are developed, the outlook for this complex natural gas resource will most likely improve, for two reasons:

1. First, the ultimate size or productivity of the unconventional gas resource base is not yet known. Improved geologic knowledge continues to expand its size by adding new gas plays. Technology progress helps increase recovery from the already defined plays. An example of this is the Barnett gas shale play in the Fort Worth Basin that is now estimated to have 20 Tcf of recoverable resource, up from 1.4 Tcf in the initial estimate prepared over a decade ago (as further discussed below).
2. Second, the economic outlook for unconventional gas resources, basins and plays will likely continue to change significantly with time. With the aid of technology, an unconventional gas basin or play can move from a high cost, marginal resource to a major, highly productive gas reserve. An example of this is the San Juan Basin coalbed methane play that originally was assessed by resource analysts as unaffordable, with costs of over $5 per million ft^3. Today, this important gas play accounts for 9 Tcf of remaining proved gas reserves plus 9 Tcf of past production (as of 2001).

At the other extreme is geopressured methane that was touted, 20 years ago, to hold 1000 years of natural gas supply. No production or reserves have been developed nor are any expected from this resource.

The sequence of resource assessments for the Barnett Shale gas play, illustrates the impact that improved knowledge and technology progress can have on the outlook for an unconventional gas play.

3.4.1 Initial Assessment

The first assessment for the Barnett Shale (Table VIII, column 1) was prepared using data available as of 1990, after the first group of 86 wells was drilled in this gas shale play. The initial wells had relatively low estimated ultimate reserves (EURs) of 0.35 Bcf per well, and the field was being developed on wide well spacings of 320 acres per well. Using this initial data, the Barnett Shale was estimated to have 1.4 Tcf of technically recoverable natural gas.

TABLE VIII

Successive Assessments of Recoverable Barnett Shale Gas Resources

	Initial assessment, 1990	USGS special assessment, 1996	Latest assessment, 1998
Development intensity (acres/well)	320	320	80–320
Completed wells			
Productive	74	180	300
Unproductive	12	30	50
Play (mi^2)	2439	2439	2439
Future wells	4792	4668	10,148
Success rate	0.86	0.86	0.86
EUR/well (Bcf)	0.35	0.84	0.35–1.50
Technically recoverable resources (Tcf)	1.4	3.4	10.0

3.4.2 Second Assessment

The second assessment by the USGS in 1996 (Table VIII, column 2) used a higher EUR per well, due to improved well-completion practices, of 0.84 Bcf per well but still assumed that wide well spacing of 320 acres would be the norm. The improved EUR per well raised the recoverable resource estimate for the Barnett Shale to 3.4 Tcf.

3.4.3 Third Assessment

The third assessment (Table VIII, column 3) incorporated a more optimum, "intensive development" model for the Barnett Shale. It recognized the very limited drainage being achieved by past Barnett Shale wells and that technology progress was leading to higher EURs in the naturally fractured portions of the play. This assessment established that 10 Tcf of technically recoverable gas exists for the Barnett Shale, with much of the economic potential of this resource residing in the higher productivity, sweet spots in the basin that would be developed at spacings of 80 acres per well. Recent updates place the recovery potential at 20 Tcf.

4. UNCONVENTIONAL GAS DEVELOPMENT IN THE UNITED STATES

4.1 Historical Review

Serious pursuit of unconventional gas in the United States began more than 20 years ago, following concerns that the nation was running out of conventional gas supplies. A Federal Power Commissions task force identified large volumes of tight sands gas in-place in three large Western basins. The Bureau of Mines noted that considerable volumes of methane were being vented for safety reasons from coal mines, wasting a valuable resource. In addition, public interest was stirred by major articles in *Fortune* and *The Wall Street Journal* that a new natural gas resource—geopressured aquifers—could provide gas for 1000 years.

These initial findings helped launch a series of more thorough assessments of unconventional gas and studies of how advances in technology could make these large in-place resources productive, as further discussed below.

4.1.1 Tight Gas Sands

Large quantities of natural gas resources were known to exist in geologically complex, tight (low permeability) formations. However, using available well completion technology, the production of gas from these tight formations, for the most part, was too low to support economic recovery. A handful of independents were exploring certain of the tight gas sand basins, with generally a poor record of success. The broad consensus was that significant advances in exploration and well completion technology would be required to develop this large, complex gas resource.

In response, the DOE, the Gas Research Institute (GRI), and industry established research and development programs to unlock the gas resource held in these tight rocks. Spurred by advances in technology, the development of tight gas sands expanded rapidly from the initial limited efforts in the Appalachian and San Juan Basins to the major Rocky Mountain gas basins and into Texas and the midcontinent. By 2001, tight gas sands were producing 3.3 Tcf per year. Proved tight gas reserves were 48 Tcf from a cumulative 80,000 drilled wells (not including the

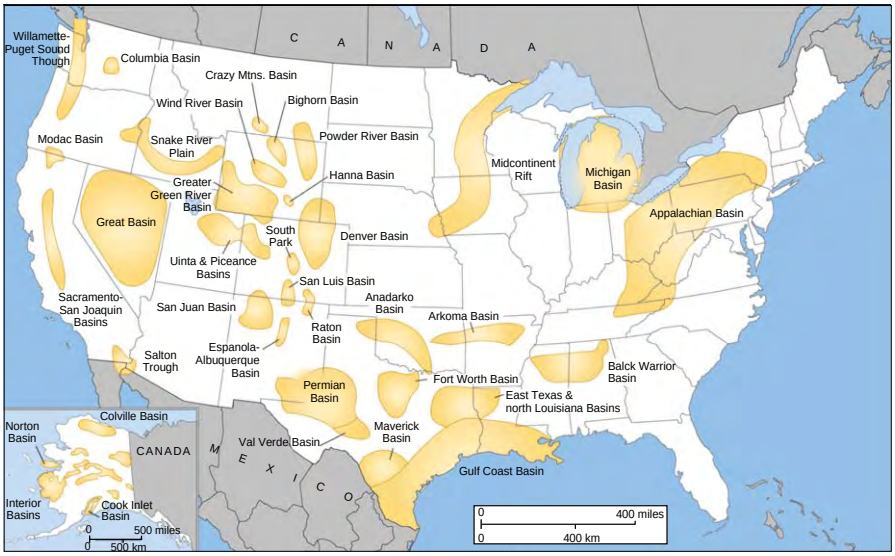


FIGURE 8 U.S. tight gas sand basins. Modified by Advanced Resources from Law (2003).

TABLE IX

In-Place and Technically Recoverable U.S. Tight Gas Sand Resources (Tcf)

Basins	Resource in-place			Remaining recoverable resources	
	FPC (1973)	Kuuskraa *et al.* (1978)	NPC (1980)	USGS[a] (1995/2002)	Advanced Resources International (2002 technology)
Rockies					
Greater Green River	240	91	136	70	108
Piceance	150	36	49	19	21
Uinta	210	50	20	[a]	11
Wind River	—	3	34	N/A	33
Denver	—	19	13	*1*	1
San Juan	—	15	3	16	13
Other					
Anadarko	—	—	—	N/A	19
Appalachia	—	—	—	*45*	16
East Texas	—	81	22	*6*	15
Gulf Coast	—	—	—	N/A	19
Northern Great Plains/ Williston	—	74	148	*43*	21
Permian	—	24	19	N/A	8
Other	—	30	480	*12*	7
Total	600	423	924	212	292

Sources. Federal Power Commission. Supply-Technical Advisory Task Force (1973); National Petroleum Council (1980). Source for values in italics: U.S. Geological Survey (1995).

[a] Included with data for Piceance Basin.

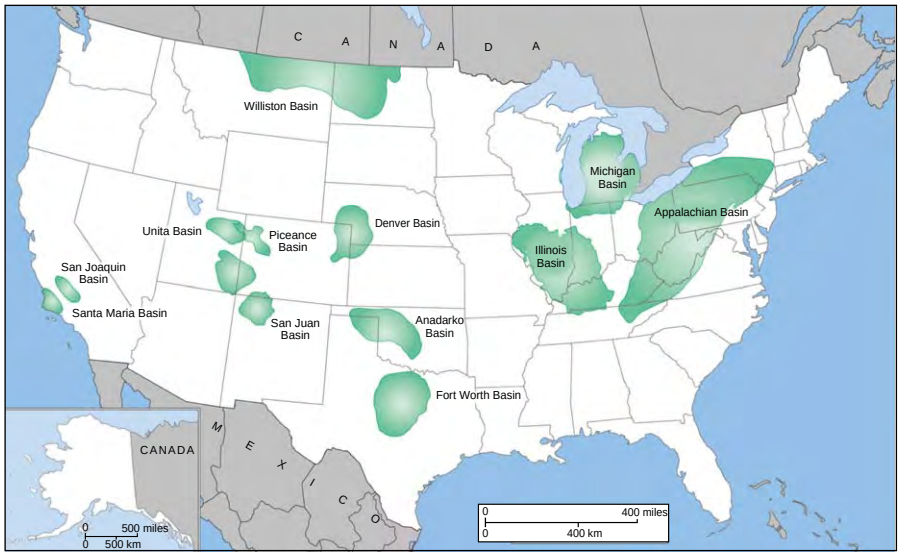


FIGURE 9 U.S. gas shale gas basins. Modified by Advanced Resources from Law (2003).

numerous, older low-producing tight gas wells in the Appalachian Basin), with approximately 101 Tcf of tight gas having been produced. Figure 8 shows the location of U.S. tight gas sand basins. Table IX provides a historical review of tight gas sand resource estimates.

TABLE X

In-Place and Technically Recoverable U.S. Gas Shale Resources (Tcf)

Basins	Resource in-place	Remaining recoverable resources	
		USGS (1995/2002)	Advanced Resources International (2002 technology)
Appraised			
Appalachian Basin (Devonian shale)	248	12	19
Michigan Basin (Antrim shale)	*35*	*19*	12
Fort Worth (Barnett shale)	*5*	N/A	17
San Juan Basin (Lewis shale)		10	10
Speculative			
Illinois Basin (Albany shale)	160	*2*	2
Other shale basins	135	*1*	2
Total	583	44	62

Source. For resource in-place data: National Petroleum Council (1992). Source for data in italics: U.S. Geological Survey (1995).

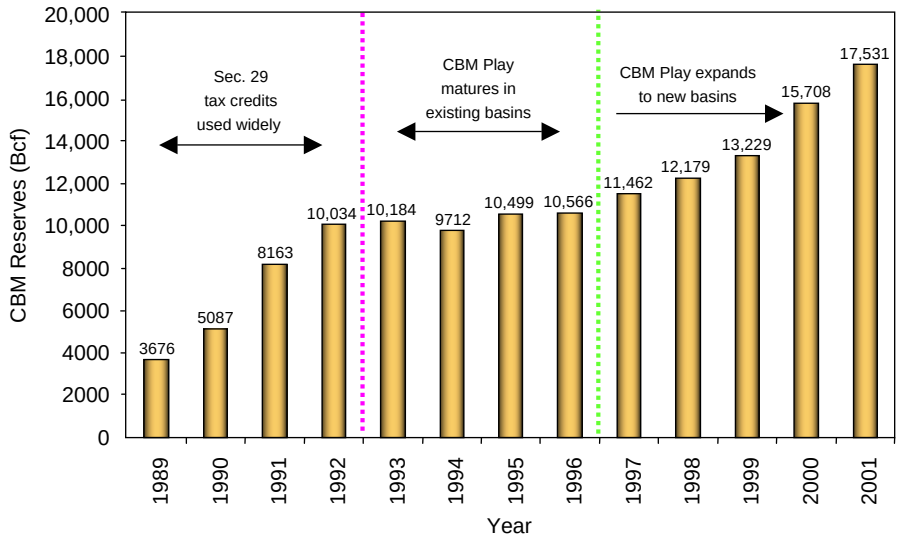


FIGURE 10 U.S. coalbed methane: 10% of U.S. natural gas proved reserves. *Source*: Advanced Research International Database.

4.1.2 Gas Shales

Twenty years ago, only the Appalachian Basin gas shales were being produced. Wells were being completed open-hole, with little definition of productive pay zones, and were being stimulated with nitroglycerine (a remnant of early 1900s technology). Much of the activity was centered in the Big Sandy area of eastern Kentucky. Little understanding existed about the key gas storage and production mechanisms or about geologically similar gas shale plays in other parts of the country. Gas shale drilling averaged only 200 wells per year and the play was in decline.

By 2001, annual gas shale production had reached 500 Bcf. Proved reserves are 7 Tcf, with another 6 Tcf having been produced. Stimulated by Section 29 tax credits and the expansion into new gas shale basins in Michigan and north Texas, 20,000 gas shale wells were drilled from 1978 to 2001 with a peak of 1700 gas shale wells completed in 1992. Figure 9 shows the location of U.S. gas shale basins. Table X provides estimates of the resource in-place, proved reserves, and remaining technically recoverable gas shale resources.

4.1.3 Coalbed Methane

The combination of building a scientific base of knowledge, developing appropriate technology, and providing economic incentives launched the birth of

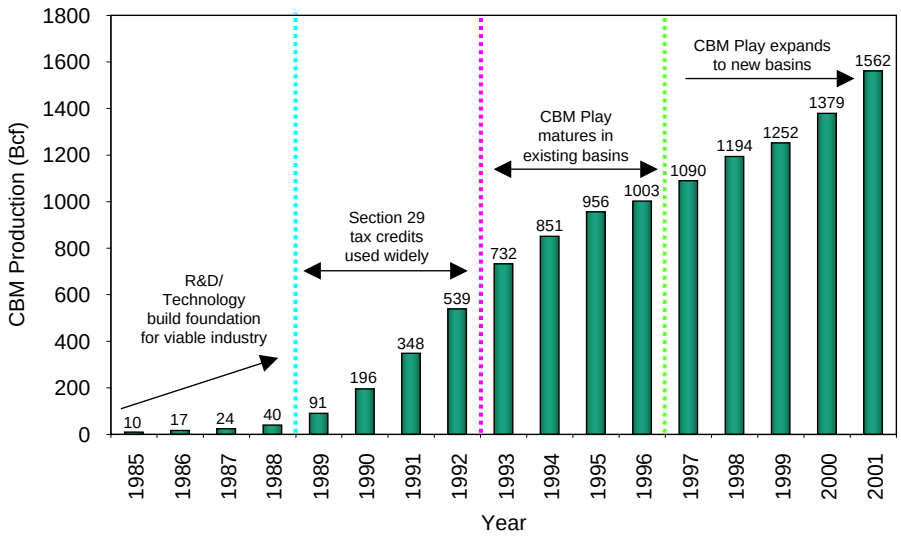


FIGURE 11 U.S. coalbed methane: 8% of U.S. natural gas production. *Source*: Advanced Resources International Database.

TABLE XI

In-Place and Technically Recoverable U.S. Coalbed Methane Resources (Tcf)

Basin	Resource in-place	Production through 2001	Proved reserves 2002	Remaining recoverable resources (2002 technology)	Ultimate resources (2025 technology)
Rockies					
Greater Green River	314	[a]	[a]	6.2	8
Uinta	10	0.3	1.7	6.5	10
Piceance	81	0.1	[b]	7.3	9
Powder River	61	0.6	2.4	36.0	50
Raton	12	0.2	1.4	6.2	9
San Juan	84	8.7	9.1	17.2	42
Other (Hanna, Wind River)	21	[a]	[a]	[a]	[b]
Other					
Northern Appalachia	61	[a]	[a]	5.0	6
Central Appalachia1.3	5	0.3	1.4	3.4	6
Warrior	19	1.3	1.2	3.2	7
Cherokee/FC/Arkoma	10	[a]	0.3	3.0	4
Illinois	21	[a]	[a]	0.6	1
Other (Gulf Coast, Western Washington	30	[a]	[a]	[a]	[b]
Total	729	11.5	17.5	94.6	152

[a] Less than 0.05 Tcf or N/A.
[b] Assuming year 2025 technology.

a new natural gas industry: coalbed methane. Much of the early development was by independent producers, such as Devon Energy and Meridian Oil, who subsequently saw their gas production and reserve holdings rise sharply.

Coalbed methane production climbed from essentially zero in the mid-1980s to 1.6 Tcf in 2001. Proved reserves are 17.5 Tcf, with another 11.5 Tcf having already been produced (see Figs. 10 and 11). The development and continuing adaptation of

technology enabled the industry to remain profitable and vigorous, even after the expiration of Section 29 tax credits in 1992. Today, several new coalbed methane basins and plays are being actively developed, including the Powder River (Wyoming), Raton (Colorado), and Uinta (Utah), providing a base for continued growth. Table XI provides estimates of the in-place and technically recoverable coalbed methane resource for the United States.

4.1.4 Geopressured Methane

An aggressive research and development program assembled considerable geologic and reservoir knowledge for this resource. And, although no commercial natural gas production was established, the research and development program helped bring a strong dose of reality and understanding regarding the viability this gas resource. It also helped dispel the speculation that "1000 years of natural gas" was at hand.

4.2 Assessment of Results

Unconventional gas offers one of the great success stories of progress in technology. A poorly understood, high-cost energy resource, one that the U.S. Geological Survey had not even included in its national appraisals of future gas resources (until their 1995 assessment), is now providing major volumes of annual gas supplies and helping meet growing domestic natural gas demand.

5. CONCLUSION

The discussion and resource estimates for unconventional gas presented in this article are a snapshot in time. The ultimate size or productivity of this geologically complex natural gas resource is not yet known. However, it is useful to document what is known as the foundation for continually improving the base of knowledge and the expectations for this large potential source of future natural gas supplies.

SEE ALSO THE FOLLOWING ARTICLES

Depletion and Valuation of Energy Resources • Natural Gas, History of • Natural Gas Industry, Energy Policy in • Natural Gas Processing and Products • Natural Gas Resources, Global Distribution of • Natural Gas Transportation and Storage • Oil and Natural Gas Resource Assessment: Classifications and Terminology • Oil and Natural Gas Resource Assessment: Geological Methods • Oil and Natural Gas Resource Assessment: Production Growth Cycle Models

Further Reading

Brandenburg, C. F., Kuuskraa, V. A., and Schraufnagel, R. A. (1998). "Coalbed Methane and Technology: Relating the U.S. Experience into International Opportunities," Presented at the 17th Congress of the World Energy Council, Houston, Texas, September 13–18, 1998.

Fisher, W. L., Pedersen, A. H., and Rosenberg, R. B. (1994). "Workshop Consensus," Proceedings of the Global Gas Resources Workshop, September 19–21, 1994. International Centre for Gas Technology, p. ix–x.

Kuuskraa, V. A. (1998). Barnett shale rising star in Fort Worth basin. *Oil Gas J.* **96,** 67–76.

Kuuskraa, V. A., and Meyers, R. F. (1983). Review of world resources of unconventional gas. *In* "Conventional and Unconventional World Natural Gas Resources" (C. Delahaye and M. Grenon, Eds.), pp. 409–458. IIASA Collab. Proc. Ser., CP-83-S4, International Institute of Applied Systems Analysis (IIASA), Laxenburg, Austria.

Law, B. E., and Curtis, J. B. (2002). "Unconventional Petroleum Systems," AAPG Bulletin, Vol. 86, No. 11, November, ISSN 0149-1423.

Nakicenovic, N., and Riahi, K. (2001). "An Assessment of Technological Change across Selected Energy Scenarios." Reprinted from a publication by the World Energy Council on behalf of the Study Group on Energy Technologies for the 21st Century. RR-02-005.

Rogner, H. H. (1997). An assessment of world hydrocarbon resources. *Annu. Rev. Energy Environ.* **22,** 217–262.

Schmoker, J. W., Quinn C. J., Crovelli, R. A., Nuccio, V. F., and Hester, T. C. (1996). "Production Characteristics and Resource Assessment of the Barnett Shale Continuous (Unconventional) Gas Accumulation, Fort Worth Basin, Texas." USGS Open-File Report 96-254. U.S. Geological Survey.

U.S. Geological Survey National Oil and Gas Resource Assessment Team (1995). "National Assessment of U.S. Oil and Gas Resources," USGS Circular 111S.1995. U.S. Geological Survey.

Natural Gas Transportation and Storage

FARUK CIVAN
University of Oklahoma
Norman, Oklahoma, United States

1. **Introduction**
2. **Natural Gas**
3. **Natural Gas Transportation**
4. **Natural Gas Storage**
5. **Gas Transportation and Storage Services, Operation, and Economic Value**
6. **Hazards from Natural Gas Transportation and Storage**
7. **Natural Gas Security and Risk Issues**

Glossary

aquifer A porous permeable sedimentary formation located belowground and containing groundwater.
base or cushion gas The minimum quantity of natural gas that must be retained in gas storage to ensure the production of natural gas from storage at a sufficient rate and pressure; usually measured in volume expressed at 15.6°C (60°F) and 1.0 atmosphere of pressure (standard conditions).
deliverability The rate at which natural gas can be withdrawn from gas storage; usually measured in volume expressed at 15.6°C (60°F) and 1.0 atmosphere of pressure (standard conditions) per unit day.
hydrate A cagelike crystalline solid material similar to ice, formed by combining gas with water.
injectivity The rate at which natural gas can be injected into gas storage; usually measured in volume expressed at 15.6°C (60°F) and 1.0 atmosphere of pressure (standard conditions) per unit day.
inventory A survey of the total amount of gas contained in storage; usually measured in volume expressed at 15.6°C (60°F) and 1.0 atmosphere of pressure (standard conditions).
liquefied natural gas (LNG) The liquid state of natural gas, essentially containing methane and little ethane.
liquefied petroleum gas (LPG) The liquid form of propane and butane.
million British thermal units (MMBtu) A measure of heat generated by gas combustion.
permeability A physical parameter expressing the fluid (gas or water) transmitting capacity of the interconnected porous space available in depleted oil and gas reservoirs and water aquifers.
pipeline A long pipe used to transmit fluids.
porosity The volume fraction of the porous space occupied by fluids (gas and water) in depleted oil and gas reservoirs and water aquifers.
reservoir An underground geologic permeable rock formation; contains fluids such as oil, gas, and water in pore spaces.
salt cavern A cavity created in a deep salt bed or dome following the dissolution of salt by circulating hot water.
standard or base conditions Thermodynamic conditions considered as the reference conditions for volume measurement, usually temperatures of 0°, 15.6°, or 25°C and 1.0 atmosphere of pressure.
storage Facilities of various types and capacities used to store natural gas.
storage measures The various means of expressing the properties and conditions of a gas storage area and its gas content.
thousand standard cubic feet (Mscf) A measure of gas volume.
total capacity The largest quantity of natural gas that can be loaded into a particular gas storage area; usually measured in volume expressed at 15.6°C (60°F) and 1.0 atmosphere of pressure (standard conditions).
working gas capacity The difference between the total and base gas capacities of gas storage.
working (top) gas The maximum quantity of natural gas that a particular gas storage can hold and still allow for circulation; usually measured in volume expressed at 15.6°C (60°F) and 1.0 atmosphere of pressure (standard conditions).

Natural gas is an important commodity traded frequently in the energy market. It can be utilized as a source of energy by direct consumption as fuel or

 © 2004 Elsevier Inc. All rights reserved.

by other means, including after conversion to hydrogen for use in fuel cells or after conversion to various chemicals. Because it is a gas and a low-density substance at ambient pressure and temperature conditions, safe and economic handling of natural gas during transportation, storage, and utilization requires special measures and sophisticated technologies. Therefore, the worldwide trade and utilization of natural gas is inherently limited by the difficulty of handling gas between the source and market.

1. INTRODUCTION

Natural gas is gaining increasingly great potential as an abundant and environmentally friendly energy source; it is available worldwide from onshore and offshore reserves in various forms, including as gases or as condensates in underground geological reservoirs, and as solid hydrates in hydrate-bearing sedimentary formations present in permafrost regions and ocean floors. Thus, technologies are being developed rapidly for its economical production and utilization. Although transportation is necessary as a means of delivering gas from production to market, gas storage ("parking" gas temporarily) at appropriate locations is instrumental in compensating between the surge and peak effects in the gas supply-and-demand value chain. Accurate measurement of natural gas using reliable meters is another important issue involved in custody transfer and inventory accounting in natural gas trading. Effective engineering and management of the gas transportation and storage facilities also include strategic security measures and emergency response procedures. Capacity and location of the source and the user, available and feasible technologies, and economics determine the means of transportation and storage used. The various types and capacities of transportation and storage facilities involved in natural gas handling and value chain are summarized in Table I.

2. NATURAL GAS

Natural gas is a mixture of hydrocarbon and nonhydrocarbon gases; these gases are produced from petroleum and hydrate-bearing formations present in geological sedimentary formations located below the ground surface and the ocean floor. Typical hydrocarbon components of natural gas

TABLE I

Means of Transportation and Storage of Natural Gas[a]

Natural gas compression	Direct gas liquefaction	Conversion to hydrate blocks or hydrate-in-oil slurry	Dissolution of petroleum gas in pentane-plus-heavier fractions of crude oil	Chemical conversion to liquid products	Electric power generation
Pipeline gas transportation with intermediate recompression	Liquid storage and loading	Transportation by tanker and pipeline	Crude transportation by tanker	Liquid transportation by tanker	Electric transmission by cable
Gas storage above ground and below ground	Liquefied gas transportation by tanker	Hydrate storage and transport to destination	Crude storage and transport to destination	Liquid storage and transport to destination	Electric distribution
Distribution	Liquefied gas unloading, storage, and transport to destination	Regasification and distribution	Regasification and distribution	Liquid distribution	
Compressed gas tank filling	Regasification and distribution				
Gas tank transportation by truck and other means					
Storage as compressed gas in tanks					
Gas distribution					

[a] From oil and gas reservoirs, as coal bed methane, as ocean–floor and permafrost hydrates, and gathered from wells.

include a significantly large percentage of methane, and relatively smaller percentages of ethane, propane, isomer and normal butanes and pentanes, hexane, heptane, and other heavier hydrocarbons. The typical nonhydrocarbon components can be classified as acid gases, such as hydrogen sulfide and carbon dioxide; inert gases, such as nitrogen and helium; odorous sulfurous gases, such as hydrogen sulfide and various mercaptans; and other impurities, such as water and mercury vapors and sulfur dissolved in gas.

The composition of natural gas dictates its behavior during the chain of processes involved in gas handling, including its transportation and storage. Its density ρ or specific volume v are given by the real gas equation of state as in Eq. (1):

$$\rho = \frac{1}{v} = \frac{m}{V} = \frac{MP}{ZRT}, \tag{1}$$

where m, V, P, and T denote the mass, volume, pressure, and temperature of the gas, M is the gas molecular mass (16 kg/kmol for methane), R is the universal gas constant (8308 J/kmol°K), and Z is the real gas deviation factor, a function of the pressure, temperature, and composition of gas, usually given by empirical correlations or charts. The density of the methane gas is about 0.65 kg/m^3 at 25°C and 1 atm under ambient conditions. The density of natural gas relative to the density ρ_a of dry air of standard composition, at standard conditions, is referred to as the specific gravity G, given by Eq. (2):

$$G = (\rho/\rho_a)_b = M/M_a, \tag{2}$$

where the molecular weight of air is $M_a \cong 29$ kg/kmol. The standard conditions b are defined as zero or 15.6°C (60°F) temperature and 1.0 atmosphere of pressure. The gravity of methane gas is 0.55. The gravity of typical natural gas is about 0.64, hence it is lighter than air. Therefore, the transportation and storage of natural gas in large mass present a challenging engineering problem, requiring sophisticated and expensive technologies. This is because the volume of natural gas per unit mass (referred to as the specific volume) is very large and its density is very low at ambient conditions. For example, 1 kg of methane gas occupies 1.5 m^3 at 25°C and 1.0 atm of pressure. Consequently, natural gas should be compressed or converted to smaller volumes by suitable means in order to be able to handle the large quantities in a convenient manner.

At elevated pressure and sufficiently low temperature conditions, natural gas can liquefy and produce a multiphase system, including a liquid phase, referred to as condensate, and a solid phase (when combined with water), referred to as a natural gas hydrate. Production of condensates and hydrates of natural gas creates various difficulties during transportation and storage of natural gas. Coping with such problems is an important issue in natural gas transportation and storage, and an economic and technological headache.

3. NATURAL GAS TRANSPORTATION

Transportation facilities of various types are used for collecting gas from distributed sources and for delivering it to end users (the customers), who utilize gas for various purposes. The worldwide trade and utilization of gas greatly depend on the availability and economics of gas transportation and distribution facilities. The primary gas transportation technologies are based on compressed gas transport, direct pipeline transfer, gas-to-liquid conversion, gas-to-solid conversion, and gas-to-power conversion.

3.1 Compressed Gas Transport in High-Pressure Tanks

Natural and petroleum gases can be transported as compressed gas in small quantities inside high-pressure tanks for use at residences, in industry, and in gas-powered vehicles. Based on Eq. (1), the volume of compressed gas at the prevailing pressure P and temperature T can be estimated by Eq. (3):

$$V = V_b Z \frac{T}{T_b} \frac{P_b}{P}, \tag{3}$$

where P_b and T_b denote the base or standard pressure and temperature, respectively, and V_b is the standard gas volume.

3.2 Direct Pipeline Transfer

Pipeline transportation is the oldest and most common technology; pipelines are widely utilized worldwide for transportation of gas. Pipeline transportation of natural gas involves a variety of equipment and instrumentation, such as a pipeline network to carry gas from different sources to the end users; meters to measure the gas quantity and flow; valves to control the gas flow; regulators to control the gas pressure; safety devices, expansion systems, and storage facilities for temporary storage

TABLE II
Worldwide Natural Gas Pipeline Extent[a]

Location	Length (km)
North America	405,980
Latin America	35,041
Europe	267,846
Africa	7160
Middle East	6854
Asia–Oceania	145,519
World total	868,400

[a] Estimated based on data in *The World Factbook* (available on the Internet through the Central Intelligence Agency Web site at http://www.cia.gov).

of gas; dehydrators to remove the moisture and prevent water condensation, hydrate plug formation, and corrosion; compressors to pressurize and "flow" the gas; prime movers and electrical control systems to operate these systems; and remote data acquisition, monitoring, and transmission systems. Table II provides worldwide statistics about the extent of natural gas pipelines based on information provided by *The World Factbook*.

One of the frequently used pipeline flow equations is the Weymouth equation, expressing the gas flow rate q_b at the base pressure and temperature conditions, given by Eq. (4):

$$q_b = \frac{\pi}{4}\frac{T_b}{P_b}\left[\frac{R(P_1^2 - P_2^2)D^5}{M_a G \bar{Z} T \bar{f} L}\right]^{1/2}. \quad (4)$$

The symbol L denotes the length of the pipeline, the gas pressures at the inlet and outlet ends of the pipeline are P_1 and P_2, D is the internal diameter of the pipe, M_a is the molecular weight of standard air, and $\bar{Z}$ and $\bar{f}$ are the average values of the real gas deviation factor and the pipe friction factor over the pipe length.

3.3 Gas-to-Liquid Conversion

Liquids can be more conveniently, safely, and economically transported in large quantities than can gases. Natural and petroleum gases can be transformed into liquids either by direct cryogenic condensation, by mixing, or by chemical conversion. Cryogenic condensation forms liquefied gases, such as liquefied natural gas (LNG), which essentially consists of methane and little ethane, and liquefied petroleum gas (LPG), which consists of the propane and butane cut. Liquefaction occurs by mixing natural or petroleum gas with the pentane-plus-heavier hydrocarbon fractions separated from an associated gas of crude oil. Chemical conversion yields liquid products such as methanol and gasoline. Chemical conversion of natural gas to liquid products requires the application of a series of complicated and expensive chemical processing methods. Therefore, chemical conversion is not a generally preferred option. Nevertheless, the liquids produced by chemical conversion are stable at ambient conditions and can be conveniently carried and safely stored at moderate near-ambient conditions. Following liquefaction, intercontinental and marine transport of liquids is accomplished by means of specially designed tankers or ships.

Transportation of gas in liquefied form is a broadly exercised option. However, this combines a series of delicate technological issues, including compression, Joule–Thomson expansion turbines, cooling and heating, fractionation, and specially designed storage tanks for tankers. Liquefied natural gas is transported as a cryogen at the normal boiling temperature at atmospheric pressure. This temperature is -161.5°C for pure methane. However, natural gas also contains small quantities of other hydrocarbon gases. Therefore, the liquefaction process undergoes a gradual condensation by cooling from heavier to lighter components at different temperatures, ending with the condensation of methane at -161.5°C. The density of the liquid methane is about 420 kg/m^3 at this condition. This means that 1 kg of liquid methane occupies a volume of only 0.0024 m^3. Hence, the volume of the liquid methane is 625 times less than the volume of the methane gas at ambient conditions.

The LNG tankers for marine transport are specially designed and constructed to carry large quantities of liquefied natural gas safely. The hull of LNG tankers usually is double, consisting of outer and inner hulls, separated with a space sufficient to protect against damage by contact or collision of tankers with other objects. Essentially, integrated and self-supporting tanks are used as containers in the conventional LNG tankers. The LNG tankers incorporating the integrated-tanks technology contain a series of rectangular tanks containing a membrane to resist the external forces. The LNG tankers incorporating the self-supporting tanks technology contain a series of individually reinforced tanks, usually spherical in shape, designed to resist the hydrostatic force of their fluid content. These tanks are placed at sufficient distances from the neighboring tanks and are insulated

individually. LNG transport by tankers also requires auxiliary services for liquefaction of natural gas, and LNG loading, storage, discharging, and regasification.

3.4 Gas-to-Solid Conversion

Under proper conditions, natural gas combines with water to form a cagelike crystalline solid material similar to ice, called a gas hydrate. The gas and water molecules are held together because of the molecular interaction forces. Hydrates can hold approximately 150 to 180 standard (at 15.6°C or 60°F and 1.0 atm) volumes of natural gas per unit hydrate volume and they are stable in the −10° to −15°C temperature range at atmospheric conditions. Natural gas hydrates exist in ocean floor and permafrost regions. Gas hydrates can also be artificially manufactured in the form of rectangular and cylindrical blocks and transported conveniently via ships and other means. Hydrates can also be transported through pipelines in the form of slurry with oil. For this purpose, the gas and water are placed into contact at conditions near 5°C and 50 atm in stirred-tank or spray-tower reactors to produce gas hydrates. This process generates some heat, which must be removed by appropriate cooling technologies. This rather new technology is under development. Hydrate transportation requires additional processes to form and dissolve the hydrates.

3.5 Gas-to-Power Conversion

Conversion of gas to electric power is accomplished in gas-powered combined-cycle electric power plants. Electricity is the most convenient means for transport, using electric transmission cables of various types. However, electric storage is a challenging technological barrier. Theoretically, 1.0 kg of methane can generate 14 kilowatt-hours (kWh) of electric energy if all the chemical energy can be converted to electric energy with 100% conversion efficiency. However, the actual energy generated by combustion of methane is less than this quantity, because of the loss of energy by friction and other means.

4. NATURAL GAS STORAGE

Temporary natural gas storage capability is essential in order to balance out the supply of natural gas with demand, and as a backup reserve to compensate for shortages and temporary interruptions of gas supply

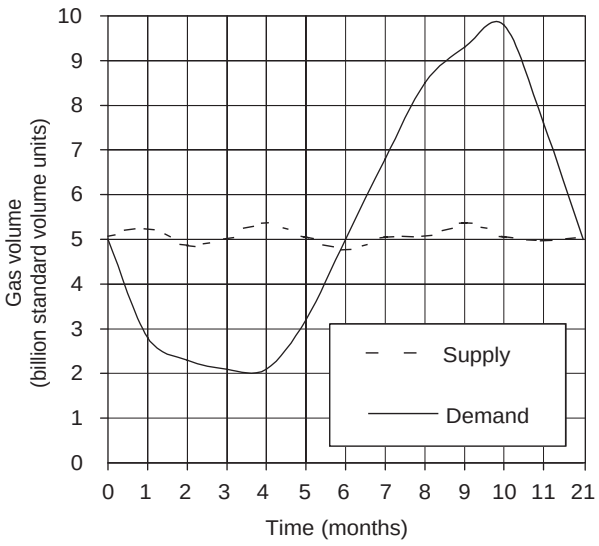


FIGURE 1 Natural gas storage and consumption based on the supply and demand trends.

or demand. Gas storage provides a convenient leverage and advantage in risk management, combating and dealing with peaks in the gas supply-and-demand chain, and accommodating gas price fluctuations. Natural gas is temporarily stored when the market is unfavorable and/or demand is low, and the stored gas is made available under favorable market and demand conditions. For example, as a heating fuel, gas is stored during summers and consumed during winters (Fig. 1). Companies providing various gas utilization services, local distribution, and intra- and interstate pipeline transportation rely on gas storage in order to alleviate the variations in the supply-and-demand chain. When storage cannot meet the demand for natural gas, liquefied natural or petroleum gases are regasified for the market.

Various features, such as capacity, accessibility, and gas injection and production rates, determine the type of gas storage for particular applications and requirements. The conventional gas storage options can be broadly classified as aboveground and underground storage facilities.

4.1 Aboveground Natural Gas Storage

The primary means of aboveground gas storage involve pipelines, cryogenic tanks, and gas hydrates.

4.1.1 Pipeline Storage

Pipeline networks can be used to provide temporary gas storage when they are not needed for gas

transmission purposes. This helps in coping with the rapid surge and peak situations in the supply-and-demand value chain over short time frames. Applying the Weymouth equation given by Eq. (4), the total quantity of gas stored in a pipeline after closing its ends can be estimated as follows:

$$V_b = \frac{\pi D^2 T_b L}{6 P_b T}\left(P_1 + P_2 - \frac{P_1 P_2}{P_1 + P_2}\right). \qquad (5)$$

4.1.2 Storage in Cryogenic Tanks

Tank storage affords small capacity and is expensive, but can be facilitated practically anywhere. Natural gas can be stored cryogenically in small quantities in specially designed storage tanks at –155° to –160°C temperatures. These tanks are well insulated against heat influx from the surroundings and therefore cooling is not required. Further, the storage tanks can be placed inside specially designed dikes to confine the liquefied gas in case of accidental spills. Cryogenic storage tanks may be constructed as double-wall or membrane tanks.

Prolonged storage of liquefied gas in storage tanks may create a hazardous condition resulting from external heating at the bottom of the tank over time. Heating can cause a density difference and stratification of the stored liquid, eventually leading to rollover and rapid mixing of the colder and heavier upper layer with the warmer and lighter lower layer of the cryogen. The rollover phenomenon can be avoided by maintaining a uniform-density liquid throughout the tank; this can be accomplished by various methods, including proper loading and discharging procedures, continuous recirculation of the liquid in the tank, and reliquefying the vapors generated by boil-off and refluxing of the liquefied vapors back into the tank.

4.1.3 Storage as a Gas Hydrate

Gas can be conveniently stored in the form of hydrate blocks. As explained for transporting gas as hydrates, typical hydrates can hold approximately 150 to 180 standard volumes of natural gas per unit hydrate volume, and they are stable in the −10° to −15°C temperature range at atmospheric conditions. This technology enables storing large quantities of natural gas in the hydrate state.

4.2 Underground Gas Storage

There are various options for underground gas storage, varying in capacity and performance, such as in-ground storage tanks, depleted and abandoned oil and gas reservoirs, aquifer storage, salt and hard-rock caverns, and abandoned mine shafts. Table III provides several statistics about the worldwide disposition of underground gas storage capacity.

4.2.1 In-Ground Storage Tanks

In-ground storage tanks and aboveground tanks are constructed similarly, but in-ground tanks are partially or fully buried in the ground.

4.2.2 Depleted and Abandoned Oil and Gas Reservoirs

Oil and gas reservoirs with their existing wells and surface handling facilities are widely used for underground gas storage after their economically recoverable hydrocarbon content has been depleted. Depleted gas reservoirs are preferred because their

TABLE III

Worldwide Underground Gas Storage Capacity[a]

Location	Storage sites (number)	Working capacity (10^9 m^3)	Total capacity (10^9 m^3)	Peak withdrawal rate (10^6 m^3/day)
United States	386	102	228	1894
Western Europe	66		86	782
Central Europe	17	7.3	—	—
New independent states	46	80.5	—	—
Russia	21	—	—	—
World	554	243	502	

[a] Based on data from an international association of natural gas industries and the Institut Français du Pétrole (IFP-CEDIGAZ). (Available at the Web site http://www.cedigaz.com.)

structure is more likely to be leakproof compared to oil reservoirs. The injectivity and deliverability of gas storage will vary with the prevailing condition and quantity of stored gas and with several other factors, including the energy or driving mechanisms (such as water drive) of the reservoir storage for gas transport; the gas and storage formation rock compressibility; the amount of base gas; the number, arrangement, and performance of the wells completed into storage; and effectiveness of the surface gas handling facilities.

Reservoir rock is a compacted sedimentary porous formation, often naturally fractured. Fractures form the highly conductive flow paths, and the interconnected pores of the rock blocks (separated by the fractures) provide the essential gas storage capability. Permeability of the storage rock formation determines the rate of gas flow through it. High-permeability reservoirs are preferred for gas storage because these can allow for high rates of injectivity and deliverability. However, over time, the performance of depleted reservoir storage can be reduced by impairment of permeability due to various damage mechanisms, such as clay swelling, fine aggregate ("fines") migration, and precipitation of scales. The performance can be restored by costly special measures, such as acid flushing and hydraulically created fracture stimulation. Depleted reservoir storage requires careful operation and mitigation in order to avoid the possibility of coproduction of water; water production can occur due to the coning phenomenon, because most reservoirs contain an underlying water table. A large quantity of base gas should be retained in underground storage in order to avoid this problem. The possibility of facilitating inert gases is being explored for use as a base or cushion gas. Typically, depleted reservoir storage involves ten to hundreds of wells, 5 million to 10 million standard cubic feet per day deliverability capability, only 33–50% of the total gas content available as working gas, a significant amount of water, and inflexibility, allowing only one to two cycles of injection and production per year with 100 to 150 days of withdrawal periods. Applying the leaky-cylindrical tank model depicted in Fig. 2, the relationship between gas flow rate and pressure is given by Eq. (6):

$$q_b = C(\bar{P}^2 - P_w^2)^n, \tag{6}$$

where $\bar{P}$ and P_w denote the average reservoir gas pressure and the gas pressure at the well bottom, respectively, and C and n denote the performance coefficient and exponent, respectively, which have

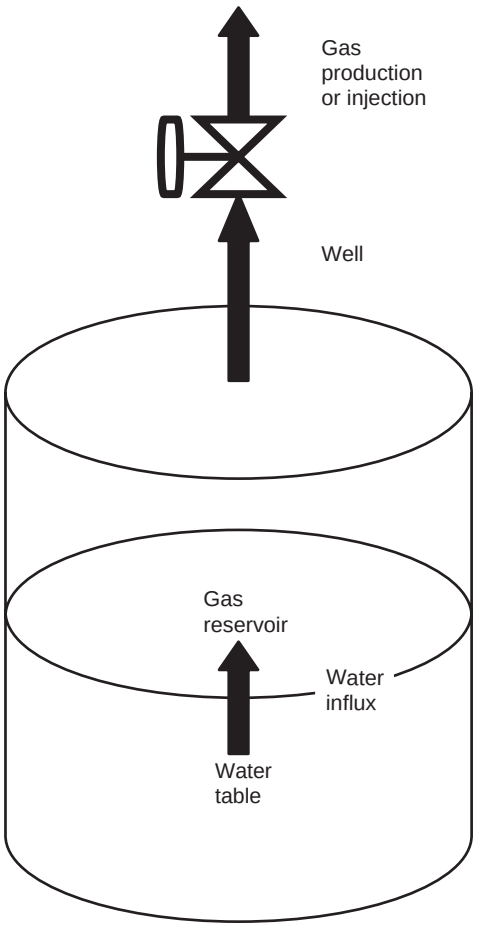


FIGURE 2 Underground gas storage reservoir and water table.

empirically determined specific values for specific storage reservoirs. Assuming constant flow rate, the cumulative production/injection gas volume is given by Eq. (7):

$$G_p = q_b t, \tag{7}$$

where G_p and q_b denote the cumulative volume and volume flow rate of the gas under standard conditions (15.6°C or 60°F and 1.0 atm), and t denotes the total elapsed time.

4.2.3 Storage in Aquifers

Suitable underground water aquifers, leakproof for gas, can be used for gas storage. Aquifer storage must have an impermeable cap rock formation to avoid gas leakage. The injectivity and deliverability of this storage depend on the gas solubility in water and the water-drive potential of the aquifer. Generally, use of aquifer storage requires a large quantity of base gas to be retained in place. Such storage needs careful

operation and mitigation in order to avoid the possibility of coproduction of water with gas. The other features of aquifer storage (including typical inventory) and the injection and production capabilities of aquifers are similar to those of depleted reservoirs.

Based on Fig. 2, the storage gas balance equation can be expressed as follows in order to relate the gas pressure in the storage to the cumulative gas production or injection:

$$\frac{P}{Z} = \frac{P_i}{Z_i}\left(1 - \frac{G_p}{G_i}\right) \Big/ \left(1 - \frac{W_e E_i}{G_i}\right), \tag{8}$$

where the subscript i denotes the values at initial or reference condition of the gas storage, and E is the gas expansion factor, given by Eq. (9):

$$E = \frac{T_b}{P_b}\frac{P}{Z}\frac{1}{T}. \tag{9}$$

The water content of the aquifer is denoted by W_e, and G_i and G_p denote the initial volume of gas stored in the storage and the cumulative volume of gas produced from the storage, respectively, both expressed at standard conditions. The water rises when gas is produced from the storage, and recedes when gas is injected. The water influx or efflux rate q_w is given proportionally to the pressure difference between the aquifer water and storage gas as

$$q_w = \frac{dW_e}{dt} = J(\bar{P}_a - P), \tag{10}$$

where W_e is the aquifer water content, t is time, J is an aquifer productivity index, P is the storage reservoir gas pressure, and $\bar{P}_a$ is the average water pressure in the aquifer.

4.2.4 Storage in Salt Caverns or Cavities

Salt caverns are created by dissolution of salt in deep salt beds and domes by hot water circulating through wells drilled into the salt formation. Storage capacity is significantly smaller in these cavities compared to the depleted reservoirs and aquifers, but cavern storage is flexible and allows for multiple cycles of gas injection and production in a year. The construction of salt cavern storage is expensive and is preferred usually only when depleted reservoir or aquifer storage opportunities are unavailable. Cavern storage provides significantly greater gas injectivity and deliverability rates in comparison to the depleted reservoirs and aquifers, and requires holding a much smaller amount of base gas in the storage to be effective. Typically, salt cavern storage involves one to three wells, 8.5 million to 17 million standard cubic meters per day deliverability capability, about 65% of the total gas content available as working gas, a negligible amount of water, and 6 to 12 cycles per year with 10 days of withdrawal periods. Simplifying Eq. (8) by neglecting the water influx effect, the gas pressure and cumulative gas production or injection at isothermal conditions are related by Eq. (11):

$$\frac{P}{Z} = \frac{P_i}{Z_i}\left(1 - \frac{G_p}{G_i}\right). \tag{11}$$

4.2.5 Hard-Rock Caverns and Abandoned Mine Shafts

The possibility of storing natural gas in properly conditioned and prepared hard-rock caverns and abandoned mine shafts is under consideration. Refrigerated mined caverns operating in the -29° to -40°C temperature range and specially lined rock caverns that can store compressed gas at elevated pressures are being explored in order to increase the capacity of flexible gas storage. However, the gas industry has not yet implemented such storage for conventional applications. The operational equation for this type of storage is similar to that for the salt cavern storage.

5. GAS TRANSPORTATION AND STORAGE SERVICES, OPERATION, AND ECONOMIC VALUE

The economic value of gas transportation and storage depends on many factors, including supply and demand trends; means of transportation and storage; measures for deliverability assurance; operational and management strategies; ability of accommodation for seasonal price fluctuations; measures for management of uncertainty, risk, surge, peaks, and pricing options; and conditions and types of available services. Mitigation of the uncertainty in natural gas price, investment, decision-making, and supply and demand is the challenge in the marketing and management issues. Development and implementation of best strategies for effective coordination of various gas sources, flexibility of transportation and storage facilities, measures for absorbing and managing peaks, and dealing with price volatility and seasonal price spreads are among the important problems. Typically, the price of natural gas remained close to $5.00 per million British thermal units (MMBtu), or $5.00 per thousand standard cubic feet (Mscf) during the most of the year 2003.

Pipeline operation is a dynamic industry, providing creative services such as gathering, lending, parking, pooling, wheeling, and interruptible and backhaul services. The cost of pipeline transportation depends on the expense, usage, distance, and types of services rendered. While providing valuable service to the suppliers and end users of gas, the gas marketing and trading industry makes a profit from the seasonal buying and selling of gas price differentials, by managing risk and taking advantage of creative services such as options, swaps, and futures.

Gas storage ensures adequate and reliable gas supply with less price volatility. Inventory accountability, deliverability assurance, and gas leakage mitigation are some of the measures of the technical, economic, and environmental performance of natural gas storage. Selection of the gas storage type, location, and operation mode is a delicate and strategic engineering issue. Dictated by the supply, demand, and market price trends, gas storage facilities can be operated at different gas injection and withdrawal modes. Simultaneous injection and production as well as consecutive injection and production modes are often resorted to in this practice. Hence, various gas storage facilities can be broadly classified as having flexible or inflexible storage capabilities with respect to the ability of providing different storage services, gas deliverability, and yearly cycling frequency. Flexible storage (e.g., salt caverns) allows multiple cycles during a year and high injection and deliverability rates, accommodating widely varying intra- and interday gas rates. Gas can be withdrawn quickly from a flexible storage within as short a time frame as a week This makes it possible to have more than 12 cycles in a year. As a result, flexible storage can capture the price differentials quite effectively over short periods of time according to the trends in the market conditions. Therefore, flexible storage has a beneficial market value when it can take advantage of small price fluctuations, but it would not have an economic value under stable market conditions. In contrast, inflexible storage (e.g., a depleted reservoir and water aquifer) may be cycled only once a year, typically with injection during summer for storage and production during winter for consumption. Consequently, while providing an advantage of storing significantly large quantities of natural gas in the storage place, depleted reservoir and aquifer storages can capture only the long-term seasonal spreads in the gas market. Hence, flexible and inflexible gas storage facilities may have different response dynamics and economic values under certain market trends.

6. HAZARDS FROM NATURAL GAS TRANSPORTATION AND STORAGE

Natural gas transportation and storage systems can create various pollution and safety problems and should be monitored, inspected, and controlled carefully in order to avoid the escape and emission of dangerous natural gas. Gas storage facilities may undergo significant variation in the gas pressure and storage stress conditions as a result of frequent injection and production under different scenarios associated with various services. Frequent overpressuring and underpressuring of gas storage can induce aging of and various types of damage to the storage facility, creating and increasing the potential for gas leak and hazards.

From the point of gas storage, depleted reservoirs, aquifers, and salt domes often contain natural defects, including natural fractures and faults, which may provide escape routes for gas. Escape, migration, and accumulation near the ground surface of large quantities of gas from storage create potential hazardous situations, such as fire and explosion, and environmental pollution problems, including noxious odors and other health hazards. Nevertheless, gas leaks from underground storage facilities are natural and unavoidable, and should be monitored carefully in order to avoid potential health and safety hazards.

7. NATURAL GAS SECURITY AND RISK ISSUES

Natural gas supply and transportation in parts of the world may be limited and/or interrupted because of political instability and unrest and frequent changes in local governments, customs, and regulations. The security and risk issues in the post-9/11/2001 world are of great concern, and ensuring the accessibility and availability of natural gas reserves and the safe operation and economic viability of natural gas transportation and storage facilities are issues of importance. Liquefied natural gas tankers, facilities for on- and offshore natural gas liquefaction, regasification, and storage, and gas pipelines are particularly vulnerable to sabotage and other man-made damages. Attacks impacting any of the storage

components may lead to rapid release of large gas clouds and to dangerous fires and explosions. Implementation of carefully designed surveillance techniques can avert such problems, and security and emergency response plans and methods for detection and deterrence of unwanted activity inside and around gas transportation and storage facilities have become an integral part of the overall natural gas handling and management task. Effective security measures require the use of highly sophisticated tools and approaches, such as the development and application of vulnerability assessment models, security vulnerability assessment of gas handling facilities, properly selected and installed security equipment, and training of personnel about safety and emergency procedures.

SEE ALSO THE FOLLOWING ARTICLES

Coal Storage and Transportation • Natural Gas, History of • Natural Gas Industry, Energy Policy in • Natural Gas Processing and Products • Natural Gas Resources, Global Distribution of • Natural Gas Resources, Unconventional • Natural Gas Transportation and Storage • Occupational Health Risks in Crude Oil and Natural Gas Extraction • Oil and Natural Gas Drilling

Further Reading

Ammer, J. R., and Sames, G. P. (2000). "Advances for Improved Storage: Deliverability Enhancement and New Storage Facilities." SPE paper 65638. Proceedings of the Society of Petroleum Engineers (SPE) Eastern Regional Meeting, Morgantown, West Virginia, 17–19 October. SPE, Dallas, Texas.

Chabrelie, M.-F. (2000). "Natural Gas in the World—2000 Survey." IFP-CEDIGAZ, Institut Français du Pétrole (IFP), Paris.

Civan, F. (1989). "Review of Methods for Measurement of Natural Gas Specific Gravity." SPE paper 19073. Proceedings of the Society of Petroleum Engineers (SPE) Gas Technology Symposium, Dallas, Texas, 7–9 June. SPE, Dallas, Texas.

Civan, F. (2000). "Reservoir Formation Damage—Fundamentals, Modeling, Assessment, and Mitigation." Gulf Publ. Company, Houston, Texas.

Clark, J. (2003). Government, industry forge partnerships for security enhancement. *Oil Gas J.* **101.24,** 20–30.

Cornot-Gandolphe, S. (1995). "Underground Gas Storage in the World—A New Era of Expansion." IFP-CEDIGAZ, Institut Français du Pétrole (IFP), Paris.

Energy Information Administration (EIA). (2002). "The Basics of Underground Natural Gas Storage." US Department of Energy, Natural Gas Analysis Publications, Washington, D.C.

Fitzgerald, A., and Taylor, M. (2001). "Offshore Gas-to-Solids Technology." SPE paper 71805. Proceedings of the Offshore European Conference, Aberdeen, Scotland, 4–7 September. SPE, Dallas, Texas.

Groves, T. (2001). Terminal LNG tank management system promotes storage safety. *World Refining*, May 2001, 46–48.

Gudmundsson, J. S., and Mork, M. (2001). "Stranded Gas to Hydrate for Storage and Transport." 2001 International Gas Research Conference, Amsterdam, November 5–8. Gas Technology Institute, Des Plaines, Illinois.

Ikoku, C. U. (1980). "Natural Gas Engineering: A Systems Approach." PennWell Publ. Company, Tulsa, Oklahoma.

Javier, M. (2002). "Decision-Making Under Uncertainty With Application to Natural Gas Utilization Investments." Master of Science thesis. University of Oklahoma, Norman, Oklahoma.

Khilyuk, L. F., Chilingar, G. V., Endres, B., and Robertson, J. O., Jr. (2000). "Gas Migration, Events Preceding Earthquakes." Gulf Publ. Company, Houston, Texas.

Mathew, J. P. (2003). Liquified natural gas: What IS it all about? *PMA OnLine Magazine*, January 2003.

Natural gas hydrate—A viable GTL alternative. (2003). *Chem. Eng. Prog.* **99**(3), 17.

Perry, W. (2003). Marine transportation facilities face New US Security Regs. *Oil Gas J.* **101.24,** 32–34.

Rojey, A., Jaffret, C., Cornot-Gandolphe, S., Durand, B., Jullian, S., and Valais, M. (1997). "Natural Gas: Production, Processing, Transport." Institut Français du Pétrole (IFP), Paris.

Tek, M. R. (1996). "Natural Gas Underground Storage: Inventory and Deliverability." PennWell Publ. Company, Tulsa, Oklahoma.

Tobin, J. (2001). "Natural Gas Transportation—Infrastructure Issues and Operational Trends." Energy Information Administration, Natural Gas Division, U.S. Department of Energy, Washington, D.C.

Tobin, J., and Thompson, J. (2001). "Natural Gas Storage in the United States in 2001: A Current Assessment and Near-Term Outlook." Energy Information Administration, U.S. Department of Energy, Washington, D.C.

U.S. Environmental Protection Agency (EPA). (1999). "National Emission Standards For Hazardous Air Pollutants: Oil and Natural Gas Production and Natural Gas Transmission and Storage, Final Rules." Federal Register, 40 CFR Part 63, Vol. 64, No 116, June 17. EPA, Washington, D.C.

Williams, D. (2001). "The Value of Gas Storage in Today's Energy Market." Technical paper, Solution Mining Research Institute (SMRI), Fall 2001 Meeting, 7–10 September, Albuquerque, New Mexico. SMRI, Richmond, Texas.

Net Energy Analysis: Concepts and Methods

ROBERT A. HERENDEEN
Illinois Natural History Survey
Champaign, Illinois, United States

1. Introduction and Background
2. Quantitative Energy Balance Indicators
3. Conceptual and Procedural Issues
4. Results and Trends
5. Concluding Comments

Glossary

absolute energy ratio (**AER**) Energy produced/energy required, including energy from the ground or the sun (dimensionless).
end use Consideration taking into account the efficiency of use of energy produced.
incremental energy ratio (**IER**) Energy produced/energy required, excluding energy from the ground or the sun (dimensionless).
internal rate of return (**IRR**) The implied interest rate if energy inputs are considered an investment and energy produced is considered a payoff (units = 1/time).
payback time The time for the accrued energy produced to balance the energy inputs (units = time).
power Energy per unit time.
power curve A graph of power inputs and outputs vs time.
premium fuel An energy type considered especially valuable. For example, oil is usually considered premium relative to coal because it is more flexible in use, less polluting, etc.

Net energy analysis of an energy technology is a comparison of the energy output with the energy needed to supply all inputs—the energy source, materials, and services—to construct, operate, and dispose of the technology.

1. INTRODUCTION AND BACKGROUND

At first glance, net energy analysis (NEA) is a natural, intuitively sensible extension of the idea of energy cost. Everything requires energy somewhere along the chain of production, including the equipment and services used to find and extract energy. In addition, energy seems to be harder to get with time: efforts to drill, dig, and dam must be in deeper and in more forbidding places for our oil, gas, coal, and hydropower (offshore for oil and gas, in mountainous terrain for coal, and into the sub-Arctic for hydropower). This in turn probably implies more energy cost, and shifts the time when this trend will hit the breakeven point. NEA seeks to find out. Some proponents of NEA claim that standard economics will not always indicate this problem (especially if subsidies to energy industries are involved, as they usually are) and that therefore net energy analysis should be carried out largely independent of economic analysis. However, the more common argument is that energy analysis is a supplement to, not a substitute for, economic analysis in decision making.

Net energy concerns peaked in the 1970s and early 1980s following the oil embargo/energy crisis years of 1973 and 1979–1980. Many NEA studies covered oil made from coal or extracted from tar sands and oil shale, geothermal sources, alcohol fuels from grain, biomass plantations, solar electricity from orbiting satellites, and nuclear electricity from ambitious proposed plant-building programs. In 1974, Federal legislation requiring net energy analysis of federally supported energy facilities was enacted. It required that "the potential for production of net energy by the proposed technology at the state of commercial application shall be analyzed and considered in evaluating proposals" [Public Law No. 93-577, Sect. 5(a) (5)]. Since that time, interest in net energy has diminished, but continues sporadically. Adherence to Public Law 93-577 was abandoned early by the U.S. Department of Energy (DOE). In 1982, responding to a criticism from the U.S. General Accounting Office that it was not performing NEAs, DOE said "[it is]

 © 2004 Elsevier Inc. All rights reserved.

DOE's view that the benefits of [NEA] are not worth the time and effort involved for general application." Net energy analysis is a type of life cycle analysis. The past decade has seen an increase in applications of the latter, but usually these have been applied to consumer products rather than to energy facilities, using toxic and other materials, water, etc., rather than energy, as numeraire.

The DOE had a point; the devil is in the details. NEA is an elusive concept subject to a number of inherent, generic problems that complicate its application. Some of these problems persist, not because they are unstudied, but because they reflect fundamental ambiguities that can be removed only by judgmental decision. The problems include all of those with energy analysis as well as others specific to the "net" question. The difficulty of specifying a system boundary, the question of how to compare energy produced and consumed at different times, and the problem of how to compare energy types of different thermodynamic qualities, density, and ease of storage all make NEA harder to perform and interpret. Some analysts therefore reject net energy analysis but support energy analysis. The latter considers the energy output to be a good or service and attaches no special significance to the fact that it is energy. In this view, it is reasonable and useful to determine how much energy is required to keep a room warm, or to produce a head of lettuce or even a ton of coal, but it is confusing at best and misleading at worst to compare outputs and inputs of a coal mine purely on the basis of energy cost. This fundamental objection attacks the very basis of NEA: NEA assumes that the economic/human life-support system can be separated into the "energy system" and the "rest of the system" and that studying the energy system as a distinct entity is valid. For many applications, this separation, although not perfectly defensible, seems to be acceptable. For others, there are stronger objections.

Little has occurred since 1982 to challenge DOE's dismissive pronouncement, and in this author's opinion NEA has very seldom, if ever, been used as a critical decision criterion. In the Grain Belt, NEA arguments have sometimes been quoted regarding ethanol production from corn, but the sorting out that occurred (with small operators using dry-milled corn shutting down while larger operators using wet milling flourished) can be understood on the basis of the monetary economies of scale, of the role of economic subsidies, and of the details of the flexibility of the two milling processes. In fairness, there is a large debt owed to NEA as a stimulant to proper thinking about resources. Predicting resource availability on the basis of past trends has been shown to be fallacious by the example of domestic U.S. oil, where increased drilling in "Project Independence" did not reverse the fact that all-time production peak was in 1971. To the extent that prediction by linear extrapolation is an economics technique and that prediction by incorporating the physical aspects of the resource is a net energy analysis technique, NEA deserves credit. But anything can be done poorly, and proper economics should incorporate physical realities. Once that is granted, the question of usefulness lies in the details of NEA.

In advance, it can be anticipated that two specific outcomes of NEA will be least controversial. If the technology appears to be a clear energy loser even after all uncertainties of technique and data are accounted for, the result is useful; the program should be scrapped. Likewise, if the technology appears an unambiguous energy producer, it is likely that NEA can now be deemphasized and the decision to proceed based on other criteria. If the technology is close to the energy breakeven point, NEA seems more appealing. Of course, in this case, a higher degree of accuracy in the result, and hence in the needed data, will be required.

2. *QUANTITATIVE ENERGY BALANCE INDICATORS*

The statement that an energy technology can be an energy winner or loser implies three assumptions. First, because energy is not created or destroyed (mass–energy in the case of nuclear energy), it would seem that at best an energy (conversion) technology can only break even. NEA usually assumes that the energy in the ground (or coming from the sun) is not to be counted as an input, i.e., is outside the system boundary. This is discussed in detail later. Second, it is usually assumed that high-quality energy (in the thermodynamic sense) is desirable, whereas low-quality energy is not. This would be covered properly if "free energy" were used instead of energy. The common (mal)practice is followed here, i.e., using "energy" actually to mean thermodynamic free energy. Third, it is assumed here that all material and service inputs to an energy facility are expressible in terms of the energy needed to produce them.

Figure 1 shows the basic NEA framework, assuming only one energy type. It is useful to have a normalized indicator of the net energy balance; two are the incremental energy ratio and the absolute

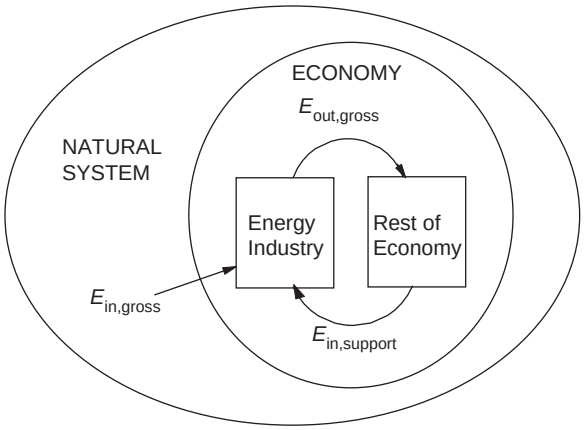


FIGURE 1 Energy flows for net energy analysis.

energy ratio:

$$\text{Incremental energy ratio} = \frac{E_{\text{out,gross}}}{E_{\text{in,support}}}$$

$$\text{Absolute energy ratio} = \frac{E_{\text{out,gross}}}{(E_{\text{in,gross}} + E_{\text{in,support}})}$$

Support energy is that obtained directly from the rest of the economy or used by the economy to provide necessary materials and service inputs. $E_{\text{in,gross}}$ is the energy from the ground or sun, that energy usually referred to as "resource." $E_{\text{out,gross}}$ is the energy supplied to the rest of the economy. The difference between the two ratios depends on the system boundary. The incremental energy ratio (IER) parallels standard economic practice in that the "cost" of exploiting a resource is the cost of extraction, and includes as inputs only those taken from the rest of the economy. The latter can be thought of as "invested" energy, and the IER is called by some workers "energy return on investment." An IER > 1 means that we have a net energy producer. The absolute energy ratio (AER) is appropriate to the more global question of the physical efficiency with which a given (say) fossil or renewable energy source can be converted into useful energy. The AER (which never exceeds 1) is useful for determining how much of a given stock of energy resource can be used by the rest of the economy.

Figure 2 illustrates the interplay of the IER and AER. Because of the possibility of energy consumption within the boundary of the energy technology, knowledge of the IER does not uniquely specify the AER, and vice versa. To exhibit how the energy flows vary and cumulate over the facility's lifetime, we can use a "power curve," shown in Fig. 3. The power curve tells "everything" (assuming that the system boundary is properly stated, etc.) but is cumbersome. It is desired to have some summary indicator of the whole time path,

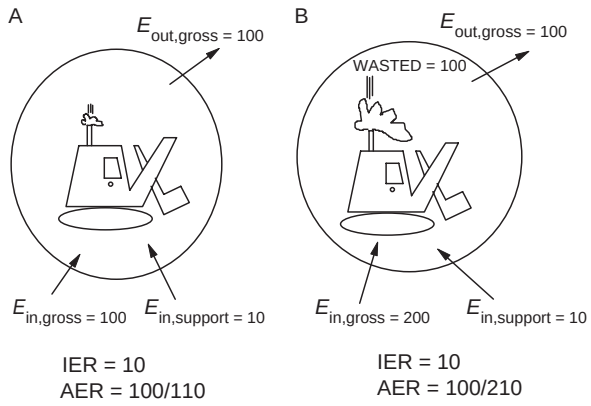


FIGURE 2 Incremental energy ratio (IER) and absolute energy ratio (AER) for two hypothetical coal mines. Both mines have the same IER, but mine B has a lower AER because of energy waste within the system boundary.

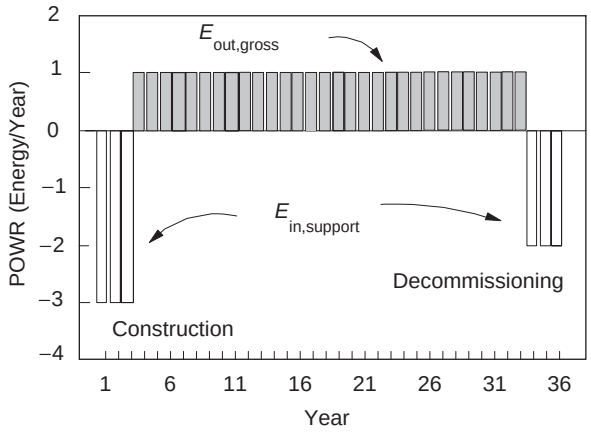


FIGURE 3 Hypothetical power curve for an energy facility through construction, operation, and decommissioning. IER = 30/[(3 × 3) + (2 × 3)] = 2. Payback time = 12 years, but it is not really "simple" because the concept does not apply to a power curve with costs following benefits. IRR = 9.1%/year.

such as (1) the energy ratio (already defined), (2) the energy payback time (the time for the accrued energy produced to balance the energy inputs), and (3) the energy internal rate of return (IRR; the implied interest rate if energy inputs are considered an investment and energy produced is considered a payoff). The formal statement is that the IRR is that value of r for which

$$\frac{\sum_{t=1}^{N} \frac{E_{\text{out,gross}_t}}{(1+r)^{t-1}}}{\sum_{t=1}^{N} \frac{E_{\text{in,support}_t}}{(1+r)^{t-1}}} = 1.$$

All three of these indicators have analogues in economics. All require summing energy over many years, and it is necessary to decide how to weigh the present against the future. The IRR by definition

explicitly assumes the standard economic practice of geometric discounting. However, for energy ratio and payback time there is no fixed rule. Most energy analysts do not discount, and that convention will be followed here. Summary indicators will be used for ease of communication but they sacrifice detail and introduce ambiguity. For example, consider simple payback time (i.e., payback time with no discounting), which is defined as that time at which the energy facility has produced as much energy as it has consumed. This definition implicitly assumes a power curve that is negative at first and then becomes, and remains, positive during facility operation. It would be less useful, and even misleading, for more complicated power curves, such as for a facility that requires significant energy inputs late in the lifetime, as for repair, maintenance, or decommissioning (see Fig. 3).

The energy ratio is also subject to difficulties in interpretation. For example, if an electric plant uses some of its own output, should that be added to the inputs or subtracted from the output? This makes no difference to payback time or to the internal rate of return, but does affect energy ratio. If the power curve is well behaved, i.e., negative at first and then positive over the whole lifetime, then the following statements are equivalent statements of positive net energy production:

- IER > 1.
- Simple payback time $<$ facility lifetime.
- Internal rate of return > 0.

3. CONCEPTUAL AND PROCEDURAL ISSUES

Several of the conceptual and procedural difficulties of NEA are summarized in Table I. It should be noted that there is considerable overlap between issues, and that several also plague economic analysis. Most are

TABLE I

Conceptual and Procedural Difficulties with Net Energy Analysis

Issue	Example	Remedied by careful statement of problem?	Comment
1. Specification of system boundary	Is sunlight "free"?	Yes	Differences of opinion persist among both analysts and users
a. Spatial	Is soil quality maintained in biomass energy production?		
b. Temporal	How much of a passive solar house is functional vs decorative?		
c. Conceptual	Should the energy cost of decommissioning a nuclear plant be included?		
	Should energy costs of labor be included?		
2. End-use consideration	Should NEA of a coal–electric plant be influenced by the efficiency of residential refrigerators? If a car gets better miles per gallon than expected, is the benefit allocatable to the ethanol-from-grain process?	No	An example of the fundamental question of to what extent the energy system is separable from the rest of the economy
3. Opportunity cost	Can energy conservation measures be thought of as producing (saved) energy and then subjected to NEA?	Yes	Difficult-to-set ground rules that are not situation-specific
4. Dynamic problem	Will a rapid buildup of many energy facilities produce net positive energy? The answer can be "no," even if the IER for each facility is >1	Yes	Raised regarding nuclear and solar energy in the 1970s but shown not to be not significant (rapid buildup did not occur)
5. Existence of more than one kind of energy	Should oil and coal be equally weighted in NEA?	Not completely	Very vexing
6. Average vs. marginal accounting	New oil wells often have a lower IER compared to old wells	Yes	Generic problem
7. Dependence of NEA on economic factors	Energy requirements for steel production depend on price of energy	No	Unavoidable because NEA is based on real technology

problems with energy analysis in general as well as with NEA specifically.

3.1 End-Use Considerations

Should NEA take into account the efficiency of use of $E_{out,gross}$ to provide an energy service such as cooling a space (e.g., a refrigerator) or moving people (e.g., an automobile)? An example of this problem is the gasohol miles-per-gallon question. It was observed in the late 1970s that cars using gasohol (a mixture of 90% unleaded gasoline and 10% ethanol) got more miles per gallon than would be expected on the basis of the enthalpy of combustion ("energy content") of the mixture as compared with that of pure gasoline. For the purpose of NEA, this could be viewed as a production of extra energy, i.e., an increase in $E_{out,gross}$. Should the credit for this increased output be allocated to the ethanol-producing process? If so, the IER increased from 1–1.7 to 1.5–2.4. Or, because the improvement is really due to the existence of the mixture, should it be inadmissible to perform the NEA of ethanol only? And what if the ethanol goes to an end use other than gasohol, for which this issue is moot? Further, this is all dependent on the type of engine. Research provides no definite answer. The user must decide, being attentive to the likelihood that in the event of controversy, different interest groups will tend to select the interpretation that best supports their cases.

3.2 Energy Conservation

NEA can be applied to energy conservation, in which case $E_{out,gross}$ is the energy saved ("neg-a-watts"), and $E_{in,support}$ is the energy required to make the insulation, the more efficient motor, etc.

3.3 The Dynamic Problem

This is an age–structure demographic issue, of concern with the rapid buildup of an energy technology. Even though a single facility may have a (lifetime) IER > 1, a program of building many facilities may be a net energy sink for many years. If the program is growing fast enough, the average facility is not even completely built! Concern over this problem was stimulated by exponential program goals for nuclear power plants (in the United States and Great Britain, doubling time $= 5$ years) and solar heating (United States, doubling time $= 3$ years). The exponential growth did not occur, and the concern has largely gone away. But it is always a potential problem.

3.4 Energy for Human Labor

There is disagreement on whether to include the energy to support human labor. The two following reasons argue against inclusion:

1. It is difficult to agree on how much of it to include. Is it the energy consequences of a worker's whole paycheck or just that portion spent for food and lodging?
2. Including labor means that the economy is viewed as producing goods only for government consumption, exports, and capital purchases. Consumption by residences is assumed to be fixed in amount and, especially, in "mix." Other energy research (for example, energy conservation) exploits the flexibility of personal consumption patterns (to purchase more insulation and less natural gas, for example). This flexibility seems to be preferable.

3.5 Different Energy Types and Premium Fuels

Not all energy types are equally useful. On average, we consider oil and natural gas more useful than coal, and the price reflects this (gasoline retails for about $10 per million Btu; coal is about $1.50 per million Btu). Various analysts have accounted for this by giving different weights to different energy types based on criteria such as scarcity or economic price. One example is calculating the IER for "premium" fuels only. Thus, if liquid fuels are defined as premium and an ethanol-from-grain plant can burn coal or crop residues, the IER will be increased over one that burns oil. This explicit introduction of nonenergy criteria again blurs the boundary of NEA.

3.6 Facility Lifetime and Other Sensitivity Issues

In Fig. 3 we see that if the energy facility has a productive life of only 15 years, instead of 30, the IER is reduced to 1, which is the breakeven point. Besides lowered lifetime, lack of reliability could cause diminished output and increased energy inputs for maintenance. Similarly, increased energy costs of decommissioning would reduce the IER; this is a real issue for nuclear power. Of course, the IER could be increased by the opposite of these trends. Explicit treatment of the uncertainty of IER is often performed. An example is an NEA of the proposed solar power satellite (SPS) of the late 1970s. This involved arrays of solar cells in geosynchronous orbit, on which electricity was

converted to microwaves and beamed to terrestrial antennas, for conversion back to electricity. Uncertainties of the lifetime and performance of photovoltaic cells in a space environment, plus relatively high energy requirements for heavy, crystal-grown silicon, which in turn required higher energy costs to place them in orbit, resulted in an IER of 2.1 ± 0.8 for the SPS. The project was not funded, though monetary issues probably dominated net energy questions. Today's solar cells are lighter and more efficient and durable, so that portion of the system would likely be more energetically favorable now.

4. RESULTS AND TRENDS

NEA has changed little since the spate of work in the 1970s and 1980s. A literature search today shows relatively few studies, and these tend to emphasize window coatings to manage solar gain and loss, a continued interest in solid biomass fuels and liquids-from-biomass, and photovoltaic units. Thus, the data in Table II, showing results for several conventional and unconventional technologies, are dominated by 25-year-old results. Table II indicates that coal tends to have IERs on the order of several 10s. Crude oil and natural gas are on the order of 10. Electricity from a fossil plant is on the order of 3 to 10. Geothermal electricity is of order 4 to 10. For solar technologies, we see that direct combustion of biomass has IERs of 3 to 20, with the energy cost of fertilization a dominant factor at the low end. Biomass liquids have IERs of 1 to 5. One study of ethanol from corn found an lER from 1 to 1.4 for total energy, which increased to 2.3–3.8 when calculated for premium fuel (liquids) only (another study 25 years later similarly found IER = 1.5). Solar space heat has IERs from 2 to 15. The high variation is partly due to climate, partly due to the boundary issue of how much of a solar house is the solar system (especially for passive solar), and partly due to the degree of coverage. The latter is another boundary

TABLE II

Selected Net Energy Results

Technology	Date	IER (unless otherwise stated)	Comment
Coal, U.S. average	1970s	37	
Eastern surface coal	1970s	43	Surface mining tends to have a higher IER compared to deep mining
Crude petroleum delivered to refinery	1970s	7	
Natural gas delivered through gas utility to user	1970s	11	
Coal mine-mouth electric power plant	1977	8	
Solar power satellite	1979	2.1 ± 0.8	Very uncertain; the IER for the photovoltaic cells alone would likely be higher today
Ethanol from corn for gasohol (gasoline:ethanol = 90:10)	1979, 1995	$\approx$1.5 to 2.5	Extremely sensitive to issues in Table I
Ceiling insulation as source of saved energy	1970s	136	Best case, previously uninsulated house
Geothermal–electric plant (vapor dominated)	1981	13 ± 4	Geothermal heat is not counted in the IER
Geothermal–electric plant (liquid dominated)	1981	4 ± 1	Geothermal heat is not counted in the IER
Short-rotation ($\approx$5-yr) trees for direct combustion	1980s	10 to 20 if unfertilized; 2 to 4 if fertilized	Unfertilized is likely unsustainable
Short-rotation biomass (trees, $\approx$5 yrs; grass, 1 yr) for direct combustion	1980s	6 if unfertilized; 2 if fertilized	Unfertilized is likely unsustainable
Residential solar space heat	1980s	2–15	Depends on climate and how well insulated the structure is (climate dependent); passive (no moving parts) tends to have higher IER compared to active
Residential solar hot water	1980s	2–5	Climate dependent
Photovoltaic cells	1976	Simple energy payback time $\approx$12 yr	
Photovoltaic cells	2000	Simple energy payback time = 1.1 to 4 yr	Payback time must be compared with device lifetime; this study concludes that the IER $\approx$30, implying a lifetime of at least 30 + yr

issue. A solar heating system that can supply year-round heat, including that one continuously overcast 5°F week in January, will require more inputs (such as more glass, thermal mass, and storage tanks). This system's IER will be lower than the IER for a system that is not expected to cover the worst periods—for which there is, say, a backup propane furnace that is not considered part of the solar heating system.

A similar comment applies to conservation as an energy source. The IER > 100 for ceiling insulation is for the first few inches of insulation in a totally uninsulated house. Putting that same insulation in a house already somewhat insulated would save less energy and give a lower IER. Also, photovoltaic cells have become better net energy producers over time as reliability has increased and material inputs have decreased.

5. CONCLUDING COMMENTS

Interpreting net energy information such as that in Table II almost instantly raises questions of the details and assumptions, which the critical reader must embrace. In spite of these, the concept (and the specter!) of the energy breakeven point and hence exhaustion is compelling enough for continued study. Perhaps the most vivid way to address this is by asking how IERs are changing over time.

SEE ALSO THE FOLLOWING ARTICLES

Aggregation of Energy • *Cost–Benefit Analysis Applied to Energy* • *Emergy Analysis and Environmental Accounting* • *Exergy Analysis of Energy Systems* • *Goods and Services: Energy Costs* • *Input–Output Analysis* • *Life Cycle Assessment and Energy Systems* • *Modeling Energy Supply and Demand: A Comparison of Approaches* • *Multicriteria Analysis of Energy* • *National Energy Modeling Systems*

Further Reading

Bullard, C., Penner, P., and Pilati, D. (1978). Net energy analysis: Handbook for combining process and input–output analysis. *Resourc. Energy* **1,** 267–313.

Chambers, R., Herendeen, R., Joyce, J., and Penner, P. (1979). Gasohol: Does it or doesn't it…produce positive net energy. *Science* **206,** 789–795.

Herendeen, R. (1988). Net energy considerations. *In* "Economic Analysis of Solar Thermal Energy Systems" (R. West and F. Kreith, Eds.), pp. 255–273. MIT Press, Cambridge.

Herendeen, R. (1998). "Ecological Numeracy: Quantitative Analysis of Environmental Issues." John Wiley and Sons, New York. See Chapter 8.

Herendeen, R., Kary, T., and Rebitzer, J. (1979). Energy analysis of the solar power satellite. *Science* **205,** 451–454.

Knapp, K., Jester, T., Milhalik, G. (2000). "Energy Balances for Photovoltaic Modules; Status and Prospects." 28th IEEE Photovoltaic Specialists' Conference, Anchorage, AK. Institute of Electrical and Electronic Engineers, New York.

Leach, G. (1975). Net energy: is it any use. *Energy Policy* **2,** 332–344.

Pilati, D. (1977). Energy analysis of electricity supply and energy conservation options. *Energy* **2,** 1–7.

Roberts, F. (ed.). (1978). "Symposium Papers: Energy Modelling and Net Energy Analysis." Institute of Gas Technology, Chicago.

Spreng, D. (1988). "Net Energy Analysis and the Energy Requirements of Energy Systems." Praeger, New York.

Whipple, C. (1980). The energy impacts of solar heating. *Science* **208,** 262–266.

Neural Network Modeling of Energy Systems

SOTERIS A. KALOGIROU
Higher Technical Institute
Nicosia, Cyprus

1. **Artificial Intelligence and Neural Networks**
2. **Biological and Artificial Neurons**
3. **Principles of Artificial Neural Networks**
4. **Network Parameters Selection**
5. **Modeling of Energy Systems Using Neural Networks**
6. **Conclusions**

Glossary

artificial intelligence The science devoted to the reproduction of the methods or results of human reasoning and brain activity.

artificial neural network A network of artificial neurons interacting with one another in a concerted manner that achieves "intelligent" results through many parallel computations without employing rules or other logical structures.

artificial neuron Element of a neural network that behaves like a neuron of the brain that possesses a number of input values to produce an output value.

backpropagation A method of supervised learning applied in multilayer feedforward neural networks. The error signal between the actual output and target is fed back through the processing layer, which tries to reduce this error by changing the connection weights.

biological neuron A brain cell that processes data and controls the response of the body. Neurons are connected to the brain and other nerve cells by transmitting connectors (axons) and receiving connectors (dendrites).

epoch A complete pass through the neural network of the entire set of training patterns.

hidden layer A neural network layer of nodes between input and output layers that contains the weights and processing data.

layers A group of neurons in a network. Layers can be input, output, or hidden.

pattern A single record or row of variables, also called an observation, that consists of inputs and outputs.

training In neural networks, training is synonymous to learning. During training, the system adapts in a specified manner so that some parts of the system suggest a meaningful behavior, projected as output.

weights Parameters for which their values are adapted during training in order to perform a task.

Artificial neural networks, or simply neural networks, are collections of small, individual interconnected processing units. Information is passed between these units along interconnections. An incoming connection has two values associated with it, an input value and a weight. The output of the unit is a function of the summed value. Neural networks, although implemented on computers, are not programmed to perform specific tasks. Instead, they are trained with respect to past history data sets until they learn the patterns presented to them. Once they are trained, new unknown patterns may be presented to them for prediction or classification.

1. ARTIFICIAL INTELLIGENCE AND NEURAL NETWORKS

For the estimation of the flow of energy and the performance of energy systems, analytic computer codes are often used. The algorithms employed are usually complicated, involving the solution of complex differential equations. These analytic programs usually require large computer resources and need a considerable amount of time to achieve convergence and thus to give accurate predictions. Instead of complex rules and mathematical routines, artificial neural networks are able to learn the key information patterns within a multidimensional information domain. In addition, neural networks are fault tolerant, robust, and noise immune. Because they are inherently noisy, data from energy systems represent good candidate problems to be handled with neural networks.

 © 2004 Elsevier Inc. All rights reserved.

The possibility of developing a machine that would "think" has intrigued humans since ancient times. In 1637, the French philosopher–mathematician Rene Descartes predicted that it would never be possible to make a machine that thinks as humans do. However, in 1950, the British mathematician and computer pioneer Alan Turing declared that one day there would be a machine that could duplicate human intelligence in every way.

Artificial intelligence (AI) is a term that generally indicates the ability of a machine or artifact to perform the same kinds of functions that characterize human thought. The term AI has also been applied to computer systems and programs capable of performing tasks more complex than straightforward programming but still far from the realm of actual thought. Thus, AI is the study of how to make computers do things similar to the way human beings do them. This term is often misinterpreted, and in order to avoid misunderstandings scholars usually employ the term computational intelligence.

It should be noted that solving a computation does not indicate understanding, something a person who solved a problem would have. Human reasoning is not based solely on rules of logic. It also involves perception, awareness, emotional preferences, values, the ability to generalize and weigh many options, evaluating experience, and many more.

Machinery can outperform humans physically. Similarly, computers can outperform mental functions in limited areas, particularly in the speed of mathematical calculations. The fastest computers are able to perform approximately 10 billion calculations per second. However, making more powerful computers is probably not the way to create a machine capable of thinking because computer programs operate according to set procedures, or logic steps, called algorithms. In addition, most computers do serial processing and operations such as those of recognition and computation are performed one at a time. The brain, however, works in a manner called parallel processing—that is, it performs a number of operations simultaneously. To achieve simulated parallel processing, supercomputers have been developed that have multiple processors that follow several algorithms at the same time.

During the past 15 years, there has been a substantial increase in interest in artificial neural networks. Neural networks are suitable for some tasks while lacking in some others. Particularly, they are good for tasks involving incomplete data sets, fuzzy or incomplete information, and for highly complex and ill-defined problems, for which humans usually decide on an intuitional basis. Additionally, they can learn from examples and are able to deal with nonlinear problems. They also exhibit robustness and fault tolerance. The tasks that neural networks cannot handle effectively are those requiring high accuracy and precision, such as in logic and arithmetic.

Neural networks have been applied successfully in various fields—engineering, mathematics, economics, medicine, meteorology, psychology, neurology, and many others. Some of the most important applications are in pattern, sound, and speech recognition; the analysis of electromyographs, cardiographs, and other medical signatures; the identification of military targets; the identification of explosives and other illegal substances in passenger suitcases; and the prediction of mineral exploration sites. They have also been used in market trends and weather forecasting, electrical and thermal load prediction, adaptive and robotic control, and many others. Neural networks are also used for energy systems process control because they can build predictive models of the process from multidimensional data routinely collected from sensors.

2. BIOLOGICAL AND ARTIFICIAL NEURONS

The basic building blocks of the nervous system are called neurons. A schematic of a biological neuron is shown in Fig. 1. Neurons are excitable cells composed of the cell body or soma (where the cell nucleus is located) and projections branching out from the soma. Each neuron has incoming fibers called dendrites and an outgoing fiber called an axon. Dendrites are input devices through which the neuron receives signals from other neurons. There is usually a single axon per neuron, and it serves as its primary output device. This may be connected to other neurons or may terminate in a muscle tissue, serving as a control signal for the muscle activity. The points of contact between an axon branch and a receiving neuron are called synapses. The change of synaptic transmission efficiency acts as a memory for past experiences. In the brain, there is a flow of coded information (using electrochemical media, the so-called neurotransmitters) from the synapses (having a density of approximately 10^4 synapses per neuron) toward the axon. The axon of each neuron transmits information to a number of other neurons. The neuron receives information at the dendrites from a large number of other neurons. It is estimated that

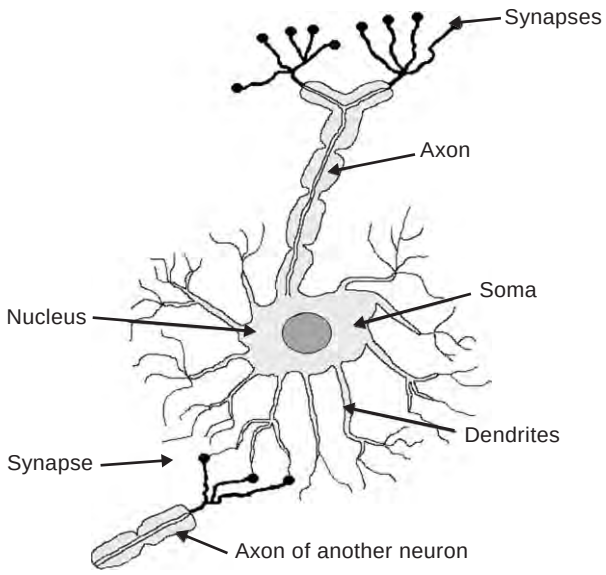


FIGURE 1 A simplified model of a biological neuron showing the soma with nucleus, the axon, dendrites, and synapses. The axon is the neuron output device and the dendrites are the input devices. The point of contact between an axon branch and a receiving neuron is called a synapse.

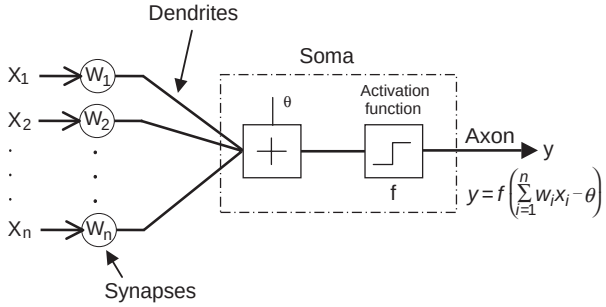


FIGURE 2 A simplified model of an artificial neuron used to model the features of the biological neuron. The synapses are the connection weights, the soma includes the summation and activation units, and the axon represents the output of the artificial neuron as in the biological counterpart. The unit step function is equal to 1 if $x \geq 0$ and is equal to zero otherwise.

each neuron may receive stimuli from as many as 10,000 other neurons. Groups of neurons are organized into subsystems and the integration of these subsystems forms the brain. It is estimated that the human brain has approximately 100 billion interconnected neurons.

Computing networks are far simpler than their biological counterparts. The first simplified neuron model was proposed by McCulloch and Pitts in 1943. Figure 2 shows a highly simplified model of an artificial neuron that may be used to stimulate some important aspects of the real biological neuron. The artificial neuron has all the features of a biological neuron (i.e., synapses, dendrites, soma, and one output line representing the axon of the neuron). The neuron is considered to be a binary threshold device with n inputs $x_1, \ldots, x_n$, one output y, a threshold level θ, and n synaptic weights $w_1, \ldots, w_n$. In such a system, excitation is applied to the input of the network. Following some suitable operation, it results in a desired output. At the synapses, there is an accumulation of some potential, which in the case of the artificial neurons is modeled as a connection weight. These weights are continuously modified during training, based on suitable learning rules.

3. PRINCIPLES OF ARTIFICIAL NEURAL NETWORKS

An artificial neural network is a group of interconnected artificial neurons interacting with one another in a concerted manner. It is in fact a massively parallel distributed processor that has a natural propensity for storing experiential knowledge and making it available for use. It resembles the human brain in two respects: The knowledge is acquired by the network through a learning process, and interneuron connection strengths known as synaptic weights are used to store the knowledge.

Artificial neural network models may be used as an alternative method in engineering analysis and predictions. They operate like a "black box" model, requiring no detailed information about the system. They imitate somewhat the learning process of a human brain because they learn the relationship between the input parameters and the controlled and uncontrolled variables by studying previously recorded data. In this way, they operate similarly to nonlinear regression, but they are much more powerful than regression analysis. Neural networks are able to handle large and complex systems with many interrelated parameters. They seem to simply ignore excess input parameters that are of minimal significance and concentrate instead on the more important inputs.

A schematic diagram of a typical multilayer feedforward neural network architecture is shown in Fig. 3. The network usually consists of an input layer, some hidden layers, and an output layer. In its simple form, each single neuron is connected to other neurons of a previous layer through adaptable synaptic weights. Knowledge is usually stored as a set of connection weights (presumably corresponding to synapse efficacy in biological neural systems).

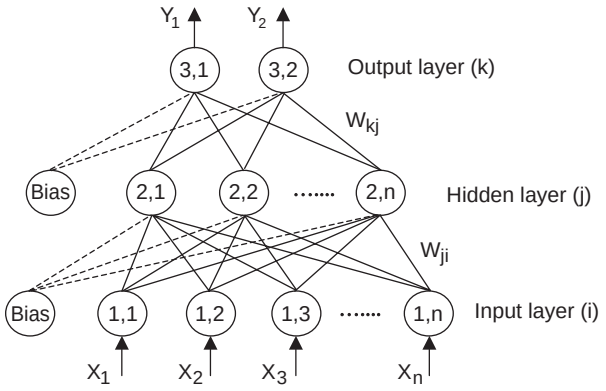


FIGURE 3 Schematic diagram of a multilayer feedforward neural network with an input layer (*i*), a hidden layer (*j*), and an output layer (*k*).

Training is the process of modifying the connection weights in some orderly fashion using a suitable learning method. The network uses a learning mode in which an input is presented to the network along with the desired output and the weights are adjusted so that the network attempts to produce the desired output. Before training, the weights are random and have no meaning, whereas after training they contain meaningful information.

Figure 4 illustrates how information is processed through a single node. The node receives weighted activation of other nodes through its incoming connections. These are first added (summation) and the result is passed through an activation function; the outcome is the activation of the node. For each of the outgoing connections, this activation value is multiplied by the specific weight and transferred to the next node.

A training set is a group of matched input and output patterns used for training the network. The outputs are the dependent variables that the network produces for the corresponding input. When each pattern is read, the network uses the input data to produce an output, which is then compared to the correct or desired output. If there is a difference, the connection weights (usually but not always) are altered in such a direction that the error is decreased. After the network has run through all the input patterns, if the error is still greater than the maximum desired tolerance, the neural network runs through all the input patterns repeatedly until all the errors are within the required tolerance. When the training reaches a satisfactory level, the network holds the weights constant and the trained network can be used to define associations in new input data sets not used to train it.

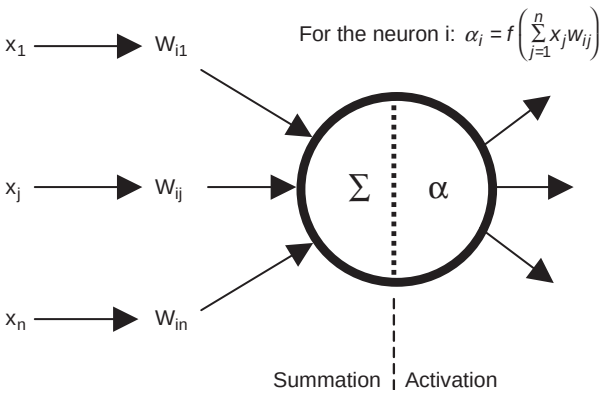


FIGURE 4 Information processing in a neural network unit.

By learning, it is meant that the system adapts (usually by changing suitable controllable parameters) in a specified manner so that some parts of the system suggest a meaningful behavior, projected as output. The controllable parameters have different names, such as synaptic weights, synaptic efficacies, and free parameters.

The classical view of learning is well interpreted and documented in approximation theories. In these, learning may be interpreted as finding a suitable hypersurface that fits known input/output data points in such a manner that the mapping is acceptably accurate. Such a mapping is usually accomplished by employing simple nonlinear functions that are used to compose the required function.

Generally, learning is achieved through any change in any characteristic of a network so that meaningful results are achieved. Thus, learning may be achieved not only through synaptic weight modification but also by modifications in network structure and through appropriate choice of activation functions, number of neurons in hidden layers, number of input parameters used, and others. By meaningful results, it is meant that a desired objective is met with a satisfactory degree of success. The objective is usually quantified by a suitable criterion or cost function. It is usually a process of minimizing an error function or maximizing a benefit function. In this respect, learning resembles optimization.

The information processing at each node site is performed by combining all input numerical information from upstream nodes in a weighted average of the form

$$\beta_i = \sum_j w_{ij}\alpha_{pj} + b_1, \tag{1}$$

where b_1 is a constant term referred to as the bias. The final nodal output is computed via the activation

function. There are many such functions, but the most common one is the logistic

$$\alpha_{(pi)} = \frac{1}{1 + e^{-\beta_i}}. \tag{2}$$

The most popular learning algorithms are backpropagation and its variants. The backpropagation algorithm is one of the most powerful learning algorithms in neural networks. Backpropagation training is a gradient descent algorithm. It tries to improve the performance of the neural network by reducing the total error by changing the weights along its gradient. The error is expressed by the root mean square value (RMS), which can be calculated by

$$E = \frac{1}{2}\left[\sum_p \sum_i \left|t_{ip} - o_{ip}\right|^2\right]^{1/2}, \tag{3}$$

where E is the RMS error, t is the network output (target), and o is the desired output vectors over all pattern (p). An error of zero indicates that all the output patterns computed by the artificial neural network perfectly match the expected values and the network is well trained. In brief, backpropagation training is performed by initially assigning random values to the weight terms (w_{ij}) in all nodes, where j is a summation of all nodes in the previous layer of nodes, and i is the node position in the present layer. Each time a training pattern is presented to the artificial neural network, the activation for each node, α_{pi}, is computed. After the output of the layer is computed, the error term, δ_{pi}, for each node is computed backwards through the network. This error term is the product of the error function, E, and the derivative of the activation function; hence, it is a measure of the change in the network output produced by an incremental change in the node weight values. For the output layer nodes and for the case of the logistic-sigmoid activation, the error term is computed as

$$\delta_{pi} = (t_{pi} - \alpha_{pi})\alpha_{pi}(1 - \alpha_{pi}). \tag{4}$$

For a node in a hidden layer,

$$\delta_{pi} = \alpha_{pi}(1 - \alpha_{pi}) \sum_k \delta_{pk} w_{kj}, \tag{5}$$

where k is a summation over all nodes in the downstream layer (the layer in the direction of the output layer), and j is the weight position in each node. The parameters δ and α are used to compute an incremental change to each weight term via

$$\Delta w_{ij} = \varepsilon(\delta_{pi}\alpha_{pj}) + m w_{ij(old)}, \tag{6}$$

where ε is the learning rate, which determines the size of the weight adjustments during each training iteration, and m is the momentum factor. It is applied to the weight change used in the previous training iteration, $w_{ij(old)}$. Both of these constant terms are specified at the start of the training cycle and determine the speed and stability of the network.

In the past few years, a large number and variety of commercial neural network programs have become available. These vary from user-friendly, step-by-step, easy to implement programs to general-purpose programs such as Matlab, in which the user must set up the neural network in the program environment and perform the necessary operations of training and testing of the network. In this case, the neural network component is a special module of the program (toolbox in Matlab) that can use the capabilities of the main program in order to input and manipulate data and produce graphs.

4. NETWORK PARAMETERS SELECTION

Although there are differences between traditional approaches and neural networks, both methods require preparing a model. The classical approach is based on the precise definition of the problem domain as well as the identification of a mathematical function or functions to describe it. However, it is very difficult to identify an accurate mathematical function when the system is nonlinear and there are parameters that vary with time. Neural networks are used to learn the behavior of the system and subsequently to simulate and predict this behavior. In defining the neural network model, first the process and the process control constrains have to be understood and identified. Then the model is defined and validated.

The neural network model and the architecture of a neural network determine how a network transforms its input into an output. This transformation is in fact a computation. The neural network architecture refers to the arrangement of neurons into layers and the connection patterns between layers, activation functions, and learning methods. Often, success depends on a clear understanding of the problem regardless of the network architecture. However, in determining which neural network architecture provides the best prediction, it is necessary to build a good model. It is essential to be able to identify the most important variables in a process and generate best-fit models. How to identify and define the best model is very controversial.

When building the neural network model, the process has to be identified with respect to the input and output variables that characterize the process. The inputs include measurements of the physical dimensions, measurements of the variables specific to the environment, and equipment-controlled variables modified by the operator. Variables that do not have any effect on the variation of the measured output are discarded. These are estimated by the contribution factors of the various input parameters. These factors indicate the contribution of each input parameter to the learning of the neural network and are usually estimated by the network, depending on the software employed.

The selection of training data is vital in the performance and convergence of the neural network model. An analysis of historical data for identification of variables that are important to the process is important. Plotting graphs to determine whether the charts of the various variables reflect what is known about the process from operating experience and for discovery of errors in data is very helpful.

The first step is to collect the required data and prepare them in a spreadsheet format with various columns representing the input and output parameters. If a large number of sequences/patterns are available in the input data file, and to avoid long training times, a smaller training file may be created to select the required parameters and the complete data set can be used for the final training.

Three types of data files are required: a training data file, a test data file, and a validation data file. The former and the latter should contain representative samples of all the cases that the network is required to handle, whereas the test file may contain approximately 10–20% of the cases contained in the training file.

The training of all patterns of a training data set is called an epoch. The training set has to be a representative collection of input/output examples. During training, the network is tested against the test file to determine its prediction accuracy, and training should be stopped when the mean average error remains unchanged for a number of epochs. This is done in order to avoid overtraining, in which case the network learns perfectly the training patterns but is unable to make predictions when an unknown training set is presented to it.

The input and output values need to be normalized. All input and output values are usually scaled individually such that the overall variance in the data set is maximized. This is necessary because it leads to faster learning. The scaling that may be used is either in the range -1 to 1 or in the range 0 to 1 depending on the type of data and the activation function used.

When using a neural network for modeling and prediction, the following steps are crucial. First, a neural network needs to be built to model the behavior of the process. Second, based on the neural network model built, the output of the model is simulated using different scenarios. Third, the control variables are modified so as to control and optimize the output.

The basic operation that has to be followed to successfully handle a problem with neural networks is to select the appropriate architecture and the suitable learning rate, momentum, number of neurons in each hidden layer, and the activation functions. This is a laborious and time-consuming method, but as experience is gathered some parameters can be predicted easily, thus shortening tremendously the time required.

In backpropagation networks, the number of hidden neurons determines how well a problem can be learned. If too many are used, the network will tend to try to memorize the problem and thus not generalize well later. If too few are used, the network will generalize well but may not have enough "power" to learn the patterns well. Determining the correct number of hidden neurons is a matter of trial and error since there is no science to it.

5. MODELING OF ENERGY SYSTEMS USING NEURAL NETWORKS

Neural networks have been applied successfully in a number of applications. The following are some of the most important:

Function approximation: In this type of problem, mapping of a multiple input to a single output is established. Unlike most statistical techniques, this can be done with adaptive model-free estimation of parameters.

Pattern association and pattern recognition: Neural networks can be used effectively to solve difficult problems in this field, such as in sound, image, or video recognition. This task can even be performed without an a priori definition of the pattern in which the network learns to identify totally new patterns.

Associative memories: This is the problem of recalling a pattern when given only a subset clue. In these applications, the network structures used are

usually complicated, composed of many interacting dynamical neurons.

Generation of new meaningful patterns: This general field of application is relatively new. Some claims have been made that suitable neuronal structures can exhibit rudimentary elements of creativity.

Neural networks have been used by various researchers for the modeling of energy systems. This section presents various such applications in a thematic rather than a chronological or any other order.

5.1 Renewable Energy Systems

In the field of renewable energy, neural networks have been used for the modeling of various components of a concentrating solar collector and for the prediction of the mean monthly steam production of a solar system. They have also been used to model the starting up of a solar steam generator, which is a particularly difficult problem because the system is working in transient conditions. In all these models, experimental data were used to train the neural networks, which were used subsequently to provide predictions for unknown cases.

Neural networks have also been used for the modeling of solar domestic water heating systems and for long-term performance prediction of the systems. In these models, performance data for 30 systems, obtained by testing the systems with the procedure suggested in international standards, together with weather conditions and some physical characteristics of the systems such as the collector area, storage tank volume and heat loss coefficient, tank type (vertical or horizontal), and type of system (open or closed) were used to train the neural network. A large variety of weather conditions were used to enable the network to accept and handle unusual cases. The trained model could be used to provide the useful energy extracted and the increase in storage tank water temperature of other unknown systems using only their physical characteristics and typical weather conditions, which reduces the experimentation required.

Other applications of neural networks in this field include the identification of the time parameters of solar collectors, photovoltaic systems, and solar radiation and wind speed prediction, which are the driving sources of renewable energy systems.

5.2 Heating, Ventilating, and Air-Conditioning Systems

In the field of heating, ventilating, and air conditioning, neural networks have been used for the modeling of houses and the estimation of the heating and cooling load, for the prediction of the air flow and pressure coefficients of naturally ventilated rooms, for the prediction of the energy consumption of passive solar buildings, and for the energy consumption optimization. The data required for the training of the neural networks were obtained from experimental measurements in actual buildings. Although most of the previously mentioned models were performed for a particular building construction, these can be developed to have a more general usefulness provided that the training databases are enriched with a variety of building constructions, orientations, and weather conditions. In such cases, neural networks need to be retrained to "learn" the new patterns.

Other applications in this field include the use of neural networks for the energy consumption prediction in commercial buildings and modeling of a room storage heater. In these models, the neural networks were used as an expert system.

Neural networks have also been used to control specific systems such as the temperature in a bus air-conditioning system according to the ambient temperature, number of passengers, and time of day. Neural networks can also be used for the modeling of refrigeration systems. In one such application, neural networks were used to model the amount of frost on the coil of a refrigerator, and based on the network results a demand defrost method was proposed.

5.3 Combustion and Incineration

In combustion, neural networks have been used mainly for the modeling and prediction of the combustion emission products. In particular, they have been used for the monitoring of combustion pollutant emissions, for the prediction of the gaseous emissions from a chain grate stocker boiler, and for the estimation of NO_x emissions in thermal power plants. They have also been used for the development of controllers for combustion systems. Applications include the predictive control of unstable combustion systems, process control for fluidized bed combustion, and optimal control of an industrial stoker-fired boiler.

For the latter application, data collected during 300 h of experiments on an actual 0.75-MW chain grate stocker boiler were used to train two neural networks. The control strategy, shown in Fig. 5, is to provide the minimum airflow required for combustion without incurring unacceptably high CO emissions and carbon in ash losses, thus improving the combustion efficiency. The system detects changes in load required and gives the matching coal feed and

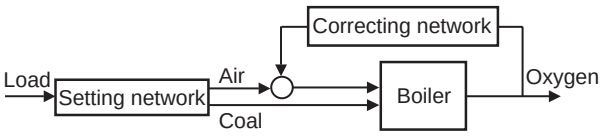


FIGURE 5 Chain grate stocker-fired boiler controller strategy.

airflow. Initially, the "setting network" acts as a look-up table to provide near optimum settings of coal feed and airflow for the desired load. After quasi-steady-state conditions are reached, the "corrective network" is activated to fine tune the airflow rate in order to optimize the combustion process on the grate.

Neural networks have also been used for the modeling of the combustion process in incineration plants, modeling of the temporal evolution of a reduced combustion chemical system, representation of the chemical reactions of combustion flames, and turbulent combustion modeling. In all the previously mentioned models, past history experimental data were used for the training of the neural networks to learn the required behavior and, according to the application, to provide the expected emissions so as to take corrective actions or to control a processes.

5.4 Internal Combustion Engines

Neural networks have been used to design engine controllers, such as a controller for idle speed regulation in internal combustion engines, a knock detector, spark advance control using cylinder pressure, and the control of combustion parameters such as spark ignition timing of petrol engines.

Neural networks have also been used for engine modeling, such as for the nonlinear time series analysis of combustion pressure data, modeling of the emissions and performance of a heavy-duty diesel engine, modeling of the volumetric efficiency and performance prediction of internal combustion engines, and modeling of a gas engine. In these models, experimental data were used for the training of the neural networks and the developed models are applicable to the particular type of engine used to collect the data. More general models can be produced by combining data from a number of engine capacity sizes and retraining the networks in order to produce general-purpose tools for the modeling of these engines.

Neural networks have also been used for the fault diagnosis of small to medium-sized diesel engines and marine diesel engines by providing an early warning of combustion-related faults. In a marine diesel engine model, the neural network was used to classify combustion quality on the basis of simulated data.

5.5 Energy Prediction and Forecasting

In the field of forecasting and prediction, neural networks have been used in power systems load forecasting and in electric power forecasting. They have also been used in tariff forecasting and energy management and in electric load forecasting in smaller capacity units such as supermarkets. The neural network models were trained with a sequence of past data recorded at the utilities and can provide either short-term or long-term predictions. Such systems can be used as effective management tools to regulate the operations of power utilities and other smaller enterprises.

6. CONCLUSIONS

From the applications described previously, it can be seen that neural networks have been applied for modeling and prediction in a wide range of fields of energy systems. Required for setting up such a neural network system are data that represent the past history and performance of the real system and a suitable selection of a neural network model. The selection of the model is usually done empirically after testing various alternative solutions with respect to network architecture and parameters. The performance of the developed neural network model is tested with unknown data collected from a real system that were not used for the training of the network.

The number of applications presented here is neither complete nor exhaustive but merely a sample that demonstrates the usefulness of artificial neural networks for modeling and prediction of energy systems. Artificial neural networks, like all other approximation techniques, have advantages and disadvantages. There are no rules regarding when this particular technique is more or less suitable for an application. Based on work published to date, neural networks offer an alternative modeling method that should not be underestimated.

SEE ALSO THE FOLLOWING ARTICLES

Complex Systems and Energy • *Computer Modeling of Renewable Power Systems* • *Fuzzy Logic Modeling of Energy Systems* • *System Dynamics and the Energy Industry*

Further Reading

Haykin, S. (1994). "Neural Networks: A Comprehensive Foundation." Macmillan, New York.

Kalogirou, S. (2000). Applications of artificial neural networks for energy systems. *Appl. Energy J.* **67,** 17–35. [Special issue].

Kalogirou, S. (2001). Artificial neural networks in renewable energy systems: A review. *Renewable Sustainable Energy Rev.* **5,** 373–401.

Ripley, B. D. (1996). "Pattern Recognition and Neural Networks." Cambridge Univ. Press, Oxford.

Rumelhart, D. E., and McClelland, J. L. (Eds.). (1986). Learning internal representations by error propagation. *In* "Parallel Distributed Processing: Explorations in the Microstructure of Cognition." MIT Press, Cambridge, MA.

Warwick, K., Ekwue, A., and Aggarwal, R. (1988). "Artificial Intelligence Techniques in Power Systems." Institution of Electrical Engineers, London.

Nongovernmental Organizations (NGOs) and Energy

BONIZELLA BIAGINI
Global Environment Facility
Washington, D.C., United States

AMBUJ SAGAR
Harvard University
Cambridge, Massachusetts, United States

1. Introduction
2. NGOs and Energy
3. Conclusion

Glossary

nongovernmental organizations (**NGOs**) National and international nonprofit organizations as well as local membership organizations, often referred to as grassroots or community-based organizations, that represent the public interest (or the interests of groups of citizens).

The term nongovernmental organization (NGO), as used in this article, refers to national and international nonprofit organizations as well as local membership organizations that are often referred to as grassroots or community-based organizations. NGOs can basically be thought of as being a third sector—separate from the state and the market—that represents the public interest (or the interests of groups of citizens). In this context, NGOs fill gaps in services provided by governments and firms and also enhance the accountability and responsiveness to citizens of these two sectors.

1. INTRODUCTION

Organizational goals of nongovernmental organizations (NGOs) include empowering people socially, economically, or politically; serving underserved or neglected populations; and protecting the local or global environment. To fulfill these objectives, NGOs are involved in a range of activities, including delivering welfare and development services, monitoring, critiquing and advocating policies, and providing information to citizens (and even to policymakers).

Although such citizens organizations have been a part of the landscape of societies for centuries, there has been a marked and rapid growth of this sector during recent years. The number of NGOs in most countries, as well as those operating transnationally, has increased substantially during the past couple of decades, as have the functions of these organizations. Consequently, these NGOs are playing an increasingly prominent role in many aspects of domestic affairs (within industrialized and developing countries) as well as in the international arena, resulting in what Edwards and Hulme referred to as the "rise and rise of NGOs" and what Salamon referred to as a "revolution." Nowhere is this transformation more prominent than in the environment and development arena. Whereas fewer than 300 NGOs attended the Stockholm Conference on the Environment in 1972, approximately 1400 NGOs registered for the Earth Summit in Rio de Janeiro, Brazil, in 1992 and roughly 18,000 NGOs attended the parallel NGO forum. More recently, more than 8000 NGOs attended the World Summit on Sustainable Development (WSSD) in Johannesburg, South Africa, and approximately 30,000 NGOs participated in the related Civil Society Forum. By now, NGOs and their networks are central players on the landscape of environmental politics and action.

At the same time, the finances underlying NGO operations have also been bolstered markedly. For example, Edwards and Hulme estimated that in

 © 2004 Elsevier Inc. All rights reserved.

1975 only 0.7% of the total development aid given by Organization for Economic Cooperation and Development (OECD) countries was channeled to NGOs, whereas by 1985 NGOs received 3.6% of development aid and by 1993–1994 they received 5.0%—with this last figure corresponding to roughly $5.7 billion. Other sources of NGO funding could have accounted for another $3 billion by the early 1990s. Of course, these numbers pertain only to funds channeled by aid and development agencies. NGOs in industrialized and developing countries also raise money from other sources, including membership fees, private foundations, individual donors, sales of publications (e.g., *State of the World* by the Worldwatch Institute), and technical services (e.g., organizing conferences and technical consulting). For example, Greenpeace raised more than four-fifths of its $143 million income in 2000 through personal donations, and the U.S. NGO Natural Resources Defense Council (NRDC) has more than 500,000 dues-paying members (see subsections on Greenpeace and NRDC in Section 2.4).

Another indicator of the growing influence of NGOs is the increasing movement of their personnel into governmental positions. This is particularly true in Western European countries, such as Germany, where members of environmental organizations formed the Green Party that has participated in governing coalitions. There have also been notable examples in the United States under Presidents Clinton and Carter, where individuals from environmental NGOs were nominated for senior governmental positions. This movement is not only into the government, as there are many former government officials now working in NGOs.

Broadly speaking, NGO activities include the following:

- Identification and articulation of needs and concerns that are being overlooked, or that are not receiving sufficient attention, by the state and the market
- Research and analysis to examine specific issues as well as potential solutions
- Advocacy, information, and technical assistance provision (e.g., coalition building) to citizens, policymakers, politicians, media, NGOs, and other groups
- Agenda setting
- Policy development
- Policy change and alternative policy formulation
- Policy implementation
- Delivery and provision of services to citizens or groups of society
- Capacity building in other groups, individuals as well as state and market actors
- Monitoring and critiquing state and market actors and activities.

The scale of NGO operations can also vary from the local (as may be common for many grassroots organizations) to the national and global. Furthermore, to fulfill their function better, NGOs often establish a wide variety of formal and informal links with other NGOs, individuals, governments and intergovernmental organizations (IGOs), development agencies, corporations, and transnational issue networks. For example, NGO relationships with governments and IGOs can be such that the NGOs may assist these latter institutions, take on tasks that have been ignored by them, lobby them to do things differently, evaluate their performance, or even challenge them, and sometimes the same NGOs may exhibit different kinds of linkages on different issues. Another important facet has been the emergence of networks and coalitions of NGOs at the national and transnational levels that have been useful for mobilizing and leveraging resources of a number of groups with common agendas or ideologies.

1.1 Greenpeace

Greenpeace is an environmental nonprofit organization that was founded in 1971 and now has a presence in 41 countries across Europe, the Americas, Asia, and the Pacific. The Greenpeace organization consists of Greenpeace International (Stichting Greenpeace Council) in Amsterdam, Netherlands, and Greenpeace offices around the world. The organization has a large base of individual supporters; as of January 2002, 2.8 million people had taken out or renewed their financial memberships within the previous 18 months.

As with many other major environmental NGOs, Greenpeace is engaged in a number of campaigns, including climate change, through which it draws attention to major environmental problems and possible solutions to them. Given the intimate relationship between energy use and climate change as well as other environmental issues, many of Greenpeace's activities focus on energy issues and, in turn, the organization has had a significant impact on various components of the energy system.

Greenpeace is actively engaged in increasing the public profile of a range of energy-related issues,

including renewable energy, nuclear energy, oil exploration, and climate change. It prepares detailed reports, commissions studies, analyzes policies, and disseminates information to the public on such issues. It also campaigns against, debates, and engages with industry (as needed). For example, whereas Greenpeace has led campaigns to prevent oil exploration in remote areas by firms such as BP or to challenge Shell on Brent Spar (the oil platform in the North Sea), it has also worked with other firms to develop new refrigeration technology. It also seeks to influence governments in international forums, such as those on climate change, as well as in the domestic arena.

Greenpeace was instrumental in the development of Greenfreeze, the world's first climate and ozone safe technology. In 1992, it brought together scientists who had extensively researched the use of propane and butane as refrigerants with an East German company, DKK Scharfenstein (now Foron), to develop this technology. It also gathered tens of thousands of preliminary orders for this new refrigerator from environmentally conscious consumers in Germany. Greenfreeze refrigerators are now sold in Western Europe, China, and Latin America.

Greenpeace does not solicit or accept funding from governments, corporations, or political parties. It relies on the voluntary donations of individual supporters and on grant support from foundations. Greenpeace's net income in 2001 was 112 million Euros (nearly exclusively from grants and donations). It spent approximately 10 million Euros on its climate campaign that year.

1.2 Natural Resources Defense Council

NRDC was founded in 1970 and now has more than 500,000 members in the United States. The organization is active in a number of areas, including clean air, energy, and water; forests and wildlife; toxics; nuclear weapons and waste; and green cities. It uses scientific analysis, legal actions, public education, and targeted grassroots efforts to promote policy changes in all of these areas.

NRDC's efforts to ensure clean air include targeting local, state, and federal governments as well as using legal avenues to promote strict pollution standards and compliance with them. NRDC also uses market-based strategies to promote new, less-polluting technologies. It also aggressively promotes renewable energy sources, such as wind and sun, as well as energy efficiency improvements in appliances, business equipment, and buildings through new and existing laws, better regulations, and financial incentives. Recently, NRDC has been paying particular attention to deregulation and restructuring of the electric industry to ensure incorporation of incentives for clean energy and efficiency.

NRDC's "earthsmartcars" campaign aims to put cleaner cars on the road by focusing on the demand side (i.e., consumer demand) as well as on the supply side (i.e., new technologies). It advocates for laws and regulations governing vehicle pollution as well as for promotion of public transportation and local planning to reduce vehicle use. NRDC has also been active in targeting diesel-fueled trucks and buses due to the particular health and environmental hazards posed by them. NRDC was instrumental in the efforts to phase out lead from gasoline, reduce acid rain, and help win adoption of national efficiency standards for consumer appliances, all of which have had a major impact on the energy system in the United States.

NRDC also has a major effort on climate change focusing on greenhouse gas (GHG) reduction domestically and internationally.

Nearly three-quarters of NRDC's 2001 operating income of $40.6 million came from membership and contributions, with foundations accounting for most of the remainder (nearly $9 million).

2. NGOs AND ENERGY

The main path through which NGOs shape the energy system is by influencing policy processes of governments and IGOs (e.g., the World Bank) that have an impact on the energy system as well as on the pace and direction of technological innovation. These include the following:

- National energy policy frameworks, including regulations, taxes, and publicly funded research
- Local and national environmental policies such as air quality regulation and GHG abatement policies
- Global environmental regimes such as the climate convention
- Contextual policies such as trade policy, export credit, and bilateral and multilateral aid that have indirect influences on global energy supply and demand

NGOs also directly participate in efforts to enhance and improve the delivery of energy services by working on relevant technologies and products as

well as on their deployment and diffusion in energy projects.

2.1 NGO Influence on National Energy Policies

Over the past few decades, NGOs have had a significant influence on shaping national or regional energy policies in most industrialized countries as well as in many developing countries. In industrialized countries, there is by now an established tradition of (mostly environmental) NGOs lobbying for domestic environmental regulations, and energy is often a priority on the agenda. In developing countries, an increasing number of NGOs have played a leading role in promoting sustainable energy technology transfer as well as the recognition of local knowledge and "home-grown" solutions in the South, critiquing existing national or local energy policies and proposing alternative policies, and participating in the design, implementation, and monitoring and evaluation of energy-related development projects. The pathways to influence have also been different in various countries. NGOs, although often with limited resources, are generally quite adroit at working within the institutional, policy, and cultural contexts within their countries and exploiting opportunities to shape energy policies through various mechanisms, as illustrated by the following examples from various countries.

2.1.1 United States

Some of the largest environmental NGOs, such as NRDC, were formed when citizens organized to fight power plants and other energy projects. Energy issues consistently appear high on the agenda for many of the largest NGOs. For example, the Sierra Club, one of the largest U.S. membership NGOs, has for years been a strong opponent of oil drilling in Alaska and a strong proponent of more stringent fuel economy standards.

The ability of NGOs to use courts as a means of influencing energy policy has been particularly important in the United States, where citizen suits to protect the environment were first permitted by judicial decisions and then enshrined in legislation such as the Clean Air Act. Environmental litigation has been used to challenge procedural deficiencies, such as lack of notice and opportunity to participate in decision making, as well as substantive outcomes, such as whether regulations are sufficiently protective of the environment consistent with legislative requirements. Litigation has also been an important educational tool in that it often generates publicity and public awareness. For example, environmental groups went to court in 2002 seeking to identify participants in meetings on energy policy with the vice president in an effort to show undue influence by industry.

An example of the high impact of environmental litigation in the United States on energy policy is a lawsuit filed by NRDC in 1983 to contest a decision by President Reagan to oppose the implementation of a law that required energy efficiency standards for refrigerators, air conditioners, and other appliances. NRDC was joined in the lawsuit by several states, including California and New York, as well as other NGOs, including a national consumer organization. The lawsuit was decided in favor of NRDC and resulted in negotiations between the plaintiffs (NGOs and state governments) and appliance manufacturers on national standards that were ultimately adopted through national legislation in 1987. It has been estimated that the standards will save U.S. consumers $200 billion through reduced energy consumption and avoided investments in power plants. U.S. appliances, which had been among the least energy efficient in the world during the 1970s, are now among the best available.

Although U.S. NGOs receive financial support from members, resources provided by private foundations have been particularly helpful in many cases. This source of funding is generally not available in other industrialized countries or is restricted from supporting litigation. The impact of foundation support for NGO efforts related to energy policy was enhanced in the United States by the founding of the Energy Foundation in 1991. The Energy Foundation was created by the McArthur, Pew, and Rockefeller foundations to provide more strategic support for utility regulatory reform and other technical issues. (Since that time, it has also received support from other foundations.) This more coordinated funding facilitated a network of lawyers and activists, at the state level, who could develop a common strategy; share data, witnesses, and legal materials; and generally work much more effectively. The Energy Foundation was also able to identify gaps in NGO work and support the creation of new organizations such as the Regulatory Assistance Project (RAP) and the Renewable Energy Policy Project (REPP).

2.1.2 Europe

NGOs have influenced domestic energy policies in many Western European countries. In Italy,

Legambiente, a leading grassroots environmental NGO, has worked on energy issues in various spheres. Started with the intellectual support of a group of scientists, many of them associated with the University of Rome "La Sapienza," Legambiente was initially very active in the antinuclear campaign. In 1986, after the accident at Chernobyl, Legambiente and other Italian NGOs promoted a referendum that resulted in a nuclear moratorium. During the 1990s, Legambiente's scientific committee, as well as its legal office, supported the work of green parliamentarians who designed and proposed Laws 9 and 10, subsequently approved by the Italian Parliament in 1991. Laws 9 and 10 introduced the "avoided cost principle" for purchase of power from nonutility generators and created significant economic incentives to produce and distribute renewable energy.

In The Netherlands, the ratio of NGOs to population is the highest in the world. Dutch NGOs have engaged in a number of activities on domestic energy policy, including public awareness, financial and fiscal incentives, and demand-side management programs. For example, in 1997, the Dutch government, the World Wildlife Fund (WWF, a major international NGO), and energy distribution companies decided to coordinate their public information campaigns. In addition, the three stakeholders joined efforts in the "Dutch Green Electricity Project" aimed at introducing a new product: "green electricity." Consumers paid 20% more, and the energy utility, in partnership with WWF, guaranteed that that this amount of electricity came from wind, small-scale hydro, biomass, and solar sources.

In Central and Eastern Europe, energy and climate NGOs have been active since the early 1980s and have significantly increased in number and activities during subsequent decades. Terra Millennium III, engaged at both the domestic and international levels, successfully lobbied the Romanian government toward an early ratification of the Kyoto Protocol. Green Action in Croatia, the Energy Club in Hungary, the Energy Efficiency Program and the Renewable and Energy Savings Center in Czech Republic, and the Polish Foundation for Energy Efficiency in Poland are among many NGOs in the region working to promote energy efficiency, renewable energy, and GHG reduction through policy research, training and education, and joint venture development.

2.1.3 India

NGOs play a prominent role in public discourse on India's development activities, including numerous aspects of the energy sector. For example, NGO opposition to the country's policy on independent power projects (IPPs), and on specific IPPs (including Enron's Dabhol project) for environmental, social, or economic reasons, has contributed to the government's reexamination of its power policies. The long and ongoing protest by the Narmada Bachao Andolan against a major dam (the Sardar Sarovar), citing reasons of insufficient attention to resettlement and rehabilitation issues as well as overall long-term effectiveness of the project, forced the World Bank to withdraw its support from the project and has rendered impossible a business-as-usual approach on hydroelectric development within the country. The International Rivers Network (IRN) has helped to support and lead a discussion on these kinds of issues in other countries as well (see subsection on IRN in Section 2.4).

Environmental NGOs have also influenced indirectly aspects of energy policy in India. For example, the Center for Science and Environment released a major study on urban air pollution to highlight the urgency of tackling transport-related pollution in Delhi and other major cities in the country. It followed this up with a strong and sustained public campaign that helped to catalyze the transition (with the support and intervention of a determined and activist judiciary) toward cleaner public transport in these cities.

2.1.4 NGOs and Nuclear Energy

No energy technology has been the subject of more NGO focus and influence than has nuclear power. The development of nuclear power has been heavily subsidized, and its further refinement and deployment is supported by a well-financed industry lobby. Nevertheless, since around 1970, nuclear energy's expansion has been blocked and even reversed (in the form of plant closures) in response to public pressure in Europe, the United States, and Japan. As noted in the World Energy Assessment, many official projections of nuclear power in 2020 envision no more, and perhaps less, total output than is the case today. NGO campaigns, coupled with the devastating accident at Chernobyl, have had a major and lasting influence on public attitudes toward nuclear power in Europe. NGOs have engaged in demonstrations (sometimes violent) at plant sites throughout Europe and the United States to highlight concerns about plant safety and waste disposal. They have also used legal means to challenge licensing and construction of plants and have prepared detailed analyses critical of the economics of nuclear power. The result has

been a significant contribution to public skepticism about the technology, ultimately leading to a refusal to accept new plants and, in some cases, pressure to cancel plants under construction or even those already operating.

A further indication of NGO influence is the widespread perception that the future revival of the nuclear industry depends as much on changes in public attitudes as on actual changes in plant operation and safety. As the World Energy Assessment states,

> *Decisions on future nuclear power will be made largely at the political level rather than on narrow technical and economic grounds. Gaining broad public support … is not simply a matter of better educating the public on the issues, which is what many in the nuclear industry believe is needed most. The industry should also seek to better understand public concerns. The industry should also consider opening up the nuclear decision-making process to diverse interest groups, so that a well informed public could ensure that its concerns are addressed every step of the way.*

2.2 Role of NGOs in Shaping International Environment/Development Policy Frameworks

A major role of NGOs in shaping energy policy is through their influence on international policy frameworks that have implications for energy systems. These include frameworks that are both legally binding, such as the UN Framework Convention on Climate Change (UNFCCC), and nonbinding, such as the Agenda 21 (the sustainable development action plan that includes significant sections about energy policies).

Energy plans and policies, previously seen as domestic issues, now have a transnational dimension with the clear recognition that the way in which energy is consumed and produced affects the regional and global environment through phenomena such as acid rain and greenhouse warming. Hence, agreements have been negotiated at the regional level, such as the Convention on Long-Range Transboundary Air Pollution (LRTAP), as well as at the global level, such as the UNFCCC.

NGOs have played a pivotal role in these negotiations and in the domestic implementation of related energy policies. At the global level, this has occurred thanks to the inclusive policy of the UN system, which allows nongovernmental groups to participate as observers in international forums. Since 1945, NGOs have been involved in consultation arrangements with the Economic and Social Council (ECOSOC) of the United Nations. Notwithstanding many difficulties and constant debate for decades with some governments that were not in favor of public participation, NGOs have increasingly gained recognition, credibility, and consensus on their involvement in UN processes, to the extent that by the early 1990s ECOSOC already included approximately 1000 NGOs. The UN Conference on Environment and Development (UNCED), or Earth Summit, held in Rio de Janeiro in 1992 was an unprecedented event with respect to NGO participation in global environmental politics. A record number of NGOs participated in the UNCED process (more than 1420 NGOs were accredited) and had the opportunity to influence the outcome of the negotiations. The UNFCCC and Agenda 21 are, among the outcomes of the Rio summit, the most relevant with respect to energy. NGOs are now considered as a major force in environmental politics. At various levels, they have helped governments and other stakeholders to identify issues and set agendas, have provided policy analysis as well as normative and regulations criteria, and have carried out monitoring and implementation activities.

Over the years, the level of expertise of NGOs has been increasingly recognized, thanks also to credible and independent policy analyses and scientific reports. Many governments, especially in developing countries, now rely on NGOs' expert support to help negotiate global environmental treaties and to help translate such agreements into domestic (energy) policies.

Worldwide, most NGOs active in global environmental politics fall into one of three categories: international NGOs (e.g., Friends of the Earth, Greenpeace, WWF, Third World Network), national organizations and think tanks, or (more recently) small and local NGOs. The last ones are particularly helped by joining umbrella organizations that work on thematic issues such as the Climate Action Network (CAN) (see subsection on CAN in Section 2.4).

2.2.1 NGOs' Involvement in the UNFCCC

The UNFCCC was adopted by more than 160 countries at the Earth Summit in Rio de Janeiro, and this agreement forged the first international work plan to address climate change. The ultimate objective of the treaty is the "stabilization of greenhouse gas concentrations in the atmosphere at a level that would prevent dangerous anthropogenic interference with the climate system." In practical terms, this statement implies a significant drop in

GHG emissions that cannot occur without significant changes and reforms in energy policies and technologies worldwide, inevitably leading to complex and contentious negotiations among countries.

NGOs have interfaced with this negotiation process in a number of ways, including directly lobbying government delegates, presenting "position papers" to the climate convention secretariat and parties, making statements during the plenary and organizing thematic "side events," using the media to exert public pressure on the delegates, and providing technical, policy, and scientific advice to the delegations (as appropriate). Although it is generally difficult to quantify the impact of NGOs, their relevant contributions to the debate, including policy analysis, position papers, fact sheets, scientific research, and economic analysis of mitigation options in the energy sector, among other inputs, have been recognized by the UNFCCC secretariat as well as by governments, which are gradually including a larger number of NGOs in their official delegations.

For example, in 1991, the Center for Science and Environment (CSE), an Indian NGO based in Delhi, published a report called "Global Warming in an Unequal World," that expressed the frustration of both Southern governments and many NGOs about the lack of attention given to the equity issue in the climate deliberations and that forcefully injected the topic into the international discussions. CSE was also asked to advise the Indian government on climate change, and it continues to have a significant influence on the country's climate policies. On the other end of the policy process, ENDA Energy, the energy branch of ENDA Tier Monde from Africa, is helping to implement the UNFCCC in African countries (see subsection on ENDA Energy in Section 2.4).

During the climate negotiations, NGOs have found numerous ways in which to exert influence. CAN NGOs publish a daily newspaper, *ECO*, that reports and comments on the details of the negotiations and is considered by many negotiators and observers to be a key source of information, policy analysis, and humor. CAN representatives also organize a daily press conference and distribute press releases commenting on the status of the climate negotiations. A Canada-based NGO, the International Institute of Sustainable Development (IISD), also publishes a more journalistic report on the negotiations, the *Earth Negotiations Bulletin*.

Over time, NGOs have gradually become more effective at networking by working together with their counterparts from various parts of the world. Increased funding from foundations and from bilateral, multilateral, and private sources, combined with North–South joint fund-raising, has increased NGOs' ability to travel and ensure a continued and consistent presence during more than a decade of climate negotiations.

2.2.2 NGOs and Agenda 21

The influence of NGOs was even more relevant in the Agenda 21 than in the climate convention due to its non-legally binding nature. Agenda 21 is intended to set out an international program of action for achieving sustainable development during the 21st century. Agenda 21 is a report of more than 500 pages and comprising 40 chapters. Energy issues are explicitly spelled out in many chapters, although there is no single energy overview or focus reflecting political differences on this topic. In the section on changing consumption patterns, NGOs provided language proposing to improve production efficiency to reduce energy and material use and to minimize the generation of waste concepts reflected in the final text of Agenda 21. The text of Agenda 21 is a particularly positive example of a collaborative approach and mutual tolerance among governments and civil society, represented mostly by NGOs.

Agenda 21 also explicitly recognizes, in its chapter 27, the role of NGOs in implementing the agenda:

> *Nongovernmental organizations play a vital role in shaping and implementing participatory democracy. Their credibility lies in the responsible and constructive role they play in society. Formal and informal organizations, as well as grassroots movements, should be recognized as partners in the implementation of the Agenda 21. … Nongovernmental organizations … possess well-established and diverse experience, expertise, and capacity in fields which will be of particular of particular importance to the implementation and review of environmentally sound and socially responsible sustainable development.*

2.2.3 NGOs and Multilateral Development Banks

During the 1980s, NGOs launched a major campaign to reform the environmental policies of the multilateral development banks (MDBs). Such a targeted reform campaign on MDBs was based not only on the size of their loan portfolios but also on the fact that MDBs act as a catalyst for additional funding from the private sector and other banks.

Energy was a key issue in this campaign. The lending decisions of the World Bank and other MDBs often favored large-scale, capital-intensive, centralized projects (e.g., large dams, fossil fuel power plants) as opposed to more decentralized,

small-scale, renewable energy or energy efficiency technologies.

Because of the significant influence of the U.S. government on the World Bank and other MDBs, a coalition of U.S.-based NGOs launched a strategy based on lobbying the U.S. Congress to influence the MDBs' lending policies through an evaluation of their environmental practices and policies. The NGOs started their campaign by convincing members of the U.S. House and Senate and their staff to hold hearings on the environmental performance of the MDBs. The U.S.-based NGOs testified, together with colleagues from developing countries, on the environmental and social impacts of the projects, especially in the energy and agriculture sectors. Recommendations on new procedures for MDB lending decisions drafted jointly by NGOs and staff of the House Banking Subcommittee were proposed to, and accepted by, the Department of the Treasury.

During the late 1980s, the campaign expanded to include an increasing number of NGOs from developing countries, economies in transition, and Western Europe. Although the campaign had a tangible influence on environmental policies of the World Bank, within the energy sector it resisted shifting its landing policy from conventional energy sources (e.g., fossil fuels, large dams). In 1989, only 2% of the World Bank's energy portfolio included end-use energy efficiency components.

Many developing country NGOs gradually increased their participation in areas of action traditionally handled by Northern NGOs. Protest against international aid agencies has demonstrated to be a very successful means to raise media, public, and political attention. In the MDB campaign, monitoring and denouncing project activities in the field became a key element of the campaign, giving developing country NGOs a pivotal role to play. In some cases, thanks to the testimony of developing-country NGOs about negative local environmental and social impacts, Northern NGOs managed to stop or at least question the project funding.

The NGO initiative calling for a World Bank reform further expanded during the 1990s, especially in 1994 when Friends of the Earth International took the lead with the "50 Years Is Enough" campaign.

The relationship between MDBs and NGOs has also changed dramatically during recent years as a result of such campaigns. The World Bank and other aid agencies have included public participation and disclosure guidelines to give more access to information and participation at all stages of project design, implementation, monitoring, and evaluation, and formalized consultations are now sometimes done on the Internet. The ongoing NGO campaign that embraces a number of MDBs' initiatives is called the "Campaign for the Reform of the World Bank," where development groups, such as Both Ends, have a leading coordinating role.

In the context of aid agencies, it is also relevant to mention the interaction between the Global Environmental Facility (GEF) and NGOs. GEF has allocated $580 million in renewable energy during the past decade or so and has engaged in bilateral consultations with NGOs since its inception. GEF public participation policy consists of information dissemination, consultation, and stakeholder participation, including those potentially affected by GEF-financed projects. The GEF Small Grant Program also includes a project portfolio that involves NGOs as central proponents, designers, and implementing stakeholders.

2.2.4 NGOs and International Finance/Aid Flows

NGOs have recently started to examine the activities of export credit and investment insurance agencies that play a major role in financing infrastructure projects in developing countries. Many of these infrastructure projects are in the energy area and, hence, have important long-term climate as well as other environmental and social implications. Many export credit agencies (ECAs) have no formal policies to evaluate such implications of the projects they fund; hence, many NGOs are building on their past work on World Bank projects to direct attention to filling this gap.

Recent examples include advocacy (so far unsuccessful) in the United States to include renewable energy targets in legislation reauthorizing the Export–Import Bank's charter and budget, establishment by the Export–Import Bank of a renewable energy advisory committee (REAC) that was pushed by advocacy groups, and greater integration of the ECA issue into climate advocacy group efforts both within the context of climate negotiations and in the runup to the WSSD.

2.3 NGOs and Energy Service Delivery

Given the centrality of energy services in improving the human condition, numerous development NGOs have, not surprisingly, been involved in efforts to increase and improve the delivery of essential energy services to poor people in many developing countries.

Thus, the main thrusts of these activities have been on improving the availability and use of energy for basic human needs, such as lighting and cooking, as well as for enhancing health services and opportunities for basic economic development. Hence, NGO efforts have aimed to enhance, and make more reliable, supplies of electricity using renewable resources (biomass, microhydro, and solar photovoltaic [PV]) as well as to make available technologies for economically or socially productive uses of energy (e.g., biomass gasifiers for small-scale industrial applications, refrigeration for health clinics) and improved cooking stoves for reducing health-adverse impacts of traditional stoves.

The presence and sophistication of NGOs involved in such activities have grown significantly over the past few decades and have often compensated for the inadequacies of the public sector in many developing countries. As mentioned earlier, given the local nature of many NGOs, development agencies interested in improving energy service delivery have also recognized their value, as have many actors in the private sector. Thus, NGOs continue to play an increasing role in this arena.

NGO involvement in energy service delivery takes many forms, but these organizations have been particularly useful at overcoming some of the major barriers—technology, finance, information—that often impede efforts to increase access to energy in rural areas.

Effective implementation of small-scale energy projects in rural areas often hinges on the availability of appropriate energy technologies. Although specific components of these technologies may be available in the commercial markets, often the development of the overall system is dependent on local resources and constraints. This necessitates making careful choices about technology design, possibly carrying out modifications to existing designs or developing new designs or components altogether. NGOs can play a useful role in such technology modification or development processes.

This is well illustrated by the case of improved cooking stove programs where, to be successful, the designs of the new stoves have had to take into account the locally available energy sources, the preferences of the users in terms of stove characteristics (not only functional but sometimes also aesthetic), and the skills and materials available locally for stove manufacture. For example, the successful Kenyan *jiko* stove was based on a design from Thailand that was modified to make it useful and appealing to local users. The training and demonstration work that underlay the development of the jiko was carried out with the assistance of the Kenya Energy and Environment Organization (KENGO), an indigenous NGO. Similarly, in the case of the improved cooking stove program in Sri Lanka, the Intermediate Technology Development Group (ITDG), a U.K.-based NGO that has been involved in energy technology projects for many decades (see subsection on ITDG in Section 2.4), played an important role in working with local firms to identify appropriate production materials and processes for indigenous production of cooking stoves. As another example, The Energy and Resources Institute (TERI, formerly known as Tata Energy Research Institute), a major NGO in India (see subsection on TERI in Section 2.4), has been involved in the development of biomass-gasifier–based technologies for thermally productive uses such as silk dyeing and cardamom drying where the design of the system, to be successful, is very much dependent on the specific needs of the small-scale enterprises that are its eventual users.

The success of the Grameen Bank in Bangladesh highlighted both the importance of microcredit in rural areas and the feasibility of NGOs and community organizations in bridging the financing gap for poor people who are deemed credit risks by banks and other traditional lending agencies. The resulting microcredit revolution has had an impact on the energy sector also where rural users were often unable to purchase energy systems for lack of financing options. NGOs, often with the help of development agencies, have begun to play a role by helping to provide microfinance to individual users or local entrepreneurs for promoting energy access. For example, Grameen Shakti (GS), a subsidiary of the Grameen Bank, has been quite successful at promoting energy service delivery through the provision of microcredit and other services (see subsection on GS in Section 2.4).

NGOs also provide other kinds of assistance to entrepreneurs and other actors in the energy service delivery arena by providing training and building capacity in technical as well as business issues. They can also play a critical role through public education and dissemination of information about energy options. In some cases, NGOs can be involved in the actual delivery of energy services. And in some of these cases, they may have functions similar to those of private energy service companies (ESCOs), although presumably maintaining their nonprofit character.

2.4 Illustrative NGOs

2.4.1 The Climate Action Network

CAN is a global network of more than 330 environment and development NGOs from some 80 countries working to promote equitable measures to limit human-induced climate change to ecologically sustainable levels. CAN is the largest recognized environmental nongovernmental network through which environmental groups coordinate on climate and energy policy at the national and international levels. CAN members work to achieve this goal through the coordination of information exchange and NGO strategy on international, regional, and national climate issues. CAN has seven regional coordinating offices that coordinate these efforts in Africa, Central and Eastern Europe, Europe, Latin America, North America, South Asia, and Southeast Asia. Diverse environmental organizations from around the globe, ranging from large international groups, such as WWF, Greenpeace, and Friends of the Earth, to small local groups in developing countries, such as Terre Vivante in Mauritania and the Green Coalition in The Philippines, work collaboratively within CAN. Its members place a high priority on both a healthy environment and development that "meets the needs of the present without compromising the ability of future generations to meet their own needs." CAN's vision is to protect the atmosphere while allowing for sustainable and equitable development worldwide.

Established in March 1989, CAN has been actively monitoring and seeking to influence the climate negotiations as well as climate-related policies and measures at the national and international levels. Member organizations include noted experts on issues of climate change science and policy. CAN is the recognized umbrella NGO in the international negotiations through which environmental groups work. CAN's credibility has been built over years of involvement at both the international and national levels. The organization's flexible structure provides an NGO forum to get together and share ideas, advocacy strategies, and information on climate change. Membership is open to nonprofit public interest organizations that are active on the global warming issue and that subscribe to the mission of CAN.

CAN's major objectives include assisting the global community in reaching the goals of the climate convention, promoting awareness and capacity building of governments and citizens worldwide on the climate issue, developing an objective understanding of climate change and its causes, sharing this information with all people (especially all members), and organizing, supporting, inspiring, and coordinating its members to take effective action on climate change.

2.4.2 ENDA Energy

ENDA Energy is a branch of the organization ENDA Tiers Monde (Environnement et Développement du tiers-monde), an association of autonomous entities coordinated by an executive secretariat, and collaborates with grassroots groups in search of alternative development models on the basis of the experience, expectations, and objectives of marginalized peoples.

ENDA Energy's work focuses on energy use and management in the African context, with an emphasis on the linkages between energy and development. ENDA Energy is also helping with the implementation in Africa of the United Nations Conventions on Climate Change and Desertification, with both areas being intimately related to the energy issue in Africa. It is also involved in the development of alternative energy technology, collecting information on energy, and implementing projects on both local and regional levels. ENDA Energy often works in partnership with other groups and through networks and jointly led projects.

ENDA Energy's team has approximately 10 permanent members, engineers, economists, and environmentalists focusing on the knowledge and planning of energy systems, the promotion of alternative energy systems suitable for the African context, and the broader challenge of integrating energy with environment for sustainable development.

Although the program originally focused mainly on energy use by grassroots communities, over time it has become involved in maintaining databases, undertaking direct fieldwork and training, and lobbying African policymakers. ENDA Energy has also helped to set up an African network, RABEDE (African Network on Bioresources–Energy–Development–Environment), which circulates information, advice, and experience on these topics through a variety of mechanisms, including the *Bulletin Africain,* the journal of the RABEDE network.

2.4.3 Grameen Shakti

GS is a subsidiary of Bangladesh's Grameen Bank, a member-owned organization founded in 1976 that provides microcredit to its members. GS was founded in 1996, when only 16% of the 21 million households were covered by the national electric

grid, to help provide renewables-based energy to rural households in Bangladesh.

By mid-2001, GS had installed more than 5800 PV home systems, with an installed capacity of 290 kWp, and planned to install 8000 systems within the subsequent 3 years. GS provides soft financing for customers as well as maintenance of its system for a year after sales. GS provides training to customers so that they can take care of minor problems themselves, and its intention is bring all customers under this program. GS also provides training to develop skilled technician-cum-retailers in the rural areas who will be able to install solar systems, provide accessories and after-sales services to PV systems buyers, and educate the rural people in renewable energy technologies. GS has trained more than 300 PV technicians under this program. GS also engages in promotion and demonstration efforts to build awareness of solar PV technology and also encourages customers to develop income generation schemes.

Under its wind program, GS is developing micro-enterprise zones using wind-based electricity generation systems; it has installed at least six turbines so far. Under its biogas program, GS is promoting the use of biodigesters to produce clean-burning biogas from cow dung and other waste biomass. GS has installed 30 biogas plants in northern Bangladesh and also provided training to customers for producing biogas from their own resources.

GS also has a research program that explores ways in which to develop appropriate renewable energy technologies (RETs) and their uses, develop methods to popularize renewable energy systems and make them accessible to large number of people, innovate financial services to facilitate rapid expansion of RETs, and develop and fabricate solar home system accessories and components to reduce total system cost.

A number of organizations have provided financial support to GS, including the International Finance Corporation, the U.S. Agency for International Development, the Swedish International Development Agency, Grameen Trust, and Grameen Fund.

2.4.4 International Rivers Network

IRN was established in 1985 in the United States as a nonprofit, all-volunteer organization of activists experienced in fighting economically, environmentally, and socially unsound river intervention projects. IRN collaborates and coordinates with local river activists worldwide. It sees itself as part of a larger citizen-based movement working on these issues and believes that its legitimacy is to be derived from the respect of individuals groups all over the world with whom it works.

IRN has a two-pronged approach that combines work on changing global policies with campaigning on specific key projects around the world. Through this approach, its campaigns for policy change are informed by specific project examples while also recognizing that a project-only approach does not address the root causes and policies that underlie unsound river interventions.

IRN works with groups from around the world on cooperative campaigns for community-based river development. This has included campaigns on the Sardar Sarovar project in India, the Three Gorges project in China, and the Lesotho Highlands Water project.

IRN also undertakes in-depth research and provides project critiques, analyses of alternatives, and activist briefings. It monitors and critiques the policies of financial institutions, including the World Bank, and provides analyses and recommendations for reforming their practices.

IRN helped to establish and continues to coordinate the International Committee on Dams, Rivers, and People (ICDRP), an informal network of NGOs and peoples movements from 13 countries. ICDRP monitored the World Commission on Dams (WCD) and facilitated civil society input into the process, and it continues to advocate for the implementation of the WCD guidelines. IRN has also supported the formation of regional networks to promote progressive change in water and energy policies.

IRN's support and revenue for the year ended December 2000 amounted to $1.2 million, of which nearly 80% came from grants and approximately 14% came from contributions. Nearly 70% of its expenses were directed toward campaigns, and just over 7% of its expenses were directed toward information and media.

2.4.5 The Intermediate Technology Development Group

ITDG, an NGO registered in the United Kingdom, was founded in 1966 by E. F. Schumacher. It has a presence in Latin America, East Africa, Southern Africa, and South Asia and focuses particularly on Peru, Kenya, Sudan, Zimbabwe, Sri Lanka, Bangladesh, and Nepal.

In these countries, ITDG works with poor communities to develop and disseminate appropriate technologies in a number of sectors, including energy,

transport, and small enterprise development. In the energy area, ITDG's efforts focus on increasing poor people's access to energy technology options. It has done this through the development and commercialization of low-cost cooking stoves as well as small-scale, off-grid, sustainable energy supply technologies such as microhydro plants, small-scale wind generators, affordable solar lanterns, and biogas plants.

As an example, in Peru, ITDG has been promoting the dissemination of microhydro energy for the past 20 years. Its integrated approach has included technology development, training, pilot projects, research on institutional issues, and advocacy work. This approach allowed ITDG to overcome the main barriers to rural electrification in Peru, namely, lack of appropriate technology, financial mechanisms, local capacity, and legal frameworks. Through a program with the Inter-American Development Bank (IDB), ITDG has implemented a small hydro revolving fund consisting of soft loans and technical assistance for isolated communities. Between 1994 and 2000, there were 21 small hydro schemes (more than 1 MW total) benefiting 15,000 people with an allocation of $700,000 that, in turn, leveraged $2.5 million from local, regional, and central governments as well as other NGOs. ITDG also helped to open a training center to build local capacity and actively disseminated the results of its effort. This program won the CTI World Climate Technology Award in 2000.

Lessons from ITDG's grassroots experience are spread through consultancy services, publishing activities, education, policy, research, and an international technical inquiries service. Thus, dissemination is an explicit part of ITDG's focus.

ITDG's income for the financial year ended March 2000 was £12.9 million (U.S. $20.8 million at then prevailing exchange rates), of which £3.6 million was from donations by individuals, companies, and trusts; £6.5 million from grants from the U.K. government, multilateral development agencies, NGOs, companies, and trusts; and the remainder from its in-house consulting, publishing, and manufacturing activities. More than half of ITDG's expenditures were devoted to its technology programs.

2.4.6 The Energy and Resources Institute

TERI was established in 1974. Although during the initial period its focus was mainly on documentation and information dissemination activities, research activities in the fields of energy, environment, and sustainable development were initiated toward the end of 1982. With a staff of some 500 drawn from multidisciplinary and highly specialized fields, TERI is the largest developing country institution focusing on issues relating to energy and environment and is involved in an enormous range of activities in these areas. Not only does TERI play a prominent role in energy, environment, and other related issues in India, it also has an international presence through its international offices and affiliate institutes.

TERI engages in research and analysis, technology development, consultancy, training and capacity building, and information collection and dissemination. It also organizes numerous seminars and workshops that bring together various stakeholders and experts on assorted topics. It also set up the TERI Institute of Advanced Studies, a degree-granting institution focusing on energy and environment technology, regulation, and governance, among other things.

TERI is also involved in analyzing policy issues on various aspects relating to energy such as global environmental issues, oil and gas resources, urban and transport systems, and economic modeling and statistical analyses.

Through its energy–environment technology division, TERI is involved in the development and dissemination of energy technologies as well as efforts to promote energy conservation, biomass energy technologies, rural energy (development and implementation of renewable energy-based programs as well as building capacity for doing so), and chemical and hydrogen energy (with a particular focus on fuel cells).

TERI's revenues have grown rapidly during the past decade or so. During the year 2000–2001, it received approximately Rs. 170 million (~U.S. $3.7 million at then prevailing exchange rates) from international sources and Rs. 110 million (~U.S. $2.4 million) from domestic sources.

3. CONCLUSION

The energy sector is often a high priority on NGOs' agendas due to the increasing demand for energy services in both developed and developing countries and its high environmental and social impacts. Consequently, NGOs from many developed and developing countries have played a substantial and generally effective role in shaping international and domestic energy agendas.

Importantly, NGOs' influence on the energy system comes through their engagement in a range of activities—from grassroots activism, to policy research, to technology development and dissemination.

TABLE I

Some Examples of NGOs That Have Had an Influence on the Energy System and/or Its Potential Trajectory within Their Home Countries or Globally and Their Main Modes of Influence

	CAN	ENDA	GS	Greenpeace	IRN	ITDG	NRDC	TERI
Research	X	X	X	X	X	X	X	X
Advocacy	X			X	X		X	
Agenda setting	X	X		X			X	X
Policy development		X						X
Policy reorientation/alternative policy formulation	X			X			X	
Policy implementation		X						
Service delivery			X			X		X
Information provision		X		X	X		X	
Technical assistance			X			X		X
Capacity building		X			X	X		
Watchdog	X			X	X		X	

Through their bottom-up approach, NGOs have shown increasing expertise, technical knowledge, and capability to monitor the implementation of energy policies, highlight problems, and suggest solutions to governments, UN agencies, multilateral aid agencies, and other energy stakeholders. And on the other end of the spectrum, many NGOs have helped local and national governments, as well as other organizations, to design improved energy policies.

Although most NGOs are involved in a number of activities at the same time (as Table I illustrates), the emphasis of their efforts is generally of a limited set of core activities that are in line with the organizations' foci. However, there are significant synergies and complementarities between the efforts of NGOs that aim to influence the energy sector within their home countries and those that aim to do so internationally. Although NGOs may collaborate and network with other like-minded NGOs in the North and South (e.g., in the MDB reform campaign), they also form partnerships with government agencies, donor organizations, and academic organizations.

Given the range of energy-related activities in which NGOs are involved, their creative strategies for influencing the energy system, and their responsiveness to the needs of citizens and marginalized stakeholders, these organizations occupy an important niche in the energy field. Through their advocating, opposing, negotiating, and consulting activities, NGOs have contributed to strengthening the links among the general public and governments, aid agencies, and other key stakeholders that act as decision makers in the energy sector and to catalyzing change. These organizations, in turn, have begun to realize the value of NGOs as allies in the design, implementation, monitoring, and evaluation of energy projects. If anything, over time, the roles played by NGOs in contributing to an evolution toward an improved, effective, and sustainable energy system will only increase.

SEE ALSO THE FOLLOWING ARTICLES

European Union Energy Policy • Global Energy Use: Status and Trends • International Energy Law and Policy • Labels and Standards for Energy • Renewable Energy Policies and Barriers • Sustainable Development: Basic Concepts and Application to Energy • United Nations Energy Agreements • World Environmental Summits: The Role of Energy

Further Reading

Anheier, H., Glasius, M., and Kaldor, M. (eds.). (2001). "Global Civil Society 2001." Oxford University Press, Oxford, UK.

Edwards, M., and Hulme, D. (1992). "Making a Difference: NGOs and Development in a Changing World." Earthscan, London.

Lin, G. (2000). Energy development and environmental NGOs: The Asian perspective. *In* "The Global Environment in the Twenty-first Century: Prospects for International Cooperation" (Pamela S. Chasek, Ed.). United Nations University Press, New York.

Hurrell, A., and Kingsbury, B. (Eds.). (1992). "The International Politics of the Environment: Actors, Interests, and Institutions." Oxford University Press, Oxford, UK.

Keck, M. E., and Sikkink, K. (1998). "Activists beyond Borders: Advocacy Networks in International Politics." Cornell University Press, Ithaca, NY.

Princen, T., and Finger, M. (1994). "Environmental NGOs in World Politics." Routledge, London.

Salamon, L. M. (1994). The rise of the nonprofit sector. *Foreign Affairs* **73**(4), 109–122.

Smillie, I., Helmich, H., German, T., and Randel, J. (Eds.). (1999). "NGOs and Governments: Stakeholders for Development." Earthscan, London.

Willetts, P. (Ed.). (1996). "The Conscience of the World: The Influence of Nongovernmental Organizations at the United Nations." Brookings Institution, Washington, DC.

Nuclear Engineering

JOSÉ N. REYES, JR. and JOHN B. KING, JR.
Oregon State University
Corvallis, Oregon, United States

1. Nuclear Power and World Electricity Generation
2. Some Basics of Nuclear Science
3. Radiation Protection
4. Design of Nuclear Power Reactors
5. Commercial Nuclear Power Plants
6. Nuclear Power Plant Safety Systems
7. Next-Generation Nuclear Power Plants
8. Nuclear Waste Disposal
9. Conclusion

Glossary

boiling water reactor A direct-cycle power system that produces steam for a turbine-generator set by boiling water in a nuclear reactor core.

breeder reactor A nuclear reactor that breeds fissile fuel through neutron absorption in nonfissile isotopes.

containment building A large concrete- and steel-reinforced building that houses a commercial nuclear steam supply system.

passive safety systems A system that provides nuclear core or containment cooling by natural processes such as buoyancy, condensation, and evaporation or by a system of prepressurized tanks filled with coolant.

pressurized water reactor A power system that produces steam for a turbine-generator set by transferring heat from a nuclear core filled with high-pressure, subcooled water to an intermediate heat exchanger, known as a steam generator.

The methods developed in 1942 by Enrico Fermi to design, construct, operate, and control the first nuclear reactor formed the basis of a new discipline in engineering, nuclear engineering. It is the purpose of this article to introduce the reader to this exciting field. Today, nuclear engineers are active throughout the world, designing a new generation of nuclear reactors that can be safely used to produce electricity or medical isotopes or even to provide the power needed for deep-space exploration. However, the guiding principle for the nuclear engineer has remained the same, to find creative ways of safely harnessing and using the energy of the atom for the benefit of mankind.

1. NUCLEAR POWER AND WORLD ELECTRICITY GENERATION

> *I am going to tantalize you with a vision of vast supplies of energy far surpassing the wildest desires of the engineer-resources so illimitable that the idea of fuel economy is not to be thought of.*
>
> *—Sir Arthur Eddington, Berlin, 1930*

Sir Arthur Eddington's general address on subatomic energy at the 1930 World Power Conference in Berlin stirred the imagination of every scientist and engineer present. The challenge was clear: find a practical means of accessing, controlling, and using the enormous energy locked in the atom as predicted by Einstein's remarkable mass–energy relation, $E = mc^2$. On December 2, 1942, Enrico Fermi transformed Eddington's visionary challenge into reality by producing the world's first controlled, self-sustaining nuclear reactor, Chicago Pile 1. Six decades later, nuclear energy now produces 16% of the world's electrical power. As shown in Fig. 1, this amount compares to 40% for coal, 15% for natural gas, 10% for oil, and 19% for renewables such as hydro, solar, wind, and geothermal power. Although the percentages for nuclear power are relatively high in countries such as France (78%), Japan (35%), Germany (32%) and the United Kingdom (27%), the United States and Canadian nuclear shares are only 20% and 15%, respectively. Although the percentage is somewhat lower in the United States, it actually boasts the largest number of nuclear power reactors in the world. That is, 104 of the 441 nuclear power reactors in the world are operating in the United

 © 2004 Elsevier Inc. All rights reserved.

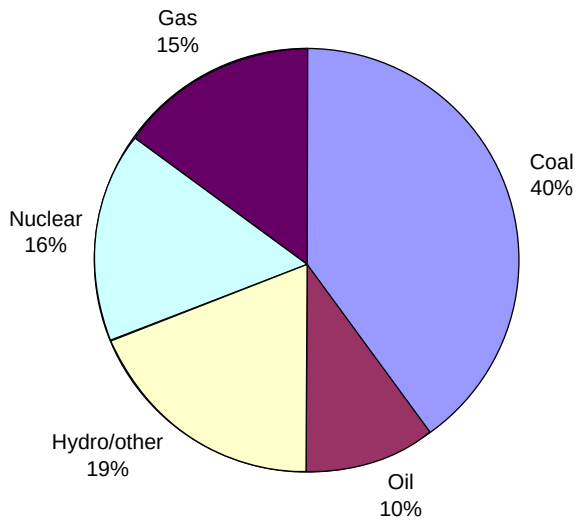


FIGURE 1 World electricity generation (statistics from the World Nuclear Association, World Energy Needs, and Nuclear Power, July 2002).

States. These reactors have a combined power output of 98.2 gigawatts (GW) of electricity, more than any other nation in the world.

2. SOME BASICS OF NUCLEAR SCIENCE

Before discussing nuclear engineering and the technology it has created, it is necessary to explain some basics of nuclear science. To this end, an atom is composed of negatively charged electrons that orbit around a positively charged nucleus. The nucleus, in turn, consists of positively charged protons and electrically neutral neutrons, both of which have masses approximately 1800 times greater than that of an electron. Thus, the preponderance of the atomic mass lies in the nucleus, which forms the tiny, central core of the atom.

Because the atom is electrically neutral, the number of electrons is identical to the number of protons, but the number of neutrons can and does vary. However, for any given chemical element, the number of protons is constant and serves to identify the element in question. For instance, all hydrogen atoms have 1 proton; all helium atoms have 2 protons, and all uranium atoms have 92 protons. Because the number of protons is uniquely linked to each element, this quantity is referred to as the atomic number Z. As already mentioned, however, for any given chemical element (i.e., for constant Z), the number of neutrons may vary. Thus, whereas all hydrogen atoms have a single proton and a single

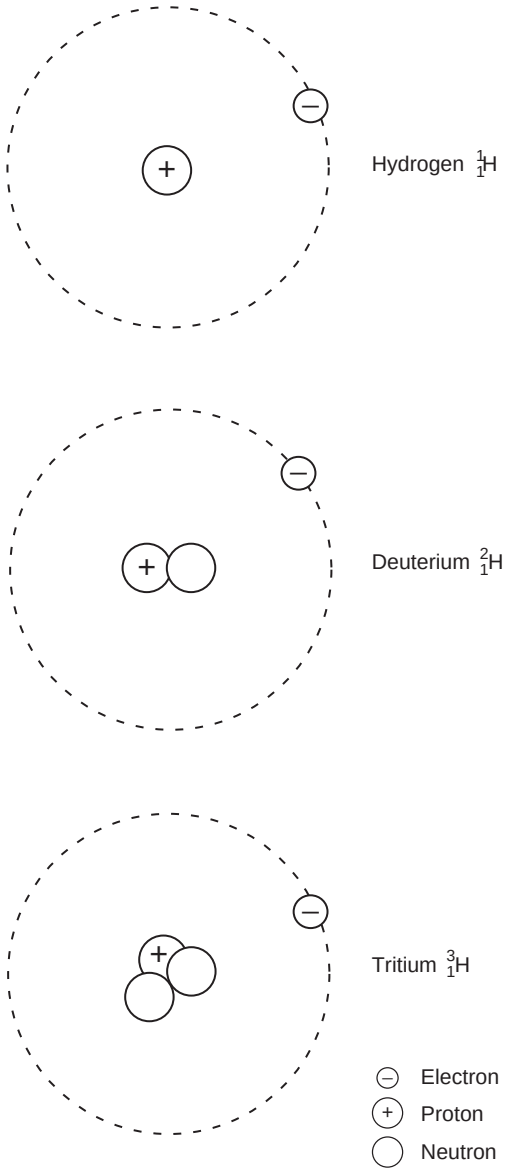


FIGURE 2 Isotopes of hydrogen.

electron, the number of neutrons varies from 0 to 2, as shown in Fig. 2. Given this fact, to uniquely specify a specific nucleus, it is necessary to specify both the number of protons (and hence the element) and the number of neutrons. Alternately, because the nuclear mass number A is specified by the total number of protons and neutrons, the nucleus can be uniquely specified by the combination of the atomic number Z and the mass number A. For instance, because all hydrogen atoms have a single proton ($Z = 1$), but can have 0, 1, or 2 neutrons, the mass number A will assume values of 1, 2, or 3, respectively. Accordingly, using the Z value for the subscript and the A value for the superscript, the

three types of hydrogen nuclei can be represented as $^{1}_{1}H$, $^{2}_{1}H$, and $^{3}_{1}H$. Moreover, because the Z value is implicit in the chemical symbol H, these nuclei may be represented more compactly as ^{1}H, ^{2}H, and ^{3}H, known as hydrogen-1, hydrogen-2 (or deuterium), and hydrogen-3 (or tritium), respectively. Finally, because these various nuclei all have the same Z value, they are all forms of hydrogen and will therefore occupy the same place in a periodic table of the elements. For this reason, they are referred to as isotopes, a word derived from the Greek terms *isos* (same or equal) and *topos* (place), which literally means "same place."

If all the isotopes of the various elements are plotted on a graph of N (the number of neutrons) vs. Z, a chart of the nuclides is obtained, a portion of which is shown in Fig. 3. The chart identifies the stable isotopes and lists essential data for each radioisotope, such as its decay constant, the type of radiation emitted, and the energy of the emitted particle. For nuclear engineers and physicists, the chart of the nuclides provides a direct means of visualizing nuclear reactions and therefore functions as a nuclear analogue of the chemists' periodic table. The known nuclides fall within a narrow band, which becomes more horizontal as N increases. Accordingly, only certain combinations of N and Z are possible, and the ratio of N to Z increases for the larger nuclei. The reason for this latter effect is that the number of protons must remain sufficiently small relative to the total number of nuclear particles, to prevent the electrostatic repulsion from exceeding the nuclear binding force.

As mentioned previously, the mass of the atom is approximately equal to the nuclear mass due to the relative lightness of the electrons. However, if the nuclear mass is measured, it is found that the actual nuclear mass is less than the total mass of its individual particles. The reason for this mass defect is that, in forming a nucleus, the particles achieve a lower energy state together than they would separately. The absolute value of this energy difference is called the binding energy of the nucleus and it is related to the mass defect through Einstein's famous formula: $E = mc^2$ (or in this case, $\Delta E = \Delta mc^2$).

To put the nuclear binding energy on a specific particle basis, the binding energy of every isotope may be obtained and divided by its mass number A to obtain the binding energy per nucleon (i.e., per nuclear particle), as shown in Fig. 4. From Fig. 4 it is seen that the binding energy per nucleon is maximized for nuclei around $A = 60$. Thus, for light nuclei, energy can be liberated by bringing the nuclei together. On the other hand, for heavy nuclei, energy can be released by splitting them apart. The former process, known as fusion, is operative in the sun and in the explosion of hydrogen bombs, but cannot yet be controlled and sustained at the moderate rates needed for power generation. The second process, by contrast, known as fission, was utilized in the first

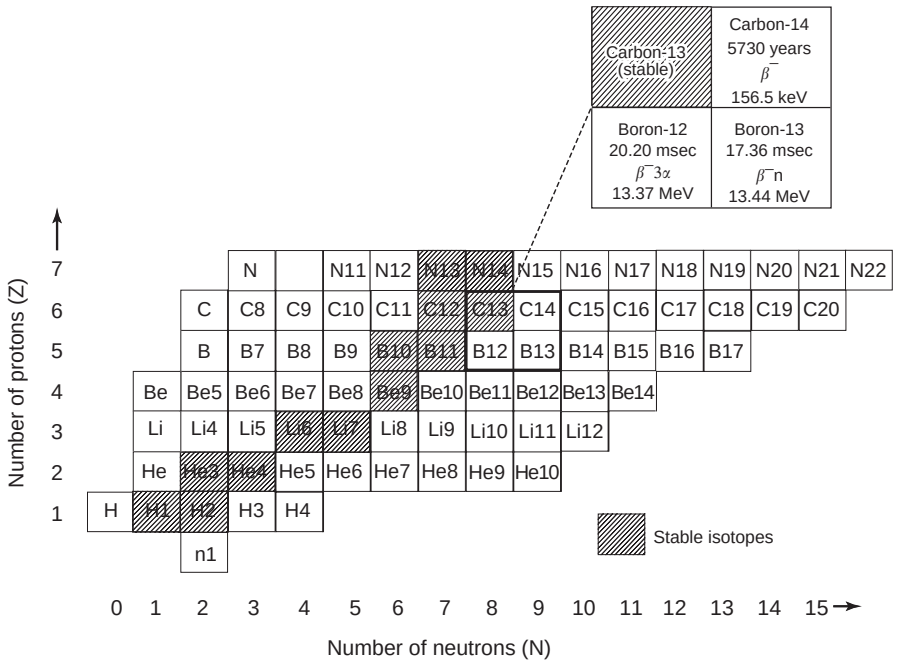


FIGURE 3 Chart of the nuclides from hydrogen to carbon.

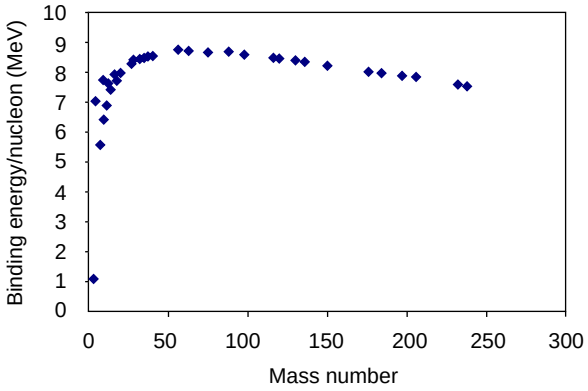


FIGURE 4 Binding energy of isotopes.

atomic bombs. Unlike fusion, fission can be controlled at sustained, moderate rates and is therefore currently utilized in nuclear power plants. For this reason, the subsequent discussion is restricted to fission technology.

With respect to fission, all nuclei are theoretically fissionable because they can be made to undergo fission by high-energy neutrons. However, some nuclei are said to be fissile, because they will undergo fission through the absorption of even slow-moving neutrons. Such slow neutrons are referred to as thermal neutrons because they have reached thermal equilibrium with their surroundings and are therefore moving as slowly as the local temperature will allow. The kinetic energy of particles is typically expressed in units of electron volts (eV), where 1 eV is the energy obtained in passing an electron or a proton through a potential difference of 1 V. Natural uranium is composed of 0.7% uranium-235, which is fissile, and 99.3% uranium-238, which is merely fissionable. Given this low percentage of the fissile isotope, natural uranium must be enriched to approximately 3% uranium-235 to serve as fuel in most reactor types. However, although the uranium-238 is not fissile, it may be transformed into fissile plutonium-239 by the absorption of a neutron. Thus, uranium-238 is said to be fertile because it can be used to breed fissile material. Given this fact, both isotopes of uranium may be utilized through special reactor designs known as breeder reactors (see later).

In addition to man-made transformations, some nuclei disintegrate spontaneously through a process known as radioactivity. As a result of radioactive decay, such nuclei emit three main types of radiation, denoted by the Greek letters α, β, and γ. In α decay, a nucleus emits an α particle, which consists of 2 protons and 2 neutrons and is identical to a helium-4 nucleus. As a result of α decay, the nuclear charge and mass decrease by 2 and 4 units, respectively. In β decay, the nucleus emits either an electron (β^-) or a positron (β^+). In β^- decay, a neutron transforms itself into a proton by emitting an electron and an antineutrino. In β^+ decay, a proton becomes a neutron by emitting a positron and a neutrino. Thus, in β decay, the nuclear charge is either increased (β^-) or reduced (β^+), but the nuclear mass remains constant. By contrast, neither the charge nor the mass of the nucleus is changed in γ decay. Rather, the nucleus gives off a high-energy photon (i.e., electromagnetic radiation) known as a γ ray. Finally, it should be mentioned that some nuclei decay by neutron emission, although this form of decay is least common. Obviously, in neutron emission, the nuclear mass is reduced without affecting the nuclear charge. As a result, the nucleus changes to a different isotope of the same element.

TABLE I

Biological Quality Factors for Various Types of Radiation

Type of radiation	Biological quality factor
X rays and γ rays	1
β particles	1–1.7
Neutrons	2–11
α particles	10
Heavy nuclei	20

With respect to radiation, a major concern is its biological impact. In this regard, the effects are found to depend on both the energy and the type of radiation absorbed. With regard to energy, the unit of absorbed dose is the "radiation absorbed dose" (rad), which is equivalent to the absorption of 0.01 joule/kilogram (J/kg) of material. However, beyond the quantity of energy absorbed, the biological effect also depends on the quality of the energy. For instance, due to their high penetrating power, γ rays deposit less energy per unit length and therefore have a smaller linear energy transfer (LET). By contrast, α particles have far less penetrating power and therefore a higher LET. Given this fact, an absorbed dose of α particles may be more dangerous than the same dose of γ rays because the energy deposition is more localized. To account for this difference, quality factors for various types of radiation have been devised, as shown in Table I. Multiplying the absorbed dose by the relevant quality factor yields the dose equivalent in "roentgen equivalent man" (rem), a measure of biological impact. Thus, based on Table I, an absorbed α dose of 1 rad would give a dose equivalent of 10 rem. The need to gain a greater

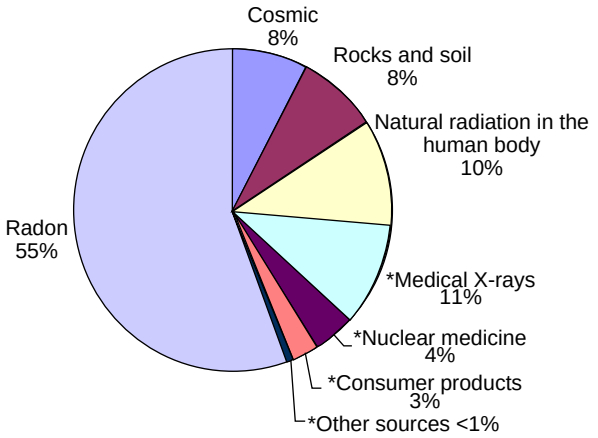


FIGURE 5 Average annual exposure to natural and man-made (asterisks) sources of radiation. The average annual dose in the United States is 360 mrem (National Council on Radiation Protection Measurements Report No. 93, 1988).

understanding of the biological effects of radiation exposure and to provide effective radiation protection methods has led to the development of a discipline known as radiation health physics.

3. RADIATION PROTECTION

Nuclear engineers work in conjunction with health physicists to assure that all activities involving radiation exposure to nuclear plant workers or to the public are kept well below the U.S. requirements stated in Title 10, Part 20 of the Code of Federal Regulations (10 CFR 20). In fact, the current industry practice is to apply the ALARA ("as low as reasonably achievable") principle to every exposure-related activity. To this end, nuclear engineers have been widely successful in designing nuclear plants that limit the dose to the public. As shown in Fig. 5, the average American receives a radiation dose of about 360 mrem (or 0.36 rem) per year. Of this exposure, the inhalation of radon gas accounts for 200 mrem per year, with another 28 mrem coming from the rocks and soil. (Radon is a decay product of the uranium naturally present in rocks and soil, and, as a noble gas, radon easily seeps into buildings.) Because both of these sources of radiation are soil based, the actual dose equivalent will vary with location. In addition to this soil-based radiation, another 27 mrem results from cosmic rays, which consist of high-energy radiation emanating from outer space. Because Earth's atmosphere acts as a filter, however, the actual exposure will vary with elevation. Beyond these external sources, the human

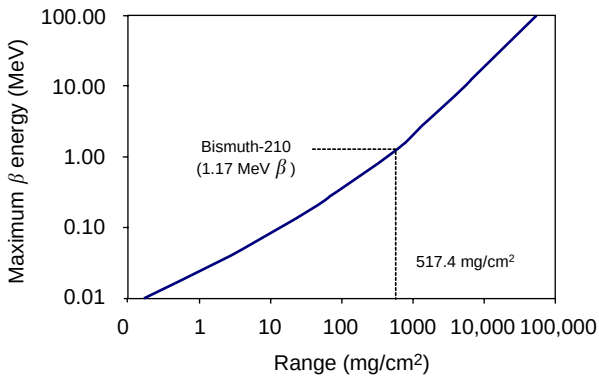


FIGURE 6 The range–energy curve for β particles.

body is also radioactive, exposing itself to about 40 mrem per year. Finally, man-made sources of radiation (such as X rays) add about 65 mrem. Nuclear power provides a mere 0.003 mrem per year, thus nuclear plants are far safer than the environment! Thus, the routine emissions of nuclear reactors are not a primary safety concern.

The three governing principles of radiation protection are time, distance, and shielding. Doses can be minimized by limiting the time of exposure, by maximizing the distance to the source, and by providing shielding to attenuate the radiation. The α particles are the easiest form of radiation to shield against. For example, an α particle with the very high energy of 4 MeV would travel only $\sim$2.3 cm in air. Hence, relatively short distances from an α source would serve as adequate protection. However, a β particle of the same energy can travel considerably farther in air, in excess of 1 m. Both α and β particles have finite ranges of travel in different materials and a significant effort has been made to document their range–energy curves. Figure 6 presents a graph of the range–energy curve for β particles. The range is defined as the shielding material density (milligrams/cubic centimeters) times the shielding material thickness (centimeters) needed to stop a β particle with a given energy. Hence, the units for the range are expressed in milligrams/square centimeter. Suppose we wish to design an aluminum shield to block the 1.17-MeV β particles emitted from a bismuth-210 source. The graph in Fig. 6 indicates that at 1.17 MeV, β particles have a range of 517.4 mg/cm^2. Dividing the range by the known density of aluminum, 2700 mg/cm^3, yields a shield thickness of approximately 0.2 cm.

Unlike α and β particles , which have finite ranges in materials, γ rays and X rays are exponentially absorbed in materials. That is, they obey the following equation, as illustrated in Fig. 7, for a

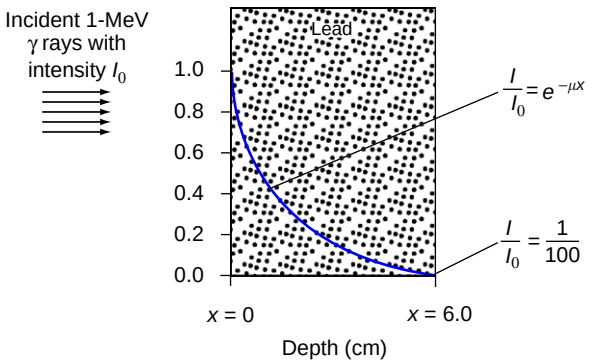


FIGURE 7 Exponential attenuation of 1.0-MeV γ rays in a lead slab.

beam of monoenergetic (single-energy) γ rays impinging on a slab:

$$I/I_0 = e^{-\mu x}.$$

This equation states that the gamma intensity, I, at a depth, x, in a shielding material is equal to the γ intensity at the front face times an exponential attenuation factor, $e^{-\mu x}$. The exponential attenuation factor is dependent on an attenuation coefficient, μ, which in turn depends on the energy of the γ ray. Values for the attenuation coefficient have been tabulated in the literature for a wide range of γ energies and materials. For example, the linear attenuation coefficient for 1.0-MeV γ particles in lead is $0.177\,\text{cm}^{-1}$. Thus, as shown in Fig. 7, 1.0-MeV γ rays would require 6 cm of lead to be attenuated by a factor of 100.

Many additional factors must be considered when designing shielding. For example, secondary radiations may be caused by the interaction of energetic γ rays with the shield material nuclei, thus requiring a greater material thickness than that predicted by the exponential attenuation equation. When shielding a radioactive isotope, consideration must also be given to its daughter products, which may have higher decay energies than the parent has. For example, strontium-90 emits a 0.54-MeV β particle, but its daughter product, yttrium-90, emits a higher energy 2.27-MeV β particle. Therefore, the shielding should be designed relative to the highest energy emission expected in the decay chain.

4. DESIGN OF NUCLEAR POWER REACTORS

Having discussed the basics of nuclear science and radiation protection, it is now possible to consider the practical application of nuclear fission as a power source. As mentioned previously, uranium-235 fissions through the absorption of thermal neutrons. Each fission, in turn, releases approximately 200 MeV of energy and two to three additional neutrons. Of these newly created neutrons, some will be lost through leakage and absorption in nonfissile material. Others, however, will be absorbed in uranium-235 nuclei to generate additional fissions and neutrons. Accordingly, it is possible to sustain the fission process via a chain reaction. If one fission induces more than one fission on average, the system is supercritical, and the fission rate will increase. On the other hand, if one fission induces less than one fission on average, the system is subcritical, and the fission rate will decrease. Finally, if each fission induces exactly one fission on average, the system is critical, and the fission rate will remain constant. In a nuclear reactor, therefore, the trick is to control the chain reaction to maintain a constant power level. In practice, this is done by the adjustment of control rods, which contain a neutron-absorbing material such as boron. By adjusting the position of the control rods, the fission rate and the corresponding heat generation rate may be increased, decreased, or held constant.

Although understanding the basic operation of a nuclear reactor is relatively straightforward, the design process is quite challenging. Nuclear engineers are tasked with calculating the time-dependent population of neutrons at each position within a nuclear reactor. The conserved quantity of interest is the neutron flux, φ. The neutron flux represents the number of neutrons per unit area, per unit time, at a given position and time inside the reactor. One method of calculation employs the neutron transport equation. This integro-differential equation is mathematically formidable and therefore not presented here. The equation, however, represents a balance of the various mechanisms by which neutrons can be lost or gained from an arbitrary volume in the system. As shown in Fig. 8, neutrons having an energy within a specified energy range and traveling in a direction within a specified range of directions can enter or leave the volume by a variety of mechanisms. The gain mechanisms include sources of neutrons, such as fission neutrons, generated within the volume, external neutrons with just the right energies and directions entering through the surface that bounds the volume, and neutrons with a different energy and direction within the volume that experience a scattering collision such that their new energies and directions fall within the specified ranges. The loss mechanisms include neutrons leaking

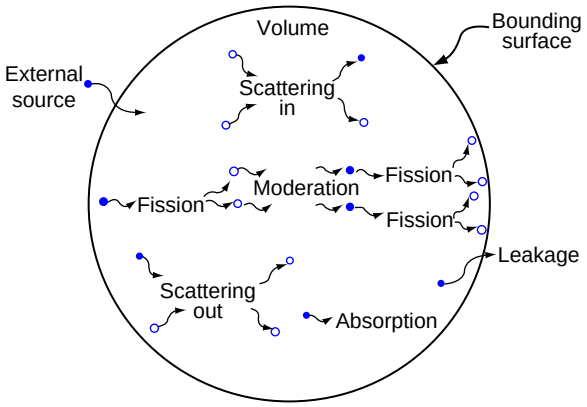


FIGURE 8 Various mechanisms by which neutrons can be lost or gained from an arbitrary volume in a reactor. Solid circles denote neutrons with energies and directions within the specified ranges. Open circles denote neutrons with energies and directions outside the specified ranges.

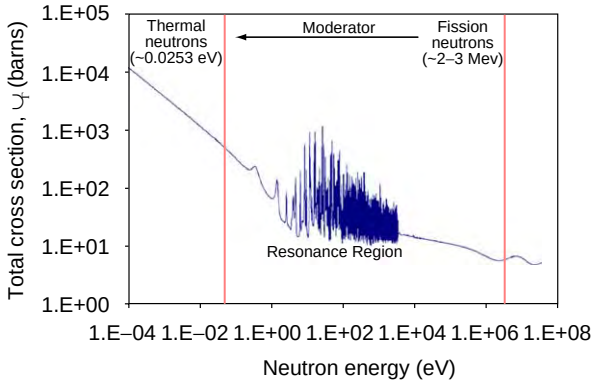


FIGURE 9 Values of total cross section for uranium-235 as a function of neutron kinetic energies.

out through the bounding surface, neutrons captured by nuclei within the volume, and neutrons within the volume that experience a scattering collision such that their new energies and directions fall outside of the specified ranges. Neutron-reflective materials such as graphite, beryllium, and heavy water (D_2O) are often used to improve fuel economy in thermal reactors. Neutrons leaking out of the reactor core interact with the reflector material nuclei and some are scattered back into the reactor. The use of reflectors allows the designer to reduce the overall dimensions of the core.

The rates at which a neutron scattering, neutron capture, or nuclei fission event occurs in a material is calculated by the reaction rate, $\Sigma\varphi$. Here φ is the neutron flux as defined previously. The reaction rate represents the frequency with which a particular reaction occurs between single-energy neutrons and the nuclei of the interacting material. It is expressed in units of reactions per unit volume, per unit time. The macroscopic cross section, Σ, consists of the nuclei density, N, of the interacting materials, in units of nuclei/cubic centimeter, times the microscopic cross section, σ, expressed in units of square centimeters. The value of the microscopic cross section varies significantly with the properties of the interacting material, the neutron kinetic energy, and the type of reaction being considered. Fission, capture, and scattering cross sections have been experimentally determined and tabulated for a wide range of materials and neutron kinetic energies. Graphs of reaction specific cross sections versus neutron kinetic energies reveal the energy ranges at which a reaction would preferentially occur in a material. For example, Fig. 9, shows the values of total cross section for uranium-235 as a function of neutron kinetic energies. The total cross section is relatively close to that of the fission cross section for this isotope. The cross sections in this graph are expressed in units called barns, where 1 barn is equivalent to 10^{-24} cm^2. The neutron kinetic energy is expressed in electron volts. Figure 9 shows that neutrons have an increased likelihood of interacting with uranium-235 nuclei at lower neutron energies. Neutrons emerge from a fission event with kinetic energies in the range of 2–3 million electron volts (2–3 MeV). By contrast, thermal neutrons at room temperature have average kinetic energies of only 0.0253 eV. Thus, to take advantage of the higher fission cross sections that arise at the lower neutron energies, thermal neutron reactors use the scattering properties of a moderator, such as graphite or water, to reduce the kinetic energy of the neutrons by eight orders of magnitude.

In summary, the neutron transport equation is a balance equation used to calculate the neutron flux as a function of position and time in a reactor. The solution to this equation is not easily obtained because its source and sink terms are given by reaction rate equations that also depend on neutron flux. In addition, the reaction rates vary with material nuclei densities and the energy-dependent neutron cross sections for fission, scattering, and capture events. A variety of numerical techniques have been developed to obtain accurate computer solutions of the three-dimensional energy-dependent neutron transport equation.

Nuclear engineers often employ an approximation to the neutron transport equation—a diffusion theory model–to predict the temporal and spatial neutron population in a reactor. Diffusion theory is much

TABLE II

Distribution of Fission Energy in a Nuclear Core[a]

Process	Percent of total reactor energy	Range
Kinetic energy of fission fragments	80.5	Very short
Kinetic energy of fast neutrons born of fission event	2.5	Medium
γ energy released at fission	2.5	Long
Kinetic energy of delayed neutrons	0.02	Medium
β decay energy of fission products	3.0	Short
Neutrinos associated with β decay Nonrecoverable	5.0	
γ energy of fission products	3.0	Long
Nonfission reactions due to excess neutrons plus β and γ decay energy of neutron absorption and γ emission (n, γ) products	3.5	Short and long

[a] Data from El-Wakil (1981).

simpler and can provide excellent results when the region of interest is not directly adjacent to a strongly absorbing medium or boundary. However, accurately predicting the fission rate of the fissionable fuels (uranium-235 and plutonium-239 produced from neutron capture in uranium-238) in the core is essential to good fuel economy and requires sophisticated modeling using benchmarked computer codes.

The amount of energy produced by fission in the core is directly related to the neutron flux profile, φ, the average energy per fission event, the number of fissionable nuclei per volume, and the effective fission microscopic cross section. Table II shows how this energy is distributed in the reactor. Most of the energy of fission, 80.5%, is associated with the kinetic energy of the fission fragments and is deposited directly in the fuel. A portion of the heat generated in a nuclear core is due to the decay of fission products. Even after the fission events in a nuclear reactor have been terminated, by the action of control rods, for example, a significant amount of decay heat continues to be generated in the core. This unique feature of nuclear heat generation imposes some very important safety constraints on the design of a nuclear power reactor. In particular, decay heat

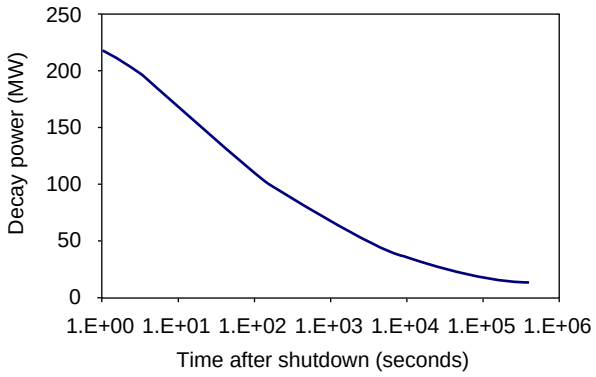


FIGURE 10 Decay power for a uranium-based thermal reactor with an initial core power of 3400 MW (thermal) and shutdown after 1 year of continual operation.

removal systems must be capable of cooling the nuclear core even after shutdown. The 1979 American National Standard for Decay Heat Power in Light Water Reactors provides the standard models useful to predict the amount of decay heat generated after reactor shutdown. The decay power is a function of the reactor power at the time of shutdown and the length of time the reactor has been at power. Figure 10 shows how core decay power varies with time after shutdown for the case of a 3400-MW thermal reactor that has been operating continuously for 1 year. One second after the reactor is shut down, the core power has dropped to $\sim$6.5% of its initial value. However, this still represents $\sim$220 MW of decay power, and significant core cooling would still be required to prevent damaging the fuel. One hour after shutdown, the core decay power is roughly 50 MW, which represents 1.5% of the original core power. Nuclear fuel is typically permitted to decay for several days before it is removed from the core as part of a refueling operation.

Having explored the fundamental concepts of nuclear engineering design as applied to nuclear reactor cores, the following section uses those concepts to describe the nuclear power plant designs that are currently employed to produce electricity.

5. COMMERCIAL NUCLEAR POWER PLANTS

The general design criteria for commercial light water reactors (LWRs) in the United States are presented in Title 10, Part 50, Appendix A of the Code of Federal Regulations (10 CFR 50 Appendix

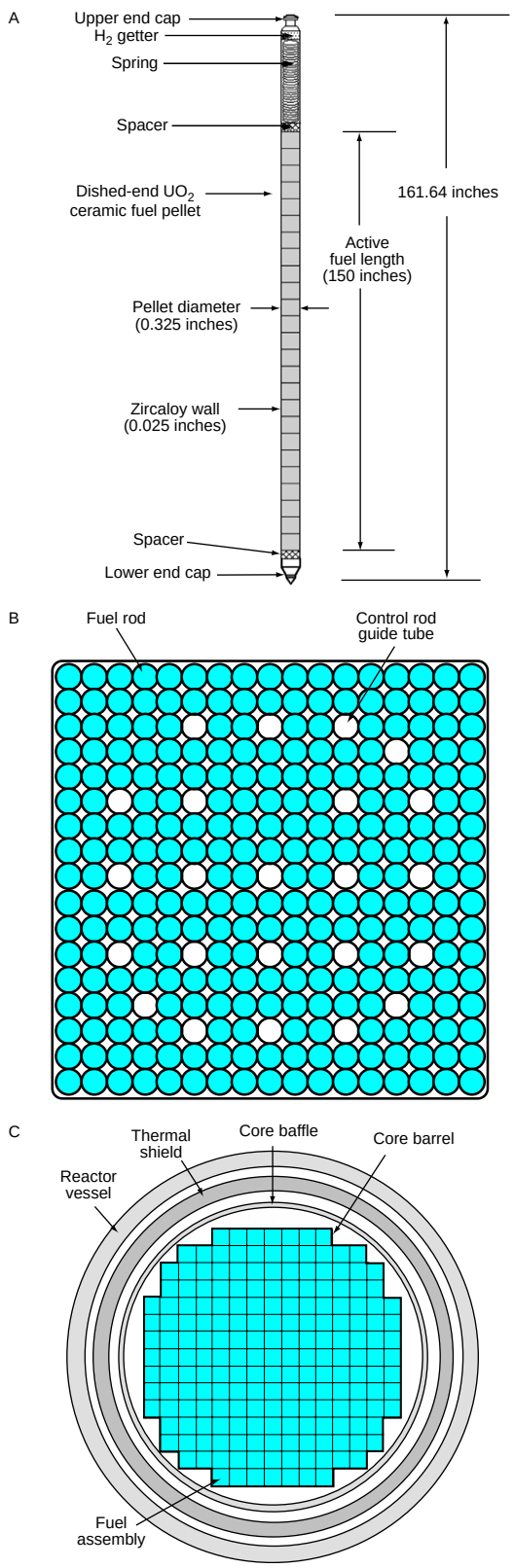


FIGURE 11 Reactor fuel geometry. (A) Fuel rod, (B) a 17×17 fuel assembly, and (C) the reactor core.

A). Inherent to these criteria is a safety philosophy known as defense-in-depth. The defense-in-depth philosophy requires that all of the safety systems be redundant (multiples of the same kind of safety system), diverse (different types of systems performing the same safety functions), and independent (a singe failure would not cause the loss of multiple safety systems). The outcome of this approach is that nuclear plants are designed with multiple barriers that prevent the release of fission products to the environment. Because the source of heat and fission products is the fuel, the discussion begins there. A typical reactor uses cylindrical fuel pellets of uranium dioxide (UO_2) enriched to about 3–5% uranium-235. Ceramic fuel is selected because of its ability to withstand very high temperatures. Each pellet is dished on the ends to allow thermal expansion and is slightly porous to trap fission gases. The pellets are stacked end to end in a long metal tube and held in place by a compressed spring. This tube, known as the cladding, provides structural support to the fuel and prevents the radioactive fission products from entering the reactor coolant. To fulfill these demands, the cladding must be a strong corrosion- and heat-resistant material as well as a poor neutron absorber. For all these reasons, a zirconium alloy called zircaloy is commonly used for cladding. As shown in Fig. 11A, taken together, the fuel pellets and the cladding form what is known as a fuel rod. The fuel rods are assembled and aligned with high precision using spacer grids to form rectangular arrays known as fuel assemblies (Fig. 11B). These assemblies, in turn, are grouped together to form a reactor core that approximates a right circular cylinder (Fig. 11C). In light water reactors, water flows through the core to remove the heat from the fuel rods. Moreover, in addition to serving as the coolant, the water also functions as the moderator, in that it acts to slow down (or moderate) the high-speed neutrons released from fission. In this regard, the moderating capacity of water stems from the fact that the average kinetic energy lost by a particle in an elastic collision is maximized when it collides with a second particle of the same mass. Accordingly, because the hydrogen nuclei have the same mass as the neutrons, light water is a very effective moderator.

With respect to LWRs, there are two major designs: the pressurized water reactor (PWR) and the boiling water reactor (BWR). As shown in Fig. 12A, the PWR utilizes three types of flow loops. In the primary loop, which passes through the reactor core, the water is highly pressurized [15.5 megapascals (MPa) = 2250 pounds per square inch

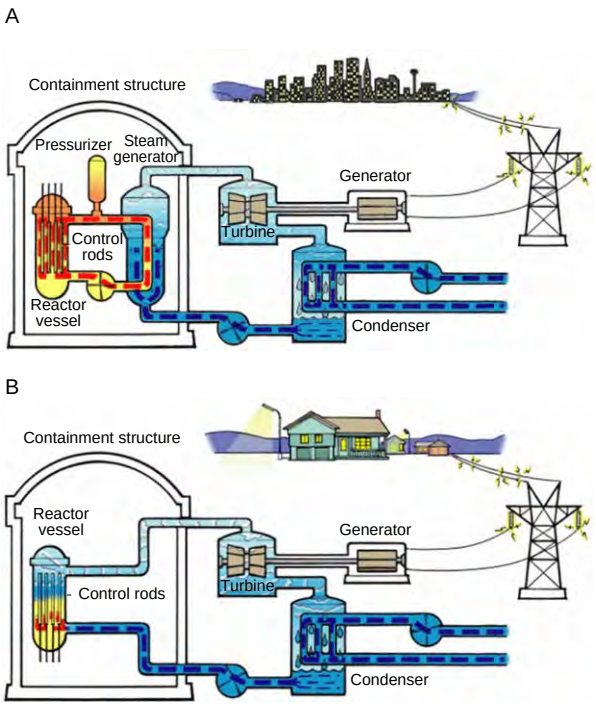


FIGURE 12 Commercial light water reactors. (A) Overall schematic for a pressurized water reactor and (B) a boiling water reactor.

gauge at atmospheric pressure (psia)] to prevent boiling. The primary coolant is heated in the core and flows through the tubes of a steam generator. The hot primary fluid gives off heat to the secondary fluid (with which it does not mix) in the steam generator, and is pumped back to the reactor to complete its circuit. The secondary fluid boils within the steam generator shell and all of the moisture is removed via separators and dryers to produce a very dry, high-pressure, steam. The steam expands to a high-velocity vapor that subsequently turns a turbine shaft connected to a generator to produce electrical power. From the turbine, the secondary fluid flows as a liquid–vapor mixture to the condenser, where it condenses to liquid by means of heat transfer to a third liquid stream. From the condenser, the secondary fluid is pumped back to the steam generator to complete its circuit. The third stream, which receives the waste heat in the condenser, generally dissipates this heat in a cooling tower or a nearby body of water such as a river or a lake. Because the PWR was the first commercially available reactor design, it is the most commonly used reactor concept worldwide.

In the BWR shown in Fig. 12B, the pressure is somewhat lower [6.9 MPa (1000 psia)], allowing the water to boil directly in the core. The moisture from the steam is removed using dryers and separators to produce a dry, high-pressure steam that goes directly to the turbine. Here the steam expands to a high-velocity vapor, which turns the turbine shaft connected to a generator to produce electrical power. From the turbine, the water flows as a liquid–vapor mixture to the condenser, where it condenses back to liquid and is then pumped back to the core. As before, the secondary stream passing through the condenser connects to a nearby body of water or a cooling tower. Although not as prevalent as the PWR, the BWR is nevertheless employed on a worldwide scale. In this regard, an improved version of the BWR, known as the advanced boiling water reactor (ABWR), is currently utilized in Japan and Taiwan.

From this description of reactors, it should be evident that many features of a nuclear power plant are identical to those of conventional power plants with the exception of the power source. Because of this exception, however, precautions must be taken to contain the radioactive fission products in the event of an accident.

6. NUCLEAR POWER PLANT SAFETY SYSTEMS

Because a nuclear explosion in a nuclear power plant is impossible due to the low fuel enrichment, the worst conceivable accident is a severe loss-of-coolant accident (LOCA), leading to a core meltdown. Although the fission process would be immediately stopped by a control rod insertion, the radioactive fission products would continue to generate decay heat in the fuel. Thus, a LOCA producing core uncovering could cause the fuel to melt. In the most extreme case, a molten mass would fall to the bottom of the reactor, melting through the reactor pressure vessel and the underlying concrete, and eventually coming to rest about 6.1 m (20 ft) underground. In the process, airborne radioactivity would likely be generated as a result of flashing steam and the gaseous products of chemical reactions. To preclude such a scenario, numerous engineered safety features have been incorporated into the design of commercial nuclear power plants. Lists of typical engineered safety features for PWRs and BWRs and brief descriptions of their functions are provided in Tables III and IV, respectively. The first four systems in each table comprise the emergency core cooling system for each design.

The zircaloy-encapsulated fuel and the primary loop can be considered the first line of defense. The reactor pressure vessel and the pipes are designed

TABLE III

Some Engineered Safety Features of Pressurized Water Reactors

Engineered safety feature	Description of function
High-pressure injection system	This system consists of multiple pumps connected to a large supply of borated water (neutron absorbing). It is capable of injecting coolant into the core at moderate flow rates while at full system pressure
Low-pressure injection system	This system consists of multiple pumps connected to a large supply of borated water capable of injecting coolant into the core at very high flow rates while at moderate system pressures
Accumulators	This system consists of a set of prepressurized tanks filled with borated water. No external power is required to inject large volumes of cold borated water into the core
Residual heat removal system	This system consists of multiple pumps capable of providing very large quantities of borated water to the core at low system pressures
Control rod shutdown margin	This system consists of an additional bank of control rods that are available to shut down the fission reaction
Auxiliary feedwater system	This system consists of multiple pumps connected to the condensate storage tank, to provide backup water for the steam generators
Containment building	This is a large-volume prestressed, posttensioned concrete building that houses the nuclear steam supply system. It is reinforced with a steel liner and steel reinforcement bars and is capable of withstanding the maximum pressures for worst-case accident conditions
Containment ventilation system	This system provides containment air cooling and uses high-efficiency filters to remove any radioactive particles inside containment
Containment sprays	This system sprays cold water into the containment to reduce containment pressure in the event of a steam release
Containment sump	This system consists of a large concrete sump at the base of the containment. It collects all of the water leaking from the reactor. It serves as the water supply for the safety injection systems during the long-term cooling phase of an accident
Redundant offsite power	This system consists of two independent sources of external AC power for all of the safety systems
Diesel generators	This system consists of multiple diesel generators (4180 V), each capable of powering all of the safety systems. In the event of a loss of offsite power, they are able to provide full power within 10 sec

with a thickness sufficient to withstand the high operating pressures. The reactor pressure vessel is made of stainless-steel-clad carbon steel with a thickness of about 25.4 cm (10 inches). However, assuming that a leak does occur, as shown in Tables III and IV, nuclear reactors have safety injection systems with independent piping, pumps, diesel generators, and water supplies. Moreover, these systems are designed to inject emergency coolant into the reactor at flow rates sufficient to meet very pessimistic conditions. For instance, the safety injection systems are designed to keep the core covered with water even if the incoming coolant pipe (the cold leg) suffers a double-ended guillotine break, with reactor coolant flowing out from both ends!

Nevertheless, assuming that such a break does occur and that the safety injection systems fail, then it would be necessary to contain the airborne radioactivity that the melting fuel would generate. For this reason, the reactor is contained in a large, dome-shaped building called the containment. The containment is a 1.2-m (4-ft)-thick, heavily reinforced and steel-lined concrete structure that is designed to withstand the internal pressures generated by flashing steam. Moreover, it is also designed to protect the reactor from outside disturbances such as earthquakes, tornadoes, hurricanes, and jet airline crashes. Consequently, when the combination of the vessel and piping thicknesses, the safety injection systems, and the rugged containment is considered, the probability of a significant release of radioactivity is nil. Although the containment design is utilized by nuclear reactors throughout most of the world, there is a notable exception. The Russian RBMK reactors (so named for the acronym in Russian, *reactor bolsoi mochnosti kipyashiy*, meaning "large power boiling reactor"), such as the one that exploded at Chernobyl in 1986, do not have a containment. Because of this difference (among others), RBMK reactors are not as safe as conventional designs and therefore do not provide an appropriate benchmark to assess worldwide reactor safety.

TABLE IV

Some Engineered Safety Features of Boiling Water Reactors

Engineered safety feature	Description of function
High-pressure injection system	This system consists of multiple pumps connected to a large supply of borated water (neutron absorbing). It is capable of injecting coolant into the core at moderate flow rates while at full system pressure
Low-pressure injection system	This system consists of multiple pumps connected to a large supply of borated water capable of injecting coolant into the core at very high flow rates while at moderate system pressures
Automatic depressurization system	This system consists of a set of valves that automatically open to reduce reactor pressure in a controlled manner
Core spray	This system consists of a set of pumps that spray water directly into the core
Control rod shutdown margin	This system consists of an additional bank of control rods that are available to shut down the fission reaction
Reactor core isolation cooling system	This system isolates the reactor from the turbines and condenser. Steam from the reactor is used to drive a turbine pump connected to the condensate storage tank to provide backup water for the core
Standby liquid control system	This system provides an alternate means of shutting down the fission reaction by injecting highly concentrated borated water into the core
Containment building	This is a large-volume prestressed, posttensioned concrete building that houses the nuclear steam supply system. It is reinforced with a steel liner and steel reinforcement bars and is capable of withstanding the maximum pressures for worst-case accident conditions
Containment sprays	This system sprays cold water into the containment to reduce containment pressure in the event of a steam release
Containment suppression pool	This system consists of a large concrete pool, or torus, at the base of the containment. It serves as a very large supply of the water for all of the safety injection systems. It also serves as a pressure suppression pool to reduce containment pressure
Redundant offsite power	This system consists of two independent sources of external AC power for all of the safety systems

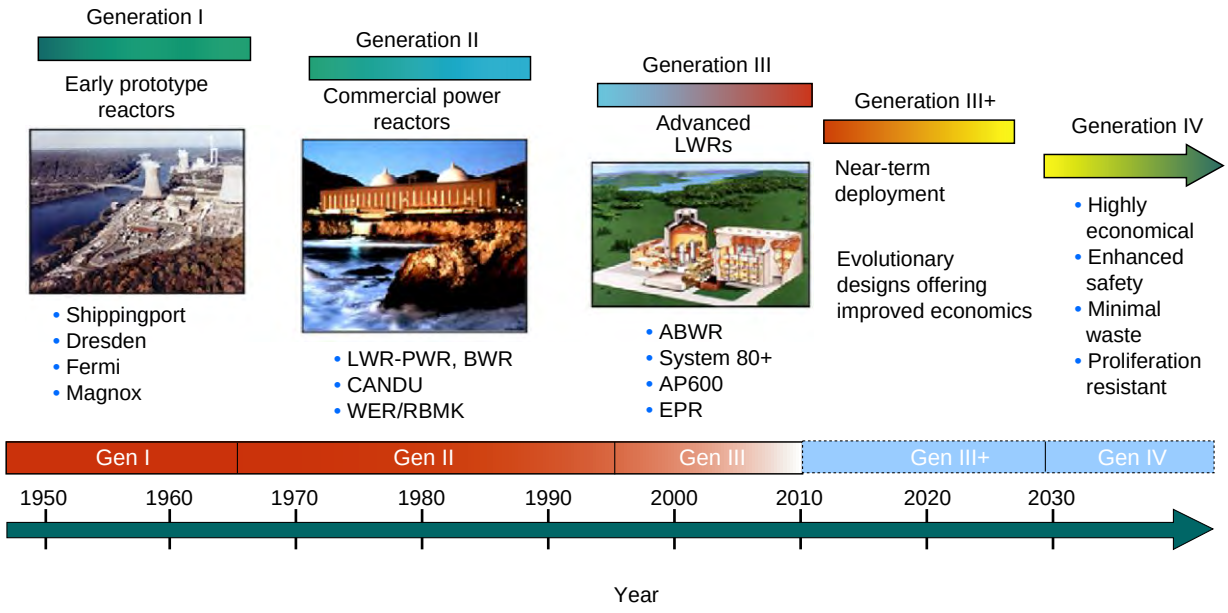


FIGURE 13 Next-generation nuclear plants, building on the experience of previous generations. LWRs, Light water reactors; PWR, pressurized water reactor; BWR, boiling water reactor; CANDU, Canadian deuterium uranium reactor; WER/RBMK, Russian light water and boiling water reactors; ABWR, advanced boiling water reactor; EPR, European pressurized water reactor. Courtesy of the U.S. Department of Energy.

7. NEXT-GENERATION NUCLEAR POWER PLANTS

Nuclear engineers are using the vast experience of two generations of reactor operation to design the nuclear power plants of the future. Figure 13 shows that a third generation of reactors will soon transition from the drawing board to reality, ready for deployment within the next 10 years. Even greater advancements are expected for the fourth-generation designs that will ready for deployment in the next 20 to 25 years. Third- and fourth-generation nuclear reactors will be safer and will exhibit a greater degree of reliability. A common design feature of these new reactors is the use of passive safety systems. Although the comprehensive systems used to safeguard today's nuclear power plants offer a very high degree of protection, even greater reliability and safety can be achieved by replacing all of the active safety systems with passive safety systems, which are driven by natural forces such as pressure and gravity. After all, whereas active pumps requiring external power can fail, gravity cannot. An example of such a system, shown in Fig. 14A, is an advanced design known as the AP600. The AP600, a 600-MWe plant designed by the Westinghouse Electric Corporation, is the first passively safe nuclear power plant certified by the U.S. Nuclear Regulatory Commission (USNRC). All of the externally powered safety systems have been replaced by elevated tanks filled with cold borated water. In the event of a LOCA, while the pressure is still high, the core makeup tanks (CMTs) begin to circulate cold water through the core by virtue of negative buoyancy. At intermediate pressures, water injection occurs via compressed nitrogen in the accumulators. Finally, at low pressures, injection from the in-containment refueling water storage tank (IRWST) occurs by virtue of gravity. Moreover, due to the containment design shown in Fig. 14B, the flashing steam is condensed on the containment walls and is recycled back to the IRWST and the lower containment compartments, to guarantee long-term cooling of the core. The heat generated by the core serves to drive all of the natural convection cooling processes. Therefore, cooling will continue as long as cooling is needed!

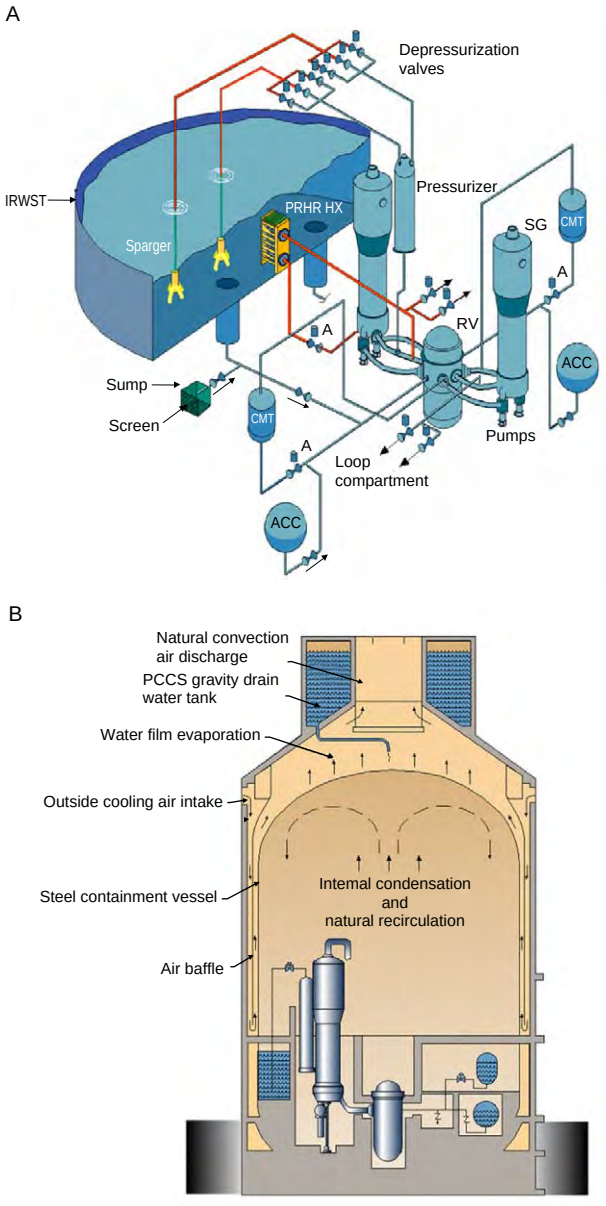


FIGURE 14 The AP600 reactor. (A) Passive safety systems. IRWST, In-containment refueling water storage tank; CMT, core makeup tank; ACC, accumulator; SG, steam generator; RV, reactor vessel; PRHR HX, passive residual heat removal heat exchanger; A, isolation valves. (B) Containment design. PCCS, Passive containment cooling system. Reproduced with permission from Westinghouse Electric Company.

Because of the novelty of the design, licensing tests to demonstrate the safety of the AP600 were conducted at various experimental facilities in Italy, Japan, and the United States. With regard to the latter, 68 tests were conducted at the Advanced Plant Experiment (APEX) test facility at Oregon State University (OSU). As shown in Fig. 15, APEX is a one-quarter-scale integral system test facility designed to simulate the passive safety systems of the AP600. The nuclear core is simulated by electric heater rods, and all the components of the passive safety system are modeled. After conducting 28 accident simulations for Westinghouse and the U.S. Department of Energy and an additional 40 confirmatory tests for

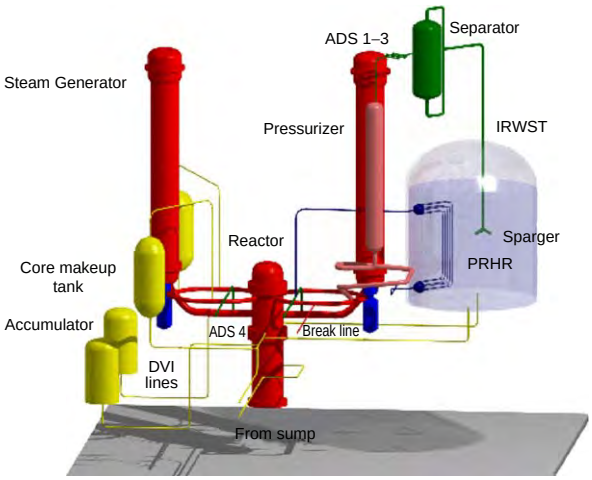


FIGURE 15 The one-quarter-scale advanced plant experiment (APEX) testing facility. IRWST, In-containment refueling water storage tank; PRHR, passive residual heat removal; ADS, automatic depressurization system; DVI, direct vessel injection. Drawing courtesy of Scott C. Franz.

the USNRC, it was concluded that even when some safety systems fail, the AP600 has sufficient redundancy to keep the core covered with water. The USNRC certified the AP600 on 15 December 1999. A larger version of the AP600, the AP1000, is currently in the licensing stage. The AP1000 is a 1000-MWe PWR with the same passive safety systems used in the AP600. To certify this design, the APEX facility at OSU has been modified by the U.S. Department of Energy (USDOE). The new facility will also be utilized to perform the confirmatory tests for the USNRC.

So far, the discussion has concerned light water reactors. In Canada, however, a reactor system has been developed that uses heavy water (i.e., water molecules with hydrogen-2, or deuterium, atoms, rather than hydrogen-1 atoms). This system is known as the Canadian deuterium uranium (CANDU) reactor. Because heavy water, compared to light water, is a much poorer neutron absorber, parasitic absorption is reduced. As a result, it is possible to use natural uranium fuel, thereby avoiding the costs of enrichment. A unique feature of the CANDU is its online refueling capability. The Canadian ACR-700 reactor is a potential candidate for near-term deployment.

To obtain thermodynamic efficiencies approaching 40%, some reactors utilize a gas cycle rather than steam. In these reactors, the gas circulates in a loop that includes both the reactor and the turbine. Two variations on this theme are the American high-temperature gas reactor (HTGR) and the British advanced gas reactor (AGR), which use helium and carbon dioxide coolants, respectively. Both systems utilize a solid graphite moderator due to the poor moderating ability of low-density gasses. The fuel rods are placed within holes in graphite blocks, with additional holes made for coolant flow. The pebble bed modular reactor (PBMR) has received significant attention because of its inherently safe fuel. The fuel consists of 0.5-mm uranium oxide particles that are triple coated with ceramic materials that will not fail even at extremely high temperatures. Approximately 15,000 particles are embedded in a single 50-mm sphere that is coated with high-purity carbon. The spheres are arranged in a critical geometry to create a nuclear core. The PBMR is a direct-cycle system that uses helium to cool the core while driving a high-efficiency Brayton gas cycle to produce electricity.

Another variation is seen in the liquid metal fast breeder reactor (LMFBR), which uses liquid sodium as the coolant. Because liquid sodium is a poor moderator of neutrons, the high-energy neutrons produced by fission stay fast to enhance breeding. For a reactor to breed, it is necessary that more fuel be produced from fertile materials than is consumed in running the reactor. Thus, for every fission, one or more of the neutrons produced must go on to breed new fuel, because exactly one of these neutrons is needed to sustain the chain reaction. Consequently, because two or more neutrons must be utilized per fission, it is necessary to maximize neutron production while minimizing leakage and parasitic absorption. Because neutron production increases, and parasitic absorption decreases, with increasing neutron energy, having a fast neutron spectrum enhances breeding. Moreover, by placing a blanket of fertile uranium-238 around the core, the leaking neutrons are utilized to breed plutonium-239 in the blanket. In this way, the LMFBR produces fissile plutonium from the nonfissile uranium-238, which constitutes 99.3% of the naturally occurring uranium. Consequently, because breeder reactors do not waste this abundant isotope, they will be able to supply world energy needs for millennia.

After an extensive review, the USDOE has narrowed their research and development efforts to focus on six generation IV reactor concepts. Each design was specifically selected because it offers a high degree of safety, reliability, proliferation resistance, economy, and efficiency. The six designs selected are the gas-cooled fast reactor (GFR), the lead-cooled fast reactor (LFR), the sodium-cooled fast reactor (SFR), the very-high-temperature reactor (VHTR), the supercritical water-cooled reactor

(SWCR), and the molten salt reactor (MSR). The VHTR design is of particular interest because its 1000°C (1832°F) outlet temperature can be used to crack water molecules to produce hydrogen for a wide range of energy-related applications, such as automobile fuel cells. All of these designs represent a significant leap forward in the development of nuclear technology and high-temperature, high-strength, materials.

8. NUCLEAR WASTE DISPOSAL

The timely deployment of the next generation of nuclear power plants will undoubtedly depend on having in place a fully functional nuclear waste repository. As shown in Fig. 16, the nuclear fuel cycle represents the life cycle of the nuclear fuel from its initial mining to its final burial. In this figure, the various waste streams are also indicated. The uranium is initially mined and separated from its ore in the milling process to produce yellowcake (U_3O_8). The yellowcake is then sent to a chemical conversion plant to produce uranium hexafluoride (UF_6) gas. Through a process of gaseous diffusion, the UF_6 is next enriched in its uranium-235 content at the enrichment plant. Essentially, because lighter

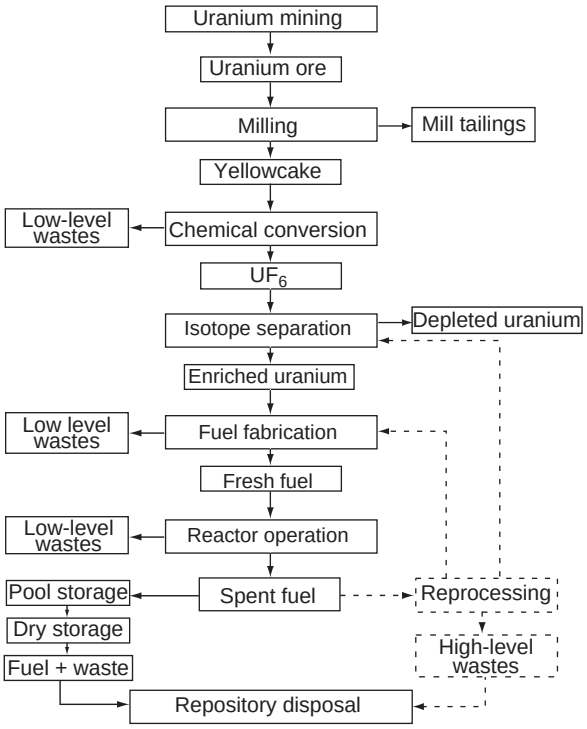


FIGURE 16 Nuclear fuel cycle.

particles diffuse more rapidly through permeable membranes, the gaseous streams crossing such membranes are progressively enriched in their uranium-235 content. Once the resulting stream is sufficiently enriched (about 3%), the gas is taken to the fuel fabrication plant to produce the uranium dioxide fuel pellets and thence fuel rods and assemblies. This fuel, in turn, is then taken to a nuclear reactor and burned for about 3 years to produce electrical power.

After the spent fuel is taken from the reactor, one of two options will be used. In the current U.S. approach, the fuel rods are stored in a water storage pool at the reactor site for several years, to allow the most intensely radioactive fission products to decay. Once these rods have sufficiently cooled, they are then removed to dry storage to await burial 300 m (1000 ft) underground in a dry geologic medium. In the second approach, the fuel rods will be taken from storage to a reprocessing plant where the uranium and plutonium will be separated out and reprocessed back into new fuel. Moreover, certain other isotopes having medical or industrial value will also be separated out. Finally, certain long-lived fission products known as actinides will also be separated out and recycled back into the fuel to be fissioned in a reactor, and thereby transformed into stable or rapidly decaying isotopes. This process of eliminating long-lived fission products through recycling and fission is known as transmutation. Because the uranium and plutonium constitute 97.1% of the spent fuel, and because the industrial isotopes and actinides constitute another 2.6%, reprocessing will utilize valuable resources within the waste while producing a 300-fold reduction in the waste volume. As a result, the annual per capita high-level waste will have the volume of an aspirin tablet. Such waste will be incorporated into a glass matrix, placed in a corrosion-resistant metal container, buried deep 300 m (1000 ft) underground in a dry geologic medium, and packed into place with bentonite clay. Moreover, because the long-lived fission products will have been separated out, the buried waste will decay more quickly. As a result, its toxicity will approach that of natural uranium within 10,000 years.

To preserve the waste intact for this amount of time, a defense-in-depth strategy will be employed. Essentially, the wastes will be mixed with molten glass, poured into a corrosion-resistant steel or titanium canister, and allowed to harden. The canisters, in turn, will be buried deep 300 m (1000 ft) in a geological medium that is both dry

and impermeable to water. In the United States, the currently envisioned site is Yucca Mountain, Nevada, where the water table is 550 m (1800 ft) below the surface and the annual rainfall is a mere 15 cm (6 inches). Thus, at a depth of 300 m (1000 ft), the waste will be safe from both rainfall and upwelling groundwater. The proposed waste disposal facility is known as a repository and will consist of a central vertical shaft that connects with horizontal tunnels. In these tunnels, the waste canisters will be placed in holes drilled into the tunnel floors and packed in with bentonite clay. Finally, when the repository becomes full, the tunnels and the central shaft will be filled in with bentonite clay to prevent the intrusion of water.

As should be apparent, a repository employs many barriers to isolate the waste. First, water must travel some distance through an impermeable geological medium. Second, the groundwater must penetrate a layer of bentonite clay, which will swell to prevent its further passage. Third, the groundwater must next penetrate a corrosion-resistant titanium or steel canister. Fourth, the water must then leach the wastes out of a solid glass matrix. Fifth, the groundwater must then carry the wastes long distances through an impermeable geological medium to reach major population centers. Sixth, the progress of some of the waste species will be delayed relative to the groundwater due to chemical reactions with the geological medium. Thus, by using multiple barriers, it will be possible to contain the waste for a long time and to delay its spread thereafter. As a result, the wastes can be kept from the public until their radioactivity has largely decayed away.

Having examined the nuclear fuel cycle, it remains to consider the transportation of radioactive materials. In this regard, the most critical issue is the possibility of radioactive releases resulting from accidents involving spent fuel shipments (for reprocessing or burial). To prevent such an occurrence, U.S. shipping casks are designed to withstand a series of four extreme tests without leaking. In order of occurrence, such tests include a 9.1-m (30-ft) fall onto an unyielding surface (i.e., from an overpass), a 1.02-m (40-inch) fall onto a 15.2-cm (6-inch) diameter pin (i.e., onto a bridge abutment), a 30-minute fire at 801.7°C (1475°F) (i.e., a crash with flammable material), and an 8-hour submersion in water (i.e., fall into a roadside creek). Moreover, such casks have been deliberately slammed into concrete barriers at 37.6 m/sec (84 miles/hour), hit broadside by locomotives going 35.8 m/sec (80 miles/hour), and then subjected to fire tests without leaking.

As a close to this discussion on waste disposal, it is fitting to consider briefly the implications of the current U.S. waste disposal policy that would result in the burial of usable fuel, including large amounts of plutonium and potentially useful medical isotopes, as opposed to spent fuel reprocessing. It is important to note that the technical feasibility of reprocessing nuclear fuel is not at issue. France has been doing so for many years. Reprocessed fuel would consist primarily of a mixture of uranium and plutonium in an oxide form, the remaining isotopes having been removed. This fuel is known as a mixed oxide fuel (MOX) fuel. Reprocessing spent fuel solely to obtain medical isotopes would not likely be cost-effective because the isotopes of interest represent only a small fraction of the total inventory. Indeed, because of their very special purity requirements, medical isotopes might still be produced more economically in a nuclear reactor dedicated solely to isotope production, even if the isotopes arise as a by-product of producing mixed oxide nuclear fuel. However, this argument should be revisited as technology advances and as the use of different medical isotopes broadens. The chief argument against reprocessing has been the concern that the plutonium-239 separated from spent nuclear fuel might be diverted to produce nuclear weapons. Although burying the spent fuel in its entirety reduces the risk of proliferation, it does not eliminate the long-term concern. It merely passes the concern to the future generations tasked with maintaining and protecting such a repository. Plutonium-239 has two viable uses: it can be used to produce a nuclear weapon or it can be fissioned in a nuclear reactor to produce electricity. Once fissioned, that quantity of plutonium is gone forever. It is interesting to note that a spent fuel reprocessing facility could also be used to convert weapons-grade plutonium to low-enriched mixed oxide nuclear fuel for electricity production, such as has been proposed for the dismantled nuclear weapons produced in the former Soviet Union. It is an encouraging commentary on society whenever swords are turned to plowshares.

9. CONCLUSION

The field of nuclear engineering encompasses nuclear science, radiation protection, and reactor core design; to understand the current state of commercial nuclear power and the role played by nuclear engineers in its development requires an understanding of these subjects. The current generation

of commercial nuclear power reactors has served humankind exceedingly well, providing 16% of the world's total electric power production. Nuclear engineers are now designing and testing innovative designs that will be deployed worldwide over the next 20 years. In particular, the fourth-generation reactors will be safer, more efficient and reliable, cost competitive, and proliferation resistant. The use of passive safety systems will reduce the complexity of the plants and will provide a new level of protection while promoting public confidence. Plans to store nuclear wastes in a secured national repository have been approved and have effectively closed the nuclear fuel cycle in the United States. Other countries have adopted a combined strategy of spent fuel reprocessing and a waste repository. The concept of using highly enriched plutonium from dismantled nuclear weapons to produce limited quantities of mixed oxide fuel for electricity production has significant merit and may lead to new discussions on fuel reprocessing.

Although the field of nuclear engineering is young relative to other engineering disciplines, few other disciplines have such a prestigious heritage. Born of Einstein's remarkable theory, fueled by Eddington's visionary challenge, and ushered into reality by Fermi's experimental genius, nuclear engineering remains faithful to its original guiding principle—to find creative ways of safely harnessing and using the energy of the atom for the benefit of mankind.

SEE ALSO THE FOLLOWING ARTICLES

Nuclear Fuel: Design and Fabrication • Nuclear Fuel Reprocessing • Nuclear Fusion Reactors • Nuclear Power Economics • Nuclear Power, History of • Nuclear Power Plants, Decommissioning of • Nuclear Power: Risk Analysis • Nuclear Proliferation and Diversion • Nuclear Waste • Occupational Health Risks in Nuclear Power • Public Reaction to Nuclear Power Siting and Disposal

Further Reading

Beckmann, P. (1985). "The Health Hazards of Not Going Nuclear." The Golem Press, Boulder, Colorado.

Binney, S. E. (1987). Nuclear energy. *In* "The Encyclopedia Americana," Vol. 23, pp. 511h–516.

El-Wakil, M. M. (1981). "Nuclear Heat Transport." American Nuclear Society, La Grange Park, Illinois.

Murray, R. L. (1994). "Understanding Radioactive Waste." Battelle Press, Columbus, Ohio.

Nuclear Fission Reactors: Boiling Water and Pressurized Water Reactors

DOUGLAS K. VOGT
Lawrence Livermore National Laboratory
Livermore, California, United States

1. **Electricity Production from Light Water Reactors**
2. **Pressurized Water Reactors**
3. **Boiling Water Reactors**
4. **Ship Propulsion**

Glossary

boiling water reactor (BWR) A light-water-cooled and -moderated reactor that operates at a pressure of ~1000 pounds per square inch at atmospheric pressure (psia). At this pressure, reactor coolant is converted from liquid water to steam by thermal energy from nuclear fission. The boiling occurs within the reactor core. Steam from the core of a BWR flows to a conventional turbine-generator system, where it is converted to electricity.

decommissioning At the end-of-life stage of a nuclear power plant, the facility contains highly radioactive components. As part of decommissioning, these components, including the reactor and other plant areas, are decontaminated to reduce the amount of radioactivity. Some components will be removed from the reactor site and buried in an engineered disposal site. Pressurized water, boiling water, and naval reactors are decommissioned.

high-level radioactive waste repository An engineered facility that isolates spent fuel or the radioactive waste from spent fuel reprocessing, safeguarding the environment.

light water Hydrogen has two stable isotopes, (1) protium (99.985% natural abundance), which has a nucleus consisting of one proton, and (2) deuterium (0.015% natural abundance), which has a nucleus consisting of one proton and one neutron. Water molecules formed from water with natural abundance hydrogen (99.985% protium) are called light water. Water molecules formed from hydrogen enriched in deuterium are called heavy water.

light water reactor (LWR) Any reactor that is cooled and moderated by ordinary or light water. Both pressurized water and boiling water reactors are light water reactors.

pressurized water reactor (PWR) A reactor that operates at high pressure (typically >2000 psia). At this pressure, PWR coolant water does not experience significant boiling. Reactor coolant is heated by thermal energy from nuclear fission, resulting in a temperature increase. Heated coolant flows to a heat exchanger with lower pressure water (~1000 psia) on the "nonreactor side." The lower pressure water is converted to steam. The steam from the steam generator flows to a conventional turbine-generator system, where it is converted to electricity.

Scientists in the former Soviet Union and in the West who developed nuclear weapons in World War II and in the early days of the Cold War realized that the tremendous energy produced by nuclear fission could be tapped either for direct use or for generating electricity. It was also clear that nuclear fission energy would allow development of compact, long-lasting power sources that could have various applications, including powering ships, especially submarines. From their roots in naval nuclear power programs, light water reactors have become an important source of electric power in many countries. Light water reactors built over the past 30 years continue to generate electricity, and new reactor designs, the advanced boiling water reactor and the advanced pressurized water reactor, are planned for operation in the 21st century.

 © 2004 Elsevier Inc. All rights reserved.

1. ELECTRICITY PRODUCTION FROM LIGHT WATER REACTORS

The first nuclear reactor to produce electricity (albeit a trivial amount) was the small experimental breeder reactor (EBR-1) in Idaho, in the United States. EBR-1 started operating in December 1951. In 1953, President Eisenhower proposed his "Atoms for Peace" program, setting the course for civilian nuclear energy development in the United States. However, it was the U.S. effort under Admiral Hyman Rickover that led to the development of the pressurized water reactor (PWR) for submarine use. As part of the PWR development, Westinghouse Electric Corporation built a land-based submarine rector prototype at the National Reactor Testing Station near Arco, Idaho. The prototype started operation in March 1953, and in 1954, the first nuclear-powered submarine, the *USS Nautilus*, was launched. In 1959, both the United States and the former Soviet Union launched their first nuclear-powered surface vessels.

The success of the submarine prototype reactor led to the U.S. Atomic Energy Commission decision to build the 90-megawatt electric (MWe) output Shippingport demonstration PWR near Pittsburgh, Pennsylvania. Shippingport started operation in 1957 and operated as a commercial nuclear power plant until 1982. In the United States, Westinghouse Electric Corporation designed the first fully commercial 250-MWe PWR, Yankee Rowe, which started operation in 1960 and operated until 1992. As the PWR was being developed, Argonne National Laboratory and General Electric Company were developing the boiling water reactor (BWR). A small prototype BWR, Vallecitos, generated electricity in Pleasanton, California from 1957 to 1963. Dresden-1, which was designed by General Electric, was the first large commercial BWR (200 MWe). Dresden-1 was constructed 50 miles southwest of Chicago, Illinois and operated from 1960 until 1978.

In 1964, the first Soviet land-based PWR power reactor began power operations in Novovoronezh (Volga region). This reactor was a new 210-MWe pressurized water reactor known as a VVER (an acronym in Russian for *veda-vodyanoi energeticheskiy reaktor*, meaning "water-cooled power reactor"). In the Murmansk region of the Arctic Northwest, a slightly bigger VVER with a capacity of 440 MWe was constructed at the Kola Nuclear Power Plant. This became a standard Russian PWR design.

By the end of the 1960s, electric utilities worldwide were placing orders for PWR and BWR units of 400 MWe to more than 1000 MWe. Other countries have built primarily PWR and BWR types, so that today, 65% of the world nuclear electric generating capacity is PWR and 23% is BWR. Overall, 16% of the world's electricity now comes from nuclear energy. Throughout the world, there are 438 commercial nuclear generating units with a total capacity of about 351,000 MWe. Worldwide, LWRs generated over 2,000,000,000 megawatt-hours (MWh) of electricity in 2001. Figure 1 shows the location of nuclear power plants worldwide that are under construction, in operation, or undergoing decommissioning. Figure 2 summarizes the

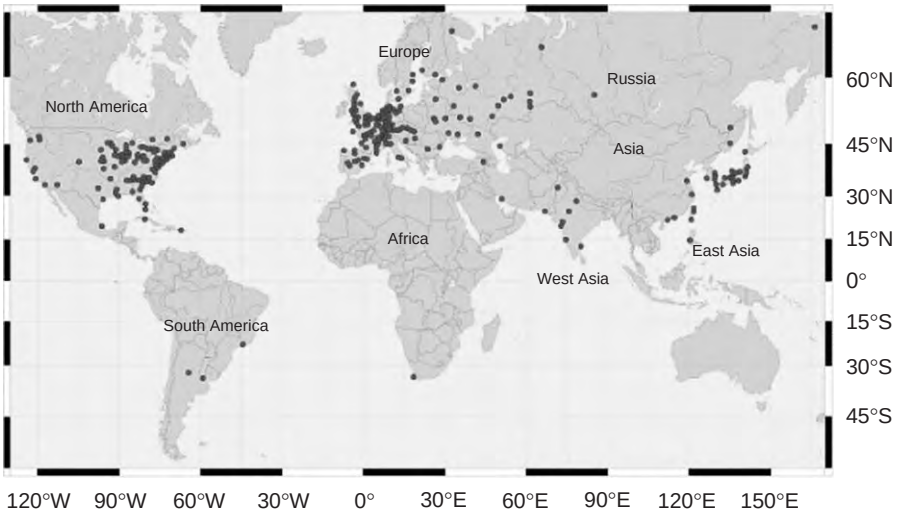


FIGURE 1 Location of nuclear power plants worldwide. Map from the International Nuclear Safety Center (Argonne National Laboratory).

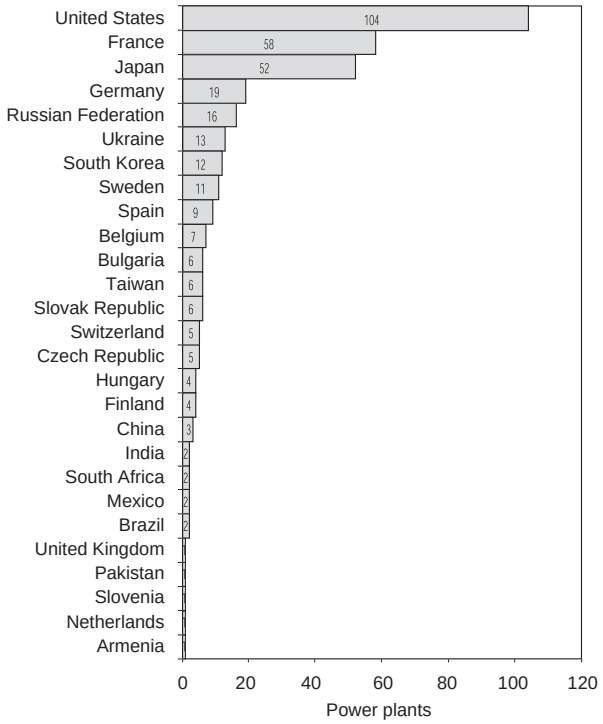


FIGURE 2 Number of light water reactor nuclear power plants worldwide.

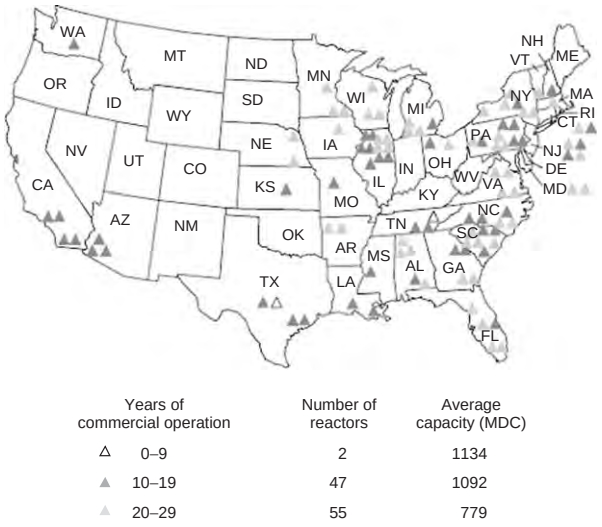


FIGURE 3 Locations of operating light water reactors in the United States as of December 2000 (there are no commercial reactors in Alaska or Hawaii). MDC, Maximum dependable capacity. This map and additional data are available on the Internet at the Web site of the U.S. Nuclear Regulatory Commission (http://www.nrc.gov).

number of operating LWR nuclear power plants in each country.

To minimize dependence on fossil fuel imports, several countries have placed significant dependence on nuclear energy from LWRs. The following countries generate more than 30% of electricity output using LWRs: France, Belgium, Slovak Republic, Ukraine, Sweden, Bulgaria, South Korea, Hungary, Slovenia, Switzerland, Armenia, Japan, Finland, and Germany. Eight countries have manufactured large LWRs for electricity generation: the United States, France, Japan, Germany, Russia, South Korea, Sweden, and the United Kingdom. Figure 3 shows the 32 states with operating LWRs in the United States.

2. *PRESSURIZED WATER REACTORS*

2.1 Plant Design Characteristics

Figure 4 provides a schematic view of a pressurized water reactor power plant. The power plant systems include a nuclear reactor and steam generator (called the nuclear steam supply system), which produce steam using energy from nuclear fission; a reactor "containment system," which limits release of radioactive material from the reactor under accident conditions; a steam-powered turbine-generator that produces electricity; and a condenser and cooling towers for heat rejection.

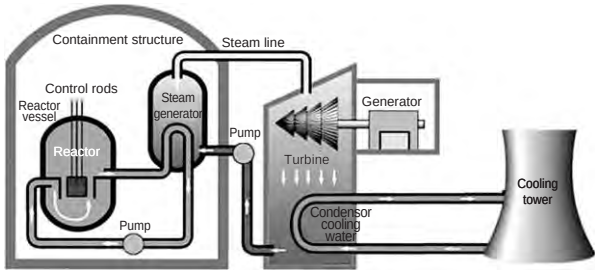


FIGURE 4 Schematic view of a pressurized water reactor.

The nuclear steam supply system includes nuclear fuel that is configured as the "reactor core" in a reactor pressure vessel. Reactor coolant water is pumped through the reactor core and heated. Heated water passes through tube-and-shell heat exchangers called "steam generators." On the reactor side of the steam generator, the water enters the inside, or tube side, of the steam generator, passes through, and eventually returns to the reactor through a closed piping system, to be reheated. On the turbine-generator side, the water passes over the outside, or shell side, of the steam generator tubes and is heated. Heat transfer through the tubes provides the energy to convert liquid water to steam on the shell side of

the steam generator. Moisture (entrained liquid water) is removed from the steam and the steam flows to a conventional steam turbine. The steam turbine is coupled to a conventional electric generator. After steam passes through the steam turbine, it passes through another heat exchanger called a "condenser," where the steam is condensed to liquid water. The condensed water from the steam turbine is then pumped through heat exchangers called "reheaters" and is returned to the steam generators in a closed piping loop. The rejected heat from the condenser system is discharged to the environment by evaporative cooling from cooling towers. The electric generator output is supplied to a local transformer yard that is connected to the local electrical grid. The reactor containment building, turbine-generator, and cooling towers are the prominent buildings visible at most nuclear power plant sites.

2.2 Fuel Design Characteristics

There are numerous PWR nuclear fuel assembly designs. However, most PWR power reactor fuel designs will be conceptually similar to the fuel assembly schematic shown in Fig. 5. The PWR uranium dioxide (UO_2) fuel pellets are nominally 0.375 inch in diameter and 0.5-inch-long solid cylinders. From 200 to 300 fuel pellets are loaded into hollow zirconium fuel tubes called fuel rods. A column of fuel pellets, typically 12 feet (144 inches) long, is loaded into a fuel rod. The PWR fuel tube is typically 150 inches in length. The nominal 6-inch length of the fuel rod that does not contain fuel pellets is called the "plenum." The plenum provides a space to accommodate fuel pellet axial thermal expansion during operation and the fission product gases that are released during power operations. A spring is inserted into the top of the fuel rod to hold the fuel pellets down in the fuel tube during transportation to the reactor and end caps are welded onto the upper and lower ends of the fuel rod. Fuel rods are inserted into spacer grids that are fabricated from a zirconium or inconel alloy. Fuel rods are held in place in the spacer grids by small leaf springs. Control rod guide tubes replace fuel rods typically at 20 to 25 fuel rod locations. The spacer grids are welded to control rod guide tubes. Some fuel assemblies may have a central fuel rod location used for in-core instrumentation. Depending on fuel assembly design, from 196 to 289 fuel rod locations are available in a PWR fuel assembly. Fuel assembly designs normally arrange fuel rods in a square array with 14, 15, 16, or 17 fuel rod locations on a side. At the top and bottom of the fuel assembly are end fittings. The end fittings are typically machined from stainless steel and provide structural strength to the fuel assembly and help control inlet and outlet flow from the reactor core during operation. In the reactor pressure vessel, 120 to 193 fuel assemblies are arranged into an approximately cylindrical geometry.

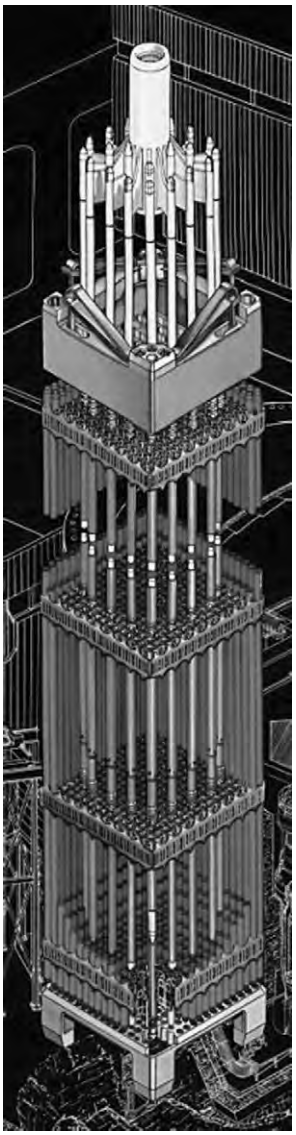


FIGURE 5 Schematic view of a pressurized water reactor fuel assembly.

Fission of uranium dioxide fuel in the reactor core is the heat source in a nuclear reactor. Fuel typically operates in a reactor for 3–5 years. During this time, the equivalent of 3–5% of the uranium atoms initially loaded into the reactor fission, leaving fission products in the fuel. Plutonium and other actinides are also produced by neutron capture. At the end of its

power-generating life, fuel is discharged from the reactor and replaced with fresh fuel. The discharged fuel is called "spent nuclear fuel" or "spent fuel." Though some fission products in spent fuel are stable (nonradioactive) or decay to nonradioactive isotopes, spent fuel remains radioactive for very long periods of time following discharge from the reactor. Spent fuel generates decay heat and penetrating radiation (high-energy γ rays and β particles). Spent fuel is stored at reactor sites for several years after discharge, typically 5 or more years. There are two methods of disposition of spent fuel. It is either shipped (in shipping packages designed for radioactive material) to a nuclear fuel reprocessing facility for recovery of residual uranium and plutonium or shipped to a high-level radioactive waste repository for disposal. Table I provides a summary of key design characteristics of a typical pressurized water reactor.

TABLE I

Pressurized Water Reactor and Fuel Design Parameters

Design parameter	Typical value
Thermal power	3411 MWt
Electrical power	1100 MWe
Reactor pressure	2250 psia
Reactor coolant inlet temperature	560°F
Reactor coolant outlet temperature	620°F
Reactor coolant flow rate	140×10^6 lb m/hr
Number of fuel assemblies in core	193
Number of steam generators	4
Steam turbine inlet temperature	541°F
Steam turbine inlet pressure	970 psia
Fuel rod array	17×17
Fuel rods per assembly	264
Control rod locations per assembly	25
Fuel assembly cross section	8.426×8.426 inches
Fuel assembly length	160 inches
Fuel rod outer diameter	0.36 inch
Fuel rod clad thickness	0.0225 inch
Fuel pellet outer diameter	0.3088 inch
Fuel pellet length	0.53 inch
Fuel assembly weight	535 kg
Fuel assembly uranium weight	423 kg

3. BOILING WATER REACTORS

3.1 Reactor Design

Figure 6 provides a schematic view of a boiling water reactor power plant. The power plant systems include a nuclear steam supply system, which produces steam using energy from nuclear fission; a reactor "containment system," which limits release of radioactive material from the reactor under accident conditions; a steam-powered turbine-generator that produces electricity; and a condenser and cooling towers for heat rejection.

The nuclear steam supply system includes nuclear fuel (discussed later) that is configured as the "reactor core" in a reactor pressure vessel. Reactor coolant water is pumped through the reactor core and heated. At the BWR system operating pressure, liquid water is converted to steam as the water is heated. Moisture is removed from the steam in a moisture separator located in the reactor pressure vessel above the reactor core. The steam flows out of the reactor building to a conventional steam turbine. The turbine-generator, condenser, and cooling towers for a BWR are similar to those for a PWR.

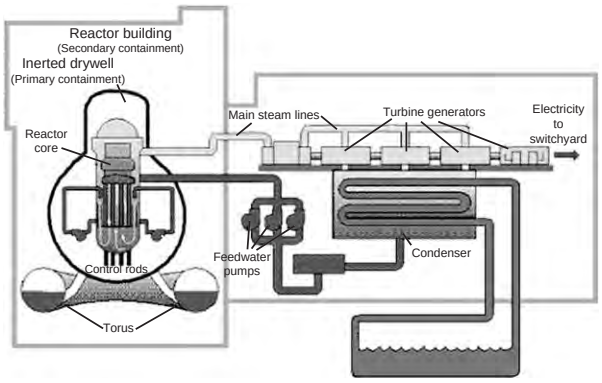


FIGURE 6 Schematic view of a boiling water reactor system.

3.2 Fuel Design Characteristics

There are numerous BWR nuclear fuel assembly designs, but most BWR power reactor fuel designs will be conceptually similar to the fuel assembly schematic shown in Fig. 7. The BWR uranium dioxide (UO_2) fuel pellets are nominally 0.4 inch in diameter and 0.4-inch-long solid cylinders. From 200 to 300 fuel pellets are loaded into hollow zirconium fuel tubes called fuel rods. A column of fuel pellets, typically 150 inches long, is loaded into a fuel rod. The fuel tube is nominally 160 inches in length. The nominal unfilled length of the fuel rod that does not contain fuel pellets is called the "plenum." The plenum provides a space to accommodate fuel pellet axial thermal expansion during operation and fission product gases that are released during power operations. A spring is inserted into the top of the

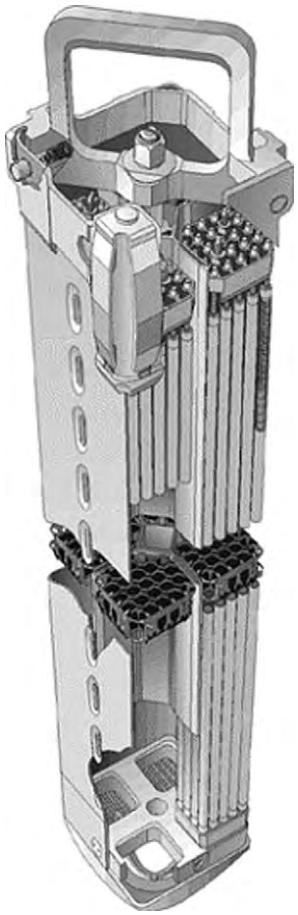


FIGURE 7 Schematic view of a boiling water reactor fuel assembly.

TABLE II

Boiling Water Reactor and Fuel Design Parameters

Design parameter	Typical value
Thermal power	3579 MWt
Electrical power	1150 MWe
Reactor pressure	1050 psia
Reactor coolant inlet temperature	533°F
Reactor coolant outlet temperature	551°F
Reactor coolant flow rate	104×10^6 lb m/hr
Number of fuel assemblies in core	800
Steam turbine inlet temperature	540°F
Steam turbine inlet pressure	965 psia
Fuel rod array	8×8
Fuel rods per assembly	62
Water rod locations per assembly	2
Fuel assembly cross section	5.4×5.4 inches
Fuel assembly length	176 inches
Fuel channel thickness	0.12 inch
Fuel rod outer diameter	0.483 inch
Fuel rod clad thickness	0.032 inch
Fuel pellet outer diameter	0.41 inch
Fuel pellet length	0.41 inch
Fuel assembly weight	220 kg
Fuel assembly uranium weight	183 kg

fuel rod to hold the fuel pellets down in the fuel tube during transportation to the reactor and end caps are welded onto the upper and lower ends of the fuel rod. Fuel rods are inserted into spacer grids that are fabricated from a zirconium or inconel alloy. Fuel rods are held in place by small leaf springs in the spacer grids. Depending on the fuel assembly design, from 36 to 81 fuel rod locations are available in a BWR fuel assembly. Most fuel assembly designs arrange fuel rods in a square array, with 7, 8, or 9 fuel rod locations on a side. Depending on the fuel assembly design, several fuel rod locations contain an open-ended zirconium alloy rod to allow water to flow the length of the fuel assembly as water boils in the rest of the reactor core. This open-ended rod is known as a "water rod." At the top and bottom of the fuel assembly are end fittings. The end fittings are typically machined form stainless steel and provide structural strength to the fuel assembly and help control inlet and outlet flow from the reactor core during operation. A BWR fuel assembly has a square cross section fuel channel that encloses the fuel assembly on all four sides. The fuel channel limits the flow of water from one fuel assembly to adjacent fuel assemblies in the reactor core. In the reactor pressure vessel, 400–800 fuel assemblies are arranged into an approximately cylindrical geometry.

As with the PWR, BWR fuel typically operates in a reactor for 3–5 years. The general characteristics of BWR and PWR spent fuels are similar. BWR spent fuel is stored on-site and is eventually shipped off-site for reprocessing or to a high-level radioactive waste repository for disposal. Table II provides a summary of key design characteristics of a typical boiling water reactor.

4. SHIP PROPULSION

Nuclear power is particularly suitable for ships, which need to be at sea for long periods without refueling, or for powerful submarine propulsion. Today, over 150 ships are powered by small nuclear reactors. The United States Navy operates some 100

nuclear ships. Though most nuclear-powered vessels are submarines, they range in type from icebreakers to aircraft carriers.

The development of nuclear propulsion marked the transition of submarines from slow underwater vessels to warships capable of sustaining 20–25 knots while submerged for months. The *Nautilus* led to the parallel development of further submarines, powered by single pressurized water reactors, and an aircraft carrier, the *USS Enterprise*, powered by eight reactor units in 1960. A cruiser, the *USS Long Beach*, followed in 1961 and was powered by two of these early reactor units. Remarkably, the *Enterprise* remains in service. Over time, standardized reactor designs for naval propulsion were developed and built by both Westinghouse and General Electric. The U.S. technology was shared with Britain; nuclear-powered ship development in France, the former Soviet Union, and China proceeded separately.

The Soviet Union built 245 nuclear submarines between 1950 and 1994. The largest submarines are the *Typhoon* class, powered by twin 190-MWt PWR reactors. The Soviet Union developed both PWR and lead-bismuth cooled reactor designs, the latter using highly enriched (>90%) fuel. Eventually, three generations of Russian submarine PWRs were utilized, the last entering service in 1987. At the end of the Cold War, in 1989, there were over 400 nuclear-powered submarines operational or being built. Some 250 of these submarines have now been scrapped and some on order were canceled, due to weapons reduction programs. Russia and the United States had over 100 each; the United Kingdom and France had less than 20 each and China had 6. The United States is the main navy with nuclear-powered aircraft carriers; both the United States and Russia have had nuclear-powered cruisers. Russia has eight nuclear icebreakers in service or being built.

Development of nuclear merchant ships began in the 1950s but has not been commercially successful. The U.S.-built *NS Savannah* was commissioned in 1962 and decommissioned 8 years later. It was a technical success, but not economically viable. The German-built *Otto Hahn* cargo ship and research facility sailed some 650,000 nautical miles on 126 voyages in 10 years without any technical problems. However, it proved too expensive to operate and was converted to diesel. The Japanese *Mutsu* was the third civil vessel. It faced technical and political problems. These three vessels used reactors with low-enriched uranium fuel.

In contrast, nuclear propulsion has proved both technically and economically feasible in the Soviet Arctic. The power levels and energy required for icebreaking, coupled with refueling difficulties for other types of vessels, are significant factors. The icebreaker *Lenin* was the world's first nuclear-powered surface vessel and remained in service for 30 years, though new reactors were fitted in 1970. It led to a series of larger icebreakers, the *Arktika* class, launched beginning in 1975. These vessels have two reactors and are used in deep Arctic waters. The *Arktika* was the first surface vessel to reach the North Pole. For use in shallow waters such as estuaries and rivers, shallow-draft *Taymyr*-class icebreakers with one reactor are being built in Finland and fitted with their nuclear steam supply system in Russia. They are built to conform to international safety standards for nuclear vessels.

Naval reactors are pressurized water types, which differ from commercial reactors producing electricity in that they have high power density in a small volume and therefore run on highly enriched uranium (>20% uranium-235) and long core lives, so that refueling is needed only after 10 or more years. New cores are designed to last 50 years in carriers and 30–40 years in submarines. Decommissioning nuclear-powered submarines has become a major task for both the United States and Russia. After removing the spent fuel, normal practice is to cut the reactor section from the vessel for disposal in shallow land burial as low-level waste.

SEE ALSO THE FOLLOWING ARTICLES

Nuclear Engineering • Nuclear Fuel: Design and Fabrication • Nuclear Fuel Reprocessing • Nuclear Fusion Reactors • Nuclear Power Economics • Nuclear Power, History of • Nuclear Power Plants, Decommissioning of • Nuclear Power: Risk Analysis • Nuclear Proliferation and Diversion • Nuclear Waste • Occupational Health Risks in Nuclear Power • Public Reaction to Nuclear Power Siting and Disposal

Further Reading

Duke Power Company (1996). "McGuire Nuclear Station Updated Final Safety Analysis Report, May 1996." Duke Power Company, Charlotte, North Carolina.

International Atomic Energy Agency (IAEA). (2001). "Energy, Electricity and Nuclear Power Estimates for the Period up to

2020." Reference Data Series No. 1, July 2001. IAEA, Vienna, Austria.
Mississippi Power and Light Company (1995). "Grand Gulf Final Safety Analysis Report, December 1995." Mississippi Power and Light Company (Entergy Corporation), New Orleans.
Uranium Information Center (2002). "Outline History of Nuclear Energy." Nuclear Issues Briefing Paper #50, February 2002. Uranium Information Center, Melbourne, Australia.
Uranium Information Center (2002). "Nuclear-Powered Ships." Nuclear Issues Briefing Paper #32, June 2002. Uranium Information Center, Melbourne, Australia.

Nuclear Fuel: Design and Fabrication

LEON C. WALTERS
Argonne National Laboratory
Idaho Falls, Idaho, United States

1. **Introduction**
2. **Design Considerations**
3. **Fabrication**
4. **Performance**
5. **Summary**

Glossary

blanket fuel Nuclear reactor fuel that contains the fertile isotopes that are bred into fissionable isotopes.
driver fuel Nuclear reactor fuel that contains the fissionable isotopes along with fertile isotopes that are bred into fissionable isotopes.
fertile isotopes Common fertile isotopes are ^{238}U and ^{232}Th.
fissionable isotopes Common fissionable isotopes are ^{235}U, ^{233}U, and ^{239}U.
natural uranium Uranium as it is mined from the earth. The composition is 99.3% ^{238}U and 0.7% ^{235}U.

Successful nuclear fuel must meet the requirements set forth by the nuclear reactor design team. These requirements include considerations of safety, economy, and ease of reprocessing or disposal. The environment within which nuclear fuel must perform, design considerations, fabrication, and performance are reviewed for nuclear fuel designs that have survived beyond the experimental stage.

1. INTRODUCTION

Economically recoverable uranium in the world is estimated to be 4 million tons. This uranium is composed of two isotopic forms, ^{235}U and ^{238}U. The two forms differ, with ^{235}U having three fewer neutrons in the nucleus. Otherwise, the two forms are identical chemically, meaning that they have the same number of electrons and protons. The uranium that is mined from the earth is called natural uranium. Natural uranium contains 0.7% ^{235}U, with the balance being ^{238}U. Some nuclear reactors were designed to operate with fuel made of natural uranium, whereas other reactor designs required fuel with a concentration of ^{235}U that is higher than 0.7%. For these latter reactor designs, the fuel had to be enriched in ^{235}U, over a range of concentrations, to as high as 93% ^{235}U for certain reactors.

Uranium that remains after the enrichment process is called depleted uranium because the ^{235}U concentration has been reduced to 0.3% or less. This depleted uranium can be converted to a fissionable isotope, ^{239}Pu, by a process called breeding. In reactors designed to optimize this breeding process, neutrons are absorbed by the nucleus of ^{238}U and the ^{238}U is converted to ^{239}Pu. It also should be pointed out that thorium can be converted to the fissionable isotope ^{233}U by the same neutron capture process. The reserves of thorium in the earth are far larger than those for uranium. Thus, if all the recoverable uranium and thorium in the world were converted to fissionable isotopes, there would be an ample energy supply for humankind for several millennia.

Reactor driver fuel is composed of the element or compounds of the fissionable isotopes ^{235}U, ^{233}U, and ^{239}Pu along with the fertile isotopes ^{238}U and ^{232}Th. Some fuels contain mixtures of more than one fissionable isotope. Reactor blanket fuel is composed of the element or compounds of the fertile isotopes ^{238}U and ^{232}Th. Fertile is a term used to indicate that the isotope can be bred to a fissionable isotope.

 © 2004 Elsevier Inc. All rights reserved.

The important nuclear processes that take place in the fuel are the following:

$$^{235}\mathrm{U} + n = \mathrm{X} + \mathrm{Y} + 2.5n + \mathrm{Energy} \quad (1)$$

$$^{238}\mathrm{U} + n = {}^{239}\mathrm{Pu} \quad (2)$$

$$^{232}\mathrm{Th} + n = {}^{233}\mathrm{U} \quad (3)$$

Most nuclear reactors deployed are fueled with uranium that either has the natural enrichment of 0.7% ^{235}U or is somewhat enriched in ^{235}U, to approximately 5%. These reactors, called thermal reactors, generate their power with new fuel according to fission reaction 1. The energy released in the fission reaction 1 is equivalent to 1 MW day for each gram fissioned. As the ^{235}U is fissioned, some of the ^{238}U in the fuel is converted to ^{239}Pu according to neutron capture reaction 2. The ^{239}Pu in turn is fissioned with approximately the same energy as for reaction 1. A fact that seems to go unnoticed is that approximately 40% of the power from a thermal reactor originates from plutonium at the end of the life of the fuel, in approximately 4 years. Thorium is converted to ^{233}U by the same process of neutron capture, reaction 3.

The X and Y of reaction 1 are called fission products. They represent most of the elements of the periodic chart, they are all radioactive, with relatively short half-lives, and they can exist in the fuel as a solid, as a gas, or, under extreme conditions, as a liquid. Their presence in the fuel and their chemical and physical behavior are critical to the performance of the fuel.

Another series of successive neutron capture reactions occurs, such as reactions 2 and 3. The products of these neutron capture reactions are the so-called minor actinides, americium, curium, and neptunium. They are also radioactive but their half-lives are extremely long. Their presence in the spent fuel is the reason for high-level waste repositories that must maintain their integrity for millennia.

The coolant in a thermal reactor is generally water or occasionally helium gas. The excess neutrons from fission reaction 1 are slowed down or moderated by nuclear collisions in the water or, in the case of the gas-cooled reactor, with graphite designed for the purpose of slowing down the neutrons. The fission reaction produces highly energetic neutrons that are slowed down to near equilibrium with the thermal motion of the surrounding material. The name "thermal" reactor thus arises by virtue of this slowing-down process. These low-energy neutrons have a higher likelihood of producing fission according to reaction 1.

The ^{238}U that is converted to the fissile ^{239}Pu is called fertile blanket fuel. The ^{232}Th that is converted to the fissile ^{233}U is also called fertile blanket material. A "fast" breeder reactor is specifically designed to optimize the conversion of fertile material to fissile material by elimination of as much moderating material as possible to keep the neutrons as energetic as possible. When designed properly, the fast breeder reactor can easily produce more fissile material from the fertile blanket material than is fissioned to produce energy. The energy density in a fast reactor is much higher than that of a thermal reactor and thus liquid metals, such as sodium, which has a much higher heat capacity than water, are used as the flowing coolant.

Chicago Pile 1 was the first nuclear reactor that achieved criticality on December 2, 1942. It was designed, built, and brought to the world's first chain reaction by a team led by Enrico Fermi. The fuel for this reactor was natural uranium dioxide pellets placed in a matrix of graphite blocks that served as the neutron moderator. The reactor produced little power and thus the fuel design was of the most rudimentary nature, being simply unclad pellets. Since that time there have been a wide variety of nuclear fuels developed, the designs for which were dependent on the mission of the nuclear reactor. Nuclear reactors have been constructed for electricity generation, breeding of uranium and thorium for the production of fissile isotopes, production of medical isotopes, tumor irradiation, and ship and space propulsion.

This article discusses the phenomena that must be considered for the design of any fuel type and then shows how the effects of these phenomena have been accommodated in the design of specific fuel types. This discussion is limited mostly to fuel types that have survived the early experimental ventures and have gone on to be candidates for commercial deployment.

2. DESIGN CONSIDERATIONS

The geometry of fuel has many variations. The radial dimension, the dimension toward the coolant, is always small and is dictated by the necessity that the energy density of the fuel during the fissioning process is enormous and a great deal of heat must be extracted to prevent the melting of the fuel. Figure 1 is a simplified drawing of a generic fuel pin, shown only

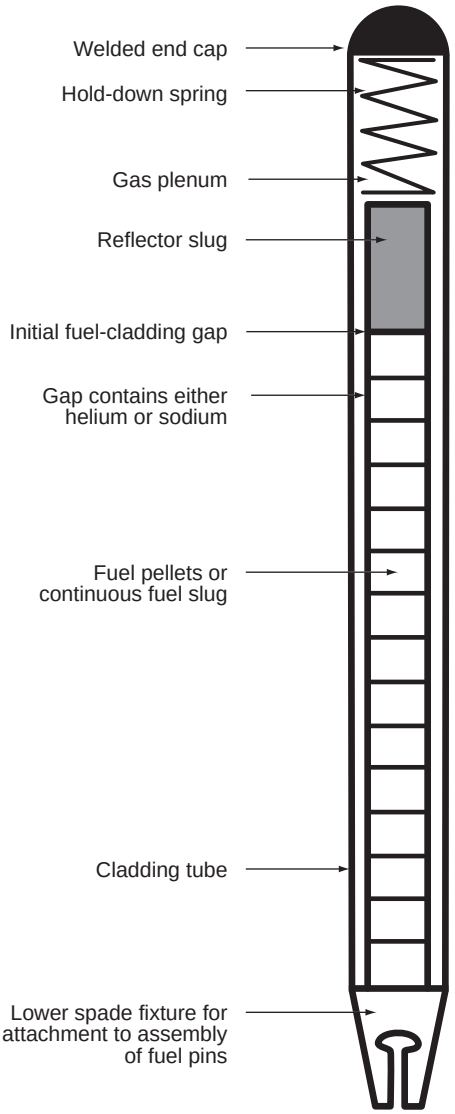


FIGURE 1 Schematic of a generic fuel pin illustrating important components.

to illustrate particular design features. Because of the small dimension between the center of the fuel and coolant, the temperature gradient is large, and this large temperature gradient causes cracking of the fuel, drives mass flow, and creates stress between the fuel and cladding due to thermal expansion differences.

The cladding is generally a metal tube that is sealed around the fuel. The cladding serves several purposes. It is usually referred to as the first barrier against the release of radioactive elements to the environment. As such, its integrity is a very important consideration. For some fuel designs, the cladding is necessary to support the fuel column and to preserve the geometry and heat transfer path of a bundle of fuel pins. Furthermore, the cladding, for many fuel pin designs, is necessary to contain a fuel transfer medium in the gap between the fuel column and the cladding. This fuel transfer medium could be either a gas, such as helium, or a liquid, such as sodium.

The environment within which the fuel pin must perform is severe. When the reactor is brought to power, a number of events occur simultaneously. High temperatures and high temperature gradients are imposed on the fuel pins. A number of high-energy particles begin disrupting the matrices of the fuel and cladding. The fission product elements cause the fuel to swell, eventually reaching the cladding. The stress generated can eventually breach the cladding. This is called the fuel–cladding mechanical interaction. Some also call this phenomenon the pellet–cladding interaction. Fission products that form gases escape from the fuel and create additional stress on the cladding as the pressure builds with time. Some of the fission products, such as iodine and cesium, may attack the cladding by corrosion reactions.

Cladding surrounding the fuel column experiences a similar set of challenges. Not only is the cladding subject to high-temperature corrosion processes from the inside but the coolant flowing on the outside of the cladding, whether it be water, liquid metal, or gas, may be corrosive to the cladding. Furthermore, the physical properties of the cladding may change dramatically under prolonged exposure to neutron irradiation. Neutron irradiation tends to embrittle most metals, some metals swell under neutron bombardment, and processes that depend on diffusion of material, such as carbide precipitation and creep, may be greatly accelerated by neutron irradiation.

Within the environment just described, the reactor designer will lay forth a set of requirements that he or she anticipates can be met by the fuel system. The inlet and outlet temperatures of the coolant will be specified, which fixes the best efficiency of the system. The reactor designer designates the exposure time of the fuel in the reactor, which is usually measured in terms of the percentage of uranium and plutonium fissioned in the fuel or the energy generated per ton of fuel. Exposure time of the fuel is important as an economic measure for the supply of fuel that must be manufactured as well as the amount of spent fuel that must be stored or reprocessed. Should the fuel be reprocessed by a particular reprocessing scheme, then the fuel must be compatible with the scheme.

The regulatory requirements demand that the fuel system survive possible off-normal events without impact to the public. Off-normal events could

involve loss-of-coolant flow conditions where the temperature of the fuel gradually increases above normal conditions over an extended period of time or more rapid events where a control rod might be inadvertently removed.

The reactor designer and regulatory system impose a set of expectations for the nuclear reactor fuel system and the environment in the reactor imposes its own set of constraints. A final qualified fuel design is always the result of many iterations and compromises. The fuel development team, usually composed of materials scientists, is supported by experts in neutronics, thermal hydraulics, core design, cladding and duct development, nuclear safety, chemistry, remote examination technology, experiment vehicle design, analytical chemistry, microscopy, and more.

3. FABRICATION

For a new fuel type, the means of fabricating the fuel is always an iterative and challenging process. Generally, small versions of the envisioned full-scale fuel pin are fabricated, irradiated in a reactor under prototypic conditions, and examined. The range of fabrication variables may be quite extensive at this experimental stage as the search proceeds for the best combination. Through a series of irradiations and examinations, the variable range is narrowed until an interim fuel fabrication specification is agreed on. At this point, a number, perhaps six, of full-size assemblies of fuel pins are irradiated under prototypic conditions to their end of life, which generally means to the point of cladding failure. Along the way, during this test period, a few of the fuel pins would be removed from the assemblies for testing under off-normal conditions in a reactor specifically designed for that purpose. If the results of these tests did not meet expectations, then there may be additional tests on improved versions. Should the results meet expectations, the fabrication specifications would be frozen. Strict quality assurance programs would be invoked in the fabrication process to ensure that the fuel performance would be repeatable.

The process of moving from a nuclear fuel concept, dictated by a reactor design, to a fully qualified and licensed fuel system takes 10 years or more. The schedule for conducting such a program is controlled by the time it takes to complete the fabrication, irradiation, and examination of a novel fuel type. Irradiation of nuclear fuel takes 4 to 5 years to complete and generally more than one iteration is required.

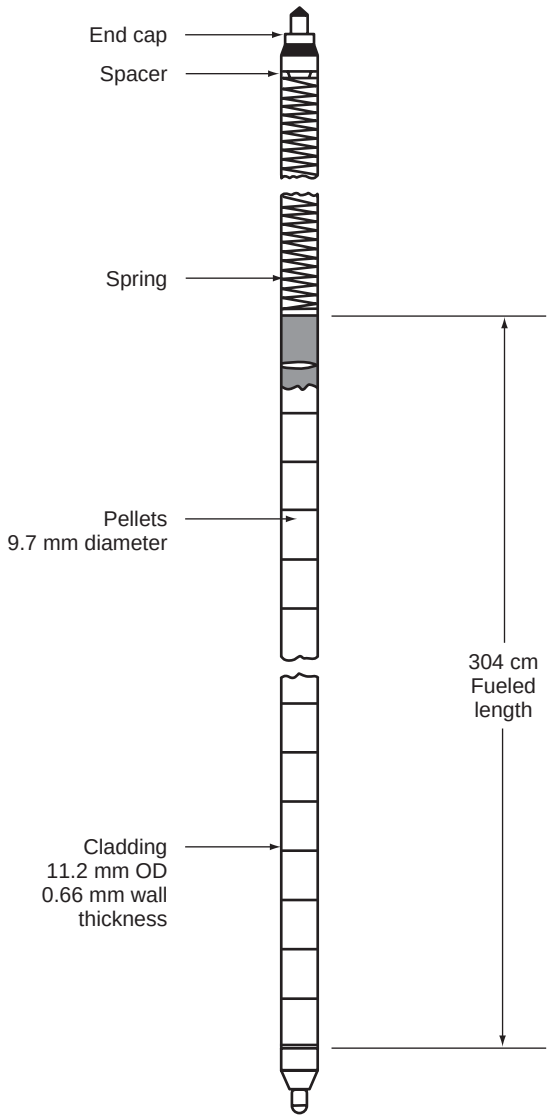


FIGURE 2 Fuel pin of a pressurized water reactor.

Figures 2, 3, and 4 show three of the many fuel types that have been developed. They were chosen because they represent classes of fuel pins that are commercially deployed or they are under active investigation for the next generation of reactors. Figure 2 shows the fuel pin utilized in a typical pressurized light water reactor. The fuel pin has evolved from nearly 50 years of utilization, first in experimental reactors and for the past 40 years in commercial reactors. The fuel column is a stack of uranium oxide or uranium and plutonium oxide fuel pellets that is held in place with springs. Prior to assembly, the cladding tube that surrounds the pellet

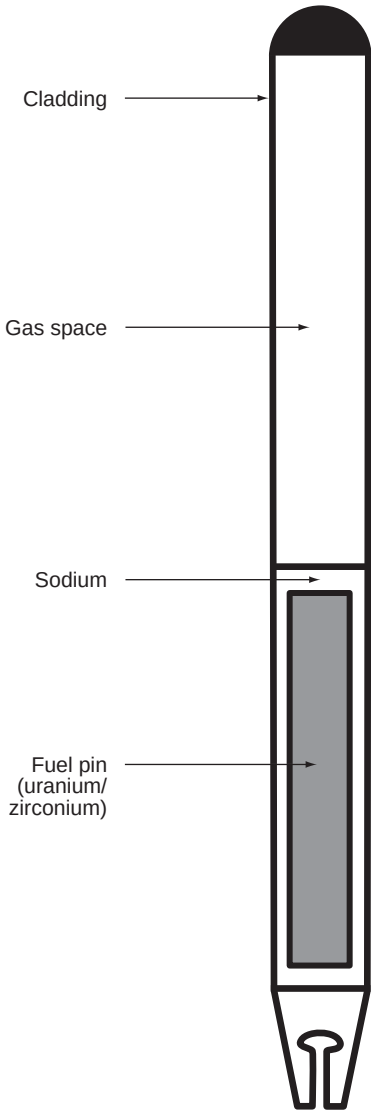


FIGURE 3 Fast reactor metal-fuel pin.

stack is filled with helium at a somewhat elevated pressure. The helium gas serves as a good heat transfer medium and the elevated pressure of the helium prevents the cladding from collapsing around the fuel column from the external pressure of the water in the pressurized water reactor. The technology for hot-pressing the oxide pellets has been well perfected to achieve high and uniform density as well as consistency in stoichiometry. The zirconium and zirconium alloy cladding has evolved over the years to a composition and heat treatment that are resistant to internal and external corrosion effects.

Figure 3 shows a fast reactor fuel pin with a metallic fuel column. The fuel column for a fast reactor fuel pin can be uranium and plutonium

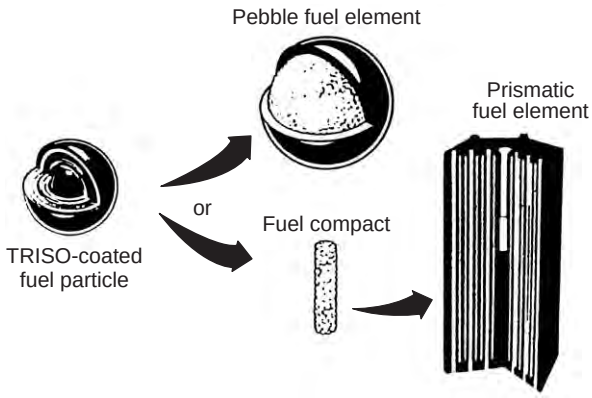


FIGURE 4 Particle fuel element.

oxide, carbide, nitride, or an alloy of uranium and plutonium. The cladding is steel. No clear choice exists for the most preferable fuel column type. Each type has its attributes. Uranium–plutonium oxide is by far the most commonly used fast reactor fuel. The PUREX (plutonium uranium extraction) fuel reprocessing technology, used for light water oxide fuel, also is acceptable for the oxide fast reactor fuel. Alternatives, but less mature reprocessing schemes, have been developed for the other fast reactor fuel types. The first cladding material for fast reactor fuel was commercial grades of austenitic stainless steel. These steels tended to swell when subjected to a high level of neutron irradiation. The steel alloys were improved to the point where little swelling occurs and the fast reactor fuels are capable of achieving very high fissile burn-up.

Most fuel pin designs are cylindrical in shape, for both fast reactor and light water reactors. Figure 4 shows the fuel design for a type of high-temperature gas-cooled reactor (HTGR) that is a departure from the cylindrical designs. These fuel types have undergone successful development in several countries; e.g., South Africa is implementing a program for commercial deployment of these HTGRs. The diameter of the fuel particle in Fig. 4 is dependent on the particular design. The central sphere, called the kernel, contains the fissile material and ranges from 150 to 500 μm, depending on the design. The kernel is composed of uranium oxide, uranium carbide, as well as thorium included with the uranium. The kernel is embedded in two layers, the BISO concept, or the now accepted three-layer version, called the TRISO concept. BISO and TRISO stand for bi-structural isotropic and tri-structural isotropic, respectively. The isotropic refers to the property of the pyrolytic layers, the bi and tri to the

number of layers. The innermost layer is porous pyrolytic carbon and is called the buffer layer. This layer supplies the void space to accommodate fission-induced swelling and gas release from the kernel. The next layer is silicon carbide, which is the load-bearing layer that contains the pressure of the fission gases. The outer pyrolytic layer protects the silicon carbide layer from reaction with the coolant impurities. Each of the layers is 30 to 50 μm thick. These small TRISO particles are further embedded either in graphite spheres for the pebble bed reactor or in cylindrical fuel compacts that are located in vertical channels in hexagonal graphite blocks in the prismatic-compact reactor. Approximately 10,000 coated particles are contained in each 6 cm diameter graphite sphere.

4. PERFORMANCE

4.1 Metal Fuel

Metal fuel was among the first fuels to be used in nuclear reactors. The Experimental Breeder Reactor I, the first nuclear reactor to generate useful electricity, on December 20, 1951, used metal uranium and plutonium fuels. Subsequently, several fast reactors used metal fuel. Metal fuel was not adopted for use in water-cooled reactors because, in the event of a cladding breach, the metal fuel was not compatible with the water coolant. The fuel will form metal hydrides and oxides when in contact with water at elevated temperature. However, metal fuel continued to be of interest for use in fast reactors.

In the mid-1960s, when the large commercial prototype fast reactors were being designed, metal fuels were largely abandoned in favor of oxide fuel. Metal fuel at that time was unable to achieve high burn-up and there was a concern that metal fuel would be unable to reach the high temperatures that reactor designers desired without the formation of a liquid phase between the fuel and the cladding. Even though the performance information on oxide was limited, it was thought that oxide fuel would not be limited in these respects.

Metal fuel possesses certain unique attributes and thus research continued to improve its performance. Metal fuel has a much higher fissile element density than all other fuels, which allows a metal fuel reactor to convert more uranium to plutonium. Metal fuel is easier and cheaper to fabricate, especially in a hot-cell environment, on reprocessed fuel. The thermal conductivity of metal fuel is higher than that of other fuels and this brings with it the possibility of designing large commercial reactors that are inherently safe, that is, reactors that will shut themselves down during an upset event, without any human intervention.

A design modification that allowed metal fuel to achieve very high burn-up was discovered. Early metal fuel was designed such that there was no gap or only a small gap between fuel and cladding. During irradiation, the accumulation of fission products would cause the fuel to swell. The gaseous fission products would nucleate into small bubbles, and when these bubbles grew into a critical size, the gas pressure in the bubble would overcome the surface tension of the uranium matrix and the fuel would swell rapidly. Compounding the swelling from the gaseous fission products was the swelling component due to the accumulation of solid fission products in the uranium matrix. The fuel swelling would cause the cladding to breach after just a few months at a burn-up of approximately 1% of the heavy metal.

A simple design change avoided the problem of early cladding breaches due to fuel swelling. The gap between the fuel column and the cladding was widened to allow the fuel to swell freely until the volume change of the fuel reached approximately 30%. When the volume change reached approximately 30%, the porosity in the fuel, caused by the fission gas bubbles, would interconnect and the fission gas would be released to a plenum volume above the fuel column. The plenum volume above the fuel was sized such that the stress on the cladding during the life of the fuel pin would remain reasonably low. Figure 5 shows the fractional gas release as a function of volume change for several metal fuel compositions. With this design change, metal fuel pins reached burn-ups in excess of 20% and were capable of remaining in the reactor for 3 to 4 years.

As with all fuel types, the fuel restructures substantially during irradiation. Figur 6 shows the cross section of a uranium–plutonium–zirconium alloy fuel pin at various elevations after 17% burn-up. The ring formation is a result of diffusion, driven by the temperature gradient, where the ring boundaries are phase boundaries. The interconnected porosity is evident throughout the cross sections. The centerline temperature of the fuel is approximately 750°C, whereas the cladding temperature is approximately 600°C.

The other perceived issue with metal fuels, that of a liquid phase forming before the design temperature is reached, never turned out to be a problem.

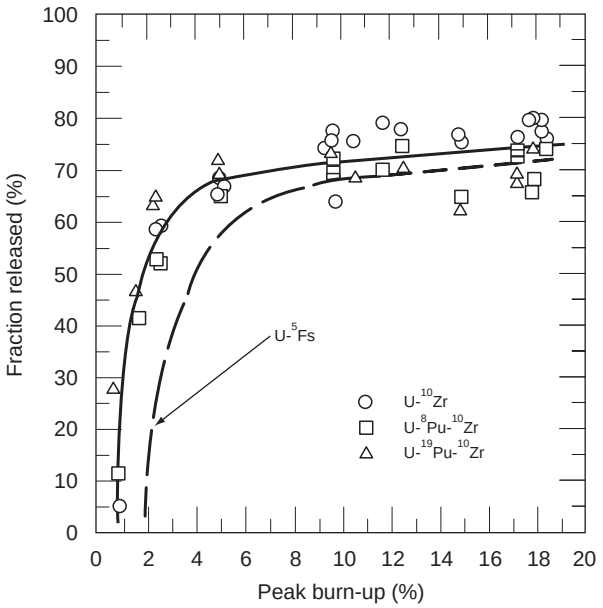


FIGURE 5 Fission gas released to the plenum, above the fuel, for various metal fuels as a function of burn-up.

Cladding steels had a peak useful temperature of no more than 600°C, even though reactor designers desired higher temperatures. At approximately this temperature, the steels used experience a decrease in yield strength such that the stress on the cladding from the fission gas and fuel swelling would cause the cladding to deform and fail. The temperature of 600°C was below the temperature where a liquid phase would form between the fuel and the cladding.

A phenomenon that persisted to limit the capability of fast reactor fuels, whether the fuel was metal or ceramic, was the swelling of the cladding and duct materials due to the neutron damage. Figure 7 shows the progressive improvement in cladding deformation. 304L SA and 316 SA are common solution-annealed austenitic stainless steels. D9 20% CW is a titanium modified 316 stainless steel that has been cold-worked, while HT-9 20% CW is a ferritic steel that has been cold-worked and has, by far, the best performance. This issue is no longer considered a limiting factor with the utilization of ferritic steels for cladding that show very little irradiation-induced swelling.

4.2 Ceramic Fuels

Uranium oxide is by far the most utilized nuclear fuel for commercial power reactors in the world. Mixed uranium and plutonium oxide fuel was the primary choice for sodium-cooled fast reactors and in addition is being increasingly used in thermal reactors as a means of reducing the increasing stockpile of reprocessed plutonium. The fuel in fast

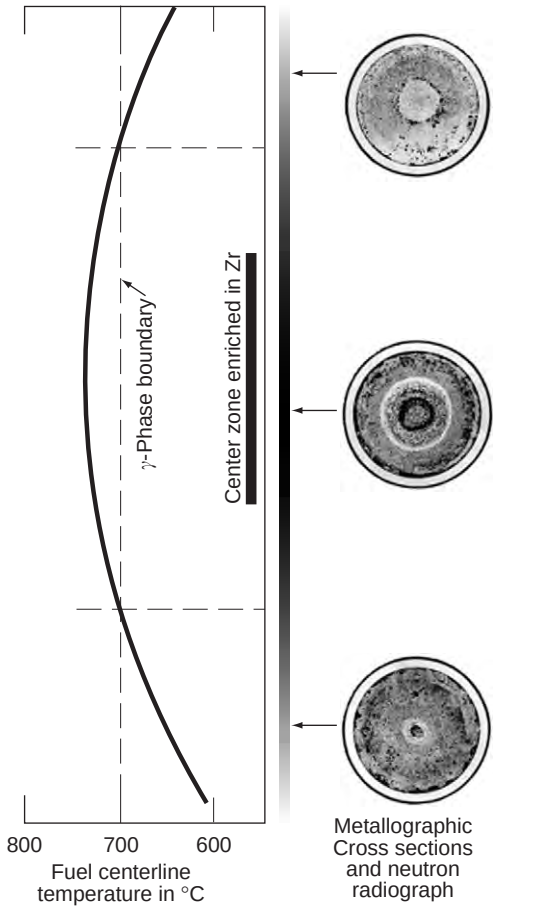


FIGURE 6 Axial temperature profile along the centerline of a U-Pu-Zr pin at 17 at.% burn-up and its relation to radial zone patterns shown in a neutron radiograph and three cross sections.

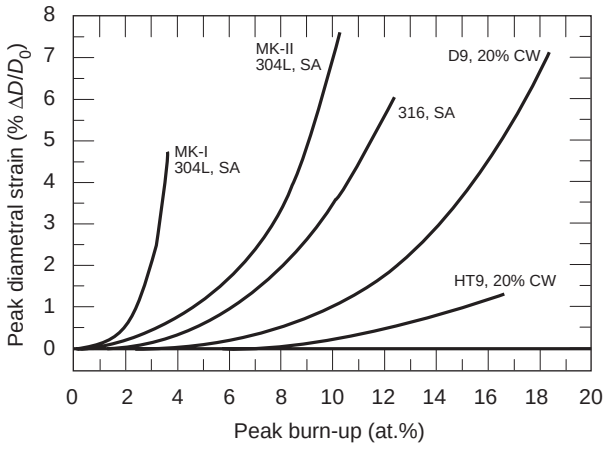


FIGURE 7 Progressive improvement in cladding deformation (swelling and creep) of metal fuel pins.

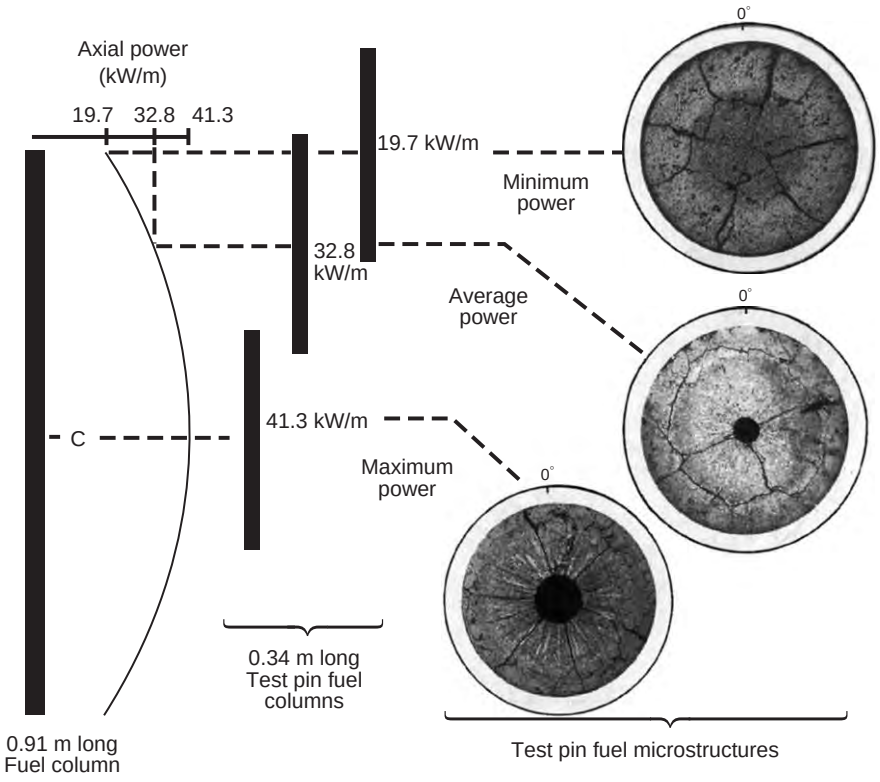


FIGURE 8 Mixed-oxide fuel restructuring versus linear heat rate.

reactors operates at a much higher fissile and energy density and thus the peak fuel and cladding temperatures are much higher. Oxide fuel columns restructure to a great extent because of the low thermal conductivity of the fuel and the high temperature gradient from the center of the fuel to the fuel–cladding interface. The peak centerline temperature of fast reactor fuel is approximately 2100°C, whereas the cladding temperatures reach 600°C in normal operation. The corresponding temperatures for a pressurized water reactor oxide fuel are 1300 and 350°C, respectively. Thus, the restructuring for fast reactor fuel is more severe. Figure 8 shows the axial variation in fuel restructuring for a fast reactor, a uranium–plutonium oxide fuel pin. At the core midplane location, the restructuring leads to a central void. The void results from the transport of the initial porosity up the thermal gradient. The high temperature results in significant grain growth. In addition to grain growth, the large thermal gradient causes a radial redistribution of oxygen and plutonium.

The fission gas that is generated behaves quite differently in thermal and fast reactor fuels. Because of the difference in fuel temperatures, the thermal reactor fuel retains most of the fission gas, whereas the fast reactor fuel releases most of the fission gas to the gas plenum in the fuel pin. The margins to failure, as a result of the gas retention or release, are considered in the design of the fuel. For thermal reactor fuel, the concern is that the fission gas will be suddenly released during an upset event and overpressurize the cladding. This eventuality has been modeled, understood, and taken into account in the allowable exposure of the fuel element. With the fast reactor fuel, the plenum must be large enough to hold the continuous generation of fission gas without overpressurizing the cladding.

Cladding attack from the volatile fission products, cesium, tellurium, and iodine, is of concern for both fast reactor and thermal reactor fuels although the controlling phenomena are different. With the zirconium and zirconium alloy claddings, early failures were due to stress corrosion cracking from the iodine. In addition, zirconium hydride formation on the outside of the thermal reactor zirconium cladding led to cladding breach. However, these problems have become well understood and were solved, so that modern thermal reactor fuel is highly reliable. For fast reactor cladding, which is steel, the volatile fission products, along with excess oxygen, tend to oxidize the cladding components, such as the

chromium. Control of the initial oxygen-to-metal ratio in the fuel minimized the extent of the oxidation to the point that the oxidation was not a limiting factor in the operation of the fuel.

Development of uranium oxide fuel for thermal reactors and uranium–plutonium fuel for both fast and thermal reactors has matured to the point where they are considered highly reliable and economic fuels. The only disadvantage of oxide fuel for fast reactors is that the fissile density is low, which makes an oxide-fueled fast reactor an inefficient breeder of plutonium.

4.3 Particle Fuels

Particle fuels have a long history of development. The Dragon reactor in the United Kingdom, the Peach Bottom and Fort St. Vrain in the United States, and the AVR and THTR-300 in Germany all operated with particle fuel. The HTTR in Japan and the HTR-10 in China are still operating with particle fuel.

Failure of a fuel pin is generally regarded as the point where the cladding fails and fission products are released to the coolant. In the case of particle fuel, failure is the point where the protective coatings around the particle fail and fission products are released to the coolant. The "ameba effect" was a failure mechanism recognized in early irradiations of particle fuel. When the particles were subjected to steep temperature gradients at high temperature or when fuel fabrication defects yielded asymmetrical kernel geometry, carbon transport from the high-temperature side of the kernel would tend to migrate to the low-temperature side. The kernel would then tend to move to the high-temperature side, ultimately breaching the carbon and silicon carbide layers. This phenomenon is well understood and is avoided by a combination of the proper kernel composition and the reduction of the imposed temperature gradients across the particle during irradiation.

Excessive anisotropy of the pyrolytic graphite layers, which surround the kernel, has led to failure and fission gas release. The need to define the exact coating conditions to avoid the critical anisotropy levels makes fabrication specifications important, where the recipes must be passed to industry and other laboratories.

Fission products react with the silicon carbide layer. In particular, the palladium group metals diffuse into the silicon carbide layer and form intermetallic compounds that can lead to failure of the layer. By keeping the temperature and time-in-reactor within known limits, failure of the layer is avoided.

Although in many respects particle fuel is complex, in both fabrication and performance the fuel has been perfected and has been shown to be highly reliable. In Germany, the fuel has been irradiated to 15% burn-up at a temperature of 1250°C. Statistical analysis has shown a failure rate for this fuel of 2×10^{-5} at the 95% confidence level. However, the German experience is better than that experienced in the development of particle fuel by other countries.

4.4 New Fuels

For completeness, new fuels that are under development are briefly discussed in this section. Proliferation issues and concerns surrounding nuclear waste repositories have given rise to a generation of special-purpose nuclear fuels. Relative to the proliferation of nuclear weapons, an issue existed regarding the diversion of highly enriched fuel, which is used in research reactors around the world. A large number of these reactors were converted to low-enriched fuel that consisted of a moderately high density of uranium silicide particles, embedded in an aluminum matrix. The fuel was fabricated in the form of plates. However, the fissile density of this fuel was not high enough to allow the conversion of the larger research reactors. To solve this problem, an effort was initiated to develop a much higher density low-enriched fuel such that these larger research reactors could be converted. This effort has been successful with the development of a fuel plate design that has metal uranium–molybdenum particles embedded in an aluminum matrix.

As discussed, the elements that dictate the longevity of a nuclear spent fuel repository are the actinides—plutonium, americium, neptunium, and curium. With fast reactors and the proper spent fuel reprocessing cycle, these elements are recycled into the reactor and fissioned to achieve useful energy. With the recent and perhaps temporary rejection of fast reactors and the associated fuel cycle due to proliferation concerns, it was proposed that neutron-generating accelerators could fission the actinides that are in the commercial thermal-reactor spent fuels and thus reduce the radioactive burden to nuclear waste repositories. New reactor fuels were required that would contain a high concentration of actinides in a matrix that was inert in the sense that the matrix would contain no uranium that could be converted to higher actinides. A number of promising options are being explored, such as the coated

particle option, where the particle kernel would be an actinide oxide or a carbide. Metal fuels are also being proposed as an option where the actinides would alloy in a matrix of zirconium. A similar option is a mixture of actinide nitrides in a matrix of zirconium nitride. Finally, a series of fuels is being studied where an oxide or other nonmetallic compound of the actinides is embedded in an inert ceramic matrix.

5. SUMMARY

Nuclear fuel development has progressed remarkably over the past 50 years. The fuel residence times in the reactors have increased by more than a factor of 20 and the reliability has increased by several orders of magnitude. Fuel is a small cost of generating power from a nuclear reactor as opposed to fuel costs from other fossil fuel energy technologies.

Nuclear fuel design and fabrication techniques will continue to evolve, with their evolution driven by economics and new uses for nuclear energy. A particular fuel design must not only meet the needs of the reactor designer but also must be economical to fabricate. Furthermore, if the fuel is to be reprocessed, it must be amenable to existing reprocessing schemes or new reprocessing technology must be developed to accommodate the fuel. Some fuel types are subjected only to once-through cycles with no reprocessing. In this case, the spent fuel and secondary containment must meet the requirements of the repository.

Looking to the future, high-temperature nuclear reactors for space travel will surely be commonplace. High-temperature fuels, such as nitride fuel, will continue to be developed for this purpose. Ground transportation will move toward hydrogen as a fuel not only to reduce carbon emissions for environmental benefits but also for many countries to reduce their dependence on foreign oil. Nuclear reactors with high outlet temperatures when coupled with thermochemical cycles will be used for hydrogen production. Here, particle fuels or yet to be developed fuels will be used in these reactors. Water resources will diminish in this century. Nuclear reactors used for seawater desalination purposes will be developed that require long-life fuel. These reactors could be deployed in developing countries, with no need for refueling for two or three decades. All of these future uses will continue to challenge fuel developers for decades.

SEE ALSO THE FOLLOWING ARTICLES

Nuclear Engineering • Nuclear Fission Reactors: Boiling Water and Pressurized Water Reactors • Nuclear Fuel Reprocessing • Nuclear Fusion Reactors • Nuclear Power Economics • Nuclear Power, History of • Nuclear Power Plants, Decommissioning of • Nuclear Power: Risk Analysis • Nuclear Proliferation and Diversion • Nuclear Waste • Occupational Health Risks in Nuclear Power • Public Reaction to Nuclear Power Siting and Disposal

Further Reading

Cahn, R., Haasen, P., and Kramer, E. (Eds.). (1994). "Materials Science and Technology, Volumes 10A and 10B, Nuclear Materials," Part I. VCH, New York.

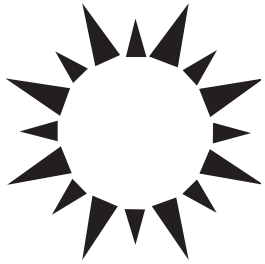
Gulden, T., and Nickel, H. (1977). Coated particle fuels. *Nucl. Technol.* **35,** 206–213.

Leggett, R., and Walters, L. (1993). Status of LMR fuel development in the United States of america. *J. Nucl. Mater.* **204,** 23–32.

Olander, D. (1976). "Fundamental Aspects of Nuclear Reactor Fuel Elements." Technical Information Center, Energy Research and Development Administration, Oak Ridge, TN.

Weisman, J., and Eckart, R. (1981). Basic elements of light water reactor fuel rod design. *Nucl. Technol.* **53,** 326–343.

Nuclear Fuel Reprocessing



HAROLD F. MCFARLANE
Argonne National Laboratory
Argonne, Illinois, United States

1. **Role of Reprocessing**
2. **Spent Nuclear Fuel**
3. **Reprocessing**
4. **Reprocessing Facilities**
5. **Reprocessing Research and Development**

Glossary

boiling water reactor A thermal-spectrum nuclear power reactor in which the cooling water is allowed to boil as it rises through the core.

burn-up The fraction of heavy metal atoms fissioned in irradiated nuclear fuel, often expressed as average energy produced per unit mass of initial uranium.

centrifugal contactor The solvent extraction equipment in which separation of the agitated phases is accelerated by centrifugal action.

cladding A metal tube that separates the nuclear material in a fuel element from the reactor coolant.

decontamination The removal of an undesirable component from a product.

depleted uranium The by-product of the uranium enrichment process; uranium containing less than 0.7% ^{235}U.

electrolytic reduction The use of an electric current in an electrolyte to promote the reduction of a ceramic fuel to a metal.

electrorefining The process of decontamination by selectively oxidizing a metal at an anode in an electrolyte while simultaneously reducing it in purer metallic form on a cathode.

enrichment The process of partial isotopic separation in order to increase the fraction of one isotope relative to another, typically ^{235}U relative to ^{238}U.

fast-spectrum reactor A nuclear reactor with a highly energetic neutron spectrum.

fissile An isotope that readily undergoes fission by thermal neutrons.

irradiated nuclear fuel The portion of nuclear fuel that has been exposed to sufficient neutron flux within a reactor to cause measurable nuclear transmutation.

melt refining The partial decontamination of metallic nuclear fuel by a process of melting and allowing fission products to volatilize or react with a substrate.

mixer-settler The equipment for solvent extraction, consisting of a region that thoroughly mixes two liquid phases and a settling region that allows them to separate again,

partitioning The separation of spent fuel into multiple like groups.

plutonium–uranium extraction (PUREX) A solvent extraction process that has been commercialized for reprocessing spent nuclear fuel.

pressurized water reactor A thermal-spectrum nuclear power reactor in which the system pressure is sufficiently high to prevent boiling of the cooling water.

pulsed column The equipment for solvent extraction, consisting of a tall pipe in which the two liquid phases flow in opposite directions, mixed by pressure pulses that force them through perforated disks inside the column.

pyrochemical A high-temperature reprocessing technique that uses selective reduction and oxidation in fused salts or metals to recover nuclear materials.

pyrometallurgical A high-temperature reprocessing technique that processes the materials in metal form.

raffinate The residue of contaminants removed from a product stream.

rare earths The lanthanide group of 14 elements.

reduction–oxidation (REDOX) The first countercurrent solvent extraction process chosen for reprocessing spent nuclear fuel.

reprocessing The separation of irradiated nuclear fuel into potentially useful product materials and waste.

sodium-cooled reactors Fast-spectrum nuclear reactors cooled with molten sodium.

solvent extraction A technique used in chemical separations involving two immiscible solvents in which one or more components are transferred between the solvents.

spent nuclear fuel Assemblies containing nuclear fuel that has reached the end of its economic lifetime.

thermal-spectrum reactor A nuclear reactor in which the neutron spectrum approaches thermal equilibrium with the core.

tonne A metric ton ($=1000$ kilograms).

transmutation The act of changing one element into another through nuclear processes.

 © 2004 Elsevier Inc. All rights reserved.

transuranic extraction (TRUEX) A solvent extraction process for transuranic elements.
tritium The radioactive hydrogen isotope with mass number 3.
vitrification The stabilization of waste in a glass matrix.
zeolite A mineral with a cagelike molecular structure; used as a catalyst or an ion exchanger.

Nuclear fuel reprocessing is the separation of irradiated nuclear fuel into potentially useful product materials and waste. The separation is accomplished by a combination of mechanical, chemical, and physical processes. Conservation of uranium resources and improved management of radioactive waste are the two primary motivations for reprocessing. In principle, almost all of the constituents of irradiated nuclear fuel can be recycled for some further use. In practice, only the recycling of plutonium and uranium has been of commercial interest.

1. ROLE OF REPROCESSING

Reprocessing is essential to closing the nuclear fuel cycle. Natural uranium contains only 0.7% ^{235}U, the fissile isotope that produces most of the fission energy in a nuclear power plant. Prior to being used in commercial nuclear fuel, uranium is typically enriched to 3–5% in ^{235}U. If the enrichment process discards depleted uranium at 0.2% ^{235}U, it takes more than 7 tonnes of uranium feed to produce 1 tonne of 4%-enriched uranium. Nuclear fuel discharged at the end of its economic lifetime contains less than 1% ^{235}U, but still more than the natural ore contains. Less than 1% of the uranium that enters the fuel cycle is actually used in a single pass through the reactor.

The other naturally occurring isotope, ^{238}U, directly contributes in a minor way to power generation. However, its main role is to transmute into plutonium by neutron capture and subsequent radioactive decay of unstable uranium and neptunium isotopes. ^{239}Pu and ^{241}Pu are fissile isotopes of plutonium that produce more than 40% of the fission energy in commercially deployed reactors. It is recovery of the plutonium (and to a lesser extent the uranium) for use in recycled nuclear fuel that has been the primary focus of commercial reprocessing. Uranium targets irradiated in special-purpose reactors are also reprocessed to obtain the fission product ^{99}Mo, the parent isotope of technetium, which is widely used in medical procedures. Among the fission products, recovery of expensive metals such as platinum and rhodium is technically achievable but is not economically viable in current market and regulatory conditions.

During the past 60 years, many different techniques for reprocessing spent nuclear fuel have been proposed and tested in the laboratory. However, commercial reprocessing has been implemented using a single type of aqueous solvent extraction technology, plutonium–uranium extraction (PUREX). Similarly, hundreds of types of reactor fuels have been irradiated for different purposes, but the vast majority of commercial fuel is uranium oxide clad in zirconium alloy tubing. As a result, commercial reprocessing plants have relatively narrow technical requirements for spent nuclear fuel that is accepted for processing.

2. SPENT NUCLEAR FUEL

The majority of commercially reprocessed nuclear fuel is discharged from pressurized water reactors, the type of reactor that accounts for two-thirds of the world's nuclear generating capacity. Most of the balance is made up of fuel from boiling water reactors, which comprise more than 20% of the capacity. Though there are many variations in fuel design for specific reactors, the fuel assemblies for these two types of nuclear reactors are similar in physical and chemical composition. (Fig. 1).

The composition of spent nuclear fuel depends on the fission energy generated within the fuel assembly during irradiation in the nuclear reactor. Burn-up, usually expressed in megawatt-days/tonne of initial uranium, is a measure of the total energy produced by a fuel assembly. Average burn-up has increased with improvements in fuel technology; 40,000 MW-days/tonne (MWd/tonne) being typical of spent fuel discharged at the beginning of the century. The mass of uranium contained in two fresh fuel assemblies for a pressurized water reactor is approximately 1 tonne. Each spent fuel assembly includes approximately 150 kg of radioactive hardware that holds the fuel pellets in place. At 40,000 MWd/tonne burn-up, the spent fuel assembly contains approximately 94.5% unused uranium, 1% plutonium, 0.2% minor actinides (neptunium, americium, and curium), and 4.3% fission products by weight. The fission products, which span the range from zinc to holmium on the periodic chart, include metals, rare earths, and gases. A small amount of tritium is also produced during ternary fission.

3. REPROCESSING

The status of large-scale commercial reprocessing is defined by the thermal oxide reprocessing plant (THORP) in Sellafield, United Kingdom, operated by British Nuclear Fuels Plc (BNFL), and the UP2 and UP3 plants in France, operated by Compagnie Générales des Matières Nucléaires (COGEMA). These plants, which are similar in design and operating principles, represent more than 90% of the world market in reprocessing services. They recover high-purity plutonium and uranium for recycle and stabilize radioactive waste in forms suitable for disposal. The fundamental process steps are storage of the spent fuel, size reduction, dissolution, chemical separations, product purification, and waste treatment. The process steps are typical of modern practices, although there are usually several practical alternatives for each step.

FIGURE 1 Boiling water reactor fuel assembly. Some fuel elements have been removed to show more structural detail. Photograph courtesy of Framatome.

3.1 Head End Processes

Storage is important to reprocessing in that it allows the spent fuel to cool sufficiently to meet the input decay heat specification. Spent fuel is typically stored in canisters that are about 6 m in length, each weighing about 100 tonnes. The canisters are placed within a storage pool in which the water quality is carefully monitored to prevent corrosion. This storage capacity, which typically provides adequate buffer for at least several months of throughput, represents a significant capital investment. With a length of 150 m, the THORP storage facility (Fig. 2) holds enough fuel for 4 years of reprocessing. When removed from the canisters, individual fuel assemblies are subjected to nondestructive examination to

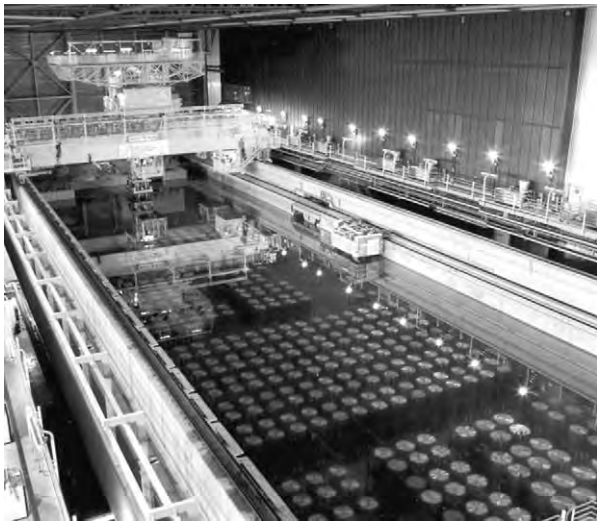


FIGURE 2 The THORP spent fuel storage pool in Sellafield, United Kingdom. Photograph courtesy of British Nuclear Fuels Plc.

verify that key nuclear characteristics match documentation that has been provided by the owner.

The first major step is mechanical disassembly to reduce the fuel to manageable dimensions and to expose the uranium inside of the metal cladding. In modern plants, the entire fuel assembly is chopped into small pieces as it moves through a shearing machine. A typical assembly contains an array of 17×17 fuel elements. Each stroke of the shearing blade cuts a few centimeters from the entire array. The end pieces, which contain no uranium, are collected separately. Shearing fuel is a dusty process in which the equipment is subjected to high stresses. Some gaseous fission products, such as krypton and xenon, are released during this operation.

The chopped fuel is transferred to the dissolver, a large tank containing hot nitric acid. Dust from the shearing operation is also delivered to the dissolver through the ventilation system. Uranium and most of the fission products dissolve readily, whereas dissolution of the plutonium proceeds somewhat more slowly. Plutonium-rich fuel grains do not always fully dissolve. For fuel irradiated to greater than 30,000 MWd/tonne, a portion of the metal fission products (e.g., ruthenium, rhodium, technetium, palladium, and molybdenum) form an insoluble alloy. The residual fission product gases and volatile fission products (e.g., tritium and iodine) are released into the off-gas system. After removal from the dissolver, the leached cladding hulls are washed with nitric acid and water, checked for residual fuel, and stored prior to processing into disposable waste. Solid particles of a few micrometers in diameter that are suspended in the dissolver liquid are removed with a centrifuge. The particles are subjected to an additional leaching step, contributing to the approximately 99.5% recovery of the nuclear material from the fuel.

The dissolver product liquor is conditioned prior to further processing. It is sparged with air at a higher temperature to remove residual iodine, then is cooled and sparged with nitrogen dioxide to adjust the plutonium from its hexavalent state to the tetravalent state, which is needed for extraction. The acidity is adjusted by the addition of HNO_3. With the final clarification of the liquor, a materials accountancy sample is taken to establish the quantity of fissile material entering the process.

3.2 Chemical Separations

The purpose of chemical separations is to partition the elements into three different product streams: a pure plutonium stream, a pure uranium stream, and a waste stream that contains the fission products and minor actinides. PUREX, with its many possible variants, is the separations process that has been universally selected for industrial reprocessing. It is a solvent extraction technique that relies on the selective transfer of components between two immiscible liquids—an aqueous phase and an organic phase. The organic phase contains tributyl phosphate (TBP) diluted to $\leq 30\%$ by volume with a hydrocarbon such as kerosene. The aqueous phase is a nitric acid solution. The two liquids are agitated into intimate contact to increase the surface area across which the species exchange can take place. They separate naturally due the difference in density, like oil floating to the top of water. Uranium and plutonium are first extracted together from the highly radioactive dissolver liquor, which simplifies the downstream separations steps. After partitioning, the separate uranium and plutonium streams are purified to the required level.

Three types of equipment are used to bring the organic and the liquid phases into contact: mixer-settlers, pulsed columns, and centrifugal contactors. Because the organic solvent is damaged by radiation and high temperature, the holdup time needs to be minimized. Multistage countercurrent pulsed columns or centrifugal contactors are preferred to mixer-settlers for high-activity separations. Mixer-settlers are used in scrubbing and stripping stages, during which the radioactivity is less intense. Pulsed columns (Fig. 3) can be several meters in height, containing dozens of mixing stages characterized by perforated disks or an alternating series of disks and rings within a tall pipe. Each pressure pulse pushes the organic phase up the column, breaking it into droplets as it passes through the perforations in the disks. These droplets contact the continuous aqueous phase as it is fed by gravity down the column. This heavier phase is removed from the bottom of the column, while the organic solvent containing the extracted product is removed from the top.

3.2.1 Initial Decontamination

Plutonium and uranium are extracted from the aqueous solution by forming nitrate complexes with TBP, leaving most of the fission products behind:

$$UO_2^{2+} + 2N_3O^- + 2TBP \rightarrow UO_2\,(NO_3)_2\,(TBP)_2$$

$$Pu^{4+} + 4NO_3^- + 2TBP \rightarrow Pu\,(NO_3)_4\,(TBP)_2.$$

Some fission products are carried over in the solvent along with the uranium and plutonium. These

FIGURE 3 Pulsed column. The boxes around the column contain sensors that measure the progress of the extraction. Photograph courtesy of COGEMA.

contaminants are eliminated in a scrubbing operation in which the solvent is washed with aqueous nitric acid. After all traces of solvent are removed, the aqueous raffinate is held in a tank pending further treatment as high-level radioactive waste.

3.2.2 Separation of Plutonium from Uranium

If the decontamination in the first extraction cycle is sufficient, uranium/plutonium partitioning can done in either a mixer-settler or a pulsed column. The uranium and plutonium pass into the contactor in the solvent stream, where uranous nitrate stabilized by hydrazine is introduced as a reducing agent for the plutonium. Plutonium in the Pu^{3+} state passes into a fresh acidic aqueous stream. The uranium, which remains in the organic solvent, goes into a mixer-settler (Fig. 4) for additional decontamination. Uranium is then removed from the solvent by dilute nitric acid. The uranium and plutonium undergo separate purification processes.

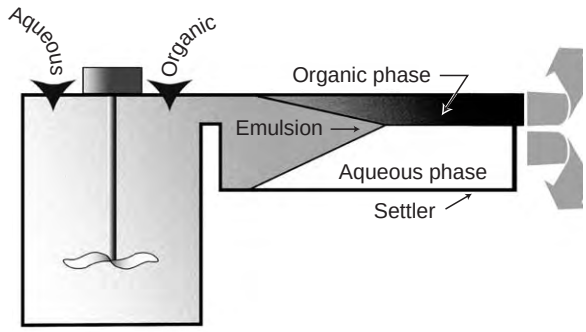


FIGURE 4 Mixer-settler concept.

3.2.3 Uranium Purification

The uranium stream passes into additional mixer-settlers, where the uranium is first extracted into the solvent and then transferred back to the aqueous phase. Through careful control of the chemistry, this series of steps removes most of the residual contaminants, principally neptunium, plutonium, technetium, and ruthenium. The uranium solution is stored in a buffer tank prior to finishing. An evaporator concentrates the uranyl nitrate after removal of any residual solvent. The concentrated nitrate solution is put through a thermal denitration process, which releases nitric acid and forms uranium trioxide. The UO_3 is packaged in stainless-steel drums for storage.

3.2.4 Plutonium Purification

The plutonium-bearing stream is washed with kerosene to remove residual TBP, then evaporated to concentrate the solution. Using pulsed columns, the plutonium is passed between the aqueous and organic phases to remove contaminants. The resulting plutonium nitrate is taken through a final kerosene wash prior to temporary storage. During the finishing process, plutonium nitrate is converted to plutonium dioxide. The properties of the PuO_2 product are adjusted to yield a powder that meets the input specifications for fuel production. The powder is packaged in a stainless-steel canister that is encased in two outer cans and transferred to a highly secure storage vault.

3.3 Waste

Radioactive waste—in gaseous, aqueous, organic, and solid forms—is generated throughout the reprocessing flow. The early experience with reprocessing for nuclear weapons production in the United States and the former Soviet Union left a legacy of inadequately managed waste and unfortunate environmental releases. However, in modern commercial plants, waste

stream management is a priority, leaving relatively small volumes for treatment and disposal.

3.3.1 Gaseous Effluents

Radioactive gases are evolved from various operations in the plant. Shearing the fuel assemblies releases some of the fission gases; most of the remaining fraction is released to the off-gas system when the fuel is dissolved. In order to avoid downstream chemistry problems, temperature and acidity are adjusted to ensure that at least 98% of the iodine is volatilized from the dissolver liquor. Some volatile ruthenium can also be released from high-temperature processes. A high-activity metal fission product that will condense on cool surfaces, ruthenium released into the ventilation system proved to be a problem in the first reprocessing plant. Ruthenium has since been controlled through process chemistry.

The noble gases krypton and xenon are released to the environment. They contribute insignificantly to the radiation dose of the workers or the public. Because noble gases are not reactive, they are difficult to trap and sequester safely. The perceived risk from a large-scale accidental release of stored krypton has outweighed that of a low-intensity continuous release. However, krypton capture and storage has been demonstrated at the small Tokai Reprocessing Plant in Japan.

Caustic scrubbers are typically used to remove the other radioactive gases and noxious gases, such as nitrogen oxides. These scrubbers are packed columns in which the gases come into contact with a countercurrent of sodium hydroxide solution. The gases are usually filtered to remove dust and small radioactive particles prior to the caustic scrub.

3.3.2 Solid Wastes

The solid wastes include primarily the cladding hulls and end pieces of the fuel as well as some plutonium-contaminated materials such as filters or process residues. The solids are compacted and stored in canisters, or they may be immobilized in concrete prior to storage. In Europe, these solids are categorized as intermediate-level wastes that require deep geologic disposal.

3.3.3 Liquid Wastes

The primary liquid waste is the first-cycle raffinate containing most of the fission products. In the United States, high-level waste is legally defined as this raffinate or any solids derived from it. The liquid is concentrated by evaporation and temporarily stored. Because the concentrated liquid is highly self-heating from the decay of ^{137}Cs, ^{90}Sr, and other fission products, the storage tanks must be cooled.

The high-level liquid waste is pumped out of storage into a treatment plant, where it is calcined and mixed with glass frit. The mixture is melted and drained into a stainless-steel canister, where it solidifies into a vitrified waste form. Lids are welded on the canisters before they are placed in cooled storage, where they remain for decades prior to final disposition. The modern French plants produce less than $0.5\,m^3$ of high-activity waste for every tonne of uranium processed, or less than 25% of the estimated volume required to store an equivalent amount of spent fuel. To the extent possible, the liquid chemical streams are cleaned and recycled in modern reprocessing plants. Diluent and solvent are both recovered during organic waste treatment. The principal wastes from this process are various salt residues from the separations steps and the decomposition products of TBP.

3.4 Environmental Effects

The modern British and French plants are able to contain their environmental effluents within strict regulatory limits. The plants are located near the coast; the principal environmental effluent is lightly contaminated water that is discharged into the ocean. Trace amounts of radionuclides are contained in the discharged water. Most of the tritium is released to the sea, because it is chemically inseparable from ordinary water. Because of its volatility, iodine is unsuitable for high-temperature waste processing. It has traditionally been diluted with the natural iodine in the seawater, as this is the safest known practice, although the search for a suitable alternative has been underway for decades.

The UP2 and UP3 plants (Fig. 5) are located at La Hague, France, where the average dose rate to the public from the environmental discharges is less than 0.01 millisievert (mSv)/year, compared to a dose from background radiation of 2.4 mSv/year. In the United Kingdom, the modern THORP plant is collocated at the Sellafield Site with an older reprocessing plant, and the peak dose to critical seafood consumers reached almost 2 mSv at that location in 1976. Today, the peak dose rate has dropped to approximately 0.1 mSv/year, a level that is due primarily to a legacy of previous seabed contamination. At both the French and the British sites, exceptionally low exposure levels were first achieved in the late 1980s.

Tight control of the radioactive inventory at plants was not the main priority in the early years of reprocessing. In the United States, reprocessing of

FIGURE 5 COGEMA reprocessing facilities at La Hague, France. Photograph courtesy of COGEMA.

short-cooled fuel resulted in release of radioactive ^{131}I between 1944 and 1947. Subsequently, the fuel was cooled for a longer time, which allowed decay of most of this isotope with its 8-day half-life. Filters and scrubbers were also added to the exhaust systems of plants to trap iodine and other radionuclides. More than 350,000 m^3 of liquid high-level waste was generated and stored in steel tanks, creating an expensive legacy that is now being addressed.

The largest environmental problems were experienced in the former Soviet Union as a result of plutonium production during the Cold War. These operations at the Mayak, Tomsk-7, and Krasnoyarsk-26 sites resulted in record releases of radioactive materials into the environment. The towns at these Siberian sites are now known as Ozersk, Seversk, and Zheleznogorsk, respectively. At Mayak, liquid wastes were discharged into Lake Karachai, the Techa River, and a series of small reservoirs on the river. At Tomsk and Krasnoyarsk, liquid wastes were discharged primarily to deep underground geologic formations, but also to surface waters.

Though the Cold War legacies from weapons plutonium production in the United States and the former Soviet Union bear scant resemblance to modern commercial reprocessing practices, their influence on public perception continues. COGEMA and BNFL prominently display up-to-date environmental information on their Internet Web sites to help allay public concern about radioactive releases.

4. REPROCESSING FACILITIES

4.1 United States

Large-scale reprocessing had its start in the Manhattan Project. During World War II, three plutonium production reactors and two reprocessing plants (the T plant and the B plant) were built at Hanford, Washington. In operation by December 1944, the T plant was the world's first reprocessing facility. Plutonium was separated from metallic uranium fuel that had been lightly irradiated in the B reactor. Though there are few similarities to today's commercial plants, some features of the early building designs by the DuPont engineers have been retained in modern facilities. The T plant was a canyonlike 245 m long, 31 m high, and only 26 m wide. Constructed of steel-reinforced concrete, the process cells were heavily shielded. Access to the process equipment was by crane through the roof of the concrete cells. The T plant used a bismuth phosphate process, selected and designed by DuPont from several separations flow sheets developed by the Metallurgical Laboratory at the University of Chicago. The B plant, the second facility based on this precipitation process, operated at Hanford before the end of the war.

Two other large reprocessing plants were also built at Hanford. Completed in 1951, the REDOX (so named because of the reduction and oxidation process involved) plant was the first reprocessing plant based on countercurrent, continuous-flow separation of plutonium and uranium. The process was developed by Argonne National Laboratory and implemented by the General Electric Company at Hanford. The PUREX (plutonium/uranium extraction) plant went into operation in 1956, applying the technology that has since been commercialized in several countries. The PUREX flow sheet was developed by the Knolls Atomic Power Laboratory and implemented at Hanford by General Electric.

The first large PUREX reprocessing plant built by the U.S. government, called the F-Canyon, was designed and constructed in 1954 by DuPont at the Savannah River site in South Carolina. PUREX was also used in specialized plants for reprocessing highly enriched uranium fuels at Savannah River (H-Canyon) and in Idaho (Idaho Chemical Processing Plant). None of these plants was designed or operated for processing commercial fuel. Notwithstanding environmental legacies, the government-owned, contractor-operated plants had an impressive production record, particularly considering the first-of-a-kind engineering that went into each plant. The REDOX plant processed more than 19,000 tonnes of spent fuel during its 16 years of operation. The capacity of the PUREX and F-Canyon plants was considerably higher, in the range of 2000–3000 tonnes of initial uranium per year of operation.

By contrast, attempts to commercialize reprocessing in the United States were unsuccessful. Of the three plants that were constructed, only one ever operated—the Nuclear Fuel Services plant at West Valley, New York. From 1966 through 1972, the West Valley plant processed fuel from several utility-owned reactors as well as some fuel from the N reactor in Hanford, Washington. With a nominal capacity of 300 tonnes/year, the plant's PUREX process worked well. However, when the plant was shut down due to changing regulatory requirements, modifications were assessed to be too expensive and the plant never restarted.

A novel 300-tonne/year plant was designed and built by General Electric in Morris, Illinois in the early 1970s. The Midwest Fuel Recovery Plant was designed to use a combined solvent extraction and fluoride volatility process rather than the complete PUREX flow sheet. However, nonradioactive testing revealed that the compact design had not incorporated sufficient buffer storage between processes. Realizing that the plant would not be capable of reliable operation, General Electric simply used the fuel storage facility, and never started the plant. Between 1970 and 1975, Allied General Services built a large 1500-tonne/year reprocessing plant in Barnwell, South Carolina. The plant underwent startup testing with fresh uranium in 1976 while the rules on plutonium conversion and high-level waste vitrification were being resolved by the Nuclear Regulatory Commission. However, in 1977, the Carter administration canceled the licensing proceedings and the plant had to be mothballed.

4.2 Western Europe

Several European reprocessing plants have been built and successfully operated, including the Eurochemic 300-tonne/year plant at Mol, Belgium, a 35-tonne/year German pilot plant, and several modern British and French commercial reprocessing plants (described in the following sections).

4.2.1 France

More than any other nation, France has successfully implemented its commercial reprocessing ambitions. As a part of the international nuclear services conglomerate AREVA, COGEMA operates the two 800-tonne/year UP2 and UP3 plants at La Hague on the Brittany coast. In addition to the French national utility Electricité de France (EDF), COGEMA's reprocessing customers include Japan, Belgium, Germany, Spain, and Switzerland. Prior to 1998, COGEMA also operated a 400-tonne/year plant in Marcoule to reprocess fuel from the French gas-cooled reactors.

4.2.2 United Kingdom

The highest capacity reprocessing plant in the United Kingdom is the 1500-tonne/year Magnox plant, which has processed fuel from the nation's gas-cooled Magnox reactors since 1964. However, these relatively old reactors are scheduled to be shut down by 2010. Once the magnesium-clad fuel has been processed, the Magnox plant will also be closed.

Completed in 1994 near Sellafield on the Irish Sea, THORP is one of the newest reprocessing plants in the world. The plant is 500 m long, 48 m high, and 120 m wide, with a chimney that stands 125 m above the landscape. The expected throughput was approximately 1200 tonnes/year, but operational difficulties, primarily with the waste vitrification plant, have not allowed it to reach that potential. The nominal capacity is now considered to be more in the range of 850 tonnes/year, though individual pieces of equipment are scaled for higher throughput. Because the capacity of the plant far exceeds the domestic need for its reprocessing services, THORP has relied on foreign contracts, particularly with Japan and Germany, for most of its business.

4.3 Asia

4.3.1 Japan

Lacking indigenous uranium resources, Japan has a long-standing policy of supporting reprocessing. Japan has both imported foreign technology and conducted a vigorous national research and development program. The Tokai Reprocessing Plant has operated with an approximately 90-tonne/year capacity since 1977. In 1993, Japan Nuclear Fuel Limited (JNFL) began constructing a large reprocessing plant near Rokkasho on Japan's northern coast (Fig. 6). The 800-tonne/year plant is scheduled to begin full operation in 2005.

The Rokkasho plant incorporates much of the French commercial technology. One notable exception is that plutonium will be coextracted with 50% uranium as an additional nuclear safeguard measure. The site includes an interim high-level waste storage facility, which in 1995 began to receive canisters of waste from Japanese fuel that had been reprocessed in Europe. The site also includes a low-level waste disposal facility, a large spent fuel storage capacity, and a uranium enrichment plant, all currently operating. An icon of modern technology, the

FIGURE 6 Rokkasho reprocessing plant on Japan's northern coast. Photograph courtesy of Japan Nuclear Fuel Limited.

Rokkasho plant provides a benchmark (Table I) for the scale of commercial reprocessing facilities.

TABLE I

Dimensions of Main Buildings at the Rokkasho Reprocessing Plant[a]

Building	Approximate floor area (m^2)	Height (stories)	
		Aboveground	Belowground
Spent fuel receiving and storage	9400	3	3
Head end	6000	5	4
Separations	5700	4	3
Purification	6500	6	3
Uranium denitration	1500	5	1
Uranium–plutonium codenitration	2700	2	2
Uranium oxide storage	2700	2	2
Uranium–plutonium mixed oxide storage	2700	1	4
High–activity liquid waste vitrification	5100	2	4
Vitrified package storage	5700	1	2
Low–activity liquid waste treatment	2600	3	2
Low–activity waste treatment	9500	4	2
Control	2900	3	2
Analytical laboratory	4900	3	3

[a] Data courtesy of Japan Nuclear Fuel Limited.

4.3.2 Russia

In Russia, as in the United States, most reprocessing experience has been with plutonium production reactor fuel. At the Mayak, Tomsk-7, and Krasnoyarsk-26 sites, more than 300,000 tonnes of uranium fuel were reprocessed between 1948 and 1996. This is an estimable production record, considering that it would take 100 years of modern commercial reprocessing to match it. The Russian production reprocessing used a somewhat unique aqueous technology based on the precipitation of sodium uranyl acetate and sodium plutonyl acetate from nitric acid solutions containing the dissolved fuel. Plutonium was separated from the uranium by reducing plutonium to the Pu^{4+} or Pu^{3+} state, which remained in solution. The acetate process created a high volume of sodium nitrate and sodium acetate waste.

A reprocessing facility for power reactor fuel was completed at Mayak in 1976. Called RT-1, the plant uses a more conventional PUREX process. It was built with a storage capacity of 400 tonnes of spent fuel. Russia is building a much larger RT-2 plant at Zheleznogorsk. To be constructed in stages, the initial operation is intended to have a capacity of 1500 tonnes/year. However, funding has been insufficient for timely completion of the plant, which is less than half finished. It has a storage capacity of 6000 tonnes, with a planned increase to 9000 tonnes.

Russia has conducted extensive research on nuclear waste management and reprocessing at many of its universities and laboratories. Since the mid-1990s, Russia has been developing plans to open an international spent fuel reprocessing center. Because the reprocessing capacity is not yet adequate, the center would initially provide spent fuel storage.

4.3.3 China and India

China has operated a reprocessing pilot plant since 1970 and has plans for constructing an 800-tonne/year facility by 2010. India is operating two small plants, each capable of about 60 tonnes/year throughput. India also has planned to install some 1000 tonnes/year of reprocessing capacity to manage the spent nuclear fuel from its growing nuclear energy infrastructure.

4.4 National Policies

The rationale for nuclear fuel reprocessing, or alternatively for deferring the practice, has been the subject of intense international discussion since the late 1970s. The crux of the debate lies in the military origin of the technology and the increasing global concern about

weapons of mass destruction. Although many refinements have been made, the technology that has been implemented commercially is fundamentally the same technology that was used to develop a military nuclear capability in the United States, the former Soviet Union, France, and the United Kingdom, and it can produce very pure plutonium.

Weapons-grade plutonium has been specified to have at least 92% of the isotope ^{239}Pu. Because of spontaneous neutron emission, radiation, and heat generation, the other isotopes of plutonium cause significant difficulties for the design, fabrication, storage, and transportation of nuclear weapons. Typical commercial spent fuel contains plutonium that has less than 80% of the ^{239}Pu isotope, so reactor-grade plutonium was initially not considered weapons usable. However, in the 1970s, weapons experts informed U.S. policymakers that it was in fact possible to make a nuclear explosive from reactor-grade plutonium, if not a military weapon. In light of this information and an unexpected nuclear test explosion by India in 1974, the U.S. position on reprocessing was reexamined during the Gerald Ford administration. In 1977, President Jimmy Carter canceled the licensing of the Barnwell plant, stranding half a billion dollars of private investment. When the reprocessing ban was quietly lifted in 1981 during the Ronald Reagan administration, there was no longer viable commercial interest in reprocessing within the United States. In 1993, President Bill Clinton reinstated the national policy against reprocessing, but in 2001 the Bush administration began to encourage research into new reprocessing technologies that would not produce weapons-suitable material.

Unfavorable economics has further fueled the debate over recycling nuclear materials. In an era of cheap uranium, fuel produced from recycled plutonium is more expensive than is standard fuel produced from enriched uranium. Though estimates of the cost premium for recycled fuel span a wide range, its actual impact on the cost of electricity production is minimal. Unlike fossil-fueled power plants, nuclear fuel accounts for no more than 20% of the total generating cost.

The original assumption in the 1960s was that the recovered plutonium would be used to fuel sodium-cooled reactors, which would improve uranium resource utilization by two orders of magnitude. The actual practice has been to recycle plutonium into existing water-cooled reactors. Primarily for safety reasons, not all of the existing reactor fleet can accept plutonium fuel. Most of those reactors that do accept the mixed plutonium/uranium oxide fuel (MOX) can use it in only one-third of the core loading. Further, the capacity of the MOX fabrication facilities has not caught up with the reprocessing capacity. As a result, there is a growing accumulation of excess civilian plutonium, which stood at more than 200 tonnes at the beginning of the 21st century.

The reprocessing debate has been predominantly a U.S. phenomenon. Approximately one-half of the spent nuclear fuel discharged annually around the world is slated for reprocessing, and the other half is slated for direct disposal. A practical argument in favor of reprocessing is that though uranium may be plentiful, land suitable for the disposal of nuclear waste is not. No nation has yet opened a deep geologic repository for high-level waste. The United States is moving toward that goal, having initiated a licensing process for Yucca Mountain, Nevada, after 20 years of study and legal challenges. If the schedule is maintained, the first fuel assemblies would be shipped in 2010.

France, Japan, and Russia have firm pro-reprocessing policies. With limited land area, more than 50 nuclear power plants, and no significant indigenous energy resources, Japan's reprocessing policy is based on a desire to have some measure of energy independence. Requiring reprocessing as a part of its national energy strategy, France has a complex spent fuel management plan that calls for leveling the total accumulation of plutonium in the entire fuel cycle before a national geologic repository is developed. Russia has expressed keen interest in providing full nuclear fuel cycle services to international clients in addition to domestic reprocessing.

4.5 International Safeguards

Commercial reprocessing plants are subject to nuclear materials safeguards verification by the International Atomic Energy Agency (IAEA) under the Treaty on the Nonproliferation of Nuclear Weapons. Entered into force in 1970, the treaty now has 188 member states. Only Israel, Pakistan, and India have not signed, although in January 2003, North Korea announced its intention to withdraw. The technical objectives of safeguards agreements established under the treaty are the timely detection of diversion of significant quantities of nuclear material from peaceful uses to manufacture nuclear weapons, and to deter such diversion by early detection. The objectives are based on the principle that a certain quantity of nuclear material is needed to manufacture a nuclear explosive device and that a

certain amount of time is needed to convert the material into a weapons-usable form.

The IAEA uses a variety of physical and chemical measurements and other sources of information to determine whether a diversion has occurred. The agency has implemented safeguards in 70 countries at more than 900 facilities, including six reprocessing plants. The IAEA also has safeguards jurisdiction over a rapidly growing inventory of plutonium in spent nuclear fuel, almost 600 tonnes in 2003.

5. REPROCESSING RESEARCH AND DEVELOPMENT

5.1 Cost Reduction

After 50 years of PUREX implementation, reprocessing is still not a mature industry. With changing objectives, evolving input fuel specifications, more stringent environmental regulations, and feedback from operating plants, the opportunities for developing process, equipment, or plant improvements are ample. Established reprocessing corporations are focused on driving down costs. Options include process modifications to make them move compact, faster, safer, less chemical intensive, more reliable, more transparent to safeguards, or easier to maintain. There is considerable incentive to reduce capital cost through simplification and smaller facilities, which could be accomplished by fewer cycles or less buffer storage. Existing plants were designed to process fuel that had cooled for only several months, whereas in practice the actual cooling time has been years, if not a decade or more. This greatly reduces the radiation and heat load requirements on the facility, which opens the door to cost-saving design changes.

As new technologies are developed, they are implemented in operating plants to the extent practicable, rather than saved for the next-generation facility. For example, utilities are routinely discharging fuel that has an average of 48,000 MWd/tonne burn-up, with 60,000 MWd/tonne expected to be achievable within a few years. Since some plants were designed with lower burn-up limits and initial enrichment limits, modifications must be made just to stay in business.

5.2 Alternative Processes

From hundreds of potential chemical separation schemes, PUREX has been adopted for commercial implementation. However, as new types of reactors and fuels are considered or as the objectives for reprocessing evolve, research is being conducted on alternative separation technologies as well as on PUREX and its aqueous siblings. Nonaqueous processes have been studied for decades because of potential advantages that include much less susceptibility to radiation and heat damage, fewer chemical steps, more compact equipment, and larger batch sizes. They have not found their way into commercial application because of disadvantages that include an inability to meet product specifications for current-generation nuclear fuel and required operation at high temperature with corrosive or reactive materials. Most nonaqueous processes can be classified as pyrometallurgical, pyrochemical, or fluoride volatility processes.

Pyrometallurgical processes include volatilization, distillation, fractional crystallization, liquid metal extraction, fused salt extraction, fused salt electrolysis, and partial oxidation. Many fission products are more volatile than uranium. Volatilization typically works by vacuum distillation of crushed, molten, or dissolved fuel to separate most of the volatile elements from the uranium. It is used to remove noble gases and iodine during conventional aqueous processing. Fractional crystallization has been tested for partitioning fission products from metallic fuel using a low-melting-temperature liquid-metal extractant and for removing bulk uranium from the dissolver liquor at very low temperature. Partial oxidation is a process in which volatile fission products are vaporized from molten metal fuel, whereas the rare earths and the more reactive fission product metals are oxidized by a sacrificial crucible liner. Argonne National Laboratory successfully conducted a pilot-scale demonstration of this melt refining process on more than 30,000 fuel pins during the late 1960s, but decontamination was considered to be too incomplete for further development.

Pyrochemical processes use selective oxidation and reduction reactions in fused salts or metals to recover uranium and plutonium. Russia has an active program that has demonstrated pyrochemical processing of uranium oxide fuels. There is also an active Japanese program to develop pyrochemical processes for removing transplutonium actinides from waste streams and for processing uranium nitride fuels.

Fluoride volatility processes take advantage of the unusual property of uranium, neptunium, and plutonium to form volatile hexafluorides. Unlike other nonaqueous processes, fluoride volatility is

more readily capable of thorough decontamination of the product materials. Fluoride volatility processes can be combined with aqueous processes, as was done in the General Electric plant at Morris, Illinois.

5.3 Partitioning and Transmutation

During the 1990s, the concept of much more aggressive nuclear waste management began to take hold. This line of research developed out of concern about leaving a legacy of radiotoxic waste that would persist for hundreds of thousands of years, the uncertainty about the future availability of deep geologic repositories, and proliferation potential of thousands of tonnes of plutonium accumulating in spent nuclear fuel in more than 30 countries. In its ultimate application, partitioning and transmutation would destroy all actinides by fission and would transmute long-lived fission products into stable elements or short-lived radioisotopes. In order to achieve transmutation of the minor actinides and some plutonium isotopes, a fast-spectrum reactor would be required.

5.3.1 Objectives of Partitioning and Transmutation

The principle behind partitioning and transmutation is that there is an optimum means of managing each group of elements in spent nuclear fuel. Within practical limits, the elements are separated into like groups as defined by hazard, transmutation potential, chemistry, heat-generation rate, proliferation risk, half-life, and energy value. Each group is managed to meet appropriate specific objectives: Uranium, which comprises 95% of the mass, can be stored until needed sometime in the future. Plutonium can be used in thermal-spectrum reactor fuel. Plutonium and minor actinides can be used in fast-spectrum reactor fuel. Highly soluble, long-lived fissions products such as ^{129}I and ^{99}Tc can be fabricated into transmutation targets or can be stabilized in custom waste forms. Short-lived waste products can be stabilized in compact waste forms. The high-heat-load fission products cesium and strontium can be separately managed until they are sufficiently cool after several decades.

Temperature is the limiting design parameter for the proposed Yucca Mountain repository. Intended to store unprocessed spent fuel, the peak mountain temperature is predicted to occur hundreds of years after closure. The transuranic elements, in particular neptunium and americium, are the principal sources of long-term heat generation (Fig. 7). A second repository, operated in a partitioning-transmutation scenario with active ventilation, could safely store the waste from five to a hundred times as many reactor-years of power production. The key is recycling the actinides, which would remove most of the volume, long-term heat generation, and long-term radiological hazard. Without actinides, the radiotoxicity of the stored nuclear waste also decays relatively quickly, becoming less toxic than uranium ore within a few hundred years.

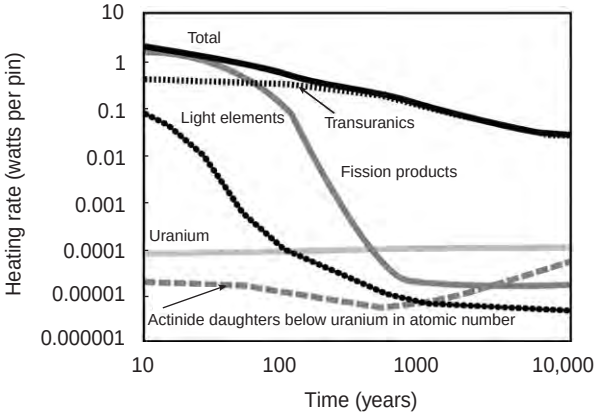


FIGURE 7 Components of the heat generation rate for spent nuclear fuel with an average burn-up of 33,000 MWd/tonne. Courtesy of Argonne National Laboratory.

5.3.2 Separations Research Programs

Japan, Russia, the United Kingdom, France, and the United States all sponsor research on advanced separations methods. Japan is pursuing both advanced aqueous and pyro processes. Russia has led pyrochemical process development. The United Kingdom is focused on improving conventional reprocessing while supporting a modest effort in pyroprocess development. France has a comprehensive program that supports a broad spectrum of research, ranging from molecular modeling to hardware improvements. The U.S. Department of Energy has consolidated all separations research into a single advanced fuel cycle program.

The U.S. aqueous process development work is aimed at determining the practicality of multiple separations. At least five different partitioning steps are being considered. After conventional nitric acid dissolution, uranium and technetium would be separately extracted, thereby removing most of the material from the main process stream. Cesium and strontium would be extracted in a second process. Neptunium and plutonium would then be coextracted to avoid clean plutonium and its

attendant proliferation risk. Transuranic extraction (TRUEX), an established process, would separate the bulk fission products from the remaining actinides. A fifth process would separate the americium and curium from the acid stream, which after that point would contain principally rare earth fission products.

This approach to multiple separations presents many technical challenges, including managing several different solvent streams and minimizing radiation damage to the organic phases while retaining fission products and high-activity α emitters through several cycles. Radiation damage to the solvent is controlled by using banks of centrifugal contactors, which minimizes the irradiation time of the solvent phases.

A second line of separations development is aimed at using a fused salt pyroprocess to partition the fuel into two product streams and two waste streams. The spent fuel would be reduced to metal by direct electrolytic reduction in a vessel containing molten LiCl. The metal product would be electrorefined (Fig. 8) to recover uranium in second vessel containing a LiCl/KCl electrolyte, in much the same way that aluminum is produced. The salt and uranium phases would be separated by vacuum distillation, followed by melt consolidation of the uranium. The cladding hulls and metals such as technetium and rhodium that are not oxidized in the salt would be withdrawn, separated from adhering salt, and processed into a metallic waste form. Argonne National Laboratory has successfully demonstrated electrorefining by processing more than 2.5 tonnes of metal fuel since 1996.

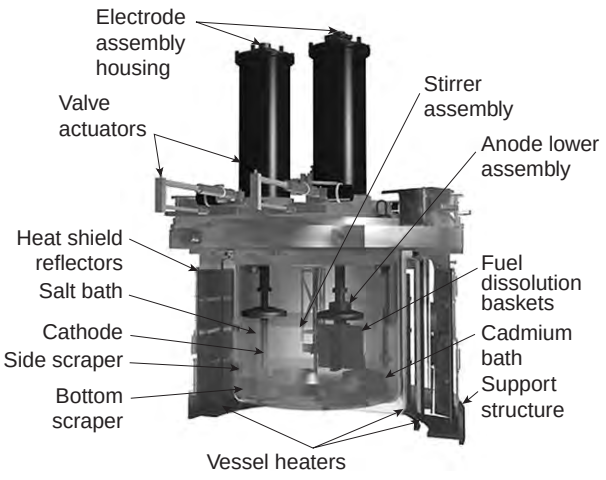


FIGURE 8 Drawing of an electrorefiner used in pyroprocess demonstrations. Courtesy of Argonne National Laboratory.

Salt containing uranium and transuranic chlorides would be treated by electrolysis or an electrometallurgical process to extract a second product containing uranium, plutonium, neptunium, americium, curium, and some rare earth fission products. This alloy could be used directly to manufacture fast-spectrum reactor fuel or it could be sent to a small purification line to be adequately decontaminated for thermal-spectrum reactor fuel. After the salt was sufficiently laden with fission products and stripped of actinides, it would be treated in a zeolite column to remove the fission product chlorides. The zeolite would be processed into a glass-bonded ceramic waste form.

5.3.3 Future Research and Development

Reprocessing research and development are limited only by the imagination of the chemist, available funding, and national policy restrictions. Ultimately, fuels, reactors, separation processes, waste forms, and geologic repositories will benefit from being designed as a secure, compatible system, rather than following independent development trajectories as they did in the 20th century.

SEE ALSO THE FOLLOWING ARTICLES

Nuclear Engineering • Nuclear Fission Reactors: Boiling Water and Pressurized Water Reactors • Nuclear Fuel: Design and Fabrication • Nuclear Power Economics • Nuclear Power, History of • Nuclear Power Plants, Decommissioning of • Nuclear Power: Risk Analysis • Nuclear Proliferation and Diversion • Nuclear Waste • Occupational Health Risks in Nuclear Power • Public Reaction to Nuclear Power Siting and Disposal

Further Reading

Benedict, M., Pigford, T. H., and Levi, H. W. (1981). "Nuclear Chemical Engineering." McGraw-Hill, Boston.

Bradley, D. J. (1997). "Behind the Iron Curtain: Radioactive Waste Management in the Former Soviet Union." Batelle Press, Columbus, Ohio.

Cochran, R. G., and Tsoulfanidis, N. (1999). "The Nuclear Fuel Cycle: Analysis and Management." American Nuclear Society, La Grange Park, Illinois.

Feiveson, H. A., von Hippel, F., and Williams, R. H. (1979). Fission power: An evolutionary strategy. *Science* **203,** 330–337.

Garwin, R. L., and Charpak, G. (2001). "Megawatts and Megatons." Knopf, New York.
Hoffman, D. C. (ed.). (2002). "Advances in Plutonium Chemistry 1967–2000." American Nuclear Society, La Grange Park, Illinois, and University Research Alliance, Amarillo, Texas.
Office of Environmental Management. (1997). "Linking Legacies: Connecting the Cold War Nuclear Weapons Production Processes to Their Environmental Consequences." DOE/EM-0319. U.S. Department of Energy, Washington, D.C.
Rasmussen, N. C. (Chair), Committee on Separations Technology and Transmutation Systems. (1996). "Nuclear Wastes: Technologies for Separations and Transmutation." National Research Council. National Academy Press, Washington, D.C.
Seaborg, E., and Seaborg, G. T. (2001). "Adventures in the Atomic Age: From Watts to Washington." Ferrar, Straus & Girox, New York.
Wilson, P. D. (ed.). (1996). "The Nuclear Fuel Cycle: From Ore to Waste." Oxford Univ Press, Oxford, United Kingdom.

Nuclear Fusion Reactors

PHILIPPE MAGAUD and G. MARBACH
Euratom–Commissariat à l'Energie Atomique
Cadarache, France

IAN COOK
Euratom–U.K. Atomic Energy Authority
Abingdon, United Kingdom

1. Introduction
2. The Fusion Reactor
3. Conclusion

Glossary

International Thermonuclear Experimental Reactor (ITER) Project whose main goal is to demonstrate the scientific and technological feasibility of fusion energy for peaceful purposes; ITER will study burning plasmas (i.e., plasmas where heating is provided mainly by the alpha particles created from fusion reactions) and will also be the first machine incorporating most of the technologies crucial for the preparation of a future fusion reactor (e.g., plasma facing components, tritium management, robotics, tests of breeding blankets).

Joint European Torus (JET) Flagship of the European Community Fusion Program; JET is the largest and most powerful fusion experiment in the world.

keV Abbreviation for kiloelectronvolt; in nuclear physics, energy is often expressed in electronvolts and multiples thereof; as a result of popular misuse of language, temperature is sometimes also indicated in kiloelectronvolts.

At the core of the sun and stars, light nuclei combine—or fuse—to create heavier nuclei. This process releases a significant quantity of energy and is the source of the heat and light that we receive. Our sun has been shining for more than 4 billion years, and according to astrophysics, it will shine for this long again before entering the phase that will lead to its extinction. Harnessing this type of reaction on Earth for the purpose of generating energy would open the way to nearly unlimited resources. This is the aim of fusion research undertaken by the leading industrial nations. After a reminder of the main principles, this article focuses on the final objective of controlled fusion research: the energy-generating fusion reactor. Many concepts are already available, whereas some would require so much progress in terms of physics and technology that they could only be implemented well into the future. An exhaustive description is not possible. Therefore, this article focuses only on the most developed: the magnetic fusion-based concepts. Worldwide studies have been done, or are ongoing, with the main goals to prove their credibility, reliability, and economic viability and that the advantages of this type of reactor in terms of safety and the environment are respected.

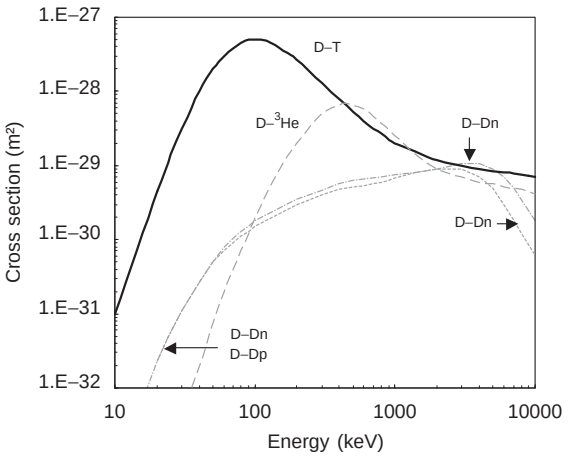
1. INTRODUCTION

1.1 Fusion Reactions

To obtain a fusion reaction, two nuclei that have a natural tendency to repel each other, both being positively charged, must be brought close enough together. Therefore, a certain amount of energy is necessary to overcome this natural barrier and reach the zone close to the nucleus where nuclear forces are capable of overcoming the barrier. The probability of crossing this barrier may be quantified by the "cross section" of the reaction. The variation of the effective cross section of several fusion reactions as a function of interaction energy is provided for light elements in Fig. 1. This gives rise to several points:

- The cross sections of fusion reactions are on the order of 1 barn ($1\ \text{barn} = 10^{-28}\ \text{m}^2$), compared with 600 barns for the fission reaction of ^{235}U obtained by thermal or slow neutron bombardment. Therefore, fusion reactions feature extremely small cross sections, even compared with other competitive processes

 © 2004 Elsevier Inc. All rights reserved.



D + T → ^{4}He (3.56 MeV) + *n* (14.03 MeV) [17.59 MeV]

D + D
- 50% → ^{3}He (0.82 MeV) + *n* (2.45 MeV) [3.27 MeV]
- 50% → T (1.01 MeV) + *p* (3.02 MeV) [4.03 MeV]

D + ^{3}He → ^{4}He (3.71 MeV) + *p* (14.64 MeV) [18.35 MeV]

Tritium production:

n + ^{6}Li → T + ^{4}He [4,78 MeV]

n + ^{7}Li → T + ^{4}He + *n* [–2,47 MeV]

FIGURE 1 Cross section of main fusion reactions.

such as ionization and coulomb scattering (collision of charged particles). This constitutes one of the inherent difficulties of fusion.

• Fusion demands high temperatures, typically more than 100 million degrees (10 keV). At such temperatures, electrons are separated from nuclei; hence, a plasma, the fourth state of matter, is formed. Plasma exists within the universe in extremely diverse forms and with extremely variable temperature and density characteristics. Plasma is the most common state of matter in the universe and has already found numerous applications in everyday life (e.g., neon tube, plasma torch).

• The most accessible fusion reaction is that involving deuterium and tritium, the two isotopes of hydrogen. Research into controlled fusion focuses on this reaction. Although deuterium is abundant, tritium exists in only trace amounts. It needs to be produced by means of nuclear reactions with lithium (Fig. 1).

1.2 Conditions Necessary for Fusion Reactions

As explained, fusion requires high temperatures. Other conditions must also be met if fusion is to be used as an energy source.

As a hot gas, plasma is never totally isolated; thus, it is subject to various losses through radiation but also through convection and conduction. In general, a time is defined synthesizing all said losses, referred to as the energy confinement time (τ_E); this is the time the plasma requires to lose its energy content if the energy sources feeding it are cut suddenly. In some ways, τ_E characterizes the quality of plasma insulation.

For the fusion reaction to be energetically viable, the energy generated by fusion reactions must at least compensate for these losses. This condition imposes a limit less than the product of energy confinement time and particle density (n) given by the Lawson criterion:

$$n \cdot \tau_E > g(T) \cdot f(Q),$$

where $g(T)$ accounts for the variation of the reaction rate with temperature T and Q corresponds to the relationship between the fusion power generated and the external power supplied to the plasma. The Q factor is frequently referred to as the energy amplification factor. There are two characteristic values of Q:

1. $Q = 1$ indicates that power produced by the plasma is equal to the power coupled to it from the exterior. This is referred to as the "break-even" point and is approached in the most efficient of the current experimental devices.
2. $Q = \infty$ indicates that external power contributed to the plasma is zero. The plasma is self-sustaining and is said to be in ignition.

For a deuterium and tritium plasma, the function $f(Q)$ is equivalent to approximately 1 for $Q = 1$ and tends rapidly toward 5 for higher values of Q. Under these conditions, and for a temperature of 10 keV, the Lawson criterion may be written

$$n \cdot \tau_E \approx 10^{20} (m^{-3} \cdot s).$$

1.3 Fusion in the Stars, Fusion on Earth

In the sun and the stars, the conditions necessary for fusion regarding temperature, density, and confinement time are maintained by gravity. This solution is impossible to implement on Earth. The fact that the Lawson criterion sets a limit, not for density and confinement time separately but rather for the product of these characteristics, allows two completely different approaches:

1. Bring a small volume of matter to a high temperature under high pressure for an extremely short time period. This is referred to as inertial confinement. The aim is to obtain the greatest possible number of fusion reactions before the

plasma disperses. Confinement time characteristics are on the order of a few picoseconds.

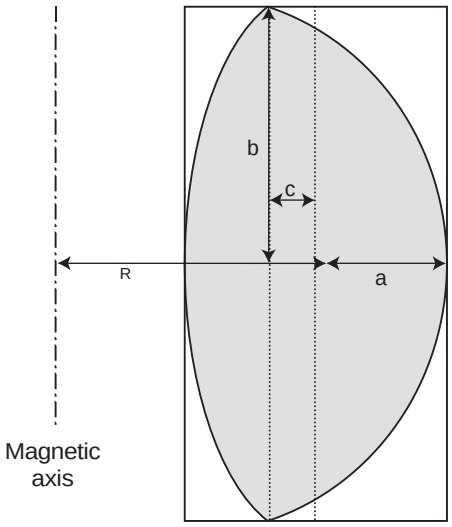
2. Trap and hold plasma at a very high temperature. This plasma is confined in an intangible box created by magnetic fields. This is referred to as magnetic confinement. In this case, the aim is to "allow time" for particles to fuse. Confinement time characteristics are on the order of a few seconds.

Even if inertial confinement energy-oriented concepts exist, research on inertial fusion is supported mainly for simulation of nuclear weapons. Magnetic fusion is being developed in an extremely open international framework due to the fact that its applications are focused solely on the production of energy. What follows is focused on this second approach.

1.4 Magnetic Boxes to Contain Plasma

Plasma is an electrically conducting fluid electrically neutral from the outside in which ions and electrons move practically independently of each other. Submerged in a magnetic field, ions and electrons will follow helical trajectories winding around field lines and will be forced to move along the field. This is the magnetic confinement principle. Initially, straight (or cylindrical) topologies were examined. However, these featured a drawback in that they let the plasma escape at extremities. To avoid this, the cylinder was closed on itself, resulting in a toric configuration, frequently characterized by the ratio between its major radius and its minor radius, referred to as the aspect ratio (Fig. 2). However, in this type of configuration, the curve (and therefore centrifugal force) and lack of homogeneity of the field (the magnetic field is stronger on the inner surface of the torus than on the outer surface) give rise to a drift of charged particles. Ions and electrons tend to separate; some move to the top and the others move to the bottom and end up leaving the magnetic trap. To compensate for this effect, the field lines were modified and made helicoidal. Thus, particles successively cross at the top and then at the bottom of the magnetic configuration. Therefore, the drift effect, which is always in the same direction, is on average compensated. This is achieved by adding another magnetic field ("poloidal" field) perpendicular to the toric field ("toroidal" field). The method implemented to generate these helicoidal field lines gives rise to two types of machines:

1. In a "tokamak," an axial current circulating in the plasma itself creates the poloidal magnetic field.



R: major radius
a: minor radius
A = R/a = aspect ratio
b/a = elongation
c/a = triangularity

FIGURE 2 Cross section of a toric fusion plasma (main geometrical characteristics).

This field becomes the secondary circuit of an enormous transformer. Thus, the tokamak is an impulse device (Fig. 3).

2. In a "stellarator," the magnetic configuration is based entirely on currents circulating in external helicoil coils. This configuration, because of its greater complexity regarding principles and geometry, is lagging behind development of the tokamak configuration. However, it does feature some inherent qualities that have motivated continuation of work in this area.

Plasma behaves like a gas and exerts a (kinetic) pressure toward the exterior, the value of which increases with temperature and density. To confine the plasma, this pressure must be balanced by pressure toward the inside. This is the function of the (magnetic) pressure exerted by the magnetic field. In practice, to avoid the appearance of instabilities, kinetic pressure must be much less than magnetic pressure (a factor of 10). Therefore, for a given magnetic field and temperature, this places a limit on density. With standard temperature (10–20 keV) and magnetic field (5–10 T) values, this limit density is on the order of $10^{20}\,m^{-3}$. As a consequence, this also sets the confinement time value required for magnetic fusion: on the order of a few seconds (refer to the Lawson criterion).

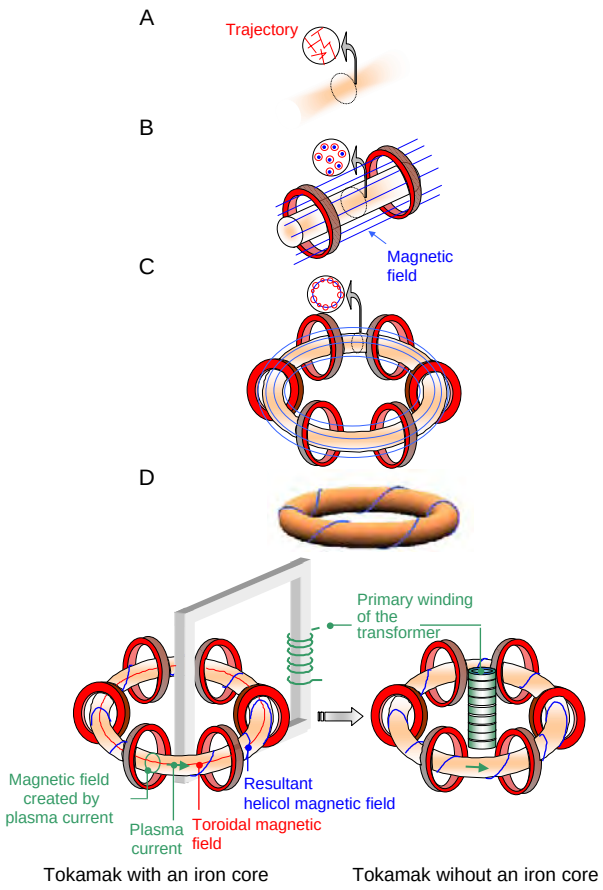


FIGURE 3 Main confinement and tokamak principle.

The stability limit is frequently characterized by the parameter β_n that is a function of plasma topology (e.g., aspect ratio, shape of the poloidal section) as well as current and pressure profiles. This parameter is on the order of 2.5 to 3.0% in current configurations but can achieve 5.0% in "advanced" configurations.

1.5 How Is Plasma Heated?

Irrespective of the manner in which the plasma is created within a confinement structure, it does not start at a temperature necessary for fusion reactions. There are three methods available to heat a plasma:

1. The current circulating in the tokamak plasma heats the plasma by the Joule effect. The latter is efficient up to a temperature on the order of 1 keV. Beyond this temperature, plasma resistivity becomes too weak and this method becomes less efficient. In a stellarator, there is no central current and, therefore, no "ohmic" heating.

2. Heating by injection of high-energy neutral particles (neutral beam heating) involves creating and accelerating an ion beam outside the confinement containment. The beam is then neutralized by charge exchange before penetrating into the plasma, where the neutral particles become ionized again and are confined by the magnetic field. Energy is then redistributed through collisions and the plasma temperature rises.

3. Plasma can absorb energy from electromagnetic waves at the characteristic frequencies of the medium. The energy of the electromagnetic waves is transferred to the plasma by antenna covering part of the confinement area. By choosing the frequency, it is possible to define the particle species (ions or electrons) that will be heated and, to some extent, the area where the wave and heating will be absorbed.

In a magnetic confinement-type thermonuclear fusion reactor, plasma temperature may be raised to a suitable level through a combination of the methods described previously. Once the number of fusion reactions becomes significant, energy contributed by helium nuclei produced that remains confined within the plasma constitutes the main heating method ($Q > 5$).

1.6 Toward Continuous Operation in Tokamak Machine

In addition to heating the plasma, the heating methods described previously allow current to be generated and, thus, allow continuous operation of the tokamak to be envisioned (despite its inherent impulse nature).

Under some conditions, the plasma may also give rise to a plasma-generated toroidal current referred to as the bootstrap current resulting from collisions between circulating electrons and locally trapped electrons. The fraction of this current with respect to total plasma current is highly dependent on temperature and density profiles. Development of reliable plasma scenarios favoring this type of current (e.g., how to maintain good profiles, where and how heat should be applied and current should be generated) is of utmost importance from the standpoint of steady-state operation, which will be the common state of a commercial fusion reactor.

2. *THE FUSION REACTOR*

A schematic diagram of a fusion tokamak-based reactor is provided in Fig. 4. A general description is

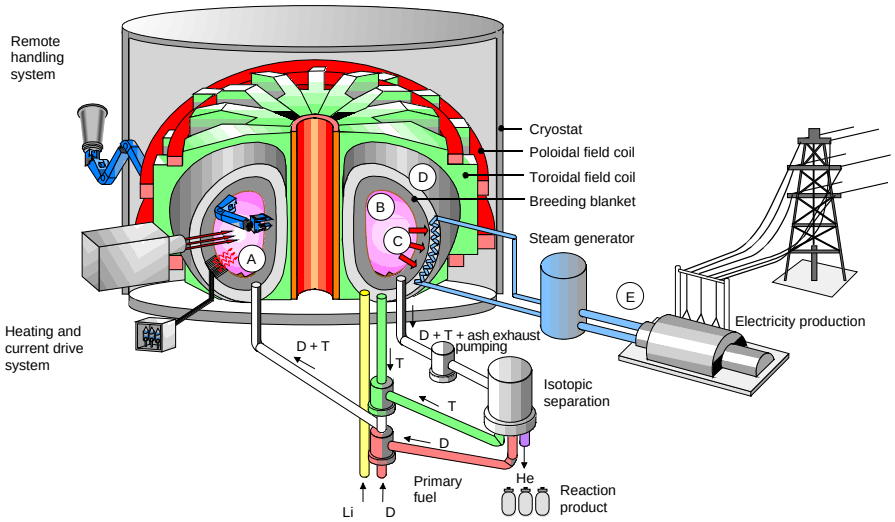


FIGURE 4 Schematic view of a fusion reactor based on a magnetic confinement: (A) vacuum chamber; (B) plasma; (C) plasma radiation; (D) blanket; (E) electricity production.

dealt with briefly here before examining the main reactor elements in detail. The deuterium/tritium fuel mixture is injected in gaseous or frozen ice pellets form into a vacuum chamber (Fig. 4A) where, by means of a confinement and heating system, it passes to the plasma state and burns continuously (Fig. 4B). The plasma produces ash (helium atoms for a D–T plasma) and energy in the form of particles and radiation (Fig. 4C). Ashes are extracted in gaseous form and are processed outside the reactor. Charged particles and radiation are absorbed, losing their energy in the first physical element surrounding the plasma (the so-called "first wall"). Neutron kinetic energy ($\sim$80% of the total energy for a D–T plasma) is converted (or even amplified through suitable nuclear reactions) into heat in the so-called "blanket" (Fig. 4D) component located after the first wall but still within the vacuum chamber. The vacuum chamber encloses the space where the fusion reaction takes place. The first wall, blanket, and vacuum chamber are cooled down by a heat extraction system. The heat is used to generate steam that supplies a conventional turbine and alternator electricity generating system (Fig. 4E).

2.1 Reactor Fuels

As already mentioned, the fusion reaction involving deuterium and tritium is the most accessible and, obviously, is the reaction on which the most credible concepts are based. Tritium is a radioactive element with a relatively short half-life (12.3 years) that will be produced on-site by neutron capture in lithium. Therefore, the primary fusion fuels are deuterium and lithium (Fig. 1), two nonradioactive elements.

Deuterium is a stable isotope of hydrogen. It is very abundant and may be extracted from seawater (33 g of deuterium per m^3 seawater) by conventional industrial processes (e.g., distillation, electrolysis, isotopic exchange). The estimated resource in oceans is on the order of 4.5×10^{13} metric tons, corresponding to an energy potential of 5×10^{11} TW per year. Bearing in mind that current global energy consumption is on the order of approximately 12 TW per year (year 1990), deuterium energy resources would exceed the lifetime of the sun ($\sim$5 billion years).

The lithium content in Earth's crust is approximately 50 ppm (0.05 g/kg). It is more abundant than tin or lead and is even 10 times more abundant than uranium (3–4 ppm). Lithium occurs naturally in a mixture of two isotopes: lithium-6 (7.5% of naturally occurring lithium) and lithium-7 (92.5% of naturally occurring lithium). Although it would be possible to use naturally occurring lithium in a fusion reactor, mixture enriched with lithium-6 (40–90% according to the design) would be preferable because, in this case, reactions are exothermic and feature higher cross sections. The enrichment processes are well known and have been validated (lithium-6 is used to produce tritium for nuclear weapons). Lithium may also be extracted from seawater that contains 0.17 g/m^3, constituting a potential reserve of 230,000 million metric tons. Proposed extraction methods are based on conventional ionic exchange, solvent extraction, or coprecipitation-type chemical processes. In a fusion reactor generating 1 GWe per year,

between 0.5 and 3.5 metric tons of naturally occurring lithium would be consumed depending on the enrichment implemented. Therefore, the use of telluric lithium would ensure reserves for several thousand years, and use of lithium extracted from seawater would push this limit to several million years.

Use of "advanced" fuels (D–D or even D–^{3}He reactions) features the dual advantage of avoiding use of tritium and production of high-energy neutrons. However, conditions regarding temperature, density, and confinement time necessary to compensate the small cross sections render their implementation extremely hypothetical and, in any case, impossible in the near future. It should also be emphasized that in a reactor based on these reactions, nearly all fusion energy produced is transferred to the first wall elements, accentuating the constraints on these components that were already very high (in the case of D–T reactions, 80% of fusion energy is carried by neutrons that give up their energy in the tritium breeding blanket).

2.2 Reactor Physics

Since the beginning of the tokamak era at the end of the 1960s, considerable advances have been made, both regarding understanding of physical phenomena and regarding technologies implemented in the construction of experimental tools. These results were obtained for numerous facilities of various sizes designed during the 1970s and operated since the beginning of the 1980s. Two highlights should be retained:

1. One is the 16 MW of fusion power obtained in the European device, JET, over approximately 1 s in 1997, with a power amplification factor, Q, of approximately 0.65, close to break-even conditions ($Q = 1$).
2. Another is long-duration plasmas produced in the Tore Supra device (Cadarache, France), fitted with superconducting magnets and plasma facing components actively cooled by water circulation. These technologies enabled several discharges of a few minutes such as the record discharge on September 18, 2002, of $4\frac{1}{2}$ min, sustained by a heating power of 3 MW and requiring to exhaust more than 750 MJ of thermal energy during the experiment (Fig. 5).

Reliable responses to most questions raised by the studies concerning controlled fusion have already been provided, albeit independently. The next step should involve integration of all these results within a single facility. This is one of the goals of the International Thermonuclear Experimental Reactor (ITER) project.

The objectives of current devices target mainly qualification of confinement scenarios selected for next generation machines. Thus, it was discovered that confinement improves with increasing machine size deteriorates with increasing additional heating power coupled to the plasma and that, under some conditions, there is a power threshold where confinement is suddenly improved. This improved confinement system is referred to as H mode (for "high confinement"), as opposed to the confinement mode obtained below the power limit, referred to as L mode (for "low confinement"). This allows the confinement time to be improved by a factor of nearly 2 with respect to L mode. Discovery of this improved confinement mode was critical, and H mode currently constitutes the reference scenario for next generation devices, notably for ITER. Stabilizing mechanisms that allow reach into H mode are not yet entirely understood but are based on stabilization of turbulence, which is the source of confinement deterioration. A transport barrier is created at the edge of the plasma, holding heat and particles in the plasma core. The most characteristic point of this scenario is the appearance of strong gradients in the edge zone. These extremely steep gradients at the edge give rise to instabilities specific to H mode, referred to as ELM (for edge localized modes). The plasma pressure profile relaxes periodically toward shallower slopes, releasing bursts of particles and heat that escape from the plasma at each ELM, thereby imposing significant constraints on vacuum chamber components.

Results acquired on a global scale from tokamak experiences were gathered in a database from which an empirical scale law was established expressing the confinement time based on the main parameters of machines and plasma. Although this method is empirical, it is extremely important to be able to extrapolate current performance levels to those of next generation machines.

Physics options adopted for reactor operation are obviously based on strong or weak extrapolations of current scenarios. They target continuous operation or at least operation with extremely high discharge durations (> 12 h) and, therefore, are based on implementation of scenarios favoring a significant fraction of noninductive current. These scenarios guarantee good confinement and control of turbulence phenomena that could damage first wall elements. Reactor size is highly dependent on these extrapolations. Diagrammatically, two major extrapolation classes limit the domain (Table I):

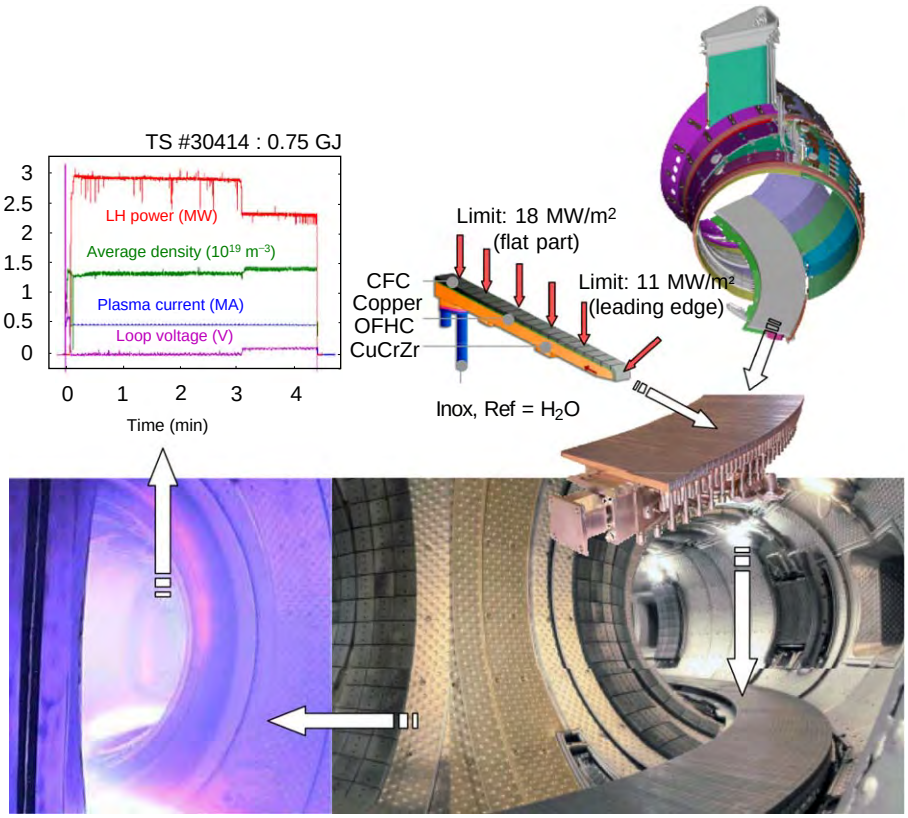


FIGURE 5 Record discharge reached in Tore Supra on September 18, 2002: a plasma discharge lasting 4 minutes 25 seconds requiring the exhaust of more than 750 MJ of thermal energy. This remarkable result was obtained due to the actively cooled plasma facing components, ensuring that every square centimeter of the wall in front of the plasma is actively cooled.

1. *Reactors involving moderate extrapolations from those adopted for ITER*. However, they imply suppression of H mode ELM and control of power transmitted to first wall components by controlling edge plasma radiation (e.g., injection of impurity). The bootstrap current fraction is generally below 50%, leading to a heating power that is quite high and, therefore, a recirculating power (power taken from the system to supply heating sources) on the order of 20%. The reactor efficiencies are similar to those acquired with conventional nuclear power plants (30–40%). Regarding size, the major radius is typically approximately 9 m (1500 MWe class).

2. *Reactors with large extrapolations from current knowledge but still credible*. These are generally accompanied by the same extrapolations at a technological level. Plasma parameters and topology are selected to maximize high-bootstrap current fraction, close to 75%, minimizing recirculating power. Instabilities are assumed to be completely controlled, as is plasma radiation at the edge where a radiative "mantle" is established without any repercussions on confinement and performance levels of the central plasma. This leads to heat fluxes on the first wall components lower than those of previous designs (factor 2) and a reactor size close to that adopted for ITER, that is, typically 6 m (1500 MWe class).

2.3 Plasma Facing Components

The plasma facing components constitute the first physical surface encountered by the plasma. Excluding accidental situations, the plasma does not come into contact with this surface. In normal operation, plasma facing components are exposed to neutronic, thermal, radiative, and thermomechanic stress loading. Thus, the management of the interface between the plasma and this wall is critical. Furthermore, the main results obtained concerning confinement, like H mode, were based on efficient management of this interface. Several types of phenomena may enter into play at the wall, for example, absorption of large quantities of gas issued by the plasma, erosion due to

TABLE I

Main Parameters of Two Generic Fusion Reactors

Parameter	ITER	Reactor type 1 (small extrapolation)		Reactor type 2 (advanced)	
Fusion power (MW)	500	3500		2500	
Electrical power (MW)	—	1500		1500	
Net reactor efficiency (percentage)	—	35–45		60–65	
Q	10	20		35	
Plasma duration	400 s	Continuous		Continuous	
Major radius (m)	6.2	9–10		6	
Minor radius	2.0	3.0–3.3		2	
Aspect ratio	3.1	3.0		3	
Plasma current (MA)	15	30		15	
Toroidal field on axis (T)	5.3	7		5.6	
Toroidal field on coil conductor (T)	12	13		13.5	
Elongation (95%/separatrix)	1.7/1.85	1.7/1.9		1.9/2.1	
Triangularity (95%/separatrix)	0.33/0.48	0.27/0.4		0.47/0.7	
Normalised β	2.0	3.3		4.5	
Average temperature (keV)	10	20		12	
Average density ($10^{20}/m^3$)	1	1.1		1.4	
Bootstrap fraction (percentage)	Experiment	35		75	
P_{ext}: heating power (MW)	70–110	250		70	
Average neutron load (MW/m^2)	0.6	~2		~2.5	
Average divertor heat load (MW/m^2)	10	10–15		5	
Armor material	Be, CFC, W	W		W	
Breeder	None	Pb-17Li	Li_2TiO_3 (pebble bed)	Li	Pb-17Li
Multiplier	None	None	Be (pebble bed)	None	None
Coolant	Water	Water	He	Self	Self
Typical outlet temperature (°C)	150	325	500	600	1100
Structure	316	Ferritic	Steel	V alloy	SiC/SiC
Reactor efficiency (percentage)	—	33	37	45	60

Note. Type 1 is a "small extrapolation" front current physics and technology. Type 2 is more advanced and is based on large (but credible) extrapolations.

the action of fast particles, heating (the heat flux can reach several megawatts/square meter). Furthermore, plasma radiation losses vary significantly with the quantity and type of impurities that plasma contains (radiation increases with the atomic number of the relevant impurity). Therefore, the plasma facing wall must be designed, on the one hand, to minimize this type of pollution and, on the other, to withstand particles and radiation from the plasma. To illustrate the problems and issues associated with these components, we propose focusing on the so-called divertor.

All charged particles present in the plasma are subject to a confinement effect caused by the magnetic field. Consequently, if no action is taken, Helium ion concentration within the plasma increases with the combustion rate. This effect must be limited because not only is the fuel diluted, but losses through radiation also increase. A specific device referred to as the divertor (Fig. 6) manages continuous extraction of reaction ash and, thus, maintains its concentration in the plasma at an acceptable level. It involves creating, by means of the poloidal magnet system, a final closed magnetic surface that stretches in two sections to the divertor. Beyond this separatrix, all magnetic surfaces are open and reach the wall. A particle first confined close to the magnetic axis moves slowly away by diffusion to reach the closed magnetic surface and will then encounter the divertor wall. Here, it is neutralized and, thus, may be extracted by pumps that continuously empty this zone. The divertor fulfills three functions that

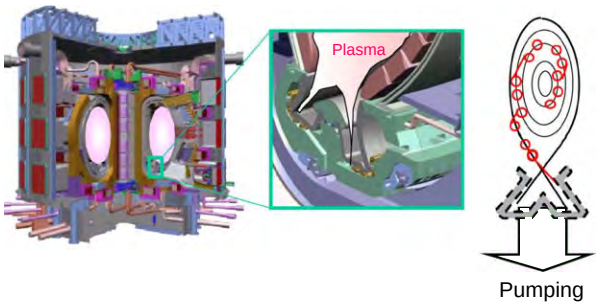


FIGURE 6 Principle of the divertor.

have extremely significant repercussions for the plasma:

- Extraction of particles during deconfinement (e.g., helium, deuterium, tritium, impurities) in a precise and controllable location
- Extraction of heat (particles striking the divertor lose their energy here)
- Control of plasma density through the divertor pumping capacity

From a technological point of view, heat extraction at the level of several megawatts per squared meter is rather challenging. Creation of a divertor presents technological difficulties associated with extraction of heat. Particle flows along the two lines of the separatrix create thermal fluxes of several megawatts per squared meter. These fluxes can be attenuated by locally injecting impurities (e.g., neon, argon) to increase radiation in this zone and by sweeping the separatrix over the entire surface of the divertor.

Components constituting the divertor are made of armor tiles mounted on a cooled structure. Numerous designs were developed through judicious selection of materials, including the armor tiles, structural materials, joining methods, and coolant.

Designs validated on current machines include a stainless steel supporting structure to which carbon fiber composite (CFC) and hardened copper (CuCrZr) armor tiles are attached and are capable of withstanding continuous thermal fluxes on the order of several megawatts per squared meter (Fig. 5).

The ITER plasma facing components design was derived from these concepts. The armor tiles will be made either of a low Z material, such as CFC or beryllium, or a refractory material, such as tungsten, for the most exposed sections of the divertor. These technologies were tested successfully in laboratories in specific installations under thermal fluxes reaching 20 MW/m^2 over thousands of cycles, but they have not yet been qualified at the industrial level.

From a reactor standpoint, objectives regarding lifetime, reliability, and safety must be satisfied, as must economic criteria, along with the constraints already imposed by plasma physics. Notably, the energy transferred to the first wall and the divertor (20% of energy released by a D–T plasma) must be recovered to be used to generate electricity. Therefore, the coolant outlet temperature must be as high as possible. Safety criteria, notably those involving authorized tritium inventories, lead to the exclusion of carbon fiber composites because these materials trap and easily retain tritium. The mechanical strength of beryllium and its erosion rate render implementation of this material in a reactor environment difficult. Tungsten is the most promising material, although its mechanical strength under irradiation is still being studied. Its implementation is the basis of plasma facing components used in several reactor designs (Fig. 7). Schematically, there are four divertor types:

1. *Designs based on extrapolations that are limited with respect to technologies implemented in ITER.* They use water under pressurized water reactor (PWR) conditions as coolant. This type of design allows the withstanding of heat fluxes of 15 MW/m^2. The structural material constituting the cooling tube is a hardened copper alloy (CuCrZr) featuring thermal conductivity properties 30 times greater than that of the stainless steel used in ITER. The armor material is tungsten jointed to a copper tube through of a compliance layer. The operation constraints of these materials, as well as deterioration of CuCrZr thermomechanical performance levels under thermal cycling, are taken into consideration. Uncertainties are also clearly identified, as is the behavior of CuCrZr under irradiation. Use of CuCrZr imposes a coolant temperature that is quite low (150°C). Optimized versions using steel and thermal barriers allow a higher coolant temperature (325°C) and improved efficiency.
2. *Designs involving helium coolant, tungsten as the armor material, and steel as the structural material.* This type of design allows resistance to heat fluxes on the order of 10 MW/m^2 while using a high-temperature coolant (500°C or even 800°C in some optimized versions).
3. *Designs involving liquid metal as the coolant and tungsten as the armor material.* These designs generally use Li–17Pb as the coolant and SiC–SiC-type ceramics as the structural materials. This would require highly significant technological advances, notably concerning improvement of thermal–mechanical properties and the resistance of these materials

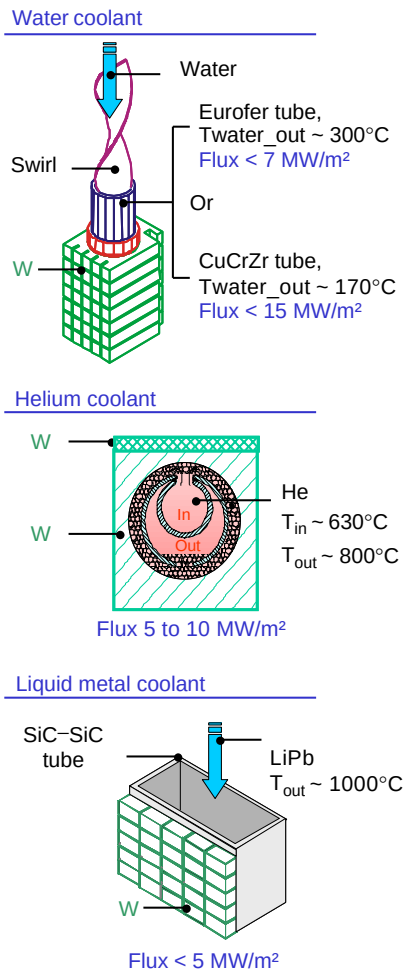


FIGURE 7 Different types of high heat flux components for divertor concepts.

when subject to irradiation. They would allow coolant temperatures of approximately 1000°C. These designs are generally intended for reactor versions that also use significant extrapolation regarding physics and adequate control of the plasma edge.

4. *Alternative designs*. There are other "nonconventional" designs in which the protective material is replaced by liquid metal (e.g., lithium, gallium) in the form of a continuously flowing film, drops, or in pool. The idea is to overcome problems of erosion, lifetime, and thermal load associated with solid materials (this type of design supports fluxes of several tens of megawatts/squared meter). However, some issues are also under investigation regarding, for example, the effects of these liquid surfaces on the plasma edge or the pumping capacity of helium and impurities.

Regarding the reactor, plasma facing components and the divertor specifically remain an open issue. The main difficulty concerns integrating complicated phenomena that control the plasma edge and technological constraints due to materials. ITER will provide some crucial answers in this area regarding feasibility and operating procedures for a reactor.

2.4 The Tritium Breeding Blanket

The element located after the plasma facing components is called the blanket (Fig. 4). This component performs several functions:

- First, the blanket provides shielding by significantly reducing neutron/gamma radiation energy to protect the outer following components (e.g., vacuum chamber, magnetic system).Its second function is to recover the energy left by the neutrons by heating materials. A coolant circulates in the structure and evacuates the heat generated to conventional equipment for electricity production.
- In the case of a reactor based on the D–T reaction, the blanket must also produce tritium necessary for fusion reactions by neutron bombardment on lithium:

$$n + 6\mathrm{Li} \rightarrow T + 4\mathrm{He} + 4,78\ \mathrm{MeV}$$

$$n + 7\mathrm{Li} \rightarrow T + 4\mathrm{He} + n - 2,47\ \mathrm{MeV}$$

One should note that the first of these two reactions is exothermic and, thus, is preferable. This requires 6Li-enriched lithium to be used. This enrichment operation is simple and already mastered in industry (for military purposes). It may also be noted that the blanket is the site of reactions generating energy that altogether account for 20% of the reactor balance. Lithium may be in a solid form (ceramic) or a liquid form (metallic alloy) depending on the blanket design implemented. The presence of a tritium breeding material does not suffice to ensure suitable production of tritium. Ended, the reaction of a tritium and deuterium atom produces only one neutron, and the reaction of this neutron on lithium produces only one tritium atom. Because neutron losses are inevitable, an adequate amount of tritium cannot be produced under these conditions. To remedy this drawback, a neutron multiplier is introduced into the blanket, allowing the appropriate balance. Lead or beryllium may be used as a multiplier. Blanket tritium-generating capacity is characterized by the self-sufficiency ratio (or tritium breeding ratio). If the tritium breeding ratio is greater than 1, the blanket produces more tritium than is consumed by the reactor.

Numerous tritium breeding blanket designs are available that fulfill the three functions just

described. They differ in terms of tritium breeding material type, coolant type, and structural material types implemented. The combinations adopted are the result of trade-offs regarding compatibility of materials with each other and the allowable operation windows (e.g., operation temperature, swelling resistance). The choice of coolant will also depend on solutions adopted for plasma facing components, with a single coolant being preferable for obvious reasons of simplicity. The choice of structural material is also vital because, in general, it is the structural material that dictates performance levels in terms of efficiency (via its maximum allowable operation temperature) and regarding lifetime of blanket elements. The main criterion is resistance under 14-MeV neutron irradiation, where a maximum of 150 dpa is usually adopted, corresponding to approximately 5 years operation at full power. Therefore, periodic replacement of blanket components is provided for in the design. The long-term induced activation of components can be tailored by proper selection of materials to avoid generation of waste that would require deep geological disposal. These materials (so-called "low-activation" materials) could be steel in which elements that are penalizing regarding activation are replaced with other elements that are more interesting but compatible from a metallurgical standpoint (e.g., nickel, molybdenum) or other material families. These are presented in order of development risk:

- Low-activation ferritic–martensitic steels have been developed to the point where their use in a nuclear component may be possible in the relatively short term. They are notably developed in Japan and in Europe, with the latter having focused its studies on said materials. A particular grade was selected (EUROFER steels) and has already given rise to significant melts (a few metric tons), allowing the manufacture of samples that were subject to numerous tests (behavior under irradiation, corrosion, welding, etc.). The operation temperature is approximately 550°C, or even 600 to 650°C, through specific optimizations (e.g., ODS steel, composition optimization). Operation temperature windows lead to efficiency of approximately 35 to 40%. The main question that still remains open concerns the effect of 14-MeV neutrons and, specifically, the effect that the formation of hydrogen and helium produced by transmutation has on embrittlement.
- Vanadium alloys are interesting in that they can be used in up to 700°C. This would allow yields of up to 45%.
- Silicone carbide composites (SiC–SiC) could operate at very high temperatures (1000°C) and, therefore, allow access to efficiency close to 60%. However, their status of development would not allow them to be used in the short term.

Schematically, there are two main types of blankets.

2.4.1 Solid Tritium Breeding Material Designs

In this type of blanket, the tritium breeding material is a lithiated ceramic (Li_2O, Li_4SiO_4, Li_2TiO_3, or Li_2ZrO_3) and beryllium is used as the neutron multiplier material. These materials are generally in the form of pebble beds (diameter on the order of millimeters). Tritium is extracted by gas circulating (He) that is in contact with the tritium breeding material and transports tritium outside the blanket. The structure is made of low-activation ferritic–martensitic steel, possibly hardened (ODS-type steel) to allow a higher coolant temperature. Coolant may be either pressurized water (typically 320°C and 15 MPa) or helium (typically 500°C and 8 MPa), allowing efficiency of between 35 and 40%. The main difficulties taken into consideration in this type of blanket design are as follows:

- Chemical compatibility of beryllium and water that may induce production of hydrogen and a risk of explosion
- Tritium permeation in water
- Evolution of pebble beds (ceramic and beryllium) when subject to irradiation
- Lower shielding capability if helium is used as a coolant
- Fuel manufacturing costs, the renewal obligation due to consumption of lithium, and fuel reprocessing to recover lithium that has not been consumed

2.4.2 Liquid Tritium Breeding Material Designs

In this type of blanket, the tritium breeding material is a liquid metal, either metal lithium, an eutectic of lithium and lead(Pb–17Li), or a molten salt composed of lithium, beryllium, and fluorine called Flibe (Li_2BeF_4). Usually, the multiplier material is incorporated into the tritium breeding material. Liquid metal circulates in the reactor and, outside the reactor, undergoes processing designed to extract tritium present in dissolved form. This also allows online control of the lithium-6 titer, thereby avoiding depletion of the fuel in lithium and the requirement to replace lithium periodically. The Pb–17Li eutectic

is frequently used due to its significant tritium production capacities, high thermal conductivity, ability to withstand irradiation damage, and inertness with air. Pb–17Li compatibility with ferritic–martensitic-type steels has been proven in up to 500°C. Use of helium as a coolant leads to difficulties in achieving self-sufficiency in tritium, and water (or even the liquid metal itself) is preferred. This gives rise to two types of liquid tritium breeding blankets:

1. *Water-cooled liquid blankets (typically 325°C and 15 MPa)*. These are generally associated with steels as the structural material and give rise to efficiency of approximately 35%.
2. *Self-cooled blankets where the tritium breeding material itself acts as the coolant*. This is implemented at high temperatures and is associated with a compatible structural material (vanadium alloy or SiC–SiC composite). Efficiency can be as high as 60%.

The economic performance and technological maturity of the various designs is provided in Fig. 8.

2.5 The Magnetic System

Coils, in which an electrical current circulates, are used to create the magnetic fields necessary for reactor operation; current is constant in the toroidal coils and variable in the poloidal coils. Toroidal coils are typically more than 10 m in height, and the largest of the poloidal coils measure more than 20 m. Currents passing through said coils are high intensity (a few tens of kiloamps) and, therefore, give rise to an electrical energy consumption problem. For a reactor under steady-state operations, implementation of conventional coils (water-cooled copper coils) severely depletes the total energy balance; therefore, superconducting coils should be used. For these coils, electrical consumption is limited to the cryogenic system that maintains coils at very low temperatures (~4 K). This technology has been implemented since 1989 to generate the toroidal magnetic field of the Tore Supra tokamak and will be applied to the ITER magnetic system.

Regarding the toroidal fields, the highest values possible must be achieved due to the relationship between kinetic pressure and magnetic pressure. Limits are imposed by the superconducting materials selected that, to preserve superconductor status, must fulfill some criteria: temperature must be less than the critical temperature, *Tc;* the magnetic field applied must be less than a certain value, *Hc;* and the current density applied must be less than the critical current density, *Jc*. Furthermore, the superconducting material is fragile and loses its superconducting properties when subject to high-stress loading. Reactor designs provide for fields on the conductor in the neighborhood of 16 T at a standard operating temperature of 4 K, values that are not far from those adopted for ITER (13 T and 4 K). Within these field ranges, the operation limits of Nb_3Sn, selected for ITER, are achieved and Nb_3Al could be used. For higher fields, implementation of immature

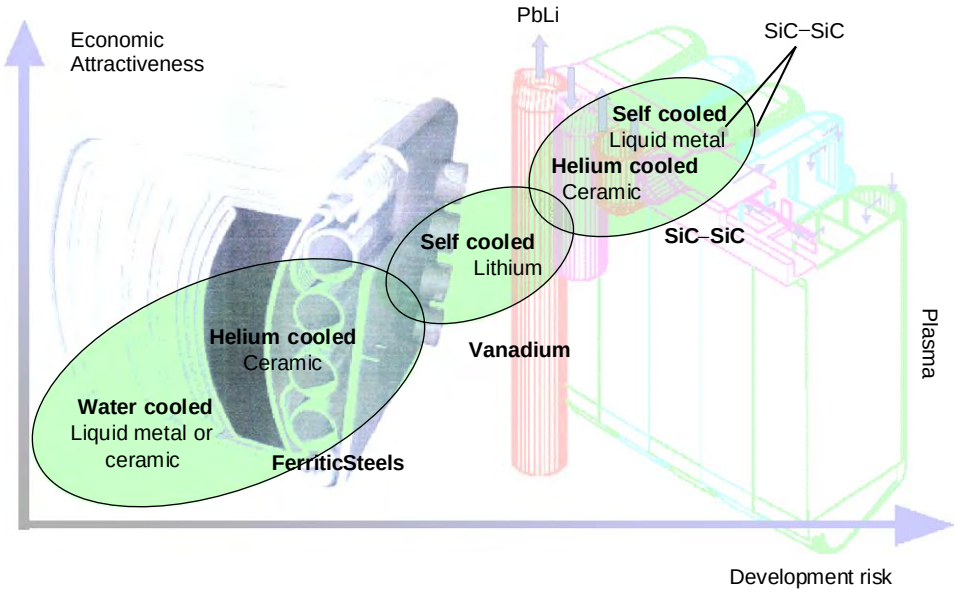


FIGURE 8 Economic attractiveness and development risk of various kinds of breeding blankets.

technologies may be considered, for example, use of Nb_3AlGe(24 T on the conductor) or even implementation of superconducting materials referred to as critical high-temperature materials.

There are several superconducting cable designs available. However, recall that the strand consists of filaments of superconducting material (NbTi or Nb_3Sn, diameter of a few microns) embedded in a copper matrix and that the total strand diameter is slightly less than 1 mm. Using this strand, a cable is made composed of more than 1000 strands positioned according to a precise architecture. The cable obtained in this manner is placed within a jacket (e.g., steel, incoloy) intended to withstand mechanical stress loading (Fig. 9). Liquid helium circulates and bathes the cable, and the entire magnetic system is placed within a huge cryostat. Manufacture of the conductor is quite sophisticated and implies other steps that are not described here, intended to ensure that superconducting properties (heat treatment) and good stability (introduction of copper strands) are achieved.

2.6 Robotized Maintenance

The operation of fusion reactors will require maintenance phases. These outage periods (scheduled or not) must be minimized to avoid penalizing facility availability (80% availability is targeted). Implementation of remote interventions is absolutely mandatory, notably to perform interventions (e.g., inspection, replacement of components) in the vacuum chamber where the radioactive environment does not allow for other solutions. Therefore, all of the objects already mentioned (e.g., plasma facing components, tritium breeding blankets) are designed for remote interventions. Even the reactor environment will be different; complex maintenance operations have already been performed successfully, for example, the fully remote exchange of the JET divertor.

Therefore, maintenance will be taken into consideration as of design, in particular for the divertor and tritium breeding blanket modules that must be replaced periodically. For example, a trade-off between segment size (limited by carrier capacity) and segment number (limited by intervention rapidity) controls the segmentation of tritium breeding blanket segmentation. Several maintenance schemes have been proposed. Their repercussions on the general design of the reactor are extremely significant. The three main schemes are as follows (Fig. 10):

1. The blanket and divertor are segmented into modules (Fig. 10A), where each module (blanket plus first wall and divertor) is removed through specific ports, the number and size of which are optimized so as to minimize intervention time.
2. The torus is divided into sectors (typically 16 sectors), with each sector including a toroidal coil, the entirety of the first wall, and the blanket (Fig. 10B). Connection interfaces are minimized, and the number of objects to be handled is reduced but leads to a sophisticated manipulation.

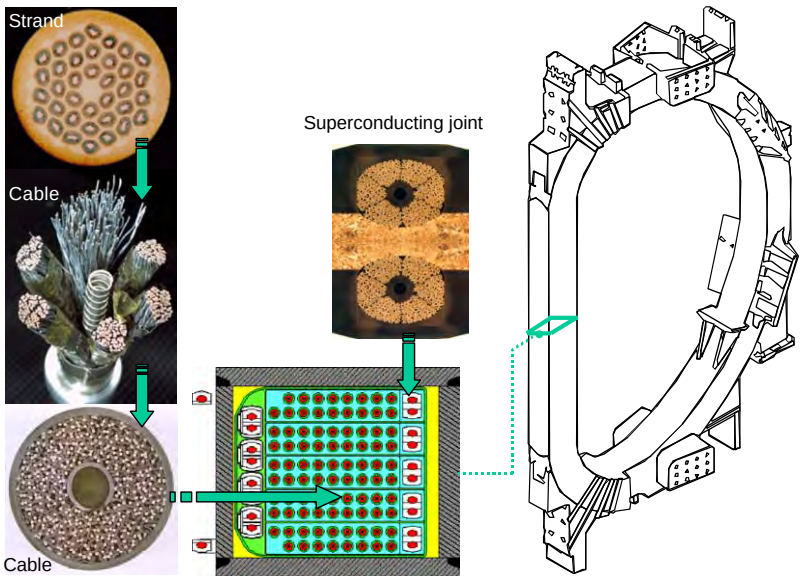


FIGURE 9 Example of superconducting cable (ITER toroidal field coil cable).

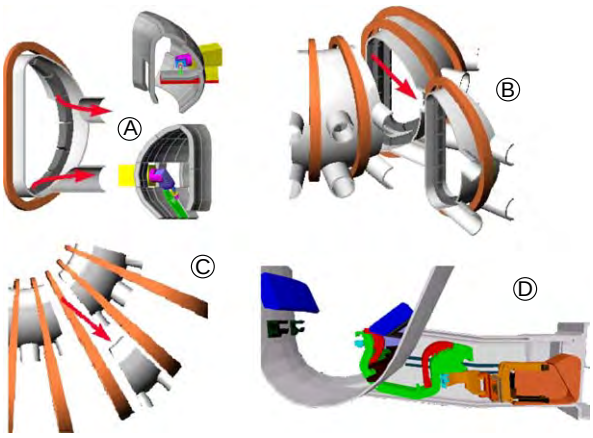


FIGURE 10 Possible maintenance schemes for divertor and blanket modularized blanket (A), large sector with toroidal coil (B) or without toroidal coil (C), and divertor maintenance (D).

3. The torus is divided into sectors that may be handled whole with the exception of toroidal coils (Fig. 10C). This assumes that toroidal coils are large. Poloidal coils are trapped and are manufactured in place. This configuration is interesting because it frees the radial space, thereby facilitating maintenance operations because a sector may be replaced in a few tens of hours.

2.7 Safety

The viability of fusion as a future energy source will be partially determined by factors associated with safety and the environment. Notably, two fundamental objectives need to be verified:

1. It must be demonstrated that no accident would require evacuation of the population.
2. Waste should not constitute a burden on future generations.

To fulfill both of these conditions, the fusion reactor falls back on its inherent advantages (e.g., the absence of any risk of a runaway reaction) and on the integration of solutions offered by implementation of structural materials referred to as low-activation materials in solution design.

2.7.1 Advantages Regarding Safety

The primary fuels of the reactor are deuterium and lithium, two nonradioactive and nontoxic bodies. Small quantities of these elements will be used (<1 kg of deuterium and <10 kg of naturally occurring lithium are enough to supply a day's generation of 1000 MWe). Therefore, management of procurement, transport, and storage of fuel would appear to be relatively simple.

Fusion reaction conditions require the use of a plasma that is not dense (a few grams of fuel in a volume of several tens of cubed meters), is extremely pure, and is at a high temperature. Therefore, the quantity of fuel in the combustion chamber during the reaction is always very low. The slightest uncontrolled interference with this environment causes the rapid cooling and automatic shutdown of the fusion reaction. There is no possibility of uncontrolled power runaway because inherent processes limit reactivity excursions of the plasma. After the plasma shutdown, the residual energy is low and the structures cannot suffer any major damage due to accidents. Thus, the main safety function to ensure is confinement (for a fission reactor, control of the reaction and removal of residual power must also be ensured). Particular attention is paid to the confinement strategy due to the presence of tritium characterized by high diffusion in most materials. Thus, design engineers have focused on minimizing the tritium inventory, both in the torus and in the fuel cycle. To minimize emissions in the event of an accident, the "multibarrier" approach is generally adopted. This provides for three confinement barriers: the vacuum vessel, the cryostat, and the reactor building itself. Analyses that were as detailed as possible concerning energy inventories, accidents likely to occur, and expected repercussions were undertaken. These works revealed that there is no accidental sequence that could impugn reactor integrity. Even in the event of a severe accident, it would not be necessary to evacuate the population around the site.

2.7.2 The Absence of Long-Term Radioactive Waste

With the exception of first start-up that requires an initial tritium load, a fusion reactor does not require transport of radioactive material.

At the end of the fusion reactor life, material surrounding the plasma comprising the reactor structure will be activated. Implementation of low-activation materials (or, more precisely, rapid decay time materials) will allow the induced radioactivity to be minimized.

The quantity of materials to be considered varies according to the reactor design, generally between 60,000 and 100,000 metric tons. After a period of 100 years following final decommissioning of the reactor, the majority (or even entirety) of materials may be considered as very low-activity waste (meeting nuclear waste declassification standards defined

by the International Atomic Energy Agency) or may be recycled in the nuclear industry. This quality may be illustrated by a strong image (Fig. 11); after 100 years of decay, the average radioactivity of fusion reactor materials is lower than that of ash from an amount of coal that would have generated the same quantity of energy (coal contains trace amounts of uranium and thorium). Elimination of fusion waste by the generation that resulted in its creation constitutes a realistic objective. Specifically, there is no requirement to provide for deep storage.

2.8 Economy

It seems premature to speak of kilowatt-hour production costs for an energy source that is still several tens of years away from commercialization. The first objective is to illustrate the repercussions of various physical variables or technological production assumptions on costs. These results, which set the relative orders of magnitude and variation directions, indeed have a direct impact on development strategy.

The second point of interest constitutes verification that the proposed device complies with market demand. Economic models used correspond to a direct extension of models used for the design, optimization, and cost calculations of current machines or projects such as ITER. From many standpoints, ITER is similar to the reactor, but above all, construction costs were calculated directly by the involved industry in Europe, Japan, Russia, and the United States. Thus, a solid evaluation basis is available and, to return to kilowatt-hours, outstanding uncertainties are significant but concern reactor availability more than the direct cost of its components.

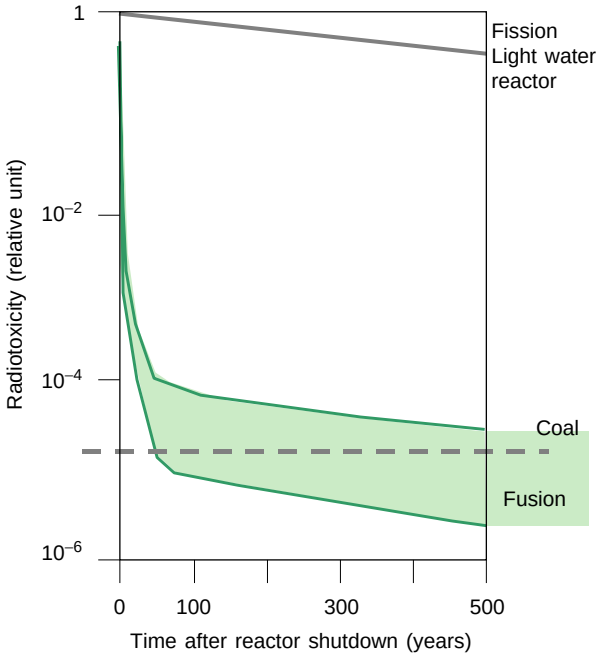


FIGURE 11 Radiotoxicity of fusion waste after reactor shutdown.

2.8.1 Direct Costs (or Internal Costs)

In terms of energy production, direct costs (or internal costs) correspond to costs invoiced to consumers. Put simply, they include plant construction costs, general operation costs, costs associated with fuel purchases, and costs of waste storage and plant dismantlement (regarding the nuclear industry).

The standard direct cost structure for various energy sources is provided in Fig. 12. It may be noted that for a fusion reactor, fuel cost accounts for less

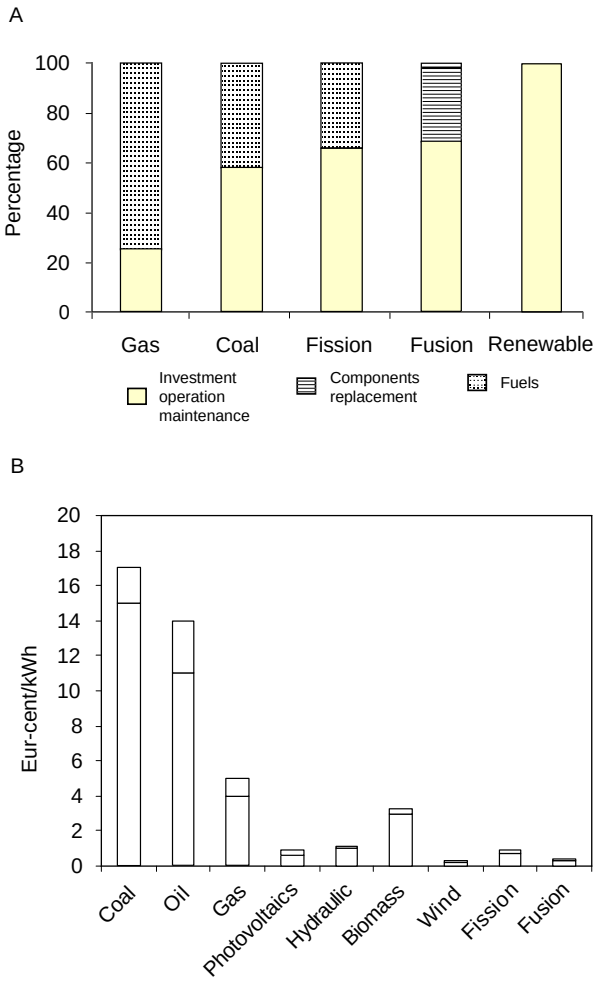


FIGURE 12 Typical breakdown of direct costs for various sources of energy (A) and comparison of externalities of several energy sectors (B).

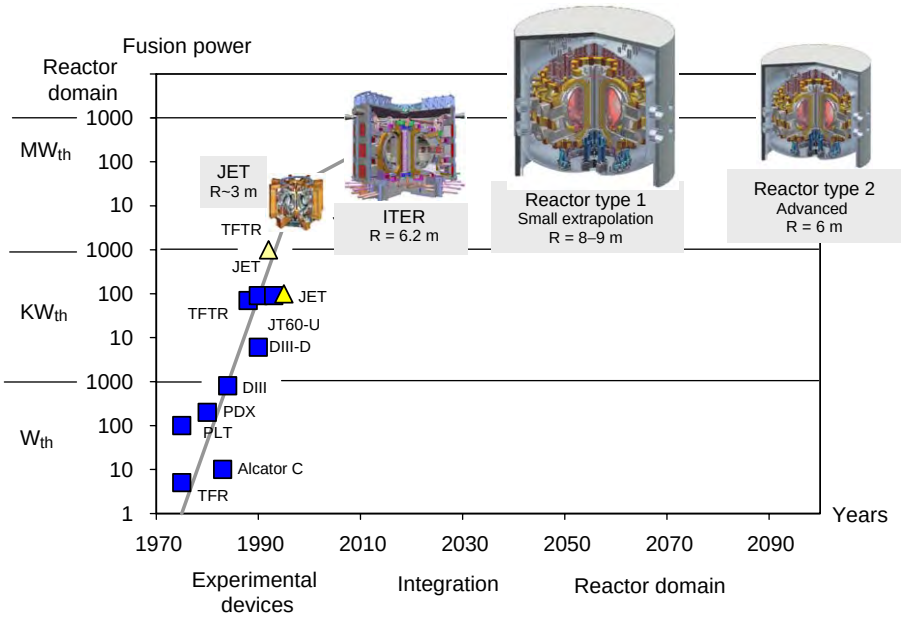


FIGURE 13 Path to the reactor.

than 1% of the total kilowatt-hour cost. Thus, the cost of energy generated by a fusion reactor is calculated by the volume of initial investment to which the cost of regular replacement of aged components is added.

The standard volume of the fusion reactor combustion chamber is on the order of $1000\,m^3$. For comparison purposes, the reactor vessel volume of a fission reactor (1400 MWe) is less than $300\,m^3$. Consequently, fission investment requirements are high simply due to facility size. The absence of fuel cost partially compensates for this drawback with respect to conventional energy sources (e.g., coal, gas, fission), but only slightly. Currently, economic calculations place the cost of a fusion kilowatt-hour at between U.S. $45 million and $90 million, with a probable value of approximately $65 million, that is, twice the cost of conventional energy production (e.g., coal, gas, fission) and between the cost of offshore wind production and photovoltaic production (assuming that in the case of these two latter sources, evaluations do not take energy storage cost into consideration). Therefore, two results should be retained. First, and contrary to what is occasionally stated, fusion reactor production costs would not automatically jeopardize the future of fusion energy. Second, the fusion energy source is characterized by a high capital cost.

2.8.2 Indirect Costs (or External Costs)

The concept of externality or external costs corresponds to a method enabling the environmental impact of an activity to be measured. The European Union has developed a method to assess externalities of an energy production system (ExternE studies). It takes into consideration identification of the emissions attributable to this system, the transfer of pollutants in the environment, and finally determination and quantification in terms of the cost of repercussions for the environment and health. This analysis is performed at all stages of the reactor type in question (e.g., fuel extraction, plant construction, operation, accident, dismantlement). For example, it allows consideration of harmful effects that mining operations have on health or of pollution associated with using fossil-based energy sources (e.g., respiratory problems). Applied to fusion, this method illustrates that the external costs of generating electrical energy by means of the fusion reactor are approximately the same as for renewables and are well below those obtained for fossil-fueled power (Fig. 12).

3. CONCLUSION

Fusion energy features numerous qualities. It is a nearly unlimited energy source that does not generate any greenhouse gas effect or environmental pollution, and it features undeniable advantages in terms of safety. With adequate design, radioactive wastes from the operation of a fusion plant should not constitute a burden for future generations.

Harnessing this energy is on a par with the advantages it features. Despite still very challenging open issues (e.g., physics of a plasma in combustion, heat extraction, manufacturing of sophisticated components such as the blanket, effect of 14-MeV neutrons on materials), scientific bases today are solid enough to realize a device demonstrating the scientific and technological feasibility of fusion energy. This is the main goal of the international ITER project.

Insertion into the energy offer could take place in the second part of the 21st century (Fig. 13), at a time when the exhaustion of conventional resources and climatic consequences of our consumption will start to have serious repercussions. It is absolutely clear that it is the duty of this generation to prepare the knowledge base and know-how that will allow tomorrow's decision makers to serenely consider all of the possible energy solutions. This is the objective of current research into fusion.

SEE ALSO THE FOLLOWING ARTICLES

Nuclear Engineering • Nuclear Fission Reactors: Boiling Water and Pressurized Water Reactors • Nuclear Fuel: Design and Fabrication • Nuclear Fuel Reprocessing • Nuclear Power Economics • Nuclear Power Plants, Decommissioning of • Nuclear Power: Risk Analysis • Nuclear Proliferation and Diversion • Nuclear Waste • Occupational Health Risks in Nuclear Power • Public Reaction to Nuclear Power Siting and Disposal

Further Reading

Borrelli, G., *et al.* (2001). "Socio-Economic Research on Fusion (SERF): Summary of EU Research 1997–2000," European Fusion Development Agreement report. European Union, Paris.

Cook, I., *et al.* (2001). "Safety and Environmental Impact of Fusion (SEIF)," European Fusion Development Agreement report. European Union, Paris.

Giancarli, L., *et al.* (2002). Candidate blanket concepts for a European fusion power plant study. *Fusion Eng. Design* **49/50,** 445–456.

International Atomic Energy Agency. (2001). "Summary of the ITER Final Design Report." ITER EDA Documentation Series No. 22. AIEA, Vienna, Austria.

International Atomic Energy Agency. (2002). "ITER Technical Basis Details." ITER EDA Documentation Series No. 24. AIEA, Vienna, Austria.

Marbach, G., *et al.* (2002). The EU power plant conceptual study. *Fusion Eng. Design* **63/64,** 1–9.

Nuclear Power Economics

GEOFFREY ROTHWELL
Stanford University
Stanford, California, United States

1. **The Nuclear Fuel Cycle**
2. **Nuclear Power Plant Technologies**
3. **Nuclear Generation Costs**
4. **Price and Net Present Value**
5. **Nuclear Power Economic Uncertainties**

Glossary

capacity factor The ratio of actual to potential energy output of a generating plant.
cost of capital A weighted average of the interest rate charged by lenders and the rate of return on investment required by investors.
decommissioning The release of a nuclear power plant from jurisdiction of the safety regulator by removing radioactive materials, such as spent nuclear fuel and the reactor core; can also involve the decontamination of all hazardous materials and the dismantling of all structures.
discounting The process of evaluating money of one period in an earlier or later period, reflecting the "time value of money." A firm's discount rate is usually equal to its cost of capital.
enrichment The process of increasing the percentage of the isotope uranium-235 (only 0.7% abundance in nature) in mined uranium (for example, to 4% for use as nuclear fuel in light-water reactors).
interest during construction (IDC) The cost of borrowing funds from lenders and investors during construction. The IDC is compounded, so it grows exponentially during construction.
levelize Transforming a dollar amount in one period into an annuity (a series of equal annual payments) over many periods.
nuclear steam supply system (NSSS) The nuclear reactor plus equipment required to transform heat from the reactor into steam (in water-cooled reactors), to drive the turbine-generator. The NSSS and the turbine-generator make up a generating unit. One or more units constitute the nuclear power plant.
Nuclear Waste Trust Fund A U.S. government fund to finance the management and disposal of spent nuclear fuel; U.S. nuclear power plant owners contribute $1/megawatt-hour of electricity produced.
reprocessing The process of removing uranium and plutonium from spent nuclear fuel, either to reduce the volume for geologic disposal or to recycle in fresh nuclear fuel.

Nuclear power economics analyzes the costs and uncertainties of generating energy with radioactive fuel in a nuclear reactor at a nuclear power plant (NPP) and compares these costs and uncertainties with other forms of energy generation. Although NPPs have been used to produce heat and to desalinate water, and might be used to produce hydrogen, commercial NPPs produce electricity, measured in megawatt-hours. As in any enterprise, generating electricity with nuclear power involves fixed and variable costs and cost uncertainties, all of which are influenced by political realities, economic conditions and policies, and social concerns.

1. THE NUCLEAR FUEL CYCLE

The generation of nuclear energy begins with uranium mining and milling. These activities are at the "front end" of the nuclear fuel cycle (Fig. 1). The cycle continues through conversion and enrichment and fuel fabrication for light water reactors (natural uranium is fuel for heavy water reactors),with each of these operations having associated costs (see Section 3.3.1). The "back end" of the cycle includes spent nuclear fuel (SNF) management, reprocessing, and high-level radioactive waste disposal. There are two types of back ends: once-through and closed cycle. Under the once-through fuel cycle, SNF is cooled at the NPP, stored or transported to a centralized facility, and finally shipped to a geologic repository, such as the one being planned in the United States at Yucca

 © 2004 Elsevier Inc. All rights reserved.

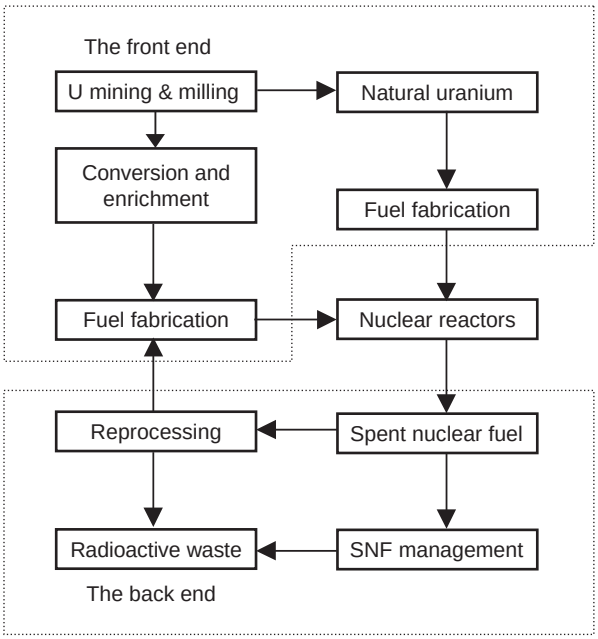
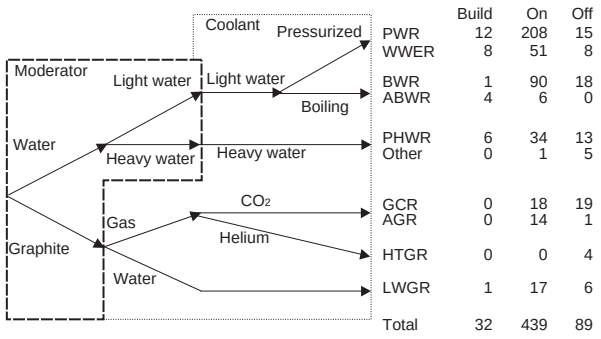


FIGURE 1 The nuclear fuel cycle. Converted and enriched uranium is used in fuel fabrication processes for light water reactors; natural uranium is used for heavy water reactors. SNF, Spent nuclear fuel.



	Build	On	Off
PWR	12	208	15
WWER	8	51	8
BWR	1	90	18
ABWR	4	6	0
PHWR	6	34	13
Other	0	1	5
GCR	0	18	19
AGR	0	14	1
HTGR	0	0	4
LWGR	1	17	6
Total	32	439	89

FIGURE 2 Commercial thermal reactor technologies. Key: Build, under construction; On, in operation; Off, retired; PWR, pressurized light water reactor; WWER, pressurized water electricity reactor; BWR, boiling light water reactor; PHWR, pressurized heavy water reactor; GCR, gas-cooled, graphite-moderated reactor; AGR, advanced GCR; HTGR, high-temperature, gas-cooled, graphite-moderated reactor; LWGR, light-water-cooled, graphite-moderated reactor. Data from IAEA (2002).

Mountain, Nevada. The closed-cycle back end involves reprocessing SNF to recycle uranium and plutonium for fuel. SNF can only be recycled in current NPPs a few times before the buildup of unstable plutonium isotopes renders the fuel unusable. Eventually, recycled SNF is separated into waste streams, some of which are processed for geologic disposition.

2. NUCLEAR POWER PLANT TECHNOLOGIES

The different types of nuclear power plant technologies are depicted in Fig. 2, which details the numbers of each type of technology (in 2002) under construction ("Build"), in operation ("On"), and permanently retired ("Off"). (Commercial fast reactors, also known as breeder reactors or liquid metal reactors, are not discussed here.) NPPs are generally classified by the moderator (to control the chain reaction) and the coolant (to remove heat). Water (light or heavy; heavy water is also known as deuterium oxide, D_2O) and graphite are used as moderators. Water (light or heavy) and gas (air, carbon dioxide, or helium) are used as coolants. Also, NPPs use either direct or indirect cycles to transfer energy to the turbine-generator to produce electricity. The direct approach allows the coolant to boil in the reactor core, creating steam to drive a turbine. The indirect approach transfers heat from a pressurized primary cooling circuit to a secondary circuit (e.g., through a steam generator) to drive the turbine.

Although many experimental reactors have combined these options, only four types of commercial thermal NPPs have been built:

1. Light water reactors (LWRs), including pressurized light-water-moderated and -cooled reactors (PWRs), Soviet/Russian pressurized water-moderated, water-cooled electricity reactors (WWERs), boiling light-water-moderated and -cooled reactors (BWRs), and advanced BWRs (ABWRs).
2. Heavy water reactors (HWRs), including pressurized heavy-water-moderated and -cooled reactors (PHWRs), such as the Canadian deuterium uranium (CANDU) reactor.
3. Gas-cooled, graphite-moderated reactors (GCRs), including early British and French GCRs and British advanced GCRs (AGRs) and high-temperature, gas-cooled, graphite-moderated reactors (HTGRs).
4. Soviet light-water-cooled, graphite-moderated reactors (LWGRs), such as the type of reactor at Chernobyl.

Because most of the currently operating NPPs are LWRs, the remainder of this article focuses on the cost of generating electricity from light water technology.

3. NUCLEAR GENERATION COSTS

Nuclear energy systems compete with other technologies in the electricity market (see Section 4). When evaluating whether to build NPPs, owner-operators consider (1) the total plant cost in millions of dollars ($M) (because capital is at risk while the plant is under construction and there are no revenues), (2) the time to build a plant (because financing costs accumulate while the plant is under construction and demand conditions can change), (3) the ratio of average fixed (capital) costs to the market price of electricity (because cash flow problems can arise with large payments to debt and equity holders), (4) the ratio of average variable (production) costs to electricity price (because this will drive dispatch under regulation or bidding strategies under deregulation), and (5) the difference between average total generation cost and electricity price (because this determines profitability and the net present value of the project). Here, a distinction is made between "fixed costs," which are sunk when the plant is completed, and "variable costs," which can be avoided if the plant were to cease production. Fixed costs (FCs) include construction costs, financing costs, and decommissioning costs. Variable costs (VCs) include fuel costs, operation and maintenance costs, and capital additions. (Throughout the following discussions, costs and prices are given in real dollars, deflated using the U.S. gross domestic product implicit price deflator, or inflated to future years assuming an appropriate inflation rate, such as 3%.)

3.1 Plant (Fixed) Costs

The electric utility industry distinguishes between base (overnight) construction cost and financing costs. Because electricity-generating plants cannot be built instantaneously, lenders require payments during construction. These payments are known as "interest during construction" (IDC). IDC is equal to base costs discounted (into the future) to the start of commercial operation. Total plant costs (PCs) are equal to base costs plus IDC. Estimated total plant costs are equal to base costs plus IDC plus contingency (discussed later). Fixed costs (or total capital costs) are equal to total plant costs plus discounted decommissioning costs.

Cost estimates also distinguish between those for the first commercial plant of a specific technology ("first-of-a-kind" plants) and later plants, known as "*n*th-of-a-kind" plants, First-of-a-kind costs include costs for research, development, demonstration, and deployment. These costs can be assigned to the first plant or amortized over a series of "transitional" plants. (However, some research, development, demonstration, and deployment costs could be subsidized with public funding, so these would not be amortized.) The *n*th-of-a-kind estimates also reflect learning during the construction of the first and transitional plants. The following discussion assumes "*n*th-of-a-kind" costs.

3.1.1 Base Construction Costs

Construction costs account for the largest portion of the cost of generating electricity from the current generation of NPPs. Cost-engineering estimates divide the construction project into a series of accounts. For example, Table I presents estimated total plant costs for a 1400-MW ABWR at a two-unit site in the United States. The accounts in this table are defined in the U.S. Department of Energy's *Energy Economic Data Base*. Direct costs include land, structures, and equipment purchased from vendors, including the nuclear steam supply system (NSSS) manufacturer. Indirect costs include construction and engineering services provided by the architect-engineer, construction manager (who can be the architect-engineer), and the owner-operator (who can be both architect-engineer and construction manager). Direct plus indirect costs equal the base construction cost.

Also included in estimated total plant costs is "contingency," which accounts for uncertainty in the cost estimate, usually as a percentage of estimated total plant costs. For example, for a simplified estimate, contingency is 30–50% of base construction cost; for a preliminary estimate, it is 15–30%; for a detailed estimate, it is 10–20%; and for a finalized estimate, it is 5–10%. Contingency can also be defined in association with the probability distribution of base construction cost. Under this interpretation, contingency is a function of the accuracy of the estimate, the confidence level of the estimate, the variance of the estimate, and the risk aversion of the estimate evaluator.

3.1.2 Financing Costs: Construction Duration and Interest During Construction

Translating base construction costs into total plant costs depends on spending in each period t during construction (CX_t, construction expenditures); construction duration, known as the lead time (LT), in years); and the cost of capital (r). A simple function for interest during construction is given by the

TABLE I

Example of a Nuclear Plant Cost Breakdown[a]

Account	Cost	Amount
	Direct costs	
21	Structures and improvements	$430
22	Reactor plant	$520
23	Turbine plant	$230
24	Electrical plant	$150
25	Miscellaneous plant	$45
26	Main heat rejection system	$45
	Total direct costs	$1420
	Indirect costs	
91	Construction services	$250
92	Engineering home office	$70
93	Field office services	$190
94	Owner's cost	$200
	Total indirect costs	$710
	Base (overnight) construction cost	$2130
	Contingency	$125
	Interest during construction at 5% (or 10%) (5-year lead time)	$282 ($564)
	One-unit cost at 5% (or 10%)[c]	$2537 ($2819)
	Plant cost at 5% (or 10%) per kilowatt (gross)[b]	$1812 ($2013)

[a] Total plant cost for a 1400-MW advanced boiling water reactor at a two-unit site built in the United States (in millions of 1998 dollars).

[b] Kilowatt (gross) does not account for electricity consumed by the plant; a 1400-MW plant typically produces 1350 MW (net).

[c] Based on data from the Nuclear Energy Agency's *Reduction of Capital Costs of Nuclear Power Plants*.

following equation,

$$\mathrm{IDC} = \sum_{t=\mathrm{LT}}^{1} \mathrm{CX}_t \left[(1 + r_t)^{(t-0.5)} - 1 \right] \qquad (1)$$

assuming expenditures are made in the middle of the period. This formula discounts construction expenditures to the start of operation. This formula can be simplified by assuming that the cost of capital is equal in each period, continuous discounting, and a function for the expenditure rate, i.e., the percentage of the base cost spent in each period. The cost of capital is generally taken as a weighted average cost of capital (debt and equity), which can be adjusted for taxes, although taxes are not discussed here. The expenditure rate can be assumed to follow a probability density function, such as the uniform or normal. Under a uniform expenditure rate with continuous compounding and constant cost of capital, the IDC can be approximated as $r \times \mathrm{LT}/2$; e.g., with $r = 10\%$ and $\mathrm{LT} = 5$ years, $\mathrm{IDC} \approx 25\%$. (If LT is expressed in months, r must be converted to the monthly cost of capital.) Of course, if the construction expenditures and the cost of capital are known *ex post*, the IDC can be calculated precisely.

This analysis shows that the cost of financing depends heavily on the length of the construction period. The start of construction can be either the date of the issuance of the construction permit or the date of the first pouring of concrete. The difference between these two can involve months of planning and site preparation. Although site preparation can be lengthy, construction expenditures and IDC are low before the pouring of concrete. The end of construction is either when the plant is connected to the transmission grid ("grid connection") or when plant operation is transferred from the builder to the commercial operator ("commercial operation") and generation near rated capacity begins. The difference between these dates can involve testing, including low-power testing and power-ascension tests; regulatory requirements, such as the completion of a final safety analysis report; and the issuance of an operating permit. This difference can be less than a month to over a year. The commercial operation date is more appropriate for the calculation of IDC. Therefore, LT in Eq. (1) is usually measured as the time from the first concrete pour to commercial operation.

Construction time increased dramatically in the United States after the accident at Three Mile Island in 1979 due to changes in safety requirements for plants still under construction. In the United States, construction time from the first pouring of concrete to the connection of the unit to the grid was 4.8 years for units completed between 1966 and 1971, 6.1 years for units completed between 1972 and 1977, 9.4 years for units completed between 1978 and 1983, and 12 years for units completed between 1984 and 1989. On the other hand, in Japan, construction times were 4 years for units completed between 1972 and 1977, 4.6 years for units completed between 1978 and 1983, 3.9 years for units completed between 1984 and 1989, and 4.6 years for units completed between 1990 and 1995. Therefore, based on the Japanese experience (where the ABWR has been built), the nuclear power

industry assumes that new (large) NPPs can be built in 4 to 5 years.

3.1.3 Decommissioning Costs

Although decommissioning costs are not spent until after the NPP has been retired, once nuclear fuel has been loaded into the reactor and the reactor experiences a self-sustaining nuclear reaction, regulatory authorities, such as the Nuclear Regulatory Commission (NRC), require that the owner-operator set aside funds to pay for the decommissioning. Officially, decommissioning refers to the release of the NPP from regulatory jurisdiction. In the United States, this means the removal of the reactor vessel, the spent nuclear fuel, and any radioactive contamination under NRC regulation. In practice, there are various levels of decontamination and dismantling after NPP retirement. These are defined by other regulators (in particular, national and state environmental regulators) or by laws defining liability for residual radioactive, hazardous, or mixed (radioactive and hazardous) contamination. These levels of decontamination can be defined as follows:

1. Black field (also known as "safe storage"). NRC regulations have not been completely satisfied; for example, spent nuclear fuel has been placed in storage on site, but the reactor has not been dismantled. The site can be used for other nuclear activities.
2. Gray field (includes the U.S. Environmental Protection Agency Superfund sites). NRC regulations have been satisfied, but hazardous waste remains; ownership transfer is therefore difficult, but the same owner can use the site for other electricity-generating activities.
3. Brown field. The site has been decontaminated and can be sold for restricted (e.g., industrial) use, but the original owner remains liable for any residual (possibly unknown) contamination.
4. Green field. There is no contamination; the site can be used for any activity.

Because of these different levels of cleanup and because the requirements for certification of decontamination vary from jurisdiction to jurisdiction (country to country and state to state), the cost of decommissioning varies widely. Decommissioning costs have been estimated from experience at completed decommissioning projects, with the application of unit-cost factors or with a cost-engineering model. The NRC relies on cost-engineering models to specify the minimum amount that should be set aside by the owner during operation for eventual decommissioning, including costs during "safe storage." These minimum amounts are $105 million (1986 dollars) for a PWR and $135 million (1986 dollars) for a BWR.

Since 1990, U.S. owners have been required to establish a nuclear decommissioning trust (NDT) fund, outside the financial control of the owner. These funds have been collected from ratepayers (like other capital expenses) and are invested at rates approximated by the interest rate on low-risk corporate bonds (e.g., an interest rate of 5%). However, the NRC does not require NDT funds to cover "the cost of removal and disposal of spent fuel or of nonradioactive structures and materials beyond that necessary to terminate the license" [U.S. Code of Federal Regulations (CFR), parts 50–75]. To account for these costs, state regulatory commissions have required higher amounts (collected from ratepayers) in the NDT than are required by the NRC.

Because of the wide variety of cost estimates, the cost of decommissioning is assumed to be one-third of the direct construction cost, discounted 40 years to the start of operations. For example, at $1420/kW (see Table I, total direct costs), decommissioning would be $473/kW. Discounted at 5% for 40 years to the start of commercial operation, the present value would be $67/kW or $94 million for a 1400-MW NPP. Thus, decommissioning costs represent the present value of decommissioning (decontamination and dismantlement) costs at the time of commercial operation.

3.1.4 Plant Costs: Empirical Evidence and Projections

The econometric analysis of U.S. NPP plant costs began in the mid-1970s. This literature focused on the issue of whether there were economies of scale in NPP generating capacity, controlling for the year of commercial operation, construction duration, and plant characteristics, such as location, number of units at the site, and NSSS manufacturer. Although early analyses found increasing returns to scale, analyses of later data could not reject constant returns to scale, primarily due to the imprecision of the estimate of scale parameters when large, expensive units came into production after the accident at Three Mile Island in 1979. For example, plant costs at Watts Bar, the last unit to attain commercial operation in the United States, were nearly $7 billion (in mixed-nominal dollars), or about $6000/kW with a construction duration of 24 years from 1972 to 1996.

Estimated costs for advanced light water reactors, based on more recent experience (primarily in Asia), are much lower. In NEA (1998), plants are either commercially available or expected to be commercially available between 2005 and 2010. Base construction cost varies between $1020 in China (using Chinese technology) and $2521 in Japan (for an ABWR).

3.2 Productivity and Average Fixed Costs

To understand fixed costs, total fixed costs can be converted into a cost per megawatt-hour (MWh) by determining an annual, uniform series of payments equal to the present value of total fixed costs at the time of commercial operation and then dividing this annual payment by the annual electricity output. Annual output is discussed first, then average fixed costs.

3.2.1 Productivity and the Capacity Factor

There are three measures of NPP performance. First, productivity refers to the ability of the NPP to produce electricity. In the nuclear power industry, productivity is usually measured by the annual capacity factor, CF:

$$\text{MWh} = \text{CF} \cdot \text{MAX} \cdot Y \tag{2a}$$

$$\text{CF} = \text{MWh}/(\text{MAX} \cdot Y), \tag{2b}$$

where MWh is annual output in megawatt-hours, MAX is the NPP's annual net maximum dependable capacity (net capacity equals gross capacity minus electricity consumed by the plant), and Y is the total hours in a year. For example, a plant with a net maximum dependable capacity of 1350 MW (equivalent to a gross capacity of 1400 MW) operating at a 90% capacity factor for 8760 hours/year generates about 10,600,000 MWh for sale per year. Second, availability measures the readiness of the NPP to generate electricity. It can be measured with time or energy units. Associated with the time definition is the service factor, SF, measured as the ratio of generating hours (H) to total hours in a year: H/Y. Third, reliability measures the certainty of NPP generation. The reliability factor, RF, is equal to (1−FR), where FR is the forced outage rate, i.e., the percentage of time devoted to recovery from a forced outage.

To relate these performance indicators to the CF, consider the following equation:

$$\begin{aligned}\text{CF} &= [\text{MWh}/(\text{MAX} \cdot \text{H})] \cdot (\text{H}/\text{Y}) \\ &= \text{CU} \cdot \text{SF}.\end{aligned} \tag{3}$$

The first term, $\text{MWh}/(\text{MAX} \cdot H)$, is the capacity utilization rate, CU, a measure of how close the NPP is to potential output when it is running. The second term, H/Y, is the service factor (SF), the percentage of the time the NPP is running. So, the capacity factor is equal to the product of capacity utilization and the service factor. Further, the service factor is equal to the product of the "scheduled availability" rate, $\text{SA} = [(Y–S)/Y]$, where S is the scheduled outage time, and the reliability factor, $\text{RF} = H/[(H+F)]$, where F is the forced outage time and $Y = H + S + F$.

This decomposition of the capacity factor into the capacity utilization rate and the service factor (and the decomposition of the service factor into the scheduled availability rate and the reliability factor) allows a detailed analysis of NPP productivity. For example, Table II presents changes in CF, CU, SF, SA, and RF for NPPs in the United States (although the decomposition holds for the performance of each unit, the average of a product is not equal to the product of the averages):

- The mean and median CF increased 2 percentage points per year, or almost 3% per year.
- Median CU increased to 99.9% in 2000.
- Median SF increased to 91.8% in 2000, i.e., an average outage time of about 1 month.
- The forced outage rate (1−RF) declined to a median value of 0.8% (=1–99.2%) in 2000, i.e., less than 3 days per year.

Most forecasts of capacity factors for new nuclear technologies assume a value equal to the best average LWR capacity factor. With median capacity factors now above 90%, forecasts for new technologies assume a lifetime average capacity factor of 90%.

3.2.2 Average Fixed Costs

To calculate average fixed costs (AFCs), total capital costs are multiplied by the capital recovery factor (CRF) to determine equal annual payments and divided by average annual output, equal to the product of the (average) capacity factor, the size of the plant, and number of hours in a year. Total fixed (capital) costs include total plant costs (PCs) plus decommissioning costs (DCs). The CRF is equal to $[r \cdot (1+r)^T]/[(1+r)^T–1]$, where r is the appropriate discount rate and T is the economic lifetime of the NPP. The appropriate discount rate is usually a weighted average cost of capital (debt and equity), consistent with the discount rate for calculating IDC. Also, the lifetime of the NPP is assumed to be the length of the operating license, which is 40 years in the United States. (However, NPPs can be closed early or their life can be extended.) If $r = 5\%$ and $T = 40$, $\text{CRF} = 5.8\%$. If $r = 10\%$ and $T = 40$, $\text{CRF} = 10.2\%$.

TABLE II

Changes in U.S. NPP Performance: 1990–2000[a]

	Capacity factor		Capacity utilization		Service factor		Scheduled availability		Reliability factor	
Year	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median
1990	67.7	70.0	90.8	94.9	73.0	75.8	75.6	78.2	96.5	98.2
1992	71.7	74.9	93.1	96.3	75.3	79.5	78.1	81.4	96.5	98.3
1994	75.5	81.4	92.9	97.2	79.7	83.3	83.1	86.6	95.7	99.0
1996	77.5	82.7	95.1	98.6	78.9	84.3	81.9	87.2	96.4	98.8
1998	80.9	87.7	94.6	99.3	81.2	88.8	83.6	89.7	98.2	99.5
2000	89.3	91.9	99.2	99.9	89.5	91.8	93.5	93.6	95.9	99.2
Increase	21.6	21.9	8.4	5.0	16.5	15.9	17.9	15.4	−0.6	1.0

[a] Values in percentages.

(Here, r is a real discount rate; the range of 5–10% is assumed to envelop most reasonable values for the average real cost of debt and equity.)

The average fixed costs for an ABWR built at a two-unit site in the United States can be calculated using the plant costs, decommissioning costs, and capital recovery costs as follows for the plant described in Table I [in millions (M) of U.S. dollars]:

$$\begin{aligned}\text{AFC} &= [(\text{PC} + \text{DC}) \cdot \text{CRF}]/\text{MWh}\\ \text{AFC} &= [(2537\text{M} + 94\text{M}) \cdot 0.058]/10,600,000\\ &= \$14.40 \quad \text{at } r = 5\%\\ \text{AFC} &= [(2819\text{M} + 94\text{M}) \cdot 0.102]/10,600,000\\ &= \$28.00 \quad \text{at } r = 10\%. \end{aligned} \tag{4}$$

Decommissioning costs are discounted at 5% to the start of commercial operation due to regulatory restrictions on NDT fund management; however, these costs are levelized at the average cost of capital, because this is the owner's opportunity cost. This example shows that the cost of capital has a great influence on the average fixed costs of NPPs.

3.3 Production (Variable) Costs

Three expenses that vary annually are fuel costs, operating and maintenance costs, and capital additions. Let average variable costs (AVCs) equal fuel costs per megawatt-hour (F) plus operating and maintenance (OM) costs per megawatt-hour plus capital additions per megawatt-hour (K):

$$\text{AVC} = F + \text{OM} + K. \tag{5}$$

Each of these costs is defined in the following section, with a discussion on how they have changed at U.S. NPPs during the 1990s.

3.3.1 Fuel Costs

Although fuel costs involve purchasing materials (e.g., uranium) and services (e.g., uranium enrichment and spent nuclear fuel management) over extended periods, fuel costs are treated as variable costs because they could be avoided if the NPP was not operating. As was seen in Fig. 1, there are three segments of the nuclear fuel cycle: the front end, in the reactor, and the back end. The front end of the nuclear fuel cycle involves uranium mining and milling, enrichment, and fuel fabrication. Uranium is primarily extracted through open-pit and underground mining. The uranium ore is crushed and leached with acids to produce a uranium oxide powder (U_3O_8), known as "yellowcake." Since the creation of a commercial market in the early 1970s, uranium production has depended on the market price of uranium. The price of uranium increased in the late 1970s, but has declined continuously since 1980 in the United States and much of the world to \$20/kg (see Fig. 3).

To increase the percentage of fissile uranium (natural uranium contains 0.7% of uranium-235, a radioactive isotope, and 99.3% uranium-238), uranium oxide is converted to uranium hexafluoride, UF_6 (a gas above 56°C). The cost of conversion has been stable for years at \$8/kg of uranium. Then the UF_6 is enriched to between 3 and 5% uranium-235 for light water nuclear reactor fuel. Enrichment is done commercially with two methods, diffusion and centrifuge. With diffusion technology, the uranium gas is pumped through a series of slightly porous barriers that allow the smaller uranium-235 to penetrate more easily than the larger uranium-238. Diffusion technology requires large amounts of electricity for pumping the UF_6. With centrifuge

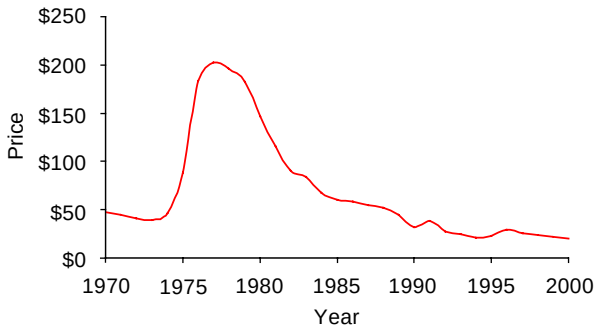


FIGURE 3 Uranium (U_3O_8) prices in kilograms/1996 U.S. dollars. Data for years 1970–1979 from NUEXCO (1979); for years 1981–2000, from Energy Information Administration (2000).

technology, the uranium gas is separated, with the lighter uranium-235 rising to the top when the UF_6 is spun. Enrichment is measured in "separative work units" (SWUs), i.e., the amount of effort required to increase the percentage of uranium-235. About 5 SWUs are needed to enrich 1 kg of uranium from 0.7 to 4% uranium-235. The SWU price has varied with changes in the price of electricity and changes in international competition. Between 1994 and 2003, SWU prices in the United States have varied between $80 and $110 (nominal dollars).

Enriched uranium is converted to uranium oxide and fabricated into nuclear reactor fuel. The cost of fuel fabrication varies between $200/kg and $400/kg of U_3O_8. Approximately 20 metric tons of uranium is consumed each year in a 1000-MW NPP. Given the cost of uranium and conversion, fuel costs are dominated by enrichment and fabrication.

The back end of the nuclear fuel cycle includes spent nuclear fuel management, reprocessing, and high-level radioactive waste disposal. The cost of each operation varies from country to country and with the value of the uranium and plutonium recovered during reprocessing. The following approximate costs prevail in the United States: (1) for storage of spent nuclear fuel (SNF) at a NPP (on site) in nontransportable casks, $250/kg; (2) for on-site storage, transportation, and storage at away-from-reactor facilities, $500/kg; (3) for all storage, transportation, encapsulation, and disposal in a geologic repository, $750/kg; and (4) for international commercial reprocessing (including transportation, recovery of uranium and plutonium, and disposal of radioactive waste), $1000/kg. United States policy calls for storage of SNF at NPPs until the construction of a repository. The repository is to be financed through the collection of $1/MWh, accumulated in the Nuclear Waste Trust Fund.

3.3.2 Operation and Maintenance Costs

Operating and maintenance expenses include labor and materials. Materials do not include equipment or other items with an economic life for more than 1 year. These items are capitalized and are known as "capital additions" (discussed later). Most of the OM expenses are fixed during the year, but they are treated as variable costs, because most could be avoided if the plant were to cease production. A majority of OM expenses pay NPP employees. There are four types of employees: administrative, operations, engineering, and maintenance, including security. These personnel work at the NPP or at corporate headquarters, for the NPP owner directly or for a subcontractor, and permanently or temporarily, e.g., during refueling outages.

3.3.3 Capital Additions

Capital additions, also known as refurbishments, include expenditures on equipment that lasts more than 1 year along with any labor that is involved in the installation or refurbishment of equipment. The largest capital additions at U.S. nuclear power plants have been the replacement of steam generators at PWRs. Capital additions are amortized over the remaining life of the plant, but because they could be avoided by retiring the plant, they are treated as a variable cost.

3.3.4 Production Costs: Empirical Evidence and Projections

Table III presents mean and median values for average variable costs at U.S. NPPs from 1990 to 2000: mean fuel expenses (F) declined from $8.51 to $4.59 and median fuel expenses declined from $7.52 to $4.40; mean OM costs declined from $19.73 to $11.47 and median OM costs declined from $17.04 to $10.47; and mean capital additions (K) declined from $5.59 to $3.00 and median capital additions declined from $4.12 to $1.27. These differences between means and medians show that unusually high costs at some NPPs have a large influence on mean variable costs.

To examine the influence of the changes in the capacity factor, CF, on AVCs, redefine the variables in Eq. (5), such that

$$\mathrm{AVC} = \underline{\mathrm{AVC}}/\mathrm{CF} = \underline{F}/\mathrm{CF} + \underline{OM}/CF + \underline{K}/CF, \quad (6)$$

where $\underline{\mathrm{AVC}}$ is the average variable cost at a 100% capacity factor and all other costs are similarly defined. For example, if AVC = $20/MWh and CF = 90%, $\underline{\mathrm{AVC}}$ = $18/MWh. Table IV presents the means and medians of $\underline{\mathrm{AVC}}$, $\underline{F}$, $\underline{\mathrm{OM}}$, and $\underline{K}$ from

TABLE III

Mean and Median Production Costs for U.S. NPPs: 1990–2000[a]

	Cost ($)		Fuel ($)		Operation/maintenance ($)		Capital additions ($)	
Year	Mean	Median	Mean	Median	Mean	Median	Mean	Median
1990	32.24	29.28	8.51	7.52	19.73	17.04	5.59	4.12
1992	31.58	27.94	7.00	6.22	19.27	16.05	6.23	4.58
1994	26.11	23.37	6.41	5.74	16.12	13.95	3.82	3.11
1996	26.59	20.31	5.70	5.27	17.55	12.38	3.11	2.27
1998	23.85	19.24	5.34	5.02	15.86	11.70	2.68	2.27
2000	19.06	17.97	4.59	4.40	11.47	10.47	3.00	1.27
Decrease	13.19	11.32	3.92	3.12	8.26	6.57	2.59	2.85

[a] All values in U.S. dollars.

TABLE IV

Mean and Median Production Costs at a 100% Capacity Factor for U.S. NPPs: 1990–2000[a]

	Cost ($)		Fuel ($)		Operation/maintenance ($)		Capital additions ($)	
Year	Mean	Median	Mean	Median	Mean	Median	Mean	Median
1990	22.50	22.12	5.99	5.39	12.90	12.14	3.90	3.09
1992	21.59	21.05	5.13	4.79	12.79	12.37	4.05	3.41
1994	19.57	19.35	4.98	4.72	11.70	11.08	2.83	2.44
1996	18.11	17.24	4.51	4.45	11.26	10.58	2.32	1.85
1998	18.06	16.76	4.39	4.31	11.42	10.77	2.04	1.85
2000	18.20	17.03	4.22	4.06	11.08	10.16	2.97	1.14
Decrease	4.30	5.09	1.77	1.33	1.82	1.98	0.93	1.95

[a] All values in U.S. dollars.

1990 to 2000. Holding the capacity factor constant, the decline in variable cost is much less dramatic: most of the decline was from an increase in the capacity factor (see Table II). For example, the decline in mean AVC was \$13.19, whereas the decline in mean AVC was only \$4.30, i.e., one-third of the decline in AVC was due to decreases in costs and two-thirds of the decline was due to increases in the capacity factor.

Although estimated AVC varies from country to country, costs per megawatt-hour are approximately \$5/MWh for fuel (including waste management and disposal costs), \$10/MWh for OM, and \$2/MWh for capital additions; i.e., near-term forecasts of AVC for both operating and new NPPs are approximately \$17/MWh. Therefore, the estimated average total costs (ATC = AFC + AVC) for new light water NPPs (following the preceding assumptions) vary between \$31.40 with a cost of capital of 5% and \$45.00 with a cost of capital of 10%.

4. PRICE AND NET PRESENT VALUE

The empirical evidence and near-term projections for total NPP plant cost, plant construction time, average fixed costs, average variable costs, and total average generation cost were examined in Section 3. In this section, these costs are compared with the price of electricity. One method for making this comparison indirectly is to calculate a levelized electricity generation cost and compare it with electricity generation prices. This is appropriate under rate-of-return regulation. Under competitive

market structures, a net present value calculation is more appropriate.

4.1 Levelized Cost

The levelized electricity generation cost (ECG) is defined by the following formula:

$$\mathrm{ECG} = \sum_t \left[C_t(1+r)^{-t}\right] / \sum_t \left[\mathrm{MWh}_t(1+r)^{-t}\right], \quad (7)$$

where the summation is over the period of construction, operation, and decommissioning, discounted to the start of commercial operation; C_t are all costs in the year t; MWh_t is electricity generation in the year t; and r is the discount rate, assumed to be equal to the cost of capital. During construction, C_t equals CX_t [see Eq. (1)]. These would be equal to base construction costs plus contingency (if an estimate) without IDC, given that discounting to commercial operation in Eq. (7) is equivalent to the IDC calculation. During operation, C_t equals AVC_t times MWh_t. During decommissioning, C_t equals DX_t, the decommissioning expenditures in year t. Finally, MWh_t is annual output during the years of operation.

Making the same assumptions as previously, for example, the following relationships apply:

- $\sum_t[C_t(1+r)^{-t}]$ during construction is equal to a total plant cost of \$2537 million for $r=5\%$ and \$2819 million for $r=10\%$, from Table I.
- $\sum_t[C_t(1+r)^{-t}]$ during operation is equal to \$17/MWh times 10,600,000 MWh, or \$180 million/year discounted over 40 years, or \$3100 million for $r=5\%$ and \$1760 million for $r=10\%$.
- $\sum_t[C_t(1+r)^{-t}]$ during decommissioning is equal to \$94 million, given restrictions on the discount rate for nuclear decommissioning trust funds.
- $\sum_t[\mathrm{MWh}_t(1+r)^{-t}]$ during operation is 10,600,000 MWh/year, discounted over 40 years, or 182,000,000 MWh for $r=5\%$ and 104,000,000 MWh for $r=10\%$.

Here, ECG = \$31.50 for $r=5\%$ and \$44.90 for $r=10\%$. These values are very close to those for the ATC (calculated previously), showing that the two techniques of levelization are equivalent.

4.2 Price Regulation and Pricing under Deregulation

Under traditional rate-of-return regulation, the price charged by an integrated electric utility would be determined by the quantity of electricity sold and the revenues required to generate electricity from its power plants and purchases made from other electricity generators. Under these conditions, the required revenues to operate the NPP would be equal to the levelized electricity generating cost.

In a competitive electricity market, the price in each period is usually determined by the bid of the highest cost producer. For example, the future price of electricity in Europe, Japan, and the United States will likely depend on the costs of generating electricity with natural gas. To approximate the price of electricity from natural gas, assume that the base construction costs of combined-cycle gas turbines (CCGTs) are \$500/kW throughout the world. With a construction time of 2 years, the IDC is approximately \$25/kW with $r=5\%$ and \$50/kW with $r=10\%$. With a 20-year capital recovery period, the CFR is 8% with $r=5\%$ and with 11.75% with $r=10\%$. The annual fixed cost is about \$42/kW for $r=5\%$ and \$65/kW for $r=10\%$. If the CCGT has a capacity factor of 90%, the average fixed cost is about \$5.30/MWh for $r=5\%$ and \$8.17/MWh for $r=10\%$. Second, the cost of natural gas per megawatt-hour depends on the heat rate. Assuming a heat rate of 7000 Btu/kWh, a natural gas price of \$4/MBtu yields an average fuel cost of \$28/MWh. Third, assuming OM plus capital additions are \$3/MWh yields a cost of generating electricity of \$36.30/MWh for $r=5\%$ and \$39.17/MWh for $r=10\%$. Therefore, with a price of natural gas at \$4/MBtu and with a cost of capital of 5%, nuclear power is cheaper than natural gas (\$31.50 $<$ \$36.30), but with a cost of capital of 10%, nuclear power is more expensive than natural gas (\$44.90 $>$ \$39.17). Of course, a complete analysis would consider all alternatives, including other nuclear technologies, coal, and renewables.

4.3 Net Present Value

Assuming natural gas sets the marginal price of electricity, a net present value analysis can determine the profitability of new nuclear power as a function of natural gas prices. Net present value (NPV) is defined as follows:

$$\mathrm{NPV} = \sum_t [(\mathrm{MWh}_t \cdot P_t) - C_t](1+r)^{-t} \quad (8a)$$

or

$$\mathrm{NPV} = \sum_t [\mathrm{MWh}_t \cdot P_t](1+r)^{-t} - \sum_t C_t(1+r)^{-t}, \quad (8b)$$

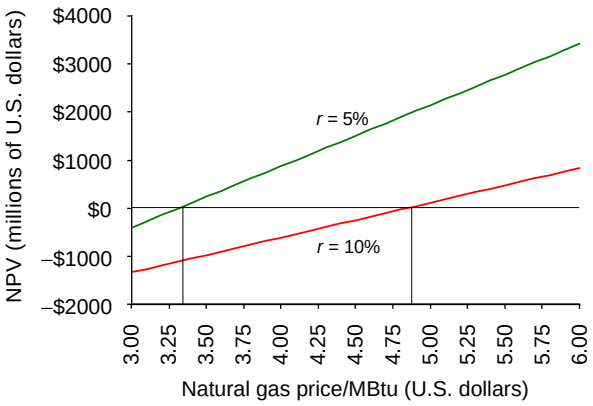


FIGURE 4 Net present value (NPV) of nuclear power as a function of the price of natural gas.

where P_t is the price of electricity and all other terms are the same as in Eq. (7). Notice that the last part of Eq. (8b) is identical to the numerator in Eq. (7). The first part of Eq. (8b) is the sum of discounted revenues. If P_t depends on the price of natural gas, NPV becomes a function of natural gas prices, holding all other variables constant (see Fig. 4). Nuclear power (under the ssumptions in Section 3) is competitive for natural gas prices above \$3.30/MBtu with a cost of capital of 5% and for prices above \$4.85/MBtu with a cost of capital of 10%.

5. NUCLEAR POWER ECONOMIC UNCERTAINTIES

If electricity generators were concerned only with cost and price estimates, more nuclear power plants would likely be built in regions of the world with high fuel costs. But electric utilities are also interested in the uncertainties associated with nuclear power economics. For example, compare the total plant cost of a NPP with a CCGT. The NPP requires an investment of \$2500 million to \$2800 million for one 1400-MW unit (under the preceding sample assumptions). If two units are required to achieve optimal scale economies at a single site, the utility faces an investment of at least \$5 billion. The CCGT can be built in units of 300 MW, requiring an investment of \$150 million. Risk-averse utilities might prefer a smaller investment even if the average total costs of the CCGT are higher. Further, the CCGT can be built in 2 years, whereas the NPP requires 5 years. During construction, interest payments to lenders accumulate exponentially. Uncertainties in the demand for electricity can also lead to larger risk premiums on capital invested in projects with long construction durations. Both effects favor technologies that can be built quickly.

Average fixed costs depend on the capacity factor. Though CCGTs are unlikely to operate at 90% capacity factors, their fixed costs are low compared to their variable costs. If NPPs are unable to achieve high capacity factors, average fixed costs increase. For example, if capacity factors at new NPPs were 70% (as they were in 1990 in the United States), average fixed costs would be \$18.40 with $r = 5\%$ and \$35.90 with $r = 10\%$. Nuclear power would not be competitive at capacity factors below 70%.

Average variable costs include fuel, OM, and capital additions. Although there are great uncertainties associated with long-term use of carbon-based fossil fuels, there are still uncertainties associated with the back end of the nuclear fuel cycle. Though most of these are political (and not economic) uncertainties, it is difficult to predict the cost of long-term geologic disposal. Also, most OM expenses are associated with labor costs. In the United States, regulation drove the increase in OM costs after the accident at Three Mile Island by requiring more employees at NPPs. Therefore, regulatory uncertainties, outside the control of the NPP owner, increase uncertainty in OM costs.

All of these uncertainties influence perception of total and average NPP generation costs by electric utilities, policymakers, and the public. The growth of commercial nuclear power depends on reducing NPP capital costs and resolving uncertainties associated with nuclear power economics. Simultaneously, reducing capital costs and uncertainties depends on new orders for NPPs. Given this dynamic interaction, market forces could lead to either the collapse or the self-sustaining expansion of the commercial nuclear power industry. This industry must change. It will be unable to do business as it has been doing under electricity deregulation and liberalization.

SEE ALSO THE FOLLOWING ARTICLES

Nuclear Engineering • Nuclear Fission Reactors: Boiling Water and Pressurized Water Reactors • Nuclear Fuel: Design and Fabrication • Nuclear Fuel Reprocessing • Nuclear Fusion Reactors • Nuclear Power, History of • Nuclear Power Plants, Decommissioning of • Nuclear Power: Risk Analysis •

Nuclear Proliferation and Diversion • Nuclear Waste • Public Reaction to Nuclear Power Siting and Disposal

Further Reading

Electric Power Research Institute (EPRI). (1997). "Technology Assessment Guide." EPRI, Palo Alto, California.

Energy Information Administration (EIA). (2000). Annual Energy Outlook. EIA, Washington, D.C.

Energy Information Administration (EIA). (2001). "The Electricity Market Module of the National Energy Modeling System: Model Documentation Report." EIA, Washington, D.C.

International Atomic Energy Agency (IAEA). (1999). "Evaluating and Improving Nuclear Power Plant Operating Performance." IAEA, Vienna.

International Atomic Energy Agency (IAEA). (2002). "Nuclear Power Reactors in the World." IAEA, Vienna.

International Atomic Energy Agency (IAEA). (2002). "Power Reactor Information System (PRIS) Database." IAEA, Vienna.

Nuclear Energy Agency. (1994). "The Economics of the Nuclear Fuel Cycle." Organization for Economic Cooperation and Development, Paris.

Nuclear Energy Agency. (1998). "Projected Costs of Generating Electricity: Update 1998." Organization for Economic Cooperation and Development, Paris.

Nuclear Energy Agency. (2000). "Uranium 1999: Resources, Production, and Demand." Organization for Economic Cooperation and Development, Paris.

Nuclear Energy Agency. (2000). "Reduction of Capital Costs of Nuclear Power Plants." Organization for Economic Cooperation and Development, Paris.

Nuclear Exchange Corporation (NUEXCO). (1979). "Monthly Report to the Nuclear Industry No. 129." NUEXCO, Menlo Park, California.

Oak Ridge National Laboratory (ORNL). (1993). "Cost Estimate Guidelines for Advanced Nuclear Power Technologies." ORNL, Oak Ridge, Tennessee.

Pasqualetti, M.,Rothwell, G. (eds.). (1991). Special issue: Nuclear decommissioning economics. *Energy J.* **12**.

Rothwell, G.(ed.). (1998). Special issue on nuclear energy in Asia. *Pacific Asian J. Energy* **8**(1).

Rothwell, G., and Gomez, T. (2003). "Electricity Economics: Regulation and Deregulation." IEEE Press with John Wiley, Piscataway, New Jersey.

Nuclear Power, History of

ROBERT J. DUFFY
Colorado State University
Fort Collins, Colorado, United States

1. **Dawning of the Nuclear Age in the United States, 1946–1954**
2. **Creating a Commercial Atom, 1954–1965**
3. **Nuclear Power Comes under Attack**
4. **Regulatory Failure and the Changing Economics of Nuclear Power**
5. **Nuclear Power in the 1980s**
6. **Nuclear Power Today**

Glossary

Atomic Energy Commission The powerful executive branch agency created in 1946 to oversee the U.S. atomic energy program; abolished in 1975.

atoms for peace The U.S. program launched in 1954 to share atomic information with European allies and to develop peaceful applications of atomic energy.

Joint Committee on Atomic Energy An unusually powerful congressional committee that dominated nuclear energy policy in the United States from 1946 to 1977.

Nuclear Regulatory Commission The agency created in 1975 to assume the regulatory duties of the Atomic Energy Commission.

Price–Anderson Act Legislation enacted in the United States in 1957 to encourage private sector interest in nuclear power by providing indemnification to the nuclear power industry in the event of a reactor accident. The law has been renewed several times over the program's history.

subgovernment Also known as policy monopolies or subsystems, these are small, stable groups of actors, both public and private, that dominate policy in specific issue areas.

The 50-year history of commercial nuclear power has been punctuated by dramatic policy changes. The first 20 years, marked by limited public participation, tight government control, and promises of clean, abundant energy, were followed by a period of intense social and political conflict over the technology's environmental and safety implications. Nuclear policy in the United States and most European nations shifted from all-out support to a more ambivalent posture, which led to a dramatic slowdown in the construction of new plants. Although nuclear plants continue to be built, public opposition and high costs are obstacles to a large-scale comeback.

1. DAWNING OF THE NUCLEAR AGE IN THE UNITED STATES, 1946–1954

1.1 The Military Roots of Commercial Nuclear Power

From its birth in the highly secretive Manhattan Project during World War II, atomic energy was defined and perceived in military terms, which meant that information regarding the atom was subject to elaborate security precautions. This concern with maintaining the "secret" of the atom set the tone for the atomic program far into the future. From the beginning, both public involvement in the development of atomic energy programs and access to information about atomic energy would be severely restricted.

When the war ended, the sense of national emergency that had driven the atomic program dissipated, and many of the top scientists who had worked on the bomb project returned to their positions in universities and the private sector. With virtually no stockpile of atomic weapons at the war's end, government officials worried that if the atomic program ground to a halt, the nation's security would be at risk. A vigorous weapons program, it was believed, was essential to the nation's defense and security, especially because it was widely recognized that the American monopoly on atomic weapons would not last forever.

 © 2004 Elsevier Inc. All rights reserved.

1.2 The Atomic Energy Act of 1946

In pursuit of these general goals, Congress in 1946 passed the Atomic Energy Act, which established the institutional framework within which atomic energy decisions would be made for approximately the next 30 years. The legislation created two bodies, the Atomic Energy Commission (AEC), which inherited ownership of all atomic materials, facilities, and information from the Manhattan Project, and the Joint Committee on Atomic Energy, which was to oversee AEC operations.

The AEC was to be the primary actor in formulating the nation's atomic energy policies. Its principal duties, as defined by Congress, were the production of atomic weapons and the fissionable materials required for their manufacture. In addition to establishing the AEC, the Atomic Energy Act mandated a government monopoly in atomic energy. Congress stipulated that responsibility for all atomic programs and information be transferred from the army to the AEC. Furthermore, the AEC was granted ownership of all fissionable materials and all facilities that used or produced them. The size and scope of the AEC's operations, together with the sweeping powers granted by Congress, clearly indicated that it was intended to be a powerful agency.

Composed of nine members from each chamber, the Joint Committee on Atomic Energy (JCAE) was clearly intended to be a powerful committee. It was, for example, the only joint committee created by statute rather than by the rules of both houses. The JCAE was also the only permanent joint committee granted the full legislative and investigative powers of a regular standing committee. Its formal status alone suggests that Congress intended to exercise control over atomic energy in general and over the AEC in particular. The Atomic Energy Act required the AEC to keep the JCAE "fully and currently informed with respect to the Commission's actions," mandated that all bills concerning the atom were to be referred to the JCAE, and it set no limits on the number of staff members or consultants the committee could hire. These sweeping and unusual grants of power, along with the JCAE's status as a special joint committee with exclusive jurisdiction over a glamorous and highly technical issue, made the committee one of the most powerful congressional bodies in the nation's history.

The normal tendency of Congress to defer to the advice of its specialized committees and subcommittees was reinforced in the case of atomic energy by several additional factors. First, atomic energy was largely perceived and justified as a defense program, with primary emphasis on its military applications. As late as June 1961, approximately two-thirds of all research and development and construction activities of the AEC's reactor program continued to serve weapons and military applications. The definition and perception of the atomic program as primarily military in character lent credibility to both the program and the Joint Committee and legitimized its exclusive policymaking role.

1.3 Secrecy and Security

Because the United States was the only nation with the atomic "secret," security was considered to be an essential part of the effort to retain the atomic monopoly. The JCAE's inclination, like that of other security and defense committees, was to keep most information classified. With its exclusive control over atomic energy matters in Congress, the committee was able to dominate the legislative aspect of policymaking in these early years. The security restrictions meant that the atomic program was obscured by a blanket of secrecy that effectively limited the number of participants in the atomic program. Nuclear power would be relatively immune to challenge so long as access to information concerning the program was so closely held.

1.4 The Creation of an Atomic Subgovernment

In many respects, the atomic energy program in the 20 years following World War II was the archetypal subgovernment. A small, cohesive, and stable group of actors consisting of the Atomic Energy Commission, the Joint Committee on Atomic Energy, the nuclear power industry, and the scientists and engineers working on the program in universities and national laboratories exercised considerable autonomy in policymaking. Each of the participants involved in nuclear policymaking during this period had either an economic, political, or organizational stake in the development and use of atomic energy. There is, however, one highly unusual aspect of this policy community. Unlike the familiar example of subsystems in which government actors had been "captured" by the very business interests they were supposed to monitor, with nuclear power there was no industry "client" group until the government created one. It was only after policymakers in the mid-1950s took steps to overcome the serious technological, economic, and political obstacles confronting the generation of electricity from fission

that a commercial nuclear power industry emerged. This unusual government–industry partnership is crucial to understanding the history of nuclear power in the United States, as well as in most other nations.

The atomic energy subgovernment was endowed with additional prestige and power because of the program's identification with national security issues. The actors in this tightly knit monopoly were united by the conviction that the development of atomic energy, first as a weapon but later as a means of generating electricity, was both necessary and desirable for the nation's welfare. Indeed, this commitment was enthusiastically embraced by those involved in formulating and implementing policy and drove the program for over 20 years. For the most part, this consensus meant that only supporters of nuclear power would be mobilized for political action. Presidents rarely became involved in the program, Congress readily deferred to the advice of the experts on the Joint Committee on Atomic Energy, and, with the exception of the nuclear power industry, interest group activity was virtually nonexistent. This one-sided mobilization ensured that generous subsidies and the absence of political conflict marked government policy during this long period.

1.5 Nuclear Power and the National Interest

In the years following the World War II, nuclear power was equated with the national interest, and it was often said that politics and atomic energy did not mix. The cold war, the prevailing faith in science, and the recognition that the atomic monopoly could not last forever all contributed to a consensus that atomic energy was instrumental in securing the nation's future. To some, the term "national interest" could be used to promote almost anything pertaining to nuclear power. Furthermore, it was often claimed that nuclear energy would revolutionize the nature of industrial society by providing a relatively cheap and abundant source of power. Nuclear energy could lead to the discovery of new production techniques, it was said, and provide cheap electricity to areas where supplies were scarce. Others claimed that nuclear power could drastically reduce the costs of production, leading to increased economic growth and an improved standard of living. In short, a vigorous atomic energy program was thought to be necessary for a nation seeking to exert military, economic, and technological leadership in the postwar era. The national interest seemed to demand that the United States explore the atom's many possibilities.

2. CREATING A COMMERCIAL ATOM, 1954–1965

The dawn of the 1950s brought a subtle but significant shift in the nation's atomic energy plans. The principal focus remained on the atom's military aspects, but government officials recognized that the so-called peaceful uses of the atom could also serve a variety of important objectives in both domestic and foreign policy. Although the cold war dictated that the program's primary purpose would continue to be weapons production, the Joint Committee began to press the AEC to put greater emphasis on developing and promoting the peaceful uses of atomic energy. At the time, the most likely prospect for peaceful applications was the production of electricity from atomic fission.

2.1 The Peaceful Atom and Postwar Politics

Despite this subtle shift in the program's avowed goals, the program was still largely justified by national security arguments. Aside from enhancing the nation's scientific reputation and providing an abundant energy source, finding a constructive use for the atom would also further the nation's postwar geopolitical aims. The explosion of the Soviet atomic bomb in 1949 had punctuated the end of the American nuclear monopoly. American officials worried that the Soviets would try to develop the peaceful atom before the United States in an attempt to reap propaganda and political benefits. This prospect was not a pleasant one for those responsible for managing the atomic program.

Although American officials were preoccupied with the Soviets, they were also concerned that other nations would seek to develop their own atomic capabilities. The events of 1949 had shown that the United States no longer had a monopoly on the atom, and that the knowledge and ability to build atomic weapons would inevitably spread to other nations. It was commonly accepted that it was only a matter of time before Great Britain and France would join the atomic club. Other nations would surely seek to do the same and launch their own atomic programs, whether the United States wanted them to or not. Faced with this unpleasant reality, the United States wanted to be in a position to control nuclear proliferation to ensure that it was compatible with American perceptions of world order. Government decision makers recognized that in order to avoid

losing influence around the globe, the United States would have to assert its leadership and share some atomic information, and secrets, with other nations. With these objectives in mind, in a speech to the United Nations in December 1953, President Eisenhower launched an initiative that came to be known as "Atoms for Peace." The program was clearly an effort to distinguish between the peaceful and military applications of the atom and to identify the United States with the peaceful applications. The Atoms for Peace plan also served a number of American foreign policy goals. The Atoms for Peace proposal, according to two researchers, was "in fact nothing less than a coherent global strategy for protecting Western Europe from Soviet domination." If the United States could develop the atom as a source of electric power, it was argued, it would be a boon to Western European industries and economies while helping to contain the Soviets.

The Atoms for Peace proposal had the added attraction of assisting American industry by providing access to potentially lucrative markets. According to the terms of the bilateral agreements signed pursuant to the program, European nations contracted with American firms for all atomic technologies and materials. The attraction of these agreements, as one official from the American contractor Babcock and Wilcox wrote to the joint committee, was that they were a "means of securing a good and early foothold" in the potentially lucrative European market for American reactor vendors. One 1955 study estimated that Europe represented a $30 billion market.

2.2 The Atomic Energy Act of 1954

In order to make more progress in developing the peaceful atom, Congress revised the Atomic Energy Act to permit greater access to nuclear information and materials. Although American firms were expressing an interest in atomic power, the calls for revising the 1946 law came primarily from the U.S. government and not from American industry, which was unprepared to assume the full responsibility for nuclear development because of the financial risks involved. In contrast to coal plants, the initial costs for nuclear plants would undoubtedly be very high, profits uncertain, and the technology new and potentially hazardous. In addition, projected demand for electricity in the early 1950s provided utilities with little incentive to build additional generating capacity. Furthermore, there were no reliable estimates of the possible damages stemming from a serious reactor accident. Nevertheless, development of the peaceful atom was assumed to be vitally important to addressing a number of problems confronting the United States at the time. If the private sector would not take the necessary steps to solve these problems, the U.S. government would. The solution to the government's problems was the Atomic Energy Act of 1954, which effectively created both commercial nuclear power and a commercial nuclear power industry. The development of commercial nuclear power in the United States, in short, emerged from an effort initiated by government rather than by private industry.

To open the door to greater industrial and international participation, Congress in 1954 loosened the restrictions and access to previously classified information, permitted the private ownership of nuclear facilities and allowed the private use of fissionable materials through a system of AEC licensing, liberalized patent laws in the atomic energy field, and offered a number of economic incentives to private industry.

2.3 The AEC's Conflict of Interest

The law also made the AEC responsible for developing, promoting, and regulating nuclear power. From the outset, this statutory conflict of interest would be problematic; it became clear that the AEC would consistently emphasize its promotional and developmental responsibilities at the expense of its regulatory duties. Congress, through the Atomic Energy Act of 1954, offered little guidance on how to resolve this dilemma. The law, for example, directed the AEC to "protect the health and safety of the public," but the statute offered little insight into how to achieve that broadly defined goal. Although such an expansive grant of authority was not unique, it would later become a problem, as the atomic program moved into its commercial phase.

2.4 Subsidies

Although the Atomic Energy Act of 1954 provided the AEC with little guidance on establishing regulations or defining safety goals, it did make clear that Congress did not want the AEC to impose burdensome regulations during the early stages of the development process. The section of the law under which all of the reactors in the AEC's Power Reactor Demonstration Program were licensed instructed the AEC to "impose the minimum amount of such regulations and terms of license as will permit the

Commission to fulfill its obligations under this Act to promote the common defense and security and to protect the health and safety of the public." In addition, members of the Joint Committee consistently pressed the AEC to move ahead with reactor development. Because reactors were not economically competitive with fossil fuel plants, however, the nation's electric utilities did not have the normal incentives to build them. As a result, the AEC had to offer a wide array of incentives, including ample financial assistance, to the nuclear industry. The AEC thus embarked on several programs designed to encourage industry participation in the commercial power program. The AEC's Power Reactor Demonstration Program, for example, assisted firms by waiving for 7 years the charges for the loan of source and special materials; by performing in its laboratories, free of charge, certain research and development work; and by agreeing to purchase technical and economic data from the participants under fixed-sum contracts.

Industry's initial response to these inducements was tepid, so the government resorted to a mix of threats and incentives to induce interest. For example, even though the AEC retained ownership of all fissionable materials, through a system of AEC licensing the Atomic Energy Act of 1954 made these materials available to firms, including both the fuel for nuclear reactors and the fissionable materials produced by nuclear reactors. Because there was still a shortage of fissionable materials for atomic weapons, the AEC agreed to purchase all of the plutonium produced in private reactors at a guaranteed price that provided firms with handsome returns.

All told, by 1962, the federal government had spent over $1.2 billion to develop reactor technology; this figure was double the amount spent by private business. One Department of Energy study conducted in the 1980s concluded that without federal subsidies nuclear electricity would have been 50% more expensive. These calculations did not include federal tax write-offs for utilities nor did they include money spent on programs with possible military implications.

2.5 The Price–Anderson Act

Even with these subsidies, the private sector was reluctant to assume responsibility for the risks involved with nuclear power. The underlying reason for this reluctance was the fear that they would be held liable for any and all damages resulting from a reactor accident. Although the AEC had an impressive safety record with its reactors, the technology was new and largely untested, and because the potential damages from an accident were unknown, the reactor vendors and electric utilities were unwilling to make a full-scale commitment to nuclear power. The industry sought, and soon received, government protection from liability claims before proceeding with any plans to build reactors. The problem was that the potential liability from a reactor accident was so large that it was unlikely that either nuclear vendors or utilities would be able to obtain the necessary insurance from private insurers. The fear of financial disaster led the industry to seek legislation limiting their liability in the event of an accident. Because the potential damages from an accident could be so high, the industry argued that indemnification, in which the government would give general protection to the atomic industry against uninsured liability to the public, was needed to allow development to proceed. Private insurers agreed to provide $60 million of protection for each nuclear facility, which was far more than had been made available to any other industry. Because it was entirely possible that damages from a reactor accident would far exceed that amount, Congress agreed to provide an additional $500 million of protection for the industry.

2.6 The First Generation of Nuclear Reactors

After adoption of the Price–Anderson Act, private sector interest in nuclear power slowly picked up. Although the liability limit eased industry fears, industry participation remained at low levels for the next few years because nuclear electricity was still much more expensive to generate than electricity produced by conventional fuels like oil and coal. However, the words and actions of government officials convinced the nuclear industry that in the future there would be nuclear power in America, with or without private sector involvement. A number of utilities decided to buy reactors in order to gain an advantage for the future; others purchased reactors because they wanted to keep the government out of the power business. The big fear of many utilities was that the federal government would build reactors on its own and give the power to public corporations. Consequently, the fear of public power led utilities to invest in reactors before they were economically competitive.

A critical moment came in December 1963, when Jersey Central Power and Light signed a contract

with General Electric to purchase a 515-MW boiling water reactor for a plant in Oyster Creek, New Jersey. This agreement seemed to herald both the arrival of nuclear power as a viable alternative for generating electricity and the beginning of widespread commercial acceptance of nuclear power in the United States. The Oyster Creek reactor was to be the first built without any direct government subsidy, and the utility claimed the plant would be more economical to operate within 5 years, compared to conventional power sources. Although utilities began to order reactors in increasing numbers, in reality nuclear fission was not yet economically competitive with fossil fuels, primarily because of the higher capital costs for nuclear plants. In fact, the Oyster Creek plant was a "loss leader" designed to persuade utilities of the viability of reactor technology. Although it has been estimated that reactor vendors lost considerable sums of money on the earliest plants, the strategy worked—utilities began purchasing increasing amounts of nuclear generating capacity.

The sudden interest of the utilities in reactors can be attributed to several factors. First, demand for electricity in the 1960s was growing at an annual rate of 7%; at that rate of growth, demand would double every 10 years. Utilities thus needed to order new plants to meet projected demand. At the same time, many public utility commissions were lowering prices, which further increased demand for electricity. Next, because profits were determined by the size of the utility rate base, utilities had an incentive to build additional generating capacity to expand their rate base. Nuclear was an especially attractive option because it was by far the most capital-intensive source of electricity. Utilities thus had a number of financial incentives to build nuclear plants. Consequently, there was a rapid increase in both the number and the size of plants ordered after Oyster Creek. For example, American utilities ordered 70 reactors between 1963 and 1967, with approximately 80% of the orders being placed in 1966 and 1967. During that same period, the average capacity per reactor jumped from 550 to 850 MW. In 1967, nuclear power accounted for slightly more than half the total generating capacity ordered that year. The market for nuclear reactors was so good that some observers refer to this period as the "Great Bandwagon Market."

But the Bandwagon Market was based on exceedingly optimistic projections and expectations of construction costs and operating performance. Because commercial generation of electricity from fission was in its infancy and still largely untested, there was little in the way of hard data concerning actual costs and performance. In the early years of the commercial program, the utilities, the Commission, and the Joint Committee lacked the technical resources to assess the projections of reactor costs and efficiency offered by the reactor vendors. Not only did they lack the resources, but the AEC and JCAE also lacked the desire to evaluate the claims carefully. In their rush to commercialize reactor technology, government officials uncritically accepted the cost and efficiency estimates.

Electric utilities in the United States and abroad, caught up in the flood of good news about nuclear power, purchased more reactors and fewer fossil plants. The rush to nuclear power on the basis of the reactor vendors' early and exceedingly optimistic price estimated eventually came back to haunt utilities that had purchased reactors. Many utilities possessed neither the technical staff nor the financial resources to build and operate nuclear reactors. It would also become clear that nuclear plants were not as efficient or reliable as vendors had claimed. The optimistic predictions for nuclear power thus succeeded in establishing a market for the technology, but they also set the stage for the future economic problems that would cripple the industry.

3. NUCLEAR POWER COMES UNDER ATTACK

In 1965, the politics of commercial nuclear power were of little concern to anyone not having a direct economic or political stake in the programs' success. By the middle of the next decade, however, the policymaking arena was crowded and complex, and few issues in American politics were more visible or controversial. What changed is that, as nuclear reactors began to be built, more people came to believe that the issue affected them. In the process, nuclear power came to be understood in new, less positive ways. Long perceived as a national security issue, in the late 1960s, nuclear power came to be seen in the context of debates over environmental protection, public health and safety, and energy supplies. Eventually, it became part of even larger debates about government regulation of business, citizen participation, and democratic governance. As a result of these changing perceptions, subgovernment members lost the power to define the issue and control the parameters of debate. The subgovernment and nuclear policy were soon radically transformed.

As understandings and perceptions shifted and became increasingly negative, new actors were drawn to the nuclear issue. The new participants included other federal agencies, officials from state and local governments, and a number of prominent citizen groups. Many of the new participants would be critical of nuclear power. The influx of new participants, who brought their own opinions, shattered the consensus that had existed within the small atomic subgovernment, disrupted established patterns of policymaking, and created a decidedly less supportive political environment for the nuclear program.

3.1 The Emergence of an Antinuclear Movement

The boom in the market for nuclear plants coincided with the rise of the environment as an issue of national importance. Subsystem members initially perceived the growing concern with environmental values as a boon to the prospects of nuclear power. The AEC and the nuclear industry actively promoted the notion that nuclear power was cleaner than coal and other sources of electricity. This makes it all the more ironic that the first consistent opposition to nuclear power emerged in response to the potential environmental consequences of reactors. For the most part, the early opposition was essentially local in nature, being directed at particular reactors and not nuclear power. More specifically, concern during this period tended to focus on the thermal pollution caused by nuclear plants' discharge of heated water into nearby lakes and rivers. The AEC initially refused to consider the environmental effects of thermal pollution, arguing that it lacked statutory authority over nonradiological environmental matters. After a 3-year battle, and much bad publicity, Congress in 1970 adopted a law requiring federal agencies, including the AEC, to consider thermal pollution issue in the course of their licensing reviews. This controversy fundamentally altered the course of nuclear politics. It was the first critical step in the redefinition of the nuclear power issue; as concerns emerged about possible radioactive emissions from reactors and the issue of nuclear waste disposal, nuclear power was increasingly understood as an environmental issue, not a national security matter.

The most significant factor in the expansion of the debate and in the demise of the atomic subgovernment was the emergence of reactor safety as a prominent issue. The controversy over safety energized the burgeoning opposition to nuclear power and thrust the issue into the media spotlight, making it impossible for public officials to ignore. The safety issue emerged and developed in much the same way the environmental issue had—as a response by local citizens to particular nuclear reactors—before rapidly escalating to include questions about the safety of nuclear plants in general. But concerns over the safety of nuclear reactors attracted more attention and generated more controversy compared to other environmental issues, largely because nuclear technology was still relatively new and unfamiliar to the American public. When questions of reactor safety emerged, the dramatic nature of the potential consequences of reactor accidents lent the nuclear issue a sense of drama and urgency that earlier concerns about thermal pollution could not. Furthermore, because the issue was so dramatic, it captured the attention of the national media. Shortly thereafter, the conflict over the safety of nuclear reactors occupied a prominent position in the public agenda. The AEC's credibility suffered a fatal blow when experts, including its own, began to disagree in public. Once some insiders began raising questions about reactor safety and the AEC's commitment to protecting it, the conflict rapidly expanded.

Antinuclear activists in the late 1960s and early 1970s were also extremely critical of the AEC's licensing procedures and rules. Critics charged, for example, that AEC procedures precluded meaningful public participation in agency proceedings while granting the nuclear industry privileged access to AEC personnel, especially in the critical early stages of the licensing process. Critics also claimed that AEC regulations were primarily responsive to the needs and desires of the industry, reflecting the commission's enthusiasm for building and licensing reactors. Perhaps most important, critics charged that the decision to approve reactors had usually been made before citizens were given an opportunity to register their opinions. They also charged that the AEC's public hearings, the principal forum in which citizens were allowed to participate, were nothing more than stacked-deck proceedings designed to foster the illusion of citizen input.

By the middle of the 1970s, the subgovernment had lost its ability to define the nuclear issue; the debate over nuclear power thus widened to include a variety of new issues. Some opponents of nuclear power were suggesting that rather than being merely a technical matter, the nuclear issue involved important social, political, and moral questions that scientists and engineers had no special ability to resolve. Antinuclear forces argued that citizens could learn enough about nuclear power to participate in a

responsible manner if given the opportunity. In this way, antinuclear forces were able to link their concerns about nuclear safety to more general concerns about accountability and responsiveness in government.

By the end of the 1970s, citizen groups were challenging, in court and in the streets, the licensing of almost every reactor. Indeed, some the largest post-Vietnam protests and demonstrations concerned nuclear power. Several rallies drew crowds estimated to be in the hundreds of thousands. Nuclear power was plagued by similar developments in Europe, especially Germany, Great Britain, France, and Sweden. In most of these nations, as in the United States, the government became less supportive of nuclear power. France, which had fewer energy options, was a notable exception to this trend, and moved ahead with an aggressive nuclear program despite widespread public opposition.

3.2 Nuclear Power and the Courts

For most of the late 1960s and early 1970s, both the AEC and the Joint Committee consistently rebuffed persons or groups seeking to participate in the formulation of nuclear policy. In a textbook case of venue shopping, critics of nuclear power turned to the courts for assistance. One of the most significant consequences of increased oversight by the courts was that the commission, and its staff, devoted greater attention to procedural rights in an attempt to ensure that its procedures were seen as fair and capable of generating a record that could withstand judicial scrutiny. Rather than run the risk of being overturned by a reviewing court, the commission would try to show that it had solicited and considered many points of view.

The consequence of increased judicial oversight was a longer and more detailed review process that led, in turn, to more stringent environmental and safety standards. In an effort to satisfy its overseers, who were demanding more comprehensive and rigorous reviews, the commission upgraded the size and quality of its technical review staff. The larger, better, and more experienced regulatory staff began demanding more detailed information from utility applicants. Moreover, the commission was cognizant of the fact that the courts were more likely to overturn standards if the commission's decision-making process was procedurally deficient. As a result, the commission standardized its licensing review process in 1972. Specifically, the commission developed standard review plans for both safety and environmental analyses. The standard review plans identified the information needed by the staff in performing their technical review of applications and suggested a format for its presentation.

Another factor contributing to the commission's increasingly tough regulatory standards was a dramatic increase in the size of nuclear reactors. At the end of 1966, the largest reactor in operation was under 300 MW, but by the end of the decade, utilities were placing orders for plants two to three times larger. The problem was that larger, more powerful reactors posed more troublesome questions from a safety perspective. The fuel in larger reactors would overheat more quickly in the event of a failure in the plant's cooling system, and larger reactors contained higher levels of radioactivity. In short, larger reactors seemed to pose a greater risk to the public health and safety, especially because utilities were pressuring the commission to approve reactors closer to population centers.

3.3 The Rise of Energy Issues

The Arab oil embargo in October 1973 and the ensuing energy crisis profoundly altered the perception of the nation's energy policies in the eyes of the public and government officials and catapulted the energy issue to the top of the nation's political agenda. Although this turn of events might have been expected to bolster the prospects of nuclear power, it actually undermined them by subjecting the various energy subgovernments to intense scrutiny. With this increased scrutiny, energy interests were placed in direct competition for federal research and development funds; nuclear power, which had received the bulk of such monies over the years, would now face stiff competition. At the same time, government officials, seeking to establish a comprehensive and coordinated national energy plan, responded by reorganizing both the administrative and legislative branches. These organizational changes, in turn, would prove to be incompatible with continued domination of energy politics by subsystems.

3.4 The Fall of the AEC and JCAE

Largely as a result of the controversies over thermal pollution and radiation, by the early 1970s the AEC was on the defensive, suffering from a deteriorating public image and an almost complete lack of credibility. Recognizing the inevitability of the separation of the AEC's regulatory and developmental functions, members of the AEC, the JCAE, and

the nuclear industry supported the Energy Reorganization Act of 1974 because they viewed it as an opportunity to emphasize the contribution nuclear power could make to solving the nation's energy problems. Hoping to enhance the prospects of nuclear power, subsystem members favored the proposal even though it meant sacrificing the AEC, whose responsibilities would be transferred to two new agencies—the Energy Research and Development Administration (ERDA) and the Nuclear Regulatory Commission (NRC).

The abolition of the AEC deprived the nuclear industry of a powerful patron and created a more complicated institutional framework for nuclear matters. Where there was once a small group of insiders restricting participation to program supporters, by the middle of the decade the range of governmental participants would be decidedly broader and less exclusive. In addition to the Nuclear Regulatory Commission and ERDA, the courts, the Environmental Protection Agency (EPA), the Fish and Wildlife Service (FWS), and state and local governments claimed jurisdiction. The AEC, for example, had not been very concerned with the environmental impacts of reactors, but the EPA and FWS were. More and more, licensing decisions were challenged in court, and state and local governments became involved through their siting and rate-making powers. For opponents of nuclear power, these multiple venues translated into new opportunities to shape nuclear decision making and to appeal unfavorable decisions to other, more receptive governmental actors. For these reasons, institutional shifts were a critical factor in the expansion of the conflict over nuclear power and in eventual shifts in government policy. Like the AEC, the Joint Committee on Atomic Energy came under attack in the 1970s for being overly protective of the nuclear power industry. The growing wave of criticism over nuclear power, coupled with the demise of the AEC, prompted increased scrutiny of the Joint Committee and fueled calls for its abolition as well. Over the next few years, the once-invincible Joint Committee struggled to ward off challenges to its authority and to the nuclear program it had supported for so long. The joint committee would not win this battle, however, and it was eliminated by Congress in 1977.

With the demise of the JCAE, longtime supporters lost control of the strategic junctures of policy-making. Responsibility for nuclear policy was transferred to some two dozen subcommittees, some of which were sympathetic to the concerns voiced by antinuclear groups. For opponents of nuclear power, this decentralization reinforced previous venue changes and created additional access points to nuclear decision making. The new structural arrangements in Congress rendered the institution less exclusive and more permeable to outsiders, but also made it more complicated. Although nuclear policy once had been shaped exclusively the Joint Committee, it now seemed that everyone had some claim to nuclear oversight. With responsibility for nuclear policy dispersed so widely, it became increasingly difficult for Congress to resolve major policy issues.

Congress in the 1970s and 1980s was inhospitable to decisive initiatives from either side, preferring to approach nuclear policy in an ad hoc manner. Still fearful of energy shortages, members of Congress were unwilling to close the door on the nuclear option, deeming it an essential short-term energy source, but at the same time they were unwilling to proceed with a full-scale nuclear program because of concerns about reactor safety. For the next few years, Congress would increase funding for both reactor safety and licensing programs and also would insist on greater oversight of the NRC and the nuclear power industry. In this way, legislators could claim that they were doing everything possible to provide the nation with safe and secure energy supplies.

4. REGULATORY FAILURE AND THE CHANGING ECONOMICS OF NUCLEAR POWER

In the late 1960s, the costs of producing electricity from coal and nuclear plants were comparable; nuclear plants were more expensive to build but had lower fuel and operating costs. By the early 1970s, utilities were ordering roughly equal amounts of coal and nuclear capacity. Within a few years, though, the bottom fell out of the market, sparing no sector of the industry. Significantly, nuclear power began its demise well before the accident at Three Mile Island in March 1979. Orders for new nuclear plants actually began to decline in 1974, and no orders would be placed after 1978. Furthermore, between 1974 and 1984, utilities canceled plans for over 100 reactors, many of which were already under construction. Some of the abandoned reactors were more than 50% complete.

The industry's economic woes were attributable to internal and external forces that caused construction and maintenance costs to skyrocket and utilities to scale back their plans to build additional nuclear

generating capacity. Among the exogenous forces, none was more important than the Organization of Petroleum Exporting Countries (OPEC) oil embargo, which sparked sudden price hikes in fossil fuels and fundamentally transformed energy consumption patterns. Higher oil prices and the desire for secure energy sources might have been expected to make nuclear-generated electricity more attractive. Instead, fears of energy shortages and rising apprehension about American dependence on imported oil prompted Americans to use less energy, thus undercutting the need for new sources of electricity.

The nuclear industry was also hurt by rapidly rising construction costs. Construction costs for all types of power plants increased between 1968 and 1977, but those for nuclear plants rose at twice the rate of coal plants and continued to increase into the next decade. By almost any measure, nuclear plants became prohibitively expensive to build. Reactors coming on-line during the 1980s, for example, cost an average of $3500/kilowatt-hour in 1983 dollars, compared with an average of $600/kWh for the 57 reactors completed before 1981. A significant portion of the increased construction costs was attributable to new regulatory requirements that increased the amount of materials, labor, and time needed to build nuclear plants. Although the nuclear industry argued that most of their problems, especially the rising construction costs, were the result of overly zealous NRC regulators, in reality, the claim of excessive regulation is simply inaccurate and ignores other more compelling explanations.

It seems clear that earlier regulatory failures and industry mismanagement were responsible for some of the cost increases. The AEC neglected its regulatory duties for most of its history and devoted the bulk of its attention and resources to working with the private sector to develop and promote the new technology. Although this process did facilitate reactor licensing in the short term, the case-by-case approach ultimately undermined the industry's long-term prospects. During this period of rapid scale-up, no two reactor designs were alike, so every application was unique, and every application review was unique. In addition, the larger plants were more complex and posed greater potential risks. This increase in the number and complexity of applications, along with the failure to standardize reactor designs, posed tremendous difficulties for the AEC staff, which lacked the resources and experience to determine if the designs were adequate and if existing regulatory standards would assure public health and safety. The staff could not keep up, and licensing review times subsequently increased. The AEC responded to the growing licensing workload by nearly doubling its staff and by developing standard review plans to formalize the review process. As the licensing and regulatory staffs became larger and more sophisticated, they adopted new environmental and safety reviews and demanded ever more detailed information from applicants. Furthermore, as the quality and quantity of reviews increased, the staff discovered that existing standards were often inadequate and began to impose more stringent safety and environmental requirements on both new and existing plants.

Other regulatory changes were a product of increased operating experience. Once a sufficient number of reactors actually came on-line in the 1970s, the commission and the industry had a chance to evaluate the sufficiency of existing standards, designs, and operating procedures. Over time, previously unanticipated equipment and design problems were discovered, and remedial steps were taken to correct them.

What these examples show is that many of the industry's economic wounds were either self-inflicted or were the result of regulatory failures by the AEC. If the industry and the AEC had not rushed to commercialize nuclear power before the technology was mature, much of the political controversy, licensing delays, and economic ills might have been avoided. A strong regulatory program might have detected potential problems early on, before reactors were built, and even more important, before they started producing electricity. If such a program had been in place, costly design changes and construction stoppages would not have been unnecessary. But that is not what happened. Instead, the discovery of unanticipated problems and reluctant admissions that some existing standards were insufficient undermined confidence in the agency and the industry and fueled the growing safety controversy.

5. NUCLEAR POWER IN THE 1980s

5.1 The NRC after Three Mile Island

At approximately 4:00 a.m. on 28 March 1979, there was a serious accident at the Three Mile Island (TMI) nuclear facility near Harrisburg, Pennsylvania. Although it would take months to reconstruct the events and circumstances that caused the accident, it was much easier to calculate the political damage to the nuclear power industry and the NRC.

Put simply, the TMI accident renewed and deepened concerns about the safety of reactors. A blue-ribbon panel appointed by President Carter to investigate the accident at Three Mile Island was highly critical of the NRC and the nuclear industry and recommended major changes in the practices and attitudes of both. After concluding that there was "no well-thought-out, integrated system for the assurance of nuclear safety within the NRC," the Kemeny Commission recommended a "total restructuring" of the agency. That did not happen, and the NRC never became an effective regulator.

In part, this was because the NRC was not really a new agency—it was merely a spin-off of the regulatory branch of the AEC. Although the Energy Reorganization Act succeeded in abolishing the Atomic Energy Commission, it never entirely displaced the commission's deeply entrenched belief in the value of nuclear power. Part of the explanation stems from the fact that the NRC was essentially a carryover from the AEC in terms of personnel, regulations, and attitudes. Indeed, the NRC's first official action was to adopt all of the AEC's rules, regulations, and standards. Throughout the 1980s, the NRC behaved much like its predecessor. The agency consistently tried to reduce licensing delays, blocked the imposition of new safety requirements, and worked to curtail public involvement in agency proceedings. To a large extent, the NRC was responding to signals from Presidents Reagan and Bush, both of whom supported nuclear power and pushed a deregulatory agenda. Reagan, in particular, tried to control the NRC by appointing commissioners who shared his aversion to regulation.

When necessary, the NRC waived and changed rules that threatened to close existing plants or to delay the licensing of new ones. In several cases when plants did not meet the regulations, the NRC simply changed the regulations to facilitate licensing. The most notorious rule change involved the Shoreham, Long Island and Seabrook, New Hampshire reactors. After the accident at Three Mile Island revealed serious flaws in emergency and evacuation planning, the NRC issued new rules that required an emergency-planning zone within a 10-mile radius of nuclear plants. The rules also mandated state and local government approval of evacuation plans for all people living within the zone. When state and local officials adjacent to the two plants refused to prepare emergency plans, the NRC issued a new rule waiving the requirement that state and local officials participate in evacuation planning.

5.2 The Problem of Nuclear Waste Disposal

The question of what to do with nuclear waste did not receive much attention in the early years of the atomic program. The AEC and others in the subgovernment were more concerned with the scientific challenges of harnessing the atom. Nuclear waste, in contrast, was generally assumed to be a less complex, essentially technical problem that would be easy to solve. Reflecting its low priority within the AEC, action on nuclear waste disposal was deferred indefinitely. This decision, like so many others in the program's early years, only made the problem worse, because the lack of an effective nuclear waste policy became one of the antinuclear movement's most potent arguments in the 1980s.

The problem for utilities was that spent reactor fuel was piling up in cooling ponds at reactor sites around the nation. The cooling ponds, however, were not designed to store waste permanently, and most could hold no more than 3 years of spent fuel. Moreover, utilities could not expand their on-site storage capacity without state regulatory approval, which was by no means assured. By the late 1970s, with many utilities worried about running out of space and facing mounting on-site storage costs, the nuclear industry sought legislative relief, arguing that a federal commitment on high-level waste was essential to its future.

After several years of deliberation, Congress acted in 1982, adopting the Nuclear Waste Policy Act (NWPA). The NWPA established detailed procedures and a schedule for the selection, construction, and operation of two permanent high-level waste repositories. Under the law, the Department of Energy (DOE) would conduct an intensive nationwide search for potential sites, which would be evaluated with a demanding set of technical criteria and guidelines, including environmental assessments. The DOE was then to compile a list of potential sites from which the President would select two for further review and site characterization, one in the eastern United States and one in the west. By all accounts, implementation of the Nuclear Waste Policy Act was a failure. Although the DOE met the act's 1983 deadline for issuing siting guidelines, virtually every other stage of the selection process was bitterly contested. All of the states considered as potential repository sites challenged the designation. With the states in open rebellion, Congress tried to repair the damage by amending the NWPA. In December 1987, Congress rejected the multiple-site

search process established in the legislation and instead designated Yucca Mountain, Nevada, as the repository site. The search for a second site in the east was abandoned.

The case of nuclear waste disposal demonstrates the increasingly decentralized nature of nuclear politics in the 1980s. The widespread search for suitable nuclear waste disposal sites touched many states and communities and brought multiple new actors, in the form of state and local governments, to the policymaking arena. Whatever their motivations, the new actors brought different perspectives and goals to the debate; indeed, most actively resisted federal efforts to site nuclear waste facilities. With states and local governments assuming a critical role in policymaking, it is clear that the conflict over nuclear waste had expanded far beyond the confines of subgovernment politics. At the same time, however, the nuclear waste issue illustrates that support for nuclear power within the federal government remained strong in the 1980s. The Department of Energy, the NRC, and Presidents Reagan and Bush supported nuclear power, and worked to reestablish a favorable political climate. With states impeding the siting of a high-level repository, the industry sought to shift the financial and legal risks of nuclear waste management to the federal government, which willingly obliged.

5.3 The Accident at Chernobyl

In 1986, a fire at a reactor in Chernobyl, in the Soviet Union, released a large radioactive cloud across much of the Ukraine and northern Europe. The world learned of the accident when monitors in Europe detected an unusual spike in radiation levels. Soviet officials initially denied that anything had happened, but eventually conceded that something had gone horribly wrong at Chernobyl. Hundreds died, thousands were evacuated, and large areas were rendered uninhabitable. The story dominated the news throughout the world and reinforced questions about the safety of nuclear reactors. Although nuclear officials in the United States and elsewhere correctly noted that the Soviet reactor had fewer redundant safety systems than reactors in their own nations, the industry suffered a severe crisis in public confidence. In the United States and many European nations, the Chernobyl accident further undermined the prospects for a nuclear resurgence. Outside of France, which remained committed to a nuclear future, few European nations built new reactors.

6. NUCLEAR POWER TODAY

6.1 Nuclear Power in the United States

American policymakers continue to be ambivalent about commercial nuclear power. Reflecting this ambivalence, nuclear policymaking in the past 20 years has been marked by incrementalism, with policymakers unwilling, or unable, to stray far from the status quo. Barring another energy or environmental crisis, it is extremely unlikely that American utilities will soon begin building any additional nuclear plants. And in the absence of a severe accident, policymakers are equally unlikely to require the shutdown of existing reactors or to rule out the possibility of future contributions from nuclear power. In the interim, the range of acceptable policy options will continue to lie somewhere in between the two extremes.

Commercial nuclear power peaked in the United States in 1990, when 112 reactors were in operation. Today, there are 103 operable reactors, in 31 states, generating approximately 20% of the nation's electricity. All told, U.S. utilities have placed orders for 259 nuclear reactors, but none has been ordered since 1978. Of the total order, 132 reactors received full power licenses; the rest were canceled. No plant has commenced operations since 1996, and 28 reactors have been permanently closed down.

Despite these downward trends, nuclear capacity in the United States is projected to increase slightly over the next 25 years, mostly due to improved performance from existing plants and fewer nuclear retirements due to licensing extensions. Owners of 10 reactors have applied for and received 20-year license extensions, and another 20 have applications pending. Owners of an additional 20 reactors are expected to apply for license extensions in the next 6 years. The George W. Bush administration is currently working to increase nuclear power's contribution to the nation's energy needs. In 2001, as part of the administration's national energy strategy, the Department of Energy issued a "roadmap" to deploy new reactors in the United States by 2010. In another important move long sought by the nuclear industry, President Bush approved the siting of a high-level waste repository at Yucca Mountain, Nevada. Although the decision has been challenged by the state, the facility is scheduled to begin operation in 2010. If the decision stands, nuclear advocates will have removed, at least temporarily, one of the major political roadblocks to the technology's ultimate fate.

6.2 Nuclear Power after September 11

The terrorist attacks of September 11, 2001 thrust nuclear power into the spotlight once again, although not in the way that industry supporters hoped. One of the planes that brought down New York's Twin Towers flew by the Indian Point nuclear facility located just 35 miles north of the city. The NRC and industry leaders were forced to confront the question of whether either group was prepared to deal with such attacks, and their responses were not encouraging. Specifically, there are concerns that both reactors and spent fuel storage pools, which are often located adjacent to reactors, are vulnerable to terrorist attack. Critics charge, moreover, that the NRC and reactor owners have consistently underestimated the risk of such attack, and that the agency's security rules are too lax and are often ignored. In designing its rules, for example, the NRC did not consider attacks by plane to be a credible threat. And although earthquakes and other natural disasters were considered in the design of spent storage fuels, they were not designed to withstand terrorist attack. According to some reports, attacks on reactors and storage facilities could have consequences worse than the 1986 accident at Chernobyl.

6.3 Nuclear Power around the World

There are now 442 nuclear plants in operation worldwide, with an additional 35 under construction. Nuclear power provides about 17% of the world's electricity. Thirty nations now have nuclear reactors, with the vast majority located in Organization for Economic Cooperation and Development (OECD) nations, where nuclear power accounts for almost 24% of all electricity. According to the Nuclear Energy Institute, the top 10 nuclear-generating nations are, in order, the United States, France, Japan, Germany, Russia, South Korea, United Kingdom, Canada, Ukraine, and Sweden. Nuclear power provides more than 75% of total electricity production in Lithuania and France, more than 50% in Belgium and Slovakia, more than 40% in Ukraine, Sweden, and Bulgaria, and about 30% in 10 other nations. The United States, with 104 operating reactors, has the most; France has 59 and Japan has 54. Japan also has plans to build up to 10 additional plants in the next 10 years.

Current projections are that global energy demand will more than double in the next 50 years, leading more nations to consider nuclear power as an answer to their energy needs. The roster of nations that now have reactors includes a growing number that have been marked by economic, political, or social instability, or that are located in volatile regions of the globe. Among the nations that have added commercial reactors in the past 5 years are Pakistan, India, South Korea, and Brazil. Some of these nations and others, including Iran and North Korea, have reactors that have been linked to nuclear weapons programs. The spread of reactor technology to such nations raises considerable concerns about nuclear proliferation and, after September 11, about reactor vulnerability to terrorism. It is an open question whether these nations have the ability to secure their reactors and the nuclear materials they use and produce. As a result, the future of nuclear power remains clouded.

SEE ALSO THE FOLLOWING ARTICLES

Coal Industry, History of • Electricity Use, History of • Hydrogen, History of • Hydropower, History and Technology of • Manufactured Gas, History of • Natural Gas, History of • Nuclear Engineering • Nuclear Power Economics • Nuclear Power Plants, Decommissioning of • Nuclear Proliferation and Diversion • Nuclear Waste • Occupational Health Risks in Nuclear Power • Public Reaction to Nuclear Power Siting and Disposal

Further Reading

Balogh, B. (1991). "Chain Reaction: Expert Debate and Public Participation in American Commercial Nuclear Power, 1945–1975." Cambridge University Press, Cambridge.

Bupp, I. C., and Derian, J.-C. (1978). "Light Water: How the Nuclear Dream Dissolved." Basic Books, New York.

Campbell, J. L. (1988). "Collapse of an Industry: Nuclear Power and the Contradictions of U.S. Policy." Cornell University Press, Ithaca, New York.

Duffy, R. J. (1997). "Nuclear Politics in America: A History and Theory of Government Regulation." University Press of Kansas, Lawrence.

Falk, J. (1982). "Global Fission: The Battle over Nuclear Power." Oxford University Press, New York.

Green, H. P., and Rosenthal, A. P. (1963). "Government of the Atom: The Integration of Powers." Atherton Press, New York.

Hewlett, R. G., and Anderson, O. E. (1962). "A History of the Atomic Energy Commission, Vol. 1, The New World, 1939–1946." Pennsylvania State University Press, University Park.

Hewlett, R. G., and Duncan, F. (1969). "A History of the Atomic Energy Commission, Vol. 2, Atomic Shield, 1947–1952." Pennsylvania State University Press, University Park, PA.

Jasper, J. M. (1990). "Nuclear Politics: Energy and the State in the United States, Sweden, and France." Princeton University Press, Princeton, New Jersey.

Nelkin, D., and Pollak, M. (1981). "The Atom Besieged: Extraparliamentary Dissent in France and Germany." MIT Press, Cambridge, Massachusetts.

Union of Concerned Scientists. (1987). "Safety Second: The NRC and America's Nuclear Power Plants" (M. Adato, principal author, with J. MacKenzie, R. Pollard, and E. Weiss). Indiana University Press, Bloomington, Indiana.

Nuclear Power Plants, Decommissioning of

REBEKAH HARTY KRIEG
Rebekah Krieg Consulting
West Richland, Washington, United States

EVA E. HICKEY and JAMES R. WEBER
Pacific Northwest National Laboratory
Richland, Washington, United States

MICHAEL T. MASNIK
U.S. Nuclear Regulatory Commission
Washington, D.C., United States

1. **Definition of Decommissioning**
2. **Decommissioning Options and Activities**
3. **Decommissioning Process**
4. **Environmental Impacts**
5. **Summary of Decommissioning Status of U.S. Nuclear Power Facilities**
6. **Conclusion**

Glossary

boiling water reactor A nuclear power reactor in which water, used as both coolant and moderator, is allowed to boil in the reactor core. The resulting steam can be used directly to drive a turbine and electrical generator, thereby producing electricity.

contamination Any undesired radioactive material or residual radioactivity that is deposited on or in structures, areas, objects, or people in excess of acceptable levels.

decontamination The reduction or removal of contaminating radioactive material from equipment, structures, areas, objects, or people.

greenfield One possible end state of decommissioning in which above-ground structures have been removed and efforts made to revegetate the site. Buildings may have been removed to below grade and then covered with soil.

high-level waste Material that consists of (1) irradiated (spent) reactor fuel, (2) liquid waste from the operation of the first cycle solvent extraction system and concentrated wastes from subsequent extraction cycles, in a facility for reprocessing irradiated reactor fuel, or (3) solids into which such liquid wastes have been converted. Primarily in the form of spent fuel discharged from commercial nuclear power reactors, high-level waste also includes some reprocessed high-level waste from defense activities and a small quantity of reprocessed commercial high-level waste.

independent spent fuel storage installation (ISFSI) A complex designed and constructed for the interim storage of spent nuclear fuel and other radioactive materials associated with spent fuel storage. The most common design for an ISFSI at this time is a concrete pad with dry casks containing spent fuel bundles.

license termination Owners of nuclear power facilities in the United States hold a license that is granted by the U.S. Nuclear Regulatory Commission. This license is not terminated until the licensee can demonstrate compliance with specific criteria that detail the amount of radioactive material that can remain at the site following decommissioning. After the termination of the license, facility owners are free to use the facility or site for any nonnuclear uses they wish to pursue.

low-level waste A general term for a wide range of radioactive wastes. Industries, hospitals, research institutions, private or government laboratories, and nuclear fuel cycle facilities (e.g., nuclear power reactors and fuel fabrication plants) using radioactive materials generate low-level waste as part of their normal operations. These wastes have many physical and chemical forms and different levels of contamination. Low-level waste usually comprises radionuclide-contaminated materials (rags, papers, filters, solidified liquids, ion-exchange resins, tools, equipment, discarded protective clothing, dirt, construction rubble, concrete, or piping).

 © 2004 Elsevier Inc. All rights reserved.

partial site release The release of a portion of an operating or decommissioning nuclear power reactor facility site for unrestricted use.

pressurized water reactor A power reactor in which heat is transferred from the core to an exchanger by high-temperature water kept under high pressure in the primary system. Steam is generated in a secondary circuit. Many reactors producing electric power are pressurized water reactors.

spent fuel The depleted nuclear fuel that has been removed from a nuclear reactor, for economic or other reasons, or because it can no longer sustain power production (cannot effectively sustain a chain reaction).

Decommissioning is a technical and administrative process by which a nuclear facility or site is safely removed from service and residual radioactivity is reduced to a level that permits release of the property for either partial or unrestricted public uses. Decommissioning is a planned, systematic activity that is both remediative and administrative. Environmental impacts are generally less during decommissioning than during construction and, in the context of water use, may be less than during operations. Nonetheless, decommissioning does impact land and water use, water and air quality, and local aquatic and terrestrial ecologies.

1. DEFINITION OF DECOMMISSIONING

During nuclear power plant reactor operation, a large inventory of radioactive fission products builds up within the fuel. Virtually all of the fission products are contained within the fuel pellets, which are encased in hollow metal rods, hermetically sealed to prevent further release. Occasionally, however, fuel rods develop small leaks, allowing a small fraction of the fission products to contaminate the reactor coolant. The radioactive contamination in the reactor coolant is the source of gaseous, liquid, and solid radioactive wastes generated in light water reactors during operation. Despite precautions to prevent the movement of contaminated material and to clean up any contaminated materials, such material is unintentionally moved through facilities by workers, equipment, liquids, and, to some degree, air motion. Contamination is most likely to occur in the reactor building, around the spent fuel pool, and around structures, systems, and components in areas near the reactor. During decommissioning, remediative measures are utilized to eliminate fission product contaminants.

The second source of radioactive material encountered in reactors during decommissioning is from activation. Activation products are created when stable substances are bombarded by neutrons. Concrete and steel surrounding the core of the reactor are the most common locations of activated material.

The spent fuel contains the largest amount of radioactive material at a facility that is being decommissioned, followed by the reactor vessel, internals, and bioshield. Systems containing smaller amounts of radioactive material include the steam generator, pressurizer, piping of the primary system and other systems, as well as the radioactive waste systems. Minor contamination is found in the secondary systems and miscellaneous piping.

The definition of decommissioning provides for a definite starting point and ending point for the decommissioning process and for the evaluation of environmental impacts from decommissioning. Decommissioning starts after a facility has been removed from service. Thus, the decision to permanently cease operation of a facility and the activities and resulting environmental impacts from permanent cessation of operation (such as the change in water temperature of the cooling water body) are not considered a part of the decommissioning process. The decommissioning process starts with a planning and preparation process and then moves through a plant transition and deactivation phase, in which fuel is transferred to temporary spent fuel pool storage, the systems are drained and may be flushed, and support systems may be installed or moved. A third phase, decontamination and dismantlement precedes the fourth phase, compliance with relevant regulations for final disposition of the site. Some activities that may occur during the decommissioning process could also occur during normal operations (such as construction, maintenance, and decommissioning of an independent spent fuel storage installation; transportation and disposal of the spent fuel, away from the reactor location; or disposal of low-level waste). These activities should not be considered exclusive to decommissioning.

2. DECOMMISSIONING OPTIONS AND ACTIVITIES

Three approaches are generally the options used for decommissioning nuclear power facilities:

1. DECON. The equipment, structures, and portions of the facility and site that contain radioactive contaminants are promptly removed or decontaminated to a level that permits cessation of operations and release of the site for public use (requires approximately 5 to 10 years).
2. SAFSTOR. The facility is made safe and stable and is maintained in that state (safe storage) until it is subsequently decontaminated and dismantled. During SAFSTOR, a facility is left intact, but the fuel has been removed from the reactor vessel and radioactive liquids have been drained from systems and components and then processed. Radioactive decay occurs during the SAFSTOR period, thus reducing the quantity of contaminated and radioactive material that must be disposed of during decontamination and dismantlement. The definition of SAFSTOR also includes decontamination and dismantlement of the facility at the end of the storage period.
3. ENTOMB. Radioactive structures, systems, and components are encased in a structurally long-lived substance, such as concrete. The entombed structure is appropriately maintained, and continued surveillance is carried out until the radioactivity decays to a safe level.

These options can be combined. For example, after power operations cease at a facility, a short storage period could be used for planning purposes, followed by removal of large components (such as the steam generators, pressurizer, and reactor vessel internals). At this point, the facility could be placed in storage for 30 years, and eventually the decontamination and dismantlement process would be completed.

2.1 DECON

Numerous activities may occur during DECON. Draining (and potentially flushing) any contaminated systems and removing resins from ion exchangers are early steps. Setup activities include establishing monitoring stations or designing and fabricating special shielding and contamination control envelopes to facilitate decommissioning activities. Reduction of the site security area (setup of new security monitoring stations), modification of the control room or establishment of an alternate control room may also occur. Site surveys are an important step. Decontamination of radioactive components, using chemical decontamination techniques also occurs. This often precedes the removal of reactor vessel and internals, removal of other large components (including major radioactive components), and removal of the balance of the primary system (charging system, boron control system, etc.). Activities related to removing other significant radioactive components are also undertaken. Decontamination and/or dismantlement of structures or buildings, temporary on-site storage of components, shipment and processing of low-level radioactive waste (including compaction or incineration of the waste), removal of the spent fuel and other non-low-level wastes to a storage facility at, or away from, the reactor site, and removal of hazardous radioactive (mixed) wastes are activities that may occur at any time during the process. Changes in management and staffing are also usually implemented.

There are several advantages to using the DECON option of decommissioning. By beginning the decontamination and dismantlement process soon after permanent cessation of operation, the available workforce can be maintained. This is advantageous because the facility staff is knowledgeable about the facility and will not require the retraining that would be necessary should the facility be placed in storage for 10 to 40 years. A second advantage may be the availability of facilities willing to accept low-level waste. Low-level waste disposal sites available to reactor facilities may be somewhat variable from year to year, and there is not always a certainty that a site will be available in future years. A third advantage is that costs for DECON may be lower than for other options because the price for decommissioning at a later date is greater, factoring in the cost of storage and inflation. Fourth, DECON also eliminates the need for long-term security, maintenance, and surveillance of the facility (excluding the on-site storage of spent fuel), which are required for the other decommissioning options. Fifth, completing the decommissioning process and being in compliance with relevant regulations for final disposition of the site allow the facility and site to become available for other purposes. The major disadvantages of DECON are higher levels of worker exposure to radiation and significant initial expenditures. Also, compared to SAFSTOR, DECON requires a larger potential commitment of disposal site space.

2.2 SAFSTOR

The SAFSTOR decommissioning option involves ensuring that the facility is in a safe, stable condition and maintaining that state for a period of time, followed by subsequent decontamination and dismantlement to levels that permit license termination. During the storage period of SAFSTOR, the facility is

left intact. The fuel has been removed from the reactor vessel and radioactive liquids have been drained from systems and components and processed. Radioactive decay occurs during the storage period, reducing the quantity of contaminated and radioactive material that must later be disposed of during decontamination and dismantlement. A number of activities typically occur during the preparation and storage stages of the SAFSTOR process. During preparation, there must be draining (and potential flushing) of some systems and removal of resins from ion exchangers, reconfiguration of the spent fuel pool cooling system, decontamination of highly contaminated and high-dose areas as necessary, performance of a radiological assessment as a baseline before storage, and removal of low-level waste that is ready to be shipped. These activities will occur at the same time or be followed by shipment and processing or storage of the fuel and other non-low-level wastes, deenergizing or deactivating systems and equipment, and reconfiguration of ventilation systems, fire protection systems, and the spent fuel pool cooling system for use during storage. Establishment of inspection and monitoring plans for use during storage, maintenance of any systems critical to final dismantlement during storage, are the final requirements of the preparatory stage. Changes in management or staffing could occur at any time.

During storage, there must be performance of preventative and corrective maintenance on plant systems that will be operating and/or functional during the storage period. There will also have to be maintenance to preserve structural integrity, maintenance of security systems, and maintenance of radiation effluent and environmental monitoring programs. Processing of any radioactive waste generated (usually small amounts) must also be done during the storage period. Following the storage period, the facility is decontaminated and dismantled to radiological levels that allow termination of the license. Activities during this period of time will be the same activities that occur for DECON.

There are several advantages to using the SAFSTOR option of decommissioning. A substantial reduction in radioactive material as a result of radioactive decay during the storage period reduces worker and public exposure to doses below those of the DECON alternative. Because there is potentially less radioactive waste, less waste disposal space is required. Moreover, the costs immediately following permanent cessation of operations are lower than costs during the first years of DECON because of reduced amounts of activity and a smaller workforce. However, because of the time gap between cessation of operations and decommissioning activities, SAFSTOR can result in a shortage of personnel familiar with the facility at the time of dismantlement and decontamination and a necessity to train new personnel. In some cases, this creates a knowledge gap because the people most familiar with the facility have long since retired by the time the storage period is completed. Other disadvantages are that during the prolonged period of storage, the plant requires continued maintenance, security, and surveillance and there are uncertainties regarding the availability and cost of future low-level waste sites, which could mean higher future costs for decontamination and dismantlement.

2.3 ENTOMB

The ENTOMB decommissioning option involves encasement of long-lived radioactive contaminants in a structurally long-lived material, such as concrete. The entombed structure is appropriately maintained and surveillance is continued until the radioactivity decays to a level permitting release of the property. The purpose of the entombment process is to isolate the entombed radioactive waste so that the facility can be released to partial or unrestricted use. Therefore, prior to entombment, an accurate characterization of the radioactive materials that are to remain is needed, and the adequacy of the entombment configuration to isolate the entombed radioactive waste must be determined.

To characterize the environmental impacts from decommissioning via the ENTOMB option adequately, two different scenarios were developed to encompass a wide range of potential options by describing two possible extreme cases of entombment. The two scenarios, “ENTOMB 1” and “ENTOMB 2,” differ primarily in the amount of decontamination and dismantlement that is done prior to the actual entombment. ENTOMB 1 assumes significant decontamination and dismantlement and removal of all contamination and activation involving long-lived radioactive isotopes prior to entombment. ENTOMB 2 assumes significantly less decontamination and dismantlement, significantly more engineered barriers, and the retention on site of long-lived radioactive isotopes. Both options assume that the spent fuel would be removed from the facility and either transported to a permanent high-level waste repository or placed in on-site storage.

ENTOMB 1 is envisioned to begin the decommissioning process in a manner similar to the

DECON option. The reactor would be defueled and the fuel initially either placed into the spent fuel pool for some period prior to disposal at a licensed high-level waste repository or placed in an on-site storage. Any decommissioning activity would be preceded by an accurate radiological characterization of structures, systems, and components (SSCs) throughout the facility. Active decommissioning would begin with draining and decontaminating of SSCs throughout the facility, with the goal of isolating and fixing contamination. The SSCs would either be decontaminated or removed and either shipped to a low-level waste burial site or placed inside the reactor containment building. Off-site disposal of resins and considerable amounts of contaminated material would occur. There would likely be a chemical decontamination of the primary system. The reactor pressure vessel and reactor internals would be removed, either intact or after sectioning, and disposed of off site.

Any other SSCs that have long-lived activation products would be removed. Interim dry storage of the vessel, vessel internals, and any other SSCs containing long-lived activation products could occur on site until a final disposal site for this waste is identified. Depending on whether the components are contaminated with long-lived radioisotopes, steam generators and the pressurizer would either be removed and disposed of off site or be retained inside the reactor containment. The spent fuel pool would be drained and decontaminated. The reactor building or containment would then be filled with SSCs contaminated with relatively short-lived isotopes from the balance of the facility. Material would be placed in the building in a manner that would minimize the spread of any contamination (i.e., dry, contamination fixed, or isolated). Engineered barriers would be put in place to deny access and to eliminate the possibility of release of any contamination to the environment. The reactor building or containment would be sealed and made weatherproof.

A partial site release could be completed for almost all of the site and the balance of the plant, thus allowing the remainder of the site to be used for other purposes. A monitoring period as long as 20 to 30 years would probably be necessary to demonstrate that the contamination was isolated and the structure was permanent. The general activities that would occur during ENTOMB 1 are planning and preparation activities, draining (and possibly flushing) contaminated systems and removing resins from ion exchangers, reduction of the site security area (optional), deactivation of support systems, and decontamination of radioactive components, including use of chemical decontamination techniques. There would also be removal of the reactor vessel and internals, removal of other large components (including major radioactive components), removal of fuel from the spent fuel pool to a storage facility at or away from the reactor site, dismantlement of remaining radioactively contaminated structures and placement of the dismantled structures in the reactor building, installation of engineered barriers and other controls to prevent inadvertent intrusion and dispersion of contamination outside of the entombed structure, and filling the void spaces in the previous reactor building structure with concrete.

ENTOMB 2 was also envisioned to begin the decommissioning process in a manner similar to the DECON option. The reactor would be defueled and the fuel initially placed into the spent fuel pool for some period prior to disposal at a permanent high-level waste repository or placed in on-site storage. Any decommissioning activity would be preceded by an accurate radiological characterization of SSCs throughout the facility. Active decommissioning would begin with the draining and decontamination of SSCs throughout the facility, with the goal of isolating and fixing contamination. The spent fuel pool would be drained and decontaminated. SSCs would either be decontaminated or be removed and either shipped to a low-level waste burial site or placed inside the reactor containment building or the reactor building. Disposal off site of resins would occur. The primary system would be drained, the reactor pressure vessel filled with contaminated material, all penetrations sealed, the reactor pressure vessel head reinstalled, and the reactor vessel filled with low-density concrete. Reactor internals would remain in place. Emphasis would be placed on draining and drying all systems and components and fixing contamination to prevent movement, either by air or liquid means. The steam generators and pressurizer would be laid up dry and remain in place. The reactor building or containment would then be filled with contaminated SSCs from the balance of the facility. Material would be placed in the building in a manner that would minimize the spread of any contamination (i.e., dry, contamination fixed, or isolated).

Engineered barriers would be put in place to deny access and eliminate the possibility of the release of any contamination to the environment. The ceiling of the containment or reactor building, in the case of boiling water reactors, may be lowered to near the refueling floor and to the top of the pressurizer for

pressurized water reactors. The cavity of the remaining structure would be filled with a low-density concrete. The resulting structure would be sealed and made weatherproof and covered with an engineered cap designed to deny access, and prevent the intrusion of water or the release of radioactive contamination to the environment.

A partial site release would be completed for most of the site and the balance of the plant, thus allowing the remainder of the site to be used for other purposes if appropriate. A site-monitoring program would likely be initiated to demonstrate the isolation of the contamination and the permanence of the structure. Monitoring could be as long as 100 years. The general activities that would occur during ENTOMB 2 are planning and preparation activities, draining (and potentially flushing) of contaminated systems and removal of resins from ion exchangers, deactivation of support systems, and removal of fuel from the spent fuel pool to a storage facility at or away from the reactor site. It would be necessary to dismantle all radioactively contaminated structures (other than the reactor building), with placement of the dismantled structures in the reactor building. The ceiling of the reactor building would be lowered to near the refueling floor (in boiling water reactors) or near the top of the pressurizer (in pressurized water reactors), and engineered barriers installed and other controls put in place to prevent inadvertent intrusion and dispersion of contamination outside of the entombed structure, filling of the cavity of the reactor building structure with low-density concrete, and placement of an engineered cap over the entombed structure to further isolate the structure from the environment.

The advantages of both ENTOMB options are reduced public exposure to radiation due to significantly less transportation of radioactive waste to low-level waste disposal sites and corresponding reduced cost of low-level waste disposal. An additional advantage of ENTOMB 2 is related to the significant reduction in the amount of work activity, and thus a significant reduction in occupational exposures, as compared to the DECON or SAFSTOR decommissioning options.

3. DECOMMISSIONING PROCESS

The decommissioning process, as shown in Fig. 1, can be divided into four stages, beginning with planning, proceeding through deactivation activities, decontamination and dismantlement and finally to administrative clearance for final disposition of the site. The first and last stages are primarily administrative.

Planning typically takes 1.5 to 2.5 years, regardless of the decommissioning option chosen. The main activities during this stage are determining the decommissioning option, making changes to the organizational structure (such as layoffs, strategic hiring, and redeploying personnel) to prepare for decommissioning activities, and starting the necessary regulatory activities. One preparation would be conducting postshutdown surveys of the facility for contamination or activation material. Planning activities vary, depending on the decommissioning option chosen and on whether the shutdown was planned or unplanned. If planned, some preparations may have already been made.

During stage 2, the plant moves from cessation of reactor operation to decommissioning activities. This stage will last from about 0.5 to 1.5 years, regardless of the decommissioning option chosen. Fuel will have to be transferred out of the reactor and into the spent fuel pool at all plants. Isolation, stabilization, or removal of all unnecessary systems is also conducted during this stage and new support systems needed for decontamination may be added. Changes to electrical systems, for instance, are common at this stage. Chemical decontamination of the primary system at this stage has resulted in reduction of total person-rem exposure during subsequent decommissioning activities. Chemically reducing levels of contaminants from piping, for instance, may significantly reduce doses to which workers are exposed and will keep costs lower.

Stage 3 involves decontamination and dismantlement of the plant for the DECON, SAFSTOR, and ENTOMB 1 options. For ENTOMB 2, stage 3 involves dismantlement of all radioactively contaminated

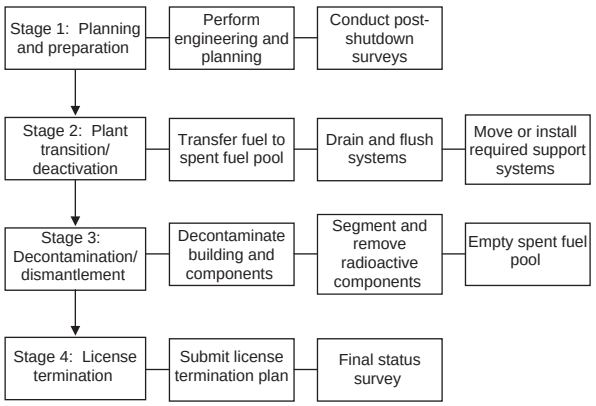


FIGURE 1 Reactor decommissioning process.

SSCs external to the reactor building and placement of these SSCs in the reactor building, followed by lowering the ceiling to the D-rings (pressurized water reactors) or the refueling floor (boiling water reactors). The greatest variability within a decommissioning stage is found in stage 3 because of the varieties of ways that systems can be dismantled. For instance, the secondary components have been dismantled at some plants in the United States, removing the moisture separators, diesel generators, and steam piping. Another plant has had its steam generators and pressurizer removed intact because transport and disposal facilities were easily available; the reactor vessel and internals were then removed whole (instead of having to be dismantled), filled with grout, welded closed, and shipped.

For the SAFSTOR option, stage 3 includes the time in storage. Decontamination and dismantlement activities can occur before, during, and after storage in SAFSTOR, with the time for stage 3 varying from a few years to more than 50 years. In both ENTOMB 1 and ENTOMB 2, grout and engineered barriers are installed. A monitoring program is also developed for the ENTOMB options. For ENTOMB 2, stage 3 involves dismantling all radioactively contaminated systems external to the reactor building and placing them in the reactor building. In pressurized water reactors, the ceiling is lowered to the D-rings; in boiling water reactors, the ceiling is lowered to the refueling floor.

Stage 4 is final administrative termination. Activities for this stage, which are similar for all options, include final site characterization, the final radiation survey, submission of a final termination plan, and a final site survey. The ENTOMB options would include both a partial site release and a detailed site monitoring program.

4. ENVIRONMENTAL IMPACTS

In order to adequately describe the environmental impacts of decommissioning, it is helpful to separate the impacts from those that occurred during site construction or operation and the immediate impacts that occur as a result of the decision to cease power operations. In general, environmental impacts occurring as a result of decommissioning a nuclear power facility are smaller than those impacts resulting from construction or operation of the facility. Immediate impacts as a result of the decision to cease power operation include the cessation of warm water flowing into the body of water that receives the plant's thermal discharge, which may result in the loss of aquatic or marine organisms that have grown to be dependent on the additional heat from the facility. Potential environmental impacts during the decommissioning process include impacts on land and water use, water quality, air quality, aquatic ecology, and terrestrial ecology, in addition to radiological, socioeconomic, and aesthetic impacts.

4.1 Land Use

Land use during decommissioning is affected by the site layout. Most sites have sufficient area existing within the previously disturbed area (whether during construction or operation of the site) and, therefore, no additional land needs to be disturbed. However, some activities projected for decommissioning are expected to require land temporarily. Temporary changes may include expanded staging and laydown areas for removal of large facility components. In addition, the large number of temporary workers needed to accomplish the major decommissioning activities may require that temporary facilities be installed for on-site parking, training, site security access, office space, change areas, fabrication shops, and other needs. Widening and rebuilding access roads or creating or expanding gravel pits for building roads may occur off site, but plants that have been decommissioned thus far have not needed such additional land.

4.2 Water Use

Nuclear reactor facilities are usually located near or adjacent to significant water bodies (aquifers, rivers, lakes, etc.) that are important to the region. Operating nuclear reactor facilities use water from multiple sources. For example, water from an adjacent lake might provide cooling water, whereas potable water may come from groundwater wells located on site. Reactor cooling is the greatest use of water at an operating reactor; other uses include waste treatment, potable water, process water, and site maintenance. In general, the impact of nuclear reactor facilities on water resources dramatically decreases after plants cease operation. The operational demand for cooling and makeup water is largely eliminated after the facility permanently ceases operation.

Most of the impacts on water resources during decommissioning of a nuclear facility are typical of those during decommissioning or construction of any large industrial facility. For example, providing water

for dust abatement is a concern for any large construction project, as is potable water usage. However, the quantities of water required for decommissioning are trivial compared to those used during operations. Water use will be greater in facilities that are undergoing decontamination and dismantlement than in those that are in the storage phase.

4.3 Water Quality

The major decommissioning activities that could affect water quality include fuel removal, stabilization, decontamination and system dismantlement, and structure dismantlement. Surface waters are most likely to be impacted either by storm water or by releases of substances during decommissioning activities. However, unintentional releases of hazardous substances have been an infrequent occurrence at decommissioning facilities. Because the focus of decommissioning is the ultimate cleanup of the facility, considerable attention is placed on minimizing spills.

Certain decommissioning activities or options may result in changes in local water chemistry. For example, dismantling structures by demolition and disposing of the concrete rubble on site increase the possibility that hydration of the concrete could locally increase the alkalinity of the groundwater.

4.4 Air Quality

Decommissioning activities that have the potential to have a nonradiological impact on air quality include (1) worker transportation to and from the site, (2) dismantling of systems and removing of equipment, (3) movement and open storage of material onsite, (4) demolition of buildings and structures, (5) shipment of material and debris to off-site locations, and (6) operation of concrete batch plants. These activities typically take place over a period of years from the time the facility ceases operation until the decommissioning is complete. The magnitude and the timing of the potential impacts of each activity will vary from plant to plant, depending on the decommissioning option selected and the status of facilities and structures at the completion of decommissioning. The most likely impact of decommissioning on air quality is fugitive dust, although it should be less than during plant construction because the sizes of the disturbed areas are smaller, the duration of activity shorter, and paved roadways available. Best management practices, such as seeding or wetting, can be used to minimize fugitive dust.

4.5 Aquatic Ecology

Aquatic ecological resources may be impacted during the decommissioning process by either direct or indirect disturbance of plant or animal communities near the facility site. Direct impacts can result from activities such as removal of structures (e.g., the intake or discharge facilities) at the shoreline or in the water; the active dredging of a stream, river, or ocean bottom; or filling a stream or bay. Indirect impacts may result from effects such as runoff. If decommissioning does not include removal of shoreline or in-water structures, very little aquatic habitat will probably be disturbed. Thus, practically all aquatic habitat that was used during regular plant operations or, at a minimum, was not previously disturbed during construction of the site will not be impacted.

4.6 Terrestrial Ecology

Similar to aquatic resources, terrestrial resources may also be affected during the decommissioning process. Direct impacts can result from activities such as the clearing of native vegetation or filling of a wetland. Indirect impacts may result from effects such as erosional runoff, dust, or noise. During decommissioning, land at the site and the surrounding terrestrial ecology may be disturbed for the construction of laydown yards, stockpiles, and support facilities. Additionally, land away from the plant site may be disturbed by upgrading or installing new transportation or utility systems. In most cases, land disturbances will result in relatively short-term impacts and ecological systems will either recover naturally or will be landscaped appropriately for an alternative use after completion of decommissioning.

4.7 Radiological

Radioactive materials are present in the reactor and in the support facilities after operations cease and the fuel has been removed from the reactor core. Exposure to these radioactive materials during decommissioning may have consequences for workers. Members of the public may also potentially be exposed to radioactive materials that are released to the environment during the decommissioning process. To date, doses from occupational exposures to individual workers during decommissioning activities at nuclear power facilities in the United States are similar to, or lower than, the doses experienced by workers in operating facilities. In addition, these

doses have been below the regulatory limits. Off-site exposures, which could affect members of the public, during periods of major decommissioning have not differed substantially from those experienced during normal operations, which are also below the regulatory limits.

4.8 Socioeconomics

There are two primary pathways through which nuclear power plant activities create socioeconomic impacts on the area surrounding the facility. The first is through expenditures in the local community by the plant workforce and direct purchases of goods and services required for plant activities. The second pathway for socioeconomic impact is through the effects on local government tax revenues and services. When a nuclear power plant is closed and decommissioned, most of the important socioeconomic impacts will be associated with plant closure, rather than with the decommissioning process. The impacts occur either through changing employment levels and local demands for housing and infrastructure, or through decline of the local tax base and the ability of local government entities to provide public services. The effect on the local tax base and public services related to closure depends on the size of the plant-related tax base relative to the overall tax base of local government, as well as on the rate at which the tax base is lost. Experience has shown that publicly owned tax-exempt plants do not have an impact through this mechanism. In addition, fully depreciated plants or a plant that is located in an urban or urbanizing area with a large or rapidly growing tax base will also not have an impact by this mechanism. A large, newer, relatively undepreciated plant, located in a small, isolated community, is much more likely to have an impact. If the plant tax base is phased out slowly after closure, the impact is more likely to be mitigated.

4.9 Aesthetic Issues

The aesthetic impacts of decommissioning fall into two categories: (1) impacts, such as noise, associated with decommissioning activities that are temporary and cease when decommissioning is complete and (2) the changed appearance of the site when decommissioning is complete. During decommissioning, the impact of activities on aesthetic resources would be temporary. The impacts would be limited both in terms of land disturbance and the duration of activity and would have characteristics similar to those encountered during industrial construction: dust and mud around the construction site, traffic and noise of trucks, and construction disarray on the site. In most cases, these impacts would not easily be visible off site. Aesthetic impacts could improve fairly rapidly in the case of an immediate DECON if the facility were dismantled, the structures removed, and the site regraded or revegetated. Impacts could also remain the same or would be similar in the case in which the structures are maintained during the decommissioning period and left standing, or throughout a long SAFSTOR or ENTOMB period. In the latter cases, the aesthetic impacts of the plant would be similar to those that occurred during the operational period.

5. SUMMARY OF DECOMMISSIONING STATUS OF U.S. NUCLEAR POWER FACILITIES

Twenty-two of the commercial nuclear reactors in the U.S. that are licensed by the U.S. Nuclear Regulatory Commission have permanently shut down and have either had their licenses terminated or are currently being decommissioned. Each facility has its own characteristics, depending on the decommissioning options being used for each facility and each facility's

TABLE I

Summary of Shutdown Plant Information

Parameter	Number
Reactor type	
Boiling water reactor	8
Pressurized water reactor	11
High-temperature, gas-cooled reactor	2
Fast breeder reactor	1
Decommissioning option	
SAFSTOR	14
DECON	7
Accident cleanup followed by storage	1
Fuel location	
Fuel on site in pool	13
No fuel on site[a]	8
Fuel on site in ISFSI[b]	1
Plan to move fuel to an ISFSI between 2000 and 2005	9

[a] Includes Three Mile Island, Unit 2, which has approximately 900 kg of fuel remaining on site due to the accident.

[b] ISFSI, Independent spent fuel storage installation.

decommissioning activities. An overview of the shutdown plants can be found in Table I, which includes 22 units shut down between 1963 and 1997. Table II summarizes important characteristics of the shutdown plants. The thermal power capabilities of the reactors ranged from 23 to 3411 MWt. The reactors operated from just a few days (Shoreham) to 33 years (Big Rock Point).

TABLE II

Permanently Shutdown Plants in the United States (2003)

Nuclear plant	Reactor type[a]	Thermal power	Shutdown date[b]	Decommissioning option[c]	Location	Fuel status and license termination date
Plants currently in decommissioning process						
Big Rock Point	BWR	240 MW	08/30/97	DECON	Michigan	Fuel in ISFSI
Dresden, Unit 1	BWR	700 MW	10/31/78	SAFSTOR	Illinois	Fuel in ISFSI
Fermi, Unit 1	FBR	200 MW	09/22/72	SAFSTOR[d]	Michigan	No fuel on site
GE-VBWR	BWR	50 MW	12/09/63	SAFSTOR	California	No fuel on site
Haddam Neck	PWR	1825 MW	07/22/96	DECON	Connecticut	Fuel in pool
Humboldt Bay, Unit 3	BWR	200 MW	07/02/76	SAFSTOR[d]	California	Fuel in pool
Indian Point, Unit 1	PWR	615 MW	10/31/74	SAFSTOR	New York	Fuel in pool
La Crosse	BWR	165 MW	04/30/87	SAFSTOR	Wisconsin	Fuel in pool
Maine Yankee	PWR	2700 MW	12/06/96	DECON	Maine	Fuel in ISFSI[e]
Millstone, Unit 1	BWR	2011 MW	11/04/95	SAFSTOR	Connecticut	Fuel in pool
Peach Bottom, Unit 1	HTGR	115 MW	10/31/74	SAFSTOR	Pennsylvania	No fuel on site
Rancho Seco	PWR	2772 MW	06/07/89	SAFSTOR[d]	California	Fuel in ISFSI/partial DECON proposed in 1997
San Onofre, Unit 1	PWR	1347 MW	11/30/92	SAFSTOR[d]	California	Fuel in pool
Saxton	PWR	28 MW	05/01/72	SAFSTOR[d]	Pennsylvania	No fuel on site/ currently in DECON
Three Mile Island, Unit 2	PWR	2772 MW	03/28/79	Accident cleanup followed by storage	Pennsylvania	Approx 900 kg fuel on site/postdefueling monitored storage
Trojan	PWR	3411 MW	11/09/92	DECON	Oregon	Fuel in ISFSI
Yankee Rowe	PWR	600 MW	10/01/91	DECON	Massachusetts	Fuel in ISFSI[e]
Zion, Unit 1	PWR	3250 MW	02/21/97	SAFSTOR	Illinois	Fuel in pool
Zion, Unit 2	PWR	3250 MW	09/19/96	SAFSTOR	Illinois	Fuel in pool
Terminated licenses						
Fort St. Vrain	HTGR	842 MW	08/18/89	DECON	Colorado	Fuel in ISFSI/license terminated in 1997
Pathfinder	BWR	190 MW	09/16/67	SAFSTOR	South Dakota	No fuel on site/license terminated in 1992
Shoreham	BWR	2436 MW	06/28/89	DECON	New York	No fuel on site/license terminated in 1995

[a] BWR, Boiling water reactor; FBR, fast breeder reactor; PWR, pressurized water reactor; HTGR, high-temperature, gas-cooled reactor.
[b] The shutdown date corresponds to the date of the last criticality.
[c] The option shown for each plant is the option that has been officially provided to the U.S. NRC. Plants in DECON may have had a short (1 to 4 years) SAFSTOR period. Likewise, plants in SAFSTOR may have performed some DECON activities or may have transitioned from the storage phase into the decontamination and dismantlement phase of SAFSTOR.
[d] These plants have recently performed or are currently performing the decontamination and dismantlement phase of SAFSTOR.
[e] Fuel to dry storage in an on-site independent spent fuel storage installation (ISFSI).

Three of the 22 plants (Fort St. Vrain, Shoreham, and Pathfinder) have completed decommissioning. Two of these three (Fort St. Vrain and Shoreham) used the DECON process for decommissioning. One facility, Shoreham, operated less than three full-power days before being shut down and decommissioned so there was relatively little contamination. One other facility, Pathfinder, was placed in SAFSTOR and subsequently decommissioned. Eleven of the plants shut down prematurely. Three Mile Island, Unit 2, ceased power operations as a result of a severe accident. Three Mile Island, Unit 2, has been placed in a monitored storage mode until Unit 1 permanently ceases operation, at which time both units are to be decommissioned.

Eleven of the permanently shutdown plants were part of the U.S. Atomic Energy Commission (AEC) Demonstrations Program, including Big Rock Point; Dresden, Unit 1; Fermi, Unit 1; GE-VBWR; Humboldt Bay, Unit 3; Indian Point, Unit 1; La Crosse; Pathfinder; Peach Bottom, Unit 1; Yankee Rowe; and Saxton. These plants were prototype designs that were jointly funded by the AEC and commercial utilities. The most recent of the Demonstration Program reactors to shut down was Big Rock Point, which operated for 33 years; it was permanently shut down in 1997.

Seven decommissioning facilities are located on multiunit sites in which the remaining units continue to operate, and one multiunit site shut down both units permanently. All eight of these licensees chose SAFSTOR as the decommissioning option. In most cases, SAFSTOR was chosen so that all units on a site could be decommissioned simultaneously. For various reasons, however, most shutdown units have done some decontamination and dismantlement. The reasons cited for choosing DECON have included the availability of low-level waste capacity, availability of staff familiar with the plant, available funding, the licensee's intent to use the land for other purposes, influence by state or local government to complete decommissioning, or a combination of other reasons.

A number of the plants have combined the DECON and SAFSTOR process either by entering shorter SAFSTOR periods or by doing an incremental DECON, allowing the plant to use resources and "decommission as they go." Sites have combined the options, usually to achieve economic advantages. For example, one site decided to shorten the SAFSTOR period and to begin incremental dismantlement out of concern over future availability of a waste site and future costs of disposal. One site that prematurely shut down had a short SAFSTOR period to allow short-lived radioactive materials to decay and to conduct more detailed planning. Safety is another reason for combining the two options. Because of seismic safety concerns, one site undertook a major dismantling project to remove a 76-m (250-foot) concrete vent stack after it had been in SAFSTOR for 10 years.

The owner of the facility determines the physical condition of the site after the decommissioning process. Some intend to restore the site to "green field" status at the end of decommissioning, whereas others may install a nonnuclear facility. Some will leave structures standing at the time of license termination, and others will not. While undergoing the decommissioning process, some have opted for partial site release to decrease the size of the site area.

6. CONCLUSION

There are environmental impacts of decommissioning a nuclear power facility. However, of the different options and processes for decommissioning a nuclear power facility, all of the options used in the United States to date have resulted in environmental impacts that are smaller than those that occur during construction and operation of the facility.

SEE ALSO THE FOLLOWING ARTICLES

Nuclear Engineering • Nuclear Fission Reactors: Boiling Water and Pressurized Water Reactors • Nuclear Fuel: Design and Fabrication • Nuclear Fuel Reprocessing • Nuclear Fusion Reactors • Nuclear Power Economics • Nuclear Power, History of • Nuclear Power: Risk Analysis • Nuclear Proliferation and Diversion • Nuclear Waste • Occupational Health Risks in Nuclear Power • Public Reaction to Nuclear Power Siting and Disposal

Further Reading

Nordhaus, W. D. (1997). "The Swedish Nuclear Dilemma: Energy and the Environment." Resources for the Future Press, Washington, D.C.

Slobodien, M. J. (ed.). (1999). "Decommissioning and Restoration of Nuclear Facilities: 1999 Health Physics Society Summer School Proceedings." Medical Physics Publ. Corp, Madison, Wisconsin.

U.S. Nuclear Regulatory Commission (NRC). (2002). "Generic Environmental Impact Statement on Decommissioning of Nuclear Facilities." NUREG-0586, Supplement 1. NRC, Washington, D.C.

U.S. Nuclear Regulatory Commission (NRC). (1997). "Final Generic Environmental Impact Statement in Support of Rulemaking on Radiological Criteria for License Termination of NRC-Licensed Nuclear Facilities." NUREG-1496, Vol. 1. NRC, Washington, D.C.

White, M. G., Norman, N. A., and Thomas, R. G. (eds.). (1995). "Decommissioning, Decontamination, and Environmental Restoration at Contaminated Nuclear Sites. American Nuclear Society, Winter Meeting 1994." American Nuclear Society, Washington, D.C.

Nuclear Power: Risk Analysis

B. JOHN GARRICK
Independent Consultant
Laguna Beach, California, United States

1. **Why Risk Assessment?**
2. **Nuclear Safety Theme**
3. **Historical Development of Nuclear Plant Safety**
4. **Nuclear Power Accident Experience**
5. **Risk Assessment Methodology**
6. **Important Applications and Benefits**
7. **Risk Assessment and Regulatory Practice**
8. **Future Direction of Risk Assessment**

Glossary

core The nuclear fuel and fission reaction region in a nuclear reactor.
core damage (melt) An accident in which the nuclear fuel is damaged beyond recovery.
defense-in-depth A combination of multiple barriers and design basis accident analysis that reasonably assures public health and safety and protection of the environment.
design basis accident An accident forming part of the basis for the design of a nuclear power plant, for which there will be reasonable assurance that there will be no damage to the nuclear fuel for that accident.
emergency core cooling system A separate and independent cooling system for the reactor core in the event of a "loss-of-coolant accident."
event tree A cause-and-effect representation of logic involving inductive reasoning.
fault tree An effect-and-cause representation of logic involving deductive reasoning.
nuclear power plant A commercial electric power-generating station that employs a nuclear reactor as the basic source of energy.
nuclear reactor An energy-generating system based on nuclear fission and a controlled self-supporting chain reaction.
pinch points The interfacing input and output states of the three models (plant, containment, site) that make up a full-scope risk assessment for a nuclear power plant.
plant damage states The output of the plant risk model representing the end states of accident scenarios that are a threat to the containment system.
quantitative risk assessment A process of systems analysis that calculates the likelihood and consequences of undesirable events and their uncertainties.
risk Answers the three questions about what can go wrong, how likely it is, and what the consequences are.
safety train A separate and independent set of engineered safety features for mitigating accidents.
scenario A sequence of events that describes the course of a nuclear plant accident.
scram The sudden shutdown of a nuclear reactor, usually by the rapid insertion of safety rods.
secondary containment An enclosure around a nuclear reactor to provide added protection from the release of radiation in the event of an accident that fails the primary containment system.
single-failure criteria A design criteria whereby the failure of a single system or piece of equipment will not result in any health and safety consequences.
trip An automatic or manual action that shuts down a system such as the reactor core or major piece of equipment, e.g., a turbine that provides the shaft power to the electric generators.

Nuclear power plant safety has been the principal driver for contemporary methods of quantitative risk assessment. Important contributions include a general definition of risk, methods for embracing and quantifying uncertainties, and the importance ranking of contributors to risk—an essential input for quantitative risk management. Most importantly, the adoption of risk assessment practices by the nuclear power industry coincides with an era of unprecedented safety in the performance of nuclear power plants.

1. WHY RISK ASSESSMENT?

The simple answer to "why risk assessment?" for nuclear power plants is that nations and the world have to make decisions about the best energy mix for

 © 2004 Elsevier Inc. All rights reserved.

the future of planet Earth. Risk to people and the environment is a fundamental attribute of societal decision making. But there is more to it than just decision making. Early in the development of nuclear power, it became clear that large inventories of radiation required a level of safety analysis beyond standard practices. Nuclear reactors were being contemplated for use in generating electricity, and safety was a concern, especially in light of the stigma of the dangers of the fission process carried over from nuclear weapons development. The nuclear power industry was forced to seek new methods of safety analysis of nuclear power plants to overcome the "fear anything nuclear" syndrome that prevailed in the minds of some members of the public. The new methods needed to provide answers to three questions: what can go wrong with a nuclear power plant, how likely is it, and what are the consequences? The traditional methods of safety analysis, although somewhat effective in answering questions about what can go wrong and the consequences, profoundly failed to answer the question having to do with the likelihood of accidents. The "likelihood" question held the key for being able to quantify nuclear power plant risk. In short, for society to have access to nuclear energy systems that have the potential to end anxieties about energy resources, the industry was forced to come up with a more convincing safety case than was possible with past methods of analysis.

The nuclear electric power industry has been the leader in the development and widespread use of quantitative risk assessment (QRA). The U.S. nuclear electric power industry gave birth to the term "probabilistic risk assessment" (PRA); the international nuclear community sometimes uses the equivalent term "probabilistic safety assessment" (PSA). The concept that appears to be best received across different industries is that of quantitative risk assessment. In this discussion, quantitative risk assessment, probabilistic risk assessment, and just plain "risk assessment" are used interchangeably.

Risk assessment has survived and flourished in the U.S. nuclear power industry because it is an exceptional tool to make better decisions. QRA was able to satisfy the desire of nuclear plant owners to have a decision tool that quantitatively allows the evaluation of various options that have multiple input variables. The most important variables to the nuclear plant owners are cost, generation, and risk (public health, worker health, and economic). Although QRA started out as a tool to address the public health risk, it facilitated evaluating an entire spectrum of variables. The industry's recovery from the Three Mile Island Unit 2 accident in 1979 was greatly aided by the use of quantitative risk assessment because of the ability to better focus on the real safety issues. In fact, the industry has had an impeccable safety record since embracing contemporary methods of quantitative risk assessment, and safety is not the only benefit that has resulted from the widespread use of risk assessment in the nuclear power industry. Risk assessment provides the ability for plant personnel to balance cost, generation, and risk. Although there is no U.S. Nuclear Regulatory Commission (NRC) requirement for an existing nuclear power plant to maintain a risk assessment, the plants do so, following general industry guidelines. The NRC does require a prospective licensee to submit a PRA with the application for any proposed new nuclear electric power unit in the United States.

2. NUCLEAR SAFETY THEME

Today, nuclear power plant safety analysis employs the most advanced methods available for assessing the health and safety of the public. Many of the methods used for nuclear power plants have been adopted by high-technology industries such as those involved in space flight, defense systems, chemical plants, refineries, offshore platforms, and transportation systems. The probabilistic concepts currently spearhead the level of sophistication of the analyses, but there are basic tenets and themes that have guided the safety management of nuclear electric power plants from the beginning. The most fundamental of these basic tenets is the concept of multiple barriers.

Multiple barriers are a concept of providing enough barriers between radiation and the environment to provide assurance that the likelihood of simultaneous breach of all barriers is remote. Examples of barriers in a nuclear power plant are high-containment-capacity fuel with cladding, an isolated reactor coolant system, primary reactor building containment, secondary building containment, and exclusion distance. Other defense mechanisms include automatic control systems, single-failure criteria (no single failure threatens fuel integrity), and recovery capabilities from equipment malfunctions. QRA provides the ability to determine what risk levels are achieved by each barrier and at what cost. The value of each barrier is placed in the context of the overall risk. A principal factor in implementing multiple barriers is choosing design basis accidents to evaluate the multiple barriers. The principle behind

the design basis accident is the requirement that the plant design incorporate the capability to withstand specific hypothetical initiating events and failures without causing damage to the nuclear fuel. This process of multiple barriers and design basis accidents is referred to by the NRC as defense-in-depth.

The defense-in-depth concept has generally been implemented through the promulgation of very specific deterministic regulations. As the nuclear regulations come under review for risk-informing the regulatory process, it is not expected that the multiple-barriers tenet of regulatory practice will change in kind, but it will change in degree with much better knowledge of the real plant risk. In particular, with the availability of much more advanced methods for calculating the value of protective barriers, the ability exists to optimize the barriers in terms of risk and cost. Furthermore, the design basis concept can be assessed for its usefulness. This is considered important, as past application of defense-in-depth in the absence of a risk-informed approach has resulted in a definite increase in the complexity and cost of nuclear power plants without commensurate improvement to the overall public health risk. Quantifying defense-in-depth is one of the most significant benefits of QRA.

3. HISTORICAL DEVELOPMENT OF NUCLEAR PLANT SAFETY

Nuclear plant safety has two major fronts—the physical system and the analysis of the physical system. On the physical system front, improvements in safety design have included the advent of secondary containment systems ($\sim$1953), the inclusion of backup safety systems known as engineered safety features, especially with respect to emergency core cooling systems and electric power ($\sim$late 1950s and early 1960s), and the introduction of separate and independent safety trains ($\sim$1970s). In the 1980s and 1990s, nuclear power plants initiated programs for scram (sudden reactor shutdown) reduction based on a complete review and analysis of operating transients. As scrams were reduced, public health risk was reduced because there were fewer departures from normal steady-state operation. Also in the 1980s and 1990s, each nuclear power plant implemented the concept of "symptom-based procedures" for accident control and installed improved simulators for operator training.

On the analysis front, many events took place leading to a greatly improved understanding of the safety of nuclear power plants. It was demonstrated that the consequences of accidents had little meaning without a better understanding of their likelihood. It became clear that it was not enough to do worst-case and maximum-credible accident analysis. Everyday transients followed by multiple failures of equipment and mistakes by operators were more likely than design basis accidents to result in reactor core damage.

The need for probabilistic analysis was recognized as early as the mid-1950s. However, detailed investigations of the probability of reactor accidents did not begin until about 1965. The first major reactor safety study to highlight the need for PRA of reactor accidents was the 1957 U.S. Atomic Energy Commission report WASH-740, "Theoretical Possibilities and Consequences of Major Accidents in Large Nuclear Power Plants." Speculative estimates were made in WASH-740 that a major reactor accident could occur with a frequency of about one chance in a million during the life of a reactor. The report went on to observe that the complexity of the problem of establishing such a probability, in the absence of operating experience, made these estimates subjective and open to considerable error and criticism. Although it did not offer many specifics, this study did create interest in probabilistic approaches, and many studies were soon to follow. These included British and Canadian efforts, probabilistic analyses of military reactors, and several studies sponsored by the U.S. Atomic Energy Commission. At about the same time, Garrick wrote a Ph.D. thesis on unified systems safety analysis of nuclear power plants based on a total systems and probabilistic approach.

The breakthrough in the probabilistic risk assessment of nuclear power plants came in 1975 with the publication of the Reactor Safety Study by the U.S. Nuclear Regulatory Commission under the direction of Professor N. C. Rasmussen of the Massachusetts Institute of Technology. This project marked a turning point in the way people analyzed the safety of complex facilities and systems. The Reactor Safety Study was followed by several major industry studies, such as the risk assessments performed on the Zion and Indian Point nuclear power plants; the new methods that were introduced in these assessments have become standards of many QRA applications. These studies provided a major breakthrough in calibrating the worth of safety features and safeguards, and therefore, the safety margins of designs.

By the 1980s, the question was no longer "why," but how soon a QRA could be developed for every nuclear power plant in the United States. That goal

has essentially been reached. The benefits of QRA for U.S. nuclear power plants have been demonstrated in terms of a reduction in frequency of core damage events (one reactor core lost in approximately the first 450 reactor years of experience versus zero reactor cores lost in over 2000 actual reactor years of experience since the Three Mile Island accident) and improved generation with a reduction in the cost of electricity. The most important benefit is nuclear power plants with reduced public health risk. QRA has been effective not only in calibrating the risk of nuclear power, but has provided better knowledge of the worth of safety systems and allowed the allocation of safety engineering resources to the most important contributors. Effective risk management of nuclear electric power plants in the United States has become a reality, not just a goal.

4. NUCLEAR POWER ACCIDENT EXPERIENCE

There have only been two accidents worldwide that have resulted in severe core damage of a nuclear power plant designed to generate electricity. The accidents involved the Three Mile Island Unit 2 plant near Harrisburg, Pennsylvania, in the United States, and the Chernobyl Nuclear Power Station in the Ukraine of the former Soviet Union. Both accidents permanently damaged the nuclear reactors involved, but only the Chernobyl accident resulted in known fatalities and injuries. The on-site consequences of the Chernobyl accident were very serious; an estimated 30 people are believed to have died from acute doses of radiation and some 300 people required hospital treatment for radiation and burn injuries. No off-site fatalities or injuries have yet been attributed to the Chernobyl accident, although the latent effects are yet to be quantified. It is important to put these two very serious accidents in context with the safety experience of the nuclear power industry. There are approximately 440 nuclear power plants in the world. Nuclear energy is just over 5% of the world primary energy production and about 17% of its electrical production. In the United States, there are 103 nuclear power plants operating, providing approximately 20% of the nation's electricity. The worldwide experience base is approaching 10,000 in-service reactor-years, of which about 3000 reactor-years is U.S. experience. The experience base is likely beyond 10,000 reactor-years if all types of reactors are included, such as research, test, weapons, and propulsion reactors. Some 70% of the nuclear power plant experience worldwide involves light water reactors, for which only one accident has occurred, Three Mile Island. This safety record is most impressive. The challenge is to keep it that way.

4.1 The Three Mile Island Unit 2 Accident

The Three Mile Island Unit 2 (TMI-2) nuclear power plant, located near Harrisburg, Pennsylvania, went into commercial operation in December 1978. The plant was designed to generate approximately 800 MW of electricity and used a pressurized water reactor supplied by the Babcock and Wilcox Company. The accident occurred on March 28, 1979. Routine mechanical malfunctions with the plant resulted in an automatic shutdown ("feedwater trip") of the main feedwater pumps, followed by a trip of the steam turbine and the dumping of steam to the condenser. The loss of heat removal from the primary system resulted in a rise of reactor system pressure and the opening of the power-operated relief valve. This action did not provide sufficient immediate pressure relief, and the control rods were automatically driven into the core to stop the fission process.

These events would have been manageable had it not been for some later problems, such as with the emergency feedwater systems. Perhaps the turning point of the accident was that the opened pressure relief valve failed to close and the operators failed to recognize it. The result was the initiation of the well-studied small loss-of-coolant accident, known as the small LOCA. The improperly open valve, together with some other valve closures that had not been corrected from previous maintenance activities, created a shortage of places to put the heat loads of the plant. The response of the plant was the initiation of high-pressure emergency cooling. High pump vibration and concern for pump seal failure resulted in the operators eventually shutting down all of the main reactor coolant pumps. It was during the time that the coolant pumps were off, for 1 to 3 hours, that the severe damage to the core took place. At about 2 hours and 20 minutes into the accident, the backup valve (known as a block valve) to the stuck-open relief valve was closed. This action terminated the small LOCA effect of the stuck-open relief valve. Although the accident was then under some level of control, it was almost 1 month before complete control was established over the reactor fuel temperature, when adequate cooling was provided by

natural circulation. The consequences of the accident were minimal in terms of the threat to public health and safety, but the damage to the reactor was too severe to recover the plant.

4.2 Chernobyl Nuclear Power Station Accident

The Chernobyl nuclear power plant in the former Soviet Union involved a 1000-MW (electrical) boiling water, graphite-moderated, direct-cycle reactor. The Chernobyl accident occurred on April 26, 1986, and was initiated during a test of reactor coolant pump operability from the reactor's own turbine generators. The purpose of the test was to determine how long the reactor coolant pumps could be operated, using electric power from the reactor's own turbine generator under the condition of turbine coast down and no steam supply from the reactor. However, the experimenters wanted a continuous steam supply, so they decided to conduct the experiment with the reactor running—a serious mistake. The test resulted in a coolant flow reduction in the core and extensive boiling. Because of the inherent properties of this reactor design, the chain reaction increases on boiling, rather than decreases as in U.S. plants, and a nuclear transient occurred that could not be counteracted by the control system. The result was a power excursion that caused the fuel to overheat, melt, and disintegrate. Fuel fragments were ejected into the coolant, causing steam explosions and rupturing fuel channels with such force that the cover of the reactor was blown off. This accident resulted in approximately 30 fatalities from acute doses of radiation and the treatment of some 300 people for radiation burn injures. The off-site consequences are still under investigation. Latent effects are expected, but they have not been quantified.

In summary, nuclear power suffered a severe setback from both of these accidents, although public support for nuclear power was already beginning to decline. Nuclear plants under construction were canceled and no new U.S. nuclear plants have been ordered since 1979. The fact that the TMI-2 accident did not result in any radiation injuries or fatalities and the Chernobyl reactor type is no longer in the mix of viable power reactors has not removed the fear that some segments of the public have of nuclear power. However, the superior performance and safety record in the United States since these two accidents has allowed the NRC to approve power upgrades and license extensions for several U.S. nuclear power plants.

5. RISK ASSESSMENT METHODOLOGY

Quantitative risk assessments of nuclear power plants are generally based on the following principles:

- The quantitative expression of risk should be in the form of a structured set of scenarios, each having a corresponding likelihood and consequence.
- Each risk scenario should take the form of a sequence of events, starting with an event that upsets an otherwise successful operation or system, and proceeding through a series of subsequent events to the end state that terminates the scenario (i.e., the consequences of the scenario).
- The set of scenarios must be complete in the sense that all of the important contributors to risk are included.
- The end states of the scenarios should reflect initial, cascading, and collateral consequences or levels of damage where appropriate.
- The scenarios must be quantified in terms of clearly defined risk measures, be realistic, incorporate uncertainties, and be based on the supporting evidence.
- The results should rank the contributors to risk in order of importance and must be presented in a way that supports decision making.

The overarching principle on which risk assessment methodology for nuclear power plants is founded is that when we ask the question "what is the risk?" we are really asking the following three questions: (1) what can go wrong, (2) how likely is it, and (3) what are the consequences? In the notation of the practitioners of risk assessment, this "triplet" definition of risk by Kaplan and Garrick is represented as follows:

$$R = \{(S_i, L_i, X_i)\}_c,$$

where S_i denotes risk scenario i, L_i denotes the likelihood of that scenario, and X_i denotes the consequences or damage level of that scenario. The angle brackets enclose the triplets, the curly brackets mean "a set of," and the subscript c denotes complete, meaning that all of the important scenarios are included in the set.

An overview of the basic structure of a nuclear power plant quantitative risk assessment is shown in Fig. 1. There are basically three models within a full-scope quantitative risk assessment—the plant model, the containment model, and the site model. The plant model begins with the consideration of different

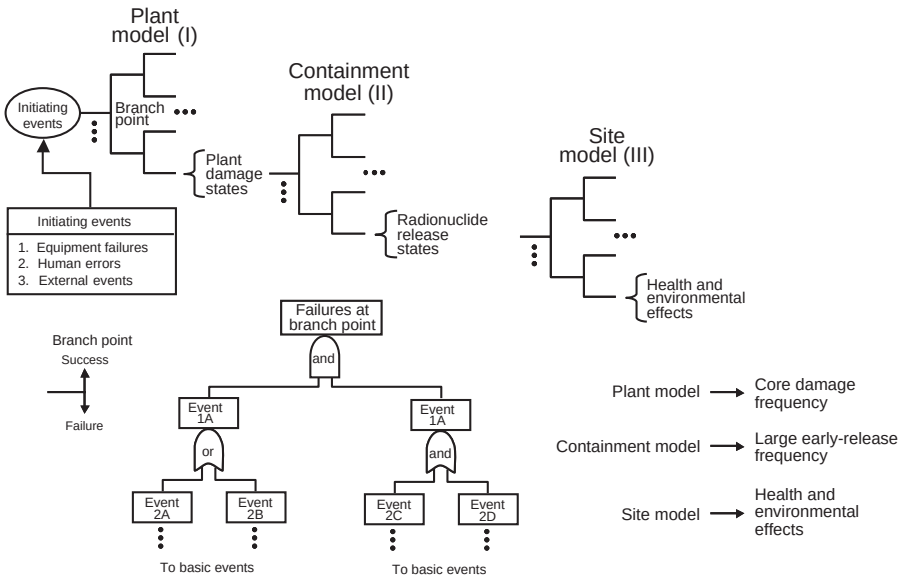


FIGURE 1 Event tree/fault tree structure of a nuclear plant quantitative risk assessment.

initiating event categories and has as output different plant damage states. The plant damage states are input to the containment model, and its outputs are different release states that become the input to the site model. The output of the site model is the calculated health and environmental risks. The interfaces between the models are "pinch-points" that allow the models to be developed independently, which greatly facilitates the organization and transparency of the analysis. Figure 1 also illustrates what is described in the next section as the "event tree," "fault tree" format for structuring scenarios. A consulting firm, Pickard, Lowe and Garrick, Inc., developed the modular event tree structure for quantitative risk assessment, a variation of the Reactor Safety Study methodology. The U.S. Nuclear Regulatory Commission and the American Nuclear Society have adopted the same three-model structure in their procedures guide on quantitative risk assessment. They identified the plant, containment, and site models cumulatively as Levels I, II, and III, respectively.

5.1 Structuring the Scenarios (S_i)

Scenario structuring encompasses the methods, algorithms, and insights needed to identify and portray the risk scenarios (S_i). Two common methods (Approach 1 and Approach 2) are used for scenario development and are sometimes referred to as the bottom-up and top-down approaches. They have the following characteristics:

1. Given a set of initiating events, the structuring of scenarios is done so the end state of each scenario is the condition that terminates the scenario. That is, the scenario determines the end state.
2. Given an end state, project backward to determine the potential scenarios that could occur to arrive at that end state.

The most common logic diagrams for representing the two methods are event trees and fault trees. An event tree starts with an initiating event and proceeds to identify the succeeding events, including branches, that eventually terminate into possible undesirable consequences. An event tree, therefore, is a cause-and-effect representation of logic. Event trees are the logic diagrams of preference for Approach 1. A fault tree starts with the end state or undesired consequence of interest and attempts to determine all of the contributing system states. Therefore, fault trees are effect-and-cause representations of logic and are the logic diagram preference for Approach 2. That is, an event tree is developed by inductive reasoning whereas a fault tree is developed by deductive reasoning. A key difference in the two representations is that a fault tree is only in "failure space" and the event tree includes both "failure and success space." The choice between the two is a matter of circumstances and preference. Often the two are used in combination such that the event tree provides the basic scenario space of events and branch points, and the fault trees are used to quantify

the "split fractions" at the branch points as illustrated in Fig. 1.

5.2 Defining Consequences (X_i)

Simply put, consequences are the end states of the scenarios. There are two perspectives of consequences and both are used in analyzing the risk of nuclear power plants. One perspective is to take each risk scenario to whatever point that is a reasonable termination of the scenario. The scenarios are then assembled by different consequences, or damage states. The other perspective is to define the damage state of interest in advance, such as "core melt," "radionuclide releases," "radiation dose to humans," "injuries," "fatalities," and "property damage," and consider only the scenarios that have as their end state the undesired damage state(s). Although examples of both approaches exist, current practice is to focus on "core melt" as the primary basis for measuring risk because core melt is a precursor to large radiation releases. One other damage state has been defined to serve as a surrogate for consequences beyond core melt, and that is "a large early-release frequency" of radiation. The choice is dependent on the requirements provided to the analyst as to how to structure the scenarios and end states.

5.3 Quantifying the Scenarios (L_i)

To quantify the likelihood, L_i, of different accident scenarios, it is first necessary to define the concept of likelihood. Most often the methodology adopts the "probability-of-frequency" principle to define likelihood. The frequency parameter is presented as a probability distribution to communicate frequency uncertainty. The actual quantification of the risk scenarios is done with the aid of the event tree (see Fig. 2). The event tree branch points are determined by actions, activities, and equipment (top events) that can alter or truncate the path of a scenario or sequence of events. An example of an action that could alter the path of a scenario is a decision by a reactor operator to shut down a cooling system. An example of an activity might be the activation of a mitigating system, such as an emergency coolant system, and an example of an equipment event might be the failure of a source of electric power. Top events are placed in the boxes across the top of the diagram (Fig. 2) and are denoted A, B, C, and D. The event tree is a powerful tool because it makes visible all of the actions, equipment, processes, events, and features that affect an event. The diagram shown in Fig. 2 has only two outcomes emerging from a branch point (e.g., success or failure). However, an event tree can have multiple outcomes from a branch point to account for different degrees of degradation of a system. An individual scenario is a single path through the tree as illustrated by the highlighted lines.

Each scenario or path through the event tree can be described by an algebraic expression (shown in Fig. 2). Using input data for the initiating event and evidence-based split fractions at the branch points, the algebraic expression can be converted to an equation for calculating the frequency of individual scenarios. The remaining step is to embed the frequencies into appropriate probability distributions to communicate their uncertainties. There are various techniques for carrying out this operation, but the one often preferred is based on Bayes' theorem. Bayes' theorem is the fundamental, logical principle governing the process of inferential reasoning. It answers the question: "How does the probability of a given hypothesis change when we obtain a new piece of evidence?"

Once the scenarios have been quantified, the results take the form of the graph in Fig. 3. Each scenario has a probability-of-frequency curve quantifying its likelihood of occurrence. Figure 3 shows the curve for a single scenario or a set of scenarios

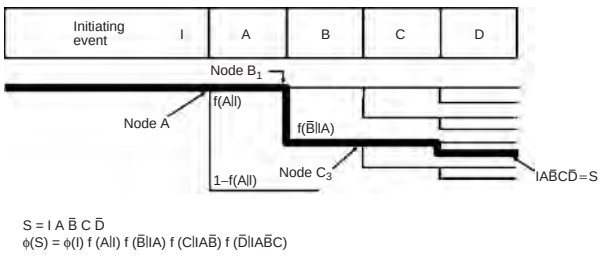


FIGURE 2 Quantification of a scenario using an event tree.

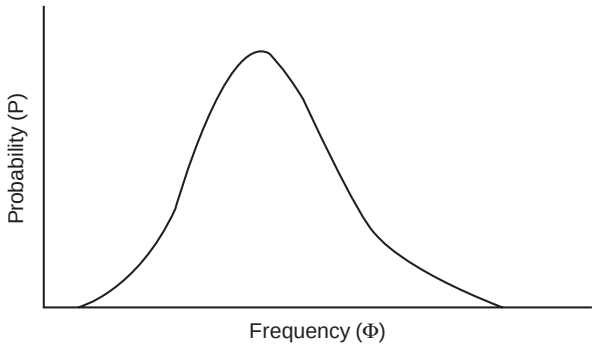


FIGURE 3 Probability-of-frequency curve for a specific consequence.

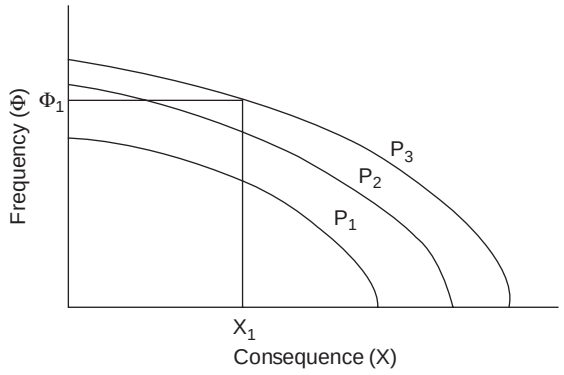


FIGURE 4 Risk curve for varying consequences, where the consequence is a variable.

leading to a single consequence. Showing different levels of damage, such as the risk of varying injuries or fatalities, requires a different type of presentation. The most common form is the classical "risk curve," also known as the "frequency-of-exceedance" curve, or the even more esoteric label, the "complementary-cumulative-distribution function." This curve is constructed by ordering the scenarios by "increasing levels of damage" and cumulating the probabilities from the bottom up in the ordered set against the different damage levels. Plotting the results on log–log paper generates curves, as shown in Fig. 4.

Although risk assessment results such as those illustrated in Figs. 3 and 4 can be beneficial in providing a perspective on the actual risks and in prioritizing contributors to risk, they are not the most important output of the risk assessment. The most important output is the revelation of the dominant contributors to the risk, which are necessary for effective risk management. The contributors are buried in the results assembled to generate such curves as those in Figs. 3 and 4. Numerous techniques can be used to extract and rank contributors. Most advanced risk assessment software packages contain algorithms for ranking the importance of contributors to a risk measure.

6. IMPORTANT APPLICATIONS AND BENEFITS

Since the Reactor Safety Study in 1975, major government and industry studies have provided the signature of what is meant by probabilistic risk assessment as practiced in the nuclear power field. These studies and applications include the joint effort of the U.S. Nuclear Regulatory Commission and the American Nuclear Society in the development of the industry's PRA procedures guide, the collaborative industry effort on the Oconee nuclear plant risk assessment, and the NRC's evaluation of five light water reactor designs culminating in the 1990 report, NUREG-1150. But the studies most responsible for specializing the Reactor Safety Study methodology to plant-specific risk assessments were the industry-sponsored Zion and Indian Point studies. The two efforts, the Reactor Safety Study and the Zion and Indian Point studies, are now discussed further.

It was the U.S. Atomic Energy Commission that undertook the Reactor Safety Study under the direction of Professor Norman C. Rasmussen of the Massachusetts Institute of Technology. The study took 3 years to complete and was a turning point in the way to think about the safety of nuclear power plants, or, for that matter, the safety of any natural or engineered system. It should be noted that the Reactor Safety Study was initiated before the Energy Reorganization Act of 1974 abolished the U.S. Atomic Energy Commission. This Act transferred to the Nuclear Regulatory Commission all the licensing and related regulatory functions assigned to the Atomic Energy Commission by the Atomic Energy Act of 1954.

The Reactor Safety Study, using the Surry nuclear power plant (pressurized water reactor) and the Peach Bottom nuclear power plant (boiling water reactor) as reference designs, calculated the risk from the operation of 100 current-design light water reactors located in the United States. The methodology was founded on the principle of a structured set of accident sequences, or scenarios. The sequences were developed using fault tree and event tree logic diagrams, quantified on the bases of the supporting evidence, and then assembled into different consequences such as core damage frequencies, radiation release fractions, and off-site radiation effects to people and property. The principal findings of the study were that the risk associated with the operation of selected nuclear power plants is extremely small and that the dominant contributor to risk is not the large loss-of-coolant accident, previously emphasized as the design basis accident. Transients and small loss-of-coolant accidents often are the major contributors to risk. The Reactor Safety Study also highlighted the important role of the reactor operators in maintaining the safety of nuclear power plants.

The public and scientific community had mixed reactions to the Reactor Safety Study. Their primary criticism of the study was that the "uncertainty analysis" was weak and the report lacked "scrutability." The initial reaction of the NRC to the

criticism was to withdraw their endorsement of the study. In spite of the criticisms of the study, there was strong support for the use of the methodology, especially from the plant owners. This favorable response, together with the fact that the Three Mile Island accident was among the event sequences addressed in the study, led the NRC to change their position and once again support the study.

The nuclear power industry was responsible for carrying the lessons learned from the Reactor Safety Study to the practical level of better understanding the safety of individual nuclear plants. Several studies followed on the heels of the Reactor Safety Study and the most comprehensive of these were the risk assessments performed on the Zion nuclear plant near Chicago and the Indian Point 2 and 3 plants near New York City. Under legal challenges at high-population-density sites, industry chose probabilistic risk assessment as a way to develop the necessary evidence of the safety of their plants. The Zion and Indian Point plants became, more or less, the test cases. The Zion and Indian Point assessments were full-scope studies, meaning that they analyzed the plant, its containment, and the off-site consequences. Challenged in the courts, the owners and operators of the plants defended the safety of their plants using the risk assessments as their primary evidence. They were successful. The conclusion was reached that the nuclear power plants at the high-population-density sites presented public health risk profiles similar to those at less populated sites. As a result, the nuclear power plants at the high-population-density sites were not shut down—a major achievement at the time.

The Zion and Indian Point studies contained many firsts, including the "triplet definition" of risk; they were the first comprehensive studies of core melt phenomena and containment response in a probabilistic format and were the first to employ a modularized event tree "pinch-point" format to represent beginning-to-end accident sequences, the first to perform uncertainty analysis at the component and basic event levels and to propagate the uncertainties through the scenarios, the first to explicitly include external events (such as earthquakes and fires) in the basic risk model, and the first to employ an atmospheric dispersion model for dose calculations that allowed for changes in plume direction. Other analytical concepts introduced by the Zion/Indian Point studies included the "probability-of-frequency" format for measuring risk and the "master logic diagram" method for determining initiating events. Many of these methods have become standards for contemporary risk assessments.

7. RISK ASSESSMENT AND REGULATORY PRACTICE

As a result of the Energy Reorganization Act of 1974, the U.S. Nuclear Regulatory Commission is responsible for licensing and regulating nuclear facilities, including nuclear power plants, and materials, and for conducting research in support of the licensing and regulatory process. The NRC's primary regulatory responsibility is to provide reasonable assurance of adequate protection of public health and safety and protection of the environment from operations and accidents involving nuclear facilities and materials. The legacy of the NRC in the development and use of risk assessment is unique for regulatory bodies. The vision of the NRC of the need for probabilistic methods was reflected in their decision to sponsor the Reactor Safety Study when such an evaluation was not required by the regulations.

During the time when the Reactor Safety Study was being reviewed and the NRC had temporarily rejected it due to early criticisms, the nuclear power industry was taking the initiative to use the methodology for better assessing the risk of its plants. The impressive results coming out of the industry studies together with other events favorable toward the Reactor Safety Study caused the NRC to again embrace the technology and to seek ways of effectively applying it. Some of the actions taken were the publishing of frequency-based safety goals(1986), requiring limited-scope individual plant examinations based on probabilistic methods to determine if any nuclear power plant was an "outlier" with respect to public health risk (1988), and, in about this same time frame, issuing two new rules having to do with the treatment of loss of all electrical power and requirements to reduce the risk of transients. Perhaps the most significant action taken by the NRC toward embracing the concept of quantitative risk assessment was the 1995 publishing of a policy statement on the use of probabilistic risk assessment methods in nuclear regulatory activity. Quoting from the PRA policy statement, "The use of PRA technology should be increased in all regulatory matters to the extent supported by the state of the art in PRA methods and data, and in a manner that complements the NRC's deterministic approach and supports the NRC's traditional defense-in-depth philosophy."

Even with what appeared to be an aggressive move on the part of the NRC toward encouraging the use of probabilistic risk assessment, there were only small changes in the regulations with respect to

the licensing process. The NRC recognized this problem and beginning in 1997 sought ways to make changes to the regulations to begin a more formal transition into a "risk-informed" approach to regulation. Several initiatives were put in place to stimulate risk-informing the regulatory process. Two examples are new rules having to do with the treatment of loss of all electrical power and requirements to reduce the risk of transients. Meanwhile, it was important for the NRC to make it clear what they mean by "risk-informed" regulation. The best answer to that came from a white paper prepared by the Commission in 1998 on "Risk-Informed, Performance-Based Regulation." Quoting from the white paper, "A risk-informed approach to regulatory decision-making represents a philosophy whereby risk insights are considered together with other factors to establish requirements that better focus licensee and regulatory attention on design and operational issues commensurate with their importance to health and safety." The current position of the NRC is a "risk-informed" approach to regulation. It is made very clear in the white paper "that the Commission does not endorse an approach that is risk-based," if what is meant by risk based is that safety decisions are solely based on the numerical results of a risk assessment.

What all this means is that the licensing process is in a transitional phase, from deterministic and prescriptive regulations, to regulations that are less prescriptive and increasingly risk oriented, but not risk based. In the meantime, license applications and amendments must be accompanied with analyses that provide risk insights, but are in compliance with the deterministic requirements still in place.

8. FUTURE DIRECTION OF RISK ASSESSMENT

The future direction of quantitative risk assessment in the risk management of nuclear power plants is dependent on several factors: advancements in understanding accident phenomena, upgrading of risk assessment methodologies, regulatory activities, and security requirements.

8.1 Advancements in Understanding Accident Phenomena

The current practice for nuclear power plant risk assessments is to make two basic calculations, core damage frequency and large-early-release frequency. In the future, calculations will need to include public health effects (immediate fatalities and latent cancer fatalities). These calculations of health effects will need to make use of the recent work with respect to realistic source terms for radiation releases. There is a lack of a dose–response model that reflects actual health effects for all levels and rates of radiation. Clearly, more accurate models of the real health effects are needed.

There is a need for realistic thermal hydraulic calculations that can form the basis of improved success criteria for the QRA models. In many cases, these success criteria are based on design basis accident calculations that do not effectively represent the performance capability of plant equipment. Design basis accident calculations also do not allow for the proper quantification of operator actions to activate equipment or to recover failed equipment. A realistic evaluation is needed of the thermal hydraulic interaction of the reactor coolant system and other fluid systems during accidents such as steam generator tube ruptures and intersystem loss of coolant accidents.

8.2 Upgrading of Risk Assessment Methodologies

QRA has greatly advanced since the breakthrough effort of the Reactor Safety Study. Improvements in the models include uncertainty analysis; the treatment of human reliability; consideration of external threats such as earthquakes, severe storms, and fires; importance ranking of contributors; and the fine-tuning of the models to better represent plant-specific details. There are still many areas for improvement. One very attractive direction would be for the risk assessments to be cast into different forms for use by different groups. The specialization could be not only for the risk assessment teams, but also for risk managers, those responsible for accident management and emergency response, and the public (risk communication). Each group has a different need for the information coming out of a risk assessment, and specializing the information by need could have a major impact on the acceptance and use of the results.

8.3 Regulatory Activities

The responsibility for the safety of the public in the United States with respect to radiation from nuclear reactors lies with the owners of the nuclear electric power plants. The owners of the nuclear units have

used their existing quantitative risk assessments to institute effective and efficient risk management programs at each nuclear power plant in the United States. Each nuclear power plant has a risk assessment that is maintained to industry guidelines. Engineers, maintenance persons, and operators at each nuclear power unit are aware of the results of the risk assessment and use them to actively manage risk. The superior safety record of nuclear electric power in the United States coincides with this new approach to risk management.

The NRC has a legal responsibility to "provide reasonable assurance of adequate protection of public health and safety." The QRAs have demonstrated that most of the deterministic-based existing regulations do not efficiently address the dominant contributors to public health risk. Some of them do not address the dominant contributors at all. The regulations must be changed such that the regulations aid both the regulator and the licensee to manage public health risk in an effective and efficient manner. The NRC is slowly moving toward regulations that address public health risk using the insights gained from the licensee QRAs. This is evidenced in the recent changes to the rules governing the monitoring of maintenance and the implementation of the Reactor Oversight Process. But the pace of change to "risk inform the regulations" has been slow. The best course is for industry to play an active role in the development of future regulations that are "risk informed" and for the NRC to allocate the necessary resources to quickly change the regulations to be effective and efficient.

8.4 Security Requirements

An important new area for receiving benefits from risk assessment has to do with combating terrorism. Nuclear plants are often mentioned as a possible target for a terrorist attack and it is important that such a threat is linked to the vulnerability of the plants. Of course, there has already been a lot of work done in this area for nuclear power plants and events such as aircraft impact have always been a consideration in the safety assessment of nuclear plants. It is just that the threat now seems more real than ever before, and the question is what risk assessment can do to help protect plants from such threats. Although a great deal is known about the vulnerability of nuclear plants, there has not yet been a systematic process of connecting such vulnerabilities to specific threats. This will require cooperation between the experts on the threat of terrorism, the intelligence community, and experts on risk assessment. Fortunately, some progress is being made, but it, too, is slower than it should be.

SEE ALSO THE FOLLOWING ARTICLES

Ecological Risk Assessment Applied to Energy Development • Nuclear Engineering • Nuclear Fuel: Design and Fabrication • Nuclear Fuel Reprocessing • Nuclear Fusion Reactors • Nuclear Power Economics • Nuclear Power Plants, Decommissioning of • Nuclear Proliferation and Diversion • Nuclear Waste • Occupational Health Risks in Nuclear Power • Public Reaction to Nuclear Power Siting and Disposal

Further Reading

Garrick, B. J. (1968). "Unified Systems Safety Analysis for Nuclear Power Plants." Ph.D. Thesis, University of California, Los Angeles.

Kaplan, S., and Garrick, B. J. (1981). On the quantitative definition of risk. *Risk Anal.* **1**(1), 11–27.

Pickard, Lowe and Garrick, Inc., Westinghouse Electric Corporation, and Fauske and Associates, Inc. (1981). "Zion Probabilistic Safety Study." Prepared for Commonwealth Edison Company, Chicago, Illinois.

Pickard, Lowe and Garrick, Inc., Westinghouse Electric Corporation, and Fauske and Associates, Inc. (1982). "Indian Point Probabilistic Safety Study." Prepared for Consolidated Edison Company of New York, Inc., and the New York Power Authority, New York.

U.S. Atomic Energy Commission (AEC). (1957). "Theoretical Possibilities and Consequences of Major Accidents in Large Nuclear Power Plants." Report WASH-740 (March, 1957). AEC, Washington, D.C.

U.S. Nuclear Regulatory Commission (NRC). (1975). "Reactor Safety Study: An Assessment of Accident Risks in U.S. Commercial Nuclear Power Plants." Report WASH-1400 (NUREG-75/014). NRC, Washington, D.C.

U.S. Nuclear Regulatory Commission (NRC). (1978). "Risk Assessment Review Group Report to the U.S. Nuclear Regulatory Commission"(H. W. Lewis, chairman). Report NUREG/CR-0400 (September, 1975). NRC, Washington, D.C.

U.S. Nuclear Regulatory Commission (NRC). (1983). "PRA Procedures Guide—A Guide to the Performance of PRAs for Nuclear Power Plants." Report NUREG/CR-2300 (January 1983). NRC, Washington, D.C.

U.S. Nuclear Regulatory Commission (NRC). (1990). "Severe Accident Risks: An Assessment for Five U.S. Nuclear Power Plants." Report NUREG-1150. NRC, Washington, D.C.

U.S. Nuclear Regulatory Commission (NRC). (1995). "The Probabilistic Risk Assessment (PRA) Policy Statement. 60 FR 42622 (August 16, 1995)." NRC, Washington, D.C.

U.S. Nuclear Regulatory Commission (NRC). (1999). "White Paper on Risk-Informed and Performance-Based Regulation. SECY-98–144 (February 24, 1999)." NRC, Washington, D.C.

Nuclear Proliferation and Diversion

H. A. FEIVESON
Princeton University
Princeton, New Jersey, United States

1. **Introduction: There Is No Such Thing as a Peaceful Atom**
2. **Nuclear Weapons**
3. **Fissile Material Production**
4. **Routes to Fissile Materials for Weapons**
5. **Nuclear Reactors and Fuel Cycles**
6. **Technical Barriers to Diversion**
7. **Institutional Barriers to Proliferation: Safeguards**
8. **Proliferation Resistance: An Initial Assessment**
9. **The Long-Term Challenge for Nuclear Power: Enhanced Proliferation Resistance**
10. **Conclusion**

Glossary

fissile material Material that can sustain an explosive chain reaction, notably plutonium of almost any isotopic composition and highly enriched uranium.
highly enriched uranium Uranium containing greater than 20% uranium-235.
nuclear proliferation The spread of nuclear weapons to states that do not now possess them or to subnational groups.
once-through fuel cycles A civilian nuclear fuel cycle in which the spent fuel from the reactor is not reprocessed and the contained plutonium separated.
proliferation resistance The array of measures that weaken the links between the civilian nuclear fuel cycle and the capabilities of a nation or subnational group to acquire nuclear weapons.
plutonium recycling The use of plutonium recovered from spent fuel in fresh fuel for reactors.
reprocessing The chemical separation of spent fuel into constituent parts, typically the contained plutonium, fission products, and uranium.
spent fuel The fuel discharged from a nuclear reactor.

This article describes the current system of safeguards on peaceful nuclear activities to prevent proliferation, gauges how well this system has succeeded and is likely to do so in the future, and examines various reactor and fuel cycle concepts that have been advanced to make nuclear power more "proliferation resistant" in the future.

1. INTRODUCTION: THERE IS NO SUCH THING AS A PEACEFUL ATOM

As early as 1904, in a talk to the Corps of Royal Engineers in England, the young scientist Frederick Soddy stated the fundamental dilemma of nuclear power in a now much-quoted passage:

> *It is probable that all heavy matter possesses—latent and bound up with the structure of the atom—a similar quantity of energy to that possessed by radium. If it could be tapped and controlled what an agent it would be in shaping the world's destiny! The man who put his hand on the lever by which a parsimonious nature regulates so jealously the output of this store of energy would possess a weapon by which he could destroy the earth if he chose.*

This dilemma—that the atom can be exploited for almost boundless energy or for almost boundless destruction—was given a more practical policy-directed twist by Enrico Fermi in 1944, soon after the Manhattan Project laid the foundations for both nuclear reactors and nuclear weapons:

"It is not clear that the public will accept an energy source that produces this much radioactivity and that can be subject to diversion of material for bombs."

It is now well understood that the fission of 3 kg of U-235 or plutonium could provide 1 day's worth of electricity for a moderate-size city or, alternatively, could generate an explosive yield of 50,000 tons of

 © 2004 Elsevier Inc. All rights reserved.

TNT, three times the destructive power of the bomb that destroyed Hiroshima. The question regarding nuclear power is whether it can continue to provide a bountiful source of energy for peaceful purposes without creating unacceptable risks of proliferation of nuclear weapons.

In this article, proliferation resistance means the panoply of measures to weaken the links between the civilian nuclear fuel cycle and the capability of a nation or a subnational group to acquire nuclear weapons. The links are several. Nuclear power programs require cadres of nuclear scientists and technicians, a network of research facilities, research reactors, and laboratories—some of which are indispensable to a nuclear weapons program and which could also serve as a cover to a parallel, dedicated weapons program. The most direct connection between nuclear power and nuclear weapons, however, is through the production and use of fissile (weapons-usable) materials, especially plutonium and highly enriched uranium (HEU).

In summary, two critical diversion and proliferation risks confront civilian nuclear power:

That countries or terrorist groups could divert fissile or radioactive materials directly from the civilian nuclear fuel cycle into nuclear explosives or (for terrorists) so-called "dirty bombs," which are conventional explosives surrounded by radioactive materials.

That countries aspiring to obtain nuclear weapons could use civilian nuclear facilities (power reactors, research reactors, reprocessing plants, uranium-enrichment plants, etc.) and trained cadres of nuclear scientists, engineers, and technicians as a cover and/or training ground for the dedicated acquisition of fissile material for nuclear weapons.

A third category of security risk posed by nuclear power is that terrorists could release substantial amounts of radioactivity through attacks on reactors, spent fuel pools, or certain other nuclear facilities. The consequences of such attacks could be devastating. However, the general issue of security of reactors and other nuclear sites raises questions outside the scope of this article and is not considered further here.

The proliferation-resistant barriers to confront these risks are both technical– (the quality of the plutonium and HEU used and produced in the fuel cycle, the degree to which the fissile material is or is not protected by a radiation shield, etc.) and institutional (e.g., safeguards, treaties, and agreements among suppliers not to export sensitive items). This article discusses both.

2. NUCLEAR WEAPONS

The critical fissile materials that can sustain an explosive chain reaction are plutonium (of almost any isotopic composition) and HEU (i.e., uranium enriched in the isotope U-235). Natural uranium contains approximately 0.7% U-235 and 99.3% U-238. Practically, uranium enriched to less than 20% U-235 cannot be used for weapons, and so-called weapon-grade uranium is generally taken to be more than 90% U-235. Although uranium with concentrations between 20 and 90% U-235 could sustain an explosive chain reaction, the critical masses needed, which go inversely as the square of the U-235 density, would be correspondingly very high. Although HEU includes all uranium above 20% U-235, here it generally refers to uranium at or above 90% U-235. Other isotopes could in principle be used in a weapon—notably U-233, americium-241, and neptunium-239—but so far these are not of practical significance. However, U-233, which is produced in a reactor containing thorium, could play a role in the future of civilian nuclear energy and is mentioned later.

Once it gathered the fissile material, a country or group could then assemble a fission weapon employing one of two fundamental designs: gun type and implosion. Fission weapons, such as were used in Hiroshima and Nagasaki, could also be used to ignite thermonuclear reactions in so-called two-stage devices, in which the fission component is termed a fission primary or "pit." Thus, a potential proliferant, even if ultimately seeking a thermonuclear weapon, will first have to develop a fission device.

In a gun device, two subcritical masses of a fissile material are explosively brought together, as along the barrel of a gun. In an implosion weapon, a subcritical sphere or shell of fissile material is imploded, increasing the density of the material to a supercritical state. For reasons discussed later, a gun-type device cannot use plutonium. The rudiments of both designs are widely known, and it must be assumed that many countries would be able to put together a team of scientists and engineers to manufacture a weapon once the fissile material was obtained. The gun-type design is particularly simple (it was used in Hiroshima and was never tested prior to that), and a subnational group may be able to use this design to produce a device, once it obtains the necessary fissile material.

Thus, the critical barrier to proliferation is acquisition of fissile material, and a central question

for civilian nuclear power is whether it provides a ready route to such acquisition.

Generally speaking, the critical masses required are 10 kg for plutonium metal and 50 kg for uranium metal (90% U-235). These are the so-called bare crits. The use of a reflector—several inches of a material that will reflect neutrons back into the weapon assembly—could cut the critical mass by as much as one-half, for example, to 4–8 kg for plutonium and 25 kg for uranium. The significant quantities specified by the International Atomic Energy Agency (IAEA) are 8 kg for plutonium and 25 kg for 90% U-235. The actual critical masses depend on the precise designs employed and composition of the plutonium and uranium, but these approximations provide a good benchmark to compare with the flows of materials in the civilian nuclear fuel cycle.

3. FISSILE MATERIAL PRODUCTION

3.1 Plutonium

Plutonium is not naturally occurring and must be produced in a reactor. Reactors dedicated to the production of plutonium for weapons have used natural uranium as fuel with low burn-up, which is the measure of how long an element of fuel is left in the reactor and therefore how much energy the element produces before it is discharged. In a reactor dedicated to plutonium production, the burn-up is typically less that 1000 thermal MW-days per ton of fuel. In such reactors, approximately 0.9 g of plutonium is produced per thermal megawatt-day. For example, the Indian CIRUS research reactor, which has a thermal power of 40 MW, at a 50% capacity factor could produce approximately 6.6 kg of plutonium per year.

Plutonium is also produced in civilian power reactors. In light water reactors (LWRs), the net plutonium production is approximately 0.3 g of plutonium per thermal megawatt-day, and in natural uranium, heavy-water reactors of the CANDU type, the production is approximately 0.5 g of plutonium per thermal megawatt-day. Thus, a 1000-MW electric (3000 thermal MW) LWR at 70% capacity factor produces approximately 230 kg of plutonium per year. In the United States, which has the equivalent of 100 1000-MW electric commercial reactors, approximately 23,000 kg of plutonium is generated in the reactors each year.

The spent fuel of all reactors contains substantial amounts of plutonium. However, as long as the plutonium remains locked in the spent fuel along with the highly radioactive fission products, it cannot be used for weapons. Furthermore, as long as the plutonium remains unseparated, it is protected from diversion by a very intense radiation field. For example, consider the dose rate from a pressurized water reactor (PWR) fuel assembly for a person at varying distances from the assembly. A typical PWR assembly weighs approximately 0.5 metric tons, has a burn-up of 35,000 MW-days per ton, and contains approximately 5 kg of plutonium. For such an assembly even for fuel out of the reactor for 15 years, a person 1 m from the assembly would receive a lethal dose in a few minutes. Moving 5 m away from such fuel would increase the time for a lethal dose to a couple of hours. After the first year of discharge, the exposure decays approximately by a factor of two every 30 years.

Separation of the plutonium has to be done in a reprocessing operation. With current technologies, the spent fuel needs to be decladded, chopped into small pieces, and dissolved in hot nitric acid to prepare for the plutonium extraction. The chemical extraction of the plutonium can then proceed by the use of blenders, pulse columns, centrifuge extractors, and the reduction and oxidation of the plutonium. All this has to be done behind heavy shielding and with remote handling. Therefore, reprocessing is not a trivial enterprise and requires substantial technical know-how.

3.2 Question of Reactor-Grade Plutonium

For a time, many in the nuclear industry believed that plutonium generated in commercial reactors could not be used for weapons. The reason for this belief (or hope) was that the so-called reactor-grade plutonium has a relatively high fraction of the isotope Pu-240, approximately 25% for plutonium from LWR spent fuel, compared to less than 6% for weapon-grade plutonium. Pu-240 fissions spontaneously, emitting large amounts of neutrons, leading to the possibility that one of the neutrons could initiate a chain reaction before the bomb assembly reaches its maximum supercritical state and thus creating a fizzle yield. Indeed, the prospect of such predetonation rules out the use of gun-type designs employing even weapon-grade plutonium, as noted earlier. This is because the assembly speed is low enough to virtually ensure a fizzle yield. Unfortunately, however, reactor-grade plutonium can be used for implosion weapons, for which the assembly speed is much greater. The issue was addressed in a 1994

National Academy of Sciences study and later described in a January 1997 U.S. Department of Energy release:

> *Virtually any combination of plutonium isotopes … can be used to make a nuclear weapon. In short, reactor-grade plutonium is weapons-usable, whether by unsophisticated proliferators or by advanced nuclear weapon states. Theft of separated plutonium, whether weapons-grade or reactor-grade, would pose a grave security risk.*

3.3 Highly Enriched Uranium

As noted previously, natural uranium contains approximately 0.7% U-235 and cannot sustain an explosive chain reaction. To obtain weapon-grade uranium, the uranium has to be enriched in isotope-separation facilities to approximately 90% U-235. This was, or is being, done in the nuclear weapon states mostly through the use of gaseous diffusion and centrifuge plants. The United States, in the Manhattan Project, and Iraq, in its nuclear program before the Gulf War, also used electromagnetic separation. All these methods require highly specialized equipment, considerable amounts of electricity (although the electricity intensity differs substantially among the different isotope-separation technologies), and a substantial effort to acquire usable quantities of weapon-usable material.

The isotope-separation work required to produce an enriched product depends on the tails assay of the depleted uranium streams but is independent of the method of separation used. It is usually measured in kilogram separative work units (SWUs). Thus, for example, to produce 1 kg of 90% uranium starting from natural uranium and with a tails assay of 0.2% requires 225 SWU and 180 kg of natural uranium (Table I). To put this in perspective, the kind of centrifuges that the Iraqis were assembling in 1990 each had a capacity of approximately 1 SWU per year. Thus, to produce one critical mass per year, for example, 25 kg of 90% uranium, would require more than 4000 centrifuges.

LWRs, the predominant commercial reactor type in the world today, and the high-temperature gas reactors currently under intensive study use low-enriched uranium (LEU) fuel, typically between 3 and 5% U-235 for LWRs and perhaps 8% for gas reactors. The production of such fuel requires isotope separation in which the uranium is enriched in the isotope U-235, with some typical data noted Table I. The isotope separation is accomplished in facilities in which in each "stage" the uranium is enriched slightly, with multiple stages connected in a "cascade" to produce the desired product. As already noted, the LEU cannot be used for weapons but the use of even LEU raises two proliferation concerns: The enrichment facilities used to produce the low-enriched fuel can be used to produce HEU, and if the low-enriched fuel is used as input to an enrichment cascade, the quantity of work required to produce HEU would be much reduced.

TABLE I

Uranium and Enrichment Requirements[a]

Feed enrichment	Product (1 kg)	Uranium feed (kg)	SWUs
Natural U	Weapon-grade U (90%)	180	225
Natural U	4% U	7.6	6.6
Natural U	8% U	15.6	17
4% U	Weapon-grade U	23.6	72
8% U	Weapon-grade U	11.3	38

[a] This is for 0.2% tails assay.

4. ROUTES TO FISSILE MATERIALS FOR WEAPONS

Although there are hybrid variants, the main routes to fissile material for weapons for an aspiring nuclear state are the following:

- Construct a dedicated uranium enrichment facility and produce HEU. All the declared nuclear weapons states—the United States, Russia, Great Britain, France, and China—did this. It is a route that was attempted by Iraq in the years leading up to the Gulf War. It was the route also followed by South Africa (before it gave up its nuclear weapon program) and, in a more ambiguous way, by Pakistan. North Korea also revealed in late 2002 that it had a program under way to produce HEU for weapons.
- Construct a dedicated production reactor and reprocessing plant, and produce and then separate plutonium. Again, all the declared nuclear weapon states did this. This route has also been tried by North Korea.
- Use a research reactor ostensibly built for peaceful purposes to produce the plutonium and then separate the plutonium in a reprocessing plant that may or may not be declared. This is the route taken by India with its CIRUS reactor and Israel at the Dimona site.

- Produce HEU at an enrichment facility nominally constructed for civilian uses. This is arguably what Pakistan did, although in the initial stages of the Pakistan program its facility was probably more accurately characterized as one dedicated to the production of HEU.
- Divert plutonium or HEU from the commercial civilian fuel cycle. As far as is nown, this has not yet happened.

One critical implication of the previous discussion is that countries could, if they wished, choose a dedicated route to nuclear weapons material. They do not have to depend on the civilian fuel cycle (although such a cycle could provide a convenient cover). As a consequence, it is not essential that safeguards provide a foolproof barrier to diversion from the civilian fuel cycle. Rather, the principal task of safeguards must be to make diversion from civilian nuclear power substantially more difficult, time-consuming, and transparent than a dedicated route.

On the other hand, it should also be understood that civilian nuclear power could provide a country with the basis for eventually a dedicated weapons program by allowing the country to train scientists and engineers, to build research facilities, and to learn techniques of reprocessing and enrichment that could later be turned to weapons uses. In this manner, a civilian program could impel a country along a path of latent proliferation, in which the country moves closer to nuclear weapons without having to make an explicit decision to actually take the final step to weapons, or at least to make transparent its intention to take such a step. Latent proliferation is particularly germane to the consideration of the spread of civilian nuclear power to countries that do not currently have any and that therefore would not have a ready infrastructure to support a dedicated route to nuclear weapons.

The routes to weapons summarized previously are for states. At the present and foreseeable state of nuclear technology, it does not appear feasible for a subnational group to produce its own fissile material. Even if it started with LEU or spent fuel, a subnational group could not realistically produce HEU or separate plutonium. The principal prospect for diversion of weapons materials from civilian nuclear programs for subnational groups is if they could steal or buy on a black market already separated fissile material—HEU or plutonium. These do not exist in the once-through fuel cycle, predominant in the world today; but HEU is used in some civilian research reactors and plutonium is separated in programs that reprocess and recycle plutonium. These potential sources of fissile material are discussed later.

Although a subnational group cannot realistically obtain fissile materials, it could conceivably acquire civilian spent fuel or high-level wastes for a dirty bomb or radiological weapon. However, the prospect that a terrorist group could use spent fuel or high-level wastes in a dirty bomb, although troubling, does not compare in severity to actual diversion of fissile material. A dirty bomb would not kill many people, and although if detonated in an urban concentration it would cause much economic dislocation, so could many other potential terrorist actions. With care, overseers of the nuclear fuel cycle could make it very difficult for a terrorist group to divert radioactive materials from nuclear facilities, especially compared to other routes terrorists could potentially take to acquire radioactive materials.

Naturally, countries or subnational groups could seek to obtain fissile materials not from peaceful nuclear activities but rather from nuclear weapons programs directly, with the security of materials in the sprawling nuclear weapons complex in Russia of particular concern. However, this possibility lies outside the scope of this article. It is critical that diversion be prevented in both the civilian and the weapons realms.

5. NUCLEAR REACTORS AND FUEL CYCLES

To understand the safeguards challenge, it is useful to understand the scope of the nuclear enterprise worldwide.

5.1 Nuclear Power Reactors

At the end of 2002, there were 442 reactors operating worldwide, with a total installed capacity of approximately 350 gigawatts electric (GW). The predominant reactor type operating today is the LWR, which is both cooled and moderated by light (ordinary) water. The LWR comprises two types of reactor: the PWR and the boiling water reactor (BWR). All LWRs use uranium fuel enriched to approximately 3–5% U-235 or mixed-oxide fuel (MOX) discussed later. Most LWRs have a capacity of approximately 1000 MW electric.

Canada and a few other countries operate heavy water reactors (HWRs), cooled and moderated by heavy water. HWRs typically use natural uranium fuel and thus do not require enrichment. This is a

proliferation-resistance plus in that a country would not have the excuse of producing enriched uranium for civilian uses. At the same time, the use of natural uranium has a drawback in that it allows a country to have independent control over its fuel cycle, not having to import enriched uranium. It would have to import or produce indigenously heavy water, however. Also, as already noted, a HWR produces more plutonium per year than a comparably sized LWR. The HWRs have capacities between 100 to several hundred megawatts.

In addition, a few other reactor types are currently deployed. Most important, the United Kingdom operates several reactors that are cooled by carbon dioxide and moderated by graphite. Some of these gas-cooled reactors use slightly enriched uranium and some use natural uranium metal. Russia operates several RBMK reactors (the Chernobyl type), which are water cooled and graphite moderated. RBMK reactors use enriched uranium.

Other reactor types are being investigated, with the belief that they might be more economic than current types and possibly also more proliferation resistant. Of special interest is the high-temperature gas-cooled reactor, including the pebble-bed modular reactor (discussed later), which is under intensive study. Also of interest is the fast breeder reactor, which is under development in Russia, Japan, and France. The type of fast reactor most fully developed is cooled by liquid sodium and uses plutonium or HEU as fuel. In such reactors, the chain reaction is carried by "fast" neutrons, which are not slowed to thermal speeds by a moderator as in the reactors described previously.

5.2 Research Reactors

In addition to the power reactors, most countries with nuclear power also operate one or more research reactors of several different types. These reactors are of concern for two reasons. First, they can be (and indeed have been) used to produce plutonium for weapons. The Indian CIRUS reactor produced the plutonium for India's first bomb, and the Dimona reactor in Israel has by all accounts been used to produce plutonium for Israeli nuclear weapons. Second, some research reactors use HEU as fuel, which could be used for nuclear weapons. This proliferation risk is discussed further later.

5.3 Uranium Enrichment Facilities

Uranium enrichment plants producing LEU are limited to only a few plants, most located in the nuclear weapons states. Minatom in Russia operates four plants. The U.S. Energy Corporation operates the Paducah plant in Kentucky. URENCO, a joint venture of three countries, operates three plants in the United Kingdom, Germany, and The Netherlands. Eurodif operates one plant in France. The total installed production capacity for enrichment is approximately 40 million SWU per year, with a current world demand of 37 million SWU.

5.4 Once-Through Fuel Cycles

Most nuclear power operates on a once-through fuel cycle in which the spent fuel is not reprocessed, and the fresh fuel for the reactors is either LEU or natural uranium, not usable for nuclear weapons. The spent fuel from the reactors for the most part remains at the reactor sites, either in spent fuel pools of water or in dry-cask storage, with the expectation that the fuel will eventually be put into deep geologic repositories (such as Yucca Mountain in the United States) for permanent disposal. In such a fuel cycle, weapon-usable material (plutonium or HEU) is never isolated.

5.5 Reprocessing and Recycling

Not all nuclear power is operated on once-through fuel cycles. The United Kingdom, France, Russia, and, to a lesser degree, Japan and India are reprocessing spent fuel, with commercial-size reprocessing plants located in Sellafield in the United Kingdom, La Hague in France, and Mayak (Chelyabinsk) in Russia. The U.K. and French reprocessing plants are reprocessing both their own spent fuel and spent fuel from other countries, notably Japan and Germany. In 1998, the total amount of plutonium separated from civilian spent fuel was approximately 195,000 kg, with most of this plutonium stored at the reprocessing plants: 70,000 kg in the United Kingdom, 76,000 kg in France, and 30,000 kg in Russia.

By 2002, approximately one-third of the spent fuel discharged from reactors each year worldwide was being reprocessed. Currently, the rate of plutonium separation is approximately 20,000–24,000 kg per year, although this may decrease during the next few years. By 2002, the total civilian inventory of separated plutonium exceeded the inventory of military plutonium, which has been estimated to be approximately 250,000 kg.

However, some of the separated plutonium is being fabricated into MOX at four plants in Europe. In 2000, these plants produced slightly less than 200,000 kg of MOX, incorporating

10,000–12,000 kg of plutonium, with MOX production capacity expected to double in the next few years. MOX is being burned in approximately 32 LWRs in France, Germany, Belgium, and Switzerland. Another 18 have been licensed to use MOX. Japan is planning to use MOX in one-third of its reactors by 2010. In almost all these cases, MOX is being used or is planned on being used in one-third of cores.

6. TECHNICAL BARRIERS TO DIVERSION

The principal technical barriers to diversion are evident from the previous discussion. They include materials barriers, such as the isotopic composition of the uranium fuel, the radiological protection provided by spent fuel, and the mass and bulk of materials that would have to be diverted. As an illustration of the last point, consider the effort that would be required to divert PWR spent fuel. Not only does the spent fuel present a strong radiation barrier to diversion but also each PWR assembly (containing approximately one critical mass of plutonium) weighs half a ton. It could not easily be handled or smuggled out of a facility. In transport, several assemblies would be put into a transport cask, perhaps weighing on the order of 100 tons—again representing a formidable barrier to diversion, certainly by a subnational group.

The technical barriers also include what might be called technology barriers, such as the complexity of facility modifications that would be required to acquire fissile material. One important illustration is to compare the relative proliferation resistance of two technologies for uranium enrichment—gaseous diffusion and centrifuges. A gaseous diffusion plant set up to produce LEU would be very difficult to convert so that the plant could instead produce HEU for weapons. In contrast, a centrifuge plant ostensibly designed to produce LEU could readily be converted to produce HEU. The reason for this is that for gaseous diffusion, it is very difficult and highly inefficient to reconfigure the cascade from one optimized to produce LEU to one designed to produce HEU, whereas it is relatively straightforward to do so for a centrifuge cascade.

7. INSTITUTIONAL BARRIERS TO PROLIFERATION: SAFEGUARDS

The Non-Proliferation Treaty (NPT), which entered into force in 1970, provides the principal institutional barrier to proliferation. The treaty defines two classes of states: the declared nuclear weapon states—the United States, Russia, the United Kingdom, France, and China—which had tested nuclear weapons before the treaty came into force, and the nonnuclear weapon states. It imposes obligations on both.

With respect to the nonnuclear weapon states, the treaty requires that each party pledge not to acquire nuclear weapons and to subject all its nuclear activities to safeguards applied by the IAEA, an autonomous organization established in 1957 under the United Nations (UN). Under the IAEA safeguards agreements signed by the parties, the parties are obliged to report all nuclear activities and to allow IAEA inspections at all declared facilities. The requirement that all of a country's peaceful activities be put under safeguards is referred to as full-scope safeguards. The purpose of the inspections is to enable the IAEA to detect any significant diversion of materials from the civilian nuclear fuel cycle, with any such detection then reported to the UN Security Council. With the exception of India, Pakistan, and Israel, all countries with any kind of nuclear program have joined the treaty, including the five declared nuclear weapon states. South Africa initially did not join the NPT and then actually manufactured a small number of nuclear weapons. However, in 1990 and 1991, it dismantled the weapons and its secret nuclear program, and it joined the NPT as a nonnuclear state in 1991. North Korea withdrew from the NPT in early 2003.

Iraq and North Korea, despite their past formal adherence to the NPT, embarked on nuclear weapon programs while still parties to the Treaty. The United States has claimed that Iran also has a clandestine weapons program, but this has been denied by Iran, and all of Iran's declared nuclear facilities remain under IAEA safeguards.

During the years leading to the First Gulf War, Iraq maintained all its declared nuclear facilities under IAEA safeguards while it was undertaking an extensive clandestine nuclear weapons program hidden from IAEA inspectors. Since 1991, of course, Iraq has been subject (with a 4-year gap between 1998 and 2002) to extraordinary inspections by the IAEA for nuclear weapons programs and a special UN inspectorate for chemical and biological weapons, with the war launched against Iraq in March 2003 in part justified by the United States, the United Kingdom, and others on grounds that Iraq was still harboring weapons of mass destruction and not cooperating with the inspection agencies.

Aside from Iraq, North Korea is the other nonnuclear party to the NPT that has most clearly sought to evade safeguards. It delayed for several years signing a safeguards agreement with the IAEA, evidently separated some plutonium in the meantime, and in 1994 threatened to withdraw from the NPT when the plutonium production was discovered. The looming crisis in 1994 was averted when, under an agreement negotiated bilaterally with the United States (the so-called Agreed Framework), North Korea agreed to place all its declared facilities under safeguards by the IAEA and to freeze (although not dismantle) its nuclear program. However, the IAEA noted in its 2001 annual report that "the Agency is still unable to verify the correctness and completeness of the initial report of nuclear material made by the DPRK and is, therefore, unable to conclude that all nuclear material subject to safeguards has been declared." In October 2002, North Korea suggested that it had a hitherto undeclared program to enrich uranium; and in December 2002 it withdrew its declared stockpile of spent fuel and its reprocessing plant from IAEA safeguards. At the time of this writing, the situation in North Korea remains unsettled.

In addition to the NPT, many nonnuclear weapon countries have joined into nuclear-free zones, in which the countries of a specific region reaffirm their decision not to acquire nuclear weapons. These zones include Latin America (Treaty of Tlatelolco, 1967), the South Pacific (Treaty of Rarotonga, 1985), Southeast Asia (Treaty of Bangkok, 1995), and Africa (Treaty of Pelindaba, 1996).

Also supplementing the NPT, countries exporting nuclear technology have entered so-called supplier agreements not to export certain specified materials or technologies to nonnuclear countries and to ensure that whatever nuclear material or equipment that is exported is under safeguards. The critical supplier agreement in place is the Nuclear Supplier Group (NSG) instituted in 1977 (when it was initially called the London Club), including all the major nuclear suppliers. The NSG has formulated a set of guidelines for nuclear transfers. The guidelines define a "trigger list" of items that suppliers agree to export to nonnuclear weapon states only when the receiving state has brought into force an agreement with the IAEA requiring the application of full-scope safeguards on all its current and future peaceful activities. Suppliers also agree to exercise restraint on the transfer of sensitive technologies, such as reprocessing and enrichment facilities. These technologies could be exported only with explicit further commitments by the recipient states, for example, not to produce HEU. Paralleling the NSG is another intergovernmental group, the Nuclear Exporters Committee, known as the Zangger Committee, set up in 1971, which also has established conditions and procedures to govern nuclear exports.

These multifaceted international arrangements have been effective. Today, essentially all nuclear facilities in nonnuclear countries (a category that does not include Israel, India, Pakistan, and North Korea) are under IAEA safeguards, which is a remarkable achievement.

Furthermore, if a state party to the NPT withdrew from the treaty, in many cases backup safeguards applied by the United States, Canada, France, or other nuclear supplier would come into force. That is, a country such as Japan could not simply shed all its safeguards obligations by withdrawing from the NPT. These residual safeguards obligations generally derive by virtue of the country importing reactor technology or nuclear fuel from supplier countries. It may be noted in passing that such backup to NPT safeguards may not persist in the future if nuclear power grows substantially and an increasing number of countries produce nuclear technologies and materials indigenously.

After the Gulf War of 1991 revealed that Iraq, although a party to the NPT and with all its declared nuclear facilities under safeguards, was pursuing a very extensive nuclear weapons program, the international community realized that safeguards would have to be extended to allow the IAEA to search for undeclared activities as well as monitoring declared sites. In response, the IAEA developed a strengthened safeguards system called INFCIRC 540, which authorizes the agency to employ environmental monitoring and other means to search for undeclared activities. This strengthened safeguard system has entered into force in several countries, including Japan and Canada. European countries have generally indicated their intention to accede to the new safeguards agreement. By the end of 2002, the United States had signed but not yet ratified. As of the summer of 2003, several countries with nuclear power ambitions had not signed, including Iran, (which has now indicated that it will sign and ratify), Egypt, Turkey, South Korea, Argentina, and Brazil.

8. PROLIFERATION RESISTANCE: AN INITIAL ASSESSMENT

How should one gauge the proliferation resistance of the nuclear fuel cycle? One starting point is an

October 2000 report by a task force under the auspices of the Nuclear Energy Research Advisory Committee (NERAC), set up by the U.S. Department of Energy, that was tasked with addressing the question of proliferation resistance. The task force considered various attributes of proliferation resistance, including the materials, technology, and institutional barriers described previously. The task force cautioned that the proliferation resistance provided by any particular barrier depends on the threat or scenario one has in mind, and so its report should be considered preliminary. Nevertheless, the summary table that the task force put forward captures the effectiveness of most of the proliferation-resistant ideas that have been put forward. Table II is a simplified form of the summary table. The table suggests two compelling conclusions: (i) that it will be extraordinarily difficult to contrive an effective proliferation-resistant nuclear fuel cycle for sophisticated states, and even difficult to do so for unsophisticated states, and (ii) that effective barriers could be erected against subnational diversions of nuclear material.

The first conclusion is not surprising. It is evident that any country with a serious nuclear energy program could convert it to weapons uses if it wished, although it might not be able to do this clandestinely. For example, even if a country operated only on a once-through fuel cycle, it could build a quick and dirty reprocessing plant able to extract plutonium reasonably rapidly within a year or less. Furthermore, it is unclear what sanctions could be brought against a country breaking out of its safeguards obligations. The IAEA is charged with bringing any violation of safeguards to the UN Security Council and conceivably the council or individual countries could take action against the violating country, but whether and how this might be done are highly uncertain. Nevertheless, a country will be loathe to be caught cheating on safeguards or to directly break a safeguards agreement, and this provides some deterrent to such actions. However, it certainly does not present an ironclad one.

In the current world environment, the potential of a state to divert weapons materials from its civilian programs is not necessarily dismaying. First, nuclear power is overwhelmingly located in a relatively few industrialized democracies, a few countries in Eastern Europe, Ukraine, Lithuania, and Russia. Of the 350-GW installed capacity worldwide, less than 10 GW is in developing countries, generally thought to be the most likely arena for further proliferation. This includes 2.3 GW in India, 2.1 GW in China, 0.9 GW in North Korea, 0.4 GW in Pakistan, and 2.7 GW in South America. For the most part, the countries with nuclear power programs are either already nuclear weapon states or countries that for whatever reason do not aspire to become nuclear weapon states. Japan, South Korea, and Taiwan have at one time or another considered nuclear weapons, but they are deeply constrained from seeking weapons. Second, as already mentioned, the civilian program, even if it could be converted to weapons purposes, might not provide a quicker or easier path to nuclear material than a dedicated route.

This much is not is great dispute. More controversial is whether and to what degree reprocessing and recycling significantly weaken the proliferation resistance of the nuclear fuel cycle.

8.1 Reprocessing and Recycling

Opponents of recycling (this author included) emphasize the proliferation risks. First, the reprocessing

TABLE II

Relative Importance of Various Barriers to a Selected Threat

	Sophisticated state, overt	Sophisticated state, covert	Unsophisticated state, covert	Subnational group
Materials/barriers				
Isotopic	Low	Low	High	High
Chemical	Very low	Very low	High	High
Radiological	Very low	Low	Moderate	High
Mass and bulk	Very low	Low	Moderate	High
Detectability	Very low	Low	Moderate	High
Technical barriers	Very low–moderate	Very low–high	Low–moderate	Moderate–high
Institutional barriers	Very low–moderate	Very low–high	Low–high	Low–moderate

Source: NERAC Task Force on Technology Oppurtunities for Increasing the Profileration Resistance of Global Civilian Nuclear Power, October 2000.

generates a tremendous quantity of separated plutonium that has to be very carefully accounted for and guarded. As already noted, at the end of 2002, well over 200 metric tons of separated plutonium was being stored in the form of plutonium oxide at several sites throughout the world, with most in France, the United Kingdom, and Russia. The material in France and the United Kingdom is reasonably secure. However, a large quantity is in Russia under less certain security, and there are appreciable quantities in Japan.

Second, the use of MOX in reactors means that there will be supplies of fresh plutonium fuel at MOX fabrication plants, reactor sites, and in transport from reprocessing plants to fabrication plants to reactors—all potential sources of diversion for subnational groups. With the recycling activities restricted to Europe, the standards of security and safeguards applied to the MOX are probably high but may not always remain so. Third, whatever the risks today, they will be multiplied if a real market develops for MOX, with middlemen and agents arranging for the purchase and sale of MOX.

Finally, safeguards at bulk-handling facilities, such as reprocessing plants and MOX fabrication plants, will be much more difficult to apply than at reactors, with far more diversion paths potentially available to a country.

These risks seem all the more exacerbating to the critics of recycling because the enterprise appears so dubious on economic grounds alone. Even if the plutonium were free (i.e., even if the MOX fabrication plants received the plutonium free, with all the costs of separating the plutonium in reprocessing plants written off), the cost of a ton of MOX fuel exceeds that of a ton of ordinary LEU at current uranium prices. Unless the cost of uranium increased very sharply, the small savings in uranium and in enrichment due to plutonium recycling cannot make up for the far greater cost of fabrication of MOX, and the economic penalty of recycling is all the more apparent if the cost of reprocessing is taken into account.

Supporters of reprocessing and recycling downplay the proliferation risks, especially since these activities are limited to Europe and Japan. They also point out that a country currently relying only on a once-through fuel cycle could build a quick and dirty reprocessing plant to separate plutonium.

Furthermore, recycling plutonium and burning it in reactors sharply reduces the buildup of plutonium in spent fuel compared to the once-through fuel cycle. This could be important because one potential risk of the once-through fuel cycle is that the plutonium contained in the spent fuel could eventually become accessible to potential proliferators as the radioactive barrier provided by the fission products decays, by a factor of approximately 10 every 100 years. Thus, potentially the geologic repositories to which the spent fuel is destined could become plutonium mines. How seriously one should take this consideration is unclear. As long as a country has an operating nuclear power program, diversion of plutonium appears far easier than it would be in a mining operation.

8.2 Research Reactors Using HEU

By far the greatest inventories of HEU in the world are in nuclear weapons and in materials recovered from dismantled nuclear weapons—on the order of 1000–2000 tons. The greatest security challenge is to secure these materials until they can be denatured (i.e., rendered not usable for weapons) by mixing the HEU with natural or depleted uranium. This denaturing of weapon-grade uranium in fact is being done in Russia, with the United States purchasing the denatured product for LEU for reactors.

The quantity of HEU at research reactors is far smaller than the weapons-derived HEU—approximately 20 tons. However, this 20 tons is widespread and also must be carefully secured. In 2001, there were approximately 40 operating HEU reactors with thermal power greater than 1 MW. Many of these are in the United States and Russia. However, several are in nonnuclear weapon states. Reactors supplied by the United States are in Australia, Belgium, Germany, Israel, The Netherlands, Portugal, and South Africa. Reactors supplied by Russia are in the Czech Republic, Hungary, Kazakhstan, North Korea, Libya, Poland, Ukraine, and Uzbekistan. There are also a small number of shutdown HEU reactors supplied by Russia that still contain some HEU.

Despite the number of HEU reactors still in operation, the situation has greatly improved during the past two decades. In 1978, the United States launched the Program for Reduced Enrichment for Research and Test Reactors, which has successfully converted many HEU reactors with thermal power greater than 1 MW in the United States and abroad to low-enriched fuel. The Soviet Union and now Russia have made similar efforts since 1978. The United States and Russia had been the principal suppliers of HEU for research reactors. Indeed, most of the reactors still using HEU were constructed before 1980. Since then, only four HEU reactors with thermal power more than 1 MW have begun

operation or are planned: a 5-MW reactor in China, a 10-MW reactor in Libya, a 10-MW reactor in Russia, and a 20-MW reactor in Germany.

Still, one of the main unfinished nonproliferation undertakings is shutting down or converting the remaining HEU reactors and removing any fresh and spent fuel inventories at the reactor sites to their countries of origin.

9. THE LONG-TERM CHALLENGE FOR NUCLEAR POWER: ENHANCED PROLIFERATION RESISTANCE

9.1 The Challenge

In the long term, if nuclear power worldwide becomes a robust contributor to a strategy to address global warming, the challenge to ensure that nuclear power is adequately proliferation resistant will be truly formidable. It is useful for illustrative purposes to take as a benchmark a global nuclear capacity of 3000 GW—an eightfold increase from today's worldwide capacity of 350 GW.

An increase of at least this magnitude will be necessary for nuclear power to make a dent in global warming. For example, under the central business-as-usual projection of the Intergovernmental Program on Climate Change (IPCC), if nuclear power increased to 3000 GW by 2075 (50% of world electricity then projected) and 6500 GW by 2100 (75% of world electricity), the total carbon emissions avoided cumulatively would be approximately 290 billion tons through 2100—only approximately one-fourth the projected cumulative carbon emissions through 2100 projected by the IPCC.

The management of a nuclear system of 3000 GW would be deeply challenging. First, whereas today the countries with nuclear power programs are either already nuclear weapon states or industrialized democracies with no current intentions to acquire nuclear weapons, an exuberant nuclear future will present a different picture. It is widely recognized that significant nuclear growth during the next half century will have to be largely in the developing countries. This is where by far the greatest increase in electricity production is projected. If one considers the top 25 countries by population projected for 2050 by the UN, among them are several that today have essentially no or a negligible amount of nuclear power: Indonesia, Nigeria, Pakistan, Bangladesh, Ethiopia, Congo, the Philippines, Vietnam, Egypt, Iran, Saudi Arabia, Tanzania, Turkey, Sudan, Uganda, Yemen, and Thailand. No doubt, several of these countries, and others less populous, will not actually develop nuclear power on a large scale. However, if nuclear power truly comes to play a substantial role in the world energy economy, it will have to be located in many of these countries and on a substantial scale. Therefore, it will be a very substantial undertaking to extent safeguards to all these countries, some of which may in time aspire to nuclear weapons.

Second, the flows of material will be immense. If based on a once-through fuel cycle using LWRs, a 3000-GW system would generate annually approximately 600 tons of plutonium and also require on the order of one-half million tons of natural uranium. If based on liquid-metal plutonium breeder reactors, it would involve the fabrication into fresh fuel annually of more than 4000 tons of plutonium (although the cumulative inventory of plutonium would be much less than that for a system based on LWRs).

9.2 Proliferation Resistance Concepts

Confronted by this challenge, nuclear engineers and scientists in the United States and abroad have been exploring advanced nuclear technologies and novel concepts to improve the proliferation resistance of nuclear power (among other goals), including reactor types and/or new fuels that allow very high burn-up and produce less plutonium than do current reactors (such as the pebble-bed high-temperature gas-cooled reactor) and that may be operated on once-through fuel cycles; breeder or particle-accelerator-driven reactors that, to the extent possible, colocate sensitive processes (such as reprocessing) with the reactor and do not separate the plutonium from other actinides; and schemes that restrict nuclear power to large, international energy parks that would then export to individual countries electricity, hydrogen, or small, sealed reactors. The reactors envisioned in this last scheme would be 40 or 50 MW and would be fueled at some central nuclear park and then sealed and sent to client countries. The reactors would have lifetime cores, not requiring refueling, and at the end of the core life (e.g., 15–20 years) would be returned to the central facility unopened. This can be called a hub–spoke configuration.

9.3 Once-Through Fuel Cycles: Thorium Fuels and the Pebble-Bed Gas-Cooled Reactor

Although the once-through concepts under consideration differ sharply in terms of reactor type,

economics, and inherent safety, the proliferation resistance attractions of the different concepts are generally similar. Two technologies, the Radkowsky thorium fuel (RTF), currently under development by the Radkowsky Thorium Power Corporation, and the pebble-bed high-temperature gas-cooled reactor illustrate the enhanced proliferation resistance promised by the new concepts.

In the short term, the RTF fuel, employing a seed-blanket concept, allows a technical fix for current LWRs aimed at making the spent fuel difficult to use for weapons purposes. The fuel is designed to operate on a denatured uranium–thorium once-through fuel cycle in current LWRs and to achieve very high burn-up. In such a fuel cycle, the reactor would generate approximately one-fifth the plutonium generated in today's LWRs per kilowatt-hour of electricity produced. Moreover, the Pu-238 concentration in the discharged fuel would be 6.5% in the seed and 12% in the blanket compared to approximately 1% in today's PWR uranium fuel. This is significant because Pu-238 has a very high specific decay heat (560 compared to 1.9 W/kg for Pu-239 and 6.8 W/kg for Pu-240). The decay heat emission for the RTF seed is approximately three times that of normal PWR reactor-grade fuel and for the RTF blanket fuel approximately six times that of PWR fuel. The higher heat loads are likely to require special heat-removal measures in the design of a weapon. The spontaneous fission rate of the RTF plutonium isotopic mixture will also be far greater than that for PWR uranium fuel, making the design of a weapon more difficult and increasing the probability of a fizzle yield for a crude weapon. The U-233 (which, like U-235, is a fissile material) generated by thorium capture of a neutron would always be denatured with a sufficient quantity of U-238 so that the uranium cannot be used for weapons, and it is contaminated with the gamma-emitting U-232 decay chain, which would make any attempt to use the material for weapons more difficult and also any diverted material more detectable.

The pebble bed modular reactor derives its proliferation resistance from the fact that the spent fuel would be high burn-up material in thousands of tiny carbon-coated spheres, making it a comparatively unattractive source from which to recover weapons-usable materials. At any given time, the core of the reactor (nominally 100–110 MW) would consist of 360,000 pebbles (60 mm in diameter), each containing approximately 7 g of uranium (at approximately 8% U-235) in the fresh fuel in 11,000 microspheres (0.9 mm in diameter). In the spent fuel, there will be plutonium. At low burn-up, however, when the weapon-grade quality of the plutonium is relatively high the plutonium content will be very low. At full burn-up, there will be more plutonium but far from weapon-grade quality, with a high fraction of Pu-238 (approximately 6%). The total content of the plutonium will be approximately 5 kg per ton of uranium fuel so that perhaps 200,000 pebbles would have to be diverted to obtain a critical mass.

There are some disadvantages. First, over time the decay of the approximately 30-year half-life fission products will lower the radiation barrier of spent fuel, whereas the decay of the 88-year half-life Pu-238 will make it easier to use the extracted plutonium for weapons. Perhaps more important, both fuel cycles use uranium that is more highly enriched than is typical today: the RTF fuel, 20% U-235 for the seed, and the pebble bed reactor, 8% U-235. Uranium enriched to 8–20% cannot be used for weapons, but the routes to weapons-grade uranium from such feed materials are easier than if one started with natural or 4% low-enriched uranium. The fabrication of the pebble-bed fuel and the fuel-handling operations will require special attention when safeguards are developed. Despite these cautions, these concepts, if they can truly be developed, promise enhanced proliferation resistance, especially during the next few decades.

In the long term, however, any nuclear system relying primarily on once-through fuel cycles will face formidable difficulties, even if sufficient uranium can be obtained from extraction from seawater or terrestrial sources to sustain a once-through cycle. (The natural uranium required will be on the order of one-half million tons annually.) For specificity, assume a 3000-GW nuclear capacity comprising half pebble-bed high-temperature gas reactors, each 100 MW, and half LWRs, each 1 GW. In such a world, there would be 15,000 pebble-bed reactors and 1500 LWRs and an enrichment requirement worldwide of approximately 400 million SWU per year. If one takes 2 million SWU per year as a nominal capacity of one enrichment plant (approximately the size of a URENCO plant), 200 such plants would be required. A 2-million SWU plant could make approximately 600 bombs per year starting with natural uranium. It could make 3500 bombs per year starting with 8% uranium, the fuel enrichment of the gas reactor fuel. Overall, the system will involve massive flows of natural and LEU and powerful economic incentive for innovation to make isotope separation cheaper and quicker. This is especially of concern because terrorist groups could

far more readily make a nuclear weapon from HEU than from separated plutonium.

Plutonium will also be a matter of concern. The spent fuel would be on the order of 50,000–70,000 tons of heavy metal per year, approximately the capacity that has been planned for Yucca Mountain (70,000 tons). Thus, nominally we can imagine one Yucca Mountain being constructed every year worldwide. Each one will have to be secured and safeguarded indefinitely since after several decades, the radioactivity surrounding the plutonium will decay substantially, making the spent fuel repositories prospective plutonium mines, as discussed earlier. Each repository (using the Yucca Mountain scale) would contain on the order of 1000 tons of plutonium-239.

9.4 Closed Fuel Cycles

Because a once-through fuel cycle will involve such prodigious use of uranium and enrichment services, it is likely that a 3000-GW nuclear power system will drive the nuclear industry toward breeder reactors. Such a development would minimize flows of uranium; but it would increase material flows of plutonium. For this reason, breeder concepts in which the plutonium is never fully separated from the fission products and/or other actinides have appeared worthy of study on proliferation-resistance grounds. Among these are integral fast reactors (IFR), which employ pyroprocessing technologies that do not separate the plutonium from the minor actinides and that integrate the reprocessing and reactor operation at a single site, and particle-accelerator driven reactors, which also use pyroprocessing, one version of which has been developed by Carlo Rubbia.

In general, scenarios to divert plutonium from these fuel cycles will require off-normal operation and specialized equipment, which in principle could be detectable by an inspection agency. On the other hand, even where pyroprocessing is used, fuel recycling equipment could be operated in a manner that would allow the extraction of large amounts of plutonium fairly quickly by adding a cleanup stage. In addition, safeguards that relied on materials accountability would be difficult to apply and therefore more than customary reliance would consequently have to be placed on containment and surveillance. Partly for these reasons, a study by the Martin-Marietta Corporation commissioned by the Departments of State and Energy in 1992 concluded that "in the absence of compelling circumstances to the contrary, access to IFR technology should be preferentially limited at the outset to countries that have good nonproliferation credentials."

This last point is the most troubling. Which countries are considered "safe" or to have "good nonproliferation credentials" will depend on who is making the judgment. It is illusory to think that the world can build a secure nuclear future that depends on two classes of countries—one where certain activities and fuel cycles are barred and one where anything, or almost anything, goes.

9.5 Hub–Spoke Configurations: International Energy Parks

One manifestation of future nuclear power is not open to many of the objections noted previously. This is to cluster all sensitive nuclear facilities in centralized, heavily guarded nuclear parks, preferably under international control. Then, electricity directly or hydrogen produced through electrolysis could be sent out from the parks. Also, the possibility of exporting small reactors from the park has been examined.

In this last case, long-life reactor cores would be assembled at the central facility. The reactors would be sealed and then exported to users in other countries, where they could be "plugged in" to the electric generation system. After some years (15–20), the core/spent fuel would be returned to the central facility or to some international spent fuel repository. During the 15–20 years of operation, there would be no refueling. In such a system, a country would need relatively few research facilities, operators, and other trained nuclear technicians and engineers.

This reactor concept has impressive proliferation-resistance credentials. These may be summarized as follows, adopting the analysis presented by engineering teams at the Berkeley Department of Nuclear Engineering, the Lawrence Livermore National Laboratory, the Argonne National Laboratory, and Westinghouse Electric Company for one version of a small reactor, the encapsulated nuclear heat source (ENHS) reactor. The following description and analysis are adapted from a paper by Ehud Greenspan *et al.* presented at the Global 2001 Conference in Paris in September 2001.

First, appropriation by a subnational group of the reactor, although it is transportable, would be a daunting challenge. The ENHS reactor is approximately 20 m long, with a 3-m diameter, and weighs approximately 200 tons. The fuel, which could either be 13% enriched uranium or a uranium–plutonium

fuel having 11 or 12% plutonium, is embedded in a mass of lead-bismuth (solid during transport and liquid during operation) throughout the core life. It would further be possible to seed the reactor with gamma-emitting cesium-137 before shipment, thus surrounding even the fresh fuel in a radiation shield. Furthermore, the ENHS does not give a country a useful source of neutrons: It is not possible to insert fertile material for irradiation. As noted previously, the core life of the reactor would be 15–20 years and during this period there would be no refueling. If operated on the hub–spoke concept, the client country would need no fuel fabrication facility and no fuel management capability. Because the reactor operates "almost autonomously," the client country would need few operators of the nuclear system. Overall, the hub–spoke concept could diminish the rationale and opportunities for a country to develop research facilities and trained cadres of scientists and technicians that could later be diverted to weapons activities.

Presumably, the client country would be able to break into the sealed reactor, but it should be possible to ensure that such an attempt to obtain nuclear fuel could not be done undetected. Moreover, the acquisition of the fuel after a break-in would probably take days to weeks.

There are some matters of concern. First, the spent fuel of the reactor (using a nominal 40-MW capacity) will contain approximately 600 kg of plutonium if the uranium fuel is used and 1.4 tons of plutonium if the uranium–plutonium fuel is used. Also, if the fuel is removed from the reactor before its complete lifetime, the plutonium could be weapon grade or close to it. The uranium fuel, if obtained by a would-be proliferant, would not be weapons usable but would require less separative work for production of weapons-grade uranium than ordinary light water uranium fuel.

These problems notwithstanding, a nuclear system based on international energy parks, if it could be developed, does present an arguably proliferation-resistant strategy for nuclear power in the long term.

Are international energy parks realistic alternatives on political and economic grounds? International energy parks run against the strong wish of many countries to become energy independent. Also, they will require either that client countries accept discriminatory restrictions on their nuclear activities not accepted by the countries hosting the nuclear parks or that all countries, including the industrialized countries, accept a high degree of international control over their nuclear energy programs. Beyond these considerations, countries will also be wary of concentrating too much of their energy future in a few places, with their attendant risks of common-mode failures, disruption of transmission lines or shipping, etc.

10. CONCLUSION

With the growth of the international terrorist threat, the diversion and proliferation risks of nuclear power are sure to gain in importance compared to the many other problems confronting nuclear power, such as economics, waste disposal, and safety. It may be that with nuclear power remaining at current levels and with almost all reactors and enrichment and reprocessing plants located in a relatively small number of highly industrialized countries, the international community will view the proliferation resistance of the current system of technologies and safeguards as tolerable, especially if next-generation reactors can allow markedly increased burn-up of fuel and if the IAEA strengthened safeguards system is widely implemented.

In the long term, if nuclear power grows substantially, the development of new proliferation-resistant technologies and fuel cycle configurations will gain new urgency. The concept under study that holds the most promise is the development of a hub–spoke arrangement in which all sensitive activities are performed at a central, perhaps international, energy park. However, such a strategy faces enormous political and practical obstacles. All the more so does the extreme of this strategy—to place all nuclear power under international control. This indeed was the central recommendation of the first comprehensive study of how to control nuclear energy, the "Report on the International Control of Atomic Energy," prepared for the Secretary of State's Committee on Atomic Energy (the so-called Acheson–Lilienthal Report) that formed the basis for the Baruch Plan for international control of nuclear weapons (submitted to the United Nations by the United States in 1946). It is worth remembering the sober view of this report:

> *There is no prospect of security against atomic warfare in a system of international agreements to outlaw such weapons controlled only by a system which relies on inspection and similar police-like methods. The reasons supporting this conclusion are not merely technical but primarily the inseparable political, social, and organizational problems involved in enforcing agreements between nations, each free to develop atomic energy but only pledged not to use bombs. So long as intrinsically dangerous activities may be carried out by nations, rivalries are inevitable and fears are engendered that place so great a pressure on a system of enforcement by police methods that no degree of ingenuity or technical competence could possibly cope with them.*

SEE ALSO THE FOLLOWING ARTICLES

Nuclear Engineering • Nuclear Fuel: Design and Fabrication • Nuclear Fusion Reactors • Nuclear Power, History of • Nuclear Power Plants, Decommissioning of • Nuclear Power: Risk Analysis • Nuclear Waste • Occupational Health Risks in Nuclear Power • Public Reaction to Nuclear Power Siting and Disposal • Radiation, Risks and Health Impacts of • Uranium Mining: Environmental Impact

Further Reading

Albright, D., Berkhout, F., and Walker, W. (1997). "Plutonium and Highly Enriched Uranium 1996." Stockholm International Peace Research Institute/Oxford Univ. Press, Stockholm/Oxford.

Bennett, R., and Kotek, J. (2002). Next generation nuclear power. *Sci. Am.* **286**(1), 70–79.

Cirincione, J. (2002). "Deadly Arsenals: Tracking Weapons of Mass Destruction." Carnegie Endowment for International Peace, Washington, DC.

Feiveson, H. A. (2001.). The search for proliferation-resistant nuclear power. *Public Interest Rep.* **54**(*5*).

Garwin, R., and Charpak, G. (2001). "Megawatts and Megatons." Knopf, New York.

Goldblat, J. (2002). "Arms Control: The New Guide to Negotiations and Agreements." Stockholm International Peace Research Institute, Stockholm.

Krass, A., *et al.* (1983). "Uranium Enrichment and Nuclear Weapons Proliferation," pp. 1–40. Stockholm International Peace Research Institute/Taylor & Francis, Stockholm/London.

Lloyd, W. R., Sheaffer, M. K., and Sutcliffe, W. G. (1993, October). Dose rate estimates from an irradiated pressurized-water-reactor fuel assembly in air, Report No. UCRL-ID-115199. Lawrence Livermore National Laboratory, Livermore, CA.

Rhodes, R., and Beller, D. (2000). The need for nuclear power. *Foreign Affairs* **79,** 30–44.

Nuclear Waste

PAUL K. ANDERSEN and ABBAS GHASSEMI
New Mexico State University
Las Cruces, New Mexico, United States

MAJID GHASSEMI
New Mexico Institute of Mining and Technology
Socorro, New Mexico, United States

1. **Nuclear Fission**
2. **Nuclear Fuel Cycle**
3. **Radionuclides in Nuclear Waste**
4. **Classification of Radioactive Wastes**
5. **Disposal of Low-Level Wastes**
6. **Management of Spent Nuclear Fuel**
7. **Disposal of High-Level Wastes**
8. **Radioactive Waste Management in the United States**
9. **Radioactive Waste Management in France**
10. **Radioactive Waste Management in Japan**

Glossary

actinides The 14 elements having atomic numbers in the range from $Z = 90$ (thorium) to $Z = 103$ (lawrencium); the element actinium ($Z = 89$) is sometimes included as well.

activation Creation of a radioactive element from a stable one by the absorption of neutrons or protons.

alpha radiation (α) Nuclear radiation consisting of helium nuclei, each comprising two neutrons and two protons.

Becquerel (Bq) The SI unit of radioactivity, equal to one disintegration per second.

bentonite A soft plastic clay formed by the chemical alteration of volcanic ash, which expands on the absorption of water.

beta radiation Energetic electrons (denoted β^-) or positrons (denoted β^+) emitted from unstable nuclei.

calcining (or calcination) The process of heating a material to a high temperature below its melting point to drive off volatile species or promote chemical changes.

centrifugal separation An enrichment process that uses a series of high-speed centrifuges to separate $(^{238}U)F_6$ and $(^{235}U)F_6$ according to the difference in their molecular masses.

clay Unconsolidated sediments consisting of particles less than 1/256 mm in diameter.

daughter A nuclide that results from the radioactive decay of another nuclide (the "parent" nuclide).

depleted uranium Uranium containing less than the natural fraction of U-235 (0.71%).

electronvolt (eV) A unit of energy, defined as the kinetic energy acquired by an electron that is accelerated through a potential difference of 1 volt; an electronvolt is equal to 1.6022×10^{-19} J.

enrichment The process of increasing the proportion of a fissile isotope in a nuclear fuel above its natural fraction.

fertile element A nonfissile element that can be made fissile by the absorption of neutrons.

fissile element A heavy element that readily undergoes fission by the absorption of thermal (i.e., slow) neutrons.

fission The splitting of a heavy nucleus into two roughly equal fragments, with the accompanying release of energetic neutrons and (sometimes) a light nucleus such as tritium.

gamma radiation (γ) High-energy photons emitted from unstable nuclei.

gaseous diffusion A separation process that depends on differences in diffusion rates of gaseous compounds through a membrane; gaseous diffusion is used to separate $(^{238}U)F_6$ and $(^{235}U)F_6$.

half-life The time required for half of the nuclei in a radioactive sample to decay.

hex Abbreviation for uranium hexafluoride, UF_6.

igneous rock Rock that solidified from molten or partly molten material.

ion exchange A separation process in which ions are removed from a liquid solution to an insoluble solid (typically a polymer resin) and are replaced by ions from the solid.

isotope Any one of multiple forms of the same element having the same atomic number Z but different mass numbers A.

leaching The removal of soluble constituents from a solid by the action of a liquid solvent.

nuclide A species of atom existing for a measurable length of time and having an atomic number Z, a mass

 © 2004 Elsevier Inc. All rights reserved.

number A, and nuclear energy distinct from any other species of atom.

radionuclide A radioactive nuclide.

radiotoxicity A measure of the potential of a radioactive substance to cause harm to living tissue by radiation after introduction into the body.

reprocessing The separation of spent nuclear fuel into useful materials (especially uranium and plutonium) and wastes.

solvent extraction A process for removing chemical solutes from a liquid solution using a second immiscible liquid phase.

spent nuclear fuel Fuel after it has been used in a nuclear reactor, containing fission products, plutonium, and unreacted uranium.

transuranic element An element having an atomic number Z greater than 92.

tuff Rock formed from consolidated volcanic ash; welded tuff results when ash particles are fused together by heat and pressure.

vitrification The process of converting or incorporating nuclear wastes into a glass for permanent disposal.

yellowcake The solid product of uranium milling, nominally U_3O_8 but actually consisting of mixed uranium oxides, hydrides, and impurities.

Radioactive waste may be defined as any material that contains a higher concentration of radionuclides than is considered safe and for which no further use is contemplated. Nuclear waste is a radioactive by-product of nuclear power generation, nuclear weapons production, and the medical and industrial applications of radioactive substances. This article focuses on radioactive waste from civilian nuclear power. In 2002, there were 438 nuclear power plants operating in 31 nations, producing more than 16% of the electricity generated worldwide. In some countries, nuclear power plants supply more than half of the electricity generated. Worldwide, nuclear power generation produces more than 200,000 m^3 of nuclear waste annually, of which less than 5% is highly radioactive. The total volume of nuclear wastes is nearly 20 million m^3.

1. NUCLEAR FISSION

All commercial nuclear power plants rely on the fission of uranium and, to a lesser extent, plutonium. (An alternative cycle, based on thorium, has been explored but never commercialized.) Naturally occurring uranium consists of two major isotopes: U-235 (0.71% by weight) and U-238 (99.28% by weight). Only U-235 is fissile, meaning that it readily undergoes fission by the absorption of slow neutrons. The fission of U-235 can be represented by

$$n + {}^{235}\mathrm{U} \rightarrow \text{fission fragments} + 2.43n.$$

A fission reaction produces two (or more) fission fragments or products and additional neutrons (2.43 neutrons on average); the latter can initiate further fission reactions. Each fission reaction releases approximately 200 MeV (million electronvolts) of energy, predominantly as the kinetic energy of the fission fragments. In contrast, the chemical combustion of a hydrocarbon fuel typically releases less than 7 eV per molecule of CO_2 formed.

The more abundant uranium isotope U-238 is not fissile. However, U-238 is fertile, meaning that it can capture neutrons and be converted to plutonium-239:

$$n + {}^{238}\mathrm{U} \rightarrow {}^{239}\mathrm{Pu}.$$

Plutonium-239 is fissile:

$$n + {}^{239}\mathrm{Pu} \rightarrow \text{fission fragments} + 2.88n.$$

2. NUCLEAR FUEL CYCLE

The nuclear fuel cycle based on uranium is depicted in Fig. 1. The complete cycle consists of nine major steps: (1) mining and milling, (2) conversion, (3) enrichment, (4) fuel fabrication, (5) reactor operation, (6) interim storage, (7) reprocessing, (8) stabilization and immobilization, and (9) final disposal.

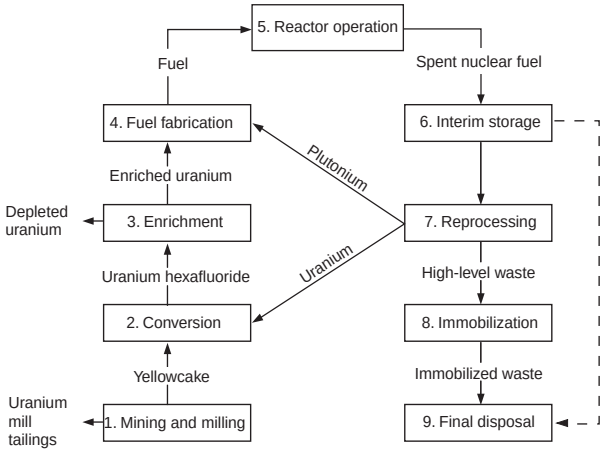


FIGURE 1 Nuclear fuel cycle based on uranium.

2.1 Mining and Milling

Uranium is excavated from open-pit or deep-shaft mines using conventional mining techniques (step 1 in the fuel cycle). The ore is crushed and the uranium is leached out using sulfuric acid or sodium carbonate and bicarbonate. The uranium is recovered from the liquid using processes such as ion exchange, solvent extraction, precipitation, and drying. The product is yellowcake, a solid consisting of U_3O_8 and various other oxides, hydroxides, and impurities.

2.2 Conversion

In the second step of the fuel cycle, yellowcake is reacted with hydrogen fluoride and fluorine to produce uranium hexafluoride (UF_6) or "hex." Depending on the temperature and pressure, hex may be handled as a solid, a liquid, or a gas.

2.3 Enrichment

Most commercial reactors require enriched fuel in which the concentration of U-235 has been increased from its natural value of 0.71% to as much as 5% by weight. (A few power reactors, most notably the early British gas-cooled reactors and the Canadian CANDU reactors, do not require enriched fuel.) Weapons-grade uranium must be enriched even further, to approximately 90% U-235. The most common enrichment processes are gaseous diffusion and centrifugation, both of which take advantage of the minute difference in mass between U-235 and U-238. The by-product from enrichment is depleted uranium, which contains less than 0.71% U-235 (typically 0.25–0.30%).

2.4 Fuel Fabrication

Fuel for commercial power reactors may contain uranium oxide (UO_2), metallic uranium, or mixed uranium and plutonium oxides ($UO_2 + PuO_2$ or "MOX"). The majority of commercial power reactors use uranium oxide, produced by chemically converting enriched UF_6 to UO_2 powder. The powder is pressed, sintered, and machined to form cylindrical pellets, which are then stacked inside metal tubes called fuel pins or fuel rods. A number of fuel pins are gathered to form fuel bundles or assemblies.

2.5 Reactor Operation

The core of a nuclear power plant that produces 1000 MW of electricity typically contains approximately 75 metric tons (165,000 pounds) of enriched uranium oxide fuel. Approximately 25 metric tons (55,000 pounds) of fresh fuel is required each year. Thus, each fuel assembly spends about 3 years in the reactor core. The heat from the fission reactions is used to generate steam for running power turbines. Most of the power output comes from the fission of U-235. However, some U-238 is transformed into Pu-239, the fission of which provides about a third of the reactor's energy output.

2.6 Interim Storage

When first withdrawn from the reactor, reactor fuel is both thermally hot and highly radioactive because of fission products. Spent nuclear fuel (SNF) assemblies are kept in cooling pools at the reactor site until their radioactivity and heat generation have declined to safer levels. This generally requires several months. (However, in the United States, spent fuel has remained in cooling ponds for decades awaiting the development of a permanent national disposal site.)

2.7 Reprocessing

Irradiated fuel from reactors may be reprocessed to remove fission products and to recover the uranium and plutonium. The fuel rods are chopped up and dissolved in acid, and the various components are separated chemically. In a true cycle, the uranium and plutonium are recycled for use as fuel. However, some countries (including the United States) employ the so-called once-through or open fuel cycle, which omits this step.

2.8 Stabilization and Immobilization

The high-level wastes from the reprocessing of SNF are mostly in liquid form. The liquid waste is calcined to produce a dry powder, which is then converted to a glass (a process called vitrification) or ceramic. (In a once-through cycle, this step is omitted because spent uranium oxide fuel is already in ceramic form.) The glass or ceramic is then sealed in corrosion-resistant canisters for long-term disposal.

2.9 Final Disposal

High-level wastes resulting from reprocessing (and SNF from a once-through cycle) must be deposited for hundreds or thousands of years to allow the radionuclides to decay to safe levels. Most proposals for final deposition envision some kind of deep

geological repository in stable crystalline rock, salt deposits, or clay.

3. RADIONUCLIDES IN NUCLEAR WASTE

More than 2500 nuclides have been discovered or created. Most of these are too short-lived to be of practical interest. The Index of Radioisotopes, published by the U.S. Nuclear Regulatory Commission (NRC), lists some 770 isotopes. Those of concern in radioactive wastes generally fall into one of three categories:

- *Activation products*. Substances used as coolants, moderators, and reactor structures can become radioactive by absorbing neutrons. These are usually light- or medium-mass nuclides.
- *Medium-mass fission products*. The fission of uranium produces more than 200 different radionuclides, covering about a third of the periodic table from copper to the lanthanides. Fission fragments tend to have neutron-rich nuclei with mass numbers from approximately 70 to 160, with peak production occurring near mass numbers 95 and 135. Most fission products are beta or beta/gamma emitters.
- *Heavy nuclei*. Nuclei having mass numbers between 232 and 246 are used as fuels for nuclear fission or are produced by neutron capture in fission reactions. Most of these are long-lived alpha emitters.

Table I lists some radionuclides of concern in nuclear wastes. These radionuclides are also discussed in detail in what follows.

3.1 Americium

Americium (Am, atomic number 95) is an artificial actinide. Its most important isotope is americium-241 (half-life 432.7 years), which is produced by the beta-decay of plutonium-241. Americium-241 is used commercially in applications such as smoke detectors, thickness gauges, and fluid level gauges. Although americium-241 has very high radiotoxicity, it is not very mobile in the environment. It decays by alpha emission to neptunium-237.

TABLE I

Some Radionuclides in Nuclear Wastes

Nuclide	Z	Half-life	Decay mode	Decay energy (MeV)	Relative radiotoxicity
Americium-241	95	432.7 years	α	5.4857	Very high
Carbon-14	6	5,730 years	β^-	0.157	Moderate
Cesium-137	55	30.17 years	β^-, γ	1.17	High
Cobalt-60	27	5.272 years	β^-, γ	2.824	High
Hydrogen-3	1	12.26 years	β^-	0.0186	Low
Iodine-129	53	16×10^6 years	β^-	0.193	Low
Iodine-131	53	8040 days	β^-	0.971	High
Iodine-133	53	20.8 hours	β^-, γ	1.76	Moderate
Krypton-85	36	10.72 years	β^-, γ	0.687	Low
Neptunium-237	93	2.14×10^6 years	α	4.957	Very high
Plutonium-238	94	87.7 years	α	5.593	Very high
Plutonium-239	94	24,110 years	α	5.244	Very high
Plutonium-242	94	376,000 years	α	4.983	Very high
Radium-226	88	1,600 years	α, γ	4.870	Very high
Radium-228	88	5.75 years	β^-	0.045	Very high
Radon-222	86	3.82 days	α	5.590	Moderate
Ruthenium-106	44	372.2 days	β^-	0.039	High
Strontium-90	38	29 years	β^-	0.546	High
Technitium-99	43	213,000 years	β^-	0.293	Low
Thorium-232	90	1.40×10^{10} years	α	4.081	High
Uranium-234	92	2.46×10^5 years	α	4.776	Very high
Uranium-235	92	7.04×10^8 years	α, γ	4.679	Low
Uranium-238	92	4.46×10^9 years	α	4.039	Low

3.2 Carbon

Carbon (C, atomic number 6) occurs in nature predominantly as the stable isotopes carbon-12 (98.89%) and carbon-13 (1.1%). Its most important radioactive isotope is carbon-14, a weak beta-emitter having a half-life of 5730 years. Carbon-14 is formed naturally in the upper atmosphere by the action of cosmic rays on nitrogen. It is also produced by neutron capture in nuclear explosions and fission reactors if nitrogen-containing materials are present. Carbon-14 can replace carbon-12 in biological systems. The radiotoxicity of carbon-14 is moderate.

3.3 Cesium

Cesium (Cs, atomic number 55) is an alkali metal having 40 isotopes, ranging in mass from 112 to 151. Of these, the fission product cesium-137 (half-life 30.17 years) is of greatest concern. Cesium-137 is one of the nuclides most responsible for the radioactivity of used reactor fuel and high-level waste (with the other being strontium-90). Cesium-137 is highly radiotoxic. It is widely used in moisture, liquid level, and thickness gauges as well as to treat cancer.

3.4 Cobalt

The transition metal cobalt (Co, atomic number 27) is brittle, hard, and magnetic. Naturally occurring cobalt, consisting of the isotope cobalt-59, is used to make very hard and corrosion-resistant alloys. The absorption of a neutron converts cobalt-59 to radioactive cobalt-60 (half-life 5.272 years), which decays by the emission of a beta particle and a high-energy gamma ray to nickel-60.

3.5 Hydrogen

Hydrogen (H, atomic number 1) has two stable isotopes: hydrogen-1 and hydrogen-2 (deuterium). Hydrogen-3 (tritium) is a low-energy beta-emitter having a half-life of 12.26 years, produced naturally by the action of cosmic rays in the upper atmosphere. Tritium also occurs as a product in a small fraction of fission reactions and is produced commercially by the irradiation of lithium-6. Tritium is most likely to be found bound to oxygen to form tritiated water, which has chemical properties that are virtually identical to those of ordinary water. Tritium is usually taken into the body by inhalation and skin absorption, although ingestion is also possible. Blood flow distributes the tritiated water evenly throughout the soft tissues of the body. Its radiotoxicity is low.

3.6 Iodine

Iodine (I, atomic number 53) is a halogen occurring below bromine in the periodic table. Two radioactive isotopes of iodine are of importance in nuclear wastes. Iodine-129 (half-life 1.6×10^7 years) and iodine-131 (half-life 8.040 days) occur as fission products. Both are beta-emitters that tend to concentrate in the thyroid. The radiotoxicity of iodine-129 is low; the radiotoxicity of iodine-131 is high. Other isotopes (e.g., iodine-123, iodine-124) are commonly used in medical imaging and treatment.

3.7 Krypton

Krypton (Kr, atomic number 36) is one of the noble gases, which as a group have little tendency to react chemically with other elements. A total of 32 isotopes of krypton have been identified, having atomic masses ranging from 69 to 100. Six of these are stable. The fission product krypton-85 is a beta-emitter having a half-life of 10.72 years. Because it is chemically inert, krypton is difficult to separate from process streams. Although krypton is not absorbed appreciably in body tissues, it may be inhaled and affect the lungs. Its radiotoxicity is low.

3.8 Neptunium

Neptunium (Np, atomic number 93) is the first transuranic element. Neptunium-239 (half-life 2.36 days) occurs as an intermediate in the formation of plutonium from uranium but is too short-lived to be of concern in most wastes. In contrast, neptunium-237 (half-life 2.14×10^6 years) is a major source of alpha radiation in spent fuel and high-level wastes. It is formed by alpha decay of americium-241 or beta decay of uranium-238. The radiotoxicity of neptunium-237 is very high.

3.9 Plutonium

Isotopes of plutonium (Pu, atomic number 94) are generated by neutron capture in uranium-238 or thorium-232. Plutonium-239 (half-life 24,110 years) is of greatest concern because it is fissionable. Other important isotopes are plutonium-238 (half-life 87.7 years), plutonium-240 (half-life 6564 years), plutonium-241 (half-life 13 years), and plutonium-242

(half-life 3.76×10^5 years). All of these isotopes have very high radiotoxicity.

3.10 Radium

Radium (Ra, atomic number 88) is a naturally occurring radioactive element. Radium-228 (half-life 5.76 years) is a decay product of thorium; radium-226 (half-life 1600 years) is a decay product of uranium. Radium-226 and its daughters (including radon-222) are the major sources of radioactivity in uranium mine tailings. Chemically, radium is an alkaline earth metal like calcium and tends to concentrate in the bone. The radiotoxicity of radium is very high.

3.11 Radon

Radon (Rn, atomic number 86) is a radioactive noble gas produced by the decay of radium. The two most important isotopes are the alpha-emitters radon-220 (half-life 54.6 s) and radon-222 (half-life 3.82 days). Being a gas, radon is highly mobile; in some areas, radon is a concern because it can accumulate in buildings. Both radon-220 and radon-222 are moderately radiotoxic.

3.12 Ruthenium

Ruthenium (Ru, atomic number 44) belongs to the platinum family of metals. It has seven stable natural isotopes. The most important radioactive isotope is ruthenium-106 (half-life 372.2 days), a fission product that decays by low-energy beta emission to rhodium-106 (a high-energy beta/gamma emitter with half-life of 30 s). Ruthenium-106 has high radiotoxicity.

3.13 Strontium

Strontium (Sr, atomic number 38) is an alkaline earth metal lying below calcium in the periodic table. Four stable isotopes are found in nature, and 16 radioactive isotopes have been produced artificially. Most of the radioactive isotopes of strontium are short-lived and of little interest. The exception is strontium-90 (half-life 29.1 years), a beta-emitter that is produced as a fission fragment. Much of the radioactivity of SNF results from strontium-90. Because its chemical properties resemble those of calcium, strontium-90 tends to concentrate in the bones. It is highly radiotoxic, decaying into yttrium-90 (itself an energetic beta-emitter of moderate radiotoxicity).

3.14 Technetium

The first artificially produced element, technetium (Tc, atomic number 43) occurs in nature only in minute amounts. At least 34 isotopes have been produced, having atomic masses ranging from 85 to 118; of these, the beta-emitter technetium-99 (half-life 2.13×10^5 years) is of special concern because it is produced in fission reactors, both as a fission fragment and as an activation product. Like iodine, technetium is absorbed by the thyroid. The radiotoxicity of technetium is relatively low.

3.15 Thorium

Thorium (Th, atomic number 90) is a naturally occurring radioactive element, several times more abundant than uranium, which has been considered as the basis for an alternative fuel cycle. The most important natural isotope is thorium-232 (half-life 14.0×10 years), which is weakly radioactive, decaying by alpha and gamma emission to radium-228. The radiotoxicity of thorium-232 is relatively low. Another isotope, thorium-230 (half-life 75,400 years) is produced naturally by the decay of uranium-234. The radiotoxicity of thorium-230 is very high. A third isotope, thorium-228 (half-life 1.9 years) also has very high radiotoxicity.

3.16 Uranium

Uranium (U, atomic number 92) is the heaviest element found in appreciable amounts in nature, making up 2 to 4 grams per metric ton of the earth's crust. Natural uranium consists of 99.283% by weight uranium-238 (half-life 4.468×10^9 years), 0.711% uranium-235 (half-life 7.04×10^8 years), and 0.005% uranium-234 (half-life 245,000 years). All three isotopes are alpha-emitters. Natural uranium has relatively low radiotoxicity.

4. CLASSIFICATION OF RADIOACTIVE WASTE

There is no generally accepted classification scheme for radioactive wastes. The following categories are commonly defined in the United States:

- Uranium mill tailings (UMTs) are the residues from the mining and milling of uranium ores. Most uranium ores in the United States contain uranium at a concentration of approximately 0.2% (2000 ppm)

by weight. Consequently, the production of 1 kg of uranium metal requires the excavation of 500 kg (or more) of ore. Milling creates additional solid and liquid wastes. Tailings generally contain relatively low concentrations of radioactive materials; however, because of the high volumes of tailings, their overall radioactivity can be high. The principal radionuclides of concern in UMTs are radium-226 (half-life 1600 years) and its daughter radon-222 (half-life 3.8 days).

- Depleted uranium, the major by-product from the enrichment process, is uranium containing less than 0.71% uranium-235. Depleted uranium currently has few civilian uses but could eventually be converted to fissile plutonium in breeder reactors. The U.S. Department of Energy (DOE) stores approximately 730,000 metric tons of depleted uranium, most of it as uranium hexafluoride.
- Low-level waste (LLW) has relatively low levels of radioactivity and contains little or no transuranic radioisotopes (i.e., radionuclides beyond uranium in the periodic table). LLW may be further divided into classes A, B, and C, with A being the least hazardous and C being the most hazardous. (Some LLW is classified as "greater than C.") Mixed low-level waste (MLLW) contains both low-level radioactive and chemically hazardous substances. Although not shown in Fig. 1, LLW is generated at every step of the nuclear fuel cycle and may include used protective clothing, rags, paper, tools, filter elements, and so on.
- Spent nuclear fuel (SNF) is the irradiated fuel discharged from nuclear reactors. It contains fission products, plutonium, and unreacted uranium and is both thermally hot and highly radioactive.
- Transuranic waste (TRUW) contains alpha-emitting radionuclides having atomic numbers greater than 92. To qualify as TRUW under U.S. law, the radionuclides must possess half-lives greater than 20 years and be present in concentrations greater than 100 nCi/g. Although their radioactivity is relatively low, the long lives of the radioisotopes require long-term storage. Virtually all of the TRUW in the United States is the by-product of nuclear weapons production.
- High-level waste (HLW) includes the residues resulting from the reprocessing of spent reactor fuel and the production of nuclear weapons. HLW contains highly radioactive fission products and small amounts of plutonium isotopes. SNF may be considered as HLW in countries that do not reprocess fuel.

Under the foregoing classification scheme, commonly followed in the United States, wastes are categorized according to their source as well as their radioactivity. However, other countries tend to categorize wastes according to their radioactivity alone. For example, in the United Kingdom, solid wastes are classified as high-, intermediate-, low-, or very low-level wastes as follows:

- Very low-level waste (VLLW) can safely be discarded as ordinary refuse. To qualify, bulk waste must contain less than 400 kBq of beta/gamma activity per 0.1 m^3 of material; individual items must contain less than 40 kBq of beta/gamma activity.
- Low-level waste (LLW) has higher levels of radioactivity than can be safely handled as ordinary refuse but not exceeding 4 GBq per metric ton of alpha radiation or 12 GBq per metric ton of beta/gamma radiation.
- Intermediate-level waste (ILW) has higher radioactivity levels than does LLW but not so high that heating becomes a concern.
- High-level waste (HLW) has such high levels of radioactivity that appreciable thermal heating occurs. Consequently, storage and processing facilities have to be designed to contain the radiation and dissipate the generated heat.

5. DISPOSAL OF LOW-LEVEL WASTE

LLW accounts for approximately 90% of the total volume of all wastes produced in the nuclear fuel cycle but only 1% of the total radioactivity. Various methods have been adopted for disposing of LLW, depending on the type of waste and its levels of radioactivity. For example, the NRC requires that Class A and B wastes be contained for at least 100 years, and Class C waste must be contained for 500 years. Combinations of natural and engineered barriers are used to prevent radioactivity from escaping to the environment. A major goal is to prevent the intrusion of ground water and surface water.

5.1 Shallow Land Burial

The cheapest way to dispose of LLW is to bury it in shallow trenches dug in stable soil (Fig. 2). A typical disposal trench may be 6 to 10 m deep, 30 m wide, and 100 to 300 m long. In some cases, the trench may be lined with a layer of clay, polymer, or concrete.

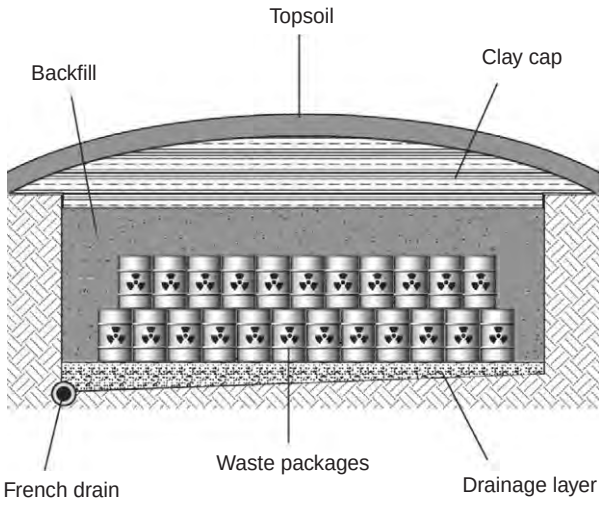


FIGURE 2 Shallow land burial for LLW.

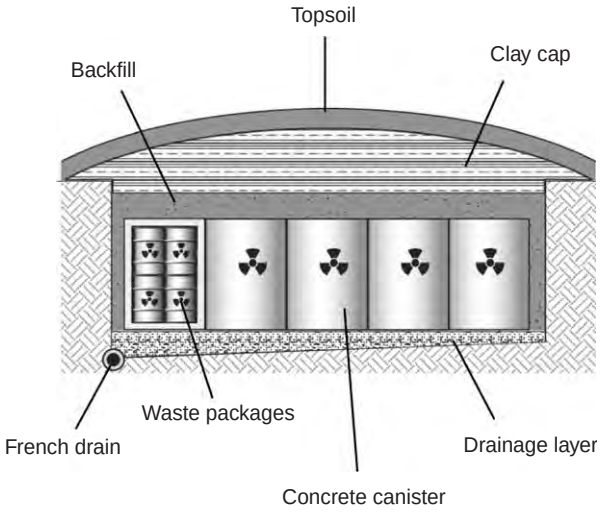


FIGURE 3 Modular concrete canister disposal.

The floor of the trench is sloped and covered with a layer of sand, gravel, or crushed stone. A drainage system keeps the waste containers dry. Once the LLW containers are in place, the trench is backfilled with sand or earth, which is then compacted and capped with a layer of clay to reduce water infiltration. The clay cap may be stabilized by a layer of topsoil planted with vegetation. Shallow land burial (SLB) has been used throughout the world.

5.2 Modular Concrete Canister

Modular concrete canister (MCC) disposal (Fig. 3) differs from SLB in that the waste containers are first placed inside concrete canisters, which are then buried in trenches. The canisters serve as an additional barrier against the release of radioactivity.

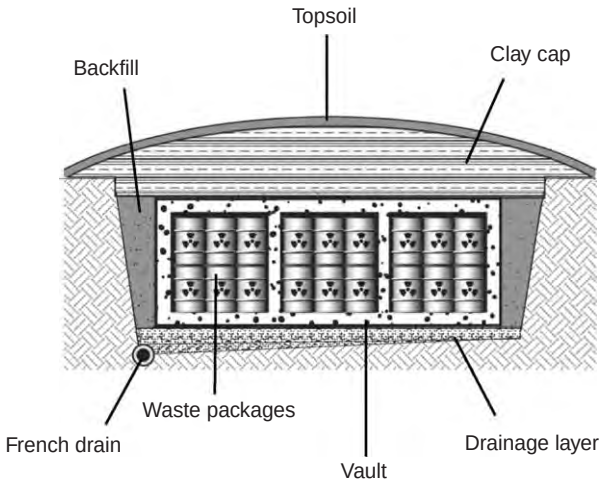


FIGURE 4 Belowground vault disposal.

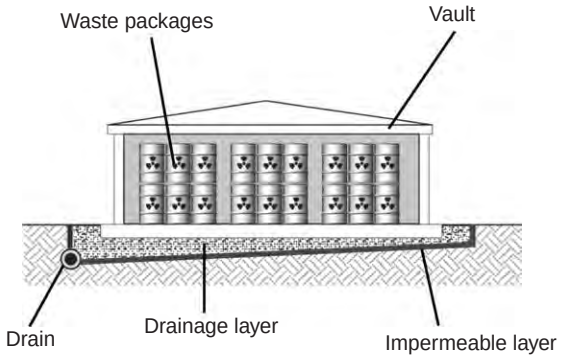


FIGURE 5 Aboveground vault disposal.

5.3 Belowground Vault

In belowground vault (BGV) disposal (Fig. 4), a specially designed concrete, metal, or polymer vault is placed in a trench below ground level. The floor of the trench is typically sloped and equipped with drains. After the waste containers are placed in the vault, the trench is backfilled with permeable sand or gravel and capped with impermeable clay, topsoil, and vegetation. BGV disposal is used in Japan at the Low-Level Waste Disposal Centre at Rokkasho.

5.4 Aboveground Vault

An aboveground vault (AGV) is a concrete, metal, or masonry structure built above ground level (Fig. 5). A vault may be a simple reinforced concrete box with a poured concrete roof, or it may be a more elaborate

structure resembling a large warehouse. The principal advantage of an AGV is that it allows easy recovery of wastes for further processing or reuse. Its principal disadvantage is that the structure is not protected by earth from the effects of weather, intrusions, or accidents. Consequently, an AGV is often considered more for temporary storage than for permanent disposal. AGVs have been used at the West Valley Demonstration Project in New York and at the Point Lepreu and Bruce sites in Canada.

5.5 Earth-Mounded Concrete Bunker

An earth-mounded concrete bunker (EMCB) is a structure built partly below ground and partly above ground (Fig. 6). Wastes are segregated according to activity; higher activity (class B/C) wastes are embedded in concrete vaults or "monoliths" below ground level, and lower activity (class A) wastes are then placed atop the vaults. The wastes are covered with an impermeable layer of clay and a layer of topsoil to form a mound or tumulus. This method is used in France at the Centre de La Manche and the Centre de l'Aube as well as in Canada at the Whiteshell Nuclear Research Establishment.

5.6 Shaft Disposal

In a typical shaft disposal (SD) scheme, a hole some 3 m in diameter is sunk 40 or 50 m into the ground (but well above the water table). The walls of the shaft may be lined with clay, metal, or concrete, and the bottom of the shaft may be left unlined for drainage.

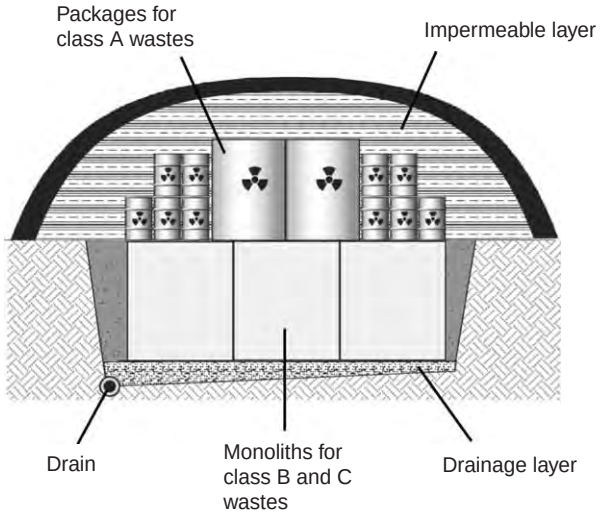


FIGURE 6 Earth-mounded concrete bunker (or tumulus) disposal of LLW.

After the waste containers have been stacked in the shaft, it is backfilled with sand or clay and capped with concrete. Shaft disposal of LLWs has been used at a number of sites in the United States and Canada.

5.7 Deep Geological Disposal

Some countries have chosen to dispose of LLW and ILW in deep geological repositories similar to those planned for HLW. These repositories are to be located in stable geological media such as igneous rock, salt, or clay.

6. MANAGEMENT OF SPENT NUCLEAR FUEL

Nuclear fuel rods are highly radioactive when first withdrawn from the reactor core because they contain fission products, principally relatively short-lived radionuclides. For this reason, the SNF is transferred to an interim storage facility adjacent to the reactor to allow the radioactivity to decline to safer levels. The radioactivity of the spent fuel decreases by approximately 99% during a year of storage.

Most interim storage facilities keep SNF in pools of water, commonly known as cooling ponds. The water serves both as a radiation shield and a heat transfer medium. Where cooling ponds have become crowded, some facilities put the SNF in dry storage instead of building new ponds. Dry storage facilities may keep SNF in concrete or metal casks, reinforced concrete vaults, or dry wells. Cooling is accomplished by natural convection of ambient air.

As shown in Table II, there is no uniform policy regarding the management of SNF. Some countries (e.g., Japan, France, Russia, the United Kingdom) currently reprocess their SNF. Other countries (e.g.,

TABLE II

National Policies on Reprocessing of Spent Nuclear Fuel

Reprocess (closed cycle)
Belgium, China, France, Germany, India, Japan, The Netherlands, Russia, Switzerland, United Kingdom
Do not reprocess (once-through cycle)
Canada, Finland, Lithuania, Slovenia, Spain, South Africa, Sweden, United States
Undecided on reprocessing
Argentina, Brazil, Czech Republic, Hungary, Mexico, Pakistan, Slovenia, South Korea, Taiwan

Canada, Finland, the United States) do not reprocess. Still other countries (e.g., Czech Republic, South Korea, Taiwan) have deferred the decision on whether to reprocess, relying on storage of SNF for the near term.

7. DISPOSAL OF HIGH-LEVEL WASTES

7.1 Objectives

Disposal of HLW (and SNF treated as waste) will require the isolation of the wastes from the biosphere for thousands of years. The principal concern is the contamination of ground water that may be transported to the biosphere. The design of a typical repository is intended to accomplish all or most of the following objectives:

- Delay the intrusion of water into the repository.
- Delay the release of radionuclides from the waste package.
- Retard the dissolution of radionuclides.
- Limit the migration of any radionuclides that escape containment.
- Dilute any escaped radionuclides to safe levels.
- Permit retrieval of the SNF for reuse.

A variety of methods for the disposal of HLW/SNF have been studied over the years. Most employ multiple barriers to provide "defense in depth." A typical scheme for HLW might involve solidification and vitrification of the waste. The vitrified wastes would then be placed inside one or more specially engineered metal, plastic, or ceramic containers. The containers would then be placed in cavities in a natural medium such as rock, clay, or salt and perhaps backfilled with a material such as bentonite, which swells when wet and slows the passage of water.

7.2 Seabed Disposal

In the past, low-level nuclear wastes have been dumped somewhat haphazardly into the ocean. Although most nations ended ocean disposal in the 1970s, several proposals have been put forth for the long-term disposal of HLW in the seabed. One approach would place the waste in specially designed canisters that would then be buried in deep sediments far from inhabited coasts. The sediments would provide a natural barrier to the movement of radionuclides, and the seawater would act as a dilution medium for any radionuclides that managed to escape containment. Seabed disposal is still considered a viable option for the future. However, an international agreement, the 1993 London Convention, banned the disposal of all radioactive wastes at sea for 25 years, after which such disposal methods may be reconsidered every 25 years.

7.3 Very Deep Hole Disposal

Nuclear waste containers could be stacked in holes as deep as 10,000 m ($\sim$6 miles) underground, putting them out of reach of flowing ground water. Despite uncertainties about the behavior of both the waste containers and the surrounding rock at such depths, recent advances in drilling technology have renewed interest in this approach.

7.4 Space Disposal

The National Aeronautical and Space Administration (NASA) and the DOE studied the possibility of disposing of nuclear waste in space. Various schemes were considered, including placing waste in orbit around the earth, the sun, or the moon; shooting waste into the sun; placing waste on the moon; and sending waste into deep space. Although space disposal would place the waste permanently outside the biosphere, it would be expensive. Also, there is the possibility of an accident during launch, which could scatter waste over a large area.

7.5 Ice Sheet Disposal

Radioactive waste could be buried in or under the ice sheets that cover much of the continent of Antarctica. Because the ice can be several kilometers thick and there is no permanent human population on the continent, the waste would be effectively isolated from human contact. However, ice disposal would be expensive because of the harsh climate and the need to transport the waste to a remote site. Climate changes also could cause melting of the polar ice, releasing the waste. Moreover, the Antarctic Treaty of 1959 prohibits the disposal of radioactive waste on the continent.

7.6 Deep Geological Disposal

Most current plans for the long-term disposal of nuclear waste involve the excavation of deep repositories in stable geological formations such as igneous rock, salt deposits, and clay. Each of these has advantages and disadvantages.

TABLE III

Plans for High-Level Waste Repositories

Country	Geological medium	Estimated opening	Status
Belgium	Clay	2035 or later	Searching for site
Canada	Granite	2035 or later	Reviewing repository concept
Finland	Crystalline bedrock	2020	Site selected (Olkiluoto)
France	Granite or clay	2020 or later	Developing repository concept
Germany	Salt	Unknown	Moratorium on development
Japan	Granite or sedimentary rock	2030 or later	Searching for site
Russia	Not selected	Unknown	Searching for site
Sweden	Crystalline rock	2020	Searching for site
Switzerland	Crystalline rock or clay	2020 or later	Searching for site
United Kingdom	Not selected	After 2040	Delaying decision until 2040
United States	Welded tuff	2010	Site selected (Yucca Mountain)

7.6.1 Igneous Rock

Intact igneous rock, such as granite or tuff, tolerates relatively high temperatures and resists the flow of water. On the other hand, hard rock may fracture, permitting the intrusion of water through the resulting fissures. Both the Yucca Mountain repository in the United States and the SFR repository in Sweden are sited in igneous rock.

7.6.2 Salt Deposits

The presence of a stable salt deposit indicates that little or no flowing water has been present at the site for geological ages. Left undisturbed, the salt should remain dry for thousands of years. Salt is easily mined. It deforms plastically under the influence of heat and pressure, eventually flowing around the waste and sealing it in a tight, resolidified mass. Salt has very low permeability. It also has a relatively high thermal conductivity, which would aid in the dissipation of heat from the waste. On the other hand, if water found its way into the repository, the resulting salt brine would tend to be highly corrosive. The Waste Isolation Pilot Plant (WIPP) in the United States is located in bedded rock salt. The Gorlben repository in Germany is located in a salt dome.

7.6.3 Clay

By definition, clays are unconsolidated sediments consisting of particles less than 1/256 mm in diameter. (Clay particles are too small to feel gritty between the fingers or teeth.) Clays are commonly composed of hydrous aluminum silicates having a sheet–silicate structure, producing grains that tend to be flat. Clay resists the intrusion of water and retards the migration of radionuclides. Like salt, clay tends to be self-sealing. Clay layers have been under consideration for repositories in Belgium, France, and Switzerland.

Table III summarizes the HLW repository plans of major nuclear countries. The nuclear waste management policies of the three leading nuclear countries (the United States, France, and Japan) are examined in more detail in the following sections.

8. RADIOACTIVE WASTE MANAGEMENT IN THE UNITED STATES

The United States has by far the largest civilian nuclear program in the world, with a total of 104 reactors having a power capacity of 97.5 GW. In 2000, these plants produced approximately 20% of the country's electricity.

8.1 History

The world's first nuclear reactor, built under the direction of Enrico Fermi, was put into operation at the University of Chicago on December 2, 1942. This led to the rapid development of an immense industrial complex for the wartime production of weapons-grade uranium and plutonium under the aegis of the Manhattan Project. On July 16, 1945—less than 3 years after Fermi's reactor achieved the first chain reaction—the first atomic bomb (a plutonium device) was exploded near Alamogordo, New Mexico. In August 1945, the United States dropped atomic bombs on Hiroshima and Nagasaki, Japan.

The U.S. Atomic Energy Commission (AEC) was established in 1946 and given the responsibility for both civilian and military nuclear programs. Initially, waste management was given lower priority than the development of nuclear weapons and (later) a nuclear power industry. It was widely expected that plants would be constructed to reprocess SNF. Uranium and plutonium would be reused; reprocessing wastes would be solidified and transferred to a permanent repository for disposal.

The National Academy of Sciences recommended in 1957, and again in 1970, that a federal waste repository be built in a stable bedded salt deposit. Accordingly, in 1970, the AEC investigated salt beds near Lyons, Kansas, as a possible site for a repository. However, the Lyons site was rejected in 1972 because of concerns about its geology and hydrology as well as the large number of boreholes left by previous oil and gas drilling in the area. In 1974, the AEC began exploring salt beds in southeastern New Mexico. The next year, a site for the WIPP was selected.

In 1974, the AEC was abolished by Congress, and its responsibilities were transferred to two new agencies. The Energy Research and Development Administration (ERDA) was created to develop and promote nuclear power, and the NRC was created to regulate civilian nuclear programs. In 1976, the U.S. Environmental Protection Agency (EPA) was given the responsibility of setting radiation protection standards for nuclear wastes. The next year, the ERDA was replaced by a new cabinet-level agency, the DOE. The new department was also given responsibility for nuclear weapons programs.

Given these organizational changes, it is perhaps not surprising that the executive branch of the federal government made little progress in formulating a coherent policy for nuclear waste management during the 1970s. Congress has repeatedly stepped in and formulated policies for waste disposal. Some of the major pieces of legislation include the following:

- The Low-Level Radioactive Waste Policy Act of 1980 requires each state to be responsible for disposing of LLW generated within its borders, either by providing a suitable disposal site or by entering into compacts with other states to develop a shared site. The disposal sites were to be in operation by 1986.
- The Nuclear Waste Policy Act (NWPA) of 1982 created the Office of Civilian Radioactive Waste Management in the DOE and gave it the responsibility for developing a system to manage SNF and HLW. The act also directed the secretary of energy to nominate five sites for study as possible SNF/HLW repositories. Based on these initial studies, the secretary was to recommend (by January 1, 1985) three of the five sites for intensive study. Finally, one site was to be recommended by March 31, 1987. The goal was to have the first repository in operation by 1998.
- The Low-Level Radioactive Waste Policy Amendments Act (LLRWPAA) of 1985 was crafted to address difficulties arising out of the Low-Level Waste Policy Act of 1980. The formation of state compacts and the development of disposal facilities had taken longer than was expected. The LLRWPAA kept open the three existing commercial disposal sites and set regulations for the disposal of LLW.
- The Nuclear Waste Policy Amendments Act of 1987 specified that Yucca Mountain be the sole site for study as an HLW repository instead of having three site studies proceeding in parallel. The amendment also canceled the requirement from the NWPA of 1982 that a second repository be created. The selection of Yucca Mountain was a political decision, not a scientific or technical one; local opposition to other sites made them less attractive politically.
- The Waste Isolation Pilot Plant Land Withdrawal Act of 1992 transferred control of the WIPP site from the Department of Interior to the DOE. The act also gave the EPA responsibility for ensuring compliance with environmental regulations.
- The Energy Policy Act of 1992 was a large and comprehensive piece of legislation covering a broad range of energy policy issues, including energy efficiency and conservation, integrated resource planning, alternative fuels, and competitive restructuring of wholesale electricity markets. The act required that the EPA set standards for the Yucca Mountain repository based on recommendations by the National Academy of Sciences. The NRC was to assume long-term responsibility for the repository.

8.2 Low-Level Waste Repositories

Six commercial LLW disposal sites have been created in the United States: Barnwell, North Carolina; Beatty, Nevada; Richland, Washington; Maxey Flats, Kentucky; Sheffield, Illinois; and West Valley, New York. All of these sites have used some form of SLB. Because of leaks, the latter three sites have been closed.

Low-level military wastes have been stored at a number of sites around the country, including Amarillo, Texas; Hanford, Washington; the Idaho National Engineering and Environmental Laboratory;

Los Alamos, New Mexico; Paducah, Kentucky; and Savannah River, South Carolina.

8.3 Waste Isolation Pilot Plant

The WIPP is an underground waste repository located in bedded salt formations in southeastern New Mexico, approximately 40 km (25 miles) east of Carlsbad. The WIPP repository was created for the permanent disposal of TRUW resulting from nuclear weapons production. It is the first deep geological repository designed specifically for long-term disposal of nuclear waste.

The waste destined for the WIPP site consists of a variety of materials contaminated with transuranic isotopes, including items such as protective clothing, laboratory equipment, used reagents, and solidified sludge. The principal radionuclides are plutonium, americium, and neptunium; small amounts of thorium, uranium, and other nuclides may also be present. Although the overall radioactivity of TRUW is much lower than that of SNF and other high-level waste, TRUW contains long-lived radionuclides that remain hazardous for thousands of years.

The WIPP repository is located in the 240 million-year-old salt beds of the Permian Salado Formation. The salt beds are the remains of a shallow saltwater sea that once covered what is now west Texas and southeastern New Mexico. The Salado Formation is 200 to 400 m thick, consisting mostly of halite (rock salt) containing thin layers (0.1–1.0 m thick) of nonhalite materials such as anhydrite (calcium sulfate) and potash.

The WIPP repository is situated 658 m (2160 feet) below the surface. It comprises a network of excavations connected to the surface by four vertical shafts: a salt-handling shaft, a TRUW shaft, an air intake shaft, and an air exhaust shaft. Underground roadways called drifts lead from the shafts to the waste disposal areas. TRUW is placed in rectangular rooms that are 4 m high, 10 m wide, and 91 m long. Adjacent rooms are separated by pillars of unexcavated salt that are 31.5 m wide. A group of seven rooms is called a panel. A total of eight panels are planned, but only one had been excavated by the time the first shipment of waste arrived in 1999. Other tunnels and rooms have been excavated for various scientific experiments.

8.4 Yucca Mountain

The Yucca Mountain repository is located in the Mojave Desert of southern Nevada, approximately 120 km (75 miles) northwest of Las Vegas. The repository is intended for the permanent disposal of SNF and HLW from both civilian and defense sources.

As planned, the Yucca Mountain repository will hold SNF that is currently kept in 131 temporary storage facilities in the United States. By 2002, these sites contained some 45,000 metric tons of SNF from commercial reactors and 2500 metric tons from defense reactors. Another 380,000 m^3 of liquid HLW from nuclear weapons production was stored at sites operated by the DOE; this waste is to be stabilized in the form of borosilicate glass.

Yucca Mountain is composed of tuff, a material formed from volcanic ash and dust. Some of the tuff is welded, meaning that the fine glassy particles of the tuff have been fused together by natural heat and pressure. The repository would be sited at a depth of approximately 350 m in a layer of welded tuff known as the Topapah Spring Member. Above and below the repository are layers of nonwelded tuffs that contain few fractures.

Yucca Mountain is located in an arid region, where the mean precipitation is less than 20 cm per year and an estimated 95% of the water runs off or evaporates before it can seep into the ground. The water table lies approximately 300 m (1000 feet) below the level of the repository. Ground water tends to flow to the south toward the Amargosa Desert, which lies in an isolated hydrological basin having no outlet to the ocean.

In July 2002, the U.S. Congress voted in favor of a resolution recommending that Yucca Mountain be considered for development as a permanent repository. At that time, no final design for the repository had been adopted, but two cases were being considered.

The base or statutory case would provide an eventual storage capacity of 70,000 metric tons of heavy metal (MTHM). The base case repository would store waste containers in 58 tunnels called emplacement drifts, each 5.5 m (18 feet) in diameter. The emplacement drifts would be spaced 81 m (266 feet) apart, and their total length would be 56,222 m (184,455 feet). An additional 12,988 m (42,612 feet) of excavations would be created for access to the repository, with another 6542 m (21,463 feet) of excavations being created for the exhaust main.

The full inventory case would have a capacity of 97,000 MTHM. It would consist of 90 emplacement drifts having a total length of 74,214 m (243,484 feet). The design could be expanded to accommodate as much as 119,000 MTHM if such expansion ever becomes necessary.

9. RADIOACTIVE WASTE MANAGEMENT IN FRANCE

France has the second largest civilian nuclear program in the world, with 59 reactors in operation. In 2000, French nuclear plants had a total capacity of 63.2 GW and produced 76% of the country's power, a higher percentage than in any other country.

9.1 Low- and Intermediate-Level Wastes

Disposal of nuclear waste is the responsibility of ANDRA (Agence Nationale pour la Gestion des Déchets Radioactifs). Waste is classified according to half-life as well as activity. Waste having half-life less than 30 years is designated as "short-lived" (SL), and waste having half-life greater than 30 years is designated as "long-lived" (LL). Four activity levels are defined: very low-level waste (VLLW), low-level waste (LLW), intermediate-level waste (ILW), and high-level waste (HLW).

The Centre de la Manche Disposal Facility, located near Cherbourg in Normandy, operated from 1969 to 1994 as a disposal site for LLW and ILW. Initially, waste was placed in simple trenches. Over the years, the EMCB concept was developed. The la Manche facility holds nearly 530,000 m^3 of waste and is no longer accepting waste shipments.

In 1992, the Centre de l'Aube Disposal Facility near Troyes began accepting low- and intermediate-level waste (LILW-SL). The Centre de l'Aube has a capacity of 1 million m^3, expandable to 2 million m^3.

9.2 Management of Spent Nuclear Fuel

SNF is kept in storage ponds for a year to allow its radioactivity to decline to safe levels, after which it is transported to a reprocessing plant.

France has reprocessed its own nuclear fuel since 1958 and has provided reprocessing services for Belgium, Germany, Japan, The Netherlands, and Switzerland. The French nuclear fuel company COGEMA (Compagnie Générale des Matiéres Nucléaires) currently operates two reprocessing plants at La Hague, west of Cherbourg in Normandy.

9.3 High-Level Wastes

HLW from reprocessing is vitrified at the reprocessing plant in La Hague. Vitrified HLW from French sources is currently stored at La Hague awaiting disposal. Vitrified HLW from foreign sources is returned to the countries of origin.

Since 1991, ANDRA and CEA (Commissariat à l'Energie Atomique) have been studying several options for long-lived radioactive waste, including deep geological disposal in clay or granite. A decision to begin construction on a deep repository could be made as early as 2006.

10. RADIOACTIVE WASTE MANAGEMENT IN JAPAN

Japan operates 53 nuclear power plants, giving it the third-largest civilian nuclear program. In 2000, Japan's nuclear plants had a total capacity of 43.5 GW, producing about a third of the country's electrical power.

10.1 Low-Level Wastes

Since 1992, Japan has operated the Low-Level Waste Disposal Centre at Rokkasho at the northern end of Honshu Island. The BGV is the preferred method of disposal. Ultimately, the facility at Rokkasho will hold 600,000 m^3 of waste.

10.2 Management of Spent Nuclear Fuel

SNF is stored in cooling ponds for 4 years or longer before being sent on for reprocessing.

Since 1977, Japan has operated a small reprocessing plant at Tokai Mura for SNF; however, most SNF is shipped to France or Great Britain for reprocessing. Japan is constructing its own reprocessing plant at the Rokkasho nuclear complex. It is expected to begin full-scale operations in 2005.

10.3 High-Level Wastes

Vitrified nuclear wastes are returned to Japan from the reprocessor. High-level vitrified wastes are sent to the High-Level Radioactive Waste Storage Facility at Rokkasho to be stored for 30 to 50 years.

In 2000, the Nuclear Waste Management Organization (NUMO) was created to develop a disposal repository in granite or sedimentary rock. Construction of the repository is expected to begin around 2030.

SEE ALSO THE FOLLOWING ARTICLES

Nuclear Engineering • Nuclear Fuel: Design and Fabrication • Nuclear Fuel Reprocessing • Nuclear Fusion Reactors • Nuclear Power, History of • Nuclear Power Plants, Decommissioning of • Nuclear Power: Risk Analysis • Nuclear Proliferation and Diversion • Occupational Health Risks in Nuclear Power • Public Reaction to Nuclear Power Siting and Disposal • Radiation, Risks and Health Impacts of

Further Reading

Bodansky, D. (1996). "Nuclear Energy: Principles, Practices, and Prospects." American Institute of Physics, Woodbury, NY.

Murray, R. L. (1994). "Understanding Radioactive Waste." 4th ed. Battelle Press, Columbus, OH.

National Research Council. (1996). "The Waste Isolation Pilot Plant: A Potential Solution for the Disposal of Transuranic Waste." National Academy Press, Washington, DC.

Wilson, P. D. (ed.). (1996). "The Nuclear Fuel Cycle: From Ore to Waste." Oxford University Press, Oxford, UK.

Obstacles to Energy Efficiency

MARILYN A. BROWN
Oak Ridge National Laboratory
Oak Ridge, Tennessee, United States

1. Background
2. The Energy Efficiency Gap
3. Obstacles to Energy Efficiency
4. The Role of Government
5. Conclusions

Glossary

efficiency gap The difference between the actual level of investment in energy efficiency and the higher level that would be cost-beneficial, from the points of view of consumers and society.

externalities The goods or services that people consume as by-products of other people's activities; such goods and services are "external" to market transactions and are therefore unpriced.

market barriers The obstacles that contribute to the slow diffusion and adoption of energy-efficient innovations.

market failures The conditions of a market that violate one or more of the neoclassical economic assumptions (e.g., rational behavior, costless transactions, and perfect information) that define an ideal market.

Homes, offices, factories, and vehicles rarely incorporate the full complement of available cost-effective, energy-efficient technologies, despite the sizable costs that inefficient designs impose on consumers and society. There is compelling evidence that vast cost-effective opportunities exist for improvements in energy efficiency, implementation of which could significantly curtail the rapid growth in energy consumption in the United States, resulting in a more competitive economy, cleaner air, lower greenhouse gas emissions, and enhanced national security. These opportunities exist because of large-scale market failures and barriers that prevent markets from operating efficiently, slowing the diffusion of innovations. Assessments of energy policies and programs suggest that public interventions can overcome many of these obstacles.

1. BACKGROUND

Examination of energy trends following the 1973/1974 oil embargo highlights the great strides in energy efficiency that have made the U.S. economy much less energy intensive today than it was in 1970. Despite this progress, an energy efficiency gap still exists. Many investments that appear to be cost-effective to the consumer are not being made in energy-efficient technologies and practices. This would not be surprising if a relatively small number of such investments were being foregone, or if only a small portion of future energy growth would be impacted by making these investments. However, numerous studies indicate the continued existence of a sizable untapped reservoir of highly cost-effective investments that could significantly cut U.S. energy use. If these investments were to occur, the result would be a more competitive economy, cleaner air, lower greenhouse gas emissions, and less dependence on foreign oil.

If an energy-efficient technology is cost-effective, why does more of it not just happen? If individuals or businesses can make money from energy efficiency, why are they all not just doing so? Why does the cost-effectiveness of an energy-efficient option not simply propel it to commercial success? These questions can be answered, in large part, by describing the range of obstacles that prevents the full exploitation of energy efficiency. Markets for energy-efficient technologies are not unique in their imperfections. Other products and services face obstacles that impede their adoption, even when their consumer economics appear to be favorable. Conditions hindering cost-effective investments in energy efficiency are particularly noteworthy because of their widespread environmental, national security, and macroeconomic repercussions.

This article begins with a series of individual technology case studies that present compelling evidence of an efficiency gap. Next, the market

 © 2004 Elsevier Inc. All rights reserved.

failures and institutional barriers that explain the existence of this gap are described. The sector differences in market failures and barriers are then characterized. This lays the groundwork for discussing potential government roles and the pros and cons of different policies and programs that could tackle barriers to energy efficiency.

2. THE ENERGY EFFICIENCY GAP

The term "efficiency gap" refers to the difference between the actual level of investment in energy efficiency and the higher level that would be cost-beneficial from the points of view of consumers and society. It is difficult to quantify the exact magnitude of this gap. One approach to characterizing its size is through modeling and scenario development. This typically involves enumerating on a technology-by-technology basis the difference between current practice and best practice, where best practice is defined as the utilization of the energy technologies with the lowest life cycle costs. Keeping in mind the natural rate of turnover of new equipment and consumer purchases, the size of the lost opportunities or the gap that exists can then be estimated. Using this technology-based accounting approach, a study by five U.S. Department of Energy laboratories concluded that removing obstacles to energy efficiency through policy interventions initiated in the year 2000 could have reduced the forecasted U.S. energy consumption in 2010 by 10%. By 2020, the reduction could grow to nearly 20%.

Modeling exercises and policy scenarios are constructive methods for bounding the potential for improved energy efficiency, but their theoretical nature may not persuade skeptics. The "natural experiment" provided by the California electricity crisis provides a basis for characterizing the efficiency gap that is grounded in actual events. Beginning in June of 2000, the State of California experienced an unprecedented electricity supply crisis. Electricity prices escalated as electricity demand exceeded supply, and rolling blackouts occurred in June 2000 and in January, March, and May of 2001. Energy efficiency and conservation played a prominent role in avoiding further problems that were forecast for the summer of 2001. In response to a vigorous energy conservation campaign, Californians averaged a 10% reduction in their electricity consumption during peak hours in the summer of 2001, compared with the previous summer. Weather data show similar temperature patterns across the two summers, thus the reduction in usage is attributable to investments in energy efficiency and altered lifestyle choices. This consumer response was prompted by a number of utility and state-run programs that promoted no-cost energy conservation actions, installation of low-cost energy efficiency measures, and investments in energy-efficient equipment, appliances, and building shell retrofits. Because a minority of households contributed the bulk of the conservation gains, the 10% reduction in consumption is probably an underestimate of the full efficiency gap.

Another way to characterize the efficiency gap is through case studies of individual technology choices. Many different case studies could be cited showing that consumers and businesses often choose not to purchase highly cost-effective energy technology. Skeptics might conclude that these technologies must carry hidden liabilities that are difficult to identify and quantify. The following case studies, however, describe efficiency gaps associated with energy-efficient options that are clearly superior to the technologies being replaced, and with no significant "hidden costs" to the consumer.

The most energy-efficient substitute for the Edison light bulb is fluorescent lighting. Fluorescents use one-quarter to one-third as much energy as incandescent bulbs, and they burn 10 to 15 times longer. A study of lighting investments using data from the Environmental Protection Agency (EPA) Green Lights Program has shown that there is a large potential for profitable energy-saving investments in fluorescent lighting that is untapped because of impediments based on organizational decision making. A key impediment is capital rationing—when a business unit is allocated a quota of capital to cover all improvements, including new product designs and cost-cutting production improvements such as energy efficiency. Another obstacle is the lack of organizational rewards for energy managers who reduce utility bills. Often the energy savings are not even measured. All fluorescents need a ballast to operate. The ballast is a device that alters the electric current flowing through the tube, which activates the gas inside, causing the tube to emit light. The most efficient ballasts today are electronic. Electronic ballasts superseded two prior technologies, the original "standard" ballast and the more efficient magnetic ballast. Efficient magnetic ballasts were commercially available as early as 1976. They were a well-tested technology, with performance characteristics equal to or better than standard ballasts by the early 1980s. By 1987, five states—including

California and New York—had prohibited the sale of standard ballasts. But the remaining three-quarters of the population chose standard ballasts over efficient ballasts by a ratio of 10:1, even though the efficient magnetic ballast typically paid back its investment in less than 2 years. The time required to establish retail distribution service networks and to gain consumer confidence were key obstacles to the adoption of the efficient magnetic ballast, as is the case with many innovations. (Since 1990, federal standards have prohibited the sale of the standard ballast.)

The purchase of home appliances illustrates additional impediments to energy efficiency. In one study by Lawrence Berkeley National Laboratory (LBNL) scientists in the 1980s, consumers were given a choice in stores throughout the United States of two refrigerators that were identical in all respects except two: energy efficiency and price. The energy-efficient model (which saved 410 kilowatt-hours (kWh) per year, more than 25% of energy usage) cost $60 more than the standard model. The energy-efficient model was highly cost-effective in almost all locations of the country. In most regions, it provided an annual return on investment of about 50%. In spite of these favorable economics, which were easily observed by the purchaser, more than half of all purchasers chose the inefficient model. The higher purchase price of the efficient model was the principal barrier to its purchase. Yet the consumer would save more money by investing in the more expensive, efficient refrigerator than by investing a similar amount in any standard investment mechanisms such as a certificate of deposit, U.S. Treasury Bill, or stock portfolio.

To enable the use of remote controls for televisions in the early 1990s, it became necessary for televisions to consume some amount of power continuously. Typical televisions with remote controls at that time used 5 to 7 W of standby power for that purpose. The Energy Star television program was able to reduce these power losses by requiring that televisions qualifying for the Energy Star label must reduce standby power to 3 W or less, a savings of roughly 50%. The resulting price increase had a payback period of 1 to 2 years for consumers. Because this savings was no more than a few dollars a year per television, there was no public outcry for manufacturers to deliver the improvement. At the same time, the aggregate savings to the nation of widespread market penetration were significant. Through the labeling program, the lack of consumer interest could be overcome. Most major manufacturers now offer such televisions, and several of them have reduced standby losses to 0.5 W.

Case studies of manufacturing facilities illustrate that barriers to energy efficiency also exist in the industrial sector. Motor systems represent the largest single end use of electricity in the American economy—23% of U.S. electricity consumption—and they present a very substantial energy efficiency potential. The results of a recent market assessment involving on-site surveys of 265 industrial facilities document that technologies offering a simple payback of 3 years or less can typically save businesses 11–18% of the energy used to drive motors. The Department of Energy (DOE) Best Practices program conducts audits, demonstrations, and technical assistance to encourage the use of proved, cost-effective technologies to improve industrial motor systems. Monitoring and validation of energy use data from these activities confirm the profitability of these investments, underscoring the large gap between current practice and economically smart investments. Limited information, expertise, and capital cause this gap.

3. OBSTACLES TO ENERGY EFFICIENCY

Numerous obstacles, including market failures and barriers, contribute to the efficiency gap (Table I). "Market failures" occur when there are flaws in the way markets operate. They are conditions of a market that violate one or more of the neoclassical economic assumptions (e.g., rational behavior, costless transactions, and perfect information) that define an ideal market. Market failures can be caused by (1) misplaced incentives, (2) distortionary fiscal and regulatory policies, (3) unpriced costs such as air

TABLE I

Market Failures and Barriers Inhibiting Energy Efficiency

Market failures (flaws in the way markets operate)	Market barriers (other factors preventing rapid adoption)
Misplaced incentives	Low priority of energy issues
Distortionary fiscal and regulatory policies	Capital market barriers
Unpriced costs	Incomplete markets for energy efficiency
Unpriced benefits	
Costly, imperfect and asymmetric information	

pollution, (4) unpriced benefits such as education, training, and technological advances, and (5) costly, imperfect, and asymmetric information. By failing to account for such market imperfections, assessments of energy policies based on neoclassical economic models often underestimate their potential benefits. It is widely argued by neoclassical economists that the existence of market failures is a prerequisite for market intervention. However, the existence of such failures is also seen as an insufficient justification for government involvement. Feasible, low-cost policies must be available that can eliminate or compensate for these market failures.

"Market barriers" refer to other obstacles that contribute to the slow diffusion and adoption of energy-efficient innovations. To the extent that it is in society's best interest to use its energy more efficiently and to reduce emissions from fossil fuel combustion, it is important to understand the full range of obstacles to energy-efficient technologies. These barriers include (1) the low priority of energy issues among consumers, (2) capital market imperfections, and (3) incomplete markets for energy-efficient features and products.

3.1 Market Failures

3.1.1 Misplaced Incentives

Misplaced incentives inhibit energy-efficient investments in each sector of the economy. This problem occurs when an "intermediary" has the authority to act on behalf of a consumer, but does not fully reflect the consumer's best interests. Examples of this failure are numerous:

- Architects, engineers, and builders select equipment, duct systems, windows, and lighting for the individuals who occupy buildings.
- Landlords purchase appliances and equipment for tenants who pay the energy bills.
- Industrial buyers choose technologies that manufacturers use in their factories.
- Specialists write product specifications for military purchases.
- Fleet managers select the vehicles to be used by drivers.
- New car buyers determine the pool of vehicles available to buyers of used cars.

The involvement of intermediaries in the purchase of energy technologies limits the ultimate consumer's role in decision making and leads to an underemphasis on life cycle costs. This is exacerbated in the construction industry, where typical fee structures for architects and engineers cause incentives to be distorted in ways that penalize efficiency. Additional first costs are typically needed to enable the installation of superior heating, ventilation, and air-conditioning systems that reduce operating costs. These additional expenditures beyond the typical "rule-of-thumb" equipment sizing used by most engineers result in a net penalty for designers of efficient systems. Even though this type of fee structure has been strongly discouraged in the United States, both the designer and procurer of design services still generally base their fee negotiation on percentage-of-cost curves.

The landlord–tenant relationship is a classic example of misplaced incentives. If a landlord buys the energy-using equipment and the tenants pay the energy bills, the landlord is not incentivized to invest in efficient equipment unless the tenants are aware of and express their self-interest. Thus, the circumstance that favors the efficient use of equipment (when the tenants pay the utility bills) leads to a disincentive for the purchase of energy-efficient equipment. The case that favors the purchase of efficient equipment (when the landlord pays the utility bills) leads to a disincentive for the tenants to use energy efficiently. About 90% of all households in multifamily buildings are renters, which makes misplaced incentives a major obstacle to energy efficiency in this segment of the housing market.

Misplaced incentives can also involve significant time lag. For instance, new car purchasers have a dominant influence on the design decisions of automakers. Yet they are not representative of the entire driving public, many of whom purchase their vehicles secondhand. In particular, compared to the average driver, new car purchasers are substantially wealthier, which skews their purchase preferences away from fuel economy and toward ride quality, power, and safety. Similar obstacles to energy efficiency exist in the secondary markets for appliances and homes.

3.1.2 Distortionary Fiscal and Regulatory Policies

Distortionary fiscal and regulatory policies can also restrain the use of efficient energy technologies. A range of these distortionary policies was recently identified in an analysis of projects aimed at installing distributed generation. Distributed generation is modular electric power located close to the energy consumer, including photovoltaics, gas turbines, and fuels cells. Regulatory barriers to these new technologies include prohibitions against their

use and state-to-state variations in environmental permitting requirements that result in significant burdens to project developers. Tariff barriers include excessive backup and standby charges as well as utility buy-back rates that do not provide credit for on-peak electricity production. An example of a distortionary fiscal policy is the tax treatment of capital versus operating costs. United States tax rules require capital costs for commercial buildings and other investments to be depreciated over many years, whereas operating costs can be fully deducted from taxable income. Because efficient technologies typically cost more than standard equipment on a first-cost basis, this tax code penalizes efficiency. Similarly, many states are uneven in their sales tax policies. For instance, most states charge sales tax on residential energy-saving devices but not on residential fuels and electricity.

Electricity pricing policies of state legislatures and regulatory commissions also prevent markets from operating efficiently and subdue incentives for energy efficiency. The price of electricity in most retail markets today is not based on time of use. It therefore does not reflect the real-time costs of electricity production, which can vary by a factor of 10 within a single day. Because most customers buy electricity as they always have (under time-invariant prices that are set months or years ahead of actual use), consumers are not responsive to the price volatility of wholesale electricity. Time-of-use pricing would encourage customers to use energy more efficiently during high-price periods. Metering, communications, and computing technologies are needed to support such dynamic pricing and voluntary-load-reduction programs. The cost of designing and installing this infrastructure represents another potential barrier to real-time pricing. Although this might be cost prohibitive for some customers, the cost of this infrastructure would be affordable for large commercial and industrial customers.

3.1.3 Unpriced Costs

A strong case can be made that fossil fuels (and other energy resources) are underpriced because market prices do not take full account of a variety of social costs associated with their use. Fossil energy using today's conversion technologies produces a variety of unpriced environmental costs (i.e., negative externalities), including air, water, and land pollution and greenhouse gas emissions. For example, more than 90% of U.S. emissions of carbon dioxide, NO_x, and SO_2 results from the combustion of fossil fuels (see Fig. 1). Other types of energy-related environmental impacts include water and ground contamination from petroleum leaks from vessels and underground storage tanks. Nuclear power production generates fission by-products that require long-term storage. Sources of renewable energy present specific environmental concerns as well, such as fish kills from hydropower, avian mortality from wind power, and particulate air pollution from biomass combustion. In addition to imposing unpriced environmental costs, oil markets also inflict unpriced national security costs. Under relatively tight market conditions, the physical concentration of oil reserves in a relatively small number of countries generates the potential for physical and price-setting supply disruptions. These market conditions impose national security costs by reducing foreign policy flexibility and complicating military strategy, especially during periods of rising oil demand and tightening world markets.

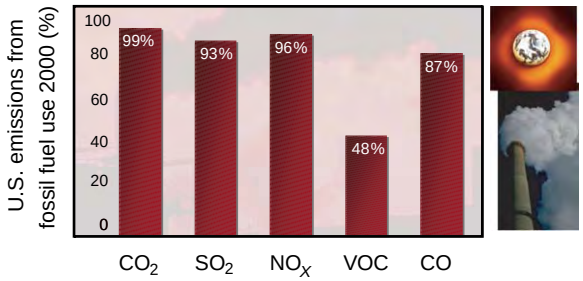


FIGURE 1 Energy-linked emissions as a percentage of total U.S. emissions. Sources: Environmental Protection Agency, *National Air Pollutant Emissions Trends Report, 1998*, EPA-454/R-00-0002-March 2000; and Energy Information Agency, *Emissions of Greenhouse Gases in the U.S., 2000*, DOE/EIA-0573(00), October 2000.

Some of these environmental costs are being addressed through regulation. Absent regulation that effectively internalizes the full costs of environmental impacts and national security, externality costs exist. When the market does not value these costs, more energy is consumed than is socially optimal. Negative externalities associated with fossil energy combustion can be "internalized" through policy interventions. Domestic carbon trading is one example of such a policy. The idea of the carbon trading system is to create fossil fuel prices that better reflect the full cost of fossil fuel consumption, causing consumers to make decisions that take into account the complete cost of the resource. These higher prices should cause consumers to use less fossil fuel. At the same time, the government-collected carbon permit revenues can be recycled to consumers.

Existing environmental control costs are embedded in some energy costs. For instance, the U.S. Environmental Protection Agency regulations enforcing the Clean Air Act and other federal legislation impose control costs on the marginal emitter of criteria pollutants such as SO_2 and NO_x. However, not all existing fossil generators incur operating cost penalties. Furthermore, there are several emissions produced by fossil fuel combustion that are not capped today. These include carbon, mercury, and smaller (2.5 μm) particulates. No costs are currently included to account for damages from these pollutants. Because energy prices do not include the full cost of environmental externalities, they understate the societal cost of fossil energy use based on today's combustion technologies. This market imperfection can be addressed through public policies and programs that bring market choices more in line with full societal costs.

3.1.4 Unpriced Benefits

Unpriced benefits also dampen the energy productivity of the economy. A public good is a good or service that has two principal characteristics. First, one person's consumption of it does not reduce the amount of it available for other people to consume. This characteristic is called "inexhaustibility." Second, once such a good is provided, it is difficult to exclude other people from consuming it, a characteristic called "nonexcludability." Because public goods are unpriced, markets tend to underproduce them.

The public goods nature of education, training, and research is an important rationale for government support. Investments by employers in creating a well-educated, highly trained workforce, for instance, are dampened because of the firm's inability to ensure that the employee will work long enough for that firm so as to repay its costs. The difficulties of selecting and installing new energy-efficient equipment compared to the simplicity of buying energy may prohibit many cost-effective investments from being realized. This is a particularly strong barrier for small and medium-sized enterprises. In many firms (especially with the current trend toward lean firms), there is often a shortage of trained technical personnel who understand and can explain the ability of energy-efficient technologies to generate a stream of cost savings that more than pay for any up-front installation premium. Government programs that pay university engineering faculty and students to conduct energy audits of industrial plants can overcome this barrier by training the next generation of energy professionals while delivering energy diagnostics and audit recommendations to plant managers.

Investing in research and development often results in benefits that cannot be fully captured by the sponsoring private entities. Although benefits might accrue to society at large, individual firms cannot realize the full economic benefits of their research and development investments. Further, companies that absorb the market risk of introducing new technologies are generally unable to reap the full benefits of their trailblazing. The payback from advances in energy-efficient technologies is not only experienced by the sponsoring company, but also flows to the public, to the company's competitors, and to other parts of the economy. The problem is especially pronounced when an industry is highly fragmented, as is the case with the construction and homebuilding industries, in which the largest 400 homebuilders serve just 40% of the residential market. Fragmentation increases the number of potential external beneficiaries of the research and development results. In highly fragmented markets, individual companies also may lack the scale of resources needed for advanced research and development investments.

The risk of innovation leakage and exploitation by competing firms puts pressure on firms to invest for quick returns. Technology innovation is typically a longer term investment fraught with risks to the investor. The result is an underinvestment in research and development from the standpoint of overall benefits to society. The problem is particularly difficult in the newly restructured electric sector, for which research and development funding has decreased dramatically. Companies will not fund the optimal societal level of basic research and development of new technologies because many of the benefits of such research will flow to their competitors and to other parts of the economy. This is true of many industries, and is one of the main rationales for government-funded long-term, precompetitive research in industries that have a vital role in the U.S. economy.

The Council of Economic Advisers estimates that the private returns from research and development are 20–30%, whereas social returns (including energy security and environmental benefits) are 50% or higher. This gap limits the extent to which the private sector can supplant a government role in maintaining nationally beneficial research and development. With the development of international markets, fragmentation is growing and industry's priorities are shifting further away from basic and

applied research and toward near-term product development and process enhancements. Over the past several decades, business spending on applied research has dropped to 15% of overall company research and development spending, whereas basic research has dropped to just 2%. In addition, corporate investments in energy research and development, in particular, are down significantly.

3.1.5 Costly, Imperfect, and Asymmetric Information

Insufficient and incorrect information can also be a major obstacle to energy efficiency. Reliable information about product price and quality allows firms to identify the least costly means of production, and gives consumers the option of selecting goods and services that best suit their needs. Yet information about energy-efficient options is often incomplete, unavailable, expensive, and difficult to obtain. With such information deficiencies, investments in energy efficiency are unlikely. Asymmetric information can also cause problems; when one party to a transaction has substantially more product information, market opportunities will tilt toward the entity with the information advantage.

It is difficult to learn about the performance and costs of energy-efficient technologies and practices because the benefits are often not directly observable. For example, residential consumers get a monthly electricity bill that provides no breakdown of individual end-uses, making it difficult to assess the benefits of efficient appliances, televisions, and other products. Appliance energy labels help, as does "Energy Star" branding, but studies have shown that many consumers do not understand them. Efficiency choices tend to be bundled with other purchases. For instance, the price paid for different levels of vehicle fuel economy is buried in base prices or in the price of complete subsystems such as engines. Further, efficiency differences are coupled with substantive differences in other critical consumer attributes such as acceleration performance, level of luxury, and vehicle handling. Reliable information on the marginal cost of fuel economy may be obtainable, but the effort required for an individual consumer to secure such information could be prohibitive.

Decision-making complexities are another source of imperfect information that can confound consumers and inhibit "rational" decision making. Even while recognizing the importance of life cycle calculations, consumers often fall back to simpler first-cost rules of thumb. Although some energy-efficient products can compete on a first-cost basis, many of them cannot. Properly trading off energy savings versus higher purchase prices involves comparing the time-discounted value of the energy savings with the present cost of the equipment—a calculation that can be difficult for purchasers to understand and compute. This is one of the reasons builders generally minimize first costs, believing (probably correctly) that the higher cost of more efficient equipment will not be capitalized into a higher resale value for the building. The complexities of decision making are one form of transaction cost. Note, however, that if consumers were extremely concerned about life cycle energy savings and were determined to base their purchasing decisions on them, product manufacturers would have a strong incentive to provide consumers with better information about energy efficiency and with clearer trade-offs. It can be argued that the lack of such information and choices is simply the consequence of consumer disinterest in using energy efficiently—the first of several market barriers.

3.2 Market Barriers

3.2.1 Low Priority on Energy Issues

Energy efficiency is not a major concern for most consumers because energy costs are not high relative to the cost of many other goods and services. In addition, the negative externalities associated with the U.S. energy system are not well understood by the public. The result is that the public places a low priority on energy issues and energy efficiency opportunities. In turn, this reduces producer interest in providing energy-efficient products. In most cases, energy is a small part of the cost of owning and operating a building, a factory, or a car. Of course, there are exceptions. For low-income families, the cost of utilities to heat, cool, and provide other energy services in their homes can be a very significant part of their income—averaging 15% compared to 4% for the typical U.S. citizen. For energy-intensive industries such as aluminum and steel, energy can represent 10–25% of their production costs. Many companies in these more energy-intensive firms have decided to incorporate energy management as a key corporate strategy. These companies have made impressive gains in efficiency. A means of sharing successful strategies and best practices can help improve overall industry performance and extend benefits to smaller, less profitable firms.

Because energy costs are typically small on an individual basis, it is easy (and rational) for

consumers to ignore them in the face of information-gathering and transaction costs. However, the potential energy emissions savings can be important when summed across all consumers. This is one reason why government agencies such as the EPA and DOE work directly with manufacturers to improve product efficiency. A little work to influence the source of mass-produced products can pay off in significant efficiency improvements and emissions reductions that rapidly propagate through the economy due to falling production costs as market shares increase.

3.2.2 Capital Market Barriers

Capital market barriers can inhibit efficiency purchases. Although, in theory, firms might be expected to borrow capital any time a profitable investment opportunity presents, in practice, firms often ration capital—that is, they impose internal limits on capital investment. The result is that mandatory investments (e.g., required by environmental or health regulations) and those that are most central to the firms' product line often are made first.

Different energy producers and consumers have varying access to financial capital, and at different rates of interest. In general, energy suppliers can obtain capital at lower interest rates than can energy consumers, resulting in an "interest rate gap." Differences in these borrowing rates may reflect differences in the knowledge base of lenders about the likely performance of investments as well as the financial risk of the potential borrower. At one extreme, electric and gas utilities are able to borrow money at low interest rates. At the other extreme, low-income households may have essentially no ability to borrow funds, resulting in an essentially infinite discount rate for investments in energy efficiency. The broader market for energy efficiency (including residential, commercial, and industrial consumers) faces interest rates available for efficiency purchases that are also much higher than the utility cost of capital. In addition, it has been shown that firms typically establish internal hurdle rates for energy efficiency investments that are higher than the cost of capital to the firm. Information gaps, institutional barriers, short time horizons, and nonseparability of energy equipment all contribute to this gap, and each is amenable to policy interventions that could move the rates down toward auto loan, mortgage, and opportunity costs. Energy prices, as a component of the profitability of an investment, are also subject to large fluctuations. The uncertainty about future energy prices, especially in the short term, seems to be an important barrier. Such uncertainties often lead to higher perceived risks, and therefore to more stringent investment criteria and a higher hurdle rate. An important reason for high hurdle rates is capital availability. Capital rationing is often used within firms as an allocation means for investments, leading to hurdle rates that are much higher than the cost of capital, especially for small projects.

3.2.3 Incomplete Markets for Energy Efficiency

Incomplete markets for energy efficiency are often a serious obstacle. Energy efficiency is generally purchased as an attribute of a product intended to provide some other service. Fuel economy in automobiles, for example, is one of a large number of features that come in a package for each make and model. If higher fuel economy were treated as an optional item, available at a higher price, then consumers would have a choice of efficiency levels. But such a separate option does not presently exist. Similar circumstances in the commercial buildings market constrain investor choices. The complexity of design, construction, and operation of commercial buildings makes it difficult to characterize the extent that a particular building is energy efficient. As a result of this host of market failures and barriers, the discount rate that consumers appear to use in making many energy efficiency decisions is higher than the interest rate at which consumers could borrow money. This discount rate gap has been widely observed in the literature and is reflected in some key energy models, such as the Energy Information Administration's National Energy Modeling System.

3.3 Sectoral Differences in Market Failures and Barriers

Each end-use sector functions differently in the U.S. energy marketplace. One of the reasons for this variation is the distinct supply chain and market structure for delivering new technologies and products in different sectors. Residential and commercial building technologies are shaped by thousands of building contractors and architectural and engineering firms, whereas the automotive industry is dominated by a few manufacturers. As a result, the principal causes of energy inefficiencies in manufacturing and transportation are not the same as the causes of inefficiencies in homes and office buildings, although there are some similarities. For example, in the manufacturing sector, investing in cost-effective

energy efficiency measures (which cut operating costs and therefore increase profits) is hampered by a common preference to invest resources to increase output and market share as a preferred route to expanding profits. Because energy uses are so diverse in the industrial sector, "one-size-fits-all" products rarely play a large role. Achieving widespread use of technological breakthroughs can require dissemination of technical documentation and technical assistance to help firms tailor the technology to their particular production processes. In the building sector, information gaps prevent the energy-efficient features of buildings from being capitalized into real estate prices. This is partly due to the lack of widely adopted building energy-rating systems. These information gaps are less characteristic of the transportation sector, in which fuel economy is well understood in terms of miles per gallon.

The end-use sectors also differ in terms of their ability to respond to changing energy prices. This is partly due to the varying longevity of the equipment that is used. For example, cars, lighting, and air conditioners turn over more quickly than do large industrial boilers, which can operate for the better part of a century. Stock turnover must be taken into account when assessing the potential market impacts of policies and programs. There are also differences in fuel flexibility. The U.S. transportation system today is relatively fuel inflexible, being primarily dependent on petroleum, whereas energy use in buildings and industry is more fuel flexible.

The vast differences in the research and development capability of the sectors also influence their ability to respond quickly to changing energy prices and market signals. The private sector as a whole spends more than $180 billion per year on research and development, dwarfing the government expenditure on all nondefense technology research and development. Of the private-sector research and development expenditure, the automobile manufacturers stand out—Ford alone spends more than $8 billion per year in research and development. Next comes the rest of the industrial sector. Here manufacturers account for a majority of research and development expenditures. In the buildings sector, the construction industry has virtually no indigenous research and development. The Council on Competitiveness estimates that the construction industry spends less than 0.2% of its sales on research and development, far less than the 3.5% that other industries spend on average.

Finally, each of the sectors is distinct in terms of the primary societal benefits from improved energy efficiencies. Fuel economy in transportation is essential to improving air quality and protecting against oil price volatility. Energy productivity in industry is essential to economic competitiveness and pollution prevention. Energy efficiency in buildings makes housing more affordable on a life cycle basis; it is also important in ensuring electric grid stability and in reducing SO_2, NO_x, and particulate matter, because most of the energy consumed in buildings is fossil-generated electricity. As a result of these differences, public policies and programs need to be customized specifically to meet the needs of each sector.

4. THE ROLE OF GOVERNMENT

The existence of market failures and barriers that inhibit socially optimal levels of investment in energy efficiency is the primary reason for considering public policy interventions. In many instances, feasible, low-cost policies can be implemented that either eliminate or compensate for market imperfections and barriers, enabling markets to operate more efficiently to the benefit of society. In other instances, policies may not be feasible; they may not fully eliminate the targeted barrier or imperfection; or they may do so at costs that exceed the benefits.

To foster energy efficiency, transaction costs need to be reduced. Several of the problems discussed herein, particularly those related to information, can be viewed as transaction costs associated with energy decision making. Examples include the costs of gathering and processing information, making decisions, and designing and enforcing contracts relating to the purchase and installation of energy-using technology. These costs are real, in the sense that they must be borne by the consumer and should be included in the cost of the energy efficiency measure. A key question is whether there are institutional interventions that can reduce these costs for individual consumers. For example, the time and effort required to find a refrigerator with the maximum cost-effective level of energy efficiency could be significant.

Information programs (e.g., fuel economy Web sites) and technical assistance (e.g., industrial energy assessments) can help make up for incomplete information by reducing the consumer's cost of acquiring and using needed information. They can also simplify decision making and can help consumers focus on energy issues that may seem small to an individual consumer but can be large from a national perspective.

The direct installation of home weatherization measures addresses the lack of access of low-income households to capital. Programs that support financing through energy services companies and utilities also address this barrier. More indirectly, but just as important, technology demonstrations provide financial markets with evidence of performance in the field, which is critical to reducing the cost of capital. For instance, electric utility companies in many regions have demonstrated the value of advanced lighting technologies through various incentive programs that have subsequently led to the widespread acceptance of these products and the increased availability of financing through mechanisms such as energy-saving performance contracts.

The public goods nature of research and development can be addressed through direct government funding. Great potential exists for public/private research and development partnerships to produce scientific breakthroughs and incremental technology enhancements that will produce new and improved products for the marketplace. If public policies reorient even a tiny fraction of the private sector's research expenditures and capability to address the nation's energy-related challenges, it could have an enormous impact. One way to reorient private-sector investments is through industry/government research and development alliances that involve joint technology roadmapping, collaborative development of advanced energy-efficient technologies, and cost sharing.

5. CONCLUSIONS

Homes, offices, factories, and vehicles rarely incorporate the full complement of cost-effective, energy-efficient technologies, despite the sizable costs that inefficient designs impose on consumers and nations. The evidence of this efficiency gap is compelling, and the reasons for it are numerous. Models of the U.S. energy system, statistical analyses, and case studies underscore the widespread existence of this gap and the array of market obstacles that cause it. By improving our understanding of these obstacles, it may be possible to fine-tune existing policies, invent entirely new types of policy interventions, and better explain their benefits to the public.

SEE ALSO THE FOLLOWING ARTICLES

Commercial Sector and Energy Use • Conservation Measures for Energy, History of • Consumption, Energy, and the Environment • Discount Rates and Energy Efficiency Gap • Economics of Energy Efficiency • Energy Efficiency and Climate Change • Energy Efficiency, Taxonomic Overview • Industrial Energy Efficiency • Labels and Standards for Energy • Market Failures in Energy Markets • Material Efficiency and Energy Use • Vehicles and Their Powerplants: Energy Use and Efficiency

Further Reading

Alderfer, R. B., Eldridge, M. M., and Starrs, T. J. (2000). "Making Connections: Case Studies of Interconnection Barriers and Their Impact on Distributed Power Project." NREL/SR-200-28053 (May 2000). National Renewable Energy Laboratory, Golden, Colorado.

Brown, M. A., Levine, M. D., Short, W., and Koomey, J. G. (2001). Scenarios for a clean energy future. *Energy Policy* **29**(14), 1179–1196.

Council of Economic Advisers (CEA). (1995). "Supporting Research and Development to Promote Economic Growth: The Federal Government's Role." CEA, Washington, D.C.

DeCanio, S. J. (1998). The efficiency paradox: Bureaucratic and organizational barriers to profitable energy-saving investments. *Energy Policy* **26**(5), 441–454.

Environmental Protection Agency (EPA). (1999). "Driving investment in Energy Efficiency: Energy Star and Other Voluntary Programs." EPA, Washington, D.C.

Geller, H., and McGaraghan, S. (1996). "Successful Government–Industry Partnership: The U.S. Department of Energy's Role in Advancing Energy-Efficient Technologies." American Council for an Energy Efficient Economy, Washington, D.C.

Greening, L. A., Sanstad, A. H., and McMahon, J. E. (1997). Effects of appliance standards on product price and attributes: An hedonic pricing model. *J. Regul. Econ.* 11(2), March

Hassett, K. A., and Metcalf, G. E. (1993). Energy conservation investment, do consumers discount the future correctly? *Energy Policy* **21**, 710–716.

Hirst, E., and Brown, M. A. (1990). Closing the efficiency gap: Barriers to the efficient use of energy. *Resourc. Conserv. Recycling* **3**, 267–281.

Horowitz, M. J., Lewis, K., and Coyle, A. (2000). Green Lights and Blue Sky: Market Transformation Revealed," 2000 ACEEE Summer Study on Energy Efficiency in Buildings, Vol. 6. American Council for an Energy-Efficient Economy, Washington, D.C.

Jaffe, A. B., and Stavins, R. N. (1994). The Energy-efficiency gap. *Energy Policy* **22**(10), 804–810.

Koomey, J., Sanstad, A. H., and Shown, L. J. (1996). Energy-efficiency lighting: Market data, market imperfections, and policy success. *Contemp. Econ. Policy* 14(3), July.

Levine, M. D., Koomey, J. G., McMahon, J. E., Sanstad, A., and Hirst, E. (1995). Energy efficiency policy and market failures. *Annu. Rev. Energy Environ.* **20**, 535–555.

Lovins, A. (1992). "Energy-Efficient Buildings Institutional Barriers and Opportunities." Strategic Issues Paper (December). E-Source. Boulder, Colorado.

Lutzenhiser, L., Gossard, M. H., and Bender, S. (2002). "Crisis in Paradise: Understanding the Household Conservation Response to California's 2001 Energy Crisis." Proceedings of the Summer Study on Energy Efficiency in Buildings, pp. 8.153–8.166. American Council for an Energy-Efficient Economy, Washington, D.C.

Martin, M., Tonn, B., Schmoyer, R., Overly, J., Schexnayder, S., and Johnson, D. (1999). "Industrial Assessment Center Program Impact Evaluation." ORNL/CON-473 (December). Oak Ridge National Laboratory, Oak Ridge, Tennessee.

Meier, A., and Whittier, J. (1983). Consumer discount rates implied by purchases of energy-efficient refrigerators. *Energy* 8(12).

Parfomak, P., and Lave, L. (1997). How many kilowatts are in a negawatt? Verifying ex post estimates of Utility conservation impacts at the regional scale. *Energy J*. 17(14).

Romm, J. J. (1999). "Cool Companies: How the Best Companies Boost Profits and Productivity by Cutting Greenhouse Gas Emissions" Island press, Covelo, California.

Ross, M. H. (1990). Capital budgeting practices of twelve large manufacturers. *In* "Advances in Business Financial Management" (P. Cooley, Ed.), pp. 157–170. Dryden Press, Chicago.

Sanstad, A. H., Blumstein, C., and Stoft, S. E. (1995). How high are options values in energy efficiency investments? *Energy Policy* **23**, 739–743.

U.S. Department of Energy (DOE), Office of Policy and International Affairs. (1996). "Policies and Measures for Reducing Energy Related Greenhouse Gas Emissions." DOE/PO–0047 (July). DOE, Washington, D.C.

U.S. Department of Energy (DOE), Office of Energy Efficiency and Renewable Energy. (2000). "Clean Energy Partnerships: A Decade of Success." DOE/EE–0213 (March). DOE, Washington, D.C.

Xenergy, Inc. (1998). "United States Industrial Electric Motor Systems Market Opportunity Assessment." Xenergy, Inc., Burlington, Massachusetts. Available on the Internet at http://www.motor.doe.gov/.

Occupational Health Risks in Crude Oil and Natural Gas Extraction

JOEL M. HAIGHT
Pennsylvania State University
University Park, Pennsylvania, United States

1. Specific Health Risks
2. General Health Risks during Specific Operations
3. Summary

Glossary

blowout prevention (BOP) equipment Usually a hydraulically controlled rubber sealing device that is pressured to close around the drill string to seal the well bore and contain the contents and pressure of a well blowout.

cellar A pit dug during the preparation of the location to be used for drilling an oil or gas well. It is created to accommodate the stack of equipment making up the well head, including the blowout prevention equipment.

corrosion inhibitor Commercially available chemical mixtures designed to protect steel or other alloys against the effects of corrosion.

diverter A sealing device and collection of valves and pipes positioned near the top of the wellhead so that in the event of a sharp, rapid pressure increase ("kick"), workers can shut in the well and open the diverter valves to direct flow and pressure away from the wellhead.

heavy metals Any of a series of dense metallic elements naturally occurring in the earth's crust, such as chromium, cadmium, lead, nickel, and iron.

hydrogen sulfide A toxic, colorless, heavier-than-air gas formed as a by-product of organic decomposition of sulfur compounds.

Raynaud's syndrome A phenomenon in which there is an abnormal constriction of the blood vessels of the fingers characterized by numbness and white color and which is exacerbated by cold conditions.

risk A measure of the combination of the severity and likelihood of the occurrence of undesirable consequences.

well program A written description of the plans involved in drilling an oil or gas well. It contains information about the size and type of rig to be used, the depth and type of well to be drilled, the expected formation pressures, drilling fluid design, casing design for the well, etc.

The extraction and production of crude oil and natural gas are complex operations that involve the potential for exposure to many health risks. This article discusses each of these individual health risks as well as the operational phases of drilling and extraction of these valuable natural resources. These risks include potential exposure to hydrogen sulfide, naturally occurring radioactive material, hydrocarbons, corrosion inhibitors, solvents, and ergonomic, biological, and physical hazards. This article also covers the drilling and completion of a well and surface production operations. This discussion includes treatment of the health risks associated with each phase of drilling and extraction operations, as well as the techniques used by industry to reduce the risk of exposure to these materials.

1. SPECIFIC HEALTH RISKS

1.1 Hydrogen Sulfide

One of the most common and toxic health risks associated with crude oil and natural gas extraction is exposure to hydrogen sulfide (H_2S). The oil and gas extraction industry has taken great pains over the years to minimize overexposure to H_2S. The industry in general has done much to design, build, and install equipment that safely contains the hydrogen sulfide; it has provided extensive education, training, and protective equipment for its employees and has installed alarm and detection systems in appropriate

 © 2004 Elsevier Inc. All rights reserved.

locations. The effort appears to have accomplished its objective as the Bureau of Labor Statistics (BLS) reports that during the years 1993–2001 there was only one year (1999) in which there were more than 50 cases of respiratory illness reported (of which only a fraction would have been from hydrogen sulfide exposure). The BLS reports similar data for the mining industry.

H_2S is a colorless, heavier-than-air gas that possesses an odor of rotten eggs at low concentrations, but deadens the sense of smell at higher concentrations. It is highly corrosive to certain metals, is generally considered more toxic than carbon monoxide, and has a toxicity similar to that of hydrogen cyanide. When it is burned, it produces sulfur dioxide (SO_2), which is also considered toxic to humans. Once H_2S is released into the atmosphere, it remains there for approximately 18 h. Its decomposition and (or) transport in the atmosphere are thought to contribute to the formation of SO_2 and sulfuric acid, which contribute to the formation of acid rain.

The main exposure route for workers in the oil and gas industry is through inhalation. At lower concentrations, H_2S acts as a local irritant, producing irritation of the eyes and mucous membranes. At higher concentrations, it can be fatal ($>$600–800 ppm). According to a 1991 report by Amdur and coworkers, at these concentrations, H_2S stops respiration at the cellular level (cells do not obtain the oxygen they need). H_2S forms a complex bond with the iron in the mitochondrial cytochromes and stops aerobic metabolism. The transfer of oxygen from the blood to the cells cannot occur. This toxic action is similar to that of cyanide. H_2S toxicity is thought to be associated with carboxyhemoglobin or methemoglobinemia. A notable trait of H_2S is that it also produces olfactory fatigue by deadening the olfactory nerve. Thus, the strong rotten egg-like odor should not be used as a measure of the presence or absence of H_2S.

The effects from short-term, high-dose exposures are well documented and have, in some cases, led to death. Long-term, low-dose exposures are of concern as well. In 1981, Ludvich reported that symptoms such as respiratory irritation, sore throat, cough, shortness of breath, accumulation of fluid in the lungs, and even memory changes are possible. One positive aspect about exposure to H_2S is that it does not accumulate in the body. With chronic exposure at very low doses at or slightly above the odor threshold of 10 ppb, effects such as headaches, nausea, appetite loss, irritability, and fatigue are possible.

H_2S is a by-product of the organic decomposition of sulfur compounds and can be found in many other settings in addition to the oil industry, such as mining, sewage treatment, and the paper industry. H_2S is found in solution in some crude oils and as a component of natural gas. It begins to flash out of solution as the oil and (or) natural gas are brought to the surface. In operations where the gas is intentionally separated from the oil as it is processed, the hydrogen sulfide will most often go with the gas fraction. This is accelerated if there is any additional heat input during extraction. Potential exposure to hydrogen sulfide can occur during drilling operations through exposure to recycled drilling muds, during exposure to the water fraction of sour crudes (sour is the name given to crude oils with a relatively higher sulfur and H_2S content), or during a blowout. Once the oil has been extracted and is being gathered, transported, and treated, there are many potential opportunities for exposure. Workers can be exposed when gauging tanks containing sour crude, when performing wellhead maintenance such as replacing valve packing, when pulling pumping rods, during entry into confined spaces such a valve pits and cellars, and during work in areas where leaks have occurred. Since H_2S is heavier than air, once a leak has occurred it tends to hug the ground and collect in low-lying areas, such as a cellar at a wellhead.

According to a 1981 study by Ludvich, since H_2S can react with some metals, corrosion and embrittlement are possible and leaks can result. H_2S causes stress cracking in certain metals used in vessels that contain or transport the crude oils containing it. It does so because the metals are subjected to high stress in corrosive environments. H_2S will also dissolve in water to form a weak acid that can cause pitting if it is also in the presence of oxygen or carbon dioxide. The most significant form of attack is the ability for H_2S to create hydrogen embrittlement. This is also known as sulfide stress cracking. In this case, brittle failure can occur rapidly. The significance of this phenomenon to human health is that H_2S creates an environment where it can cause significant leaks in containment vessels. Releasing H_2S to the work environment increases the risk of exposure. There are many alloys that have been developed and successfully used to contain hydrogen sulfide (e.g., Monel 400 and K-500, Incoloy 800 and 825, Inconel 600, 625, 718, and X-750, and Hastelloy C, G, and X). These and other appropriate alloys should be considered when designing a system to store, transport, or process sour crude oil or natural gas containing hydrogen sulfide.

Additional protection schemes used in minimizing the risk of exposure include the extensive use of

hydrogen sulfide detection equipment positioned around potential leak or collection areas. This detection equipment is connected to dedicated alarm systems that alert workers in the area that there has been a release or leak. On receiving this alarm, workers are able to don appropriate respiratory equipment and respond to stop the leak or safely evacuate the area before exposure occurs. The use of supplied or self-contained breathing-air respiratory protection is common in areas where exposure to hydrogen sulfide is possible. This can be anywhere in the oil field; however, the protection is generally offered near wellheads, in motor control centers, around oil treatment plants, and in any low-lying area that might require entry by workers.

1.2 Naturally Occurring Radioactive Material

The risk of exposure to naturally occurring radioactive material (NORM) must be considered in discussions of the oil and natural gas extraction industry. NORM is found to be associated with barium sulfate ($BaSO_4$, also known as barite). This is sometimes a component found in drilling mud. In a 1994 report, Wilson explains that as drilling mud or fluids are circulated down hole, sometimes barite is lost in the rock formation as drilling progresses. It can later be brought to the surface in the production stream as the pressure profile in the well or in the formation changes over time. $BaSO_4$ will also absorb radium as it precipitates out on tubing, casing, and production equipment when the well is flowing. The barium sulfate is deposited on the piping and tubing along the production string. This way, it is exposed to any material contained in the production stream as it passes up the production string. Occasionally, the production stream will contain radium, which is then absorbed onto the $BaSO_4$ deposits. This can result in radioactive deposits that require protection for workers having to handle equipment or properly dispose of it on dismantling at the end of its useful life. NORM production also seems to increase as brine production increases in the well. In this case, the NORM may accumulate in the drill string.

Natural gas flows from the formation and to the surface at high flow rates under turbulent conditions. Wilson further explains that radon gas is often mixed with the natural gas as it flows into the production system. Radon's half-life is only ~ 3.8 days and it then decomposes to radioactive lead, Pb^{210}. The route the natural gas takes as it travels up the production tubing and into the transportation and processing system is sometimes found to have a thin layer of Pb^{210} on the surfaces of the piping and tubing. This thin accumulation layer of ~ 0.004 in. can also affect the disposal of this equipment at the end of its useful life.

In the petroleum industry, it is estimated that the typical radium 226 concentration in oil sludge and scale is 155 pCi/g. It can be as high as 40,000 pCi/g. It is also estimated that the petroleum industry generates as much as 360,000 m^3 of NORM waste each year. Proper disposal is critical to reducing the risk of exposure.

According to Wilson, in oil and gas extraction, radiation exposure levels could be expected to be relatively low and to occur over a long period (career). The long-term effects of radiation exposure are difficult to determine. However, the delayed effects of radiation exposure are generally thought to be the precursor of cancer and genetic defects. In general, exposure to radiation at high enough levels is thought to damage or change cells in the body. Cell changes or damage can occur at almost any place in the body. Examples include the blood system (anemia due to loss of red cells and lower resistance from loss of white cells), the lymphatic system (signs of hemorrhaging in lymph nodes and weight loss in the spleen), the digestive tract (nausea and vomiting from damage to the lining of the organs in the digestive canal), and the lungs (lung cancer).

To reduce exposures, many considerations must be incorporated into well programs and drilling operations. The best approach to exposure reduction is through time, distance, and shielding. Exposure dose follows the inverse square law, so basically, doubling the distance to the source cuts the dose in half. Other interventions applied include the use of shielding, warning labels, and personal protective equipment, such as gloves, coveralls, and boots. Full-face air-purifying respirators with appropriate NORM cartridges are also worn. An active detection program is important to the protection of the workers. As radioactive material is detected, the workers should be removed from the exposure site as soon as possible to minimize the time of exposure. NORM-contaminated equipment as well as sludge, scale, and other deposits must be disposed of properly in designated disposal facilities.

1.3 Hydrocarbons

The hydrocarbons associated with oil and natural gas extraction are generally thought to be of lower health risk than many of the contaminants found in or

associated with oil and gas production. Methane, ethane, propane, and butane are found in a vapor state under atmospheric conditions. This makes them, in theory, the easiest to be exposed to through inhalation. According to the American Conference of Governmental Industrial Hygienists' 2001 Threshold Limit Values for Chemical Substances and Physical Agents Guide, methane and ethane are considered nothing more than simple asphyxiants and propane and butane have threshold limit values–time-weighted average (TLV–TWA) of 2500 and 800 ppm respectively. Thus, exposure to these fractions is considered a low health risk.

In general, as the carbon number of the crude oil fractions increases, the health risk also increases. The TLV–TWA for pentane, heptane, and octane ranges from 600 ppm for pentane (five carbons) to 300 ppm for octane (eight carbons). This is not true for the normal hexane isomer (*n*-hexane, six carbons). It is a higher health risk, with a TLV–TWA of 50 ppm. Exposure to these fractions would occur on contact exposure to liquid crude oil during spills or leaks; however, a portion of each of these components would likely volatilize depending on the atmospheric conditions. This would increase the potential for inhalation exposure to occur. It is important to point out that crude oil exposure would have to be carefully measured to determine whether a person was overexposed due to the combined exposure to several components or fractions at one time. Crude oil (all fractions still bound up in solution) does not necessarily behave in the body in same way as the individual fractions would.

Some of the potential health effects from hydrocarbon exposure include central nervous system depression, vomiting, and mucous membrane irritation. Since the TLV–TWAs are relatively high for most of the hydrocarbon fractions, the incidence of overexposure during oil and gas extraction is less common than exposure to hydrocarbon distillates and hydrocarbon contaminants present during oil and gas refining.

Although the crude oil is not refined, some of the health effects associated with these crude oil fractions and "reformates" can still generate health effects. Benzene is a carcinogenic component of petroleum and can represent a significant percentage of a barrel of crude oil. Benzene can attack the blood system if exposures are high enough. The inhalation route is considered the most critical and most common since benzene has a relatively high vapor pressure under ambient conditions. It can be absorbed through the skin, so excessive or long-term skin contact exposure should be avoided. Fortunately, in crude oil, benzene is not present in its concentrated form during extraction operations. Therefore, this naturally reduces exposure potential.

Exposure control programs involve the application of leak minimization principles, working systems to manage confined space entry and hot work, detection equipment, exposure monitoring, and the use of respiratory protection as appropriate.

1.4 Heavy Metals (Chromium, Nickel, Iron, Lead, etc.)

Heavy metals are naturally occurring elements within the earth's crust. Those that would be of the greatest health risk during oil and gas extraction include lead, chromium, cadmium, nickel, and iron. There are many more. These metals are redistributed throughout the environment by geologic and biologic cycles. Heavy metals become potential exposure agents with an associated health risk to oil and gas workers during the extraction process. In 1991, Devereaux reported that the metals reside in the rock formations below the ground's surface. On drilling or extraction, heavy metals become entrained in or included with drilling fluids, oil, water, brine, and rock cuttings that are returned to the surface. Exposure to these metals in solid form would not be significant at this point, as it is not expected that workers would regularly ingest these materials. Exposure could occur if materials containing the metals became airborne either from an aerosolized mist in a spill or pressurized leak or from the creation of a dust cloud of dried material. This would make inhalation of the material possible. Although the risk of exposure to heavy metals during most extraction activities would generally be considered low, burning the oil, such as is sometimes done after a flow test at a production well, would increase the risk of exposure. Metal fumes would be created as the heat of the fire evaporates the metals and they then quickly condense to respirable-size particles. These particulates would become entrained in the hydrocarbon smoke and potentially be inhaled by workers in the area.

Metals are neither created nor destroyed in the body. Some are considered essential to the diet in trace amounts. When the "trace" amount is exceeded, not only are there immediate health risks to be concerned about, but some metals accumulate in the body unless intervention is implemented, such as administration of chelating agents. Chronic effects from exposure to heavy metals are usually of greater

concern than acute exposure. This concern is based on the accumulation of metals in the body over time. In the oil and gas industry, acute exposures are relatively easy to avoid, since in most circumstances the metals are in solid form. The potential health effects from chronic or acute exposure to heavy metals are many and varied, depending on the metal involved. Some of the effects include chronic pulmonary disease and chronic renal disease from cadmium; peripheral neuropathy and anemia induced by lead; dermatitis caused by nickel; and renal failure and hepatic cirrhosis due to iron (acute exposure). Chromium is a carcinogen and each of the metals listed above exhibits carcinogenic effects.

During drilling operations, the risk of exposure to heavy metals in drilling fluids, brine, and rock cuttings is higher than during extraction and production operations. This is true because the drilling system is much less closed than the production system. However, during drilling, the metals are in solid form, with the highest risk of exposure coming from exposure to metal-containing mist or dust as noted above. The exposure risk can be minimized here by wearing proper protective equipment, such as coveralls, gloves, and eye protection. It may also be appropriate to wear respiratory protection when exposure to metal-containing dust is possible. Proper maintenance of the drilling and circulating equipment will also help minimize the risk of exposure by reducing the likelihood for drilling fluid, brine, water, and oil leaks.

For production operations workers, the highest risk of exposure comes when burning metal-containing crude oil following a flow test. The risk of exposure to heavy metals can be minimized in this case by finding alternate means to dispose of the flow test oil or using appropriate respiratory protection such as an organic vapor air-purifying respirator. Alternate disposal for the oil may include discharge into a clean holding tank and redirecting the oil back into the production stream for processing.

1.5 Corrosion Inhibitors

The equipment used in oil and gas extraction is exposed to many different types of fluids and other materials. Some of these materials are not compatible with the steels and other alloys used in manufacturing oil and gas handling equipment. Corrosion can result and subsequent leakage of the production stream is then possible. The equipment is designed using materials that can withstand attack from many different contaminants and process streams under many different conditions. To supplement this, many companies use corrosion inhibitors to help protect their equipment. These inhibitors are used as part of a regular, on-going treatment program that complements the equipment's preventive maintenance program. Many different corrosion inhibitors are safe and present little health risk. However, in years past, two materials that were used for this purpose contained either hydrazine or arsenic. These materials are no longer widely used, if at all, but residual material is or can still be left on equipment internals. Hydrazine and arsenic have been identified as carcinogens. The inorganic compound hydrazine has been shown to induce pulmonary tumors in sensitive strains of mice at high dose levels and arsenic has been associated with the development of lung and skin cancer and lymphomas in humans. This association has been shown for the trivalent, inorganic form of arsenic in drinking water.

Since these materials are no longer widely used in corrosion inhibitors, it is likely that the risk of exposure is becoming lower as time passes. However, it may still be safe to suggest that if there is evidence of buildup of any kind on the internals of drilling, extraction, or processing equipment, one would do well to ascertain whether the buildup contained arsenic or hydrazine. The literature suggests that the highest exposures to these materials occur through ingestion of drinking water. If effort is made to avoid ingestion (ensuring the drinking water system does not contain harmful corrosion inhibitors) and skin contact (protective gloves, coveralls, etc.), the risk of exposure is reduced. Dismantled equipment, left out in the open air, may contain arsenic- or hydrazine-contaminated scale or buildup. If this material dries out, it could become airborne if disturbed. Small particles could be picked up in the wind and dispersed into work areas. Removal of the equipment from work areas or removal of any potentially contaminated buildup on equipment would help to reduce the risk of exposure.

1.6 Organic Solvents (Diesel Fuel, Kerosene, etc.)

Nearly everyone is exposed to solvents in some form or fashion over his or her lifetime. In the crude oil and natural gas extraction industry, exposure to solvents can potentially be high. Since some of the contaminants and components of crudes, lubricants, drilling fluids, brine, and other materials can collect and build up on equipment, solvent use is often required to

remove deposits and clean down-hole and surface handling equipment. Many solvents have been made to be environmentally friendly and health-friendly. However, some solvents used in the oil and gas industry are not. Solvents such as gasoline, diesel fuel, and kerosene are often used. Exposure to solvents can occur through inhalation and skin contact. Concentration and time of exposure are two important considerations, as is the chronic or acute nature of the exposure. Since there are so many different solvents in use, with varying health effects, it is difficult to characterize the health effects of all solvents. In general, solvents can dissolve protective oils in the skin, drying it out to the point where cracking occurs. Cracking allows solvent to be more readily absorbed into the system on continued exposure. Inhalation exposures to solvents (gasoline, diesel, kerosene, etc.) at high enough concentrations can result in dizziness, nausea, headache, and other central nervous system effects.

Solvents are used in the oil and gas extraction industry to remove buildup of hydrocarbons and other contaminants from equipment (valve bodies, pump impellers, shafts, etc.). This is done by first placing equipment in vats, tubs, or other containers of the solvent and then, through rinsing, brushing, or other active means, the buildup or contaminant is removed. There are a number of types of solvent application devices and methods, such as high-pressure sprays as well as soaking.

The type of application is critical to assessing and protecting against exposure. For any application that aerosolizes the solvent, ventilation or respiratory protection or both should be used. For liquid contact, gloves and eye, face, and body protection (aprons, coveralls, etc.) should be worn at all times when solvents are being handled. When and where possible, it is effective to reduce exposure time and opportunities to a minimum. Many solvents are flammable (especially diesel and kerosene), so appropriate fire precautions should be taken. The ventilation system should be rated for flammable materials and the potential for static spark generation should be minimized. Proper disposal of spent solvents is important to keeping exposures low.

1.7 Ergonomic, Physical, and Biological Health Risks (Noise, Heavy Lifting, Mosquitoes, Vermin, Heat, and Cold)

The oil and gas drilling and extraction industry has, like most industries, become more automated over the past 100 years. Even though there is not as much lifting and handling of heavy equipment and supplies as there once was, there are still ergonomic and physical health risks associated with drilling and extraction operations. With the need for compressors, pumps, turbines, steam generators, drilling motors, and rotary tables, etc., noise levels around drilling and extraction operations can be high. Noise levels around compressors, turbines, and oil-fired steam generators can be as high as 95 to 100 dBA (decibels measured on an A-weighted scale). Noise exposure around equipment during drilling and completion activities can be of greater concern than during production operations because workers are exposed for longer and more continuous periods. It is a 24 h operation and workers sometimes remain in high-noise-level areas for their full shift. Oil and gas production operations do involve the use of compressors, turbines, and steam generators. Even though the sound levels are similar to those around drilling operations, exposure periods are usually shorter and more intermittent. Many companies make an effort to use only equipment that is designed and built to generate lower noise levels, but this does not always completely reduce the noise levels to acceptable levels. The goal of many companies is to reduce noise levels and exposures to the point where they do not have to implement the Occupational Safety and Health Act's Hearing Conservation Standard. For time-weighted average exposures above 85 dBA, implementation of this standard is mandatory in the United States. Hearing protection, training, and audiometric testing can supplement sound level reduction interventions.

Although the drilling and extraction operations have become automated over the years, there is still a need for the transport, installation, application, and handling of heavy, bulky, and awkward-to-move equipment and supplies. Equipment and supplies, such as large wrenches, drums of chemicals, chains, rigging equipment, and valves, can be heavy or require awkward body positions to lift or handle. Opening and closing large valves can present an ergonomic risk. There is a risk of back and shoulder strains in this type of operation. At times, work on a drilling rig requires the use of supplied-air respiratory equipment. Although this respiratory equipment is designed and built to be worn and carried, when it is combined with the carrying and use of other tools over a long shift, the risk of muscle strain or, at least, fatigue is present. Although automation and lift-assist devices have been successfully implemented and used in the oil and gas extraction industry, many

tasks still require manual handling and brute strength (e.g., manipulating and attaching drill pipe from the monkey board in the drilling derrick). The search for improved means and use of automated and lift devices continues. In the meantime, education and training can go a long way to helping workers understand their physical limitations and how to accomplish their job without straining or stressing their muscles and bones.

The climate in oil and gas extraction areas can be extreme and present certain health risks. The hot humid climates of Africa, South America, and the South Pacific region present the risk of heat stress. Without proper acclimatization, hydration, and work/rest regimens, workers risk the onslaught of heat-related illness, including the potential for heat stroke. Care must be taken when tasks are to be performed under high heat and humidity conditions. The cold in areas such as Siberia, the north slope of Alaska, and western Kazakhstan also presents the risk of frostbite and freezing. In these climates, it is also possible to have difficulty operating equipment that requires manual dexterity. If gloves make it difficult to operate small equipment, such as drills, torquing guns, and wrenches, workers will be more likely to remove gloves meant to protect them from the cold. In extreme cold, blood flow to the extremities is decreased and the vibration of the equipment then contributes to the development of Raynaud's syndrome (an abnormal constriction in the blood vessels of the fingers when exposed to cold). This phenomenon results in a loss of feeling in the fingers. Adequate, layered clothing helps protect workers who must work in the cold environment. Providing heated shelters, as well as hand- and foot-warming devices to allow workers temporary and regular relief from the cold, can reduce the risk of frostbite, Raynaud's syndrome, and freezing.

Oil and gas drilling and extraction operations are found all over the world. Some of these locations present a health risk due to the occupation or infestation by insects and other vermin. The risk of contracting mosquito-borne malaria is present in some locations, such as sub-Saharan Africa. According to local government environmental agency representatives, the fleas on the rats and other rodents of western Kazakhstan reportedly carry various diseases. Mosquito control programs involving spraying and minimization of standing water can help reduce the risk of exposure to malaria. Workers can also reduce the risk of contracting malaria through prophylactic drug regimens. Efforts to reduce the encroachment of disease-carrying rodents into living areas of drilling and production operations involve the use of poisons, traps, and continued food removal and habitat destruction.

2. GENERAL HEALTH RISKS DURING SPECIFIC OPERATIONS

2.1 Drilling Operations—Land Based

Once it has been determined that a particular formation has the necessary characteristics for oil or gas production, and the necessary permits have been approved, a well program is developed. This program contains various technical information about how the well is to be drilled, such as how the rig will be positioned, depth of the well, drilling fluid design, maintenance requirements, geological information, casing design, and kick tolerances. It also contains information about the hazards that are expected to be encountered. This is critical to the overall program as some of the hazards have the potential to result in loss of life.

2.1.1 Location Preparation and Driving the Conductor

The first step in drilling an oil or gas well is the preparation of the location for the drilling rig. During this phase, the site is cleared, the boundary is marked, and access roads are installed. Depending on the geographical area, consideration must be given to hazards such as mosquito-borne malaria, unexploded war ordinance, and hot and cold ambient conditions. During this phase, a number of pits are dug. These excavations include a rig pit to accommodate blowout prevention equipment. This pit is called the "cellar" and it is lined with concrete. Health-related concerns here include the potential for collapse of the sides of the pit as workers are preparing it for concrete. In the event of a leak in production equipment, the pit provides a location for the potential collection of the heavier-than-air, toxic hydrogen sulfide. After completion of drilling and commencement of production, work inside the cellar requires diligent application of a confined-space-entry permit program including air testing and use of appropriate respiratory protection.

A waste pit is dug for disposal of drill cuttings after they have been separated from the drilling fluids. In desert locations, a concrete-lined water pit is constructed and a water well is drilled to provide water for the drilling operations. Extensive quantities of water are required for mixing drilling mud and

cement. At this point, conductor pipe is driven into the ground with a pile driver. This provides the structure for the beginning of drilling operations. The process generates high-impact noise levels. This is common to pile driving operations in general. Since the exposure limit for impact noise is 140 dBA, there is probably less concern for overexposure to noise at this point in the operation. The health hazards associated with this phase of the operation are minimal; however, there is a health risk concern when oil begins to flow. The pits are low-lying areas that present a place for heavier-than-air gases and vapors such as H_2S to collect.

2.1.2 Moving the Rig on Location and Attaching the Diverter

The rig substructure is constructed of steel beams in a large frame, which supports the rig floor. On land, the rig floor is generally ~15–20 ft above the ground level. The derrick rises up from the substructure above the rig floor (sometimes as much as 100 ft or more). The drill floor supports the rotary table, which turns the drill string. The structure is secured with large steel pins. Although much of this work is done with cranes, winches, and other lift devices, there is much manual lifting and manipulating of heavy equipment during this phase. The risk of musculoskeletal sprains and strains is high during this operation. Once the traveling block and block line are in place, the rig is ready to start operating.

During drilling operations, there is always a danger of an unexpected "kick." The formation releases pressure and flow as formation restrictions are cleared. If the pressure exceeds the gravity head and friction pressure of the drilling fluids, the pressure and flow can escape up the well. This potentially exposes the workers on the rig floor and the environment to high-pressure, potentially high-temperature fluids, hydrocarbons, and hydrogen sulfide. One of the tools drillers use to minimize the risk of this exposure is to install a flow diverter above the conductor pipe. When the rig is ready to start, a diverter is attached to the conductor pipe. This diverter contains a large rubber seal forced under hydraulic pressure in and around the drill string to seal it. Two large pipes are installed under this seal. They lead away from the rig in opposite directions. If, during drilling, there is a kick in the well below the conductor pipe, the flow is directed away from rig when the diverter is closed and the valve on the pipe leading from the rig is opened. The flow is usually directed to a containment system, where the release of energy and flow can be controlled and contained.

After the conductor pipe has been driven into the ground with the pile driver, the inside of the pipe will be full of rocks, dirt, and other rubble. This is removed by drilling. During this process, the drill bit is rotated as it is lowered into the rubble. The rocks are removed as drilling mud is pumped through the drill string and the mud rises back up the annular space between the drill string and the conductor pipe, carrying the rocks with it. As the conductor pipe area is drilled, there is a potential to penetrate a shallow pocket of gas. This gas pocket can contain hydrocarbon or hydrogen sulfide and if it is penetrated, the pressure, hydrocarbons, and other contaminants can be released to the surface where the workers can be exposed. It is easier to recognize and control this release if the size of the drill hole is kept small. A smaller drill hole would help to keep the initial flow small. This would allow a few extra seconds for the drillers to close the diverter and to begin rapidly introducing drilling fluid down the drill string to control the pressure and flow. This gas release can be of great enough intensity that removal of large amounts of rock also introduces the potential danger of projectiles.

2.1.3 Spudding and Drilling the Well

When the drill bit reaches the bottom of the conductor pipe and begins the drilling process, this phase is called spudding the well. As the well bore is drilled during the spudding phase, rock cuttings are generated and must be removed from the hole. This is done by pumping drilling fluid down through the hollow drill string, out the holes in the bottom of the drill bit, and back up the annular space between the hole and the drill pipe. This phase continues until the borehole is deep enough for the surface casing.

When spudding is complete, drilling of the first hole section and installation of the intermediate casing sections begin. The drilling process is the same in that a rotating bit at the bottom of the drill string removes rock cuttings from the bottom of the hole while drilling fluid is pumped through the drill string. In this section of the well, the surface casing is installed and cemented in. The blowout prevention (BOP) equipment will also be installed and tested. This equipment helps to minimize the potential for exposure to the unexpected sudden release of pressure and well fluids during drilling operations. As the automatically controlled seal material is tightened around the drilling and production equipment to

form a seal, the blowout is contained in the well and kept away from workers.

When the BOP equipment is installed and tested, the casing is drilled out and the strength of the formation must be tested. Testing the strength of the formation is necessary to help ensure that it will support the well without collapsing. This test is done by first drilling out a cement valve called a float shoe at the bottom of the surface casing and then removing all cuttings and approximately 15 ft of new rock below the bottom of the surface casing. Then the drill bit is removed to a point above the BOP equipment. The BOP equipment is closed and the hole is incrementally pressurized by adding drilling fluid a little at a time until the pressure created by the addition of fluid to a closed system begins to reduce as more fluid is added. This tells the drillers that some "leak-off" to the formation is occurring at that pressure. The addition of more fluid at this point could fracture the formation and weaken it. This pressure is determined to be the formation strength. Although there is always a risk of release of pressure and fluids from the formation, the risk of exposure of workers to hydrogen sulfide, hydrocarbons, etc., at this point, is low due to the containment provided by the BOP equipment.

When the BOP equipment is operational and the formation has been tested, the program is ready for drilling of the intermediate hole sections. This is where the diameter of the well bore is reduced and the depth increased significantly. With the intermediate casing sections installed, the production hole section is next. This is the smallest diameter section and is generally the bottom of the well. This section coincides with the presence of both liquid and gaseous hydrocarbons. This is also the point where blowouts, the release of hydrogen sulfide and other hydrocarbons, and exposure to drilling fluids and formation contaminants become more likely. Hydrocarbon and other gases sometimes dissolve in the drilling fluids and then escape at the surface in circulation tanks. This introduces the risk of exposure to the workers as well as the risk of fire should flammable hydrocarbon gases find an ignition source. Throughout this phase, casing and liner pipe are being installed, cemented in place, and sealed.

2.2 Drilling Operations—Offshore

Offshore drilling often involves drilling several wells from one platform. This requires directional drilling to access several locations in the formation without the expense and risks of a production platform at every well location. The health risks for offshore drilling are, for the most part, the same as they are for onshore drilling. The contaminants present in the underground formations will be the same. Hydrogen sulfide exposure will be of concern, as will hydrocarbon, heavy metals, and NORM exposure. The differences would be related to the isolated and confined location. If there are contaminant leaks or releases, there is no immediate escape. Exposure to a cold environment in locations such as the North Sea and the Maritime Provinces in Canada would be an additional risk brought on by exposure to the cold-water environment.

2.3 Surface Production Operations

Surface production operations include production of the oil and natural gas at the wellhead, collection of produced oil and gas at collection manifolds, and transport to the production treatment facility. Control stations at well sites are enclosed and often climate-controlled spaces where the instrumentation that controls the well operation is housed. In the event of a leak, the control stations can become a collection point for hydrogen sulfide or other hydrocarbons. These spaces are often fitted with hydrogen sulfide, oxygen, and flammable material detectors with alarms to indicate when concentrations of these contaminants may be an exposure concern for operators. As mentioned earlier, cellars or pits dug to support the drilling operation may be left at a wellhead. These also serve as a collection point for hydrogen sulfide and other heavier-than-air contaminants should a leak develop.

As production at a particular well progresses over time, there is a need to test its production capability. Flow tests are conducted to test this capacity. During this test, oil is allowed to flow into a pit or similar reservoir. The flow is measured over time to gauge this flow capacity and ultimately to ascertain the well's production rate to help determine production life. This oil is sometimes burned as a means of disposing of it. This practice potentially exposes workers to hydrogen sulfide, sulfur dioxide (converted from hydrogen sulfide in the combustion process), carbon smoke particulates, and unburned hydrocarbons. Under normal conditions, the oil is transported through pipelines to collection manifolds until it results in one stream entering the treatment facility.

As oil flows through the pipelines and collection manifolds, two situations that can have an impact on human health occur. In some climates, and with some

crude oils, hydrates can precipitate out of the oil and collect in the pipelines. This can occur under low-temperature conditions, or when there are changes in pressure, or at higher salt concentrations. These hydrates can plug the lines and trap pressure. If they free themselves as thermal cycling occurs throughout a day, they can be forced rapidly through the pipe by the pressure behind them. This, in essence, is an out of control projectile that can rupture pipelines. If a line ruptures, workers in the area could be struck by the hydrate material or piping fragment, injured by the moving pipeline, or exposed to the hydrocarbons or hydrogen sulfide that would be released in a rupture. Methanol is sometimes injected into the line to help dissolve the hydrate buildup before it becomes a plug in the line. Although done under procedure-driven control, this practice has the potential to expose the workers to alcohol or fire.

Lines also become plugged over time with sediment and other contaminants and occasionally must be cleared. The method of such clearing is called "pigging" the line. A barrel-shaped device of various designs and sizes, with various scraping mechanisms positioned around the device's circumference (called a pig), is pressured through the line to remove any blockage or restrictions in the line. The pig is slightly smaller in diameter than the internal diameter of the line. It is loaded at a closed pig-launching station, the loading door is closed, the production line is opened, and the pig is pressured through the line with natural gas, nitrogen, or some other nonreactive pressurizing material. The sludge or blockage ahead of the pig is either routed back into the production stream or removed and disposed of at the pig-receiving station. At the receiving station, the pig is removed after the production side is closed and the pressure is bled off through a vent. This operation clears the line and improves flow efficiency.

The health risks associated with this operation include potential exposure to high pressure, hydrogen sulfide, or hydrocarbons. If an attempt is made to open a pig trap while it is still pressured, the door would swing open rapidly; serious injury is likely. Because the system is opened during launching and receiving, the possibility of exposure to hydrogen sulfide and other hydrocarbons is increased.

The goal of the production facility is to separate the production stream into oil, gas, and water. This is achieved through a series of separators, which make use of pressure differentials to separate oil and gas and make use of viscosity and gravity differences to separate oil and water. Sometimes the gas contains hydrogen sulfide, which must be further processed. Natural gas is sometimes distilled to separate methane, ethane, propane, and sometimes butane into separate streams. Prior to this step, depending on the hydrogen sulfide content in the production stream, which can be as high as 15–25% by volume, the gas must pass through an absorption process. This step strips the hydrogen sulfide from the gas by placing it in contact with an amine or similar absorptive chemical with an affinity for hydrogen sulfide. The amine or other absorptive chemical must also be treated to remove the H_2S so it can be reused. The chemical is passed through a "regeneration" process, such as distillation, to remove the H_2S. The gas containing the hydrogen sulfide and other acid gases, collectively referred to "sour gas," are processed in sulfur treatment units utilizing heat and reaction with catalysts to convert the hydrogen sulfide to elemental sulfur. This elemental sulfur can then be sold as a product. At several stages in the sulfur treatment process, small quantities of hydrogen sulfide and sulfur dioxide can still be present. Exposure is still a possibility; therefore, the use of appropriate respiratory protection during sulfur blocking or loading may be necessary depending on the concentration.

The crude oil fraction, which is usually considered the main product, is sometimes passed through desalting equipment, undergoes solids centrifuging, and undergoes distillation to remove salts, unwanted produced solids, and any residual gases. All of these steps are not necessary for all crude oils. Some crude oils need only undergo static retention time in a holding tank to allow the water to settle out and it is then ready for sale. This would be true for crude oils containing less gas and hydrogen sulfide. Although the health risks at this stage are relatively low compared to stages where the hydrogen sulfide concentration is high, an appropriate level of precaution is necessary. Proper respiratory protection should be nearby and calibrated and operational leak detection equipment should be functioning.

3. SUMMARY

The extraction and production of crude oil and natural gas are complex operations that involve the potential for exposure to many health risks. Each of these individual health risks has been well studied over the years. Sound-exposure reduction programs have been established and successfully implemented in many oil and gas operations around the world. The risk of exposure to hydrogen sulfide, NORM,

hydrocarbons, solvents, heavy metals, and other hazards has been successfully managed in the oil and gas industry. These risks have been managed by properly designing, installing, and using equipment that contains these contaminants. Oil and gas companies have reduced exposures by providing appropriate detection and alarm equipment that helps to shut down equipment that is malfunctioning or leaking or by alerting workers of a leak or other operational problem. Additional management of these risks has been achieved by implementing effective permit to work programs and by providing appropriate protective equipment and educating workers so that they are better able to recognize and avoid the risk of exposure to the contaminants found during crude oil and natural gas extraction.

The crude oil and natural gas industry has done well to minimize the overall health risks associated with drilling and extraction operations. Even though the severity of exposure to contaminants such as hydrogen sulfide or sulfur dioxide is quite high, the likelihood of exposure has been greatly reduced through the diligent efforts of this industry.

SEE ALSO THE FOLLOWING ARTICLES

Crude Oil Releases to the Environment: Natural Fate and Remediation Options • Crude Oil Spills, Environmental Impact of • Natural Gas Processing and Products • Natural Gas Transportation and Storage • Occupational Health Risks in Nuclear Power • Radiation, Risks and Health Impacts of

Further Reading

Amdur, M. O., Doull, J., Klaassen, C. D. (Eds.). (1991). "Casarett and Doull's Toxicology: The Basic Science of Poisons, 4th ed. Pergamon, New York.

American Conference of Governmental Hygienists. (2001). "Threshold Limit Values for Chemical Substances and Physical Agents—Biological Exposure Indices." ACGIH Worldwide, Cincinnati, OH.

Arnold, K., and Stewart, M. (1998). "Surface Production Operations, Design of Oil-Handling Systems and Facilities." 2nd ed. Gulf Publishing, Houston, TX.

Bureau of Labor Statistics, United States Government, Web site available at http://www.bls.gov.

Devereux, S. (1999). "Drilling Technology in Non-technical Language." PennWell, Tulsa, OK.

Ludvich, D. (1981). "The Hydrogen Sulfide Technical Manual." Safety Technology and Oilfield Protectors, Lafayette, LA.

Mandavia, S. (2001). "Toxicity, Hydrogen Sulfide" (D. C. Lee, J. T. VanDeVoort, J. Benitez, J. Halamka, and R. J. Roberge, Eds.). eMedicine Web site available at http://www.emedicine.com.

McGavran, P. (2001). "Literature Review of the Effects Associated with the Inhalation of Hydrogen." Idaho Department of Environmental Quality, Boise, ID.

Perrin, D. (1999). "Oil and Gas Field Development Techniques, Well Completion and Servicing." Editions Technip. Paris and Institut Français du Petrole, Rueil-Malmaison, France.

Plog, B. A. (Ed.). (1988). "Fundamentals of Industrial Hygiene," 3rd ed. National Safety Council, Chicago, IL.

Williams, P. L., James, R. C., Roberts, S. M. (Eds.). (2000). "Principles of Toxicology Environmental and Industrial Applications," 2nd ed. Wiley, New York.

Wilson, W. F. (1994). "N. O. R. M. A Guide to Naturally Occurring Radioactive Material." PennWell, Tulsa, OK.

Occupational Health Risks in Nuclear Power

DAVID B. RICHARDSON
University of North Carolina, Chapel Hill
Chapel Hill, North Carolina, United States

1. **Ionizing Radiation**
2. **Radiation Exposure and Radiation Dose**
3. **Characteristics of Radiation Exposures in the Nuclear Power Industry**
4. **Biological Effects of Ionizing Radiation**
5. **Health Risk Following Occupational Exposure to Ionizing Radiation**
6. **Radiation Hazards in the Context of the Nuclear Fuel Cycle**

Glossary

absorbed dose A measure of the joules of radiation energy per kilogram of matter.
alpha particle A positively charged nuclear particle consisting of two protons and two neutrons emitted during radioactive decay.
beta particle A high-speed electron emitted during radioactive decay.
effective dose A measure calculated by multiplying equivalent dose by a tissue-specific weighting factor that is intended to take into account differences in sensitivity of different organs and tissues of the human body to the stochastic effects of ionizing.
equivalent dose A measure calculated by multiplying absorbed dose by a radiation-specific weighting factor that is intended to take into account differences in the density of ionization produced by different types of ionizing radiation.
gamma ray High-energy photons of electromagnetic radiation.
ionizing radiation Photons of high-energy electromagnetic radiation and particle forms of radiation that have sufficient energy to produce ions by removing electrons from atoms or molecules.

This discussion focuses on the health risks faced by workers in the nuclear power industry who operate and maintain nuclear reactors used for electrical power generation. These workers may encounter a range of hazards on the job, including exposures to solvents, welding fumes, and asbestos, as well as occupational risks for traumatic injury and electric shock. In addition, many workers in the nuclear power industry are exposed to radiological hazards arising from external sources of ionizing radiation and from internal deposition (via inhalation, ingestion, or absorption) of radionuclides. Understanding of the effects of these occupational exposures comes from a number of disciplines, including biology, toxicology, and epidemiology. Biology and toxicology provide evidence derived from laboratory research about the effects of ionizing radiation exposure on molecules, cells, and experimental animals. Epidemiology, in contrast, provides evidence about radiation health effects derived from studies of disease in human populations. The radiological hazards associated with nuclear power generation are notably different from the occupational hazards associated with other methods for power generation and consequently will be the focus of this review.

1. IONIZING RADIATION

The term ionizing radiation refers to photons of high-energy electromagnetic radiation as well as corpuscular or particle forms of radiation. Occupational exposure to ionizing radiation occurs in the nuclear power industry as a result of radioactive decay of unstable isotopes of elements. Certain isotopes of elements, such as uranium and plutonium, can emit alpha particles (two protons and two neutrons) during radioactive decay. Alpha emitters are primarily found among the elements that are heavier than

 © 2004 Elsevier Inc. All rights reserved.

lead, although some isotopes of elements lighter than lead decay by alpha emission. Radioactive decay in most elements that are lighter than lead occurs via beta decay in which an electron, called a beta particle, is ejected from the nucleus. Workers employed at heavy water nuclear power reactors, for example, may have substantial potential for exposure to beta radiation from tritium (a radioactive isotope of hydrogen). Gamma radiation often accompanies alpha and beta particle emission. The emission of neutrons occurs as the result of nuclear reactions (for example, fission of uranium-235 nuclei).

As ionizing radiation is transmitted through human tissue, energy is transferred to the tissue. This can lead to ionization of the atoms or molecules in the surrounding tissue. Alpha particles, which are relatively heavy and have a double positive charge, readily transfer energy to human tissue, resulting in the slowing of the alpha particle along its path. The energy of the alpha particles will almost entirely be deposited in the outer (dead) layers of skin if a person is externally exposed to alpha radiation. In contrast, if a radionuclide that emits alpha particles is deposited within the human body (for example, through inhalation, ingestion, or injection), damage will occur to the living cells of organs and tissues. Beta particles, which are electrons traveling at high speeds, can penetrate to a greater depth in human tissue than alpha particles. The distance that a beta particle travels through human tissue depends on the particle's energy and can range from several millimeters (for beta particles of medium energy) up to a depth of 2 cm in human tissue (for high-energy beta particles). Neutrons, which are particles without an electrical charge, can also travel short distances through human tissue. Neutrons produce ionization indirectly, by capture (for low-energy neutrons) or collision with surrounding material (for higher-energy neutrons). Gamma rays are high-energy photons of electromagnetic radiation and can completely penetrate the human body, depositing energy and causing ionization of atoms and molecules along the photons' path.

2. *RADIATION EXPOSURE AND RADIATION DOSE*

Radiation dose may be expressed in terms of absorbed dose, equivalent dose, and effective dose. An absorbed dose is the amount of radiation energy absorbed per unit mass of an organ or tissue. In 1976 the gray was adopted as the International System of Units (SI) unit of absorbed dose; one gray is equal to one joule of radiation energy per kilogram of matter. Absorbed dose is also sometimes reported in a unit known as the rad, which is an acronym for radiation absorbed dose. One rad is equal to 1/100th of one gray.

A measure of equivalent dose to an organ or tissue is a quantity that takes into account differences in ionization density for different types of radiation (e.g., alpha particles, beta particles, and gamma rays). Equivalent dose is calculated by weighting the absorbed dose to an organ or tissue by a radiation weighting factor that reflects the relative biological effectiveness of the photon or particle that produced the ionization within the tissue (it is assumed that the biological effect associated with a photon or particle is related to the ionization density). Values for this radiation weighting factor are assumed to be independent of the organ or tissue that is exposed as well as the biological end point under consideration. For outcomes that are assessed at the cellular or molecular level, such as DNA damage, it has been observed that a given absorbed dose of ionization radiation produces a greater amount of damage when the dose is due to alpha radiation than photons of electromagnetic radiation. For human health outcomes, such as cancer, however, there is uncertainty about the relative biological effectiveness of different forms of ionizing radiation because of the limitation of available empirical evidence about the magnitude of radiation-cancer associations for absorbed doses of different types of ionizing radiation. The International Commission on Radiological Protection currently recommends a radiation weighting factor of 1 for photons and electrons and a radiation weighting factor of 20 for alpha particles. While absorbed dose is a physical quantity, equivalent dose estimates may change as recommendations about weighting factors are revised. Equivalent dose is expressed in the unit sievert (Sv). An earlier unit used to express equivalent dose was the rem, which is an acronym for radiation equivalent dose; one rem is equal to 1/100th of one Sv.

A measure of effective dose is intended to express the biological effect associated with irradiation, taking into account differences in radiosensitivity of the body's various organs and tissues. Effective dose is calculated by weighting the equivalent dose in a tissue or organ by a tissue weighting factor that reflects the contribution of the tissue to the detriment of health when the body is uniformly irradiated. Doses to tissues and organs such as the lung are assigned a greater weight than doses to tissues and organs that are assumed to be relatively insensitive to adverse

effects of ionizing radiation. Effective dose is expressed in the unit Sv, and, like equivalent dose estimates, effective dose estimates may change as recommendations about weighting factors are revised.

3. CHARACTERISTICS OF RADIATION EXPOSURES IN THE NUCLEAR POWER INDUSTRY

Workers in commercial nuclear power plants are primarily exposed externally to gamma rays from activation products, such as Co-60, as well as from fission products, such as Zr-95 and Cs-137, when fuel failures occur. On average, workers in the commercial nuclear power industry accrue effective doses of 2-3 mSv per year. However, the average dose reported for nuclear power plant workers varies by reactor type as well as by job activity. For commercial boiling water reactors in the United States, the reported average annual measurable effective dose per monitored work has fallen between 1973 and 2001 from approximately 8.5 mSv to approximately 1.7 mSv; for pressurized water reactors over the same period, the average annual measurable effective dose per work has fallen from 10.0 mSv to 1.7 mSv. Maintenance work, particularly during outages (when refueling, maintenance, inspections, design modifications, and other activities take place), tends to be associated with relatively high accrued doses. It is estimated that most occupational radiation exposure are accrued during these special maintenance activities. Furthermore, the energy and geometry of radiation fields typically differs for exposures accrued during outage work when compared to routine operations. Doses tend to be accrued at higher dose rates and in an anterior/posterior geometry during outage work. In contrast, lower dose rate exposures accrued in a rotational geometry tend to be accrued during routine operations. The characteristics, in terms of energy and geometry, of radiation fields may be important for critical evaluation of reported occupational doses. For workers in areas where there is potential for occupational exposure to external sources of ionizing radiation, doses are typically monitored by a dosimeter that is worn on the chest. The response of a radiation dosimeter is influenced by the energy and geometry of the radiation field, and doses may be missed if the dose is exclusively delivered to an extremity of the body.

Other characteristics of the work environment can also affect radiation doses received on the job; these include the design of the facility, the rate of fuel failure, and radiation protection practices (particularly during repair work). In many countries, including the United States, contract workers are hired to perform short-term maintenance tasks and to provide additional support during outages. Record keeping and retrospective evaluation of accrued doses for contract workers is often more difficult than for plant and utility personnel because contractors may work at more than one facility per monitoring year. Workers employed at multiple facilities in a year are often referred to as transient workers. Analyses of transient worker data in the U.S. nuclear power industry in 2001 indicate that their average measurable effective dose per worker (3.7 mSv) tend to be higher than the average for all workers in the industry. In 2001, transient workers' doses accounted for 53% of the industry's collective dose. A small proportion of nuclear industry workers are exposed to neutrons on the job, although neutron doses account for only a small proportion of the total collective dose accrued by nuclear power plant workers.

Internal deposition of radionuclides occurs among some workers employed in the commercial nuclear power industry. In the United States in 2001, for example, the licensed operators of commercial nuclear power reactors reported occupational intake by ingestion of radionuclides including CO-60, RU-106, PU-238, AM-241, and CM-243. Reported occupational exposure by inhalation at power reactors included CO-58, CO-60, AM-241, CM-242, H-3, and CU-64. At heavy water reactors, workers may accrue significant internal doses from tritium, a beta emitter produced by neutron activation of the heavy water (deuterium). On average, internal exposure contributes about 30% of the annual collective occupational radiation effective dose among workers employed at heavy water reactors.

The International Commission on Radiological Protection recommends that occupational doses be limited to 20 mSv per year. Currently U.S. regulations require that employers limit the total effective dose equivalent from internal and external sources to 50 mSv per year. Few workers in the commercial nuclear power industry accrue doses in excess of this annual regulatory limit, although relatively high annual doses may be received by workers at nuclear power facilities in situations of accidents and remedial activities.

While average annual radiation dose estimates for workers in the nuclear power industry are low, if employment in the industry continues for several decades some workers may accrue cumulative doses of a hundred mSv or more. In a study of workers

employed at a U.S. nuclear power facility that started operation in 1975, over an average of 5.4 years of follow-up, the average cumulative external radiation dose accrued by plant and utility workers was 2 mSv and the average dose accrued by contract workers was 30 mSv. In contrast, long-term workers (those employed in the industry for 15 years or more) accrued an average cumulative dose of 111 mSv.

4. BIOLOGICAL EFFECTS OF IONIZING RADIATION

As charged particles, such as alpha and beta particles and uncharged neutrons and gamma rays, pass through living tissue energy may be transferred to the tissue. Although usually involving only a small fraction of all cells in the body, this interaction can produce harmful health effects. It is believed that the primary mechanism by which biological damage occurs is via the creation of ionized atoms and molecules that become chemically reactive. This can occur directly via ionization of a critical molecule, such as DNA, or indirectly via ionization of nearby molecules, such as water.

4.1 Deterministic and Stochastic Effects

At relatively high doses, exposure to ionizing radiation produces adverse consequences such as nausea, radiation burns, cataracts, decreased fertility, spontaneous bleeding, and, at very high doses, death. An acute equivalent dose of 150 mSv or more to the testes is likely to produce temporary sterility; at high doses, permanent sterility may occur. An acute equivalent dose of 500 mSv or above to the lens of the eye will lead to detectable opacities; such a dose to the bone marrow will lead to depression of red blood cell production. Cell killing is crucial to the development of most of these effects. These are referred to as deterministic effects; the incidence and severity of the deterministic effects observed in an individual exposed to ionizing radiation tends to increase with increasing dose. The rate at which the radiation dose is accrued is also an important factor in the production of deterministic effects; delivery of the same magnitude of radiation dose over a longer period of time reduces the likelihood of producing these adverse effects.

At lower radiation doses, the primary concern is that ionizing radiation increases the probability of certain stochastic effects, most notably cancer and genetic effects. Unlike deterministic effects of ionizing radiation, the severity of these stochastic effects is not believed to be related to the radiation dose; rather, the radiation dose is believed to increase the probability that one of these stochastic effects will occur in an exposed population. Current regulatory guidelines are premised on the assumption that any dose of ionizing radiation will increase the risk of stochastic effects.

4.2 Genetic and Somatic Effects

The damage to chromosomal DNA produced by exposure to ionizing radiation may lead to adverse genetic effects as well as somatic effects. Genetic effects are those that are caused by mutations of any germ cell (sperm or oocytes), which led to permanent changes in genetic material that are subsequently inherited by offspring of the exposed individual. Since ionizing radiation may damage chromosomal material, preconceptual exposure to ionizing radiation may lead to mutations in genetic material that will be transmitted to the exposed person's children. Genetic effects may result in decreased fertility, birth defects, or a chromosomal defect impacting offspring and future generations. Such effects have been documented in laboratory experiments involving irradiation of nonhuman animals; however, heritable genetic effects of ionizing radiation exposure have thus far not been documented in human populations such as the (F1) cohort of A-bomb survivors.

Human exposure to ionizing radiation may also lead to adverse somatic effects. Somatic effects are those mutation events that occur in the cells of tissue and organ systems other than the germ cells and that led to health outcomes such as cancer. Although a full understanding of the molecular basis of cancer has not been developed, point mutations, which are single base nucleotide changes in DNA, are believed to be involved in the malignant transformation of cells as part of a multistep process of carcinogenesis. Since ionizing radiation can damage the genetic material of any cell in the body, it has been suggested that radiation-induced cancer can occur in most, if not all, sites or tissues of the body although the radiosensitivity of tissues may vary.

5. HEALTH RISK FOLLOWING OCCUPATIONAL EXPOSURE TO IONIZING RADIATION

The primary quantitative basis for regulations that are used for radiation protection efforts in the

nuclear power industry are risk estimates derived from the life span study (LSS) of atomic bomb survivors in the cities of Hiroshima and Nagasaki. When the atomic bombs were dropped in 1945, the estimated population of Hiroshima was about 330,000 and the estimated population of Nagasaki was about 250,000. It is not reliably known how many people survived these attacks; however, beginning in 1950 during U.S. military occupation of Japan, a systematic survey of atomic bomb survivors was undertaken. A nationwide survey of A-bomb survivors who were still alive in October 1950 was conducted as a supplement to Japanese decennial census. The survey indicated that 159,000 people who had been exposed in Hiroshima and 125,000 people who had been exposed in Nagasaki were alive at the time of the census. Subsequently, individuals were assigned radiation dose estimates based on information collected by questionnaire on survivors' location and position, as well as consideration about shielding and other factors influencing radiation dose. Recent analyses, using DS86 dose estimates, have included 75,991 A-bomb survivors. Among that group of survivors, 34,272 survivors had estimated doses equal to 0 Gy.

These data have been the basis for estimation of the long-term cancer risk following radiation exposure at a wide range of ages; risk estimates for noncancer causes of death have also been reported. Results are often reported as the estimated excess relative risk of disease or death per sievert radiation dose (ERR_{1Sv}). Estimates of excess relative risk per unit dose are derived using a model in which it is assumed that the ratio of the radiation-induced cancer risk to the baseline cancer risk (the risk for those unexposed to radiation) is constant over time. An $ERR_{1Sv} = 0.40$, for example, indicates that a group of people exposed to 1 Sv is estimated to have a 40% higher cancer mortality rate than a comparable group that was not exposed to radiation. In workers' compensation cases, a related measure is sometimes used, which is the probability of causation (or assigned share); this is the probability that a cancer that has occurred was caused by radiation and is equal to $ERR/(1 + ERR)$.

Based on follow-up of atomic bomb survivors through 1990, the estimated increase in all solid cancer mortality was 45% per Sv ($ERR_{1Sv} = 0.45$). The excess relative risk for the incidence of all solid tumors (as opposed to cancer mortality) diagnosed between 1958 and 1987 among members of the LSS was $ERR_{1Sv} = 0.63$. The larger estimate when evaluating cancer incidence than when evaluating cancer mortality has been attributed to greater diagnostic accuracy of the incidence data and the lack of full representation of radiosensitive but relatively nonfatal cancers, such as breast and thyroid, in the mortality data. Analyses of leukemia incidence reported strong evidence of radiation-induced risks for leukemia, with ERR at 1 Sv for leukemia subtypes (ALL, AML and CML, respectively) were 9.1, 3.3, and 6.2. Lung ($ERR_{1Sv} = 0.95$), breast ($ERR_{1Sv} = 1.59$), and thyroid ($ERR_{1Sv} = 1.15$) were also sites with relatively large estimates of association.

There are limitations to the study A-bomb survivors, however, which led to uncertainty about the true magnitude of hazard posed by occupational radiation exposures. One source of uncertainty relates to the appropriateness of using information derived from studies of the effects of high dose-rate acute exposures received by the atomic bomb survivors to derive risk estimates for workers who are typically exposed to relatively low-dose rate chronic exposures. This requires extrapolation from high-dose effects to low-dose effects (and from the settings of an acute exposure to settings of protracted or chronic exposures). Applying risk estimates derived from the LSS also requires the assumption that the effects of ionizing radiation on the population of Japanese survivors exposed in 1945 apply to contemporary populations of workers. There are also uncertainties about the validity of extrapolating over time and between countries that have different background cancer incidence and mortality rates. In addition, questions have been raised about the validity of conclusions drawn from a study in which the study population is restricted to those people who survived not only the initial atomic blast(s) but also epidemics of infectious diseases, malnutrition, and the devastation of local infrastructure. It is unknown how large an impact this initial selection effect may have on estimates of cancer risk derived from these survivors. Finally, questions persist about uncertainty and bias in radiation dose estimates for A-bomb survivors; the accuracy of survivors' dose estimates depends on the accuracy of the information about location of survivors that was collected by questionnaire, as well as on assumptions about shielding and other factors that influenced survivors' doses.

Given these concerns, studies of workers who have been occupationally exposed to ionizing radiation are also of substantial interest for considerations about worker protection. Studies of workers can provide direct evidence about low dose chronic radiation exposures accrued in the settings of

regulatory interest. Furthermore, occupational studies of radiation health effects offer the opportunity to study humans who can be followed through employment records and occupational medical departments, who were often individually monitored for radiation exposures, and who can be traced through pension benefits and other record sources. Studies of workers in the nuclear power and nuclear weapons industries have focused primarily on the effects of external penetrating radiation, in part because external doses easier to quantify than are doses from internally deposited radionuclides. A notable limitation of studies of workers in the nuclear industry, however, is the relatively low statistical power of most studies when compared to the LSS. Most occupational cohorts include fewer people than the LSS, and the magnitude of radiation doses accrued by workers in the nuclear industry is far lower than the doses accrued by many people in the LSS. Consequently, in most studies, the confidence intervals of the radiation risk estimates are much wider than those derived from analyses of the LSS. However, continued follow-up of these occupational cohorts will continue to increase the weight of evidence derived from study populations that can provide direct evidence about the health effects associated with chronic low level radiation exposures. Ongoing studies of workers employed at nuclear power plants offer the possibility of contributing to this literature and may have greater statistical power than studies of nuclear weapons workers given the higher average annual exposures reported for workers in the nuclear power industry.

A number studies of workers in the nuclear weapons industry have investigated the relationship between recorded occupational radiation doses and mortality. Some of these studies have observed significant increases in all cancer mortality with increasing cumulative radiation doses, which were typically accrued at dose rates below current radiation protection guidelines, while other studies have reported no evidence of radiation-induced cancer. Among those studies that reported evidence of radiation-induced cancer among U.S. nuclear workers are studies of Oak Ridge National Laboratory (ORNL), Fernald, and Rocketdyne. Significant positive trends in cancer mortality with radiation dose have also been reported among workers in the United Kingdom's Atomic Weapons Establishment (AWE) and among workers included in the Canadian National Dose Registry. Significant association between leukemia and radiation dose were observed in studies of workers employed at Rocketdyne and Mound in the United States, Sellafield in the United Kingdom, and AECL in Canada, as well as in the combined analyses of data from workers cohorts in the United Kingdom.

For example, in a study of 124,743 workers in the United Kingdom's nuclear power, fuel cycle, and weapons production industries, the association between external radiation dose and all cancer mortality was $ERR_{1Sv} = 0.09$ (90%CI, −0.28, 0.52), and for leukemia it was $ERR_{1Sv} = 2.55$ (90%CI, −0.03, 7.2). A study of radiation-mortality associations among 206,620 Canadian workers, which included people employed in a variety of (nonmining) occupations in which monitoring for radiation exposure occurs, reported that for males the ERR_{1Sv} for all cancers was 3.0 (90%CI, 1.1, 4.9). No statistically significant association between radiation dose and mortality from all cancers was seen in a combined analysis of radiation-mortality associations among 95,673 nuclear industry workers in the United States, the United Kingdom, and Canada; the ERR_{1Sv} for all cancers excluding leukemia was −0.07 (90%CI, −0.4, 0.3). Under a 2-year lag assumption the ERR_{1Sv} for leukemia excluding CLL was 2.2 (90%CI, 0.1, 5.7). Although these risks estimates are lower than the linear estimates obtained from studies of atomic bomb survivors, they are compatible with a range of possibilities, from a reduction of risk at low doses, to risks twice those on which current radiation protection recommendations are based. Variation in findings between cohorts may reflect complex patterns of bias related to selection into cohorts (as well as selection within cohorts in specific jobs or areas), exposure measurement error, and confounding exposures. Differences in radiation-mortality associations between cohorts could also reflect differences between study populations in the distribution of modifying factors.

Summary estimates of radiation-cancer associations are by definition a weighted average of the dose response relationships for subgroups with different susceptibilities. Differences in sensitivity to the carcinogenic effects of ionizing radiation may arise from genetic predispositions as well as from induced susceptibility from exposure to nonradiological occupational or environmental agents. Some markers of innate differences between people in radiosensitivity have been identified; for example, specific germline mutations are associated with increased vulnerability to the carcinogenic effects of radiation. More generally, there are observations that suggest

that the process of aging may itself influence susceptibility to radiation-induced cancer.

For radiation protection purposes, consideration of age at exposure is crucial. For example, it is important to limit the radiation doses accrued by pregnant females in the workplace because the developing human fetus which is believed to be extremely sensitive to the carcinogenic effects of ionizing radiation. Strong evidence of the radiosensitivity of the human fetus comes from the Oxford Survey of Childhood Cancers (OSCC), which was a national case-control study that included children under 16 years of age in Great Britain who died from any form of malignancy and a matched control (a living child of the same sex, living in the same administrative district in which the death occurred, and born on approximately the same day as the deceased child). The most recent analyses of the OSCC were based on data for 14,759 traced cases who died at ages 0 to 15 years during 1953 to 1979 and their matched controls. Findings from the OSCC have led to estimates of the excess relative risk for childhood cancer at 1 milligray that range from 0.038 to 0.051 (depending on fetal dose estimates and the birth cohorts included in analyses).

For occupational exposures accrued by adults who are not pregnant, age at exposure may also be an important consideration, although there is divergent evidence about age-related changes in sensitivity to the carcinogenic effects of ionizing radiation among adults. For all cancers, estimates of excess relative risk per unit dose in the LSS decline as age at exposure increases. In contrast, some studies of nuclear workers exposed to chronic low levels of ionizing radiation suggest that at older adult ages sensitivity to the carcinogenic effects of ionizing radiation increases. Such increases in radiosensitivity have been postulated to reflect declines in the efficiency of most biological systems, including the body's cellular repair and immune systems. Increases in radiation-cancer associations with age at exposure have been reported in studies of workers employed at Hanford, Oak Ridge National Laboratory, Fernald, and Rocketdyne, although evidence of variation in radiation effects with age at exposure was not found in the combined analyses of U.S., UK, and Canadian nuclear workers. For lung cancer, there is some evidence of increase in the magnitude of the radiation-cancer association with age at exposure in the LSS as well as in several occupational cohort studies. Again, such variation in findings between cohorts may reflect patterns of bias, exposure measurement error, and confounding exposures; furthermore, estimates of the excess relative risk of cancer associated with doses accrued in very narrow time windows, such as those dose response estimates reported for occupational exposures accrued at very old ages, are expected to be particularly unstable.

Effects of internal exposure to ionizing radiation from internal deposition of radionuclides encountered in the nuclear power industry are less well studied in occupational epidemiology. It is known that alpha and beta radiation are highly effective at ionizing and that their effects on cellular molecules can be extremely damaging. There is clear evidence of mutations and structural changes to DNA resulting from alpha radiation; therefore, these exposures are presumed to carcinogenic. Studies of workers in the former Soviet Union nuclear weapons complex provide evidence of increased cancer following relatively high doses from plutonium exposure; however, internal dosimetry is complex, particularly when reconstructing historical dose estimates using records for these workers.

While accidents can result in larger occupational exposures to ionizing radiation, epidemiological investigations of accidental exposures to ionizing radiation have typically been hampered by poor dosimetry data. The most notable of these was the nuclear accident at the Chernobyl power station on April 26, 1986, which resulted in large-scale occupational exposures to ionizing radiation. More than 600,000 workers participated in the clean-up after the accident. Only a small proportional of these workers, often referred to as "liquidators," received individual dosimeters and, among those who were monitored, problems arose in recording and maintaining dosimetric information. Cleanup workers have been reported to exhibit substantially increased incidence of chromosomal aberrations; while studies of cytogenetic damage can offer some information about an individual's exposure, such approaches are poorly suited to systematic investigation of populations. In a nested case control study that included 34 leukemia cases occurring more than 2 years after the accident and 136 controls, no difference was found in official dose estimates. In contrast, increases in the incidence of leukemia have been reported among cohorts of liquidators in Russia, Belarus, and Ukraine. However, the results of these studies must be interpreted with caution because of problems related to the accuracy of diagnoses and differences in surveillance and registration of cancer cases among liquidators compared to population referents.

6. RADIATION HAZARDS IN THE CONTEXT OF THE NUCLEAR FUEL CYCLE

The activities in the nuclear power industry are dependent on the work done in early and later stages of the nuclear fuel cycle. The nuclear fuel cycle begins with the work of uranium mining and milling, uranium enrichment, and either fabrication of nuclear fuel for reactor operations or weapons production. Workers are also employed in the transportation and disposal of wastes and, in some cases, the reprocessing of spent nuclear fuel. These activities take place in different geographic locations and involve workers employed in many different industries and also involve potential occupational exposures to ionizing radiation. The United Nations Scientific Committee on the Effects of Atomic Radiation has published a useful review of the radiation doses accrued by workers in each of stages of the nuclear fuel cycle. For most workers employed in the nuclear fuel cycle, radiation doses tend to be accrued at low dose rates and may include external as well as internal sources of radiation exposure.

Experimental evidence indicates that DNA damage can be produced by an individual radiation track, suggesting that even very low doses of radiation produce observable biological effects. However, there remains uncertainty about the magnitude of risk for adverse health effects associated with low levels of ionizing radiation and uncertainty about the factors that influence individual susceptibility to radiation-induced disease. A better understanding of the occupational risks associated with chronic low level occupational radiation exposures should come from ongoing studies of workers in the nuclear industry, which, it is hoped, will help to inform worker protection and compensation programs.

SEE ALSO THE FOLLOWING ARTICLES

Nuclear Engineering • *Nuclear Power, History of* • *Nuclear Power Plants, Decommissioning of* • *Nuclear Power: Risk Analysis* • *Nuclear Proliferation and Diversion* • *Nuclear Waste* • *Occupational Health Risks in Crude Oil and Natural Gas Extraction* • *Public Reaction to Nuclear Power Siting and Disposal* • *Radiation, Risks and Health Impacts of*

Further Reading

Committee for the Compilation of Materials on Damage Caused by the Atomic Bombs in Hiroshima and Nagasaki. "Hirosima, Nagasaki: The Physical, Medical, and Social Effects of the Atomic Bombings." (1981). Iwanami Shoten, Tokyo.

International Agency for Research on Cancer. "Ionizing Radiation, Part 1: X- and Gamma-Radiation, and Neutrons." (2000). IARC Press, Lyon, France.

National Research Council, Committee on the Biological Effects of Ionizing Radiation (BEIR V). (1990). "Health Effects of Exposure to Low Levels of Ionizing Radiation (BEIR V)." National Academy Press, Washington, DC.

United Nations Scientific Committee on the Effects of Atomic Radiation. (2000). "Sources and Effects of Ionizing Radiation." United Nations, New York.

U.S. Nuclear Regulatory Commission. (2001). "Occupational Radiation Exposure at Commercial Nuclear Power Reactors and Other Facilities, 2001: Thirty-Fourth Annual Report." Report No. NUREG-0713, Vol. 23. U.S. Nuclear Regulatory Commission, Office of Nuclear Regulatory Research, Washington, DC.

Ocean, Energy Flows in

RUI XIN HUANG
Woods Hole Oceanographic Institution
Woods Hole, Massachusetts, United States

1. **Basic Conservation Laws for Ocean Circulation**
2. **Energetics of the Oceanic Circulation**
3. **Heat Flux in the Oceans**
4. **The Ocean Is Not a Heat Engine**
5. **Mechanical Energy Flows in the Ocean**
6. **Major Challenge Associated with Energetics, Oceanic Modeling, and Climate Change**
7. **Conclusion**

Glossary

Ekman layer A boundary layer within the top 20 to 30 m of the ocean where the wind stress is balanced by Coriolis and frictional forces.

mixed layer A layer of water with properties that are almost vertically homogenized by wind stirring and convective overturning due to the unstable stratification induced by cooling or salinification.

thermocline A thin layer of water at the depth range of 300 to 800 m in the tropical/subtropical basins, where the vertical temperature gradient is maximum.

thermohaline circulation An overturning flow in the ocean driven by mechanical stirring/mixing, which transports mass, heat, freshwater, and other properties. In addition, the surface heat and freshwater fluxes are necessary for setting up the flow.

wind-driven circulation A circulation in the upper kilometer of the ocean, which is primarily controlled by wind stress. In addition, this circulation also depends on buoyancy forcing and mixing.

Oceanic circulation plays an important role in the climate system because it transports heat and freshwater fluxes meridionally. There is a huge heat flux through the upper surface of the ocean, but the ocean is not a heat engine; instead, the ocean is driven by external mechanical energy, including wind stress and tides. Although mechanical energy flux is 1000 times smaller than the heat flux, it controls the strength of the oceanic general circulation. In particular, energy input through wind stress is the most important source of mechanical energy in driving the oceanic circulation.

1. BASIC CONSERVATION LAWS FOR OCEAN CIRCULATION

Oceanic circulation is a mechanical system, so it obeys many conservation laws as other mechanical systems do, such as the conservation of mass, angular momentum, vorticity, and energy. It is well known that these laws should be obeyed when we study the oceanic circulation either observationally, theoretically, or numerically. However, this fundamental rule has not always been followed closely in previous studies of the oceanic circulation. As a result, our knowledge of the energetics of the ocean circulation remains preliminary.

1.1 Mass Conservation Law

This is one of the most important laws of nature. However, mass conservation has not been applied in many existing numerical models. In fact, most models are based on the Boussinesq approximations, which consist of three terms: mass conservation is replaced by a volume conservation, *in situ* density in the horizontal momentum equations is replaced by a constant reference density, and tracer prognostic equations are based on volume conservation instead of mass conservation. For models based on the Boussinesq approximations, heating (cooling) can induce an artificial loss (gain) of gravitational potential energy. Furthermore, precipitation (evaporation) is often simulated in terms of the equivalent virtual salt flux out of (into) the ocean; thus, an artificial sink of gravitational potential energy is generated in the models.

 © 2004 Elsevier Inc. All rights reserved.

1.2 Angular Momentum Conservation Law

In a rotating system, the linear momentum is not conserved; instead the angular momentum conservation is more relevant (i.e., the rate of change of the angular momentum of the oceanic general circulation must be balanced by torques). The angular momentum conservation law has been used extensively in the study of the atmospheric circulation. The Antarctic Circumpolar Current is the corresponding part in the oceans where the angular momentum conservation law dictates the balance.

1.3 Potential Vorticity Conservation Law

The potential vorticity equation can be obtained by cross-differentiation of the horizontal momentum equations and subtracting the results; it has been extensively used in oceanographic studies. According to the theory, total potential vorticity is conserved in the interior of a stratified fluid, and the only source of potential vorticity must be on the outer surfaces of the stratified fluid.

1.4 Energy Conservation Law

Motions in the oceans must be associated with friction; thus, to maintain circulation in the ocean, mechanical energy source is required to balance the loss of mechanical energy. Energy source and dissipation are the most important means of controlling the oceanic circulation system. However, over the past years the mechanical energy balance of the oceanic general circulation has been overlooked, and there has been very little discussion about such a fundamental issue. In fact, energy balance equations for the oceanic general circulation in most previously published books or papers are incomplete because one of the most important terms in the mechanical energy balance, energy required for vertical mixing in a stratified environment, is not included. Therefore, there is no complete statement about mechanical energy balance for the oceanic general circulation. In addition, most numerical models do not conserve total energy, especially in the finite difference scheme in time stepping. Thus, energetics diagnosed from these models are unreliable.

2. ENERGETICS OF THE OCEANIC CIRCULATION

Energy in the ocean can be classified into four major forms: the gravitational potential energy, kinetic energy, internal energy, and chemical potential, as shown in Fig. 1. Although the chemical potential can be considered as part of the internal energy, it is listed as a separate item in the diagram.

There are four sources (sinks) of gravitational potential energy generated by external forcing. A major source is the tidal force, which is due to the temporal variation of the gravitational forces of the moon and the sun. Although tidal motions have been studied separately from the oceanic general circulation in the past, it is now recognized that a strong

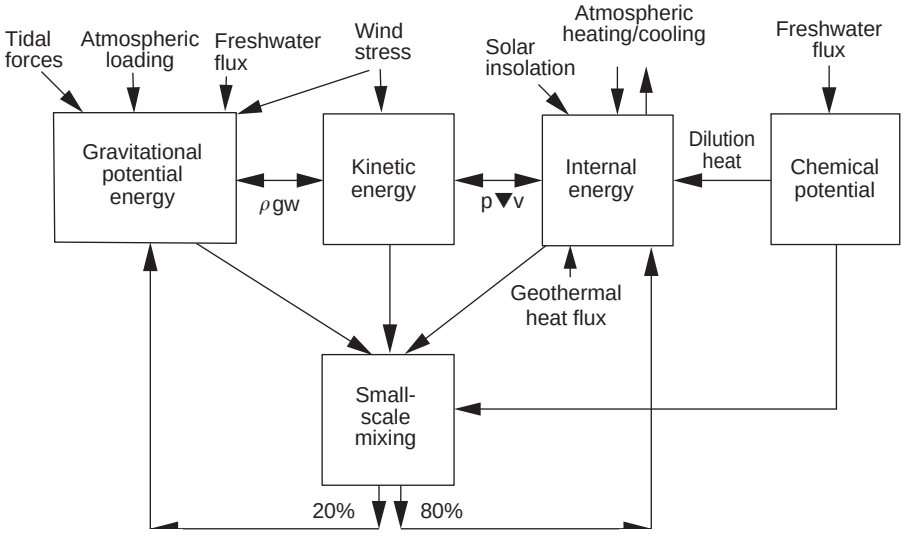


FIGURE 1 Energetics diagram for the ocean circulation.

connection exists between the tidal motions and the oceanic general circulation.

In addition, changes of the sea level atmospheric pressure (called the atmospheric loading) can generate gravitational potential energy. The total amount of this energy flux remains unclear. Furthermore, freshwater flux across the air-sea interface can generate gravitational potential energy. This turns out to be a small sink of energy, and the global integrated amount of energy flux is about $-0.007TW$ ($1TW = 10^{12}$ *Watts*). The reason for this loss of gravitational potential energy is that evaporation (precipitation) prevails at subtropics (high latitudes) where sea level is high (low) due to warm (cold) temperature. Wind stress is one of the most important sources of mechanical energy for the ocean. In fact, wind stress generates surface waves, thus providing a source of mechanical energy, which is equally partitioned into the gravitational potential energy and kinetic energy. The other source of kinetic energy is the energy input from the wind stress to the surface geostrophic currents and the ageostrophic currents or the Ekman drift.

The external sources of internal energy include three terms: the solar radiation, heat exchange to the atmosphere, and geothermal heat flux. Heat exchange to the atmosphere includes latent heat flux, long wave radiation, and sensible heat flux. Geothermal heat flux through the sea floor is much smaller than the air-sea heat fluxes, but it is one of the important factors controlling the deep circulation in the abyssal ocean.

The source of the chemical potential is associated with the salinity, which is generated by freshwater flux through the air-sea interface in forms of evaporation and precipitation. The general role of chemical potential in the oceanic circulation remains unclear. One of the major roles played by the chemical potential is driving the molecular mixing of salt.

2.1 Energy Conversion

Energy in any given form can be converted into other forms through specific processes, as highlighted in Fig. 1. For example, gravitational potential energy and kinetic energy can be transferred to each other through the vertical velocity conversion term. In addition, kinetic energy can be converted into gravitational potential energy through small scale mixing by internal waves and turbulence. On the other hand, gravitational potential energy can be converted into kinetic energy of meso-scale eddies through baroclinic instability. This may be one of the major pathways of energy flow in the ocean.

2.2 Conversion from Kinetic Energy to Gravitational Potential Energy

In a turbulent ocean, the conversion from kinetic to potential energy is represented by $\overline{\rho w g}$, where the overbar indicates the ensemble mean. This exchange can be separated into several terms:

$$\overline{\rho w g} = \bar{\rho}\bar{w}\bar{g} + \overline{\rho' w'}\bar{g} + (\overline{\rho' w'} + \rho' \bar{w} + \bar{\rho} w').$$

The first term on the right-hand side represents the kinetic to potential energy conversion associated with the ensemble mean density and velocity fields. The second term on the right-hand side represents the contribution due to turbulent mixing. In a stratified fluid, upward motion is usually correlated with positive density anomaly, so turbulent mixing in a stratified ocean increases the gravitational potential energy of the state, and thus requires a source of kinetic energy. The rest of the terms on the right-hand side represent the kinetic to potential energy conversion due to the time-dependent component of the gravitational force, such as the tidal forces.

In the stratified oceans, the vertical eddy mixing rate κ can be related to the energy dissipation rate through

$$\kappa N^2 = a\varepsilon, N^2 = -\frac{g}{\rho_0}\frac{\partial \rho_\theta}{\partial z},$$

where N^2 is the buoyancy frequency, ε is energy dissipation rate, and $\alpha \approx 0.2$ is an empirical coefficient. Therefore, the decline in turbulent kinetic energy does not entirely become "wasted" heat. In fact, a small percentage of the turbulent kinetic energy "dissipation," represented by the empirical coefficient a, is actually transformed to large-scale potential energy. The energy required to sustain mixing is one of the most important dynamic factors controlling the meridional overturning rate.

2.3 Conversion of Internal Energy to Mechanical Energy

Internal energy can be converted to kinetic energy through the compression term. Since sea water is nearly incompressible, this conversion rate is rather low. In addition, it is very important to note that energy transformation is not unconditional. Instead, the conversion between internal energy and mechanical energy is subject to the second law of thermodynamics—that is, mechanical energy can be converted entirely into internal energy; however, internal energy can be converted into mechanical energy only partially, with the upper bound set by the

Carnot efficiency. As will be discussed later, however, the ocean is not a heat engine; therefore, the equivalent thermal-mechanical energy conversion efficiency is negative.

3. HEAT FLUX IN THE OCEANS

Oceans play an important role in the earth's climate system in many ways. In fact, most of the solar radiation penetrates the atmosphere and is first absorbed by the ocean and the land and then sent back to the atmosphere in forms of latent heat, long wave radiation, and sensible heat. To a large degree, these processes are controlled by the oceanic circulation. In the quasi-steady state, all heat flux absorbed by the ocean must be released back into the atmosphere. The spatial distribution of air-sea heat flux is one of the critical components of the earth's climate system.

3.1 Heat Flux through the Upper Surface of the Ocean

The dominating form of energy flux in the oceans is the heat flux. Solar radiation on Earth can penetrate the atmosphere, with only a small fraction being absorbed by the greenhouse gases, such as water vapor and CO_2. Since oceans cover about 70% of the earth, most of the solar radiation is directly absorbed by the oceans (total amount of 52.4*PW*).

At low latitudes, solar radiation absorbed by the ocean exceeds the heat flux from the ocean into the atmosphere, including the latent heat flux, long wave radiation, and the sensible heat flux. Thus, the ocean at low latitudes gains net heat through the upper surface, and this extra heat is transported poleward and released at middle and high latitudes. The most important areas of the ocean-to-atmosphere heat release include the Gulf Stream, Kuroshio, and the Northern North Atlantic.

3.2 Poleward Heat Fluxes in the Climate System

Heat fluxes in the current climate system consist of three components: the atmospheric sensible heat flux, the oceanic sensible heat flux, and the latent heat flux associated with the hydrological cycle in the atmosphere-ocean coupled system (Fig. 2). The sum of these three fluxes is the total poleward heat flux diagnosed from satellite measurements. The ocean plays the role of heat storage and heat redistributor. The most important part for climate study is the total

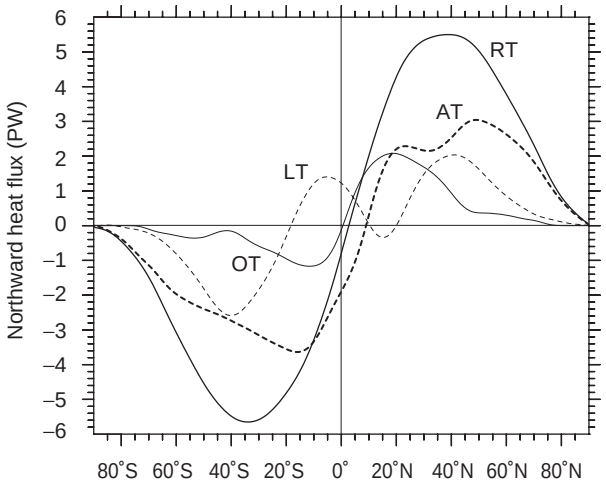


FIGURE 2 Northward heat flux in PW: RT indicates the heat flux required for radiation balance as diagnosed from satellite data; AT indicates the atmospheric sensible heat flux; OT indicates the oceanic sensible heat flux; and LT indicates the latent heat flux associated with the hydrological cycle.

amount of poleward heat flux associated with the oceanic current, which is about 2*PW*. Note that the latent heat flux, which has been traditionally associated with the moisture flux in the atmosphere, can be considered as an atmosphere-ocean coupled mode.

Although the atmospheric sensible heat flux is the largest component in both hemispheres, it is not much larger than the other components. Thus, the total oceanic contribution to the poleward heat flux may consist roughly of 50% of the total heat flux.

The poleward sensible heat flux in the ocean can be separated into two major components: the flux associated with the meridional overturning cell and the flux associated with the horizontal gyration. Each component can be further divided into the contribution by the mean Eulerian flow and eddies. For example, poleward heat flux in the North Atlantic reaches its maximum around 25°N, where both the mean meridional overturning and horizontal wind-driven gyre contribute. In comparison, however, the contribution by eddies is relatively small. In the North Pacific, the mean meridional overturning cell induces an equatorward heat flux; however, the sum of heat fluxes is still poleward because it is associated with the domination of the wind-driven gyre.

3.3 Total Heat Flux through the Oceans

The oceanic to atmospheric heat flux consists of three components: the latent heat flux, 31.1*PW*; the long wave radiation, 17.8*PW*; and the sensible heat flux, 3.5*PW* (Fig. 3). Although air-sea interface heat flux is

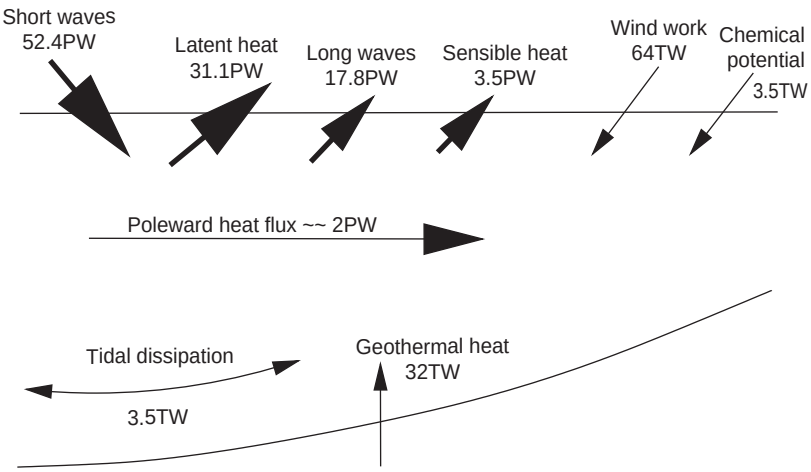


FIGURE 3 A diagram of external sources of energy for the ocean circulation.

of major importance to climate study, it does not directly drive the oceanic general circulation. Instead, air-sea heat flux is controlled by the oceanic circulation or, more accurately, controlled by the atmosphere-ocean-land coupled system. From the energetic point of view, the oceanic circulation is governed by many laws of conservation, in particular the energy conservation law. The thermohaline boundary conditions at the sea surface are preconditions for the thermohaline circulation only, while the ocean circulation is primarily controlled by the sources and sinks of mechanical energy, with the total air-sea heat flux as the by product of the circulation.

In Fig. 3, energy fluxes into the ocean in other forms are also included because this additional energy input eventually turns into internal energy and is released in to the atmosphere in forms of heat flux. The amount of geothermal heat flux is about $32TW$, which is eventually lost in to the atmosphere through the air-sea heat exchange. Furthermore, there are three other fluxes: the tidal dissipation, $3.5TW$; the chemical potential generated through evaporation and precipitation; and the wind work, $64TW$. Thus, the total outgoing heat flux is the sum of the incoming heat flux, $52.4PW$, plus four additional energy sources: 64 (wind) $+ 32$ (geothermal) $+ 3.5$ (tides) $+ 3.5TW$ (chemical potential) $= 103TW$.

4. THE OCEAN IS NOT A HEAT ENGINE

4.1 Is the Ocean a Heat Engine?

The atmosphere and oceans work together as a heat engine, if we neglect the small contribution of tidal energy to the circulation in the oceans. The atmosphere can be considered as a heat engine, which is driven by differential heating, with an efficiency of 0.8% (the corresponding Carnot efficiency is about 33%). In this sense although the oceans are subject to differential heating similar to the atmosphere, they are not at all a heat engine. As will be discussed here, the driving force for the oceanic circulation is the mechanical energy in the forms of wind stress and tides. In comparison, differential heating is only a precondition for the thermohaline circulation, and not the driving force of the oceanic general circulation. Thus, the ocean is not a heat engine; instead, it is a machine driven by external mechanical energy that transports thermal energy, freshwater, CO_2, and other tracers.

4.2 Sandstrom's Theorem

Sandstrom theorem states: A closed steady circulation can be maintained in the ocean only if the heating source is situated at a level lower than the cooling source. Sandstrom's theorem can be demonstrated by laboratory experiments, as shown in Fig. 4. In the first experiment, the heating source is put at a level higher than the cooling source. There is no detectable circulation, and a stable stratification can be observed between the heating and cooling levels. Theoretically, there is still a circulation driven by the molecular mixing, but it is so sluggish that there is practically none. In the second experiment, the heating source is put at a level lower than the cooling source. A vigorous circulation can be observed between the levels of heating and cooling, just as we put a pot of water over the kitchen range.

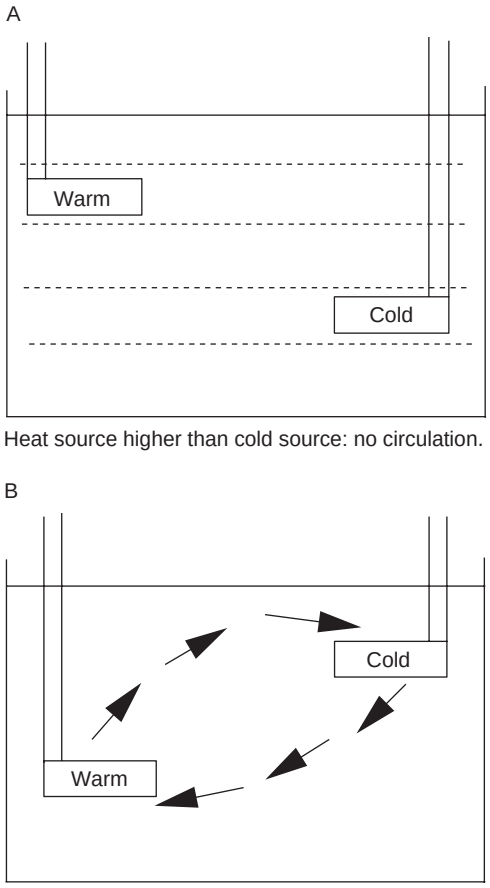


FIGURE 4 Laboratory experiments demonstrating the Sandstrom theorem.

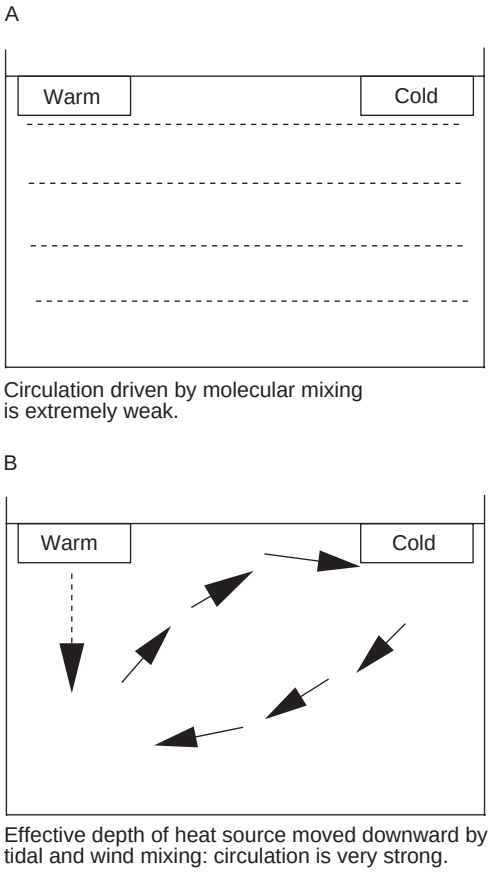


FIGURE 5 Application of the Sandstrom theorem to the world's oceans.

Sandstrom's theorem applies to both the atmosphere and oceans. The atmosphere is heated from below and cooled from above; thus, it is a heat engine. On the other hand, the ocean is heated and cooled from the sea surface. Since heating and cooling take place at the same level, there is no circulation driven by the thermal forcing (Fig. 5A). Under such a condition, thermal forcing can create a very thin thermocline, on the order of meters on the surface of the oceans. Below such a thin thermocline, temperature is virtually constant, equal to the coldest temperature set up by the deepwater formation at high latitudes. Since there is no density difference at depth, there is only a very sluggish circulation. However, the effective depth of heating is moved downward by wind and tidal mixing. The large horizontal density gradient thus created drives a strong circulation in the world oceans (Fig. 5B).

The essential point is that to maintain a quasi-steady circulation in the ocean, external sources of mechanical energy are required to balance the loss of mechanical energy caused by friction associated with motion. Although there is a tremendous amount of thermal energy going through the ocean, it cannot be converted into mechanical energy needed to sustain the oceanic circulation, as dictated by the fundamental law of thermodynamics. Therefore, instead of utilizing the huge amount of thermal energy going through the ocean, the maintenance of the oceanic circulation depends on the availability of external mechanical energy, with a flux rate of three to four orders of magnitude smaller than the thermal energy fluxes (Fig. 3).

5. MECHANICAL ENERGY FLOWS IN THE OCEAN

Mechanical energy sources and sinks for the oceans are illustrated in Fig. 6. Wind stress applies to the

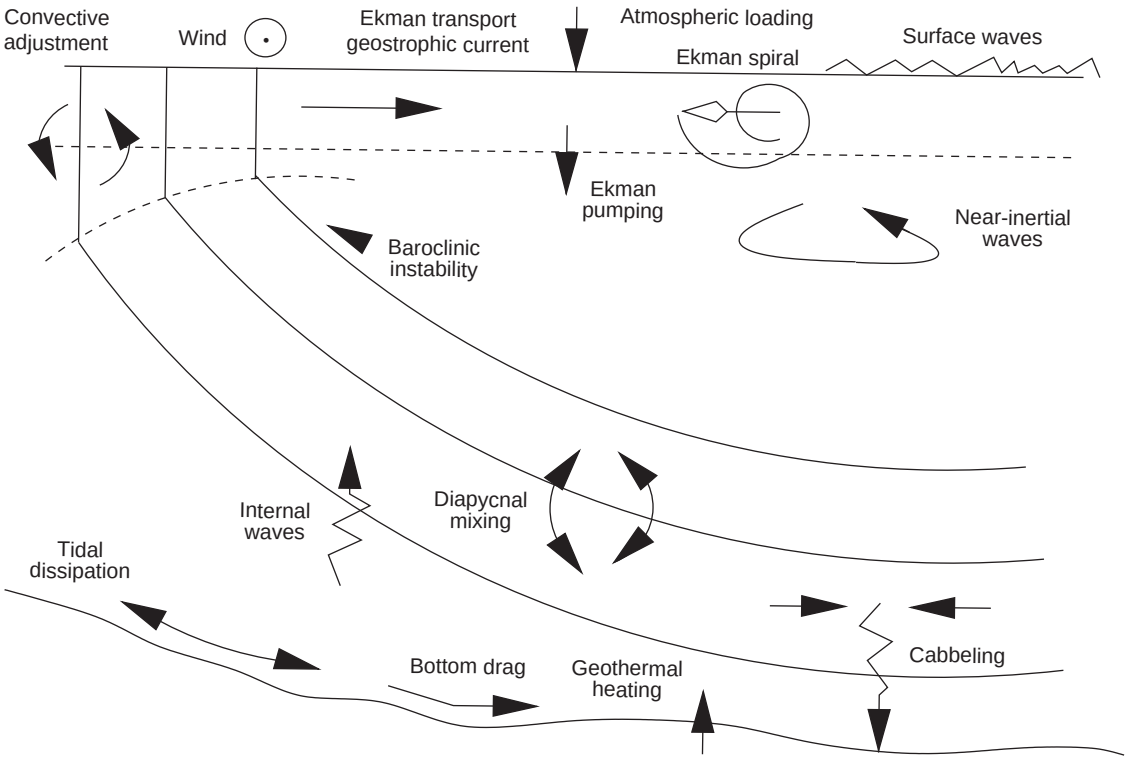


FIGURE 6 Mechanical energy diagram for the ocean circulation.

surface of the ocean and drives both surface currents and waves. The mechanical energy input from wind stress into the ocean W_{wind} is defined by

$$W_{wind} = \langle \sigma_{ij} \rangle \cdot \vec{u} = \vec{\tau} \cdot \vec{U}_0 + \overline{\vec{\tau}' \vec{u}'_0} + \overline{p' w'_0},$$

where $\vec{\tau}$ and $\vec{U}_0$ are the spatially averaged tangential stress and surface velocity respectively, $\vec{\tau}'$ and $\vec{u}'_0$ are the perturbations, and p' and w'_0 are perturbations of the surface pressure and velocity component normal to the surface. The perturbations are defined in terms of the scale of the surface wavelength.

The first term on the right-hand side is the wind stress work of the quasi-steady current on the surface, and quasi-steadiness is defined in comparison to the timescale of the typical surface waves. $\vec{U}_0$ has two components: the geostrophic current and the Ekman drift:

$$\vec{U}_0 = \vec{U}_{0,G} + \vec{U}_{0,Ekman.}$$

5.1 Geostrophic Current

Energy input through the surface geostrophic current can be calculated as the scaler product of wind stress and surface geostrophic velocity. The recent calculation by Carl Wunsch puts the global total as $1.3TW$. This energy is directly fed into the large-scale current, so it can be efficiently turned into gravitational potential energy through the vertical velocity conversion term.

5.2 Surface Drift

There is a frictional boundary layer (Ekman layer) on the top of the ocean, in which the wind stress is balanced by frictional force. The surface drift is called the Ekman drift. The exact amount of wind stress energy input through the Ekman spiral is unclear, and the preliminary estimate made by Wei Wang and Rui Xin Huang is about $2.4TW$ for the frequency higher than 1/(2 days). In addition, there is a large amount of energy input through the near-inertial waves, which is due to the resonance at a frequency of $\omega = -f$; that is, for inertial frequency of the clockwise (anticlockwise) wind stress in the Northern (Southern) Hemisphere. Recent calculations by Watanaba and Hibiya gave an estimate of $0.7TW$.

5.3 Geopotential and Kinetic Energy of the Wind-Driven Circulation

The convergence of the Ekman flux gives rise to a pumping velocity at the base of the Ekman layer W_E,

which is responsible for pushing the warm water into the subsurface ocean and thus forming the main thermocline in the subtropical ocean. The main thermocline is a relatively thin layer of water at the depth of 500 to 800 m in the upper ocean where the vertical temperature gradient is maximum. Since salinity contrition to the density is relatively small in most places of the oceans, the thermocline is also close to the pycnocline where the vertical gradient of density is maximum. Ekman pumping sets up the wind-driven circulation and the associated bow-shaped main thermocline in the world's oceans. In this process, gravitational potential energy is increased, which is a very efficient way of converting kinetic energy into gravitational potential energy. Simple scaling analysis shows that the amount of available gravitational potential energy associated with the main thermocline is about 1000 times larger than the amount of kinetic energy associated with the mean flow of the wind-driven circulation.

5.4 Loss of Gravitational Potential Energy through Convective Adjustment

Another important process taking place in the mixed layer is the convective adjustment due to cooling and salinification. According to the Sandstrom theorem, if both heating and cooling apply to the same level (the sea surface), there is no gravitational potential energy being generated by the thermal forcing alone. However, heating and cooling in the mixed layer is not a linear process. During the cooling process, water at the sea surface becomes heavier than water at the depth, thus a gravitational unstable stratification appears. Due to this unstable stratification, a rapid convective adjustment process takes place and density becomes nearly homogenized in the upper part of the water column. During this process, gravitational potential energy of the mean state is converted into kinetic energy for turbulence and internal waves (Fig. 7A).

As a result, the effective center of cooling is not at the sea surface; instead it is located halfway of the well-mixed layer's depth. Because of this asymmetry associated with cooling/salinification, surface buoyancy forcing gives rise to a sink of gravitational potential energy. The total amount of energy loss through this process remains unclear. A preliminary estimate based on the monthly mean climatology by Rui Xin Huang and Wei Wang for this sink term is about 0.24*TW*. However, calculation in which the diurnal cycle is resolved may give rise to a much larger value.

5.5 Baroclinic Instability

The steep isopycnal surface in the oceans, such as that along the meridional edges of the wind-driven gyre, is unstable due to baroclinic instability. Because of this, a large amount of gravitational potential energy of the mean state can be converted into the eddy kinetic/potential energy (Fig. 7B). It is estimated that the amount of eddy kinetic energy is about 100 times larger than the kinetic energy of the time-mean flow. Despite great effort in carrying out field observations and numerical simulation, there is no reliable estimate on the total amount of eddy kinetic energy in the world's oceans. However, satellite observations have provided global distribution of both the mean and eddy kinetic energy on the sea surface. The ratio of these two forms of kinetic energy is on the order of 100, consistent with

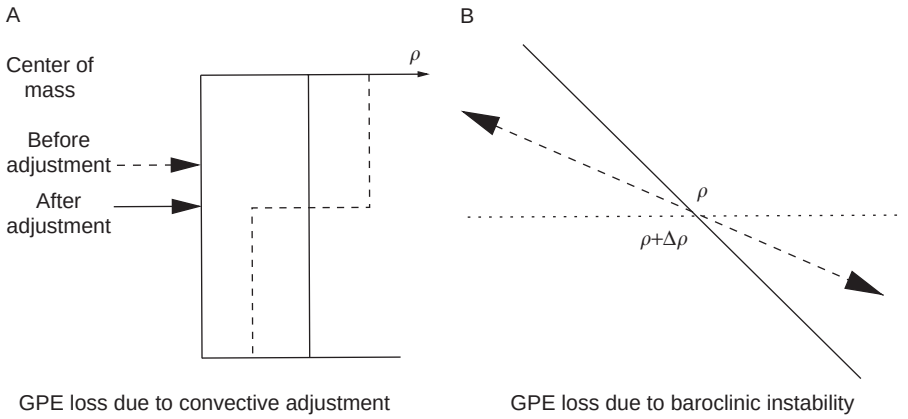


FIGURE 7 Gravitational potential energy loss due to convective adjustment and baroclinic instability.

theoretical prediction based on scaling. The rate of energy conversion over the world's oceans through the baroclinic instability remains unknown, and the recent estimate by Rui Xin Huang and Wei Wang is about $1.1TW$. Since most eddy energy is dissipated through small-scale processes, the conversation from the mean state to eddies is considered as a sink of the gravitational potential energy of the mean state.

5.6 Surface Waves

Another major source of energy caused by wind stress is fed through the surface waves. Wind stress drives surface waves in the oceans, and this energy input can be treated as the form drag for the atmospheric boundary layer. The rate of energy input through surface waves remains unclear; however, the preliminary estimate by Wei Wang and Rui Xin Huang is about $60TW$. Although this seems to be a large amount of energy, it is believed that most of it is transformed into long waves and propagates away from the source area. It is likely that only a small fraction of this energy is dissipated locally through wave breaking and white capping, but a major fraction of this incoming energy may be dissipated to remote places in forms of swell. These long waves have a rather low dissipation rate, their energy is likely to dissipate cascading into other forms of energy or through wave breaking along the beaches in the world's oceans; however, the details of this dissipation mechanism remain unclear.

5.7 Diapycnal and along Isopycnal Mixing

Tracers, including temperature and salinity, are mixed in the oceans through internal wave breaking and turbulence. Mixing can be classified into diapycnal mixing and along isopycnal mixing. Since along isopycnal mixing involves the least amount of gravitational potential energy, it is the dominating form of mixing on large scales.

On the other hand, in stratified fluid vertical mixing increases gravitational potential energy because light fluid is pushed downward and heavy fluid is pushed upward. Although molecular diffusion can play the role of mixing, the corresponding rate is too small and thus the result is irrelevant to the oceans. In the upper ocean, below the mixed layer, diapycnal mixing rate is on the order of $10^{-5}\,m^2s^{-1}$, which is much larger than the mixing rate due to molecular mixing. In other places of the oceans, the mixing rate can be much higher than this background rate. The strong diapycnal (or vertical) mixing in the oceans is driven by strong internal wave and turbulence.

The energy sources supporting diapycnal mixing in the oceans include wind stress input through the sea surface and tidal dissipation, which will be discussed in the next section. Turbulent mixing associated with fast current, especially that associated with flow over sill and the down-slope flow afterward, can provide a strong source of energy supporting mixing. There is much observational evidence indicating that mixing can be on the order of 10^{-3}–$10^{-1}m^2s^{-1}$. The mixing rate can be as large as $0.1\,m^2s^{-1}$ within the Romanche Fracture Zone, where water drops more than 500 m within 100 km of a downward flow.

Internal lee waves generated by flow over topography, such as the Antarctic Circumpolar Current, are probably one of the most important contributors.

The total amount of energy supporting diapycnal mixing in the oceans remains unclear because mixing is highly nonuniform in space and time. According to the theory of thermohaline circulation, the meridional overturning rate is directly controlled by the strength of the external mechanical energy available for supporting mixing in the oceans, including mixing in the mixed layer and mixing in the subsurface layers. Therefore, understanding the physics related to the spatial and temporal distribution of mixing is one of the most important research frontiers in physical oceanography.

5.8 Tidal Dissipation

The primary source of mechanical energy supporting mixing likely comes from both the wind stress applied to the ocean's surface and tidal dissipation in the deep ocean. The total amount of tidal dissipation in the world's oceans is $3.5TW$ (Fig. 8), which is calculated from accurate tracking of the moon's orbit. The spatial distribution of tidal energy dissipation remains inaccurate. Based on satellite altimeter data assimilation, Walter Munk and Carl Wunsch estimated that $2.6TW$ is dissipated within the shallow seas of the world's oceans, and the remainder $0.9TW$ is believed to be distributed in the deep ocean (Fig. 6). Tidal dissipation in the deep ocean takes place primarily over rough topography, where barotropic tidal energy is converted into energy for internal tides and internal waves that sustain the bottom-intensified diapycnal mixing on the order of $10^{-3}\,m^2s^{-1}$. The conversion of kinetic energy to gravitational potential energy through

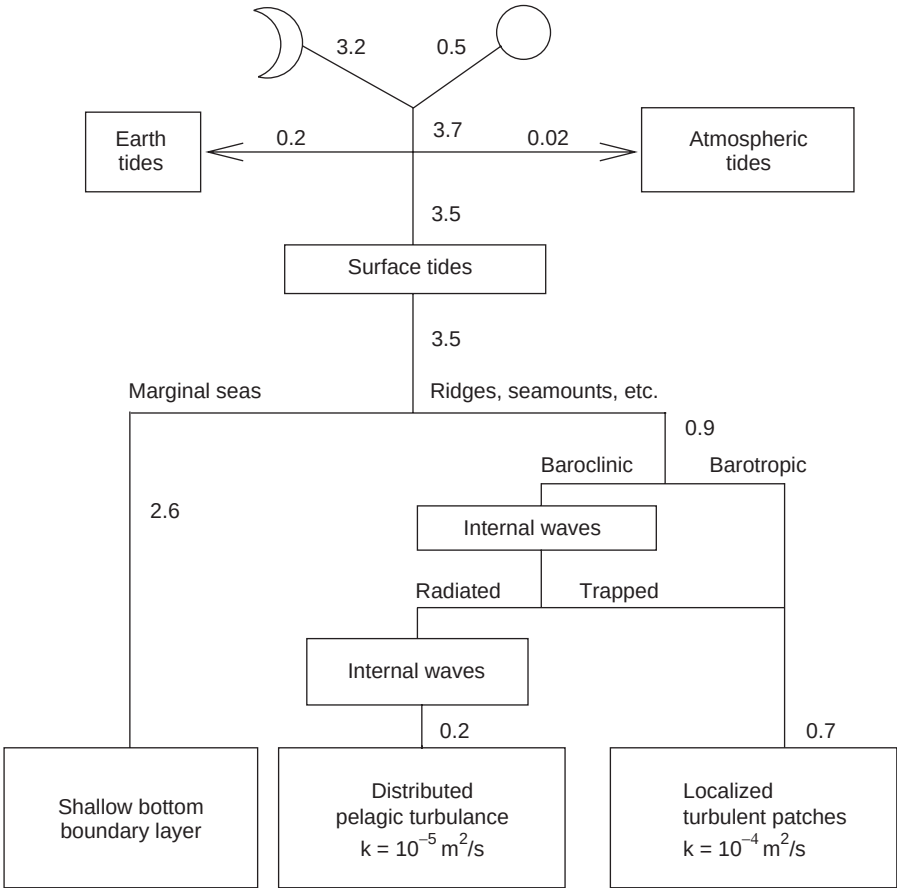


FIGURE 8 Tidal dissipation diagram, unit in *TW*. Modified from Munk and Wunsch (1998).

internal waves and turbulence is of rather low efficiency, typically on the range of 20%; thus, the amount of gravitational potential energy generated by tidal dissipation in the deep ocean is about 0.18*TW*.

Tidal dissipation has varied greatly over the geological past, thus the energy supporting diapycnal mixing may vary as well. The reader is referred to the review by Kagan and Sundermann.

5.9 Geothermal Heat

In addition, geothermal heat or hot plumes provide a total heat flux of 32*TW*. Although this is much smaller than the heat flux across the air-sea interface, it may be a significant component of the driving force for the abyssal circulation. Since the geothermal heating is applied at great depth, and the corresponding cooling takes place at the sea surface, geothermal heat can be more efficiently converted into gravitational potential energy. The rate of this conversion is about 0.05*TW*, which is a small term compared to other major terms; nevertheless it is not negligible, especially for the abyssal circulation and temperature distribution in the abyss.

5.10 Bottom Drag

Oceanic currents moving over rough bottom topography must overcome bottom or form drag, as shown in Fig. 6. The total amount of bottom drag for the world's oceans circulation remains unclear; however, the preliminary estimate for the open oceans is about 0.4*TW*.

5.11 Atmospheric Loading

Finally, sea level atmospheric pressure varies with time, and the sea surface moves in the vertical direction in response. As a result, changes in sea level atmospheric pressure can input mechanical energy into the oceans. The total amount of energy due to this source for the world's ocean circulation remains unclear; however, the preliminary estimate for the open oceans is about 0.04*TW*.

5.12 Miscellaneous

Many other mechanisms contribute to the mechanical energy balance in the oceans.

Cabbeling. Due to the nonlinearity of the equation of state, water density is increased during both diapycnal and isopycnal mixing. The newly formed water with higher density sinks is called cabbeling, as shown in Fig. 6. As a result, gravitational potential energy is converted into energy for internal waves and turbulence. The total amount of gravitational potential energy loss associated with cabbeling in the oceans remains unclear.

Double diffusion: Seawater contains salt, thus it is a two-component chemical mixture. On the level of molecular mixing, the heat diffusion is 100 times faster than the salt diffusion. The difference in heat and salt diffusion for laminar fluid and turbulent fluid environment is one of the most important aspects of the thermohaline circulation in the oceans.

Double diffusion in the oceans primarily manifests itself by two types: salt fingers and diffusive convection. Salt fingers appear when warm and salty water lie over cold and freshwater. The mechanism that drives the instability is illustrated in Fig. 9A, assuming a vertical pipe connects the cold and freshwater in the lower part of the water column with the warm and salty water in the upper part of the water column. The wall of the pipe is very thin, so it allows a rather efficient heat flux into the pipe thus warming up the cold water. At the same time, the low salt diffusivity preserves the freshness of the ascending water parcel. Thus, the buoyancy difference drives the upward motion of water inside the pipe, and a self-propelled fountain can be built in the ocean.

The subtropical gyre interior is a salt-finger's favorite place, where strong solar insolation and excessive evaporation leads to warm and salty water above the main thermocline. In this salt-finger favorite regime, warm and salty fingers move downward and cold and fresh plumes move upward. Since heat is 100 times more easily diffused between the salt fingers and the environment, salt fingers lose their buoyancy and continue their movement downward. In this way gravitational potential energy released is used to drive the mixing. According to the new insite observations in the upper thermocline of the subtropical North Atlantic, the mixing rate for the temperature (salinity) is on the order of $4 \times 10^{-5} m^2 s^{-1} (8 \times 10^{-5} m^2 s^{-1})$, and the equivalent density mixing rate is negative. The total amount of gravitational energy loss due to salt fingering in the world's oceans remain unknown.

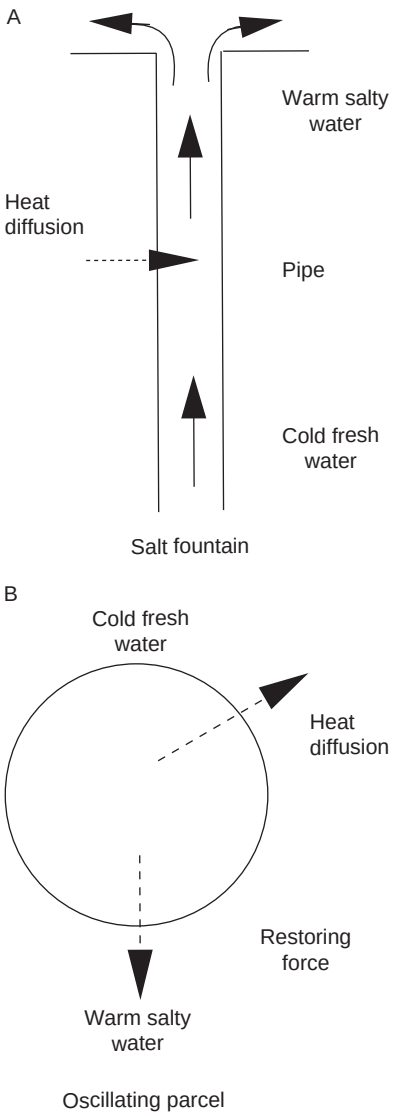


FIGURE 9 Two possible cases for double diffusion in the oceans.

The other possible double diffusive process is the diffusive convection which takes place if cold and freshwater overlay warm saltwater. The system is relatively stable and allows an oscillatory instability (Fig. 9b).

5.13 Attempt at Balancing the Mechanical Energy in the Ocean

The balance of mechanical energy, including both the kinetic energy and gravitational potential energy, is shown in Fig. 10. It is clear that at this time we do not know even the lowest-order balance of mechanical energy. There is much kinetic energy input into

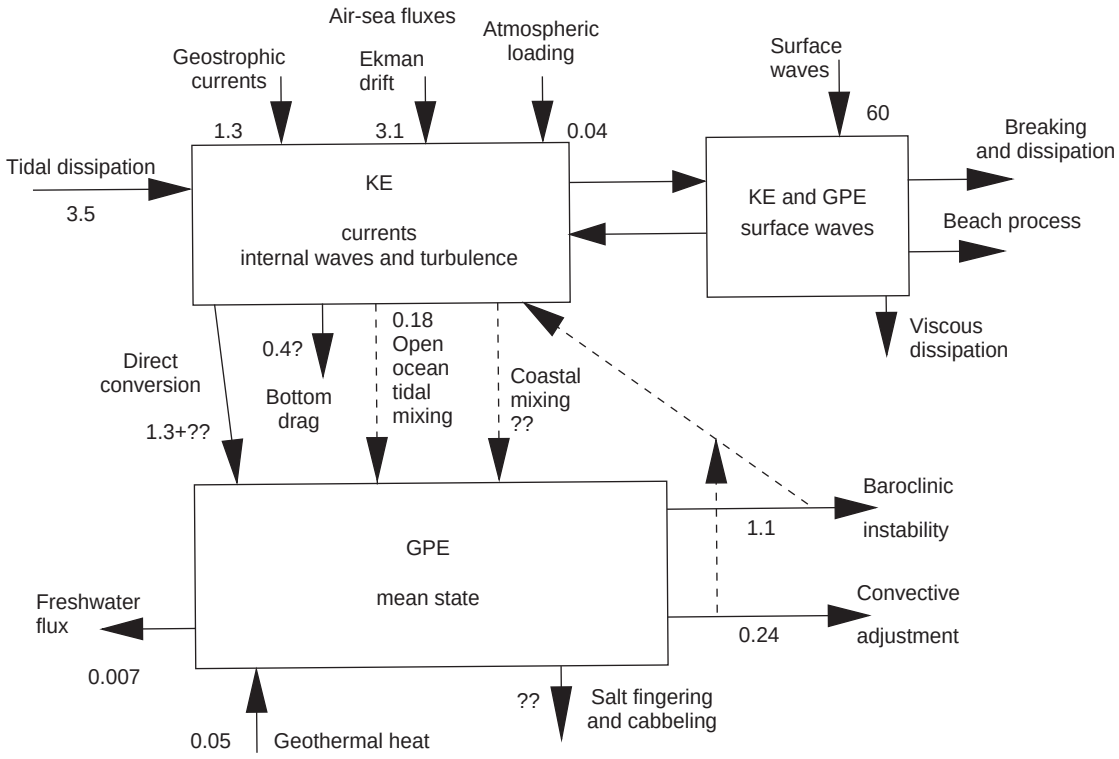


FIGURE 10 Mechanical energy balance for the world's oceans, unit in *TW*.

the ocean, but it is not clear how such energy is distributed and eventually dissipated in the oceans. On the other hand, there are two major sinks of gravitational potential energy, but it is not clear how energy is transported and supplied to such energy sinks. It is fair to say that most of the energy fluxes listed in Fig. 10 are accurate to the factor of 2 only. More accurate energy pathways and estimates require further study.

6. MAJOR CHALLENGE ASSOCIATED WITH ENERGETICS, OCEANIC MODELING, AND CLIMATE CHANGE

6.1 Scaling Laws Governing the Meridional Circulation

Scaling analysis shows that the meridional mass flux and heat flux are linearly proportional to the mechanical energy available for mixing. Thus, the amount of energy available for supporting mixing is what really controls the strength of circulation, so more energy available for mixing means stronger circulation.

6.2 Energetics and Numerical Simulation of Ocean Circulation

Since we do not well know the distribution of mechanical energy available for mixing, many oceanic general circulation models of the current generation are based on rather simple parameterization of subgrid mixing. A common practice in oceanic general circulation modeling is to choose a diapycnal mixing rate and let the models run; the diapycnal mixing rate remains the same all the time, even under different climatic conditions. In fact, no modelers have ever checked how much energy is required to support mixing and circulation and, most important, whether this amount of energy is really available.

An interesting and important question is how to parameterize mixing in climate-related models. Should we use the same mixing rate or same mixing energy when the climate is drifting? Clearly, both a fixed mixing rate and a fixed mixing energy represent the two extremes, and the reality is likely to be between these two assumptions.

It is speculated that when we have the new generation of the Oceanic General Circulation Model (OGCM), in which the mixing parameterization is more rational, climate variability induced by something like the CO_2 doubling may be different

from whatever we have learned based on existing models.

7. CONCLUSION

In summary, energetics of oceanic circulation is of capital importance in understanding ocean circulation and climate. Energetics of the oceanic circulation is one the most important research frontiers. There are many unanswered fundamental questions, and at this time, we do not have the lowest-order balance. Most flux terms cited here should be treated as preliminary estimates, which will certainly be improved in the near future. Since circulation in the oceans is primarily controlled by energy available for mixing, further study of energetics will bring about major breakthroughs in our understanding of the ocean circulation and climate.

SEE ALSO THE FOLLOWING ARTICLES

Aquaculture and Energy Use • Conservation of Energy Concept, History of • Desalination and Energy Use • Earth's Energy Balance • Fisheries and Energy Use • Lithosphere, Energy Flows in • Ocean Thermal Energy

Further Reading

Bryden, H. L., and Imawaki, S. (2001). Ocean heat transport. *In* "Ocean Circulation and Climate" (G. Siedler, J. Church, and J. Gould, Eds.), International Geophysical Series, pp. 455–474. Academic Press, New York.

Garrett, C., and St. Laurent, L. (2002). Aspects of ocean mixing. *J. Oceanogr. Soc. Japan* **58,** 11–24.

Gill, A. E., Green, J. S. A., and Simmons, A. J. (1974). Energy partition in the large-scale ocean circulation and the production of mid-ocean eddies. *Deep-Sea Research* **21,** 499–528.

Huang, R. X. (1998). Mixing and available potential energy in a Boussinesq ocean. *J. Phys. Oceanogr.* **28,** 669–678.

Huang, R. X. (1999). Mixing and energetics of the thermohaline circulation. *J. Phys. Oceanogr.* **29,** 727–746.

Kagan, B. A., and Sundermann, J. (1996). Dissipation of tidal energy, paleotides, and evolution of the earth-moon system. *Adv. Geophysics* **38,** 179–266.

Munk, W. H., and Wunsch, C. (1998). The moon and mixing: Abyssal recipes II. *Deep Sea Research I* **45,** 1977–2010.

Trenberth, K. E., and Caron, J. M. (2001). Estimates of meridional atmosphere and ocean heat transport. *J. Climate* **14,** 3433–3443.

Wunsch, C. (1998). The work done by the wind on the oceanic general circulation. *J. Phys. Oceanogr.* **28,** 2331–2339.

Wunsch, C. (2001). Ocean observations and the climate forecast problem. *In* "Meteorology at the Millennium" (R. Pearce, Ed.), Royal Meteorological Society, pp. 233–245. Academic Press, New York.

Ocean Thermal Energy

DON E. LENNARD
Ocean Thermal Energy Conversion Systems Ltd.
Kent, United Kingdom

1. **Definition and Basis of OTEC**
2. **The Beginnings of OTEC**
3. **Uses of OTEC Power—and Other Products**
4. **Variations on OTEC Plants and Factors Affecting Them**
5. **Technology of OTEC Plants**
6. **Some OTEC Projects and Designs**
7. **Cost, Economy, and Competitiveness**
8. **Markets and the Way Ahead**

Glossary

closed cycle When the working fluid of the heat engine cycle is fully contained and isolated, changing from liquid to gaseous phases and back again.

deep ocean water applications (DOWAs) Alternative products and processes that can be derived from an ocean thermal energy conversion plant.

exclusive economic zone (EEZ) The area/volume of sea/sea bed, and its resources, that can be claimed by a nation under the UN Convention on the Law of the Sea.

International OTEC/DOWA Association (IOA) The worldwide grouping for all aspects of ocean thermal energy conversion and deep ocean water applications.

ocean thermal energy conversion (OTEC) The technique of extracting energy from the vertical temperature difference in the oceans.

open cycle When seawater is the working fluid, is used in free flow, and is not recirculated.

The surfaces of the oceans capture huge amounts of solar energy—some thousands of times more than the energy consumed by our current civilization—with most of this being stored in the form of thermal energy in the surface layers of the oceans. Those surface layers do not mix freely or easily with the deeper waters, which are much colder. Ocean thermal energy conversion (OTEC) is based on the extraction of energy from the temperature difference existing between those warm surface waters of the oceans in extensive tropical and subtropical areas and the deep waters in those same areas that flow from the polar regions.

1. DEFINITION AND BASIS OF OTEC

Ocean thermal energy conversion (OTEC) is based on the extraction of energy from the temperature difference existing between those warm surface waters of the oceans in extensive tropical and subtropical areas (approximately latitudes 25° north to 25° south) and the deep waters in those same areas that flow from the polar regions (predominantly the Antarctic). Surface temperatures can reach as high as 29°C in the Pacific, whereas the temperature at a depth of 1000 m will typically be 4°C, although this temperature can sometimes be found at depths as shallow as 700 m. Figure 1 shows the general scale and distribution of this thermal resource. The typical design case for an OTEC plant assumes a temperature difference of 20°C.

2. THE BEGINNINGS OF OTEC

Ocean thermal energy conversion (OTEC) has existed for more than 120 years, but its technological status now owes very much to developments in offshore activities associated with the oil and gas industries during the latter part of the 20th century—more particularly as exploitation of those fossil fuel resources has moved into deeper waters and harsher climates.

The OTEC power circuit is that of a standard heat engine cycle, but where the temperature difference is much less than that in the case of a steam engine or an internal combustion engine. It was in 1881 that the Frenchman Arsène d'Arsonval suggested using

 © 2004 Elsevier Inc. All rights reserved.

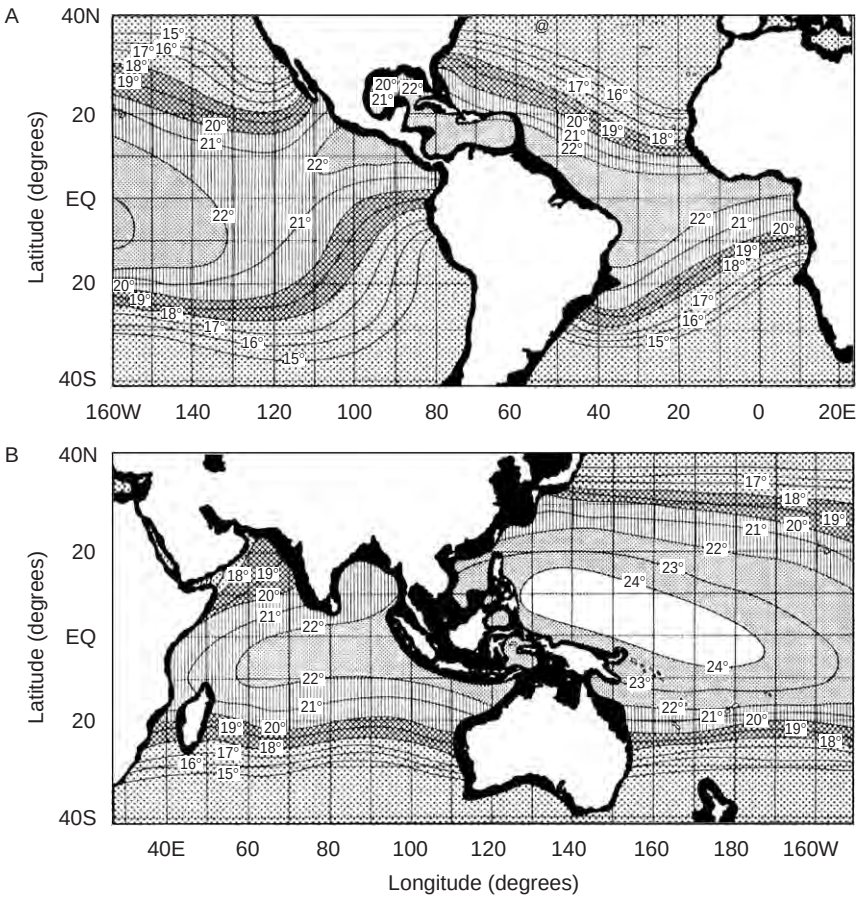


FIGURE 1 The ocean thermal resource: Worldwide distribution. Average temperature differences (in °C) between the surface and a depth of 1000 m. From U.S. Department of Energy/Ocean Data Systems Inc.

any working fluid having an appropriate vapor pressure at a temperature close to that of warm seawater. After cooling in a surface condenser, the liquid working fluid is circulated by pumping to the evaporator, where the fluid changes to a gaseous phase with a considerable increase in volume and pressure, passes through a turbine connected to an electrical generator, and is then recycled on through the condenser. This concept, known as closed cycle OTEC, is shown in Fig. 2. Working fluids, including ammonia, propane, butane, and freon, are suitable. Closed cycle OTEC requires the design and construction of large surface heat exchangers that are close to—some say beyond—the limits of currently available technology to enable the construction of modular closed cycle OTEC plants if the capacity is in excess of approximately 100 MW.

A second option was proposed by another Frenchman, George Claude, during the late 1920s. In Claude's OTEC concept, the working fluid is the vapor formed by the warm seawater itself when boiled in an evaporator maintained at an appropriate low pressure. Water vapor is condensed either through a direct contact condenser by mixing with cold seawater or directly through a surface condenser. In the latter case, desalinated water is a by-product of the thermal process. In neither case is the condensed vapor reintroduced to the working fluid circuit. Claude's concept, known as open cycle OTEC, is shown in Fig. 3. The main technical difficulty for open cycle OTEC plants comes from the low vapor pressure of the working fluid; this requires very large turbines. Available technology suggests that the construction of modular open cycle OTEC plants with capacities of tens of megawatts is currently possible.

3. USES OF OTEC POWER—AND OTHER PRODUCTS

There are significant energy requirements in the tropical and subtropical areas where OTEC is most

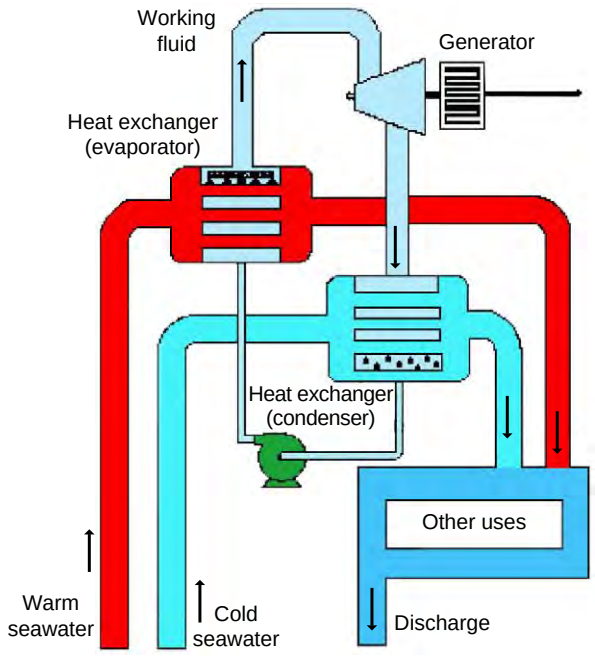


FIGURE 2 Closed cycle OTEC: Power circuit. From National Energy Laboratory of Hawaii Authority.

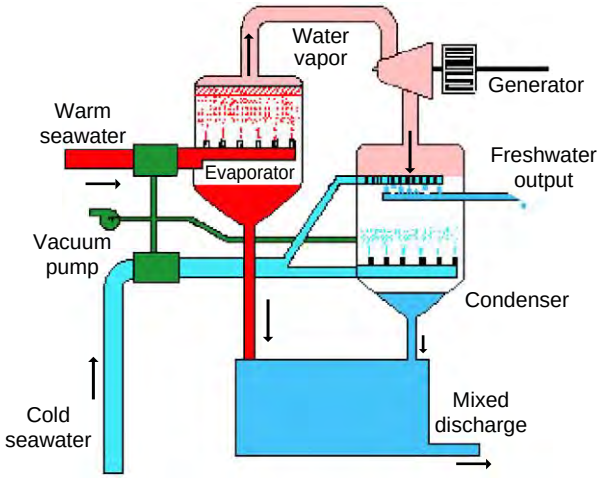


FIGURE 3 Open cycle OTEC: Power circuit. From National Energy Laboratory of Hawaii Authority.

attractive, but there also are a number of possible "by-products" from OTEC, the benefits from which are also considerable. Usually referred to as deep ocean water applications (DOWAs), examples of these include aquaculture and the production of potable water (Fig. 4 shows a flow chart for such a combination). Potable water can be an immediate product of open cycle OTEC where evaporated warm seawater is condensed, whereas in the case of a closed cycle OTEC plant, low-temperature distillation/multistage flash techniques can be used. Aquaculture and potable water production fit in well with the large quantities of both warm and cold water moved by an OTEC plant and can be incorporated separately or together. In the limiting case, all of the deliverable power can be used to drive pumps and other equipment for the desalination and aquaculture activities. The cold deep water for the OTEC plant is free of pathogens and rich in nutrients, and phytoplankton can convert the nutrient organic materials on which shellfish can then feed and grow. In addition, other fish and seaweed will grow at greatly enhanced rates. The cold water is also a source of pharmaceutical activities and products.

A further DOWA is use of the cold water, still cool after passing through the condenser, for agricultural purposes by passing the cool water through pipes buried in the soil, thereby enabling products of temperate climates, such as vegetables and tomatoes, to be grown in tropical zones—an application already occurring.

Other DOWAs for OTEC make use of the electrical power *in situ* to achieve, alternatively or additionally, air conditioning or energy-intensive products. For example, hydrogen may be produced by electrolyzing water, and that version of OTEC power can then be stored and transported to market—usually to industrialized nations outside the OTEC zones—as liquid hydrogen. Alternatively again, hydrogen may be produced as an intermediate product, being used in turn to produce ammonia. Currently, the use of ammonia fertilizers is determined in part by production capacity from natural gas. The use of such fertilizers in the developing world, much of it in the tropical and subtropical zones where OTEC processes are available, could make a major contribution to world food production.

Finally, mention should be made of the production of some metals from their ores—aluminum from bauxite is a good example—that requires considerable power input. The location of an OTEC plant onshore by the bauxite deposits, or transportation of the deposits to an offshore plant with refined metal being shipped back on the return journey, are other, probably longer term options for application of OTEC power.

Although the economic viability of these other processes varies during the first years of the 21st century, with some being economically attractive and others being less so, this range of possible uses of OTEC power gives flexibility to the ultimate uses of that power and gives rise to a diversified market. It is

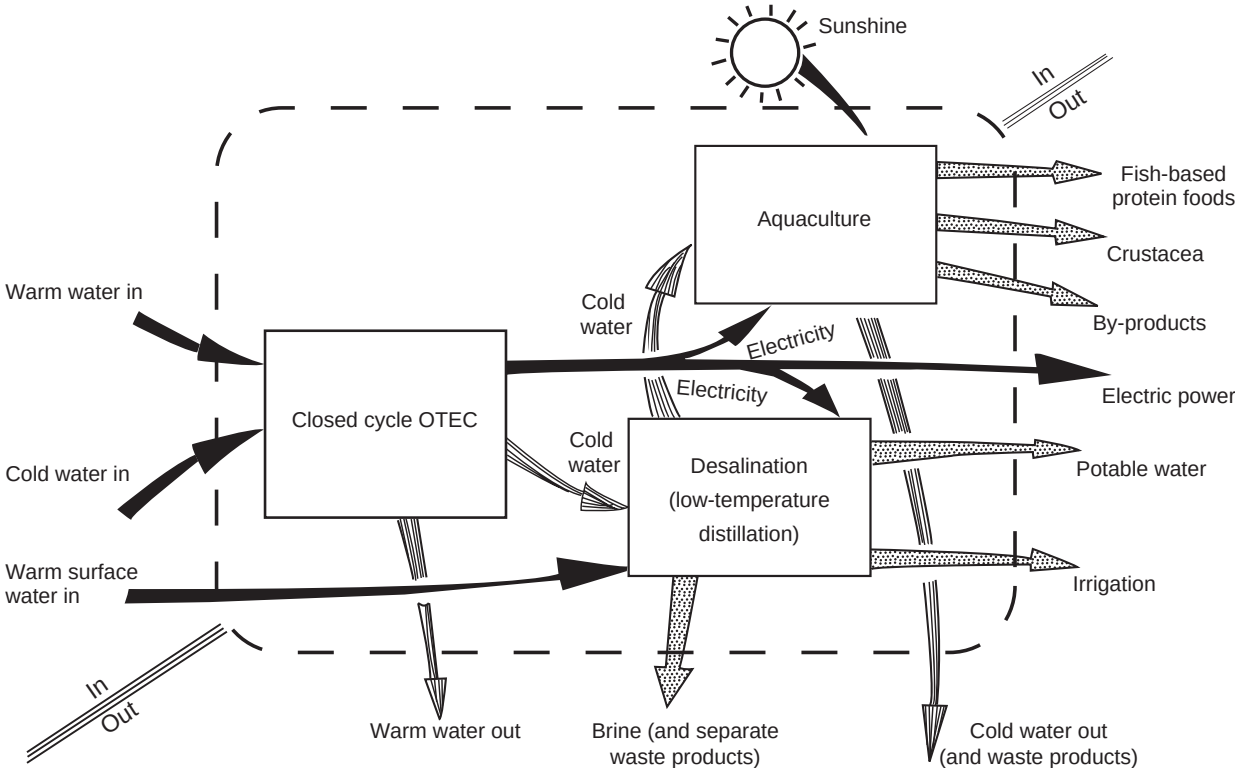


FIGURE 4 OTEC deep ocean water applications (DOWAs): Flow diagram of electricity, aquaculture, and potable water outputs.

this variety of markets, coupled with the technological developments of the offshore oil and gas activities, that encouraged a number of countries to become involved in OTEC developments during the final quarter of the 20th century, including Belgium, China, Finland, France, India, Indonesia, Ivory Coast, Jamaica, The Netherlands, Norway, Sweden, Taiwan, the United Kingdom, the United States, and the Soviet Union. Some of these developments are described later.

4. VARIATIONS ON OTEC PLANTS AND FACTORS AFFECTING THEM

Although based on well-established heat engine cycles, the OTEC energy extraction process has an overall practical efficiency of some 2.5% due to the relatively small operating temperature differential, which has to be compared with a figure of typically 30% or more for a fossil fuel power station. To obtain the 2.5% efficiency, the standard design case temperature difference of some 20°C is required, but clearly the larger the temperature difference, the greater the efficiency of the plant. An increase of temperature difference of 1°C results in an increase in efficiency of approximately 10%. Because the cold water is at depths of up to 1000 m, and because the flow of water required per megawatt of power output is in the range of 4 to 8 m^3/s, the scale of the plant will be apparent. Therefore, practical realization—engineering of the system—is a key issue in considering the future of OTEC and is a major element in its economics.

Although the thermal resource is particularly relevant to developing countries, there are many other factors to be considered before it can be said that a particular country or location is suitable for an OTEC installation. These include distance from shore to the thermal resource, depth of the ocean bed, depth of the resource, size of the thermal resource within the exclusive economic zone (EEZ), replenishment capability for both warm and cold water, currents, waves, hurricanes, sea bed conditions for anchoring, sea bed conditions for power cables of floating plants, present installed power and source, installed power per head, annual consumption, annual consumption per head, present cost per unit (including any subsidy), local oil, gas, and/or

coal production, scope for other renewables, aquaculture potential, potable water potential, and environmental impact.

Given satisfactory figures from these assessments, the forms of OTEC plant to be used, including combinations with other relevant DOWAs, are varied and include the following:

- Closed cycle and open cycle
- Floating but constrained to maintain station
- Land based
- Shelf/Tower based

Within the floating concept, a number of options are available, including the following:

- Ship-shape hull
- Semi-submersible
- Submersible
- Guyed tower
- "Other options" being regularly developed in the offshore oil and gas sector

And then, there is the question of whether the OTEC plant is for one or more of the following:

- Generating electricity
- Aquaculture
- Desalination processes
- Other DOWAs

In the longer term, there can be "grazing" OTEC plants—plants that move freely with currents or under their own power—for use as a source of power to the following:

- Recover and process deep ocean minerals
- Produce liquid hydrogen, liquid ammonia, and the like for trans-shipment to other areas to be used in industrial processes and power generation
- Process at sea ores and other raw materials that are energy intensive in the extraction of the refined product

Such plants will be dependent for their operation on a broad agreement to, and application of, the Law of the Sea Convention.

It can be seen that the options for OTEC plants are considerable and that optimization for a particular site involves the review of many variables.

5. TECHNOLOGY OF OTEC PLANTS

Major parts of an OTEC plant involve the use of well-developed materials and equipment, many of these having also been tried, tested, and found to be satisfactory in the type of marine environment where OTEC plants will operate, with predictable performance and economy. Because many of the components of an OTEC plant, whether land based or floating, are well-established technology, OTEC has an advantage over some of the other renewable energies. But there are exceptions. Four specific items—cold water pipe, heat exchangers, moorings, and power transmission to shore—have been the subjects of extensive research and development (R&D) programs to determine in detail the technical appropriateness and economy of preferred solutions. However, the availability of new materials and powerful analytical techniques are considerably easing remaining problems; and there is substantial scope for technology transfer from other industries, including not only offshore oil and gas but also aerospace, among others.

For the cold water pipe, the quantities of water involved and their momentum are substantial. To keep loads in the seawater circuit at acceptable levels, it is generally accepted that the velocity in the pipe should not exceed approximately 1.5 m/s, although some design studies allow for a figure as high as 4.0 m/s. The pipe must certainly be capable of repair during its working life, and because a number of potential sites for OTEC are in geographical areas subject to hurricanes, it should also have the ability to be structurally disconnected from the main OTEC hull on fairly short notice where a floating design is under consideration. The combination of these factors tends toward a pipe design that is not monolithic but rather has the ability to be sectioned.

Over the pipe length of 1000 m (or more for a land-based version), subsurface currents will have different magnitudes and directions. Therefore, calculations of loading will be dependent on wind, sea state and current conditions, and extensive analysis will be required to determine the critical load conditions. Much work on this has already been done, and representative practical tests have also been carried out in Japan and the United States.

For the heat exchangers, operational efficiency—of both the evaporator and the condenser—is fundamental to the economic success of the OTEC plant. Achievement of the plant's operating efficiency of approximately 2.5% is highly dependent on the heat transfer rate in both the evaporator and the condenser. Also, at the low operating temperature differences in these items, the maintenance of optimum efficiency is dependent on minimizing biofouling and corrosion products on the surfaces of the heat exchanger materials. Because of the small

temperature difference, the heat exchanger surfaces must be very large to achieve practical power outputs, although for a 10-MW size the 50,000-m^2 surface area required is no bigger than the largest currently produced, particularly if modular units are used. Full commercial sizes of OTEC power plants, typically 25 to 40 MW for floating plants, will require correspondingly larger heat exchanger surfaces. Altogether, these requirements result in the heat exchangers being very expensive items. Some estimates place approximately 40% of the total cost for early OTEC plants against the heat exchanger system, whereas others estimate less than 25%. With a titanium–aluminum sandwich, estimated costs are lower, but this material combination is not yet available in a fully commercialized form. France, Canada, the United States, and the United Kingdom have undertaken extensive development programs for OTEC heat exchanger applications.

In the case of moorings, the floating variant of an OTEC plant, delivering electrical power to shore, must maintain station to provide a safe and secure path for the electrical power cable. Position fixing can be by mechanical means (e.g., cables, chains), by dynamic positioning (e.g., thrusters), or perhaps (in the future) by use of hydrodynamic forces.

It is again environmental forces from wind, current (which will vary in magnitude and direction with depth), and wave drift (the part of the wave spectra's momentum transferred to the floating plant) that give rise to substantial design problems. Hurricanes are not infrequent at many potential OTEC locations, and environmental loading must include their effect in terms of both wind and wave drift forces. Wave drift forces are often the major loading on conventional large floating structures, and there is a large range in the values of each of the forces when calculated from current theories. It follows that, with variation in calculated values, considerable work must be undertaken if sensible design values are to be obtained. The inefficiency of design that would follow from overestimates of environmental loads would have a very undesirable effect on the economy of an OTEC plant.

Although it is clear that specific values of mooring loads can be calculated only once the mooring system has been designed and vice versa, two related general points about the mooring can be made. First, it should locate the plant within a prespecified watch circle. In turn, the watch circle should be determined, at its maximum, by the strength and limit to freedom of response of the electric power cable and, at its minimum, by the limitations on load that can be applied by the mooring system to the OTEC plant (e.g., hull, cold water pipe).

Existing "deep water" moorings make use of wire rope or chain, and the historical knowledge of these materials has made them strong contenders for OTEC application, although the weight of both wire rope and chain becomes excessive. The use of synthetic fiber ropes has been greatly extended during recent years for conventional marine applications, and these ropes are now strong contenders.

Finally, for power transmission, it should be noted that although the electrical generating system for an OTEC plant is essentially "state of the art," the particular environmental location and corresponding conditions do introduce constraints. The transmission for floating plants requires developments, to cater particularly to the riser section from sea bed to plant as well as to the sea bed cable running from the riser to shore and the grid, although the growth in offshore wind farms, albeit in shallower water, is providing relevant operating experience.

Apart from the required mechanical strength of cable for this application, its prime function is, of course, to transmit electrical power at a high efficiency from source to grid, and the distance offshore determines whether transmission is by direct or alternating current. In terms of efficiency alone, the alternating current (AC) option is preferred for shorter transmission distances (up to $\sim$30 km), with losses at approximately 0.5% of input power per kilometer. For greater distances, the direct current (DC) is preferred, with losses much less at approximately 0.01%/km but with the additional inversion and conversion losses of approximately 2%. Distances as great as 400 km have been suggested as possible with the use of DC, but the loss of efficiency, as well as the cost of manufacture and installation of the cable, eliminates much if not all of the benefit of an OTEC installation at those distances.

For OTEC applications, with design depths initially specified at 2000 m, cable development itself is not anticipated to present problems, although long-term resistance to pressure at that depth will require validation. Laying of the cable is not expected to present insuperable difficulties either. The problem that will require particular attention is cable junctions at depth, including the junction with the riser cable. In addition, protection and repair procedures for the cable in shallower waters, where it is brought ashore and where anchors, trawl boards, and other hazards (natural as well as those connected with human activity), will present a greater variety of

problems than those in the deeper waters. Again, offshore wind experience is providing guidance.

6. SOME OTEC PROJECTS AND DESIGNS

There have been a number of designs, tests of components, and operations of small demonstration plants over the past 50 years or so, and some of these are briefly described in this section.

During the 1950s, France designed open cycle OTEC plants for two of its then overseas territories: Ivory Coast and Guadeloupe. During the late 1970s and on into the 1980s, designs for a 5-MW plant in Tahiti were completed, tests were undertaken on heat exchangers, and a range of cold water pipe designs were studied.

In the United States, also during the late 1970s, "mini-OTEC" was constructed on a moored barge and operated in Hawaiian waters with a 670-m cold water pipe. This successfully demonstrated the principles of OTEC, generating gross power of a little less than 100 kW. This was followed up with "OTEC-1," operating on a converted tanker, specifically to check out the effectiveness of heat exchangers but also enabling cold water pipe and mooring details to be developed. Since then, Hawaii has maintained an active interest in OTEC, principally through the Natural Energy Laboratory of Hawaii Authority and the Pacific International Center for High-Technology Research, and during the 1990s it built and operated a land-based, 255-kW gross open cycle OTEC plant that produced net power. It is also in Hawaii that the majority of the practical DOWA activities have been developed, including pharmaceuticals, many on a fully commercial basis.

Japanese interest in OTEC developed through the 1970s, culminating in a land-based, 100-kW gross closed cycle OTEC plant on the island nation of Nauru in the Pacific, located very close to the equator with an excellent thermal resource, and a net output of more than 30 kW was achieved. Among other Japanese programs with industrial involvement, Saga University has played a major and ongoing role. At the turn of the 21st century, it was involved with the Indian proposal for a 1-MW floating OTEC plant that, after testing, is due for installation in an Indian Ocean island group.

European industry, through the international R&D organization Eurocean, completed a design for an OTEC/desalination/aquaculture plant during the 1980s, and at the same time The Netherlands completed a 250-kW design for installation in the Indonesian island of Bali.

Also during the 1980s, a U.K. combined industry–academic team design was proposed for a 10-MW floating closed cycle OTEC plant for island applications, including hydrographical surveys for the preferred location in the Caribbean where cold deep water comes close to the shore (Fig. 5).

During the 1990s, the European Union interest in OTEC and DOWAs took the form of a review of all options, bringing OTEC experts to Brussels, Belgium, from around the world. Although Europe itself has no home market for OTEC given that there is no

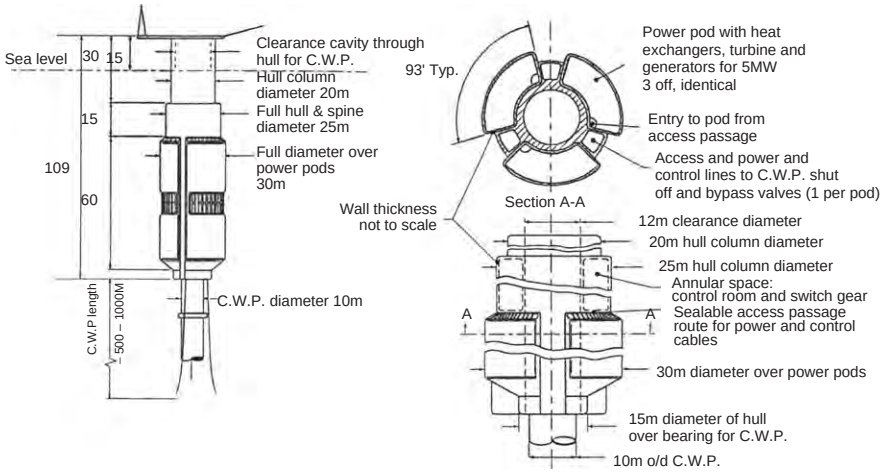


FIGURE 5 General arrangement of 10-MW closed cycle floating OTEC plant. All dimensions are in meters. From Ocean Thermal Energy Conversion Systems Ltd.

adequate thermal resource, there are a number of national dependencies, as well as many (mainly island) countries receiving aid from the European Union through the Lomé and Cotonou agreements, that are located where the OTEC resources are very strong.

Taiwan has a rapidly growing energy demand, increasing up to 10% per year, and for nearly a quarter of a century has developed ideas for OTEC, particularly for application off its east coast where plans for a 5-MW plant exist for a number of sites. During the mid-1990s, a 20-year master OTEC plan for Taiwan was produced, envisaging a 100-MW demonstration plant being developed, with a challenging total OTEC capacity for the country of 3200 MW in 20 years.

Information on all current developments in OTEC and DOWAs is available through the International OTEC/DOWA Association (IOA), which operates out of Taiwan and has a worldwide membership that includes most people involved in OTEC/DOWA activities.

7. COST, ECONOMY, AND COMPETITIVENESS

Because of the diffuse nature of most renewable energies, the size of these renewable energy generation systems is large when compared with fossil fuel generators, and the capital costs of the former are proportionately large as a result. Therefore, on the basis of capital cost alone, renewables (including OTEC) compare poorly with oil-fired power generation. Oil-fired plants would typically cost a few hundred dollars per installed kilowatt, whereas OTEC would cost a few thousand dollars per installed kilowatt. The case for cost "comparability" is crucial to the acceptability of OTEC. Fuel oil has a substantial price, whereas OTEC fuel is free. Maintenance costs for well-developed oil-fired plants are low, whereas maintenance costs for low-efficiency OTEC plants are relatively high. In addition, the high capital cost of OTEC means high total interest charges to be serviced in comparison with the interest charges for oil-fired plants.

All of these points must be incorporated into the financial calculation if an accurate and realistic cost comparison is to be made. Also, and almost uniquely, OTEC is a base load system. Output varies little whether the sun shines or it is nighttime because the thermal resource of the oceans is so massive and the change between summer and winter is typically no more than 10%. Therefore, OTEC has advantages over other renewables that require buffer storage to achieve base load, with commensurate added costs.

A review of past national and international programs of work shows that a variety of assumptions were made, resulting in overlap between, for example, costs for floating OTEC plants and those for land-based ones, plants with closed working cycles and those with open ones, and simple designs using established technology and technologically advanced designs. Beyond confirming a reduction in costs with size, there is little in the way of conclusions to draw from these comparisons.

The following figures are based on the design of a 10-MW demonstration plant, shown as a general arrangement in Fig. 5. As a demonstrator, it has three 5-MW power pods, with the third pod for development or use if either of the two main power production pods has to be shut down at any time. The floating plant consists of a single cylindrical hull in concrete, heat exchangers in plate form constructed from a titanium–aluminium sandwich, a cold water pipe of 1000 m length in fiber-reinforced plastic, moorings in wire and chain, and transmission to shore over a distance of 10 km, resulting in a capital cost figure of $94 million or $9400/kW. The percentage costs of major components are shown in Table I.

The calculations for this 10-MW floating closed cycle OTEC plant, designed for a Caribbean or South Pacific island where the temperature difference varies between 23 and 21°C between summer and winter, show a generating cost of $0.18/kWh using the 21°C temperature difference. If potable water is a by-product (highly desirable for both of these locations), the generating cost falls to $0.11/kWh with the potable water costed at $0.80/m^3. Island costs for

TABLE I

Component Costs for 10-MW Floating OTEC Plant (Percentages)

Site-specific data	2
Heat exchangers	23
Cold water pipe	6
Moorings	5
Electrical transmission (sea bed and riser)	8
Pumps, turbines, generators, and control	13
Hull (including warm water circuit)	20
Installation and maintenance	5
Start-up and test	8
Contingency	10

potable water are in the range of \$0.40 to \$1.60/m^3, so the figure used here is seen to be at the conservative end of the spectrum. Although there are islands where land fuel costs are three to four times those in developed countries, for the two islands considered here, the uplift is only 75%. So, instead of \$20/barrel, oil would be \$35/barrel. At that price, electrical generation from an oil-fueled plant would be costed at \$0.09/kWh. Therefore, the OTEC demonstrator plant is seen to be approaching cost competition with such a plant on these islands. General engineering experience suggests that by the 8th to 10th production plant, capital costs will be reduced and generating costs may be expected to fall by some 35%. If these reductions are achieved, OTEC will become competitive with oil-fired plants for many island locations. In these examples, no benefit has been given to OTEC for the "environmental plus" that it has in common with many other renewables. One recommendation of the 1992 Earth Summit in Rio de Janeiro, Brazil, was the introduction of a carbon tax for fossil fuels, but to date this has not been applied. If such a tax is brought into use, as may well be the case during the first decade of the new century, all renewables, including OTEC, will benefit further in terms of competitiveness with hydrocarbons.

In addition to these simple costing figures, it is essential that an operator or utility also sees this new technology as attractive. For the example given here, a notional return of 20.4%, corresponding to a real return of 14.7%, is calculated, and these are reasonably attractive returns. So, for both the consumer and the plant operator, OTEC is beginning to look attractive. But history shows that new technology has considerable difficulty in attracting financing for the first examples, and OTEC is unlikely to be an exception to this. Therefore, it seems essential that the first two or three plants of realistic size—approximately 10 MW if of the floating variety—will need to be funded by governments or international funding agencies such as the World Bank, the Asian Development Bank, or the European Bank for Reconstruction and Development. After that, and if the figures cited here are achieved, the financial business sector may be expected to step in and support further installations.

8. MARKETS AND THE WAY AHEAD

All surveys indicate that the world demand for energy will continue to increase considerably, and the greater part of that increase is expected in developing countries. At the same time, it is anticipated that the percentage of "new" energies will also grow, from a near zero figure at the end of the 20th century to 6% by 2020. This translates into new energies of some 12,000 MW per year averaged over the period from 2000 to 2020. Capital costs for natural temperature difference equipment, due to the low efficiency, is on the order of \$5,000 to \$10,000/kW, some 10 times the capital cost for conventional power systems. Therefore, the funding of new energies equates to a total sum of \$60 billion to \$120 billion each year—very substantial business by any standards. Therefore, an important point to note is that this is, for the construction, operational, and financing sectors, an activity of very considerable interest. But this business will develop only if it is economically attractive to the utilities that will invest in and operate the OTEC plants. As indicated earlier, that situation is rapidly approaching.

Of course, regardless of the feasibility of OTEC-generated electricity, there must be a demand for it at a price it can meet or better. Superposition of a standard atlas onto the contours of Fig. 1 will illustrate which countries can benefit most from the OTEC resource, and it is apparent that, with significant exceptions, the great majority are developing nations rather than developed ones and that many of these are rather small islands (Table II). For this reason, many early OTEC plants will be small—typically less than 10 MW rather than on the order of 50 MW. The majority of island countries that do have adequate demand currently rely on imported oil for generating power and are paying high prices for this oil due to delivery charges. Therefore, in the shorter term, it is islands that provide the main market for OTEC plants, particularly those islands

TABLE II

Island Nations with High OTEC Resources

Bahamas
Cayman Islands
Fiji
Guam
Jamaica
Pacific Islands (Trust Territories)
Palau
Papua New Guinea
Seychelles
St. Lucia
Trinidad and Tobago

that have a growing power demand. Although the rate of growth of power demand in many independent developing countries should be dependent on the price of the power generated, surveys indicate that the greatest increase in power demand is in those developing nations, where incidentally potable water is also a high priority—an ideal combination for OTEC. Table II lists island nations where OTEC has particular application, and one of these intends to meet most of its power requirements—30 MW—from OTEC by 2015.

The growing acceptance of the 200-mile EEZ is also particularly relevant to oceanic island nations, although the benefit of that is likely to relate more to grazing OTEC plants that seem unlikely to materialize during the first quarter of the new century. Nevertheless, this potential application of OTEC is of considerable importance to island nations because it could be a substantial tool in enabling them to benefit from the resources within the EEZ.

The key to further progress, now that so much relevant R&D has been completed, is the construction of representative scale demonstration plants. An examination of island sites has indicated that floating OTEC plants of 10 MW (and smaller land-based plants) can produce electricity at prices that will be competitive with those of oil-fired generating systems for the best situations. However, the fully economic sizes for general application in the tropics and subtropics will be 25 to 40 MW for floating plants and 8 to 10 MW for land-based plants.

The way ahead is for the construction of demonstration-sized plants in one or more island nations, financed through international funding agencies, to achieve cost-competitive, environmentally friendly base load power—and, probably, potable water as well.

SEE ALSO THE FOLLOWING ARTICLES

Desalination and Energy Use • Geothermal Direct Use • Geothermal Power Generation • Ground-Source Heat Pumps • Lithosphere, Energy Flows in • Ocean, Energy Flows in • Thermal Energy Storage

Further Reading

Avery, W., and Wu, C. (1994). "Renewable Energy from the Ocean: A Guide to OTEC." Oxford University Press, Oxford, UK.

Energy Commission. (1994). "Master OTEC Plan for the Republic of China." Ministry of Economic Affairs, Republic of China.

Gauthier, M., Golmen, L., and Lennard, D. (2000). Ocean thermal energy conversion (OTEC) and deep ocean water applications (DOWA). *In* "New and Renewable Technologies for Sustainable Development" Conference Proceedings. European Community, Madeira Island, Portugal.

Lennard, D. (2001). Ocean thermal energy conversion. *In* "Survey of Energy Resources." World Energy Council, London.

Vadus, J. (1997). A strategy for OTEC commercialization. *In* "Proceedings of the 1997 International OTEC/DOWA Association," pp. 235–247. IOA, Singapore.

Vega, L. (1994). Economics of ocean thermal energy conversion. *In* "Proceedings of the 1994 International OTEC/DOWA Association." IOA, Brighton, UK.

Oil and Natural Gas Drilling

J. J. AZAR
University of Tulsa
Tulsa, Oklahoma, United States

1. **The Rotary Drilling Process**
2. **Purpose of Drilling**
3. **Rotary Drilling Requirements**
4. **Rig Systems**
5. **Drilling Fluids and Hydraulics**
6. **Drill Bits**
7. **Drill String**
8. **Horizontal and Multilateral Well Drilling**
9. **High-Angle/Extended Reach Drilling**

Glossary

annulus The space between the outer wall of the drill string and the wall of the well bore; it provides a passageway for fluids and drilled rock cuttings to return to the surface.
blowout When a sudden intrusion of formation fluids into the well bore occurs and the preparation to prevent or control the kick fluid has failed.
development well drilling Drilling in an area where wells have been drilled before and geological information has been established.
draw-work Part of the rotary rig hoisting system that provides the power to move loads in and out of the wellbore.
drill bit A tool specially designed to efficiently make a hole in a rock formation by shearing, crushing, or shoveling.
drill string Consists of all the tubular connected by tool joints.
drilling fluid A liquid, gas, or gasified liquid that is used in drilling to perform several functions such as cuttings removal, cooling, and stabilization of well bore.
drilling hydraulics A branch of engineering that describes drilling fluids behavior while in motion.
exploration drilling Drilling in an area where the geological conditions are not known.
high-angle/extended reach drilling (HA/ERD) When the horizontal departure over the true vertical depth of the well is greater than 2.0 and the angle of inclination is more than 63° from vertical.
horizontal drilling Consists of drilling horizontally (approximately 90° from vertical) into the reservoir rock.
horsepower The rate of doing work or transferring energy that is equivalent to moving 33,000 pounds 1 foot per minute.
kick When formation fluid enters into a hole unintentionally; the fluid is what is called kick fluid.
marine riser Large diameter steel casing with buoy sections that serves as an extension of a well bore from sea bottom to rig surface area when drilling offshore.
multilateral drilling Consists of drilling several laterals from a main trunk drilled into a reservoir; the trunk may be vertical, horizontal, or directional.
rotary drilling A process of boring a hole into the ground by applying a force and right-hand rotation on a drill bit and circulating a drilling fluid from surface through a drill string and back to surface through the annulus.

Digging or boring holes into the earth subsurface in search of water and food dates back to prehistoric times. The tools used were very primitive in nature and included stones, sticks, and bones. Published records on hole digging in search of brine and possibly oil have been reported as early as 1500 BC. The term "drilling," which during the early days simply meant boring a hole into the earth subsurface, today is defined as an elaborate "well construction" (Fig. 1). Around 1800, the art of drilling and its tools were developed to some extent, and it is known that by this time drilling for oil had been conducted in Japan, France, and Burma as well as other places around the world. Although oil was first discovered in America in 1814, the first well purposely drilled for commercial oil was the Drake well of 1859 (depth of 59 feet and estimated production of 20 to 30 standard barrels/day). While rotary drilling was being developed in other countries (mainly France), cable tool drilling (percussion-type tools) was being developed in the United States prior to the 1900s. Cable tool rigs were used to drill the most oil wells in the world by far until the 1920s. In the Corsicana oilfield in Texas, discovered during the 1890s, cable tool rigs were not successful in drilling the unconsolidated

 © 2004 Elsevier Inc. All rights reserved.

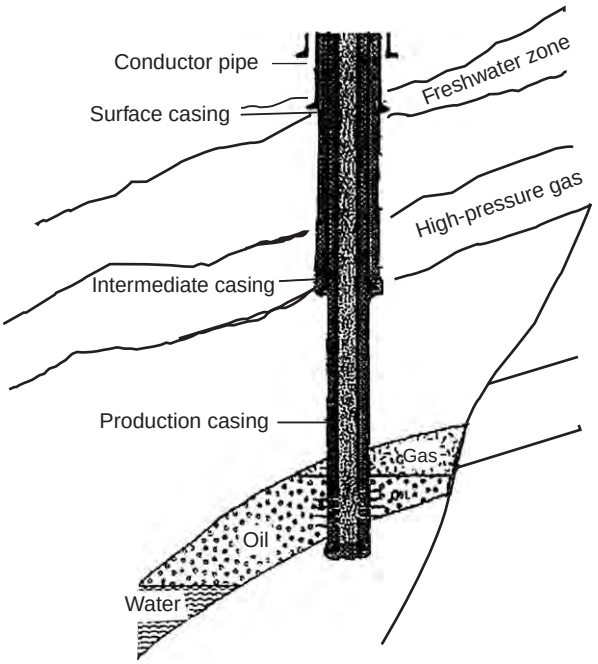


FIGURE 1 Well construction.

formations, thereby bringing wet rotary drilling rigs into the region. In 1901, the famous Lucas Spindletop well was drilled using a rotary rig system and was completed at an estimated oil production of 75,000 to 80,000 standard barrels/day at a total depth of approximately 1100 feet. This is the famous well that gave birth to the oil and gas industry, as we knowit today, and has had the greatest impact on the economy and culture of the entire world since then. There have been many technological advances in drilling over the years. The most striking advances, which have occurred during the past decade or so, allowed the oil and gas industry to reach targets below and away from a surface location in excess of 38,000 feet (12 km) and moving farther

1. THE ROTARY DRILLING PROCESS

Whether drilling vertical, directional, horizontal, or multilateral wells (Fig. 2) for the exploitation of hydrocarbons, the elements needed to drill wells successfully and economically are the same but differ in requirements to achieve them). These elements, as shown in Fig. 3, include the following:

- Force acting downward on a rock-cutting tool called a drill bit

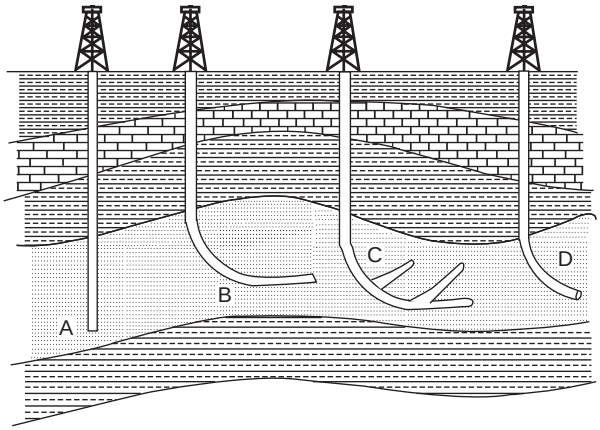


FIGURE 2 Well types: (A) vertical, (B) horizontal, (C) multilateral, and (D) directional.

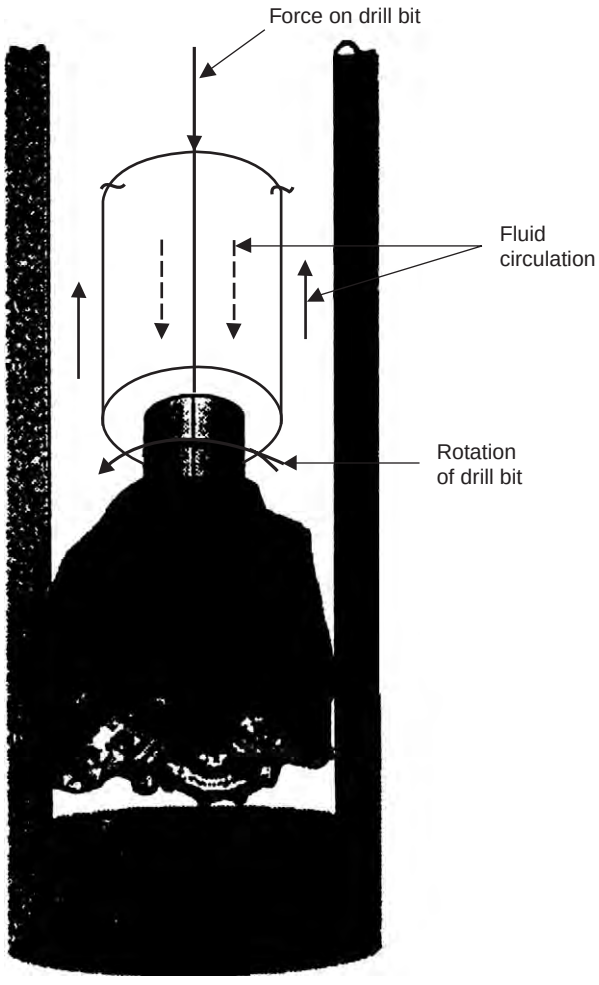


FIGURE 3 Elements required for drilling a hole.

- Right-hand rotation of the drill bit
- Circulating fluid (gaseous, liquid, or gasified liquid) from the surface through a tubular (drill

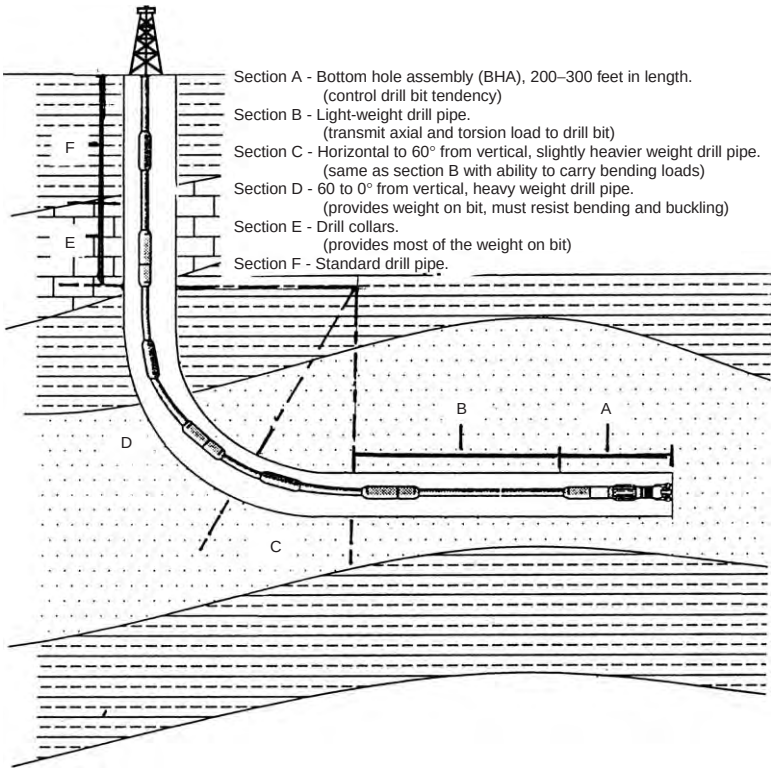


FIGURE 4 Drill string configuration in horizontal well drilling.

string) to hole bottom and back to the surface through an annular space

1.1 Force

The force on the drill bit is provided by slacking off tension on the drilling line through the drawwork part of the rig hoisting system. In vertical well drilling, this force is provided by the amount of certain length of heavy-weight pipes (drill collars) located directly above the bit that is allowed to go in compression as a result of slacking-off surface tension. When a well is being drilled at some inclined angle from vertical, heavy-weight pipes (e.g., drill collars, heavy-weight drill pipes) are no longer placed in the highly inclined sections; rather, they are placed in sections of lesser inclination or near vertical. The reason is that the heavier the tubulars in the highly inclined section of the hole above the bit, the greater the frictional forces on the drill string during sliding and/or rotation, possibly leading to the inability of advancing the drill bit to reach the intended target zone (Fig. 4).

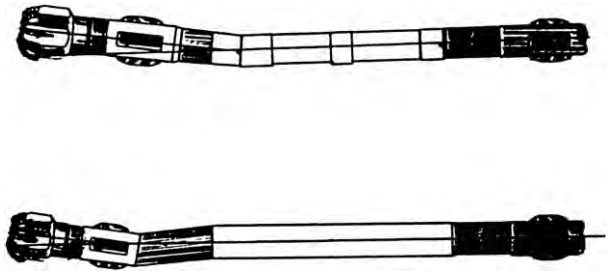


FIGURE 5 Positive displacement mud motors.

1.2 Rotation

Drill bit rotation may be induced from either a surface location or a subsurface location. From the surface, bit rotation can be provided through the use of conventional rotary table in conjunction with a kelly and kelly-drive system or a top drive motor. Subsurface rotation is provided through the use of down-hole mud motors (Fig. 5). For a given applied weight (force) on the drill bit and a desired rotary speed, the required rotary power is equal to the torque, mainly generated by the bit, multiplied by the corresponding rotary speed. However, when the well is inclined and bit rotation is provided

from the surface, the rotary power required is bit torque plus generated torque due to the frictional forces between contact surface of drill string and walls of the wellbore (hole) multiplied by the corresponding rotary speed. Therefore, in directional well drilling, rotary power requirement can be in excess of six times that in vertical well drilling, depending on hole inclination angle. As hole angle increases from vertical (0°) to horizontal (90°), the drag and torque will increase accordingly due to frictional forces. In the high inclined portion, including horizontal of the well, the effective weight of the drill string components will practically all go into friction drag and torque, and that represents undesirable drilling conditions. For excessive drag, advancing the bit may become the limiting factor in reaching the intended target. For excessive torque, the rig available rotary power may become the limiting factor in rotating the bit on bottom.

1.3 Circulation

While drilling, both heat and rock cuttings are being generated continuously. If the process of making a hole is to be conducted successfully, a fluid has to be circulated continuously from the surface to the bottom of the hole and back to the surface again to dissipate heat and remove rock cuttings. The various functions of drilling fluids are discussed later.

2. PURPOSE OF DRILLING

Drilling wells in an oil patch may be for the purpose of exploration, development, infill, injection, or reentry into existing wells. Exploration drilling is conducted to find and evaluate potential reserves in a reservoir rock. Development drilling is conducted to develop a field with proven reserves to its fullest extent of total production. Injection well drilling is conducted for the purpose of storing hydrocarbons (mainly gas), disposing of unwanted produced waters, or regaining the reservoir pressure by injecting water, steam, or gas into the reservoir. Reentry well drilling is conducted on existing wells for the purpose of deepening to lower pay zones, recompleting, or drilling laterals from the main trunk. Infill drilling is conducted to replace depleted wells or to drill additional wells so as to sustain production economics of the field.

3. ROTARY DRILLING REQUIREMENTS

Drilling a hole requires manpower and hardware systems. The manpower includes a drilling engineering group and a rig operating group. The first group provides engineering support for all drilling programs required to successfully and economically meet the objectives of the well being drilled. The second group is responsible for day-to-day operations of the rig and effective implementation of the drilling programs. It consists of tool pushers, drillers, and various rig crew personnel (e.g., roughnecks, derrickmen, enginemen). The hardware systems that make up a rig, whether for drilling offshore or onshore wells, are generally the same:

- Power generation system
- Hoisting system
- Rotary system
- Fluid circulation system
- Blowout prevention system
- Drilling data acquisition and monitoring system

Drilling rigs are classified as onshore (land) or offshore (marine) rigs. Their main features are mobility, safety, flexibility and well depth capability. The only major differences in offshore rigs are an additional system referred to as marine rise (extension pipe between rig floor and seabed) and in floating drilling a motion compensating system to keep the rig on location. Modern land rigs are built in units and skid-mounted so that they can be easily transported from one drilling site to another. Once on location, the rig components are assembled (rigging up) to drill the well. Offshore rigs are mounted on manmade surfaces that are sea bottom-supported or floating units. These units may be mobile or permanently fixed on location. The mobile units include semi-submersibles, jack-ups, and ships. Exploration drilling is always done using mobile units. Once a field has been economically proven in reserves, a stationary platform is permanently constructed to support further drilling, production operations, and living (Fig. 6).

4. RIG SYSTEMS

4.1 Power Generating System

The power system on rotary rigs is to provide sufficient power to allow all essential drilling operations to be conducted safely and efficiently. Depending on total well depth, rig power requirement may range

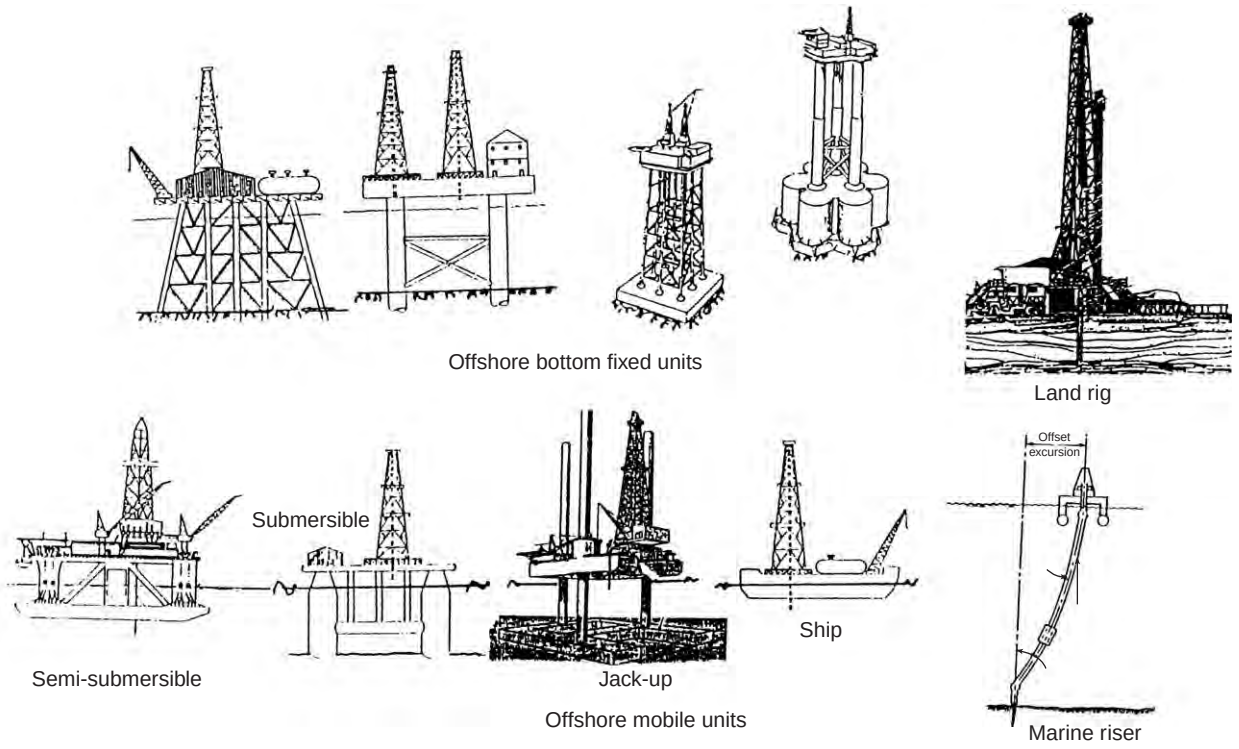


FIGURE 6 Various drilling rigs.

from few hundred horsepower to in excess of 5000 horsepower. Rig power is generated mostly by burning diesel fuel or sometimes on location by natural gas (if available). The power is transmitted to various rig systems (e.g., hoisting, circulation, rotary) by means of mechanical or electric drives. With mechanical drives, the power is taken directly from engines through chains or belts to particular rig systems. With electric drives, the engines drive electric generators to produce electricity, and electric motors (mainly the direct current type) are used to drive various rig systems.

4.2 Hoisting System

The primary function of the hoisting system is to lower and raise tools and equipment (e.g., drill string, casings, measuring tools) into and out of the wellbore during drilling, testing, and casings operations. The major components consist of the draw-work, crown block, traveling block, hook and swivel, drilling line, and elevator (Fig. 7).

The draw-work provides the hoisting and breaking power to enable movement of loads in and out of the hole. Its components include the drum, brake, transmission, and cathead. The blocks, which are pulley-type devices with drilling line strung between them, provide a mechanical advantage that allows raising very heavy loads out of the hole at optimum tripping-out speed.

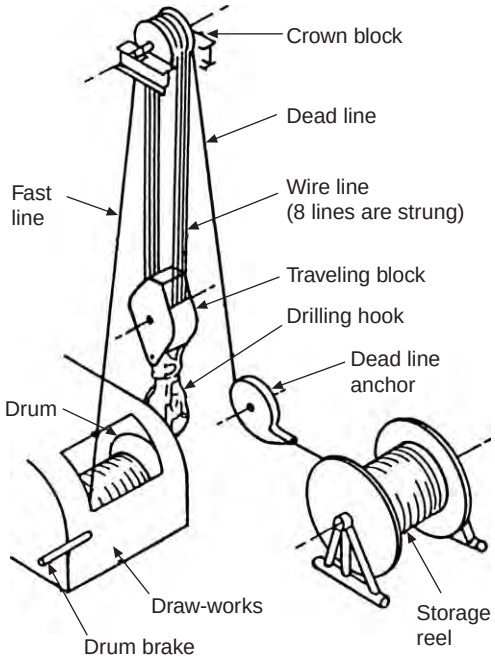


FIGURE 7 Hoisting system components.

4.3 Fluid Circulating System

The function of the circulating system is to allow the movement of drilling fluids from the surface, through the drill string, to the bottom of the hole, and back to the surface through the annular space. The main components include mud pumps, gas/air compressors, the drill string, mud tanks and treatment equipment, annular space, and all surface tubular connections (Fig. 8). Pumps used in rotary drilling are of the positive displacement type. They consist of pistons with rods, cylinders (liners), and intake/discharge valves. Duplex pumps have two liners and move mud on the forward and backward strokes. Triplex pumps have three liners and move mud only on the forward stroke. Air/gas compressors are used only when drilling with gaseous or gasified fluids. The drill string consists of tubular sections connected together by threaded joints. It also includes the drill bit. The annulus is the space between the outer wall of the drill string and the wall of the hole and provides the passageway for the drilling fluids and the drilled rock cuttings to return to the surface. In offshore drilling, the marine riser is used to extend the wellbore from the sea bottom to the rig floor.

4.4 Rotary System

The rotary system includes all components that allow drill bit rotation on the bottom. These are the drill string, swivel, rotary table, kelly, kelly bushing, and drive. Today, top drive motors, especially in offshore drilling, are becoming more popular and provide bit rotation instead of the conventional rotary table. They allow rotation and circulation not only during drilling but also during tripping operations. In addition, they allow full stand-length connections during drilling as compared with one length of drill pipe when using a rotary table. This adds more safety and higher drilling efficiency. In directional well drilling, down hole mud motors are sometimes used for bit rotation (Fig. 9).

4.5 Blowout Prevention System

The function of the blowout prevention system is to control the movement of kick fluids (formation fluids that enter the wellbore) during drilling, tripping, and casing operations. The system must allow (1) shutting the well at the surface, (2) safely removing kick fluids out of the wellbore, (3) replacing original drilling fluid with higher density fluids to prevent any further formation fluid intrusion, and (4) moving pipe in and out of the hole while under pressure (stripping operations). The basic components of the system include the blowout preventer stack or "BOP" (e.g., annular preventer, ram preventers, spools, internal preventers), the casing head, flow and choke lines and fittings, kill lines and connections, separators, and accumulators (Fig. 10).

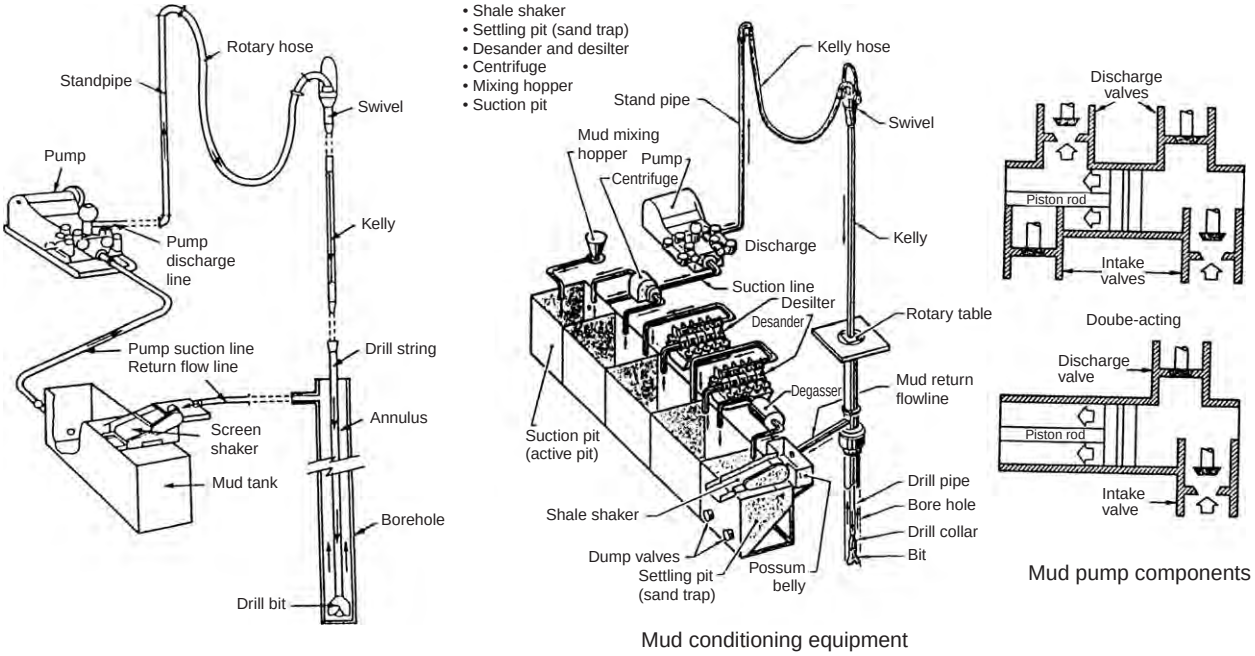


FIGURE 8 Fluid circulating system components.

The annular preventer, with reinforced rubber packing, will shut the annular space around any part of the drilling string in the hole regardless of shape or size. The ram preventers are of three types: pipe, shear, and blind rams. The pipe rams are like annular preventers; however, they have an opening that fits only around the body portion of drill pipe. The shear rams are two solid semicircular steel discs with a knife-like cutting edge that can totally shear (cut) the pipe into two parts and, hence, shut in the well above the lower part of the drill string left in the hole. The blind ram is similar to shear ram but with no knife-like cutting edges; therefore, it is activated only if there is no drill string inside it, thereby causing the well to be shut in. Internal preventers are like gate valves in that they open if the flow of fluid is downward and shut in if the fluid reverses its flow direction. Drilling spools are casing-type fittings placed in the BOP stack to provide space between two consecutive pipe rams for temporary storage of drill pipe tool joints that do not pass through the pipe ram when it is closed during stripping operations. Spools also provide attachments for choke lines and kill lines.

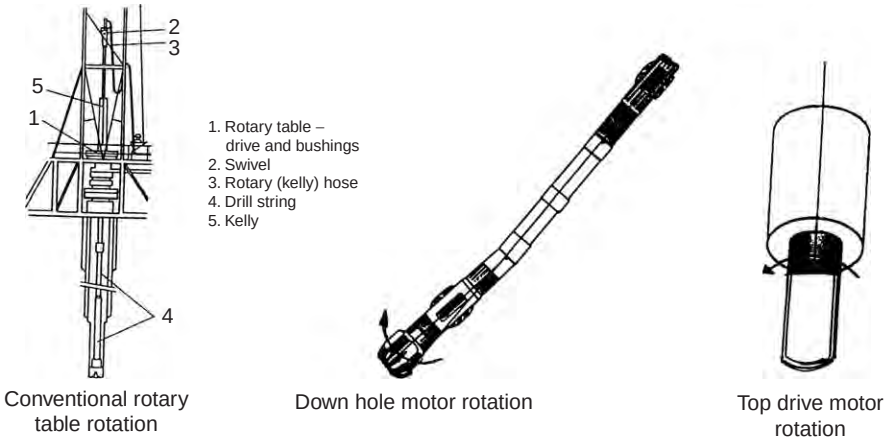


FIGURE 9 Rotary system.

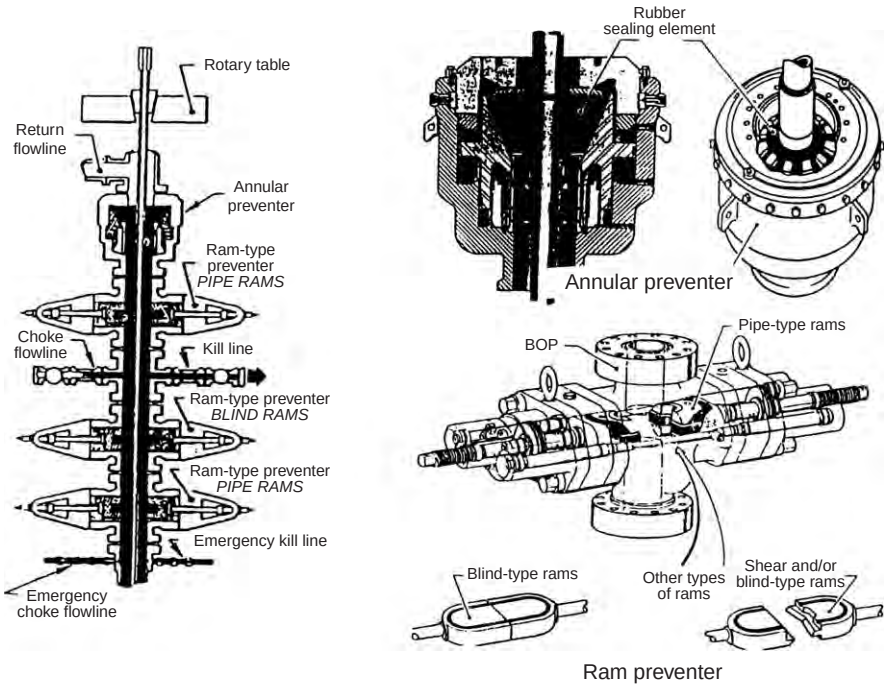


FIGURE 10 Blowout prevention system components.

4.6 Drilling Data Acquisition and Monitoring

The data acquisition and monitoring system consists of the devices (both surface and subsurface) used to monitor, analyze, display, record, and retrieve information regarding all aspects of the drilling operations. Parameters of interest include the rate of penetration (ROP) of drill bit into the rock; hook load of suspended equipment; hole depth; mud pump delivery pressure, flow rate, and speed (strokes per minute); torque; weight on bit (WOB); rotary speed and hoisting speed; mud properties; formation evaluation; and *in situ* bottom hole information (e.g., stresses, temperature, pressure). In the oil and gas industry, this system is regarded as the functional pulse of the drilling process. Costly drilling problems, such as well kicks, potential pipe sticking, lost circulation, excessive drag and torque on the drill string, and hole instabilities, can be minimized or prevented by a highly automated data acquisition and monitoring system.

5. DRILLING FLUIDS AND HYDRAULICS

5.1 Drilling Fluids

5.1.1 Functions

A drilling fluid is defined as a gaseous, liquid, or gasified liquid substance used in rotary drilling to perform any or all of the various functions required to successfully drill a usable hole for the production of oil or gas at the lowest overall well cost. These functions are (1) drilled cutting removal and bit cleaning, (2) containment of formation fluids, (3) hole stabilization prior to casing/cementing, (4) cooling and lubrication, (5) suspension of desired added solids, (6) reduction in casing and drill string suspended heavy weight, and (7) allowance in formation evaluation. The failure of a drilling fluid to perform any of its required specified functions can lead to very costly drilling problems and even a potential loss of the well. Therefore, the understanding and applications of each of the drilling functions are of the utmost importance. The following subsections provide a brief description of these functions.

5.1.1.1 Cleaning

Drill bit cleaning. For a drill bit to penetrate the rock efficiently, its cutting surfaces (e.g., teeth, compact diamond cutters, impregnated small diamond matrix) have to be cleaned continuously to prevent clogging and balling. In general, the lubricity and level of hydraulics (amount of fluid flow rate) are responsible for this function.

Bottom hole cleaning. Removal of drilled cuttings from underneath the bit as soon as they are generated is critical for enhancing the ROP and, thus, reducing the cost of drilling. In addition to impeding the drilling rate, if these cuttings are not removed immediately from underneath the bit, they will be regrinded to very small fines, making their removal from the mud both difficult and expensive. Therefore, the drilling fluid must provide hydraulics at the bottom beneath the bit to create high enough cross-flow velocities to sweep generated cuttings efficiently.

Annular hole cleaning. To avoid very costly problems such as mechanical pipe sticking, high torque and drag on the drill string, running casings/cementing difficulties, logging difficulties, and slow drilling, the efficient removal of cuttings to the surface through the annular space is a vital function of the drilling fluid. Some of the factors that are known to affect hole cleaning include fluid annular velocity, mud rheology, hole inclination angle, drill string rotation, annular eccentricity, ROP, and the generated cuttings themselves (e.g., size, shape, density)

5.1.1.2 Containment of Formation Fluids In conventional drilling, the drilling fluid density must be in excess of the equivalent formation fluid pressure density so as to prevent any intrusion of formation fluids into the wellbore (kicks). For example, if it is known that the formation pressure is "Pff" at a given well depth, a drilling fluid must be selected to have a density that will exert a higher pressure or "Pdf." This is referred to as overbalanced drilling. The pressure difference between Pdf and Pff is the overpressure in the wellbore and is referred to as the overbalance pressure.

5.1.1.3 Wellbore Stabilization Before drilling, the rock strength at some depth is in equilibrium with the *in situ* rock stresses and is free of contact with foreign fluids. However, while drilling, the balance between rock strength and the *in situ* stresses is disturbed. In addition, foreign fluids are introduced and an interaction process with the formation begins. This may lead to various types of wellbore instabilities such as hole closure, hole enlargement, hole fracturing, and hole collapse. Hole closure can cause excessive torque/drag, mechanical pipe sticking, and casing landing problems. Hole enlargement can cause hole deviation problems and difficulties in

cementing, logging, and removing annular cuttings. Wellbore fracturing can cause lost circulation, which in turn may lead to kick occurrence. Wellbore collapse can cause mechanical pipe sticking and possible total loss of the well. Therefore, to ensure wellbore stability, the density, chemical composition, and hydraulics of the drilling fluid must be selected and maintained properly during the drilling of a specific hole interval.

5.1.1.4 Cooling/Lubrication Considerable heat is generated from frictional contact of the rotating drill string and drill bit against the wellbore walls. This heat, if not dissipated continuously, can cause premature failure of drill bit and drill string components. Therefore, continuous circulation of a fluid during drilling to dissipate the heat is necessary. Lubrication of a drilling fluid is essential only in directional well drilling, specifically when far-reaching drilling is being conducted.

5.1.1.5 Aid in Formation Evaluation During drilling, it is important to identify the type of formation that is being drilled and to know the fluids it contains. The fluid in use must provide the means of allowing tools, such as logging tools, to obtain this information.

5.1.1.6 Suspension of Desired Solids When the density of a liquid must be increased by adding high-specific gravity solids, such as barite and iron oxide, the drilling fluid must suspend these solids under static conditions (so as not to allow settling) and dynamic conditions (so as not to allow sagging).

5.1.2 Selection

The governing criterion in the design of the drilling fluid program is to achieve the lowest overall cost per foot for the well. There are several considerations that enter into the selection of the drilling fluid to drill a specific hole interval. These include well type (e.g., exploratory well, development well), formation type (e.g., shale, anhydrite, salt, high temperature, lost circulation, producing/nonproducing), casing program, corrosion, rig, environmental impact, and water composition/availability.

5.1.3 Types

Drilling fluids are classified as gaseous, liquid, or gasified liquids. Gaseous drilling fluids include dry air/gas, mist, and foam. Liquid drilling fluids can be clear water (fresh or sea) or mud (water base or oil base). Gasified drilling fluids are mud with air/gas mixed. Nitrogen is used most commonly, especially in underbalance drilling (drilling while producing).

5.1.4 Properties

The drilling fluid properties that have the most impact on the drilling efficiency are density, viscosity, fluid loss (loss of liquid phase into the formation), chemical composition, and amount of solids. In general, low density, low viscosity, and low solids are most desirable. Chemical composition is very important when drilling certain types of shale and when drilling the reservoir rock.

5.1.5 Additives

To make up a drilling mud or maintain it during drilling, several additives may be required: viscosity building (e.g., bentonite or attapulgite clay, polymers), viscosity reducing (e.g., phosphates, tannates, lignites, lignosulfonates, sodium polyacrylates), weighting (e.g., barite, iron oxide, dissolved salts), fluid loss reducers (e.g., bentonite, starch, polymers), emulsifiers (e.g., oil in water, water in oil), lost circulation materials (e.g., granules, fibrous, flakes), and special additives (e.g., flocculants, corrosion control, defoamers, pH, alkalinity).

5.2 Drilling Fluids Hydraulics

Hydraulics is an area of drilling that concerns the interrelated effects of viscosity, mud weight, and flow rate on the drilling fluid performance in motion of its various functions. Conditions that may be ideal for the performance of one function may be detrimental for another. Thus, the conditions selected must represent a compromise that will lead to the best outcome for a given drilling operation. In general, the hydraulics program design deals with the following field operations: (1) during drilling and kick control and (2) during tripping and casings. During drilling, rig hydraulic horsepower (flow rate times corresponding pressure) must be sufficient to allow fluid circulation at a desired flow rate so as to achieve those fluid functions that will lead to successful drilling. These functions include hole bottom cleaning (drill bit hydraulics) and annular hole cleaning (cuttings transport). Pump hydraulic horsepower that must be delivered is the sum of power related to friction pressure losses of the moving fluid throughout the circulating system and power due to accelerating the fluid across bit nozzles. The first step in hydraulic program design is to determine an optimum flow rate, Q_{opt}, and corresponding optimum nozzle size diameters that will

ensure optimum bottom hole cleaning, thereby leading to a higher rate of rock penetration. However, this optimum flow rate must be equal to or less than some maximum flow rate, Q_{max}, and equal to or greater than some minimum flow rate, Q_{min}. Limiting the maximum flow rate are the maximum available rig hydraulic power, the resulting equivalent circulating density (ECD) of the fluid in motion, and potential hydraulic hole erosion. Unlike static fluid columns where the column pressure is directly related to fluid density, in a moving fluid column, the pressure is related to fluid density and friction pressure losses. During drilling, this will lead to an ECD that may be high enough to exceed the formation fracture gradient and, therefore, cause lost circulation. This same situation can also arise during kick control operations. The annular fluid velocity needed to efficiently transport rock cuttings to the surface governs the minimum flow rate. Friction pressure losses in annular space during tripping and casings operations can cause an increase in mud weight while moving pipe downward into the fluid column (surge pressures) and a decrease while moving pipe out (swab pressures). In the former case, as during drilling, this may cause formation fracturing. In the latter case, the circulating mud pressure may get lower than the formation fluid pressure, and this can lead to kick occurrence.

6. DRILL BITS

The drill bit is a tool whose only function is to break rocks into fragments for the process of making a hole in the earth subsurface. Rocks can vary in strength from very soft to very hard, making it necessary to have different types of bits that can be selected for a particular type of rock and strength.

6.1 Bit Types

Drill bits are grouped into two categories: roller cone and drag (Fig. 11). Roller cone bits are of two types: milled-tooth and insert. They are designed to drill rocks of various types and strengths. The milled-tooth bits are generally used when tooth wear is not a critical issue, whereas the insert bits with tungsten carbide inserts are used when tooth wear becomes the limiting factor in the life of the bit. Both the milled-tooth and insert bits have bearings wear that limits their operating lives. The rock-cutting mechanisms of roller cone bits use a shoveling (gouging) action in soft formations and a crushing (chiseling) action in hard formations. Drag bits are of two types as well: polycrystalline diamond cutters (PDCs) and diamond matrix body. The major differences from the roller cone bits are the rock-cutting mechanisms and a lack of bearings. The PDC bit cutting mechanism is rock shearing, whereas the diamond matrix is rock grinding. PDC bits are used in all nonabrasive rocks regardless of strength, whereas the diamond matrix bits are generally used in extremely hard abrasive formations. Drill bit selection is based mainly on past bit performance records. Past experience and learning are key factors in selecting the right bit for the right hole interval to be drilled.

6.2 Drill Bit Operating Parameters

The proper selection of bit operating parameters is of utmost importance in minimizing drilling cost.

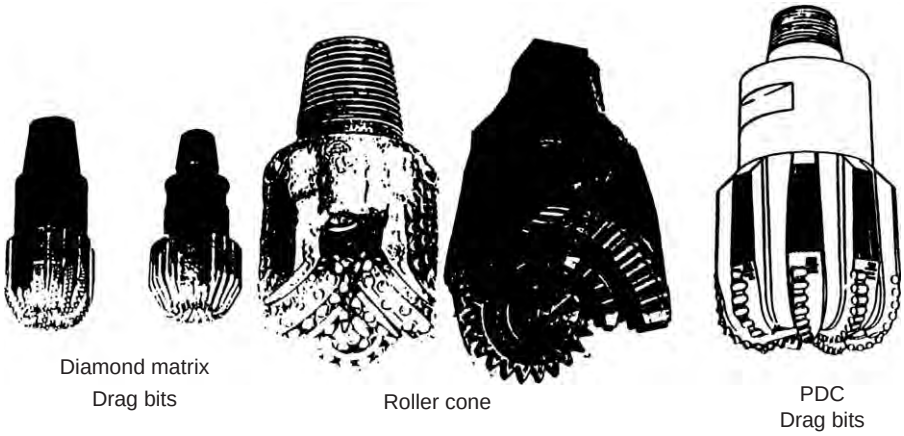


FIGURE 11 Various types of drill bits.

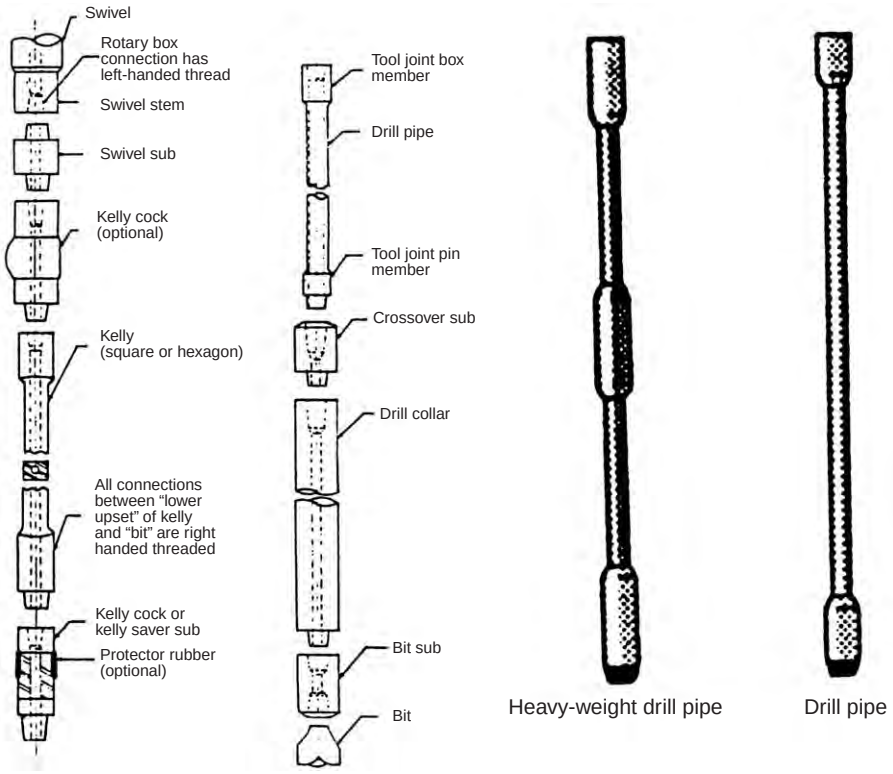


FIGURE 12 Drill string components.

These parameters include WOB, rotary speed in revolutions per minute (rpm), and the hydraulic parameters (flow rate and nozzle size). To achieve the lowest dollar per foot overall drilling cost, an optimum WOB and revolutions per minute with optimum bit hydraulics must be selected. In general, drag bits require lower weight and higher revolutions per minute than do roller cone bits. The hydraulic design is similar to both.

7. DRILL STRING

The drill string is an assemblage of tubular sections (approximately 30 feet in length) connected by threaded joints. It includes light- and heavy-weight drill pipe and very heavy drill collars. Its functions include the following:

- Provide and transfer weight to the drill bit
- Transfer rotation to drill bit if down hole motor rotation is not provided
- Provide a conduit for fluid circulation.

It must safely carry high axial loads, bending loads, torsional loads, and pressure, and it must resist buckling. The bottom hole assembly (BHA) portion of the drill string is that part behind the drill bit that provides directional control and weight to the bit (Fig. 12).

8. HORIZONTAL AND MULTILATERAL WELL DRILLING

8.1 Definition

Horizontal drilling is a process of directing a bit to drill along a horizontal path oriented approximately 85 to 95° from a vertical direction, whereas multilateral drilling is a process of drilling lateral holes from a main trunk such as horizontal, vertical, or directional (Fig. 13).

8.2 Purpose

The interest in drilling horizontal and multilateral wells is attributed to the following major factors:

- Ability to orient the borehole to accommodate parallel or orthogonal hydraulic fractures
- Ability to improve overall hydrocarbon recovery
- Ability to orient the wellbore in the direction of maximum stability

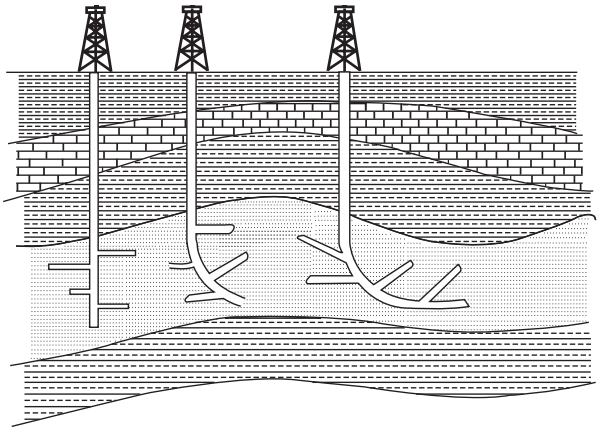


FIGURE 13 Multilateral wells.

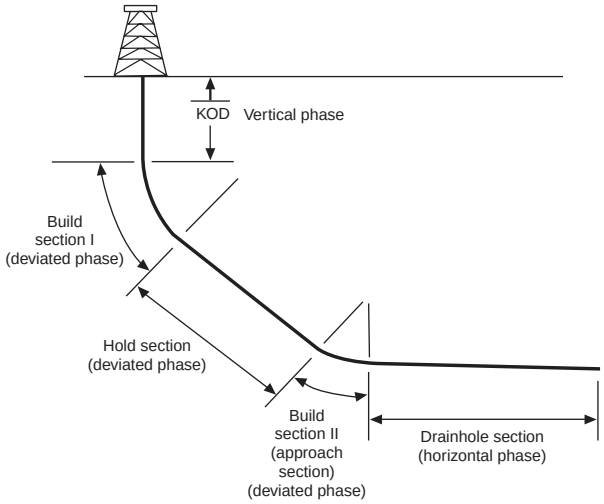


FIGURE 14 Horizontal drilling phases.

- Ability to reduce the number of wells in the development of a whole field
- Ability to enhance primary and secondary production due to lower drawdown pressures.

8.3 Applications

The reservoir rocks that are potential candidates for horizontal and multilateral drilling applications include the following:

- Naturally fractured reservoirs
- Low permeability reservoir (<1 md)
- Channel sand and reef reservoirs
- Coal bed methane reservoirs
- Reservoirs with potential water and/or gas coning problems
- Thin reservoirs
- Economically inaccessible reservoirs.

8.4 Drilling Methods

Horizontal well drilling consists of three stages: The first stage is a vertical hole section that starts at the surface and stays vertical down to a predetermined depth (kickoff depth). The second stage is a deviated section that begins at the kickoff depth and ends at an entry point into the reservoir at approximately 90° from vertical. The third stage is drilling the horizontal section into the reservoir (Fig. 14). The second stage (deviated section drilling) may be accomplished in one of four ways (Fig. 15):

Simple build curve. Continuously build at a constant build rate to reach the reservoir.

Ideal build curve. Continuously build at a constant build rate until a certain predetermined depth, and then change to another build rate to reach the reservoir.

Simple tangent build curve. Build at some preselected build rate until a predetermined hole inclination angle is reached, hold for a certain tangent length, and then build again at the same build rate as the first to reach the reservoir.

Complex tangent build curve. This is the same as the simple tangent except that the second build is at a different build rate than the first.

The drilling methods are identified by the amount of build rate angle used to drill the curved section. These include the following:

- Long radius drilling (build rate angle is < 6°/100 feet of drilled section)
- Medium radius drilling (build rate angle is between 7 and 35°/100 feet of drilled section)
- Short radius drilling (build rate angle is between 1.5 and 3.0°/1 foot of drilled section)
- Ultra-short build curve (theoretically no build curve section)

The most commonly used methods in the field are the medium and long radius methods. In long radius drilling, the tools can be similar to those used in vertical well drilling. However, for better control of the preselected well path and the ability to use measurement while drilling (MWD) tools and logging while drilling (LWD) tools, down hole mud motors or rotary steering systems are used.

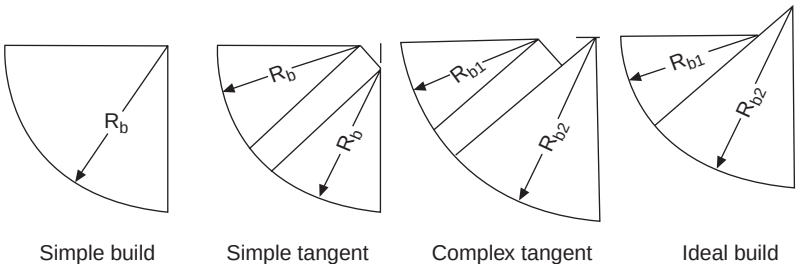


FIGURE 15 Build design curves.

9. HIGH-ANGLE/EXTENDED REACH DRILLING

9.1 Definition

High-angle (HA) wells are those having angles of inclination in excess of approximately 63° from vertical. Extended reach drilling (ERD) is where the ratio of well horizontal departure over its true vertical depth exceeds 2.0. Record horizontal reach with today's drilling technology has been reported in excess of 33,000 feet for a true vertical depth of 5000 feet.

9.2 Purpose

HA/ERD has unveiled worldwide opportunities to economically recover hydrocarbon reserves from single- or dual-field locations that might not have been possible otherwise. As in horizontal well drilling, it reduces the impact on the environment and overall cost significantly due to a reduction of the number of wells needed to develop an entire field or fields. The success of, and rapid increase in, HA/ERD activities is attributed to the tremendous amount of technology breakthroughs (engineering and tools), use of the learning process, effective teamwork, implementation of innovative ideas, preplanning, and real-time field monitoring of data while drilling.

9.3 Critical Issues in HA/ERD

Two of the most critical issues in HA/ERD are the associated inherent increase in torque/drag and the hydraulic requirement for drilled cuttings removal. In addition, there are other constraints that are related to drilling mechanics, field operating practices, logistics, and some formation characteristics. The following are brief descriptions of the concerns regarding these issues.

9.3.1 Torque/Drag

Efforts to reduce torque and drag, and therefore to increase target reach, have been in the proper selection of wellbore trajectory, increases in mud lubricity, use of rotary steerable systems, and development of nonrotating drill pipe protectors. Also of importance are models that enable engineers to predict torque and drag for effective preplanning in HA/ERD activities.

9.3.2 Top Drive Systems

Top drive systems are now available in the capacity range of 60,000 to 70,000 foot pounds and may soon reach higher capacities. The problem is matching the drill string torque strength with that of the top drive. There are some ways in which to increase drill string torque capacity such as reducing drill pipe tension capacity, and this in turn will allow an increase in tool joint strength, use of high-fiction thread compound that will allow higher torque with the same stresses, use of double-shoulder tool joints that will provide higher torsional capacity), and use of high-strength drill pipe (150,000–165,000 psi yield strength) that will provide higher torque capacity.

9.3.3 Wellbore Integrity

Successful HA/ERD is predicated on engineers' ability to properly manage the stability of the wellbore. Major factors that affect wellbore structural integrity are local area geology characteristics, *in situ* rock stresses, shale activity level, well inclination angle, and azimuth orientations. The primary means of managing wellbore stability are by use of proper mud density, use of mud filtrate that is compatible with that of the formation being drilled, proper selection of casing setting depths, advanced wellbore model analysis, real-time monitoring (both MWD and LWD) related to borehole closure, breakouts, enlargements, clay swelling, and pore pressure buildup. Data acquisition and data analysis are key factors in combating wellbore instability problems.

9.3.4 Drilling Fluids for HA/ERD

The key issues that need to be addressed in selection and design of the drilling fluids for HA/ERD are lubricity (high) to reduce frictional forces and proper viscosity (generally low shear rate rheology, high flow rates, drill string rotation) to enhance annular hole cleaning. In addition, the design of drilling fluids must have properties that will be least damaging to the pay zone and will provide the best wellbore stability for long open hole exposure. It must also have the least impact on the environment.

SEE ALSO THE FOLLOWING ARTICLES

Markets for Petroleum • Natural Gas, History of • Natural Gas Transportation and Storage • Occupational Health Risks in Crude Oil and Natural Gas Extraction • Oil and Natural Gas Exploration • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Oil and Natural Gas: Offshore Operations • Oil Industry, History of • Oil Pipelines • Oil Recovery • Oil Refining and Products • Public Reaction to Offshore Oil

Further Reading

Aadnoy, B. R. (1997). "Modern Well Design." Gulf Publishing, Houston, TX.

Bourgoyne, A. T., Millheim, K. K., Chenevert, M. E., and Young, F. S. (1986). "Applied Drilling Engineering." SPE textbook series. Society of Petroleum Engineers, Richardson, TX.

Lummus, J. L., and Azar, J. J. (1986). "Drilling Fluids Optimization: Practical Field Approach." PennWell Publishing, Tulsa, OK.

Rabia, H. (1985). "Oilwell Drilling Engineering." Graham & Trotman, London.

Oil and Natural Gas: Economics of Exploration

EMIL ATTANASI and PHILIP FREEMAN
United States Geological Survey
Reston, Virginia, United States

1. **Introduction**
2. **Fundamental Concepts**
3. **Discovery Process of Conventional Oil and Gas**
4. **Exploration: Individual Firm Decisions**
5. **Industry: Exploration and Public Policy**
6. **U.S. Exploration: Market Failure and Regulation**
7. **International Oil and Gas Exploration Provisions**
8. **Trends in Exploration and Discovery**
9. **Conclusions**

Glossary

barrel of oil equivalent Approximate energy equivalence relations where one barrel of crude oil = 6 thousand cubic feet of gas or 1.5 barrels of natural gas liquids.
basin-centered accumulation A regionally extensive and typically thick zone or unit of hydrocarbon saturated low-permeability rock in the deep, central part of a sedimentary basin.
common property resources Resources not under the control of a single authority and where access is unrestricted or cannot easily be restricted.
continuous-type accumulation A regionally pervasive hydrocarbon accumulation that is not bounded (delineated) by a hydrocarbon-water contact that commonly occurs independent of structural and stratigraphic traps.
externality Situation where the welfare of an individual depends not just on his own actions but directly on actions of another independent economic agent, such as when pollution upstream affects the welfare of downstream water users.
inferred reserves Expected cumulative additions to proved reserves in oil and gas fields discovered as of a certain date, where such additions come by extensions, additions of new pools, or enhancing hydrocarbon flow to the well.
known recovery The sum of a field's past production and its current estimate of proved reserves. Known recovery is an estimate of field size.
market failure Inability of traditional markets to efficiently allocate the use of a resource.
proved reserves Estimated quantities of hydrocarbons that geologic and engineering data demonstrate with reasonable certainty to be recoverable from identified fields under existing economic and operating conditions.
user costs The value of all future sacrifices (or the value of opportunities foregone) associated with production and use of a particular unit of *in situ* resource now rather than in some future period.

Exploration is the search for undiscovered oil and gas resources that have development and production costs no higher than the expected costs associated with adding and producing reserves from known deposits. This article explains from theoretical and empirical perspectives the economic forces that govern oil and gas exploration. It discuses public policy issues associated with oil and gas exploration. The concluding section summarizes statistics relating exploration efforts to discoveries in the United States and areas outside of the North America.

1. INTRODUCTION

In many areas of the world outside of North America and Europe, natural gas is still not an economic commodity because there are no local natural gas markets. On a Btu basis, natural gas is almost four times more costly to transport than oil, so transporting gas to international markets can be prohibitively expensive. Most of the description of the exploration/discovery process applies to both oil and gas. At the heart of the discussion, however, is the assumption that gas is commercially valuable. In areas outside of North America and Europe, it can be expected that reporting of some gas discoveries and

reserves is incomplete because the gas discovered is not commercial.

The exploration decisions of firms in a competitive industry lead to a level of exploration where the marginal finding (discovery) costs are related to "user cost." User cost is defined as the value of all future sacrifices associated with production and use of a particular unit of *in situ* resources now, rather than in some future period. Marginal finding costs and exploration productivity are important leading indicators of the long-term trend in the industry's life cycle and in the future availability of oil and gas supplies. The theme of this article is that information generated by the exploration discovery process for oil and gas provides the basic data for both industry and society to assess future resource availability and for planning and formulating a reasoned energy policy.

2. FUNDAMENTAL CONCEPTS

"Proved reserves" are estimated quantities of hydrocarbons that geologic and engineering data demonstrate with reasonable certainty to be recoverable from identified fields under existing economic and operating conditions. Estimates of proved reserves are important because they are the leading indicators of short-term sustainability of oil and gas production. No more than 10 to 15% of proved reserves (and often much less) can be extracted annually to avoid reservoir damage. So proved reserve levels limit annual production to amounts well short of known recoverable resources.

The objective of oil and gas exploration is to identify resources that can be added to proved reserves. These resources should be of such a quality and quantity that the resource can be commercially developed immediately. Industry exploration consists of a variety of activities that include surveying of surface geology, processing and interpreting newly collected geophysical data, reprocessing and interpreting previously collected data, acquiring mineral rights and access, taking subsurface core samples, and finally drilling exploratory oil and gas wells.

Exploration wells drilled to find reserves are classified by risk level. Categories for exploration wells are extension or outpost wells, new pool tests (shallower or deeper pool), and new field wildcat wells. Even the standard infill development well is not without risk; some development wells are drilled that fail to make contact with the producing reservoir. Risk, or the probability of failure, on average tends to increase from infill development wells to exploratory wells represented by extensions and outposts, new pool tests, and then to new field wildcat wells. Historically, an exploratory well is classified as a new field wildcat well when it is drilled at least two miles from the nearest productive accumulation. Other exploratory wells that find reserves through new pool tests and extension drilling commonly add reserves to fields already discovered.

Figure 1 is a schematic showing the different types of test wells (exclusive of new field wildcat wells). In the United States in the past two decades, the additions to reserves that are derived strictly from new field wildcat wells represent a relatively small proportion of annual additions to proved reserves. Most of the annual additions to proved reserves are the result of the drilling of well types represented in Fig. 1. The reserves found are credited to oil and gas fields that are already discovered and are recorded and published as extensions and revisions annually by the Energy Information Administration. The application of fluid injection programs and well stimulation also add reserves to discovered fields. These procedures facilitate flow of hydrocarbons fluids through the production well bore.

Estimates of sizes of oil and gas fields are commonly based on known recovery. A field's known recovery is defined as the sum of its past production and most recent estimate of proved reserves. When additions to proved reserves are credited to a field already discovered, its field size based on the estimate of known recovery is said to grow. Resource analysts call hydrocarbon resources that are expected to be added to the proved reserves of discovered fields inferred reserves. The movement of known hydrocarbon

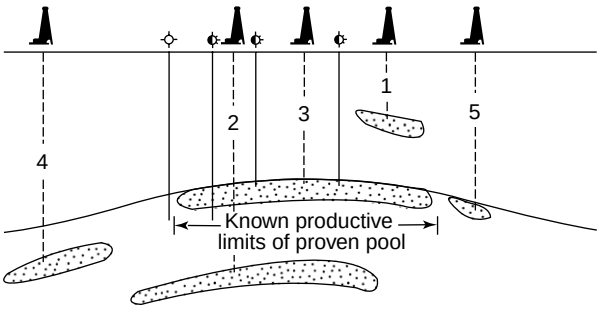


FIGURE 1 Schematic of wells leading to additions to reserves in discovered fields; (1) shallower pool test, (2) deeper pool test, (3) infill well, (4) new pool test, (5) extension or outpost. In practice, the operator or regulatory body may classify the accumulations penetrated by wells 1 through 5 as a single field or as more than one field. Recognition of the relationship among the accumulations could also be further complicated by the order in which the wells were actually drilled. Modified from Drew, 1997.

resources from the inferred to proved reserve category commonly requires exploratory drilling.

There may be some ambiguity whether new reserves are classified as either from new fields or from field growth (see Fig. 1). The U.S. Department of Energy defines a field as an area consisting of a single or multiple reservoirs all related to the same individual geologic structural feature and/or stratigraphic condition. There may be two or more reservoirs in the field that may be separated vertically or laterally by impermeable strata. From a practical standpoint, the definition of an oil or gas field is not exact. The assignment of pools to fields may be for convenience in regulation, or it may be an artifact of the discovery sequence of the pools, rather than for well-defined geologic reasons. Depending on the time sequence of discovery, accumulations shown in Fig. 1 may be assigned to a single field or to more than one field.

Figure 1 shows discrete oil or gas accumulations that are considered to be conventional and producible with conventional methods. Continuous-type oil or gas deposits are identified regional accumulations and are analogous to low-grade mineral deposits. An example is the Wattenberg continuous-type nonassociated gas accumulation that extends over an area of more than 1000 square miles in the Denver basin. The boundary of a discrete accumulation is usually defined by a hydrocarbon/water contact, but there is no such clear-cut boundary for a continuous-type accumulation (see Fig. 2). The geographic locations of continuous-type accumulations are generally known although their extent may be uncertain.

Continuous-type hydrocarbon accumulations commonly have reservoir properties that impede the flow of hydrocarbon fluids. Most of these accumulations require that extra measures be taken to induce additional flow to the well bore in order to attain commercial production rates. These measures may include drilling wells with horizontal sections or fracturing the formation to create passageways for the resource to migrate to the well bore. Unconventional or continuous-type gas accumulations have received increasing attention by the U.S. domestic petroleum industry during the past decade. Basin-centered gas accumulations (Fig. 2) may be partially developed by the recompletion and stimulation of wells that have depleted shallower conventional pools. If new wells target the continuous-type accumulation, the drilling risk is dominated by failure to stimulate or complete the well so that commercial flow-rates are attained, rather than by missing hydrocarbon-charged sedimentary rocks. Most of the wells drilled in continuous-type accumulations are classified as development wells. In this article, only exploration for conventional discrete accumulations is discussed.

Outside of the United States and Canada, oil and gas exploration is still directed to discrete conventional accumulations. Nonetheless the *in situ* hydrocarbon resources in the continuous-type accumulations are substantial, and the costs incurred in drilling and producing such resources provide a backstop or upper bound to the costs that should be incurred in finding and developing discrete conventional accumulations.

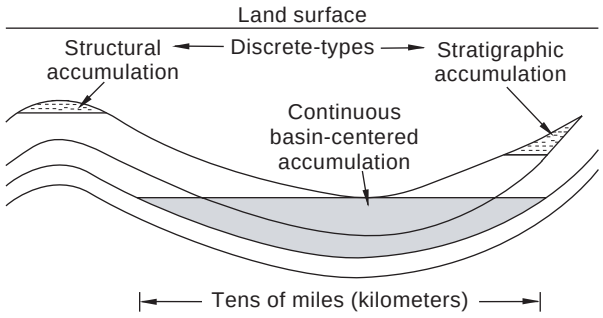


FIGURE 2 A schematic diagram that shows the geologic setting of a continuous-type accumulation in relation to discrete conventional accumulations in a structural trap and in a stratigraphic trap. Data from U. S. Geological Survey National Oil and Gas Resource Assessment Team, 1995.

3. DISCOVERY PROCESS OF CONVENTIONAL OIL AND GAS

Physical properties of the occurrence of oil and gas fields tend to cause regularity in the process of oil and gas discovery. The regularity provides the basis of predicting yields of future exploration from estimated undiscovered resources. Oil and gas deposits occur in sedimentary basins. Within a basin, discoveries are typically classified into sets of geologically similar deposits that are called petroleum plays. The petroleum industry uses the concept of the petroleum play, commonly identified in terms of a geologic formation or strata, as a basis to classify targets in a basin geologically.

The observed size-frequency distributions of discoveries in most petroleum plays and provinces are highly skewed. A very small proportion of the accumulations contain most of the discovered resource. Figure 3 shows the frequency-discovery size distribution of oil and gas fields discovered in the Permian basin of the United States through 1996, and Table I shows volumes of hydrocarbons

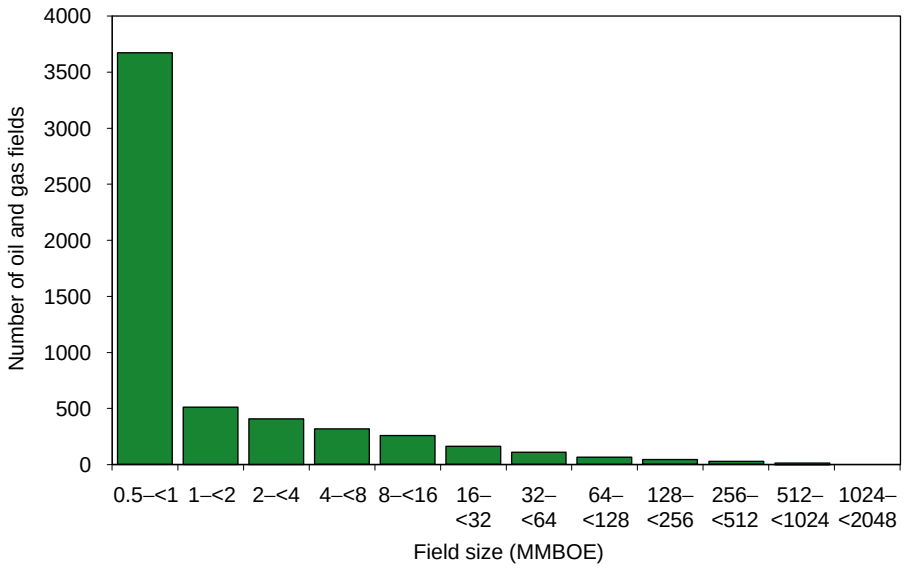


FIGURE 3 Histogram showing the size-frequency distribution of oil and gas fields in the Permian basin discovered through 1996. One barrel of oil equivalent = 6 thousand cubic feet of gas. MMBOE = millions of barrels of oil equivalent. Data are from the Energy Information Administration; field data are from 1998 issue of Oil and Gas Integrated Field File.

TABLE I

Proportion of Total Petroleum in the Permian Basin Contained in Various Size Classes of Fields

	Fields		Petroleum	
Size class[a] (BOE)	Number	Cumulative percentage	Percentage in size class	Cumulative percentage
2048–4096	1	0.02	4.86	4.86
1024–2048	4	0.10	12.14	17.01
512–1024	14	0.38	19.05	36.05
256–512	27	0.93	17.43	53.48
128–256	35	1.64	11.23	64.71
64–128	61	2.87	9.92	74.63
32–64	91	4.71	8.01	82.64
16–32	133	7.40	5.61	88.25
8–16	216	11.76	4.66	92.91
4–8	267	17.15	2.85	95.76
2–4	344	24.11	1.85	97.61
1–2	432	32.83	1.15	98.75
0.5–1	481	42.55	0.64	99.40
0.25–0.5	495	52.56	0.33	99.73
0.125–0.25	483	62.32	0.16	99.89
0.0625–0.125	451	71.43	0.08	99.97
<0.0625	1414	100.00	0.03	100.00

[a] In millions of barrels of oil equivalent (BOE) where 1 barrel of oil = 6 thousand cubic feet.

associated with each discovery size class. The Permian basin, located in Southwest Texas and Eastern New Mexico, is the most prolific oil-producing basin in the conterminous United States. The largest 27 fields (or 0.55% of the fields) contain more than half of the hydrocarbons discovered to date, whereas the 1414 fields in the smallest size class (less than 60 thousand barrels of oil equivalent,

where one barrel of crude oil equals 6000 cubic feet of gas or 1.5 barrels of natural gas liquids) account of only 0.03% of the hydrocarbons.

The shape of the Permian basin discovery size frequency distribution is typical of most plays and basins; most of the hydrocarbons are found in a few very large accumulations or fields. Exploratory drilling success-ratios (successful wells to wells drilled) provide no information about the sizes of future discoveries. The large number of small accumulations determines success ratio levels. The largest field, containing nearly 2.6 billion barrels of the oil equivalent, is at least six orders of magnitude larger than the fields in the smallest size class. If the x-axis had not been in log base 2, the size classes for the smallest category and the largest category could not have been graphed without a break in the axis.

Assuming the surface expression of a deposit is roughly proportional to its volume, then even with random drilling the average discovery sizes would decline with equal increments of exploratory drilling, as the largest deposits are found early in the discovery process. For example, the surface area of the 6 billion barrel East Texas field covers more than 200 square miles. Any improvement over purely random drilling quickens the early discovery of large fields and accelerates the decline in the discovery rate. Figure 4 shows the progression in the average discovery sizes (in barrels of oil equivalent) from 1920 to 1996 in 5-year increments for the Permian basin. The regularity of the discovery process allows the expected yields of future exploration to be computed with simple analytical models. The discovery rate (yield per exploratory well) is controlled by the sizes of fields discovered and the order of discovery.

The exploration histories of many of the basins in the United States follow predictable patterns. New-field wildcat wells were drilled in unproven plays at low but irregular rates. After a significant discovery is made, there is typically an influx of new prospectors and much higher levels of drilling; similar to the 19th century gold rushes. Drilling rates eventually weaken as returns deteriorate and exploration in other plays or basins becomes more attractive. As other plays in the basin are tested, then the process repeats itself following another significant discovery in an unproven play. An efficient exploration process, where the larger accumulations are found early, results from physical characteristics of the size distribution of oil and gas deposits in nature even when wells are randomly sited. If the industry is reasonably efficient in finding the largest and lowest cost accumulations early in the discovery process, then the past discovery sequence provides information useful in estimating the magnitude and

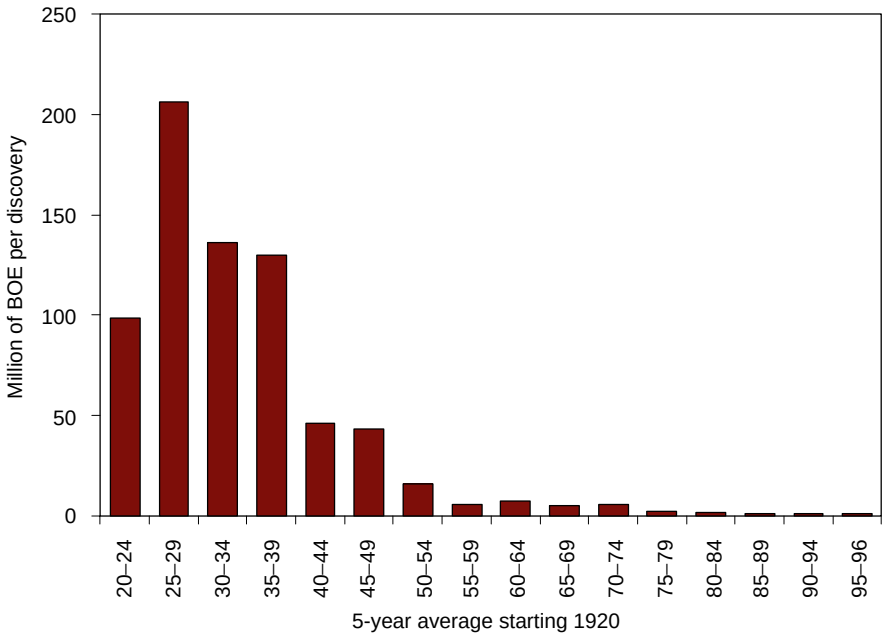


FIGURE 4 Average sizes of fields discovered in 5 year intervals in the Permian basin between 1920 and 1996. Field data are from the Energy Information Administration, 1998 issue of Oil and Gas Integrated Field File. 1 barrel of oil equivalent (BOE) equals 6 thousand cubic feet of gas.

characteristics of the undiscovered resources. For areas where field size is narrowly defined (such as the sum of only proved reserves plus past production), adjustments may be required to the estimated sizes of recent finds to allow for their full delineation so as not to overstate discovery decline in sizes.

In the United States, a relatively small proportion of the known sedimentary basins (or provinces) contain most of the hydrocarbons discovered to date. Figure 5 shows the distribution of known recovery (past production and proved reserves of oil, gas, natural gas liquids) in fields discovered through 1998, expressed in barrels of oil equivalent for the U.S. onshore basins located in the conterminous United States. The distribution represents 57 onshore lower 48 provinces that were delineated and assessed by the U.S. Geological Survey for the 1995 National Assessment of Oil and Gas Resources. Table II lists

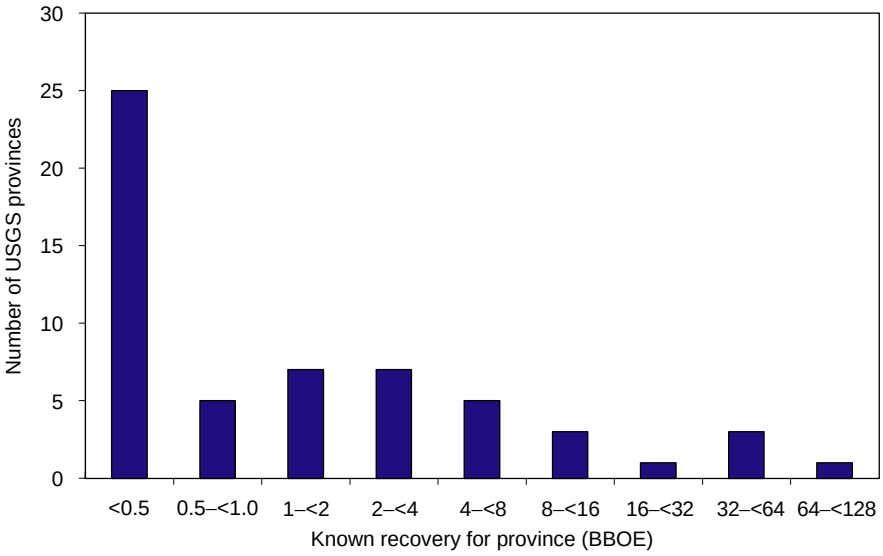


FIGURE 5 Histogram showing the cumulative discovered through 1998 hydrocarbon volumes-frequency distribution for U.S. basins through 1998. BBOE = billions of barrels of oil equivalent. Data from NRG Associates, 2000.

TABLE II

The Ten Leading U.S. Onshore Petroleum Provinces of the Conterminous 48 States

	Units[a] of oil and gas				
	(BBO)	(TCF)	(BBL)	(BBOE)	Discovery date[b]
1. Gulf Coast	25.8	265.8	8.3	75.6	1901
2. Permian Basin	34.4	102.3	6.5	55.7	1920
3. Anadarko Basin	5.0	157.1	5.6	34.8	1916
4. Miss-La. Salt Domes and E. Texas	16.5	84.6	3.5	32.9	1895
5. San Joaquin	15.4	13.2	0.8	18.2	1887
6. Appalachian Basin	3.4	43.0	0.0	11.0	1871
7. Los Angeles	9.0	7.5	0.4	10.5	1880
8. San Juan	0.3	45.5	1.5	8.8	1921
9. Bend-Arch Ft Worth	4.9	13.0	0.9	7.7	1902
10. Cherokee	6.3	2.3	0.0	6.7	1873

[a] Units abbreviations: BBO = billions of barrels of crude oil, TCF = trillions of cubic feet of gas, BBL = billions of barrels of natural gas liquids, BBOE = billions of barrels of oil equivalent. Gas and natural gas liquids conversion factors the following: 1 barrel of oil equivalent = 6 thousand cubic feet gas = 1.5 barrels of natural gas liquids.

[b] Discovery date refers to the discovery date of the first field of at least 10 million barrels of oil equivalent.

From NRG Associates, Significant oil fields in the United States 2000, database.

the known recovery for the ten leading onshore provinces and the discovery date of the first field having at least ten million barrels of oil equivalent. Four of the 57 provinces with the largest known recovery account for 63% of the petroleum discovered to date, and the top nine have more than 80% of the discovered petroleum. On a worldwide basis, relatively few provinces contain most of the hydrocarbons discovered to date.

Commercial petroleum was discovered in the United States in the middle of the 19th century. Just 50 years later, at the end of 1901, petroleum provinces containing more than half the known petroleum recovery (oil and gas) in the onshore lower 48 states had reported at least one discovery of at least 10 million barrels of oil equivalent. By the end of 1920, the basins that contained 94% of oil and gas discovered through 1998 had already been explored sufficiently, so that in each basin at least one field of 10 million barrels of oil equivalent had been discovered. The exploratory drilling through 1920, however, represented only a small percentage, less than 5%, of total exploratory drilling through 1998. Most onshore exploratory drilling since 1920 in the lower 48 states has been follow-up drilling in these and other less prolific basins.

Exploration access, distance from market, and technology are factors determining the order in which basins are explored. In the very early periods of the U.S. petroleum industry, search intensity was influenced by the costs of transporting and marketing petroleum. Some of the remote prolific basins were not explored as early as those closer to market areas.

4. EXPLORATION: INDIVIDUAL FIRM DECISIONS

Exploration expenditures are direct investments made by individual firms for the purpose of locating unidentified, but potentially commercial quantities of oil and gas. For an individual firm, exploration often begins with a hypothesis about the formation of hydrocarbon accumulations in a specific geologic setting. The process typically starts with a literature search to identify potential geographic locations where the hypothesis might be tested. The selection of a target area includes review of geologic maps, reconnaissance-type geochemical data and geophysical data compilation from the open literature and from commercial vendors. Fieldwork is often required to verify interpretations and to collect additional data to identify specific drilling targets. During drilling, data are continuously collected and interpreted to maximize the chances of locating significant accumulations. With the location of hydrocarbons, additional geologic and economic analysis will determine chances of a commercial discovery.

The economic theory that explains how firms set the optimal level of exploration expenditures is complex and beyond the scope of this article. The discussion that follows is highly simplified, but readers primarily concerned with public policy and trends in exploration and discoveries may wish to skip it.

Economic determinants of exploration investment are typically studied in the context of a representative firm's planning strategy. Suppose we consider a representative firm that searches for and produces oil in several areas over several time periods. The optimal plan tells the firm how much oil to produce each time period and the level of investment expenditures in production facilities and exploration to make (both regionally and temporally) so as to maximize the present value of its net cash-flow stream over the planning period.

In general, the optimal strategy is to increase oil production (extraction per time period) until the marginal cash profit (revenue minus operating costs) is balanced with the present value of the user costs. Simply, user costs represent the opportunity costs associated with producing and using a unit of resource now rather than postponing it. Specifically, it consists of the sum of (1) the future value of a unit of production retained in the ground instead of produced and (2) the cost, in terms useful life of production facilities if the unit of resource is produced at the present time. The value of retaining the unit in the ground is linked to the ease of replacement of that unit. For example, the value of postponing production may be small if discovery costs are low, that is, if replacement is easy, or if future resource prices are expected to decline because a cheaper substitute can fill demand. Marginal discovery or finding costs are sometimes used to approximate user costs. The greater discovery costs to replace that unit, the greater the user costs that pushes the representative firm to postpone production of that unit.

Abstracting from effects of uncertainty for the moment, for each time period exploration expenditures should be set to where the marginal discovery costs of new reserves equals the marginal value of newly discovered reserves. This value, in turn, is related to the user costs associated with the extraction

rule discussed in the previous paragraph. If the firm can put value on additional variables related to, for example, its experience in an area, then it should try to include these benefits in its decision.

Exploration strategy is coordinated with production and investment strategies to ensure that overall costs are minimized and the present value of net income optimized for the entire planning period. For an individual firm operating in multiple areas, if its threshold hurdle rate (required rate of return on capital) cannot be achieved in the region, or if it has a host of other more attractive opportunities elsewhere, then for that area, the firm will withdraw exploration, sell off prospects, and deplete producing properties.

Uncertainty in the returns to exploration effort leads to a departure from the strategy described earlier. The magnitude of this effect depends on a firm's risk tolerances, how uncertainty affects marginal product of exploration, and the perceived risk of ruin. However, viewed at the industry level, there is a distribution of industry hurdle rates of return and risk tolerances. This distribution ensures that information generated by the exploration process—that is, marginal discovery costs—is not an artifact of a single firm's behavior, but is representative of the industry and thus has general value for assessing future industry prospects.

Though many authors have proposed elegant theoretical models of how firms make extraction, investment, and exploration decisions, there is still a surprising lack of empirical research that tests and verifies the hypotheses that have been generated.

5. INDUSTRY: EXPLORATION AND PUBLIC POLICY

At the industry level, production, investment, and exploration patterns are examined for their consistency with the representative firm's optimal strategy and in the context of certain public policy questions. For example, the government might institute incentives designed to encourage domestic oil and gas exploration effort in order to maintain domestic production capacity. In addition to direct tax incentives and subsidies for exploration, the other government policy instruments include access to minerals on public lands, research and development policy, import quota/tariff policy, and direct price controls. All of these instruments have, to some extent, been used during the history of the U.S. oil and gas industry.

The standard formulation of the economic theory of exhaustible resources, as proposed by economist Harold Hotelling, asserts that the net prices of *in situ* finite resources should increase at the rate of at least the current rate of interest. Hotelling's theory rests on the assumption of a fixed stock of reserves that can be produced at will. However, proved reserves of oil and gas can be augmented through exploration or through more intense development. Production rates of oil and gas wells are constrained or limited by reservoir pressures. It is not surprising that many empirical studies have failed to verify the naïve predictions of the Hotelling theory. When exploration is modeled at the industry level, the derived theoretical implications are rarely unambiguous.

A broader question relating to public policy is whether industry exploration costs and actions accurately indicate future prospects of oil and gas. If the finding costs accurately reflect the costs associated with replacement costs of current proved reserves, then trends in such costs could foreshadow future supply problems. Finding costs, however, are not directly observable from industry or government data sources. It may take years for new finds to be sufficiently developed so that their size can be accurately estimated. Though discovery costs are not directly observable, the locations, search intensity, and the apparent risks the industry accepts reflect its assessment of remaining domestic and worldwide oil and gas exploration prospects.

If it is assumed that each firm will equalize the marginal costs of additional reserves either through exploration, intensive development that adds reserves, and purchase, then one can infer from market transactions prices the approximate industry finding costs. Empirical studies of oil and gas property transactions values show a rather consistent pattern. Developed in-ground oil and gas reserve prices hover at about one-third wellhead prices or at half of net prices, which are the wellhead prices minus all operational costs including taxes.

For the individual firm, exploration effort can be expected to increase until marginal finding costs equal user costs. User cost represents the benefits of adding resources to reserves. Empirically, user costs are measured as the difference between the in-ground market value of the developed proved reserves and the development costs. User costs embody the future price expectations of market agents, just as the valuation of properties embodies future reserve price expectations.

For the U.S. petroleum industry, oil price expectations are set in world markets but natural gas price

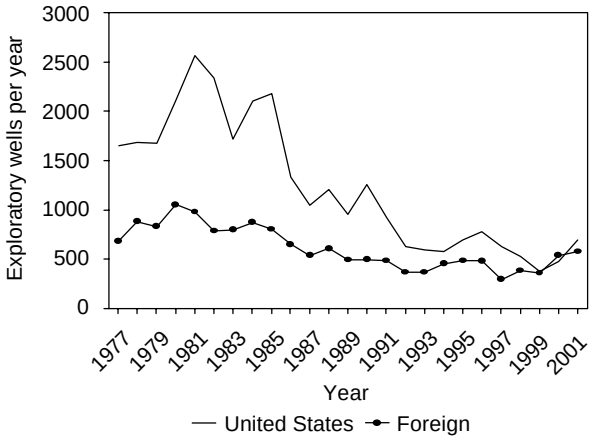


FIGURE 6 Number of total oil and gas exploration wells drilled by the group of Financial Reporting System (FRS) companies in the United States and foreign countries. Data from EIA, Performance profiles of major energy producers 2001, 2003.

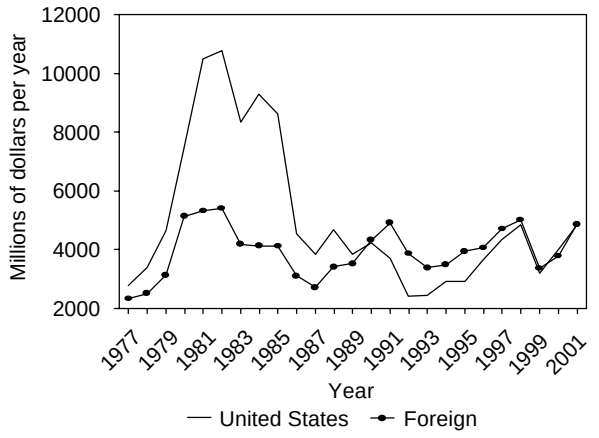


FIGURE 7 Total oil and gas exploration expenditures by the group of Financial Reporting System (FRS) companies in the United States and foreign countries. Data from EIA, Performance profiles of major energy producers 2001, 2003.

expectations are set by the expected demand supply interaction within North America. The unprecedented increase in perceived U.S. user costs during the late 1970s and early 1980s led to record drilling levels and oil finding costs in the lower 48 states but failed to significantly increase proved reserves. User costs as well as exploration activity declined in the United States as world crude oil prices plummeted in 1986 to half the levels that prevailed in 1985.

The U.S. Energy Information Administration compiles and reports the financial accounts, called the Financial Reporting System (FRS), for the group of firms representing the largest domestic crude oil producers. Although there have been changes in the composition of the FRS firms over the years, these firms are the leading U.S. oil and gas producers. The report includes financial expenditures on exploration including numbers of wells drilled, oil and gas production by the country where the producing property is located, and investment expenditures.

Figures 6 and 7 show U.S. and foreign exploratory wells and expenditures reported for the FRS firms. During the early 1980s, both domestic exploratory drilling and exploration expenditures substantially exceeded exploration outside the United States. Since the late 1990s, domestic and foreign exploration drilling and expenditures by these firms have been about equal. The volume of crude oil production from foreign properties for the FRS firms has steadily increased, and in 1997 it exceeded oil production at U.S. properties. Within the United States since 1993, more than half the FRS firms' domestic exploration expenditures have targeted offshore prospects, principally in the deep water of the Gulf of Mexico. In summary, the industry exploration trend demonstrated by the FRS group of firms (representing the industry leaders) has been to focus exploration effort in the deepwater offshore areas and countries outside of the United States where the opportunity to discover large accumulations is greater.

6. U.S. EXPLORATION: MARKET FAILURE AND REGULATION

The U.S. petroleum industry is a mature industry. U.S. crude oil production reached its peak in 1970, and gas production reached its peak in 1971. In 2000, the United States produced 5.86 million barrels per day of crude oil, or 8.7% of the world's production. It also produced 20.1 trillion cubic feet of gas, which represents about 23% of the world's gas production. During the history of oil and gas exploration in the United States, the combination of private ownership of mineral rights and the competitive market forces produced excesses in exploration and production that ultimately resulted in government regulation. Actions that were rational from an individual firm's perspective turned out to be very costly from society's standpoint. Although some of these inefficiencies are in the past, they are reviewed here to provide insight into the regulatory environment.

6.1 Externalities and Regulation

In the United States, mineral rights can be owned either privately and by the federal and state governments. Government-owned mineral rights are typically

transferred to the private sector for exploration and production. Several systems were used to transfer mineral rights to the private sector in the past. A comprehensive analysis of these systems with respect to economic efficiency and government's ability to capture economic rents is beyond the scope of this article. However, where significant hydrocarbons are expected, the method now used to transfer of oil and gas rights to the private sector is the sale of leases through a competitive bidding system.

Private ownership of mineral rights resulted in what economists refer to as market failures and led to governmental regulations to remedy these failures. Because an oil or gas pool can extend over several parcels, there could be a number of different property owners with a claim to the oil or gas in a common pool. This is the common property or pool externality. An externality occurs when the welfare of an individual depends directly not just on the individual's activities, but also on activities controlled by some other economic agent. For example, pollution generated upstream affects the welfare of downstream water users.

In the absence of regulation, oil and gas resources were subject to the rule of capture. Historically, the owner who first discovered the pool would start immediately to intensively drill and produce the resource. In the subsurface, oil and gas flows to where there is a reservoir pressure differential; so by quickly drilling production wells, the first producer could drain resources from neighboring owners. The neighboring owners must exploit their part of the pool as quickly as possible just to avoid losing resources to the early producer. This practice accelerated exploration and resulted in too much exploration and too rapid production of the resource. The accelerated production reduced the quantity of the resource that could ultimately be recovered from the pool.

Even when oil prices collapsed, production would not decline because each producer feared having one's own part of the pooled drained by a neighbor. In the leading oil-producing states, government regulators instituted pro-rationing that limited production for each well to an allowable rate per month and set minimum spacing for wells. Well-allowable rates were set to restrict oil supply in order to establish a floor on wellhead prices. For many states, pro-rationing remained in effect until the early 1970s, when the first Arab oil embargo occurred. State regulatory commissions and the federal government now require that production of newly discovered pools be unitized among the owners to avoid the common pool externality and to assure that potentially recoverable resources are not lost by early overproduction.

Information is crucial to the competitive exploration process. During the period when the rule of capture rewarded early exploration, a system of scouting services was developed in the industry. The domestic and international scouting services report to subscribers land acquisition transactions, well-permitting applications, drilling activities, and post-drilling actions of firms exploring an area. The information generated by discoveries and dry holes is valuable to neighboring leaseholders and other exploration firms who have not taken the risk of actually drilling. The benefits from observing the outcome of drilling without bearing the risk is an externality known as an information spillover. In fact, some Canadian provincial governments lease tracts in a checkerboard fashion to maximize their benefits as the ultimate owner of the resource of such information spillovers.

With the general adoption of forced unitization of new discoveries during the last half of the 20th century, major integrated companies followed the strategy of taking large land positions around a prospect and farming out small parcels to an independent firm (operator) in exchange for drilling a risky wildcat well. In many cases, the independent operator has a lower cost of capital than the major firm, because the independent can fund the well through limited partnerships that provide tax benefits to individual investors. With this system, the firm that actually drilled the discovery well received only a small fraction of the actual discovery. For example, in the Denver basin from 1949 through 1974, the independent firms were credited with drilling discovery wells for fields representing 75% of the reserves but ended up owning only 21% of the reserves in those deposits. The remaining reserves were credited to the major companies because of their dominant land positions.

6.2 Federal Policies to Promote Exploration

The federal government has encouraged oil and gas exploration during most of the 20th century by favorable income tax treatment, specifically through a provision known as the percentage depletion allowance. Favorable tax treatment for petroleum investments enabled promoters and operators of independent firms to raise significant amounts of money by offering tax shelter plans to fund this

drilling. During the peak years of exploratory drilling from 1981 to 1982 in the United States, most of the exploration wells drilled were funded through such arrangements. This funding was reduced dramatically when the Tax Reform Act of 1986 significantly reduced tax rates and rescinded part of the tax code favorable to such tax shelter funds.

During the 1950s and the following decade, the federal government imposed quotas on imported crude oil. The stated purpose was to maintain domestic crude oil production capacity for national security reasons. The quotas had the effect of insulating domestic prices from lower world prices and providing the necessary incentives for exploration and production of domestic resources. Import quotas were rescinded as domestic oil production approached its peak in 1970 and it was clear that supply would not grow as rapidly as U.S. demand.

7. INTERNATIONAL OIL AND GAS EXPLORATION PROVISIONS

Since the early part of the 20th century, British, Dutch, and French oil companies have explored for and produced oil internationally, starting with their former colonies. After World War II, all major U.S. companies initiated international exploration to various degrees.

Outside the United States, national governments typically control ownership of mineral rights. The rights to exploration concession areas are negotiated with the central governments. Although some of the European countries with North Sea production follow an auction system, in some ways similar to the U.S. system, they still maintain tight control of resource development. Many leading oil and gas-producing countries require separate exploration and production agreements. For some countries, including those in the Middle East, Mexico, and Venezuela, petroleum exploration and production is almost entirely controlled by the state oil company, and foreign companies, if allowed at all, are used on a contract basis.

For countries that permit foreign participation, exploration concessions typically require a work commitment including investment in data collection and a specified number of wells drilled in exchange for the right to explore. The concession areas are sufficiently large so that a field will have a single owner, thereby eliminating common pool externalities. In addition, the concession contract will specify that all data collected by the concessionaire should revert to the government for use by subsequent concessionaires. The agreement will also specify what rights, if any, the exploration firm might have in the event of a discovery. Once a discovery is made, a production concession is negotiated, often requiring the private company to share production with the state oil company or to assume the role of contract producer with payments taken in kind as a percentage of production.

Each government has a different set of contract provisions and production taxes designed to attain a specific set of policy goals. These goals may include training and employment of local workers, construction of infrastructure, and provision of social services to the local communities. Moreover, the governments typically reserve the right to tailor provisions on an individual basis. A description of these provisions is well beyond the scope of this article.

8. TRENDS IN EXPLORATION AND DISCOVERY

8.1 Drilling Activity

Summaries of the total numbers of wells drilled outside the United States are published by country in the journal *World Oil* and the American Petroleum Institute's *Petroleum Data Book*. Data show that the percentage of worldwide annual drilling represented by the wells drilled in the United States went from 80% in 1970 to 30% in 2000. In 1950, the United States accounted for 52% of the world's crude oil production, but by 2000 its share had fallen to less than 9%. The time profiles of new field wildcat wells (successful and dry) drilled each year since 1950 for each of three country groups are shown in Fig. 8. The drilling records for some countries are undoubtedly incomplete, but the graph illustrates the decline in U.S. new field wildcat drilling as well as the decline in the U.S. share of world new field wildcat drilling after 1981. The U.S. share declined from over 90% in 1950 to around 55% by the late 1990s.

8.2 Past Discoveries

There are no internationally accepted standards for reporting reserves or sizes of discoveries. In some countries such as the United States and Canada, reserve definitions are very conservative because they are tied to production facilities and economic conditions. In other countries, the definition of reserves is broader, reflecting the volumes of

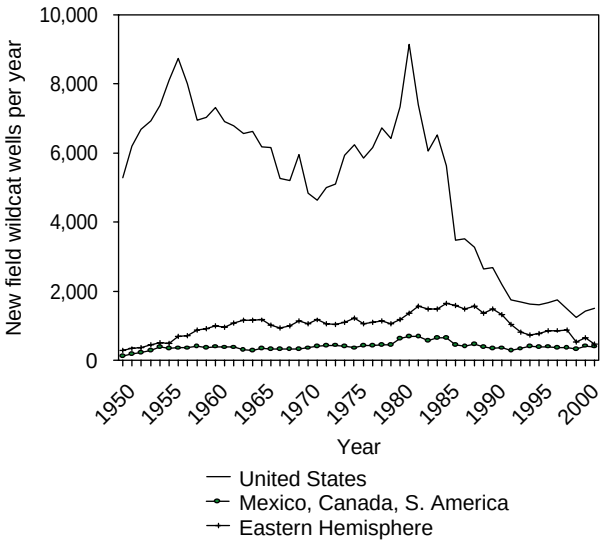


FIGURE 8 Time profile of new-field wildcat wells drilled in the United States, Eastern Hemisphere countries, and the group consisting of South America, Canada, and Mexico from 1950 to 2000. Data from I.H.S. Energy Group, International Petroleum Exploration Production Data Base, 2002, 2002 API Petroleum Data Book.

hydrocarbons that are technically recoverable without reference to being commercial. The discovery data presented in Figs. 9 and 10 must be viewed with this caveat in mind.

Figure 9 shows the annual oil discovery rate, averaged over 5-year intervals, that started in 1915 and annual production. The reported discoveries are in fields that can be produced with conventional production methods. The dominance of the Eastern Hemisphere is striking. It is likely that new discoveries will be revised upward as these fields are delineated and developed, so the sharp decline in overall discoveries is likely to be moderated. The peak levels of discovery from 1955 to 1965, however, will not be repeated. During this period, the Middle East accounted for about half of the discoveries worldwide. World production declined in response to the economic recession between 1980 and 1985, but it began to increase after 1986. It now appears to substantially exceed the past two decades of discoveries.

Figure 10 shows the natural gas discovery rate and production for the world. The Eastern Hemisphere discoveries dominate. The best years for discoveries of gas worldwide were from 1965 to 1970. During this period, the former Soviet Union accounted for two-thirds of all gas discoveries. During the 1970s, many of the largest gas discoveries were in the offshore areas of the North Sea and the Middle East. Figure 10 shows that gas production accelerated

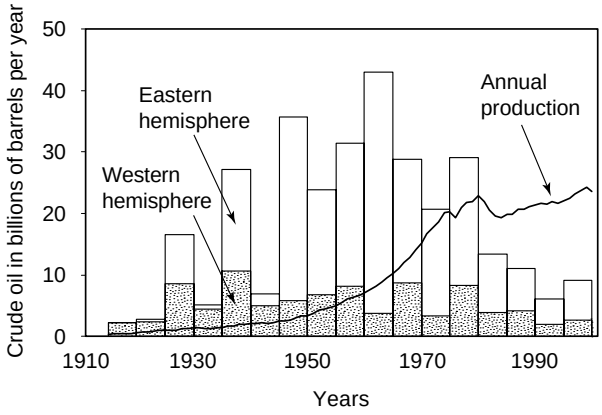


FIGURE 9 Estimated world annual crude oil discovery rate averaged over 5-year periods and annual production. Bottom section of bar is Western Hemisphere and top is Eastern Hemisphere. From U.S. Energy Information Administration, I.H.S. Energy Group, International Petroleum Exploration and Production Data Base, 2002, and Canadian Association of Petroleum Producers, *Statistical Handbook,* 2002.

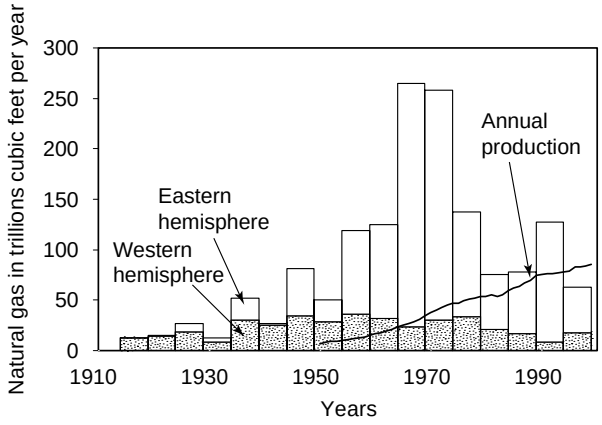


FIGURE 10 Estimated world annual natural gas discovery rate averages over 5-year periods and annual production. Bottom section of bar is Western Hemisphere and top is Eastern Hemisphere. From U.S. Energy Information Administration, I.H.S. Energy Group, International Petroleum Exploration and Production Data Base, 2002, and Canadian Association of Petroleum Producers, *Statistical Handbook,* 2002.

during the late 1980s. Estimates of the size of gas reserves are probably underreported because gas is not commercially marketable in many parts of the world. Consequently, the reported discovery rates shown in Fig. 10 are probably underestimated.

9. CONCLUSIONS

This discussion has identified the determinants of oil and gas exploration, considered the effects of

mineral property ownership on exploration rates, and examined empirical data showing the worldwide oil and gas discovery rates. Because of the way oil and gas occurs in nature, the statistics generated by the discovery process provide the basis for predicting the potential future availability of resource supply from conventional accumulations. As discovery rates inevitably decline and discovery costs increase, resources that are brought to market will, in all likelihood, come increasingly from oil and gas accumulations that are unconventional.

SEE ALSO THE FOLLOWING ARTICLES

Markets for Natural Gas • Markets for Petroleum • Natural Gas Resources, Global Distribution of • Oil and Natural Gas Drilling • Oil and Natural Gas Exploration • Oil and Natural Gas Leasing • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Oil and Natural Gas Resource Assessment: Classifications and Terminology • Oil-Led Development: Social, Political, and Economic Consequences • Oil Price Volatility • Petroleum Property Valuation

Further Reading

Adelman, M. A. (1992). Finding and development costs in the United States 1945–1986. *In* "Advances in the Economics of Energy and Resources" (J. Maroney, Ed.), pp. 11–58. JAI Press, Greenwich, CT.

Adelman, M. A. (1993). Modeling world oil supply. *Energy J.* **14,** 1–32.

Adelman, M. A., De Silva, H., and Koehn, F. (1991). User cost in oil production. *Res. Energy* **13,** 217–240.

Adelman, M. A., and Watkins, G. C. (1997). The value of United States oil and gas reserves: Estimation and application. *In* "Advances in the Economics of Energy and Resources" (J. Moroney, Ed.), pp. 131–183. JAI Press, Greenwich, CT.

Attanasi, E. D., and Root, D. H. (1995). Petroleum reserves (oil and gas reserves). *In* "Encyclopedia of Energy Technology and the Environment," pp. 2264–2277. John Wiley and Sons, New York.

Bohi, D. R., and Toman, M. A. (1984). Analyzing nonrenewable resource supply. Resources for the Future, Washington DC.

Devarajan, S., and Fisher, A. (1982). Exploration and scarcity. *J. Pol. Econ.* **90,** 1279–1290.

Drew, L. J. (1990). "Oil and Gas Forecasting." Oxford University Press, New York.

Drew, L. J. (1997). "Undiscovered Petroleum and Mineral Resources: Assessment and Controversy." Plenum, New York.

Energy Information Administration (2002). U.S. Crude oil, natural gas, and natural gas liquids reserves, 2002, Annual Report. DOE/EIA-0216(03).

Energy Information Administration (2003). Performance profiles of major energy producers, 2001, January, DOE/EIA00206(01) at http://www.eia.doe.gov/emeu/perpro/perfpro2001.pdf.

Hotelling, H. (1931). The economics of exhaustible resources. *J. Pol. Econ.* **39,** 137–175.

Klett, T. R., Ahlbrandt, T. S., Schmoker, J. W., and Dolton, G. L. (1997). Ranking of the world's oil and gas provinces by known petroleum volumes. U.S. Geological Survey Open File Report 97–463, one CD-ROM. Washington, DC.

McDonald, S. L. (1994). The Hotelling principle and in-ground values of oil reserves. *Energy J.* **15,** 1–17.

Securities and Exchange Commission (1981). Regulation S-X Rule 40–10, Financial Accounting and Reporting Oil and Gas Producing Activities, Securities and Exchange Commission Reserves Definitions, Bowne & Co. Inc., March 1981, New York.

Society of Petroleum Engineers (1987). Definitions for oil and gas reserves, Journal of Petroleum Technology, May 1987, pp. 557–578.

U.S. Geological Survey Assessment Team (1995). 1995 National Assessment of United States Oil and Gas Resources. U.S. Geological Survey Circular 1118. Washington, DC.

Yergin, D. (1991). "The Prize." Simon & Shuster, New York.

Oil and Natural Gas Exploration

MARLAN W. DOWNEY
Southern Methodist University
Dallas, Texas, United States

1. General Characteristics of Oil and Natural Gas Exploration
2. Fundamental Concepts Guiding Exploration for Oil and Natural Gas
3. Initiation of Exploration
4. Regional Studies
5. Prospect Definition
6. Search Techniques for Oil and Natural Gas
7. Assessing Risk in Exploration
8. The Results: Evaluating Exploration Effectiveness
9. Summary

Glossary

field A commercially valuable subsurface accumulation of hydrocarbons.
permeability A measure of the transmissibility of fluids in the void space in a rock.
petroleum A naturally occurring mixture of liquid hydrocarbons.
porosity Percentage of void space in rock volume.
reservoir A rock layer with substantial porosity and permeability.
wildcat An exploration well (supposedly one drilled so far from civilization that the drillers can hear the cries of wildcats).

Exploration for oil and natural gas is a search for buried treasure, on a huge scale, for commercial purposes. Exploration is a business-oriented search for new hydrocarbon assets. Geology is a science; geophysics is a science; exploration is a business. It is not difficult to find traces of oil and natural gas; it is very difficult to find sufficient quantities of hydrocarbons to make a profit.

1. GENERAL CHARACTERISTICS OF OIL AND NATURAL GAS EXPLORATION

A romantic image of exploration is one of hardy drillers risking fortunes on the luck of a wildcat well. The large risks and the financial dangers are still very much a part of modern exploration, but all the tools of modern technology are coupled with the strengths of business analysis to quantify and diminish the risks of exploration.

Exploration accomplishments come with high costs and high risks. Some idea of the risk of exploration investments, relative to other investments, may be gained by understanding that banks do not loan money for exploration investments. Hardy investors in exploration oil and gas projects are effectively lenders-of-last-resort, and such investors demand the potential for high returns to balance the exceptional risks. When a company drills a $20 million "dry hole," a well that fails to find significant oil or gas, there is no salvage value and no optional use for the dry hole: It is a $20 million loss.

In a typical year, approximately 10 billion barrels of oil and 70 trillion cubic feet of natural gas are discovered in the world. Assuming a value of $25 per barrel for oil and $3.00 for 1000 cubic feet of gas, it can be seen that $460 billion dollars worth of new assets are discovered each year by oil and gas exploration. The rewards for exploration success can be huge, but costs are high and failure is common.

The exploration task is far greater than discovering new subsurface reservoirs of hydrocarbons; it must be that of discovering those hydrocarbons in an economically efficient manner. Successful oil and gas exploration owes as much to efficient business practices as to technical excellence.

The first step in exploration is always an innovative idea about where oil or gas might be found. The pioneering geologist Wallace Pratt once said, "Oil is first found in the minds of men." It is hoped that such

 © 2004 Elsevier Inc. All rights reserved.

ideas arise from fundamental knowledge of the rock layers in the earth, their suitability for generating and retaining oil and gas accumulations, and an analysis of the clues available to suggest that proper conditions are actually present in a particular area. Oil and gas accumulations in the earth are extremely valuable. The discovery of new accumulations (fields) generates great wealth. Oil and gas fields are very difficult to find, and exploration—a systematic search for new oil and gas accumulations—depends on proper use of technology, acquisition of new data, and the deductive analysis of earth clues in a manner reminiscent of a detective.

2. FUNDAMENTAL CONCEPTS GUIDING EXPLORATION FOR OIL AND NATURAL GAS

All oil and much natural gas are created by deep burial and heating of organic matter contained within sedimentary rocks. As this organic matter is converted to oil and gas by burial and heating in the earth, oil and gas are forced out of the organic-rich sedimentary layers and expelled into open fractures or porous transmissive zones. These fractures and transmissive zones in the earth provide paths for hydrocarbon migration. The oil and gas may be retained and trapped in subsurface reservoirs or may continue to leak upward to the surface of the earth. Such leaks of hydrocarbons to the surface have been forming tar pits, oil lakes, and gas seeps for millions of years.

If the migrating hydrocarbons encounter an impervious layer of rock, they will be prevented from escaping to the surface. If the impervious layer roofs a rock layer with porosity, the migrating petroleum may fill the pore space of the reservoir and displace some of the water originally filling the pores of the rock. If the roofing layer and the associated reservoir rock have been folded in a domal shape, the hydrocarbons may fill the container formed by the domed surface. The oil, natural gas, and water in the reservoir strata have distinctly different densities. Gas floats on oil; oil floats on water. In a domal structure, then, the lighter gas will occupy the pore spaces in the uppermost part of the structure and be underlain by the oil, which in turn is underlain by water.

The two most fundamental search elements in exploration for oil and gas are(i) a knowledge of where oil and gas have been generated and provided to migration routes and (ii) the location of specific traps that could interrupt migration and retain petroleum.

3. INITIATION OF EXPLORATION

The exploration process can start with a new idea, new information, or a new area becoming available for search. If exploration is initiated because a new international search area becomes available for leasing and drilling, it is quite likely that the new search area will be leased to whomever bids the most for the right to explore. In such new areas, most bidders will have a similar set of information obtained from the host country. A fascinating area of operations research derives from the competitive aspects of bidding under uncertainty. An insight into this rich area of decision making is given by understanding that any winning bidder (the one who bid higher than anyone else) often bids more than the lease is worth. The more bidders, the more likely that the winning bidder has overbid. If 20 companies, with similar knowledge, submit bids for the undrilled area, what is the chance that the winning bidder (who values the undrilled area higher than 19 other knowledgeable bidders) has still managed to bid less than the true value of the lease? This chance ranges between slim and none. In the oil and gas business, this situation of winning the bid (but paying too much) is described by Capen as the winner's curse.

When exploration is initiated because of exclusive access to new information, the company with the new information has some remarkable advantages in finding and valuing new exploration projects. This competitive advantage is one of the reasons that large companies spend huge amounts of money in acquiring new data for their exclusive use and advantage.

Sometimes, exploration is initiated simply because a team or unique individual has extracted a novel concept, a new interpretation, from generally available data. Such individuals or teams are rare and extraordinarily valuable; they spawn new directions for exploration and give incredible competitive advantages to their employers. Such great oil finders are like master chess players, seeing more deeply, more comprehensively, than others.

4. REGIONAL STUDIES

Exploration for oil and gas in the earth can be usefully subdivided into two scales of effort: regional

evaluation and prospect definition. In the first effort, data are gathered to indicate whether specific regions of the earth are likely to contain hydrocarbons. In the second effort, an attempt is made to decide exactly where to drill to find oil or gas, within the general region that appears favorable for containing subsurface accumulations of hydrocarbons.

Regional studies attempt to answer the following questions: Where are the organic-rich layers in the subsurface? What is their richness and character? Where in the earth have they been buried to sufficient temperatures to generate oil and gas? What was the time at which the source rocks were heated and provided petroleum? What traps were available to retain the generated hydrocarbons at the time the hydrocarbons were generated and migrating?

Geologists sometimes use the phase "petroleum systems analysis" to describe the type of studies that must be made to determine whether a region should be expected to contain accumulations of hydrocarbons. This concept of the petroleum system, pioneered by Magoon and Dow, recognizes that accumulations of hydrocarbons in the earth require a special combination of unique circumstances before hydrocarbons can be generated and be available at the right time to fill thick porous reservoirs contained within traps. If the hydrocarbons are not available, the reservoirs and traps will be barren. If the porous reservoirs are absent, the traps will only contain hydrocarbons in rocks unsuitable for production. If the traps are not present, the hydrocarbons will move through the porous reservoir and continue to the surface of the earth. All of the necessary parameters must be present simultaneously at the time of generation and accumulation and persist until the present day.

One simple test for the presence of a working petroleum system is a documentation of a well in the area that flowed hydrocarbons to the bore hole. Although this flow of hydrocarbons may be trivial from a commercial standpoint, it is of great significance in indicating that a working petroleum system is likely. From such an observation, the presence of hydrocarbons trapped in a reservoir is demonstrated and not merely guessed at or hoped for.

In the final stage of regional studies, the geologist and the geochemist may make a mass balance calculation of the total amount of hydrocarbons expected to have been generated from the petroleum source rocks in an area to compare with the amount and type of hydrocarbons already discovered in an area. If these regional studies are promising, the task of the explorationist can shift to a search for specific "prospects" suitable for drilling and evaluation in the region.

5. PROSPECT DEFINITION

5.1 Prospect Varieties

There are two general varieties of hydrocarbon traps: structural and stratigraphic. However, many combinations of these trap types are known.

5.1.1 Structural Traps

Structural traps involve folds of subsurface rock layers that provide interruptions to the continued upward migration of oil and gas. If hydrocarbons are migrating in an uninterrupted porous reservoir, beneath a layer of sealing rock, they may move hundreds of miles laterally and up to the surface, unless collected along the way by a concave fold (anticline) in the earth layers. (In an analogous manner, on the surface of the earth, depressions in the earth's surface collect the flow of streams to provide the accumulations of water we know as lakes.) Each concave-downward trending fold attracts the hydrocarbons, which are buoyant and floating on the subsurface waters in the reservoirs. As each concave-downward fold is filled to capacity with the migrating hydrocarbons, the excess hydrocarbons may spill from an edge of the confining surface and continue to move upward.

Structural traps have two significant attractions to explorers: The structural fold is readily located by various techniques, and the structural fold perturbs the moving hydrocarbons and may act to focus and collect the buoyant hydrocarbons.

Another common type of structural trap, the fault trap, is found where alternating layers of reservoirs and seals are broken by faults that cause the reservoir layers to become offset against sealing layers on the up-dip side. These fault traps are important containers of hydrocarbons.

5.1.2 Stratigraphic Traps

Stratigraphic traps occur in great varieties and combinations. A large proportion of the world's oil and gas is contained in stratigraphic traps. Essentially, a stratigraphic trap results from a lateral change in the subsurface rock strata—for example, where a porous and permeable rock layer gradually becomes nonporous and impermeable in an up-dip direction. Hydrocarbons migrating up-dip in the porous strata will collect at the place where the

hydrocarbons are impeded by the seal. Stratigraphic traps are more complex and more difficult to find than structural traps, and they require special understanding of such reservoir characteristics as the lateral extent of the reservoir and its variance laterally and vertically. Whether hydrocarbons are trapped at the up-dip edge of a reservoir depends on the pore throat openings characterizing the up-dip barrier and on the buoyancy properties of the migrating hydrocarbons. The buoyancy pressure of the hydrocarbon column is dependent on the density difference between the hydrocarbon and water and on the height of the hydrocarbon column. If increasingly more hydrocarbons migrate into a stratigraphic trap, increasing the height of the hydrocarbon column and its buoyancy pressure, the hydrocarbons may eventually be forced into the fine pore throats of the up-dip barrier and move through it.

As Downey pointed out, it is easy to find conditions in which tiny quantities of hydrocarbons are entrapped because small accumulations lack buoyancy pressure to overcome even minor barriers. It becomes progressively more difficult to trap larger accumulations of hydrocarbons because large columns of hydrocarbons exert large buoyancy forces against the pore throats of the potential up-dip seal. Seal risk becomes more important as larger fields are sought.

6. SEARCH TECHNIQUES FOR OIL AND NATURAL GAS

6.1 Geophysical Prospecting Techniques

6.1.1 Seismic Surveys

Prospects (hypothesized accumulations) are most commonly located and described by geophysical techniques. The most useful of these geophysical techniques involves penetrating the rock layers with acoustic waves and recording, at the surface, the reflection of these acoustic waves from subsurface rock interfaces. These seismic surveys require immense compute power and technical sophistication, but the general approach is not unlike the use of sonar to find schools of fish or to depict the form of the sea bottom beneath a boat. Generally, the structural form, strata thicknesses, and the general characteristics of the subsurface rocks can readily be determined from information provided by seismic investigations.

In the past, seismic data were obtained by using linear tracks, providing two-dimensional (2D) data; modern seismic work favors collection and analysis of data in three dimensions, giving unrivaled detailed pictures of the subsurface. Of course, the cost to acquire 3D data is much more than that of the 2D data set, and careful analysis must be made to ensure that the value obtained from the 3D data set justifies its costs. In addition, a 3D set may be repeated over time to monitor changes in seismic reflection character. Such 3D data sets are called 4D sets and can be very useful for review of *in situ* changes in field characteristics. The depletion of hydrocarbons in various segments of a field can be monitored over time and extraction efficiencies maximized.

In favorable circumstances, the acoustical data collected from the seismic work may directly indicate whether the subsurface layers contain oil and gas. The discovery of seismic techniques (bright spots) for direct detection of underground oil and gas by M. C. Forrest of Shell Oil in the late 1960s revolutionized exploration.

6.1.2 Gravity and Magnetic Surveys

Gravity surveys provide measurements of variations in the earth's gravity at a number of locations in a region. These gravity variations represent changes in the density of the rock column under the measuring site and are helpful as a quick, inexpensive way of recognizing the presence of thick sections of sedimentary rock. Gravity surveys are most helpful in regional studies, but in special cases, such as for areas where prospects are created by large, buoyant salt domes, gravity can be very useful to define the distribution and structural attitude of the salt masses, which are less dense than the adjacent rocks.

Magnetic surveys are also helpful in regional studies to provide an estimate of depth to magnetic basement. On the prospect scale, magnetic investigations can readily reveal areas of igneous intrusion; the igneous dikes and sills appear as strong magnetic anomalies.

6.2 Surface Geochemical Techniques

Surface measurements of hydrocarbons are often a useful supplement to other exploration search techniques. The presumption is that tiny leaks of hydrocarbons are occurring above subsurface oil and gas accumulations, and with proper collection and analysis of soils and soil gases these microseeps can be recognized and used to pinpoint the subsurface origin of the leaks. The problem with the proper

utilization of surface geochemical techniques is not one of adequate sensitivity but one of differentiating hydrocarbons leaking from buried accumulations from hydrocarbons abundantly generated by plants and living organisms and hydrocarbons provided by surface contamination.

6.3 Geologic Techniques

Geological techniques in exploration have changed greatly during the past 50 years. Originally, the geologist searched surface exposures for oil and gas shows and mapped folds of surface rocks. Later, the geologist emphasized subsurface data such as recognition of oil and gas shows from rock cuttings from wells, identified the reservoir characteristics of subsurface layers, and integrated geophysical data with electric well surveys. The modern geologist has added expertise in organic geochemistry, geophysics, petrophysics, and economics. One of the greatest values of a modern petroleum geologist is his or her ability to visualize the subsurface in three dimensions using a knowledge of rock facies, environments of deposition, and sequence stratigraphy so as to reason from negative data (where oil is not) where oil should be.

6.4 Dowsing Techniques

Dowsing and other prospecting techniques calling on supernatural powers or divine insights are of absolutely no value. Dowsers say that they can feel emanations from oil and gas fields, underground water currents, and deposits of precious metals. There are many practitioners, devotees, and believers, but there is no science. No such technique has ever been shown to work at levels better than pure chance in any controlled trial. The Randi Institute maintains a $1 million prize to anyone who can demonstrate such remarkable powers in controlled circumstances. It remains uncollected.

6.5 Exploration Drilling Techniques

Drilling is costly and is generally reserved to test the best of the exploration prospects. Drilling of the exploratory test wells, the wildcats, is typically the most expensive procedure in the exploration cycle. Drilling represents the ultimate test of the oil finders' hypothesis that the mapped prospect may contain hydrocarbons.

Onshore drilling is generally accomplished with a moveable drilling rig. The drilling rig has three basic components: the derrick, which provides height for hoisting the drill pipe as the well is drilled; the powered rotary table, which clasps and turns the drill pipe with its attached drill bit; and the "mud" pumps, which provide power to force the mud through the drill pipe and drill bit. The drilling mud is really a sophisticated mixture of expensive chemicals designed to create a temporary skin for the walls of the newly drilled hole, act as coolant to the drill bit, and to provide adjustable pressure control for restraining any encountered earth pressures. The mud is carefully weighted with additives, such as barium carbonate, so that the weight of the column of mud in the drill pipe is sufficient to hold down any earth pressures encountered and prevent a blowout.

The procedure for drilling a hole into the earth involves rotating a dangling hollow drill pipe, which has a perforated bit at the end. As the upper end of the drill pipe is held firmly within the rotary table on the drilling floor, the drill bit on the down-hole end of the drill pipe rotates and grinds on the rocks on the bottom of the hole. Mud is pumped at high pressure down the hollow pipe and out the orifices in the bit at the bottom of the hole. The mud jets out of the orifices in the drill bit and assists in the drilling process. On the return trip up the annulus outside the drill pipe, the mud serves to carry out the rock fragments created by the drilling process. These rock chips are screened from the mud system at the surface, rinsed, and examined by the well-site geologist for rock characteristics and oil shows. The cleansed mud is carefully checked to ensure that its physical and chemical properties are exactly what is required, and then it is recirculated down the drill pipe.

As the drill bit becomes worn, the entire drill pipe must be hoisted from the hole; each 90-ft section of pipe is unscrewed from the upper end of the dangling drill pipe and stacked in the derrick. When the bottom end of the pipe is pulled onto the rig floor, the old bit is removed, a new bit is added, and the 90-ft sections of pipe are sequentially reattached to the upper end of the drill string as the bit is lowered to the bottom to continue drilling. In the colorful language of the oil fields, this long and difficult procedure is called "making a trip."

If the well has been drilling at 15,000 ft below the derrick floor, the round trip involves 332 separate connection activities by the roughnecks, the skilled drilling team.

The fundamental mechanics of drilling an oil well are more akin to rotating a weight on a string than pressing on a brace and bit because the miles-long

drill pipe has enormous weight but little compressive strength. At specific planned depths, the drill pipe and bit are hoisted out of the hole and long lengths of steel casing are lowered and set in the hole. Cement is pumped down the hollow casing until it flows out the bottom of the casing and up the annulus between the casing and the earth layers. After the cement cures, it bonds the outside of the casing to the bore hole walls. The steel casing can now protect the drilling operation from influxes of fluids or rocks caving from previously penetrated layers of rock. The drill pipe and bit are lowered into the hole again and drilling is recommenced, with the upper portion of the hole "cased off" and protected. This casing procedure may be repeated many times during the drilling of the well, with each set of casing nestling inside the others.

Offshore drilling utilizes the same basic drilling procedures as those of onshore drilling but involves a number of special and expensive modifications. Offshore drilling requires a secure place to set the drilling rig as well as space for the staff, equipment, and supplies.

In waters shallower than 600 ft, it is common to use a large floating barge, called a jack-up, with long legs on each corner. The jack-up is towed into position with the legs elevated into the air. After the jack-up is positioned over the prospect, the legs are jacked down into the sea bottom, and the barge with the drilling rig is lifted and supported above the water on the legs. After drilling is completed, the legs are jacked up, the barge is let down to float on the sea, and the barge is towed to another location.

In deep waters (1000–12,000 ft), very large quantities of oil and gas must be discovered by the explorers to pay for the immense costs of drilling. Large deep-water drilling vessels can cost more than $500 million to construct. Ordinary expenses to operate the deep-water drilling program can easily be $250,000 a day. Costs to drill wildcat wells in deep water have sometimes exceeded $100 million per well.

The drilling rig for deep water is installed on a specially designed mobile vessel, providing crew quarters and containing all necessary equipment for months of work. Such drilling vessels may have numerous thrusters for lateral positioning and be capable of automatically adjusting their position to keep the drilling rig directly over the drilling target. The drill bit and drill pipe are lowered thousands of feet to the sea floor, and drilling is commenced toward targets that may be tens of thousands of feet below the sea floor.

7. ASSESSING RISK IN EXPLORATION

The major reason for failure in exploration is misunderstanding risk. Most exploration ventures fail. Proper selection of exploration projects can lessen the average risk in the portfolio. Selection of exploration investments involves an estimation of the exploration costs to determine success, an estimation of the probability of success, and the value of the project, given success. It is generally understood that proper estimation of the probability of success is one of the most important elements in maintaining a high-quality portfolio of exploration projects. Probabilities of success for various exploration ventures in a portfolio can range over three orders of magnitude. These estimations of probability of success must be made in a consistent manner, with geologic and mathematical rigor.

A number of approaches to estimating probability of success are utilized by various groups. In all cases, the historical wildcat success rates in a particular province should be compiled for comparison to any estimates of future success. Past history does not define future probabilities of success, but historical knowledge acts to constrain wildly optimistic forecasts.

The most useful general approach is outlined by Capen and is easy to understand because it mimics a board game with a "wheel of chance" and a spinner. For each prospect risk analysis, the wheel of chance is divided into two segments, whose areas denote the historical probability that a wildcat would or would not expect to encounter an accumulation of hydrocarbons in the region being analyzed. If one-fourth of the wildcats in the region have encountered some sort of accumulation of hydrocarbons, one-fourth of the wheel of chance would be noted as "oil and gas," and three-fourths of the wheel would be identified as "dry hole." One time in four, the spinner, representing chance, would land on the quarter circle called oil and gas. Although this example indicates that one wildcat out of four might be expected to find an oil and gas accumulation, remember that most oil and gas accumulations in the earth are very small.

Oil and gas accumulations have a nearly log-normal frequency distribution in the earth. As a result, we can divide the quarter circle area of oil and gas with a log scale along the quarter circle circumference, and estimate the differing likelihood that the spinner might alight on the segment (probability) of 500 million barrels versus the segment (probability) of a 10-million barrel or smaller field (Fig. 1).

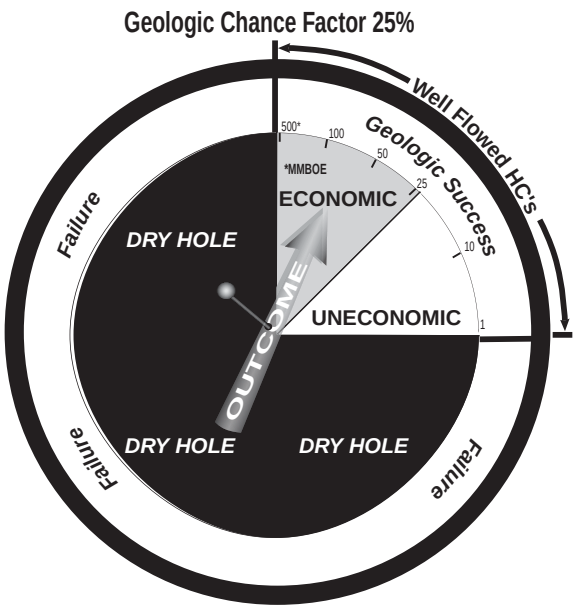


FIGURE 1 "Wheel of chance" for assessing probability of success in an exploration drilling venture; after Capen (1992).

Lastly, we can determine the minimum size of accumulation that can be commercially developed. If we are interested in making money in oil and gas exploration, we must analyze the probability of finding a field at least as large as the minimum developable size. If our minimum developable field size is 10 million barrels of oil, then we need to determine the probability that our next wildcat will encounter at least 10 million barrels of oil. Using the probability of commercial success, the expected cost of discovery, and the present value worth of the target, we can estimate the risk discounted present value worth of the exploration project.

7.1 Contractual Risk in Exploration

At the inception of exploration, an agreement must be negotiated that provides for the recovery of costs, taxation, and the share of the production retained by the explorer. These agreements or production contracts typically extend for the life of the production. In all cases outside the United States in which significant exploration finds have been made, the contract has been unilaterally altered in favor of the host country at some time during the life of the contract. Even within the United States, modifications to previous oil and gas contracts with state and federal governments are common. This certainty of some degree of contract abrogation is a fact of life and another risk to be assessed by the oil and gas investor.

8. THE RESULTS: EVALUATING EXPLORATION EFFECTIVENESS

8.1 Reserve Replacement

Large oil and gas companies can be considered as powerful search engines for the discovery of new assets each year. A large oil company that produces 500 million barrels of oil each year to supply its refinery and marketing system should find at least 500 million barrels of new oil each year or its asset base will diminish and the value of its shares may fall. The ratio of new reserves found to those produced is called the reserve replacement ratio, and it is closely monitored as one indication of the efficiency of the exploration process for individual companies.

8.2 SEC Finding Costs

Each publicly traded company in the United States is required to provide a yearly report to the Securities and Exchange Commission (SEC) that includes specific data on current-year cost of finding oil and gas. These company reports to the SEC are greatly relied on by analysts but can be very misleading. This is because the SEC finding cost reports mandate a comparison between the amount of oil and gas volumes "booked" each year by each company compared to the amount of money that each company spent in the current year on exploration. The problem for the unwary analyst is that the volumes booked in the current year may have been found many years ago with past exploration expenditures; the exploration expenditures of the current year may have found volumes that will not be booked as producible reserves until several years in the future. The oft-quoted SEC exploration finding cost, dollars spent on exploration in the current year divided by volumes booked in the current year, is an unreliable and often meaningless number. Exploration efficiency is extremely important to know and very difficult to decipher from such published information.

8.3 Internal Measures of Finding Costs

The internal finding costs of a company are determined by comparing each year's exploration expenditures to the volumes thought to have been discovered during the same year. These internal exploration finding costs, properly created and regularly monitored, are a very useful monitor of exploration efficiency.

8.4 Financial Measures of Exploration Efficiency

The perfect method for evaluating exploration efficiency might seem to be a simple comparison of the money spent each year on exploration versus the present value worth of the petroleum assets discovered each year. An indication of the problems with this method is suggested if we rephrase the financial measure more precisely and state, "a comparison of the money spent each year on exploration versus an estimate of the present value worth of petroleum assets discovered." Much of the apparent rigor of the financial measure disappears when it is realized that the financial measure depends on guesses as to oil and gas prices many years in the future, which are multiplied by guesses as to the amount and timing of the future production.

8.5 The "Gold" Standard: Risk-Discounted Expectation versus Actual

The best technique for monitoring exploration efficiency (the gold standard) is a cumulative comparison of the risk-discounted volume expectation of each wildcat versus the actual volume outcome. Exploration efficiency can be easily monitored by a cumulative plot of barrels predicted to be found (and barrels actually found) recorded against the wildcat well sequence. This graph provides two cumulative curves: the volumes expected to have been found and the volumes actually found. Such a plot also illustrates the need for an adequate number of wildcat chances to obtain success. Over any significant period of time, the sum of the predicted volumes must match the actual outcomes. This gold standard is a rigorous test that compares what the exploration team promised it was going to find versus actual outcome over time (Figs. 2 and 3).

8.6 Finding and Development Costs

One of the most useful measures of the performance of an oil company over a long period is given by the company's average costs to find and produce oil and gas fields on a per barrel basis. Companies with low finding and development costs have a considerable advantage in the marketplace. Prices received by the producer for its oil and gas may increase or decrease, but companies that are able to deliver oil and gas at low finding and development costs are always the strongest.

Current finding and development costs represent the outcomes of decisions and actions made many months, perhaps many years, earlier. Low finding and development costs reflect well on the company and, to a lesser degree, on the current management.

8.7 Reserve Appreciation

A final caution in evaluating exploration efficiency is more subtle to understand and more powerful in its effect: What volumes are actually being discovered by current exploration? Numerous "hindcasting" studies of exploration have demonstrated that the oil volumes thought to have been

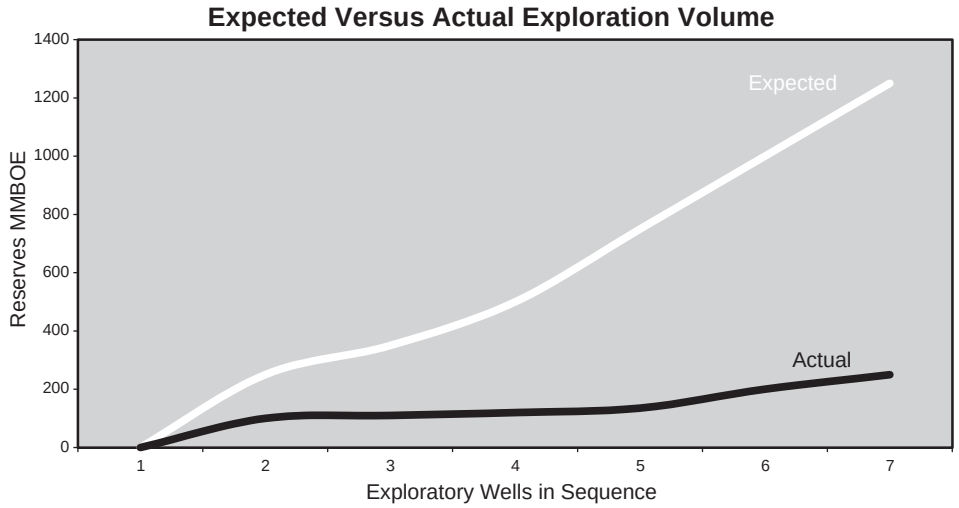


FIGURE 2 Volumes predicted to be discovered versus volumes actually discovered; unsuccessful exploration portfolio.

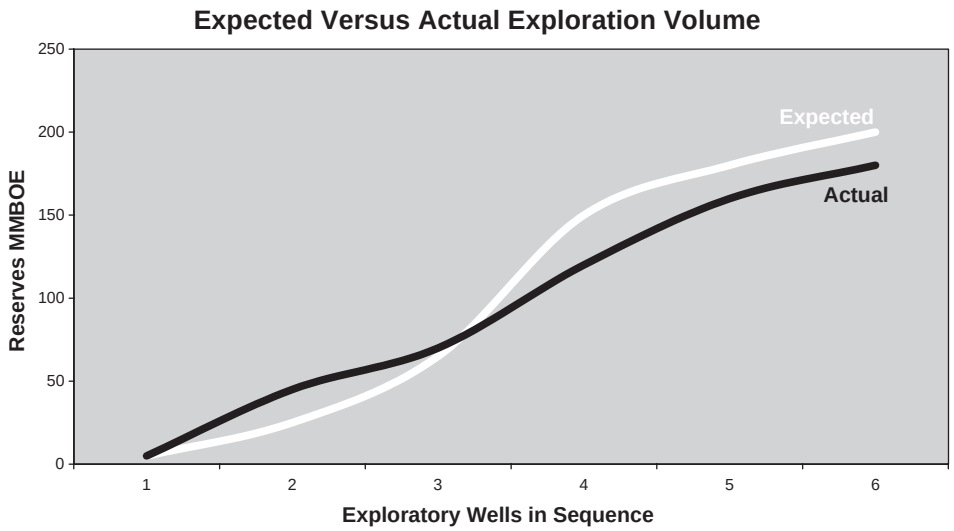


FIGURE 3 Volumes predicted to be discovered versus volumes actually discovered; improved portfolio.

found by exploration continue to increase with time and field production, until the ultimate production from oil fields is seen to dwarf the early, ultraconservative estimates made at the time the fields were discovered. How great is this underestimation of exploration discoveries as a fraction of ultimate oil recovery? On an overall basis, the early reserve estimations appear to more than triple after long-term production gives a revised picture of ultimate oil reserves. (Techniques for determining natural gas volumes are far more accurate and early exploration estimates appear to understate ultimate gas reserves by only 15–20%). Why is this understatement of new reserves so prevalent? Two explanations may suffice: one economic and one technical.

At the time of first development of a newly discovered field, huge sums of money must be invested before there is any production or payback. A 100-million barrel oil field may require $200–800 million for drilling wells and providing facilities for production. Estimates of field size, made at this early time for the purpose of justifying development costs, must be very conservative because company management must guarantee sufficient reserves to cover the projected development expenses. Such early estimates should properly be considered as being 95% certain of at least the volumes, rather than expectation volumes.

Second, once the wells are drilled, the field facilities are in place, and costs recovered, incremental volumes become available from secondary, higher cost hydrocarbon-bearing horizons, whose volumes often dwarf those of the originally productive horizons. The ultimate effect of this general understatement of discovery volumes is to underestimate the value of the assets discovered each year by exploration.

8.8 Size Distribution of Oil and Gas Fields

In any sufficiently large area where oil and gas fields occur, the size of the individual oil and gas fields will be found to conform to a near log-normal size distribution. In lay terms, this means that there will be very few large fields, a substantial number of medium fields, and a great number of very small fields. This remarkably consistent outcome is a result of some subtle mathematics. Each oil field depends on the simultaneous occurrence of a number of parameters, such as the amount of hydrocarbon charge available, suitable reservoirs to contain the hydrocarbons, a trap configuration to hold the hydrocarbons, a seal to confine the hydrocarbons, and sufficient permeability to allow the hydrocarbons to flow to the surface at commercial rates. The log-normal field size distribution, ubiquitous in nature, is understandable as a result of multiplication of the probability distributions describing the critical parameters.

9. SUMMARY

Finding traces of oil and gas is not difficult; making money finding oil and gas is very difficult. Exploration

is the business of efficiently finding oil and gas so that the new discoveries can be developed and produced. Successful exploration requires a fundamental knowledge of how oil and gas are generated and how they migrate and accumulate. A number of search techniques are available to explorers and the proper use of these technologies is the responsibility of the technical staff in the oil companies. In addition, oil and gas exploration must be conducted in a cost-efficient manner so that the exploration product, the field discoveries, will be highly profitable and will properly reward the oil company for its assumption of large financial risk.

SEE ALSO THE FOLLOWING ARTICLES

Markets for Natural Gas • Markets for Petroleum • Natural Gas Resources, Global Distribution of • Oil and Natural Gas Drilling • Oil and Natural Gas: Economics of Exploration • Oil and Natural Gas Leasing • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Petroleum Property Valuation • Risk Analysis Applied to Energy Systems

Further Reading

Attanasi, E. D., and Root, D. H. (1994). The enigma of oil and gas field growth. *Am. Assoc. Petroleum Geol. Bull.* **78,** 321–333.

Capen, E. C. (1992). Dealing with exploration uncertainty: Business of petroleum geology, AAPG Treatise of Petroleum Geology. American Association of Petroleum Geologists, Tulsa, OK.

Downey, M. W. (1984). Evaluating seals for petroleum accumulations. *Am. Assoc. Petroleum Geol. Bull.* **68**(11), 1752–1763.

Forrest, M. C. (2000, May). "Bright" ideas still needed persistence. *Am. Assoc. Petroleum Geol. Explorer,* 4–12.

Magoon, L. B., and Dow, W. G. (1994). The petroleum system: From source to trap, AAPG Memoir 60. American Association of Petroleum Geologists, Tulsa, OK.

Oil and Natural Gas Leasing

DENNIS D. MURAOKA
California State University, Channel Islands
Camarillo, California, United States

1. **Current and Historical Acreage and Production from Federal Oil and Gas Leases in the United States**
2. **The Legal Framework for Federal Oil and Gas Leasing**
3. **Federal Oil and Gas Leasing Procedures**
4. **The Effectiveness of Federal Oil and Gas Leasing Procedures**
5. **State and Local Government and Environmental Concerns Regarding Federal Oil and Gas Leasing**

Glossary

bonus payment A one-time, nonrefundable cash payment made by a lessee to the lessor at the time a lease is issued. It is usually the bid variable for competitive leases.

competitive lease A lease issued in an area believed to contain crude oil or natural gas. Such leases are auctioned to the highest qualified bidder.

economic rent The difference between the total revenue generated by a lease and the total necessary cost needed to generate this revenue.

noncompetitive lease A lease issued for lands not believed to contain crude oil and natural gas.

outer continental shelf The shallow portion of the continental margin that extends from the coastline to the more steeply inclined continental slope. These offshore lands fall under the jurisdiction of the federal or state governments.

primary lease term The time period specified in a lease wherein the lessee has the exclusive right to explore for crude oil or natural gas.

rental payment An annual payment by the lessee to the lessor determined on a per acre basis from the time the lease is issued to keep the lease in force.

royalty payment A payment in value or in kind by a lessee to the lessor calculated as a percentage of the gross income from a lease.

transfer payment A payment that redistributes wealth from one sector or group within the economy to another.

Federal, state, and local governments own and manage over 30% of the surface area of the United States. In addition, government agencies oversee vast offshore lands adjacent to the U.S. coastline. These holdings place vast quantities of crude oil, natural gas, and other natural resources under public stewardship. Rather than develop these resources as public sector enterprises, government agencies have typically transferred subsurface mineral rights to the private sector while retaining the ownership of the public lands. The contract used to convey these rights is a lease. In the case of a crude oil and natural gas lease, the government (the lessor) conveys to individuals or businesses in the private sector (the lessees) the right to explore for and extract the oil and gas from beneath the government lands. In return, the lessees agree to compensate the government with a series of payments described in the lease a greement. The leasing procedures described below pertain to federal oil and gas leases. State and local government leasing procedures vary widely. Nevertheless, they include the same or similar components as their federal counterpart.

1. CURRENT AND HISTORICAL ACREAGE AND PRODUCTION FROM FEDERAL OIL AND GAS LEASES IN THE UNITED STATES

As of December 31, 2000, the federal government had contracted with the private sector for oil and gas production from federal lands in 31 states (Table I). There were 23,844 producing or producible leases covering an area of 12.6 million acres of federal and Native American lands. The average federal onshore lease covered 530 acres. In addition, the federal government had lease contracts for submerged lands in the Gulf of Mexico and lands offshore from California and Alaska. There were 1833 leases offshore leases covering 8.7 million acres of offshore land. The average offshore lease was 4749 acres—much larger than the average onshore lease.

 © 2004 Elsevier Inc. All rights reserved.

TABLE I

Federal Onshore and Offshore Oil and Gas Leases as of December 31, 2000

Onshore	Number of producing and producible leases	Acreage leased (acres)
Alabama	24	13,822
Alaska	36	64,178
Arizona	15	69,096
Arkansas	192	90,092
California	323	77,478
Colorado	2193	1,893,447
Florida	3	3476
Illinois	8	1581
Kansas	446	122,597
Kentucky	61	35,481
Louisiana	223	79,222
Maryland	2	34,941
Michigan	70	78,800
Mississippi	94	44,245
Missouri	1	200
Montana	1756	883,473
Nebraska	25	37,275
Nevada	37	23,124
New Mexico	6464	3,989,351
New York	4	1009
North Dakota	572	320,969
Ohio	167	30,472
Oklahoma	2579	272,268
Pennsylvania	60	24,680
South Dakota	76	34,777
Tennessee	7	2446
Texas	207	110,678
Utah	1931	1,129,060
Virginia	15	11,119
West Virginia	159	149,693
Wyoming	6094	3,017,958
Onshore subtotal	23,844	12,647,008
Offshore		
Alaska	4	20,481
California	43	217,668
Gulf of Mexico	1786	8,467,355
Offshore subtotal	1833	8,705,504
Onshore and offshore total	25,677	21,352,512

Source. Reprinted from U.S. Department of the Interior (2001). "Mineral Revenues 2000," Minerals Management Service, Herndon, VA.

Federal oil and gas leases contribute over one-third of the crude oil and natural gas produced in the United States. Federal leases produced 689 million barrels of oil or approximately 32.4% of the total production in the United States in 2000 (Fig. 1). Similarly, federal leases produced 7.1 billion MCF (thousand cubic feet) of gas or approximately 35.5% of the United States total production of gas in 2000 (Fig. 2). To put these production totals into perspective, the population of the United States in 2000 was 281.1 million people. Thus, in the year 2000, federal leases produced approximately 2.5 barrels of oil and 25 MCF of gas for every person in the United States.

The cumulative production from federal oil and gas leases is enormous. From 1920 to 2000, the production of crude oil from federal leases totaled 23.7 billion barrels. The total value of this production was $273.2 billion. Natural gas production from federal leases totaled 199.1 billion MCF with a total value of $305.5 billion.

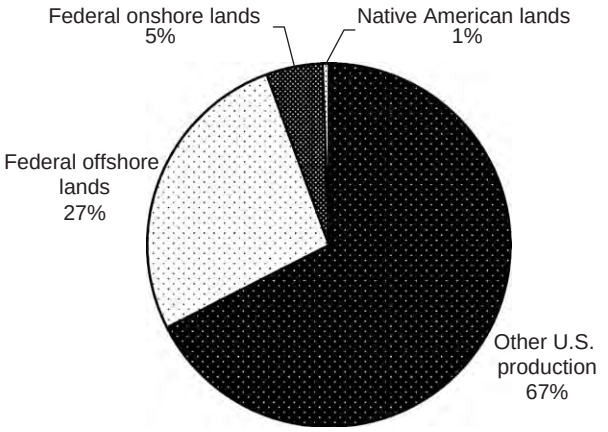


FIGURE 1 Crude oil production in the United States in 2000.

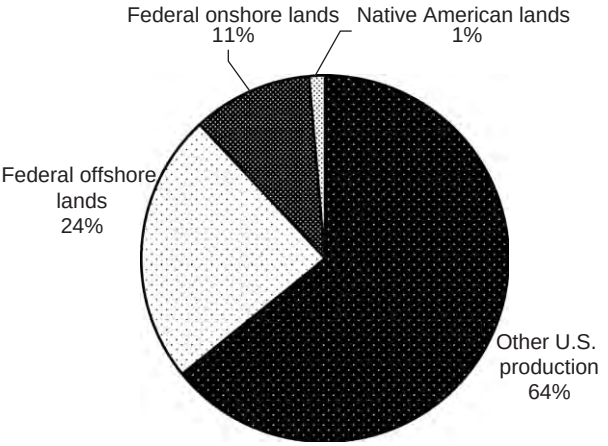


FIGURE 2 Natural gas production in the United States in 2000.

2. THE LEGAL FRAMEWORK FOR FEDERAL OIL AND GAS LEASING

The push to develop federal lands for oil and gas production followed shortly on the heels of World War I. The United States found itself in an energy shortage in 1919 that would extend into the 1920s. In response to this shortage, Congress passed the Mineral Leasing Act (Table II). President Woodrow Wilson signed this landmark legislation into law in 1920. Although the legislation has since been amended many times, many of the provisions of the original law are still in effect today. The Mineral Leasing Act of 1920 authorized the Department of the Interior to lease certain federal onshore lands for the production of oil and gas and other hard minerals. In addition, the law established many of the terms, conditions, and procedures used to issue federal oil and gas leases. The Mineral Leasing Act (for Acquired Lands) of 1947 allowed for the leasing of additional federal lands including national forests and national grasslands.

Offshore oil and gas development began in 1896 with the drilling of the first offshore well, which was drilled in the tidelands adjacent to Summerland, California (near Santa Barbara). The possibility of mining offshore lands sparked a decades-long debate as to who controlled these lands, with both the federal government and the governments of the adjacent states claiming ownership. The issue was finally resolved with the passage of the Submerged Lands Act of 1953. This statute assigns the rights to offshore lands from the coastline and seaward for 3 nautical miles (1 nautical mile is approximately 1.15 miles) to the adjacent state except for Texas and the west coast of Florida. These two states hold the rights to the submerged lands from the coastline and seaward for 9 nautical miles. The federal government

TABLE II

Important Legislation Affecting Federal Oil and Gas Leasing

Mineral Leasing Act	Originally passed in 1920 and later amended, this law authorizes the leasing of federal lands for the development of crude oil, natural gas, coal, sulfur, phosphate, potassium, and sodium
Mineral Leasing Act (for Acquired Lands)	Originally passed in 1947 and later amended, this law authorizes the leasing of federal-acquired lands; acquired lands include national forests, national grasslands, etc.
Submerged Lands Act	Passed in 1953, this law grants jurisdiction to the submerged lands from the shoreline seaward for 3 nautical miles (3.45 miles) to the coastal states, except for the coasts of Texas and Florida, where jurisdiction extends 9 nautical miles (10.35 miles)
Outer Continental Shelf Lands Act	Originally passed in 1953 and later amended, this law authorizes the Department of the Interior to lease submerged lands under federal jurisdiction for mineral development
Clean Air Act	Passed in 1970 and later amended, this law regulates the emission of air pollutants from industrial activities
Marine Mammal Protection Act	Passed in 1972, this law provides legal protection for marine mammals and their habitats
Endangered Species Act	Passed in 1973, this law provides legal protection for endangered species and their habitats
Clean Water Act	Passed in 1977, this law regulates the discharge of pollutants into the surface waters of the United States
Federal Oil and Gas Royalty Management Act	Passed in 1982, this law ensures that royalties from federal leases are properly collected and accounted for
National Fishing Enhancement Act	Passed in 1984, this law allows offshore platforms to be used as artificial reefs
Oil Pollution Act	Passed in 1990, this law prohibits the discharge of oil into federal waters and ensures that owners and operators have sufficient resources for oil spill cleanup should such discharges occur
Energy Policy Act	Passed in 1992, this law sets the primary lease term for onshore oil and gas leases at 10 years
Deepwater Royalty Relief Act	Passed in 1995, this law allows the Department of the Interior to grant royalty relief and mandates royalty relief under certain conditions for Gulf of Mexico leases in water depth of 658 ft (200 m) or greater.
Federal Oil and Gas Royalty Simplification and Fairness Act	Passed in 1996, this law improves the management of royalties from federal leases.

Source. Reprinted from the U.S. Department of Interior, Minerals Management Service, U.S. Offshore Milestones and Bureau of Land Management Web site "Minerals and Mining Laws and Regulations," available at www.ut.blm.gov/infocenter/info&gleasing.html.

holds the rights to the offshore lands extending beyond for a distance to 200 miles offshore. These submerged lands are known as the "outer continental shelf." As a practical matter, the suitability of these lands for oil and gas production is limited by drilling technology. The record for deepwater drilling is a well drilled in 9687 ft of water in the Gulf of Mexico in 2001. Continued advances in drilling technology will allow for exploration at greater and greater water depths.

The Outer Continental Shelf Lands Act of 1953 authorized the Department of the Interior to lease outer continental shelf mineral rights to the private sector. This law, as amended, specifies the terms, conditions, and procedures for federal offshore leasing.

Federal leasing progressed rapidly following the enactment of the Outer Continental Shelf Lands Act, first in the Gulf of Mexico and then offshore from California. In 1968, the federal government issued leases in the Santa Barbara Channel. On January 28, 1969, the fifth well drilled from a platform in the Santa Barbara Channel blew out and spilled crude oil for the next 11 days. Media coverage of the resulting oil spill outraged the nation. It is widely believed that this outrage was one of several factors that led Congress to enact a series of environmental laws, many of which affect onshore and offshore oil and gas operations. These laws include the Clean Air Act of 1970, the Marine Mammal Protection Act of 1972, the Endangered Species Act of 1973, the Clean Water Act of 1977, and later, the Oil Pollution Act of 1990.

There has long been concern that the federal government has not received all of the revenue that it is due from its leasing activities (see Section 4.2.3). The Federal Oil and Gas Royalty Management Act of 1982 and the Federal Oil and Gas Royalty Simplification and Fairness Act were enacted to ensure that the government accurately accounted for and collected all payments from lessees for federal oil and gas leases.

3. FEDERAL OIL AND GAS LEASING PROCEDURES

The terms of federal oil and gas leases depend on factors such as whether the lease is competitive or noncompetitive and whether the lease is for onshore or offshore lands. Nevertheless, each federal lease contains certain common elements including the term of the lease, the method for granting the lease, and the types of payments made by lessees. These elements are described below. This description is followed by discussion of how these elements are put together to form the various types of federal oil and gas leases.

3.1 The Primary Lease Term

When the federal government issues an oil and gas lease, it conveys to the lessee the exclusive right to explore and develop the leased land for a predetermined period of time, usually 5 to 10 years. This time period is called the "primary lease term." If a lease begins production during the primary lease term, the lease term is extended for as long as production continues. It should be noted that during the primary lease term, the lessee is typically not obligated to begin exploration and may at any time surrender the lease to the government.

3.2 Methods for Granting Federal Oil and Gas Leases

When more than one party expresses an interest in a lease tract, how does the federal government determine which party will be awarded the tract? In these circumstances, the rights to the lease are auctioned to the highest qualified bidder. Such is the case with all onshore leases where there is a high likelihood that oil and gas will be discovered in commercial quantities and with all offshore leases. For the vast majority of leases, an upfront, cash payment called the bonus payment is the bid variable. Auctions for onshore leases are conducted using oral bidding. Offshore leases are auctioned using sealed bidding.

3.3 Types of Payment for Federal Oil and Gas Leases

There is a series of payments made by lessees to the federal government for the exclusive right to explore and develop a lease tract. These payments include bonus, rental, and royalty payments.

3.3.1 Bonus Payments

A bonus payment is a one-time, nonrefundable, cash payment made by a lessee to the government at the time a lease is issued. It is typically the bid variable at competitive auctions; that is, bidders make their offers in the form of the bonus payment that they will make if they are the successful bidder. Bonus payments are nonrefundable. In the event that a tract is dry or does not contain oil or gas in commercial quantities and the lease is surrendered

to the government, the bonus payment is not refunded. The Department of the Interior accepted bonus payments of $52.2 million for onshore leases during fiscal year (FY) 2000. There were 2311 leases issued, with an average bonus payment of $22,626. Offshore leasing activity was limited to the Gulf of Mexico in FY 2000. There were two lease sales that included 553 lease tracts and 2.9 million acres of submerged lands. The Department of the Interior accepted $441.8 million in bonus payments for these tracts. The average bonus payment was $798,912 per tract. Total onshore and offshore bonus payments exceeded $494 million in 2000.

3.3.2 Rental Payments

Lessees make annual rental payments to the federal government to retain their leases. The amount of the annual rental payment is determined by federal law and is calculated on a per acre basis ranging from as little as $0.25 per acre to as much as $5 per acre. For offshore leases, rental payments cease when production begins. For onshore leases, rental payments continue until the leases are surrendered. Rental payments for onshore oil and gas leases added $44.5 million to government coffers in FY 2000. Rental payments for federal offshore leases totaled $207.8 million. Total onshore and offshore rental payments exceeded $250 million in 2000.

3.3.3 Royalty Payments

In addition to bonus and rental payments, lessees also typically make payments that vary directly with production. The most widely used payment of this type it the royalty.

The term "royalty" harkens back to an era of powerful monarchies. It is, quite literally, "a payment to the crown" and it was a payment to the monarch for using his or her land. In modern usage, royalty payments are payments by lessees to the government calculated as a percentage of gross output. These payments are generally made in monetary terms, but a royalty may be paid in kind. Royalty rates are set by federal law and are typically one-eighth (12.5%) to one-sixth (18.67%) of production. Royalty payments from federal oil and gas onshore oil and gas leases totaled $967.8 million in calendar year 2000. Offshore oil and gas lease royalties totaled $4.1 billion over the same period. The combined total onshore and offshore oil and gas royalties exceeded $5 billion in calendar year 2000. These payments exceeded the receipts from bonus and rental payments by a ratio of 6 to 1.

A minimum royalty payment is sometimes included in a lease contract. As noted earlier, once production begins, lessees make royalty payments to the federal government. A minimum royalty is closely related to a rental payment in that it is determined on a per acre basis. When a minimum royalty payment is included in the lease contract, the lessee must pay the actual royalty payment based on gross production or the minimum royalty payment calculated on a per acre basis, whichever is greater.

A sliding scale royalty is included in some lease contracts. With a sliding scale royalty, the royalty rate varies directly with the level of production; that is, as the production from a lease increases, the royalty rate increases and vice versa. In the case of leases with low production, this royalty relief allows the leases to continue to produce longer than they would have been able to had a higher fixed royalty remained in effect. Similarly, the Deepwater Royalty Relief Act of 1995 provides royalty relief for deepwater (and, therefore, high exploration and development cost) leases.

3.3.4 Typical Federal Lease Procedures and Terms

The federal government uses different combinations of the terms and procedures described above when leasing federal lands for oil and gas production. The leasing terms and procedures for onshore public domain lands and offshore lands are summarized in Table III and are described below.

3.3.4.1 Onshore Public Domain Leases The leasing of onshore public domain lands begins when private firms and individuals express an interest in the leasing the lands. The private sector "nominates" lands in which it has an interest and then the Department of the Interior determines whether to issue leases for the nominated lands. In making this decision, the Department of the Interior evaluates nominated lands in terms of its multiple-use objectives for these federal lands. Should the government decide to lease the lands, it advertises the lease sale. An exact description of the lands to be leased and the terms of the lease are included in the advertisement. For federal onshore leases, an oral auction is conducted to determine who shall win the rights to the mineral rights to the lands—the customary bid variable is a cash bonus bid. Only citizens of the United States or corporations organized under the laws of the United States may bid for federal oil and gas leases. Foreign citizens may hold lease interests only as stockholders in U.S. corporations.

TABLE III

Summary of Federal Oil and Gas Lease Procedures and Terms

	Oil and gas leases on offshore lands
Method of award	Competitive sealed bidding
Customary bid variable	Bonus payment
Customary royalty rate	Flat rates of 12.5% or 16.67% in amount or value of production
Rent and minimum royalty	Rent $3 to $5 per acre
	Minimum royalty $3 to $5 per acre postdiscovery
	Rent $10 per acre for drainage leases
	Minium royalty of $10 per acre for drainage leases
Lease term	Primary lease term 5 to 10 years (8 to 10 years in water depths of 400 m or more); continued if producing in commercial quantities
Size of lease	2500 to 5760 acres unless a larger area is needed for a production unit
	Oil and gas leases on onshore lands: Competitive leases
Method of award	Competitive oral bidding
Customary bid variable	Bonus payment
Customary royalty rate	Leases issued from 5/3/45 to 12/22/87: oil 12.5 to 25% in amount or value of production, depending on production per well per day for the calendar month
	Leases issued after 12/22/87: flat rate of 12.5% in amount or value of production
Rent	Leases issued before 9/2/60: $0.25 to $1 per acre
	Leases issued from 9/2/60 to 12/22/87: $2 per acre
	Leases issued after 12/22/87: $1.50 per acre for the first 5 years, $2 per acre for subsequent years
Lease term	Primary lease term 10 years; continued if producing in commercial quantities
Size of lease	Leases issued before 12/22/87: 640 acres or less
	Leases issued after 12/22/87: maximum 2560 acres for lower 49 states and maximum 5760 acres in Alaska
	Oil and gas leases on onshore lands: Noncompetitive leases
Method of award	Noncompetitive, by application
Customary bid variable	Not applicable
Customary royalty rate	Flat rate of 12.5% in amount or value of production
Rent	Leases issued before 9/2/60: $0.25 to $1 per acre
	Leases issued from 9/2/60 to 2/1/77: $0.50 per acre
	Leases issued from 2/1/77 to 12/22/87: $1 to $2 per acre for the first 5 years, $2 per acre for subsequent years
	Leases issued after 12/22/87: $1.50 per acre for the first 5 years, $2 per acre for subsequent years
Lease term	Primary lease term 10 years; continued if producing in commercial quantities
Size of lease	640 acres minimum, 10,240 acres maximum within 6-mile square

Successful lessees must make annual rental payments of $1.50 per acre for the first 5 years of the lease and $2.00 per acre thereafter. A fixed royalty rate of one-eighth of all production (12.5%) is also required. The primary lease term for onshore leases is 10 years as required by the Energy Policy Act of 1992 and may be extended for as long as production continues. It is interesting to note that bidders are not allowed to conduct exploratory drilling on the lands prior to the auction. Bidders must instead rely on interpretations of seismic surveys and other geophysical tools to evaluate the oil and gas potential of a lease tract. A lease issued in this fashion is a competitive lease.

In some instances, the prospects for oil or gas discovery may be low and the lease sale will attract no bidders. In this case, the lease is issued on a first-come, first-served basis and the lease is deemed "noncompetitive." The terms for a noncompetitive onshore lease are the same as those of a competitive onshore lease except that there is no cash bonus payment. The Federal Onshore Oil and Gas Leasing

Reform Act of 1987 requires that before lands can be leased on a noncompetitive basis, they must first be offered on a competitive basis.

3.3.4.2 Federal Offshore Leases The terms and procedures used to issue offshore leases are similar to those of competitive onshore leases, but with some important differences. As with onshore leasing, the offshore leasing process begins with nominations from the private sector. The customary bid variable is the cash bonus; however, unlike onshore leases, offshore leases are issued using a sealed bid auction. The rental rate for offshore leases is \$3 to \$5 per acre unless the lease is designated as a drainage lease, in which case the annual rental rate is \$10 per acre. A drainage lease is a lease issued for lands containing a known oil field. Successful lessees in this case will "drain" the known oil field. With offshore leases, rental payments cease once production begins. There is a minimum royalty payment that is calculated using the same formula as the rental payment. A fixed royalty rate of one-eighth (12.5%) or one-sixth (16.67%) is required from production. The primary lease term is 5 to 10 years in shallow waters and 8 to 10 years in waters exceeding 400 m in depth. Except for the largest oil companies, firms may bid jointly on offshore leases and leases are transferable to other qualified firms. The Outer Continental Shelf Lands Act requires the issuance of all offshore leases on a competitive basis.

4. THE EFFECTIVENESS OF FEDERAL OIL AND GAS LEASING PROCEDURES

Economic theory suggests that the effectiveness of federal oil and gas leasing procedures should be judged using the following criterion: Federal oil and gas leasing procedures should maximize the value of the public lands and collect this value for the American people.

This criterion has several important implications for economic behavior. First, to maximize the value of the land, only those oil and gas exploration and production activities will be undertaken that are such that the additional economic benefits of the oil and gas produced (or marginal benefits) exceed the additional economic costs needed in their production (or marginal costs). Next, these activities will be optimally timed. There are many alternative time paths for oil and gas exploration and production for any given lease. Each alternative time path should be evaluated in terms of its net present value using an appropriate social discount rate and the alternative that yields the highest net present value should be undertaken. By foregoing those activities whose marginal costs are not justified by their marginal benefits, and by the optimal timing of these activities, the value of the land is maximized. Assuming that all of the benefits and costs of oil and gas exploration, development, and production are borne by competitive lessees, and assuming that social and private discount rates closely approximate one another, the industry behavior will maximize the value of the lands to society. If a portion of the benefits or costs is external to the firm, or if private and social discount rates diverge, then industry behavior will maximize profit, but may maximize the value of the lands to society. The case of external costs is discussed in Section 5. Finally, the leasing procedures should collect any surplus value generated by the leases for the American people. This surplus value, also known as economic rent, is the difference between the total revenue generated by the lease and the total necessary costs needed to generate this revenue. Included in the total cost is a normal rate of return on investment for the lessee. In economic theory, a normal rate of return on investment is properly thought of as a cost because such a return could have been earned in an alternative investment (it is an opportunity cost). Any administrative and compliance costs associated with the leasing procedures reduce the value of the leases and, therefore, reduce the economic rent. Thus, an ideal leasing procedure holds administrative and compliance costs to a minimum. In summary, leasing procedures should maximize the economic rent from public lands and collect this rent for the American people.

4.1 The Effectiveness of the Primary Lease Term

Federal oil and gas leases are issued with a primary lease term to encourage lessees to conduct exploration activities expeditiously. To the extent that the primary lease term alters the timing of exploration, it reduces the economic rent that can be derived from a lease. To maximize economic rent, the primary lease term should be lengthened so that it does not alter lessee behavior or should be eliminated altogether.

4.2 The Effectiveness of the Different Types of Payment for Federal Oil and Gas Leases

There are four types of payments made by lessees to the government for the mineral rights on public lands: bonus, rental, royalty, and profit-share payments. Although the lessees view these payments as costs of production, in a social sense they are not economic costs like expenditures for labor and equipment to drill an exploratory well. They are transfer payments that redistribute income from the private to the public sector. As such, it would be ideal if these payments did not alter lease operating decisions.

4.2.1 The Effectiveness of Bonus Payments

Such is the case with bonus payments. A bonus payment is a one-time payment made at the time a lease is issued. Hence, it is does not affect future marginal exploration and development decisions. The exploration, development, and production decisions that maximize the value of the lease to the lessee also maximize the value of the lease to society.

4.2.2 The Effectiveness of Rental Payments

Rental payments affect marginal decisions and alter the productivity and life of a lease. It is interesting to note that the effects of rental payments differ for onshore versus offshore leases. For onshore leases, rental payments continue for the life of the lease. They are viewed as a marginal cost of operating a lease for one more period and can be avoided by surrendering a lease. As such, they may lead to the abandonment of a lease that would be profitable in the absence of the rental payment. In certain circumstances, a lease may be abandoned as "dry" when in fact it would be developed in the absence of a rental payment. The value of the output lost to premature abandonment or nondevelopment of a lease is a loss to society and wastes a valuable natural resource.

For offshore leases, rental payments cease when production begins. As such, a rental payment can be avoided by accelerating exploration. In both onshore and offshore leases, rental payments alter the behavior of lessees and reduce the economic rent that can be derived from the federal lands.

Administrative and compliance costs associated with rental payments are not as small as those associated with bonus payments, but still are not high. This is because the amount of the annual rental payment is not recalculated each year and does not vary with production.

4.2.3 The Effectiveness of Royalty Payments

Like rental payment, royalty payments are perceived as additions to the marginal cost of production. As such, they also can lead to the premature abandonment or nondevelopment of a lease that should otherwise be developed, reducing the economic rent from the lease. Reducing the royalty rate as production decreases can mitigate the adverse effect on total production. The Outer Continental Shelf Lands Act (as amended) and the Deepwater Royalty Relief Act include royalty rate reductions for offshore leases.

Royalty payments also alter the timing of exploration and production activities by lessees. With other factors held constant, businesses in general and lessees in particular will realize revenues as early as possible and delay costs for as long as possible. Because lessees perceive royalty payments as marginal costs, they delay production.

Unlike bonus and rental payments, royalty payments are contingent on production. The administrative and compliance costs associated with bonus payments are negligible and those associated with rental payments are small. Such is not the case with royalty payments because they depend on actual production. For the government to ensure that it is receiving the full amount of the royalty payments it is due, it must monitor actual production. Concern over the accuracy of royalty payments led to the passage of the Federal Oil and Gas Royalty Management Act of 1982 and the Federal Oil and Gas Royalty Simplification and Fairness Act of 1996.

4.3 The Effectiveness of Different Methods for Granting Federal Oil and Gas Leases

The bonus payment is the bid variable for federal offshore leases and the auction is conducted using sealed bidding. This combination is highly effective in maximizing and collecting economic rent for leased lands. As noted earlier, bonus payments do not affect decisions at the margin and, therefore, do not alter the time path for exploration and production, thereby maximizing the economic rent. With a sealed bid auction, bidders submit their bids in a sealed envelope. On the day of the auction, the bids are opened and the lease is awarded to the highest qualified bidder. The bidders need not be present to win and may not bid again if their initial bids are not high enough to win the auction. At the time of the auction, bidders are unaware of the number of competing bids. Because of this, they prepare bids that approximate the economic rent. Indeed, concerns have been raised that this

method of leasing is too successful and that winning bids will systematically come from a bidders who are too optimistic about the prospects for the lease. These "lucky" bidders have bid too much and will find that they are operating their leases at a loss. This phenomenon is called the "winner's curse." Bidders are aware of this phenomenon and can be expected to adjust their bids accordingly.

Oral bid auctions are used to issue onshore leases. Oral bidding is also effective in collecting economic rent if there are large numbers of bidders and if the bidders do not collude with one another. With an oral bid auction, the winning bid is just above the valuation of the second high bidder. That is, the winning bidder must outbid the nearest competitor, but may not need to bid the full valuation of the lease.

Experiments have been conducted with alternate bid variables including the royalty rate and profit-share rate. These bid variables neither maximize economic rent nor collect it and should be avoided.

5. STATE AND LOCAL GOVERNMENT AND ENVIRONMENTAL CONCERNS REGARDING FEDERAL OIL AND GAS LEASING

The production of oil and gas from federal lands yields large net benefits to the nation through the revenue it generates for government coffers and through the oil and gas it provides for the American economy. Despite these benefits, federal oil and gas leasing does not enjoy universal support. The most vocal critics of the federal leasing program are state and local governments and environmental groups. Their opposition stems from the fact that not all of the costs of federal oil and gas leasing operations are internalized by the lessees.

Lessees bear the lease acquisition costs and the costs of the labor and capital needed for lease exploration and development, but may not bear the cost of local public services including providing police and fire protection, building and maintaining the transportation infrastructure, and providing public education. Whereas the benefits of a federal lease projects are spread across the entire nation, these external costs are borne disproportionately by those citizens living near the projects.

The opposition from environmental groups to federal oil and gas leasing is based on two interrelated concerns. First, lessees do not bear the external cost resulting from the air, water, and noise pollution generated by lease operations. To the extent that the lessees do not internalize these costs, environmental concerns are well founded. Second, public lands often have multiple conflicting uses and it may be difficult to determine which use yields the highest value to the nation. For example, certain public lands may be used to produce oil, gas, or

TABLE IV

Selected Formulas for Disbursements of Revenues from Federal Oil and Gas Leases

	Oil and gas leases on offshore lands
States	27% of royalty, rent, and bonus revenues to the affected state
Special Purpose Accounts	$150 million annually to the Historic Preservation Fund
U.S. Treasury	The remainder of the revenue to the General Fund of the U.S. Treasury
	Public domain lands
States	50% to the state in which the lease is located except Alaska
	90% to the state of Alaska if the lease is located in Alaska
Special Purpose Funds	40% to the Reclamation Fund except if the lease is located in Alaska in which case no revenues are directed to the Reclamation Fund
U.S. Treasury	10% to the General Fund of the U.S. Treasury
	National forests
States	25% to the state(s) in which the forest is located
Special Purpose Funds	None
U.S. Treasury	75% to the General Fund of the U.S. Treasury

Source. Reprinted from U.S. Department of the Interior, Minerals Management Service (2001). "Mineral Revenues 2000," Table 30, pp. 97–98.

wood products, for recreation, or as wildlife preserves. The value of the lands in oil and gas production or timber production is relatively easy to determine because of the markets for these end products. The value of the lands in recreational uses or as wildlife preserves is more difficult to determine. In fact, there is no general agreement that the value of these and other ecosystem services can be measured in monetary terms. Techniques have been developed to estimate the value of public lands in these circumstances, but these techniques are in their infancy and the estimates they generate are subject to debate. Indeed, disagreements between the oil and gas industry and environmentalists on the appropriate use of public lands can be traced to disagreements as to the value of the lands in alternate uses.

Local governments and environmental groups remain active in their opposition to federal oil and gas leasing. They have enjoyed some successes in limiting federal oil and gas leasing, as evidenced by the fact that very few offshore leases have been issued in the past decade and the repurchase by the federal government of leases offshore from Florida in 2002.

A possible solution to the concerns of state and local governments regarding the provision of public services may be found in the way that the federal government disburses oil and gas lease revenues. A summary of revenue disbursement formulas is found in Table IV. These formulas are set by federal law. State and local opposition to federal oil and gas leasing will lessen if a greater percentage of lease revenues are directed to the affected geographic area.

Environmental concerns may be taken into account by using environmental laws to internalize the external costs of federal oil and gas leasing. Federal leasing should take place only when the benefits of the leasing exceed all economic costs (including environmental costs). Unfortunately, some environmentalists and environmental groups would like to halt federal oil and gas leasing regardless of the benefit to the nation. Such a prohibition on leasing reduces the welfare of the nation.

SEE ALSO THE FOLLOWING ARTICLES

Markets for Natural Gas • Markets for Petroleum • Natural Gas, History of • Oil and Natural Gas Drilling • Oil and Natural Gas: Economics of Exploration • Oil and Natural Gas Exploration • Oil Industry, History of • Oil Price Volatility • Oil Recovery • Oil Refining and Products • Petroleum Property Valuation • Strategic Petroleum Reserves

Further Reading

McDonald, S. L. (1979). "Leasing of Federal Lands for Fossil Fuel Production." Published for Resources for the Future by the Johns Hopkins University Press, Baltimore, MD.

Mead, W. J., *et al.* (1985). "Offshore Lands: Oil and Gas Leasing and Conservation on the Outer Continental Shelf." Pacific Institute for Public Policy Research, San Francisco, CA.

Minerals Management Service, U.S. Department of the Interior (2001). "Mineral Revenues 2000." Herndon, VA.

Minerals Management Service, U.S. Department of the Interior (2001). "Outer Continental Shelf Oil and Gas Leasing Procedures," Outer Continental Shelf Report MMS 2001-076. New Orleans, LA.

Oil and Natural Gas Liquids: Global Magnitude and Distribution

THOMAS S. AHLBRANDT
U.S. Geological Survey
Denver, Colorado, United States

1. Introduction
2. Worldwide Distribution
3. Impact of Distribution
4. Data Sources

Glossary

continuous accumulations Petroleum that occurs in extensive reservoirs and is not necessarily related to conventional structural or stratigraphic traps. These accumulations of oil and/or gas lack well-defined down-dip petroleum/water contacts and thus are not localized by the buoyancy of oil or natural gas in water.

conventional accumulations Petroleum that occurs in structural or stratigraphic traps, commonly bounded by a down-dip water contact, and therefore affected by the buoyancy of petroleum in water.

crude oil A mixture of hydrocarbons that exists in a liquid phase in natural underground reservoirs and remains liquid at atmospheric pressure after passing through surface separation facilities. Crude oil may also contain some nonhydrocarbon components; referred to as oil in this article.

cumulative production Volumes of oil and natural gas liquids that have been produced.

endowment The sum of cumulative production, remaining reserves, mean undiscovered recoverable volumes, and mean additions to reserves by field growth.

field A contiguous area consisting of a single reservoir or multiple reservoirs of petroleum, all grouped on, or related to, a single geologic structural and/or stratigraphic feature.

future petroleum The sum of the remaining reserves, mean reserve growth, and the mean of the undiscovered volume. Cumulative production does not contribute to the future petroleum. The terms future oil, future liquid volume, or future endowment are sometimes used as variations of future petroleum to reflect those resources that are yet to be produced.

liquid petroleum A term used to represent crude oil combined with natural gas liquids.

natural gas liquids (NGL) Those portions of reservoir gas that are liquefied at the surface in lease separators, field facilities, or gas processing plants. Natural gas liquids are heavier homologs of methane (including ethane, propane, butane, pentane, natural gasoline, and condensate).

oil field A field producing oil and gas is termed an oil field when the petroleum contained within has a gas/oil ratio (GOR) of less than 20,000 cubic feet per barrel. If the GOR $>20{,}000$ cubic feet per barrel, the field is called a gas field.

remaining reserves Recoverable oil and NGL volumes that were originally present and have not yet been produced.

reserve The estimated quantities of petroleum expected to be commercially recovered from known accumulations relative to a specified date, under prevailing economic conditions, operating practices, and government regulations. Reserves are part of the identified (discovered) resources and include only recoverable materials.

reserve (field) growth The increases of estimated petroleum volume that commonly occur as oil and gas fields are developed and produced.

resource A concentration of naturally occurring solid, liquid, or gaseous hydrocarbons in or on the earth's crust, some of which is currently or potentially economically extractable.

undiscovered resources Resources inferred from geologic information and theory to exist outside of known oil and gas fields.

Although their abundance and potential shortages are debated, under most energy projections, fossil fuels will remain the dominant source of energy for the remainder of this century. Their distribution, quantities, and availability will thus continue to be of

paramount importance for world economies. Policy decisions made in the United States and elsewhere depend on detailed knowledge of world petroleum resources. Therefore, petroleum resources are periodically reassessed, not just because new data become available and better geologic models are developed but also because many nongeologic factors determine which part of the crustal abundance of petroleum will be economic and acceptable over some foreseeable future. Misunderstanding of the reserve and resource terminology and the nature of resources allow those with fixed points of view to argue vehemently over the size of remaining resources.

1. INTRODUCTION

Detailed petroleum information was published in a world assessment of petroleum resources conducted by the U.S. Geological Survey (USGS) in its World Petroleum Assessment 2000 (WPA). This assessment provides estimates of the quantities of conventional technically recoverable oil, gas, and natural gas liquids (NGL) outside the United States that have the potential to be added to reserves. Many individuals and groups, such as the International Energy Agency and the Energy Information Administration of the Department of Energy, use this report as the baseline for world petroleum resources. This assessment is also the basis for estimates of reserve growth and undiscovered resources for this article. Reserves and cumulative production data are derived from the IHS Energy database.

1.1 Resource Classification

This article classifies liquid petroleum resources into four categories: cumulative production—those volumes of oil and NGL that have already been produced; reserves—volumes of liquids remaining in already discovered conventional oil fields that are known with reasonable certainty; reserve growth (also known as field growth)—those volumes in already discovered conventional oil fields that are not known with certainty but can potentially be added to reserves in the future; and undiscovered resources—volumes of oil and NGL that have not yet been discovered. These volumes are estimates and subject to change through time in light of new geologic and engineering data as well as changing economic conditions. Additionally, as oil or NGL are discovered or produced, volumes from one category may shift to another category (e.g., from reserves to cumulative production). Many terms are used to identify categories and subcategories of reserves and resources; such terms are not always used consistently and must be considered cautiously. In this article, only the four categories defined are used and their sum is the endowment.

Oil resources in continuous accumulations, such as heavy oils, bitumens, tar sands, and oil shales, are generally classified and evaluated separately. Such continuous resources were identified, but not formally assessed, in USGS WPA 2000; they are not included in this article. This article considers only conventional liquid petroleum.

1.2 Uncertainty of Estimates

Estimates of oil and NGL are subject to considerable uncertainty. This is true for all categories, including estimates of past production. The uncertainty of these volumes leads to probabilistic approaches to estimating reserves, reserve growth, and undiscovered resources. The World Petroleum Congress/Society of Petroleum Engineers/American Association of Petroleum Geologists issued a reserve and resource classification system that recognized this uncertainty in all categories, including reserve estimates. Their classification uses the terms proved, probable, and possible: A 90% certainty is assigned to proved reserves, a 50% certainty to probable reserves, and a 10% certainty to possible reserves. In this article, the means of the probabilistic estimates are used.

Recent studies demonstrate that reserve growth is a worldwide phenomenon and must be considered in any analysis of world petroleum resources. As with other estimates, for reserve growth, the mean estimates are used here.

2. WORLDWIDE DISTRIBUTION

Remaining reserves and cumulative production data as of 1/1996 were obtained from the IHS Energy database. Two major components of potential petroleum resources for the world were assessed in USGS WPA 2000: undiscovered conventional resources and reserve (or field) growth. The assessed areas include 128 of the 937 geologic provinces defined by the USGS WPA 2000, and the U.S. provinces assessed by the USGS in 1995 and the MMS in 1996. This study includes more than 95% of the conventional petroleum resources of the world (Fig. 1).

Potential additions to reserves from reserve growth were estimated at the world level only, but for this

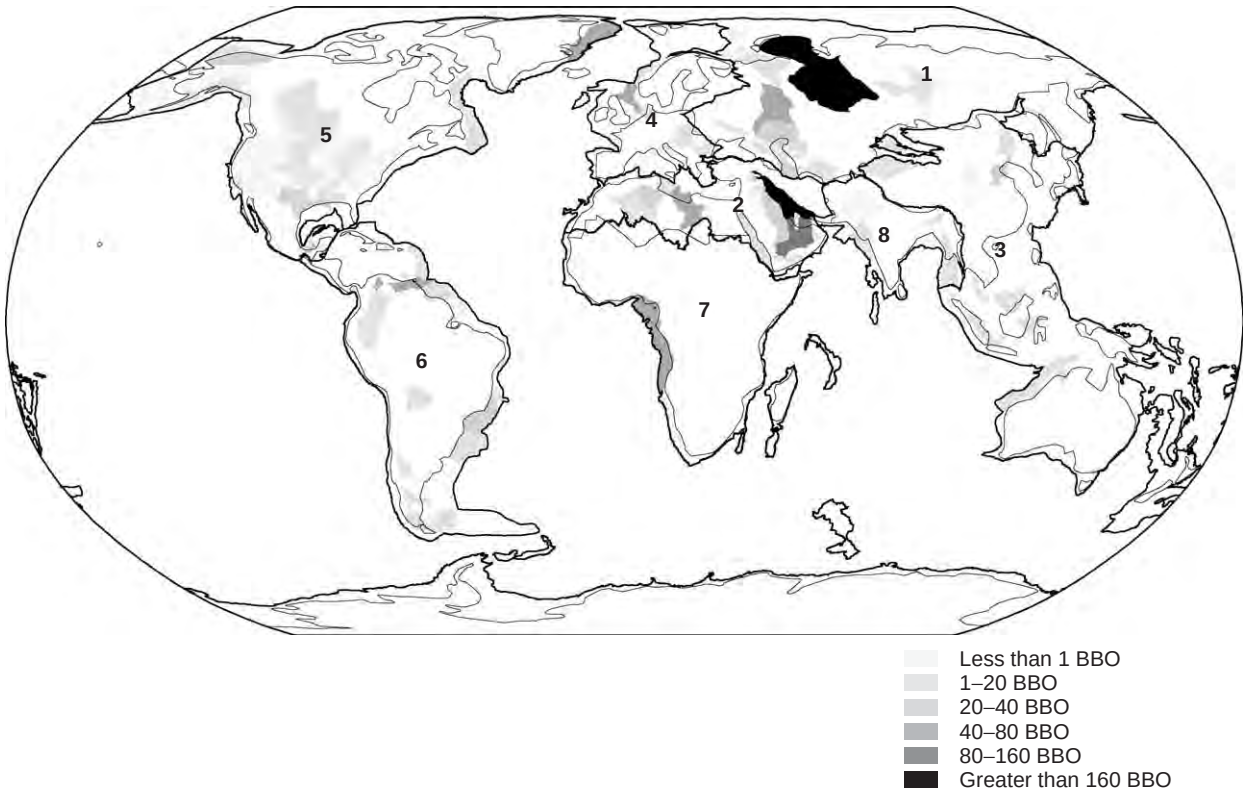


FIGURE 1 Conventional oil resources of the world [billions of barrels (BBO)] as assessed by the USGS World Petroleum Assessment (2000). Darker shading indicates greater resource volume. Data include cumulative production and remaining reserves from IHS Energy as of January 1996 and the mean estimate of undiscovered conventional oil resources of the world as assessed by the USGS World Petroleum Assessment Team in 128 provinces of the world outside the United States, including the U.S. province assessments as assessed by the USGS in 1995 and the Minerals Management Service in 1996. USGS regions: 1, Former Soviet Union; 2, Middle East and North Africa; 3, Asia Pacific; 4, Europe; 5, North America; 6, Central and South America; 7, sub-Saharan Africa and Antarctica; 8, South Asia.

article they have been allocated to the regional level (Fig. 2; Tables I and II). Reserve growth for the world was first estimated by the USGS in WPA 2000, but it has also been estimated by British Petroleum, the International Energy Agency, and individual researchers. A hierarchical assessment system was used in the USGS WPA 2000, with the foundation being the assessment unit, a subdivision of a total petroleum system, which was the formally assessed entity. These basic estimates were then allocated to the appropriate total petroleum systems, provinces, region, and finally aggregated to the world level.

Approximately 3 trillion barrels of oil (TBO) is estimated to be the world's petroleum endowment of recoverable oil and NGL; this endowment includes cumulative production, remaining reserves, reserve growth, and undiscovered resources (mean estimate). This volume includes the U.S. estimates made by the USGS in 1995 for both the onshore and state waters and by the Minerals Management Service (MMS) in 1996 for the federal offshore U.S. World oil reserves are currently 1.1 TBO but were 0.96 TBO as of January 1996; world consumption is approximately 0.028 TBO per year according to the IHS database. Half of the world's undiscovered liquid petroleum potential is offshore. Arctic basins are the next frontier, with approximately 25% of undiscovered petroleum resources (Fig. 1).

The estimate of oil endowment of 3 TBO is comparable to 1998 a estimate by British Petroleum of 3.3 TBO at the mean and the 2002 estimate by J. S. Edwards of 3.7 TBO at the median. The USGS mean reserve growth estimate for the world is 667 billion barrels of oil (BBO). For the world's oil fields in 1998, British Petroleum estimated reserve growth to be 500 BBO, and the International Energy Agency estimated it to be between 418 and 1136 BBO.

This section discusses the worldwide distribution of conventional oil and NGL resources. Much of the data for this section are from the USGS

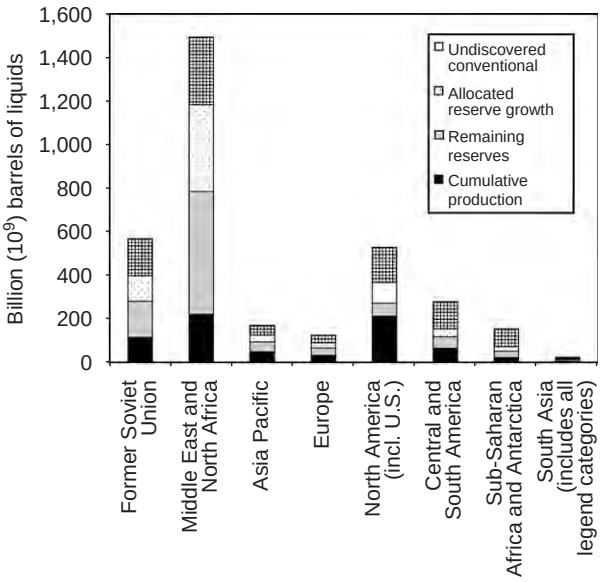


FIGURE 2 Conventional liquid petroleum endowment (cumulative production, remaining reserves, mean estimate of reserve growth, and mean estimate of undiscovered resources) for the eight regions of the world as defined by the USGS World Petroleum Assessment (2000).

WPA assessment, the USGS 1995 U.S. national assessment, the MMS federal offshore assessment, the U.S. Energy Information Administration, the International Energy Agency, and the British Petroleum statistical review.

Table I shows the volumes of conventional oil and NGL for the eight USGS world regions. The geographic distribution of the world's oil endowment for USGS-defined provinces within the eight regions is shown in Fig. 1. Table II combines the regional oil and NGL data into liquid petroleum resources in each region. The regions are discussed in decreasing order of their total resource volume. Because reserve (field) growth was estimated by the USGS in 2000 at the world level only, this regional analysis includes the first discussion of reserve growth by the world's regions, and Tables I and II and Fig. 1 give a complete analysis of the regional oil and/or liquid petroleum endowment.

Information presented in Figs. 1–3 and Tables I–V supports several observations. First, liquid petroleum resources are unevenly distributed throughout the world. Three regions contribute more than 77% of the world's liquid petroleum endowment; the Middle East/North Africa contributes the greatest volume (44.7%), followed by the former Soviet Union (17%) and North America (15.8%). Second, four regions of the world export most of the liquid petroleum (Middle East/North Africa, the former Soviet Union, Sub-Saharan Africa, and Central and South America in descending order) to the remaining five regions. Third, North America, the oldest producing province, still produces nearly 20% of the world's liquid petroleum. Fourth, Europe is the next to last ranked region, with only 3.8% of the world's liquid petroleum endowment, and it is a substantial net importer of these resources. Finally, reserve growth may significantly contribute to the endowment in three regions: the Middle East/North Africa, the former Soviet Union, and North America (Fig. 2).

2.1 Middle East and North Africa

With 44.7% of the world's liquid petroleum endowment and 58.8% of the remaining liquid resources, this region will continue be the largest source of oil and NGL. Recent discoveries in the Arabian Peninsula, the Gulf of Suez, and the Algerian basins confirm the high liquid petroleum potential of the region. For example, discoveries in the Trias/Ghadames Basin in the past 5 years have more than doubled the volume of all previous discoveries in the province. Multi-billion barrel discoveries have been made in other basins in this region, including the Azadegan field in Iran, the Kra Al Maru field in the northern Arabian Gulf, many satellite fields to Ghawar field in Saudi Arabia, and significant (1 billion barrels or larger) discoveries in the Sirte Basin in Libya. Detailed reserve growth studies demonstrate very large potential for discovered but undeveloped reserves, particularly in Iraq and Iran. With 312 BB of undiscovered liquid petroleum and reserve growth potential of 398 BB (Table II), this region will continue to be preeminent for liquid petroleum resources (Tables III and IV).

2.2 Former Soviet Union

This region has 17% of the world's liquid petroleum endowment. Recent oil discoveries in the Caspian region demonstrate the high liquid petroleum potential of this region—for example, in the North Caspian Basin, the Kashagan field, for which estimates range well above 20 BB, and the Azeri-Chirag-Gunashli complex in the Caspian, for which estimates are at least 5 BB. Frontier provinces in the offshore Arctic of the former Soviet Union (FSU), such as the Laptev Sea, may have significant petroleum potential but are unevaluated. More mature provinces, including the West Siberian,

TABLE I

World Oil Resources and NGL Resources by Region (in Million Barrels)[a]

	Oil					NGL				
Region[b]	Cumulative production	Remaining reserves	Allocated reserve growth	Undiscovered resources	Endowment	Cumulative production	Remaining reserves	Allocated reserve growth	Undiscovered resources	Endowment
1. Former Soviet Union	112,106	152,173	104,537	115,985	484,801	1801	14,753	9,071	54,806	80,431
2. Middle East and North Africa	218,644	528,620	363,141	229,882	1,340,287	1413	35,376	21,750	81,747	140,286
3. Asia Pacific	47,722	39,980	27,465	29,780	144,947	1022	4,906	3,016	15,379	24,323
4. Europe	31,235	27,572	18,941	22,292	100,040	649	6,400	3,935	13,667	24,651
5. North America (including U.S.)[c]	208,832	59,143	95,341	153,491	516,807	1980	1,462	899	7,853	12,194
6. Central and South America	64,096	50,272	34,535	105,106	254,009	486	2,016	1,239	20,196	23,937
7. Sub-Saharan Africa and Antarctica	22,437	26,049	17,895	71,512	137,893	10	2,932	1,803	10,766	15,511
8. South Asia	5,131	7,071	4,858	3,580	20,640	24	467	287	2,604	3,382
Total (including U.S.)	710,203	890,880	666,712	731,628	2,999,423	7385	68,312	42,000	207,018	324,715

[a] All estimates are mean estimates. Oil and NGL resource totals from U.S. Geological Survey; cumulative oil production and remaining reserves from IHS Energy as of January 1996. Allocated mean reserve growth from U.S. Geological Survey estimates; allocated (for non-U.S.) proportional to mean volumes of remaining reserves.

[b] USGS region and code.

[c] For U.S. portion of North America, oil and NGL estimates are from USGS (1995) and MMS (1996).

TABLE II

World Petroleum Liquids (Oil and NGL Combined) Resources by Region[a]

Region[b]	Million barrels						%					
	Cumulative production	Remaining reserves	Allocated reserve growth	Undiscovered resources	Endowment	Future resources	World cumulative production	World remaining reserves	World reserve growth	World undicovered resources	World future resources	World's endowment
1. Former Soviet Union	113,907	166,926	117,502	170,791	569,126	455,219	15.87	17.40	16.10	18.20	17.32	17.01
2. Middle East and North Africa	220,057	563,996	398,421	311,629	1,494,103	1,274,046	30.67	58.80	54.58	33.20	48.48	44.66
3. Asia Pacific	48,744	44,886	31,504	45,159	170,293	121,549	6.79	4.68	4.32	4.81	4.63	5.09
4. Europe	31,884	33,972	23,581	35,959	125,396	93,512	4.44	3.54	3.23	3.83	3.56	3.75
5. North America (including U.S.)[c]	210,812	60,605	96,240	161,344	529,001	318,189	29.38	6.32	13.18	17.19	12.11	15.81
6. Central and South America	64,582	52,288	37,061	125,302	279,233	214,651	9.00	5.45	5.08	13.35	8.17	8.35
7. Sub-Saharan Africa and Antartica	22,447	28,981	20,364	82,278	154,070	131,623	3.13	3.02	2.79	8.77	5.01	4.61
8. South Asia	5,155	7,538	5,326	6,184	24,203	19,048	0.72	0.79	0.73	0.66	0.72	0.72
Total (including U.S.)	717,588	959,192	730,000	938,646	3,345,426	2,627,838	100.00	100.00	100.00	100.00	100.00	100.00

[a] All estimates are mean estimates. Oil and NGL resource totals from U.S. Geological Survey; cumulative oil production and remaining reserves from IHS Energy as of January 1996. Allocated reserve growth from U.S. Geological Survey estimates; allocated (for non-U.S.) proportional to volumes of remaining reserves.

[b] USGS region and code.

[c] For U.S. portion of North America, oil and NGL estimates are from USGS (1995) and MMS (1996).

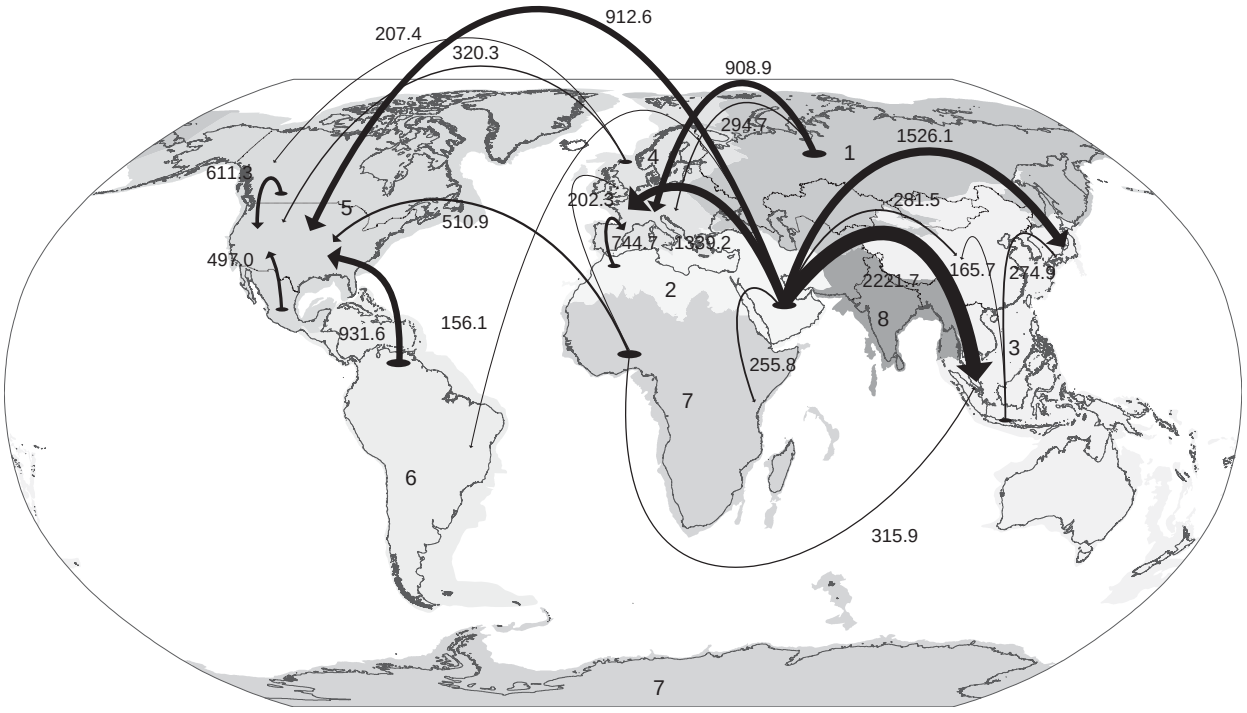


FIGURE 3 Major oil trade movements in million barrels of oil for 2000. Modified from British Petroleum statistical review to unit measurements in millions of barrels assuming 33 API gravity. USGS regions: 1, Former Soviet Union; 2, Middle East and North Africa; 3, Asia Pacific; 4, Europe; 5, North America; 6, Central and South America; 7, sub-Saharan Africa and Antarctica; 8, South Asia.

TABLE III

World Oil Production in 2000[a]

Region[b]	Oil production	Oil consumption	% world production	% world consumption	% surplus or deficit
1. Former Soviet Union	2,978	1,340	10.63	4.83	+5.80
2. Middle East and North Africa	9,877	2,292	35.24	8.26	+26.98
3. Asia Pacific	2,662	6,655	9.50	23.97	−14.47
4. Europe	2,562	5,522	9.14	19.89	−10.75
5. North America (including U.S.)	5,588	8,684	19.94	31.27	−11.33
6. Central and South America	2,547	1,874	9.09	6.75	+2.34
7. Sub-Saharan Africa and Antarctica	1,520	469	5.42	1.69	+3.73
8. South Asia	296	930	1.06	3.35	−2.29
Total (including U.S.)	28,029	27,767	100.00	100.00	

[a] Year 2000 production and consumption from U.S. Energy Information Administration. All volumes in MMBO. NGL data not available.
[b] USGS region and code.

Timan-Pechora, and Volga Ural Basins, still have significant oil potential. The East Siberia Basin contains the discovered but undeveloped Yurubchen-Takhoma oil field (~10 BB) and very large heavy oil deposits, neither of which were formally assessed in 2000 by the USGS. This region will remain a significant source of liquid petroleum resources as demonstrated by the estimated undiscovered resource of 171 BB and reserve growth of 117 BB (Table II).

TABLE IV

Top 20 Producers of Liquid Petroleum (Crude Oil and Natural Gas Liquids), 2000[a]

Country	Million barrels per day
Saudi Arabia	9.12
United States	9.06
Russia	6.71
Iran	3.78
Mexico	3.48
Norway	3.32
China	3.25
Venezuela	3.14
Canada	2.76
Iraq	2.59
United Arab Emirates	2.57
United Kingdom	2.55
Kuwait	2.25
Nigeria	2.15
Brazil	1.54
Indonesia	1.51
Libya	1.47
Algeria	1.44
Oman	0.94
Qatar	0.87

[a]Data from the Energy Information Administration (2001).

TABLE V

Top 20 Consumers of World Petroleum Liquids, 2000[a]

Country	Million barrels per day
United States	19.70
Japan	5.53
China	4.78
Germany	2.77
Russia	2.50
Brazil	2.16
South Korea	2.15
Canada	2.07
France	2.02
Mexico	1.99
India	1.99
Italy	1.87
United Kingdom	1.72
Spain	1.46
Saudi Arabia	1.30
Iran	1.08
Indonesia	1.04
Australia	0.86
The Netherlands	0.85
Taiwan	0.85

[a]Data from the Energy Information Administration (2001).

2.3 North America

The major oil-producing regions in Canada, Mexico, and the United States are relatively mature areas in terms of exploration and development compared to much of the rest of the world. Despite this maturity, North America still produces 20% of the world's liquid petroleum, second only to the Middle East/North Africa, which provides 35.2%. North America has 15.8% of the world's liquid petroleum endowment (which includes all resource categories for oil and NGL), but it has only 12.1% of its future liquid petroleum potential (which excludes cumulative production). These facts indicate that North America has utilized a significant portion of its liquid petroleum endowment. Since North America is the largest consumer of liquid petroleum with 31.3% of world consumption, this region will remain a major importer of liquid petroleum. However, significant new discoveries are still being made, particularly offshore in the Gulf of Mexico and on the North Slope of Alaska. For example, Thunder Horse, Atlantis, and Green Canyon 640 fields are new discoveries in excess of a half billion barrels in the Gulf of Mexico. Multiple discoveries in excess of several hundred million barrels on the North Slope of Alaska may provide significant additions to reserves. In Canada, the eastern offshore continental shelf remains prospective for petroleum, as evidenced by new discoveries, mostly gas, but the multi-billion barrel Hibernia field indicates the presence of significant liquid petroleum resources. Despite these successes, the North American region imports 10.8% more liquid petroleum than it produces and will probably continue to import. Canadian heavy oil and tar sand deposits in Alberta will be critical to hydrocarbon supply in the region. It is estimated that 10% of North America's oil production will come from Alberta's oil sands by 2005. They are estimated to contain 1.6 trillion barrels of oil, of which 311 billion barrels are recoverable. These sands are considered to be a continuous deposit and as such were not assessed by the USGS in 2000, and they are not included in the figures or tables in this article.

Northeast Greenland was included in this region and as such added considerably to the potential of North America, although the potential in Greenland is hypothetical and there is an accessibility risk of developing resources due to its hostile environment.

Numerous geologic structures that occur offshore are thought to be on the western side of the North Sea producing area. These structures may contain oil, as evidenced by oil accumulations exhumed onshore in Greenland. The same geologic settings are prolific liquid petroleum provinces in the North Sea.

With undiscovered resources of 161 BB and potential reserve growth of 96 BB, this region remains an important producer of liquid petroleum (Tables II and IV).

2.4 Central and South America

With 8.4% of the world's liquid petroleum endowment, and consuming only 6.8% of the world's liquid petroleum, this region is a petroleum supplier to the remainder of the world. The Atlantic offshore of South America has become a prolific hydrocarbon producer. The success of the offshore areas of Brazil in the Campos Basin, with oil discoveries greater than 1 BBO at Roncador and Barracuda fields and a 600-million barrel discovery in the Santos Basin, demonstrate the significant liquid petroleum resources offshore. Elsewhere along this Atlantic margin, large undiscovered liquid petroleum resources were estimated for offshore areas of Guyana and Surinam. The discovery of hydrocarbons in deep water of the south Atlantic (approximately 4000-m water depths) in the Campos Basin make both the offshore area of South America and counterpart basins along the western margin of Africa one of the most active and successful exploration areas. This significant potential is reflected in the estimate of 125 BB of undiscovered liquid petroleum resources and another 37 BB of potential reserve growth of liquid petroleum (Table II).

Although not assessed, enormous heavy oil deposits are known in the Orinoco heavy oil belt. These resources are estimated at more than 1 trillion barrels of oil in place, comparable to the Alberta oil sands. The recoverable amount from the Orinoco heavy oil belt varies depending on the gravity of the oil, depth, and location.

2.5 Asia/Pacific

With only 5.1% of the world's liquid petroleum endowment and a huge population to support, the Asia/Pacific region is the second largest consumer of liquid petroleum (24%) but produces only 9.5% of it, resulting in the largest percentage deficit (−14.5%) of liquid petroleum of any region of the world (Tables III and IV). Overall, estimates for undiscovered liquid petroleum resources in this region have been reduced from previous estimates to 45 BB, with 31 BB of potential reserve growth (Table II). Much of this reduction resulted from the individual reassessment of Chinese provinces. Extensive exploration in China has achieved only modest success. China has been a net oil importer since 1993, and the consequent reduced likelihood of adding substantial new reserves will make China increasingly competitive elsewhere for petroleum in the world. Modest increases of undiscovered oil were estimated for the northwest shelf of Australia and deep-marine reservoirs in the Kutei and Sabah offshore basins of Indonesia and Malaysia.

2.6 Sub-Saharan Africa and Antarctica

With only 1.7% of the world's liquid petroleum consumption and 4.6% of the world's endowment, this region is an important supplier of liquid petroleum (Tables III–V). Continued success in the Niger Delta, as well as in deep-water reservoirs especially offshore of Cameroon and Angola, make this region particularly important for liquid petroleum. Many new oil discoveries of more than a half billion barrels of oil have been made offshore of western Africa. The undiscovered liquid petroleum resource estimate of 82 BB and the estimated 20 BB of potential reserve growth demonstrate the significant potential volumes of this region (Table II). Examples of discoveries in excess of 1 BB include Girassol and Dalia fields in Angola. In terms of liquid petroleum export potential, this region ranks third behind the Middle East/North Africa and the FSU (Table III).

Antarctica was not assessed by the USGS in 2000 because it will probably not be accessible for petroleum exploration, due to international agreements, in the next few decades.

2.7 Europe

The liquid petroleum endowment of Europe ranks next to last at 3.8%, and Europe has the second largest production versus consumption deficit (−10.8%). In the North Sea, the region has benefited from the latest exploration and development technology in developing relatively young petroleum provinces. With fields performing much longer and at better rates than once anticipated, Europe remains heavily dependent on foreign supplies of liquid petroleum on a percentage basis and will likely have a continuing reliance on foreign supplies (Tables III–V)

With 36 BB of undiscovered liquid petroleum resources and 24 BB of potential reserve growth (Table II) this region will continue to be an important provider of liquid petroleum, as demonstrated in Table IV.

2.8 South Asia

The South Asia region is next to last in liquid petroleum consumption and has the smallest endowment (0.7%; Table II). It produces approximately 1% of the world's liquid petroleum but consumes 3.3%, thus running a deficit (−2.3%) that is fourth largest in the world (Table III). This region generally has relatively low oil potential but significant gas potential. The liquid petroleum potential resides mostly in the Bombay High, Irrawaddy, and Assam provinces. Recent discoveries in the thrust belt of Pakistan provide optimism but will not meet the increasing liquid petroleum needs of this region. The large population, the need for economic growth, and the current deficit consumption will require considerable and increasing imports of liquid petroleum. The estimated undiscovered liquid petroleum resources of 6.2 BB and reserve growth potential of 5.3 BB are modest.

3. IMPACT OF DISTRIBUTION

As can be seen in Figs. 1 and 2 and Table II, the largest liquid petroleum endowments are concentrated in a few areas, primarily the Middle East, the FSU, and, to a lesser degree, North and South America, North Africa, and sub-Saharan Africa. This uneven distribution of liquid petroleum results in considerable trading of oil as demonstrated in Fig. 3 (data for 2000 to be consistent with the resource data). Generalized oil trade movements depicted in Fig. 3 demonstrate the dominance of the Middle East and FSU, with significant contributions from South America and West Africa. For the United States and Europe, significant interregional trading is required to meet the enormous liquid petroleum consumption. These two regions account for most of the oil trading; that is, North America and Europe consume 51.2% (North America, 31.3%; Europe, 19.9%) but only produce 29% of the resources, ensuring the continuation of large amounts of imports (Table III). Tables IV and V demonstrate that some large producing and consuming nations rank high on both lists, such as the United States, Russia, China, Canada, the United Kingdom, Brazil, and Saudi Arabia. The tables also show that many small producers are large consumers, such as Japan, South Korea, Taiwan, Germany, France, Italy, Spain, and India. This imbalance of supply and consumption drives an increasing degree of world liquid petroleum trading, as shown for 2000 in Fig. 3.

The USGS estimates that future liquid petroleum supplies for the world are significant (Tables I and II). According to 2001 IHS Energy data, world oil reserves are at all-time highs (more than 1.1 trillion barrels). Annual oil consumption is approximately 0.7% of the endowment (28 BBO/year). The world has used approximately 23% of its total oil endowment of approximately 3 TBO. In summary, liquid petroleum remains relatively abundant and will continue to be a dominant source of energy along with natural gas and coal throughout this century. However, these resources are unevenly distributed geographically, and the successful transition to other forms of energy will require a continuing global trading environment and continued investment to access the world's undiscovered and reserve growth resources of liquid petroleum.

Various groups have projected the time at which peak oil production might occur using the USGS data as a reference case. For example, the Energy Information Administration estimates a world oil peak in approximately 2036–2040. However, in 2002, Albert Cavallo suggested a 2015–2020 peak for oil production from non-OPEC countries. Cavallo's analysis suggests even greater dependence on the relatively localized conventional oil resources. Increasingly, natural gas resources, only 7% of which are utilized outside the United States, will be developed and produced. Additionally, development of heavy oil in the Orinoco and tar sand deposits in Alberta is already occurring and may increase. Other remote heavy oil deposits, such as those of the East Siberian Basin in the FSU, can be developed once they are viewed as economic.

4. DATA SOURCES

Information about liquid petroleum resources is dynamic and commonly proprietary. Fortunately, several sources of information are easily accessible. Much of the data on discovered resources in this article was obtained from the U.S. Energy Information Administration. Their Web site (www.eia.doe.gov) provides extensive data on resources and use for all energy sources (not just petroleum) for the United States and the rest of the world. The data for reserve

growth and undiscovered resources are derived primarily from the USGS's 2000 world assessment. The USGS also has a Web site (energy.usgs.gov) containing extensive data for the United States and the rest of the world.

Additional information can be found in several petroleum industry magazines (e.g., *Oil and Gas Journal* and *World Oil*) that publish annual statistical summaries. BP also produces an annual statistical summary and has a Web site (www.bp.com/centres/energy) with much data.

SEE ALSO THE FOLLOWING ARTICLES

Markets for Petroleum • Nationalism and Oil • Oil and Natural Gas: Economics of Exploration • Oil and Natural Gas Resource Assessment: Classifications and Terminology • Oil Industry, History of • Oil-Led Development: Social, Political, and Economic Consequences • Petroleum Property Valuation • Strategic Petroleum Reserves

Further Reading

British Petroleum (2001). BP Statistical Review of World Energy. BP, London. [Available at http://www.bp.com/centres/energy].

Cavallo, A. J. (2002). Predicting the peak in world oil production. *Nat. Resour. Res.* **11,** 187–195.

DeGolyer, and MacNaughton. (2000). "Twentieth Century Petroleum Statistics," 56th ed. DeGolyer & MacNaughton, Dallas, TX.

Edwards, J. S. (2002). Twenty-first century energy: Transition from fossil fuels to renewable, nonpolluting energy sources. *In* "Sustainability of Energy and Water through the 21st Century" (L. C. Gerhard, P. P. Leahy, and V. J. Yannacone, Eds.), Proceedings of the Arbor Day Farm Conference, pp. 37–48. Kansas Geological Survey/American Association of Petroleum Geologists (AAPG) and the AAPG Division of Environmental Geosciences, Lawrence, KS/Tulsa, OK.

Energy Information Administration (2001). "International Energy Outlook 2001." U.S. Department of Energy, Washington, DC. [Available at http://www.eia.doe.gov].

IHS Energy (2001). "Petroleum Exploration and Production Database." IHS, Denver, CO. [Database available from IHS, P.O. Box 740619, Houston, TX 77274–0619]

International Energy Agency (1998). "World Energy Outlook." International Energy Agency/Organization for Economic Cooperation and Development, Paris.

International Energy Agency (2000). "World Energy Outlook." International Energy Agency/Organization for Economic Cooperation and Development, Paris.

International Energy Agency (2001). "World Energy Outlook—2001 Insights." International Energy Agency/Organization for Economic Cooperation and Development, Paris.

Masters, C. D., Attanasi, E. D., and Root, D. H. (1994). *In* "World Petroleum Assessment and Analysis: Proceedings of the 14th World Petroleum Congress," pp. 529–541. Wiley, London.

Minerals Management Service (1996). "An assessment of the undiscovered hydrocarbon potential of the nation's outer continental shelf, OCS Report MMS 96–0034." Minerals Management Service, Washington, DC.

Schollnberger, W. E. (1998, November). Projections of the world's hydrocarbon resources and reserve depletion in the 21st century. *Houston Geol. Soc. Bull.*, 31–37.

U.S. Geological Survey (1995). "National assessment of United States and gas resources,." Circular 1118. U.S. Geological Survey, Washington, DC.

U.S. Geological Survey World Energy Assessment Team (2000). U.S. Geological Survey world petroleum assessment 2000—Description and results, Digital Data Series DDS-60. U.S. Geological Survey, Washington, DC. [Available at http://energy.usgs.gov]

Oil and Natural Gas: Offshore Operations

RON BAKER
University of Texas
Austin, Texas, United States

1. History
2. Scope
3. Exploration
4. Drilling Rights
5. Drilling Rigs
6. Drilling a Well
7. Formation Evaluation
8. Development Drilling
9. Well Completion
10. Production and Workover
11. Oil and Gas Transportation
12. Summary

Glossary

bit The cutting or boring element used in drilling oil and gas wells. The bit consists of a cutting element and a circulating element. The circulating element permits the passage of drilling fluid and utilizes the hydraulic force of the fluid stream to improve drilling rates.

Christmas tree The control valves, pressure gauges, and restrictions (chokes) assembled at the top of a well to control the flow of oil or gas after the well has been drilled and completed.

drill collar A heavy, thick-walled tube, usually steel, placed between the drill pipe and the bit in the drill stem to provide weight on the bit.

drilling fluid A fluid (either gas or liquid) used in rotary drilling. Drilling fluid serves many functions—for example, lifting cuttings out of the wellbore and carrying them to the surface, keeping the hole from caving in, controlling pressure found in underground rock formations, and cooling and lubricating the bit and drill stem.

drill pipe Heavy seamless tubing used to rotate the bit and circulate the drilling fluid.

drill stem All members of the assembly used for rotary drilling from the surface to the bottom of the hole. Includes drill pipe and drill collars.

glycol dehydrator A processing unit that removes all or most of the water from gas. A typical glycol unit consists of an absorber, in which wet gas is put into contact with glycol to remove the water (glycol absorbs the water), and a reboiler, which heats the wet glycol to remove the water from it so the glycol can be reused.

jackup drilling rig A mobile bottom-supported offshore drilling structure with legs that support the deck and the hull. When positioned over the drill site, the bottom of the legs rests on the seafloor. A jackup is towed to position with its legs up; once the legs are firmly positioned on the bottom, the deck and hull height are adjusted and leveled.

mud The liquid usually placed in steel tanks, or pits, on a rig, which is circulated through the wellbore during rotary drilling operations. In addition to bringing cuttings to the surface, mud cools and lubricates the bit and drill stem, holds formation pressures in check, and prevents the hole from caving in.

platform rig An immobile offshore structure from which development wells are drilled and produced. Platform rigs may be built of steel or concrete.

semisubmersible drilling rig A floating offshore drilling unit that has pontoons and columns that, when flooded, cause the unit to submerge in the water to a predetermined depth.

wellhead The equipment installed at the surface of a wellbore. A wellhead includes such equipment as the casinghead and tubinghead, from which special hangers and pressure-tight seals suspend the well's casing and tubing.

Extracting oil and gas from layers of rock that lie beneath oceans and seas presents special problems to oil producers that they do not face on land sites. This article examines some of the special conditions the marine environment imposes on finding, producing, and transporting oil and gas. The petroleum industry produces oil and gas from special layers of rocks called reservoirs. Like a multilayered cake, additional

 © 2004 Elsevier Inc. All rights reserved.

beds of rock lie above and below these reservoirs. And, like the frosting on a cake, a relatively thin layer of ground sometimes covers the rock layers. On the other hand, the "frosting" may not be dry land; it may be water instead. Since oceans and seas cover about three-fourths of the earth, it is no surprise that water also covers rock layers.

1. HISTORY

Edwin Drake drilled the first oil well in the United States in 1859 on a piece of land near Titusville, Pennsylvania. It was only 38 years later, in 1897, that another enthusiast drilled the first offshore well in the United States. He drilled it off the coast of Southern California, immediately south of Santa Barbara in the town of Summerland.

In addition to a pleasant, sunny climate, Summerland had numerous springs. These springs did not, however, produce water; instead, crude oil and natural gas bubbled from them. One person, H. L. Williams, was knowledgeable about extracting oil from the earth. So he drilled several wells in the vicinity of the springs. These early wells were successful and, as a result, he and others drilled many more in the area.

After drilling several wells, they noticed that those nearest the ocean were the best producers. Eventually, they drilled several wells on the beach itself. At this point, however, the Pacific ocean stymied them. Experience convinced them, however, that more oil lay in the rock formations below the ocean. The question was how to drill for it.

Williams came up with the idea of building a wharf and erecting the drilling rig on it. His first offshore well, drilled from a wharf made of wooden pilings and timbers, extended about 300 ft (90 m) into the ocean. On the end of the wharf, Williams erected a drilling rig and used it to drill the first offshore well in the United States. As expected, it was a good producer and before long the entrepreneurs built several more wharves. The longest stretched more than 1200 ft (nearly 400 m) into the Pacific.

2. SCOPE

Offshore activities take place in the waters of more than half the nations on Earth. No longer do primitive, shore-bound wooden wharves confine offshore operators. (Operators are the oil companies that produce oil and gas.) Instead, they drill wells from modern steel or concrete structures. These structures are, in many cases, movable. What is more, they can float while being moved, and often while drilling. Further, offshore rigs can drill in waters that are more than 10,000 ft (more than 3000 m) deep and as far as 200 miles (over 300 km) from shore.

Drilling and producing oil and gas wells are important phases of offshore operations, but the scope goes further. Offshore operations also include exploring—looking for likely places where oil and gas may exist in the rock formations that lie beneath the surface of the oceans, seas, gulfs, and bays. In addition, offshore operations include transporting oil and gas—moving them from their points of production offshore to refineries and plants on land.

3. EXPLORATION

Deep reservoirs buried in rock layers beneath an ocean do not indicate their presence on the surface. Explorationists cannot therefore find them by direct observation. To locate likely rock formations that contain hydrocarbons—oil and gas—companies often employ seismic surveying. Seismic surveying provides a considerable amount of information about subsurface formations. And modern computing power reveals reservoirs that the technology of the past could not locate. Besides powerful computers to interpret the findings, seismic surveying depends on the fact that rock layers reflect sound waves. That is, sound bounces off the layers like an echo. Oil seekers use this reflection to draw a sort of underground picture.

As mentioned before, rocks occur in layers, one on top of the other, much like the layers of a cake. In offshore seismic exploration, a boat trails a special noisemaker that generates a loud, low-frequency sound. The loud, low-frequency sound waves go down through the water and into the rock layers. They then continue traveling downward through the many layers for thousands of feet (meters).

Where one kind of layer ends and another begins, the boundary reflects some of the sound waves. The reflected sound waves travel back up through the layers and into the water. In the water, sensitive detectors called hydrophones, which the boat also tows, pick up the sound waves. The reflections from shallower layers arrive at the hydrophones sooner than the reflections from deeper layers.

The arrival times of each reflection indicate the depth of the layers, giving explorationists a cross-sectional view of the subsurface. Expert interpreters

using powerful computers examine a seismic section and determine whether a geological structure may contain hydrocarbons. Geologists call this cross section a seismic section (Fig. 1).

4. DRILLING RIGHTS

An oil company that wishes to drill for and possibly produce hydrocarbons must obtain the rights to do so. In most cases, the company must obtain these rights from the country in whose waters the oil and gas activity will take place. In the United States, most offshore drilling occurs in the Outer Continental Shelf, or OCS. The Outer Continental Shelf starts at the point offshore where state ownership of the water and of any minerals (such as hydrocarbons) under the water ends. For instance, the state of Texas owns the mineral rights from shore to a distance of 9 miles (14.4 km). Louisiana, on the other hand, owns the mineral rights from shore to a distance of 3 miles (4.8 km).

The U.S. Department of the Interior, Minerals Management Service (MMS), regulates the OCS. The OCS ends where international waters begin or where treaties with other countries, such as Mexico and Canada, establish the end of U.S. jurisdiction. Thus, to explore in OCS waters of the United States, the oil company or companies (sometimes several companies form partnerships) must obtain the right to do so from the federal government. Companies usually obtain this right by bidding on offshore blocks, or tracts, offered for sale by the government at various times. The company or companies that bid the most money are the ones most likely to win the right to drill for and produce any hydrocarbons discovered on the block.

In countries other than the United States, the system is similar, except that most other nations own all the mineral rights from the shoreline out to international waters or the waters of bordering countries. For example, in the North Sea, which is bordered by the United Kingdom, Norway, Denmark, Germany, and the Netherlands, the area is divided into sectors. Each country controls its own sector. In this area, the country in whose sector the drilling site lies grants the oil companies concessions to drill.

5. DRILLING RIGS

Once an oil company obtains the right to drill a wildcat, or exploratory, well to confirm the existence of hydrocarbons in a rock layer, it must then select some type of drilling rig. More often than not, the oil company will use a mobile offshore drilling unit, which is shortened to MODU and is pronounced "moe-doo" (Fig. 2). Rig owners can move mobile

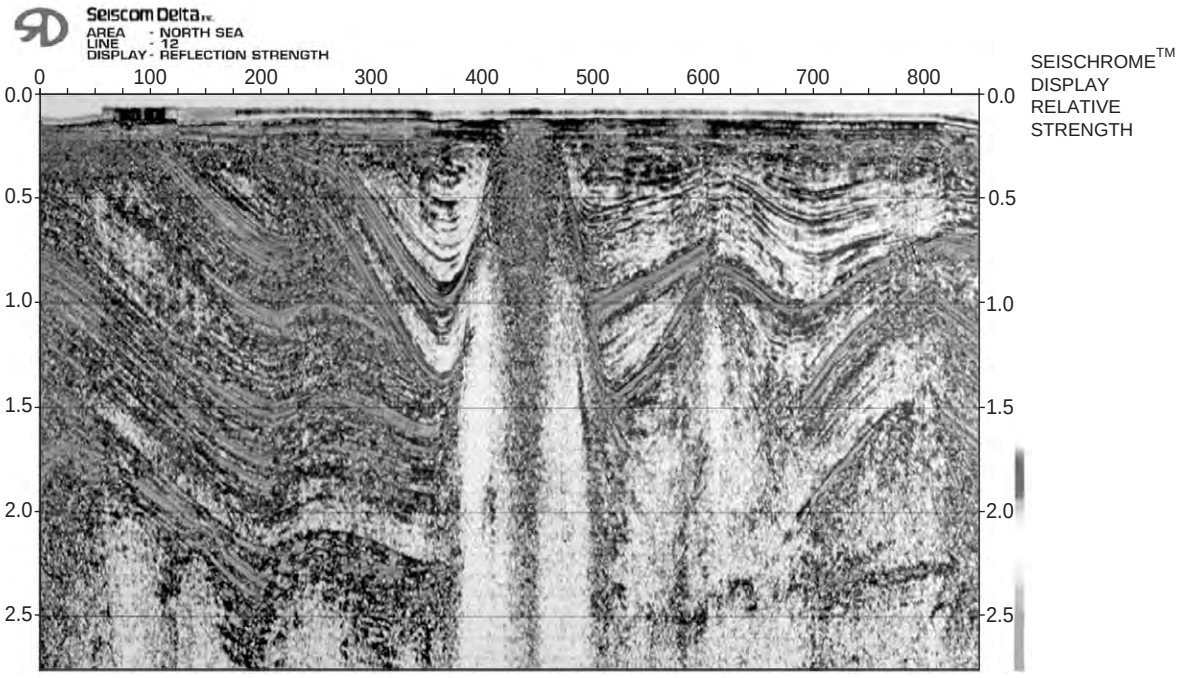


FIGURE 1 A seismic section gives a view of subsurface formations.

FIGURE 2 A mobile offshore drilling unit (MODU).

offshore drilling units from one drill site on the water to another. A rig has to be mobile because, after it finishes drilling one exploratory well, a crew has to move it to another site—perhaps nearby, perhaps far away—to drill another. Oil operators use two basic types of MODUs to drill most offshore wildcat wells: bottom-supported units and floating units. Bottom-supported units include jackups. Floating units include semisubmersibles and drill ships (Fig. 3).

5.1 Jackups

Jackups are popular because they provide a stable drilling platform. They are stable because part of their structure firmly contacts the bottom of the ocean. They can also drill in relatively deep water: up to 350 ft (100 m) deep. What is more, towboats can easily move a jackup from one location to another.

Most jackups have open-truss legs. In this design, manufacturers construct the legs from tubular steel members. They then crisscross the members to make very strong, but relatively lightweight, legs. Open-truss legs look somewhat like the towers that carry high-voltage electric lines across the countryside. The legs of a jackup pass through openings in a barge

FIGURE 3 A drill ship.

FIGURE 4 A special ship carries a semisubmersible rig.

hull. Three or four legs are common, but some designs call for more or fewer. The deck of the barge serves as the platform for drilling equipment and other machinery. When a crew moves a jackup, it raises, or jacks, the legs up out of the water so that the barge floats. With the legs completely out of the water, the rig movers can transport the entire rig to the drilling location. Usually rig movers tow the rig.

Sometimes, large ships with flat decks move jackups. Crew members submerge the ship so that its deck is below the water. They then maneuver the rig onto the submerged ship's deck. With the rig in place, crew members pump out the water from the ship, which allows it to float. Especially for long-distance moves, ships can carry the jackup at speeds faster than boats can tow it. Ships can also carry semisubmersibles (Fig. 4).

Once the jackup rig is on location, crew members jack the legs down until they come into firm contact with the seafloor. Then they jack up the barge hull onto the legs until it is clear of the water's surface and well above high waves.

5.2 Floating Units

Floating rigs include drill ships and semisubmersibles. Generally, oil companies most often use drill ships to drill wells in the deep, remote waters of the world. Semisubmersibles can drill in seas too rough for drill ships. Drill ships have a ship-shaped hull that allows the drilling contractor (the rig owner that the oil company hires to drill) to store a large amount of supplies and equipment on board the vessel. Thus, a drill ship can stock up in port, sail to the drilling location, and drill for long periods without being restocked.

Semisubmersibles have large pontoon hulls and big columns, which the rig operator submerges to a depth below the water's surface. Submergence makes semisubmersibles the most stable of floating drilling units. Even though semisubmersibles cannot carry as much drilling supplies as a drill ship—they do not have a hull for storage—drilling companies often have to use them to drill in rough seas.

6. DRILLING A WELL

A writer once described a drilling rig as a portable hole factory. He pointed out that the sole purpose of a rig was to make holes in the ground, and since drilling contractors had to drill holes at different locations, the rig had to be movable. The writer wasn't too far off the mark. A rig—be it large or small, on land or offshore—has one main job: drilling wells.

6.1 Bits and Drilling Fluid

The job of a rig and the people who work on it is to put a drill bit in the earth and turn it. A bit is a hole-boring device (Fig. 5). Pressed very hard against the ground and turned, or rotated, it makes a hole. The teeth on the bit grind and gouge the rock into small pieces. These pieces of rock, or cuttings, must be moved out of the way so the bit teeth can be constantly exposed to uncut rock. To move cuttings away from the bit, the rig pumps a special liquid called drilling fluid, or mud, down to the bit. Most often, drilling mud is a mixture of water, special clays, and certain minerals and chemicals.

Special openings in the bit called nozzles eject drilling mud out of the bit with great speed and pressure. These jets of mud lift the cuttings off the bottom of the hole and away from the bit. With the cuttings out of the way, they do not interfere with the bit teeth's ability to drill. Mud also carries the cuttings up the hole and to the surface for disposal.

FIGURE 5 A bit has teeth (cutters) to penetrate a formation and make a hole.

The mud pump in the circulating system moves drilling mud down the hole through special pipe called drill pipe and drill collars, which are called the drill string. Mud then shoots out of the bit through jets, picks up the cuttings, and carries them to the surface. On the surface, special equipment removes the cuttings. Then, the mud pump recirculates the clean mud back down the hole.

6.2 Rotating System

Part of keeping the bit drilling involves rotating it, so the rig needs a rotating system. Commonly, offshore rigs use a top drive to rotate the bit. Top drives have a very powerful built-in electric motor (some have more than one motor). The motor or motors rotate a drive shaft. Crewmembers make up the top of the drill string to the drive shaft. When the driller activates the top drive's motor and drive shaft, the drive shaft rotates the attached drill string and bit.

6.3 Power System

Any drilling rig needs power—power to actuate the mud pumps, to hoist the drill string, and to run all

the machinery on the rig. Several large diesel engines usually provide the power. A diesel engine is an internal-combustion engine, which means that it runs because a mixture of fuel and air burns inside the engine. Electricity transmits engine power—that is, the diesel engines put out mechanical power. The mechanical power turns electrical generators. The generators produce electricity, which heavy-duty cables carry to powerful electric motors on the equipment needing power. Drillers control the transfer of power with controls and switches from their position on the rig floor.

6.4 Hoisting System

Another major system on the rig is the hoisting system, which is a very large block-and-tackle system (Fig. 6). It consists of the drawworks, or hoist; a mast or a derrick; the crown block; the traveling block; and wire rope. It suspends the drill pipe and drill collars in the hole. (Drill collars are extra heavy pipe that is used to put weight on the bit to make it bite into the formation and drill.)

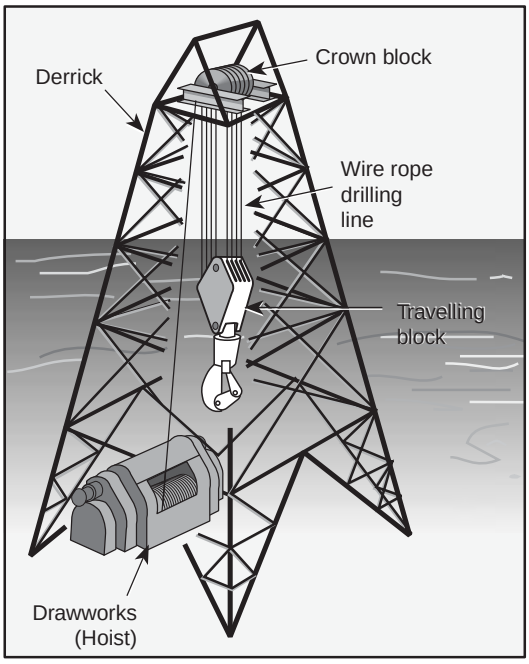


FIGURE 6 The hoisting system consists of several components, all of which serve to suspend and move drilling equipment in and out of the hole.

7. FORMATION EVALUATION

Whether the oil company drills an exploration well from a bottom-supported or a floating unit, it must evaluate prospective formations—that is, the company must determine whether the well has struck oil or gas in sufficient quantities to justify further reservoir development. Because the cost of developing, or exploiting, offshore reservoirs is so high, the quantity of oil or gas must be large. In fact, companies sometimes abandon reservoirs discovered offshore even though they contain a fair amount of oil and gas. Evaluation involves measuring certain properties of a potential reservoir rock and obtaining samples of the rock and any fluids that may be in it.

7.1 Well Logging

Well logging is an evaluation method in which a logging crew lowers a special tool, a sonde, into the well and then pulls it back up. As the sonde passes the formations on its way up the wellbore, it senses and measures electrical, radioactive, and acoustic (sound) properties of the rocks. Measuring and recording these properties provides a good deal of information about the rocks. The sonde transmits its measurements via conducting cable to the surface, where computers record them. Experts analyze the resulting recording, the log (Fig. 7). Analysis involves careful study of the curves on the log. The curves indicate whether hydrocarbons exist in the formations investigated and to what extent.

7.2 Coring

Coring is another method of formation evaluation. The operator—the oil company— obtains one type of core sample by lowering a sidewall sampler into the borehole. At the desired interval, the driller activates the sampler to fire several small cylinders into the wall of the hole. The cylinders penetrate a short distance into the rock and cut a small core. When crewmembers start to remove the tool from the hole, cables attached to the cylinders pull them from the wall of the hole, and the cores come up the hole with the tool. Geologists can then examine the small cores and analyze them for indications of the character of the formation.

7.3 Drill Stem Testing

A drill stem test (DST) is a method of formation evaluation. The test allows fluids in the formation being evaluated to flow into the drill stem and to the surface. Pressure-recording devices in the drill stem test tool measure and record pressure in the well while it is both flowing and not flowing (shut in).

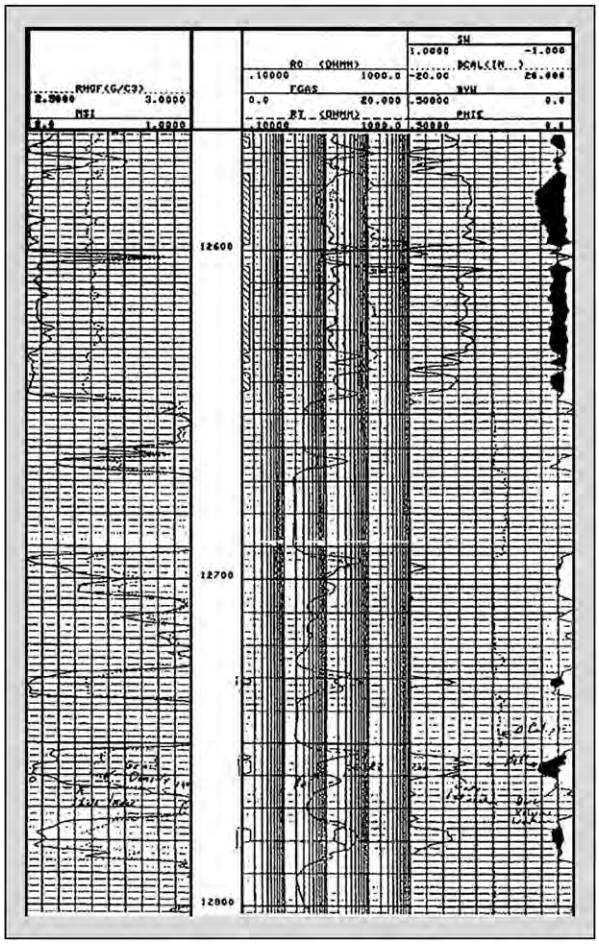


FIGURE 7 Curves on this well log reveal information about formations penetrated by the hole.

A DST can also obtain actual fluid samples from the formation. From detailed analysis of the pressure recordings and of the fluids, the operator can infer a great deal of information about the reservoir.

8. DEVELOPMENT DRILLING

If tests on an exploratory well prove favorable, the oil operator usually drills and evaluates several additional wells in or near the reservoir. These additional wells are appraisal wells. The operator drills appraisal wells to confirm further that the reservoir contains enough hydrocarbons to justify the enormous expense of developing it. If the appraisal wells reveal that the reservoir contains enough hydrocarbons, then development drilling may occur. Development drilling is the drilling of several wells into a reservoir to extract hydrocarbons discovered by exploratory wells and confirmed by appraisal wells.

FIGURE 8 A self-contained platform.

Operators use many kinds of rigs to drill development wells. A common type in U.S. waters is the platform rig. Operators also use mobile offshore drilling units to drill development wells, the same rigs they often employ to drill exploratory and appraisal wells.

8.1 Drilling Platforms

Oil companies, or operators, drill a large number of offshore development wells from totally fixed, self-contained platforms. The operator drills, completes, and produces all the development wells from the platform, thus eliminating the need for rig mobility. Modern platforms house all the drilling equipment, production equipment, crew quarters, offices, galley, and recreation rooms (Fig. 8). Manufacturers build some of steel and some of concrete.

A steel-jacket platform consists of a jacket—a tall, vertical section made of tubular steel members—which is the foundation of the platform. Piles driven into the seabed support it. The jacket extends upward so that the top rises above the waterline. Additional sections on top of the jacket provide

space for crew quarters, a drilling rig, and all the equipment needed to drill. The height of the jacket depends on the water's depth. Some of the tallest stand in water over 1000 ft (300 m) deep.

In the North Sea, rig builders make concrete gravity platforms from steel-reinforced concrete. They construct them in deep, protected waters near shore. They then float and tow to the site in a vertical position (Fig. 9). Some have three or four tall concrete caissons, or columns, that resemble smokestacks. At the site, the operator floods the caissons, and the platform submerges to come to rest on bottom.

One end of the caissons rests on the seafloor. The rig builders install drilling equipment, crew quarters, and production equipment on a platform on the other end, well above the waterline. Concrete platforms weigh a tremendous amount, so builders do not have to pin them to the bottom with pilings. The force of gravity alone keeps them in place.

8.2 Mobile Offshore Drilling Units

In some cases, an operator uses a MODU, such as a jackup, to drill development wells. For example, the operator can erect a relatively small platform jacket, only large enough to provide space for drilling and production wellheads on the site. In such cases, the drilling contractor places the MODU next to the jacket and drills and completes the wells on it. After the wells are drilled, the contractor moves the MODU away and, if necessary, the operator erects an additional platform nearby to house production equipment and quarters for the production crew. A system of two relatively small platforms and a MODU can reduce costs below those of a single large drilling and production platform. Operators can also use mobile rigs to drill development wells into a reservoir, perhaps into parts of it that are not accessible from a platform located over another section. They drill these satellite wells one at a time—that is, the MODU drills one well and then moves to drill other wells at different locations over the reservoir. This method of drilling a single development well and then moving the rig is practical only with MODUs.

FIGURE 9 Boats tow a concrete platform to a site in the North Sea.

Sometimes it is not economically efficient to drill several single development wells at various points into a reservoir. In such cases, operators can use MODUs with subsea templates. A subsea template is an arrangement of special equipment the operator places on the seafloor. The template allows the operator to drill, complete, and produce several wells from the relatively small point on the seafloor that the template covers. The operator places the MODU, usually a floater, on the water's surface over the template, and drills and completes several wells on the template.

9. WELL COMPLETION

After operators drill a development well into the reservoir and run and cement the final string of casing, they must complete the well. The operator must provide a way for the reservoir fluids to enter the well and reach the surface.

9.1 Perforating

To open the well to the reservoir, the operator can perforate the casing and cement that line the well drilled into the reservoir. That is, a perforating company can make several small holes in the casing and cement. Hydrocarbons flow from the reservoir into the well through the perforations.

To perforate a well, a perforating operator lowers a special perforating gun to a point opposite the producing formation. The operator fires the gun, and several small but intense explosions penetrate the casing and cement and travel a short distance into the reservoir. These small explosions are shaped charges—explosives that release a high-energy charge of gases and particles. These gases and particles travel in one direction and penetrate casing, cement, and formation. Reservoir fluids then enter the well through the perforations.

9.2 Tubing and Packers

Usually, operators run a special string of pipe inside the casing. This special pipe is tubing. Reservoir fluids flow up the well through the tubing that is inside the

cemented casing. Operators produce most wells through tubing instead of directly through the casing. They produce through tubing for several reasons. For one thing, the crew does not cement the tubing in the well. As a result, when a joint of tubing fails, as it almost inevitably will over the life of the well, the operator can easily replace the failed joint. Since casing is cemented, it is very hard to replace. For another thing, tubing allows the operator to control the well's production by placing special tools and devices on or in the tubing string. These devices allow operators to produce the well efficiently. Casing does not provide a place to install tools or devices required for production. In addition, the operator usually installs safety valves in the tubing string. These valves automatically shut off the flow of fluids from the well if damage occurs at the surface. Finally, tubing protects casing from the corrosive and erosive effects of produced fluids. Over the life of a well, reservoir fluids tend to corrode metals with which they are in contact. Producing fluids through tubing, which the operator can easily replace, preserves the casing, which is not so easy to repair or replace.

Crewmembers often install a packer near the bottom of the tubing above the producing zone. A packer is a length of pipe through which well fluids flow. The packer has gripping elements that latch onto the inside wall of the casing to help anchor the packer. Rubber sealing elements on the packer, when expanded, provide a pressure-tight seal between the tubing and casing. The packer forms a fluid-tight seal between the tubing and the casing. Thus, when reservoir fluids flow into the well, the packer forces them to flow into the tubing.

9.3 Surface and Subsea Completions

Completing a well also involves the control of fluids as they flow from the top of the well. The operator can place the devices that control fluid flow either above the surface of the water or onto the seafloor. In a surface completion, the operator places the equipment that controls the flow of hydrocarbons from the well above the waterline. Operators usually complete the wells they drill from a platform on the platform itself. In a subsea completion, the operator places the equipment that controls the flow of hydrocarbons from the well on or below the seafloor.

In a subsea completion, the operator locates the collection of special valves and fittings that controls the flow of hydrocarbons from the well—the so-called Christmas tree—at the top of the well on the seafloor. A subsea Christmas tree can be wet or dry. If wet, the operator installs it on the wellhead where it remains exposed to the surrounding water. If dry, the operator installs the Christmas tree in some type of housing that isolates it from the water, often at atmospheric pressure.

Divers may maintain and repair wet trees; sometimes, the operator utilizes remotely operated vehicles (ROVs) to make repairs. Operators can maintain some dry trees with small submarines called submersibles or diving bells to gain access to the housing around the Christmas tree. Workers enter the housing without any special clothing or underwater breathing equipment.

10. PRODUCTION AND WORKOVER

Offshore production involves a wide range of techniques and equipment. The operator uses the techniques and equipment to get oil and gas out of the reservoir and into a transportation system. As produced fluids flow up the well and to the surface, the operator has to separate oil, gas, and water, and must handle each according to its individual requirements. As a result, a typical production platform may have a large amount of equipment and several people to successfully extract, treat, and move oil and gas to shore for refining and processing.

10.1 Reservoir Drive Mechanisms

Natural pressure in the reservoir forces oil and gas to flow from a reservoir into a well and to the surface. The main sources of this pressure are gas, water, or both in combination. Four kinds of natural drive mechanisms occur with oil and gas: dissolved-gas drive, gas-cap drive, water drive, and combination drive (Fig. 10).

Dissolved-gas drive is present when all or almost all of the hydrocarbons in the reservoir are liquid in its undrilled state. When the operator drills a well into such a reservoir, the well creates an area of reduced pressure. It is somewhat like jamming a straw into a bottle of liquid and sucking on the straw. Sucking on the straw reduces pressure and liquid flows up the straw. Similarly in a reservoir, reducing the well's pressure causes some of the lighter liquid hydrocarbons to become gaseous. The gas comes out of solution and, as it does, it drives oil into the well and to the surface.

Gas-cap drive results when a reservoir contains so much gas that all of it is not dissolved in the oil. Because gas is lighter than oil, it usually rises to the

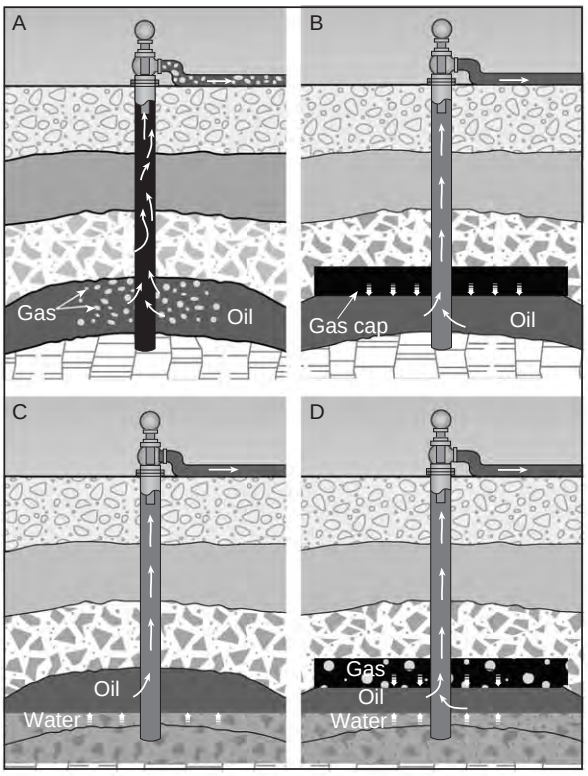


FIGURE 10 Dissolved gas drive (A), gas-cap drive (B), water drive (C), and combination drive (D).

top of the reservoir and forms a cap over the oil. When the operator completes a well into such a reservoir, the gas cap expands. The expanding gas drives oil into the well and to the surface. Water drive occurs when large amounts of water—usually salt water—lie beneath the oil in a reservoir. When the operator drills a well, the water in the reservoir moves toward the well and drives oil ahead of it. Combination drive exists when both gas and water provide energy to drive oil to the surface. Gas—in the form of a cap or dissolved in the oil or a combination of the two—and water lying below the oil work to lift oil to the surface. Combination drive is the most common drive mechanism.

10.2 Handling Oil, Gas, and Water

Oil and gas seldom reside in a reservoir by themselves. Water and other impurities like sediment usually coexist with them. As operators produce the reservoir, these impurities come to the surface with the oil and gas. Operators must separate and remove the foreign material from the hydrocarbons. They must also separate the oil and gas from each other before they can transport them to shore for refining and processing. Operators must separate and remove impurities—especially sediment and water (S&W)—because the impurities take up space in a tanker or pipeline that valuable hydrocarbons could otherwise occupy. Furthermore, water can corrode practically any steel surface, including that of a tanker compartment or a pipeline.

Operators have to separate oil from gas primarily because each has different handling requirements. For instance, the operator usually has to put gas under additional pressure to put it into a pipeline or special tanker for transport to shore facilities. Because oil is a liquid, it cannot be compressed, and therefore must be pumped into a pipeline or tanker. To separate and handle oil, gas, and water, operators use several techniques and a variety of equipment. They usually locate the equipment on the same platform from which they drilled the wells. But once they start producing the wells, the platform becomes a production platform and operators use the derrick for such purposes as raising and lowering equipment to work over the well. Because it's not being used strictly for drilling at this time, the derrick can be used to run special tools into wells that are producing but that need cleanout or repair.

In a subsea completion, the operator routes oil, gas, and water into flow lines that run from the wells to a production riser. Production flows up the riser to a floating buoy. An oil tanker can then tie up to the buoy, take on the produced fluids, and transport them to shore for separation and treatment. In some cases, the operator moors a processing and storage ship to the buoy. Equipment on the ship separates, treats, and stores the reservoir fluids. When oil fills up the ship's tanks, the ship transports it to a receiving terminal on shore. Operators sometimes use this system in deep waters where erecting a platform is not economically feasible.

In shallower waters, operators can send production from subsea wells to a platform. Once on the platform, equipment separates the fluids and offloads the oil onto a tanker or injects it into a pipeline.

Water produced with oil and gas occurs in three ways: (1) as free water, (2) combined with oil in an emulsion, or (3) mixed with gas as vapor. Free water is water not tightly bound to the oil and gas. It readily settles out if the operator gives the produced fluids a chance to remain stationary for a short period in a tank or vessel. Thus, one type of equipment producers sometimes use on an offshore production facility is a free-water knockout (FWKO). A free-water knockout is usually a cylindrical vessel. The operator pipes well fluids that

contain relatively large amounts of free water after they leave the Christmas tree into the FWKO. An FWKO can be either vertical (standing upright) or horizontal (lying on its side). Vertical designs take up less space but cannot handle as much volume as horizontal ones. Horizontal vessels expose the fluids inside them to more surface area. The design selected depends primarily on the volume, or quantity, of the fluids being produced and the amount of space available on the platform.

When well fluids go into a free-water knockout, the free water, since it is heavier than oil or gas, falls to the bottom of the vessel. The operator then removes it and disposes of it properly. If significant amounts of gas are present in the fluid stream, the manufacturer can equip the knockout to handle the gas. In such cases, the gas rises to the top of the knockout. The producer removes the gas and sends it to other equipment on the platform for further handling. Oil generally winds up in the middle. The operator pipes the oil to equipment on the platform for further handling.

Sometimes, water tightly binds to the oil in the produced fluids. Usually, this water spreads out (disperses) in the oil as small droplets that cannot readily settle out. This combination of oil and water is an emulsion. The operator specially treats emulsions to remove the emulsified water from the oil. To treat an emulsion, producers use special vessels—heater-treaters. They pipe the emulsion into the heater-treater and use chemicals, electricity, heat, or all three to break the emulsion. Breaking the emulsion means making the water droplets coalesce (merge) and settle out of the oil. Usually, the operator adds chemicals to the emulsion and pipes it into the heater-treater. Heater-treaters, like free-water knockouts, can be either vertical or horizontal. Chemical companies formulate the chemicals to make the small water drops merge into large drops. Large water drops fall from the oil to the bottom of the treater where the operator removes them.

Electricity can also cause small water drops to merge and fall to the bottom of a treater. If an emulsion, usually with emulsion breaking chemicals in it, passes over a high-voltage electric grid in the treater, the water droplets merge and fall. Electrostatic treaters use electricity to treat, or break, an emulsion.

Heat reduces the viscosity, or thickness, of oil. Since water can fall out of a heated, thin oil more easily than it can out of a cool, thick oil, operators sometimes use heater-treaters in the emulsion-breaking process. In such cases, a fire tube, which is an enclosed gas flame, heats the emulsion while it is in the vessel. Operators often use chemicals and electricity in addition to heat.

10.3 Separation of Oil and Gas

To separate oil and gas from one another, the producer usually employs a separator. Most separators are vertical or horizontal, although a spherical (ball-shaped) design is available. All provide an enclosed space into which the operator pipes produced fluids. Given a chance to remain stationary for a short period, the force of gravity makes the fluids separate. Some are two-phase separators, which means that only oil and gas or emulsion and gas separate inside them. Others are three-phase separators, which means that oil (or emulsion), gas, and water separate inside them. If only relatively small amounts of free water are in the produced fluids, or if some free water remains after the fluids pass through a free-water knockout, then the producer may use a three-phase separator.

As separation occurs inside the separator, the gas goes to the top of the vessel. The operator pipes it to additional gas-handling facilities on the platform—to a glycol dehydrator, for example. If water separates out, operators remove it from the bottom and dispose of it. They remove oil from the middle (or bottom, if they are using a two-phase separator) and transport it to shore if the oil does not require further treatment. Usually, however, water is emulsified in the oil, so the operator pipes this emulsion to treaters elsewhere on the platform. After the treaters break out most of the emulsified water, the operator transports the treated oil to shore.

10.4 Artificial Lift

As natural reservoir pressure produces hydrocarbons, this natural pressure declines. Eventually, the pressure dwindles to the point that it can no longer lift oil to the surface. Natural reservoir pressure usually depletes long before the operator is able to extract all the hydrocarbons from the reservoir. So operators turn to artificial lift to produce at least some of the remaining hydrocarbons. In artificial lift, the operator applies some form of external energy to the well. The applied energy must reduce pressure at the bottom of the well. It must reduce the bottom-hole pressure enough so that the small amount of remaining reservoir pressure can force more reservoir fluids into the well and lift them to the surface. Offshore, the most common form of artificial lift is gas lift. In gas lift, compressors inject natural gas into

a well. This gas lowers pressure at the bottom of the well. Gas is a very lightweight, or low-density, fluid. A low-density fluid like gas does not exert as much pressure as a liquid. Injecting gas to the bottom of the well forces most of the liquids out of the well so that a mixture of liquid and gas remains. This liquid-gas mixture is lighter than the original liquid column and thus exerts less pressure on the reservoir at the bottom of the well. Since reservoir pressure is now higher than pressure inside the well, fluids from the reservoir flow into the well and to the surface.

10.5 Improved Recovery Techniques

Artificial lift cannot produce all of the hydrocarbons that reside in a reservoir. In spite of every effort, a fair amount of oil is left behind even after artificial lift is applied. The laws of physics, the behavior of fluids in a reservoir, and the nature of the reservoir itself all combine to make it impossible to extract every bit of oil in a reservoir. In some reservoirs, up to 75% of the hydrocarbons originally in place are left behind. On the average, an oil company can produce only about 33% of the total amount of oil in the reservoir. In an attempt to recover at least some of the remaining oil, companies can apply improved recovery, which is the application of various fluids, chemicals, and heat to a reservoir. Improved recovery techniques, if successful, can remove additional amounts of oil from the reservoir that is left behind by natural and artificial lift.

10.6 Well Servicing and Workover

As wells produce and start aging, downhole equipment wears out, sand from the reservoir gets into the well and restricts fluid flow, gas-lift valves malfunction, corrosion eats away metal parts, and other problems occur. In some cases, the operator hires a well service and workover contractor, a company that specializes in offshore well repair, to repair or maintain the well. In other cases, the operator turns to the company's own employees who have the expertise and equipment to do it themselves. In either case, the goal of well servicing and workover is to fix whatever is wrong and restore production to a normal rate or to a rate as near normal as possible.

Operators and contractors use several techniques, tools, and methods to repair or maintain wells. On large, self-contained platforms, the operator often retains the derrick used for drilling the original wells and uses it for servicing those same wells. For instance, if part of the tubing string fails—maybe corrosion eats away part of it—a repair crew has to pull the entire string out of the well. The crew then replaces the damaged portion and run the string back in. Having the original derrick in place makes pulling and running tubing relatively easy. On the other hand, if the platform is small and has no derrick or if a satellite well needs to have its tubing pulled, then the operator may move in a mobile workover rig. Both jackup and semisubmersible workover rigs are available for servicing and workover.

Special wireline tools can run, manipulate, and pull much of the equipment operators place in the well when they complete it. Wireline is strong, fine wire that spools onto a reel. Repair crews can attach various tools and devices to wireline and lower or pull it from the well in one continuous movement. Using wireline is much faster than running and pulling a tubing string. Since wireline is one continuous strand, crewmembers do not need a derrick to run tools and equipment on it.

For servicing subsea completions, repair crews can pump special tools down the flow line and into the tubing string to do a specific job. After they complete the job, they circulate the tool back to the surface. Such pumpdown tools are often called through-the-flow-line tools, or TFL tools for short. TFLs have simplified remedial work on subsea completions. Also, operators use them on surface completions as well.

11. OIL AND GAS TRANSPORTATION

After operators produce, separate, treat, and, in some cases, store oil and gas on an offshore production facility, they must ultimately send them to shore. Once on land, refineries and processing plants make products that practically everyone in the world uses. Companies normally transport oil and gas in two ways: in pipelines or in tanker ships. A pipeline is a series of connected pipes through which a company sends oil or gas. One end of the pipeline originates at a production facility and the other end terminates at a facility on shore. Pumps or compressors inject the oil or gas into the pipeline.

Building an offshore pipeline is a complex operation. It often involves the laying of several miles (kilometers) of large-diameter pipe on the seafloor under adverse conditions. Usually, the pipeline construction company coats all but the ends of the pipe with concrete. Concrete protects the pipe, helps prevent corrosion caused by seawater, and

weights the pipe so that it will stay on bottom. The construction crewmembers leave the pipe ends free of concrete so they can weld the joints together. They then coat the welded joints to prevent them from corroding. If the pipeline carries crude oil, the terminal on shore will probably consist in part of several large storage tanks. From these tanks, land pipelines send oil to refineries. A gas pipeline generally terminates at a facility capable of processing the gas. Gas processing involves recovering the heavier hydrocarbon components of natural gas, such as propane and butane. After the processing plant breaks down the gas into its components, it then pipelines the components to other plants for additional processing or to consumers.

An offshore pipeline construction company can lay an offshore pipeline in several ways. Commonly, companies will use a special barge, a lay barge. It floats on the water's surface, or, if the water is rough, it floats in a semisubmerged state, much like a semisubmersible drilling rig. Think of a lay barge as a floating offshore assembly line. Each step in the operation—welding, inspection, wrapping, laying—occurs at a station on the barge as a special machine conveys the pipe along the barge's length. Ideally, once the operation starts, it won't stop until the line is completely laid. Unfortunately, bad weather and unforeseen events can bring things to an abrupt halt and increase what are already very high costs.

A lay barge carries a large number of pipeline joints. It also carries the personnel and equipment necessary to weld the joints together, inspect them for integrity, apply a protective coating on the welded joints, and place the pipe on the seafloor. Further, as the welding machines join each joint, the barge has to be able to move forward so that it lays the pipe in a steady and continuous motion with as few delays as possible. Since the lay barge usually cannot carry enough pipe to complete the job, supply boats shuttle pipe to the barge from a shore base.

Sometimes, it is not feasible for a company to build a pipeline to carry oil to shore. For example, if an area's production is not very large, the expense of building a pipeline could be more than the income received from the sale of oil. Or building a pipeline at a particular site could be physically impossible. Whatever the reason, when companies cannot pipeline oil to shore, they often use oil tanker ships instead. In the case of an offshore gas field, the operator usually must build a pipeline because no other practical way exists to transport gas. In a few cases, however, a company may build an offshore gas-liquefaction facility—where the facility cools natural gas to an extremely low temperature and liquefies it. The company can then transport the liquid gas to shore via tankers specially designed to handle liquefied natural gas (LNG). On the other hand, a company can produce oil from a platform or subsea wells and pipe it to a special mooring system. A tanker ship can tie up to the system, take on oil, and then transport the oil to an onshore terminal for unloading.

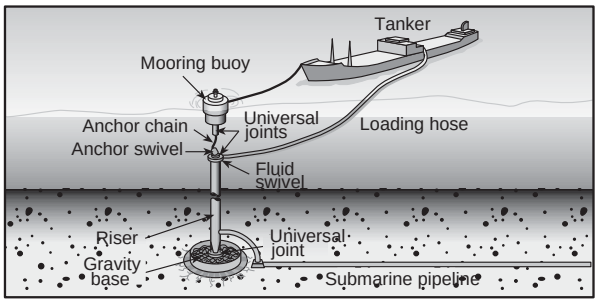


FIGURE 11 Single point buoy mooring (SPBM) system.

The most common type of mooring system is single-point buoy mooring (SPBM) (Fig. 11). An SPBM system is an installation in the sea where a tanker can take on a crude oil cargo. It is called a single-point system because it provides only one point to which the tanker ties up, or moors. Oil to the SPBM system comes from wells on a platform or from subsea wells.

12. SUMMARY

Offshore operations cover a broad range of activities that include exploration, exploration drilling, development drilling, production, workover, and transportation. Offshore exploration includes the use of seismic surveys. Seismic survey methods yield evidence of the presence of hydrocarbons in formations that lie below the oceans and seas of the world. Should exploration reveal the likelihood of hydrocarbons at a certain site, then drilling a well becomes necessary. Often, operators use a mobile offshore drilling unit—a MODU—to drill the well. Because companies can evaluate a well after they drill it, they can find out for sure whether oil and gas reside in a potential reservoir. Evaluation techniques such as logging, drill stem testing, and coring can confirm that the exploration well has penetrated a hydrocarbon-bearing formation.

If the wildcat well shows what appear to be ample quantities of oil and gas, a company may drill appraisal wells. Appraisal wells further confirm that

the reservoir holds enough hydrocarbons to justify the enormous expense of developing the reservoir. If commercial amounts of oil or gas do exist in a reservoir, the company must develop the reservoir. It may move a self-contained platform to the site and erect the platform. The platform may be a rigid steel or concrete structure. Usually, the operator drills several development wells from the platform. On the other hand, the operator may place a subsea template on the seafloor, and drill the wells from a MODU. Whether from a platform or on a template, the operator drills and completes several wells to produce hydrocarbons from the reservoir.

Producers may drill satellite wells where they cannot get to the reservoir from a platform or a template. Satellite wells and wells drilled on a template are often subsea completions—that is, producers install the Christmas trees on or below the seafloor. Wells that operators complete on platforms are usually surface completions: they install the Christmas trees on a platform deck above the waterline.

The production phase of offshore operations includes the installation of special pieces of equipment to separate and treat crude oil, natural gas, and water. In most cases, operators install the equipment on a platform—usually the same platform they drilled the wells on. Sometimes, however, they install the equipment on a special processing and storage ship. They most often use such ships with subsea completions.

Operating companies may have to put wells produced from platforms and from special ships on artificial lift. They will most likely use gas lift, which allows wells to produce after natural drive depletes. In some cases, applying additional recovery techniques to extract still more hydrocarbons from an aged and dying reservoir may be possible.

Once a company completes and puts development wells on production, it will need to service and repair the wells to keep them producing at the desired rate. In many instances, operators can use wireline and pumpdown repair methods, but in other instances, they may have to move in a semisubmersible or jackup well service and workover unit.

Finally, a company may build a pipeline to transport the oil and gas to refineries or plants for the extraction of useful products. Or perhaps tanker ships will shuttle oil to shore. If so, the operator will install special mooring systems. Offshore operations occur all over the world because it is beneath the oceans and seas where the petroleum industry will find most of the world's future hydrocarbons.

SEE ALSO THE FOLLOWING ARTICLES

Crude Oil Spills, Environmental Impact of • *Gas Hydrates* • *Natural Gas, History of* • *Occupational Health Risks in Crude Oil and Natural Gas Extraction* • *Oil and Natural Gas Drilling* • *Oil and Natural Gas Exploration* • *Oil and Natural Gas Liquids: Global Magnitude and Distribution* • *Oil Industry, History of* • *Oil Pipelines* • *Oil Recovery* • *Public Reaction to Offshore Oil*

Further Reading

Baker, R. (1983). "Oil and Gas: The Production Story." Petroleum Extension Service, University of Texas at Austin, Austin, TX.

Baker, R. (2001). "A Primer of Offshore Operations." Petroleum Extension Service, University of Texas at Austin, Austin, TX.

Gray, F. (1995). "Petroleum Production in Nontechnical Language." PennWell, Tulsa, OK.

Hosmanek, M. (1984). "Pipeline Construction." Petroleum Extension Service, University of Texas at Austin, Austin, TX.

Hyne, N. J. (1995). "Nontechnical Guide to Petroleum Geology, Exploration, Drilling, and Production." PennWell, Tulsa, OK.

Johnson, D. E., and Pile, K. E. (1988). "Well Logging for the Nontechnical Person." PennWell, Tulsa, OK.

Kennedy, J. L. (1993). "Oil and Gas Pipeline Fundamentals." PennWell, Tulsa, OK.

Link, P. K. (2001). "Basic Petroleum Geology." Society of Petroleum Engineers, Richardson, TX.

Van Dyke, K. (1996). "A Primer of Oilwell Service, Workover, and Completion." Petroleum Extension Service, The University of Texas at Austin, Austin, TX.

Van Dyke, K. (1997). "Fundamentals of Petroleum." Petroleum Extension Service, University of Texas at Austin, Austin, TX.

Oil and Natural Gas Resource Assessment: Classifications and Terminology

T. R. KLETT
United States Geological Survey
Denver, Colorado, United States

1. **Introduction**
2. **Commodities Considered in Oil and Natural Gas Resource Assessments**
3. **Classification of Oil and Natural Gas Accumulations**
4. **Assessment Units**
5. **Exploration and Discovery History**
6. **Assessment Concepts**
7. **Parameters for Classification of Oil and Natural Gas Resources**
8. **Classification of Oil and Natural Gas Resources**

Glossary

accumulation One or more reservoirs of petroleum that have distinct trap, charge, and reservoir characteristics. For purposes of assessment, an accumulation is treated as a single entity. An accumulation may encompass several fields or be equivalent to a single field or reservoir; alternatively a field may be equivalent to one or several accumulations.

assessment An evaluation or appraisal of the quantity, quality, and composition of technically or economically recoverable petroleum or in-place petroleum that provides information on which an estimate of value of the resources on land and in offshore areas can be made.

field An area consisting of a single reservoir or multiple reservoirs of petroleum, all grouped on, or related to, a single geological structural and (or) stratigraphic feature.

petroleum A collective term for crude oil, natural gas, natural gas liquids, and tar.

reserves The estimated quantities of petroleum of a specified date, expected to be commercially recovered from known accumulations under prevailing economic conditions, operating practices, and government regulations. Reserves are part of the identified (discovered) resources and include only recoverable quantities.

reserve growth That part of the identified resources over and above measured reserves, estimated to be added to existing fields and reservoirs through delineation of additional in-place oil and natural gas, improved recovery efficiency, and revisions resulting from recalculation of viable reserves under dynamically changing economic and operating conditions.

resource A concentration of naturally occurring solid, liquid, or gaseous hydrocarbons in or on the earth's crust, some of which is, or potentially is, economically extractable.

The United States Geological Survey (USGS) periodically assesses petroleum resources of the world. An assessment of petroleum resources is an evaluation or appraisal of the quantity, quality, and composition of technically or economically recoverable petroleum or of in-place petroleum that provides information on which an estimate of value of the resources on land and in offshore areas can be made. The goal of USGS assessments is to provide impartial, scientifically based, societally relevant petroleum-resource information essential to the economic and strategic security of the United States. Resources must be continuously reassessed in light of new geologic knowledge, progress in science and technology, and shifts in economic and political conditions. Assessments of resources may involve either discovered quantities (reserves) or undiscovered resources. Assessed petroleum resource volumes are estimates, not measurements. In resource-assessment studies, appraisal is synonymous to, and used interchangeably with, assessment. Appreciation refers to reserve growth, the increases in reserves of fields or accumulations through time with continued development and production.

1. INTRODUCTION

Society's dependence on petroleum, as both a fuel and a feedstock for materials used in industry and agriculture (such as plastics, medicine, and fertilizer), requires knowledge of oil and natural gas availability. Availability can be considered in terms of geology and economics. Geologic availability concerns the quantity and quality of petroleum and is the most fundamental characteristic that governs commercial use. Assessment of geologic availability of petroleum requires basic knowledge of the geology.

Societal concepts of the magnitude of petroleum resources commonly are based on economic rather than geologic availability. However, economic availability is evaluated within the frame of petroleum estimated to be geologically available. Several categories of petroleum volumes in terms of geologic availability exist and evaluations of economic availability may be very different depending on which is used. Because major political and economic decisions are based on the economic availability of petroleum, the classification system and terminology used to report estimated volumes of petroleum, therefore, must be precisely defined.

This article provides classifications and terminology used by the USGS for the assessment of petroleum resources. First, an approach to petroleum resource assessments is presented, followed by descriptions of petroleum commodities, spatial geologic units (assessment units), petroleum accumulations, and supporting exploration and discovery data. Assessment concepts used by the USGS are then presented, followed by classifications of estimated petroleum resources. A comparison of resources and reserves definitions is made between the older joint U.S. Bureau of Mines and U.S. Geological Survey (USBM/USGS) classification system and the newer classification system of the Society of Petroleum Engineers and World Petroleum Congresses (SPE/WPC) and Society of Petroleum Engineers, World Petroleum Congresses, and American Association of Petroleum Geologists (SPE/WPC/AAPG), which is representative of the industry sector. Other organizations and agencies in the United States, such as the Potential Gas Committee, the Department of Energy's Energy Information Administration, and the Department of Interior's Minerals Management Service, have similar classifications for resources and reserves, but are not included in this article.

The classifications and terminology used in this article are intended to represent standard usage by the U.S. oil and natural gas industry, the resource-assessment community, and the USGS. The USGS adopted many of the definitions given by the American Petroleum Institute. Nevertheless, the classifications and terms in this article do not represent an exhaustive compilation from all sources and other professionals, agencies, and organizations may define some terms used in this article differently.

2. COMMODITIES CONSIDERED IN OIL AND NATURAL GAS RESOURCE ASSESSMENTS

Petroleum, as used in this article, is a collective term for crude oil, natural gas, natural gas liquids (NGL), and tar. The petroleum commodities considered in USGS assessments are crude oil, natural gas, and NGL, all of which are typically produced from the subsurface through wells. Petroleum commodities produced by other means require different assessment methodologies. Such commodities have not been appraised in recent USGS assessments; they include natural gas dissolved in geopressured brines, resources in tar deposits, oil shales, and gas hydrates.

Crude oil, according to the American Petroleum Institute (API), is defined as a mixture of hydrocarbons that exists in a liquid phase in natural underground reservoirs and remains liquid at atmospheric pressure after passing through surface separation facilities. Although oil is mainly a mixture of complex hydrocarbon compounds, having carbon/hydrogen ratios ranging typically from 6 to 8, it may also contain some nonhydrocarbon components, such as sulfur, nitrogen, oxygen, vanadium, and nickel. Oil has specific gravities ranging typically from 0.76 (55° API gravity) to 1.00 (10° API gravity). API gravity, defined by the API, is a measure of the density of oils.

The volumes reported as crude oil generally include (1) liquids technically defined as crude oil, (2) small amounts of hydrocarbons that exist in the gaseous phase in underground reservoirs but are liquid at atmospheric pressure after being recovered from oil well gas in lease separators, and (3) small amounts of nonhydrocarbons produced with the oil.

The American Petroleum Institute defines natural gas as a mixture of hydrocarbon compounds and small quantities of various nonhydrocarbons existing in the gaseous phase or in solution with oil in natural underground reservoirs under reservoir temperature and pressure conditions and produced as a gas under standard temperature and pressure conditions [60°F or 15°C and 1 atm (14.7 lbs/in.^{2} or 101.325 kPa)].

Natural gas is principally methane, but may contain ethane, propane, butanes, and pentanes, as well as certain nonhydrocarbon gases, such as carbon dioxide, hydrogen sulfide, nitrogen, and helium.

Natural gas can be associated with or dissolved in oil accumulations (associated/dissolved gas or wet gas) or not associated with any liquid hydrocarbons (nonassociated gas or dry gas). Associated gas is free natural gas, commonly known as gas-cap gas, which overlies and is in contact with crude oil in the reservoir. Dissolved gas is natural gas in solution with crude oil in the reservoir under reservoir temperature and pressure conditions. Associated/dissolved gas is commonly reinjected into the reservoir to maintain a pressure drive for oil production and may therefore not be an economically feasible resource in a near-term assessment time frame. Nonassociated gas is free natural gas that is not in contact with crude oil in the reservoir.

Natural gas liquids, as defined by the American Petroleum Institute, are those portions of reservoir gas that are liquefied at the surface in lease separators, field facilities, or gas processing plants. Natural gas liquids are the heavier homologues of methane (including ethane, propane, butane, pentane, natural gasoline, and condensate) and are in a liquid phase under surface conditions [that is, standard temperature and pressure conditions; 60°F or 15°C and 1 atmosphere (14.7 lbs/in.^{2} or 101.325 kPa)]. The specific gravity of NGL is generally less than 0.76 (greater than 55° API gravity).

3. CLASSIFICATION OF OIL AND NATURAL GAS ACCUMULATIONS

Assessed petroleum commodities may be categorized with respect to accumulations as (1) oil in oil accumulations, (2) natural gas in oil accumulations (associated/dissolved natural gas), (3) NGL in oil accumulations, (4) natural gas in gas accumulations (nonassociated natural gas), and (5) liquids (both crude oil and NGL) in gas accumulations.

An accumulation is one or more reservoirs of petroleum that have distinct trap, charge, and reservoir characteristics. For purposes of assessment, an accumulation is treated as a single entity. An accumulation may encompass several fields or be equivalent to a single field or reservoir; alternatively, a field may be equivalent to one or several accumulations. The USGS identifies two major types of petroleum accumulations based on geology: conventional and continuous (Fig. 1). Conventional accumulations are associated with structural or stratigraphic traps, commonly bounded by a down-dip water contact, and therefore affected by the buoyancy of petroleum in water. Continuous accumulations are areally extensive reservoirs of petroleum not necessarily related to conventional structural or stratigraphic traps. These accumulations lack well-defined down-dip petroleum/water contacts and thus are not localized by the buoyancy of oil or natural gas in water. Examples of continuous accumulations include "tight gas reservoirs," coal bed

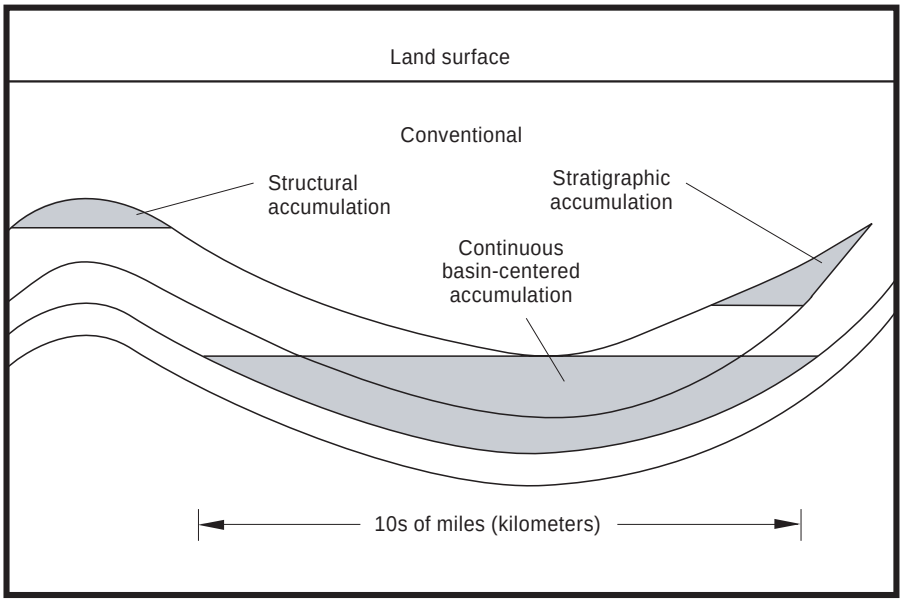


FIGURE 1 Conventional and continuous accumulations. Modified from Schmoker (1996).

gas, oil and gas in shale, oil and gas in chalk, basin-centered gas, and shallow biogenic gas.

A continuous accumulation can be treated as a collection of petroleum-containing cells. The USGS defines a cell as a subdivision or area within a continuous accumulation having dimensions related to the drainage areas of wells (not to be confused with finite-element cells as used in reservoir modeling). Cells differ from wells by having areal properties, whereas wells are considered areally as points and a cell may contain more than one well. Two types of cells are recognized: tested cells (those evaluated by drilling) and untested cells. Tested cells are categorized as productive, having at least one well with reported production, and nonproductive, evaluated with drilling but where none of the wells have reported production. Untested cells have not been evaluated by wells.

The term unconventional has been loosely used interchangeably with continuous when describing these accumulations. Continuous accumulations are defined by geologic criteria, whereas unconventional accumulations are characterized by a variety of standards, such as arbitrary matrix-permeability limits, special regulatory status, the need for unusual (at the time) engineering techniques, or difficulty of physical access (for example, polar regions or deep water).

Oil and gas accumulations, both conventional and continuous, are treated separately in the USGS assessment process. Gas/oil ratios (GOR) are calculated for each accumulation to identify or categorize the major commodity (oil or gas). An oil accumulation is defined as having a GOR less than 20,000 ft^3 of gas/barrel of oil; a gas accumulation is defined as having a GOR equal to or greater than 20,000 ft^3 of gas/barrel of oil. Coproduct ratios (NGL/natural gas ratio; and petroleum liquids/natural gas ratio) are also calculated to aid in the assessment of those associated resources on an accumulation level. Petroleum liquids are undifferentiated oil and NGL.

Accumulations can contain various production elements having spatial characteristics that include fields, reservoirs or pools, and wells.

A field is an area consisting of a single reservoir or multiple reservoirs of petroleum, all grouped on, or related to, a single geological structural and (or) stratigraphic feature. Individual reservoirs in a single field may be separated vertically by impervious strata or laterally by local geologic barriers. When projected to the surface, the reservoir(s) within the field can form an approximately contiguous area that may be circumscribed. Common field classifications include discovered (located by exploratory drilling), under development (undergoing methods or additional drilling to bring into production), producing, and abandoned.

The volume of petroleum produced within a given year is called annual production. The successive sum of annually produced volumes is the cumulative production to that date. Recoverable oil and gas volumes that have not yet been produced are called remaining reserves. Remaining reserves and just "reserves" are used interchangeably; reserves include only recoverable quantities. The recoverable volume is the sum of cumulative production and remaining reserves. The recoverable volumes are typically well known within a specified uncertainty range and provide a measure of the size of a field (field size). These recoverable volumes may also be described as known recoverable volumes (those reported at any given time) and total (ultimate) recoverable volumes.

A reservoir, as defined by the American Petroleum Institute, is a porous and permeable underground formation containing an individual and separate natural accumulation of producible petroleum, which is confined by impermeable rock or water barriers. A reservoir is typically a single hydraulically connected unit with a single natural pressure system. Pool is an older term for reservoir, but is still in use by some. A pool, however, may consist of more than one reservoir in certain situations.

Wells are drilled to find, delineate, and produce petroleum from accumulations. Some wells are used to inject fluids to enhance productivity of other wells in the accumulation. Wells may be classified with respect to drilling goals. Such classifications are called initial classifications and are (1) exploratory wells, which include new-field wildcat, new-pool wildcat, deeper pool test, shallower pool test, outpost or extension tests; (2) development and injector wells; (3) service wells, which include observation wells; and (4) stratigraphic tests. A final classification of status may be assigned to a drilled well, which includes oil well, gas well, multiple completions, junked or abandoned, and dry hole. Additionally, a classification of petroleum content is commonly recorded for wells, which includes oil, gas, oil and gas, dry, oil shows, and gas shows.

The volumes of petroleum from producing wells are reported as annual and cumulative production, as for fields. However, the total volume of petroleum estimated to be recovered from wells is called the estimated ultimate recovery. The estimated ultimate recovery of a well is typically calculated by extrapolation of its past production performance.

4. ASSESSMENT UNITS

In order to organize, evaluate, and delineate areas and rock volumes for assessment, the USGS developed a hierarchical scheme of geographic and geologic units. This scheme consists of regions, geologic provinces, total petroleum systems, and assessment units. Regions serve as organizational units and geologic provinces are used as prioritization tools. Assessment of undiscovered resources is carried out at the level of the total petroleum system or the assessment unit.

A geologic province is an area, having characteristic dimensions of hundreds of kilometers, that encompasses a natural geologic feature (for example, a sedimentary basin, thrust belt, or accreted terrane) or some combination of contiguous geologic features. Geologic province boundaries are delineated along natural geologic boundaries, although in some places they may be located arbitrarily (for example, along specific water-depth contours in the open oceans).

Total petroleum systems and assessment units are delineated within each geologic province considered for assessment. It is not necessary for the boundaries of total petroleum systems and assessment units to be entirely contained within the geologic province.

The USGS identifies "total" petroleum systems for the assessment process, rather than the narrowly defined petroleum systems delineated by known or discovered accumulations or production. The definition of a total petroleum system is similar to that for a petroleum system, except that undiscovered accumulations are included in total petroleum systems. The total petroleum system includes all genetically related petroleum that occurs, or is estimated to occur, in shows and accumulations, both discovered and undiscovered, which have been generated by a pod or by closely related pods of mature source rock. Particular emphasis is placed on the similarities of petroleum fluids within total petroleum systems. Total petroleum systems exist within a limited mappable geologic space, together with the essential mappable geologic elements necessary for the accumulation of petroleum (source, reservoir, seal, and overburden rocks). These essential geologic elements control the fundamental processes of generation, expulsion, migration, entrapment, and preservation of petroleum within the total petroleum system.

In areas or rock volumes where more than one source rock exists and significant mixing of petroleum occurred, identification of individual total petroleum systems may not be possible or practical for the assessment processes. In such areas or rock volumes, a composite total petroleum system may be defined. The concept of composite total petroleum systems is applied when more than one source rock is assumed to charge the accumulations.

Assessment units are identified to organize the assessment process and serve as the basic area or rock-volume entities for resource appraisal. An assessment unit is a mappable part of a total petroleum system in which discovered and undiscovered oil and gas accumulations constitute a single relatively homogenous population such that the methodology of resource assessment is applicable. An assessment unit might be equivalent to a total petroleum system, but total petroleum systems may be subdivided into two or more assessment units such that each assessment unit is sufficiently homogenous in terms of geology, exploration considerations, accessibility, and risk to be assessed individually.

Some USGS assessments identify plays to organize the assessment process and serve as the basic area or rock-volume entities for resource appraisal. A play is a set of known or postulated oil and gas accumulations sharing similar geologic, geographic, and temporal properties, such as source rock, migration pathway, timing, trapping mechanism, reservoir rocks, and hydrocarbon type. A play may differ from an assessment unit in that it can include one or more assessment units. More commonly, an assessment unit includes one or more plays. Other USGS studies use prospects as the basic units for assessment. A prospect is a location where an accumulation is expected to exist and geologic information and economics justify the drilling of a wildcat or test well.

The terms conventional and continuous are used to describe assessment units and plays and are assigned primarily to describe their respective accumulations.

A subdivision of continuous accumulations or continuous assessment units is the "sweet spot." A sweet spot is an area within a continuous accumulation where production characteristics are relatively more favorable than elsewhere.

5. EXPLORATION AND DISCOVERY HISTORY

Petroleum resource assessments are based, partly or wholly, on analysis of the process by which accumulations within a given area are discovered (discovery

process). With conventional accumulations, typically, the largest and shallowest accumulations are found relatively early in the exploration history, followed by the discovery of smaller accumulations. The smaller accumulations found later in the history typically require more time and exploration effort, including the number of exploratory wells drilled for discovery.

Exploration- and discovery-history data are used to model the discovery process, by which the number and sizes of yet to be found (undiscovered) accumulations are estimated. These data are graphed and analyzed statistically. Graphs of accumulation-size distribution and of the number of discovered accumulations with respect to size class portray the sizes of past discoveries. The difference between the discovered size distribution and an assumed distribution for the parent population of accumulations can provide an estimate of the number of undiscovered accumulations for each of the size classes.

Graphs of accumulation sizes with respect to year of discovery or amount of exploration effort provide indications of success and discovery maturity and their extrapolation allows inferences concerning numbers and sizes of remaining (undiscovered) accumulations. Success commonly can be described as one of two ratios: (1) the number of discoveries per number of years of exploration and (2) the number of discoveries per number of exploratory wells drilled. During initial exploration, success ratios may increase with time due to the experience gained by exploration. After initial exploration, the success ratio typically decreases with time as an area becomes more mature with respect to exploration (discovery maturity). However, new exploration-play concepts can offset or reverse a decreasing trend in success ratio.

6. ASSESSMENT CONCEPTS

The USGS incorporates the concepts of forecast span, minimum accumulation size or minimum recovery per cell, and probability of occurrence, or risk, in the assessment process.

In making assessments, the USGS considers only those quantities of oil and gas resources that have the potential to be added to reserves within a specified forecast span. Petroleum resources can be ranked to provide a visual representation of practicability of extraction in the form of a triangle or pyramid (a concept of J. K. Gray). A relatively small volume of high-quality resources is present near the top, with increasing volumes of progressively lower quality and less practicable resources advancing toward the base (Fig. 2). For an assessment, a slice can be drawn through the pyramid at some quality level, which defines the resource of oil or gas that is estimated to become economic within the foreseeable future. The position of the slice can be regarded as the forecast span, whereby resources below the slice are not assessed. The slice may move lower into the pyramid on successive assessments, as originally less practicable and poorer quality resources are deemed potentially economic over time.

Oil and natural gas from undiscovered conventional accumulations and the development of cells within continuous accumulations must exceed or equal a selected minimum recoverable volume. A minimum recoverable volume is selected to avoid inclusion of large numbers of volumetrically insignificant accumulations or cells. In the time frame of the assessment, volumes of oil and natural gas from very small accumulations or cells are unlikely to contribute significantly to the total resources.

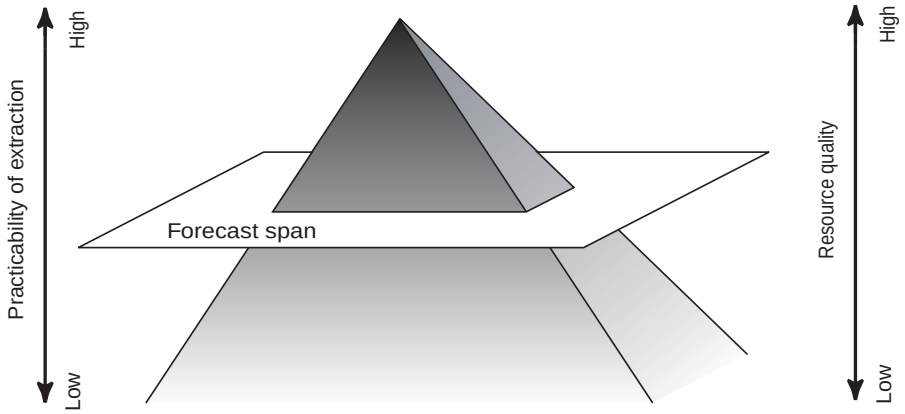


FIGURE 2 Pyramid of undiscovered petroleum resources.

Risking elements are used by the USGS to address the question of the probability of occurrence of at least one undiscovered accumulation of minimum volume or greater, or cell of minimum total recovery (or estimated ultimate recovery) or greater, somewhere in the area or rock volume being assessed that has the potential to be added to reserves in the time frame of the assessment. Probabilities for the occurrence of adequate charge, adequate rocks (reservoirs, seals, traps, etc.), and adequate timing of geologic events (burial, maturation, migration, trap formation, uplift, etc.) define geologic risk. The probability that essential petroleum exploration and production activities will be possible in the specified forecast span of the assessment, at least somewhere in the area or rock volume, is also assigned, defining access risk. Each element in this risking structure applies to the area or rock volume as a whole and is not equivalent to the percentage of the area or rock volume that might be unfavorable in terms of geology or access.

Estimates of oil and natural gas resources are termed conditional if oil and natural gas are assumed to be present in the subject area or rock volume to be assessed. Conditional estimates do not incorporate the risk that the area or rock volume may be devoid of oil or natural gas. If a risk that oil and natural gas may not be present is incorporated in the resource estimates, then these are termed unconditional or risked estimates. Statistically, the unconditional (risked) mean may be determined through multiplication of the mean of a conditional (unrisked) distribution by the related probability of occurrence.

7. PARAMETERS FOR CLASSIFICATION OF OIL AND NATURAL GAS RESOURCES

A resource is a concentration of naturally occurring solid, liquid, or gaseous hydrocarbons in or on the earth's crust, which is, or potentially is, economically extractable. Resources may be classified in terms of economic recoverability, geologic certainty, and uncertainty of the resource quantity.

The 1980 USBM/USGS classification system is based on two parameters whereby resources are classified by feasibility of economic recovery and degree of geologic certainty or assurance (originally called degree of uncertainty of existence) (Fig. 3). The SPE/WPC/AAPG have proposed a new system whereby petroleum resources may be classified by three parameters: feasibility of economic recovery, degree of geologic assurance, and degree of uncertainty of estimated resource quantities (Fig. 4). Although some differences exist between the two systems, the classification schemes are comparable.

Resources include reserves and all other petroleum accumulations that may eventually become available—including known accumulations that are not recoverable under current economic conditions or technology, or unknown accumulations of varying degrees of richness that may be inferred to exist, but not yet discovered. Therefore, resources can be classified in terms of geologic assurance as discovered (identified) and undiscovered.

According to the USBM/USGS, identified resources are those whose location, grade (for mineral

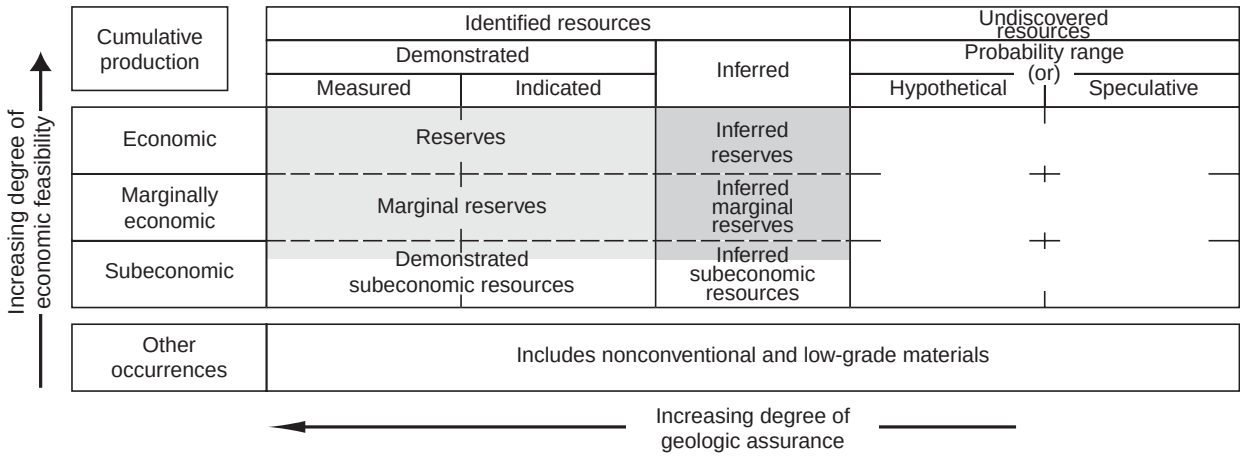


FIGURE 3 Diagram showing the resource classification system of the USBM/USGS. The terminology used in this classification system is from McKelvey (1972, 1973), but modified by Brobst and Pratt (1973) and adopted with minor changes for joint use by the USBM/USGS (1976, 1980).

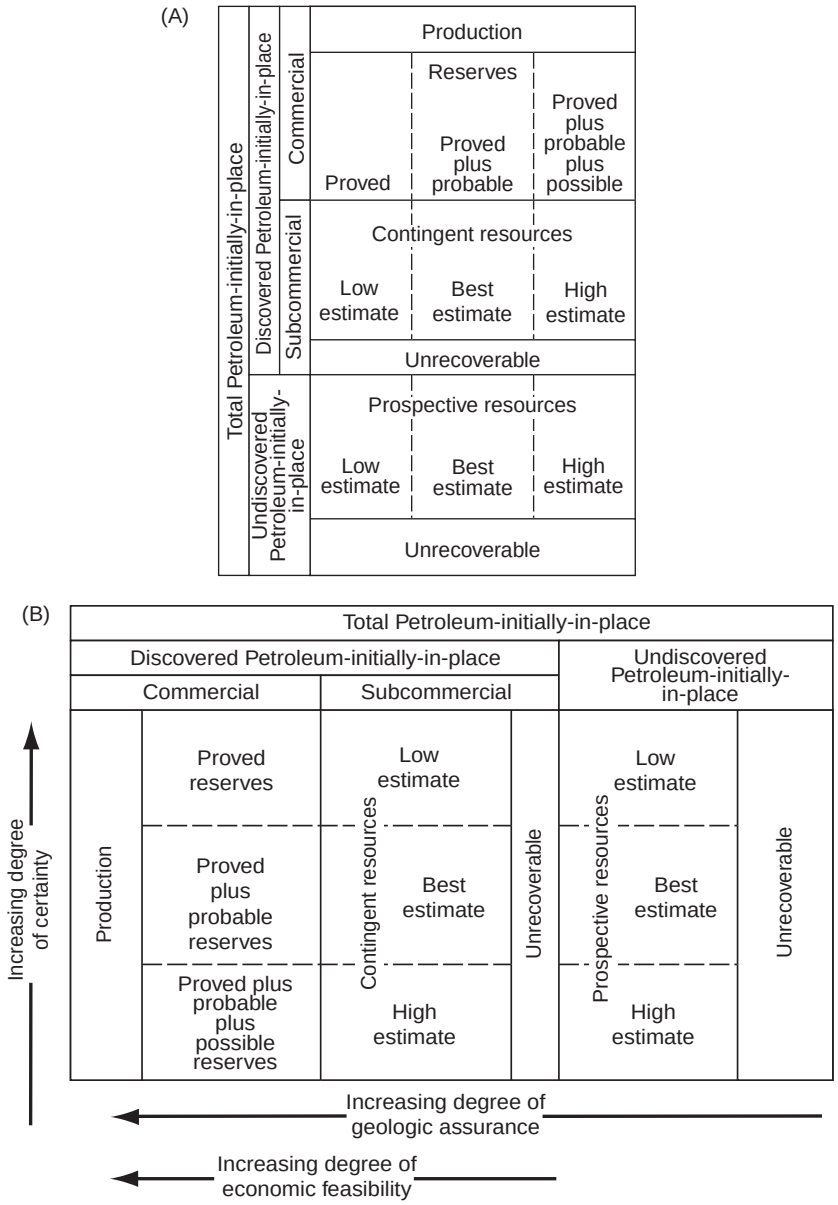


FIGURE 4 (A) Diagram showing the resource classification system of the SPE/WPC/AAPG. Reproduced from the Society of Petroleum Engineers, World Petroleum Congresses, and American Association of Petroleum Geologists (2001), "Guidelines for the Evaluation of Petroleum Reserves and Resources," with permission. Copyright Society of Petroleum Engineers. (B) Same diagram as in A, but slightly modified for comparison with Fig. 3.

resources), quality, and quantity are known or estimated from specific geologic evidence. Undiscovered resources are those that are inferred from geologic information and theory and comprise accumulations that are separate from identified resources [that is, existing outside of known oil and (or) natural gas accumulations]. Similarly, the SPE/WPC/AAPG define identified resources as discovered petroleum-initially-in-place. This is the quantity of petroleum that is estimated, on a given date, to be contained in known accumulations, plus those quantities already produced therefrom. In addition, undiscovered petroleum-initially-in-place is defined as that quantity of petroleum that is estimated, on a given date, to be contained in accumulations yet to be discovered.

Resources are also classified in terms of feasibility of economic recovery as economic or commercial and subeconomic or subcommercial (Fig. 3). A third classification, marginally economic, had been used

by the USGS. The USBM/USGS classified both discovered (identified) and undiscovered resources in terms of economic recoverability. In contrast, the SPE/WPC/AAPG classify only discovered resources in terms of economic recoverability (Fig. 4).

A degree of uncertainty is typically reported for the estimated quantities of discovered and undiscovered resources that are potentially recoverable. The uncertainty of estimated resource quantities may be expressed probabilistically, either as a range or single-value statistic such as a mean, mode, median (P_{50}), or some other percentile.

8. CLASSIFICATION OF OIL AND NATURAL GAS RESOURCES

Discovered (identified) and undiscovered resources are incorporated in the classification systems of both the USBM/USGS and the SPE/WPC/AAPG. The USBM/USGS divide discovered (identified) resources in terms of geologic assurance into measured, indicated, and inferred classes (Fig. 3). Demonstrated resources are the sum of measured and indicated resources.

The estimated economically recoverable portion of discovered (identified) resources is classified as reserves. The part of the discovered (identified) resources from which reserves are estimated is called the reserve base (Fig. 3, shaded and striped areas). Discovered resources, in the SPE/WPC/AAPG classification, that are subeconomic or subcommercial are called contingent resources. This classification system also includes all produced petroleum and identified unrecoverable petroleum as discovered resources. The USBM/USGS system does not include the amount of past production as part of the existing identified resource. However, the produced volume of petroleum may be used to better understand the amount of resources.

Reserves are that part of the identified (discovered) resources that could be economically extracted or produced at the time of determination and include only recoverable quantities. Similarly, the SPE/WPC and SPE/WPC/AAPG define reserves as the estimated quantities of petroleum expected to be commercially recovered from known accumulations relative to a specified date. The American Petroleum Institute adds that reserves are anticipated to be commercially recoverable under prevailing economic conditions, operating practices, and government regulations. Uncertainty of reserve estimates depends on the amount of geologic and engineering data and interpretations available at the time of the estimate. Based on the level of uncertainty, reserves are defined in classes of (1) proved (most certain), (2) probable, and (3) possible. Although the measured, indicated, and inferred classes of reserves are assigned to reflect geologic assurance, these classes have been loosely interchanged with, respectively, the proved, probable, and possible classes.

Proved petroleum reserves are the estimated quantities of crude oil, natural gas, and NGL, which geological and engineering data demonstrate with reasonable certainty to be recoverable in future years from known reservoirs under existing economic and operating conditions, that is, prices and costs prevailing at the time the estimate is made. The SPE/WPC and SPE/WPC/AAPG define proved reserves similarly, but with the addition that the estimate is made from a given date forward and under current economic conditions, operating methods, and government regulations.

Reserves are considered proved if economic producibility of the reservoir is supported by either actual production or conclusive formation tests. (See Fig. 5.) According to the SPE/WPC and SPE/WPC/AAPG, there should be at least a 90% probability that the quantities of oil or natural gas actually recovered will equal or exceed the estimate. Proved reserves (informally called 1P reserves) are typically reported for fields in the United States.

Proved reserves can be categorized as (1) developed—those expected to be recovered from existing wells; or (2) undeveloped—those expected to be recovered from new wells, deeper wells, well recompletions, or on installation of production or transportation facilities.

Probable reserves are those unproved reserves that are more likely than not to be recovered. Probable reserves should have at least a 50% probability that the quantities actually recovered will equal or exceed the sum of estimated proved plus probable reserves (Fig. 4). Volumes of proved plus probable reserves are greater than those of proved reserves alone. Proved plus probable reserves (informally called 2P reserves) are typically reported for fields in countries other than the United States.

Possible reserves are those unproved reserves that have at least a 10% probability that the quantities actually recovered will equal or exceed the sum of estimated proved plus probable plus possible reserves. Proved plus probable plus possible reserves (informally called 3P reserves) are not used in USGS assessments.

In contrast to reserves, contingent resources are estimated to be potentially recoverable from known

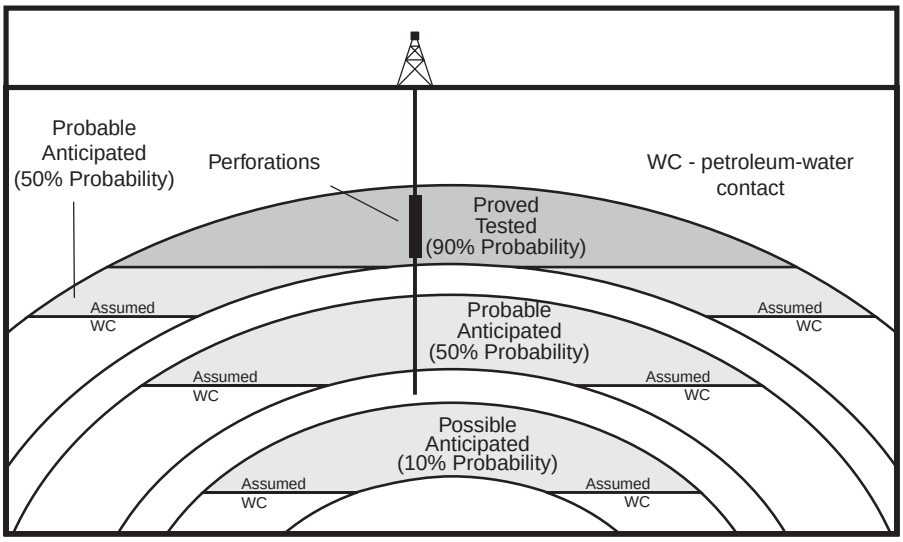


FIGURE 5 Proved, probable, and possible reserves.

(discovered) accumulations, but are not considered to be commercially recoverable.

Successive estimates of the total crude oil, natural gas, and natural gas liquids to be recovered in fields and reservoirs generally increase through time with continued development and production, the result commonly being additions of reserves. These additions are directly related to increases in the total size (cumulative production plus remaining reserves) of the field or reservoir. Reserve growth (also called field growth, reserve appreciation, and ultimate recovery appreciation) is therefore that part of the identified resources over and above measured reserves, estimated to be added to existing fields and reservoirs. Reserve growth occurs for a variety of geologic, engineering, operational, and economic reasons, including the following: (1) delineation of additional in-place oil and natural gas, including addition of new reservoirs and extensions; (2) improved recovery efficiency; and (3) revisions resulting from recalculation of viable reserves under dynamically changing economic and operating conditions.

The USBM/USGS divided undiscovered resources into hypothetical and speculative classes to reflect geologic assurance (Fig. 3). Hypothetical resources are undiscovered resources that may be reasonably expected to exist under geologic conditions analogous to those in known producing districts or regions. Speculative resources are undiscovered resources that may exist elsewhere, in districts or regions with no discovered resources. In the SPE/WPC/AAPG classification, undiscovered resources are divided into prospective resources, which are considered to be potentially recoverable, and resources that are unrecoverable (Fig. 4).

The USGS assesses the quantities of oil, gas, and NGL that are technically recoverable and have the potential to be added to reserves within some specified future time frame. The USGS defines technically recoverable resources as resources in accumulations producible using current recovery technology and in 1995 the U.S. Geological Survey National Oil and Gas Resource Assessment Team reported that "these are oil and natural gas resources that may be produced at the surface from a well as a consequence of natural pressure within the subsurface reservoir, artificial lifting of oil from the reservoir to the surface, and the maintenance of reservoir pressure by fluid injection. These resources are generally conceived as existing in accumulations of sufficient size to be amenable to the application of existing recovery technology."

USGS assessments consider three types of technically recoverable resources: (1) undiscovered conventional accumulations of oil and natural gas; (2) additions of oil and natural gas from untested cells within continuous accumulations; and (3) the potential future additions to reserves of known conventional accumulations by reserve growth. Each of these resource types requires a different technique for evaluation and assessment.

The oil and natural gas from undiscovered conventional accumulations are clearly equivalent to the undiscovered resource classification, whereas oil and natural gas from development of continuous accumulations may be equivalent to both discovered

and undiscovered classes (Figs. 3 and 4). Oil and natural gas from field or reserve growth would include inferred resources using the USBM/USGS classification (Fig. 3) or possible reserves and contingent resources using the SPE/WPC/AAPG classification (Fig. 4).

Although based on the best geologic and historical information and theory available, assessed petroleum volumes are unknown quantities, not measurements, and therefore should be expressed with probability distributions of uncertainty.

Acknowledgments

The definitions of the terms crude oil, natural gas, natural gas liquids, field, reservoir, reserves, and proved reserves are from "Standard Definitions of Petroleum Statistics" (1995), 5th ed. Reproduced courtesy of the American Petroleum Institute.

SEE ALSO THE FOLLOWING ARTICLES

Biomass Resource Assessment • Markets for Natural Gas • Markets for Petroleum • Natural Gas Resources, Global Distribution of • Oil and Natural Gas: Economics of Exploration • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Oil and Natural Gas Resource Assessment: Geological Methods • Oil and Natural Gas Resource Assessment: Production Growth Cycle Models • Petroleum Property Valuation

Further Reading

Ahlbrandt, T. S., and McCabe, P. J. (2002). Global petroleum resources: A view to the future. *Geotimes* **47,** 14–18.

American Petroleum Institute. (1995). "Standard Definitions for Petroleum Statistics." 5th ed. American Petroleum Institute, Washington, DC.

Attanasi, E. D. (1998). "Economics and the 1995 Assessment of United States Oil and Gas Resources." U.S. Geological Survey Circular 1145. U.S. Geological Survey, Washington, DC.

Brobst, D. A., and Pratt, W. P. (1973). Introduction. *In* "United States Mineral Resources" (D. A. Brobst and W. P. Pratt, Eds.), pp. 1–8. U.S. Geological Survey Professional Paper 820. U.S. Geological Survey, Washington, DC.

Dolton, G. L., Carlson, K. H., Charpentier, R. R., Coury, A. B., Crovelli, R. A., Frezon, S. E., Khan, A. S., Lister, J. H., McMullin, R. H., Pike, R. S., Powers, R. B., Scott, E. W., and Varnes, K. L. (1981). "Estimates of undiscovered recoverable conventional resources of oil and gas in the United States." U.S. Geological Survey Circular 860. U.S. Geological Survey, Denver, CO.

Gautier, D. L., Dolton, G. L., Takahashi, K. I., and Varnes, K. L. (eds.). (1996). "1995 National Assessment of United States Oil and Gas Resources—Results, Methodology, and Supporting Data," U.S. Geological Survey Digital Data Series DDS-30, Release 2 (CD-ROM).

Lahee, F. H. (1944). Classification of exploratory drilling and statistics for 1943. *Am. Assoc. Pet. Geol. Bull.* **28,** 701–721.

Levorsen, A. I. (1967). "Geology of Petroleum." W. H. Freeman, San Francisco, CA.

Magoon, L. B., and Dow, W. G. (eds.). (1994). "The Total Petroleum System—From Source to Trap." American Association of Petroleum Geologists Memoir 60. American Association of Petroleum Geologists, Tulsa, OK.

Magoon, L. B., and Schmoker, J. W. (2000). The total petroleum system—The natural fluid network that constrains the assessment unit. *In* "U.S. Geological Survey World Petroleum Assessment 2000—Description and Results (U.S. Geological Survey World Energy Assessment Team, Eds.), U.S. Geological Survey Digital Data Series DDS-60 (CD-ROM).

Masters, J. A. (1979). Deep basin gas trap, western Canada. *Am. Assoc. Pet. Geol. Bull.* **63,** 152–181.

McCabe, P. J. (1998). Energy resources—Cornucopia or empty barrel? *Am. Assoc. Pet. Geol. Bull.* **82,** 2110–2134.

McKelvey, V. E. (1972). Mineral resource estimates and public policy. *Am. Scientist* **60,** 32–40.

McKelvey, V. E. (1973). Mineral resource estimates and public policy. *In* "United States Mineral Resources" (D. A. Brobst and W. P. Pratt, Eds.), pp. 9–19. U.S. Geological Survey Professional Paper 820. U.S. Geological Survey, Washington, DC.

Schmoker, J. W. (1996). Method for assessing continuous-type (unconventional) hydrocarbon accumulations. *In* "1995 National Assessment of United States Oil and Gas Resources—Results, Methodology, and Supporting Data" (D. L. Gautier, G. L., Dolton, K. I. Takahashi, and K. L. Varnes, Eds.), U.S. Geological Survey Digital Data Series DDS-30, Release 2 (CD-ROM).

Schmoker, J. W. (1999). "U.S. Geological Survey Assessment Model for Continuous (Unconventional) Oil and Gas Accumulations—The 'FORSPAN' Model." U.S. Geological Survey Bulletin 2168. U.S. Geological Survey, Washington, DC.

Society of Petroleum Engineers and World Petroleum Congresses. (1997). "Petroleum Reserves Definitions." Society of Petroleum Engineers, Richardson, TX. Available on the World Wide Web at www.spe.org.

Society of Petroleum Engineers, World Petroleum Congresses, and American Association of Petroleum Geologists. (2001). "Guidelines for the Evaluation of Petroleum Reserves and Resources." Society of Petroleum Engineers, Richardson, TX. Available on the World Wide Web at www.spe.org.

U.S. Bureau of Mines and U.S. Geological Survey. (1976). "Principles of the Mineral Resource Classification System of the U.S. Bureau of Mines and U.S. Geological Survey." U.S. Geological Survey Bulletin 1450-A. U.S. Geological Survey, Washington, DC.

U.S. Bureau of Mines and U.S. Geological Survey. (1980). "Principles of a Reserve/Resource Classification for Minerals." U.S. Geological Survey Circular 831. U.S. Geological Survey, Washington, DC.

U.S. Geological Survey National Oil and Gas Resource Assessment Team. (1995). "1995 National Assessment of United States Oil and Gas Resources." U.S. Geological Survey Circular 1118. U.S. Geological Survey, Washington, DC.

U.S. Securities, Exchange Commission. (2002). Regulation S-X. U.S. Securities and Exchange Commission SEC 1887 (08-2002).

Oil and Natural Gas Resource Assessment: Geological Methods

DONALD L. GAUTIER
U.S. Geological Survey
Menlo Park, California, United States

1. Introduction
2. Assessment Methodology

Glossary

play A volume of rock having common geologic attributes with respect to oil and gas accumulations.
reserves That part of a resource that can be produced under existing conditions of technology and economics.
resources The natural occurrence of potentially useful hydrocarbon compounds.
trap The geometric arrangement of porous and permeable reservoir rocks and permeability barriers to hydrocarbon migration that results in an oil and/or gas accumulation.

1. INTRODUCTION

Since the beginning of the petroleum age, the utility of oil and gas has inflamed the desire to understand and quantitatively predict the resources yet to be found. Alongside the search for petroleum, the sciences of geology, geophysics, and organic geochemistry have grown to maturity in the effort to predict petroleum ahead of the drill. Consequently, the properties of petroleum are understood in great detail. The geological requirements for large oil and gas accumulations are quite specific and, as a result, resources are highly uneven in their distribution. Of the hundreds of sedimentary basins in the world, 5 of them contain almost one-half of the known petroleum and fewer than 50 basins contain well over 90% of the known oil and gas (Fig. 1).

Geological methods of resource assessment are based on the simple understanding that petroleum is a natural component of the earth and that its quantity, quality, and distribution are the result of geological processes. Geological methods of resource assessment attempt to use the scientific underpinnings of geology and the empirical results of the search for petroleum to predict the oil and gas accumulations yet to be found.

1.1 Geological Controls on the Distribution of Oil and Gas Resources

Nearly all petroleum geologists agree that oil is formed from the thermo-chemical transformation of sedimentary organic matter. John Hunt states this view succinctly: "The overwhelming geochemical and geological evidence from both sediment and petroleum studies of the past few decades clearly shows that most petroleum originated from organic matter buried with the sediments in sedimentary basins."

Knowing the organic origin of oil and gas has led to extensive investigations of the properties of

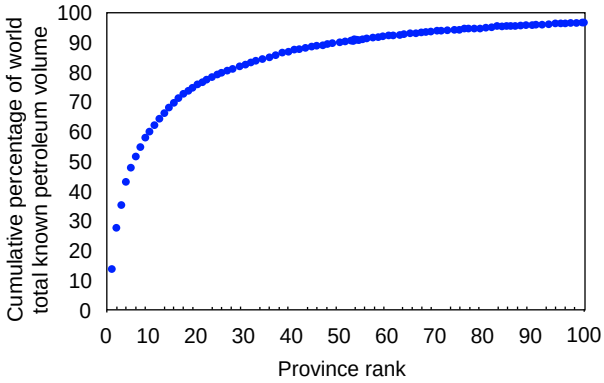


FIGURE 1 Cumulative percentage of world total known petroleum volume by top 100 ranked geologic provinces. Reprinted from U.S. Geological Survey World Assessment Team (2000).

carbon compounds in sedimentary sequences and to the concept of organic sedimentary facies. Recognition that areas of highly successful petroleum exploration (so-called fairways) are geographically related to places where particular types of organic-matter-rich sediments have been buried and heated (petroleum kitchens) has resulted in increased success in explorations.

Although an adequate concentration of organic matter is necessary, that alone is not sufficient for the existence of exploitable hydrocarbon accumulations. Oil generation occurs only from sedimentary rocks with certain types and concentrations of sedimentary organic matter under particular conditions of temperature and burial (the petroleum source rock). When the source rock attains the temperature required, hydrocarbon fluids and gases are squeezed out by compaction and expelled into adjacent permeable beds, such as porous limestone or sandstone.

Oil and gas have much lower density than the water that ordinarily fills the pore networks of sedimentary rocks. As a result, hydrocarbons that have been generated and expelled from the heated organic matter migrate upward under the effects of buoyancy. Most hydrocarbons generated from sedimentary organic matter probably migrate to the surface of the earth, where they form natural oil or gas seeps and are lost to erosion and oxidation. Less commonly, the geologic environment through which the hydrocarbons are moving forms a barrier to migration. A permeability barrier that slows or halts petroleum migration is called a seal and the porous rocks that contain petroleum are reservoirs. Taken together, the geometric relationship of reservoirs and seals is the trap. Structural traps, stratigraphic traps, hydrodynamic traps, and diagenetic traps are four well-known types of barriers to migration. A trap is the geometric arrangement of porous and permeable reservoir rocks and permeability barriers to hydrocarbon migration that results in an oil and/or gas accumulation.

Unless unusual conditions exist, oil, gas, and water are vertically arrayed in the trap in order of their buoyancy, with oil overlying water (the oil–water contact) and gas overlying oil (the gas–oil contact). The pay zone (the rocks from which oil and gas can be produced) may extend from the lowest part of the oil–water contact to the top of the gas accumulation. The volume of oil and gas in the trap (the size of the accumulation) depends on the availability of hydrocarbons, the properties of the reservoir, the dimensions of the trap, and the permeability of the seal that slows further migration. Giant oil fields are found in close proximity to high-quality source rocks, where large geologic structures have an effective seal and where thick reservoirs with high porosity are present. Petroleum geologists and geophysicists are devoted to the search for such accumulations.

1.2 Resources to Reserves—The Intersection of Geology, Technology, and Economics

Despite the fundamental importance of understanding the geological distribution of oil and gas resources, a great deal of additional knowledge and effort is necessary for successful (profitable) production. Many petroleum geologists can cite interesting but frustrating anecdotes of "geological success" in which months or years of research and expensive exploratory drilling resulted in the discovery of hydrocarbons as predicted by geological considerations, but where technological requirements or market forces made production impossible. The petroleum accumulation may be in water too deep, or it may consist of hydrocarbons that are too viscous to flow to the wellbore, or numerous other possibilities may render the accumulation impossible to produce profitably.

For the purposes of resource analysis, a determination of the geological distribution of hydrocarbon molecules in the earth's crust is insufficient. An attempt must be made to estimate that part of the resource that can be actually utilized. In a geologically based resource assessment, it is essential to distinguish between resources and reserves:

Resources are the natural occurrence of potentially useful hydrocarbon compounds.

Reserves are that part of a resource that can be produced under existing conditions of technology and economics

Resources are geological, but reserves are that subset of the resource that exists at the intersection of geology, available technology, and favorable economic conditions. In order to convert resources into reserves, three conditions must be met. First, the hydrocarbons must be present in geologically defined and identifiable accumulations. Second, technology must be available through which such accumulations can be exploited. Third, the economic situation of the accumulation must be such that the application of technology can result in profitable sale of the produced commodity. The principal goal of

geologically based resource assessments is the prediction of the sizes and numbers of such potentially developable accumulations in advance of their discovery.

A well-known example of the difference between resources and reserves comes from the Green River Basin (GRB) of Wyoming. Geological investigations of the GRB suggest that enormous quantities of natural gas may be present in thick sedimentary sequences containing gas-prone coals and carbonaceous mudstones (the source rocks) and low-permeability, nonmarine sandstones (the reservoirs). The total volume of natural gas resources in the GRB has been estimated to exceed 5000 trillion ft^3 (TCF). This volume of natural gas is roughly equivalent to the total proved natural gas reserves of the world! However, initial resource assessments suggested that less than 120 TCF of this huge resource is recoverable using available technology, and of the technically recoverable resources, only a few TCF can be produced economically. In this example, the volumetric difference between resources and reserves ranges over three orders of magnitude.

In as much as hydrocarbon accumulations have been successfully explored and developed, a record of the conversion of geological resources to exploitable reserves exists in the form of drilling records, production histories, and reported field sizes. In combination with geological interpretation, analysis of exploration and production data provides a basis for prediction. Geological insight tempered with exploration and production experience facilitates and underlies all geological approaches to resource assessment.

2. ASSESSMENT METHODOLOGY

A wide variety of geologically based methodologies have been employed over the years to estimate quantities of undiscovered accumulations. In each case, geological information in some combination with exploration history (as opposed to direct extrapolation of drilling statistics, well production profiles, or economic projections) has been used to estimate future additions to oil and gas reserves. In all geologically based assessment methods, drilling and production data are viewed in an explicitly geological context. Geological understanding is necessary for resource prediction.

Geologically based methodological approaches can be grouped into a few basic themes:

1. Direct assessment and geological consensus techniques;
2. Trap-oriented assessments, including prospect appraisal, deposit simulation, and play analysis; and
3. Source rock/fluid geochemistry-oriented approaches, including petroleum system and material balance methods.

Each of these themes is typically used to some degree in the geologically based approaches and they are usually used in combination with one or more other techniques. In addition, geological information is more or less necessary to many numerically based, historical performance and areal or volumetric yield methods of assessment.

2.1 Direct Assessment and Geological Consensus Methods

Direct assessment is a subjective approach, in which knowledgeable individuals (experts) directly provide their views of the resource potential of a specified place or subject. Typically, a group of experts is assembled and these experts independently or jointly review the geological information available for a region, basin, prospect, or other subject of interest to the assessment. This review may include previous assessments, including nongeological assessments. Each expert arrives at an individual opinion or position concerning the resource potential. This opinion may be expressed as a simple point estimate or as a range of possibilities. Individually determined estimates can be statistically combined into a final probability density function that captures the full range of opinions held by the group.

In the geological consensus approach, which is a variant of the direct assessment method, the assembled group may review the individual estimates and make revisions where necessary. Depending on the rules established for the assessment, each member of the group may be required to agree to reach a consensus value, which is often achieved only after vigorous debate. Alternatively, final results may reflect the average of the group's opinions or they may agree that outlying high and low estimates will be omitted from the final results.

This direct, subjective approach is sometimes, perhaps derisively, called the Delphi method, after the famous Greek oracle located near the ancient city of the same name. The title rightly emphasizes the highly subjective nature of this approach, within which little or no documentation is required and

precisely reproducible results may be impossible to achieve. Direct assessment by experts is thus only as good as the expertise of the individuals involved. As noted by White and Gehman, "one must know how expert are the experts in order to assess the assessment."

However, when the experts are indeed expert, the method readily captures their views. Direct assessment has the great advantage, when compared to other methodologies, of being applicable to any and all levels of data availability. An expert group can equally well assess a frontier province with little or no data or a highly developed and explored play, where thousands of wells and fields have been documented. The direct assessment method also has the advantage of being a rapid, relatively simple (and therefore inexpensive) procedure, one that is particularly effective for capturing the knowledge of highly divergent positions or organizations. The direct assessment approach has been used to great effect, for example, by the National Petroleum Council and by the Department of Energy Oil Resources Panel, when it was desired to represent the views of a wide range of experts from diverse organizations to form a group position on various issues.

2.2 Trap-Oriented Assessment

In this approach, which may include prospect appraisal, deposit simulation, and play analysis, specialists focus on anticipated geological properties of individual accumulations or collections of accumulations. Resource estimates are derived directly from geological concepts and principles, commonly by means of standard equations employed by petroleum engineers to evaluate discovered accumulations. This approach is commonly data- and labor-intensive, relying on seismic geophysical data to resolve individual prospects to be recommended for drilling. Alternatively, a more regionally based perspective may be used to predict the undiscovered resources in collections of prospects. In any case, the goal is to integrate the best available earth-science information and then to derive volumetric estimates from that information. In the play or prospect approach, the conditions of oil and gas entrapment are emphasized. In particular, the properties of potential reservoir rocks, the timing of formation of structural features that might serve as traps, and the efficiency of sealing lithologies are regarded as crucial to predicting volumes of undiscovered resources.

Many resource assessments have been based on the play and prospect concept. Although other interpretations exist, most resource analysts agree that a play is a volume of rock having common geologic attributes with respect to oil and gas accumulations. Common attributes usually include reservoir rock, trapping mechanism and geometry, and seal properties, as well as maturation and migration timing and geographic extent. Common, specific source rocks for hydrocarbon generation may or may not be essential to play definition.

Play analysis is the systematic attempt to estimate undiscovered resources within a well-defined play. Although in practice play-based assessments vary considerably in approach and results, most share the goal that variations in results should be the consequence of availability and quality of geological information rather than of the methodology.

Play analysis is derived directly from the exploration methods used by petroleum companies, but as an assessment technique, play analysis originated with methodology developed by the Geological Survey of Canada in the early 1970s. The approach has been applied in many guises since.

Many common features of play analysis can be illustrated by an example from a U.S. Geological Survey study of Arctic National Wildlife Refuge (ANWR) Area 1002 in Alaska.

For the ANWR assessment, plays were defined based on geological conditions; in most cases, the stratigraphic relationships of reservoir rocks and postulated trapping mechanisms were sufficient to distinguish among plays. For each play, the number of prospects (potential oil or gas accumulations) was estimated from interpretation of seismic data and geologic inference. The uncertainty in the number of prospects was evaluated, as was the range of values for each of the various geologic characteristics in the analysis. The link between the geological observations in related wells and rock outcrops and the geophysical measurements was provided by standard equations developed by the petroleum engineers working on the project.

Using Monte Carlo simulation (a commonly used numerical probabilistic technique), the numbers of prospects and their characteristics were combined with the range of (simulated) oil and gas accumulations. This simulation leads to the name "deposit simulation approach." In the ANWR assessment, the geologic prospect characteristics were as follows: area of closure (the map view of the boundaries of the trap), reservoir thickness, reservoir porosity, water saturation, trap fill, and the distribution of

oil versus gas, along with some expected properties of trapped hydrocarbon fluids (such as sulfur content or oil density).

The initial step, as outlined in Fig. 2, was to define plays. In play definition, geologists specify the geologic properties of the play in terms of reservoirs, seals, and traps. Following play definition, the number and size of potential petroleum accumulations were estimated for each play as a probability distribution, based on the range of expected values for geologic characteristics such as reservoir thickness or porosity. These distributions were limited by a chosen minimum field size (the size of the smallest potentially viable accumulation in the play). The resulting distributions were then subjected to a risking procedure designed to weigh the likelihood that geologic conditions were favorable to generate at least one accumulation larger than the minimum field size. Geologic risk was evaluated at both the play and the prospect levels and the resulting distributions and risks were combined to make an estimate of in-place petroleum resources for each play. Next, an estimated recovery factor (the fraction of oil in the trap that can actually be extracted by specified types of technology) is applied to the in-place petroleum resources of each play to calculate the technically recoverable petroleum.

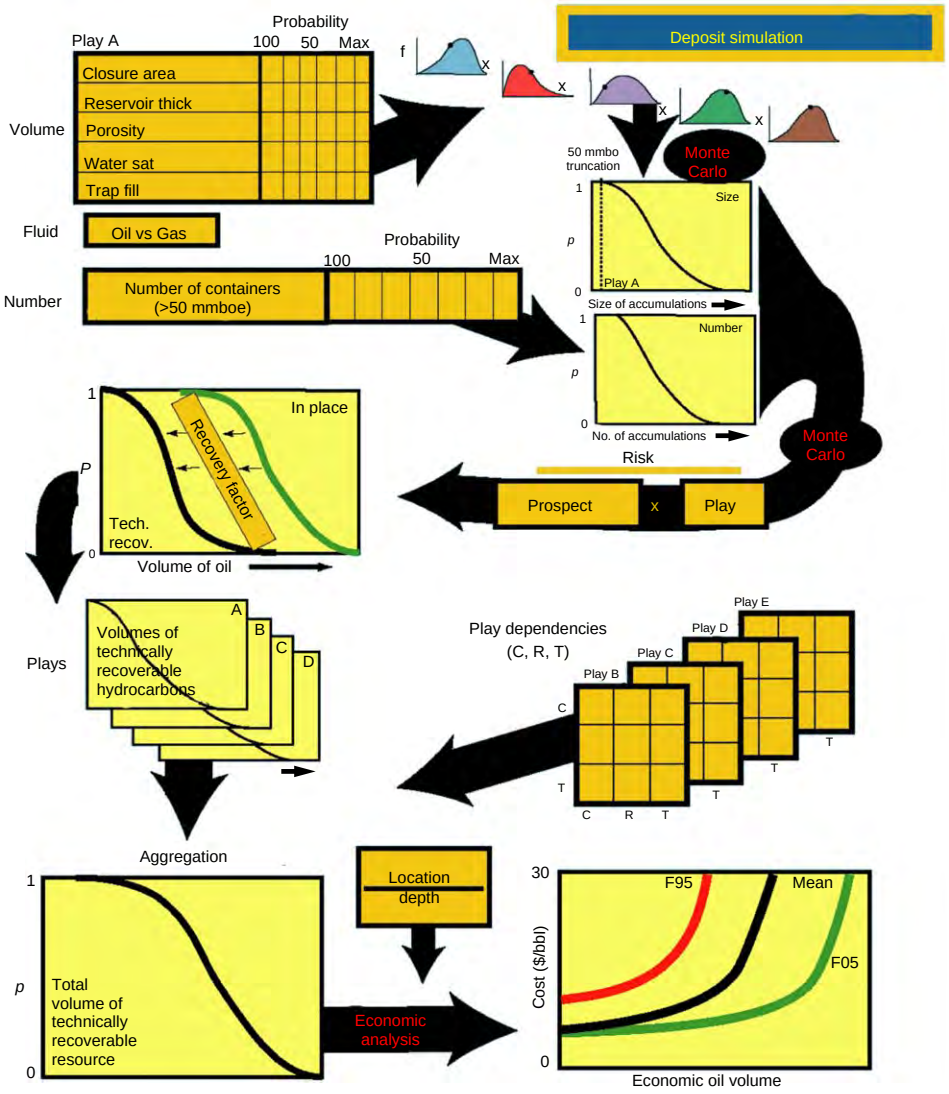


FIGURE 2 Flow chart showing the steps followed in the U.S. Geological Survey assessment of the Arctic National Wildlife Refuge 1002 Area. Reprinted from Bird (1999).

Estimates for each play are aggregated to determine the total resources of the area of assessment. Such aggregates are sensitive to geologic dependencies between the plays. For example, if two plays share the same source rock, inadequate charge in one play might indicate inadequate charge in all plays sharing that source rock. Such dependencies need to be considered and accounted for in the prospect simulation.

This sort of play and prospect simulation approach can be applied to frontier areas with little or no drilling data (as is the case with ANWR) or can be applied to areas of extensive exploration, where well data serve to reduce the uncertainty associated with the numbers and properties of known and undiscovered accumulations.

2.3 Source Rock Geochemistry and Material Balance Methods

Source rock geochemical and material balance methods of resource assessment are an attempt to quantitatively estimate volumes and distributions of hydrocarbons based on physical and chemical principles. In contrast to deposit simulation or play analysis, which emphasize traps and reservoirs, the source rock geochemical/material balance approach focuses first on the properties and processes of the petroleum source rocks and the related fluid system.

Although organic geochemistry has been used in the analysis of petroleum and petroleum products since the earliest days of the industry, systematic applications of geochemical methods to resource evaluation have come about only since the development of laboratory equipment suitable for the routine analysis of kerogens (specific organic components of sedimentary rocks, some of which yield hydrocarbons during thermal maturation) and oils. Such work began in the laboratories of major oil companies during the 1960s and 1970s. At that time, researchers were confident that they finally were beginning to understand the fundamental conditions of oil generation, expulsion, and migration.

2.4 Source Rock-Based Approaches: Material Balance Methodology

In the material balance approach, the volumes of hydrocarbons available to charge traps (the charge risk) are estimated through a series of estimates and calculations. J. T. Smith and his colleagues at Shell's Bellaire Research Center near Houston, Texas, developed one of the earliest versions of the approach in the 1970s. The Shell group based their calculations on their definition of the petroleum system. The petroleum system was defined to include all related thermally mature source rocks, migration paths, reservoirs, seals, and hydrocarbons. The material balance calculation followed an explicit sequence of steps:

1. The description and definition of the petroleum system, based on concepts and data of the relationships between postulated source rocks and oils and gases inferred to have been generated from those source rocks.
2. A fetch area (G) of hydrocarbon generation is estimated for the defined petroleum system.
3. Primary migration efficiency (PM), the fraction of generated hydrocarbons that actually are expelled from the source rock, is estimated for the system.
4. The fraction of expelled hydrocarbons that moves through permeable migration conduits of the petroleum system is evaluated. This fraction is usually referred to as the secondary migration (SM).
5. An explicit estimate is made of loss (L) along the migration pathway because secondary migration incurs significant loss of hydrocarbons due to residual saturation along the migration route.
6. The charge available to the traps of the petroleum system is then calculated:

$$\text{Charge} = G(PM)(SM)L.$$

Advanced versions of the basic material balance approach have been some of the most complex and data-intensive methods ever applied to resource analysis. For example, the Prospect Appraisal by Quantitative Calibration System used by Royal Dutch/Shell in The Hague during the 1980s employed a Bayesian, calibrated geologic model. In this system, input parameters include charge, structure, reservoir properties, and retention (seal efficiency). The model follows a scheme of Bayesian updates and uses equations derived from geologic calibration studies, which in turn relied on statistical analysis of extensive data sets from Shell's worldwide exploration experience. The model was thus conceptually based on material balance, but was broadly experimental and empirical in its calibration and application.

These geochemical mass-balance approaches have the advantage of relying on concepts of physical geological processes, which, although subject to error, are likely to be less influenced by preconceptions than more subjective methods such as a Delphi assessment. The mass-balance method is scientifically elegant in its attempt to base resource evaluation on scientific principles. Geochemists and research-oriented geologists often favor this approach because, in principle, uncertainty in resource estimation can be systematically reduced through scientific investigation.

The principal disadvantages of the material balance approach are that it is time-consuming, data-intensive, expensive, and difficult to verify. Material balance studies, by definition, are state-of-the-art, with scientific updates constituting a fundamental part of the assessment process. Even after decades of research, uncertainties concerning types and distributions of organic matter and the physics of thermal maturation are great. Models of secondary migration are highly uncertain due to complexities in mapping fluid conduits and determining flow properties. Uncertainties surrounding the thickness, depth, porosity, seal competence, and timing of traps are as great in this approach as in the trap-based methodologies. For all these reasons and more, verifiable geochemical mass balance assessments remain a desirable, but as yet unattainable goal.

2.5 Resource Appraisal by Petroleum System

Some organizations have sought to combine the strengths of the geochemical material balance concepts with the empirical drilling and discovery data of the play- and prospect-based approaches. A widely (but by no means universally) adopted definition is as follows:

The petroleum system is a natural system that encompasses a pod of active source rock and all related oil and gas and which includes all the geologic elements and processes that are essential if a hydrocarbon accumulation is to exist.

The petroleum system as defined by Magoon and Dow is thus similar to the definition used at Shell in the 1970s. For each system, the geographic and stratigraphic extent of the fluids and source rocks is mapped and the maps provide the framework within which plays or assessment units can be evaluated. The petroleum system becomes the primary conceptual driver for resource analysis and evaluation of charge risk becomes a critical concern. Once the petroleum system is defined and subdivisions of assessment units or plays are identified, resource assessment can proceed in a manner similar to that of the prospect simulation or play analysis methods. By directly relating exploitable accumulations to source rocks and their burial history, the play or prospect analysis is made more scientifically grounded in those geological processes responsible for petroleum formation. Despite sharing many common features with play analysis, the petroleum system approach unequivocally places emphasis on the source rock–hydrocarbon fluid relationship and all analyses are based on this pairing.

2.6 Significance and Use of Geological Methods in Historical Performance, Volumetric Yield, and Discovery Process Modeling Methods

Historical performance, volumetric yield, and discovery process modeling methods of resource analysis depend largely on the evaluation of drilling statistics, discovery rates, productivity rates, and known field size distributions. Although they do not incorporate geological information directly, each method explicitly or implicitly relies on geological information for success.

In the historical performance approach, historical data of drilling, discovery, and production may be mathematically fitted to various logistic or growth functions from which past performance can be extrapolated into the future. Probably the most famous of all oil and gas assessment methodologies is the historical performance model developed by M. King Hubbert, who extrapolated discovery and production data to more or less successfully predict the year of maximum U.S. oil production (1970). Hubbert's methods have had a resurgence of popularity as they have been applied to evaluations of global oil and gas data to predict future world oil production. The historical performance methods do not directly incorporate geological information and it is important to note that such techniques cannot be used in frontier areas where no exploration or production has occurred. Moreover, historical performance methods are really adequate only in areas in their later stages of exploration and development or in regions where a geologic analogue of the historical data can be reliably applied.

Discovery process modeling is a sophisticated approach to resource assessment, in which numbers and sizes of accumulations are extrapolated to future discoveries. Most discovery process models are based on numbers of wells drilled and areas remaining unexplored. As with other historical performance methods, discovery process modeling works best in areas of well-defined, consistent geology and a reasonably long history of exploration and development.

In the areal or volumetric yield methods, resources are calculated as the product of the land surface area or volume of rock under assessment, multiplied by a yield factor determined from analogous areas or rock volumes. An attempt is made to determine the quantities of oil yet to be found in untested areas or rock volumes, based on the volumes previously found in geologically similar settings. The results of such studies are typically expressed in terms of volumes of oil or gas per unit area or per unit volume. As with the historical performance models, yield methods do not directly utilize geologic information but rely on geological considerations in defining areas to be assessed. Areal or volumetric yield methods were used by some of the pioneering scientists in the field of resource assessment. Examples of early works include the studies of Weeks, Zapp, and Hendricks. Early approaches included worldwide average yields expressed in terms of barrels of oil per square mile of sedimentary rock, applied over the sedimentary cover of the entire planet.

Although such broadly generalized approaches are rarely used alone, the concept of undiscovered quantities of hydrocarbons per unit of untested rock remains fundamental to geologically based methods. The success of the areal or volumetric yield approach depends critically on the selection of meaningful geologic analogues; the method works only to the degree that explored and unexplored areas are comparable. The problem, of course, is that no two geologic provinces are identical.

SEE ALSO THE FOLLOWING ARTICLES

Biomass Resource Assessment • Markets for Natural Gas • Markets for Petroleum • Natural Gas Resources, Global Distribution of • Oil and Natural Gas: Economics of Exploration • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Oil and Natural Gas Resource Assessment: Classifications and Terminology • Oil and Natural Gas Resource Assessment: Production Growth Cycle Models • Petroleum Property Valuation

Further Reading

ANWR Assessment Team (1999). "The Oil and Gas Resource Potential of the Arctic National Wildlife Refuge 1002 Area, Alaska," U.S.G.S. Open File Report 99-34. U.S. Geological Survey. [2 CD-ROMs]

Attanasi, E. D. (1998). "Economics and the 1995 National Assessment of United States Oil and Gas Resources." U.S. Geological Survey Circular 1145, U.S. Geological Survey.

Bird, K. J. (1999). Assessment overview. *In* "The Oil and Gas Resource Potential of the Arctic National Wildlife Refuge 1002 Area, Alaska" (ANWR Assessment Team, Eds.), U.S.G.S. Open File Report 99-34, pp. AO1–56. U.S. Geological Survey. [2 CD-ROMs]

Brekke, H. (1989). Play analysis of the Norwegian North Sea: Appendix 12. *In* "Proceedings of the Sixth Meeting of the Working Group on Resource Assessment, Bangkok, Thailand, September 1989," United Nations Committee for Co-ordination of Joint Prospecting for Mineral Resources in Asian Offshore Areas (CCOP), CCOP Technical Secretariat.

Campbell, C. J., and Laherrer, J. (1998). The end of cheap oil. *Sci. Am.* **278,** 78–83.

Deffeyes, K. (2002). "Hubbert's Peak: The Impending World Oil Shortage." Princeton University Press, Princeton, NJ.

Demaison, G. J., and Huizinga, B. J. (1991). Genetic classification of petroleum systems. *Am. Assoc. Petroleum Geol. Bull.* **75,** 1626–1643.

Dolton, G. L., Bird, K. J., and Crovelli, R. A. (1987). Assessment of in-place oil and gas resources. *In* "Petroleum Geology of the Northern Part of the Arctic National Wildlife Refuge, Northeastern Alaska," US Geological Survey Bulletin 1778 (K. J. Bird and L. B. Magoon, Eds.), pp. 277–298. U.S. Geological Survey.

Gautier, D. L., Dolton, G. L., Varnes, K. L, and Takahashi, K. I. (1995). "National Assessment of Oil and Gas Resources," U.S. Geological Survey Digital Data Series 30. U.S. Geological Survey. [CD-ROM]

Hunt, J. M. (1996). "Petroleum Geochemistry and Geology." 2nd ed. Freeman, New York.

Law, B. E., Spencer, C. W., Charpentier, R. R., Crovelli, R. A., Mast, R. F., Dolton, G. L., and Wandrey, C. J. (1989). Estimates of gas resources in over pressured low-permeability Cretaceous and Tertiary sandstone reservoirs, Greater Green River Basin, Wyoming, Colorado, and Utah. *In* "Gas Resources of Wyoming: 40th Wyoming Geological Association Field Conference, Casper, Wyoming, 1987" (J. L. Eisert, Ed.), pp. 39–62.

Magoon, L. B., and Dow, W. G. (1994). "The Petroleum System—From Source to Trap." American Association of Petroleum Geologists Memoir 60, American Association of Petroleum Geologists, Tulsa, OK.

National Petroleum Council. (1991). "U.S. Arctic Oil and Gas." National Petroleum Council, Washington, DC.

Oil Resources Panel. (1993). "An Assessment of the Oil Resource Base of the United States," U.S. Department of Energy Publication DOE/BC-93/1/SP. U.S. Department of Energy.

Sluijk, D., and Nederlof, M. H. (1984). Worldwide geological experience as a systematic basis for prospect appraisal. *In* "Petroleum Geochemistry and Basin Evaluation," American

Association of Petroleum Geologists Memoir 35 (G. Demaison, and R. J. Murris, Eds), pp. 15–26. American Association of Petroleum Geologists.

Ulmishek, G. F., and Klemme, H. D. (1990). "Depositional Controls, Distribution, and Effectiveness of World's Petroleum Source Rocks." U.S. Geological Survey Bulletin 1931, U.S. Geological Survey.

U.S.G.S. World Assessment Team. (2000). "U.S. Geological Survey World Petroleum Assessment 2000—Description and Results," U.S. Geological Survey Digital Data Series DDS-60. U.S. Geological Survey. [4 CD-ROMs]

White, D. A. (1988). Oil and gas play maps in exploration and assessment. *Am. Assoc. Petroleum Geol. Bull.* **72,** 944–949.

Oil and Natural Gas Resource Assessment: Production Growth Cycle Models

JEAN LAHERRERE
Association for the Study of Peak Oil and Gas
Uppsala, Sweden

1. **Quality of Discovery and Production Data**
2. **How Many Models?**
3. **Cases**
4. **Other Models and Ultimate Assessments**
5. **Conclusions**

Glossary

creaming curve Displays the cumulative discoveries versus the cumulative number of exploratory wells (new field wildcats).
cumulative production The addition of all production for a given area at a given date, from the start of production to this date.
Hubbert curve A bell-shaped curve representing annual production versus time.
parabolic fractal A display in a size rank (by decreasing size), log–log format showing a curved plot.
production growth cycle Where production starts at zero and goes up and down to end back at zero again, but possibly with several peaks.
proven reserves Those quantities of petroleum that, by analysis of geological and engineering data, can be estimated with reasonable certainty to be commercially recoverable from a certain date forward from known reservoirs and under current economic conditions, operating methods, and government regulations.
reserve additions For a given year, the addition in the reserve reporting by means of adjustments, revisions of past estimates (net of sales and acquisitions of fields), and new field discoveries.
ultimate recovery The final cumulative production when production is completely abandoned.

The goal of oil and gas modeling is to forecast future production, and the aggregation of this future production represents reserves. Oil has to be discovered before it is produced. Discovery (i.e., reserves) and production need to be studied together. A model has to be represented by the simplest equation that best fits the past discovery and production, starting when the volume was significant. The value of a model is mainly the value of the data. Some modeling may be ignored, not because of the model but rather because the data are poor or incomplete. There is no theoretical justification for any model; only the fit between the model and the past data justifies its value.

1. QUALITY OF DISCOVERY AND PRODUCTION DATA

Oil production has several problems in terms of definition. Oil is not produced; rather, it is extracted, except when it is synthesized from coal or natural gas by chemical reactions. The term "oil" may represent crude oil only (including lease condensate for some countries such as the United States), totaling approximately 65 million barrels/day (25 billion barrels/year) of the world supply, or the demand for all liquids (including synthetic oil, condensate and natural gas plant liquids, and refinery gains), totaling approximately 75 million barrels/day (28 billion barrels/year). Oil can be reported in volume (cubic meters or barrels) or in weight (tonnes or metric tons), making the world total difficult to estimate when density is not well known or is ignored. Furthermore, some oil is lost (e.g., 2 billion barrels during the Gulf War), stolen, or unaccounted for when "cheating" occurs due to quotas. The inaccurate reporting by the International Energy Agency (IEA) in 1998 led to 600 million "missing

 © 2004 Elsevier Inc. All rights reserved.

barrels" between supply and demand, giving a false impression of an abundance of oil and a low price of $10/barrel. Only scout companies (as petrologistics) can provide fairly accurate volume estimates of the oil being shipped.

Models need to study the past reserves with time. Reserves are, in fact, the addition of all future production until the end. The production of today was part of the reserves of yesterday. It is imperative that reserves and production represent the same product. In reality, they do not. The most often used databases are the *Oil & Gas Journal* (OGJ) published every December, the *BP Statistical Review* (BP) published every June, *World Oil* (WO) published every August, and the Organization of Petroleum Exporting Countries (OPEC) (Table I).

For the world's oil, the discrepancies are striking. In addition to unrealistic figures, BP specifies that tar sands are not included in their reserves but that they are included in their production. Most published data are incoherent and/or badly defined. It is possible that the authors of these publications do not understand the importance of accuracy and/or do not want to define what is measured because they do not know.

Oil discovery has to be studied carefully. Unfortunately, the reporting of reserves to the public is a political act, and operators are reluctant to tell the truth if doing so can damage their images. Operators in the United States have to comply with Securities and Exchange Commission (SEC) rules and report only proven reserves, thereby omitting probable reserves, whereas the rest of the world reports both proven and probable reserves (close to the mean or expected value). The SEC shuns the probabilistic approach for known reserves, whereas the 2000 U.S. Geological Survey (USGS) largely uses probabilistic simulation to evaluate the undiscovered reserves. The volume of proven reserves is left to the appreciation of what is a "reasonable certainty to exist," and this can vary between 51 and 99%. (The same definition of "reasonable certainty of no harm" is used by the U.S. Food and Drug Administration [FDA] to allow the sale of new products.) Figure 1 plots the remaining oil reserves from political sources (being current proven) published by OPEC, BP, OGJ, and WO; these reserves are completely different from the technical reserves that are the present mean values backdated to the year of discovery and are confidential. Statistically, the mean (expected) values do not grow despite the fact that some values will increase and some will decrease. On the contrary, proven reserves are very conservative and, thus, are

TABLE I

Common Databases for Oil Production and Reserves

	Reserves (billions of barrels)	Production (millions of barrels/day)
OGJ end 2002	1 212.880 852	66.042 7
BP end 2002	1047.7	73.935
OGJ end 2001	1 031.553 477	63.694 8
BP end 2001	1050.3	74.350
WO end 2001	1 017.763 1	67.700 444
OPEC end 2001	1 074.850	65.498

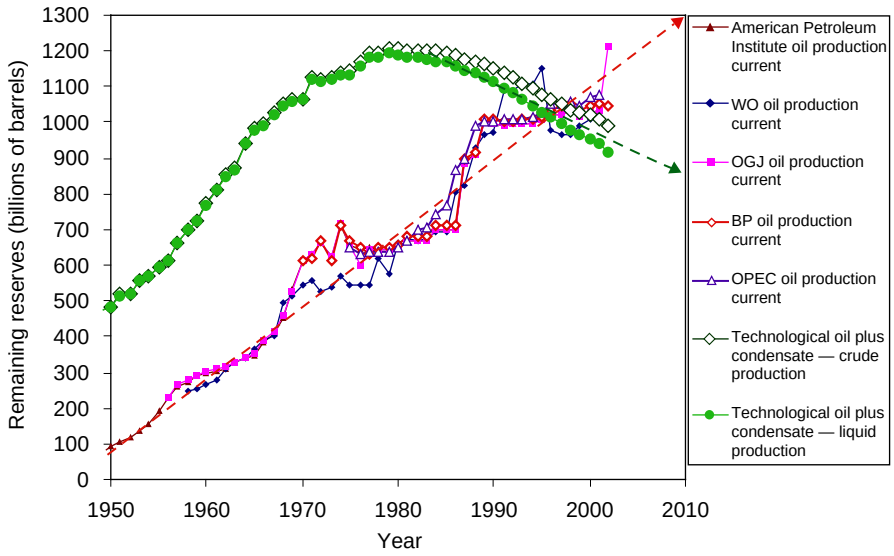


FIGURE 1 World remaining oil reserves from political and technical sources.

designed to generate growth, and this pleases most everyone (e.g., operators, bankers, shareholders). Unfortunately, although during the past quarter-century the U.S. reserve additions have come mainly from revisions of past discoveries as positive revisions were twice as large as negative revisions, the U.S. Department of Energy's (DOE) 2001 annual report indicated that negative revisions were larger than positive ones, with the Minerals Management Service (MMS) in the Gulf of Mexico reporting the same things since 1998. Only in 1990 did DOE publish "mean" estimates of annual discoveries from 1900 to 1988. These values were used, as were the annual new discoveries up to 2001, to get the present U.S. mean values by combining them with MMS reserve growth. It is unfortunate that such mean value data known by DOE are kept confidential. It is also unfortunate that the United States uses an old method to evaluate its reserves rather than using the approach used by the rest of the world.

The technical reserves decline since 1980 indicates that discovery has been lower than production since that time, but in fact the political reserves jumped by 300 billion barrels during the second half of the 1980s when OPEC quotas were decided on the reserves. OPEC countries have increased their reserves by 50% during some years (except in the neutral zone owned 50/50 by Saudi Arabia and Kuwait because the two owners disagree on the year of increase) when in fact no significant discoveries occurred during those years. Exxon–Mobil publishes world oil discoveries that stand in very close agreement with the technical reserves displayed in the Fig. 1, declining since 1980 as annual discovery became smaller than annual production.

Recently, OGJ increased Canada's reserves by 175 billion barrels to include the tar sands (extracted mainly by mining) that were previously excluded as unconventional because the Orinoco extra-heavy oils are now being produced in very conventional ways. BP, which follows OGJ for OPEC members, refused to include this large increase for the tar sand reserves even though the oil from the tar sands was included in BP production.

2. HOW MANY MODELS?

There is no theory to fully explain the best model to assess oil and gas discovery and production. The best approach is to deal with the most homogenous data covering the most natural area. In the past, exploration basins were explained in terms of tectonics. Today, they are described as "petroleum systems" in terms of generation of hydrocarbons because the most important factor is the source rocks that have generated oil and gas. Gathering fields by large petroleum systems gives much better results than does gathering fields by country. However, exploration activities depend strongly on the country (e.g., opening of blocks, fiscal terms). It is better to deal with large areas such as continents. The data should also be considered over the full life of exploration; too many studies consider only the past decade or so. The best way is to study the past and extrapolate toward the future with the best simple model. However, finding a good fit for a particular set of data does not mean that this model is necessarily the solution; perhaps another model will provide a better solution. Comparing oilfield distributions can be done with different models such as log–normal and parabolic fractal. The Gulf of Mexico, Niger delta, and Saharan Triassic petroleum systems were compared in a study. When comparing with a log–normal model, the Gulf of Mexico is different from the Niger delta and Saharan Triassic (which are similar). This model considers mainly the frequency of activity, and the Gulf of Mexico is explored intensively in search of small fields, whereas in Niger and the Sahara only large fields are drilled. When comparing with a parabolic fractal model, the Gulf of Mexico and Niger are similar in that they are dispersed habitats with many large fields of similar size, whereas the Sahara is a concentrated habitat with few giant fields. Finding a solution with one model should not keep one from looking at other models to find another solution given that nature is not linear and often involves several solutions.

The range of production and discovery models is large (e.g., creaming curve with hyperbola, cumulative production and discovery with logistic curve, annual discovery and production with normal curve or derivative of logistic (Hubbert curve), parabolic fractal for field size rank in a log–log display, log–normal distribution, stretched exponential). Each case is particular and cannot be generalized. Instead of speaking about the strengths and the weaknesses of each model, many cases should be shown to determine exactly how good the model is.

One of the most famous models for oil production is the Hubbert model, which is a bell-shaped curve that peaks at its midpoint. Hubbert was a brilliant geophysicist (for Shell and USGS) but had a short temper. He is well known by geologists for his theory on hydrodynamics, that is, tilting the water contact in oilfields when the aquifer is strongly moving. He is

also well known in the oil industry for his forecast in 1956 that U.S. oil production would peak in 1970. His theory was rejected when published but was acknowledged later when the U.S. production did in fact peak in 1970. Buzz Ivanhoe created a Hubbert Center for petroleum supply studies at the Colorado School of Mines and issues a quarterly newsletter.

2.1 Hubbert Curve: One Peak?

Hubbert, in a famous 1956 paper, contended that oil has to be found first before it is produced and that the production pattern has to be similar to the discovery pattern. He stated that production starts from zero, rises to a peak, and declines back to zero, but he did not give any equation except the bell-shaped curve. He said that the curve will likely be symmetrical. His Hubbert curve was obviously drawn by hand with an abacus, and he measured the volume by the surface below his curve, using as a unit a square in billions of barrels in the corner of the graph. He did not show any graph with several peaks, although he did not dismiss such a possibility.

Hubbert's 1956 forecast was based on the assessment that the U.S. oil ultimate could be 200 billion barrels, giving a peak in 1970 (but also 150 billion barrels). It was not until the 1980s that he wrote that his curve was the derivative of the logistic function. The logistic function was introduced in 1864 by the Belgian mathematician Verhulst as a law for population growth. The equation for the cumulative production (CP) for an ultimate U is

$$CP = U/\{1 + \exp[-b(t - tm)]\},$$

where tm is the inflexion point (corresponding to the peak time for the annual production). A population with a constant growth rate (e.g., bacteria) displays an exponential growth until a point where the population reaches the limit of resources and starts a constant decline to stabilize at a constant level. A constant growth is impossible in a limited universe. Bacteria doubling each half-hour, without being constrained by the food resource, will reach the limit of the solar system in a week and the limit of the universe in 11 days.

The derivative of the logistic curve is a bell-shaped curve very close to a normal (Gauss) curve. Its equation for the annual production P, peaking at a value Pm for the time tm, is as follows:

$$P = 2Pm/\{1 + \cosh[-b(t - tm)]\}.$$

When plotting the annual/cumulative production in percentage versus the cumulative production, if the production follows a derivative of logistics, the plot is linear. Using the mean values for the continental United States (48 states) as indicated previously, the plot is nearly linear from 1937 to 2001. The linear extrapolation toward zero indicates an ultimate of approximately 200 billion barrels. Hubbert was right on the peak in 1970 when he used an ultimate of 200 billion barrels (a rounded-up value because he knew that the accuracy was low). In Fig. 2, the linear trend

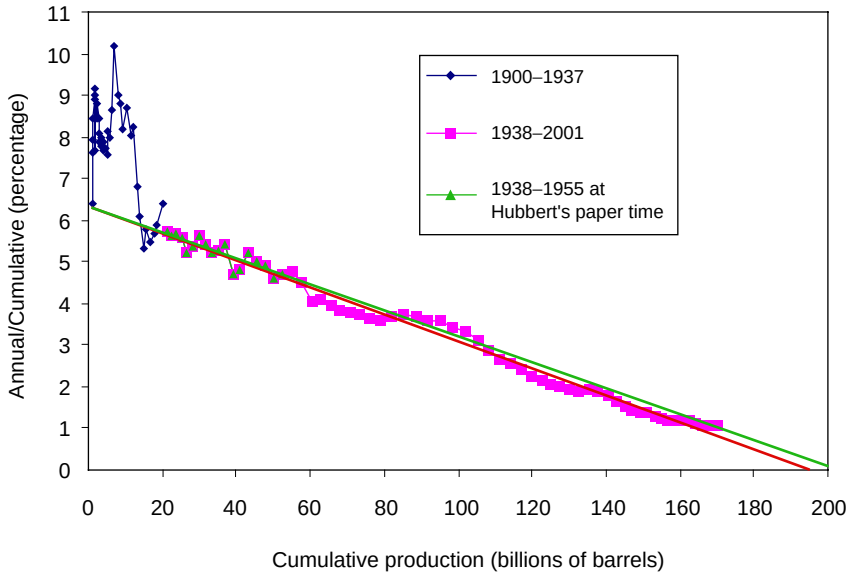


FIGURE 2 Continental U.S. (48 states) oil production: Annual/Cumulative versus cumulative.

from 1938 to 1955 (at the time of the Hubbert forecast) already indicates this 200-billion barrel ultimate. Hubbert was lucky that the real value was close to a rounded number.

Plotting the annual mean discovery, it is easy to draw a Hubbert curve that fits the discovery data (smoothed over a 5-year period) and has an ultimate of 200 billion barrels (Fig. 3). U.S. discovery in the 48 mainland states peaked around 1935 (the largest oilfield, in East Texas, was discovered in 1930) at a level of 3.2 billion barrels/year. This Hubbert discovery curve, shifted by 30 years, fits the production curve perfectly. Why is oil production in the continental United States as symmetrical in the rise as in the decline?

Plotting cumulative production versus time shows a perfect fit using one logistic curve with an ultimate of 200 billion barrels, whereas cumulative discovery needs two logistic curves: the first with an ultimate of 150 billion barrels and the second with an ultimate of 50 billion barrels.

Oil production in the mainland United States comes from more than 22 000 producers (a very large number), and randomness has to be taken into account, as in the air where there is a very large number of molecules with a random Brownian move that gives a perfect law among pressure, volume, and temperature. According to the central limit theorem (CLT), in probability, the addition of a large number of independent asymmetrical distributions gives a normal (symmetrical) distribution. The large number of U.S. independent producers leads to random behavior, and the aggregation of the very large number of fields is normal. In terms of the previous modeling of production by a bell-shaped curve, it occurs that the model seems wrong around 1930 (Great Depression), 1950 (pro-rationing); and 1980 (oil prices) as well as when political or economic events obliged all operators to act in the same direction so that randomness cannot apply.

It is easier to work on cumulative discovery (CD) and production (CP) as small details are smoothed. In the continental United States, the discovery can be well modeled with two logistic curves for an ultimate of 200 billion barrels when only one logistic curve fits the production well (Fig. 4).

2.1.1 Multiple Cycles

But there are few places with such a large number of producers and continuous activity for more than a century as in the continental United States. Most other countries display several cycles of exploration activity and then of production. One good example is the nearly depleted oil production in France. A graph of the 1995 "World Oil Supply 1930–2050" for France was illustrated in a 1999 OGJ article by the author titled "World Oil Supply: What Goes Up Must Come Down, but When Will It Peak?" France's oil production was modeled with two cycles fitting the past production (up to 1994) and forecasting approximately 10 million barrels/year in 2000. Figure 5 is an update with the real 2000 value of exactly 10 million barrels. France's oil production, if a new cycle is not found, will cease in 2010. The modeling of the production gives a good fit for an 800-million barrel ultimate, whereas the ultimate

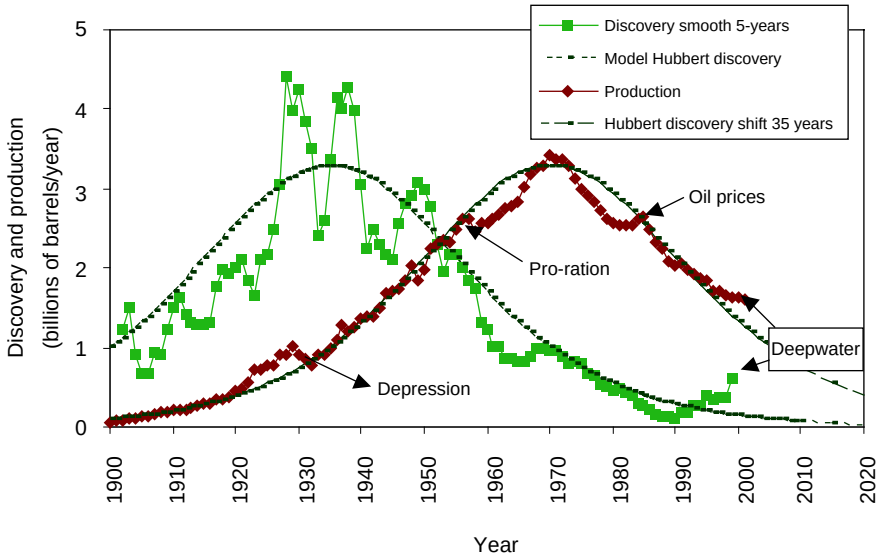


FIGURE 3 Continental U.S. (48 states): Annual oil discovery and production.

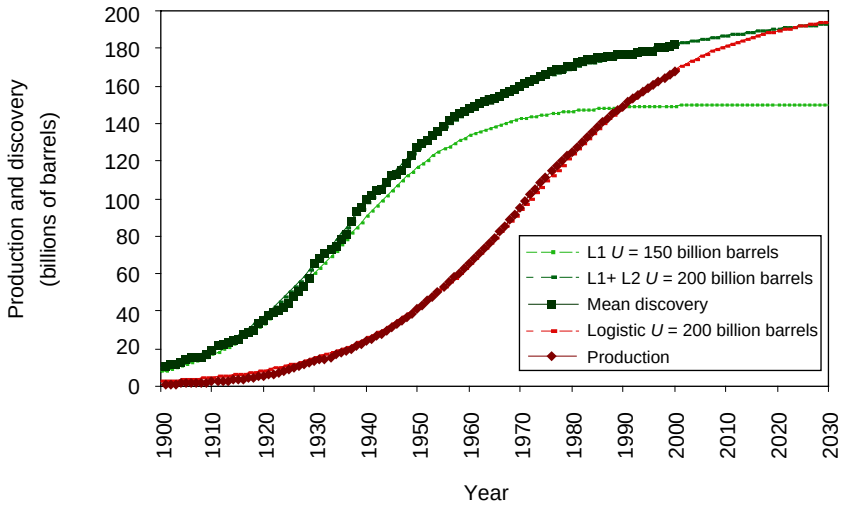


FIGURE 4 Continental United States (48 states): Cumulative oil discovery and production.

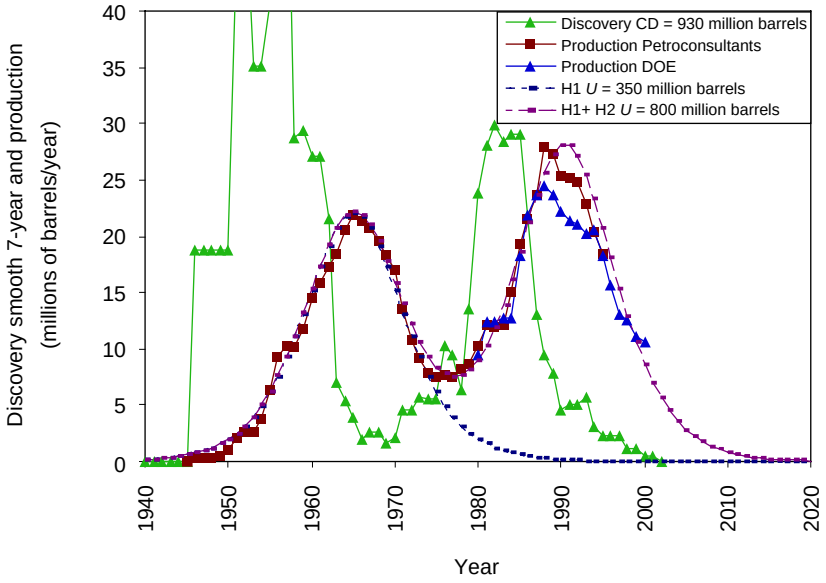


FIGURE 5 France oil discovery and production.

from discovery is higher at approximately 950 million barrels. Discoveries seems to be somewhat overestimated. But the fit between production data between Petroconsultants and DOE is poor for the first half of the 1990s, confirming the poor quality of published data.

2.2 Ultimate

Drawing a good model for oil production implies estimating the oil ultimate, for example, 200 billion barrels for the continental United States. The plot of annual/cumulative percentage versus cumulative needs to be linear to give a reliable value, but that is not always the case. The best approach is the creaming curve. This curve was "invented" by Shell during the 1980s, displaying the cumulative discoveries versus the cumulative number of new field wildcats (NFWs) so as to eliminate the ups and downs of exploration when plotting versus time. When NFWs are not available, the plot versus number of discoveries (in time sequence) can give good results. Some economists call the creaming curve the curve of cumulative discoveries

versus the rank of the field by decreasing order (as in the fractal display). This is an inaccurate label because the curve varies with new discoveries as rank changes.

2.2.1 Creaming Curve

All of the creaming curves the author modeled were easily fitted with several hyperbolas, each representing a new cycle (as shown previously for the United States). The only problem is needing to guess whether a new cycle is possible, and only geologists can answer that by studying the geological potential. In the case of the Middle East, the last cycle is from 1974 to 2002 (28 years) corresponding to 2500 NFWs having an ultimate of 50 billion barrels with 830 discoveries, whereas earlier (70 years) the ultimate was 820 billion barrels corresponding to 1400 NFWs and 340 fields. This means that from 1974 to 2002, the success ratio was 33% but there was a potential of 120 million barrels per discovery, whereas earlier the success ratio was only 24% but the average field was approximately 600 million barrels. It is obvious that between 1974 and 2002, more fields were found but were much more smaller. It is inaccurate to say, as is often done, that the Middle East was poorly explored and has a huge potential (Fig. 6). The only poorly explored place is the Western Desert in Iraq, but the petroleum system is perfectly known and is mainly gas prone, the oil potential is small, and there is almost no deepwater.

2.2.2 Parabolic Fractal

Another way in which to assess the ultimate of a petroleum system (taking only a part of the system will disturb the natural behavior) is to plot in fractal display the field size versus the rank of the field by decreasing order in a log–log format. The linear fractal (power law), as developed by Benoit Mandelbrot, is only a theoretical interpretation given that every natural system displays a curved pattern. Natural distributions are fractal, and self-similarity is the rule of nature. On an illustration of a geological outcrop, a scale is needed. Fractal equations are used in movies to create "natural background."

Every decade, the study of the evolution of discoveries leads to the drawing of the parabola representing the ultimate in the ground. The "yet to be found" can be deducted from the "yet in the ground" with some economic and time constraints. The example of the Niger delta petroleum system (covering three countries) shows that the 1960 pattern was quickly replaced by a pattern where the largest fields already found did not change much and only smaller fields were added (Fig. 7). The parabola represents the ultimate in the ground representing a large number of small fields that will be never discovered. The slope at the beginning (for the first 10 fields) characterizes the habitat, where it is a dispersed habitat (the slope is nearly flat), in contrast to a concentrated habitat (e.g., the Saharan Triassic, which shows a steep slope).

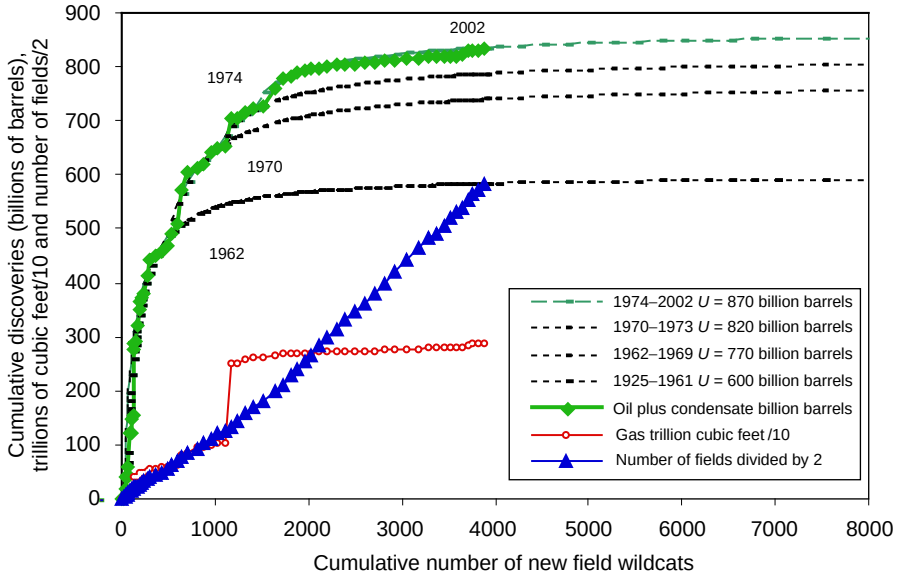


FIGURE 6 Middle East oil and gas creaming curve.

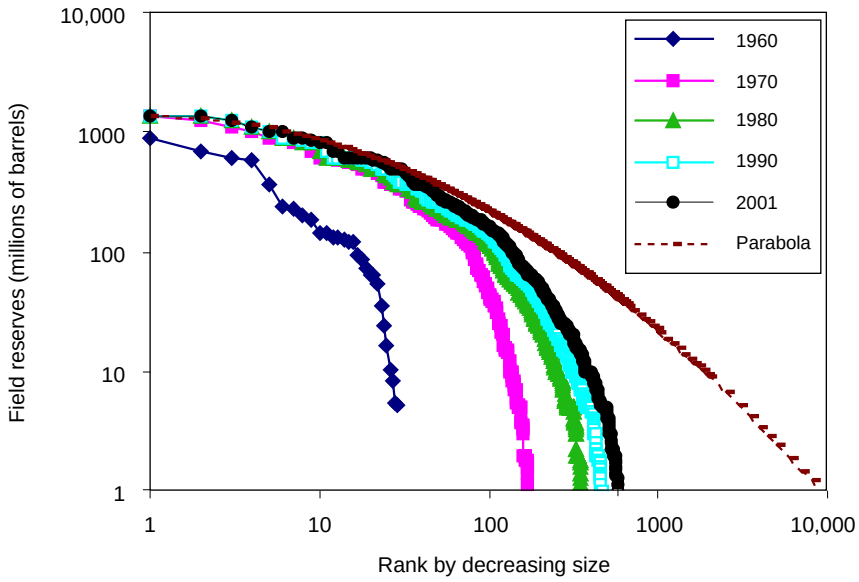


FIGURE 7 Niger delta oilfield size fractal display.

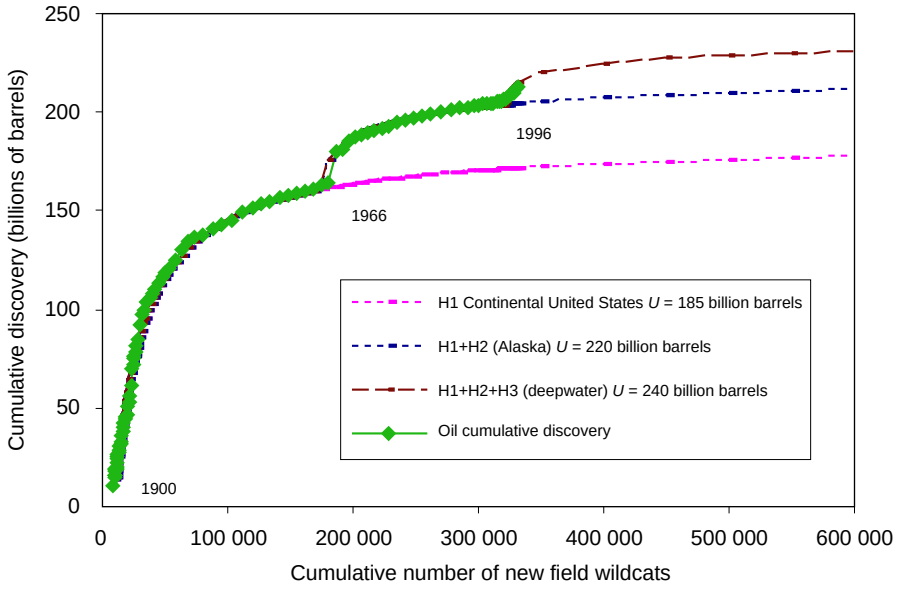


FIGURE 8 U.S. oil creaming curve.

3. CASES

3.1 United States

The creaming curve for the full United States (all 50 states) displays three cycles, well fitted to data with hyperbola curves. It is easy to recognize a first cycle for the continental United States (48 states), a second cycle for Alaska, and a third cycle for deepwater. Without a new cycle, the U.S. oil ultimate is approximately 225 billion barrels. The creaming curve represents the well-known law of diminishing returns in mineral exploration. Big gold nuggets are found first, and fine particles are found at the end. The creaming curve indicates that finding an additional 10 billion barrels will require more than 50 000 NFWs (Fig. 8).

The annual/cumulative versus cumulative plot is nearly linear from 1940 to 2001, confirming an ultimate of approximately 220 billion barrels, very different from the ultimate estimated by the USGS 2000 report of an unrealistic 362 billion barrels (of

which 83 billion barrels were undiscovered and 76 billion barrels were of reserve growth). This annual/cumulative versus cumulative plot could be good in the case of one cycle, but it could prevent the new cycle from being seen coming or the deviation from the logistic function.

With a U.S. ultimate of 220 billion barrels coming from the production trend, it is easy to plot a logistic curve for the cumulative production. But it is interesting to compare it with the cumulative mean discovery. This mean discovery comes from the 1990 Energy Information Administration (EIA) report and the last annual reports that are grown to get the mean. This estimate is questionable, but it is striking to compare discovery shifted by 30 years with the production. Despite the jump due to Prudhoe Bay, the shifted discovery fits the production very well.

The published annual proven reserves plus cumulative production are plotted in Fig. 9 and in fact follow roughly in parallel the cumulative production, and a linear extrapolation for the last 30 years will go to 280 billion barrels in 2030. This is much lower than the estimates of the 2000 USGS with an ultimate in 2025 of 362 billion barrels (of which 76 billion barrels is for the so-called reserve growth), and even the ultimate without reserve growth at 296 billion barrels looks too high. The proven reserves are a poor estimation of the future and should be replaced by the practice of using mean reserves, as is done by the rest of the world.

3.2 World Oil

The world cumulative conventional oil discovery and production are fairly well modeled, with one logistic curve having an ultimate of 2 trillion barrels but with different slopes (Fig. 10). Again, the USGS ultimate of 3.012 trillion barrels looks unrealistic in light of the past mean discovery, but when the questionable reserve growth is removed, it reduces to a more acceptable 2.3 trillion barrels.

The cumulative discovery can be broken down by size of fields, separating the large fields of more than 2 billion barrels (which make approximately half of the total volume), the giants (500–2000 million barrels), the majors (100–499 million barrels), the rest (0–99 million barrels) (Fig. 11). All curves can be modeled well with one logistic, with the largest having reached close to the ultimate first. It is the principle of the creaming curve or the law of diminishing returns. Global results give a simple model as it smoothes the discrepancies.

The logistic model can also model the total primary energy consumption. (Unfortunately, these data omit the muscular energy from humans and animals—energy that built many empires—with only two cycles with an ultimate, in fact a peak, of 12 billion tonnes.) The second cycle is associated with the demographic explosion (Fig. 12).

When the ultimate has been assessed, for the world liquids, as 3 trillion barrels, the best way in which to plot the possible future production is as

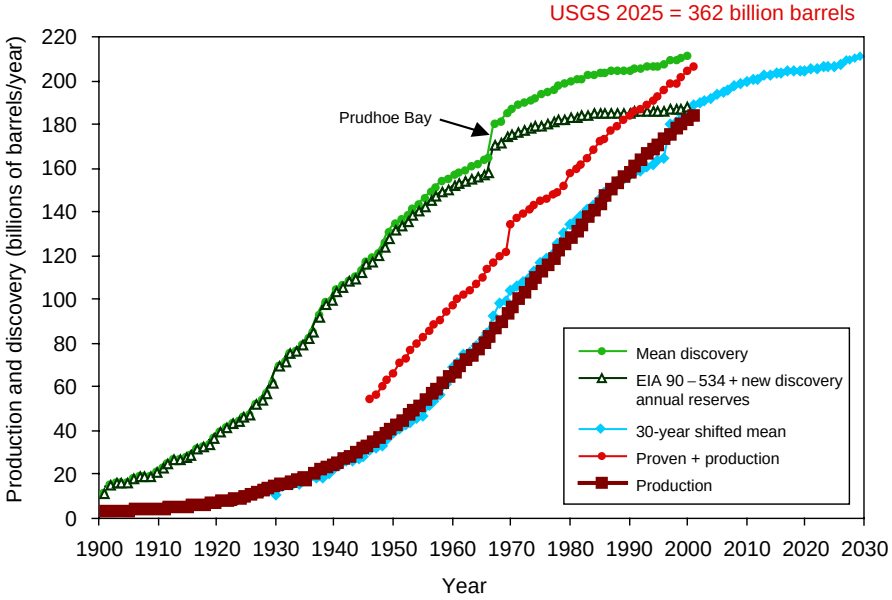


FIGURE 9 U.S. cumulative oil production and shifted mean discovery.

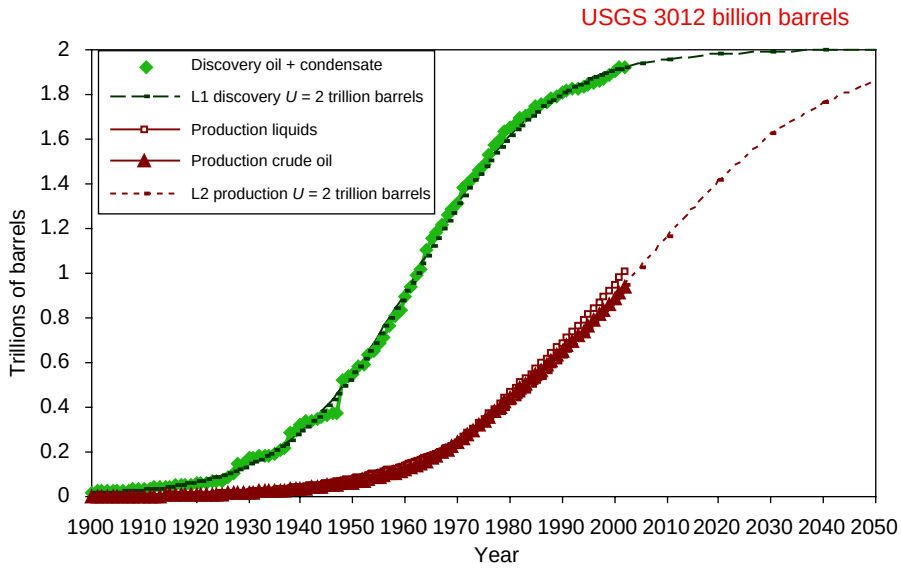


FIGURE 10 World cumulative conventional oil discovery and production.

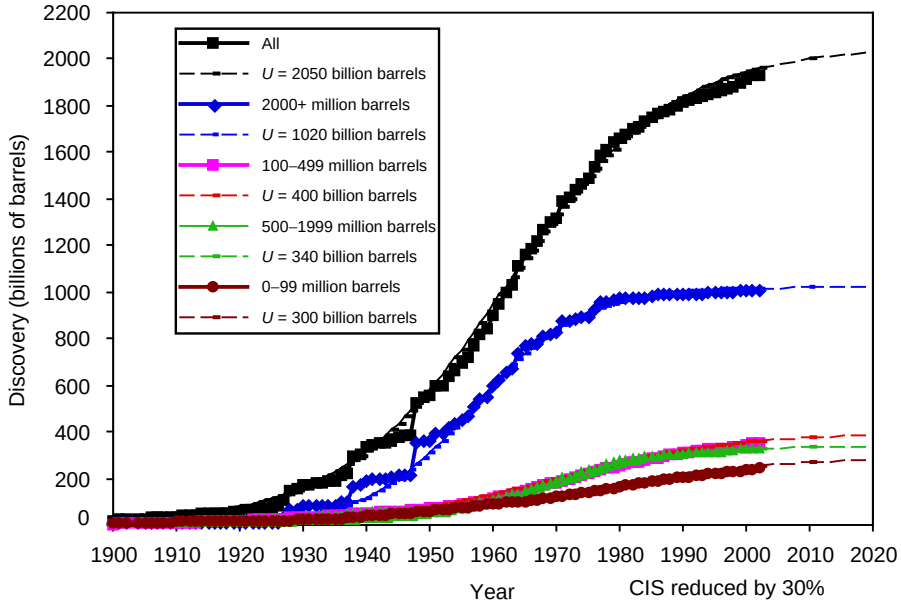


FIGURE 11 World oil cumulative discovery by field size.

being an addition to the OPEC Hubbert curve (ultimate 1.3 trillion barrels) and the non-OPEC curve (ultimate 1.6 trillion barrels), with the assumption that the production would be constrained not by the demand but rather by the supply. A smooth curve is better than some scenarios, as Wood noted from the USGS 2000 report, where many scenarios are drawn with a constant growth up to a peak that is followed by a constant decline. Such an angular pattern looks unnatural. The peak could be approximately 90 million barrels/day by around 2015. It is likely that the demand will put down the supply curve given that the official scenarios of oil production are based on constant growth corresponding to economic growth of more than 3% per year for the next 20 years or so, leading to a DOE forecast of 120 million barrels/day in 2025 (without any sign of peaking). This level appears to be nearly impossible to reach from the standpoint of technical data.

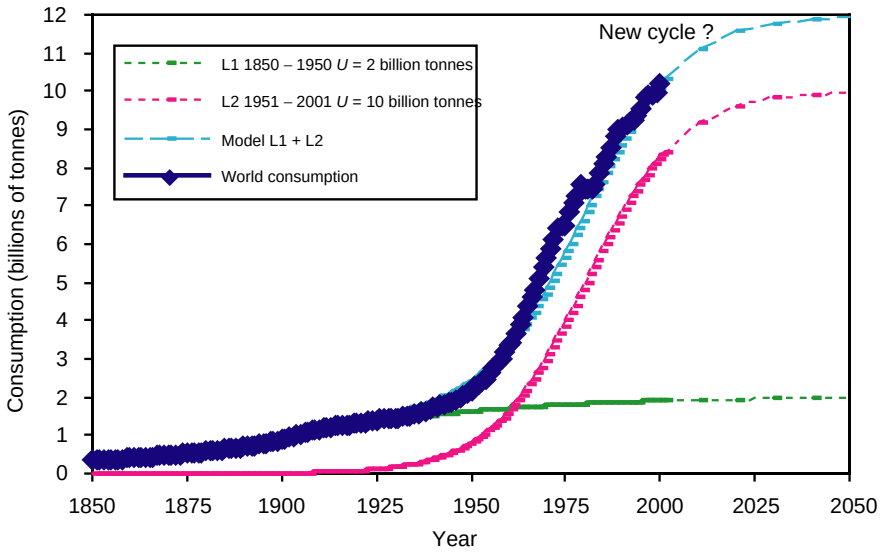


FIGURE 12 World primary energy consumption.

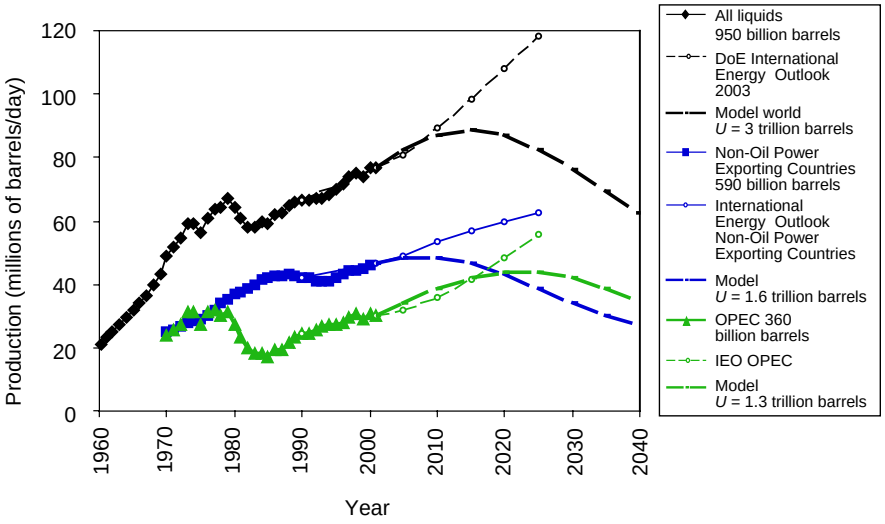


FIGURE 13 World liquids production and forecasts.

If the demand stays at a level of approximately 80 million barrels/day (a bumpy plateau), the real decline will be after 2020 (Fig. 13). The supply could constrain the demand only after 2020. If the supply is high enough to satisfy demand until that time, oil prices could be chaotic unless OPEC is able to keep its price mechanism working. Low prices will lead to more demand and then to a lack of supply.

For natural gas, the ultimate is 10 pounds/cubic foot for conventional gas and 12 pounds/cubic foot for all gas, giving a peak around 2030 of roughly 130 trillion cubic feet/year (Fig. 14). DOE's 2003 International Energy Outlook forecasts 180 trillion cubic feet/year in 2025 with no sign of slowing down. An ultimate with 9 pounds/cubic foot will give a peak of approximately 110 trillion cubic feet/year around 2020.

3.3 Hubbert Peak, Midpoint, and Constraints for the World

In the media today, the Hubbert symmetrical curve (and peak) is often found to describe the coming oil decline, and many believe that the Hubbert peak arriving at the midpoint (when half of the ultimate is

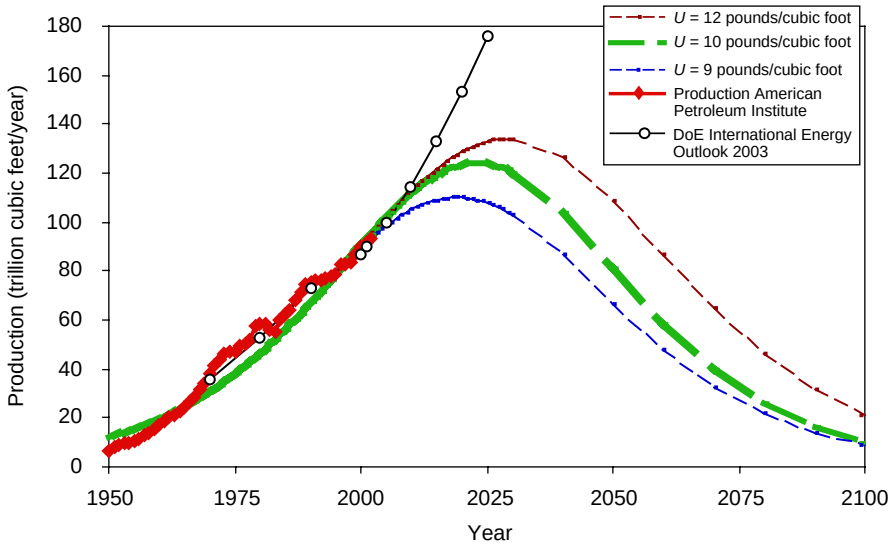


FIGURE 14 World natural gas production and forecasts.

produced) is the rule. A single symmetrical curve occurs only in places where exploration and production are active, without any interruption or any constraints, as in the continental United States and (probably) Norway. But even if so far there has been no constraint from the oil supply (except for a short time given that transport needs some time), there are many cases of constraint from the demand. Production was reduced in 1930 in Texas (with the creation of the Texas Railroad Commission, which was the model for the Organization of Petroleum Exporting Countries [OPEC] in 1967) as the discovery of the East Texas oilfield led to the fall of the oil price from \$1/barrel to \$0.1/barrel and in 1979 as the demand fell due to energy savings in light of the high oil prices. The world oil production peaked in 1979, and it took 15 years to get back to that same level. Figure 13 shows the forecast if there is no demand constraint, but despite the official forecast of 90 million barrels/day in 2010 and 100 million barrels/day in 2015, the supply could provide only 90 million barrels/day in 2010 and 2015 due to the anticipated problems in the world economy for the next decade. It is likely that the demand will constrain the supply given that the potential for oil savings is huge, particularly in the United States, where energy consumption is double that in Europe for a similar lifestyle. The supply will not really constrain the demand until after 2020.

Hubbert curves for a world liquid ultimate of 3 trillion barrels are plotted first in H1 as fitting the quick rise from 1950 to 1979 and giving a peak around 2000 at 140 million barrels/day, second in H2 as fitting the past 20 years or so and giving a peak in 2012 at 80 million barrels/day (but the model is too high before 1950), and in two cycles in H3 + H4 as fitting the data since 1950 with a peak around 2020 at 90 million barrels/day. The forecast in Fig. 15 (coming from a long study and adding several Hubbert curves) gives a peak at 90 million barrels/day around 2015, but the midpoint is not reached until 2019 because the global curve is not symmetrical.

3.4 North America Gas

Whereas there is only one oil market because transport is cheap, there are three gas markets because transport is 6 to 10 times more expensive. The North American gas market has been supplied only locally. The mean gas discovery (2P = proven + provable) of the United States, Canada, and Mexico is shifted by 20 years to fit the production that is plotted as raw, dry, and dry minus unconventional (coalbed methane and tight gas). The graph in Fig. 16 was done in 2000 when production was still rising. The author's forecast at that time was for a soon-to-come sharp decline.

The combined U.S. and Canada natural gas discovery is a perfect logistic curve for an ultimate of 1700 trillion cubic feet (Fig. 17), but the USGS 2000 report gave an ultimate of 2118 trillion cubic feet because USGS assumed that the reserve growth in the United States would bring 355 trillion cubic feet without saying how much in Canada. Removing this reserve growth, the ultimate is then 1763 trillion cubic feet, in close agreement with the estimate. The

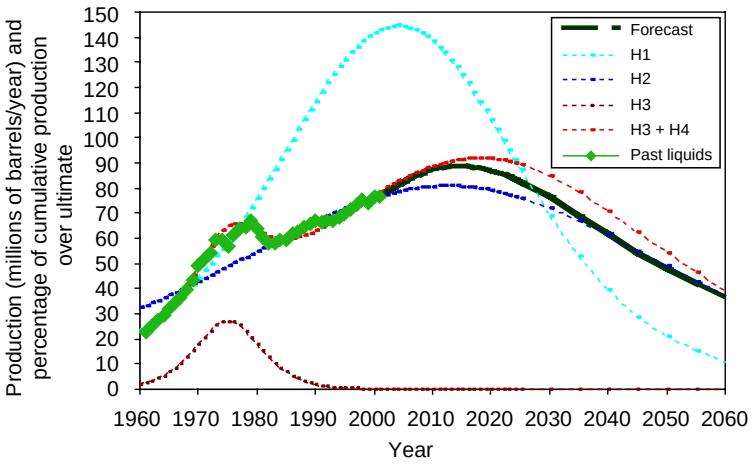


FIGURE 15 World liquids production with forecast and Hubbert curves for 3 trillion barrels.

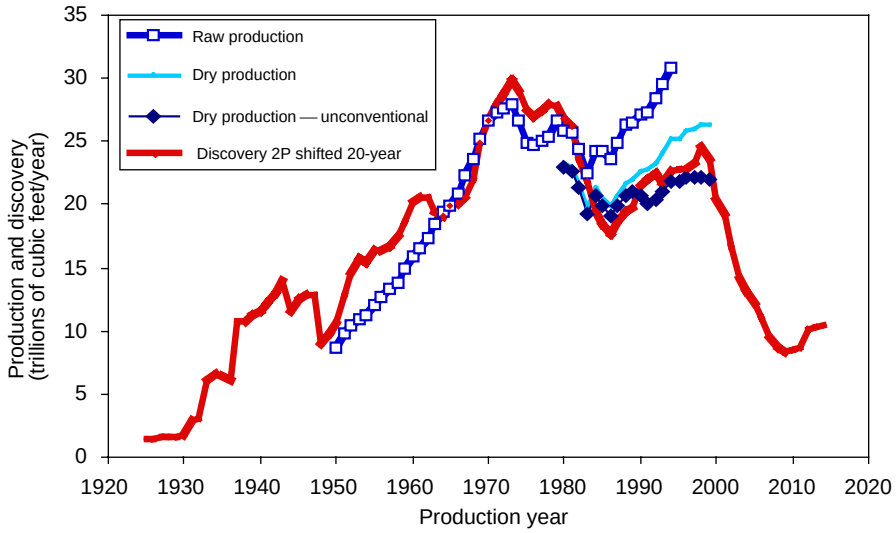


FIGURE 16 U.S., Canada, and Mexico natural gas: production and shifted discovery.

production is following the same logistic curve with a shift of 33 years. It means that the future gas production seems to be fairly settled for the next 30 years or so.

This confirms that the best modeling is not a model but rather a correlation of the production with a shifted discovery. Without any assessment of the ultimate, the part of the shifted discovery beyond current year is the best forecast.

4. OTHER MODELS AND ULTIMATE ASSESSMENTS

Albert Bartlett has written several articles on the Hubbert peak, but he uses a normal curve for annual production instead of a derivative of logistic, with the latter being better in theory assuming the influence of randomness in countries where there is a large number of producers. But there are many countries with few producers. The cumulative production is then modeled by an integral of the normal curve, and this is more complex to handle than the logistic function.

Richard Duncan has developed some software (a kind of "black box") modeling the coming peak, but because he has no access to the confidential discovery values, his model relies only on past production. He received some help from Walter Youngquist in getting some geology into his modeling. Duncan compared three different forecasts by Campbell, Duncan, and Smith, but the values of the

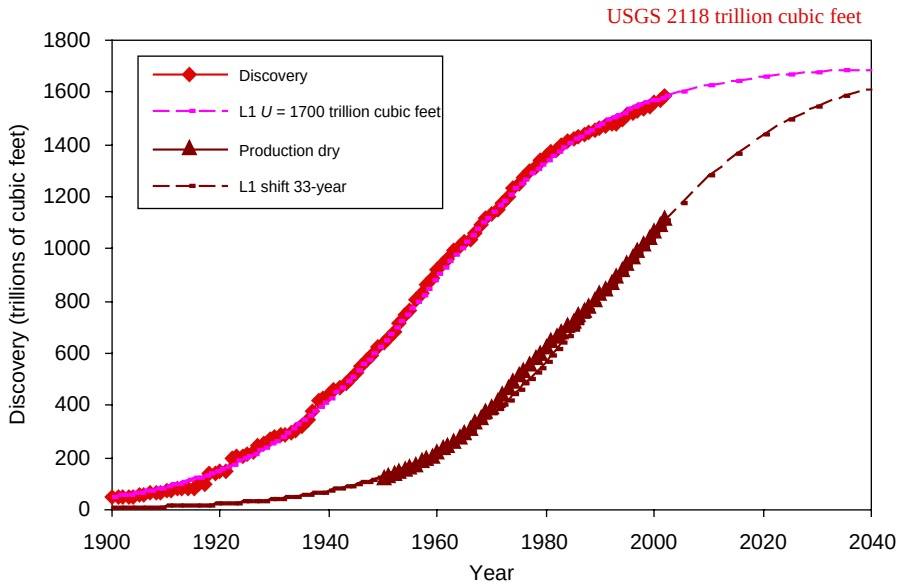


FIGURE 17 U.S. and Canada natural gas cumulative discovery and production.

three forecasts starting in 2000 are significantly different, meaning that they do not deal with the same product and so lead to differences due not only to the models.

Michael Smith published a report titled "Analysis of Global Oil Supply to 2050," which provides forecasts up to 2100 for every producing country based on an ultimate and a constant decline from the peak. The peak of conventional and unconventional oil is forecasted at approximately 90 million barrels/day in 2012. It is a very thorough work, updating a similar report written in 1994 by Campbell.

Ken Deffeyes wrote a good book, *Hubbert's Peak,* in which he stated that, despite having discussions and eating lunch many times with Hubbert, Deffeyes did not dare to ask him about the earliest roots of his prediction. Deffeyes uses the annual/cumulative versus cumulative plot, and Hubbert could have done so as well in 1956. Hubbert was clever enough to use rounded figures because he was aware of the poor accuracy of any oil data. But he apparently used his knowledge about discovery more than he used a plot on production.

The base of a production forecast is the ultimate (often called the estimated ultimate recovery [EUR]). Geologists have published their estimates for the world since 1942, and for the past 30 years or so the estimates have ranged from 1600 to 3000 billion barrels (according to various definitions). In 1977, the French Petroleum Institute (IFP) used a Delphi approach by asking a large number of experts first with an individual inquiry and second with an inquiry after giving the results of the first inquiry. USGS used the Delphi process for its assessment of undiscovered reserves for their world petroleum estimates of 1984, 1987, 1992, and 1994.

The USGS 2000 report did a very good job of defining and describing the petroleum systems of the world with the help of the oil industry. But this industry help stopped when evaluating the potential of undiscovered reserves (due to confidentiality). Instead of using a Delphi process in light of the lack of experts, for each area, only one geologist was asked to give on a single sheet of paper (called the seventh approximation) six values of the minimum, median, and maximum numbers of the undiscovered fields as well as of the size. Then, Monte Carlo simulations were used to give a full distribution of the undiscovered reserves. It was the way in which East Greenland was provided with 47 billion barrels of mean undiscovered oil.

A kind of poor model used by the industry is the R/P, which is the ratio of the remaining reserves on the date versus the annual production of the current year. It is given as 40 years for oil, but this ratio is poor for two reasons: (1) the reserves are the political data rather than the technical data and (2) the model assumes that the production will stay flat for 40 years and drop suddenly to zero during the 41st year. This model is far from the Hubbert model and far from production reality.

5. CONCLUSIONS

There is no theory to justify any model for oil or gas production. The best approach is to use the best data (for reserves, it has to be the backdated mean and never the proven value) for the longest possible period on the largest possible area and to try the simplest model of second degree. Having modeled thousands of curves, it is amazing to the author to find that any event, particularly annual discovery and production, can be modeled with several symmetrical bell-shaped curves (as a Fourier analysis for a permanent wave) and that oil and gas cumulative discovery and production can be modeled fairly with several logistic curves versus time and with several hyperbolas versus the cumulative number of NFWs.

But the best tool is to find the best fit between the shifted (by a certain number of years varying between 5 and 35) annual discovery and the annual production, with this shift allowing the future production to be predicted by following the part of the discovery beyond today.

The Reserve (future production) estimate depends mainly on the reservoir characteristics that are mainly geological, and it is why the recovery factor ranges from 3% (fractured tight reservoir) to 85% (very porous reef). But economy and technology cannot change geology. Do not underestimate the geology.

SEE ALSO THE FOLLOWING ARTICLES

Biomass Resource Assessment • Markets for Natural Gas • Markets for Petroleum • Natural Gas Resources, Global Distribution of • Oil and Natural Gas: Economics of Exploration • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Oil and Natural Gas Resource Assessment: Classifications and Terminology • Oil and Natural Gas Resource Assessment: Geological Methods • Petroleum Property Valuation

Further Reading

Bartlett, A. A. (1999). An analysis of U.S. and world oil production patterns using Hubbert-style curves. *Math. Geol.* **32,** 1–17.

Campbell, C. J., and Laherrère, J. H. (1995). "The World's Oil Supply, 1930–2050." Petroconsultants, Geneva, Switzerland.

Duncan, R. C. (2000). The peak of world oil production and the road to the Olduvai Gorge. Paper presented at Geological Society of America, "Summit 2000," Reno, NV.

Duncan R. C. (2003). Three world oil forecasts predict peak oil production. *Oil Gas J.*, May 26, pp. 18–21.

Hubbert, M. K. (1956). Nuclear energy and fossil fuels. *In* "Drilling and Production Practice," American Petroleum Institute, proceedings of spring meeting, San Antonio, TX, pp. 7–25.

Ivanhoe, L. J. (1995). Future world oil supplies: There is a finite limit. *World Oil*, October, pp. 77–88.

Laherrère, J. H. (1996). "Distributions de type fractal parabolique dans la Nature." www.oilcrisis.com/laherrere/fractal.htm.

Laherrère, J. H. (1999). World oil supply: What goes up must come down—When will it peak? *Oil Gas J.,* Feb. 1, pp. 57–64. http://dieoff.com/page177.htm.

Laherrère, J. H. (2000a). Distribution of field sizes in a petroleum system: log–normal, parabolic fractal, or stretched exponential? *Mar. Petrol. Geol.* **17,** 539–546. www.elsevier.nl/cgi-bin/cas/tree/store/jmpg/cas_free/browse/browse.cgi?year=2000&volume=17&issue=4.

Laherrère, J. H. (2000b). Learn strengths, weaknesses to understand Hubbert curve. *Oil Gas J.* April 17, http://dieoff.com/page191.htm.

Laherrère, J. H. (2001). Forecasting future production with past discoveries. Paper presented at OPEC seminar, "OPEC and the Global Energy Balance: Towards a Sustainable Energy Future," Vienna, Austria. www.hubbertpeak.com/laherrere/opec2001.pdf.

Laherrère, J. H. (2003). "Future of Oil Supplies." Swiss Institute of Technology, Zurich, Switzerland. www.hubbertpeak.com/laherrere/zurich/pdf.

Laherrère, J. H., and Sornette, D. (1998). Stretched exponential distributions in nature and economy: Fat tails with characteristic scales. *Eur. Phys. J. B* **2,** 525–539.

Longwell, H. J. (2002). The future of the oil and gas industry: Past approaches, new challenges. *World Energy* **5**(3). www.worldenergysource.com/articles/pdf/longwell_we_v5n3.pdf.

Masters, C.D., Attanasi, E.D., Root, D. (1994). World petroleum assessment and analysis. *In* "Proceedings of the 14th World Petroleum Congress," vol. 2, pp. 529–542. Wiley, Stavanger Norway.

Perrodon, A., Laherrère, J. H., and Campbell, C. J. (1998). "The World's Non-conventional Oil and Gas." Petroleum Economist, London.

Smith, M. (2002). Analysts claim early peak in world oil production. *Oil Gas J.,* August 26, pp. 33–36.

U.S. Department of Energy/Energy Information Administration. (1990). "US Oil and Gas Reserves by Year of Field Discovery." DOE/EIA, Washington, DC.

U.S. Department of Energy/Energy Information Administration. (2003). "IEO 2003." www.eia.doe.gov/oiaf/ieo.

U.S. Geological Survey. (2000). "World Petroleum Assessment." http://greenwood.cr.usgs.gov/energy/worldenergy/dds-60/espt4.html.

Wood, J. (2000). "Long Term World Oil Supply (a resource base/production path analysis)." www.eia.doe.gov/pub/oil_gas/petroleum/presentations/2000/long_term_supply/longtermoilsupplypresentation.ppt.

Youngquist, W. (1997). "Geodestinies: The Inevitable Control of Earth Resources over Nations and Individuals." National Book, Portland, OR.

Oil Crises, Historical Perspective

MAMDOUH G. SALAMEH
Oil Market Consultancy Service
Haslemere, Surrey, United Kingdom

1. **Introduction**
2. **The Road to the First Oil Crisis**
3. **The Oil Weapon**
4. **The Embargo**
5. **The Second Oil Crisis**
6. **Anatomy of Oil Crises**
7. **Third Oil Crisis?**
8. **Implications for the Global Economy**

Glossary

aramco The Arabian American oil company that obtained the concession for oil exploration in the Eastern Province of Saudi Arabia in 1933. It was originally made up of four American oil companies: Standard Oil of California (Socal), Texaco, Standard Oil of New Jersey (then called Esso and later changed to Exxon Mobil), and Socony-Vacuum (Mobil). In 1976, it was purchased by the Saudi government and renamed "Saudi Aramco."

conventional crude oil The oil produced from an underground reservoir, after being freed from any gas that may have dissolved in it under reservoir conditions, but before any other operation has been performed on it. In the oil industry, simply termed "crude."

proven reserves The quantities that geological and engineering information indicates with reasonable certainty can be recovered from known reserves under existing economic and operating conditions.

renewable energy An energy source that does not depend on finite reserves of fossil or nuclear fuels, such as solar energy, wind energy, biomass, hydroelectric power, and hydrogen. All of these renewable sources involve the generation of electricity.

shale oil A distillate obtained when oil shale (a rock of sedimentary origin) is heated in retorts.

ultimate global reserves This is the amount of oil reserves that would have been produced when production eventually ceases.

unconventional oil Oil that has been extracted from tar sands, oil shale, extra heavy oil, and the conversion of natural gas to liquids, known collectively as synthetic fuels or "synfuels."

The 20th century was truly the century of oil. Though the modern history of oil begins in the latter half of the 19th century, it is the 20th century that has been completely transformed by the advent of oil. Oil has a unique position in the global economic system. One could not imagine modern societies existing without oil. Modern societies' transportation, industry, electricity, and agriculture are virtually dependent on oil. Oil makes the difference between war and peace. The importance of oil cannot be compared with that of any other commodity or raw material because of its versatility and dimensions, namely, economic, military, social, and political. The free enterprise system, which is the core of the capitalist thinking, and modern business owe their rise and development to the discovery of oil. Oil is the world's largest and most pervasive business. It is a business that touches every corner of the globe and every person on earth. The financial resources and the level of activity involved in exploring, refining, and marketing oil vastly exceed those of any other industry. Of the top 20 companies in the world, 7 are oil companies. Human beings are so dependent on oil, and oil is so embedded in daily life, that individuals hardly stop to comprehend its pervasive significance. Developing nations give no indication that they want to deny themselves the benefits of an oil-powered economy, whatever the environmental questions. In addition, any notion of scaling back the world's consumption of oil will be influenced by future population growth.

1. INTRODUCTION

No other commodity has been so intimately intertwined with national strategies and global politics

 © 2004 Elsevier Inc. All rights reserved.

and power as oil. In World War I, the Allies floated to victory on a wave of oil. Oil was central to the course and outcome of World War II in both the Far East and Europe. One of the Allied powers' strategic advantages in World War II was that they controlled 86% of the world's oil reserves. The Japanese attack on Pearl Harbor was about oil security. Among Hitler's most important strategic objectives in the invasion of the Soviet Union was the capture of the oilfields in the Caucasus. In the Cold War years, the battle for the control of oil resources between international oil companies and developing countries was a major incentive and inspiration behind the great drama of decolonization and emergent nationalism.

During the 20th century, oil emerged as an effective instrument of power. The emergence of the United States as the world's leading power in the 20th century coincided with the discovery of oil in America and the replacement of coal by oil as the main energy source. As the age of coal gave way to oil, Great Britain, the world's first coal superpower, gave way to the United States, the world's first oil superpower.

Yet oil has also proved that it can be a blessing for some and a curse for others. Since its discovery, it has bedeviled the Middle East with conflict and wars. Oil was at the heart of the first post-Cold War crisis of the 1990s—the Gulf War. The Soviet Union—the world's second largest oil exporter—squandered its enormous oil earnings in the 1970s and 1980s in a futile military buildup. And the United States, once the largest oil producer and still its largest consumer, must import almost 60% of its oil needs, weakening its overall strategic position and adding greatly to an already burdensome trade deficit—a precarious position for the only superpower in the world.

The world could face an energy gap probably during the first decade of the 21st century once global conventional oil production has peaked. This gap will have to be filled with unconventional and renewable energy sources. A transition from fossil fuels to renewable energy sources is, therefore, inevitable if humans are to bridge the energy gap and create a sustainable future energy supply. Sometime in the 21st century, nuclear, solar, geothermal, wind, and hydrogen energy sources may be sufficiently developed to meet a larger share of the world's energy needs. But for now humans will continue to live in an age of oil. Oil will, therefore, still be supplying a major share of the global energy needs for most, perhaps all, of the 21st century and will continue to have far-reaching effects on the global economy.

2. THE ROAD TO THE FIRST OIL CRISIS

One distinctive feature dominated the global economic scene in the decades following World War II. It was the rising consumption of oil. Total world energy consumption more than tripled between 1949 and 1972. Yet that growth paled in comparison to the rise in oil demand, which during the same period increased more than $5\frac{1}{2}$ times over. Everywhere, growth in the demand for oil was strong. Between 1948 and 1972, consumption tripled in the United States, from 5.8 to 16.4 million barrels/day—unprecedented except when measured against what was happening elsewhere. In the same years, demand for oil in Western Europe increased 15 times over, from 970,000 barrels/day to 14.1 million barrels/day. In Japan, the change was nothing less than spectacular; consumption increased 137 times over, from 32,000 barrels/day to 4.4 million barrels/day.

The main drivers of this global surge in oil use were the rapid economic growth and the cheap price of oil. During the 1950s and 1960s, the price of oil fell until it became very cheap, which also contributed mightily to the swelling of consumption. Many governments encouraged its use to power economic growth and industrial modernization. There was one final reason that the market for oil grew so rapidly. Each oil-exporting country wanted higher volumes of its oil sold in order to gain larger revenues.

In the buoyant decades following World War II, oil overtook coal as the main fuel for economic growth. Huge volumes of oil surged out of Venezuela and the Middle East and flowed around the world. Oil was abundant. It was environmentally more attractive and was easier and more convenient to handle. And oil became cheaper than coal, which proved the most desirable and decisive characteristic of all. Its use provided a competitive advantage for energy-intensive industries. It also gave a competitive edge to countries that shifted to it. And yet, there was a haunting question: How reliable was the flow of oil on which modern societies had come to depend? What were the risks?

Among the Arabs, there had been talk for more than a decade about wielding the "oil weapon." This was their chance. On June 6, 1967, the day after the start of the Six-Day War, Arab oil ministers, members of Organization of Arab Petroleum Exporting Countries, formally called for an oil embargo against countries friendly to Israel. Saudi Arabia, Kuwait, Iraq, Libya, and Algeria thereupon banned shipments to the United States and Great Britain.

By June 8, the flow of Arab oil had been reduced by 60%. The overall initial loss of Middle Eastern oil was 6 million barrels/day. Moreover, logistics were in total chaos not only because of the interruptions but also because, as in 1956, the Suez Canal and the pipelines from Iraq and Saudi Arabia to the Mediterranean were closed. The situation grew more threatening in late June and early July when, coincidentally, civil war broke out in Nigeria, depriving the world oil market of 500,000 barrels/day at a critical moment.

However, by July 1967, a mere month after the Six-Day War, it became clear that the selective Arab oil embargo was a failure; supplies were being redistributed to where they were needed. And by the beginning of September, the embargo had been lifted.

The 1970s saw a dramatic shift in world oil. Demand was catching up with available supply and the 20-year surplus was over. As a result, the world was rapidly becoming more dependent on the Middle East and North Africa for its oil.

Oil consumption surged beyond expectation around the world, as ever-greater amounts of petroleum products were burned in factories, power plants, homes, and cars. The cheap price of oil in the 1960s and early 1970s meant that there was no incentive for fuel-efficient automobiles. The late 1960s and early 1970s were also the watershed years for the domestic U.S. oil industry. The United States ran out of surplus capacity. In the period 1957 to 1963, surplus capacity in the United States had totaled approximately 4 million barrels/day. By 1970, it had declined to only 1 million barrels/day. That was the year, too, that American oil production peaked at 10 million barrels/day. From then on, it began its decline, never to rise again. With consumption continuing to rise, the United States had to turn to the world oil market to satisfy its needs. Net imports tripled from 2.2 million barrels/day in 1967 to 6 million barrels/day by 1973. Imports as a share of total oil consumption over the same years rose from 19 to 36%.

The disappearance of surplus capacity in the United States would have major implications, for it meant that the "security margin" on which the Western world depended was gone. For the United States, it marked a shift from: (1) oil self-sufficiency to reliance on Middle East oil; (2) being a major exporter to becoming a major importer; (3) loss of the ability to control the global oil markets at a time when Middle East oil producers (Arab Gulf producers) began to assert themselves on the global markets; and (4) inability to provide stand-by supply to its allies in an emergency. This meant a major shift from energy security to vulnerability and dependency. Indeed, the razor's edge was the ever-increasing reliance on the oil of the Middle East. New production had come from Indonesia and Nigeria (in the latter case, after the end of its civil war in early 1970), but that output was dwarfed by the growth in Middle Eastern production. Between 1960 and 1970, Western world oil demand had grown by 21 million barrels/day. During that same period, production in the Middle East (including North Africa) had grown by 13 million barrels/day. In other words, two-thirds of the huge increase in oil consumption was being satisfied by Middle East oil.

In a wider sense, the disappearance of surplus capacity caused an economic and geopolitical transformation of the global oil market. In an economic sense, the "center of gravity" of oil production, energy security, and control of global oil supplies had shifted from the United States to the Middle East. In a geopolitical sense, the oil revenue and the global dependence on Middle East oil provided the Arab Gulf producers with unprecedented political influence. This they channeled into support of their political aspirations as they did during the 1973 war and the resulting oil embargo.

Another disturbing development was that the relationship between the oil companies and the producing nations was beginning to unravel. In gaining greater control over the oil companies, whether by participation or outright nationalization, the exporting countries also gained greater control over prices. The result was the new system that was forged in Tehran and Tripoli, under which prices were the subject of negotiation between companies and countries, with the producing countries taking the lead in pushing up the posted price.

However, the supply–demand balance that emerged at the beginning of the 1970s was sending a most important message: Cheap oil had been a tremendous boon to economic growth, but it could not be sustained. Demand could not continue growing at the rate it was; new supplies needed to be developed. That was what the disappearance of spare capacity meant. Something had to give, and that something was price.

By 1972, many experts reckoned that the world was heading for an acute oil shortage in a few years. The signs of a shortage were visible everywhere. The demand for oil in the summer of 1973 was going above the wildest predictions—in Europe, in Japan, and most of all in the United States. Imports from the

Middle East to the United States were still racing up: production inside the United States was still falling. In April, U.S. President Nixon lifted restrictions on imports of oil, so the Middle East oil flowed in still faster. There was a new danger sign when Kuwait decided in 1972 to conserve its resources and to keep its production below 3 million barrels/day.

As late as 1970, there were still approximately 3 million barrels/day of spare capacity in the world outside the United States, with most of it concentrated in the Middle East. By the second half of 1973, the spare capacity had shrunk to only 500,000 barrels/day. That was just 1% of world consumption. With a 99% capacity utilization and a 1% security margin, the oil supply–demand balance was indeed extremely precarious.

In June 1973, as prices were zooming up, the Organization of Petroleum Exporting Countries (OPEC) summoned another meeting in Geneva to insist on another price increase because of the further devaluation of the U.S. dollar. The radicals—Algeria, Libya and Iraq—were pressing for unilateral control of price, but eventually OPEC agreed on a new formula that increased prices by another 12%. By September 1973, for the first time since the founding of OPEC, the market price of oil had risen above the posted price. It was a sure sign that OPEC was in a very strong bargaining position. Armed with this knowledge, OPEC invited the oil companies to meet them in Vienna on October 8 to discuss substantial increases in the price of oil.

In this atmosphere of crisis, the oil company delegates prepared to confront OPEC in Vienna on October 8. And then, just as they were leaving for Vienna, Egypt and Syria invaded Israeli-occupied territories. There was war.

While the shortage loomed, the Arabs were at last achieving closer unity. They were determined to use oil as a weapon against Israel and by 1973 the militants were being joined by Saudi Arabia. The very fact that Saudi Arabia had become the largest oil exporter made King Feisal more vulnerable in the face of his Arab colleagues and the danger of an embargo more likely, for he could not afford to be seen as a blackleg.

The international oil order had been irrevocably changed. However, it was not only a question of price, but of power. The extent of dependence by the industrial countries on oil as a source of energy had been exposed and the practicality of controlling supply as means of exerting pressure for raising the price of oil had been dramatically demonstrated. Although the oil weapon had not worked in 1967, the "rationale of those who called for its use as a weapon in the Middle East conflict has been strengthened in current circumstances."

3. THE OIL WEAPON

Contrary to popular belief, the Americans, not the Arabs, were the first to wield the oil weapon. They used it against Japan when on July 25, 1941, the United States announced a freeze on Japanese funds necessary for Japan to buy American oil, which, in practice, meant an embargo on oil. The embargo was the result of Japanese military aggression in Asia.

Increasingly worried about a cut-off of oil supplies from the United States, Tokyo instituted a policy to establish self-sufficiency and to try to eliminate dependence on U.S. oil supplies. In 1940–1941, it was energy security that led Japan to occupy the Dutch East Indies and take control of its oilfields. Indeed, the U.S. oil embargo was an important factor leading Japan to attack Pearl Harbor, bringing the United States into World War II. Oil had been central to Japan's decision to go to war.

Ever since the 1950s, the Arab world had been talking about using the oil weapon to force Israel to give up occupied Arab territories. Yet the weapon had always been deflected by the fact that Arab oil, though it seemed endlessly abundant, was not the supply of last resort.

In June 1967, 2 days into the Six-Day War, the Arabs wielded the oil weapon when they imposed an oil embargo against the United States and Great Britain. However, by July 1967, it became clear that the Arab embargo had failed. The Arabs would have to wait for another chance to wield the oil weapon again. That chance came their way when just moments before 2:00 P.M. on October 6, 1973, Egyptian and Syrian armies launched an attack on Israeli-held positions in the Sinai and the Golan Heights. Thus began the October War or, what became known as the Yom Kippur War, the fourth of the Arab–Israeli wars—the most destructive and intense of all of them and the one with the most far-reaching consequences. One of the most potent weapons used in this war was the oil weapon, wielded in the form of an embargo—production cutbacks and restrictions on exports—that, in the words of Henry Kissinger, "altered irrevocably the world as it had grown up in the postwar period."

The embargo, like the war itself, came as a surprise and a shock. Yet the pathway to both in retrospect seemed in some ways unmistakable. By

1973, oil had become the lifeblood of the world's industrial economies and it was being pumped and circulated with very little to spare. Never before in the entire postwar period had the supply–demand equation been so tight, while the relationships between the oil-exporting countries and the oil companies continued to unravel. It was a situation in which any additional pressure could precipitate a crisis—in this case, one of global proportions.

With supply problems becoming chronic in the early 1970s, talk about an energy crisis began to circulate in the United States. There was agreement, in limited circles, that the United States faced a major problem. Price controls on oil, imposed by Nixon in 1971 as part of his overall anti-inflation program, were discouraging domestic oil production while stimulating consumption. The artificially low prices provided little incentive either for new exploration or for conservation. By the summer of 1973, United States oil imports were 6.2 million barrels/day, compared to 4.5 million barrels/day in 1972 and 3.2 million barrels/day in 1970. The oil trade journal *Petroleum Intelligence Weekly* reported in August 1973 that "near-panic buying by the U.S., the Europeans, and the Japanese was sending oil prices sky-rocketing."

As global demand continued to rise against the limit of available supply, market prices exceeded the official posted prices. It was a decisive change, truly underlining the end of surplus. For so long, reflecting the chronic condition of oversupply, market prices had been below posted prices, irritating relations between companies and governments. But the situation had reversed and the exporting countries did not want to see the growing gap between the posted price and the market price go to the companies. Wasting little time, the exporters sought to revise their participation and buy-back arrangements so that they would be able to obtain a larger share of the rising prices.

One of the principal characteristics governing the operations of the oil industry is that it generates an important "economic rent" or "oil surplus," the appropriation of which involves three players: the exporting countries, the consuming countries, and the multinational oil companies. Both the exporting countries and the consuming countries are effectively staking a claim to the significant element of economic rent built into the price of oil. For the exporters, such as the Arab Gulf producers, oil remains the single most important source of income, generating approximately 85 to 90% of their revenues. Significantly, consumer countries have always looked on oil as an important source of taxation since demand for it is relatively inelastic; that is to say, it varies little as the price changes. These countries have more maneuverability when the price of oil is low. This was amply demonstrated when, as a result of the oil price collapse in 1986, many of them took the opportunity to raise tax rates on petroleum products. This practice was, to a lesser extent, in operation in the 1970s but has accelerated since 1986 in Europe, with tax levels on petroleum products reaching between 80 and 87%. In other words, the sharp increases in taxes by the consuming countries were intended to cream off more of the rent. Is it any wonder that the United Kingdom has been for years earning far more revenue from taxes on petroleum products than from its North Sea oil exports?

On September 1, 1973, the fourth anniversary of Muamer Qaddafi's coup, Libya nationalized 51% of those company operations it had not already taken over. The radicals in OPEC—Iran, Algeria, and Libya—began pushing for a revision in the Tehran and Tripoli agreements. By the late summer of 1973, the other exporters, observing the upward trend of prices on the open market, came around to that same point of view. They cited rising inflation, the dollar's devaluation, and also the rising price of oil. By September 1973, the Saudi oil minister Sheikh Ahmed Zaki Yamani was able to announce that the Tehran Agreement was dead. Even as the economics of oil were changing, so were the politics that surrounded it—and dramatically so.

By April of 1973, President Anwar Sadat of Egypt had begun formulating with Syria's Hafez Al-Asad strategic plans for a joint Egyptian–Syrian attack against Israel. Sadat's secret was tightly kept. One of the few people outside the high commands of Egypt and Syria with whom he shared it was King Feisal. And that meant oil would be central to the coming conflict.

In the early 1970s, as the market tightened, various elements in the Arab world became more vocal in calling for use of the oil weapon to achieve their economic and political objectives. King Feisal was not among them. He had gone out of his way to reject the use of the oil weapon. "It was not only useless," he said, "but dangerous even to think of that. Politics and oil should not be mixed." Yet, by early 1973, Feisal was changing his mind. Why?

Part of the answer lay in the marketplace. Much sooner than expected, Middle Eastern oil had become the supply of last resort. In particular, Saudi Arabia had become the marginal supplier for everybody,

including the United States; American dependence on the Gulf had come not by the widely predicted 1985, but by 1973. The United States would no longer be able to increase production to supply its allies in the event of a crisis and the United States itself was now, finally, vulnerable. The supply–demand balance was working to make Saudi Arabia even more powerful. Its share of world exports had risen rapidly from 16% in 1970 to 25% in 1973 and was continuing to rise.

In addition, there was a growing view within Saudi Arabia that it was earning revenues in excess of what it could spend. Two devaluations of the U.S. dollar had abruptly cut the worth of the financial holdings of countries with large dollar reserves, including Saudi Arabia.

The changing conditions in the marketplace, which with each passing day made the Arab oil weapon more potent, coincided with significant political developments. By the spring of 1973, Sadat was strongly pressing Feisal to consider using the oil weapon to support Egypt in a confrontation with Israel. King Feisal also felt growing pressure from many elements within his kingdom and throughout the Arab world. Thus, politics and economics had come together to change Feisal's mind. Thereupon the Saudis began a campaign to make their views known, warning that they would not increase their oil production capacity to meet rising demand and that the Arab oil weapon would be used, in some fashion, unless the United States moved closer to the Arab viewpoint and away from Israel.

On August 23, 1973, Sadat made an unannounced trip to Riyadh to see King Feisal. He told the king that he was considering going to war against Israel. It would begin with a surprise attack and he wanted Saudi Arabia's support and cooperation. He got it.

On October 17, 1973, 11 days into the war, Arab oil ministers meeting in Kuwait agreed to institute a total oil embargo against the United States and other countries friendly to Israel. They decided to cut production 5% from the September level and to keep cutting by 5% in each succeeding month until their objectives were met. Oil supplies at previous levels would be maintained to "friendly states." One clear objective of the plan was to split the industrial countries right from the start.

On October 19, Nixon publicly proposed a $2.2 billion military aid package for Israel. In retaliation for the Israeli aid proposal, Saudi Arabia had gone beyond the rolling cutbacks; it would now cut off all shipments of oil, every last barrel, to the United States. The oil weapon was now fully in battle—a weapon, in Kissinger's words, of political blackmail. To counter that blackmail, Kissinger called for the industrialized nations to meet in Washington, DC at the earliest possible moment. He wanted the oil-consuming West to make a united stand against the Arabs.

4. THE EMBARGO

The embargo came as an almost complete surprise despite the evidence at hand: almost two decades of discussion in the Arab world about the oil weapon, the failed embargo in 1967, Sadat's public discussion of the "oil option" in early 1973, and the exceedingly tight oil market of 1973. What transformed the situation and galvanized the production cuts and the embargo against the United States was the very public nature of the resupply of ammunitions and armaments to Israel and then the $2.2 billion aid package.

On October 21, Sheikh Yamani met with the president of Aramco, Frank Junkers. Using computer data about exports and destinations that the Saudis had requested from Aramco a few days earlier, Yamani laid out the ground rules for the cutbacks and the embargo the Saudis were about to impose. He told Junkers that any deviations from the ground rules would be harshly dealt with.

At the time of the embargo, the management of the Saudi oil industry was in the hands of Aramco (the Arabian–American Oil Company), the joint venture between Standard Oil of California (Socal), Texaco, Standard Oil of New Jersey (then called Esso and later changed to ExxonMobil), and Socony-Vacuum (Mobil).

In 1948, U.S. imports of crude oil and products together exceeded exports for the first time. No longer could the United States continue its historical role as supplier to the rest of the world. That shift added a new dimension to the vexing question of energy security. The lessons of World War II, the growing economic significance of oil, and the magnitude of Middle Eastern oil reserves all served, in the context of the developing Cold War with the Soviet Union, to define access to that oil as a prime element in Western security. Oil provided the point at which foreign policy, international economic considerations, national security, and corporate interests would all converge. The Middle East would be the focus. There the oil companies were already building up production and making new arrangements to secure their positions.

In Saudi Arabia, development was in the hands of Aramco. The company understood from the time it obtained the Saudi oil concession in 1933 that the

concession would always be in jeopardy if it could not satisfy the expectations and demands of King Abdul Aziz Ibn Saud, the founder of Saudi Arabia, and the royal family. Since then, it has worked tirelessly to enhance Saudi oil reserves and production and build terminals and pipelines for exporting the oil worldwide.

In October 1972, Sheikh Yamani negotiated a participation agreement between Saudi Arabia and Aramco. It provided for an immediate 25% participation share, rising to 51% by 1983. Aramco had finally agreed to participation with Saudi Arabia because the alternative was worse—outright nationalization.

In June 1974, Saudi Arabia, operating on Yamani's principle of participation, took a 60% share in Aramco. By the end of the year, the Saudis told Aramco that 60% was simply not enough. They wanted 100%. An agreement to that effect was eventually reached in 1976 between Aramco and Saudi Arabia, almost 43 years after the concession was granted. By then, the proven reserves of Saudi Arabia were estimated at 149 billion barrels—more than one-quarter of the world's total reserves.

But the agreement did not by any means provide for a severing of links. Thus, under the new arrangement, Saudi Arabia would take over ownership of all Aramco's assets and rights within the country. Aramco could continue to be the operator and provide services to Saudi Arabia, for which it would receive 21 cents per barrel. In return, it would market 80% of Saudi production. In 1980, Saudi Arabia finally paid compensation, based on net book value, for all Aramco's holdings within the kingdom. With that, the sun finally set on the great concessions.

The Saudis had already worked out the embargo in some detail. They insisted that on top of the 10% cutback, Aramco must subtract all shipments to the United States, including the military. The Saudis asked Aramco for details of all crude oil used to supply American military forces throughout the world. The details were provided and the Saudis duly instructed Aramco to stop the supplies to the U.S. military. The situation was serious enough for Washington to ask whether British Petroleum (BP) could supply the U.S. Sixth Fleet in the Mediterranean.

At the beginning of November 1973, only 2 weeks after the initial decision to use the oil weapon, the Arab oil ministers decided to increase the size of the across-the-board cuts. This resulted in a gross loss of 5 million barrels/day of supply from the market. This time, however, there was no spare capacity in the United States. Without it, the United States had lost its critical ability to influence the world oil market. And with the price of a barrel of oil skyrocketing, the oil exporters could cut back on volumes and still increase their total income.

The panic and shortage of oil supplies caused by the embargo led to a quadrupling of crude oil price and precipitated a severe recession, which adversely affected the economies of the industrialized nations. Panic buying meant extra demand in the market. The bidding propelled prices even further upward. The posted price for Iranian oil, in accordance with the October 16 agreement, was $4.50 per barrel. In December, it sold for $22.60.

The oil crisis had far-reaching political and economic effects. The quadrupling of prices by the Arab oil embargo and the exporters' assumption of complete control in setting those prices brought massive changes to the world economy. The combined oil earnings of the oil exporters rose from $23 billion in 1972 to $140 billion by 1977. For the industrial countries, the sudden hike in oil prices brought profound dislocations. The oil rents flooding into the treasuries of the exporters added to a huge withdrawal of their purchasing power and sent them into deep recession. The U.S. gross national product (GNP) plunged 6% between 1973 and 1975 and unemployment doubled to 9%. Japan's GNP declined in 1974 for the first time since the end of World War II. At the same time, the price increases delivered a powerful inflationary shock to economies in which inflationary forces had already taken hold. President Nixon later commented that: "The oil embargo made it transparently clear that the economies of Western Europe and Japan could be devastated almost as completely by an oil cutoff as they could be by a nuclear attack."

On March 18, 1974, the Arab oil ministers agreed to end the embargo after the United States warned that peace efforts between the Arabs and Israel could not proceed without the lifting of the embargo. After two decades of talk and several failed attempts, the oil weapon had finally been successfully used, with an impact not merely convincing, but overwhelming, and far greater than even its proponents have dared to expect. It had transformed world oil and the relations between producers and consumers and it had remade the international economy. Now it could be resheathed. But the threat would remain.

5. THE SECOND OIL CRISIS

The Iranian revolution was at the heart of the second oil crisis. The astronomical oil price rises of 1979

and the emergence of the Rotterdam "Spot Market" were a direct consequence of the Iranian revolution. In January 7, 1978, a Tehran newspaper published a savage attack on an implacable opponent of the Shah, an elderly Shiite cleric named Ayatollah Ruhollah Khomeini, who was then living in exile in Iraq. This journalistic assault on Khomeini set off riots in the holy city of Qom, which remained his spiritual home. Troops were called in and demonstrators were killed. The disturbance in Qom ignited riots and demonstrations across the country, with further dramatic clashes and more people killed. Strikes immobilized the economy and the government and demonstrations and riots went on unchecked.

All through 1978, the internal political strife against the Shah's regime and the political drama that was unfolding simultaneously in Paris and Tehran were pushing Iran toward an explosion.

It had become evident in the mid-1970s that Iran simply could not absorb the vast increase in oil revenues that was flooding the country. The petrodollars, misspent on grandiose modernization programs or lost to waste and corruption, were generating economic chaos and social and political unrest throughout the nation. Iranians from every sector of national life were losing patience with the Shah's regime and the rush to modernization. Grasping for some certitude in the melee, they increasingly heeded the call of traditional Islam and of an ever more fervent fundamentalism. The beneficiary was Ayatollah Khomeini, whose religious rectitude and unyielding resistance made him the embodiment of opposition to the Shah's regime. The Iranian oil industry was in a state of escalating chaos. The impact of the strikes was felt immediately. On October 13, 1978, workers at the world's largest oil refinery, in Abadan, suddenly went on strike at the instigation of the exiled Ayatollah Khomeini, who was inciting them from Paris in recorded speeches on cassettes smuggled to Iran. Within a week, the strike had spread throughout most of the oil installations and Iran was, for all intents and purposes, out of the oil business.

Iran was the second-largest exporter of oil after Saudi Arabia. Of the 5.7 million barrels/day produced in Iran, approximately 4.5 million barrels/day were exported. By early November 1978, production decreased from 5.7 million barrels/day to 700,000 barrels/day. And by December 25, Iranian oil exports ceased altogether. That would prove to be a pivotal event in the world oil market. Spot prices in Europe surged 10 to 20% above official prices. The ceasing of Iranian exports came at a time when, in the international market, the winter demand surge was beginning. Oil companies, responding to the earlier general softness in the market, had been letting their inventories fall.

On December 26, Khomeini declared, "as long as the Shah has not left the country there will be no oil exports." On January 16, 1979, the Shah left the country. And on February 1, Khomeini landed in Tehran.

By the time the Khomeini regime decided to resume pumping, they could not restore production to prerevolutionary levels because the Ayatollah kicked out of Iran all the Western companies that operated the Iranian oilfields. During the production disruption, these fields lost gas pressure and the lack of maintenance made matters worse. The Iranians apparently did not have enough technical know-how to maintain or operate the fields. In effect, the Iranians pulled the rug from under the feet of the oil market, the world panicked, and the prices started to hit the roof.

Up until September 1977, there was actually a glut of oil in the market, which meant that the continuous rise in oil prices had come to a halt. The oil surplus was due to a number of factors. North Sea oil production increased much faster than expected, Mexico became a large oil exporter, and in July 1977 oil from Alaska started its long-awaited flow to the United States. At the same time, Saudi Arabia, the world's largest oil exporter, was pumping 8.5 million barrels/day in the first half of 1977 in an attempt to frustrate demands for high price increases by some members of OPEC. Another reason is that demand for crude oil in Europe barely increased from the 1976 levels, as the economies there recovered more slowly than expected from the worldwide recession that followed the Arab oil embargo 3 years earlier. Also, the demand for gasoline in the United States, expected to rise 3% in 1977, grew only 2% due to the introduction of new fuel-efficient American cars.

However, between 1974 and 1978, the world oil supplies seemed to have reached some kind of uneasy stability. Before Iranian oil production ceased, the world oil output approximately matched demand at 63 million barrels/day. By December 26,1978, when all oil exports from Iran ceased, the world oil market had a shortfall of 5 million barrels/day. Part of the shortfall had been made up by other OPEC members. This left a shortfall of 2 million barrels/day.

Yet when measured against world demand of 63 million barrels/day, the shortage was no more than 3%. Why should a 3% loss of supplies have resulted in a 150% increase in the price? The answer was

panic. The rush to build inventories by oil companies resulted in an additional 3 million barrels/day of demand above actual consumption. When added to the 2 million barrels/day of net lost supplies, the outcome was a total shortfall of 5 million barrels/day, which was equivalent to approximately 8% of global consumption. The panic buying more than doubled the actual shortage and further fueled the panic. That drove the price from $13/barrel to $34/barrel.

As the world oil shortage became more serious, there was frenzied activity on the Rotterdam Spot Market, which rapidly became the barometer of the crisis and also an indicator of the extreme price levels, and this market became the new frontier of the oil trade. Many factors conspired in early 1979 to make the Spot Market more excitable. The New Iranian regime decided to sell more and more oil on a day-to-day basis instead of on term contracts, so that in effect became part of the Spot Market. The Japanese had been dependent on Iranian oil and as the crisis deepened, they came unstuck. It was Japan that led the panic buying and by May 1979 spot crude leapt up to $34.5/barrel. The Spot Market became still more excited as the Europeans tried to replenish their declining oil stocks. American oil companies with refined oil in the Caribbean began shipping it to Rotterdam instead of the east coast of the United States.

Throughout the cold spring of 1979, the oil shortage was being felt across the industrial world and by the week beginning May 14, the Spot Market began to go crazy. Oil deals by the major oil companies at the Spot Market helped push the price further up. It became evident that the major oil companies previously sold cargos of their cheap oil bought out at the long-term contract prices, making huge profits. They were now rushing in to buy oil for storage because they were not sure how long the Iranian oil exports cutoff would last. The next big move in the Spot Market came in October 1979, when the spot prices hit $38/barrel. Then, in early November, 90 people, including 63 Americans, were taken hostage at the U.S. Embassy in Tehran. An international crisis with grave military overtones suddenly erupted onto the world scene. Spot crude hit $40/barrel.

However, the panic buying of 1979–1980 would become the glut of 1980–1986. Prices would eventually tumble. By the summer of 1980, oil inventories were very high; a pronounced economic recession was already emerging; in the consuming countries, both product prices and demand were even falling; and the inventory surplus continuing to swell. Now it was the buyers' turn to walk away from contracts and the demand for OPEC oil was decreasing. Indeed, in mid-September, a number of OPEC countries agreed to voluntarily cut back production by 10% in an effort to firm prices.

But as the world oil market was becoming calmer, on September 22, 1980 Iraqi troops began an attack on Iran. The outbreak of war threw the oil supply system into jeopardy, threatening a third oil crisis. In its initial stages, the Iran–Iraq War abruptly removed almost 4 million barrels/day of oil from the world market—6% of world demand. Spot prices jumped up again. Arab light crude oil reached its highest price ever: $42/barrel. Panic was once again driving the market. However, supply from other sources was making up for the lost output from Iran and Iraq. Within days of the war, OPEC producers raised their production. At the same time, production in Mexico, the North Sea, Alaska, and other non-OPEC countries was continuing to increase as well. Non-OPEC producers, anxious to increase market share, were making significant cuts in their official prices. As a result, OPEC's output in 1981 was 26% lower than the 1979 output and in fact was the lowest it had been since 1970.

In retrospect, one important question presents itself. Could the astronomical rise in the oil prices have been held in check? Sheikh Yamani seems to think so. He expressed the view that if the United States government and the governments of other major consuming nations intervened at the time and forbade the oil companies from trading in the Spot Market, the prices could have been checked and the panic would have ended.

6. ANATOMY OF OIL CRISES

The world witnessed two oil crises during the last 30 years of the 20th century. The topic will return to the front pages soon, however, because the world may be confronted by a third oil crisis. This third crisis promises to be similar to but have a more modest economic impact than the two previous crises, unless the price of oil hits the $50/barrel mark.

6.1 Definition of Oil Crises

For economic purposes, an oil crisis is defined here as an increase in oil prices large enough to cause a worldwide recession or a significant reduction in global real gross domestic product (GDP) below projected rates by two to three percentage points.

The 1973 and 1979 episodes both qualify as oil crises by this definition. The 1973 oil crisis caused a decline in GDP of 4.7% in the United States, 2.5% in Europe, and 7% in Japan. According to the U.S. government, the 1979 increase in oil prices caused world GDP to drop by 3% from the trend.

The price increase following the first oil crisis raised consumer payments for oil by approximately $473 billion (in real 1999 U.S. dollars), whereas the second oil crisis increased consumer expenditure by $1048 billion. By contrast, the oil price hikes during 1999–2001 raised consumer expenditure by $480 billion.

6.2 Characteristics of Oil Crises

The 1973 and 1979 crises shared four characteristics. First, the disruption in oil supplies occurred at a time of rapid expansion in the global economy. The rapid economic growth fueled greater consumption of oil. In the 5 years that preceded the 1973 crisis, global oil consumption had grown from 38.3 million barrels/day in 1968 to 52.7 million barrels/day in 1972, an average annual increase of 7.5%. Similarly, in the 5 years preceding the 1979 crisis, global consumption had risen from 53 million barrels/day in 1974 to 63 million barrels/day in 1978, an average annual increase of 3.8%.

Second, both disruptions occurred at a time when the world crude oil production was operating at virtually full capacity. Global capacity utilization reached 99% in 1973, with OPEC accounting for 56% of total production. The second oil crisis had seen a deficit of 5 million barrels/day resulting from the disruption in Iranian oil production. Third, each crisis took place at a time when global investment in oil exploration had been declining, making it impossible to achieve a speedy increase in non-OPEC production. In both 1973 and 1979–1980, the global oil industry was at the end, rather than the start, of a new surge in non-OPEC output. Fourth, in both crises, OPEC members had made a deliberate decision to reduce oil production in order to achieve political ends.

6.3 Underlying Factors

In a tight oil market, any of the following underlying factors could, individually or collectively, trigger a price escalation reminiscent of the spot market prices of 1979 and precipitate an oil crisis.

6.3.1 The Global Oil Demand

Economic growth and population growth are the most important drivers behind increasing global energy demand. The global economy is projected to grow by 3.2% per annum, on average, to 2025. Global GDP is projected to rise from $49 trillion in 2000 (year 2000 dollars purchasing power parity) to $108 trillion in 2025 and $196 trillion in 2050.

World population is expected to grow from 6 billion in 2000 to 8 billion in 2020. The population growth among the 4.8 billion people living in developing countries is estimated at 1.7% per annum. This compares with an average 0.3% per annum in the developed countries. Expanding industrialization and improving standards of living will contribute significantly to the growing energy demand.

The developed countries produce approximately one-third of global oil but consume two-thirds, whereas the developing countries produce two-thirds but consume only one-third (see Table I). Annual per capita consumption in the developing countries is 2 barrels/year. This compares with 14.2 barrels/year in the developed countries and 25 barrels/year in the United States.

The International Energy Agency (IEA) and the U.S. Department of Energy forecast that the world oil demand will grow from 76.1 million barrels/day in 2001 to 95.8 million barrels/day in 2010 and 115 million barrels/day in 2020, with Middle East producers having to meet the major part of the additional demand. However, this will depend on sufficient investment to expand production capacity (see Table II).

6.3.2 The Global Sustainable Productive Capacity

Sustainable productive capacity is here defined as "being attainable within thirty days and sustainable

TABLE I

World Crude Oil Production versus Demand in 2001

Region	Production (billion barrels)	Share (%)	Demand (billion barrels)	Share (%)
Developed countries	8	30	18	64
Developing countries	19	70	10	36
World	27	100	28	100
United States	2	7	7	25

Sources: British Petroleum Statistical Review of World Energy, June 2002; and U.S. Energy Information Administration, June 2001.

TABLE II

World Oil Demand and Supply (Million Barrels/Day), 2000–2020

	2000	2001	2005	2010	2020
World demand	76.2	76.1	83.5	95.8	114.7
World supply					
Non-OPEC	45.2	46.1	44.7	43.6	49.6
OPEC	29.3	30.0	36.0	45.9	51.1
Stock change	−1.7	—	—	—	—
Synfuels	1.2[a]	1.3[a]	1.8[a]	2.7[a]	4.2[a]
Total supply	76.2	76.1	80.7	89.5	100.7
Global oil deficit	—	—	−2.8	−6.3	−14.0

Sources: U.S. Department of Energy; British Petroleum Statistical Review of World Energy, July 2002; and International Energy Agency.

[a] Synfuel oil production is already included in non-OPEC supply figures.

TABLE III

OPEC Sustainable Capacity and Capacity Utilization (Million Barrels/Day) in January 2001

Country	Capacity	Production	Capacity utilization
Algeria	0.880	0.860	98%
Indonesia	1.300	1.300	100%
Iran	3.500	3.500	100%
Iraq	3.000	2.900	97%
Kuwait	2.200	2.200	100%
Libya	1.450	1.450	100%
Nigeria	2.100	2.100	100%
Qatar	0.720	0.700	97%
Saudi Arabia	9.250	9.000	97%
United Arab Emirates	2.400	2.400	100%
Venezuela	2.900	2.900	100%
Total	29.700	29.310	99%

Sources. Energy Intelligence Group's "Oil Market Intelligence"; Petroleum Review, April 2000; International Energy Agency; and Author's projections.

for three months." OPEC's sustainable productive capacity stood at 29.7 million barrels/day in January 2001 with a 99% capacity utilization (see Table III). The organization's immediately available spare capacity stood then at 390,000 barrels/day. However, with the production cutbacks since March 2001, spare capacity has risen to 4 million barrels/day.

The capital costs of maintaining and expanding OPEC's capacity over a 5-year period are estimated at $112 billion, money that members do not have. These projected costs are based on the member countries' planned capacity increase of 7 to 36.7 million barrels/day.

There is no non-OPEC spare capacity. The financial incentive of high oil prices and firm demand mean that every non-OPEC oilfield is being exploited and any capacity brought onstream is being utilized as quickly as possible.

6.3.3 The Ultimate Global Proven Reserves

World ultimate conventional oil reserves are estimated at 2000 billion barrels. This is the amount of production that would have been produced when production eventually ceases. Different countries are at different stages of their reserve depletion curves. Some, such as the United States, are past their midpoint and in terminal decline, whereas others are close to midpoint, such as the United Kingdom and Norway. The U.K. sector of the North Sea is currently at peak production and is set to decline at approximately 6% per year. However, the five major Gulf producers—Saudi Arabia, Iraq, Iran, Kuwait, and United Arab Emirates—are at an early stage of depletion and can exert a "swing" role, making up the difference between world demand and what others can supply. They can do this only until they themselves reach their midpoint of depletion, probably by 2013.

The expert consensus is that the world's midpoint of reserve depletion will be reached when 1000 billion barrels of oil have been produced—that is to say, half the ultimate reserves of 2000 billion barrels. With 935 billion barrels already produced, this will occur probably between 2004 and 2005. The yet-to-find (YTF) oil reserves are estimated at 280 billion barrels (see Table IV). As the world production peak approaches, the oil price will soar.

However, if the potential of unconventional oil, such as tar sand oil and extra heavy oil, is included, amounting to 572 billion barrels, then the midpoint of depletion could be delayed for a few more years—but not beyond 2010.

In 1956, the geologist M. King Hubbert predicted that U.S. oil production would peak in the early 1970s. Almost everyone, inside and outside the oil industry, rejected Hubbert's analysis. The controversy raged until 1970, when U.S. production of crude oil started to fall. U.S. production peaked at 9.64 million barrels/day in 1970 and has been falling since then, reaching 5.77 million barrels/day by 2001. Hubbert was proven right and his bell-shaped curve became a useful tool of oil production analysis.

TABLE IV

Ultimate Global Reserves of Conventional Oil and Depletion Rate (End of 2001)

	Volume or rate	Description
Ultimate reserves (billion barrels)	2000	Amount of production when production ceases
Produced so far (billion barrels)	935	Until the end of 2001
Yet-to-produce (billion barrels)	1065	Ultimate reserves less produced
Discovered so far (billion barrels)	1720	Produced plus remaining reserves
Yet-to-find (billion barrels/year)	280	Ultimate reserves less discovered
Discovery rate (billion barrels/year)	7	Annual additions from new fields
Depletion rate (%)	3	Annual production a percentage of yet-to-produce

Sources. United States Geological Survey, and British petroleum Statiscal Review of World Energy, June 2002.

Around 1995, several analysts began applying Hubbert's method to world oil production. Based on Hubbert's pioneering work and an estimated 1.8–2.1 trillion barrels of ultimate reserves, they established that the peak production year will be sometime between 2004 and 2009. If the predictions are correct, there will be a huge impact on the global economy, with the industrialized nations bidding against one another for the dwindling oil supply. One promising oil province that remains unexplored is the Spratly Islands in the South China Sea, where exploration has been delayed by conflicting claims to the islands by six different countries. Potential reserves in the disputed territories are estimated at multibillion barrels of oil and gas. But even if the South China Sea oil reserves are proven, they could hardly quench China's thirst for oil. By 2010, China is projected to overtake Japan to become the world's second largest oil importer after the United States.

Another promising province is the Caspian Basin. Estimates of 40 to 60 billion barrels as the ultimate reserve base of the Caspian region are judged to be reasonable by most geologists familiar with the region. Apart from the limited size of the reserves, the area's oil is very costly to find, develop, produce, and transport to world markets. Projected Caspian Sea oil production of 2–3 million barrels/day by 2010 can be achieved only when prices exceed $20/barrel (in real terms). Oil prices will be the key factor in the expansion of Caspian Sea oil.

6.3.4 New Oil Discovery Rates

The widely held view that improved seismic surveying and interpretation have improved drilling success rates is not borne out by the level of discoveries during the period 1992–2001 (see Table V).

TABLE V

Global Crude Oil Reserves Additions[a] (Billion Barrels) 1992–2001

Year	Added in year	Annual production	As % of annual production
1992	7.80	23.98	33
1993	4.00	24.09	17
1994	6.95	24.42	28
1995	5.62	24.77	23
1996	5.24	25.42	21
1997	5.92	26.22	23
1998	7.60	26.75	28
1999	13.00	26.22	50
2000	12.60	27.19	46
2001	8.90	27.81	32
1992–2001	77.63	256.87	30
Average	7.76	25.83	30

Sources. IHS Energy Group's 2002 World Petroleum Trends Report, and British Petroleum Statistical Review of World Energy, 1993–2002.

[a]Excluding the United States and Canada.

The race for reserves is, therefore, on. With the demand for oil envisaged for the next 10 years, the world will consume an average 30 billion barrels per year over that period. If the global oil industry wants to replace this consumption with new reserves without diluting the world's existing proven reserves of some 1 trillion barrels, it must find an additional 300 billion barrels of new oil in the next decade—a daunting challenge indeed.

6.3.5 Global Reserve Depletion Rate

Globally the reserve depletion rate is generally calculated at 3%. This means that to sustain the

world's current 76 million barrels/day consumption at that level, approximately 4 million barrels/day of new capacity is needed every year. Against this, the world production capacity has remained static or even declined while consumption has been increasing.

This means that the Middle East producers, with 65% of the world's proven reserves and just one-third of global production, will assume clear-cut leadership of the supply side of the oil market.

7. THIRD OIL CRISIS?

The parallels between current conditions and the early 1970s are unnerving. The conditions that made the 1973 and 1979 oil crises possible exist in the early 2000s. First, world oil consumption is growing rapidly and is projected to continue expanding. World oil demand grew by 2.4% in 2000, 2.2% in 2001 with a projected additional annual increase of 1.8% until 2005.

Second, every indicator points to the fact that production from OPEC will not rise substantially. OPEC's maximum production capacity is estimated at 29.7 million barrels/day. The organization must produce approximately 30 million barrels/day to maintain a supply–demand balance under the current level of stocks. Yet, an output level of 30 million barrels/day may be beyond the organization's reach or, if it is feasible, achievable only with production from Iraq.

Third, the pattern of low-level industry investment in 2000–2001 is also very similar to that observed during the first and second oil crises. Oil companies' capital investment in exploration in 2000/2001 was slow in the aftermath of low oil prices during 1998–1999. Capital spending on exploration and production by the supergiants—Exxon Mobil, BP Amoco, and Shell—fell 20% to $6.91 billion in the first half of 2000 from a year earlier. Most companies decided to wait before boosting investment, preferring instead to buy back stocks. This has reduced the capital investment available for production and production capacity expansion.

Fourth, as in 1973 and 1979, the causes of this crisis will be a reduction in OPEC production and stagnation in non-OPEC's taking place at a time of rapid economic growth.

Fifth, as in 1973 and 1979, it will be a major political event that could send oil prices rocketing and thus precipitate a third oil crisis. The 1973 Arab–Israeli war was behind the first oil crisis, whereas the Iranian revolution was at the heart of the second. Similarly, a war against Iraq could precipitate a third oil crisis.

Such conditions make for a truly tight market with the potential for developing into a genuine crisis. But even if the world manages to escape a third oil crisis, oil-market turmoil can be expected to last until 2004.

However, a growing number of opinions among energy experts suggest that global conventional oil production will probably peak sometime during this decade, between 2004 and 2010. Declining oil production will cause a global energy gap, which will have to be filled by unconventional and renewable energy sources.

Various projections of global ultimate conventional oil reserves and peak years have been suggested by energy experts and researchers between 1969 and 2002. The extreme end of opinion is represented by the United States Geological Survey (USGS), IEA, and the U.S. Energy Information Administration (EIA) (see Table VI).

The estimate by EIA is so implausibly high that it can be ignored, whereas the USGS estimate includes 724 billion barrels of YTF reserves. Such huge YTF reserves require discovering an additional amount of oil equivalent to the entire Middle East. But since 90% of global conventional oil has already been found, trying to find 724 billion barrels of new oil is not only an exceptionally daunting task, but virtually impossible.

However, such estimates are of only limited relevance. What is important when attempting to identify future supplies are two key factors: the discovery rate and the development rate, and their relationship to the production rate.

The technology for extracting oil from tar sands, oil shale, and extra heavy oil, known collectively as synfuels, exists but extraction costs are high. Synfuel oil is usually 3 times as labor-to-energy intensive and 10 times as capital-to-energy intensive as conventional oil.

Whereas some—and possibly a great deal—of unconventional oil (synfuels) will eventually be available, there will not be enough to replace the shortfalls in conventional oil. Synfuels will be hard-pressed to meet 3% of the global oil demand in 2010 and 4% in 2020, because of the slow extraction rate and the huge investments needed.

In 2002, only 35,000 barrels/day of natural gas to liquid oil is produced worldwide. This is projected to rise to 685,000 barrels/day by 2010, equivalent to 0.7% of global demand. The constraint, however, might be the very large capital commitment. For a

TABLE VI

Various Projections of Global Ultimate Conventional Oil Reserves and Peak Year (Billion Barrels)

Author	Affiliation	Year	Estimated ultimate reserves	Peak year
Hubert	Shell	1969	2100	2000
Bookout	Shell	1989	2000	2010
Mackenzie	Researcher	1996	2600	2007–2019
Appleby	BP	1996		2010
Invanhoe	Consultant	1996		2010
Edwards	University of Colorado	1997	2836	2020[a]
Campbell	Consultant	1997	1800–2000	2010
Bernaby	ENI	1998		2005
Schollenberger	Amoco	1998		2015–2035[a]
IEA	OECD	1998	2800	2010–2020[a]
EIA	DOE	1998	4700	2030[a]
Laherrere	Consultant	1999	2700	2010[a]
USGS	International Department	2000	3270	—
Salameh	Consultant	2000	2000	2004–2005
Deffeyes	Princeton University	2001	1800–2100	2004

Sources. Various.

Note. BP, British Petroleum; DOE, U.S. Department of Energy; EIA, U.S. Energy Information Administration; IEA, International Energy Agency; OECD, Organisation for Economic Cooperation and Development; USGS, U.S. Geological Survey; ENI, Ente Nazionale Idrocarburi.

[a] These ultimate reserve estimates include extra heavy crude, tar sands, oil shale, and also projected production of gas-to-liquid oil.

production of 400,000 barrels/day, this would amount to $8 billion.

In 2000, renewable energy sources contributed 2% to the global primary energy demand. However, by 2025 they are projected to contribute 4%, rising to 5% by 2050 (see Table VII).

Fuel-cell motor technology will eventually have a great impact on the global consumption of gasoline and diesel. But it could take years before hydrogen-powered cars dominate the highways and certainly not before they are able to compete with today's cars in terms of range, convenience, and affordability.

Fossil fuels, with a growing contribution from nuclear energy, will, therefore, still be supplying the main share of the global energy needs for most—perhaps all, of the 21st century.

TABLE VII

World Primary Energy Consumption[a] (mtoe), 2000–2050

	2000	2025	2050
Primary energy	9631	16,618	19,760
Oil	3835	6429	7344
Natural gas	2190	4760	6207
Coal	2136	3283	3037
Nuclear	636	733	825
Hydro	617	818	1440
Renewables[b]	217	595	907
Renewables: % of total	2	4	5

Sources. Shell International, Scenarios of 2050; Organization of Petroleum Exporting Countries Statistical Review of World Energy, June 2002; and U.S. Energy Information Administration International Energy Outlook 2002.

[a] mtoe = Million tonnes oil equivalent.

[b] Excluding hydro (i.e., hydroelectric power).

8. IMPLICATIONS FOR THE GLOBAL ECONOMY

Twenty years ago, oil crises wreaked havoc on the world economy. Today, although oil remains important, a new crisis will have a much more modest impact because of the diminished role of oil in the global economy, unless the price of oil rises to $50/barrel.

One reason is that oil represents a much smaller share of global GDP today than it did during the two previous crises. Another reason is that the economy is more open, which makes it harder for companies to pass on cost increases to customers. A third reason is that customers can hedge against price increases.

Finally, economies around the world have become more adaptable in their use of every resource, including oil. Consequently, increases in oil prices trigger substitutions of natural gas or any other energy sources in manufacturing. For instance, if one takes the case of the United States, which accounts for 25% of the global oil consumption, it is found that a rise in oil prices today is much less important to the U.S. economy than it has been in the past. The decline in importance of oil prices can be seen from Table VIII, which shows the share of crude oil in nominal GDP.

Table VIII shows that crude oil accounted for 4% of GDP at the time of the first oil crisis, 6% at the time of the second oil crisis, and 2.4% at the time of the Gulf War. In 2001, it accounted for 2%. However, if the price of oil hits the $50/barrel mark in 2002, then crude oil will have the same share in the U.S. economy as in 1974, with adverse economic implications for the U.S. and the global economy.

Despite the above, higher oil prices still are very important to the world economy. A new report by the International Monetary Fund (IMF) notes that a war on Iraq could be the final straw for an international economy still struggling with the aftermath of September 11, the 20% drop in world share prices this year, and the implosion of the Argentine and Brazilian financial services. The IMF argues that the risks are "primarily on the downside" even before the possibility of a conflagration in Iraq and the Middle East is taken into account.

The IMF warns of the consequences for the oil markets of a war in the Gulf region. It says that a $15/barrel increase in the price of oil would be a severe blow to the global economy, knocking at least 1 percentage point off global GDP and sending the world economy spiraling toward recession.

Certainly, all the evidence from past conflicts in the region, including the 1973 Arab–Israeli War, Iranian revolution, and the Gulf War of 1990–1991, suggests that each of these conflicts has been followed by recession and unemployment, particularly in the West.

This time around, the scenario is even more worrying. Financial markets are already paralyzed as a result of the burst of the technology bubble and the mammoth corporate accounting frauds in the United States.

Before Washington's decision to up the ante with Iraq, there was some hope at the IMF that there would be some kind of bounce-back for the global economy in 2003. The world economy was projected to pick up speed, growing by 3.7% in 2003, after slipping this year to 2.4%. The United States was forecast to pick up from a 1.4% expansion this year to 2.3% in 2003.

But higher oil prices could make all the difference. The IMF notes that oil prices began to surge in the second half of August 2002 as the U.S. administration began to talk up the war with Iraq. The price of crude oil on both sides of the Atlantic is already above the $22 to $28 per barrel price range built into most major economic projections. For oil-importing countries, higher oil prices act as an immediate tax on consumption and business. An oil price hike makes it less likely that consumers will spend and business will invest.

With the global economy struggling to pick up momentum, a prolonged period of high oil prices would, without doubt, delay an upturn. It would almost certainly give another leg to the downturn in share prices—in 2003, down by 40% since the peak in March 2000—hitting confidence hard.

Although the loss of Iraqi oil supplies over the short term might make little significant difference, it is the collateral damage that could deliver the most serious blow. There is the possibility that Iraq might seek to disable the critical Saudi and Kuwaiti oilfields, as in the Gulf War, or that the Arab countries—outraged by America's attack on one of their own—might seek to impose an oil embargo or refuse to increase supplies to make up for the loss of Iraqi output.

The result would be a hike in prices on world oil markets and disruption to an already misfiring global economy. However, once the war is over, there may well be cheaper and more plentiful oil as Iraq's

TABLE VIII

Oil in the U.S. Economy: Consumption, Prices, and Percentage of GDP

	Oil consumption (million barrels/day)	Oil prices ($/barrel)	Nominal GDP ($ billion)	Oil consumption as percentage of GDP
1974	16.65	9.07	1382	4.0
1980	17.06	28.07	2784	6.3
1990	16.99	22.22	5744	2.4
1999	19.36	17.60	8857	1.4
2000	19.63	27.72	9224	2.2
2001	19.90	25.93	9462	2.0
2002[a]	20.18	30.00	9611	2.3
2002[a]	20.18	50.00	9611	4.0

Sources. Courtesy of PKVerleger LLC; British Petroleum Statistical Review of World Energy, June 2002; U.S. Energy Information Administration Annual Energy Outlook 2002.

[a] Estimates.

production is restored. But by then the damage to a rickety world economy will have been inflicted.

SEE ALSO THE FOLLOWING ARTICLES

Geopolitics of Energy • Markets for Petroleum • Nationalism and Oil • National Security and Energy • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Oil Industry, History of • Oil-Led Development: Social, Political, and Economic Consequences • Oil Price Volatility • OPEC, History of • OPEC Market Behavior, 1973–2003 • Strategic Petroleum Reserves • War and Energy

Further Reading

Deffeys, K. S. (2001). "Hubbert's Peak: The Impending World Oil Shortage." Princeton University Press, Princeton, NJ.

Nixon, R. (1980). "The Real War." Sidgwick & Jackson, London.

Robinson, J. (1988). "Yamani: the inside story." Simon & Schuster, London.

Salameh, M. G. (1990). "Is a Third Oil Crisis Inevitable." Biddles, Guildford, UK.

Salameh, M. G. (1999). Technology, oil reserve depletion and the myth of reserve-to-production (R/P) ratio. OPEC Review, June 1999, pp. 113–124.

Salameh, M. G. (2001). Anatomy of an impending third oil crisis. *In* Proceedings of the 24th IAEE International Conference, April 25–27, 2001, Houston, TX, pp. 1–11.

Salameh, M. G. (2001). "The Quest for Middle East Oil: the U S Versus the Asia-Pacific Region." International Asian Energy Conference, August, Hong Kong.

Sampson, A. (1980). "The Seven Sisters." Hodder & Stoughton, London.

Verleger, P. K. (2000). "Third Oil Shock: Real or Imaginary?" Consequences and Policy Alternatives. International Economics Policy Briefs, No. 00-4, Institute for International Economics, Washington, DC.

Yergin, D. (1991). "The Prize: The Epic Quest for Oil, Money and Power." Simon & Schuster, New York.

Oil Industry, History of

AUGUST W. GIEBELHAUS
Georgia Institute of Technology
Atlanta, Georgia, United States

1. **Petroleum Prior to the Modern Era**
2. **Oil's Age of Illumination**
3. **Competition and Monopoly**
4. **Technological Change and Energy Transition**
5. **Abundance, Scarcity, and Conservation**
6. **The Globalization of the Industry**
7. **International Competition and War**
8. **The Middle East, OPEC, and the Energy Crisis**
9. **Oil in the Modern Economy**

Glossary

barrel The international standard of measure for crude oil and oil products equivalent to 42 U.S. gallons and used since first defined in the U.S. oil fields in the 1860s.

cracking Various refining processes using heat, pressure, and catalysts to change less volatile and heavier petroleum fractions into compounds with lower boiling points.

crude oil Unrefined petroleum as it comes from the well, consisting almost entirely of carbon and hydrogen compounds and varying widely in appearance, color, odor, and the presence of sulfur, nitrogen, oxygen compounds, and ash.

illuminating oil Common 19th-century name for kerosene, the fraction of crude oil between gas oil and gasoline on the refiner's scale, usually ranging between 105 and 300°F; represented the first major commercial petroleum product and defined the early decades of the industry.

law of capture Legal principle emanating from the courts in the 19th century that encouraged the rapid drilling and pumping of oil from underground pools; a major cause of wasteful practice in the early oil industry.

Organization of Petroleum Exporting Countries (OPEC) Organized in 1960 to establish a united front against the power of the world's multinational oil companies, it became a force to be reckoned with in the 1970s and remains a relevant player in the economics of world oil.

refining Processes employed to transform crude oil into useful commercial products; includes fractional distillation, cracking, purifying, and treating.

rotary drilling Method for drilling deep holes in search of petroleum or gas in which a drill bit is attached to a revolving drill pipe; used first in the early 20th century.

Spindletop Enormous oil strike or "gusher" on the Texas Gulf Coast near Beaumont in 1901 that changed the face of the oil industry. Economic forces unleashed brought a market erosion of Standard Oil dominance and the birth of oil's new "age of energy" with the increased production of fuel oil and gasoline.

Standard Oil Parent firm of the John D. Rockefeller oil interests, initially capitalized as an Ohio corporation, reorganized as a New Jersey holding company in 1899, and dissolved by the U.S. Supreme Court in a famous antitrust decision in 1911.

Most global citizens today view the international oil industry as an enormous industrial power whose financial profits have arisen from the utilization of one of nature's gifts, often through shrewdness and technological innovation but sometimes also as a result of monopolistic practices and exploitation of Third World resources. During the time of the price dislocations of 1973 and 1979 that defined the "energy crisis" of the decade, many pointed to domination of an oligopoly of multinational firms defined by British journalist Anthony Sampson as the "seven sisters." An aura of monopolistic competition and "big business" has surrounded this industry from its beginnings, as the Rockefeller Standard Oil empire remains imprinted as a past symbol of rampant corporate power. However, the history of this industry has been much more complex than a narrative of wealth accumulation by a few major companies. It represents a story of vision, risk, business success, and often disaster played out in an increasingly multinational arena. Oil has emerged as the dominant fuel consumed in the United States, the country in which the industry took life in the modern era, but domestic crude oil production peaked in 1970 and has declined ever since. Once the "Saudi Arabia of the world," the United States

 © 2004 Elsevier Inc. All rights reserved.

became a net importer of petroleum at the historical moment that an oil regime had come to dominate world energy. The energy crises of 1973 and 1979, the Persian Gulf War of 1991, and the invasion of Iraq in 2003 have all highlighted the geopolitical importance of Middle Eastern oil, and the major vertically integrated, multinational companies have become almost equal players with governments in this arena.

1. PETROLEUM PRIOR TO THE MODERN ERA

Energy historians have emphasized the concept of transition from one energy regime to another as an analytical tool. One model, for example, suggests that early mankind entered the first transition from hunting and gathering to the harvesting of biomass energy through agriculture and silviculture. This relatively straightforward and simple energy form dominated in antiquity, the medieval period, and well into the early modern era. The era of the industrial revolution saw the transition from biomass harvesting to the exploitation of fossil energy in the form of coal. The shift from a predominant reliance on coal to an overwhelmingly significant role for oil and natural gas has defined the last transition. Two aspects derive from this analysis. The first is to understand how very brief has been the "oil age" within the context of human history (approximately 140 years, from 1859 to the present). The other is the speculative argument that since all fossil fuels (coal, oil, and natural gas) will eventually be depleted, we must plan as a global society for the next transition to a more sustainable energy source. These arguments have most often focused on one type of "renewable" fuel or another, whether it be, in the words of energy analyst Amory Lovins, hard or soft. For example, many advocates of nuclear energy still argue that fission breeder reactors promise a potential renewable energy source, but one that many social scientists would criticize as undesirable or hard because of the great potential for environmental degradation. For others, more environmentally benign, soft solutions, such as solar, wind, tidal, or geothermal energy or renewable biomass fuels such as ethanol, provide the answer. These arguments must be left for another discussion, but all experts agree that the age of overabundant, cheap petroleum will end at some point; the problem is to determine when this will occur.

Prior to the industrial era, petroleum had a very limited utility. Records show human uses for oil in the form of natural seepage of asphaltic bitumen as far back as ancient Mesopotamia prior to 3000 BC. Employed as a mastic in construction, a waterproofing for ships, and a medicinal poultice, this oil apparently did have a small market. Later, Near Eastern cultures successfully distilled natural crude oil to obtain lamp fuel and the basic ingredient for "Greek fire," a military incendiary introduced with great effect by the Byzantine Empire against the rigging of attacking ships. In the West, limited supplies of crude oil seepage were confined mostly to medicinal use. Native Americans had also discovered the medical benefits of crude oil, and when Edwin L. Drake and associates decided to drill for oil in northwestern Pennsylvania in 1859, they were aware that local tribes had been using oil for a very long time.

A market for medicinal oil developed by the early 19th century in Pennsylvania and a number of entrepreneurs entered the business. Samuel M. Kier of Pittsburgh, the most successful of these businessmen, established a successful medicinal market for petroleum or rock oil before the Civil War. These original "snake oil salesmen" claimed a host of cures for rock oil to be used both externally and internally. In a parallel development, a small Russian industry developed around the presence of natural seepage in the Baku region, initially also for medicinal use. However, it took a 19th-century shortage of lamp oil, precipitated by a whale oil crisis and relative scarcity of coal oil, to spark the modern petroleum industry.

2. OIL'S AGE OF ILLUMINATION

In the 1850s, experiments with refined petroleum demonstrated that it could serve as a satisfactory lamp oil (in kerosene form). The problem was that it was not available in suitable quantities. The drilling of water wells in northwest Pennsylvania had uncovered the presence of oil along with brine in a number of instances. Why couldn't one drill for oil? A group of investors hired "Colonel" Edwin L. Drake to obtain leases and attempt to drill for oil using the standard percussion techniques then used for drilling for water. On August 27, 1859, Drake's rig struck oil at 69 ft near Oil Creek, a short distance from the town of Titusville, Pennsylvania. The oil age was born. The market for illuminating oil was

insatiable and the growth of the industry was sudden and dramatic. Crude oil production increased from only 2000 barrels in 1859 to 4.8 million barrels a decade later and 5,350,000 barrels in 1871. From the beginning, the demand for illuminating oil reflected both the burgeoning domestic population and economy and an expanding export business. Between the 1860s and 1900, crude oil and oil products exported ranged from one-third to three-fourths of total U.S. production.

This new industry initially operated in a classically competitive fashion. Ease of entry defined all aspects of the business, and capital investment was minimal. Refining technology was primitive, consisting of heating crude in a still in order to obtain the desired kerosene fraction; other constituent fractions such as gasoline were often run into streams or onto the ground in ditches. The evolution of the legal doctrine of law of capture by the Pennsylvania courts also contributed to wasteful practices in the oil fields. Several different landowners or leaseholders might sit astride an underground oil pool. Drawing an analogy with game being legally captured if it were lured onto another's land, and citing precedent in English common law, the courts upheld the right of the property owner to "capture" oil that had migrated under the owner's land. Limited understanding of oil geology at that time mistakenly led to the belief that oil flowed in underground rivers. Each owner or leaseholder rushed to drill and pump before their neighbors depleted the oil pool. Oil markets became characterized by alternate periods of oil glut followed quickly by scarcity. This wasteful practice continued as the oil frontier moved westward from Pennsylvania, Ohio, and Indiana in the latter 19th century to Texas, California, and the midcontinent in the 20th century.

3. COMPETITION AND MONOPOLY

The young John D. Rockefeller entered this rough-and-tumble business in 1863. He had cut his teeth as a partner in the wholesale grocery business in Cleveland, doing well as purveyor of supplies to Ohio military units. However, Cleveland possessed several locational advantages that encouraged its development as an important refining center. Located on Lake Erie, close to the Pennsylvania oil fields, it had railroad trunkline connections to major eastern markets, a readily available workforce, and sources of financial capital. The industry, then as today, consisted of four main functional sectors: production, refining, transportation, and marketing. Production entails all the activities involved in "getting it out of the ground." This includes exploration, drilling, leasing, pumping, primary and secondary recovery techniques, and all other associated activities. Refining entails all relevant processes developed to obtain usable products from crude oil in its natural state. Initially, this included only fractional distillation, the heating of crude to boil off its constituent parts, but has evolved today to include a host of sophisticated specialty practices. Transportation, the third element of the industry, encompasses the movement of crude oil from the oil field to the refinery and the shipment of refined product to market. In the beginning of the industry in the 1860s and 1870s, this meant by barge, railroad car, or horse-drawn wagon. Oil transport evolved to include petroleum pipelines, tanker trucks on the highway, and fleets of tankers on the high seas. Marketing envelops the distribution and sale of petroleum products to the consumer. 1870s marketing might have consisted of a Standard Oil horse-drawn tank wagon selling branded "illuminating oil" or kerosene in neighborhoods; today, we think of well-advertised gasoline stations and slick television commercials designed to get us to use one firm's brand over another's. Although we think of the industry today as consisting of a few vertically integrated firms (those that are engaged in all four major functions from production through marketing), the oil fraternity has always contained a strong element of independents. These individual producers, refiners, pipeline operators, and marketers often viewed themselves as opponents of the big companies such as Standard Oil and they still represent a distinct voice in the industry.

Rockefeller's early endeavors were concentrated in the refining sector of the industry, achieving growth through horizontal combination; he did not adopt a strategy of vertical integration until later. First as a partnership, then organized as a corporation in 1870, Rockefeller and associates' Standard Oil Company proceeded to achieve what was later termed the conquest of Cleveland. Using tactics that one could define as either shrewd or unethical, Rockefeller succeeded in dominating the Cleveland refining business. Concentrating on transportation as the key to control, Standard used its size to obtain preferential shipping arrangements with the railroads through the use of rebates and then sought to eliminate competition with a series of pooling and cartel agreements. Rockefeller's attempts to control

the refining business with the South Improvement Company and later the National Refiners Association remain textbook examples of business strategy in the late 19th century. When Rockefeller found informal agreements inadequate to achieve tight industry control, he changed his approach to one of merger and acquisition.

By 1878, Standard controlled more than 90% of total U.S. refining capacity. To formalize this economic empire, Rockefeller and associates created the Standard Oil Trust Agreement in 1882, later reorganized as the Standard Oil Company (New Jersey) as a legal holding company under New Jersey law in 1899. Rockefeller integrated backward in the 1880s by acquiring producing properties in Ohio, into transportation with his own pipelines, and forward into marketing with his branded cans of kerosene for domestic sale and export. The Standard near monopoly became the hated enemy of independent oil producers, who found their prices driven down by Rockefeller buying power, and by smaller operators, who had been forced to sell out in the face of price-cutting attacks. As a public outcry against the power of large Gilded Age corporations emerged, "muckraking" exposés by Henry Demerest Lloyd and Ida M. Tarbell, a daughter of an independent oilman, painted Standard Oil as a hated symbol of monopoly in America.

Standard had been under legal attack in state courts by its enemies and competitors for years, but the passage of the Sherman Antitrust Act in 1890 had lain the groundwork for federal challenge. With a change in political climate during the presidency of Theodore Roosevelt toward progressivism, federal investigations went forward, culminating in the forced dissolution of Standard Oil into several constituent companies in 1911. This did not hurt Rockefeller personally too much because he remained a stockholder in these many companies, and historians have argued that the former monopoly had simply transformed into an oligopoly of large vertically integrated companies. For example, Standard Oil (New Jersey) suffered for a time because it was a large refiner and marketer but lacked integrated production facilities. Soon, however, it rectified this situation with the acquisition of oil production in the Oklahoma and Texas fields. Similar strategies had enabled most of the spun-off firms to emerge as strong vertically integrated units by the early 1920s. However, market forces other than court decisions had already begun to change the U.S. oil industry in significant ways.

4. TECHNOLOGICAL CHANGE AND ENERGY TRANSITION

Economists Harold F. Williamson and Ralph Andreano have argued that by the eve of the 1911 dissolution decree the competitive structure of oil already had been altered, the key factor being the discovery of vast supplies of oil on the Texas Gulf Coast at Spindletop in 1901. If the Drake well in 1859 had trumpeted the birth of the illuminating oil industry, Spindletop marked the birth of its new age of energy. Spindletop and subsequent other new western fields provided vast oil for the growing economy but also enabled firms such as Gulf, Texas, and Sun Oil, independent of the Standard interests, to gain a substantial foothold. The gulf region was underdeveloped and attractive to young companies, and the state of Texas was hostile toward the Standard Oil monopoly. Asphaltic-based Spindletop crude made inferior grades of kerosene and lubricants but yielded a satisfactory fuel oil. New firms, such as Gulf and Texaco, easily integrated forward into the refining, transportation, and marketing of this fuel oil in the coal-starved Southwest. Located near tidewater on the Gulf Coast, small investments in pipelines also enabled operators to get their crude to seagoing tank ships, which could then carry it to other markets. The Sun Oil Company of Pennsylvania, a long-time competitor of Standard, built a robust business by shipping Texas crude to its refinery outside Philadelphia.

In 1900, Standard Oil (New Jersey) controlled approximately 86% of all crude oil supplies, 82% of refining capacity, and 85% of all kerosene and gasoline sold in the United States. On the eve of the 1911 court decree, Standard's control of crude production had declined to approximately 60–65% and refining capacity to 64%. Moreover, Standard's competitors now supplied approximately 70% of the fuel oil, 45% of the lubricants, 33% of gasoline and waxes, and 25% of the kerosene in the domestic market. Newly formed post-Spindletop companies such as Gulf and Texaco, along with invigorated older independent firms such as Sun and Pure, had captured significant market share. Meanwhile, Standard, heavily invested in the traditional kerosene business, was slow to move into the production and marketing of gasoline.

At approximately the same time, the automobile, which had appeared in the 1890s as a novelty, had begun to shed its elitist image. The introduction of mass-produced and relatively inexpensive vehicles, led by the 1908 Ford Model T, very quickly

influenced developments in the oil business. In 1908, the total output of the U.S. auto industry was 65,000 vehicles. In less than a decade, Ford alone sold more than 500,000 units annually. Within that same decade, both the volume of production and the total value of gasoline passed those of kerosene. By the turn of the century, the increasing electrification of urban America and widespread use of the Edison incandescent lamp were already a worry to oil companies whose major product was kerosene. Now the oil industry was becoming increasingly concerned about how it could boost production of gasoline to meet demand.

As the oil industry faced real or apparent shortages of supply, one response was technological. New production methods such as the introduction of rotary drilling early in the century increased crude supplies, and refinery innovations enabled crude stocks to be further extended. However, as the demand for gasoline increased, new oil discoveries in Texas, California, and Oklahoma in the early 1900s proved insufficient. Once discarded as relatively useless, this lighter fraction typically constituted 10–15% of a barrel of crude oil. Refiners stretched the gasoline fraction by including more of the heavier kerosene, but this resulted in an inferior motor fuel. One could enrich the blend of gasoline with the addition of what would later be termed higher octane product obtained from selected premium crudes (e.g., from California) or "natural" or "casinghead" gasoline yielded from highly saturated natural gas. The most important breakthrough, however, occurred with the introduction of thermal cracking technology in 1913 by a Ph.D. chemist employed by the Standard Oil Company (Indiana), William Burton. The industry had employed light cracking, the application of heat to distillation to literally rearrange hydrocarbon molecules, since the 1860s to obtain higher yields of kerosene from feed stock. By dramatically increasing the temperature and pressure of his cracking stills, Burton discovered that he could double the output of gasoline obtained over previous fractional distillation and cracking methods. An additional advantage was that this gasoline was of generally superior quality. Although the problem of "knocking" in the internal combustion engine was not yet understood fully, it would become a major technological challenge in the 1920s as higher compression, higher performance engines appeared.

Oil's growth as an energy source between 1900 and 1920 coincided with an enormous increase in total U.S. energy consumption. Samuel J. Schurr and Bruce Netschert calculated that total energy use increased by 123% during this span of two decades. Although oil companies in the Southwest had begun to charge that "King Coal's" reign was over, aggregate statistics show that the industry's share of total energy consumption had only risen from 2.4% of the total in 1900 to 12.3% in 1920. Coal remained dominant, with its share of total energy consumption actually slightly increasing from 71.4 to 72.5%. The actual transition of fossil fuel dominance from coal to oil would come later in the century, but the foundation had been set.

5. ABUNDANCE, SCARCITY, AND CONSERVATION

From its earliest beginnings in the 19th century, the oil industry had to deal with the recurrent feast or famine that accompanied the alternate discovery of new fields followed by their depletion. In "Special Report on the Petroleum of Pennsylvania," compiled in 1874 by geologist Henry Wrigley, it was argued that "we have reaped this fine harvest of mineral wealth in a most reckless and wasteful manner." John F. Carl of the Pennsylvania Geologic Survey concluded in 1886 that "the great Pennsylvania oil fields, which have supplied the world for years, are being exhausted, and cannot respond to the heavy drafts made upon them many years longer, unless reinforced by new deposits from deeper horizons." History soon demonstrated that many of these concerns were unfounded, as the discovery of new producing fields once again replaced scarcity and high crude prices with glut and price depression. However, the projections of future supply in the United States published by the U.S. Geologic Survey in 1908 predicted total depletion of U.S. reserves by 1927, based on then current levels of consumption, and the secretary of the interior warned President Taft in 1909 of an impending oil shortage. Driven by the rush to get oil out of the ground by the law of capture, the industry's production sector remained chaotic and highly unpredictable as the "forest of derricks" that defined each field moved ever westward. Within the context of this volatile business one can read the arguments of the Standard Oil (New Jersey) attorneys defending the company from antitrust attack in a different light. They maintained that Rockefeller's strategy of consolidation, denounced as evil monopoly, in fact represented a rational attempt to impose order and stability on an unstable industry.

A new era of conservation as well as the first significant enforcement of antitrust law had arrived on the political scene at the beginning of the 20th century with the presidential administration of Theodore Roosevelt. It is important to differentiate the conservation approach, best defined as wise utilization of natural resources, from the preservationist approach of Roosevelt contemporary John Muir, founder of the Sierra Club. Roosevelt and his key advisers, such as chief forester of the United States Gifford Pinchot, did not want to lock away America's resources but did favor planning and responsible exploitation. One should interpret the efforts of the U.S. Geological Survey and the Bureau of Mines (created in 1910) in this light. These two institutions provided some attention to petroleum-related issues, but the creation of the Bureau of Mines' petroleum division in 1914 signaled a heightened federal government interest in oil conservation. Work centered at the division's Bartlesville, Oklahoma Petroleum Experiment Station focused on reservoir behavior, efficient drilling practices, well spacing, and secondary recovery techniques and disseminated knowledge to industry through influential Bureau of Mines publications.

In the 1920s, petroleum conservationists began to encourage unitization or the unit management of oil pools as a central alternative to the wasteful law of capture. This approach encouraged each pool to be operated as a cooperative unit, with the individual leaseholder's percentage share defining the amount of oil that could be pumped out. Operators would drill fewer wells, produce oil at a controlled rate, and either use or return to the producing zone under pressure all natural gas obtained along with the oil rather than the standard practice of flaring it or venting it into the atmosphere. Unitization was supported by the scientific principle that all recoverable oil in its undisturbed state contains gas in solution and that this solution has a lower viscosity, lower specific gravity, and lower surface tension than gas-free oil. It had long been known that natural gas had the same expulsive function in the well that carbon dioxide has in a shaken soda bottle, but reduced viscosity of the oil meant more fluidity and ultimately enhanced recovery. There were champions of this approach in the early 1920s within the industry and the government, among them Henry L. Doherty of the Cities Service Company and George Otis Smith, Director of the U.S. Geologic Survey. However, Doherty failed in efforts to get the American Petroleum Institute, the powerful trade association of the industry founded after World War I, to support the practice. The industry only began to come around after discoveries of huge amounts of oil in the midcontinental United States in the mid- to late 1920s. Doherty's plea that wise management led to the elimination of economic as well as physical waste began to receive attention.

The problem in the United States was that many operators were involved in the production of any given pool. This was free enterprise at work accompanied by individualistic opposition to government mandate to enforce unit operation. Because of the differing circumstances in most foreign fields, however, a different story unfolded. The Anglo-Persian Oil Company (APOC), like other European firms, had lagged behind its U.S. counterparts in a number of technical areas, including the adoption of thermal cracking and gas absorption plants to obtain rich "natural gasoline." However, because APOC had a monopolistic position in the huge Iranian fields it was able to unitize its operations. By carefully regulating the number of wells drilled, APOC could conserve the volume of natural gas in solution with the petroleum and operate the pool in the most efficient manner to sustain long-term recoverability.

The U.S. oil industry was still operating within the schizophrenic world of alternate scarcity and overabundance. A perceived scarcity of petroleum reserves at the beginning of the 1920s, coupled with increased demand, stimulated other technological developments. The Bureau of Mines conducted extensive research into enhanced and secondary recovery techniques such as the waterflooding of older fields to boost production. Integrated firms developed their own thermal cracking technologies in efforts to circumvent the Burton patents held by Standard of Indiana and increase their own gasoline output. There was also brief flirtation with alternative liquid fuels in the early part of the decade. Standard Oil (New Jersey) marketed a 25% ethanol–gasoline blend in 1922–1923 and later obtained the basic German patents for coal liquefaction in a licensing with the I. G. Farben interests. There was also a boom in Western shale oil in the 1920s, which has left an interesting if eccentric history. All these liquid fuel alternatives would soon prove unnecessary as new strikes of oil again dampened anxiety about shortages.

By 1930, oil's share of total U.S. energy consumption had increased to an impressive 23.8%, the most dramatic shift being in gasoline consumption. In 1920, oil-based motor fuel represented only 2.2% of total oil consumption compared with 48.1% for fuel oil. By 1930, gasoline's percentage

had increased to 42.6% of the total and fuel oil's relative figure had shrunk to 43.5%. Kerosene production and sale had decreased sharply. However, traditional problems still plagued the industry. The opening of the East Texas field in 1930 again demonstrated that abundance was a double-edged sword. As unrestricted flow accelerated, the price of crude that had sold for $3 a barrel in 1919 had decreased to 10¢ in the summer of 1931. Voluntary or even compulsory unitization of pools was powerless in the face of such a major problem of overproduction. A new emergency approach developed in the states to combat this economic disaster—regulated prorationing, or the enforced limiting of production at each wellhead. Initially introduced under martial law by the Texas and Oklahoma governors in 1931, legislation in 1932 ceded the power to regulate production to the Texas Railroad Commission and the Oklahoma Corporation Commission. These actions did provide some stability: The price of crude increased to 85¢ a barrel by 1934. Federal New Deal legislation supported and augmented state practice as Franklin Roosevelt's National Recovery Administration (NRA) supported "fair competition" under the NRA Oil Code and policed the interstate shipment of "hot oil," oil produced illegally above state-mandated prorationing quotas. When the Supreme Court struck down the NRA in 1935, senators and congressmen from the oil-producing states swiftly moved to fill in the gap of cooperative regulation with passage of the Connolly Hot Oil Act and the Interstate Oil Compact Commission in 1935. These laws would serve to regulate production and combat "economic waste" (e.g., low prices) to the present day.

6. THE GLOBALIZATION OF THE INDUSTRY

At the end of World War I, British Foreign Secretary George Curzon stated that "the allies had floated to victory on a sea of oil." The major world navies had converted their coal-fired boilers to fuel oil in the critical years before the war and the noninterrupted flow of fuel for these vessels did indeed prove crucial. Winston Churchill, who had pushed for the conversion of the Royal Navy to fuel oil, was also the major player in the British government decision to take a controlling 51% financial position in the Anglo-Persian Oil Company (BP) in 1914. Unlike much of the activity of the former European colonial powers, the history of U.S. operations abroad has centered on the efforts of private firms obtaining concessions and the government has never owned or directly participated financially in actual operations. The U.S. State Department and other agencies have at various times, however, given strong support to these private efforts. Initial U.S. exploratory ventures occurred in Mexico early in the 20th century and in Venezuela in the 1920s. The most important producing area in the world today, the Middle East, had remained a British sphere of influence for a very long time. U.S. firms first obtained concessions in Iraq (Mesopotamia) in 1928 and in 1934 Standard of New Jersey (Exxon) and Standard of New York (Mobil) acquired one-fourth of the Iraqi Petroleum Company. Standard Oil of California (Chevron) gained leases in Bahrain Island in 1927 and Saudi Arabia in 1933, selling half of its rights to Texaco in 1936 when the two companies jointly formed the Arabian-American Oil Company (ARAMCO). The large vertically integrated major firms had control of large domestic supplies, but they viewed the obtaining of foreign crude as a rational economic strategy for their foreign marketing activities and as a hedge against continuing predictions of long-term depletion of U.S. supplies.

Oil imports into the United States prior to 1945 remained low, rarely exceeding 5% of total consumption. However, from 1945 to 1959 the total demand for petroleum products increased by approximately 80%, the single largest share being for gasoline as the automobile culture emerged triumphant in the postwar era. By the 1950s, cheaper, imported oil accounted for approximately 12% of needs and domestic U.S. producers were becoming alarmed by the challenge. Moreover, the Internal Revenue Service had ruled that U.S. companies operating abroad could deduct all foreign oil royalties and taxes from their corporate income tax bill, thus providing a huge financial incentive to pump Middle Eastern crude. President Dwight Eisenhower responded to political pressure from independent oilmen, most of them in western states, by asking the majors to implement voluntary import curbs. In 1959, acknowledging the failure of such voluntary measures, Eisenhower imposed mandatory import quotas.

Despite these quotas, as demand continued to expand in the gas-guzzling United States, imports became seen as the only way to meet needs. When the Nixon administration lifted quotas in 1973, oil imports as a percentage of total U.S. consumption dramatically increased, reaching 38% in 1974. At

the same time, a regulated price structure for domestically produced oil, a carryover from Nixon's 1971 anti-inflation pricing policies, discouraged exploration and drilling at home. Imports increased to approximately one-half of total U.S. consumption by 1979 and have continued at these levels or above.

7. INTERNATIONAL COMPETITION AND WAR

World War II further crystallized the strategic importance of oil to international security, and those nations with ready access to supply were greatly advantaged. When Hitler invaded Poland in 1939, fully one-third of the total oil capacity with which Nazi Germany went to war was represented by synthetic oil derived from brown coal (approximately 19 million barrels per year). Germany's need for oil reserves is evident in Hitler's decision to invade Soviet Russia in June 1941. Most are aware of the German thrust toward Moscow that was only thwarted with Hitler's legions in view of the capital. An equally and perhaps more important southern line of attack aimed at the rich oil-producing properties in Baku in the southern Caucasus is less known. The United States' decision to embargo shipments of high-octane aviation fuel to Japan in 1941 was a significant factor in the Imperial government's decision to launch the war at Pearl Harbor and to attack Dutch interests in the East Indies in search of oil to fuel its military machine.

World War II also represented an important watershed in the history of the U.S. petroleum industry. As the nation mobilized for conflict, there was a validation of the New Deal consensus that a cooperative relationship between business and government represented the best way to achieve and maintain a stable industry in the wake of the inherently uncertain realm of unpredictable resources and rising demand. Much of the suspicion toward New Deal regulation was superseded by a new spirit, imbued with patriotism, but also pleased that production curtailment was being replaced by dramatic efforts aimed at increasing production. The wartime Petroleum Administration for War (PAW) fostered cooperation in pipeline construction, the patent pooling of technical processes, and the development of new catalytic cracking plants needed to produce huge quantities of 100 octane aviation gasoline demanded by new high-performance aircraft. If the allies had indeed floated to victory on a sea of oil during the 1914–1918 conflict, one can make a good case that they flew to victory in the 1939–1945 war.

A council of oil executives, the Petroleum Industry War Service Committee, was formed to advice the PAW, headed by Secretary of the Interior and New Deal zealot Harold Ickes. The oil czar's new image of working with industry to boost oil production contrasted with his previous reputation among oilmen as Horrible Harold. In 1946, the success of this industry advisory group was so valued by government and industry that President Harry Truman continued it in peacetime with the creation of the National Petroleum Council (NPC). The NPC has continued to function as a consulting and advisory body to government and as a generator of reports and studies on all aspects of the oil industry. It has primarily worked in coordination with the Department of the Interior and, after 1977, the Department of Energy.

Another major outcome of the war was an increased awakening of U.S. interest in foreign oil resources and the laying of the foundation for the structure of oil importation that characterizes the industry today. During World War II, the protection of Middle East oil reserves became an essential aim of U.S. foreign policy.

After the war, the exploitation of oil abroad continued to expand at the same time that conservation in the form of market-demand prorationing continued at home. When Chevron and Texaco invited Mobil and Exxon into partnership with ARAMCO in 1946, the U.S. government response was to grant clearance to the companies from antitrust prosecution, citing the value to the national interest. However, U.S. oil policy was ambivalent during the period of the early Cold War. The State Department continued to support aggressive U.S. firms operating abroad, but the industry was still split between independent companies, which were primarily engaged in the domestic industry, and the larger major multinational firms, which were seeking to expand their interests overseas.

8. THE MIDDLE EAST, OPEC, AND THE ENERGY CRISIS

In the 1950s and 1960s, an increasing number of former European colonies in Asia and Africa were gaining independence. In other areas where colonial powers had not maintained political control in the

strictest sense for some time, such as the Middle East, multinational oil companies effectively controlled the destinies of these nations through their economic power. The formation of the Arab League in 1945 and the first Arab–Israeli war in 1948–1949, which erupted in response to the birth of modern Israel, also pointed to a growing Arab nationalism. With Israel's victory, the Palestinians were now without a homeland and this issue has dominated Middle Eastern politics ever since. The first direct threat to Western oil interests, however, came not from an Arab state but from a strategically situated nation whose interests have often coincided with her Arab neighbors, Iran.

Iranian Prime Minister Mohammed Mossadegh's nationalization of British Petroleum's properties in 1951 led to the first real postwar oil crisis. This nationalist challenge to Western interests, coupled with the proximity of Iran to Soviet Russia, prompted the United States to act. The Central Intelligence Agency took a direct hand in a political counterrevolution that overthrew Mossadegh and installed the U.S.-backed Shah, Reza Pahlavi, to the throne in 1953. With U.S. State Department support, U.S. companies now obtained 40% of a new foreign consortium established under the Shah to exploit Iranian oil. The Shah remained America's man in Iran until his overthrow during the Iranian revolution of 1979. During that time, private oil interests were served well with access to Iranian crude, and the State Department enjoyed Persian Gulf stability through military and economic support of the Shah.

A new breed of Arab leaders, such as Egypt's Gamal Abdel Nasser, also symbolized this new nationalistic and independent attitude. Nasser's seizure of the Suez Canal in 1956 was the precipitating factor in the second Arab–Israeli war. Libya's Colonel Muammar Qaddafi's ascension to power in 1969 represented an even more aggressive leadership and an assertion of Muslim values. Superimposed over all these events was the playing out of Cold War tensions as both the Soviet Union and the United States sought to assert their influence in this vital area. In the midst of this changing political landscape, representatives of the major petroleum-producing countries meeting in Baghdad, Iraq, in 1960 formed the Organization of Petroleum Exporting Countries (OPEC). This new cartel of countries, not companies, viewed itself as a counterweight to the international cartel of multinational oil companies later christened the Seven Sisters (British Petroleum, Royal Dutch/Shell, Exxon, Chevron, Mobil, Texaco, and Gulf). Arab spokesmen among OPEC began to talk about "using the oil weapon" but little happened until the outbreak of the third Arab–Israeli conflict in 1967. On June 6, 1967, the day after hostilities began, Arab oil ministers called for an embargo on oil shipped to nations friendly to Israel. The flow of Arab oil was reduced by 60% by June 8 and a crisis threatened. Although there were examples of property destruction and temporary economic dislocation, the embargo was a failure in the wake of the brief 6-day war. The oil companies were able to weather the storm and the major losers were the Muslim nations, which lost tremendous oil revenue during the embargo. The situation would not be the same when similar problems occurred in 1973.

Israel had defeated the Arabs with relative ease during the three previous Middle Eastern conflicts since World War II. Despite their huge oil reserves, it appeared that the Arab nations were no military match for a modern technologically equipped military power with American and European support. The Yom Kippur War that broke out in the fall of 1973 had a different outcome. Catching Israel by surprise on Judaism's high holy day, Egypt and Syria inflicted heavy casualties on Israel in 3 weeks of bloody fighting. At the time the war broke out in October 1973, OPEC was meeting in Vienna. OPEC took two strong measures: They voted to raise the posted price of oil from $3 to $5 a barrel, and they announced an embargo against countries supporting Israel. The Nixon administration had been caught off guard at a point when the sitting president was fighting to keep his job in the midst of the Watergate scandal. Secretary of State Henry Kissinger performed brilliantly in negotiating a cease-fire through rounds of shuttle diplomacy that saw him flying from capital to capital. Although military operations halted at precisely the time when Israel appeared to be gaining the upper hand in the conflict, Kissinger's immediate goal was to cease hostilities. Further diplomacy brought a peace settlement, but the world would experience the aftermath of the embargo and price shocks of October 1973 for the rest of the decade.

High oil prices fueled the increasing problem of stagflation (inflation coupled with stagnating economic growth) that haunted the remaining months of the Nixon presidency until his resignation over Watergate-related matters in August 1974. A similar set of problems limited the effectiveness the of brief presidency of Gerald R. Ford. The presidential election of 1976 brought a Washington outsider, James Earl "Jimmy" Carter, to office and energy policy emerged as a central issue of his

administration. The multinational oil companies had come increasingly under attack in the 1970s on charges of complicity in the OPEC embargo, the taking of "obscene" profits when the market price of crude tripled, and for opposing alternative forms of energy out of selfish economic motive. There was a growing literature on multinational firms generally that asked the question, To whom do these firms owe loyalty? Examples of the fuzzy nature of this question existed particularly with regard to instances of the behavior of oil companies in the international arena. For example, there were many complaints that ARAMCO, ostensibly a U.S. company, had fully cooperated with Saudi Arabia and OPEC during the embargo. In another case, there was evidence that Gulf Oil had supported one faction in the Angola civil war that controlled its area of business operation, whereas the U.S. government backed a competing rebel group. The political economy of world oil had become incredibly more complicated, and the aims of the multinational oil industry had become ever more intertwined with international diplomacy.

9. OIL IN THE MODERN ECONOMY

President Carter urged the American people to view the 1970s energy crisis as the moral equivalent of war. The problem was that the idea never really took hold. To be sure, there had been frightening events in 1973–1974, such as citizens freezing to death for want of fuel oil in New England and a man shot to death in California when he cut into the line at a service station. However, by 1977 oil prices had begun to stabilize as price incentives for new exploration in Alaska, Mexico, and the North Sea along with completion of the Alaska pipeline kicked in. President Carter's and Energy Secretary Schlessinger's appeal for energy conservation had also begun to pay off. Perhaps the most significant legacy of Carter's policies will be his deregulation of oil and natural gas that reintroduced a program of more realistic market pricing. Strong political support for investment in alternative energy technologies, such as solar power, synthetic fuel, wind energy, and biomass conversion, attracted much publicity and enthusiasm but faded amid cynical charges that the energy crisis had been contrived by the oil companies and with the return of relatively lower oil prices. Nuclear power advocates believing that Carter, a nuclear engineer, would give a boost to that industry had their hopes dashed with the Three Mile Island disaster of March 28, 1978—a blow from which the industry has never fully recovered. Negotiations with the Saudi and Iranian governments in 1978 led to a realization that it was in the best economic and political interests of the oil producers to curtail prices and it appeared that oil prices had achieved stability. However, the international oil economy would have a more cruel blow in store for Jimmy Carter—one that probably cost him reelection in 1980.

President Carter visited Iran and congratulated the Shah as a force of stability in the Persian Gulf. Soon after, when the Iranian people overthrew the Shah in 1979, halting Iranian oil production in the process, another major spike in prices resulted. As had occurred in 1973, the consuming countries seemed unable to compensate for what was, on a global basis, only a small reduction in supplies. Jimmy Carter's political fate became sealed when the United States allowed the Shah into the country for medical and humanitarian reasons and the U.S. embassy was stormed and 50 hostages were ultimately held captive. Candidate Ronald Reagan benefited from a number of Carter's difficulties, including very high interest rates and an apparent weak policy abroad, but high oil prices coupled with the embarrassment of the hostage crisis comprised a fatal mix. The revolutionary government of Iran rubbed salt in the wound when it released the hostages on the day of Ronald Reagan's inaugural as president. President Reagan also benefited from a weakening of OPEC solidarity as members sold oil at below posted prices in order to obtain badly needed money. In 1982, non-OPEC production exceeded OPEC production for the first time as production from the North Sea, Soviet Union, Mexico, Alaska, Egypt, Angola, and China added to world reserves. Although OPEC still remains a relevant player in the political economy of oil, its role is much diminished from its position in the early 1970s. The long-running Iran–Iraq war caused price dislocations when it first started in 1980, but it actually bolstered world supplies in the long term because both nations were forced to sell above quota for needed revenue.

The industry has also had to deal with an increasing number of environmental issues in recent decades. One can date offshore drilling technology, the placing of wells in submerged lands, to the early years of the 20th century, but the technique has only become of critical importance recently. Operators in Santa Barbara, California, for example, had drilled more than 200 wells from piers extending out over the water in 1906, and drilling had been attempted off the coast of Peru that same year. Further experiments went forward in the United States and

abroad during the 1920s and the 1930s, but it would not be until after World War II that offshore drilling began to take on the importance that it has today with deeper drilling techniques. Drilling off the Texas Gulf Coast, for example, became very extensive after the war. However, the events of January 1969 off the coast of Santa Barbara sent shocks throughout the industry. When the drilling of an offshore well in the Santa Barbara channel ran into difficulties that sent a slick of heavy crude onto 30 miles of beaches, there was a huge public outcry. It was also in 1969 that BP, Humble Oil (Exxon), and Atlantic Richfield announced plans to construct a pipeline across the Alaskan tundra aimed at transporting oil from Prudhoe Bay, north of the Arctic Circle, to the port of Valdez. The ensuing debate delayed the project, brought cost overruns, and pitted the environmental movement against the industry in a very public way before the project was finally completed in 1977. Things quieted down with the wide recognition of the part that Alaskan oil had played in dampening high oil prices. Then the Exxon Valdez incident occurred. Oil spills from seagoing tankers have always presented a risk, but the size of the new generation of supertankers exacerbated concerns. When the supertanker Exxon Valdez went aground in Alaska's Prince William Sound spilling 240,000 barrels of Prudhoe Bay oil in 1989, environmental critics of the industry received a further boost. Although it is environmentally risky, deep-water drilling as is occurring in the North Sea, wilderness pipelines, and the passage of supertankers on the oceans have become essential elements of commerce in world oil.

The industrialized nations have enjoyed a relatively more benign oil economy in the 1990s and into the early new century, but the historic pattern of dependency on oil controlled by the producing nations remains a reality. In 2003, the United States imported 55% of its oil; in the crisis year of 1973, imports amounted to only 35%. The Iraqi invasion of Kuwait on August 2, 1990, precipitated a series of problems in the Persian Gulf region. Future historians will interpret the defeat of Saddam Hussein in the 1991 Gulf War by the first Bush administration as but a chapter in the events that led to the U.S. invasion of Iraq in the spring of 2003 directed by his son. Neither in 1991 nor in 2003 was oil the sole factor in precipitating international crisis, but one cannot ignore the extremely significant role that it has played. Critics point to the very close ties between the administration of the second President Bush and the oil industry, and many have uttered the protest refrain of "No Blood for Oil" with reference to the war in Iraq. However, in a figurative sense oil has become the blood of international commerce and industry more than ever as we embark on a new century.

One of the main business trends in the global oil industry in recent years exhibits a touch of irony for the United States. This is the movement toward megamergers initiated by the acquisition by Chevron (Standard Oil of California) of Gulf Oil in 1984 and continuing through 2000. The demise of Gulf occurred after swashbuckling financier T. Boone Pickens' small Mesa Petroleum Company shocked the industry with a takeover bid for the large multinational. Chevron's ultimate acquisition of Gulf was followed by a series of additional moves, including the Royal Dutch/Shell group acquiring that part of Shell USA that it did not already own; BP's merger with Sohio (Standard Oil of Ohio); the joining of Exxon (Standard Oil of New Jersey) and Mobil (Standard Oil of New York); Chevron's merger with Texaco; and the merger of BP, Amoco (Standard Oil of Indiana), and ARCO (Rockefeller's old Atlantic Refining Company). The notorious Seven Sisters of the 1970s had shrunk considerably by the turn of the century. Historical irony lies in this partial reassembling of the pieces of the old Rockefeller Standard Oil Company, which had been fragmented into 34 separate companies in 1911.

Why this rush toward consolidation? In general terms, this oil industry behavior relates to the broader trend of U.S. merger activity beginning in the 1980s and present today. The argument that firms required a leaner, more competitive profile to respond to the challenges of the new global economy became most prominent, and a more benign antitrust environment created by recent Republican administrations in Washington has abetted the trend. For the oil industry specifically, there are other reasons cited. One of these is the argument that the companies representing the consuming nations need to have a stronger bargaining position with OPEC and other producing groups. Others see consolidation as a key to mobilizing more effective resistance to world environmental movements, such as the Kyoto Protocol, which have called for the reduction of heat-trapping greenhouse gas emissions caused by the burning of coal, oil, and gasoline. Certainly, environmental groups have been among those most critical of the industry mergers. Some industry spokesmen have referred to declining profits in the wake of relatively lower world oil prices as the reason for moving toward the cost reductions resulting from merger. Within the context of the

history of the oil industry, one might ask simply "how big should big oil be?"

Oil supplies will eventually become exhausted; this is an accepted reality. However, there remain huge supplies of petroleum and a world economy that depends on access to it. There will be an historic transition to other energy sources in the future and the individual companies that comprise the industry may very well have diversified to the extent that they will become part of this new political economy. The history of the oil industry has been relatively brief in years, approximately a century and a half, but it has been a richly diverse and fascinating story that is still not concluded.

SEE ALSO THE FOLLOWING ARTICLES

Coal Industry, History of • Electricity Use, History of • Geopolitics of Energy • Hydrogen, History of • Manufactured Gas, History of • Nationalism and Oil • Natural Gas, History of • Nuclear Power, History of • Oil Crises, Historical Perspective • Oil-Led Development: Social, Political, and Economic Consequences • OPEC, History of • OPEC Market Behavior, 1973–2003 • War and Energy

Further Reading

Brantly, J. E. (1971). "History of Oil Well Drilling." Gulf, Houston, TX.

Chernow, R. (1998). "Titan: The Life of John D. Rockefeller, Sr." Random House, New York.

Clark, J. G. (1987). "Energy and the Federal Government: Fossil Fuel Policies, 1900–1946." Univ. of Illinois Press, Urbana.

Clark, J. G. (1990). "The Political Economy of World Energy: A Twentieth Century Perspective." Univ. of North Carolina Press, Chapel Hill.

Gorman, H. S. (2001). "Redefining Efficiency: Pollution Concerns, Regulatory Mechanisms, and Technological Change in the U.S. Petroleum Industry." Univ. of Akron Press, Akron, OH.

Olien, R. M., and Olien, D. D. (2000). "Oil and Ideology: The Cultural Creation of the American Petroleum Industry." Univ. of North Carolina Press, Chapel Hill.

Painter, D. S. (1986). "Oil and the American Century: The Political Economy of U.S. Foreign Oil Policy, 1941–1954." Johns Hopkins Univ. Press, Baltimore.

Pratt, J. A., Becker, W. H., and McClenahan, W. Jr. (2002). "Voice of the Marketplace: A History of the National Petroleum Council." Texas A&M Univ. Press, College Station.

Schneider, S. A. (1983). "The Oil Price Revolution." Johns Hopkins Univ. Press, Baltimore.

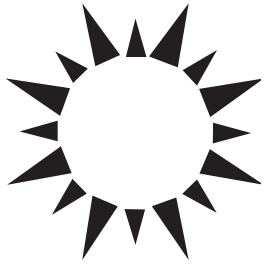
Vietor, R. H. K. (1984). "Energy Policy in America Since 1945." Cambridge Univ. Press, New York.

Williamson, H. F., and Daum, A. (1959). "The American Petroleum Industry: The Age of Illumination, 1859–1899." Northwestern Univ. Press, Evanston, IL.

Williamson, H. F., Andreano, R. L., Daum, A., and Klose, G. C. (1963). "The American Petroleum Industry: The Age of Energy, 1899–1959." Northwestern Univ. Press, Evanston, IL.

Yergin, D. (1991). "The Prize: The Epic Quest for Oil, Money & Power." Simon & Schuster, New York.

Oil-Led Development: Social, Political, and Economic Consequences



TERRY LYNN KARL
Stanford University
Stanford, California, United States

1. **Introduction**
2. **Definitions: Oil Dependence, Resource Curse, Dutch Disease, and Rentier States**
3. **Poverty and Social Welfare Consequences of Oil-Led Development**
4. **Oil-Related Changes in Social Structure**
5. **The Rentier State**
6. **Oil, Political Stability, and State Capacity**
7. **Social and Environmental Impacts at Regional and Local Levels**
8. **Petroviolence and Civil War**
9. **Conclusion**

Glossary

corruption Though often used interchangeably with rent seeking, corruption is more narrowly defined as the misuse of public power or resources for private gain, and it is generally illegal.

Dutch disease Named after the negative effects of the North Sea oil boom on industrial production in The Netherlands, this phenomenon occurs when resource booms cause real exchange rates to rise and labor and capital to migrate to the booming sector. This results in higher costs and reduced competitiveness for domestically produced goods and services, effectively "crowding out" previously productive sectors.

oil-led development An overwhelming dependence on revenues from the export (and not the internal consumption) of petroleum, as measured by the ratio of oil and gas to gross domestic product, total exports, and the contribution to central government revenues.

rent In Adam Smith's classic definition, this is unearned income or profits "reaped by those who did not sow." According to economists, rents are earnings in excess of all relevant costs, including the market rate of return on invested assets. They are the equivalent of what most noneconomists consider to be monopoly profits.

rentier state A state that lives from externally generated rents rather than from the surplus production of the population. In oil-exporting states, this is measured by the percentage of natural resource rents in total government revenues.

rent seeking The efforts, both legal and illegal, to acquire access to or control over opportunities for earning rents. In oil-dependent countries, rent seeking refers to widespread behavior, in both the public and the private sectors, aimed at capturing oil money through unproductive means.

resource curse The negative growth and development outcomes associated with minerals and petroleum-led development. In its narrowest sense, it is the inverse relationship between high levels of natural resource dependence and growth rates.

Proponents of oil-led development believe that countries lucky enough to have "black gold" can base their development on this resource. They point to the potential benefits of enhanced economic growth and the creation of jobs, increased government revenues to finance poverty alleviation, the transfer of technology, the improvement of infrastructure, and the encouragement of related industries. But the experience of almost all oil-exporting countries to date illustrates few of these benefits. To the contrary, the consequences of oil-led development tend to be negative, including slower than expected growth, barriers to economic diversification, poor social welfare indicators, and high levels of poverty, inequality, and unemployment. Furthermore, countries dependent on oil as their major resource for development are characterized by corruption and

 © 2004 Elsevier Inc. All rights reserved.

exceptionally poor governance , a culture of rent seeking, often devastating economic, health, and environmental consequences at the local level, and high incidences of conflict and war. In sum, countries that depend on oil for their livelihood eventually become among the most economically troubled, the most authoritarian, and the most conflict-ridden in the world.

1. INTRODUCTION

Oil is a commodity with special characteristics. These include its unique role as both the common natural heritage of a country and the motor of global industrialization, its depletability, its price volatility and consequent boom–bust cycles, its especially high capital intensity and technological sophistication, its enclave nature, and the exceptional generation of profits that accrue to the state and to private actors. The combination of these factors produces what has been called "the paradox of plenty" or the "resource curse." This is not due to the resource per se, which is simply a black and viscous substance, and it is not inevitable. A resource boom can be beneficial or detrimental: Norway, an oil exporter, has used the benefits of North Sea petroleum to earn the highest place on the United Nations Development Program's list of best development performance, whereas other exporters, such as Nigeria and Angola, are clustered near the bottom. Instead, what matters for the social consequences generated by petroleum dependence are, first, the type of preexisting political, social, and economic institutions available to manage oil wealth as it comes onstream and, second, the extent to which oil revenues subsequently transform these institutions in a rentier direction. Because almost all proved oil reserves exist in less developed countries, where administrative institutions tend to be weak (only 4% can be found in advanced industrialized democracies), the probability of the resource curse is exceptionally high.

2. DEFINITIONS: OIL DEPENDENCE, RESOURCE CURSE, DUTCH DISEASE, AND RENTIER STATES

Mineral (especially oil-led) development is often promoted as a key path for developing countries seeking sustained economic growth. But the oil-led development model of today is significantly different from the role that energy played in the late 19th and early 20th centuries in the United States, Canada, and Australia. In those earlier and more successful experiences, mining and oil exploitation contributed only a very small percentage of total economic output, never dominated exports, and never came anywhere near the magnitude of dependence that characterizes contemporary oil-led development. While leaving a considerable regional impact, oil and minerals were never the motor of development.

Today, to the contrary, oil-led development means that countries are overwhelmingly dependent on revenues gleaned from the export of petroleum. This dependence generally is measured by the ratio of oil and gas exports to gross domestic product; in countries that are sustained by petroleum rents, this figure ranges from a low of 4.9% (in Cameroon, a dependent country that is running out of oil) to a high of 86% (in Equatorial Guinea, one of the newest oil producers). Dependence is also reflected in export profiles, with oil in dependent countries generally making up 60–95% of a country's total exports. Oil-dependent countries can be found in all geographic regions of the world, although they are most commonly associated with the Middle East and, more recently, Africa.

Oil-dependent countries suffer from what economists call the "resource curse." In its simplest form, this refers to the inverse association between growth and natural resource abundance, especially minerals and oil. This association repeatedly has been observed across time and in countries that vary by population size and composition, income level, and type of government; it is so persistent that has been called a "constant motif" of economic history. Specifically, countries that are resource poor (without petroleum) grew four times more rapidly than resource-rich (with petroleum) countries between 1970 and 1993, despite the fact that they had half the savings. Similar findings have been replicated through a study of the members of the Organization of Petroleum Exporting Countries (OPEC), using a different and longer time period, from 1965 to 1998. OPEC members experienced an average decrease in their per capita gross national product (GNP) of 1.3% per year during this period, whereas lower and middle-income developing countries as a whole grew by an average rate of 2.2% per year over the same time. Moreover, studies show that the greater the dependence on oil and mineral resources, the worse the growth performance. Finally, countries dependent on the export of oil have not only performed

worse than their resource-poor counterparts; they have also performed far worse than they should have given their revenue streams.

The causes of this resource curse are a matter of debate, but the negative association between growth and oil and mineral wealth is not attributed to the mere existence of the natural resource. Oil in itself cannot encourage or hinder growth. Instead, this association is less direct and, though the weight of various specific causal mechanisms is still debated, it is generally attributed to some combination of the factors. First, oil windfalls can hurt other sectors of the economy by pushing up the real exchange rate of a country's currency and thus rendering most other exports noncompetitive, a phenomenon called the "Dutch disease." The reduced competitiveness in agricultural and manufacturing exports then "crowds out" other productive sectors and makes the diversification of the economy particularly difficult. This in turn reinforces the dependence on oil and, over time, it can result in a permanent loss of competitiveness.

Second, the long-term price deflation and price volatility of the international primary commodities market hinders economic development. Since 1970, this volatility has grown worse, and oil prices are twice as variable as those of other commodities. This means that oil economies are more likely to face more frequent economic shocks, with their attendant problems, and they are especially susceptible to acute boom–bust cycles. This oil price volatility exerts a strong negative influence on budgetary discipline and the control of public finances as well as on state planning, which subsequently means that economic performance deviates from planned targets by as much as 30%. Volatility also exerts a negative influence on investment, income distribution, and poverty alleviation.

Third, the enclave nature of the industry combined with its capital intensity fosters especially weak linkages to the broader economy and does little to create employment. Because oil is the world's most capital-intensive industry, the sector creates few jobs per unit of capital invested, and the skills required by these jobs often do not fit the profile of the unemployed. If growth in the oil sector had a significant multiplier effect, this would not be such a great problem, but the productive linkages between this sector and the rest of the economy tend to be weak. Furthermore, the opportunities for technology diffusion are very limited, and so is infrastructure development. Downstream processing industries have typically not emerged, and when they do, they are often at a competitive disadvantage.

Perhaps most important, petroleum may be one of the hardest resources to utilize well; countries dependent on oil exports seem particularly susceptible to policy failure. The reason lies in the weakness of preexisting institutions in places where oil for export is found, their frequently authoritarian character, and the ease with which they can be transformed by an overwhelmingly powerful export sector. Generally, oil rents produce a rentier state—one that lives from the profits of oil rather than from the extraction of a surplus from its own population. In rentier states, economic influence and political power are especially concentrated, the lines between public and private are very blurred, and rent seeking as a wealth creation strategy is rampant. Rentier states are notoriously inefficient because productive activity suffers and self-reinforcing "vicious" development cycles can set in. Together, all of these factors slow growth, raise powerful barriers to the diversification away from petroleum dependence, and produce the skewed development patterns described by the resource curse.

3. POVERTY AND SOCIAL WELFARE CONSEQUENCES OF OIL-LED DEVELOPMENT

One of the most important social consequences of the resource curse is that oil-exporting countries have unusually high poverty rates, poor health care, high rates of child mortality, and poor educational performance given their revenues—outcomes that contradict the beliefs about what should happen within oil-exporting countries. Though it is true that most forms of primary commodity dependence are associated with poverty, not all commodities are equally culpable. Countries dependent on agricultural commodities tend to perform better with respect to poverty, minerals in general are linked to high levels of poverty, and oil dependence in particular is correlated with low life expectancy and high malnutrition rates.

Oil dependence has an ambiguous relationship with poverty alleviation, and this is related to the boom–bust cycles accompanying dependence on the resource. At the beginning of oil exploitation for export, per capita income rises during the "euphoric," or "boom," period. Especially in the initial stages of production for export, petroleum revenues initially transform a society, often suddenly and dramatically. Employment increases, infrastructure is

improved, and per capita income grows rapidly. Thus, for example, per capita oil exports for North Africa and the Middle East soared from $270 in 1970 to $2042 in 1980, and this fueled accelerated economic activity. But the failure to diversify from oil dependence into other self-sustaining economic activities, especially agriculture and labor-intensive industry, becomes a significant obstacle to pro-poor development. Over time, as booms peter out, oil exporters are plagued by (often sudden) declines in per capita income. In North Africa and the Middle East, for example, per capita oil exports plunged from the 1980 high of $2042 in 1980 to $407 by 1992 as oil prices dropped and population increased. Paradoxically, in what seems to be the midst of plenty, a high percentage of people living in oil-exporting countries tend to remain poor or suffer from dramatic shifts in their welfare that ultimately leave them in poverty. Thus, despite significant rises in per capita income, over the past several decades, all oil-dependent countries have seen the living standards of their populations drop, and sometimes drop very dramatically. This boom–bust cycle affects even the world's richest oil exporters. In Saudi Arabia, for example, where proved reserves are the greatest in the world, per capita income has plunged from $28,600 in 1981 to $6800 in 2001.

For many countries, including Algeria, Angola, Congo, Ecuador, Gabon, Iran, Iraq, Kuwait, Libya, Qatar, Saudi Arabia, and Trinidad Tobago, this plunge has been very severe, moving real per capita incomes back to the 1970s and 1980s. For a few countries, most notably Nigeria and Venezuela, the growth of poverty has been catastrophic; in these cases, real per capita income has plummeted to 1960 levels. It is almost as if 40 years of development had not taken place. In Nigeria, the disparity between oil wealth and poverty is especially notable. Despite the fact that over $300 billion in oil profits has been generated over the past 25 years, the proportion of households living below the United Nation's absolute poverty line of $1/day grew from 27% in 1980 to 66% by 1996. Income disparities are shocking: the richest 10% controls 40% of the country's wealth and its poorest 20% has a share of just 4.4%. But oil dependence is associated with more than sudden shifts in levels of poverty and exceptionally low living standards for much of the population in petrostates. It is also linked to unusually high rates of child mortality, child malnutrition, low life expectancy, poor health care, and reduced expenditures on education. In countries dependent on oil and/or minerals, both infant mortality and life expectancy at birth are worse than in non-oil and non-mineral countries at the same income levels. Simply put, when taken as a group, the greater the dependency of a country on oil, the greater the likelihood that children born there will die at birth, will have poorer health care, nutrition, and education than their resource-poor counterparts, and will die sooner, if they survive at birth.

The statistics are startling. For each 5-point increase in oil dependence, the under-5-years-old mortality rate rises by 3.8 per 1000. This may be due to the fact that oil dependence is also negatively correlated with health care expenditures. Paradoxically, the more countries are dependent on oil, the less they spend on health as a percentage of gross domestic product (GDP). In Nigeria, for example, the government spends about $2 per person per year on health care, which is far less than the $34 per year recommended for developing countries by the World Health Organization. But poor child welfare performance is also due to higher malnutrition rates that exist in oil-dependent countries. Indeed, once the effects of per capita income are taken into account, for every 5-point rise in oil dependence, there is a corresponding 1% rise in the percentage of children under 5 years old who are malnourished. Compare, for example, the global average of 26.5 malnourished children per 1000 to the 37.7 per 1000 rate in oil-rich Nigeria.

Given the available resources, education also performs worse than expected, affecting future prospects for growth. Countries that are dependent on natural resources, inadvertently or deliberately, neglect the development of their human resources by devoting inadequate attention and expenditure to education. Thus, school enrollments tend to be lower than in their non-resource-rich counterparts. In the OPEC countries, for example, 57% of all children go to secondary school compared with 64% for the world as a whole; OPEC spends less than 4% of the GNP on education compared with almost 5% for the world as a whole (in 1997 figures). The explanation for poor educational performance in oil-exporting countries is not clear. Perhaps because the high skill level needed by oil-rich countries in their leading sector can be bought or imported, their governments do not face the same urgent educational imperatives and may underrate the need for strong educational policies. Flooded with easy money, they may perceive more urgent needs than the long-term investments in education that results in long-term development benefits.

4. OIL-RELATED CHANGES IN SOCIAL STRUCTURE

Dependence on petroleum skews the social structure of countries. Because of the enormous capital and technological resources necessary to exploit this resource, foreigners (principally oil companies) become a dominant, if not *the* dominant internal social force, especially at the beginning stages of development. This has important implication for the creation of a domestic entrepreneurial class. Though foreign companies may form partnerships with domestic elites, their overwhelming economic presence and capital and technological advantages mean that domestic entrepreneurs have less opportunity to develop on their own. To be successful, the domestic businesses must forge close ties either to the state or to foreign capital, or they may be marginalized, e.g., merchants in Middle East oil-exporting countries. This pattern exists for other types of primary commodity exporters, but it is more exaggerated in oil-exporting countries because domestic capitalist economic groups, notoriously concentrated in monopolies or oligopolies, are dependent on oil rents and the political power arrangements that choose their distribution through patronage. Thus, instead of a capitalist class, the *nouveau riche*, who are fabulously and ostentatiously rich and owe their success to the state, characterize oil states. But because this wealth is the result of a windfall and privileged links to the state and because it may be largely independent of merit-based efforts made by citizens, this pattern of wealth creation encourages rent seeking as well as a tendency to live beyond one's means.

Middle and professional classes are also shaped by dependence on oil exports as the engine of the economy. Labor markets in oil-exporting countries tend to offer only three major types of jobs—oil related, public sector, and private services—and this retards the growth of a large middle and professional class. When these groups do appear, they differ from other middle and professional classes because their job prospects and standard of living are directly linked to the fortunes of the major export industry, petroleum, thus they are exceptionally vulnerable. During boom periods (for example, the 1970s and early 1980s), jobs and wealth are readily available for the educated, but during bust cycles, middle and professional classes may be educated but have few job opportunities and little prospect for wealth. The outcome is often intense social and generational tension, especially in urban areas, as the population and number of educated grow and employment shrinks. This is most notable in the Middle East, where the young generation of urban middle sectors has seen their parents' situation deteriorate and has arrived too late for economic benefits or adequate social services in stagnant economies.

At the same time, the formation of a broad-based urban working class is compromised. Because oil employs relatively few workers and their skill level must be especially high, and because the rest of the labor market is skewed, dependence on oil fosters a type of labor aristocracy that is separate from most of the workforce. Though this separation, delineated by educational and skill level, can be found in most developing countries, it is especially notable in oil-exporting countries because they have among the fastest rural-to-urban migration rates in the world. Rural poor not only experience the normal economic pulls to cities, but oil rents also create exaggerated expectations of new opportunities—even as the Dutch disease begins to bias against agriculture and agrarian interests. So rapid is the outflow from the countryside that some landlords, most notably those in Iran, have been compelled to import foreign workers to till their lands. This especially rapid rural-to-urban migration means that cities are filled with a relatively small middle and professional class when compared to the vast majority of underskilled and underemployed workers.

Finally, exceptional levels of in-migration characterize oil states. This is encouraged by the structure of the labor market as well as by the pull of oil wealth. In some cases, migration has dramatically altered the profile of petrostates. Most of the oil-exporting countries in the Gulf region, for example, have more foreign than citizen residents! Somewhere between 50 and 90% of private sector workers in the Gulf are foreigners. There are 6 million foreigners among Saudi Arabia's 18 million residents, and foreigners are 98% of manufacturing workers, 97% of construction workers, and 93% of the service sector. This extensive in-migration further distorts levels of inequality because immigrants are generally paid less, compared to nationals. Saudi youth, for example, often expect to earn at least SR2500 ($670) a month for unskilled work, whereas foreigners filling these jobs typically earn SR500–SR1000 a month.

This peculiar social structure is linked to a specific "culture of oil" that permeates all social classes and groups as well as the state. As novelists, journalists, and other observers have repeatedly illustrated, petroleum creates a world of illusion because some people become wealthy without effort. This means

that work ethics are undermined and negative attitudes toward certain forms of work, especially manual labor, can prevail in many oil-exporting countries. This in turn can translate into lower levels of productivity than those found in comparable resource-poor states. As novelists, journalists, and other observers of oil countries have repeatedly noted, states and people that experience a sudden influx of income they did not work hard to get have not usually developed the fiscal and financial discipline or work habits normally required to get and keep such windfalls. They tend to become accustomed to relatively high salaries and little work. For this reason, employers in oil-exporting countries report that they prefer foreign workers who will work harder for less money, seemingly grateful that they may be earning five times the salary possible in their country of origin.

Embedded in this social structure are two fundamental cleavages. First, because the West dominates the oil industry through both production and consumption, a Western definition of modernity based on the market rather than indigenous cultures is transmitted; indeed, oil exporters may experience the most accelerated Westernization and may be the most exposed to Western influence when compared to non-oil countries. At least part of the country will be linked together through this form of modernization, e.g., technocrats, public sector employees, and educated elites. But precisely because oil development accelerates the rate of change, because oil countries are so exposed to the West, and because the discontinuities provoked by petroleum wealth are so great, the failure of the promise of an apparently easy modernization may give rise to decidedly conservative anti-Western movements based on different principles for the organization of economic life, as in Algeria and Iran, or distinctive traditional notions about the depletion of finite resources, as among the U'wa indigenous people in Colombia.

Second, though the inequalities created by oil-led development appear to be at about the same levels as in non-oil states with similar incomes, people in oil-exporting countries may experience these inequalities very differently because they occur in what is widely perceived to be a rich country. The visibility of oil wealth may compound the problem. Where traditional practices are essentially conservative and egalitarian, as in some Latin American indigenous groups, or where religious practices emphasize the need to redistribute income fairly, avoid earning interest, take care of the poor, and prohibit waste and idleness, as in the Islamic world, the cultural shock can be especially powerful. When rulers appear to be wasteful, despotic, and dominated by foreigners, this can produce an especially potent political mix.

5. THE RENTIER STATE

Ineffective and inefficient governance, perhaps more than any other factor, may explain the extent of poverty in oil-dependent countries, but this, too, is related to the presence of oil. Because the revenue base of the state *is* the state, oil rents affect state capacity. Oil dependence skews the institutional development of the state because oil rents weaken agencies of restraint. In resource-poor countries, intense population pressure on scarce resources reduces the tolerance for inefficiency and predation, and the economy cannot support extensive protection or an overexpanded bureaucracy. But in oil states, the brake of scarcity does not exist. Instead, oil dependence encourages the expansion of states into new arenas while weakening opportunities to strengthen administrative capacities, especially non-oil-based tax systems, merit-based civil services, and the rule of law—fundamental elements for creating efficient states.

The impact of oil rents on effective governance has a pernicious effect on the quality of administrative institutions in less developed countries, regardless of whether they are democratic or authoritarian. First, because oil states do not have to extract the majority of their resources from their own populations, they do not have to build the institutional capacities that have historically been required by such extraction. This means that they are denied the information that is generated by a robust tax bureaucracy, and they are also denied the incentives for innovation within a civil service that stems from scarcity. Even where state capacity embedded in tax authorities may have previously existed, oil rents tend to be undermining. With the discovery of oil, these tax authorities are often disbanded because they appear to be no longer necessary. Second, because windfall gains that arise from petroleum encourage rent-seeking behavior, the state becomes a type of "honey pot" in which competing interests try to capture a significant portion of resource rents by capturing portions of the state. A vicious cycle results in which all actors try to gain parts of the bureaucracy while governments, in turn, reward their supporters by funneling favors their way. But this means that the public sector tends to lack the corporate cohesiveness and authority necessary to exercise effective public policy.

Finally, if the state is especially weak and the target of capture, it is also especially overloaded. Oil revenues are the catalyst for a chronic tendency of the state to become overextended, over-centralized, and captured by special interests. This can be seen through the accelerated and especially large growth of the public sector, the overextension of public expenditure, and the unusually extended periods of protection for import-competing sectors. Yet, without the institutional and administrative capacity to cope with this enhanced state role, this overextension is a formula for ineffectiveness.

The most telling indicator of declining state capacity is the loss of fiscal control, measured by overspending and soaring debt as well as by the inability of oil states to reform themselves. This is because oil states degenerate into sophisticated mechanisms to transfer resources from the primary sector to politically influential urban groups, especially as windfall gains provoke a type of "feeding frenzy" to capture petrodollars. This does not occur to the same extent when labor-intensive activities drive economic growth, centering on food and agricultural products, for example, in part because they tend to generate fewer rents. The political competitions for resource rents (when combined with the often nontransparent mechanisms for distributing them) have important efficiency costs. For example, they make it more difficult for governments to moderate spending in response to the price volatility of petroleum, thereby further distorting the economy. In general, oil rents permit incapable state institutions to endure, and ineffective policies to persist, considerably longer than they can in less resource-rich countries. To avoid unpopular reforms, governments use their oil as collateral for borrowing abroad or intensify the squeeze on the export sector. Petrodollars simply permit more scope for cumulative policy error.

States that have the greatest resource endowments, and especially oil-exporting countries, also have the extraordinarily high levels of corruption—a reality confirmed by stunning quantitative evidence and numerous case studies. With incomes of the order $35 billion/year for Mexico; $30 billion/year for Venezuela; $22 billion/year for Nigeria, the temptations for abuse are immense, and with weak state capacity and rule of law in place, there is little institutional restraint. "People rob," one finance minister of an oil-exporting country remarked, "because there is no reason not to." Oil rents and institutional weakness form a vicious cycle. Quantitative evidence suggests that the extent of corruption is higher in countries in which civil service recruitment and promotion procedures rely less on merit-based considerations; where this is the case, efforts to reform the civil service are blocked in order to sustain patterns of corruption. At its worst, this can degenerate into a "corruption trap," whereby payoffs at the top of political and business institutions encourage the corruption of others until a large percentage of public and private sector figures are involved, as the case of Nigeria demonstrates.

Corruption takes place not only at the production and exports stage through secret signature bonuses and opaque financial arrangements, but also as a result of extremely high-level and difficult-to-absorb investments at the "upstream" stage as well as at the trading, or "downstream," stage, where massive resources tend to disappear through price transfers that are difficult to track. Though transactions are obviously clandestine, evidence of oil-related corruption abounds in both the private sector and the state. The former president of the French state oil company, Elf Aquitaine, is charged with presiding over the commission payments on oil deals with African countries. Mobil oil executives are charged with illegal payments in Kazakhstan. In Angola, more than $1 billion/year of oil revenues disappeared between 1996 and 2001, a full one-sixth of the national income, in a country where more than 70% of the population lives on less than $1/day.

Corruption contributes to the resource curse. Rulers will support policies that produce personalized rents even if these policies result in lower overall social welfare; because they need to share these rents with supporters and subordinates, the level of distortion can be very great. Policy choices are deformed in a number of ways. First, when huge oil rents are present, officials tend to favor larger public sectors with overly excessive regulatory interventions that enhance opportunities for rent seeking. Second, policy choices are distorted toward the financing of megaprojects in which payoffs can be more easily hidden and the collection of bribes is easier, whereas productive long-term investment remains undersupplied. Highly capital-intensive and specialized one-of-a-kind designs may be favored so that there are no reliable cost benchmarks; for example, an aluminum smelter was built for $2.4 billion in Nigeria even though it served no valid development objective and its cost was between 60 and 100% higher than the cost of similar plants elsewhere. Infrastructure and defense projects are also favored over health and education expenditures, thereby reducing the quality of public services as well

as lowering the quality of public infrastructure. Most important, corruption affects both economic growth and income levels. Economists estimate, for example, that Venezuela's average GDP growth rate would be raised by some 1.4% annually had it reduced its corruption to the level of corruption in Chile.

6. OIL, POLITICAL STABILITY, AND STATE CAPACITY

Oil and centralized rule seem to go together; and oil and democracy do not generally mix. Political scientists have repeatedly documented this relationship through case studies, and they have found a robust and statistically significant association between oil dependence and authoritarian governments. Oil appears to impede the appearance of democracy in most cases, especially in the Middle East and North Africa, though it facilitated democratization in Venezuela.

The hindering of democratization seems to occur primarily through different, though related, mechanisms. The first is based on how rentier states collect revenues. Because these states live from oil rents rather than from direct taxation, they are likely to tax their populations lightly or not at all. Thus, they are unusually detached from and unaccountable to the general population, and their populations, in turn, are less likely to demand accountability from and representation in government. In effect, the vital link between taxation and representation is broken. Studies have shown, for example, that the governments of Kuwait and Qatar became less accountable to the traditional merchant class in this way. Even in Venezuela, where some type of democracy exists, the lack of taxation has made both representation and state accountability less than expected.

A second causal mechanism depends on how regimes spend state revenues. Oil wealth produces greater spending on patronage, which, in turn, weakens existing pressures for representation and accountability. In effect, popular acquiescence is achieved through the political distribution of rents. Oil states can buy political consensus, and their access to rents facilitates the cooptation of potential opponents or dissident voices. With basic needs met by an often generous welfare state, with the absence of taxation, and with little more than demands for quiet and loyalty in return, populations tend to be politically inactive, relatively obedient and loyal, and levels of protest remain low, at least as long as the oil state can deliver. Thus, for long periods, an unusual combination of dependence, passivity, and entitlement marks the political culture of petroleum exporters. This is especially the case in smaller exporting states such as the Gulf monarchies, where oil reserves per capita are 43 times those of large exporting states such as Algeria, Indonesia, Nigeria, Venezuela, and Iran, and where such costly distributive policies can be sustained for a longer time. Regimes have even used their largess to prevent the formation of social groups, independent from the state, that might someday prove to be political challengers or to rid themselves of already existing challengers—a phenomenon that has been documented (during various historical periods) in Venezuela, Algeria, Iraq, Iran, Kuwait, and Qatar. In the latter two countries, for example, the political distribution of oil rents eliminated the influence of the merchant class in decision making, leaving the rulers with no real political opponents that could base themselves in a social class. In Iran, under the Shah, the agricultural class was simply transformed into urban commercial (and dependent) interests through the politically judicious use of oil rents. But the spending of oil rents supports repression as well as cooptation to keep authoritarian rulers in power. Not surprisingly, then, oil dependence is closely associated with military spending and the creation of vast repressive apparatuses. This is in part due to the fact that superpowers are wary of letting oil reserves fall out of the control of their allies and into the hands of possible opposition groups. As a group, oil exporters spend much more money and a greater percentage of their revenues on their military and security forces than do non-mineral-dependent countries. For example, whereas the average developing country spends about 12.5% of its budget on the military, Ecuador, in contrast, spends 20.3% and Saudi Arabia spends a whopping 35.8%. The extent of militarization is stunning. In the decade from 1984 to 1994, for example, the share of annual military expenditures as a percentage of total central government expenditures in OPEC countries was three times that of the developed countries, and two to ten times that of the non-oil-dependent developing countries.

For these reasons, oil revenues tend to lend support, at least in the short to medium term, to whatever type of regime is in place, whether it is the occasional democracy or the more likely authoritarian ruler. Though all states may use their fiscal powers to reduce dissent through coercion or cooptation, oil wealth provides states with exceptional possibilities to do so—a phenomenon that has

been observed throughout the Middle East and in Mexico and Venezuela. Thus oil wealth is robustly associated with more durable regimes, and oil dependence is a positive predictor of durability. Even though authoritarian regimes in general are more likely to fall during economic crises, oil-based authoritarian regimes have some cushion from this general rule. Regimes such as Suharto's in Indonesia, Saddam Hussein's in Iraq, Gomez's in Venezuela, and the long-lived monarchies of the Persian Gulf (all of which lasted at least three decades) are representative of this unusual durability. Even if power shifts from one type of authoritarian rule to another (or to some form of elite democracy), political elites inherit the power that comes from control over the process of rent distribution because they control the state during windfalls, and they can consolidate this form of control through their allocative power. Thus, oil rents initially help regimes to consolidate, then enable them to endure for unusually long periods, and even enable them to persist during periods of bust.

The norm of regime stability is only part of the story. Richly detailed case studies of Nigeria, Venezuela, and Iran show that oil can help to undermine political stability over time, especially in authoritarian regimes. Virtually all oil-rich states tend to face significantly higher levels of social protest when oil revenues fall, and some of these regimes collapse. Where regimes have developed mechanisms of social control, permit rotation in power, or have sources of legitimacy that are not based on oil rents, they are more likely to endure through boom–bust cycles. But where initial oil exploitation coincides with regime and state building, non-oil-based interests do not form and patronage rents may be the main glue of the polity. Under these circumstances, these regimes are especially fragile and vulnerable during oil busts.

7. SOCIAL AND ENVIRONMENTAL IMPACTS AT REGIONAL AND LOCAL LEVELS

The exploitation of oil has a profound regional and local impact, and from the standpoint of the majority of the local population, this impact is alarming. Rather than bring prosperity to a region, as is often the claim, the boom–bust cycle associated with petroleum dependence is magnified. Over time, localities where oil is actually located, compared to the rest of the country, tend to suffer from lower economic growth, lower per capita incomes, greater dislocations, higher environmental and health hazards, and higher levels of conflict. Economically, petroleum fails to offer long-term sustainable employment alternatives at the local level, but it can seriously disrupt preexisting patterns of production. The promise of new jobs that new oil exploitation seems to offer typically attracts large numbers of migrants to an exploitation area. The rapid influx of people and the higher relative salaries of oil project workers inflate the local prices of key goods and services, bringing about a significant increase in the cost of living, even for those who do not share in the benefits of an oil project. For example, the municipality of Yopal, in the state of Casanare, Colombia, abruptly filled with migrants hoping to find employment at salaries three to four times the minimum wage, even before nearby massive oil fields at Cusiana-Cupiagua came onstream. Rents and prices increased 300%, virtually overnight. But because most jobs created by the petroleum industry are temporary or seasonal in nature, and because the growth in jobs generally occurs only during the exploration phase (land needs to be cleared or equipment transported), the industry actually offers comparatively few jobs over time. Thus, though discoveries trigger massive changes, beginning with the influx of workers seeking employment on the construction of roads, pipelines, and other infrastructure, these increased employment opportunities do not last; employment levels tend to decline dramatically when infrastructure construction is complete. These problems are compounded by the expropriation of arable land for resource extraction activity and by environmental damage, which promote a shift away from subsistence agriculture. The resulting instability in employment and income and food instability stress the local economy.

The social fabric of oil localities also changes, as disparities in income emerge and migrants pour in, often from other countries, ethnic groups, or religions. After the construction phase has been completed, the most likely local result of an oil boom (along with higher than average local inflation, increased migration, chronic underemployment, and food shortages) is increased prostitution, autoimmune disease syndrome (AIDS), and crime. Original residents who may not have been able to share in oil benefits increasingly clash with "newcomers," as they see their own ways of life greatly disrupted. This is the case of the Bakola/Bagyeli pygmies, for example, an ethnic minority in the region around Kribi,

Cameroon, who depend on forest products and hunting for their subsistence. They claim that the Chad–Cameroon pipeline construction has destroyed their medicinal plants and fishing and game areas, benefiting foreign workers and the Bantu population without providing meaningful compensation to them. The adverse impact on public health near oil localities is especially great. The migration of workers and the conditions of their housing lead to an increase in the incidences of communicable diseases, such as AIDS, other sexually transmitted diseases, tuberculosis, and cholera. Along the Chad–Cameroon pipeline, for example, temporary encampments have led to the rise of prostitution and, consequently, the appearance of human immune deficiency virus (HIV)/AIDS.

The environmental dimension of oil exploration is a chief cause of social dislocation. Hazardous wastes, site contamination, and the lack of sufficient protection of surface and subsurface waters, biodiversity, and air quality (both in the immediate vicinity of the oil project and in relation to global concerns such as ozone-depleting substances and greenhouse gases) have endangered the health of local populations near oil installations and pipelines; local livelihoods such as farming and fishing have been destroyed. Local communities have reported a sharp rise in infantile leukemia near oil facilities. This disruption is most profound among ethnic minorities and indigenous peoples who live off the land and whose customs and traditions may also be threatened. In Ecuador, the Cofan Indian Tribe has reported the contamination of its drinking supply. In Colombia, where at least 2.1 million barrels of petroleum have been spilled since 1987 (approximately 11 times as much oil as was spilled in the Exxon Valdez disaster of 1989), severe damage to this tropical ecosystem includes air pollution, land clearings, water contamination, soil erosion, sedimentation, and the disturbance of wildlife habitats. Petroleum wastes wash directly into local waterways, and Colombia's Institute of Natural Resources (INDERENA) has repeatedly condemned the presence of high concentrations of heavy metals and toxic polycyclic aromatic hydrocarbons, which are 300 times higher than drinking water standards allow in northern Colombia and 50% higher than international standards allow for oil discharges to surface waters.

The fate of the Niger Delta region, where exploration began in 1958, is the best known example of the local impact of oil exploration. Although 2 million barrels are pumped out of the Niger Delta mangrove swamps every day, providing Nigeria with a large share of its GDP, over 90% of its export earnings, and almost all its tax revenues, the people in the region have barely benefited. Despite producing energy for the country and the world, many of the Nigerians do not even have electricity. Though compensation has paid for land acquisition and oil spillages have aided some individuals from the Ogoni minority, whose land is affected, the local economy and the environment have been devastated. Gas flaring has permanently scorched the earth, destroying food crops and rendering farmlands barren. Some scientists believe that the incomplete combustion of the flares has resulted in acid rain that, in turn, has damaged crops and drinking water. Oil spillages (an average of three per month) and ruptured pipelines (either from improper maintenance or sabotage) have destroyed streams, farmlands, and aquatic life. Thousands of villagers have been killed in pipeline explosions resulting from leaks, including over 700 people in one leak alone in October 1998. This has made unlivable the Alwa Ibom community in Iko, a once-thriving economically stable and self-supporting community. By most calculations, the region remains one of the most economically backward and politically marginalized in the country. As popular protest against the activities of oil companies rises and security forces are increasingly called on for protection of facilities, it is also one of the most conflict-ridden and politically explosive regions.

8. PETROVIOLENCE AND CIVIL WAR

Natural resources and war are linked, but oil plays a special role in this relationship. Economists have found that high levels of primary commodity export dependence are associated with civil war, but petroleum dependence is even more likely to be associated with conflict compared to any other commodity. Countries dependent on oil are more likely to have civil wars than are their resource-poor counterparts; these wars are more likely to be secessionist and they are likely to be of even greater duration and intensity compared to wars where oil is not present. Evidence of this relationship is both statistical and case study based.

Because oil produces such high rents, it can be the main impetus for going to war, either directly or indirectly. Oil revenues may be the catalyst for a conflict that might not otherwise have happened. In

the Congo Republic, for example, an opposition group received $150 million funding from the French oil company, Elf Aquitaine, to support a takeover of the government so that the company could receive more favorable treatment under the new regime (which it subsequently did). The payment financed a 4-month war that resulted in 10,000 dead and the destruction of parts of Brazzaville. More frequently, the impact of oil on the outbreak of civil conflict is more indirect, i.e., the result of long-standing grievances over land expropriation, environmental damage, corruption, or the maldistribution of resources. This is especially true during bust cycles, as economic opportunities dry up. Recent civil wars and violent conflicts in oil-exporting countries have occurred in Algeria (since 1991), Angola (1975, 2002), Indonesia/Aceh (since 1986), Yemen (1990–1994), the Sudan (since 1983), Nigeria (1980–1984), Iraq (1985–1992), and the Republic of Congo (1997, 1999). Though cross border wars, e.g., the Iraqi invasion of Kuwait, have also occurred, the powerful association is with civil wars.

Secessionist wars are statistically more frequent in oil-exporting than in non-oil-exporting countries. When secessionist movements are present, the likelihood of conflict is especially high because the promise of oil wealth appears to make viable a secession that might not seem possible in poorly endowed areas. Not surprisingly, when oil is regionally concentrated and when benefits accrue to the nation while most adverse effects are local, secessionist wars are more likely. Examples abound. In the Sudan, war was triggered by President Numeiry's decision to place newly discovered oil fields, in the country's Christian south, under the control of the Muslim north. In Indonesia, the Aceh Freedom Movement has denounced the government for stealing Aceh's oil and natural gas resources as a main reason for its separatist struggle, and it has used the analogy of Brunei to convince its followers that Aceh could be equally as rich. In Nigeria, Biafra's move to secede occurred only after the government had made fiscal decisions that treated oil as a centralized, rather than a regional, asset. In this way, fights over oil revenues may become the reason for ratcheting up levels of preexisting conflict in a society.

Oil dependence is associated with particularly intense conflict. Because petroleum is so technologically sophisticated and requires so much capital, it is not easily extracted and transported; it is not "lootable" like drugs or gems. This means that it is difficult for rebels or generally unskilled groups to exploit, but governments can use this wealth to attempt preemptive repression. This is the case in Sudan, for example, where the government tried to forcibly clear entire populations away from the oil fields and the pipelines. Oil rents paid for the destruction of crops and razing of houses as well as for widespread terror against the local population. Oil's nonlootability also means that separatist conflicts (like that of the Sudan) may be especially bloody and intractable where petroleum deposits coincide with the presence of minority groups. When more straightforward fights over the distribution of oil rents between groups can be resolved by a new pattern of distribution, this is often not the case in separatist wars. But when oil is involved, such struggles are generally resolved only by the seizure of control of oil fields by rebels and a subsequent declaration of autonomy, or by the government's total defeat of the minority located near the oil fields.

Oil dependence, like dependence on other mineral resources, is associated with civil wars of long duration. Wars are expensive to pursue, and both governments and rebels can use oil rents to finance their armies. Because petroleum is transported generally through pipelines, it can be easily disrupted, and pipelines are an invitation to extortion. In Colombia, for example, oil revenues support the government's battle against rebel movements, but because petroleum must be transported to the coast through two pipelines that are both over 400 miles long, there are almost unlimited opportunities for rebels to extract "protection rests" and other forms of resources from oil companies. In 2000 alone, the pipelines were bombed 98 times and kidnappings for ransom were frequent; according to one estimate, rebel groups have managed to earn an estimated windfall of $140 million annually.

9. CONCLUSION

More than any other group of countries, oil-dependent countries demonstrate perverse linkages between economic performance, poverty, bad governance, injustice, and conflict. This is not due to the resource per se, but to the structures and incentives that oil dependence creates. Various proposals exist to mitigate this "paradox of plenty," including demands for revenue transparency by oil companies and exporting governments, revenue management schemes, stabilization funds to mitigate price shocks, reforms of taxation and civil service, and the democratization and deconcentration of both the industry and the exporting countries. Without the

implementation of reforms, the consequences of oil dependence will continue to be adverse.

SEE ALSO THE FOLLOWING ARTICLES

Development and Energy, Overview • Geopolitics of Energy • Markets for Petroleum • Nationalism and Oil • National Security and Energy • Oil and Natural Gas Exploration • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Oil Industry, History of • Oil Price Volatility • Resource Curse and Investment in Energy Industries • Sustainable Development: Basic Concepts and Application to Energy

Further Reading

Auty, R. (1993). "Sustaining Development in the Mineral Economies: The Resource Curse Thesis." Routledge, London.

Auty, R. (ed.). (2001). "Resource Abundance and Economic Development." Oxford Univ. Press, Oxford.

Beblawi, H., and Luciani, G. (eds.). (1987). "The Rentier State." Croom Helm, New York.

Chaudry, K. (1997). "The Price of Wealth: Economies and Institutions in the Middle East." Cornell Univ. Press, Ithaca, New York.

Bannon, I., and Collier, P. (eds.). (2003). "Natural Resources and Violent Conflict: Actions and Options." World Bank, Washington, D.C.

Collier, P., and Hoeffler, A. (1998). On economic causes of civil war. *Oxford Econ. Pap.* **50,** 563–573.

Gary, I., and Karl, T. (2003). "Bottom of the Barrel: Africa's Oil Boom and the Poor." Catholic Relief Services, Baltimore, Maryland.

Gedicks, A. (2001). "Resource Rebels: Native Challenges to Mining and Oil Corporations." South End Press, Cambridge, Massachusetts.

Gelb, A. (1988). "Oil Windfalls: Blessing or Curse?" Oxford Univ. Press, Oxford, United Kingdom.

Glyfason, T. (2001). Natural resources, education and economic development. *Eur. Econ. Rev.* **45**(4–6), 847–859.

Human Rights Watch. (1999). "The Price of Oil: Corporate Responsibility and Human Rights Violations in Nigeria's Oil Producing Communities." Available on the Internet at www.hrw.org.

Karl, T. (1997). "The Paradox of Plenty: Oil Booms and Petro-States." University of California Press, Berkeley, California.

Karl, T. (1999). "The perils of petroleum: Reflections on the paradox of plenty. *In* "Fueling the 21st Century: The New Political Economy of Energy." [Special edition of *J. Int. Affairs* **53** (No. 1).]

LeBillon, P. (2001). The political ecology of war: Natural resources and armed conflict. *Pol. Geogr.* **20,** 561–584.

Leite, C., and Weidmann, J. (1999). "Does Mother Nature Corrupt? Natural Resources, Corruption, and Economic Growth." IMF Working Paper WP/99/85. Inernational Monetary Fund Publ., Washington, D.C.

Mahdavy, H. (1970). Patterns and problems of economic development in rentier states: The case of Iran. *In* "Studies in the Economic History of the Middle East (M. A. Cook, ed.). Oxford Univ. Press, Oxford, United Kingdom.

Ross, M. (2001). "Extractive Sectors and the Poor." Oxfam America Report. Oxfam, London.

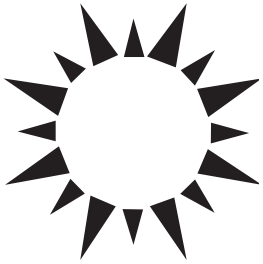
Ross, M. (2001). Does oil hinder democracy? *World Pol.* **53**

Sachs, J., and Warner, A (1999). The big push: Natural resource booms and growth. *J. Dev. Econ.* **59,** I.

Smith, B. (2004). Oil wealth and regime survival in the developing world, 1960–1999. *Am. J. Pol. Sci* (in press).

Oil Pipelines

PRASANTA KUMAR DEY
University of the West Indies
Bridgetown, Barbados



1. **Introduction**
2. **Pipeline System Design Issues**
3. **Pipe Coating System Selection**
4. **Pipeline Route Selection**
5. **Construction Management Issues**
6. **Risk-Based Inspection and Maintenance**

Glossary

corrosion One of the major causes of pipeline failure is corrosion, an electrochemical process that changes metal back to ore. Corrosion generally takes place when there is a difference of potential between two areas having a path for the flow of current. Due to this flow, one of the areas loses metal.

corrosion protection system The combination of coating and cathodic protection.

Environmental Impact Assessment (EIA) A planning aid concerned with identifying, predicting, and assessing impacts arising from proposed activities such as policies, programs, plans, and development projects that may affect the environment.

pipe coating system To protect pipelines from corrosion, pipes are coated with chemical compounds, which work as insulator. The coating system commonly uses coal tar enamel, fusion bonded epoxy (FBE), and three-layer side extruded polyethylene (3LPE) as coating materials.

pipeline failure The use of the term failure in the context of pipelines varies from country to country and across organizations. For example, in Western Europe, any loss of gas/oil is considered a failure, whereas in the United States, an incident is considered to be a failure only when it is associated with the loss of commodity and also involves a fatality or injury, or damage of more than $50,000. In India, generally any loss of commodity, however small, is considered to be a failure. Hence, failure is defined as any unintended loss of commodity from a pipeline that is engaged in the transportation of that commodity.

pipeline system Consists of oil terminal (offshore/onshore), tank farm, originating pumping station, intermediate pump station, scrapper station, pipelines connecting terminal, tank farm, and stations.

risk-based maintenance A selecting and prioritizing method of inspection and maintenance practices (tools and techniques) for facilities, equipment, and machines on the basis of their failure characteristics (likelihood and severity of failure).

Socio-Economic Impact Assessment (SEIA) A study for assessing, in advance, the social and economic consequences of the initiation of any industrial activity to the human population. SEIA forms a part of the EIA, which is mandatory for certain industrial projects in many countries.

strategic sourcing An institutional process of developing partnership among all stakeholders across the supply chain.

Oil pipelines are the nervous system of the oil industry, as they transport crude oil from sources to refineries and petroleum products from refineries to demand points. Therefore, the efficient operation of these pipelines determines the effectiveness of the entire business.

1. INTRODUCTION

Cross-country oil pipelines are the most energy-efficient, safe, environment-friendly, and economic way to ship hydrocarbons (gas, crude oil, and finished products) over long distances, either within the geographical boundary of a country or beyond it. A significant portion of many nations' energy requirements is now transported through pipelines. The economies of many countries depend on the smooth and uninterrupted operation of these lines, so it is increasingly important to ensure the safe and failure-free operation of pipelines.

Pipelines are the irreplaceable core of the U.S. petroleum transportation system and hence the key to meeting petroleum demand. Without oil pipelines, petroleum products would not reach the millions of

 © 2004 Elsevier Inc. All rights reserved.

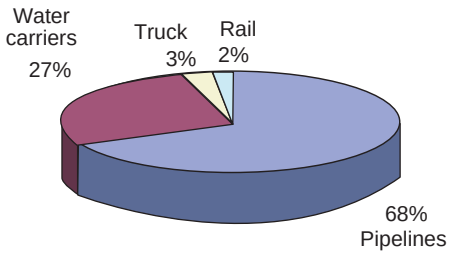


FIGURE 1 Domestic shipments of petroleum products in the United States.

consumers in all 50 states. Oil pipelines transport roughly two-thirds of the petroleum shipped in the United States; they deliver more than 14 billion barrels of petroleum per year. Figure 1 shows the domestic shipments of petroleum in the United States using various modes. Figure 2 shows both the crude and petroleum products pipelines network in the United States.

While pipelines are one of the safest modes of transporting bulk energy, and have failure rates much lower than the railroads or highway transportation,

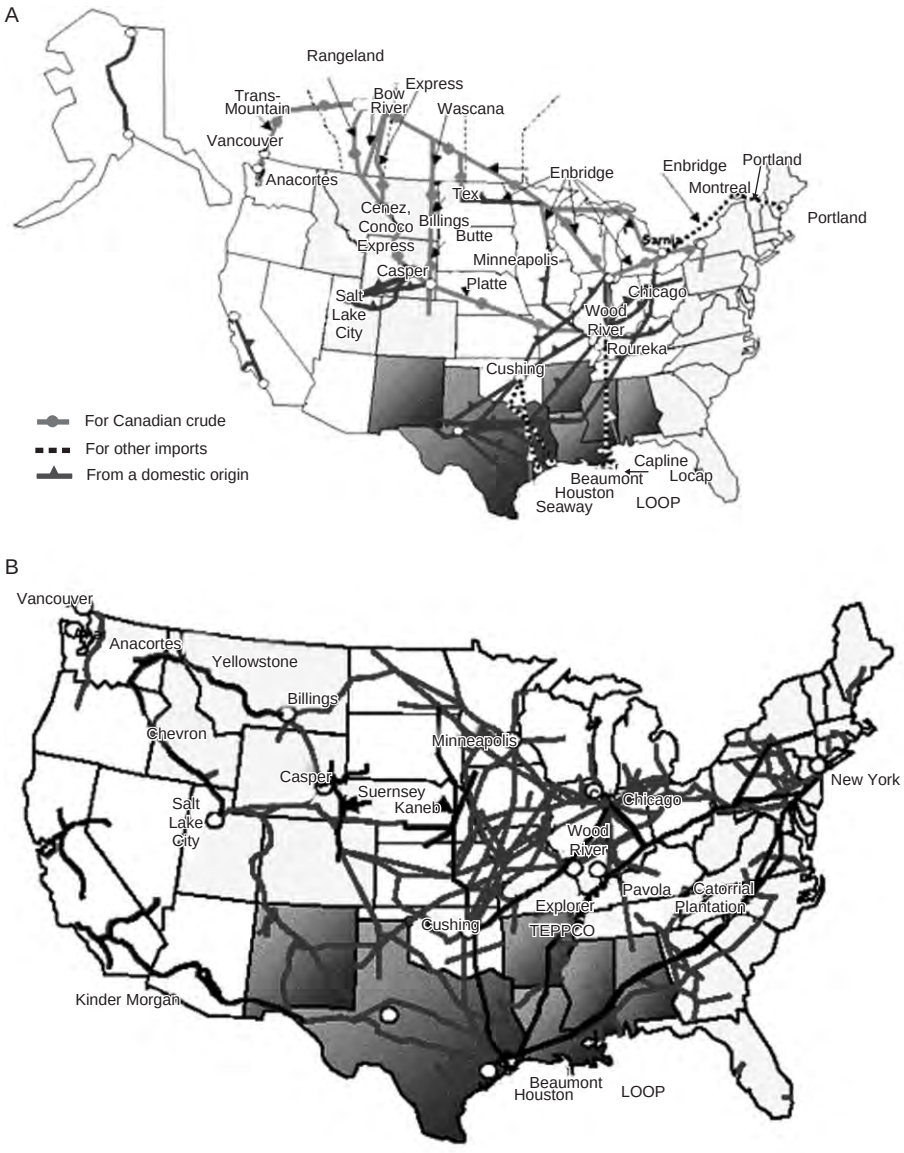


FIGURE 2 Crude and petroleum products pipelines network in the United States. The crude oil pipeline length is 55,000 miles, and the products oil pipeline length is 95,000 miles. (A) Selected crude oil truckline systems. From www.pipeline101.com. (B) Major refined products pipeline. State shading shows Petroleum Administration for Defense Districts (PADDs). From Allegro Energy Group (2001).

TABLE I

Length of Pipeline and Service (000s km) in Western Europe

	1971–1975	1976–1980	1981–1985	1986–1990	1991–1995	1996–2000
Crude	7.9	8.7	8.7	8.0	8.6	8.1
Product	7.4	8.6	9.3	15.6	20.8	21.7
Hot	0.7	0.7	0.6	0.5	0.5	0.5
Total	16.0	18.0	18.7	24.1	29.9	30.3

	The spillage frequencies by cause and their proportions shown over three decades					
	1971–1980		1981–1990		1991–2000	
Cold pipelines	Spillages per 1000 km	% of total	Spillages per 1000 km	% of total	Spillages per 1000 km	% of total
Third party	0.31	41%	0.19	38%	0.14	41%
Natural	0.04	5%	0.02	3%	0.01	2%
Corrosion	0.12	16%	0.12	24%	0.07	20%
Operational	0.06	7%	0.06	12%	0.03	8%
Mechanical	0.23	31%	0.11	22%	0.10	30%

Source. Concawe (2002).

failures do occur, and sometimes with catastrophic consequences. In 1993 in Venezuela, 51 people were burned to death when a gas pipeline failed and the escaping gas ignited. In 1994, a 36-in. (914 mm) pipeline in New Jersey failed, resulting in the death of one person and more than 50 injuries. Similar failures also have occurred in the United Kingdom, Russia, Canada, Pakistan, and India. While pipeline failure rarely cause fatalities, disruptions in operation lead to large business losses. Failures can be very expensive and cause considerable damage to the environment. The safety performance of the oil pipeline industry in United States has improved substantially. All oil pipelines together had 129 spills larger than 50 barrels (2100 gallons) in 2001. Deaths and injuries resulting from oil pipeline transportation are rare but do occur occasionally. Table I shows the statistics of spillage incidents of Western European cross-country oil pipelines over the past 30 years along with the information about the pipeline network over the past 30 years.

Achieving pipeline throughput depends on failure-free operations, which are feasible through efficient pipeline system design, appropriate coating system selection, best pipeline route selection, effective construction planning, and an optimum maintenance policy.

2. PIPELINE SYSTEM DESIGN ISSUES

Figure 3 shows a schematic diagram of a typical pipeline system. Crude oil is received from an offshore terminal/oil field and transported to the tank farm and from the tank farm to the refinery through pipelines for refining. Depending on the distance between the tank farm and refinery, hydraulic gradient, and pipe size, the number of pump stations is designed to facilitate the transportation. Refinery produces petroleum products through distillation processes and they are stored into various product tanks. From these tanks, products are pumped via products pipelines to different destination as per demand. Intermediate pump stations are designed in accordance to length of the pipeline, hydraulic gradient, and pipe size.

For a given distance in the same terrain and constant flow rate, with the increase of pipe size (diameter), the power requirement reduces as pipe friction reduces with pipe diameter. This increases capital expenditure but reduces operating cost. Therefore, pipeline operators require determining the size (diameter) of the pipeline with the consideration of life cycle costing (capital cost and operating cost) of the project. Figure 4 demonstrates the relationships between pipe diameter and the total cost of the pipeline project. Accordingly, the number of pump stations and pumping units and their capacities are derived. However, with the increase in oil demand, pipeline throughput (flow rate) must be enhanced, which can be done in one of the following ways:

- Augmenting capacity of the pumping unit (pump and prime mover), either by replacing or modifying configuration
- Constructing additional intermediate pump stations

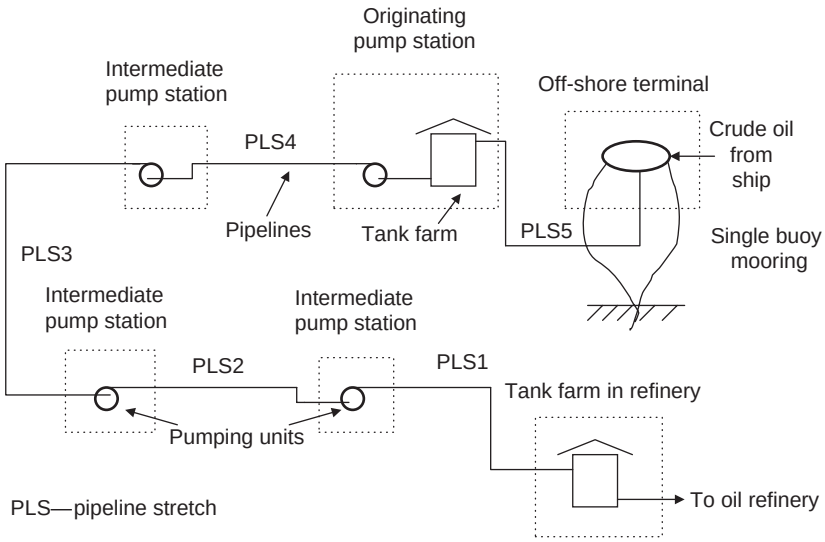


FIGURE 3 Schematic of cross-country crude oil pipelines.

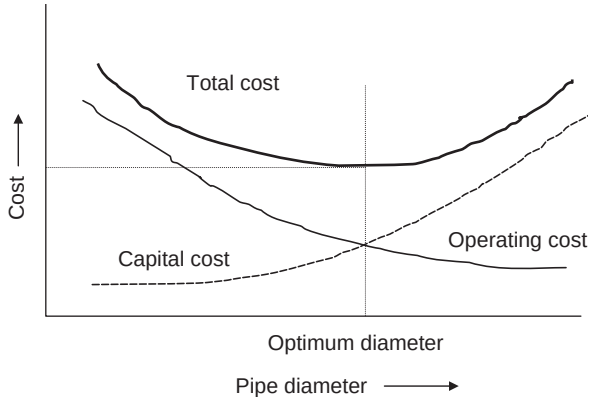


FIGURE 4 Relationship between pipeline diameter and total cost (capita and operating) of pipeline.

- Replacing the existing pipeline with a bigger pipeline
- Laying twin line/loop lines

Each method has its own pros and cons. Therefore, consideration of projected oil demand during the feasibility study leads to the design of a better pipeline system by making a rationale decision on pipe size (diameter and thickness), number of pump stations, and number and capacity of pumping units.

3. PIPE COATING SYSTEM SELECTION

Pipeline coatings are the primary means for protecting pipelines against external corrosion. In fact, the life of pipelines depends mostly on coating materials and their applications. These not only reduce the chance of pipeline failure from external corrosion considerably, but also reduce inspection and maintenance cost throughout their lives. Hence, selection of the right type of coating materials for pipelines improves the overall productivity of the pipeline organizations. Wide varieties of coating materials are available for cross-country petroleum pipeline application. The coal tar enamel (CTE), three-layer side extruded polyethylene (3LPE) coating, and fusion bonded epoxy (FBE) coatings are commonly used by pipeline operators. They are characterized by varied physical and chemical properties, and each has its own varied application procedure. Although materials and application cost are the major factors for coating selection, factors like soil characteristics, environment concern, maintainability, and safety are also getting increased attention from pipeline operators. All these factors again behave in a particular manner when cathodic protection is applied on a coated pipeline. Coating behavior will to a large extent also depend on cathodic protection parameters. Therefore, to achieve corrosion protection for the design life of a pipeline, the corrosion protection system—the combination of coatings and cathodic protection—has to be designed, specified, applied, constructed, maintained, and managed correctly.

Study shows that FBE coating is mostly employed by the pipeline operators compared to CTE and 3LPE tape coatings, primarily due to its superior performance with respect to soil stress resistance,

adhesion, and cathodic protection shielding, although FBE is cost intensive for both materials and application. It has excellent compatibility with cathodic protection and stands up well against high temperature. However, it is weak from the point of view of impact damage and moisture absorption. There are instances of poor performance of the coating system using FBE, mainly due to faulty application. In the early days of using FBE, a number of problems were found and recognized. Some of these are as follows:

- Poor adhesion on one side of the longitudinal seam weld. Caused by inadequate surface preparation and/or low temperature of application.
- Pinholes in or roughening of the coating. Caused by inadequate surface preparation (salts).
- Craters/blistering of the coating. Caused by extraneous material (e.g., rubber) trapped in the coating as it gels, or mill scale trapped in the steel and giving off gasses while coating is curing.
- Foaming. Can be caused by high levels of moisture in the powder or too high an application temperature.
- Strain lines/cracks in the coating. Caused by hydrostatic testing at very low temperatures.
- Blisters in the parent coating when induction heating field joints. Can be caused by moisture uptake, trapped moisture, contamination under the coating, or incorrect temperature profile during coating (cold ends).
- Loss of adhesion at the cutback. Can be caused by moisture uptake or contamination under the coating.
- Blistering of the coating soon after application over the circumfrential weld—joint coating. Following cellulosic welding techniques, hydrogen gas can be evolved.
- Leopard spotting in the ground. Caused by the formation of corrosion products underneath coating that has lost adhesion to the substrate, disbanded, or delaminated. This usually occurs at pinholes.

CTE coatings have a long history of pipeline protection. CTE has good performance record across the world. It has excellent electrical insulating property and low permeation properties that resist bacterial attack and resist solvent action of petroleum oils. It is less expensive compared to other types of coating materials. Pipelines with CTE coating can be easily maintained as field repair of external coating can be easily organized. A temporary operations facility can be built near the site, as manual and semiautomatic application of coating is feasible. This reduces the logistic expenditure considerably. The performance of CTE depends mostly on its application. Hence, there are instances of pipeline failures due to poor application CTE coating. Handling during storage, transportation to the site, and handling during construction also contribute pipeline failure. CTE suffers from drawbacks such as the negative environmental impact as it produces fumes, which is injurious to health during application. CTE coatings are subjected to soil stress, which can cause coating to wrinkle, crack, disband, and leave exposed steel surfaces. Additionally, if water penetrates under disbanded coatings, cathodic protection shielding can be a problem. Cathodic protection requirement generally increases as the pipelines with CTE coatings age. Operating temperature of pipelines with CTE coating should not go beyond 65°C.

The 3LPE coatings (3-layer, low density, polyethylene) called for 50 to 75 microns of fusion bonded epoxy powder, ~150 microns of a polyethylene co-polymer adhesive and an overall total film thickness of 2.3 mm using a low-density polyethylene (PE). As low-density polyethylene (LDPE) has a higher rate of moisture vapor transmission, it has more chance of failure. Adhesion is the major problem of this type of coating. Coating performance mostly depends on surface preparation, application, and storage facilities of coated pipes.

Selection of a specific coating for a pipeline is governed by the factors like coating characteristics, application complexity, environment and safety aspects, the physical facility requirement for coating application, and the cost of materials and application.

The coating is primarily characterized by its soil stress, adhesion, and cathodic protection shielding properties. Soil stress is the amount of pressure or pull applied to the coating surface after being buried. This can cause the coating to wrinkle, sag, or crack. Different coating materials have different soil stress properties. Coating that is least affected by soil stress performs better than coating that is more affected by soil stress. Adhesion is the measure of bond between the coating materials and pipe materials. Hence, coating with a better bond is treated as superior. If the coating loses adhesion to the pipe materials, corrodants enters into the annular space between the pipe surface and the disbond coating. In this situation, corrosion takes place if the disbonded coating does not permit entry of the cathodic protection current. This phenomenon is called

cathodic protection shielding and depeds on the type of coating. A coating that has lesser tendency to shield cathodic protection current is preferred.

Different coating materials have different application procedures and requirements. Pipe surface preparation is an important activity for improving adhesion. FBE and CTE require more stringent surface preparation than other types of coatings. Some coating types are vulnerable from handling and construction damage and call for advanced logistic facilities for transporting to the site for construction. Pipe-to-pipe welding onsite requires field joint coating for all types of coating. Hence, preference is given to the type coating that has better compatibility to a wide range of field coating systems. Depending on the application complexity, the coating types can be classified and prioritized. None of the coating material has infinite life and is liable to get damaged for various reasons. Hence, repair and maintenance should be taken up for effective operations of the pipeline system. The coating that is compatible with a wide range of repair and maintenance practices is preferred over others.

The physical facility requirement for the coating application is an important consideration for selecting a coating type. This depends on plant, logistics, and warehouse requirements. The selection of technology (automatic, semiautomatic, and manual) has a strong bearing on it. Pipeline operators often choose to construct a temporary coating plant near to the construction site so as to get the advantage in expensive logistics and warehousing. However, this reduces automation and, in turn, productivity. Some coating applications are not at all technically and economically feasible through temporary operations facilities. Although through superior technology, good quality coating can be produced in a competitive price through scale economy, it increases logistics and warehouse cost considerably.

Environmental friendliness and the tendency to be less hazardous to health are other factors to be considered when selecting the right coating, as all pipe coating production technologies affect the environment negatively and present considerable health hazards to the operators. Many countries have their own standard environmental regulations, which restrict the use of some coating materials. Manual application always results in more hazard to operators than automatic and semiautomatic application. Hence, even in a labor-intensive economy, preference should be given for selecting a coating that calls for less human intervention during application. Additionally, adherence to a stringent inspection requirement is another factor that can attract considerable attention to a specific coating system.

Along with these factors, materials cost and application cost are to be considered while selecting the coating type for any pipeline system. Figure 5 shows the selection model in an integrated framework.

The following steps will ensure that the pipeline coating system is effective and improves pipeline performance:

- Select the best coating system for specific project.
- Develop specifications for application.

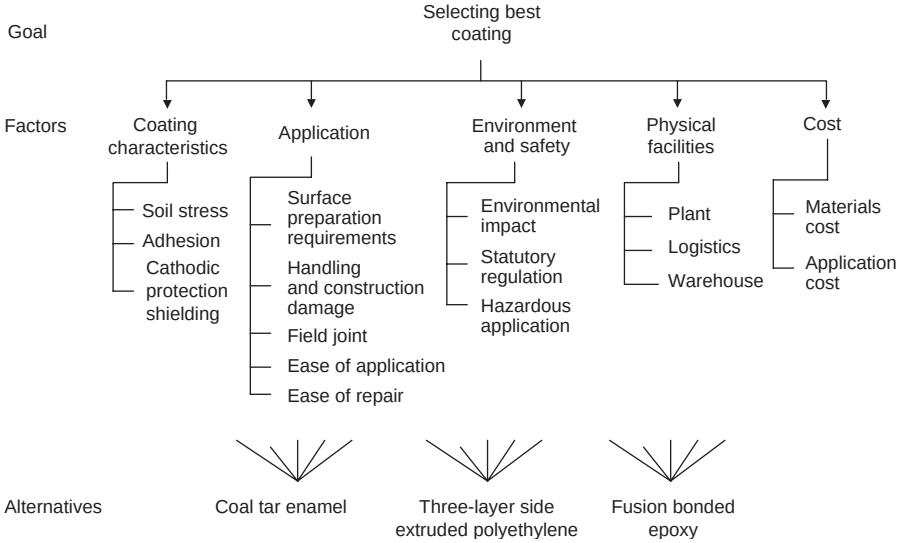


FIGURE 5 Coating system selection model.

- Customize production facilities and develop inspection procedures (project specific).
- Provide storage and logistics for coated pipes.
- Organize field joint coating and subsequent maintenance activities.

4. PIPELINE ROUTE SELECTION

Pipeline route selection plays a major role when designing an effective pipeline system, as not only the health of the pipeline depends on its terrain, but also the productivity of pipeline business depends on length, operability, constructability, and maintainability.

Cross-country petroleum pipeline route selection is governed by the following goals:

- Establish the shortest possible route connecting the originating, intermediate, and terminal locations.
- Ensure, as far as practicable, accessibility during operation and maintenance stages.
- Preserve the ecological balance and avoid/minimize environmental damage; the route should be kept clear of forests as much as possible.
- Avoid populated areas for public safety reasons.
- Keep rail, road, river, and canal crossings to the bare minimum.
- Avoid hilly or rocky terrain.
- Avoid a route running parallel to high-voltage transmission lines or D.C. circuits.
- Use an existing right of way, if possible.
- Avoid other obstacles such as wells, houses, orchards, lakes, or ponds.

These characteristics must be determined by a reconnaissance survey and the goal of finding the shortest possible route is always important. There are many examples of loss of productivity due to the wrong pipeline route selection.

A specific stretch of pipeline in eastern India (102 km, 12.75 in./324 mm in diameter, 0.25 in./6.35 mm wall thickness, API 5L X-46 pipe grade) was commissioned in September 1977. The pipeline is protected against external corrosion by an impressed current cathodic protection (CP) system, over and above the coal tar enamel external coating. The right of way includes a waterlogged of 30 km area that remains under water for 6 to 7 months in a year, with a water depth varying between 3 and four 4 m. In this 30-km stretch, extensive external corrosion has occurred. Between pipeline chainage (distance from a reference point) of 11 km to 41 km, two leaks were encountered, and at 203 locations, sleeving (reinforcement cladding) has been carried out due to heavy external corrosion. The pipeline operates at a lower throughput than its designed capacity; it is estimated that this pipeline incurs a loss of US$ 2.5 million per year on average (US$ 2.3 million in lost production due to reduced capacity operation plus US$ 0.2 million for additional maintenance cost). If the pipeline had been rerouted during the initial design phase, the capital cost would have increased by US$ 2 million (US$ 1 million for additional mainline cost, plus US$ 1 million for an additional station). The benefit of an alternative route is clear.

Another example of poor route selection is a 40-km stretch of coal belt along the route of a petroleum pipeline in eastern India. This stretch is vulnerable to failure due to third-party activities (coal mining) in the surrounding area. A risk analysis study determined that the probability and severity of a failure here were very high, with an expected cost of US$ 20 million. Rerouting the pipeline during the design stage would have cost US$ 1.25 million. Avoiding this area would have been cost-effective.

Petroleum pipelines are designed to carry a specific capacity for a specific life period and rely on a forecast of a supply-demand scenario of petroleum products. It is quite likely that during a pipeline's life span, the capacity of the line may need to be increased to maximize returns commensurate with petroleum product demand. However, if the pipeline route does not provide adequate room for augmentation, this cannot be accomplished.

Pipeline route selection should be governed not only by factors like the shortest total distance, approachability, and constructability; factors such as operability, augmentation capability, and maintainability also need to be considered. The following steps can be taken to select the best pipeline route:

- Identify a few feasible alternate pipeline routes using a geographic information system (GIS).
- Develop a database for all alternate feasible pipeline routes for analysis. Table II shows a typical database for four alternate pipeline routes.
- Identify factors that lead to selecting the best pipeline route among the alternatives. Table III shows a list of factors and subfactors that are identified by the experienced pipeline operators for a typical pipeline route selection problem.
- Select the optimum route using the preceding information with active involvement of the concerned stakeholders.

TABLE II

Pipeline Database

Description	Route I	Route II	Route III	Route IV
Throughput (MMTPA)[a]	3	3	3	3
Length (km)	780	1000	750	800
Number of stations[b]	3	3	3	3
Terrain detail (km)	430	785	570	770
Normal terrain	2	5	45	15
Slushy terrain	—	1	3	2
Rocky terrain	—	5	7	2
Forest terrain	3	4	5	1
River crossing	330	200	120	10
Populated area	15	—	—	—
Coal belt area				
Soil conditions	Less corrosive soil	Less corrosive soil	Corrosive soil for slushy terrain	Less corrosive soil
Third-party activity	More due to coal belt and populated area	More due to populated area	More due to populated area	—
Chances of pilferage	Higher due to populated area	Higher due to populated area	Higher due to populated area	—

[a] MMTPA = million metric tons per annum.
[b] 1-originating pumping station, 1-intermediate pump station, 1-terminal delivery station.

TABLE III

Factors and Subfactors

1 Pipeline length
2 Operability
  a Hydraulic gradient
  b Expansion possibility
3 Maintainability
  a Corrosion
  b Chances of pilferage
  c Third party activities
4 Approachability
  a Proximity to railway and highway
  b Terrain characteristics
5 Constructability
  a Statutory clearances from various authorities
  b Mobilization
  c Availability of power/water
  d Ease of construction
6 Environmental friendliness

- Carry out a reconnaissance survey.
- Finalize the route through a detailed survey.

The following paragraphs describe each factor and subfactor.

Pipeline length governs the capacity requirement of almost all the equipment for the entire pipeline project, as pipeline head loss is directly proportional to the length of the pipeline. Hence, the shorter the length of a pipeline, the lower the capital cost of the project and vice versa.

The hydraulic gradient is a major factor in selecting prime mover power for pipeline operations, as a negative hydraulic gradient demands higher prime mover power. Similarly, more route diversion causes more friction loss, resulting in higher prime mover power for the same throughput. These result in more capital investment.

A pipeline is designed for specific throughput in line with demand; a pipeline may need to be augmented in the future to cope with the demand for maximizing profit. Therefore, expansion/augmentation capability is one attribute of a properly designed pipeline. In addition to improving the existing prime mover capacity, a pipeline can also be augmented by installing more pumping stations along the route and laying loop lines/parallel lines. Therefore, the pipeline route should be chosen with a view to its augmentation capability.

Although pipelines are designed with adequate safety factors, they are subject to failure due to various reasons. Pipeline corrosion, pilferage, and third-party activities are factors that may create

quantum throughput loss along with chances of disaster; therefore, these factors should be carefully considered during the feasibility study. In a decision model, these factors may influence the selection of a specific route.

One of the major causes of pipeline failure is corrosion, an electrochemical process that changes metal back to ore. Corrosion generally takes place when there is a difference of potential between two areas having a path for the flow of current. Due to this flow, one of the areas loses metal.

External interference is another leading cause of pipeline failure. It can be malicious (sabotage or pilferage) or be caused by other agencies sharing the same utility corridor. The latter is known as third-party activity. In both cases, a pipeline can be damaged severely. External interference with malicious intent is more common in socioeconomically backward areas, while in regions with more industrial activity, third-party damage is common.

All activities, industrial or otherwise, are prone to natural calamities, but pipelines are especially vulnerable. A pipeline passes through all types of terrain, including geologically sensitive areas. Earthquakes, landslides, floods, and other natural disasters are common reasons for pipeline failures.

Although a cross-country petroleum pipeline is buried underground, the right of way should allow uninterrupted construction activities as well as operation, inspection, and maintenance. The ideal pipeline route should be along a railway track or a major highway. This is not always possible due to the long length of pipelines, which may require river crossings and passage through forests, deserts, and so on. Therefore, a pipeline route with better approachability gains an edge over other routes.

Laying a pipeline across states/province or national boundaries requires permission from statutory governmental authorities. Stringent safety and environmental stipulations are sometimes hindrances to project activities.

Mobilization is a major construction activity. One factor in pipeline routing is the provision for effective mobilization by the contractor. Distance to market, the availability of power and water, and the number of skilled and unskilled workers are typical requirements for starting effective construction activities.

Pipeline construction methods vary greatly with terrain conditions. For example, laying a pipeline across a river requires horizontal direction drilling (HDD), while laying pipe across a rocky area requires rock-trenching techniques. Therefore, location characteristics are a major cost component of pipeline construction. Inappropriate route selection can cause major time and cost overruns.

Poor construction, combined with inadequate inspections and low-quality materials, also contributes to pipeline failures. Other reasons include human and operational error and equipment malfunctions. Computerized control systems considerably reduce the chance of failure from these factors.

Pipelines handle hazardous petroleum products. Although pipelines are designed with safety features, failure is not uncommon. Sometimes failures result in a release of large quantities of petroleum products into the environment. If this should happen, a pipeline located in a remote area is less of a safety concern.

Figure 6 depicts a pipeline route selection model.

5. CONSTRUCTION MANAGEMENT ISSUES

Estimated miles of natural gas, crude oil, and refined-products pipelines under way or planned for construction outside the United States and Canada total 75,995 miles (122,276 km) (Pipeline and Gas Industry, 2001). Table IV shows the regional breakup of pipelines under construction or planned.

A 600-mile (1072 km) Chad–Cameroon pipeline has been planned to commission by 2004. The pipeline would increase government revenues by 45 to 50% per year and allow it to use those resources for important investments in health, education, environment, infrastructure, and rural development, necessary to reduce poverty. However, the plan suffers from tremendous negative impacts like social and ecological risk as the pipeline passes through environment-sensitive zones, loss of livelihood and resettlement, and loss of biodiversity. Moreover, the selection of a pipeline route had political intent. Although the status of the project is quite impressive, the project is vulnerable with respect to financial and political risk due to the involvement of various international organizations including the World Bank, ExxonMobil of the United States, Petronas of Malaysia, and ChevronTexaco of the United States.

The government of Georgia has authorized the construction of the Baku-Tbilisi-Ceyhan oil pipeline. The document was signed by environment minister Nino Chkhobadze in the presence of the country's president and minister of state. Authorization was given provided a number of ecological conditions are complied with; otherwise the authorization may be automatically annulled, the environment ministry

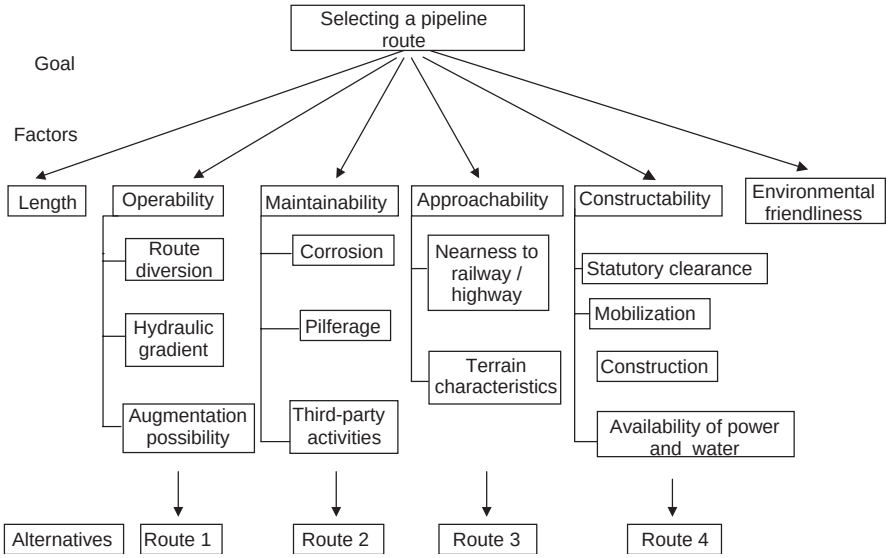


FIGURE 6 Model for selecting oil pipeline route.

TABLE IV

Outside United States/Canada – 75,995 Miles of Pipelines Under Way/Planned/Proposed

	Gas		Crude		Products		Other		Totals		
Country	**Current**	**Future**	**Current**	**Future**	**Current**	**Future**	**Current**	**Future**	**Current**	**Future**	**Totals**
Europe	2704	14,990	568	1830	40	9			3312	16,829	20,141
Middle East	878	9786		2548		284			878	12,618	13,496
Africa	1278	3442	725	511	484	500			2487	4453	6940
South Pacific	1017	6816		120		101			1107	7037	8144
Far East	649	9384	206	2341		565			855	12,290	13,145
Mexico/Central America	101	2717		41		346			101	3104	3205
South America	745	6889	477	746	590	1113	364		2176	8748	10,924
Totals	7372	54,024	1976	8137	1204	2918	364		10,916	65,079	75,995

Source. Pipeline and Gas Industry (2001).

reported. Thus, the pipeline will not be laid via sanitary zones of water bodies, including the Tsalkskoye water reservoir and Lake Tabatskuri. The same rule will be applied to the Borjomi Gorge, where popular mineral water is produced. It has not been disclosed how much longer the pipeline will become as a result. Trying to allay the public anxiety because of the pipeline being laid in the Borjomi Gorge, where famous mineral springs are situated, the president of Georgia, Eduard Shevardnadze, said in an interview on Georgian national radio that the project's investors had proposed almost "unprecedented" measures for environmental protection. This clearly indicates construction management issues of mega pipeline projects.

A cross-country petroleum pipeline construction project is characterized by the complexity of its execution with respect to lack of experience in relation to certain design conditions being exceeded (water depth, ground condition, pipeline size, etc.), the influence of external factors that are beyond human control, external causes that limit resource availability (of techniques and technology), various environment impacts, government laws and regulations, and changes in the economic and political environment. Cost and time overruns and the unsatisfactory quality of a project are the general sources of disappointment to the management of a pipeline organization.

When it comes to projects that involve worldwide resources and stakeholders, project managers know

the drill: plan early, rally support, and appease critics. Without a proactive strategy, ventures could languish and excessive costs could accrue. This is specially the case in pipeline projects. Stakeholders must be wooed before the engineering and logistics of pipelines can be considered. If authorized, the project is studied and designed before managers are in a position to recruit well-trained workers, synchronize management philosophies, and work to ensure the timely arrival of materials, parts, and equipment.

Oil pipelines are environmentally sensitive because they traverse through varied terrain covering crop fields, forests, rivers, populated areas, desert, hills, and offshore. Any malfunction of these pipelines may cause a devastating effect on the environment. Hence, the pipeline operators plan and design pipeline projects with sufficient consideration of environmental and social aspects along with the technological alternatives. Traditionally, in project appraisal, the optimum technical alternative is selected using financial analysis. Impact assessments (IA) are then carried out to justify the selection and subsequent statutory approval. However, IAs often suggest alternative sites or alternate technology and implementation methodology, resulting in a revision of the entire technical and financial analysis. This problem can be resolved by carrying out technical analysis, social assessment, and environmental impact assessment together in an integrated framework. This will help planners to select the best project among a few feasible projects. Subsequent financial analysis then justifies the selection.

Once a market need for a new pipeline is established, potential routes are determined by looking at the predicted flow of crude oil from sources to refinery, and petroleum products from refinery to demand points. Potential routes are then chosen based on pipes and construction costs, projected residents, and probable returns. Certainly, selecting the shortest distance between two points and routing along a railway or major highway is preferable, but it is not always possible because of wildlife and aquatic concerns, as well as issues involving right of way. Project managers must therefore approach the regulators with jurisdiction over such proposals and learn the potential pitfalls. Field surveys and environmental impact assessments must be performed. While the degree of permissiveness varies in different region of the world, overall there must not be any significant effect on either the natural habitat or the landowners who lease their right of way. Typically after months—if not years—of wrangling, the common solution is to reroute the pipeline to accommodate all concerns.

The success of any project is determined by whether it is completed on time, stays within budget, and achieves the requisite performance (technical requirement). The main barriers for achievement are changes in the project environment. The problem multiples with the size of the project as uncertainties in project outcome increases with size. Large-scale construction projects are exposed to uncertain environments because of such factors as planning and design complexity, the presence of various interest groups (project owner, owner's project group, consultants, contractors, vendors, etc.), resources (materials, equipment, funds, etc.) availability, the climatic environment, the economic and political environment, and statutory regulations.

Although risk and uncertainty affect all projects, size can be a major cause of risk. Other risk factors include the complexity of the project, the speed of its construction, the location of the project, and its degree of unfamiliarity.

As oil pipeline construction projects are characterized by all of these factors, a conventional approach to project management is not sufficient, as it does not enable the project management team to establish an adequate relationship among all phases of project, to forecast project achievement for building confidence of project team, to make decisions objectively with the help of an available database, to provide adequate information for effective project management, and to establish close cooperation among project team members. Moreover, oil pipeline construction should be completed on a fast track, not only to achieve faster return on investment but also to reduce the negative impact of the project, as considerable negative impacts can occur during the construction of a pipeline, especially in highly populated areas and in crop fields. To achieve success, therefore, the pipeline operators must increasingly adopt and manage construction risks effectively and manage procurement through strategic sourcing.

5.1 Construction Risk Management

Figure 7 shows a flowchart for construction risk management, and Fig. 8 shows the work breakdown structure of an oil pipeline project. The risk factors and subfactors are identified with the involvement of executives working in projects through brainstorming sessions. In the brainstorming session, executives were given a checklist of risk, which was used initially to identify risk factors and subfactors; then, through group consensus, the executives developed a

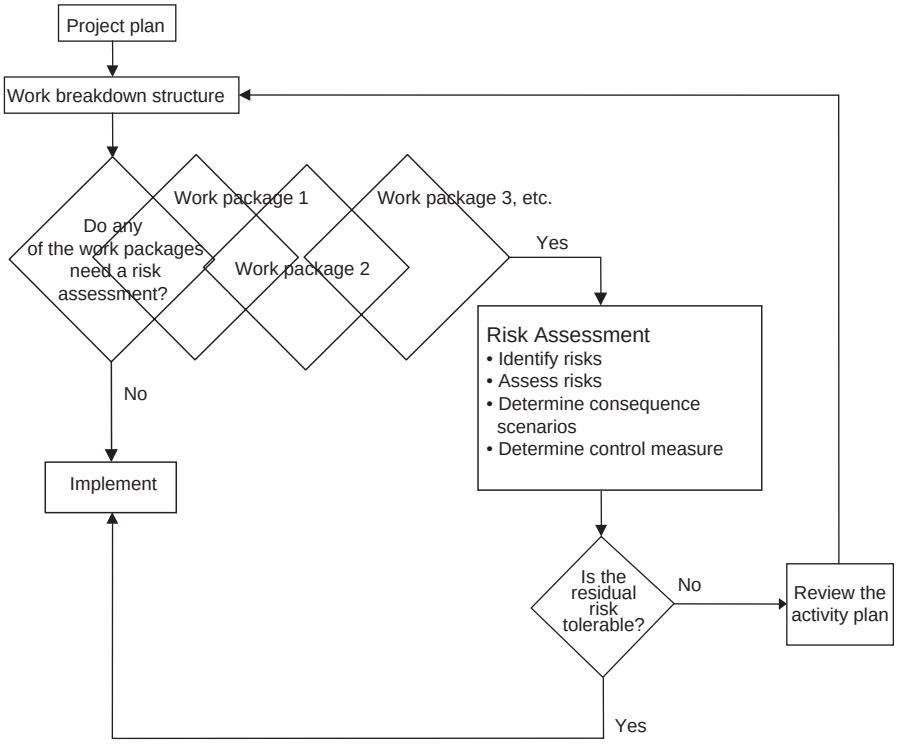


FIGURE 7 Construction risk management flowchart.

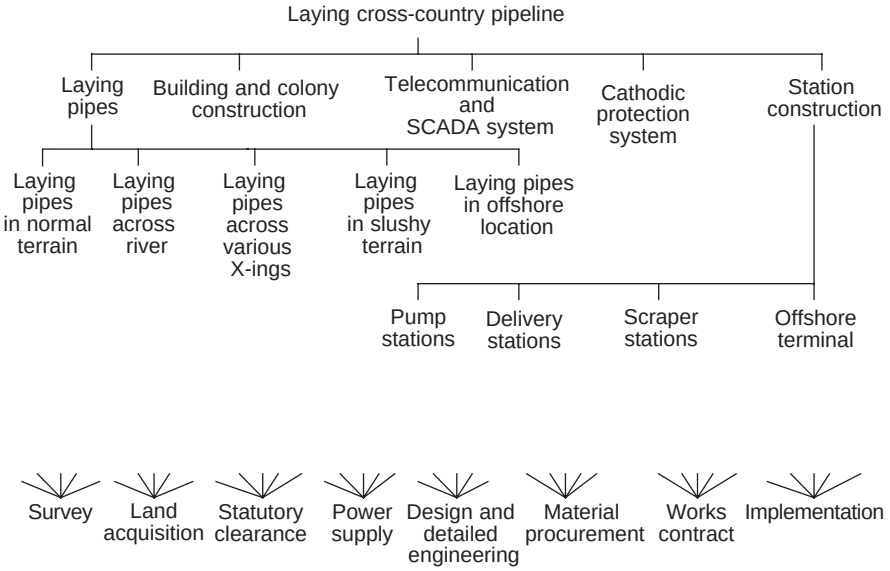


FIGURE 8 Work breakdown structure.

risk structure. The following are the risk factors and subfactors of a typical pipeline construction project:

1. Technical risk
   a. Scope change
   b. Technology selection
   c. Implementation methodology selection
   d. Equipment risk
   e. Materials risk
   f. Engineering and design change
2. Acts of God
   a. Natural calamities normal

   b. Natural calamities abnormal
3. Financial, economical, and political risk
   a. Inflation risk
   b. Fund risk
   c. Changes of local law
   d. Changes in government policy
   e. Improper estimation
4. Organizational risk
   a. Capability risk of owner's project group
   b. Contractor's failure
   c. Vendor's failure
   d. Consultant's failure
5. Statutory clearance risk
   a. Environmental clearance
   b. Land acquisition
   c. Clearance from chief controller of explosive (CCE)
   d. Other clearance from government authorities

A model for determining the risk level of a pipeline project is shown in Fig. 9.

Risk analysis consists of determining the likelihood of occurrence and severity of risk factors. Experienced pipeline executives take part in analysis, and an experienced-based method is often adopted to determine both the probability and the severity of each risk factor. The risk analysis results clearly derive responses, which are to be implemented to achieve project success.

5.2 Strategic Sourcing

Construction procurement is another issue that requires attention to achieve success in pipeline construction, as various specialized services and unique items are required from wide varieties of suppliers, contractors, and consultants across the world.

Table V demonstrates how strategic sourcing provides a paradigm shift from the typical L1 (offering work to lowest quoted bidder) environment.

In the pipeline industry, the following typical opportunities exist:

- Procurement of pipes, valves, station equipment, and construction services across a geographically diverse pipeline network not fully centralized
- Traditional approach to purchasing geared toward issuing quotes, buying from the lowest bidder, and negotiating contracts frequently
- Insufficient coordination between purchasing and engineering to determine the most cost-optimal solution
- Lack of long-term relationships with most suppliers
- Suboptimal utilization of procurement scale
- Globalization of pipeline industry

Strategic sourcing enables companies to enjoy significant cost benefits without painful headcount reductions. It views the project procurement function

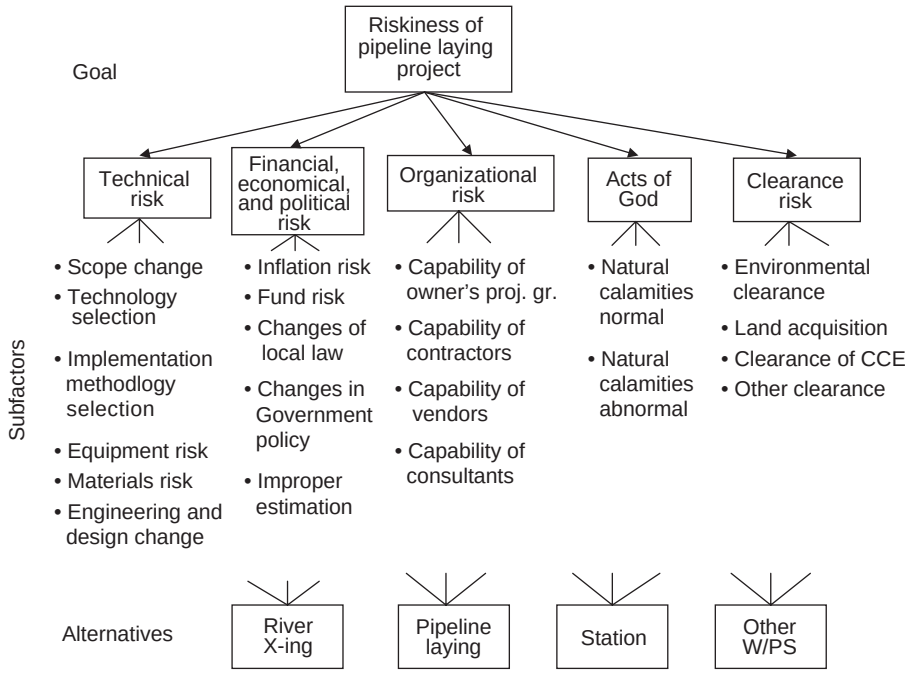


FIGURE 9 Model for determining riskiness of pipeline project.

from transactional to strategic. In the transactional perspective, procurement is viewed as being made up of people who place and transact orders, whereas in the strategic perspective, procurement is viewed as a knowledge-based buying organization. It focuses on total cost of ownership as opposed to price paid per commodity. It develops an optimal mix of supply relationships to provide competitive advantage.

A 15 to 30% cost reduction in procurement is feasible using strategic sourcing concepts. Reduction of the vendor base and consolidating volumes can yield a 5 to 15% cost improvement, and reduction of acquisition costs through structural change in the vendor relationship yields a 10 to 20% cost savings. Additionally, there are the following benefits:

- Enhancing product/service quality, reliability, and deliverability
- Leveraging supplier's technology and innovation capabilities
- Building internal sourcing capabilities that can be transferred throughout the organization.

The formation of winning alliances among client, consultants, contractors, and suppliers requires a total mindset and understanding of the changing role. Table VI demonstrates the transformation of the consultant's/contractor's/supplier's role for forming winning alliance. The winning alliances are made of partners with proven expertise, common experience, compatible business culture, and objectives.

Both global and local components are equally important for the client to bring international resources to be transformed at a local scene, and the output of the process has to serve local business needs or local market demand. Good global market coverage enables a client to establish links with potential consultants, contractors, and suppliers to procure globally. The development in technology and communication enables the client to exploit these opportunities and build global networks to gain a competitive edge.

Well-established local links provide the client with an understanding of the socioeconomic environment within which it has to operate. Alliances with local

TABLE V

Comparison of Strategic Sourcing Environment with Typical "L1" Environment

Typical "L1" environment	Strategic sourcing environment
Clearly defined requirements and specifications	Develop deep understanding of requirements — value engineer to identify optimal value/cost trade-offs
Open bid process with little/no ability for suppliers to offer alternative designs/specifications – purchase price focus	Develop deep understanding of supply industry product and service offerings, and performance drives of key suppliers
Elaborate internal controls and rigid processes to prevent changes/modifications to stated specifications	Develop robust total cost of ownership model, make buying decisions based on total cost over relevant time horizon
Typically, short-term contracts (1 year or less) that invite frequent competition and limit long-term supplier investment	True apples-to-apples comparison

TABLE VI

Transformation of the Consultant's/Contractor's/Supplier's Role

	Delivering value to a client	Coproducing value with a client	Delivering value to the consumers
Scope	Product delivery	Product and service delivery	Business function delivery
Horizon	Construction project	Product life cycle	Business project life-cycle consortia
Risk allocation	Client	Shared	Consortia
Competence	Understanding client's need	Understanding client's business	Understanding socioeconomic need
Dominant culture	Bidding culture	Cooperation	Partnering
Performance criteria	Design, construction, and product performance	Operational performance	Business and social performance
Alliances	Operational/tactical	Tactical/strategic	Strategic
Benefit of partnering	Cost savings	Improved performance	Competitive advantage
Extent of partnering	Within construction supply chain	Includes client/operator/lender	Includes stakeholders

Source. Mitrovic (1999).

consultants, contractors, and supplier organizations may also help market penetration by better addressing a customer's requirement with cost-effective design and delivery.

Close links with local and global consultants, contractors, and suppliers, preferably developed through long-term partnership relations, provide a client with a better chance to understand the business goals of other stakeholders within a dynamic business environment.

6. RISK-BASED INSPECTION AND MAINTENANCE

Traditionally, most pipeline operators ensure that during the design stage, safety provisions are created to provide a theoretical minimum failure rate for the life of the pipeline. Safety provisions are considered when selecting pipes and other fittings. To prevent corrosion, a pipeline is electrically isolated by providing high-resistance external coating materials. As a secondary protective measure, a low-voltage direct current is impressed in the pipe at a precalculated distance to transfer any corrosion that occurs due to breaks in the coating caused by a heap of buried iron junk, rails, and so on. This is called impressed current cathodic protection. The quality of the commodity that is being transported through the line is also ensured, and sometimes corrosion-preventing chemicals (corrosion inhibitors) are mixed with the commodity. To avoid deliberate damage of the pipeline in isolated locations, regular patrolling of the right of way from the air as well as on foot is carried out, and all third-party activities near the route are monitored.

Various techniques are routinely used to monitor the status of a pipeline. Any deterioration in the line may cause a leak or rupture. Modern methodologies can ensure the structural integrity of an operating pipeline without taking it out of service.

The existing inspection and maintenance practices commonly followed by most pipeline operators are formulated mainly on the basis of experience. However, operators are developing an organized maintenance policy based on data analysis and other in-house studies to replace rule-of-thumb-based policies. The primary reasons for this are stringent environmental protection laws, scarce resources, and excessive inspection costs. Existing policies are not sharply focused from the point of view of the greatest damage/defect risk to a pipeline. The basis for selecting health monitoring and inspection techniques is not clear to many operators. In many cases, a survey is conducted over an entire pipeline or on a particular segment, when another segment needs it more. Avoidable expenditures are thus incurred.

A strong motivation exists, therefore, to derive a technique that will help pipeline operators select the right type of inspection/monitoring technique for segments that need it. A more clearly focused inspection and maintenance policy that has a low investment-to-benefit ratio should be formulated.

Figure 10 demonstrates an inspection and maintenance system of oil pipelines and its relationship with pipeline operations. The following steps demonstrate a risk-based inspection and maintenance method for oil pipelines.

Step 1: Entire oil pipeline is classified into a few stretches.
Step 2: All information related to the pipeline, including the terrain detail under study, is prepared and documented section wise.
Step 3: The risk factors that can cause failures are identified. Generally, pipelines fail because of one of these reasons: corrosion (internal and external), free span, external interference, construction and materials defects, acts of God, or human and operational error.

One of the major causes of pipeline failure is corrosion, an electrochemical process that changes metal back to ore. Corrosion generally takes place when there is a difference of potential between two areas having a path for the flow of current. Due to this flow, one of the areas loses metal.

Internal corrosion is normally caused by chemical reaction between the pipeline material and the fluid or gas, such as CO_2, H_2S, and O_2. Corrosion can take the form of general or localized metal loss from pipe, and it may also give rise to cracking in pipeline material. The rate of corrosion depends on pipe materials, type of products being transported through pipelines, and corrosion inhibitor. The four major reasons for internal corrosion are as follows:

- Sweet corrosion is caused by the presence of carbon dioxide dissolved in the fluids, also called carbonic acid corrosion. This form of corrosion is typically slow, localized, and has pitting attack on pipeline metal.
- Sour corrosion is caused by hydrogen sulfide in the fluids. This form of corrosion causes rapid failure by cracking of the pipe metal.

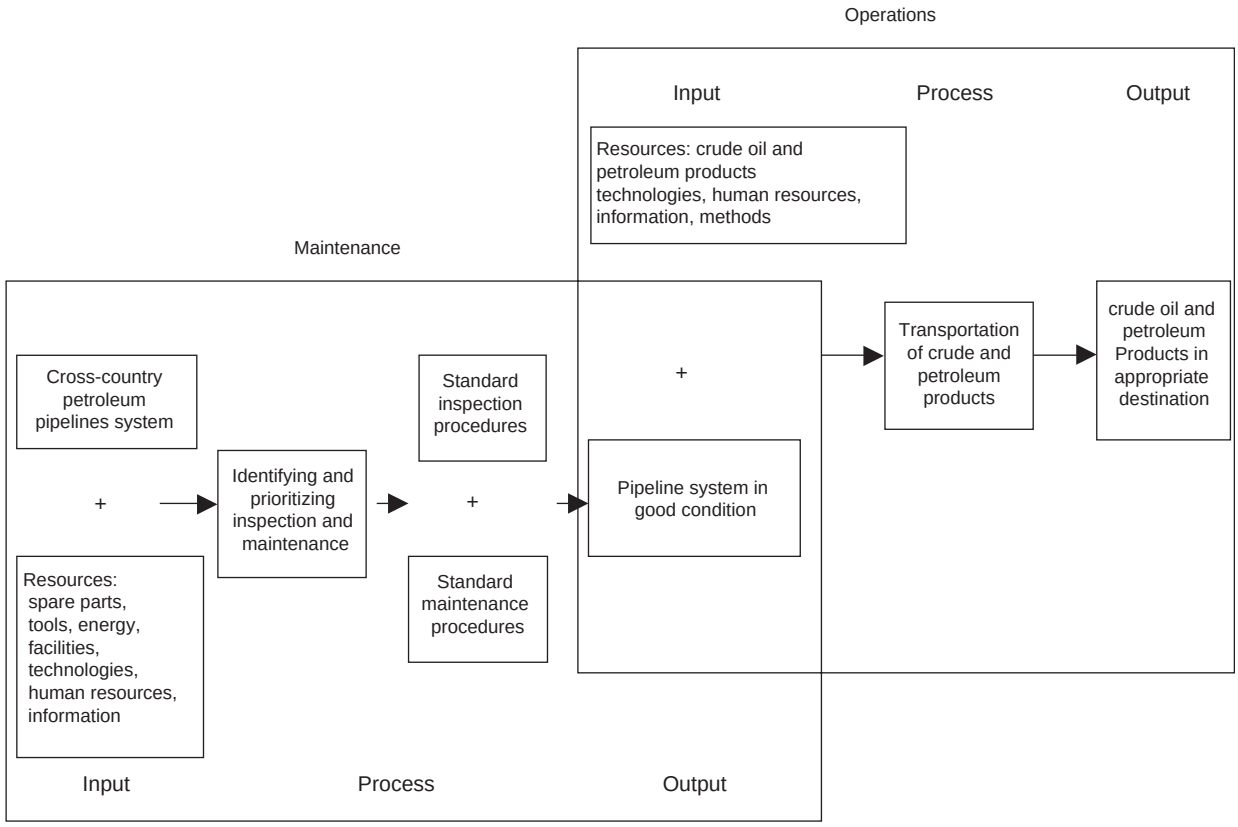


FIGURE 10 Relationship between operations and maintenance of cross-country petroleum pipeline system.

- Corrosion mainly results from the presence of oxygen, carbon dioxide, and hydrogen sulfide in the water, which comes from some oil fields.
- Microbiological corrosion results from the growth of bacteria in the pipeline.

Chemical reaction between pipe metal and seawater causes an external corrosion. The rate of external corrosion depends on the condition of the coating and cathodic protection. The external corrosion also includes external erosion, which is caused by solid substances in the seawater when they come in contact with pipelines.

A free span is another criteria that can cause pipeline failure. A free span is an area where the pipe does not have support because of subsidence. If the free span is long, pipe self-weight and earth load for onshore pipeline, and pipe self-weight and ocean currents might damage the pipeline.

External interference is another leading cause of pipeline failure. It can be malicious (sabotage or pilferage) or may be caused by other agencies sharing the same utility corridor. The latter is known as third-party activity. In both cases, a pipeline can be damaged severely. External interference with malicious intent is more common in socioeconomically backward areas, while in regions with more industrial activity, third-party damage is common.

All activities, industrial or otherwise, are prone to natural calamities, but pipelines are especially vulnerable. A pipeline passes through all types of terrain, including geologically sensitive areas. Earthquakes, landslides, floods, and other natural disasters are common causes of pipeline failures.

Poor construction, combined with inadequate inspections and low-quality materials, also contributes to pipeline failures. Other reasons include human and operational error and equipment malfunctions. Computerized control systems considerably reduce the chance of failure from these factors.

Human and operational errors are another sources of pipeline failure. Inadequate instrumentation, a foolproof operating system, lack of standardized operating procedures, untrained operators, and so on are the common causes of pipeline failure due to human and operational errors.

Step 4: The next step is to form of a risk structure (Fig. 11) based on the identified risk factors.

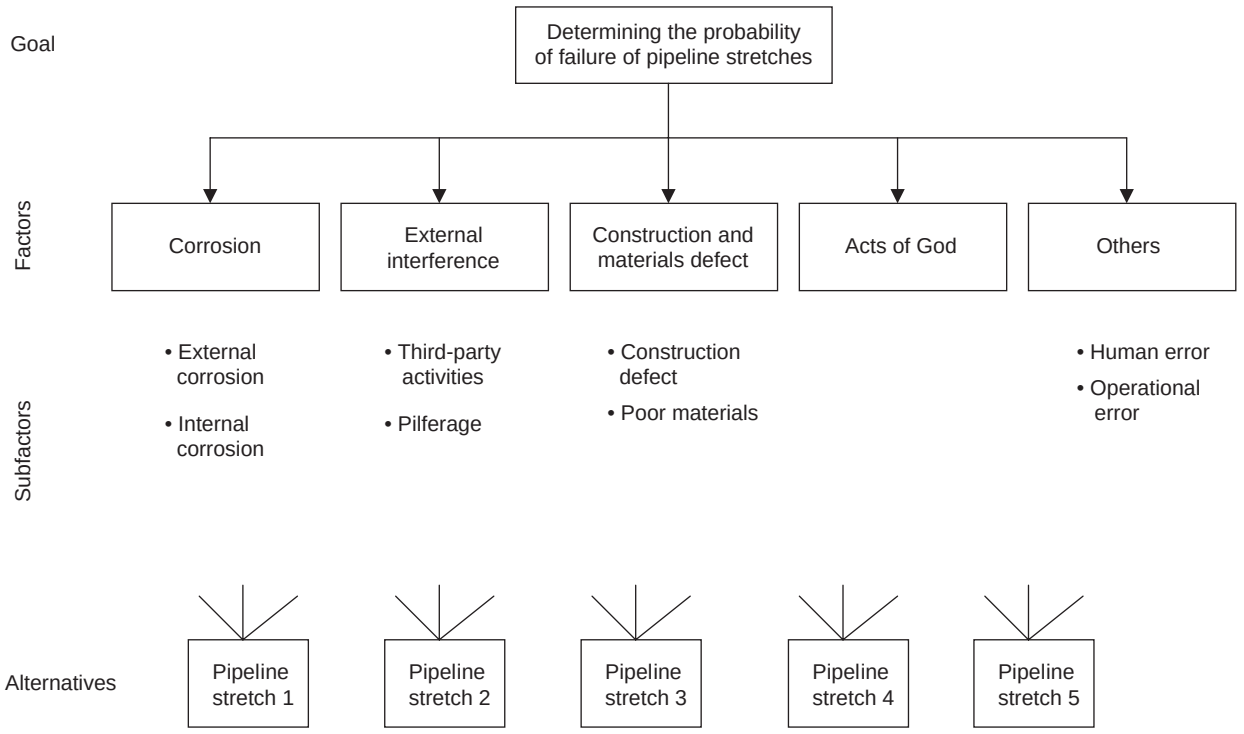


FIGURE 11 Risk structure to determine failure probability of various stretches.

Step 5: In this step, the likelihood of pipeline failure due to each factor is determined. The alternative pipeline stretches are compared with respect to each risk factor to determine the likelihood of failure for each pipeline stretch.

Step 6: In this step, specific inspection/maintenance requirements are determined for specific segments of pipelines from the likelihood of failure data; this is done to mitigate risk.

Step 7: The last step demonstrates the cost-benefit analysis of a suggested inspection and maintenance strategy along with a cost-effective insurance plan for the pipeline. The basis of the insurance premium depends on the likelihood of its failure, the expected failure cost in a given period, the risk perception of the management/organization, and the inspection/maintenance programs undertaken.

Advantages of this risk-based inspection and maintenance model are as follows:

- Reducing subjectivity in the decision making process when selecting an inspection technique
- Identifying the right pipeline or segment for inspection and maintenance
- Formulating an inspection and maintenance policy
- Deriving the budget allocation for inspection and maintenance
- Providing guidance to deploy the right mix of labor in inspection and maintenance
- Enhancing emergency preparations
- Assessing risk and fixing an insurance premium
- Forming a basis for demonstrating the risk level to governments and other regulatory agencies

If a pipeline system is designed, constructed, and operated ideally, many inspection and maintenance problems will not crop up. The overall performance of pipeline operations and maintenance would be improved through the following actions:

- Pipeline routes are to be decided on the basis of the life cycle costing approach, not on the basis of the shortest route.
- The maintenance characteristics of the pipeline are to be considered along with pressure and temperature parameters while designing pipe thickness for various stretches of pipeline.
- Pipeline coating shall be selected on the basis of the terrain condition, the environmental policy of the organization, the cost of coating materials, the construction methodology, inspection, and the maintenance philosophy.

- The construction methodology of a pipeline in a critical section is to be formulated during the feasibility stage of the project, and this shall commensurate with the design and operational philosophy of the pipeline as a whole. Factors such as availability of technology, availability of consultants, contractors and vendors, experience of owner project group, government regulations, and environmental requirements throughout the life of pipeline are to be rationally considered when selecting the best construction methodology.
- Networking in pipeline operations demands a foolproof mechanism in the system for minimizing operational and human errors. Improved instrumentation shall be designed that is commensurate with the design philosophy of the entire pipeline system.
- All pipeline operators are to be suitably trained in pipeline operation before taking charge of specific pipelines. Pipeline simulation training may be one of these areas. Criticality of pipelines and the expertise of personnel are to be considered when finding staff for pipeline operations.

SEE ALSO THE FOLLOWING ARTICLES

Crude Oil Releases to the Environment: Natural Fate and Remediation Options • Crude Oil Spills, Environmental Impact of • Occupational Health Risks in Crude Oil and Natural Gas Extraction • Oil and Natural Gas Drilling • Oil and Natural Gas: Offshore Operations • Oil Industry, History of • Oil Recovery • Oil Refining and Products • Petroleum System: Nature's Distribution System for Oil and Gas • Public Reaction to Offshore Oil

Further Reading

Awakul, P., and Ogunlana, S. O. (2002). The effect of attitudinal differences on interface conflicts in large scale construction projects: A case study. *Construction Manage. Econ.* **20,** 365–377.

Calvin, H., and Dey, P. K. (2002). Social impact assessment of a sewerage project in Barbados. *Impact Assess. Proj. Appraisal* **20**(2).

Dey, P. K., and Gupta, S. S. (2000). Decision support system yields better pipeline route. *OilGas J.,* PennWell, **98**(22), 68–73.

Dey, P. K., and Gupta, S. S. (2001). Risk-based model aids selection of pipeline inspection, maintenance strategies. *Oil Gas J.,* PennWell, **99**(28), 39–67.

Dey, P. K. (2003). Analytic hierarchy process analyses risk of operating cross-country petroleum pipelines in India. *Natural Hazard Rev.,* ASCE, **4**(4), 213–221.

Dey, P. K. (2002). An integrated assessment model for cross-country pipelines. *Environmental Impact Assess. Rev.* **22**(6), 703–721.

Dey, P. K. (2001). Decision support system for project risk management: A case study. *Manage. Decision,* MCB University Press **39**(8), 634–649.

Dey, P. K., and Gupta, S. S. (2001). Feasibility analysis of cross-country pipeline projects: A quantitative approach. *Project Manage. J.* **32**(4), 50–58.

Dey, P. K., Tabucanon, M. T., and Ogunlana, S. O. (1994). Planning for project control through risk analysis: A case of petroleum pipeline laying. *Int. J. Project Manage.* **12**(1), 23–33.

Dey, P. K., Tabucanon, M. T., and Ogunlana, S. O. (1996). Petroleum pipeline construction planning: A conceptual framework. *Int. J. Project Manage.* **14**(4), 231–240.

Dey, P. K. (2001). Integrated approach to project feasibility analysis. *Impact Assess. Proj. Appraisal* **19**(3), 235–245.

Dey, P. K. (February, 2004). Decision support system for inspection and maintenance of cross-country petroleum pipeline. *IEEE Transactions Engineering Manage.*, in press.

Lyons, D. (February, 2002). Western European cross-country oil pipelines 30-year performance statistics, Report no. 1/02, CONCAWE, Brussels.

Mitrovic, D. (1999). Winning alliances for large scale construction projects on the world market. *In* "Profitable Partnering in Construction Procurement" (S. O. Ogunlana, Ed.), pp. 189–199. E & FN Spon. London and New York.

Montemurro, D., and Barnett, S. (1998). GIS-based process helps Trans Canada select best route for expansion line. *Oil Gas J.,* **22**.

Ogunlana, S. O., Yisa, S., and Yotsinsak, T. (2001). An assessment of people's satisfaction with the hearing on the Yadana gas pipeline project. *J. Environ. Monitoring Assess.* **72,** 207–225.

Pipeline and Gas Industry (2001). "Worldwide Pipeline Construction Report," Vol. 84, no. 11, November–December.

Ramanathan, R., and Geetha, S. (1998.). Socio-economic impact assessment of industrial projects in India. *Impact Assess. Proj. Appraisal* **16**(1), 27–32.

U.S. Department of Transportation (1995). *Pipeline Safety Regulation,* October 1.

Oil Price Volatility

M. J. HWANG
West Virginia University
Morgantown, West Virginia, United States

C. W. YANG
Clarion University of Pennsylvania
Clarion, Pennsylvania, United States

B. N. HUANG
Providence University
Shalu, Taiwan

H. OHTA
Aoyama Gakuin University
Tokyo, Japan

1. **Introduction**
2. **Microeconomic Foundation of the Price Elasticity of Demand and Its Estimation**
3. **Volatility of an Oil Price Change and Macroeconomic Activity**
4. **Conclusion**

Glossary

crude oil price Crude oil is sold through many contract arrangements, in spot transactions, and on futures markets. In 1971, the power to control crude oil prices shifted from the United States to OPEC when the Texas Railroad Commission set prorating at 100% for the first time.

demand elasticity The relative responsiveness of quantity demanded to changes in price. The price elasticity of demand is important in affecting the pricing behavior of OPEC.

economic impact The study of oil price shocks and their effects on economic activities.

oil price volatility The volatility of crude oil prices is influenced more by both demand structure and shifting demand and supply conditions and less by the cost of producing crude oil. Oil price volatility creates uncertainty and therefore an unstable economy.

Organization of Petroleum Exporting Countries (OPEC) Formed in 1960 with five founding members: Iran, Iraq, Kuwait, Saudi Arabia, and Venezuela. By the end of 1971, six other countries had joined the group: Qatar, Indonesia, Libya, United Arab Emirates, Algeria, and Nigeria. OPEC effectively controlled oil prices independent from the United States during the Arab oil embargo.

The world oil price has been extremely volatile in the past three decades. The cartel pricing is largely affected by the aggregate demand the cartel faces and related elasticity. The stable and unstable cases of price elasticity of demand are investigated in this article to shed light on the seemingly mysterious Organization of Petroleum Exporting Countries pricing behavior. We estimate the elasticity of oil demands in the U.S. market (the world's largest energy consumer) and use this information to probe and predict movements in the market price of crude oil. The volatility of crude oil prices creates uncertainty and therefore an unstable economy. Employing recent data, our empirical results indicate that a higher oil price seems to have a greater impact on the stock market than on the output market.

1. INTRODUCTION

The world oil price has been extremely volatile in the past three decades. It was as low as $2.17 per barrel in 1971 but spiked to $34 in 1981. It soared to approximately $40 per barrel toward the end of February 2003. The Organization of Petroleum Exporting Countries (OPEC) price increases of the

 © 2004 Elsevier Inc. All rights reserved.

1970s drove Western industrialized economies into recession. During the 1950s and 1960s, many competitive independent producers characterized the crude oil industry. The demand facing the individual producer is relatively elastic, even though the demand curve for the entire industry is rather inelastic. The price of crude oil was close to the marginal cost during that period of time. OPEC emerged as an effective cartel in approximately 1971 when it successfully raised the pattern of world prices with the Tehran and Tripoli agreements. With the Arab–Israeli war of 1973, its consequent oil embargo, and the nationalization of oil production in member countries, the structure of the cartel provided the means to raise prices substantially from $4.10 per barrel in 1973 to $11.11 in 1974 to reap a monopoly profit.

With the beginning of political problems in Iran, another large increase in oil prices occurred in 1978 when Iranian production declined from a peak of 6 million barrels per day to 0.5 million barrels. Even though half of the reduction in Iranian oil was offset by expanded production by other OPEC producers, the effect was immediate and substantial, causing the price to increase to $13.49 in 1978, because the elasticity of demand was rather inelastic in the short term. Because of the war between Iraq and Iran, Iraq's crude oil production declined by 2.7 million barrels per day and Iran's production declined by 600,000 barrels per day. As a result, the price of crude oil more than doubled from $13.49 per barrel in 1978 to $34 in 1981.

Higher oil prices create inflationary pressure and slow worldwide economic activities. The recession of the early 1980s reduced demand for oil. From 1982 to 1985, OPEC tried to stabilize the world oil price with low production quotas. These attempts were unsuccessful because various members of OPEC produced beyond their quotas, causing crude oil prices to decrease below $10 per barrel by mid-1986. In particular, Saudi Arabia's increases in oil production frequently depressed the oil price. The price of crude oil remained weak until the start of the Gulf War in 1990, when it eclipsed $40 per barrel. Because of the uncertainty associated with the invasion of Kuwait by Iraq and the ensuing Gulf War, the oil price spiked again to $34 in late 1990. After the war and the recession of the early 1990s, the crude oil price began a steady decline. The economy, however, started to turn around in 1994. With a strong economy in the United States and a booming economy in Asia, increased demand led to a steady price recovery well into 1997. The financial crisis and subsequent economic setbacks in Asia started in 1997 and oil prices plummeted to approximately $10 in late 1998 and early 1999. With the recovery in Asia and the decrease in oil quotas by OPEC, the price has recovered to approximately $30 a barrel in recent years. In 2002, for the first time, Russia's oil production surpassed that of Saudi Arabia, signaling a more complicated pricing scheme for oil.

The literature on the volatility of crude oil prices relates oil price changes either to the effect of the price elasticity of demand or to the instability of the market structures. It is apparent that the stable price is established through the equilibrium of total world demand and supply, including OPEC and non-OPEC production. In the short term, the price change is largely impacted by the immediate substantial increase or decrease in oil production from OPEC members and political events. The cartel's pricing policy is largely affected by the aggregate demand it faces and related elasticity.

The volatility of crude oil prices in turn creates uncertainty and therefore an unstable economy. In his pioneering work, Hamilton indicated that oil price increases have partially accounted for every U.S. depression since World War II. Many researchers, using different estimation procedures and data, have tested the relationships between the oil price increases and many different macroeconomic variables. Using a multiequation statistical approach incorporating the U.S. interest rate, oil price, industrial production, and real stock returns with daily data from 1947 to 1996, Sadorsky found that oil price volatility does have an impact on real stock returns. Recently, emphasis has shifted to the asymmetry of the impact of oil price shocks on economic activities and on stock markets. Mork was the first to provide the asymmetry of oil price shocks or its volatility on economic activities. Using data from industrial nations, Mork and Olson verified that there is a negative relationship between an oil price increase and national output, whereas no statistical significance can be attributed to them when the oil price declines. Lee *et al.* estimated oil price shocks from a generalized econometric model and investigated the impacts of positive and negative oil price shocks on economic activities. They came to the same conclusion that positive shocks have a statistically significant impact on economic activities, whereas negative shocks have no significant impact.

This article examines the volatility of crude oil prices by first determining the potential maximal price that OPEC can extract based on the microeconomic foundation of the elasticity theory proposed by

Greenhut *et al*. The market structure of OPEC, the stable and unstable demand structure, and related elasticity of demand are discussed. In particular, the theory of unstable price elasticity of demand helps explain some of the pricing behavior of OPEC. The price elasticity of demand is then estimated to shed light on the volatility of oil prices. This article further investigates the significance of changing oil prices on the economy by examining the relationship between oil price volatility and industrial production and/or the stock market.

2. MICROECONOMIC FOUNDATION OF THE PRICE ELASTICITY OF DEMAND AND ITS ESTIMATION

Consider a cartel (e.g., OPEC) whose objective is to maximize joint profit:

$$\pi = PQ - \sum_{i=1}^{n} \mathrm{TC}_i(q_i) = PQ - \mathrm{TC}(Q), \qquad (1)$$

where P is a uniform cartel price with $P' < 0$, $Q = \Sigma q_i$ is the sum of outputs of n cartel members (q_i), and TC_i is the total cost function of cartel member i. Note that OPEC behaves much like a monopoly or an effective cartel despite occasional squabbles over output quotas. Under recent OPEC arrangements, if the price becomes too low, OPEC would reduce output by half a million barrels per day ($\Delta Q = 500{,}000$) at one time. A 500,000-barrel decrease (increase) in Q is approximately 2% of the total cartel output, and each member must accept a 2% decrease (increase) in its quota in the case of a tightly controlled cartel. Within this framework, it is the total cartel output Q, instead of q_i, that plays a crucial role in the pricing decision.

The first-order condition requires

$$\pi' = QP' + P - \mathrm{MC} = 0, \qquad (2)$$

where $\pi' = d\pi/dQ$ is the marginal profit.

Equation (2) states that the marginal revenue (MR) equals the common cartel marginal cost (MC, or horizontal summation of marginal cost curves for all cartel members) in equilibrium. Note that if some MC_i's exceed the going market price, these members have no role to play in the pricing decision. However, this is not likely the case for OPEC because production costs fall far short of the market price.

Substituting $\mathrm{MR} = P(1-1/e)$ into Eq. (2) yields

$$P(1 - 1/e) = \mathrm{MC}$$

or

$$P(1 - 1/e) = P/k, \qquad (3)$$

where $k = P/\mathrm{MC}$. Solving for the profit maximizing price elasticity leads to

$$e^* = k/(k - 1). \qquad (4)$$

The second-order condition plays a critical role in describing the switching demand conditions. Here, we classify two major demand cases via expanding the elasticity theory formulated by Greenhut *et al*. in terms of the second-order condition.

By using the chain rule, we have

$$\frac{d^2\pi}{dQ^2} = \frac{d\pi'}{dQ} = \frac{dP}{dQ}\left(\frac{\eta - (1-e)}{e} - \frac{-e}{\beta k}\right) < 0, \qquad (5)$$

where $\beta = (dQ/d\mathrm{MC})(\mathrm{MC}/Q)$ is the output elasticity on the cartel marginal cost curve, and the price elasticity of elasticity $\eta = (de/dP)(P/e)$ measures the percentage change in price elasticity from the percentage price change. Alternatively, by taking the derivative of the marginal profit π' with respect to price P, the second-order condition in Eq. (5) appears as

$$\frac{d\pi'}{dP} = \frac{\eta - (1-e)}{e} + \frac{e}{\beta k} > 0. \qquad (6)$$

For a given output elasticity β, if the marginal cost is an insignificant portion of price (i.e., a large k) or the marginal cost is constant ($\beta \to \infty$, as is true in a large-scale crude oil production with a sizable fixed cost), the second term on the right side of Eq. (6) plays a trivial role and can be ignored. It follows that $d\pi'/dP > 0$ for $d\pi'/dQ < 0$, and it is sufficient that relation A holds:

$$\frac{d\pi'}{dP} > 0 \qquad \text{if } \eta > 1 - e,\ \eta = 1, \text{ or } \eta > 1. \qquad (\mathrm{A})$$

The residual term $e/\beta k$ would reinforce the relation, but its magnitude may be insignificant, with k and/or β being large enough.

On the other hand, the unstable relation B would follow if the second-order condition is not satisfied:

$$\frac{d\pi'}{dP} < 0 \qquad \text{if } \eta < 1 - e. \qquad (\mathrm{B})$$

Relation A indicates that no matter what the value of e is, the marginal profit π' will increase with price less than, proportionately equal to, or more rapidly than price if and only if $1-e < \eta < 1$, $\eta = 1$, or $\eta > 1$, respectively. If prior market conditions establish a price at a level at which the elasticity of demand is less than unity and MR is below MC (i.e., the marginal profit being negative), an OPEC price hike

could be favorable. Under relation A, an increasingly elastic demand at higher prices (e.g., $\eta > 0$) would generally create conditions in which alternative fuels become more competitive following the price hike. The desired result under profit-maximizing principles is to have the marginal profit increase to 0 and elasticity close to $e^* = k/(k-1)$ depending on the ratio k ($k = P/\mathrm{MC}$).

Since the marginal cost for OPEC countries is very low relative to price, and k is accordingly large, e^* should be very close to 1. Under the condition $\eta = 1$ or $\eta > 1$, the increase in elasticity at a higher price is proportionately equal to or greater than the increase in price, and we reach the point at which negative marginal profit increases to zero more rapidly than under $1 - e < \eta < 1$. On the other hand, if the price elasticity of demand is greater than unity and MR is greater than MC, it becomes advantageous for OPEC to lower its price. In general, the elasticity of demand would converge toward $e^* = k/(k-1)$ and the marginal profit to zero. That is, the market system would adjust itself under stable relation A.

Relation B indicates that if $\eta - (1 - e) < 0$, the marginal profit will decline (increase) as the price of crude oil is raised (lowered). This relation is an unusual and basically unstable case. Relation B also implies that the second-order condition in Eqs. (5) or (6) is not satisfied. The elasticity will therefore diverge from $e^* = k/(k-1)$. A cartel can benefit from an increase in price if demand is inelastic. Under the unstable relation B, the elasticity of demand decreases markedly as the price is raised. Inasmuch as marginal profit decreases and is negative, output will be curtailed and the price will be raised further. This can occur only in unusual circumstances (e.g., the energy crisis of 1973–1974). It is not a mathematical curiosity to have relation B—that is, an unusually strong convex demand curve. This could be a direct result of a structural break in the world oil market from a preembargo competitive oil market to an after-embargo cartel market. For example, the market demand changed from an elastic competitive market demand (before the 1973 embargo) to an inelastic cartel demand (after the embargo, from point F to C in Fig. 1).

The regime shift occurred when OPEC emerged as an effective cartel in 1971 after the Tehran Agreement. The ensuing Arab–Israeli War of 1973, subsequent oil embargo, and the nationalization of oil production in member countries enabled the cartel to raise the price of crude oil from $2 per barrel in 1971 to $34 in 1981. The price skyrocketed again to $37 in 2000 and $39.99 in late February

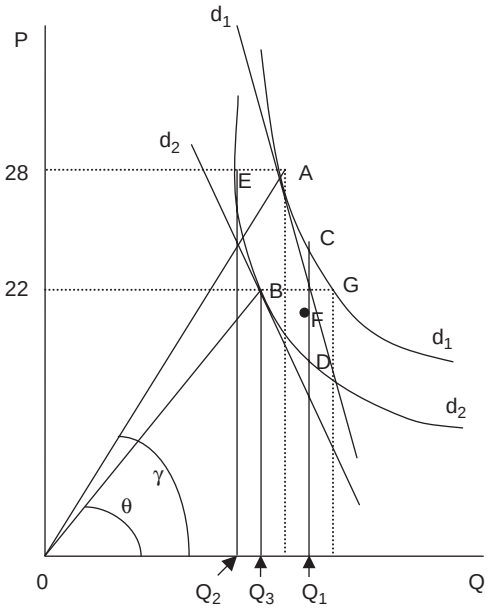


FIGURE 1 Switch in demand structure for oil. F, preembargo point; C, after-embargo point; A and E, upper limit of the price band; B and G, lower limit of the price band; A to B, market positions due to recession; C to D, market positions due to recession without a production cut.

2003. The market changed from a competitive one to one in which OPEC was able to exercise a considerable degree of monopoly power. Demand, in turn, changed from a relatively elastic individual market demand curve to a relatively inelastic industry cartel demand curve. The OPEC countries could therefore raise the oil price even further under this structural break in demand. This was the direct result of the unstable relation B.

In order to estimate the demand relations, we use data from "The Annual Energy Review" and "The Economic Report of the President." The price of coal (PC) is measured in cents per million Btu (cost, insurance, and freight price to electric utility power plants), as are prices of petroleum (P) and natural gas (PN). The quantity of petroleum consumption (Q) is in quadrillion Btu (10^{15}) and real gross domestic product is used to reflect income (Y). The sample period extends from 1949 to 1998. As with other economic variables, the unit root (unstable variables or not stationary) property needs to be examined before the estimations. Note that we fail to reject the null hypothesis of a unit root for all the variables even after the logarithmic transformation. That is, these variables are not stationary for the ordinary least squares technique: A technique is needed that relates a set of independent variables and dependent variables in terms of a linear equation. Since all the

variables in our model are not stationary, we apply the two-step technique by Engle-Granger to explore the cointegration relation (comovements of unstable variables). The cointegration model, used in the presence of unstable variables, suggests the use of the error correction model (ECM) in estimating the demand relation. That is, demand for petroleum can be appropriately formulated as

$$dQ_t = \mathrm{ECM}_{t-1} + d\mathrm{P}_t + d\mathrm{PC}_t + d\mathrm{PN}_t + dY_t + v_t, \tag{7}$$

where the prefix d denotes first difference of the variables (in logarithm); Q_t denotes consumption of crude oil; ECM_{t-1} is representing past disequilibrium; P_t denotes the price of crude oil; Y_t is income; PC_t is the price of coal; and PN_t represents the price of natural gas. Despite its complexity, the model is more general because it allows past disequilibrium or perturbations to be included in explaining oil consumption. The estimated results for the entire sample period are shown as follows:

$$dQ_t = \underset{(-1.56)}{-0.1645}\ \mathrm{ECM}_{t-1} - \underset{(-3.22)}{0.0502}\ d\mathrm{P}_t + \underset{(1.02)}{0.0275}\ d\mathrm{PN}_t - \underset{(-0.58)}{0.0210}\ d\mathrm{PC}_t + \underset{(6.89)}{0.7664}\ dY_t + \underset{(5.33)}{0.7462}\ \mathrm{AR}(1) - \underset{(-0.45)}{0.0054} + e_{1t} \tag{8}$$

$$\bar{R}^2 = 0.71,$$

where $\bar{R}^2$ is the adjusted R square, which reflects general fitness of the model; t statistics in parentheses indicate the significance of estimated coefficients. Within the framework of the partial adjustment model (or model with lagged dependent variable) with ECM, the price elasticity tends to converge approximately to $-0.198 = [-0.0502/(1-0.7426)]$ when estimated in the level of first difference. Similarly, we report the estimation results for the two-subsample periods (1949–1975 and 1976–1998):

$$dQ_t = \underset{(-2.72)}{-0.1182}\ \mathrm{ECM}_{t-1} - \underset{(-2.52)}{0.1778}\ d\mathrm{P}_t - \underset{(-1.22)}{0.060}\ d\mathrm{PN}_t + \underset{(1.92)}{0.1150}\ d\mathrm{PC}_t + \underset{(4.26)}{0.6097}\ dY_t + \underset{(2.89)}{0.0230} + e_{2t} \tag{9}$$

$$\bar{R}^2 = 0.70$$

and

$$dQ_t = \underset{(-2.83)}{-0.1643}\ \mathrm{ECM}_{t-1} - \underset{(-3.06)}{0.0625}\ d\mathrm{P}_t + \underset{(2.23)}{0.0673}\ d\mathrm{PN}_t + \underset{(0.26)}{0.0403}\ d\mathrm{PC}_t + \underset{(6.61)}{1.3127}\ dY_t - \underset{(-5.11)}{0.0366} + e_{3t} \tag{10}$$

$$\bar{R}^2 = 0.76,$$

The price elasticity of demand for petroleum within the entire sample period tends to approach -0.198 $[-0.0502/(1-0.7462)]$ in the long term, as indicated by Eq. (8). Thus, there appears to be enough room for further price hikes because the estimated price elasticity is not close to $k/(k-1)$ or approximately -1.05 in the empirical estimation as predicted by our theory.

The short-term price elasticities before and after the structural break can thus be estimated as -0.1778 and -0.0625, respectively. The empirical results indicate that (i) the elasticity after the energy crisis decreased in absolute value from 0.1778 to 0.0625 (a 65% decrease), clearly an unstable case of $\eta < 0$ (i.e., the price elasticity decreases as price increases), and (ii) there seems to be considerable room for a price hike because the long-term elasticity from Eq. (8) of -0.198 falls far short of $k/(k-1)$.

Neither the long-term price elasticity for the entire sample period (-0.198) nor the short-term price elasticity after the structural break (-0.0625) are close to the theoretical limit: $k/(k-1) \approx 1.05$ as is expected by a profit-maximizing cartel. The discrepancy may be explained by the significant income elasticity of 1.3127 in Eq. (10). Since the business cycle is both inevitable and unpredictable, a recession could certainly shift the demand curve to the left (Fig. 1). Continuous and gradual price hikes without disruptions in demand would render price elasticity toward $k/(k-1)$ in the long term. A recession would generally depress both price and quantity, as shown in Fig. 1 from point A to B. Given that the price elasticity can be geometrically measured as $e_A = \tan\gamma/\text{slope}\ (d_1)$ at point A, the size of the price elasticity does not appear to have changed appreciably. That is, $e_A = \tan\gamma/\text{slope}\ (d_1)$ and $e_B = \tan\theta/\text{slope}\ (d_2)$ are similar because both $\tan\theta$ and the slope of the line d_2 have decreased. Unless the U.S. economy is recession-proof in the long term, it does not seem possible that the long-term price elasticity would approach $k/(k-1)$ as implied by the theory. The wild swings in oil price after 1973 speak to the fact that demand for and supply of crude oil are not independent of political events. The

significant change in the oil price was and can be detrimental to the suppliers with limited capacity and relatively higher marginal cost. In contrast, producers with lower marginal cost and greater capacity (e.g., Saudi Arabia) would benefit from greater output quota. To prevent such violent price changes, it is advantageous to OPEC to have a price band in which the price elasticity is not too low. However, can $k/(k-1) \approx 1.05$, a theoretical limit developed previously, ever be reached? There are at least three reasons in favor of this argument. First, switching to alternative fuels becomes more plausible at high oil prices and thereby tends to push long-term price elasticity close to $k/(k-1)$. Second, noncartel output (e.g., 5–7 million barrels per day from Russia) in the long term can present a credible threat when the price is high enough. This renders OPEC's residual demand curve flatter for every high price. Third, significant *ad valorem* tariffs in G8 countries would effectively pivot the demand curve downward.

Substantial price fluctuations have been witnessed for the past three decades, especially since the 1973 energy crisis. Based on our estimate, the average price elasticity was very low, −0.1778 before 1975, but the change in market structure has given rise to the unstable demand case. It has been borne out with the short-term elasticity of −0.0625 after 1975. Notice that the negative and significant ECM coefficient (−0.1643) in Eq. (10) suggests a convergent trend toward equilibrium values after the structural break. In the absence of major war, it does not seem plausible that the price of oil will significantly exceed $40 due to the income effect in the demand structure. The price elasticity, hovering at −0.0625 after the structural break, suggests that the oil price would more likely approach the upper limit of the price band ($28) than the lower limit ($22) if OPEC members strictly adhered to the production cut (e.g., from Q_1 to between Q_2 and Q_3; Fig. 1) in the presence of a recession. The result is borne out because oil prices lingered around $28 at the beginning of May 2001, 1 month after the production cut. As the recession deepened after the September 11, 2001, terrorist attacks, the price of crude oil dropped to $20 in November 2001 when the OPEC cartel was reluctant to further reduce production. Failing to do so would result in a drastic price reduction (from point C to point D in Fig. 1) as occurred before. On the other hand, the upward pressure on price could be too irresistible because −0.0625 is far below the theoretical limit of −1.05, as inelastic demand promotes price increase.

3. VOLATILITY OF AN OIL PRICE CHANGE AND MACROECONOMIC ACTIVITY

Given that oil is of great importance in the production process, its impact on industrial output cannot be ignored. Furthermore, the growing magnitude of oil price changes reflects the major role of uncertainty in statistical modeling. For instance, Lee *et al*. found that uncertainty in oil price (conditional variance of an oil price change) could significantly impact economic growth. In particular, one unit of positive normal shock (a large price increase) gives rise to a decreased output, whereas one unit of negative shock does not necessarily lead to an increased output. Similarly, the impact of an oil price change on the financial market has received increased attention. Sadorsky was among the first to apply a formal statistical model to investigate such an impact. To capture the uncertainty inherent in the macroeconomic foundation, we need to first define the real oil price after adjusting for the consumer price index or roilp_t. To ensure stationarity property, let Δroilp_t represent the first difference of the logarithmic transformation on roilp_t. One of the common ways to measure the volatility of oil price is to apply an econometrical model called autoregressive and conditional heteroskedastic (ARCH) by Engle or its generalized version, the GARCH model, by Bollerslev. The GARCH model shown below is fairly general in describing the behavior of a volatile time series such as oil price in macroeconomics:

$$\begin{aligned}\Delta\text{roilp}_t &= a_0 + \sum_{i=1}^{p} \phi_i \Delta\text{roilp}_{t-1} + \varepsilon_t - \sum_{j=1}^{k} \theta_j \varepsilon_{t-j} \\ \varepsilon_t &= z_t \sqrt{h_t}\, z_t \sim N(0,1) \\ h_t &= \alpha_0 + \beta_1 h_{t-1} + \alpha_1 \varepsilon_{t-1}^2, \end{aligned} \tag{11}$$

where h_t is the conditional variance often used to represent volatility, and $z_t = \varepsilon_t/\sqrt{h_t}$ denotes a standardized disturbance term. Note that optimal lags of the GARCH (1,1) −ARMA(p, q) are selected based on the criterion that a series correlation (correlation via residuals) is not found in z_t and z_t^2. Finally, z_t is normally distributed with zero mean and unit variance or $N(0,1)$. Even though h_t is frequently chosen to measure the volatility, the major problem is that variance masks the direction of changes to which shocks are administered. The asymmetric impact—only positive shocks (or price increases) slow economic activity and/or the stock market—begs the use of z_t instead of h_t as a proxy for

volatility. As such, we use z_t as the proxy in our research model.

Monthly data of industrial production (ip_t), real stock return ($\Delta rstkp_t$), consumer price index (cpi_t), exchange rate ($exch_t$), interest rate (r_t), and stock indices from July 1984 to March 2002 were obtained from the International Financial Statistics (IFS) data bank. Data for oil prices are from "Energy Prices and Taxes" published by the International Energy Agency. Note that we include exchange rates because oil is exported to many industrialized economies. In addition, the interest rate is incorporated to account for monetary policy. Major industrialized countries included in the study are the United States, Canada, Italy, Japan, Germany, and the United Kingdom.

As in many other time series analyses, we need to examine the existence of a unit root (unstable) on the variables of lip_t, $lrstkp_t$, lr_t, and $lexch_t$. The prefix l indicates the logarithmic transformation of the original variables, used to damp unstable variables. If these variables are $I(0)$ or integrated of order zero indicating stationarity or stability of the variable, we could add z_t to form a five-variable VAR model. However, if these variables are $I(1)$ or integrated of order 1 (not stationary), we need to first examine the potential existence of any cointegration relation (or comovements of unstable variables). A five-variable vector error correction model can be formulated in the presence of such a cointegration relation. Otherwise, the four-variable (first difference) model along with z_t would suffice to analyze the direction of causality among the variables.

The techniques for examining variable stationarity by Phillips and Perron and by Kwiatkwoski *et al*. are applied to the four variables for the six countries. The result is unanimous in that all the variables are of $I(1)$ or not stationary. Consequently, we examine the potential existence of cointegration relations via the trace method developed by Johansen. An examination of the trace statistics suggests that no cointegration relation exists for the five countries except Japan, and two sets of cointegration relations exist for Japan. That is, we can include the past disturbance term ECM_{t-1} in the model to explain industrial production for Japan as follows:

$$\begin{aligned}\Delta lip_t = a_4 + r_4 ECM_{t-1} + \sum_{i=1}^{k} b_{4i}\Delta lr_{t-i} \\ + \sum_{i=1}^{k} c_{4i}\Delta lexch_{t-i} + \sum_{i=1}^{k} d_{4i} z_{t-i} \\ + \sum_{i=1}^{k} f_{4i}\Delta lip_{t-i} + \sum_{i=1}^{k} g_{4i}\Delta lrstkp_{t-1} + e_{4t},\end{aligned} \tag{12}$$

where ECM_{t-1} is the error correction term representing the long-term cointegration. On the other hand, the five-variable vector autoregression model is applied to the other five countries as follows (only the stock return equation is shown):

$$\begin{aligned}\Delta lrstkp_t = a_5 + \sum_{i=1}^{k} b_{5i}\Delta lr_{t-i} + \sum_{i=1}^{k} c_{5i}\Delta lexch_{t-i} \\ + \sum_{i=1}^{k} d_{5i} z_{t-i} \\ + \sum_{i=1}^{k} f_{5i}\Delta lip_{t-i} + \sum_{i=1}^{k} g_{5i}\Delta lrstkp_{t-i} + e_{5t}.\end{aligned} \tag{13}$$

The optimal lag lengths are determined based on the Akaike's Information Criterion, and directions of the Granger causality can be analyzed by the impulse response function.

To purge serial correlation in z_t of Eq. (11), for more accurate estimation, a lag of 5 on autoregression and a lag of 1 on moving average are needed.

TABLE I

Results of the Granger Causality[a]

H_0	Canada	Germany	Italy	Japan	United Kingdom	United States
z_t not $\rightarrow \Delta lr_t$	Cannot reject	Cannot reject	(+) Reject*	Cannot reject	Cannot reject	Cannot reject
z_t not $\rightarrow \Delta lexch_t$	Cannot reject	Cannot reject	Cannot reject	Cannot reject	Cannot reject	Cannot reject
z_t not $\rightarrow \Delta lip_t$	Cannot reject	Cannot reject	(−/+) Reject**	Cannot reject	(+) Reject***	Cannot reject
z_t not $\rightarrow lrstkp_t$	Cannot reject	(−) Reject**	(−) Reject*	Cannot reject	Cannot reject	(−) Reject***

[a] "not $\rightarrow$" means "does not Granger cause." The Granger causality carries the information that variable x causes variable y, whereas the ordinary regression model provides x associates y. A rejection of H_0 implies the existence of a Granger causality. *, **, and *** denote 1, 5, and 10% significance level, respectively. The signs in parentheses denote the direction of the causality from the impulse response analysis. −/+ indicates the impact is first negative and then turns positive.

The result indicates that the volatile behaviors of oil prices are best described by an ARCH model for Canada due to its insignificant β_1. However, the GARCH model is preferred for the remaining countries, of which Japan has the strongest persistence in oil price volatility (i.e., $\alpha_1 + \beta_1 > 1$). The United States has the smallest persistence, with $\alpha_1 + \beta_1 = 0.7866$. Except for Japan, in which $\alpha_1 > \beta_1$ (unexpected shock impacts current conditional variance more than past conditional variance), the reverse is true for other countries. Note that the GARCH model is selected so that z_t and z_t^2 are free of serial correlation, an undesirable property in a regression analysis.

It is not surprising that the greatest z_t (greatest change in oil price) value occurred during the Gulf War in 1990. Table I reports the causality results from the statistical models. The lengths of impacts can be determined from the analysis. In the case of Germany, the volatility of a real oil price change exerts a negative impact on stock returns for three periods before it tapers off to zero. For Italy, the volatility of a real oil price change leads the monthly interest rate to change positively for 7 months, the monthly stock return to change negatively for approximately 4 months, and the monthly industrial production to change negatively in the first month but positively during the next 3 months.

Surprisingly, the same volatility exerts positive impacts on industrial production (3 months) for the United Kingdom before leveling off to zero. The favorable impact of higher oil prices on industrial output can be attributed to the fact that Britain is one of the non-OPEC producers: A higher oil price commands the greater revenue ($\eta < 0$) from export, which can in turn be reinvested. For the United States, the volatility of a real oil price change leads to negative stock returns for 3 months before diminishing gradually to zero. The same volatility has no appreciable impacts on either industrial production or stock returns for Canada and Japan. It seems that the volatility of oil price changes leads to negative stock returns in three of the six countries. It affects industrial production in only two countries, including a positive impact in the United Kingdom.

4. CONCLUSION

Recent studies have highlighted the role of oil price uncertainty. The results seem to favor the long-held conjecture that higher oil prices have a greater impact on the stock market than on the output market. This is supported in our model for the United States. Higher oil prices seem to have a lower impact on major economies such as those of the United States and Japan due to the presence of strategic oil reserve. A strategic oil reserve is essential to absorb the shocks from an excessive price hike. At the heart of the volatility analysis is the concept of oscillating price elasticity at the microeconomic foundation developed by Greenhut *et al*. The existence of unstable price elasticity indeed sows the seed of inherent volatile oil price changes, which can readily inflict shocks to the economy, especially a relatively small economy whose demand for oil depends solely on import. Our volatility model, employing the most recent data, supports this conclusion. The impacts on the interest rate and other monetary variables remain a challenging research avenue for the future.

SEE ALSO THE FOLLOWING ARTICLES

Business Cycles and Energy Prices • Energy Futures and Options • Inflation and Energy Prices • Markets for Petroleum • Oil and Natural Gas: Economics of Exploration • Oil and Natural Gas Leasing • Oil Crises, Historical Perspective • Oil Industry, History of • Oil-Led Development: Social, Political, and Economic Consequences

Further Reading

Alhajji, A. F., and Huettner, D. (2000). OPEC and world crude oil market from 1973 to 1994: Cartel, oligopoly, or competitive? *Energy J.* **21**(3), 31–58.

Bollerslev, T. (1986). Generalized autoregressive heteroskedasticity model. *J. Econometrics* **31,** 307–327.

Burbridge, J., and Harrison, A. (1984). Testing for the effect of oil price rises using vector autogressions. *Int. Econ. Rev.* **25**(1), 459–484.

Dargay, J., and Gately, D. (1995). The response of world energy and oil demand to income growth and changes in oil prices. *Annu. Rev. Energy Environ.* **20,** 145–178.

Department of Energy, Energy Information Administration (1998). *Annu. Energy Rev.*

Engle, R. F. (1982). Autoregressive conditional heteroskadasticity with estimates of variance of United Kingdom inflation. *Econometrica* **50,** 987–1007.

Engle, R. F., and Granger, C. W. J. (1987). Cointegrantion and error correction: Representation, estimation, and testing. *Econometrica* **55,** 251–276.

Ferderer, J. (1996). Oil price volatility and the macroeconomy. *J. Macroecon.* **18**(1), 1–26.

Gately, D. (1993). The imperfect price—Reversibility of world oil demand. *Energy J.* **14**(4), 163–182.

Greene, D. L. (1991). A note on OPEC market power and oil prices. *Energy Econ.* **13,** 123–129.
Greene, D. L., Jones, D. W., and Leiby, P. N. (1998). The outlook for U.S. oil dependence. *Energy Policy* **26**(1), 55–69.
Greenhut, M. L., Hwang, M. J., and Ohta, H. (1974). Price discrimination by regulated motor carriers: Comment. *Am. Econ. Rev.* **64**(4), 780–784.
Griffin, J. M. (1985). OPEC behavior: A test of alternative hypotheses. *Am. Econ. Rev.* **75**(5), 954–963.
Griffin, J. M., and Teece, D. J. (1982). "OPEC Behavior and World Oil Prices." Allen & Unwin, London.
Hamilton, J. D. (1983). Oil and the macroeconomy since World War II. *J. Political Econ.* **99**(2), 228–248.
Huang, R. D., Masulis, R. W., and Stoll, H. R. (1996). Energy shocks and financial markets. *J. Future Market* **16**(1), 1–27.
Hwang, M. J. (1982). Crude oil pricing in the world market. *Atlantic Econ. J.* **10**(2), 1–5.
Johansen, S. (1988). Statistical and hypothesis testing of cointegration vectors. *J. Econ. Dynamic Control* **12,** 231–254.
Kwiatkwoski, D., Phillips, C. B., Schmidt, P., and Shin, Y. (1992). Testing the null hypothesis of stationary against the alternative of a unit root. *J. Econometrics* **54,** 159–178.
Lee, K., Ni, S., and Ratti, R. A. (1995). Oil shocks and the macroeconomy: The role of price variability. *Energy J.* **16**(4), 39–56.
Mork, K. A., and Olson, M. H. T. (1994). Macroeconomic responses to oil price increases and decreases in seven OECD countries. *Energy J.* **15**(4), 19–35.
Papapetrou, E. (2001). Oil price shocks stock market: Economic activity and employment in Greece. *Energy Econ.* **23,** 511–532.
Phillips, P. C. B., and Perron, P. (1989). Testing for unit root in time series regression. *Biometrika* **75,** 335–346.
Sadorsky, P. (1999). Oil price shocks and stock market activity. *Energy Econ.* **21,** 449–469.
U.S. Government Printing Office (1998). "Economic Report of the President." U.S. Government Printing Office, Washington, DC.
Yang, C. W., Hwang, M. J., and Huang, B. N. (2002). An analysis of factors affecting price volatility of the U.S. oil market. *Energy Econ.* **24,** 107–119.

Oil Recovery

RUSSELL T. JOHNS
University of Texas at Austin
Austin, Texas, United States

1. Importance of Oil and Gas Recovery
2. Overview
3. Petroleum Reservoir Fluids
4. Thermodynamics and Modeling of Oil and Gas
5. Primary Recovery Processes
6. Secondary and Tertiary Recovery Processes

Glossary

enhanced oil recovery processes Those processes that stimulate oil recovery by the injection of materials not normally present within the reservoir; usually initiated as a secondary or tertiary recovery method.
miscibility Occurs between two fluids when no distinct interface exists when mixed in any proportion.
oil saturation The ratio of the oil volume that occupies the pores to the volume of the pores.
permeability The ability of fluids to flow through pores of the rock that is largely dependent on the size, shape, and connectivity of the pores.
porosity The ratio of the pore volume of the rock where fluids reside to the bulk volume of rock that includes rock grains and pores.
primary recovery processes Those processes that recover oil by relying on the natural energy available in the reservoir and adjacent aquifers.
recovery efficiency or factor The fraction of the original volume of oil in place that is recovered.
reservoir The portion of porous and permeable rock that contains oil, gas, or water that typically consists of multiple layers of sandstones, limestones, or dolomites from a few meters to nearly a thousand meters thick.
secondary and tertiary recovery processes Those processes that stimulate oil recovery by external injection of fluids to increase the energy available in the reservoir following primary recovery.

Oil recovery is about the efficient and cost-effective recovery of oil from subsurface reservoirs. The design of recovery methods to satisfy these objectives is the responsibility of petroleum engineers. Petroleum engineers handle the analysis, design, and implementation of recovery projects, which include topics in reservoir, drilling, petrophysical, and production engineering.

1. IMPORTANCE OF OIL AND GAS RECOVERY

1.1 World Energy Consumption

Oil and gas recovery are, and will remain, critical to the energy needs of the world. Figure 1 shows the world's energy consumption for the five major energy sources. In 2002, nearly two-thirds of the world's energy supply came from oil and gas recovery, and oil was the single most important source of energy in the world.

Oil and gas are expected to remain important far into the future. World energy consumption is expected to increase by nearly 58% between 2001 and 2025. Over that period, oil and gas energy sources will remain at nearly two-thirds of the world's energy needs (Fig. 2). Renewables such as wind, solar, and geothermal energy currently represent approximately 2.5% of the world's energy consumption and are not expected to increase dramatically over this period.

1.2 Environmental Benefits

Petroleum engineers have also been pivotal in solving complex environmental problems by extending oil recovery technology to the effective cleanup of aquifers contaminated with toxic chemicals. An example of such an extension is the use of surfactants to clean up aquifers contaminated with nonaqueous phase liquids such as trichloroethane. Surfactant enhanced aquifer remediation, has been very effective in the cleanup of aquifers where conventional cleanup methods have failed.

 © 2004 Elsevier Inc. All rights reserved.

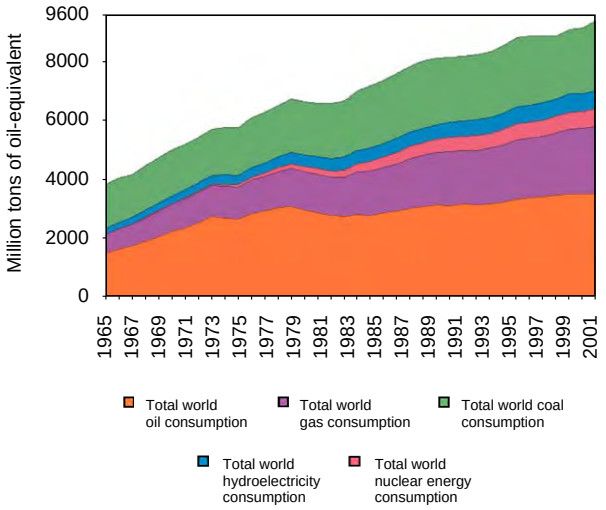


FIGURE 1 Historic world energy consumption of primary energy sources (in millions of tonnes oil-equivalent). From British Petroleum (2003).

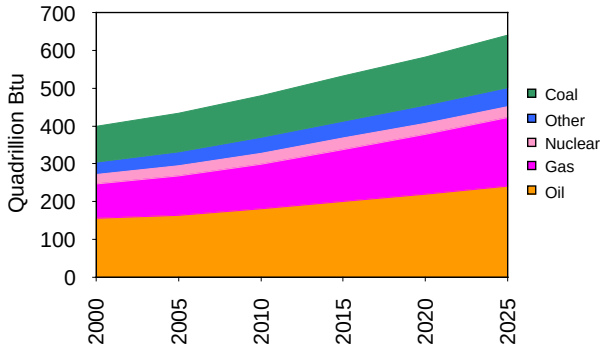


FIGURE 2 Projected world energy consumption by energy source (in quadrillions of Btu). Data from Energy Information Administration. (2003). "System for the Analysis of Global Energy Markets." EIA, Washington, DC.

Achieving high oil recovery by good engineering practices has additional environmental benefits. For example, the use of enhanced oil recovery (EOR) techniques extends the field life beyond its primary recovery period, which lessens the need for new field developments. In addition, EOR reduces greenhouse gas (GHG) emissions by sequestration of carbon dioxide, and injection of hydrocarbon gases that would otherwise be flared.

2. *OVERVIEW*

Subsurface oil reservoirs are found at depths at or near the Earth's surface to several kilometers below. Oil reservoirs are not pools or lakes of oil below the surface; the oil resides in pores within rock, and this greatly complicates recovery. Most oil very near the Earth's surface is heavy oil because lighter components in the oil (e.g., methane, ethane) migrate into the atmosphere over time, leaving only the heaviest components behind. However, the processes involved in the migration and formation of oil reservoirs is only partially understood.

Recovery depends largely on the nature of the reservoir and the fluids. Reservoir heterogeneities, such as layering within the reservoir, play a significant role in how quickly the oil (or gas) is produced. For example, oil in low-permeability layers is difficult to recover due to its resistance to flow. Reservoirs may also contain isolated compartments or fractures of high permeability. Fractures, in particular, require good characterization to reduce uncertainty in the recovery predictions. The composition of the oil, the reservoir temperature and pressure, and the distribution of the oil, water, and gas within the reservoir are also important to the selected recovery process. Thus, the most important step in oil recovery design is to characterize accurately the reservoir formation and its fluid properties.

One of the most important outcomes of any characterization is to determine the oil target for a recovery process by calculation of the volume of the original oil in place (OOIP). Only with a reliable OOIP number can petroleum engineers evaluate the success of a particular recovery process.

Surface seismic borehole cores, pressure transient and tracer tests, and geophysical methods (e.g., well logging) are often used to make a three-dimensional characterization of the oil reservoir. Well logging is a geophysical method that introduces a device, called a sonde, down a borehole to measure, using electrical and nuclear methods, properties such as porosity and fluid saturations. However, the characterization of a reservoir is a daunting task given that reservoirs are highly heterogeneous, and these methods sample limited and varying volumes of reservoirs at differing degrees of resolution.

The number and placement of wells in the reservoir is also important to recovery. Wells can be perforated over a selected depth interval where it is believed that good oil recovery will occur. Advances in drilling engineering allow for wells to be drilled a great length horizontally through thin oil layers. Where multiple wells had to be drilled in the past, now sometimes only a few wells are required. Advances in offshore drilling have also opened up new possibilities for finding and producing oil reservoirs. Offshore wells can now be drilled in oceans that are more than a mile deep.

Once the reservoir and its fluids are characterized, a recovery process is selected based on economic and engineering constraints. Primary recovery methods are those that are initially used to recover as much of the oil as possible by drilling production wells (boreholes) in predetermined patterns. Primary recovery eventually becomes uneconomical as the reservoir fluids lose energy (or pressure) over time as the oil is produced. Some reservoirs require pumps or other methods to assist in recovery, especially at later stages of production. Secondary and tertiary recovery methods, such as water and gas injection, are typically used (if cost-effective) after oil recovery from primary methods has peaked and a significant decline is taking place. Either new injection wells are drilled or some of the production wells are converted into injection wells in a predetermined pattern partially dictated by the pattern of wells used for primary recovery.

An accurate estimation of the ultimate recovery efficiency is one of the primary goals of petroleum engineers. The recovery efficiency or factor, which is a fraction between 0 and 1, is defined as the ratio of the volume of oil produced (or recovered) to the volume of OOIP.

The most important aspects of oil recovery from an economic viewpoint are the rate and volume of oil recovered. A decision about which recovery process to use is not based solely on what could be recovered but rather on economic limits and rankings, risk factors, and available assets. Besides safety and environmental concerns, engineering designs are driven by the desire to rapidly produce the oil in the most cost-effective way.

3. PETROLEUM RESERVOIR FLUIDS

The composition of oil and gas in a reservoir (or the fluids injected) is important in the design of recovery processes. Several experimental methods based on true boiling point distillation and gas chromatography are available to determine the composition of the fluids. Current methods can report molar compositions for individual carbon numbers up to approximately C_{35} and can also give PNA (paraffins, naphthenes, and aromatics) distributions. Older reports generally report only paraffin hydrocarbon components up to carbon numbers of C_6, and the rest are lumped into a C_{7+} fraction or pseudo-component. Ideally, compositional analyses should give the mole fractions, boiling points, specific gravity, critical properties, and molecular weights of as many components as possible so that equations of state (EOSs), which are mathematical models of the oil and gas phase behavior, can be constructed with reasonable accuracy and reliability. This is a daunting task given that many isomers exist for carbon numbers greater than C_6. For example, the paraffin hydrocarbon decane (C_{10}) has 75 different structural configurations, each one of which has different physical properties. Thus, components with a similar number of carbon atoms are often lumped into one pseudo-component with their average properties reported.

Petroleum reservoir fluids are classified as black oils, volatile or near critical oils, retrograde gas condensates, wet gases, and dry gases. The fluid composition determines the phase envelopes (the boundary of the two-phase region), which are composed of the dew point and bubble point curves. The reservoir fluid type is determined relative to the phase envelope (Fig. 3). A single phase, usually gas or liquid, exists outside the phase envelope. Inside the envelope, two phases coexist at equilibrium. The dew point and bubble point curves are separated by the critical point (labeled C in Fig. 3), where the bubble point curve is to the left of the critical point. The dew point curve gives the pressure at which the first small drop of liquid forms as the pressure is increased at constant temperature. The bubble point curve gives the pressure at which the first small bubble of gas forms as the pressure is decreased at constant temperature. The critical point is the temperature and pressure at which the two phases and their properties become identical. The critical point is very important in the determination of miscibility.

Four phase envelopes are shown in Fig. 3, each of which is for a given reservoir fluid composition. For a

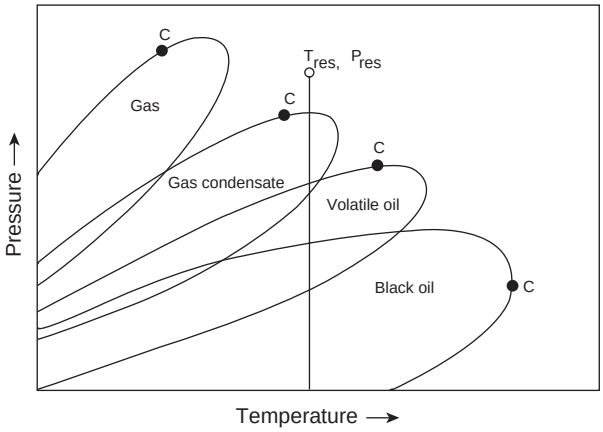


FIGURE 3 Reservoir fluid types and phase envelopes. From Pederson, Fredenslund, and Thomassen (1989).

fluid composition that gives the phase envelope at the upper left-hand side of the figure, the fluid is a gas at the reservoir temperature and pressure shown and remains a gas as the reservoir pressure is reduced. The reservoir pressure is typically reduced during production, especially for primary depletion drives. The reservoir temperature is essentially constant so long as fluids such as steam are not injected. For a fluid composition that gives the next phase envelope from the left in Fig. 3, the reservoir pressure intersects the dew point curve twice. Thus, as pressure is reduced during production, the liquid saturation will first increase and reach a maximum but will subsequently decrease to zero. Reservoir fluids that behave in this way are known as retrograde gas condensates. For the other two fluid compositions (and their corresponding phase envelopes), the fluids are single-phase oils, known as volatile or near critical oils and black oils. For both of these oils, the bubble point curve is first intersected as pressure is reduced at a pressure known as the bubble point pressure or saturation pressure, followed by the intersection of the dew point curve. Thus, the oil saturation continuously decreases, whereas the gas saturation continuously increases. Black oils contain a greater mole fraction of heavier components than do the other fluid types, and these heavier components will remain in the liquid phase even for very low pressures.

4. THERMODYNAMICS AND MODELING OF OIL AND GAS

Numerical simulation is often used to predict oil recovery from a reservoir, and the calculation of phase behavior is an integral part of those predictions. Phase behavior describes the complex interaction among physically distinct, separable portions of matter called phases that are in contact with each other. Typical phases are solids, liquids, and vapors. Phase behavior plays a vital role in many petroleum applications such as EOR, compositional simulation, geochemical behavior, wellbore stability, geothermal energy, environmental cleanup, multiphase flow in wellbores and pipes, and design of surface facilities.

Thermodynamics, which is central to understanding phase behavior, is the study of energy and its transformations. Thermodynamics began as the study of heat applied to steam power but was broadened substantially by J. W. Gibbs during the mid- to late 1800s. Thermodynamics describes the energy changes that occur during phase transformations; it is used to predict the phase separations that occur. The most fundamental idea in thermodynamics is the conservation of total energy—the first law of thermodynamics. The first law is based on our everyday observation that for any change in thermodynamic properties, total energy, which includes internal, potential, kinetic, heat, and work energy, is conserved. The second fundamental idea in thermodynamics is the total entropy balance or the second law of thermodynamics. Entropy is a thermodynamic property that expresses the unidirectional nature of a process.

4.1 Phase Equilibrium Thermodynamics

Gibbs's most significant contribution was the development of phase equilibrium thermodynamics applied to multicomponent mixtures, particularly the concept of chemical potential. The concept of chemical potential leads to the simple result that, at equilibrium, the chemical potential of each component must be the same in all phases ($\mu_{iL} = \mu_{iV}$).

Phase equilibrium thermodynamics seeks to determine properties such as temperature, pressure, and phase compositions that establish themselves after all tendencies for further change have disappeared. Figure 4 is a schematic showing a closed container of liquid and vapor that could represent one grid block in a numerical simulation. Given constant and known temperature, pressure, and overall molar fractions or compositions (z_i where $i = 1,\ldots,n_c$), the fundamental task in petroleum phase behavior calculations is to quantify the phase molar fractions (phase splits L and V) and the phase equilibrium molar fractions of the vapor (y_i where $i = 1,\ldots,n_c$) and liquid phases (x_i where $i = 1,\ldots,n_c$) that form at

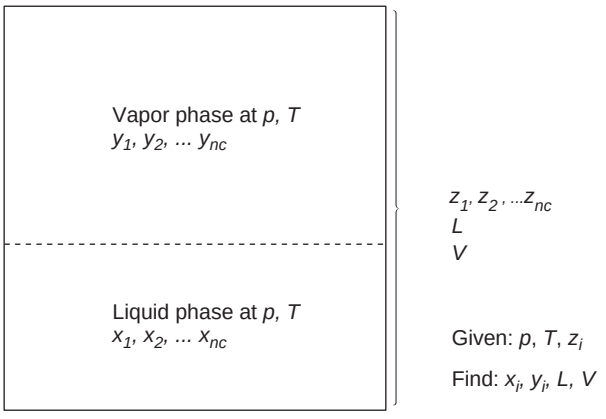


FIGURE 4 Vapor–liquid equilibrium at constant pressure, temperature, and overall molar composition. A dashed line shows a distinct interface between the two phases.

equilibrium. z_i is the ratio of the moles of component i in all phases to the total moles of all components in all phases, x_i is the ratio of the moles of component i in liquid to the total moles of liquid, y_i is the ratio of the moles of component i in vapor to the total moles of liquid, and n_c is the number of components. L is the ratio of the total moles of liquid to the total moles of all phases, and V is the ratio of the total moles of vapor to the total moles of all phases. The phases are assumed to be homogenous in that parameters such as pressure, temperature, density, viscosity, and phase compositions are uniform throughout the phases. By satisfying the equilibrium conditions that $\mu_{iL} = \mu_{iV}$, the vapor–liquid equilibrium properties can be calculated to determine the compositions of vapor and liquid in equilibrium along with the phase molar splits and volumes.

Oil recovery processes are never in equilibrium, and it would appear that equilibrium thermodynamics is not very useful. The concept of local equilibrium is used to overcome this apparent limitation of thermodynamics. Equilibrium at a point in a reservoir, termed local equilibrium, often applies when internal relaxation processes are rapid with respect to the rate at which changes are imposed on the system. That is, equilibrium thermodynamics can be applied locally over small volumes of the reservoir, even though pressure and other gradients remain across the reservoir. In reservoir simulation, the small volumes are taken to be grid blocks. The size of the grid blocks must be sufficiently small so that good accuracy is obtained.

4.2 Equations of State

An accurate characterization of phase behavior is critical to the prediction of oil recovery. Continuous experimental data covering the full range of pressures and compositions that may occur in a reservoir are never available. Thus, mathematical models that are "tuned" to the available experimental data are used to predict the phase behavior. EOS models are used for this purpose. The simplest and most fundamental EOS is the ideal gas equation, where the pressure, volume, and temperature of a fluid are related by $pV = RT$, where p is pressure, V is molar volume (inverse of molar density), R is gas constant, and T is temperature.

For compositional simulation purposes, more complex EOSs that accurately represent, in a continuous way, both the oil and gas phases are used. The two most widely used cubic EOSs in the petroleum industry are the Peng–Robinson EOS (PREOS) and modified versions of the Redlich–Kwong EOS (RKEOS). These two cubic EOSs are sufficiently simple that they can be used in reservoir simulations with as many as hundreds of thousands of grid blocks so long as the number of components that represent the fluid is less than approximately 10.

An example of a cubic EOS is the PREOS, which is expressed for a pure fluid as

$$p = \frac{RT}{V-b} - \frac{a\alpha}{V(V+b)+b(V-b)}$$

where parameters a and b are functions of the component's critical temperature and pressure and α is a temperature-dependent function. This type of cubic EOS represents both the liquid and vapor in a continuous way. For example, when density decreases (or V is large), the PREOS approaches the correct limit of the ideal gas equation. For the other limit where V approaches b, the cubic EOS approaches an incompressible liquid in that the pressure increases rapidly with small changes in the density.

For fluid mixtures, the same form of the cubic EOS is used, but mixing rules determine the EOS parameters. Binary interaction parameters are used to account for nonideal interactions between different species. Volume shift parameters adjust liquid phase densities calculated from the EOS to experimentally measured densities. The phase viscosities are estimated from empirical relations and properties calculated from the EOS. Critical volumes of the components can be used to adjust the phase viscosities to match experimentally measured viscosities.

The accuracy of any EOS depends on its ability to model the attractions and repulsions among molecules over a wide range of temperatures and pressures. EOS models are empirical in that they do not attempt to model the detailed physics; rather, they attempt to model only the cumulative effect in terms of a small number of empirical parameters. In general, cubic EOS models of the Peng–Robinson type are more accurate for molecules with weak polarity or chemical interactions, and this explains why water, a polar substance, is difficult to model with a cubic EOS. Mixtures that contain alcohols, bases, organic or inorganic acids, and electrolytes are other examples of fluids that are not accurately modeled by cubic EOS models of the type shown here.

Fluid compositions may vary from one location in the reservoir to another. In such cases, multiple EOS characterizations may be required. Compositional variations can occur for a variety of reasons. For example, gravity can cause vertical compositional

gradients, where heavier components become more concentrated at greater depths. Composition variations in a reservoir can also be caused by thermal gradients.

4.3 Flash Calculation

The procedure for equilibrium calculation of a two-phase mixture shown in Fig. 4 is called a flash calculation. For an equilibrium flash calculation, the pressure, temperature, and overall mole fractions are specified. The general procedure for a flash calculation is as follows:

1. Make an initial guess of the K values, where $K_i = y_i/x_i$. Most EOS programs use empirical correlations to estimate the phase equilibrium molar fractions based on K values.
2. Calculate x_i and y_i using a material balance procedure such as the Rachford–Rice procedure. Typically, a nonlinear material balance equation is solved by Newton–Raphson iteration.
3. Calculate the cubic EOS parameters (e.g., a, b). This step is straightforward and depends on the selected EOS and its associated mixing rules. The critical temperatures, pressures, and acentric factors for each component, among others, are needed to calculate the EOS parameters.
4. Solve the cubic EOS for the molar volumes V of the liquid and vapor.
5. Calculate the chemical potentials (or, alternatively, the fugacities) of each component in each phase. The selected cubic EOS is used to determine an expression for the fugacity of a component in a phase.
6. Check whether equilibrium has been reached. The correct equilibrium solution is found when $\mu_{iL} = \mu_{iV}$ (or, alternatively, the component fugacities are equal) for each component relative to an acceptable tolerance.
7. If the tolerance has not been satisfied, update the K values and repeat steps 2 to 6. Successive substitution or Newton–Raphson iteration is often used to update the K values.

4.4 Simplified Fluid Characterization

Phase behavior calculations require that components and their properties be known. However, crude oils typically have hundreds of components, making the preceding flash calculation procedure computationally too intensive even for numerical simulation. Thus, components are often lumped into 10 pseudo-components or less to characterize the *in situ* fluid. The characterization usually takes the following three steps:

1. Analyze the hydrocarbon components in the *in situ* fluid using analytical techniques such as chromatography and distillation.
2. Separate and lump the measured components into as few pseudo-components as is possible without sacrificing the required accuracy. The properties and selection of the pseudo-components are determined using a variety of methods and empirical correlations. The required pseudo-component properties are those needed for the cubic EOS calculations such as critical temperature, pressure, and acentric factor.
3. Adjust the pseudo-component properties so that the EOS will match all available pressure–volume–temperature (PVT) experiments. This process, which typically uses nonlinear regression, is known as EOS tuning. The objective of tuning is to match fluid properties and phase equilibria accurately over the range of pressures and compositions expected to occur in a reservoir. EOS tuning is nearly always needed due to the inherent uncertainty in the properties estimated from step 2, especially for the heavier components. Binary interaction parameters between the heaviest and lightest pseudo-components are typically the first parameters to be adjusted, although other parameters (e.g., the acentric factors) might need tuning.

The selection of pseudo-components and their properties is likely not unique, as is often the case when numerous model parameters are estimated by fitting measured data with nonlinear regression. Care should be taken to avoid estimates in the pseudo-component properties that are unphysical and to reduce the number of parameters fitted. Furthermore, the final EOS characterization is most accurate in the range of the experimental phase behavior data that have been measured and fitted. Phase behavior data that covers, to the extent possible, the conditions that occur in the reservoir should be collected.

4.5 Pressure–Volume–Temperature Experiments

Routine PVT experiments give volumetric information over a range of composition and pressure conditions expected for conventional methods such

as water flooding and pressure depletion. Typical routine experiments include differential liberation, constant composition, constant volume depletion, and separator tests. These data are used for tuning EOSs for black and volatile oils. Differential liberation simulates depletion of black oil below its bubble point pressure as gas is evolved. Constant composition expansion tests simulate black oil depletion from above to below the bubble point pressure. Constant volume depletion experiments simulate depletion of a volatile oil below its bubble point pressure or a retrograde gas condensate below its dew point. This test can provide an approximate estimate of oil recovery that is expected during depletion. Finally, separator experiments simulate single- or multiple-stage separation processes at the surface. This test can provide gas/oil ratios (GORs) and formation volume factors (B_o) and molecular weight of the oil that remains in the surface stock tank. The oil formation volume factor is the ratio of the volume of oil at reservoir conditions to the volume of oil in the stock tank. The stock tank conditions are typically at atmospheric pressure and 15°C, which are referred to as standard temperature and pressure (STP) conditions. Compositions of the gas and liquid phases that separate at each simulated stage are measured from separator tests.

When tertiary recovery methods (e.g., gas injection) are used, other PVT experiments are required for EOS tuning that may better represent the compositions expected in the reservoir during recovery. These tests include swelling tests and multiple-contact experiments. Swelling tests measure phase behavior data in a plot that is commonly called a pressure composition diagram in thermodynamics. The injection gas is mixed with reservoir oil along a dilution line to compositions beyond near critical conditions. Phase equilibrium compositions are measured for several overall compositions along with the liquid saturation. The swelling or increase in the formation volume factor as gas is added is also measured. Multiple-contact tests simulate the gas/oil contacting process that occurs in a reservoir. Forward experiments simulate multiple contacting in a vaporizing gas drive in that fresh gas repeatedly contacts equilibrium oil. A backward multicontact experiment simulates contacts that occur in a purely condensing gas drive, where equilibrium gas repeatedly contacts fresh oil. However, multiple-contact tests do not adequately simulate the combined condensing–vaporizing drives that commonly occur during enriched gas or carbon dioxide floods.

5. PRIMARY RECOVERY PROCESSES

Primary recovery processes are those that recover oil by relying on the natural energy available in the reservoir and adjacent aquifers. Primary recovery relies on the expansion of *in situ* fluids or rocks to produce oil (or gas); specific mechanisms include solution gas drive, gas cap drive, natural water drive, and rock compaction. A combination of these recovery mechanisms occurs in most reservoirs, but often one of them is predominant. Gravity can also assist in the recovery, depending on the fluids and the reservoir geometry. Detailed numerical simulations of these processes are often performed to predict oil recovery.

When oil and gas are recovered during primary production and depletion, oil, gas, and some water are expelled from the reservoir, thereby reducing the reservoir pressure and causing the remaining oil, gas, and water to expand so as to fill the pore spaces vacated by the fluids removed. When the reservoir is hydraulically connected to aquifers or water-bearing layers, water may flow or influx into the oil reservoir as the reservoir pressure drops during production. The recovery factor for all of these processes depends largely on the reservoir heterogeneities and properties, reservoir pressure and temperature, reservoir fluid composition, production well patterns, and economics.

The goal of any recovery process is to maximize the oil recovery at the stock tank. The formation volume factor is a convenient measure of the volume of oil that may exist at stock tank conditions compared with its volume *in situ*. For example, for an undersaturated black oil (i.e., an oil that exists at a reservoir pressure above its bubble point pressure), the formation volume factor is always greater than 1.0 because the oil contains solution gas. Typical formation volume factors for black oils in oilfield units are approximately 1.2 to 1.6 reservoir barrels (RB)/stock tank barrel (STB). Volatile oils have even higher formation volume factors than do black oils; thus, shrinkage is even more of a concern.

As oil is produced, the oil pressure is lowered below its bubble point pressure and solution gas evolves from the oil, shrinking the oil volume (see Fig. 3). The solution GOR is a measure of the volume of solution gas in standard cubic feet (SCF) per STB.

5.1 Solution Gas Drive

Many oil reservoirs are initially undersaturated with dissolved gas in that the reservoir pressure is above the bubble point pressure. For such reservoirs, if

there is no water influx from the periphery of the reservoirs, the oil and water compressibilities provide the primary energy for recovery. That is, the liquids expand as pressure is decreased, causing oil to flow to the production wells and to the surface. Because the fluid compressibilities are small, there is often a rapid decline in average reservoir pressure during the liquid expansion period until the bubble point pressure is reached. Often, little water is produced because the initial or connate water saturation is typically small, making the resistance to water flow large.

Once the average reservoir pressure drops below the bubble point pressure, solution gas evolves from the oil. The quantity of gas that evolves depends on the oil composition, that is, whether it is a black oil or a volatile oil. The solution gas that develops in the pores provides additional energy beyond liquid expansion with which to recover oil. Because gas is more compressible than liquids, the average reservoir pressure does not decrease as rapidly as it did when the pressure was above the bubble point pressure. When the reservoir pressure eventually becomes very low, the only remaining source of energy for recovery is often gravity drainage of oil to the wells. Gravity drainage occurs slowly and pumping is often required to flow oil to the surface.

Volumetric and material balances are often used to determine the initial oil in place and to give an approximate estimate of the potential oil recovery. For a reservoir under pressure depletion with little water influx, produced oil must be replaced by gas. Thus, the gas saturation increases while the oil saturation decreases; therefore, the volume of expansion is equal to the volume of production. At abandonment (i.e., when the oil is no longer recovered in economical quantities), the recovery in STB per acre-foot is given by the difference in the OOIP and the oil in place at abandonment,

$$7758\phi\left[\frac{(1-S_w)}{B_{oi}}-\frac{\left(1-S_w-S_g\right)}{B_o}\right],$$

where B_{oi} is the initial formation volume factor of the reservoir oil, S_w is the initial water saturation, ϕ is the porosity, B_o is the formation volume factor of the oil at abandonment conditions, and S_g is the gas saturation at abandonment conditions. Recoveries from solution gas drives are typically small and in the range of 10 to 30% of the OOIP.

An accurate determination of the recovery factors requires measuring the fluid saturations. Core experiments or core samples taken from the reservoir are often used to measure the water and gas saturations that go into the volumetric balances. Other techniques for measuring saturations include tracer tests and (more commonly) well logging.

5.2 Gas Cap Drive

Some oil reservoirs occur at conditions below the bubble point pressure so that an initial gas cap has formed over geological time. In such cases, the oil is initially saturated with dissolved gas and the production is primarily controlled by the size and expansion of the gas cap (Fig. 5). Because gas compressibility is large compared with liquid compressibilities, a small gas cap can provide significant pressure maintenance and contribute to increased oil recovery. For example, as production proceeds and the reservoir pressure is slowly reduced, the gas cap expands, displacing oil downward toward the perforated intervals of the production wells. The recovery efficiency is particularly enhanced if the reservoir is tilted so that gravity maintains a stable gas/oil interface during production. The field life can be extended during the later stages of production by recompleting deeper or sealing the perforation interval of high-GOR wells. The ultimate recoveries by gas cap expansion can be in excess of 50% of the OOIP with gravity assistance but are typically in the range of 25 to 35%.

5.3 Natural Water Drive

For reservoirs that have natural water influx (water drive), the reservoir pressure during oil production can often be maintained, providing a significant increase in energy in which to recover oil. The rate of pressure decline will depend on the volume of water that flows inward to the reservoir from its boundaries compared with the volume of oil, water, and gas that is recovered or expelled via production wells. High recoveries are obtained especially for tilted reservoirs where the water can move vertically upward, thereby displacing the oil (Fig. 5).

Like gas displacing oil downward during a gas cap drive, water that flows into the reservoir displaces oil, but a significant amount of oil can be left behind when it is trapped by capillary forces in what is termed residual oil. Therefore, for natural water drive reservoirs, the recovery is limited by the residual oil saturation. Residual oil saturations depend on the geometry of the pores, among other factors, and are commonly in the range of 10 to 40%.

A simple volumetric balance can sometimes be used to estimate the maximum recovery for a water

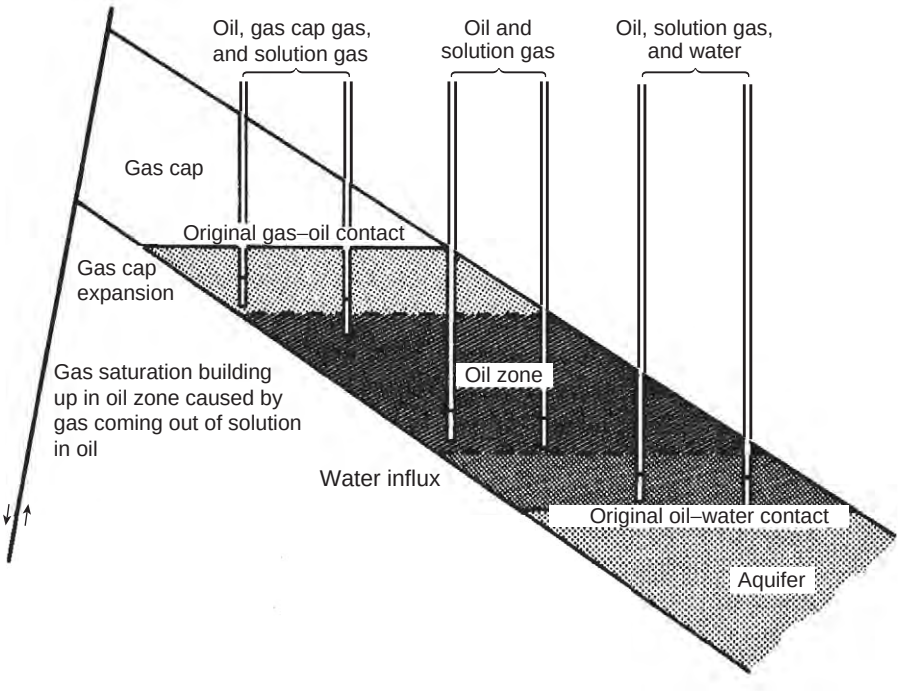


FIGURE 5 Primary recovery for a tilted reservoir under simultaneous gas cap drive and water drive. From Craft and Hawkins (1991).

drive. For example, when the reservoir pressure is maintained at high levels, no significant amount of gas evolves in the reservoir during production and the recovery in STB per acre-foot is given by

$$\frac{7758\phi(1 - S_w - S_{or})}{B_{oi}},$$

where S_{or} is the residual oil saturation. The recovery efficiency for natural water drive reservoirs is typically greater than that under solution gas drive and can be in excess of 50% of the OOIP.

5.4 Compaction Drive

The compressibility of the rock can also affect recovery, but typically this type of drive is secondary to other mechanisms. However, for some shallow reservoirs, compaction drives can provide a significant increase in primary recovery. Ultimate recovery factors of up to 50% for shallow reservoirs have been reported in a few cases.

The principles behind compaction drives are simple, but they are often difficult to model accurately. As the pressure is reduced, the pore volume of the rock may decrease (i.e., compact), and this can provide energy to squeeze the oil and displace it toward production wells. Compaction can lead to subsidence at the surface, and this can be a major problem.

6. SECONDARY AND TERTIARY RECOVERY PROCESSES

Secondary and tertiary recovery processes are typically defined as those that add external energy to the reservoir to recover additional oil or gas. The most common type of secondary recovery process is that of water flooding, in which water is injected into the reservoir and displaces oil toward the production wells. As the name implies, water flooding as a secondary process typically occurs after primary production, although this is not necessarily so. Other secondary and tertiary processes that are commonly used include gas injection (both miscible and immiscible), microemulsion and polymer flooding, *in situ* combustion, and steam flooding.

Tertiary recovery processes often use EOR technology. EOR processes are those that stimulate oil recovery by the injection of materials not normally present within the reservoir. EOR techniques allow the economic value of existing fields to be maximized through increased oil recovery and field life extension. There are also environmental benefits associated with the use of EOR, as described previously.

6.1 Factors Controlling Recovery Efficiency

The recovery efficiency that results from secondary and tertiary recovery processes is the product of the displacement efficiency and the volumetric sweep efficiency (i.e., $E_R = E_D E_V$). The volumetric sweep efficiency E_V is defined as the ratio of the oil volume contacted by the recovery process to the volume of OOIP (Fig. 6). The fraction of oil that is contacted by the recovery process and ultimately recovered or displaced is defined as the displacement efficiency (E_D).

6.1.1 Reservoir Mixing

Mixing or dilution of injected fluids with reservoir fluids can cause the displacement efficiency E_D to decrease. Mixing in a reservoir can be caused by dispersion, diffusion, or cross-flow. Cross-flow is the process by which fluids move from one reservoir layer to another due to capillary, viscous, or gravity forces. Diffusion is the spreading of components in a single-phase fluid caused by random molecular movement. Dispersion is the spreading of components in a single-phase fluid caused by streamline splitting and dead-end pores during advection.

6.1.2 Fluid Mobility

The sweep efficiency E_V depends on the mobility of the injected fluids as well as the well pattern. The mobility is defined as the ratio of the permeability to the fluid viscosity. Injected fluids of high mobility can flow through the high-permeability layers of the reservoir more easily. Conformance methods are often used to improve sweep efficiency by controlling where fluids are injected (or produced) or by injecting fluids that alter how the fluids move through the reservoir layers. Gels, polymers, and foams are examples of such fluids.

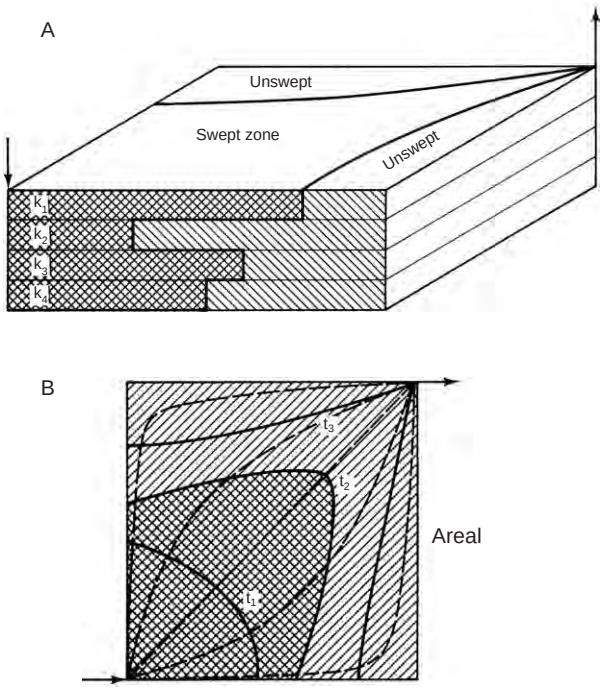


FIGURE 6 Schematic of sweep efficiency in a layered reservoir: (A) vertical and (B) areal. From Lake (1989).

6.1.3 Fluid Density

The density of the injected fluids also affects the sweep efficiency. Light fluids can rapidly migrate upward in a gravity tongue, bypassing oil, while denser fluids can migrate downward. In such cases, heterogeneities may improve the sweep efficiency by not allowing fluids to migrate vertically upward (or downward). Water alternating gas injection (WAG), in which water and gas are injected in alternating cycles, is often used to improve the sweep efficiency by increasing the mixing of reservoir fluids with injected fluids. However, increased mixing can significantly reduce the displacement efficiency. Typically, there is a trade-off between the displacement and sweep efficiencies in maximizing oil recovery.

6.1.4 Multiphase Flow and Relative Permeability

Darcy's law governs flow of oil, water, and gas in porous media. Darcy's law says that the flow rate at any point in the reservoir is given by the fluid pressure gradient, the viscosity of the fluid, and its effective fluid permeability. That is, when gravity is neglected,

$$q = \frac{k}{\mu}\frac{dp}{dx},$$

where q is the volumetric flux of the fluid in meters per second, k is the effective permeability of the fluid, μ is the fluid viscosity, and dp/dx is the fluid pressure gradient. Therefore, for oil, the volumetric flux through the reservoir will be smaller for viscous or heavier oils than for lighter oils.

The effective permeability is the product of the absolute permeability of the rock and the relative permeability. The absolute permeability of the rock is typically measured with 100% water or gas occupying the pore space. However, when two or more phases are present in the pores, flow of each phase is restricted by the other phases. As the saturation of the water or gas phases is increased and the oil saturation decreases, the resistance for oil to flow increases until oil no longer flows at its residual oil saturation.

Flow resistances also depend on the "wettability" of the rock. For water-wet rocks, the water resides in the smaller pores and on the surfaces of the rocks, whereas the oil is in the center of the larger pores. Fluids that are in large pores can flow easier than can those in smaller pores. Reservoirs can be water wet, oil wet, or mixed wet, although the wetting nature can vary significantly within the reservoirs.

A measure of the flow resistance is given by the phase relative permeability, which is usually a number between 0 and 1. Therefore, the effective permeability of a fluid is the absolute permeability multiplied by its relative permeability. The concept of effective permeability explains why oil recovery for a water flood or an immiscible gas flood is not 100%, even when sweep is nearly 100%.

6.2 Water Flooding

Water injection as a secondary method of recovery is very similar to the recovery process that occurs from natural water influx. The main advantage here is that the volume and placement of the water can be controlled so that sweep efficiency is improved and reservoir pressure is maintained. Polymers are sometimes injected with the water to increase the water's viscosity and to decrease its mobility. Sweep efficiency can be significantly improved with polymer injection.

6.3 Immiscible Gas Injection

Immiscible gas injection in the reservoir is similar to the recovery process that occurs during a gas cap drive. Like water flooding, the volume and placement of gas can be controlled to improve sweep efficiency and maintain reservoir energy or pressure. Typical gases for immiscible flooding are methane, nitrogen, carbon dioxide, and air. Many of these gases are not completely immiscible with the oil. For example, carbon dioxide nearly always has some limited miscibility with the oil and, therefore, can swell the oil and reduce its viscosity, both of which can improve recovery. However, carbon dioxide is relatively expensive to inject as an immiscible gas and is generally not used in this way today.

6.4 Miscible Gas Injection

A much more efficient and commonly used method today is to inject gas that is miscible with the oil at the reservoir temperature and pressure of interest. The oil recovery for miscible gas floods can increase significantly beyond primary recovery processes because the interfacial tension between the residual oil and injected gas is reduced to zero. As the interfacial tension between the fluids is reduced, capillary pressure is decreased and the residual oil saturation decreases in many cases to values of less than 1 or 2%.

One miscible fluid used today is carbon dioxide, which is injected as a supercritical fluid at temperatures and pressures beyond its critical point (Fig. 7). Other gases that can be miscible with oil are methane and nitrogen, although these gases require much higher reservoir pressures to achieve miscibility. The type of gas used also depends on its cost and its availability at the field site.

There are two types of miscibility that can occur between resident oil and injected gas. Fluids can be multicontact miscible or first-contact miscible. First-contact miscibility occurs when the gas and oil are miscible when mixed in any proportions. This type of miscibility is ideal but is seldom achieved given the cost and availability of gas components in the field. Multicontact miscibility occurs more often and is the process by which gas and oil develop miscibility in the reservoir by repeated contact with each other. Historically, the multicontact miscible process was thought to occur either by vaporizing contacts or by condensing contacts. In vaporizing drives, fresh gas mixes with equilibrium oil and vaporizes the intermediate-weight components in the oil, whereas in condensing drives, fresh oil mixes with equilibrium gas and condenses intermediate-weight components in the gas into the oil. Vaporizing drives occur when components, such as methane and nitrogen, are injected at pressures above the minimum pressure for miscibility (or MMP). Condensing drives were thought to occur when gases are sufficiently enriched with hydrocarbon components, such as ethane and propane, to a composition called the minimum enrichment for miscibility (or MME).

The drive mechanisms for both enriched hydrocarbon gas floods and carbon dioxide floods have been shown recently to be combined condensing/vaporizing. In the combined condensing/vaporizing process, the gas at the leading edge of the gas front swells oil, while residual oil is vaporized by gas at the trailing edge. Miscibility is approached in the reservoir between the condensing and vaporizing regions.

Multicontact miscibility is only approached, and is never actually attained, in a reservoir. Dilution of the gas by mixing of resident oil with injected gas

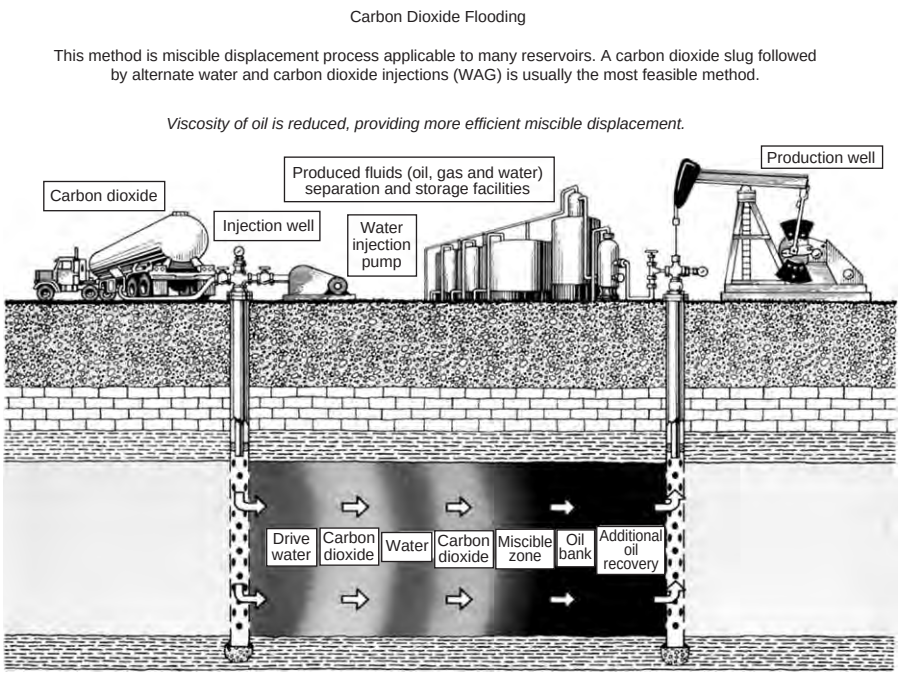


FIGURE 7 Schematic of a miscible gas flooding process. Drawing by Joe Lindley, U.S. Department of Energy, Bartlesville, OK. From Lake (1989).

always causes some two-phase flow or partial miscibility to occur. Fortunately, achieving near miscibility often is sufficient to greatly enhance recovery by a reduction in residual oil saturation. Floods are generally defined as miscible when the reservoir pressure is above the thermodynamic MMP or the composition of the injected gas is beyond the MME.

The estimation of the MMP and MME has been historically determined from slim-tube experiments, numerical simulations with fine grids, or correlations. In slim-tube experiments, oil recovery from the injection of gas into rock samples of thin diameter is plotted as a function of the pressure or enrichment. A bend in the recovery curve is thought to occur at the MMP or MME. However, both slim-tube experiments and numerical simulations can have a large uncertainty in these parameters due to the effect of dispersion on reducing the recovery. Recent methods have allowed for analytical calculation of these important parameters, but these methods rely on an accurate fluid characterization from an EOS.

6.5 Microemulsion Flooding

Microemulsion flooding is the injection of a microemulsion into the reservoir to reduce the interfacial tension that traps residual oil. Microemulsions are surfactant stabilized dispersions of water and hydrocarbons. Micelles are formed at sufficiently high surfactant concentrations that are capable of solubilizing fluid within their core centers. One significant advantage of microemulsions is that they have viscosities significantly greater than water and, hence, can improve sweep efficiency.

Microemulsion flooding is more complex to perform than are other EOR methods and, thus, requires significant technical expertise. The injected slug of microemulsion depends on many parameters, including reservoir temperature, water salinity, and crude oil type. Because of its complexity, cost, and other concerns, microemulsion flooding is not used often for oil recovery. However, an extension of this method, called surfactant enhanced aquifer remediation (SEAR), is being used today to effectively clean up shallow contaminated aquifers that otherwise would remain contaminated.

6.6 *In situ* Combustion

In situ combustion is a process in which ignition of coke deposits left by the crude oil is initiated at an air injection well, causing a combustion front to propagate into the reservoir (Fig. 8). The coke region is the zone where carbonaceous material has been deposited as a result of thermal cracking of the crude

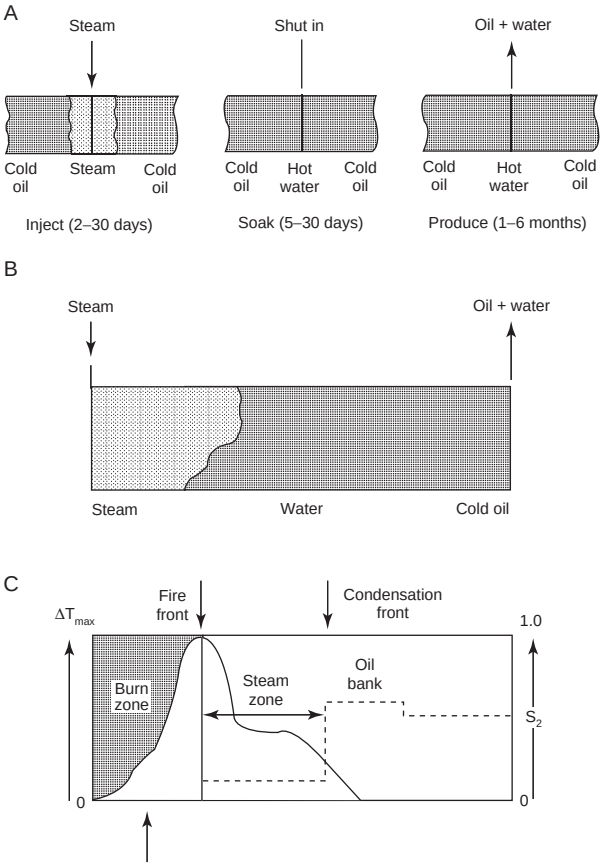


FIGURE 8 Thermal methods for improved recovery: (A) steam soak (or huff n' puff), (B) steam drive, and (C) *in situ* combustion. The oil saturation is denoted as S_2. From Lake 1989, *Enhanced Oil Recovery.*

oil. In the zone immediately ahead of the hot combustion zone, the temperature is increased significantly and reaches a plateau based on the condensation of water at the reservoir pressure.

Recovery is increased primarily due to two main factors. A steam zone develops in front of the combustion zone, causing a steam flood of the oil. A small amount of carbon dioxide is also formed, increasing vaporization and swelling of the oil. *In situ* combustion is not widely used today primarily due to the economic risk involved in burning the oil and keeping the combustion front ignited. However, air injection as an immiscible gas flood has been reported to accidentally cause *in situ* combustion, and this process may occur more often than is realized.

6.7 Steam Flooding

In contrast to *in situ* combustion, steam flooding is a method that relies on the external injection of steam or hot water into the reservoir. Two commonly used forms of steam injection are the steam drive, in which steam is continuously injected, and steam soak or cyclic steam injection, which involves repeated injection of steam and water (Fig. 8). Cyclic steam injection is also sometimes known as "huff n' puff." Steam aids oil recovery by reduction in the oil viscosity, reduction in the residual oil saturation, thermal expansion effects, and steam distillation. Steam flooding is very common for heavy oil reservoirs in which steam causes a significant reduction in the viscosity of the heavy oils. Steam flooding is currently the most used EOR technology and has enjoyed significant commercial success.

SEE ALSO THE FOLLOWING ARTICLES

Crude Oil Spills, Environmental Impact of • Oil and Natural Gas Drilling • Oil and Natural Gas Exploration • Oil and Natural Gas: Offshore Operations • Oil Pipelines • Oil Refining and Products • Petroleum System: Nature's Distribution System for Oil and Gas

Further Reading

Bradley, H. B. (ed.). (1987). "Petroleum Engineering Handbook." Society of Petroleum Engineers, Richardson, TX.

British Petroleum. (2003). "Statistical Review of World Energy." BP, London.

Craft, B. C., and Hawkins, M. F. (revised by R. E. Terry) (1991). "Applied Petroleum Reservoir Engineering," 2nd ed. Prentice Hall, Englewood Cliffs, NJ.

Craig, F. F., Jr. (1971). "The Reservoir Engineering Aspects of Waterflooding," SPE Monograph Series, vol. 3. Society of Petroleum Engineers, Richardson, TX.

Firoozabadi, A. (1999). "Thermodynamics of Hydrocarbon Reservoirs." McGraw–Hill, New York.

Jarrell, P. M., Fox, C. E., Stein, M. H., and Webb, S. L. (2002). "Practical Aspects of CO_2 Flooding," SPE Monograph Series, vol. 22. Society of Petroleum Engineers, Richardson, TX.

Lake, L. W. (1989). "Enhanced Oil Recovery." Prentice Hall, Englewood Cliffs, NJ.

McCain, W. D., Jr. (1990). "The Properties of Petroleum Fluids." 2nd ed. PennWell Publishing, Tulsa, OK.

Pederson, K. S., Fredenslund, A., and Thomassen, P. (1989). Properties of Oils and Natural Gases, Contributions in Petroleum Geology and Engineering, vol. 5. Gulf Publishing, Houston, TX.

Petroleum Extension Service. (1983). "Oil and Gas: The Production Story." University of Texas at Austin, Austin, TX.

Petroleum Extension Service. (2001). "A Primer on Oil Well Drilling," 6th ed. University of Texas at Austin, Austin, TX.

Stalkup, F. I., Jr. (1984). "Miscible Displacement," SPE Monograph Series. Society of Petroleum Engineers, Richardson, TX.

Oil Refining and Products

ABDULLAH M. AITANI
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia

1. **Overview of Refinery Processing**
2. **Crude Oil and Products**
3. **Light Oil Processing**
4. **Heavy Distillate Processing**
5. **Residual Oil Processing**
6. **Treating Processes**
7. **Environmental and Future Issues**

Glossary

alkylation A process using sulfuric acid or hydrofluoric acid as a catalyst to combine light olefins and isobutane to produce a high-octane product known as alkylate.

°API gravity A scale of liquid specific gravity (SG) that indicates the lightness or heaviness of hydrocarbons, defined by [(141.5/SG) −131.5].

catalytic cracking A process for the breaking-up of heavier hydrocarbons into lighter hydrocarbon fractions by the use of heat and catalysts.

cetane number A measure of ignition quality for kerosene, diesel, and heating oil, using a single-cylinder engine.

coking A process for thermally converting and upgrading heavy residues into lighter products and by-product petroleum coke.

crude oil A complex mixture of hydrocarbons containing low percentages of sulfur, nitrogen, and oxygen compounds and trace quantities of many other elements.

deasphalting A process for removing asphaltic materials from reduced crude, using liquid propane to dissolve nonasphaltic compounds.

hydrocracking A process used to convert heavier feedstock into lower boiling point, higher value products. The process employs high pressure, high temperature, a catalyst, and hydrogen.

hydrodesulfurization A catalytic process for the removal of sulfur compounds from hydrocarbons using hydrogen.

isomerization A catalytic process for the conversion and skeletal rearrangement of straight-chain hydrocarbons into branched-chain molecules of higher octane number.

methyl tertiary butyl ether (MTBE) An ether added to gasoline to raise octane number and enhance combustion.

octane number A measure of resistance to knocking of gasoline under laboratory conditions that simulate city driving conditions.

olefins Unsaturated hydrocarbons, such as ethylene and propylene, that have a double carbon bond, with the molecular formula C_nH_{2n}.

paraffins Saturated aliphatic hydrocarbons with the molecular formula C_nH_{2n+2}.

reforming A process for the transformation of naphtha into products with higher octane number. Reforming comprises isomerization, cracking, polymerization, and dehydrogenation.

visbreaking A low-temperature cracking process used to reduce the viscosity or pour point of straight-run residues.

This article discusses the various aspects of petroleum refining and oil products as a primary energy source and as a valuable feedstock for petrochemicals. The main objective of refining is to convert crude oils of various origins and different compositions into valuable products and fuels having the qualities and quantities demanded by the market. Various refining processes, such as separation, conversion, finishing, and environmental protection, are presented and briefly discussed. The ever-changing demand and quality of fuels, as well as environmental issues and the challenges facing the refining industry, are also highlighted. Environmental regulations have played a significant role in the progress of the refining industry and may even change the competition between petroleum and other alternative energy sources.

1. OVERVIEW OF REFINERY PROCESSING

1.1 Introduction

Refining is the processing of crude oil into a number of useful hydrocarbon products. Processing utilizes

 © 2004 Elsevier Inc. All rights reserved.

chemicals, catalysts, heat, and pressure to separate and combine the basic types of hydrocarbon molecules naturally found in crude oil into groups of similar molecules. The refining process also rearranges their structures and bonding patterns into different hydrocarbon molecules and compounds. Therefore, it is the type of hydrocarbon (paraffinic, naphthenic, or aromatic) and its demand that configure the refining industry. Petroleum refining has evolved continuously in response to changing demands for better and different products. The trend in demand has also been accompanied by continuous improvement in product quality, such as octane number for gasoline and cetane number for diesel.

The original requirement was to produce kerosene for household use, followed by the development of the internal combustion engine and the production of transportation fuels (gasoline, diesel, and fuels). Refineries produce a variety of products including many required as feedstocks for the petrochemical industry. Early refining was the simple distillation (fractionation) of crude oil followed by the development in the 1920s of the thermal cracking processes, such as visbreaking and coking. The processes crack heavy fuels into more useful and desirable products by applying pressure and heat.

In the early 1940s, the catalytic processes were developed to meet the increasing demand of gasoline and higher octane numbers. Processes such as catalytic cracking, alkylation, isomerization, hydrocracking, and reforming were developed throughout the 1960s to increase gasoline yields and improve antiknock characteristics. Some of these processes also produced valuable feedstocks for the modern petrochemical industry. In the 21st century, the refinery uses an array of various catalytic and noncatalytic processes to meet new product specifications and to convert less desirable fractions into more valuable liquid fuels, petrochemical feedstocks, and electricity. The refinery has shifted from only physical separations to something close to a chemical plant.

1.2 Refinery Operations

Modern refineries incorporate fractionation, conversion, treatment, and blending operations and may also include petrochemical processing. Most light distillates are further converted into more usable products by changing the size and structure of the hydrocarbon molecules through cracking, reforming, and other conversion processes, as discussed further in this article. Figure 1 presents a typical scheme of a high-conversion refinery.

Various streams are subjected to various separation processes, such as extraction, hydrotreating, and sweetening, to remove undesirable constituents and improve product quality. In general, petroleum refining operations can be grouped as follows:

- Fractionation (distillation) is the separation of crude oil in atmospheric and vacuum distillation towers into groups of hydrocarbon compounds of differing boiling-point ranges called “fractions” or “cuts.”
- Light oil processing prepares light distillates through rearrangement of molecules using isomerization and catalytic reforming or combination processes such as alkylation and polymerization.
- Heavy oil processing changes the size and/or structure of hydrocarbon molecules through thermal or catalytic cracking processes.
- Treatment and environmental protection processes involve chemical or physical separation, such as dissolving, absorption, or precipitation, using a variety and combination of processes including drying, solvent refining, and sweetening.

1.3 World Refining

World refining capacity reached 81.9 million barrels/day in 2002 and is expected to increase by 4.3% per year to 115 million barrels/day by 2020. Table I presents a regional distribution capacity of major refining processes. There are 722 refineries in 116 countries, with more than 200 refineries in Asia-Pacific regions alone. The United States has maintained its leading position as the largest and most sophisticated oil-refining region. Approximately 25% of worldwide refining capacity is located in North America and another 25% is located in Asia followed by 17% in Western Europe. Remaining regions process approximately 33% of the world’s crude oil in medium-type conversion refining schemes. World capacities of various refining processes are presented in Table II. Hydrotreating alone represents approximately 50% of total crude capacity, whereas all catalytic processes represent approximately 82% of crude capacity.

In general, the refining industry has always been characterized as a high-volume, low-profit-margin industry. World refining continues to be challenged by uncertainty of supply, difficult market conditions, government regulation, availability of capital, and slow growth. Although shipping of refined products has been increasing over the years, a close connection remains between domestic markets and domestic

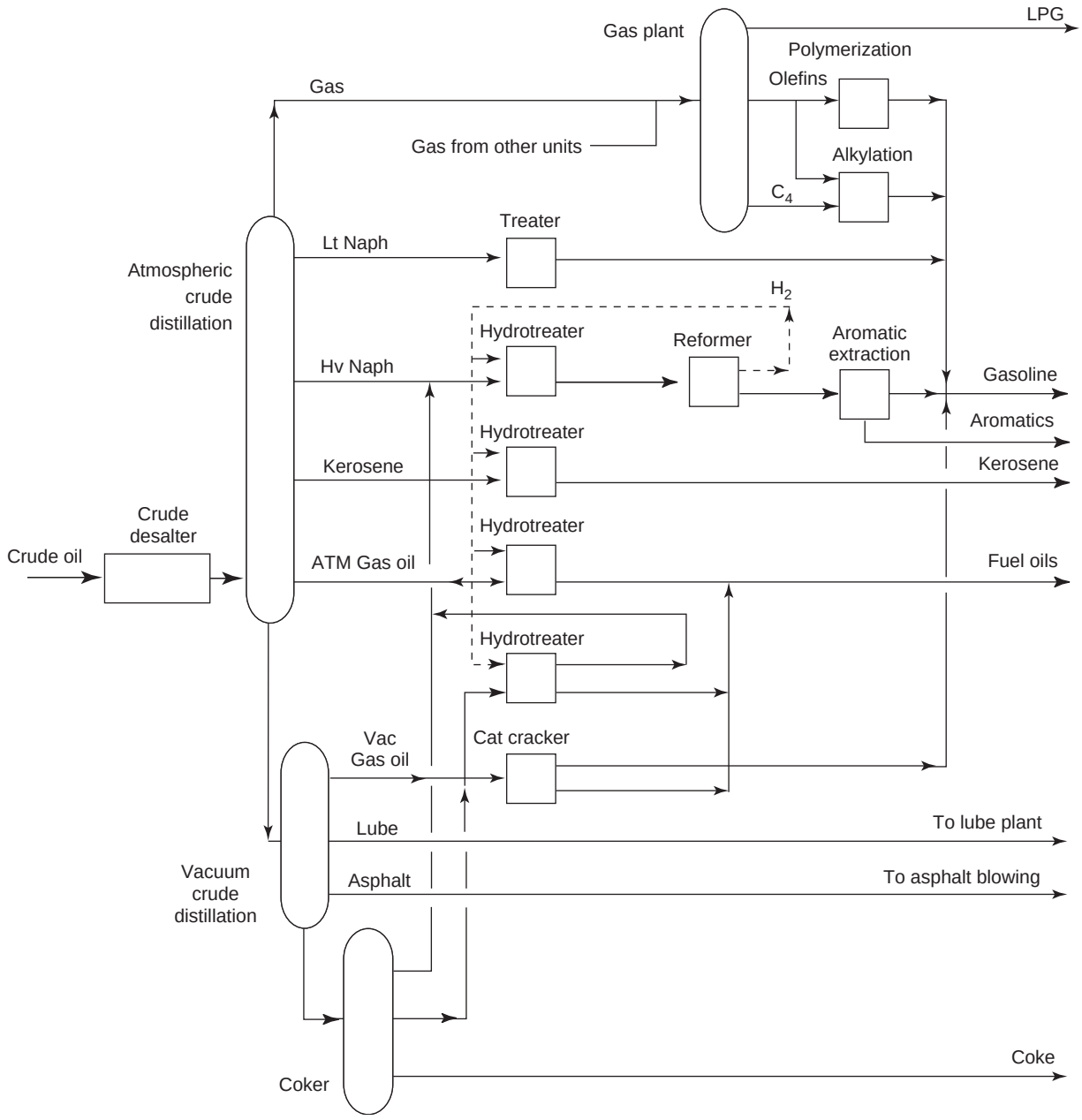


FIGURE 1 Schematic diagram of a high-conversion refinery. Cat cracker, catalytic cracker; Hv Naph, heavy naphtha; Lt Naph, light naphtha; LPG, liquefied petroleum gas; Vac gas oil, vacuum gas oil; ATM gas oil, atmospheric gas oil.

production. This explains the large differences in refinery schemes from one country to another and from one region to another.

2. CRUDE OIL AND PRODUCTS

2.1 Type and Composition of Crude Oils

Crude oil is a mixture of hydrocarbon compounds such as paraffins, naphthenes, and aromatics plus small amounts of organic compounds of sulfur, oxygen, and nitrogen, in addition to small amounts of metallic compounds of vanadium, nickel, and sodium. Although the concentration of nonhydrocarbon compounds is very small, their influence on catalytic petroleum processing is large. There are a large number of individual components in a specific crude oil, reaching approximately 350 hydrocarbons and approximately 200 sulfur compounds. A specific crude oil can comprise a very large number of

TABLE I

Regional Distribution of Refining Operations Worldwide

Region	No. of refineries	Regional capacity	
		million barrels/day	%
North America	160	20.3	24.8
South America	69	6.7	8.2
Western Europe	105	14.6	17.8
Eastern Europe	95	10.6	12.9
Asia	202	20.2	24.7
Middle East	46	6.3	7.7
Africa	45	3.2	3.9
Total	722	81.9	100

TABLE II

World Refining Processes and Their Share of Crude Oil Capacity

Process	Capacity (million barrels/day)	% of crude capacity
Vacuum distillation	26.7	32.6
Coking	4.2	5.1
Visbreaking	3.7	4.5
Catalytic cracking	14.2	17.3
Naphtha reforming	11.2	13.7
Hydrocracking	4.4	5.4
Hydrotreating	38.4	46.9
Alkylation	1.9	2.3
Polymerization	0.2	0.2
Aromatics production	1.2	1.5
Isomerization	1.5	1.8
Oxygenates	0.3	0.4

compounds that are not easily identifiable or quantifiable. Most of these compounds have a carbon number less than 16 and these form a relatively high proportion of crude oil.

The elemental composition of crude oils depends on the type and origin of the crude; however, these elements vary within narrow limits. The proportions of these elements in a typical crude are 84.5% carbon, 13% hydrogen, 1–3% sulfur, and less than 1% each of nitrogen, oxygen, metals, and salts. The physical properties of crude oils vary within a wide range. Crude oils are defined in terms of API (American Petroleum Institute) gravity: the higher the API gravity, the lighter the crude. Crude oils with low carbon, high hydrogen, and high API gravity are usually rich in paraffins and tend to yield greater proportions of gasoline and light petroleum products; those with high carbon, low hydrogen, and low API gravities are usually rich in aromatics.

Crude oils can be classified in many different ways, generally based on their density (API), sulfur content, and hydrocarbon composition. Condensate ranks highest, with densities reaching more than 50° API, whereas densities of heavy crudes may reach as low as 10° API. In general, refinery crude base stocks consist of mixtures of two or more different crude oils. There are more than 600 different commercial crude oils traded worldwide. In 2002, world oil production reached 66 million barrels/day, 40% of which is produced by members of the Organization of Petroleum Exporting Countries. Despite all energy alternatives, crude oil will remain the world's primary energy source, constituting approximately 40% of world energy up to the year 2020.

2.2 Crude Oil Processing

As a first step in the refining process, water, inorganic salts, suspended solids, and water-soluble trace metal contaminants are removed by desalting using chemical or electrostatic separation. This process is usually considered a part of the crude distillation unit. The desalted crude is continuously drawn from the top of the settling tanks and sent to the crude fractionation unit. Distillation of crude oil into straight-run cuts occurs in atmospheric and vacuum towers. The main fractions obtained have specific boiling-point ranges and can be classified in order of decreasing volatility into gases, light distillates, middle distillates, gas oils, and residue. The composition of the products is directly related to the characteristics of the crude processed. Desalted crude is processed in a vertical distillation column at pressures slightly above atmospheric and at temperatures ranging from 345 to 370°C (heating above these temperatures may cause undesirable thermal cracking). In order to further distill the residue from atmospheric distillation at higher temperatures, reduced pressure is required to prevent thermal cracking. Vacuum distillation resembles atmospheric distillation except that larger diameter columns are used to maintain comparable vapor velocities at the reduced pressures.

2.3 Transportation Fuels

Major oil products are mainly transportation fuels that represent approximately 52% of total worldwide oil consumption. Gasoline and diesel are the main concern, along with a large number of special

petroleum-based products. Crude oil type and refining configuration determine the quantity and quality of oil products. Tables III and IV present some typical data on the volume of atmospheric distillate cuts and refined products derived from the processing of Arabian light crude oil. For example, approximately 80–85 vol% of the refined products produced in a medium-type conversion refinery have a boiling temperature lower than 345°C compared to an amount of 55 vol% found in Arabian crude oil (see Tables III and IV). Almost half of the products are gasoline and lighter distillates. The demand for transportation fuels and petrochemical feedstocks has been increasing steadily compared to the decreasing demand for heating fuels and residual fuel oil, which are being replaced by natural gas.

2.3.1 Gasoline

Motor gasoline is the highest volume refinery product, having a mixture of hydrocarbons with boiling points ranging from ambient temperature to approximately 205°C. It flows easily, spreads quickly, and may evaporate completely in a few hours under temperate conditions. It is highly volatile and flammable. Gasoline is made up of different refinery streams, mainly straight-run naphtha, isomerized C_5/C_6 paraffins, reformate, hydrocracking, fluid catalytic cracking (FCC) gasoline, oligomerate, alkylate, and ethers. The most environmentally friendly gasoline comes from branched paraffins. The important qualities for gasoline are octane number (antiknock), volatility (starting and vapor lock), vapor pressure, and sulfur content (environmental control). Additives are often used to enhance gasoline performance and to provide protection against oxidation and rust formation. Table V presents some typical data for current and future specifications for gasoline in Europe.

TABLE III

Atmospheric Distillates Derived from Arabian Light Crude Oil

Product name	Boiling point (°C)	Volume (%)
Light naphtha	10–90	8
Heavy naphtha	90–200	21
Kerosene	200–260	11
Gas oil	260–345	15
Residue	345+	45

TABLE IV

Typical Refined Products Derived from Arabian Light Crude Oil (Medium-Conversion Refinery)

Product name	Carbon no.	Boiling point (°C)	Volume (%)
Gases	1–4	−40–40	5
Gasoline	5–10	40–200	45
Kerosene/jet	10–16	150–260	5
Fuel oil	20–70	200–345	25
Residue	>70	>345	20

2.3.2 Diesel Fuel

Diesel fuel is usually second in volume next to gasoline. Diesel blend consists of cuts from atmospheric distillation, hydrocracking, FCC light cycle oil, and some products obtained from visbreaking and coking. The main property of diesel fuel for automotive engine combustion is cetane number, which is a measure of engine start-up and combustion. Diesel fuel and domestic heating oils have boiling point ranges of approximately 200–370°C. The desired properties of these distillate fuels include controlled flash and pour points, clean burning, and no deposit formation in storage tanks. Sulfur reduction and cetane improvement have been extensively investigated to produce ultralow-sulfur diesel (ULSD). Meeting future specifications for ULSD of 10–15 ppm sulfur will require significant hydrotreating investment. Table VI presents some typical data for current and future specifications for diesel fuel in Europe.

2.3.3 Jet Fuel (Kerosene)

Jet fuel is the third most important transportation fuel. It is a middle-distillate product that is used for jets (commercial and military) and is used around the world in cooking and heating (kerosene). When used as a jet fuel, some of the critical qualities are freeze

TABLE V

Current and Future Specifications for Gasoline

	Year	
Specification[a]	2000	2005
Sulfur (ppm)	150	30–50
Benzene content (vol%, maximum)	1	1
Aromatics (vol%, maximum)	42	35
Olefins (vol%, maximum)	18	15
Oxygen (wt%, maximum)	2.7	2.3
RVP (kPa)	60	50
RON/MON (minimum)	95/85	95/85

[a]RVP, Reid vapor pressure; RON, research octane number; MON, motor octane number.

TABLE VI
Current and Future Specifications for Diesel Fuel

Specification	Year 2000	2005
Sulfur (ppm)	350	50
Specific gravity (maximum)	0.845	0.825
API (minimum)	36	40
Cetane number	51	54
Distillation T_{95} (°C, maximum)	360	360
Polycyclic aromatic hydrocarbons (PAH wt%, maximum)	11	1

point, flash point, and smoke point. Commercial jet fuel has a boiling point range of approximately 190–275°C and that of military jet fuel is 55–285°C. Kerosene, with less critical specifications, is used for lighting, heating, solvents, and blending into diesel fuel. *n*-Paraffins in the range C_{12}–C_{14} may be extracted from kerosene for use in the production of detergents.

2.4 Other Oil Products

2.4.1 Refinery Gases

Refinery gases are the lightest hydrocarbons containing a mixture of gases from methane to liquefied petroleum gas (LPG) and some pentanes. The gases are processed to separate LPG, which consists principally of propane and butane. Other refinery gases include lighter paraffins, unsaturates, and hydrogen sulfide. LPG is used as fuel and as an intermediate in the manufacture of olefins and selected petrochemical feedstocks. Butanes are also used in the manufacture of ethers and to adjust the vapor pressure of gasoline. LPG is also used in transportation and in domestic and household applications.

2.4.2 Petrochemical Feedstocks

Light olefins from FCC and benzene, toluene, xylenes (BTX) aromatics from naphtha reforming are the main petrochemical feedstocks derived from refineries. These products are the basis for integrating refining and petrochemical operations. Olefins include propylene and butanes, whereas benzene and xylenes are precursors for many valuable chemicals and intermediates, such as styrene and polyesters.

2.4.3 Residual Fuel Oil

Residual fuel oil is the least valuable of the refiner's products, selling at a price below that of crude oil. Residual fuels are difficult to pump and may be heavier than water; they are also difficult to disperse and are likely to form tar balls, lumps, and emulsions. Many marine vessels, power plants, commercial buildings, and industrial facilities use residual fuels or combinations of residual and distillate fuels for heating and processing. The two most critical properties of residual fuels are viscosity and low sulfur content for environmental control.

2.4.4 Coke and Asphalt

Petroleum coke is produced in coking units and is almost pure carbon with a variety of uses from electrodes to charcoal briquets. Asphalt is a semisolid material produced from vacuum distillation and is classified into various commercial grades. It is used mainly for paving roads and roofing materials.

2.4.5 Lubricants

Vacuum distillation and special refining processes produce lube-oil base stocks. Additives such as antioxidants and viscosity enhancers are blended into the base stocks to provide the characteristics required for motor oils, industrial greases, lubricants, and cutting oils. The most critical quality is a high viscosity index, which provides for greater consistency under varying temperatures.

3. LIGHT OIL PROCESSING

3.1 Catalytic Naphtha Reforming

Catalytic naphtha reforming combines a catalyst, hardware, and processing to produce high-octane reformate for gasoline blending or BTX aromatics for petrochemical feedstocks. Reformers are also the source of much needed hydrogen for hydroprocessing operations. Reforming reactions comprise cracking, polymerization, dehydrogenation, and isomerization, which take place simultaneously. Universal Oil Products (UOP) and Axens-Institut Français du Pétrole (IFP) are the two major licensors and catalyst suppliers for catalytic naphtha reforming. Reforming processes differ in the mode of operation [semiregenerative or continuous catalyst regenerative (CCR)], catalyst type, and process engineering design. All licensors agree on the necessity of hydrotreating the feed to remove permanent reforming catalyst poisons and to reduce the temporary catalyst poisons to low levels. There are more than 700 reformers worldwide, with a total capacity of approximately 11.2 million barrels/day. Approximately 40% of this capacity is located in

North America, followed by 20% each in Western Europe and Asia-Pacific regions.

3.1.1 Reforming Processes

Reforming processes are generally classified into semiregenerative, cyclic, and CCR. This classification is based on the frequency and mode of regeneration. The semiregenerative process requires unit shutdown for catalyst regeneration, whereas the cyclic process utilizes a swing reactor for regeneration in addition to regular reactors. The continuous process permits catalyst replacement during normal operation. Globally, the semiregenerative scheme dominates reforming capacity at approximately 57%, followed by the continuous regenerative process at 27% and the cyclic process at 11%. Most grassroots reformers are designed to use continuous catalyst regeneration.

The semiregenerative process is a conventional reforming process that operates continuously over a period of up to 1 year. Conversion is kept more or less constant by raising the reactor temperatures as catalyst activity declines. The cyclic process typically uses five or six fixed catalyst beds, similar to the semiregenerative process, with one additional swing reactor, which is a spare reactor. CCR is characterized by high catalyst activity with reduced catalyst requirements, more uniform reformate of higher aromatic content, and high hydrogen purity. Figure 2 presents a schematic diagram of a CCR process. The continuous process represents a step change in reforming technology compared to semiregenerative and cyclic processes.

3.1.2 Reforming Catalysts

Since the 1950s, commercial reforming catalysts have been essentially heterogeneous monometallic compounds and are composed of a base support material (usually chlorided alumina) on which platinum metal was placed. These catalysts were capable of producing high-octane products; however, because they quickly deactivated as a result of coke formation, they required high-pressure, low-octane operations. In the early 1970s, bimetallic catalysts were introduced to meet increasing severity requirements. Platinum and another metal (often rhenium, tin, or iridium) account for most commercial bimetallic reforming catalysts. The catalyst is most often presented as 1/16, 1/8, or 1/4 in. Al_2O_3 cylindrical extrudates or beads, into which platinum and other metal have been deposited. In commercial catalysts, platinum concentration ranges between 0.3 and 0.7 wt% and chloride is added (0.1–1.0 wt%) to the alumina support (η or γ) to provide acidity.

3.2 Isomerization

Isomerization is an intermediate, fed preparation-type process. There are more than 200 units worldwide, with a processing capacity of 1.5 million barrels/day of light paraffins. Two types of units exist: C_4 isomerization and C_5/C_6 isomerization. A C_4 unit will convert normal butane into isobutane, to provide additional feedstock for alkylation units, whereas a C_5/C_6 unit will isomerize mixtures of C_5/C_6 paraffins, saturate benzene, and remove naphtenes.

Isomerization is similar to catalytic reforming in that the hydrocarbon molecules are rearranged, but unlike catalytic reforming, isomerization just converts normal paraffins to isoparaffins. The greater value of branched paraffins over straight paraffins is a result of their higher octane contribution. The

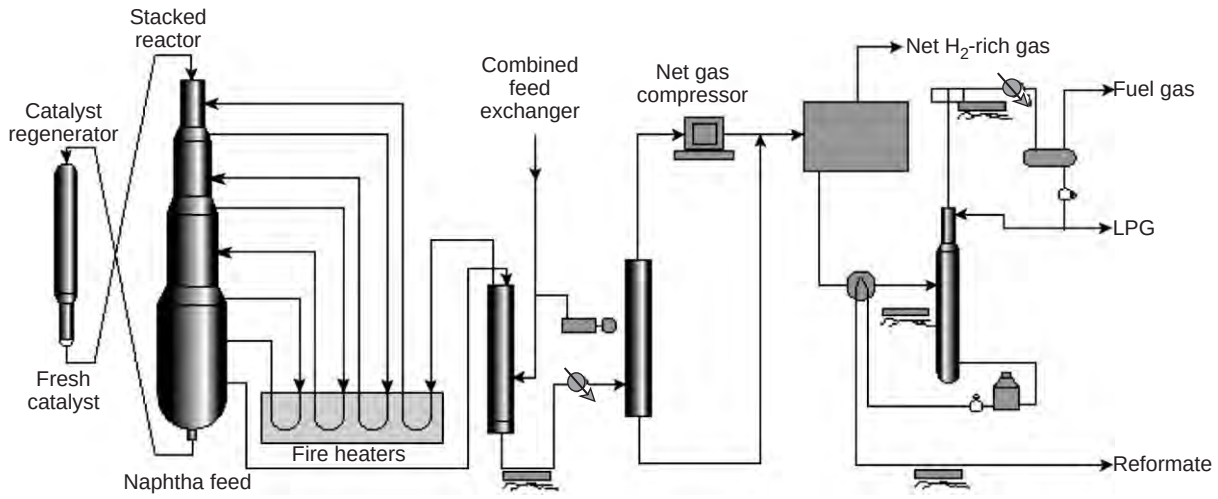


FIGURE 2 Flow scheme of a continuous catalyst regenerative naphtha reformer. LPG, liquefied petroleum gas.

formation of isobutane is a necessary step to produce alkylate gasoline or methyl tertiary butyl ether (MTBE). The extent of paraffin isomerization is limited by a temperature-dependent thermodynamic equilibrium. For these reactions, a more active catalyst permits a lower reaction temperature and that leads to higher equilibrium levels. Isomerization of paraffins takes place under medium pressure (typically 30 bar) in a hydrogen atmosphere.

C_4 isomerization produces isobutane feedstock for alkylation. Platinum or another metal catalyst, alumina chloride, is used for the higher temperature processes. In a typical low-temperature process where only aluminum chloride is used, the feed to the isomerization unit is *n*-butane or mixed butanes combined with hydrogen (to inhibit olefin formation). C_5/C_6 isomerization increases the octane number of the light gasoline components *n*-pentane and *n*-hexane, which are found in abundance in straight-run gasoline. The basic C_5/C_6 isomerization process is essentially the same as butane isomerization.

3.3 Alkylation

Alkylation is the process that produces gasoline-range compounds from the combination of light C_3–C_5 olefins (mainly a mixture of propylene and butylene) with isobutene. The highly exothermic reaction is carried out in the presence of a strong acid catalyst, either sulfuric acid (H_2SO_4) or hydrofluoric acid (HF). The world alkylation capacity stands at 1.9 million barrels/day and new grassroots units have been constructed in many refineries worldwide, especially those with FCC units. The alkylate product is composed of a mixture of high-octane, branched-chain paraffinic hydrocarbons. Alkylate is a premium clean gasoline blend, with octane number depending on the type of feedstocks and operating conditions. Research efforts are directed toward the development of environmentally acceptable solid superacids capable of replacing HF and H_2SO_4. Much of the work is concentrated on sulfonated zirconia catalysts.

3.3.1 Sulfuric Acid Process

In H_2SO_4-based alkylation units, the feedstock (propylene, butylene, amylene, and fresh isobutane) enters the reactor and contacts the concentrated sulfuric acid catalyst (in concentrations of 85–95%). The reactor effluent is separated into hydrocarbon and acid phases in a settler and the acid is returned to the reactor. The hydrocarbon phase is hot water washed with a caustic compound for pH control before being successively depropanized, deisobutanized, and debutanized. The alkylate obtained from the deisobutanizer can then go directly to motor gasoline blending.

3.3.2 Hydrofluoric Acid Process

In a typical HF process, olefin and isobutane feedstock are dried and fed to a combination reactor/settler system. The process is operated at temperatures attainable by cooling water and at higher pressures to keep fluid in the liquid state. The reactor effluent flows to a separating vessel, where the acid separates from the hydrocarbons. The acid layer at the bottom of the separating vessel is recycled. Propane with a trace amount of HF goes to an HF stripper for HF removal and is then defluorinated, treated, and sent to storage. Isobutane is recycled to the reactor/settler and the alkylate is sent to gasoline blending.

3.4 Etherification

Etherification results from the selective reaction of methanol or ethanol to isobutene. The ether products, such as MTBE or other oxygenates, are used as components in gasoline because of their high octane blending value. The refinery capacity of oxygenate units is approximately 266,000 barrels/day, with almost all units associated with alkylation processes. The exothermic reaction is conducted in liquid phase at 85–90°C over a highly acidic ion-exchange polystyrene resin catalyst. The reaction is very rapid and equilibrium is limited under typical reaction conditions. A combination of catalytic distillation is applied to remove the product as vapor, thereby driving the reaction to almost 100% conversion. The etherification process is needed to supply oxygenates to meet the specifications of reformulated gasoline (minimum 2.7 wt% oxygenate content). In general, MTBE is the preferred oxygenate because of its low production cost and convenient preparation route relative to those of other ethers.

3.5 Polymerization and Dimerization

Catalytic polymerization and dimerization in petroleum refining refer to the conversion of FCC light olefins, such as ethylene, propylene, and butenes, into higher octane hydrocarbons for gasoline blending. Polymerization combines two or more identical olefin molecules to form a single molecule, with the same elements being present in the same proportions as in the original molecules. Light olefin feedstock is pretreated to remove sulfur and other undesirable

compounds. In the catalytic process, the feedstock either is passed over a solid phosphoric acid catalyst on silica or comes into contact with liquid phosphoric acid, where an exothermic polymeric reaction occurs. Another process uses a homogenous catalyst system of aluminum-alkyl and a nickel coordination complex. The hydrocarbon phase is separated, stabilized, and fractionated into LPG and oligomers or dimers.

4. HEAVY DISTILLATE PROCESSING

4.1 Catalytic Hydrotreating

Catalytic hydrotreating is a hydrogenation process used to remove approximately 90% of contaminants, such as nitrogen, sulfur, oxygen, and metals, from liquid petroleum fractions. These contaminants, if not removed from the petroleum fractions as they travel through the refinery processing units, can have detrimental effects on the equipment, the catalysts, and the quality of the finished product. In addition, hydrotreating converts olefins and aromatics to saturated compounds. World capacity of all types of hydrotreating stands at approximately 38.3 million barrels/day. Hydrotreating is used to pretreat catalytic reformer feeds, saturate aromatics in naphtha, desulfurize kerosene/jet and diesel, saturate aromatics, and pretreat catalytic cracker feeds. It also includes heavy gas oil and residue hydrotreating as well as posthydrotreating of FCC naphtha.

Hydrotreating for sulfur or nitrogen removal is called hydrodesulfurization or hydrodenitrogenation. Hydrotreating processes differ depending on the feedstock available and the catalysts used. Mild hydrotreating is used to remove sulfur and saturate olefins. More severe hydrotreating removes nitrogen and additional sulfur and saturates aromatics. In a typical catalytic hydrotreater unit, the feedstock is mixed with hydrogen, preheated in a fired heater (315–425°C), and then charged under pressure (up to 68 atm) through a fixed-bed catalytic reactor. In the reactor, the sulfur and nitrogen compounds in the feedstock are converted into H_2S and NH_3. Hydrodesulfurized products are blended or used as catalytic reforming feedstock. Hydrotreating catalysts contain cobalt or molybdenum oxides supported on alumina and less often nickel and tungsten.

4.2 Fluid Catalytic Cracking

Catalytic cracking is the largest refining process for gasoline production, with a global capacity of more than 14.2 million barrels/day. The process converts heavy feedstocks such as vacuum distillates, residues, and deasphalted oil into lighter products that are rich in olefins and aromatics. There are several commercial FCC processes that are employed in world refineries with major differences in the method of catalyst handling. FCC catalysts are typically solid acids of fine particles, especially zeolites (synthetic Y-faujasite), aluminum silicate, treated clay (kaolin), bauxite, and silica-alumina. Zeolite content in commercial FCC catalysts is generally in the range of 5–20 wt%, whereas the balance is a silica-alumina amorphous matrix. Additives to the FCC process make up no more than 5% of the catalyst and they are basically used as octane enhancers, metal passivators, and SO_x reducing agents and are used in CO oxidation and for gasoline sulfur reduction.

The FCC unit comprises a reaction section, product fractionation, and a regeneration section. In principle, the reactor (riser) and the regenerator form the catalyst circulation unit in which the fluidized catalyst is continuously circulated using air, oil vapors, and steam as the conveying media. Figure 3 presents a schematic of a typical FCC process. The operating temperatures of the FCC unit range from 500 to 550°C at low pressures. Hydrocarbon feed temperatures range from 260 to 425°C, whereas regenerator exit temperatures for hot catalyst range from 650 to 815°C. Several operating parameters, mainly temperature, affect overall conversion and it is essential to determine which product slate is needed so that process conditions are appropriately selected.

4.2.1 Reaction Section (Riser)

A typical FCC unit involves mixing a preheated hydrocarbon charge with hot, regenerated catalyst as it enters the riser leading to the reactor. Major process variables are temperature, pressure, catalyst/oil ratio, and space velocity. Hydrocarbon feed is combined with a recycle stream within the riser, vaporized, and raised to reactor temperature by the hot regenerated catalyst. As the mixture moves up the riser, the charge is cracked at approximately 110 kPa and the residence time is on the order of 1 s. In modern FCC units, almost all cracking occurs within the riser. The cracking continues until the oil vapors are separated from the catalyst in the reactor cyclones.

Cracking reactions are endothermic; the energy balance is obtained by the burning of catalyst-deposited coke in the regenerator. Both primary and secondary reactions take place during catalytic

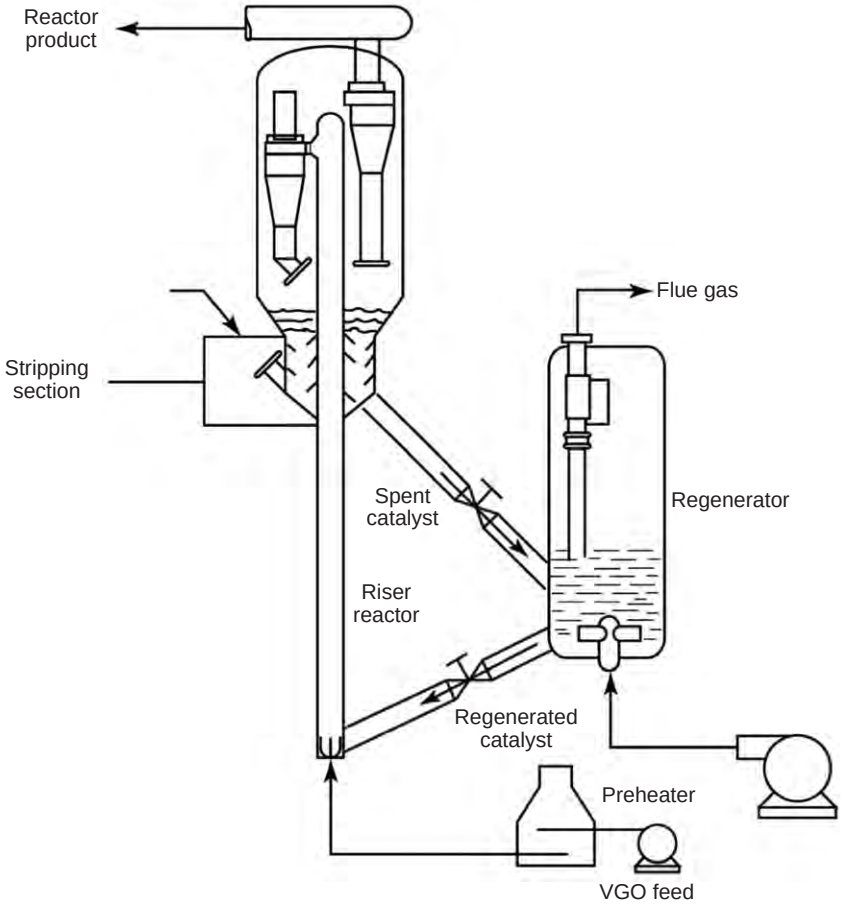


FIGURE 3 Schematic diagram of fluid catalytic cracking process. VGO feed, vacuum gas oil feed.

cracking. Primary reactions are the result of the cracking of paraffins, alkyl naphthenes, and alkyl aromatics. In general, all cracking reactions are characterized by the production of appreciable amounts of corresponding olefins. During the reactions, however, approximately 40% of the sulfur in the FCC feed is converted to H_2S, which is easily removed. Much of the ongoing research is directed to the removal of the remaining sulfur in FCC gasoline.

4.2.2 Product Fractionation

Cracked hydrocarbons are separated into various products. The resultant product stream from the reaction section is charged to a fractionating column, where it is separated into fractions, and some of the heavy oil is recycled to the riser. The main FCC products are LPG, the gasoline fraction, and light cycle oil. By-products include refinery gases, residue (slurry), and coke. Since the FCC unit is the major source of olefins in the refinery (for the downstream alkylation unit or petrochemical feedstock), an unsaturated gas plant is generally considered a part of it.

4.2.3 Regeneration Section

Spent FCC catalyst is regenerated by burning off deposited coke to carbon dioxide The catalyst flows through the catalyst stripper to the regenerator, where most of the coke deposits burn off in the presence of preheated air. The carbon content of the regenerated catalyst is generally kept at the lowest level to achieve selectivity benefits. Catalyst circulation and coke yield determine the temperature at which the regenerator is operated. Maximum regenerator temperatures are limited by mechanical specifications or sometimes by catalyst stability. The temperature in the regenerator reaches almost 650°C due to the exothermic nature of coke-burning reactions. Spent catalyst is continuously removed and fresh catalyst is added as makeup to optimize the cracking process. This added catalyst is in effect the main determinant of catalyst activity. The typical

catalyst makeup requirement is approximately 0.1 kg per barrel of total feed.

4.3 Hydrocracking

Catalytic hydrocracking of heavy petroleum cuts is an important process for the production of gasoline, jet fuel, and light gas oils. Some hydrocracking processes also allow the production of a highly purified residue, which can be an excellent base for oils. The process employs high pressure, high temperature, a catalyst, and hydrogen. In contrast to FCC, the advantage of hydrocracking is that middle distillates, jet fuels, and gas oils of very good quality are provided. In general, hydrocracking is more effective in converting gas oils to lighter products, but it is more expensive to carry out.

Hydrocracking is used for feedstocks that are difficult to process by either catalytic cracking or reforming, since these feedstocks are usually characterized by a high polycyclic aromatic content and/or high concentrations of the two principal catalyst poisons, sulfur and nitrogen compounds. These feedstocks include heavy gas oils, FCC cycle oils, deasphalted oil, and visbreaker or coke gas oil. The process depends largely on the nature of the feedstock and the relative rates of the two competing reactions, hydrogenation and cracking. Heavy aromatic feedstock is converted into lighter products under a wide range of very high pressures (70–140 atm) and fairly high temperatures (400–820°C), in the presence of hydrogen and special catalysts.

4.3.1 Hydrocracking Process

Hydrocracking is a two-stage process combining catalytic cracking and hydrogenation, wherein heavier feedstocks are cracked in the presence of hydrogen. The reaction typically involves a reactor section, gas separator, scrubber for sulfur removal, and product fractionator. The reactor section contains a multicatalyst bed that can be of the fixed-bed or ebullated-bed type and some employ on-stream catalyst addition and withdrawal to maintain catalyst activity.

4.3.2 Hydrocracking Catalysts

The catalysts used in hydrocracking are all of the bifunctional type, combining an acid function and a hydrogenating function. The acid function is carried by supports with a large surface area and having a superficial acidity, such as halogenated aluminas, zeolites, amorphous silica-aluminas, and clays. The hydrogenating function is carried either by one or more transition metals, such as iron, cobalt, nickel, ruthenium, rhodium, palladium, osmium, iridium, and platinum, or by a combination of molybdenum and tungsten. The conventional catalysts of catalytic hydrocracking are made up of weak acid supports. These systems are more particularly used to produce middle distillates of very good quality and also, if their acidity is very weak, oil bases. Amorphous silica-aluminas serve as supports with low acidity. These systems have a very good selectivity in middle distillates and the products formed are of good quality. The low-acid catalysts among these can also produce lubricant bases.

5. RESIDUAL OIL PROCESSING

5.1 Solvent Deasphalting

Solvent deasphalting (SDA) is a separation process that represents a further step in the minimization of residual fuel. Figure 4 presents a schematic diagram of a typical SDA process. The process takes advantage of the fact that maltenes are more soluble in light paraffinic solvents than asphaltenes. This solubility increases with solvent molecular weight and decreases with temperature. There are constraints with respect to how deep an SDA unit can cut into the residue or how much deasphalted oil (DAO) can be produced. These constraints are related to the DAO quality specifications required by downstream conversion units and the final high-sulfur residual fuel oil stability and quality.

SDA has the advantage of being a relatively low-cost process that has the flexibility to meet a wide range of DAO qualities. The process has very good selectivity for asphaltenes and metals rejection, some selectivity for carbon rejection, and less selectivity for sulfur and nitrogen. It is most suitable for the more paraffinic vacuum residues as opposed to the high-asphaltene-, high-metal-, high-concarbon-containing vacuum residues. The disadvantages of the process are that it performs no conversion, produces a very high viscosity by-product pitch, and where high-quality DAO is required, SDA is limited in the quality of feedstock that can be economically processed.

5.2 Visbreaking

Visbreaking is the most widespread process for noncatalytic mild conversion of residues, with a world capacity of 3.7 million barrels/day. The process is designed to reduce the viscosity of atmospheric or vacuum residues by thermal cracking. It produces 15–20% of atmospheric distillates with proportionate reduction in the production of

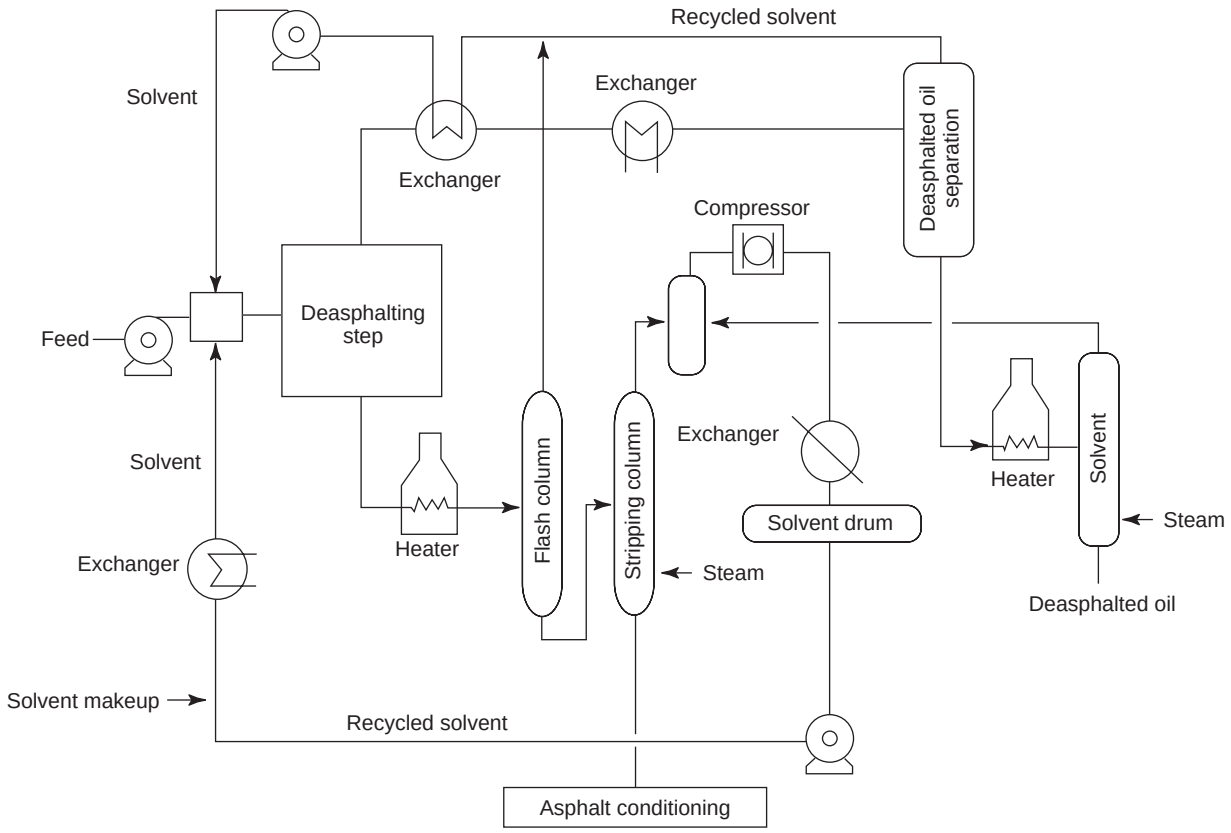


FIGURE 4 Schematic diagram of a typical solvent deasphalting process.

residual fuel oil. Visbreaking reduces the quantity of cutter stock required to meet fuel oil specifications and, depending on fuel oil sulfur specifications, typically reduces the overall quantity of fuel oil produced by 20%. In general, visbreakers are typically used to produce vacuum residues.

The process is available in two schemes: coil cracker and soaker cracker. The coil cracker operates at high temperatures during a short residence time of approximately 1 min. The soaker scheme uses a soaking drum at 30–40°C at approximately 10–20 min residence time. The residue is rapidly heated in a furnace and then cracked for a specific residence time in a soaking zone under proper conditions of pressure and temperature. The soaking zone may be within the heater or in a separate adiabatic soaking drum. The cracked residue leaves the soaking zone after the desired degree of reaction is reached and is quenched with gas oil to stop the reaction and prevent coking.

5.3 Coking

Approximately 90% of coke production comes from delayed coking. The process is one of the preferred thermal cracking schemes for residue upgrading in many refineries, mainly in the United States. The process provides essentially complete rejection of metals and concarbon while providing partial or complete conversion to naphtha and diesel. World capacity of coking units is 4.2 million barrels/day (approximately 54% of this capacity is in U.S. refineries) and total coke production is approximately 172,000 tons/day. New cokers are designed to minimize coke and produce a heavy coker gas oil that is catalytically upgraded. The yield slate for a delayed coker can be varied to meet a refiner's objectives through the selection of operating parameters. Coke yield and the conversion of heavy coker gas oil are reduced, as the operating pressure and recycle are reduced and, to a lesser extent, as temperature is increased.

5.4 Residue Hydrotreating and Residue FCC

Refineries that have a substantial capacity for visbreaking, solvent deasphalting, or coking are faced with large quantities of visbreaker tar, asphalt or

coke, respectively. These residues have high viscosity and high organic sulfur content (4–6 wt%), with primary consequences reflected in the potential for sulfur emissions and the design requirements for a sulfur removal system. Sulfur content is also important from the standpoint of corrosion, which requires proper selection of design materials and operating conditions. Other properties of residues include their high heating value due to the high level of fixed carbon that results in a higher yield of syngas per ton of residue processed. Moreover, residues have low volatile matter and ash content as well as little to no oxygen content, resulting in low reactivity.

Residue hydrotreating is another method for reducing high-sulfur residual fuel oil yields. Atmospheric and vacuum residue desulfurization units are commonly operated to desulfurize the residue as a preparatory measure for feeding low-sulfur vacuum gas-oil feed to cracking units (FCC and hydrocrackers), low-sulfur residue feed to delayed coker units, and low-sulfur fuel oil to power stations. Two different types of processing units are used for the direct hydrotreating of residues. These units are either a down-flow, trickle phase reactor system (fixed catalyst bed) or a liquid recycle and back-mixing system (ebullating bed). Economics generally tend to limit residue hydrotreating applications to feedstocks containing less than 250 ppm nickel and vanadium.

Residue FCC (RFCC) is a well-established approach for converting a significant portion of the heavier fractions of the crude barrel into a high-octane gasoline-blending component. In addition to high gasoline yields, the RFCC unit produces gaseous, distillate, and fuel oil-range products. The RFCC unit's product quality is directly affected by its feedstock quality. In particular, unlike hydrotreating, RFCC redistributes sulfur, but does not remove it from the products. Consequently, tightening product specifications have forced refiners to hydrotreat some, or all, of the RFCC's products. Similarly, in the future, the SO_x emissions from an RFCC may become more of an obstacle for residue conversion projects. For these reasons, a point can be reached where the RFCC's profitability can economically justify hydrotreating the RFCC's feedstock.

6. TREATING PROCESSES

6.1 Hydrogen Production

Refineries are experiencing a substantial increase in hydrogen requirements to improve product quality and process heavy sour crudes. Hydroprocessing and saturation of aromatics and olefins will accelerate the demand for hydrogen within the refinery. Catalytic naphtha reforming alone is not able to meet refinery hydrogen requirements. A survey on world refining indicated that the capacity of supplementary refinery hydrogen, produced mainly by steam reforming of methane, reached 337 million m^3/day (11,880 million ft^3/day-MMcfd) in 2002 compared to 110 million m^3/day in 1990. There is a growing recognition that there will be a significant future shortage of refinery hydrogen supply. Specific hydrogen production units, such as steam methane reformers or those carrying out partial oxidation of heavy residues, will have to be built.

6.2 Residue Gasification

The gasification of refinery residues into clean syngas provides an alternative route for the production of hydrogen and the generation of electricity in a combined turbine and steam cycle. Compared to steam-methane reforming, gasification of residues can be a viable process for refinery hydrogen production when natural gas price is in the range of $3.75–4.00/million British thermal units (MMBtu). The largest application of syngas production is in the generation of electricity power by the integrated gasification combined cycle (IGCC) process. Consumption of electricity in the modern conversion refinery is increasing and the need for additional power capacity is quite common, as is the need to replace old capacity. The design of a residue gasification plant requires a careful matching and integration of the various process steps to ensure optimum performance of the whole system. In general, the IGCC plant consists of several steps: gasification, gas desulfurization, and a combined cycle. The technologies of the gasification and the combined cycle are well known; the innovation, however, is their integration in order to maximize the overall IGCC efficiency.

6.3 Aromatics Extraction

BTX aromatics are high-value petrochemical feedstocks produced by catalytic naphtha reforming and extracted from the reformate stream. Whether or not other aromatics are recovered, it is sometimes necessary to remove benzene from the reformate in order to meet mandated specifications on gasoline composition. Aromatics production in refineries reached 1.2 million barrels/day in 2002. Most new aromatic complexes are configured to maximize the yield of benzene and paraxylene and sometimes orthoxylene. The solvents used in the extraction of

aromatics include dimethylfrmamide, formylmorpholine, dimethylsulfoxide, sulfolane, and ethylene glycols.

6.4 Sweetening

Sweetening is the removal of contaminants such as organic compounds containing sulfur, nitrogen, and oxygen; dissolved metals and inorganic salts; and soluble salts dissolved in emulsified water from petroleum fractions or streams. A variety of intermediate and finished products, including middle distillates, gasoline, kerosene, jet fuel, and sour gases, are sweetened. Treating can be accomplished at an intermediate stage in the refining process or just before the finished product is sent to storage. Choices of a treating method depend on the nature of the petroleum fractions, the amount and type of impurities in the fractions to be treated, the extent to which the process removes the impurities, and end-product specifications. Treatment materials include acids, solvents, alkalis, oxidizing agents, and adsorption agents.

6.5 Sulfur Recovery

Sulfur recovery converts hydrogen sulfide in sour gases and hydrocarbon streams to elemental sulfur. Total sulfur production in world refineries reached approximately 64,000 tons/day in 2002 compared to approximately 28,000 tons/day in 1996, corresponding to a yearly growing recovery rate of 20%. In other words, in 2002 an average refinery recovered 0.8 kg sulfur from one processed barrel of crude oil compared to less than 0.4 kg sulfur recovered in 1996. This indicates the increasing severity of operations to meet stringent environmental requirements. The most widely used sulfur recovery system is the Claus process, which uses both thermal and catalytic conversion reactions. A typical process produces elemental sulfur by burning hydrogen sulfide under controlled conditions. Knockout pots are used to remove water and hydrocarbons from feed gas streams. The gases are then exposed to a catalyst to recover additional sulfur. Sulfur vapor from burning and conversion is condensed and recovered.

6.6 Acid Gas Removal

Amine plants remove acid contaminants from sour gas and hydrocarbon streams. In amine plants, gas and liquid hydrocarbon streams containing carbon dioxide and/or hydrogen sulfide are charged to a gas absorption tower or liquid contactor, where the acid contaminants are absorbed by counterflow amine solutions [i.e., monoethanol amine (MEA), diethanol amine (DEA), methyl diethanol amine (MDEA)]. The stripped gas or liquid is removed overhead and the amine is sent to a regenerator. In the regenerator, the acidic components are stripped by heat and reboiling and are disposed of, and the amine is recycled.

7. ENVIRONMENTAL AND FUTURE ISSUES

7.1 Environmental Issues

Refiners are faced with many environmental, economic, and operational issues. Environmental legislation is a growing concern, driving changes in product specifications, product markets, and refinery operating practices. Strict product quality specifications and severe emission and discharge limits have economic impact on the average refiner. In the near future, the following environmental trends will continue to grow, but they will not create significant changes in oil consumption patterns: (1) the production of clean transportation fuels according to new specifications and (2) refinery operation within strict emissions regulations.

The configuration of many refineries has changed substantially, mainly due to the declining quality of crude oil supply and environmental regulations. In retrospect, refinery changes brought about by the variations in crude supply and composition were evolutionary, whereas environmental regulations were revolutionary.

7.1.1 Clean Transportation Fuels

Since 1990, government agencies have imposed strict environmental restrictions on transportation fuels to improve product quality specifications. Fuel reformulation is being discussed all over the world. Automotive manufacturers are demanding lower gasoline sulfur levels and lower driveability indices. Refiners must improve air quality by delivering clean products that minimize emissions of toxic and hazardous hydrocarbons. Gasoline and diesel formulations have been already changed in many countries and will change even more in the coming years.

Refining is faced with huge investments to meet new stringent specifications for sulfur, aromatics, and olefin content. Gasoline sulfur reduction is centered around the FCC unit employing feed pretreatment or gasoline posttreatment. The reduced demand for ethers, such as MTBE in gasoline for oxygenate content, necessitates the utilization of branched paraffin isomer products of alkylation and isomerization. For diesel fuel, this means a sulfur content

less than 30 or even 15 ppm, an increase of the cetane number, a reduction in polyaromatic content, and T_{95} point limitation. To fulfill all these legislative and regional requirements, refiners must either revamp existing units or invest in new hydroprocessing and hydrogen production units.

7.1.2 Refinery Emissions

Refiners must comply with various environmental regulations to reduce all types of pollutants in their waste gas as well as wastewater systems. Most concerns involve the emissions of SO_x, NO_x, CO, hydrocarbons, and particulates. The oxides are present in flue gases from furnaces, boilers, and FCC regenerators. Tail gas treatment and selective catalytic reduction units are being added to limit SO_2 and NO_x emissions. Water pollutants include oil, phenol, sulfur, ammonia, chlorides, and heavy metals. New biological processes can be used to convert H_2S or SO_x from gaseous and aqueous streams. Spent catalysts and sludges are also of concern to the refinery in reducing pollution. Some spent FCC catalysts can be used in cement but other catalysts that contain heavy metals need special treatment before proper disposal.

7.2 Future Refining Issues

World refining has been adapting to ongoing product changes and environmental challenges. Transportation fuels with approximately free sulfur will be needed to satisfy the demand of the automotive industry to reduce emissions from internal combustion engines. There will be an increased demand for alkylate and isomerate gasoline as well as deep-desulfurized diesel. This will increase the hydrogen content in gasoline, enhance combustion, and reduce the levels of carbon dioxide emissions.

The introduction of fuel cells as a feasible way to fuel zero-emission vehicles is a major challenge to oil companies and refiners. Virtually every major automotive manufacturer has a fuel-cell program and most claim production readiness by 2005. Refiners need to adapt to this technology in the future, especially regarding new fuels needed for fuel cells. Fuel-cell vehicles need hydrogen generated on-board or carried in either compressed or liquid form. The latter calls for a global hydrogen infrastructure. The use of hydrocarbons and specifically gasoline to generate hydrogen offers many economic advantages such as the availability of a ready-made global fueling infrastructure.

The huge technological challenges associated with the transfer to a hydrogen economy necessitate an efficient and better use of hydrocarbon resources to compete with renewable energy sources. Refiners need to enhance and integrate their business with chemical production and power generation. In the long run, the refinery will produce not just fuels, but also chemicals and electricity.

SEE ALSO THE FOLLOWING ARTICLES

Coal Preparation • Oil and Natural Gas Drilling • Oil and Natural Gas Exploration • Oil and Natural Gas: Offshore Operations • Oil Pipelines • Oil Recovery • Petroleum System: Nature's Distribution System for Oil and Gas

Further Reading

Aitani, A. (1995). Reforming processes. *In* "Catalytic Naphtha Reforming" (G. Antos *et al.*, Eds.), pp. 409–436. Dekker, New York.

Farrauto, R., and Bartholomew, C. (1997). "Fundamentals of Industrial Catalytic Processes," pp. 519–579. Blackie Academic and Professional, London.

Gary, J., and Handwerk, G. (2001). "Petroleum Refining Technology and Economics." 4th ed. Dekker, New York.

Heinrich, G. (1995). Introduction to refining. *In* "Petroleum Refining" (J. P. Wauquier, Ed.), pp. 365–413. Editions Technip, Paris.

Hoffman, H. (1992). Petroleum and its products. *In* "Riegel's Handbook of Industrial Chemistry" (J. Kent, Ed.), 9th ed., pp. 490–496. Van Nostrand Reinhold, New York.

Le Page, J. P., Chatila, S., and Davidson, M. (1992). "Resid and Heavy Oil Processing." Editions Technip, Paris.

Maples, R. (2000). "Petroleum Refinery Process Economics." 2nd ed. PennWell Books, Tulsa, OK.

Martino, G., and Wechem, H. (2002). "Current Status and Future Developments in Catalytic Technologies Related to Refining and Petrochemistry," Review and Forecast Paper, 17th World Petroleum Congress, Rio de Janeiro, Brazil, September 2002.

Meyers, R. (1997). "Handbook of Petroleum Refining Processes." 2nd ed. McGraw-Hill, New York.

Penning, T. (2001). "Petroleum Refining: A Look at the Future, Hydrocarbon Processing," February, pp. 45–46.

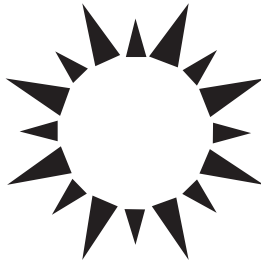
Silvy, R. (2002). Global refining catalyst industry will achieve strong recovery by 2005, Oil & Gas Journal, pp. 48–56. September 2, 2002.

Speight, J., and Ozum, B. (2002). "Petroleum Refining Processes." Dekker, New York.

Stell, J. (2002). Worldwide refining survey. Oil & Gas Journal, December 23, 2002, pp. 68–70.

Oil Sands and Heavy Oil

WORLD ENERGY COUNCIL
London, United Kingdom



1. History
2. The Resource Base
3. Extraction and Conversion Technology
4. Economic Issues
5. Environmental Issues
6. Country Profiles

Glossary

estimated additional reserves The amount, expressed as tonnage of recoverable synthetic crude oil or raw bitumen (additional to the proved recoverable reserves), that is of foreseeable economic interest. Speculative amounts are not included.

natural bitumen Comprises bitumen or other petroleum with very high viscosity (contained in bituminous sands, oil sands, or tar sands), which is not recoverable by conventional means; the petroleum is obtained either as raw bitumen (through *in situ* recovery) or as a synthetic crude oil (via an integrated surface mining plus upgrading process).

proved amount in place The tonnage of natural bitumen that has been carefully measured and assessed as exploitable under present and expected local economic conditions with existing available technology.

proved recoverable reserves The tonnage of synthetic crude oil or raw bitumen that has been carefully measured and assessed as recoverable under present and expected local economic conditions with existing available technology.

Oil sands are deposits of bitumen, viscous oil that must be treated in order to convert it into an upgraded crude oil before it can be used in refineries to produce gasoline and other fuels. Natural bitumen and heavy oil are closely related types of petroleum, differing from each other, and from the petroleum from which they are derived, only to the degree by which they have been degraded. This alteration, through bacterial attack, water washing, and inspissations, has resulted in severe loss of the light ends of the petroleum, notably the paraffins, and subsequent relative enrichment of the heavy molecules, leading to increased density and viscosity. Of these molecules, the asphaltenes are very large and incorporate nonhydrocarbons such as nitrogen, sulfur, oxygen, and metals, in particular nickel and vanadium. The result of this chemistry is an array of problems beyond those encountered with conventional petroleum with respect to exploitation, transportation, storage, and refining. This, of course, is reflected in the increased cost of extraction and processing and the physical limitations on production capacity.

1. HISTORY

The Bible has numerous references to the use of pitch, which is asphalt or bitumen in its soft state. It also is called "slime" in some translations of the Bible. Noah waterproofs the Ark with pitch: "And God said to Noah, '… Make yourself an ark of Gopher Wood; make rooms in the ark, and cover it inside and out with pitch.'" (Genesis 6:13–18 RSV). The reed basket that carries the infant Moses into the Nile River is waterproofed with pitch: "… And when she could hide him no longer she took for him a basket made of bulrushes, and daubed it with bitumen and pitch; and she put the child in it and placed it among the reeds at the river's brink." (Exodus 2:1–4 RSV)

Indeed, throughout ancient times, the Elamites, Chaldeans, Akkadians, and Sumerians mined bitumen. Mesopotamian bitumen was exported to Egypt, where it was used for various purposes, including the preservation of mummies. The Dead Sea was known as Lake Asphaltites (from which the term asphalt was derived) because of the lumps of semisolid petroleum that were washed up on its shores from underwater seeps. Bitumen was used as mortar for bricks, as caulking for ships, in road building, for

 © 2004 Elsevier Inc. All rights reserved.

bonding tools, in inlaid work and jewel setting, and for waterproofing baskets and mats. Artistic and religious objects were carved from bitumen-impregnated sands.

The fur trader Peter Pond is the first European to describe the vast oil sands (tar sands) of Athabasca, Canada. In 1788, he wrote: "At about 24 miles from the fork (of the Athabasca and Clearwater Rivers) are some bituminous fountains into which a pole of 20 feet long may be inserted without the least resistance. The bitumen is in a fluid state and when mixed with gum, the resinous substance collected from the spruce fir, it serves to gum the Indians' canoes. In its heated state it emits a smell like that of sea coal." In 1913, Sidney Ells, an engineer for the Canadian Federal Department of Mines, advocates the hot water flotation method of separating bitumen from the tar sands of Athabasca. He is the first to bring out samples for laboratory testing as a road paving material. Oil sands cannot compete economically with imported asphalt and the project is dropped. His is one of several unsuccessful attempts to develop the Athabasca sands. Decades later, oil sands will re-emerge as a viable energy source. In 1967, the first barrel was shipped from the Great Canadian Oil Sands Project, the world's first commercially successful operation to tap the rich Athabasca oil sands in Alberta, Canada. The project, owned by the Sun Oil Company (later Suncor Energy), ushered in an era of rapid development of the oil sand resource base.

2. THE RESOURCE BASE

Table I lists estimates of the magnitude of oil sand and heavy oil resource base. It is estimated that there are 1.7 trillion barrels of oil in the oil sands of Canada and that 15% (255 billion barrels) of the total oil in place is recoverable. Canada accounts for approximately 75% of the world's oil sand resources. Other countries and regions that have significant, but more modest, resources include the United States, China, Eastern Europe and the former Soviet Union, the Caribbean Basin, and Pakistan (see Fig. 1). Approximately 700,000 barrels per day of Canadian oil sands are being produced. This supply is divided into two categories: "oil sands *in situ*" (often referred to as bitumen) and "oil sands mining." These two categories reflect the method of recovery. The bitumen is extracted by injecting very hot steam into the rock formation to heat the oil, lower its viscosity, and allow it to flow more like conventional oil. Slightly more than one-half (approximately 400,000 barrels per day) of Canadian oil sands production is derived from the more expensive oil sands mining method.

TABLE I

Natural Bitumen: Resources, Reserves, and Production in 1999

		Millions of tons			
	Recovery method	Proved amount in place	Proved recoverable reserves	Estimated additional reserves	Production in 1999
North America					
Canada		45,300	979		30.0
	Surface	9000	747		15.9
	In situ	36,300	232		14.1
United States of America	*In situ*			4231	
South America					
Venezuela	*In situ*	3880	373	118	3.3
Europe					
Romania	*In situ*	4	1		
Middle East					
Jordan	Surface	40	5		

Note. The data shown above are those reported by World Energy Council Member Committees in 2000/2001. They thus constitute a sample, reflecting the information available in particular countries; they should not be considered as complete, or necessarily representative of the situation in each region. For this reason, regional and global aggregates have not been computed.
World Energy Council Member Committees have been unable to provide sufficient numerical data for extra-heavy oil to be included in this table.

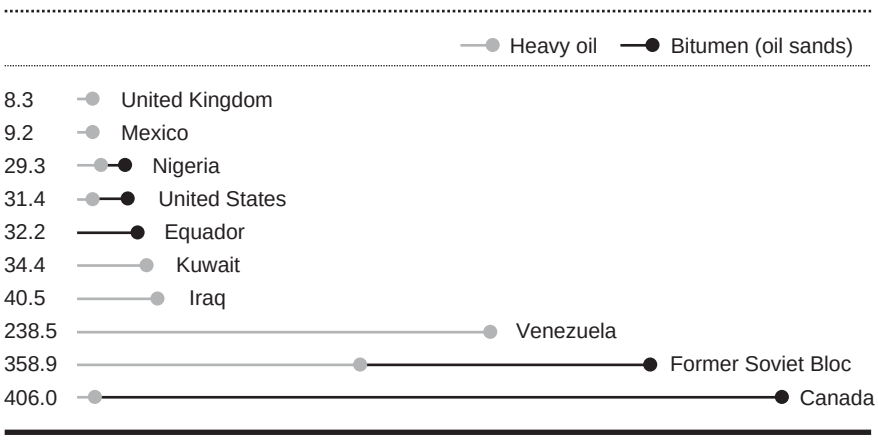


FIGURE 1 World oil sands (bitumen) and heavy oil resources, in billions of cubic meters. Reprinted from Petroleum Communication Foundation, Alberta, Canada, with permission.

Those deposits that are close enough to the surface are actually mined.

The heavy crude oil deposit of the Orinoco Oil Belt, a part of the Eastern Venezuela basin, represents nearly 90% of the known heavy oil in place. Between them, these two deposits, each located up-dip against a continental craton, represent approximately 3600 billion barrels of oil in place. This is only the remaining, degraded remnant of petroleum deposits once totaling as much as 18,000 billion barrels.

Heavy oil is recorded in 219 separate deposits; some of these are different reservoirs in a single field, some are producing, and some are abandoned. The deposits are found in 30 countries and in 54 different geological basins, with 11 of the deposits located offshore and 5 partially offshore. The following data are average values. Most of the reservoirs are sandstone at depths of 5400 ft, thicknesses of 126 ft, porosities of 21%, and permeabilities of 1255–6160 millidarcies. The American Petroleum Institute (API) gravity is 8° and viscosity is 22,700 cP. The viscosity varies from 6000 at 70°F, to 4600 at 100°F, to 1400 at 130°F, and the gas/oil ratio is only 431. The chemical data for the whole crude demonstrate the processing difficulties with heavy crude. The Conradson Carbon is 11.5 wt%, asphaltenes are 16 wt%, sulfur is 4.69 wt%, nickel is 260 ppm, and vanadium is 972 wt%. Especially significant are the residuum yield of 62 vol% and its specific gravity of 1.06 and Conradson Carbon of 17.8 wt%. These data suggest for extra-heavy crude a content of 56 wt% asphalt and 23 wt% coke.

Oil resource data are very incomplete but those that are available for heavy crude, especially data from Canada, the United States, and Venezuela, are as follows, all in millions of barrels: original oil in place, 2,133,912; cumulative production, 17,214; annual production, 432; reserves, 45,575; and probable reserves, 193,203.

3. EXTRACTION AND CONVERSION TECHNOLOGY

Oil sands recovery processes include extraction and separation systems to remove the bitumen from the sand and water. There are two basic methods of extraction. To depths of approximately 150 ft, the bitumen and rock may be surface-mined, with the bitumen subsequently separated from the rock by a hot water process. The mining operation begins with clearing trees and brush from the site and removing the overburden that sits above the oil sands deposit. This includes the topsoil, sand, clay, gravel, and muskeg (a swamp or bog in which there are many layers of decaying matter, often covered with sphagnum moss). The topsoil and muskeg are stockpiled so that they can be replaced as sections of the mined-out area are reclaimed. The rest of the overburden is used to reconstruct the landscape.

At the processing plant, the mixture of oil, sand, and water goes into a large separation vessel. A thick froth is skimmed off, mixed with a solvent, and spun in a centrifuge to remove water and dissolved salts from the bitumen. The solvent is recycled. The sand and water are waste products, known as tailings, that fall to the bottom of the separation vessel.

The sand is returned to the mine site to fill in mined-out areas. Water from the extraction process, containing sand, fine clay particles, and traces of

bitumen, goes into settling ponds. Some bitumen may be skimmed off the ponds if it floats to the surface. The sand sinks to the bottom and bacteria digest the remaining bitumen, but the fine clay particles stay suspended for some time before slowly settling. Water is recycled back to the extraction plant for use in the separation process.

For deeper deposits, *in situ* extraction is required. In cyclic steam stimulation, high-pressure steam is injected into the oil sands formation for several weeks. The heat softens the bitumen and the water vapor helps to dilute and separate the bitumen from the sand grains. The pressure also creates channels and cracks through which the bitumen can flow to the well. When a portion of the reservoir is thoroughly saturated, the steam is turned off and the reservoir "soaks" for several weeks. This is followed by the production phase, when the bitumen flows, or is pumped up, to the same wells to the surface. When production rates decline, another cycle of steam injection begins. This process is sometimes called "huff and puff" recovery.

A second *in situ* method is steam-assisted gravity drainage (SAGD), in which pairs of horizontal wells are drilled near the base of the bitumen deposit. Steam is injected into the injector well that is placed approximately 5 m above the producer well. The steam rises and heats the bitumen, which flows down under the force of gravity to the lower producer well from which it is pumped to the surface. The bitumen is either upgraded at the site, or at a regional upgrader, or mixed with diluent and shipped to a refinery. Several pilot projects have tested the SAGD process and several commercial-scale projects are in the construction, engineering design, and regulatory approval stages.

4. ECONOMIC ISSUES

How much does recovery from oil sands cost? Supply costs are expressed as "full cycle" costs. They include all costs associated with exploration, development, and production; capital costs; operating costs; taxes and royalties; and a 10% real rate of return to the producer. Capital costs average $5 to $9 per barrel and operating costs average $8 to $12 per barrel. Such costs are presented as a range, reflecting the variance in reservoir quality, depth, project size, and operating parameters. The remainder of the supply cost is dominated by the cleaning and upgrading methods that are required to turn a very low quality hydrocarbon into more conventional oil that can be accepted by a refinery. Such methods include the removal of sulfur, heavy metals, and noncombustible materials, as well as conversion to a more hydrogenated and lighter hydrocarbon. These costs are typically in the $3 to $5 per barrel range. None of the aforementioned costs include transportation to market. Suncor Energy opened the upgrading units of its "Project Millennium" in Alberta with production costs of approximately $9 per barrel. The company's near-term goal is to lower production costs to $5.50 per barrel, which would make Suncor the lowest cost oil producer in North America.

5. ENVIRONMENTAL ISSUES

The extraction and processing of oil sands require huge volumes of water, creating the potential for various forms of water pollution, including possible contamination of ground and surface water with dissolved solids and toxic metal compounds produced from processed materials. Processing operations also release various air pollutants, such as sulfur oxides, hydrogen sulfide, nitrogen oxides, particulate matter, and carbon dioxide. Mining and processing disturbs large amounts of land. Approximately 2 tons of oil sand must be extracted and processed to make one 159-liter barrel of crude oil.

6. COUNTRY PROFILES

6.1 Canada

The National Energy Board (NEB) distinguishes between two types of nonconventional oil obtained from deposits of oil sands, defining them as follows:

Bitumen (also known as crude bitumen) is "a highly viscous mixture, mainly of hydrocarbons heavier than pentanes. In its natural state, it is not usually recoverable at a commercial rate through a well."

Upgraded crude oil (also known as synthetic crude) is "a mixture of hydrocarbons similar to light crude oil derived by upgrading oil sands bitumen."

Canada's "discovered recoverable resources" of oil sands bitumen are quoted by the NEB as 49 billion m^3 (over 300 billion barrels), of which approximately 475 million m^3 had been produced by the end of 1999. Of the remainder (shown as "proved amount in place" in Table I), 9650 million m^3 (9 billion tons) consists of synthetic crude recoverable through mining projects and 38,850 million m^3

(36.3 billion tons) consists of crude bitumen recoverable through *in situ* extraction.

Within these huge resources, the "remaining established reserves" at the end of 1999 (shown as "proved recoverable reserves" in Table I) have been assessed by the Canadian Association of Petroleum Producers (CAPP) as 799.9 million m^3 (equivalent to approximately 747 million tons) of mining-integrated synthetic crude oil and 248.1 million m^3 (approximately 232 million tons) of *in situ* bitumen.

The major deposits are in four geographic and geologic regions of Alberta: Athabasca, Wabasca, Cold Lake, and Peace River.

Although the existence of oil sands deposits was noted in the 18th century, it was not until 1875 that a complete survey was undertaken and it was only in the 20th century that exploitation was embarked on. The deposits range from being several hundred meters below ground to surface outcroppings. The extraction of bitumen from the oil sands was initially based on surface mining but *in situ* techniques became necessary in order to reach the deeper deposits.

There was much experimentation with oil sands technology in the first half of the 20th century but it was not until the effects of the economic climate of the 1950s and early 1960s began to be felt that commercial development became viable. The Government of Alberta's oil sands development policy was announced in 1962 and the Great Canadian Oil Sands Project (GCOS) was conceived and approved. The ownership of GCOS passed to Sun Oil Company and in 1967 the world's first integrated oil sands production and upgrading plant was started up by Suncor (formerly Sun Oil).

Suncor's area of operation, 40 km north of Fort McMurray, is within the Athabasca deposits. The processing capability of the original Oil Sands Plant has steadily increased and the expansion of the Steepbank Mine (on the opposite side of the Athabasca River) resulted in record production of 105,600 barrels/day in 1999. At the beginning of 1999, the company announced its Project Millennium, a phased series of expansions to the Steepbank mine, adding bitumen extraction plants and increasing upgrader capacity. The first phase was expected to increase production to 130,000 barrels/day by 2001 and the second phase to 225,000 barrels/day in 2003. In 2000, the establishment (subject to the necessary approval) of an *in situ* project at Firebag (40 km northeast of the Oil Sands Plant) was announced. It is planned that Firebag, in conjunction with the open-pit mining operation, will result in production reaching 260,000 barrels/day in 2004. Through a combination of mining and *in situ* development, Suncor envisages an oil sands production of 400,000–450,000 barrels/day in 2008.

Syncrude, a joint venture with 10 participants (Imperial Oil, a subsidiary of Exxon, is the majority shareholder with 25% of holdings), operates the Lake Mildred plant, also 40 km north of Fort McMurray. Production began in 1978 and, using open-pit mining methods, the shallow deposits are recovered for bitumen extraction and the production of upgraded crude oil. Gross production was 223,000 barrels/day in 1999. A new project, the Aurora mine, a 35 km extension from Lake Mildred, opened in August 2000. The mine's output is partially processed on-site and then pipelined to the upgrader for further treatment. In 1999, a major expansion to Syncrude's upgrading capacity was approved by the federal government; construction began in 2001. It is planned that the work under development will result in a capacity of approximately 350,000 barrels/day in 2004.

The Cold Lake oil sands deposits area is operated by Imperial Oil. The company began commercial development in 1983 and has since gradually expanded facilities; the total production of bitumen in 1999 was 132,000 barrels/day. Imperial Oil plans to bring further expansion on-stream so that as of late 2002, bitumen production would have increased by 30,000 barrels/day.

Commercial production of Shell Canada's Peace River *in situ* deposits (northwestern Alberta) began in 1986. Bitumen production capacity is set at approximately 12,000 barrels/day although during the year 2000 the actual production from existing wells was considerably lower. In an attempt to boost declining bitumen production, Shell announced in late 2000 that it would drill 18 new wells.

Albian Sands Energy, a joint venture, has been created to build and operate the Muskeg River Mine on behalf of its owners: Shell Canada (the majority shareholder, with 60% of holdings), Chevron Canada (20% of holdings), and Western Oil Sands (20% of holdings). The mine, already under construction, is located 75 km north of Fort McMurray (Athabasca). In addition, a pipeline is to be constructed to link the mine to an upgrader to be built next to Shell's Scotford refinery. The start-up of the project was scheduled for late 2002, with anticipated production of 155,000 barrels/day of bitumen.

Taking into account all operations, total output from Canadian oil sands in 1999 was 323,000 barrels/day of synthetic crude and 244,000 barrels/day of crude bitumen from the *in situ* plants; together

these represented 22% of Canada's total production of crude oil and natural gas liquids.

6.2 Trinidad and Tobago

The famous Pitch Lake at La Brea (named after the Spanish word for tar or pitch) was reputedly discovered at the end of the 16th century. Trinidad Lake Asphalt, a semisolid emulsion of soluble bitumen, mineral matter, and other minor constituents (mainly water), was mined and used as a road surfacing material as long ago as 1815.

The lake contains 10 million tons of reserves, which at the current rate of extraction are expected to last for another 400 years. Lake Asphalt of Trinidad and Tobago (1978) Ltd. (TLA), a state-owned company, produces between 10,000 and 15,000 tons per annum and exports most of this amount, after removal of water, etc.

In combination with bitumen (asphalt) from refined crude oil, the product has featured significantly in the road construction industry over a long period of time and in many countries. In addition to mining it, TLA distributes the natural bitumen and has incorporated it into a range of paints and coatings. The company has also developed a process for making cationic bitumen emulsions. Production of these emulsified bitumen, water, and soap solutions began in late 1996 and they are used widely throughout the industrialized world in place of solvent-based bitumen emulsions.

6.3 United States of America

Distillation of tar sands, occurring as a surface outcrop in California, was carried out in the 1870s. During the following century, efforts were periodically made to establish the industry in both California and various other states, but the availability of low-priced, indigenous conventional oil meant that there was never a persistently strong incentive for the development of tar sands deposits.

The United States classifies tar sands as being measured or demonstrated ("the bitumen resource based on core and log analyses") or as being speculative or inferred ("the bitumen that is presumed to exist from reported tar shows on drillers' lithological logs and/or geological interpretations").

The tar sands resource of 58.1 billion barrels (22.4 "measured" and 35.7 "speculative") is widely distributed throughout the country, with 33% located in Utah, 17% in Alaska, and the remaining 50% in California, Alabama, Kentucky, Texas, and other states. There are eight giant (>1 billion barrels) deposits of natural asphalt *in situ*, which represent nearly 80% of the total U.S. demonstrated and inferred resources.

The geological conditions of the Utah deposits have meant that recovery is difficult and expensive. Likewise, the Texan deposits, mostly deep and relatively thin, have also proved difficult to recover. The only state where small volumes of tar sand hydrocarbons are being produced from subsurface deposits (associated with heavy oil) is California.

Gilsonite (a naturally occurring solid hydrocarbon) is being produced by three companies from a number of vertical veins in the Green River Formation and overlying Eocene Uinta Formation in Uintah County, eastern Utah. Production figures for the gilsonite district are not available, but probably total several hundred thousand tons per year. Gilsonite is used in a variety of specialty products, such as printing inks, paints, and protective coatings, drilling and foundry sand additives, and briquetting.

6.4 Venezuela

There are vast deposits of bitumen and extra-heavy oil in the Orinoco Oil Belt (OOB) in eastern Venezuela, north of the Orinoco river. The original-oil-in-place of the extra-heavy oil reservoirs of the OOB has been estimated at 1200 billion barrels, with some 270 billion barrels of oil recoverable. Venezuela's total proved reserves of crude oil (76.8 billion barrels at the end of 1999) include 35.7 billion barrels of extra-heavy crudes.

There are four joint ventures for the exploitation of extra-heavy crude. Petróleos de Venezuela (PDVSA), the state oil company, has a minority interest in all four and all are at different stages of development:

The Hamaca project (a joint venture between Phillips Petroleum, Texaco, and PDVSA) has been delayed, owing to financing problems, but is planned to produce 190,000 barrels/day.

The Sincor Project (a joint venture between TotalFinaElf, Statoil, and PDVSA) was reported to have started bitumen production in December 2000, with its upgrading plant having been scheduled to come on-stream a year later. The project is planned to produce 180,000 barrels/day.

Production from the Petrozuata project, a joint venture between Conoco and PDVSA, has begun and had reached its target of 120,000 barrels/day by February 2001. Work to enable production to increase to 150,000 barrels/day is under way. An upgrader processes the 120,000 barrels/day of 9° API

oil, turning it into 103,000 barrels/day of lighter, synthetic crude, some of which is used as refinery feedstock to obtain gasoline and diesel for the domestic and export markets. Beginning early in 2001, the remainder was shipped to the United States for processing into higher value products.

The Cerro Negro is a joint venture project between ExxonMobil, Veba, and PDVSA. Output was expected to rise from 60,000 barrels/day in 2000 to 120,000 barrels/day by March 2001, following the completion of a new coking unit.

In the early 1980s, Intevep, the research affiliate of the state oil company PDVSA, developed a method of utilizing some of the hitherto untouched potential of Venezuela's extra-heavy oil/natural bitumen resources. Natural bitumen (7.5°–8.5° API) extracted from the reservoir is emulsified with water (70% natural bitumen, 30% water, $<1\%$ surfactants), the resulting product being called Orimulsion. Orimulsion can be pumped, stored, transported, and burnt under boilers using conventional equipment with only minor modifications. Initial tests were conducted in Japan, Canada, and the United Kingdom and exports began in 1988.

Orimulsion is processed, shipped, and marketed by Bitúmenes del Orinoco S.A. (Bitor), a PDVSA subsidiary, but with the fuel's relatively high sulfur content and its emission of particulates, Intevep continues to seek improvements in its characteristics in order to match increasingly strict international environmental regulations. Bitor operates an Orimulsion plant at Morichal in Cerro Negro, with a capacity of 5.2 million tons per year. The company hopes to produce 20 million tons per year by 2006.

Following manufacture at the plant, the Orimulsion is transported by pipeline approximately 320 km to the José export terminal for shipment. During the 1990s, other markets were developed and Barbados, Brazil, Canada, China, Costa Rica, Denmark, Finland, Germany, Guatemala, Italy, Japan, Lithuania, Northern Ireland, the Philippines, South Korea, Taiwan, Thailand, and Turkey either consume or are considering consuming the product. In 1999, 4.9 million tons of Orimulsion were exported, bringing the cumulative total to greater than 27 million tons.

In addition to being used in conventional power plants using steam turbines, Orimulsion can be used in diesel engines for power generation, in cement plants, as a feedstock for Integrated Gasification Combined Cycle, and as a "reburning" fuel (a method of reducing NO_x by staging combustion in the boiler).

Acknowledgments

This article is adapted from the chapter on Natural Bitumen and Extra-Heavy Oil in the "World Energy Council 2001 Survey of Energy Resources." Reprinted by permission of World Energy Council, London, United Kingdom.

SEE ALSO THE FOLLOWING ARTICLES

Gas Hydrates • Natural Gas Resources, Unconventional • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Oil Recovery • Oil Shale

Further Reading

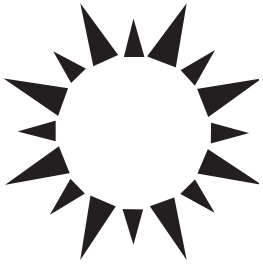
Alberta Environment. (1999). "Guidelines for Reclamation to Forest Vegetation in the Alberta Oil Sands Region." Conservation and Reclamation Information Letter. C&R/IL/99-1. Calgary.

Ferguson, B. G. (1986). "Athabasca Oil Sands: Northern Resource Exploration, 1875–1951." Canadian Plains Research Center Regina, Saskatchewan.

Shih, S.S., and Oballa, M. C. (Eds.). (1991). "Tar Sand and Oil Upgrading Technology." American Institute of Chemical Engineers, New York.

Oil Shale

JOHN R. DYNI
U.S. Geological Survey
Denver, Colorado, United States



1. Introduction
2. What Is Oil Shale?
3. Age and Origin of Oil Shale
4. Oil Shale Types
5. Retorting Methods
6. Oil Shale Deposits
7. Summary of Resources
8. Oil Shale Production
9. Environmental Concerns
10. Future Trends

Glossary

by-products Mineral commodities and other products that are produced from an oil shale operation. These may include products from retorting oil shale, such as sulfur and ammonium sulfate, or extraction of metals from the raw or spent shale.

char The noncombusted portion of the organic matter that remains with the mineral residue of oil shale after retorting.

Fischer assay A laboratory method of determining the quantity of synthetic oil that a sample of crushed oil shale will produce upon heating in a laboratory bench-scale retort.

hydroretort A retort in which oil shale is thermally decomposed in the presence of hydrogen under pressure.

kerogen The same as organic matter as defined in this glossary.

***in situ* retort** A method of retorting oil shale underground that offers the advantages of minimum materials handling and reduction in atmospheric pollutants.

lacustrine pertains to lakes; a depositional environment for some oil shales.

oil shale resource The total amount of shale oil in a deposit, including those portions that are not economically or technologically recoverable at present.

organic matter The portion of an oil shale composed mainly of carbon, hydrogen, and oxygen that can be retorted to yield oil, gas, water, and char.

overburden The unconsolidated sediments and bedrock that overlie an oil shale or mineral deposit.

overburden ratio The ratio of the thickness of overburden to the thickness of a mineable bed of oil shale or other mineral.

retort A closed vessel wherein crushed oil shale is heated to a temperature at which the oil shale will thermally decompose the organic matter into oil, gas, and water.

spent shale The mineral waste from retorting oil shale that usually includes some char derived from the organic matter.

Resources of oil shale are abundant in the world. Oil shale deposits range from Cambrian to Tertiary in age and were formed in a variety of geologic environments, including continental bogs and swamps, commonly in association with coal deposits, lacustrine (lake), and marine environments. Some deposits are too small to have economic value, whereas others with great potential value underlie thousands of square kilometers and are hundreds of meters thick. Some deposits of oil shale are interbedded and commingled with minerals that have current or potential economic value, including trona ($Na_2CO_3 \cdot NaHCO_3 \cdot 2H_2O$), nahcolite ($NaHCO_3$), and dawsonite [$NaAl(OH)_2CO_3$]. Other deposits contain large, but generally low-grade, resources of metals, notably uranium, molybdenum, vanadium, copper, zinc, chromium, and nickel.

1. INTRODUCTION

Oil shale is a low-unit-cost energy commodity that requires mining and retorting thousands of tons of oil shale per year to achieve profitability. For example, one short ton of a commercial grade oil shale from the Green River Formation in Colorado assays about 35 U.S. gallons of shale oil per short ton of rock (approximately 145 liters per metric ton), or about 0.83 U.S. barrel of oil. About one-fifth of the rock by weight is organic matter that can be retorted to obtain oil and gas. The remaining four-fifths of the

oil shale is inert mineral matter that is mined with the organic matter, passed through the retort, and disposed of in an environmentally acceptable way.

Developing a commercial oil shale operation includes opening a mine, building rock-crushing facilities, constructing a retort (or several retorts) of 10,000 or more bbls per day, and perhaps adding a hydrotreating plant for upgrading the shale oil. The costs for such a facility would be many millions of dollars. Additionally, acquiring water rights and bonus bids for oil-shale tracts on Federal lands increase start-up costs. Various factors preclude shale oil from developing a major share of the today's fossil energy market, including using today's technology, high capital costs, relatively low shale-oil production rates, yields of low-hydrogen shale oil from the retrot that needs upgrading for shipment to an oil refinery, disposal of spent shale, and high operating costs. However, new innovative technologies of mining and retorting may reduce the operating costs to a point where oil shale would become competitive in price with petroleum and natural gas.

However, some countries have oil shale industries, including Australia, Brazil, China, Germany, and Estonia. The largest oil shale operation by far is in Estonia, where approximately 12–13 million metric tons of oil shale is mined per year. Approximately 80% is burned as fuel in two large electric power plants located in northeastern Estonia, 15% is retorted for shale oil, and the remainder is used to manufacture cement. Other countries, including Sweden, Great Britain (Scotland), and Canada, have had oil shale industries in the past, but these have closed owing to the availability of cheaper supplies of petroleum. Some countries, including Israel, Jordan, and Morocco, which have sizable deposits of oil shale but lack other fossil fuels, are interested in developing their oil shale resources.

This article defines oil shale and discusses its origin, reviews the resources of world oil shale deposits, and outlines the methods of extraction of combustible products from oil shale. Environmental problems associated with the commercialization of oil shale are briefly reviewed. Future trends in the utilization of oil shale as a resource of energy, petrochemicals, and other by-products are discussed.

2. *WHAT IS OIL SHALE?*

Oil shale is commonly defined as a fine-grained sedimentary rock containing organic matter that will yield abundant amounts of oil and gas upon destructive distillation. Most of the organic matter is insoluble in ordinary organic solvents; therefore, it must be decomposed by heating to release oil and combustible gases.

Oil shales range widely in organic content and oil yield. Commercial grades of oil shale as determined by the yield of shale oil range from approximately 100 to 200 liters per metric ton (mt) of rock. The U.S. Geological Survey has used a lower limit of 40 liters/mt, whereas other researchers have suggested a limit as low as 25 liters/mt.

How much oil and gas (or heat energy) an oil shale can produce depends not only on the amount of organic matter present in the rock but also on the composition of the organic matter as well as the geothermal history of the deposit. Some deposits have been so altered by geothermal heating that the organic matter has been degraded to the point that it will yield little oil and gas. Commercial deposits, therefore, are those that have not been subjected to strong geothermal alteration by deep burial depth or thermally altered by magmatic sources.

2.1 Determination of Oil Shale Grade

When oil shale is heated in a closed vessel (i.e., a retort) in the absence of air to approximately 500°C, the organic matter thermally decomposes into oil, gas, water, and char. The grade of an oil shale is commonly determined by the Fischer assay method. This method is a laboratory procedure for measuring the amounts of oil, water, "gas plus loss," and residue composed of mineral matter and char that a 100-g sample of crushed oil shale will produce in a small retort when heated under a set of standard conditions. Although the Fischer assay has been widely used to determine the grade of many oil shale deposits, it has several drawbacks. The amount of char in the spent shale as well as the composition of the gases produced by the Fischer assay are not determined. Typically, the gases consist mainly of methane, with some heavier alkanes, hydrogen, and carbon dioxide. The char may be burned for heat energy. Therefore, the Fischer assay does not measure the total available energy in the oil shale.

An additional complication is that certain types of retorts or retorting processes can produce more oil than that reported by Fischer assay. One such process is hydroretorting, whereby the crushed oil shale is retorted in a hydrogen atmosphere under pressure. This method has been demonstrated in the laboratory to produce more than 200% of Fischer assay yields for some types of oil shale.

2.2 Products from Oil Shale

Oil obtained by retorting oil shale is essentially a synthetic crude oil that requires refining, like any petroleum crude oil, to produce transportation fuels, heating oil, asphalt, and other petrochemicals. Some oil shales are burned directly as a low-grade fuel in electric power plants or are even utilized in the manufacture of cement. By-products obtained during retorting of oil shale include sulfur and ammonium sulfate.

Some oil shales have notably high concentrations of potentially valuable metals and minerals. The Swedish Alum shale (an oil shale), which contains Europe's largest low-grade resource of uranium, was mined for alum salt from the 1600s to the early 1900s. During World War II oil was produced from the Alum shale for military use. Shale oil production continued until 1966, when the operation ceased due to import of cheaper supplies of petroleum. The Alum shale was also mined for uranium from 1965 to 1969.

A resource estimated at 29 billion metric tons of water-soluble nahcolite ($NaHCO_3$) is commingled with thick, high-grade deposits of oil shale in the Green River Formation in Colorado. The nahcolite (but not the oil shale) is solution mined for sodium bicarbonate and soda ash (Na_2CO_3). Enormous deposits of water-soluble trona interbedded with oil shale, marlstone, and mudstone are present in the same formation in southwestern Wyoming. The Wyoming trona deposits, the world's largest known mineral resource of soda ash, are mined by several companies. Sodium bicarbonate and soda ash are important industrial chemicals that are used domestically and are exported to many countries for the manufacture of glass, detergents, and many other industrial products.

3. AGE AND ORIGIN OF OIL SHALE

Oil shale deposits range from Cambrian to Tertiary in age and are found in many countries. Oil shales are formed in a variety of geologic environments, including swamps and bogs (often in association with coal deposits), lakes of small to very large size, and marine settings commonly covering hundreds to thousands of square kilometers.

Much of the organic matter in oil shale is derived largely from various types of freshwater and marine algae. For example, much of the organic matter in the lacustrine Green River oil shale of Eocene age in the western United States is thought to be derived from blue-green prokaryotic algae (cyanobacteria). Probably other bacteria including sulfate reducers were also abundant, as evidenced by the presence of iron sulfide minerals and the absence of sulfate salts that should have precipitated during the evaporative stages of the Green River lakes. Algal precursors also make up much of the organic matter in marine oil shales. Specific types of algae comprise the bulk of the organic matter in some oil shales. Fragments of land plant material derived from stems, leaves, pollen, spores, and cuticles can be carried into a lake or sea by streams and be deposited with the remains of the aquatic microorganisms. Such fragments can often be recognized under the microscope and can help determine the paleogeographic setting and identify precursors of the organic matter in an oil shale deposit.

Some conditions favorable for the formation and preservation of the organic matter include bodies of water in which current and wave action are minimal and the oxygen content of the bottom waters is low. The waters are commonly thermally or chemically stratified and may contain relatively high levels of H_2S, which is toxic to many bottom-feeding organisms.

4. OIL SHALE TYPES

In early studies many local, often colorful, names were applied to specific deposits of oil shale. Some examples include kukersite, kerosene shale, bituminite, alum shale, cannel coal, wollongite, albertite, and candle coal. With modern analytical tools including blue fluorescence microscopy, it has been found that lipid-rich organic components of the organic matter fluoresce in bright colors, making it possible to identify some of the organisms from which the organic matter was derived. The method has proved useful in developing the following rational classification of oil shales.

Organic-rich sedimentary rocks can be classified as bitumen-impregnated rocks, humic coals, and oil shales. Based of their origin, three major types of oil shale are recognized: terrestrial, lacustrine (lake), and marine oil shale. These types can be subdivided in part on the basis of organic petrography into cannel coal (terrestrial oil shale), lamosite and torbanite (lacustrine oil shale), and kukersite, tasmanite, and marinite (marine oil shale) (Fig. 1).

Cannel coal (also known as sapropelic coal) is derived from planktonic debris, spores, pollen, and other very fine detritus of aquatic plants deposited in more or less open bodies of water in a swamp or bog under

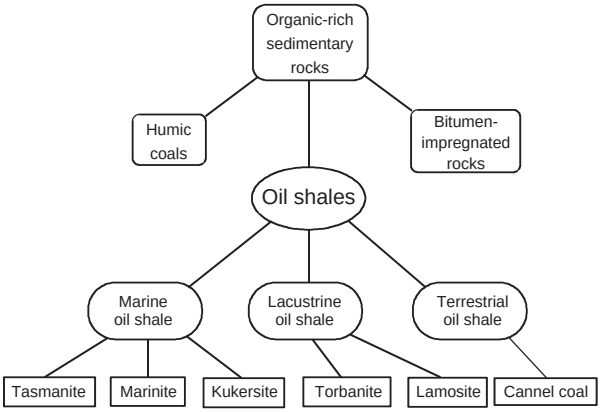


FIGURE 1 Classification of oil shales.

anaerobic conditions. The organic matter is rich in lipid and proteinaceous material, giving rise to a high oil yield upon retorting. Cannel coals can be closely associated with humic coals that originate from peat derived from terrestrial plants growing in swamps.

Lacustrine oil shales are derived largely from algae and associated phytoplankton in lakes whose waters range from fresh to saline under dysaerobic to anaerobic conditions. Two types of lacustrine oil shale are recognized: lamosites and torbanites. Lamosites are oil shales derived from lacustrine algae and other phytoplankton. Some of the organic matter may include leaf and stem fragments, pollen, and spores derived from land plants carried by streams into the lake. Fossil remains of clams, mollusks, fish, and other aquatic fauna can be found in some deposits. However, such organisms are usually sparse in parts of the oil shale further from shore, where the waters were stagnant and deficient in oxygen. Thin, wavy laminae and streaks of lipid-rich algal remains are common. Some thinly laminated black algal mats may also be present. The Green River oil shale deposits of the western United States are lamosites.

Torbanite is a lacustrine oil shale in which a substantial amount of the organic matter is derived from the alga *Botryococcus* or related species of lacustrine algae. Terrestrial plant debris is a common constituent of the organic matter. Some torbanites form very high-grade oil shales yielding as much as 45–55% shale oil by weight of the rock.

Marine oil shales include marinite, tasmanite, and kukersite. These are formed in restricted seas in which current and wave action is minimal, and the waters are commonly low in oxygen content. The most common type of marine oil shale is marinite. Tasmanite and kukersite are specific to certain types of algae.

Marinites are derived mainly from marine phytoplankton. Some aquatic organisms that can be identified under the microscope include tasmanitids, *Pediastrum* (a marine green alga), and spores. A significant portion of the organic matter in marinites consists of very fine-grained, poorly fluorescing material called bituminite. Its origin is not clear, but it may consist of degraded algal and bacterial detritus. Land plant detritus may be more or less abundant depending on the location of the accumulating organic-rich sediments in the sea relative to land areas.

Tasmanite is a variety of marine oil shale containing a large organic component derived from tasmanitids, a thick-walled, unicellular, marine green alga. Kukersite is another variety of marine oil shale whose organic matter is derived mainly from a specific type of green alga. There is an important deposit of kukersite in northeastern Estonia.

5. RETORTING METHODS

Extraction of energy (oil, gas, and heat) from oil shale takes essentially two forms: (i) the recovery of shale oil and combustible gas by retorting or (ii) burning the oil shale directly for heat energy. Two basic methods of recovering oil from oil shale are surface retorts and *in situ* methods. Surface retorts include horizontal or vertical units in which oil shale, crushed to an appropriate size (commonly 1- to 10-cm particles), is heated to approximately 500°C. The organic matter thermally decomposes into vapors of oil and combustible gases, which are recovered. Part of the organic matter forms a char that remains with the discharged spent shale after retorting. The char can be burned for heat energy useable elsewhere in an oil shale facility (Fig. 2).

A variety of surface retorts have been used to recover shale oil. Some are batch retorts, whereby the retort is charged with a given amount of crushed shale, which is retorted, and oil vapors are collected from a section of the retort and cooled and condensed to liquid. The spent shale is removed from the retort before the next batch of raw oil shale is introduced. Other retorts operate on a continuous feed basis whereby a stream of crushed shale moves through the retort and the spent shale exits at the other end of the retort. One type of experimental "hydroretort" has been designed to operate under a hydrogen atmosphere at an elevated pressure. Hydroretorting increases yields for some oil shale deposits by 200% or more compared to yields obtained by Fischer assay.

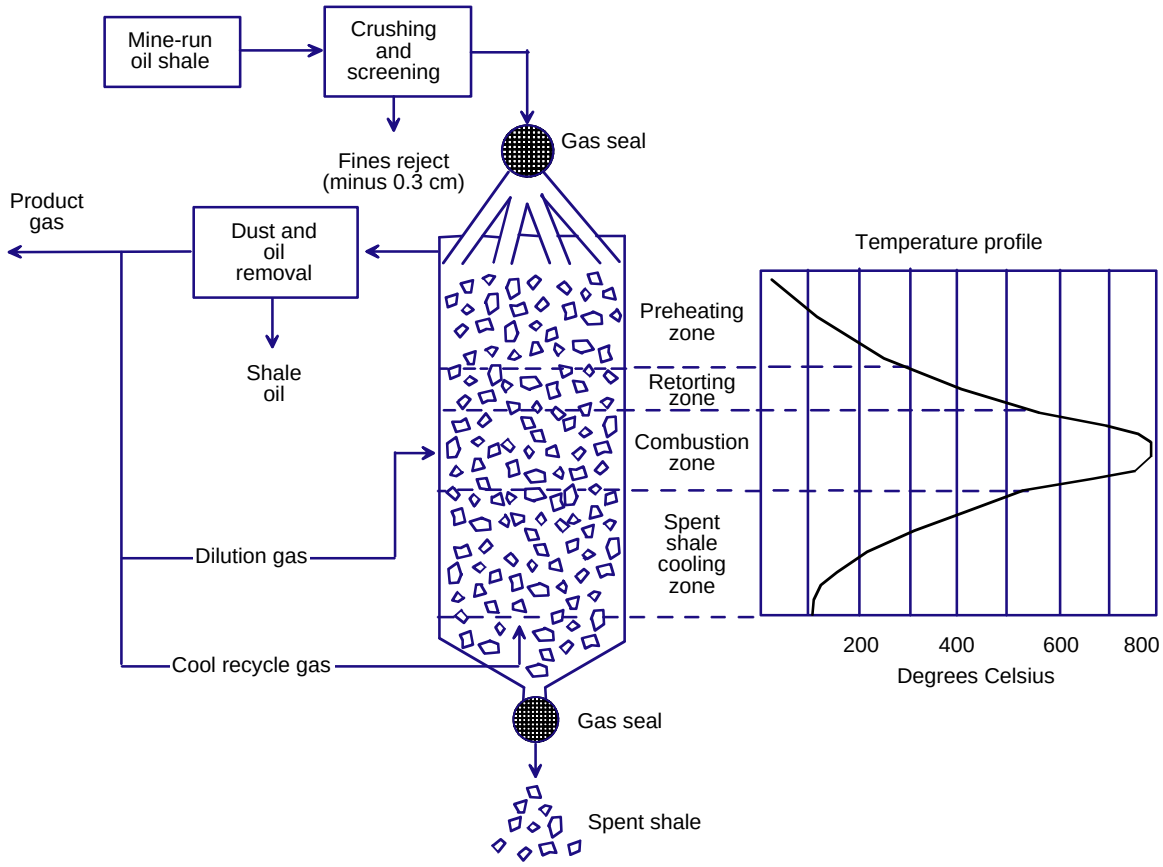


FIGURE 2 Schematic cross section of typical oil shale retort.

In Estonia, two types of retorts are in use: the Kiviter and Galoter retorts. The former retort can process 2.5- to 12.5-cm particles of oil shale but not smaller ones. The Galoter retort was developed to process smaller sized particles ranging up to 2.4 cm. The Tasiuk retort, originally designed to process Canadian tar sands, is used to process oil shale from the Stuart deposit in Queensland, Australia. It is a horizontal unit that uses part of the spent shale for preheating the raw shale. In Brazil, the vertical Petrosix retort, based on an early U.S. Bureau of Mines design, produces oil from the Iratí oil shale. In Colorado, the Tosco II horizontal retort utilized preheated ceramic balls to retort the crushed oil shale. The Unocal II retort, a vertical unit, was unusual in that the crushed shale was pumped into the bottom of the retort and the spent shale was discharged at the top.

Shale oil may also be recovered by *in situ* methods, such as retorting the oil shale through boreholes drilled into an oil shale deposit. The borehole is a retort in which the oil shale is heated by one of a variety of methods to thermally decompose the organic matter to oil and gas. The oil shale may be retorted by electrical resistance heating, radio-frequency heating, propane-fired burner units, or heat exchangers whereby the source of the primary heat for the exchanger may be electrical heating by fossil fuels or by a small nuclear reactor.

Modified *in situ* retorting involves mining galleries above and below a column of oil shale several tens of meters wide by several hundred meters deep and rubblizing the column of oil shale between the upper and lower galleries by explosives. The column of broken oil shale, which serves as a retort, is ignited at its top. Air is pumped into the top of the retort to sustain a fire front that moves downward through the rubblized oil shale. The shale oil vapors that form in the heated zone ahead of the fire front move downward and collect as oil in the bottom gallery, where it is pumped to the surface.

Several problems have been encountered with field tests using modified *in situ* retorting. The distribution of particle sizes of oil shale after blasting is not optimal because the largest particles are not completely retorted. The fire front does not move

uniformly downward through the rubblized shale but follows channels through the retort and bypasses unretorted shale. Further field experiments are needed to resolve these problems.

6. OIL SHALE DEPOSITS

Oil shales are widely distributed throughout many countries. Although the total world resource is unknown, estimates are as high as 500 or more billion metric tons of shale oil. How much of these resources can be recovered with today's mining and processing technologies is highly speculative, but if as much as 20% of these resources is recoverable, 100 billion metric tons of shale oil would still represent an enormous fossil energy resource.

Some deposits have been relatively well investigated by geologic mapping and by core drilling. Others are only poorly known, but some of these may prove to be major world-class deposits when they are more thoroughly explored. Some better known deposits are described in detail here to better acquaint the reader with the wide variety of oil shale deposits that exist.

6.1 United States

The United States is endowed with two exceptionally large resources of oil shale: the oil shale deposits of the lacustrine Green River Formation of Eocene age in Colorado, Wyoming, and Utah and the marine Devonian black shales of the eastern United States. Smaller deposits of oil shale of submarginal economic value are found elsewhere in the United States.

6.1.1 Green River Formation

Lacustrine sediments of the Green River Formation were deposited in two large inland lakes that occupied as much as 65,000 km^2 in several structural–sedimentary basins in northwestern Colorado, northeastern Utah, and southwestern Wyoming during Early through Middle Eocene time. Lake Gosiute, which occupied the Green River and Washakie Basins in southwestern Wyoming, was separated from Lake Uinta to the south in Colorado and Utah by the Uinta Mountain Uplift and its eastern extension, the Axial Basin Anticline (Fig. 3). The two lakes were in existence for more than 10 million years during a period of warm temperate to subtropical climatic conditions. Episodic volcanism introduced much volcanic ash to the lakes and their drainage basins. The warm alkaline lake waters together with water-soluble nutrients derived from the ash, provided excellent conditions for the abundant growth of blue-green algae (cyanobacteria). At times, the lake basins were closed and drier conditions prevailed, which led to precipitation of billions of metric tons of soluble sodium salts, including halite, nahcolite, trona, and shortite ($Na_2CO_3 \cdot 2CaCO_3$). In the Piceance Creek Basin in northwestern Colorado, thick high-grade oil shale commingled with 29 billion tons of nahcolite was deposited.

At times, the Green River lakes hosted a variety of fishes, rays, bivalves, gastropods, ostracods, and other aquatic fauna. Land areas peripheral to the lakes supported a large and varied assemblage of land plants, insects, amphibians, turtles, lizards, snakes, crocodiles, birds, and numerous mammalian animals.

Numerous attempts to establish an oil shale industry in the western United States, especially in northwestern Colorado, have been unsuccessful. The Unocal oil shale operation was the last major project to suspend operations in 1991. Construction costs for the oil shale facility were $650 million. The company produced 657,000 tons of shale oil, which was shipped to Chicago for refining under a price-support program subsidized by the U.S. government. The average rate of production during the last months of operation was approximately 857 tons of shale oil per day. Several major oil companies continue to hold large tracts of oil shale lands and water rights in northwestern Colorado for future development.

Shale oil resources in the Green River Formation in Colorado are estimated at 171 billion tons of shale oil. The Green River shale oil resources in northeastern Utah are estimated at 48 billion tons. In southwestern Wyoming, the Green River shale oil resources are estimated to be $\geq$35 billion tons. Additional oil shale resources are also present in the Washakie Basin east of Rock Springs, Wyoming.

6.1.2 Eastern Devonian Black Shales

Black organic-rich marine shale and associated sediments of late Devonian and early Carboniferous age underlie approximately 725,000 km^2 in the eastern United States. These shales have been exploited for natural gas for many years, but they have also been considered as a potential resource of shale oil and uranium.

The Devonian black shales were deposited in a large epeiric sea that covered much of the middle and eastern United States east of the Mississippi River (Fig. 4). This area included the broad, shallow, interior

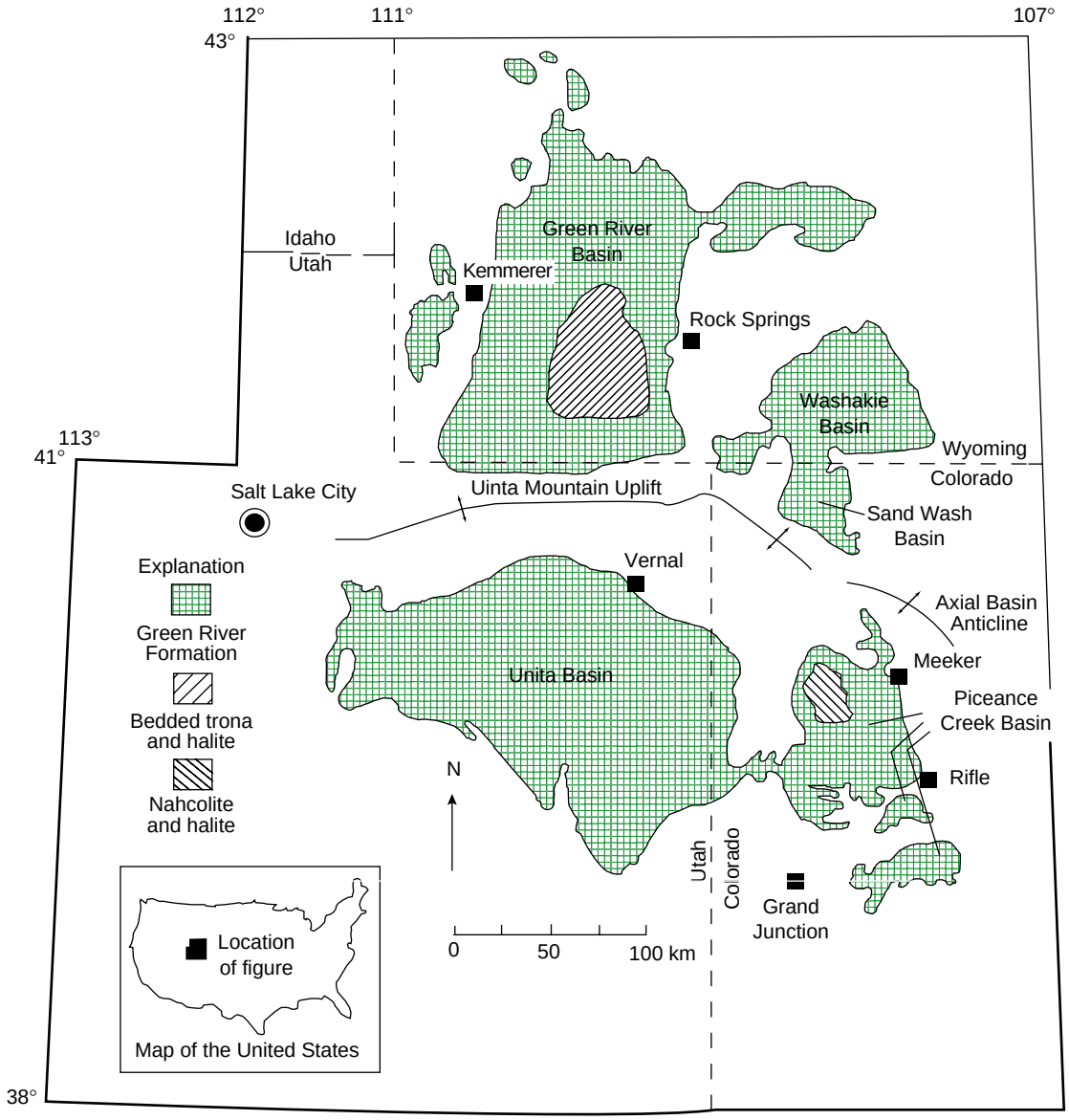


FIGURE 3 Areal distribution of the Green River Formation containing oil shale and sodium carbonate resources in Colorado, Wyoming, and Utah.

platform to the west, which grades eastward into the Appalachian Basin. The depth to the base of the Devonian black shales ranges from zero for surface exposures to 2700 m along the depositional axis of the Appalachian Basin. The oil shale resource is in that part of the basin where the organic content is highest and where the black shales are closest to the surface.

The late Devonian sea was relatively shallow with minimal wave and current activity. A large part of the Devonian black shales is amorphous nonfluorescing material, although a few structured marine fossil organisms, including *Tasmanites*, *Botyrococcus*, and *Foerstia*, have been recognized. Conodonts and linguloid brachiopods are sparingly distributed through some beds. Although much of the amorphous organic matter (bituminite) is of uncertain origin, it is generally believed that it was derived from planktonic marine algae.

In distal parts of the Devonian sea, the organic matter accumulated very slowly with very fine-grained clayey sediments in poorly oxygenated waters free of burrowing organisms. Thirty centimeters of such oil shale, on a compacted basis, are estimated to represent as much as 150,000 years of sedimentation.

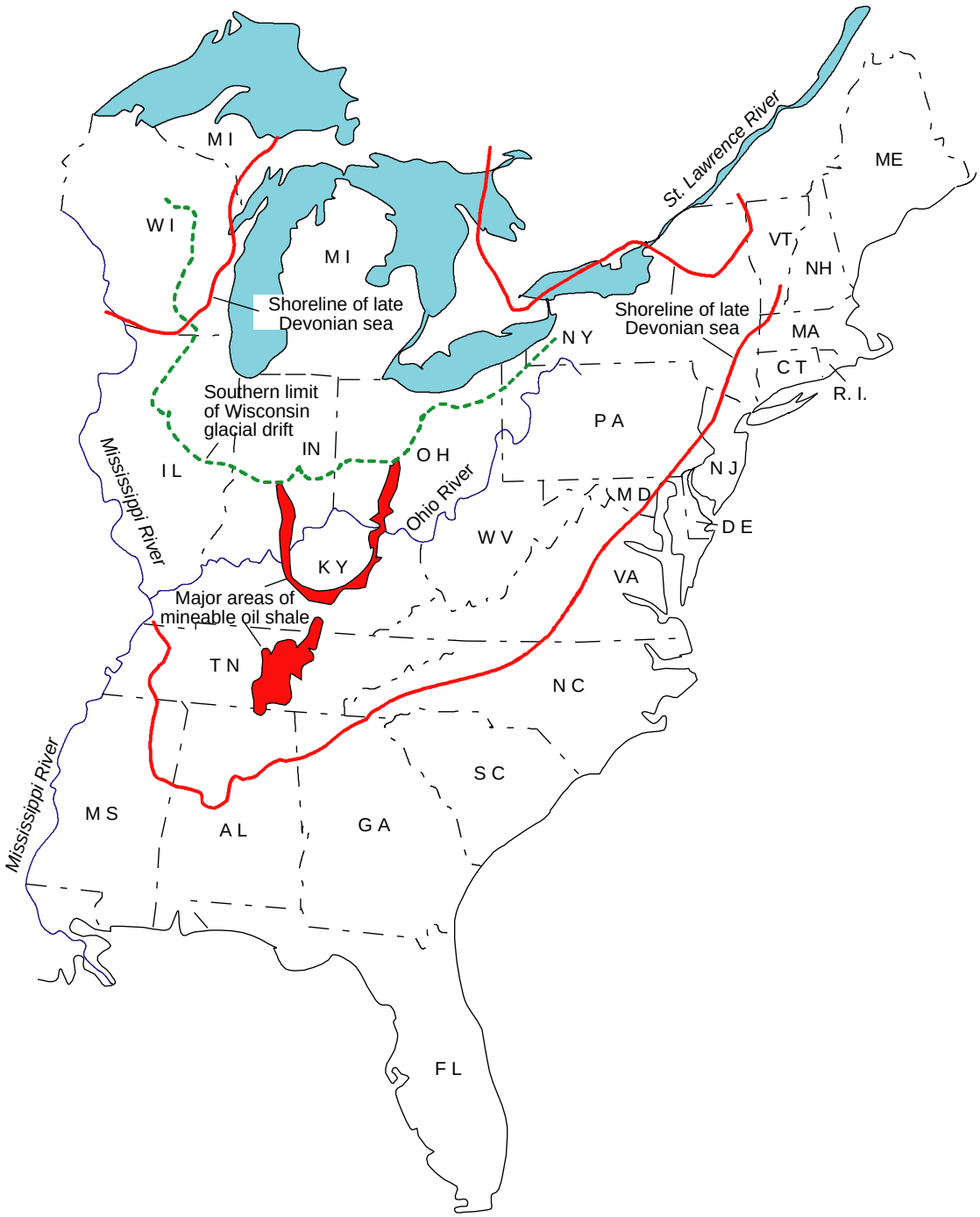


FIGURE 4 Map of the eastern United States showing the approximate limit of the Devonian sea, the southern limit of Pleistocene glacial drift, and the area of near-surface mineable Devonian oil shale.

The black shales thicken eastward into the Appalachian Basin due to increasing amounts of clastic sediments shed into the sea from the Appalachian highland east of the basin. Pyrite and marcasite are abundant accessory minerals, but carbonate minerals comprise only a minor fraction of the black shales.

The organic matter in the Devonian black shales yields only half as much oil as an equal weight of organic matter in the Green River oil shale. This

difference is attributable to the type of organic carbon found in each oil shale. The organic matter in the Devonian oil shale has a higher ratio of aromatic to aliphatic carbon than that of the Green River oil shale. As a consequence, laboratory analyses show that the Devonian shale produces much less oil and more methane and char than the Green River oil shale. However, the oil yield from the Devonian shale can be increased by more than 200% by hydroretorting.

The mineable shale oil resource of the Devonian black shales (Fig. 4) is estimated to be 423 billion barrels (~61 billion tons). This figure is based on the following criteria: organic content ≥10 weight percent, overburden ≤60 m, stripping ratio ≤2.5:1, thickness ≥3 m, mining by open pit, and processing by hydroretorting.

6.2 Other Deposits

Some examples of oil shale deposits in other countries, including several countries that have oil shale industries, follow.

6.2.1 Australia

The oil shale deposits of Australia range from small noneconomic occurrences to deposits large enough to be exploitable. The demonstrated oil shale resources for 12 deposits are estimated to be 58 billion tons, from which 24 billion barrels (~3.1 billion tons) of shale oil are estimated to be recoverable. The Australian deposits, which range in age from Cambrian to Tertiary, are of diverse origin. These deposits are located in eastern Queensland, New South Wales, South Australia, Victoria, and Tasmania. Those having the best potential for development are the Tertiary lacustrine oil shale deposits in Queensland. One of these, the Stuart deposit, contains an estimated 36.4 billion barrels (~5.2 billion tons) of shale oil. An open-pit mine and demonstration retorting facility have been constructed. The first shipment of Stuart shale oil to market was made in 2001, and as of February 2003, a total of 700,000 barrels (~100,000 tons) of shale oil have been produced and marketed.

6.2.2 Brazil

At least nine occurrences of oil shale ranging from Devonian to Tertiary in age have been reported in different parts of Brazil. Of these, oil shale of the Permian Iratí Formation has the greatest potential for development because of its accessibility, grade, and widespread distribution. The formation crops out in the form of a large "S" extending from the northeastern part of the state of São Paulo southward for 1700 km to the southern border of Rio Grande do Sul and into northern Uruguay. In the state of Paraná, in the vicinity of the town of São Mateus do Sul, two economic beds of oil shale have been identified. The upper one, 6.5 m thick, and the lower one, 3.2 m thick, are separated by approximately 8 m of barren shale and limestone. The two beds are mined by open pit. An *in situ* resource of approximately 11.5 billion tons of shale oil has been identified.

The Iratí oil shale is dark gray, brown, and black, very fine grained, and laminated. Clay minerals comprise approximately 70% of the shale, with minor contributions of detrital quartz, feldspar, and pyrite; carbonate minerals are sparse. Unlike the Devonian black shales of the eastern United States, the Iratí oil shale is not notably enriched in metals. The oil shale is high in sulfur (approximately 4%) and the moisture content is approximately 5%. Gross heating value (dry basis) is approximately 1500 kcal/kg of shale.

The origin of the Iratí oil shale is unclear. Some researchers have argued that the oil shale was deposited in a fresh to brackish water lacustrine environment. Others have suggested that the oil shale originated in a partially enclosed intracontinental marine basin of reduced salinity connected to the open sea that formed after the close of late Carboniferous glaciation.

Development of the Iratí oil shale began at São Mateus do Sul, where a prototype retort and processing plant were constructed. The facility began operations in 1972 with a design capacity of 1600 tons of oil shale per day. In 1991, an industrial-size vertical retort, 11 m in diameter by 60 m high, with a design capacity of 6200 tons of oil shale per day, was put into operation. Between 1980 and 1998, approximately 1,550,000 tons of shale oil was produced, with by-products of sulfur, liquefied petroleum gas, and fuel gas.

6.2.3 Estonia

In Estonia, two kukersite deposits are recognized: the Middle Ordovician Estonia deposit, which is continuous with the Leningrad deposit in Russia, and the somewhat younger Middle Ordovician Tapa deposit. Together, they occupy more than 50,000 km^2.

In the Estonia deposit, there are as many as 50 beds of kukersite, commonly 10–40 cm in thickness, and kerogen-rich limestone that alternates with fossiliferous fine-grained limestone. The organic

content of individual beds of kukersite can reach 40–45 weight percent of the oil shale and yield as much as 320–500 liters of oil per ton of rock. In the operating mines, the ore zone is approximately 2.5 m thick and has a heating value of 2440–3020 kcal/kg. Most of organic matter in kukersite is derived from the fossil marine green alga *Gloeocapsomorpha prisca*, which has affinities to the modern cyanobacterium *Entophysalis major*, an extant species that forms algal mats in intertidal to very shallow subtidal marine waters. The mineral fraction of the kukersite consists of calcite (>50%), dolomite (<10–15%), and quartz, feldspar, illite, chlorite, and pyrite (<10–15%).

The interbedded limestone and kukersite sequence was deposited in a shallow subtidal marine basin adjacent to a shallow coastal area on the north side of the Baltic Sea near Finland. The abundance of marine macrofossils and low pyrite content of the sequence suggest oxygenated waters, unlike most depositional environments for oil shale. The widespread continuity of the even thin beds of kukersite suggests negligible bottom currents.

The kukersite deposits of Estonia have been known since the 1700s, but their exploitation did not start until World War I as a result of fuel shortages brought about by the war. Full-scale mining began in 1918 when 17,000 tons of oil shale was mined by open pit. By 1940, annual production of oil shale had reached 1.7 million tons. However, it was not until after World War II, during the Soviet era, that production increased dramatically, peaking in 1980 when 31 million tons of oil shale was mined by 11 open-pit and underground mines. The annual production of oil shale decreased to 14 million tons in 1995 and has ranged from 10 to 15 million tons until 2000. Approximately 80% of the mined kukersite is burned as fuel in two large electric power plants, 15% is retorted for shale oil to make a variety of petrochemicals, and a few percent is used to manufacture cement and other products. In 1997, the Estonian oil shale industry was sustained in part by state subsidies amounting to approximately $10 million. Mineable reserves of kukersite are estimated to be 6 billion tons. Since 1917, 850 million tons of oil shale has been mined.

The Tapa deposit is still in the prospecting stage and the resources are yet to be evaluated. Available data suggest that the deposit is of lower grade than the Estonia deposit.

The marine *Dictyonema* shale of Early Ordovician age underlies most of northern Estonia and is probably equivalent to the upper part of the Alum shale of Sweden. Until recently, little was known about the *Dictyonema* shale because it was covertly mined and processed for uranium at a secret plant located at Sillamäe in northeastern Estonia during the Soviet era. The shale ranges from less than 0.5 to more than 5 m in thickness. A total of 22.5 tons of uranium was produced from approximately 272,000 tons of Dictyonema shale.

6.2.4 Israel

Twenty marinite deposits of Late Cretaceous age have been found in Israel and approximately 12 billion tons of oil shale reserves has been identified. The organic content of the marinites is relatively low, ranging from 6 to 17 weight percent, and the oil yield is only 60–71 liters/ton. The moisture content is high (~20%), as is the carbonate content (45–70% calcite) and sulfur content (5–7%). Some of the deposits are close enough to the surface to be mined by open pit. A commercially exploitable bed of phosphate rock, 8–15 m thick, underlies the Misor Rotem oil shale mine in south central Israel. The oil shale from this mine has a heating value of 650–1200 kcal/kg. It was burned in a fluidized bed boiler to provide steam to operate a turbogenerator in a 25-MW electric power plant. The plant began operating in 1989 but is now closed.

6.2.5 Jordan

Approximately 26 occurrences of oil shale are found in Jordan, but relatively few have been studied in detail or explored by core drilling. The deposits commonly occur in fault-bounded synclines (grabens) covered by unconsolidated gravel and silt with lenticular beds of marl and limestone and, locally, by volcanic rocks. The oil shales are marinites of Late Cretaceous to Early Tertiary age. Eight deposits that have been explored to some extent yield approximately 7–11% oil by weight, and their average thickness ranges from 30 to 136 m. Resources in terms of shale oil for 6 of these 8 deposits range from 74 million to 2.15 billion tons.

Overburden ratios range from 0.4:1 to 2.2:1. The oil shales are low in moisture, averaging approximately 2–5.5 weight percent. The sulfur content for some of the oil shales is approximately 0.3–4.3 weight percent. Metals of possible economic value, including uranium, molybdenum, vanadium, chromium, and nickel, occur in concentrations ranging from approximately 20 to 650 parts per million. Surface water supplies are scarce in Jordan; therefore, groundwater aquifers will need to be

tapped for an oil shale industry. Because Jordan has few resources of oil and gas and no coal, it has been actively exploring its oil shale deposits for their possible use as fuel in electric power plants.

6.2.6 China

Two noteworthy deposits of oil shale are found at Fushun in Liaoning Province in northeastern China and at Maoming in the western part of Guangdong Province approximately 350 km west of Hong Kong. At Fushun, the oil shale overlies a thick deposit of subbituminous to bituminous coal of Eocene age. The oil shale is lacustrine in origin. The oil shale ranges from 48 to 190 m in thickness. At the West Open Pit coal mine, it is 115 m thick. The lower 15 m consists of low-grade, light-brown oil shale and the upper 100 m consists of richer grade, brown to dark brown finely laminated oil shale. The contact between the underlying coal and oil shale is gradational, suggesting a peat swamp environment that gradually subsided and was replaced by a lake in which the oil shale was deposited. The oil yield ranges from 5 to 16 weight percent. The oil shale resources in the vicinity of the mine are estimated to be 260 million tons, of which 90% is thought to be mineable. The total resource of the Fushun deposit is estimated to be 3.6 billion tons of oil shale. The first production of shale oil at Fushun began in 1930. In the mid-1990s, 60 retorts, each with a capacity of 100 tons of oil shale per day, produced 60,000 tons of shale oil per year.

The deposit of oil shale at Maoming is also Tertiary in age and lacustrine in origin. Reserves estimated at 5 billion tons are reported in an area 50 km long, 10 km wide, and 20–25 m thick. The oil yield of Maoming oil shale is 4–12 weight percent and averages 6 or 7%. The ore is yellowish brown with a bulk density of approximately 1.85 g/cc. The oil shale contains 72% mineral matter, 11% moisture, and approximately 1% sulfur, and it has a heating value of approximately 1750 kcal/kg (dry basis).

6.2.7 Russia

More than 80 deposits of oil shale have been identified in Russia. The kukersite deposit near St. Petersburg, which is continuous with the Estonian kukersite deposit, is burned as fuel in the Slansky electric power plant near St. Petersburg. Oil shale in the Volga–Pechersk oil shale province was used as fuel in two electric power plants but is no longer burned due to high SO_2 emissions.

7. SUMMARY OF RESOURCES

Estimating the total world resources of oil shale presents a number of problems. Many deposits are poorly explored and analytical data are sparse, making it difficult to determine the grade and size of the resource. Core drilling and laboratory analyses of oil shale samples are expensive, which lessens the likelihood of conducting exploration programs, especially in Third World countries with limited budgets. However, there has been considerable exploration in some countries with good deposits of oil shale but that are not well endowed with other fossil fuels, such as Australia, Jordan, Israel, and Morocco.

Resource data are reported in a variety of ways: as volumetric data (i.e., gallons and liters per metric ton), gravimetric data (U.S. short tons and metric tons), and heating values (British thermal units per pound of oil shale, kilocalories per kilogram, or megajoules per kilogram of oil shale). To bring some uniformity to the data, Table I lists the resources only in U.S. barrels and metric tons of shale oil. To do this, it was necessary to convert volumetric to gravimetric units, which requires knowledge of the specific gravity of the shale oil. With some exceptions, when the specific gravity is unknown, a value of approximately 0.910 was assumed.

The quality of the resource estimates given in Table I ranges from poor to satisfactory. Therefore, the estimates should be regarded with caution as to their accuracy. The best estimates are those given for deposits that have been explored by moderate to extensive core drilling, including the Estonia kukersite deposit; some of the Tertiary deposits in eastern Queensland, Australia; several deposits in Israel, Jordan, and Morocco; and the Green River deposits in the western United States.

The total shale oil resources in 39 countries is estimated at 487 billion tons (3 trillion barrels). Because many deposits are not included in this estimate, this total is a conservative number that should increase significantly when lesser known deposits have been explored more thoroughly.

How much of the world oil shale resources can be recovered is based on numerous factors. The size, depth, thickness, and grade of a deposit are, of course, important. However, conflicts with surface uses of lands overlying a deposit, such as towns and parks, could restrict areas that could be mined. Improved mining and processing technologies, such as *in situ* energy recovery and hydroretorting methods, could make some deposits attractive for development that would otherwise be subeconomic.

TABLE I

Shale Oil Resources of Some World Oil Shale Deposits

Country, region, or deposit	In-place shale oil resources		Country, region, or deposit	In-place shale oil resources	
	$\times 10^6$ bbls	$\times 10^6$ tons		$\times 10^6$ bbls	$\times 10^6$ tons
Argentina	400	57	Yarmouk[a]	Large	
Armenia	305	44	Kazakhstan	2837	400
Australia			Madagascar	32	5
New South Wales	40	6	Mongolia	294	42
Queensland	31,041	4432	Morocco		
South Australia	600	86	Tarfaya, zone R	42,145	6448
Tasmania	48	7	Timahdit	11,236	1719
Austria	8	1	New Zealand	19	3
Belarus			Poland	48	7
Pripyat Basin	6988		Russia		
Brazil	82,000	11,734	Central Basin	70	10
Bulgaria	125	18	Olenyok Basin	167,715	24,000
Burma	2000	286	St. Petersburg kukersite	25,157	3600
Canada			Timano–Petshorsk Basin	3494	500
Manitoba–Saskatchewan	1250	191	Turgai and Nizneiljsk deposit	210	30
New Brunswick	286	40	Volga Basin	31,447	4500
Nova Scotia	1705	244	Vyshegodsk Basin	195,967	2800
Ontario	12,000	1717	Other deposits	210	30
Chile	21	3	South Africa	130	19
China	16,000	2290	Spain	280	40
Egypt	5700	816	Sweden	6114	875
Estonia			Syria	Large?	
Dictyonema shale	12,386	1900	Thailand	6400	916
Estonia deposit	3900	594	Turkey	1985	284
France	7000	1002	Turkmanistan and Uzbekistan		
Germany	2000	286	Amudarja Basin	7687	1100
Hungary	56	8	United Kingdom	3500	501
Iraq			United States		
Yarmouk[a]	Large		Eastern Devonian shale	189,000	27,000
Israel	4000	550	Elko Formation	228	33
Italy	73,000	10,446	Green River Formation	1,734,000	258,000
Jordan			Heath Formation	180,000	25,758
Attarat Umm Ghudran	8103	1243	Phosphoria Formation	250,000	35,775
El Lajjun	821	126	Uzbekistan		
Juref ed Darawish	3325	510	Kyzylkum Basin	8386	1200
Sultani	482	74	Zaire	100,000	14,310
Wadi Maghar	14,009	2149			
Wadi Thamad	7432	1140	Total	3,323,910	487,002

[a] The Yarmouk deposit lies across the Jordanian–Iraqi border.

8. OIL SHALE PRODUCTION

A number of countries have had oil shale industries in the past. Some countries have restarted oil shale operations following periods of no activity. Australia, for example, produced shale oil from several deposits in the 1800s, but these closed when the high-grade ore was worked out. During the past

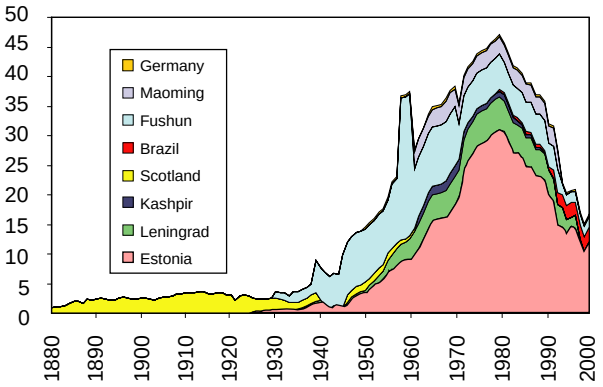


FIGURE 5 Production of oil shale in millions of metric tons from Estonia (Estonia deposit), Russia (Leningrad and Kashpir deposits), Scotland (Lothians), Brazil (Iratí Formation), China (Maoming and Fushun deposits), and Germany (Dotternhausen).

several years, a new attempt has been made to establish an oil shale industry in Queensland. Brazil has been operating an oil shale industry since the 1980s. Estonia, however, has one of the longest continuous oil shale operations, beginning during the World War I. Despite a major effort to establish an oil shale industry in Green River deposits in the western United States by the federal government, the effort failed in part to low petroleum prices and the lack of a technology to profitably recover shale oil from the deposits.

Figure 5 shows the production of oil shale from eight deposits in six countries from 1880 to 2000. World production, much of it from the Estonian kukersite deposit, peaked in 1980 when approximately 47 million tons of oil shale was mined. Production declined to a low of 15 million tons in 1999.

9. ENVIRONMENTAL CONCERNS

Open-pit mining and surface retorts present the most problems of environmental concern. Dust from mining operations can be a significant air pollutant and hazardous to workers, although wetting mine roads and piles of waste rock reduces the amount of dust. Crushing operations may also give rise to dusty conditions, but again, wetting the materials during handling reduces the amount of dust from such operations. Spent shale from some retorts contains significant quantities of phenols and other types of soluble organic compounds that are potential health hazards. Surface and groundwaters infiltrating waste piles of spent shale and overburden materials can leach soluble organic compounds, sulfates, chlorides, and carbonates that could pollute freshwater streams and lakes. Some trace elements found in some oil shales may be toxic to humans and animals if they are released to the atmosphere or into streams and lakes.

In situ operations offer the possibility of reducing air and water pollutants by leaving the overburden intact as well as leaving most or all of the spent shale underground. A detailed understanding of the subsurface geology is required to ensure that the potentially hazardous waste products of an *in situ* operation remain underground and will not surface later through springs in contact with the spent shale. Other favorable aspects of *in situ* retorting include minimum disturbance of the surface, a smaller labor force, reduced capital costs for plant construction and operation because surface retorts are not required, and possibly higher quality shale oil from subsurface retorting.

10. FUTURE TRENDS

Estonia, which is endowed with important resources of oil shale, will probably continue to operate an oil shale industry for years. Strides have been made to reduce air pollution from burning and retorting oil shale and from leaching deleterious substances from huge waste piles of spent shale in Estonia. Oil shale operations in China, Germany, Brazil, and Australia may also continue for some years and possibly expand if economic conditions are favorable.

Although current oil shale operations constitute a small industry on a worldwide basis, the enormous resources of oil shale suggest that it will find a place in the production of energy, petrochemicals, and by-products in the future. New mining and processing technologies, especially *in situ* techniques that minimize the handling of large amounts of raw and spent shale, could provide the impetus for oil shale development.

SEE ALSO THE FOLLOWING ARTICLES

Gas Hydrates • Natural Gas Resources, Unconventional • Oil and Natural Gas Drilling • Oil and Natural Gas: Economics of Exploration • Oil and Natural Gas Liquids: Global Magnitudes and Distribution • Oil Crises, Historical Perspective • Oil Recovery • Oil Refining and Products • Oil Sands and Heavy Oil • Strategic Petroleum Reserves

Further Reading

Allred, V. D. (1982). "Oil Shale Processing Technology." Center for Professional Advancement, East Brunswick, NJ.

Crisp, P. T., *et al.* (1987). "Australian Oil Shales; A Compendium of Geological and Chemical Data." University of Woolongong, Departments of Geology and Chemistry, New South Wales, Australia.

Dyni, J. R., Anders, D. E., Rex, and R. C., Jr. (1990). Comparison of hydroretorting, Fischer assay, and Rock–Eval analyses of some world oil shales. *In* "Proceedings of the 1989 Eastern Oil Shale Symposium." pp. 270–286. Institute for Mining and Minerals Research, University of Kentucky, Lexington.

"Eastern Oil Shale Symposium Proceedings" (1981–1993). Institute for Mining and Minerals Research, University of Kentucky, Louisville. [Includes papers on geology, mining, and processing of oil shale with emphasis on the Devonian black shales of the eastern United States.]

Hutton, A. C. (1987). Petrographic classification of oil shales. *Int. J. Coal Geol.* **8**, 203–231.

Kattai, V., Saadre, T., and Savitski, L. (2000). "Estonian Oil Shale." Eesti Geoloogiakeskus, Tallinn. [in Estonian with an extended English summary].

"Oil Shale" (serial publication). Estonian Academy, Publishers, Tallinn. [Includes papers, mostly in English, on geology, mining, and processing of oil shale.]

"Oil Shale Symposium Proceedings" (1964–1992). Colorado School of Mines, Golden. [Includes papers on geology, mining, and processing of oil shale with emphasis on the Green River oil shale deposits of Colorado, Wyoming, and Utah.]

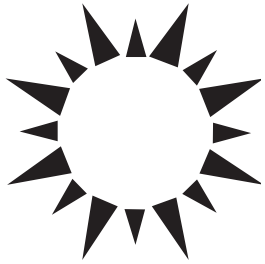
Roen, J. B., and Kepferle, R. C. (eds.). (1993). "Petroleum geology of the Devonian and Mississippian black shale of eastern United States, Bulletin 1909, Chapters A–N." U.S. Geological Survey, Denver, CO.

Russell, P. L. (1990). "Oil Shales of the World, Their Origin, Occurrence, and Exploitation." Pergamon, New York.

Smith, J. W. (1980). Oil shale resources of the United States. *Mineral Energy Resour.* **23**(6), 1–20.

OPEC, History of

FADHIL J. CHALABI
Centre for Global Energy Studies
London, United Kingdom



1. **The Oil Industry Structure under the "Seven Sisters"**
2. **OPEC's Formative Years**
3. **A Short Period of Cooperation between Producers and Oil Companies**
4. **The Oil Price Shocks**
5. **OPEC as the Last Resort Oil Supplier: The Quota System**
6. **The Backlash**
7. **OPEC's Limitations in Oil Supply Management**
8. **OPEC and a Future of Uncertainty**

Glossary

downstream In the oil industry, refers to refineries and the distribution of oil products.

horizontal integration A system that secured cooperation among the major oil companies in the upstream phase of the industry and allowed for the planning of crude oil production according to the needs of each company.

oil concessions A system that prevailed in the Middle East and Venezuela, whereby the major oil companies were given the exclusive right to explore, find, develop, and export crude oil and where the companies jointly owned an "operating company" in each country and held a quasi-monopoly among the producing/exporting companies.

OPEC basket price A "basket" of seven crudes as a reference price, namely Sahara Blend of Algeria, Dubai Crude, Saudi Arabia's Arab Light, Minas of Indonesia, Bonny Light of Nigeria, Tia Inana of Venezuela, and Isthmus of Mexico.

per barrel income Oil revenue per barrel of oil after accounting for the cost of production.

posted price A fictitious fixed price, set by the major oil companies until 1960, that served as a reference point for calculating taxes and royalties paid by them to the host countries.

"Seven Sisters" Refers to the seven major international oil companies that, until the early 1970s, rigorously controlled the world oil industry outside the United States and former Soviet Union (British Petroleum, Mobil, Socal, Jersey [Exxon], Shell, Gulf, and CFP [Total]).

upstream In the oil industry, refers to crude oil production, that is, investments in exploring, finding, and developing crude oil.

vertical integration A system by which each operating company obtained crude oil, destined for its own refineries, so as to plan crude oil production in line with the needs of those refineries.

Delegates from five major oil-producing countries—Iran, Kuwait, Saudi Arabia, Venezuela, and Iraq—met in Baghdad and announced the foundation of the Organization of Petroleum Exporting Countries (OPEC) on September 16, 1960. At that time, no one could have foreseen that this event would play such a crucial role, some 10 years or so later, in reshaping the world energy scene and would have such an impact on the world economy. The event at the time passed nearly unnoticed, except perhaps by the specialized petroleum media. It took place after two successive reductions in what was then called the posted price of oil, which previously was set and posted unilaterally by the major international oil companies, usually referred to as the "Seven Sisters." The first cut, in February 1959, was by as much as U.S. \$0.18/barrel (or $\sim$10%) for Gulf oil, from \$2.04 to \$1.86/barrel. The second price reduction of 7% was decided by Exxon in August 1960, and the other companies followed suit. The OPEC delegates' immediate and prime objective was to safeguard their member countries' oil revenue against any further erosion as a result of the companies' deciding to cut prices further. According to the profit-sharing agreements put into force during the early 1950s between the holders of oil concessions in these Middle East areas, on the one hand, and their host countries, on the other, the latter's per barrel revenue of exported oil was determined at 50% of the official (i.e., "posted") price minus half of the cost of producing that barrel. Therefore, any fluctuation in that price, whether upward or downward, changed

 © 2004 Elsevier Inc. All rights reserved.

the per barrel revenue accordingly. When the oil companies undertook price cuts, the per barrel income of these countries was reduced by more than 15%, compared with the price before the two cuts. Until the late 1950s, the official posted price was more of a tax reference price, on which host countries received their taxes, than a real market price resulting from the exchange of oil between buyers and sellers in a free market. For the companies, on the other hand, the tax paid to host countries, together with the production costs, determined the operators' tax-paid cost, meaning the cost incurred by the oil companies as a result of producing one barrel of oil plus tax and royalties paid to the host governments. For this reason, the first resolution adopted by the newly formed organization emphasized that the companies should maintain price stability and that prices should not be subjected to fluctuation. It also emphasized that the companies should not undertake any change of the posted price without consultation with the host countries. Furthermore, the resolution accentuated the necessity of restoring prices to the pre-1959 level. Other wider objectives were also articulated and incorporated into the organization's five resolutions, namely that it would endeavor to have a price system that would secure stability in the market by using various means, including the regulation of production, with a view to protecting the interests of both oil consumers and oil producers and guaranteeing stable oil revenues for the latter. Moreover, OPEC announced that the real purpose of the organization was to unify the oil policies of member countries so as to safeguard their interests individually and collectively.

1. THE OIL INDUSTRY STRUCTURE UNDER THE "SEVEN SISTERS"

Prior to the formation of the Organization of Petroleum Exporting Countries (OPEC), and until the late 1950s, the international oil industry had been characterized mainly by the dominant position of the major multinational oil companies through a system of oil concessions granted to the major oil-producing countries, according to which these companies were interlinked in the "upstream" phase of the industry, that is, exploration, development, and production of crude oil. Through joint ownership of the holding companies that operated in various countries, they were able to plan their production of crude oil according to their requirements. The oldest example of this kind of "horizontal integration" (i.e., with the companies being interlinked in the upstream phase) was the Iraqi Petroleum Company (IPC), formerly the Turkish Petroleum Company, which was a non-profit-making holding and operating company owned by the Anglo–Iranian Oil Company (British Petroleum [BP]), Compagnie Française Pétrole (CFP or Total), Royal Dutch Shell, Standard Oil of New Jersey (Exxon), Mobil, and the Gulbenkian Foundation (known as "Mr. 5 percent," Gulbenkian had been instrumental in negotiations for an agreement, prior to World War I, between the Ottoman Empire and German and British interests involved in the exploitation of Iraqi oil).

Each of these companies was a shareholder in other countries in the Middle East. For example, BP owned half of the Kuwait Oil Company and all the oil of pre-Musadeq Iran (although its holding was reduced to 40% in the consortium that was founded after Musadeq) as well as shares in Qatar Petroleum Company and Abu Dhabi Petroleum Company, whereas the American company Esso (Exxon) had 30% of Aramco in Saudi Arabia as well as shares in Qatar Petroleum, Abu Dhabi Petroleum, and the like. This type of interlinking enabled them to control and manage crude oil supplies worldwide, along with the bulk of oil exports from the major oil-producing countries, so that oil trading became a question of intercompany exchange with no free market operating outside the companies' control whereby crude oil was exchanged between sellers and other buyers. At the same time, each of these "sisters" had its own "downstream" operations—transportation, refining, oil products, and distribution networks—that made them "vertically" integrated. This compact system of horizontal and vertical integration allowed the companies to plan for their future crude oil requirements in line with their downstream requirements, that is, the amount of oil products needed by each according to its market outlets in the countries to which crude oil was shipped.

However, by the second half of the 1950s, this compact system of complete control of production and intercompany exchanges began to weaken with the appearance of independent oil companies searching for access to cheaper crude oil. This new development led to the creation, albeit on a limited scale, of a free market for buying and selling crude oil outside the control of the major companies. The state-owned Italian company, ENI, was able to find

its way in investing beyond the control of the Seven Sisters by offering different and better financial terms than the oil concessions. Similarly, many independent oil companies began to invest outside the sisters' control by offering apparently better financial terms than the oil concession system.

At the same time, oil was discovered in new producing areas, as in the case of Libya, where in addition to Exxon Mobil, Occidental was an important American independent company. These developments led to a greater amount of crude oil being produced outside the system, controlled by the "majors" and offered for sale to independent refiners with no access to crude oil. Furthermore, crude oil from the then Soviet Union began to make its way into a free market for sale at competitive prices, and this country needed to offer attractive discounts to move its oil in the market. Other factors also played roles. One was a measure taken by the United States to limit the entry of crude oil into the market so as to protect its oil industry. This policy evolved toward the end of the 1950s, when the U.S. government's oil import policy predicated itself on the premise that foreign oil should be a supplement to, and should not supplant, American oil. This deprived many independent American oil companies that had access to foreign crude oil from entering the U.S. market, a situation that led to their having to sell surplus crude outside the major oil companies' system. A market price began to shape up in the form of discounts off the posted price set by the major oil companies. On the other hand, a new pattern of tax relationship emerged following the entry of newcomers investing in oil in new areas such as Libya, where the government's per barrel share was calculated not on the official posted price (which was a fixed price) but rather on the basis of a price realized in the free market (which was always below the posted price). In this way, when the realized price fell, the host government's per barrel share fell accordingly and the tax cost for the new producer would correspondingly be lower, thereby providing a good margin of profit with which to reinvest in new production capacity for crude oil.

In this way, the free market, where crude oil is bought and sold physically and freely, started to grow. Obviously, the larger this market, the less it was dominated by the major oil companies, whose share in world production outside the United States and Soviet Union now came under threat from the newcomers. As explained by Edith Penrose in her book, *The Large International Firm in Developing Countries: The International Petroleum Industry,* in 1950 the Seven Sisters owned 85% of the crude oil production in the world outside Canada, the United States, the Soviet Union, and China. By 1960, this share had fallen to 72%, to the benefit of the newcomers, whose share had correspondingly grown from 15% to 28%. The fall in the Majors' share of refinery capacity was even more pronounced: from 72% to 53% during the same period – again to the benefit of the independent oil companies. Against this atmosphere of increasing competition and the threat to the dominant position of the major companies by newcomers and the Soviet Union, (the latter of which was selling crude at massive discounts), the major oil companies had to cut the price posted by them to reduce the tax paid cost. This was actually pushed by Esso (Exxon) and followed by the other companies; hence, the cuts in the posted price by 1959 and 1960 were a countermeasure to the newcomers' increased market share. In other words, if this mounting competition with price cuts in a free market had not been curbed by restrictive measures, this situation would have led to a harmful level of competition and continued price cuts.

2. OPEC'S FORMATIVE YEARS

The timely formation of OPEC put an end to this increasingly damaging level of competition with its adverse effect on the position of the major oil companies and on the oil industry in general. The first step that OPEC took was to enlarge its membership, and it was only shortly after its foundation that Qatar joined the organization, followed by Indonesia, Libya, the Emirates of Abu Dhabi, Algeria, Nigeria, Ecuador, and Gabon (although Ecuador and Gabon left OPEC during the 1990s). The enlargement of the OPEC base at this stage served to strengthen its negotiating power vis-à-vis the oil companies. In addition to this enlargement, OPEC took certain measures that helped greatly in putting an end to price erosion. These involved, among other things, the unification of the tax system, that is, levying taxes from the oil companies that operated in member countries' territory so as to make them conform to those applied in the Gulf (i.e., by using the fixed official posted price as a reference for the calculation of tax and royalties and not a realized market price). This measure put an end to the tax system in new areas where tax and royalties were paid by investors on the basis of realized market prices and not fixed posted prices, a system that tends to put new companies in a better competitive position than the

major companies. In so doing, OPEC succeeded in curbing the newcomers' ability to invest in expanding capacity and acquire an enlarged market share at the expense of the majors.

According to OPEC resolutions, Libya amended its tax system and put it in line with that of the Gulf. Obviously, this measure, adopted by Libya in response to OPEC, had the effect of limiting the expansion of new companies' investments as they began paying higher taxes than they did previously. This OPEC measure, while serving the producing exporting countries by consolidating price stability, also benefited the major oil companies by preventing their dominant competitive position in the oil industry from eroding further.

Another measure taken by OPEC that was of no less significance in consolidating the price structure was the amendment to the concessions' tax regime by which the payment of "royalties" (paid by investors to the owners of oilfields) was treated separately from the taxation or what was called the "expensing of the royalty" (i.e., the royalty to be paid by the concessionaires but considered as part of the production costs for the purpose of calculating taxes). Among the established rules in the oil industry, as implemented in the United States since the early part of the 20th century, royalties had to be paid to the landowners regardless of whether profits were realized or not. Taxes are levied on profits that have to be paid in addition to royalties. When the 50/50 profit sharing agreements between host governments and the major oil companies were concluded during the 1950s, the 50% belonging to the host countries covered both the royalty at 12.5% and the tax at 37.5% of the posted price.

OPEC asked the companies to implement this principle in the oil concessionary system operating in the areas, with 50% tax being paid after the host countries had received the royalty of 12.5% of the posted price. The companies agreed with this, provided that the royalty would be considered part of the cost of production. This meant that only half of the royalty would be payable to the host countries from royalties; hence, the total government share increased from 50 to 56.25% of the posted price. Although this increase had to be implemented gradually by allowing temporary decreasing discounts to the companies, it nevertheless had the effect of consolidating the price structure by increasing the tax paid cost. This measure had the effect of strengthening the level of realized market prices because the higher the tax paid cost, the higher the price floor, making price cuts more difficult.

It is clear that the mere existence of OPEC, whose raison d'être was to put an end to price cuts by the companies, helped not only to safeguard the producers' interests but also to strengthen the price structure and put an end to the expansion of the competitive market. No less important were the consequential measures taken by OPEC that had the effect of serving the mutual interests of the oil producers and major oil companies. It was because of these measures that the majors were able to maintain their leading role in the industry and to weaken the role of the newcomers. During the first 10 years of its life, OPEC placed a limitation on the expansion of the free market, which was the source of price instability. Without OPEC and its measures during those early years, other changes taking place in the structure of the oil industry at that time would have further encouraged the newcomers, thereby increasing instability in the market.

In fact, the position of the major companies improved after the creation of OPEC, and to quote Penrose again, "The share of the major oil companies in the upstream part of the industry increased from 72% in 1960 to 76% in 1966. Similarly, in the downstream operations, their share increased from 53% to 61%, respectively." This meant that the shares of the newcomers had to decline from 28 to 24% in upstream activity and from 47 to 39% in downstream activity. In effect, OPEC had reversed the trend of increasing the share of the newcomers and limiting their power.

Although OPEC had failed to restore the posted prices to their levels prior to the 1959 and 1960 price cuts, as was the initial objective of the organization when it was founded in Baghdad, Iraq, in 1960, the formation of OPEC and the measures it took simply constituted a form of amendment of oil concessions. Prior to OPEC, host countries were not partners in their own industry and their role did not exceed the levying of taxes from the oil companies, the latter of which, according to the terms of the concessions, had the absolute freedom to decide whatever was suitable to their own interests in all matters relating to the oil industry in territory covered by concessions.

Another episode that would later play a pivotal role in changing the structure of the oil industry and the nature of the organization was OPEC's "declaratory statement of petroleum policy in member countries" in June 1966, which became a historical turning point for OPEC in that it laid down basic principles that would later influence all of the major events of the 1970s. The statement emphasized the right of producing countries to fix unilaterally the

price of their own oil in an endeavor to give the states a greater role in the development of hydrocarbon resources and to accord the states the right to participate in concession-holding agreements. The statement focused on amending the concession system in the light of changing circumstances.

The first 10 formative years of OPEC constituted a change in the relationship between the oil producers and the oil companies in a way that served both interests in the changing conditions of the industry. Although it is true that OPEC was still a dormant partner during those years, within 10 years this partnership had evolved into an active partnership, shifting from the function of simple tax collectors into that of active negotiating partners in determining prices.

3. A SHORT PERIOD OF COOPERATION BETWEEN PRODUCERS AND OIL COMPANIES

After these first 10 years, things took another turn. At the OPEC Conference in Caracas, Venezuela, in December 1970, it was decided to enter collectively into negotiations with the oil companies with a view to raising the price of oil and the tax ratio. Negotiations with the oil companies were first conducted in Tehran, Iran, by delegates from the six OPEC member countries bordering the Gulf: Iran, Iraq, Kuwait, Saudi Arabia, Qatar, and the United Arab Emirates. After several rounds of negotiations, they succeeded in taking a firm collective position to raise the fixed price by approximately $0.35 and increasing the tax ratio from 50 to 55%.

The Tehran Agreement was concluded in mid-February 1971 but was applied retrospectively from January 1 of that year. It was meant to be for 5 years, after which it was to be renegotiated. In addition to price and tax amendments, the Tehran Agreement provided for the adjustment of the price upward (on an annual basis) to reflect the impact of world inflation, together with a fixed annual increase. Later, and following the two successive devaluations of the U.S. dollar in 1972 and 1973, two agreements were negotiated with the companies and signed by the same signatories of the Tehran Agreement to adjust oil prices and, thus, to preserve the purchasing power of the other currencies relative to the U.S. dollar. The fact that all prices are denominated in U.S. dollars means that any depreciation of the value of the U.S. dollar could lead to a loss in real value of oil revenues when used to pay OPEC countries' imports from non-dollar countries. The dollar negotiations took place in Geneva, Switzerland, and the Geneva Agreements (known as Geneva 1 and Geneva 2) laid down the formulas with which to correct the oil price (agreed to in the 1971 Tehran Agreement), whether upward or downward, in relation to the movement of the value of the U.S. dollar vis-à-vis other major currencies (U.K. sterling; French, Swiss, and Belgian francs; German Deutschemark; Italian lira; and Japanese yen).

4. THE OIL PRICE SHOCKS

However, after $2\frac{1}{2}$ years of price stability and cooperation between OPEC and the oil companies, the history of OPEC and the oil industry took a dramatic turn, unleashing huge price volatility. In June1973, OPEC decided to reopen negotiations with the companies to revise upward the Tehran Agreement price in the light of prevailing market conditions. By that time, market prices had reached such high levels that the Tehran Agreement price was left lagging behind, and OPEC argued that the additional windfalls gained by the companies ought to have been shared with the producers. Accordingly, negotiations took place in Vienna, Austria, in early October 1973 but failed as the oil companies refused OPEC's demand for an increase in the price set by the Tehran Agreement. This provoked OPEC into fixing the price of its oil unilaterally and independently of the oil companies' participation. At a meeting in Kuwait on October 16, 1973, that was attended by the same OPEC countries bordering the Gulf, a decision was announced to increase the oil price by 70% so that the OPEC price was fixed at $5.40/barrel.

At the time of this decision, war was raging between Egypt and Syria, on the one hand, and in the state of Israel, on the other. When Egyptian forces succeeded in crossing the Suez Canal into occupied Sinai, the reaction of the United States and The Netherlands was to rally to the side of Israel. This led to mounting pressure from Arab public opinion to use Arab oil as a "political weapon" to coerce the United States into changing its hard-line policy of continued support for Israel and culminated with the Arab member countries of OPEC meeting to decide to place an oil embargo on both the United States and The Netherlands and to undertake successive cuts in production. It is little wonder that, following that fateful decision, the oil market prices

soared—unchecked and uncontrolled. Meanwhile, the Shah of Iran began a campaign within OPEC to raise the price substantially. At this time, the shah, having enjoyed the support of the West, appointed himself as the "policeman of the Gulf," a role that he saw as justifying his need to raise huge sums of cash with which to arm his country to the teeth. The reality was that the shah wished to control the area and present Iran as the most dominant power in the whole of the Gulf region.

This drove the Shah of Iran at first to put pressure on the oil companies to raise production until he realized that the most effective means of achieving this was by simply pushing for ever higher prices. He couched this in terms of "a new discourse of oil": "a noble commodity that should not be burnt" but instead should be reserved for "higher, nobler purposes such as petrochemicals."

By the time that OPEC's ordinary (winter) meeting was convened in Tehran in December 1973, market prices had already exceeded $15/barrel as a result of the Arabs' so-called oil embargo. The Shah of Iran took every advantage of this and succeeded in increasing the price from $5.40 to $10.84/barrel (or an increase of 140%). Relative to the Tehran Agreement and the oil prices during the summer of 1973, this constituted a 400% price increase, causing an enormous shock across the market. Various measures were taken by the consumer nations to slow down oil consumption such as limiting vehicular speed and implementing other energy-saving measures.

The vast majority of OPEC members tended to emulate the Shah of Iran in his quest for price escalation, but without fully comprehending the long-term reaction of such a sudden dramatic increase in oil prices and the negative impact on world demand for OPEC oil. For their part, the Saudis, aware of the economic implications and concerned about the consequential negative impact of price escalation on their own oil, were worried about such sharp oil price increases. One who was particularly worried was Sheikh Ahmed Zaki Yamani, the then Saudi oil minister, who consistently expressed his opposition to any further price increase. However, in OPEC price discussions during the late 1970s, especially during the oil market developments in the wake of the Iranian Revolution, the Saudi kingdom became more amenable to political pressures exerted by the others and, hence, finally acquiesced to most of the price decisions.

The price shock proved to be an inducement for greater efficiency in fuel use so that the amount of gasoline used per mile driven was in continuous decline. Coupled with this and consumer countries' fiscal policies, which involved dramatically increasing taxes on gasoline and other petroleum products (especially in Europe), oil became an extremely expensive commodity for end consumers. The resultant dramatic slowdown in the growth of oil consumption indicated how the impact of the price shock took its toll on Western Europe to a far greater extent than on the rest of the world. During the 1960s and early 1970s, oil consumption in Western Europe had grown at an average rate of more than 8% per year, but after the price shock, oil consumption fell from a peak of 15.2 million barrels/day in 1973 to 13.5 million barrels/day 2 years later, settling at approximately 14 million barrels/day thereafter. Conversely, the impact of higher prices on oil consumption in the United States and Japan was far less than in Western Europe, mainly due to those two countries' far lower level of taxation on oil products.

With a lead time, the price shock heralded a process of structural change in the world oil industry. Significantly, the emerging energy-saving campaigns led to diminishing oil consumption while achieving the same level of economic growth. Prior to the price shock, an increase in gross domestic product (GDP) had entailed an equal increase in oil consumption, but after the shock, the relationship between economic growth and oil consumption changed. A process of gradual "decoupling" of oil consumption from economic growth would not be attenuated later with the advent of even higher oil prices; that is, less oil consumption per unit of GDP became a permanent feature.

No sooner had the world started to live with the high oil prices than another, even greater price shock occurred in June 1979, the effect of Ayatollah Khomeini's revolution in Iran and, prior to that, the all-out strike of Iran's petroleum industry, causing a sudden interruption of more than 4 million barrels/day of Iranian oil and consequent market chaos with unparalleled high prices. Although the disappearance of Iranian oil from the market was soon compensated for by the increase in production from Saudi Arabia, Kuwait, and Iraq, the oil market continued to be turbulent.

OPEC's initial reaction to the rocketing prices on the world market showed some degree of rationality. Because market prices were soaring well beyond the OPEC price, and so were generating great windfalls for the oil companies, the organization (assuming that the situation was only temporary and that production could be increased from other Gulf countries to compensate for the disappearing Iranian oil)

considered adding temporary market premia, intended to be abolished once the market had returned to normal conditions. However, the continued raging markets caused many OPEC members to clamor for changing the initially rational decision and to instead add the premia increase to OPEC's fixed price on a permanent (rather than temporary) basis.

By the time OPEC met in Caracas in December 1979, OPEC's official prices had soared to more than $24/barrel, and with the markets still in turmoil, OPEC's members continued to add price increments, so that when OPEC met in Bali, Indonesia, a year later, the majority decided to fix the price of OPEC oil at $36/barrel as of January 1981. However, Saudi Arabia, aware of the negative effects of higher prices on oil consumption and the world economy, refused to raise the price beyond $32/barrel. Thus, OPEC arrived at a situation whereby a two-tiered pricing system prevailed.

The second oil price shock had an even greater negative impact on world oil consumption, and the soaring oil price served only to promote further world oil supplies from outside OPEC as well as alternative sources of energy. An ineluctable backlash lay in store for OPEC when the oil price across the world market began to fall, led by non-OPEC producers, in particular the North Sea.

5. OPEC AS THE LAST RESORT OIL SUPPLIER: THE QUOTA SYSTEM

OPEC had to either stick with its official price by reducing its output or else follow the market and reduce its price to retain its share of the world oil market. OPEC chose the first option in resorting to the "quota" system, based on the concept of its member countries, as a whole, producing oil in volumes designated only to fill the gap between, on the one hand, total world demand and, on the other, world supplies from outside OPEC; that is, OPEC chose to be the last resort producer or what was termed the "swing producer." This meant that at a certain demand level, the higher the non-OPEC supplies, the lower the OPEC total quotas. Given increasing oil supplies from outside OPEC in the face of weakening world demand, the organization's market share would correspondingly fall naturally and continuously.

Although first discussed in 1982, the OPEC quota system was formally adopted in March 1983 at an extraordinary meeting held in London that set a production ceiling of 17.5 million barrels/day, reducing the OPEC price to a unified $28/barrel. This system proved to be even more anomalous given that Saudi Arabia did not have a quota share and was left to be the "swing" producer within a range of 5 million barrels/day. OPEC decided to fix its price and take a reference price based on the crude of Arab Light API 34 f.o.b. Ras Tanura, which Saudi Arabia became committed to charging the oil companies. Meanwhile, the prices of other crudes were to be fixed upward or downward of that reference price, taking into account the quality of the crude and its geographical location in Saudi Arabia. In other words, OPEC now became the swing producer, bearing the burden of price defense at the expense of its world market share; whereas Saudi Arabia bore the brunt by becoming the swing producer within OPEC, absorbing this fall in world demand for OPEC oil far more than did the other OPEC members. This created a situation in which the 12 member countries of OPEC were producing practically the amount of oil commensurate with their allocated quotas, whereas Saudi Arabia was forced to swing downward relative to the fall in the call on total OPEC production, seeing its production in 1985 fall to approximately 2.5 million barrels/day, or roughly one-quarter of its production capacity (leaving three-quarters unproduced). The fall in Saudi Arabia's production was so great that it caused a problem of insufficiency of associated gas for the generation of electricity for water desalination plants.

With this continued decline in the call on its oil, OPEC saw its own market share fall from 62% during the mid-1970s to 37% in 1985, all as a result of increasing supplies from outside OPEC coupled with the fall in demand as a result of high prices and improving energy efficiency in industrial nations. For example, North Sea production (both British and Norwegian) leaped to 3.5 million barrels/day in 1985, up from 2.0 million barrels/day in 1975, and continued to rise thereafter. The reason behind this was that OPEC's high prices provided such a wide profit margin for oil investors that high-cost areas such as the North Sea became not only feasible but highly lucrative areas for continued reinvestment. Because OPEC adhered to the system of fixed price and swing production, any additional oil coming from outside OPEC would first capture its share in the market before buyers resorted to OPEC oil. Also, the greater the supplies of non-OPEC oil, the less OPEC oil that was on the market to meet world demand.

Furthermore, the two oil price shocks triggered enormous investments in scientific research to

improve the efficiency of technology in finding, developing, and producing oil and, thus, in reducing the high cost of upstream operations in these new areas, thereby adding for the oil companies an even higher margin of profit with which to reinvest in high-cost areas.

By 1985, OPEC production was indicative of the backlash suffered by OPEC at the hands of its own price policies. In less than 6 years, OPEC's total production had fallen from 31 million barrels/day to roughly half that amount, and this led to a situation that was increasingly difficult to sustain. Consequently, things got out of hand to the point where Saudi Arabia decided to drop its system of selling oil at fixed prices in accordance with the 1973 quota system and instead adopted market-oriented pricing, or "net-back value," meaning that it obtained the price of its crude from prevailing prices of oil products in the main consuming areas minus the cost of refining, transporting, and handling oil. Saudi Arabia's production started to rise quickly, and by the summer of 1986, a "free-for-all" situation prevailed.

6. THE BACKLASH

This led to a severe collapse in the oil price so that during the summer of 1986 Arab Light was selling at less than $8/barrel. However, this price collapse generated mounting pressure on Saudi Arabia to forgo its refusal to adopt the quota system. The pressure came first from the OPEC members that had seen their revenues fall so drastically. A large pressure group formed within OPEC (led by Iran, Algeria, Libya, and Venezuela) that was politically powerful enough to campaign against Saudi Arabia, which was supported (within OPEC) only by Kuwait. Meanwhile, pressure came from external quarters when the high-cost producers in the United States were forced to shut down their expensive oil wells, and Texas, where oil is the spur behind all economic activity, sank into a deep recession. Investments in high-cost North Sea oil were similarly jeopardized by the drastic fall in oil prices. This additional external pressure finally pushed Saudi Arabia into changing its policy and agreeing to readopt the quota system with a fixed price for its oil at $18/barrel.

In January 1987, OPEC changed its pricing system of taking Saudi Arabia's Arab Light as a reference price for other OPEC crudes and replaced it with a "basket" of seven crudes as a reference price: Sahara Blend of Algeria, Dubai Crude, Saudi Arabia's Arab Light, Minas of Indonesia, Bonny Light of Nigeria, Tia Inana of Venezuela, and Isthmus of Mexico. However, by 1988, OPEC had effectively abandoned the fixed price system, substituting it with an agreement on target prices so that supplies could be regulated to maintain that target. This new system saved OPEC from the headache of having to agree to a fixed price for all other crudes in the light of their quality and geographical locations. However, OPEC has kept to its system of quotas and a total ceiling corresponding to the difference between total global demand and world oil supplies outside OPEC. There is no doubt that this new pricing system made decisions on the OPEC price more flexible than the earlier system, which had caused several litigious problems and endless discussions among member countries about the appropriate price for each crude relative to the reference price.

In deciding to set the price of oil, OPEC had taken on the heavy burden of oil price management with little success in achieving price stability. Yet this had been the original objective on which the organization, when founded, predicated itself. Instead, when OPEC took over oil pricing control in 1973, enormous price volatility ensued, as is clearly shown in Fig. 1. For example, the price of Arab Light crude skyrocketed from $2 to $5/barrel during the summer of 1973 to approximately $40/barrel only 7 years later and then plummeted to roughly $8/barrel 5 years after that.

7. OPEC'S LIMITATIONS IN OIL SUPPLY MANAGEMENT

In OPEC's management of oil supplies, its limitations were to prove to be serious, with the first and foremost drawback being the intense politicization of the organization that became especially evident once OPEC took over the pricing of oil, at total variance with the original objective when the organization was formed. OPEC oil became a political commodity subject to political pressures and maneuvers. In general, politicians seek immediate or short-term benefits with little regard for the long-term consequences of their decisions. The politicization of OPEC first became evident when the Shah of Iran sought to escalate prices without due regard for the impact of higher prices on world oil demand and the supplies from outside the organization. The heads of state of OPEC's members became involved in the

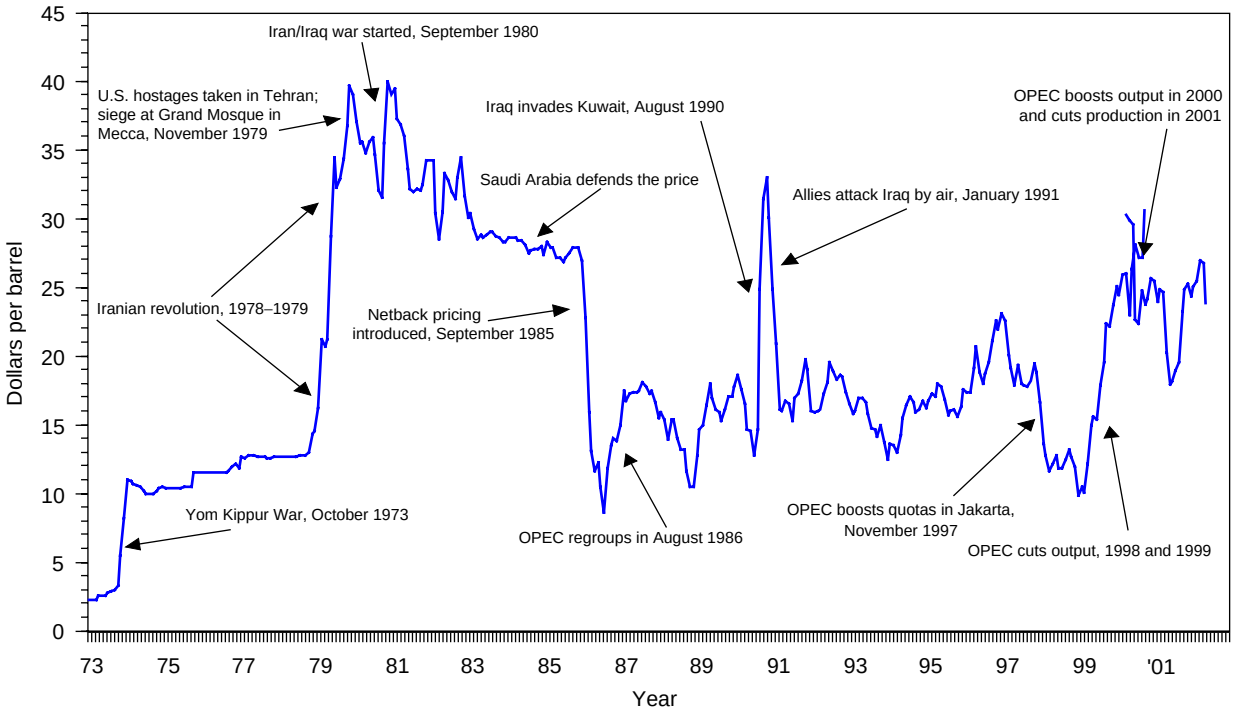


FIGURE 1 Arab Light Spot Prices, 1973–2002 (dollars/barrel). Data from British Petroleum, Center for Global Energy Studies, and Middle East Economic Survey.

organization's oil policies without due regard for each member country's relative importance in terms of oil reserves, in a way as to allow low-reserve countries (e.g., Algeria, Libya) to put political pressure on their high-reserve fellow members to comply with their demands. Thus, OPEC's decision-making process had little to do with sound economics.

OPEC's second limitation is that its decisions on the oil price are pushed by political events that invariably lead to an interruption of oil supplies. Figure 1 shows how the wide fluctuations in the oil price have been caused primarily by political events rather than by market forces. World oil markets can be seen as simply reacting to the interruption of supplies caused by political events. As shown in the figure, the first price shock was the result of the political decision of the Arab oil embargo in 1973, and the second price shock was caused by the Iranian Revolution in 1979. Prices then rose steeply again during the Iraqi invasion of Kuwait in 1990. In 2003, the war with Iraq, followed by lack of security of Iraqi oil pipelines and supplies, once again posed a renewed threat of oil shortfall, forcing market prices upward.

A third and major limitation placed on OPEC's role as an instrument for price stabilization lies in the disintegration of the organization's crude oil production within the international oil industry. This began when OPEC first wrested control of the oil industry from the international oil companies, creating a predicament quite contrary to the situation that prevailed when the oil companies controlled the upstream in response to downstream requirements, all of which were controlled by the major oil companies in a way as to create a stable market by avoiding a shortage or surplus of crude oil. In OPEC's case, decisions on crude oil production are not organically related to downstream requirements. OPEC production is based simply on the difference between world demand and the production from outside the organization, without any knowledge of the downstream exigencies of consumer countries. This kind of disintegration has inevitably led to price volatility in the world oil market.

OPEC's fourth limitation is that OPEC members are entirely dependent on their oil revenues to meet their foreign currency requirements. This dependence, which amounts to more than 90% of their external trade, means that OPEC's decisions are, in reality, dictated by the short-term financial requirements of its member countries. Moreover, during the 30 years or so since the first price shock, OPEC countries have done practically nothing to diversify their economies and reduce their dependence on oil. These characteristics continue to subject OPEC's

decisions to the pressing short-term needs of its member countries' budgets without looking into any form of economic rationale.

Fifth among OPEC's limitations as an efficient oil supply manager is the frequent failure of some member countries to abide by the quota system. It is noticeable that there is always a difference between the actual OPEC production quotas and the official ones. The degree of quota compliance within OPEC is reinforced when oil prices are very low, affecting its member countries' budgets; conversely, the degree of compliance falls off when oil prices are high.

Perhaps an even more significant limitation to OPEC's decision making is the organization's heterogeneous nature. Not having a homogenous entity has proved to be a major weakness because there are enormous differences among the member countries. For example, those with low oil reserves always seek higher prices so as to maximize their oil revenues by increasing the per barrel income; also, because their production capacity is so limited, they do not even care about the market share or the long-term effects of high prices on demand and supply. Conversely, member countries with large reserves (e.g., Saudi Arabia) in principle have regard for their market share to maximize income from their larger volume and, thus, higher market share. OPEC has in the past tried to formulate a long-term strategy but has never succeeded due to these conflicts of interest within the organization.

8. OPEC AND A FUTURE OF UNCERTAINTY

What future awaits OPEC? So far, OPEC has managed to survive many crises, even those of the 1980s (Table I). Regardless of the problems facing the organization's quota system, OPEC continues to function. Despite differences in the long-term interests of its members, what unites and motivates OPEC, enabling it to reach unanimous decisions, is short-term oil revenue and financial gain. However, its power has been eroded gradually over time as a result of its own policies. OPEC price practices, which amount to deriving short-term gain at the expense of long-term benefit to the oil industry as a whole, have altered the entire world energy structure and made oil far less dominant in its share of energy consumption, to the benefit of natural gas and nuclear power. This is especially evident in the cases of both Western Europe and Japan. Table II shows the extent to which oil has been losing its share in total energy consumption. Prior to OPEC's price shocks, oil had accounted for 62% of Western Europe's total consumption, yet by the year 2000, oil's share had fallen to 42%. Meanwhile, natural gas's share of total energy increased from 10 to 22%, and nuclear power's share increased from 1 to 13%, during the same period. The shift is more dramatic in the case of Japan, where oil's share of total energy has diminished from 77 to 50%. Meanwhile, its share of natural gas increased from 2 to 13%, and its share of nuclear power increased from 1 to 14%, during the same period.

In other words, OPEC's price shocks triggered a process in the industrialized world of gradually shifting from oil to alternative energies. This shift will accelerate in the future for reasons related to both environment and technological development. Although the United States still refuses to ratify the Kyoto Protocol (the international agreement that aims at a drastic reduction in carbon dioxide emissions through policy measures to reduce the consumption of fossil fuels), there is mounting pressure in both Europe and Japan to comply and, thus, reduce the consumption of oil in favor of more environmentally friendly resources, in particular natural gas, nuclear power, and (in the future) renewable energy resources. Meanwhile, technology favors these shifts because it reduces the costs of otherwise expensive alternatives. For example, the hybrid vehicle consumes much less gasoline than does the standard combustion engine, and eventually fuel cell-driven vehicles will further shrink the demand for oil.

Moreover, OPEC's high price policies have led to heavy investment in oil resources in high-cost areas outside OPEC while technological progress in the upstream has reduced the costs of finding and developing oil, thereby making investments in these alternative high-cost areas very lucrative. The amount of oil that has been produced outside OPEC has been increasing enormously. Excluding the former Soviet Union, non-OPEC oil supplies increased from 18,904 million barrels/day in 1973 to 35,569 million barrels/day in 2000. After 1982, when OPEC had opted to be the last resort supplier, the increase in non-OPEC oil supplies has been at the expense of OPEC's market share, the decline in which has occurred specifically during the same period. The less its market share, the more difficult it is to manage oil supplies at a certain stabilized price level. In addition to the price effects, geopolitics has played a significant role in changing the map of world oil supplies.

TABLE I

OPEC: A Chronology of Events

1960 (September 16)	In Baghdad, Iraq, the foundation of OPEC was announced by the five major oil-producing countries (Iran, Iraq, Kuwait, Saudi Arabia, and Venezuela)
1961	Qatar joined OPEC, followed by Indonesia and Libya (1963), the Emirates of Abu Dhabi (1968), Algeria (1969), Nigeria (1971), Ecuador (1973), and Gabon (1975). The latter two members left OPEC during the 1990s
1966 (June)	OPEC issued a "declaratory statement of petroleum policy in member countries," a historical turning point, as it laid down basic principles that would later influence all major events of the 1970s. The statement emphasized the right of producing countries to fix unilaterally the price of their oil, in an endeavor to give the states a greater role in the development of hydrocarbon resources and to give the states the right to participate in concession-holding agreements. The emphasis was on amending the concession system in the light of changing circumstances
1970 (December)	In Caracas, Venezuela, the OPEC Conference was held. A decision was made to negotiate collectively with the oil companies to establish a general increase in posted prices and to raise the tax ratio to a minimum of $55
1971 (February)	In Tehran, Iran, the Tehran Agreement (to be of 5 years' duration and then renegotiated) was the outcome of this first round of negotiations with the oil companies, conducted by delegates from the six OPEC member countries bordering the Gulf (Iran, Iraq, Kuwait, Saudi Arabia, Qatar, and the United Arab Emirates). In addition to price and tax amendments, the Tehran Agreement provided for the adjustment of the price upward (on an annual basis) to reflect the impact of world inflation, together with a fixed annual increase
1972 and 1973	Two successive devaluations of the U.S. dollar led to two further agreements (in Geneva, Switzerland) to adjust prices and preserve the purchasing power of the other currencies relative to the U.S. dollar. Known as the Geneva Agreements ("Geneva 1" and "Geneva 2"), these two agreements laid down the formulas with which to correct the oil price (agreed in the 1971 Tehran Agreement), whether upward or downward, in relation to the movement of the value of the U.S. dollar relative to other major currencies (U.K. sterling; French, Swiss, and Belgian francs; German Deutschemark; Italian lira; and Japanese yen)
1973 (June)	OPEC decided to reopen negotiations with the companies to revise upward the Tehran Agreement oil price in the light of prevailing market conditions, where oil prices had reached such high levels that the Tehran price was left lagging behind, and OPEC also argued that additional windfalls gained by the companies ought to have been shared with the producers
1973 (July)	In Vienna, Austria, OPEC negotiations with the oil companies failed, with the latter refusing OPEC's demand for an increase in the price set by the Tehran Agreement. This provoked OPEC into fixing the price of its oil unilaterally and independently of the oil companies' participation
1973 (October 16)	In Kuwait, the same OPEC Gulf members decided to increase the oil price by 70% so that the OPEC price was fixed at $5.40/barrel
1973 (October)	The Arab oil embargo was implemented
1973 (December)	In Tehran, OPEC's ordinary (winter) meeting was held. Spurred on by the Arab oil embargo, the Shah of Iran pushed for increasing the price yet further by 140%, from $5.40 to $10.84/barrel, a 400% increase relative to the Tehran Agreement
1979	During early part of year, an all-out petroleum industry strike took place in Iran, forcing oil prices upward to an unparalleled level
1979 (June)	The Ayatollah Khomeini's revolution and political coup occurred in Tehran
1979 (December)	In Caracas, with OPEC's official prices having soared to more than $24/barrel, OPEC members continued with price increments
1981 (January)	In Bali, Indonesia, the majority of OPEC members decided to fix the price of OPEC oil at $36/barrel. Saudi Arabia refused to raise the price beyond $32/barrel, giving rise to a two-tiered pricing system
1983 (March)	In London, an extraordinary OPEC meeting was held. Instead of substantially reducing its price to retain its share of the world oil market, OPEC stuck with its official price by reducing its output through a "quota" system. OPEC opted to be the last resort supplier with a production ceiling of 17.5 million barrels/day. The official OPEC price was now set at a unified $28/barrel
1985	The extent of the petroleum backlash was evident when OPEC total production had already fallen by half from 31 million barrels/day
1986	A "free-for-all" situation prevailed with a rapid rise in Saudi Arabia's production, culminating in an oil price collapse in July, with Arab Light selling at less than $8/barrel
1986	Pressure groups form within OPEC and among other countries to push Saudi Arabia into readopting the quota system with a fixed price for its oil at $18/barrel

continues

Table I continued

1987 (January)	OPEC changed its pricing system (of taking Arab Light as a reference price for other OPEC crudes) and replaced it with a "basket" of seven crudes as a reference price (Sahara Blend of Algeria, Dubai Crude, Saudi Arabia's Arab Light, Minas of Indonesia, Bonny Light of Nigeria, Tia Inana of Venezuela, and Isthmus of Mexico). However, by 1988, OPEC had effectively abandoned the fixed price system, substituting it with an agreement on "target prices" so that supplies could be regulated to maintain that target
1990 (August)	Iraq invaded Kuwait. Oil prices rose steeply. UN sanctions were imposed on Iraq
1991 (January)	The Gulf War with Iraq took place
1997 (September)	The Kyoto Protocol compelled signatory nations to set a deadline for reducing carbon dioxide emissions to diminish global warming. This favored the adoption of alternative cleaner energies
2003 (March 20–April 17)	The second Gulf War involving Iraq took place

TABLE II

Percentage Shares of Total Energy Consumption 1973–2001

	Western Europe			Japan			North America		
	1973	1985	2001	1973	1985	2001	1973	1985	2001
Oil	62	45	42	77	57	48	45	41	39
Natural gas	10	16	22	2	10	14	30	26	25
Coal	20	25	14	15	20	20	18	24	23
Nuclear	1	10	13	1	10	14	1	6	8
Hydro	7	3	8	5	2	4	6	3	5

Source. Data from British Petroleum.

The most critical issue currently is the question of the security of Gulf oil supplies, with this area being considered prone to political instability, a factor affecting oil supplies and prices. This geopolitical factor had added momentum to greater investment in oil supplies from outside the Gulf. The understanding reached in 2002 between the United States and the Russian Federation indicates how U.S. policy is oriented toward reducing dependence on the Gulf in favor of Russian and Caspian Sea oil as well as toward obtaining increased oil supplies from West Africa (mainly Nigeria, Angola, and Equatorial Guinea) and from offshore oil production in the Gulf of Mexico.

An even greater contributory factor to the weakening of OPEC's control over oil prices is the huge investment that some member countries (e.g., Nigeria, Algeria) have designated for themselves to increase their production capacities. Once these expansion programs come on-stream, it is very difficult to perceive that OPEC can continue to subject these countries to the restrictions of its production quota regime for the sake of shoring up the oil price.

Perhaps even more significant will be the question of Iraq, a country extremely rich in oil deposits that have been discovered but have remained undeveloped and that could provide a huge increase in that country's production capacity. It has been estimated that, given favorable political and financial conditions, Iraqi production could exceed 6 million barrels/day and could even reach 8 million barrels/day. It is obvious that the organization would be unable to restrict Iraqi production after the country's dire economy has been so severely held back since the Gulf War and UN sanctions.

From the preceding picture, the prospects for the future of OPEC may appear grim. However, in the event of the organization's collapse, oil prices would immediately collapse across the world markets, affecting all investments in high-cost areas outside OPEC. Such a scenario would renew the demand for cheaper OPEC oil, and for geopolitical reasons, this would not be favored by the developed nations. For this reason, it is very difficult to make any accurate prediction concerning the future of OPEC and the political and economic forces that may reshape the future status of the organization.

The fall of Saddam's Ba'athist regime in Iraq will have far-reaching effects on OPEC and on the oil industry in general. Iraq's oil potential is huge and could ultimately reach production capacity levels near those of Saudi Arabia. During the 1970s, many important giant oilfields were discovered but remained undeveloped due to Saddam's wars and the UN sanctions. In a matter of a few years, it would be technically feasible to reach a capacity of more than 6 million barrels/day. But meanwhile, the current administration is primarily concerned with restoring production and exports from the existing oilfields together with plans to rehabilitate the oil sector.

The first phase of developing Iraqi oil is to restore Iraq's capacity to its 1990 pre-UN sanctions level of 3.5 million barrels/day. Even with huge investments, this may take at least 2 years. It is not yet clear what policy would be adopted for the next phase, which is the expansion of Iraq's oil capacity to 6 million barrels/day or even to 8 million barrels/day. It goes without saying that foreign investment in Iraq would not be possible without first establishing a credible and stable Iraqi government with a sound legal system that would safeguard the interests of foreign investors.

SEE ALSO THE FOLLOWING ARTICLES

Economic Geography of Energy • *Geopolitics of Energy* • *Global Energy Use: Status and Trends* • *International Energy Law and Policy* • *National Energy Policy: United States* • *Nationalism and Oil* • *National Security and Energy* • *Oil Crises, Historical Perspective* • *Oil-Led Development: Social, Political, and Economic Consequences* • *Oil Industry, History of*

Further Reading

Amoco, B. P. (1970–2001). "Annual statistical reviews." – BP Amoco, London.

Center for Global Energy Studies. (1990–2002). "Global Oil Reports." CGES, London.

Chalabi, F. J. (1980). "OPEC and the International Oil Industry: A Changing Structure." Oxford University Press, Oxford, UK.

Chalabi, F. J. (1989). "OPEC at the Crossroads." Pergamon, Oxford, UK.

Penrose, E. (1968). "The Large International Firm in Developing Countries: The International Petroleum Industry." Allen & Unwin, London.

Sampson, A. (1975). "The Seven Sisters." Hodder & Stroughton, London.

OPEC Market Behavior, 1973–2003

A. F. ALHAJJI
Ohio Northern University
Ada, Ohio, United States

1. Introduction
2. Overview
3. Oligopoly Models
4. Competitive Models
5. OPEC and the Cartel Status
6. Conclusion

Glossary

cartel A group of producers that collude to reduce production and increase prices in order to maximize the wealth of the group.
monopoly A market with a single producer that sells a product having no close substitute.
nationalization The host government takeover of foreign companies' operations.
netback pricing A method of determining the wellhead price of oil and gas by deducting from the downstream price, the destination price, the transportation price, and other charges that arise between wellhead and point of sale.
oligopoly A market with few competing producers.
OPEC basket price The average price of seven marker crudes.
OPEC quota A specified maximum amount of oil production assigned by OPEC to each of its members.
Organization of Petroleum Exporting Countries (OPEC) Historically, it consisted of 13 countries: Saudi Arabia, Kuwait, Iraq, Iran, Qatar, Abu Dhabi, Libya, Algeria, Nigeria, Venezuela, Indonesia, Gabon, and Ecuador. Ecuador and Gabon withdrew from the organization in 1992 and 1995, respectively.
posted price A price of a barrel of oil that the host nations used as a base to collect taxes from the international oil companies, regardless of the market price. In most cases, posted prices were higher than market prices by a wide margin. After nationalization, countries decided to make it the market price.
price band The minimum and maximum for the average OPEC basket price, set by OPEC in 2000.
Seven Sisters Refers to the seven international oil companies that dominated the world oil industry in the 1930s and 1940s: Exxon, British Petroleum, Shell, Gulf, Texaco, Socal, and Mobil.
user's cost The cost of replacing oil that has already been produced. It is a concept used only with nonrenewable resources.

Two supply factors are known to influence oil prices and dominate research activities: scarcity and market power. Scarcity increases prices above marginal cost by the discount rate, even in a competitive market. In noncompetitive markets, prices are usually higher than marginal cost. However, events of the past 80 years indicate that the market power of the Texas Railroad Commission, international oil companies, and Organization of Petroleum Exporting Countries (OPEC) has had a more significant impact on oil prices than scarcity. For this reason, most researchers focus on OPEC when they discuss world oil markets of the past 30 years. Nevertheless, this focus on market power without regard to depletion and technical issues has polarized academic and political circles to the extent that it is almost impossible to reach a consensus on OPEC and its role in world oil markets. It is also difficult to discuss OPEC behavior without putting analyses in their historical context. It is necessary to understand the political, economical, social, and philosophical factors that led to the establishment of OPEC and its role today. Oil is not only a nonrenewable resource but also a political and strategic commodity.

1. INTRODUCTION

Some researchers argue that oil is a special commodity that requires intervention and depletion management.

The oil industry is characterized by booms or busts, which calls for additional controls not provided by market forces. This extra control seeks to avoid or minimize the boom-or-bust cycle. Standard Oil, the Texas Railroad Commission, the Seven Sisters, and the Organization of Petroleum Exporting Countries (OPEC) have all exercised it with varying degree of success. OPEC has been the least successful and failed to manage depletion. Some researchers even argue that U.S. embargos on some of the oil-producing countries are a form of oil market management to support oil prices. OPEC does not have grand strategies and plans. It pragmatically reacts to various political, economical, and natural factors that influence its members, but the impact of these factors varies among members at any given time. Thus, OPEC must reconcile the divergent interests of its membership through meetings and negotiations.

The focus of this article is OPEC market behavior. Therefore, only market models introduced in the past 30 years are discussed. An overview of the many oil market models developed from 1973 to present shows that these models divide easily into oligopoly models and competitive models. Most researchers use oligopoly models to explain OPEC behavior and assume that OPEC, or a subset of it, is a cartel that seeks monopoly profits by changing production and influencing prices. Others argue that the world oil market is competitive and that factors other than cartelization better explain oil price increases. For each side, there are various models that explain OPEC behavior.

A few researchers argue that OPEC is a wealth maximizer. Most researchers argue that OPEC functions not to maximize wealth but to maximize revenues and achieve other social and political purposes. These contrasting views resulted from the emphasis of most researchers on market power rather than on market failure and depletion management.

Section 2 presents an overview of OPEC and non-OPEC production, world oil demand, oil prices, and OPEC's actions. Section 3 reviews and critiques the oligopoly models of OPEC. The competitive models are presented in Section 4. Section 5 provides various points of view regarding the application of the term cartel to OPEC.

2. OVERVIEW

Figures 1 and 2 show oil prices, OPEC production, non-OPEC production, and world demand for oil between 1973 and the first quarter of 2003. Major

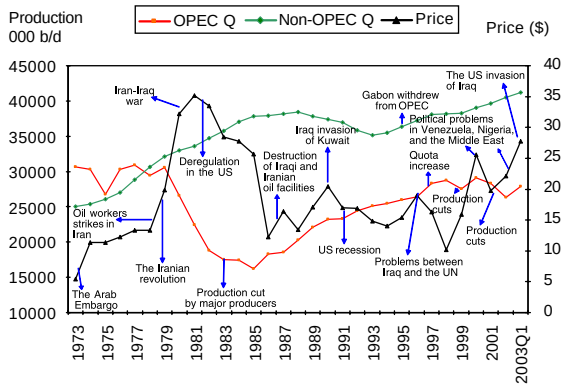


FIGURE 1 Oil prices, OPEC and non-OPEC production (thousands b/d), and major events, 1973–2003. Data from *Monthly Energy Review*, Energy Information Administration, July 2003. Data for 2003 are for the first quarter.

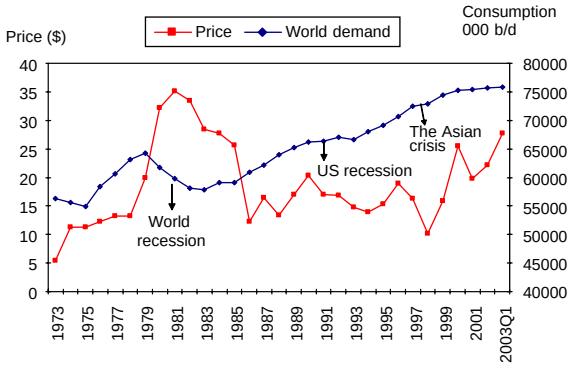


FIGURE 2 Oil prices and world demand (thousands b/d), 1973–2003. Data From "BP Statistical Review of World Energy" (2003) and Energy Information Administration (2003). Data for 2003 are estimates for the first quarter.

political events and some OPEC actions are plotted in Fig. 1 to illustrate their impact on prices and oil production. OPEC as an organization did not cause energy crises, oil price spikes, and price volatility. However, various political incidents that involved individual OPEC members did contribute to energy crises and market instability. Researchers provide various models and explanations for such changes in production and prices. From the information in Figs. 1 and 2, the period between 1973 and 2003 can be divided into six phases: 1973–1978, 1979–1981, 1982–1986, 1987–1991, 1992–1997, and 1998 to the present.

2.1 The First Oil Crisis: 1973–1978

Oil prices increased in 1973 when, on October 19, several Arab countries—all of them allies of the

United States—imposed an embargo on the United States and The Netherlands for their support of Israel in the 1973 war. Although the embargo led to a mere 5% decrease in world oil supply and lasted only 2 months, oil prices increased fourfold. Theory predicts that oil prices will return to their original level when an embargo is lifted, but prices in 1974 remained high. World oil demand decreased at first because of substitution, conservation, and lower economic growth. However, it continued its upward trend once the economies of the consuming countries absorbed the impact of the energy shock.

This production cut by some OPEC members and the subsequent spike in prices led some researchers and observers to conclude that OPEC had become a powerful cartel. They point out that non-OPEC members behaved differently and increased production to take advantage of the higher oil prices. Others believe that different factors led to the production cut and that price increases were not related to cartelization, including changes in property rights, the limited capacity of the economies of OPEC members to absorb productive investment, environmental concerns about natural gas flaring by some OPEC members, panic and stockpiling in the consuming countries, and U.S. wellhead price control. This last factor kept U.S. oil prices below the world price, decreased U.S. oil production, increased U.S. demand for oil, and increased oil imports.

2.2 The Second Oil Crisis: 1979–1981

Oil prices increased again in 1979 as a result of the Iranian revolution. Only Saudi Arabia and some non-OPEC members increased their production to compensate for the lost Iranian oil. Prices continued to rise when the Iraq–Iran war started in September 1980. Iraqi and Iranian oil production came to a complete halt by the end of 1980 and the world lost 9 million barrels per day (mb/d). Coupled with panic and stockpiling, spot oil prices reached historic levels. However, OPEC contract prices were much lower, especially those of Saudi Arabia. These prices were set by OPEC members, not the market.

Conservation, substitution, and a worldwide recession brought a substantial decline in world oil demand in 1980 and 1981. Meanwhile, non-OPEC production continued to increase. As a result, OPEC output and market share decreased. Two major factors affected oil prices and OPEC behavior in 1981: price deregulation in the United States and price agreement among OPEC members. The agreement set price differentials for each OPEC crude relative to the Saudi marker. This agreement indicated that Saudi Arabia was not willing to defend prices alone. It wanted other members to share this burden. The agreement ensured that if the price of the Saudi marker declined, the price of other member's crudes would decline too. Therefore, countries had to cooperate with Saudi Arabia and cut output to preserve prices. This was the first serious coordinated attempt by OPEC to influence world oil markets. Market management requires price control, output control, or both. OPEC chose the first method in 1981. By the time OPEC agreed on prices, a decrease in demand led Nigeria unilaterally to cut its prices by $5.50/b. First Iran and then other OPEC members followed. As a result, the price agreement collapsed.

2.3 The Oil Glut: 1982–1986

A decline in world oil demand and an increase in non-OPEC production led to a sharp decrease in OPEC market share between 1982 and 1986. The fundamentals of world oil markets changed drastically to the extent that massive production cuts by some OPEC members were not enough to prevent an oil price collapse, as shown in Fig. 1.

The breakdown of the first price agreement in 1981 led OPEC to consider another method of market control: production management. For the first time, OPEC established a loose quota system in March 1982. In this agreement, Saudi Arabia was not assigned a quota with the understanding that it would behave as a swing producer to balance the world oil market. In addition, Iraq was allowed to produce more than its quota. Many countries were dissatisfied with the system, which broke down almost immediately. OPEC members violated their quota by more than 2 mb/d by the end of 1982.

The 1983–1986 period witnessed a sharp decrease in OPEC market share, especially that of Saudi Arabia; an increase in cheating by OPEC members; a decline in world demand; and an increase in non-OPEC production. Consequently, OPEC was forced to reduce its posted marker prices and lower its production ceiling. Saudi Arabia continued to defend oil prices by cutting production until its output declined to only 2 mb/d, an amount insufficient to sustain the Saudi economy. For the first time since the establishment of the quota system, OPEC assigned Saudi Arabia a quota of 4.35 mb/d in order to protect Saudi Arabia's market share. This figure was less than half of what Saudi Arabia was producing in 1980. In addition, OPEC abandoned the posted price

mechanism. It abandoned Saudi Arab Light as the marker crude in favor of market-oriented netback prices. Prices were no longer administered by OPEC through the posted prices mechanism. OPEC decided to take the market price and produce accordingly. As a result of these developments, the price plummeted to single digits in August 1986.

This collapse reversed historical trends of increased substitution and conservation. It stimulated the world demand for oil, as shown in Fig. 2. It also decreased non-OPEC production by shutting down stripper wells in the United States and other countries.

2.4 Calm before the Storm: 1987–1991

Four factors brought about increased oil prices in 1987: major damage to Iraqi and Iranian oil facilities, less cheating by OPEC members, reduction in non-OPEC output, and an increase in oil demand. Netback pricing was abandoned by OPEC members by 1988 in favor of a sophisticated pricing formula that linked OPEC crude prices to market prices.

Oil prices improved by 1989, and OPEC regained market share because of a decline in non-OPEC production and an increase in demand. Prices increased to more than $40/b when Iraq invaded Kuwait in August 1990. The world lost 4.5 mb/d, but Saudi Arabia, along with other oil-producing countries, was able to compensate for the loss of Iraqi and Kuwaiti oil. Although oil prices remained relatively high due to the Gulf War, the decline in world demand and the U.S. postwar recession resulted in reduced oil prices in 1991. A United Nations (UN) embargo on Iraqi oil went into effect in 1991. Iraqi oil did not find its way into world markets legally until 1996. The UN lifted the embargo in April 2003 after the U.S. invasion of Iraq.

2.5 Post Gulf War: 1992–1997

A combination of factors caused oil prices to plunge in 1993, including the expected return of Iraqi exports to world oil markets, surging North Sea output, and weak demand. At the end of 1997, OPEC decided to increase its quota by 10% for the first half of 1998 on the erroneous assumption that Asian oil demand would continue to grow. Unfortunately, the Asian financial crisis of 1997 proved OPEC wrong. Oil prices began to decrease. OPEC ignored many signals at that time. For example, the increase in oil prices in 1996 was caused not only by an increase in Asian demand but also by the delay in signing the oil-for-food program between Iraq and the UN.

2.6 Recent Developments: 1998–Present

Despite very strong economic growth in the United States and many European countries in 1998, oil prices continued to decrease until they reached single digits in the first quarter of 1999. Analysts cited four reasons for this decline: the Asian financial crisis, the increase in OPEC and non-OPEC production, the increase in Iraqi exports, and two consecutive mild winters in the consuming countries.

OPEC efforts to halt the 1998 decline in oil prices failed despite two announced cuts and the promise of output cuts by some non-OPEC members. The collapse of oil prices in the first quarter of 1999 forced OPEC and non-OPEC members to announce production cuts by 2.104 mb/d in March 1999. OPEC decided to continue the cuts when it met in September of that year, and oil prices continued to rise, as shown in Fig. 1.

The sharp increase in oil prices in 2000 forced OPEC to increase its quotas in March, June, September, and October. However, these increased quotas did not translate into actual production. Many OPEC members lacked the spare capacity. Three factors characterized 2000: frequent meetings, frequent quota increases, and the unofficial creation of a price band. In October, OPEC activated the unofficial price band for the first time and increased production by 500,000 b/d.

OPEC set up a price band mechanism during its March 2000 meeting. According to the mechanism, OPEC would increase production automatically by 500,000 b/d if the average OPEC basket price remained above $28/b for 20 consecutive trading days. Production would be cut by a similar amount if the basket price remained below $22/b for 10 consecutive trading days. The automatic adjustment was later abandoned so that OPEC members could adjust production at their discretion. The creation of the price band is OPEC's fourth serious attempt to manage world oil markets. At first, OPEC tried to control prices through creating price differentials for each OPEC crude in 1981. The failure of this pricing mechanism led OPEC to adopt a quota system to control output in 1982. The failure of this move led OPEC to adopt netback pricing in 1985. Theoretically, netback pricing with the quota system in place should have enabled OPEC to control both prices and outputs. However, netback pricing, a market-oriented pricing mechanism, annulled the quota system, which is a form of market management. OPEC failed to control the market for the third time. Creating the price band may have enabled OPEC for

the first time to use both outputs and prices to manage world oil markets.

The sharp increase in oil prices in 1999 and 2000 led to lower demand and higher non-OPEC production in 2001. Consequently, oil prices declined and OPEC was forced to cut production twice—in January by 1.5 mb/d and in March by 1 mb/d. The demand for oil declined again at the end of 2001 after the September 11 terrorist attack on the United States. The decline in air travel and lower economic growth were the primary reasons for the decline in world oil demand. Consequently, OPEC agreed in November 2001 to cut production by 1.5 mb/d effective January 1, 2002, but only if non-OPEC producers cut their output by 500,000 b/d as well. With non-OPEC cooperation, the cut was implemented, although data indicate that non-OPEC cooperation may have been merely symbolic.

Despite lower world economic growth in 2002 and higher world oil production, political problems pushed oil prices higher. Iraq halted its official oil exports in April for 1 month to protest the Israeli incursion into the West Bank. Meanwhile, an attempted coup forced Venezuelan President Hugo Chavez to resign on April 12, only to resume his presidency on April 14.

Oil prices had been increasing since May 2002. A general fear that the war with Iraq would disrupt the Middle Eastern oil supplies initiated the increase. Labor strikes in Venezuela that halted oil exports forced prices to increase to more than $30/b in December 2002. Both Iraq and Venezuela are major oil producers and OPEC members.

OPEC met in December 2002 and decided to take unusual action—to cut oil production by increasing quotas and cutting production at the same time. OPEC members were violating their quota by 3 mb/d. OPEC decided to increase quotas by 1.3 mb/d and cut production by 1.7 mb/d. When prices continued to increase, OPEC met a month later and increased production by 1.5 mb/d.

Oil prices continued to increase in 2003. The possibility of a U.S. invasion of Iraq and political tensions in several oil-producing countries kept prices high despite the substantial increase in OPEC production and low economic growth in the consuming countries. Analysts attributed the increase in prices to the "war premium," which is the difference between market prices and what prices would have been without the threat of war in Iraq.

Contrary to expectations, oil prices started to decline when President Bush issued an ultimatum to the president of Iraq, Saddam Hussein. He gave him 48 h to leave Iraq or face war. Prices continued to decline during the war. Several factors contributed to this decline. First, the war premium disappeared as traders became certain and optimistic about the outcome of the war. Second, several countries started selling their oil that was stored in the Caribbean and on large tankers in the hope that prices would be higher during the war. Third, several countries increased their oil production, especially Saudi Arabia, as shown in Fig. 3. Fourth, traders were convinced that the Bush administration would use the Strategic Petroleum Reserves if prices increased. Fifth, traders expected Iraq to start producing in large quantities very soon after the end of the war.

Oil prices did not decline substantially as many experts expected. The delay in Iraq exports and lower inventories kept prices at approximately $30/b in the United States. Despite talk of production cuts, only Saudi Arabia cut production by 200,000 b/d by May, whereas most members increased production.

3. OLIGOPOLY MODELS

Production by non-OPEC members, which comprises approximately 60% of world production, eliminates the monopoly models for OPEC. Monopoly in the literature refers to a single producer controlling the whole market. Hence, the focus is on oligopoly models, in which the world oil market is shared between a dominant producer and the competitive fringe. Most models that assign market power to OPEC or to some subset of its members are categorized as one of three models: OPEC as a dominant producer, the OPEC core as a dominant producer, or Saudi Arabia as a dominant producer. Although OPEC literature frequently contains expressions such as "monopoly," "monopoly power," and "monopoly profit," researchers are actually endorsing oligopoly models.

3.1 OPEC as a Dominant Producer

This model assumes that OPEC is a monolithic cartel in which countries have unified goals and collectively set the price for oil. Non-OPEC oil producers are the competitive fringe. The demand for OPEC's oil is the residual demand, the difference between world demand and non-OPEC supply. OPEC sets the price where its marginal revenue equals its marginal user's cost. The competitive fringe equates that price with their marginal user's cost and supplies accordingly; OPEC supplies the rest.

The model assumes that OPEC became a powerful cartel in 1973. OPEC was able to cut production, increase prices, and transfer wealth from the oil-consuming countries to the oil-producing countries.

This model may explain changes in OPEC and non-OPEC production, world demand, and changes in oil prices from 1973 to the present. As shown in Fig. 1, OPEC initially cut production and prices continued to increase between 1973 and 1981. The higher price was a self-defeating strategy for OPEC. It decreased demand through conservation, substitution, and fuel utilization. In addition, it increased non-OPEC supplies. The remaining residual demand decreased. OPEC lost its market share in the early 1980s. Consequently, oil prices and OPEC oil export revenues declined.

The lower price in the mid-1980s stimulated world demand and slowed the growth of non-OPEC production. Such development increased the residual demand for OPEC's oil and increased oil prices. The same situation explains the behavior of world oil markets from the 1990s to the present.

3.1.1 Criticism of the Model

This model does not apply to OPEC for the following reasons:

1. OPEC does not meet the characteristics of a cartel as stated in the economic literature.
2. The model assumes that production decisions are made by OPEC authority and not by the countries. However, several examples show that countries have acted unilaterally.
3. Historical evidence indicates that production cuts in the 1970s were related to the deterioration of technology after nationalization in several countries, such as Libya and Venezuela.
4. Statistical tests do not support this model.

3.2 The OPEC Core as a Dominant Producer

The failure of the previous model to fit OPEC behavior led researchers to introduce more sophisticated models. To be more realistic, researchers acknowledge that OPEC has a market power and its members have different interests. Three models have been introduced: the two-part cartel, the three-part cartel, and the core as monolithic cartel.

3.2.1 Two-Part Cartel Model

This model assumes that OPEC is a nonuniform cartel. OPEC is divided into two groups—saver and spender countries. The saver countries are the United Arab Emirates (UAE), Iraq, Kuwait, Libya, Qatar, and Saudi Arabia. The spender countries are Algeria, Ecuador, Indonesia, Iran, Nigeria, and Venezuela. The two groups can compete or collude with each other. In the case of competition, the saver countries will act as a dominant producer and set the price that maximizes profit. The spender countries will act as a competitive fringe. In the case of collusion, both groups will form a monolithic cartel. The outcome will be similar to the outcome in the monolithic cartel model discussed previously.

Applying the model to Figs. 1 and 2, the saver countries that set the world price made production cuts, whereas spender countries continued to increase their production. Once the quota was implemented in 1982, the saver countries cut production, whereas the spender countries violated their quotas. As world demand decreased and non-OPEC production increased, the core's market share shrank. Later, a trend reversal led to an increase in prices and an increase in the core's share. A similar explanation can be used to illustrate changes in output and prices up to the present.

3.2.2 Three-Part Cartel

OPEC is divided into three groups:

1. The cartel core countries: Kuwait, Libya, Qatar, Saudi Arabia, and the UAE. These countries have vast oil reserves, small populations, and flexible economic development. The core acts as the dominant firm and sets the price of oil. The cartel core nations carry excess capacity in order to maintain prices during emergency demand.
2. The price-maximizing countries: Algeria, Iran, and Venezuela. These countries have relatively low oil reserves and large populations with potential for economic development. These countries aim for higher prices by cutting production.
3. The output-maximizing countries: Ecuador, Gabon, Indonesia, Iraq, and Nigeria. They have limited reserves, large populations, and a pressing need for economic development. These countries sell at any price, even in a weak market.

Applying this model to Figs. 1 and 2 explains OPEC behavior in a way similar to that of the previous model. However, this model shows that some countries pushed for higher prices when prices declined. This situation may explain the reversal of oil prices in 1987, despite the fact that Saudi Arabia had more than doubled its production by that time.

3.2.3 Core as a Monolithic Cartel

This model differs from the previous two models in that it does not assume that OPEC is a cartel, but that the core countries behave as a cartel that sets the price while all other producers, including the rest of OPEC, behave as a competitive fringe. The choice of core members (Kuwait, Qatar, Saudi Arabia, and the UAE) is not based on reserves and population but on economic and political factors. This model may explain OPEC behavior in the same manner as before, especially in the OPEC dominant producer model.

3.2.4 Criticism of the Model

This model does not match actual OPEC behavior for the following reasons:

1. The division of OPEC into two and three groups was subjective. Researchers did not base the division on theoretical, behavioral, or statistical analyses. They used factors that are not directly related to production decisions, such as reserves and population. Theoretically, production capacity, not reserves, has an impact on production. Statistically, various tests indicate that neither reserves nor population has a direct impact on production levels and prices.
2. Historical data do not support this model. Countries within each group did not act in unison, and some countries switched positions occasionally.
3. According to the three-part cartel model, the price-maximizing countries cut production with higher oil prices. At the same time, the core increases its production to offset the reduction, and vice versa in the case of low prices. Historical data do not support such conclusions.
4. For the core to be a cartel, it must operate as a dominant producer. Statistical tests do not support such behavior.
5. Given the sheer size of Saudi production relative to the production of other core members, it is very difficult to distinguish between the impact of the core and the impact of Saudi Arabia.

3.3 Saudi Arabia as the Dominant Producer

This model assumes that OPEC is not a cartel. Saudi Arabia acts as the dominant producer. All other producers behave as a competitive fringe. Saudi Arabia sets the price. All other producers expand their output to the point at which that price equals their marginal user's cost. Saudi Arabia then supplies the rest to meet the world demand.

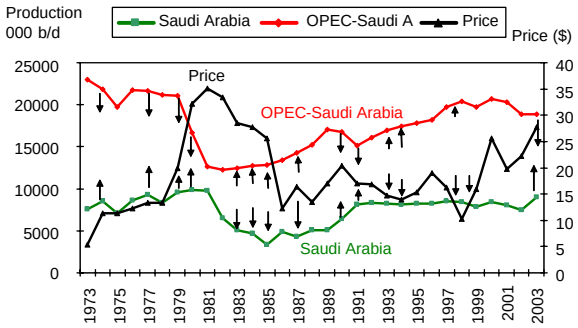


FIGURE 3 Saudi Arabia and other OPEC countries production (thousands b/d) and oil prices. Arrow direction indicates that Saudi Arabia's actions differ from those of other OPEC members. Source: Energy Information Administration (2003, January).

This model could be the most plausible for explaining OPEC's behavior in the past 30 years. It gained support over the years even from researchers who first concluded that OPEC was a cartel. Looking at Figs. 1 and 2, this model can explain the changes in the world oil market in a similar way as the OPEC as a dominant producer model. The only difference is that production cuts and increases are made by Saudi Arabia, not by OPEC.

There are several factors that lend support to this model. First, historical data indicate that only Saudi Arabia has large excess production capacity. It is the only country to voluntarily reduce capacity and production. The 1 mb/d increase in Saudi production before and during the invasion of Iraq led to a $9/b price decline from preinvasion levels. The price during the invasion of Iraq was at least $15/b less than the price that analysts expected.

Second, Saudi Arabia's output is negatively correlated with that of other OPEC members, as seen in Fig. 3. The opposite arrows point to the direction of change in production. For example, in 1998, Saudi production decreased (arrow is pointing down) and the production of the rest of OPEC members increased (arrow is pointing up).

Third, Saudi Arabia was the only OPEC member not assigned a quota when OPEC implemented the 1982 quota system, so Saudi Arabia could act as "swing producer" to stabilize the world oil market.

Finally, statistical tests indicate that Saudi Arabia has operated on the elastic part of its demand curve since 1973. Operation on the elastic part of the demand curve is a requisite for the dominant producer model to apply.

3.3.1 Criticism of the Model

This model does not apply to OPEC for the following reasons:

1. The model assumes that the behavior of other OPEC members is similar to that of non-OPEC members. Historical data do not support this assumption.
2. The model does not explain why the Saudis pressed especially hard to convince non-OPEC members, such as Mexico, Norway, and Russia, to cooperate in production cuts. However, political factors may explain such behavior.
3. Figure 3 indicates that the model may not fit after 1999. Data show concerted efforts between Saudi Arabia and other OPEC members to cut production. Although Fig. 3 shows that Saudi Arabia increased production while the rest of OPEC decreased production in the first quarter of 2003, the decrease occurred because of the U.S. invasion of Iraq. Data indicate that all remaining OPEC members increased production during this period.
4. Saudi political objectives contradict this model. Political scientists who try to explain OPEC behavior claim that Saudi Arabia usually deviates from its dominant producer position in order to achieve political goals. These models assume that political interaction among OPEC members is important for cartel stability. These models are based on the idea that wealth maximization, security, and political influence are to some degree substitutes for each other. Thus, the decision to choose, for example, more security necessarily implies less wealth. In this case, political models and wealth maximization models offer divergent interpretations and predictions.

4. COMPETITIVE MODELS

The lack of support for various oligopoly models led researchers to introduce a variety of models that assume a competitive world oil market. Based on this view, five different models are presented to explain OPEC behavior.

4.1 Competitive Model

Oil prices increased in 1973 not only because of the embargo but also because of panic, stockpiling, and a spot market in which refiners sold futures crude oil. Rather than refining it, they used the crude as a commodity to gain quick profit.

Panic, stockpiling, spot markets, and speculation decreased supply and increased demand, which resulted in very high oil prices. In the same manner, oil prices increased in 1979, 1990, 1996, 2000, 2002, and 2003: A political problem led to a series of actions that lowered supply and increased demand. According to this model, there was no role for OPEC as an organization in the output disruption.

Applying this concept to Figs. 1 and 2, the boycott of 1973 shifted the supply curve to the left and increased oil prices. A domino effect followed: Panic and speculation shifted the demand to the right, refiners sold the crude and shifted supply farther to the left, and prices continued to increase. The Iranian revolution and the Iran–Iraq war caused the same chain reaction, which led to the second oil crisis.

In the early 1980s, the demand decreased and world supply increased, which led to lower oil prices. Increased demand and lower world supply, mostly from non-OPEC countries, resulted in increased oil prices in the second half of the 1980s. In the same manner, this model explains changes in prices and production in 1990, 1996, 1998, and 2000–2003.

4.1.1 Criticism of the Model

This model does not apply to OPEC for the following reasons:

1. In a competitive market, output will increase with higher prices. This is not the case for OPEC. It either cut or maintained output when prices were increasing, as shown in Fig. 1.
2. The model does not explain why some OPEC members, mainly Saudi Arabia, cut production in the mid-1980s while non-OPEC countries increased production.
3. It does not explain why the behavior of Saudi Arabia is different from that of other countries.
4. Statistical tests reject this model.

4.2 Property Rights Model

Production cuts by some OPEC members in the early 1970s were associated with changes in property rights when the host countries claimed ownership of the oil reserves either through nationalization or gradual ownership of the operating international oil companies (IOCs). This change in ownership led to uncoordinated production cuts. The host countries' discount rates were lower than those of the IOCs. In other words, foreign companies wanted to get as much money as possible from their concessions

before they left, which resulted in increased production. When the host countries took over, they wanted to produce less in order to conserve the resource for future generations and to maximize efficient rate of production per oil field.

Applying this model to Figs. 1 and 2, one sees that 1973–1974 was a transition period. It reflected a sharp increase in government ownership during which a sharp decrease in production caused a switch to a higher price path. Even during the 1979 price shock, some researchers argued that production cutbacks were the result of ownership transfers from companies to host countries during that period.

Results of statistical tests show "mixed performance" for OPEC members. The property rights model cannot be rejected for Iran, Kuwait, Libya, Nigeria, Qatar, and Venezuela. The model was rejected for the rest of OPEC members.

Researchers who introduced this model acknowledge that production cuts led to a panic in the market and an increase in demand, which increased oil prices even further. They also acknowledge that the limited absorptive capacity of OPEC members sustained prices after the transfer of ownership. As shown for the next model, this limitation led to additional production cuts.

4.2.1 Criticism of the Model

This model does not apply to OPEC for the following reasons:

1. The model cannot explain changes in production and prices after property transfer ended in the late 1970s. However, one might argue that oil-producing countries changed their discount rates as a result of "property reevaluation" or a change in "demand expectations."
2. Some researchers claim that historical facts contradict this model. For example, production increases in the 1950s and 1960s may have resulted from the host countries' desire to increase production. However, this criticism ignores the fact that IOC payments to the host governments during that period were a function of production levels. Countries collected more revenues as production increased. Therefore, they pushed the IOCs to increase production.
3. Some researchers argue that the production cut was not related to a lower discount rate by the countries. Rather, lack of technology caused it. The IOCs were deploying the latest technology in the early 1970s to enhance production and take advantage of rising oil prices. After nationalization, state-owned companies were not able to deploy the advanced technology that had been used by the IOCs, which resulted in lower production, especially in old oil fields.

4.3 The Target Revenue Model

This model was introduced to explain why oil prices were sustained at their postembargo levels despite the official lifting of the embargo in March 1974. According to this model, political events and changes in property rights in 1973 caused prices to increase. Oil prices did not decline after the end of the embargo because countries, without coordination or collusion, cut their production further as prices increased.

Most OPEC members had primitive economies that could not absorb the additional revenues from higher oil prices. Countries had two options: transfer the extra revenues to foreign financial markets or keep the extra dollars in cash for future use. For three reasons, OPEC members chose to do neither. First, in the 1970s, real returns on investments in the West were very low, if not negative. Second, some members feared asset confiscation by Western governments, especially after the freezing of Iranian assets by the United States. Third, dollar devaluation and inflation would have reduced the value of any cash revenues saved. These reasons led some oil-producing countries to cut production. They believed that oil in the ground was "the world's best investment."

The result of output cuts is a backward-bending supply curve for prices higher than those needed to support the target revenue. If there are a sufficient number of countries with backward-bending supply curves, then the market supply curve will take the same shape.

Most researchers agree that this model may explain OPEC behavior in the short term but only in some periods during the 1970s. Recent research indicates that, even today, the model applies only to centrally planned and isolated oil-producing countries, regardless of their OPEC membership. Statistical tests show that this model applies only to two OPEC members, Libya and Iran. Both countries are centrally planned and isolated by UN and U.S. embargoes.

4.3.1 Criticism of the Model

This model does not apply to OPEC for the following reasons:

1. The backward supply curve can hold only for a very short period of time. Prices will continue to move along the supply curve.

2. Some researchers argue that the model is not stable. If we have a backward supply curve, an increase in the negatively sloped demand will lead to two points of equilibrium with different production levels and different prices. However, some researchers argue that the equilibrium is stable. Producers in this case will choose the lower production level to conserve the resource.

3. Historically, some researchers rejected this model because of its assumption that foreign investment is not a viable alternative to domestic investment. They cited Kuwait's success in building a financial empire in the West. They also cited data illustrating that every OPEC member has an overseas investment. This objective ignored the possibility that political reasons may prevent some countries from acquiring access to foreign financial markets.

4. All statistical tests in the literature reject this model for OPEC as a whole. However, it cannot be rejected for some countries, such as Iran and Libya.

4.4 The Purchasing Power Explanation

Although not a formal model, it explains why some OPEC members behave competitively, violate their quota, and decline to coordinate production with Saudi Arabia. According to this explanation, some OPEC members increase production and violate their quota if the purchasing power of their oil exports is decreasing as a result of variations in major world currency exchange rates and higher export prices of industrial countries.

If oil production cuts lead to higher oil prices and higher export prices for the products of industrial countries, it is not in the interest of the member to participate in production cuts. The purchasing power of its oil exports would decline. Although this concept may apply to all imports of goods and services, the situation is more severe if the oil-producing country is importing large amounts of petroleum products. If an OPEC member must import petroleum products, production cuts will increase the cost of imports and force the real value of oil exports to decrease.

Although Indonesia is an OPEC member, its large population and relatively small reserves force it to import crude oil and other refined products. Imports of crude and refined products represent approximately 8% of its total imports. During the fourfold increase in oil prices in the 1970s and the threefold increase in 1980, the value of these imports increased substantially. As a consequence, the real value of Indonesian oil exports decreased. This decline may have forced Indonesia to increase its oil production and ignore its OPEC quotas in order to increase its purchasing power. Data indicate that Indonesia did not coordinate production with other OPEC members. Studies indicate that the importation of crude oil and other oil-intensive products was the primary reason for the decrease in the purchasing power of Indonesia's oil exports.

Historically, seven OPEC members imported crude oil and petroleum products on a scale large enough to affect their purchasing power: Algeria, Ecuador, Gabon, Indonesia, Nigeria, the UAE, and Venezuela. These imports may explain the competitive behavior of these countries.

4.5 The Foreign Investment Explanation

This theory was introduced to explain the behavior of Kuwait. It assumes that OPEC is not a cartel and attempts to explain why Kuwait does not coordinate production with Saudi Arabia, as indicated by the production data of both countries.

Kuwait's behavior can be explained by its foreign investments. The amount of foreign investment does not play a role in oil production decisions. The type of investment does. At first, Kuwait invested in upstream oil operations in the North Sea. At that time, it was in Kuwait's interest to cut production or sustain it to obtain higher prices for its oil in Kuwait and the North Sea. Not until the 1980s did Kuwait extend its investment to downstream operations to include thousands of gas stations throughout Europe and refineries in Singapore.

Since governments control retail prices of petroleum products in Europe, it was in Kuwait's interest to violate its OPEC quota and increase production. Lower crude oil prices from increased production enabled Kuwait to generate a large profit margin from retail outlets, where product prices did not change. The profit from lower crude oil prices outweighed the loss from selling the Kuwaiti crude at a lower price. In addition, Kuwait's return on investment exceeded its revenues from crude exports throughout the 1980s.

5. OPEC AND THE CARTEL STATUS

Despite the widespread use of the word *cartel* by media outlets and politicians, most researchers today, even those who assigned cartel status to OPEC in the 1970s, believe that OPEC is not a "monolithic" cartel. However, events since 1999 show that OPEC's

behavior may have changed now that a price band is in place. It holds almost monthly meetings instead of the usual quarterly meeting and makes frequent output adjustments. Figure 3 indicates that since 1999, production adjustment by Saudi Arabia matched that of other OPEC members. Some researchers believe that the recent changes in OPEC behavior constitute a de facto cartel. Others argue that Saudi Arabia is still the dominant player that initiates most output changes. They point out that production cuts by OPEC and non-OPEC members were not voluntary but rather the result of technical, political, and natural factors. The arguments of each group are summarized next.

5.1 OPEC as a Cartel

Those who argue that OPEC is a cartel cite the following reasons:

1. Historical records indicate that OPEC was modeled after the Texas Railroad Commission. Even today, its objective is to coordinate production among members.
2. Since its inception, OPEC members have been meeting periodically to coordinate policies, production, and prices.
3. Historically, OPEC cut production to increase prices, as indicated in Fig. 1.
4. OPEC currently uses a quota system and a price band, two important cartel characteristics.
5. The price of oil is much higher than marginal cost.
6. OPEC members have been able to transfer massive amount of wealth from the consuming countries since 1973.

5.2 OPEC as a "Commodity Producer Club" But Not a Cartel

The cartel theory states that there are seven characteristics that must exist in a group of producers in order to be labeled a cartel: A cartel must assign quotas to its members, monitor members to avoid violations, punish violators, target a minimum price, take action to defend the price, have a large market share, and divide votes among members based on their market share. OPEC did not meet any of these characteristics during its peak of power between 1973 and 1981. OPEC assigned quotas in 1982 but failed to enforce them. OPEC did not set a price band until 2000. It might have been able to defend prices since 2001 with the help of technical, political, and natural factors that prevented many oil-producing countries from increasing production. OPEC never had a punishment mechanism or an effective monitoring system such as that of the Texas Railroad Commission, which used the National Guard and Texas Rangers to enforce its rules and cut production in Texas and Oklahoma. In addition, OPEC's market share is relatively small (approximately 40%). Even today, each OPEC member has one vote regardless of its production and reserves. Unlike the Seven Sisters, OPEC did not divide the world market and did not control investment. The market share of the Seven Sisters was more than double that of OPEC.

By U.S. standards, OPEC does not fit the description of a cartel. Monopolization has two elements: the acquisition of monopoly position and the intent to monopolize and exclude rivals. Official U.S. judicial records indicate that a monopoly position has been associated with companies whose market share approached or exceeded 80%. This is double the current OPEC market share and much larger than the 54% market share achieved in 1974. OPEC has thus failed, by a wide margin, to acquire a monopoly position. As for intent to acquire a monopoly and to exclude rivals, OPEC has never taken the steps mentioned previously to be a successful cartel. OPEC was labeled a cartel because it supposedly increased prices, but how could OPEC exclude rivals by increasing oil prices? Figure 1 shows that higher oil prices bring higher production, more producers, and more competition. OPEC did not act as a monopolist. OPEC has increased production many times in recent years to stabilize prices. It has become an active participant in the world community through negotiations with consuming countries and active aid programs to developing countries. Many researchers today, including lawyers and policy makers, view OPEC as an agent of stabilization in a volatile world. In fact, the Bush administration praised OPEC for its efforts to prevent prices from increasing to record levels during the invasion of Iraq.

None of the statistical tests in the literature support the cartel model for OPEC. Researchers who conducted these tests concluded that OPEC is a "partial," "weak," "clumsy," or "loose" cartel. Such conclusions are mere journalistic expressions. No known theory in the economic literature supports them. The results of so-called statistical tests are not acceptable. For example, applying the same tests to non-OPEC countries shows that non-OPEC producers fit the

cartel model better than OPEC members. In addition, these tests focus on parallel behavior, not on cartelization. Parallel behavior may exist in any market, including competitive ones.

The cartel model requires OPEC to operate on the elastic portion of its demand curve. Studies show that OPEC operated on the inelastic portion of its demand curve during its power peak in the 1970s.

Coordination among governments and commodity producer clubs is well-known. This coordination is not labeled as cartelization, despite periodic meetings and clear goals to curb production and maintain prices.

Researchers who attempt to prove that OPEC is a cartel by focusing on the difference between marginal cost and the price of oil ignore the basic principles of natural resource economics. The price of a nonrenewable resource in a competitive market reflects the marginal cost and the discount rate chosen by the producer. Therefore, in a perfectly competitive market the price of any natural resource is higher than its marginal cost, unlike renewable resources, the price of which equals the marginal cost. Hence, focusing on the difference between marginal cost and price to prove market power is conceptually wrong unless the researcher has perfect estimates of the user's cost.

Researchers who believe that OPEC is not a cartel argue that there are other explanations for OPEC wealth transfer that are not related to cartelization. Some researchers argue that it was both the market power of Saudi Arabia and U.S. price controls that led to wealth transfer to OPEC. Various studies show that U.S. oil price controls suppressed U.S. production, increased the world demand for oil, and raised Saudi Arabia's output. This unintended effect enabled OPEC members, mainly Saudi Arabia, which increased its production substantially in the late 1970s, to transfer wealth from the consuming countries to the producing countries.

Historical records indicate that OPEC was reacting to economic and political events rather than being a proactive cartel that consistently followed polices that would maximize its wealth. OPEC reacts pragmatically to compromise on the conflicting interests of its members. In addition, OPEC pricing has always followed the spot market.

6. CONCLUSION

There are two prominent supply factors that explain the behavior of OPEC: scarcity and market power. The focus of most researchers on market power without regard to scarcity and technical issues and their political and social implications has polarized academic and political circles to the extent that it is almost impossible to reach a consensus on OPEC and its role in world oil markets.

Review of various models introduced to explain OPEC behavior indicates the following unequivocal conclusions:

1. Significant production by non-OPEC members rules out the use of monopoly models.
2. Historical data and research studies indicate that the competitive model does not apply to OPEC.
3. Only oligopoly models may explain OPEC behavior. However, theory, historical data, and statistical tests do not support the monolithic cartel model for OPEC.
4. No single model can explain OPEC behavior even at particular periods of time.

One way to explain the behavior of OPEC and non-OPEC producers, world oil demand, and oil prices is to synthesize these models based on the results of various studies. Such a synthesis indicated that OPEC is driven by political and technical influences as much as by the inexorable forces of economics and the market.

This synthesis must tell the story of the past 30 years without violating logic, theory, data, or history. The behavior of OPEC can be explained using the dominant producer model for Saudi Arabia or a dynamic core that changes members periodically and by using various models to explain the behavior of other OPEC members. In this case, other OPEC members follow various models at various points of time, such as the property rights model, the target revenue model, the purchasing power explanation, the foreign investment explanation, and even pure competitive behavior. At different times, OPEC, or a subset of it, may look like a monolithic cartel when, coincidentally, various models necessitate the same action. In this case, countries may cut or increase production jointly. For this reason, single-equation models, correlation tests, and cointegration tests indicate parallel action but not cartelization.

However, one should not forget the role of political and natural events, panic, stockpiling, and speculation and their impact on oil markets. OPEC, like the Texas Railroad Commission, was established to protect property rights and to bring order to a chaotic situation that was leading to tremendous wastage and losses. OPEC aid to Third World

countries makes OPEC look more like a world-class intergovernmental organization than merely a cartel that cuts production and increases prices. Production increases by OPEC members to combat the increasing prices that resulted from the labor strikes in Venezuela and Nigeria and the war in Iraq indicate the significant role that OPEC plays in stabilizing oil prices, especially given that political turmoil, not OPEC production cuts, caused world energy crises.

The shape and role of OPEC may change in the future. Some members may leave the organization when they become net oil importers. Others may leave if they privatize their oil industry and become unable to force private companies to follow OPEC quotas. On the other hand, new members may join the organization. Some current OPEC members may increase their capacity substantially. Others may even build excess capacity and compete with Saudi Arabia. Regardless of these changes, OPEC will keep playing an important role in world oil markets. Lower oil prices and the withdrawal of some members will not lead to the demise of this powerful organization.

The impact of the growing importance of natural gas on OPEC remains to be seen. Four OPEC members are among the world's largest natural gas players: Iran, Algeria, Qatar, and Libya. The emphasis on natural gas may either weaken or strengthen OPEC. Only time will tell.

Acknowledgments

I thank Michael Bunter, Collin J. Campbell, Juan Pablo Perez Castillo, Melaku Geboye Desta, Dermot Gately, Luis Lugo, Adam Sieminski, Thomas Walde, and an anonymous referee for their helpful comments on the manuscript.

SEE ALSO THE FOLLOWING ARTICLES

Economic Geography of Energy • Geopolitics of Energy • Global Energy Use: Status and Trends • National Energy Policy: United States • Nationalism and Oil • National Security and Energy • Oil Crises, Historical Perspective • Oil Industry, History of • Oil-Led Development: Social, Political, and Economic Consequences • War and Energy

Further Reading

Adelman, M. A. (1995). "The Genie out of the Bottle: World oil since 1970." MIT Press, Cambridge, MA.

Alhajji, A. F., and Huettner, D. (2000a). OPEC & other commodity cartels: A comparison. *Energy Policy* **28**(15), 1151–1164.

Alhajji, A. F., and Huettner, D. (2000b). Crude oil markets from 1973 to 1994: Cartel, oligopoly or competitive? *Energy J.* **21**(3), 31–60.

Alhajji, A. F., and Huettner, D. (2000c). The Target revenue model and the world oil market: Empirical evidence from 1971 to 1994. *Energy J.* **21**(2), 121–144.

Crèmer, J., and Salhi-Isfahani, D. (1991). "Models of the Oil Market." Harwood, New York.

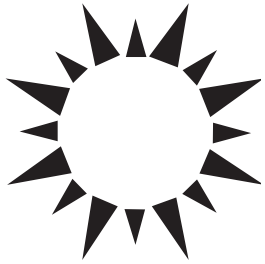
Griffin, J., and Teece, D. (1982). "OPEC Behavior and World Oil Prices." Allen & Unwin, Boston.

Mabro, R. (1998). OPEC behavior 1960–1998: A review of the literature. *J. Energy Literature* **4**(1), 3–26.

Noring, Ø. (2001). "Crude Power, Politics and the Oil Market." Tauris, London.

Stevens, P. (2000). "The Economics of Energy," Vols. 1 and 2. Elgar, Cheltenham, UK.

Origin of Life and Energy



RONALD F. FOX
Georgia Institute of Technology
Atlanta, Georgia, United States

1. **Stellar Nucleosynthesis of the Elements**
2. **Relative Abundance of the Elements**
3. **Gibbs Free Energies of Formation**
4. **The Monomer-to-Polymer Transition: No Life to Life**
5. **Energy: Redox Reactions, Thioesters, and Phosphate**
6. **Banded Iron Formations in the Geological Record**
7. **The Importance of D-Orbitals**
8. **Thermal Energy and Synthesis**
9. **The RNA World**

Glossary

dehydration condensates Polymers in which the monomer linkages are equivalent to the removal of a molecule of water between the joined monomers.

Gibbs free energy A type of thermodynamic energy containing internal energy, entropic energy, and pressure volume energy that is applicable for systems at constant temperature and pressure.

macromolecular Pertaining to molecules of large size having masses in the thousands to millions of Daltons (1 Dalton is 1 atomic mass unit). All biologically important macromolecules are polymers, molecules made from linking together monomers by means of dehydration linkages. There are three major types: proteins made from amino acid monomers, polysaccharides made from simple sugars, and polynucleotides made from mononucleotides.

micells Self-assembled monolayered closed structures made from fatty acids dissolved in water.

phosphate A molecule made from phosphorus, oxygen, and hydrogen that is an ion when dissolved in water with the formula HPO_4^{2-}.

protein biosynthesis The complex mechanism by which proteins are synthesized, amino acid by amino acid, in accord with a sequence of nucleic acid bases in a messenger RNA molecule that has been transcribed from a DNA gene. Many proteins and RNA molecules are involved in the apparatus that performs this function.

redox energy The energy of electrons transferred with them during electron transfer reactions.

stromatolites Layered fossil rocks believed to be the remains of mats of bacterial colonies topped and fed by phototrophic organisms, perhaps the ancestors of contemporary cyanobacteria.

thioester An energy-rich compound involving sulfur covalently linked to the carbon of a carbonyl group.

vesicle A bilayer, self-assembled, closed structure made from glycerophospholipids possessing two fatty acid side chains.

The phrase origin of life refers to those natural geophysical processes that may have occurred on primitive Earth approximately 3.5–4.0 billion years ago that gave rise to life. Presumably there are no extant representatives of the earliest forms of life since they were surely driven to extinction by more advanced forms very early in the evolution of life. Fossils do not preserve macromolecular structures inside cells so that no direct fossil evidence exists. Even putative cellular fossils from 3.5 billion years ago, the stromatolites, are not absolutely proven to be what they appear to be. On the other hand, their similarity to contemporary formations deposited by bacterial mats topped with cyanobacteria, which may not be too different from the primitive organisms of 3.5 billion years ago, suggests that the metabolic and protein biosynthesis macromolecular machinery of contemporary cells was already developed at that time. How these macromolecular structures came into being is the problem of the origin of life.

In this context, energy refers to a variety of forms, both chemical and physical, that are relevant in the context of the origin of life. These include (i) fusion energy released during the formation of the nuclei of elements during stellar nucleosynthesis; (ii) ultraviolet (UV) light energy of the sun that excites electrons into excited states, thereby providing them with redox energy, for example, during the oxidation of iron from ferrous iron (Fe^{2+}) to ferric iron (Fe^{3+});

 © 2004 Elsevier Inc. All rights reserved.

(iii) thioester compounds that are rich in Gibbs free energy so that they spontaneously engage in a variety of reactions and syntheses; (iv) phosphate compounds that serve to activate intermediates in the synthesis of polymers; and (v) thermal energy or heat. The problem for the origin of life and energy is to understand how these different forms of energy are connected and transduced from one kind to another and promote the origin and evolution of the macromolecular processes of living matter. In contemporary organisms there is also the fundamental importance of chemiosmotic proton energy, but this form of energy likely evolved relatively late in the origin of the first cells.

1. STELLAR NUCLEOSYNTHESIS OF THE ELEMENTS

Understanding why life is made out of the elements that are observed to be most abundantly present in organisms is aided by considering the origin of the elements during stellar nucleosynthesis. These elements are hydrogen (H), carbon (C), nitrogen (N), oxygen (O), phosphorous (P), sulfur (S), chlorine (Cl), sodium from natrium (Na), magnesium (Mg), potassium from kalium (K), calcium (Ca), and iron from ferrum (Fe). Although several other elements are essential for many contemporary organisms, such as iodine (I), selenium (Se), and cobalt (Co), primitive life almost certainly could have arisen based on the primordial dozen elements listed previously. These elements are the available building materials for life because they are the most abundant naturally occurring elements.

During the lifetime of a typical star, tremendous changes in pressure, mass density, and temperature take place inside the star. Temperatures between 10^7 and 10^9 K occur as well as densities from 10^2 to 10^5 g/cm^3 (ordinary liquid water has a density of 1.0 g/cm^3). These extreme conditions promote a series of nuclear fusion processes that are organized around the nucleus of helium, called an alpha particle, that contains two protons and two neutrons. This is a particularly stable configuration of nucleons (protons or neutrons) and its formation from free but very hot protons (the H nucleus) releases vast amounts of energy, thereby driving the formation process. Subsequent fusion involves the incorporation of more protons into the alpha particles and tends to give rise to nuclei that are simple integer multiples of alphas, the nuclei of C, O, Mg, silicon (Si), S, Ca, and Fe. Notice the presence of Si and the absence of N, P, Na, K, and Cl. The processes of synthesis of the alpha multiple nuclei generate these additional elements, albeit in lower amounts than the true alpha multiples. As discussed later, the d-orbital properties of Si make it unsuitable for the basic metabolic constituents of cells, and it instead has a very strong tendency to combine with O to form enormous three-dimensional inert silicate compounds such as quartz. Fusion energy release leads to the elements of the first four periods of the periodic table, the most abundant elements in the cosmos, and these are the building blocks of living matter.

2. RELATIVE ABUNDANCE OF THE ELEMENTS

By weight, three-fourths of the earth's crust is made of O (50%) and Si (25%), primarily as silicates. The oceans are mostly O (86%) and H (11%), primarily as water. Although abundance is important, fitness is just as important. Si is 135 times more abundant than C anywhere on Earth even remotely suitable for life, and it shares with C the tendency to form four covalent bonds, but it is highly unfit as a constituent of living matter compared to C. C can combine with other types of atoms to form myriads of small molecules, some of which involve double bonds. It combines to make strong stable bonds. Si, however, combines most strongly with O to make highly polymeric inert structures called silicates. Si–Si bonds are highly susceptible to attack by H_2O, NH_3^+, and O_2, unlike C–C bonds, which are stable. Only when structural material in the form of spines and spicules is used by living organisms such as cacti and sponges does Si actually occur in living matter. The metabolic pathways, however, are devoid of Si, using C compounds instead. These differences between C and Si have to do with their location in the second and third periods of the periodic table, respectively.

H, C, N, and O make up approximately 99% of the constituents of all living organisms. As members of the first and second periods, they are the most abundant elements and they combine to make short, strong, stable bonds. P and S, although less abundant, are nevertheless more abundant then heavier elements, and because they reside in the third period they possess unfilled d-orbitals. These orbitals confer on them the ability to engage in transfer reactions that mediate energy changes. S in the form of the thiol group, –SH, contributes to thioester-based energy

transactions, and P combines with O to form phosphate, which contributes to energy activation for the promotion of polymerizations. The most stable nucleus of all, and the end product of the fusion processes in stellar nucleosynthesis, belongs to Fe. This element is ideally suited as a conveyer of redox energy. Generation of redox energy, the energy of excited electrons, is caused by sunlight.

3. GIBBS FREE ENERGIES OF FORMATION

Once the elements have formed and planetary systems have accreted around Sun-like stars, temperatures become low enough for the spontaneous formation of many types of small molecules. This spontaneity is measured by the Gibbs free energy of formation from the elements. In a spontaneous process at constant temperature and pressure, the Gibbs free energy must decrease. This says nothing about rates. Without catalysts, a long period of time may be required before the minimum in free energy is achieved. The previous statement is the second law of thermodynamics expressed in terms of the Gibbs free energy rather than the entropy. This difference is a result of applying the second law to a system in contact with both temperature and pressure reservoirs and applying it to an isolated system, respectively.

In Table I, the Gibbs free energy of formation from the elements for a variety of compounds is listed for the standard conditions of a temperature of 298 K and atmospheric pressure. The table lists the name of the substance; its empirical formula; its molecular weight; its free energy of formation as $-\Delta G^0_{298}$, where the minus sign is explicitly included (a positive listing in the table means a negative free energy of formation); and its free energy of formation per gram as $-\Delta G^0_{298}/g$. By convention, the elements are given the value 0.0.

Notice how many important biological compounds are listed in Table I. The strongest tendencies toward formation from the elements (i.e., the most positive entries in the table) occur for aqueous ions, quartz, calcium phosphate, and water; however, most entries have values between 0.5 and 2.0. The table lists simple sugars, amino acids, a purine (xanthine), and numerous compounds that occur in metabolism. Clearly, life evolved from materials that were naturally occurring spontaneous products of the most abundant elements. Their formation from the elements is driven by a decrease in Gibbs free energy.

4. THE MONOMER-TO-POLYMER TRANSITION: NO LIFE TO LIFE

Life as we know it is distinguished by the presence of macromolecules: proteins, polysaccharides, and polynucleotides. These provide structural elements, catalysts, energy storage, and information storage and processing components for the cell. The dynamical cellular processes inside the cell that we recognize as living properties are almost exclusively associated with these polymers.

Polymers are chains of monomeric constituents that have been joined together by covalent bonds. In the cases of proteins and polynucleotides, these chains are linear, whereas in the case of polysaccharides they are branched. In all cases, the chemical structure of the linkages that join monomers into polymers is that of dehydration condensates. This means that the linkage between any two adjacent monomers anywhere in the chain can be broken by hydrolysis (i.e., cleavage by a molecule of water; literally "hydro-lysis"). Most cells are 80–95% water, which means that the tendency toward hydrolysis is strong, although perhaps not rapid unless catalysts are available. An acid environment and digestive enzymes are needed by our stomachs to process food because the spontaneous rates of hydrolysis are too slow, but these processes are spontaneous because they are attended by a decrease in Gibbs free energy. Thus, the problem for organisms is the reverse—that is, the synthesis of polymers from monomers, an uphill process thermodynamically that requires the input of Gibbs free energy. It is for this reason that the metabolic pathways are organized around the generation of energy-rich compounds, particularly adenosine triphosphate (ATP), the almost universal currency of free energy for the cell.

The transition of matter from no life to life is in essence the transition from monomers to polymers. The monomers are formed spontaneously from the elements and have negative free energies of formation relative to the elements. The polymers, however, have positive free energies of formation relative to the monomers from which they are synthesized. Most would appear in Table I with a value of 0.9 ± 0.15. Making the "peptide" bond between amino acids in proteins costs approximately 2 or 3 kcal/mol, the "glycosidic" bond between sugar monomers in polysaccharides costs approximately 3 or 4 kcal/mol, and the "phosphodiester" bond between mononucleotides in polynucleotides costs approximately 5 kcal/mol. In each case, activation is essentially a phosphate activation, slightly disguised

TABLE I

Gibbs Free Energy of Formation for a Variety of Compounds[a]

Compound	Empirical formula	Molecular weight (g/mol)	$-\Delta G^0_{298}$ (kcal/mol)	$-\Delta G^0_{298}/g$ (kcal/g)
Acetaldehyde	C_2H_4O	44	33.4	0.759
Acetic acid	$C_2H_4O_2$	60	94.7	1.578
Acetate (aq)	$C_2H_3O_2^-$	59	89.0	1.508
Acetyl CoA	$C_{23}H_{39}O_{18}N_7P_3S$	825	89.4	0.108
Alanine	$C_3H_7O_2N$	89	88.7	0.997
Ammonium ion (aq)	NH_4^+	18	19.0	1.056
Arginine	$C_6H_{15}O_2N_4$	175	126.7	0.724
Asparagine	$C_4H_8O_3N_2$	132	125.8	0.953
Aspartate	$C_4H_6O_4N^-$	132	167.4	1.268
Calcium (c)	Ca_n	40*n*	0.0	0.000
Calcium (aq)	Ca^{2+}	40	132.2	3.305
Calcium hydrogen phosphate	$CaHPO_4$	138	401.5	2.909
Carbon (c, graphite)	C_n	12*n*	0.0	0.000
Hydrogen carbonate (aq)	CHO_3^+	61	140.3	2.300
Carbon dioxide (g)	CO_2	44	94.2	2.141
Carbon dioxide (aq)	CO_2	44	92.3	2.098
Chlorine	Cl_2	71	0.0	0.000
Citrate	$C_6H_5O_7^{3-}$	189	278.7	1.475
Creatine	$C_4H_9O_2N_3$	131	63.1	0.482
Cysteine	$C_3H_7O_2NS$	121	81.2	0.671
Dihydroxyacetone phosphate	$C_3H_7O_6P$	170	308.9	1.817
Erythrose 4-phosphate	$C_4H_9O_7P$	200	343.8	1.719
Ethanol	C_2H_6O	46	43.4	0.943
Formaldehyde	CH_2O	30	31.2	1.040
Formic acid	CH_2O_2	46	85.1	1.850
Formate (aq)	CHO_2^-	45	83.8	1.862
Fructose	$C_6H_{12}O_6$	180	218.7	1.215
Fructose 6-phosphate	$C_6H_{13}O_9P$	260	420.0	1.615
Fructose bisphosphate	$C_6H_{14}O_{12}P_2$	340	621.3	1.827
Fumarate	$C_4H_3O_4^-$	115	144.3	1.255
Galactose	$C_6H_{12}O_6$	180	220.6	1.226
Glucose	$C_6H_{12}O_6$	180	219.1	1.217
Glucose 6-phosphate	$C_6H_{13}O_9P$	260	420.5	1.617
Glutamate	$C_5H_8O_4N$	146	166.5	1.140
Glutamine	$C_5H_{10}O_3N_2$	146	125.4	0.859
Glycerol	$C_3H_8O_3$	92	116.7	1.268
Glycerol phosphate	$C_3H_9O_6P$	172	319.2	1.856
Glycine	$C_2H_5O_2N$	75	90.0	1.200
Glyceraldehyde 3-phosphate	$C_3H_7O_6P$	170	307.1	1.806
Hydrogen (g)	H_2	2	0.0	0.000
Hydrogen cyanide	HCN	27	−28.7	−1.063
Hydrogen sulfide	H_2S	34	6.5	0.191
Hydronium ion (aq)	H_3O^+	19	56.7	2.984
Hydroxyl	HO^-	17	37.6	2.212
Iron (c)	Fe_n	55.8*n*	0.0	0.000
Iron (II) (aq)	Fe^{2+}	55.8	20.3	0.364
Iron (III) (aq)	Fe^{3+}	55.8	2.5	0.045

continues

Table I continued

Compound	Empirical formula	Molecular weight (g/mol)	$-\Delta G^0_{298}$ (kcal/mol)	$-\Delta G^0_{298}/g$ (kcal/g)
Isocitrate	$C_6H_5O_7^{3-}$	189	277.1	1.466
Isoleucine	$C_6H_{12}O_2N$	131	82.2	0.627
α-Ketoglutarate	$C_5H_4O_5^{2-}$	144	190.7	1.324
Lactate	$C_3H_5O^{3-}$	89	123.4	1.387
Lactose	$C_{12}H_{22}O_{11}$	332	362.0	1.090
Leucine	$C_6H_{13}O_2N$	131	85.1	0.650
Magnesium (c)	Mg_n	24.3n	0.0	0.000
Magnesium ion (aq)	Mg^{2+}	24.3	109.0	4.486
Methane	CH_4	16	12.1	0.756
Methanol	CH_4O	32	41.9	1.309
Nitrate ion (aq)	NO^{3-}	62	26.4	0.426
Nitrite ion (aq)	NO^{2-}	46	8.2	0.178
Nitrogen	N_2	28	0.0	0.000
Oxalate	$C_2O_4^{2-}$	88	161.2	1.832
Oxaloacetate	$C_4H_2O_5^{2-}$	130	190.4	1.465
Oxygen	O_2	32	0.0	0.000
Phenylalanine	$C_9H_{11}O_2N$	165	49.5	0.300
Dihydrogen phosphate ion (aq)	$H_2PO_4^-$	97	271.2	2.825
Hydrogen phosphate ion (aq)	HPO_4^{2-}	96	261.4	2.723
Phosphate ion (aq)	PO_4^{3-}	95	245.1	2.580
Phosphoric acid	H_3PO_4	98	−274.1	−2.797
Phosphorus (c, white)	P_n	31n	0.0	0.000
Phosphoenolpyruvate	$C_3H_5O_6P$	136	303.3	2.230
Potassium (c)	K_n	39.1n	0.0	0.000
Potassium ion (aq)	K^+	39.1	67.5	1.726
Pyruvate	$C_3H_3O_3$	87	113.4	1.303
Quartz (c)	$(SiO_2)_n$	60n	192.4n	3.207
Ribose 5-phosphate	$C_5H_{11}O_8P$	230	382.2	1.662
Ribulose 5-phosphate	$C_5H_{11}O_8P$	230	381.7	1.660
Sedoheptulose 7-phosphate	$C_7H_{15}O_{10}P$	290	457.1	1.576
Serine	$C_3H_7O_3N$	105	122.1	1.163
Silicon (c)	Si_n	28n	0.0	0.000
Sodium (c)	Na_n	23n	0.0	0.000
Sodium ion (aq)	Na^+	23	62.6	2.722
Sodium chloride (c)	NaCl	58.5	91.8	1.569
Succinate	$C_4H_4O_4^{2-}$	116	164.9	1.422
Succinyl CoA	$C_{25}H_{40}O_{20}N_7P_3S$	882	164.0	0.186
Sucrose	$C_{12}H_{22}O_{11}$	342	370.7	1.084
Sulfate ion (aq)	SO_4^{2-}	96	177.3	1.847
Sulfite ion (aq)	SO_3^{2-}	80	118.8	1.485
Sulfur (c, rhombic)	S_n	32n	0.0	0.000
Threonine	$C_4H_9O_3N$	119	122.9	1.033
Tryptophan	$C_{11}H_{12}O_2N_2$	204	29.9	0.147
Tyrosine	$C_9H_{11}O_3N$	181	92.5	0.511
Urea	CH_4ON_2	60	48.7	0.812
Valine	$C_5H_{11}O_2N$	117	86.0	0.735
Xanthine	$C_5H_5O_2N_4$	153	33.3	0.218
Xylulose	$C_5H_{10}O_5$	150	178.7	1.191
Water	H_2O	18	56.7	3.150

[a] Abbreviations used: aq, aqueous; c, crystaline; g, gaseous.

one way or another. Indeed, for proteins the intermediate-activated species is the aminoacyl-adenylate, a phosphate that is part of adenine monophosphate (AMP). The adenine of ATP can be substituted for by uracil, thymine, cytosine, and guanine, generating the activated monomers UTP, TTP, CTP, and GTP, from which, along with ATP, all DNA and RNA chains can be made. In these substances, the sugar ribose occurs for the RNA case and the sugar deoxyribose occurs for the DNA case. When deoxyribose is involved, the notation dATP, dGTP, dCTP, and dTTP is used. The activation of sugar monomers involves uridine diphosphate (UDP) derivatives. Thus, in all three cases—proteins, polysaccharides, and polynucleotides—the activated monomers are phosphorylated compounds. A single phosphate is used for aminoacyl-adenylates that are AMP derivatives, a diphosphate is used for UDP monosaccharides that are UDP derivatives, and triphosphates are used for UTP, ATP, dATP, TTP, CTP, dCTP, GTP, and dGTP that ultimately are all generated from ATP by phosphate transfers.

The required extra free energy is ultimately provided by ATP through formation of activated monomers that, once activated, spontaneously condense into polymers but only with the aid of complex catalytic molecular machinery. Thus, dehydration and activation are universal structural and synthetic themes for all life as we know it. Moreover, the use of energy-rich phosphate compounds to achieve activation is also universal. The problem with regard to the origin of life is how these phosphate compounds came into being before there were enzymes to catalyze energy metabolism. From this perspective, this is the fundamental "chicken-and-egg" problem.

Elaborate macromolecular structures have evolved that make activation possible in all extant cells. Many enzymes support a complex chain of oxidation reactions that harvests redox energy from high-energy electrons—not high energy in the physicist's sense but high energy in the sense of biochemistry. The electrons may be excited initially by the intense UV irradiating the nascent Earth. Also, reduced compounds with high electrical potentials (in Volts) can be generated in gentle thermal processes that dry out the reaction mixture. Drying means that water is leaving the system as steam and is taking with it the thermodynamic barrier to dehydration condensations. In the redox case, the electrons engage in a series of electron transfers, called redox reactions, that can couple to other energy-requiring processes such as monomer activation.

The purest form of redox process is of the type exemplified by the ferrous–ferric transition:

$$Fe^{2+} \Leftrightarrow Fe^{3+} + e^{-}$$

There are many instances of this type of redox step in the electron transport chains of bacteria and in mitochondria and chloroplasts, the energy-transducing organelles of eukaryotic cells. Two major classes of Fe-based transfers are found in all forms of life: those involving iron–sulfur proteins and those involving cytochrome, a heme-Fe-containing protein. These different species of Fe-containing proteins contain individuals with redox potentials spanning the range from free Fe (0.771 V) to ferredoxin–Fe (−0.420 V). This permits evolution of a chain of many relatively small steps in the total decrease of Gibbs free energy, affording many opportunities for coupling to other processes.

The most rudimentary alternative to the simple electron transfers of Fe is cotransfer of a proton with an electron (i.e., a hydrogen atom in pieces). The ubiquitous ubiquinone species, found in all organisms possessing electron transport chains, are of this type and are responsible for the mechanism of chemiosmosis in membrane-bound electron transport chain complexes. When the fundamental electron–proton, current–current coupling occurs in mitochondria, a decrease of Gibbs free energy during the electron transfer is partially used to increase Gibbs free energy for a steepening concentration gradient (that for a proton means a pH gradient) and an increasing transmembrane electrical potential because the proton is positively charged. This generates the chemiosmotic "proton-motive force," which is actually an electrical potential for protons (in Volts), not a force. Many processes, most notably transport of all sorts of molecules across the membrane, are driven by this form of cellular energy rather than by ATP. Indeed, ATP is synthesized chemiosmotically in bacteria, mitochondria, and chloroplasts by rotary enzyme complexes. ATP is ideally suited for activating monomers for the synthesis of polymers. However, there are some ATP-driven ion transporters as well. Since chemiosmosis probably evolved after life began, phosphate energy was the universal currency of energy for the origin of life, but only with the aid of another sort of energy discussed later.

The evolution of the cell membrane is unknown. Contemporary cellular membranes are primarily made from the class of lipids called glycerophospholipids, which contain two fatty acid side

chains and a phosphorylated alcohol such as choline, ethanolamine, inositol, or serine. The fatty acid side chains contain 14–24 carbon atoms. An elaborate enzyme complex catalyzes a long sequence of reactions in order to make these highly specialized lipids. Published models for an early evolution of a primitive precursor to these glycerophospholipids, in an abiotic geochemical thermal process, make an earlier evolution of chemiosmosis more plausible.

Glycerophospholipids are needed to form lipid bilayer membranes. These structures, the essence of the cellular envelope, spontaneously self-assemble from aqueous solutions of lipid molecules. The more easily generated fatty acids, or single-tailed lipids, produced by early Earth's geochemophysical processes only form micells instead of the larger and more versatile bilayer vesicles. Did the cellular membrane structure evolve before life evolved, during the prebiotic chemical evolution phase? Did it require a sophisticated metabolic machinery before its molecular constituents could be generated in sufficient quantities to comprise cells? Wouldn't such an apparatus necessarily have to be contained within a membranous boundary in order to keep enzyme and substrate together? This is another chicken-and-egg question.

Evidence suggests that before there was life as we know it, there were morphological structures resembling cells, the bilayer vesicles, and there was a source of chemical energy through coupling to the electron transfer processes. These electron transfers run downhill spontaneously but through couplings to other processes, they are generative of high-energy chemicals. The oxidation of carbohydrates by molecular O_2 releases enormous amounts of Gibbs free energy (approximately 700 kcal/mol for 1 mol of glucose and 6 mole of O_2), approximately 40% of which is harvested by the pathways of glycolysis, generation of acetyl CoA, the citric acid cycle, and the electron transport chain.

Carbohydrates are made by plants using the energy of sunlight, now in the visible part of the spectrum rather than in the primordial UV. From Table I, it is seen that carbohydrates, sugars, have less negative Gibbs free energies of formation than an equivalent amount of methane and carbon dioxide. The composition of many sugars can be written as a simple integer multiple, n, of formaldehyde [i.e., as $(CH_2O)_n$]. Thus, the point is made by the fact that the process

$$CH_4 + CO_2 \rightarrow 2CH_2O$$

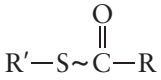
is attended by an increase in Gibbs Free energy of approximately 43.9 kcal/mol. It can be argued that sugar formed by whatever means can spontaneously convert back to methane and carbon dioxide by the inverse of the process shown previously, a process referred to as carbon disproportionation. Since this releases energy, it has been argued that it is a possible source of prebiotic energy. O_2 is also made by plants. This means that glycolysis (that does not directly involve O_2) is more primitive than the citric acid cycle or the electron transport chain (that are directly linked to O_2). Interestingly, glycolysis couples the carbon disproportionation of glucose to the generation of high-energy phosphate in the form of ATP. The dimer, pyrophosphate, is the energy-rich portion of ATP and is sufficiently energy rich that it alone can drive synthetic processes such as polymerizations. Pyrophosphate, as well as other polyphosphates, can also be made abiotically by simply using gentle heating to dryness or by coupling to thioester generation, as will be discussed later.

There might well have been myriads of micron-sized bilayer microspheres along with abundant reducing potential from high-energy electrons and abundant pyrophosphate before there was true cellular polymer synthesis. This micron-scale, energy-rich environment appears ideally suited for the evolution of polymerization machinery. Only within such a simple energy-rich environment can the required variety of polymers form at all. Before true macromolecular polymers emerged, a smaller class of mixed oligomers, the coenzymes, appear to have emerged. Coenzymes are the active catalytic moieties of many enzymes. They contain small molecular components called vitamins as well as ribose, phosphate, and various other constituents, usually approximately six units of various kinds per coenzyme. The presence of phosphate, indeed pyrophosphate, and sulfur in many of them suggests that these species are relics of an early stage of chemical evolution when energy transactions were already dominated by thioesters and phosphates. Since the coenzymes are thermodynamically uphill from their constituents, because they contain exclusively dehydration linkages, they could come into being only if the chemical milieu was one rich in energy-transducing species. With their emergence, a rudimentary metabolism, still devoid of modern protein catalysts, could have also emerged. This chemical environment, perhaps housed in a membranous vesicle, would provide the setting for the ultimate development of polymer synthesis.

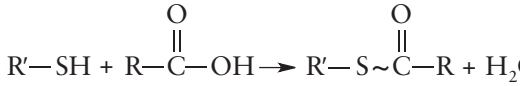
5. ENERGY: REDOX REACTIONS, THIOESTERS, AND PHOSPHATE

The nature of life as we know it is dominated by the properties of polymers, most notably the proteins and the polynucleotides. They are responsible for the catalytic and informational characteristics of cells. For them to exist at all in an aqueous environment, Gibbs free energy is required to overcome the thermodynamic barrier to their synthesis. The energy for monomer activation, the prerequisite to polymer synthesis, is supplied by phosphate compounds. The concentration of inorganic phosphate in all bodies of water on contemporary Earth is 10^{-5} times that in living cells. This makes it extremely unlikely that pyrophosphate was immediately available on the primitive earth for the purpose of monomer activation. A more primitive, and abundant, supply of energy was needed first to couple the redox energy generated by UV and Fe to synthetic chemical energy. The solution to this puzzle appears to be the thioester.

The basic structure of a thioester is



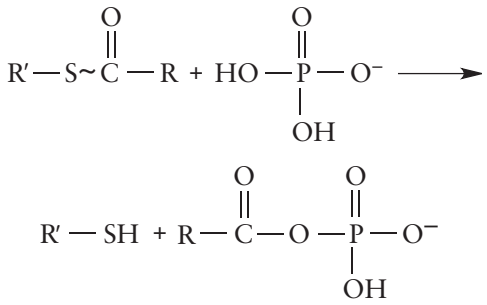
where *R′* and *R* represent any of a great variety of molecular groups, or residues. The tilde denotes the high-energy nature of the S and C bond. These energy-rich compounds can form from carboxylic acids and thiols without the aid of organismic enzymes by either of two energy-requiring mechanisms. The generic reaction is a dehydration reaction:



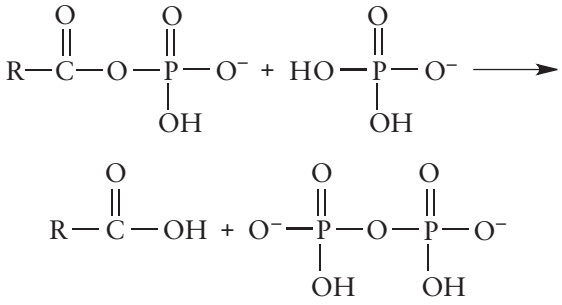
Variations on this theme in which the carboxylic acid on the left-hand side is an aldehyde or an α-keto acid are also possible, in which case two electrons and two protons or two electrons, two protons, and a carbon dioxide molecule are the by-products, respectively, rather than a molecule of water. One of the two energy-requiring mechanisms is simple heating of the thiol and the carboxylic acid at low pH. This promotes dehydration. The other mechanism works well for an aldehyde and a thiol, or an α-keto acid and a thiol, wherein the oxidation by two Fe^{3+} ions takes away the freed electrons, leaving as by-products the thioester and two protons or two protons and a carbon dioxide molecule, respectively. The ferric ions for this mechanism are the product of UV irradiation of Fe^{2+}, as discussed previously. It is even possible at room temperature and neutral pH to spontaneously form thioesters from glyceraldehyde and a thiol.

Once formed, thioesters may be reduced to form reduced organic compounds not naturally produced abiotically. Reductive carboxylation to form dicarboxylic acids is also possible. Reductive amination of an α-keto acid will produce an α-amino acid. The electrons needed for these reductions, and many others, may come from Fe^{2+} irradiated by UV. These reactions represent precursors to many contemporary metabolic reactions that today are catalyzed by enzymes. In these modern mechanisms, the role of a thioester intermediate is often retained in the active site of the enzyme. An example of this can be found in glycolysis.

In order to get phosphate into the picture, the first step is the formation of acyl-phosphates. These form spontaneously through phosphorolytic attack of a thioester by inorganic phosphate:



These acyl-phosphates are highly reactive and very energy rich. Acetyl-phosphate is an example of such a compound in contemporary metabolism. Inorganic phosphate attacks the acyl-phosphate to form pyrophosphate, a process that has also been shown to occur without enzymes.



Thus, the transduction of energy from redox energy, or heat, through thioester intermediates and into pyrophosphate was plausible on the primitive earth before life per se had arisen. A large variety of molecules can form from the thioesters and reducing potential generated from UV and Fe oxidation, which are still biologically relevant today. Indeed, life as we

know it is consistent with the proposition that it is these naturally occurring compounds that form the basis for life.

It is popular to refer to various stages of the evolution of life as "worlds" (e.g., the RNA world, discussed later). The earlier stage of chemical evolution described previously can be called the iron–thioester–pyrophosphate world. It represents the fundamental significance of energy transductions underlying the ultimate development of polymers and the overwhelming importance in energy transductions of three elements: P, S, and Fe. However, because P is so scarce, this world must have been preceded by the iron–thioester world. This world lasted for some time and provided a mechanism for the slow but steady accumulation of phosphate. Once the pyrophosphate world emerged, which was really the iron–thioester–pyrophosphate world, the origin of coenzymes could leave its signatures.

6. BANDED IRON FORMATIONS IN THE GEOLOGICAL RECORD

Banded iron formations are worldwide layered sedimentary deposits rich in iron that range in age from 1.5 to 3.8 billion years old, the age of the earliest known rocks. They contain 30–60% Fe^{3+}. No molecular oxygen was present in the earth's atmosphere until 1.5–2.0 billion years ago. Thus, the oxidant earlier could not have been O_2. O_2 is now the ultimate oxidant for nearly all organisms, but its function as such depends on a complicated enzyme complex. In contrast, the ability of UV to oxidize Fe^{2+} is geophysically natural and even autocatalytic. It has been demonstrated in the laboratory at acid, neutral, and alkaline pH. Since Fe^{3+} can serve as an oxidant for other molecules, Fe^{2+} can be regenerated. This creates an iron cycle. Depending on the stage of chemical evolution for all other molecular species, the amount of Fe^{3+} in any geological era could have gone through cycles as well, leading to the layers of magnetite minerals as the banded iron formations.

In an aqueous environment, the UV oxidation of Fe^{2+} forms Fe^{3+} and is followed by coprecipitation of ferric and ferrous oxides (Fe_2O_3 and FeO, respectively). These oxides form from the iron ions and water molecules releasing protons as a by-product as well as the mineral precipitate, magnetite:

$$2Fe^{3+} + Fe^{2+} + 4H_2O \rightarrow 8H^+ + [Fe_2O_3][FeO]. \quad (3)$$

The released protons are available for the reduction of a large variety of substances and could have been responsible for the production of such species as H_2, CH_4, HCN, and NH_3 as long ago as 3.8 billion years. These reduced species are otherwise not readily produced in other abiotic processes and are essential for the subsequent chemical evolution of amino acids and nucleoside bases, the purines and pyrimidines. It should be noted that UV energy was one of the most abundant forms of energy on the primitive earth.

Fe^{3+} may well have served as an early electron acceptor and thereby supported the oxidative synthesis of thioesters, as was argued previously. Iron–sulfur proteins are among the most ancient known proteins and may be relics of an earlier iron–thioester–pyrophosphate world but clearly not of the iron–thioester world.

7. THE IMPORTANCE OF D-ORBITALS

It is thought that the singular importance of P and S in energy transactions and the noteworthy unfitness of Si for metabolic chemistry are due to the properties of elements in the third period of the periodic table and also to d-orbitals. These electron states appear in the third period of the periodic table but are unoccupied in Si, P, and S. Moving from left to right within a period, atoms actually get smaller even though they get larger as the period increases. Thus, all atoms in the third period are larger than those in the second period, but within the third period the size order is Si > P > S. Like C, N, and O, P and S can form double bonds, but Si is too large to do so. Thus, Si cannot form the vast variety of compounds that C does and is restricted to single bonds, forcing it into immense three-dimensional polymers of $(SiO_2)_n$. Moreover, its unoccupied d-orbitals afford an opportunity for attack of the relatively open Si–Si bonds by nucleophiles such as O_2, H_2O, and NH_3. In contrast, the bonds of P and S are tighter and their relatively smaller size allows them to form double bonds. The same d-orbitals that are such a liability for Si are instead an asset for P and S. For phosphates in particular, the result is a chemical group that is easily transferred while relatively impervious to chemical attack. In contrast, the smaller atoms of H, C, N, and O produce very stable bonds that would not be suitable for easily transferred groups. Thus, phosphate appears to be a natural choice for the activation group in synthetic biochemistry. The d-orbitals make it pentavalent and this contributes

to the resonance enhancement of its energy transactions that confer on it its high-energy status.

8. THERMAL ENERGY AND SYNTHESIS

Thermal energy is of central importance for the origin of life in two distinct ways. At the molecular level, it gives rise to a very robust energetic activity called Brownian motion. The collisions between water molecules and other molecules are very energetic and are responsible for the Brownian motion of other molecules. This Brownian motion makes proteins fluctuate vigorously in conformation and promotes the catalytic activities of these molecules. Indeed, without the Brownian motion, the protein enzymes would lose their functionality. Brownian motion is also responsible for very rapid mixing of species within a micron-sized environment. The other importance for thermal energy is that it can overcome the thermodynamic barrier to dehydration polymerizations. Since this barrier is a result of the overwhelming presence of water molecules and the fact that all polymer linkages are dehydration bonds, thermal energy (i.e., heat) can remove the water molecules if applied in a gentle enough manner. Too much heating will simply destroy the molecules and create a gooey, tarry mess. For origin of life considerations, it is this second aspect of thermal energy that is of synthetic importance.

A prime example of the synthetic power of thermal energy is the thermal synthesis of polyamino acids, called thermal proteinoids. This type of polymer synthesis works especially well if the amino acid mixture is rich in aspartic acid and glutamic acid, not unlike the composition of many contemporary biological proteins. It is also assisted by the presence of phosphate, which by itself yields polyphosphates when heated. Probably, the polyphosphates formed in the amino acid mixture help promote condensation of the amino acids. Such phosphate-enhanced reaction mixtures yield thermal proteinoids at temperatures as low as 60°C if heating is allowed for 150 h. It is also of interest that when the thermal proteinoids are returned to water after synthesis, they spontaneously form uniform populations of micron-diameter microspheres. Since many amino acid residues are lipophilic, these thermal polymers have some properties in common with true glycerophospholipids. In particular, the proteinoid microspheres are bilayered vesicles rather than micells. The proteinoids also possess a number of relatively weak catalytic abilities covering the basic reaction steps that make up metabolism. All of this together strongly suggests a natural mechanism for the formation of catalytic microsphere environments in which the iron–thioester–pyrophosphate world could gain a foothold. Although the proteinoid microsphere membranes are not nearly as impermeable as the membranes of modern organisms, they would nevertheless trap any polymers formed within them from monomers that could freely enter them from the environment. With the evolution of a more sophisticated metabolism, incorporation of primitive lipid analogs would make the membranes much less permeable and permit the establishment of chemiosmotic mechanisms. The proteinoid model has received a great deal of criticism in the literature. However, a close study of the primary literature presenting the experimental results that support the model clearly refutes much of the criticism.

9. THE RNA WORLD

Since the discovery of ribozymes (RNA molecules with enzymatic activity) in the early 1980s, it has been very popular to speak of the origin of life in terms of the RNA world. The basic notion is that RNA simultaneously has the ability to serve as the genetic material and as the first enzymes. Thus, the answer to the question, Which came first, proteins or polynucleotides?, receives the emphatic answer, polynucleotides, specifically RNA. This question is an old chicken-and-egg question, and opinion has alternated between proteins first and polynucleotides first for some years. A more evenly balanced view seems warranted given the difficulties surrounding the chemical evolution of mononucleotides, especially their important ingredient, the ribose molecule, that so far is not easily produced in abiotic experiments. Previously, it was clearly established that the spontaneous production of polymers is thermodynamically inhibited. RNA polymers require phosphodiester linkages, the dehydration linkages for polynucleotides, and these are uphill in Gibbs free energy. This situation is doubly a problem for polynucleotides because they are made from monomers, the mononucleotides, which are themselves oligomers (i.e., very small polymers). ATP, for example, contains adenine, ribose, and three phosphates. Each component is linked to its neighbors by dehydration linkages for a total of four in the case of ATP or any other mononucleotide triphosphate.

Thus, even the oligomeric monomers for polynucleotide synthesis are uphill in Gibbs free energy relative to their constituents. Given these thermodynamic barriers, coupled with the difficulty explaining an easy natural origin for ribose, it is highly unlikely that an RNA world was the beginning stage for life as we know it. For this reason, the RNA world section of this article is the last section, not the first.

In the preceding sections, it was argued that an energy-transducing chemical evolution preceded the true emergence of life. Before the RNA world could have emerged, the iron–thioester–pyrophosphate world had to be in place, perhaps already encapsulated inside microspheres with proteinoid and/or lipid precursor membranes. Recall that phosphate, an essential component of polynucleotide phosphodiester backbones, is scarce on the earth and only through the aid of thioesters does it seem likely that it was recruited from the geophysical environment into the biological realm, as described previously. If the question is asked, Which came first, the RNA world or the iron–thioester–pyrophosphate world?, then it should be clear that it was the latter.

Once the essential energy requirements for the transition from monomers to polymers are established, the RNA world becomes possible. In this sense, the origin of life, the transition from monomers to polymers, coincides with the emergence of the RNA world. The problem then shifts from one of origins to one of evolution. The key issue is that of protein biosynthesis by a gene-directed mechanism. The evolution of this mechanism must account for the emergence of ribosomes, transfer RNA (tRNA), messenger RNA (mRNA), ribosomal RNA, and the aminoacyl-tRNA synthetases (aaRSs). The aaRSs of contemporary organisms are a diverse and complex group of enzymes responsible for attaching the cognate amino acid to a tRNA with a specific anticodon. The mechanism of this recognition process in contemporary organisms is far from completely understood, and all of the structural and evolutionary evidence strongly points to a long evolution of the mechanism and its components. This unsolved problem constitutes the big gap in our comprehension of the contemporary genetic mechanism.

Models for how these complex components could have evolved from much more primitive precursors have been proposed. They purport to explain (i) why the genetic code is a three base code with much degeneracy in the third codon and why the codons for the amino acids are what they are; (ii) why the N-terminus to C-terminus direction of the gene-directed protein is colinear with the $5' \rightarrow 3'$ direction of the cognate mRNA; (iii) why L-amino acids and D-ribose are used (i.e., whether this is a relative relationship that could as easily have been D-amino acids with L-ribose instead or an absolute necessity); and (iv) how gene-directed synthesis of proteins got started before there was the complex array of protein and RNA catalysts that currently make the process work. There are many other related issues. One published model proposes that the first genes and mRNAs were one and the same molecule, an RNA, and that protein synthesis took place directly on this RNA when activated amino acids interacted directly with the 2′–OH groups of the ribose moieties to form amino acyl ribose esters. Through a subsequent conformation change, the esterified amino acids link up to form small proteins and release the RNA for another round of synthesis. The RNA has served as both gene and messenger in this model. The activated amino acids might be carboxyl phosphates generated from pyrophosphate or they might be thioesters. Subsequent evolution of DNA, lacking the 2′–OH groups, would separate the roles of information storage in DNA and information translation in the RNA. This very simple genetic system would be poised for the subsequent evolution of ribosomes, tRNAs, and aaRSs. Such models provide conceptual insight into origins and evolution and also provide a number of challenges for experimentalists. However, advances in biotechnology make possible a number of experiments to test the ideas just presented.

SEE ALSO THE FOLLOWING ARTICLES

Earth's Energy Balance • Ecosystems and Energy: History and Overview • Heterotrophic Energy Flows • Lithosphere, Energy Flows in • Ocean, Energy Flows in • Photosynthesis and Autotrophic Energy Flows • Sun, Energy from • Thermal Energy Storage • Work, Power, and Energy

Further Reading

Berg, J. M., Tymoczko, J. L., and Stryer, L. (2002). "Biochemistry." Freeman, New York.

Calvin, M. (1969). "Chemical Evolution." Oxford Univ. Press, New York.

Deamer, D., Kuzina, S. I., Mikhailov, A. I., Maslikov, E. I., and Seleznev, S. A. (1991). Origin of amphiphilic molecules and their role in primary structure formation. *J. Evol. Biochem. Physiol.* **27**, 212–217.

de Duve, C. (1991). "Blueprint for a Cell: The Nature and Origin of Life." Patterson, Burlington, NC.

Fox, R. F. (1988). "Energy and the Evolution of Life." Freeman, New York.

Fox, S. W., and Dose, K. (1972). "Molecular Evolution and the Origin of Life." Freeman, San Francisco.

Fowler, W. A. (1967). "Nuclear Astrophysics." American Philosophical Society, Philadelphia.

Harold, F. M. (1986). "The Vital Force: A Study of Bioenergetics." Freeman, New York.

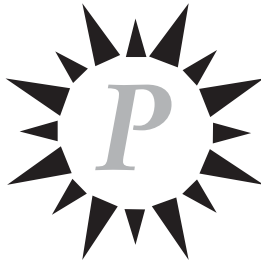
Joyce, G. F., and Orgel, L. E. (1999). Prospects for understanding the origin of the RNA world. *In* "The RNA World" (R. F. Gesteland, Ed.), 2nd ed., pp. 49–77. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Schopf, J. W. (1999). "The Cradle of Life: The Discovery of the Earth's Earliest Fossils." Princeton Univ. Press, Princeton, NJ.

Wald, G. (1962). Life in the second and third periods; Or why phosphorus and sulfur for high-energy bonds? *In* "Horizons in Biochemistry" (M. Kasha and B. Pullman, Eds.), pp. 156–167. Academic Press, New York.

Weber, A. L. (1998). Prebiotic amino acid thioester synthesis: Thiol-dependent amino acid synthesis from formose substrates (formaldehyde and glycolaldehyde) and ammonia. *Origins Life Evol. Biosphere* **28,** 259–270.

Passenger Demand for Travel and Energy Use



ANDREAS SCHAFER
Massachusetts Institute of Technology
Cambridge, Massachusetts, United States

1. **Determinants of Passenger Travel Energy Use**
2. **Travel Demand and Mode Choice**
3. **Vehicle Technology**
4. **Human Factors**
5. **Passenger Travel Energy Use and Emissions**
6. **Reducing Passenger Travel Energy Use and Emissions**
7. **Outlook**

Glossary

energy intensity Energy use per unit service (here passenger-kilometer).

light-duty vehicle Individual passenger transport modes consisting of mainly automobiles and those light trucks (e.g., minivans, vans, sport utility vehicles, pickup trucks) used for passenger travel.

occupancy rate The number of occupants in a vehicle weighted by the driving distance, that is, the ratio of passenger-kilometers and vehicle-kilometers.

passenger-kilometer The product of the number of passenger trips and the average distance traveled, where 1 passenger-km corresponds to one passenger traveling a distance of 1 km and so on.

vehicle-kilometer The product of the number of vehicle trips and the average distance traveled, where 1 vehicle-km corresponds to one vehicle traveling a distance of 1 km and so on.

Among the many factors determining travel demand, growth in income and growth in population are the two single most important drivers of human mobility. In addition, on a highly aggregate, world regional level that is necessary to estimate world passenger travel energy use, mainly four constraints determine travel demand and mode choice. Two of the constraints are so-called travel budgets; although a country's population consists of many individuals with very different preferences, the average expenditure shares of money and time are roughly constant. Travel demand and mode choice also depend on transportation prices, that is, the unit cost of travel to the consumer. As this article shows, for some modes, especially aircraft, these unit costs fell sharply during the second half of the 20th century, contributing to a rapidly rising demand for that mode. Finally, travel demand is also determined by land use settings; humans in densely populated areas rely more heavily on mass transit modes, whereas those living in countries with very low-density urban areas enjoy greater use of automobiles.

1. DETERMINANTS OF PASSENGER TRAVEL ENERGY USE

Getting from travel demand to energy use requires multiplying the product of per capita passenger-kilometer (*pkm/cap*) and population (*pop*) with energy use per vehicle-kilometer traveled (*E/vkm*) and dividing by the occupancy rate (*pkm/vkm*). Equation (1) describes the overall identity of passenger travel energy use (*E*) as it may apply to, for example, automobile transport:

$$E = \frac{E}{vkm} \cdot \frac{vkm}{pkm} \cdot \frac{pkm}{cap} \cdot pop. \qquad (1)$$

Note that most of the factors in Eq. (1) are a function of additional variables. For example, energy use per vehicle-kilometer traveled (*E/vkm*) is determined by vehicle size, technology, payload, and driving characteristics. Also note that many of the factors in the equation are interrelated. For example, increasing the number of persons sharing a vehicle (i.e., decreasing *vkm/pkm*) reduces *E/pkm* but increases *E/vkm* because the vehicle is now transporting a

 © 2004 Elsevier Inc. All rights reserved.

heavier load. Simultaneously, a greater number of people sharing a car reduces the number of vehicles on the road, and this—as a first-order effect—increases traffic flows and, thus, reduces *E/vkm*. In the subsequent sections, this article examines each of these components more closely. It begins with the growth in per capita travel demand (*pkm/cap*), moves on to vehicle technology (*E/vkm*), and finally examines trends in human factors that are reflected in the inverse vehicle occupancy rate (*vkm/pkm*) but also in the choice of technology and driving characteristics (*E/vkm*).

2. TRAVEL DEMAND AND MODE CHOICE

2.1 Factors Affecting Travel

Growth in income and growth in population are the most important drivers of passenger mobility. Higher income allows households to afford motorized transportation and, ultimately, to own automobiles that enable more trips and longer distances within a given amount of time compared with most other modes of transport. In contrast, population represents a scaling factor of total mobility. In addition, travel demand is determined by socioeconomic variables, including age, gender, and profession. For example, males travel more than females, and people in their 20s and 30s are more mobile than other age groups. Further determinants include land use settings, supply variables (e.g., accessibility to transport systems), prices for (transportation-related) goods and services, and political and institutional forces. Although this list of factors affecting travel demand could be extended, quantifying total energy use for, and emissions from, passenger travel requires a more aggregate perspective of travel demand.

Hence, in addition to the two driving forces, growth in gross domestic product (GDP) per capita and growth in population, aggregate passenger travel demand depends mainly on four variables: the share of money dedicated to transportation, the share of time dedicated to transportation, land use characteristics, and prices for travel by mode. Yacov Zahavi was first to hypothesize constant levels of time and money expenditure shares per traveler (i.e., a person who conducts at least one mechanized trip per day) and per household, respectively. Zahavi has used the two budgets in urban transportation models. Subsequently, other researchers have generalized them to the average person and found similar levels of stability at national levels. More recently, both budgets were applied to project future levels of travel demand worldwide.

Figures 1A and 1B illustrate the stability of the two travel budgets for a wide range of income levels. As shown in Fig. 1A, people dedicate roughly 5% of their daily time (or 1.2 h) to transportation on average, independent of whether they live in African villages or in the highly industrialized world. Although travel time and money expenditures are highly variable on an individual level (reflecting widely differing preferences and constraints across travelers), mean values of these expenditures are much more stable for groups of people. Thus, the variability of travel time expenditures is in part dependent on the aggregation level. Variability is highest across cities (empty circles in the figure), already smaller across countries (full circles), and should be negligible on a world regional level. In addition to the constant travel time budget, humans spend a predictable share of income on travel, ranging from 3 to 5% in the case where they predominantly depend on public transportation modes (i.e., the developing world). This share increases to 10 to 15% if they own at least one automobile (Fig. 1B). Mainly because of these budgets' skewed distribution over the population, attempts to explain the underlying forces for constant averages have not been convincing until recently.

Because of their fundamental nature, the two budgets govern the daily travel behavior of virtually each person irrespective of cultural, geographical, and economic settings. While these budgets are very similar across national boundaries, economic and other settings can vary substantially and thus cause differences in averages of travel demand indicators. For example, in very low-income rural areas of developing countries, nearly all trips are done within a distance of 5 km—the distance that can be covered on foot within the travel time budget. With rising income, humans can afford faster modes of transport and, thus, increase the geographical area of daily interaction. In the high-income, automobile-dominated industrialized countries, more than 95% of all trips occur within a distance of 50 km—the distance that can be covered by an automobile within the travel time budget.

In addition to the two travel budgets, travel behavior is determined by land use characteristics. Extreme differences in population densities can have a sensible impact on the choice of transportation mode. Because of severe space constraints and the

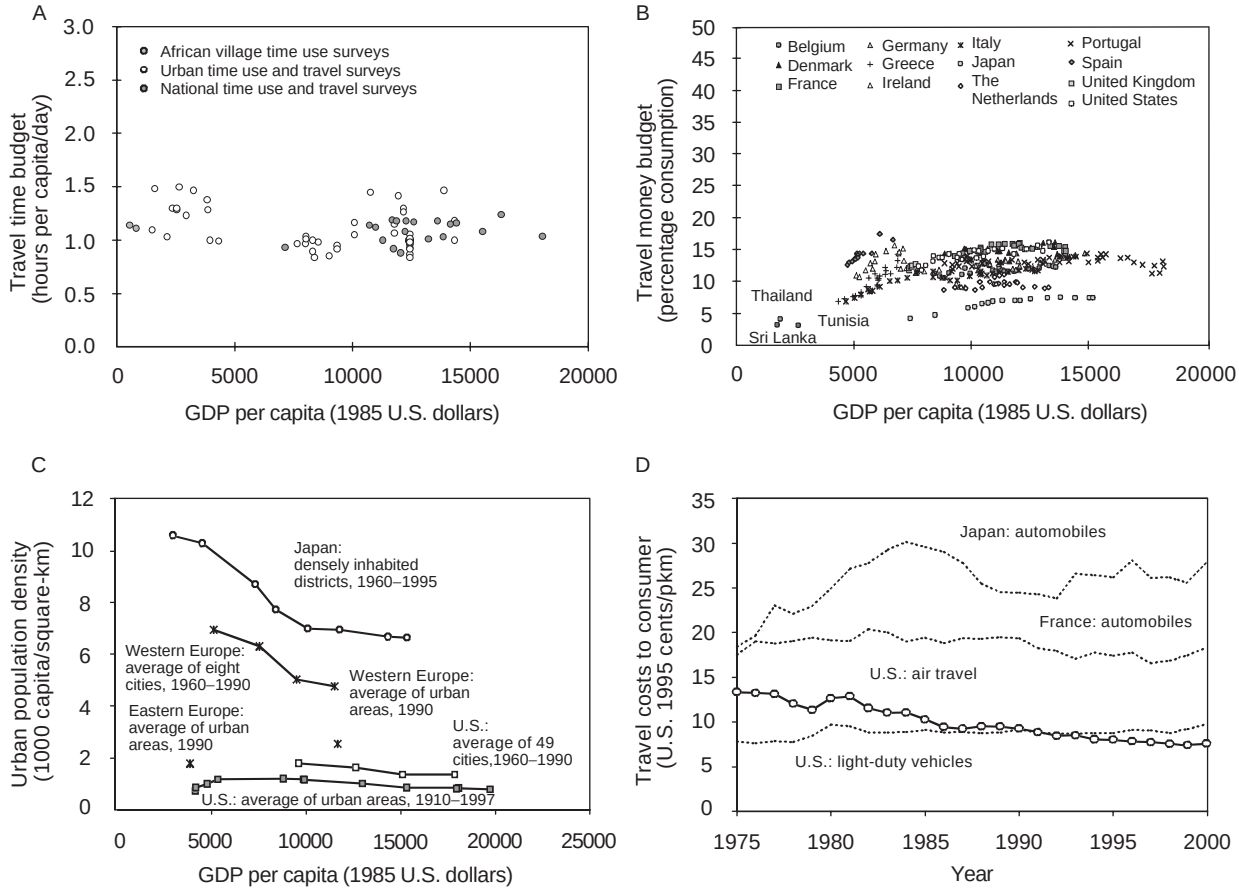


FIGURE 1 Fundamental determinants of total travel demand and mode choice. (A) Average travel time expenditures per capita and day (travel time budget). (B) Share of consumption dedicated to transportation on average (travel money budget). (C) Declining urban population densities in various parts of the industrialized world. (D) Average price of travel per passenger-kilometer for major transportation modes in the United States, France, and Japan.

associated high costs of operating a private motor vehicle, very high-population density areas are served mainly by public transportation modes. In contrast, lower density cities are dominated by private vehicles, which provide the highest levels of flexibility in that type of environment. Thus, the ongoing process of suburbanization in higher income countries, enabled mainly by the automobile, necessarily reduces the relative importance of public transport modes. Figure 1C illustrates the long-term process of urbanization and the subsequent sprawl in the United States beginning after World War II, the period of mass motorization. Suburbanization has occurred in all other parts of the industrialized world, mostly at significantly higher population densities.

The final determinant of aggregate travel behavior is travel costs to the consumer. Figure 1D reports time trends for air travel (United States) and automobile transport (France, Japan, and United States). While absolute automobile travel costs have remained stable in France and the United States, real per capita income has increased by more than 50% during the same time. In Japan, where automobile travel costs have increased by nearly 30% since 1975, real per capita income has increased by 80%. Together the trends in travel costs and income imply that automobile travel has become increasingly affordable everywhere. From a cross-country perspective, differences in driving costs remain related to the availability of land and associated prices. Costs per passenger-kilometer for automobile travel are lowest in the United States, the country with the lowest urban population densities, and are highest in highly dense Japan. (The low gasoline taxes in the United States have contributed to the comparatively low cost of driving and the associated low population densities.) Whereas automobile travel costs have remained roughly constant in the United States, costs

of air travel have actually declined, from U.S. 12.1 cents/pkm in 1978 (the beginning of airline deregulation) to 7.5 cents/pkm in 2000.

Together, the two driving forces (income and population) and the four variables describe a sufficiently complete picture of worldwide travel demand and mode choice.

2.2 Total Mobility

Translating income into a roughly proportional amount of travel expenditures, a stable travel money budget implies that total travel demand rises approximately in proportion to income. (The initially increasing share of the travel money budget, which results in travel expenditures rising faster than income, is used to finance the transition from lower cost public modes to higher cost automobiles [Fig. 1]; thus, travel demand also roughly grows in proportion to income during the period of rising money budgets.) Figure 2 illustrates the tight relationship between income (approximated by GDP per capita) and per capita traffic volume (e.g., automobiles, buses, railways, aircraft) for 11 world regions and the world from 1960 to 1990. Although the illustrated growth relationship holds for all regions, differences in economic structures, prices, land use settings, culture, and values can cause differences in per capita travel demand at a given GDP per capita level. For example, lower population densities along with lower prices for owning and operating an automobile, but also differences in cultural values such as higher labor flexibility, have caused per capita traffic volume to be nearly 40% higher in North America than in Western Europe at a GDP per capita level of U.S. $10,000.

Table I reports traffic volume per capita and totals for all 11 world regions for 1960 and 1990. Historically, most of the world's traffic volume has occurred in the industrialized world, accounting for slightly more than half of the world passenger traffic volume in 1990. However, because the highest growth in income and population has occurred in the developing world, culminating with a nearly 12-fold rise in the Other Pacific Asia region (the rapidly developing Asian economies), those countries are increasing their share in world mobility. Currently, most of the traffic in the developing world occurs in rapidly growing urban areas (where income is higher), including the many "megacities," where the vehicle fleet grows much faster than transportation infrastructure, leading to extreme levels of congestion, urban air pollution, accidents, and other transportation-related problems.

2.3 Mode Choice

Whereas a constant travel money budget translates rising income into growing travel demand, a constant travel time budget requires the growing traffic volume to be satisfied by faster modes of transport. The associated continuous shift toward increasingly fast transportation modes with rising per capita

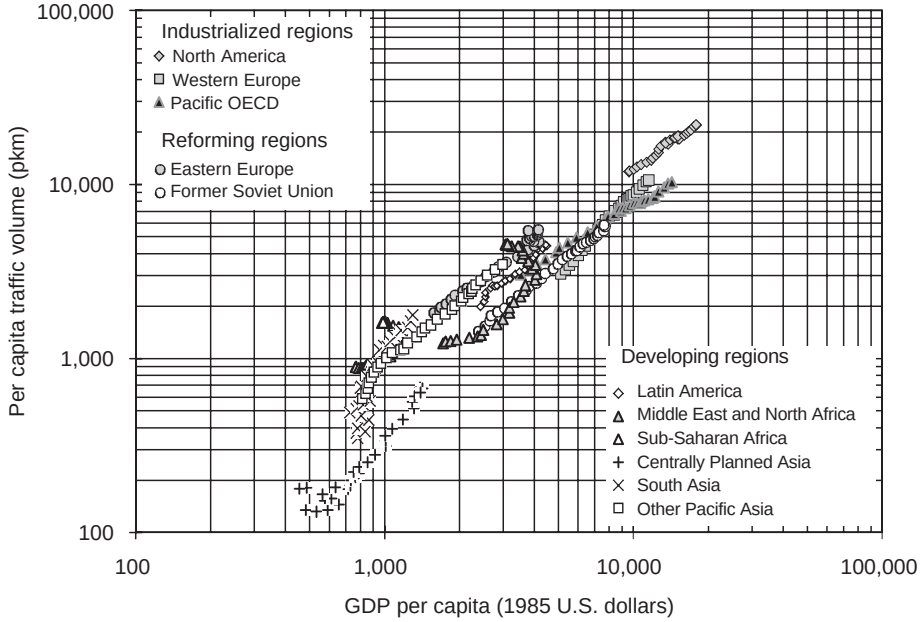


FIGURE 2 Evolution of per capita traffic volume over GDP per capita for 11 world regions, 1960–1990.

TABLE I

Traffic Volume (per Capita and Total) in 11 World Regions, 3 Meta-regions, and the World: 1960 and 1990

	1960			1990		
		Absolute			Absolute	
	pkm/cap	Billion pkm	Percentage world	pkm/cap	Billion pkm	Percentage world
North America	11854	2384	43.5	22078	6193	26.7
Pacific OECD	3074	1106	20.2	10622	4696	20.2
Western Europe	3025	323	5.9	10294	1482	6.4
Total industrialized	4400	3813	69.6	14276	12372	53.3
Eastern Europe	1824	181	3.3	5389	666	2.9
Former Soviet Union	1419	295	5.4	5796	1631	7.0
Total reforming	1550	477	8.7	5672	2297	9.9
Centrally Planned Asia	1222	140	2.6	4546	1244	5.4
South Asia	898	193	3.5	1614	811	3.5
Other Pacific Asia	152	109	2.0	637	805	3.5
Sub-Saharan Africa	349	200	3.7	1778	2015	8.7
Middle East and North Africa	587	125	2.3	3470	1459	6.3
Latin America	1980	424	7.7	5094	2228	9.6
Less developed countries	582	1191	21.7	2125	8562	36.9
World	1814	5481	100.0	4382	23231	100.0

traffic volume is shown in Fig. 3 for five world regions. The modal trajectories in this figure report three stages of the evolution of motorized mobility.

During the first stage, at low levels of economic development and travel demand, low-speed collective modes (railways and buses) account for virtually all motorized transport services, as seen in the South Asia region. At that initial stage, a substantial share ($>50\%$) of all passenger-kilometers are covered by nonmotorized modes. Individual and high-speed means of transport still account for less than 10% of total passenger-kilometers and serve only a small share of the population. (Due to the lack of data, nonmotorized modes and motorized two- and three-wheelers were excluded from that analysis.)

During the second stage, at higher levels of income and travel demand ($\sim$2000 passenger-km per capita), the relative importance of low-speed collective modes saturates and declines due to the strongly rising demand for increasingly affordable automobile travel. Without a significant share of automobile transport, these elevated demand levels could not be reached within a fixed travel time. In Eastern Europe, the rise of the automobile occurred despite constrained access to individual modes of transport in a socialist system. Apparently, these fundamental dynamics of the transportation system could not be easily eliminated by transportation policies. High-speed transport systems still have a small share.

During the third stage of mobility development, above demand levels of 6000 to 10,000 passenger-km per capita, the share of automobile traffic volume saturates and subsequently declines. The high demand levels now require a substantial share of air traffic and/or high-speed ground transportation systems. The rising significance of high-speed travel was enabled by continuously declining air travel costs (Fig. 1D). In North America, where all passenger traffic is split between automobiles and aircraft, the latter, offering still greater mobility through significantly higher speeds, is strongly increasing its market share at the expense of passenger cars. At these high levels of per capita traffic volume, ordinary rail and bus services maintain a low share in special market niches (mainly high-density population areas).

The specific modal shares are also determined by urban land use characteristics. In North America, where urban population density is low ($\sim$800 people/square-km in U.S. urban areas, as reported in Fig. 1C), the share of automobile traffic volume was saturated at nearly 90% in 1960. In more densely populated Western Europe, with about four times the density levels of U.S. cities, automobiles achieved only a 72% share of total traffic volume. Correspondingly, in the Pacific Organization for Economic Cooperation

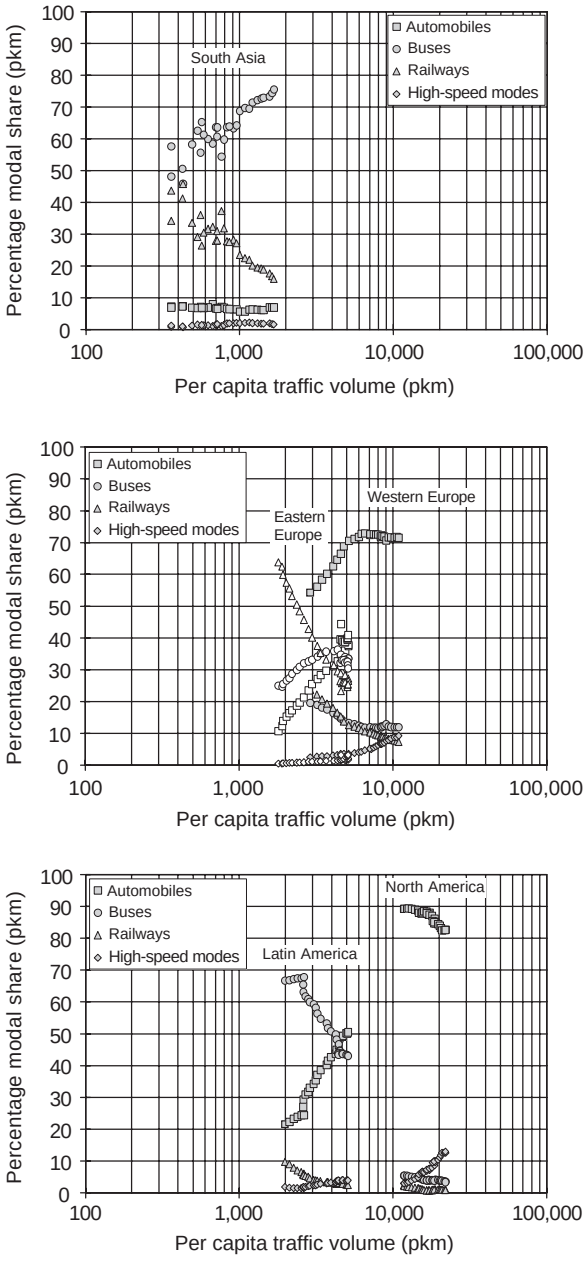


FIGURE 3 Three stages characterizing the evolution of motorized mobility: relative importance of transport modes (in percentage passenger-kilometers) over passenger-kilometers per capita, 1960–1990.

and Development (OECD) region, which is dominated by Japan and its high population density (five to six times the U.S. level), the relative importance of automobile traffic volume is likely to peak below the Western European level. Primarily as a consequence of different types of land use, investments in transportation infrastructures also have been different across regions (e.g., a high share of railways in high-density Pacific OECD [essentially Japan], a low share of railways in low-density North America).

On a lower level of aggregation, the previously described mode shifts occur within two separate transportation markets. Whereas the substitution of automobiles for public transportation occurs mainly in expanding urban areas, the shift from automobiles to high-speed modes of transport occurs exclusively in longer distance intercity travel.

2.4 Travel Purposes

Nearly all travel is done for satisfying a purpose, whether it is getting to work, pursuing personal business-related activities such as shopping, or enjoying leisure. Figure 4 shows that the observed growth in travel demand follows regular patterns when disaggregated into its purposes. At low levels of income and low levels of daily distance traveled, people undertake less than two trips per day (e.g., Delhi suburbs during the late 1970s). The associated per capita traffic volume is less than 15 passenger-km per day (5500 passenger-km per year). At such low mobility levels, travel serves satisfying basic needs nearly exclusively. One trip in a day is dedicated to a combination of work (short-term survival) and education (longer term well-being), and roughly half a trip is dedicated largely to personal business (essentially shopping at local markets), on average. With rising per capita traffic volume and automobile ownership, additional trips for personal business (e.g., shopping, health care, religious services) and leisure (including holidays) are undertaken. At the high income levels comparable to those of OECD countries, people make more than three trips per day, devoting approximately one trip to work or education, one to two trips to personal business, and one trip to leisure. The highest trip rate can be observed for the United States, with the highest level of passenger-kilometers per capita, Here, personal business trips account for nearly half of all trips made, and work trips only account for 20% of all trips. At such high mobility levels, some travel is also done for its own sake. According to the 1995 U.S. Nationwide Personal Transportation Survey, pleasure driving accounted for 28 billion passenger-km or 0.7% of total passenger traffic volume.

3. VEHICLE TECHNOLOGY

Any movement requires energy. Basic laws of physics imply that vehicle energy requirements rise with the

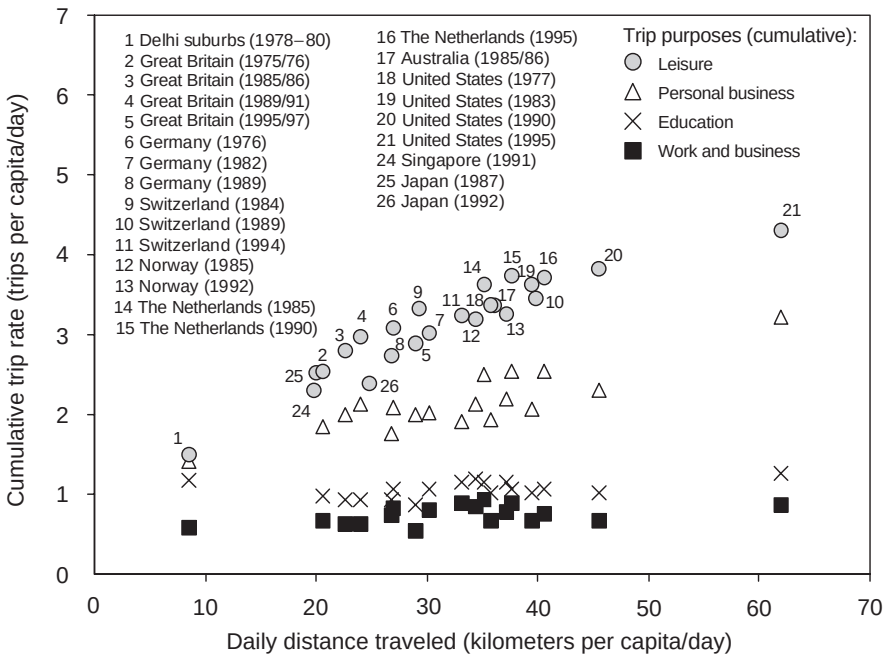


FIGURE 4 Growth in trips per capita per day, by purpose, over daily distance traveled.

TABLE II

Major Characteristics of Transportation Modes in the United States

	Mean travel speed (km/h)	Mean trip distance (km)	Mean occupancy rate (pkm/vkm)	Mean energy use/intensity (MJ/vkm)	(MJ/pkm)
Walking	4.7	0.8	N/A	N/A	0.21
Bicycle	11	2.1	1.0	0.07	0.07
Motorcycle	52	16	1.2	1.6	1.4
Buses	24	14	16.8	18	1.1
Railways	10–46	6–168	23.4	45	1.9
Automobile	54	15	1.6	3.7	2.3
Personal truck	54	15	1.6	4.6	2.9
Aircraft	550	1216	91.1	220	2.4

Note. Mean travel speeds and mean trip distances are derived from the 1995 U.S. Nationwide Personal Transportation Survey. Mean occupancy rates, energy use, and energy intensity for motorized modes are derived from aggregate statistical references and related to 2000. The conversion between energy use and energy intensity does not account for extra passenger weight. Ranges in speed and mean trip distance for railways reflect tramways (lower number) and intercity rail (high number). Other figures are averages.

square of speed, whereas vehicle power requirements (i.e., energy use per unit time) increase with the cube of speed. Because travel speeds can differ substantially across transport modes, understanding differences in vehicle energy use requires taking into account vehicle operation characteristics. Table II reports major operation characteristics and the approximate level of energy use for various transport modes. As would be expected, energy intensity rises with speed provided that comparisons are made on a consistent basis. The mean vehicle speed and distance, here derived from the 1995 U.S. Nationwide Personal Transportation Survey, cover a range of two to three orders of magnitude. Although the coefficient of variation of the mean trip distance is much greater than 1.0 for all motorized surface transport modes, these modes cannot be easily substituted due to very different service characteristics (including speed) and human factors.

The energy use and intensity figures of the mechanized modes in Table II relate to the United States in 2000 and cannot reflect the enormous progress that was achieved in vehicle technology. Although the fundamental principles of surface vehicle and aircraft propulsion have remained unchanged over many decades, the performance of these systems and their integration into the vehicle have improved continuously. These improvements were compounded by ever further gradual reductions in all types of vehicle resistance, resulting in significant reductions in fuel consumption overall. An illustrative case is the automobile. If adjusting for changes in vehicle mass, fuel consumption per kilometer traveled and vehicle empty mass of the average new light-duty vehicle (LDV) sold have declined significantly—in the United States, from 8.2 to 5.5 L/100 ton-km between 1975 and 2000 alone. However, human factors did not allow a full translation of these gains into real reductions of transportation energy intensity. These and other barriers are described in the following section.

4. HUMAN FACTORS

Human factors, here defined as how technology is being used, generally include a wide range of behavior-related activities, from declining vehicle occupancy rates to the process of decision making when buying a new vehicle in a market offering models with a wide range of fuel consumption. This section briefly describes the most important components.

4.1 Declining Vehicle Occupancy Rates

In the United States, an average of 2.20 persons were sharing an automobile in 1969. That number declined to 1.90 in 1977, 1.75 in 1983, 1.64 in 1990, and 1.59 in 1995. Not taking into account secondary effects, that 28% decline in occupancy rate caused a 38% increase in passenger travel energy use, with all other factors being equal. Such continuous decline in occupancy rates was caused by the rising share of women in the labor force, the increase in vehicle fleets, and the decrease in average household size. These somewhat interrelated forces were strongest during the second half of the 20th century. Between 1969 and 1995 alone, the share of women in the labor force increased from 38 to 46%, the growth in the number of vehicles per household increased from 1.16 to 1.78 (reducing the need for sharing a vehicle), and average household size declined from 3.16 to 2.65 (reducing the opportunity for sharing a vehicle). Given that these trends' rate of change will diminish in future, vehicle occupancy rates will necessarily level off at a rate well above 1.0. A further factor that contributes to stabilizing occupancy rates is the inherently higher rate of leisure-based trips, the most rapidly growing trip purpose. Whereas automobiles carry only 1.2 occupants in work-related travel on average, the occupancy rate is 2.0 for leisure travel. These U.S. figures and trends are very consistent with those in other parts of the industrialized world.

Occupancy rates have also declined for buses and trains, due mainly to the shift toward faster and more flexible modes. For example, the average U.S. transit bus carried 12.9 passengers per bus in 1970, and that number declined to 9.2 in 2000. Occupancy rates have increased only in air traffic and currently account for approximately 70% in U.S. domestic travel. That comparatively high rate can be attributed to preferable service characteristics (high speed) in combination with declining fares (Fig. 1D).

4.2 Shift toward Larger and More Powerful Vehicles

The average 1975 model year U.S. LDV had an empty mass of 1.84 tons and an engine power of 102 kW, translating into a vehicle power/mass ratio of 55 kW/ton, and fuel consumption was 15.1 L/100 km. As a result of higher oil prices and tight fuel economy standards, the curb weight of the average new LDV declined to 1.43 tons and the power/mass ratio increased only slightly to 60 kW/ton by 1987, while fuel consumption declined by more than 40% to 8.8 L/100 km. However, with the onset of declining oil prices and the absence of a further tightening of fuel economy standards after 1987, curb weight rose to 1.66 tons, the power/mass ratio rose to 79 kW/ton, and fuel consumption rose to 9.2 L/100 km in 2000. The renewed growth in these performance indicators after 1987 is due mainly to the rising share of personal trucks, which already account for half of all LDVs sold in the United States today. In addition, shifts toward larger vehicles within the automobile segment further contributed to the growth in vehicle mass and engine power after 1987. These shifts can be attributed to consumer desires for increased safety, acceleration capability, and convenience (e.g., space, climate control, entertainment and navigation systems). Obviously, such an increase in consumer attributes, which has also been observed in many

other industrialized countries (albeit to a lesser extent), partly compensates the reduction potential in vehicle fuel consumption.

4.3 Other Factors

In addition to the decline in occupancy rates and the shift toward larger and more powerful vehicles, further human factors exist. Among those, perhaps the most significant one is driving behavior. Although little empirical evidence exists in that area, the shift toward more powerful vehicles generally allows more rapidly changing driving maneuvers that increase vehicle energy use per unit distance traveled. It is likely that a significant fraction of drivers adapt their driving style to the changing technology.

4.4 The Offsetting Effect of Human Factors

As a simplified illustration of the offsetting effect of human factors, energy intensity of the U.S. LDV fleet (the largest single energy consumer in the U.S. transportation sector) is decomposed into reductions in vehicle fuel consumption, relative growth in vehicle size classes, and declining occupancy rates. LDV energy intensity can be formulated as follows:

$$\left(\frac{E}{pkm}\right)_{LDV} = \sum_i \frac{E_i}{vkm_i} \cdot \frac{vkm_i}{vkm} \cdot \frac{vkm}{pkm}, \qquad (2)$$

where, E_i/vkm_i is vehicle energy use of automobiles and light trucks, vkm_i/vkm is the share of each of the two segments in kilometers driven, and vkm/pkm is the inverse average occupancy rate, assumed to be identical for all LDVs. Differentiating Eq. (2) and reorganizing the terms results in the percentage change of LDV energy intensity

$$\left[\frac{d\left(\frac{E}{pkm}\right)}{\frac{E}{pkm}}\right]_{LDV} = \sum_i \frac{d\left(\frac{E}{vkm}\right)_i}{\left(\frac{E}{vkm}\right)_i} \cdot \frac{E_i}{E} + \sum_i \frac{d\left(\frac{vkm_i}{vkm}\right)}{\frac{vkm_i}{vkm}} \cdot \frac{E_i}{E} + \frac{d\left(\frac{vkm}{pkm}\right)}{\frac{vkm}{pkm}}. \qquad (3)$$

The application of Eq. (3) to the U.S. LDV fleet from 1970/1971 to 1999/2000 is depicted in Fig. 5. During the entire 31-year period, the change in human factors (here explicitly the relative growth in travel demand and decline in vehicle occupancy rates) was greater than zero; thus, human factors at least partly offset LDV fuel efficiency improvements. Until 1978, human factors more than offset fuel efficiency

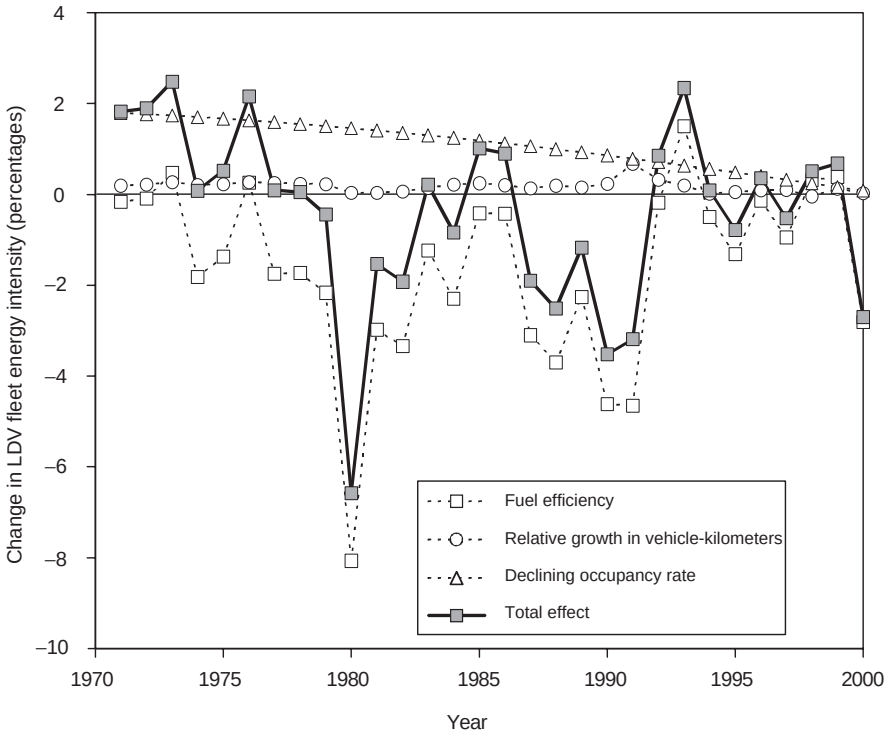


FIGURE 5 Annual percentage change in light-duty vehicle (LDV) energy intensity as the total of changes in technology improvements, shift in vehicle segments, and occupancy rate.

improvements occurring in the vehicle fleet, raising its average energy intensity. Because of drastic vehicle fuel efficiency improvements caused by the combination of the second oil shock and tight fuel economy standards, average energy intensity declined between 1979 and 1984. However, during that period, human factors offset that decline in energy intensity by more than 1%. Starting in 1985, LDV energy intensity declined only gradually, increasingly due to only modest improvements in energy efficiency.

5. PASSENGER TRAVEL ENERGY USE AND EMISSIONS

Figure 6 reports total passenger travel energy intensity (all modes) for France, Germany, and the United States over time. In France and Germany, where passenger travel energy intensity has remained roughly constant, improvements in fuel efficiency have been completely offset through a combination of human factors and shifts toward faster and more energy-intensive transportation modes. In contrast, in the United States, passenger travel energy intensity declined from 3.2 MJ/pkm in 1977 to 2.5 MJ/pkm in 1991, due mainly to drastic LDV and passenger aircraft fuel efficiency improvements. Multiplying average passenger travel energy intensity by growth in passenger-kilometers leads to total passenger transport energy use. Thus, in light of a nearly 90% increase in passenger traffic volume, U.S. passenger travel energy use grew from 10.4 EJ in 1970 to 17.8 EJ in 1990, an increase of more than 70%.

On a global scale, total transportation final energy use grew from 35 EJ in 1971 to 75 EJ in 2000, with approximately 60% of that amount corresponding to passenger travel. In contrast to criteria pollutants (e.g., carbon monoxide, lead, nitrogen oxide, ozone, particulate matter, sulfur dioxide), which can be decoupled from energy use through adjustments in fuel composition, engine combustion settings, and exhaust gas catalysts, carbon dioxide emissions are strongly linked to energy use via the fuel's carbon content. Thus, because oil products supplied approximately 97% of transportation energy, passenger travel carbon emissions increased in the same proportion as did energy use, that is, from approximately 0.4 billion tons to 0.9 billion tons.

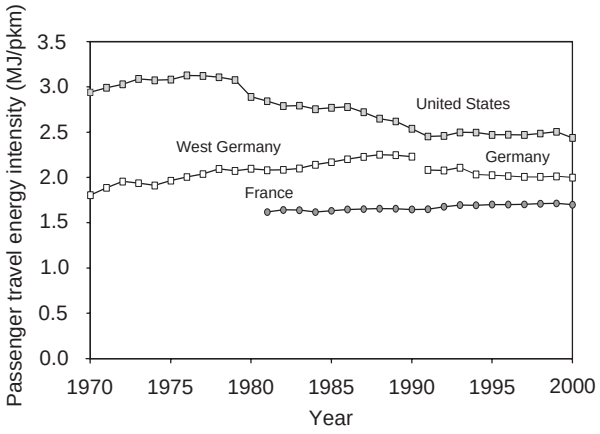


FIGURE 6 Evolution of total passenger travel energy intensity in France, Germany, and the United States.

6. REDUCING PASSENGER TRAVEL ENERGY USE AND EMISSIONS

According to Eqs. (1) and (2), passenger travel energy use can be controlled through improving energy efficiency and shifting toward less energy-consuming modes (reducing *E/vkm*) and reducing vehicle-kilometers traveled (reducing *vkm*). Various kinds of emissions, most notably greenhouse gases (GHGs), can be further reduced through burning less carbon-containing fuels or through burning zero-carbon fuels.

6.1 Reducing Vehicle-Kilometers Traveled

Transportation demand management (TDM) aims at reducing passenger travel through increasing the use of mass transit systems, increasing vehicle occupancy rates, substituting travel by electronic means, and reducing travel needs by land use planning. In addition, TDM includes a range of administrative measures, including more flexible work schedules, auto-restricted zones, and parking management. Common to most of these measures individually is their small potential, resulting in part from fundamental travel behavior and societal countertrends. The small potential of TDM is also partly inherent; electronic substitutes offer only a small reduction potential in traffic and associated energy use, even if assuming a relatively high percentage of the labor force working a significant amount of the time at home. (If 20% of the entire labor force worked an average of 2.5 days per week at home, work travel would be reduced by only 10%, and in light of the small 20% share of work trips in the United States, total vehicle trips would be reduced by a mere 2%.) Although the potential of each individual measure is limited, a sensible bundle of measures promises a significantly larger impact on travel and energy use.

Such packages of measures will be most effective if they contain pricing instruments that directly charge travelers for using transportation infrastructure, for example, through time- and day-dependent measures (e.g., road pricing) or rigid measures (e.g., variabilization of fixed costs, i.e., reducing vehicle taxes and/or insurance payments but increasing fuel costs). Unfortunately, such direct measures often meet public resistance and, thus, require long time periods for implementation.

6.2 The Promise of Technology

In the past, technology change has proven to be the most promising tool for reducing energy use and emissions. However, because human factors in combination with the growth in total travel demand can easily offset gains achieved by more fuel-efficient technology, noticeable reductions in passenger travel energy use require radical technology change. Such drastic change was offered by the three-way catalyst that has reduced vehicle tailpipe emissions of carbon monoxide, unburned hydrocarbons, and nitrous oxide by two orders of magnitude. Because reductions of that magnitude are not possible for energy use (given that a large share of a country's population would have to move on foot or by bike, as can be seen from Table II), noticeable reductions in energy use also require appropriate policy measures not only to push more fuel-efficient technology into the market but also to mitigate other human factors and demand growth at least in part.

Despite significant technological progress that has already occurred in improving fuel efficiency of all transport systems, a substantial potential in reducing transport sector energy use still exists. Perhaps the largest potential is offered by LDVs, which currently consume approximately 80% of U.S. passenger transport energy use. LDV fuel efficiency can be increased by increasing the efficiency of the drivetrain (engine plus transmission) and by reducing the amount of energy necessary to move the vehicle (driving resistances). Although the single largest reduction in energy use is expected to come from the drivetrain, the largest total potential will result from a combination of drivetrain measures and a reduction in driving resistances. Following such a holistic approach, individual energy savings potentials are compounded. If maintaining consumer attributes of current vehicles, fuel consumption can be reduced by 50 to 70%. Slightly lower potentials for reducing fuel consumption exist for urban buses and, over the longer term, for passenger aircraft.

6.3 Barriers

In addition to the human factors discussed previously, there are several barriers to reducing passenger travel energy use. One barrier reflects the higher retail price of more fuel-efficient transport technology. In the case of LDVs, this retail price increase can be drastic. If dividing into thirds the difference in fuel consumption between the average new vehicle on the road today and a highly fuel-efficient vehicle consuming 3 to 4 L/100 km, the first third of the potential for reducing energy use comes at a comparatively low cost. Whereas the second third already requires an increase in the vehicle retail price by up to 20%, the final third comes at a very high price (up to 30%). Because empirical evidence suggests that consumers may be willing to amortize extra costs for a more fuel-efficient vehicle over a few years only, penetrating highly fuel-efficient technologies into the market requires government action such as higher gasoline prices, incentives for purchasing more fuel-efficient technology, and some kind of regulation.

However, even if tight policy measures force more fuel-efficient technology into the transportation sector, increasingly slow fleet turnover significantly delays the translation of new vehicle fleet energy use to that of the entire vehicle fleet. Along with the above-discussed decline in vehicle energy use per ton-kilometer, continuous technological progress increased the average vehicle lifetime from 11 years for a 1966 model to 15 years for a 1990 model. (As a result, the median age of the U.S. automobile fleet nearly doubled from 4.9 years in 1970 to 8.8 years in 2000.) Thus, the associated time for a (nearly) complete phase-out of the entire vehicle fleet increased from 18 years (1966 fleet) to 21 years (1990 fleet). In light of these increasingly long time scales, reductions in fuel consumption need to be carried out rather soon to experience midterm effects.

7. OUTLOOK

Taking into account the previously discussed determinants of passenger demand for travel and energy use as well as the barriers for introducing more fuel-efficient vehicles, passenger travel energy use will continue to increase worldwide. Growth will be largest in those parts of the developing world where income and population increase most rapidly. Building on a continuation of the illustrated trends in travel demand, mode choice, and energy intensity, one recent study projected world passenger travel demand

to rise by a factor of 4.5 through 2050. During the same period, energy use and carbon dioxide emissions would increase by a factor of 3.5. To prevent such drastic growth in energy use and GHG emissions, technology for drastically reducing passenger travel energy use must be made available on a large scale during the coming years. Because of the extra investments for fuel-saving technologies and the high consumer discount rate, their large-scale introduction requires government action. The long turnover rates of transportation technology require such reductions to be implemented rather soon to experience noticeable reductions in energy use and (cumulative) GHG emissions during the coming decades. However, whether technology change alone can be the ultimate answer to reducing passenger travel energy use and GHG emissions remains to be seen.

SEE ALSO THE FOLLOWING ARTICLES

Alternative Transportation Fuels: Contemporary Case Studies • *Bicycling* • *Intelligent Transportation Systems* • *Internal Combustion Engine Vehicles* • *International Comparisons of Energy End Use: Benefits and Risks* • *Leisure, Energy Costs of* • *Lifestyles and Energy* • *Motor Vehicle Use, Social Costs of* • *Transportation and Energy, Overview* • *Vehicles and Their Powerplants: Energy Use and Efficiency*

Further Reading

Davis, S. C. (various years). "Transportation Energy Data Book." Oak Ridge National Laboratory, Oak Ridge, TN. www-cta.ornl.gov/cta/data/index.html.

Greene, D. L. (1996). "Transportation and Energy." Eno Transportation Foundation, Washington, DC.

Rennie, J., *et al.* (eds.). (1997). The future of transportation [special issue]. *Sci. Am.* **277** (4).

Schafer, A., and Victor, D. (2000). The future mobility of the world population. *Trans. Res. A* **34**(3), 171–205.

Hellman, K. H., and Heavenrich, R. M. (2003). "Light-Duty Automotive Technology and Fuel Economy Trends: 1975 through 2003." EPA 420-R-03-006. U.S. Environmental Protection Agency, Washington, DC.

Peat Resources

ANNE JELLE SCHILSTRA
University of Groningen
Groningen, The Netherlands
MICHIEL A. W. GERDING
Historian, Province of Drenthe
Assen, The Netherlands

1. Peat: A Short Introduction
2. Peat as an Energy Source
3. Environmental Concerns
4. The History of the Use of Peat for Fuel

Glossary

bog (raised) A mire raised above the surrounding landscape and fed only by precipitation.
boreal Biogeographical zone between the temperate and the subarctic zones.
fen Peatland, which in addition to precipitation, receives water that has been in contact with mineral soil or bedrock.
mire A peatland where peat is currently being formed and accumulating.
peat Vegetable matter accumulated and preserved in anaerobic waterlogged areas.
turf The material noun for dried peat.

Peat is the dead vegetable matter that has accumulated in waterlogged anaerobic layers over periods of often thousands of years. The slow diffusion of oxygen in water limits the decay of organic matter so that the cycle of carbon fixation by photosynthesis and usual decomposition of the material after dying is incomplete. The mires or peatlands where peat is formed are quite extensive in Canada, Alaska, Siberia, northern Europe, and some tropical countries. In boreal areas especially, *Sphagnum* species contribute to peat formation, increasing the thickness of the peat layer annually by 0.2–4 mm. In the tropical peatlands, rainforest trees produce the litter and roots that may accumulate in permanently waterlogged forest floors.

1. PEAT: A SHORT INTRODUCTION

A traditional distinction is made between bogs and fens: bogs rise above their surroundings and are fed by rainwater only. Due to the capillary properties of peat, the water level can be drawn up so that soggy lens-shaped domes hundreds of meters wide and up to 12 m deep can be formed. By weight, the peat from bogs may consist of 98% water and only 2% solid matter. Peat that is formed in shallow lakes and that may eventually fill them completely is usually called a fen. In suitable circumstances, the peat blanket can grow sideways and can eventually even cover nearby hills. The solid matter may constitute approximately 10% of the peat. See Fig. 1. Undisturbed peatlands are ecosystems where the wet environment and the lack of nutrients have led to the evolution of a characteristic specialized flora and fauna.

Boreal peatlands may date from the end of the last ice age, approximately 10,000 years ago. But obviously the upper parts and the fringes can be of quite recent origin. As a high groundwater level is crucial for the continuing existence of peatlands, interference with the water level has dramatic consequences. Drainage will end the accumulation of new peat, but in contrast may enhance the growth of other plants; for example, trees that remain rather stunted on intact peatlands due to the shallowness of the aerobic top layer (the acrotelm) can grow normally when, after drainage, the top of the formerly anaerobic layer (the catotelm) is oxygenated. Another consequence of drainage is that the layers now freshly exposed to oxygen will start to decompose. Whereas the methane (CH_4) emissions that characterize many healthy mires are halted, the emission of carbon dioxide (CO_2) greatly increases: the mire has changed from a net carbon sink to a

 © 2004 Elsevier Inc. All rights reserved.

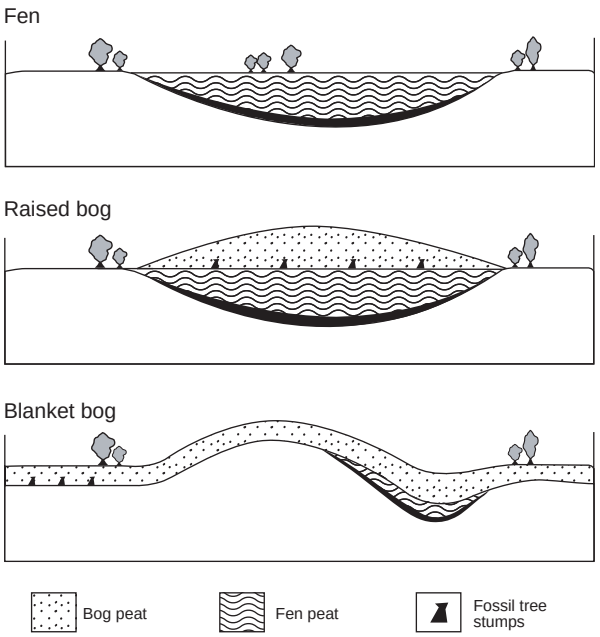


FIGURE 1 A fen can develop into a raised bog. Only precipitation feeds the living surface. A bog may also grow sideways and cover nearby hills. The bottom of the bog may be thousands of years old.

carbon source. When forest grows on a drained peatland, it will temporarily fixate CO_2 faster than the peat underneath oxidizes. But eventually, when the forest approaches maturity and the net uptake of CO_2 diminishes, CO_2 emissions from the peat will out-compete the forest uptake.

The question of whether or not peat is a fossil fuel has been around for some time. Much younger than brown coal or lignite, peat is rather "old" when compared to renewable energy sources, such as wood from production forests. Designating peat as a renewable or fossil fuel may have implications; the power it generates may be labeled "green" or not. It can convincingly be argued, however, that peat is neither a fossil fuel nor a renewable energy source. Some agreement has been found by labeling it a "slowly renewable biofuel."

2. PEAT AS AN ENERGY SOURCE

Estimations on the total area of peatlands are rather varied, in part because countries use different definitions of peatlands. A rough estimation would be almost 4 million km^2 (3% of the earth's land area), not including the more than 2.4 million km^2 of terrestrial wetlands. Canada, Alaska, Scandinavia, and Russia have huge boreal peatlands and in Indonesia and South America extensive tropical peatlands can be found. Estimations of the amount of carbon stored in peatlands also varies, but a safe guess would be 270–370 Gt, (1 Gt = 1 gigaton = 10^9 tons), approximately one-third of the global soil (nonfossil), carbon store or two-thirds of the atmospheric carbon pool. Annual addition to this carbon store is approximately 40–70 million tons, small compared to the 100 Gt that is exchanged annually worldwide between the atmosphere and the terrestrial ecosystems. In Europe and Africa, more than one-half of the original peatlands have been used or destroyed, whereas in the Netherlands, for example, only 1% is left. Agriculture (50%), forestry (30%), and peat extraction (10%) are the main causes for global anthropogenic peatland losses.

In Fig. 2, some properties of peat and other fuels are depicted. As can be seen, the heating value of dry peat is between those of wood and brown coal. The water content of peat is an important factor for its practical heating value. Peat taken from undrained peatlands may have a water content like that of a very wet sponge. Preparing peat for industrial use then amounts largely to the efficient drying of the substrate. Peatlands to be harvested are usually drained for several years by parallel ditches 20 m apart, cleared of growth, and leveled. The soil is then roughened for a couple of centimeters and left to dry in the wind and the sun. After a suitable time, the loose peat is collected by giant pneumatic harvesters or brushed into ridges, after which the next layer is prepared to dry. After additional drying when necessary, the peat is moved by truck or narrow-gauge railway to a collecting area or directly to the power plant. Clearly, the annual harvest strongly depends on the weather; a given area may yield in a good year several times more peat than in a bad year. For example, in 1997 Finland harvested 10.1 million tons of peat for energy production but in 1998 it harvested only 1.5 million tons.

In Finland, Ireland, and Sweden, peat is also used for strategic reasons. With little or no local fossil fuel deposits, fuel diversification that includes local sources seems important. Also, in remote areas where peat is easily available, it is sometimes the obvious choice for an energy source. An additional argument for the use of peat for power generation is found as a means to stimulate local economy and labor opportunities. The Edenderry peat electricity plant (127 MW) in central Ireland is an example of this policy. Modern large-scale power plants are based on fluidized bed combustion technology that

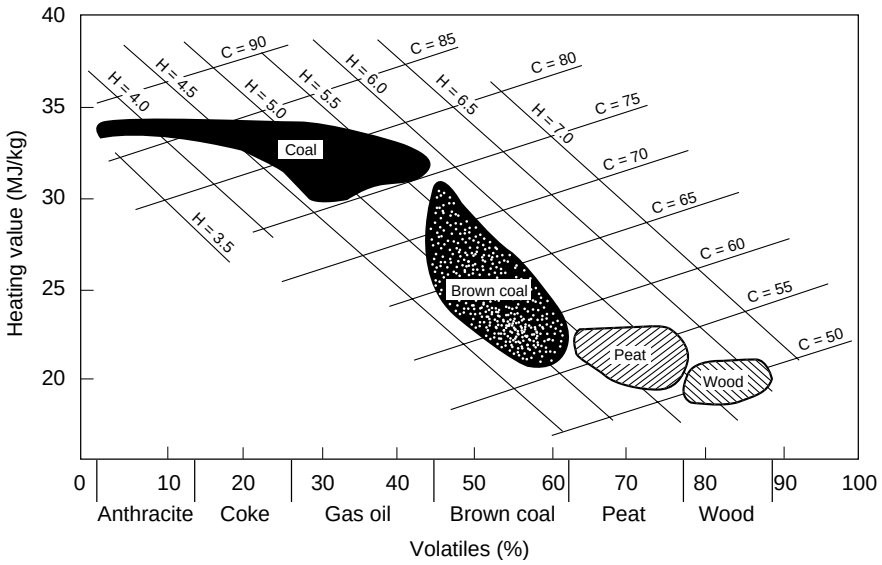


FIGURE 2 Fuel properties of dry peat compared to other fuels. Carbon and hydrogen content by weight is indicated by the sloping lines. Natural gas (CH_4, by weight 75% carbon and 25% hydrogen) lies outside the frame.

results in higher efficiencies, lower emissions, and the capability to include other fuels such as wood chips.

Finland, Ireland, and Sweden are the only industrialized countries that use peat with a state of the art technology. In Ireland, 11% of the electrical power is generated by peat-fueled stations; in Finland, approximately 6% of the electrical power is generated by peat-fueled stations, and wood is often co-fired as well. In Sweden, less than 1% of the electrical power is generated by peat-fueled stations.

3. ENVIRONMENTAL CONCERNS

The use of peat for energy has a number of environmental problems in common with the use of fossil fuels. Of global concern is the emission of CO_2. Plants that are the source of renewable biofuels recycle atmospheric CO_2 so that solar energy can in principle be harvested and used in a carbon-neutral way. Growing and harvesting peat similarly would require a cycle length of thousands of years, quite impractical. Nor is the fact that other living peatlands continue to accumulate atmospheric CO_2 faster than peat is converted to CO_2 in power stations a valid argument for "sustainability." The sulfur content of peat is low but varies; 0.2% by weight of dry matter is a representative value. Local effects are extensive. Harvesting of peat is akin to open-cast mining. After harvesting, which can take 20 years, the remaining area can be put to good use, for rewetting and regrowth of peat, for forestry, or for agriculture. In the Netherlands, for example, in major parts of the west of the country, agriculture is actually "after use" of peatlands where most peat was removed centuries ago. Only small areas where accumulating peat is left (a total of $\sim 10\,km^2$) have to be managed and protected in order to keep its natural values from further deteriorating. Largely due to the oxidizing of the former peat cover, approximately 26% of this country lies below sea level and 66% would be inundated regularly without the protection from dikes and dunes. The harvesting of peat as a natural resource generations ago has as a consequence that the Dutch today must face the risk of flood disasters and the costs of preventing them.

4. THE HISTORY OF THE USE OF PEAT FOR FUEL

Fens and bogs, peatlands, are in the words of Julius Caesar "land that can not be walked upon and water that can not be sailed on." The fact that it can be used for fuel was first recorded by Plinius the Elder, who, in his *Naturalis Historia* (77 AD), mentions people who dug lumps of earth that they dried in wind and sunshine and then set fire to in order to cook their food and warm their houses. He writes about a people living in the delta of the river Rhine, now the Netherlands. Apart from this singular quote, almost nothing is recorded about the use of peat for

fuel until well into the Middle Ages, and then only from the Low Countries (Flanders and the Netherlands). Yet millions of hectares of peatlands existed north of the Alps.

In the history of the use of peat for fuel, it is necessary to make a distinction between peat digging for domestic use only and peat production on a larger scale in a businesslike operation intended for larger markets. This section concentrates on the latter. Outside the Netherlands, the literature on peat and its history appears to be very scarce or virtually nonexistent. A comprehensive overview can be given only of the Low Countries.

The reason for the widespread use of peat for fuel in the Netherlands and Flanders lies in the lack of firewood on the one hand and the abundance of waterways (rivers and canals) on the other hand. The relatively low cost of transportation of the bulky material combined with the engineering skills of the Dutchmen in controlling the water and managing waterways were crucial in the proliferation of the use of peat for fuel. From the early Middle Ages until well into the 20th century, peat (turf) was *the* source of "fossil" energy in the Netherlands.

The Flanders cities of Antwerp, Ghent, and Brugge were the first to invest in large-scale peat production from ca. 1250 AD onward. From there it spread northward, so that in the 14th century the fens in a large part of Holland, the most important province of the Netherlands (with cities such as Amsterdam, Leiden, Delft, and Haarlem), were brought into production. A crucial step in the development was the introduction in 1513 of the "dredger" (bagger-beugel), a bucket on a long wooden pole of approximately 4 m. Now it was no longer necessary to drain the fens first before the peat could be cut and dried. The wet peatmud could be dredged directly from the water and stored in small boats and then spread out on nearby fields to be cut and dried. Production soared and coincided with the strong economic and demographic growth of the Netherlands in the 16th and 17th centuries ("Golden Age"). It has been argued that the Dutch Golden Age was born from peat. If the total energy supply had relied on wood, one-quarter of the present area of the country would have to have been covered completely and permanently with production forest. Furthermore, the absence of waterways would have required approximately 100,000 horses for the transport of fuel and a further 600,000 ha of cultivated area to feed them. Thus, the energy attainability gave the Dutch Republic a tremendous advantage.

The dredging of low-peat turf had one major disadvantage: the loss of land was enormous. Over time, lakes replaced pastures and the unstable banks threatened the nearby countryside. Strict regulations were issued and taxes were levied to form funds from which new polders could be formed. This, and the demand for energy, caused the Dutch merchants to look elsewhere in the country. In the northeastern part of the country, 140,000 ha of good-quality black peat was still available. From the middle of the 17th century onward, this was the most important production area. In contrast to the dredging of low-peat turf, the high-peat areas had to be drained first. Thus, an infrastructure of waterways had to be build in order to drain the bogs and thereafter to transport the turf. This required large investments, which the rich merchants of Holland had in abundance. Also, local nobility and city and regional governments took part in the peat companies ("Veencompagnieën"), which were in effect mainly canal companies. Most of the work in the peateries itself was performed by subcontractors and private entrepreneurs (Fig. 3).

The contribution of peat from the north of the country to the national energy supply is a good example of the importance of peat in the Dutch energy supply. The darker areas are former peat bogs, the lighter areas are fens that have been excavated since the 16th century.

In the period between 1550 and 1950, peat digging played an important role in the northern parts of the Netherlands. At the outset, this was limited, but from the 17th century onward, its role increasingly grew in importance. During these 400 years, approximately 100,000 ha of raised bog peat and 42,000 ha of fenland peat was converted to dry peat fuel. Approximately 16.6×10^{11} turfs were dug, representing a combustion value of 3.5×10^{9} GJ, which is comparable with 8300 million tons. This peat was excavated from 62 raised bog areas and 11 fenland areas. An average depth of 2 m (12 layers) of winnable peat was obtained.

The most important reason for the long-term development of production is the growth in population. Until 1850, peat was by far the most important source of energy for just about all sectors of society, not just for domestic needs such as cooking and heating, but also for crafts and industry, especially those industries directly linked to the growth in population such as brick and tile production, brewing, baking, and distilling. Until 1650—it can be assumed—when peat played an important role in the export industry, production was not directly related

FIGURE 3 Peat-containing soils in the Netherlands. Most of the country was covered by peat in Roman times. Ages of use have removed most of the peat, leaving a country where 26% of its area lies below sea level. The darker areas are former peat bogs, the lighter areas are fens that have been excavated since the 16th century.

to the internal growth of the population. From the 19th century onward, coal began to assume a more important role in the satisfaction of energy requirements, but until at least World War I the growth in population was so strong that peat continued to prosper. Thereafter no longer. The continuous growth from 1750 onward was over and peat began to play an increasingly marginal role, leading to a complete dependence on coal. Nevertheless, peat production during the interwar years, in addition to production in many crisis years, underwent many years of reasonable growth, especially in the second half of the 1930s. After a temporary boom during World War II, the role of peat was completely finished by 1950.

4.1 Other Countries: Germany

Dutch peateries were taken as an example by the German governments of the northwest part of the country, especially Niedersachsen, where enormous amounts of peat bogs were to be found also. Yet the German initiatives never reached the scale of the Dutch. But the cities of Bremen, Oldenburg, and Emden in the 18th and 19th centuries managed to be self-sufficient in their energy supply because they could benefit from nearby peat deposits. The city of Hamburg lacked these advantages and constituted an important market for turf from Ostfriesland and Groningen. The oldest peat colony in Germany was Papenburg, founded in the 17th century. Other colonies followed, but never on the grand scale of the Dutch examples. Instead, Germany had in the 18th and 19th centuries a policy of erecting "Moorkolonien," settlements of paupers from the cities on the edge of bog areas who were forced to become small subsistence farmers.

4.2 British Isles

Although many peat bogs and fens were to be found in England and especially in Scotland, peat digging never developed on a businesslike scale in these

countries. Late in the 19th century, reports were published to promote such activities, for instance, the erection of electricity plants in Scotland fueled by peat, but they never came into being.

4.3 Ireland

In Ireland, the case was quite different. The country has a long tradition of using peat for fuel. There is evidence that it was already used as such in prehistoric times. Peat cutting increased in the 17th century due to the disappearance of forests and turf was the primary source of energy for country dwellers as well as city folk. Production reached its peak in the first half of the 19th century before the great potato famines of the 1840s. Until the early decades of the 20th century, all of the production was manual and carried out by individual turf cutters, who sold their product to boat-owners who transported the turf to Dublin by way of the Grand Canal, which was extended through the great Bog of Allen in 1786. In the west also, much of the turf was carried to the towns by boat. Many people in the Irish countryside as well as in the cities had individual rights to cut turf for their own domestic use, although they were not allowed to sell it. In 1912, 15,000 tons of turf was brought to Dublin on boats that could carry turf loads of 30–60 tons. In the first four decades of the 20th century, cooperative turf societies were established in order to promote the production and harvesting of hand-cut turf and to facilitate its direct sale by the producers. During World War II, turf became of strategic importance when coal supplies dried up and the entire fuel needs of the country depended on turf. In these years, thousands of Dubliners made their way into the Dublin mountains to cut their own peat.

After the war, the Bord na Móna was formed and took over all turf-cutting schemes at the government's request. The company developed into the world's second largest peat company. Most of the fuel peat goes to power plants in the country, which take up approximately 80% of the total peat production. Between 1946 and 1995, Bord na Móna produced ca. 170 million tons of peat for fuel.

4.4 North America

Although a great deal of Canada consists of peatlands, peat for fuel has never been an issue there or in the United States either. Peat is harvested only for horticultural purposes. The same can be said about Central and South America.

4.5 Baltic and Scandinavia

In most of the Baltic states, peat was used for fuel on a very small scale and then only for domestic use. Businesslike activities such as those in Germany, Ireland, and the Netherlands were absent. Peat production for horticultural purposes was, and still is, much more widespread. Only in Finland has there been production of peat for fuel since the 1970s. Since the oil crisis of 1973, the Finnish government labeled peat as a strategic good. In the middle of the country, a power station is fueled by peat.

4.6 Russia

In Russia, formerly the largest user of peat for energy, peat consumption represents only 0.3% of the primary energy source.

SEE ALSO THE FOLLOWING ARTICLES

Biomass Combustion • Biomass for Renewable Energy and Fuels • Coal Resources, Formation of

Further Reading

de Zeeuw, J. J. (1978). Peat and the Dutch Golden Age: The historical meaning of energy-attainability. A.A.G. Bijdragen, Vol. 21, pp. 3–33. University of Wageningen, Wageningen, the Netherlands.

Gerding, M. A. W. (1995). Vier eeuwen turfwinning (Four ages of peat digging). A.A.G. Bijdragen, Vol. 35, University of Wageningen, Wageningen, the Netherlands.

Joosten, H., and Clarke, D. (2002). "Wise Use of Mires and Peatlands: Background and Principles Including a Framework for Decision-making." International Mire Conservation Group and International Peat Society. Jyväskylä, Finland.

Lappalainen, E. (1996). "Global Peat Resources." International Peat Society, Jyskä, Finland.

Schilstra, A. J. (2001). How sustainable is the use of peat for energy production? *Ecol. Econ.* **39**, 285–293.

Petroleum Property Valuation

JAMES L. SMITH
Southern Methodist University
Dallas, Texas, United States

1. **Impact of Property Valuations on the Petroleum Industry**
2. **Discounted Cash Flow (DCF) Analysis: The Problem Simplified**
3. **Special Characteristics of Petroleum Properties**
4. **Incorporating the Value of Real Options**
5. **Portfolio Analysis: The Combined Value of Multiple Properties**
6. **Conclusion**

Glossary

capital asset Any equipment, facility, or plant capable of generating a long-lived stream of future income.

discounted cash flow (DCF) A method for estimating the present value of future cash flows that adjusts for the time value of money and the degree of uncertainty surrounding future receipts.

discount rate The factor by which expected receipts in a future period are reduced to reflect the time value of money and unpredictable variation in the amount ultimately received.

diversifiable risk A source of unpredictable financial performance that varies randomly from one investment to another and therefore averages out, rather than accumulates, over all of an investor's holdings.

fair market value The price expected to be paid for any asset in a voluntary exchange between an independent buyer and independent seller.

Monte Carlo analysis A method for assessing the magnitude and implications of risk by simulating possible outcomes via random sampling from the probability distribution that is assumed to control underlying risk factors.

net present value (NPV) A measure of the value of a project obtained by discounting the stream of net cash flows (revenues minus expenditures) to be received over the entire life of the project, based on the time profile and riskiness of net receipts.

nondiversifiable risk A cause of unpredictable financial performance that tends to impact all of an investor's holdings in the same direction or manner.

proved reserves The volume of petroleum resources in a developed field that are reasonably expected to be recoverable given current technology and prices.

real options The general phenomenon by which the value of physical assets depends on, and therefore may be enhanced by, management's ability to modify or postpone investment and operating decisions based on the receipt of new information.

risk premium That portion of the discount rate that compensates the investor for the inherent unpredictability of future financial returns, as opposed to the pure time value of money.

Petroleum property valuation refers to the analytical procedure by which the commercial value of oil and gas fields is assessed. This assessment provides to prospective buyers and sellers, and other interested parties such as lenders and tax assessors, an estimate of the fair market value of underground deposits of oil and gas—the amount for which they might be bought or sold. The value of an underground deposit is directly related to the ultimate value of whatever petroleum may be extracted in the future, but because the future is uncertain, the value of the property is subject to various sources of risk that stem from geological as well as economic factors. To be useful, the valuation procedure must take proper account of the unpredictable fluctuations that would cause field development, operations, and performance to deviate from the expected outcome. This task represents a complex analytical problem that has challenged traditional valuation methods. To meet the challenge, new and highly innovative techniques have been developed and are finding widespread use.

 © 2004 Elsevier Inc. All rights reserved.

1. IMPACT OF PROPERTY VALUATIONS ON THE PETROLEUM INDUSTRY

The petroleum industry's reliance on accurate and reliable valuation methods is apparent. The risk of paying too much for an acquired property, or selling for too little, is always present. The frequency and size of property transactions underscores the importance of getting the valuations right. Since 1979, for example, more than 5000 parcels of existing oil and gas fields have been sold in the United States, bringing more than $600 billion of revenue to the sellers. The negotiations that support these exchanges hinge on finding valuation methods that both sides deem acceptable. In addition, investments are made by the petroleum industry to acquire leases in raw acreage where no oil or gas field is known to exist. The U.S. government is a major source of such properties and has raised in excess of $60 billion since 1954 (when the federal leasing program was initiated) by auctioning petroleum exploration and development rights on federally owned lands and the outer continental shelf. The petroleum industry makes comparable investments on a regular basis to acquire oil and gas leases from private landowners and the states, as well. The ability to value petroleum properties accurately therefore plays a critical role in determining the financial success or failure of oil and gas producers.

2. DISCOUNTED CASH FLOW (DCF) ANALYSIS: THE PROBLEM SIMPLIFIED

In some respects, an oil field is no different than any other capital asset and valuation techniques for petroleum properties are therefore similar to procedures used in other sectors of the economy. A capital asset represents any long-lived investment in productive facilities that have the potential to generate a future stream of earnings. If those earnings are not large and predictable enough to justify the initial expenditure, the investment should not be made. Intuitively, the value of the capital asset may be thought of as the extent to which anticipated cash receipts outweigh the initial expenditure. Measuring and weighing the projected cash flows therefore forms the heart of the valuation problem.

2.1 Projecting Cash Flows

The projected cash flow stream is a composite forecast that results from many separate assumptions concerning physical attributes of the oil field and the economic environment in which it will be produced. The number of wells and size of facilities required to delineate and develop the field, in conjunction with the presumed cost level for drilling services and oil field equipment, will roughly determine the scope and timing of initial expenditures. The magnitude and duration of cash inflows (sales revenue minus operating cost) are determined by a further set of assumptions regarding the flow rate from individual wells (and the rate at which production will decline as the field is depleted), the quality and price of produced oil and gas, necessary operating and maintenance costs required to keep wells and field plant facilities in order, and the level of royalties and taxes that are due to lessors and governmental authorities. Thus, the projection of net cash flow for the field as a whole is the combined result of many interrelated but distinct cash flow streams. Some components are fixed by contract and can typically be projected with relative certainty (e.g., royalty obligations and rental payments), but others require trained guesswork that leaves a wide margin of error (e.g., future production rates and oil price trends). It seems reasonable that those components of the cash flow stream that are known with relative certainty be given greater weight in figuring the overall value of the field, but as discussed further later in this article, properly executing this aspect of the valuation procedure was hardly practical until so-called options-based valuation methods were developed.

2.2 Discounting Cash Flows

A dollar received (or paid) in the future is worth less than a dollar received (or paid) today because of the time value of money. Cash in hand can be invested to earn interest, and therefore will have grown in value to outweigh an equivalent amount of cash to be received at any point in the future. If the relevant periodic rate of interest is represented by the symbol i (e.g., $i = 10\%$), then the present value of a dollar to be received t periods hence is given by $PV(i, t) = 1/(1+i)^t$. This expression is referred to as the discount factor, and i is said to be the discount rate. The discount factor determines the relative weight to be given to cash flows received at different times during the life of the oil field. Cash flows to be received immediately are given full weight, since

$PV(i, 0) = 1$, but the weight assigned to a future receipt declines according to the amount of delay. Thus, the net present value (NPV) of an arbitrary cash flow stream represented by the (discrete) series of periodic receipts $\{CF_0, CF_1, CF_2,..., CF_T\}$ is computed as the sum of individual present values:

$$NPV = \sum_{t=0}^{T} \frac{CF_t}{(1+i)^t}. \quad (1)$$

It is quite common to perform this computation on the basis of continuous discounting, where the periodic intervals are taken to be arbitrarily short (a day, a minute,..., an instant), in which case the discount factor for cash to be received at future time t declines exponentially with the length of delay: $PV(i,t) = e^{-it}$. Therefore, when the cash flow stream is expressed as a continuous function of time, NPV is reckoned as the area under the discounted cash flow curve:

$$NPV = \int_0^T CF_t \cdot e^{-it} dt. \quad (2)$$

It is apparent, whether the problem is formulated in discrete or continuous time, that correct selection of the discount rate is critical to the valuation process. This parameter alone determines the relative weight that will be given to early versus late cash flows. Since exploration and development of oil and gas fields is typically characterized by large negative cash flows early on, to be followed after substantial delay by a stream of positive cash flows, the choice of a discount rate is decisive in determining whether the value of a given property is indeed positive. The extent to which discounting diminishes the contribution of future receipts to the value of the property is illustrated in Fig. 1, which shows the time profile of cash flows from a hypothetical oil field development project. With no other changes to revenues or expense, the property's net present value is reduced by a factor of ten, from nearly $1 billion to roughly $100 million, as the discount rate is raised from 8% (panel a) to 20% (panel b). These panels also illustrate how discounting affects the payback period for the property in question (i.e., the time required before the value of discounted receipts finally offsets initial expenditures): 7 versus 11 years at the respective rates of discount.

With so much at stake, the selection of a discount rate cannot be made arbitrarily. If the discount rate is not properly matched to the riskiness of the particular cash flow stream being evaluated, the estimate of fair market value will be in error. Because no two oil fields are identical, the appropriate discount rate may vary from property to property. A completely riskless cash flow stream (which is rare) should be discounted using the risk-free rate, which is approximated by the interest rate paid to the holders of long-term government bonds. Cash flow streams that are more risky, like the future earnings of a typical oil field, must be discounted at a higher rate sufficient to compensate for the owner's aversion to bearing that risk. The degree of compensation required to adequately adjust for risk is referred to as the risk premium and can be estimated from market data using a framework called the capital asset pricing model.

One important implication of the capital asset pricing model is that diversifiable risks do not contribute to the risk premium. A diversifiable risk is any factor, like the success or failure of a given well, that can be diluted or averaged out by investing in a sufficiently large number of separate properties. In contrast, risks stemming from future fluctuations in oil prices or drilling costs are nondiversifiable because all oil fields would be affected similarly by these common factors. The distinction between diversifiable and nondiversifiable risk is critical to accurate valuation, especially with respect to the exploratory segment of the petroleum industry: although petroleum exploration may be one of the riskiest businesses in the world, a substantial portion of those risks are diversifiable, thus the risk premium and discount rate for unexplored petroleum properties is relatively low in comparison to other industries.

The appropriate discount rate for the type of petroleum properties typically developed by the major U.S. oil and gas producers would be in the vicinity of 8 to 14%. This is the nominal rate, to be used for discounting cash flows that are stated in current dollars (dollars of the day). If future cash flow streams are projected in terms of constant dollars (where the effect of inflation has already been removed), then the expected rate of inflation must be deducted from the nominal discount rate, as well. Cash flow streams derived from properties owned by smaller or less experienced producers who are unable to diversify their holdings, or in certain foreign lands, may be deemed riskier, in which case the discount rate must be increased in proportion to the added risk.

Not only does the appropriate discount rate vary according to property and owner, but the individual components of overall cash flow within any given project are likely to vary in riskiness and should, in principle, be discounted at separate rates. It is fair to say, however, that methods for disentangling the separate risk factors are complex and prone to error, and it is common practice to discount the overall net

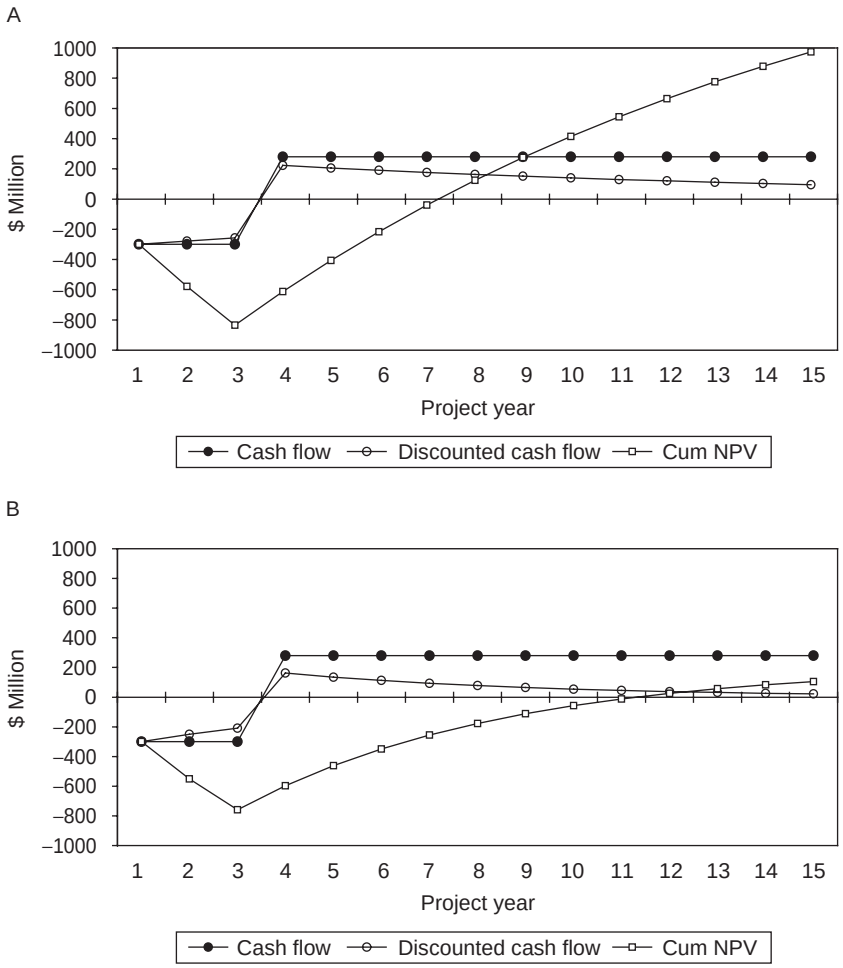


FIGURE 1 Discounting diminishes the contribution of future receipts. The illustrations show a hypothetical net cash flow stream, heavily negative at the outset, followed by consecutive years of positive operating revenues. When discounted at the higher rate, it takes longer for the same revenue stream to offset initial expenditures, resulting in a lower cumulative net present value. (A) Net cash flow and NPV at 8% discount rate. (B) Net cash flow and NPV at 20% discount rate.

cash flow stream at a single rate that reflects the composite risk of the entire project. In many applications, the error involved in this approximation is probably not large. Moreover, new insights regarding the valuation of real options (to be discussed later) provide a procedure in which it is appropriate to discount all cash flow streams at the same rate, which circumvents entirely the problem of estimating separate risk-adjusted discount factors.

3. SPECIAL CHARACTERISTICS OF PETROLEUM PROPERTIES

Most oil and gas fields share certain physical and economic characteristics that strongly influence the pattern and behavior of cash flows, and therefore value. Although the following factors are not necessarily unique to the valuation of petroleum properties, their influence is of sufficient importance to justify more detailed discussion.

3.1 Exploration and Development Risk

Whether at the stage of exploration or development, investments made for the purpose of exploiting an underground petroleum deposit often go awry. Technical failures or economic circumstances may block the recovery of any resources from the property in question. During the past 30 years, 72% of all exploration wells and 19% of all development wells drilled in the United States have

resulted in so-called dry holes, which is industry parlance for a well that is unsuccessful. The causes of failure are numerous, ranging from barren geological formations, to deficiencies in the quality of the deposit that preclude recovery at reasonable cost, to the technical failure or breakdown of drilling equipment. In all such cases, the initial investment is forfeited—written off as the cost of a gamble.

The risk of dry holes is incorporated in the valuation process directly, by assigning appropriate weight to a zero-payoff outcome. After this modification, the NPV expression given previously, compare Eq. (2), would appear thus:

$$NPV = CF_0 + (1 - p_{DH}) \times \sum_{t=1}^{T} \frac{CF_t}{(1+i)^t}, \qquad (3)$$

where CF_0 represents the cost of drilling, p_{DH} represents the probability of a dry hole, and the $\{CF_1,\ldots, CF_T\}$ represent expected future cash flows contingent on success of the well. If the risk of failure is diversifiable, which is certainly true of drilling undertaken by large, publicly owned companies, it does not contribute to the risk premium associated with the property, which means that the discount rate (i) would not be affected by the presence or size of p_{DH}.

Drilling risk can take more complex and subtle forms than the simple dichotomy between success and failure. Outcomes (size of deposit, daily flow rates, gas/oil ratio, etc.) depend on many underlying factors that are not known with certainty but which are amenable to probabilistic analysis. The influence of uncertainty regarding these factors can be quantified via Monte Carlo analysis, wherein projected cash flows and the implied value from Eq. (3) are recomputed under a broad range of possible scenarios. This exercise yields a probability-weighted average outcome, which is the best single indicator of property value. Monte Carlo analysis also reveals the potential range of error in the valuation due to uncertainty in the underlying geological and economic parameters.

A particular nomenclature developed by the petroleum industry permits the degree of exploratory and development risk associated with a given property to be quickly assessed. The "proved reserves" category is the most certain because it includes only those resources that have already been delineated and developed and shown to be economically recoverable using existing technology under prevailing market conditions. Although the outcome of drilling may already have been resolved, proved reserves are not entirely free of risk due to the continuing possibility that price and cost fluctuations will impact future cash flows. These factors represent nondiversifiable sources of risk that influence selection of the discount rate. In practice, the value of proved reserves reported in a producer's financial statements must be reestimated and updated each year to reflect the impact of price and cost fluctuations and any new conclusions about reservoir behavior that emerge from ongoing field operations.

Resources categorized as probable and possible are successively further removed from having been tested, let alone proven, by the drill bit. Their valuations are accordingly subject to increasing (and potentially very large) margins of error. In addition, the reader must be warned that, although this resource classification scheme has been endorsed by the World Petroleum Congress, adherence to the definitions may vary in different parts of the world.

3.2 Oil and Gas Equivalents

Many petroleum properties contain natural gas (and various natural gas liquids) as well as oil, a circumstance that complicates the valuation process. Although the thermal energy content of a barrel of oil is roughly six times that of an mcf (thousand cubic feet) of gas, the market values of the two rarely, if ever, stand in that ratio, as illustrated in Fig. 2. Lack of pipeline facilities is one factor that tends to depress the value of gas deposits, which are more difficult and costly to bring to market than oil. The unconcentrated form in which natural gas occurs (low energy density per unit volume) also renders it

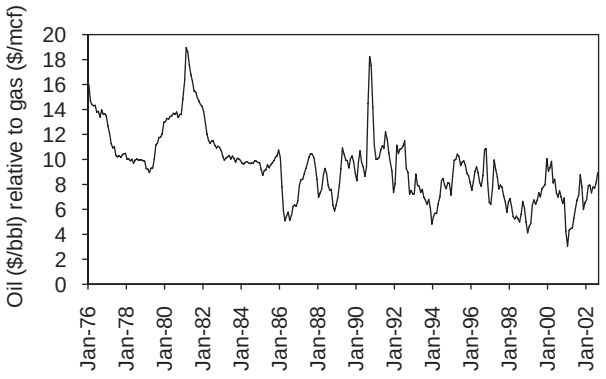


FIGURE 2 Historical fluctuation in the price of oil relative to natural gas. If the two fields traded at parity in terms of raw energy content, one barrel of oil would sell for roughly six times as much as one mcf of natural gas. Historically, oil has tended to trade at a premium, but the ratio is highly variable.

unsuitable for many uses (e.g., as transport fuel), and this tends to further depress its value. Consequently, the relationship between the value of oil and gas deposits is not stable but fluctuates markedly through time and space, depending on the relative demand for, and supply of, the two fuels in regional markets.

The value of any petroleum property will depend on the specific quantities of oil versus gas that are present. While it is common to see properties described in terms of the combined amount of "oil equivalents" (which usually means that each mcf of gas has been counted as one-sixth of a barrel of oil—based on the heat equivalencies of the two fuels), there is no reliable basis for such aggregation, to the extent that the chairman of a leading international oil company has proclaimed that oil- and gas-equivalents simply do not exist. What was meant, of course, is that any calculation of oil equivalents is a garbling of information that only serves to obscure the value of the underlying property. Oil equivalents cannot be compared for purposes of valuation to an equal volume of oil reserves—their values would not be the same. Nor can the oil equivalents of one property be compared to the oil equivalents of any other property where oil and gas reserves are present in different proportions. To do so risks a gross miscalculation of value. As long as consumers continue to distinguish between the two types of hydrocarbons, so must the owners and operators of petroleum properties.

3.3 The Volatility of Commodity Prices and Property Values

Compared to most commodities, oil and gas exhibit highly volatile price movements. Daily, annual, and monthly swings are unusually large relative to the base price levels for both fuels, which puts a large portion of the value of any petroleum property at risk.

The price of a barrel of oil at the wellhead differs, however, from the value of a barrel of oil in the ground, primarily because the reserve cannot be produced and delivered to a buyer instantaneously. This difference is evident in Fig. 3, which contrasts annual changes in the value of oil and gas reserves with corresponding changes in the wellhead price levels of these two commodities. Two things are apparent. First, *in situ* values (i.e., the value of petroleum reserves in the ground) are much smaller than wellhead prices. Over the past 10 years, oil and gas reserves have sold on average for only about 22% and 36% of their respective wellhead prices. Second, the *in situ* values are much more stable than wellhead prices. The year-to-year price change for oil in the ground averages (in absolute value) roughly 13%, versus 20% for changes in price at the wellhead. For gas, the contrast is even stronger: year-to-year price changes averaged 11% in the ground versus 24% at the wellhead.

The relationship between *in situ* values and wellhead prices, although complex and ever-changing, can be understood via a simple model of production from a developed oil field. Let q_0 represent the initial level of production, which is presumed to decline continuously at the rate a due to natural pressure loss in the reservoir as depletion proceeds; thus production at time t is given by $q_t = q_0 e^{-at}$. Assume further that the expected wellhead price of oil remains fixed over the relevant horizon at P per barrel, and that unit production costs amount to C per barrel. Using Eq. (2), the net present value of the property can then be calculated:

$$\begin{aligned} \text{NPV} &= \int_0^T q_0(P-C)\cdot e^{-(a+i)t}dt \\ &\approx \int_0^\infty q_0(P-C)\cdot e^{-(a+i)t}dt \\ &= \frac{q_0(P-C)}{a+i}, \end{aligned} \tag{4}$$

where the approximation is justified by the fact that oil fields are long lived. The volume of reserves (R) in the deposit is given by total production over the life of the property:

$$R = \int_0^T q_0 \cdot e^{-at}dt \approx \int_0^\infty q_0 \cdot e^{-at}dt = \frac{q_0}{a}, \tag{5}$$

which means the rate of extraction from reserves is given by the decline rate: $q_0 = aR$. After substituting this expression for q_0 into (4) and dividing by R, we obtain the *in situ* value (V) of a barrel of reserves:

$$V = \frac{NPV}{R} = \frac{a}{a+i}(P-C). \tag{6}$$

Equation (6) says quite a lot about the value of a producing property. To be concrete, let us set the production decline rate equal to the discount rate (10% is a realistic number for both), and set production costs equal to one-third of the wellhead price. After simplification, the relationship between *in situ* values and wellhead prices then reduces to $V = P/3$, which illustrates the petroleum industry's traditional one-third rule: The value of a barrel in the ground is worth roughly one-third of the price at the wellhead.

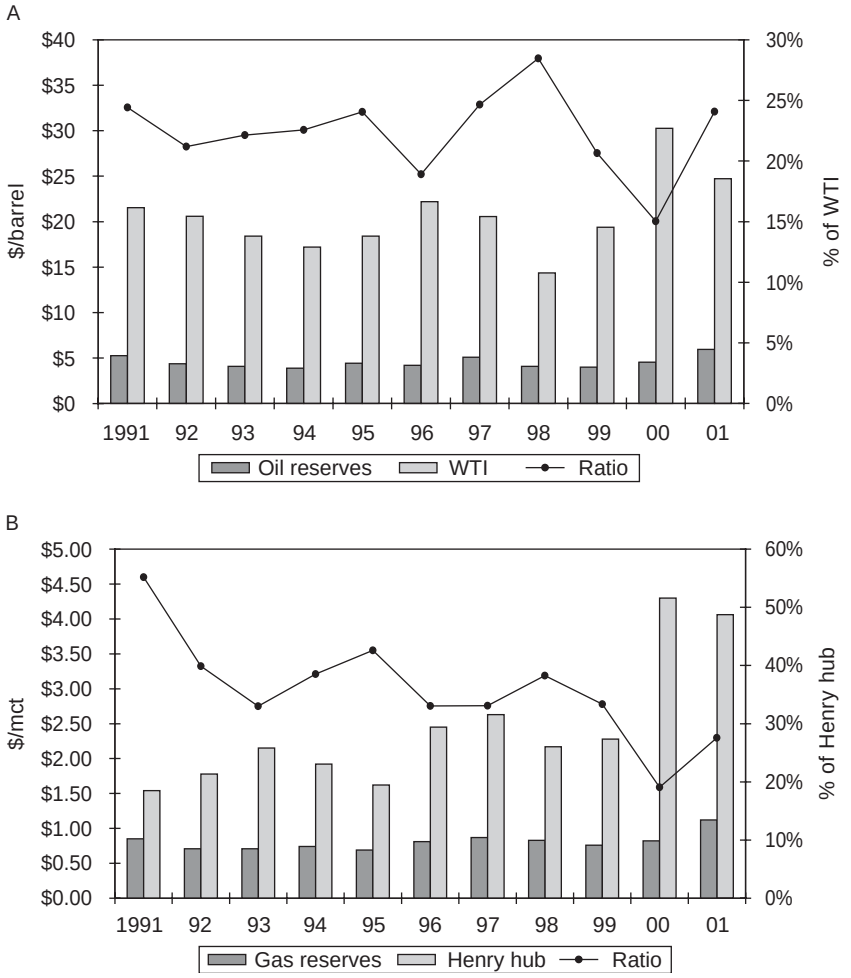


FIGURE 3 Value of oil and natural gas, *in situ* versus wellhead. *In situ* valuations are determined by the average price of reserve transactions tabulated by Cornerstone Ventures, L. P. Wellhead values are determined by the prices of WTI (oil) and Henry Hub (gas) contracts on the New York Mercantile Exchange. From Exhibits 4 and 5, Cornerstone Ventures, "Annual Reserve Report, 2001" (dated February 28, 2002), with permission.

Like any rule of thumb, the one-third rule is often wrong, as the numbers in Fig. 3 demonstrate, but it does point to a general tendency. Moreover, the derivation provided in Eqs. (4) to (6) allows one to anticipate when and why deviations would arise. Reserves that are extracted more rapidly (like natural gas, for example) would tend to sell for more than one-third of the wellhead price. To see why, simply substitute $a > i$ into (6). This confirms a pattern that was evident in Fig. 3: for gas, the value of reserves is consistently a larger fraction of wellhead price than for oil. It is also evident that the value of reserves should move inversely with the level of operating costs, which is why mature fields are eventually abandoned as it becomes more expensive to extract the oil.

What remains to be seen is why *in situ* values are less volatile than wellhead prices. According to the one-third rule, every 10% rise in wellhead price would be matched by a 10% rise in the value of reserves: the volatilities should be the same, but they are not. The explanation stems from the nature of commodity price movements, and the difference between random walk and mean-reverting processes, as illustrated in Fig. 4. A random walk process tends to wander off, rather than to return to its starting point. Any chance departure from the existing price level tends to become permanent. A mean-reverting process tends to be self-correcting; any succession of upward price movements increases the chance of future downward movements, which are required to restore the price to its former level.

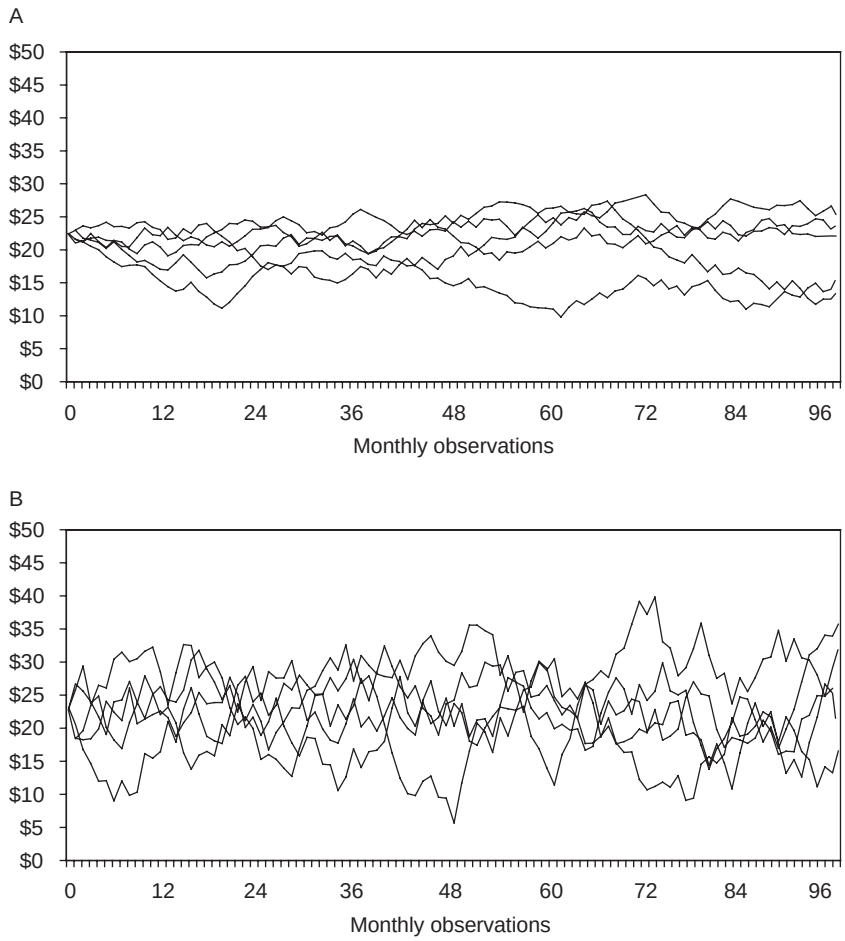


FIGURE 4 The illustrations show five examples each of a random walk sequence and a mean-reverting sequence. Simulations performed by the author. (A) Random walk price trends. (B) Mean-reverting price trends.

Whereas returns on investments in the stock market tend to follow a random walk, the prices of major commodities appear to be mean reverting, which is consistent with the view that the forces of supply and demand tend to keep commodity prices from drifting permanently away from their equilibrium levels. Mean reversion also implies that short-term fluctuations in the price of oil and gas at the wellhead are likely to be reversed in due course. Since the value of a reserve is determined by a combination of current and future prices, the long-term stability provided by mean reversion tends to dampen the impact of short-term commodity price movements. Equation (6), which uses a single value (P) to represent both the current and future commodity price level, is unrealistic in this regard; if prices do not follow a random walk, it gives accurate valuations only when the current wellhead price happens to correspond to the long-term level.

3.4 The Relationship between Reservoir Engineering and Property Valuation

To this point, we have taken the projection of future production, and therefore costs and revenues, as being determined exogenously. Subject to certain physical constraints, however, the rate of production is actually determined by petroleum engineers who design and install facilities with a view to maximizing the value of the field. Property valuation therefore rests implicitly on the assumption that production operations are optimized, and that process of optimization must itself be conducted within the valuation framework.

To illustrate, let us return to the previous example of an oil field subject to exponential decline. Based on our assumption that the discount rate and decline rate both equal 10%, and that operating costs amount to one-third of the wellhead price, we

determined the value of the property to be $P/3$ per barrel of reserves. Now, imagine that our team of reservoir engineers has identified an alternative drilling pattern that would double the extraction rate ($a = 20\%$), with no sacrifice in total reserve volume—the same amount of oil would be produced, but faster. However, this alternative development strategy would also require the expenditure of an additional \$2 per barrel of reserves. Should it be adopted by management? If so, what would be the impact on the value of the property?

The valuation framework, represented in this case by Eq. (6), supplies the answer to these questions. After incrementing the extraction rate to 20%, but leaving all else unchanged, we find the value of the reserve to be $4P/9$ under the alternative drilling pattern, an increase of $P/9$ per barrel. This is a sensible strategy only if the gain outweighs the incremental cost of \$2 per barrel. Thus, we are justified in pursuing the faster, but more costly, production program if, but only if, the wellhead price is expected to exceed \$18/barrel.

Although the example may seem overly simplified, it illustrates an essential point: the value of a petroleum property is neither fixed nor guaranteed, and certainly it is not determined by geology and commodity prices alone. Value depends on management's willingness to identify alternative development concepts and production strategies, and the ability to adapt flexibly to changes in the economic environment. Management that falls short in this regard is bound to leave some portion of a property's potential value on the table.

4. INCORPORATING THE VALUE OF REAL OPTIONS

The discounted cash flow (DCF) technique is versatile, but not without limitations. To project and properly discount future cash flows requires a forecast of petroleum prices, some disentangling of myriad risk factors that impinge on individual components of the cash flow stream, and a correct view as to when each step in the enterprise will transpire. If prices were stable, these requirements would be less of a burden. For the petroleum industry, however, and particularly since the rise of OPEC in the 1970s, the degree of guesswork and resulting scope for error can be painfully high.

To alleviate these problems, advances have exploited the options approach, a technique developed in the 1970s as an alternative to the DCF method. The options approach is a simple but brilliant innovation that was devised initially to value certain types of purely financial assets (e.g., stock market puts and calls). When extended to the problem of valuing tangible investments (bricks and mortar, steel and concrete), the technique is referred to as the real options approach. The profound importance and broad impact of these advances in valuation methodology were quickly recognized, as reflected by the award of the Nobel Memorial Prize in Economics in 1997.

In some situations, the real options approach circumvents all three of the difficulties noted previously: it simplifies the problem of adjusting for risk, provides a suitable forecast of future prices, and dispenses with any rigid or preconceived timeline for project activities. The last aspect is especially critical and gives the method its name. As noted in the preceding section, a portion of the value of any property is dependent on managerial flexibility in the design and timing of project components. The options approach assumes not that management will precommit to a fixed and rigid schedule of drilling and production, but that instead management will react rationally to future events as the project unfolds. Pertinent decisions can be taken at numerous points in the execution of any project, and the essence of the options approach is to recognize that management will make those decisions when the time comes using the information then on hand—and not before.

A simple example gives the flavor of this approach. Consider an owner who holds a 2-year lease on an *undeveloped* oil field, one that has been appraised as holding 100 million barrels of recoverable oil, but the required wells and production facilities have not yet been installed. Suppose installation of productive capacity is estimated to cost \$5.50 per barrel of reserves. The owner of the lease then holds a development option: by incurring an expenditure of \$550 million, the owner will acquire 100 million barrels of developed reserves. If the value of a developed reserve (*in situ*) is expected to exceed \$5.50 per barrel, the expenditure seems justified; otherwise it does not.

As we have seen (see Fig. 3), there is an active market in developed reserves. Suppose those transactions reveal the value of developed reserves to be, say, only \$5 per barrel. Moreover, suppose the historical volatility seen in that market indicates a 50% chance that the value of developed reserves will rise or fall by \$1 each year (a random walk). Thus, the value of developed reserves is expected to remain at \$5 in the

future, albeit with unpredictable variations around that level.

It might appear that development of the reserves in question would be inadvisable, and that a 2-year lease on the property would therefore have no value. It is certain, at least, that immediate development of the reserves would effect a $50 million loss ($500 million minus $550 million), and that prospects for development are not expected to improve on average for the duration of the lease. To conclude from these facts, however, that the lease has no value is to ignore the value of active management. If provisions of the lease do not compel management to develop the reserves, then the lease has considerable option value and should command a relatively high price despite the sometimes unfavorable environment. Indeed, the fair market value of the property would amount to $31 million.

This result is obtained by a straightforward application of the options approach, as diagrammed in Fig. 5. The figure shows a binomial tree that charts the possible future values of developed reserves. Starting from $5 per barrel, the value would either rise or fall (with equal probability) to $6 or $4 after the first year. Continued volatility could carry it to either $3 or $7 by the end of the second year, but the most likely value would be $5 (because there are two price paths leading to that level). Each box in the tree represents a decision node: a point at which management must decide on the basis of available information whether or not to exercise the option to develop the reserves.

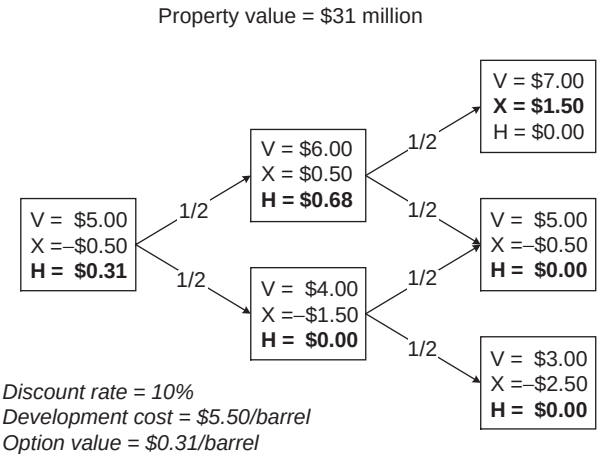


FIGURE 5 Oil field development option. The value of a 2-year field-development lease is calculated by working backward, from right to left, through the binomial tree. The owner must decide, at each node, whether it is more profitable to develop the reserves immediately or to hold off. In this example, the value of developed reserves is assumed to follow a random walk, starting from the level of $55 per barrel.

At the end of the second year, as the lease is about to expire, that decision is straightforward. If the value has reached $7, then development generates an immediate profit of $1.50 per barrel. The only alternative is to allow the lease to expire, which would generate no profit. If the value were lower ($5 or $3), however, then it would be better to hold off—which guarantees zero profit but avoids a loss. Thus, if the reserves had not already been developed, it is clear how management should proceed when the lease is about to expire. In the diagram, the boldface entry at each decision node reflects the optimal choice of action, either **X** for exercise or **H** for holding off. The number shown beside each symbol represents the profit that would be earned via that course of action.

Knowing the end point is critical because it allows us, by working backward, to evaluate the property at each of the earlier decision nodes as well. Suppose, for example, that we find at the end of the first year the value of developed reserves has risen to $6. Immediate development would generate a profit of $0.50 per barrel. The alternative is to hold off, to wait and see what the second year may bring. But we have already evaluated the two possibilities: reserve value will either rise to $7 (which allows a profit of $1.50) or fall to $5 (where profits are $0). The two possibilities are equally likely, so the expected value is simply the average, $0.75. Even after discounting at 10% to compensate for the extra year's delay, the present value of holding off at the end of year 1 is $0.68 (= $0.75/1.10), which exceeds the profit from immediate development. Thus, at the decision node in question, although the option to develop immediately is said to be "in the money," it should not be exercised. This illustrates a more general principle that is not so easy to accommodate within the DCF framework: delay may be advisable even when immediate action appears profitable.

By the same routine, and always working right to left, the other nodes can be completed. Of particular interest is the first node, which represents the property's current valuation based on all available information. While it was evident before we began that immediate development would bring a loss of $0.50 per barrel, we now see that price volatility (and management's capacity to react appropriately to future price changes) adds value to the property. The value of $0.31 per barrel that we now assign is the present value of the average of the two outcomes that will be realized by the end of the

first year (either \$0.68 or \$0.00), each discounted at the rate of 10% to allow for one year's delay: $\$0.31 = \frac{1}{2}(\$0.68 + \$0.00)/1.10$.

The options approach provides answers to some additional questions that would be difficult to address via the DCF method. Specifically, the relationship between the length of lease (term to expiration of the development option), the degree of price volatility, and property value is developed explicitly. Extending the lease term (i.e., adding nodes) and/or increasing the degree of future price volatility (i.e., spreading the tree) has the effect of increasing the upside potential of a given property and can only increase (never decrease) its value. It is a simple matter to reconstruct and recompute the binomial tree under varied assumptions and thereby chart the impact of these parameters.

The method could have been illustrated just as well using a trinomial tree, for which the price movement at each node is either up, constant, or down. In practice, the time step between nodes is taken to be relatively small, in which case the final results are invariant to the particular structure of branching that is adopted. Regarding the discount rate and volatility parameters that are required to value the development option, it has been shown that if the analyst follows a certain formulation to measure the volatility (range and probability of future price movements) from historical market prices, then it is appropriate to use the risk-free rate of discount. This aspect of the options approach frees the analyst from the need to separately figure risk adjustment factors for each component of the cash flow stream using the capital asset pricing model, and from the necessity of preparing a subjective forecast of future prices. Many variations on the basic option framework are possible, including, for example, applications to unexplored properties, already-producing properties, and special formulations that are designed to capture random walk, mean-reverting, and other types of price fluctuations in the underlying resource.

5. PORTFOLIO ANALYSIS: THE COMBINED VALUE OF MULTIPLE PROPERTIES

We have so far examined the problem of valuing a single petroleum property—one oil field considered in isolation of other similar properties that might be included in a transaction or already held by a potential buyer. Valuing a collection, or portfolio, of such properties raises additional issues, some of which push to the very limits of modern techniques. It is useful to distinguish three cases, based on the extent to which the properties are related and whether or not they can be exploited sequentially.

5.1 Properties with Independent Values

If the values of the separate properties are believed to be statistically independent, then the single-property methods described previously can be applied directly and not much else need be said. The value of the whole portfolio would equal the sum of the parts. To satisfy the independence criterion, however, the outcome (i.e., net cash flow) of each property must be uncorrelated with the others. If there are common economic risk factors (e.g., price and cost levels) on which all the properties depend, their values are unlikely to be independent. Common geological factors could also create dependence, as when an exploratory failure on one property is deemed to decrease the probability of success on others.

5.2 Properties with Dependent Values

This case seems more complex, and therefore potentially more interesting. However, dependence among properties does not by itself necessarily require any revision to the valuation method. The whole will still be equal to the sum of the parts, at least if the properties will be exploited simultaneously. By this, we mean that knowledge of the outcome from any one property is not available in time to alter management's plan for exploiting the others. Thus, the properties are operated as if their outcomes are independent, even if an underlying correlation does exist.

5.3 Dependent Properties Exploited Sequentially

If outcomes are dependent and it is possible to exploit the properties sequentially, then the valuation problem changes fundamentally. Management will seek to use whatever information is gleaned from earlier outcomes to enhance subsequent decisions. Statistical dependence implies the existence of relevant information spillovers that facilitate this practice. Thus, it is possible for the value of the portfolio to exceed the sum of the individual parts—the value of acquired information making up the difference. Application of the techniques discussed

previously therefore provides only a lower bound for the combined value.

Valuation models that incorporate information spillovers can become enormously complex as the number of properties increases. The central issue can be illustrated, however, by a simple example. Consider an oil producer who owns two properties, neither of which has been drilled. Let the cost of each exploratory well be \$2 million and the value of each underground deposit (if confirmed by exploratory success) be \$10 million. Finally, we assume that dry hole risk is 40% for each property. What is the value of this portfolio and of its two components?

Based on individual analysis, the value of each component would appear to be \$4 million ($=0.6\times$ \$10−\$2). By taking the sum of its parts, the value of the portfolio would then be appraised at \$8 million. If the drilling outcomes of the two properties are independent, or if they are dependent but drilled simultaneously, this is a correct analysis and we are finished. But, suppose the drilling results are highly dependent (perhaps both geological prospects can be traced to a common sedimentary source), such that the outcome of the first foretells the outcome of the second. If the first confirms a deposit, of which the probability is 60%, the other would also be drilled and the second deposit confirmed too, giving a combined value of \$16 million ($=\$10-\$2+$ \$10−\$2). On the other hand, if the first well fails, of which the probability is 40%, the second would not be drilled, limiting the combined loss to \$2 million.

By recognizing and exploiting the information spillover, management will on average make \$8.8 million ($=0.6\times\$16–0.4\times\$2$), which exceeds by 10% the combined value of the individual properties. To achieve this result, however, the properties must be exploited sequentially, rather than simultaneously. Sequential investment creates an option, while dependence creates the information spillover that gives value to the option. Although our example is very simple, the phenomenon it describes is quite general. Similar results are obtained whether the dependence among properties is complete or partial, symmetric or asymmetric, positive or negative. In all such cases, the sum of property values reckoned individually provides only a minimum valuation for the combined portfolio.

6. CONCLUSION

Uncertainties that range from measurable price volatility to seemingly imponderable geological gambles constantly beset the petroleum business, and such factors will always challenge the accuracy of even the most advanced valuation methods. Although margins of error are inherently large, it is not too much to ask that petroleum property valuations be correct at least on average. Analytical methods have made some marked progress toward that goal, but the remaining obstacles are not inconsequential. We can reasonably expect the quest for improved valuations of petroleum properties to sustain basic research into some of the most fundamental methods of financial economics well into the future.

Acknowledgments

I sincerely thank G. Campbell Watkins and an anonymous referee for reading an earlier draft of this article and for their many helpful suggestions.

SEE ALSO THE FOLLOWING ARTICLES

Depletion and Valuation of Energy Resources • *Discount Rates and Energy Efficiency Gap* • *Energy Futures and Options* • *Investment in Fossil Fuels Industries* • *Markets for Petroleum* • *Oil and Natural Gas: Economics of Exploration* • *Oil and Natural Gas Liquids: Global Magnitude and Distribution* • *Value Theory and Energy*

Further Reading

Adelman, M. A. (1990). Mineral depletion, with special reference to petroleum. *Rev. Econ. Stat.* **72**(1), 1–10.

Adelman, M. A., and Watkins, G. C. (1997). The value of United States oil and gas reserves: Estimation and application. *Adv. Econ. Energy Res.* **10**, 131–184.

Dixit, A. K., and Pindyck, R. S. (1995). The options approach to capital investment. *Harvard Business Rev.* **73**(3), 105–118.

Jacoby, H. D., and Laughton, D. G. (1992). Project evaluation: A practical asset pricing method. *Energy J.* **13**(2), 19–47.

Lerche, I., and MacKay, J. A. (1999). "Economic Risk in Hydrocarbon Exploration." Academic Press, San Diego, CA.

McCray, A. W. (1975). "Petroleum Evaluations and Economic Decisions." Prentice-Hall, Englewood Cliffs, NJ.

Paddock, J. L., Siegel, D. R., and Smith, J. L. (1988). Option valuation of claims on real assets: The case of offshore petroleum leases. *Quarterly J. Economics* **103**(3), 479–508.

Pickles, E., and Smith, J. L. (1993). Petroleum property valuation: A binomial lattice implementation of option pricing theory. *Energy J.* **14**(2), 1–26.

Pindyck, R. S. (1999). The long-run evolution of energy prices. *Energy J.* **20**(2), 1–27.

Petroleum System: Nature's Distribution System for Oil and Gas

LESLIE B. MAGOON
U.S. Geological Survey
Menlo Park, California, United States

1. **Importance of the Petroleum System**
2. **Essential Elements and Processes of the Petroleum System**
3. **Petroleum System Folio Sheet**
4. **The Petroleum System as a Working Hypothesis**

Glossary

active source rock A source rock that generates petroleum, either biogenically or thermally. If a source rock was active in the past, it is either inactive or spent in the present.

burial history chart A burial history curve or geohistory diagram constructed to show the time over which hydrocarbon generation occurs. Depicts the essential elements and the critical moment for the system.

critical moment The time that best depicts the generation-migration-accumulation of hydrocarbons in a petroleum system. A map and cross section drawn at the critical moment best shows the geographic and stratigraphic extent of the system.

essential elements The source rock, reservoir rock, seal rock, and the overburden rock of a petroleum system. Together with the processes, essential elements control the distribution of petroleum in the lithosphere.

events chart A chart for a petroleum system showing when the essential elements and processes took place as well as the preservation time and critical moment of the system.

generation-migration-accumulation One petroleum system process that includes the generation and movement of petroleum from the pod of active source rock to the petroleum show, seep, or accumulation. The time over which this process occurs is the age of the petroleum system.

geographic extent The area over which the petroleum system occurs, defined by a line that circumscribes the pod of active source rock as well as all the discovered petroleum shows, seeps, and accumulations that originated from that pod. The geographic extent is mapped at the critical moment. Also the **known extent**.

level of certainty The measure of confidence that petroleum from a series of genetically related accumulations originated from a specific pod of active source rock. Three levels used are known (!), hypothetical (.), and speculative (?), depending on the level of geochemical, geophysical and geological evidence.

overburden rock The sedimentary rock above which compresses and consolidates the material below. In a petroleum system, the overburden rock overlies the source rock and contributes to its thermal maturation because of higher temperatures at greater depths. An essential element of the petroleum system.

petroleum A mineral oil occurring in subsurface rocks and at the surface which is a naturally occurring mixture of hydrocarbon and nonhydrocarbon compounds. It may occur in the gaseous, liquid, or solid state depending on the nature of these compounds and the existent conditions of temperature and pressure. Common synonyms are **hydrocarbons** and **oil and gas**.

petroleum system The essential elements and processes as well as all genetically related hydrocarbons that occur in petroleum shows, seeps, and accumulations whose provenance is a single pod of active source rock. Also called **hydrocarbon system** and **oil and gas system**.

petroleum system age The time over which the process of generation-migration-accumulation of hydrocarbons in the system takes place on the events chart.

petroleum system name A compound name that includes the source rock in the pod of active source rock, the reservoir rock containing the largest volume of petroleum, and the level of certainty of a petroleum system; for example, the Mandal-Ekofisk(!) petroleum system.

petroleum system processes The two processes of trap formation, and the generation-migration-accumulation. The preservation, degradation, and destruction of

 © 2004 Elsevier Inc. All rights reserved.

petroleum is omitted as a process as it generally occurs after a petroleum system is formed (preservation time). Together with the essential elements, the processes control the distribution of petroleum in the lithosphere.

pod of active source rock A contiguous volume of source rock that is generating and expelling petroleum at the critical moment and is the provenance for a series of genetically related petroleum shows, seeps, and accumulations in a petroleum system. A pod of mature source rock may be active, inactive, or spent.

preservation time The time after generation-migration-accumulation of petroleum takes place and encompasses any changes to the petroleum accumulations up to the present day.

reservoir rock A subsurface volume of rock that has sufficient porosity and permeability to permit the migration and accumulation of petroleum under adequate trap conditions. An essential element of the petroleum system.

seal rock A shale or other impervious rock that acts as a barrier to the passage of petroleum migrating in the subsurface; it overlies the reservoir rock to form a trap or conduit. Also known as **roof rock** and **cap rock** an essential element of the petroleum system.

source rock A rock unit containing sufficient organic matter of suitable chemical composition to biogenically or thermally generate and expel petroleum. An essential element of the petroleum system.

stratigraphic extent The span of lithologic units that encompasses the essential elements within the geographic extent of a petroleum system. It can be displayed on the burial history chart and cross section drawn at the critical moment.

trap Any geometric arrangement of rock regardless of origin that permits significant accumulation of oil, or gas, or both in the subsurface and includes a reservoir rock and an overlying or updip seal rock. Types are stratigraphic, structural, and combination traps. Trap formation is one of the petroleum system processes.

A petroleum system encompasses a pod of active source rock and all related oil and gas, and it includes all the geologic elements and processes that are essential if a hydrocarbon accumulation is to exist. Petroleum here includes (1) high concentrations of any of the following substances: thermal and biogenic gas found in conventional reservoirs as well as in gas hydrate, tight reservoirs, fractured shale, and coal; and (2) condensate, crude oil, and asphalt found in nature. System describes the interdependent elements and processes that form the functional unit that creates hydrocarbon accumulations. The essential elements include a petroleum source rock, reservoir rock, seal rock, and overburden rock; whereas the two processes are trap formation, and the generation, migration (primary and secondary), and accumulation of petroleum. These essential elements and processes must occur in time and space so that organic matter included in a source rock can be converted into a petroleum accumulation. As used in this article, petroleum, hydrocarbons, and oil and gas are synonyms and are included under the term petroleum. Petroleum originally referred to crude oil, but was later broadened to include all naturally occurring hydrocarbons, whether gaseous, liquid, or solid. Geochemically, hydrocarbon compounds are those that contain only hydrogen and carbon, such as aromatic hydrocarbons and saturated hydrocarbons. Hydrocarbon compounds are in contrast to nonhydrocarbon compounds, or those that contain nitrogen, sulfur, and oxygen (N, S, O). Hydrocarbon and nonhydrocarbon compounds are both found in crude oil and natural gas, but hydrocarbon compounds usually predominate. Over the past 20 years, whenever the term hydrocarbons has been used without modifiers it is meant to be synonymous with petroleum. When oil and gas are used together, it collectively refers to crude oil and natural gas in any proportion. Condensate is a gas phase in the accumulation and in a liquid phase at the surface; either way it is considered petroleum as is solid petroleum (i.e. natural bitumen, natural asphalt, bitumenous sands, etc.).

1. IMPORTANCE OF THE PETROLEUM SYSTEM

Petroleum or oil and gas accumulations mostly occur in sedimentary rocks and less frequently in igneous and metamorphic rocks. These seemingly contradictory occurrences are easily explained when petroleum is viewed in the context of a subsurface fluid distribution network whose origin is an active source rock and whose destination is a trap or seep. An active source rock is sufficiently organic-rich that when heated high enough over sufficient time injects petroleum into the adjacent country rock. This largely thermal process of primary migration or expulsion gives way to the physical process of secondary migration outside the active source rock because of buoyancy and diffusion. Depending on the extent of lateral permeability of the country rock, the expelled petroleum will either stay near the site of expulsion or migrate up dip to the nearest trap. If the nearest trap is filled to the spill point, the spilled petroleum will continue to migrate up dip to the next trap and so on until migration ceases or it seeps to the earth's surface.

The extent, connectivity, and type of the conduits and traps radiating up dip from the active source rock determine the nature and area over which these hydrocarbon fluids migrate. This hydrocarbon distribution network is an integral part of the petroleum system; it can be mapped in space and time using geochemical, geological, and geophysical information.

Even though the petroleum system was introduced in 1972, the concept was largely overlooked until the early 1990s. In the past decade, many petroleum systems have been studied and results published in journals, magazines, and presented at scientific meetings. Petroleum geologists now recognize that the sedimentary basin provides the rock framework in which a hydrocarbon fluid distribution network resides (Fig. 1). It is this fluid system that petroleum geologists need to understand if their exploration of the play and prospect are to be successful.

2. ESSENTIAL ELEMENTS AND PROCESSES OF THE PETROLEUM SYSTEM

2.1 Source Rock

A source rock contains sufficient organic matter of suitable chemical composition to biogenically or thermally generate and expel petroleum. Only one active source rock interval occurs in each petroleum system. The two important elements in organic matter are carbon and hydrogen; the carbon provides the skeleton on which the hydrogen is attached. The oxidation of hydrogen provides the energy so the more hydrogen per unit of carbon, the better quality the source rock.

The higher the concentration of carbon in organic matter, the richer is the source rock. Carbon is measured in weight-percent of rock matrix. A source rock needs at least 2 wt % of organic carbon before it can expel petroleum. Organic carbon content (TOC) for a petroleum source rock commonly exceeds 3 wt % and can get as high as 10 wt % or more.

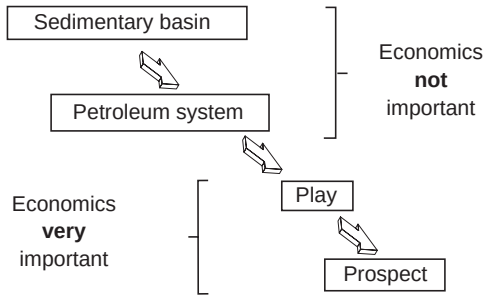


FIGURE 1 Four levels of petroleum investigation. From Magoon, L. B., and Dow, W. G. (1994). The petroleum system—From source to trap. Memoir 60, p. 4 (as Figure 1.1). AAPG © 1994. Reprinted by permission of the AAPG whose permission is required for further use.

Hydrogen content is measured in hydrocarbons per unit of organic carbon (mg HC/g TOC) and is a measure of source rock quality, the higher the hydrogen index (HI), the better the quality. In addition, the HI indicates the type of hydrocarbon fluid expelled. For example, HI greater than 300 (mg HC/g TOC) and less than 900 expel mostly oil with some gas, HI between 200 and 300 expel some light oil but mostly gas, and HI between 50 and 200 is just natural gas. As a source rock thermally matures and expels petroleum, the HI is reduced to below 50 so the difference between the immature HI and HI of the depleted source rock is a measure of the amount of hydrogen or petroleum expelled. For example, if an immature source rock has an HI of 500 and the HI of the spent source rock is 100, the 400 mg HC/g TOC of oil and gas was expelled as compared to an immature HI of 200 and the spent HI is 100, the amount of expelled gas is 100 mg HC/g TOC.

The presence of porphryns from plant chlorophyll as well as other complex molecules that are the building blocks of single and multicellular plants and animals are proof that the organic matter in source rocks originates from the biosphere. These so-called biomarkers indicate the origin and sometimes the age of the source rock and the origin of the petroleum.

2.2 Overburden Rock

Overburden rock is the sedimentary rock above which compresses and consolidates the material below. In a petroleum system, the overburden rock overlies the source rock and contributes to its thermal maturation because of higher temperatures at greater depths. The age and thickness of the overburden rock determines the burial rate and thereby influences the heating rate. Frequently, the reservoir and seal rocks immediately overly the source rock and are thus included within the overburden rock. The close proximity of the source and reservoir rocks allow the expelled petroleum immediate access to a migration conduit. The deposition and erosion of the overburden rock through time determines the slope and form of the burial history curve of the source rock through petroleum expulsion. Rapid burial with little erosion is best for the preservation of the petroleum system.

2.3 Reservoir Rock

A reservoir rock is a subsurface volume of rock that has sufficient porosity and permeability to permit the migration and accumulation of petroleum under adequate trap conditions. Porosity is a measure in percentage of pore volume or size of holes or vugs per unit volume of rock. For example, a well-sorted sand in a 300 ml container will hold about 100 ml of water in its pore space, or a porosity of 33%. If petroleum is present, it can also occupy this pore space. During burial of this sand, compaction reduces this porosity substantially to where only a small percentage of porosity is left.

Permeability is a measure of the connectivity of pores in the subsurface. The sand in the container has narrow pore throats between the large pores that allows fluid to pass from one pore to another. Permeability is measured in millidarcies (md) or Darcy (1000 md) of these narrow throats. Commonly, permeability in the range of 100 to 500 md are reasonable values for a petroleum reservoir rock. Values over a Darcy are exceptional. Fractures have infinite permeability.

Any lithology can have porosity and permeability. Siliciclastic sandstone is most likely to have vertical and lateral porosity and permeability over a relatively large distance. Frequently, carbonate rock has the most complex patterns of porosity, frequently vuggy porosity with little or no permeability, but dolomitized zones can have lateral porosity and permeability. Volcanic rock can be vuggy where hot gases bubble from molten rock. Metamorphic and other sedimentary rocks can acquire secondary porosity because of weathering or groundwater movement. All these rock types can be shattered so that unconnected pores can be connected with fractures.

Major and minor reservoir rocks are determined from the percentage of in-place petroleum that originated from a particular pod of active source rock. If the volume of in-place petroleum is unavailable, the volume of recoverable hydrocarbons are used. All oil and gas fields included in a petroleum system are listed and the original in-place (recoverable) hydrocarbons are determined by stratigraphic interval. The volumes of in-place hydrocarbons for each stratigraphic interval are added up, and the percentage of in-place hydrocarbons by stratigraphic interval is determined. Usually, one stratigraphic interval contains most of the in-place hydrocarbons, so this interval is the major reservoir rock. Reservoir rocks that contain minor or lesser amounts of in-place hydrocarbons are the minor reservoir rocks. The major reservoir rock is the second name used in the name of the petroleum system. The major reservoir rock indicates the optimum migration path between the pod of active source rock and the traps that include the major reservoir rock. The minor reservoir rock indicates the least effective migration path, or it indicates a migration path that should be studied for overlooked prospects. Major and minor reservoir rocks should be included on the petroleum system events chart.

2.4 Seal Rock

Seal rock is a shale, evaporite, or other impervious rock that acts as a barrier to the passage of petroleum migrating in the sub-surface; it overlies the reservoir rock to form a trap or conduit. The seal rock is also known as roof rock or cap rock. A dipping reservoir rock overlain by a seal rock provides a migration path for petroleum.

2.5 Trap Formation

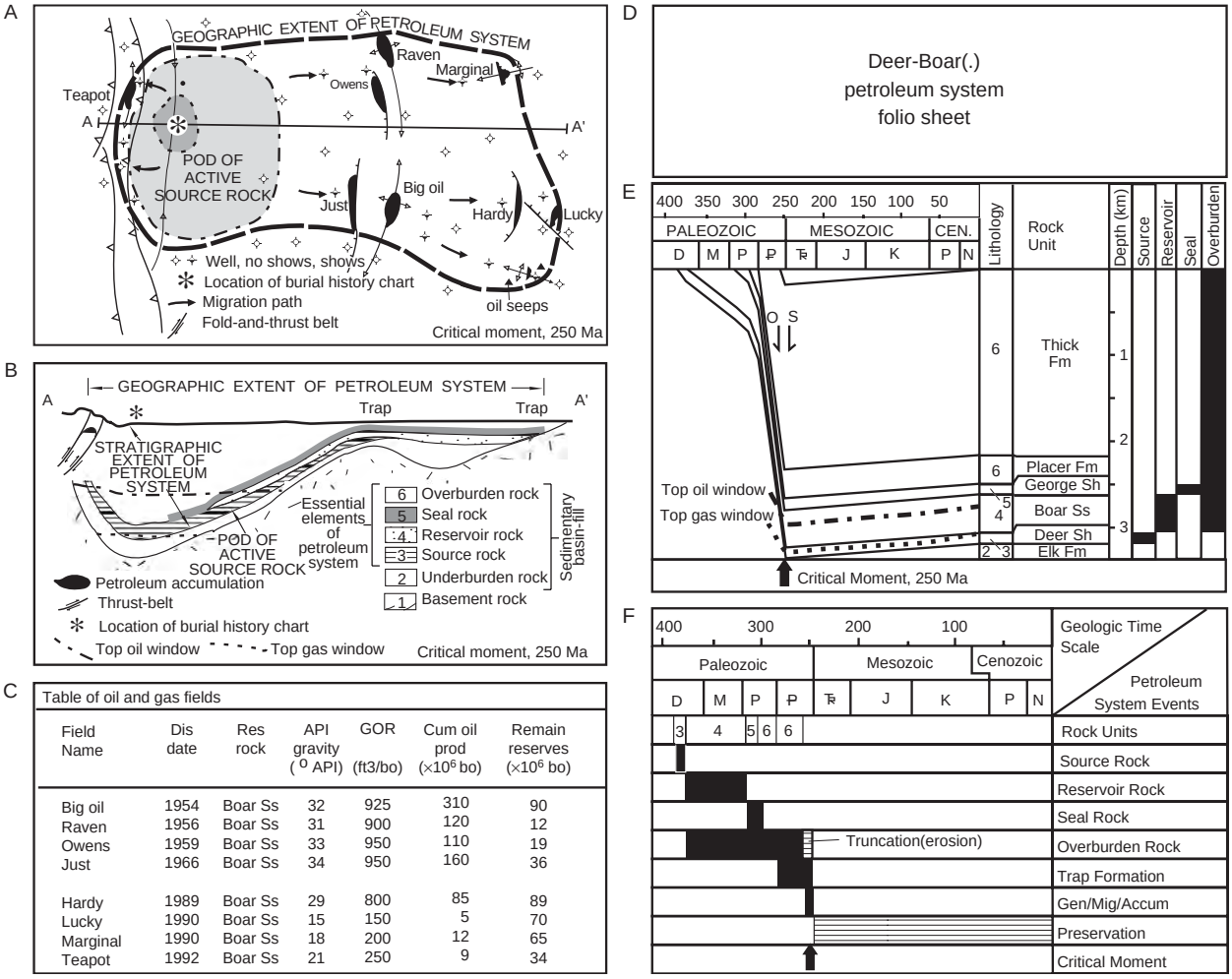
Any geometric arrangement of rock regardless of origin that permits significant accumulation of oil or gas, or both, in the subsurface and includes a reservoir rock and an overlying or updip seal rock is a trap. Types are stratigraphic, structural, and combination traps. The formation of traps by deposition, tension or compression is one of the processes needed to create a petroleum system. Trap formation must occur before or during petroleum migration in order to have an accumulation.

2.6 Generation-Migration-Accumulation

One petroleum system process that includes the generation and movement of petroleum from the pod of active source rock to the petroleum show, seep, or accumulation. The time over which this process occurs is the age of the petroleum system. Even though each increment of this process can be studied separately, it is the petroleum in the accumulation, seep, or show in a well that is proof that a petroleum system exists.

3. PETROLEUM SYSTEM FOLIO SHEET

The petroleum-system folio sheet is a graphical way to depict the geographic, stratigraphic, and temporal evolution of the petroleum system (Fig. 2; Table I).



Field Name	Dis date	Res rock	API gravity (^{O} API)	GOR (ft3/bo)	Cum oil prod (×10^{6} bo)	Remain reserves (×10^{6} bo)
Big oil	1954	Boar Ss	32	925	310	90
Raven	1956	Boar Ss	31	900	120	12
Owens	1959	Boar Ss	33	950	110	19
Just	1966	Boar Ss	34	950	160	36
Hardy	1989	Boar Ss	29	800	85	89
Lucky	1990	Boar Ss	15	150	5	70
Marginal	1990	Boar Ss	18	200	12	65
Teapot	1992	Boar Ss	21	250	9	34

FIGURE 2 (A) Map showing the geographic extent of the so-called Deer-Boar(.) petroleum system at the critical moment (CM, 250 Ma). Thermally immature source rock is outside the oil window. The pod of active source rock lies within the oil and gas windows. (B) Cross section showing the stratigraphic extent of the so-called Deer-Boar(.) petroleum system at the critical moment (250 Ma). Thermally immature source rock lies up dip of the oil window. The pod of active source rock is down dip of the oil window. (C) Table of oil and gas fields in the Deer-Boar(.) petroleum system, or the accumulations related to one pod of active source rock. (D) The name of the petroleum-fluid system. (E) Burial history chart shows the critical moment (250 Ma) and the time of oil generation (260-240 Ma) for the so-called Deer-Boar(.) petroleum system. This information is used on the events chart. All rock unit names used here are fictitious. (F) Events chart shows the relationship between the essential elements and processes as well as the preservation time and critical moment for the so-called Deer-Boar(.) petroleum system. (G) The text and other figures needed to describe the petroleum system. From Magoon, L. B., and Dow, W. G. (2000). Mapping the petroleum system—An investigative technique to explore the hydrocarbon fluid system. Memoir 73, p. 55 (as Figure 1). AAPG © 2000. Reprinted by permission of the AAPG whose permission is required for further use.

The folio sheet includes (1) the petroleum system map (Fig. 2A), (2) the petroleum system cross section (Fig. 2B), (3) a table of genetically related accumulations (Fig. 2C), (4) the petroleum system name (Fig. 2D), (5) a burial history chart located over the pod of active source rock (Fig. 2E), and (6) an events chart to summarize the history of the petroleum system (Fig. 2F). In the ideal case, this folio sheet summarizes the detailed work of many specialists and provides the supportive information for the petroleum system map needed to evaluate the play and prospect.

3.1 Petroleum System Map

The petroleum system map (Fig. 2A; Table I) shows (1) the geographic extent of the petroleum system; (2) the pod of active source rock; (3) the genetically

TABLE I

Components of a Complete Petroleum System Study

Figure or table	Information required	Purpose
Petroleum system map	Locate petroleum fields included in system; Indicate whether oil or gas; Indicate surface and subsurface oil or gas shows included in system; Indicate direction of petroleum migration; and Indicate distribution of pod of active source rock.	*Geographic extent of a petroleum system* at critical moment shown by circumscribing the pod of active source rock and the outer limit of migrated hydrocarbons; *Source rock name* from pod of active source rock; and *Petroleum system burial history chart* location.
Petroleum system table	List all oil or gas fields by petroleum system; Indicate discovered reserves and in-place petroleum by stratigraphic unit; trap type; reservoir rock name, age, and lithology; and seal rock name, age, and lithology; For oil field indicate GOR, API gravity, sulfur content, and Pr/Ph ratio; and For gas field indicate GOR, $\delta^{13}C$, and gas wetness $[(C_1/C_1-C_4)\times 100]$.	*In-place petroleum* for mass balance calculation; *Reservoir rock name* from that reservoir rock with the highest percentage of in-place petroleum; *Seal rock* most commonly occuring in trap; and *Oil or gas province* from average GOR.
Petroleum system cross section	Structural and stratigraphic information such as deformation style and rock units; Indicate oil window and gas window; Indicate petroleum shows and accumulations; Draw at critical moment; and Indicate direction and conduits for petroleum migration.	*Stratigraphic extent of petroleum system* at the critical moment by identifying the base of the pod of active source rock or base of hydrocarbon column, whichever is deeper; *Geographic extent of petroleum system*; *Pod of active source rock* shown; *Overburden rock* shown; and *Petroleum system burial history chart* location.
Petroleum system burial history chart	Stratigraphic units penetrated in well; Time rock-units were deposited; Thickness of rock units; Names of rock units; Lithology of rock units; Present day thermal maturity profile; Present day geothermal gradient; Computer program to determine time and depth for oil window and gas window; and Indicate essential elements of petroleum system.	*Petroleum system events chart information* determined from chart, such as onset and end (at uplift) of petroleum generation, and critical moment; *Essential elements of petroleum system* shown; *Oil window* depth for cross section at critical moment; and *Gas window* depth for cross section at critical moment.
Petroleum system events chart	Age of essential elements of petroleum system; Onset and end of trap formation; Onset and end of petroleum generation, migration, accumulation; Preservation time of petroleum system; and Critical moment.	*Petroleum system events chart* summarizes in one diagram the essential elements and processes of the system, as well as the preservation time and critical moment.
Petroleum-petroleum correlation	Geochemical evidence, such as bulk properties, biological markers, and isotopic data to show that more than one petroleum accumulation came from the same source rock (but not necessarily the same pod of active source rock).	*Geographic and stratigraphic extent of a petroleum system* is established with this geochemical correction in concert with the structure and stratigraphy of the pod of active source rock and the adjacent traps.
Petroleum-source rock correlation	Geochemical evidence, such as biological marker and isotopic data, to indicate a certain petroleum originated from a specific source rock.	*Level of certainty* is established using geological and geochemical evidence and indicates the confidence that a specific source rock expelled a given petroleum.
Mass balance calculation	TOC and Rock-Eval pyrolysis; Source rock density; and Volume of pod of active source rock.	*Mass of petroleum generated* to determine petroleum system generation-accumulation efficiency (GAE).

Source. From Magoon, L. B., and Valin, Z. C. (1994). Overview of petroleum system case studies. Memoir 60, p. 332 (as Table 20.2). AAPG © 1994. Reprinted by permission of the AAPG whose permission is required for further use.

related oil and gas accumulations, shows, and seeps; (4) the location of the burial history chart; and (5) the location of the petroleum-system cross section. Usually this map is drawn at the present day, but may be refined later to depict the critical moment or that time when most of the hydrocarbons were generated and accumulating, especially if the petroleum system is old. This map is in contrast to a geologic map which depicts rock units and geometry at the surface or an oil and gas field map.

3.1.1 Accumulations, Shows, and Seeps

The genetically related oil and gas fields or petroleum accumulations, by inference, originated from the pod of active source rock. So start with an oil and gas field map of the petroleum province. The accumulations shown on an oil and gas field map need to be grouped into one or more possible petroleum systems based on their geographic and stratigraphic locations, and the bulk and geochemical properties of the fluids in each accumulation (Fig. 2A). For example, the accumulations in the fictitious Deer-Boar(.) petroleum system include Teapot in the thrust belt; Raven, Owens, Just, and Big Oil in the anticlines adjacent to the foreland basin; and Marginal, Hardy, and Lucky, the most distal accumulations. These widely spaced accumulations are in the same stratigraphic interval, the Boar Sandstone, and all have similar geochemical properties, as known from wells that sample these accumulations.

The exploratory wells indicate that some of the wells have oil shows in the Boar Sandstone. One surface seep occurs in the southeast portion of the map where the Boar Sandstone crops out. These oil occurrences have geochemical similarities with the oil accumulations. Based on this information, these oil occurrences are judged to belong to the same petroleum system.

3.1.2 Pod of Active Source Rock

The pod of active source rock is a contiguous volume of source rock that generated and expelled petroleum at the critical moment and is the provenance for a series of genetically related petroleum shows, seeps, and accumulations in a petroleum system. A pod of mature source rock may be active, inactive, or spent. There is only one pod of active source rock for each petroleum system.

The kerogen type for the thermally mature source rock has been shown by numerous investigators to control the type and volume of petroleum expelled. Other investigators have provided explanations of the tools and techniques to characterize and map the pod of active source rock. The pod of active source rock (also referred to as the kitchen or oil and gas window) is a required feature of the petroleum system map because of its genetic control on petroleum accumulations.

For the fictitious Deer-Boar(.), the pod of active source rock is in the western part of the map area and is mapped using a dashed line (long and short dashes, Figs. 2A and 2B). The Deer Shale is considered the most likely source rock because it is geographically near and stratigraphically below the Boar Sandstone, the reservoir rock for the accumulations, shows, and seep. Thermal maturity data for the Deer Shale indicate it is thermally mature in the foreland basin, but immature in the thrust belt toward the west and on the craton toward the east. The dashed lines correspond to the thermal maturity contours for the top of the oil window (long and short dashes) and the top of the gas window (short dashes) (Figs. 2A and 2B).

3.1.3 Geographic Extent

The geographic extent of the petroleum system at the critical moment is described by a line that circumscribes the pod of active source rock and includes all the discovered petroleum shows, seeps, and accumulations that originated from that pod. The critical moment is a snapshot of the petroleum system at the time when most of the hydrocarbons were generated and migrating. A map of the Deer-Boar(.) petroleum system, drawn at the end of the Paleozoic time (250 Ma), includes a line that circumscribes the pod of active source rock and all related, discovered hydrocarbons. This area represents the geographic extent or known extent of the petroleum system (Fig. 2A).

To draw this petroleum system map, we assume that all the petroleum occurrences emanate from the same pod of active source rock. The pod of active source rock is shown as a shaded area, and the depth to the source rock is assumed to be close to the thickness of overburden rock which is contoured in thousands of meters. Based on the locations of the accumulations, drill-stem tests, and exploratory wells with and without oil shows the geographic extent of the petroleum system is drawn. Using the guidelines discussed earlier, the location of the cross section and the burial chart are indicated. This petroleum system map emphasizes the short migration path of the discovered oil fields and that the long migration path to the northwest lacks an oil field. This map strongly suggests that the most likely place to find undiscovered oil is above or close to the pod of active source rock and along the preferential

migration paths as expressed by the geographic extent of the petroleum system.

3.1.4 Location of Cross Section

The cross section is placed on the map so that it passes through the largest accumulations, thickest overburden rock, and extends over the entire geographic extent of the petroleum system (Figure 2B). If possible, start with an available present-day cross section unless the petroleum system is so old or structurally altered that a restored cross section representing a previous time is required to depict the time when most of the hydrocarbons migrated and accumulated. Here, a cross section at the end of the Paleozoic (250 Ma) was used (Fig. 2B). The largest accumulations are included because they are usually located on the simplest, most efficient migration path from the pod of active source rock. A transect through the thickest overburden rock shows the most likely area where the source rock is thermally mature and, therefore, the provenance of the hydrocarbons. The cross section should transect the entire the petroleum system so that the basis for the geographic extent can be demonstrated.

3.1.5 Location of Burial History Chart

The location of the burial history chart is along the cross section line within the pod of active source rock (Fig. 2B, E). At this location, the source rock must be thermally mature (active or spent) otherwise petroleum would be absent in the conduits or migration paths. The reconstruction of the burial history provides the basis for the times of the onset (O) of generation-migration-accumulation, the depletion (D) of the source rock, and the critical moment (CM).

3.2 Table of Fields

The table showing all the oil and gas accumulations included in the folio sheet provides important information about the petroleum system (Fig. 2C; Table I). First, the discovery dates and sizes of the fields are useful for field-size distributions and discovery-rate modeling. Second, the complexity of the hydrocarbon plumbing system is suggested by the number of reservoir rocks. One reservoir rock for all fields indicates a simple plumbing system, whereas many reservoir rocks indicates a more complicated system. Third, the size of the petroleum system and the generation and expulsion efficiency can be determined by using the total volume of recoverable oil and gas for all fields. Finally, the reservoir rock with the highest percentage of oil or gas reserves is to be used in the petroleum system name. For example, if all the oil is in the Boar Sandstone, it is included in the name (Fig. 2D).

For example, the Deer-Boar(.) is a 1.2 billion barrel petroleum system with a simple plumbing system. This size designation using recoverable petroleum is most useful to the explorationist who is interested in (1) comparing the sizes of different petroleum systems to rank or plan an exploration program and (2) comparing the field-sizes in a petroleum system to determine the most likely prospect size and what volumes can be produced. However, the size (volume of recoverable petroleum) of the petroleum system needed for material balance equations is quite different. Here, three additional types of information are estimated: the in-place petroleum for each field, what is left behind along the migration path, and what was lost in surface seeps and exhumed accumulations. These estimates are made by the investigator. This volume of petroleum can then be compared to the estimated volume of petroleum generated in the active source rock.

3.3 Cross Section

The petroleum system cross section (Fig. 2B; Table I), drawn at the critical moment, or time when most of the hydrocarbons were generated, shows the geographic and stratigraphic extent of the petroleum system and how each rock unit functions within the system to distribute the oil and gas. Stratigraphically, the petroleum system includes a petroleum source rock, reservoir rock, seal rock, and overburden rock. This cross section is in contrast to structural or stratigraphic cross sections.

The presence of adequate overburden rock in the correct geometry provides (1) the burial needed to thermally mature the source rock, (2) the up dip vector needed for oil and gas to migrate to shallower depths, (3) the burial depth variations needed to form traps for petroleum accumulations, and (4) the burial depth of accumulations that allow for the preservation or biodegradation of oil. If the history of the petroleum system is to be correctly modeled, the age, thickness, and erosional history of the overburden rock is required. The cross section, drawn to represent the end of the Paleozoic (250 Ma), shows the geometry or structural style of the essential elements at the time of hydrocarbon accumulation, or critical moment, and best depicts the stratigraphic extent of the system (Fig. 2B).

3.4 Petroleum System Name

The name of a petroleum system labels the hydrocarbon-fluid system or distribution network (Fig. 2D) in the same way the name Colorado River designates an aqueous distribution system, the river and its tributaries. The name of the petroleum system includes the geological formation names of the source rock followed by the major reservoir rock (Fig. 2C) and then the symbol expressing the level of certainty. For example, the Deer-Boar(.) is the name of a hydrocarbon fluid system whose source rock, the Deer Shale, most likely generated the petroleum that charged one or more reservoir rocks, which in this case is the Boar Sandstone. It is the major reservoir rock because it contains the highest percentage by volume of hydrocarbons in the petroleum system.

A petroleum system can be identified at three levels of certainty: known, hypothetical, and speculative. At the end of the system's name, the level of certainty is indicated by (!) for known, (.) for hypothetical, and (?) for speculative. The symbol indicates the level of certainty that a particular pod of active source rock has generated the hydrocarbons on the table of accumulations (Fig. 2C). In a known petroleum system, a well-defined geochemical correlation exists between the source rock and the oil accumulations. In a hypothetical petroleum system, geochemical information identifies a source rock, but no geochemical match exists between the source rock and the petroleum accumulation. In a speculative petroleum system, the link between a source rock and petroleum accumulations is postulated entirely on the basis of geological or geophysical evidence.

In certain frontier areas of the world, especially in the offshore, the stratigraphic units are poorly understood and frequently undesignated. Here, the judgment of the investigator is required. The geologist should avoid using ages, such as Jurassic, in the name because it fails to uniquely identify the petroleum system. Other situations arise where it is difficult or confusing to follow the naming rules. For example, when a rock unit that includes both the source and reservoir forms more than one petroleum system, the same name might be used more than once, such as Monterey(!) petroleum systems. Here, a geographic modifier can be used to differentiate the systems. Another naming problem arises in frontier areas where formal rock units are lacking, so only ages or geophysically mapable units are used. A geographic name or the name of an accumulation in the petroleum system may be used. If it is impossible to follow the formal designation of a petroleum system, the investigator should select a unique name that identifies the fluid system, not the rock units.

3.5 Burial History Chart

The purpose of the burial history chart is to show the essential elements, and three important hydrocarbon events, which are (1) the onset (O) of generation-migration-accumulation, (2) the partially spent or depleted (D) active source rock, and (3) the critical moment (CM) of the petroleum system (Fig. 2E; Table I). The top of the oil and gas windows, and the lithology and name of the rock units involved should also be shown. This chart uses sedimentologic and paleontologic evidence in the overburden rock to reconstruct the burial or thermal history of the source rock. The onset of generation-migration-accumulation usually occurs when the source rock reaches a thermal maturity at a vitrinite reflectance equivalence of 0.6% R_o and ends when the source rock is either uplifted before all the hydrocarbons can be expelled or is depleted as the source rock is more deeply buried. The location of the burial history chart is shown on the petroleum system map and cross section.

In this example, the Deer Shale (rock unit 3) is the source rock, the Boar Sandstone (4) is the reservoir rock, the George Shale (5) is the seal rock, and all the rock units above the Deer Shale (4, 5, 6) comprise the overburden rock. The burial history chart is located where the overburden rock is thickest, and indicates that the onset of generation-migration-accumulation started 260 Ma in Permian time and was at maximum burial depth 255 Ma. Oil generation ended about 240 Ma because the source rock is depleted. The critical moment as judged by the investigator is 250 Ma because modeling indicates most ($>50\%$) of the hydrocarbons have been expelled and are accumulating in their primary traps. However, the investigator would be correct to have chosen anytime between 250 and 240 Ma, but 250 Ma was chosen because the best geologic information is available to reconstruct the map and cross section. The time of generation-migration-accumulation ranges from 260 to 240 Ma and is the age of the petroleum system.

As mentioned in the cross section discussion, knowing the age and thickness for each increment of overburden rock is crucial for any modeling exercise. Each increment needs to be bounded by time lines whose dates are supported by paleontologic dates, isotopic age dates, or other means of dating strata. As the number of increments in the overburden rock

increases, the details of the burial history of the source rock will be better understood. For instance, in the previous example, the overburden rock is Permian, undifferentiated. Suppose, however, that paleontologic dates indicate that 95% of the overburden rock is Early Permian and that the rest is Late Permian. This increase in increments in the overburden rocks will change the time when the underlying source rock becomes thermally mature. In this example, the change in time that the source rock becomes mature might be considered insignificant, but in other examples the difference could be significant.

3.5.1 Location of Burial History Chart

The burial history chart chosen to show the three hydrocarbon events for a petroleum system should be located in the pod of active source rock where, in the judgment of the investigator, much of the oil and gas originated. Usually this location is down-dip from a major migration path to the largest fields.

Petroleum systems are seldom so simple that only one burial history chart adequately describes the same three hydrocarbon events for every location in the pod of active source rock. The investigator chooses the burial history curve that best suits the purpose. If the investigator is presenting (oral or written) a petroleum system investigation, he or she would use the burial history curve down-dip from a major migration path to the largest fields. However, if the subject is a play or prospect, the burial history curve would be down-dip on a suspected migration path to the area of the play or prospect.

3.5.2 Critical Moment

The generation, migration, and accumulation of oil and gas in a petroleum system never starts when the source rock is being deposited and seldom extends into the present day. If a source rock is deposited in the Paleozoic, it may be the Mesozoic before it becomes thermally mature and charges adjacent traps, and by the Cenozoic, this source rock is probably depleted. The time over which the process of generation-migration-accumulation takes place could be tens of millions of years. This is a long period of time to chose from if an investigator needs to select the most appropriate moment during this process to make a map and cross section that shows the petroleum system when most ($>50\%$) of the hydrocarbons were migrating and accumulating. To help the investigator with this important exercise, the critical moment was introduced and incorporated into the petroleum system folio sheet.

Geologists use the concept of the critical moment for other exercises. Whenever a map, such as an isopach map, is constructed, it is frequently reconstructed to its original thickness at the moment of deposition. The kinematic development of a fold-and-thrust belt occurs over many millions of years, but it is frequently represented by one cross section, or a snapshot in time. A structural cross section of a fold-and-thrust belt reconstructed at the end of the Cretaceous, for example, utilizes the critical moment concept.

The critical moment is the time that best depicts the generation-migration-accumulation of hydrocarbons in a petroleum system. This definition needs an explanation and an example to be better understood. A moment is a brief, indefinite interval of time that is of particular importance. For a camera to take a picture, a moment is less than a second. In geology, the further one goes back in time, the interval becomes thousands or even millions of years. For the petroleum system, moment relates to the shortest measurable time. Critical refers to the moment that best shows, in the judgment of the investigator, the process of generation-migration-accumulation.

Best is a keyword in this definition. Best contains the criteria the investigator should use to select the appropriate moment. The best time needs to fulfill several criteria: (1) it must be within the age of the petroleum system; (2) it must be when most, or more than half, of the hydrocarbons are migrating and accumulating; and (3) it must be shown as an arrow, not an interval, on the burial history and events charts.

The critical moment of a petroleum system can vary depending on its purpose. If the purpose is a petroleum system case study, then the critical moment should be representative of the entire system. However, if the purpose is related to an exploration play or prospect, then the critical moment should be for that part of the pod of active source rock most likely to charge the traps in the play or prospect. Depending on the size, thickness, and variation of the thermal maturity of the pod of active source rock and the objective of the investigator, these could be different best times, none of which are incorrect. In fact, the investigator may need to make numerous burial history charts of a large, thick pod of active source rock that has a wide range of thermal maturity to determine which best moment properly depicts generation-migration-accumulation for the intended audience.

The burial history chart omits important information available in most modeling software packages that explains the critical moment from a different perspective (Fig. 3). This graph shows cumulative volumes of generated or expelled oil and gas.

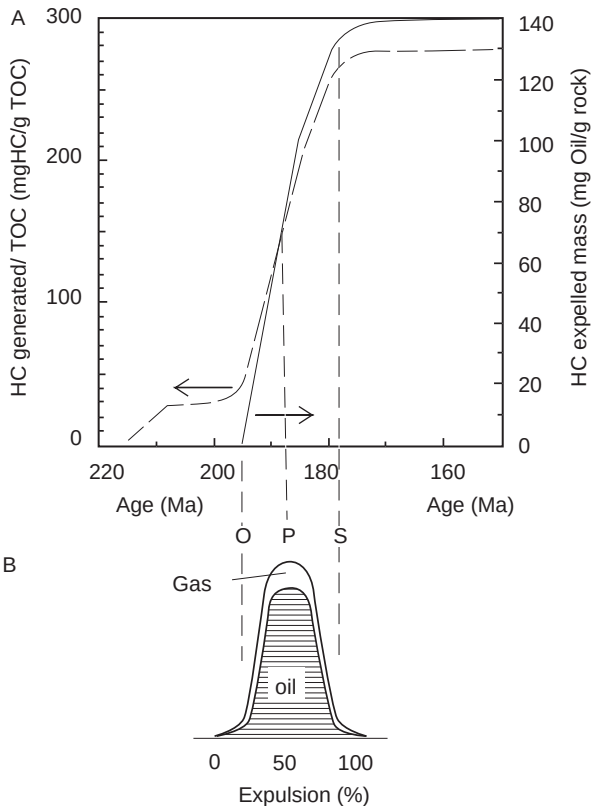


FIGURE 3 (A) Cumulative curves showing the time over which hydrocarbons (HC) are generated and expelled. (B) Distribution curve for oil and gas (HC) expulsion using the same information and showing the onset (O) of expulsion, peak (P) expulsion, and depletion (D) of the source rock. The critical moment (CM) is selected to be any time between P and D. From Magoon, L. B., and Dow, W. G. (2000). Mapping the petroleum system—An investigative technique to explore the hydrocarbon fluid system. Memoir 73, p. 64 (as Figure 7). AAPG © 2000. Reprinted by permission of the AAPG whose permission is required for further use.

Wherever the curves are horizontal, no volume is being added. This graph shows the onset of generation (dashed line) precedes expulsion (solid line) by almost 20 million years. According to this graph, most expulsion occurs more than 15 million years from 195 to 179 Ma. When this cumulative expulsion curve is transformed to a curve showing the distribution of expulsion, it shows peak expulsion at 187.5 Ma (Fig. 3). At this time, at least half of the petroleum is migrating so the critical moment can be selected by the investigator as any time between 187.5 to 179 Ma.

3.6 Events Chart

The petroleum-system events chart shows the temporal relationship of the rock units, essential elements, processes, preservation, and critical moment for each petroleum system in bar graph form (Fig. 2F; Table I). The events chart concept is flexible and is used as a risk chart to evaluate plays and prospects

The events chart shows the following items (Fig. 2F). The rock units include those that are within the stratigraphic extent of the petroleum system. On a certainty scale, the ages of the four essential elements (source, reservoir, seal, and overburden rocks) are usually better established from paleontology or radiometric age dates than those associated with the two processes (trap formation and generation-migration-accumulation). Determining the times over which the traps form and hydrocarbons accumulate (generation-migration-accumulation) is more interpretive because there is less precise temporal information about these processes. Therefore, risk or uncertainty with regard to the times over which the two processes takes place is higher or less certain than for the better established times of development of the four essential elements. This certainty relationship is important if a similar chart is constructed for a complementary play/prospect or assessment unit.

When an events chart is constructed for a complementary prospect, it becomes a risk chart. The risk chart is derived from the petroleum-system events chart which, in turn, is derived from the summation of the events chart for each oil and gas field in the petroleum system. These oil or gas fields are successful prospects. Unsuccessful prospects are dry holes. For example, if a risk chart for a prospect is similar to the petroleum-system events chart, then it can be concluded that this prospect is more likely to contain petroleum than one that has a dissimilar risk chart. Conversely, if an events chart is constructed for each dry hole within a petroleum system, they should be dissimilar from producing fields. This dissimilarity indicates where the greater uncertainty lies. Used this way, the events chart is a useful analytical tool to deal with uncertainty or risk.

One important issue this simple bar graph addresses is as follows: For an evolving petroleum system to effectively trap migrating hydrocarbon fluids, the trap forming process must occur before or during the generation-migration-accumulation process in order for petroleum to accumulate.

When constructing an events chart, these rules should be followed. First, there is only one pod of active source rock for each petroleum system. Second, every effective reservoir rock needs a seal, no matter how thin. Third, show only reservoir rocks

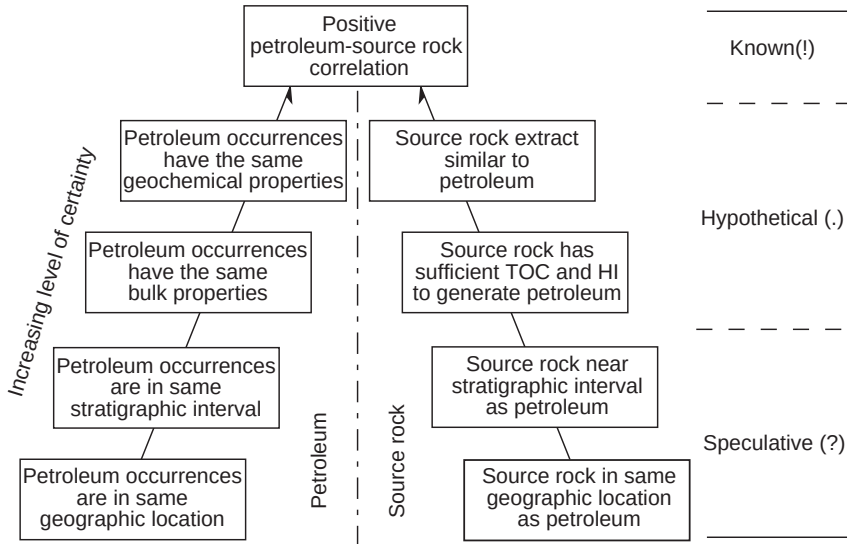


FIGURE 4 The logic sequence used for the levels of certainty from guessing to knowing. From Magoon, L. B., and Dow, W. G. (2000). Mapping the petroleum system—An investigative technique to explore the hydrocarbon fluid system. Memoir 73, p. 56 (as Figure 2). AAPG © 2000. Reprinted by permission of the AAPG whose permission is required for further use.

that contain petroleum accumulations, shows, or seeps. Fourth, show eroded overburden rock with hatcher lines so that it can be incorporated in the modeling exercise. Fifth, the best information for timing of trap formation comes from oil and gas fields. Sixth, the best information for generation-migration-accumulation comes from geological and geochemical information about the source rock that are then incorporated into the burial modeling and kinetics. This information indicates the onset, peak, and end of generation-migration-accumulation or when the active source rock is depleted (spent) or uplifted (inactive source rock). This process takes place over a relatively short period of time. Seventh, preservation time, by definition, starts when generation-migration-accumulation ends and continues to the present. Some petroleum systems have no preservation time. Last, when the critical moment occurs is, as discussed earlier, in the judgment of investigator, but modeling software packages are useful tools as they show the time over which expulsion from the pod of active source rock occurs.

4. THE PETROLEUM SYSTEM AS A WORKING HYPOTHESIS

A petroleum system investigation starts with a working hypothesis for generation, migration, and entrapment of petroleum in a province, based on available geological and geochemical data, which evolves as more data becomes available (Fig. 2A, B). The investigator starts with an oil and gas field map and related field data for the petroleum province of interest. The geographic location of the accumulations is important because accumulations located close together are more likely to have originated from the same pod of active source rock (Fig. 4). Accumulations that occur in the same or nearly the same stratigraphic interval are also likely to be from the same active source rock. In contrast, accumulations separated by barren rock sections are presumed to have originated from different pods of active source rock. Accumulations of widely differing bulk properties, such as gas versus oil, API oil gravity, gas-to-oil ratios, and sulfur contents, may also be presumed to originate from different pods of active source rock. Detailed geochemical data on oil and gas samples provide the next level of evidence for determining whether a series of hydrocarbon accumulations originated from one or more pods of active source rock. Last, comparing the geochemistry of oils and gases to possible source rocks provides the highest level of certainty as to which active source rock generated which oil or gas type.

By acquiring and organizing information that addresses these issues of location, stratigraphic position, and geochemistry, an investigator can take a working hypothesis of how a particular petroleum system formed to increasing levels of certainty (Fig. 2; Table I). The investigator organizes the

information on the oil and gas accumulations into groups of like petroleum types on the oil and gas field map, cross section and table (Figs. 2A, 2B, and 2C).

With this step completed, the investigator then locates all surface seeps on the oil and gas field map, which now becomes the petroleum system map. The seeps with the same geochemical composition as the subsurface accumulation provides geographic evidence for the end point of a migration path. The stratigraphic unit from which the fluid emanates can be compared to the stratigraphic unit in which oil and gas accumulations are found to determine the complexity of their migration paths. If the stratigraphic units are the same, then the migration paths are simple. If they are different, migration may be more complex. Geochemical information from seeps can be compared with that of discovered accumulations in order to link the seep fluid to the proper petroleum system.

Oil and gas shows in exploratory wells are added to the petroleum system map and cross section to better define the migration paths. As this map and cross section evolve, the investigator is encouraged to anticipate how the final map will look based on the framework geology and petroleum fluid information. Intuitively, exploration risk is high if the petroleum system is complicated and hence less predictable; risk is lower if the petroleum system is simple and thus more predictable.

After similar hydrocarbon fluids in the petroleum system have been mapped individual oil and gas accumulations are tabulated to better understand the size (by volume) and complexity of the petroleum system. The petroleum system table is organized by stratigraphic interval in each field (Fig. 2C). These stratigraphic intervals are zones, members, and formations that produce or contain measurable amounts of oil and gas. The table should include age of the stratigraphic interval, API gravity and sulfur content of the oil, gas-to-oil ratio (GOR), cumulative amount of oil and gas produced, and the remaining amount of oil and gas that can be produced. Other information the investigator may chose to include are lithology, gross and net thickness, porosity and permeability of the reservoir rock, geometry, closure, and area of the trap, and detailed geochemistry of the oil and gas. The information included in the table will depend on what is available and the objectives of the investigation. The required information is used to determine the size (by volume) of the petroleum system, which reservoir rock name is to be used in the name of the petroleum system, and to evaluate the complexity of the migration path.

Now the provenance or origin of the petroleum is mapped as the pod of active source rock. Only one pod of active source rock occurs in each petroleum system. A pod is a contiguous body of source rock that has or is expelling oil and gas. Because this pod has thickness and area, it can be mapped using well, outcrop, or geophysical data. When an organic-rich rock is in close, or reasonably close, proximity, both stratigraphically and geographically, to oil and gas accumulations, shows, or seeps, it is tentatively correlated with those fluids. Based on seismic, well, or outcrop data, the likelihood of correlation increases when the source rock's burial depth is known to reach 3 km, which in the experience of the authors is a reasonable minimum burial depth for thermal maturity or when burial depth modeling indicates a source rock is in of below the oil window. The correlation gains certainty if the source rock, by vitrinite reflectance or some other analytical technique, is established as being thermally mature. If the kerogen type of the source rock is consistent with that of the oil and gas, then confidence increases that the source rock is correctly correlated. If the geochemical composition of the organic matter in the source rock compares favorably with the migrated petroleum, the oil-source rock correlation is considered a match. Using seismic, well, and outcrop data, the suspected or confirmed active source rock is mapped as a contiguous, three-dimensional body, or pod, on the petroleum system map and cross section.

In this manner, the petroleum system map and cross section evolve, as the working hypothesis is taken to successive level of certainty. To further refine this work, a burial history chart and events chart are constructed and the petroleum system is named. This article discussed each of these petroleum system components in sequence, but they are frequently developed in parallel, and their relationship to each other is considered so that the petroleum system can be properly mapped. To organize these components, the petroleum system folio sheet is used.

SEE ALSO THE FOLLOWING ARTICLES

Natural Gas Resources, Global Distribution of • *Oil and Natural Gas Drilling* • *Oil and Natural Gas Exploration* • *Oil and Natural Gas Liquids: Global Magnitude and Distribution* • *Oil Recovery* • *Oil Refining and Products*

Further Reading

Dow, W. G. (1974). Application of oil-correlation and source-rock data to exploration in Williston basin. *American Assn. Petroleum Geol. Bull.* **58**(7), 1253–1262.

Law, B. E., and Curtis, J. B. (2002). Unconventional petroleum systems. *American Assn. Petroleum Geol. Bull.* **86**(11), 1851–1999.

Magoon, L. B. (1995). The play that complements the petroleum system–A new exploration equation. *Oil Gas J.* **93**(no. 40), 85–87.

Magoon, L. B., and Beaumont, E. A. (1999). Petroleum system. *In* "Exploring for Oil and Gas Traps" (E. A. Beaumont and N. H. Foster, Eds.), pp. 3.1–3.34. American Association of Petroleum Geologists Treatise of Petroleum Geology, Tulsa, OK.

Magoon, L. B., and Dow, W. G. (1994). The petroleum system–From source to trap. American Assn. Petroleum Geol. Memoir **60,** 655.

Magoon, L. B., and Dow, W. G. (2000). Mapping the petroleum system–An investigative technique to explore the hydrocarbon fluid system. *In* "Petroleum Systems of South Atlantic Margins" (M. R. Mello and B. J. Katz, Eds.). American Association of Petroleum Geologists Memoir **73,** pp. 53–68.

Magoon, L. B., and Schmoker, J. W. (2000). The total petroleum system–The natural fluid network that constrains the assessment unit. *In* "World Petroleum Assessment 2000" (U.S. Geological Survey World Energy Assessment Team, eds.). U.S. Geological Survey DDS-60, 2 CD-ROMs.

Magoon, L. B., and Valin, Z. C. (1994). Overview of petroleum system case studies. American Association of Petroleum Geologists Memoir **60,** 332.

Philanthropy and Energy

ERIC HEITZ
The Energy Foundation
San Francisco, California, United States

1. Introduction to Energy and Philanthropy
2. Qualities of Effective Philanthropy
3. The Merits of Energy as a Philanthropic Endeavor
4. Recent History: Energy and Philanthropy
5. Conclusions: The Future Role for Foundations in Energy

Glossary

energy/environmental nexus The fact that the world's energy problems cause the majority of the world's environmental problems.
foundation A philanthropic organization; assets remain tax free as long as 5% of those assets are given for charitable and educational purposes each year.
leverage The ratio of philanthropic investment and the quantifiable change that results.
philanthropy (1) A desire to improve the material, social, and spiritual welfare of humanity; (2) a term applied to institutions that disperse wealth in pursuit of these goals.
sustainable energy The energy that is produced and used in ways that support long-term human development in all its social, economic, and environmental dimensions.
Tax Reform Act of 1969 A reform of U.S. tax code, codifying the rules under which U.S. foundations operate.

Smart philanthropy can play a productive role in U.S. and global energy policy. It can help spur markets for the next generation of clean energy technologies that address the energy/environment nexus, especially global warming pollution. This article offers some principles for focused philanthropy in energy, discusses the merits of energy as a philanthropic endeavor, and outlines the recent history of energy and philanthropy in the United States.

1. INTRODUCTION TO ENERGY AND PHILANTHROPY

The best philanthropy involves the search for cause, an attempt to cure evils at their source.

—John D. Rockefeller, Sr.

Philanthropy's impact on society is both underappreciated and misunderstood. At their worst, the individual foundations that comprise the world's philanthropic sector are opaque institutions with vague goals; their funds disappear with no apparent trace. At their best, they are strategically driven, staffed by experts in the field, and can claim major contributions to the fields of human health and medicine, agricultural science, social policy, education, art, and, more recently, energy and the environment.

In essence, foundations are nonprofit institutions that disperse accumulated wealth. John D. Rockefeller and Andrew Carnegie, who built institutions to give away their considerable wealth at the end of the 19th century, are credited with birthing the modern grant-making foundation. The Tax Reform Act of 1969 codified their favorable tax treatment in the United States. In short, assets converted to foundation status are not taxed as long as 5% of those assets are given out each year for "charitable and educational purposes." Coupled with the growth in wealth in the last half of the 20th century, these rules spurred rapid growth of the foundation sector. Since 1975, the number of foundations has more than doubled and total giving (in constant dollars) has more than tripled. According to a report published by the American Association of Fundraising Counsel (AAFRC) Trust for Philanthropy, *Giving USA 2002*, over 56,000 registered foundations in the United States collectively gave out $27.6 billion in funding. Figure 1 shows their investments. Of the total, only 7% was for environmental purposes, and

 © 2004 Elsevier Inc. All rights reserved.

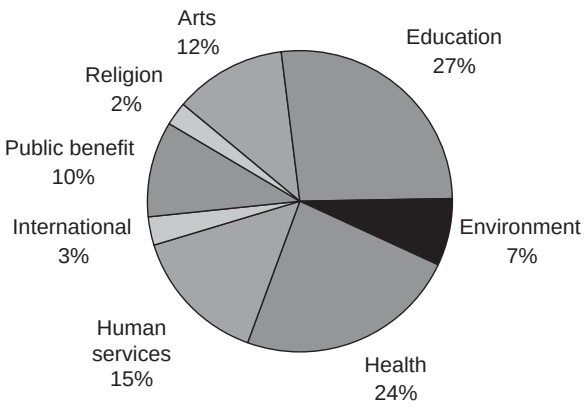


FIGURE 1 Philanthropic investment by field.

a fraction of that focused on energy. This investment pales in comparison to the nation's $700 billion annual energy bill, raising the question, "What is the role for philanthropy in energy?" Some answers are proposed here, first by offering some characteristics of good philanthropy, followed by looking at the merits of energy as a philanthropic investment strategy, including some recent results, then concluding with some thoughts on the future role for philanthropy in energy.

2. QUALITIES OF EFFECTIVE PHILANTHROPY

Foundations do not have the resources to fund directly solutions to the world's energy challenges, whether that means transitioning to a clean energy future or simply providing basic energy services in developing countries. Focusing on the United States, where the challenge is to transition to a clean and low-carbon energy system, total foundation investments in energy pale in comparison to the nation's energy research and development budget of $1.3 billion, which in turn is dwarfed by our annual energy bill. Given the scale of the investment flows compared to the size of the philanthropic resources, the theory of change adopted by all serious philanthropy work in energy is simple: public policy will change investment patterns, public or private, which will drive change (the U.S. experience is the focus here, but the same theory holds for international energy grant making). Since the 1970's, smart philanthropy has proved this approach, affecting policies that in turn change investment streams, often with a billion-to-one dollar or more leverage. Of course, it is impossible in a scientific way to attribute these changes solely to foundation investments; political change is too complex. But careful students of the best philanthropy find remarkable returns, as is discussed in Section 4. So the question looms, "what makes good philanthropy?" Borrowing liberally from the writings of Hal Harvey, Director of the Hewlett Foundation's Environment Program, and Dr. Kenneth Prewitt, former Vice President of the Rockefeller Foundation, the following list of answers is offered:

1. Target root causes: To begin, it is essential, as John D. Rockefeller, Sr. advised, to target root causes. Alleviating symptoms will never change the underlying dynamic that created them; it may bring solace, but it will not speed change. This is why energy makes such an interesting philanthropic topic: it is the root cause of many environmental, economic, health, and international security problems facing the world.
2. Be focused: The biggest mistake many philanthropies make is to dabble in many areas instead of going deep in one area. This advice holds with respect to the choice of topics (e.g., marine conservation, forest preservation, energy) and choice of philanthropic means (e.g., research and analysis, policy development, policy advocacy, grassroots advocacy, or media work). It is very difficult to understand the political dynamics of any given sector. Deep insight is required to understand capital flows, to learn who the "brain trust" for a specific field is, and to learn the inclinations of the key decision makers. To do this in many fields at once is difficult. With respect to means, learning to fund the best research demands different skills, then choosing an effective grassroots advocate or trying to influence the media. The best philanthropies—those with a track record of measurable change—focus narrowly on a few topics and specialize in a few means.
3. Hire professional staff: To see the best prospect for change requires expertise equivalent to the best businesses or advocates in the field. The myth that giving away money is easy and thus any generalist can take up the task is one of the most harmful approaches to quality philanthropy.
4. Be strategy driven: Thinking through, writing down, and peer reviewing an explicit strategy of change is essential to good philanthropy. There are many paths to change. Philanthropies should know precisely what path they seek and what series of steps has to happen to affect change. Foundations should expose the premise of their strategy for debate and adjust their strategies, refining them to adapt to changing conditions as needed.

5. Scale to the problem: Solving large, complex problems, such as the nexus of energy/environment problems, begs for large philanthropic investments. Though the scale of philanthropic investments pales in comparison to the $700 billion per year energy business, by focusing resources on key venues such as public utilities commission hearings or, in the transportation sector, air pollution authorities, foundation dollars can have a large impact.

6. Think long term: No serious energy or environmental problem can be solved overnight. Changing investment patterns, technologies, or consumer choices is a slow process, sometimes frustratingly so. Foundations, unlike most of the for-profit world, do not have next quarter's bottom line to define today's agenda. The best philanthropy takes the long view, investing for decades, not years.

3. THE MERITS OF ENERGY AS A PHILANTHROPIC ENDEAVOR

3.1 Energy as an Environmental and International Security Nexus

The goal of most philanthropy is to make the world better, whether improving conditions of the poor, stamping out disease, protecting endangered habitats, or increasing international security. Energy is interesting for philanthropy because it is simultaneously a key driver of economic growth and is at the center of Earth's most pressing environmental and security problems.

3.1.1 Environment

Fossil energy—coal and oil—has fueled tremendous economic growth over the past century, but at substantial, and growing, environmental cost. The U.S. Environmental Protection Agency (EPA) and Department of Energy (DOE) statistics show energy use is the leading cause of greenhouse gas emissions, smog, acid rain, air toxics, oil spills, and nuclear waste (Fig. 2). None of these problems can be solved without new energy trajectories. If the United States and the rest of world ramp up fossil fuel consumption to fuel further economic growth, there will be potentially catastrophic global environmental consequences. China, for example, is expected to surpass Western Europe's carbon emissions before 2010 and will surpass the United States to become the world's largest carbon emitter by about 2030. The Energy Information Administration (EIA) *International Energy Outlook 2003* report shows that Brazil and India are on similar trajectories for rapid fossil energy growth (Fig. 3).

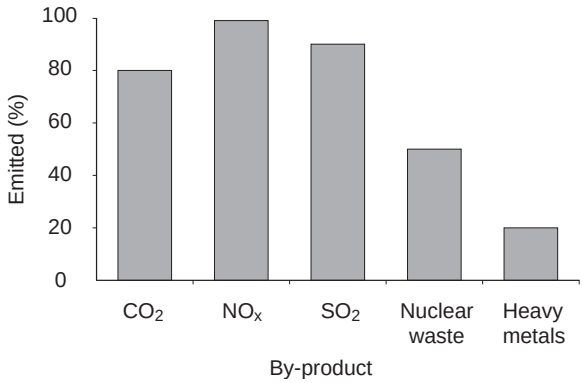


FIGURE 2 Percentage of emissions into the environment from energy use by-products.

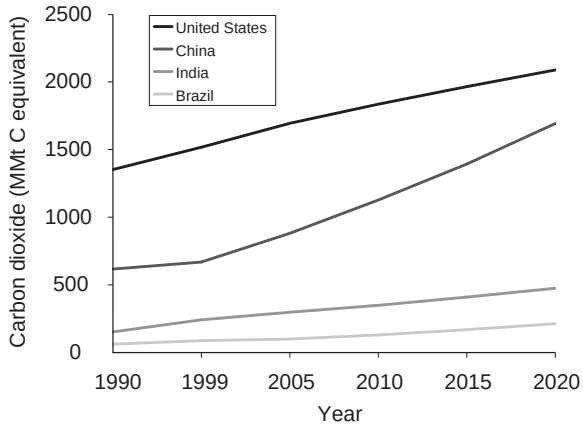


FIGURE 3 Projected carbon dioxide emissions. MMt C, million metric tons of carbon.

The energy crises of the 1970s inspired great concern about running out of fossil fuels. Geology and technology have since revealed that there are vast reserves of coal, and large reserves of natural gas and oil. A major crisis in energy, instead, turns out to be that the capacity of the environment—our atmosphere, forests, soils, rivers, and oceans—to absorb the wastes of our energy-intensive society is threatened and will suffer long before energy supplies will be exhausted. *Laissez-faire* economics, wherein prices rise due to supply scarcity and so drive investment in new energy options, will not solve this dilemma. Sensible policy is needed to move developed and developing nations to a sustainable energy path. Foundations can help create and advance new policies.

3.1.2 International Security

The 2003 war in Iraq, a country situated on the world's second largest remaining oil reserves, brings

home the international security implications of the U.S. economy's dependence on imported oil. The world's largest remaining oil supplies and, even more important in the short term, the swing production that most influences world market price, are in the politically volatile Middle East. We import more oil today than at any time in U.S. history. Worse, the latest government forecast by the Energy Information Administration, *Annual Energy Outlook 2003,* projects a 60% increase in domestic oil use over the next 25 years, with imports rising to 67.8%. And some developing countries are on even faster oil consumption growth trajectories. China's automobile market, for example, is growing at almost 35% per year. In the International Energy Agency report, *China's Worldwide Quest for Energy Security*, projections show oil imports increasing to 8 million barrels per day (mbd) by 2020 (from 0 mbd in 1993), further increasing competitive pressures on diminishing world oil supplies.

3.2 Promising Technology Solutions

The compelling question is whether it is possible to build a benign energy system given a large and still-growing world population and universal aspirations for higher income. The answer to that question must be found in technology transformation. Our society needs to invent a low-carbon, low-pollution affluence. The energy field is replete with examples of how this might be done and, over the past several decades, examples of how philanthropy can help accelerate change. It is clear that technologies are available to reduce drastically pollution and carbon from buildings, industry, and transportation. Recent philanthropic efforts focus on catalyzing this technology transformation in the building sector, transportation field, and power sectors. There are promising trends in each sector.

3.2.1 Buildings

The case for technology transformation is perhaps easiest to make in the buildings sector. It is neither difficult nor costly to design a building that uses only nominal energy for heating and cooling and that requires only a quarter of the electricity of a conventional building. In the mid-1980s, California revamped its building codes toward this end. Title 24, as the resulting codes were named, reduces energy use in newly constructed buildings by some 75% compared to precode buildings. Appliance standards are another example of the transformative power of new technologies. Refrigerators sold today are larger and have more features, yet use one-third the energy compared to 20 years ago. State appliance standards, eventually adopted by the federal government, drove these savings, with billion dollar savings for consumers. The American Council for an Energy Efficient Economy estimates that existing standards—for refrigerators, air conditioners, lights, washings machines, etc.—will save the equivalent of the energy generated by 400 large power plants, at net consumer savings of $160 billion by 2020. Standards being contemplated for adoption by the federal government today could save consumers another $75 billion dollars by 2020. Foundations have been the largest nongovernmental supporters of advancing such standards.

3.2.2 Transforming the Automobile

Transportation has the highest potential for technological advancement. The automobile is on the verge of a technological revolution. Today's automobiles are between 15 and 20% efficient; the lion's share of each gallon of gasoline goes directly out the exhaust pipe. Existing new technologies can double the efficiency of the engine, and double again the efficiency of the driveline, resulting in cars that exceed 100 miles per gallon with only modest air pollution (Fig. 4). High-efficiency tires, drivelines, and aerodynamic shells can reduce power requirements by a factor of two to four. Hybrid drives, which couple the efficiency of electric drives with the range and power output of liquid or gaseous fuels, can double engine/driveline efficiency. And fuel cells, which convert gas or liquid hydrogen directly into electricity at high efficiency with no moving parts and virtually no pollution, offer a very clean transportation future.

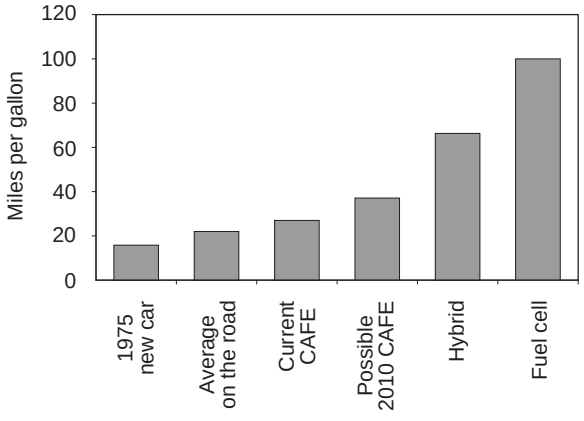


FIGURE 4 Automobile fuel efficiency. CAFE, Corporate average fuel economy (standards imposed under the Energy Policy and Conservation Act of 1975).

Toyota's Prius offers a hybrid drivetrain, including regenerative braking. The Prius gets almost 50 miles per gallon, but otherwise has features and performance similar to those of a Toyota Corolla. The Prius is exceeding all sales predictions, and Toyota recently doubled production volume. Honda has come out with a hybrid version of its popular Civic model, which gets 45 miles per gallon. Ford expects to introduce a sport utility vehicle (SUV) hybrid that is projected to obtain 40 miles per gallon, more than double today's average SUV mileage. Toyota is expected to offer an SUV on its Lexus 330 platform as well. GM has also announced hybrid versions. Foundations have a role in making the analytic case for these technologies and in spurring the policies that spur their commercialization.

3.2.3 Cleaner Power

Renewable energy technologies have enjoyed dramatic price reductions over the past two decades. Wind power has dropped from 25¢/kilowatt-hour to less than 3¢/kWh today. The central region of the United States is known as the "Saudi Arabia of wind energy." The 11 contiguous states from the Dakotas to Texas have the wind energy potential to supply more than the total current electricity consumption of the entire United States (the states are Montana, Wyoming, North Dakota, South Dakota, Minnesota, Iowa, Nebraska, Kansas, Colorado, Oklahoma, and Texas). Driven largely by state policies spurring wind investments—which foundation-supported advocates championed in many states—wind is the fastest growing source of generation in the United States. California, the world's fifth largest economy, has adopted a goal of 20% renewable power generation by 2015. This policy will spur nearly $10 billion in new wind investments. New York has announced a goal of 25% renewable energy by 2010, and 12 other states have similar renewable energy goals, which are driving new markets for renewable energy developers (Fig. 5).

4. RECENT HISTORY: ENERGY AND PHILANTHROPY

4.1 The Ford Foundation Energy Policy Project

Major foundation investments in energy began in the 1970s, driven partly by the emerging environmental movement and more directly by the Organization of Petroleum Exporting Countries (OPEC) oil crisis. Under McGeorge Bundy, former National Security Advisor to Presidents Kennedy and Johnson, The Ford Foundation launched The Energy Policy Project in 1971, ultimately spending $4 million on a series of energy studies. The most influential study, published in 1974, was *A Time to Choose*, by S. David Freeman. It was pathbreaking in two respects. First, it argued that economic growth and energy demands need not grow in lockstep—conservation and efficiency could decouple them. Second, it laid out the policies—from auto fuel efficiency regulations, to appliance efficiency standards, to federal energy efficiency research and development—that continue to be the mainstays of the U.S. energy efficiency policy today.

The Ford Foundation followed *Time to Choose* with two other major studies. In early 1979, they released *Nuclear Power Issues and Choices*, which concluded that, despite its many problems, nuclear energy should be a major power source in the future. Most important today, it argued that plutonium processing in breeder reactors was both uneconomic and created too great a danger of plutonium prolifera-

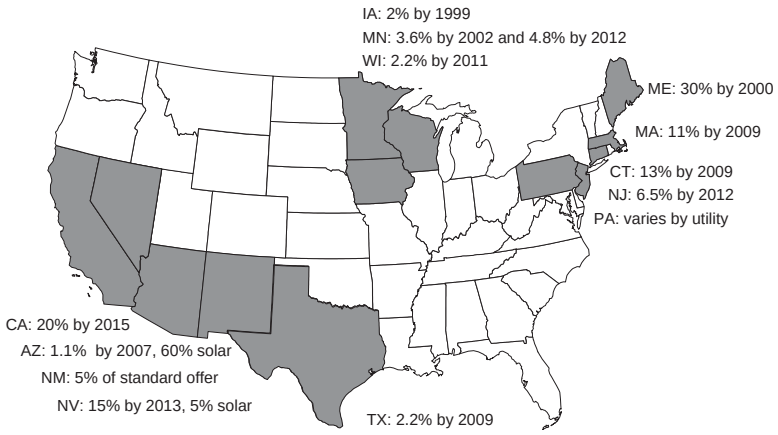


FIGURE 5 State renewable energy policies.

tion. One of the lead authors of the study, Dr. Spurgeon Keeny, Jr., argues that the study was a major factor in President Jimmy Carter's decision to defer commercial reprocessing of plutonium indefinitely. A third Ford Foundation study, released in 1979, *Energy, the Next Twenty Years*, was the least influential of the three studies; it concluded that the energy problem was essentially economic in nature. The U.S. would not run out of energy (given coal and oil shale reserves) any time soon, but prices would inevitably rise. Its main conclusion, that society would benefit from having the price of energy reflect its true value, continues to be the holy grail of energy economists today.

4.2 The Next Wave: From Studies to Advocacy

The premise of The Ford Foundation's investments in the 1970s was that serious new policy studies—the equivalent of investing $14 million today—would capture high-level decision-maker attention and move the country along a more "intelligent" energy path. As the decade progressed, and as the environmental movement matured, a handful of foundations began to invest based on a different premise—namely, that direct advocacy was needed to change policies, which, in turn, would shape energy investments for the country. The crucible in which these early advocacy efforts were born was the nuclear power debate. Small, environmentally focused foundations such as the Mertz Gilmore Foundation, Levinson Foundation, Nathan Cummings Foundation, New-Land Foundation, List Foundation, J. M. Kaplan Fund, Compton Foundation, and the Rockefeller Family & Associates supported technical advocacy in state regulatory proceedings to (1) attack nuclear industry cost and safety claims and (2) make the case for cheaper and cleaner alternatives, particularly natural gas power plants and energy efficiency. Historians attribute the decline of the nuclear industry to many factors, including cost overruns, the public movement against nuclear power, and the fear of accidents such as the 1979 Three Mile Island incident. The early foundation-funded advocacy efforts deserve some of the credit as well. They also spawned a new approach to foundation-supported advocacy that continues to this day. The leaders in the energy policy advocacy work included the Mertz Gilmore Foundation, an early supporter of global warming policy work and one of the first foundations to offer grants in support of renewable energy. Other foundations taking on energy issues at this time were the Jessie Smith Noyes Foundation, the Evergreen Fund, the Changing Horizons Charitable Trust, the Conservation and Research Foundation, and the Aria Foundation.

Following on the heels of Earth Day 1990 and the Gulf War in 1990/1991, which put global warming and oil dependency in the public eye, the presidents of the John D. and Catherine T. MacArthur Foundation, the Rockefeller Foundation, and the Pew Charitable Trusts launched a special-purpose foundation focused entirely on the energy/environmental nexus. Opening its doors in January 1991, the Energy Foundation received a $100 million commitment for its first decade. Focused entirely on advancing energy efficiency and renewable energy, the Energy Foundation's first decade concentrated primarily on advancing state policies to spur commercialization of clean energy technologies. In the utility sector, for example, foundation-supported advocates in a dozen states advanced policies to make energy efficiency investments for utilities as profitable as building power plants. By 1994, U.S. utility investments in energy efficiency nearly tripled to $2.7 billion per year, halving growth in electric demand. Even after a major restructuring of the utility industry, which slashed all capital investments by utilities, spending on energy efficiency by utilities today exceeds $1 billion per year.

The Energy Foundation also invested heavily in state efforts to advance the cleanest and most efficient cars. Beginning in California in the early 1990s, and moving to New York, Massachusetts, Maine, and Vermont, scientists, energy experts, and grassroots groups worked in state air regulatory proceedings, arguing that only the most advanced vehicle technologies would clean the air adequately to meet public demands for healthy air. Because California, New York, Massachusetts, Maine, and Vermont represent 20% of the U.S. automobile and light-truck market, new tailpipe requirements captured the attention of the automobile industry. California's low-emission vehicle and zero-emission vehicle standards, and the subsequent adoption of the standards by New York and Massachusetts, are credited by many experts as the primary drivers spurring new investments in hybrid technology and fuel cell vehicles. Today, the Toyota Prius and Honda Civic hybrids are five-passenger production cars that are superclean and get nearly 50 mpg in real-world driving. Toyota's goal is to sell 300,000 hybrids annually worldwide by the end of this decade. GM, Ford, and Chrysler have all announced plans for hybrid vehicles as well. The newly revised California Air Resources Board rules, released April 2002, called for minimum annual sales of 110,000 hybrids by 2009.

These two examples—working toward reduction of electric demand and advocating for fuel efficiency—are indicative of Energy Foundation work. The foundation continues to invest in the power sector to advance efficiency and renewable energy, to promote superefficient building technologies, and to spur commercialization of advanced-technology vehicles. The Energy Foundation has evolved into a project of major foundations advancing clean energy technologies. In 1996, the Mertz Gilmore Foundation joined, and in 1998, the McKnight Foundation joined. The David and Lucile Packard Foundation joined in 1999, launching the China Sustainable Energy Program, which helps China promote energy efficiency and renewable energy technologies and employs nine Chinese nationals in a Beijing office. The William and Flora Hewlett Foundation joined in 2002, supporting advanced vehicle technology work in China and the United States as well as work to advance clean energy in the western United States. The Energy Foundation annual budget is now around $20 million per year.

The Energy Foundation is the only special-purpose energy philanthropy in the United States, but it is by no means alone in supporting energy work. Under its Environment Program, the Pew Charitable Trusts has been a major investor making the case for action on global warming emissions. At the federal level, this includes the Pew Center on Global Climate Change (www.pewclimate.org), which built a coalition of major businesses that endorse action on climate and which has been a player in international deliberations and U.S. policy debates. The Pew Charitable Trusts, independent of the Energy Foundation, also built a state network of advocates focused on the health and environmental damages from the many Eisenhower-era coal power plants that still generate a significant share of U.S electric power. Many of these power plants—in Texas, Ohio, Minnesota, and North Carolina—have been cleaned up or shut down as a result, and the federal debate to cap power sector emissions and, possibly, control carbon emissions from power plants is active. The William and Flora Hewlett Foundation is also a major new investor in energy work, beyond its investment in the Energy Foundation. Hewlett is focused on advanced vehicle technology in Mexico and Brazil, as well as on energy production on western U.S. lands. The Turner Foundation has also made energy investments, as have the Rockefeller Brothers' Fund and the W. Alton Jones Foundation. Major regional funders have also invested in energy, especially the Joyce Foundation in the midwestern United States, the John Merck Fund in New England, and the Bullitt Foundation in the Pacific northwest, among others. This author estimates that annual philanthropic support for energy (broadly defined to include global warming work) in the United States is about $40 million per year in 2003.

5. CONCLUSIONS: THE FUTURE ROLE FOR FOUNDATIONS IN ENERGY

Foundations have made good progress on the energy front, but huge gaps remain. Promising technologies in use and production today prove that the prospects for an affluent but low-carbon society are good. But the U.S. and the world economies continue to run primarily on oil and coal. As the largest consumer and producer of energy in the world, the United States has the responsibility to lead the transition to cleaner technologies. Philanthropy can play a role in this transition by spurring policies that commercialize the best technologies.

But U.S. leadership is not enough. Rapidly developing economies, such as those of China, India, and Brazil, cannot afford to follow our path of energy consumption—they need opportunities to leapfrog to advanced technologies that offer higher standards of living but with radically reduced energy use. Here, too, philanthropic opportunities in energy abound, with potential billion-to-one payoffs. In essence, we are confronted with the ultimate "diffusion of innovations" challenge. Technology learning curves—for clean technologies such as wind power, photovoltaics, or superefficient gas turbines—tell us that each cumulative doubling of production can reduce prices by 20–30%. In turn, lower prices speed the penetration of these new energy technologies into U.S. and world markets. The history of the past three decades suggests energy policy, in the absence of higher price signals driven by energy scarcity, will be the major factor in how fast we move down these clean technology cost curves and toward a transformed energy economy. They also suggest that smart philanthropy can play a productive role.

SEE ALSO THE FOLLOWING ARTICLES

Equity and Distribution in Energy Policy • *Nongovernmental Organizations (NGOs) and Energy* • *United Nations Energy Agreements* • *World Environment Summits: The Role of Energy*

Further Reading

Center on Philanthropy at Indiana University. (2002). "Giving USA 2002." American Association of Fundraising Counsel (AAFRC) Trust for Philanthropy, Indianapolis.

Energy Foundation. (2004). Available on the Internet at http://www.ef.org/.

Environmental Grantmaking Foundations. (2003). "Environmental Grantmaking Foundations," 2003 ed. Resources for Global Sustainability, Inc., Cary, North Carolina.

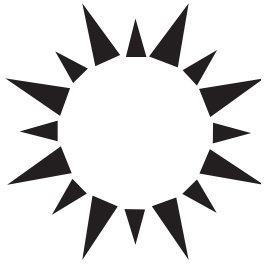
Harvey, H. (1999). The importance of focus in effective philanthropy. *In* "New Directions for Philanthropic Fundraising, The Changing World of Foundation Fundraising: New Challenges and Opportunities" (S. A. Glass, Ed.). Jossey-Bass, San Francisco. Chap. 2.

Nielsen, W. A. (1996). "Inside American Philanthropy: The Dramas of Donorship." University of Oklahoma Press, Norman.

Prewitt, K. (1987). Foundations. *In* "The Nonprofit Sector: A Research Handbook," (W. W. Powell, Ed.). Yale University Press, New Haven Publication forthcoming.

Prewitt, K. (1995). "Social Sciences and Private Philanthropy: The Quest for Social Relevance. Series on Foundations and Their Role in American Life. 15." Indiana University Center on Philanthropy, Indianapolis.

Photosynthesis and Autotrophic Energy Flows



JED P. SPARKS
Cornell University
Ithaca, New York, United States

1. **Introduction**
2. **The Photosynthetic Process**
3. **Chemoautotrophy**
4. **Photorespiration**
5. **Photosynthetic Pathway Variation**
6. **Global Carbon Cycling: Photosynthetic Considerations**
7. **Global Climate Implications**
8. **Conclusion**

Glossary

adenosine triphosphate (ATP) Energy-bearing molecule formed during the light reactions through the phosphorylation of adenosine diphosphate. The energy bound in ATP is used as the primary energy source for most biological reactions.

C_4 photosynthesis In order to reduce the amount of carbon lost during photorespiration, some plant have developed C_4 photosynthesis. These plants initially carboxylate the acceptor molecule phosphoenolpyruvate to form four-carbon organic acids (hence the term C_4 photosynthesis). These molecules are transported internally to specialized bundle sheathe cells. There, the plant enzymatically breaks down the acid and releases CO_2. The CO_2 is then fixed again using the normal enzymes of the dark reactions. As a result, the concentration of CO_2 at the point of the dark reactions is much higher than that of the surrounding atmosphere and effectively eliminates photorespiration.

chemoautotrophs Organisms that use CO_2 as a source of carbon, but rather than using light energy they obtain reducing power and energy from the oxidation of inorganic compounds.

dark reactions The process of using the reducing power and energy produced during the light reactions to fix inorganic material into high-energy organic molecules.

light reactions The process of coupling light energy to the creation of chemical reducing power and energy. This process occurs in the thylakoid membrane in higher plants.

nicotinamide adenine dinucleotide phosphate (NADPH) Reducing molecule formed during the light reactions. Reducing energy of NADPH is used to reduce (i.e., store energy) inorganic material making organic high-energy molecules.

oxidation–reduction reaction Chemical reaction involving the transfer of electrons. Reduction is the transfer of an electron from a donor molecule to an acceptor molecule. During the reaction, the donor is said to become oxidized and the acceptor reduced.

photorespiration The RuBP carboxylase–oxygenase enzyme also exhibits oxygenase activity (i.e., it will catalyze the reaction of O_2 with RuBP). This leads to the loss of a CO_2 molecule previously fixed and decreases the efficiency of photosynthetic carbon assimilation.

photosynthesis The process of converting light energy to organic chemical energy.

RuBP carboxylase oxidase Arguably one of the most important enzymes on Earth. RuBP carboxylase oxidase (or Rubisco) catalyzes the reaction of CO_2 and an acceptor molecule, ribulose bisphosphate (RuBP), to form 3-phosphoglyceric acid (3-PGA). This is the initial carboxylation step of the dark reactions of photosynthesis.

Photosynthesis and the resultant autotrophic energy flow is a multistep pathway that is arguably the most important set of chemical reactions on earth. The reduction of compounds (i.e., the storing of energy in chemical bonds) is the foundation of all biological metabolism. Photosynthesis and, to a lesser extent, chemoautotrophy are the processes of this reduction. The capture of light energy or the oxidation of inorganic material produces energy and reducing power to form high-energy molecules necessary for life. This article gives a general introduction and overview of photo- and chemoautotrophic reactions and the organisms who perform these reactions, and

 © 2004 Elsevier Inc. All rights reserved.

it briefly outlines the recent human impact on photosynthetic processes and global carbon cycling.

1. INTRODUCTION

Photosynthesis or photoautotrophism is the mechanism of converting light energy to organic chemical energy performed by certain bacteria, algae, and plants. The photosynthetic process consists of both physical and chemical reactions enabling organisms to exploit an energy resource external to the Earth's system: the sun. The organic molecules produced by these organisms provide the energy powering nearly all living organisms. In short, photosynthetic reactions directly or indirectly provide energy for most biological reactions and may be expressed as:

$$\text{Inorganic molecule} + \text{light energy} \rightarrow \text{high-energy organic molecule}$$

According to the laws of thermodynamics, complex systems will tend to move from a state of organization (i.e., high energy) to a random state (i.e., low energy) through time until equilibrium is reached. In the case of living organisms, equilibrium between the organism and the environment is the death and decomposition of the organism. Therefore, to stay alive an organism must consume energy to reverse the thermodynamic direction. Living organisms are not in balance in terms of thermodynamics and require energy to maintain structure, accumulate matter, conduct metabolism, grow, and, eventually, reproduce.

All nonphotosynthetic organisms, such as animals, fungi, and bacteria, are dependent on organic materials originally generated through photosynthetic (or chemosynthetic) reactions. Without this supply of high-energy molecules, organisms would not survive. Only those organisms able to utilize the supply of energy from the sun increase the total free energy of the earth and increase the energy available to organisms unable to use either chemical or solar energy sources, including humans.

The photosynthetic process is found in some bacteria, cyanobacteria (often referred to as the blue-green algae), algae, and higher plants. Given a supply of sunlight, photosynthetic organisms can survive, grow, and reproduce using only inorganic forms of matter. Molecules, known as pigments, that absorb light and convert physical energy to chemical energy have allowed the evolution and proliferation of complex multicellular organisms, dependent on light as their ultimate energy source.

2. THE PHOTOSYNTHETIC PROCESS

In simple terms, photosynthesis is the process by which light energy is used to convert chemical substances to a more energy-filled state. In all photosynthetic organisms, it involves the capture of the energy of a photon of light by a pigment molecule, the formation of an excited electron, and the use of the excited electron to reduce an acceptor substance and to form reduced or energy-rich molecules (Fig. 1). Photosynthesis is often thought of as a two-stage process. The first stage, the light reactions, is the process of coupling light energy to the creation of chemical reducing power and energy. The second stage, the dark reactions, is the process of using the reducing power and energy produced during the light reactions to fix inorganic molecules into organic molecules. This two-stage process is outlined in detail in the following sections.

2.1 Oxidation–Reduction Reactions

The foundations of photosynthesis are a series of oxidation and reduction reactions. Very simply defined, reduction is the transfer of an electron (e^-) from a donor (M_d) molecule to an acceptor molecule

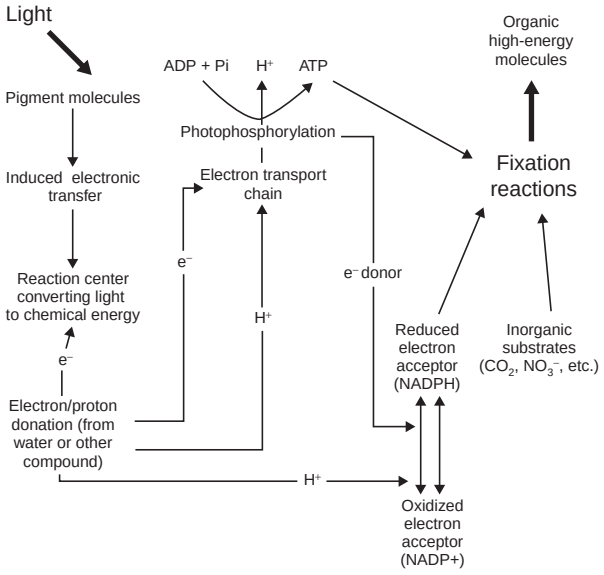


FIGURE 1 The basic reactions found in most photosynthetic organisms.

(M_a). During the reaction, the donor is oxidized and the acceptor is reduced. An electrically neutral compound becomes negatively charged and will extract a proton from water to restore electrical neutrality:

$$M_d + M_a \rightarrow = M_d^+ + M_a^-;\ M_a^- + H^+ \rightarrow M_aH;\quad M_d^+ + e^- \rightarrow M_d.$$

Reduction and oxidation reactions are fundamental to the photosynthetic process. The primary reaction of photosynthesis is the linking of light energy to a biochemical process. This is accomplished by the transfer of electrons from a light-excited chlorophyll pigment to an acceptor molecule. The acceptor molecule (in higher plants, a quinone) is reduced and the chlorophyll oxidized. Electrons are then donated from various sources (water in the case of higher plants) to the oxidized chlorophyll, reducing it and allowing the process to repeat.

The separation of the photosynthetic process into the light and dark reactions is primarily a historical artifact and has led to confusion regarding the use of terms. The generation of both the reducing [nicotinamide adenine dinucleotide phosphate (NADPH)] and the energy-containing molecules [adenosine triphosphate (ATP)] requires light energy, and these are logically referred to as the light reactions. The NADPH and ATP molecules are subsequently used to reduce other inorganic substances and convert them to organic molecules. With an adequate supply of NADPH and ATP, the chemical reactions of photosynthesis will proceed irrespective of light. The dark reactions of photosynthesis can and do occur in the light and darkness.

2.2 Capture of Light Energy and Electron Donation: The Light Reactions

Pigment molecules of various kinds allow photosynthetic organisms to both capture light and convert the captured light energy into chemical energy. Light-harvesting pigments (primarily the chlorophylls) capture light from the sun and form excited pigments. The combined excitation energy of multiple pigment molecules is focused in a single pigment molecule known as the reaction center (RC) and used to reduce acceptor molecules. The light-harvesting apparatus (the RC plus multiple accessory pigments) or antenna pigments absorb light and excitation energy is produced. This excitation energy moves through the antenna molecules to the RC where an electron is excited and lost. The electron that is lost from the pigment is then replaced by an electron from another source in the environment. In photosynthetic bacteria, a wide variety of compounds may donate electrons to the oxidized pigment. In cyanobacteria, algae, bryophytes, ferns, and higher plants water is used as the source of electrons and diatomic oxygen (O_2) is released:

$$2\,H_2O + \text{light energy} + \text{photosynthetic pigments} \rightarrow O_2 + 4H^+ + 4e^-.$$

The most common pigment molecule in higher plants is chlorophyll, but there are many other pigments involved in photosynthesis across different organisms, including phycocyanin, phycocoerythrin, and allophycocyanin in blue-green and red algae and carotenoids and xanthophylls in most higher plants and some algae and bacteria.

In both bacteria and higher photosynthetic organisms the excited electrons are used to form reducing molecules, either nicotinamide adenine dinucleotide (NADH) or, more commonly in photosynthetic organisms, NADPH. Simultaneously, through electron transport, protons accumulate on one side of a cellular membrane, forming a proton concentration gradient, the energy of which is used to produce ATP through the phosphorylation of adenosine diphosphate. ATP and NADPH provide reducing power and energy to nearly all biological energy and matter transformation reactions.

2.3 Reduction of Inorganic Material: The Dark Reactions

All organic biomass created by photosynthetic organisms is generated using reductant (NADPH) and energy (ATP) from the light reactions. This energy (i.e., ATP and NADPH) of photosynthesis may eventually be used to reduce a range of organic substances. The most common compound reduced is carbon dioxide (CO_2), leading to the synthesis of common sugars (CH_2O) and other organic molecules based on carbon. The fixation of carbon is the focus of this section. However, ATP and NADPH from the light reactions are also used to assimilate other inorganic compounds. Nitrate ions (NO_3^-) are reduced to ammonia, which is synthesized to amino acids and, eventually, proteins:

$$NO_3^- + 9\,H^+ + 8\,e^- \xrightarrow{\text{nitrate and nitrate reductase enzymes}} NH_3 + 3\,H_2O.$$

Sulfate ions must also be reduced before entering biological metabolism:

$$SO_4^{2-} + 9\,H^+ + 8\,e^- \xrightarrow{\text{enzymes}} HS^- + 4\,H_2O.$$

Photosynthesis, although almost always associated with the fixation of carbon, is a process with many possible products and it participates in many facets of metabolism.

The dark reactions of photosynthesis in higher plants occur within the chloroplast. The rate of CO_2 fixation is determined by the enzyme activity, the supply of CO_2, and the conditions within the chloroplast. The assimilation of carbon dioxide is a cyclical process. The assimilation (i.e., fixation or reduction) of CO_2 can be simply described as

$$\begin{aligned} &3\,CO_2 + 9\,ATP + 6\,NADPH + 5\,H^+ \\ &\quad \rightarrow C_3H_5O_3P\ (\text{simple sugar}) + 9\,ADP \\ &\quad + 8\,P_i + 6\,NADP^+ + 3\,H_2O + \text{energy}. \end{aligned}$$

This scheme is a simplification listing only the substrates and energy inputs. The process includes 11 enzymes catalyzing a 13-step carboxylation cycle. In the initial carboxylation reaction, the enzyme ribulose bisphosphate (RuBP) carboxylase catalyzes the reaction of CO_2 and an acceptor molecule, RuBP, to form 3-phosphoglyceric acid (3-PGA). RuBP is then regenerated from 3-PGA in reactions consuming NADPH and ATP. For each CO_2 assimilated, a minimum of three ATP and two NADPH are needed from the light reactions. The roles of the dark reactions are therefore to regenerate RuBP for additional CO_2 assimilation and to produce simple sugars.

Interestingly, the enzyme RuBP carboxylase exhibits dual activity. RuBP carboxylase will catalyze CO_2 assimilation as described previously or the assimilation of O_2—oxygenation. Because of this dual functionality, the enzyme is often referred to as RuBP carboxylase oxygenase or Rubisco. The ability of Rubisco to catalyze oxygenation reactions has implications for photosynthetic energy flows and will be considered later.

2.4 Variations in Photosynthesis among Biological Organisms

The most primitive type of photosynthesis is thought to be that exhibited in the halophilic bacterium, *Halobacterium halobium*. These organisms have patches of the pigment bacteriorhodopsin on their outer membrane. This pigment captures light and transports H^+ from inside the cell to the outside. This process forms a pH/electrical gradient that the organism then uses to synthesize ATP when the H^+ diffuses back into the cell. This is the simplest form of photosynthetic reaction and is not associated with more complex electron transport systems found in other photosynthetic organisms.

Most photosynthetic bacteria exhibit a slightly more complex photosynthetic system in which electron transport is used to form a pH gradient for ATP synthesis. An electron transport chain greatly increases the efficiency of ATP synthesis and allows for the generation of NADPH. To facilitate an electron transport chain, the outer cell membrane of most photosynthetic bacteria is folded, creating an enclosed space where a chemical gradient can be generated between the pocket and the interior of the cell. Within these pockets, bacterial chlorophyll and associated pigments bound in the membrane capture light and form a reduced acceptor. Simultaneously, the transported electrons reduce quinone compounds within the membrane that transport H^+ from the interior of the cell to the pockets. A pH gradient is created and the diffusion of H^+ back into the cell drives the synthesis of ATP. In contrast to *Halobacterium*, these more complex photosynthetic bacteria utilize electron transport and produce NADH in addition to ATP.

The prokaryotic cyanobacteria (blue-green algae) produce oxygen during photosynthesis similar to higher plants and have a membrane structure similar to that of the complex bacteria: extensive membranes arranged around the exterior of the cell in sheets. The red algae are also similar to the complex bacteria. The membranes have regularly arranged patches of pigment called phycobilisomes. The phycobilisomes contain a grouping of pigments linked to a single reaction center. There are two types of complexes in the membranes, one associated with electron transport (photosystem I) and the other associated with water oxidation and electron further transport (photosystem II). As in the bacteria, transport of electrons leads to the synthesis of reductant (NADPH) and the formation of a pH gradient, which drives the synthesis of ATP.

In higher algae, bryophytes, ferns, and higher plants, the photosynthetic membranes (thylakoids) are enclosed within a double-membraned organelle known as a chloroplast. Multiple chloroplasts can be found within a single plant cell. The inside of the thylakoid membrane is called the lumen and is chemically equivalent to the outside of the cell. There are two photosystems, as in the cyanobacteria, but the higher photosynthetic organisms use some form of chlorophyll as the primary pigment and a variety of secondary pigments. Also, as in the cyanobacteria, electron transport leads to the reduction of an acceptor, the formation of a pH gradient, and the synthesis of ATP.

3. CHEMOAUTOTROPHY

Not all organisms require light to reduce inorganic molecules. These organisms are referred to as chemoautotrophs and are found entirely within the bacteria. Chemoautotrophy has received considerable attention since the discovery of several deep-sea vent communities in the Galapagos rift in 1977. The areas surrounding these sea vents contain entire communities based on energy from chemoautotrophic bacteria; primarily sulfur oxidizers. These bacteria use CO_2 as a source of carbon and obtain reducing power and energy from the oxidation of elemental sulfur, hydrogen sulfide, or thiosulfite. The submarine volcanic vents with which these organisms are associated discharge large quantities of sulfide-rich warm water. Although unique in the sense of an entire community based on chemoautotrophically derived energy, deep-sea vent communities are not the only environments to support chemoautotrophic bacteria. Other sulfur-oxidizing bacteria are known in deep freshwater lakes, in surface springs, and in caves. In addition, other chemosynthetic bacteria are known to oxidize ammonium (NH_4^+), nitrite (NO_2^-), iron (Fe^{2+}), hydrogen (H_2), or carbon monoxide (CO). Of these, the nitrifying bacteria, which oxidize ammonium to nitrite and nitrite to nitrate, are among the most ecologically important organisms on Earth due to their role in the cycling of global nitrogen between organic and inorganic pools. Without them, much of the nitrogen necessary for plant growth would be unavailable.

4. PHOTORESPIRATION

The oxygenase activity of the RuBP carboxylase–oxygenase enzyme is significant because it leads to the loss of a CO_2 molecule previously fixed and decreases the efficiency of photosynthetic carbon assimilation. When RuBP catalyzes an oxygenase reaction, it produces a phosphoglycolate (PG) and 3-PGA. The PG then moves out of the chloroplast and is metabolized by the glycolate pathway and leads to the eventual release of CO_2. This is a form of respiration to the plant and offsets the amount of carbon fixed during assimilation. This process occurs during the light and is directly associated with photosynthesis; it is therefore referred to as photorespiration. The amount of photorespiration (i.e., the affinity of Rubisco for oxygen as a substrate) is dependent on temperature and the ratio of O_2 to CO_2 in the chloroplast. Early in Earth's history when the photosynthetic reactions were thought to have evolved, the ratio of O_2 to CO_2 concentration was relatively low compared to today. Under these conditions, photorespiration is negligible. However, in the current atmosphere, in which the partial pressure of O_2 is 1000 times higher than that of CO_2, 20–30% of the fixed carbon may be lost to photorespiration. Some plants have developed mechanisms that elevate the concentration of CO_2 inside the leaves, thereby decreasing or eliminating photorespiration. These specialized mechanisms are discussed in the following section.

5. PHOTOSYNTHETIC PATHWAY VARIATION

To minimize the effects of photorespiration, some plants have developed anatomical and biochemical characteristics allowing them to concentrate CO_2 at the site of Rubisco activity. The traditional photosynthetic dark reactions are often referred to as C_3 photosynthesis because the initial product of carboxylation is the three-carbon compound, 3-PGA. Some plants in hot and dry environments, in which photorespiratory loss of carbon is very high, have developed an alternative pathway, the C_4 pathway, to minimize carbon loss. The internal leaf anatomy and biochemistry of a C_4 plant is significantly different than that of a C_3 plant (Fig. 2). C_3 plants have vascular bundles surrounded by a layer or sheathe of large colorless cells. Other cells, known as palisade cells, contain chlorophyll and are arranged against the

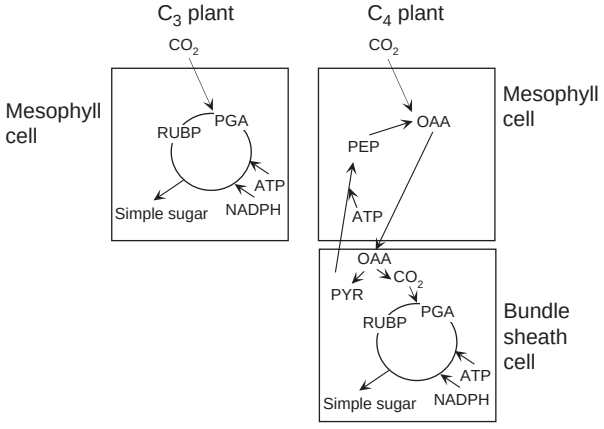


FIGURE 2 The leaf anatomy and biochemistry of a typical C_3 plant compared to a typical C_4 plant. The specialized structure and biochemistry allow C_4 plants to concentrate CO_2 inside the leaf and avoid losing carbon through photorespiration.

upper edge of the leaf surface. In contrast, plants utilizing the C_4 pathway have vascular bundles surrounded by a sheathe of cells rich in chlorophyll. C_4 plants initially carboxylate the acceptor molecule phosphoenolpyruvate (PEP) using the enzyme PEP carboxylase to form oxaloacetate, which is then converted to one of two four-carbon organic acids, malic or aspartic acid (hence the term C_4 photosynthesis). These molecules are transported internally to the specialized bundle sheathe cells. There, the plant enzymatically breaks down the malic and aspartic acid to release the fixed CO_2. The CO_2 is then fixed again using the normal enzymes of C_3 photosynthesis to form three-carbon molecules. C_4 plants do not replace the C_3 cycle; the PEP in conjunction with the bundle sheathe anatomy acts as a CO_2 pump to concentrate CO_2 at the site of Rubisco fixation in the bundle sheathe cells. As a result, the concentration of CO_2 within the bundle sheathe cells is much higher than that in the surrounding atmosphere and effectively eliminates photorespiration.

Under scarce water conditions, a third photosynthetic pathway has developed: crassulacean acid metabolism (CAM). These plants tend to be succulents and members of the plant families Cactaceae, Euphorbiaceae, and Crassulaceae. The CAM pathway is similar to the C_4 pathway in that CO_2 is first fixed into a four-carbon acid, malate, using the enzyme PEP carboxylase. The malate is then enzymatically cleaved, yielding a CO_2 molecule, which then participates in the traditional dark reactions of C_3 photosynthesis. Unlike C_4 plants, in which these two steps are physically separated in the mesophyll and bundle sheath cells, in CAM plants both steps occur in the mesophyll cells but at different times. The CAM plant opens its stomata at night, taking up CO_2 and fixing it to malate using PEP carboxylase. During the day, the plant closes its stomata and converts the malate into CO_2, which is then fixed using the normal C_3 dark reactions. By only opening stomata at night when the temperature is lowest and relative humidity highest, CAM plants conserve water. Although CAM plants use water very efficiently, the pathway is energetically expensive and is not usually observed except in very dry environments.

6. GLOBAL CARBON CYCLING: PHOTOSYNTHETIC CONSIDERATIONS

The global cycling of carbon involves both biological and physical processes. Only the biological components are discussed here. Plants and other photosynthetic organisms assimilate CO_2 and eventually convert the CO_2 to simple sugars. From simple sugars, plants synthesize a variety of compounds and store them in the form of plant tissue. When herbivores consume plant material, simple sugars are synthesized and transformed to other carbon-based compounds. Predators feeding on the herbivores redigest and again transform the carbon. At each step, some of the carbon returns to the atmosphere in the form of CO_2 as a by-product of respiration. The remainder is incorporated and temporarily stored in living biomass. Carbon contained in living organisms is eventually released through decomposition. The rate of release is highly variable and depends on both environmental and biological conditions. In tropical forests, most of the carbon in detritus is rapidly recycled. In drier, cooler regions considerable quantities of carbon are stored in the soil for long periods of time.

Similar cycling takes place in aquatic environments. Phytoplankton assimilate carbon dioxide that has diffused into the upper layers of water and dissolved carbonate and bicarbonate ions and convert it to organic sugars. The carbohydrate produced eventually passes through the aquatic food web. Significant portions of carbon, bound as carbonate in shells and exoskeletons of organisms, settle to the lake or ocean floor. This carbon is largely isolated from biotic activity and becomes incorporated into bottom sediments. It may then enter a long-term pool and may not recycle to the atmosphere again until it resurfaces as limestone through geological time.

The total global carbon pool involved in global carbon cycling is estimated to be approximately 55,000 Gt (1 Gt = 1 billion metric tons). Fossil fuels, the result of ancient photosynthesis, account for approximately 10,000 Gt. The ocean contains approximately 35,000 Gt, primarily as bicarbonate (HCO_3^-) and carbonate ions (CO_3^-). Dead organic matter in the ocean contains 1650 Gt of carbon. Living biomass in the ocean, primarily phytoplankton, is 3 Gt. Terrestrial ecosystems contain approximately 1200 Gt as dead organic matter and 550 Gt as living matter. The atmosphere holds approximately 740 Gt of carbon.

Although global carbon cycling is a combination of both physical and biological processes, photosynthesis is obviously a central controlling factor. In order to understand the global carbon cycle, we must understand how photosynthetic reactions occur over broad temporal and spatial scales.

Human activities have severely altered the global carbon cycle during the past century, with unknown future consequences.

7. GLOBAL CLIMATE IMPLICATIONS

The flux of carbon among terrestrial, aquatic, and atmospheric pools is at least partially controlled by photosynthetic processes. The relative carbon pool sizes have major effects on global climate. The CO_2 flux among land, sea, and atmosphere has, since the industrial revolution, been disturbed by a rapid injection of carbon dioxide to the atmosphere from the burning of fossil fuels and from the clearing of forests. Both the fossil fuel and biomass carbon are relatively long-lived pools with long residence times. However, human activity has volatilized this carbon originally fixed by photosynthesis into the atmosphere. At the beginning of the industrial revolution, atmospheric CO_2 was approximately 260 ppmv. Since then, input of CO_2 has increased this concentration to approximately 372 ppmv in 2002.

Only approximately 40–50% of the total CO_2 injected into the atmosphere remains there. What happens to the rest is a subject of current research and not completely understood. Most likely, the ocean is a large sink for this excess carbon. The other potential major sink is the terrestrial biota. Terrestrial ecosystems, especially forests, do have the ability to store some carbon through the photosynthetic production of biomass. However, whether this storage is greater than the contribution of carbon through respiration is not clear in many terrestrial systems. A complete mechanistic understanding of the global carbon cycle has yet to be achieved, and this will be an area of active scientific endeavor for many years. However, there is very little disagreement that atmospheric CO_2 has increased dramatically during the past century and that if atmospheric CO_2 concentrations continue to increase at the present rate, it will exceed 400 ppmv by 2030. In view of our dependence on fossil fuels and the growing body of evidence linking atmospheric CO_2 and climate, understanding the fate of excess atmospheric carbon is one of the most pressing scientific research needs.

In addition to potentially changing the earth's climate system, rising atmospheric CO_2 concentrations directly impact the photosynthetic process and may increase the primary production of photosynthetic organisms in some systems. Satellite imagery and remote-sensing information for the past two decades indicate as much as a 6% increase in total production of terrestrial systems, primarily in the tropics. Smaller scale manipulative experiments, such as free air CO_2 enrichment (FACE) studies, have yielded more complex results. Most FACE studies have observed an initial increase in productivity but then a stabilization in productivity over time. The mechanisms behind these observations are not clear, but it seems likely that some other limitation besides CO_2 is eventually expressed in most terrestrial systems. The processes and feedbacks of carbon cycling, of which photosynthesis is a major component, are very active areas of research.

8. CONCLUSION

The photosynthetic process is literally the most important energy transfer on Earth. The fixation of carbon by photosynthetic organisms provides the high-energy molecules needed to sustain nearly all organisms on Earth. Photosynthesis is a two-stage process: (i) harnessing of light energy to produce reducing power (NADPH) and energy (ATP) and (ii) use of the reducing energy to fix organic molecules. We usually associate photosynthesis with the reduction of atmospheric CO_2 to organic sugars. However, the energy generated during the light reactions of photosynthesis can be used to reduce other compounds (e.g., nitrogen and sulfur). Photosynthetic reactions have, in part, controlled the relative carbon pool sizes on Earth. Human activities, especially fossil fuel consumption and biomass burning, have led to an injection of carbon as CO_2 into the atmosphere. This rapid increase in carbon in the atmosphere has the potential to dramatically alter future Earth systems function and climate. A complete mechanistic understanding of global carbon cycling including photosynthetic processes is imperative.

SEE ALSO THE FOLLOWING ARTICLES

Carbon Sequestration, Terrestrial • *Conversion of Energy: People and Animals* • *Earth's Energy Balance* • *Food Capture, Energy Costs of* • *Heterotrophic Energy Flows* • *Origin of Life and Energy* • *Photosynthesis, Artificial* • *Sun, Energy from* • *Thermoregulation*

Further Reading

Davis, G. R. (1990). Energy for planet Earth. *Sci. Am.* **263,** 20–27.

Hall, D. O., and Rao, K. K. (1999). "Photosynthesis." Cambridge Univ. Press, Cambridge, UK.

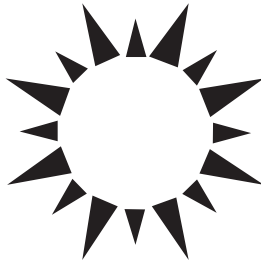
Laisk, A., and Oja, V. (1998). "Dynamics of Leaf Photosynthesis: Rapid-Response Measurements and Their Interpretations." CSIRO, Collingwood, Australia.

Lawlor, D. W. (2001). "Photosynthesis." Springer-Verlag, New York.

Wigley, T. M. L., and Schimel D. S. (2000). "The Carbon Cycle." Cambridge Univ. Press, Cambridge, UK.

A0005 # *Photosynthesis, Artificial*

NATHAN S. LEWIS
California Institute of Technology
Pasadena, California, United States



1. Types of Artificial Photosynthetic Devices
2. Basic Processes of a Photoelectrochemical Energy Conversion Device
3. Photoelectrolysis Using Metal Oxides and Other Semiconductors
4. Approaches to Constructing Stable, Efficient Photoelectrochemical Cells

Glossary

G0005 **anode** The electrode in an electrochemical cell that supplies electrons into the external circuit. This electrode results in removal of electrons from the electrolyte at the electrode/electrolyte interface.

G0010 **cathode** The electrode in an electrochemical cell that injects electrons into the electrolyte.

G0015 **electrochemical cell** A complete electrical circuit, consisting of two electrodes, the anode and the cathode, an electrolyte in contact with the electrodes, and a redox species that can carry electrical charge in the electrolyte solution.

G0020 **electrolyte** A liquid containing an ionic salt that allows electrical conduction in the liquid.

G0025 **photoelectrode** A material that absorbs incident light and produces current through an external circuit.

G0030 **photoelectrolysis of water** A specific photoelectrochemical cell in which water is oxidized, forming oxygen gas, at the anode, and water is reduced, forming hydrogen gas, at the cathode.

G0035 **redox couple** A pair of chemical species that is involved with donating and/or accepting electrons from electrodes in an electrochemical cell. The redox couple consists of a chemically reduced species that can donate electrons to an electrode, thereby forming the chemically oxidized form of the redox couple. The oxidized species of the redox couple can accept electrons from an electrode, forming the chemically reduced form of the redox couple.

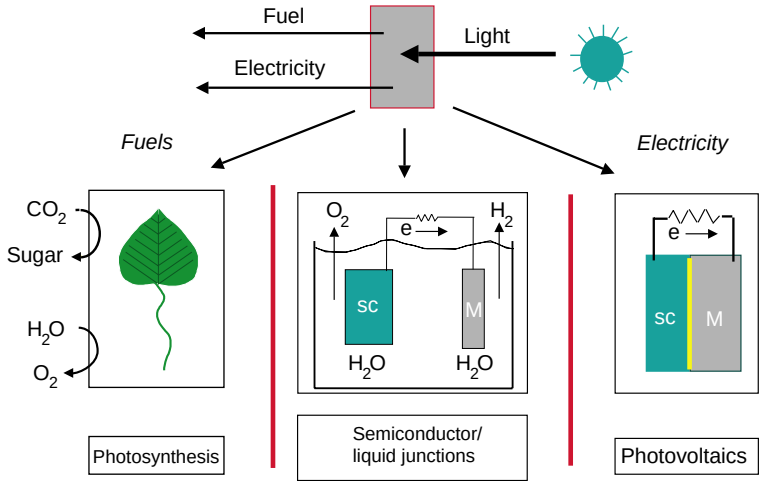
P0005 The global primary energy consumption from all sources in 1998 was approximately 400×10^{18} J, which equates to a mean energy consumption rate of 12.8×10^{12} W or 12.8 terawatts (TW). By comparison, the net solar power striking the surface of the earth is approximately 1.2×10^{5} TW. A diurnally averaged 20 TW of power could be obtained if approximately 0.16% of the land on Earth were devoted to energy farms that converted solar energy at 10% net conversion efficiency. Biomass farms generally display net solar energy conversion efficiencies of 0.1–0.5%. In contrast, the theoretical maximum energy conversion efficiency for a simple light-absorbing solar energy conversion device is more than 30%, and photovoltaic cells that produce electricity from sunlight have demonstrated efficiencies in the field in excess of 20%. Numerous laboratories throughout the world are therefore building artificial photosynthetic systems with the hope of understanding, and ultimately improving on, the energy conversion properties of the natural photosynthetic process.

1. TYPES OF ARTIFICIAL PHOTOSYNTHETIC DEVICES S0005

P0010 Several molecular systems have been constructed that mimic various aspects of photosynthesis. Two of these utilize molecular systems that are derived from natural photosynthesis but that incorporate chemically based modifications to produce artificial photosynthetic devices. These devices use artificial photosynthetic pigments to drive chemical reactions across lipid bilayers or use noble metal catalysts to change the function of the photosynthetic process to produce hydrogen and oxygen instead of sugars and oxygen. Neither of these systems are sufficiently robust to be operated for extended periods of time as energy conversion devices, but they have shown that it is possible to produce artificial photosynthetic assemblies that function well in a laboratory setting.

P0015 The third system, semiconductor photoelectrochemistry, utilizes a nonbiological approach to produce fuels in an entirely artificial photosynthetic system. Photoelectrochemical systems use sunlight to supply part or all of the energy required to generate the storable chemical fuels (Fig. 1). Photoelectrochemical

 © 2004 Elsevier Inc. All rights reserved.



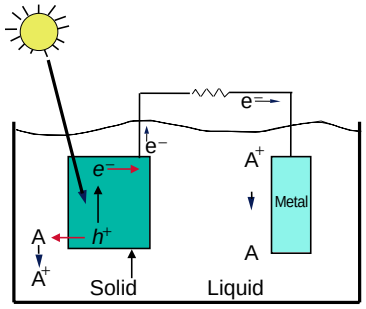
F005 ***FIGURE 1*** Known methods for transforming solar energy into chemical fuels or electrical energy. Photosynthesis stores approximately 3% of the incident solar energy in the form of chemical fuels for use by the plant. Photovoltaics produce electrical energy with up to 25% efficiency. Wet solar cells based on semiconductor/liquid junctions produce either storable chemical fuels or electrical energy, depending on their mode of operation. Efficiencies of these artificial photosynthetic systems can be greater than 15%.

systems can also be utilized to provide electrical energy. Photoelectrochemical energy conversion systems, although a relatively new technology, have demonstrated laboratory solar energy conversion efficiencies of $>16\%$ for electricity formation and $>10\%$ for fuel formation during the sunlight-assisted splitting of water into $H_2(g)$ and $O_2(g)$. In this respect, they represent a promising method for obtaining renewable, storable energy from the sun.

P0020 A photoelectrochemical cell is closely related to a battery, fuel cell, or other type of electrochemical system in that it is composed of two electrodes and an electrolyte. Unlike batteries or fuel cells, in which the energy is initially stored in the reactants and is then released through the electrical discharge circuit during the formation of chemical reaction products, photoelectrochemical cells utilize the input optical energy to drive electrochemical reactions (Fig. 1). To construct a photoelectrochemical cell, one of the metal electrodes of a conventional electrochemical cell is replaced by a semiconductor electrode. Light absorbed by the semiconductor electrode produces a current through the cell, which results in the production of chemical fuels and/or electrical energy by the device.

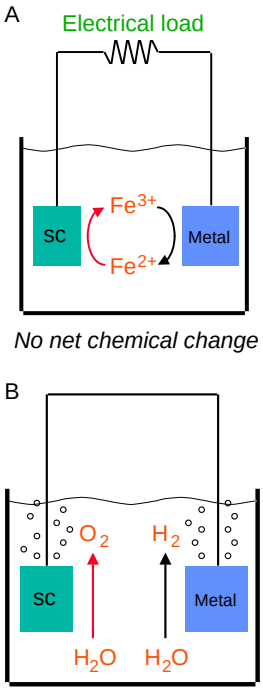
S0010 2. *BASIC PROCESSES OF A PHOTOELECTROCHEMICAL ENERGY CONVERSION DEVICE*

P0025 The basic processes that occur in such a system are well understood. The semiconductor electrode efficiently absorbs light, producing an excited electronic state. In this excited state, the electron and the electron vacancy (the "hole") are both more energetic than they were in their respective ground states. The photoexcited electrons and holes are generally not tightly bound to an individual atom or set of atoms in the solid. These delocalized charge carriers readily move through the semiconductor and are therefore available to perform chemical reactions at the solid/liquid interface (Fig. 2). When the photogenerated electrons and holes drive chemical reactions at the electrodes in the cell, energy from sunlight is stored in the system.



F0010 ***FIGURE 2*** Basic operation of a photoelectrochemical cell. The semiconducting electrode absorbs energy from sunlight and produces a current through the electrochemical cell. The current produces electrical and/or chemical energy in the system.

P0030 If the chemical reaction at one electrode is exactly the opposite of the reaction at the other electrode, no net chemical change will occur in the electrolyte. The overall process is then conversion of light into



F0015 ***FIGURE 3*** Two modes of operation for wet solar cells based on photoactive semiconductor materials. In both possibilities shown, the semiconductor (sc) absorbs the light and produces energetic electrons that can flow through the circuit. The differences between the cells result from the types of chemical reactions that occur in the liquid as a result of this electron flow. (A) A wet solar cell that produces electrical power from incident sunlight. The semiconductor produces excited electrons that flow through the circuit, doing work on an electrical load. This load could be a light bulb, battery, or other type of electrical machinery. The electrons are then injected into the solution at the metal electrode and produce a chemical reaction in the liquid. In this example, the injected electrons convert one form of iron (Fe^{3+}, with three ionized electrons) to another (Fe^{2+}, with only two ionized electrons). The Fe^{2+} then moves to the semiconductor electrode, carrying with it the injected electron. To complete the circuit, the Fe^{2+} then donates this electron to the semiconductor electrode. This regenerates the Fe^{3+} state in the solution. Thus, no net chemical change occurs in the liquid, but the light has produced a current through the electrical device in the circuit. (B) A wet solar cell that produces both electrical power and chemical fuels from incident sunlight. The processes of light harvesting and current flow through the circuit are very similar to those in A, but the chemical reactions are different at the two electrodes. In this example, the injected electrons react with the water to produce gaseous hydrogen (H_2), which bubbles out of the solution. To complete the electrical circuit, the semiconductor electrode must be provided with electrons. Here, they come from the water. Removal of electrons from the water produces oxygen gas (O_2), which also bubbles out of the solution. Thus, illumination of this cell with light produces both current flow and gaseous chemical fuels. The gases could be collected and stored for later use, with the stored energy being recovered either in a fuel cell or by burning the hydrogen and oxygen in an engine to regenerate the water.

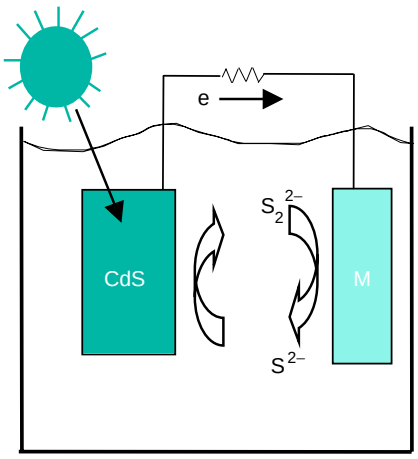
electricity. This mode of photoelectrochemical cell operation resembles that of a photovoltaic device and is designated as a regenerative photoelectrochemical cell. Alternatively, if the chemical reactions are different at each electrode, chemical fuels will be produced by the light-induced current flow. This type of system is designated as a fuel-forming photoelectrochemical cell. Both regenerative and fuel-forming photoelectrochemical cells have been demonstrated in the laboratory. A particularly appealing situation is depicted in Fig. 3, in which water is the liquid and current flow through the circuit simultaneously produces electricity, hydrogen gas, and oxygen gas.

P0035 Closer consideration of a photoelectrochemical cell makes it clear that efficient light absorption alone is not sufficient to obtain efficient device operation. If the charged carriers created by optical excitation of the semiconductor had no preference regarding their direction of motion, equal numbers of excited electrons would travel in clockwise and counterclockwise directions through the circuit in the cell. The equal magnitudes, but opposite signs, of the currents would lead to a net cancellation of current in the wire, and no net energy output would be obtained from such a cell.

P0040 Charge separation, however, is very efficiently effected in a well-designed photoelectrochemical cell. The key to directing the charge carriers preferentially through the circuit lies in the formation of a strong electric field at the semiconductor/liquid interface (Fig. 4). This electric field either attracts or repels electrons from the interface, depending on the polarity of the field. Fortunately, maintaining the field does not require a constant energy input from an external source: This electric field occurs spontaneously whenever two phases with different chemical properties are brought into contact. In semiconductor/liquid junctions, the interfacial fields are generally so strong that any photogenerated electrons and holes are rapidly driven in opposite directions, and charge carrier recombination is thereby suppressed.

S0015 *3. PHOTOELECTROLYSIS USING METAL OXIDES AND OTHER SEMICONDUCTORS*

P0045 Certain materials, generally metal oxides, can sustain the unassisted photoelectrolysis of water into H_2 and O_2 (see Fig. 3). The energy contained in the photons that are absorbed by these materials can in many cases be efficiently converted into energy stored into



F0020 ***FIGURE 4*** Schematic of a wet semiconductor/liquid-based solar cell with a stabilizing agent in the solution. The light is absorbed by the semiconductor and produces excited electrons. Although CdS is shown, the semiconducting material could be silicon, gallium arsenide, titanium dioxide, or other suitable samples. The excited electrons are directed into the wire by the electric field because the negatively charged electrons are repelled by the negative charges that reside on the liquid side of the semiconductor/liquid boundary. The electrons then flow clockwise through the circuit, performing work through a load. They then reach the metal electrode and are forced to undergo a chemical reaction with the liquid in order to continue the circuit. The chemical produced from this reaction then moves through the liquid and reacts with the semiconductor electrode. This restores the electron to its original position and completes the current loop. The stabilizer is needed to accept and donate electrons at the two electrode/electrolyte interfaces to prevent corrosion or passivation of the two active electrodes in the cell.

the chemical fuels produced by the photoelectrolysis system. However, these materials have optical absorption thresholds (band gaps) that are too large to permit effective absorption of the visible and infrared portions of the solar spectrum, thereby wasting a large fraction of the incident solar energy and yielding overall energy conversion efficiencies of $<1\%$. Modification of metal oxide materials has been claimed to be much more promising, with a modified TiO_2 photoanode yielding 8% efficiency in the photoassisted splitting of water into H_2 and O_2.

P0050 Metal chalcogenides have been investigated as photoanodes because the lattice S, Se, or Te atoms are weaker oxidants than lattice oxygen atoms and thus would enable more efficient utilization of the incident solar energy. Photoelectrochemical studies of metal chalcogenide electrodes, however, have revealed materials that are better absorbers than TiO_2 and therefore can provide more desirable semiconductor electrodes in photoelectrolysis systems. However, these materials are unstable in contact with water. There is no fundamental thermodynamic reason that this should be the case; rather, it is a limitation of the known materials systems in photoelectrochemistry.

4. APPROACHES TO CONSTRUCTING STABLE, EFFICIENT PHOTOELECTROCHEMICAL CELLS

S0020

4.1 Regenerative Cells in Aqueous Solution

S0025

P0055 Several approaches are being pursued to circumvent the problems described previously with the direct, efficient photoelectrochemical splitting of water. An alternative approach to minimize aqueous-based corrosion of semiconductor electrodes has been to search for water-soluble molecules that exchange electrons with the photoelectrode much more readily than does water. Such molecules, if present in the liquid in sufficient concentration, could prevent electrode corrosion by reacting with the photogenerated holes in the electrode before the holes can participate in the corrosion process. If the oxidized form of these corrosion-inhibiting reagents also provides a useful fuel, the system is even more desirable.

P0060 This approach has been successfully implemented for the II–IV semiconductors, such as CdS, CdSe, and CdTe. These materials can be stabilized in aqueous solutions through addition of sulfide to the solution to provide another source of reduced sulfide to compete with electrode corrosion (Fig. 4). In this manner, a stable regenerative photoelectrochemical cell can be constructed from an otherwise unstable semiconductor electrode. Long-term stability for several years has been demonstrated for such systems. The use of the stabilizing agent eliminates the possibility of production of desirable chemical fuels from the photoelectrochemical process but in return provides a stable photoelectrochemical cell for the conversion of light into electrical power. For example, stable, 16% efficient cells have been obtained using gallium arsenide in contact with an aqueous medium containing high concentrations of selenide ions.

P0065 $CuInSe_2$ photoanodes have yielded $>10\%$ efficiency for the production of I_2 from aqueous iodide ions. The production of I_2 is more desirable than the formation of Se (as in the GaAs system) because the I_2 could provide a useful fuel to charge an *in situ* storage battery. If properly coupled with a metal electrode, this $CuInSe_2$ cell could produce H_2 and I_2

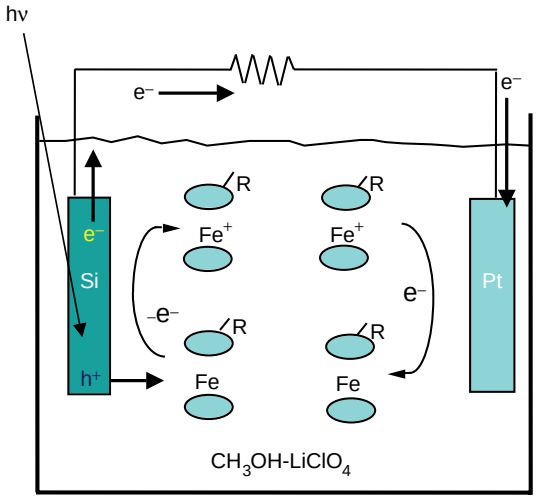
from aqueous HI. The stored chemical energy in the H_2 and I_2 would then be recovered at a later time, much as would be done if H_2 and O_2 gases had been obtained from a water-splitting process. Similar results for I_2 production have been obtained for $MoSe_2$ electrodes and for *in situ* electrical storage using the alloy material $CdSe_{0.65}Te_{0.35}$. Si electrodes have been covered with small metal islands to produce Br_2 from aqueous HBr with high efficiency, and $>10\%$ solar energy conversion efficiencies have been demonstrated for the formation of $H_2(g)$ from water at p-type InP electrodes. The key to making these fuel-forming systems more practical is the efficient coupling to the other, nonsemiconducting electrode: If this can be done without significant additional electrical energy input, fuel production could be sustained in a very efficient manner.

S0030 ## 4.2 Regenerative Cells in Nonaqueous Solvents

P0070 For some materials, such as CdS, several stabilizer/electrolyte combinations have been identified that are stable on the months/years timescale, whereas for others, such as Si photoanodes, no reagent has been discovered that provides stable operation even on the timescale of minutes in aqueous solution. A general approach to circumvent the instability of electrode materials such as Si and GaAs in aqueous solutions is simply to eliminate the corrosive liquid, water, from the environment of the photoelectrochemical cell. If the water in the cell is replaced by alcohol or another organic liquid, stable photoelectrode operation can readily be obtained, provided that the organic solvent also does not effect corrosion of the solid. Using this approach, stable photoelectrochemical cells have been obtained with Si, GaAs, InP, and other semiconductors whose optical properties are well matched to the solar spectrum.

P0075 In a nonaqueous photoelectrochemical cell, photogenerated charge carriers in the semiconductor are not scavenged by reaction with the solvent but are transferred to a chemical reagent that has been deliberately added to the electrolyte solution. This chemical stabilizing agent accepts an electron (is reduced) at one electrode and donates that electron (becomes oxidized) at the other electrode. The stabilizing agent carries the electrical charge between the plates of the electrochemical cell but undergoes no net overall chemical change during extended cell operation. Stabilizers should ideally be species that rapidly exchange electrons with both the semiconductor and the metal electrodes in the cell. Rapid reaction with the semiconductor is required to compete effectively with corrosion reactions that will arise from the inevitable presence of trace quantities of water in the cell, whereas rapid electrochemical reaction with the metal electrode will minimize resistive losses from the charge transfer step. Nonaqueous semiconductor/liquid junctions are almost universally restricted to operating as regenerative photoelectrochemical cells because the reduced and/or oxidized forms of the stabilizers are rarely useful fuels. The stabilizing reagents are therefore most effectively utilized when they are added to the cell only once to facilitate the sustained stable production of electrical energy from an otherwise unstable semiconductor electrode.

P0080 An advantage of nonaqueous photoelectrochemical cells is that they allow utilization of many semiconductors that are not stable in any known aqueous electrolyte. An example is provided by the properties of Si photoanodes, which are readily stabilized for months when the solvent is methanol or ethanol in the presence of organometallic stabilizing agents (Fig. 5). Another advantage of using nonaqueous solvents is the ability to use a variety of stabilizing agents and to vary their chemical properties widely while still obtaining stable



F0025 ***FIGURE 5*** Stabilization of semiconductor electrodes in nonaqueous solvents. Although unstable in aqueous solutions, many semiconductors can be stabilized in nonaqueous electrolytes. In this example, the Si electrode is in contact with a nonaqueous liquid that contains a high-boiling solvent (propylene carbonate), salt to provide electrical conductivity (lithium perchlorate), and redox ions to carry the electrical charges through the liquid (ferrocene, which is an organometallic complex of Fe^{2+} that is soluble in organic solvents). The cell is 14% efficient in converting incident solar energy into electrical energy.

electrical currents from the photoelectrochemical device. This can be extremely useful in optimizing the energy conversion efficiency parameters of a photoelectrochemical system. For example, in the water splitting systems discussed previously, significant energy is wasted when there is a large energy difference between the levels of electrons or holes in the photoelectrode and the oxidation/reduction potentials for the water splitting reaction. Such energy differences can be minimized, for example, using Si/alcohol contacts because the energy level of the stabilizer can be changed until it closely matches the desired energy level of the semiconductor. This allows optimization of the cell efficiency through simultaneous chemical design of the electrolyte, photoelectrode material, and stabilizing agent.

P0085 These nonaqueous photoelectrochemical devices can yield relatively high solar energy conversion efficiencies: 12–14% efficiency has been obtained in unoptimized photoelectrochemical cells compared to the 15–20% efficiency that is obtained for production-line solid-state Si and GaAs photovoltaics. The nonaqueous photoelectrochemical cells also offer an opportunity to readily change the chemistry at the solid/liquid interface and to investigate the effect of these changes on the electron yield of the system. Although much remains to be learned about the principles that govern the behavior of these cells, many of them can properly be denoted as technical successes because they are efficient and stable in sunlight.

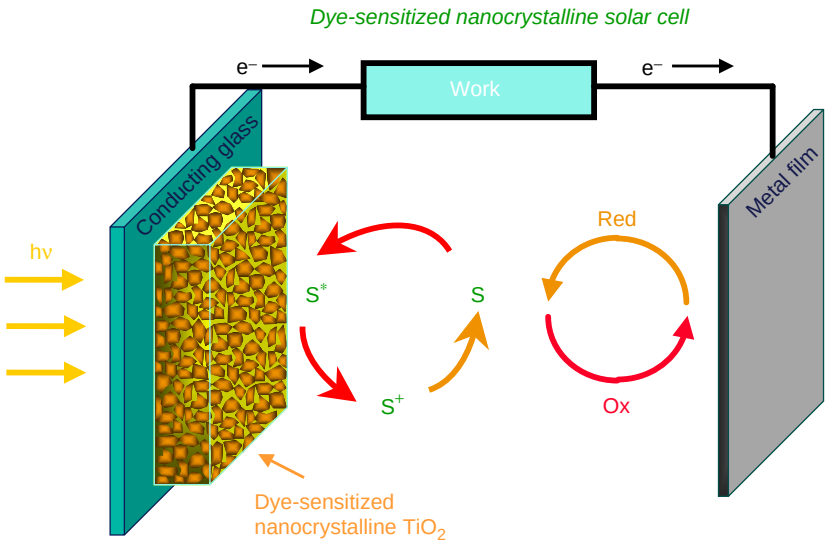
S0035 ## 4.3 Dye Sensitization of Stable Electrode Materials

P0090 Another approach to artificial photosynthesis using semiconductor photoelectrochemistry is to extend the optical response of photoelectrode materials that are known to be stable in water, such as TiO_2. When a highly colored dye is coated on the surface of a semiconductor, the absorption of a white (i.e., high band gap) material such as TiO_2 can be extended into the visible region of the spectrum. This type of dye sensitization is analogous to the chemistry that forms the basis of the modern color photographic process, with the solid being semiconducting AgBr and with various dyes providing the color response of the AgBr film emulsion.

P0095 Most dye-coated, single-crystal semiconductor electrodes exhibit solar energy conversion efficiencies of less than 0.1%. Even with the most highly colored molecules (i.e., dye molecules having very large optical absorption coefficients), several layers of dye are required to absorb all the incident light. However, only the dye layer closest to the electrode surface is generally effective in injecting an excited electron to the electrode. Since only a monolayer of dye is effectively useful for light absorption, most of the photons incident onto the cell are either transmitted through the device or absorbed by dye molecules that do not contribute to photocurrent flow.

P0100 To deal with this issue, very rough, high-surface-area, dye-coated titanium dioxide electrodes are used. In this system, the TiO_2 consists of crystallites with a diameter of approximately 10 nm. A sintered film of these nanocrystals, although relatively transparent, possesses such a high surface area that only a monolayer of adsorbed dye molecules can absorb $>90\%$ of the photons incident onto the sample. The excited state of the adsorbed dye is sufficiently reducing that it will donate an electron to the conduction band of the TiO_2. The presence of the TiO_2, or similar semiconducting material, is crucial because the TiO_2 accepts the photogenerated electrons from the excited state of the dye and directs the injected electron toward the electrical contact, preventing recombination with the oxidized form of the dye or with the species in the electrolyte solution. The dyes investigated to date, including Ru(polypyridyl) complexes and chlorophylls, have extended the absorption range of TiO_2 into the visible region of the spectrum and have provided much better overlap of the electrode response with that of the solar spectrum. Furthermore, the excited states of these adsorbed dye molecules have yielded very effective charge separation, producing photocurrent in the TiO_2 with near unity quantum yield. The efficiency of the system is thus not limited by the charge collection and electron injection process but by the voltage that can be obtained from the specific system of concern.

P0105 This system resembles the molecular photosynthetic process in that the photochemical reactions in the cell are completely dictated by the reactivity of the adsorbed dye molecules. The solid induces the charge separation and dictates the directionality of current flow, but the adsorbed dye molecules determine the course of the remaining steps in the energy conversion cycle. After injection of an electron into the solid, the oxidized dye molecule receives an electron from a soluble electron donor, regenerating the original form of the dye. The electrical circuit is then completed by injection of the electron into the solution at the other electrode in the system (Fig. 6).



F0030 ***FIGURE 6*** A wet solar cell made from a dye-sensitized semiconductor. A colored dye is used to allow normally transparent semiconductors to absorb sunlight in the visible region of the electromagnetic spectrum. The high surface area of the mesoporous solid allows nearly complete absorption of the incident light by a monolayer of dye adsorbed onto the surface of the semiconducting solid. The excited electrons produced from absorption by the dye are injected into the semiconductor and then make their way through the circuit as in a conventional wet cell. The dye is a ruthenium-based metal complex, and the support is a high-surface-area TiO_2 material, related to the inexpensive pigment commonly used in white paint. Solar energy conversion efficiencies as high as 10% have been reported in test devices of this type of potentially inexpensive solid/liquid junction.

P0110 Some of the Ru(polypyridyl) dyes investigated have oxidation potentials that are very close to those required to form $O_2(g)$ from water. Thus, the problems encountered in the photoelectrochemistry of metal oxide electrodes, involving severe energy loss between the level of the photogenerated holes and potential for the oxidation of water to $O_2(g)$, can be avoided, in principle, in the dye-sensitized system. The electronic energy levels of the adsorbed dye molecules are simply tuned chemically until they are close to that of the desired chemical reaction in the cell. Unfortunately, oxidation of water to oxygen gas is a multielectron process and is generally very difficult to accomplish at a rapid rate. To quickly regenerate the original dye molecule for reuse in another photochemical cycle, another molecule, called a supersensitizer, must be added to the dye-sensitized photoelectrochemical cell. The supersensitizer donates an electron to the oxidized form of the dye, regenerating the original form of the dye and producing the oxidized supersensitizer. The oxidized supersensitizer is then reduced at the second electrode, yielding a regenerative photoelectrochemical cell for the entire photoelectrochemical cycle.

P0115 The TiO_2 dye-sensitized photoelectrochemical cell has been modified and improved by using nonaqueous solvents. Nonaqueous solvents allow greater stability of the system and also a wider choice of supersensitizers. This system is essentially an extension of the strategy for producing stable photoelectrochemical cells involving the use of redox stabilizers in nonaqueous electrolytes, but in which relatively expensive single-crystal semiconductors have been replaced by less expensive, dye-sensitized nanocrystalline photoanode materials. Using this approach, solar energy conversion efficiencies of up to 10% have been reported. Relative to conventional photovoltaic cells, it has been claimed that this type of photoelectrochemical system has cost advantages, including the use of inexpensive raw materials, the lack of difficult fabrication steps or expensive crystal purification processes, and the ability to implement the design simply over large areas. The system indeed appears promising for electrical energy generation, provided that efficiencies can be improved and that light-induced degradation of the dye and the other constituents in the cell can be kept to levels required for long-term, maintenance-free operation of the system.

P0120 As can be seen, there are many possibilities for improving these photoelectrochemical cells, and no one strategy has emerged as the clear choice. We know a great deal about the fundamental operating

principles of each component of these artificial photosynthetic systems. Advances have also been made in practical efforts to improve their energy conversion efficiencies and to increase their production rates of chemically desirable fuels. In some cases, the systems show promise in the laboratory as highly efficient energy conversion devices. Even for these systems, choices must still be made to obtain the correct combination of materials that might comprise a complete artificial photosynthetic system. Siting considerations indicate that at 10% overall conversion efficiency, a relatively small fraction of the suitable land area available globally would be required to produce significant amounts of power relative to current global fossil energy consumption. Artificial photosynthetic systems, however, are in their infancy, and the resultant energy costs that might be realized from a fully developed artificial photosynthetic technology are difficult to estimate. Possible advantages of the approach relative to photovoltaics for production of electricity are the ability to directly form fuels from an inherently intermittent energy source as well as the ability to use inexpensive raw materials as the light absorbers without suffering large decreases in energy conversion efficiency of the resulting photoelectrochemical energy conversion devices. The environmental impacts of these systems would be relatively benign provided that inert raw materials, such as Si and TiO_2, were used as the basic photoelectrode materials. The actual decision to construct a commercial system will also require input regarding a number of market-related factors, including the price of alternative energy sources, the value of the fuels produced by the photoelectrochemical cells, and the cost in both dollars and energy of fabrication and maintenance of the system. There are similar constraints for all alternative energy sources, but the promise and demonstrated efficiency of these devices is too great to rule out these interesting artificial photosynthetic devices as energy options for the future.

SEE ALSO THE FOLLOWING ARTICLES

Photosynthesis and Autotrophic Energy Flows • Photovoltaic Materials, Physics of • Solar Fuels and Materials

Further Reading

Bard, A. J., and Fox, M. A. (1995). Artificial photosynthesis: Solar splitting of water to hydrogen and oxygen. *Accounts Chem. Res.* **28,** 141–145.

Finklea, H. O. (1988). "Semiconductor Electrodes." Elsevier, New York.

Fujishima, A., and Honda, K. (1972). Electrochemical photolysis of water at a semiconductor electrode. *Nature* **238,** 37–38.

Gerischer, H. (1970). Semiconductor photoelectrochemistry. *In* "Physical Chemistry: An Advanced Treatise" (H. Eyring, D. Henderson, and W. Yost, Eds.), pp. 463–542. Academic Press, New York.

Hodes, G., Fonash, S. J., Heller, A., and Miller, B. (1985). Photoelectrochemical cells based on polycrystalline semiconductors. *Adv. Electrochemistry Electrochem. Eng.* **13,** 113–158.

Koval, C. A., and Howard, J. N. (1992). Electron transfer at semiconductor electrode–liquid electrolyte interfaces. *Chem. Rev.* **92,** 411–433.

Nozik, A. J. (1978). Photoelectrochemistry: applications to solar energy conversion. *Annu. Rev. Phys. Chem.* **29,** 189–222.

O'Regan, B., and Grätzel, M. (1991). A low-cost, high-efficiency solar cell based on dye-sensitized colloidal TiO_2 films. *Nature* **353,** 737–740.

Parkinson, B. A. (1983). An overview of the progress in photoelectrochemical energy conversion. *J. Chem. Education* **60,** 338–340.

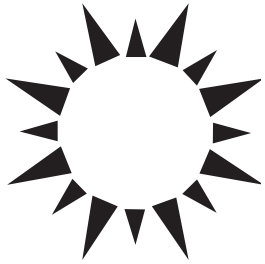
Rajeshwar, K. (1985). Materials aspects of photoelectrochemical energy conversion. *J. Appl. Electrochem.* **15,** 1–22.

Tan, M. X., Laibinis, P. E., Nguyen, S. T., Kesselman, J. M., Stanton, C. E., and Lewis, N. S. (1994). Principles and applications of semiconductor photoelectrochemistry. *Prog. Inorganic Chem.* **41,** 21–144.

Tributsch, H. (1986). Photoelectrolysis and photoelectrochemical catalysis. *In* "Modern Aspects of Electrochemistry" (J. O. M. Bockris and B. E. Conway, Eds.), pp. 303–355. Plenum, New York.

Wrighton, M. S. (1979). Photoelectrochemical conversion of optical energy to electricity and fuels. *Accounts Chem. Res.* **12,** 303–310.

Photovoltaic Conversion: Space Applications



PETER A. ILES
Arcadia
California, United States

1. **Special Features of Space Photovoltaics**
2. **Evolution of Space Solar Cells**
3. **Progress in Photovoltaics**
4. **Future Trends in Space Photovoltaics**
5. **Conclusion**

Glossary

air mass zero (AMO) Location outside Earth's atmosphere.
antireflecting coating Dielectric layers deposited on solar cells to reduce reflection of incident sunlight.
back surface field (BSF) Electric field at the back surface of a solar cell.
back surface reflector (BSR) Mirror formed at the back surface of solar cells to increase reflection of unabsorbed sunlight from the back surface.
beginning of life (BOL) Period immediately after spacecraft is launched.
conversion efficiency Fraction of the power output of a solar cell to the incident power from the sun, expressed as a percentage.
coverglass Transparent cover bonded to the front surface of solar cells.
end of life (EOL) Period after solar cells have withstood defined operating conditions in space, mostly exposure to radiation.
epitaxial layers Layers grown to replicate the crystal lattice properties of a substrate material.
metal organic chemical vapor deposition (MOCVD) Deposition of epitaxial layers using gaseous elements, obtained mainly from compounds consisting of organic chemical groups (e.g., methyl and CH_3) and a metal.
organometallic vapor phase epitaxy (OMVPE) Similar to MOCVD.
PN junction Voltage barrier formed in a semiconductor, providing the conversion of optical (photonic) energy into electrical energy.
P/N solar cell The illuminated surface is a P-type emitter layer formed over an N-type base. N/P solar cells have opposite layer polarity.
substrate The starting material for forming solar cells; also refers to the material to which solar cells are bonded to form panels (or arrays).

Photovoltaic arrays on satellites and spacecraft provide electric power for on-board equipment and communication with control centers on Earth. Before 1957, when the first satellite, Sputnik I, was launched, there was much debate about the best way to generate electric power on satellites. Most conventional fuels do not function well in the vacuum of space, and even if usable, they cannot operate for extended periods without exceeding the weight limits set by the propulsion rockets and posing serious safety concerns. Sunlight is the obvious power source in space, removing the need to launch fuel with the spacecraft. In 1954, Bell Telephone Laboratories (BTL) announced that it had made a significant increase in the output from silicon solar cells. Four years later, commercial versions of BTL cells were successfully used on the second U.S. satellite Vanguard I to provide power for a radio-tracking signal that lasted for several years. Following this successful demonstration, space photovoltaic (PV) power was rapidly accepted and used widely. For more than 40 years, arrays of solar cells have delivered power flawlessly for long periods. PV power is one of the major reasons for the successful performance of thousands of satellite. Satellite applications have become familiar, including worldwide telecommunications, sensing of weather and other Earth surface geographic conditions, and navigation using the constellation of global positioning satellites (GPS). Satellites have also performed many scientific missions, carrying astronomical telescopes and exploring around or on the moon

 © 2004 Elsevier Inc. All rights reserved.

and Mars and diverse locations toward or away from the sun. Military satellites have collected large amounts of data for surveillance or information gathering. Space PV continues to be an active niche market, supplied by organizations throughout the world. The high visibility of space programs has publicized the many advantages of PV-generated power.

1. SPECIAL FEATURES OF SPACE PHOTOVOLTAICS

Space photovoltaic (PV) differs from terrestrial PV in several ways.

1.1 Components of a Space Solar Power System

Space solar power systems comprise the following components:

- *Solar cells*: The most important components are the solar cells, which are connected together in series-parallel circuits to supply the required voltages and currents.
- *Coverglasses*: Thin transparent covers of glass or quartz are adhesively bonded to the cells to reduce their temperature by increasing the reradiation of infrared wavelengths into space. Covers also protect the cell from damage caused by low-energy protons and have an optical coating to reject ultraviolet (UV) wavelengths to prevent UV darkening of the adhesive.
- *Metal interconnectors*: Thin metal foils, 100–150 μm thick, are soldered or welded to electrically connect the contact pads on the cells. Placing all the contact pads on the back surface of the cells facilitates automated bonding. The metal-to-cell bonds must be robust for high reliability.
- *Array substrates*: Interconnected cells are bonded to lightweight substrates to form a panel or array. Aluminum honeycomb and graphite-epoxy composites are used most. Kapton has been used on fold-out arrays to reduce weight.
- *Deployment and tracking mechanisms*: In some cases, cells are attached to the outer surface of the satellite (body mounted), but they are usually mounted on wings or paddles. Wings are folded during launch and deployed by releasing spring clamps, or by folding out or rolling out, with means to lock the array in position. Sun tracking mechanisms keep the solar arrays facing the sun.
- *Power storage*: Chemical batteries store PV-generated energy and stabilize the electric power fed to the satellite circuitry. Many charge–discharge cycles can occur, and to operate in the vacuum, custom-manufactured sealed batteries are used with nickel–cadmium or lithium electrodes.
- *Electric power distribution*: Autonomous power conditioning equipment controls the battery charge conditions and allocates power to the circuits in the satellite.

1.2 Variety of Locations

Locations where spacecraft must operate are classified into four broad categories:

- *Near-Earth*: In the vicinity of the earth, there are five types of satellite orbits:
  1. Low earth orbit (LEO), altitude 150–1000 miles.
  2. Medium earth orbit (MEO), altitude 1000–22,000 miles.
  3. Geosynchronous earth orbit (GEO), altitude approximately 23,000 miles. Worldwide communications requires 3 satellites in GEO, 15–20 satellites in MEO, and 50 or more satellites in LEO.
  4. Orbital transfer missions (e.g., LEO to GEO).
  5. Cislunar to near the moon, 23,000–250,000 miles distant.
- *Near-Sun*: These include missions toward Venus, Mercury, or even closer to the sun.
- *Far-Sun*: Missions beyond the earth's orbit, including those to Mars, the asteroid belt, Jupiter, and Saturn.
- *Landers*: These are bases on the moon or Mars or for landing on comets or asteroids. Manned bases require a large amount of power.

1.3 The Space Environment

Space contains only minute amounts of gases. In this ultrahigh vacuum, heat cannot be removed from the spacecraft by conduction or convection but only by radiation. The power received from the sun and the solar array temperatures at any location depend on the distance from the sun. In the vicinity of the earth, outside the gaseous atmosphere, the incident solar output as a function of wavelength is shown as solar extraterrestrial in Fig. 1. This solar spectrum is called air mass zero (AMO), and the power density is approximately 1365 watts per square meter (Wm^{-2}).

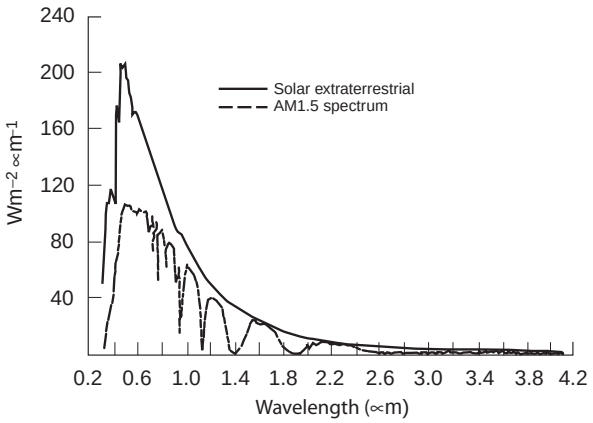


FIGURE 1 Solar spectrum incident on the earth. Solar extraterrestrial, outside gaseous atmosphere; AM1.5 spectrum, at the earth's surface.

Figure 1 also shows the spectrum received at the earth's surface (AM1.5), with the losses in the atmosphere reducing the power density to 800–1000 Wm^{-2}. All solar cell conversion efficiencies given in this article are AMO efficiencies, given by the percentage ratio of the electric power output of the cell, held at a test temperature of 30°C, to the incident power, 1365 Wm^{-2}, from a light source that simulates the AMO spectrum.

1.3.1 Charged Particles

Exposure to charged atomic particles emitted from the sun is not a problem, except during short periods of solar flares or when satellites move through the Van Allen Belts (VABs), in which the particles have been trapped by the earth's magnetic field. A series of satellites measured the energies and concentrations of charged particles in the belts. The outer VAB, at an altitude of approximately 14,000 miles, is more intense than the inner VAB, which has an altitude of approximately 3000 miles. The relative exposures to radiation in LEO, MEO, and GEO are low, high, and moderate, respectively. To estimate the power lost by solar cells from particle bombardment during a mission, the accumulated dosage for the specific mission orbit is calculated and used in ground radiation tests of cells to estimate the end of life output.

1.3.2 Solar Storms

It is recognized that very intense emissions from active regions on the sun can cause serious problems for both terrestrial and satellite power systems. The emissions include X-rays, bubbles of plasma ejected from the sun's corona (coronal mass ejection), and high-energy charged particles, mostly protons and electrons. The plasma generates shock waves that accelerate atomic particles, including those trapped in the VAB, and it also depresses the earth's magnetic field to altitudes at which it affects satellites in GEO orbits. The charged particles cause electrostatic charging problems in solar arrays and in semiconductor circuitry. Failures in satellite attitude controls, especially if they are based on magnetometers, lead to loss of power when the solar arrays are not facing the sun and to reduced communication with the satellite antennas. The large increase in charged particles causes significant power loss from radiation damage in the solar panels.

These solar storm activities have occurred mostly during solar maxima, but problems are observed during quieter periods. Following disruption of several communications satellites, with worldwide impact, additional satellites have been deployed to monitor the sun's activity to provide early warning, allowing some preventive measures to be taken.

1.3.3 Other Orbital Effects

Satellites in low LEOs experience atmospheric drag, causing the satellites to move to lower orbits and eventually fall to the earth. To reduce the amount of booster rocket fuel needed to offset drag effects, the area of the solar arrays must be as small as possible.

For orbits below GEO, satellites pass into the shadow of the earth for part of the orbit. For LEO, shadowing for 30 min during each 90-min orbit causes temperature swings from 60 to −150°C. In 1 year, satellites in LEO can experience 9000 of these temperature cycles. For reliable operation during these cycles, qualified robust procedures must be used to deposit and bond the metallic contacts.

Solar arrays may experience impact damage from many thousands of pieces of space debris left by earlier satellites. The debris must be monitored, shielded, or avoided, especially for manned satellites.

1.4 Remote Operation with High Reliability

The PV power system must be highly reliable because its performance is essential to the success of the satellite mission. PV arrays usually operate unattended, although there have been a few cases of repair or replacement. The solar cells and the rest of the array are manufactured to comprehensive specifications, controlled jointly by the manufacturers and the users. The successful record of space PV arrays is

attributed to this tight process control and detailed documentation.

2. EVOLUTION OF SPACE SOLAR CELLS

Solar cells, which convert sunlight to electrical energy, are the most important component in space PV power systems. High beginning-of-life (BOL) efficiencies are needed because a limited area is available for mounting cells and arrays on satellites. High end-of-life (EOL) efficiencies and very high reliability when operating in space are also required.

2.1 Silicon Space Solar Cells

Silicon (Si) has a band gap (1.15 eV) slightly below the optimum value (1.35 eV) that gives highest PV efficiency. For many years, Si cells had the highest efficiency because of the high-quality, single-crystalline Si developed by the semiconductor industry. The efficiency of the P/N Si space cells derived from the Bell Telephone Laboratory cells was improved from 7 to 10%, and they were used until the early 1960s. After the discovery of the VABs, ground-based particle accelerator tests showed that N/P Si cells had superior resistance to radiation damage. N/P Si cells were quickly adopted and were steadily improved to become the workhorse solar cells used in space.

2.1.1 Choice of Silicon

After extensive testing, simple crystals of Si grown by the Czochralski method (Cz-Si) were found to give cells with the best combined BOL and EOL performance. Two Si resistivity levels were used most, 2 and 10 ohm-cm, containing boron (the P-type impurity), at concentrations of approximately 150 and 30 parts per billion, respectively.

2.1.2 Silicon Cell Improvements

Improved performance was obtained from cells that were manufactured under space-qualified manufacturing methods and that were also compatible with the design and assembly of space solar arrays. One improvement formed shallower PN junctions with reduced concentrations of impurity (phosphorus) in the N-layer. Back surface fields (BSFs) were incorporated to reduce loss of charge carriers by recombination at the back surface. Reflection of incident sunlight was reduced by depositing multilayer antireflection layers, or by texturing the front surface by preferential etching, to form many small random pyramids. Cell temperatures in space were reduced by increasing the reflectivity of the back surface (BSR). Optimized gridline contacts were evaporated on the front surface, and the adhesion of the metal-to-Si contacts was increased.

For some special array designs, or to facilitate automated bonding procedures, pad contacts to the N and P regions were both located on the back surface of the cell. If the back contact was also formed in a grid pattern, the cells were bifacial, generating power when illuminated from the front and back. Bifacial cells can generate additional power in space, converting albedo reflection from clouds or the oceans. Thin Si cells (down to 70 μm) were manufactured. Table I shows production median BOL and EOL efficiencies for the three N/P Si cell designs used most until the 1980s and that are still used when they meet all mission goals. The EOL conditions specified in Table I are also used for EOL efficiencies mentioned in the following sections for other types of cells.

2.1.3 Advanced Silicon Cells

Research centers, principally the University of New South Wales and Stanford University, made significant increases in the efficiency of Si cells (to 19–20% AMO) by using high-quality float-zone refined Si (Fz-Si), by forming regular patterns of pyramids on the front surface, and by "passivating" the front and back surfaces to reduce recombination of charge carriers. Very clean processing conditions were maintained throughout processing to preserve the high quality of the Si. When advanced cells were adapted for space use, improved Cz-Si was used, with slightly lower BOL efficiency (17–18%). EOL efficiency was 12%,

TABLE I

Performance of Most-Used Silicon Space Cells[a]

Cell structure	BOL efficiency (%)[b]	EOL efficiency (%)[c]
2 ohm-cm, BSR	13.5	10.0
10 ohm-cm, BSF, BSR	14.8	9.6
10 ohm-cm, BSR	12.5	10.1

[a] Abbreviations used: BOL, beginning of life; EOL, end of life; BSR, back surface reflector; BSF, back surface field; AMO, air mass zero.

[b] AMO illumination; cell temperature, 30°C.

[c] AMO illumination; cell temperature, 30°C; after exposure to IE15 1 MeV electrons/cm^2.

higher than that of standard Si cells, but the large excess BOL efficiency resulted in problems in the power conditioning equipment. Figure 2 shows the structure of an advanced space Si cell. These cells have been used in applications when radiation exposures were moderate, but they were largely superseded by III–V cells.

2.2 III–V Space Cells

Compounds with elements in group III and group V in the periodic table are semiconductors, like the group IV elements Si and Ge, and have a wide range of properties.

2.2.1 Gallium Arsenide Solar Cells

Gallium arsenide (GaAs) has a band gap of 1.42 eV, close to the value giving peak solar cell efficiency. High-efficiency GaAs cells had been demonstrated, but the space cell community made significant improvements in forming large-area, high-efficiency GaAs cells. The important advance was the ability to grow the GaAs layers by organometallic vapor phase epitaxy (OMVPE) in large production-scale reactors that accommodated an area of 1500 cm^2 of GaAs substrates per run and gave very uniform layers over large areas. In a further advance, the GaAs substrates were replaced by Ge substrates, which have atomic lattice spacing very similar to that of GaAs, leading to growth of high-quality epitaxial GaAs layers. Ge substrates cost less and are stronger mechanically, allowing fabrication of GaAs/Ge large-area, thin GaAs cells. GaAs/Ge cells with an area of more than 36 cm^2 with median BOL efficiencies of 18 to 19% were developed, with superior radiation resistance (EOL efficiency, 14%) and less fall-off in output when operated at elevated temperatures. The improved performance resulted in significant reduction of area and weight of the solar array, making GaAs/Ge cells cost-effective for the PV system, even though they may cost 5–10 times as much as Si cells. These system advantages resulted in rapid acceptance, especially for commercial communications satellites and defense satellites.

2.2.2 Multijunction Solar Cells

Tandem cells made by stacking cells with appropriate band gaps have higher efficiency. After groups at the National Renewable Energy Laboratory successfully grew gallium-indium-phosphide ($GaInP_2$) cells over GaAs cells, on the same GaAs substrate, these $GaInP_2$/GaAs cells were adapted by the space cell community using Ge substrates and production-level OMVPE growth. These dual-junction (DJ) cells eventually attained BOL efficiencies as high as 25.5% and EOL efficiencies of approximately 20%. Additional research led to formation of a PN junction in the Ge substrate during growth of the DJ cell. The resulting triple-junction (TJ) cells have median efficiencies of approximately 27.5% and have been modified to have superior radiation resistance, with EOL efficiencies of approximately

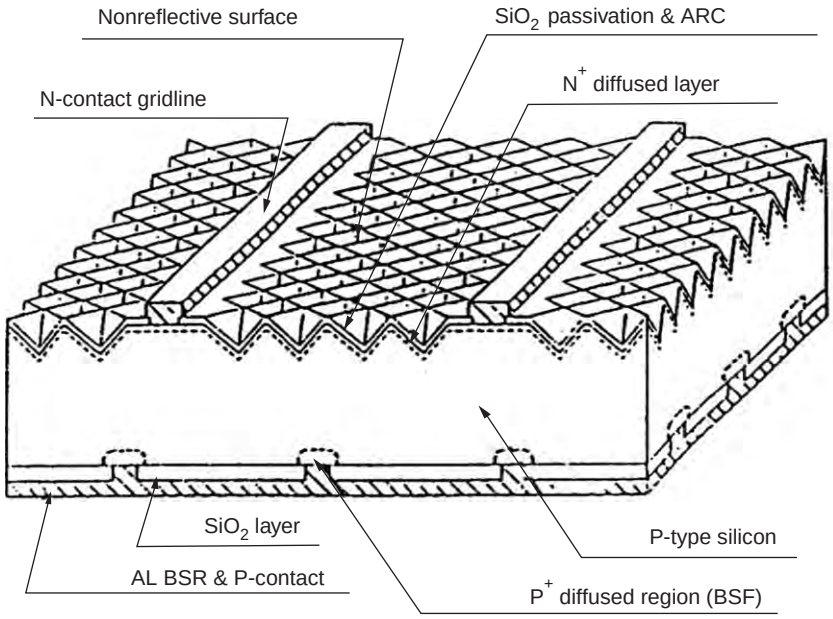


FIGURE 2 Structure of advanced space silicon cell.

23%. Figure 3 shows the cross section of a TJ cell. In addition to the three PN junctions, there were many other layers for passivating interfaces and surfaces and for effectively conducting current through the cell. The growth and properties of all these layers were systematically optimized and controlled by computerized growth programs. Consistency of performance was checked by advanced characterization procedures after each reactor run. The costs resulting from the higher complexity were slightly higher than for GaAs/Ge cells but were offset by the superior performance.

Testing showed that when some multijunction cells in a series string were shadowed, they often degraded. To prevent this shading loss, each MJ cell required its own bypass diode. At first, discrete diodes were used, but now each TJ cell includes a bypass diode formed during OMVPE growth of the cell.

Development of four-junction cells promises efficiencies of 33–35%, but to date efficient four-junction cells have had severe material problems.

2.2.3 Indium Phosphide-Based Cells

Cells containing indium and phosphorus (InP and $GaInP_2$) have high radiation resistance and show recovery of output after radiation. These cells, especially InP grown on Si substrates, are candidates for orbits with very high radiation levels, including MEO or near to Jupiter.

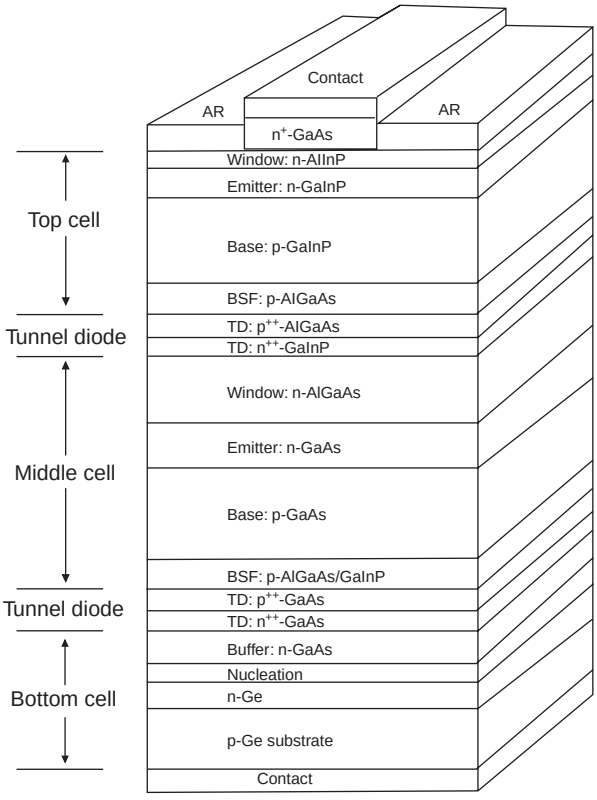


FIGURE 3 Cross section of triple-junction solar cell.

2.3 Thin Film Cells

Thin film (TF) cells, such as amorphous Si or copper-indium-diselenide ($CuInSe_2$), have advantages of low mass, easy deployment, and low cost. However, TF cells with efficiencies of 8–10% AMO have required the use of heavy substrates, which offsets their advantages. TF cells may find use for Mars bases or orbital transfer.

2.4 Solar Cells with Concentrators

With slight redesign, mostly of the front contact gridlines, all solar cells have higher efficiency when incident sunlight is concentrated by lenses or mirrors. Near-Earth operation is limited to low concentrations (five suns) because of the difficulty in removing excess heat from the arrays. However, for far-Sun missions, lightweight concentrators have been used successfully to increase the low intensity and low temperatures leading to higher solar array output.

2.5 Adapting to a Variety of Missions

All solar cells can be modified to perform a wide range of different space missions, but the TJ cells have proved to be most effective. Triple junction cells are versatile, being effective for far-Sun journeys and on the Martian surface. They also perform well in near-Sun operation. Closer to the sun, array temperatures are reduced by tilting arrays away from direct sunlight. At temperatures of approximately 120°C, Ge cells become inoperative and TJ cells become DJ cells, still with high efficiency and sufficient output at the increased intensities.

3. PROGRESS IN PHOTOVOLTAICS

3.1 Power Needs

Photovoltaic power systems have evolved to meet the increasing demands for electric power on a wide range of spacecraft. Cell efficiencies, array designs, and reliability have steadily improved (Table II). Until the 1980s, Si solar cells were used most, and since then GaAs/Ge or multijunction (MJ) cells have been used to meet the specified power needs. Most

TABLE II

Chronological Trends in Space Solar Cell Efficiency (%)[a]

	Year									
Cell type[b]	1960	1965	1970	1975	1980	1985	1990	1995	2000	2002
A	12	12.2	12.5	12.8	13.0	13.0	13.0	13.0	13.0	13.0
B	13	13.2	13.5	13.8	14.0	14.0	14.0	14.0	14.0	14.0
C	13.5	13.8	14.0	14.2	14.5	14.8	15.0	15.0	15.0	15.0
D						16.5	17.0	17.8	18.0	18.0
E						17.0	18.0	19.2	20.0	20.0
F							20.5	21.8	23.5	25.5
G								23.8	25.5	27.5

[a] Air mass zero, 30°C.

[b] A, silicon, 10 ohm-cm, back surface reflector (BSR); B, silicon, 2 ohm-cm, BSR; C, silicon, 10 ohm-cm, BSR, back surface field (BSF); D, silicon, 10 ohm-cm, advanced design; E, gallium arsenide (on GaAs or Ge substrate); F, dual junction (GaInP2/GaAs); G, triple junction (GaInP2/GaAs/Ge).

satellites have performed successfully with PV power levels of a few kilowatts or less. To handle increased data requirements, communications satellites must deliver power levels up to 15 kW, using MJ cells and sometimes low-concentration mirrors. Even higher power levels are needed for constellations of communications satellites. An example is Iridium, comprising 66 satellites, each delivering approximately 4 kW for a total of approximately 250 kW.

Manned space stations require high power. The International Space Station uses large-area, coplanar back contact Si cells specially designed for deployment of large-area, lightweight arrays. Two wings, each approximately 100 ft long and 40 ft wide, deliver approximately 150 kW, and plans call for adding further wings to produce more than 400 kW. This is approximately 10,000 times the average power available on satellites in the early 1960s. The estimated working life of these cells is 10–20 years.

Some well-publicized spacecraft provide good examples of adapting PV technology to suit the mission. The Mars Pathfinder lander derived approximately 200 W from the unfolded panels and 16 W from the Sojourner robotic rover. GaAs/Ge cells were used, and the output of one cell was monitored to measure the gradual loss of power as Martian atmospheric dust settled on the cell.

Other satellites have approached or landed on comets or asteroids. Deep Space I traveled approximately 100 million miles into the asteroid belt. Its PV arrays consisted of MJ cells with lightweight Fresnel lens concentrators, giving approximately 30-times concentration. In addition to operating on-board equipment, the PV arrays powered an ion engine that propelled the satellite on its journey. This solar electric propulsion technology is promising for travel within the inner Solar System.

3.2 Interaction with Mainstream Photovoltaics

There has always been close cooperation between groups developing space cells and those developing terrestrial cells. For many years, Si space cells had the highest efficiency, but terrestrial Si cells were developed with significantly increased output. Space PV has been a highly visible showcase for high-efficiency cells. The process procedures for Si space cells were the most consistent available for objective evaluation of the terrestrial solar cell potential of many forms of solar-grade Si. High-efficiency MJ cells designed and operated under incident sunlight concentrated 500–1000 times have promise for large-scale terrestrial PV applications. III–V semiconductor optoelectronic technology has benefited from PV experience with large-scale OMVPE reactors.

4. FUTURE TRENDS IN SPACE PHOTOVOLTAICS

4.1 To 2012

In the next decade, there will be continued and increased demand for PV power on a wide range of missions. Higher power levels will be required, especially for communications satellites and manned stations. Plans call for exploration of Mars or the

moon to prepare for later manned bases. Development of very small satellites used for smart sensing will begin.

Advanced array designs will be developed, and solar electric propulsion will be used for travel to distant destinations. Some satellites will have to operate in high-radiation fields, including ME orbits. For these missions, solar cells with very high radiation resistance will be needed, probably InP or $GaInP_2$ cells, which have high tolerance to radiation damage as well as the ability to recover later while illuminated by the sun.

Thermophotovoltaic cells developed from III–V solar cell technology may be used to convert heat generated by radioactive isotopes to provide electric power for satellites journeying to distant planets such as Pluto or beyond.

Space-qualified four-junction cells with efficiencies in the mid-30% range will continue to be developed, with effort directed to resolve the severe materials problems involved. The need for higher efficiencies may be relaxed somewhat if transponders can be built to use less power or if electronic circuit designs can further compress the data handled on the satellite.

Current solar cells and arrays and their manufacturing base can probably meet all space power needs in the next decade.

4.1.1 Unmanned Aerial Vehicles

Several applications currently provided by satellites are being planned for unmanned aerial vehicles (UAVs), which include high-flying unmanned airplanes, airships (blimps), or balloons. PV arrays provide propulsion power for the airplanes and airships and power the electronic equipment on the UAVs. UAVs operate at altitudes as high as 100,000 ft for periods of months, and they can be used for surveillance, crop monitoring, or as communications links. The solar arrays on UAVs must be highly efficient and lightweight, and current space-use solar cells are best suited to this application. Current UAV plans call for 50 kW per vehicle, in the next decade increasing to several hundred kilowatts of PV power for constellations of UAVs.

4.2 After 2012

In this period, other power generators, possibly nuclear, will be developed and become available for use in space. It is difficult to predict the needs for space PV beyond 2012, but PV will continue to play a major role. There will be continued extension of most of the previous missions. Manned bases on Mars or the moon will require large amounts of PV power, and it is an advantage that PV systems are modular, allowing large power stations to be assembled during a series of flights.

The problems in manufacturing four-junction cells will be resolved, providing cells with efficiencies of approximately 35%. Long term, there are exciting possibilities to increase efficiencies perhaps as high as 60–90%. These cells, called "third generation" by Martin Green, will require advanced control over semiconductor material properties and possibly application of nanotechnology. Resolution of these material problems and validation for space use are not anticipated until well after 2012.

After 2012, significant advances will be made in the use of small satellites (called microsats or nanosats). These nanosats will act as smart sensors, operating in autonomous clusters or "swarms."

In the future, there may be some very large-scale applications of PV power in space. These projects will require significant advances in solar cell and array technology and use of very sophisticated manufacturing methods. Some of these applications may involve assembly in space, and in the distant future solar cells may be manufactured using lunar materials. The following sections discuss some of these large-scale applications.

4.2.1 Space Power Satellites

Peter Glaser proposed placing large satellites in geosynchronous orbits to provide terrestrial power. The satellites would carry very large-area solar arrays, generating many megawatts (or even gigawatts) of electric power. This power would be converted to microwave power, to be beamed to Earth, and converted back to electric power. In principle, the satellites receive full AMO sunlight for almost 24 h per day. However, this advantage is offset by the conversion losses incurred in beaming the power to Earth. Power satellites are more attractive for regions that have low sunlight levels or restricted areas for deploying PV arrays. There are many problems to be resolved, including the cost and complexity of building and maintaining large arrays in space and in controlling and beaming large amounts of power. Also, at this future time, improved terrestrial PV power generators will be widely available at reduced costs, requiring assessment of whether the many problems with space power satellites justify their use to meet terrestrial needs.

Even more ambitious long-term proposals call for installing several bases containing PV solar arrays on the moon. The electricity would be carried by buried

lines to microwave generators beaming power to the earth.

4.2.2 Space Colonies

Colonies living in satellites located at stable positions in space, between the earth and the moon, will require large amounts of power to sustain life and operate equipment. Preliminary plans call for PV power combined with solar–thermal converters.

4.2.3 Space Cities

Arthur Clarke's concept of cities located at GEO altitudes uses space elevators to transport materials and electric power to orbit. This is a very speculative project, requiring materials that can support structures 23,000 miles long. If successful, it would require removing most of the LEO and GEO satellites operating at that time.

5. CONCLUSION

For almost all of the space age, PV has been a prominent and essential key to successful operation of thousands of satellites performing a wide variety of activities. This continuing niche market has shown great resilience, sustained by the manufacturing of space solar cells with significantly increased efficiency and high reliability. Steady improvement in space cells has served as a focus for general research on high-efficiency cells. The programs to develop III–V-based cells demonstrated that large-area, good-quality crystalline layers could be grown in large-capacity OMVPE reactors. Pioneering work led to the use of Ge substrates for these cells. Some of the experience gained with space cells designed to operate at high concentrations has led to possible terrestrial applications. The unique need to deal with the degrading effects of charged particle radiation on space cells has added considerable understanding of the effect of radiation on semiconductor properties. Current manufacturing technology for space cells and arrays can deliver power for most satellites planned for the next decade and probably beyond. Based on the good record to date, PV power will continue to make major contributions to many exciting space projects.

SEE ALSO THE FOLLOWING ARTICLES

Lunar–Solar Power System • Photovoltaic Energy: Stand-Alone and Grid-Connected Systems • Photovoltaic Materials, Physics of • Photovoltaics, Environmental Impact of

Further Reading

Bailey, S. G., and Flood, D. J. (1998). Space photovoltaics. *Prog. Photovoltaics Res. Appl.* **6**(1), 1–14.

Flood, D. J., and Brandhorst, H. (1987). Space solar cells. *Curr. Topics Photovoltaics* **2,** 143–202.

Green, M. A. (2001). Third generation photovoltaics: Ultra-high conversion efficiency at low cost? *Prog. Photovoltaics Res. Appl.* **9,** 123–126.

Iles, P. A. (1998). From Vanguard to Pathfinder: Forty years of solar cells in space. *In* "Conference Record of the Second World Conference and Exhibition on Photovoltaic Solar Energy Conversion, Vienna, Austria." Vol. 1, pp. LXVII–LXX. European Commission, Ispra (VA), Italy.

Iles, P. A. (2000). Future of photovoltaics for space applications. *Prog. Photovoltaics Res. Appl.* **8**(1), 39–51.

Iles, P. A. (2001). Evolution of space solar cells. *Solar Energy Mater. Solar Cells* **68,** 1–13.

Perlin, J. (1999). "From Space to Earth." Aztec, Ann Arbor, MI.

Rauschenbach, H. S. (1980). "Solar Cell Array Design Handbook." Van Nostrand Reinhold, New York.

Smith, K. D., Gummel, H. K., Bode, J. D., Cuttriss, D. B., Nielsen, R. J., and Rosenzweig, W. (1963). The solar cells and their mounting. *Bell System Tech. J.* **XL11,** 1765–1816.

Smits, F. M. (1976). History of silicon solar cells. *IEEE Trans. Electron Devices* **ED-23,** 640–643.

Photovoltaic Energy: Stand-Alone and Grid-Connected Systems

TIM MEYER
Fraunhofer Institute for Solar Energy Systems ISE
Freiburg, Germany

1. **Applications of Photovoltaic Systems and Markets**
2. **System Concepts and Special Aspects**
3. **Components of PV Systems**

Glossary

autonomous electricity supply system Any electricity supply system that operates independently from the grid, be it in remote locations, mobile, or grid-independent applications for which a grid is nearby.

backup generator Controllable source of electricity in photovoltaic hybrid systems (e.g., based on a piston-type engine).

charge controller Electronic device to control the charging and discharging of batteries and thus protect the batteries.

grid-connected photovoltaic system A photovoltaic system that feeds power into the electricity grid using an inverter.

hybrid system Autonomous electricity supply system based on at least two sources of electricity, such as photovoltaic, wind, or diesel generator, always including a battery storage.

inverter Power electronic device to convert direct current (DC) to alternating current (AC).

maximum power point (MPP) The point of operation at which a photovoltaic module provides the maximum power output.

minigrid An AC electricity grid that forms an island and is not or not always part of the regular grid. Minigrids can be used to power small regions or villages or to maintain the power supply in a region when the main grid fails.

solar home system (SHS) Autonomous DC power supply system for light, radio, or TV used for rural electrification of single houses.

stand-alone PV system Autonomous electricity supply system based on photovoltaics as the power source.

Power supplies based on photovoltaics (PV) were first used commercially for space applications and small appliances, such as pocket calculators and watches. Because of lower prices of PV modules and growing experience in PV system design, higher power ranges have become technically and economically viable.

1. APPLICATIONS OF PHOTOVOLTAIC SYSTEMS AND MARKETS

There are a vast number of applications for photovoltaic (PV) technology, covering all kinds of autonomous systems as well as grid connection (Fig. 1). The power range is extremely wide due to the modularity of PV power, from milliwatts to megawatts. To improve the economics or to use the variety of available energy sources, in a power range of kilowatts and higher so-called hybrid systems are often used. They combine two or more power generators, such as PV, wind turbines, diesel generators, or microhydro generators.

Driven by large market introduction programs, an increasing share of PV capacity has been installed in grid-connected applications. In 2000, these exceeded 50% of the market of the 20 International Energy Agency member countries for the first time (Fig. 2). Due to strong interest in PV technology, further growth of grid-connected systems is expected. However, in the long term the larger economic potential lies in off-grid power supplies. It is estimated that approximately 2 billion people live without electricity from the grid. Autonomous power supplies based on PV or PV hybrid systems are cheaper than diesel systems or grid connection in many of these cases. Socioeconomic and sociotechnical issues are the main barriers to large-scale market takeoff.

 © 2004 Elsevier Inc. All rights reserved.

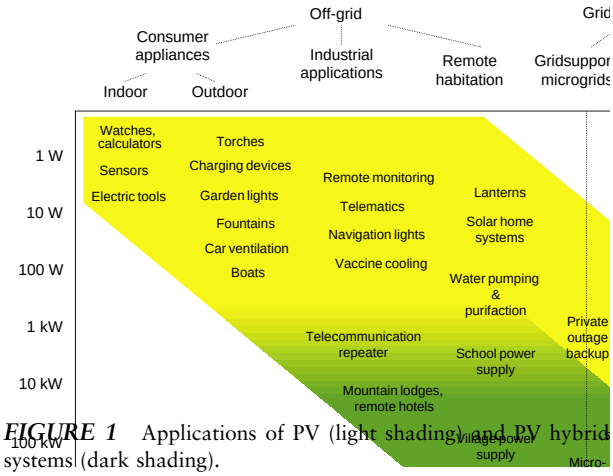


FIGURE 1 Applications of PV (light shading) and PV hybrid systems (dark shading).

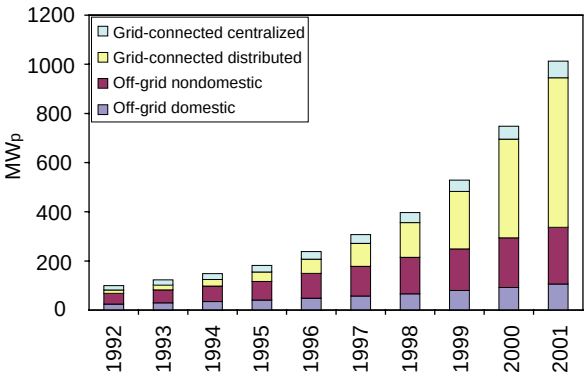


FIGURE 2 PV market development in the IEA countries (Australia, Austria, Canada, Denmark, Finland, France, Germany, Israel, Italy, Japan, Republic of Korea, The Netherlands, Norway, Portugal, Sweden, Switzerland, the United Kingdom, and the United States). Markets for rural electrification are not included. Source: IEA PVPS Programme.

2. SYSTEM CONCEPTS AND SPECIAL ASPECTS

2.1 Autonomous Systems

2.1.1 PV Products and Applications Based on PV Only

Simple autonomous power supplies consist of a PV generator, a charge controller, and a storage unit. However, since batteries have a complex behavior with respect to aging, state of charge, temperature dependencies, etc., the charge controller can have a simple voltage-dependent charge/discharge control or complex algorithms.

Figure 3 shows a solar home system (SHS) as an example of an autonomous power supply system

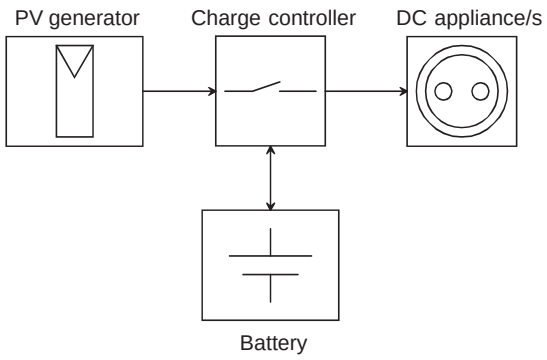


FIGURE 3 Typical configuration of a solar home system.

based on PV only. Typical component sizes are 50 W for the PV module and a 12-V, 70- to 120-Ah lead acid battery; direct current (DC) appliances for this system are compact fluorescent lamps, a radio, and a black-and-white TV set. Typical electricity consumption of such systems is up to 250 Wh per day. In some electrification programs, alternating current (AC) appliances are used in SHSs instead of DC appliances. This requires an additional inverter with a peak power in the range of 150–500 W.

2.1.2 DC-Coupled PV Hybrid Systems

If larger power or energy demand is to be satisfied, additional generators must be introduced into the system. This particularly applies if solar irradiation varies widely throughout the year. To guarantee a reliable power supply during wintertime, large PV generators and storage units are required in these cases. Wind, hydro, or diesel generators are also typically used (Figs 4–6).

Introducing additional components can lower overall electricity cost but also increase the need for energy management and maintenance. Energy management considers the profile of electricity generation from PV or other renewable energy sources, the battery characteristics, and the energy demand of the appliances. To economically optimize energy and battery management, predicting the fluctuating sources (PV, wind, and hydro) can be helpful. The appliances can be either DC or AC. In the latter case, an inverter is used to provide AC electricity. Additionally, an AC bypass from the diesel generator can be used as backup for AC appliances.

2.1.3 AC-Coupled PV Hybrid Systems

Instead of coupling generators and battery storage via a DC bus, hybrid systems can also be AC coupled. The extend to which DC busses are avoided can vary. Fig. 7 shows a system layout in which the

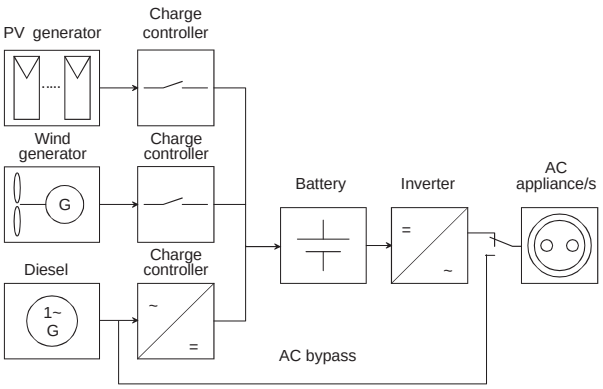


FIGURE 4 Typical layout of a DC-coupled hybrid system.

FIGURE 5 The Rotwand house in the German Alps. Electricity is provided with a PV, wind, and diesel hybrid system in operation since 1992.

PV generator and battery are connected via a DC bus and additional generators are AC coupled to the system. Using the concept of modular system technology, DC coupling is completely avoided and the battery is equipped with extra power electronics (Fig. 8). Modularity adds to the cost of additional equipment and power conversion losses.

In AC-coupled hybrid systems, an island grid is built. There are different ways to achieve this, particularly concerning stability and fast voltage control. In the most common design, one component

FIGURE 6 A telecommunication repeater station powered by a PV hybrid system.

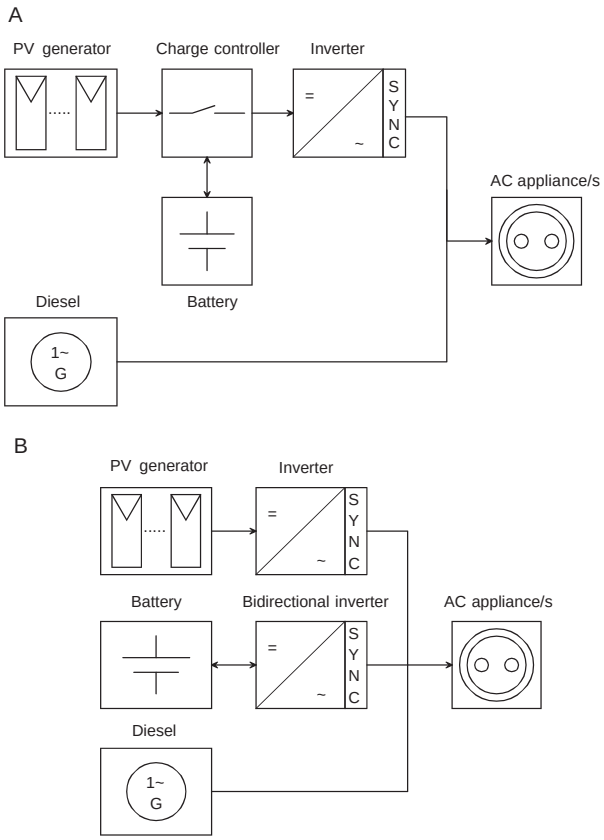


FIGURE 7 Typical layouts of AC-coupled systems (A) with PV generator and battery coupled via a DC bus and (B) following the modular system technology concept.

acts as the voltage source and constitutes the grid, setting voltage and frequency (e.g., 230 V and 50 Hz). The other AC generators then feed in their power, acting as current sources. Reactive power demand (either capacitive or inductive) is provided by the generator acting as a voltage source. The advantage

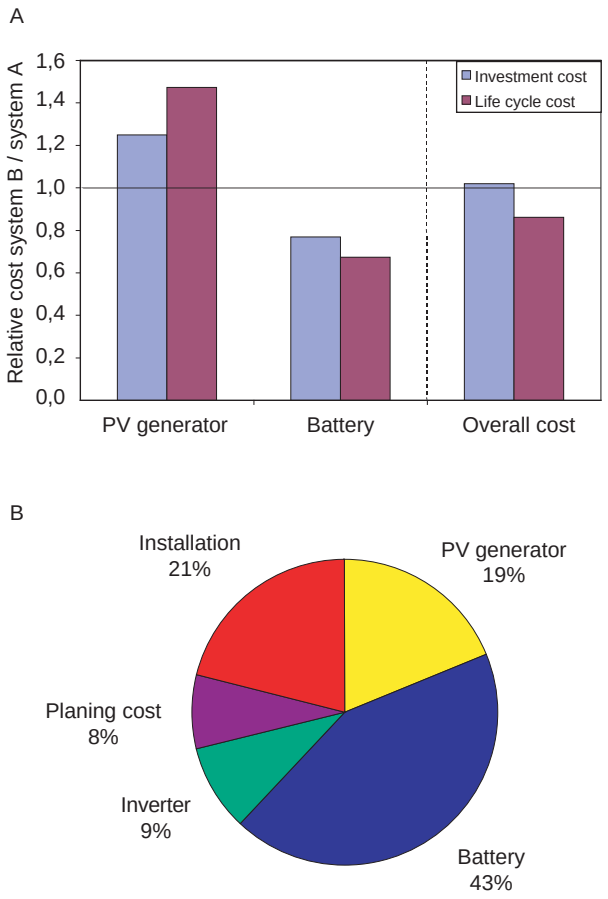


FIGURE 8 (A) Comparison of calculated costs for two systems in Mexico. The systems will provide 1500 kWh per year for 20 years. The interest rate is 6%. System A is optimized toward minimum investment cost and system B toward minimum life cycle cost. In system B, costs are shifted toward the PV generator to reduce the battery stress. Although both systems have approximately the same investment cost, system B has 14% less life cycle cost. The more complex hybrid systems are, the more benefit such economic optimization can yield. (B) Shares of life cycle cost for system A. The battery accounts for more than 40% of overall life cycle cost.

is a relatively easy control structure. However, overall system efficiency may be negatively affected, for example, if a diesel generator is used in building the grid and is mainly running idle just to maintain the grid voltage. Furthermore, if fast load or generation changes occur, the constituting component (diesel or inverter) has to cope with them solely and provide all extra power.

As for large electricity networks, new control concepts distribute this burden to several or all generators, which also build the grid (i.e., constitute the AC voltage). Using droop characteristics for frequency and voltage control is one approach to synchronize the voltage control without the need for very fast and excessive communication between components. The first systems of this type have recently been introduced into the market.

Independent of how fast voltage control is achieved, in DC hybrid systems and AC-coupled systems an overall energy management system determines which components operate and at which power level, thus optimizing energy and battery management.

2.1.4 Main Design Criteria for Hybrid Systems and Advantages and Disadvantages of Different System Concepts

As the number of different types of generators bound together increases, so does the complexity of the system. This also applies if several generators of one type are introduced into the system (e.g., in the case of expansion). Thus, depending on the type of application and the prospect for changing needs in power supply following installation, a number of criteria must be considered during the concept design phase. Some of these are contradicting, so trade-offs must be made.

One of the most expensive components of such systems, the PV generator, has a long lifetime of approximately 20 years. Thus, in order to fully make use of this investment, many systems are designed for such long lifetimes, with the following specifications:

- Lowest possible life cycle cost: The cost structure of autonomous power supply systems is often complex and strongly dependent on operation schemes and depreciation times (e.g., 20 years).
- Stability against load and generation variations: Daily, monthly, and yearly profiles of the load as well as generation by renewable generation vary widely over time. Furthermore, over the years, the amount of energy needed changes in most applications. For example, additional loads might be connected over the years. In applications such as lighting systems in bus stations, rare blackouts are acceptable. In many others, they are not. The system design must be stable against such variations and have a minimum of costly overcapacity (e.g., during summer).
- Modularity and flexibility: If load or generation change too much, the system hardware needs to be changed. Usually, this implies an expansion due to increased energy demand. It must be possible to adapt the system with minimal engineering effort regarding the hardware and the energy management algorithm.
- Maintenance and repair: Over the course of many years of operation, maintenance is crucial. The

system must be as easy to maintain as possible. Also, the availability of spare parts may be crucial.

- Voltage stability and quality of power supply: Quality of power supply does not only include the availability of power and energy but also the quality and stability of the voltage. Especially in AC-coupled systems, this implies strong demand for fast current and voltage control.
- Socioeconomics: In most applications, users interact with autonomous power supplies. Be it in remote houses, village power systems, or for SHSs, without taking sociotechnical and socioeconomic factors into account, the system will fail over the long or short term. This is of particular importance for rural electrification.
- In certain applications such as telematics and signaling, in which blackouts can lead to accidents, the reliability and safety of power supply are of utmost importance.

Many of these demands are contradictory, making optimization complex and susceptible to changed boundary conditions. Thus, new system designs have been developed to more easily and flexibly optimize systems during the planning phase as well as during operation. The designs for energy management constitutes the most important system characteristics and are thus described in more detail.

2.1.5 Energy Management in Off-Grid Systems

Energy management is usually achieved by one central intelligence. All system parameters, such as generator and storage sizes, storage characteristics and charging/discharging strategies, weather forecast (e.g., for PV and wind), and user profiles are fed into the system before operation. Using these parameters and actual measured values of voltages, currents, temperature, PV generation, etc., the system operation is optimized. In addition to standard software, fuzzy logic and artificial neural networks are also used. However, expansion or changes to the system require changes to these system parameters. Furthermore, the system operator requires data that are often not well specified on data sheets, particularly for the battery and the backup generators. Thus, determining a proper operating strategy is difficult.

Alternative methods from automation and control in industrial applications or from management of distributed generation try to decentralize as much intelligence as possible, combining a central control unit with one decentralized control unit per generator, storage, or load. Corresponding software concepts can vary from hierarchically controlled communication and interaction to software agents. In all these cases, component information is stored in a device preferably provided by the component manufacturer. Using measured values, this device is also able to compute cost functions and the availability of the generator or storage (e.g., using the filling state of the diesel tank and maintenance intervals). For these calculations, some information may be provided by the central control unit, such as weather forecasts or the cost for maintenance personnel to visit the site of operation. All cost functions and limiting conditions from the components are then used to optimize the system operation (e.g., using virtual trading schemes). Thus, there is no need to change the energy management parameters when changing the system design. New components are registered and removed components are deregistered automatically and provide data and cost functions in a generic format. The virtual stock market takes into account all registered players at a certain time. Such flexible systems are currently under development and not available on the market.

2.1.6 Cost Structure of Autonomous Power Supply Systems

Autonomous power supply systems often have long lifetimes or depreciation times of up to 20 years. Thus, to develop a mature system design, life cycle cost needs to be considered rather than investment cost. The following are especially important to consider: the complex dependency of aging of the battery on the operating conditions (batteries need to be replaced every 4–8 years), repair or replacement of electronic components, costly maintenance in remote locations, and the cost of financing and penalties for electricity demand that may not be met (if allowable). Because weather profiles differ from year to year, the technical optimization should be based on long time series of 20 years. The difference between systems optimized toward investment cost and those optimized toward life cycle cost is shown in Fig. 8. The main difference between system designs is that, if optimized toward minimum life cycle cost, the battery costs are reduced by appropriate design and operation strategies and by the expanded lifetime or more efficient use of the battery. It is important to note that in autonomous systems the PV generator is not the largest cost factor and that the overall life cycle cost of the system is typically approximately three times more than the pure investment cost (considering interest payments, replacement of components, maintenance and repair, fuels, etc.). When designing an autonomous power

supply, the choice of power generators and their sizes should be determined only with respect to overall lifetime costs. All options should be considered. Nevertheless, in many cases PV or PV hybrid systems are the most cost-effective solution.

2.1.7 Socioeconomic Aspects of Autonomous Systems and Rural Electrification

Autonomous power supplies are the most promising technology for the 2 billion people living without electricity today. During the past few decades, in all regions of the world many projects have demonstrated the technology and its viability. However, a large number of these projects were not sustainable due to equipment failure after a few years, payments not being made by the rural customers, or electrification programs not being accepted by the population as a viable alternative to grid connection. From such projects, many lessons have been learned providing valuable feedback for the design of components and systems as well as business models, infrastructure programs, and policy.

In rural electrification, it is of utmost importance to include cultural aspects in technology and business model design. Sustainable projects and thus sustainable business require not only a screening phase of the respective region and a planning and implementation phase, but also sociotechnical monitoring during operation and ongoing analyses to adapt the strategy. Companies and staff have to learn the specifics of rural electrification and the sociocultural backgrounds of the people, and they must understand that rural electrification is a process. The following issues are crucial:

- There is often a gap between the energy demand and the supply by the PV or PV hybrid system after a short period of time. Once electricity is available, people buy new appliances and consume more energy than expected. This particularly applies if there is little "ownership" for the system (e.g., if a village power supply is used by many users). In this case, the cause (too high energy consumption) and the effect (blackout) are not easily linked to the individual user. In any case, if demand is too high and cannot be met, dissatisfaction is created, possibly leading to stoppage of payments, etc. Although difficult, the future electricity consumption needs to be evaluated before the system is installed.
- Due to erroneous handling of components and systems, these systems degrade or break more quickly.
- Replacement of components is time-consuming and expensive. It is vital to develop a minimum density of systems within a certain area and to install an adequate infrastructure. This includes measures for capacity building, such as training of technical staff and users.
- Lack of financing, especially after several years of service, is often experienced. There might be clear reasons, such as dissatisfaction of customers, but even if the systems run properly, payments tend to decline significantly after several years. Financing schemes and business models need to take this into account.

2.2 Grid-Connected Systems

Grid-connected systems constitute the largest market share for photovoltaics. Major driving forces are market introduction programs in Japan, Germany, and the United States. These systems are targeted at households. Typical system sizes range from 1 to 5 kW. However, there are commercial actors and many systems more than 10 kW and up to several megawatts have been built (e.g., utility owned or joint ownership). The power of inverters that is required to feed the PV electricity into the grid ranges from 100 W to more than 100 kW.

The smallest inverters are module integrated ones, with each single PV module having its own inverter. Main advantages are lower losses at partial shading and lower mismatch losses between modules due to individual maximum power point (MPP) tracking and less demanding cabling because standard AC cables can be used. DC cabling and switches are more expensive because in the case of faults arcs can cause fire. On the other hand, maintenance and repair of many small inverters are more difficult and communication between inverters (e.g., to determine the overall output) requires higher effort. Cost benefits have not yet been realized. It has also been reported that large numbers of module integrated inverters influence each other through increased harmonics on the AC line. Today, few systems are built with module integrated inverters.

The most common system is shown in Fig. 9. A number of PV modules are connected in series to form a string, typically with output voltages at the MPP of 700 V. One or more parallel strings are connected to one inverter to feed the power into the grid. Depending on regulations and inverter power, either one-phase or three-phase inverters are used. Often, large systems are built using many inverters in the range of a few kilowatts. Again, the modularity of PV electricity is apparent.

In addition to the modular concept shown in Fig. 9, for large PV systems, large inverters are used with up to 250 kW or even 1 MW per unit. Both designs, few large inverters or many medium-size inverters, are motivated by cost and neither is superior by principle. The first makes use of the economies of scale, and the latter makes use of the economies of quantities.

The generation profile of PV systems strongly depends on location. In the Northern Hemisphere, most of the energy of PV systems is yielded at partial load, as shown in Fig. 10. Thus, high inverter efficiency at partial load is mandatory. Furthermore, PV inverters with a nominal power rating between 10 and 20% below the peak power of the PV generator are typically used. The inverter is able to handle short periods of overload situations and the energy loss is small compared to the cost reduction due to the smaller inverter.

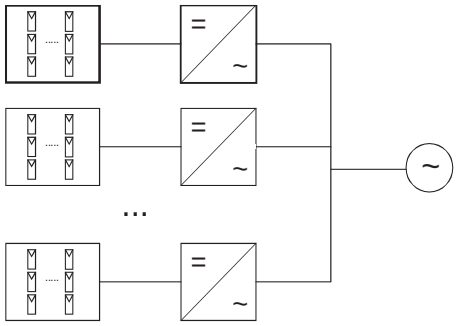


FIGURE 9 Configuration of a grid-connected PV system consisting of several strings.

The performance ratio is used as a measure of the quality of PV systems that is site independent with respect to insolation. The performance ratio is defined as the annual AC energy output of a PV system divided by the annual global insolation in the module plane weighted with the nominal PV module efficiency. Good systems achieve a performance ratio of approximately 0.75. For larger commercial power plants, values of approximately 0.80 should be obtained. The performance ratio aggregates most component and system characteristics: mismatch losses between modules and strings, real module efficiency compared to nominal efficiency, inverter efficiency, and quality of MPP tracking.

2.2.1 Safety

Technical standards, and thus system layouts and inverter concepts, differ from country to country.

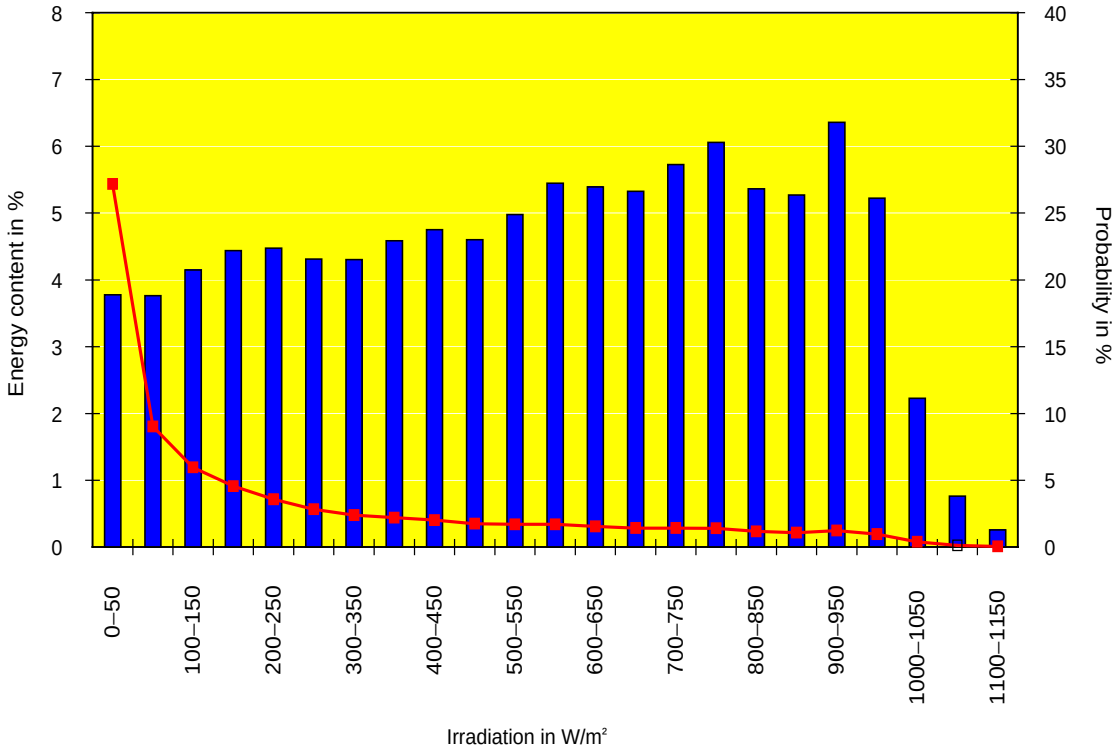


FIGURE 10 Distribution of the energy yield of a PV system in southern Germany. The bars show the energy content for each class of irradiation, and the line shows the probability of the occurrence of this class with respect to time. Most of the electricity is generated at partial load.

Whereas in the United States earthing of the generator is mandatory, in European countries the use of isolated modules (safety class II) is standard. This also allows different inverter designs, particularly transformerless inverters, which have been increasingly used due to their lower weight and higher efficiency.

In addition to earthing, protection against islanding as well as fault current detection in systems with transformerless inverters are necessary, and these are treated rather differently. Islanding is a phenomenon in which during a blackout distributed power sources such as PV systems keep a grid segment—an island—live. This could pose a danger to users, maintenance personnel, and equipment. However, the prerequisite for islanding is that active and reactive power supply by the distributed sources and the active and reactive power demand in the segment be very close. Thus, to prevent islanding it is sufficient to measure voltage levels and frequency and automatically disconnect the inverter if the standard margins are left. If the maintenance personnel follow the safety rules, PV systems do not pose any danger. However, often much more complex methods are used to prevent islanding. With standardization of PV system technology, this might change.

2.2.2 Metering and Cost

To meter the electricity generated, either an extra meter is installed or the electricity is fed, for example, into the house wiring that supplies existing loads (net metering). In the latter case, if PV generation exceeds the actual load, either the meter runs backwards or an extra meter measures the excess electricity.

The generation cost of PV electricity in grid-connected systems varies between approximately 30 €c in sunny regions and 60 €c in northern Europe if good practice is assumed. Unlike in autonomous systems, the PV modules contribute the largest share of life cycle cost. Further cost reduction, however, is achievable using inverter technology and better installation practices.

2.2.3 Building Integration Aspects

Solar energy by nature is a distributed energy source. In order to optimally use existing infrastructure and not unnecessarily seal additional areas, building integration is the best option for grid-connected PV. The PV modules can also provide the following:

- Weather protection (wind and water)
- Sun protection (glare and heating of buildings)
- Use of daylight
- Noise protection
- Act as a blind
- Shielding of electromagnetic fields
- Aesthetics

As such, there are a large number of PV module technologies involving isolating and nonisolating modules, different mounting principles, different module and cell colors, etc.

The cost of integrating PV in rooftops and facades and that of the required standard material can be regarded as a credit. Particularly for facade integration, for which costly material such as marble or metal constructions are often replaced, this can compensate a large share of the overall system costs. This remains true even though facade-integrated PV modules have a significantly lower energy yield due to the nonoptimal tilt angle. In northern Europe, the energy yield is reduced by approximately 30–40% (Fig. 11).

For roof and facade integration, shading of PV modules and operation temperature are of special importance. If one module of a string of modules is shaded, the output of the whole string can be dramatically reduced, depending on the connection scheme and the use of bypass diodes. Thus, particularly for PV facades, which often experience complex shading situations from neighboring buildings or trees, the system must be designed carefully. System parts that are shaded during certain times should have a separate inverter and should be made less susceptible through the use of additional bypass diodes.

Because temperature affects the performance of PV modules, rear ventilation by natural convection is

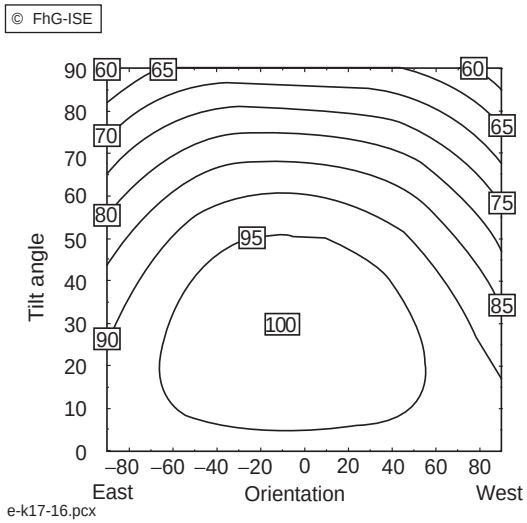


FIGURE 11 Maximum annual PV module yield as a function of orientation and tilt angle in Freiburg, Germany (48°N, 7.8°E).

another issue to consider for building integrated systems. The lower the temperature, the higher the yield. The difference in performance between PV modules with good ventilation and those with bad ventilation can be up to 10%.

2.2.4 Distributed Generation Aspects

Building integration of PV modules can provide significant additional value, as described previously. Another dimension of added value can be tapped if the service provided to the grid is considered. However, improperly integrated PV can also pose additional burden to the grid and thus cause additional cost. Possible factors are as follows:

- Correlation of PV generation and load (reduced peak load in distribution grids)
- Increased security of supply in weak grids (continued operation during blackouts in remote areas)
- Reduced distribution losses in cables and transformers
- Reduced load of transformers, cables, switches, etc. leading to reduced aging
- Deferral of investments for capacity expansion of existing grids
- Active compensation of reactive power or harmonics
- Reduced risk due to short lead times for planning and modularity of PV technology

The actual added or reduced value of grid-connected systems has not been studied. However, several case studies show that at least part of this potential can be tapped.

3. COMPONENTS OF PV SYSTEMS

3.1 PV Modules

Crystalline silicon solar cells comprise approximately 90% of the market share of PV modules. These PV modules typically consist of 36 cells, but other configurations can be bought off the shelf. The current-voltage curve of a 36-cell module fits very well to 12-V lead acid batteries, so simple autonomous power supplies can be built. However, the exact number of 36 cells is more historically motivated. With improved cell technology, even at higher ambient temperatures the module voltage would still be high enough to fully charge a lead acid battery if 32 cells were connected in series. Depending on climatic conditions, even less cells could be used.

The yield of a PV system is usually measured in kilowatt-hours per kilowatt peak (kWh/kW_p) and depends on several factors:

- Cell and module technology used
- Location/climatic conditions
- Tilt angle and orientation
- Mechanical installation (facade or roof integration, freely ventilated modules etc., influencing module temperature)
- Efficiency of system components (MPP tracking and conversion efficiency)
- Type of system

The area needed for 1 kW of PV power mainly depends on the cell and module technology used. Module efficiencies vary from several percent (for some thin-film modules) to approximately 16 or 17% for monocrystalline silicon solar cells. Typical module sizes for grid connected systems range from 50 to more than 300 W. Especially for solar lanterns and lighting, modules in the range of 10–20 W_p are available. Because 36 cells per module are standard, smaller modules are made from cells that are cut into half or quarter cells. Thus, small PV modules with crystalline silicon solar cells are more expensive. This does not apply to thin-film modules.

The performance of PV modules depends on their operation temperature. Depending on the technology used, the temperature coefficient is between -0.2 and -0.5%/K. Thus, the type of installation and degree of rear ventilation are important.

PV modules are rated in watt peak under standard test conditions (STCs). The STCs are defined as 1000 W/m^2 insolation with a AM solar spectrum of 1.5 and 25°C module temperature. However, these conditions rarely occur in real systems. Thus, PV generators can actually yield significantly less or more energy than rated, depending on the location, climatic conditions, etc. However, there is no widely accepted alternative rating model that takes into account more realistic reporting conditions. Furthermore, there is relatively large uncertainty of the measurements at the manufacturing plant and on site. Spectral mismatch, temperature distribution, etc. comprise approximately 3–5% of uncertainty. Overall deviations in systems measurement can be 5–10%.

Whereas in grid-connected systems usually all energy can be fed into the grid, in autonomous systems part of the energy yield in summer cannot be used due to limited storage capacity. Thus, the tilt angle of PV modules in autonomous systems is usually optimized for large energy yield during winter. The overall yearly energy yield is reduced.

Fig. 11 shows the maximum energy yield of PV modules as a function of orientation and tilt angle for the Northern Hemisphere (48°N, 7.8°E).

3.2 Batteries

Batteries are a central component in autonomous power supply systems based on PV or PV hybrids. As discussed previously, batteries contribute the largest share of cost of autonomous systems if life cycle cost is considered. Thus, except for small, portable systems, lead acid batteries are the most common technology used. They are the most cost-effective and have good electrical characteristics (e.g., energy efficiency in PV systems can be as high as 85–90%). Several different technologies for electrodes and electrolytes are available. Thus, for all applications suitable products can be chosen with a wide range of electrical and lifetime characteristics.

For reliable and economic operation of autonomous systems, the charging/discharging principle and system operation strategies are highly relevant in order to make full use of the installed capacity and reduce aging of the battery. In particular, most IV (current phase, voltage phase) charging strategies in autonomous systems rarely fully charge a battery if non-optimal conditions are assumed (e.g., during periods of reduced PV generation). Deep discharge in winter or during longer periods with little generation from PV or wind generators accelerate aging processes.

Depending on the application, different classes of systems can be defined that determine how the battery is stressed (Fig. 12). In addition to the dimensions of the system, the choice of battery type and operation strategy is strongly influenced by these classes.

3.3 Charge Controllers

Charge controllers ensure that the battery in autonomous systems is treated as well as possible, particularly that deep discharge and overcharging are avoided. Furthermore, some devices provide information about the battery state of charge. Modern charge controllers use charging and discharging procedures that take into account the battery temperature and the charging and discharging currents over time. Today, IV charging is state of the art, but new charging strategies are being developed [e.g., IVI (current phase, voltage phase, current phase) charging].

There are two basic designs for charge controllers in PV applications: series or parallel regulators. A series regulator adds a voltage drop between the PV module and battery, whereas a parallel regulator draws current from the PV module bypassing the battery (Fig. 13). Hybrid configurations have also been developed. In all cases, voltage regulation can be performed statically or can be achieved as an average by pulse width modulation. The advantage of the latter is reduced power losses in the

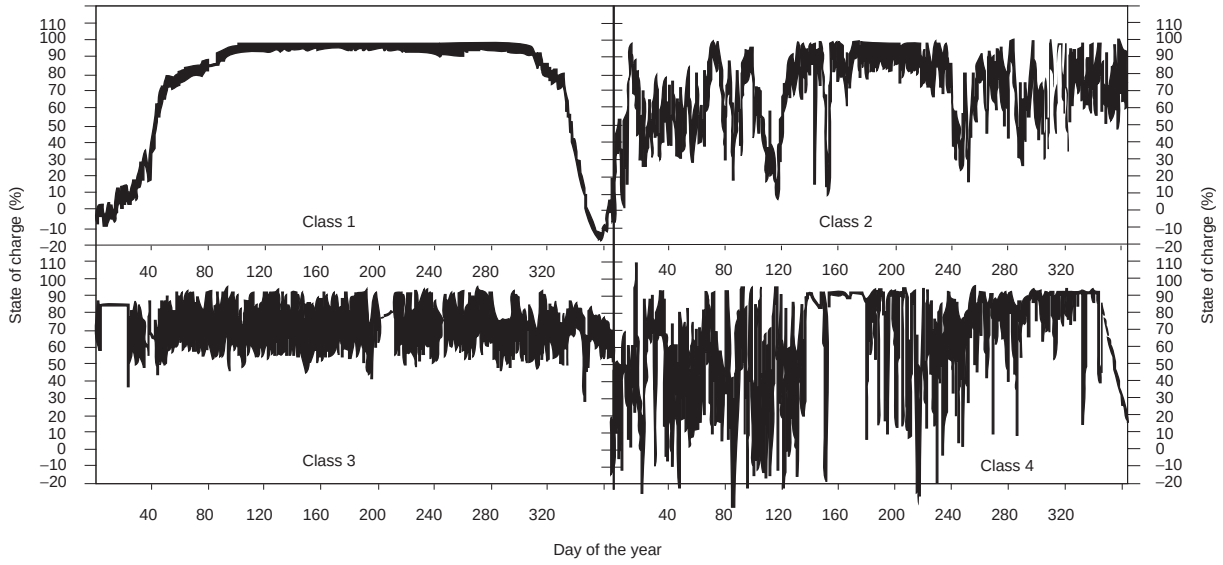


FIGURE 12 State of charge of the battery during a year for four typical operating conditions. The annual energy turnover ranges from approximately 20 times the nominal capacity in class 1 to 180 times in class 4. This demonstrates the different requirements for the batteries under different system designs.

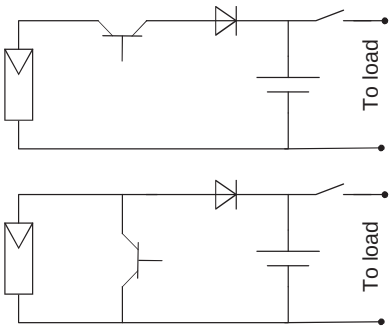


FIGURE 13 Basic principles of charge controllers: series (top) or parallel (bottom) regulator. To avoid deep discharge, the load can be disconnected. A series diode prevents reverse power flow through the PV generator.

semiconductor switch for the microcycles in the battery current and electromagnetic interference (EMI).

The most important functions of charge controllers are to

- Avoid high voltages during charging of the battery (temperature compensates cutoff voltage by −4 to −6 mV/K per battery cell)
- Avoid deep discharging of the battery
- Avoid extended periods without full charging of the battery
- Slightly overcharge once a month to ensure that the battery is fully charged and to promote gassing if applicable

Since some of these aims are contradictory, suitable compromises must be found (e.g., between avoiding high voltages to reduce corrosion and occasionally overcharging the battery to reduce sulfation).

Charge controllers should have a low internal consumption, high efficiency (96–98%), and make reverse poling impossible. They should withstand overvoltages (at least twice the open circuit voltage of the solar generator) and ambient temperatures of 0–50°C. Some charge controllers use DC/DC converters to adapt PV module and battery voltage. They can also perform active MPP tracking, whereas with most charge regulators the PV module is operated at the level of the actual battery voltage.

Additional functions, such as starting backup generators and stirring the electrolyte, might be added for systems of higher complexity. In any case, providing information about the state of charge or even state of health of the battery is crucial in order to enable the user to learn how to operate the system effectively, to establish optimal energy management, and to avoid detrimental deep discharging of the batteries.

3.4 Inverters

Inverters are used to convert the DC output of PV or a battery to AC electricity, either to be fed into the grid or to form an island system. There are many different types of power electronic topologies used in the market. Early PV systems were equipped with thyristor inverters, which were commutated by the grid. Due to their poor voltage and current quality (such inverters have very large harmonics), they have been replaced by self-commutating inverters (with IGBT or MOSFET as semiconductor switches). Only in the 100-kW power range are thyristor inverters still used. The first self-commutating inverters performed sine modulation of the output current on the primary side and had a 50-Hz transformer to adapt the voltage level to grid voltage. Such inverters still comprise a significant market share. Some manufacturers offer devices with high-frequency transformers instead, thus reducing the efficiency and the cost. In countries in which earthing of the PV generator is not mandatory, transformerless inverters are increasingly used. They are lighter and more efficient, but particular care must be taken with regard to EMI and fault current detection.

Whereas the first PV inverters were often adaptations of existing inverters for electrical drives, today's devices are manufactured taking into account the complexity and demands of specific PV applications. Inverter efficiency is generally important in PV applications but of particular significance at partial load because the bulk of the energy is yielded at partial load. Furthermore, with a large-area PV generator coupled to the DC side of the inverter and the public grid on the AC side, stricter standards have to be met with regard to harmonics and EMI.

Design criteria and functions of PV inverters are as follows:

- Efficiency: well above 90% efficiency at 5% of nominal load
- Cost
- Weight and size: inverters with a 50-Hz transformer typically weigh more than 10 kg/kW, sometimes causing a problem with handling
- Voltage and current quality: harmonics and EMI
- Overload capability: approximately 20–30% for grid-connected inverters and up to 200% for short-term overload for island inverters

- Precise and robust MPP tracking (reliably finding the overall MPP in partial shading situations)
- Supervision of the grid, safety/ENS (Einrichtung zur Netzüberwachung)
- Data acquisition and monitoring

To optimize inverters following these partially contradictory aims, the design of the power stages and control algorithms are increasingly more integrated. To implement suitable control schemes, digital signal processors are increasingly used.

To account for the fact that PV electricity is yielded following a typical distribution, in Europe a weighted efficiency measure is used to aggregate the efficiency curve of inverters—the European efficiency. Good inverters in the kilowatt power range have a European efficiency of 92–96% and up to 97% for large central inverters.

Inverters also perform MPP tracking in order to optimally operate the PV generator. Many different algorithms are used for MPP tracking for the following:

- Precision: high precision of MPP tracking requires high-precision measurement components
- Finding the global maximum power output in the case of partial shading, when a local maximum can occur
- Quickly adapting the MPP to changes in insolation (e.g., if clouds pass by).

3.4.1 Inverters for Island Systems and Minigrids

Inverters for island systems and minigrids are very similar to inverters for grid connection with regard to the power stages. Most inverter topologies are suitable for island operation as well. The major prerequisite is the capability to provide reactive power. The actual design of the power stages still differs significantly because in island systems the inverter requires a much higher overload capability. Furthermore, fast voltage and current control and higher dynamics of load profiles require appropriately dimensioned power stages.

Thus, there are major differences concerning control algorithms for voltage and current control. Particularly when several inverters are to be connected in parallel in island systems or when inverters are to be combined with other AC generators, achieving voltage stability is a complex task.

3.5 Backup Generators

In PV hybrid systems, backup generators are used to provide additional electricity in times of low insolation, to increase system availability, and to reduce overall electricity cost. In most medium or large power applications, diesel generators are the cheapest backup solution (backup capacity of 5 kW or more). In many countries, subsidized prices for diesel contribute significantly. However, the major disadvantages of piston-type backup generators are their need for maintenance every few hundred hours of operation and bad efficiency during operation at partial load, which is the common operation mode. In remote areas, access to the systems is often difficult or time-consuming, thus imposing extra cost for maintenance and fuel transport. The availability of spare parts and competent maintenance staff can be a major problem in developing countries in which diesel or PV–diesel hybrid systems are used for electrification of villages.

A major problem regarding relatively small hybrid power systems is the lack of suitable backup generators in the power range of 1 kW or less. Thus, oversized piston-type generators are sometimes used for applications in which 1-kW backup capacity would be sufficient. Due to the poor efficiency of diesel and Otto engines under partial load conditions, the overall system efficiency of such systems is not satisfactory.

In small hybrid systems, thermoelectric converters with a backup capacity of 10 to hundreds of watts are also used. In the medium term, fuel cells are an interesting alternative to small diesel or thermoelectric units because they promise high efficiency and low maintenance cost. However, the problem of hydrogen supply needs to be resolved.

SEE ALSO THE FOLLOWING ARTICLES

Batteries, Overview • Photovoltaic Conversion: Space Applications • Photovoltaic Materials, Physics of • Photovoltaics, Environmental Impact of • Solar Cells • Solar Energy, History of • Solar Thermal Power Generation

Further Reading

Erge, T., and Sick, F. (1996). "Photovoltaics in Buildings—A Design Handbook for Engineers." James & James, London.

Gabler, H., Preiser, K., Bopp, G., Sauer, D. U., and Schmidt, H. (1998). Energy storage in photovoltaic stand-alone energy supply systems. *Prog. Photovoltaics* **6**, 271–291. [Special issue].

Luque, A., and Hegedus, S. (eds.). (2003). "Handbook of Photovoltaic Engineering." Wiley, New York.

Photovoltaic Materials, Physics of

BOLKO VON ROEDERN
National Renewable Energy Laboratory
Golden, Colorado, United States

1. Solar Cell Operation
2. Crystalline Silicon
3. Thin-Film Silicon
4. Polycrystalline Thin Films
5. Concentrator Schemes and III–V Solar Cells
6. Emerging Schemes
7. Future Potential (Physical Limitations of Materials)

Glossary

amorphous silicon An amorphous semiconductor that has allowed fabrication of moderately high-efficiency thin-film solar cells using a deposited layer of silicon–hydrogen alloy. It lacks crystallinity or the exact periodical arrangement of atoms.

buffer layer A thin layer inserted near the junction of a solar cell to enhance its performance, mostly by enhancing the voltage of a solar cell.

cadmium telluride A thin-film semiconductor that has allowed fabrication of high-efficiency thin-film solar cells.

copper indium diselenide A chalcopyrite semiconductor that has allowed fabrication of very high-efficiency thin-film solar cells.

diode A rectifying semiconductor junction. If designed appropriately, it will deliver electric power upon illumination and function as a solar cell.

silicon The most common material used to build semiconductor devices and solar cells.

solar cell A solid-state semiconductor device used for the direct conversion of light into electric power (i.e., photovoltaic energy conversion).

At the beginning of the 21st century, the direct conversion of sunlight into electricity (solar electricity) via semiconductor devices called solar cells or photovoltaic modules was commercially established. The total number of photovoltaic modules produced worldwide (approximately 500 MW in 2002) was much too small to be considered energy significant. However, energy significance can be expected between 2010 and 2015, when annual production levels of several 1000 MW per year are reached, based on expected annual growth of 20–30%. Assuming a conversion efficiency of 10%, 1000 MW of photovoltaic module power will cover an area of 10 km^2 or approximately 4 square miles. In a sunny location, an area of approximately 160 × 160 km (approximately 25,000 km^2 or 10,000 square miles) would be required to generate from solar power all the electricity currently consumed by the United States. Therefore, a tremendous industrial scale-up must occur to accomplish these goals. The material candidates likely to be used for achieving such commercial potential and the associated physics that relates to solar cell performance and performance limitations are discussed in this article.

1. SOLAR CELL OPERATION

1.1 Construction of the Solar Cell Device

A solar cell is a device that allows the collection of excess electric carriers, called electrons and holes, because of a built-in potential. When semiconductors are used to build such a device, the operation is determined by the bandgap [E_g, usually given in units of electron volts (eV), or voltage multiplied by the charge of the electron e] or energetic separation of the conduction and valence bands. Light is absorbed when the quantum energy of the light (the photon energy) is greater than the bandgap. Under the influence of the built-in potential and, to a lesser degree, gradients in the carrier concentration, carriers will flow from one side of the cell to the other. Electrons will move along the conduction band edge, and holes will move along the valence band edge. Figure 1 represents a solar cell in its most generic form. The schematic depicts the physical composition of layers, using terminology that typically describes

 © 2004 Elsevier Inc. All rights reserved.

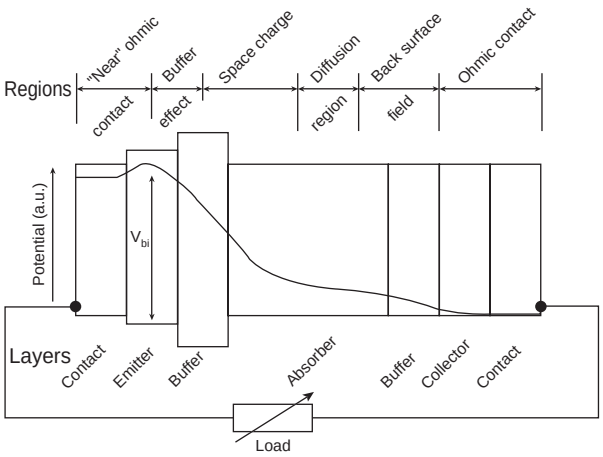


FIGURE 1 Schematic of a p-i-n diode. The potential is shown for no-lead (i.e., open-circuit) conditions.

thin-film solar cells. The terms along the top of the schematic indicate the electronic regions of the cell. The built-in potential, V_{bi}, will determine the maximum voltage that the cell could theoretically deliver. For a homojunction cell such as a crystalline silicon cell, the emitter and collector are usually formed by heavily doping the silicon p-type and n-type (as shown for the right-hand collector in the schematic). However, in heterojunction cells, the bandgaps of the materials making up the junction may vary (as shown for the left-hand side, i.e., the emitter in the schematic). For a homojunction cell, the maximum value of V_{bi} is approximately, but less than, E_g/e. In heterojunction solar cells the determination of V_{bi} is less straightforward because the bandgaps of the emitter, buffer, and absorber layers are different. In this instance, in addition to the value of E_g of each layer, the relative alignment of the conduction or valence band edges depends on the "affinity" or work function of each layer. However, in all cells, the potential profile is a smooth continuous curve passing through the device. Emitter and collector layers with low conductivities may not have enough lateral conductivity to collect the currents at a metal gridline. In these instances, the insertion of a transparent conductive layer is required. SnO_2, ZnO, or indium tin oxide layers are used as transparent conductors. Often, this layer has a double function, acting as an antireflection coating in addition to providing the electrical contact.

The potential profile through the cell will also be determined by the load. Figure 1 depicts the device near open-circuit conditions. When one device contact is shorted to the other one, they have to come to the same potential, forcing a lowering of the electrical potential on the left-hand side down and/or rising on the right-hand side. Traditionally, much analysis went into identifying collection losses within a specific layer or region of a cell. Regions are often characterized by the material properties of the layer making up most of it. It is important to realize, however, that the distinction of such regions in terms of layers alone is somewhat arbitrary. Manipulation of the physical construction of one region of a solar cell or merely a change in the operating condition of the solar cell device will affect what occurs in adjacent regions.

1.2 Solar Cell Operation

The device shown in Fig. 1 is basically a p-i-n diode, where the emitter and collector doping define the junctions. The cell absorber layer can be lightly doped, undoped (i = intrinsic), or, in some instances, show a naturally prevailing doping that cannot be easily controlled by adding chemical dopants. If a bias is applied to such a diode in the dark (replacing the "load" in Fig. 1 with a power source), the current flow is blocked for one polarity of the applied voltage (reverse direction) and increases exponentially for the opposite polarity. Such "forward" current flow can be approximated by the diode equation:

$$I(V) = I_o[\exp(V/nk_BT) - 1], \qquad (1)$$

where I and V are the currents and voltages, respectively; n is the diode quality factor (a fitting parameter); k_B is the Boltzmann constant; and T is the temperature in Kelvin. The diode quality factor is often interpreted to account for "losses" or nonideal diode behavior. Some theories predict that the factor should fall between 1 and 2 depending on the recombination loss mechanism encountered. In practice, there is little correlation between solar cell performance and the actual n values used in fitting experimentally observed I–V characteristics. For moderately high-efficient cells, unpredicted values of $n > 2$ have been observed. The I_o factor depends on the height of the barrier, which in a homojunction cell is given by

$$I_o = I_{oo} * \exp(-E_g/mk_BT), \qquad (2)$$

where m is a fitting parameter and I_{oo} a constant. Very often, the sum of two superimposed exponential curves is used to fit experimental $I(V)$ data. Theories have been developed to relate the magnitude of the fitting parameters to material properties, most commonly those of the absorber layer, which is often

assumed to be the most important factor controlling the performance of a solar cell.

When the diode is illuminated, a current flows in the reverse direction under the influence of the built-in field. If the device is short-circuited, the current density measured (per unit area of the cell surface) is called the short-circuit current density (J_{SC}). When the device is not connected to a load, the open-circuit voltage (V_{OC}) is measured. When the device is connected to a load, an *I–V* curve, as shown in Fig. 2 for a $CuInGaSe_2$-based solar cell, is obtained. Following the curve, a value for J and V can be found where the power output $J_{mp}V_{mp}$ is maximized (mp = maximum power). The ratio of $J_{mp}V_{mp}/V_{OC}J_{SC}$ is called the cell fill factor (FF). The power P generated by a solar cell is $P = FF \times V_{OC} \times J_{SC}$. The cell efficiency, η, relates the fraction of the total power of the incident sunlight (the solar constant, $S = 1\,kW/m^2$) to the electrical power generated by the cell (strictly speaking, under normal incidence and for a cell operating at a temperature of 20°C). Efficiency $\eta = P$ (kW/m^2)/S (kW/m^2) and is usually reported as a percentage. It is important to note that there can be a difference in the current–voltage behavior of the dark and the light current, commonly causing a "crossover" of the measured curves. Such observations clearly complicate the way in which solar cell efficiency can be related to the material properties used in a solar cell.

Another important property frequently measured is the solar cell collection efficiency and its spectral dependency. This is important because sunlight has a wide energetic distribution of photon energies. In the ideal case, all photons with energies larger than the bandgap E_g of the absorber generate one electron-hole pair, and ideally these can be separated and collected. Photons with energies $< E_g$ are not absorbed, thereby not creating carriers. Photons with energies much larger than E_g still only generate one electron-hole pair, and the excess energy is lost through thermalization because the carriers can be transported and collected only at the band edges. This fundamental limitation to solar cell efficiency is referred to as the Shockley–Queisser (SQ) limit after the authors who first considered this effect. Because of the SQ limitation, there is an optimum bandgap for the cell absorber; such an optimum is found by trading off the inability to absorb the low-energy solar photons in larger E_g absorbers and the need to increase the E_g and built-in potential (and thus voltage) of the cell. The optimum bandgap for a single absorber is approximately 1.4 eV. However, for E_g values between 1.0 and 1.8 eV, the dependency of the cell efficiency on the bandgap is not very strong, making all materials within this bandgap range candidates for solar cell absorbers. Because the electronic properties of a semiconductor may change as the bandgap is altered (e.g., by alloying), absorbers used in high-efficiency cells are often selected by making a compromise between bandgap and electronic performance.

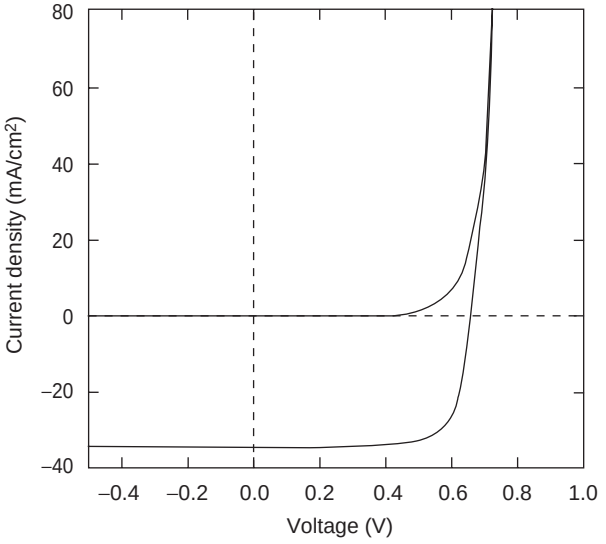


FIGURE 2 I–V curves under light and in the dark for a $CuInGaSe_2$ solar cell. Parameters are as follows: $V_{OC} = 0.66$ V; $J_{SC} = 34.3\,mA/cm^2$; FF = 77%; eff. = 17.4%.

A demonstrated way to overcome the SQ limitation is to stack two (or more) cells with different absorber bandgaps on top of each other. Thereby, the higher energy photons are absorbed in a higher bandgap top cell(s) (delivering higher voltage), whereas the low-energy photons pass through the top cell(s) and are converted in the lower E_g bottom cell(s).

Figure 3 shows the quantum efficiency (QE) of the cell from Fig. 2. The cell response is determined for light of single wavelength (or a narrow band of wavelengths), measuring the relative response with monochromatic or narrow-bandwidth light. Ideally, this curve would be almost a rectangle. In practice, optical and electrical collection losses determine the shape. There are also reflection, shadowing, and sometimes absorption losses as light enters the cell. It is often straightforward to account for such optical losses. It is more difficult to analyze the losses to the QE response that result from incomplete carrier collection. These losses may depend on the light penetration into the absorber—that is, how far away from the emitter (or collector, if the cell has a back-surface

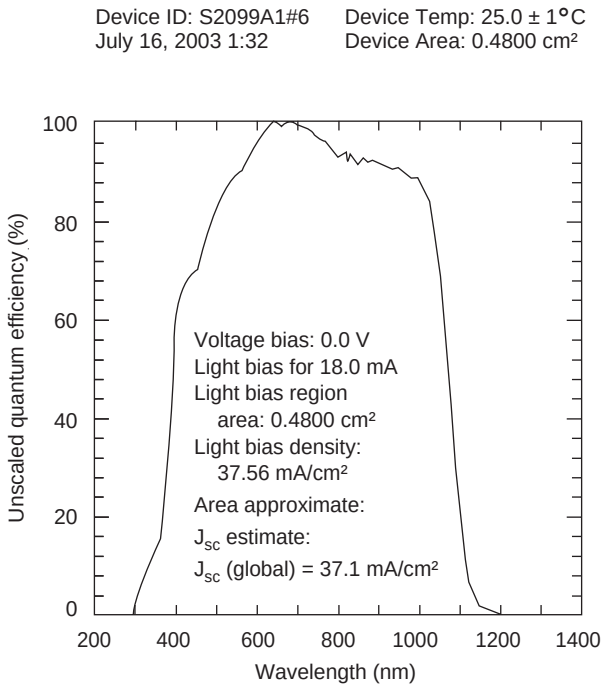


FIGURE 3 Spectral response for the cell shown in Fig. 2. The curve shows the relative response at each wavelength relative to the strongest response that defines the 100% value.

field) the carriers were generated. Except for amorphous and nanocrystalline thin-film silicon, organic, and dye-sensitized TiO_2 (Grätzel) solar cells, usually the absorber layers are thick enough to absorb most of the incident sunlight with photon energies $> E_g$. To first order, the shape of the long-wavelength fall-off of the QE is determined by the absorption edge of the absorber material. More detailed analyses reveal that the response depends on the details of cell design and the experimental conditions used, such as probe and bias light intensities, bias light wavelength, and electrical bias. For example, in crystalline Si solar cells, the edge of the long-wavelength QE response is determined by the doping of the wafer "base" (absorber). The base doping (at the levels typically used) does not change the optical properties of the Si wafer, but the response changes as if E_g were increased.

1.3 The Role of Buffer Layers

Solar cells are made from a wide variety of materials, and traditionally, less-than-ideal performance was attributed to very specific imperfections and defects found in the materials used to build the cells. It is also of interest to identify and characterize mechanisms and observations applicable to cells made from different materials. For many cell technologies, it has been a challenge for the experimental device fabricator to obtain cells with high V_{OC}. The opposite problem typically exists when a device structure is run in a computer model. The most important factor determining V_{OC} is the bandgap of the layers involved, so the initial modeling calculations often predict V_{OC} values greater than those observed experimentally. Although modeling may introduce ad hoc thin defective layers at the interface to match calculated and experimentally observed V_{OC} values, researchers usually maximize V_{OC} by employing grading schemes for the electronic and optical properties of the layers used near the junction. In thin-film solar cells, these schemes are typically referred to as buffer layers. In crystalline Si- and GaAs-based solar cells, these schemes are often referred to as contact passivation. It can be argued that a unique voltage-enhancing mechanism is operational in all solar cells built from different semiconductor materials.

Figure 4 presents a unifying buffer layer concept that can explain how buffer layers enhance solar cell voltages. Traditionally, it is assumed that the recombination loss current j_j would be determined by recombination through surface or interface defect states near the junction. Figure 4 suggests that, instead, the transport properties or carrier mobilities of the materials used to build the junction should be reduced (e.g., by using materials with potential fluctuations of the band edges that have significantly reduced carrier mobilities). This configuration reduces access of carriers to such voltage-reducing loss pathways. Solar cell fabrication depends on robust, repeatable, and scalable processes for creating an optimum or near-optimum grading of the transport properties in the materials used to build the junction regions of the cell. Ideally, the carrier-transport mobilities should be sufficient to allow carrier collection in the direction of the applied field but no larger. The principle can be best understood by comparing a Schottky barrier (SB; metal/semiconductor) crystalline Si junction and a metal/insulator/semiconductor (MIS) junction. Although working SB solar cells can be fabricated, they suffer from low cell voltage. On the other hand, MIS devices can attain voltages comparable to the best achievable by other junction-formation techniques. Although it is sometimes argued that the oxides provide passivation of the interface, it is well-known that for crystalline Si one or two atomic layers of thermally grown oxide provide excellent surface passivation. However, for

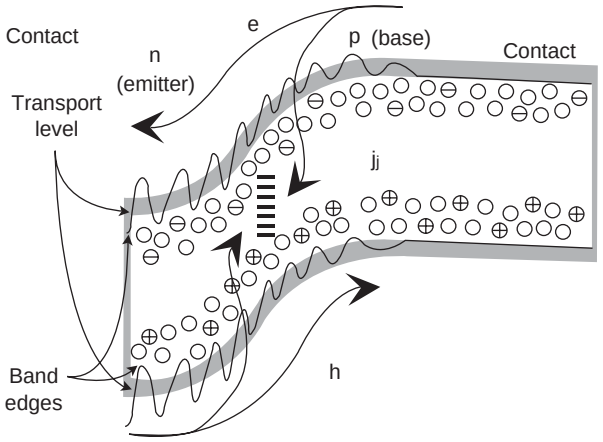


FIGURE 4 A unifying buffer-layer concept that can explain how buffer layers enhance solar cell voltages.

an MIS structure to exhibit much larger V_{OC} and device efficiency than the equivalent SB structure, the oxide buffer layer must be approximately 1.5 nm thick, nearly preventing carrier transport through this thin insulating layer. In a diffused crystalline Si p-n junction, the existence of a specific buffer layer is not generally acknowledged; however, a low-mobility buffer is likely present (i.e., a compensated layer as the phosphorus emitter diffuses into the boron-doped base). A buffer layer concept helps explain the trade-off between cell current and voltage that is often observed in the experimental research on optimizing solar cells. Because of its resistive nature, the buffer will cause a larger drop in the electric field near the junction, thereby reducing the extension of the space charge region in the absorber layer. Traditional cell analyses account for the measured device behavior in terms of surface and absorber recombination losses but cannot provide an explanation for the commonly occurring current-for-voltage trade-off observed in the optimization of experimental and commercial solar cells. It is important to realize that adding too much of a buffer layer can have disastrous consequences and ruin the solar cell performance. For industrial purposes, the thickness and electrical properties of the buffer layers must be uniformly and repeatedly obtained over large areas to achieve a consistent product.

2. *CRYSTALLINE SILICON*

2.1 Workhorse of the Industry

Most photovoltaic modules are made from crystalline silicon cells. Each cell is made from wafers produced either from Czochralski single crystals or from large-grain polycrystalline ingots. Many photovoltaic manufacturers produce their own wafers in order to have tight control over manufacturing cost and wafer quality. In fact, in the early stages of technology development, inventions focused on deriving economical and performance advantages predominantly from the wafer-preparation process. The crystal growth processes appear to have become less critical in recent years because the structural perfection may be a lesser factor for cell performance than initially thought. Instead, much manufacturing cost reduction was derived from changing the wafering technology by switching from inner-diameter saws cutting one wafer at a time to wire saws slicing several feet of ingot in one cut. This has resulted in an increase in the number of usable wafers obtained per centimeter of ingot from 8–12 to 16–22, with typical wafer thicknesses at or slightly less than 300 μm and a further decreasing trend. In principle, wire sawing allows the cutting of much thinner wafers. However, it remains a challenge to process these thinner wafers and avoid excessive breakage during cell processing and module assembly.

Smaller, but not insignificant, portions of crystalline Si cells are based on wafers grown as silicon ribbon or as thick silicon films deposited on conducting ceramic substrates. These technologies avoid the wafering steps, and the performance of Si ribbon cells has reached levels comparable to those of wafer Si-based technologies. Of interest is that these various approaches still continue to be developed and used in parallel. Approximately 15 years ago, when today's manufacturers or their predecessors began to develop the different processes, each had hoped that their preferred method of making the Si substrate would result in a unique cost/performance advantage. It has not become evident which of these approaches is the most cost-effective. The reason for this may lie in the fact that the quality of the wafer plays a lesser role than originally thought. Instead, it became much more important to learn how to process each wafer into a high-efficiency solar cell.

2.2 Silicon Solar Cell Operation

For a crystalline Si solar cell, the complicated structure of Fig. 1 is often approximated in terms of a bulk (wafer) lifetime and a surface recombination velocity. Another frequently used parameter is the (zero-field) diffusion length as a measure for collecting carriers generated in the diffusion region of the absorber. Such characterization omits a crucial parameter: the cell voltage. For diffusion-processed cells,

V_{OC} depends critically on the boron doping level of the wafer base. Typical doping levels for optimum cell performance are 0.2–2 Ωcm. It is typically found that higher doping levels of the wafer or absorber increase V_{OC}, whereas they decrease J_{SC}. For the highest efficiencies, a compromise must be made.

Although most commercial cells are manufactured by diffusing emitters and collectors, higher performing champion Si solar cells often have a different cell structure. Point-contact or buried laser grooved-contact cells (referred to as passivated emitter, rear locally diffused cells) have reached efficiency levels of 24%. Because the localized contacts require large diffusion lengths to collect carriers over even greater distances, these high-efficiency cells can only be made with single-crystal "float zone" or Czochralski wafers. Many analyses have correlated V_{OC} values with the hole density in the absorber and have ascribed lower than optimum V_{OC} values to recombination losses in the absorber. With the realization of the so-called HIT (heterojunctions with intrinsic thin layers) solar cell developed by Sanyo Corporation, this interpretation becomes challenging. The HIT cell, a heterojunction Si cell, uses amorphous silicon emitter, collector, and buffer layers and n-type Czochralski wafers as the absorber. One practical advantage of this cell type is a temperature coefficient approximately half that in typical diffused cells, giving these cells an advantage at typical operating temperatures under full sunlight. The latter observations suggest that cell voltages and their temperature coefficients are predominantly determined by the details of junction formation and not by losses in the wafer or absorber.

The use of buffer layers in crystalline Si HIT cells supports the universality of a voltage-enhancing buffer layer concept. In diffusion-processed and screen-printed cells, fire-through nitride passivation layers are used. A nitride layer between the emitter and the metal grid contacts, frequently referred to as a passivation layer because of the high hydrogen contents of the nitride layers, can also be seen as a buffer concept. Finally, the sensitivity of V_{OC} to the boron base doping level in diffusion-processed cells can also be understood in terms of a buffer layer (in this instance, a low-mobility compensated layer that must be formed while the phosphorus emitter is diffused into the boron-doped p-type base).

2.3 Defect Engineering

Most crystalline Si cell manufacturers produce their cells by phosphorus diffusion for the emitter and use aluminum alloy for the collector. It has become well accepted that the Si material can be upgraded during the cell-processing step. Figure 5 shows a pyramid schematic of how fundamental starting-material properties, substrate preparation, and processing during solar cell fabrication all strongly affect the resulting solar cell efficiency. Upgrading during cell processing often occurs automatically as the cell emitter, collector, and passivation and antireflection layers are formed, and it is usually referred to as phosphorous and/or aluminum gettering. Optimizing these gettering processes was critical in achieving today's high commercial cell efficiency levels (typically 12 to >15%). Some manufacturers use additional process steps such as hydrogenation treatments for defect passivation. Such upgrades are commonly believed to lower the defect densities or the effectiveness of defects.

It is difficult *a priori* to distinguish whether the cell processes only affected the defect properties in the wafer (absorber region) or whether manipulation of the junction and buffer regions affected the space-charge region of the cell, improving the carrier collection from an unaltered absorber. The interactive nature of defects and impurities is well accepted among researchers. The lifetimes or diffusion lengths not only depend on the impurity levels (e.g., the concentration of a lifetime-killing metal impurity) but also depend, in a complicated way, on the presence of other inactive impurities, such as oxygen or carbon. Impurities interact with point or extended defects such as dislocations and structural defects in polycrystalline materials. The effectiveness of impurities in promoting recombination losses is strongly determined by

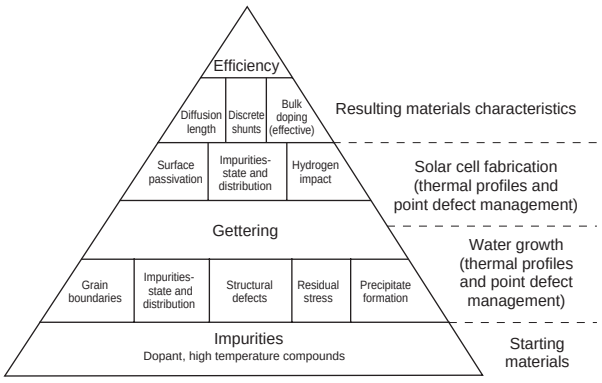


FIGURE 5 Pyramid schematic of how fundamental starting-material properties, substrate preparation, and processing during solar cell fabrication all strongly and interactively affect the resulting solar cell efficiency. Reproduced with permission from Rand *et al.* (2002).

gettering and passivation effects. In summary, it is difficult to describe solar cell performance simply in terms of defect and impurity concentrations in the Si wafer or sheet material alone.

3. THIN-FILM SILICON

3.1 Amorphous Silicon

Amorphous silicon solar cells are commercially available and can be produced on a variety of substrates ranging from glass to flexible thin foils. Cells are built in p-i-n or n-i-p configurations, where p and n represent thin doped (amorphous or nanocrystalline) layers, and the absorber layer is an intrinsic undoped layer. Cells are deliberately made thin (<300 nm thick) such that the space-charge region extends through the entire absorber (i-layer). The very small thickness is possible because amorphous silicon has stronger light absorption than crystalline Si (a direct bandgap) and light-trapping schemes have been developed to enhance the optical absorption.

The semiconducting properties of this material were discovered when films were prepared by plasma-enhanced vapor deposition of silane (SiH_4) gas. It was established that the optical and electrical properties are dominated by approximately 10 atomic-% (at.%) hydrogen in the film. It was also discovered that the material could be doped n-type (usually with phosphorus) and p-type (usually with boron), making this material much like a crystalline semiconductor. Until then, there was debate about whether amorphous semiconductors could, in principle, be doped. Some argued that the flexibility of the amorphous network would prevent three-valent acceptor or five-valent donor atoms from being incorporated into the four-valent Si network, which is necessary to produce additional conduction electrons or holes. Doping changes the resistivity of amorphous silicon by more than seven and six orders of magnitude for n- and p-type doping, respectively.

Unhydrogenated amorphous silicon films can also be prepared, for example, by evaporating silicon onto a substrate below 400°C. However, such films are very defective and show high conductivity. It is generally accepted that the quality of such films is low because the very high density ($>10^{20}/cm^3$) of silicon "dangling bonds" acts as a defect. With hydrogenation, their density is orders of magnitude lower (10^{15}–$10^{17}\,cm^{-3}$). Device-grade intrinsic hydrogenated amorphous silicon (a-Si:H) has the following properties: bandgap, 1.6–1.8 eV; dark conductivity, 10^{-9}–$10^{-12}\,\Omega$cm; and photoconductivity under 1 sun intensity light of 10^{-5}–$10^{-6}\,\Omega$cm. The conductivity activation energy ranges from 0.65 to 0.9 eV.

Amorphous silicon degrades under illumination. In a solar cell, all three parameters—V_{OC}, FF, and J_{SC}—can be affected and usually decrease upon illumination. In intrinsic layers, the degradation affects the photoconductivity (which decreases), and the midgap density D_o, which increases. The properties of the light-induced D_o defect (e.g., its paramagnetic spin resonance signal) are virtually indistinguishable from those of the native D_o defect found in the annealed state in much lower concentrations. Fortunately, the degradation appears to be self-limiting and stabilizes. The exact rates of degradation depend on the material properties and the conditions (temperature and light intensity) used for light soaking. Annealing (for 5–150 min at 150–180°C) essentially restores the light-soaked material (or cell) to its original state. With 1 sun intense light, most of the degradation will occur within several tens to hundreds of hours. This effect is referred to as the Staebler–Wronski mechanism.

It is clear that the effect, although influenced by the presence of impurities, is inherent to the amorphous silicon– (or germanium–) hydrogen alloy system. There is no sure mitigation scheme to avoid degradation in high-quality a-Si:H, but the details of materials preparation can noticeably influence the amount of material property or solar cell degradation observed. Because of the intrinsic nature of the effect, it is common to examine stabilized materials and solar cell performance or efficiency.

In the past, the near-midgap dangling bond defect was used as a strong quality indicator. Specifically, models were made that correlated such defect density with solar cell fill factors and the photoconductive properties of the layer. However, careful quantitative work later showed that the solar cell fill factor or the photoconductivity (PC) of the intrinsic layer are not controlled by a simple, single, straightforward mechanism. In the past, it was assumed to be the generation of D_o defects. It has been confirmed that the kinetics with which observables such as D_o, PC, or cell FF change upon light soaking or annealing show differences, and there is a lack of quantitative correlation between D_o generation and PC and FF losses but a better correlation between PC and FF. Such findings may explain why materials with lower stabilized D_o densities have not led to improved stabilized solar cell performance.

Also, degradation and annealing behavior in most instances can be classified into two mechanisms: fast (low-temperature annealable) and slow (below 80°C nonannealable) mechanisms. Any comprehensive Staebler–Wronski model must account for both mechanisms, which appear to be linked. The effects of the two mechanisms on degradation and annealing are large enough to affect the power output of a-Si:H-based modules by $\pm 10\%$. Interest in mitigating Staebler–Wronski degradation remains high because if it were possible to maintain today's solar cell performance in the higher annealed state, 16% efficient multijunction cells and 10% efficient commercial modules would likely be a reality. However, all data suggest that the degradation is intrinsic to a-Si:H and related alloys.

The prevailing mechanisms suggested in the literature implicate weak Si–Si bonds and weakly bonded hydrogen in the material as the cause for the instability. Experimentally, the instability appears not to depend systematically on hydrogen content of the material or on how tightly hydrogen is bonded. In fact, some of the more stable (deuterated and high hydrogen dilution, lower substrate temperature) solar cell recipes were achieved using materials that contain larger amounts of loosely bonded hydrogen.

Amorphous silicon solar cell technology has realized the first low-cost multijunction bandgap commercial solar cells. To make the cells more sensitive to red light, a-Si:H is alloyed with germanium (Ge). Especially for Ge contents >50 at.% (corresponding to a bandgap of <1.45 eV), the solar cell performance decreases significantly. A plausible explanation for this phenomenon is the decreased carrier collection due to lower carrier mobilities or higher recombination losses. This decreased carrier collection is mitigated by making the narrow bandgap cells very thin (150 nm or less) and by grading the Ge content in the intrinsic layers, such that the highest Ge content is found near the p-i interface of the junction. Use of Ge alloying for multi-bandgap cells has increased the stabilized device performance (e.g., on the order of 1 absolute efficiency point when an a-Si:H/a-Si:H cell and a-Si:H/a-$Si_{0.5}Ge_{0.5}$:H cell are compared). However, Ge alloying has enhanced the cell performance less than expected from the reduction in bandgap because of deterioration of the electronic properties with Ge alloying and significant light-induced degradation in such alloys, which can be offset only by using a very thin, Ge-graded, narrow bandgap cell structure. Attempts were made to use carbon alloys (a-SiC:H) to widen the bandgap of the top cell of a triple-junction device to increase the voltage of this component cell. Such efforts have been abandoned because a-SiC:H cells showed aggravated light-induced degradation, and it was not possible to increase the voltage of these cells beyond the best values achievable with optimized a-Si:H intrinsic layers (1.0 V). There is no consensus as to which microscopic mechanisms cause the electronic deterioration upon alloying with Ge, C, or, even more significant, other elements.

3.2 Crystalline Silicon Films

There are two approaches for using thin silicon films for photovoltaic applications. When amorphous silicon deposition techniques are used with excessive hydrogen dilution, the resulting films become nanocrystalline (nc-Si). The term microcrystalline has also been used for such films; however, grain sizes are always $\ll 1\,\mu$m. Typical crystallite sizes are on the order of tens of nanometers, and to have reasonable electronic properties, a few atomic percent of hydrogen must be incorporated in such films. Solar cells with efficiencies of approximately 7% have been demonstrated using low-temperature deposition approaches.

Conventional thinking caused researchers to search for higher solar cell efficiency using Si films with much larger grain sizes. However, experimental work established a lack of improvement of cell efficiency with grain size in the range of 100 nm to 1 mm. To achieve grain sizes >1 mm, either high-temperature recrystallization (remelting) or epitaxial growth on a larger grain silicon substrate are required. Such methods require a substrate that can withstand the high temperatures involved and that is chemically inert so as not to release impurities into the silicon film. Such requirements may eliminate commercial glasses as a substrate. In the laboratory, graphite and ceramic substrates have been used with some success.

First, the low-temperature approaches are discussed. The solar cell efficiencies achieved appear too low to justify the development of a single-junction nanocrystalline solar cell technology. The main application is replacing the bottom cell in a-Si multijunction cells. To function as a bottom cell, the nc-Si cell must generate $>24\,\text{mA/cm}^2$ and have a sufficiently high red response to generate $>8\,\text{mA/cm}^2$ in a triple-junction stack cell or $>12\,\text{mA/cm}^2$ for a double-junction device. With a high-quality back reflector for optical enhancements, current densities

of approximately 24 mA/cm^2 were demonstrated with nc-Si absorbers $\leq 2\,\mu$m thick.

How nanocrystalline cells operate is very poorly understood. Preparation of nc-Si films and cells is extremely sensitive to deposition conditions. Physical properties that may change are the crystalline orientation, luminescence properties, and conductivity of the films. When such changes occur, solar cell properties also change, often dramatically. However, it is difficult to pinpoint how the properties of the films determine cell performance. Rather, whenever the absorber properties change, it is necessary to readjust the junctions (p, n, and buffer layers) used in the cell. It appears from the optimization of nc-Si cells that a regime is often encountered in which higher J_{SC} values are obtained by sacrificing V_{OC} or vice versa. Such a trade-off between voltage and current is not uncommon for both thin-film and wafer-based crystalline silicon. The highest voltages for the nc-Si cell, in conjunction with the required high J_{SC} values, remain to be determined. Better optimization and better understanding of the mechanism leading to such a trade-off are both required.

Cells spanning the amorphous to nanocrystalline transition (adding a gradually increasing nanocrystalline phase to the intrinsic layer) have been prepared. Cells with a large fraction of amorphous phase in their i-layer show large voltages but also light-induced degradation. Light soaking actually enhances V_{OC} but decreases FF. With such unstable i-layers, the high J_{SC} values required for multijunctions could not be demonstrated. It is important to note that even the stable cell recipe with the largest grain size and best-ordered i-layers did not result in the highest efficiency.

Two additional universal findings about nc-Si cells should be mentioned. First, as for a-Si:H alloys, the best cell results are obtained when the i-layer deposition rate is slow (<0.5 nm/s). However, some deposition methods [such as hot-wire chemical vapor deposition (CVD)] have shown that different film structures and modest cell efficiencies (routinely, approximately 5%) can be obtained at higher deposition rates (1–2 nm/s). Second, it has been observed that some nc-Si films are porous and therefore subject to postdeposition oxidation. After oxidation, solar cell performance degrades severely. Such films must be avoided because it has been reported that oxidation occurs even when such devices are capped, for example, by a p-layer in a solar cell.

Of concern with regard to nc-Si is that in recent years cell efficiency records have been somewhat stagnant (approximately 10%). Given the current performance, it appears that there is opportunity to incrementally improve a-Si:H-based multijunctions by using the nc-Si cell as a bottom cell. Given the large variety of nc-Si films that can be produced, it may well be possible that cell recipes, at greater deposition rates, can be found that result in a breakthrough for even higher cell efficiency.

Next, the higher temperature, large-grain crystalline Si films are discussed. Two approaches have resulted in larger grain sizes. One approach to thin-film solar cells relies on melt recrystallizing a CVD-deposited, $\sim$100-μm-thick growth layer on a ceramic or other substrate. Then, active solar cell layers are deposited epitaxially by CVD. The active Si layers used are a 5-μm-thick back-surface field and contact layer (boron-doped, $\sim 0.1\,\Omega$ cm) and a very lightly boron-doped, 30- to 35-μm-thick absorber layer. The cell emitter is formed by a P-diffusion process, resulting in a cell with the following parameters: $\eta = 9.1\%$, $V_{OC} = 0.54$ V, FF $= 0.73$, and $J_{SC} = 23.1$ mA/cm^2. Because the ceramic substrate is insulating, monolithic interconnection of cells can be accomplished and has been demonstrated. Interestingly, the current generated and collected in these absorbers is similar to the current densities measured in the 1.5- to 2-μm-thick nc-Si absorbers. This may be because in thin, optically enhanced Si film cells, carrier collection is quite efficient, whereas in thicker films carrier collection remains a challenge for carriers generated some distance (more than a few micrometers) from the junction.

Other deposition approaches, such as using an iodine vapor transport reaction, provide a process for growing large-grain films in the 750–950°C substrate temperature range. Cell efficiencies of these larger grain silicon films have not met expectations. One reason may be that grain size is not as much a driver of material quality and cell performance as has been hitherto assumed. The behavior of the nanocrystalline cells (in which smaller grain size leads to higher cell performance) suggests that there may be a genuinely beneficial aspect of polycrystallinity. For intermediate grain size silicon films, it may be that this effect is lost, and the carrier collection is still too impeded compared to that of single- or large-grain (>1 mm) crystalline Si wafers.

4. POLYCRYSTALLINE THIN FILMS

Since approximately 1990, solar cell efficiencies of $>13\%$ have been obtained with copper indium diselenide ($CuInSe_2$ or CIS) and cadmium telluride

(CdTe) polycrystalline thin-film semiconductors. These materials have few other applications. They are attractive as solar cell absorbers because they perform well. Cell efficiencies on the order of 19% for CIS and 16% for CdTe have been demonstrated. These materials were mostly developed using heuristic solar cell optimization schemes, with a lagging understanding of the fundamental physical properties of the materials and cells.

4.1 Copper Indium Diselenide

CIS was developed after it was realized that CdS/CuS solar cells developed earlier, despite promising cell efficiency (10% demonstrated), were not viable because of their propensity to become shunted while operating. This shunting was related to copper migration. On the other hand, the ternary $CuInSe_2$ and quaternary $CuIn_{1-x}Ga_xSe_2$ compounds (x is typically 0.2–0.3 for high-efficiency cells) do not suffer from such intrinsic degradation. Early on, it was realized that to obtain high solar cell performance, the films had to be nonstoichiometric. Specifically, for good cell performance, the Cu/(In + Ga) atomic ratio must be between 0.86 and 0.96. The Se content of the film must be slightly higher than 2. Films for solar cells are deposited by a variety of methods, including coevaporation from metal or metal–selenide precursors. It is beneficial to have some excess selenium available during deposition. However, it is also possible to react mixed-phase metal or metal–selenide precursors into a single-phase (or seemingly single-phase) $CuInGaSe_2$ layer in a selenization reaction in the presence of Se vapor or H_2Se gas. Another approach is to produce absorbers from nanoparticle precursors. Many deposition recipes for high-efficiency cells call for ramping both the substrate temperature and the ratios of the metal or metal–selenide precursors during film growth such that the Cu deficiency is greatest near the surface of the film. The addition of Ga has made it easier to achieve higher efficiency cells, both because the bandgap of the absorber increases from approximately 1.0 to 1.15 eV and because Ga affects the electronic properties in a way that is favorable to cell performance, especially the V_{OC} parameter. For the best cell results, Ga incorporation in the absorber should be uniform or in a "notch" distribution (higher concentrations near the front and back of the absorber). The best solar cell results are obtained for films deposited or selenized at temperatures higher than 500°C. Unfortunately, some deposition approaches lead to an undesirable high accumulation of Ga mostly near the back surface of the absorber. In these instances, sulfur has been used to form $Cu(In,Ga)Se_{2-x}S_x$ pentenary compounds to achieve the bandgap widening and desirable electronic modification near the front surface of the absorber. However, experimental results suggest that S alloying leads to somewhat inferior cell performance compared to optimized Ga alloying.

CIS solar cells are built up on a molybdenum (Mo)-coated soda-lime glass substrate. The use of Mo metallization is important to achieve high cell performance, but the reason for this is poorly understood. It is well established that the addition of some sodium during film growth is quite beneficial; Na had been incorporated inadvertently when soda-lime glass substrates were used. Sodium can be added during absorber growth to achieve high performance on Na-free glasses or on Mo-coated metal or high-temperature resistant plastic foils.

CIS devices use a ZnO/CdS heterojunction emitter layer. The best and most consistent solar cell results are obtained when a thin CdS layer is deposited using a wet chemical bath deposition (CBD) process. The reason for this is debated in the literature, and no consensus has been reached. Alternatives to CdS (such as resistive oxides or ZnS) have been explored with some success but have not achieved the same champion cell efficiencies or processing robustness as CBD CdS. The ZnO contact is often prepared in the form of a graded layer or bilayer, where the part of the film in contact with the CdS is deliberately made more resistive. It can be argued that in a CIS solar cell the optimum junction profile is affected by the resistive ZnO layer, the CdS or alternative buffer layer, and the surface portion of the $Cu(In,Ga)Se_2$ absorber. This makes the cell optimization process very interdependent; that is, if one of the layers in the cell is altered, all other layers (their properties and thicknesses) must be reoptimized. In other words, there is no unique optimum for the properties of each layer. Conventionally, it is assumed that cells fabricated from more perfect layers result in higher cell efficiency. For CIS and CdTe solar cells, cells deposited as multicrystalline thin-film layers on low-cost substrates have achieved greater device performance than solar cells made on (bulk or epitaxial) single-crystal versions of these materials.

4.2 CdTe Solar Cells

CdTe was developed after it was realized that a number of deposition approaches, such as

evaporation, sublimation, sputtering, or electrodeposition, resulted in promising devices (10% demonstrated). Early on, researchers realized that to obtain high solar cell performance, the films had to be heat treated after exposing them to a $CdCl_2$ solution or vapor. Initially, growth of the crystallite grain size and Cl incorporation into the film were observed. However, it was later found that the $CdCl_2$ heat treatment is also essential for cells with CdTe absorber layers grown at high temperature having even larger grain sizes. Treatments are carried out in the range of 360–500°C. CdTe always grows as a stoichiometric compound that is much less toxic than Cd metal. There is debate as to whether oxygen incorporation (doping) of the CdTe layer is necessary or beneficial.

CdTe cells usually use a SnO_2/buffer/CdS heterojunction emitter layer. It is important to keep the thickness of the CdS as thin as possible to avoid optical absorption losses from this layer. However, reducing the thickness too much often leads to reduced V_{OC} and FF. The buffer layers are often undoped SnO_2 layers. Some alternative buffer layers, such as zinc stannate or CdS:O (the latter prepared as an amorphous layer that crystallizes during the cell anneal step), may ultimately lead to higher cell performance. It is generally accepted that some intermixing of the CdS and CdTe takes place and that such intermixing may positively affect cell performance.

The rear-contact preparation for CdTe is more an art than a science. Unsuitable contacts result in very poor device performance. At first, for laboratory cells, a bilayer of copper and gold was used with good success. All contact applications generally benefit from acid etching the CdTe device. This has been shown to leave a Te-enriched surface. The Cu/Au contacts, in general, show poor long-term stability and are not suitable for commercial product. Graphite pastes were found to be another good contacting scheme; graphite paste ("dac") that is typically doped with Cu has been applied to the etched surface. Usually, conductive silver ink is applied as a secondary contact to the graphite layer. This type of contact performs well but is less desirable as a contacting scheme for large-area commercial modules. Because contact formation was difficult due to the nearly intrinsic nature of the CdTe layer, Cu-doped ZnTe layers have also been successfully developed as a back contact. Although they are thought of as completing the junction, it has been reported that the choice of metal used to contact this ZnTe layer can affect the solar cell properties. Commercial manufacturers have developed proprietary thin-film contacting schemes.

It is remarkable that the performance of a CdTe solar cell is very strongly affected and can be dominated by either junction (i.e., the emitter or the collector). Early commercial CdTe modules have shown some change (sometimes an initial improvement, but also degradation) in performance. Laboratory cells are typically stressed at 60–90°C and show some loss of power (0–30%) after such stressing, unless the cells are processed in a very narrow, empirically established process window. The stability is affected by the details of both the front and back contact as well as by the $CdCl_2$ anneal treatment of the CdTe. The rapid degradation of cells using Cu/Au contacts suggests Cu diffusion or migration as a major cause. Further studies have revealed, however, that although too much Cu applied with the back contact leads to degradation, variations for modest amounts of Cu (e.g., in Cu-doped graphite or Cu-doped ZnTe contacts) can increase or decrease cell stability. Cells made without any Cu deliberately added were also found to degrade. Thus, minimizing or eliminating the Cu doping of the back contact is not a sure way to eliminate the degradation. It appears that cells and modules with sufficient commercial stability ($<$10–20% degradation over 20 years of operation) can be manufactured, but each manufacturing process must be tested to eliminate processes that result in too much degradation.

5. CONCENTRATOR SCHEMES AND III–V SOLAR CELLS

For large centralized photovoltaic arrays, it has been proposed to minimize the use of semiconductor material by focusing the sunlight through lens or mirror systems onto a solar cell mounted at the focal point of the concentrator. Such systems must be "tracked" (i.e., rotated around two axes to remain in focus with the sun). This approach opens up the possibility of using very high-efficiency solar cells that would be too costly to incorporate into flat-plate arrays for terrestrial power production. Because of the complexity and maintenance requirements of mechanical trackers, this approach is only viable for large systems in which hundreds of kilowatts or perhaps megawatts of power are required. An advantage of this scheme is that the efficiency of high-efficiency solar cells increases further under concentrated sunlight. A drawback is that such

systems have some optical and system losses and produce power only from direct sunlight (i.e., when the sun is directly visible), whereas flat-plate systems still convert the 5–50% of sunlight available during cloudy conditions when the sun is partially or totally obscured.

The solar cells for such concentrator systems are either high-efficiency single-crystal silicon solar cells (20–27% conversion efficiency under concentration) or GaAs-based multijunction solar cells that have demonstrated efficiencies well above 30% under concentration. Champion Si cells and GaAs cells have attained close to theoretically possible efficiency values, and the challenge is to incorporate such cells into even higher efficiency multijunction solar cells.

The GaAs or III–V semiconductor community has developed near-ideal high-efficiency multijunction solar cells. Typically, these cells are built as homojunctions. For high performance, combinations of a lower doped n-type homoemitter and a higher doped (AlGaAs) heteropassivation layer are used. Without concentration, these high-efficiency cells are also used commercially for space cells, for which their high efficiency and radiation resistance outweigh their much higher cost. The cells are grown in thin-film form using heteroepitaxy on a Ge substrate. A popular cell structure is the GaInP/GaAs/Ge triple-junction cell, providing the appropriate lattice-constant match to Ge and GaAs for successful heteroepitaxy growth. In principle, although for the III–V semiconductors there are a wide variety of compounds providing a very large range of E_g (while still being lattice matched to Ge and GaAs), some materials clearly work better than others for solar cells. There has been no success in achieving high cell performance for a 1-eV bandgap cell using lattice-matched $Ga_{1-x}In_xAs_{1-y}N_y$ alloys. Even for this near-ideal and well-characterized semiconductor system, it is unknown which physical/chemical absorber properties determine high and low solar cell performance.

6. EMERGING SCHEMES

6.1 Dye-Sensitized

A solar cell efficiency greater than 10% has been reported by Michael Grätzel *et al.*, who invented and optimized a dye-sensitized solar cell at the Swiss Federal Institute of Technology. The dye-sensitized cell is an electrochemical cell in which an electrolyte, rather than a conductor or semiconductor, is used to make one of the junctions of a solar cell. The absorber layer consists of a deliberately porous SnO_2/titanium dioxide (TiO_2) coating, for which the surface of the TiO_2 is on the order of 1000-fold enhanced over the surface of a smooth film. This $\sim$100-nm-thick TiO_2 is coated with a very thin ($\sim$1-nm-thick) dye. The best performance has been achieved with ruthenium (Ru)-containing dyes, such as Ru bipyridyl or Ru terpyridine complexes that absorb near-infrared light. The contact is made with an iodine redox electrolyte and a second piece of SnO_2-coated glass as the electrical contact. Note that this type of cell is not actually wet. The electrolyte simply has to wet the dye, requiring only drops of electrolyte per 100-cm^2 device area. Nevertheless, when multiple cells are made on one substrate with monolithic interconnection, this device must be tightly sealed not only around the perimeter but also between adjacent cells. In an attempt to simplify fabrication, solid or gel electrolytes have been investigated but have resulted in inferior cell performance compared to cells with liquid electrolytes.

The prevailing physical description of cell operation assumes that in this device the photogenerated electrons and holes are already separated upon generation in the physical phase boundary of the dye. This creates a photo-induced potential gradient that drives electrons and holes in opposite directions.

6.2 Other Schemes

Although semiconducting organic polymers have been known for some time, only recently has there been a focus on using these materials for preparing solar cells. Their physics and limitations are poorly understood. The junction is made from an electron conductor (e.g., perylene benzimidazole) and a hole conductor (e.g., poly-3-hexylthiophene). It is generally believed that in these materials the absorbed light creates excitons that are transported to the planar interface at which the carriers are separated. The efficiency levels achieved (up to 4% in a few instances but typically lower) suggest that further development is required before these will become candidates for terrestrial power application. Interest is high because in recent years there have been significant advances with regard to organic light-emitting diodes in terms of luminosity.

6.3 Theoretical Concepts

Several schemes have been proposed to overcome the SQ limit discussed previously. Aside from multijunction solar cells, which have been produced and

allowed moderate efficiency gains, schemes have also been proposed in which suitable arrangements in single-junction devices may lead to cell efficiencies exceeding 80%. One proposal suggests that sub-bandgap light can be absorbed via defect states (impurity solar cells). Other schemes propose using the energy that is greater than the corresponding bandgap of the absorber to create additional electron hole pairs (hot carrier cells). These schemes rely on separating the generation and collection processes of carriers, and they assume lossless carrier collection. Experimentally, attempts to realize such schemes have been unable to verify the expected significant enhancement of the short-circuit current density or spectral response.

7. FUTURE POTENTIAL (PHYSICAL LIMITATIONS OF MATERIALS)

There is good empirical knowledge that some materials work better than others for solar cells. In the search for new materials, materials with the desired bandgap, electronic properties allowing carrier separation and collection, and electronic properties that allow robust junction-fabrication schemes must be found. Researchers with a background in GaAs-based semiconductors have also stressed the importance of matching the lattice constants to achieve perfect epitaxial growth.

Achieving the desired bandgap (e.g., by alloying different materials) is possible most of the time. Cell efficiencies hinge on the electronic properties, which are often not conducive for achieving high cell performance. Often, it is difficult to determine what causes low cell efficiency. Structural perfection and lattice matching are requirements for GaAs-based solar cells. However, in other materials systems disordered polycrystalline semiconductor films lead to solar cell performance equal to or exceeding that which can be achieved from single crystalline versions of these materials.

Robust junction-formation processes remain a challenge, although nature has provided some schemes for either homo- or heterojunctions that are quite reproducible, controllable, and manufacturable. It is becoming increasingly accepted that solar cell optimization is a very interdependent process; that is, if one layer of the cell is altered, leading to lower solar cell performance, this can be fixed by reoptimizing the processing of other layers in the cell. In other words, cell optimization cannot rely on combining individually optimized layers to achieve high performance. Rather, the correct combination of absorber, junction, and contacting layers is required to achieve high solar cell performance.

SEE ALSO THE FOLLOWING ARTICLES

Materials for Solar Energy • Photovoltaic Conversion: Space Applications • Photovoltaic Energy: Stand-Alone and Grid-Connected Systems • Photovoltaics, Environmental Impact of

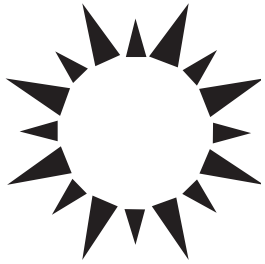
Further Reading

Luque, A., and Hegedus, S. (2002). "Handbook of Photovoltaic Science and Engineering." Wiley, New York.

Rand, J., Rozgonyi, G., Lu, J., and Reedy, R. (2002). Characterization of silicon-film sheet material. Conference Record of the 29th IEEE Photovoltaic Specialists Conference, New Orleans, May 19–24, pp. 98–101.

Shockley, W., Queisser, H. J. (1961). Detailed balance limit of efficiency of p–n junction solar cells. *J. Appl. Phys.* **32**, 510–519.

Photovoltaics, Environmental Impact of



VASILIS M. FTHENAKIS
Brookhaven National Laboratory
Upton, New York, United States
ALEC O. BULAWKA
U.S. Department of Energy
Washington, D.C., United States

1. **Overview of Hazards in Photovoltaic Manufacture**
2. **Crystalline Silicon Solar Cells**
3. **Amorphous Silicon Solar Cells**
4. **Cadmium Telluride Solar Cells**
5. **Copper Indium Diselenide Solar Cells**
6. **Gallium Arsenide High-Efficiency Solar Cells**
7. **Operation of PV Modules**
8. **Decommissioning PV Modules**
9. **Environmental Benefits of Photovoltaics Operation**
10. **Conclusion**

Glossary

emergency response planning guideline-2 (ERPG-2) The concentration of a hazardous material below which nearly all people could be exposed for up to 1 h without irreversible or other serious health effects or symptoms that would impair their ability to take protective action.

immediately dangerous to life or health concentration (IDLH) The maximum concentration of a hazardous material from which one could escape within 30 min without any symptoms impairing escape or any irreversible health effects.

threshold limit value—ceiling The concentration of a hazardous material that should not be exceeded during any part of the working exposure.

threshold limit value, short-term exposure level (STEL) The maximum concentration of a hazardous material to which workers can be exposed for up to 15 min, provided not more than four excursions per day are permitted with at least 60 min between exposures, and provided that the daily permissible exposure limit is also not exceeded.

threshold limit value, time-weighted average (TLV-TWA) The time-weighted average threshold concentration of a hazardous material above which workers must not be exposed during work shifts (8 h/day, 40 h/week).

Photovoltaic (PV) technologies have distinct environmental advantages for generating electricity over conventional power technologies. The operation of PV systems does not produce any noise, toxic emissions, or greenhouse gases; regardless of the composition or technology used, PV generation of electricity is a zero-emissions process. Photovoltaic modules are durable and do not produce emissions even during extreme conditions (accelerated aging with low Earth orbit radiation and thermal cycles from 80 to -80°C). Photovoltaic energy can help meet the growing worldwide demand for electricity and can do so without incurring the high economic and environmental costs of burning fossil fuels and installing power lines. Compared to burning coal, every gigawatt-hour of electricity generated by PVs would prevent the release of approximately 1000 tons of carbon dioxide, 10 tons of sulfur dioxide, 6 tons of nitrogen oxides, and 0.4 tons of particulates.

However, as is the case with any energy source or product, there are also environmental, health, and safety (EHS) hazards associated with the manufacture of solar cells. The PV industry uses toxic and flammable substances, although in smaller quantities than do most other industries, and using hazardous chemicals can entail occupational and environmental hazards. Addressing EHS concerns is the focus of numerous studies of the National Photovoltaic EHS Assistance Center at Brookhaven National Laboratory, which operates under the auspices of the U.S. Department of Energy. These studies are proactive and ongoing as new technologies, materials, processes,

 © 2004 Elsevier Inc. All rights reserved.

and increased production levels are encountered. The activities of the center are coordinated with the U.S. PV industry, which is very sensitive and responsive to EHS issues. This article provides a concise summary of EHS issues pertaining to the manufacture of crystalline silicon, amorphous silicon, copper indium diselenide, copper indium gallium diselenide, gallium arsenide, and cadmium telluride solar cells, all of which are commercially available.

1. OVERVIEW OF HAZARDS IN PHOTOVOLTAIC MANUFACTURE

A variety of materials are used in manufacturing photovoltaic (PV) cells, including materials widely used in commercial applications (e.g., aluminum and glass), chemicals that have seen extensive use in the semiconductor industry (e.g., silane and hydrofluoric acid), and some chemicals that are unique to the PV industry [e.g., copper indium selenide (CIS) and cadmium telluride (CdTe)]. For such new materials, the U.S. Department of Energy has collaborated with the National Institutes of Health to study their potential health hazards. Table I presents the main hazards associated with specific technologies, and Table II shows that these hazards differ for different technologies and deposition processes. Potential risks are related to the materials' toxicity, corrosivity, flammability, and explosiveness.

2. CRYSTALLINE SILICON SOLAR CELLS

2.1 Occupational Health Issues

In manufacturing wafer-based crystalline silicon (*x*-Si) solar cells, there are some occupational health issues, including potential chemical burns and inhalation of fumes from the hydrofluoric acid, nitric acid (e.g., HNO_3), and alkalis (e.g., NaOH) that are used for cleaning the wafers, removing dopant oxides, and cleaning the reactor. Dopant gases and vapors [e.g., phosphorus oxychloride ($POCl_3$) and diborane B_2H_6] are also hazardous if inhaled. $POCl_3$ is a liquid, but in a deposition chamber it can generate toxic P_2O_5 and Cl_2 gaseous effluents. Inhalation hazards are controlled with properly designed ventilation systems in the process stations. Other occupational hazards are related to the flammability of silane (SiH_4) and its by-products used in silicon nitride deposition.

TABLE I

Major Hazards in PV Manufacturing

Module type	Potential hazards
x-Si	HF acid burns, SiH_4 fires/explosions, Pb solder/module disposal
a-Si	SiH_4 fires/explosions
CdTe	Cd toxicity, carcinogenicity, module disposal
CIS, CGS	H_2Se toxicity, module disposal
GaAs	AsH_3 toxicity, As carcinogenicity, H_2 flammability, module disposal

2.2 Public Health and Environmental Issues

No public health issues were readily identifiable with this technology. The environmental issues are related to generating liquid and solid wastes during the slicing, cleaning, and etching of wafers and during the processing and assembling of solar cells.

The *x*-Si PV industry has initiated programs of waste minimization and hence studies environmentally friendlier alternatives for solders, slurries, and solvents. Success was reported at both laboratory and manufacturing scales in reducing the caustic waste generated by etching. Other efforts for waste minimization include recycling stainless-steel cutting wires, recovering the SiC in the slurry, and in-house neutralization of acid and alkali solutions.

Finally, the content of Pb in solder in many of today's modules creates concerns about the disposal of modules at decommissioning. One U.S. company is using Pb-free soldering and has offered assistance to the rest of the industry for adopting this option. It is likely that the *x*-Si industry will gradually move away from Pb–Sn solder.

3. AMORPHOUS SILICON SOLAR CELLS

Amorphous silicon (a-Si), CdTe, CIS, and gallium arsenide (GaAs) are thin-film technologies that use approximately 1/100 of the costly PV material used on *x*-Si.

3.1 Occupational Safety Issues

The principal safety hazard of this technology involves the use of SiH_4 gas, which is extremely pyrophoric. The lower limit for its spontaneous

TABLE II

Hazardous Materials Used in PV Manufacturing[a]

Material	Source	TLV-TWA[b]	STEL[c]	IDLH[d]	ERPG-2[e]	Critical effects
Arsine	GaAs CVD	0.05 ppm	—	3 ppm	0.5 ppm	Blood, kidney
Arsenic compounds	GaAs	0.01 mg/m^3	—	—	—	Cancer, lung
Cadmium compounds	CdTe and CdS deposition $CdCl_2$ treatment	0.01 mg/				m^3 (dust) 0.002 mg/ m^3 (fumes)
—	—	NA	Cancer, kidney			
Carbon tetrachloride	Etchant	5 ppm	10 ppm		100 ppm	Liver, cancer, greenhouse gas
Chlorosilanes	a-Si and *x*-Si deposition	5 ppm	—	800 ppm	—	Irritant
Copper	CIS deposition	1 mg/				m^3 (dust) 0.2 mg/ m^3 (fumes)
—	100 mg/m^3	—				
Diborane	a-Si dopant	0.1 ppm	—	40 ppm	1 ppm	CNS, pulmonary
Germane	a-Si dopant	0.2 ppm	—	—	—	Blood, CNS, kidney
Hydrogen	a-Si deposition	—	—	—	—	Fire hazard
Hydrogen fluoride	Etchant	—	C[f], 3 ppm	30 ppm	20 ppm	Irritant, burns, bone, teeth
Hydrogen selenide	CIS sputtering	0.05 ppm	—	1 ppm	—	Irritant, GI, flammable
Hydrogen sulfide	CIS sputtering	10 ppm	15 ppm	100 ppm	30 ppm	Irritant, CNS, flammable
Indium compounds	CIS deposition	0.1 mg/m^3	—	—	—	Pulmonary, bone, GI
Lead	Soldering	0.05 mg/m^3	—	—	—	CNS, GI, blood, kidney, reproductive
Nitric acid	Wafer cleaning	2 ppm	4 ppm	25 ppm	—	Irritant, corrosive
Phosphine	a-Si dopant	0.3 ppm	1 ppm	50 ppm	0.5 ppm	Irritant, CNS, GI, flammable
Phosphorous oxychloride	*x*-Si dopant	0.1 ppm				Irritant, kidney
Selenium compounds	CIS deposition	0.2 mg/m^3		1 mg/m^3		Irritant
Sodium hydroxide	Wafer cleaning		C[f], 2 mg/m^3	10 mg/m^3	5 mg/m^3	Irritant
Silane	a-Si deposition	5 ppm	—	—	—	Irritant, fire and explosion hazard
Silicon tetrafluoride	a-Si deposition	—				
Tellurium compounds	CIS deposition	0.1 mg/m^3				CNS, cyanosis, liver

[a] Abbreviations used: CVD, chemical vapor deposit; NA, not available; CNS, central nervous system; GI, gastrointestinal tract.

[b] The threshold limit value, time-weighted average (TLV-TWA) is defined by the American Conference of Governmental Industrial Hygienists (ACGIH) as the time-weighted average threshold concentration above which workers must not be exposed during work shifts (8 h/day, 40 h/week).

[c] The threshold limit value, short-term exposure level (STEL) is defined by the ACGIH as the maximum concentration to which workers can be exposed for up to 15 min, provided not more than four excursions per day are permitted with at least 60 min between exposures, and provided that the daily permissible exposure limit is also not exceeded.

[d] The immediately dangerous to life or health concentration (IDLH) is defined by the National Institute for Occupational Safety and Health (NIOSH) as the maximum concentration from which one could escape within 30 min without any symptoms impairing escape or any irreversible health effects.

[e] The emergency response planning guideline-2 (ERPG-2) is defined by the American Industrial Hygiene Association as the concentration below which nearly all people could be exposed for up to 1 h without irreversible or other serious health effects or symptoms that would impair their ability to take protective action.

[f] The threshold limit value—ceiling (C) is defined by the ACGIH and the NIOSH as the concentration that should not be exceeded during any part of the working exposure.

ignition in air is 2 or 3%, depending on the carrier gas. If mixing is incomplete, a pyrophoric concentration of SiH_4 may exist locally, even if its concentration in the carrier gas is less than 2%. At silane concentrations $\geq$4.5%, the mixtures were found to be metastable and ignited after a certain delay. In an accident, this occurrence could be extremely destructive because the protection provided by venting would be ineffective. In addition to SiH_4, hydrogen used in a-Si manufacturing is flammable and explosive. Most PV manufacturers use sophisticated gas-handling systems with sufficient safety features to minimize the risks of fire and explosion. Some facilities store silane and hydrogen in bulk from tube trailers to avoid the need to frequently change gas cylinders. A bulk module typically contains eight cylindrical tubes that are mounted onto a trailer suitable for short- and long-range transport. These modules carry up to 3000 kg of silane. Another option utilizes a single, 450-liter cylinder mounted on a skid, which contains up to 150 kg of silane (minibulk). For additional safety, these storage systems are equipped with isolation and flow-restricting valves.

Bulk storage decreases the probability of an accident since changing the trailer is an infrequent event that is executed in a precise, well-controlled manner under the supervision of the plant's management, safety officials, the gas supplier, and local fire department officials. On the other hand, if an accident does occur, the consequences can be much greater compared to those involving gas cylinders. Silane is used mainly in glow discharge deposition at very low utilization rates (e.g., 10%). To the extent that this rate increases in the future, the potential worst consequences of an accident will be reduced.

Toxic doping gases (e.g., AsH_3, PH_3, and GeH_4) are used in quantities too small to pose any significant hazards to public health or to the environment. However, leakage of these gases can cause significant occupational risks, and management must provide continuous vigilance to safeguard personnel.

3.2 Public Health and Environmental Issues

Silane, used in bulk quantities in a-Si facilities, may pose hazards to the surrounding community if separation zones are inadequate. In the United States, the guidelines of the Compressed Gas Association specify minimum distances from places of public assembly that range from 80 to 450 ft, depending on the quantity and pressure of silane in the containers used. The corresponding minimum distances to the plant's property lines are 50–300 ft. These prescribed separation distances are considered sufficient to protect the public under worst-condition accidents. No environmental issues have been identified with this technology.

4. CADMIUM TELLURIDE SOLAR CELLS

4.1 Occupational Health Issues

In CdTe manufacturing, the principal concerns are related to the toxicity of the feedstock materials (e.g., CdTe, CdS, and $CdCl_2$). The occupational health hazards presented by Cd and Te compounds in various processing steps vary as a function of their specific toxicity, physical state, and the mode of exposure. No clinical data are available on the human health effects of exposure to CdTe. The National Institute of Environmental Sciences conducted a limited number of laboratory tests comparing the acute toxicity of CdTe, CIS, and copper indium gallium diselenide (CGS). These tests showed that of the three compounds, CdTe has the highest toxicity and CGS has the lowest. No comparisons were made with the parent Cd and Te substances. Cadmium, one of the CdTe precursors, is an extremely hazardous material. The acute health effects from inhaling it include pneumonitis, pulmonary edema, and death. However, CdTe is insoluble to water; as such, it may be less toxic than cadmium. This issue needs further investigation.

In solar cell manufacturing facilities, workers may be exposed to Cd compounds through inhalation of contaminated air and by ingestion from hand-to-mouth contact. Inhalation is probably the most important pathway because of the greater potential for exposure and the higher efficiency of absorption of Cd compounds through the lung than through the gastrointestinal tract. The physical state in which the Cd compound is used and/or released to the environment is another determinant of risk. Processes in which Cd compounds are used or produced in the form of fine fumes or respirable particulates present larger health risks than when they are in liquid or solid forms. Similarly, operations involving volatile or soluble Cd compounds (e.g., $CdCl_2$) must be more closely scrutinized. Hazards to workers may arise from preparing feedstock, fume/vapor leaks, etching of excess materials from panels, maintenance

operations (e.g., scraping and cleaning), and handling of waste. Caution must be exercised when working with this material, and several layers of control must be implemented to prevent employees from being exposed. In general, the hierarchy includes engineering controls, personal protective equipment, and work practices. Area and personal monitoring would provide information on the type and extent of workers' exposure, assist in identifying potential sources of exposure, and provide data on the effectiveness of the controls. The U.S. industry is vigilant in preventing health risks and has established proactive programs in industrial hygiene and environmental control. Workers' exposure to Cd in PV manufacturing facilities is controlled by rigorous industrial hygiene practices and is continuously monitored by medical tests, thus minimizing potential health risks.

4.2 Public Health and Environmental Issues

No public health issues have been identified for this technology. Environmental issues are related to the disposal of manufacturing waste and end-of-life modules; these are discussed in Section 8.

5. COPPER INDIUM DISELENIDE SOLAR CELLS

5.1 Occupational Health and Safety

The most common processes for forming CIS solar cells are the coevaporation of Cu, In, and Se and the selenization of Cu and In layers in an H_2Se atmosphere. The toxicity of Cu, In, and Se is considered mild. There is little information on the toxicity of CIS. Laboratory animal studies have shown CIS to have mild to moderate toxicity in the respiratory track; in comparing CIS, CGS, and CdTe, CIS was found to be less toxic than CdTe and somewhat more toxic than CGS.

The selenium threshold level [threshold limit value, time-weighted average (TLV-TWA)] of $0.2\,mg/m^3$ was set to prevent systemic toxicity and to minimize the potential for irritation of the eyes and upper respiratory tract. Interestingly, selenium is an essential element in the human diet and in small doses has exhibited cancer-preventive properties.

Although elemental selenium is only mildly toxic, hydrogen selenide is highly toxic. It has an immediately dangerous to life and health (IDLH) concentration of only 1 part per million (ppm). Hydrogen selenide resembles arsine physiologically; however, its vapor pressure is lower than that of arsine and it is oxidized to the less toxic selenium on the mucous membranes of the respiratory track. Hydrogen selenide has a TLV-TWA of 0.05 ppm to prevent irritation and the onset of chronic hydrogen selenide-related disease. To mitigate hazards from H_2Se, the CIS deposition system should be enclosed under negative pressure and be exhausted through an emergency control scrubber. The same applies to the gas cabinets containing H_2Se cylinders in use.

The options for substitution, isolation, work practices, and personnel monitoring discussed for CdTe are also applicable to CIS manufacturing. In addition, the presence of hydrogen selenide in some CIS fabrication processes requires pertinent engineering and administrative controls to safeguard workers and the public against exposure to this highly toxic gas.

5.2 Public Health and Environmental Issues

Potential public health issues are related to using hydrogen selenide in facilities that employ it as a major feedstock material. The hazards can be minimized by using safer alternatives, limiting inventories, using flow-restricting valves, and employing other safety options. Emissions of hydrogen selenide from process tools are controlled with either wet or dry scrubbing. Also, some facilities have scrubbers that can control accidental releases of this gas. Environmental issues are related to the disposal of manufacturing waste and end-of-life modules; these are discussed in Section 8.

6. GALLIUM ARSENIDE HIGH-EFFICIENCY SOLAR CELLS

6.1 Occupational Health and Safety

Metal organic chemical vapor deposition (MOCVD) is the most common process for fabricating III/V high-efficiency PV cells. It employs as feedstock arsine and phosphine, which are highly toxic hydride gases. Similar to the handling of silane and hydrogen selenide, the safe use of these hydrides requires several layers of engineering and administrative controls to safeguard workers and the public against accidental exposure. Such requirements pose financial demands and risks that could hinder scaling up of the

technology to multi-megawatt levels. Part of the problem is that use of the hydrides in MOCVD is highly ineffective. Only approximately 2–10% are deposited on the PV panels because a 10–50 times excess of V to III (As to Ga) is required. Metal organic compounds are used more effectively, with their material utilization ranging from 20 to 50%. In scaling up to 10 MW/year production using MOCVD, the current designs of flat-plate III–V modules require several tons of arsine and hydrogen. These quantities can be effectively delivered only with tube trailers, each carrying 2 or 3 tons of gases. The potential consequences of a worst-case failure in one of them could be catastrophic. However, it is more likely that terrestrial systems will be concentrators, not flat plates, because the former would be less expensive to manufacture. PV cells concentrate light between 100 and 2000 times. Therefore, a possible practical strength is 500 times concentrators; the material requirements for producing such concentrators are 600 times less than those needed for flat plates, and worst consequences are equally reduced.

The best way to minimize both the risks associated with certain chemicals and the costs of managing risk is to assess all alternatives during the first steps of developing the technology and designing the facility. Studies by Komeno and coworkers have shown that these hydrides may be replaced in the future by using tertiary butyl arsine (TBAs) and tertiary butyl phosphine (TBP). There appear to be no intrinsic technical barriers to growing PV-quality GaAs with TBAs and GaAsP or $GaInP_2$ with TBP. However, until substitutes are tested and implemented, it is prudent to use arsine and phosphine from commercial, reduced-pressure containers. Research efforts are under way in Europe to replace hydrogen by inert nitrogen. Apparently, there is no inherent reason to prohibit such a substitution. However, since molecular hydrogen decomposes to some extent, and atoms participate in the gas-phase chemistry, the PV research community is challenged with learning how to optimize III–V growth conditions with nitrogen.

7. OPERATION OF PV MODULES

The operation of PV systems does not produce any emissions. Although some toxic materials are imbedded in PV modules, these are used in extremely small quantities (e.g., 5–10 g/m^2) and they are safely encapsulated within sheets of glass or plastic. Toxic compounds cannot cause any adverse health effects unless they enter the human body in harmful doses. The only pathways by which one might be exposed to any materials from a PV module are by accidentally ingesting flakes or dust particles or inhaling dust and fumes. The PV material layers are stable and solid, and they are encapsulated between thick layers of glass or plastic. Unless the module is finely ground, dust particles cannot be generated. All the PV materials discussed herein have a zero vapor pressure at ambient conditions. Therefore, it is impossible for any vapors or dust to be generated during normal use of PV modules.

The potential exists for exposure to toxic vapors via inhalation if the modules are consumed in a residential fire and people inhale the smoke from the fire. However, common residential fires in the United States are not likely to vaporize CdTe and GaAs layers; flame temperatures in roof fires are in the 800–900°C range. The melting point of CdTe is 1041°C, and evaporation occurs at 1050°C in open air and at approximately 900°C under nonoxidizing conditions. The melting point of CdS is 1750°C and that of GaAs is 1238°C. CIS starts evaporating at 600°C, but only a 20% weight loss was measured at 1000°C. Experiments at Brookhaven National Laboratory showed that CdTe modules with both back and front glass effectively encapsulate Cd when heated up to 1100°C.

8. DECOMMISSIONING PV MODULES

Photovoltaic modules may contain small amounts of regulated materials, which vary from one type to another. Environmental regulations can determine the cost and complexity of dealing with end-of-life PV modules. If they are classified as hazardous according to federal or state criteria, then special requirements for material handling, disposal, record keeping, and reporting increase the cost of decommissioning the modules. In doing so, the principal concern regards the presence of Cd in CdTe and CdS solar films and of Pb in *x*-Si modules if they contain Pb-based solder. Should these modules end in a municipal waste incinerator (MWI), the heavy metals could gasify and a fraction of the gases will be released into the atmosphere. If the MWI is equipped with an electrostatic precipitator (ESP), this fraction can be as small as 0.5%, with the balance of the heavy metals remaining in the ash. The ash must be disposed of in controlled landfills.

Should the modules end up in municipal landfills, there is the potential for the heavy metals to leach

out into the soil. The leachability of metals in landfills is characterized by two elution tests: the U.S. Environmental Protection Agency Toxicity Characterization Leachate Profile (TCLP) and the German DEV S4 (Deutsches Einheitsverfahren). Both tests assume a worst-case scenario for such potential leaching. In these tests, small pieces ($<1\,cm^2$) of broken modules are suspended and rotated in an eluent for 24 h. The metals present in the eluent are then measured and compared with limits prescribed by each testing protocol. If the metals' concentration exceeds the limits, the modules are demonstrating the metals' leachability and may need to be recycled or disposed of in a hazardous waste landfill. If the metals are not leaching in excessive quantities, the modules can be disposed of in a commercial landfill. Early CdTe modules failed the TCLP and the DEV tests, while Apollo modules produced by BP Solar have been reported to pass the TCLP. In exploratory tests with a few commercial *x*-Si modules, some failed the TCLP limit for Pb by approximately 30%. No data were found on TCLP testing of currently produced *x*-Si modules. CIS modules passed the test for Se in a first set of experiments. No data were found for GaAs modules. The a-Si modules contain very little hazardous material and easily pass the test. The TCLP test is a conservative one that requires breaking up the whole module into very small pieces, whereas under normal conditions the PV layer is sandwiched between two layers of glass and reasonably isolated from the environment.

The ultimate solution to PV waste and end-of-life management is recycling the useful materials. Studies at Brookhaven National Laboratory and Solar Cells, Incorporated showed that recycling is technologically and economically feasible based on the current collection/recycling infrastructure and on emerging recycling technologies. Metals from used solar panels in large centralized applications can be reclaimed in metal smelting/refining facilities that use the glass as a fluxing agent and recover most of the metals by incorporating them in their product streams. In dispersed operations, small quantities and the high costs of transportation make this option relatively expensive. Research supported by the U.S. Department of Energy Small Business Research Initiative (SBRI) program aims the development of technologies for hydrometallurgical separation that may be used for both small-scale (in-house) and large-scale recycling. The PV industry is investigating these options as part of its proactive long-term environmental strategy to preserve the environmental friendliness of solar cells. Although the PV industry will not face this problem on a large scale before 2020, today's choices of materials and module designs may very well set a precedent for the future. The PV industry may choose to recycle spent modules, even if there are no regulatory issues related to their disposal, because recycling may improve market penetration.

9. ENVIRONMENTAL BENEFITS OF PHOTOVOLTAICS OPERATION

Carbon dioxide emissions are zero during the operation of PV systems because they require little or no maintenance or oversight. Some CO_2 emissions can be attributed to manufacturing because it takes energy to make a module. Thus, PV systems require some input energy, which they pay back early in their lifetime. This energy payback (EPB) was the focus of studies by Alsema, Frankl, Kato, and others. In the past, PV manufacture had a high EPB because the technology was immature and burdened by energy-intensive processing. Existing PV systems have an EPB of 3–10 years; future technological options likely will reduce the system's EPB to 1–3 years. In this range, the amount of CO_2 displaced by PV during a 30-year lifetime outdoors is 90–97% compared to the CO_2 of the energy it offsets. For example, assuming that U.S. energy generation causes 160 g carbon equivalent of CO_2 per kilowatt-hour of electricity, then a 1-kW PV array in an average U.S. location would produce approximately 1600 kWh each year and 48,000 kWh in 30 years. This array would avoid approximately $(48{,}000 \times 0.95) \times 160$ g/kWh, or approximately 7 metric tons of carbon equivalent during its useful life. Building integrated photovoltaics may further shorten the period of EPB by avoiding the energy required for structural supports. Exact predictions require a knowledge of the mix of energies with which PV will be manufactured and the mix that it will displace (which will vary with location, application, and date). The U.S. Photovoltaic Industry Roadmap projected that by 2020 the domestic PV industry could provide up to 3.2 GW of new peak electricity, increasing to approximately 25 GW by 2030. These projections correspond to an average growth in PV capacity of 25% per year. The estimated mixture of applications is one-half AC distributed generation, one-sixth AC wholesale grid generation, and one-third DC and AC value applications.

Fthenakis and Morris used the MARKAL MACRO model to test these projections. MARKAL is a multiperiod, linear programming model that

captures the complex interrelationships of energy systems and their environmental emissions across the spectrum from energy resources to energy services. The model minimizes the total cost of the energy system over a multiyear period, subject to constraints, such as limitations on pollution emissions or on the growth rates of technologies. The analysis includes the entire energy system: extracting the resources; processing operations (such as refineries); electric, gas, and oil transmission and distribution operations; and end-use technologies, such as lighting, industrial heat production, and transportation technologies (from automobiles to ships and airplanes).

All energy resources on both the supply and demand sides, as well as energy efficiency and conservation technologies, compete in an even-handed manner. The model and its associated database are continually upgraded. Current projections are based on the 2001 MARKAL database, designed to closely match the U.S. Energy Information Administration's "Annual Energy Outlook 2001" except for the PV and wind data, which were obtained from the Department of Energy (DOE) and the Electric Power Research Institute (EPRI). There are 922 technologies in the MARKAL database, including demand devices such as residential heating, commercial refrigeration, and lighting. The database includes 52 electricity-generating technologies. MARKAL also projects pollution emissions based on the characteristics of the individual technologies in the energy mixture.

Table III shows the PV input data for the model. This information derives from the PV Roadmap for 2000–2020 and from the DOE/EPRI for 2025 and 2030 (the Roadmap stops at 2020). We assumed that the peak electricity demand largely overlapped the peak power demand. The only constraint on the model was the upper bound on growth of 25% per year. An average U.S. insolation of 1800 kW/h/m^2 was assumed; the PV penetration in regions of high solar insolation could be much greater than the average (Table IV).

Carbon and other gas emissions are an important part of the model's output. PV technology under the predicted rates of implementation in the United States would accelerate a steadily increasing displacement of carbon emissions. In 2030, this displacement is projected to be 17 million tons/year, and during the entire 30-year period PV would displace a total of 128 million metric tons of carbon (Table V).

In addition, in 2030 alone, PV would displace 169,000 tons of SO_2, 103,000 ton/year of NO_x, and 6200 tons of particulates (Table VI).

TABLE III

Cost Assumptions for PV ($/Wp) Input Values in MARKAL

	2000	2005	2010	2015	2020	2025	2030
$/Wp	6.1	4.3	3.0	2.1	1.5	1.0	0.9

TABLE IV

Total U.S. Capacity and % PV for the MARKAL Roadmap Case

	2000	2005	2010	2015	2020	2025	2030
Total capacity (kW/h/m^2)	768	821	895	972	1033	1129	1252
PV capacity (kW/h/m^2)	0.08	0.32	1.19	3.84	11.93	36.6	111.9
% PV	0.01	0.04	0.13	0.40	1.15	3.24	8.94

TABLE V

Displacement of Carbon Emissions by PV (Million Metric Tons/Year)

	2000	2005	2010	2015	2020	2025	2030
Carbon displaced	0.02	0.06	0.21	0.64	1.98	5.80	17.02

TABLE VI

Displacement of Other Emissions by PV (Thousands of Tons/Year)

	2000	2005	2010	2015	2020	2025	2030
NO_x	0.17	0.70	2.23	5.92	15.60	39.8	103
PM10	0.005	0.02	0.08	0.22	0.89	2.10	6.22
SO_2	0.23	0.82	2.62	7.45	21.3	59.4	169

This analysis addresses only the grid-connected portion of the Roadmap in the highly competitive U.S. market; it does not examine the PV prospects of the promising markets of the developing world.

10. CONCLUSION

The manufacture of PV modules uses some hazardous materials, which can present health and safety risks if adequate precautions are not taken. Routine conditions in manufacturing facilities should not pose any threat to health and the environment.

Hazardous materials could adversely affect occupational health and, in some instances, public health during accidents. Such hazards arise primarily from the toxicity and explosiveness of specific gases. Accidental releases of hazardous gases and vapors can be prevented by using safer technologies, processes, and materials, better material utilization, and enhancing employee training and safety procedures. As the PV industry vigilantly and systematically approaches these issues and mitigation strategies, the risk to the industry, the workers, and the public is continually minimized.

Photovoltaics in the energy market will replace fossil fuel combustion technologies and will subsequently reduce the discharges of carbon dioxide, sulfur dioxide, nitrogen oxide, and particulates throughout the region where they are installed. Based on energy–environmental–economic analysis, we predict that with the current price projections for the energy technologies in the U.S. energy mixture, photovoltaics could reach a capacity of 12 GW by 2020 and 112 GW by 2030, eliminating a significant amount of pollutants from the U.S. energy cycle.

SEE ALSO THE FOLLOWING ARTICLES

Arid Environments, Impacts of Energy Development in • Electricity, Environmental Impacts of • Photovoltaic Conversion: Space Applications • Photovoltaic Energy: Stand-Alone and Grid-Connected Systems • Photovoltaic Materials, Physics of • Solar Cells

Further Reading

Alsema, E. A. (1998). CO_2 and energy payback of PV. *In* "Photovoltaics and the Environment" (V. Fthenakis, *et al*., Eds.), Workshop Report BNL-52557. Brookhaven National Laboratory, Upton, NY.

Frankl, P., and Gamberale, M. (1998). Analysis of energy and CO_2 aspects of building integration of photovoltaic systems. *In* "Photovoltaics and the Environment (V. Fthenakis, *et al*., Eds.), Workshop Report BNL-52557. Brookhaven National Laboratory, Upton, NY.

Fthenakis, V. M. (2000). End-of-life management and recycling of PV modules. *Energy Policy* **28,** 1051–1058.

Fthenakis, V. M. (2001). Multi-layer protection analysis for photovoltaic manufacturing. *Process Safety Prog.* **20**(2), 1–8.

Fthenakis, V. M., and Gonsiorawski, R. (1999). Lead-free solder technology from ASE Americas, Workshop Report BNL-67536. Brookhaven National Laboratory, Upton, NY.

Fthenakis, V. M., and Morris, S. (2003). MARKAL analysis of PV capacity and CO_2 emissions' reduction in the U.S. *In* "Proceedings of the World Conference on Photovoltaic Energy Conversion (WCPEC-3), May 11–18, 2003, Osaka, Japan."

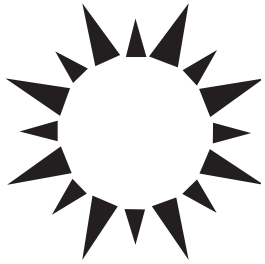
Fthenakis, V., Morris, S., Moskowitz, P., and Morgan, D. (1999). Toxicity of cadmium telluride, copper indium diselenide, and copper gallium diselenide. *Prog. Photovoltaics* **7,** 489–497.

Kato, K., Hibino, T., Komoto, K., Ihara, S., Yamamoto, S., and Fujihara, H. (2001). A life cycle analysis on thin film CdS/CdTe PV modules. *Solar Energy Materials Solar Cells* **67,** 279–287.

Komeno, J. (1994). Metalorganic vapor phase epitaxy using organic group V precursors. *J. Crystal Growth* **145**(1–4), 468–472.

Tamanini, F., Chaffee, J. L., and Jambar, R. L. (1998). Reactivity and ignition characteristics of silane/air mixtures. *Process Safety Prog.* **17**(4), 243–258.

Physics and Economics of Energy, Conceptual Overview



KARL-ERIK ERIKSSON
Karlstad University
Karlstad, Sweden

1. Introduction
2. Physical Aspects of Energy
3. Cosmic Origin of Exergy Resources
4. Energy in Society
5. Energy, Values, and Economics
6. The Evolutionary Perspective

Glossary

entropy A measure of the disorder or unavailability of energy within a closed system. More entropy means less energy is available for doing work.

First Law of Thermodynamics The internal energy in a system isolated from other systems remains constant (energy is conserved).

Law of Conservation of Matter During an ordinary chemical change, there is no detectable increase or decrease in the quantity of matter.

Second Law of Thermodynamics The principle that the disorder of an isolated system (entropy) increases to a maximum.

1. INTRODUCTION

The use of energy in human society is a many-faceted subject. Many fields of knowledge are involved in the understanding of this subject. Energy is a fundamental concept in physics. When dealing with energy in society, a certain usefulness is assumed, a set of qualities of energy forms that are also possible to describe in physical terms. Energy is procured from natural systems and given back to natural systems after transformations in the society. These processes are economic and biological as well as technical in nature and they are often regulated by society. The technical processes can be designed or analyzed on the basis of the physical sciences. Exchanges and regulations can be analyzed in terms of values and costs/risks. Here, environmental and health aspects become intertwined with technical, economic, and judicial aspects. On a fundamental level, there are the value aspects, in particular, ethical aspects.

One must keep in mind that all these aspects are relevant. In science, one often must simplify, and there are often temptations for scientists to adopt one aspect as being supreme and to let it dominate the description or analysis. Simplified pictures of this kind can be very pedagogical, but they may also lead to misconceptions that can be very harmful. One way to check one's understanding is to ask what the system/process or the problem in question looks like on another level. Traditionally, in science, analyzing in terms of subsystems or subprocesses has been a very successful approach. It has become increasingly clear that it may also be necessary to analyze a certain system or process in relation to the larger context, sometimes several larger contexts, in which it is embedded.

In the case of energy in society, from looking mainly at technical systems for providing primary energy, one has been forced to ask whether the whole societal energy system can be sustainably embedded in the ecosphere. For nuclear energy, one has been forced to ask whether the bulk handling of radioactive material can be performed safely in the present and future world with its conflicting interests and unforeseeable risks. The time scale involved in considering the handling of nuclear waste is long not only compared to lead times in energy technology but even compared to the know human history.

One system boundary that must be broken immediately is the one between energy and matter. First, most of the crucial problems of the energy system are material-related. In the case of energy and climate, it is the transition of carbon from its original

 © 2004 Elsevier Inc. All rights reserved.

state in fossil coal, oil, or gas to atmospheric carbon dioxide that is crucial. Second, a very large part of the energy use in society is related to material transformations or transports of materials. Energy use in the aluminum and cement industries are two examples of this.

Physical processes of conversions of energy and materials are fundamental for the metabolism of living organisms and for the metabolism of human society. Different forms of energy and matter have different physical (thermodynamic) values and also different degrees of usefulness and economic value.

Sections 2 and 3 deal with basic concepts and principles related to the physical aspects of energy. Section 4 deals with the use of energy in society and the interface between physics and economics. Section 5 discusses energy in connection to values and economic principles. The two aspects are brought together in Section 6, where the evolutionary perspective on the use of energy and matter meets the human responsibility of today's resource handling on Earth.

2. PHYSICAL ASPECTS OF ENERGY

The scientific field that deals with transformations of energy and materials is thermodynamics, which describes macroscopic quantities of energy and matter. The theoretical way to understand this is statistical; the underlying theory is statistical mechanics, which analyzes populations of many particles and their motions and interactions.

In the late 1950s, E. T. Jaynes showed how statistical mechanics can be interpreted in terms of another relatively new theory, C. E. Shannon's theory of information, basically a theory of signals. It deals with probability distributions, conjectured or observed. Information theory is of fundamental interest for describing an important use of energy in nature and society, the build up of structures.

In information theory, an answer, "yes" or "no," or, equivalently, a binary figure, 0 or 1, carries one bit of information, against a background of total uncertainty with *a priori* probability $\frac{1}{2}$ for 0 or 1. The information carried by a certain symbol having a priori probability p is (in bits)

$$^2\log\frac{1}{p}.$$

Clearly, when $p = \frac{1}{2}$, the result is unity.

In a text, the different letters appear with different probabilities (different frequencies), and hence they carry different amounts of information. In an n-letter alphabet, the letters can be denoted by $i\,(i = 1, 2, \ldots, n)$. Let p_i be the probability that, in a randomly chosen space, the letter that one finds there is i. Weighing the information gain in finding the letter i, $^2\log\frac{1}{p_i}$ with the probability p_i for its appearance, and summing over all letters, one obtains the expected information gain, i.e., the present information deficit,

$$S(p) = \sum_{i=1}^{n} p_i \, ^2\log\frac{1}{p_i}.$$

This is the entropy of the probability distribution $p = \{p_i\}$.

From looking just at single letters in a text, one can continue and look at pairs of neighboring letters or, more generally, at strings, i.e., sequences, of increasing length. In a text in a certain language, there are correlations between neighboring letters; i.e., certain letter combinations occur more frequently than others.

If there are two probability assignments, $p^{(0)} = \{p_i^{(0)}\}$ and $p = \{p_i\}$, and the latter is based on more knowledge than the former, then the information gain in going from the former to the latter is

$$K(p^{(0)}; p) = \sum_{i=1}^{n} p_i \left[-\left(^2\log\frac{1}{p_i} - ^2\log\frac{1}{p_i^{(0)}} \right) \right]$$
$$= \sum_{i=1}^{n} p_i \, ^2\log\frac{p_i}{p_i^{(0)}}.$$

(The information difference is minus the information deficit difference; moreover, the best probability assignment $p = \{p_i\}$ must be used to give the weights.)

Here, $K(p^{(0)}; p)$ is the relative entropy (or Kullback information) of the probability distribution p with respect to the reference distribution $p^{(0)} = \{p_i^{(0)}\}$. The Kullback information has the property that it is never negative and it is zero only when the probability distributions coincide.

These two concepts, entropy and relative entropy, are the basic concepts of information theory. They are very flexible and can be readily used in information theory, statistics, linguistics, and physics, as well as in other sciences.

As already mentioned, one may generalize from a string of discrete letters, i.e., a text, to a lattice in two or more dimensions. An important example of a two-dimensional lattice is a picture image, consisting of pixels, in black/white or color. Generalizing to continuous systems in any number of dimensions is also possible. In classical physics, one has vectors in the so-called phase space of positions and momenta of many particles, and the resulting theory is, as

mentioned already, statistical mechanics. Here, probability assignments represent macroscopic physical states. Extension into quantum mechanics is quite straightforward but is not discussed here. It should be mentioned, though, that this extension is essential for several important energy technologies.

Entropy is thus a fundamental quantity not only in information theory but also in statistical mechanics and thermodynamics. The two entropies are identical for a physical system except for a constant factor,

$$S_{\text{thermodyn}} = (k\ln 2)S_{\text{info}},$$

where k is Boltzmann's constant, and

$$k\ln 2 = 1.0 \times 10^{-23}\,\text{JK}^{-1}\,\text{bit}^{-1}.$$

Energy is a fundamental physical quantity. It is closely related to time and momentum is closely related to space. In fact, the quantity representing energy in quantum mechanics also has the function of generating displacement in time and momentum generates spatial displacement. Conservation of energy, the first law of thermodynamics, is related to the fact that the laws of nature are invariant under time displacements. Similarly, conservation of momentum (the law of inertia) and conservation of angular momentum follow from the universality of physical laws in space.

Matter, under the conditions of the universe of today, is made up of nucleons (neutrons and protons, the constituents of atomic nuclei) and electrons. Three quarks constitute a nucleon. The mass of an electron is approximately 1/200 of the mass of a nucleon. In particle physics, as studied at accelerator laboratories, a large variety of particles made up of quarks and anti-quarks are created and investigated. This includes the formation of matter and antimatter. Also leptons, the particle family to which the electron belongs, can be similarly produced and studied.

Electric charge is conserved. The same is true for lepton number and quark number with extremely good accuracy (except possibly for extreme situations, such as in the very early universe). Since accelerator laboratory reactions are rare, this means that for all practical purposes, one can regard the number of nucleons as constant and consider this the law of conservation of matter.

The state of a system of many particles is usually not exactly known. Even if the laws of motion were exactly known, the uncertainty would still grow in time due to delicate dependence on unknown parameters of the initial state. Hence, for any system left by itself, the entropy increases. This is the second law of thermodynamics.

When the entropy of a closed system has reached its maximum value, the system will remain in this state; i.e., it is in its equilibrium state. Hence, the deviation of entropy from its maximum value is a measure of the degree of nonequilibrium.

Another useful measure related to this is the Kullback information or relative entropy, defined above. Where there is disequilibrium, there is, at least in principle, the possibility of extracting work. Consider an ideal process, i.e., a reversible and entropy non-producing process, to extract mechanical work (which is of zero entropy) from a system, while it approaches equilibrium. Then, the amount of extractable work from the system, called the exergy E of the system, is proportional to the Kullback information,

$$E = (k\ln 2)T_0K(p^{(0)};p).$$

Here, p is the original state of the system, and $p^{(0)}$ is the equilibrium state of temperature, T_0, reached reversibly under the extraction of the amount E of mechanical work.

This may all seem very theoretical, but it is applicable to all sorts of disequilibria, such as differences in temperature or in pressure or chemical disequilibria. It is thus generally applicable to physical and technical systems and processes for energy and material transfers and conversions.

Earth is situated in a radiation flow from the Sun and it radiates this energy as heat into the interstellar space. The geophysical systems and life on Earth's surface convert radiated energy coming from the Sun at an average temperature of 5800 K. Approximately 30% is reflected. After conversions, involving the water cycle and the carbon cycle, the absorbed energy leaves Earth as radiation at an average temperature of 255 K (−18°C). This can be viewed as a huge heat engine with an available exergy power of 115,000 TW. Of this, the global society uses less than 0.01%.

In terms of information inflow or, rather, flow of information capacity, the radiation exchange of Earth amounts to 4×10^{37} bits per second. Of the huge inflow of information, some has been coded into the structure of life and the complex molecular interplays within the cell, also into the regulatory mechanisms within organisms, between individuals of a population, and between species of an ecosystem. Life has interacted with and changed the geophysical systems in a mutually adaptive evolution. Earth has a large system of biogeochemical

cycles on which life depends. Except for the influence of actions of the human society, in particular the industrial society, this system would be relatively stationary.

There are many examples of structure formation in systems with a through-flow of energy (Fig. 1). The whole living world contains an enormous richness of such phenomena on various scales. There are also many non-living examples, such as cloud formation and the formation of snowflakes.

A classroom experiment of self-organization, easy to repeat, is the set of reactions in a solution of chemicals, devised by Belousov and Zhabotinsky. The solution contains an energy-rich substance and substances that become involved in nonlinear reaction kinetics and diffusion. Small thermal fluctuations in concentrations are sufficient to trigger the system to form visible geometrical structures that are slowly changing and eventually decaying.

Figure 2 shows a diagram of exergy (information capacity) flows in such a self-organizing system. The same principle applies also to spontaneous structuring on various levels in cosmic, stellar, and biological evolution. A through-flow of energy can give a net inflow of exergy. In a self-organizing system, there are flows driven by gradients, differences in concentration, pressure, or temperature. Such motions are irreversible processes involving entropy production.

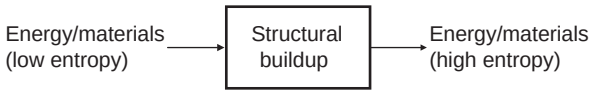


FIGURE 1 Diagram of structure formation of a system in a flow of energy or of energy with matter. Due to differences in entropy between the flows, there is a net inflow of exergy or information capacity. There is also entropy production (dissipation) within the system.

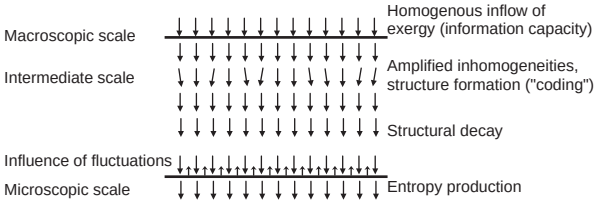


FIGURE 2 Diagram indicating flow of information from the macroscopic scale to the microscopic scale via an intermediate scale with structure formation triggered by fluctuations. Structure formation can take place on different scales. Structure formation and decay are both present in the flow from macro to micro.

In practice, of course, no process is totally ideal/reversible; there is always production of entropy. Therefore, not all of the exergy present in a system can be extracted as work. The exergy loss in a conversion process is proportional to the entropy produced in the process,

$$W = E - T_0 \Delta S.$$

(Here, E the original exergy, W is the extracted work, ΔS is the entropy production, and T_0 is a reference temperature, usually the ambient temperature or the equilibrium temperature.)

Thus, exergy can be lost but not created. It is something valuable, also economically. Therefore, it is useful as a physical resource measure. Expressed in thermodynamic quantities, the exergy is

$$E = U - T_0 S + p_0 V - \sum_i \mu_{i0} N_i,$$

where U is the internal energy, S is the entropy, V is the volume, and N_i are the amounts of different chemical components. Furthermore, T_0 is the reference temperature, p_0 is the reference pressure, and μ_{i0} are reference chemical potentials. The exergy is thus context-dependent through the choice of reference system.

Apart from nuclear reactions and decays, chemical elements are conserved. Thus, chemical elements, energy, and exergy can be subject to bookkeeping; i.e., balance accounts can be set up for inflows, outflows, and internal movements and conversions. Clearly, such bookkeeping can easily be generalized to include nuclear reactions and decays, for instance, in the nuclear industry. If the context dependence of exergy is to be avoided, entropy and entropy production can be taken into account instead. Physical bookkeeping can be done for natural systems, such as an ecosystem, a population, an individual organism, or a cell, but also for production/consumption units, such as a factory, a machine, a car, or a household.

The disequilibrium between the dense and hot Sun and the cold interstellar space functions like the poles of a large battery. Earth is situated in a radiation flow, an energy current from the hot pole to the cold pole of this battery. On Earth, this current charges secondary batteries and secondary currents: temperature and pessure differences, differences in gravitational potentials between high-altitude dams and the sea; the water cycle, incuding rivers and ocean currents, winds, and waves. Many of those secondary batteries and currents are used by the societal energy systems.

3. COSMIC ORIGIN OF EXERGY RESOURCES

It may be natural to ask for the origin of the Sun/space battery. How was it charged? The answer is that two disequilibria developed in the early universe, one nuclear and one gravitational. Both were related to the expansion of the universe and with this the changing equilibrium conditions (and with this the maximum possible entropy).

According to the standard theory of the early universe, nuclear matter existed as nucleons and was in thermal equilibrium 0.01 s after the Big Bang. The universe continued to expand and when the temperature passed below the dissociation temperatures for ordinary nuclei, the possible equilibrium state for nuclear matter became iron nuclei (^{56}Fe). At that time, approximately $\frac{1}{8}$ of the nucleons were neutrons; the rest were protons. Only the first steps of nuclear reactions could take place under the prevailing conditions. During the fourth minute, neutrons joined with protons to form helium (^{4}He). After this, almost all nuclear matter was in the form of helium $\left(\frac{1}{4}\right)$ and the remaining hydrogen, i.e., protons $\left(\frac{3}{4}\right)$. The conditions needed for continued nuclear reactions were not prevalent. The nuclear fuel of the early universe was thus saved for later burning in stars. During the first hours of the universe, through the cooling by cosmic expansion, the difference widened between the actual state of nuclear matter and the equilibrium state. The nuclear matter had become "supercooled" through cosmic expansion and nuclear exergy had been created.

In the early universe, matter was suspended in radiation. When the temperature dropped below the ionization temperature of helium and hydrogen a few hundred thousand years after the Big Bang, neutral atoms could form and radiation became more weakly coupled to matter. Then, gravitational clumping was possible and gravitational exergy was created. Matter could assemble on various levels to become galaxies or solar systems. Stars could start to form and in them, energy was converted into radiation through nuclear burning, which produced the various chemical elements. Matter thrown out from exploding stars was reassembled in later generations of stars, such as the Sun. What now remains of the original radiation is the 3 K cosmic background radiation.

With self-organizing systems, the old clockwork metaphor of physical science had to give way to a new view of dynamics, supported by a vastly increased capacity to study processes through computer simulations. The clockwork itself is a dissipative (entropy-producing) system: when it runs down, it converts mechanical energy into heat. Even if fundamental physical laws, as in quantum field theory, may be elegantly summarized in a principle of least action, actual physics is still indeterministic due to uncertainties in initial conditions and external influences. The old deterministic way of viewing physics may have had too strong an impact on other sciences, providing them with obsolete ideals.

4. ENERGY IN SOCIETY

It is of interest to consider the main uses of exergy in society. They can be summarized as follows:

1. to replace losses that occur through
   i. leakages (e.g., heat leaking in or out through roofs or walls),
   ii. friction (e.g., in machines or vehicles), and
   iii. material degradation and dispersion (through wearing and tearing and corrosion); and
2. to steer irreversible processes of flows from one state to another (e.g., matter moving from high to low pressure or thermal energy moving from hot to cold).

With this kind of description, the places to look for improved efficiency become more obvious.

Physical bookkeeping, as described above, is then of interest for analyzing the efficiency of a system or for studying a system as part of a larger system, such as a community as part of an ecosystem in which it is embedded. In the 1970s, Robert Ayres developed physics-based principles and schemes for book keeping of the turnover of energy and materials in society.

Human society is a major component in Earth's cycles of nitrogen and sulfur. Humans are systematically interfering with the carbon/oxygen flows by bringing fossil carbon into the atmosphere as CO_2. Thereby, the natural greenhouse effect that makes Earth inhabitable is enhanced, resulting in global warming and risks for major climatic changes.

Some material flows are crucial for the functioning of the ecosphere. Others are toxic or geophysically harmful and must be avoided. This means that, in material balance accounts, some flows must be more carefully guarded than others. There is no single quantity that can function as an overall index. Good overviews are needed, but the generalists cannot do away with the detailed knowledge provided by the specialists.

Robert Ayres also pointed out that energy and matter are never consumed; both remain after consumption. Even the matter of our bodies is continuously exchanged; it is a temporary loan to be ultimately paid back. Exergy, i.e., energy quality, is consumed, however, through entropy production.

What is useful is a flow of services that is enjoyed, directly or in the use of products. Thus, "product" is an intermediary concept; the concept of basic economic importance is "service." It is useful to think of temporary leasing of products (carrying energy and matter) as a kind of service being consumed.

Figure 3 shows a diagrammatic representation of one unit in societal resource use and service production. In Fig. 4, production of a displacement service (transport or storage) is similarly represented.

Strictly speaking, a production process is not completed until all the flows have been duly processed in society or nature. This means that in the case of fossil fuels, the relaxation mechanisms of the atmosphere and climatic systems are to be viewed as part of the process of producing energy services. This is anomalous in the sense that completion of the production process must continue long after the consumption of the energy services. The waste handling in the nuclear energy industry, as mentioned, is anomalous in the same way.

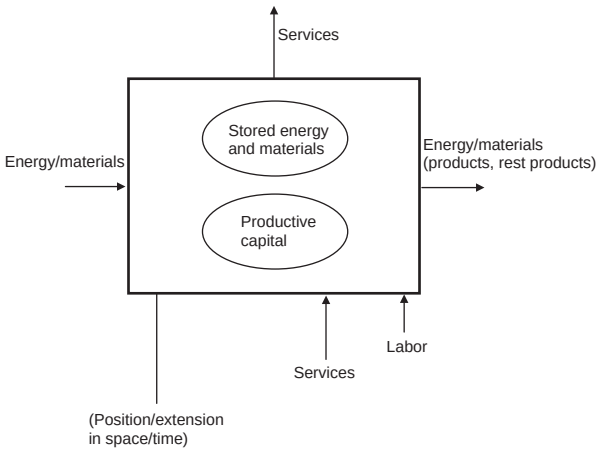


FIGURE 3 The metabolism of a society can be viewed as being built up of transformation activities like the one represented here. Physical flows are shown as horizontal arrows. Primary inflows come from natural systems and the final physical outflows also end up in natural systems. Only services are finally consumed. Products are viewed as intermediate entities. Leasing of a product over a time period is viewed as a service. Here labor is represented as nonphysical, since its human and economic aspects are more important than its physical aspects.

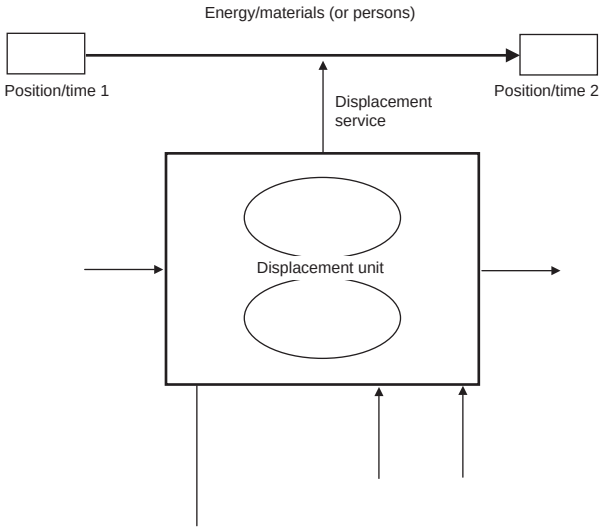


FIGURE 4 Diagram of a transport or storage (displacement) production unit, providing a transport or storage service.

Exergy was noted above as a resource measure. It is very general but still it cannot serve as a total resource measure, because it does not include all important aspects of the actual context. Exergy is needed to run everyday systems that make bodies, homes, and societies function. It is used (and most of it is consumed) for both maintenance and buildup.

Sadi Carnot's classical problem of maximizing the work output from heat can be viewed as an engineering problem with a physical solution but also as an economization problem. Thus, exergy expresses thermodynamic value but can also carry economic value.

To determine the space available for the totality of human activities (Daly's issue of scale, a reasonable size of one's ecological footprint or environmental space), one must analyze how the global society as a physical and ecological system can be sustainably embedded in the ecosphere. The global society is clearly far from static. A more just distribution of resources involves expansion, which must be compensated for by structural changes and increased efficiency ("factor 4" or "factor 10").

Very important discussions are going on in connection to the United Nations system, United Nations conferences and international negotiations, as well as in environmental science and in development research and in non-governmental organizations working with these issues. Consumers' responses to environmental and fair-trade rating of products have taken global issues into everyday life.

As mentioned above the exergy flow into Earth through its radiation exchange (Sun/space) is vast. There are large areas where solar energy could be efficiently harvested, in areas where people live as well as in uninhabited areas. Hydrogen can be expected to become a light and convenient energy carrier. The amount of exergy used in human society could be collected by photovoltaics in a rather limited desert area. The transition from fossil fuels to solar energy is at least physically feasible.

From the 2002 United Nations Conference on Environment and Development in Johannesburg, South Africa, the Swedish Ministry of Environment summarizes the discussion within the Forum on Science, Technology, and Innovation for Sustainable Development as follows (translation from Swedish by K.-E.E.):

> *From frontline research it is reported that it is possible to develop the technologies needed for the whole of humankind to enjoy a high standard of living. Access to clean water and good sanitation, clean energy, food and good health conditions for everybody is totally feasible. In this development, one must take as points of departure the ecological systems that form the basis for all this. Research and new investments will be needed. Sufficient amounts of capital and other resources are available but have to be redistributed—a massive resource mobilization is necessary from the public sphere as well as the private sphere.*
>
> *—Kjellén et al., 2002, Summary*

It would be good to have a clearer picture of a future sustainable resource use technology and the action space allowed for, say, an average family in a just global society.

The movement away from dumping CO_2 into the atmosphere may for some time involve sequestering of CO_2 into underground deposits, but in the long run, the necessary decarbonization must take place through abandoning of fossil carbon and a transition to renewable primary energy.

It is worth noting that, apart from biomass, renewable energy leads not only to decarbonization but also to dematerialization by largely removing material energy carriers. Wind energy and solar energy can be harvested without anything that corresponds to hydropower dams or nuclear reactors, and when a material energy carrier is needed, it can be hydrogen. More sophisticated material quality management can be achieved through repair, reuse, and recycling and through substitution of organic materials and light metals for heavier metals.

It is important to note that several materials that are candidates for energy converters and energy storage (photovoltaics, fuel cells, and batteries) are problematic. Some are quite rare in Earth's crust; some have toxic effects and should not be allowed into the ecosphere. Fortunately, silicon, still a top candidate for photovoltaics, is both abundant and nontoxic. Exactly which technological combinations will prove sustainable in the long term cannot be foreseen as yet.

One can ask what the resource system of a just global society would look like and how it could be organized administratively and economically to serve the average family mentioned above. It should be possible to use physical, technical, and ecological constraints as inputs in economic models and deduce relative prices, probably quite different from today's prices.

5. ENERGY, VALUES, AND ECONOMICS

We have seen that exergy is basic for life and activities in nature and society. Exergy is a measure of value in physical terms. However, the more complex a system or a process becomes, the less can be said by physics. Even the term "complexity" is problematic. There are several definitions of complexity as a quantitative concept in information theory. These definitions capture important aspects, but there is no mature theory of complexity. For the understanding of complex systems and processes, one needs the life sciences, social sciences, and humanities. What about economics, in particular, energy economics?

Some years ago, in a draft of the economics chapter for a report from the Intergovernmental Panel on Climate Change (IPCC), in an estimate of the costs of climate change, a statistical life in a poor country was valued as $\frac{1}{15}$ of a statistical life in a rich country. Willingness to pay for life insurance was part of the basis for this finding.

This led to very strong criticism by many scientists and finally the discussion convinced the authors that a change was necessary in the principles for making cost estimates in a global perspective. Willingness (or rather ability) to pay for life insurance could no longer be considered a reasonable basis for value. The presently dominant ideology that attributes a fundamental significance to market and money was shaken, at least temporarily. If the IPCC draft had taken into account that the marginal value of money is greater for a poor person than for a rich person, then a factor of approximately 15 would have appeared and canceled the factor $\frac{1}{15}$.

All this shows that moral values can be deeply involved in dealing with economic evaluations of energy systems.

If the idea that economic value is tied to human welfare is taken seriously, then this has very strong implications. The global welfare deficit can be computed for the current very unequal income distribution with a conventional welfare function. The deficit can then be expressed as a cost in money terms, using money value at the average income level. The cost turns out to be on the same order of magnitude as the total global income or larger. Thus, if inequality were included in measures of economic performance, economic descriptions would change drastically.

This discussion shows that it is important to test the validity limits of reasoning and concepts. It is especially so with economics, considering the dominant position of economic rationality for decision-making on almost all levels.

An important study in this context attempted to arrive at an economic value of nature. This work showed that large modifications are needed in economics in order to include the value of natural systems. It also showed that the uncertainties are large. Economics also has difficulties in handling not only natural capital, but also human capital and social and cultural capital and the value flows associated with these.

Either such value aspects should be included in economic considerations or those value aspects should define the boundaries within which economic considerations are to be made. As the above example from the IPCC draft shows, pure economic considerations, not governed by human value, can become immoral. The conclusion is that physical and economic aspects are both very important, but one should be well aware of their limitations.

Efficiency is a topic where physics and economics meet. Carnot's efficiency problem opened a new branch of physics: thermodynamics; at the same time, it was an economic problem. Thermodynamic efficiency can be economically visible through an exergy price.

There are also other links. Distributional inefficiency is related to technical (and thermodynamic) inefficiency: poor people may not afford investments to improve efficiency and rich people may not feel the fees introduced to discourage them from waste. Organizational efficiency and social efficiency are also relevant aspects to be included in an efficiency analysis.

Justice between generations can be seen as an intertemporal efficiency. If the time anomalies in production described above are to be taken seriously, then, for consistency, there should be no discounting of the future; the *a priori* discount must be zero.

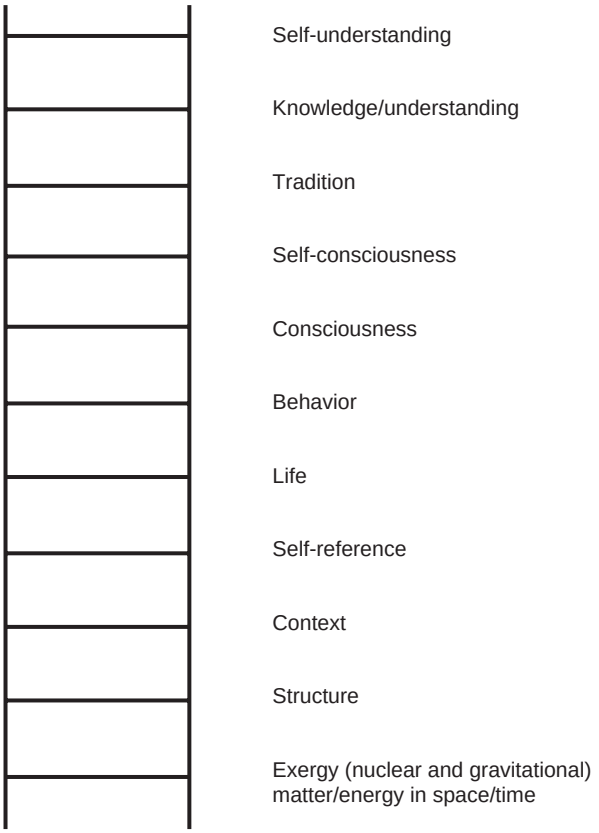


FIGURE 5 Diagram of an evolution ladder with emergent system characteristics.

6. THE EVOLUTIONARY PERSPECTIVE

Since the very early beginnings of the universe, transformations of energy and matter have involved the emergence of new systems and phenomena. The history of the universe is a history of creation. Structure was created or became manifest in early clouds on a different scale, to become solar systems or galaxies.

The surface of Earth developed conditions favorable for the richness of carbon chemistry, a wealth of possible forms and processes of varying permanence and flexibility. Self-organization also came to involve self-reference, i.e., systems carrying blueprints of themselves. Accidental changes in such blueprints sometimes led to more efficient modified systems. Reproduction, life, variation, selection, and

evolution became possible. Evolution is a process of accumulating information, selected on the basis of its functional advantages for reproduction. The story is as represented in Figs. 1 and 2, repeated on many levels and many scales. Information capacity (exergy) became coded into relatively stable information carrying a meaning.

The ladder in Fig. 5 is an attempt to identify steps in evolution, starting with the cosmic origin of the universe in the Big Bang. The steps may be quite ambiguous, but they are meant to convey the evolutionary perspective of modern science. We human beings may be seen as organs, through which the universe/cosmos can view itself and reflect over itself.

Not knowing of evolution in other parts of the universe, it is reasonable to ask what evolutionary step could or should be taken by humans on Earth. Information and communication technology reaches almost everywhere on Earth and is becoming global in character. Technically, humankind is one unit. Ecologically, we humans are strongly dependent on one another in the treatment of Earth. However, economically, politically, and ideologically, we are still divided. In the quoted summary from Johannesburg, adequate resources for a good life for everybody, including clean energy, is described as a realistic possibility. The next evolutionary step on the ladder would therefore be to unite, to develop and share a common self-understanding, and to act together and organize a good life on Earth.

SEE ALSO THE FOLLOWING ARTICLES

Complex Systems and Energy • Earth's Energy Balance • Economic Thought, History of Energy in • Entropy and the Economic Process • Evolutionary Economics and Energy • Exergoeconomics • Exergy Analysis of Energy Systems • Information Theory and Energy • Input–Output Analysis • Thermodynamics and Economics, Overview • Thermodynamics, Laws of • Work, Power, and Energy

Further Reading

Andersson, B. A., and Jacobsson, S. (2000). *Energy Policy* **28,** 1037.

Daly, H. E., Cobb, J. B., Jr. (1989). "For the Common Good." Beacon, Boston, MA.

Eriksson, K. E. (1994). *Ric. Econ.* **48,** 341

Eriksson, K. E. (2002). Lecture notes from the 4th Regional College on Modelling, Simulation, and Optimization, Accra, Ghana, 2002, National Centre for Mathematics and Science, Accra, Ghana.

Eriksson, K. E., Islam, S., and Skagerstam, B. S. (1982). *Nature* **296,** 540.

Eriksson, K. E., Lindgren, K., and Månsson, B. (1987). "Structure, Context, Complexity, Organization." World Scientific, Singapore.

Holmberg, J., Lundqvist, U., Robèrt, K. H., and Wackernagel, M. (1999). *Int. J. Sust. Dev. World Ecol.* **6,** 17.

Kjellén, B., Näslund, S., and Olsson, K. (2002). "Forum on Science, Technology and Innovation for Sustainable Development, Johannesburg, 2002, Swedish Ministry of Environment, Internal Report" [in Swedish].

Meyer, A. M., *et al.* (1995). *Nature* **378,** 433.

Råde, I., and Andersson, B. A. (2001). *J. Power Sources* **93,** 55.

Rees, W. E., and Wackernagel, M. (1994). *In* "Investing in Natural Capital: The Ecological Economics Approach to Sustainability" (A. M. Jansson, M. Hammer, C. Folke, and R. Costanza, Eds.), p. 362. Island Press, Washington DC.

Robèrt, K. H., Holmberg, J., and von Weizsäcker, E. U. (2000). *Greener Manage. Int.* **31,** 25.

von Weizsäcker, E. U., Lovins, A. B., and Lovins, L. H. (1997). "The Factor four—Doubling Wealth, Halving Resource Use." Earthscan, London.

Wall, G. (1985). "Exergy—A Useful Concept," Ph.D. thesis. Department of Physical Resource Theory, Chalmers, Göteborg, Sweden.

Plastics Production and Energy

MARTIN PATEL and NITIN MUTHA
Utrecht University
Utrecht, The Netherlands

1. Introduction
2. Method and Scope
3. Results
4. Discussion
5. Conclusion

Glossary

bio-based plastics Plastics that are fully or partially produced from biomass-derived feedstocks (renewable raw materials).
cradle-to-factory gate System covering all process steps from the extraction of resources to their transformation (e.g., in refineries) and the conversion to monomers and polymers.
cradle-to-grave Same as cradle-to-factory gate but including also all activities related to waste management (due to the diversity of application areas, the use phase is often excluded in comparisons for polymers).
finite energy Nonrenewable energy sources (i.e., mainly fossil and nuclear fuels).
petrochemical plastics Plastics that are produced from petrochemical feedstocks.
plastics Macromolecular materials produced from petrochemical feedstocks or renewable raw materials by means of polymerization.

Worldwide, the production of polymers totaled approximately 180 million tonnes in 2001. Of all man-made bulk materials, the total energy input for plastics production is estimated to be fifth after that for cement, steel, paper/board, and bricks/tiles. In the past few decades, plastics production outpaced that of all other bulk materials, with the average worldwide growth rate for plastics 5.8% per year between 1990 and 2000. During the same period, the growth rate of plastics production exceeded gross domestic product growth [3.6% per annum (PA)] by more than 2% PA. Plastics enjoy high growth rates both in low-income and high-income countries (e.g., approximately 40% PA in Western Europe versus approximately 300% PA in India between 1990 and 2000). Given the very low per capita consumption rates in developing countries (India, 3.5 kg/capita; China, 12 kg/capita; world, 24 kg/capita), plastics production can be expected to continue to grow considerably in the next decades.

1. INTRODUCTION

Among the bulk plastics, use of polyolefins has increased most in the past decade, with particularly high growth rates for polypropylene (PP) and linear low-density polyethylene (LLDPE). Polystyrene (PS) and polyvinyl chloride (PVC), on the other hand, have grown at a rate approximately the world average. Table I provides an overview of products that are typically manufactured from these and other types of polymers.

The overwhelming share of all plastics is produced from petrochemical feedstocks (more than 99.9%). This is the consequence of the growth of the petrochemical industry in the early 20th century, whereas the first man-made polymers were derived from biomass resources. However, since the 1980s and 1990s, increased use of bio-based polymers has been observed in certain application areas. In 1998, production of bio-based polymers in the European Union (EU) was approximately 25,000 tonnes. The total market is projected to grow autonomously to approximately 500,000 tonnes until 2010 [28% per annum (PA), 1.5% market share] and to approximately 1 million tonnes if supportive policies and measures are implemented (36% PA, 3% market share). So far, there are no policies and measures at the EU level, whereas national legislation provides advantages for bio-based polymers in certain areas (e.g., due to the packaging ordinance in Germany).

 © 2004 Elsevier Inc. All rights reserved.

TABLE I

Overview of Typical Applications by Types of Polymers

Polymer	Applications
Epoxy resin	Industrial floorings, foams, fiber-reinforced plastics
High-density Polyethylene	Moulded products, heavy-duty films
Linear low-density polyethylene	Films (food packaging), rotational moulded products (e.g., tanks)
Low-density polyethylene	Food packaging
Melamine formaldehyde resin	Decorative laminates, lacquers
Phenol formaldehyde resin	Laminates for printed circuit boards
Polyacrylonitrile	Fibers
Polyamide (nylon)	Fibers, plastic gears, tire cord
Polycarbonate	Safety goggles, safety transparent shields, machinery housings
Polyethylene terephthalate	Blow-moulded bottles
Polypropylene	Injection-moulded products (e.g., technical parts for washing machine), films (e.g., for cigarette wrappers), monofilaments, ropes
Polystyrene	Moulded containers, lids, jars, bottles, foamed plastics
Polyurethane	Foams, coatings, adhesives and elastomers
Polyvinyl chloride	Extrusion products (wire and cable, conduits, profiles)
Unsaturated polyesteralkyd resin	Fibers, coatings, adhesive
Urea formaldehyde resin	Paper coating, adhesive

Most of the bio-based plastics that are currently on the market are biodegradable, but there are also increasingly more bio-based materials that are not biodegradable (e.g., composites of natural fibers and petrochemical polymers).

Starch polymers represent by far the largest group of commercially available bio-based plastics. Other prominent bio-based plastics are polylactic acid (PLA), polytrimethylene terephthalate, and polyhydroxyalkanoates (PHA).

Due their importance, starch polymers, PLA, PHA, and natural fiber composites, which are all manufactured at an industrial scale, are discussed in this article in addition to conventional petrochemical plastics. The first three materials are biodegradable; this is not the case for the natural fiber composites and the petrochemical plastics covered.

Given the dynamic developments discussed previously and the variety of plastics on the market, it is of great relevance to provide an overview of the energy use related to plastics. If the total life cycle of plastics is taken into account—comprising all process steps from the extraction of resources to polymer production and plastics processing and waste management—plastics account for approximately 4% of the total primary energy use in high-income countries (4.2–4.4% in Western Europe; data exclude the use phase). More than 90% of this is required for all processes related to plastics production, on which we will focus in this article (approximately 4.0% of total fossil fuel use in Western Europe). In the future, improved waste management, including recycling, reuse, and advanced energy recovery, may contribute substantially to materials and energy savings. These options and the attendant potential for energy saving are not discussed in this article.

2. METHOD AND SCOPE

This article is based on a review of energy analyses and life cycle studies on petrochemical and bio-based plastics. The studies reviewed basically follow the principles of life cycle assessment (LCA), which is a standardized method to quantify environmental impacts. In practically all LCA studies, energy use is included in the set of indicators, which typically also includes the release of greenhouse gases, the acidification and eutrophication potentials, waste generation, and other environmental impacts.

As in most LCA studies, we focus on the use of finite (i.e., nonrenewable) energy. This comprises fossil fuels and, to a lesser extent, nuclear fuels. Large amounts of renewable resources (especially biomass derivatives) are used for the production of bio-based plastics, but the quantities used are not discussed in this article. We focus on finite energy because of the limited resource perspective, the high dependence on foreign supplies, and the generally higher environmental impacts than in the case of renewable energy.

As indicated previously, we focus on energy use for plastics production, also referred to as cradle-to-factory gate analyses. This type of analysis includes all processes, including the extraction of energy resources, their transformation (mainly in refineries and power stations), the conversion of the feedstocks to monomers and polymers, and the manufacture of the required auxiliaries (Fig. 1, gray box). Processing of polymers to final products, their use phase, and

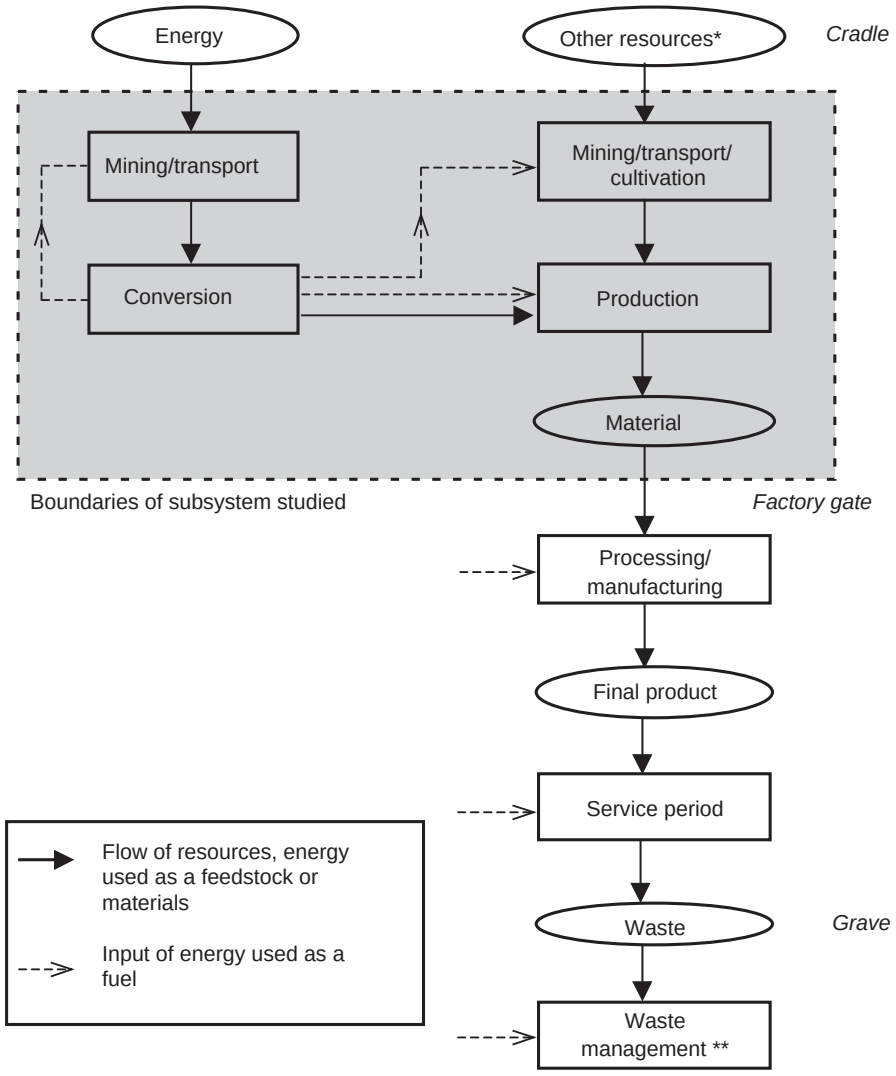


FIGURE 1 Overview of the entire system and the subsystem discussed in this article (gray box). *Energy analyses do not account for the consumption of resources other than energy, but they do include the energy requirements to process these "other resources." **Landfilling, incineration, and recycling (not shown).

the waste management stage are not included in the system boundaries (in contrast, studies that do include the latter are referred to as "cradle-to-grave analyses).

Fossil fuels are used both as a feedstock for chemical reactions and to cover the energy requirements of the production processes and transportation (Fig. 1). In the case of bio-based plastics, the biomass-derived inputs are almost exclusively used as a feedstock but efforts are being made to integrate waste and by-products in the energy supply.

As a basis for all comparisons, 1 tonne of plastics in the form of pellets (granules) has been chosen as the functional unit. This is a convenient basis since it allows the comparison across types of polymers, but it also has some drawbacks.

3. RESULTS

A distinction can be made between the use of energy resources as feedstocks and as process energy (in the form of electricity, steam, or direct fuel use). Reporting of feedstock energy is complicated by the fact that in some processes, part of the feedstock is used to fuel the process (e.g., in steam crackers for olefin production and in ammonia production). As a consequence, feedstock energy use should be

defined precisely in every study, but this is rarely done (if feedstock and process energy are distinguished at all). For this reason, we do not distinguish between these two types of energy use and exclusively report total energy use (cradle-to-factory gate energy).

However, there is a relatively safe way to address the issue by making use of information on the heating value of plastics. The heating value represents the amount of feedstock energy that is physically fixed in the plastics (Table II). Hence, the remainder (i.e., the difference between cradle-to-factory gate energy and the heating value) can be defined as the process energy consumed in the various production steps.

In the following sections, we discuss energy requirements for the production of petrochemical plastics and bio-based plastics.

3.1 Petrochemical Plastics

Table III provides an overview of cradle-to-factory gate energy requirements of 21 petrochemical primary plastics (pellets and granules; compare Table I for typical applications). The studies identified differ with regard to the geographic scope (only high-income countries are represented) and the reference year (the oldest source dates back to 1980). There is also a major difference concerning the coverage of types of plastics: Whereas some of the sources specifically focus on plastics production and/or recycling, others cover the most important building materials or even all major bulk materials.

A key source used frequently in Europe (also in LCA software tools) is the Eco-Profiles published by the Association of Plastics Manufacturers in Europe (APME). Table III shows two data sets from APME, one of which was published in 1999 and the other pre-1999. Other widely disseminated data sets are those released by the U.S. Office of Industrial Technology (OIT). It is interesting to note that substantial corrections have been made in the revised APME data set and that the values in the two data sets published by the OIT differ considerably (Table III). The difference in values across the other sources is even larger.

We calculated the standard deviation for those plastics for which at least 7 independent sources are available (polyethylene, polyvinyl chloride, polystyrene, polyethylene terephthalate, polyurethane, and nylon 6,6) and determined values ranging between 18 and 23% of the arithmetic mean. Assuming a normal distribution, the cradle-to-factory gate energy requirements have a probability of 68% within a range of $\pm 23\%$. For the remaining 14 plastics, there are less than 7 independent sources. The data from approximately half of these fall into the uncertainty range established for the 7 plastics, whereas the spread is larger for the others. Due to the low number of independent sources, quantification of the uncertainty range for the 14 plastics is hardly possible.

Especially for sources covering a larger number of materials, it is possible to make a statement about the relative position of the values: The data reported by APME (both data sets) are in the medium range, whereas the values published by Kindler and Nikles are partly in the medium range and partly on the higher side (PE is an exception since it is on the lower side). Data from studies by Alcorn and by Shutov are on the higher side, whereas those published by Patel and by the Plastic Waste Management Institute, Japan, are in most cases lower than those in the other sources (with the exception of nylon 6,6).

For the data published by the APME and by Patel, relatively detailed information is available about the assumptions and system boundaries that allow discussion of the reasons for the discrepancy. Except for ammonia, nylon-6,6, polyethylene, and polyvinyl chloride, the energy requirements determined for Germany are generally lower than those determined by the APME for Western Europe. The main reasons are as follows:

TABLE II

Heating Value of Bio-Based and Petrochemical Polymers[a]

Polymer	Lower heating value, GJ/tonne (dry matter)
Starch polymers	13.6
Polyhydroxybutyrate	22.0
Polyhydroxyvalerate	25.0
Polylactic acid	17.9
Lignin (picea abies)	24.2
China reed	18.0
Flax	16.3
Hemp	17.4
Kenaf	16.5
Polyethylene	43.3
Polystyrene	39.4
Polyethylene terephthalate	22.1
Polyvinyl chloride	17.9

[a]Heating values calculated according to Boie, compare Reimann and Hämmerli, 1995.

- The data for Germany refer to the net calorific value (NCV; lower heating value), whereas the APME publications use gross calorific values (GCVs; higher heating value). The difference is within 10%.
- The APME data are based on surveys for all of Western Europe. They may therefore include some smaller, older plants that are less efficient than the average of all units in Germany in the mid-1990s.
- The APME data accounts for the various shares of purchased versus autogenerated electricity, including the respective efficiencies for the sites and the public grids in the various countries covered. In contrast, exclusive autogeneration of power in efficient combined heat and power (CHP) plants has been assumed for plastics production in Germany.
- The efficiencies for steam generation assumed in the APME publications are smaller than the average assumed in the calculations for Germany.

There are also several aspects that have not been taken into account in the calculations for Germany but that are covered by the APME data:

- For specific processes, data were only available for the selectivities and not for the yields of the products. In these cases, the calculations are based on selectivities instead of yields, thus assuming a conversion of 100% (yield = selectivity × conver- × conversion). Consequently, material efficiency is slightly overestimated and therefore the results for the energy requirements in Germany are somewhat underestimated.
- The assumed losses in the steam system (7.5%) in plastics production in Germany may be too low, as indicated by the literature data on the losses related to degasification of the boiler feed water, elutriation of the boiler, and noncovered recycling of condensate indicate. On the other hand, the German Association of Chemical Industry estimates that the average steam losses are 2.5–4% (G. Thomas, personal communication, 1999).
- Facilities outside the battery limits (e.g., flue gas scrubbers and sewage treatment plants) are not included in the system boundaries and, consequently, they are not included in the data for Germany.
- Ancillary materials (e.g., packaging materials or lubricants), catalysts, and other materials used in small amounts have not been taken into account in the calculations for Germany.
- It is possible that the operating rates assumed in the calculations for Germany are higher than in reality. Moreover, the process data used for Germany refer to continuous operation with no maintenance intervals (i.e., they exclude energy-intensive reactor start-ups and shutdowns).
- Transport energy used in the extraction and supply of resources has been accounted for (Fig. 1). However, due to the lack of detailed data, transport between all the subsequent conversion steps was not included in the calculations for Germany. Since the majority of the products studied were transported in pipelines, which does not consume a large amount of energy, and since transport energy generally only represents a small share of the total energy consumption, the errors resulting from this procedure can be assumed to be negligible.

To summarize, it may well be that the data for Germany underestimate the real energy requirements. However, the different choices made regarding the aspects listed previously and the differences in the years and regions analyzed largely explain the difference compared to the APME data. The results for Germany are therefore considered to be representative estimates for this country in the mid-1990s.

Despite the ranges in values in Table III, most studies confirm that relatively little energy is required to manufacture the bulk plastics PE and PVC. In contrast, the energy required for the manufacture of engineering and other plastics can be more than twice as high.

3.2 Bio-Based Plastics

Contrary to petrochemical plastics, the feedstocks used to produce bio-based plastics are derived from agricultural crops (e.g., maize, wheat, or sugar beet) or from suitable waste streams (e.g., from the dairy industry). Regarding petrochemical plastics, processing, cultivation, and transportation almost exclusively rely on finite energy sources, particularly fossil fuels. For bio-based plastics, the cultivation of crops (or the supply of suitable waste streams) results in an energy demand that is absent in the case of petrochemical plastics. Moreover, feedstock preparation and product separation may be clearly higher in the case of bio-based plastics due to relatively low yields and high water content in the raw material and product streams. It is thus not obvious from the outset that cradle-to-factory gate energy requirements for bio-based plastics are lower compared to those for petrochemical plastics.

However, Table IV shows that this is the case for nearly all of the bio-based plastics studied: thermoplastic starch (TPS), starch copolymers [blends of TPS with polycaprolactone (PCL) or polyvinyl

TABLE III

Energy Requirements for the Production of Petrochemical Plastics[a]

Plastic	Total primary energy use (cradle-to-factory gate) (GJ/t)											
	APME (1999)[b] Western Europe	Pre-1999 APMEb and BUWAL Western Europe	Curlee and Das (1998) USA	Howard (1995) UK	Kindler and Nikles (1980) Germany	Alcorn (1998) New Zealand	OIT (2002) USA	OIT chemical industry profile reports USA	Patel (1999) Germany	PWMI Japan (1999) Japan	Shutov (1999) USA	Worrell *et al.* (1994) The Netherlands
Epoxy resin	140.7	141.7[c]							107.1			
Melamine resin									79.9			
Nylon-6					156			124.3	122.7			
Nylon-6,6	142.3	143.6[d]			154		241.7	135.1–157.4	163.0		147.6	
Phenol formaldehyde resin					82				60.0		89.2	
Polyacrylonitrile									69.3			
Polyacrylate	PMMA, 102.3	PMMA, 111.7[e]			PMMA, 91				82.6[f]			
Polycarbonate	116.8	116.3[g]			107				80.3		158.5	
Polyethylene (PE)				85				74.1–80.5	64.6			
Polyethylene, high-density	*79.9*	*81.0*[b]	*84.8*		*70–72*	*103*	*131.5*			*65.7*	*98.1*	
Polyethylene, low-density	*80.6*	*88.6*[b]	*89.5*		*69*	*103*	*136.0*			*69.2*[i]	*103.2*	*67.8*
Polyethylene, linear low-density	*72.3*	*83.0*[b]					*121.1*					
Polyethylene terephthalate (PET)	76.2–77.5[j]	81.7–83.8[j]	113.2		84–108	106			59.4	60.0[k]	106.4	78.2
Polypropylene			79.5		73	64[l]	126.0	54.5			95.3	63.2
Polystyrene	86.7[m]	86.3[n]	79.7	96	80–82			100.0–104.5	70.8	68.5	117.1	82.7
Polyurethane (PUR)	104.4–104.9[o]	104.4–104.9[p]		65	98	74			75.5[q]		73.7	
Polyvinyl acetate		66.8[r]							58.2			
Polyvinyl chloride	*56.6*[s]	*56.7*[t]	59.5	59	53	70	34.6	44.8	53.2	45.4	79.0	52.4

Synthetic rubber			64.4[u]	SBR, 77.5
Unsaturated polyester/ alkyd resins	78	93.7–100.4	64.5[v]	
Urea resin			50.3	

[a] Cradle-to-factory gate nonrenewable energy including both feedstock energy and process energy. Abbreviations used: APME, Association of Plastics Manufacturers in Europe; BUWAL, Bundesamt für Umwelt, Wald and Landschaft (Swiss Agency for the Environment, Forests, and Landscape); OIT, Office of Industrial Technologies; PWMI, Plastic Waste Management Institute; PMMA.

[b] APME also provides data on renewable energy that are outside the scope of this article (see text). However, since the share of renewables is very low (mostly 0.3–0.4%), these quantities have not been deducted in this column. This has the advantage that the data listed here can be quickly compared with the original source without further calculations.

[c] APME Eco-profiles, Report 12 (1997).

[d] APME Eco-profiles, Report 15 (1997).

[e] APME Eco-profiles, Report 14 (1997).

[f] Modeled as polymethyl acrylate.

[g] APME Eco-profiles, Report 13 (1997).

[h] APME Eco-profiles, Report 3 (1993).

[i] LLDPE included.

[j] Lower value for amorphous resin, higher value for bottle grade PET.

[k] Bottle-grade PE.

[l] In some tables of the same source reported to be 95.4 GJ/t.

[m] Figure for GPPS (general purpose polystyrene). For comparison, HIPS (high impact polystyrene), 92.2 GJ/t.

[n] Figure for GPPS. For comparison, HIPS, 90.7 GJ/t.

[o] Lower value for flexible PUR foam, higher value for rigid PUR.

[p] Lower value for flexible PUR foam, higher value for rigid PUR. APME Eco-profiles, Report 9 (1997).

[q] 50% rigid polyurethane, 50% flexible polyurethane.

[r] APME Eco-profiles, Report 6 (1994).

[s] Figure refers to suspension PVC, which is the general-purpose grade and is used for most mass applications. For comparison, bulk PVC (used for specific types of hard sheets and bottles); 63.1 GJ/t; emulsion PVC, 67.7 GJ/t.

[t] APME Eco-profiles, Report 6 (1998). Figure refers to suspension PVC, which is the general-purpose grade and is used for most mass applications. For comparison; bulk PVC (used for specific types of hard sheets and bottles); 63.2 GJ/t; emulsion PVC, 68.0 GJ/t.

[u] Weighted average of the German production of SBR (styrene butadiene rubber), BR (butyl rubber), NBR (nitrile butadiene rubber), and CR (chloroprene rubber) in the mid-1990s.

[v] Modeled as resin synthesized from phthalic anhydride (1 mol) and propylene glycol (2 mol).

TABLE IV

Energy Requirements for the Production of Bio-Based Plastics[a]

Type of plastic	Cradle-to-gate finite energy use (GJ/t)[b]	Reference
Bio-based plastics (pellets)		
TPS	25.4	Patel *et al.* (1999)
TPS	25.5	Dinkel *et al.* (1996)
TPS	25.4	Dinkel *et al.* (1996)
TPS (maize starch + 5.4% maize grit + 12.7% PVOH)	18.9	Würdinger *et al.* (2002)
TPS + 15% PVOH	24.9	Patel *et al.* (1999)
TPS + 52.5% PCL	48.3	Patel *et al.* (1999)
TPS + 60% PCL	52.3	Patel *et al.* (1999)
Starch copolymer foam grade	32.4–36.5	Estermann *et al.* (2000)
Starch copolymer film grade	53.5	Estermann and Schwarzwälder (1998)
PLA	54	Vink (2002)
PHA by fermentation	81	Gerngross and Slater (2000)
PHA, various processes	66–573	Heyde (1998)
Natural fiber composites		
ABS copolymer[c]	117	Wötzel *et al.* (1999)
Hemp fiber/epoxy composite[c]	89	Wötzel *et al.* (1999)
Glass-fiber reinforced polypropylene[d]	93	Corbière-Nicollier *et al.* (2001)
China reed reinforced polypropylene[d]	61	Corbière-Nicollier *et al.* (2001)
Petrochemical polymers		
HDPE	79.9	APME (Boustead) (1999)
LDPE	80.6	APME (Boustead) (1999)
LLDPE	72.3	APME (Boustead) (1999)
PET (bottle grade)	77	APME (Boustead) (1999)
PS (general purpose)	87	APME (Boustead) (1999)
Expanded PS	84	APME (Boustead) (1999)
Petrochemical copolymers		
PCL	83	Estermann and Schwarzwälder (1998)
PCL	77	Kopf (1999)
PVOH	102	Estermann *et al.* (2000)
PVOH	58	Kopf (1999)

[a] Cradle-to-factory gate nonrenewable energy including both feedstock energy and process energy. Complete bibliographic information on the references cited in this table can be found in Patel *et al.* (2003). Abbreviations used: TPS, thermoplastic starch; PVOH, polyvinyl alcohol; PCL, polycaprolactone; PLA, polylactides; PHA, polyhydroxyalkanoates; ABS, acrylonitrile, butadiene, styrene; HDPE, high-density polyethylene; LDPE, low-density polyethylene; LLDPE, linear low-density polyethylene; PET, polyethylene terephthalate; PS, polystyrene; APME, Association of Plastics Manufacturers in Europe.

[b] Total of process energy and feedstock energy. Finite (nonrenewable energy) only; i.e., total of fossil and nuclear energy. In the cradle-to factory gate concept, the downstream system boundary coincides with the output of the polymer or the end product. Hence, no credits are ascribed to potential benefits in the use phase to valuable by-products from waste management (steam, electricity, and secondary materials).

[c] Application studied: Interior side panel for passenger car; ABS stands for acrylonitrile-butadiene-styrene.

[d] Application studied: Transport pallet.

alcohol (PVOH), polylactides (PLA), and the natural fiber composites. For polyhydroxyalkanoates (PHA), the energy use strongly depends on the process and feedstocks used and ranges between values that are somewhat lower to those that are clearly higher than for petrochemical plastics. To facilitate direct comparison, Table IV also provides the energy requirements for the most important petrochemical polymers and for PCL and PVOH, which are both used as copolymers for starch plastics. Life cycle practitioners consider the data for PCL and PVOH to be subject to major uncertainties.

Most of the materials listed in Table IV can be produced according to the current state-of-the-art. Future options (e.g., the production of PHA in crops and the medium- and long-term production processes for PLAs) are not listed in Table IV. As an exception, technical problems still need to be resolved for pallets reinforced with china reed (material stiffness).

It can be concluded from Table IV that the cradle-to-factory gate energy requirements of pure TPS are 65–75% below those for PE. For starch polymers with high amounts of petrochemical polymers, the energy requirements are approximately 30% lower than those for polyolefins. However, the applications for pure starch polymers and blends with small amounts of petrochemical copolymers are limited due to inferior material properties (especially sensitivity to moisture). With larger amounts of petrochemical copolymers, polymers with wider application profiles can be produced. This allows to obtain higher energy savings in absolute terms while the marginal energy savings decrease with higher shares of petrochemical copolymers.

For PLA, data are only available from a source (Cargill Dow), according to which cradle-to-factory gate energy requirements are 20–30% less than those for PE.

The results reported for PHA vary greatly. Cradle-to-factory gate energy requirements in the best case (66.1 GJ/t) are 10–20% lower than those for PE. The upper range in Table IV refers to substrate supply from fossil-based feedstocks (e.g., methanol).

It can be concluded that of all bio-based plastics studied, starch polymers perform best in environmental terms under the current state-of-the-art, with some differences among the various types of starch polymers. Compared to starch polymers, the environmental benefits seem to be smaller for PLA. For PHA, the environmental advantage seems to be very small compared to that for conventional plastics in the best case; in the worst case, it can be substantially higher. For both PLA and PHA, the production method and the scale of production can influence decisively the cradle-to-factory gate energy requirements.

The extent to which natural fibers (e.g., hemp and flax) can replace fiberglass (which is heavy and energy intensive to produce) determines the net environmental benefits. Interesting possibilities are also available.

Since most of the data in Table IV refer to the current state-of-the-art, technological progress, improved process integration, and various other possibilities for optimization will likely provide more favorable results for bioplastics in the future.

4. DISCUSSION

In the preceding section, all data were compared on the basis of 1 tonne of plastics. Although this is a convenient choice, it suffers from the drawback of not taking differences in material properties and the resulting parameters at the end use level into account. For cradle-to-factory gate analyses, this mainly concerns (i) materials processing, in which the amount of material required to manufacture a certain end product and the energy use in the processing step may differ depending on the type of plastic (e.g., bio-based vs petrochemical plastics), and (ii) transportation energy, which can be substantial for end products with a low bulk density such as packaging chips (so-called loose fill).

On the other hand, these aspects can only be fully taken into account by analyses dedicated to specific end products (e.g., plastic bag), whereas the choice of one mass unit of polymers as a functional unit allows a preliminary conclusion about the advantages or disadvantages in terms of energy use. If, for example, the energy requirements for a given polymer are relatively high at the level of primary materials (pellets), there is a good chance that this will also be true at the end product level.

The studies reviewed in this article reflect the situation in different countries. The types of feedstocks used and the processes applied can vary from country to country or from continent to continent. For example, olefin production in steam crackers is mainly based on naphtha in Europe and Japan, whereas in the United States natural gas-based ethane crackers dominate. Differences in processes and feedstocks may result in different yields and energy efficiencies. Due to the competition in the sector, the differences in efficiencies of the production processes can be assumed to be relatively low across high-income countries. This may be less the case for power production: Here, the requirements of (finite) energy resources depend to a large extent on the type of fuel used (e.g., coal with relatively low efficiencies vs natural gas with high ones), the share of CHP, and the percentage of renewable energy (generally very low but high in Norway and Switzerland).

Direct comparison of energy data is complicated by the fact that both the NCV and the GCV are used in this type of analysis but that the choice is not always clearly reported. This also concerns the

inclusion or exclusion of ancillary materials, catalysts, and additional facilities such as sewage treatment plants that generally have a much lower impact on the results. This can also be assumed for different reference years as long as they do not differ by more than a decade. Larger discrepancies among the studies may accrue from different allocation procedures for processes with larger amounts of coproducts. Also in these cases, the concrete choices are hardly ever explained in detail.

An attempt should be made to consider these aspects when using cradle-to-factory gate energy data for further analysis and when interpreting the results and drawing conclusions. Moreover, whether the waste management system should also be included in the system boundaries (cradle-to-grave system) should be carefully considered.

5. CONCLUSION

By combining the cradle-to-factory gate energy data with current production data, the global primary energy use for the production of polymers is estimated to be approximately 15,000 PJ or 3.6% of the global primary energy use. Although the uncertainty of the production data is unknown, the specific energy requirements for industrialized countries have been estimated with a probability of 68% within a range of $\pm 23\%$. Since the energy efficiency in developing countries is generally lower, the global value can be considered a conservative estimate. In view of the high current and anticipated growth of plastics production, the attendant energy use is very likely to increase substantially. Basically, four strategies can be distinguished to minimize this growth:

1. Energy efficiency: The most straightforward strategy is increased energy efficiency in the petrochemical processes and in polymerization and increased use of production routes and intermediates with low overall energy requirements. For example, PVC, which was formerly manufactured from energy-intensive acetylene, is currently mainly being produced from ethylene. Research and development efforts are directed toward the use of ethane as a feedstock.
2. Material substitution: With advances in polymer science, it may become increasingly possible to replace energy-intensive engineering plastics (such as polyamides) with modified bulk polymers (e.g., polyolefins), which are more energy efficient to produce.
3. Waste management: Various types of recycling, the reuse of plastic components, and energy recovery can contribute to lower life cycle energy requirements and environmental impacts.
4. Bio-based plastics: As apparent from Section III, the substitution of bio-based raw materials for petrochemical feedstocks may provide an interesting potential for energy saving (Table IV).

Energy analysis can be a useful tool for establishing the relative benefits of specific measures of these strategies.

When conducting an energy analysis, the researcher is often confronted with the fact that different sources provide different values for a certain product and that the deviation between them can be substantial (Table III). This is generally due to different system boundaries (e.g., concerning the years and geographical regions studied and the technologies and processes covered). It is a prerequisite for a reliable analysis that the cradle-to-factory gate data used have been determined in a consistent manner. In contrast, comparative assessments that use energy data originating from various sources can easily result in distorted conclusions given the large data ranges shown in Table III (differences of a factor of two in extreme cases). These data ranges and the corrections made by the APME indicate that considerable uncertainties continue to exist despite the fact that petrochemical polymers are manufactured by use of mature technologies that are applied globally with only limited variations. On the other hand, there is good agreement concerning the relative position of the various plastics.

As a consequence, further analyses on cradle-to-factory energy requirements for plastics should be conducted. For the time being, it is suggested to use, wherever available, data sets that are valid for the country (e.g., Germany) or region (Western Europe) under investigation. In order to check the reliability of the results, it is recommended to conduct sensitivity analyses using other data sets that have been determined in a consistent manner for a different region than the one investigated. Alternatively, calculations on error propagation can be conducted.

SEE ALSO THE FOLLOWING ARTICLES

Aluminum Production and Energy • Cement and Energy • Glass and Energy • Material Efficiency and Energy Use • Material Use in Automobiles • Steel Production and Energy

Further Reading

Alcorn, A. (1998). "Embodied Energy Coefficients of Building Materials," 3rd ed. Victoria University of Wellington, Centre for Building Performance Research, Wellington, New Zealand. [http://www.arch.vuw.ac.nz/cbpr/index_embodied_energy.html]

Association of Plastics Manufacturers in Europe (1993–1998). Eco-profiles of the European plastics industry (various reports on different types of plastics). Prepared by I. Boustead for the Association of Plastics Manufacturers in Europe, Brussels, Belgium.

Association of Plastics Manufacturers in Europe (1999). "Eco-Profiles of Plastics and Related Intermediates." Association of Plastics Manufacturers in Europe, Brussels, Belgium. [(http://lca.apme.org].

Ayres, R. U. (1995). Life cycle analysis: A critique. *Resour. Conserv. Recycling* **14,** 199–223.

Curlee, T. R., Das, S. (1998). The viability of recycling plastic by tertiary processes. *J. Solid Waste Technol. Management* **25**(3/4), 148–160.

European Commission (2001). "ECCP (European Climate Change Programme)—Long Report." European Commission, Brussels, Belgium. [http://europa.eu.int/comm/environment/climat/eccp_longreport_0106.pdf].

Howard, N. (1995). Environmental impact of building materials, Volume D, plastics and elastomers, Project Report 14. Construction Industry Research and Information Association, London.

International Monetary Fund (2001). "World Economic Outlook." International Monetary Fund, Washington, DC.

International Organisation for Standardisation (1997). "Environmental management—Life cycle assessment—Principles and framework, ISO 14 040 (DIN EN ISO 14 040)." Deutsches Institut für Normung, Berlin.

International Organisation for Standardisation (1998). "Environmental management—Life cycle assessment—Goal and scope definition and inventory analysis, EN ISO 14 041." Deutsches Institut für Normung, Berlin.

International Organisation for Standardisation (1999). "Environmental management—Life cycle assessment—Life cycle impact assessment, EN ISO 14 042 (draft)." Deutsches Institut für Normung, Berlin.

International Organisation for Standardisation (1999). "Environmental management—Life cycle assessment—Life cycle interpretation, EN ISO 14 043 (draft)." Deutsches Institut für Normung, Berlin.

Joosten, L. (2001). The industrial metabolism of plastics. Analysis of material flows, energy consumption and CO_2 emissions in the life cycle of plastics, PhD thesis. Utrecht University, Utrecht, The Netherlands.

Kindler, H., and Nikles, A. (1980). Energieaufwand zur Herstellung von Werkstoffen—Berechnungsgrundsätze und Energieäquivalenzwerte von Kunststoffen. *Kunststoffe* **70**(12), 802–807.

Office of Industrial Technologies (2000). "Energy and Environmental Profile of the U.S. Chemical Industry." U.S. Department of Energy, Washington, DC. [http://www.oit.doe.gov/chemicals/pdfs/profile_chap1.pdf].

Office of Industrial Technologies (2002). "Energy, Environmental, and Economics (E3) Handbook." Energy Efficiency and Renewable Energy Network, U.S. Department of Energy, Washington, DC. [http://www.oit.doe.gov/E3handbook].

Patel, M. (1999). Closing carbon cycles: Carbon use for materials in the context of resource efficiency and climate change, PhD thesis. Utrecht University, Utrecht, The Netherlands. [ISBN 90-73958-51-2; http://www.library.uu.nl/digiarchief/dip/diss/1894529/inhoud.htm]

Patel, M., Bastioli, C., Marini, L., Würdinger, E. (2003). Life-cycle assessment of bio-based polymers and natural fibres. *In:* "Biopolymers" (A. Steinbüchel, Ed.), Vol. 10. Wiley–VCH, Weinheim, Germany.

Plastic Waste Management Institute, Japan (1999). Report on survey of LCI data of petrochemical products. Overview—LDPE, HDPE, PP, PS, EPS, PVC and B-PET. Plastic Waste Management Institute, Japan, Tokyo. [Available from Hirokazu Hayashi, WIC, Senior Chief Researcher Office of Research & Information, Daiichi-Kaneichi Building, 3-32-3, Hongou Bunkyou-ku, Tokyo 113-0033, Japan]

Reimann, D. O., and Hämmerli, H. (1995). "Verbrennungstechnik für Abfälle in Theorie und Praxis." Schriftenreihe Umweltschutz, Bamberg, Germany.

Shutov, F. (1999). Effective energy and gas emission savings using plastics waste recycling technologies. Expert group meeting on industrial energy efficiency, cogeneration, and climate change. Environmental Sustainability Division/Kyoto Protocol Branch/United Nations Industrial Development Organisation. Vienna International Center, Vienna, Austria, December 2-3.

VKE working group statistics and market research (German Association of Plastics Manufacturers) (2002). Plastics: Business data and charts. German Association of Plastics Manufacturers, Frankfurt, Germany. [http://www.vke.de/pdf-files/widat_engl.pdf]

Worrell, E., van Heijningen, R. J. J., de Castro, J. F. M., Hazewinkel, J. H. O., Beer, J. G. de, Faaij, A. P. C., and Vringer, K. (1994). New gross energy-requirement figures for materials production. *Energy* **19,** 627–640.

Polar Regions, Impacts of Energy Development

BRUCE C. FORBES
University of Lapland
Rovaniemi, Finland

1. **Regions of Impact**
2. **Geography, Accessibility, and Economics of Energy Reserves**
3. **Types of Impact**
4. **Ecological Implications of Energy Development**
5. **Sociocultural Implications of Energy Development**

Glossary

corridor Linear infrastructure (e.g., road, railway, and power line) that cuts across otherwise intact vegetation/soil cover, such as closed boreal forest or tundra.

cumulative impacts Responses to disturbance, often relatively indirect, that differ from simple direct and additive effects by accumulating in space and time, sometimes in unforeseen ways.

disturbance Any sudden deviation (e.g., a reduction in species richness or biomass) from a reference state, which need not be static. A disturbance regime has four basic dimensions: frequency, intensity, scale (extent), and timing (season).

permafrost Perennially frozen ground; a substrate that is in a state at or below 0°C (32°F) for 2 or more years. It can be composed of soil, bedrock, water/ice, or any combination of these. Moving generally from north to south, the Arctic can be divided into regions of continuous, discontinuous, and sporadic permafrost.

rehabilitation Used as a broader term than revegetation to denote the establishment of productive and functioning systems and may include revegetation, erosion control, or fish transplanting efforts.

restoration A more restrictive term referring to the long-term process of reestablishing the original composition, structure, and function to a disturbed site.

revegetation Refers to the establishment of vegetation, regardless of composition, on a disturbed area. Can be either natural (unassisted) or assisted.

succession Generally involves changes in vegetation species composition or in the relative numbers of individual plants through time at a given locale. It is often, but not always, precipitated by disturbance.

thermokarst The process by which characteristic landforms result from the melting of ice-rich permafrost.

tundra From the Finnish word *tunturi,* for sparse barrens near the timberline. Tundra is the Russian translation of this and refers more to the treeless zones beyond the latitudinal or arctic timberline as opposed to altitudinal timberline.

The polar regions hold enormous potential for the production of energy, mainly in the form of crude oil, natural gas and gas condensates, coal, and hydroelectric power. Some coal deposits in arctic Russia have been under industrial exploration and exploitation since the 1930s, and in other countries sources of hydropower and petroleum have also been under development for several decades. The Arctic contains by far the greater proportion of proven and economically viable deposits. In Antarctica, the Antarctic Treaty currently bans mining of any kind, although coal has been found in two regions and it has been suggested that rock formations there may well contain oil or gas. However, even if sizable reserves were discovered, and no exploratory drilling has been done, it is highly unlikely that the deposits could be recovered economically or safely. This article therefore focuses on the impacts of energy development in the Arctic, with special emphasis on oil and gas production, the main source of energy-derived revenue as well as a major cause of concern for environmental and sociocultural degradation. Examples are drawn from past and current energy developments from throughout the circumpolar North.

1. REGIONS OF IMPACT

The way we view energy development in the circumpolar North has partly to do with where we

 © 2004 Elsevier Inc. All rights reserved.

come from. Americans tend to think of Alaska as the Arctic. Because of the construction of the Trans Alaska Pipeline (TAPS) in the 1970s, the oil spill from the Exxon Valdez in Prince William Sound in 1989, and the bitter fight over opening the Arctic National Wildlife Refuge (ANWR) to oil development during the past decade, many Americans are familiar with the image of Alaska as an important oil-producing state. In neighboring Canada, increasingly more Canadians are probably aware of the rekindled prospects for a major gas pipeline from the Beaufort Sea south via the Mackenzie River valley to central Alberta. Talk to British or Scandinavian nationals about oil and one will hear about the declining production from the North Sea. Also, whereas most Russians know of the political power wielded by the former state monopoly, Gazprom, relatively few outside of international business/finance circles recognize this company's name.

The fact is that arctic energy development has become an increasingly multinational domain driven by a high-tech, expensive, and growing network of extraction and transport infrastructure that spans virtually the entire Northern Hemisphere. Drivers in Europe fill up their cars with gasoline refined from crude oil pumped from beneath reindeer pastures in western Siberia. Homes in Chicago are heated in winter with natural gas piped halfway across Canada from the mountains of northern British Columbia. Village power plants in central Russia burn coal mined at a former gulag camp in the open tundra well north of the Arctic Circle in the faraway Komi Republic. New England buys surplus electricity generated from huge artificial lakes in the James Bay region of northern Quebec. A hydropowered smelter under construction in the highlands of eastern Iceland will soon be providing aluminum to the world market. A fleet of icebreakers and ice-strengthened tankers is being constructed to carry oil via Russia's northern sea route (NSR) to markets in Europe and the United States. In other words, arctic energy exemplifies the process of globalization.

To understand the impacts of energy development in the Arctic we must consider first the geography, accessibility, and economics of energy reserves within the Arctic rim countries: Russia, Finland, Sweden, Norway, Iceland, Greenland, Canada, and the United States. Nonrenewable and renewable resources vary widely in quantity and quality across the circumpolar North, as do the financial and human resources available to exploit them, not to mention the prevailing environmental conditions. Some countries, such as Finland and Sweden, have little or no petroleum resources of their own and must import virtually all gasoline and fuel oils. They rely on nuclear plants, hydropower, coal, and even the burning of their plentiful woodchips to produce sufficient electricity for lighting and heating during the long, cold, and dark winters. Iceland has abundant geothermal energy available locally, but it also relies heavily on hydropower and imported petroleum to meet its energy needs. In contrast, Russia is considered the Saudi Arabia of natural gas and may one day be Europe's prime supplier.

2. GEOGRAPHY, ACCESSIBILITY, AND ECONOMICS OF ENERGY RESERVES

2.1 Alaska

Interest in petroleum development in Alaska dates to the early 20th century, when representatives of the U.S. Geological Survey sent to explore the North Slope provided reports about oil seeping out of the ground in places on the coastal tundra beyond the Brooks Range. In 1923, President Harding ordered that a 37,000 square mile area be set aside as National Petroleum Reserve No. 4. After the outbreak of World War II, the U.S government became concerned about the domestic supply of oil and began seismically exploring the huge area. During the period 1944–1953, dozens of test wells were drilled in the zone of continuous permafrost. At the time, the area was extremely remote, and with no roads most travel across the tundra was necessarily by helicopter and tank-like off-road vehicles with metal tracks instead of wheels. The oil fields found were disappointingly small, although six sizable gas fields were discovered. It was not until January 1968, after 10 years of exploration, that Atlantic Richfield Company (ARCO), Humble Oil, and, later, British Petroleum (BP) tapped into a field estimated to contain at least 5–10 billion barrels (bbl). The strike at Prudhoe Bay turned out to be the largest ever made in North America and plans were drawn up for an 800-mile pipeline to extend from the North Slope to the ice-free port of Valdez.

Because the pipeline had to cross land that Alaska Natives claimed, the discovery at Prudhoe Bay immediately created pressure for a land claims settlement to access the oil fields. The Alaska Native Claims Settlement Act (ANCSA) of 1971 transferred title to 18 million ha of land and $962.5 million to Alaska Natives. The land and money were not given

directly to individual Natives but, rather, the act divided the state into 12 regions and required that a state-chartered Native corporation be formed for each region with the provision that undeveloped Native corporation land was to be untaxed until 1991. There were problems of inequity because of the wide variation in exploitable natural resources. Natives who held claim over what became the North Slope Borough stood to profit from oil field leases far more than their neighbors. Although originally skeptical of Native claims regarding the critical importance of subsistence livelihoods, in 1987 some key amendments were made to ANCSA by which Congress recognized that the Native subsistence economy and culture are indeed significant factors in the management of Native lands and so attempted to provide for their protection. Subsistence provisions were also set forth in the Alaska National Interest Lands Conservation Act of 1980.

Since coming on line, the Prudhoe Bay/Kuparuk oil fields, via TAPS, have accounted for approximately 25% of the oil produced in the United States. Approximately 85% of the Alaskan state budget is derived from taxes and royalties collected on oil production. BP Exploration (Alaska), Incorporated and Phillips Alaska (formerly ARCO Alaska) are the two principal oil field operators and are both among the top 10 private employers in the state. At its peak in 1987, the North Slope produced an average of 1.61 million bbl of oil per day (bpd). In 2002, daily output declined to approximately 0.54 million bpd. Estimates of the recoverable reserves from the so-called 1002 area of ANWR vary widely from 4.3 to 30 billion bbl. This range has partly to do with economics because when market prices are less than $35 per barrel, a larger volume of technically recoverable oil is considered to be economically recoverable. However, politics also plays a role in that different players prefer to publicize higher or lower estimates. In comparison, the great Prudhoe Bay field was originally estimated to hold approximately 9.2 billion bbl in total reserves, but this was later increased to 23 billion bbl of oil and 26 trillion ft^3 of natural gas.

2.2 Canada

Canada is the world's third-largest natural gas producer, after Russia and the United States, with approximately 50% of its 6.5 trillion ft^3 of annual production destined for the U.S. market. Overall, Canada supplies 16% of the gas consumed in the United States, double its share a decade ago. In the Canadian Arctic, the major petroleum reserves lie in the Mackenzie River valley, delta, and Beaufort Sea region.

In the early 1900s, native Dene first led geologists to the site they called Le Gohlini, or "where the oil is," near Norman Wells on the lower Mackenzie River just north of the Arctic Circle. This led to the drilling of the first test well in 1919 by Imperial Oil, which confirmed the presence of rich oil fields. World War II led the U.S. Army to oversee the fast-track construction of the strategically located Canadian Oil Pipeline (Canol) almost 3000 km from Norman Wells to Whitehorse, Yukon. However, completion of the pipeline in less than 2 years coincided with the end of the war, and the project was quickly abandoned. In the early 1980s, Imperial Oil brought its Norman Wells operation to full capacity, and oil began flowing through the 870-km Norman Wells pipeline to Zama, Alberta. As of 2003, the pipeline averaged 10 million bbl per year (bpy) and was Canada's fourth-largest oil-producing field. Natural gas is also produced at Norman Wells, mainly for local use and for reinjection to enhance oil recovery.

Farther north, in the 1970s, some reserves of oil and gas were located in the high Arctic by Panarctic Oils. Feasibility studies revealed that a pipeline was not economically viable. Since the 1980s, there have been some shipments of oil eastward from Panarctic's Bent Horn field on Cameron Island and several experimental shipments of oil westward from Gulf Oil's Amauligak field in the Beaufort Sea. However, Canada's oil production is primarily south of the 60th parallel and gas comprises a much larger component of the nation's petroleum production. In recent years, exploration has been expanding off the coasts of Newfoundland and Labrador, where the federal and provincial governments are trying to replace the declining fishing industries with energy prospecting. Canada is also considering lifting a moratorium on oil and gas development off British Columbia.

The vast natural gas reserves of the Mackenzie River delta region were discovered in 1971. Whereas the Prudhoe Bay area of Alaska is estimated to contain as much as 100 trillion ft^3 of natural gas, an additional 64–73 trillion ft^3 lies in the combined onshore and offshore reserves of the Mackenzie delta and adjoining Canadian portion of the Beaufort Sea. A proposed 3500-mile Mackenzie River valley pipeline linking the American and Canadian fields would cost an estimated $26 billion and take 7 years to build. A shorter pipeline exploiting only the Mackenzie delta field would cost approximately $3 billion. The latter pipeline would have a capacity of approximately

1 billion cubic feet per day, equal to approximately 2% of U.S. consumption, whereas a pipeline joining northern Alaska would carry 4–5 billion ft^3 per day.

For the longer pipeline, energy company engineers have been studying three possible routes from the Arctic Ocean to connect via Norman Wells to pipelines in Alberta. Proponents of an Alaskan route estimate that they could start pumping gas south in 6 years. They believe that they can work more quickly because permits and an international regulatory framework have existed since the late 1970s to place a gas pipeline parallel to the TAPS, which would then run along the Alaska–Canada Highway. Another plan would cut a new overland route across the Brooks Range and northern Yukon, and a third route would run up the Mackenzie River valley and have a submerged section running 15 ft below the floor of the Arctic Ocean for 300 miles to connect the Mackenzie River delta to Prudhoe Bay.

The latter route is the one officially favored by the government of the Northwest Territories, which has frequently been at odds with the government of the Yukon Territory over the choice of a pipeline route. The northern Yukon is estimated to have reserves of 3.3 trillion ft^3 in the Eagle Plains and Peel Plateau and could potentially contribute 200 million ft^3/day to a Mackenzie Valley pipeline. The main hub for regional development and administrative activities is Inuvik, Northwest Territories, with a population of 3000, making it Canada's largest community north of the Arctic Circle. Like Prudhoe Bay, it is accessible year-round by road from the south.

As of early 2003, the Aboriginal Pipeline Group was raising money to cover its share of the costs for preliminary design and regulatory work, regardless of the chosen route. Meanwhile, TransCanada PipeLines Ltd. was positioning itself to play the dominant role in delivering arctic gas to southern markets, although Calgary-based Enbridge was working to be heavily involved as well. In June 2003, a consortium consisting of TransCanada, Imperial Oil, and ExxonMobil's Canadian unit completed a partnership agreement with the Aboriginal Pipeline Group to build an 800-mile portion of the Mackenzie Valley pipeline.

2.3 Greenland

Greenland may be sitting on oil reserves comparable in size to those of the entire North Sea according to the Home Rule government's Bureau for Minerals and Petroleum (BMP). North Sea fields have produced approximately 29 billion bbl of petroleum to date and proven remaining reserves stand at more than 15 billion bbl. A consortium led by Norway's Statoil and U.S. Phillips Petroleum has been awarded exploration licenses in waters off the southwest coast of Greenland. The BMP acknowledges that gigantic fields would be necessary to economically justify opening up the area.

2.4 Russia

Large-scale energy exploration and exploitation in arctic Russia have been going on for several decades. A coal mine at Vorkuta in northernmost Komi Republic began production as a gulag camp under Stalin in the late 1930s with a railway link to Moscow and other Russian regions. It grew quickly to a large complex of connected mining towns with an eventual peak population of 280,000 and coal production of 22 million tons per year by the late 1980s. Since 1991, the population has declined by approximately 30% and coal production has fallen well below levels from the mid-1970s. Petroleum has also long been under production in the Russian north, especially in the northern boreal and subarctic lowlands of western Siberia. Western companies began exploring projects in Russia even before the collapse of the Soviet Union, but more than a decade later they can point to few successes after combined investments of more than $5 billion.

ExxonMobil, the world's largest oil firm, is focused mainly on its huge oil and gas project on Russia's far eastern island of Sakhalin and is therefore not directly involved in any truly arctic ventures. Shell, the world's second largest oil firm, signed a memorandum in the late 1990s to develop oil and gas condensate deposits at Gazprom's huge Zapolyarnoye gas field in western Siberia, but the deal was never finalized. BP, the world's third largest oil firm, has had extensive projects in the Russian north via its 50% interest in Tyumen Oil (TNK), which produces 1.2 million bpd and has reserves of 5.2 billion bbl. In June 2003, U.K. Prime Minister Blair and Russian President Putin jointly announced the creation of a new oil company, TNK-BP, which will lead to an additional $6.15 billion investment by BP in Russia. TotalFinalElf, the world's fifth largest oil company, is developing a medium-sized oil field in Russia's north together with Norway's Norsk Hydro, Kharyaga, which produces 30,000 bpd and holds 710 million bbl of reserves. Total also acquired in 2000 the giant but remote Vankor field in eastern Siberia and has plans to develop with Gazprom the giant offshore arctic Shtokmanovskoye gas field. Other firms

actively engaged in Russian energy production and transport include ConocoPhillips, Finland's Fortum group, and Germany's Ruhrgas, which are involved in the financing, planning, and building of pipelines to European markets, mainly via the terminal facility at Primorsk, 100 miles (160 km) northwest of St. Petersburg.

In general, Russia faces enormous transport problems in getting its petroleum resources, for the most part deep in Siberia, to export markets. This makes scenarios for shipping oil to North America profitable only at times of high prices. Most of Russia's oil and gas goes to Europe. In the short term, U.S. oil companies could raise Russia's oil output by investing more in Russia. However, U.S. companies have been reluctant to invest there because of shifting tax codes and regulatory policies and have called for binding legislation before investing more money in the country. Specifically, they have been unwilling to do this without production sharing agreements, which Russia's Duma has been reluctant to grant foreign firms.

2.4.1 Murmansk Terminal

As the world's No. 2 oil exporter, Russia is seeking new markets outside Europe, whereas the United States, the world's largest oil consumer, wants to cut its reliance on the oil-rich but volatile Middle East. Under this scenario, Murmansk would become the main port for oil export to the United States. By 2010, Russian companies hope to cover 10–15% of crude oil consumption in the United States. Indeed, in November 2002 Presidents Bush and Putin pledged to strengthen energy ties between the two countries. The project is part of a blossoming relationship between the former Cold War rivals. However, Washington wants more than simply increased Russian oil imports out of its new energy relationship with Moscow. It is also keen to open Russia's vast oil reserves to U.S. investors.

In November 2002, Russia's top four oil companies—Lukoil, Yukos, TNK (since merged with BP), and Sibneft (since merged with Yukos)—announced their plan to sign a preliminary deal to build a 1-million bpd capacity oil-export terminal in Murmansk, Russia's only ice-free Arctic port. Oil to supply Western Europe and North America is expected to come primarily from the Timan–Pechora fields, at least initially. Timan–Pechora has one of the region's largest known reserves of oil, estimated to be approximately 126 billion bbl. Currently, ConocoPhillips produces approximately 25,000 bpd (10 million bpy) from its Polar Lights facility, a joint venture with Lukoil along the border between the Nenets Autonomous Okrug (NAO) and Komi Republic. It also wants to develop a neighboring area, Northern Territories, believed to contain 4.25 billion bbl of crude under Russian standards, which tend to give larger estimates than Western rules. Transport would be either by ice-strengthened tankers loaded at the Lukoil-operated Varandei terminal in the NAO or via a proposed 80-million tons per year pipeline.

There are two options for the pipeline route. The first route is western Siberia–Usa–Yaroslavl–Archangelsk–Murmansk. It is 2500 km long, including a short section which runs along the floor of the White Sea. The second route is western Siberia–Ukhta–Murmansk, which is 3600 km long. This route is fully on the ground and allows additional switching of the Kharyaginskoe field of Lukoil and TotalFinaElf in the NAO to the transportation system. There are also proposals to modernize an existing pipeline, Uhkta–Yaroslavl, which transports oil from the Lukoil fields in the Komi Republic. The costs of the project are estimated to be $3.4 billion for the first route and $4.5 billion for the second route.

This is the first project of its kind and scale in Russia in which the companies plan to undertake the development essentially on their own. This would mean little or no participation by the state, which normally defines oil transportation via the trunk pipelines as a monopoly activity. To date, the fact that pipelines are built solely by state monopoly Transneft has allowed President Putin to maintain control over the up and coming petroleum oligarchs. However, most analysts insist that the development of such a grandiose project as construction of a large oil export facility in Murmansk, in combination with development and transport of the hydrocarbon reserves of Timan–Pechora, Yamal, and the Barents seabed, will not be possible without any state support. It is widely acknowledged that Russia has already reached an oil production bottleneck and cannot increase export capacity before expanding its transport infrastructure (ports, pipelines, and railways) and replacing a great deal of obsolete technology.

Russia's daily output is 8 million bpd, of which approximately 5.1 million bbl is exported. In comparison, U.S. domestic crude oil production averaged 5.8 million bpd in 2002, whereas oil imports averaged 9 million bpd for the same year, spiking to 9.4 million bpd in October. It is widely believed that the Russian arctic seabed may well become one of the main sources of oil in the 21st century. Moreover, oil resources of the North Sea are

steadily decreasing and it is estimated that reserves of the Norwegian sector are sufficient for only another 25 years and that Europe will soon import up to 85% of its energy supplies. Currently, Russia provides only approximately 20% of Europe's natural gas needs and 16% of its oil.

2.4.2 Yamal

In response to the situation described previously, Gazprom and Rosneft Oil Company and its affiliates for the exploitation of arctic oil and gas resources created Sevmorneftgaz as a joint enterprise. At the same time, many agreements between Russian and foreign consortiums have come and gone since the collapse of the Soviet Union. Amoco Eurasia (since merged with BP), for example, pulled out of Yamal in 1996 after several years of negotiations and feasibility studies. One reason why major Western companies are prepared to risk a return to Russia is the price of oil. Diving to $10 a barrel in 1998, it has climbed steadily since and, despite a downward dip after the terrorist attack of September 11, 2001, has been buoyed by tensions in the Middle East.

In January 2002, Gazprom and the administration of the Yamal–Nenets Autonomous Okrug (YNAO) sent the federal government a draft law intended to make Yamal a "region of strategic interests." The aim is to ensure the company's goal of increasing its annual output to 530 billion cm by 2010 as production from its massive Urengoi and Yamburg fields begins to decline. In doing so, Gazprom demonstrated its intention to concentrate on developing its Yamal reserves over the natural gas potential of eastern Siberia. One reason is the relative ease of transporting gas from the Yamal Peninsula. Whereas eastern Siberia is almost devoid of infrastructure, the established Yamal–Europe pipeline is capable of pumping 18 billion cm of gas annually, which could increase to 60 billion cm once a second line is completed, most likely via Poland.

There are 25 known large fields of natural gas on Yamal, with estimated total reserves of 10.4–13 trillion cm and hypothetical reserves of up to 50 trillion cm. Developing them to the extent that it would be possible to begin export of shipments in commercial quantities would require more than $40 billion according to Gazprom's estimates from the early 1990s. By March 2003, Gazprom's board of directors had plans to adopt a "program for the comprehensive industrial development of the Yamal Peninsula fields."

On Yamal Peninsula, the Bovanenkovo and Kharasavey gas condensate fields are the most thoroughly explored and best prepared for production. Their development scenario calls for the construction of a liquefied natural gas (LNG) plant at the Kharasavey field and with LNG shipments by the NSR. This program will become the foundation of the company's strategic development until 2010. As such, the Yamal Peninsula and the estuaries of the Ob and Taz Rivers are expected to be major regions for the expansion of Gazprom's gas production. Adopting the Yamal Development Program will mean that other large fields, especially the gigantic Shtokmanovskoye field in the Barents Sea, will be developed in the more distant future. The simultaneous implementation of two major projects would be too costly for Gazprom and might result in intense competition between gas sellers in domestic and foreign markets.

2.4.3 Shtokmanovskoye

Discovered in Soviet times, the Shtokmanovskoye gas field in the Barents Sea is estimated to hold 112 trillion ft^3 (3.2 trillion cm) of natural gas and 170 million bbl of crude oil and is due onstream in 2010 under a $19- to $21-billion development scheme. It is believed to contain the world's largest natural gas deposits. Houston-based ConocoPhillips, the world's sixth largest energy company, and TotalFinalElf each want a 25% stake in the Shtokmanovskoye project. However, as with the terminal facility in Murmansk, there is substantial skepticism among the major foreign players as to whether the project can really be developed.

As a long-term strategy, Gazprom must either focus on the development of midsized fields and deeper horizons of currently developed fields in the Nadym–Pur Taz District of the YNAO or undertake the full-scale development of fields on the Yamal Peninsula. The third option is the development of arctic offshore fields such as Shtokmanovskoye, but as of summer 2003 the Yamal Peninsula appeared to be the main development target until at least 2010.

3. TYPES OF IMPACT

The causes of environmental degradation commonly associated with the petroleum industry are many and have historically included drilling pads and roads and the excavation of the gravel and sand quarries necessary for their construction; blowing sand and dust from quarries and especially roads; housing; processing facilities; pipelines and power transmission lines; waste fluids from drill holes; oil spills; marine, air, and ground traffic, including rutting from off-road tracked vehicles; and seismic survey

trails. Some of these impacts (e.g., vehicle tracks) have become relatively rare occurrences in the North American Arctic since the institution of strictly enforced mitigation regulations in the 1970s. On the other hand, most of these impacts continue to be common and widespread in the Russian Arctic despite the existence of relevant regulations since the 1980s, primarily due to a lack of enforcement.

In March 2003, a panel of experts convened by the U.S. National Academies published a major review titled "Cumulative Environmental Effects of Oil and Gas Activities on Alaska's North Slope." They reported on the overall influence of 35 years of oil exploration and production activities in and around Prudhoe Bay and classified the impacts into four categories: effects on the physical environment, effects on vegetation, effects on animals, and effects on the human environment. The panel's aim was to assess not only known impacts but also future cumulative effects should production expand, for example, eastward into ANWR. The 1.5-million-acre coastal plain of ANWR occupies 5% of Alaska's North Slope and is somewhat smaller than the state of South Carolina. It encompasses six ecological zones and is considered to contain the greatest biological diversity of any protected area in the Arctic. This is probably due to its location within what is known as Beringia, a region at the crossroads of two continents during the ice ages of the Pleistocene. As a result of this geography, Alaska is home to a number of species associated primarily with North America or Eurasia as well as a fair share of endemic species.

Less tangible than the obvious impacts described previously, but also acknowledged by the panel, was that the "wilderness" or "wildlands" value of ANWR would invariably be degraded by petroleum development. This would violate, in their words, "perhaps the most prominent of the values ANWR was established to protect." Although many Alaska Natives disagree over the appropriateness of wilderness designation, there is general agreement that industrial "structures and activities [would] also violate the spirit of the land, a value that is reported by some Alaska Natives to be central to their culture."

In terms of physical and vegetation impacts throughout the Arctic, perhaps the most important factor distinguishing energy development there from that of other regions is the presence of permafrost. The late arctic–alpine biologist Dwight Billings referred to permafrost as the "glue" that binds arctic tundra ecosystems together. Permafrost exerts major influence over surface and subsurface hydrology, and thus affects substrate temperatures and the pattern and productivity of vegetation. A large share of the global carbon budget is stored in permafrost, and there is a strong linkage between carbon dioxide (CO_2) exchange and methane (CH_4) formation and emission rates in tundra ecosystems. However, as frozen ground warms sequestration processes are affected and tundra soils can turn from carbon and methane sinks to sources. The atmospheric input of methane from permafrost regions has been estimated to be 20–40 Tg CH_4/year, which is almost 25% of the total emissions from natural sources globally. It is well-known that the types of surface disturbance associated with energy development lead to soil warming and can easily lead to thermokarst, thus threatening the maintenance of permafrost in affected areas.

In addition to direct disturbances of the ground surface, other less visible impacts can accumulate over time. These may occur independently of each other or may be exacerbated through synergy among various proximal effects. These indirect impacts, which fall under the umbrella of cumulative impacts, are well documented, and although they were formerly unforeseen, we can now predict them in many cases. According to Skip Walker of the Institute of Arctic Biology at the University of Alaska Fairbanks (UAF), cumulative impacts differ from simple additive effects in five major ways: (i) The disturbances are so close in time that the system cannot recover (time-crowded perturbations); (ii) the disturbances are so closely spaced that the effects are not fully dissipated between them (space-crowded perturbations); (iii) combinations of disturbances produce effects that are different from those of any single disturbance (synergistic effects); (iv) the effects are delayed in time or space from the original disturbance (indirect effects); and (v) large regions are affected by many seemingly insignificant perturbations (nibbling), somewhat analogous to the process of fragmentation during deforestation.

The panel convened by the National Academies concluded that the effects of oil field infrastructure are

> *manifest not only at the physical footprint itself but also at distances that vary according to the environmental component affected. Effects on hydrology, vegetation, and animal populations occur at distances up to several kilometers, and cumulative effects on wildland values, especially visual ones, extend much farther, as can the effects on marine mammals of sound caused by some offshore activities. All effects attributable to the structures and the activities associated with them accumulate, and many will persist as long as the structures remain, even if industrial activity ceases.*

During the construction of TAPS, more than 78,500 ha (785 km^2) of land was directly disturbed, of which 31,700 ha (370 km^2), or 40%, was for gravel borrow sites or quarries extracted from both surface sources and streambeds. According to agronomist Jay McKendrick, Professor Emeritus at UAF, this extraction of gravel fill for drill pads and roads is the most extensive tundra disturbance. It has been calculated that approximately 0.04% of Alaska's Arctic Coastal Plain (or 0.1% of the area between the Colville and Canning Rivers) has been affected by gravel fill. In the Russian Arctic, by comparison, exploration and development activities at a single gas field at Bovanenkovo on the Yamal Peninsula had by 1990 resulted in the loss of 127,000 ha (1270 km^2) of tundra and 0.14% of the total area on Yamal. On neighboring Taz Peninsula, the figure has reached 1.5%, or approximately 6000–7000 km^2. Geographer Arkady Tishkov estimates that the area of destroyed ecosystems within the tundra zone of Russia comprises 8% of its total territory.

Regarding spills of liquids, from 1974 to 1977, when TAPS was first built and operated, more than 16,000 hydrocarbon spills, totaling more than 22,200 bbl, occurred along the pipeline route. During 1985 and 1986, 952 spills were reported on the North Slope, totaling 6137 bbl. In the Russian north, 7–20% of the oil extracted each year is lost through accidents along oil pipelines. A large spill of 103,000–126,000 tons near Usinsk in 1994 was reported in the world media and generated international concern, albeit briefly. However, such events are not uncommon and occur every 2 or 3 years according to Russian geographer Gregory Vilchek. Because Murmansk does not lie in the permafrost zone, its proposed oil export terminal would therefore mainly be a threat to coastal and marine ecosystems via potential spills. On the other hand, the proposed supply pipeline across the Kola Peninsula from the fields at Timan–Pechora (via Archangelsk) would certainly affect a large swath of land.

Impacts associated with energy development are complex in that various human activities can influence ecosystems simultaneously and can have both immediate catastrophic and long-term effects, leading to the aforementioned cumulative impacts. Sometimes, in practice, it is difficult to distinguish between direct and indirect impacts and scientists may use different methods for classifying disturbances. For example, Russian authors distinguish three main classes of disturbed areas: *ochagovyi* (local), *lineinyi* (linear), and *fonovyi* (spatial). The most striking example of the first type is sites surrounding petroleum boreholes (drill sites). Transport corridors appearing in connection with road and pipeline construction constitute linear disturbances. Large territories affected by air pollution (e.g., coal burning) are examples of spatial disturbances.

4. ECOLOGICAL IMPLICATIONS OF ENERGY DEVELOPMENT

The previous section described the major existing and planned energy developments around the circumpolar North. What are the implications of these various schemes for procuring enough energy at affordable prices? Is it possible to be efficient economically as well as to mitigate major environmental degradation? What are the true environmental costs of energy extraction along the entire route from source to consumption? What if societies and cultures other than those producing or buying the energy are adversely affected during the exploration and extraction phases to the extent that their health and traditional livelihoods are threatened? These are far from trivial questions.

The following example illustrates the scope of the process of just one contemporary energy production scheme in which the full costs are not necessarily accounted for by either the producers or the end users. During a debate on whether to build a fifth nuclear reactor in Finland, some "environmental" opponents declared that purchasing natural gas from neighboring Russia would be a cleaner and safer alternative. However, others pointed out two problems with this reasoning, which they criticized as the rationale of "out of sight, out of mind." First, much of the gas consumed in northern and central Europe is derived from fields in western Siberia, where indigenous groups dependent on traditional livelihoods (e.g., reindeer herding, hunting, gathering, and fishing) have been powerless to stop widespread habitat degradation within their ancestral homelands. Second, Russian pipeline construction standards have traditionally been well below Western levels, so that much of the gas ends up leaking into the atmosphere during transport. When burned properly, natural gas emits approximately 60% less carbon dioxide to the atmosphere than coal and approximately 30% less than oil. However, gas released directly to the atmosphere as methane has many more times the heat-trapping capacity than CO_2, making it a potentially powerful contributor to the greenhouse effect. These are just some of the

issues that need to be considered during the ongoing development of international energy networks.

Anthropogenic activities in the Alaskan Arctic up to the present time have generally created small but locally intense terrestrial disturbances ($<1\,km^2$ and often $<100\,m^2$), and in total they occupy small proportions ($<10\%$) of oil and gas fields. These small-scale disturbances may affect wildlife out of proportion to their spatial extent by creating microscale heterogeneity with patches that can either attract or repel animals. As mentioned previously, in portions of Arctic Russia, such as northwest Siberia, the realized and potential impacts are greater by an order of magnitude as a result of little or no adherence to or enforcement of regulations aimed at mitigation.

Much of the debate concerning the ecological implications of energy development revolves around habitat loss in the North American Arctic. Habitat is the place where an organism is found, including both biotic and abiotic factors. In general, the distribution of a population within its habitat at a given time is a measure of habitat value. Loss of certain portions of the species' habitat may affect the species more immediately and more severely than loss of other portions of its habitat. Loss of high-value habitats could have immediate adverse consequences for the populations that they support. Russia faces not only major losses of habitat but also the direct losses of wildlife and local species extinctions because rampant poaching of fish, wildlife, and semidomestic reindeer in and around arctic petroleum fields and along transport corridors is a major problem.

Petroleum deposits in western Siberia were first identified and tapped during Soviet times, mainly in the more accessible subarctic zones south of the arctic tree line. Those original oil and gas fields were developed recklessly with little or no concern for the technical longevity of the fields or the environmental damage resulting from the exploration and exploitation phases. It is certain that Russia will need foreign expertise in both extraction and environmental mitigation, in addition to capital investments, to develop new fields in the Arctic.

In terms of raising this capital, it makes little difference who becomes Gazprom's partner—Russian oil companies, major gas traders such as Germany's Ruhrgas, or multinationals such as BP or Shell. However, the environmental implications are potentially profound because the traditionally state-owned Russian monopolies have had few incentives for environmental protection. Western companies, on the other hand, are highly competitive and increasingly sensitive to their image. For example, BP Amoco PLC is one of two dominant oil companies on Alaska's North Slope and has been trying to enhance its image globally during the past several years, beginning with a new sunburst logo and slogan "Beyond Petroleum" in summer 2000. Although BP pumps more than 300,000 bpd from Alaska, it has made efforts elsewhere to increase its production of cleaner-burning natural gas over crude oil and also invested heavily in solar power. Nevertheless, in December 2002, a U.S. judge tightened controls over BP by requiring the company to provide unrestricted access to its Alaskan operations to verify compliance with federal, state, and local environmental, health, and safety laws. The merger of BP with TNK raises hope for mitigation as its new Russian fields are developed.

In North America, proponents of the future development of ANWR argue that the lessons learned during the past 30 years, combined with new exploration and drilling technologies, allow engineers to design oil fields with an ecological footprint half that used in the Prudhoe Bay oil field. However, opponents point out that the switch from gravel to winter ice roads is leading to the depletion of surface and groundwater in some areas. Furthermore, the cumulative impacts from past development described previously are not yet fully understood, although it is clear that even small disturbed patches of vegetation and soils can persist indefinitely.

It is known that environmental regulation can make a positive difference. Studies of disturbance associated with exploratory well sites in northern Yukon Territory found significantly less terrain damage and thermokarst topography after the introduction of territorial land-use regulations in the early 1970s. Improvements in industry operating conditions (e.g., the use of vehicles equipped with low-pressure tires) and strict application of the land-use regulations (e.g., the restriction of cross-tundra vehicle movement to winter months) are important factors in the reduced impact observed. In this light, the potential for the most serious terrain and environmental damage is now associated with the drilling operation, which may sometimes extend into the critical summer months, and the disposal of waste drilling fluids. That the disturbances in the Yukon were neither as obvious nor as severe as in the Mackenzie River delta and the islands of the High Arctic was attributed, respectively, to the relative absence of ice-rich permafrost and the ameliorated subarctic climate conditions.

With regard to wildlife, the impacts of the oil development at Prudhoe Bay on the caribou of

the Central Arctic herd (CAH) are heatedly debated. On the one hand, the CAH has increased from 3000 animals in 1973 to 29,000 today, and the animals can be seen grazing among the oil field infrastructure in summer, even using the roads for relief from insect harassment. On the other hand, the population growth coincides with a period of favorable climate conditions that may not last. Studies have shown that mothers pregnant with calves tend to stay at least 4 km from roads and that the calving grounds have shifted away from the infrastructure over the years from the most productive habitats to less productive areas. At the same time, the Porcupine caribou herd (PCH), which migrates into ANWR from Canada, has been declining since 1989. There are concerns that if the refuge is developed the additional stress will cause an increase in calf mortality and a dramatic reduction in herd size. Concerning the proposed pipeline route, some analysts believe that using the TAPS and Alaska Highway corridor would have less impact because it would primarily use existing corridors. Opponents believe that a submerged Mackenzie Valley pipeline could endanger both marine and terrestrial ecosystems.

Concerning offshore development, the World Wildlife Fund and others have acknowledged that safe transportation with minimum pollution poses serious challenges for environmental organizations in general and marine environmental science in particular. The Arctic Operational Platform project, funded by the European Union during 2002–2004, is intended to demonstrate the development capacity of the NSR. The port at Varandei in NAO remains the only Russian Arctic stationary terminal for oil transportation and is expected to grow. Other terminals, such as the one on nearby Kolguev Island, are functioning only on a temporary basis.

There is concern that Gazprom is launching into the gas liquefaction business in an area with some of the most extreme natural and climatic conditions in the world. Frozen seas would complicate shipment of liquids and LNG via the NSR. The cold climate means that the biodegradation of oil spills on both land and in ice-covered waters would occur at a much slower rate than in more temperate ecosystems. Most of the spills of liquids on Alaska's North Slope have consisted of refined petroleum products and occurred on water or in gravel pads, although some have occurred on terrestrial vegetation. Oil does not penetrate deeply into saturated soils, but spills on dry sites are absorbed by mosses and the underlying organic material and mineral soils. Sedges and willows are the first vascular plants to reappear following a terrestrial oil spill. Recovery from diesel spills is extremely slow, showing little change after approximately 30 years and probably after even half a century. By comparison, recovery from spills of crankcase and crude oil in the low Arctic generally occurs more quickly (<30 years) in wetland habitats. Recovery can take much longer in the colder, drier habitats of the high Arctic, not to mention in ice-covered waters.

NSR transport projects supporting the development of the Bovanenkovo and Kharasavey gas fields on Yamal require the construction of a port at Kharasavey for the import of extraction equipment and the export of liquids. At the same time, a railroad is being laid from the Ob River gulf to Bovanenkovo and a parallel service road is planned to continue from Bovanenkovo to Kharasavey. There has been a great deal of damage along this corridor, especially due to quarrying. All quarries—sand, gravel, and rubble—are slow and difficult to revegetate either naturally or with assistance and are therefore generally kept to a minimum by companies active in the North American Arctic. In the best case scenario, a complete plant cover is possible over a period of 20–30 years, although costly subsidies of nutrients, organic matter, and water are typically required to achieve this. Furthermore, the majority of plants that are either planted intentionally or eventually colonize the site naturally are not likely to be those that occurred there originally.

Another factor to consider when dealing with tundra vegetation is climate change. A warming arctic climate may offer both opportunities for revegetation efforts and risks from migrating exotic species that may compete with native species. However, with or without climatic change, we cannot expect the distribution of weedy plants, or even whole plant communities, to remain static at their northern limits. There is concern that expansion of corridors and increasing human relocation from southern regions will likely continue to provide a ready source of introduced species. A recent effort by the United Nations Environment Program to model anthropogenic impact, GLOBIO, predicted that by 2050 more than 50% of the Arctic may be affected by critical levels of disturbance. Much of the impact is expected to occur along the growing network of transport corridors and related infrastructure (e.g., housing, power lines, and airports).

Soon after the strike at Prudhoe Bay, oil companies and TAPS arranged for environmental studies of their own and supported research at the Institute of

Arctic Biology at UAF into revegetation and rehabilitation methods for tundra disturbed during the course of engineering operations and infrastructure development. Based on early experiments, seeding of three grasses [*Arctagrostis latifolia* (Var. Alyeska), *Festuca rubra* (Var. Arctared), and *Poa glauca* (Var. Tundra)] became a standard practice. On medium-texture, mesic soils, seeding and fertilizing usually produced grass stands within three growing seasons. In the long term, seeding these commercial grasses has frequently delayed natural recolonization where seeded grasses established and persisted. Examples are overburden stockpiles at gravel mines. Where such stockpiles were seeded, the commercial grasses have resisted invasion by other tundra plant species for at least 15 years. It is believed that competition from the seeded grasses is responsible for the delay of indigenous plant invasion during the process of succession.

Although there are still important gaps in our understanding of natural tundra ecosystems, it is becoming increasingly clear that there are varied and critical functions to be served by restoration procedures in disturbed ecosystems. Among the most important applications are the prevention of erosion, rehabilitation of degraded wildlife habitat, and protection of permafrost and associated benefits from carbon sequestration. Most applications to date have taken place in subarctic and low Arctic ecosystems. In the early years, research was necessarily concentrated on assisting recovery (i.e., establishing a simple plant cover as quickly as possible). Recently, however, ecological restoration has come to be seen as an aim in itself. Although plants are selected primarily to prevent erosion, many of the chosen species also provide important food for wildlife. Since it is standard practice to apply organic matter and/or chemical fertilizers to subsidize the initial stages of growth, managers need to be wary about the access of herbivores (e.g., caribou and reindeer) to sites either naturally recovering from disturbance or actively revegetated.

5. SOCIOCULTURAL IMPLICATIONS OF ENERGY DEVELOPMENT

In the discussion of the impact of development of ANWR, referred to by some as the Serengeti of the North, much of the focus has been on the area's wildlife and wildlife habitat. As a result, the human dimension is often overlooked. However, in North America, as in Russia, many communities depend to a greater or lesser extent on fish and game for nutritional and cultural purposes. Just as onshore drilling can negatively affect calving caribou and therefore hunting opportunities, offshore oil developments may prove destructive to populations of fish, marine mammals, or birds on which local residents depend for subsistence needs. Similarly, oil or gas operations may generate forms of water or air pollution that are physically harmful to local residents. Political scientist Oran Young of the University of California at Santa Barbara observes that large-scale development involving oil or hydroelectric power in the far north may also severely disturb or distort important social institutions or central elements of the preexisting indigenous way of life, such as subsistence economies, cultural norms of sharing, and traditional authority structures. Similarly, construction projects associated with the development of northern resources may produce severe stresses on delivery systems for health, education, and welfare even in larger modern settlements.

Relative to their North American counterparts, Russia's arctic indigenous groups, the so-called small-numbered peoples of the North, lack significant political clout. Researcher Gail Osherenko of the University of California at Santa Barbara suggests that title to land may be essential to their cultural survival in the face of extensive arctic resource development. Unfortunately, despite the advent of glasnost and perestroika, there are still risks for speaking out on such issues. Some of the more vocal political activists in western Siberia have been jailed.

For groups that still practice traditional livelihoods, such as nomadic reindeer herding, the last generation has been one of adaptation to significant upheaval. The Yamal Nenets, for example, have survived first Tsarist and later Soviet dreams of establishing state and religious authority over even the most remote human populations. However, nothing has challenged them like the ongoing search for petroleum beneath their ancestral lands. The data discussed previously concerning the amount of destroyed tundra and forest–tundra are of critical sociocultural significance because virtually all of the lands once comprised reindeer grazing pasture. A culture that has changed little in 1000 years is suddenly endangered because of the havoc wrought by petroleum development. In addition to environmental degradation, the infusion of southern workers to the oil and gas fields has ushered in a host of health and demographic problems that did not exist 20 years ago.

Concerning development of the NSR, Alona Yefimenko of the Arctic Council's Indigenous Peoples' Secretariat is skeptical about measures to protect indigenous livelihoods. She believes that "practically, most of us know that real consultation and equal participation and representation by indigenous peoples requires a legal and political infrastructure that does not exist." In her view, the existing examples of energy development in Russia's north show "a very sad picture": ruined reindeer pastures, deserted oil and gas derricks, roads dissecting reindeer migration routes, etc. Nenets leader Dmitri Khorolya, president of the Reindeer Herders' Union of Russia and director of Yarsalinski sovkhoz, the largest collective management unit on Yamal Peninsula, has voiced similar alarm. In June 2001, he observed that

> *the vulnerable ethnic–economical systems of reindeer peoples are frequently exposed to hard market conditions, particularly where oil and gas mining has become the principal factor in the development of Arctic areas. Industrial activity in the North has resulted in the destruction of many thousands of hectares of reindeer pasture. The process is continuing. In some regions pasture degradation threatens preservation of reindeer husbandry and the anxiety of reindeer herders for their future should be heard by the world community*

The pasture degradation results not only from the direct impacts by industry. Additional losses occur as the animals are forced onto progressively smaller parcels of land each year, resulting in sometimes severe overgrazing and desertification of the remaining pastures.

As in Russia, the proven and potential natural gas reserves of arctic Canada occur mainly on lands inhabited by indigenous peoples or First Nations. The difference is that the political clout of the Canadian First Nations was sufficient to halt development of the Mackenzie Valley pipeline in the 1970s. However, most of the First Nations groups that once opposed drilling on legal, environmental, and cultural grounds have abandoned their protests and now favor exploiting their region's energy resources. As a result, the government of the Northwest Territories is currently pressing Ottawa for an ambitious road and bridge construction program that would lower production costs. Stephen Kakfwi, premier of the Northwest Territories, predicts that the pipeline will soon be built and that "many communities will become twice the size, and all will have paved roads." His recently elected counterpart in the Yukon, premier Dennis Fentie, is similarly supportive of the development plan.

Native attitudes toward the development of ANWR are somewhat less homogeneous. The inland Gwich'in settlements of northern Yukon and Arctic Village in neighboring Alaska depend on the caribou of PCH and are bitterly opposed to any further development of the coastal plain. At the same time, the Iñupiat of Kaktovik on the coast of Alaska are just as strongly in favor of development.

SEE ALSO THE FOLLOWING ARTICLES

Arid Environments, Impacts of Energy Development in • Climate Protection and Energy Policy • Crude Oil Spills, Environmental Impact of • Energy Development on Public Land in the United States • Environmental Injustices of Energy Facilities • Gas Hydrates • Natural Gas Resources, Global Distribution of • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Oil and Natural Gas: Offshore Operations • Oil Pipelines • Wetlands: Impacts of Energy Development in the Mississippi Delta

Further Reading

Crawford, R. M. M. (Ed.) (1997). "Disturbance and Recovery in Arctic Lands: An Ecological Perspective." Kluwer, Dordrecht, The Netherlands.

Forbes, B. C. (1999). Reindeer herding and petroleum development on poluostrov Yamal: Sustainable or mutually incompatible uses? *Polar Record.* **35,** 317–322.

Forbes, B. C., McKendrick, J. D. (2002). Polar tundra. In: Davy, A. J., Perrow, M. (Eds), Handbook of Ecological Restoration, Cambridge Univ. Press, Cambridge, UK. pp. 355–375.

Jorgenson, M. T., and Joyce, M. R. (1994). Six strategies for rehabilitating land disturbed by oil development in arctic Alaska. *Arctic* **47,** 374–390.

Khitun, O., Rebristaya, O. (2002). Anthropogenic impacts on habitat structure and species richness in the West Siberian Arctic. *In* "Wilderness in the Circumpolar North" (A. E. Watson, L. Alessa, and J. Sproull, Eds.), pp. 85–95. U.S. Department of Agriculture/Forest Service, Rocky Mountain Research Station, Ogden, UT.

Nellemann, C., *et al.* (2001). "GLOBIO Global Methodology for Mapping Human Impacts on the Biosphere: The Arctic 2050 Scenario and Global Application." United Nations Environment Program, Nairobi, Kenya.

Orians, G., *et al.* (2003). "Cumulative Environmental Effects of Oil and Gas Activities on Alaska's North Slope." National Academies Press, Washington, DC.

Rasmussen, R. O., and Koroleva, N. (Eds.) (2004). "Assessing Social and Environmental Impacts in the North: Methodologies in Evaluation of Socio-economic and Environmental Consequences of Mining and Energy Production in the Arctic and Sub-Arctic." Kluwer, Dordrecht, The Netherlands.

Reynolds, J. F., and Tenhunen, J. D. (Eds.) (1996). "Landscape Function and Disturbance in Arctic Tundra." Springer-Verlag, Berlin.
Sumina, O. I. (Ed.) (2000). "Research on Anthropogenic Impacts in the Russian Arctic," Arctic Centre Reports 29. Univ. of Lapland, Rovaniemi, Finland.
Truett, J. C., Johnson, S. R. (Eds.) (2000). "The Natural History of an Arctic Oil Field: Development and the Biota." Academic Press, San Diego.
Vilchek, G. E., and Bykova, O. Y. (1992). The origin of regional ecological problems within the northern Tyumen Oblast, Russia. *Arctic Alpine Res.* **24,** 99–107.
Walker, D. A., and Walker, M. D. (1991). History and pattern of disturbance in Alaskan Arctic Terrestrial ecosystems: A hierarchical approach to analysing landscape change. *J. Appl. Ecol.* **28,** 244–276.

Population Growth and Energy

GAYL D. NESS
University of Michigan
Ann Arbor, Michigan, United States

1. **A Basic Observation**
2. **Early Modern Population Growth and Energy Use**
3. **1700–1950: The First Modern Population Growth Spurt**
4. **1950–2050: The Second Modern Population Growth Spurt**
5. **Modern Population Growth and Energy**

Glossary

crude birth rate The number of births per 1000 population.
crude death rate The number of deaths per 1000 population.
demographic transition A transition from high to low birth and death rates.
epidemiological transition A transition from high, variable death rates to low, stable death rates.
infant mortality rate The number of infant (under 1 year of age) deaths per 1000 live births; this is considered one of the best overall measures of a society's level of health and welfare.
least developed regions The United Nations designation for the world's poorest 48 countries, mostly from sub-Saharan Africa, South Asia, Laos and Myanmar in Southeast Asia, and Haiti in the Caribbean.
less developed regions The United Nations designation of poorer countries; includes most of Asia, except Japan; Africa, except South Africa; and Latin America.
more developed regions The United Nations designation for the world's wealthy, industrial nations.

The basic outlook of a long-term positive relationship between population growth and energy use is complicated by observations of different historical periods. The early modern period, in the 15th and 16th centuries, saw a new technology that harnessed wind for sails, bringing an increased population growth in Europe and Asia and a demographic collapse in the Americas. Modern population growth from the 18th through the 20th centuries was based on fossil fuels and industrialization, with distinct differences between the North Atlantic community and the rest of the world. The present is marked by great increases in both population and energy use at the global level, with considerable variance at the level of individual countries.

1. A BASIC OBSERVATION

1.1 A Close Correlation

The basic picture of population growth in our era, the past 1000 years, is a familiar one. Figure 1 traces only the past 1000 years. The margins could easily be pushed back 100 or 200,000 years without altering the basic pattern. We have had thousands of years of slow growth, followed by exponential rates of very rapid growth over only the past three centuries. This basic pattern of growth, presented in Fig. 1, can be related to the pattern of growth in energy consumption. For this measure, a surrogate, carbon emissions, provides a picture of the past two centuries. Although this does not directly indicate energy use, it indicates the major source of energy for the past two centuries. Combustion releases carbon into the atmosphere in the form of carbon dioxide. First wood, then coal, and finally oil were the major sources of energy for the recent past (Fig. 2).

The correspondence of the two growth patterns seen in Figs. 1 and 2 is striking and indisputable. Together, these charts suggest that population growth produces increased energy consumption, or that increased energy consumption produces population growth. To be sure, if population growth were considered to occur with a constant technology (for example, using wood for heating and cooking), the juxtaposition of the growth curves would be wholly understandable. More people means more wood is needed for heating and cooking. But there are complicating conditions.

 © 2004 Elsevier Inc. All rights reserved.

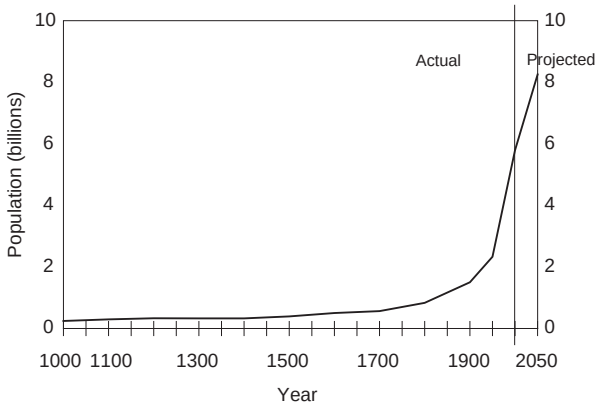


FIGURE 1 World population growth, 1000–2050. Adapted from Ness (1993).

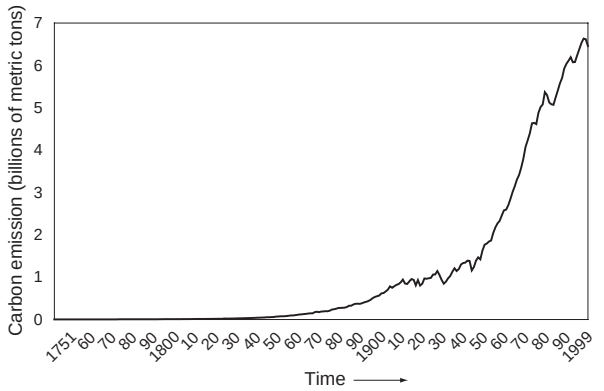


FIGURE 2 World atmospheric carbon emissions, 1751–1999. Adapted from CDIAC (2001).

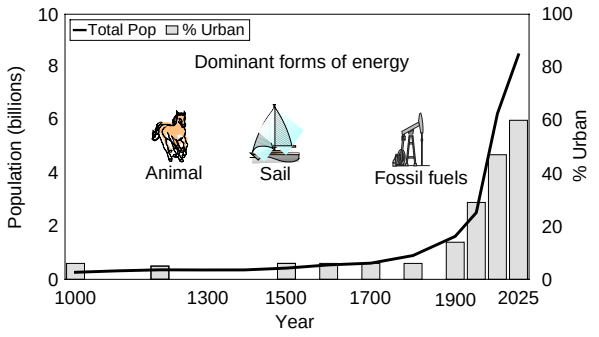


FIGURE 3 World population growth, urbanization, and energy changes.

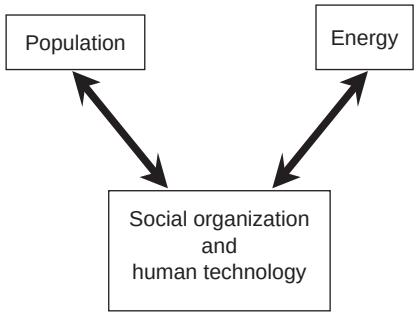


FIGURE 4 Population and energy relationships.

1.2 Population Growth and Energy: A Complex Relationship

Ben Franklin, for one among many, showed how humans could build a more efficient stove, using less wood for more heat. Thus a change in technology can change the tight correspondence between population and energy use. This can be seen more clearly if the data in Fig. 1 are complicated by adding an indication of major changes in energy technology and social organization. This is seen in Fig. 3, which shows that the recent pattern of population growth was accompanied by a revolution in energy, from animal to sail to fossil fuels, and by an even more radical revolution in human social organization, from rural agrarian society to urban industrial society. This implies a complex change in both energy converters (animals, sailing ships, and heat engines) and sources of energy (animal feed, wind, and carbon sequesters). Obviously, the causal paths run in both directions. It is this observation that has led to a more complex picture of the relation between population growth and energy. It is a perspective borrowed from the discipline of human ecology. The basic model is seen in Fig. 4, but can easily be stated here: There is no direct relationship between population growth (or decline) and energy. All relationships between population and energy flow through and are mitigated by some form of human technology and human social organization.

The complex population and energy relationship can be seen in three distinctive historical periods over the past five centuries. The first period shows a small population growth spurt in the 16th century after the revolutionary progress in wind energy opened the world to a new trading system. The second shows how the modern "North Atlantic" population community grew to its present state, related to the beginnings of the fossil fuel energy system. The third focuses on the past half-century, when a new and more rapid period of population growth outside the North Atlantic Community occurred. This has been accompanied by a worldwide intensification of energy

use as the less developed countries of the world began to push for modern economic development. It has also been accompanied by both a demand and a vision for a new energy revolution that will lead to "sustained development." In all of these illustrations, it can be seen how human social organization and human technology affect the relationship between population growth and energy.

2. EARLY MODERN POPULATION GROWTH AND ENERGY USE

Perhaps one of the most dramatic illustrations of the complex relationship between population and energy is to be found in the 15th and 16th centuries. As Fig. 2 suggests, there was a great energy transformation in the 15th and 16th centuries and a slight increase in population growth rates. Until 1500, most of world trade and production were carried on by animal and human power. The Silk Route moved goods (and microbes, of which more shortly) between Asia, the Middle East, and Europe for at least the past 2000 years. In all this period, the Americas were isolated and free of Euro-Asiatic contact. That changed in the latter part of the 15th century. Columbus made the sea link between Europe and the Americas in 1492; Vasco da Gama made the sea link between Europe and Asia in 1498, and Magellan's ship completed the first circumnavigation of Earth in 1523. In the words of J. H. Parry, the west "discovered the seas" and tied the continents together in what would become a single ecosystem.

These new sea links were based on a remarkable revolution in technology, especially in naval architecture, between 1400 and 1450. This has been called the transformation from oars and swords to guns and sails. Until the 15th century, Western European, especially Mediterranean, naval architecture was based on two types of vessels. For war, there was the oar-propelled "galley," manned by slaves or freemen moving what was basically an infantry to engage in hand-to-hand combat on the sea. For trade, there were crude single-masted, square-rigged ships, often using only a large oar over the lee stern for steering. In the first half of the 15th century, the Portuguese developed the caravel, a revolutionary three-masted ship. The square sail amidships was balanced with a lateen sail in the stern and a third sail forward; all supported by rigging between the three masts. Ultimately, the hull was raised higher out of the water and fitted with canons. This new ship proved exceptionally powerful, using wind and gunpowder to carry men and goods and to fight at sea throughout the world. It did more as well; it provided the basic design for sailing ships that grew larger and larger for the next 500 years. With this new rigging design, masts could be extended upward and increased in number, adding acres and acres of wind-catching sail. An immediate result of this new energy transformation included two dramatically different impacts on world population. The population growth rates in Europe and Asia rose from their long-term rate of less than 0.1% per annum to about 0.2% per annum. In the Americas, on the other hand, there was a massive demographic collapse.

European and Asian populations grew as new crops from the Americas quickly found their way across the seas and raised the carrying capacity of the land. Potatoes, sweet potatoes, beans, ground nuts, tomatoes, chili peppers, and corn became standard new crops in Asia and Europe, enhancing the calorie-producing power of the land, reducing mortality, and increasing population growth rates. This was an early example of what has come to be called the epidemiological transition. Historically, death rates have been both high and variable. In agrarian societies, annual changes in the weather could produce large swings in the death rate. A long winter and bad summer meant reduced nutrition and higher death rates. Better weather the following year would see an increase in nutrition and a substantial reduction in the death rate. With modernization, the death rates declined, and they also became more stable.

In return for the new crops, the Americas were given Euro-Asian microorganisms. Two millennia of Euro-Asian trade had moved diseases back and forth and had created a degree of immunity in those populations. Lacking such contact for more millennia, the Americas developed no such immunities. Measles, typhoid, and small pox proved to be deadly imports to the Amerindian populations. Within a century of Columbus' landing, the entire native population of the Caribbean was wiped out. The large and concentrated Aztec and Inca populations were literally decimated. Later, the same thing would happen to the smaller and more scattered tribal Indian societies in the rest of South and North America. This is perhaps one of the most dramatic illustrations of the complex nature of the impact of energy changes in population. The specific outcome was the result of a technological change that increased energy use, which had dramatically different impacts on different populations. The sail revolution can be argued to involve a decline in

energy use, at least in the human and animal energy used to produce and transport goods. It can also be seen as a net gain in energy: the energy gained by sails in comparison with the energy required to harness the wind. It also overcame a major limitation of the swift and maneuverable galleys, which were tied to the land because of the food and water required to feed the men who moved them. In any event, this was a technological innovation that made extensive and efficient use of a form of energy previously used only sparsely and inefficiently. This energy transformation directly produced an increase of population growth in one part of the world, and a demographic collapse in another.

TABLE I

Regional Distribution of World Population, 1500–2050

	Year[a]					
Region	**1700**	**1800**	**1900**	**1950**	**2000**	**2050**
World population[b]	610	902	1622	2515	6057	9322
Europe	20%	20%	24%	23%	12%	6%
North America	<1%	1%	5%	7%	5%	5%
Asia	68%	69%	60%	55%	61%	58%
Africa	10%	8%	7%	9%	13%	21%
Latin America	2%	2%	4%	7%	9%	9%

[a] Percentage of worldwide population in a given year.
[b] Millions.

3. 1700–1950: THE FIRST MODERN POPULATION GROWTH SPURT

The early modern population growth spurt brought European and Asian population annual growth rates to about 0.2% by 1700. From this point, Asian growth stagnated for the next two centuries while Europe began a new growth spurt. This was the result of what has come to be known as the demographic transition, or the transition from high to low birth and death rates. From about 1700, mortality in Europe began a long-term secular decline. Trade, agricultural, and industrial revolutions, plus a moderate global warming, gradually increased the standard of living, which lowered crude death rates from about 25–30 to 10 by about 1900. It was especially the industrial revolution that saw a dramatic rise in energy consumption. The new steam engines were powered first by wood, then the dramatic shift was made to fossil fuels, coal, and, later, oil and natural gas. Here increased energy use reduced mortality and led to increased population growth.

Although the mortality rate was decreasing throughout the 18th and much of the 19th centuries, the past levels of high fertility (crude birth rate: 30–35) remained high, producing a period of rapid population growth. In 19th-century Europe and its American extension, annual rates of population growth were about 0.8–1.0%. By the end of the 19th century, however, birth rates began to decline, coming into line with low death rates by about 1950. Thus this early demographic transition, from high to low birth and death rates, was completed in about two centuries. This coincided with the transition from rural agrarian to urban industrial society, which, among other things, transformed children from an economic asset to an economic liability. This helps to explain the decline in fertility. In addition, the period saw the expansion of the European population to the Americas, and especially to North America, creating a new "North Atlantic" community marked by the new form of urban industrial society. Thus again, the relationship between population growth and energy appears as a complex and indirect one. Increased energy consumption came with the expansion of the machines of the new industrial society. Greater energy consumption implied lower costs of transportation and higher levels of human productivity. This facilitated urban living, implying both a decline in the economic value of and an increase in the economic costs of children. This ultimately led to lower levels of fertility and population growth. Thus, increased energy use led first to more rapid population growth, and finally to a slowing of population growth. This population growth spurt was region specific. The most rapid growth occurred in Europe and North America, changing the regional distribution of the world's population. This can be seen in Table I.

4. 1950–2050: THE SECOND MODERN POPULATION GROWTH SPURT

The demographic transition that was completed in the Western World by the mid-20th century is now taking place throughout Asia, Latin America, and Africa. In its population reports, the United Nations refers to this part of the world as the less developed regions (LDRs). The World Bank and others refer to it as "the world of developing or less developed"

countries. The "least developed" regions, consisting basically of much of South Asia and sub-Saharan Africa, are a subgroup of the LDRs. The United Nations contrasts this with what are called the more developed regions (MDRs), i.e., the North Atlantic community, Australia, New Zealand, and Japan.

The pattern of this new demographic transition is the same as in the past: mortality decreases first while fertility remains high, producing a period of rapid population growth. Then fertility declines, coming into line with the lower mortality, and the overall growth rate slows. Although the pattern is the same, the rates, absolute numbers, and timing are radically different. First, levels of mortality and fertility in the LDRs at the beginning of the transition were much higher than they were at the beginning of the earlier MDR transition. Second, the mortality decline in the LDRs came very rapidly, following major medical advances that reduced mortality rapidly after World War II. Declines that took two centuries in the MDRs now came in two to four decades. This produced a much higher rate of population growth in the LDRs. Whereas annual population growth rates in the past transition were about 1% or less, population growth rates in this current transition are 2.5% and higher. Third, the fertility decline that is now taking place in the LDRs can occur more rapidly than it could in the past. Past fertility declines that might have required one or two generations in the past can now take place in two decades, or one generation.

The major cause of the change in the speed of the transition lies in two technological breakthroughs and associated changes in public policy and large-scale organizations. In the middle of the 20th century, a new medical technology was developed that could attack infectious diseases. Antibiotic drugs could stop infections and chemical pesticides could kill the vectors of infectious diseases such as malaria. These new techniques could be carried rapidly to the LDRs via newly created worldwide organizations such as the United Nations World Health Organization, nongovernmental organizations, and expanding private market organizations. The result was an exceedingly rapid decline in mortality and a rapid rise in population growth rates. World population growth rates stood at 1.79% in 1950/1955. They peaked at 2.04% in 1965/1970, then began to decline. For the LDRs, however, the peak in 1965/1970 was 2.51%. The MDR population growth rates peaked at 1.20% per year in 1950/1955 and have been declining steadily since then.

The modern fertility decline that is now occurring, and has been completed in many of the LDRs, can also occur more quickly than in the past. This is due in large part to the development of non-coital-specific contraceptive technology, which emerged only in the 1960s. Like new medical technology, new contraceptive technology can be disseminated by large-scale organizations, which developed rapidly after 1960 in the form of private and public national and international family-planning programs. Moreover, these technological and organizational developments were accompanied by a radical change in national and international population policies. Most national population policies before 1952 were pronatalist. Governments supported high fertility and often banned programs and activities that promoted birth control or deliberate fertility limitation. This changed in 1952, when India became the first government to announce a national policy to reduce population growth by reducing fertility in marriage. This began what has come to be called the "antinatalist policy revolution." Comparable policy changes became common in the rest of Asia, followed by Latin America and Africa in the next three decades. In 1968, the United Nations resolved that it is a basic human right that couples be allowed to choose the number and timing of their children and to be assisted in doing so. Today most countries in the LDRs have policies that promote or allow the promotion of fertility control programs and contraceptive distribution. Notable exceptions are found primarily in select group of Muslim countries, though many Muslim countries, such as Bangladesh, Indonesia, Iran, and Tunisia, actively promote effective national family planning programs.

Just as the first modern population growth spurt differed by world region, the current transition also differs by world region. Asia led the way in policy and reproductive behavior changes, followed by Latin America and then by Africa, which still lags considerably behind in fertility decline and thus shows continuing high growth rates. The resulting changing rates of population growth in these regions over the past 50 years can be seen in Table II.

5. MODERN POPULATION GROWTH AND ENERGY

The relation between population growth and energy currently shows two different patterns, at the global level and at individual country levels. At the macro (worldwide) level, there is a close correspondence between population growth and energy consumption,

TABLE II

World Population Growth Rates by Region, 1950–2050[a]

Region[b]	Annual average (%)					
	1950–1955	1965–1970	1970–1975	1980–1985	1990–1995	2000–2005
World	1.79	2.04	1.93	1.71	1.49	1.23
MDRs	1.20	0.83	0.78	0.59	0.44	0.16
LDRs	2.06	2.51	2.35	2.06	1.76	1.48
Asia	1.93	2.42	2.25	1.85	1.58	1.26
Africa	2.17	2.59	2.61	2.87	2.54	2.33
Latin America	2.65	2.58	2.45	2.08	1.72	1.42

[a] *Source*. United Nations, *World Population Prospects*, the 2000 Revision.
[b] MDRs, More developed regions; LDRs, less developed regions.

with some highly suggestive changes in the past two decades. At the micro (individual country) level, there is considerable variance in the relationship. Each requires some discussion and illustration.

5.1 Global-Level Population and Energy

For the past two centuries, there has been a very close correspondence between world population growth and energy consumption, as measured by atmospheric carbon emissions. Figure 5 shows the overall global relationship; Fig. 6 shows the relationship between carbon emissions and emissions per capita of world population. Both figures show a clear positive relationship between population growth and energy consumption. Energy consumption grew only slowly in the 19th century but showed exponential growth in the 20th century. The urban industrial society based on fossil fuels brought with it higher standards of living and substantial population growth. In the 20th century, energy consumption grew more rapidly than population did, imposing an increase in per capita energy consumption. This has begun to change in the past two decades, however, because per capita consumption, or carbon emissions, appear to have reached a plateau and may even be declining slightly. As shall be seen shortly, this declining per capita consumption has a number of causes. First, there has been a change from more polluting to cleaner fuels, so that carbon emissions can decline even as actual energy consumption rises. Second, there has been a structural change in the industrial world, showing a decline in the manufacturing and agriculture industries, based on the proportion of wealth they generate, and a rise in the service industries. The service industries are less

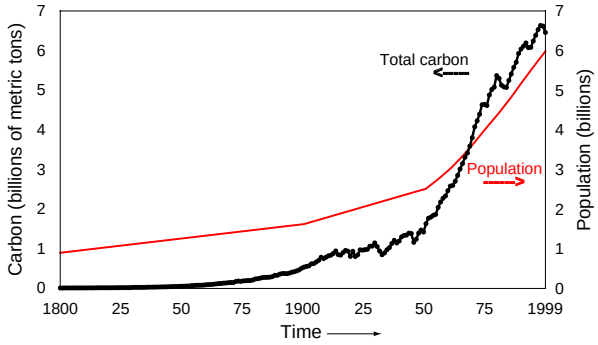


FIGURE 5 World atmospheric carbon emissions and total population.

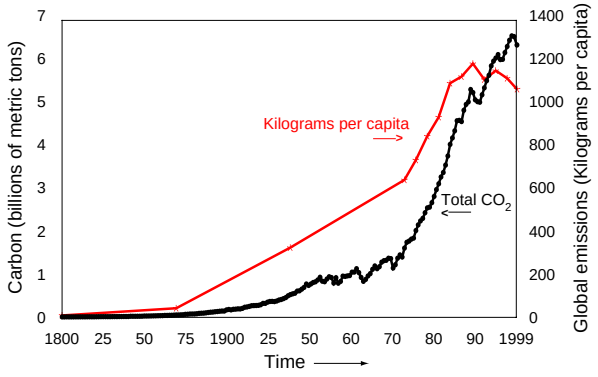


FIGURE 6 World total and per capita atmospheric carbon emissions. Adapted from CDIAC (2002) and Ness (1993).

energy demanding compared to the older manufacturing industries, thus energy use can decline even as population grows and economic development progresses. Finally, the technology of energy consumption has become much more efficient, requiring less energy for a growing population and growing wealth.

None of these changes has been universally constant, however. The specific conditions of individual countries and blocks of countries have broad effects. Individual countries differ by resource endowments, by levels of development, and by specific energy policies, both explicit and implicit. To understand these patterns requires turning to a series of individual country illustrations and comparisons.

5.2 Individual Country Variance

There are three important points to note here. First, the wealthy countries differ from the poor countries in levels of energy consumption and in the trajectories of growth. Second, wealthy countries differ greatly among themselves in energy consumption and its trajectory. Third, the less developed countries also differ among themselves in levels of energy consumption and patterns of population change.

5.2.1 Wealthy Countries and Poor Countries

The first point can be seen readily in Table III. The populations of the more developed regions are growing only very slowly; their per capita energy consumption is about six times as high as that in the LDRs. At the same time, energy consumption is actually declining slightly, though it can be seen that this is due largely to Europe's substantial 20% decline from 1987 to 2000. This was primarily due to the economic collapse of the Soviet Block countries, however, as well as improvements in efficiency in other parts of Europe. Compared to the MDRs, the LDR population and per capita energy use are growing much more rapidly.

In the proportion of energy from renewable sources, there is also a substantial difference between the wealthy and poor countries. Poor countries derive about a quarter of their energy from renewable sources, and for the poorest, the proportion can reach 90%; the rich countries take only 6% from such sources. But the differences are even greater than this implies. A great part of the renewable source in the poor countries is wood, implying a low technology. In the wealthy countries, much of the renewable sources are high-technology sources such as waste recovery, hydroelectric, solar, and wind.

TABLE III

Population and Energy Consumption by World Regions[a]

			Energy consumption, 1997					
						Nonrenewable (%)		
Region[b]	Population (millions)	Growth (%)	Total[c]	Change (%)	Per capita (kg)	Total	Nuclear	Renewable (%)
World	6057	1.35	9,521,506	+16	1635	86	7	14
MDRs	1191	0.3	5,827,461	−4	4505	94	10	6
LDRs	4865	1.62	3,631,617	49	803	75	1	25
Asia	3414	1.1	2,958,844	36	899	79	4	21
Japan	127	0.26	514,898	37	4085	96	16	4
Europe	727	−0.04	2,553,858	−20	3507	95	12	5
France	59	0.37	247,534	19	4233	93	42	7
MENA	432	2.28	525,927	58	1388	96	0	4
SSA[d]	651	2.56	—	—	—	—	—	—
Nigeria	114	2.74	88,652	36	853	78	0	22
S. Africa	43	1.57	107,220	15	2766	89	3	11
N. America	314	1.04	2,400,174	16	7947	94	8	6
C. America	173	1.52	1,98,317	26	1202	82	1	18
S. America	346	1.38	379,732	35	1148	72	1	22
Oceania	31	1.37	118,305	31	4038	91	0	9

[a] *Source*: *WRI World Resources 2000–2001*, World Resources Institute, Washington, D.C.
[b] MDRs, More developed regions; LDRs, less developed regions; MENA, Middle East and North America; SSA, sub-Saharan Africa.
[c] Thousand metric tons oil equivalent.
[d] Complete data for Africa are not available; Nigeria illustrates conditions in sub-Saharan Africa.

Economic development and technology are the forces driving the changing patterns of energy consumption, and they come with low population growth rates, which have even turned negative in some cases.

5.2.2 Differences among Wealthy Countries

Most wealthy countries have seen an increase in energy efficiency, implying a slower growth of per capita energy consumption and even a reduction in the energy costs of increasing wealth. Figure 7 illustrates this process over a long time series in the United States. Over the past 130 years, U.S. real per capita output has steadily increased. At the same time, the national population has grown from 40 million to 280 million. Carbon emissions from fossil fuel emissions per $1000 GNP have increased steadily throughout the first two decades of the 20th century then began a steady decline. Many countries experienced similar trajectories as coal was replaced with oil and natural gas and combustion processes were made cleaner and more efficient.

In the latter half of the 20th century, wealthy countries began to differ markedly in their patterns of energy consumption. As can be seen in Table III, Europe has begun a substantial reduction, 20% in the past 13 years, of total energy consumption. As noted, however, this comes primarily from the Soviet economic collapse, which has also been accompanied by a dramatic reduction in population growth rates, even to actual declines in population. Still, both Japan and Europe have much lower levels of per capita energy consumption than do the United States and Canada, for what is essentially the same standard of living. France has lowered fossil fuel consumption with a large shift to nuclear power, which now accounts for about 42% of domestic energy consumption. Japan has also built a substantial nuclear facility, which gives it about 16% of its energy. Both Europe and Japan levy heavy taxes on gasoline consumption, leading to more extensive use of smaller and more efficient vehicles. The United States oil tax policies do not encourage the kind of efficiencies found in Japan and Europe. Thus, low population growth rates in the wealthy countries are accompanied by a wide range of different policies and technologies. These produce very different levels of energy consumption.

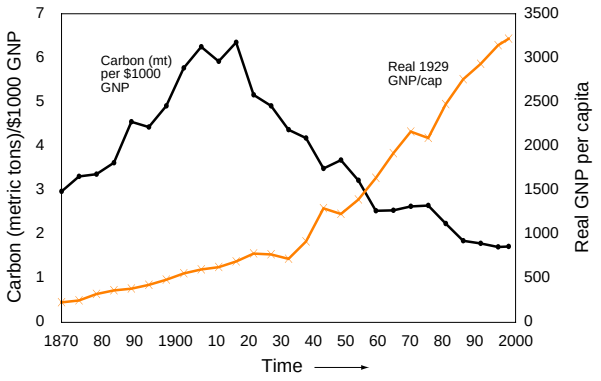


FIGURE 7 United States real gross national product (GNP) per capita and carbon emissions per $1000 real GNP.

5.2.3 Differences among Poor Countries

There are considerable differences among poor countries in the relationship between population growth and energy. This, too, is due to different ecosystems and resource endowments, but more importantly to different policies with respect to economic and social development and environmental protection. Almost all poor countries officially espouse the aim of promoting economic development. Most also promote, either actively or passively, some form of fertility limitation to reduce population growth and promote human health. They differ remarkably, however, in their capacities to turn those aims into effective programs. Thus they differ in the population growth–energy relationship. Table IV uses three pairs of roughly comparable countries, chosen primarily for their differences in levels of welfare, as measured by the infant mortality rate.It shows their population levels and growth rates, energy consumption, and level of economic development.

Bangladesh and Pakistan are two large, poor, Muslim countries in South Asia. They share a long history of Hindu–Muslim culture and more than two centuries of British colonial rule. Both countries actively espouse aims of economic development and human welfare. Both are experiencing the recent demographic transition now sweeping the less developed regions. Bangladesh has made more progress than has Pakistan in reducing its fertility and thus its population growth rate. This has also given it a slightly higher level of human health and welfare, as indicated by its lower infant mortality rate. But its per capita rate of energy consumption is only two-thirds that of Pakistan, and its energy consumption growth rate is also only two-thirds that of Pakistan. It is slightly poorer than Pakistan, but at the moment its rate of economic development is substantially greater than Pakistan's. Here, then, slower population growth and lower economic development are

TABLE IV

Poor Country Comparisons on Population, Welfare, Energy, and Economic Development

Country	Population (millons)	Growth (%)	IMR[a]	Energy consumption			GNP per capita[c]	PPP per capita[d]	Growth (%)
				Total[b]	Growth (%)	Per capita (kg)			
S. Asia									
Bangladesh	137,439	2.12	79	21,894	27	179	370	1680	3.3
Pakistan	141,256	2.66	95	42,048	39	292	420	1920	0.9
SE Asia									
Thailand	62,806	1.34	25	46,166	124	773	1970	6550	0.9
Myanmar	47,749	1.48	92	12,249	14	279	300	600	0.5
Africa									
Zimbabwe	12,627	1.91	65	8154	9	727	480	2340	−9.8
Mozambique	18,292	2.31	137	6994	−1	379	210	1000	6.7
Latin America									
Costa Rica	226	2.48	12	1157	14	309	3950	8080	−1.0
Guatemala	11,385	2.64	42	4433	47	421	1670	3850	−0.06

[a] IMR, Infant mortality rate.
[b] Thousand tons oil equivalent.
[c] GNP, Gross national product.
[d] PPP, Purchasing power parity, expressing gross national income in terms of local buying power.

associated with slightly lower levels of energy consumption.

Thailand and Myanmar also share a long history of Theravada Buddhist culture in a riverine wet rice ecosystem. Myanmar came under British colonialism in the early 19th century, achieving independence in 1948, after about a century of colonial rule; Thailand managed to stave off colonial penetration through astute diplomacy, balancing British and French overseas interests. Both countries emerged in 1950 with strong aims of promoting economic development and human welfare. Thailand has been far more successful than Burma in both its development and population policies. With one of the world's most effective national family planning programs, it brought fertility from high traditional levels to below replacement in just two decades. Its associated health and education programs have given it a much higher quality of life, with an infant mortality rate about a quarter of that in Myanmar. It is wealthier and has experienced much higher rates of economic development, and its energy consumption has greatly outstripped that of Myanmar. Both countries began in 1950 with roughly equal levels of energy consumption; today Thailand's total consumption is almost four times Myanmar's and its per capita level is about three times Myanmar's. In this case, much slower population growth and more rapid development are associated with greatly advanced energy consumption.

In southeastern Africa, Zimbabwe and Mozambique share roughly similar landscapes. Until relatively recently, Zimbabwe had a relatively progressive and effective program for economic development and social welfare. Its fertility limitation program helped bring down fertility and the rate of population growth, producing a welfare benefit, as seen in its lower infant mortality rate, which is only about half that in Mozambique. Zimbabwe is substantially wealthier and consumes almost twice as much energy per capita as does Mozambique. As in Southeast Asia, slowed population growth and greater welfare are associated with higher level rates of growth in energy consumption.

Finally, Costa Rica and Guatemala share the topical ecosystem of Central America, the long period of Spanish colonialism, and almost two centuries of political independence that failed to bring much development until the past half-century. During this time, Costa Rica embarked on a program of democratic economic development and welfare promotion that has made it one of the most advanced societies in Central America. It has had an effective fertility limitation program that has brought down the population growth rate and improved human welfare. Guatemala lagged

considerably, under both external pressures and an internally repressive government. Costa Rica is also substantially wealthier than Guatemala, yet Guatemala consumes about a third more energy per capita than Costa Rica, and its growth rate of energy consumption is about four times that of Costa Rica. Here, then, is a case in which slower population growth and higher wealth and welfare are associated with lower levels of energy consumption.

Although these are only illustrations, they show clearly how the relationship between population growth and energy consumption can be highly complex and highly varied. The near future holds promise of some striking changes in both population and energy. Population growth rates have been declining worldwide since 1965 and are expected to continue that decline. Current United Nations projections foresee a world of some 9.3 billion in 2050. The MDRs will continue to grow very slowly to 2025, when they are projected to reach 1.218 billion, after which they will decline slowly. The LDRs, on the other hand, are expected to continuing growing from their current 4.9 billion to 8.1 billion in 2050, when they are expected to be still growing at 0.57%. Because the biannual projections over the past few decades have consistently reduced the expected totals and growth rates, it is quite possible that the current estimates will turn out to be too high as well. The rising LDR population will surely increase the demand for more energy. It remains unclear what that energy will be. Certainly fossil fuels will continue to play a major part in consumption. But just as clearly, major changes can be expected. There is likely to be a gradual increase in energy efficiency, which will reduce the energy costs of economic development. There may also be a major revolution to renewable sources of energy, which will dramatically change the population–energy connection. It is expected, however that that connection will continue to be indirect, complex, and many faceted.

SEE ALSO THE FOLLOWING ARTICLES

City Planning and Energy Use • Consumption, Energy, and the Environment • Cultural Evolution and Energy • Development and Energy, Overview • Ecological Footprints and Energy • Economic Growth and Energy • Global Energy Use: Status and Trends • Heat Islands and Energy • Land Requirements of Energy Systems • Sustainable Development: Basic Concepts and Application to Energy • Urbanization and Energy

Further Reading

Carbon Dioxide Information Analysis Center (CDIAC). (2001). "Fiscal Year 2001 Annual Report." CDIAC, Oak Ridge National Laboratory, Oak Ridge, Tennessee.

Carbon Dioxide Information Analysis Center (CDIAC). (2002). "Global, Regional and National Annual CO_2 Emissions from Fossil Fuel Burning, Cement Production and Gas Flaring: 1751–1998." CDIAC, Oak Ridge National Laboratory, Oak Ridge, Tennessee.

Cipolla, C. M. (1965). "Guns, Sails and Empire: Technological Innovation in the Early Phases of European Expansion, 1400–1700." Minerva Press, New York.

Coale, A., and Watkins, S. C. (eds.). (1986). "The Decline of Fertility in Europe." Princeton University Press, Princeton, New Jersey.

Livi-Bacci, M. (1992). "A Concise History of World Population [transl by Carl Ipsen]." Blackwell, Cambridge, Massachusetts.

McEvedy, C., and Jones, R. (1978). "Atlas of World Population History." Penguin Books, New York.

Ness, G. D. (1993). The long view: Population environment dynamics in historical perspective. *In* "Population Environment Dynamics: Ideas and Observations" (G. D. Ness, W. D. Drake, and S. R. Brechin, Eds.), pp. 33–56. University of Michigan Press, Ann Arbor.

Ness, G. D., and Ando, H. (1984). "The Land is Shrinking: Population Planning in Asia." Johns Hopkins University Press, Baltimore.

Parry, J. H. (1974). "Discovery of the Seas." Dial Press, New York.

United Nations. (2000). "World Population Prospects, the 2000 Revision" [revised every 2 years]. United Nations, New York.

Viola, H. J., and Margolis, C. (eds.). (1991). "Seeds of Change: Five Hundred Years Since Columbus." Smithsonian Institution Press, Washington D.C.

Potential for Energy Efficiency: Developing Nations

RANDALL SPALDING-FECHER
University of Cape Town
Rondebosch, South Africa

JOYASHREE ROY
Jadavpur University
Calcutta, India

YANJIA WANG
Tsinghua University
Beijing, China

WOLFGANG LUTZ
Energy Strategies for Sustainable Development
Ter Aar, The Netherlands

1. **Introduction: The Context for Energy Efficiency**
2. **Types of Potential for Increased End-Use Energy Efficiency**
3. **Economic Potential for Energy Efficiency by Major Developing Regions**

Glossary

economic potential The energy savings achieved if all new investment and retrofit were based on the most energy-efficient technologies that were cost-effective at that time.

energy intensity The amount of energy required to produce a unit of economic output.

market potential The energy saving that is expected to be realized given the existing energy prices, consumer preferences, and energy policies.

suppressed demand When the demand for energy services is low due to income, access, and infrastructure constraints, not because of consumer preferences.

technical potential The achievable energy savings that result from introducing the most energy-efficient commercial or near-commercial technology at a given time, without taking into account the costs or the life of the equipment to be replaced.

welfare potential The economic potential for energy savings when external costs and economic opportunity costs are included.

This article presents the potential for improvements in energy efficiency in developing countries. It begins by discussing the relevant development context as well as additional barriers that energy-efficiency programs face in poor countries. Then, regional overviews of the potential for energy efficiency within key sectors are presented.

1. INTRODUCTION: THE CONTEXT FOR ENERGY EFFICIENCY

1.1 The Development Context in the South: Poverty and Suppressed Demand for Energy Services

The context for end-use energy efficiency improvements in developing nations is their current energy poverty and the large share of the world's population that uses almost no commercial energy. Per capita consumption of primary energy in Africa, for example, is only 13% of the average in Organization for Economic Cooperation and Development (OECD) countries, whereas for all developing countries it is less than one-fourth (Fig. 1). The gap is even greater for electricity consumption (Fig. 2).

It is common in many poor countries for families not to be able to afford enough energy to heat their

 © 2004 Elsevier Inc. All rights reserved.

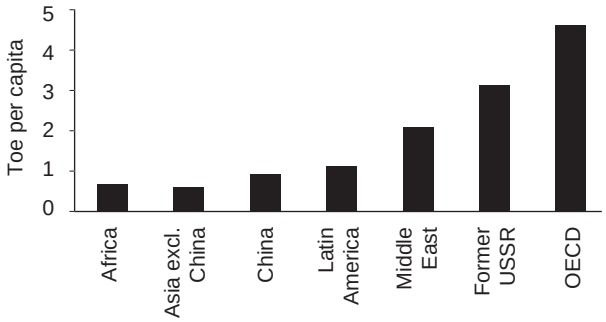


FIGURE 1 Total primary energy consumption per capita. Toe, tons of oil equivalent.

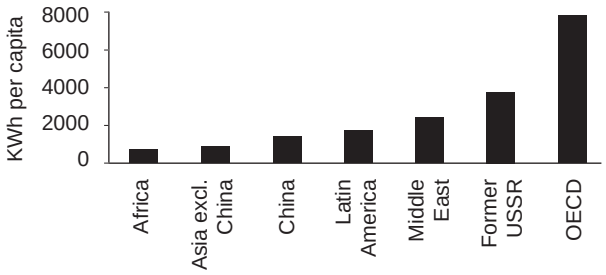


FIGURE 2 Electricity consumption per capita.

homes comfortably and cook sufficient food, let alone benefit from high-quality lighting or television. This energy poverty implies massive unmet (or suppressed) demand for energy services in these countries, with demand constrained by lack of energy infrastructure as well as income. In many African countries, for example, less than 10% of the population has access to electricity, and it is common for more than 80% of residential energy demand to be met by fuel wood.

Given these low levels of energy consumption, how do we know that there is a potential for improving end-use energy efficiency? First, it must be remembered that although developing countries' share of fossil fuel consumption is small, rapid population and economic growth will result in a substantial increase in this share in the early 21st century. From 14% of world primary energy demand in 1971, developing countries will account for 40% by 2010 if present trends continue. Even with aggressive policies to promote energy efficiency, their energy demand is likely to increase 5- to 10-fold during the next 30–40 years, resulting in a 3-fold increase in world energy demand.

Second, one measure of the potential for energy efficiency improvements in developing nations is their much higher energy intensity [energy use per unit of gross domestic product (GDP)] compared to

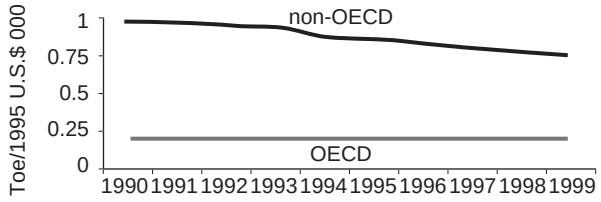


FIGURE 3 Energy intensity (energy use per unit GDP) in the developing and industrialized world.

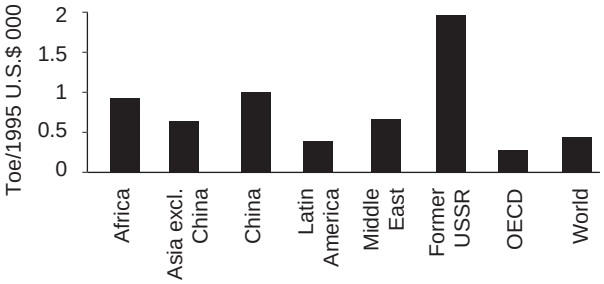


FIGURE 4 Energy intensity by region (energy use per unit GDP).

that of industrialized countries. As Fig. 3 shows, based on International Energy Agency data, energy use per dollar of GDP in the developing world is almost three times that of the OECD countries. If we consider the difference in purchasing power of developing currencies, this picture changes, with developing countries being only approximately 30% more energy intensive. For products that are traded, particularly those exported by developing countries, purchasing power parity does not really matter. In other words, if India is exporting steel and using two or three times as much energy to make a dollar's worth of steel, that is a very high cost and a competitive disadvantage. Of course, it also represents an opportunity to increase energy efficiency in a cost-effective manner. Even within the developing world, however, there are vast differences in energy intensity, as shown in Fig. 4.

1.2 Additional Barriers to Energy Efficiency in Developing Nations

Although barriers to the uptake of energy efficiency have been generally described elsewhere, they take a different form in many developing countries, and some are unique to developing countries. Before discussing different regions, however, it is useful to highlight some of the more significant barriers in developing countries.

1.2.1 Affordability and Financing

Investment in energy efficiency improvements is often constrained by the limited and irregular cash flow in poor households and difficulties in accessing additional finance; they frequently have to rely on hire purchase agreements or "loan sharks." Poor households sometimes invest in fuels and appliances that, ironically, are inefficient in both energy and economic terms. Poverty dictates overlooking long-term factors such as the life cycle costs, efficiency, and safety of an appliance.

For fuel purchases as well, households are often constrained by income flows, with a decision on purchase depending on the availability of money at a particular time. When there is little, kerosene and coal stoves are used because one can buy these in small amounts. Householders know that buying fuels this way is expensive, but not enough money is available to buy in bulk.

Even within industry, a shortage of capital or high interest rates are significant barriers to normal equipment upgrades, not to mention the purchase of more expensive energy-efficient equipment. This means that the rate of capital stock turnover is lower than for similar industries in industrialized countries.

1.2.2 Information

Information and awareness programs need to take into account the wide range of stakeholders involved in the delivery of energy efficiency services. For example, poor consumers may have low levels of literacy or need materials in local languages. Other stakeholders include local government, utilities, manufacturers, developers, builders, and nongovernmental organizations—all with different informational requirements. Often, awareness programs tend to provide broad conceptual information when the target audience requires more sophisticated information. Where people are already aware of broad notions of energy efficiency, for example, information may be needed on practical implementation, such as how to build an energy-efficient house or install a ceiling, where to access financing, how to mobilize capital to finance energy efficiency for the poor, and how to facilitate community participation.

1.2.3 Physical Access to or Availability of Fuels

In certain circumstances, low-income households are unable to secure the best mix of fuels because certain fuels are not readily available. For example, although kerosene networks are good in many countries, with both fuels and appliances being readily available, many poor household do not have electricity connections. Although a few developing countries have started to develop residential bottled gas networks, this is the exception rather than the rule. The coverage of distribution networks is often weak and inhibited by poor transport infrastructure and a lack of access to transport.

1.2.4 Split Incentives: Construction, Ownership, and Use

In the delivery of housing, those making the initial capital investment in constructing a house are generally separate from those who will live in it and pay the operating costs. It is not common, therefore, to find developers investing in energy efficiency. The consequence is that the low-efficiency house has high operating costs. This is a major problem in developing countries with large public housing programs because the government priority is to deliver as many houses as quickly as possible; for both government and contractors, the tendency is to cut corners and minimize initial cost. It is also a problem in industrialized countries.

1.2.5 Lack of Tenure and Urban/Rural Commitment

In many urban informal settlements, appliance ownership is constrained by space and tenure problems. Lack of space, for example, means that in informal electrified settlements and backyard shacks, more households own two-plate electric stoves than stoves with ovens. More important, the tenure problems in informal unplanned settlements and shacks play a direct role in purchasing electrical appliances or other expensive investments in efficiency. Migrant workers continue to play a large role in many countries' urban communities. They are committed to their rural households and view urban life as temporary and thus tend to save or remit money for the maintenance of the rural households, which obviously implies an even lesser willingness or ability to pay a higher cost for efficient appliances.

1.2.6 Multiple Fuel Use and Household Needs

Multiple fuel use means that households use more than one fuel for the same end use. It is common to use gas, kerosene, coal, and/or electric stoves for cooking. Gas might be used for specific tasks (such as cooking special quick foods) and kerosene appliances used for foods that take a longer time to cook. Therefore, although gas may be a more energy-efficient cooking fuel than kerosene or electricity (and healthier than kerosene), households may not completely stop using kerosene after they buy a gas

stove. For example, they may use kerosene for cooking certain foods or if they run out of gas and do not have enough cash to buy an entire gas canister since kerosene can be purchased in small quantities. Also, in some contexts one appliance serves more than one purpose, as in the case in which a kerosene or coal stove is used for cooking and space heating.

1.2.7 The Symbolic Value of Appliances

Not surprisingly, poor consumers do not simply consider the economics of appliance choices when making decisions: Symbolic value can be as important as functional value. For example, many formal households tend to replace their nonelectrical appliances with modern and sophisticated appliances immediately after electrification, not simply because electricity is cleaner and a more convenient fuel. There is a general perception that nonelectrical appliances are not appropriate for formal households. Having electric appliances attracts respect and envy, as symbols of modernity and comfort, and many people will go far to acquire them: The bigger the appliance, the better. Therefore, consumers may not be attracted to smaller, more efficient appliances unless other features enhanced the sense of "modernity."

1.2.8 Technical and Managerial Skills

Compared to industrialized countries, the industrial and commercial sectors in developing countries have lower levels of skill in both the technoeconomic dimension of energy efficiency and the managerial skill to develop and administer programs. There may be many qualified analysts and professionals, but resources are spread too thinly across huge economies and social needs. This is particularly a problem for small to medium firms.

2. TYPES OF POTENTIAL FOR INCREASED END-USE ENERGY EFFICIENCY

As a background to describing the potential for increasing end-use energy efficiency, it is important to understand the different types of energy efficiency potential. The most important categories are technical, ecnomic, and market (Fig. 5):

- Technical potential is the achievable energy savings that result from introducing the most energy-efficient commercial or near-commercial technology at a given time, without taking into account the costs or the life of the equipment to be replaced. Realizing the full technical potential means scrapping existing appliances and equipment, although sometimes a phased technical potential is described, meaning that all new equipment is replaced with the most energy-efficient technology.

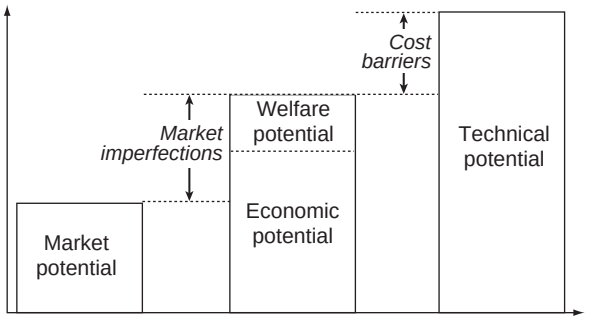


FIGURE 5 Types of potentials for energy savings.

- Economic potential is the energy savings achieved if all new investment and retrofit were based on the most energy-efficient technologies that were cost-effective at that time. Cost-effectiveness at a minimum must include the energy and operating savings on the benefits side and the additional capital, any additional maintenance or control costs, and future replacement costs on the cost side. There are several perspectives on cost-effectiveness. For an individual firm or household, we would only consider the costs and benefits mentioned previously. In a macroeconomic perspective, however, we would need to consider the costs of implementing the energy efficiency programs. We would also include the benefits that could accrue to the broader economy, such as changes in technology costs over time through learning and economies of scale, or other spillover effects and nonenergy productivity gains. Finally, if we consider the societal perspective, we should include the avoided external costs of the energy saved by end-use efficiency measures. This would increase the economic potential since the benefits (avoided costs) increase. This is called the welfare potential of energy savings.
- Market potential is the energy saving that we expect to realize given the existing energy prices, technology, consumer preferences, and energy policies. This definition must take into account the social obstacles to energy efficiency and other market imperfections (i.e., no internalizing external costs) that block the realization of the full economic potential. In addition, market potential may consider whether energy users choose to use their energy

savings to buy more of what is essentially a less expensive service. This is called the take back or rebound effect.

These definitions also help us understand that the role of energy policy in promoting end-use energy efficiency is to both remove the market barriers that keep market potential below economic potential and address cost issues(e.g., through targeted research and development) that can bring the economic potential closer to the phase-in technical potential. Later, when we discuss the potential for energy-efficient improvements, we refer to economic potential unless otherwise noted.

3. ECONOMIC POTENTIAL FOR ENERGY EFFICIENCY BY MAJOR DEVELOPING REGIONS

3.1 India

3.1.1 Trends in Energy Intensity and Consumption per Capita

India, the world's second most populous nation, has seen its population explode from 300 million in 1947 to approximately 1 billion today. Energy consumption has grown rapidly through the process of development. Aggregate consumption has increased approximately 50% during the past decade. The share of commercial energy consumption increased from 26% in 1950–1951 to approximately 60% in 1994–1995. With an elasticity of commercial energy demand (i.e., the percentage change in energy demand for a 1% change in income) of approximately 0.92, and with the objective of rapid economic growth of 6% per annum, a high growth in commercial energy demand is expected. India is the second largest commercial energy consumer within non-OECD East Asia (after China), comprising 19% of the region's total primary energy consumption.

The current low per capita energy consumption (3% of that of the United States and 9% of that of Korea) is an indicator of huge potential for growth in energy demand. Despite an annual approximate 5% growth rate in electricity generation, unmet demand for electricity continues. Although 75% of India's population lives in rural areas, only 29% of rural households have electricity supply and there is no power supply network in many rural areas. In addition to geographical constraints, sparse distribution of dwellings, low ability to pay, and low load make extending the distribution network either impossible or uneconomic for many remote villages. Noncommercial energy still meets 40% of the total energy consumption, which shows the extent of demand that awaits to be met by commercial and more efficient forms of energy.

Energy demand in India has always been greater than supply. Thus, in view of shortages and restrictions, the past trends in consumption of commercial energy do not really represent the growth of energy demand but reflect the growth of its availability. Energy consumption by fuel is dominated by coal (54%) and oil (31%). There is wide scope for enhancing end-use energy efficiency. In addition to replacing outdated technologies and removing market barriers such as pricing distortions, overall economic potential would increase by simply changing the scale of production of many production activities. Such potential exists in all of the energy-using sectors.

Sectoral demand for energy arises mainly from lighting and cooking demand in the residential sector, irrigation and other operations in the agricultural sector, transport of passengers and freight in the transport sector, and fuel requirements in the industrial sector. Fig. 6 shows the use of energy by sector, whereas Fig. 7 shows the trends in energy intensities of these sectors. Decomposition analysis of the energy intensity trends shows that in India increasing economic activity is the main driving force for rising energy consumption, not increasing energy

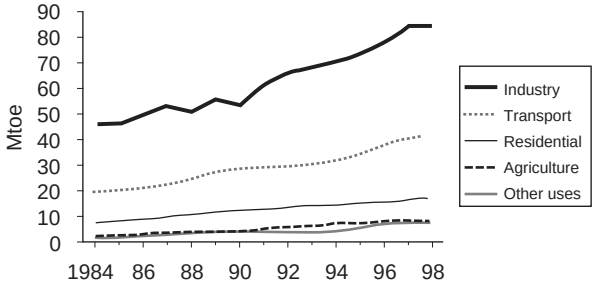


FIGURE 6 Trends in consumption of commercial energy in India, 1984–1998.

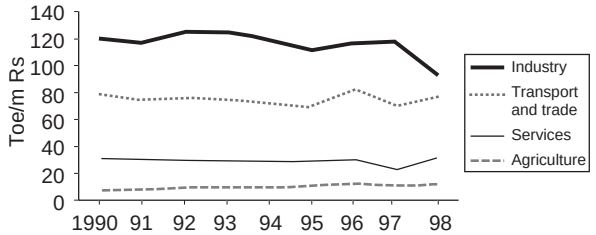


FIGURE 7 Sectoral energy intensities in India (energy/real GDP).

intensity. Since it is difficult to restrict energy demand due to increased output or activity directly at the early stage of development, improved energy efficiency is essential.

A number of efforts to improve energy efficiency, especially through exploration of technical potential, have resulted in a declining trend in energy intensity. The Eighth Plan (1992–1997) included a National Energy Efficiency Program that has the potential to save 5,000 MW in the electricity sector and 6 million tons of oil in the petroleum sector over the 5 years. India established an energy efficiency center as a major agency to implement programs and an energy conservation fund to finance them. On the other hand, electricity prices are far below marginal costs throughout most of the country, as are prices for natural gas and coal. This is a substantial barrier to energy efficiency investments and essentially subsidizes energy sources with environmental and health costs from pollution. One of the major challenges in a country with such high unmet demand is how to correct prices and provide incentives for investment in energy efficiency while protecting the social needs of the poor. A subsidy program for efficient vehicles or appliances without a move toward cost-reflective prices, for example, could substantially increase overall energy demand, as consumers buy more of a much needed service, and not reduce environmental impacts from the sector.

3.1.2 Industrial Sector Potential

In general, Indian industry is highly energy-intensive and its energy efficiency is far below that of industrialized countries. For example, in nitrogenous fertilizer manufacturing, older manufacturing units generally require 8.4–10.5 tons of oil equivalent (toe) per ton of ammonia produced, whereas the standard achieved in a plant in Bristol, England, is only 4.6 toe. An improvement in energy efficiency also includes energy conservation undertaken by a producer or a consumer that reduces losses of energy. In the electricity sector, it has been estimated that older power plants in many developing countries consume 18–44% more fuel per kilowatt hour of electricity produced than plants in industrial countries. Since India has vast reserves of coal, this is the primary fuel in the industrial sector. India has always relied heavily on coal for generating electricity with a low average conversion efficiency of 23%. There is potential for at least a 10% improvement in the conversion efficiency of existing plants. Greater energy efficiency can be implemented through technological enhancements, more research and development, improved energy management, and better operational practices.

A World Bank-sponsored study (1995) of energy demand in five major Asian countries (China, India, Indonesia, Korea, and Thailand) concluded that the industrial sector has the largest potential for energy conservation. This sector accounts for the largest share of final energy consumption in India and is dominated by energy-intensive industries. Higher energy intensities are ascribed to (i) slow capital stock turnover and utilization of out-of-date processes, (ii) poor energy management, and (iii) lack of incentives for firms to invest in energy efficiency due to low energy prices. In addition, scale is a major issue in many industries. In nearly all states in India for the cement and paper sector, and in a few states for aluminum, increasing the scale of production could reduce inefficiency in energy use.

We can also consider a recycling strategy in which secondary materials are fed into the original process, thereby lowering per unit energy consumption. Examples include the substitution of waste paper for paper pulp and using recycled glass and plastics. To improve overall industrial energy efficiency and reduce pollution, there is greater potential from adopting available clean technologies at the beginning of industrial operation rather than "end-of-pipe" solutions, which can actually increase overall energy use.

Technical energy efficiency improvement potential is shown by comparing specific energy consumption (i.e., energy consumption per unit of output) in industrial subsectors. For example, the energy efficiency of ammonia plants can be increased by revamping and modernization. A typical energy efficiency revamp of a plant would reduce the energy consumption for individual plants installed before 1980 by 0.8–2.2 Mtoe and for plants installed between 1980 and 1990 by approximately 0.6–0.7 Mtoe, depending on the feedstock used. The age of the technology, the scale of the plant, and management practices have a major impact on energy efficiency of the overall process. Energy savings potentials are highest for naphtha and fuel oil/low-sulfur heavy stock-based plants built before 1980. With current installed capacity of 5.09 Mt production, energy savings in pre-1980 ammonia plants alone would account for 6.02 Mtoe per year.

In the iron and steel sector, there is a large potential for efficiency improvement. Worldwide, specific primary energy consumption is decreasing with rising scrap-based electric arc furnace production. In India, however, due to a scarcity of scrap metal, this may not be a good option. Still, efficiency

improvement may be achieved indirectly by improving power-generation efficiency that feeds iron and steel plants through modernization and captive combined heat and power generation. In addition, technology conversion (e.g., to induction furnaces) and retrofitting, recycling, and waste heat recovery promise major savings. Although most measures are cost-effective and provide net benefits within a short time, only a few have been or are currently being implemented in the iron and steel sector. Lack of dissemination of information on energy-efficient technology as well as specific information on saving and benefits of saving potential contribute to the low uptake of energy efficiency improvements.

For the cement industry, there is not much scope for improvement in dry processes, but there is more in semidry and wet processes. Besides the technical potential when compared to world best-practice technology, efficiency improvements can be achieved through increased use of additives such as fly-ash and blast furnace slag in manufacturing cement. Structural change toward more production of portland slag cement reduces energy consumption, and using a higher share of blast furnace slag compared to clinker leads to lower energy intensity as well. In India, 20% of final energy could be saved by this twofold structural change, which would decrease the clinker-to-cement ratio by 20%. It has also been estimated that better maintenance and monitoring of plant activity can minimize downtime of machinery and plant, thus avoiding excess energy needed for restarting the process.

In the aluminum sector, the potential for energy efficiency improvement varies across companies. For example, whereas the National Aluminum Corporation could achieve 17%, the Hindusthan Aluminum Corporation could achieve 40%. Energy savings potential in smelting range from approximately 16 to 30%. Better monitoring and control can conserve an additional 2 or 3%. Potential from retrofitting ranges from 10 to 15%, and the costs of this retrofitting and modernization are 25–35% lower than the cost of building an entirely new plant. Table I summarizes the industrial sector potentials for energy efficiency improvements by comparing Indian specific energy consumption to worldwide best practice. Worldwide best practice reflects the best available technology adjusted for India-specific structural factors. For paper, best practice reflects the average energy consumption abroad.

Research on the responsiveness of the Indian industry to changes in energy prices indicates that price-based policies could be effective in improving energy efficiency. Pricing policy in the fertilizer sector, for example, has shown that energy price increases would push less productive and inefficient plants (mostly smaller ones) out of the market or force them to immediately improve productivity and efficiency. It has been estimated that with all other economic variables constant, India industry could achieve worldwide best practice in specific energy consumption by 2005 or 2010 with an average annual nominal energy price increase of approximately 6 or 4%, respectively, measured as an increase in the fuel price index relative to other input prices.

3.1.3 Residential and Commercial Sector Potential

In the residential and commercial sector, the demand for energy is mainly for lighting, cooking, space cooling, and, to some extent in higher altitudes, space heating. The residential sector in India accounts for 40–50% of the total energy consumption. For cooking, 77% of households in India use traditional fuels such as firewood, animal dung, and agricultural residues, 0.3% use electricity, 4% use coal or charcoal, 8% use liquefied petroleum gas (LPG), 7% use kerosene, and 0.5% use biogas. Experience

TABLE I

Specific Energy Consumption in Indian Industry and Worldwide Best Practice, 1993

		India			India
Sector	Units	Average energy consumption	Best practice (marginal plant)	Worldwide best practice	Savings potential (%)
Aluminum	GJ/ton aluminum	86–100	87	65.2–81.9	5–35
Cement	GJ/ton cement	4.00	~2.93	3.06	24
Fertilizer	GJ/ton ammonia	39.2	36.4	28	29
Iron and steel	GJ/ton crude steel	35.4	31.4	17.1	52
Pulp and paper	GJ/ton paper	51.6–80.0	NA	32.0–40.9	~43

tells us that in the ladder of fuel-use transition, households usually move from traditional fuels to kerosene/coal (depending on availability), LPG, and electricity; the use of traditional fuel is indeed an indicator of unmet demand for more efficient fuels.

In India, urban and rural energy problems differ. In urban areas, introducing energy-efficient lamps, better space-cooling devices, and solar passive building structures can enhance efficiency in electricity and fuel use enormously. In the absence of any major demand-side management through policies or appliance standards, however, there is hardly any incentive for the households or commercial sector to enhance efficiency in use. Adopting energy-efficient appliances could provide up to 50% savings in electricity consumption in these end uses. This is particularly important in India, where demand for electricity is expected to grow by 7% per year and the gap between peak demand and peak supply today is more than 11,000 MW.

In rural areas, where almost 75% of the Indian population lives, biomass (primarily firewood) accounts for more than 90% of energy needs. It may not be possible to replace this entirely with more modern energy forms and developing modern biomass technologies will be a key element for addressing the needs of rural areas. One of the important constraints in the wider utilization of renewable energy is the high initial capital cost. India has potential for renewable energy generation, particularly wind and hydropower, and programs in these areas, aided by Global Environment Facility (GEF) funding, are under way. At the same time, India presents a unique opportunity for the large-scale commercial exploitation of biomass gasification technology to meet the needs of the agricultural and rural residential and small-industries sector. The gasifier power could solve the problems of technical and economic inefficiencies to a large extent, thereby reducing the social cost of power supply. Currently, commercial activity in most rural areas depends on power supply from the privately owned, decentralized, diesel power-generating sets, whereas residential lighting is typically met with stand-alone kerosene lamps. Use of kerosene for lighting is particularly inefficient, with only 1% of the energy in the kerosene producing useful light. The Ministry of Non-Conventional Energy Sources aims to replace low-efficiency kerosene lamps in rural areas with efficient renewable lighting technologies.

3.1.4 Transport Sector Potential

Transport shows a constantly rising trend in fuel demand. Decomposition analysis shows that it is the rising demand for overall transportation service that is driving up fuel demand. With relatively low substitution possibilities and high income elasticity for privately owned vehicles, it is expected that fuel demand in the transport sector will increase along with vehicle ownership for years to come. In recent years, fuel switching has been attempted by introducing compressed natural gas in public transport in metropolitan cities. Fossil fuel-based transportation may also be reduced through development of the rail network. Pricing reform is playing an important role in reducing inefficient use of energy. In terms of technical interventions, studies have shown that inefficient use of fossil fuels often results from poor design of public buses, which are not using best-practice technologies. Efficiency in transportation fuel can also be increased remarkably by improving the urban road conditions and reducing traffic congestion.

3.1.5 Agricultural Sector Potential

In the agriculture sector, power use for irrigation holds significant potential for improving energy efficiency. Currently, diesel generator-based water pumps with efficiencies of less than 33% are used to draw underground water; an adequate electricity supply would allow a shift to more efficient electric pumps. Where electricity is now available for agriculture, it is heavily subsidized and charges are not linked to consumption, leading to inefficient use. Correction of prices would lead to increased efficiency through better water and irrigation management. Just an important, more efficient use of water for irrigation would decrease agricultural energy demand for the same agricultural output.

Table II summarizes the potential for energy efficiency improvements in India based on a wide range of studies.

3.2 China and Southeast Asia

3.2.1 Trends in Energy Intensity and Consumption per Capita

China has demonstrated significant improvements in energy efficiency since the 1980s. Energy use per unit GDP declined from 2.18 toe per thousand U.S. dollars in 1980 to only 0.73 in 2000, as shown in Fig. 8. During this period, commercial energy consumption increased 2.3 times, whereas the GDP increased more than six times. However, commercial energy intensity in China is double that of OECD countries and triple if noncommercial energy is included (Fig. 4). Many factors affect energy intensity, including economic

TABLE II

Summary of Energy Efficiency Potentials for India

Sector	Technological options/area	Economic potential (% of current use)
Industry	Fertilizer	29
	Cement	20–24
	Electrical	17
	Thermal	27
	Pulp and paper	20–40
	Textiles	23
	Iron and steel	15–50
	Aluminum	5–35
	Refineries	8–10
	Brick-making	15–40
	Foundries	30–50
	Industrial cogeneration	3500 MW (sugar)
Residential	Lighting	10–70
	Refrigerator	25
	Air-conditioning	10
Agriculture	Pump sets	25–55
Transport	Two- and three-wheelers	25
	Cars	7.5–10
	Trains (diesel)	5–10
	Trains (electric)	5–10

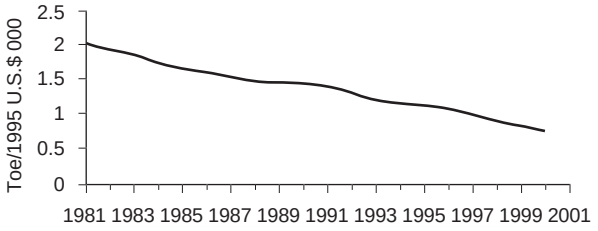


FIGURE 8 Energy intensity (energy/GDP) trends in China, 1981–1999.

structure, product mix, exchange rates, energy resource mix, and technical efficiency. In general, industrial sectors, especially heavy industry such as the iron and steel industry and cement industry, consume more energy per unit of economic output than do light manufacturing, finance, and service sectors. In China, for example, the relative energy consumption per unit of value added in primary (e.g., resource extraction and beneficiation), secondary (e.g., manufacturing and construction), and tertiary (e.g., finance and services) sectors was 12, 68, and 20, respectively, in 1999.

Changes in economic structure have contributed significantly to China's energy intensity improvement. Fig. 9 shows that economic structure has

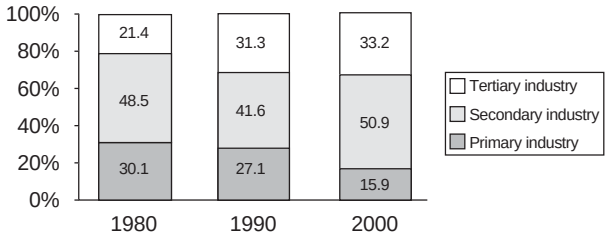


FIGURE 9 Composition of GDP in China, 1980–2000.

changed in China, with a decline in primary industry replaced by rapid growth in the tertiary sector. This trend is expected to continue, and the share of tertiary activities in the total economy will increase. Even within the secondary sector, subsectors with higher value-added and lower energy intensity, such as computer manufacturing and telecommunications equipment production, are developing faster than the traditional heavier manufacturing industries.

China is one of the few countries where coal is the dominant resource for energy use, which is part of reason for China's high energy intensity. For example, even using the latest technology, the highest energy efficiency for combined heat and power generation based on coal as a resource is 60%. Natural gas, on the other hand, can obtain conversion efficiencies of more than 80%. China is currently promoting clean energy production and importing cleaner technologies, which will result in energy efficiency improvements. In Southeast Asian countries, energy intensity is much lower than that of China. Because of the financial crises in the mid-1990s, however, the primary energy consumption per GDP increased, as shown in Fig. 10.

To provide a regulatory framework for energy efficiency interventions, many Southeast Asian countries and China have enacted a range of energy efficiency laws, with differences based on culture, political ideology, and existing institutional structures. Although the scope varies, common provisions include mandatory auditing for industries, mandatory energy building codes, developing performance standards for energy-using equipment, information dissemination, and capacity building. In most cases, implementing the legislation is carried out by existing energy agencies involved in energy efficiency promotion. In Thailand, an energy conservation fund to finance activities and programs for energy efficiency promotion has been set up, whereas for other countries the law specifies budget allocations for energy efficiency programs. For countries without specific legislation, energy efficiency activities are limited to programs that do not require specific

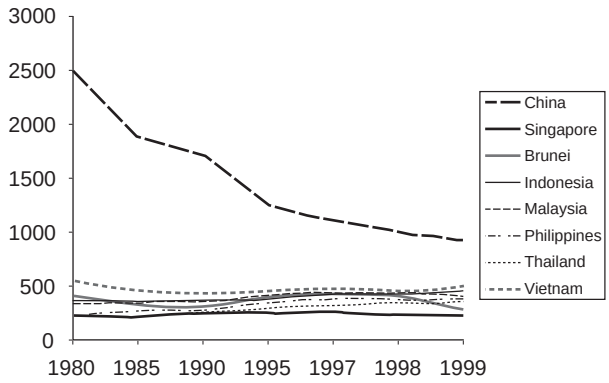


FIGURE 10 Primary energy consumption per unit of GDP.

regulation, such as publicity campaigns, demonstration projects, and programs that can be funded by external grants and loans, government budgets, or payments by consumers and end users. Some energy efficiency programs, such as mandatory appliance standards, appliance labeling, and building energy codes in the Philippines, Malaysia, and Singapore, are implemented through regulation but without reference to any specific law. A wide variety of policy instruments are used in promoting energy efficiency and accelerating the diffusion of energy-efficient technologies. These include appliances standards and building energy codes, appliance labeling, and financial incentives.

Since opening up to the outside world, the Chinese government has paid significant attention to energy conservation projects, including energy policy objectives to "carry out conservation and development of resources simultaneously but giving priority to conservation." To address energy shortages, environmental protection, and sustainable development, the Energy Conservation Law was passed in 1997, providing broad guidance for energy efficiency improvement in all sectors. Since then, governments at national, provincial, and local levels and industrial associations have issued a number of energy efficiency policies and adopted measures for the law's implementation.

3.2.2 Industrial Sector Potential

China's industrial sector is extremely energy intensive and accounted for 76% of the country's total energy used in 1997. It is estimated that the overall final energy use efficiency in the industrial sector was 46% in China, 20% lower than that of OECD countries. The largest energy-consuming industries are chemicals, ferrous metals, and building materials. The energy use per unit of product output provides a better comparison than energy use per unit of production value added. Because of outdated production technologies and processes, slow capital stock turnover, small-scale operation, and poor management and training of staff, most industrial enterprises in China consume more energy to produce a unit of product than applies in developed countries. Even within China, the product energy intensities are quite different in various enterprises. Table III compares energy intensity for selected products between developed countries and China.

There are several technologies available for energy efficiency improvement in industrial sectors. For example, in the cement industry, wet to dry process conversion, on-site cogeneration of heat and power, and coal ash utilization can significantly decrease energy consumption. Based on the available research, it is reasonable that energy efficiency in industry could be increased by 15–20% in China, with a similar situation in Southeast Asia (Table IV).

3.2.3 Residential and Commercial Sector Potential

In 1999, the Chinese residential sector consumed 78 Mtoe commercial energy, accounting for 14% of total final consumption of commercial energy. Including biomass, total residential final consumption was 289 Mtoe. The commercial sector accounts for a further 3% of commercial final consumption. Most residential energy consumption is for space heating, followed by cooking and lighting. Electricity consumption for appliances is increasing rapidly, and the use of air-conditioning is especially important for commercial buildings and it is being used increasingly in households as well.

The rate of construction of new buildings is high due to increases in population, rising standards of living and demands for more living space per person, and the continuing process of urbanization as people migrate toward areas with greater economic development. The majority of the new buildings, however, have poor energy performance. This is not unlike what happened in Europe in the years of reconstruction after World War II. The urgency of supplying new housing and commercial space, pressure to hold down initial investment costs, and low energy prices mean that little attention is paid energy performance. Designs are not adapted to local climatic conditions and ignore the potential for day lighting in commercial buildings.

As a result, energy costs are much higher than expected on economic grounds, even without considering the positive environmental and health impacts of better building design. In the northern part

TABLE III

Energy Intensity in China and Industrialized Countries for Selected Products

Product	Specific energy use	Country	Year
Steel (GJ/ton steel)	60	China (average)	1980
	41		1996
	32		1999
	21	China (best)	2001
	25	United States	Mid-1990s
	18	Japan	
	21	Sweden	
Cement (GJ/ton cement)	6	China (large and medium-sized)	1980
	5		1996
	4	United States	Mid-1990s
	3	Japan	1990
Paper (GJ/ton paper)	46	China	1997
	21	OECD countries	
Ethylene (GJ/ton)	36	China	
	26	OECD countries	
Glass plate (GJ/box weight)	0.8	China	
	0.4	OECD countries	
Ammonia (GJ/ton)	41	China	
	28	OECD countries	
Copper (GJ/ton)	40	China	
	24	OECD countries	
Cotton yarn (kWh/ton)	2300	China	
	2100	OECD countries	

of China, energy consumption for space heating during winter is approximately twice as high as in areas with similar climatic conditions in industrialized countries. As the new building energy code is implemented, new buildings are expected to have a 20% energy use savings compared to existing buildings.

3.2.4 Transport Sector Potential

In recent years, the number of automobiles has increased dramatically in China, especially private cars, but also those for freight and passenger transport. Compared to rail, road transport uses much more energy but is generally faster and more convenient. Energy intensity for overall transport sector increased from 19 kg oil equivalent per ton-kilometer (kgoe/t-km) to 23 kgoe/t-km between 1990 and 1997.

Railway transport in China has been undergoing a transition from steam to diesel and electric locomotives, which has greatly improved energy efficiency because diesel and electric locomotives are 30 and 25% thermally efficient, respectively, compared to 8% for steam locomotives. This shift decreases the energy needed to move passengers and freight by approximately three-fourths. Nonetheless, China's energy consumption per ton-kilometer of freight is still 60% higher than that in the United States and 30% higher than in Germany.

Improved energy efficiency in road transport will derive from using more new automobiles and improving the transportation management system. These trends are already starting to emerge in China in fright transport, with the energy use per ton-kilometer decreasing 15–20% since 1980. For passenger transport, however, energy intensity (measured as energy per passenger-kilometer) has actually increased during the past two decades as a result of lower load factors for public buses.

3.2.5 Agricultural Sector Potential

As China continues to modernize the agriculture sector, mechanization will result in increased energy use. Nevertheless, significant potential for increasing energy efficiency can be achieved with more efficient water-pumping and agricultural equipment. The Chinese government is generally emphasizing rural energy development as supplying more clean and convenient energy to farmers but not the efficient use of that energy.

3.3 Africa

The major challenge for energy development in Africa is how to increase most Africans' access to modern energy services while taking into consideration the economic costs and environmental impacts of these energy sources. Africa has vast energy resources but extremely limited energy and industrial development and the lowest rates of access to electricity in the world.

3.3.1 Trends in Energy Intensity and Consumption per Capita

As mentioned previously, energy consumption in Africa is among the world's lowest, although energy

TABLE IV

Summary of Energy Efficiency Potentials for China and Southeast Asia

Sector	Technological options/area	Southeast Asia 2020 (%)	China 2010 (%)
Industry	Fertilizer	20	10–20
	Iron and steel		15–25
	Nonferrous metals		3–16
	Cement		10–20
	Pulp and paper		20–40
	Electric motors		15–20
	Textiles		15–28
	Aluminum		5–20
	Refineries		5–10
	Brick-making		17–32
	Foundries		8–14
	Ethylene		10–30
	Calcium carbide		10–22
	Sulfate		14–25
	Caustic soda		10–30
	Glass		22
	Ceramics		30
Residential	Electric appliances	20–60	
	Illumination	20–60	
	Lighting		10–40
	Refrigerators		10–15
	Air-conditioning		15
	Washing machines		15
	Cooking appliances		20–40
	Space heating appliances		10–30
Transport	Total transport	2275 PJ	
	Two- and three-wheelers		10–15
	Cars		5–15
	Trains (diesel)		8–14
	Trains (electric)		10
Agriculture	Pump sets		20–50
	Motors		10–30

intensity is relative high. Despite the low base, primary energy consumption per capita has increased by less than 1% per year during the past two decades (Fig. 11). At the same time, energy intensity has actually increased, so African countries need more energy to produce $1 of GDP (Fig. 12). This is true even considering the purchasing power parity difference between Africa and the rest of the world. Data for the continent, however, obviously hide vast differences between countries (Fig. 13).

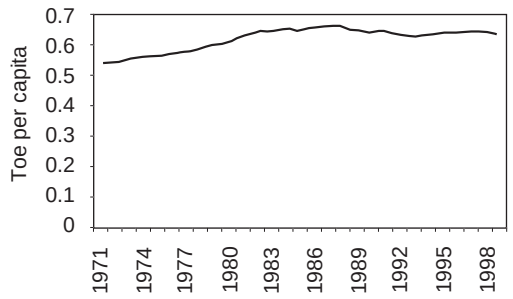


FIGURE 11 Energy consumption per capita in Africa.

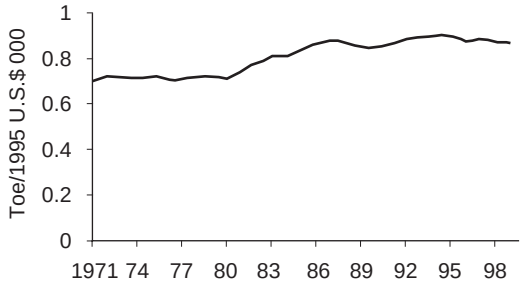


FIGURE 12 Energy intensity trends in Africa (energy use per unit GDP).

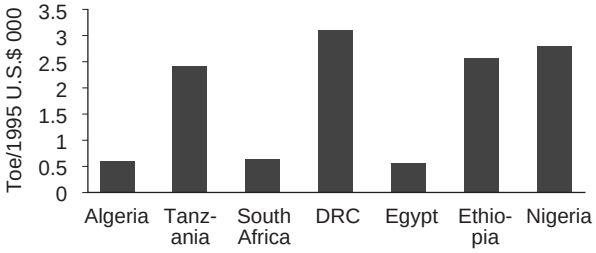


FIGURE 13 Energy intensity in African countries (energy use per unit GDP).

Although a number of African countries have included energy efficiency as part of their national energy policies, few have taken concrete steps toward legislation or specific programs and measures to promote it. Only Tunisia and Ghana have implemented mandatory appliance standards, whereas South Africa is implementing a mandatory commercial building code and has a voluntary residential building energy guideline. Ghana has created an independent body with multistakeholder support called the Energy Foundation, which has been highly successful in implementing energy efficiency programs in a range of sectors. Egypt is also considering a similar path and has set up an energy efficiency council to help develop a national energy efficiency strategy.

3.3.2 Industrial Sector Potential

Many African countries have relatively small industrial sectors, with the exception of large resource extraction industries such as mining and oil production. Even those countries, such as South Africa and some in the north, that have larger industrial sectors tend to rely on heavy, energy-intensive industrial subsectors and processes. Moreover, several southern African countries' dependence on coal, which tends to have a lower conversion efficiency in industrial boilers and power stations, also implies significant opportunities for cost-effective energy savings. An example of this is shown in Fig. 14, which compares the specific energy intensity (i.e., energy use per unit of sectoral output) of Zimbabwe to benchmark international energy intensity levels in industrialized countries.

Key industrial end uses for which industrial energy efficiency can be improved include electric motors, compressed air, lighting, process heating (boilers), and efficiency heating, ventilation, and air-conditioning (HVAC) systems. A wide range of studies suggest that the economic potential for energy efficiency in the industrial sector is 20–30%. In other words, achieving these savings is possible without increasing the cost of production, given current energy prices and technology availability.

Another useful set of examples derives from analysis of major industries in South Africa, in which better compressed air management, lighting, boilers, and changing to variable-speed drives can save 20% of the energy use for those end uses at paybacks of less than 5 years.

An important area that is not covered in the summary table for Africa is the possibility of fuel switching, particularly from coal or oil to natural gas for industrial boilers. As Table V shows, gas boilers are considerably more energy efficient. Although electrical process heat is efficient at the point of use, the large losses in power stations mean that the overall system efficiency is closer to 25%.

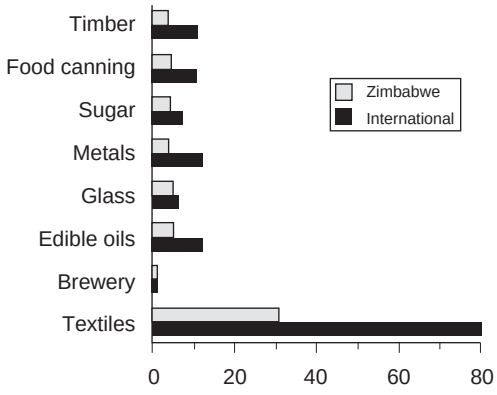


FIGURE 14 Examples of relative energy use efficiencies in industry (GJ/ton).

3.3.3 Residential and Commercial Sector Potential (Buildings and Appliances)

The majority of households in Africa rely primarily on biomass for energy services, particularly cooking. Given that traditional use of wood and charcoal for cooking is only 15–20% efficient, whereas kerosene, LPG, and electric stoves are approximately 40, 55, and 65% efficient, respectively, an important intervention is to move households to commercial fuels for cooking. In addition, improved biomass stoves can increase efficiencies to 25–30%. The problem with assessing the economics of these options is that traditional fuel wood is generally collected and thus outside the commercial energy market. Even purchasing an improved biomass stove costs money, and purchasing fuels for stoves adds even more to household energy costs. Nevertheless, kerosene is preferred over fuel wood in most areas where it is available, and a number of successful programs to increase the use of LPG and electricity for cooking have been implemented, such as in Senegal, Botswana, and South Africa.

Where electric appliances are used in the home or for commercial applications, the potential for savings is significant. Switching to compact fluorescent lighting, for example, can save two-thirds of the lighting energy use, and proper passive solar design of houses along with ceilings and insulation (especially in low-cost housing) can save 60–70% of space-heating energy from a variety of sources. Commercial energy intensity, as with industrial energy use, tends to be high, with many opportunities for improving the efficiency of HVAC systems, lighting, and office

TABLE V

Energy Efficiency of Boilers by Fuel

Fuel technology	Overall efficiency including steam system losses (%)
Electrical process heating	75
Coal-fired boilers	60
Oil-fired boilers	65
Wood-fired boilers	60
Gas-fired boilers	70

equipment. Many housekeeping measures can save energy with almost no capital cost.

3.3.4 Transport Sector Potential

As with industry, South Africa and most north African countries have much higher levels of motorization than other countries in Africa. Road transport is the main mode in Africa. Most vehicles are imported and generally old and poorly maintained, which decreases their efficiency, as does their operation in congested urban areas and on poor-quality road networks. Improving the quality of the road network, maintaining vehicles properly, and shifting road transport to rail could all significantly improve the energy efficiency of transport. Providing affordable, safe public transport systems would also shift passenger travel away from road traffic and increase energy efficiency as well as reduce traffic congestion. The estimates in Table VI are based on a combination of shifting to public transport and improved road and rail transport efficiency.

TABLE VI

Summary of Energy Efficiency Potentials for Africa

Sector	Technological options/area	Economic potential (% of current use)
Industry	Total industry	15–35
	Industry, compressed air	>20
	Industry, variable-speed drives	>20
	Industry, efficient motors	>1–5
	Industry, efficient HVAC	>10
	Industry, efficient boilers	15–20
	Iron and steel	7
	Cement	10–15
	Aluminum (secondary)	45
	Refineries	6
	Inorganic chemicals	19
	Consumer goods	25
	Food	16–30
	Cogeneration	20–25
Residential	Total energy	5–20[a]
	Electric appliances	11–25
	Lighting, compact fluorescent lights	50–70
	Refrigerators	30
	Space heating	60–70
Commercial/ public	Commercial air-conditioning (heat and cool)	50
	Commercial lighting	10
	Computers	10
		20–25
Transport	Cars, road system	30
	Total transport	30
	Road transport	30

[a] Market potential.

3.4 Latin America

During the last decade of the 20th century, the energy sector in Latin America underwent profound changes. In many countries, formerly state-owned monopolies were privatized, electricity and gas sectors restructured, and energy markets liberalized. Foreign direct investment in the energy sector in Latin America has been higher than in any other region in the world. Interconnections between countries are increasing and the market penetration of natural gas is changing the energy mix in many countries. Although energy prices are declining due to competition and alternative sources of commercial energy, many of the notorious problems of energy supply in the region have remained as serious and urgent as they were before the reform process started. Among these problems are (i) disparities of energy supply in more developed urban and less developed rural areas, (ii) insufficient access to clean and convenient energy services by the urban and rural poor, (iii) poor quality of service and protection of consumers, and (iv) only marginal attention to energy efficiency and decentralized renewable energies in official mainstream energy policies. The efficient use of energy in the whole supply chain, from energy extraction and transformation to the end user, could be drastically improved to allow for the expansion of energy services while safeguarding sustainable economic and social development.

3.4.1 Trends in Energy Intensity and Consumption per Capita

Energy intensity in Latin America is substantially higher than in industrialized countries but lower than that of many developing regions (Fig. 4). At the same time, energy consumption per capita is only one-fourth that of industrialized countries (Fig. 1). As shown in Figs. 15 and 16, the energy intensity of Latin American economies has remained quite constant during the past 30 years, whereas the total primary energy supply per capita has increased only 30% during this period, with stagnation during the 1980s and early 1990s. Both energy intensity and per capita consumption differ substantially among different countries of the region (Fig. 17).

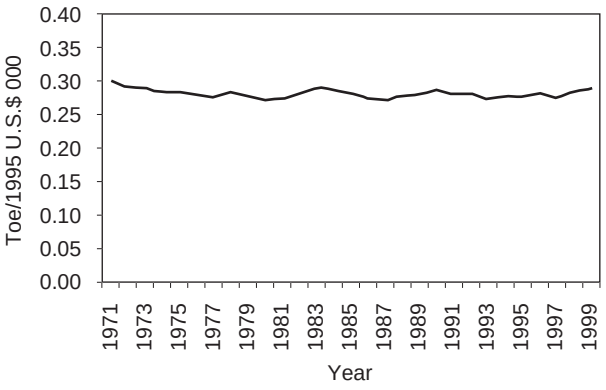


FIGURE 15 Energy intensity in Latin America.

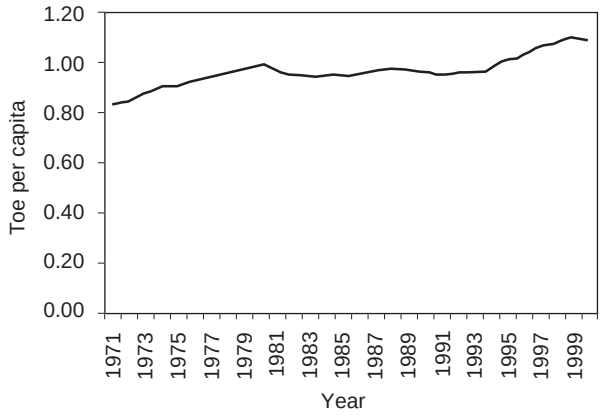


FIGURE 16 Primary energy consumption per capita in Latin America.

The use of noncommercial energy such as fuel wood is still high, particularly in rural areas, whereas low-grade and polluting fuels, such as charcoal and kerosene, are widely used in low-income urban areas. In Peru, for example, 63% of the energy consumption in the residential and commercial sectors in 1995 was fuel wood, 13% kerosene, and 6% animal dung compared to 9% electricity, 6% LPG, and 0.1% natural gas.

During the 1980s and 1990s, energy efficiency programs and legislation were developed and implemented in various countries of the region. Pioneers in this area are Brazil and Mexico, followed in the early 1990s by Costa Rica and later by other countries, such as Argentina, Colombia, Chile, Ecuador, and Peru. These energy efficiency programs were financed partly from own national sources but also with contributions from international donors. Although foreign funds have been crucial, the lack of commitment of national governments has in some cases impeded the wider uptake of the energy

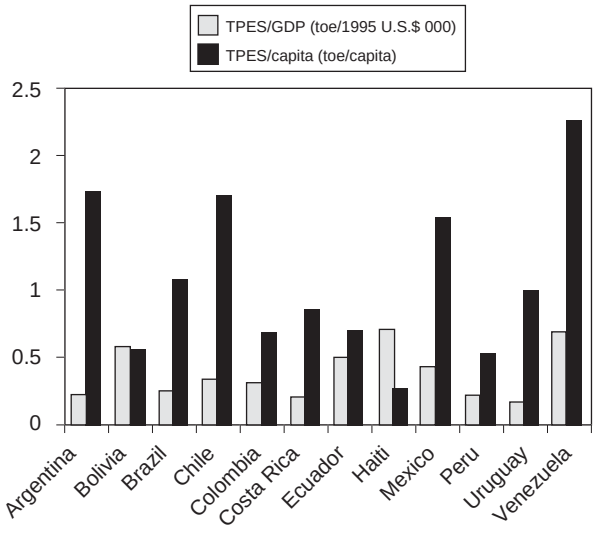


FIGURE 17 Energy intensity and per capita consumption in various Latin American and Caribbean countries, 1999.

efficiency message by the market and society. Brazil, Colombia, Costa Rica, and Peru have introduced new specific legislation for the promotion of energy efficiency and renewable energies, and bills for energy efficiency laws have been presented in other countries. These energy efficiency laws basically define the authority of governments to act in specific areas, such as information dissemination, energy efficiency standards and labeling, mandatory energy audits, the promotion of energy service companies or fiscal incentives. For example, although the elaboration of technical standards and energy efficiency labeling schemes is gaining momentum in various Latin American countries, important and visible schemes have been implemented only in Brazil and Mexico. Although the application of the standards in Mexico is mandatory and includes minimum levels of energy efficiency, the labeling scheme applied in Brazil has been voluntary. However, mandatory minimum efficiency standards are under serious consideration following the energy crisis in 2001 (Table VII).

3.4.2 Industrial Sector Potential

The market opening during the 1990s has forced many industries in Latin America to modernize their production processes in order to compete with foreign companies in domestic and international markets. Foreign direct investment has taken place in many industrial sectors and substantial technology transfer has been achieved. The modernization of industrial processes often involves higher energy efficiency and fuel switching. An example is the

TABLE VII

Summary of Energy Efficiency Potentials in Latin America

Sector	Technological options/area	Economic potential (% of current use)
Industry	Industry, efficient motors and drives	15–30
	Industry, process heat	20–50
	Industry, refrigeration	15–40
	Industry, efficient boilers and steam systems	20–40
	Iron and steel	10–30
	Cement	10–40
	Food and beverages	20–30
	Textiles	20–30
	Cogeneration	20–30
Residential	Total residential	20–40
	Electric appliances (general)	20–40
	Lighting, compact fluorescent lights	30–50
	Refrigerators	35–50
	Space heating	25–50
Commercial/ public	Total commercial/public	15–40
	Air-conditioning	20–30
	Commercial lighting	30–40
	Public lighting	20–45

large-scale substitution of pyrometallurgical processes in the Chilean copper industry by modern, more efficient electrochemical processes. Whereas large industries are successfully adapting to more efficient processes, small- and medium-scale industries in Latin America often face many problems adapting to the more competitive environment.

The energy-intensive sectors cement, iron and steel, chemicals, and food and beverages consume approximately 60% of industrial energy in Latin America. Several studies carried out during the 1990s estimate that the energy conservation potential in these activities is 10–30%. The same range of savings potential also applies for the textile, other nonmetallic minerals, and machine building industries, with the majority of measures having favorable payback periods. Electrical motors and drives, process heat, and refrigeration are important areas of industrial energy end use. The energy savings potential in the field of electrical motors and drives is up to 30%, depending on the kind of measures taken. Losses in industrial thermal processes are as high as 70%, and energy conservation potentials in industrial steam systems are as high as 40%. There is also a significant potential for industrial cogeneration in Latin America, which is largely untapped because of unfavorable legal frameworks.

3.4.3 Residential and Commercial Sector Potential (Buildings and Appliances)

Residential, commercial, and public buildings offer high energy savings potentials throughout Latin America. Building codes, where existing and enforced, do not normally include requirements concerning thermal insulation. This is particularly relevant for housing for the poor, where low-cost materials and designs are usually applied. Forthcoming standards for energy certification of residential, commercial, and administrative buildings in Chile are an encouraging step in the right direction.

Another area with major energy savings potential is household appliances. The widespread use of low-grade fuels, such as kerosene, fuel wood, and charcoal, in rural and poor urban households leads not only to high levels of pollution but also to high energy losses (up to 90%). Studies indicate that the economic potential for energy efficiency improvements in residential end uses such as cooking, lighting, and refrigeration is 20–50%. On a national basis, the Costa Rican government estimates that the combined energy savings potential of efficient refrigeration and lighting systems in the residential and commercial sectors is 12–22% until 2015, taking into consideration different levels of efficiency and market penetration of the equipment. Similar potentials for end-use efficiency (typically 20–40%) exist in commercial and public buildings (e.g., shopping centers and hotels) and in street lighting.

Social problems, such as poverty and unequal distribution of income, are an important barrier to a more rigorous transformation of the residential end-use sector. Attention to energy efficiency is dampened because of more urgent, basic service needs, and new, efficient equipment and housing are often not affordable to low-income households. Several projects in the region funded by the GEF focus on end-use efficiency in the residential, commercial, and public sectors, particularly on efficient residential and street lighting.

3.4.4 Transport Sector Potential

Road transport is the most important transport mode in Latin America, with problems such as traffic congestion, air pollution in cities, and poor fuel economy due in part to the age of the car and truck

fleet. Effective transportation systems combined with supportive urban planning, building on the model originally established in Curitiba, Brazil, in the 1970s, and the increasing use of compressed natural gas and LPG in passenger and goods transport in Argentina, Brazil, and other countries are examples of measures taken to mitigate this situation. Nevertheless, energy use in the transport sector is growing and will further grow with increasing welfare and consumption levels. Measures to reduce the energy intensity of the transport sector would require improving the quality of roads, implementing effective regulation to maintain vehicles properly, as well as incentives to use cleaner and more efficient fuels and cars, introducing effective transport systems in more cities and revitalizing railway transport in Latin America. Energy savings potential due to more efficient driving habits and preventive maintenance have been estimated to be 15–25% of the fuel consumption in passenger cars, trucks, and buses.

SEE ALSO THE FOLLOWING ARTICLES

Cultural Evolution and Energy • Development and Energy, Overview • Energy Efficiency, Taxonomic Overview • Energy Ladder in Developing Nations • Global Energy Use: Status and Trends • Indoor Air Quality in Developing Nations • Industrial Energy Efficiency • National Energy Policy: Brazil • National Energy Policy: China • National Energy Policy: India • Obstacles to Energy Efficiency • Sustainable Development: Basic Concepts and Application to Energy • Women and Energy: Issues in Developing Nations

Further Reading

Bhagavan, M. R., and Karekezi, S. (eds.). (1992). "Energy Management in Africa." Zed Books, London.

China International Engineering Consultant Corporation. (1997). "Comparison between China and Foreign Countries on Unit Energy Consumption of Energy Intensive Industries." State Development Planning Commission, Beijing.

Das Gupta, M., and Roy, R. (2000). Manufacturing energy use in India: A decomposition analysis. *Asian J. Energy Environ.* **1**(2), 223–250.

Das Gupta, M., and Roy, J. (2002). Estimation and analysis of carbon dioxide emissions from energy intensive manufacturing industries in India. *Int. J. Energy Environ. Econ.*

Geller, H., Jannuzzi, G., Schaeffer, R., and Tolmasquim, M. T. (1998). The efficient use of electricity in Brazil: Progress and opportunities. *Energy Policy* **26**(11), 859–872.

International Energy Initiative (2001). *Energy Sustainable Dev.* **4**(4). [Special issue on the work of the Working Group on Energy Strategies and Technologies of the China Council for International Cooperation on Environment and Development]

Jochem, E. (2000). Energy end-use efficiency. *In*: "World Energy Assessment: Energy and the Challenge of Sustainability" (United Nations Development Program, United Nations Department of Economic and Social Affairs, and World Energy Council, Eds.), pp. 173–217. United Nations Development Program, New York.

Karekezi, S., and MacKenzie, G. A. (1993). "Energy Options for Africa: Environmentally Sustainable Alternatives." Zed Books, London.

Liu, Z., Sinton, J. E., Yang, F., Levine, M. D., and Ting, M. (1994). "Industrial sector energy conservation programs in the People's Republic of China during the Seventh Five-Year Plan (1986–1990), Report LBL-36395." Lawrence Berkeley Laboratory, Berkeley, CA.

Ministry of Environment and Energy (XXXX). "Programa Nacional de Conservación de Energía 2001– 2006." National Commission for Energy Conservation, San Jose, Costa Rica.

Ministry of Mines and Energy. (1998). "Plan de Eficiencia Energética para el Mediano y Largo Plazo, 1999–2009." Ministry of Mines and Energy, Lima, Peru.

Mongia, N., Sathaye, J., and Mongia, P. (1994). Energy use and carbon implications in India: Focus on Industry. *Energy Policy* **22**(11), 894–906.

OLADE/CEPAL/GTZ (1997). Energy and Sustainable Development in Latin America and the Caribbean: Approaches to Energy Policy. OLADE/CEPAL/GTZ, Quito, Ecuador.

OLADE/European Commission (1995, January). Proyecto "Conservación de Energía en la Industria: Esquemas Subregionales de Financiamiento."

Roy, J., Sathaye, J., Sanstad, A., Schumacher, K., and Mongia, P. (1999). Productivity trends in Indian energy intensive manufacturing industries. *Energy J.* **20**(3), 33–61.

Sathaye, J. A., and Ravindranath, N. H. (1998). Climate change mitigation in the energy and forestry sectors of developing countries. *Annu. Rev. Energy Environ.* **23,** 387–437.

Spalding-Fecher, R., Clark, A., Davis, M., and Simmonds, G. (2003). The economics of energy efficiency for the poor: A South African Case Study. *Energy Int. J.* **27**(12), 1099–1116.

Tata Energy Research Institute. (1999). "TERI Energy Database Directory and Yearbook 1998–99." Tata Energy Research Institute, New Delhi.

Thorne, S. (1995). "Energy Efficiency Potential in the South African Economy: A Review." Energy and Development Research Centre, University of Cape Town, Cape Town, South Africa.

Trikam, A. (2002). Greenhouse gas mitigation options in the industrial sector. *South African J. Econ. Management Sci.* **5**(2), 473–498.

Winkler, H., Spalding-Fecher, R., Tyani, L., and Matibe, K. (2002). Cost–benefit analysis of energy efficiency in urban low-cost housing. *Dev. Southern Africa* **19**(5).

Prices of Energy, History of

PETER BERCK
University of California, Berkeley
Berkeley, California, United States

MICHAEL J. ROBERTS
U.S. Department of Agriculture[1]
Washington, D.C., United States

1. **The Long Historical Sweep**
2. **The Cost of Finding Oil**
3. **The Organization of Petroleum Exporting Countries and the Oil Embargoes**
4. **Coal Prices**
5. **Natural Gas Prices**
6. **Energy Prices to Users**
7. **A Market for "Energy": Cointegration of Prices**
8. **Prices and the Nonrenewable Resource Price Theory**
9. **Conclusion: Supply Factors and Demand Factors**

Glossary

capital asset pricing model Predicts that an asset's risk-adjusted rate equals the risk-free rate plus an asset's beta multiplied by the expected return on the market portfolio (aggregate wealth). An asset's beta is its covariance with the market portfolio divided by its variance.

cointegration The property that a linear combination of data series is stationary while the individual series is not stationary.

marginal cost The cost of producing one additional unit of a good.

stationarity The property that the expected mean and variance of a series do not change over time.

The earliest energy product for which there are prices is wood. Wood has other, and higher valued, uses than as a source of energy. It is also expensive to transport per British thermal unit provided, hard to burn efficiently, nonuniform in composition, and very high in potentially polluting by-products. For these reasons, it has been supplanted in the energy-intensive developed world. However, it remains an important source of fuel in less developed countries. Coal rapidly supplanted wood as the fuel of the industrial revolution, and as a product of British society (and its American offshoot) there are long series of data available on the price of coal, or coales as it was first called. The percentage share of coal in the energy market has decreased for reasons of transport and environment. Efficient transport for coal is now provided by dedicated trains (unit trains), although water transport, including canals, was the principal method until the late 19th century. The need for expensive fixed transport has made coal hostage to railroad pricing policies and to government regulations. Water power was also an important industrial energy source in the early industrial era, but it was harnessed in a local fashion (mills on streams), and so we are unaware of price series for this direct use. Later, of course, water was used to produce electricity, and there are ample records of electric prices and estimates of hydropower costs. Nuclear power is likewise used to manufacture electricity, although it has been used only since the late 20th century. Price records are for electricity, regardless of its source, although cost estimates are available for nuclear power. The gas industry began with manufactured gas, usually derived from coal, which was used for lighting. In the United States, Baltimore was the first city to have piped gas in 1816. The transmission of natural gas began in 1872, but the 8-in. pipeline from Greenstown, Indiana, to Chicago in 1891 was the first large commercial natural gas venture. Natural gas, lacking the carbon monoxide of manufactured gas, was suitable for heating. Natural gas is now an important source for

[1] The views expressed are those of the authors and do not necessarily reflect the views or policies of the U.S. Department of Agriculture.

 © 2004 Elsevier Inc. All rights reserved.

heating and power generation. Its emission profile is more favorable than that of coal or oil, which helps account for its growing usage. Price series for natural gas are not nearly as long as for coal because large-scale commercial use only began in the 20th century. Oil is the last fuel mineral that we consider. The first oil well was drilled in 1859 near Titusville, Pennsylvania, and oil rapidly became the fuel mineral of choice. Oil burns cleaner and is easier to transport than coal. It is also easier to use because oil furnaces have no need for ash removal or other hand work. Oil price history is well recorded. Because of Standard Oil, the Texas Railroad Commission, and the Organization of Petroleum Exporting Countries, oil prices are part of modern political debate and consciousness.

1. THE LONG HISTORICAL SWEEP

Figure 1 presents Hausman's real price index series for wood and coal for Great Britain from 1450 to 1989. Hausman normalized the index to 100 in 1600 in the belief that by then coal was cheaper than charcoal so that the figure would understate the advantage of coal over wood as a fuel. There is not sufficient evidence concerning heat content to normalize the two series in that fashion. The series are meant to capture the price of the fuel source and are net of transportation or taxes. The figure clearly shows the increasing price of wood and decreasing price of coal in the 1600s. In fact, coal does not begin to rapidly increase in real price until nearly the beginning of the 20th century. In addition to cartels, strikes, and recessions, which account for some of the peaks and dips in Fig. 1, the major change in the U.K. coal industry was nationalization after World War II. The overall increase in 20th-century price is due to the quality of U.K. coal deposits and the depletion of more favorable seams. This is peculiar to the United Kingdom.

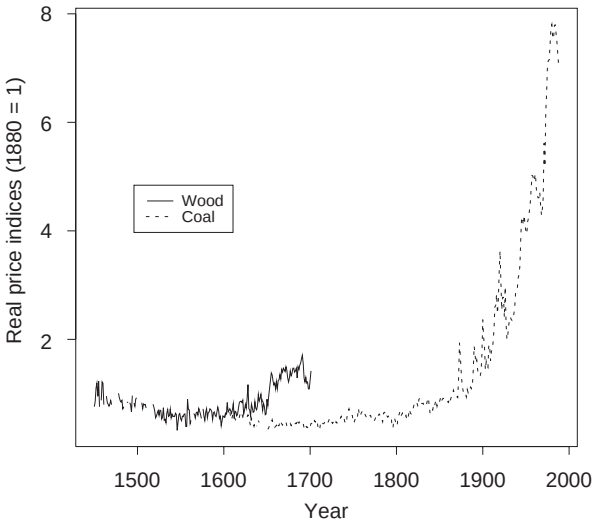


FIGURE 1 Historical prices for Great Britain fuel—wood and coal. Adapted from Hausman (1995).

Figure 2 shows the prices of the dominant energy minerals in the United States from 1870 to the present. The prices are in real 1970 dollars per British thermal unit (Btu). The price of oil decreased significantly from 1870 to the energy shocks of the 1970s and 1980s and then returned to prices not unusual in the early 20th century. A more complete discussion of the energy shocks is provided later. Coal prices are less than those of oil, reflecting oil's greater portability and cleaner burning. Coal prices increased along with oil prices after the 1970 energy price increases. Natural gas was regulated until 1979 and then deregulated during the next decade. Particularly after 1970, the clean burning characteristic of natural gas made it a desirable fuel. This property and the end of regulation account for the change in the relative prices of coal and oil.

Figure 3 shows U.S. energy consumption by source, and it indicates the increasing importance of natural gas, oil, and coal and the decreasing

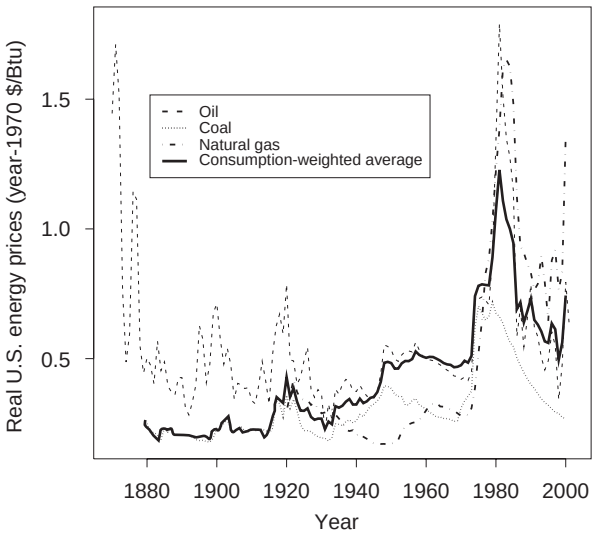


FIGURE 2 Historical prices in the United States, 1870–2001. Annual data on current prices for the raw resource products were obtained from the Energy Information Administration (EIA) and converted to 1970 dollars using the consumer price index. These values were converted to approximate dollars per Btu based on the following data from the EIA: In 1970, electric utilities paid on average approximately 28¢ per million Btu of natural gas, 31¢ per million Btu of coal, and 42¢ per million Btu of petroleum.

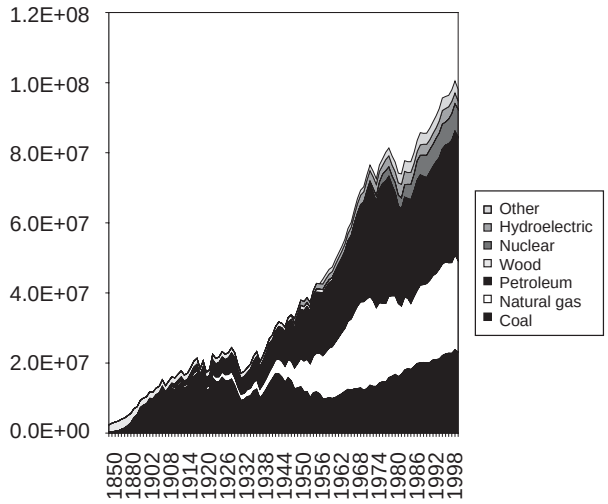


FIGURE 3 U.S. energy consumption by source (trillion Btus). Data from the Energy Information Administration and historical statistics.

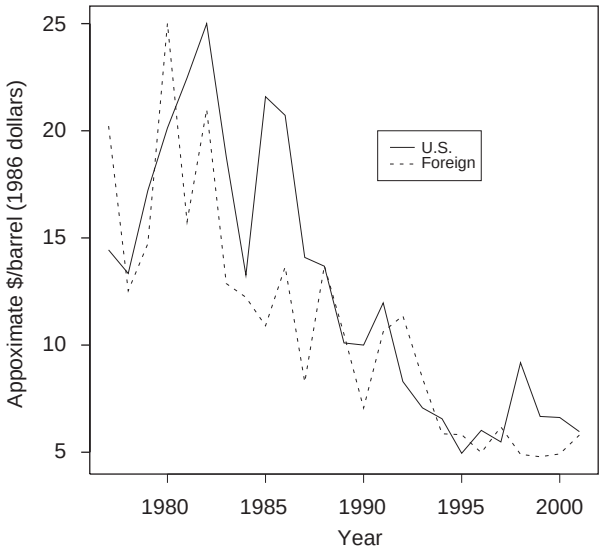


FIGURE 4 Cost to find oil and gas. Estimates for the cost to find oil were derived from data obtained from the Energy Information Administration, Form EIA-28 (Financial Reporting System). The left axis represents total expenditures on exploration and development divided by new additions to reserves. Total reserves include both oil and natural gas. Natural gas reserves were converted to barrels using the average September 1998 prices for natural gas and oil.

importance of wood. Coal was the first energy mineral to produce significant energy output. Coal energy usage increased slowly during the 20th century. At the beginning of the century, coal accounted for approximately 73% of U.S. energy use, with wood accounting for nearly all the rest. Today, coal, gas, and petroleum account for approximately 22.6, 24, and 39.4% of use, respectively, with the remainder coming from renewable sources, primarily nuclear and hydropower.

2. THE COST OF FINDING OIL

The private cost to the economy of using a barrel of oil is the cost of finding a barrel of new reserves of the same quality to replace the used-up barrel. If the oil business were competitive, which it is not, then the market price of oil would be the finding costs plus the development costs plus the marginal costs of production. For this monopolistic business, it is the cost of finding oil that provides the most information about oil's ultimate scarcity. The cost series for finding oil are necessarily very incomplete and lack certainty. The series developed by the Energy Information Administration (EIA) comprise expenditures on exploration and development divided by oil and gas reserve additions. The latter series is subject to considerable uncertainty since for many years it is unknown how much oil (and gas, valued as oil equivalent) was actually found. Additions to reserves can be dominated by changes in the estimates of existing fields, whereas it is the cost of developing new sources that is of most interest. Information was collected for the years 1977–2001 for both foreign and domestic activities of U.S. reporting firms and is plotted in Fig. 4. For 1977, the cost was calculated in 1986 dollars as $14/barrel U.S. and $20/barrel foreign. Domestic costs peaked at $25 in 1982, declined to $4.95 in 1995, increased again, and then settled at $6/barrel in 2001; foreign costs mostly declined during the period and ended at $6. It is believed that the cost of finding and developing oil in the Persian Gulf is less than these figures.

3. THE ORGANIZATION OF PETROLEUM EXPORTING COUNTRIES AND THE OIL EMBARGOES

On October 17, 1973, the Organization of Petroleum Exporting Countries (OPEC), which was originally formed in 1960, announced a progressive embargo on the sale of oil to the United States and other countries that supported Israel in the 1973 Yom Kippur war. They also announced a 70% increase in the reference price for crude oil, an artificial price that determines how much oil companies must pay in taxes. Given that the United States bought only 5% of its oil from

OPEC, an embargo of the type announced was unlikely to inconvenience the United States or substantially raise prices. However, Saudi Arabia decreased its production 35% below the September output and increased the price an additional 6¢. Combined with the nationalization of oil company interests, the cutbacks in production resulted in major increases in the price of oil and a major redistribution of the rents from oil production to the oil-producing companies. In retrospect, it is clear that these actions by Saudi Arabia, with some support from the rest of OPEC, were largely commercial in nature and U.S. attempts at bilateral relationships and other diplomacy had no effect. Production limitation became an effective tool for raising prices far above discovery and production costs.

The next increase in prices occurred in 1978–1981 as a result of a secret OPEC program to impose production controls and due to turmoil in Iran, including the Iran–Iraq war. When Iran's oil output decreased, Saudi Arabia did not increase its output. With Iran not producing at all in January 1979, Saudi Arabia decreased production by 2 million barrels per day on January 20. Its output vacillated for the next several months, always well below capacity, with the result that oil prices topped $20/barrel in October 1979. Prices increased until 1982, when the Saudi export price reached $33.50/barrel. In this period of war-induced uncertainty, oil companies and oil users hoarded significant amounts of oil.

Adelman attributes the price decline in 1982 to the end of a speculative bubble. The continued slow decline until 1986 was the result of increased supply from non-OPEC members and the consequence of lagged demand elasticity. In 1986, prices declined considerably, presumably a function of the substantial non-OPEC output. The lowest price was reported by Adelman as $6.08 for Saudi Light; in August of 1986, OPEC again set production quotas, with cooperation from some non-OPEC members.

The next major price event was the Gulf War in 1991, again producing a spike in prices. The war directly ended the sales by Iraq and Kuwait and spurred the building of precautionary supplies. Because there was sufficient capacity without the output of these two countries, prices rapidly subsided.

4. COAL PRICES

In the time of Queen Elizabeth I, the coal trade in England (i.e., coal sold into the London market) was monopolized. Monopolization was possible because the London market was served by two rivers, the Tyne and the Wear, and the coal lands adjacent to these rivers. The land involved was not so extensive as to defy ownership by an oligopoly. The key issue in the early coal trade was water transport, with coal lands situated near canals or rivers with major markets being valuable and other lands being of lesser value.

Although oligopoly was in existence as early as 1517, it was formalized with royal charter as a monopoly of coal and also of political power in Newcastle by Queen Elizabeth on March 22, 1600. Other important dates cited by Sweeney include an act of Parliament to make restriction of trade in coal illegal in 1710, an allocation document fixing the quantities of coal among major owners in 1733, a lack of regulation in the period 1750–1771, a resumption of regulation in 1771 or 1772, and on-and-off regulation through 1850. Although the detailed history is not provided here, the British coal series should be interpreted in the light of the on-and-off nature of the quantity regulation. Price changes may reflect temporary changes in the Lerner index as much as changes in capacity or demand.

Coal prices in the United States during the 20th century were subject to a very different sort of government regulation. Because of the need to reduce sulfur emissions in the air, the U.S. government required new and substantially modified power plants to reduce sulfur emissions in such a way that there was no advantage to the use of low-sulfur rather than high-sulfur coal. In effect, since 1990 the emissions permit trading program no longer disadvantages low-sulfur coal as a way to reduce emissions.

5. NATURAL GAS PRICES

In the United States, natural gas was regulated and sold at a low price until gradual deregulation from 1979 to 1989. Thus, the prices after 1989 were free to increase in response to increased demand, driven by the desire to switch from expensive oil to gas and by environmental restrictions that allow the burning of gas but restrict oil and coal. For the data plotted in Fig. 2 prior to 1979, the correlation between gas price changes and oil price changes is 0.067; after deregulation, the correlation is 0.605.

6. ENERGY PRICES TO USERS

Although there is a world market in many of the energy minerals, national governments' tax and

allocation policies create very different prices for end users in different countries. In general, the United States has among the lowest prices of consuming nations to end users of energy products, and all countries generally price consumer products higher than industrial products.

In comparing prices across countries, one must keep in mind the problems of converting currencies and the qualities of the fuel. The following discussion is in terms of conversion at the official exchange rate as opposed to purchasing power parity. The exchange rate is the best choice for industrial uses, especially where the end products (e.g., cars, steel, and airplanes) are widely traded. Particularly in countries less well attached to the international trading system, exchange rates can be very different than purchasing power parity, and the use of dollars as the official exchange rate would provide a poor estimate of the burden on consumers of energy purchases relative to a basket of other goods. In order to make industrial and household prices comparable, we use exchange rate conversion throughout. There is also an issue of heat equivalent [or heat equivalent net of the energy needed to remove water from the fuel, gross caloric value (GCV) and net caloric value (NCV), respectively.] When comparing prices across fuels (oil versus coal) or across different types of the same fuel (U.S. bituminous coal versus Czech brown coal), the most accurate comparisons are in terms of ton of oil equivalent or NCV. In order to use tax information, we use the reported metric tons (tonnes).

Perhaps the most traded energy commodity is fuel oil. The following section reviews the evidence that fuel oil is one international market; that is, the prices in different countries increase and decreases simultaneously, with differences only in transportation costs. For 2001, the range of prices for high-sulfur fuel oil for industry, in U.S. dollars per tonne, was 103 to 258. The lowest price in the International Energy Administration's (IEA) data set was for Mexico, which is a net exporting nation. (The data do not include the Gulf States.) The U.S. price was $147/tonne, that of France was $154.4/tonne, and that of Turkey was $181.7/tonne, whereas the Organization for Economic Cooperation and Development (OECD) average was $189/tonne. France levied taxes of $16.6/tonne, whereas the United States levied no taxes; therefore, the pretax price in France was cheaper than in the United States. Turkey's tax was $42 and its pretax price was cheaper than that of the United States, which should not be surprising given that Turkey is a transshipment point of Mideast oil.

Automobile fuel is refined from oil. Apart from issues of taxation, the prices of automobile fuel should be very similar throughout the world, just as the ex-tax prices of fuel oil are very similar. Consumers pay the price with tax, which varies greatly from country to country. Of the countries that report a price for regular unleaded gasoline, the United States has the lowest at $0.39/liter. Germany has a price of $0.90 and Denmark is highest at $0.97. France reports a premium leaded price of $1.00 and Turkey $0.99, and Mexico reports a premium unleaded price of $0.65 per liter. This variance in pricing is attributable to taxation. Although it is not reported in the IEA data, there can also be a vast difference in the attributes of gasoline. In order to meet its obligations under the Clean Air Act, the state of California requires the use of a much more refined gasoline than is used in the rest of the United States and this gasoline carries a hefty premium in price.

Coal is a basic energy source both for industry and for electric generation. Coal is more difficult to transport than oil (railroads and slurries versus pipelines) and major deposits are often not near ocean transport, unlike oil. Coal is also more likely to be used without much preprocessing (for pollution control purposes, it can be washed, whereas crude oil is always refined), so coal must be matched much more closely to boilers than crude oil needs to be matched to its ultimate customers. For these reasons, there should be a greater variance in the underlying price of coal than there is in the underlying price of oil. Again examining IEA data, the prices for steam coal (for industry or electric generation as opposed to coking or metallurgical coal), the U.S. price was $36/tonne and that of Japan was $32/tonne, with many other countries in the $30 range. Germany was highest with a price of $52/tonne. Coal can also be measured in tons of oil equivalent to compensate for different caloric outputs per ton, and the pattern is much the same.

Electricity is generated from a variety of processes, including burning of coal, oil, and natural gas, nuclear reactors, hydropower, wind, photovoltaic, and even tide power. A country's electricity prices to industry and to consumers represent a mix of the costs of these methods. Countries such as Norway, or areas of countries such as the U.S. Pacific Northwest, that have significant surpluses of hydropower can have very low costs of production, as can countries that are willing to burn coal, particularly with minimal pollution control. Variation in electric rates across countries is thus explained by natural abundance of cheap methods, preferences for clean air, and taxation. In

2000, the U.S. price for households was $0.082/kWh (excluding tax) and it was $0.042/kWh for industry. Denmark had the highest household price at $0.197/kWh, and the Slovak Republic had the cheapest at $0.05/kWh. The Slovak price is probably related to the burning of brown coal, with its concomitant problems of air pollution. Norway's household price was $0.057, reflective of both hydropower and petroleum reserves. Industrial prices in Europe are much lower than household prices, largely due to taxes. France's industrial price was $0.036/kWh compared to its household price of $0.102/kWh. Italy, Japan, and Turkey are outliers in the IEA data set because they had industrial electric prices of $0.089, $0.143 (1999), and $0.080/kWh, respectively.

Natural gas for households on a toe-NCV basis also show great variation among countries, with the U.S. price at $417, slightly above the OECD average of $391 in 2001. This is quite different from 1994, when the U.S. price was $274 compared to an OECD average of $383. Presumably, this change in relative price is related to shortages of both pipeline space and gas at the wellhead. Despite the clear preference in the United States to generate electricity with natural gas for air quality reasons, we expect that the price will decrease. With a price of $788, Denmark has the highest price in the dataset.

Taxation represents a large percentage of the delivered price of energy-containing products, particularly consumer products. Automotive fuel taxes represent half or more of the value of the product in all but 4 of the 29 countries included in the IEA dataset. In the United Kingdom, they are 76.9% of the value.

The earlier years in the IEA data set tell much the same story as the later years. Pricing to consumers tends to have more taxation and more idiosyncratic variation than pricing to industries.

7. A MARKET FOR "ENERGY": COINTEGRATION OF PRICES

Energy is not a homogeneous resource and the prices of the various energy minerals do not necessarily track each other well over time. Oil is the most transported of the energy minerals, much of it originating in the Persian Gulf and transported by tanker. The differences in the prices of oil at different ports can be explained largely by transport costs. One way to test the hypothesis that oil is traded in a single market is to determine whether the series are cointegrated. Two price series are cointegrated if there exists a linear combination of the two series that is stationary (the mean and variance do not change over time). In other words, the two series move together.

An example of this type of test is to regress the price of Saudi Light landed in southern Europe on the price of the same oil landed in Rotterdam and then test the residuals of the regression to determine if they are stationary. Given the oil series are both stationary in differences, a finding that the residuals of this regression are also stationary means that the southern Europe price series is a linear function of the Rotterdam series, and any remaining differences between the two series can be accounted for by transport costs. It means that oil price shocks affect both series at the same time. Cointegration tests for oil at different ports find cointegration and support the notion that oil, in Morris Adelman's words, is a single large pool.

Coal contributes to world energy somewhat less than oil and so if there were one price for "energy," then oil and coal prices would also be expected to be cointegrated. Similarly, coal prices in different localities would be cointegrated. There is little evidence that coal prices move together or that they move with oil prices. Figures 1 and 2 show the prices of coal for the United Kingdom and the United States. In these long series, the price in the United Kingdom increases approximately eightfold, with the strong trend being upward. The price in the United States shows a peak in the early 1980s and then a sharp decline to its previous value. It is dubious whether the U.S. price is differenced stationary rather than just stationary. The U.K. price is clearly stationary only in differences. Thus, these two prices are not cointegrated. Similarly, there is little evidence that the German and U.K. prices are cointegrated.

Although the US coal and oil prices are not found to be cointegrated using the normal testing procedure, Fig. 2 indicates that the coal price does bear some relationship to the oil price, very notably in the price run-up in the late 1970s. One explanation for the difference in the U.S. coal and oil prices is that coal is sold on long-term domestic contracts and is not subject to supply disruptions. As a result, there is no incentive to hoard coal, whereas hoarding oil before the Gulf War was a major cause of the price increase.

There is a world price of oil, but the oil price is not closely related to the prices of coal in various countries, nor are the prices of coal closely related to each other. Energy is not one great pool and there

is no "price of energy," just a multiplicity of prices of energy minerals.

8. PRICES AND THE NONRENEWABLE RESOURCE PRICE THEORY

There is no convincing evidence of a trend in oil, gas, or coal prices. One theory on these prices that has empirical support is that the first differences are stationary with mean drifts of zero. In this case, there would be no deterministic trend in the prices, simply random variation. The best estimate of future prices would be today's price. Another possibility is that the prices revert to a quadratic trend that is stochastic. This view would lead to a prediction of a slight increase in prices. Based on the historical evidence in the time series for energy prices, there is no good evidence for forecasts of increasing price.

Examining the oil price series in Fig. 2, it appears that oil prices were distinctly declining in the early years of exploitation, then leveled off, and since 1980 have increased. Since 1980, prices appear to have followed a quadratic trend. Since 2002, prices seem to have reverted to pre-Iraq war levels. We conclude that analysis of the price path of oil does not lead to a prediction of increased prices.

The lack of a clear trend—deterministic or stochastic—in energy prices at first seems at odds with the theory of natural resources. Theory predicts that prices for these commodities will ultimately rise at the rate of interest, a classic finding by Hotelling. Hotelling reasoned that if one extracted a barrel of oil today and put the money in the bank for 1 year, then one should have exactly the same financial outcome as if one extracted the barrel in 1 year. If this equality were not true, then one would be better off extracting all of one's oil in just one of the time periods, whereas we observe extraction in all time periods. A barrel extracted today would earn its price, $R(0)$, and after 1 year at interest rate i it would be worth $R(0)(1+i)$. This must be equal to the value extracted in 1 year's time, $R(1)$ and $R(1)=R(0)(1+i)$. This theory predicts that mineral prices will rise at the rate of interest. However, the price that is referred to is the seldom-observed price of minerals still in the ground, which is called the mineral rent. The price of minerals that we observe, P, is the mineral rent R plus the marginal extraction costs c. The theory is not so clear about minerals that have already been mined or lifted, which is the series that we have presented.

Since energy mineral prices do not increase at 7–11%, which is the normal rate of return for financial assets, economists have posited that either technical progress or a decline in grade account for the slow or nonexistent rates of appreciation of energy mineral prices.

One possibility is that there is technical progress in the mining technology; that is, that c decreases over time. Then the marginal cost of mining would decline and the value of the underlying mineral would increase. The observed price is the sum of the two prices and so could increase or decrease. However, since the extraction costs are decreasing and the underlying price is increasing, eventually the extraction costs would be insignificant and the price increases in the rent would be dominant. This process requires the rent to become dominant; however, if it is true that undiscovered, but easily discoverable, reserves are very large relative to demand, then rents will be very small and prices will not exhibit an upward trend for a very long time.

Flat prices can also be reconciled with the Hotelling theory by considering the fact that depleting a resource affects extraction. The first units extracted are presumably the least costly, with each unit becoming progressively more costly as cumulative extraction progresses. In this case, part of the opportunity cost of extracting a unit today is increased cost in the future. Rents increase more slowly than the interest rate, but the difference is exactly offset by an increase in future costs. Thus, in the long term, prices still increase at the rate of interest, even if rents do not. The evidence for grade depletion as a worldwide phenomena is thin—there are substantial high-grade coal seams in the United States, and extraction costs (and finding costs) for oil in the Gulf remain very low.

Another way to reconcile the flat price path for energy minerals with the theoretical prediction of price increase is to calculate the required interest rate using the same logic used to determine the needed rate of return for common stocks.

In most of the theoretical literature, the interest rate is taken as a fixed market price of time preference: the marginal value of a dollar received tomorrow relative to its value today, with all else remaining the same. In a world of uncertainty, however, the marginal value of returns depends not only on the timing of returns but also on uncertain contingencies. Losing money on a stock market investment is likely to be more painful if one also loses one's job but less so if growth in one's home equity more than compensates for stock market losses. The fair rate of return of one investment

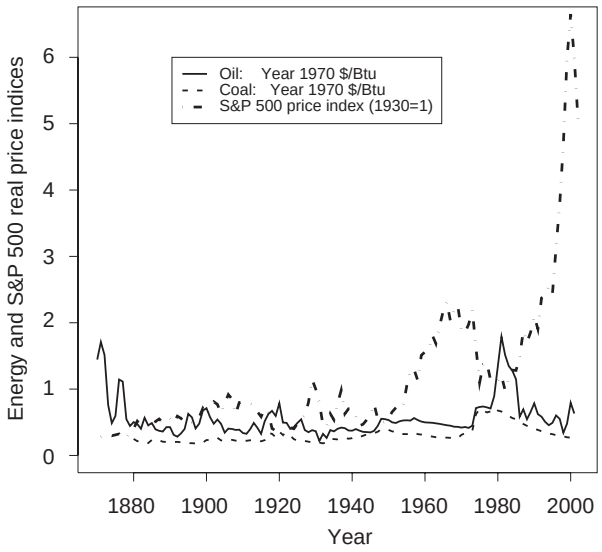


FIGURE 5 Long-term stock market returns and energy prices. The coal price and oil price data are those plotted in Fig. 2. The long historical series of the Standard & Poor's 500 was obtained from Robert Shiller (http://aida.econ.yale.edu/~shiller/data.htm).

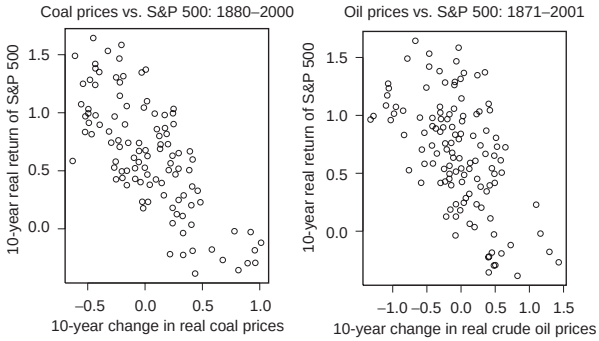


FIGURE 6 Long-term stock market returns and energy price changes. Returns on the Standard & Poor's 500 assume annually reinvested dividends. The long historical series of the Standard & Poor's 500 was obtained from Robert Shiller (http://aida.econ.yale.edu/~shiller/data.htm).

therefore depends on how its returns are correlated with those of other investments.

Interest rates reported for treasury bills and other assets are typically low-risk investments that yield the same amount regardless of other contingencies. In contrast, a decision not to extract and sell a unit of oil today is an investment with a return that depends on future oil prices, which are volatile and uncertain. Thus, interest rates on treasury bills or other low-risk investments may serve as a poor proxy for the rate of return that should be applied to nonrenewable resources.

The capital asset pricing model (CAPM) explains the difference in average returns across investments according to their differential risks. The model assumes that individuals, with all else being the same, prefer greater expected returns and lower risk. To mitigate risk, individuals invest in large portfolios of assets in order to average away the idiosyncratic risk of each asset against the idiosyncratic risk of all other assets. In equilibrium, individuals thus do not care about idiosyncratic risk; the only risk that matters is that pertaining to aggregate wealth—that associated with the entire portfolio of assets.

Theory predicts that assets having returns with a strong positive correlation with wealth should have high average returns. The strong positive correlation of these assets implies that they pay off most when wealth is high and marginal utility of wealth is low. Conversely, assets with returns having a negative covariance with wealth will earn low or even negative average returns. These assets provide insurance against aggregate losses, yielding higher outcomes when wealth and consumption are low and marginal utility is high.

Energy prices have a strong negative correlation with the aggregate economy and aggregate stock market returns and therefore should have a risk-adjusted average rate of return that is lower than that of low-risk investments such as treasury bills. The relationship between energy prices and the stock market, illustrated in Figs. 5 and 6, has received less attention but is more closely related to CAPM. The correlation between the 10-year return of the Standard & Poor's 500 and the 10-year change in coal prices is less than −0.7. Simple CAPM-based estimates of the risk-adjusted rates of return for oil, coal, and natural gas, regardless of the time horizon, are between −1 and 2% and are not statistically different from 0%. As long as price fluctuations for these fossil fuels are strongly negatively correlated with national wealth, theory does not predict that the prices will trend up.

9. CONCLUSION: SUPPLY FACTORS AND DEMAND FACTORS

The past century of energy prices is a history of shifts in supply factors, shifts in demand factors, and policies, such as taxes and trade restrictions. However, the most notable price spikes, including those that occurred following coal strikes in the early part of the 20th century and the oil embargo, OPEC activities, and Iraq wars, are all attributable to shocks that reduced supply or anticipated supply. The observation that supply factors were most

relevant to historical price fluctuations mirrors the negative relationship that has been documented between oil prices and the macroeconomy. Because energy resources are essential inputs, energy price shocks caused by unanticipated shifts in supply cause reductions in aggregate output. If aggregate demand shocks were driving price increases then the relationship between energy prices and aggregate output and stock market returns would be positive, not negative.

The history of energy prices therefore stands in contrast to the history of energy consumption, which has grown steadily with aggregate output and demand. Over the long term, prices have trended flat and have been highly volatile, with the drivers of price shocks stemming from shifts in supply. Supply volatility has caused prices to be negatively correlated with aggregate output and wealth, which in turn explains why prices have not trended up.

SEE ALSO THE FOLLOWING ARTICLES

Business Cycles and Energy Prices • Depletion and Valuation of Energy Resources • Economics of Energy Demand • Economics of Energy Supply • Energy Futures and Options • Inflation and Energy Prices • Innovation and Energy Prices • Oil Price Volatility • OPEC Market Behavior, 1973–2003 • Stock Markets and Energy Prices

Further Reading

Adelman, M. (1993). "The Economics of Petroleum Supply." MIT Press, Cambridge, MA.

Adelman, M. (1995). "The Genie out of the Bottle." MIT Press, Cambridge, MA.

Berck, P., and Roberts, M. (1996). Natural resource prices: Will they ever turn up? *J. Environ. Econ. Management* **31,** 65–78.

Hausman, W. J. (1995). Long term trends in energy prices. *In* "The State of Humanity" (J. L. Simon, Ed.), pp. 280–286. Blackwell/Cato Institute, Oxford, UK.

International Energy Agency (1984–2003). "Energy Prices and Taxes." Organization for Economic Cooperation and Development, Paris.

Livernois, J., and Martin, P. (2001). Price, scarcity rent, and a modified r-percent rule for nonrenewable resources. *Can. J. Econ.* **34,** 827–845.

Sweezy, P. (1938). "Monopoly and Competition in the English Coal Trade, 1550–1850." Harvard Univ. Press, Cambridge, MA.

Public Reaction to Electricity Transmission Lines

LUTHER P. GERLACH
University of Minnesota
Minneapolis, Minnesota, United States

1. **Opposition to Electricity Transmission Lines Is Significant, Organized, and Informed**
2. **What Is the Significance of Opposition to Electricity Transmission Lines?**
3. **How Do Transmission Line Opponents Organize?**
4. **What Reasons Do People Give for Their Opposition to Transmission Lines?**
5. **Need: The First Defense against Transmission Line Siting**
6. **"Size" of the Transmission System as a Reason**
7. **Structure and Restructuring of the Electricity Utility Industry as Reasons**
8. **Sentiments about Visual Impacts and Loss of Control as Reasons**
9. **Health, Safety, and Electric and Magnetic Fields as Reasons**
10. **Summary**

Glossary

alternating current (AC) Flow of electricity that constantly changes direction between positive and negative sides. Most transmission lines are AC.

direct current (DC) Electric current that flows constantly in one direction. A significant case of opposition to a transmission line in the l970s was to a DC system from North Dakota to Minnesota.

electric and magnetic fields (EMF) Invisible lines of force that surround any electrical device, such as a transmission line, electrical wiring, or an appliance. There is controversy about the health effects of exposure to these fields, especially magnetic fields.

electricity utility A firm established to provide electricity services. The traditional structure is one of vertical integration of generating, transmitting, and distributing services. The new structure in many countries separates these functions.

opposition group People organized to contest facility siting and building. They are usually organized in local groups interacting with others in communication networks. Groups usually include persistent activists as cadre and episodic participants in support.

siting authority The government entities authorized to make decisions about siting.

siting process The process of permitting or approving transmission or other facilities, including determining public need for them and their location or route.

transmission lines Heavy wires that carry large amounts of electricity over long distances from a generating station to places where electricity is used. The lines are held high above the ground on tall structures called towers in the United States and pylons in the United Kingdom.

Opposition is the most significant public response to transmission lines in Western industrial democracies. The study of this opposition reveals how people contest and resist transmission line siting by organizing in groups and networks, and by developing and communicating arguments about why the line is not needed and how it negatively impacts people and environment.

1. OPPOSITION TO ELECTRICITY TRANSMISSION LINES IS SIGNIFICANT, ORGANIZED, AND INFORMED

In the United States, the United Kingdom, and Canada and probably in other Western, industrially advanced democracies, the most observable and significant response of people directly affected by the siting and building of electricity transmission lines in their communities is their opposition to these projects. This article examines this public response of

 © 2004 Elsevier Inc. All rights reserved.

opposition. It considers how people organize to oppose and examines the main reasons that people and groups give for their opposition. Major reasons given for opposition are as follows: (1) need and alternatives, (2) size/length and voltage, (3) structure of industry, (4) sentiments, visual impacts and loss of control, (5) health and safety risks, particularly from electric and magnetic fields, (6) fairness, (7) environmental justice, (8) environmental quality, and (9) siting process. This article examines the first five reasons and mentions the others. These reasons constitute a system of interdependent ideas. As people oppose the siting of transmission facilities, they produce, learn, and communicate these reasons. People build local groups to contest a particular transmission project and build global networks to spread and exchange information. The article describes this organization, illustrating communication with examples of the flow of information about electric and magnetic fields.

2. *WHAT IS THE SIGNIFICANCE OF OPPOSITION TO ELECTRICITY TRANSMISSION LINES?*

People can respond in several ways to the information that an electricity utility proposes to site (locate and build) transmission facilities (line and/or substations) in or near their communities. They can actively oppose the line, passively accept it, or actively support it. This article is mostly about active opposition, also termed resistance or protest. It is about opposition in the 1970 s in United States and from the 1990 s through the time of this writing, 2003, in the United States, Canada, and the United Kingdom. Examples from case studies in these countries illustrate opposition. This section on significance introduces many of the examples used throughout this article and such examples are indicated by an asterisk.

Findings for this article were obtained by the author, using field research, conducting a document and media study, and, for cases since the 1990s, corresponding by e-mail and gathering information from the Internet. Sources include representatives of the main parties to transmission controversies: electricity utilities and trade associations and their consultants, government siting authorities, and especially groups opposing transmission siting. Efforts to obtain information from electricity utilities, trade associations, and citizen groups on the European continent, Japan, and India through e-mail inquiries were unsuccessful. Thus, their stories are not represented in this article. This article presents the ideas of people, especially of line opponents, but does not evaluate their accuracy. The article is about what people say and do, not what they should say and do.

This article is about opposition because this is the most observable significant response. Utility executives, consultants, government officials, and energy researchers state that opposition to the siting and constructing of transmission facilities has been one of three main obstacles to the construction of enough such facilities to meet their need and avert what some refer to as congestion of transmission. Many of these sources state that a second obstacle is the way local and state rules and processes to regulate and approve transmission facilities delay or make approval uncertain. Many also state that a third obstacle is the uncertainty about who will own and operate the transmission system and who will pay for grid expansion and improvement. This third obstacle has become important since the 1990s, as the electricity industry has been restructuring so that it is partly regulated and partly competitive. Opposition, approval process, and economic and structural uncertainty affect one another. Approval processes provide opportunities for people to organize and oppose. Opposition, including litigation, can delay approval and force modifications in transmission design and routes and the prospect of this increases uncertainties, deters investment, and provokes project cancellations.

2.1 What Utilities, Government, Consultants and Researchers Say

In 2000, 23 persons across the United States affiliated with electric utilities and related energy-producing and marketing firms, energy-related government agencies, industrial consumers, and environmental groups were interviewed about the need for and obstacles to the expansion of electricity transmission capacity by Eric Hirst, a consultant on electric industry restructuring, for his July 2000 report to the Edison Electric Institute (EEI), the trade association of the investor-owned electricity utilities. Hirst's interviewees mention the following as cases of transmission facilities not being built because they were opposed and/or did not meet requirements for approval.

- A 765 kV line proposed by American Electric Power (AEP) in Virginia and West Virginia (*AEP 765);

- Two lines [230 kV (Chisago) and 345 kV (Arrowhead-Weston)] between Minnesota and Wisconsin (*A-W 345);
- A 500 kV line between Georgia and Michigan;
- Expanded interfaces between Indiana and Michigan;
- Expanded interfaces between Pennsylvania–New Jersey–Maryland and New York and between New York and New England [notably, three projects led by Connecticut Light and Power (CL&P): *CL&P 345 & cable];
- Additional transmission between Wyoming and eastern Colorado;
- Additional transmission facilities needed to serve the growing loads of Boston, New York City, Long Island, San Francisco, and San Diego.

EEI official David Owens, in the July/August 2001 EEI Electric Perspectives, attributes the problems of building new transmission facilities to the many challenges in regulatory review and approval. However, he indicates that these challenges are importantly driven by the public sentiment against transmission expansion and the involvement of government agencies, courts, federal and tribal governments, and competing interest groups. The public sentiment against transmission expansion, he says, is represented by the phrases NIMBY (Not in My Backyard), NOPE (Not on Planet Earth), and BANANA (Build Absolutely Nothing Anywhere Near Anyone). In other words, permitting processes enable people to express what has become a generic opposition to siting new facilities. Owens also provides examples of transmission projects that have been delayed for many years because of opposition and process, mentioning some of the cases listed by Hirst: the project by Xcel Energy (Chisago-Apple River) to link transmission facilities in Minnesota and Wisconsin, proposed in 1996, and the project by AEP to link West Virginia and southwestern Virginia through a 115-mile, 765 kV system proposed in 1990. He also notes the rejection by the Connecticut Siting Council of a cable to run below Long Island Sound, linking Long Island to Connecticut. By summer 2003, the Chisago project had been cancelled, the AEP outcome was uncertain, and the Long Island cable projects were on hold. The AEP and Long Island projects are mentioned in news reports about the August 2003 Northeast Blackout as examples of how parochial interests block construction of interstate transmission. The AEP and Long Island cases are also mentioned in a May 2002 National Transmission Grid Study for the U.S. Department of Energy (DOE) as examples of how needed interstate transmission projects are significantly delayed through siting processes. The study indicates that the Federal Energy Regulatory Agency (FERC) may need to expedite siting.

In October 2001, Alan Schiber, Chairman of the Ohio Public Utilities Commission, testified before the U.S. House Committee on Energy and Commerce on behalf of the National Association of Regulatory Commissioners about transmission siting, especially about efforts to increase the federal role in this. He disagreed with "the presumption in Washington that the reason transmission systems are not being expanded or managed sufficiently to meet transmission need and prevent congestion is that State eminent domain and siting authority is inadequate." Instead, "a major impediment to siting energy infrastructure in general and electric transmission in particular is the great difficulty in getting public acceptance for these facilities." No matter where siting responsibility falls, with state or with federal government, siting energy infrastructure is not easy or amenable to a "quick fix." Schiber notes that people object to siting such infrastructure because they do not want to bear the impacts or the costs of the facilities (even though they do want the electricity).

Basin Electric Power Cooperative, which serves member systems in North Dakota, where it is based, and eight other western states, tells its customers in newsletters of its success in building needed transmission facilities, notably in 2003 a direct current system intertie near Rapid City, South Dakota. In interviews, its officers explain that public opposition is not an obstacle to transmission siting and that they successfully negotiate agreements with landowners (although Basin Electric newsletters do refer to the problems of the "not in my backyard" resistance to siting new transmission). Instead, they say, "A primary obstacle to the construction of transmission is the regulatory uncertainty associated with the deregulation effort and improper pricing methodologies for the high-voltage transmission system." In their October 2000 newsletter, Basin Electric's Vice President of Government Relations summarizes the consequences of the uncertainties by rhetorically asking: "Why would you build it if you don't know that you can use it and everybody has the same rights to it?" Basin's marketing manager asks, "Who should pay for it? Will there be a return on investment?" To obtain a return, Basin Electric calls for pricing in which all users pay their *pro rata* share of the average cost of the grid, a method known as postage stamp pricing. This differs from the currently dominant license plate pricing, which recovers costs for transmission facilities from

local ratepayers. Basin indicates that postage stamp pricing is opposed by parochial interests, presumably by states or localities with traditional low-cost electricity that fear that expansion of the grid will raise their costs by exporting their supply.

2.2 Opposition in the 1970s

Opposition to electricity transmission and recognition of its significance is not new. In the 1970s, electricity utilities began building high-voltage transmission systems to carry electricity long distances and people strongly resisted some of these projects. A notable example is that of the 1974–1979 resistance to the siting and building of a $\pm$400 kV direct current (DC) line across west-central Minnesota by the rural electric cooperatives United Power Association and Cooperative Power Association, collaborating as Cooperative United, or CU (*CU $\pm$400 kV DC). Farmers, rural townsfolk, and urban allies fought the line in the meetings and hearings of the siting process and by litigating in the courts, demonstrating in the fields, confronting surveyors, officials, local sheriffs, and state police, sabotaging towers, facing arrest, dealing with federal agents, and more. Reporters for state and national media covered the events. Researchers studied and reported the case in journals, books, and films, with some researchers calling it an example of a grassroots social movement. The line was built, but the struggle against it forced changes in legislation about siting such facilities.

Another case of note was resistance to the siting of a 155-mile, 765 kV alternating current (AC) line by the Power Authority of the State of New York (PASNY), connecting its system with that of Hydro-Quebec, in Canada. This project (*PASNY 765), proposed in 1973 and argued through hearings for several years, was also protested militantly through some of the same actions as occurred in Minnesota. Minnesota and New York power line resisters learned from one another through telephone, mail, exchange of newsletters, and visits.

Opposition made it difficult to build electricity transmission systems from the l970s on, in Minnesota, New York, and elsewhere. In its August 22, 1988 issue, *Newsweek* reported that there was a virtual moratorium in the United States on building new high-voltage power lines because of "community opposition and environmental worries."

2.3 Opposition Reported Worldwide on the Web

Opposition is not limited to the United States, as power line opponents are pleased to announce. From the late l980s through 2000, people opposing siting by Ontario Hydro of a 500 kV AC transmission line through their neighborhood in Bridlewood, Ontario, Canada (*Bridlewood O-Hydro 500) provided an information service to tell their local colleagues and later others across the world about why and how transmission lines were being opposed almost everywhere. This service appeared first as newsletters and later as a Web site dedicated to spreading the word about struggles against power lines and other technologies that produce electromagnetic fields. The site briefly summarizes approximately 50 cases in which the siting or operation of electricity transmission facilities was opposed by community groups or permitting authorities. Many of these cases are reported for localities across Canada and the United States, but also reported are cases in the Philippines, Spain, Morocco, Australia, Italy, Venezuela, Greece, Mexico, Ireland, and Antigua. The summaries are brief and not sourced. One purpose of this list and others like it is to advertise that opposition to transmission lines is legitimate and common.

The Bridlewood site was discontinued in 2000, a few years after the Hydro-Quebec transmission line was approved. However, people in southeastern suburbs of St. Paul, Minnesota, opposing Xcel Energy's project to upgrade a 115 kV power line in their neighborhood (*Xcel 115 upgrade), include the Bridlewood list of world power line struggles in their Web site. These protesters of Xcel have been in the news for contesting the upgrading project through 4 years of hearings and litigation. Their Web site is an important node in a worldwide network of transmission line opponents. One of these is the group Revolt, organized in North Yorkshire, United Kingdom, in 1991 to resist siting by the National Grid, a transmission only firm, of a 75 km, 400 kV AC link from Lackenby south to Shipton, near the city of York (*North Yorkshire 400). Their resistance apparently helped delay project approval until 2002.

Note that above-mentioned cases marked with an asterisk are discussed throughout this article to illustrate opponent ideas and organization. The designations, e.g., AEP 765, CU $\pm$400 kV DC, are the author's for this article and not necessarily used by utilities or government.

2.4 Counteropposition

Opposition is the main observable and most significant public response to transmission projects; however, there are some instances in which people

have organized as apparent grassroots proponents of specific transmission projects. Some of these proponents are criticized by opponents as front groups for industry, or "Astroturf lobbyists."

The 1970s CU ±400 kV DC case provides an instance of this. Midway during the conflict, scores of farmers and other rural people across Minnesota came to the state capital in St. Paul for a Line up for Power/Pro-Power demonstration of support for the line. These were customers and employees of local rural electric cooperatives that were served by the two associations building the line. Their cooperatives organized the event, bused them to it, fed them, and prepared placards for them to carry with slogans proclaiming the need for more power and mocking the slogans of line opponents. Opponents confronted these proponents on the steps of the capital and in debates with legislators. Most proponents seemed uncomfortable in their role and unable or unwilling to argue for the line as ably or fervently as opponents argued against it. The cooperatives did not repeat this effort.

In the late l990s, as many people and groups in West Virginia and Virginia were resisting the proposal by AEP to build a 115-mile-long 765 kV transmission line through their neighborhoods, others organized to promote this AEP proposal. Ron Nixon reports in the December 7, 1997 *Roanoke Times* that the promoters organized as the Coalition for Energy and Economic Revitalization (CEER). CEER's leader, Bill Tanger (who also chairs the Roanoke Valley's Sierra Club), called CEER "the largest grass-roots coalition in Virginia's history." Its aim, he says, is to prevent widespread electrical blackouts in southwestern Virginia and to foster economic development in Virginia and West Virginia. However, power line opponents claim that CEER people are paid by AEP and that they are not grassroots activists, but rather a front group for AEP and what public relations watchdogs term Astroturf lobbyists. The purpose of such lobbying is to make a planned program look like a spontaneous explosion of community support. Regulatory documents show that AEP paid more than $500,000 to CEER through Tanger and another $200,000 to the advertising agency that he runs. In contrast, the AEP line opponents spent nearly $300,000 over 6 years waging their resistance against the line.

2.5 Windmill Advocates

There is another type of advocacy for transmission line construction that is not considered a front for industry, but is supported by those otherwise opposed to new electricity generation and transmission. This is advocacy for facilities to transmit electricity from windmills and other renewable or "green" sources, particularly if these facilities are owned by farmers, ranchers, communities, and small businesses rather than large utilities. Wind/alternative energy advocates in the United States and Europe have long complained that transmission facilities are not adequate to incorporate electricity from these alternatives into the grid. In Minnesota, North Dakota, and South Dakota, states with considerable wind resources, some of the same people and groups who have opposed transmission lines from conventional generation facilities have supported the expansion of transmission and generation when these include electricity from wind turbines. In January 2003, Minnesota regulators approved a proposal by Xcel Energy to build high-voltage transmission facilities in southwestern Minnesota under conditions that require it to speed up its purchase of wind-generated energy by several years. The routes for the line remained to be permitted. Michael Noble, who directs an organization promoting sustainable energy, Minnesotans for an Energy-Efficient Economy, says that unlike power lines from conventional sources, which are controversial, this power line system will be supported by environmentalists and local communities benefiting from the generation and sale of wind energy. One strong supporter is George Crocker, who helped lead urban allies of west-central Minnesota farmers in their l970s fight against the CU line and helped lead people in Minnesota and Wisconsin in resisting the Chisago and Arrowhead-Weston lines in the 1990s and early 2000s. Crocker says these activities are consistent because all help farmers and other ordinary people to make a living and persuade industry to improve environmental quality.

In summary, opposition is the most significant and observable public response to the siting of electricity transmission lines and opposition combines with siting approval processes and financial and regulatory uncertainties to block or delay the construction of new transmission systems.

3. HOW DO TRANSMISSION LINE OPPONENTS ORGANIZE?

The siting of a transmission line is usually opposed by many groups, some organized specifically and locally, others participating as allies advocating related causes. Groups are composed of many

persons and leaders who vary in their commitment and persistence. They oppose transmission lines for many reasons and interact in multi-issued communication networks. They act as interest or pressure groups to influence political decisions more than they act as a movement to change society and culture, but they have the potential to promote such change.

3.1 Participation in Groups: Cadres and Peripherals

Some people oppose transmission facility siting as individuals, but most do so as participants in groups. Over time, the organization of opposition is often composed of many groups of varying types that interact in networks. Key groups are those organized among neighbors to resist the line and its impacts in their locality. These local groups may cooperate through umbrella organizations to contest the line over its whole length and in state or other centralized hearings or litigation.

Participants in the groups organized to fight the line vary in their depth of commitment and the time and effort that they devote to the fight. The result is that a local struggle against a transmission system is usually waged by a few core activists and a larger number of peripheral participants. Core activists represent the group consistently in public hearings and media interviews and constitute its cadre. They organize and direct the campaigns through which peripheral participants raise money, sign petitions, send letters, lobby officials, participate in hearings, demonstrate, and monitor surveying and building. Core activists also manage group newsletters and Web sites. The numbers of people participating peripherally oscillate widely. Hundreds come out for some events, notably important public hearings, but retire when their campaign seems either to have succeeded, for example, in forcing a rerouting of the line, or to have failed. However, hundreds again appear when they learn about new threats. Furthermore, some people on the periphery become more committed and join the core, even leading it when others tire.

3.2 Groups Divide and New Groups Form

People oppose transmission systems for many reasons or concerns and core groups may try to embrace the widest array of these. However, when people feel their concern is ignored, they may form a new group to express it. Also, people form new groups when they differ over tactics or use a different attorney. Furthermore, people tend to form groups in each of the many localities potentially impacted by a transmission project and as routes are decided or changed, those who escape impact often drop out and those who are newly impacted launch their defense.

3.3 Groups Organize a Division of Labor

Core activists usually organize a division of labor, allocating roles according to the capabilities and experiences of their participants. Often they also organize a weak hierarchy using the institutional models familiar to them. Someone becomes chair or president, another becomes secretary, and another becomes treasurer. Furthermore, core activists also name their group, commonly in words that describe their concerns and whose initials form a colorful acronym. Groups establish officers and organizational names in part to facilitate their participation in official public hearings and litigation, to raise and account for money, and sometimes to give their participants the legal protection of incorporation.

3.4 Many Groups with Diverse Issues Collaborate in Networks

Groups organized specifically to fight the line are often joined in resistance by representatives of groups already established to advocate other causes deemed related, such as environmental protection, environmental justice, native rights, or green energy. Also, representatives of established civic and commercial associations and local governments sometimes support opposition efforts. These various parties exchange information about transmission and companion issues with persons and groups across nations and the world sharing these concerns. The Internet facilitates this communication. A group's Web site details its resistance efforts and also links to Web sites of other groups that oppose transmission and generation facilities, promote alternative energy, or warn about the effects of electric and magnetic fields on health.

The groups organized locally and voluntarily to oppose a transmission line project are often criticized as having but a single issue, stopping a line, but over time some group participants work on many issues related to the generation, transmission, and use of electricity and the impacts of electricity technology on health and welfare. Opponents embrace many

issues because they find that the best way to oppose a specific energy project is to propose alternatives to it and to debate energy policy and the culture of energy use. Another reason people expand issues is that during their campaigns against the line, they collaborated with other groups concerned with these other matters. For instance, people organized to oppose the Arrowhead-Weston transmission line in Minnesota and Wisconsin allied with an environmental justice group, Just Energy, and the Split Lake Cree First Nation in Manitoba, Canada, to protest the impact on the Cree of the hydroelectric dams likely to transmit electricity over the lines. Thus, the line protesters expand their issues to include environmental justice and native rights.

3.5 Pressure Groups or a Social Movement?

Environmental justice and native rights activism, like environmentalism, are generally understood to be social movements. In contrast, transmission opposition groups are not referred to as movements, but rather as interest groups or pressure groups, terms that refer to their efforts to influence specific, local political decisions. A movement can usefully be defined as a group of people who are organized for, ideologically motivated by, and committed to a purpose that implements some form of personal and/or social and cultural change, who are actively engaged in the recruitment of others, and whose influence is growing in opposition to the established order within which it originated or which has controlled it. Transmission line opponents become more like a movement to the extent that they not only try to influence decisions about the siting of a specific line, but try to change energy policy and the culture of energy use among their fellows and elsewhere. They become more like a movement when they recruit ever more people to participate with them in promoting these changes and when their local groups interact with others holding similar ideas across the world. This potential is not yet realized.

Transmission line opponents are sometimes labeled environmentalists, participants in the environmental movement. However, protecting the environment is only one of the reasons people give for opposing transmission lines. As noted below, line opponents often complain that their farm or community is sacrificed to a transmission route because environmentalists and environmental protection legislation have exempted parks, forests, wetlands, or even scenic highway corridors from inclusion in the route.

3.6 Example of Organization and Constituency

Rural England Versus Overhead Transmission Lines (Revolt) provides an example of organization and constituency. Revolt has been led by a small core group since its founding in 1991 to oppose a 400 kV transmission system proposed to be sited in North Yorkshire, United Kingdom, by the National Grid. The Revolt chair, M. J. O'Carroll, writes that the core has worked especially with approximately 1000 households and 100 landowners and with them has obtained 8000 written formal objections and two petitions of 12,000 and 14,000 signatures opposing the proposals. Revolt describes its participants as a widespread mixture of people, urban and rural, living inside and outside the region affected by the line, as well as overseas. They range in income from affluent, to middle income, to poor; some are large rural landowners, some live in small towns and suburbs, and some are residents of government housing. Revolt has also been joined by local government authorities of main and minor parties. Its participants oppose the lines for many reasons. Some from urban Middlesbrough support the campaign because they value the surrounding countryside for recreation; others have different reasons. In the countryside, landowners and other residents worry that the lines will lower their property values, but some residents of a prosperous suburb of Middlesbrough favor the new transmission system because it will be followed by - the removal of an older 275 kV line from their neighborhood.

Even though the line was approved and energized by 2003, a few Revolt leaders continue to communicate ideas about transmission lines and other electricity and energy issues through their Web site. The site is linked to the sites of transmission line opponents around the world. Revolt has been an important node in the network of electricity transmission line opposition and contributes ideas about change, which inform a movement.

4. WHAT REASONS DO PEOPLE GIVE FOR THEIR OPPOSITION TO TRANSMISSION LINES?

People express many reasons for opposing transmission lines. What are these reasons and how do utilities and their consultants interpret these?

4.1 Understandable Feelings?

The director of transmission and distribution operations for the Tampa Electric Company (TECO) tells reporters for the August 18, 2003 *Tampa Tribune* "I understand how people feel. No one would want a (power line) pole in their back yard." TECO officials say that they had to build the line and its 125 ft tall, 3 ft wide concrete poles through Egypt Lake because the neighborhood grew up around their substation and right of way. They affirm that they strive to be a good corporate citizen (by lessening the impacts of facility construction), but have an obligation to provide economical and reliable power. Described by the article as "repentant," the officials say they are open to discussing how they can make the process of siting facilities work better.

4.2 Not in My Backyard?

One of the ways electricity utilities try to improve process and public relations is by having consultants explain why and how people oppose their projects. One such consultant, Anne Gunning, writes in a 2001 Edison Electric Institute pamphlet that whereas many project opponents may resist simply because they are opposed to anything in their backyard, increasingly people identify "real reasons" for their opposition. Gunning lists the popular acronyms that are used (by critics to explain resistance to their projects) and suggests that these acronyms and reasons have evolved. In the 1970s, there was NIMBY. In the 1980s, there was BANANA. In the 1990s, there was LULU (Locally Undesirable Land Use), NOPE, and CAVE (Citizen's against Virtually Everything). In the 2000s, there is TARR (There Are Real Reasons). TARR, according to Gunning, is something new and important, which utilities must recognize. Dismissing opposition as NIMBY will not advance a project, she says. People believe that science, law, or some other "real" proof is on their side, even though they also often agree that new energy infrastructure is needed. Gunning writes that the concerns can be about money, environment, health, or quality of life. Although she advises utilities to recognize this, she also states that the people holding these concerns "will not change their minds based on evidence." Furthermore, her presumption that it took people until the 2000s to base their opposition on real reasons rather than selfishness or obstructionism is incorrect. People have argued real reasons throughout the history of opposition and certainly since the l970s.

4.3 Many Reasons Summarized

People express their many reasons for opposing transmission line siting in many forums. In three of these, Web sites, public siting hearings, and agreements between utilities and publics, they summarize their reasons.

4.3.1 Save Our Unique Lands

Save Our Unique Lands (SOUL) was formed in 1999 to oppose a 345 kV transmission line, 250 miles long, from the Arrowhead substation in Duluth, Minnesota, to Wasau, Wisconsin. The line is proposed by Allete (formerly Minnesota Power) and Wisconsin Public Service. These utilities named the project "Power Up Wisconsin," reflecting their claim that this will ease a power shortage in Wisconsin and improve the reliability of the transmission system. SOUL opposes the line for many reasons, synthesizing these on its home page. Combining reasons of voltage, length, appearance, need, and fairness, SOUL refers to this as " a giant 345,000 volt 'bulk transfer' electric and fiber-optics line, proposed by two private for profit utilities, to slice a 250 mile permanent scar across the face of beautiful northern Wisconsin." Adding concerns about environmental justice, equity, and property rights, it calls the line a huge "extension cord," planned to cross the Ceded Territory of the Chippewa in order to transfer massive electric sales to points south and east and on through Chicago. It warns that the project's "great negative impacts of health risks, property devaluation and ruination of the 'north woods' would be carried on the backs of the 11,500 private property owners and the tribes, as the companies use eminent domain to 'take' private lands." There is "blood on the lines," because the electricity sold on the lines comes from "two devastating sources." One is coal plants in North Dakota, whose emissions add to global warming and deposit mercury in the northern lakes, contaminating fish consumed by women and children. Another source is the hydroelectric dams of Manitoba Hydro, "mega-dams," which have flooded an area the size of Minnesota belonging to the Cree. According to SOUL, traditional Cree communities have experienced "inhuman suffering" because of this. And, according to SOUL, the power carried by the lines is not needed "to keep the lights on" in Wisconsin," but instead is to be used to power cities elsewhere as well as proposed copper–nickel–iron mining in northern Wisconsin, mining that SOUL and its allies also oppose.

The line siting authority, the Wisconsin Public Utilities Commission (WPUC), did not determine that SOUL's reasons were sufficient to stop the line project, but instead found that the line was needed and approved a specific route for it by October 2001. But in November 2002, the American Transmission Company, a transmission-only firm, which will build and operate the line, filed a petition to reopen the case because its new estimated cost for the project was $420 million rather than the approved cost (of $165 million). The WPUC will go to new hearings and opponents have this cost increase as a new reason to object.

4.3.2 Revolt

In the United Kingdom, the group Revolt (it does not capitalize its acronym) also effectively summarizes on its Web site some of its main reasons for opposing the National Grid's proposed 400 kV transmission line in North Yorkshire:

1. It is not needed for public electricity supply, only for profit (for the electricity industry).
2. It will ruin large areas of treasured countryside.
3. It will ruin some families by massively devaluing their homes, without compensation.
4. It will promote uneconomic bulk transport of power from north to south, wasting valuable energy in generation and transmission, and cause environmental damage.
5. This conflicts with new government energy policy favoring efficient local generation combining cooling and heating.
6. It will cost consumers dearly, in capital to build and in revenue for increasing operating costs both for the line and for knock-on grid developments throughout England.
7. The great majority of landowners refused to grant wayleaves (rights of way), despite outrageous inducements.
8. There is overwhelming public and all-party local government opposition.
9. People fear possible health risks from electromagnetic fields, which new evidence is reinforcing.

4.3.3 Hearing Testimony Opposing AEP 765 kV

In 1998 and 1999, public hearings were held in localities in Virginia impacted by two possible routes for a 765 kV line and related facilities proposed by a subsidiary of AEP. In his October 2000 report of the hearings, the hearing examiner describes the witnesses, noting that over 500 witnesses testified. Some of these witnesses were members of a number of groups organized expressly to oppose the line in various localities. Some were members of local governments along the routes. Others represented established civic organizations, such as historic preservation committees. The examiner writes that the overwhelming majority of the public witnesses were opposed to any kind of transmission line in southwestern Virginia and he summarizes their reasons as follows: The line is viewed as a symbol of corporate greed at the expense of the cultural attachment of the people to their land and the scenic beauty of the region. The land is mostly pastoral—rolling fields and hills with mountain backdrops, and cattle and sheep farms with streams running through them. The line is described as a means of transporting electricity generated in the Midwest to sell in the Northeast, enriching corporate buccaneers and their stockholders. Risks to health and environment include contamination of ground and surface water from herbicides sprayed to keep rights-of-way cleared of vegetation, increased cancer caused by electromagnetic fields (EMF) radiating from the line, increased air and water pollution from the coal-fired generating plants providing power to be transported on the line, and noise from the line. Building the lines and towers will threaten caves and sensitive rock formations and old growth forests. Demand-side management and local generation provide alternatives to the line. The routes being considered are unfair in that they skirt National Forest land at the expense of the homes and land of ordinary people and in some places at the expense of poor people. The project will harm tourism and its benefits to local economy. Additional reasons were also given.

4.3.4 Agreement between Hydro-Quebec and Farmers in Canada

A summary of many of the impacts that transmission systems have on agriculture is found in the December 2000 statement of agreement between the Quebec farmers' association [Union des producteurs agricoles (UPA)] and the giant utility, Hydro-Quebec, in Canada. Hydro-Quebec explains that it is a government corporation that provides electricity primarily through hydroelectric plants often located more than 500 km from consumers. To reach consumers, it has constructed a vast power transmission system and its transmission lines run from east to west across many farms and woodlands. Hydro-Quebec and UPA agree that constructing such lines can have the following temporary and permanent impacts:

Temporary impacts during construction include:

- Impacts associated with the staking of rights-of-way;
- Reduced crop yields due to soil compaction;
- Alternation of underground or surface drainage systems;
- Alteration of irrigation systems;
- Damage to ditches;
- Broken fences, which can hinder livestock control;
- Noise produced by construction machinery, which can affect poultry and fur-bearing animals;
- Disruption of crop operations;
- Loss of time (during negotiations, for example);
- Loss of revenue (cash flow) while awaiting compensation;
- Impacts on areas or elements located outside the right-of-way, such as damage to farm roads, debris from tree felling, ruts and soil compaction, damaged trees, waste materials; and
- Construction debris and other waste materials.

Permanent impacts include:

- Loss of farmland or woodland;
- Loss of revenue, which could compromise the operation's profitability;
- Loss of time (e.g., time spent in negotiations or driving around the structure);
- Risk of farm machinery running into the structures;
- Creation of enclaves;
- Usage restrictions and other limitations associated with easements;
- Alteration of irrigation systems;
- Changes to crop operations;
- Impossibility or increased danger of using airplanes in agriculture.

The document also mentions that studies are being conducted to identify and analyze the biological impacts of electromagnetic fields on human and animal health.

4.4 Rules, Process, and Influence Are Also Reasons for Opposition

The impacts identified in Quebec are like those expressed by farmers in west-central Minnesota in the l970s about the CU $\pm$400 kV DC line. The Minnesotans, however, complained that their concerns about these impacts were not heeded by the utilities, the siting officials, or the environmental and transportation agencies. Like some in southwestern Virginia, the Minnesota line protesters complained that these agencies were constrained by environmental protection legislation to be more concerned about protecting parks, forests, wetlands, and highway corridors from power lines than protecting farmland. In other words, it was not only the decisions about the line that were wrong, it was also the process of making the decisions that was wrong. This is a common complaint in transmission line cases. Often opponents of transmission siting decisions also complain that even though they have presented many good reasons why a transmission project should be denied or fundamentally changed, ultimately the project is approved, perhaps with minor modifications. Government siting authorities say that they do heed the concerns of opponents. But they face a dilemma. On the one hand, people show that they want and need more electricity at an affordable cost. On the other hand, people do not want to be impacted by the facilities required to provide this electricity. They must reconcile the diverse interests and also manage the conflicts among these interests. Facility opponents may recognize the dilemma, but also suspect that the siting process favors energy companies, with their vast financial and political resources, and favors growth in electricity supply, since this accompanies desired economic growth.

In short, people present many reasons for opposing transmission lines. Some reasons concern the rules and processes used to make the decisions and many reasons concern the characteristics of the line, the need for it, and its impacts on environment but especially on human health and welfare. In the next section, this article explains in more detail how people argue against the line with respect to its need, its size, and the structure of the utility industry; its appearance and threat to their sense of well-being; and its threat to their health and safety.

5. NEED: THE FIRST DEFENSE AGAINST TRANSMISSION LINE SITING

Opponents of transmission lines quickly learn that the first defense against the line project is to declare that is not needed. If it is not needed to serve the public interest, then there is no reason for government to allow the electricity utility or transmission company to march the line over private property or community resources, using the power of eminent domain, and impacting environment, health, and welfare. Such logic is institutionalized in government

siting regulations. In the United States, state siting authorities decide whether an electricity-generating plant or transmission line is needed before, or as, they decide where to place the facility and whether the impacts on environment and people are acceptable or can be made so. The authorities usually decide this and award "certificates of need" and "environmental compatibility" through siting processes at the state and local levels. Processes usually include public participation, such as meetings in which the public is informed about the project and hearings in which the public critiques it. States vary in the structures and procedures that they use to make decisions about need and route, with some using one agency and others using multiple bodies.

5.1 Need Is Public

Need is defined as public. That is, siting authorities must determine that the facility serves the public interest rather than the interest of the utility. In the United States, it is has been the states that approve and regulate facilities within their borders and the people whose interests that must be satisfied by a decision by state authorities are the residents of this state. Opponents often talk about this when they ask questions about transmission systems that cross state lines. In Connecticut, for example, opponents of a projected transmission system that they suspect to be for the electricity supply of New York reminded their siting council at a 2002 hearing that "the state legislature has given the council jurisdiction to objectively balance the statewide public need for adequate and reliable services at the lowest reasonable cost to consumers with the need to protect the environment and ecology of the state for the construction and operation of facilities under its jurisdiction."

The utilities applying for a facility will also argue the public interest, claiming that they have the state-mandated requirement to meet the demand for electricity and that this demand is demonstrated by the evidence of public use. The president and chief operating officer of the Wisconsin utility proposing the Minnesota–Wisconsin Arrowhead-Weston line explains (in an EEI brochure about facility siting) that "customers expect electricity to be there when it's needed, so we have to make sure the delivery system can fulfill their expectation." How is expectation measured? Regulations tell the WPUC that the proposed facilities are to satisfy the *reasonable* needs of the public for an *adequate* supply of electricity (author's italics). Permitting agencies also often require that the applicant demonstrate that demand cannot be met more cost-effectively through other measures, such as more local generation, instead of long-distance transmission, or energy conservation and load management. Utilities, especially their economic and technical specialists, present detailed statistics and forecasts to argue their case before the permitting agency.

However, the utilities must also communicate need to the public, including those opposing or considering opposing the line. This public, particularly the opponents, will not only respond to claims of the utility, it will offer counterarguments, including counter-statistics to refute projections of use and counter-concepts about expectations and the meaning of reasonable and adequate. It was partly public opposition to electricity facilities and critique of need beginning in the l970s that led legislators in Minnesota to require that utilities in the 2000s show that their calculations of need include consideration of conservation and other alternatives to a proposed facility.

5.2 Need Is Argued Objectively and Subjectively and on Three Fronts

Need is argued using numbers. Numbers, statistics, graphs, calculations, and projections seem central to debates about need. The Connecticut Siting Council, for instance, puts public need and benefit at the top of its list of factors, which can be objectively presented and argued. However, need is also argued subjectively. For instance, the meaning of reasonable and adequate is open to debate, much as is the difference between needs and wants. Power line opponents certainly are willing to debate these and other questions about need to further their efforts to stop the line. Opponents challenge project need because this is a major defense and because they can find much about need to challenge even though proponents present need as manifest and can interpret opposition as "NIMBY" selfishness. There are three ways in which people contest the need for a project. One is by contesting projections of local electricity demand and calculations of supply, arguing instead that the electricity will be exported for a profit. Another is by claiming that energy demand can be met with alternative technologies or reduced through conservation. People also reject the rule that once a certificate of need is granted, need can no longer be contested.

5.2.1 Contesting Supply, Demand, and Fairness

Opponents contest the calculations that the utilities and government regulators use to argue that

electricity demand is increasing and must be met with new supply. Opposition groups often include people with the knowledge and skills to do this well. At a mid-l970s public meeting in Minnesota on the CU proposal for a ±400 kV DC line, one of the proposal opponents, a farmer with a doctorate in physics and previous employment in the aerospace industry, critiqued CU's extrapolations from the electricity consumption curves in west-central Minnesota. Speaking for the county group No Power Line, he argued that the curves showed a leveling of demand during the early 1970s so that power line construction could be postponed until the mid-1980s, when the technology of superconductivity would be advanced enough to put the line underground.

In 2002 hearings before the Connecticut Siting Council, Woodland Coalition opponents of the CL&P proposals to construct a 345 kV transmission system in southwestern Connecticut also contested the utility's projections. One said in written testimony: "If the definition of need is to be defined by the anticipated power consumption of southwest Connecticut, further supported by CL&P's own calculations of power consumption growth over the past 20 years, then I assert to this council that CL&P has fallen woefully short of proving its need for the amount of capacity expansion that it seeks." This critic goes on to argue that there is no need for an eightfold increase in generation capacity unless the council has expanded its definition of need to include connection to a regional and national power grid. It is this issue of transmission of electricity across jurisdictions that informs the most powerful challenge of need and its determination, namely, that the need being served is someone else's.

The Woodlands Coalition line opponents complain that the utilities will use the new facilities to sell electricity to Long Island and New York. Connecticut people have paid or will pay for the generating and transmitting facilities and will suffer the impacts of these on their environment, health, and quality of life. Long Islanders have not done their fair share of building energy facilities. One of the Connecticut line opponents used his expertise as a financial analyst to argue that the utilities want to build the transmission not to meet local need but to make a profit. He finds this particularly true of the unregulated energy trading company component of Northeast Utilities, the firm that will build and operate the transmission systems. He says that it is as wrong to give an unregulated and for-profit utility the rights of Eminent Domain as it would be to grant this right to Wal-Mart.

Similarly, opponents of other interstate transmission projects contest need on the grounds that these projects serve the financial interests of utilities. To this, they also add charges of unfairness. The utilities will provide electricity to those who did not pay for the facilities and, in addition, by exporting electricity, local prices will increase. Although utilities customarily argue that there is a need for more transmission, some, as noted for Basin Electric, also say that deregulation has made it unclear whether those firms and customers who pay for transmission benefit fairly from it.

In the l970s, well before deregulation, the west-central Minnesota farmers and rural townspeople opposing the ±CU line also complained that the line was not needed locally but instead was being used to power "the Cities," of Minneapolis and St. Paul. Sometimes these rural Minnesotan's pictured the people of the Cities, especially the suburbs, as using electricity wastefully, but they also recognized that many of these people were their relatives and tried to enlist their aid in fighting big industry and distant government. This theme of energy flowing away, to cities, is a continuing one, expressed in 2000, for instance, in the portrayal of the Arrowhead-Weston Line as a giant extension cord to Chicago. Even without deregulation, utilities and siting agencies would have difficulty persuading people that transmission through their communities served their local needs.

One main response of the utilities to such problems is to argue that transmission across states does meet local needs by contributing to overall system reliability. A number of Edison Electric Institute publications during the l990s make this argument. But it is an argument that is as likely to be as resisted as are other arguments for facility need, especially if the utility industry is increasingly put on a free market basis. James Creighton, a consultant on public participation for electricity utilities, wonders what will happen if state regulatory bodies say that it is unnecessary to issue a certificate of need for transmission lines because the market can decide need. This will reduce regulatory costs and may seem to speed up decision-making, something that some political leaders advocate. However, Creighton warns, the public is not accustomed to letting the market decide transmission need. It will challenge need and either the state will return to the issue or a plethora of local government entities will act to manage these challenges. The alternative to this is that the federal government, perhaps through the FERC, will act to determine need anywhere in the

country and perhaps also to make other transmission facility siting decisions. This has been advocated by some. But, as has been noted, state authorities resist this federalization because it threatens their traditional rights and they know it would be resisted by the public. Transmission line siting processes could become not only a battleground for making decisions about the need for and siting of electricity facilities but about state's rights.

5.2.2 Conservation and Alternatives

Another argument against the need for a new facility is that energy demand can be met with alternative technologies or reduced through conservation.

In the l970s CU ±400 kV DC case in Minnesota, line opponents pointed to technologies such as underground lines using superconductivity; wireless broadcasting of electricity using the ideas of Nikolai Tesla; and local generation using windmills, biomass, photovoltaic cells, and low-head hydroelectric dams. The quest for alternatives encouraged the power line opponents to ally with advocates of "soft," renewable energy sources.

Opponents of the generation and transmission of electricity through conventional technologies continue in the early 2000s to call for alternatives and to work with proponents of the latter. Sometimes the demand for alternatives seems initially a device to contest the need for a conventional overhead transmission line and legitimate resistance to it. However, advocates for alternatives have used the threat of public opposition to conventional facilities to induce utilities to include increasing amounts of wind and conservation in their energy mix. Mention has been made of the way that wind advocates work with utilities in the United States to support the need for more transmission lines if these also carry electricity from wind turbines owned and operated by farmers, ranchers, and rural communities. Efforts are also well under way in the United Kingdom and Europe to advance the use of wind resources by developing better ways to transmit electricity from these sources. However, in the United Kingdom, leaders of the group opposing the 400 kV line in North Yorkshire write that because wind resources are intermittent, incorporating electricity from them into a reliable grid is not efficient. Utilities must build too much infrastructure to do this, including more transmission lines. This argument is consistent with Revolt's criticism that large-scale transmission systems with remote power generation are inefficient. They lose vast amounts of energy transmitting over long distances. The need for new lines can be diminished by reducing this inefficiency and improving the efficiency of energy use. Energy supply can be improved by using new technologies that combine energy functions. Wind should be used to make a locally storable form of energy, such as hydrogen by electrolysis of water.

5.2.3 Contesting Rules about When Need Is Argued

In addition to contesting the need for a facility, its opponents also often argue that because of procedural rules they are unable to debate need at appropriate times. On the one hand, people criticize efforts to site facilities before need is demonstrated. The case of the CU ±400 kV DC line in 1970s Minnesota illustrates this. Because new state legislation governing energy facility siting had not yet fully taken effect, the siting authorities held hearings on where to route the line before they could hold hearings to determine the need for it. Those opposing the line argued that this made the hearings not legitimate and continually tried to question need during the routing hearings.

However, even though need is usually determined before debates about routes and impacts take place, opponents often never agree that there is this need or that they had adequate opportunity to argue against it. For one, fewer people come to the hearings to decide need than to those to decide a specific route. People are notified that they might be affected by a facility to meet such need because they are in a corridor from which a route could be selected. However, most do not come to hearings on need, apparently because they are busy, the corridor is wide, the threat of a route over them is distant, and the subject of need is general. Those few who do attend speak of the problems they have in persuading their neighbors to join them.

On the other hand, the rule that need, once decided, can no longer be at issue does not stop people from arguing against it when they are defending themselves against the actual threat of a line overhead. Furthermore, the rule itself is questioned. In early 2003, the Minnesota Center for Environmental Advocacy (MCEA) praised a change in the rules governing energy facility siting for not allowing approval of the construction of a new power plant or line when a less environmentally damaging alternative exists. MCEA argues, however, that continued useful and fair consideration of the alternatives is jeopardized by the rule. It envisages a situation in which people do not participate in deliberations on need because they do not know that this concerns them. When people find that the

facility may be built next to them, they can no longer raise the question of need and no longer do this by identifying and arguing for alternatives, such as conservation or different technologies. MCEA will likely lobby the legislature to change the rules.

6. "SIZE" OF THE TRANSMISSION SYSTEM AS A REASON

People find reasons and organizational opportunities to oppose transmission systems in what some term the size of the system and others term design and configuration. Size includes the width and height of towers, the length of the line, the voltage carried, and the complexities of the system. People can protest towers and lines of any size as invasive, but they usually find more reasons to protest larger towers and higher voltages. Furthermore, longer lines are more likely to impact more people, increasing the numbers and diversifying the skills of those who start or join local groups. Siting processes usually vary according to the size of the voltage and length of the line, with official opportunities for public participation (and thus opposition) often increasing as voltage and length rise over a certain threshold. States and other political units vary in their rules about size thresholds and in their criteria and procedures for making decisions. Thus, as lines extend across jurisdictional borders, they increase the exposure of line builders to different rules and procedures. Examples of how opposition groups respond to size are presented below.

6.1 High Voltage, Remote Siting, Long Distances

In their 1977 study for the Minnesota Department of Health on the health and safety effects of high-voltage transmission lines, researchers attribute much of the public worry about these effects to the fact that by the l960s and early l970s the electricity utility industry was constructing or proposing transmission systems much higher in voltage and much longer in reach than the public had hitherto experienced. These were 500 and 765 kV AC lines, then designated extra-high-voltage alternating current, and ±400 kV DC lines, designated high-voltage direct current. Physical efficiencies, new technologies, and perceived new needs and opportunities gave utilities incentives to use such systems. Demand for electricity was growing exponentially; however, utilities found it increasingly difficult to build generating plants near the main load (use) centers because of public opposition and regulations to protect air and water quality. One answer was to build power plants in remote locations. One such place was the Great Plains, where there were resources of coal or lignite fuels to be mined, river water for turbine cooling, and air that could suffer some degradation within clean air rules. Another was in Manitoba or Quebec, Canada, where there were rivers to be dammed for hydroelectric generation. New technologies enabled firms to build higher voltage lines from generators to load centers, taking advantage of physical efficiencies. The capacity of a line to carry power varies approximately as the square of its operating voltage. Thus, doubling voltage provides four times the power transfer on the same right-of-way.

6.2 1970s: CU ±400 kV DC

Responding to these conditions, two rural electric cooperatives based in Minnesota, Cooperative Power and United Power, collaborated through the Rural Electric Administration to plan and eventually build a system of generation and transmission that was termed innovative even by critics. The electricity is generated at a 1100 MW Coal Creek station along the Missouri River at Underwood, North Dakota, fueled by lignite coal mined in a giant strip mine next to the station. The alternating current produced by the generator is converted into direct current and transmitted on a ±400 kV DC power line 430 miles eastward to Delano, Minnesota, just west of the Twin Cities of Minneapolis and St. Paul, where it is converted to alternating current for distribution to customers of the cooperatives, organized as CU. CU officials could take pride in this large size, long expanse, high voltage, and complexity. But these characteristics provoked and facilitated opposition.

The opponents criticized the technology and built organization along its likely routes. They said that when CU used the designation ±400 kV DC, they were trying to hide the fact that this meant the voltage between the two pairs of wires of the system is 800,000 V, not 400,000 V. If the voltage difference between one of the pairs of wires and the ground is +400 kV and the voltage difference between the other pair of wires and the ground is −400 kV, the total, they say, is 800 kV. Only one other system in the United States, on the West Coast, has as high a voltage. CU's opponents said that they should not be made guinea pigs for new power line technology.

People organized opposition groups in different localities along the long corridor of the line in Minnesota and also shared information with a group in North Dakota that opposed the strip mine and power plant. These different groups contested the project locally, using particular approaches, but also cooperated. People also attacked the need for and fairness of the line by declaring that its 430 miles ends not in their lands, but in the suburbs of Minneapolis and St. Paul—the Cities. Yes, they need electricity and are willing to bear the impacts for such, but this is electricity for others. Some pictured these suburban and urban others as energy wasters.

6.3 1970s: PASNY 765 kV AC

In another case of remote siting and long-distance transmission in the early 1970s, PASNY proposed a 765 kV AC line to connect PASNY's system with that of the Canadian utility, Hydro-Quebec, rich in power from its dams. In testimony in April 1977 before a New York State Assembly committee reviewing protested siting decisions, a protester recalled that when she and her neighbors in upper New York heard about the proposed line through a series of meetings in 1973, l974, and l975 they were "unconcerned," supposing it would be "another power line like the 230,000 volt line in our area." But when she and others learned that it would be much higher voltage, they worried. Furthermore, when they looked at the long route of the line, they complained that it serves not them, in upper New York state, but New York City.

6.4 1990s and 2000s: A-W 345 kV, AEP 765 kV, CL&P 345 kV

Voltage and length continue to matter in the 1990s and 2000s. People organized as SOUL to fight the A-W 345 kV line in Minnesota and Wisconsin picture it as a giant extension cord for Chicago. Another example is found in the response to the AEP 765 kV line from West Virginia to Virginia. Persons interviewed by the *Roanoke Times* reporters R. Nixon and L. Caliri for their 1997–1998 series on the AEP project worry about the size of the line. The operator of an apple orchard notes that he and his workers already put up with a 138 kV line in the orchard and a 345 kV line nearby. But these are "nothing" compared to the 765 kV line. "Nobody's going to want to work under all that line." A vice president of the National Real Estate Appraisal Institute said that studies in some Florida counties show that high-voltage lines lower values in and near the easement.

In Connecticut in the early 2000s, the Woodlands Coalition opposed three related line-building projects of CL&P. Testifying before the state siting authority, a coalition activist who holds a Ph.D. in physics and heads a technology firm says that a 345 kV line is qualitatively different from 115 kV and lower voltage lines. It can produce a corona discharge, which produces potentially dangerous ionizing radiation, a health risk. The chair of the Norwalk Shellfish Commission testified that a 345,000 V system is used for long-distance transmission across open spaces and would be unsuitable for developed southwestern Connecticut.

6.5 Xcel 115 kV, More Risk from EMF?

Not all transmission line opponents point to high voltage as the problem. In St. Paul, Minnesota, the Power Line Task Force is contesting a proposal by a Xcel Energy to upgrade from a single to a double circuit its 115 kV transmission line running through their neighborhoods. Its Web site explains how 115 kV lines can be as dangerous as high-voltage lines, of 345 kV and above. Overall risks do rise as voltage increases. However, risks from EMF are associated with the amount of current flowing through the line rather than voltage. To deliver a given amount of power, utilities must push more current through low-voltage lines than through high-voltage lines. The magnetic field under a 115 kV line is often greater than that under a 345 kV line. Also, high-voltage lines are customarily built on wider rights-of-way than low-voltage lines so people live closer to them than they do to high-voltage lines. Thus, the site declares, low-voltage transmission lines (up to 69 kV) generally endanger human health more than do high-voltage lines.

In summary, people find reasons and opportunities in the size of transmission systems to oppose their siting.

7. STRUCTURE AND RESTRUCTURING OF THE ELECTRICITY UTILITY INDUSTRY AS REASONS

The structure of the electricity utility industry and changes in this structure provide people with reasons to oppose transmission lines.

7.1 Deregulation

In the United States in the 1970s, protesters argued that the electricity utilities were becoming too large and too vertically integrated. The CU project in North Dakota and Minnesota is an example. With loans and urging from the federal Rural Electric Administration, two large rural electric cooperative associations combined with a coal company to mine and burn coal and to generate, transmit, and distribute electricity. However, in accord with changes in federal and state legislation, the U.S. electricity industry has, since the early 1990s, been changing its structure away from such vertical integration, regulated primarily at the state level, to one shaped by competitive markets. By 2002, 23 states and the District of Columbia had enacted restructuring legislation and most other states were initiating or considering such a change. Utilities are dividing into regulated and nonregulated entities and are separating generation from transmission and transmission from distribution. Firms whose sole function is to build and operate transmission systems are established. Regulation is giving way to competitive markets. Markets enable firms specializing in trading to buy electricity where it is cheapest and sell it anywhere—if they can transmit it. Such transmission requires that transmission systems at the state level be integrated into regional and national grids. This calls for the building of new transmission lines. These are opposed.

In the United Kingdom, the electricity industry says that their country is a model in the development of open energy markets. A privatization process begun in 1990, shaped by the Electricity Act of 1989, has created a competitive market with open access to transmission and distribution networks, particularly in England and Wales. The grid owners (National Grid for transmission and 12 regional distribution companies for distribution) are regulated monopolies, whereas generation and supply are privatized.

Opponents of specific transmission projects find reasons for opposition in these structural features and changes. Examples of these reasons are noted in this article. In essence, people argue that new transmission lines are not needed to serve the interests of those impacted by the transmission and generation project, or the general public, but rather to profit the few who own the transmission, generation, or trading systems. This idea underlies arguments about need, fairness, and environmental justice.

7.2 Expanding Federal Control

Another structural issue that affects opposition is the projected expansion of the role of the federal government in siting and permitting electricity facilities and designing the system of electricity marketing. As argued in the 2002 National Transmission Grid Study for the U.S. DOE, the U.S. transmission system is a national public good that must be modernized by building new facilities. The study does not directly attribute failure to build these to opposition, but instead it focuses on flaws in the approval process that, among other things, fails to obtain public acceptance and fails to overcome the competing interests of local authorities and state and federal agencies, for instance, in land use. It also finds fault in business uncertainties, such as those mentioned by Basin Electric. As a remedy, the DOE expects that the FERC can play a larger role in siting and making rules. One role could be to grant designated entities the right of eminent domain to acquire property for rights-of-way when states do not. There are many who oppose such FERC expansion. The objections of the National Association of Regulatory Commissions have been noted. Also, some groups resisting specific transmission projects object to the expansion of the FERC and to deregulation and seek to influence state and federal legislation accordingly.

8. SENTIMENTS ABOUT VISUAL IMPACTS AND LOSS OF CONTROL AS REASONS

Power line opponents describe the way that transmission lines and towers mar landscapes and even utility executives agree that these facilities are “ugly.” Opponents also feel the lines and towers symbolize their loss of control.

8.1 Visual Impacts in Rural England

On their Web site, Revolt writers warn that the visual impact of 229 pylons—many taller than York Minster or the Statue of Liberty—will be catastrophic. They write that the overhead line will run through countryside of outstanding beauty, skirting a national park and impinging on ancient woodland, historic buildings, and sites of special scientific interest. They include a poem that William Wordsworth wrote in 1802, praising the magnificence of the horizon that the giant pylons will deface.

In interesting contrast, the National Grid Web site, under the title "Overhead or Underground?" includes a 1932 poem by Stephen Spender that likens pylons to whips of anger, to nude, giant girls that have no secret, and to the quick perspective of the future. However, National Grid also asks: "Everything being equal, wouldn't we all prefer electricity cables to be buried out of sight than straddling the countryside on big, ugly pylons?" Since everything is not equal, undergrounding has its own environmental, financial, and operational costs, so that it is usually not a viable option for high-voltage transmission lines. National Grid says that undergrounding may be the best option where overhead lines are not practical, as in densely populated areas. However, National Grid did agree to underground a small segment of the North Yorkshire 400 kV line because officials agreed with opponents that the pylons and line at this segment would jeopardize the views of some especially cherished landscape features of rural and parkland and that thousands of people enjoyed these views. Revolt's chair writes that one reason that Revolt has attracted the participation of many city people not under the line is that they do not want the countryside marred.

Appreciation of rural landscapes is strong in English culture. This motivates groups organized specifically to protect the quality of life and environment in rural England to include in their programs opposition to overhead transmission lines. For instance, the Campaign to Protect Rural England and the Friends of the Lake District and Cumbria have campaigned vigorously over past decades for overhead wires to be undergrounded to preserve the English countryside, especially special landscape areas. They argue that this has become economically feasible.

8.2 Visual Impacts in Minnesota

It is not only in the English countryside that people object to the appearance of transmission lines and towers. In their l970s protest of the CU line, rural Minnesotans sounded somewhat like rural Englanders in decrying the way that the towers and lines would ruin their landscape, a landscape that they appreciated for its open farm fields and rolling hills dotted with lakes and woods. Expressing the sentiments of many, one said, "You see, these wide open spaces are why we're here. If you have these big old monsters, you suddenly have an encroachment on your space. And it would destroy our peace of mind. It would destroy the quality of our life."

8.3 Lines and Towers as Symbols of Invasion and Loss of Control

Such words point to another reason that people oppose transmission lines. They represent the invasion and taking of their land by powerful others. In their 2000 environmental impact assessment for the Arrowhead-Weston line, analysts for the Wisconsin Public Utilities Commission recognize aesthetics and visual impacts as legitimate public concerns. They recognize that the aesthetic effects not only interfere with public enjoyment of landscapes, but also represent an invasion of landowner rights.

Under the category of "social concerns," the analysts note that people may feel that they have lost control of their lives when the towers and line are built on their land. When people learn that the line may expose them involuntarily to electric and magnetic fields and that these fields may pose risks to their health, they may interpret this as a threat to their freedom to choose their risks. The comments of power line opponents agree with these assessments. Landowners can be compensated with money for taking their land, but this may not compensate for the intangible threats to their sense of well-being.

8.4 Political Leverage of Sentiments

Sentiments about the ugliness of transmission lines and loss of power and control are important forces motivating people to oppose transmission projects. Utility officials and government regulators may accord them some legitimacy. However, public participation specialist James Creighton, claims that risks to health and safety are more potent politically than ugliness and invasiveness. He intimates that protesters know this and base their objections on such risks.

The article next examines health concerns, notably those about electric and magnetic fields.

9. HEALTH, SAFETY, AND ELECTRIC AND MAGNETIC FIELDS AS REASONS

Opponents of overhead transmission lines often claim that these lines and supporting facilities put their health and safety at risk. Of the health risks, those of exposure to EMF have become the most argued, controversial, and politically significant. Electricity utilities and health departments define

EMF as invisible lines of force that surround any electrical device, such as a power line, electrical wiring, or an appliance. Electric fields are produced by voltage and easily shielded by objects. Magnetic fields are produced by current and these fields pass through most materials. Both fields weaken with increasing distance from the source. EMF has become an issue in transmission line siting because findings from research beginning in the early 1970s have suggested that magnetic fields may be associated with increased risks of cancer and perhaps other health problems. Opponents of transmission lines have learned about this, spread word about it across their networks, argued it in siting hearings, and helped to make the effects of EMF an issue in science and energy policy.

9.1 EMF Concerns in 1970s in Minnesota and New York

This story begins in the mid-1970s in Minnesota, where people opposed the proposed ± 400 kV DC line of CU. As part of their effort, these opponents identified others who were opposing transmission lines elsewhere and they also gathered information about transmission systems and their impacts. Several individuals in the group No Power Lines (NPL) read papers on long-term exposure to low-level electric and magnetic fields by Andrew Marino, a biophysicist and attorney, and R. Becker, a medical doctor, and an English translation of a paper in Russian that mentioned Soviet studies on railroad workers exposed to fields from a 500 kV substation. The Marino and Becker papers reported that rats exposed to the fields suffer adverse health effects. The Soviet studies suggested adverse effects on humans.

In 1975, NPL participated in public hearings held by the Minnesota Environmental Quality Council (EQC) to make decisions about the CU proposal. NPL introduced as evidence correspondence from Marino and the Soviet paper. EQC officials and the CU lawyer discounted the latter because it referred to a study not in evidence, a translation, and not the original paper of the original study. The EQC gave more weight to the analysis by an expert witness for CU, which found no significant biological effects from exposure to electric and magnetic fields. NPL participants were angered by this decision and by the way their evidence was dismissed and their abilities demeaned.

NPL and other groups continued their fight sufficiently to keep the EMF and other health and safety issues alive even after EQC issued a construction permit for the CU line. In September 1976, the groups, organized as a regional coalition, petitioned the EQC to "modify, revoke, or suspend" the permit on the grounds that the line "would be dangerous to the public health and the environment." They cited the direct testimony of Marino and Becker before the New York Public Service Commission, which, responding to its own power line protest, was conducting major hearings on the health and safety effects of two 765 kV transmission lines in New York State. The EQC Board denied this petition from CURE. However, the Board told the Minnesota Department of Health (MDOH) to conduct its own study of effects. Protesters said findings would arrive too late to stop the project. The MDOH study, a survey of the literature, was reported in October 1977, after construction was under way. Its authors concluded that adverse biological effects had not been demonstrated to exist either empirically or in the laboratory nor had they been explained by a reasonable hypothesis. They also declared it a logical impossibility to state conclusively that there are no such effects.

A main source of information for the MDOH researchers was the Common Record Hearings of the Public Service Commission (PSC) of New York State. In 1975–1976, PSC conducted an exhaustive inquiry through testimony and cross-examination of 26 scientists and engineers into the health and safety effects of two proposed 765 kV AC transmission lines in New York. One would connect the system of PASNY with that of Hydro-Quebec. Strong resistance to this line did much to persuade the PSC to conduct this inquiry.

At the hearings and in subsequent rebuttals of the PSC findings, Marino and Becker testified that biological effects might occur from exposure to high-voltage transmission line electric and magnetic fields and recommended that the 765 kV lines as designed not be built. However, PASNY presented four specialists in medicine, radiation biology, and biophysics who testified to the contrary. The PSC Board decided to authorize the line. However, it reserved the right later to require protective measures to protect the public health and safety from exposure to fields, including a wider right-of-way than proposed and reasonable restrictions on operating voltage and loading. In contrast to its Board, the PSC staff found that "the 765 kV line will probably cause biological effects in humans exposed to them on a chronic basis." It recommended relocation of residences and more research. The PSC decision was

challenged and testimony rebutted. It was not until January 1977 that the hearing was closed. Construction began but resistance continued, sometimes through militant demonstrations in the fields. These decisions and events were reported in the media and Minnesota opponents of the CU line learned about them through their New York contacts.

The Minnesota opponents continued to resist the line for 4 years after the EQC granted the construction permit in 1976. They held mass rallies at construction sites and the state capital. They picketed and harassed surveyors and workers, threatening some with violence. They confronted local sheriffs and then state police. Someone shot through the windshield of a security guard's pickup truck. Over 120 opponents were arrested but the charges against most were dropped and only 4 were convicted. As the line was constructed and tested in 1978 and 1979, some opponents shot out thousands of insulators and, over many months, some toppled 14 towers by removing bolts. The opponents called this sabotage, likening it to resistance against an invader and claiming that it was necessary to protect them from the intolerable effects of the line on their health. These effects, they declared in letters to their legislators, the governor, and the state energy agency, included headaches, skin rashes, nosebleeds, respiratory problems, dizziness, and fatigue. In addition, some reported that their livestock were aborting fetuses and producing less milk. Officials held hearings at which people told of these effects. A Department of Health researcher referred to studies by University of California (Berkeley) researchers to raise the possibility that the effects were caused by a concentration of positive ions in the air in the vicinity of a $\pm$400 kV DC line. These might cause a secretion of serotonin, a neural hormone that can produce the symptoms mentioned by the farmers. Power line opponents and officials then argued over how best to investigate these health effects and make decisions about the future of the line. The opponents said that if the line were allowed to continue operating while its effects on their health were being studied, they would be human guinea pigs. Officials did not agree to stop operation, but they did survey residents along the line about their interpretations of health effects and they did hire the Berkeley researchers to prepare updated reports about transmission line effects. Meanwhile, CU and state officials asked the federal government to help them by federalizing the line under the Rural Electric Administration and to make attacks on it a federal offense, subject to FBI investigation. No one was arrested or charged for toppling the towers. By 1981, the line was completed and overt resistance had ended; the $\pm$400 kV DC system has been operating since then and, CU says, serving Minnesotans well. The new health effects studies were inconclusive, proving neither that people suffered adverse effects under the line nor that they did not. By the early 2000s, the debate about these effects had diminished in Minnesota, although some former opponents continue to say that the effects are real.

9.2 EMF Concerns in Ontario, Canada, 1984–2000

In Canada, residents of Bridlewood, Kanata, Ontario, outside Ottawa, began opposing proposals by the electricity utility, Ontario Hydro, to site two 500 kV lines through their neighborhood in 1984, approximately 2 years before they learned about the health effects of electromagnetic radiation and referred to these in lobbying against the line. According to their Web-published "History of the Bridlewood Hydro Line Struggle," these power line opponents learned about EMF risks by reading a 1986 article on the health effects of electromagnetic radiation in the *Ottawa Citizen*. This news, as well as the failure of efforts to stop the line through local government proceedings, led them to mobilize against the line as the Bridlewood Residents Hydro Line Committee (BRHLC). This group tried to stop the line by communicating information about its costs and risks, lobbying Ontario government, and suing in the courts. They also picketed a visit of the Premier and blockaded Ontario Hydro construction vehicles. However, the lines were approved and built. Yet, BRHLC leaders say that they did have some victories—notably, the raised awareness of the EMF health and safety issue, locally, nationally, and internationally. They talked about the issue on national television news programs and spread word about it through handouts and, later, the Internet.

Two registered nurses in the group researched health effects through libraries. Two electric engineers in the group worked with Ontario Hydro technical officials to gather and assess information. Others established a community information network through Internet sites using the new National Freenet. There, people and groups exchanged information about transmission line issues, particularly about EMF. Eventually this was institutionalized in a Web site as the Bridlewood Electromagnetic Fields Information service. The service includes an

extensive bibliography on electromagnetic fields and health. Its hundreds of references include reports in major science journals and utility industry publications of research findings about the characteristics and effects of EMF and popular articles warning about the risks of exposure to EMF.

9.3 EMF Issues on the Internet

Although the BRHLC lost its fight against Ontario Hydro, a few participants continued to maintain the information service site and the bibliography. When the original administrator of the site resigned in February 2000, others maintained it. BRHLC also founded an open, unmoderated mailing list on Yahoo for the discussion of electromagnetic radiation. Participants around the world tell about the health effects of EMF from transmission lines and other sources, especially cell phone transmission masts; about their struggles against the siting of these facilities; and about technical details, such as regulations governing exposure to EMF. In 2000, the site's founder discontinued participation, but others, graduate engineers, carried on. They consider it important that "the EMF community" has this as a means of communication that belongs to it "collectively" and that is "open and unfiltered." Thus, even though a particular struggle against a transmission line ends in defeat, some participants continue to express their concerns about transmission line issues.

9.4 EMF and Transmission Line Opposition Networking, 1999–2003

The Bridlewood Electromagnetic Fields Information Service Web site also contains links to other groups opposing transmission lines, just as these groups link to Bridlewood. EMF effects provide one of the main concerns that these various groups can share. As mentioned, one of the linked groups is the Power Line Task Force in the southeastern suburbs of St. Paul, Minnesota, which from 1999 through at least 2003 challenged the project by Xcel Energy to upgrade its 115 kV transmission line running through their neighborhoods. The group leader, a financial planner with a doctorate in economics, established a Web site to publicize the group's fight, but also to spread word of EMF issues around the world. The group objects to the size of the towers of the Xcel project; however, its Web site chiefly attacks the effects of EMF. Xcel Energy insists that the upgrading will reduce exposure to these fields, but the protesters disagree. They cite studies, particularly one by the California Department of Health in 2001, that they interpret to show that EMF can cause childhood leukemia, adult brain cancer, Lou Gehrig's disease, and miscarriage. Xcel Energy disagrees.

The powerlinefacts Web site links to the Web site of the group Revolt, just as Revolt links to it. The Web site of Revolt critiques the many impacts of a 400 kV transmission system proposed in 1991 by the National Grid for North Yorkshire and in operation in 2003. Revolt's Health Issues pages links to some 20 sites on EMF subjects. Some offer published scientific assessments of health effects, such as a much-quoted report released in 1999 by the U.S. National Institutes of Environmental Health Sciences and reports of the International EMF Project of the World Health Organization. Other links are to media reports of these assessments, new research findings, scholarly discourse about EMF effects, and risk management theory and policy.

9.5 Scientific Studies of EMF and How They Are Interpreted

Revolt, Powerlinefacts, the Bridlewood Bibliography, and other opponent publications refer to the reports of scientific studies about EMF and other transmission line health effects. Since the l980s, there have been many such studies. Researchers from many disciplines, especially epidemiology, cellular biology, biostatistics, and risk assessment, have reported their findings in scholarly journals. Governments, health agencies, and research institutes at national, local, and international levels have convened panels of researchers to assess these findings. Utilities, their opponents, and siting authorities all refer to these studies to make their arguments or decisions.

Many of these studies were reviewed, summarized, and synthesized in 2002 by an Interagency Working Group in Minnesota and presented as a White Paper on Electric and Magnetic Field Policy and Mitigation Options to inform decision makers. This White Paper exemplifies many EMF studies. It distinguishes among findings from epidemiological studies, findings from laboratory studies, and the determination of a biological mechanism for how magnetic fields may cause effects, notably cancer (childhood leukemia and chronic lymphocytic leukemia in occupationally exposed adults). The paper concludes that epidemiological studies have mixed results. Some show no statistically significant association between exposure

to EMF and health effects. Some show a weak association. Newer laboratory studies fail to show such an association. Furthermore, the studies, with one notable exception, fail to establish a biological causal mechanism. Thus, most studies conclude that there is insufficient evidence to prove an association between EMF and health effects. However, many studies also conclude that there is insufficient evidence to prove that EMF exposure is safe.

These conclusions lend themselves to diverse interpretation and each major party in transmission line siting cases imposes its own meaning on them according to its interests and understandings. In essence, opponents say that these research studies and assessments show that the risks of EMF from power lines are significant enough to make government regulators deny approval of projects that site overhead transmission lines near human habitation or at least delay approval until more studies are conducted to determine risk. A weak association still means that there is risk and it is unjust to impose this risk involuntarily on people. Some also say that the applicant utility must prove that the transmission lines will not put people's health at risk from EMF before a project is given approval, rather than require that the opponents prove that there is a risk. Utilities say that the reports demonstrate that the risks from EMF are not sufficient to deny approval. They also disagree that the burden of proof of no risk be on them. Government siting authorities usually say that the reports show that EMF risks are not sufficient to deny approval. Some also recognized that if the burden of proof of no harm is shifted to the applicant, this could establish a new principle in regulatory policy, the precautionary principle. However, the siting authorities sometimes indicated that exposures would be mitigated if later scientific studies showed this was necessary. Siting authorities, legislatures, and international agencies continue to request such studies.

Meanwhile, scientists continue their studies, but their contributions seem likely to give the parties to the issue more to argue about. Findings and analysis reported by researchers at Bristol University in the United Kingdom provide an example. It will be recalled that scientists could not determine a biomechanical explanation of how EMF might produce health effects. However, on the basis of their research in the l990s, including 2000 experiments near pylons in 1999, Denis Henshaw and colleagues at Bristol hypothesized a biomechanism. They report that high-voltage AC transmission lines produce corona ions that charge particles (aerosols) from automobile exhaust and other pollutants in the atmosphere, near and downwind of the pylon. The deposition of these particles on the skin and in the lungs is increased, elevating the risk of diseases such as skin and lung cancer. Henshaw's propositions have been communicated to the public through the mass media. Revolt and other opponents of transmission lines mention the "Henshaw hypothesis" on their Web sites and the hypothesis has been used by opponents to bolster their testimony against line siting.

The media reports that scientists working for the electricity industry refuted the Henshaw results. The industry also notes in public announcements that the U.K. National Radiological Protection Board found an earlier small-scale study of the same phenomena by Henshaw to be flawed in method and theory and its results "implausible and highly speculative." In addition, industry and some government regulators point to the results of other studies that offer epidemiological and laboratory evidence that they say contradicts the implications of Henshaw, including studies for the National Radiological Protection Board (NRPB) by the renowned Sir Richard Doll.

In 2003 in the United States, researchers at the Electric Research Power Institute (EPRI) requested funds to assess the Henshaw Hypothesis. EPRI is the research arm of the electricity utilities industry. The debate about the hypothesis and the effects of EMF exposure is likely to continue. For example, Revolt reports that Henshaw's research provides further evidence that the hormone melatonin, involved in the body's circadian rhythms as well as in certain types of cancer, is also affected by magnetic fields. Revolt says uncertainties about the effects of EMF call for the use of the precautionary principle in siting transmission lines. Referring to positions taken by the European Court of Justice, Revolt says precaution means that in the face of such uncertainty, preventive measures may be taken without having to wait until the reality of the risks become apparent.

9.6 Perception and Communication of EMF Risks

Some analysts argue that the public is becoming unduly alarmed about the risks of exposure to EMF because they do not understand the meaning of statistical associations and probabilities as reported erroneously in the news. In an online review in the *British Medical Journal*, Simon Chapman, Professor of Public Health at University of Sydney, Australia,

discusses responses to news of a 2001 NRPB study on EMF risk. This study estimated that additional risk from power lines meant an extra case of childhood leukemia every 2 years: from approximately 1 in 20,000 to 1 in 10,000. Chapman writes that this is a very small risk, but that the media in Australia and the United Kingdom reported it under alarming headlines such as "Power lines double cancer risk." Chapman writes that because of such exaggerations of EMF effects, anxious Australians living near power lines have called cancer agencies asking whether they should sell their house or have their children tested. Others, such as John Farley in his Quackwatch Web site, attribute public fears about EMF to the exaggerations of quacks, who sell devices to detect EMF or medicines to cure it, or to the alarmist writings of journalist Paul Brodeur, who claims that government and industry are conspiring to hide the truth about EMF risks.

The American Physical Society (APS) also finds the risks of EMF effects exaggerated and not understood. In 1995, APS, the major professional association of physicists in the United States, published its official assessment of the effects of EMF on health. It stated that unsubstantiated claims that EMF are threats to health provoke fears among people and these fears lead to expensive mitigation efforts and, in some cases, to lengthy and divisive court proceedings.

As noted above, health agencies at the state, national, and international levels seek to communicate what is known about EMF effects in order to facilitate decision-making. However, D. G. MacGregor, P. Slovic, and M. G. Morgan suggest from their psychometric research on perception and communication of risks that even careful and scientific presentation of the uncertainties about EMF exposure will increase people's concerns and that people interpret new evidence according to naive conceptions about how EMF interferes with normal biology.

9.7 Is EMF Fear a Cause or an Effect of Opposition to Transmission Siting?

The Chapman and APS statements indicate that it is fear of EMF that causes people to oppose transmission lines. It will be recalled that utility consultants say that people oppose transmission lines because they are ugly or simply because they do not want them in their "backyard," but express concern about EMF and other health effects as their reason because health is politically more powerful and legitimate. Which of these suggestions about cause is correct? Probably neither. It is difficult to determine cause-and-effect relationships in complex social phenomena (just as it is difficult to determine these in complex biological phenomena). This article shows how concern about the effects of EMF is part of an array of concerns held and learned by transmission line opponents. It is impossible to know how much it is concern about EMF that induces people to oppose transmission siting or how much it is concern about other impacts that induces people to learn about EMF risks and present these as reasons for opposition. When people work with others to oppose transmission line siting, and do so in contest with proponents of the line, they engage in a process of cultural learning. They learn why and how to oppose. They develop a system of reasons and other ideas about why the transmission project is wrong and as they do they also develop an organization to express their ideas and fight to stop or alter the project. Some of them continue to work on EMF and other electricity and energy issues after their fight ends. It usually ends because the line project is approved. But the project is usually approved only after their resistance has long delayed it and altered it.

10. SUMMARY

Organized opposition, informed by key ideas or reasons, has been the most observable and significant response to efforts to site and construct electricity transmission lines in the United States, the United Kingdom, and Canada. This is probably also the case elsewhere in Western industrial democracies, that is, where authorities permit people to oppose projects important to industry and government. Opposition is one of three major obstacles to the siting and constructing of new transmission lines. Opposition affects and is affected by the other two obstacles, namely, the complexities of official processes of obtaining approval and the uncertainties of obtaining sufficient return on investment. People organize their opposition to a transmission project by forming a local group that interacts in communication networks with other groups opposing the project and with groups elsewhere holding related concerns. People develop many reasons for opposing the line and communicate these in many forums, especially in public hearings, Web sites, and newsletters. Their main reasons are these: The line is not needed

because there are alternatives to it in generation, transmission, and consumption. The line is too high in voltage and carries electricity to other states or regions. It will be used chiefly to profit its owners rather than to serve the public it impacts. It is ugly and invasive and represents a loss of control. Unfairness and injustice mark the way the line was approved and in how its impacts and the impacts of generating its sources of electricity are distributed. Building and operating the line damages agriculture, farm economy, and the environment. The line threatens the health and safety of people living along it, especially from its electric and magnetic fields. People develop these reasons into a system of related ideas as they argue with utilities and siting authorities. They are usually understood to act as interest or pressure groups. However, these groups might become parts of a movement of change as they work to change the culture of energy use as well as energy policy.

SEE ALSO THE FOLLOWING ARTICLES

Electromagnetic Fields, Health Impacts of • Environmental Injustices of Energy Facilities • Public Reaction to Energy, Overview • Public Reaction to Nuclear Power Siting and Disposal • Public Reaction to Offshore Oil • Public Reaction to Renewable Energy Sources and Systems

Further Reading

Caliri, L., Nixon, R., *et al*. (1997/1998). A Question of Power, The Roanoke Times, http://www.newrivervalley.com. [Series on the AEP 765 kV Case]

Casper, B. M., and Wellstone, P. D. (1981). "Powerline: The First Battle of America's Energy War." University of Massachusetts Press, Amherst, MA. [The CU ±400 kV Case].

Gerlach, L. P. Energy wars and social change. *In* "Predicting Sociocultural Change" (S. Abbot and J. van Willigan, Eds.), Southern Anthropological Society Proceedings No. 13. University of Georgia Press, Athens, GA. [The CU ±400 kV Case]

Gerlach, L. P. (1981). Protest movements and the construction of risk. *In* "The Social and Cultural Construction of Risk" (B. Johnson and V. Covello, Eds.), pp. 103–145. D. Reidel, Dordrecht, the Netherlands/Boston, MA.

Gerlach, L. P. (1999). The structure of social movements: Environmental activism and its opponents. *In* "Waves of Protest: Social Movements since the Sixties" (J. Freeman and V. Johnson, Eds.), pp. 85–97. Rowan and Littlefield, New York.

MacGregor, D. G., Slovic, P., and Morgan, M. G. (1994). Perception of risks from electromagnetic files: A psychometric evaluation of a risk-communication approach. *Risk Anal.* **14**, 815–828.

Minnesota State Interagency Working Group on EMF Issues. (2002). "A White Paper on Electric and Magnetic Field (EMF) Policy and Mitigation Options." http://www.health.state.mn.us/divs/eh/radiation/emf/.

Opposition Group Web Sites

Bridlewood Residents Hydro Line Committee (BRHLC): http://www.feb.se/Bridlewood/

Power Line Task Force: http://www.powerlinefacts.com

Rural England Versus Overhead Transmission Lines (Revolt): http://www.revolt.co.uk

Save Our Unique Lands (SOUL): http://www.wakeupwisconsin.com

Woodlands Coalition: http://www.woodlandscoalition.com

Public Reaction to Energy, Overview

ERIC R. A. N. SMITH
University of California, Santa Barbara
Santa Barbara, California, United States

1. **Definition and Importance of Public Opinion on Energy**
2. **The Public's Knowledge about Energy**
3. **The Public's Environmentalist but Conflicted Tendencies**
4. **Public Opinion on Energy and Demographic Characteristics**
5. **Public Opinion and Political Orientations**

Glossary

economic issues Political issues that relate to the distribution of wealth in society.
party identification A way in which people think of themselves; a psychological and emotional attachment to a political party.
social issues Political issues that relate to values and morals.

The term public opinion refers to the factual beliefs and preferences about issues of interest to the general population. Whether people believe that oil exists in the Arctic National Wildlife Refuge (a factual question) and whether people believe that the refuge should be opened to oil exploration (a preference) are typical examples of public opinion questions. Public opinion is normally measured using survey methods. Pollsters contact randomly selected samples of adults and ask them a series of questions to measure their opinions on current policy issues. The results of these surveys often play a major role in policy making.

1. DEFINITION AND IMPORTANCE OF PUBLIC OPINION ON ENERGY

What the public thinks about energy is important because public opinion is a major force influencing public policy on energy production and consumption. The fact that elected officials generally follow their constituents' wishes on high-profile issues is well documented by political scientists. Studies certainly show the influence of major industries, such as petroleum and nuclear power, but that influence generally occurs on low-profile issues. Most studies show that on major issues that receive a good deal of press attention, such as nuclear power and offshore oil development, the public generally gets what it wants.

Academic studies aside, observers can easily see the influence of public opinion in the cases of both nuclear power and offshore oil development. Public opposition to nuclear power, especially after the 1979 Three Mile Island accident, helped block the construction of new nuclear power plants. In the aftermath of the accident, public support for nuclear power dropped sharply, and every proposal for a new nuclear facility was met with mass protests. Public opposition, coupled with soaring construction costs, ended the growth of the nuclear power industry. No more nuclear power plants were ever ordered.

In the case of offshore oil drilling, the California public has moved through intermittent periods of quiet acceptance and intense opposition during the past century. The 1969 spill in the Santa Barbara Channel crystallized public opposition, and—some historians argue—helped launch the modern environmental movement. During the 1980s and 1990s, public opinion polls consistently showed overwhelming opposition to further drilling. In order to curry favor with California voters, President George H. W. Bush ordered a moratorium on new oil drilling leases in federal waters in 1990. President Clinton continued it in his administration, and President George W. Bush has allowed the policy to stand. In the case of offshore drilling along the Florida coast, public opinion against oil drilling was so intense that in June 2002 President Bush ordered the Interior

 © 2004 Elsevier Inc. All rights reserved.

Department to buy back some offshore oil leases that had been sold to oil companies and asked Congress to pass legislation to repurchase other leases. The official explanation for the buy-back from Department of Interior Secretary Gail Norton was that offshore oil development in Florida was opposed by "public opinion."

One can also see the influence of public opinion on policy makers in every energy crisis from that of 1973–1974 to the gasoline and electricity price spikes of 2000–2001. Every time shortages appeared and prices soared, the public demanded laws to lower energy prices. Politicians leapt to do the public's bidding, despite the fact that lowering energy prices encourages consumption, making the energy situation worse.

The initiative process gives voters a measure of direct control over public policy. In the 23 states and hundreds of counties and cities with the initiative process, voters can do more than demand that their elected officials pass the laws they want: Voters can pass the laws themselves. In these cases, public opinion controls energy policy. In California, for example, voters have passed both local and statewide initiatives limiting oil development. Although antinuclear power groups were not able to persuade California voters to pass an initiative to block nuclear power in the state in 1976, they came close enough to scare the state legislature into passing tough nuclear safety legislation that effectively blocked efforts to build more nuclear power plants. Moreover, the threat of an initiative vote against any controversial, proposed power source is clearly a deterrent to potential investors.

What the public wants does not always determine public policy. The influence of public opinion varies depending on a number of factors, including the attention given to the policy by the news media and the public, how much the public cares about the policy, expert opinion, and the interest groups arrayed for and against various policy choices. Nevertheless, in some circumstances, the public can get exactly what it wants—even against seemingly overwhelming opposition from the oil and gas, nuclear power, or electricity industries and their political allies. The public's policy preferences may not always be wise or well informed, but they can be influential.

2. THE PUBLIC'S KNOWLEDGE ABOUT ENERGY

One of the best known findings of public opinion research is that most of the public pays little attention to politics and, consequently, few people know much about politics or public policy. The list of facts not known by a majority of the public is stunning. What do the words "conservative" and "liberal" mean? What is the federal minimum wage? What majority is needed in the House and Senate to override a presidential veto? Who is the Speaker of the House of Representatives? Who is the Chief Justice of the U.S. Supreme Court? In national, representative surveys of adults, fewer than half the respondents offered correct answers to any of these questions.

These findings should not be exaggerated. The public is not ignorant or stupid. In fact, only a small percentage of the population knows virtually nothing about politics. Rather, most people do know something about politics, just not a great deal. The public's modest level of knowledge stems from the fact that few people are interested in politics. In areas that people do care about—their jobs, their communities, their hobbies, and their favorite movies or sports teams—people know a great deal. Politics and public policy just do not draw much attention or interest.

Given these findings, one should not expect the public to know much about energy policy. There are a few aspects of the energy situation that almost everyone should grasp—for example, that the United States imports oil from the Middle East, whether gas prices are rising or falling, wind mills can be used to generate electricity, or waste from nuclear power plants is radioactive and hazardous. However, most aspects of energy policy are complicated and difficult to understand without some specialized knowledge or training. How much would new oil drilling in the Arctic National Wildlife Refuge or along the Florida or California coasts add to national oil supplies? Can the United States ever become energy independent? What are the effects of deregulation of the wholesale electricity market on prices? What are the trade-offs between clean air and electricity production? How safe is nuclear power? What is global warming, what causes it, and what are the policy options for doing something about it? None of these questions can be answered without a fair amount of knowledge. Because most energy policy questions are complicated, we should not expect the general public to understand them.

Table I presents selected survey results illustrating the public's lack of knowledge about energy matters. Questions 1 and 2 show that despite the Arab–Israeli wars, the gas lines, and the sharp gasoline price hikes, in the 1970s only approximately half the public realized that the United States had to import oil to meet its energy needs. The failure to realize this

TABLE I

Selected Knowledge Questions, U.S. Samples

Question	Year: % correct
1. Do you think the United States has to import oil to meet our needs, or do we produce as much oil as we need?	1977: 48 1979: 51 1980: 63 1991: 50
2. Using the card, at the present time, how much of the oil used in the United States would you say we have to import from other countries? (Five choices)	1977: 29 1978: 30 1981: 27
3. About what proportion of its oil would you say the United States imports from other countries? Would you say two-thirds or more, about half, about one-third, or less than one-third?	1984: 33 1990: 49
4. There is a debate about whether this country should become less dependent on oil imported from foreign countries. Please give your best estimate of which phrase comes closest to describing how much of our oil comes from foreign countries—about a fourth, about half, about three-fourths.	2002: 34
5. Here is a list of different companies. All of them have operations here in the United States. But, would you go down that list and for each one would you tell me whether, to your knowledge, the main headquarters is located here in the United States or located in a foreign country?…	1978: 78 1986: 76
5a. Exxon?	1978: 17
5b. Shell?	1986: 19
6. From what you've heard or read, do you think a nuclear power plant accident could cause an atomic explosion with a mushroom-shaped cloud like the one at Hiroshima.	1979: 33
7. To your knowledge, what percentage of the nation's electric power is currently supplied by nuclear power plants?	1979: 5 1986: 6
8. In 1979, about 20 years ago, there was a nuclear power plant breakdown near Harrisburg, Pennsylvania. Would you happen to remember the name of that nuclear plant, or not?	1999: 38

Sources. Items 1–3 and 5–7 are from Delli Carpini, M.X., and Keeter, S. (1996). "What Americans Know about Politics," Appendix 3; Yale Univ. Press, New Haven; CT. Item 4 is from the Associated Press poll, February 22–27, 2002 ($N = 1016$). Item 8 is from the Gallup poll, March 19–21, 1999 ($N = 1018$). Both Items 4 and 8 are reported by Polling Report.Com (http://www.PollingReport.com/enviro.htm). Note: Delli Carpini and Keeter do not identify the surveys or sample sizes for the surveys, but all surveys were conducted by major reputable firms and are archived by the Roper Center for Public Opinion Research.

fundamental fact about the U.S. energy situation during the 1970s may seem astonishing, but it is consistent with other findings about the level of the public's knowledge. Large blocks of information are missing, and only a relatively small number of people have much detailed knowledge about any public policy issue. The other results in Table I add to the picture of the public's limited knowledge about energy policy.

Questions 3 and 4 show that the public lacks knowledge about our dependence on foreign oil continues. Question 5 shows that despite the political rhetoric about foreign and multinational oil companies, only a small portion of the public can identify them.

The next question about the possibility that a nuclear power plant could explode in a mushroom-shaped cloud is particularly revealing. The question was asked soon after the Three Mile Island accident. Even though television, newspapers, and news magazines reported extensively on what might happen during the Three Mile Island nuclear power plant accident, only one-third of the public knew that a "mushroom-shaped cloud" was not possible. The Jack Lemon/Michael Douglas/Jane Fonda movie, "The China Syndrome," which premiered just before the Three Mile Island accident and offered a clear description of what might happen in the event of a nuclear power plant accident, may have been a critical and box office success, but an educational success it was not.

Table II presents results from a survey of Californians. The first question in Table II asks how long the world's oil supply will last. This is a difficult question because the topic has not been the subject of many news reports since the 1970s and because experts disagree. It has been the subject of some news coverage, however. Coincidentally, it was the subject of a *Scientific American* article published only weeks before the survey. Despite this, half the respondents admitted they did not know the answer and declined to guess. The answers from the other half ranged from less than 5 to more than 1000

TABLE II

Selected Knowledge Questions, California Sample

Question	% correct
1. As you know, the amount of oil in the world is limited. Do you know roughly how many years it will be before experts believe the world will run out of oil? [Probe if necessary]: Just your best estimate.	23
2. Do you happen to know whether it is currently safer to transport oil using oil tankers or using oil pipelines? By safer, I mean the way which is least likely to result in a major oil spill.	58
3. How often do you think a typical offshore oil platform along the California coast is likely to have a major oil spill—once every 5 years, once every 10 years, once 20 years, once every 40 years, or less frequently than this?	15
4. When a major oil spill occurs, how much threat does it pose to human life—a great deal, some, only a little, or no risk at all?	6

Source. California Offshore Oil Drilling and Energy Policy Survey, March 1998, conducted by the Field Institute ($N = 810$).

years. A reasonable estimate is somewhere between 50 and 100 years. Taking any number in this range as correct, we see that only 23% of our sample could make a reasonable guess.

The next three questions in Table II differ from the previous questions because they not only measure factual knowledge but also the public's perceptions of risk. In the abstract, this difference may sound trivial, but in practice people who feel threatened respond by demanding that politicians do something about the risks. In the case of oil, those demands are to limit or end offshore oil development.

The first risk question asks respondents whether it is safer to transport oil by using oil tankers or pipelines. The answer, pipelines, should be relatively well-known because most of the world's catastrophic oil spills have been tanker disasters, including such well-known ships as the Castillo de Bellver, the Amoco Cadiz, the Torrey Canyon, and the Exxon Valdez. Because tanker spills have received so much news coverage, one might expect most of the public to know this answer. The survey, however, showed that only 58% of the California respondents correctly identified pipelines as the safest way to transport oil. That is only slightly higher than the 50% who would have gotten the answer correct if they had guessed based on coin tosses.

The second risk question asks respondents to estimate how frequently a typical offshore oil platform will have a major oil spill. The public's answers to this question reveal seriously exaggerated fears. Twenty-three percent of the California public said that they expected a major oil spill every 5 years from a typical platform. An additional 25% said that they expected such a spill once every 10 years. Although the exact meaning of "major oil spill" is not stated in the question, by any definition major oil spills from a typical platform are far less frequent than once every 10 years. If we were to use 1000 barrels of oil as the standard for a major oil spill, we would have to conclude that the correct answer—given by only 15% of the sample—was "less than once every 40 years" because there were only 11 spills of that size from all U.S. offshore platforms from 1964 to 1992 (and none since 1980). Even using a far smaller number of barrels as the standard would still yield less than once every 40 years as the correct answer because of the huge number of offshore oil platforms in operation in U.S. waters (more than 3800 platforms in 1990). This exaggeration of risk, or lack of knowledge, presumably contributes to the public's opposition to further offshore oil drilling.

The answers to the next question about oil spills also help us understand people's fears. The question asks how much threat a major oil spill poses to human life. Again, we see that the public seriously exaggerates the threat posed by oil. Thirty-one percent of the statewide sample responded that a major oil spill would pose a serious threat to human health. An additional 36% believed that it would pose some threat. In fact, oil spills pose only a negligible threat to human life. They certainly pose a substantial threat to many other kinds of life but not to human beings.

The public's low level of knowledge about energy policy is important because of the public's influence on policy makers. When the public forms opinions and makes demands on politicians based on misinformation, public policy may suffer. As the data in this section show, the public is often misinformed about even the most basic aspects of energy policy.

3. THE PUBLIC'S ENVIRONMENTALIST BUT CONFLICTED TENDENCIES

When the public does turn its attention to energy issues, it tends to prefer environmentalist policies. That is, the public generally favors conservation;

favors alternative, sustainable energy sources (wind and solar power); and opposes conventional energy development (oil, nuclear power, and coal). However, the public also displays a clear preference for government action to maintain low energy prices, despite the fact that low prices lead to excessive and wasteful energy consumption. These inconsistent preferences seem to result from a mix of people's misunderstandings about the energy situation and their distrust of the energy industry.

Evidence for the environmentalist tendencies of the public can be seen in a variety of survey results. Figure 1 shows the percentages of people who said in response to a series of Gallup polls that they preferred more conservation or more production as a solution to U.S. energy problems. The solid line represents the percentage indicating that they preferred conservation, and the dotted line represents the percentage indicating that they preferred production. Even at the height of the gasoline price hikes in 2001, the public preferred conservation by a substantial margin. By March 2002, when prices had stabilized and the public was becoming accustomed to higher prices, conservation was preferred by a margin of 60 to 30%. (The remaining 10%, not shown in the figure, either responded "both" or "neither.") Figure 2 shows the results of a related question in which respondents were asked in March 2001 and March 2002 whether they would give priority to development of energy supplies or protection of the environment. The environment won by a 52 to 40% majority. Figure 3 presents another trade-off. Respondents were asked whether they believed "protection of the environment should be given priority, even at the risk of curbing economic growth, or economic growth should be given priority, even if the environment suffers to some extent." By wide and consistent margins from 1984 to 2000, the polls showed a strong public preference for environmental protection. Support for environmental protection ranged from 61 to 71%, whereas the number giving priority to economic growth ranged from only 19 to 32% and was only 28% when the question was last asked in April 2000.

On more specific policy questions, polls show that the public strongly supports investment in environmentally friendly energy sources and generally opposes the development of any conventional energy

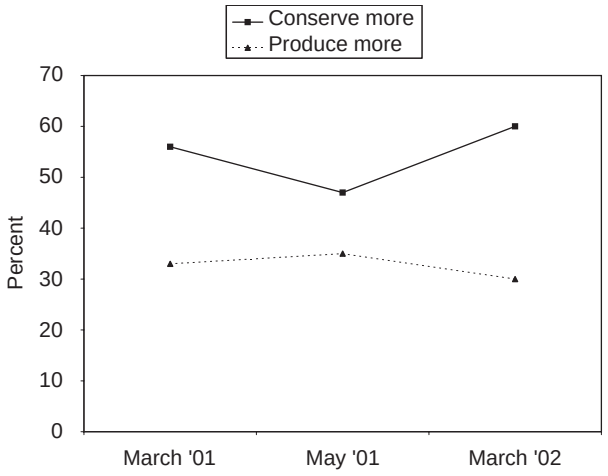


FIGURE 1 Energy solution: More conservation or more production. Source: Gallup polls.

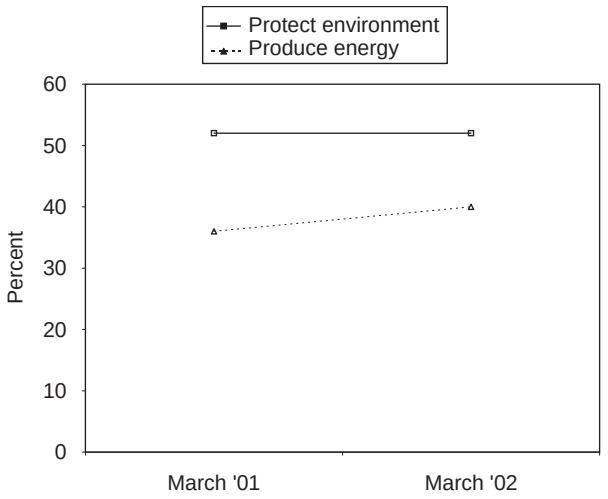


FIGURE 2 Environmental protection vs energy production. Source: Gallup polls.

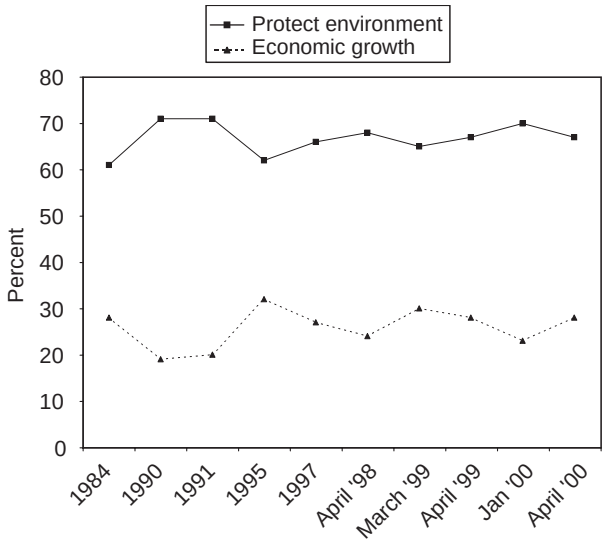


FIGURE 3 Environmental protection vs economic growth. Source: Gallup polls.

sources that are criticized by environmental groups. Table III shows the public's response to several policy proposals for dealing with U.S. energy problems. The public is nearly unanimous in calling for more support for solar, wind, and fuel-cell technology. Requiring automobile companies to build more fuel-efficient cars is also strongly supported. The proposals that fail to win majority backing are drilling for oil in the Arctic National Wildlife Refuge and nuclear power.

Further questions in other polls show majorities opposing drilling for oil and natural gas along the California coast or weakening environmental protections, such as automobile or industrial emissions, in order to increase energy production.

Despite evidence that a substantial majority of the public holds environmentalist views, there are signs that the public does not consistently follow through on those stated preferences. The public supports conservation in principle but opposes a variety of conservation policies because they place burdens on consumers. Simply stated, the public wants the benefits of cheap, plentiful energy but does not want to pay the costs.

The best behavioral evidence of the public's conflicted views can be seen in sales figures for sport utility vehicles (SUVs). According to a March 2000 Gallup poll, approximately one-fourth of U.S. households own at least one SUV. Moreover, in 2000—a year of sharply rising gasoline prices—nearly half of all new car sales were gas-guzzling SUVs, minivans, or light trucks. Thus, although the public says that it wants fuel-efficient cars, it actually buys fuel-inefficient cars (and trucks and SUVs).

Survey data also reveal the public's desire for low-cost energy. During the spring of 2000, when oil prices surged, a Gallup poll found that 80% of the public favored lowering state fuel oil taxes, 75% favored reducing federal gasoline taxes, and 59% favored drawing oil from the U.S. Strategic Petroleum Reserve to put more oil on the market and cut prices. In a similar vein, when asked whether they favored or opposed the proposal of "giving tax breaks to provide incentives for drilling for more oil and gas in the U.S.," Gallup poll respondents in March 2001 favored the idea by a 53 to 43% margin. In addition, the proposal to set "legal limits on the amount of energy which average consumers can use" was solidly opposed by a 62 to 35% margin. The 2000–2001 electricity crisis in California provides further evidence of the public's desire for cheap energy. A May 2001 Field Institute poll asked Californians, "Is it a good thing or a bad thing for the federal government to cap energy prices?" Seventy percent said it was a good thing.

Most economists tell us that capping or subsidizing energy prices during shortages is foolish. If prices are allowed to rise freely, consumers will use less energy, and the energy problems will be resolved by free market forces. In contrast, if prices are capped or if energy consumption is subsidized, more energy will be used and shortages will worsen—potentially leading to lines at gasoline stations, electricity blackouts, or other problems.

Unlike economists, most of the public likes energy subsidies and dislikes free market solutions to energy shortages, but this does not mean that people are rejecting free market economics. Instead, many people believe that energy shortages are faked by energy companies to increase prices and profits. A 1974 Roper poll showed this when it asked, "Some people say there is a real shortage of gasoline and fuel oil because demand has outrun supply. Others say there really isn't a shortage of gasoline and fuel oil and the big companies are holding it back for their own advantage. What do you think—that there is or is not a real shortage of gasoline and oil?" Seventy-three percent said there was no real shortage. During the next energy crisis in March 1979, the CBS/*New York Times* poll asked, "Do you think the shortage of oil we hear about is real or are we just being told there are shortages so oil companies can charge higher prices?" Sixty-nine percent responded that they were just being told that shortages existed, and another 11% said they were not sure. Only 20% believed the shortages existed. Similar survey findings were reported during the gasoline price hikes immediately before the 1991 Persian Gulf War, during the 2000–2001 gasoline price hikes, and during the 2001 California electricity crisis.

TABLE III

Public Support for Energy Alternatives

Alternative	Favor (%)	Oppose (%)
1. Investments in new sources of energy such as solar, wind, and fuel cells	91	8
2. Mandating more energy-efficient cars	77	20
3. Opening up the Arctic National Wildlife Refuge in Alaska for oil exploration	44	51
4. Increasing the use of nuclear power as a major source of power	42	51

Source. Gallup poll, November 8–11, 2001.

Apparently, the public's reasoning is that the energy industries cause supply shortages and high prices by manipulating energy markets. In other words, high prices are the result of illegal or unethical activity. Given this situation, people can easily jump to the conclusion that the government should step in and fix the problem with price caps, subsidies, or other measures.

Here, the public's lack of knowledge about energy policy plays a role. In every energy crisis the United States has faced, politicians and activists have accused energy companies of manipulating prices. Certainly in the case of the oil crises, when people who do not even know that the United States needs to import oil to meet its energy needs hear those charges, they will be especially likely to believe them. More broadly, a mix of people's misunderstandings about the energy situation and their distrust of the energy industry may lead to poor judgments about the best energy policies for the government to adopt.

Public opposition to increases in energy prices affects more than government behavior during energy crises. Proposals for a carbon or Btu tax, such as the one offered by President Clinton in 1993, face a skeptical public. The idea behind Btu taxes is to increase the price of fossil fuels, causing consumption to decline and consumers to switch to using other fuel alternatives. Such proposals face two problems in winning public support. First, the proposals are complicated and require at least a moderate understanding of both economics and energy. Second, the proposals call for raising energy prices. These are both serious obstacles to Btu taxes.

In summary, the public generally favors environmentalist policies such as conservation and the development of alternative, sustainable power sources. However, the public also wants the government to take steps to guarantee low, stable energy prices. The contradictions among these views remain unresolved.

4. PUBLIC OPINION ON ENERGY AND DEMOGRAPHIC CHARACTERISTICS

To understand the sources of people's opinions about energy policy, it helps to examine the larger context of public opinion across a range of issues. Public opinion scholars have found that the best way to describe public opinion on domestic issues is to divide issues into two broad categories—economic and social. Economic issues are those that relate to the distribution of wealth in society—the tax system, Social Security, Medicare, welfare, regulation of businesses and unions, etc. Social issues are those that relate to values and morals—abortion, birth control, free speech, civil rights, women's rights, sexual preferences, etc. This distinction is useful for two reasons. First, people tend to have similar opinions on the two types of issues. That is, if someone is liberal on one social issue, he or she is likely to be liberal on others. Second, the two types of issues have different causes and demographic correlates. Opinions on economic issues generally vary with income and self-interest, whereas opinions on social issues usually vary with education and age.

At first glance, energy policy would seem to be a classic economic issue. Energy production is a huge economic sector. Most disputes over energy policy involve the government regulating private businesses. Regulations affect company profits, create new jobs for some people, and eliminate existing jobs held by other people. From pollution controls to protections for endangered species, regulations on the energy industry have substantial economic impacts.

In some respects, attitudes toward energy policy do respond to events in ways that are consistent with people looking at energy policy as an economic policy. An excellent example of this is the way public support for offshore oil drilling along the California coast parallels the inflation-adjusted price of gasoline. Figure 4 shows a fairly close fit between support for offshore oil drilling and the real price of gasoline. When the price of gasoline increased in the late 1970s, so did support for offshore oil drilling. When the price of gasoline decreased in the 1980s, support for offshore oil declined as well. The 1989 survey shows a sharp decline in pro-development feelings that is not matched by gasoline prices, but it does

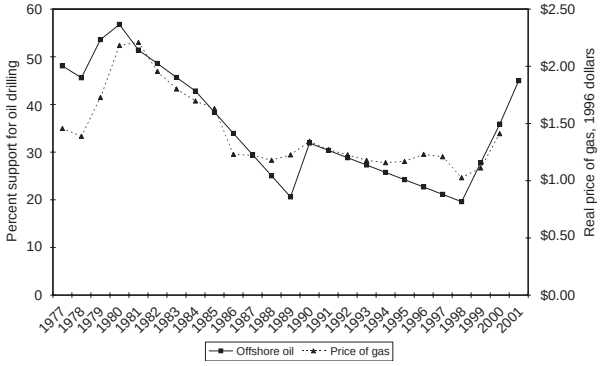


FIGURE 4 Support for oil drilling and the price of gasoline.

correspond to the Exxon Valdez oil spill. The post-1998 increase in support again parallels the increase in gasoline prices.

There is also some individual-level evidence that people think of energy issues in economic terms. For example, as the distance people commute to work increases, they become more likely to favor offshore oil drilling and other steps to increase oil production and cut prices. That is, economic self-interest affects attitudes toward energy policy in some cases.

However, the classic patterns of economic issues do not appear. Opinions on energy issues generally do not vary with income, occupation, or other measures of social class. A few studies have found that attitudes on energy policy vary with income, but these studies are outnumbered by the studies that found no relationships.

For the most part, attitudes toward environmental issues, including energy issues, behave like attitudes toward most social issues. Two key indicators of this pattern are the relationships that energy attitudes have with education and age. On virtually every social issue, the most liberal people in the population are the well educated and the young. The basis for the education–social issue relationship seems to be that education brings people into contact with new ideas and new types of people and broadens the perspectives of the well educated. This process develops a greater tolerance for people who are different—blacks, the disabled, gays and lesbians—and a greater appreciation of ideas such as the value of the environment.

Age is also related to opinions on social issues, but not because age causes beliefs or opinions. Instead, age is related to opinions on social issues because age reflects the generations in which people grew up and were socialized. Throughout the past century, the United States has been moving in a more socially liberal direction. Ideas such as racial and gender equality, which were once considered radical, are now generally accepted. Ideas such as gay rights, which were once never even discussed, are now on the leading edge of social change. Similarly, the modern environmental movement began in the 1960s. Therefore, people who grew up and were socialized before the 1960s are less likely to hold environmentalist views than those raised during the environmental era.

The relationship between education and the trade-off between protecting the environment or finding new sources of oil and gas is typical of relationships between education and questions about energy policy. Figure 5 presents the results of an April 2001 *Washington Post* poll, which showed that 51% of Americans with a high school education or less said that environmental protection was more important than finding new oil sources, whereas 58% of college graduates held that view. The better educated the respondents were, the more they favored protecting the environment.

The relationship between age and the environment–energy sources trade-off in the *Washington Post* poll is even stronger. As Fig. 6 shows, 66% of the 18- to 30-year-old respondents favored protecting the environment, whereas only 38% of those 61 or older did so—a 28% difference across the

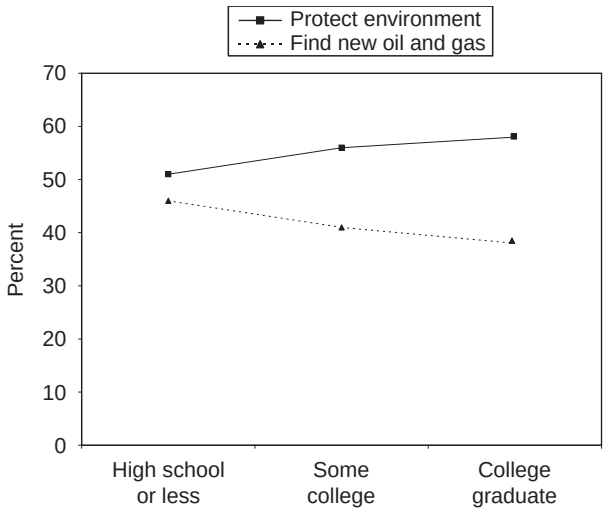


FIGURE 5 Energy vs environmental protection by education. Source: *Washington Post* poll, April 2001.

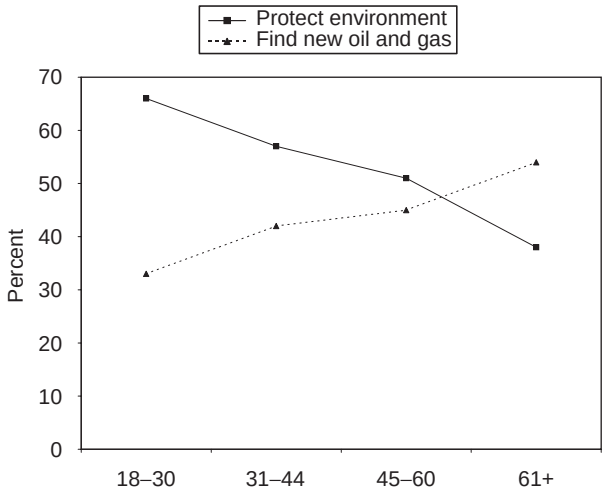


FIGURE 6 Energy vs environmental protection by age. Source: *Washington Post* poll, April 2001.

generations. This finding is typical in that age has been found to be the demographic characteristic yielding the strongest, most consistent differences on environmental issues.

Other demographic variables are also related to opinions on environmental and energy-related issues, although not as strongly or consistently as education and age. Gender is usually related to attitudes toward energy issues, but only if the issue entails some aspect of risk to people. When risk is involved, women tend to take the pro-environmental stand. For example, women tend to oppose nuclear power more than men do because many people in the public regard nuclear power as risky. In contrast, men and women usually hold similar opinions about offshore oil drilling because it is not seen as inherently risky by the public.

Race, ethnicity, and whether one lives in a rural or urban area have also been found to relate to environmental and energy issues in a number of studies, but the relationships do not consistently fall in a single direction. That is, in some studies blacks have been found to hold more pro-environmental views than whites; in other studies, the reverse pattern has been found. Studies of urban and rural dwellers yield similar results. The relationships often can be seen to depend on the exact issue in question. Location of hazardous waste dumps, for example, is of more concern to people living near proposed sites than those living at great distance. The demographic characteristics of those living near the dumps, therefore, may determine what studies find. Consequently, broad generalizations are difficult to make.

5. PUBLIC OPINION AND POLITICAL ORIENTATIONS

A second way to explain public opinion on energy issues is to examine people's political orientations and worldviews. By far the most important political orientation is party identification. Party identification is a way in which people think of themselves—a psychological and emotional attachment to a political party. People identify with political parties in the same way they identify with ethnicities or religions. Native-born Americans typically acquire a party identification from their parents by age 9 or 10—long before they understand anything about the substance of politics.

The importance of party identification stems from the way in which it influences people's opinions and behavior. Party identification influences people's opinions because it guides their decisions about what to believe or disbelieve when they listen to the news or political discussions among friends. People who think of themselves as Democrats, for example, are far more likely than others to believe Democratic politicians and doubt Republican politicians. Similarly, Republicans trust fellow Republicans but not Democrats. As a result, both groups' view of reality is filtered through their partisan beliefs. In addition, party identification influences people's values. When a politician of one's political party says that preserving the Alaskan wilderness is vital, or that America desperately needs Alaskan oil, one's opinion is likely to be swayed. As a result of differing beliefs about the facts and differing opinions, the behavior of Democrats and Republicans differs as well. In voting and in a wide range of political activities, Democrats and Republican normally choose opposing sides.

In the world of politics, the Republican Party generally represents business interests and the wealthier segments of society, whereas the Democratic Party represents the working-class and lower income segments. Because environmental regulations are constraints on businesses, Republicans have usually been far less supportive of them than have Democrats. As a result, Republicans have generally leaned in a pro-development direction, whereas Democrats have generally leaned in a pro-environment direction.

These partisan patterns can easily be seen both in national political disputes and in public opinion data on energy issues. The fight over whether to open the Arctic National Wildlife Refuge for oil exploration and drilling is a good example. In the 2000 presidential election, the issue that drew the most public attention was the rising price of gasoline. Other issues garnered more press coverage, but drivers were reminded of the high prices every time they filled their cars' gas tanks. As a result, public interest remained high throughout the year. Both presidential candidates responded to the public's concerns. Governor George W. Bush, a former Texas oil man, declared that the way to lower prices was to produce more and, in particular, to open the Arctic National Wildlife Refuge in Alaska for oil drilling. Vice President Al Gore countered that the way to cut gasoline prices was to conserve energy and to pressure major oil-exporting countries such as Saudi Arabia to produce more oil for the world market. He stoutly opposed relaxing environmental standards to make it easier to drill for oil in the United States, and he insisted that the Arctic National Wildlife Refuge be kept closed to oil exploration.

The partisan division among the two parties' presidential candidates is mirrored in public opinion

data. Figure 7 shows the partisan breakdown of support for the proposal to deal with the energy situation by "oil exploration in the Alaskan Arctic Wildlife Refuge." Republicans supported the idea by a 54 to 43% margin, but Democrats opposed it by a 65 to 27% margin. People who did not identify with either political party—that is, those who call themselves independents—fall between Republicans and Democrats in their policy views.

Similar partisan differences exist across many energy-related and environmental issues. The same May 2001 Gallup poll found sharp differences regarding whether to invest more money in natural gas pipelines (favored by 76% of Republicans but only 54% of Democrats), whether to drill for natural gas on federal land (favored by 75% of Republicans but only 55% of Democrats), and whether to increase the use of nuclear power (favored by 59% of Republicans but only 45% of Democrats). Other national and state polls routinely find similar partisan differences, with Republicans typically comprising the group that prefers conventional energy development and Democrats comprising the group that prefers conservation and alternative energy sources.

Another political orientation that helps explain public opinion on energy and environmental issues is ideological self-identification—that is, whether people consider themselves to be liberal, moderate, or conservative. Ideological self-identification is a curious characteristic. Pollsters ask respondents a question such as "Generally speaking, in politics do you consider yourself as conservative, liberal, middle-of-the-road, or don't you think of yourself in those terms?" The terms liberal and conservative are in wide use in public debate. However, repeated studies for more than 50 years show that only approximately half the public can offer even crude definitions of what the labels mean. Typically, approximately one-third of respondents say "I don't know" when asked for a definition. Another one-sixth will offer definitions that are not even remotely recognizable. Nevertheless, the ideological labels that people give themselves predict their opinions relatively well, including their opinions on many energy issues.

Environmental problems emerged as political issues relatively recently in comparison with traditional economic conflicts over labor laws, welfare systems, and tax codes. Nevertheless, when they did begin to draw the public's attention in the 1960s, liberal leaders argued for laws imposing clean air standards, clean water standards, and other pro-environmental regulations. Conservative leaders argued against allowing the government to impose new regulations on businesses. Environmentalism quickly became an established part of liberalism, whereas pro-development views became part of the conservative agenda. This liberal–conservative distinction on environmental issues spread quickly among both political leaders and the public.

Figure 8 shows the ideological divisions over the dispute about offshore oil and gas drilling along the California coast. Seventy-seven percent of the people who say they are "strong liberals" oppose more drilling, whereas only 34% of strong conservatives oppose it. On the pro-development side, the figures

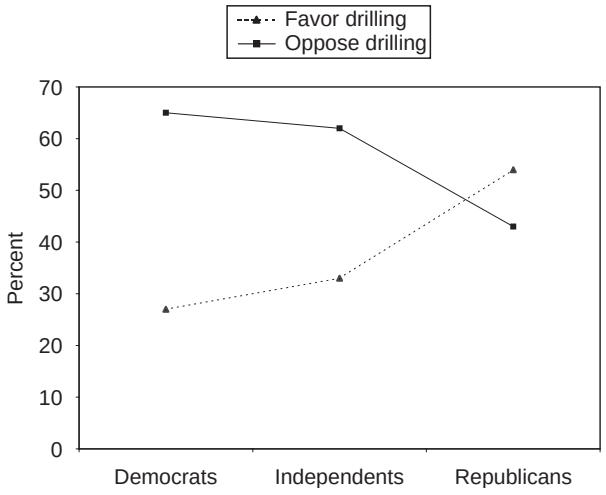


FIGURE 7 Support for oil drilling in the Arctic National Wildlife Refuge by party. Source: Gallup poll, May 2001.

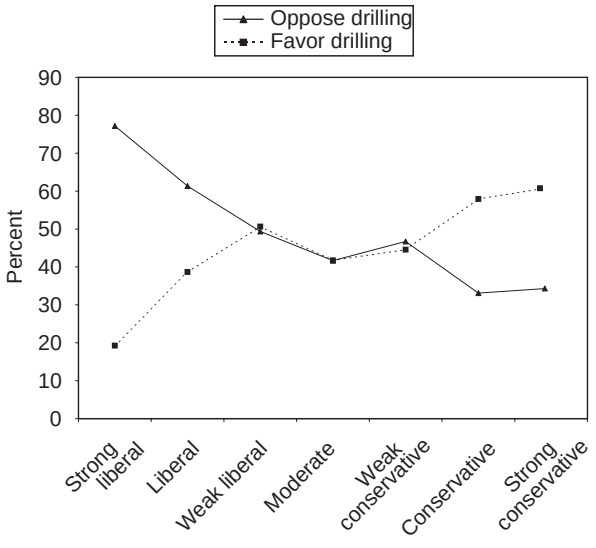


FIGURE 8 Californians support for offshore oil drilling by ideology. Source: Field Institute poll, May 2001.

are reversed, with 61% of strong conservatives but only 19% of strong liberals favoring more drilling.

Taken together, party identification and ideological self-identification do a much better job of explaining public opinion on energy issues than do demographic variables. Nevertheless, opinions on environmental and energy-related issues are not understood or explained as well by public opinion scholars as opinions in other areas (e.g., welfare and civil rights). Given the importance of public opinion to policy makers, research in this area continues.

SEE ALSO THE FOLLOWING ARTICLES

Energy Development on Public Land in the United States • Hydropower Resettlement Projects, Socioeconomic Impacts of • Lifestyles and Energy • Media Portrayals of Energy • Public Reaction to Electricity Transmission Lines • Public Reaction to Nuclear Power Siting and Disposal • Public Reaction to Offshore Oil • Public Reaction to Renewable Energy Sources and Systems

Further Reading

Dunlap, R. E., and Mertig, A. G. (eds.). (1992). "American Environmentalism: The U.S. Environmental Movement, 1970–1990." Taylor & Francis, Philadelphia.

Freudenburg, W. R., and Rosa, E. A. (eds.). (1984). "Public Reactions to Nuclear Power: Are There Critical Masses?" Westview, Boulder, CO.

"The Gallup Poll Monthly." Gallup Organization, Princeton, NJ.

Jones, R. E., and Dunlap, R. E. (1992). The social bases of environmental concern: Have they changed over time. *Rural Sociol.* **57**, 28–47.

Kempton, W., Boster, J. S., and Hartley, J. A. (1995). "Environmental Values and American Culture." MIT Press, Cambridge, MA.

Richman, A. (1979). The polls: Public attitudes towards the energy crisis. *Public Opin. Q.* **43**, 576–585.

Smith, E. R. A. N. (2002). "Energy, the Environment, and Public Opinion." Rowman & Littlefield, Boulder, CO.

Public Reaction to Nuclear Power Siting and Disposal

EUGENE A. ROSA and JAMES RICE
Washington State University
Pullman, Washington, United States

1. **Background**
2. **Nuclear Generation**
3. **Evolving Contexts**
4. **Public Reactions to Nuclear Power in the United States**
5. **Environmental Consciousness and Nuclear's Fortunes**
6. **The Present and the Future**
7. **Nuclear Imagery**
8. **Nuclear Waste**
9. **Conclusion**

Glossary

energy The capacity for doing work. Energy may be stored, and it may be transferred from one form into another. Forms of energy include electrical, chemical, thermal, and mechanical.

global warming Atmospheric gases such carbon dioxide and other greenhouse gases make the earth warmer than would direct sunlight alone by trapping heat from the sun as it is reradiated off the earth's surface and absorbed in the atmosphere rather than escaping into space.

greenhouse gases Atmospheric gases that trap the heat of the sun in the earth's atmosphere. Carbon dioxide, methane, nitrogen oxides, and ozone are common examples.

nuclear energy Power obtained by splitting (fission) or joining (fusion) atoms of radioactive materials, such as uranium.

nuclear waste The materials, both liquid and solid, left over from the production of nuclear energy. Nuclear waste exhibits varying amounts of radiation and can be harmful to humans.

power The rate at which energy flows.

radiation The flow of energy via electromagnetic waves. Radiation is produced by radioactivity.

radioactivity Produced when nuclei are undergoing transformation and emitting radiation.

reactor A device in which a controlled nuclear chain reaction can be maintained to produce heat. The heat is then typically used to produce steam, which in turn is used to run conventional turbine generators to produce electricity.

uranium A radioactive element found in ore. Uranium atoms can be split to create energy.

The harnessing of the atom, opening the way for the generation of electricity with nuclear energy, was met with a technological enthusiasm perhaps unmatched in modern history. Nuclear energy was a technology with unprecedented potential. Early in its development, nuclear power was claimed to be the cheapest form of electricity and was projected to provide a substantial portion of the electricity needs in the advanced nations of the world. History, however, conspired to stall the march of nuclear power with the result that these projections were never realized despite the technology's continuing potential. Standing in the way of its destined path to triumph, a variety of unanticipated technological, social, and political obstacles cloud the technology's future nearly everywhere. This article examines the key social obstacles leading to the stalemating of nuclear power.

1. BACKGROUND

Although born of war, nuclear fission technology was soon converted to the "peaceful" generation of electricity. The emergence of nuclear power as an energy source and the high hopes it engendered were hardly surprising in an evolutionary context. Passing through stages from early to mature industrialization was synonymous with a substitution process—the substitution of more readily abundant, less expensive, energy potent fuels for dwindling, more expensive, less potent fuels. With the appearance of fission technology, the generation of commercially produced electricity seemed the rightful heir to propel this evolutionary process forward. After all,

 © 2004 Elsevier Inc. All rights reserved.

the technology could convert puny amounts of raw fuel into usable energy so vast that it would be "too cheap to meter," displacing the reliance on fossil fuel electricity in the process. However, standing in the way of nuclear power's destined path to success were a wide variety of unanticipated technological as well as social and political obstacles that cloud the technology's future nearly everywhere.

The tension between the technology's potential and its public acceptance was captured in the most repeated quote in nuclear history–by Alvin Weinberg, the doyen of nuclear power:

> *We nuclear people have made a Faustian bargain with society. On the one hand, we offer—in the catalytic burner—an inexhaustible source of energy. But the price that we demand of society for this magical energy source is both vigilance and a longevity of our social institutions that we are quite unaccustomed to.*

Another price demanded of any centralized technology promoted by the national state is acceptance by its citizens. Public reactions to nuclear technology in the United States and elsewhere produced a checkered history for the technology that evolved in five overlapping phases: an era of initial enthusiasm, a rejection at the local level, a rise in general public concern and the assertion of alternative technologies, a period of ambivalence, and the current era of persistent pragmatism. Earlier phases prefigured later ones. Therefore, the evolution of public opinion about nuclear power in the United States can be understood as the dynamic and historically contingent intersection between the actions of the federal government, private utilities, industrial contractors, social movement organizations, and the general public in relation to this complex technological system perceived as promising but risky.

The decade of the 1960s, the earliest phase of nuclear power development in the United States, was unquestionably one of unbridled optimism, based, as it was, on the promise of nearly unlimited supplies of cheap electricity. During this period the federal government crystallized its roles as both promoter (with massive investments in R&D and industry subsidies that may total more than $1 trillion) and as regulator of the technology—roles it continues to fulfill today. The eventual recognition of the inherent conflict of interest in the two roles led to their separation in 1974 by the creation of the Energy Research and Development Administration to fulfill the promotion role (in 1977 to be absorbed into the newly created Department of Energy) and the Nuclear Regulatory Commission to fulfill the regulatory role. During this optimistic phase, the nuclear industry was most vigorous in its promotion of the technology. For example, beginning in 1963, plants were offered to utilities for the first time on a strictly commercial basis with so-called "turnkey" contracts, so named because the purchasing utility was guaranteed that the plant would be completed at a fixed price and needed to do little more than "turn the key" to begin operations.

As the initial developer of nuclear technology, the United States became the world leader in efforts to build nuclear power plants, to commercialize them, and to promote the technology among the industrial nations. It continues to exercise major influence on the industry worldwide. Also, its commitment to nuclear, with only slight variation, has persisted through 10 successive presidential administrations. Hence, this early and lengthy experience with nuclear energy makes the United States an especially suitable focal point for examining the evolution of public acceptance of the technology. Although this examination focuses principally on the U.S. experience, it closely reflects the experience of other countries, such as Germany, Canada, Japan, the United Kingdom, and leading promoter France, although France exhibits certain deviations from overall patterns.

2. NUCLEAR GENERATION

In January 2003, 31 countries worldwide produced electricity with nuclear energy in 441 power reactors. World generating capacity totaled 359 GW (359,000 MW), representing 16% of total worldwide electricity production. With the completion of the 32 plants currently under construction, generating capacity is expected to increase by 7.5%, and it may increase another 10% if the 32 additional planned plants (35% larger, on average) are eventually completed.

Table I presents data for each of the world's nuclear nations. In terms of total generating capacity, the United States leads the world with 98.2 GW (104 reactors), followed by France with 63.1 GW (59 reactors), Japan with 44.3 GW (54 reactors), Germany with 21.3 GW (19 reactors), Russia with 20.1 GW (30 reactors), South Korea with 14.9 GW (18 reactors), the United Kingdom with 12.1 GW (31 reactors), Ukraine with 11.2 GW (13 reactors), and Canada with 10 GW (14 reactors). As a percentage of total electricity, Lithuania leads the way with more than 80% of its electricity produced by nuclear, followed closely by France (the largest

TABLE I

Nuclear Capacity and Operations of the World's Nuclear Nations

Country	Nuclear generating capacity (Gwe)	No. of nuclear power reactors (as of February 2003)	Nuclear as a % of total electricity generation (as of 2002)
Argentina	0.935	2	7.2
Armenia	0.376	1	40.5
Belgium	5.760	7	57.3
Brazil	1.901	2	4
Bulgaria	2.722	4	47.3
Canada	10.018	14	12.3
China	5.318	7	1.4
Czech Republic	3.468	6	24.5
Finland	2.656	4	29.8
France	63.073	59	78
Germany	21.283	19	29.9
Hungary	1.755	4	36.1
India	2.503	14	3.7
Japan	44.287	54	34.5
South Korea	14.890	18	38.6
Lithuania	2.370	2	80.1
Mexico	1.360	2	4.1
The Netherlands	0.450	1	4
Pakistan	0.425	2	2.5
Romania	0.655	1	10.3
Russia	20.793	30	16
Slovak Republic	2.408	6	54.7
Slovenia	0.676	1	40.7
South Africa	1.800	2	5.9
Spain	7.574	9	25.8
Sweden	9.432	11	45.7
Switzerland	3.200	5	39.5
Taiwan	4.884	6	21
Ukraine	11.207	13	45.7
United Kingdom	12.052	31	22.4
United States	98.230	104	20.3
Total	358.46	441	30[a]

[a] 16% of world total.

exporter of nuclear energy) at approximately 80%, Belgium at 57%, Ukraine at 46%, Sweden at 45%, Bulgaria at 47%, Switzerland at 40%, Japan at 34%, South Korea at 38%, the United States at 20%, Canada at 12%, and Russia at 16%.

Generating 16% of all electricity consumed in the world, nuclear power clearly remains a significant energy source. However, the significance diminishes considerably when placed in the context of projections made in the 1960s and 1970s during the heyday of the technology. One estimate projected 3000 plants worldwide by 2000 (nearly 7 times actual capacity), whereas another estimate, the 1972 Atomic Energy Commission's most optimistic estimate for the United States alone, called for 1200 plants by the same year (nearly 12 times actual capacity). Not a single plant was ordered in the United States after 1978, and every plant ordered after 1974 was eventually canceled so that not a single reactor built in the United States was ordered after 1973.

A number of industrial nations have chosen to forego the nuclear option, especially in the post-Chernobyl era: Australia, Austria, Denmark, Italy, Ireland, and Norway. Sweden continues to operate its 11 nuclear reactors (1 was closed in 1999) and has extended the 2010 date calling for the shutdown of all plants due to a 1980 national referendum, but it continues its policy of phasing out nuclear power. Germany, an early enthusiast for nuclear energy, passed legislation in 1998 for the eventual phase out of nuclear power but subsequently modified the policy in 2001 to limit the operational lives of existing plants to 32 years, thereby deferring any immediate closures. Japan, 80% dependent on imported energy, has scaled back its design for nuclear because of a September 1999 accident in Tokai, Mura, Ibaraki Prefecture, where two workers died from severe radiation and dozens of others were exposed.

A clear exception to this general pattern for the trilateral nations is Finland, the first nation in the world to have an approved site for the disposal of high-level nuclear waste and whose parliament in mid-2002 approved the building of a fifth nuclear power plant. Finland's success with its nuclear program may be the harbinger for a jump-start of the sagging nuclear programs of the majority of Western nations, or it may be an anomaly stemming from other unique features of Finnish history (e.g., its recent transformation from an agrarian to industrial society) and its unitary form of democracy.

Apart from Japan, nuclear capacity continues to grow in Asia (at the world's fastest rates), especially in China, the Republic of Korea, Taiwan, India, and Pakistan. As of 2002, more than two-thirds of the more than 32 reactors in 11 countries under construction throughout the world were located in Asia. Whether these nations will continue to experience growth uninterrupted by social and political obstacles or whether they will encounter difficulties similar to those experienced by Western

nations will be the subject of considerable scrutiny throughout the world.

In summary, although nuclear energy has assumed a significant and durable role in the energy mixes of a number of nations, its share of that mix is far short of the projections of three decades ago and with current technology it is projected to grow only moderately or even decline by midcentury. Furthermore, nuclear power is stalemated in many countries. What accounts for the striking underperformance of this technology? A number of factors conspired to derail nuclear's trajectory to predicted success: shifting energy markets, geopolitics, capital shortages, historical turns, and a failure to gain widespread public support. To what extent will these factors persist to obstruct a nuclear future?

A proper assessment of the future prospects for nuclear power requires consideration of those features of the historical context that are undergoing change. Two features loom particularly large in this respect, an internal one and an external one. Recognizing fundamental flaws in past reactor designs, such as the mushrooming of alternative designs and plant sizes and of limitations in safety systems, the world nuclear industry is working toward the development of generation IV reactors, a series of designs that are simpler, more fuel efficient, and arguably safer. The second, larger context is the scientific consensus over the warming of the planet due to the accumulation of greenhouse gases, especially CO_2. The evolution of these contexts is discussed in the next section.

Despite past experience and what appears to be a propitious new environment for nuclear power, the future of the technology will rest, ultimately, on public acceptance. Hence, the remainder of this article is devoted to the evolution of public attitudes toward nuclear power and to the implications the historical patterns have for the future of the technology. It is organized around three general themes: the evolutionary status of public attitudes toward the technology, the willingness by publics to see nuclear growth, and how publics view the technology abstractly.

3. EVOLVING CONTEXTS

3.1 New Reactor Designs

Nuclear power reactors are typically categorized in generational terms, reflecting evolutionary changes in design, size, and operational uses. Generation I reactors, the earliest to be developed in the 1950s and 1960s, were typically small (often based on submarine designs), with few still in operation today. Generation II reactors are the most common type of reactor in use today, dominated by U.S. light-water technology. Generation III reactors, an advanced design, are still in their incipient stages of development, with only a handful in operation in Japan. Generation IV reactors are the most advanced designs currently under development by 12 nuclear nations.

Comprising more than six different designs and systems, generation IV reactors all have in common higher (often much higher) operational temperatures than those of reactors currently operating. For some designs, this will allow for hydrogen production. Among the many other benefits arguably resulting from this generation of reactors are simpler designs and operations, thereby increasing safety; improved fuel efficiency, thereby improving the competitiveness of nuclear electricity; and use of a closed fuel cycle, thereby reducing the amount of high-level wastes to be managed.

Whether the increased safety of this new generation of reactors will attract widespread support for nuclear power will depend on a safety record that confirms the safety claims and whether publics trust the proponents of the new designs.

3.2 Global Climate Change

The emergence of a scientific consensus over global climate change has shed a new spotlight on nuclear power. Because it does not emit the principal greenhouse gas, CO_2, nuclear power is now viewed by many observers as a principal source of future energy poised for renewed, rapid growth. If there were dramatic new growth in nuclear power it could indeed reduce the rate of CO_2 emissions, but it should be recognized that nuclear is not the sole solution to global warming. Nuclear energy currently accounts for only 2.5% of total world energy consumption (8% in the United States and 33% in France), and its growth is limited by electricity's meager 16% share (the only type of usable energy produced by nuclear) of world energy consumption. The 16% share could increase as a function of electricity's growing fraction of total energy use, but nuclear's potential for addressing climate change issues nevertheless remains relatively small.

Whether public recognition of the seriousness of climate change and nuclear's ameliorative role in curbing it will swing the pendulum of public opinion to majority support is a difficult question to assess. Among the many unresolved issues the question raises is how citizens will balance a very diffuse and

perhaps distant risk whose observable consequences are not immediately known (climate change) against a very proximate risk (nuclear accidents) for which some of the consequences are known. The social psychological literature suggests that the proximate risk would trump the diffuse one.

4. PUBLIC REACTIONS TO NUCLEAR POWER IN THE UNITED STATES

4.1 Long-Term General Trends

4.1.1 Building More Nuclear Power Plants

As noted previously, the ultimate fate of nuclear power rests with public acceptance. The longest running nuclear time series, providing the best available window for assessing the evolution of public opinion, concentrates on attitudes toward the continued growth of nuclear power. Similar, but not identical, growth questions have been asked by leading polling organizations Louis Harris and Associates and Cambridge Reports Research International. The evolution in opinion about nuclear plant construction, including opinion reversal, can best be seen by charting trend lines for these data series. Figure 1 presents a trend line from the Harris data and Fig. 2 presents a trend line from the Cambridge data.

The Harris item asking for opinions about building more nuclear plants was first asked in 1975 and repeated until 1990. In general, until the late 1970s a majority of Americans supported, often by wide margins, growth in nuclear power (an exception is the April 1976 result showing a plurality of support). Then, immediately following the accident at Unit 2 at Three Mile Island (TMI) in Pennsylvania in April 1979, this support eroded considerably but still consistently exceeded opposition. With the May 1979 survey, support had regained its majority position but appeared to be standing on much shakier ground than before the accident: Over the course of 12 surveys conducted over the next 22 months, support for nuclear power, although generally consistent, vacillated between bare majority and plurality. In one case, May 1980, opposition actually exceeded support by

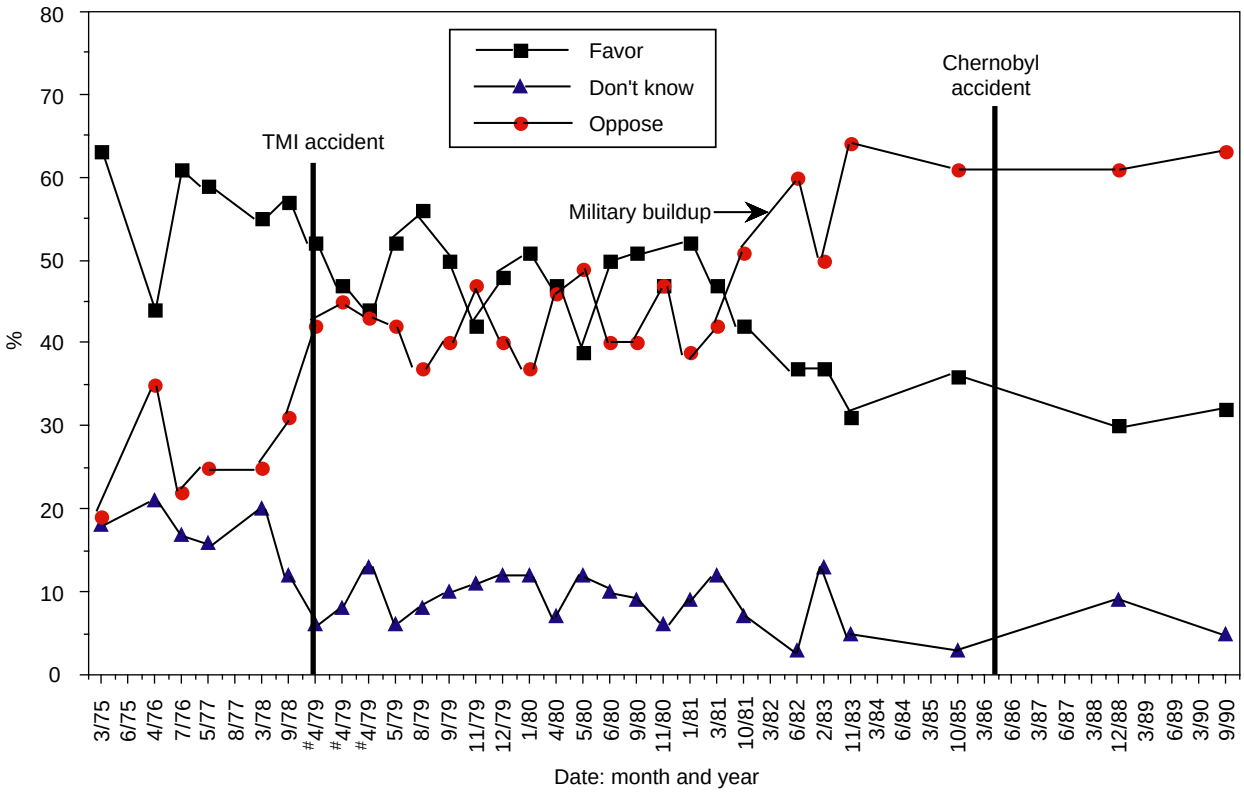


FIGURE 1 Public attitudes toward building nuclear power plants in the United States. Question: "In general do you favor or oppose the building of more nuclear power plants in the United States?" TMI, Three Mile Island. #Indicates multiple surveys in the same month. Source: Louis Harris and Associates national samples.

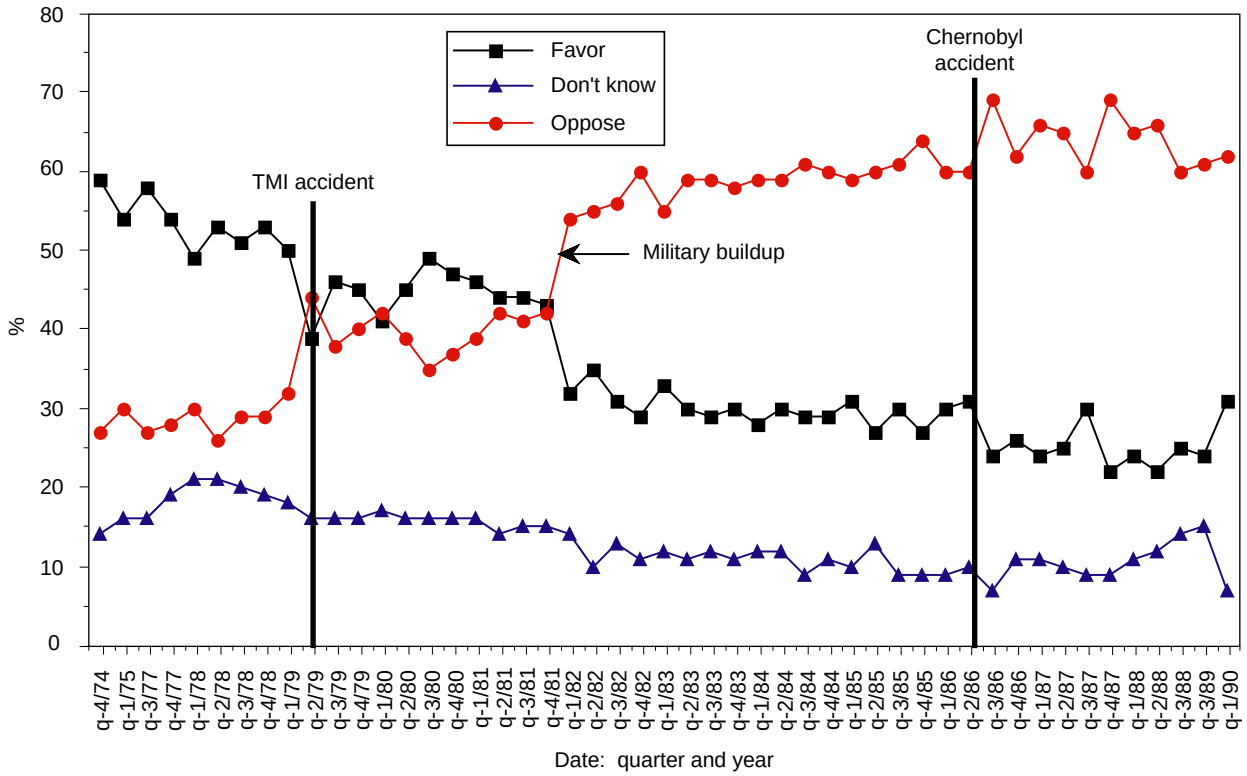


FIGURE 2 Public attitudes toward building nuclear power plants in the United States. Statement: "I am going to read you a list of proposals for dealing with the energy crisis, and I'd like you to tell me whether you generally favor or oppose each one … Building more nuclear power plants." TMI, Three Mile Island. Source: Cambridge national samples.

10 percentage points, but this was a temporary setback lasting for only a short period of time.

In contrast, the October 1981 result showing majority opposition signaled not just a temporary setback but also a break with the past—indeed, a reversal in public preference for nuclear growth. Since then, opposition has consistently exceeded support. In fact, over the past 4 surveys spanning 7 years, opposition exceeded support by a margin of approximately 2:1, a complete reversal from the earliest Harris survey.

Cambridge, likewise, queried the public repeatedly about whether to build more nuclear power plants. In the fourth quarter of 1974, Cambridge began asking Americans their views about building more nuclear power plants as a way of dealing with the energy crisis. That Cambridge contextualized its question within a concern for the energy crisis is perfectly understandable given that it first polled the public on the subject in the months following the Arab oil embargo of 1973–1974 when national agendas were dominated by the idea of an "energy crisis." This contextualization distinguishes the Cambridge series from the Harris one, and one might expect it to have a noticeable affect on the overall trends. With the notion of crisis in mind, citizens might be willing to accept any energy source that would mitigate or eliminate the crisis, including nuclear power. It might also be the case that in later surveys, long after the idea of crisis had receded from public consciousness, the effect could be in the opposite direction: Because a crisis no longer exists, there is no urgent need to build more nuclear power plants.

Figure 2, presenting the Cambridge trends, exhibits similar patterns to the Harris results; nuclear power enjoyed consistently strong support throughout the 1970s, at least up until the TMI accident in April 1979. The effect of the accident on public attitudes is apparent, as opposition exceeded support for the first time in the second-quarter 1979 survey. By the third quarter of that year, however, support had rebounded sufficiently to outpace opposition, but not by as large a margin as before the accident. Furthermore, while support remained consistent during the next $2\frac{1}{2}$ years, covering 10 surveys, it was only a plurality of support, and sometimes barely that. Then, sustaining the pattern of the Harris results, a sharp reversal in opinion appeared

(showing up for Cambridge in the first quarter of 1982), with a majority of respondents expressing opposition to further nuclear plant construction. What is most noteworthy about this shift from majority support to majority opposition is the fact that it endured across all subsequent surveys over the course of 8 years. Furthermore, for most surveys the ratio of opposition to support approaches or exceeds 2:1—again, a near-perfect reversal of public preference in the early 1970s. Finally, the substantive similarity between the Harris and Cambridge results suggest that wording effects, if there are any, do not markedly alter the overall patterns.

In general, despite the slight wording changes between the polling organizations, a common and persistent pattern of growing opposition to the building of more nuclear power plants emerges. The pattern comprises three distinct stages: an early stage in the 1970s when Americans were enthusiastic about nuclear growth, a second stage of ambivalence following TMI when a less enthusiastic plurality of citizens consistently supported growth, and a third stage emerging in the early 1980s when a decisive majority of Americans opposed building more nuclear power plants in the country. Apace with the decline in support for nuclear power plants was the decline in the proportion of "don't knows," reflecting, perhaps, the eventual crystallization of attitudes.

A focus on the timing of the reversal illuminates still another striking feature in the overall pattern: the change in attitudes, with opposition overtaking support, occurred more than 4 years before the Chernobyl nuclear accident in the Ukraine. This was apparently due to a renewed fear of nuclear war stimulated by the Reagan administration's bellicose posture toward the Soviet Union and the accompanying buildup of nuclear arms. Whatever the reason for the striking shift in attitude, it set in motion a steady growth in opposition that continued to the very end of each data series. With such a high level of opposition, there was little room for further opposition, so the Chernobyl accident evidently did not have a marked impact on public attitudes: At the time of the accident, opposition already exceeded support (best seen in the Cambridge results in Fig. 2) by nearly two to one. Whatever effects Chernobyl may have had, they were apparently negligible compared to the sustained long-term decline in support for new nuclear power plants.

Although these data provide a clear picture of the evolution of opposition to nuclear power, the question about building more nuclear plants has not been repeated for more than a decade. Given the propitious new winds for nuclear power noted previously, how can we be certain that opposition to the building of more plants remains at this high level? Four polls address this question. A March 1999 poll of registered voters found 60% opposed and 26% in favor of building more nuclear power plants. A March 2001 Gallup poll of all adults found less but still majority opposition (51 vs 44%). Two 2001 polls of all adults by ABC News/*Washington Post* sustain the conclusion that opposition to the building of nuclear power plants is still in the majority. In its January poll, 60% were either somewhat or strongly opposed, whereas 37% were somewhat favorable or strongly favorable. Its June poll revealed some relaxing of opposition but still with a majority (52%) opposed, whereas 41% were in favor.

Clearly, the American public does not support the construction of additional nuclear power plants. This is the unmistakable message contained in the longest running time series assessing nuclear opinion, reaffirmed by recent survey data.

The public opinion experience of other countries deviates somewhat from the pattern in the United States but available data suggest that citizens in most nuclear nations have lingering reservations about the technology. For example, although opposition to nuclear power in Sweden has softened during the past two decades, opposition still outpaces support for the technology. Available public opinion data in Japan show a very similar pattern to that of the United States, except that opposition's overtake of support did not occur until the Chernobyl accident in 1986, and recent data show the Japanese to be more fearful of a nuclear accident than any other nation.

In comparison to the nuclear stalemate in much of the world, France has experienced nearly uninterrupted success with its nuclear program. This has repeatedly prompted nuclear proponents to point to France as the exemplar of not only a successful nuclear program but also a model for other nuclear nations to emulate. Whether France is an appropriate model for other nations to follow is subject to considerable debate. Key differences distancing France from other nuclear programs include its monopoly ownership by a single government utility, Electricité de France; its unique system of training for its technological elite; its strategy of adopting a single design for all its reactors; its practice of siting multiple reactors at a single location; its unitary democratic structure; and its policy style (called technocratic etatism), in which individual liberties are subordinated to the will of the state.

Added to these obvious differences are others more subtle but no less important to France's nuclear success. In public opinion polls, French citizens are consistently more supportive of nuclear than their American counterparts. What accounts for this persistent pattern? In the only comparative opinion study between French and American publics it was found that perceptions of the risk of nuclear power were virtually identical between the two publics, thereby eliminating risk as an explanation for the national differences. French citizens distance themselves from Americans in their strong belief in the need for nuclear power, their more fatalistic viewpoint about its risks, their greater trust of nuclear management, and especially their greater trust in experts and civil authority (a feature of etatism). There are likely similar differences between the French and citizens of other nuclear democracies.

4.2 Proximate Trends

4.2.1 Nuclear Power as a Source of Electricity

The sole commercial use of nuclear power is the generation of electricity. Despite the long-term trends showing a growing dissatisfaction with the building of more nuclear power plants in the United States, perhaps the public is more supportive of the technology when it is connected directly to its instrumental purpose: the generation of electricity. In a poll taken soon after the Chernobyl accident, the public was evenly split on the use of nuclear power to generate electricity, but a strong 55% majority favored nuclear in a follow-up poll in 1989. Other polls on public attitudes toward nuclear power as an electricity source followed. For all three polls conducted between 1992 and 1993, a majority of respondents favored (strongly or somewhat) the use of nuclear power to generate electricity. It should be noted, however, that these polls did not ask for unqualified support for nuclear power but as one of the ways of providing electricity. Americans are apparently aware that nuclear is already an important energy source for the generation of electricity. Two surveys in 1986 and 1989 asked how important nuclear was in meeting the nation's electricity needs today. A sizable majority in both (70%) agreed that it was important. The most recent data show a considerable softening of support, although the general pattern is sustained.

4.2.2 Local Siting of Power Plants

It is one thing to have an opinion toward the construction of nuclear power plants in the abstract; it is another thing to be confronted with the prospect of having a plant built nearby. Figure 3 presents trend data from national samples that asked about having the local electric company build a nuclear power plant in the area.

The early siting of nuclear power plants was hardly a controversial matter. Indeed, in five of the earliest polls 55% responded favorably, nearly twice the level of opposition. However, opposition to local nuclear facilities began to increase as early as 1974. By 1976, a slight plurality, and by 1978—before the TMI accident—a clear majority of respondents opposed the construction of a nuclear power plant within 5 miles; in less than a decade the opposition reached 73%. Polling asking whether citizens would be willing to have a nuclear power plant within 10 miles of their community attracted 54% opposition in 1999 (compared to 21% in favor) and 50% opposition in 2000 (compared to 22% in favor). In a 2001 poll, 63% of the adults sampled were somewhat or strongly opposed to the construction of a nuclear power plant "in their area," whereas 33% were somewhat or very favorable to the idea.

It might be expected that, as had been the case regarding the building of nuclear power plants in the abstract, Americans would be more tolerant of local nuclear plants if they were reminded that the purpose of the technology was to provide electricity. Cambridge and Gallup, using an identically worded item, attempted to assess this possibility in a series of polls from 1986 to 1990. The most popular of three possible responses, attracting near-majority to clear-majority support, was to "reserve judgment"—perhaps reflecting, on the one hand, a reluctance to reach a firm position until apprised of the specific circumstances surrounding such a decision and, on the other hand, a willingness to leave open the nuclear option as a hedge against future electricity shortages. Of those who had evidently formed an opinion about a local plant, opposition consistently outweighed support, but by a declining margin over time.

4.2.3 Local Host Communities

As suggested by the data (Fig. 3) about the local siting of nuclear power plants, in the early years of nuclear commercialization communities hosting plants were quite supportive of the technology. Indeed, a number of communities actively sought to attract nuclear power plants for economic reasons and as a matter of technological pride. This early enthusiasm is sustained by the trend data in Fig. 4, a summary of 36 host community surveys.

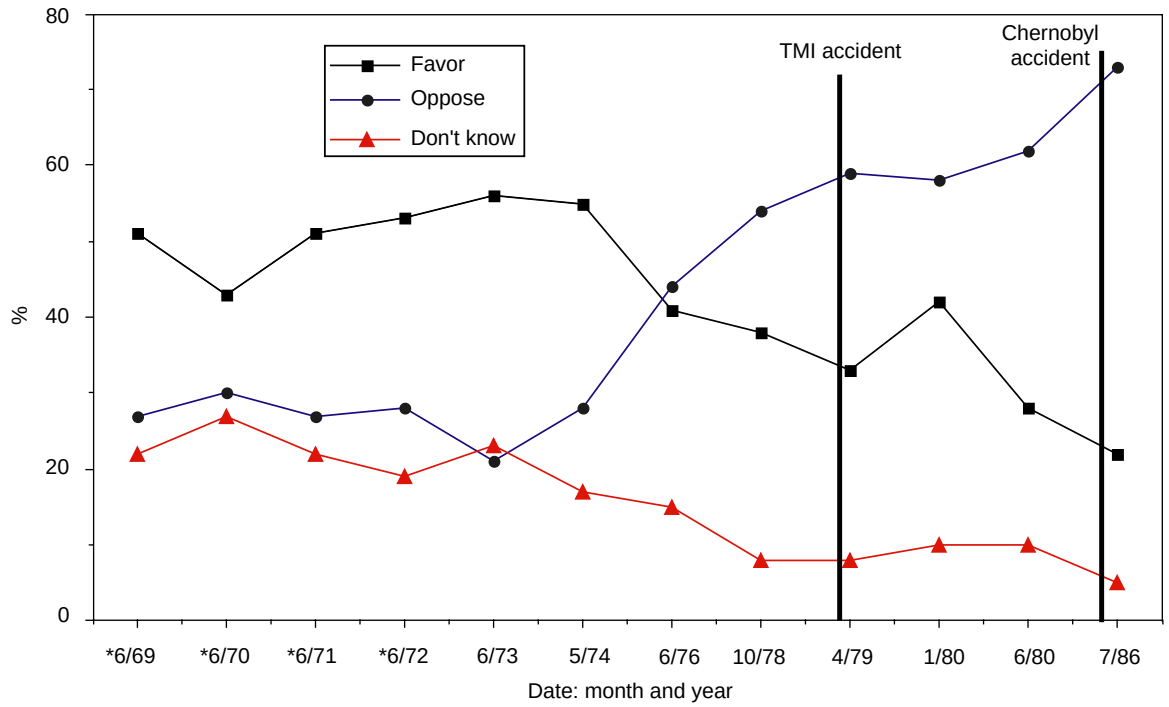


FIGURE 3 Public attitudes toward construction of local nuclear power plants. Question: "Suppose your local electric company said it wanted to build a nuclear power plant in this area. Would building such a plant be all right with you, or would you be against it?" TMI, Three Mile Island. *, Month survey conducted not reported in original sources. Midyear dates were assumed. Source: Cambridge national samples. Boulder, CO.

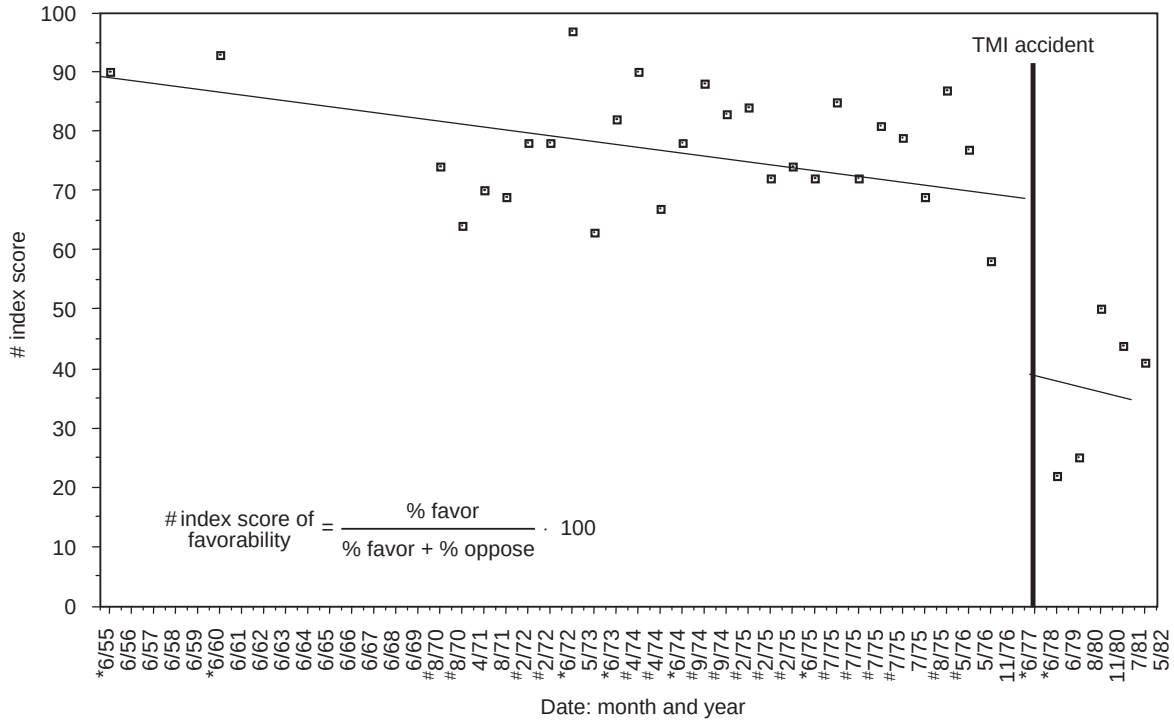


FIGURE 4 Host community attitudes toward nuclear power plants before and after the Three Mile Island (TMI) accident. *Month survey conducted not reported in original sources. Midyear dates were assumed. #Indicates multiple surveys in the same month. Source: Cambridge national samples.

Post-TMI data suggest, however, that host community support can no longer be taken for granted. In none of the surveys conducted before the TMI accident was opposition found among more than 33% of respondents; in no known survey after the accident has there been less than 50% opposition. Regression analysis shows that more than 70% of the explained variance across the surveys is due to the TMI accident. Host communities in the United States, once supportive of nuclear facilities in their vicinity, now show considerable reservation about having those facilities nearby. Also, the siting of nuclear facilities is a persistent problem in a number of nuclear nations.

5. ENVIRONMENTAL CONSCIOUSNESS AND NUCLEAR'S FORTUNES

The presented trend data clearly illustrate the dramatic impacts key events (the TMI accident and military buildup) had on public opinion about nuclear power. However, preceding the reversal in general attitudes by several years was growing disenchantment with local nuclear facilities—the principal fault line in the growth of nuclear power. What accounted for this early disenchantment?

In the late 1960s and early 1970s, the Western world experienced a profound cultural change in collective consciousness in which environmental impacts, previously a set of disparate concerns, coalesced into an environmental consciousness. Signature features of this change were the passage of the National Environmental Protection Act in 1969 in the United States, a model for similar legislation elsewhere, and the first Earth day in 1970, the largest public demonstration in the history of the planet—attracting many millions of participants worldwide. Although these events did not immediately lead to a reversal in nuclear's fortunes, they established a cultural framework of unprecedented scrutiny of the technology. Early local opposition to nuclear power plants reflected the emergent consciousness with environmental issues such as thermal pollution (release of heat to rivers and lakes) and threats to marine species as the rallying cries for protest.

When the accidents at TMI and Chernobyl took place, publics everywhere were predisposed to see the events in the most negative light. Another element of the new consciousness, later to impact nuclear directly, was ever-greater demands on policy makers to more actively engage the public in decisions regarding large-scale technologies.

The emergence of a consciousness potentially critical of nuclear energy was one side of the energy supply coin. Missing was the other side of the coin that provided a coherent alternative, a revolutionary vision of an energy supply mix absent a nuclear component. In the 1970s, many scholars were attracted to the issue, but none could match the impact of E. F. Schumacher's "appropriate technology" vision or Amory Lovins's "soft path" vision. Now, instead of a focus on nuclear technology versus its alternatives, the focus was broadened to a criticism of the megascale, centralized energy supply systems of industrial societies. Lovins went further, developing a detailed blueprint of renewable energy systems absent a nuclear component.

6. THE PRESENT AND THE FUTURE

6.1 Pragmatism

Does the sizable opposition to the building of more nuclear power plants translate into a deep-seated aversion to the technology, into an unreflective rejection of all matters nuclear or even a willingness to get rid of the technology completely? Is the opposition the result of sensible caution about a large-scale technology that is neither simple nor risk-free? In one survey, respondents were asked whether nuclear power was a good choice, a realistic choice, or a bad choice. Strong pluralities to majorities consistently said it was a realistic choice. The remaining 40–50% of the public was more likely to say it was a bad choice, but sometimes by only narrow margins.

With regard to eliminating nuclear power, only minorities of Americans have supported this option. The most frequently chosen option, attracting sizable pluralities to majorities, favors the status quo: Let the existing nuclear plants operate, but do not build any more. Taken together, the picture that emerges from these items is not one of a knee-jerk public passionately committed to the riddance of nuclear energy but of a pragmatic public weighing technological risks against unforeseeable future energy needs. As a hedge against a shortfall in future energy supply, the public appears willing to leave the nuclear option open.

6.2 Nuclear Futures

What does the public believe the future holds for nuclear power? The longest running time series, the

Harris and Cambridge ones, asking about building more nuclear plants are of limited use in addressing this question because the time orientation of these items is ambiguous. The ambiguity can be avoided with questions that specifically refer to the future, as have a number of polls.

Since 1985, Americans have repeatedly been asked how important they think nuclear energy plants will be in meeting the nation's electricity needs in the years ahead. The trend data of responses to this question are presented in Fig. 5.

It is clear that sizable majorities of Americans see an important role for nuclear power in the future. Three-fourths of respondents have consistently thought nuclear power will play a very important or somewhat important role in meeting the nation's future electricity needs. This finding is consistent with public knowledge of the importance of nuclear-generated electricity and with the public's pragmatic acceptance of the status quo. The picture changes somewhat when Americans are asked whether the nation should rely more on nuclear power in the future. In two polls, respondents were slightly more inclined to opt for less nuclear power in the future.

6.3 Nuclear as a Trade-Off Choice

Nearly all imported oil is used for petroleum products and not for the generation of electricity. Nevertheless, Americans have been asked whether they would favor nuclear power if it would reduce the nation's dependence on foreign oil. Presented with such a trade-off, majorities of the public favor nuclear power—often by sizable margins.

The modern trade-off context is the question about substituting nuclear for fossil fuels (especially coal) in the generation of electricity to reduce greenhouse gas emissions. What is the level of public support for nuclear power to address global warming? Data directly addressing this question are meager. Surveys from the 1990s show that majorities believed that nuclear energy could cut greenhouse gas emissions and air pollution and vast majorities would favor using nuclear energy to cut greenhouse gases and air pollution. Recent data, however, contradict that finding. Asked whether they would favor nuclear power as a means "for dealing with the pollution that causes climate change," a majority of citizens (55 vs 42%) were opposed. Thus, we have a far from complete answer to this question.

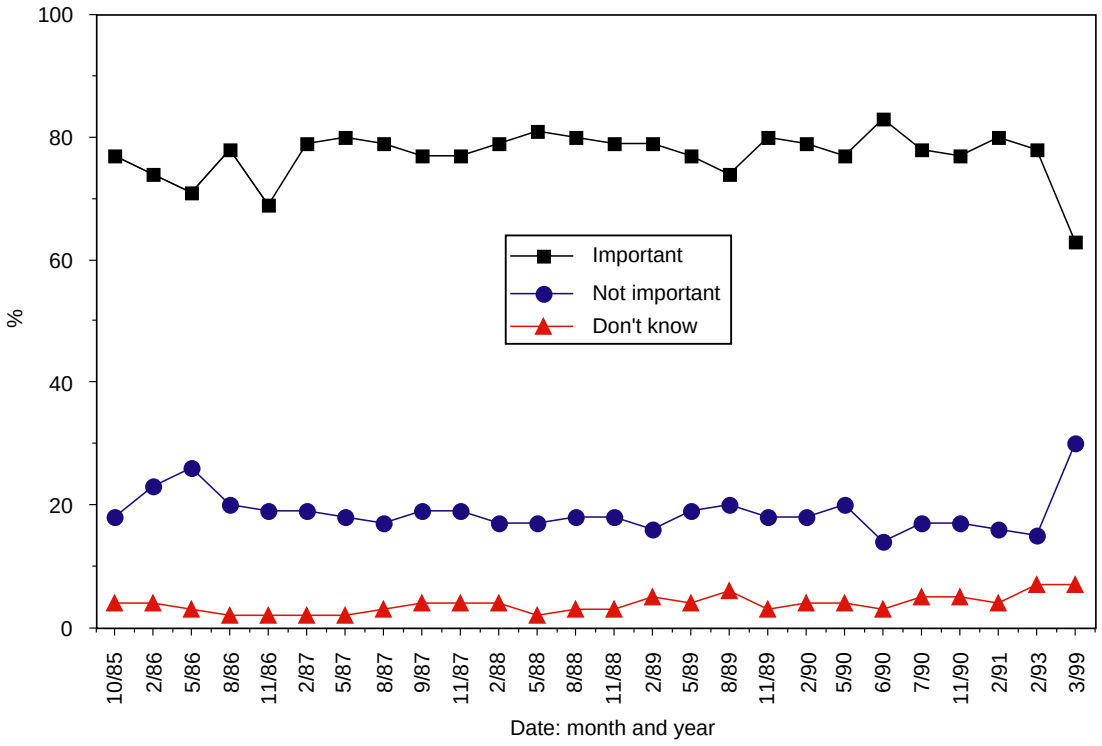


FIGURE 5 Importance of nuclear power in the future. Question: "How important do you think nuclear power plants will be in meeting this nation's electricity needs in the years ahead?" Source: Cambridge reports and Gallup national samples.

6.4 Interpreting the Public Mood

Taken together, the overall results may appear contradictory, perhaps even irrational. On the one hand, solid majorities of the American public oppose the construction of more nuclear power plants and are likewise opposed to their local siting. On the other hand, equally solid majorities believe nuclear power will be an important energy source in the nation's future, although this is not necessarily their first choice. What accounts for the apparent contradiction? In our view, the results do not reveal a contradiction but instead a pragmatic logic whereby the public distinguishes between nuclear power in principle and in practice.

For nuclear power, it appears that the collective public's preference for specific policy actions can be quite different from its general attitudes. Thus, Americans appear generally to support a role for nuclear power in the future and are solidly behind the idea of leaving the nuclear option open—perhaps as a trump card against possible future energy shortages. However, when it comes to the specific instrumental means for achieving that future—namely, the siting and construction of nuclear power plants—they are solidly opposed. What are the key impediments to the realization of the public's expected nuclear future?

7. NUCLEAR IMAGERY

Underlying attitudinal responses and global assessments of nuclear energy are a variety of cognitive filters as well as images that shape thoughts and dispositions toward the technology. Although nuclear energy is Janus-like, with inherent benefits and risks, there is considerable evidence showing negative imagery is much deeper than positive imagery, overpowering the positive with salient thoughts of danger, toxicity, death, sickness, destruction, pollution, and the like. Some researchers have gone further, arguing that nuclear facilities, especially nuclear waste facilities, attract stigma—a deeply discrediting or devaluing feature that marks them for special concern or avoidance. Although the claim for stigma may stretch the available evidence, it is unequivocally clear that images associated with things nuclear are predominately negative and militate against the unreserved acceptance of the technology. There is little doubt that the future of nuclear energy is heavily dependent on overcoming this deep negative imagery.

8. NUCLEAR WASTE

One of the unavoidable by-products of nuclear-generated electricity is the production of a variety of highly radioactive wastes. Originally believed to be a tractable, and solvable problem, a solution to the disposition of nuclear waste still eludes the nuclear establishment worldwide. Although there is a universal scientific consensus that deep geological isolation is the preferred solution to the disposition of wastes—especially high-level wastes with their extremely long radioactive half-lives—no nuclear nation has built a geologic repository or has emplaced any waste. With few exceptions (e.g., Finland), no nation has even been able to achieve a sufficient level of political and public support to reach an uncontested final decision on the siting of a repository. In the meantime, the waste produced by nuclear energy accumulates in pools at reactor sites or at military installations.

A failure to solve the waste problem proves to be especially challenging in gaining public acceptance of the technology. Publics everywhere hold deep negative images of nuclear waste—even deeper than for nuclear power—and are opposed to siting a repository not only in their immediate vicinity but also just about anywhere. When asked to indicate the closest distance they would be willing to live in the vicinity of a variety of facilities, a nuclear waste facility typically ranks last—in one survey at 200 miles away (twice the distance from the next most undesirable facility, a chemical waste landfill, and three to eight times the distances from oil refineries, nuclear power plants, and pesticide manufacturing plants).

Particularly revealing of the depth of negative imagery associated with nuclear waste are results from a survey in which citizens were asked to associate freely about the concept of a nuclear waste repository. The vast majority of images elicited were extremely negative, punctuated by such categories as dangerous/toxic, death/sickness, environmental damage, and bad/negative.

Thus, a failure to solve the nuclear waste problem could be the Achilles' heel of nuclear power, ensuring a continuation of the technology's cloudy future or worse—its ultimate failure. That the ultimate success of nuclear power is conditioned by a solution to the waste problem is underscored in a 2001 Eurobarometer poll conducted in 15 European Union (EU) countries. When asked, "If all waste is managed safely, should nuclear power remain an option for electricity production in the EU?" 51% of the 16,000 respondents answered in the affirmative, whereas

near equal percentages rejected this option (25%) or had no opinion (24%). This conditional favorability, however, is tempered by the strong plurality (46%) of the same respondents who said they thought that there was no safe way to dispose of the waste.

9. CONCLUSION

If it were written in ancient times, the history of nuclear power in the United States and elsewhere would be written as a drama between two mythical figures: the Phoenix and Hydra. Like the mythical Phoenix that arose from its own ashes, the nuclear establishment—a combination of scientific, commercial, and policy interests—counters each setback to nuclear's fortunes with either organizational or technical fixes. The emergence of a fourth generation of reactors, presumably much safer than their predecessors, is the most recent manifestation of this pattern. In contrast, since the mid-1970s the general public has viewed nuclear power as a many-headed Hydra in which as soon as one of its concerns is addressed—or recedes from consciousness—two others spring up to take its place.

Since the first sustained chain reaction at Stagg Field at the University of Chicago in 1942, commercial nuclear power has experienced paroxysms of enthusiasm and growth, followed by extended periods of reflection and reconsideration. Public support for the technology began with unrestrained enthusiasm, only to be followed by extended periods of doubt and reservation. Currently, two clouds of uncertainty continue to shroud public acceptance of the technology: the safety of nuclear plants and the challenge of disposing of nuclear wastes, especially high-level wastes.

Whether public acceptance can be regained with a new generation of reactors with advanced designs—argued to be inherently safer than the current generation of reactors—will depend on a sequence of contingencies. It will depend on whether the claims about safety are credible, whether credible arguments about safety can overcome the deep negative imagery that shrouds the technology, whether alternative technologies emerge and are economically competitive, and on the vagaries of the future geopolitics of energy supplies. It will also clearly depend on solving the nuclear waste problem.

Irrespective of the regaining of public trust for nuclear power and its management, it is unlikely that the technology will be abandoned worldwide—at least not in the foreseeable future. Too much has already been invested in nuclear infrastructure (including many national and international institutions), too many national states see a significant role for nuclear in their future energy mix, and too many powerful technological elites remain committed to promoting its growth. However, many questions remain. Will nuclear power remain in steady state (eventually resulting in its demise as the current inventory of plants reach the end of their usable lives) or grow again? Will it be able to grow absent a successful program for the disposal of nuclear wastes? If it does grow, will it be in a supportive public climate or a climate of continuing tension, forcing the government to take a more active role in overcoming or circumventing those tensions—perhaps with legislative or regulatory actions that erode the democratic process? All these questions embed even deeper and more challenging ones about technological knowledge, historical contingency, political culture and climate, and the evolution of democracies. It will be the answers to these hoary questions that will determine the future of nuclear power.

SEE ALSO THE FOLLOWING ARTICLES

Environmental Injustices of Energy Facilities • *Nuclear Engineering* • *Nuclear Power: Risk Analysis* • *Occupational Health Risks in Nuclear Power* • *Public Reaction to Electricity Transmission Lines* • *Public Reaction to Offshore Oil* • *Public Reaction to Renewable Energy Sources and Systems*

Further Reading

Bupp, I. C., and Derian, J.-C. (1981). "The Failed Promise of Nuclear Power: The Story of Light Water." Basic Books, New York.

Dunlap, R. E., Kraft, M. E., and Rosa, E. A. (eds.). (1993). "The Public and Nuclear Waste: Citizens' Views of Repository Siting." Duke Univ. Press, Durham, NC.

Freudenburg, W. R., and Rosa, E. A. (eds.). (1984). "Public Reactions to Nuclear Power: Are There Critical Masses?" Westview /AAAS.

Garwin, R. L., and Charpak, G. (2001). "Megawatts and Megatons: A Turning Point in the Nuclear Age." Knopf, New York.

Jacob, G. (1990). "Site Unseen: The Politics of Siting a Nuclear Waste Repository." Univ. of Pittsburgh Press, Pittsburgh, PA.

Jasper, J. (1990). "Nuclear Politics: Energy and the State in the United States, Sweden, and France." Princeton Univ. Press, Princeton, NJ.

Lovins, A. (1976). Energy strategy: The road not taken. *Foreign Affairs* **55**, 65–96.
Rosa, E. A., and Dunlap, R. E. (1994). Nuclear power: Three decades of public opinion. *Public Opin. Q.* **58**, 295–325.
Schwartz, S. I. (1998). "Atomic Audit." Brookings Institution Press, Washington, DC.
Weart, S. (1988). "Nuclear Fear: A History of Images." Harvard Univ. Press, Cambridge, MA.

Public Reaction to Offshore Oil

WILLIAM R. FREUDENBURG
University of California, Santa Barbara
Santa Barbara, California, United States

ROBERT GRAMLING
University of Louisiana, Lafayette
Lafayette, Louisiana, United States

1. **Introduction**
2. **Historical Overview**
3. **The U.S. Tidelands Controversies**
4. **The International Growth of Offshore Technologies**
5. **The Evolving Range of Views Throughout the World**
6. **Factors Associated with Differing Viewpoints in the United States**

Glossary

attitudes Unlike physical objects, attitudes cannot be seen or weighed. Instead, they are more like the biological concept of life—difficult to define but sufficiently important and pervasive to have attracted considerable study. In general, they are seen as being like opinions but more deeply connected to one's overall orientation toward the world (and hence less likely to change) while not as deeply connected as values. In most of the technical literature, attitudes are seen as having three components—relatively factual beliefs, more emotional evaluations, and in many but not all definitions, "conative" or behavioral orientations, involving intentions to act in certain ways but not others.

Continental Shelf The relatively shallow ocean-bottom lands close to the edge of a continent. In many but by no means all cases, these lands do form a literal shelf, sometimes extending for dozens or even hundreds of kilometers beyond dry land before dropping off into deeper waters. In the United States, the outer Continental Shelf includes the ocean-bottom lands beyond the limits of state jurisdiction (generally 3 miles from land).

Dutch disease A term that is sometimes used to describe almost any case of economic weakness in the face of resource exports but that more formally refers to economic weaknesses resulting from twin problems experienced in The Netherlands in the wake of its North Sea oil discoveries: (i) a significant increase in the strength of national currency (making its other exports more expensive) and (ii) a tendency for the oil sector to draw people and capital away from other sectors of the economy, such as manufacturing and agriculture.

economic multiplier Many approaches to socioeconomic impact assessment divide local or even national economies into two broad sectors: economic base (or export) activities, which bring cash into an area through the selling of goods and services to people who live elsewhere (e.g., agriculture, oil extraction, and tourism), and nonbasic (or service) activities, which can create additional jobs by keeping the cash circulating locally (e.g., when people use that cash while visiting local shoe repair shops, restaurants, grocery stores, etc.). The ratio between basic and nonbasic activities provides the economic multiplier, although actual multipliers from new activities (e.g., new oil drilling) have often proven to be much lower in practice, in recent decades, than the ones that have been put forth in the proponents' economic models.

social multiplier Just as one economic base job can help to support other nonbasic jobs in the same community or region, one person who works for a given industry is likely to have many friends and relatives nearby, particularly in the case of an established industry. Since people tend to care not just about their own well-being but also about the well-being of their friends and acquaintances, there can be much stronger political support for an important industry (e.g., agriculture in Iowa and oil in Texas) than might be expected simply on the basis of the number of people or jobs supported by that industry directly.

tidelands controversies A series of battles during the mid-20th century involving political control over (and tax revenues from) new activities of offshore oil drilling in U.S. waters. The controversies were largely brought to an end by two pieces of legislation, one of which gave control and revenues to specific states for drilling taking place in territorial waters (generally those within 3 miles of shore) and the other of which granted control and revenues to the federal government for activities farther offshore.

 © 2004 Elsevier Inc. All rights reserved.

It was more than a century ago when humans first thought of the idea of drilling through water, as well as land, in search of oil, and during the course of this century, the attitudes or views about potential oil developments have varied widely. Rather than representing relatively random reactions, or anything approaching universal approval or opposition, local attitudes tend to vary more systematically across time and space. These attitudes, in turn, shape the policies and policy actions that can cause offshore oil development to be easier or more difficult to pursue. This article discusses the historical background and geographic extent of offshore oil development and the importance of three factors with regard to a region's attitudes toward offshore development: the region's historical relationship to oil development; the biophysical environment of the region, such as its coastal and marine topography; and relevant sociocultural conditions, such as educational levels, social networks, and prior economic adaptation.

1. INTRODUCTION

The first offshore drilling was actually done from a set of oil derricks along a pier that jutted into the Pacific Ocean off the shores of Summerland, California, southeast of Santa Barbara, in 1898, and the first offshore oil controversy erupted a few miles away the very next year. When an oil company began to construct an oil derrick off the shores of Montecito, California—the highly affluent Santa Barbara suburb that is adjacent to Summerland and that occupies the few miles of coastline between Summerland and Santa Barbara—a local mob, described approvingly on page 1 of the *Santa Barbara Morning Press* the next day as "a party of the best known society men of Santa Barbara, armed to meet any resistance," attacked the rig and tore it down.

The local "society men" seem not to have suffered any noteworthy legal repercussions from their actions, despite having been so well-known, but oil companies did: The graphic expression of local attitudes proved effective in blocking further drilling along that stretch of coastline for decades to come. Elsewhere in the world, however, for at least the next 50 years, most such efforts to find oil—offshore as well as on—were much more likely to be the focus of celebration, not protest. Had this article been written in 1953 rather than in 2003, it almost certainly would have included a discussion of what were known at the time as the "tidelands controversies," but as noted later, these controversies had to do not with opposition to offshore oil but with the virtual opposite, involving competition over which political jurisdictions would be able to enjoy the accompanying tax benefits.

Today, offshore exploration and production have spread throughout the world, taking place off the coasts of Africa, Australia, Central and South America, and Europe. In the United States, offshore oil is limited to a handful of rigs off Alaska's North Slope, 21 rigs off California (all in the Santa Barbara region, just a few miles away from Summerland and Montecito), and a vastly larger number of less famous oil rigs (approximately 3500) in the Gulf of Mexico, where some platforms are producing oil in water that is 2 miles deep.

As these patterns suggest, public attitudes toward offshore oil activities are far more diverse than once might have appeared to be the case. To understand what has happened in the century or more since the derricks first appeared on the Summerland pier, and to understand the variations in public attitudes toward offshore oil activities, it is useful to begin with a brief history. Following the historical summary, we present an overview of the variations that can be found in some of the key offshore regions of the world and then provide a more detailed examination of what appear to be the two polar cases—the two regions, both in the same country, where attitudes appear to be the most and the least favorable.

2. HISTORICAL OVERVIEW

The original impetus behind the search for offshore oil came just 17 months after the displeasure of the Montecito mob, when the world's first major petroleum reservoir, Spindeltop, was tapped on January 10, 1901, near Beaumont, Texas. Compared to the petroleum reservoirs that had been found in Pennsylvania and elsewhere during the latter part of the 19th century, Spindeltop was huge; its discovery transformed the petroleum industry, both in the United States and throughout the world. For purposes of this article, however, the key point is that the reservoir proved to be located under the limestone cap that covered the top of a salt dome. Not surprisingly, this discovery led to interest in other such domes, many of which are located off the shores of Louisiana and Texas.

Despite the early discoveries off the shores of south-central California—or perhaps in part because

of the reactions they inspired—it was not until the 1920s that offshore oil development began to be pursued in earnest. The pursuit took place largely under the waters of Lake Maraicaibo in Venezuela and more extensively in the marshes and the shallow, protected waters of the Gulf of Mexico. By far the greatest development took place along the Louisiana deltaic plain—the lands that were formed by the delta of the Mississippi River—once the development of nondestructive exploratory (seismic) techniques allowed oil companies to locate and explore the numerous salt domes along the Louisiana and Texas coasts of the Gulf of Mexico.

At the time, it may not have occurred to many of the oil companies that explored the Louisiana marshes, or the people who lived nearby, that they were involved in the early stages of developing what would become one of the great industrial innovations of the 20th century. What almost certainly did occur to them was that it was no simple matter to drill oil wells through the unstable soils of the region, especially in that the best oil prospects were often situated in marshes, swamps, and the shallow waters of bays and lakes. The initial techniques involved the construction of a platform on pilings that were driven into the coastal marshes, with drilling equipment brought to the site on barges and installed on the platform. From there, drilling proceeded much as it would have been done on land, except that transportation to and from the site was often by boat, and additional costs were created by the need to build a platform for each exploratory well. In addition, pilings driven into the thick layers of silt have no solid foundation, relying entirely on the friction between the pilings and the mucky bottom. Since drilling machinery produces significant vibrations, up to 250 pilings were needed for a single exploratory platform. As interest in deeper prospects increased, these early techniques became too costly, since the nonrecoverable cost of platform construction was a constant regardless of the outcome of the drilling.

The technological breakthrough that facilitated large-scale exploration came in 1933 when the Texas Company (later Texaco) introduced the "submersible" drilling barge: A drilling rig was simply mounted on a barge, which could be towed to a drilling site and flooded. Sitting on the shallow bottom, the barge provided a stable base for drilling; once the job was done, the barge could be refloated and towed to a new location. Lake Pelto, Louisiana, behind Last Island, east of the mouth of the Atchafalaya River, was the site of the first successful test of the technology, opening a new era in marine drilling. Within a decade, these rigs were in use throughout the marshes and shallow coastal waters in Louisiana and Texas, sparking a growing support sector for marine drilling.

Throughout the 1930s and 1940s, the submersible drilling rigs evolved, becoming larger and more powerful, allowing them to tap the deeper oil-bearing formations along the Gulf Coast. Steam engines were replaced by diesel-electric rigs, and the tungsten carbide drilling bit was introduced in 1952. By 1955, inland submersibles were setting world drilling depth records of 22,500 ft. Much of the initial exploration, drilling, and production occurred in eastern St. Mary Parish, Louisiana, and residents of nearby towns, primarily Morgan City and Berwick, appear to have been eager to help provide labor and support to the oil firms that were drilling in the challenging areas nearby.

As inland drilling rigs were evolving, the focus was shifting offshore. In 1946, the Magnolia Petroleum Company (which later became part of Mobil), using pilings driven into the sea bottom, constructed a platform in the open Gulf on a site leased from the state of Louisiana. The site was south of Morgan City, 5 miles from the nearest land. Although the attempt was an economic failure (a dry hole), it was the key forerunner of what most observers today would consider to be a truly offshore effort, and it was a technological success in that it demonstrated that exploration in the open Gulf was feasible.

Following Magnolia's lead, and using the same type of technology, Kerr–McGee established the first producing well on the outer Continental Shelf in 1947, 12 miles off the Louisiana coast, almost due south of the mouth of the Atchafalaya River, on what is now Ship Shoal Block 32. This sparked additional interest offshore, and by the middle of 1948, in addition to offshore production by Kerr–McGee, 13 locations were in various stages of development off the coasts of Louisiana and Texas and plans were under way for drilling in 14 other locations. Within the next decade and a half, production platforms would be found in more than 200 ft of water, although the offshore bottoms of the Gulf of Mexico were sufficiently gradual that these platforms were being placed many miles offshore.

3. THE U.S. TIDELANDS CONTROVERSIES

The virtual explosion of offshore drilling activities during this era provided the backdrop for what came to be known during the Truman administration and

afterwards as the Tidelands controversies. These controversies, which continued until the subsequent Eisenhower administration pushed two new laws through Congress, were focused squarely on offshore oil drilling, but contrary to what many present-day readers might expect, they had almost nothing to do with public opposition. Instead, they had to do with what political body would reap the benefits of offshore oil—the states or the federal government.

In addition to the oil-friendly states of Texas and Louisiana, which were engaged in legal battles with the federal government, California and Florida—despite the intense opposition that both states would later express toward offshore oil and that at least some areas of California had shown a half century before—also showed strong interest in the offshore oil industry. As one observer informally summarized the tenor of that earlier time, looking back from a perspective of three decades later, "The basic attitude was, 'Good luck—hope you find some oil—and if you do, send us a check.'" The main concerns of the time had to do with whether the regulations and the public revenues from offshore oil would be the business of the federal government or the states.

Contrary to the Truman administration, which had inspired the controversies by favoring federal control, the Eisenhower administration espoused a "states' rights" position. This position aligned well with the preferences of the oil companies (which expected more favorable regulatory policies from the states than from the federal government) and with the relevant states (which were mainly located along the Gulf of Mexico and which were keenly interested in the public revenues that oil development could generate). The Eisenhower administration's first piece of legislation thus inspired considerable support from the relevant states and the oil companies by giving the states the legal and financial rights to all oil developments taking place within 3 miles of their shores. The second piece of legislation, which appears to have been the focus of much less attention at the time, gave the federal government the rights to the undersea lands that were farther away from the continent, along what came to be called the outer Continental Shelf, or the subsea lands that were out beyond the 3-mile limit of the states' jurisdictions.

4. THE INTERNATIONAL GROWTH OF OFFSHORE TECHNOLOGIES

Throughout the 1960s and early 1970s, most of the technology and the expertise still came from the Louisiana and Texas shores of the Gulf of Mexico. In 1968, for example, the Glomar Challenger, a state-of-the-art drill ship from Orange, Texas, began a series of worldwide trips for the Scripps Institute, taking deep-bottom cores for research purposes and using technology that would have been inconceivable only a few years before. Before the trips were over, the Glomar Challenger would drill a hole almost 3000 ft deep in an ocean-bottom location that was more than 16,000 ft below the surface of the sea.

As exploration and production continued to expand, however, offshore energy development came to be an increasingly worldwide phenomenon. By the mid-1960s, exploratory drilling had begun in the Persian Gulf, along the western shores of Africa (especially off Nigeria), in the Far East (Indonesia, Japan, Borneo, and Australia), and along the Cook Inlet in Alaska. By 1967, 75 countries were exploring for offshore oil and gas, and 20 of these were producing oil and/or gas offshore.

It was the North Sea experience, in particular, that may have provided the clearest early warnings of what would follow. Interest in offshore petrochemical extraction from the North Sea had begun in the late 1950s, just a few years after the passing of the Eisenhower-era legislation in the United States, and growing significantly in 1959 with the discovery of the Groningen gas field off Holland. By 1962, when the size of the Groningen field was better understood, that region saw a multimillion-dollar drilling effort, which was in full swing by the mid-1960s, with the first North Sea production in 1967. As offshore drilling efforts continued to expand, however, the offshore industry found itself in waters that were literally as well as figuratively deeper and more hostile. The loss of the drilling rig Sea Gem in the North Sea in 1965 indicated the dangers associated with the new environment, and increased emphasis was placed on adapting the Gulf of Mexico technology to more hostile conditions.

Particularly after the 1969 Santa Barbara oil spill just a few miles away from the first offshore oil derricks—in a location where the Nixon administration had insisted on proceeding with oil production despite local worries about just the kind of spill that actually occurred—the major oil companies devoted increasing attention to other areas of the world, reinforcing the already aggressive efforts by drilling and support companies that were searching for offshore discoveries. By the mid-1970s, for the first time since the earliest days of offshore drilling,

the majority of the world's offshore drilling operations were located in countries outside the United States, with approximately 75% of the activity in other countries, and with the North Sea being the prime area.

As previously mentioned, however, in testing the waters of the North Sea, oil companies encountered much colder temperatures than they had experienced in the Gulf of Mexico, politically as well as climatically. Rather than being able simply to "send a check" after finding oil, the oil companies (which in the case of Norway included a state-owned oil company) needed to spend years in negotiations with the citizens and leaders of the region. In both of the nations exercising political jurisdiction over key areas of the North Sea—Norway and Great Britain—affected citizens expressed a range of concerns and demonstrated a level of political resolve that would have been almost unimaginable along the Gulf of Mexico in earlier times. In a nation that takes great pride in its long history of Scandinavian distinctiveness, its widespread public participation, and its careful planning for environmental preservation, Norway understandably provided a very different context for operations from that of the oil industry's previous experiences in the Gulf of Mexico. Similarly, in Great Britain (particularly in Scotland and the Shetland and Orkney Islands, which would have had the most to gain or lose from offshore development), there were extensive discussions of the kinds of conditions under which the development of North Sea oil would be allowed to proceed.

The net results of the extensive discussions in both Norway and Great Britain included numerous measures for environmental protection, for local as well as national economic benefits, and for local socioeconomic impact mitigation, which were seen as innovative or even revolutionary by many analysts, especially against the backdrop of previous U.S. experiences. Norway, in particular, emphasized the combining of environmental protection with economic development measures (including a state-owned oil company, but also a number of private enterprises) that eventually led the country to become one of the world's leaders in offshore technology, allowing it to emulate the Gulf Coast pattern of selling not only its oil but also its technology. Several years later, when development began on the giant Hibernia oil field off the coast of Newfoundland, Canada, most observers saw far more similarity to the North Sea model than to previous experiences along the Gulf of Mexico.

5. THE EVOLVING RANGE OF VIEWS THROUGHOUT THE WORLD

There is no single or universal attitude toward offshore development. Instead, as can be seen by contrasting the regions that represent the two most extreme ends of the continuum—a task to which we will soon turn—the attitudes that are encountered are likely to reflect the specific physical, historical, and socioeconomic conditions that are in effect for any given place and time. To the extent to which there are shared characteristics across locations, most of the main reasons for supporting oil developments in a given locality are economic ones, and most of the reasons for opposing the same developments are environmental ones. Even this generalization, however, is oversimplified. As discussed later, many of the regions that have hosted oil development have had fewer economic benefits than they expected. However, at least in the decades since the Santa Barbara oil spill of 1969, the most often feared environmental problem of offshore drilling, the oil spill, has proved to be quite rare—at least in offshore drilling and production, as opposed to the far more numerous spills associated with the transportation of oil, particularly the well-known tanker spills from the Exxon Valdez, the Braer, and other such ships. Particularly in industrialized nations, the environmental and economic challenges and problems that are of greatest concern to nearby residents may have little to do with oil spills, having to do instead with the onshore activities that support the offshore platforms. Some of these activities and businesses, including those involving diesel mechanics, suppliers of drilling muds, and the many tons of pipes and other supplies, can conflict with other ways of using prized coastal locations, many of which are also quite important economically as well as esthetically, with uses ranging from fishing to high-value homebuilding.

It is important to be particularly cautious, however, about drawing many firm conclusions about public attitudes in less developed countries, because systematic research on attitudes in most such countries is relatively rare. Far from displaying high levels of concern or sensitivity toward local attitudes, in fact, many of the less prosperous nations that have developed their oil resources have shown tendencies toward political repression that many observers have found disturbing. So pronounced is this tendency—onshore as well as offshore—that after examining cross-national data from 113 nations, Michael Ross published his findings under the title "Does Oil Hurt

Democracy?" The answer, he concluded, appears to be in the affirmative.

In broad terms, to the extent to which the reactions of many different nations have been examined, the most common focus has involved the financial and, to a lesser extent, the political implications of offshore development. As a useful simplification, three main patterns have emerged across time. The first and by far the most common pattern in some respects reflects a combination of the earlier tidelands controversies in the United States and the efforts by Norwegian leaders to use planning and politics to extract greater economic benefits from the vast flows of revenue that are involved in offshore oil development. The second involves an underlying pattern of concern regarding the kinds of democracy-harming tendencies noted by Ross. The third pattern provides an ironic contrast against the first, involving a problem that is sometimes called Dutch disease, as will be discussed later.

In Venezuela, the discovery of oil under the waters of Lake Maraicaibo in the early 1920s was followed by more than a half century of economic growth, during which oil revenues were used to provide a wide range of economic development benefits. For most of the next 50 years or more, Venezuela was seen by most observers as one of Latin America's most notable models of democratic stability and social peace, and the common conclusion was that the political as well as the economic benefits could be traced in large part to oil-based prosperity. Since the oil price collapses of the 1980s, however, Venezuela's political stability has been seen by some to be as questionable as its prosperity, with clear signs of tension between the middle/upper classes and Hugo Chavez (the current President at the time of this writing in 2003). Recent assessments are much more likely to call attention to class warfare than to social stability.

A similar pattern is visible in Indonesia, which provided the lion's share of "Dutch" oil for Royal Dutch Shell for nearly half a century, during which time the former colony was generally known as the Dutch East Indies, before achieving independence after World War II. For a period of several decades after the former colony became an independent nation, its leaders were seen by most observers as showing considerable political acumen in their efforts to obtain economic and developmental benefits from indigenous oil resources. Rather than attempting to "nationalize" or take over the oil companies that were operating in Indonesia at the time, Ibnu Sutowo—a close personal associate of Presidents Sukarno and Suharto and the key force behind Indonesia's state-owned oil company, Pertamina—emphasized a more gradual approach. One component involved the motto "Learn while you work, work while you learn" for the Indonesian nationals who continued to work in the oil fields until the old patterns of colonial dominance (and the old colonial Mining Law of 1899) were finally brought to an end. Another, more tangible component involved the relatively gradual process of ending the earlier patterns, beginning when Article 33 of the new Constitution of 1945 declared all of Indonesia's natural resources to belong to its people but not fully completed until approximately 20 years later, when the Petroleum Act of 1960 and then the production sharing contract arrangement of 1966 ultimately came into effect.

Even after Indonesia passed its new law in 1960, the major oil companies were not taken over by the Indonesian state; instead, the companies experienced a transition from a previous legal status of "concessionaire" (i.e., enjoying concessions that had been granted under Colonial Dutch rule) to a new status of being contractors to the state enterprise, Pertamina. The transition was completed after Sutowo and the Indonesian state instituted production sharing contracts, which required foreign contractors to bear the full cost of exploring for oil but reserved approximately 40% of the oil for Pertamina and the Indonesian government. Under other contract provisions, most of the lands that were leased to the oil companies reverted back to the government within a decade or less, and the Indonesian share of the profits would increase as world oil prices increased.

Not surprisingly, it appears that at least the financial implications of Indonesia's increasing control over its oil and the associated revenues won considerable praise from most Indonesians for many decades. Indonesia has also been seen by many scholars as an exception to some of the threats to democratic development that are of concern to authors such as Ross. The oil revenues were reinvested in a wide range of other enterprises that were seen as contributing both to the development of the growing national economy and to the integration of what is not only the most populous petroleum-exporting nation but also the most heavily populated Islamic nation on Earth. This latter point is a major one, given that Indonesia embraces what most Western observers find to be a surprisingly diverse and far-flung collection of peoples. It encompasses more than 1000 islands and hundreds of native languages spread across a distance of more than

3000 miles of land and water. Through investments in everything from schools to steel factories, Indonesia used its oil revenues as an important resource for building a stronger and more prosperous nation.

Particularly in the later years of the Suharto era, however, Indonesians became increasingly critical of the "crony capitalism" and corruption that appear to have become increasingly pervasive, both in Pertamina and in the broader society, and of the increasing extent to which the income derived from offshore oil and gas fields was used to support military repression. As suggested by the fact that some of the more troubled areas of Indonesia also happen to have rich offshore petroleum deposits—including Aceh, East Kalimantan, and the now-independent nation of East Timor—the degree of Indonesian "exceptionalism" may be less than previously thought.

Finally, off the western shores of Nigeria, very few observers have seen the kinds of exceptionalism once thought to exist in Indonesia and Venezuela. Instead, particularly after what the Nigerian government characterized as the "execution" of the noted human rights activist, Ken Saro-Wiwa, there has been extensive international attention to what many have seen as heavy-handed and antidemocratic central government repression of local opposition to offshore oil development and its attendant onshore impacts, particularly among the native people of the area, the Ogoni.

Economic benefits have sometimes proved elusive, as well. Although the vast sums of money involved in offshore oil development often lead to expectations for rapid economic growth in oil-exporting countries, careful analyses have tended to find much less desirable outcomes, economically as well as politically.

Perhaps the best known description of the economic problems is the notion of Dutch disease. Given its catchy alliteration, this phrase has been attached to almost any example of economic weakness in the face of resource exports, but more formally the phrase is understood to involve the combination of two factors that appear to have led to a generally unexpected slowdown in the economy of The Netherlands, even in the face of that country's increased revenues from its own North Sea petroleum reserves. First, a sharp increase in exports led to a significant increase in the strength of the national currency—a fact that meant other exports from the country became more expensive for people who lived outside The Netherlands, reducing international demand for Dutch products. Second, the booming resource sector was seen as drawing people and capital away from other sectors of the economy, such as manufacturing and agriculture, raising the effective production costs in the other sectors over time. These concerns led to deliberate efforts, particularly in Norway, to offer increased inducements for investment in non-oil sectors—efforts that may well have been warranted in any case. However, subsequent analyses have indicated that unexpected combinations of rich resources and poor people appear to be more widespread than can be explained by the Dutch disease hypothesis alone.

6. FACTORS ASSOCIATED WITH DIFFERING VIEWPOINTS IN THE UNITED STATES

There is no one or universal pattern of public attitudes toward offshore oil development. Instead, attitudes vary widely in response to local conditions and the nature of oil development that is proposed. Of all the places that have been studied systematically, however, it appears that the widest variations in attitudes are found between two different areas of the same country—northern California and southern Louisiana. It is clearly not the case that the opposition is a straightforward function of the environmental damage that has been created in each region. Indeed, with the possible exception of the Santa Barbara oil spill (which took place along California's south-central coastline, in any event), the environmental impacts of offshore oil development have actually been far more significant and widespread in southern Louisiana than in northern California. Instead, the variations in public attitudes between these two regions tend to illustrate the broader points we are making in this article, reflecting the historical, physical, and social characteristics of the regions in question. Thus, a comparison of the two regions provides a useful way to bring this article to a close.

Historical factors are important both because of changes in broader historical factors and local cultural characteristics, and because differing regions experience an activity (in this case, offshore oil) during different periods of that historical transition, in effect resulting in different local histories. When oil exploration started in Louisiana's coastal marshes in the 1920s, the prevailing attitude toward the environment almost everywhere in the United States was an exuberant one, involving the "conquest of nature" and the fresh legacy that remained from the mythic

conquest of the western frontier. Although environmental concern was not completely unknown, it certainly had little of the character or strength that it would have a few decades later. In nearby Florida, for example—where later decades would see a level of organized opposition to offshore oil drilling comparable to that of California—essentially all of the offshore lands that were seen as potential oil-bearing prospects were leased to oil companies by the state of Florida well before anyone there had ever heard of Earth day.

In addition however, important aspects of the development of offshore technology were unique to southern Louisiana. When oil exploration began to move into the coastal marshes, the marshes were viewed not as ecologically vital wetlands but as hostile territory that needed to be subdued for human benefit. Louisiana also had a long history, dating back to the early 1800s, of resource harvest (fish, shrimp, oysters, and fur cypress lumber) in coastal areas. Furthermore, oil and these earlier extractive activities, especially fishing, proved to be highly compatible in several important ways. First, oil companies had no experience with coastal and offshore waters, but the fishing industry did. When Magnolia Petroleum Company drilled the first offshore well in 1946 south of Morgan City, boats from the local shrimp fleet were hired for the initial survey and the transportation of workers and supplies. Local people supplied the navigation knowledge for early offshore oil, and some of them later became crew and supply boat captains.

Second, the scheduling of offshore oil work allowed for individuals to maintain traditional coastal occupations. The movement into the marsh and estuaries of the Louisiana deltaic plain was large scale and relatively permanent. However, since it involved environments in which human settlement was not really possible, and in which distances and transportation difficulties meant that commuting had to be limited, an innovative form of concentrated work scheduling soon evolved. The common pattern for offshore work became one in which employees meet at a prearranged site to "go offshore" either by boat or, recently, by helicopter. The stay offshore varies, but is typically 7, 14, or 21 days. Following the stay offshore, the employees return to the meeting site and have a period of time off, typically the same length as the stay offshore. Because of this schedule, many such workers chose to continue their traditional occupations. This meant that boats, gear, and knowledge were maintained and did not fall into disrepair, as might have occurred if oil jobs in Southern Louisiana had been scheduled for five days a week. When the oil patch crashed in the mid-1980s, many coastal residents found it useful to fall back on these skills and equipment to feed their families.

In addition, oil development in Louisiana started gradually, and it grew as the product of local innovation, which understandably generated considerable pride. In contrast, when the first offshore development appeared in California in the late 1960s, not only was environmentalism emerging as a national trend but also the huge rigs that were by that time associated with offshore oil were seen by a largely urban population as foreign and threatening. When the Nixon administration insisted on moving ahead with offshore leasing just a few miles away from the beaches of Summerland and Montecito in the face of stiff local opposition, and after the Santa Barbara oil spill a few years later provided a tangible indicator of the validity of local concerns, Californians' opposition to further offshore oil activities became all the more intense.

Physical factors, or to use the more technical terminology, coastal geomorphology, can also be seen to differ quite distinctly between the two states. There are fundamental differences between the Louisiana and California coasts, both in coastal topography and in marine topography. The coastal regions of Louisiana differ from those of California, as well as from most other areas of the United States, in three main ways: Coastal wetlands limit land-based access to the coast, estuaries offer many opportunities for harbor space, and low relief and low energy levels characterize most of the coast. As a result, the distribution of populations and roadways in Louisiana is very different from that found in most coastal states. Few of the state's population lives on or near the coast and, with the exception of Grand Isle, it is often almost impossible to get anywhere near the coast by road. Most of the coast is lined with a broad and virtually impenetrable band of coastal wetlands that are far better suited for the abundant fish and wildlife of the region than for human habitation. In most coastal states, especially Florida and California, where proposals for outer Continental Shelf (OCS) development have met with intense opposition, the situation is completely different. Most of the population lives on or near the coast, and virtually all of the coast is readily accessible by road. In California, Florida, and much of the rest of the coastal United States, the coast is seen as a valuable public resource, a thing of beauty, and a source of popular recreation—in part because the coast is actually seen by so many people, so often.

Because of the coastal wetlands in Louisiana, not only is the coast rarely seen, but local residents' descriptions of their state's coastal regions are more likely to involve alligators and mosquitoes than spectacular scenery.

The second key feature of Louisiana's coastal topography is the presence of an extensive estuary system. Although it is difficult to reach the Louisiana coast from land, it is easy to do so from the water. Louisiana is characterized by numerous waterways that intersect the highway network farther inland, and that provide coastal access for marine interests. To those familiar with other coastlines, the concept of harbor is probably a common one. A harbor is generally thought of as an enclosed, protected water body with orderly docks, designed to maximize the number of boats that can be stored in the limited protected waters. Outside of the New Orleans metropolitan area, the concept is little used in coastal Louisiana, where people simply live among their boats. There is no shortage of places to put a boat in coastal Louisiana. The natural levees of the many bayous that traverse the area provide innumerable places both to dock boats and to live. People do both. The geomorphology of the Louisiana coast even allows oil field and fishing vessels to use many of the same facilities. There have been a few cases, such as Morgan City in the boom years of the 1970s and early 1980s, in which the demand for dock space became so intense that the shrimpers were crowded out. Generally, however, fishing and oil field boats have in the past, and do today, coexist peacefully. Along the coast of California, by contrast, most of the available harbors are small, and dock space is quite limited. A 250-ft offshore supply vessel represents a considerable threat to local fishing interests.

The third key feature is that Louisiana is a region of extremely low relief and generally low energy. When Louisiana residents do reach the shoreline of their state, they see flat water and flat land, with comparatively low-energy beaches. They see water that is seldom clear because of the discharge of silt by the Mississippi and Atchafalaya Rivers. In areas of the country that are farther removed from the "muddy" Mississippi, including the areas to the east in Florida, the waters are generally clearer. In northern California, residents frequently observe the power of the sea where it meets the coast, and they often comment on it with what they call "awe." In contrast, the Gulf of Mexico off of Louisiana is generally a more sedate body of water, and even when the Gulf does inspire awe, such as during tropical storms and hurricanes, its displays have rarely been seen by people who are alive today. As noted previously, only a small percentage of the population lives on the Gulf, and a combination of modern weather forecasting and centuries of experience have led to routine evacuations of coastal and low-lying areas as tropical storms approach.

The distinctiveness of the Louisiana environment, however, also extends offshore and to what is literally below the surface—that is, the marine topography. Most people in Louisiana do not see offshore oil as an issue, and many are particularly bewildered by the fact that fishing interests in other locations have expressed so much concern over offshore developments, given that marine use conflicts have been so notably absent in the Gulf. Many fishing boats, both commercial and recreational, will actually tie up to the oil rigs to fish, and in New Orleans the Aquarium of the Americas includes a large display, complete with the title "From Rigs to Riches," pointing out the advantages of oil platforms as a form of habitat for many of the fish species of the Gulf. Given the Louisiana experience, why would there be so much opposition to oil structures from fishing interests in other regions? The reasons are not so obscure as they first appear. In addition to the historical factors previously noted, there are two key aspects of the marine environment that have helped to limit the potential for conflicts between OCS development and other marine activities, particularly fishing, off the coast of Louisiana. One has to do with the gradual slope of bottoms along the central Gulf of Mexico, and the other has to do with the presence of silt bottoms that limit the number of obstacles likely to be encountered.

Off the coast of Louisiana, the gradual slope of the Continental Shelf presents very different conditions from those found in areas where offshore oil development has been most contentious. In most areas of the country, particularly along the Pacific Ocean, the Continental Shelf drops off much more dramatically into the ocean basin, meaning that it is also much narrower than the shelf that exists in the Gulf.

As a result of this difference in sea-bottom slopes, the available area of the Louisiana Continental Shelf is far larger than is the case in California, reducing significantly the probability of use conflicts. The gradual slope also reduces the number of problems that are likely to be created by any given obstacle. Even if a fishing boat needs to make a quarter-mile detour around oil operations, for example, there will be little significant impact on the boat's ability to keep its nets in contact with the sea floor. In contrast,

in areas such as the Pacific Ocean off the California coast, the steeper slope makes the Louisiana experience largely irrelevant in two ways. First, the actual area available for use is smaller, meaning that even an apparently small loss of area for OCS activities can have a proportionately major impact on the area available for fishing in a given region. Second, given that bottom-dragging operations need to work along a contour line, following sea bottoms at a given depth, the presence of an oil platform can mean that fishing boats would need to make a detour that would correspond to a difference of several hundred feet in water depth, effectively precluding the option of "fishing around" many such structures.

An additional feature of the marine environment of the Gulf of Mexico involves the presence of silt bottoms. Although the heavy silt discharges by the Mississippi and Atchafalaya Rivers mean that the water is seldom clear, reducing many concerns over esthetic impacts, the nature of the bottom also means that fishing operations encounter few obstacles such as rock outcroppings that would lead to the loss of nets and other gear. In regions such as California, the frequent presence of rocky outcroppings can severely limit the ability of fishing boats to change their trawl runs.

One consequence of the difference in bottom types is that, to reiterate a point that often comes as a surprise to the opponents of oil development in other regions, oil rigs actually can provide a significant advantage for fishing operations on silt bottoms of the Gulf of Mexico. Certain types of commercially important fish can only survive in the kinds of habitat known collectively as hard substrate—rocky bottoms, reefs, rock outcroppings, and the like. In the central Gulf of Mexico regions where oil development activities have been the most intense, natural outcroppings of this sort are so rare along the predominantly silty bottoms that oil-related structures now comprise approximately 25% of all hard substrate. In effect, the oil rigs serve as artificial reefs, concentrating and probably increasing the fish populations, which is precisely why it is quite common to see fishing boats literally tied up to oil rigs in search of fish.

In summary, in Louisiana the coastline is inaccessible to most land-based populations, and accordingly low in social salience, while offering enough potential harbor space to meet the needs of offshore oil development as well as potentially competing uses such as fishing operations. The offshore seafloors, meanwhile, tend to have such gradual slopes as to offer vast areas of virtually level bottoms, relatively free of obstacles but also so devoid of natural reefs that the oil rigs provide a valuable service for fishing operations. As was the case for the historical factors, most of these characteristics are almost precisely reversed for the coastal regions of much of the nation, particularly the northern California coast; the very considerations that have contributed to the ready acceptance of offshore oil in Louisiana tend to exert the opposite effect in California.

Social factors provide a third broad category of reasons why the attitudes toward offshore development can differ so widely, even within the context of a single country. As a useful simplification, four social factors appear to have encouraged the easy acceptance of offshore oil in Louisiana relative to what might be expected in other areas of the United States: the average educational levels, the patterns of social contacts, the importance of prior extractive industries, and the potential for overadaptation that characterized the coastal regions of Louisiana as offshore activities were first being developed.

Given that support for environmental protection is quite strong in the United States, as it is in most countries of the world today, very few sociodemographic characteristics show strong correlations with environmental concerns. For example, studies show that blacks are as supportive of strong environmental controls as whites, and poor people tend to be as supportive as wealthier people. One consistent exception, however, has to do with educational levels, with better educated persons generally expressing higher levels of environmental concern. Thus, it may be significant that, particularly as offshore development was picking up speed in the 1930s and 1940s, coastal Louisiana had some of the lowest educational levels in the country. In St. Mary parish, the scene of initial offshore activity, only 47.2% of the adult population had 5 years of education in 1940, and only 12.2% had graduated from high school. Other rural areas of southern Louisiana had similarly low educational levels. In comparison, more than 78% of adults in the United States had a high school education or more by the 1990 Census.

Second, the other industries that most characterized coastal Louisiana at the time of initial offshore development were ones that involved extractive uses of the coast. Like oil development, they involved the extraction of raw materials from nature. Local residents obtained products from the Atchafalaya Basin (cypress lumber, fish, crawfish, water fowl, and moss for furniture stuffing) and from the coastal marsh (furs, shrimp, and oysters). The export of such raw materials had provided the mainstay of the

economy in coastal Louisiana for almost a century prior to offshore development. In areas where extractive activities are dominant, as in Louisiana in the 1940s and 1950s, there tends to be less resistance to additional extractive activity. In most coastal regions of the United States today, and certainly in those regions of California where proposals for offshore oil have been met with the stiffest resistance, the economy has come to be far more dependent on the amenity values of the coast than on its extractive values. The proportion of the population depending on the extraction of coastal resources is small and shrinking, whereas not just tourism but also other amenity-related forms of development, including many high-value economic activities, depend on high levels of environmental amenities. These growing economic interests place high importance on maintaining relatively pristine environmental conditions. Thus, in California's coastal areas, and increasingly in other areas as well, the likelihood of finding support from extractive workers can be expected to continue to decline.

Third, social interaction patterns can exert powerful influences on individuals' attitudes. One particularly powerful, if often overlooked, pattern involves what we have called a social multiplier effect: Even if a given individual does not work in the offshore oil industry, that person's attitudes may be affected by whether his or her friends and relatives do. Given the historical, biophysical, and other social factors summarized previously, the "average" resident of coastal Louisiana in the 1940s would be expected to have known many friends and neighbors who were employed in the oil industry. By the 1980s or 1990s, it was virtually impossible to live in southern Louisiana and not know someone who was employed in an oil-related enterprise. In addition, as reported in the 2000 Census, Louisiana residents were more likely to live in the state in which they were born than were residents of any other state. In contrast, in most of the coastal regions of the United States today, the situation is the opposite. Not only do these regions have few residents who are involved in oil or gas extraction but also they have high numbers of people who have moved to the coastal regions precisely because of their desire for high environmental quality. This tendency is especially likely in areas such as northern California, where there is no current base of employment directly dependent on local oil and gas development.

Finally, an additional way to understand the uniqueness of Louisiana's compatibility with offshore development is to examine the degree of adaptation that has taken place. As the various components of the human environment become adapted to a given form of development activity, there is a tendency for new skills, knowledge, tools, networks, and other resources to be built up around that activity. Three potentially problematic results, however, can also occur. First, any type of development narrows a region's options because time, resources, and human capital are devoted to a particular developmental scenario, sometimes limiting the options for alternative scenarios. We call this process developmental channelization, although others have used terms such as path dependency to denote similar patterns. Second, as part of this process, adaptations to earlier forms of development may be lost (sometimes quickly and sometimes across generations). Third, nearby communities can run the risk of overadaptation to the new development activity. Particularly if that new activity is not a sustainable one, then when it ceases or declines, communities or regions may be left in the position of being less able to survive in their environment than they would have been before the new development occurred. Thus, to a large extent coastal Louisiana

TABLE I

Factors That Led to Positive Evaluations of Offshore Oil in Louisiana[a]

Historical factors
Early historical era—"age of exuberance"
Temporal priority of development
Incremental onset and evolution of industry
Local origins and development of technology
Biophysical factors
Coastal topography
Broad coastal marshes that preclude coastal highways and population concentrations
Estuary systems that include many potential harbors
Low relief and energy levels
Marine topography
Broad, gradual slopes that increase area and decrease significance of conflicts with platforms
Silt bottoms with few outcroppings and obstacles
Social factors
Low educational levels
Extractive orientation toward biophysical environment
Favorable patterns of contact with oil industry personnel
Extensive prior adaptation to oil development

[a] Adapted from Freudenburg and Gramling (1994).

has become not just adapted to but also dependent on its offshore petroleum activities, and it is difficult to be too critical in evaluating an activity from which one has benefited, particularly if the activity is one on which one remains dependent for any hopes of comparable future prosperity. This, coupled with the industry's generally good record in terms of offshore environmental accidents, has tended to bolster the perception of offshore activity in the Gulf as involving relatively low levels of risk. Table I summarizes these factors. As a simplified but generally helpful rule of thumb, where more of these factors are present, attitudes toward offshore development can be expected to be more positive.

SEE ALSO THE FOLLOWING ARTICLES

Crude Oil Releases to the Environment: Natural Fate and Remediation Options • Oil and Natural Gas Drilling • Oil and Natural Gas: Offshore Operations • Oil-Led Development: Social, Political, and Economic Consequences • Public Reaction to Electricity Transmission Lines • Public Reaction to Energy, Overview • Public Reaction to Nuclear Power Siting and Disposal • Public Reaction to Renewable Energy Sources and Systems • Wetlands: Impacts of Energy Development in the Mississippi Delta

Further Reading

Ahmad Khan, S. (1994). "Nigeria: The Political Economy of Oil." Oxford Univ. Press, Oxford, UK.

Andersen, S. S. (1993). "The Struggle over North Sea Oil and Gas: Government Strategies in Denmark, Britain, and Norway." Scandinavian Univ. Press/Oxford Univ. Press, New York.

Baldwin, P. L., and Baldwin, M. F. (1975). "Onshore Planning for Offshore Oil: Lessons from Scotland." Conservation Foundation, Washington, DC.

Barnes, P. (1995). "Indonesia: The Political Economy of Energy." Oxford Institute for Energy Resources, Oxford, UK.

Freudenburg, W. R., and Gramling, R. (1994). "Oil in Troubled Waters: Perceptions, Politics, and the Battle over Offshore Drilling." State Univ. of New York Press, New York.

Gramling, R. (1996). "Oil on the Edge: Offshore Development, Conflict, Gridlock." State Univ. of New York Press, New York.

Gramling, R., and Freudenburg, W. R. (1996). Crude, coppertone, and the coast: Developmental channelization and the constraint of alternative development opportunities. *Soc. Nat. Resour.* **9**, 483–506.

Jones, R. E., and Dunlap, R. E. (1992). The social bases of environmental concern: Have they changed over time? *Rural Sociol.* **57**(1), 28–47.

Molotch, H., Freudenburg, W. R., and Paulsen, K. (2000). History repeats itself, but how? City character, urban tradition, and the accomplishment of place. *Am. Sociol. Rev.* **65**(6), 791–823.

Okome, O. (ed.). (1999). "Before I Am Hanged: Ken Saro-wiwa—Literature, Politics, and Dissent." Africa World Press, Trenton, NJ.

Ross, M. L. (2001). Does oil hurt democracy. *World Politics* **53**, 325–361.

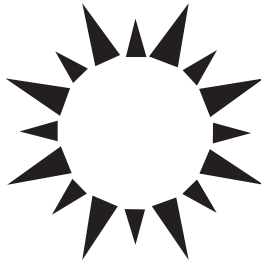
Seda, F. S. S. E. (2001). "Petroleum paradox: Natural resources and development in Indonesia, 1967–1997." Unpublished PhD dissertation, University of Wisconsin, Madison.

Seyfrit, C. L. (1988). A need for post-impact and policy studies: The case of the Shetland Experience. *Sociol. Inquiry* **58**, 206–215.

Smith, D. A. (1994). Uneven development and the environment: Toward a world-system perspective. *Humboldt J. Soc. Relations* **20**(1), 151–173.

Storey, K. J., and Shrimpton, M. (1987). "Planning for Large-Scale Construction Projects: A Socio-economic Guide for Communities, Industry and Government." Environmental Studies Research Funds, Ottawa, Ontario, Canada.

Public Reaction to Renewable Energy Sources and Systems



TIMOTHY C. COBURN
Abilene Christian University
Abilene, Texas, United States
BARBARA C. FARHAR
University of Colorado
Boulder, Colorado, United States

1. **Definition of Renewable Energy**
2. **Political, Philosophical, and Economic Perspectives**
3. **The Evolving Consumerism toward Renewable Energy**
4. **Government Support for Renewable Energy Initiatives**
5. **Trends in the Demand for Renewable Energy and Renewable Energy Systems**
6. **Consumer Attitudes toward Renewable Energy and Renewable Energy Systems**
7. **Stated Willingness to Pay for Electricity Derived from Renewable Energy**
8. **Gauging Actual Participation of Consumers in Renewable Energy Programs**
9. **Public Response to Renewables in the Transportation Energy Sector**
10. **Factors That Motivate Public Response to Renewables**
11. **Modeling Public Response to Renewables**
12. **Conclusion**

Glossary

biomass Organic materials such as woodchips, twigs and stems, grasses, weeds, aquatic plants and other quick-growing plants, and animal wastes that can be used to produce fuels or electricity through a process of fermentation, distillation, or incineration.

cogeneration The simultaneous production of several types of energy; a process that converts a fuel into both thermal and electrical energy.

ethanol Nontoxic alcohol produced from sugar, starch, or other carbohydrates through the process of fermentation; sometimes referred to as grain alcohol or ethyl alcohol; used as a fuel or fuel additive to improve oxygenation.

fuel cell A device that converts the energy of a fuel directly into electricity and heat without combustion but through the process of oxidation; originally developed for the space program and now largely intended for use in automobiles.

geothermal energy Energy in the form of natural heat originating from within the earth and contained in rocks, water, brines, or steam.

green power Electricity produced from renewable sources or in such a way that the process results in little or no adverse impact on the environment.

green pricing Strategies and formulas used by utility companies to recoup the costs of producing electricity from renewable sources, often by charging customers premiums or incremental amounts on their monthly utility bills; customer participation in green pricing programs is voluntary.

grid-tied photovoltaics (GPV) Photovoltaic systems designed to operate in parallel with and interconnected to the electric utility grid.

hydroelectric power or hydropower Energy produced by the force or pressure of falling water.

landfill gas A gas consisting mostly of methane that is generated as a by-product through the natural decomposition of organic material at landfill disposal sites.

methanol A toxic alcohol derived largely from natural gas or coal; sometimes referred to as the simplest alcohol, and other names include methyl alcohol, carbinol, and wood alcohol. It is used as an antifreeze, solvent, or fuel/fuel additive.

nuclear fission The process of splitting the nucleus of an atom into two parts, either spontaneously or as the result of an impact from another particle, resulting in the release of energy.

ocean energy Energy produced by the force or pressure of wave action.

photovoltaics (PV) Passive devices constructed from semiconducting materials that absorb sunlight and convert it into heat or electricity.

renewable energy Heat, electricity, or power produced from renewables or renewable sources, such as

 © 2004 Elsevier Inc. All rights reserved.

wind, the sun, subsurface steam, and water/wave action.

renewables Naturally occurring resources or materials used in energy, fuel, or electricity production that can be regenerated or replenished over time.

solar energy Energy from the sun that is converted into heat or electricity through passive collector devices.

sustainable In the context of energy resources, consumption in a manner that does not result in depletion and/or use of natural assets that are renewable; in the context of the environment, usage or treatment of Earth and its natural and physical assets in a manner that does not deplete, damage, or destroy.

wind energy Heat or electricity produced by the force or pressure of wind or air currents.

Whereas most energy streams are derived from sources that are exhaustible (e.g., fossil fuels), renewable energy is produced from sources that can be replenished over time. Faced with declining availability of some nonrenewable resources and with growing concerns about their environmental and public health impacts, society has become increasingly more attracted to renewable energy that exists in greater abundance and has minimal adverse physical effects. In response to a measurable shift in public opinion over the past two decades, many initiatives have been undertaken to expand the use of renewable energy sources and systems. The results of numerous polls and surveys indicate that citizens throughout the industrialized world now desire greater reliance on renewable and are willing to invest in their development.

1. DEFINITION OF RENEWABLE ENERGY

Renewable energy means different things to different people. Although there is little argument as to what energy is, even in its myriad forms, the term renewable energy conjures up a more diverse assortment of images.

In the simplest terms, energy is the capacity of a physical system to do work, or alternatively, to produce heat. Energy is a characteristic of the physics of a system. Renewable, on the other hand, pertains to the origin of the energy; thus, ideas and opinions about renewable energy are often confounded with information (or misinformation) about the source.

Renewable energy is energy that is derived from a supply that is constantly and naturally replenished over a relatively short time. Hence, any discussion of renewable energy is ultimately reduced to a discussion of renewable resources as they are derived from geophysical processes: sunlight, wind, falling water, sustainable biomass, wave motion, tides, and geothermics. Some entities, such as the Texas Renewable Energy Industries Association, describe all renewable energy as being directly or indirectly derived from the sun or natural movements and mechanisms of the environment. Others—private, commercial, governmental, and international—have their own definitions. One of the more visible characterizations has been provided by the U.S. Department of Energy (DOE), which segregates renewable energy into six components: bioenergy derived from biomass resources; geothermal energy derived from the earth's own heat-producing processes; hydrogen, hydropower, or hydroelectric power; ocean energy; solar energy; and wind energy.

2. POLITICAL, PHILOSOPHICAL, AND ECONOMIC PERSPECTIVES

Given the absence of a uniform definition, it is not surprising that the larger energy community is somewhat unclear about the role renewable energy plays in the broad arena of energy policy and economics, and that markets for renewable energy and related systems have yet to fully mature. In addition, because public perception is so often impacted by media coverage, or the lack of it, uncertainty and misunderstanding about renewable energy among the general public might be a perfectly legitimate response.

Renewable energy evokes many different metaphors and perceptions. Some people think of renewable energy as a nondepletable, environmentally friendly alternative to conventional fossil fuels, whereas others consider it to be an expensive, high-tech, and possibly unreliable proposition. Some feel threatened by renewable energy and view it as detrimental to current energy markets and jobs, but others (e.g., residents of California) believe it represents a sound investment for long-term economic and energy security.

For many years, renewable energy advocates have been unjustifiably associated with the more radical component of the environmental movement. Although this impression has moderated, the political and economic struggles over energy supply, demand, efficiency, and sustainability continue even today in countries throughout the world, and the role to be played by renewable energy, along with the concept of energy efficiency, is at the very heart of many of them.

One has to look no further than the ongoing conflict in the United States over drilling for oil in the Alaska National Wildlife Reserve (ANWR) to understand what is at stake with regard to jobs, markets, and the domestic economy. Hence, within the greater energy complex, the views and claims of renewable energy and energy efficiency advocates have sometimes been mistrusted—even totally dismissed—because, on the surface, they have seemed counter rather than complementary to conventional energy policy and wisdom. However, from a retrospective viewpoint, there has been a not-so-subtle shift in public and corporate attitudes toward renewable energy, particularly in the past decade. In the historical arena in which plentiful, inexpensive fossil energy was taken for granted, renewable energy was not even on the map. However, in the more contemporary worldview that is tempered by international conflict, global warming, economic uncertainty, environmental and public health concerns, and rolling electricity blackouts, renewable energy has a more prominent place at the table, even among those who are steeped in the fossil fuel tradition. As is the case for many things, it is a matter of perspective. It is the changing perspective regarding renewables within the energy community that holds such great interest from both the social and economic standpoints.

3. THE EVOLVING CONSUMERISM TOWARD RENEWABLE ENERGY

The posture of consumers is unquestionably changing. Although some politicians and corporate executives would still have the public believe that renewable energy is a bad investment, the more pragmatic and economically savvy consumers of the early 21st century are not buying the argument. It seems they are not quite as ill informed as they were once thought to be. First detected in the late 1980s, a distinctively proactive stance toward energy derived from renewable resources is now readily apparent in the results of almost all public opinion energy research. This change, which has been described as "the quiet revolution in renewable energy," has occurred at the grassroots consumer level with little fanfare and surprising determination, catching the corporate energy establishment somewhat off guard. Despite earlier technological missteps, consumers now see reliance on alternative energy sources and systems as fundamental to maintaining the social norms and economic well-being of society. Consumers seem poised to rapidly adapt their lifestyles to a wide range of alternative energy technologies, from wind turbines to fuel cell vehicles and rooftop solar electric panels.

Some of the attitudinal and behavioral change is undoubtedly due to initiatives by consumers to seek more balance in their lives, a greater degree of economic security, and more independence and control over their own futures. In addition, the post-baby-boom generation has simply been exposed to more technological innovation, and individuals in this age group have come to understand that technological expansion is a major economic force throughout the world.

4. GOVERNMENT SUPPORT FOR RENEWABLE ENERGY INITIATIVES

The change is also partly attributable to substantial government investment in research and development activities that serve not only as the catalysts for new ideas and equipment but also as incubators for fledgling industries that create jobs. Through government mandates, rebates, tax credits, and other incentives, use of renewable energy and renewable energy systems is proliferating at both the local and the national levels. In the United States, all federal agency vehicle fleets must now contain specified percentages of alternative fuel vehicles, and energy consumption at federal facilities and installations must include specified percentages from renewable sources. As an example of the results of such a proactive government stance, late in 2002 Dyess Air Force Base in Abilene, Texas, became the first U.S. facility to be fully dedicated to wind power. Programs, situations, circumstances, and events of this nature, being publicized through the media, definitely have a positive, although perhaps unquantifiable, impact on consumer awareness and behavior.

Throughout the United States and in other countries as well, government agencies have engaged in a fairly robust educational blitz aimed at promoting the development and use of renewable resources. Such educational initiatives begin at the elementary school level in an effort to establish an attitude toward renewable energy that gives rise to lifelong patronage and support. Numerous advocacy groups, such as the Colorado Renewable Energy Society, the American Council for Renewable Energy, the Clean Energy Funds Network, and Citizens for Renewable Energy in Canada, have taken up the charge and play a major role in the overall educational process. Increasingly more public and grassroots community events, such as the 2003 Renewable Energy Roundup

and Green Living Fair in Fredericksburg, Texas, and the 2004 Renewable Energy and Sustainable Living Fair in central Wisconsin, are being organized and promoted, although the total number of individuals who participate in these activities and their reasons for doing so have not been researched and are unknown.

In the United States, the federal government also provides grants directly to individual citizens and small businesses to encourage the use and development of renewable energy. The U.S. Department of Agriculture sponsors a grant program to assist farmers, ranchers, and rural small businesses in developing renewable energy systems and energy efficiency improvements. The DOE has also announced grant awards to eight Native American tribes to develop renewable energy projects on their lands.

In addition, state and regional governments have been prime marketers of information about renewables. Many, if not all, U.S. state energy offices actively promote the use of renewable energy through Internet sites and other media. The events in California during the past few years have brought renewable energy to the forefront and have impacted public opinion about energy (at least in that area of the United States) in ways that might not otherwise have been possible.

5. TRENDS IN THE DEMAND FOR RENEWABLE ENERGY AND RENEWABLE ENERGY SYSTEMS

With all this activity and an increasingly pro-renewables consumer stance, one might expect actual use of renewable energy to be skyrocketing. In some ways, this is true, but in others it is not. The U.S. Energy Information Administration (EIA) reports in its "Renewable Energy Annual" that in 2001, total renewable energy consumption (solar energy, biomass, geothermal energy, conventional hydroelectric, and wind energy) declined by more than 12%, resulting in a decrease in the share of total U.S. energy consumption to 6%, the lowest level in 12 years (Fig. 1). These percentages are somewhat misleading because much of the change is due to a sharp reduction in the use of hydropower resulting from environmental and political pressures to eliminate dams in the Pacific Northwest. Shipments of solar thermal collectors, on the other hand, surged from 8.4 million square feet in 2000 to 11.2 million square feet in 2001 (Fig. 2), whereas total renewable electric generating capacity increased by a modest 1803 MW to 96,741 MW in 2001 (Fig. 3) due largely to increased wind power generation. The latest data available on total shipments of geothermal heat pumps indicate that they declined from 41,679 in 1999 to 35,581 in 2000 (Fig. 4). A new DOE initiative pertaining to geothermal heat pumps, planned to begin in 2004, may reverse this trend.

A major reason why renewable energy consumption has not exhibited larger increases is that utility

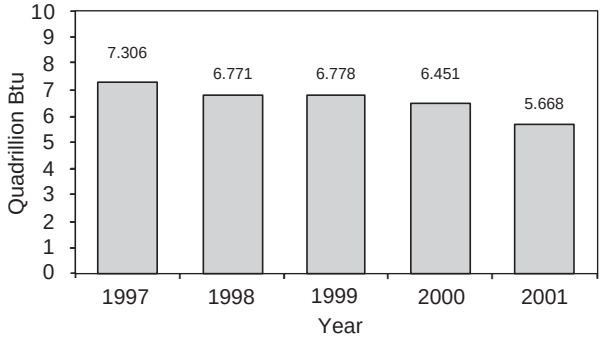


FIGURE 1 Total U.S. consumption of renewable energy, all sources, 1997–2001. Data for 2001 were preliminary at the time of reporting. Data from the U.S. Energy Information Administration (2002b).

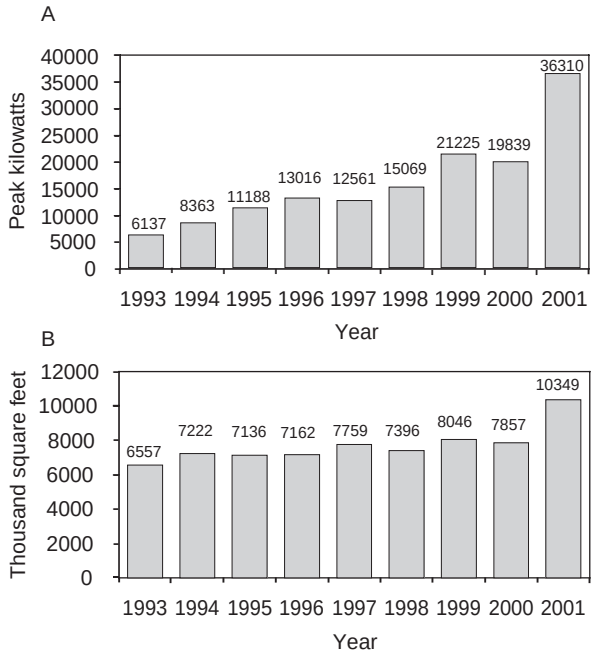


FIGURE 2 Annual total domestic shipments (U.S. states and territories, excluding shipments for space/satellite applications) of photovoltaic cells and modules (A) and solar thermal collectors (B), 1993–2001. Total shipments are net of exports. Data from the U.S. Energy Information Administration (2002b).

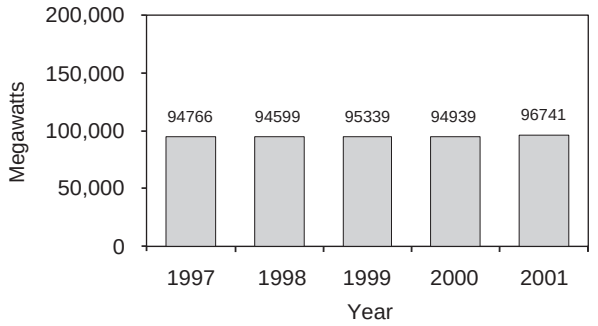


FIGURE 3 United States. electric net summer capacity attributable to renewables, all sources, 1997–2001. Data for 2001 were preliminary at the time of reporting. Data from the U.S. Energy Information Administration (2002b).

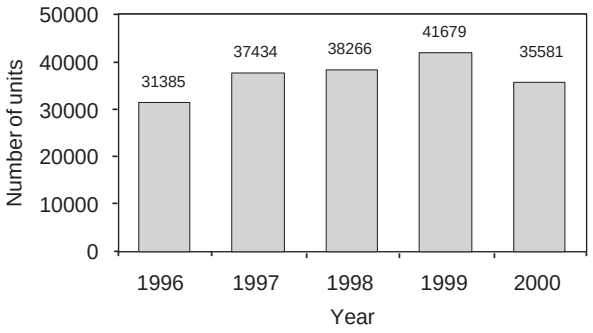


FIGURE 4 Annual total shipments of geothermal heat pumps, 1996–2000. Data from the U.S. Energy Information Administration (2002b).

companies have been slow to incorporate renewables into their resource base, citing economic factors associated with developing infrastructure and reliable resource supplies. Another reason is that utility regulators have been slow to adopt pro-renewable energy policies, and local government agencies have been inconsistent in their pursuit of economic incentives that promote market development. Although it is true that sufficient infrastructure may not yet exist to support continuous growth in the renewable energy market, this is really a proverbial "chicken-and-egg" problem. Markets cannot develop unless investments are made, and investments cannot be made unless the market is thought to be of sufficient size to justify them.

The utilities, however, being driven by their own consumer research, are finally getting the message. Utility customers are telling their electricity suppliers that they want renewables in the resource base. During the past few years, approximately 85 utilities in the United States have begun offering electricity generated from wind and solar power to their residential and commercial customers. The DOE provides a variety of information in both written and electronic form about successful ventures of local utilities into renewable energy. As a result, two phenomena are occurring: Energy supply is becoming more diverse and less centralized, and the demand for microturbines, rooftop-mounted solar panels for water heating and electricity generation, and other decentralized power sources is intensifying, in both the residential and the commercial markets.

6. CONSUMER ATTITUDES TOWARD RENEWABLE ENERGY AND RENEWABLE ENERGY SYSTEMS

Public response to renewable energy sources and systems is commonly measured through market and opinion surveys. Many such surveys pertaining to renewable energy have been conducted during the past 20 or more years by various organizations with different agendas. The quality of such surveys is highly variable due to inconsistency in survey format, question phrasing, and selection of respondents. Attitudinal responses are also affected by respondents' knowledge about specific energy sources since respondent knowledge is often uneven. In addition, almost all the results are descriptive in nature and provide little in the way of inferences with regard to the real populations of interest. Furthermore, consumer attitudes about renewable energy can span a wide variety of topics, so the research results are not always directly comparable. Table I provides comparative information about the operational and statistical aspects of a cross section of market and opinion research efforts at the local, regional, and national levels that pertain to renewable energy. Data obtained in objective, national surveys are thought to be the most reliable and unbiased.

Despite the acknowledged difficulties with some surveys, a trend has emerged with regard to consumer attitudes toward renewable energy. Some of the most telling results have been obtained from utility market surveys, many of which have been conducted by and for individual utility companies. Two 1999 reports summarize some of the findings. The first report, prepared at the National Renewable Energy Laboratory (NREL), contains an aggregate summary of results obtained in 14 different "best practices" quantitative surveys encompassing 12 utility service territories in five western/southwestern

TABLE I

Comparative Information on the Operational and Statistical Aspects of a Cross Section of Market and Opinion Research Efforts (Regional, Local, and National) Pertaining to Renewable Energy, 1995–2002

Year	Survey organization	Sponsor	Thematic content[a]	Population coverage	Subjects[b]	Survey mode	Margin of error (%)[c]
1995	Research/Strategy/ Management, Lanham, MD	Sustainable Energy Coalition, Takoma Park, MD	A	Nationwide	1000	Telephone	3.10
1997	Action Marketing Research, Minneapolis, MN	Environmental Law & Policy Center, Chicago, IL	A	Minnesota	900[d]	Telephone	Unknown
1997	Mellman Group, Washington, DC		A	Nationwide	800	Telephone	3.50
1998	The Eppstein Group, Ft. Worth, TX	Houston Lighting & Power, Houston, TX	A	Houston	>200	Deliberative poll	N/A
1998	Center for Deliberative Polling, University of Texas, Austin	TU Electric, Texas	A, B	TU Electric customers	Unknown	Deliberative poll	N/A
1998	Center for Deliberative Polling, University of Texas, Austin	Union Pacific Resources, Ft. Worth, TX	A	Texas	1207	Unknown	<3.0
1998	The Eppstein Group, Ft. Worth, TX	Southwest Public Service (SPS), Amarillo, TX	A, B	SPS customers	Unknown	Deliberative poll	N/A
1998	Research/Strategy/ Management, Rockville, MD	Sustainable Energy Coalition, Takoma Park, MD	B	Nationwide	1003	Unknown	3.10
1998	The Texas Poll, Scripps Howard and Office of Survey Research, University of Texas, Austin	Sierra Club, Lone Star Chapter, Austin, TX	B	Texas[e]	1004	Telephone	3.00
2000	International Consumer Research, Media, PA	Deloitte & Touch, New York, NY	A	Nationwide	600[f]	Unknown	Unknown
2001	Mellman Group, Washington, DC		B	Nationwide	1000	Telephone	3.10
2001	The Gallup Organization, Lincoln, NE		B	Nationwide	512[g]	Telephone	5.00
2001	Iowa State University, Ames, IA		B	Iowa farm producers	2060[b]	Unknown	Unknown
2002	Opinion Dynamics Corporation, Cambridge, MA	MA Technology Collaborative, Boston, MA	A, B	Massachusetts[i]	650	Telephone	3.84
2002	ORC/Macro Intl, Calverston, MD	NRG Systems, Hinesburg, VT	B	Vermont	402	Telephone	Unknown

[a] A, willingness to pay for green pricing; B, energy policy issues or emphasis of utilities on renewables.
[b] Unless specified otherwise, simple random sampling of registered voters is presumed.
[c] Usually stated as an overall rate, with subgroup rates being higher. Based on a statistical confidence level of 95%.
[d] Equal allocation among some counties and cities.
[e] Adults over the age of 18.
[f] Consumers (undefined).
[g] Adults.
[h] Based on multiplying reported response rate, 61%, times the original sample size (3376).
[i] Massachusetts residents.

states (results from an Electric Power Research Institute study are also included). Drawing on the approach and research used at NREL, the second report, prepared at the Bonneville Power Administration (BPA), contains a similarly constructed summary of 17 surveys or focus group studies conducted in the Pacific Northwest.

Included in the NREL report are the results for representative questions abstracted from the 14 surveys. One such question asked the customers to indicate the extent to which they favor various energy resources. The percentages of respondents indicating that they "somewhat favor" or "strongly favor" solar energy, wind, natural gas, geothermal energy, nuclear fission, and coal were 93, 91, 83, 71, 31, and 24%, respectively, suggesting a higher degree of favorability toward renewables than more traditional energy sources. Another question from this group of surveys asked customers to indicate which energy sources their utilities should use to generate power. Of the residential customers that responded, 88% said their utilities should use renewables (solar, wind, geothermal, or landfill gas) and 85% said their utilities should use hydropower, whereas only 55, 39, and 14%, respectively, said their utilities should use natural gas, nuclear fission, and coal. In a third question, customers were asked to indicate the source from which they preferred their electricity to be generated. Of those responding, 41% said they would choose wind and solar energy as their electricity source, 35% said they would choose natural gas, 9% said they would choose nuclear fission, and only 5% said they would choose coal.

Similar results are reported in the BPA document. In a question included in the 1997 Environmental Issues and Tracking Study commissioned by Portland General Electric and summarized by BPA, customers were asked to indicate the single additional energy source they would like to see their electric company develop during the next 5–10 years. Sixty-nine percent chose solar, wind, geothermal, or hydroelectric, 11% chose cogeneration using waste heat, 9% selected natural gas-powered turbines, 2% selected fuel cells, 2% chose coal, and 7% indicated some other source or did not know. The BPA report also summarized the results from a 1997 survey commissioned by Conservation and Renewable Energy Systems, an energy and conservation consortium of public utilities in Washington State. In that study, 88 and 89% of residential and commercial customers, respectively, indicated that utilities should include renewables as one of their energy sources. Even when residential customers were told that "renewables may cost more today but might make sense in the long run," there was no change in the percentages favoring them.

These comprehensive assessments reveal a repeated and consistent pattern that strong majorities of utility customers in various regions of the United States prefer renewables to other available energy sources and favor utility use of renewable resources to produce electricity. This pattern corroborates findings in polls conducted at the national level during a period of 23 years that Farhar summarized and described in a 1996 publication for the Renewable Energy Policy Project (REPP) in Washington, DC.

Some of the most recent evidence about consumer attitudes toward renewables was obtained in a 2002 survey conducted by the Wells Fargo Public Opinion Research Program (Wells Fargo PORP) at the University of Colorado, Denver. In this case, 95% of respondents disagreed with the statement, "It really makes no difference to me how my utility generates electricity." Of these same respondents, 82% indicated that utilities should focus on renewable resources as opposed to fossil resources to generate additional electric power; and when asked which source they would most like to see developed, 82% chose a renewable one (wind, solar, or hydropower). Furthermore, 84% indicated that the state's utilities should use "somewhat more" or "a lot more" renewable energy than they currently are using, 62% indicated that the state legislature should require utilities to use more renewable energy, and 78% said they would support a legislative mandate for utilities to produce 10% of their electricity from renewables within the next 10 years. Finally, strong majorities of respondents in this study rejected the notions that (i) it is too soon for utilities to invest in renewable energy, (ii) renewable energy is not worth developing because there is not enough of it, and (iii) wind farms are ugly.

The opinions and attitudes of European consumers toward renewable energy appear to mimic those of citizens in the United States. In a 2002 survey conducted by the European Commission, 69% of respondents said they supported more research for renewable energy, whereas the percentages of respondents who said they supported more research for natural gas, oil, and coal were only 13, 6, and 5%, respectively. On the other hand, only 40% stated they believe renewable energy would be the least expensive source in 2050. Like so many surveys of its type, this research does not directly address consumer attitudes toward actual use of renewables in power generation. It does, however,

serve to highlight the tendency of consumers to favor renewables when all other things are equal.

Information on consumer acceptance of specific renewable energy technologies and systems is more limited. Perhaps the most recent and extensive study about a particular technology, grid-tied photovoltaics (GPV), was conducted at NREL in 2000. In this survey of homeowners in Colorado, 68% of respondents favored GPV being made more widely available, but there was clearly greater favorability toward than familiarity with the technology. In the same study, participants were asked to identify which electricity source—coal, hydropower, natural gas, nuclear fission, oil, or solar energy and wind—they think of most when presented with each of several descriptive statements. Table II indicates the percentage of respondents who selected solar energy and wind with respect to each of the statements, providing information on a somewhat broader spectrum of attitudes toward renewable energy than presented in the other work.

Several surveys have been conducted in the United Kingdom concerning consumer favorability toward wind energy development. In one such survey, the University of Wales surveyed public opinions on behalf of the British Broadcasting Corporation at three different sites. At all three sites, majorities of greater than 60% of participants responded in favor of wind development despite aesthetic concerns associated with equipment installation. Environmental superiority to other fuels was most often selected as the main attribute of wind energy.

Collectively, this evidence suggests that a sizable percentage of consumers are favorably disposed toward renewable energy, particularly from the environmental standpoint; consumers and utility customers want to see more renewable energy use; and markets for renewable energy systems are growing.

TABLE II

Percentages of Colorado Respondents to a Survey about Grid-Tied Photovoltaics Who Associate Wind and Solar Energy with Several Economic and Environmental Statements

Statement	Wind and solar (%)
Best for the environment	79
Safest	70
Most abundant	43
Makes the United States most self-reliant	35
Least expensive	24
Will help the most in meeting the growing demands for energy	41
Stimulates economic development the most	24
Diversifies and extends our energy mix the most	52
Improves the quality of our air the most	80
Adds the most new high-technology jobs	40
Use of this energy source will increase the most in the 21st century	60
Best for the U.S. economy	32
Most positive overall reaction	51

Source: Farhar and Coburn (2000).

7. STATED WILLINGNESS TO PAY FOR ELECTRICITY DERIVED FROM RENEWABLE ENERGY

Although measuring consumer attitudes and perceptions about renewable energy is relatively straightforward, quantifying actual program participation or likely purchases of renewable energy and associated equipment is not. A consumer's proactive stance toward renewable energy does not always lead to direct or explicit actions that represent true consumption behavior. The reasons are numerous and complex, although some are more apparent than others. A full discussion is beyond the scope of this article.

One of the more obvious reasons is that the cost of producing electricity from renewable resources may be more than the cost of producing it from fossil fuels, and so the prospect of higher monthly utility bills sometimes dampens the enthusiasm of consumers despite their personal preferences. Consequently, a disparity exists between consumers' stated preferences for renewables or green pricing programs and their actual conduct/behavior when it comes time to participate or financially underwrite resource and infrastructure development.

Potential consumer actions are difficult to quantify directly. Consequently, stated "willingness to pay" is often used as a surrogate indicator of likely behavior in many avenues of consumer research. Although imperfect as a measure of actual behavior, willingness to pay is a notion that addresses the idea of potential consumer response to technological enhancements and it allows businesses, industries, and governing bodies to determine the probable extent to which consumers might embrace products and technologies in the marketplace. Willingness to pay, particularly in the context of green pricing, has been used extensively in energy market and opinion

surveys during the past decade as a proxy for consumers' actual purchase of electricity generated from renewable resources.

The 1999 NREL report described previously provides an authoritative summary of stated willingness to pay for electricity from renewable resources. As previously noted, this work encompasses the results of 14 different surveys and an Electric Power Research Institute study. In all the surveys reviewed, majorities of respondents indicate they are willing to pay at least a modest incremental monthly amount for electricity generated from renewable resources. The sizes of these majorities range from 52 to 95% without prior coaching or training regarding the economic or technological issues. The percentages are even higher when such information is made available.

Based on the similarity in the nature of the survey findings, an aggregate residential customer willingness-to-pay curve can be developed representing the incremental monthly amount that consumers would be willing to pay for electricity produced by their utility companies from renewable sources (Fig. 5). As expected, the percentage of consumers who are willing to pay more decreases sharply as the amount increases. Nonetheless, based on this analysis, approximately 70% of consumers could be expected to report a willingness to pay at least $5 more per month for electricity produced from renewable resources, and approximately 38% of consumers could be expected to report a willingness to pay at least $10 more per month. Although willingness to pay is not the same as actual purchase, these percentages are not trivial and they represent fairly solid market potential.

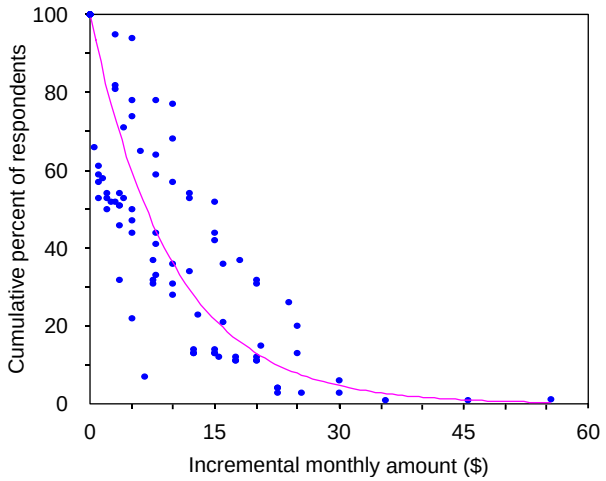


FIGURE 5 Aggregated willingness-to-pay curve for residential customers. The equation of the curve is $Y = 100e^{-104}M$, where Y is the cumulative percentage of respondents and M is dollars more per month ($R^2 = 0.76$). Source: Farhar (1999).

The results pertaining to stated willingness to pay from the utility-sponsored surveys summarized in the 1999 NREL report corroborate other survey results. Farhar's 1996 report on behalf of REPP indicates that 57–80% of the respondents in national polls state they are willing to pay more for "cleaner" electricity or electricity generated from "environmentally friendly" sources. The 2000 NREL study of Colorado homeowners pertaining to GPV indicates that 76% of respondents are willing to pay at least $1 more per month for electricity generated from renewable sources. Further corroborating evidence from polls of adults in four midwestern states is reported in a 1998 study conducted by Tarnai and Moore at Washington State University, along with information contained in the 1999 BPA report.

Some results that specifically pertain to wind power program participation are presented in a 1998 report by the Energy Center of Wisconsin. In contrast to the opinion and market research surveys described in the 1999 NREL and BPA studies, this one is more experimental in nature, the research is more controlled, and the analysis is more rigorous. Of the 761 respondents who were asked hypothetically if they would be willing to pay an additional amount on their monthly utility bill to obtain wind power, 43% said yes. Of these, 61% said they would be willing to pay at most an additional $2 per month or $24 per year. Fifty-eight percent said they would be willing to pay at most $4 more per month or $48 per year, 50% said they would be willing to pay at most $8 more per month or $96 per year, and 18% said they would be willing to pay at most $12 more per month or $144 per year. These results compare closely with those noted previously (Fig. 5).

Perhaps the most recent research in the United States on willingness to pay for renewables is found in the study conducted by Wells Fargo PORP. In this study, 55% of respondents indicate they are willing to pay an additional $10 or more on a typical monthly bill of $55 for electricity generated from renewables, whereas 18% are willing to pay nothing more. Two-thirds (67%) of the respondents indicate that any additional cost of using renewables to generate electricity should be spread across all consumers.

Current knowledge about U.S. consumers' stated willingness to pay for renewables closely parallels research findings in other industrialized countries, although more information seems to be available for

the United States. For example, in a small survey conducted on behalf of New Zealand's Energy Efficiency and Conservation Authority, more than half of the respondents indicated a willingness to pay 5% more for wind-generated electricity, and approximately 20% were willing to pay an additional 15%.

Irrespective of issues about comparability, the combined message from all these studies is that a high percentage of consumers nationwide, and perhaps even in other areas of the world, are indeed willing to pay more on a monthly basis for electricity produced from renewable sources. Of all possible indicators, this one signifies a fairly positive public response to renewable energy sources and systems in the general context of green pricing.

The finding suggests that markets for renewable energy and renewable energy systems do exist and that they can be expanded through strategic planning and implementation.

8. GAUGING ACTUAL PARTICIPATION OF CONSUMERS IN RENEWABLE ENERGY PROGRAMS

As suggested previously, engaging consumers in renewable energy participation has been the subject of much thought and investigation. However, because it is largely proprietary in nature, information about actual consumer participation in renewable energy programs and their purchases of renewable energy systems is difficult to obtain and assess. There are both operational and technical difficulties to overcome as well as definitional inconsistencies. For several years, the EIA has tracked shipments of solar PV cells/modules and solar thermal collection devices as a proxy for consumer demand (both private and commercial); as noted previously, these shipments were at an all-time high in 2001. Such equipment includes solar hot water devices. In fact, the EIA reports that in 2001, the residential sector replaced the industrial sector as the largest market for PV cells and modules. This evidence alone suggests, at least indirectly, that consumer demand for renewable energy systems is on the rise.

In addition, since April 2000, NREL staff have compiled and reported a "top ten" list of green pricing programs in the United States in each of the following categories: renewable energy sales, total number of customer participants, and customer participation rates. More than 300 utilities in 32 states offer green pricing programs. In December 2002, the number of patrons in the 10 largest programs with respect to customer participation ranged from 5644 to 72,732, for a total of 201,305 (approximately 42,000 of which are low-income customers in Los Angeles who do not pay a premium for green power). Participation rates in the programs with the 10 highest figures ranged from 3.0 to 5.8%.

Other entities throughout the world are also focusing more formally on the engagement of consumers and measurement of their participation in renewable energy programs. The National Institute for Research Advancement (NIRA) in Japan devoted an entire issue of its journal, *NIRA Policy Research*, to renewable energy development and citizen participation in the context of wind power.

Some of the most interesting and promising research concerning consumer attitudes and participation pertains to the new-home construction industry. Under the auspices of its Zero Energy Home (ZEH) program, DOE is promoting the construction of new single-family homes that combine available renewable energy technologies and energy-efficiency construction techniques to consume zero net energy from the utility grid. Both qualitative and quantitative research is currently being conducted at NREL on the consumer response to ZEH-like homes in San Diego, California. Although proprietary studies of homebuyers are relatively widespread in the building industry, this is some of the earliest work on customer response to ZEHs, particularly in a geographic locale in which volatile energy costs have been a political and economic concern.

The research involves direct interviews with purchasers of high-end homes (in excess of $500,000) in a large development in which more than 250 highly energy-efficient homes with GPV may eventually be sold. Almost all the homes will have solar water heating as well. Initial interviews with the home buyers have targeted the following issues: reasons for purchase, preferences for standard or optional systems, views on energy efficiency and solar energy features, satisfaction with utility bills, aesthetics, construction and technological problem areas, impacts on energy-related behavior, homeowner knowledge, and overall homebuyer satisfaction.

Early findings were described in a paper presented at the 2002 Summer Study conference of the American Council for an Energy Efficient Economy. Among homeowners with GPV, the availability of solar features was frequently cited as one of the top three reasons for purchase. Some of these buyers said they were looking for a hedge against rising energy

costs and some were motivated by their own environmental values. Although indicators of satisfaction can be misleading in the early stages of technology deployment, all the individuals interviewed indicated they were satisfied and that they would make the purchase again. Their feedback, although expected to be representative of highly positive and proactive early adopters, provided key insights into the motivation of consumers to become involved with renewable energy technology and the marketability of ZEHs. The realization that so many of these homes would actually be sold is further evidence of the existence of consumer demand for renewable energy and renewable energy systems.

Both GPV and solar water heating are included in the Million Solar Roofs initiative directed by DOE's Federal Energy Management Program. The objective of the Million Solar Roofs program is to encourage and support the installation of 1 million solar energy systems on residential, commercial, and public sector buildings by 2010. If successful, the program will have had a major impact on public perception and renewable energy markets in the United States.

Solar water heating is a more mature technology than GPV, having been available in several forms for more than 20 years. Studies on the market response to solar water heating have been conducted by several entities. The actual size of the installation base is difficult to reliably estimate, but anecdotal estimates for the United States range from 1 to 2 million systems. In the international arena, the International Energy Agency conducted a 5-year program begun in 1998 to enhance the market for solar domestic hot water systems in Canada, Denmark, Germany, The Netherlands, and Sweden. With the ultimate objective of 10,000 installations, this program is similar to the Million Solar Roofs initiative in the United States. Results of the project have not yet been made publicly available. Based on a March 2002 survey, the Australian Bureau of Statistics estimates that in Northern Territory, approximately 53% of households have solar hot water systems, as do approximately 16% of households in Western Australia (although this number is down from its peak of approximately 20% in 1999). Although these percentages seem robust, they represent only approximately 4% of all households in Australia.

Finally, the 1998 study by the Energy Center of Wisconsin concerning wind power program participation reports that 24% of 647 survey respondents who were asked to make an actual commitment to pay more to receive wind power agreed to do so (as previously noted, in another group of 761 individuals in the study that were asked, hypothetically, about their willingness to pay for green pricing, 43% responded affirmatively). When presented with a menu of monthly payment amounts, 47% committed to pay up to $2 more per month or $24 per year for wind power (compared to 61% of the other group who said, hypothetically, that they would be willing to pay up to $2 more per month). Thirty-five percent committed to pay up to $4 more per month (compared to 58% of the other group), 23% committed to pay up to $8 more per month (compared to 50% of the other group), and 8% committed to pay up to $24 per month (compared to 31% of the other group). A third group of 237 individuals were asked to pay for a fixed-size block of wind power of their choice every month, and surprisingly, more than one-third (37%) agreed to do so. The minimum block size was 50 kWh per month at an estimated incremental cost of $2, and the maximum block size was 600 kWh per month at an estimated incremental cost of $24. Other block sizes in multiples of 100 kWh were available at $4 per 100 kWh per month.

As expected, the percentages of respondents in the Wisconsin study who actually agreed to sign up were lower than the corresponding percentages of respondents who were asked about their hypothetical willingness to pay. This is the only publicly available research that involves a direct comparison of actual consumer behavior regarding renewables to presumed, hypothetical behavior, and it is the only publicly available work that attempts to measure the difference. Nonetheless, it seems clear that consumers throughout the world are utilizing or purchasing renewable energy systems for personal use and that the trend toward greater acceptance and deployment is a positive, increasing one.

9. PUBLIC RESPONSE TO RENEWABLES IN THE TRANSPORTATION ENERGY SECTOR

The foregoing discussion focuses on electricity and power generation, particularly as they relate to the residential and commercial energy sectors. However, attitudes toward renewable resources carry over to other energy sectors as well. With respect to transportation energy, a decade-long emphasis on the development and proliferation of alternative fuels and

alternative fuel vehicles has produced consumer responses not unlike those exhibited toward electricity and power generation. Again, only limited information is available, and it must be noted that not all alternative fuels are derived from renewable resources.

Although the rate of growth since 2000 has slowed, the total number of alternative fuel vehicles in use has increased steadily in the past decade in response to direct support, mandates, and research and development activities of government agencies throughout the world. Whereas more alternative fuel vehicles have certainly been placed in service during this time, record keeping has also improved, so the estimated annual rates of increase, particularly in the early 1990s, may not be completely accurate. The EIA estimates that in 2002 there were approximately 500,000 alternative fuel vehicles in use in the United States, of which approximately 60% were privately owned and operated. Approximately 82,000 of these were estimated to be ethanol-fueled vehicles (E85 and E95 combined; E85 is an 85/15% mixture of ethanol and gasoline). In comparison, the EIA reported only approximately 45,000 alternative fuel vehicles in 1993, of which only approximately 450 were ethanol fueled.

Of all the fuels classified by the DOE as alternative fuels, only ethanol can be specifically classified as a renewable transportation fuel. Ethanol is derived mostly from corn. Methanol might also be regarded as having been derived from a renewable resource, depending on its composition. The number of methanol-fueled vehicles in use in the United States has declined rapidly since 1997 to fewer than 6000 in 2002, whereas the number of ethanol-fueled vehicles has increased. In fact, no new methanol-fueled vehicles are being manufactured. Among all types of alternative fuel vehicles, the number of ethanol-fueled vehicles is growing at one of the fastest rates, particularly in the Midwest, where corn is plentiful (Fig. 6).

Not all alternative fuel vehicles operate exclusively on alternative fuel (alternative fuel vehicles that operate exclusively on their target fuel are referred to as dedicated vehicles; those that do not are said to be nondedicated). Ethanol vehicles, for example, can typically operate on either ethanol or gasoline (these are often referred to as flexible fuel vehicles). Information about the percentage of time that such vehicles actually operate on ethanol, especially privately operated vehicles, is sketchy.

Despite the growth in alternative fuel vehicles, the total number is still only a fraction of the entire in-use vehicle population. The EIA estimates that in the United States, alternative fuel vehicles represent only approximately 0.2% of total registered on-road vehicles. Furthermore, the amount of alternative fuel consumed annually in the transportation sector is less than 3% of total fuel consumption, and the percentage has declined since 1999. The EIA estimated U.S. alternative fuel consumption in 2001 to be 4.4 billion gasoline-equivalent gallons, of which ethanol accounted for 1.12 billion gasoline-equivalent gallons. A large percentage of ethanol is used as an additive in oxygenated gasoline formulations; thus, by default, many Americans—indeed many consumers throughout the world—are ethanol users.

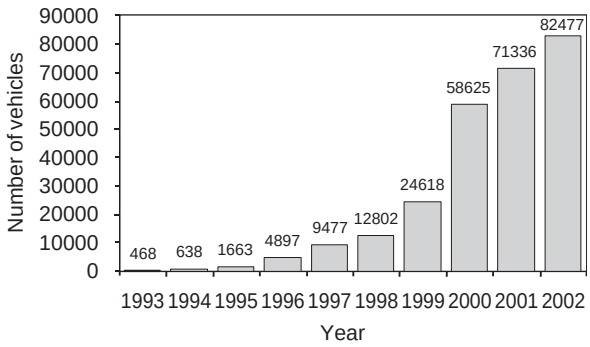


FIGURE 6 Annual numbers of ethanol-fueled vehicles (E85 and E95) combined. Totals for 2001 and 2002 are estimated. Data from the U.S. Energy Information Administration (2002b).

A number of studies have been conducted regarding market penetration of alternative fuel vehicles. For example, in 1996 the EIA reported on the markets for alternative fuel vehicles in general, whereas an earlier study conducted at the University of California at Davis focused specifically on alcohol-fueled vehicles.

Three NREL reports from the mid-1990s provided synopses about the preferences and satisfaction of drivers of alternative fuel vehicles. The first one described the results of a survey of U.S. federal fleet drivers in which respondents were asked to compare the performance of alcohol-fueled vehicles to that of other types. Approximately 90% of respondents assigned overall ratings of "very good" or "excellent" to alcohol-fueled vehicles, and similarly high percentages of drivers indicated they were leaning toward being "satisfied" or "very satisfied" with their vehicles. Approximately 87% of drivers of ethanol-fueled vehicles and approximately 67% of drivers of methanol-fueled vehicles indicated they would recommend their vehicles to someone else. A second report described the results from a corresponding survey of federal fleet managers. A high percentage of

these individuals (approximately 70%) indicated that drivers of the alcohol vehicles in their fleets were neutral toward them (a somewhat less enthusiastic response than from the drivers), but more than 80% of the fleet managers indicated they received approximately the same number of complaints from drivers of alcohol-fueled vehicles as they did from drivers of gasoline-fueled vehicles. The final report described the results of a survey of drivers of state and city government fleet vehicles. More than 80% of the drivers of alcohol-fueled vehicles assigned them overall ratings of excellent or very good, and almost all drivers said they were about the same as gasoline-fueled vehicles. Approximately 70% of the drivers of alcohol-fueled vehicles said they were very satisfied with their vehicles, and approximately the same percentage said they would recommend their vehicles to other drivers.

A number of other surveys of fleet managers have been conducted, both by commercial entities and by professional research organizations. However, many such studies have statistical and operational shortcomings, and they are not specifically concerned with alcohol-based (renewable) fuels.

Detailed information regarding the preferences and perceptions of owners/drivers of privately owned alcohol-fueled vehicles is not available, although anecdotal evidence suggests similarly high levels of satisfaction. After all, individuals are unlikely to purchase these kinds of vehicles for private or business use if they do not already have a positive attitude toward them. Likewise, no recent data concerning willingness to pay for alcohol-fueled vehicles are available.

Like the consumer information concerning heat and power generated from renewables, this collective body of evidence suggests the public response to automobiles capable of operating on renewable fuels is positive. More expansive growth seems to hinge primarily on vehicle availability in the marketplace, accessibility to adequate fuel (adequate refueling infrastructure, ease of refueling, etc.), and both vehicle and fuel price.

Fuel cell technology promises to revolutionize automotive power systems and substantially reduce the environmental impact of transportation emissions. Fuel cells also show promise as power plants for other appliances and devices such as computers. Simply stated, fuel cells are devices that produce electrical energy from a fuel using an electrochemical reaction, or the process of oxidation, rather than combustion. The function of a fuel cell is similar to that of a battery. Although various fuels can be used, hydrogen is the primary choice because the only by-products of the reaction are water, heat, and electricity. In a direct-hydrogen fuel cell vehicle, the controlled reaction between hydrogen and oxygen is believed to be more efficient than the process of internal combustion, in which the fuel (e.g., gasoline or diesel) is actually burned. Other fuels besides hydrogen can be used—particularly gasoline, diesel, ethanol, and methanol—but they all have to first be converted to hydrogen onboard in order for the fuel cell process to work. Safety concerns surrounding the transportation of hydrogen onboard private and commercial vehicles is the primary reason for considering other fuels, and there is considerable interest in renewable fuels such as ethanol and methanol because of their environmentally friendly characteristics. With support from various government agencies and other funding organizations, most of the automotive industry is engaged in fuel cell technology research. Honda and Toyota are the industry leaders; both companies have manufactured a limited number of fuel cell vehicles for sale (Toyota already has a sizable market, particularly in Japan, for its hybrid electric vehicles). Fuel cell systems for heavy vehicles and buses are also being researched by various organizations, such as Georgetown University and NREL. Although an important objective of many fuel cell vehicle research programs, there has been little investigation into public acceptance or consumer demand, and only limited information is available in the public arena. Consequently, only time will tell whether fuel cell technology can live up to its billing and supplant other power systems that are already deeply entrenched in the transportation sector.

10. FACTORS THAT MOTIVATE PUBLIC RESPONSE TO RENEWABLES

The question as to what motivates consumers to become personally involved with renewable energy is both interesting and multifaceted. The driving forces are myriad, but some of the more obvious ones are individualism and personal security, economics, environmental concern, the pursuit of a technocentric lifestyle, and even altruism. Although a direct linkage between such influences and actual purchase behavior is difficult to establish, survey data support the notion that individuals who are more favorable toward renewable energy also tend to be more sensitive to, or disposed toward, one or more of these factors.

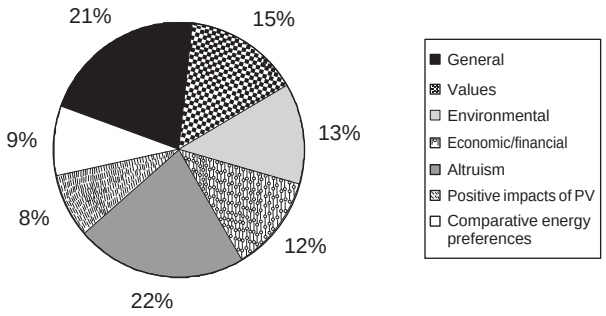


FIGURE 7 Percentages of references by participants in focused interviews to factors that motivate GPV adoption; $n = 3465$ total references by 120 participants in a purposive sample of residential energy customers in Colorado. Source: Farhar and Burhmann (1998).

A fairly extensive analysis of consumer motivation to adopt renewable energy technology is reported in the qualitative work by Farhar and Burhmann leading up to the 2000 NREL study of public response to GPV in Colorado. The percentages of all references in focused interviews to various motivating factors are summarized in Fig. 7.

The 2000 NREL study subsequently reports the percentages of respondents who indicated various perceived benefits of GPV to be important. These percentages are summarized in Table III. Taken together, the results of these two studies suggest that consumers' motivations to adopt renewable energy technology (e.g., GPV) are quite varied but perhaps more generally focused on economics and the environment.

The link between energy preferences and environmental concerns has been known for some time, but with increased emphasis on global warming and the greenhouse effect, the linkage is becoming more pronounced. National surveys conducted during the past 10 years indicate that fairly strong majorities of Americans believe there is a connection between energy usage and risk to the environment, and in the 2000 NREL study Colorado homeowners are shown to associate energy usage with environmental threat.

Although economic issues are always on the minds of consumers, they have perhaps been brought to the forefront by the recent energy shortages and price spikes in California. A 2001 brief prepared by Farhar on behalf of the Roper Center describes the results of several national polls conducted prior to the September 11, 2001, terrorist attack that indicate increasing concern among Americans about energy costs and the impact that worldwide shortages might have on them.

TABLE III

Percentages of Survey Respondents Who Indicated Several Perceived Benefits of Grid-Tied Photovoltaics (GPV) to Be Important

Perceived benefit of a GPV system	% of respondents indicating importance[a]
Result in long-term energy cost savings	86
Help reduce air pollution and acid rain in our area	84
Conserve natural resources	85
Benefit future generations	82
Reduce electricity bills right away	82
Help protect the environment	82
Increase resale value of home	79
Have electricity during a power outage	76
Help reduce global warming	75
Pay for itself over system lifetime	72
Sell excess electricity back to ultility company	71
Increase self-sufficiency	70
Increase diversity of energy sources	70
Help create jobs for Colorado	68
Good for Colorado's economy	68
Encourages others to replace gas and coal use	62
Feels good to do this	51
Increase awareness of household energy use	46
An opportunity to make a difference in community	44
Help create and expand PV market	38
A new technology to enjoy	33
Profitable for utility companies	20
First on block to have a PV system	10

Source. Farhar and Coburn (2000).

[a] On a scale of 1 to 10, with 10 being "very important," the percentage of all respondents who chose 7–10.

Research conducted at the University of Michigan prior to the catastrophic events of 2001 suggests, however, that economics may not be the sole motivating factor. In 1998, Clark et al. noted that the motivation for individuals to participate in initiatives such as green power programs is more complex, encompassing both intrinsic and extrinsic factors. Recent research concerning the motivations for corporations and other nonresidential customers to become involved in green power complements the University of Michigan findings. A 2001 report

prepared by Lawrence Berkeley National Laboratory (LBNL) in conjunction with Ed Holt & Associates indicates that altruism is the most significant factor, particularly among smaller companies and firms. Although conventional wisdom might suggest that economic gain is the principal motivating factor, this work implies otherwise.

Taken together, this body of evidence suggests that economics, environmental concerns, and national security and self-sufficiency will be the forces driving consumer demand for renewable energy and renewable energy systems in the foreseeable future. This conclusion is not surprising, particularly to those who have broad knowledge of the issues facing the energy establishment. What is surprising is the rapidity and certainty with which the interaction of these forces is changing the worldwide political, industrial, and economic landscape. The wildcard in the proliferation of renewables may be the speed with which science, engineering, business, and industry can bring products to the marketplace to capture the imagination of a public eager to try new inventions that have the potential to reduce personal uncertainty.

11. MODELING PUBLIC RESPONSE TO RENEWABLES

Surprisingly, little has been done in the way of formally modeling or predicting the public response to renewable energy and renewable energy systems. The 1998 report by the Energy Center of Wisconsin pertaining to wind power program participation purports to develop predictive models, but no such work beyond descriptive percentage distributions is actually presented. The 2000 NREL report presents the results of an extensive multivariate analysis of the survey data aimed at characterizing the Colorado market for GPV, which in context can be used to predict consumer participation. Beyond these two endeavors, little has been done to specifically predict public participation or response. On the other hand, there have been many efforts to develop economic models and market penetration models pertaining to renewable energy and renewable transportation fuels, although some of this knowledge is becoming dated. A collaborative partnership among NREL, LBNL, and Ed Holt & Associates resulted in forecasts of growth in green power markets in the United States.

The problem is that modeling and predicting public response to any technological intervention is difficult. Societal, cultural, economic, and other forces all interact to affect the responses of ordinary citizens. In addition, the identification and selection of appropriate measurements, and the collection of data in sufficient amounts and of sufficient quality, are substantial impediments to research. Finally, as previously suggested, behavior and attitudes are not necessarily directly related, and without a specific definition of what actually constitutes the public response, the ideas of modeling and prediction are necessarily reduced to purely theoretical exercises or to the reporting of mostly empirical, possibly indirect, observations.

12. CONCLUSION

Public attitudes toward renewable energy and renewable energy systems have shifted during the past 10–20 years, and they continue to evolve. The change is observable in ongoing opinion and market research conducted by government agencies, business entities, and respected research organizations throughout the world. Several factors are responsible. Most significant, perhaps, is a growing recognition that worldwide demand for energy—specifically, electricity—is surging and that traditional energy resources are limited. Recent economic and political instability, both globally and domestically, have created a milieu of uncertainty that accentuates public concern about the ability of government and industry to supply the energy necessary to meet societal needs. In addition, there is increasing evidence of global climate change, and consumers are becoming more convinced of its link to the consumption of traditional fossil fuels and more apprehensive about its impact on the environment and public health.

Nearly all the evidence of an attitude shift is empirical, and more needs to be done from an inferential and predictive standpoint. In addition, there are acknowledged issues of data quality and comparability. However, the amount of evidence has grown to the point that the positive trend toward renewables in public opinion and consumer attitudes cannot be denied. Operationalizing a favorable outlook into actual product purchases and program participation is a different matter, indeed; but here, too, there is substantial evidence that markets are developing and that consumers are increasingly acting on their convictions that renewable energy is a viable approach to sustainable economic and environmental security. As in the case of all technological ventures, price, availability of products,

accessibility to infrastructure, and reliability and performance of systems will be key to further consumer participation and expansion of markets.

SEE ALSO THE FOLLOWING ARTICLES

Public Reaction to Electricity Transmission Lines • Public Reaction to Energy, Overview • Public Reaction to Nuclear Power Siting and Disposal • Public Reaction to Offshore Oil • Renewable Energy and the City • Renewable Energy Policies and Barriers • Renewable Energy, Taxonomic Overview

Further Reading

Adamson, K. A. (2003). An examination of consumer demand in the secondary niche market for fuel cell vehicles in Europe. *Int. J. Hydrogen Energy* **28**(7), 771–780.

Appelbaum, L., and Cuciti, P. (2003). "Electricity Generation, Renewable Energy and Environmental Considerations: Results of a Colorado Survey." University of Colorado at Denver, the Wells Fargo Public Opinion Research Program, Denver, CO.

Australian Bureau of Statistics (2002). "Environmental Issues: People's Views and Practices," ABS Catalogue No. 4602.0. Australian Bureau of Statistics, Canberra, Australia.

Champ, P. A., and Bishop, R. C. (1998). "Wind-Power Program Participation: Developing Predictive Models," Report 179-1. Energy Center of Wisconsin, Madison, WI.

Farhar, B. C. (1993). "Trends in public perceptions and preferences on energy and environmental policy," NREL/TP-461-4857. National Renewable Energy Laboratory, Golden, CO.

Farhar, B. C. (1996). Energy and the environment: The public view, Issue Brief No. 3. University of Maryland, Renewable Energy Policy Project, College Park, MD. [Available: http://www.repp.org]

Farhar, B. C., (1999). "Willingness to pay for renewable electricity: A reveiw of utility market research," TP-550-26148. National Renewable Energy Laboratory, Golden, CO.

Farhar, B. C., and Burhmann, J. (1998). "Public response to residential grid-tied PV systems in Colorado: A qualitative market assessment," TP-550-24004. National Renewable Energy Laboratory, Golden, CO.

Farhar, B. C., and Coburn, T. C. (2000). "Colorado homeowner preferences on energy and environmental policy," NREL/TP-550-25285. National Renewable Energy Laboratory, Golden, CO.

Farhar, B. C., and Houston, A. H. (1996). "Willingness to pay for electricity from renewable energy," NREL/TP-260-21216. National Renewable Energy Laboratory, Golden, CO.

Farhar, B. C., *et al.* (2002). Market response to new zero energy homes in San Diego, California. In "Proceedings of the 2002 ACEEE Summer Study." American Council for an Energy Efficient Economy, Asilomar, CA.

Ferguson, E. G. (1999). "Renewable Resources and Conservation: what consumers want." Bonneville Power Administration, Portland, OR.

Gielecki, M., *et al.* (2001). Incentives, mandates, and government programs for promoting renewable energy. In "Renewable Energy 2000: Issues and Trends," DOE/EIA-0628. U.S. Energy Information Administration, Office of Coal, Nuclear, Electric, and Alternative Fuels, Washington, DC. [Available: http://www.eia.doe.gov.cneaf/solar.renewables.rea_issues/reaissues_sum.html]

Rauluk, V. (2001). "Development of markets for solar water heating systems," Venture Catalyst Inc. final report 1998–2001 prepared for the U.S. Department of Energy, DOE/EE10357-F. U.S. Department of Energy, Washington, DC.

U.S. Energy Information Administration (2002a). "Alternatives to Traditional Transportation Fuels 2000." U.S. Energy Information Administration, Washington, DC. [Available:http://www.eia.doe.gov/cnea/alternate/page/datatables/atfl13_00.html]

U.S. Energy Information Administration (2002b). Renewable energy annual 2001, with preliminary data for 2001, DOE/EIA-0603. U.S. Energy Information Administration, Office of Coal, Nuclear, Electric, and Alternative Fuels, Washington, DC. [Available:http://www.eia.doe.gov/cneaf/solar.renewables/page/rea_data/rea_sum.html]

Whalen, M., and Coburn, T. (1997). "Perspectives on AFVs: 1996 federal fleet driver survey," NREL/BK-540-22721. National Renewable Energy Laboratory, Golden, CO.

Wiser, R. H., *et al.* (2001a). "Public goods and private interests: Understanding non-residential demand for green power," LBNL-47300. Lawrence Berkeley National Laboratory, Berkeley, CA.

Wiser, R. H., *et al.* (2001b). "Forecasting the growth of green power markets in the United States," NREL/TP-620-29408. National Renewable Energy Laboratory, Golden, CO.

Radiation, Risks and Health Impacts of

RICHARD WILSON
Harvard University
Cambridge, Massachusetts, United States

1. **Introduction**
2. **Effects at High Doses (> 3 Sv)**
3. **Effects at Moderate to High Doses (0.3–3 Sv)**
4. **The Effects of Dose Rate**
5. **Effects at Low and Very Low Doses**
6. **Weighting Factors (W_r), or Relative Biological Effectiveness**
7. **Policy Implications**
8. **Public Perception**

Glossary

absolute risk model A model in which the increase in the rate of cancer, or other disease, is independent of the natural incidence at that age and place.

absorbed dose The amount of energy absorbed in a material. The Système International unit for absorbed dose is named after the medical physicist Professor L. H. Gray. The gray (Gy) represents an energy absorption of 1 joule per kilogram of absorbing material, i.e., 1 Gy = 1J/kg. However, an older unit, the rad, is still often used, particularly in the United States (100 rad = 1 Gy).

ALARA An acronym for "as low as reasonably achievable," a criterion for judging radiation protection; alternatively, ALAP (as low as practicable) or ALARP (as low as reasonably practicable).

assigned share See **probability of causation**.

equivalent dose (radiation weighting factor) An expression introduced to take into account the dependence of harmful biological radiation effects on the type of radiation being absorbed. The same absorbed dose delivered by different types of radiation may result in different degrees of biological damage to body tissues. The total energy deposited is not the only factor that determines the extent of the damage. The equivalent dose is therefore a measure of the risk associated with an exposure to ionizing radiation. Risks due to exposures to different radiation types can be directly compared in terms of equivalent dose. The unit of equivalent dose is the sievert (Sv), defined for a given type of radiation by the relationship $H = DW_r$, where H is the equivalent dose in sieverts, D is the absorbed dose in grays, and W_r is a numerical weighting factor. In older centimeter-gram-second (CGS) units, the term "roentgen equivalent man" (rem) is still often used (1 Sv = 100 rem).

excess relative risk (ERR) The excess (additional) risk caused by radiation, divided by natural risk (incidence).

ionizing radiation The radiation with individual quanta of high enough energy to ionize an atom.

LD_{50} The "least dose" at which 50% or more of people are affected (usually producing, but not always, death).

probability of causation The estimated probability that an event (such as cancer) was caused by a postulated action (usually radiation exposure).

radiogenic cancers Occurrence of cancers at a body site where it is known that radiation can cause an increase; most leukemias are radiogenic cancers, whereas chronic lymphocytic leukemia is not.

relative biological effectiveness (RBE) An alternative description for the weighting factor W_r.

relative risk model Extrapolation of the ratio of a risk to the risk from natural causes at a certain age and place; this model is in contrast to the absolute risk model.

Each and every one of us is exposed to natural ionizing radiation, all the time, from the sky, from the air, from the ground, and even from our own bodies. Mankind can also produce ionizing radiation. This was realized when X rays were discovered by Professor Roentgen in 1895; radioactivity was discovered by Becquerel soon thereafter, and scientists began to notice that exposure to radiation had harmful effects—radiation was noted to burn, damage, and even destroy tissue. These effects were used very soon (about 1905) for beneficial purposes, to destroy an unwanted skin tumor. Almost from the

 © 2004 Elsevier Inc. All rights reserved.

beginning arose the dilemma: excessive radiation can be bad for you, but it can also be good for you. The focus here is on assessment of the risks and the benefits of radiation.

1. INTRODUCTION

The most important medical application of radiation is the use of X rays for diagnostic purposes. There is an enormous advantage to a physician, and hence to the patient, of knowing exactly where a bone or blood vessel lies in the body. Early on, this realization seemed to outweigh any small risk that continued use of X rays might entail. This crude risk–benefit comparison probably contributed to the fact that it was many years before people acted to reduce radiation risk. The benefit seemed so great that it was forgotten that risk could be reduced without changing the benefit. In the 1920s and 1930s, a number of pioneer radiologists suffered severe injuries, and some even died, as a result of prolonged exposure to dangerously high intensities of X rays. These early workers in the field had no means of accurately measuring the harm caused by radiation and they often used inferior measurement equipment even though better equipment was available; to establish radiation exposure, they depended on unreliable effects, such as the degree of skin reddening caused by the exposure, or on timing the exposure from a certain type of X ray machine.

Other adverse effects of radiation were noted in workers involved in several industrial or clinical uses of radiation in the first half of the 20th century. In the 1920s, watch dials were painted with a luminous paint containing an alpha particle emitter, usually from radium. Most watch factory painters sucked the paintbrush to obtain a fine point, thereby absorbing radium in their mouth, ultimately in their bones. In clinics and hospitals, physicians suffered from excess radiation when they administered large doses of radiation to cure painful diseases such as ankylosing spondylitis. X-Ray operators failed to shield themselves and the parts of the patient's body that were not under study. Radiologists would sometimes look straight at a fluorescent screen and therefore straight into the X-ray tube.

The radiation dose for a typical chest X ray was about 1 rem until about 1960; with good shielding and use of image intensifiers it can now be (and could have been 50 years before) 7 mrem. The big impetus toward understanding the health effects of radiation was the massive expansion of nuclear technology following the development of nuclear reactors, and the explosion of the atomic bombs in 1945. Since then, radiation has become inextricably linked with the atomic bomb and its adverse effects, much to the detriment of benign uses. The fact that radiation is associated with bombs leads many people to believe that radiation is uniquely dangerous. That is far from the case. For example, the largest and most important study of radiation-induced cancer comes from the study of Hiroshima and Nagasaki victims, among whom 500 or so excess cancers (more than expected from the nonexposed population) were observed. Yet that many excess lung cancers were found in a separate study unrelated to exposure from the bombs. One crucial feature of such data is that radiation does not seem to cause any medical or biological effect that cannot also occur naturally. It merely increases the probability of the effect. This fact is very important, both for understanding possible dose–response curves and for deciding what, if anything, to do about radiation at low doses. It also leads us to ask a general question: Does radiation add an additional probability of developing cancer (an absolute risk model) or does it multiply the probability that is there from natural or other causes (a relative risk model)? Although both models are discussed in most summary reports, it is noteworthy that the emphasis has changed from an absolute risk model (in 1970) to a relative risk model with an increase in calculated risk—particularly in aging population groups.

2. EFFECTS AT HIGH DOSES (>3 Sv)

When radiation is given over a short period (less than a day) at doses greater than 200 rem, the primary effect is serious destruction of blood and tissue. Two situations can be distinguished: radiation exposure of a small area of the body and exposure of a large percentage of the body surface. A recent example of the first situation arose in the republic of Georgia, of the former Soviet Union, when a soldier picked up a large strontium 90 source that had been used for a thermoelectric generator and left behind by the Soviet army. The soldier's hand was treated for a year in the hospital of the Curie Institute in Paris before the soldier fully recovered.

In the past, misuse of X rays led to many situations in which a hand or other extremity was

heavily exposed (up to 200 Sv = 20,000 rem). At this level, tissue is destroyed but can be surgically replaced. High exposures of limited quantities of tissue occur in radiotherapy. The aim in radiotherapy is to provide a large enough dose to destroy a tumor, without destroying neighboring tissue. Typical doses to the tumor are 60 Sv (= 6000 rem). If the dose is given in several fractions, the neighboring healthy tissue can recover. In contrast, a much smaller dose given to the whole body can destroy the whole organism. Ordinarily, the dose from which half of the recipients will die, the LD_{50}, is about 2.5 Sv (= 250 rem), but the LD_{50} can be extended up to about 5 Sv in aggressive treatment. Following high-dose radiation, blood transfusions to bring in healthy blood have been successful, and bone marrow transplants to replace destroyed bone marrow were tried in a few cases of heavily exposed workers after the 1986 Chernobyl nuclear accident. A couple of the Chernobyl patients improved, but an equal number were probably killed by the treatment. At 5 Sv (= 500 rem), the gut and many other body organs begin to fail. Death is not immediate but takes up to a week. One famous victim, the Canadian Dr. Louis Slotin, a scientist at the Omega Site Laboratory in Los Alamos, New Mexico, lived for a month, and actually wrote a description of the incident while he was dying, after an accident with supercritical plutonium in 1946 exposed him to ~10 Gy of gamma and neutron radiation.

3. EFFECTS AT MODERATE TO HIGH DOSES (0.3–3 Sv)

3.1 Chronic vs. Acute Effects

Daily doses lower than those required to destroy tissue, can, if repeated over a long period, cause health effects. In contrast to the acute effects from exposure over a short exposure, these chronic effects often involve whole body doses accumulated over years, of 10 Sv and more. From general biological principles, the following effects of continued exposure to even moderate doses of radiation might be expected: induction or aggravation of cancer, induction or aggravation of heart disease, induction or aggravation of other diseases, production of genetic anomalies that can be passed on to another generation, and induction of birth defects by genetic mutation. Most of the studies address only cancer induction, and the data do indeed suggest that cancer induction is the dominant effect. The data are better for cancer than for the other outcomes largely because the observed effects are greater. Several groups of radiation-induced cancer (e.g., leukemia, solid cancers, lung cancer) can be distinguished based on their different characteristics.

3.2 Chronic Radiation Sickness

Normally, any worker who receives a high radiation dose on the job is prohibited from participating further in any work that involves radiation exposure. A case history in which this practice was ignored occurred in the late 1950s, in the former Soviet Union, where there was a great push to build an atomic bomb to "balance power" with the United States. For many years, it had been rumored that workers at the Mayak nuclear reprocessing plant in the southern Urals received large radiation doses. Attempts from the western countries to discover the facts about what happened at Mayak were fruitless until 1991, when a description was published in Russian in the journal *Priroda*. The workers at Mayak had continued working even after being highly exposed, and over a 2-year period received doses up to 10 Sv (= 1000 rem). The first diagnosed effect was a drop in blood cell count and general sickness. This was called chronic radiation sickness, in contrast to acute radiation sickness, which is caused by large singular doses. Radionuclides from the Mayak chemical processing facility were dumped into neighboring Lake Karachay. In 1955 and 1956, water overtopped the reservoir and spilled into the Techa River, which drains into the Arctic Ocean. Villagers drank the water and ingested many radionuclides. For 40 years, the health of 30,000 villagers around the Techa River has been studied and chronic radiation sickness has been diagnosed. Around the Techa River, the exposures were mostly internal; the internal organ dose is used in describing leukemia incidence. In contrast, "solid" cancer diagnoses depend on external doses, which are far less easily determined. The internal Techa River exposure determination was based on detection of the "bone seeker" strontium-90; doses were fairly accurately determined by examination of radioactivity of teeth and other bones.

3.3 Leukemia

Although there have been studies of the effects of radiation on people for 100 years, the most important studies are the studies of the consequences of the Hiroshima and Nagasaki atomic bombs. The

research on the survivors has involved many good scientists, considerable effort, and considerable expense, mostly at the Radiation Effects Research Foundation in Hiroshima. The exposures occurred many decades ago, but an increase in cancers over that expected in the general population is still occurring. Therefore, the most recent research is of greatest importance. In addition, the radiation dose to which the population was exposed is uncertain. It was derived from measurements at other (similar) nuclear explosions, and based on the neutron dose by measuring long-lived radioactivity in the region.

Leukemia was the first cancer to be noticed in the atomic bomb survivors. Leukemias began to appear 5 years after the exposure and continued to appear for another 20 years, after which the incidence of leukemias (almost) ceased. Radiation induces leukemias more readily than other cancers. Therefore, leukemias are often considered to be a "marker" for radiation effects. But the variation with age is clearly in great contrast to that of the "solid" cancers, and in the elderly even an "absolute risk" model would overpredict the number.

3.4 Solid Cancers (Excluding Lung Cancer)

Solid cancers occur in tissues of the body and are not connected with blood components. They do not appear until 20 years after radiation exposure and have been noted to continue to appear for 50 years. The increase of cancer with radiation dose seems to follow a "relative risk" pattern whereby the risk, after the latent period of 20 years, is increased by a constant fraction relative to the background at that age.

3.5 Lung Cancers

Lung cancers are of particular interest because the lung is the point of entry for a major radiation exposure. In order to understand the effect on humans of alpha particle irradiation, there have been several animal studies that show an increased incidence of cancer related to radiation dose. The animal studies showed a quadratic relationship. Therefore researchers initially looked for a quadratic relationship, and not a linear relationship, in human lung cancers. Several cohorts have provided useful information.

One of the surprises of the atomic age (i.e., since 1945) was that an excess of lung cancers appeared among uranium miners. In some of the early studies, lung cancers seemed to appear only among Caucasian miners and not among Native American Indians. There was some speculation of a racial difference in sensitivity. This speculation was proved incorrect both by follow up and by observations that there were fewer tobacco smokers among the Indian miners, compared to the Caucasian miners. Most analysts have assumed that lung cancers are due to exposure to radon gas in the mines rather than to any other pollution in the mines (dust, etc.), although there is still room to question this. The main concern is that inhaled particles or radionuclides might cause lung cancer. The generally accepted effect of radon gas is that radon alpha particles and "daughter" products irradiate the lung and induce cancer. This suggests that exposures to other alpha particle emitters also should produce lung cancers.

There have been several studies of the health of persons who worked with, and possibly inhaled, plutonium. Lung cancer is the principal effect that may be caused by plutonium inhalation. Although there have been many persons working with plutonium, in Los Alamos (New Mexico), Rocky Flats (Colorado), and Hanford (Washington), safety precautions have kept exposures small except in one cohort, the workers at Mayak (the former Soviet Union) in the 1950s. Because plutonium deposits in the bone, it was possible in most cases at Mayak to make accurate estimates of exposure by looking at the deposited radioactivity in the bone, either in *vivo*, or after autopsy. The highly exposed group of Mayak workers developed lung cancer at twice the rate of the less exposed workers, whose lung cancers are assumed to be caused by cigarette smoking. The first studies found agreement with the expected quadratic variation of incidence with dose, but later studies showed that a linear relationship is more probable.

There have also been "ecological" studies of the relationship between the average lung cancer incidence in a population and the average concentration of radon gas. In the most complete of these studies, radon gas was measured in half a million U.S. households and data were compared to the average lung cancer incidence in the region. To the surprise of the author of the studies, the lung cancer rate was found to be lower in high-radon areas compared to low-radon areas. The effect is highly significant. There is argument among experts as to whether these data indeed show that inhaled radon gas is beneficial. On one side, modern epidemiologists argue that such "ecological" studies show little by themselves. All they show is that the average lung cancer incidence rate $\langle i \rangle = -k * \langle c \rangle \times \langle p \rangle$, where $\langle c \rangle$ and $\langle p \rangle$

are the average concentration and the average population density. They do not show that $c = k * a * b$. Moreover, there is, in their view, too much room for bias. In particular, the study includes both smokers and nonsmokers intermingled, and the largest cause of lung cancer is known to be cigarette smoking. On the other side, scientists note that data are data; the data do not fit the regulatory model proposed by the U.S. Environmental Agency (EPA), and that model must be incorrect or at least inadequate. The general consensus is that, at the concentration levels of concern, radon gas cannot be the major contributor to lung cancer, either for smokers or nonsmokers.

3.6 Childhood Cancers

In the period from 1940 to 1970, it was common to give an X ray to pregnant women to discover any problems with the infant fetus. This dosed the fetus with radiation approaching 0.01 Sv ($= 1$ rem). In a classic study (the Oxford study), it was found that the probability of developing childhood leukemia in the age group 0–8 years increased with the number of X rays. This effect was not observed in the Hiroshima–Nagasaki studies. It is usually accepted that these leukemias are caused by something that happened *in utero*. But there is a possibility that causality proceeds in the opposite direction. A physician noticing an abnormality leading to cancer might call for an X ray. The correlation might then merely be between the abnormality and a subsequent cancer. Childhood cancers occur up to 7 years from birth, so cancers from the Chernobyl accident should have been seen. A small effect is possible; it would not be larger than, and could be smaller than, the effect suggested by the Oxford study. It would be especially interesting to examine this more carefully by looking at all children of the women who were pregnant and who were exposed at Chernobyl or any other radiation accident.

3.7 Thyroid Cancers

In any nuclear reactor accident, one of the first determinations must be whether any radioactive iodine was released. The short-lived iodine isotope can deposit in pastures; if cows ingest contaminated pasture grass, the iodine will concentrate in their milk. If children drink the milk, the iodine will concentrate in the thyroid, resulting in a large radiation dose, up to 100,000 times normal. This happened at an accident in Windscale, U.K. in 1956, and at Chernobyl in 1986. The first preventive measure must be to break the route of contamination by stopping all consumption of milk from the affected area. This was done at Windscale, but not at Chernobyl because of a desire to maintain secrecy. Many western scientists brought this to the attention of the scientists and politicians in the former Soviet Union, but when thyroid cancers were diagnosed among children starting in 1979, it was a great surprise.

Thyroid cancers, particularly among children, are usually not fatal. They are closely linked with benign thyroid nodules, so it is hard to get a diagnosis that is consistent across many groups. Initially (in 1989 and 1991), therefore, an increase in children's thyroid cancers in the former Soviet Union was attributed to a sudden increase in attention paid to the problem by physicians. However, this was soon shown to be incorrect. By 2002, over 1500 childhood thyroid cancers were seen in Belarus, Ukraine, and Russia, with excess relative risk of over 50. Fortunately, there has been considerable attention paid to these by western physicians, and only 11 fatalities have occurred so far. Increases in childhood thyroid cancers have not been seen elsewhere. Now that many years have passed since the Chernobyl accident, an increase in adult childhood cancers might be expected. Indeed, an increase has been seen, but such increases are hard to interpret, because of the twofold effect of increased attention to the problem by physicians and the fact that it is hard to develop consistent diagnostic criteria. One Russian study claims to account for the difficulties by searching for a dose-related effect, but none was found. Nonetheless, whether adult thyroid cancers have truly increased since 1979 (when an apparent increase was observed) remains a matter of controversy.

3.8 Heart Disease and Other Chronic Outcomes

A careful examination of the Hiroshima–Nagasaki data shows an increase in several other medical outcomes. In particular, heart disease appears with a frequency of about one-third the frequency of cancer.

3.9 Genetic Effects

There is a general belief, based both on general biological principles and on animal experiments, that radiation can cause genetic defects, but no well-authenticated examples attributable to radiation have been observed in humans, even in the Japanese

atomic bomb survivors. Again, the numerical value of the dose response is controversial. However, for both cancer and genetic defects, the disagreement is over very small numbers, i.e., whether the effects around 1 mSv (= 100 mrem) are very small or very, very small—much smaller per capita compared to the cancer effects.

3.10 Growth and Development of Children

A statistically significant adverse effect of radiation has been found in the Radiation Effects Research Foundation (RERF) data on children whose parents were irradiated while they were *in utero*. Those whose parents were exposed within 1500 m of the hypocenter of the Hiroshima bomb were on average 2.25 cm shorter, 3 kg lighter, and 1.1 cm smaller in head circumference compared to those exposed further away. Although (just) statistically significant, this effect needs confirmation in other cohorts before it can be generally accepted. This has not been confirmed in any other cohort. However, although other exposed groups exist, such as those exposed near Chernobyl or at the Techa River, there has been no systematic study that either confirms or denies the relationship. Another effect that is perhaps associated with (or a consequence of) the reduction in head size was seen in the RERF data. There is a statistically significant reduction in Intelligence Quotient (IQ) among children exposed *in utero*. In other cases, there is severe mental retardation. Although these effects on children are generally believed to have a threshold, the data on reduction in IQ are also consistent with a linear relationship between reduction and dose.

4. THE EFFECTS OF DOSE RATE

The previous discussion of the cancers caused by radiation exposure implies that it is the accumulated dose that is important. This seems to be a good first approximation. But from general principles one might guess that a high radiation dose given at a low rate over a period of years might have a different (probably smaller) effect than the same dose given in a short time, although if the time period is long, on the order of a lifetime, the difference is likely to be small. Data from animal exposures to radiation show that there is a reduction in cancers (for the same total dose) at low dose rates. A dose rate reduction factor (DRRF) is usually introduced to describe this. From animal data, a DRRF of between 2 and 10 is found and can be used for radiation protection purposes. Situations of accumulated radiation doses of 0.5 Sv (= 50 rem) exist, including the Mayak and Techa River cohorts noted earlier and the population exposed to Cs-137 subsequent to the Chrenobyl accident. From the Techa River and the Mayak data it seems that the DRRF for leukemia is about 3, but with a large error band (perhaps from 2 to 6). The DRRF for solid cancers is about 1, with a much larger error band ranging up to a factor of 10.

The importance of the data from these accident situations in which large cumulative doses, but not large acute doses, were received can not be overestimated, as they are useful for predicting the effect of nuclear accidents, as distinct from nuclear explosions. However, the data are not accurate enough to be used directly, and so at the present time, the world must depend primarily on the Hiroshima-Nagasaki data with an associated DRRF.

5. EFFECTS AT LOW AND VERY LOW DOSES

5.1 Early Belief in Beneficial Effects

Soon after natural radioactivity was discovered in the 1890s, there arose a general belief that exposure to moderate levels of radiation was therapeutically beneficial. This belief was not unique to radiation exposure, and many known poisons were given for therapeutic purposes. The small amounts of arsenic in Fowler's solution cured many disorders, and arsenic was also used in larger doses to cure syphilis before other, safer antibiotics were available. Communities with mineral hot springs added "radium" to their names to attract customers. The name Radium Hot Springs appeared in many places. As late as 1980, the radioactivity level of some European mineral waters was listed in advertising as a positive feature. However, the data supporting the suggested beneficial effect are scanty. Although there were several reports of cures, it is unclear whether these cures were due to radium, to other minerals in the water, or merely to the beneficial effects of a long soak in hot water.

Some old experiments on animals suggested that moderate radiation doses led to a reduction in disease and to a prolongation of life. But these reports have not been replicated with recent protocols, nor have there been human studies with modern

epidemiological methods. It is completely consistent that radiation would cause a beneficial effect for one medical outcome and an adverse effect for another one. It seems very plausible that radiation could reduce infectious disease by destroying the germs. But in the last 100 years, infectious diseases, with the exception of autoimmune disease syndrome (AIDS), have been controlled by other methods, so that even if it is true that radiation reduces the effect of infectious disease, it may now be irrelevant. It may be relevant for AIDS, but no study has been made. Alcohol in high doses is well known to cause cancer and has a narcotic effect at still higher doses. But at low doses—up to one drink a day—alcohol reduces the risk of stroke.

Since 1945, there has been a search for some chemical that can reduce the effects of radiation. A few substances can cause excretion of undesired radionuclides. It has been suggested that the cellular molecule MDC1 (mediator of DNA damage checkpoint protein 1) acts like an emergency service in cells to detect and repair damage to DNA caused by radiation. It is too early to say whether the effect is significant.

Although it is generally assumed that radiation effects at any site or organ will follow similar dose–response relationships, this is not necessarily true. The National Toxicology Program database lists a couple of dozen chemicals that decrease the likelihood of cancer in rats or mice when applied at one site, but increase it when applied at another.

5.2 The Definition of Low Dose

The concept of what constitutes a low dose has been modified considerably over the past 50 years. In the first half of the 20th century, physicians believed that 0.01 Sv (= 1 rem) was a low dose. In 1945, a typical chest X ray administered a dose of 0.01 Sv and at least one jurisdiction (United Kingdom) went as far as to propose mandating such a X ray every year to help fight tuberculosis. Most experts now believe that this would have caused an excess number of cancers. The proposed U.K. bill died in the House of Lords because of the objections of the physicist Lord Cherwell. A mandate for a yearly chest X ray would now be more reasonable. Chest X rays are now routinely performed with a dose 150 times smaller. Also, a skin test for tuberculosis antibodies now makes a chest X ray unnecessary in most cases.

The idea that 0.01 Sv per year is a low dose has changed. In 1987, the U.S. Nuclear Regulatory Commission proposed to label "Below Regulatory Concern" a radiation dose exposure of no more than 1 mrem (0.00001 Sv) to any person. The proposal was withdrawn after some vocal public opposition. This dose is 1000 times smaller than what was medically accepted before 1945! Moreover, natural background exposures are a few hundred millirems, or 100 times this amount, and vary considerably from place to place and person to person. For scientific and political reasons, "low dose" now means doses as low as, and usually well below, background doses. But scientifically there is still no generally accepted evidence for an effect, beneficial or adverse, of a radiation dose this small for any medical outcome

5.3 Dose–Response Curves of Acute and Chronic Radiation Exposure

Acute effects seem to occur only with radiation doses above a definite threshold dose. They are sometimes called deterministic effects. When a population is exposed and the dose is below that threshold, no one is hurt; everyone is hurt when exposure is above that threshold. This threshold dose seems to be independent of gender, genetic predisposition, and of racial characteristics. Even skin effects, which might be expected to be affected by pigmentation, seem to be independent of race. There are small variations and a dose–response curve is usually drawn as in Fig. 1. The dose at which 50% of persons in a population are affected is called the LD_{50}. For whole body exposure, the LD_{50} is about 2.5 Sv (= 250 rem) and persons exposed at half this dose survive. At a dose that is one-fourth of the LD_{50}, it is generally accepted that acute effects will not occur.

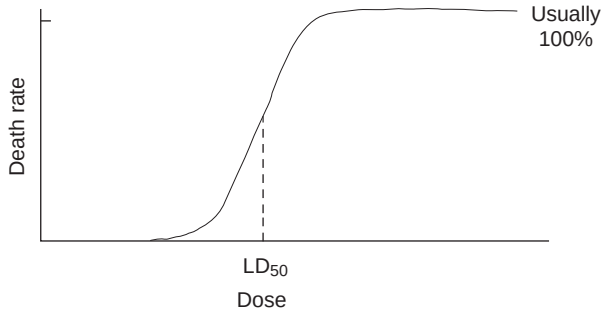


FIGURE 1 Dose–response curve, graphically representing the acute effect of a lethal dose of radiation. The LD_{50} is the dose at which 50% of persons in a population are affected. A single acute dose or an accumulated acute level of radiation over a short time causes death, whereas one-tenth the lethal dose, repeated 10 times, does not cause death.

In contrast, the chronic radiation effect of cancer seems to follow a different dose–response curve. For those dosed at the LD_{50}, cancer seems unimportant. Even of those surviving after an exposure at the LD_{50}, only about 10–20% will develop cancer. However, in contrast to acute effects, at a quarter of this dose (0.5 Sv), many persons will develop cancer. Moreover, no one has been able to determine which persons in an exposed group will develop cancer and which will not. The dose–response relationship follows a flatter curve. Various possibilities to extrapolate high doses from the data are shown in Fig. 2. These include the linear; the superlinear, in which the effect at low doses is higher than the linear; the threshold, with zero below the threshold; and one with hormesis (from the Greek, *hormaein*, to excite), which is a beneficial effect below the threshold. That a linear curve is theoretically possible was noted in 1922 by the physicist Geoffrey Crowther, who produced the first theory of radiation-induced cancer. The idea is that when a cell is ionized by radiation it will initiate a cancer. The probability of ionizing a cell in a given time is clearly proportional to the radiation intensity and hence affords a theory that cancer induction is linear with dose even at low doses. But this theory in its simplest form cannot be true. Cosmic rays and background radiation ionize millions of cells every day and yet lifetime cancer incidence is only about 30% in the U.S. population. Other effects must modify this idea drastically. Cells can be repaired; they can be excreted without leading to a cancer, and so on. Whether cancer incidence is linear with dose depends, therefore, on whether these important mechanisms vary with dose. A discussion of these

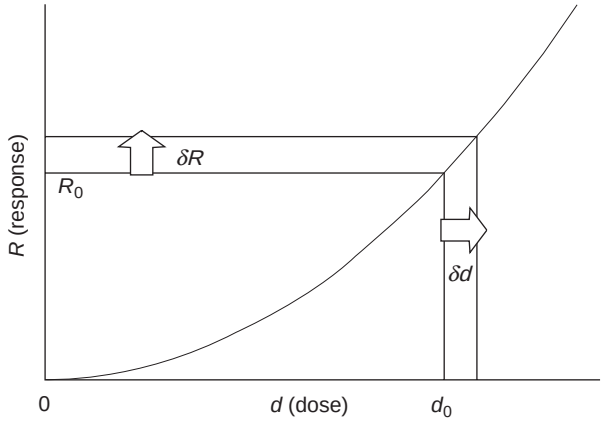


FIGURE 2 Threshold dose–response relationship($R = Ad^3$), with axes shifted to R_0 and d_0. Note that δR_0 is proportional to δd_0.

becomes very complex. Nonetheless, it is therefore generally believed that the development of cancer, whether induced by radiation or other agents, natural or anthropogenic, is basically a stochastic process.

The discussion in the preceding paragraph does not, by itself, tell us the behavior of the dose–response curve at background levels. But it raises the possibility that it is a linear dose–response, down to zero additional dose. This is the assumption that is now usually made for all cancers.

5.4 Variation of Cancer Incidence with Background Exposure

Mankind is exposed to a variety of radiation from naturally occurring sources, often called background exposures. These background exposures have been extensively measured and vary widely over the globe. The average North American or Canadian receives about 0.4 mSv (= 40 mrem) per year from internal radioactivity and cosmic rays, 0.6 mSv (= 60 mrem) per year from medical and dental diagnostics, and about 0.02 mSv per year from other miscellaneous man-made sources. Among these sources, nuclear power stations contribute less than 0.001 mSv. Collectively, background radiation exposures total about 1 mSv (= 100 mrem) per year, although persons living at high altitude have a higher radiation exposure than do those at sea level. Many studies have therefore attempted to relate health indices to altitude to deduce the effect of radiation. In many studies, cancer mortality seems to be lower in the high-altitude areas. In particular, the radiation levels in the Rocky Mountain states are higher than they are in the Gulf Coast states, yet the cancer rate is lower. This effect may be seen throughout the United States and Canada, but confounding factors may exist. Many Mormons, who do not smoke or drink alcohol or coffee, live in Utah and mountain states and seem to have half the cancer rate of their non-Mormon neighbors. Many analysts conclude that the only fact of importance from studies of geographical variation is that radiation at levels of about 2–3 mSv per year, or 0.1–0.2 Sv (10–20 rem) per lifetime, is not an important factor in developing human cancers compared to other factors.

In addition to high altitudes, there are other areas of high radiation background in the world. High levels of ambient radiation are encountered in Brazil, China, and India, for example. A population of several hundred thousand people living in coastal

areas of Kerala, South India, is exposed to natural radiation levels up to 70 mSv (=70 rem) per year, with half the population exposed to doses higher than 2.6 mSv (=260 mrem) per year. The observed incidence of cancer is about three-quarters of that in North America. This cannot be discounted as a consequence of early deaths due to other causes, because the population's life expectancy is 74 years for women and 69 years for men—only a few years less than life expectancy in North America. There has been no observation of an elevated incidence of malformations in newborns in this population. Although most experts are unwilling to accept these data as definitive evidence for a beneficial effect at low doses of radiation, these studies do show unequivocally that ordinary people who do not worry about the geographical distribution of cancer rates, which is almost everyone, should not worry about radiation at these doses.

5.5 General Argument about Effects of Low Doses

When radiation doses of concern are considerably greater than background doses, it is possible to discuss logically the effects of radiation, independently of whatever causes the background. In discussing radiation doses at levels comparable to the background, or even 100 times smaller, it is necessary to consider "introduced" and background doses together. However, very few scientists and scientific studies do this logically. The most important fact in such a discussion is that introduced radiation does not seem to induce any cancers that are not seen in the absence of radiation and are therefore also assumed to be caused in the population by "background" processes. The common idea that there is a "radiation cancer" that a physician can diagnose as distinct from any other cancer is false. The assignment of causation—that a cancer was caused by radiation—depends not on a physician's observations but on knowledge of likely exposure. Because the cancers produced by radiation and those produced by background are indistinguishable, it is likely that, at some stage in the tumor formation process, the background and radiation doses act in the same way. This is an assumption that has not been refuted, although of course it is an assumption, the validity of which must constantly be questioned. Whatever the basic biological process relating a dose to cancer, Taylor's theorem then tells us that a differential linearity results, provided that the radiation dose acts similarly to background. This argument is a general one and can apply to outcomes other than cancer. It can also apply to other cases such as respiratory problems caused by air pollution or cigarette smoking. This then makes it mandatory for any discussion of behavior related to low doses to include a discussion of what causes the natural background of cancers. It should also have implications for public policy and public perception. It is usually assumed that chronic radiation sickness and the effect on growth and development (Section 3.10) are threshold-type effects following the dose–response curve of Fig. 1. But the general argument applies to more than cancer. Either of these two medical outcomes could follow a linear dose response. In the Techa River study of chronic radiation sickness, for example, high radiation exposure was explicitly included in the definition of the disease, making it impossible to study the dose–response below that screening level.

The cancers from introduced radiation and background radiation cannot be distinguished by pathologists with typical, currently available tools. But it is possible that they can be distinguished, in principle, in which case the preceding argument would not apply. One, but only one, study suggests that the chromosome structure of an acute myelogenous leukemia (AML) occurring subsequent to, and presumably caused by, radiotherapy was appreciably different from cases of AML that occur "naturally." If this turns out to be a general result, the low-dose extrapolation arguments must be drastically reconsidered.

5.6 Epidemiological Studies of Low-Dose Behavior

There is great difficulty in measuring any effect, positive or negative, at the low radiation doses at issue. First, a study of 10,000 patients is only just enough to measure an effect, either beneficial or adverse, of 6% with a statistical significance of three standard deviations. Second, at levels of a few percent, other nonstatistical errors enter into any study. To measure reliably an increase in cancer rate of 10% would necessitate careful study of nearly 1000 patients with the disease. Another important effect is that if there is differential uncertainty in measuring the radiation dose, any structure in the dose–response curve tends to get washed out, and a threshold-type curve may appear to be linear.

The earliest studies of the survivors of the atomic bombs in Hiroshima and Nagasaki suggested (with

weak statistical significance, because they are on the edge of the natural background) a threshold at a dose of about 0.2 Sv (= 20 rem), or even hormesis. The data from the most recent follow up are better fit with linearity down to a dose of 0.05 Sv (= 5 rem). The recent analysis takes all the data together in a manner similar to a maximum likelihood analysis. It is well known that a maximum-likelihood approach can pull out of the "noise" those signals that are not clear when binned data are taken, and a Gaussian approximation is used. It is, however, admitted that there may well be nonstatistical errors that invalidate this conclusion. The authors of the recent research warn against concluding that there is any heath effect of radiation, adverse or beneficial, below 0.1 Sv (< 10 rem).

One of the first studies to suggest a threshold involved the study of bone cancers among the radium dial painters of the early 20th century. The dose to which they were exposed can be measured very reliably after diagnosis, either *in vivo* or in autopsy, because the radium stays in the bone. Two groups of dial painters have been studied, European and American. Data from the U.S. study (Fig. 3) indicated no cancers at a low dose. Linearity is statistically excluded. However, this is for alpha particle exposure and may not be generally applicable.

Workers at nuclear power plants and workers at other fuel cycle facilities have also been exposed to radiation at levels above background. A United Nations agency, the International Agency for Research on Cancer (IARC), compiled a database of 90,000 workers in the nuclear industries of the United States, Canada, and the United Kingdom. In their compilation, several workers received cumulative doses up to 0.2 Sv (= 20 rem). The data alone do not establish a clear link to any form of cancer. This is not merely a "meta-analysis" of several papers but is a combined study of those groups for which the data are deemed reliable. The additional number of leukemias seen is consistent with an extrapolation from the number at higher doses (but the lower 95th percentile of the number is close to zero). The number of additional "solid" cancers is close to zero but the upper 95th percentile is close to the linear extrapolation from higher doses. This suggests that numbers five times larger, as sometimes suggested by individual studies at a single location, are definitively excluded.

Data on workers who helped to clean up Chernobyl and live in the Russian Republic have been collected by the Medical Research Laboratory in Obninsk, Russia. There is a small, marginally

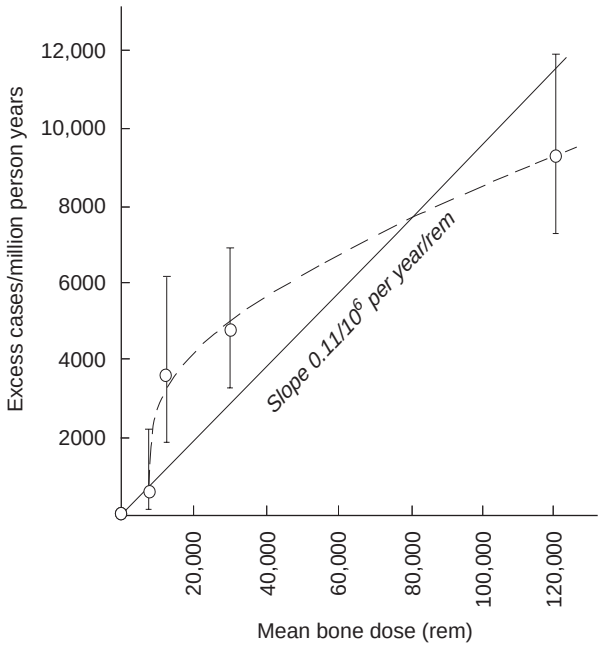


FIGURE 3 Dose–response data for a bone cancer study at Argonne National Laboratory. Study participants were exposed to radium-226 during a 20-year period (1915–1935); follow-up assessments were continued through 1971. The group included watch dial painters and some patients given radium therapeutically. Error bars include the 90% range.

significant, increase in leukemia, in agreement with the calculated linear response. Data on lung cancer among another group of cohorts, the previously discussed Mayak workers, are particularly interesting. The first studies on the Mayak workers suggested that the dose–response relationship was quadratic with dose (in qualitative agreement with animal data), and therefore a low-dose effect approaching zero seemed to make sense. But a more careful look at the data suggests that a linear dose–response relationship fits the data better. However, the doses are not low and a threshold or reduced effect at lower doses is possible.

Attempts have been made to use data on therapeutic uses of radiation to address the issue. There are few such studies with a large enough data sample. Some of these suggest that there is a dose below which there is no effect. But there are potential biases in such a study in that physicians may only look at cases in which the doses are high. There is no consensus that systematic errors are so well controlled that the studies can be used to discuss directly the shape of the dose–response curve below a total dose of 0.2 Sv (= 20 rem).

Only recently have there been studies of the effects of radon on people in residential situations. There are

two types of studies. One, an "ecological" study, compares the average lung cancer rate in a community to the average radon concentration in the houses of that community. The average lung cancer rate decreases with increasing radon concentration. It would be a logical non sequitur to derive directly from such a study the relationship of the probability of an individual person succumbing to cancer and the radon concentration to which that individual is exposed (the dose–response relationship). To make such a conclusion is sometimes called "the ecological fallacy." However, it is legitimate to compare any set of data with a theory, and if the data do not fit, the theory must be wrong. In particular, the theory esposed by the U.S. EPA, which uses a linear dose–response relationship, cannot be correct. However, the EPA error may be in not including other risk factors for lung cancer that are more important than radon. Another type of study is the "retrospective cohort" approach. There are a few such studies, in which a group of people are followed and the individual doses from radon are estimated. These are free from the ecological fallacy, but there are few data in the low-dose region, to which 90% of Americans are exposed. It is important to realize that any conclusion about the risk at low doses (that is, doses below natural background level) derived from these studies is dependent on an extrapolation, which may not be in direct disagreement with the ecological studies that use more people and are therefore statistically more significant. Most experts agree that below a radon concentration of about 20 pCi/liter, no one is sure which type of study is correct.

6. WEIGHTING FACTORS (W_r), OR RELATIVE BIOLOGICAL EFFECTIVENESS

As a first approximation, the effect of radiation on people depends only on the energy deposited per unit volume (measured in grays or rads). However, radiobiological reasoning suggests that particles traveling at speeds less than the velocity of light are relatively more effective at producing radiation effects, and a weighting factor, W_r, sometimes called the radiobiological equivalent (RBE), must be used. For high-energy X ray and gamma ray exposures, $W_r = \mathrm{RBE} = 1$. Most of the studies of effects of radiation on people are associated with ionization by particles traveling at speeds near the velocity of light, and the ionization per unit of energy loss is at a minimum. For example, the radiation from the Hiroshima and Nagasaki bombs consisted mostly of high-energy gamma rays. Data on the correct value to use for this factor for other situations are sparse. The factors are usually estimated from animal data and contain large errors. This is clear from Table I, which summarizes the latest data.

TABLE I

Approximate Radio Biological Equivalent (RBE) Factors (Simplified)

Radiation type	RBE factor
Photons	
E > 250 keV	1.0 (definition)
30 keV < E < 250 keV	1.9
E < 30 keV	2.4
Electrons	
E > 60 keV	1.0
15 keV < E < 60 keV	1.9
E < 15 keV	2.7
Neutrons	
E > 20 MeV	2.2
2 MeV < E < 20 MeV	4.4
0.1 MeV < E < 2 MeV	7.8
10 keV < E < 100 keV	3.4
E < 3.4 keV	1.7

Modified from the SENES study (see Further Reading) for probability of causation.

7. POLICY IMPLICATIONS

7.1 ALARA

The first step in discussing a sensible public policy is to summarize the data in quantitative terms. This is repeatedly done by various agencies. In the United States, the important agencies include the National Council on Radiation Protection and Measurements and the National Academy of Sciences, Committee on Biological Effects of Ionizing Radiation. In the United Kingdom, the agencies include the National Radiological Protection Board. Internationally, there is the United Nations Subcommittee for the Effects of Atomic Radiation. These organizations produce a set of predictions that, for want of a better word, is often called the "establishment" position. This establishment position assumes, for example, that the dose–response curve is linear at low doses, and provides data on the best estimate of the slope of this dose–response curve. As noted previously, this assumption remains unproved and is probably unprovable in any direct way.

The details of the establishment predictions are complex. The predictions depend on location of exposure, age at exposure, duration of exposure, rate of exposure, and gender. They concentrate almost exclusively on effects on cancer, but there are differences between the various cancer body sites. Scientists, however, tend to summarize the effects by a risk coefficient that is a sum over all fatal cancers, a weighted average over the population and for exposure to whole body irradiation. The cancer risk then is about 1/20 to 1/50 of lifetime cases per person-sievert (1/2,000 to 1/5,000 per person-rem). Alternatively stated, the dose to add 1% to the cancer incidence (which is a relative risk of 0.03) is between 0.2 and 0.5 Sv (20–50 rem). Of course, for some cancer body sites and for some ages the relative risk (RR) is much greater.

Starting in 1927, the International Commission on Radiological Protection (ICRP) noted that no threshold had been found below which adverse effects on health are known not to occur. They recommended that "no radiation exposure shall be permitted without expectation of some benefit," a view that survives today. Attempts have been made over the years to define the ICRP precautionary principle more clearly. It was first suggested that radiation exposures be kept as low as practicable (ALAP). In the early 1970s, scientists, unhappy with the vagueness of the terms, replaced ALAP with "as low as reasonably achievable" (ALARA) (economic and other factors taken into account). Europeans instead adopted "as low as reasonably practicable" (ALARP). It is unclear why anyone thought that this wording was an improvement.

In 1975, after a 2-year-long public hearing, the Atomic Energy Commission (later the Nuclear Regulatory Commission) attempted a quantitative definition of ALARA. It decided that radiation dose should be reduced if the cost per unit of integrated radiation dose (integrated over the whole population) is less than $1,000 per person-rem. (The present limit, which has been updated in response to political correctness, a change of units of radiation exposure, and inflation, is $200,000 per person-sievert.) Using a widely used establishment slope of the dose–response relationship for ionizing radiation and cancer, this corresponds to a cost of about $6,000,000 per cancer averted. It is interesting to note that the U.S. EPA has recently, and apparently independently, come up with a number of $6,100,000 as the "value of a statistical life" to be used in discussing costs to avert a cancer, for which the EPA also assumes a linear low-dose response.

On face value, this rule would suggest that an attempt to reduce exposure that costs more than $200,000 per person-sievert is overreacting to the radiation risk, and also that any such attempt should be enforced only "on the average." The reality is more complicated. In disposal of nuclear wastes, for example, this rule has not been applied and already much more money has been spent than the rule suggests. Also, it is clear that exposure to radiation of someone aged 70+ years is far less important than it is to someone aged 40 years. A 70-year-old has a shorter time to live, and the 20-year latent period for most cancers almost guarantees that he or she will not get a radiation-induced cancer.

Moreover, people have traditionally been allowed higher exposures in their work, with the obvious compensation of having a job. Thus, the National Council for Radiation Protection and Measurement (NCRPM) recommended in 1990 that the rule for occupational exposure be closer to $2,000 per person-sievert ($20 per person-rem). The radiation protection authorities are just beginning to consider such niceties, which are easily handled by standard tools of decision analysis. Despite these complexities, the application of "precautionary" policies for radiation protection has been closely tied to cost–benefit analysis, and to the conservative (but difficult-to-test) assumption that the risks from exposure at low levels is linearly proportional to the exposure.

Professional societies, unhappy with the present use of ALARA for very low doses, have emphasized the uncertainty surrounding the possible effects caused by low doses of radiation. For example, the Health Physics Society issued the following statement:

> *In accordance with current knowledge of radiation health risks, the Health Physics Society recommends against quantitative estimation of health risks below an individual dose of 5 rem (= 0.05 Sv) in one year or a lifetime dose of 10 rem (= 0.1 Sv) in addition to background radiation. Risk estimation in this dose range should be strictly qualitative, accentuating a range of hypothetical health outcomes with an emphasis on the likely possibility of zero adverse health effects. The current philosophy of radiation protection is based on the assumption that any radiation dose, no matter how small, may result in human health effects, such as cancer and hereditary genetic damage. There is substantial and convincing scientific evidence for health risks at high dose. Below 10 rem (which includes occupational and environmental exposures), risks of health effects are either too small to be observed or are nonexistent.*

7.2 Probability of Causation

When a person has cancer, society or the individual wants to know the cause and who to blame. In most

cases, the cause is, of course, not known and there may be multiple causes. It is not appropriate to calculate the "risk" of getting cancer (or indeed any other disease) when it is already known that the person has cancer, because the risk in such a situation is unity! But it is appropriate to use a different concept, the "probability of causation." Loosely, the question to ask concerns what the probability is that the cancer was caused by the particular agent of concern. The answer can be derived from the general ideas of probability using the concepts of conditional probability. The probability of causation (POC, or sometimes PC) is defined as follows:

$$\mathrm{POC} = \frac{\left(\begin{array}{c}\text{Risk of cancer from the}\\ \text{cause under study}\end{array}\right)}{\left(\begin{array}{c}\text{Risk of cancer from all causes}\\ \text{including the cause under study}\end{array}\right)}$$

There have been many theoretical criticisms of this concept and some scientists prefer the expression "assigned share." If the radiation effect is beneficial, the probability can be negative, which makes no sense. If radiation is only one of several causes, the sum of the various POCs for different possible causes can be greater than unity. But with the approximations usually used, POC makes sense. Synergistic effects are usually ignored, and the dose–response slope is linear and positive. Then the probability is positive. To evaluate this, the risk from causes other than the one under consideration is usually taken as the natural incidence of the cancer in question the assumption is then made that the population group from which the incidence data are derived is representative of the individual whose risk is evaluated in the numerator. Dividing through by the natural incidence, we derive the equation $\mathrm{POC} = \mathrm{ERR}/(1 + \mathrm{ERR})$, where ERR is the excess relative risk for the cancer in question. Of course, ERR is a function of age and gender.

In the 1950s, the United States conducted many atomic bomb tests in the atmosphere in Nevada. In some tests, U.S. army veterans were deliberately situated as observers to give them experience and ability to respond in battle without panic. In 1985, some of these veterans, and, later, residents of the states downwind of the test explosions, put pressure on Congress to give them compensation for the assumed damage. Because, as noted previously, there is no such thing as a "radiation cancer," the U.S. Congress requested the National Cancer Institute to develop a set of tables to determine the probability that a person's cancer was due to radiation exposure. These tables were published, but have since been updated into a user-operated computer program. The tables assume a linear dose–response relationship. There is disagreement on how to use the POC concept. In the United Kingdom, partial compensation is given to persons whose POC is greater than 20%. Procedures in the United States are governed by the legal maxim that "it is more likely than not" that a particular action or dose was the cause of the injury. This translates to POC $>50\%$. But few veterans would be compensated under such a rule, and it was desired politically to be generous to the veterans. An uncertainty distribution is now requested and the screening demands that only if POC is $<50\%$ with 99% certainty shall the effect be ignored.

A more logical approach would be to note that for an individual it does not matter whether a cancer developed because of bad luck in the stochastic probability or because the actual risk coefficient was on the high side of the uncertainty distribution. Then the two distributions should be combined (folded) into one. This has not been used in any situation to date.

8. PUBLIC PERCEPTION

As already noted, public perception of the effects of radiation exposure changed dramatically in the second half of the 20th century, presumably as a result of the atomic bombs in August 1945. This, combined with an unattainable desire for "zero risk," leads to considerable logical problems. For example, it has been proposed in one jurisdiction that in dismantling any building containing radioactive material, nothing with any measured radiation levels shall be placed in an ordinary landfill. If that were applied to cemeteries, it would be impossible to bury anyone, because radioactivity in a body can be easily measured. Nor could anyone be cremated, because radioactivity in the effluent would cause the crematorium to be shut down. However, zero risk is usually used to apply only to a particular action or process a particular person or group dislikes, for which a different logic applies.

SEE ALSO THE FOLLOWING ARTICLES

Air Pollution from Energy Production and Use • *Air Pollution, Health Effects of* • *Electromagnetic Fields,*

Health Impacts of • *Gasoline Additives and Public Health* • *Nuclear Proliferation and Diversion* • *Nuclear Waste* • *Occupational Health Risks in Nuclear Power*

Further Reading

Boice, J. D., Jr., and Fraumeni, J. F., Jr. (1984). "Radiation Carcinogenesis: Epidemiology and Biological Significance." Raven Press, New York.

International Atomic Energy Agency (IAEA). (1997). "Doses of Ionizing Radiation: Biological Effects and Regulatory Control." IAEA-TECDOC-976, IAEA-CN-67/63, pp. 223–226. IAEA, Vienna.

International Atomic Energy Agency, Vienna, Austria. Web site: http://www.iaea.org (Accessed in March 2003.)

International Committee on Radiation Units and Measurement. Web site: http://www.icru.org (Accessed in March 2003.)

Mettler, F. A., and Upton, A. C. (eds.). (1995). "Medical Effects of Ionizing Radiation," 2nd ed. W. B. Saunders, Philadelphia.

National Council on Radiation Protection and Measurements (NCRPM). (1995). "Principles and Application of Collective Dose in Radiation Protection." Report No. 121. NCRPM, Bethesda.

National Council on Radiological Protection and Measurements, United States. Web site: http://www.ncrp.com (Accessed in March 2003.)

National Radiological Protection Board, United Kingdom. Web site: http://www.nrpb.org.uk (Accessed in March 2003.)

National Research Council, U.S. National Academy of Sciences. (1990). "Health Effects of Exposure to Low Levels of Ionizing Radiation." BEIR V, 1990. National Academy Press, Washington, D.C.

Pierce, D. A., Shimizu, Y., Preston, D. L., Vaeth, M., and Mabuchi, K. (1996). Studies of the mortality of atomic bomb survivors. Report 12, Part I. Cancer: 1950–1990. *Radiat. Res.* **146,** 1–27.

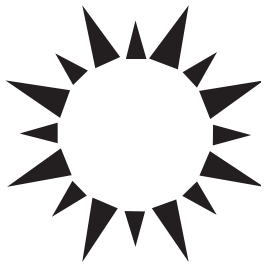
Radiation Effects Research Foundation, Japan. Web site: http://www.rerf.or.jp (Accessed in March 2003.)

SENES, Oak Ridge, Inc., Center for Risk Analysis. Probability of causation calculation and numbers. Web site: http://www.senes.com (Accessed in March 2003.)

Shapiro, J. (1990). "Radiation Protection: A Guide for Scientists and Physicians." 3rd ed. Harvard University Press, Cambridge.

United Nations Scientific Committee on the Effects of Atomic Radiation. (2001). "Report to the General Assembly, with Scientific Annex." UNSCEAR 2001. United Nations, New York.

Rebound Effect of Energy Conservation



HORACE HERRING
The Open University
Milton Keynes, United Kingdom

1. **Implications of the Rebound Effect**
2. **Price Elasticity Effects**
3. **Transformation Effects**
4. **Theoretical Argument**
5. **Empirical Evidence**
6. **Policy Implications**
7. **Conclusion**

Glossary

energy services The services that consumers want from the supply of electricity and other fuels: warm homes, cold beers, lighting, mobility, etc. A combination of energy supply and appliances (boilers, refrigerators, lamps, cars, etc.) provide these services.

price elasticity The ratio of the percentage change in energy use to the percentage change in energy price.

The rebound effect (or take-back effect) is the term used to describe the effect that the lower costs of energy services, due to increased energy efficiency, have on consumer behavior both individually and nationally. Simply stated, the rebound effect is the extent of the energy saving produced by an efficiency investment that is taken back by consumers in the form of higher consumption, in the form of either more hours of use or a higher quality of energy service.

1. IMPLICATIONS OF THE REBOUND EFFECT

Implications of the rebound effect are demonstrated by the following example: If a 18-W compact fluorescent bulb is replaced by a 75-W incandescent bulb, the energy saving should be 76%. However, this is seldom the case. Consumers who realize that the light now costs less per hour to run are often less concerned about switching it off; indeed, they may intentionally leave it on all night. Thus, they "take back" some of the energy savings in the form of higher levels of energy service (more hours of light). This is particularly the case when the past level of energy services, such as heating or cooling, was considered inadequate.

The rebound effect is a phenomenon based on economic theory and long-term historical studies, but as with all economic observations its magnitude is a matter of considerable dispute. Its significance for energy policy has increased during the past few decades with the claim by energy analysts in the 1970s, and later by environmentalists in the late 1980s, that increasing energy efficiency would lead to reduced national energy consumption and hence lower greenhouse gas emissions. Whether this claim is feasible depends crucially on the extent of the rebound effect: If it is small ($<100\%$) energy efficiency improvements will lead to lower energy consumption, whereas if it is large ($>100\%$) energy consumption will be higher. Note the use of the relative terms lower and higher: What exactly they are relative to has often been left unstated and has been a cause of much confusion in energy policy debates. Sometimes it refers to current energy consumption, whereas at other times it refers to a reduction in the future rate of growth in energy consumption.

The claim that increasing energy efficiency would lead to reduced national energy consumption was first challenged by Len Brookes in 1979 in his review of Leach's pioneering work, "A Low Energy Strategy for the UK," when he criticized Leach's approach to estimating national energy savings because of its

 © 2004 Elsevier Inc. All rights reserved.

failure to consider macroeconomic factors. This was followed in the early 1980s by similar criticism by Daniel Khazzoom of the work of Amory Lovins. The criticism of Brookes and Khazzoom was called the Khazzoom–Brookes (KB) postulate by the economist Harry Saunders in 1992. The KB postulate (sometimes referred to as the Jevons paradox) may be described as follows: Those energy-efficiency improvements that, on the broadest considerations, are economically justified at the microlevel lead to higher levels of energy consumption at the macrolevel than in the absence of such improvements.

The debate grew more intense in the early 1990s, spurred by global warming concerns and policy debate over the role of energy efficiency. It was conducted mostly in the pages of the academic journals *Energy Policy* and *Energy Journal* but also spread to the pages of *The New York Times* in late 1994 and the leading UK science magazine *New Scientist* in 1998. It culminated in two special issues of journals in 2000 devoted to aspects of the rebound effect—*Energy Policy* and *Energy and Environment*—to which many of the protagonists in this debate contributed. There is no dispute that the rebound effect exists; what is at the core of the conflict is its magnitude, and the two opposing positions can be summarized as follows:

1. Energy use is higher than if there had been no efficiency response—a position maintained by Len Brookes and, in some circumstances, Harry Saunders.
2. Energy use is lower than if there had been no efficiency response—a position maintained by Lee Schipper and colleagues.

Each side has supported its case with a mix of theoretical argument and empirical observations, based on historical data on energy use. A key problem in resolving the two positions is that it is not possible to perform control experiments to determine whether energy use is higher or lower than if there had been no efficiency improvements; after all, there is only one past. Attempts to estimate the overall magnitude of the rebound effect, using theoretical economic models based on neoclassical growth theory, have proved inconclusive, with the results dependent on assumptions about the elasticity of substitution of energy for other factors of production—the greater the elasticity, the greater the rebound. Another problem is that the rebound effect differs by energy end use and sector of the economy, and the response at the microeconomic level (the consumer) is different from that at the macroeconomic level (the national economy). Also, the rebound effect can vary between countries, depending on the energy costs and unmet demand for energy services; thus, it may be considerably higher in developing countries than in the United States.

However, all agree that the rebound effect is closely linked to the elasticity of substitution, which is a component of the widely measurable phenomenon of price elasticity. The higher the observed price elasticity of energy services, the greater the rebound effect. This leads to the paradoxical position that any imposition of energy or carbon taxes, in the hope of reducing energy use, would have its impact severely blunted by the rebound effect if consumers have a high price elasticity.

2. PRICE ELASTICITY EFFECTS

Rebound effects are due to the increased use of energy services caused by the reduction in their effective price due to greater efficiency (price elasticity) and can be classified into five categories:

1. Direct effects due to the desire of consumers to use more of any commodity due to its lower price
2. Income-related effects due to the fact that with lower energy price, more income is available to spend on energy
3. Product substitution effects due to the inclination to substitute energy services for other final consumption goods when energy services become less expensive than other final goods
4. Factor substitution effects due to the inclination to substitute energy for other factors of production (capital, labor, materials, time, etc.) in the production of final goods
5. Transformational effects due to long-term changes in the economy caused by changes in technology, consumer preferences, and even social institutions brought about by the substitution of energy for other factors of production.

The effect of perceived lower costs on energy use is termed price elasticity, which is the ratio of the percentage change in energy use to the percentage change in energy price. Price elasticities vary by commodity and over time, depending on the ability of consumers to respond to price changes through changes in behavior, substitution of alternatives, or technical change. It is important to distinguish price-induced efficiency gains from non-priced-induced gains. The former are caused by factor substitution and may involve costs to the economy (such as was caused by the oil price increases in the 1970s and

may result from the imposition of energy taxes). Non-priced-induced gains are due to technological improvements, and the rebound effects from them are of most concern to energy analysts.

Much evidence on the magnitude of the direct rebound effect derives from U.S. transportation studies for which there is good statistical data on miles traveled per vehicle and petrol consumption. The results indicate that the number of vehicle miles traveled will increase (or rebound) 10–30% as a result of improvement in fuel efficiency. Similar results are obtained for domestic energy services, such as space heating (Table I). However, the long-term price elasticity is considerably higher because the ability of consumers to respond to changes in prices is much greater: For private transport in Europe, the rebound (or long-term price elasticity) is estimated to be 30–50%.

Expansion of service demand can also take the form of consumer preference for higher levels of quality, comfort, and performance. For instance, the oil price increases of the 1970s stimulated R&D on more efficient car engines providing more miles per gallon of gas (petrol); when oil prices declined in the late 1980s car manufacturers offered consumers more powerful and better equipped cars (e.g., sport utility vehicles) with the same fuel costs but lower miles per gallon. Here, consumers appear to take the possible savings in the form of higher levels of performance or quality. Another area in which consumers "trade up" is substituting energy for time: faster cars and planes, labor-saving domestic appliances, fast food, and private rather than public transport.

An important impact of the effect of higher efficiency, and consequent lower cost of energy services, is on marginal consumers—those who could not previously afford the service. The appliance market, such as for air-conditioning, will expand as the cost of energy service declines, with the magnitude depending on the elasticity of service demand. This is because consumers' choice of the efficiency of an equipment depends on factor costs, and more efficient equipment will increase the demand for the service because of lower marginal costs per unit of service.

The second effect on consumers of lower energy costs is the income effect—their ability to spend the monetary savings on other goods and services in the economy, which will contribute to economic growth. What they spend this discretionary income on depends on their current income levels: Those with low incomes will use it for basic goods and those with higher incomes on luxury services. However, almost all goods and services involve some energy consumption in their production, distribution, and consumption, and their energy intensity varies markedly. Schipper argues that the income effect is very low because only 5–15% of income saved is respent on indirect energy use (with air travel at the high end).

What the monetary savings is spent on is crucial to the concept of sustainable consumption—that is, consumption that has low environmental impact (i.e., low material and energy content).

TABLE I

Estimates of the Direct Rebound Effect by End Use for the U.S. Residential Sector[a]

End use	Rebound effect (%)	No. of studies
Space heating	10–30	26
Space cooling	0–50	9
Water heating	<10–40	5
Lighting	5–12	4
Car transport	10–30	22

[a] From Greening, Greene, D., and Difiglio, C. (2000). Energy efficiency and consumption—The rebound effect—A survey. *Energy Policy* **28**, 398.

3. TRANSFORMATION EFFECTS

Probably the greatest effect (in the long term) of lower costs of energy services is on the direction and pace of technical change and innovation in the economy. New goods and services will be introduced to take account of the possibility of lower costs, and mass markets will be created by the continuing decline in operating costs bought about by efficiency improvements. As Len Brookes comments.

> *It is inconceivable that we should have had the high levels of economic output triggered by the industrial revolution if energy conversion had stayed where it was at the beginning of the 19th century. Energy productivity and the productivity of other factors of production fed on one another with rising energy efficiency contributing to rising productivity of other factors of production—labor and capital—and rising output contributing to rising energy efficiency by way of embodied technical progress. Without this interactive process we should not have had, in the meantime, the very large increases in energy consumption alongside large improvement in energy conversion efficiency.*

An illustration of these very large increases in energy consumption and energy efficiency due to

technical change is given here for public lighting in Great Britain for the period 1920–1995. This sector is chosen because there are good statistical series for electricity consumption, lamp efficiency, and road mileage. These statistics show an increase of more than 30-fold in electricity consumption for public lighting and 20-fold in lamp efficiency but less than 50% in road mileage. The net result of this increase in consumption and efficiency since the early 1920s is that light intensity, in terms of lumens per mile, increased more than 400 times and increased 4-fold since 1960. During the same period, energy intensity, in terms of MWh per mile, increased 25-fold, with a 250% increase since 1960 (Fig. 1).

Thus, it seems that most of the tremendous increase in lamp efficiency has been in the form of higher levels of service, both in more miles illuminated and in higher illumination levels, and not in the form of lower consumption. Interestingly, both consumption and efficiency seem to have leveled off in the 1990s, whereas during a period of rapid growth in consumption from 1960 to 1980, lamp efficiency increased by 50%.

This process of technical change is aided by manufacturers that continually seek efficiency improvements in order to lower consumer costs and create new mass markets. In conjunction with this, utilities welcome efficiency improvements to lower the cost of their product: Better to sell much at a low profit margin than little at a high margin. The massive expansion of electricity consumption in the 20th century was fueled by the continual decrease in electricity prices, bought about by a nearly 10-fold increase in generation efficiency, which spurred the development of new electrical goods and services: electric lighting in the 1900s, domestic refrigeration in the1930 s, TV in the1950 s, microwaves and videos in the1980 s, and computers and the Internet in the 1990s.

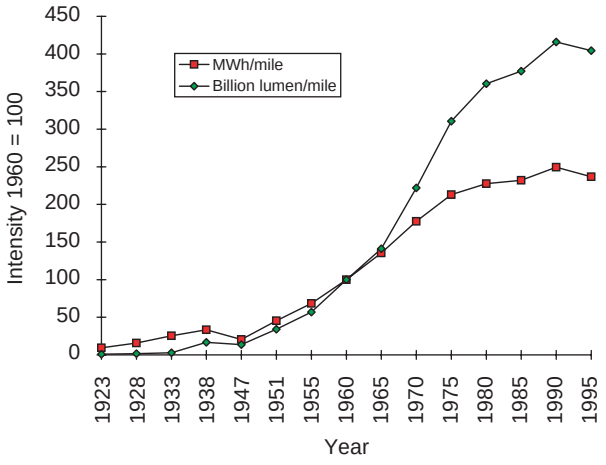


FIGURE 1 Changes in energy and light intensity for public lighting in Great Britain, 1920–1995. Reprinted from Herring, H. (1999). Does energy efficiency save energy? The debate and its consequences. *Appl. Energy* **63**, 209–226, Copyright 1999, with permission from Elsevier.

How much the improvement in energy efficiency per se stimulates economic growth is the subject of much discussion, and it is explored later.

4. THEORETICAL ARGUMENT

Key contributors in the debate over the rebound effect are economists, who seek to identify through economic theory and mathematical models the conditions under which the rebound effect may be greater than 100% (a state termed backfire). Debate by economists over the effect of efficiency improvements is long-standing, and current economists all acknowledge the work of one of the founding fathers of energy economics, Stanley Jevons, who in his classic work "The Coal Question," first published in Great Britain in1865, argued that "it is a confusion of ideas to suppose that the economical use of fuel is equivalent to diminished consumption. The very contrary is the truth."

Jevons pointed out that

> *the reduction of the consumption of coal, per ton of iron, to less than one-third of its former amount, was followed, in Scotland, by a 10-fold increase in total consumption, between the years 1830 and 1863, not to speak of the indirect effect of cheap iron in accelerating other coal-consuming branches of industry.*

Thus, Jevons laid the foundation for the idea that the rebound effect (in energy markets) could be greater than 100%, which is sometimes referred to as the Jevons paradox. The conditions under which it may occur are explored by the Austrian economist Franz Wirl and the U.S. economists Harry Saunders and Richard Howarth using extensive mathematical notion. Wirl, in his book "The Economics of Conservation Programs," argues that the extent to which more efficient appliance raises the service demand depends on elasticity of service demand:

> *The most important consequent is that (ceteris paribus) a more efficient appliance raises service demand such that actual conservation falls short of an engineering efficiency comparison. Indeed "conservation" in the sense of improving energy efficiency may raise energy demand, if demand for service is elastic.*

He demonstrates that whereas higher prices reduce unambiguously demand, the impact of efficiency is

ambiguous and an improvement in efficiency need not lower consumption. He argues, "A necessary and sufficient condition for excluding this so-called conservation paradox is that service demand is inelastic, which in turn is equivalent to the requirement that the elasticity of marginal benefit is less than -1."

He concludes that the exclusion of his conservation paradox requires a price-inelastic service demand, as for price-elastic services an efficiency improvement of 1% will lower the cost of the service by 1%, which will induce an expansion of the service of more than 1%. Hence, the expansion of service more than outweighs the conservation due to the efficiency improvement. Thus, for the rebound effect to be less than 100%, Wirl concludes that the elasticity of marginal benefit has to be less than -1 (i.e., elasticity of service demand is less than 1, which is inelastic).

Wirl observes that the rebound from efficiency improvement on electricity demand works through increasing the service demand. However, in practice some of this feedback may not lead to a physical expansion of the service demand but to upgrading, for example, cars with more features and power (larger and faster). He makes a crude assessment of this rebound effect, arguing that it is small for energy services close to saturation levels (and with low marginal costs), such as stereos or TVs, but larger for services with substantial marginal costs, such as space heating, air-conditioning, transport, and perhaps lighting. However, in the former case the absolute conservation potential is small, whereas in the latter the potential is large. This reflects the fact that consumers are more sensitive (and responsive) to end uses having a large energy cost. Thus, Wirl concludes that the rebound effect seems important for services with a significant conservation potential but negligible for services with a minor conservation potential in terms of kilowatt-hours.

Work by Saunders and Howarth, using neoclassical growth theory, seeks to identify the conditions under which backfire (rebound effect $>100\%$) occurs. The results are sensitive to the choice of production function, and for some functions they depend on the assumptions made about the elasticity of substitution between energy and other factors. The use of the Cobb–Douglas production function, with the assumption of an elasticity of substitution of 1, validates the KB postulate, although many energy analysts question its conceptual underpinnings, which rest on neoclassical economic principles. In contrast, Richard Howarth uses a Leontieff production function that assumes zero elasticity of substitution, and he argues, based on this model, that improved energy efficiency cannot give rise to increased energy use except under implausible assumptions.

Saunders lists the cases and circumstances in which the KB postulate is or is not validated. He is developing a simple economic model (or first-order tool) that will allow economists to take an arbitrary specified new efficiency technology and deploy it against an arbitrary specified production function to generate a prediction of the resulting change in fuel use, including possible rebound effects. Preliminary results show that energy efficiency technologies that enhance other factors of production may produce backfire, and that there is a marked difference in rebound between industrial sectors and between the same industries in different countries.

The impact of energy efficiency technologies in stimulating the use of other factors of production has been termed the Cashmir effect, with the observation that a technical fix to improve energy productivity can produce a bonus in materials productivity. As Ronald Sutherland comments, "Energy efficiency increases, not as a direct effort to reduce energy use, but as a result of overall productivity improvements in all inputs. In economic terms capital and energy appear to be complements not substitutes."

To resolve this issue is no easy task. Richard Howarth remarks,

> *Sorting out the empirical dimensions of the Khazzoom–Brookes hypothesis … would require detailed models that merge engineering approaches to energy efficiency, micro-engineering studies of the demand for energy services, and macroeconomic models of savings and investments. The construction of such models is an ambitious task.*

5. EMPIRICAL EVIDENCE

A number of studies have been performed to determine the impact of improved efficiency on energy use. In the United States, a major study on energy processes during the period 1880–1970 was done by Sam Schurr, whereas Lee Schipper and colleagues have done much pioneering work on energy trends for Organization for Economic Co-operation and Development countries on a sectorial level for the period 1970–1995.

However, the problem is in measuring energy efficiency because its two indicators, energy intensity (energy use per unit output) and the energy coefficient (the output elasticity of energy consumption), can give false signals. Schurr maintained that allying capital and labor inputs with new injections

of energy into economic systems can increase the productivity of both capital and labor. This results in a decrease in energy intensity due to a larger denominator in the form of higher economic output. This can deliver a false message when in fact there has been no change in the efficiency of conversion of fuel to useful heat and work. Nevertheless, it is accepted that there is a steady long-term trend in efficiency improvement in the economy due to the vintage effect. That is, the tendency for new plant and appliances to be more efficient than those they replace. Thus, it is limited by the rate of stock turnover and the rate of additions to stock, generally due to economic growth.

Schurr's empirical findings indicated that for the period 1920–1953, new technologies, often using electricity, not only increased the productivity of labor and capital but also improved energy productivity (i.e., reduced energy intensity). Energy efficiency improved at the same time that energy consumption and economic output increased. However, total output grew at a faster rate than the rate at which energy intensity declined, so total energy consumption increased. It was only in exceptional circumstances, such as the 1979 oil price hike, that energy productivity exceeded multifactor productivity, which actually declined in 1979 due to economic recession.

Other work examining U.S. energy data includes that by Bill Hogan and Dale Jorgenson, who studied time series data on energy intensity for a large number of sectors and found that there is a trend of increasing energy intensity, once price effects are carefully taken out.

The extent to which energy efficiency stimulates economic activity, and hence increased energy use, is contentious, but the effect is probably small. An examination of trends in energy intensity in International Energy Agency countries for the period 1970–1995 by Schipper and Grubb concluded that feedback effects were small in mature sectors of mature economies and only potentially large in a few cases. Their thesis is that the improvement in efficiency per se is only a small part of the reason why total energy use may have increased.

They remark that the scale of the system keeps increasing with population, household formation, and increasing incomes and sectorial output; thus, energy use must increase. However, they conclude that energy efficiency improvements, as measured by lowered energy intensities, almost always led to lower use than otherwise, and such energy efficiency improvements do restrain energy growth.

The assumption that a major part of the rebound effect will automatically manifest itself as increased output or activity levels, and thus be visible through higher energy intensities, may be incorrect. The rebound could, as Wirl points out, take the form of higher quality, performance, and comfort. A key problem in resolving the KB postulate, which is at heart a standard empirical question, is that it is not possible to run historical control experiments on society to determine whether energy use is higher or lower than if there had been no efficiency gains. This means that straightforward observation of visible macroeconomic variables, such as energy intensities, cannot provide much explanation. Instead, it may be more useful to use more sophisticated econometric tools, such as generalized technology production and cost functions, to analyze macro- or sectorial-level statistical data.

6. POLICY IMPLICATIONS

Improving energy efficiency has become, in many countries, a key part of the national strategy to deal with global warming and is founded on the belief that it would lead to reduced national energy consumption and hence lower greenhouse gas emissions. There is a long policy history of seeking national improvements in energy efficiency as a solution to environmental problems, but ultimately efficiency gains have been used to encourage economic growth and progress rather than to reduce consumption. The end result has been that economic growth has outpaced the rate of efficiency increase, and total energy consumption has increased. For instance, in the United Kingdom between 1965 and 1998, efficiency (as expressed by energy intensity, a rough proxy) doubled—a factor two improvement. However, gross domestic product increased by 147%, so total energy consumption increased by 25%. Thus, at current rates of efficiency improvement, it is feasible for there to be a factor four (or fourfold) improvement in the next century. However, as the UK Royal Commission on Environmental Protection comments,

> *There will continue to be very large gains in energy and resource efficiency but on current trends we find no reason to believe that these improvements can counteract the tendency for energy consumption to grow. Even if energy consumed per unit of output were reduced by three-quarters or factor four, half a century of economic growth at 3% a year (slightly less than the global trend for the past quarter century) would more than quadruple output, leaving overall energy consumption unchanged.*

This unequal race between energy efficiency and economic growth is well appreciated by many environmental economists and sociologists concerned with resource consumption. For instance, Mathias Wackernagel and William Rees, in their book "Our Ecological Footprint," explicitly make the point that technological efficiency may actually lead to increased net consumption of resources. They conclude

> *Ironically then, it is precisely the economic gains from improved technical efficiency that increase the rate of resource throughput. Microeconomic reality demands that these efficiency gains be used to short-term economic advantage. Far from conserving natural capital or decreasing ecological footprints, this leads to higher consumption.*

Promoting energy efficiency may not necessarily be the best way to save energy or reduce pollution. It may actually encourage energy consumption by conveying the message that consuming increasing amounts of energy is acceptable as long as energy is consumed by technologies that have been deemed efficient. Mithra Moezzi, a U.S. energy analyst, believes that changes in consumer perception and behavior are required, and that overreliance on technical solutions is mistaken. For instance, she advocates that more emphasis should be given to absolute, rather than relative, energy consumption in energy labels and standards. A larger refrigerator may be more efficient but also consumes more than a smaller one. She remarks,

> *The standards and guidelines inherent in energy policies reveal moral and technical judgments about what is "good" and what is possible. When evaluating any particular standard, guideline, or other message related to energy conservation or efficiency, it is critical to consider not only "energy saved" in a narrow sense, but what underlying messages are being conveyed and how they might affect the cultural perceptions of efficiency and consumption in short and long terms. By choosing specifications that reflect absolute rather than solely relative consumption, policies may encourage an ideological as well as practical shift in perceptions of energy use.*

Society has generally preferred technical or economic solutions, and scientists and engineers have instead emphasized the technical possibilities of a shift to less resource-intensive types of consumption (and also the development of nonfossil energy sources). One solution they advocate is the concept of service efficiency, which may be defined as providing a maximum of useful energy services to consumers using the minimum of materials and energy use. There is an extensive literature on this concept and there have been many attempts to design new types of products that deliver energy services using innovative combinations of market goods and services and household labor.

Service efficiency in itself is not a panacea for sustainable consumption because the gains are easily offset by an increase in the number and variety of products consumed. Also, as research on the cultural aspects of consumption has shown, it is necessary to understand how and why we consume. This is why innovations that are meant to be efficient and to reduce the need for resources often have the opposite effect. As F.-J. Radermacher, a German sociologist, remarked, :

> *The trap that we have fallen into again and again over the course of technical progress consists of our always using progress on top of whatever went before (the rebound effect). This effect predicts that market forces and humanity's apparently unlimited capacity for consumption will use new technology to convert more and more resources into more and more activities, functions, services, and products.*

It is often argued that information technologies such as computers will reduce material and energy consumption through substituting "virtual" for real experiences and goods. However, Finnish sociologists Heiskanen and Pantzar ask,

> *Will the information superhighway do away with the urge to travel? Will consumers actually substitute one good for another, or will they want to have it all: the television on, the newspaper on the table, and electronic news pointlessly self-scanning as the consumer of all this information dozes on the couch?*

They make the comparison with claims made in the early 1980s about the paperless office, which never happened and, in fact, office paper consumption has increased.

7. CONCLUSION

A key debate about energy efficiency concerns the extent of the rebound: To what extent does energy use increase due to efficiency improvements? Does greater efficiency lead to higher or lower energy use than there would have been without those improvements? This debate, however, is difficult to resolve, as it is with all macroeconomic arguments, since we cannot conduct such (economic) experiments on society. Thus, the rebound argument is contentious and difficult to prove either way. However, it may be possible in the long term, using good empirical research based on sound theoretical tools, to reach strong conclusions about whether it is true (or

untrue) in certain places, at certain times, or under certain conditions and thus determine its overall economy-wide impacts, but there will never be absolute scientific certainty.

Nevertheless, the school of argument that links improvements in energy efficiency with increased energy consumption—the KB postulate—has a strong claim based on the historical record dating back to the industrial revolution and the tremendous economic growth that has followed factor 10 efficiency improvements in the steam engine and other power devices. To what extent energy efficiency technologies stimulated economic growth is uncertain, and economic history, evolutionary economics, and institutional economics could shed more light on this issue than theoretical models based on neoclassical growth theory. The KB postulate may be rooted in the vast structural change in technologies, lifestyles, and social institutions that characterized the past two centuries rather than in a future world determined to reduce, or at least curtail, energy use. We need not repeat past patterns of economic growth and energy use, particularly if we practice sustainable consumption and emphasize quality of output and increased leisure time rather than quantity of output.

The rebound effect is strongly influenced by the difference in substitution potentials (elasticities) among all the factors (energy, labor, capital, and materials) and these change over time. It is very likely that energy efficiency technologies that enhance other nonenergy factors will generate substantial rebound, which will be the case in many industries. It will take some time and much research before it is known definitively whether the KB postulate is valid, and the result is unlikely to be clear-cut. It is likely that the extent of the rebound will vary depending on energy end use, industrial sectors, and even countries. Some undoubtedly will show backfire, some just modest rebound, and some may even show superconservation (i.e., a negative rebound with energy use decreasing more than the efficiency increases).

The task for energy researchers is thus to establish exactly where energy efficiency will be most effective rather than just promote energy efficiency as the solution to global warming.

SEE ALSO THE FOLLOWING ARTICLES

Consumption, Energy, and the Environment • Cost–Benefit Analysis Applied to Energy • Economics of Energy Efficiency • Energy Efficiency, Taxonomic Overview • Information Technology and Energy Use • Lifestyles and Energy • Obstacles to Energy Efficiency • Potential for Energy Efficiency: Developing Nations

Further Reading

Binswanger, M. (2001). Technological progress and sustainable development: What about the rebound effect? Ecol. *Econ.* **36,** 119–132.

Brookes, L. (1979). A low energy strategy for the UK by G. Leach *et al.*: A review and reply. *Atom* **269,** 3–8.

Herring, H. (2000a). Is energy efficiency environmentally friendly? *Energy Environ.* **11,** 313–326.

Herring, H. (Ed.) (2000b). How green is energy efficiency? *Energy Environ.* **11**(5).

Howarth, R. (1997). Energy efficiency and economic growth. *Contemporary Econ. Policy* **15,** 1–9.

Khazzoom, J. D. (1980). Economic implications of mandated efficiency standards for household appliances. *Energy J.* **1,** 21–39.

Moezzi, M. (1998). The predicament of efficiency. *In* "Proceedings of the 1998 ACEEE Summer Study on Energy Efficiency in Buildings, August 1998," pp. 4273–4282. Available: http://enduse.lbl.gov/Info/Pred-abstract.html.

Sanne, C. (2002). Willing consumers—Or locked-in? Policies for a sustainable consumption. *Ecol. Econ.* **42,** 273–287.

Saunders, H. (1992). The Khazzoom–Brookes postulate and neoclassical growth. *Energy J.* **13,** 131–148.

Schipper, L. (Ed.) (2000). On the rebound: The interaction of energy efficiency, energy use and economic activity. *Energy Policy* **28**(6/7).

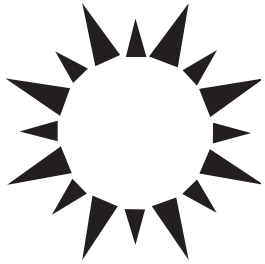
Schurr, S. H. (1990). "Electricity in the American Economy: Agent of Technological Change." Greenwood, Westport, CT.

Sutherland, R. (1996). The economics of energy conservation policy. *Energy Policy* **24,** 361–370.

Wirl, F. (1997). "The Economics of Conservation Programs." Kluwer, Dordrecht, The Netherlands.

Recycling of Metals

ERNST WORRELL
Lawrence Berkeley National Laboratory
Berkeley, California, United States



1. Introduction
2. Production of Metals
3. Recycling of Metals
4. Energy Use and Recycling
5. Future of Recycling

Glossary

home scrap Scrap produced at the metal-producing plant as the result of inefficiencies or material losses in the metals-producing plant.
new scrap Scrap that is produced at a manufacturing industry using metal in its production process.
old scrap Scrap recovered from a waste stream after consumption of the metal-containing products. This is also called postconsumer scrap.
primary production The production of metals from primary resources (i.e., ore).
recycling rate The relative amount of metal recycled as a function of apparent domestic consumption of the metal. Apparent consumption is the sum of domestic production and imports minus exports of the metal.
secondary production The production of metals from recovered resources such as scrap.

Metals are used in virtually all applications and are produced in virtually all countries. The production of metals consumes approximately 7% of all the energy in the world. The most important metals are iron and steel, aluminum, copper, and a few others. Metals are produced from primary resources (ore) and from secondary sources such as scrap metal. Because of the high energy intensity and costs of metals, metal recycling is an efficient and effective way to reduce energy consumption in the metal industries. Recycling rates vary for the various metals. Energy savings vary for the various metals depending on the type of metal, amount of metal recycled, and type of scrap recycled.

1. INTRODUCTION

Metals have contributed to the technological progress of society for an extraordinary long time and are still a key material in our society. This is unlikely to change in the future. However, the source of the metals has and will change in the foreseeable future.

Historically, metals had such a high value that they were reused and recycled. However, in modern society increasingly more metals are used in a dissipative way. The use of metals has also rapidly increased throughout society. This had led to the use of lower concentrations of ores for the production of metals. For example, copper is a relatively scarce metal. At the beginning of the 20th century, copper concentrations in ore were more than 2%. Currently, copper concentrations of mined ores are approximately 0.5% or less. This means that more than four times as much ore must be processed to produce the same amount of copper. The use of lower quality resources may have led to increased costs of these materials if it were not for technological progress in mining, ore beneficiation, and metal production technology. However, with the lower concentrations found in ores used today the potential for environmental pollution has increased. It is estimated that annually more than 6.4×10^9 metric tons of ore is extracted from the earth, of which more than 90% is discarded as mine and mill wastes. With the increasing volume of ores used, the amount of energy needed to extract and process the ores has also increased dramatically. This is unsustainable from an environmental and (long-term) economic perspective.

Recycling allows recovering part of the metals entering the industrial society, reducing the need for the extraction of resources from the natural environment. Recycling will save natural resources, reduce pollution, and save energy. The energy savings depend on the energy needed to recover and upgrade the metals from the waste streams compared to the

 © 2004 Elsevier Inc. All rights reserved.

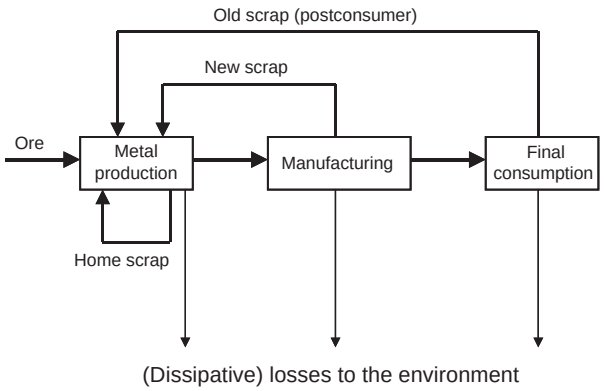


FIGURE 1 Schematic representation of scrap qualities and types.

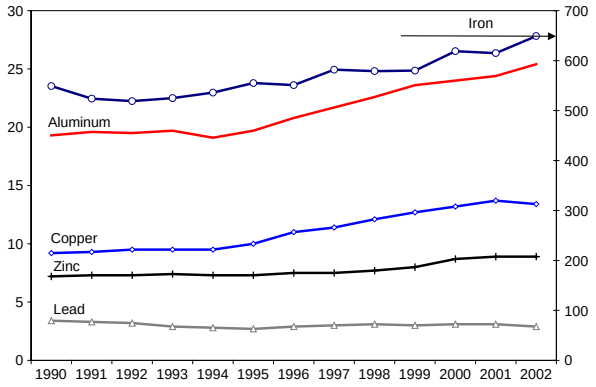


FIGURE 2 Global primary production of iron, aluminum, copper, zinc, and lead from 1990 to 2002. Iron production is depicted on the right-hand axis, whereas production of the other metals is depicted on the left-hand axis.

energy needed to extract the ores from the natural environment.

Primary production involves metals produced from ore, whereas secondary production involves recycled metals or scrap. This term is not an expression of the quality of the final product. Recycled metals can come from three streams: home scrap, new scrap, and old scrap (postconsumer) (Fig. 1). Home scrap is produced in the metal-producing plant and is the result of inefficiencies or material losses in the plant. New scrap is scrap that comes from the manufacturing industry that purchased the metal. Both for home and for new scrap, the material composition is well-known, and therefore these types of scrap demand a higher price in the global scrap market. Old scrap is scrap recovered from a waste stream after consumption of the metal-containing products. The exact composition of old scrap is not known, which may limit recycling due to contamination with other metals that may be difficult or impossible to remove in the recycling process.

Metals and scraps are typically traded on a global scale. Hence, relative prices of primary metals and scrap are determined by global developments. For example, the economic crisis in Asia in the late 1990s triggered a decline in the world prices of steel scrap because Asia is a large net importer of steel scrap.

2. PRODUCTION OF METALS

Although society uses many metals, the following metals represent the bulk of the metals used: iron (steel), aluminum, copper, lead, and zinc. This article focuses on these five metals. Figure 2 presents the global primary production of the five metals. Iron (steel) clearly dominates the metal consumption on a global scale, representing approximately 90% of all metal consumption. All ores contain the oxide of the metal. The metal oxides are reduced to metallic form to be used by society. Because the chemical reduction process is endothermic, and because many reduction processes take place at high temperatures, it requires large amounts of energy. Besides a high energy intensity, the processes used to produce primary metals are also very polluting. Pollution may be due to the concentration of the ores, the smelting process (or processes producing inputs for the smelting, such as coke making to make iron), or energy generation.

Iron is the main metal component of steel. Global steel production exceeds 900 million tonnes, whereas global iron production was 649 million tonnes in 2002. Accounting for material losses in the processes, the difference between the amount of steel and iron produced is met by recycled scrap. Steel is produced in more than 60 countries. The main iron and steel producers are China, Japan, the United States, Russia, South Korea, and Germany. Iron ore contains on average approximately 58% iron. Iron is produced in two forms: pig iron (more than 93% of global iron production) and direct reduced iron (DRI). Pig iron is produced in a blast furnace using coke in combination with injected coal or oil to reduce sintered or pelletized iron ore to pig iron. Limestone is added as a fluxing agent. Coke is produced in coke ovens. Reduction of the iron ore is the largest energy-consuming process in the production of primary steel. DRI is produced by reduction of the ores below the melting point in small-scale plants (<1 million tonnes/year) and has different properties than pig iron. DRI production is growing

and approximately 5% of the iron in the world is produced by direct reduction, of which more than 90% uses natural gas as a fuel. Steel is made by reducing the carbon content of the iron. Primary steel is produced by two processes: the open-hearth furnace (OHF; 4% of world steel production in 2002) and the basic oxygen furnace (BOF; 60% of 2002 world steel production). The OHF is still used in different configurations, mainly in Eastern Europe, India, and China. The BOF process is rapidly replacing OHF worldwide because of its greater productivity and lower capital costs. In addition, this process needs no net input of energy and can even be a net energy exporter. Hot pig iron is mixed with up to 30% recycled scrap. The process operates through the injection of oxygen. Secondary steel is produced in an electric arc furnace (34% of 2002 world steel production) using scrap. Scrap is melted and refined using a strong electric current. DRI can be used to enhance product quality. Several process variations exist and fuels can be injected to reduce electricity use. The steel is then cast and shaped into different products. Traditionally, primary steel was the steel of choice to make thin and high-quality products such as hot and cold rolled sheets. Through improved quality control of the steel manufacturing process and technological development in casting and rolling, today secondary steel is also used to produce hot and cold rolled steel.

Aluminum is one of the most abundant metals in the earth's crust. Aluminum is found as bauxite, which contains approximately 25% alumina. Hence, for every ton of aluminum approximately 8 tons of bauxite is processed. Global aluminum production was more than 25 million tonnes in 2002. Primary aluminum is produced in 43 countries. The main producers are Russia, China, the United States, Canada, Australia, and Brazil. Bauxite is first concentrated in the Bayer process to produce pure aluminum oxide, or alumina. The Bayer process uses fuels and electricity. The alumina is then processed using the Hall–Heroult process. Used for more than 100 years, the Hall–Heroult process is an electrolysis process. The alumina is dissolved in molten cryolite and electrolyzed using carbon anodes. This is a very energy-intensive process, using large amounts of electricity (generally 13 or 14 MWh/tonne of aluminum). The production of carbon electrodes from tar pitch also consumes a considerable amount of energy. Because of the high electricity intensity, aluminum production is increasingly more concentrated in countries and areas with low electricity costs. Secondary aluminum is made by melting and refining aluminum scrap. Because the scrap quality is often difficult to determine and to control, primary aluminum is typically used for high-quality applications such as thin sheet and foil, whereas secondary aluminum is mainly used for castings. Aluminum recycling rates vary for different countries depending on the primary aluminum products used. In the United States, approximately 3 million tonnes of aluminum scrap has been recovered, of which 60% came from new scrap and 40% from old scrap.

Copper has been the material of choice for electrical applications. However, competition from aluminum has reduced the role of copper in this market. Still, copper remains the third most important metal. Global primary copper production in 2002 was 13.4 million tonnes. Primary copper production is concentrated in less than 40 countries. The leading primary producers are Chile, Japan, China, and the United States. Copper ores that are currently mined have a very low concentration, typically less than 0.5%. This means that for every ton of copper more than 200 tons of ore is mined, transported, and processed. At the mine site, the ore is crushed and ground to concentrate the copper. The copper is extracted from the ore through leaching and other processes. The concentrate contains approximately 25% copper. The extracted copper oxide is then reduced in smelters at high temperatures to produce blister or anode copper. Different smelting processes are used throughout the world. The blister copper is further refined electrolytically. The smelting and refining process is highly energy-intensive. Secondary copper is produced by mixing some recovered scrap in the smelter, and it is then further refined. In 2001, only 12% of all copper was produced from secondary copper.

Zinc and lead are often found in the same ores. However, primary production is often separated. Zinc is the fourth most consumed metal in the world. Global primary production is estimated at approximately 9 million tonnes in 2002. Primary zinc is produced in 35 countries. The main producing countries are China, Canada, Australia, Japan, and the United States. Zinc is used for many applications and a relatively large portion of zinc consumption is for dissipative uses such as galvanizing steel and for chemicals. Zinc is mined from a variety of ores in coproduction with other metals. The zinc is leached from the ore and concentrated. The concentrate is often shipped to the zinc smelter. The zinc is smelted and refined in a electrolysis process using more than 4 MWh/tonne of zinc. Secondary zinc is often melted

and refined at separate plants and can also be recycled by primary zinc plants in the electrolysis process. Often, the secondary smelters produce zinc-based chemicals. In 2001, approximately 30% of the world zinc production was from secondary materials, including brass, die casting scrap, flue dust, galvanizing residues, and zinc sheet.

Lead is used for many applications, but the main market is lead–acid batteries (used in the automotive industry). Batteries represent approximately 90% of the market for lead. Lead production has remained stabile for the past decade at approximately 3 million tonnes/year. Although demand for lead–acid batteries has increased, many other uses for lead have declined (e.g., as an additive in gasoline and paint) due to environmental and health concerns related to the dissipative use of lead. Lead is found in ores with cometals such as zinc. The lead ore is most often concentrated at or near the mine. Primary lead production is concentrated in 30 countries, with the largest producers being China, Australia, Canada, the United States, the United Kingdom, and Germany. The lead concentrate is smelted at high temperatures in a blast furnace. The product of the blast furnace is refined into the final lead metal. Lead is most often processed in secondary lead smelters. Most of the recycled lead is from old scrap recovered by specialized lead-recovery companies.

3. RECYCLING OF METALS

Recycling and recovery rates vary for the different metals. The recovery rate is defined as the amount of metal that is recovered after use, whereas the recycling rate is determined by the amount of metal actually recycled. These rates can be expressed as a function of apparent domestic consumption, domestic production, or the amount of discarded metal. Different researchers and countries use different system boundaries and definitions for both of these terms. Hence, it is difficult to easily compare the recycling rates between different countries and studies. Therefore, the following discussion focuses on available data for the United States collected and analyzed by the U.S. Geological Survey (USGS). The USGS expresses recycling rates as a function of apparent domestic consumption of the metal. Apparent consumption is the sum of domestic production and imports deducting exports of metal. The USGS includes both new and old scrap in the definition of recycled metals.

Globally, more than 370 million tonnes of iron and steel scrap is recycled by the iron and steel industry. In the United States, 71 million tonnes of steel scrap is recycled, compared to an apparent consumption of 119 million tonnes in 2001. The recycling rate is estimated to be 60%. Figure 3 presents the recycling rates for the major metals in the United States. Because steel production dominates the production of metals, the recycled amount of steel also dwarfs the volumes of other metals recycled. However, as Fig. 3 demonstrates, the relative recycling rate is also high compared to that of other metals. Only lead has a higher recycling rate. Unfortunately, no data are available on the contribution of old and new scrap for iron and steel. Because most steel is used in durable products (e.g., appliances, automobiles, and construction) as well as packaging, it is relatively easy to recover and recycle. Recycling rates of iron and steel vary by the metal market. It is estimated that 65% of all the steel in an automobile produced in the United States is recycled steel. However, the recycling rate of packaging cans is probably lower because not all communities have systems in place to collect and separate the steel cans. Note that tin can also be recovered from packaging cans in specialized detinning plants. After recovery of the tin, the steel is recycled in a steel plant.

Aluminum is recovered in almost all market segments in which aluminum is used. However, most secondary smelters focus on production of aluminum die casting for the automotive and appliance industry. There are no data available for the global amount of aluminum recycled. In 2001, almost 3 million tons of scrap was recovered in the United

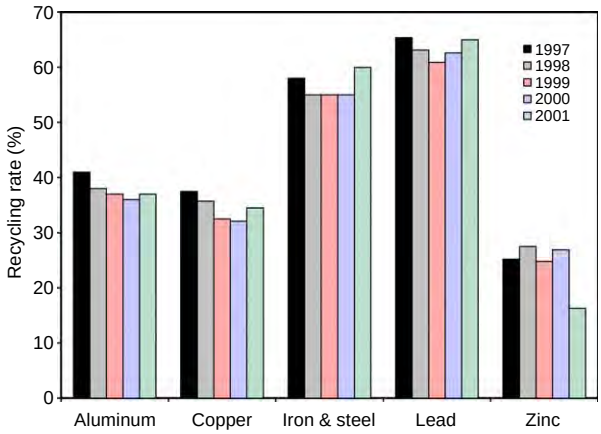


FIGURE 3 Recycling rates for selected metals in the United States for the period 1997–2001. Recycling rate is expressed as the amount of recycled metal as a share of total apparent consumption in the United States. From the U.S. Geological Survey.

States, representing a recycling rate of 37% (Fig. 3). Approximately 60% of the recycled scrap is from new (manufacturing) scrap and 40% from old scrap. In the United States, the major source of aluminum scrap is the aluminum beverage can, representing almost half of all old scrap. In 2001, 55.6 billion aluminum cans were recycled, which is equivalent to a recycling rate of 55% when compared to the manufacture of cans. Figure 4 depicts the relative shares of new and old scrap for the metals discussed in this article.

Global copper recycling was estimated to be 37% of the world copper consumption in 1997. In the United States, the total amount of recycled copper was estimated to be slightly more than 1.1 million tonnes in 2001. The recycling rate is relatively low (34.5%; Fig. 3) due to the dissipative use of copper as well as decreasing world market prices for copper. The recycling of copper concentrates on new scrap, whereas the recovery from old scrap has decreased in the past 5 years. Most copper is recovered from alloys and is mostly used by brass mills. New copper scrap represents more than 70% of all recycled copper.

Lead recycling demonstrates different dynamics than most other metals because most of the lead is recovered from postconsumer wastes (i.e., old scrap). This is due to the reduced dissipative use of lead and the concentration of lead consumption for automotive lead–acid batteries. In most industrialized countries, efficient and effective collection networks exist to recover the old batteries that are recycled by dedicated recycling companies. In the United States, approximately 96% of all recycled lead is from batteries and only 4% is recovered from new scrap (Fig. 4). In 2001, 1.1 million tonnes of lead was recycled, whereas the total apparent consumption was 1.7 million tonnes. With a recycling rate of 65% (Fig. 3), lead is currently the most recycled metal in the United States.

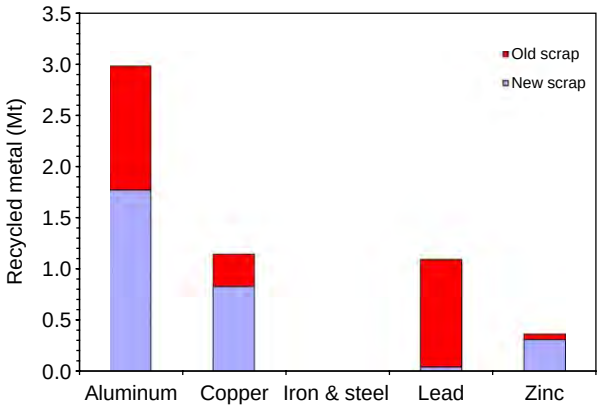


FIGURE 4 Recycled metal flows in the United States for 2001, subdivided into old (postconsumer) and new scrap. From the U.S. Geological Survey.

Zinc use is dominated by dissipative uses such as galvanizing steel products. Hence, the recycling of postconsumer wastes of zinc is very limited. Zinc can be recovered from the waste of steel recycling. Zinc is recovered from the dust of electric arc furnaces (secondary steel production) and may be recovered from the sludge from primary steel-making plants. New technological developments aim at stripping zinc from steel. Hence, zinc recycling concentrates on the use of new scrap. This limits the net recycling rate of zinc. The declining recycling of old scrap in the United States has led to a reduced recycling rate of 16% (compared to as high as 28% in 1998), of which 86% is from new scrap (Fig. 4). Worldwide, approximately 30% of all zinc is produced from secondary materials.

This discussion of the five metals has highlighted the differences in recycling rates and the impact of the dominant uses on the recycling rates. However, it also shows that relatively large shares of (postconsumer) metal wastes are not recovered. This demonstrates the potential for future increases in recovery and recycling rates, which will be especially important because the production of home and new scrap will likely decrease due to improved material efficiencies of metal production and processing.

4. ENERGY USE AND RECYCLING

Metal production comprises approximately 7% of global energy use, making the sector one of the largest energy-consuming industries. Energy consumption is dominated by the production of iron and steel. An estimated 4% of world energy use is consumed in the production of iron and steel. Table I summarizes the estimated energy intensities for selected metals. The energy consumption data include the energy used in mining of the ores. It is important to note that there are relatively large differences in the energy intensity and efficiency of industrial processes in different countries and for plants within a country. Differences in energy use for metal production in various countries are due to differences in the raw material used (i.e., ore and scrap), the mix of products produced, material losses in the production process, and the energy efficiency of the production facilities. These four factors also contribute to continuous changes in energy intensity.

TABLE I

Energy Intensities of Primary and Secondary Metal Production[a]

Metal	Primary production (million metric tons)	Energy intensity, primary production (GJ/tonne)	Energy intensity, secondary production (GJ/tonne)	Energy savings of recycling (%)
Iron and steel	649	26	12	54
Aluminum	25.4	213	11	95
Copper	13.4	91	13–37	60–85

[a] Values are approximate and will vary by product and production location. Energy intensity is expressed as primary energy. Electricity is converted to fuels used to generate the power.

It has been estimated that the energy intensity in the U.S. iron and steel industry during the period 1980–1991 changed mainly due to improved energy and material efficiency and a change toward more recycling. However, in China, during the same period only energy efficiency improvement contributed to the change in energy intensity because it has less domestic scrap available. Good statistical information on energy use does not exist for all metal industries or countries.

Table I provides estimates for the energy intensity of primary and secondary metal production. The high energy use and energy intensity make this sector an important candidate for energy conservation policy. In metals production, energy can be saved through (i) improving energy efficiency of the production process, (ii) improving material quality and material efficiency, (iii) reuse of metal-containing products, and (iv) recycling.

Table I also provides estimates for the relative energy savings of recycling metals. The data demonstrate that the largest energy savings can be achieved in the recycling of aluminum. Energy savings in aluminum recycling is estimated to be 95%. For steel, the energy savings exceeds 50% of the energy inputs of primary steelmaking. These data demonstrate the importance of recycling strategies in an integrated energy and environmental policy framework.

Increased recycling will reduce the energy intensity of metal production. The net reduction in overall energy use in a country or the world will also depend on the growth in the demand for metals and other materials. Globally, there is strong growth in the consumption of all metals, which will likely offset the energy savings of recycling. However, without recycling the total energy needs for the increased mining and primary production would be much higher than currently expected.

Several countries have started to develop integrated energy policies that account for the potential energy savings of increased recycling and material efficiency efforts. Other countries have started to account for the energy savings as part of the rationale for recycling programs. Several studies have shown the benefits of sustained recycling programs and efforts to reduce future energy use and CO_2 emissions, highlighting the importance of integration of recycling in energy policies.

The energy consumption data in Table I exclude the energy needed to collect, recover, and transport the metals. Although considerable amounts of energy can be used to collect and recover postconsumer and old scrap, these are generally small compared to the energy savings. The efficiency and effectiveness of postconsumer programs will vary depending on the type of collection system operated, sustained lifetime, geography of the area served, as well as cultural characteristics. Based on the energy used in collecting and energy savings of recycling, an "energy optimal" recycling strategy can be designed. Such a strategy may well result in different optimal recycling rates based on the regional characteristics (Fig. 5). This optimum will change with changes in technology performance. A study of aluminum recycling in Germany showed that the current recycling rate of 59% could be improved to 79% to optimize energy savings from the recycling scheme using currently available technology.

Technology development will affect the energy consumption of primary and secondary production. New energy-efficient technologies are being developed that can substantially reduce the energy consumption in steelmaking through integration of different production steps (e.g., casting and shaping of steel products) and new designs of processes (e.g., electric arc furnaces with scrap preheating). For aluminum, new developments aim to reduce the

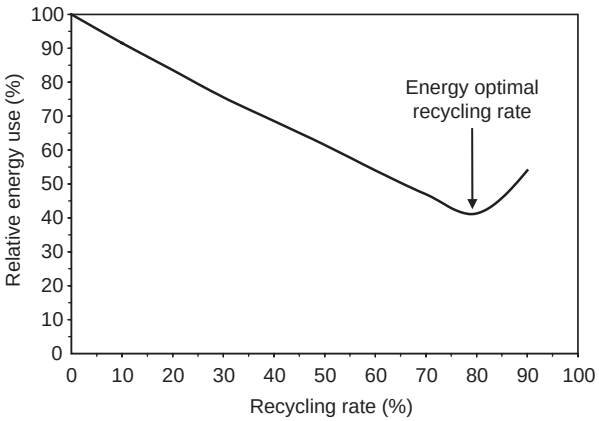


FIGURE 5 Schematic representation of energy use as a function of the recycling rate. The vertical axis depicts the amount of energy used for the production of a given volume of metal relative to the amount of energy needed to produce the primary metal (0% recycling is 100% of the primary energy use). Energy use includes the energy needed for collection, recovery, transport, and recycling of scrap.

electricity consumption of primary aluminum smelting through new electrode development and, in the long term, the development of totally new smelting processes to break through the thermodynamic barrier of the Hall–Heroult process. Secondary aluminum production is also improved through more efficient furnaces and improved yield of secondary melting furnaces. Similar developments are found for the other metals. New developments in metal recovery and separation technology will increase the recovery efficiency, improve quality control, and reduce energy intensity.

5. FUTURE OF RECYCLING

The future of metal recycling will be shaped by different driving forces. The main driving forces are depletion of high-quality ores, technology innovation, the availability of secondary resources, the costs of energy, and a drive toward more sustainable industry practices as part of sustainable development.

Depletion of high-quality ores has driven metal producers to extract ores with increasingly lower metal concentrations. Copper is an excellent example of the depletion of high-quality ore, such that ores currently exploited have concentrations less than 0.5%, whereas it was 2% at the beginning of the 20th century. This had led to exploitation of high-quality ore reserves at remote locations. Decreasing ore quality may not apply to all metals. For example, iron ore and bauxite reserves are abundant and are found in many locations. However, there is competition to use the highest quality ore first. Also, geopolitical factors may play a role in the sourcing of raw materials and hence the costs of primary production. This process will increase the costs of primary metals and make recycling more attractive.

On the other hand, the increasing costs of ore extraction are also driven by technology innovation. Innovation and the introduction of new technology in the metals extracting and primary metals industry have reduced the cost of metal production over time. Therefore, although current copper ore contains on average only one-fourth of the copper that the ore contained a century ago, copper is now cheaper. Continued innovation in the metals industry will drive down the costs of primary production. However, the same applies to secondary metal produc-tion technology. Innovative processes are decreasing the costs of metal recovery and recycling while simultaneously improving the quality of the secondary metals. The rapid growth of the secondary steel industry in the 1990s in the United States is an example of how innovative technology and low resource (scrap) costs gave secondary steel producers a competitive edge over primary (integrated) steel producers. Within a decade, the secondary steel industry expanded to provide almost half of the steel production in the United States.

As previously discussed, the sources of scrap vary for the different metals. Most secondary metal producers prefer new scrap because the exact composition of the metal is known, enabling the production of high-quality secondary metals. However, due to innovations improving the energy efficiency of metal production and processing, the amount of new scrap will further decline in the future. Hence, future recycling efforts will focus more on old (postconsumer) scrap. Costs, quality, and availability will become more important factors with regard to secondary production. Innovative technologies may alleviate some of the quality and costs problems, whereas policies aimed at the reduction of dissipative use of metals and improved scrap recovery efficiency are needed to contribute to sustainable secondary metal production.

The high energy intensity of primary metals production makes this type of production much more vulnerable to energy price changes and supply disruptions. Already, primary production of the most energy-intensive metals such as aluminum takes places in countries with low-cost energy sources. The changing energy markets in the United States have contributed to the closing of primary aluminum smelters in the

northwest. The increased importance of climate change in energy policy worldwide will affect the costs and opportunity to use fossil energy. The primary metals industry will be affected more intensively by these changers than the secondary producers.

Finally, a drive toward sustainable development of our industrial society will affect the metals industry. Climate change is just one of the challenges for sustainable development because primary metals production is also a source of many other environmental pollutants. Historically, primary metal producers were major sources of acidic air pollution (from copper, nickel, and other metal smelters) and metal dust emissions. Primary metal production also results in large volumes of solid and fluid wastes (e.g., red mud from bauxite processing and sludge containing arsenic from metal leaching processes). A drive toward sustainable business practices will drive innovation but will also support further recycling efforts as well as material substitution and product changes. This will lead to an increased importance of metal recycling.

It is difficult to assess the contribution of the different drivers to the future of the metals industry. However, it is clear that metal recycling will play a more important role in the future, leading to further reductions in energy intensity in metal production.

SEE ALSO THE FOLLOWING ARTICLES

Aluminum Production and Energy • Material Efficiency and Energy Use • Material Use in Automobiles • Recycling of Paper • Reuse and Energy • Steel Production and Energy

Further Reading

Ayres, R. U. (1997). Metals recycling: Economic and environmental implications. *Resour. Conservation Recycling* **21,** 145–173.

Fenton, M. D. (1998). "Iron and Steel Recycling in the United States in 1998." U.S. Geological Survey, U.S. Department of the Interior, Reston, VA.

Melo, M. T. (1999). Statistical analysis of metal scrap generation: The case of aluminium in Germany. *Resour. Conservation Recycling* **26,** 91–113.

Michaelis, P., and Jackson, T. (2000). Material and energy flows through the UK iron and steel sector, Part 1: 1954–1994. *Resour. Conservation Recycling* **29,** 133–158.

Quinkertz, R., Rombach, G., and Liebig, D. (2001). A scenario to optimise the energy demand of aluminium production depending on the recycling quota. *Resour. Conservation Recycling* **33,** 217–234.

Ruth, M. (1998). Dematerialization in five U.S. metals sectors: Implications for energy use and CO_2 emissions. *Resour. Policy* **24,** 1–18.

Tilton, J. E. (1999). The future of recycling. *Resour. Policy* **25,** 197–204.

U.S. Geological Survey (2001). "Recycling—Metals." U.S. Geological Survey, U.S. Department of the Interior, Reston, VA.

Worrell, E., van Heijningen, R. J. J., de Castro, J. F. M., Hazewinkel, J. H. O., de Beer, J. G., Faaij, A. P. C., and Vringer, K. (1994). New gross energy-requirement figures for materials production. *Energy Int. J.* **19**(6), 627–640.

Worrell, E., Price, L. K., Martin, N., Farla, J. C. M., and Schaeffer, R. (1997). Energy intensity in the iron and steel industry: A comparison of physical and economic indicators. *Energy Policy* **25**(7/8), 727–744.

Recycling of Paper

LINDA GAINES
Argonne National Laboratory
Argonne, Illinois, United States

1. **Introduction**
2. **Production and Recycling Statistics**
3. **Primary Paper Production**
4. **Use of Recovered Paper**
5. **Conclusion**

Glossary

chemical pulp Pulp obtained by digestion of wood with solutions of various chemicals. Chemical pulps are used to make shipping containers, paper bags, printing and writing papers, and other products requiring strength.

corrugated container A box manufactured from containerboard, layers of linerboard, and one layer of medium. The layers are combined and corrugations pressed into the medium.

de-inking A process in which most of the ink, filler, and other extraneous material are removed from recovered paper, resulting in a pulp that can be used in the manufacture of new paper.

direct energy Fuel or electric energy consumed during a process.

indirect energy Energy required to produce a material or chemical input to a process.

kraft paper Paper made from wood pulp by a modified sulfate pulping process. Its natural unbleached color is brown, but it can be bleached to white. It is used as a wrapping paper or converted into bags, envelopes, or coated paper.

kraft (sulfate) pulp A strong papermaking fiber produced by the kraft chemical process in which the active cooking agent is a mixture of sodium hydroxide and sodium sulfide. The term kraft is derived from a German word meaning strong.

mechanical pulp Wood pulp manufactured by mechanical processes. The resulting paper is opaque and has good printing properties, but it is weak and discolors easily when exposed to light due to residual lignin. Uses include newsprint, specialty papers, tissue, toweling, paperboard, and wallboard.

newsprint A lightweight paper, made mainly from mechanical wood pulp, engineered to be bright and opaque for the good print contrast needed by newspapers.

paper The name for all kinds of matted or felted sheets of fiber (usually vegetable, but sometimes mineral, animal, or synthetic) formed on a fine screen from a water suspension.

primary fiber Pulp made from trees or other vegetable matter rather than recovered materials. Also called virgin fiber.

pulp Fibrous material prepared from wood, cotton, grasses, etc. by chemical or mechanical processes for use in making paper or cellulose products.

pulp substitutes A combination of paper and paperboard trimmings from paper mills and converting plants.

recycled or secondary fiber Cellulose fiber reclaimed from waste material and reused, sometimes with a minor portion of virgin material, to produce new paper.

After paper products have served their intended uses, there are several alternative paths for material disposition. These include (generally, but not necessarily, in order of decreasing energy saved) reuse, recycling to the same product, recycling to a lower valued product, combustion for energy recovery, incineration without energy recovery, and landfill. The best option depends on the decision criterion selected.

1. INTRODUCTION

After paper products have served their intended uses, there are several alternative paths for material disposition. These include (generally, but not necessarily, in order of decreasing energy saved) reuse, recycling to the same product, recycling to a lower valued product, combustion for energy recovery, incineration without energy recovery, and landfill. Solid-waste management is an important issue in the United States, and the priorities are a subject of

 © 2004 Elsevier Inc. All rights reserved.

debate. According to the U.S. Environmental Protection Agency (EPA), "Recycling (including composting) is the preferred waste management option to further reduce potential risks to human health and the environment, divert waste from landfills and combustors, conserve energy, and slow the depletion of nonrenewable natural resources." Combustion for energy recovery is generally not counted toward material recovery goals, probably because it is viewed as a form of disposal and is therefore assumed to waste resources and have negative environmental effects. However, cocombustion in coal-fired plants or combustion in appropriately pollution-controlled waste-to-energy (WTE) plants is safe, avoids landfill costs, and can displace fossil fuels. In some cases, more fossil fuels can be displaced by combustion than by recycling. This article evaluates alternative product life cycles and discusses trade-offs among possible goals. For instance, office paper recycling conserves trees but may require greater fossil fuel input than virgin paper production. Technical and institutional barriers could arise as recycling rates increase. The main areas of concern are energy and materials use and landfill volume, but key economic and environmental factors are also identified.

2. PRODUCTION AND RECYCLING STATISTICS

In the period from 1990 to 2000, production of paper and paperboard in the United States increased from 80.5 to 96.4 million tons. This represents a 6% per capita increase, from 644 to 683 pounds per person per year, despite predictions that increased use of electronic communications would move us toward a "paperless society." Since most paper products have very short use times, postconsumer paper waste generated during the same period tracked production closely, increasing from 72.7 to 86.7 million tons. Paper and paperboard have accounted for approximately 37% by mass [and approximately an equal percentage of volume because they have close to the average municipal solid waste (MSW) density] of MSW during the period, according to EPA statistics.

This discarded material represents a valuable resource and a costly liability to discard, and efforts to utilize it have increased. The percentage of postconsumer material recovered from 1990 to 2000 doubled from 20.2 to 39.4%. Note that the paper industry reports higher recovery numbers because it includes in-house scrap, such as pulp substitutes, and materials such as unsold magazines that are recycled immediately without actually being used, as well as postconsumer material. More paper is recovered from MSW streams than all other materials combined. Corrugated boxes (recovered at a rate of 71%) account for 54% of all paper recovered, and newspapers (with a 58% recovery rate) account for another 22% of recovered paper. Office papers (of which 54% are recovered) account for another 10%; no other category contributes as much as 5%. Therefore, energy and materials implications of paper recycling are of special interest.

A significant quantity of wastepaper is exported. Some is mixed waste, sent to low-labor-cost countries where it can be economically sorted by hand, but higher grade materials are also included. Export of high-grade wastes could adversely affect recycled fiber quality and has implications for energy use. Similarly, imports have energy use implications. In general, imports of finished products displace domestic resources, but they worsen the trade balance. The costs and benefits of imports and exports of finished products and scrap must both be considered.

3. PRIMARY PAPER PRODUCTION

More than 48% of the raw material fiber used for papermaking in the United States in 2002 was recovered or reclaimed, with recovered paper contributing 37.7% and wood residues 10.6%. The total has been more or less constant, with the percentage of recovered paper increasing and that of wood residues declining recently due to their increased use in oriented strand board and composite panels. Primary fibers are a renewable, domestic resource, much of which is currently supplied by tree plantations. Short-rotation forestry can be used to further increase the productivity of these plantations. Therefore, use of primary fibers need not mean cutting old-growth natural forests, with the implied loss of habitat, wilderness, and recreation land. However, monoculture may have environmental implications of its own, as does recycling; these should be taken into account during decision making about paper production and recycling policy.

For purposes of this article, virgin paper can be classified into two main types that have different chemical compositions and physical properties. In reality, the industry produces a large variety of products, but these two classes cover most of them.

The two types are distinguished by the processes used to make the pulp: chemical or mechanical. Kraft pulp (the most common type of chemical pulp) and thermomechanical pulp (TMP), plus recycled fiber (chemical or mechanical), are the largest inputs to U.S. papermaking (85% of total pulp input), and their use is growing. This article focuses on these types. The energy and materials inputs to production of the different types of pulp differ, although both start with wood as the basic raw material. Papers made from different types of pulp are often kept separate in recycling because of the differences in material properties, but the secondary processing sequences are similar.

Most discussions of energy use in papermaking are confusing because of a failure to define exactly what is meant by energy. It is important to keep track of the form in which energy is used. The major inputs to a paper mill are wood, purchased energy in the form of electricity and fossil fuels, and water. Some components of the wood are used as a raw material and some as by-product fuel. What the industry generally reports is the process energy use at the mill, which is the sum of the purchased energy and the process by-product energy (part of the energy originally in the wood) that is used. It is more useful to define total energy input to include the entire energy content of the wood as well as the purchased fossil fuels and electricity because the wood input could otherwise have been used as a fuel. Some of this energy is not used up; it is still available from combustion of the final product. (The difference between the total energy input and the residual energy content of the product is equal to the process energy as tracked by the industry.) Total energy input, purchased and process energy, and the net nonwood energy consumption for typical paper production and recycling are shown in Table I for discarding, burning, and recycling 1 ton of printing and writing paper ("office" paper) and newsprint (representing typical products from chemical and mechanical pulps). The table shows that total energy input is reduced by recycling, but that nonwood energy use is actually increased by recycling of office paper. Recycling office paper may actually require more fossil fuel purchases than producing it from virgin materials because the recycling process does not generate by-product fuels and must therefore use purchased energy.

Caution is in order regarding energy consumption in paper mills. The numbers cited are representative of typical U.S. industry practice for specific grades and can be used to provide an understanding of the factors involved in recycling decisions. However, actual paper mill energy use varies widely, depending on many factors, including the age of the equipment, the product and its degree of bleaching, the geographical location, and the season. For instance, recycling of paperboard requires less energy than recycling of office paper and may therefore be preferable to combustion. It is hoped that the American Forest and Paper Association will publish actual industry data that will allow more accurate analysis. In addition, different mills use different mixes of purchased and self-generated electricity, affecting both the fuel mix and the primary fuel usage for generating electricity. Also, the fuel mix for electricity generation varies from utility to utility, affecting the actual primary energy usage from electricity purchases (or the fuel displaced if electricity is sold back to the grid). Table I was prepared using the U.S. average generation mix, but decisions for specific areas should employ actual local utility data.

Energy and resources used for papermaking in other areas of the world could be quite different from those used in the United States. Therefore, the conclusions of this article may not necessarily be applicable to other regions. For example, in Europe, where wood resources are more limited, clay fillers are used more extensively in the production of newsprint to produce a lighter weight, less wood- and energy-intensive product. Nonwood plant fiber pulp is used extensively in China, whereas very little is used in the United States and Europe. Nonwood plants are fast growing and produce a different fiber than slower growing trees. Many nonwood fibers are short like hardwood fibers, but others are much longer and must be shortened mechanically or chemically to optimize their papermaking value. Process energy requirements differ with fiber characteristics.

3.1 Chemical Pulping

Chemical pulping separates wood chips into their components of cellulose, hemicellulose, and lignin and retains the cellulose fibers while rejecting most of the rest; therefore, the yields are less than 50%. However, much of the material removed is used for process fuel, so energy purchases are low. The two main types of chemical pulp are kraft (sulfate) and sulfite, indicating the types of chemicals used. Most printing and writing papers, packaging, and paperboard are made from kraft pulp because of its superior strength and economics. Sulfite pulp has a relatively small and declining market share and will not be discussed further. The total direct process

TABLE I

Energy Required to Supply 1 Ton of Paper to Consumers[a]

Path	Wood[b] (dry tons)	Heat of combustion of wood (10^6 Btu)	Purchased fuel (10^6 Btu)	Purchased electricity (kWh)	Process energy[c] (10^6 Btu)	Energy recovered (10^6 Btu)	Energy displaced (10^6 Btu)	Net total energy input (10^6 Btu)	Nonwood energy input (10^6 Btu)
Wood to kraft[d] to landfill	2.2	37.4	9.5	495	36.5	0	0	52.1	14.7
Wood to kraft to WTE	2.2	37.4	9.5	495	36.5	15.6	10.9	52.1	3.8
Recycled fiber to kraft recycling	0.33	5.6	7.4	656	14.3	0	0	19.9	14.3
Wood to news to landfill	1.05	17.9	0	3079	32.3	0	0	50.2	32.3
Wood to news to WTE	1.05	17.9	0	3079	32.3	16	11.9	50.2	20.3
ONP to news to recycling	0.21	3.6	5.0	1436	20.1	0	0	23.7	20.1

Source. Gaines and Stodolsky (1993).

[a] Abbreviations used: WTE, waste to energy; ONP, old newsprint.

[b] Wood for kraft paper includes bark, which is used as by-product fuel. Newsprint input excludes bark, generally used elsewhere in integrated mills.

[c] Electricity converted at 10,500 Btu/kWh; values are lower if some of the electricity is cogenerated.

[d] Kraft refers to printing and writing grades; other paper or board types would have slightly different energy requirements.

energy requirement reported by the industry for pulping and making of printing and writing papers averages approximately 26×10^6 Btu/ton, but the energy required varies greatly with the product and with the degree of energy recovery included, ranging from a high of approximately $30–35 \times 10^6$ Btu/ton for kraft board to a low of $10–12 \times 10^6$ Btu/ton for recycled tissue. Process energy use is generally higher for products requiring extensive bleaching. Using paper that has undergone less bleaching could reduce fuel purchases more than recycling. There is a public misconception that virgin paper is white and recycled paper is brown. However, it is the degree of bleaching that determines brightness, so virgin paper can be brown and recycled paper can be white. Reducing bleaching also reduces environmental impacts. The primary purchased energy requirement for kraft paper, over half of which is bleached, is typically approximately 15×10^6 Btu/ton.

3.2 Mechanical Pulping

Products made from mechanically produced fibers include newsprint, high-bulk news (pocket books), coated publication paper, and chipboard. Newsprint is generally made from TMP, although some other mechanical pulps are still used, and some kraft pulp may be added to increase strength. The energy required to make 1 ton of newsprint (both direct and indirect energy) is all purchased (no process by-product energy) and ranges between 27.6 and 39.4×10^6 Btu/ton. This range primarily results from variation in pulping energy and electrical power-generation efficiency. It is assumed that all steam needed by the process comes from the waste heat generated in the TMP process. Therefore, these estimates include electrical power only. The mid-range estimate is 33.5×10^6 Btu/ton, which is calculated on the basis of typical pulping energy and 50% in-plant power generation.

4. USE OF RECOVERED PAPER

4.1 Paper Recycling

Wastepaper is pulped mechanically and possibly de-inked before bleaching and papermaking. These upgrading processes, used to recycle printing and writing papers, remove ink and any coatings or fillers that were used in the paper. They are reasonably energy-intensive, using approximately 2×10^6 Btu/ton (excluding energy embodied in chemicals). However, they are essential for upgrading of recycled material for utilization in high-grade products, such as printing and writing paper. High-grade recycled products therefore may have higher energy requirements than shown in this analysis and yield lower energy savings. Segregated wastepaper streams such as paperboard, newsprint, or corrugated generally do not require de-inking.

For de-inked fiber of modest quality, the yield is typically approximately 67%. This means that each ton of wastepaper yields 0.67 tons of fiber

and 0.33 tons (dry basis) of sludge, which is sent to the landfill. Sludge is difficult to dewater and at best only approximately 50% solidifies, resulting in a mass to the landfill of 0.66 tons. Thus, for every ton of de-inked fiber, 1 ton of sludge goes to the landfill or is available for combustion. Although the Office of Technology Assessment classified this sludge as noncombustible, it can be gasified or dried with stack gas and then burned. The sludge has a heating value of approximately 14×10^6 Btu/ton (dry basis) from the fiber and ink content. There was previously concern about toxic metals from printing inks in the sludge (or in the ash from sludge or MSW combustion); however, the inks that contained these materials have been phased out in the United States.

The production of secondary paper results in fiber loss. Therefore, virgin fiber is needed to maintain a constant supply of paper to the consumer. Very little newspaper is reported to be rejected at the materials recovery facility, but 15–22% of the fiber input to the mill is lost in the recycling process because of fiber breakage. There is no way to avoid this loss. Thus, the recovery rate for collected material is 78–85%. The energy needed to produce the primary pulp that makes up for the loss must be included in the energy needed to recycle. Direct energy purchases by a modern secondary newsprint mill are typically approximately 16×10^6 Btu/ton of recycled paper product; the total primary energy needed, including transportation and embodied energy in chemicals, is approximately 20.8×10^6 Btu/ton. The energy needed to supply newsprint to the consumer, assuming 22% fiber loss, is thus 23.6×10^6 Btu/ton ($0.78 \times 20.8 \times 10^6$ Btu/ton for the recycled material plus $0.22 \times 33.5 \times 10^6$ Btu/ton for the makeup material).

For recycled paper products, all of the energy is generally purchased, although some mills do generate some of their own energy by combustion of sludge. This can reduce energy purchases by as much as a factor of two. Technically, if paper is recycled within an integrated mill, some of the energy used may have been generated as a by-product of other processes within the mill. However, it cannot be credited to the recycling process and should still be considered as imported or purchased.

Recycling processes may also differ by region. Most European and Japanese mills started using flotation technology for de-inking before U.S. mills did so. However, at least one major U.S. mill still uses washing to de-ink the recovered fiber. The flotation process became dominant in Europe mainly because of effluent regulations (washing utilizes more water) and the higher recovery yields of fines and fillers (the high levels of clay fillers in European paper necessitate high levels of recovery to maximize yield).

4.2 Fiber Energy Recovery: Use of Wastepaper as Fuel

Energy can also be saved by burning waste newsprint or kraft paper (instead of recycling it) and recovering energy as electrical power or steam. This option should be called fiber energy recovery and is not a type of disposal. It should not be confused with incineration, in which wastes are burned but no energy is recovered. Approximately 10% of old newsprint (ONP) in the United States is removed from the waste stream and utilized for fiber energy recovery. The energy generated (normally electrical power) displaces energy produced from fossil and nuclear fuels at central-station power plants. When 1 ton of paper is burned in a WTE plant, 11.3×10^6 Btu of energy is displaced. If paper (news or kraft) is pelletized (either alone or as part of mixed waste) for cocombustion in coal-fired power plants, the generation efficiency of electricity produced from the waste is that of the power plant, rather than the lower efficiency of the WTE plant, and more energy is displaced. In cocombustion, each ton of paper displaces 16.8×10^6 Btu of coal. In addition, combustion can be accomplished in an environmentally sound manner; emissions of sulfur oxides (SO_x) are actually reduced by cocombustion of waste pellets, so more costly pollution controls can be avoided.

Wastepaper can be burned in either existing woodwaste- or coal-fired boilers or in new, efficient boilers designed to burn recovered fiber/wastepaper. Depending on the type of boiler, it may be necessary to pelletize the paper fuel to ensure clean and complete combustion. The energy produced can displace fossil fuel use in pulp and paper mills, cement kilns, and other industrial or commercial facilities. If electricity is cogenerated, any excess power generated can be sold back to the grid, thereby displacing additional fossil fuels. Although total energy use may not be reduced by this strategy, renewable biomass energy is substituted for fossil fuels, and the result is a reduction in emissions of CO_2 as well as SO_x, NO_x, and volatile organic compounds. In addition, the volume of solid waste is reduced. Pelletization of mixed waste paper (MWP) requires the addition of moisture; one manufacturer supplies this via paper recycling sludge, which is otherwise a disposal problem.

This technology has been used in at least one large paper mill; the boiler technology is commercially available. The boilers should be no more costly than those firing fossil fuels, although retrofits and pelletizing equipment could be needed and add to total costs. An economic factor that could hinder use of this technology is the extremely volatile price of wastepaper. This is, in turn, related to the varied mandates for recycled content in paper. Recognition of fiber energy recovery as an acceptable form of recycling could help reduce this barrier. One potential drawback to combustion for energy recovery is that it could result in competition with paper recycling mills for feedstock. This could drive up the price of recovered paper and jeopardize the economic viability of both industries.

A key institutional and economic factor is whether the paper to be burned is classified as a solid waste. If it is so defined, more stringent emission requirements may apply to the boiler in some states, and additional, costly pollution control measures may be required, negatively impacting the process economics. Requirement of additional controls is inappropriate because combustion of wastepaper actually reduces boiler emissions; therefore, either used paper should be classified as fiber fuel (with less stringent regulations) and not as a solid waste or facilities switching to it should be exempted from equipment installations specified in waste-combustion new source performance standards (NSPS) or from the "Best Available Control Technology" review on fuel switching if they can demonstrate that they have reduced their emissions to levels at or below those of the NSPS.

Compared to production of recycled office paper using fossil fuel, production of virgin office using wastepaper as the fuel would reduce CO_2 emissions by approximately 2 tons per ton of paper, in addition to reducing fossil fuel use. The result is that paper is produced with zero or negative net CO_2 impact. This is because biomass is a renewable resource. It is assumed that trees cut down to produce paper will be replanted and again sequester the CO_2, so there is no net CO_2 impact on the environment. The sulfur content of paper is much lower than that of the coal or oil it displaces; SO_x emissions may be reduced by more than expected by simply displacing coal or oil with paper because of reactions with compounds in the paper. NO_x emissions are also reduced, as is the quantity of solid waste generated.

Chesapeake Paper Products Company (no longer in business) estimated energy use and emissions for a boiler that would have burned recovered fiber at its pulp and paper mill in West Point, Virginia, as well as for an efficient new mill. In both cases, combustion of recovered fiber would have reduced emissions and fossil fuel use. Georgia Pacific Corporation cofired wastepaper pellets with low-sulfur coal in a water-tube stoker boiler at its mill in Jarratt, Virginia. It compared steady-state operation of the boiler with 100% coal feed to that with 30% coal and 70% wastepaper pellets (by weight); measured emissions of CO_2, CO, SO_2, NO_x, and hydrocarbons; and analyzed the ash. Results were very favorable.

In addition, feasibility studies were completed of combustion of MWP at three facilities in Washington State: a paper mill, a hospital, and a cement kiln. Shredded MWP could be fired in a cement kiln without the need to pelletize it, with no additional controls required because of the nature of the existing process. Hospitals and paper mills would require additional pollution controls if the paper were classed as a waste. Economics were found to be favorable if disposal costs were avoided. MWP was obtained from recycling brokers for proximate and ultimate analysis, and it was found to have acceptable fuel characteristics, but insufficient funding was available for test burns. However, data from a test burn of 50% coal/50% pellets at the University of Wisconsin showed reductions in emissions of SO_2, NO_x, and metals compared to 100% coal. Additional feasibility studies with favorable results were carried out in Canada, Oregon, and Wisconsin. A small college in Iowa has used pellets to supply much of its steam.

4.3 Effects of Recycling

The maximum purchased energy that could be saved by recycling newspapers in the United States, assuming a constant supply to the consumer, is 150×10^{12} Btu/year or the equivalent of approximately 8 million tons of coal annually ($[33.5 \times 10^6$ Btu/ton-23.6×10^6 Btu/ton$] \times 15.0 \times 10^6$ tons). The savings at current levels of recycling are lower. Similar energy savings could be achieved by burning ONP and recovering energy that would displace fossil fuels in electricity generation. However, if newspaper is recycled, the trees that would have been used for replacement newspaper are also saved. Therefore, because trees are a valuable resource, recycling of newspaper makes sense and should be encouraged. The purchased energy difference between recycling and WTE options for newspaper is

small, but the total energy input, including wood, is reduced by recycling of newsprint.

However, the situation for office paper is very different. In a modern kraft mill, paper can be made from wood with little purchased energy because of the use of by-product fuels, whereas recycling requires significant energy purchases because no fuel by-products are produced when wastepaper is pulped. Therefore, a decision to recycle office paper is a decision to use fossil fuels in place of wood. Recycling paper does keep it out of landfills, but so does combustion in a WTE or power plant, and energy recovery displaces additional fossil fuel. Combustion of paper in MSW (excluding newsprint) could displace more than 650×10^{12} Btu/year of fossil and nuclear fuels ($[60 \times 10^6$ tons$] \times [15.6 \times 10^6$ Btu/ton$] \times [0.7$, the relative efficiency of WTE plants to utilities]). This is equivalent to approximately 30 million tons of coal annually or approximately 3% of U.S. annual coal consumption.

The best option for paper depends on the objective. In the case of office paper, there is a choice. To save fossil energy and reduce landfill volume, burn wastepaper. To save trees, recycle paper. Therefore, it may be inappropriate for waste management planners and policymakers to exclude combustion for energy recovery from the acceptable waste disposition alternatives, as proposed U.S. legislation did very explicitly. Combustion for energy recovery (as opposed to incineration) as an appropriate option to recycling has been accepted in Europe, however, with The Netherlands adopting a 40% goal and Sweden 60%.

Significant energy conservation could be achieved by reuse of some paper products, such as corrugated boxes. Newspapers could also be reused. There are bins located in subway stations for collection of newspapers. The message on the bins is "Read, Ride, Recycle." They could also include "Reuse" in their list. Newspapers left by early commuters could be read by those arriving later. One possible mechanism would be to hire some of the people already soliciting money to reassemble and sell complete used papers at a reduced price. This would benefit the worker and the environment to a much greater extent than would recycling the paper after one use.

Another factor to be considered in decisions regarding paper production and disposition is emission of CO_2 to the environment, which is of concern because of possible global warming. Growing trees respire (take in) CO_2 at a rate that varies with their age. The new trees in the plantations that replace the ones cut for fiber take in more CO_2 than do the old forests. A study for Electric Power Research Institute (EPRI) estimated the effects on CO_2 emissions of various materials options for paper production and disposition. This study concluded that producing paper from plantation trees and burning the wastepaper for energy recovery is preferable, from the standpoint of greenhouse gas emissions, to all other options, including recycling. Thus, the same option for disposition of office paper is chosen if greenhouse gas reduction is the objective as when fossil fuel use and old-growth forest cutting are to be minimized.

Because a large amount of wastepaper is exported, the energy implications of used paper exports are considered briefly. For most categories of paper, U.S. consumption exceeds production, so net exports of waste imply an increase in primary production for domestic use. The balance of trade is better served by exporting higher value-added finished products. Export of old newsprint without importing newsprint increases domestic energy use and cutting of trees and does not serve the goal of energy conservation. Although export does avoid landfill, so would combustion or recycling, which would recover the embodied energy domestically. For office paper, an increase in primary production over secondary production implies greater wood use but smaller fossil fuel purchases. Therefore, if domestic energy conservation is the primary goal, it may be appropriate to encourage exports of waste office paper but not of old newspaper.

5. CONCLUSION

The most beneficial product life-cycle depends on the goals to be achieved. Trade-offs may have to be made, as in the case of kraft paper, in which trees and fossil fuels are traded off in the recycling decision. Therefore, the objectives of decision makers must be well defined before appropriate recommendations can be made. The optimal mix of reuse, recycling, and combustion for the United States may not necessarily correspond to the optimal mix for Europe and other areas of the world. Reuse is often overlooked but may be the best option when possible.

Technical and institutional barriers to increased recycling rates must be considered before mandating high recovery levels. High rates mean either separation of mixed waste, with its attendant technical difficulties, or development of improved source-separation procedures. The success of source-separation programs depends on the willingness of the

public to cooperate. Such incentives as bottle deposits have had some success; the limits of public cooperation need to be explored. In general, the cleanest, most easily recognizable waste products are recovered first. As recovery rates rise due to government mandates, economic pressures, or possibly public awareness, the quality of the material recovered will decline, and the energy and costs required to produce an acceptable product will rise. Eventually, a break-even point will be reached, at which the benefits of recycling are outweighed by the cost. This break-even point depends on the material and on the measures of cost being used.

Recycling of newspaper saves trees (a material and an energy resource) and fossil fuels, whereas combustion only displaces fossil fuels; recycling is preferable. Production of paper from chemical pulps, on the other hand, depends primarily on wood inputs for production but on fossil fuels for recycling, and the product can be burned to displace fossil fuels. More energy could be recovered by combustion of office paper than by recycling. In addition, kraft paper represents approximately 70% of the volume of material discarded, so it must be kept out of landfills. Combustion and recycling prevent paper from filling up landfills. A life cycle for office paper that included virgin production from dedicated plantations (to save old-growth forests and their value for recreation and wildlife habitat) and product combustion for energy recovery would minimize fossil fuel use as well as landfill volume. This is, in essence, substitution of biomass for fossil fuels. On the other hand, a life cycle including recycling minimizes the use of trees. The best answer depends on the grade of paper, the location, the fuel mix for the plants and the electricity supply, and the details of the process design. All of the alternatives must be evaluated objectively before decisions are made.

Acknowledgments

I thank the many helpful people in the paper industry who provided real-world feedback on the information in this article, enabling important updates and clarifications. Thanks especially to Benjamin A. Thorp at Georgia Pacific, who provided information, references, and personal contacts. Those who provided particularly useful input include Tom Brown at Caraustar, Jim Ramp at SP Newsprint, and Dedric Bergstrom at Glatfelter.

SEE ALSO THE FOLLOWING ARTICLES

Material Efficiency and Energy Use • Recycling of Metals • Reuse and Energy

Further Reading

American Forest and Paper Association web site: http://www.afandpa.org.

Butner, R. E., Kangas, M., and Stapley, C. E. (1993). Paper recycling: Impact on electricity-use, electro-technology opportunities, Report RP-3228-06. H. A. Simons Strategic Services Division for Electric Power Research Institute, Decatur, GA.

Environmental Defense Fund, Duke University, Johnson & Johnson, McDonald's, The Prudential Insurance Company of America, and Time Inc. (1995, December). White paper No. 3, lifecycle environmental comparison: Virgin paper and recycled paper-based systems. (Updated February 2002).

Gaines, L. L., and Stodolsky, F. (1993, December). Mandated recycling rates: Impacts on energy consumption and MSW volume, Report ANL/ESD-25. Argonne National Laboratory, Argonne, IL.

Office of Solid Waste and Emergency Response, U.S. Environmental Protection Agency (2002, June). Municipal solid waste in the United States: 2000 facts and figures, Report EP1530-R-02-001. Office of Solid Waste and Emergency Response, U.S. Environmental Protection Agency, Washington, DC.

U.S. Department of Commerce, Bureau of the Census (2002). "Statistical Abstract of the United States: 2002," 122nd ed. U.S. Department of Commerce, Washington, DC.

Refrigeration and Air-Conditioning

BOGDAN D. HORBANIUC
Technical University of Iaşi
Iaşi, Romania

1. **Introduction, Applications, and Historical Background**
2. **Basic Principles of Refrigeration**
3. **Reversed Thermodynamic Cycles**
4. **The Air–Standard Refrigeration Cycle**
5. **Vapor–Compression Refrigeration Cycles**
6. **Absorption Refrigeration Systems**
7. **Steam Jet Refrigeration**
8. **Other Methods of Refrigeration**
9. **Cryogenics**
10. **Air-Conditioning**

Glossary

air conditioner System composed of a refrigeration machine, heat exchangers, air ducts, filters, fans, and/or a humidifier that supplies air at specified parameters (temperature, humidity, and purity).

air-conditioning The simultaneous control of the temperature, humidity, and cleanliness of the air in a space as required either by the thermal comfort of persons that live or work in the respective space or by specified technological prescriptions.

Carnot cycle Thermodynamic cycle having theoretically the highest efficiency; used as a criterion in evaluating the effectiveness of a real cycle.

coefficient of performance Thermodynamic efficiency of a refrigeration cycle, expressed as the ratio between the heat removed from the low-temperature body and the external work consumed in order to remove that heat and to transfer it to the high-temperature body (surroundings).

cryogenics The deep-refrigeration domain involving temperatures less than 120 K.

refrigerant Substance in the fluid state that transfers to the surroundings the heat removed from the cooled space.

refrigeration The creation and maintenance within a medium of a temperature lower than that of the surroundings.

refrigeration system Complex aggregate of interrelated machines, heat transfer equipment, and control devices that together accomplish the refrigeration task.

The second law of thermodynamics, as stated by Clausius, asserts that heat can be transferred from a body of low temperature to a body of higher temperature only if external work is expended. The high-temperature body is called cooled medium (space) and the fluid that carries the removed heat and transfers it to the surroundings is called refrigerant. To be able to accomplish these tasks, the refrigerant must perform a sequence of thermodynamic processes the effect of which is the refrigeration of the cooled medium. The temperature domains of refrigeration correspond to air-conditioning (approximately 5°C), industrial refrigeration (0 to −150°C), and cryogenics (−150 to −273°C).

1. INTRODUCTION, APPLICATIONS, AND HISTORICAL BACKGROUND

The applications of refrigeration cover almost all domains of human activity: the food industry (processing, storage, and transportation of aliments), air-conditioning, ice manufacture, chemical industry (removal of the heat of reaction or of the heat of mixing, salts separation from solutions, and gases liquefaction), civil engineering and mining (freezing of moisture in unstable water-bearing soils), the manufacturing industry (heat treatments and metal cutting), the pharmaceutical industry, biomedical applications (blood and tissue storage and hypothermia in surgery), and fundamental research (atomic and nuclear physics).

 © 2004 Elsevier Inc. All rights reserved.

In 1755, William Cullen obtained ice by evaporating water under low pressure. In 1805, Oliver Evans described theoretically the operation of the refrigerant's vapor compressor. J. Leslie developed the first absorption refrigeration system using the water–H_2SO_4 solution in 1810. In 1823, Michael Faraday published the first classical work on gases liquefaction and liquefied SO_2, H_2S, CO_2, and NH_3. Sadi Carnot described the refrigeration cycle in 1824. The first patent on a mechanically driven refrigeration system belongs to Jacob Perkins (1834, London), who proposed ether as a refrigerant. In 1844, John Gorrie, an American physician, constructed the first compression refrigeration machine using air as refrigerant. The first viable commercial system was put into operation in 1857 (James Harrison and D.E. Siebe) using ethyl as a refrigerant. The first absorption refrigeration machine was patented by Ferdinand Carré in 1859 and was designed to use the water–ammonia solution. At the beginning, these machines were used to produce ice. More than 600 absorption refrigeration machines had been sold in the United States by 1876. A. C. Kirk patented the reversed Stirling air refrigeration cycle in 1861. Carl von Linde (Germany) developed the first air liquefaction machine in 1895, and in 1908 Kammerlingh Onnes (Holland) liquefied helium. The first primitive version of a household refrigerator dates from 1898. In 1910, Maurice Leblanc (France) constructed the first steam jet refrigeration machine. Migley, Henne, and McNary (United States) used chlorofluorocarbons (CFCs) for the first time as refrigerants in 1930, and in 1933 the adiabatic demagnetization of paramagnetic salts (theoretically developed in 1926 by Debye and Giauque) was applied by W. F. Giauque and D. P. MacDougall to obtain temperatures of approximately 0.7 K. The lithium bromide–water solution was utilized after 1940 in industrial applications of the absorption refrigeration cycle (the Carrier and Trane systems). Small refrigeration machines that work on the basis of the reversed thermoelectric effect have been in use since 1958 to cool electronic components. In the 1980s, the adiabatic demagnetization of the nuclear spin system of copper allowed a temperature of 10^{-8} K to be reached.

The applications of refrigeration in the area of food processing and preservation have given impetus to the economic development of the meat-, fruits-, and vegetables-producing countries (usually located in the warm climate zone), allowing them to export these goods long distances. The evolution of the perishable foodstuff refrigeration techniques was consequently strongly related to the development of these countries. In 1873, Charles Tellier presented to the Academy of Sciences a note regarding fresh meat refrigeration. In 1875, the ship Paraguay, equipped according to Carré's specifications, transported refrigerated meat at -30°C from Buenos Aires to Le Havre, France. In 1876, Tellier uses for the first time an absorption refrigeration system to preserve meat, fruits, and vegetables on the ship Le Frigorifique during its journey from Rouen, France, to Buenos Aires. In 1882, Argentina used for the first time refrigeration systems in slaughterhouses.

Refrigerated fruits were marketed for the first time in the United States in 1910. In approximately 1930, refrigerated vegetables were marketed for the first time. Refrigeration trucks have been in use in the United States, Italy, and Great Britain since 1933. In 1938, the Hoveman quick refrigeration method (direct contact under vacuum of the foods contained in latex sacks) was developed.

2. BASIC PRINCIPLES OF REFRIGERATION

The refrigerant passes through the refrigeration system by continuously removing heat from the cooled medium and rejecting this heat to the surroundings (atmospheric air). To accomplish its task, the refrigerant must be in a fluid state. Due to the fact that the latent heat of vaporization–condensation is much higher than the sensible heat, the change of state of the refrigerant is preferred in order to maximize the transferred heat (the refrigeration effect). Consequently, in most refrigeration systems the cooling effect is achieved by the evaporation of a liquid, a process that extracts the amount of heat necessary to vaporize the fluid. By using a fluid that boils at a low temperature, heat is removed from the respective body (space) by cooling it. The greater the removed heat (latent heat of vaporization), the more intense the cooling effect. The refrigeration processes that occur in such a manner that the refrigerant undergoes a single time the heat removal from the cooled medium are called open-cycle processes. Some examples of open-cycle processes are (1) Water evaporation in atmospheric air: water cools the air by removing from it an amount of heat equal to the latent heat of evaporation, and (2) Endothermic mixtures: by using substances that require heat to mix, the formation of the mixture removes the heat from the space that

contains the mixture. Such endothermic mixtures contain different percentages of water and salts such as sodium chloride, ammonium chloride, calcium chloride, and ammonium carbonate.

Refrigeration processes that occur when the refrigerant, after undergoing a sequence of physical transformations, regains its initial state and repeats the same sequence of transformations a number of times are called closed-cycle processes. The basic refrigeration techniques (vapor–compression refrigeration, absorption refrigeration, and steam jet refrigeration) are examples of closed-cycle processes.

2.1 Methods of Refrigeration

Thermodynamic methods

- Open-cycle processes (evaporation of a liquid and refrigerant mixtures)
- Closed-cycle processes (mechanical vapor–compression, absorption refrigeration, steam jet refrigeration, and expansion of a compressed gas in centrifugal field of forces)

Electrical and magnetic techniques

- Reversed thermoelectric effect
- Adiabatic demagnetization of a paramagnetic salt

Methods used in cryogenics

- Gas throttling (the Joule–Thomson effect)
- Expansion of a gas with work output
- Reversed Stirling cycle
- Linde method
- Claude method
- Diffusion pump method.

The refrigeration capacity has been defined as the heat rate necessary to freeze a short ton (907 kg) of ice in 24 h. One ton of refrigeration = 12,000 Btu/h = 3.51 kW.

3. REVERSED THERMODYNAMIC CYCLES

Thermodynamic cycles allow the reciprocal conversion in a continuous manner of thermal energy into other kinds of energy. A work-supplying cycle is referred to as the power cycle (Fig. 1). Such cycles are traced clockwise and they remove heat from the hot reservoir, supply work, and reject heat to the cold reservoir. A reversed cycle is traced counterclockwise and consumes external work in order to cool the cold reservoir by transferring heat to the high-temperature reservoir (Fig. 1).

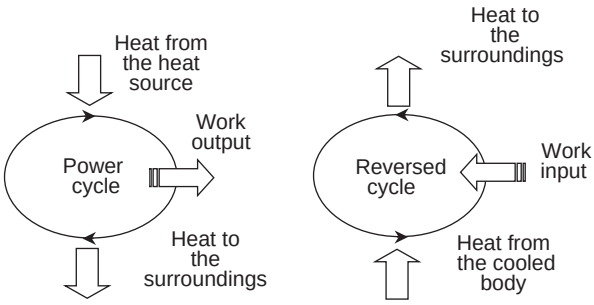


FIGURE 1 Power and reversed cycles.

Reversed cycles are used either to extract heat from bodies colder than the surroundings (cooling effect), referred to as refrigeration cycles, or to heat spaces by extracting heat from low-temperature sources (heating effect), called heat pump cycles.

The efficiency of a cycle is evaluated by the coefficient of performance (COP). In general, the effectiveness of a cycle is the ratio of the desired effect to the energy expended to obtain this effect. In the case of power cycles, the desired effect is the net work transfer, and the expended energy is the heat extracted from the high-temperature reservoir. For the reversed cycles, the desired effect depends on the purpose of the cycle, whereas the energy expense is the same, namely the external work input:

$$\mathrm{COP} = \frac{\text{desired (useful) effect}}{\text{work input}}.$$

If the purpose is the refrigeration of a space, then the desired outcome is the refrigeration effect (i.e., the heat removed from the low-temperature body):

$$(\mathrm{COP})_{\mathrm{refrig}} = \frac{\text{refrigerating effect}}{\text{work input}}.$$

The actual COP of the refrigeration systems is on the order of 3.

On the other hand, if the purpose of the cycle is to heat a space, the heating effect (the heat rejected to the high-temperature body) represents the required result:

$$(\mathrm{COP})_{\text{heat pump}} = \frac{\text{heating effect}}{\text{work input}};\ (\mathrm{COP})_{\text{heat pump}} = 4-5.$$

The reversed Carnot cycle is the reversible (ideal) cycle that has the thermodynamically highest possible COP and is composed of two isothermal and two isentropic (adiabatic) processes. Figure 2 shows the Carnot refrigeration (a–b–c–d–a) and the Carnot heat pump (1–2–3–4–1) cycles traced in the temperature–entropy plot. The processes that compose

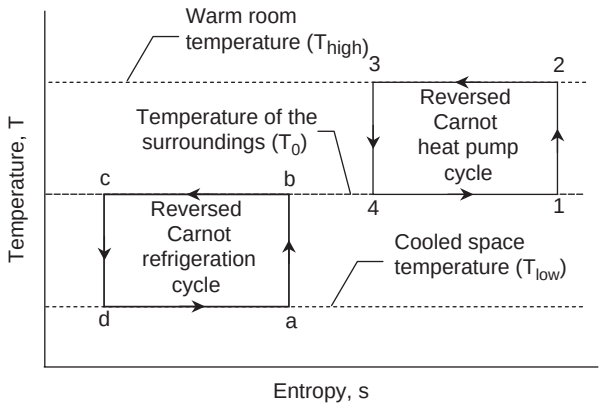


FIGURE 2 Reversed Carnot cycles.

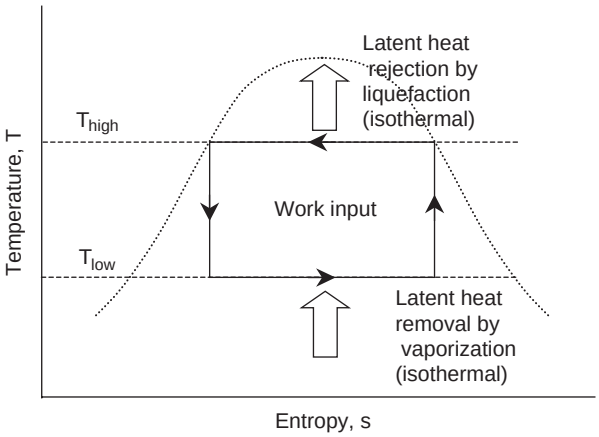


FIGURE 3 Reversed Carnot refrigeration cycle with liquefaction and vaporization of the refrigerant.

the Carnot refrigeration cycle are the adiabatic (isentropic) compression a–b from T_{low} to T_0, the isothermal heat rejection b–c to the surroundings at T_0, the adiabatic (isentropic) expansion c–d from T_0 to T_{low}, and the isothermal heat removal d–a from the cooled space at T_{low}. The COP of such a cycle is

$$(\text{COP})_{\text{refrig}}^{\text{Carnot}} = \frac{1}{(T_0/T_{\text{low}}) - 1}$$

The Carnot heat pump cycle consists of the adiabatic (isentropic) compression 1–2 from T_0 to T_{high}, the isothermal heat rejection 2–3 to the warm room at T_{high}, the adiabatic (isentropic) expansion 3–4 from T_{high} to T_0, and the isothermal heat removal 4–1 from the surroundings at T_0. The COP of a Carnot heat pump cycle can be expressed as

$$(\text{COP})_{\text{heat pump}}^{\text{Carnot}} = \frac{1}{1 - (T_0/T_{\text{high}})}$$

Perfect isothermal processes cannot be obtained with gases as working fluids since such processes require either infinitely low rates of expansion or compression to maintain the temperature rigorously constant at a finite heat transfer rate or infinitely high rates of heat transfer at finite rates of expansion or compression. The only way to obtain the isothermal processes of a reversed Carnot cycle is to use a fluid as refrigerant that liquefies from the gas state by ejecting its latent heat of vaporization at the higher temperature and that vaporizes at the lower temperature of the cycle by removing from the cooled space an amount of heat equal to the latent heat of vaporization. This cycle is represented in Fig. 3.

There are two major drawbacks to the reversed Carnot cycle with phase-change refrigerant:

1. The adiabatic compression between T_{low} and T_{high} occurs in the liquid–vapor region, which is why it is called wet compression. Due to the presence of the liquid phase, the liquid droplets may damage the compressor blades (erosion and pitting).

Solution: The compression should occur in the superheated vapor region.

2. The expansion involves a low-quality liquid–vapor mixture and therefore the turbine's efficiency is extremely low. Although this process provides work for the compressor, the price to be paid in terms of efficiency and difficult operation conditions does not justify the presence of the expander–turbine. The advantage of the net work production is surpassed by the losses.

Solution: The expansion process in the turbine is sacrificed in favor of a more thermodynamically effective process: throttling. The expander is replaced by a throttling valve, which destroys the potential energy of pressure and reduces the refrigerant's pressure from p_{high} to p_{low} but makes the system more reliable and simple.

The result is the so-called vapor–compression refrigeration cycle.

4. THE AIR–STANDARD REFRIGERATION CYCLE

The adiabatic expansion of a gas is accompanied by a significant decrease in temperature. Therefore, this phenomenon can be used in refrigeration cycles. When refrigerant leakage may occur in the cooled space occupied by living beings, or when air-conditioning is required (such as in the case of aircraft), air is the recommended working fluid. Other advantages result from the fact that air is available in unlimited quantities and involves no

manufacturing costs. Due to the low COP of the gas refrigeration systems, their use is recommended only when very well justified.

The air–standard refrigeration cycle is the reversed Brayton cycle. The direct Brayton cycle is the basic cycle of operation of gas turbine power systems.

4.1 Reversed Brayton Cycle

The cycle is composed of two isentropic processes (compression and expansion), an isobaric heat rejection to the surroundings, and an isobaric heat removal from the cooled space.

The reversed Brayton cycle requires relative high pressure ratios and large volumetric flow rates to achieve modest refrigeration capacities. For usual applications, vapor–compression or absorption refrigeration systems are more suitable from the standpoint of refrigeration capacity per system size. However, by modifying the scheme gas refrigeration systems can be used to liquefy air. Use of a regenerative heat exchanger, which cools the air below the temperature of the surroundings, leads to an important temperature decrease after the expansion, which means that the cooling effect is much more intense.

Figure 4 shows the scheme of the system and the temperature-entropy plot of the cycle. The processes that comprise the cycle are the adiabatic compression (1–2) of the air from p_{low} to p_{high} (in the compressor), the isobaric heat rejection (2–3) to the surroundings, the adiabatic expansion (3–4) of the air from p_{high} to p_{low} (in the turbine), and the isobaric heat removal (4–1) from the cooled space. The turbine supplies a fraction of the necessary work input for the compressor.

The functions of the environmental control system of the aircraft are to adjust the cabin temperature, pressure, and humidity; cool avionics; defog the windshield; and supply conditioned air to various systems and subsystems. The heat that is removed from the previously mentioned parts of the aircraft is transferred to the ambient air and to the vehicle's fuel supply. Figure 5 shows the schematic of the air refrigeration system for aircraft cooling and the cycle's *T–s* (temperature-entropy) diagram.

Atmospheric air is absorbed by the compressor, and it heats up and rejects heat to the surroundings. The cooled air expands in the turbine and its temperature decreases under the prescribed cabin temperature. The cooled air leaving the expander is mixed with the warm cabin air, and the result is refreshed air with the imposed temperature. In addition, a fraction of the cooling air is directed to the avionics and other systems that require cooling. The dashed line in Fig. 5 represents a fictitious process necessary to close the cycle.

The COP of the cycle strongly depends on the pressure ratio p_{high}/p_{low}: The lower the ratio, the higher the COP. Consequently, the compression ratio is limited to the values in the range 3–5. On the other hand, this fact leads to a decrease of the refrigeration effect (the heat extracted from the cooled body). To increase this amount of heat, regeneration is required: The air leaving the cooled space cools the air that leaves the heat transfer surface of the hot end of the refrigeration system. The air leaving the cooled space (state 1) is cold enough to cool the stream of air that has rejected heat to the ambient (state 3). Consequently, a regenerative heat exchanger is placed before the expander in order to achieve this supplementary cooling.

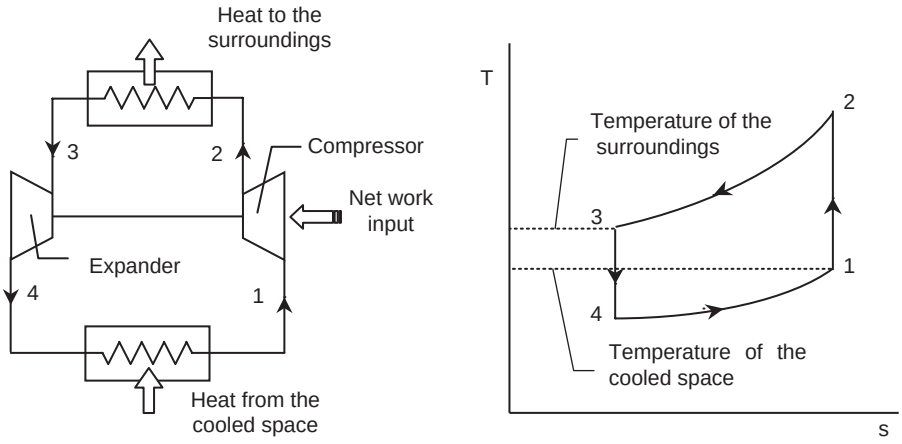


FIGURE 4 Scheme of the typical reversed Brayton refrigeration system and the temperature–entropy diagram of the cycle.

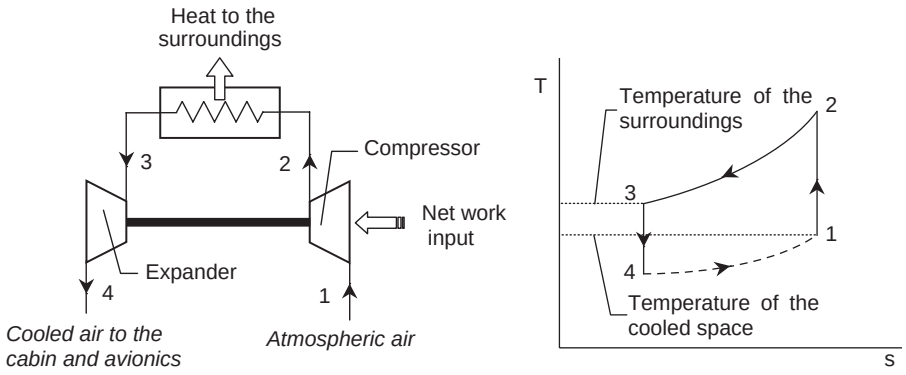


FIGURE 5 Open-cycle arrangement of the reversed Brayton refrigeration system for aircraft cooling.

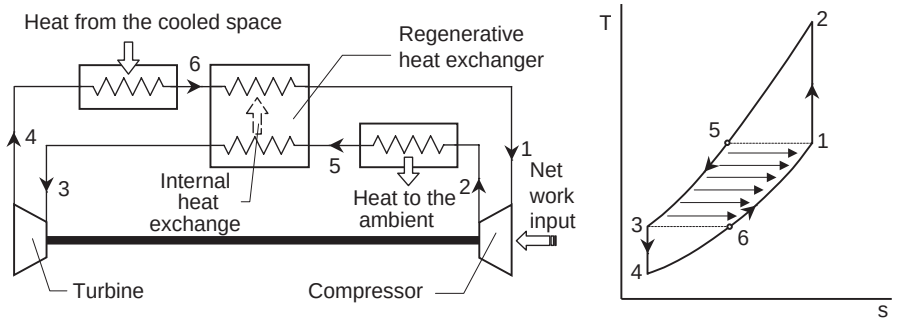


FIGURE 6 The air cycle with heat regeneration.

Figure 6 shows the reversed Brayton refrigeration system with heat regeneration. The processes are basically the same as for the reversed Brayton cycle. In addition, an internal heat exchange occurs in the regenerative heat exchanger: The air that rejected heat to the surroundings undergoes a supplementary cooling (process 5–3) by rejecting heat to the air that leaves the cooled space (process 6–1).

In order to reduce the work required by the system and thus to increase the COP, compression may be divided into two stages with intercooling. The pressure ratio for the two compression stages (compressors) must be the same so as to minimize the work input.

4.2 Bootstrap System

When using ram air as a heat sink to reject the generated heat within an aircraft, the corresponding takeoff gross weight (TOGW) penalty can be defined as the sum of the ram air circuit equipment weight and the additional fuel. Optimizing the aircraft performance means minimizing the ram air circuit TOGW penalty, and consequently the most light-weight and simple systems are required. This is the case for air–standard refrigeration systems. Due to severe weight limitations, most of the aircraft cooling systems use the open-loop bootstrap configuration.

Open-loop bootstrap systems use bleed air from the aircraft's engine compressor. This stream is compressed by the bootstrap compressor, cooled by the ram air, and passes through a turbine that drives the bootstrap compressor. The temperature of this stream after passing through the turbine is lower than the temperature of the load and thus cooling is performed and heat is rejected overboard together with the air.

Figure 7 shows a schematic of an open-loop bootstrap system with heat regeneration. Air is compressed in the aircraft engine compressor from the ambient pressure to the combustor pressure. In the lower stages of the engine compressor, a fraction (called bleed air) is extracted at an intermediate pressure and is sent to the bootstrap compressor, where its pressure and temperature are increased. Next, the air stream passes through the regenerative cross-flow heat exchanger, where it undergoes a preliminary cooling by rejecting heat to the "used" air that leaves the cabin. In the ram air heat exchanger, the ram air coming from the surroundings cools the air. In the bootstrap turbine, air is expanded to the cabin pressure and to a temperature lower than

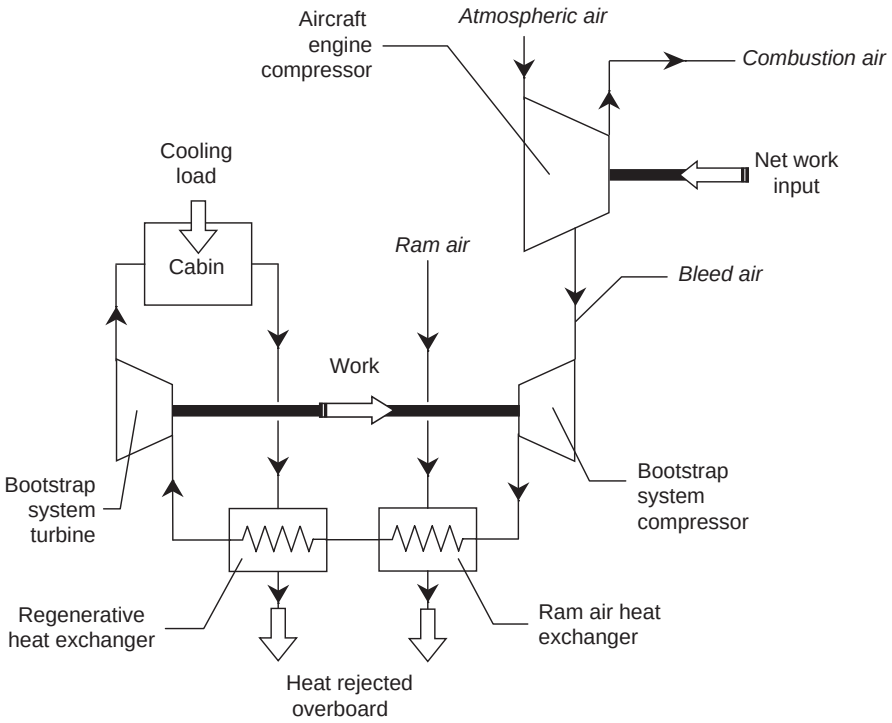


FIGURE 7 Open-loop bootstrap system for an aircraft.

the imposed cabin temperature. By flowing through the cabin and the aircraft systems, it removes the cooling load and afterward is sent to the regenerative heat exchanger to cool the incoming air. Finally, the used air is thrown overboard, as is the ram air. In the open-loop bootstrap system, the power supplied by the bootstrap turbine equals the shaft work rate absorbed by the bootstrap compressor.

5. VAPOR–COMPRESSION REFRIGERATION CYCLES

As discussed previously, the major drawbacks of the reversed Carnot cycle with vaporization and liquefaction make the Carnot refrigeration machine inappropriate for practical applications. By replacing the turbine (expander) with a throttling device and moving the compression into the superheated vapor domain, these shortcomings are eliminated and the result is the vapor–compression refrigeration (VCR) cycle.

5.1 Basic Vapor–Compression Refrigeration Cycle

The four basic components of a VCR system (Fig. 8) are discussed here. The compressor increases the refrigerant's pressure from p_{low} to p_{high} (the isentropic process 1–2 in Fig. 8) and consequently its temperature becomes higher than that of the high-temperature region. In the condenser, the refrigerant rejects to the surroundings (or to a cooling medium) its latent heat of condensation and liquefies (the isobaric process 2–3). The expansion (or throttling) valve reduces the refrigerant's pressure (the isenthalpic process 3–4) from p_{high} to p_{low}, which is characterized by a boiling point below the temperature of the cooled space. In the evaporator, the refrigerant boils by removing heat from the refrigerated space (the isobaric process 4–1) and turns into superheated vapor that is sucked by the compressor. This process is repeated (Fig. 8).

The condenser and the evaporator are heat exchangers and the expansion valve is a capillary tube or a short-tube orifice. The throttling process occurs at constant enthalpy, which is why it is called isenthalpic. In modern refrigeration systems, an active throttling device that senses the temperature in the evaporator replaces the capillary tube and adjusts the pressure drop in order to control the cooling effect.

The expansion of the high-pressure liquid in the expansion valve results in a low-quality, two-phase fluid. The COP of the vapor–compression

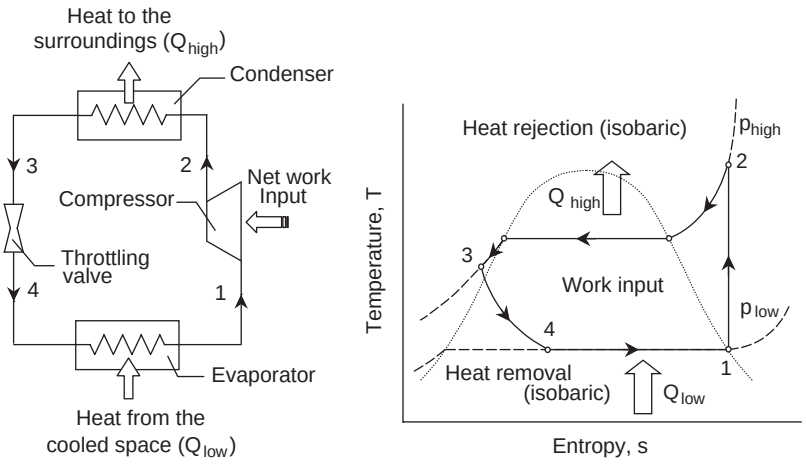


FIGURE 8 Scheme of the basic vapor compression refrigeration system and the cycle representation in the temperature–entropy diagram.

refrigeration system is given by

$$\text{COP} = \frac{h_3 - h_1}{h_2 - h_1}$$

where h is the specific enthalpy.

Irreversibilities inevitably occur in actual processes, and therefore the actual performance of the vapor–compression refrigeration system is less effective than that of the ideal cycle. Such an irreversible process is the nonisentropic compression in the compressor, which results in larger work input and an additional vapor superheating that increases h_2 and thus decreases the COP. Another effect that reduces the COP (h_3 decreases) is the slight liquid supercooling after the condensation due to the variation in temperature of the surroundings.

The lowest cooling temperature of the system is determined by the pressure ratio of the compressor and by the characteristics of the refrigerant. In a simple cycle, typical cooling temperatures vary from -50 to 2°C. Reciprocating compressors operate with pressure ratios up to 9:1. Outside the previously mentioned temperature range, either the compressor performance becomes unacceptably poor or the risk of excessive heating endangers lubrication, bearings, etc. By staged compression, the inferior temperature limit is lowered to -100°C. There are limitations regarding the refrigerant's characteristics: the high specific volume at low temperatures, which necessitates large size equipment; the too small specific volume of the condensed liquid; and the too high condensing pressure. In this case, the solution is to use cascaded systems with two or more refrigerants, each of which operates in a separate cycle. The advantage of this approach is the possibility to choose for each cycle the best refrigerant from the standpoint of properties, which ensures the best possible operation. With cascaded systems, the lowest temperature that can be obtained is approximately -150°C.

Although centrifugal compressors may operate with pressure ratios up to 18:1, due to the low efficiency caused by the high temperature at the end of compression, the practically acceptable pressure ratio is the same as that for reciprocating compressors. To increase the efficiency, interstage cooling is used in the so-called economizers. In such equipment, relatively cool gas is mixed with the hot gas resulting from the compression process.

5.2 Refrigerants

More than 30 refrigerants are known, of which 10 are the most used. Refrigeration involves a large variety of applications and parameters. Consequently, one refrigerant cannot meet all the requirements. This is why there are different refrigerants.

5.2.1 Requirements

The general criteria to be met by a refrigerant candidate are

- High latent heat of vaporization in order to minimize the amount of refrigerant used
- Freezing point below the operating temperature at the cold end of the system
- Condenser pressure sufficiently low to avoid too high mechanical stresses in the components of the refrigeration system

- Evaporator pressure above the atmospheric pressure in order to eliminate the risk of leakage of air and moisture into the refrigeration system
- Critical parameters (temperature and pressure) well above the operating ones
- Low specific volume in the suction section of the compressor; large specific volume requires a larger size compressor and high cross section areas of the low-pressure piping
- Low manufacturing cost
- Good heat transfer properties (thermal conductivity, specific heat, and viscosity)
- Nonflammable, nontoxic, noncorrosive, and nonexplosive.

The toxicity of refrigerants, combined with their flammability, is rated by a class system. Class A designates refrigerants that are nontoxic at concentrations less than 400 parts per million (ppm) and class B refers to refrigerants that are toxic at less than 400 ppm in the air. Odorless toxic refrigerants should contain a warning agent (e.g., acrolein, which is an irritant to eyes). The flammability is evaluated by numbers ranging from 1 to 3 (1, nonflammable; 2, moderately flammable; 3, highly flammable). For instance, a refrigerant rated A2 indicates that it is nontoxic at less than 400 ppm and moderately flammable. Ammonia is rated B2.

5.2.2 Classification

The refrigerants are classified as showed in Table I. Almost half of the vapor–compression refrigeration systems use ammonia as refrigerant. Other working fluids usually referred to as halocarbons or halogenated hydrocarbons are the CFCs and the hydrochlorofluorocarbons (HCFCs), which are generally

TABLE I

Classification and Physical Properties of the Refrigerants

Class	Chemical name	Chemical formula	Number designation	Normal boiling point (°C)	Critical temperature (°C)	Critical pressure (bar)
Inorganic compounds	Water	H_2O	—	100	374	221.29
	Ammonia	NH_4	R-717	−33.35	132.4	113
	Carbon dioxide	CO_2	R-744	−78.5	31	73.3
	Sulfur dioxide	SO_2	R-764	−10.01	157.5	78.8
Hydrocarbons	Methane	CH_4	R-50	−161.5	−82.5	44.4
	Ethylene	C_2H_4	R-1150	−103.5	9.5	51.2
	Ethane	C_2H_6	R-170	−88.6	32.2	48.9
	Propylene	C_2H_6	R-1270	−47.7	91.4	46
	Propane	C_3H_8	R-290	−42.3	96.8	42.6
	Isobutane	C_4H_{10}	R-600	−11.7	113.7	36.7
Chlorofluorocarbons	Tetrachloromethane	CCl_4	R-10	76.7	283.14	45.6
	Trichlorofluoromethane	CCl_3F	R-11	23.7	214.1	34.1
	Dichlorodifluoromethane	CCl_2F_2	R-12	−29.8	112	41.15
	Chlorotrifluoromethane	$CClF_3$	R-13	−81.5	28.8	38.6
	Tetrafluoromethane	CF_4	R-14	−128	−45.5	37.5
	1,1,2-Trichlorotrifluoroethane	CCl_2FCClF_2	R-113	47.7	214.1	34.1
	1,2-Dichlorotetrafluoroethane	$CClF_2CClF_2$	R-114	3.5	145.7	32.8
	Monochloropentafluoroethane	$CClF_2CF_3$	R-115	−38	80	32.4
Hydrochlorofluorocarbons	Dichloromonofluoromethane	$CHCl_2F$	R-21	8.9	178.5	51.66
	Chlorodifluoromethane	$CHClF_2$	R-22	−40.8	96	49.36
	Dichloromethane	CH_2Cl_2	R-30	39.2	235.4	59.7
	Chloromethane	CH_3Cl	R-40	−24	143	66.8
	Chlorodifluoroethane	CH_3CClF_2	R-142	−9.25	136.45	41.5
	Trifluoroethane	CH_3CF_3	R-143	−47.6	73.1	37.8
	Difluoroethane	CH_3CHF_2	R-152	−25	113.5	44.9
Azeoltropes mixture	R-22, R-115	—	R-502	−45.6	90	42.6

nonflammable and nontoxic. CFCs and HCFCs have proven to be harmful to the ozone layer and contribute to the greenhouse effect responsible for global warming. For this reason, national and international regulations have restricted their use, recommending their substitution with other fluids that are not harmful.

Hydrocarbons such as propane and butane are cheap and suitable as refrigerants, except they are flammable. However, careful handling, a good system design, and the avoidance of leakage by appropriate safety measures reduce or eliminate the risks of explosion.

Mixtures of two or more refrigerants are also used. They are of two types: zeotropic mixtures, which exhibit a saturation temperature range (called temperature glide) at constant pressure, and azeotropic mixtures, which behave like a single component, evaporating or condensing at a fixed temperature at constant pressure (Table I).

5.2.3 Environmental Problems, Regulations, and Substitution Agents

In 1986, world production of CFCs was 950,000 tons. The highest CFC quantities are used in air-conditioning (house and automotive); in the United States and Japan, 40 and 25%, respectively, of R-12 production in 1986 were directed to this area of applications. Automobile production in 1990 was 40 million pieces (40% air-conditioned).

In 1974, Rowland and Molina discovered that CFCs destroy the ozone layer, which acts like a shield for life on Earth against the cosmic ultraviolet (UV) rays. In 1985, a 40% depletion of the ozone layer was observed. The mechanism of the phenomenon, which takes place in the stratosphere, consists of the reaction (triggered by UV radiation) of a molecule of CFC with ozone (O_3), the result of which is the liberation of an atom of chlorine, which combines with ozone to form oxygen and a molecule chlorine–oxygen. This molecule may react with a free atom of oxygen, resulting in molecular oxygen and a free chlorine atom, which again reacts with ozone and the process is repeated.

The ozone-depletion potential (ODP) of a refrigerant defines the potential damage caused by the refrigerant molecule to the ozone in the stratosphere and is evaluated based on the estimated life of the refrigerant in the atmosphere. It seems that ozone depletion is proportional to the squared concentration of chlorine in the atmosphere. This concentration is approximately 3 parts per billion in volume, and it is estimated that it will double by 2020. The life span in the upper atmosphere is 120 years for R-12, 90 years for R-113, and 65 years for R-11. CFC molecules are nontoxic and chemically stable (inert) in the lower atmosphere and in the troposphere (up to 10 km in altitude). They become chemically active in the stratosphere, where they are decomposed by the cosmic UV radiation and are destroying the ozone layer.

Another harmful effect of the CFCs and HCFCs relates to the fact that they are opaque to the infrared radiation emitted by the earth as a result of its exposure to the sun's thermal radiation. As a result, infrared thermal radiation is confined in the system Earth–atmosphere and thermal energy is accumulated instead of being emitted, leading to global warming. The previously described mechanism constitutes the greenhouse effect (the glass covering a greenhouse acts in a similar manner by heating the space of the greenhouse).

The intensity of this phenomenon associated with a CFC or HCFC is evaluated by means of the halocarbon global warming potential (HGWP). For instance, a molecule of R-12 has the same effect as 5000 molecules of CO_2.

CFCs exhibit the highest ODP and HGWP. HCFCs are less stable and therefore their ODP and HGWP are correspondingly lower. Hydrofluorocarbons (HFCs) have zero ODP since fluorine does not react with ozone, but some of them are characterized by equal or larger HGWP than that of the HCFCs. Hydrocarbons have zero ODP and HGWP. Figs. 9 and 10 illustrate the relative ODP and HGWP of some of the most popular CFCs.

Due to their harmful effects, regulations have been issued by governments in order to ban or to reduce the use and production of these substances. In the

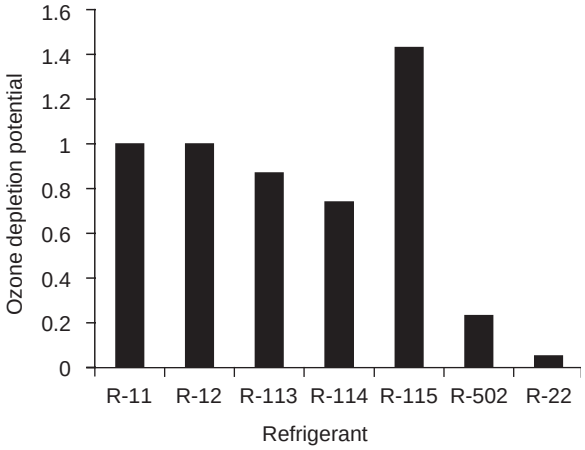


FIGURE 9 The relative ozone-depletion potential of some of the most used CFCs (normalized with respect to R-11).

late 1970s, the United States and Canada prohibited the use of CFCs as aerosols. In 1987, the Montreal Protocol was signed by 43 countries, which agreed to freeze the CFCs production at the 1990 level and to gradually decrease production by 20% by 1994 and by 50% by 1999. In the United States, the production of CFCs was to cease by 1996 and that of HCFCs is scheduled to cease by 2020. The European Community, after amendments made in London in 1990 and introduced in the Copenhagen Protocol in 1992, decided to eliminate the use of CFCs by 1995. The most promising substitution agents seem to be the HCFCs and the HFCs. Unfortunately, by increasing the amount of hydrogen in the molecule, the flammability increases, whereas by increasing the amount of fluorine, the HGWP increases. Therefore, the Copenhagen Protocol states that the HCFCs should play only a transitory role by 2030. After this deadline, only HFCs will be accepted.

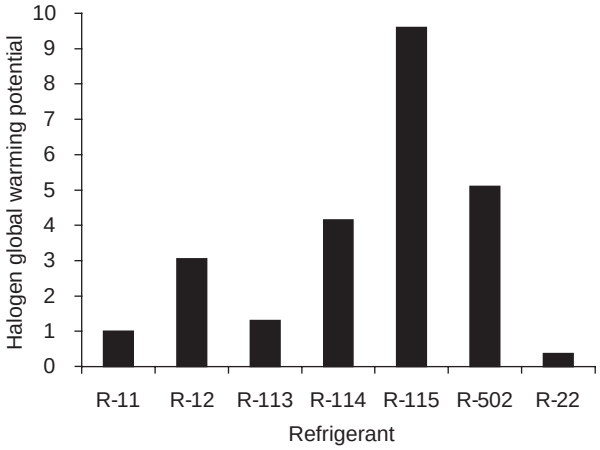


FIGURE 10 The relative halogen global warming potential of some of the most used CFCs (normalized with respect to R-11).

The refrigerants to be replaced are R-10–R-12 and R-113–R-115. Proposed substitution agents (all of them HFCs and HCFCs) are R-141b, R-123, R-22, R-125, R-143, and R-124:

R-123 ($CHCl_2CF_3$) replaces R-11 and R-113: ODP = 0.5 and HGWP = 0.1.
R-141b (CH_3CCl_2F) replaces R-11 and R-113: ODP = 0.08 and HGWP = 0.092.
R-134a (CH_2FCF_3) replaces R-12: ODP = 0 and HGWP = 0.285.
R-290 (propane) replaces R-12: ODP = 0 and HGWP = 0.

An overview of the substitution solutions that are considered is presented in Table II.

TABLE II

Substitution Substances for the CFCs[a]

	Transition refrigerants			Chlorine-free alternatives for medium and long term			
		HCFCs mixtures			HCFCs mixtures		
Refrigerants to be replaced	Molecular HCFCs	Type	Composition	Molecular HFCs	Type	Composition	Halogene-free refrigerants
							Propane (R-290)
		R-401A		R-134a			
R-12	R-22	R-401B	R-22/152a/124	R-152a	—	R-134a/152a	
							Isobutane (R-600)
		R-402A		R-32			
					R-404A	R-143a/125	
		R-402B	R-22/125/290				
					AZ50	R-143a/125/132	
R-508	R-22			R-125			R-290, NH_3
					FX40	R-143a/125/132/134a	
		R-403A	R-22/218/290				
					RX4	R-32/125/134a	
		R-403B		R-143a			
				R-134a	AC9000	R-32/125/134a	R-290, NH_3
R-22	R-22	—	—				
				R-32	AZ20	R-32/125	

[a] Abbreviations used: CFCs, chlorofluorocarbons; HCFCs, hydrochlorofluorocarbons; HFC, hydrofluorocarbons.

5.3 Variations of the Vapor–Compression Refrigeration System

Some solutions will improve the performance or the flexibility of the vapor–compression refrigeration systems. The most used are presented here.

5.3.1 Multiple Evaporators

These systems operate at more than one evaporator temperature. The stream of liquid refrigerant coming from the condenser is divided and each fraction is throttled at a different pressure, which determines the evaporation temperature. The higher pressure fractions, after turning into vapor in the corresponding evaporators, pass through valves in which they expand in order to match the lowest pressure and to allow mixing. Such systems are used to cool different temperature loads.

5.3.2 Multistage Compression

The reason why such systems are used is to minimize the compressor work input by means of intercooling (gas cooling) between the compression stages. Vapor coming from the evaporator is compressed adiabatically by the first-stage compressor (Fig. 11). The heated vapor that results from the compression process mixes in the gas cooler with the relatively cool vapor coming from the condenser at $p_{intermediate}$. The mixture is compressed in the second stage up to the condenser pressure, p_{high}. After rejecting heat, the vapor stream (whose temperature is lower than the temperature after the first compression stage) is divided into two streams: the main one, which goes to the main expansion valve and to the evaporator, and the secondary one, which is directed to the gas cooler. Since its pressure p_{high} exceeds $p_{intermediate}$, a throttling valve is necessary to reduce it to $p_{intermediate}$ (Fig. 11).

5.3.3 Cascaded Systems

In multistage systems, the pressure ratio p_{high}/p_{low} cannot be increased above a certain limit, which is imposed by the refrigerant's triple and critical points (p_{low} must be higher than p_{triple} and p_{high} must be less than $p_{critical}$). These limits impose a pressure range that is different for each refrigerant. Moreover, an evaporator pressure below 0.1 bar is not recommended due to the bulky compressor cylinders required for the high specific volume of the refrigerant at such a low pressure and to the special safety measures that have to be considered in order to prevent air from leaking into the installation. Therefore, for a three-stage vapor–compression refrigeration system, the highest acceptable value for the overall pressure ratio is 100–110. The temperature gap corresponding to this value is approximately 100–120°C; consequently, the minimum refrigeration temperature to be achieved is approximately

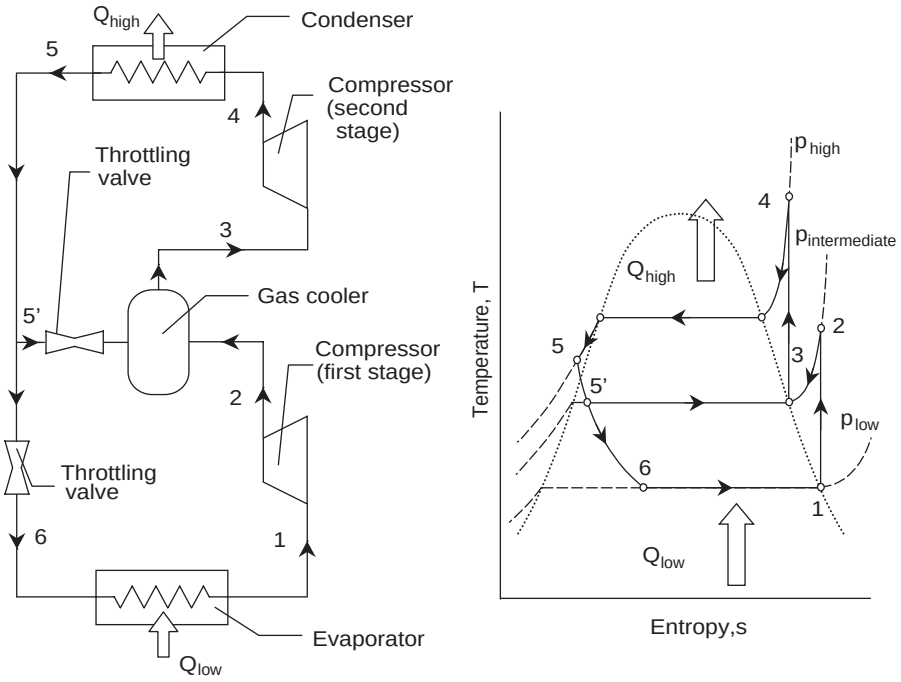


FIGURE 11 Scheme and cycle representation of a two-stage vapor-compression refrigeration system.

−70 to −80°C, considering that the condensation temperature varies from 30 to 40°C. When a single system has reached its limits, to achieve a more effective refrigeration effect, a chain of refrigeration systems using different refrigerants is necessary. Such a configuration is referred to as a cascaded system. Each of the components of a cascaded system may be single stage or multistage. With such configurations, temperatures of approximately −200°C are achievable. For instance, by cascading four systems—ammonia for the range 30 to −33°C, followed by ethylene (−28 to −104°C), methane (−99 to −160°C), and nitrogen—a temperature of −196°C can be obtained. The components in a cascade system are linked by heat exchangers, which operate as an evaporator for the "upper" component and as a condenser for the following system in the sequence.

6. ABSORPTION REFRIGERATION SYSTEMS

The idea of using an absorption fluid as a refrigerant carrier derived from the drawback of VCR systems that the gas compression requires a high work input. A pump that requires practically no work to increase the pressure in the refrigeration system replaces the complicated and work-consuming compressor.

There are two major advantages of absorption refrigeration systems (ARSs) compared with VCRs: No CFCs or HCFCs are used as refrigerants, and they use heat from different sources, such as combustion, industrial processes, waste heat (an economical solution for recovery), or solar heat.

6.1 Principles of Absorption Refrigeration

The cost of the energy supplied to the compressor represents the cost of operating a VCR system. This is why a great deal of effort has been dedicated to designing work-reducing techniques. Whatever the work-saving methods involved, as long as a gas must be compressed, a large amount of input work is required. On the other hand, increasing a liquid's pressure involves practically no work since liquids are almost incompressible. The idea of compressing a liquid in order to save work led to the development of absorption cycles. The price to be paid results from the fact that in order to accomplish the cycle, the refrigerant has to be absorbed by the liquid before compression and must be released after. The absorption process at constant pressure involves a heat exchange: When the liquid (absorbent) absorbs the refrigerant, it releases heat (absorption heat) that has to be removed because the increase in temperature results in a decrease in the solubility of the refrigerant. When the solution containing the absorbed refrigerant must liberate it, heat input from an exterior source is necessary. Therefore, heat represents the primary energy used by ARSs.

The solution poor in refrigerant is called weak solution, whereas the solution rich in refrigerant is referred to as strong solution. An ARS is composed of a generator, in which heat is added to the strong solution to release refrigerant vapor; a condenser, throttling valve, and evaporator, which perform the same functions as for VCR systems; a pump that increases the pressure of the solution; and an absorber, in which the weak solution coming from the generator absorbs the refrigerant vapor by rejecting the absorption heat to a cooling medium (usually water). Fig. 12A shows the scheme of an ARS. There are two paths:

The refrigerant path: The high-pressure vapor leaves the generator (state 1), condenses in the condenser by rejecting heat to the surroundings, and leaves it in state 2. Its pressure is reduced in the first throttling valve (state 3) and it evaporates at low pressure in the evaporator by removing heat from the cooled space. State 4 corresponds to saturated low-pressure vapor. The vapor is absorbed in the absorber by the weak solution.

The solution path: The pressure of the strong solution (state a) is increased by the pump (state b). The absorption occurs in the absorber and is facilitated by the heat removal performed by the water that flows through the cooling coil. The strong solution enters the generator, where heat is added in order to perform the separation process of the vapor from the solution. The high-pressure vapor leaves the generator, and the weak solution that results from the separation process (state c) is throttled in the second throttling valve to the absorber pressure (state d).

To save energy, a regenerative heat exchanger (also called economizer) may be placed on the solution path (Fig. 12B). The weak solution that leaves the generator still contains enough thermal energy to preheat the weak solution entering the generator. Moreover, the process of solution enrichment occurring in the absorber generates heat that must be removed. Precooling of the weak solution can improve this process. The strong solution

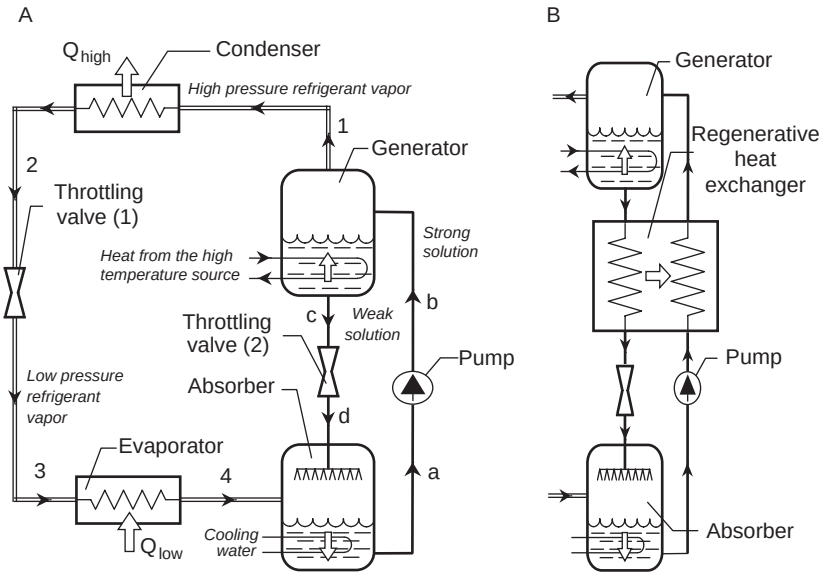


FIGURE 12 Scheme of an absorption refrigeration system: (A) without heat regeneration; (B) with heat regeneration (only the solution path is schematized).

preheating before the generator saves energy by "returning" heat to the separation process that occurs in the generator.

The COP of the ARS is

$$\mathrm{COP} = \frac{Q_{\mathrm{low}}}{Q_{\mathrm{high}} + W_{\mathrm{p}}},$$

where Q_{low} is the heat removed from the low-temperature body, Q_{high} is the heat rejected to the surroundings (the high-temperature body), and W_{p} is the work input of the solution pump.

Supposing that the heat transfer processes are isothermal and reversible and that the pump work input is negligible, the previous expression of the COP becomes

$$\mathrm{COP} = \frac{T_{\mathrm{high}} - T_{\mathrm{A}}}{T_{\mathrm{high}}} \times \frac{T_{\mathrm{low}}}{T_{\mathrm{A}} - T_{\mathrm{low}}},$$

where T_{low} and T_{high} are the temperatures of the cooled space, respectively, of the surroundings, and T_{A} is the temperature of the heat removal in the absorber. Note that this equation represents the upper limit of the COP (under ideal conditions). Due to the irreversibilities, the actual COP is less than this maximum value. In terms of actual COP, ARSs are generally poor compared with VCRs. For instance, the COP of a lithium bromide–water system is approximately 0.6, which can be improved by different techniques but with increased capital costs.

Most ARSs use the pairs water–ammonia and lithium bromide (LiBr)–water. They can be coupled with vapor–compression refrigeration systems and are used especially in the chemical industry and in air-conditioning. Their main advantage is the simultaneous refrigeration and warm water supply. The most common pairs used in ARSs are presented in Table III.

TABLE III

Absorbent–Refrigerant Pairs Used in Absorption Refrigeration Systems

Absorption Medium	Refrigerant
Water	Ammonia
Water	Methyl-amine
Lithium chloride or bromide	Water
Paraffin oil	Toluene or Pentane
Sulfuric acid	Water
Dimethyl ether or tetraethylene glycol	Methylene chloride
Sodium or potassium hydrate	Water

6.2 Water–Ammonia System

This system is used especially for small refrigerators with gas or electric heating. There are three major differences with respect to the LiBr–water systems:

1. The water–ammonia systems operate at high pressures (4–5 bar), whereas the LiBr–water machines operate at 0.04–0.08 bar.

2. Whereas LiBr is not volatile, water releases vapor during the distillation process, and therefore a rectifier is required before the condenser to separate water vapor from ammonia vapor.

3. The cooling medium for the water–ammonia system is air, whereas for the LiBr–water machines water is used as coolant.

6.2.1 Platten–Munters (Servel) System

This is a water–ammonia ARS used especially for household refrigerators and has no moving parts (pumps). The main feature of this system is the presence of an inert gas (hydrogen) that has two main functions: (i) It plays the role of an equalizing medium for the pressures in the system and thus it eliminates the necessity of a pump between the absorber and the generator, and (ii) it acts as a carrier for the refrigerant.

The pressure is the same in the entire system, namely the condenser pressure. Ammonia has different partial pressures in different regions of the machine. The ammonia vapor that leaves the generator and the rectifier chases the hydrogen from the condenser. After the condenser, liquid ammonia directly enters the evaporator, without any throttling. Here, ammonia evaporates and its vapor diffuses among the hydrogen molecules. The evaporation occurs due to the difference between the partial pressure of the refrigerant vapor in the close vicinity of the free surface of the liquid and the partial pressure of the refrigerant vapor in the hydrogen–ammonia mixture. This mixture is heavier than pure hydrogen. It moves downward and leaves the evaporator entering the absorber, where the weak solution absorbs the ammonia vapor from the mixture, which becomes lighter and returns to the evaporator. The strong solution leaving the absorber passes through a coil, in which a fraction of it releases vapor that "pumps" the solution to the generator by the thermosyphon effect. The heating of the generator is performed either electrically or with combustion gases. The separation process occurs in the generator and the weak solution returns to the absorber by gravity. The ammonia vapor enters the condenser and the cycle is repeated.

6.3 Lithium Bromide–Water System

The refrigeration systems belonging to this family can be classified from the standpoint of the origin of the heat source in indirect-fired systems, which use steam or hot liquid; direct-fired systems with heat from combustion gases; and heat-recovery systems, which use waste heat. Alternative heat sources, such as solar or geothermal energy, are also used.

The absorbent (LiBr) acts as a refrigerant carrier from the low-pressure region to the high-pressure region of the machine. For water chilling, the COP of a LiBr–water ARS varies in the range 0.65–0.7. Heat input temperatures for single-stage machines are usually 95°C, whereas for two-stage ones they are approximately 150°C.

ARSs that use this pair usually serve as water chillers for air-conditioning applications. Water is chilled to approximately 5°C. The condenser temperature is 40–45°C, whereas the evaporator temperature is 4–7°C. Temperatures below the water's triple point are not reachable since water (the refrigerant) freezes.

Low-power units used in unitary air conditioners are natural gas fired. Very high-power units with a refrigeration capacity of up to 6 MW use 1.5–2 bar of steam as a heat source and are cooled with water at 26–30°C.

6.3.1 Carrier Absorption Refrigeration System

ARSs operate at two pressure levels, p_{high} and p_{low}. The pressure is p_{high} in the generator and condenser and p_{low} in the evaporator and absorber. Therefore, the components operating at the same pressure level may be grouped in the same vessel. The Carrier system uses this arrangement, as shown in Fig. 13.

The LiBr–water solution boils in the generator by removing heat from the high-temperature heat source (steam, hot water, solar heat, or waste heat). Only water vapor is released. The vapor enters the condenser, where it condenses by rejecting the latent heat of evaporation to the cooling water that flows through a coil. The resulting liquid refrigerant passes through a throttling valve, where its pressure is reduced and it turns into a liquid–vapor mixture, which enters the evaporator. Here, the mixture evaporates by removing heat from the cooled agent (chilled water). To improve the process, the liquid fraction of refrigerant is recirculated by a refrigerant pump and sprayed onto the heat transfer surface of the cooling coil. The vapor passes to the absorber, where it is put into contact with the sprayed weak solution and is absorbed. The strong solution that results is pumped and flows through the economizer, where it removes heat from the downstream flowing weak solution and enters the generator. The process is then repeated.

The coolant (water) first flows through the absorber, where it removes the absorption heat from

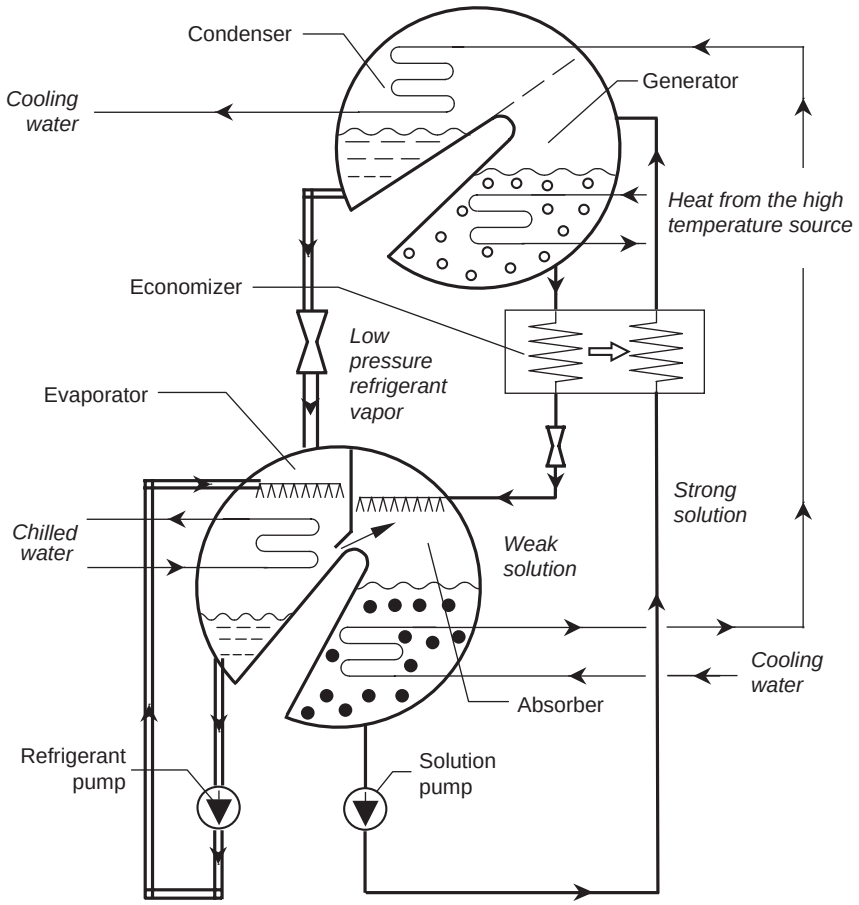


FIGURE 13 Scheme of the Carrier absorption refrigeration system.

the solution, and subsequently goes to the condenser, where it is heated by the condensing refrigerant.

In contrast to the Carrier system, which uses two vessels, in the Trane system a single vessel contains all the components of the absorption refrigeration machine. A comparison between the water–ammonia and the LiBr–water systems is provided in Table IV.

7. STEAM JET REFRIGERATION

The specific feature of the steam jet refrigeration (SJR) is the use of water as refrigerant. From the thermodynamic standpoint, water is a very good refrigerant since it has a high latent heat of evaporation and it is cheap, nontoxic, nonflammable, and available in unlimited quantities. On the other hand, water exhibits two important shortcomings: very low saturation pressures and very high specific volumes at usual temperatures. For these reasons, the use of water as refrigerant is limited to temperatures in the range 8–11°C and very rarely below 4°C. To obtain an evaporator temperature of 7°C, p_{low} must be approximately 1 kPa. Consequently, very large specific volumes (on the order of 129 m^3/kg) are involved in the compression process, and only centrifugal compressors are able to manipulate them. Another option is to use steam ejectors, which represent an economically convenient solution when steam is available and is not produced just for this purpose.

SJR is actually VCR with the mechanical compressor replaced by an ejector (or jet compressor). Although much less efficient than centrifugal compressors, jet compressors are simple (no moving parts) and inexpensive. The condenser operates at subatmospheric pressure and air leakage into the condenser is to be expected. For this reason, a secondary ejector driven by a fraction of the steam main stream removes the noncondensable gases. The high-pressure steam used to compress the low-pressure steam coming from the evaporator must also condense; therefore, the heat transfer surface area of the condenser must be large enough to

TABLE IV

Comparison between the Water–Ammonia and Lithium Bromide–Water systems

Criterion	Water–ammonia	Lithium bromide–water
Saturation temperatures difference	Relatively low	Very large
Refrigerant vapor purity	Poor	Very high
	The vapor released in the generator is a mixture of water and ammonia vapor and therefore a rectifier is required.	The vapor in the generator is pure water vapor. No rectifier is required.
Absorption rate	High	Low
	No recirculation is needed.	LiBr slowly absorbs water and therefore a recirculation of the weak solution in the absorber is necessary.
Toxicity	High	Relatively low
Corrosion	Noncorrosive toward ferrous metals	Corrodes nonferrous metals
	Ammonia corrodes copper and its alloys.	
Heat supply shutoff	Accepted	Not accepted
	The cooling of the solution due to the lack of heat input does not result in crystallization, which occurs at temperatures far below the ambient temperature.	The strong solution crystallizes at ambient temperature. In such situations, the installation has to be emptied and the solution is stored in heated reservoirs.

achieve this task. As a consequence, the size of the condenser must be three or four times the size of a VCR condenser.

The capacities of SJR systems are in the range 35–3500 kW. Typical applications of such systems are foods drying, concentration of fruit juices, and dehydration of chemicals. Direct vaporization is used in order to cool the substance; therefore, no heat exchanger or indirect cooling process are required, making SJR more suitable than VCR.

SJR systems are simple, require low capital costs and a low operating cost if using waste steam, reliable (safe) due to the lack of moving parts, and use an inexpensive refrigerant (water). On the other hand, they require a high consumption of steam and cooling water for the condenser, are difficult to operate under conditions of variable load, and have a limited area of applications compared with the other refrigeration systems. Their only worthwhile application is cooling by evaporation in drying processes.

A SJR system is presented in Fig. 14. The liquid water entering the evaporator evaporates at constant pressure by removing heat from the cooled body. The vacuum generated in the ejector by the high-pressure steam sucks the resulting vapor. The compression effect is accomplished in the jet compressor's diffuser section by the conversion of the kinetic energy of the jet into potential energy of pressure. The high-pressure vapor condenses in the condenser and the resulting liquid is divided into two streams: One fraction is pumped into the steam boiler, where it turns into steam, and the other passes through the throttling valve and evaporates in the evaporator by removing heat.

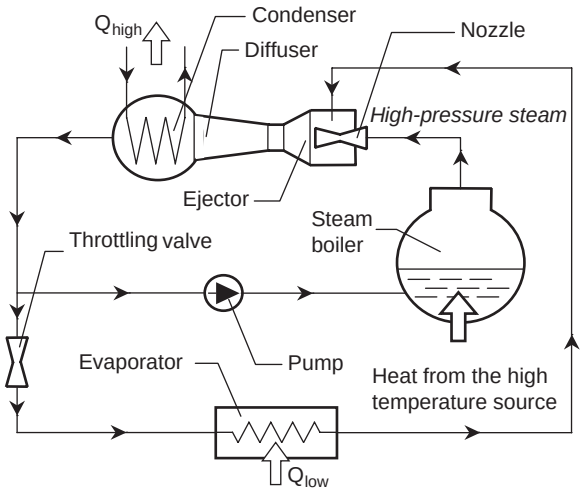


FIGURE 14 Steam jet refrigeration system.

Adjustment of the refrigeration power can be performed by varying the number of ejectors working at the same time (gross tuning) and by means of the steam flow rate per ejector (fine-tuning).

8. OTHER METHODS OF REFRIGERATION

Other refrigeration techniques are used, but only within a narrow area of applications: electronic components cooling, air liquefaction, refrigeration of small spaces, and aerospace.

8.1 Reversed Stirling Cycle (Philips Machine)

The power Stirling cycle was patented in 1816. The reversed cycle was applied in refrigeration after World War II by the Philips Company. Basically, the cycle is composed of two isothermal and two isometric (constant-volume) processes.

An interesting attribute of the Stirling cycle is that it occurs in a closed system: The working fluid is contained in the machine and there is no exchange of substance between the exterior and the refrigeration system. Another beneficial consequence is the possibility to choose the optimal fluid (from the thermodynamic standpoint) since cost is of no importance in this case due to the small quantity of refrigerant required. Moreover, because under ideal conditions all heat transfer processes between the system and the heat source and sink occur isothermally, the cycle's efficiency equals the Carnot efficiency. Obviously, under actual conditions, the Stirling's cycle efficiency is less than the ideal one due to the irreversibilities that inherently occur in every process.

The main feature of the cycle is heat regeneration, namely the internal flow of heat, which is first stored in a structure called a regenerator and is discharged later to the gas circulating through this unit. The regenerator consists of a heat-storage material (ceramic mesh, wire, or metallic gauze) through which the gas is forced back and forth during the isochoric processes. The gas rejects heat to storage medium (cooling) or removes heat from it (heating). In this manner, heat is "lent" to the storage mass of the regenerator during the isochoric gas cooling and is taken back later from it during the isochoric gas heating.

The machine (Fig. 15) is of the reciprocating type and consists of a cylinder with two pistons (A and B) and a regenerator (R). By means of a sophisticated mechanical system, the two pistons perform interrelated motions that allow the processes that comprise the cycle to occur.

The principle of the cycle operation is schematized in Fig. 15 in correspondence with the *T–s* plot of the processes. The first process is the isothermal compression (1–2), which occurs as piston B is moved to the left while the coolant removes the heat that is generated as a consequence of the compression work and thus the temperature T_{high} of the gas remains constant. During this time, piston A is held in position. The following process (2–3) is performed by moving both pistons at the same rate to the left. The volume of gas remains constant and the evolution is isochoric (isometric). The gas is forced

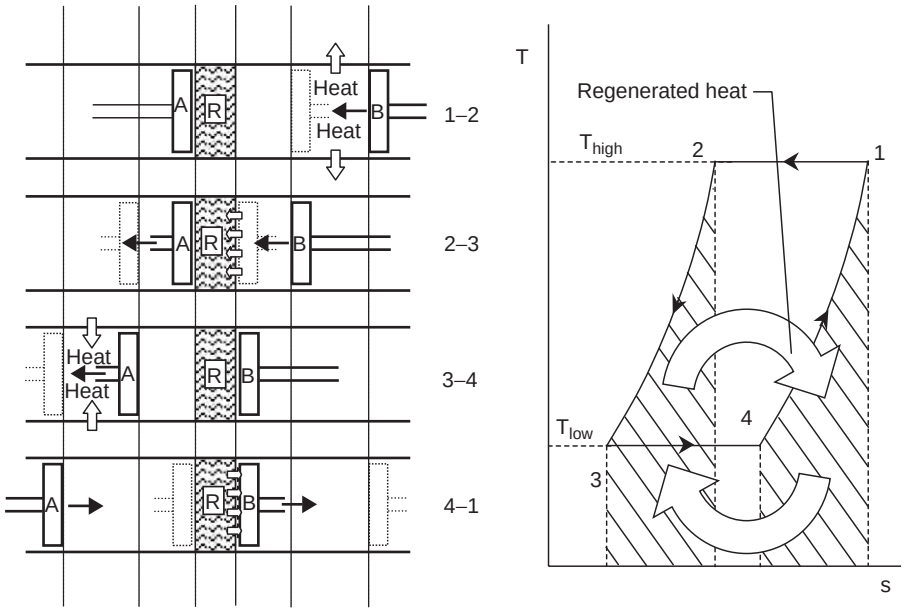


FIGURE 15 Schematic representation of the reversed Stirling cycle operation.

through the regenerator and rejects heat to the storage medium, which is colder as a result of the previous passage of gas. Isothermal expansion (3–4) follows, which is performed by blocking piston B and moving piston A to the right while heat is removed from the cooled space at T_{low} = constant. The last process of the cycle is isochoric gas heating (4–1). Both pistons move at the same rate to the right and the volume of gas is constant. By passing through the regenerator R, the cold gas removes the heat that has been stored during the previous isometric process and regains the temperature from the beginning of the cycle (T_{low}).

The Philips machine is used for temperatures in the range 223–13 K (−50 to −260°C). Most of its applications are in cryogenics. The actual machine, designed by Köhler and Jonkers, differs from the previously discussed scheme from the standpoint of the pistons' movement, which is continuous (the pistons of the ideal machine move intermittently). Helium and hydrogen are the usual working fluids. For temperatures higher than −130°C, air may also be used. The most usual applications are in the area of scientific research, gas liquefying, etc. The Philips refrigeration machine is characterized by very low losses and therefore its efficiency in the temperature range −150 to −90°C is approximately 0.4. A machine with a relatively small size (cylinder inner diameter = 70 mm and piston stroke = 50 mm) can liquefy approximately 5 kg of air per hour, with the necessary power input being approximately 6 kW.

8.2 Open-Cycle Systems

8.2.1 Gas Expansion in a Field of Centrifugal Forces (Ranque Effect)

The French physicist Ranque observed this phenomenon for the first time in 1933, which is why it is referred to as the Ranque effect. The vortex tube device of a cylindrical tube is equipped with a tangential nozzle and a cone, as in Fig. 16. As the gas expands in the nozzle, a very intense circular current develops, characterized by a strong radial velocity gradient. The compression of the gas close to the tube wall heats the gas, whereas the circulation becomes longitudinal near the tube's axis. As a result, a separation of streams occurs: The hot one leaves the tube at the right end and the cold one is collected at the exit orifice from the left end. The axial position of the cone controls the flow rate and thus the temperatures of the two streams. Turbulence and viscosity are the source of the phenomenon: The kinetic energy of the gas particles has important variations, and thus the gas is compressed and expanded adiabatically, performing a refrigeration cycle that pumps heat to the peripheral regions of the gas by cooling the regions situated on the tube axis (Fig. 16).

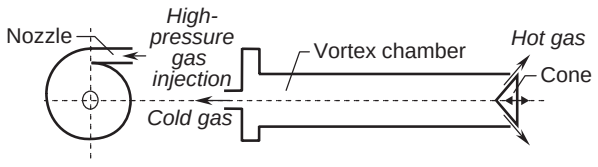


FIGURE 16 The vortex (Ranque) tube.

The parameters that influence the performance of the vortex device are the tube's geometry (length and diameter), the diameter of the cold gas escape orifice, the nozzle's section area and type, the temperature and pressure of the compressed gas, and its physical properties. For instance, for an air pressure of 11 bar at an inlet volumetric flow rate of 30.5 m^3/h, by using a tube with a certain geometry one can obtain a fraction equal to 0.2 from the total flow rate at −41°C.

8.2.2 Reversed Thermoelectric (Peltier) Effect

The direct thermoelectric effect was discovered in 1821 by the German physicist Thomas J. Seebeck. It consists of the apparition of an electromotive force as the junctions of two wires made of different electrical conductors (the ensemble is called thermocouple) are placed in mediums with different temperatures.

The reversed thermoelectric effect was discovered in 1834 by the French physicist Jean-Charles Anastase Peltier. By connecting a thermocouple to a direct current (DC) source, one junction heats and the other cools. By inverting the current's direction, the cold and hot junctions switch functions.

The apparition of a temperature difference between two junctions involving different metallic conductors when connected to a DC source is referred to as the Peltier effect. The explanation resides in the fact that different electrical conductors exhibit different energy levels of their conduction electrons. When forced to pass from the high-energy region to the low-energy region under the influence of an electric field, the electrons reject their excess energy as heat. This phenomenon occurs at the boundary between the two conductors, which are put into contact by welding. The opposite process takes place at the other junction: Low-energy electrons absorb from the surroundings the necessary energy to pass to the high-energy region, and this phenomenon manifests as heat removal from the environment.

The cooling effect is poor in conductors, whereas in semiconductors its intensity is important. The maximum cooling efficiency is reached when one of the semiconductors is of the n type and the other is of the p type. Typical p-type semiconductors used for this purpose are alloys of Sb_2Te_3 and Bi_2Te_3, and n-type semiconductors are alloys of Bi_2Te_3 and Bi_2Se_3. With such couples, temperature differences of up to 60°C have been obtained. By connecting more thermoelectric couples in series (battery arrangement: the cold end of one of them cools the hot end of the other), the temperature gap can be increased.

Thermoelectric refrigeration batteries are used in low-power as well as high-power applications. For small systems, optimization aims to get the highest power possible by sacrificing efficiency. High-power systems are designed to achieve the highest performance in terms of efficiency. Other potential applications of thermoelectric refrigeration systems are small air-conditioning systems, household refrigerators, and aerospace.

The cooling effect is reduced by heat generation at the passage of the electric current through a conducting body (the Joule–Lenz effect) and by the conduction heat transfer between the two junctions.

The advantages of thermoelectric refrigeration systems result from the fact that they require no refrigerant, have no moving parts, are simple, the refrigeration power can be as small as necessary, the refrigeration element has a very small size (applications in surgery, electronics, and dentistry), and the refrigeration temperature can be easily controlled by varying the current's intensity. On the other hand, they have poor shock resistance and are expensive.

9. CRYOGENICS

The term cryogenics derives from the Greek *kryos*, for icy cold. The upper limit that characterizes the cryogenic region of the temperature scale is subjective. If the delimitation criterion takes into consideration practical reasons, then this upper limit can be established at a higher than usually declared value (120 K) in order to include natural gas liquefaction. The cryogenics region of the temperature scale 0–120 K can be divided into subregions according to the criteria of the cryogenic method and refrigerant as follows:

- 63–120 K: The industrial cryogenics domain. The cryogenic methods are based on thermodynamic state-change processes of the usual refrigeration type, as described previously.
- 1–60 K: Thermodynamic methods combined with additional vacuum processes that characterize the domain of liquefaction of hydrogen, helium, neon, nitrogen; also of industrial importance.
- 0.3–1 K: The domain of laboratory cryogenics.
- <1 K: The cryogenics domain of advanced scientific research. The methods used are of the magnetic type. Temperatures on the order of 10^{-8} K have been obtained. The third law of thermodynamics (Nernst, 1906) states that absolute zero cannot be reached whatever method used.

Cryogenics is most important in industrial applications: liquefaction, gas separation, and superconductivity. Cryogenics is also on the cutting edge of scientific research: nuclear and elementary particles physics, physics of the solid, cryobiology, and cryosurgery.

9.1 Cryogenic Cycles and Methods

Some of the previously described refrigeration techniques are also applicable to the cryogenic temperatures' domain. In gas-expansion processes, the gas, after being compressed and rejecting heat, expands in an expansion machine. Compared with throttling, this method has a higher cooling intensity, expressed as the temperature drop per pressure drop (degrees/bar). In the cryogenic cycles without the refrigerant's phase change, the latter remains gaseous as it undergoes the processes that comprise the cycle. Usually, these cycles correspond to low- and medium-power refrigeration systems (also known as gas refrigeration machines) and are used in laboratory applications. The most used is the Philips machine, whose cycle (the reversed Stirling cycle) was previously described. By using helium as refrigerant, such machines can reach 77 K in a single stage and 12 K in two stages.

In the following sections, specific cryogenic methods are presented. Some of the most used cryogenic agents are shown in Table V along with some important physical properties.

9.1.1 Joule–Thomson Effect

By being throttled, the gas undergoes a pressure reduction and under certain conditions its temperature decreases. The phenomenon is called the Joule–Thomson effect and is accompanied by a temperature change of the gas. If the process occurs within a temperature and pressure range that ensures that the

TABLE V

Main Cryogenic Agents and Important Physical Properties

Cryogen	Normal boiling point (K)	Critical temperature (K)	Critical pressure (kPa)	Triple point temperature (K)	Triple point pressure (kPa)
Helium	4.22	5.28	227		
Hydrogen	20.39	33.28	1296	14.0	7.2
Nitrogen	77.33	126.17	3385	63.22	12.55
Carbon monoxide	82.11	132.9	3502	68.11	15.38
Argon	87.28	151.2	4861	83.78	
Oxygen	90.22	154.8	5081	54.39	0.14
Methane	111.72	190.61	4619	90.67	11.65
Krypton	119.83	209.4	5488	116.0	73.22

effect is positive (the gas temperature decreases as its pressure is reduced), then throttling results in gas cooling and can be successfully used to liquefy gases. The following requirements must be met:

1. The gases involved must have very low saturation temperatures and pressures.
2. In order to bring the gas into a state characterized by a positive Joule–Thomson effect, a foregoing preparation is necessary: the gas has to be precooled in auxiliary classical refrigeration systems.
3. To obtain very low temperatures, very high pressure drops are necessary.

For instance, for hydrogen, the pressure value before throttling must equal 110 bars after a precooling below 190 K (approximately −83°C) to obtain a positive Joule–Thomson effect.

9.1.2 Linde Method

Basically, the configuration of the cryogenic system is that of a classical VCR machine. The specific features of the cryogenic system are the aftercooler that replaces the condenser and the counterflow heat exchanger placed between the aftercooler and the throttling valve. In the aftercooler, the gas is brought to the ambient temperature. The function of the counterflow heat exchanger is to perform a supplementary cooling of the high-pressure gas that leaves the aftercooler. The coolant is cold gas collected from the separation vessel placed after the throttling valve. A low-quality liquid vapor mixture results after the throttling process. The mixture enters a separation vessel, where the liquid phase is gravitationally collected at the bottom and withdrawn, whereas the vapor occupies the space above the liquid. The vapor (which is at the saturation temperature) is circulated through the counterflow heat exchanger and removes heat from the stream of high-pressure gas coming from the aftercooler. This supplementary cooling significantly increases the liquefaction efficiency.

Adding a second throttling and a precooling of the compressed gas after the compressor can improve the basic configuration described previously. Such improvements reduce the necessary power input of the cryogenic system.

9.1.3 Claude Method

This technique uses an expander and is an improved variation of the Linde method. By expanding a gas in a machine, the work output combined with the pressure decrease leads to a stronger cooling than that of the simple pressure destruction by throttling. Consequently, after being cooled in the aftercooler, the gas undergoes a primary cooling in a counterflow heat exchanger by rejecting heat to a stream of cold gas coming from the second-stage counterflow heat exchanger. After this primary cooling, the gas stream is divided: A fraction enters the second-stage heat exchanger, whereas the remainder expands in the expander, resulting in a colder gas, which is returned to the main coolant flow path that enters the second-stage heat exchanger. The vapor collected at the upper end of the separation vessel is directed to the final counterflow heat exchanger, where it cools the gas stream that leaves the second-stage heat exchanger. By cascading several Linde and Claude cryogenic systems operating with different refrigerants, one can obtain liquefaction of helium and an additional cooling in helium down to 0.7 K.

9.1.4 Diffusion Pump Method

This device is used to liquefy helium and is based on the active coal's property of adsorbing or desorbing

helium. Active coal has a very porous structure, which allows it to adsorb gases in large amounts. The quantity of gas adsorbed increases as temperature decreases and as pressure increases. As a consequence, a gas is adsorbed at low temperature. By heating the adsorbent, the gas is released. This property allows the adsorbers to operate in adsorption–desorption cycles. The adsorption process is exothermic (i.e., it releases heat), whereas the desorption is endothermic, requiring heat in order to occur. These phenomena are used to liquefy helium in cryogenic systems in which the compressor has been replaced by a diffusion pump.

The active coal is saturated with helium at normal or higher pressure and the heat of adsorption is evacuated by cooling. The system (coal + adsorbed helium) is then thermally insulated and undergoes a process of desorption by creating vacuum in the space that confines it. As a result, the temperature decreases, and the refrigeration effect occurs. Using such a technique, a temperature of 2 K has been obtained in liquid helium.

9.1.5 Thermomagnetic Effect

A substance that is not a magnet in the absence of a magnetic field is called paramagnetic. Only few substances are paramagnetic: the crystalline hydrates of some rare minerals (e.g., $Gd_2SO_4 \cdot 8H_2O$) or the salts of the metals from the platinum or aluminum groups.

When subjected to an exterior magnetic field, the microscopic dipole magnetic moments of the paramagnetic material oppose this action and consequently the magnetic field performs work to change their orientation. At the macroscopic level, this phenomenon manifests as heating of the substance. If one removes this heat with a coolant, the magnetization process will be isothermal. Switching off the magnetic field, the substance undergoes a demagnetization process and the energy stored in the substance as a consequence of its opposition to the action of the field is transferred to the magnetic dipoles to regain their chaotic orientation. If the substance is thermally insulated, the result is an intense cooling. The phenomenon is called the thermomagnetic effect and the method is referred to as adiabatic demagnetization of paramagnetic substances.

A second phenomenon occurs during the magnetization–demagnetization cycle of the paramagnetic body. The internal energy of a substance depends not only on the kinetic and potential energies of the molecules but also on the magnetization. As magnetization increases, more order is transferred to the system by orienting the elemental magnets along the magnetic field lines and thus the chaotic movements of the molecules are limited. The number of degrees of liberty of the thermal motion is correspondingly reduced. By this action, an internal magnetic field develops, called the internal Lorentz field. The internal energy of the body decreases as a consequence of magnetization. During the adiabatic demagnetization phase, not only does the magnetization work have to be redistributed to the molecules to regain their initial chaotic orientation but also the amount of internal energy corresponding to the degrees of liberty lost during the magnetization process has to be redistributed. An additional temperature decrease occurs due to this internal energy redistribution process.

By using the adiabatic demagnetization of paramagnetic salts, temperatures in the range 0.3–10^{-3} K can be obtained and maintained for a short time.

To obtain temperatures less than 10^{-3} K, the adiabatic demagnetization of the nuclear state of a paramagnetic salt has been used. The phenomenon is basically the same as in the case of the atomic paramagnetism. Using adiabatic demagnetization of the nuclear spin system of copper, a temperature of 10^{-8} K was attained in the 1980s.

9.2 Applications

9.2.1 Air Separation

Air components such as oxygen, nitrogen, and argon are used in different applications. Nitrogen is used in metallurgy (as inert atmosphere), electronics, chemical processes, and deep refrigeration.

Oxygen is used in metallurgy, welding, chemical industry, and medical applications. Argon is used as inert atmosphere in electric bulbs and welding.

Dry air is a mixture of three major gases (in moles, 78.03% nitrogen, 20.99% oxygen, and 0.934% argon) and other components, such as carbon dioxide, rare gases, hydrogen, and helium (all less than 0.05%). Air liquefaction constituted the first application of the Linde method in 1895. Before being processed, air is dried up and carbon dioxide is removed since water vapor and CO_2 freeze during air liquefaction, endangering the operation of the air-separation unit. The separation of the three main components is based on the distillation process. A two-staged distillation column operating at two pressure levels separates oxygen and nitrogen, whereas argon is distilled in a separated column. A typical air-separation plant supplies liquid and

gaseous oxygen, which are collected at the lower end of the low-pressure stage of the column, and liquid and gaseous nitrogen, which are extracted from the medium- and high-pressure regions of the second-stage column. A turbine expander supplies the refrigeration effect necessary for the air separation. Less important refrigeration is obtained by means of the Joule–Thomson effect.

Total air-separation plants provide oxygen, nitrogen, argon, and rare gases (He, Ne, Kr, and Xe).

9.2.2 Liquefaction of Natural Gas

Two reasons justify the liquefaction of natural gas. First, the demand is not constant, exhibiting a peak in midwinter. To meet this increased demand, the distribution piping should be designed accordingly, which involves high capital costs. It is more economical to design the distribution grid according to a medium demand, to store liquefied gas in large storage tanks, and to vaporize and add it to the network when necessary. Second, the oil/gas field is usually far away from the users or in a location that does not economically or technically justify a pipeline for collecting the extracted gas (e.g., offshore extraction platforms). In such cases, gas is liquefied and transported to a facility near the user area, where it is vaporized and injected into the distribution network.

The refrigeration effect necessary for liquefaction is achieved by a variety of methods. In some cases, methane is used as a refrigerant in intermediate refrigeration.

9.2.3 Helium Liquefaction

Helium was the last of the gases to be liquefied in 1908 by Kammerlingh Onnes. Helium is used in welding, aerospace, superconductivity, inert atmosphere applications, medicine, and advanced scientific research. Helium's concentration in air is well below the economically acceptable level. However, it is present in much higher concentrations (up to 15%) in natural gases. Helium's liquefaction requires special cryogenic techniques since its saturation temperature is the lowest among all substances (Table V). Due to the fact that at room temperature helium has a negative Joule–Thomson effect (it heats when throttled), previous precooling below the inversion temperature (the temperature above which the Joule–Thomson effect changes slope) is necessary. This is performed by means of a Claude process, which brings the gas to 20 K at 3000kPa. Liquefaction is then accomplished by means of an expansion valve.

9.2.4 Superconductivity

Helium has two stable isotopes, He_I (atomic mass = 3) and He_{II} (atomic mass = 4), which are present in different concentrations. Its normal boiling point is 4.23 K, and below this temperature it is a colorless lightweight liquid with a very low viscosity. The separation of the two isotopes can be done at temperatures lower than 2.17 K, which is the threshold below which He_I is in a normal state, whereas He_{II} is in the superfluid state. Helium is the only substance that can maintain its liquid state at near absolute zero and has no triple point. Solid helium can be obtained only under high-pressure conditions (at $T < 1.0$ K, the pressure must be higher than 2500 kPa to obtain solid He_{II} or higher than 3000 kPa for He_I). The higher the temperature, the higher the pressure: For instance, at $T \approx 60$K, the solidification pressure is approximately 10^6 kPa.

Superfluidity is a special property of liquid He_{II} at $T < 2.17$ K. The liquid's unusual behavior consists of the fact that its flow velocity through capillary tubes depends only on the temperature and the tube diameter, and not on the pressure drop and the tube length. The flow occurs without friction, with the viscosity of the liquid becoming extremely low. The phenomenon was discovered in 1932 by the Russian physicist P. L. Kapitsa, who called it superfluidity. At 5.072 kPa and $T = 2.174$ K, helium's thermal conductivity exhibits a radical change: It suddenly becomes 3×10^6 times higher than the thermal conductivity of other liquids at the same temperature (liquid helium included) or 1000 higher than copper's conductivity at ambient temperature.

In normal electrical conductors, the resistance decreases as temperature decreases. In 1911, by studying the electrical behavior of a mercury sample in liquid helium, the Dutch physicist Kammerlingh Onnes discovered that the sample's electrical resistance disappeared suddenly. It was the first time that the phenomenon called superconductivity was noticed. Further research on this phenomenon was performed and it was found that it occurs in a wide range of metals and alloys. Practically, the resistance equals zero. This fact has been verified by immersing a metal ring in liquid helium and injecting an electric current. The current can run through the ring undiminished for months.

Superconductivity is used mainly in magnet applications (plasma physics and electrical machines). An interesting application may be the use of superconductive magnets for magnetic levitation in high-speed railroads (500 km/h).

Other potential applications are in the field of electric energy transport. Power plants are situated as close as possible to the fuel source, which is generally far away from users. Consequently, the energy has to be transported through metallic cables. Due to the losses caused by the metal's electrical resistance, economic limitations dictate that the transported power at 1.5×10^6 V be less than 10^6 kW. The use of superconductive cables may be economically acceptable for energy transport at distances on the order of 200 km for AC and 50–800 km for DC. In the case of very high-power superconductive energy transport, the economy may be approximately 5 or 6 MWh/km.

For obvious economic reasons, the research in the field focuses on different alloys and semiconductor combinations in order to develop substances that reach the superconductive state at the highest possible temperatures, preferably 100 K (liquid nitrogen).

10. AIR-CONDITIONING

The following are the most important developments in the evolution of air-conditioning and air-conditioning systems:

At the end of the 19th century:

- Founding of the scientific hygiene (Max von Pettenkofer, Germany)
- Research on air exchange, hygrometry, and air moisture
- The use of electric motors for air fans in large buildings; air filtration
- By 1890, air humidification with steam and later by water spraying; beginnings of air-conditioning.

20th century:

- By 1900, the first air-conditioning systems with preheater, humidifier, and reheater (United States); W. H. Carrier (1876–1950), father of air-conditioning
- The first temperature and humidity control systems (pneumatic and electrical)
- By 1920, the upsurge of air-conditioning; central air-conditioning systems for comfort (theaters, office buildings, and conference halls) and industry, especially for hygroscopic materials processing (tobacco, paper, and textiles)
- By 1930, the first unitary systems that include all the components; the use of freons for the refrigeration side; room air conditioners

TABLE VI

The Standard Dry Air Composition

Component	Volume fraction
Oxygen	0.2095
Nitrogen	0.7809
Argon	0.0093
Carbon dioxide	0.0003

After 1945, very rapid development of air-conditioning systems: dual-path systems, air-and-water systems with induction devices, and fan–coil units.

10.1 Basic Principles

The standard dry air composition as established in 1949 by the International Joint Committee on Psychrometric Data is presented in Table VI. The water vapor content of the air may vary from zero (dry air) to a maximum (saturated air) that depends on the temperature at a certain atmospheric pressure.

The humidity ratio W of the air is the ratio of the mass of water vapor (m_v) to the mass of dry air (m_a):

$$W = \frac{m_v}{m_a}\left[\frac{\text{kg water vapor}}{\text{kg dry air}}\right].$$

The relative humidity ϕ is the ratio of the mass of water vapor (m_v) contained in the air to the mass of water vapor at saturation (m_s) at the same temperature:

$$\phi = \frac{m_v}{m_s}.$$

The relative humidity varies in the range 0–1 (0, dry air; 1, saturated air).

The dew point temperature is the temperature below which water vapor condensation occurs if the air is cooled at constant humidity ratio. The dew point temperature can also be the air saturation temperature at the same pressure and humidity ratio as the considered mixture.

Due to the moisture present in the air, the heating and cooling processes undergone by the air are accompanied by changes in its relative humidity and/or humidity ratio. Consequently, the design of air processing to meet certain requirements regarding its temperature and moisture must take into consideration the heat and mass transfer associated with these state changes. Moreover, in certain cases (such as cooling below the dew point temperature at constant humidity ratio), water vapor condensation will result and liquid water will leave the air.

10.1.1 Human Comfort

The main parameters that define the air quality in a space are temperature, humidity, air purity (involving dust, microorganisms, and odors), thermal radiation, and air motion.

The factors that influence the subjective sensation of comfort in a space are of extreme variety and depend on sex, age, the nature of the human activity in that space, etc. A 100% satisfaction regarding comfort is an unrealistic objective due to the previously mentioned considerations. This is why the American Society of Heating, Refrigerating and Air Conditioning Engineers Comfort Standard 55 defines an environment as thermally acceptable if at least 80% of the occupants consider it acceptable from the thermal standpoint.

The normal temperature of the human body is approximately 36.9°C and is maintained at this constant value by physiological mechanisms. If the temperature and humidity of the environment allow these mechanisms to accomplish their duty, then the subjective perception of comfort is obtained.

There are basically two ways by which the human body maintains its temperature at a quasi-constant level:

1. Metabolism, which increases or decreases the rate of internal heat production as the body temperature decreases or increases, and which depends on the level of effort that characterizes the human activity.
2. Control of the heat transfer between the body and the environment: If the external temperature is low, the body's heat losses are reduced by reducing the intensity of blood circulation (vascular constriction at the skin level). If the exterior temperature is too high, blood vessels that irrigate the skin dilate, allowing a higher blood flow and consequently a higher rate of heat rejection. This cooling mechanism is combined with the heat rejection by means of the latent heat of evaporation of the sweat.

The heat release of the human body can vary from 75 (sleep) to 700 W (sustained hard work), in accordance with the degree of effort of the activity. The corresponding limits of the humidity release by perspiration are 0.05 and 0.45 kg/h, respectively.

The relative humidity of the air is not directly perceived as such, but it contributes to the general sensation of comfort by encouraging or hampering sweating. This is why a higher temperature is endurable when air is dry, whereas moderate temperatures seem intolerable if the relative humidity approaches 100%.

In all cases, the air motion is of great importance since it intensifies the convection heat transfer of the body. An air velocity of approximately 0.25 m/s is acceptable from the standpoint of thermal comfort. Another factor that determines the thermal comfort is thermal radiation, which becomes important if there are surfaces in the environment whose temperatures are different from the air temperature.

These factors act in strong correlation with each other, which is why their influence must always be considered in an overall manner. For example, Figs. 17 and 18 present the combined action of air temperature and relative humidity and air temperature and velocity, respectively, in defining the comfort domain.

The air-conditioning systems are classified according to the following criteria:

Application: Comfort and technological
Air pressure: High and low (classical)

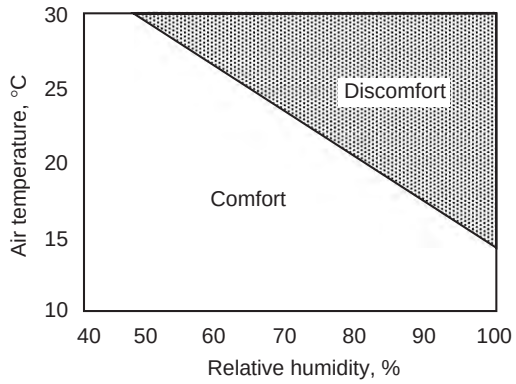


FIGURE 17 The comfort and discomfort domains resulting from the combined action of the air temperature and relative humidity.

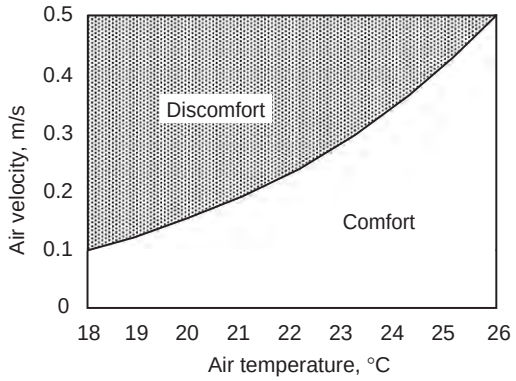


FIGURE 18 The combined action of air velocity and temperature in defining the comfort and discomfort domains.

Flow rate: Constant and variable
Heat transfer agent: All air, air and water, or all water.

The design of an air-conditioning system must consider many variables that determine the air parameters, such as solar radiation, air infiltration, heat gains due to conduction and convection from boundaries and internal objects, and latent heat and humidity gains caused by water-containing objects. All these heat and moisture sources determine the cooling load, which must be equal to the heat rate that has to be removed from the space by the air-conditioning equipment. This requirement is rarely met since the cooling load is almost permanently variable as a consequence of the ever-changing conditions (season, solar radiation, ambient temperature, wind, occupancy of the space, daily and weekly work cycle, variable operation of the heat and moisture-generating equipment, etc.). Therefore, it is recommended that the air-conditioning system should be designed for a peak cooling load and it will be operated intermittently at loads less than the peak (design) value.

The indoor design conditions are 25°C and 50% relative humidity. The necessity of a good match with the cooling load has determined the design of packed refrigeration systems of different refrigeration powers, which are delivered ready to be integrated in the air-conditioning systems:

- VCR systems equipped with reciprocating compressors: power range, 11.6–464 kW
- VCR systems with centrifugal compressors: power range, 348–2320 kW
- Lithium bromide–water absorption refrigeration systems: power range, 116–5800 kW
- Steam jet refrigeration systems: power range, 10–6000 kW.

10.2 Temperature and Humidity Requirements for Different Activities

The function of air-conditioning systems is to control the temperature and humidity of a space in order to either meet the comfort requirements of the personnel in that space or create the microclimate in an industrial space. The thermal comfort conditions for a space impose a temperature within the range 20–25°C, and the relative humidity must be in the range 35–65%. Many industrial processes involve technological prescriptions that impose special temperature and/or humidity values for the air, in some cases in very narrow ranges. Table VII illustrates these requirements for some of the most common domains.

TABLE VII

Temperature and Relative Humidity Requirements for Some Activities

Domain	Activity	Temperature (°C)	Relative humidity (%)
Libraries	Book storage	18–25	40–50
Museums	Painting galleries	18–24	40–55
Beverages	Brewing	10–15	80–85
Rubber	Vulcanization	26–28	25–30
	Surgical material	24–33	25–30
Chocolate	Packaging	18	55
Mechanical	Precision assembly	20–24	40–50
Pharmaceuticals	Manufacturing	21–27	10–50
Tobacco	Tobacco storage	32	85–88
	Cigarette manufacturing	21–24	55–65
	Packaging	23	65
Textiles	Cotton spinning	24–27	50–65
	Wool spinning	24–27	60–80

10.3 Basic Central Air-Conditioning

This is a single-zone air-conditioning system and supplies a constant or variable air volume at low, medium, or high pressure. Usually, the system is located outside the air-conditioned area.

The typical components of air-conditioning equipment are air filters; a refrigeration unit; cooling coils; heating coils using steam, hot water, or electricity; an air humidification unit; air ducts; water pipes; fans; and control equipment.

Usually, central systems supply to an entire building both warm air for heating during the cold season and cold air for cooling in the warm season. The outdoor air undergoes a series of heating and cooling processes combined with humidification and/or drying, with a resulting stream of processed air that meets the comfort requirements. Only the air-conditioning subsystems that comprise the general system are discussed further.

The air-handling components of such systems can be arranged in various configurations according to the specific requirements of the space to be air-conditioned. The air-conditioning system must be easy to maintain and must accomplish its task in the simplest configuration possible. The central systems

may be classified as all-air systems, air-and-water systems, and all-water systems.

10.3.1 All-Air Systems

These systems supply only air to the conditioned space and process it in order to offset the maximum cooling load from the building. For spaces in the building whose cooling load is less than the maximum one, a reheat of the air can be provided by adequate heating equipment.

A typical all-air system is shown in Fig. 19. All-air systems are of two types: single-path and dual-path. In single-path systems, a single stream of air flows through the heating and cooling coils that are arranged in series. In dual-path systems, the heating and cooling coils are placed on separate paths and the processed air stream results by blending the heated and cooled steams.

All-air systems are used in buildings that have several spaces that each require a separate control of the air's temperature and humidity, such as hospitals, office buildings, hotels, and stores. All-air systems are used wherever a close temperature and humidity control is required (e.g., hospital operating rooms, computer rooms, and textile and tobacco industries).

10.3.2 Air-and-Water Systems

These air-conditioning systems were developed and implemented after World War II due to the need to reduce the space required for the air ducts in buildings. Their typical use is in high-rise structures. A central unit processes the minimum amount of fresh air at 14–16°C, called primary air, and distributes it under a relatively high pressure to the rooms and spaces to be air-conditioned. Due to the higher pressure, the air velocity through the ducts is high, and therefore the flow section area of the ducts is small, as is the space required. As a result, higher power fans are required. Water (called secondary water) is also processed in the central unit. Primary air and secondary water are distributed to terminal units in the conditioned spaces throughout the building. The room terminal may be an induction unit or a fan–coil unit. The primary air supplied by the central air-conditioning system is mixed in the room terminal units with the return (secondary) air and distributed in the rooms. The advantage of this system is that the temperature can be adjusted in each room. However, this configuration complicates the piping because each room or space must have its own heating and cooling pipes (water is the heat transfer agent), and consequently the costs are higher.

The induction unit uses the ejection effect produced by the high-pressure primary air that flows through nozzles and sucks the interior (return) air, which is forced through heat exchangers that cool or heat it as required. The induction units are usually installed under a window or in an overhead location. They have some advantages, such as individual room temperature adjustment, less space required, easy maintenance (no moving parts), long life span, and smaller size of the central unit due to the small volume of primary air to be processed. On the other

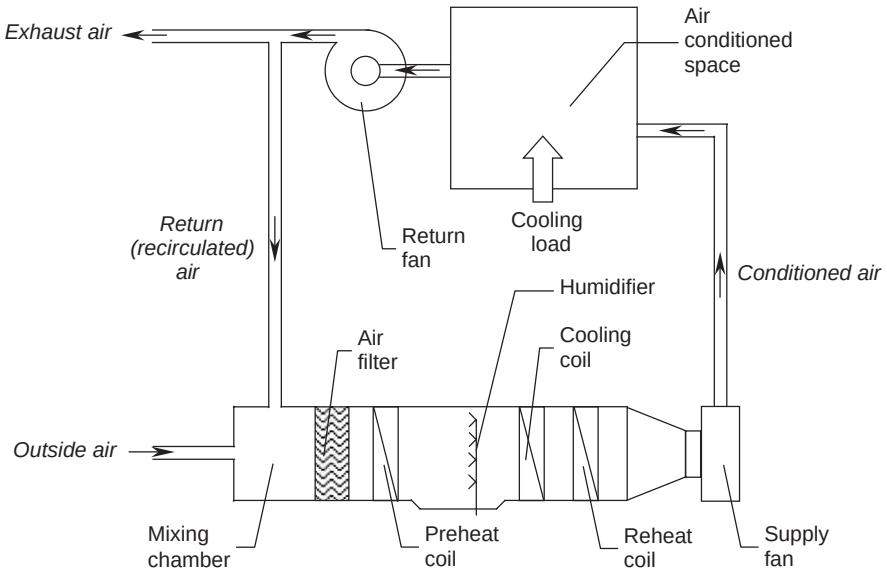


FIGURE 19 Scheme of an all-air central air-conditioning system.

hand, fouling of the secondary air side of the cooling or heating coils is possible and high energy consumption is necessary.

The typical components of a fan–coil unit are a filter, finned-tube coils, and a fan. Hot or cold water circulates through the coils. The main drawback is the noise produced by the fan. Some configurations contain an additional electric heater, which is used during the season transition to avoid changeover problems.

The air-and-water systems are used when no accurate control of the humidity is required (schools, office buildings, hotels, hospitals, etc.).

In all-water systems, the only cooling agent is water, which flows through finned coils, and individual fans blow air across the coils. These systems are recommended for their flexibility, but only air cooling and dehumidification are possible, unless a central humidifier is provided. The typical room terminal unit in this case is the fan–coil unit.

10.4 Unitary Air Conditioners

Unitary systems (also called multiple-packaged unit systems) are intended to match the air quality criteria for separate (individual) spaces. Their design allows a large variety of configurations and placements. The arrangement may be single or split, placement may be wall mounted or suspended from ceilings, cooling may be done with an evaporative condenser, they may be water or air cooled, etc. The main advantage of the unitary systems is flexibility. They operate more economically than central systems in large spaces, in which the placement of the units can cover the local cooling loads in a very effective manner.

Unitary systems (cooling capacity, 2.5–400 kW) are suitable when low initial cost, lower performance, and simple installation are required. Typical applications are office buildings, motels, and stores.

They are manufactured in different configurations in order to meet the multitude of possible requirements and applications. Each system consists of filters, heating and cooling coils, fans, one or more refrigeration compressors, and control equipment. Sometimes, multipackaged units for perimeter rooms are used in combination with central air-conditioning to increase flexibility.

Outdoor unitary systems are intended to cool or heat an entire building and consist of unitary systems, air ducts, and temperature and humidity control equipment. Usually, they are mounted on the roof, but in certain situations they can be mounted on the ground level.

Indoor unitary air-conditioning systems are located in the service areas of the building, close to the conditioned space, and are composed of an indoor unit with a water- or an air-cooled condenser, air ducts, and temperature and humidity control equipment.

The split configuration consists of two units connected by the refrigerant pipes:

1. The exterior unit (usually mounted in a console on the external face of the wall, on the ground, or on the roof), which contains the compressor, a fan, and the condenser
2. The interior unit, which consists of the evaporator, the reheat coil, the humidifier, and the fan

The advantages of unitary air conditioners are as follows:

- Simple temperature control by the room occupant
- One room can be heated or cooled independently from other rooms in the building
- Malfunction affects only one room
- Low initial cost
- Space economy
- Individual energy metering possible
- Immediately available.

Disadvantages include:

- Limited performance options and air distribution control
- No possibility of humidity control
- Higher energy consumption due to the usually lower efficiency

10.5 Room Air Conditioners

Room air conditioners (cooling capacity, 1.2–10.5 kW) are intended for air cooling, humidifying, and cleaning and are designed for mounting in windows, as consoles, or through special openings in walls. They have low installation costs and a relatively low price, are reliable and easy to maintain, and no or a minimum number of air ducts are required.

The window-mounted air-cooling unit (Fig. 20) only cools the air, without controlling its relative humidity. The refrigeration power of such units varies from 1 to 8 kW. The condenser is air cooled and the fan blows the air through the evaporator and the condenser. Window-mounted room air conditioners are usually used when low cost, simplicity of

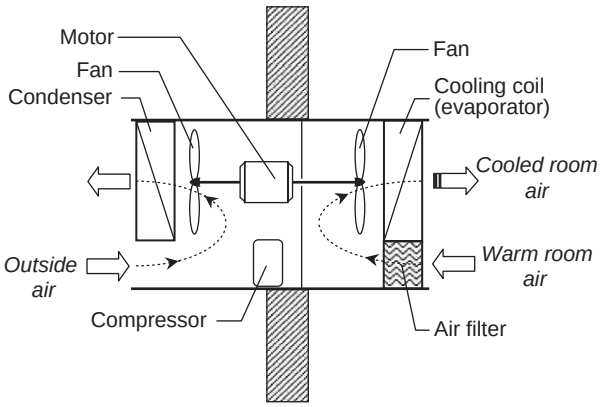


FIGURE 20 Typical window-mounted room air conditioner.

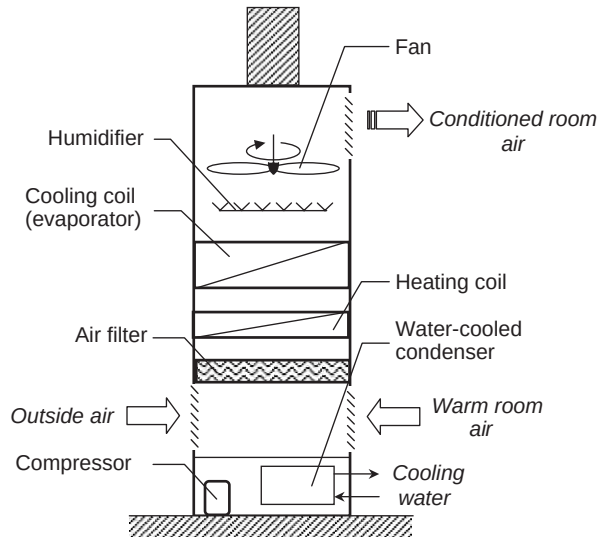


FIGURE 21 Typical through-the-wall room air-conditioning unit.

operation, and maintenance are primary criteria in selecting a candidate system (motels, low-cost apartments, and hotels), as auxiliaries to central systems, or as a backup for certain spaces when the basic system is down.

Through-the-wall mounted air conditioners (Fig. 21) contain a complete air-cooling and air-handling system packaged in one unit. They are used in buildings that require several temperature control zones (office buildings, hotels, and schools) and because less space is required compared with a central system, installation is relatively simple, and initial costs are generally lower than those of a central system. However, their noise level generally limits their applications.

SEE ALSO THE FOLLOWING ARTICLES

Air Pollution, Health Effects of • District Heating and Cooling • Electricity Use, History of • Heat Transfer • Indoor Air Quality in Developing Nations • Indoor Air Quality in Industrial Nations • Solar Cooling, Dehumidification, and Air-Conditioning • Thermal Comfort • Thermodynamic Sciences, History of

Further Reading

Anderson, E. E. (1993). "Thermodynamics." PWS, Boston.
Bejan, A. (1988). "Advanced Engineering Thermodynamics." Wiley, New York.
Howell, J. R., and Buckius, R. O. (1987). "Fundamentals of Engineering Thermodynamics." McGraw-Hill, New York.
Kutz, M. (Ed.) (1998). "Mechanical Engineers' Handbook". Wiley, New York.
Moran, M. J., and Shapiro, H. N. (1999). "Fundamentals of Engineering Thermodynamics." Wiley, New York.
Parsons, R. A. (ed.). (1987a). "ASHRAE Handbook." HVAC Systems and Applications. American Society of Heating, Refrigerating and Air-Conditioning Engineers, Atlanta.
Parsons, R. A. (ed.). (1987b). "ASHRAE Handbook." Equipment American Society of Heating. Refrigerating and Air-Conditioning Engineers, Atlanta.
Parsons, R. A. (ed.). (1990). "ASHRAE Handbook." Refrigeration Systems and Applications. American Society of Heating, Refrigerating and Air-Conditioning Engineers, Atlanta.
Sonntag, R. E., and Van Wylen, G. (1982). "Introduction to Thermodynamics Classical and Statistical." Wiley, New York.

Remote Sensing for Energy Resources

FLOYD F. SABINS
Remote Sensing Enterprises, Inc.
Fullerton, California, United States

1. Remote Sensing Technology
2. Oil Exploration
3. Environmental Applications

Glossary

anticline An arch in sedimentary strata. Oil may occur in the crest of anticlines.
electromagnetic spectrum The continuous sequence of electromagnetic energy arranged according to wavelength or frequency.
Landsat The unmanned Earth-orbiting satellites operated by NASA that acquire repeated remote sensing images.
radar Acronym for radio detection and ranging. Radar is an active form of remote sensing that records wavelengths from millimeters to centimeters.

Remote sensing is defined as the science of acquiring, processing, and interpreting images from satellites and aircraft that record the interaction between matter and electromagnetic energy. Remote sensing images of the earth are acquired in three wavelength intervals, or regions, of the electromagnetic spectrum. The visible region ranges from 0.4 to 0.7 μm and is divided into the blue, green, and red bands. The infrared (IR) region ranges from 0.7 to 30 μm and is divided into the reflected IR and thermal IR portions. The reflected IR portion ranges from 0.7 to 3.0 μm; the energy is predominantly reflected sunlight at wavelengths longer than visible light. The Landsat and other satellite systems acquire valuable images in the visible and reflected IR regions. The thermal IR portion ranges from 3.0 to 15.0 μm; the energy is radiant, or heat, energy. Thermal IR images have considerable potential for oil exploration in arid and semiarid terrains; however, the method has been underutilized in recent years largely because suitable images are lacking.

Visible, reflected IR, and thermal IR images are acquired in the passive mode by remote sensing systems that simply record the available energy that is reflected or radiated from the earth. The microwave region ranges from 0.1 to 30 cm; images are primarily acquired in the active mode called radar. An aircraft or satellite radar antenna transmits short bursts of microwave energy toward Earth. The energy interacts with materials and is returned to the antenna system, which records the information for later display as an image. Remote sensing is a rapidly evolving science with new imaging systems deployed on an irregular schedule by various nations and private concerns. The World Wide Web is a primary source of information about this dynamic science. Table I lists representative Web sites, many of which are linked to additional sites.

1. REMOTE SENSING TECHNOLOGY

The major application of remote sensing to energy is for oil exploration and, to a lesser extent, assessing the environmental impact of oil-related operations. The two most widely and successfully used systems are Landsat and radar.

1.1 Landsat Thematic Mapper Images

Landsat is an unmanned NASA satellite that orbits the earth in a sun-synchronous pattern at an altitude of 705 km.The satellite orbits over almost the entire earth in 16 days. If an area is cloudy during one overflight, suitable images may be acquired during subsequent 16-day cycles. Beginning in 1982, the

 © 2004 Elsevier Inc. All rights reserved.

second-generation Landsats 4 and 5 carried the thematic mapper (TM), which is a multispectral scanner that records images in seven separate wavelength intervals, or bands, for each scene. Bands 1–3 record blue, green, and red light. Bands 4, 5, and 7 record reflected IR energy. Band 6 records thermal IR energy, but it is rarely used for exploration. Table II lists the bands and their characteristics. Figure 1 shows reflectance spectra of vegetation and common sedimentary rocks together with wavelengths recorded by TM bands. The spectra of the different rocks are very similar in the visible bands but have major differences in the reflected IR bands; therefore, bands 4, 5, and 7 are especially useful for mapping different rock types.

Spatial resolution of the visible and reflected bands is 30 m. The current Landsat 7 carries the enhanced thematic mapper with an additional band 8 (0.52–0.90 μm), which has a spatial resolution of 15 m and can be convolved with the other bands during computer processing to improve their apparent spatial resolution.

Figure 2 is a subscene from a TM band 4 image of central Wyoming. The subscene displays less than 0.5% of the area covered by the full scene (170×185 km). Figure 3 shows the major features in the image and is generalized from a detailed geologic map that was interpreted for the full scene. The image includes four anticlines, two of which (Gebo and Little Sand Draw) are oil fields. These fields were discovered prior to the launch of Landsat, but this example demonstrates the value of the images for oil exploration.

Any three TM bands may be merged in any combination of blue, green, and red to produce color composite images that are far superior to panchromatic images, such as Fig. 2. TM images of the world are available at modest prices from the U.S. Geological Survey EROS Data Center in Sioux Falls, South Dakota (Table I). The images are distributed in digital format and must be computer processed to produce images. A variety of commercial desktop computers and software are available for this purpose.

Multispectral scanners, such as TM, record up to approximately 10 bands of data with wavelength intervals on the order of 0.10 μm. Hyperspectral scanners are a special form of multispectral scanners

TABLE I

Representative Remote Sensing Web Sites[a]

Web site	Web address	Comments
American Society for Photogrammetry & Remote Sensing	www.asprs.org/resources.html	Enter keyword "Satellites" for charts and tables describing past, current, and future remote sensing satellites
Canada Center for Remote Sensing	www.ccrs.nrcan.gc.ca	Canadian counterpart of EROS Data Center
EROS Data Center	edc.usgs.gov	Search and other aerial photographs, Landsat, radar, and other images
EROS Data Center DAAC	edcdaac.usgs.gov	Search and order SIR-C, ASTER, AVHRR, TIMS, MODIS, TERRA
Jet Propulsion Laboratory homepage	www.jpl.nasa.gov	Home page for JPL images: A VIRIS, TIMS, ASTER, SIR, others
JPL MASTER site	masterweb.jpl.nasa.gov	View and other MASTER aircraft hyperspectral images
JPL Shuttle Imaging Radar (SIR) site	southport.jpl.nasa.gov	SIR and Shuttle Radar Topography Mission (SRTM 2000)
NASA home page	www.nasa.gov	Links to NASA headquarters and centers
Purdue University	dynamo.ecn.purdue.edu/~biehl/multispec	Source for Multispec freeware for processing image data
Radarsat International	www.rsi.ca	Search and order information for worldwide satellite radar images
Space Imaging	www.spaceimaging.com	Search and other information for IKONOS high-resolution images
SPOT Image	www.spot.com	Search and order information for SPOT satellite images
USGS Spectroscopy Laboratory	speclab.cr.usgs.gov	Basic reference for hyperspectral imaging

[a] From Sabins (1997), Table 1–8.

TABLE II

Landsat Thematic Mapper (TM) Spectral Bands[a]

TM band, Landsat 4 and 5	ETM band, Landsat 7	Wavelength (μm)	Characteristics
1 (spatial resolution 30 m)	1(spatial resolutions 30 m)	0.45–0.52	Blue green; maximum penetration of clear water; useful for distinguishing soil from vegetation and deciduous plants from coniferous plants
2 (30 m)	2 (30 m)	0.52–0.60	Green; matches green reflectance peak of vegetation, which is useful for assessing plant vigor
3 (30 m)	3 (30 m)	0.63–0.69	Red; matches a chlorophyll absorption band that is important for discriminating vegetation types
4 (30 m)	4 (30 m)	0.76–0.90	Reflected IR; useful for determining biomass content and for mapping shorelines
5 (30 m)	5 (30 m)	1.55–1.75	Reflected IR; indicates moisture content of soil and vegetation; penetrates thin clouds
6 (120 m)	6 (60 m)	10.40–12.50	Thermal IR; nighttime images useful for thermal mapping and for estimating soil moisture
7 (30 m)	7 (30 m)	2.08–2.35	Reflected IR; especially useful for distinguishing rock types and for mineral exploration
	8 (15 m)	0.52–0.90	Visible and reflected IR; may be convolved with other bands to improve apparent spatial resolution

[a] From Sabins (1997), Table 3–4.

that record many tens of bands with wavelength intervals on the order of 0.01 μm. The Web sites for the Jet Propulsion Laboratory (JPL), U.S. Geological Survey Spectroscopy Laboratory, and EROS Data Center (Table I) describe the major hyperspectral systems. Hyperspectral images have had limited applications to oil exploration.

In addition to Landsat, a number of other Earth-orbiting remote sensing satellites have been launched and others are planned. The Web site of the American Society for Photogrammetry and Remote Sensing (Table I) summarizes the characteristics of these satellites.

1.2 Radar Images

Radar is an active form of remote sensing that is shown diagrammatically in Fig. 4. Pulses of microwave energy, at wavelengths of a few centimeters to a few tens of centimeters, are directed to the terrain from a fixed antenna carried on an aircraft or satellite. Energy is directed downward toward the side of the ground track. Each radar pulse "illuminates" a narrow strip, or scan line, of terrain normal to the track. The depression angle is measured between the horizon and the center of the pulse of radar energy. Figure 4 shows a single scan line. The incident energy interacts with the terrain and portions are returned to the radar system, which records the data. Three characteristics of terrain determine the intensity of the radar return:

1. Dielectric constant, which is largely a function of moisture content.
2. Surface roughness at a scale of centimeters. Smooth surfaces, such as calm water, have dark signatures. Rough surfaces, such as forests and coarse gravel, have bright signatures.
3. Topography at a scale of meters to kilometers. Figure 4 shows that slopes facing toward the radar antenna have bright signatures called highlights. Slopes facing away from the antenna have dark signatures called shadows.

Radar images have two major advantages over visible and reflected IR images, such as Landsat. Because radar is an active system that provides its own source of energy, images can be acquired day or night and through clouds, rain, and fog. Aircraft radar images are typically acquired with a low depression angle of 20° or less (Fig. 4), which causes strong highlights and shadows from topographic features that may express geologic features. Spatial resolution of modern synthetic aperture radar (SAR) images ranges from meters to tens of meters, even at satellite altitudes. Topographic maps are produced from repeated images using the process called interferometeric SAR (IFSAR), which is described on the JPL Shuttle Imaging Radar Web site (Table I).

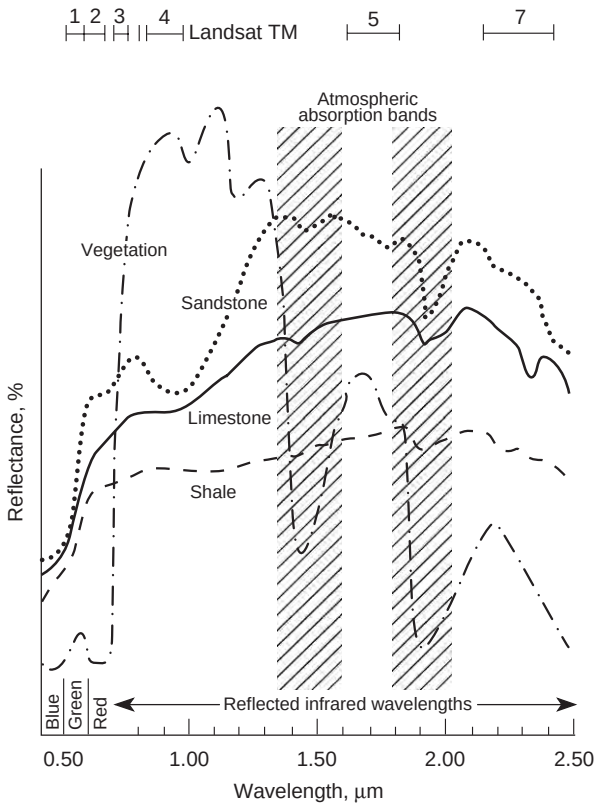


FIGURE 1 Reflectance spectra of vegetation and sedimentary rocks. Positions of TM bands are shown. Note that rock spectra have more distinguishing characteristics in the reflected IR region than in the visible region. Atmospheric absorption bands are caused by water vapor in the atmosphere, which prevents image acquisition. From Sabins (1997, Fig. 3-1).

FIGURE 2 Landsat TM band 4 (0.76–0.90 μm) image of subscene in central Wyoming. This subscene (190 km^2) is taken from a TM scene that covers 31,450 km^2. From Sabins (1997, Fig. 3-7D).

Figure 5 is a radar image along the south flank of the Papuan Fold and Thrust Belt in Papua New Guinea (PNG). Figure 6 is a generalized geologic map interpreted from the image. The terrain consists of northwest-trending anticlines and faults that form topographic ridges and scarps. Radar illumination was toward the north, which forms bright highlights on southwest-facing slopes and dark shadows on northeast-facing slopes.

The Web sites for the EROS Data Center and JPL provide examples and information on aircraft and satellite radar images acquired by U.S. government agencies. The Radarsat Web site provides access to satellite radar images acquired by Canada.

1.3 Digital Image Processing

Except for conventional aerial photographs, all remote sensing images are recorded in digital format and are computer processed to produce images for interpretation. The three functional categories of image processing are

1. Image restoration, which compensates for data errors, noise, and geometric distortions introduced during the scanning, recording, and playback operations.
2. Image enhancement, which alters the visual impact that the image has on the interpreter in a manner that improves the information content.
3. Information extraction, which utilizes the decision-making capability of the computer to recognize and classify pixels on the basis of their digital signatures.

The Purdue University Web site (Table I) provides image processing software.

2. OIL EXPLORATION

The major energy application of remote sensing is for oil exploration. In the early stages of an exploration

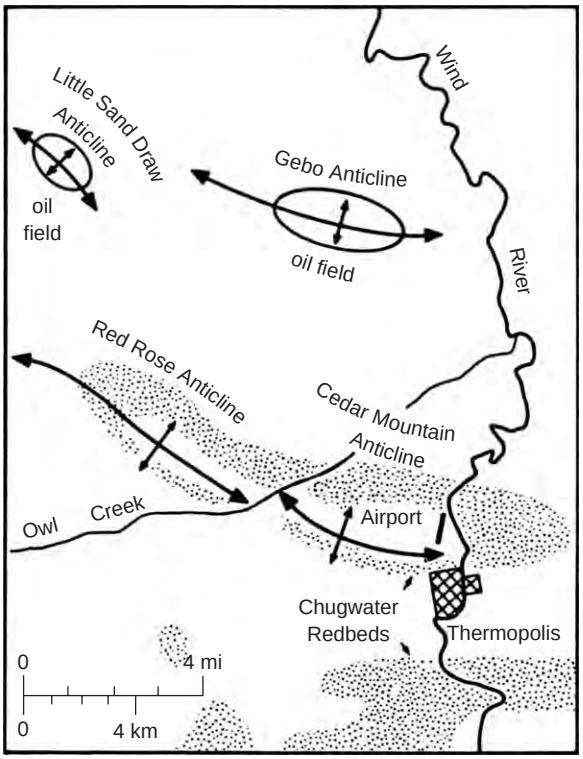


FIGURE 3 Map showing major features in Wyoming subscene. From Sabins (1997, Fig. 3-7H).

project, images are used to outline sedimentary basins and potential drilling prospects. Geologic fieldwork and geophysical surveys are used to verify and refine the prospects prior to drilling exploratory wells. In 1978, the Chevron Corporation (now ChevronTexaco) utilized images from an early version of Landsat to help discover the first oil fields in the Sudan. Remote sensing has subsequently played a key role in major discoveries in PNG and the Central Arabian Arch. The PNG project area is a tropical rain forest with complex geology, whereas the Arabian project is a desert with relatively simple geology. These two case histories illustrate the successful application of different remote sensing technologies in widely different geologic and environmental settings.

2.1 Papua New Guinea Project

In 1984, the Chevron Corporation explored two petroleum prospecting licenses (PPLs) in the eastern portion of the island of New Guinea, which is part of the nation of PNG. The PPLs cover an area of 14,400 km^2 that spans the boundary between the Fly–Strickland Lowlands in the southwest and the Papuan Fold and Thrust belt in the northeast. Oil

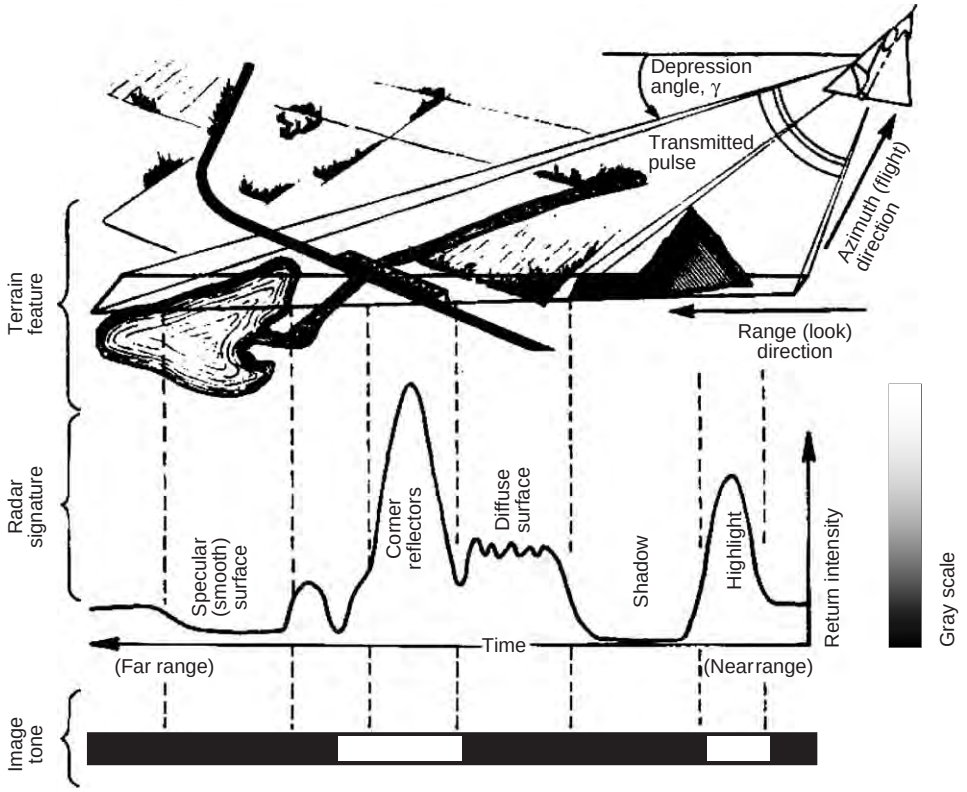


FIGURE 4 Terrain returns and image signatures for a pulse of radar energy. From Sabins (1997, Fig. 6-3).

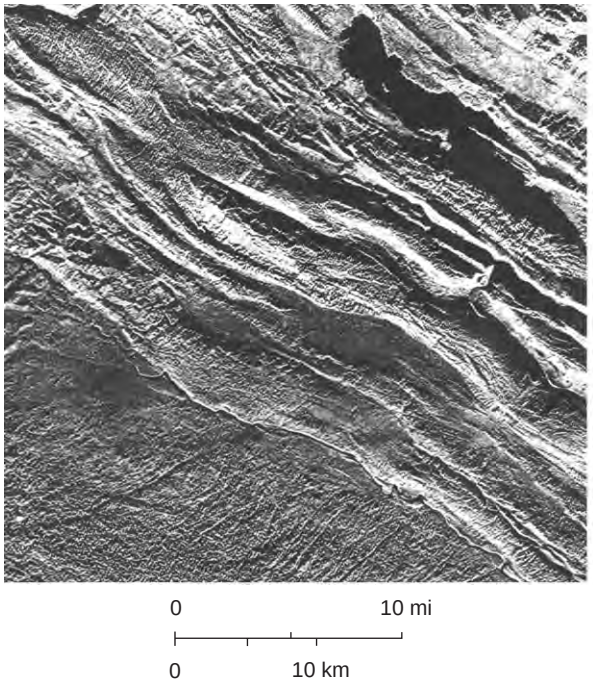


FIGURE 5 Aircraft radar image of the Lake Kutubu area of Papua New Guinea. From Sabins (1997, Fig. 10-8).

companies have explored and drilled the region for many years because of the numerous surface seeps of oil and gas and hydrocarbon shows in exploratory wells. When this project began, 144 wells had been drilled in PNG, but no commercial oil fields had been discovered. This remote area has a canopy of tropical rain forest and a hot, humid climate with persistent cloud cover and heavy rainfall. The terrain is dominated by steep ridges and deep river channels. Much of the region is underlain by limestone that weathers into steep pinnacles and pits, called karst topography, that cannot be traversed by vehicles. Reliable seismic data were unobtainable because of the rugged terrain and the weathering effects caused by pervasive subsurface solution cavities.

At the beginning of the project, Chevron decided to rely on radar images and limited fieldwork. A contract radar survey aircraft covered the area with a series of parallel image strips that were compiled into mosaics and interpreted to produce geologic maps. Based on these maps, the first well was drilled on a promising structure. The well was dry but had encouraging shows of oil. Exploration then shifted southeast to the region of Lake Kutubu. Figure 5 is

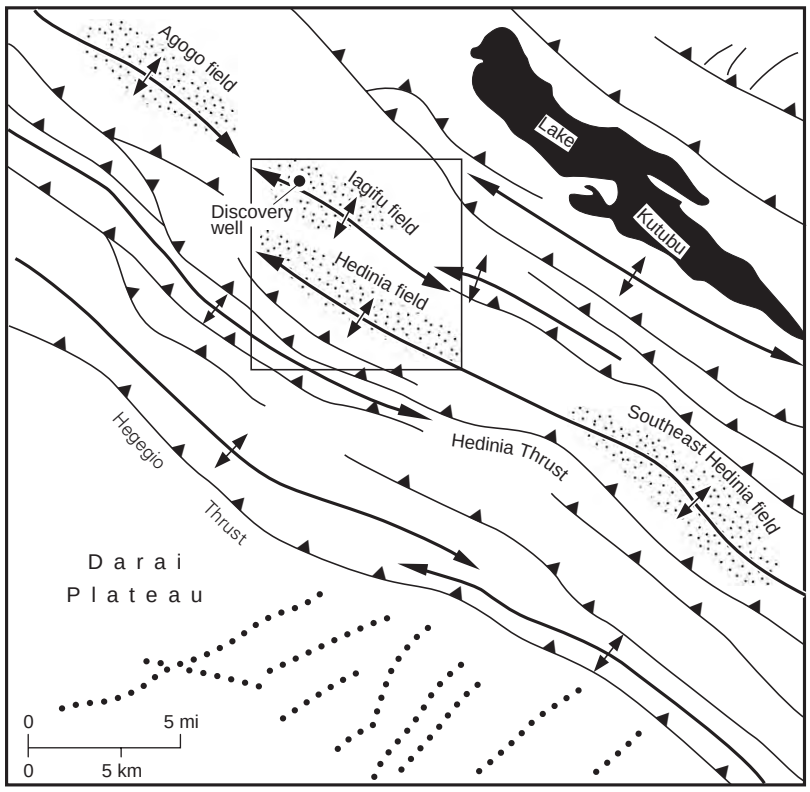


FIGURE 6 Geologic interpretation of radar image of the Lake Kutubu area of Papua New Guinea. From Sabins (1997, Fig. 10-9).

the radar image and Fig. 6 is the interpretation map for the area. Based on the radar interpretation and limited fieldwork, the first commercial oil discovery in PNG was made in March 1986 on the Iagifu Anticline at the location shown in Fig. 6. Subsequently, Chevron has discovered several additional commercial fields. A number of promising structures remain to be drilled in the region.

2.2 Central Arabian Arch Project

For a number of years, the Arabian American Oil Company (Aramco) was restricted to explore only within the "Retained Areas" of existing production that are located along the Arabian Gulf. Once the giant fields were drilled, no new fields were discovered through 1986. At that time, the government directed the company, now called Saudi Aramco and owned by the government, to explore the vast region of the Kingdom with oil potential that lies outside the Retained Areas. In early 1988, Saudi Aramco asked the Chevron Remote Sensing Research Group to conduct a remote sensing analysis of this region. The initial phase of the study covered the Central Arabian Arch, an area of approximately 200,000 km^2. The two objectives of the project were to map the regional geology and identify surface features that may be the expression of oil prospects.

The climate, topography, and geology of Saudi Arabia are the opposite of those in PNG. The arid climate and lack of vegetation are ideal for using Landsat images. The geologic structure is simple and uniform, in contrast to the complex faulting and folding in PNG.

Landat TM bands 2, 4, and 7 were combined in blue, green, and red to produce color images that were interpreted to produce regional geologic maps at a scale of 1:250,000 (1 in. = 4 miles). Aside from the sand seas, bedrock is well exposed. Geologic structure consists of regional dips of approximately 1° to the northeast or east on opposite flanks of the Central Arabian Arch. Distinctive color signatures on the images enabled the interpreters to divide several formations of sedimentary rocks into two or more mappable subdivisions, which proved valuable for identifying oil prospects.

Identifying potential drilling prospects was challenging because the geologic structure is dominated by broad uniform dipslopes that extend for several hundred kilometers along the strike and several tens of kilometers in the downdip direction. Unlike PNG, there are no obvious folds. Within the Retained Areas, oil fields are trapped in gentle anticlines formed above high-angle faults that offset the underlying rocks. The anticlines grade upward into flattening of the regional dip to form subtle structural terraces at the surface. We used this information to develop a three-dimensional geologic model for recognizing potential oil prospects on the images. A number of prospects were interpreted and were checked by geologists in the field. Saudi Aramco then followed up with seismic surveys of the most promising prospects.

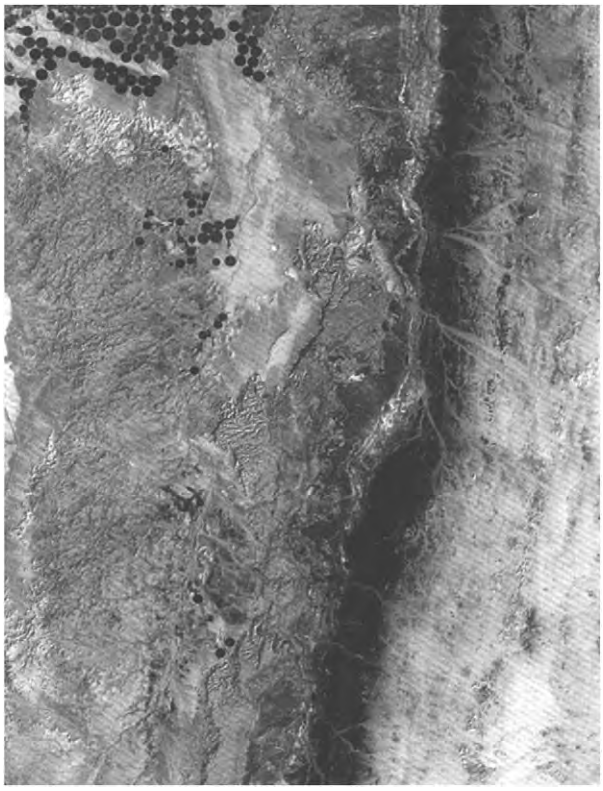


FIGURE 7 Landsat TM image of the Raghib prospect, Central Arabian Arch. From Sabins (1997, Plate 20-D).

Figure 7 is the TM image of the Raghib prospect located 110 km southeast of Riyadh. This black-and-white version lacks much of the information shown in the original color image that was used for our interpretation. Figure 8 is the geologic interpretation map. Jurassic and Cretaceous strata dip 1° east. Arcuate scarps form the downdip margin of an anomalous topographic depression that is mantled by windblown sand. Figure 9 is a subsurface structure map from a seismic survey that defines a northwest-trending anticline that is 35 km long and 10 km wide. The axis of the anticline is annotated on the Landsat interpretation map (Fig. 8). The topographic depression on the image coincides with the

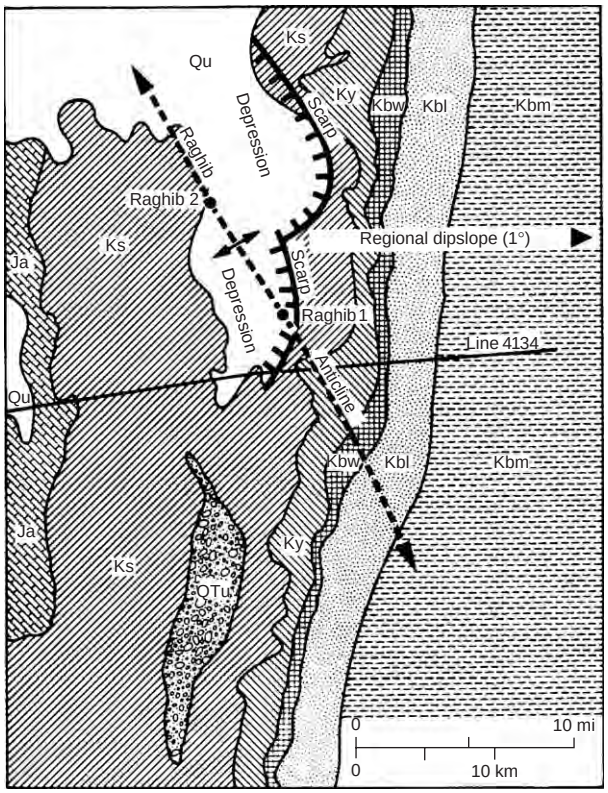


FIGURE 8 Geologic interpretation of the Raghib prospect. Raghib 1 is the discovery well and Raghib 2 is the confirmation well for the Raghib oil field. The patterns and symbols designate geologic formations. From Sabins (1997, Fig. 10-20).

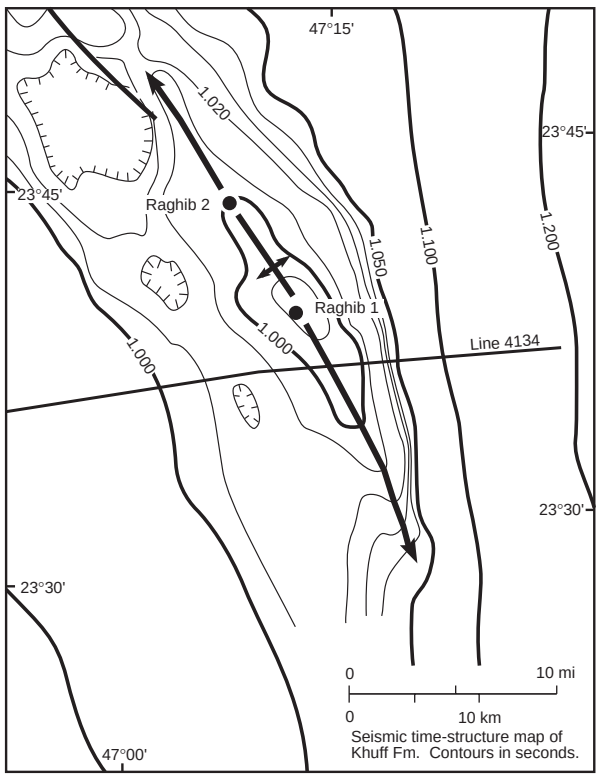


FIGURE 9 Seismic structure map of the Raghib oil field. The area coincides with the image (Fig. 7) and geologic map (Fig. 8). Contours are drawn on top of the Khuff Limestone (Permian age) and delineate a closed anticline that coincides with the surface feature. From Sabins (1997, Fig. 10-21).

crest of the subsurface anticline. In late 1989, Saudi Aramco drilled the Raghib 1 exploratory well located on the crest of the subsurface anticline (Fig. 8) and discovered a new oil field. A number of prospects remain to be explored.

2.3 Operational Applications

In addition to defining prospects, images are a valuable asset to field operations in regions that lack reliable maps. The cost and effectiveness of seismic surveys can be improved by information on topography and trafficability of the survey area. In the Sudan, Landsat images provided the only reliable base maps. The images were also digitally processed to map the different terrains, such as rivers, marshes, uplands, and papyrus swamps, that required different vehicles for the seismic surveys. In the rugged terrain of Yemen, topographic maps were compiled from satellite images and used to lay out the optimum pattern of seismic survey lines. In older oil fields, well sites and pipelines may be mapped inaccurately but can be correctly located with high-resolution satellite images. The cost is much less than that of a conventional survey on the ground.

2.4 Direct Detection of Oil

In PNG, Saudi Arabia, and elsewhere, remote sensing is used to recognize the surface expression of the subsurface structures (folds, faults, and domes) in which oil is trapped. Oil is not detected. In a number of oil deposits, some oil leaks to the surface to form oil seeps. During the early days of the oil industry, exploratory wells were commonly located at seeps. In a few fields the leaking oil has interacted chemically with the surface rocks to alter their mineralogy and color. Some investigators have attempted to identify seeps and altered rocks as an exploration method called direct detection. To date, the results of this method have been inconclusive.

3. ENVIRONMENTAL APPLICATIONS

Offshore oil spills are predominantly caused by tanker accidents and, to a lesser extent, drilling and production activities. Remote sensing images are used to detect spills and to monitor and direct cleanup operations. Table III lists the interactions between oil and electromagnetic energy at wavelength regions ranging from ultraviolet (UV) to radar. Figure 10 shows these interactions diagrammatically. The following sections describe the detection of oil spills on remote sensing images acquired in the various spectral regions. In offshore areas, oil may leak from deposits beneath the seafloor and form slicks on the water. Remote sensing of these natural slicks has been used to aid offshore exploration.

3.1 Ultraviolet Images

The UV spectral region spans wavelengths from 0.03 to 0.40 μm. Wavelengths shorter than 0.30 μm are absorbed by the atmosphere; therefore, only wavelengths from 0.30 to 0.40 μm are used in remote sensing of the earth. Figure 10A shows that incoming UV radiation from the sun stimulates oil to fluorescence at wavelengths from 0.30 to 0.45 μm (long-wavelength UV to visible blue). UV images are the most sensitive remote sensing method for monitoring oil spills and can detect films as thin as 0.15 μm. Daylight and very clear atmosphere are essential to acquire UV images. UV energy is strongly scattered by the atmosphere, but useable images can be acquired from altitudes below 1000 m. Cameras and scanners are passive systems that operate in daytime to record fluorescence stimulated by the sun. Active UV systems have been developed for aircraft use. A laser irradiates the water with UV energy that stimulates any oil to fluoresce. The fluorescence is recorded as a spectrum or as an image. The spectrum can be compared with a reference library of spectra to identify the oil in the spill. These active systems can acquire images both day and night but must operate at low altitudes under clear conditions.

3.2 Visible and Reflected Infrared Images

The visible spectral region spans wavelengths from 0.40 to 0.70 μm and is detectable by the eye. The reflected IR region extends from 0.40 to 3.00 μm, which is beyond the sensitivity range of the eye. The interaction between oil and electromagnetic energy in the visible and reflected IR regions is determined by the absorption and reflection of sunlight (Fig. 10B). Various types of aerial photographs record visible wavelengths and the shortest IR wavelengths (0.70–0.90 μm). Landsat TM records three visible and three reflected IR bands in digital format that is suitable for computer processing. During the invasion of Kuwait in January 1991, Iraqi forces deliberately released 4–6 million barrels of crude oil into the Arabian Gulf, where it covered approximately 1200 km^2 of water and 500 km of the Saudi Arabian coastline. This spill is the largest on record; for comparison, the 1989 Exxon Valdez spill in Alaska (<300,000 barrels) was an order of magnitude smaller. Repeated TM images tracked the extent and movement of the Kuwait spill.

3.3 Thermal IR Images

The thermal IR region spans wavelengths from 3.0 to 14.0 μm (Table III) and consists of energy radiated, or emitted, from materials. Figure 10C shows the interaction of thermal IR energy with oil and water. Both liquids have the same kinetic temperature of 18°C because they are in direct contact. Remote sensing systems, however, record radiant temperature. Radiant temperature differs from kinetic temperature because of the property called emissiv-

TABLE III

Remote Sensing Interactions and Systems for Detecting Oil on Water[a]

Remote sensing region	Wavelength	Oil property detected	Oil signature	Water signature
Ultraviolet	0.3–0.4 μm	Fluorescence stimulated by sunlight or laser	Bright	Dark
Visible and reflected infrared	0.4–3.0 μm	Solar reflectance	Brown, black	Blue, green
Thermal infrared	3.0–14.0 μm	Radiant temperature, controlled by emissivity	Dark (cool)	Bright (warm)
Radar	3.0–25 cm	Dampening of capillary waves	Dark (smooth)	Bright (rough)

[a] From Sabins (1997), Table 3–4.

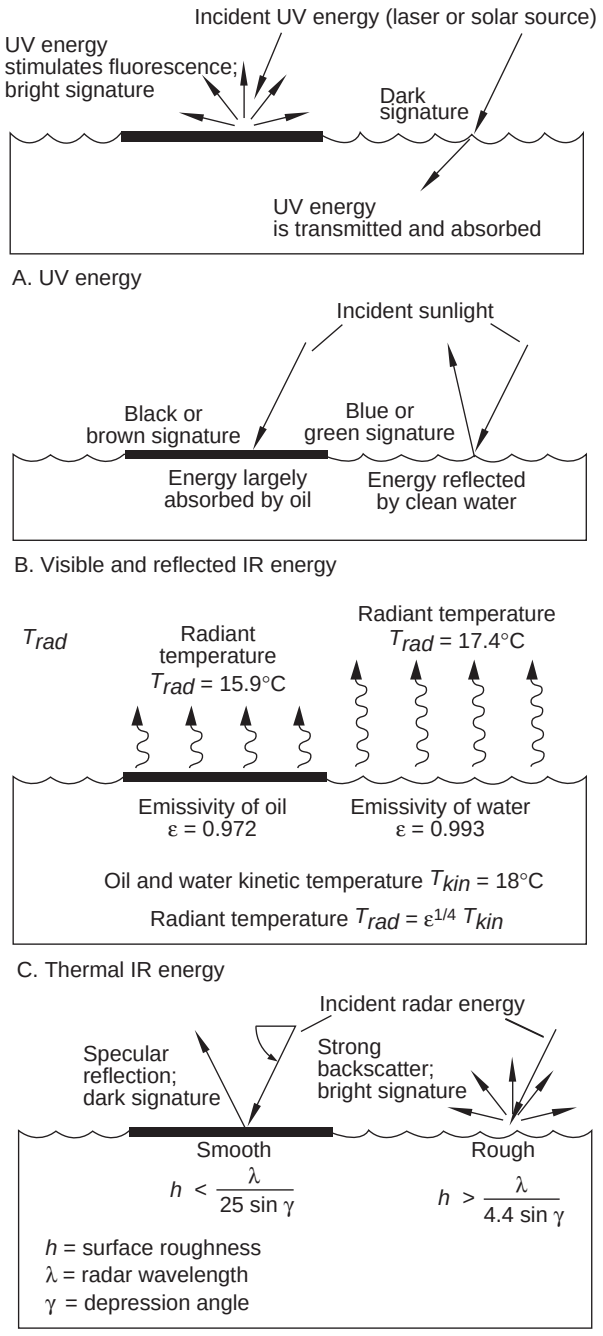


FIGURE 10 Interaction mechanism between water and oil spills. From Sabins (1997, Fig. 9-30).

ity, which expresses the ability of a material to emit radiant energy. Figure 10C shows emissivity values for oil (0.972) and water (0.993) and the equation for calculating radiant temperature. Substituting these values into the equation (for a kinetic temperature of 18°C) results in a radiant temperature of 17.4°C for water and 15.9°C for oil. The difference of 1.6°C in radiant temperature between oil and water is readily measured by thermal IR systems, which are typically sensitive to temperature differences of 0.1°C. Thermal IR images acquired from aircraft and satellites record oil slicks with cool signatures that are distinct from the warmer signatures of clean water.

3.4 Radar Images

The interaction between the surface and incident radar energy is very sensitive to variations in roughness of the surface, measured in millimeters and centimeters. Rough surfaces scatter large amounts of energy back to the antenna and have bright signatures. Smooth surfaces reflect the energy away from the antenna and have dark signatures. In Fig. 10D, clean water has characteristic ripples and capillary wavelets that cause a bright radar signature. An oil spill dampens the waves and causes a smooth surface, or slick, with a dark radar signature.

SEE ALSO THE FOLLOWING ARTICLES

Crude Oil Spills, Environmental Impact of • *Gas Hydrates* • *Oil and Natural Gas Exploration*

Further Reading

Berger, Z. (1994). "Satellite Hydrocarbon Exploration." Springer-Verlag, Berlin.

Jensen, J. R. (1986). "Introductory Digital Image Processing—A Remote Sensing Perspective." Prentice-Hall, Englewood Cliffs, NJ.

Prost, G. L. (1993). "Remote Sensing for Geologists." Gordon & Breach, Lausanne, Switzerland.

Sabins, F. F. (1997). "Remote Sensing—Principles and Interpretation." 3rd ed. Freeman, New York.

Renewable Energy and the City

PETER DROEGE
University of Sydney DEM Australia/World Council for Renewable Energy
Sydney, Australia

1. **Energy Technology and Modern Urbanization**
2. **Historical Perspectives**
3. **Innovation in Urban Renewable Energy Technology**
4. **The Urban Renewable Energy Portfolio**
5. **Climate Stability in Urban Development**
6. **Urban Renewable Energy Policy and Practice Examples**

Glossary

city An urbanised area managed and represented by one or several local governments, culturally and communally understood as a city, with specific administrative and political boundaries.

city region The general urbanized, ecological, and economical area constituting and surrounding one or several urban nuclei, all or any of which may be defined as a city.

climate-stable practice A city's climate-stable practice is a practical commitment to lower greenhouse gas emissions by a target date such as the year 2050 to an amount that is proportionally in keeping with the globally sustainable level.

carbon-dioxide-equivalent (CO_2-e) Expresses the presence of all effective greenhouse gases as an amount that would be required in CO_2 to achieve the same effect.

greenhouse gas (GHG) Human activity effected gases that trigger global warming are carbon dioxide (CO_2), methane (CH_4), nitrous oxide (N_2O), chlorofluorocarbons, especially CFC-11, HCFC-22, and CF_4. In this article their emission levels are expressed as CO_2-e.

solar In its wider definition connotes all aspects of energy sources, which can be traced to the action of the sun. These include solar thermal, photovoltaic, biomass, bio energy, wind and wave energy, and, strictly speaking, also fossil fuels. In the pure sense, solar describes renewable forms of energy that neither create greenhouse gas emissions nor nondegradable or toxic waste.

urban(ized) area A builtup space of higher density and production and consumption functions that is distinct from rural or undisturbed natural areas.

Cities are intensive centers of social activity and the central feature in the economical technology of civilizations. Cities directly express the energy resource framework and infrastructure available at any given time in history. Global acceleration in city formation and urban growth can be ascribed to a number of technological innovations, none more profound in their impact than the rise and spread of fossil fuel resources, products, systems, and tools. The resulting massive threats—CO_2 emissions–induced climate change and the inevitably approaching plateau in worldwide fossil fuel production—raise the vexing issue of cities' overwhelming dependence on oil and other hydrocarbon energy sources as well as the question of how and when a postfossil fuel urban era can commence and whether this necessary historical transformation can take place in a stable manner.

1. ENERGY TECHNOLOGY AND MODERN URBANIZATION

The age of industrialization came into full force through the modern exploration and use of fossil fuels. As one of its most striking phenomena, the rapid expansion of cities throughout the late 19th and the 20th centuries was a direct outcome of the fossil fuel economy as well. The development and operation of cities and urbanized areas presently absorbs roughly three-quarters of the world's fossil fuel production, including intra- and interurban transport. This indicates cities' massive role in the worldwide carbon consumption and emission regime: fossil fuels, prime source of anthropogenic CO_2 emissions, supply 85% of total global commercial energy use—and their consumption is increasing at a rapid rate.

Yet economic regions, nations, and cities worldwide are under great pressure to find alternative

 © 2004 Elsevier Inc. All rights reserved.

sources. Cities will have to be powered differently within the lifetime, prompted by a historical coincidence of carbon emissions–induced climate change, the great likelihood of an impending plateau in fossil fuel production—while energy demand continues to rise dramatically, prompting steep price rises—and mounting environmental damage due to the use of oil and coal. The use of renewable and distributed micropower systems is already on the rise, but the speed of change is seen as too low by many to meet global goals in time to avert serious crises. Combined with large-scale facility efficiency and energy conservation programs, the introduction of renewable energy generation technologies in cities offers one set of profound opportunities as well as challenges, and the reform of transport systems, settlement patterns, and land-use practices provides another.

Cities, towns, and other urban communities are regarded as settings for coordinated efforts aimed at global renewable energy technology introduction and carbon emissions reduction programs. However, substantial organizational and cultural barriers militate against immediate, wider change. Among these loom large the subsidiary regard in which cities are held in the traditional hierarchical frames of the international arrangements that deal with global issues. Another challenge is the energy-blind nature of contemporary physical planning, typically very short planning horizons and political uncertainties that prevail on the local level. In terms of policy development, measurement techniques, and planning reality, an extraordinary gulf exists between the global nature of greenhouse gas impacts and fuel depletion prospects and the local reality that represents both final impact and original source of globally experienced change.

Due to their fundamental fossil fuel reliance, most capital outlays in new buildings, infrastructure, and other urban facilities risk early obsolescence: urban areas, infrastructure systems, and individual buildings and other facilities that are based on fossil fuel construction, maintenance, and operation are likely to be rendered dysfunctional well before the end of their life. All basic urban communication infrastructures, both traditional (such as roads, rail, air, and seaports) and advanced (such as telecommunications), have been nurtured in a world of near-absolute fossil fuel dependency. Indeed, the Internet, a vital global and largely urban network of networks, relies largely on fossil fuel operated hardware, conduit-based webs, and wireless communication systems. If global communications are to be sustained reliably beyond the middle of the 21st century, they must be powered by distributed, ubiquitous, and redundant renewable power supply systems. The required technology arrays with the concomitant policy and action frameworks do in principle exist to bring about manageable postfossil fuel urban environments.

2. HISTORICAL PERSPECTIVES

2.1 The Rise of the Fossil City

Modern cities have grown on a rich fossil fuel supply base, and especially dependent are the largest, most rapidly growing urban agglomerations. The very logic of their global rise and regional spread is founded on the availability of powerful, centralized, and inexpensive fuels—coal, petroleum, and natural gas—yielding fossil urban structures and patterns based on fossil transport, fossil construction machinery, and fossil industrial systems and manufacturing processes. Intensive economies and labor markets are clustered around the centralized and networked city regions, anchored by heavy investments in infrastructure: power, transport, and communications. This has boosted the primacy of cities over—and ultimately detachment from—agrarian hinterlands.

The new cities of the 19th and 20th centuries—and the very cultures they engendered—were a product of the rising combustion economy. The sheer size of London exploded when coal-fired power made this a technological possibility. Historian Lewis Mumford called the phenomenal urban transformations taking place in industrializing countries between 1820 and 1900 “Coketown” or “Paleotechnic Paradise.” Pre- and post—World War I modern city innovations in the Soviet Union, the United States, Europe, and across Asia were brought to life by the electrifying jolt of the new energy technologies. In the 1940s and 1950s, industrial and government-led transport policies revolutionized urban America away from rail-based modes to technologies promoting petroleum, automobile, and tire consumption. The rising car dependency in most cities, combined with the spread of electrified suburban households under expanding power grids, constitutes one of the most vexing challenges of managing urban environments worldwide today.

Fossil fuel–powered industrial transformations that underpinned broad urban structure change were celebrated by the leaders of the great modern design

movements, from the beginning of the 20th century: Italian futurism, constructivism in the early Soviet Union, De Stijl in the Netherlands, the Bauhaus in Germany, the declarations of the International Modern Architecture Congresses (CIAM—Les Congrès Internationaux de l'Architecture Moderne, 1928–1956) and the International Style that spread from the United States throughout the industrialized world.

The fossil machine age is an outgrowth of earlier stages of the Industrial Revolution when water power was a main driver of textile mill operations. Boosted by the industrial application of coal and electricity, it gave rise to Frederick Winslow Taylor (1856–1915) and Henry Ford's (1863–1947) ideas about the mechanization of manufacturing, as noted by Gideon in 1948. In their wake, the increasingly urban and automated global production-consumption systems of the advanced industrial age boosted global power use at an exponential rate.

But industrialization as powered by electricity, coal-fired steam engines, and petroleum combustion motors also meant the rapid growth of cities, driving the search for innovation in urban form. A plethora of utopian premonitions spread as a result, as revolutionary as the technological changes that triggered them. Peter Kropotkin's (1842–1921) and other anarcho-syndicalist influences on ideas about ideal communities were a direct outgrowth of this era; so were Sir Ebenezer Howard's (1850–1928) concepts and the rise of the Garden City movement. The genesis of the Regional Plan Association of New York and its seminal plan of 1929 were of this lineage, as was Frank Lloyd Wright's (1869–1959) Broadacre City (first described in his 1932 work *The Disappearing City*). Ludwig Hilbersheimer's (1885–1967) mass housing concepts and General Motors' Futurama pavilion at the 1933 "Century of Progress" World Exposition in Chicago boldly anticipated the great car-friendly cities to come. Brasilia, Canberra, and all of the British New Towns are fossil technology plans come to life.

The urban environmental phenomenon of the modern suburb followed. Early ex-urban subdivisions sprung up as aspiring new communities along the tracks of electric tramways, such as the famed streetcar suburbs of many North American, European, and Australian cities such as Brookline, Newton, or Jamaica Plain in Boston, or Balmain, Rozelle, or Mosman in Sydney. Soon thereafter the combustion-engine driven vehicle created a new urban reality altogether: the new car suburbs expanding along motorways. Mass media–driven promotion of a new fossil fuel–based lifestyle spawned the automobile and petrochemical product–based industrial civilization after World War II. This new vision became expressed in a new set of urban form aspirations. Many of these early premonitions proved to be very much practical concepts: already very early in the 20th century the great electric inventions of the telephone and the elevator opened the door to skyscraper cities, first as utopian visions and later as pervasive reality.

Modern carbon-based energy civilization is epitomized in its great landmarks with the seminal work of migrant Bauhaus architects such as Walter Gropius and Mies van der Rohe. Swiss architectural pioneer Le Corbusier's ideas about the radical modernization of cities are widely regarded the most influential of these, boldly calling for the razing of prefossil cities, as exemplified by his *tabula rasa* concepts: the Parisian Plan Voisin; La Ville Radieuse; the famed lone pilot project for a mass housing block, Unite d'Habitation; and in some realized urban projects such as in the Alsatian town of St. Di*é*. His plan for Chandigarh, the joint capital of the Punjab and Haryana states established in 1953, stands out among the league of pioneering fossil fuel based urban innovations. This simple yet revolutionary doctrine was soon implemented in cities in the United States, the United Kingdom, and countries around the world. It brought urban destruction and blight through wholesale urban renewal projects in what were identified as inner-city slum areas: the neighborhoods of the poor and disenfranchised and the prefossil relics they inhabited.

The postwar years in Europe saw powerful expansion plans realized, taking advantage of the spreading webs of electrification that crisscrossed former farmlands. The Greater London strategy, the European satellite city concepts, and the Stockholm and Copenhagen finger plans gave bold plan form to a wider, worldwide turn to life in cities, driven by structural shifts and opportunities brought about by the rise of the fossil fuel economy and its promises.

Yet the new vision of renewed and more orderly, healthier, and socially more equitable cities did not escape the environmental reality of advanced industrialization. With the burning of fossil fuel arrived sulfur dioxide, nitrous oxide, and a host of other toxic gases as well as carcinogenic airborne combustion particles, creating the smog and air pollution crises of the 1960s and 1970s. With the exception of the greenhouse gases themselves, these have largely been conquered as hazards in the industrialized world while still plaguing large cities of the developing world.

Mounting calls for more densely concentrated and public transport-based forms of urban development since the 1970s were early signs of a declining stage of the fossil-fuel economy when savings in energy conservation and fuel efficiency began to be identified as the cheapest, fastest, and most immediately useful means to reduce emissions. Denser cities were shown to be more fuel-efficient, while car-dependent, low-density urban structures incapable of sustaining public transport have come to be understood as a major hindrance to achieving sustainable urban life.

It is widely understood, however, that efficiency and conservation measures are not sufficient to halt the global rise in CO_2 emissions in the long term. Massive substitutions of conventional energy technologies conveyance with renewable systems of energy supply and use have long been accepted as essential. For these new technologies to be in place in time to be of significant use, they need to be widely introduced as soon as possible. In this sense, transit-oriented development and attempts at re-creating preindustrial urban patterns such as neo-traditional design and other approaches are important interim measures.

Both nature and form of buildings were changed by the development and spread of the revolutionary new fuels and the industrialized manufacturing processes they engendered. Urban environments and architecture changed radically in the new fossil age, breaking with all earlier traditions. The Modern Movement stylistically refined mass applications in electricity, power machinery, air conditioning, industrial steel products, advancing glass technology, mass produced curtain walls, prefabricated building systems, and a number of other highly energy-intensive technologies. The new thinking about buildings, highlighted by the International Style, applied to aesthetically refined, ornament-free, skeletal, industrialized, and largely corporate structures. It became the new global aesthetic of the possible as the advanced fossil age dawned in cities around the world. Buildings became disconnected from their climatic and cultural context due to the end of local resource dependency. An interest in vernacular form and regionalism emerged as a counter-reaction and an attempt to recapture the local language of form: the regional building traditions that had been lost in industrializing change. This movement remained a stylistic artifice since it failed to address the fundamental dynamics of design and development under conditions of unsustainable energy resource practice.

2.2 Local Measures: Fossil Fuel Conservation versus Replacement

As a reaction to these historical perspectives, the building industry has witnessed a certain rise in design responses to regional climatic conditions, as part of a powerful efficiency and conservation push since the 1970s. More recently, the zero-emissions house and office building has become a design concept as part of strategies to introduce renewable energy as an increasing contributor to managing urban energy supplies. In its pure form it is to function without primary fossil or nuclear energy supply input and has been the focus of worldwide research and development efforts for several decades. Also addressed is the question of how to reduce the large amount of fossil energy that is embodied in the building materials themselves, the services contributing to their making, and the energy implications of the very form and design of neighborhoods and cities. The energy household of cities is expressed by the sum total of all goods and services consumed in a given location: hence location-based emissions reduction measures are a primary target of urban energy reform.

The urban transport question, too, has opened up a powerful energy dimension, with two countervailing fundamental positions emerging. The energy conservation strategy points to the massive sprawl and wastefulness engendered by modern car-based cities, with the least dense consuming the highest amount of gasoline per capita. The public transit oriented development and urban consolidation or compact city policies of the 1980s resulted from these observations. The other argument is based on a general conviction that the age of individual transport is inevitably going to stay. It sees the future of personal transport to lie in biofuel and hydrogen combustion or fuel cell–powered vehicles. In this scenario the advocates of denser, more complex urban settings serving cultural and social objectives would lose out, since it paints the prospect of a solar-hydrogen energy revolution that could also accelerate urban sprawl.

2.3 Global Aspects

Global urbanization as a product of the fossil-fuel economy is expressed in the ready availability of inexpensive fossil power for all urban infrastructures, telecommunications, and freight and human transport systems on surface, sea, and air. These bundles of infrastructure at once link cities globally and bypass their regions. As a consequence, globalized urban systems are inherently more vulnerable to the

decline of global fossil fuel supplies than those that rely more on their local and regional human and land resources.

The deployment of renewable energy technologies on a large and especially urban scale has the potential of bringing about a time of differentiated globalization, marked by a distinction between local supplies of food and elementary goods on the one hand and the global trade in services on the other. This condition as brought about by nonfossil production modes would be characterized by a rise of regional economies in support of urban centers, based on regional resources such as productive land for food, biomass, and wind energy production. New ways of reknitting central cities with their regional economies and related spatial structures are already being pursued by a number of communities. These are based on older principles of rural urban support economies, boosting the primary sectors of agriculture and forestry. Cities around the world have started to make concrete links between their renewable energy needs and potential regional resources capable of meeting that need. This movement is also beginning to help spawn new indigenous manufacturing and advanced industry sectors in renewable energy production, supply, and services.

In urban industrial terms, initiatives fall into two categories: new technology development and the accelerated market uptake of applications. Technology push and supply from the 1970s through the 1990s were limited to a number of limited-scale industry efforts and pockets of largely government-sponsored research and development projects. This history is often marked by a mismatch with market demand, especially given the powerful subsidies granted to the fossil energy sector. Yet the international and domestic policy and pricing environment of the early 2000s promises to change fast, heralding massive urban technological and practice changes and a natural integration of technology development and markets.

While a few local urban systems may seem relatively safe from a terminal fossil fuel shock through their reliance on coal, hydroelectric, nuclear or bioenergetic power, no currently utilized alternative energy source alone can keep operational the vast majority of cities. Also, the interconnectedness of the global system makes implausible the survivability of regional pockets of self-sufficiency. The only viable option to secure the continuity of urban civilization in this century is a system-wide turn to a broad portfolio of renewable energy sources based on an overwhelming availability of solar, wind, wave, hydrogen-based, and other renewable forms of energy generation, combined with massive conservation and efficiency innovations.

3. INNOVATION IN URBAN RENEWABLE ENERGY TECHNOLOGY

Most contemporary fuel sources are due to expire well within the 21st century as widely available and predominantly peacefully contested sources, and much of this reality will become globally pervasive within these next decades. Even conservative industry, national, and international governmental sources estimate that oil will expire by 2050—the depletion of the more easily accessible supply sources is likely to take place already in the 2020–2030 time frame. Given rising use rate scenarios, natural gas is likely to evaporate as a commercially available source by 2040. The oldest and most problematic source of fossil fuel supply, coal, is expected to expire as a globally available source before 2100—provided no large-scale efforts of energy fuel substitution are pursued to stretch its deposits.

Uranium supplies are also estimated to reach depletion levels by the mid-2030s, even if the daunting cost, security, safety, disposal, and public acceptance dilemmas did not weigh so profoundly against their use. Significant growth rates in uranium usage are not anticipated, given the risks involved and a waning willingness among many governments to pursue nuclear fission further. Nuclear fusion, more than half a century after its arrival as a great modern dream, still struggles with enormous conceptual and technical flaws, and the problems associated with huge financial, environmental, and political costs. These include the spectre of broad public resistance to a new generation of nuclear super-reactors.

This outlook is as troubling as it is underemphasized in its public significance. Fossil fuels are the lifeblood of modern economies and the foundation of the petrochemical, transport, and industrial production systems. Hence they also sustain the spatial centers of global civilization: cities and city regions. The paradox is perplexing but can be explained by the daunting magnitude of the problem. While the modern city in its history, form, and growth dynamics is inconceivable without these fuels, a persuasive agenda to deal with the prospects equitably, humanely, and positively still eludes global and local planners alike.

Yet the most optimistic plans about the powering of urban areas describe entire cities as net renewable energy producers. This idea requires a rethinking of

urban-regional alliances as well as an adoption of increasingly firm industry promotion practices. The Australian city of Melbourne, as an example, is in the process of investing in renewable energy producers with the dual aim to reduce its fossil fuel dependency and to promote the development of more advanced industries capable of competing nationally as well as internationally.

In a technologically advanced renewable economy energy supplies no longer exclusively depend on large, centralized supply models but can unfold in a more diverse and differentiated manner, in keeping with the principles of convergence, blurring the conventional distinctions between different technologies and production and consumption. Traditional energy users become net energy generators, for instance, through solar systems or zero-emission, renewable-source based hydrogen fuel cells in vehicles capable of powering homes and external machinery.

Systems convergence dynamics also point to a merger of information technology, telecommunications, and energy systems. While some electric utilities already lease their grids for information transmission purposes, emerging technology goes much further: future energy systems are ubiquitous and pervasive, and networked into virtual power plants. Ubiquitous energy management connotes the hopeful notion that in a renewable-energy based economy a myriad of small and medium-scale providers of energy services could replace the system of large-scale centralized ones. In an urban system this could operate both at the high end described and at low levels of technological sophistication.

Possible high-technology directions contain features derived from information technology and telecommunications mergers, or the blending of these new technologies with energy production and consumption modes residing everywhere, in applications that range from personal equipment to cars and facilities. At the lower end of the scale, in a distributed yet low-cost and low-maintenance environment, small hydropower and solar systems are seen to be capable of leveraging access to the global information network for small remote communities in developing nations.

4. THE URBAN RENEWABLE ENERGY PORTFOLIO

A number of renewable energy technologies and related approaches and systems is available to supplement and even supplant conventional fuel sources in cities. As a main distinguishing feature when compared to more conventional systems (fossil, nuclear, large hydro, municipal waste incineration), they all vary greatly in scope, style, and systems features, and their economic and physical attributes varies too, including their visual attractiveness. Together, they present an urban power paradigm shift, in the move from the large and passively sourced supply networks to a much wider and dynamic array, or portfolio of systems, many of which are in the control of individual users.

Because the legislative, practical, and financial energy supply structure in a number of countries is not traditionally geared to cope with such simple systems as decentralized small cogeneration plants or grid-feeding domestic solar photovoltaic or solar thermal power plants, significant challenges lie ahead, waiting to be overcome. Fortunately, numerous urban applications are in place, largely in Europe and the United States, ready to be used as best-practice examples.

Buildings, facilities, infrastructures, and even cultural expectations associated with the city are intrinsically linked to and based on the industrial paradigm engendered by fossil fuels. Energy efficiency and conservation have an extraordinarily powerful role to play in reducing the historical dependency of cities on massive fossil fuel input. Ambient temperature stabilizing landscape corridors, native garden-based suburbs, vegetation-covered roofs and walls, natively landscaped atriums, urban agriculture, and better water management all contribute to reducing urban energy input requirements, in this case for cooling.

In supplementation of efficiency and conservation measures, the available power-generation techniques in cities include those shown in the Table I. Many sources listed under Bio/waste are not strictly renewable or environmentally sustainable in their use, especially those relying on waste incineration. They are nevertheless included because many countries do account for them in their renewable, sustainable, new, or alternative energy statistics, as does the International Energy Agency (IEA). Unfortunately, these de facto nonrenewable and carbon emitting sources represent the lion's share of many statistically declared sustainable energy components. New, more efficient fossil fuels systems such as advanced coal and gas applications are excluded altogether.

5. CLIMATE STABILITY IN URBAN DEVELOPMENT

Worldwide, municipal policies are marked by a slow but steady growth in commitment to planned urban

TABLE I

Renewable Power-Generation Techniques and Applications Relevant to Urban Systems

Solar	Photovoltaics (PV)	PV panel and array systems	Building/infrastructure
		PV concentrator	Industrial/commercial
		Building-integrated photovoltaics	
	Solar-thermal	Parabolic trough concentrator	
		CLFR technology[a]	
		Parabolic dish/sterling concentrator	
		Power tower/MTSA[b]	
		Solar chimney	
	Solar water heater		Buildings/facilities
	Solar pond		Parks
Wind	Turbines	Large turbine	Grid/region
		Offshore wind farm	
		Small turbines/vertical axis turbine	
Bioenergy	Biofuels/wood	Black liquor	Power stations: Grid
		Forestry crops	
		Forestry residues	
		Timber/paper industrial wastes	
	Biofuel/agro	Agricultural crops	
		Agricultural residues	
		Bagasse	
		Animal waste	
		Food industrial waste	
	Biofuel/municipal	Municipal solid waste	
		Sewage	
		Landfill gas	
Hydropower	Hydro-electricity	Conventional hydropower	Grid/rivers/streams
		Mini and micro hydropower	
Geothermal		Large geothermal	Grid
		Heat pump	Distributed
Ocean		Ocean thermal	Grid
		Tidal power	
		Wave energy	
	Hydrogen	Combustion	Transport
		Fuel cell	Distributed (heat/power/transport)

Assembled by Hiroyuki Akagawa, Technical Research Institute Obayashi Corporation, University of Sydney (2002).
[a]CLFR, compact linear solar reflector.
[b]MTSA, multitower solar array.

energy practice. As global agreements promise to become firmer under mounting pressures to resolve the greenhouse challenge, and as nations and regional policy units such as the European Union begin to establish clearer greenhouse, energy, and environment frameworks, local institutional pressure mounts to arrive at unprecedented arrangements in urban emissions and energy practice.

As cities are increasingly seen as global emissions performers, distributed energy managers and potential credit trading entities, a new science and engineering practice is on the rise that interprets cities and city regions as power systems, resource flow units, and point-of-emission locations. The best-known urban environmental and community planning models are based on relative improvement targets and selective means of accounting for greenhouse gas (GHG) emissions, setting the stage for the phasing in of renewable energy systems on an urban scale.

A number of place-based emissions allocation techniques and action approaches are in use. These

include the United States Environmental Protection Agency's support made available to states in compiling GHG inventories, in producing action plans, and in staging demonstration projects. They also include the emissions approximation system developed for the Cities for Climate Protection (CCP) program operated by the International Council for Local Environmental Initiatives (ICLEI). Another accounting approach based on the behavior of large geographical units—one degree of longitude by one degree of latitude—is a program called Global Change in Local Places developed by the Association of American Geographers. There is the potential to apply aspects of the Advanced Local Energy Planning approach developed under the International Energy Agency's (IEA) Building and Community Systems program, or the physical model of the economy developed by the Australian Commonwealth Science, Industry and Research Organization. Another ambitious model is a total-flow, end-consumption, and value-based accounting system advanced at the University of Sydney. It also aims to compare, evaluate, and advance a range of methods.

While most GHG emission sources and mitigation efforts are inherently local, their effects and most easy modes of measurement operate globally. However, the identification of globally diffused GHG levels carries little practical local meaning. For cities to become active and integral participants of any global action program, they need to opt for one or the other GHG accounting or allocation method that reliably links local practice to global aims. A persuasive argument holds that only systems reliably measuring contributions to climate stability are valid and more specifically approaches that embrace a 3.3-ton of a 1990 carbon dioxide-equivalent target by 2050. This target is based on a fundamental equity calculation: that on a per capita basis each person has only an annual 3.3 ton emissions allowance, if oceans and forests are to neutralize excessive carbon emissions. In contrast, Australia and the United States approach 30 tons per annum per person, while most developing countries, including India and China, lie well below this sustainable 3.3 t/p/a level.

An ideal method has not yet been found, and in pursuing it, cities need to apply certain performance criteria. To be informative, universally acceptable and suitable as a basis for coordinated climate stable practice such model would need to be comprehensive, precise, and accurate. To be useful it also has to be pragmatic, affordable, and easily replicable across a large number of communities or geographic units. Another challenge lies in avoiding a new information poverty barrier. It could develop were local carbon trading to be introduced as based on highly sophisticated and expensive forms of data gathering—poorer nations have less capacity to cope with the technical challenges involved in measuring emissions performance, let alone reliably monitor its progress over time.

The single most widely promoted method is that of ICLEI's CCP program. It is introduced by participating local governments as the first of CCP's five steps to better GHG performance: namely the arrival at an energy and emissions inventory, forecast, reduction target, target achievement plan, and active implementation program to achieve measurable GHG reductions. At these relatively early days of the CCP program measurable success in reaching significant implementation goals has been elusive, while the other milestones have been reached to lesser and varying degrees. It is hoped that in future rounds of development, in association with other methods and given a better understanding of the key barriers to success, a breakthrough in implementation can be achieved. The very imperfections of the method and its omission of significant emission sources makes the CCP system attractive in practical terms but perceived as scientifically ambiguous in globally credible accounting contexts, and of limited efficacy for local improvement monitoring over time, when absolute performance is to be measured.

There are a number of obstacles on the path to a transformed urban energy practice, although none is insurmountable in principle. The most formidable is the fact that local governance structures are not usually geared to end-state or long-range planning, hence the adoption of long-term accounting methods is often resisted. The effective allocation of incentives and accountabilities is difficult too. Notwithstanding the efforts described earlier, there is the persistent question of how to convincingly disentangle resource flows in a local government area not only for carbon accounting but also for scenario modeling purposes, and how to identify agency accountabilities for savings and reduction.

Despite these challenges, the very consideration of broad energy and greenhouse action programs opens up new dimensions for government. It can provide a new development perspective and drive regional realignments, the formation of state initiatives and industrial alliances. It can mean a new sense of empowerment and the opportunity for economic strengthening as well as engender competitive advantages of a more promotional nature.

Governments discover that energy responses can be scaled well to local development paths and growth options. In newly industrializing countries with interest in maintaining considerable levels of autonomy such as China and India, there is a good fit with local traditions and climatic and economic conditions. Local industry can get stimulated while environmental policies find a practical agenda. Comprehensive renewable energy practice on the local level directly reinforces broad quality-of-life aims, as well as healthy water, clean food, and sustained biodiversity levels.

Besides the aim of responsibility devolution in pursuit of global greenhouse objectives, another government reform agenda item is being satisfied as well: better accountability in public service. Outcome-geared reform and strengthening of local governance in delivering on performance commitments could well be a direct outcome of an indexed and measurable sustainable energy practice. Local consensus, too, is expected to be forthcoming more easily as agendas become more clearly understood and made visible in tangible improvements.

6. URBAN RENEWABLE ENERGY POLICY AND PRACTICE EXAMPLES

In the absence of useful and established patterns of practice, a search is under way for new means of reconciling local government's sectorial concerns, technological opportunities, and shifts in energy markets with global environmental imperatives. There are several hopeful initiatives, focusing squarely on comprehensive local action in urban and regional conversion to a renewable energy supply base.

6.1 Urban Programs: Barcelona's Solar Ordinance

An increasing number of cities search for paths to energy sustainability. One of the most progressive of these is Barcelona. The city has calculated its very favorable solar energy supply potential and in 2000 introduced a regulatory system requiring households and industrial users to provide at least 60% of their hot water requirement utilizing solar systems. The scheme is focused, specific, and practical and based on a precise accounting of domestic energy production and utilization profiles.

6.2 Professional Practice: The European Charter for Solar Energy in Architecture and City Planning

The charter, signed by 30 eminent architects and developed by these under the leadership of German architect Thomas Herzog, was released in 1996, at the Fourth European Conference Solar Energy in Architecture and City Planning, chaired by then minister for construction Klaus Töpfer and parliamentarian Hermann Scheer. The charter was advanced in historical contradistinction to the fossil-age Charter of Athens, referred to earlier. It is comprehensive in attempting to deal with most physical planning challenges in an energy-conscious manner. The charter focuses on the role of planners, the building site, the design and construction process, buildings in use, and the city as a sustainable planning challenge.

6.3 Regional Policy: The Eurosolar Guide to State Politicians—A State Program for Renewable Energy

The renewable energy technology and policy organization Eurosolar has in 2001 issued a blueprint program, for the use by regional politicians in the implementation of renewable energy. It has ten focal areas or considerations: (1) the role model played by state government, (2) the role of state energy agencies, (3) the strengthening of municipal power companies and the placement of energy supply into community hands, (4) research and education, (5) renewable energy in regional development, (6) strengthening of agriculture, (7) the orientation of land-use planning and building toward renewable energy, (8) transport and traffic systems, (9) investment and finance, and (10) state government awards for the exemplary use of renewable energy.

6.4 City Guidance: The World Council for Renewable Energy's Solar City Program

The Solar City approach originally emerged from a new generation of International Energy Agency (IEA) research and development work to pursue citywide applications as integral to the main planning agenda. It is now adopted by the World Council for Renewable Energy (WCRE) as a basic policy and planning basis for cities. It recommends that renewable energy technologies and other means of greenhouse gas

emissions reduction and absorption should be applied in a coherent spatial and social context, as well as within a finite and community-wide time frame.

For application by local institutions, the focus is on the energy supply and technology side, within a comprehensive town planning and design strategy that includes institutional arrangements. The goal is to develop energy technology applications and emissions accounting systems along with performance targets linked to urban development and reform initiatives. Also important are strategies that are based on a consideration of urban-rural linkages and value land use and transport investment choices with a view to their contribution to long-range energy and resource self-sufficiency goals.

The Solar City planning approach presents five areas of focus to cities and city-regions, to be advanced simultaneously. This recommended program constitutes a well-rounded framework for transforming any city's energy regime in a systematic and comprehensive manner.

6.4.1 Solar City Strategies

It is recommended to identify local planning and development approaches that are conducive to the introduction of renewable energy technologies, within a broadly energy-conscious community development approach. To be addressed are strategy, planning tools, organizational arrangements, legislation and standards, incentive structures, public information, and exemplary municipal practice.

By introducing improved ways of adopting renewable energy technologies, a municipal program is to be designed to contribute to climate-stable practice in the building and property development industry, land-use planning, and infrastructure development. It should be framed in ways that help strengthen local governments' efforts to build enlightened community performance and household preferences.

Cities and towns can pursue renewable energy programs through the following methods:

- Direct legislation and standards
- The provision of incentives and disincentives
- Corporate capital asset practice, power purchasing, and pricing
- Institutional reform and improved strategic and general planning practices
- Community action, industry alliances, information, and education

Guidance provided to participating cities is to investigate each of these in detail and develop advanced means of building improved urban practice approaches.

6.4.2 Setting Targets

The aim is to introduce, enhance, and evaluate suitable approaches that help explain the role of renewable energy technologies in the broader urban energy context. It is important to include both renewable energy technology implementation targets as well as absolute climate-stable carbon dioxide-equivalent emissions measures aiming at a specific future date.

Planning methods based on energy technology introduction and emissions accounting methods may deploy backcasting approaches. This involves the development of alternative urban development growth and technology transformation scenarios, then establishing milestones for technology innovation and emissions to determine workable reduction rates over time.

6.4.3 Urban Renewable Energy Technology, Systems, and Industry Development

The aim is to work with cities in advancing renewable energy technologies and systems and to help promote the renewable energy industry in a way that allows it to serve as model for the rest of the national urban system. The emphasis is therefore on market-led approaches of technology system development and deployment, through pricing, investment, electricity purchasing policies, information, model action, and other means.

It is recommended that optional paths are developed, evaluated, and implemented, suitable for the informed and broad introduction of renewable energy technology portfolios and the use by city governments, municipal utilities, businesses, industries, and households. Special emphasis is to be placed on microgeneration and distributed low-energy production in buildings, facilities, and urban infrastructures. Emerging and potentially competing solar and other renewable energy systems and related urban services are to be assessed for their systematic introduction in ways that are meaningful to cities' development agendas—physical planning, sustainability objectives, organization, services—and their value in targeted emissions reductions.

Outcomes are expected to include technology, systems, and industry development options, suitable for selective and targeted implementation in general and specific action plans—answering the following questions: What can city governments in collaboration with industry and constituent urban

communities do to advance the direct use of renewable energy sources? How can these be useful to residential neighborhoods, industry, and transport? How can they support the generation of electricity in quantity, through solar, wind, biomass, geothermal, and sustainable hydroelectric power? And how can cities as organizations contribute to the development and deployment of industrial and residential application, such as stand-alone power generators, heat pumps, photovoltaics, solar hot water production, or solar cooling?

6.4.4 Understanding Best Practice

The recommendation is to make accessible and apply useful lessons from related initiatives domestically and worldwide. This is to be achieved by studying successful practice in integrated urban energy planning, management, and projects. The activities include an identification of scope and criteria for evaluation, information gathering and documentation, study and evaluation, analysis and description, case study development, and communication and dissemination. The scope encompasses technologies, management practices, and growth strategies. Several main categories of case studies are to be differentiated: comparable cities, urban precincts, and settlement projects but also development policies and programs.

6.4.5 Learning from Action

This important final recommendation is to creatively reflect on program experience derived from the application of such programs. This will help develop a shared understanding of the barriers to community, institutional, industrial, and technological change, with a view toward the planned phasing in of solar and other renewable energy sources on an urban and regional scale.

This structured approach will not only be useful to any city engaged in a renewable energy implementation program, but also be of value in the application of lessons and methods across the national urban system.

SEE ALSO THE FOLLOWING ARTICLES

Biomass for Renewable Energy and Fuels • City Planning and Energy Use • Computer Modeling of Renewable Power Systems • Public Reaction to Renewable Energy Sources and Systems • Renewable Energy in Europe • Renewable Energy in Southern Africa • Renewable Energy in the United States • Renewable Energy Policies and Barriers • Renewable Energy, Taxonomic Overview • Renewable Portfolio Standard • Urbanization and Energy

Further Reading

Byrne, J., Wang, Y.-D., Lee, H., and Kim, J.-D. (1998). An equity- and sustainability-based policy response to global climate change. *Energy Pol.* **26**(4), 335–343.

Calthorpe, P. (1993). "The Next American Metropolis: Ecology, Community, and the American Dream." Princeton Architectural Press, New York.

Capello, R., Nijkamp, P., and Pepping, P. (1999). "Sustainable cities and energy policies." Springer, Berlin.

Gideon, S. (1948). "Mechanization Takes Command: A Contribution to Anonymous History." Oxford University Press, New York, NY.

Girardet, H. (1999). Creating sustainable cities. *In* "Schumacher Briefings: 2." Green Books, Devon.

Howard, E. (1902). "Garden Cities of Tomorrow." Swan Sonnenschein, London.

International Energy Agency (2002). "Renewable Energy Almanac." IEA, Paris.

Kates, R. W., Mayfield, M. W., Torrie, R. D., and Witcher, B. (1998). Methods for estimating greenhouse gases from local places. *Local Environ.* **3**(3).

Lenzen, M. (1997). Energy and greenhouse gas costs of living for Australia during 1993/4. *Energy* **23**(6), 487–586.

Mumford, L. (1961). "The City in History." Secker and Warburg, London.

Newman, P. W. G., and Kenworthy, J. R. (1987). "Gasoline Consumption and Cities: A Comparison of U.S. Cities with a Global Survey and Some Implications." Murdoch University, Murdoch, WA.

Organization for Economic Cooperation and Development (1995). "Urban Energy Handbook: Good Local Practice." OECD Publication and Information Centre, Paris.

Roulet, C.-A. (2001). Solar energy and the global heat balance of a city. *Solar Energy* **70,** 255–261.

Sartogo, F. (1998). Town planning and the use of renewable energies near Rome. *Renewable Energy* **15,** 395–400.

Scheer, H. (2002). "The Solar Economy. Renewable Energy for a Sustainable Global Future." Earthscan, London.

Steemers, K. (2003). Energy and the city: Density, buildings and transport. *Energy Bldgs.* **35,** 3–14.

Waals, J. F. M. v. d. (2000). The compact city and the environment: A review. *Tijdschrift voor Economische en Sociale Geografic* **91,** 111–121.

Renewable Energy in Europe

JOACHIM NITSCH, WOLFRAM KREWITT, and OLE LANGNISS
German Aerospace Center
Stuttgart, Germany

1. **Introduction**
2. **Present Use of Renewable Energy Sources in Europe**
3. **Policies to Foster Renewable Energies in Europe**
4. **Potentials, Costs, and Environmental Characteristics of Representative Technologies**
5. **Prospects of RESs in Europe**

Glossary

biomass All organic substances that can be used for energy production, including industrial, commercial, and agricultural wood and plant residues; municipal organic waste; animal manure; and crops directly produced for energy purposes. Biomass is available as solids (wood, straw), liquids (biofuels), and gases (biogases).

EU-15 The 15-member European Union countries (Western Europe is used as a synonym for the EU-15 extended by Switzerland, Norway, and Iceland).

final energy The energy used by consumers in final form: electricity, heating fuels, and transportation fuels.

geothermal energy Natural underground heat used for heating purposes either directly or by means of heat pumps (near-surface heat and hot underground water) and for electricity generation (heat extraction from hot rock).

German Renewable Energy Resources Act An act passed in March 2000 that obliges grid operators to buy electricity generated from renewable energy sources at a given price (feed-in model). The price, which depends on the technology and the year the power-generating installations were built, decreases for installations built subsequent to implementation of the act. The overall costs are passed on equally to all final consumers of the electricity.

greenhouse gases Airborne gases contributing to climate change, including water vapor, carbon dioxide, methane, nitrous oxide, and synthetic, volatile chemicals (e.g., tetrachloro- and tetrafluoromethane, sulfur hexafluoride, and hydrofluoroalkanes). The most important contribution to climate change comes from carbon dioxide.

primary energy The energy directly embodied in natural resources, including fossil fuels (coal, lignite, crude oil, and natural gas), nuclear energy (in the form of heat produced during fission), and renewable energy sources.

renewable energy sources All natural energy flows that are inexhaustible (i.e., renewable) from an anthropogenic point of view: solar radiation; hydropower; wind, geothermal, wave, and tidal energy; and biomass.

1. INTRODUCTION

There are various and somewhat complementary reasons to foster the growth of renewable energy sources (RESs) in Europe. A major incentive for RES policies in the past two decades has been to reduce the environmental impact of energy use both locally (e.g., pollutant emission reduction) and globally (e.g., greenhouse gas reduction). In some countries, concerns about the safety of nuclear power generation have motivated the search for RESs. Another motivation for replacing foreign fossil and nuclear fuels with domestic RESs relates to security issues and Europe's growing dependency on foreign energy sources. In 1999, half of the energy demand of countries in the 15-member European Union (EU-15) was met by imported energy, and this share is expected to increase to 70% within the next two to three decades. In particular, the dependence on oil could reach 90% by 2020. Further motivating studies on developing RESs are the social and economic benefits, such as the creation of jobs, the creation of new industries with good export prospects, and the positive structural effects on regional economies. The characteristics of RES development lead to public benefits in all of these areas. Additionally, the growing integration of Europe

 © 2004 Elsevier Inc. All rights reserved.

continues to highlight the importance of the future development of RESs.

Currently, individual national goals and political measures substantially determine the actual growth of RES industries. In addition to national goals for RES development, the EU Parliament and the EU Commission have defined broader goals for the contribution of RESs by 2010. Other European long-term energy scenarios, from 2030 to 2050, also deal with RESs, and some call for future energy systems based on significantly high contributions from RESs. In addition, the EU affects the promotion of national RES programs within a legal framework. The Commission is anxious to coordinate national promotion efforts, which currently vary greatly.

Markets for RESs are increasingly being established throughout Europe; the dynamically growing wind power industry has been particularly active in this regard. It is foreseeable that a uniform strategy will be developed within this decade for further promoting and implementing mid- and long-term goals for RES technologies in Europe. This process, which was an essential point on the agenda of the Johannesburg World Summit for Sustainable Development in 2002, is presently supported by a growing awareness of the importance of increasing the development of RES technology worldwide. In summary, there is every indication that Western Europe will strengthen its position as the world's leading region in utilization and commercialization of RESs in the next few years.

2. PRESENT USE OF RENEWABLE ENERGY SOURCES IN EUROPE

2.1 Energy Sources

In Europe, modern RES technologies were explored thoroughly for the first time after the oil embargo/price crisis of 1973. Notably, market introduction of renewable energy technology started in about 1985, but the RES industry has become vital only during the past decade. Accordingly, relevant statistical RES data have been systematically compiled only over the past decade, although reliable and consistent statistical RES data, collected since 1989, exist for all 15 countries of the European Union and for Western Europe (defined here as the EU-15 plus Switzerland, Norway, and Iceland). Earlier data from various European countries are fragmentary, inconsistent, and less reliable. Most energy source data are available up to 2001; the country-specific data have been compiled up to 1999/2000. Only a few eastern European countries have provided RES data, and their RES contributions today are still very small.

Of renewable energy sources presently utilized in Europe, as well as in other continents, two dominate: hydropower and firewood (for traditional heating purposes). Both resources were the main basis of human energy supply before industrialization, and even with the emergence of fossil and nuclear energies, both have retained importance. In 1970, hydropower met 30% of European electricity demand; by 2000, this had decreased to 17%. The current degree of exploitation of hydropower and firewood today is considered high, thus the contribution of these two energy sources cannot be increased considerably in the future. Primary energy production from RESs in Europe has risen continuously since 1990 (Fig. 1). Within the past decade, production increased by 40%, accounting for 6.4% of total primary energy production in 2001. The current RES production of 3850 PJ/year (a petajoule is 1000 trillion, or 10^{15}, joules) is provided mainly by biomass (62%) and by hydropower (31%). Geothermal energy (4%) and wind energy (3%) contribute far less. The usable final energy production from RESs amounts to 3350 PJ/year, corresponding to 8.5% of the total final energy consumption in the EU-15 countries. Vehicle fuels contribute only 30 PJ/year; usable RES heat (1880 PJ/year) meets 9.5% of total heat demand, and RES electricity (1460 PJ/year $\approx$ 406 TWh/year; 1 terawatt-hour = 1×10^{15} kWh) comprises 15% of total electricity generation (Fig. 2). About 80% of this electricity comes from hydropower. The use of biomass (10%) occurs predominantly in wood-fueled power plants (60%), although biogas and sewage gas (10%) and municipal waste (30%) also contribute significantly. The contribution of wind energy has grown considerably, to 6.6%, and geothermal energy contributes 1.2%; solar electricity presently contributes an insignificant amount (0.1%). The dominance of biomass in heat production is notable. Individual firewood heating contributes about 60%, and another 20% comes from wood- and waste-fired district heating systems; industrial wood-fired heating systems contribute 18% and biogas contributes 2%. Solar and geothermal energy contribute only small amounts to heating (about 1% each).

2.2 Use Different Countries

Renewable energy sources contribute to the energy supply in individual EU member states in quite

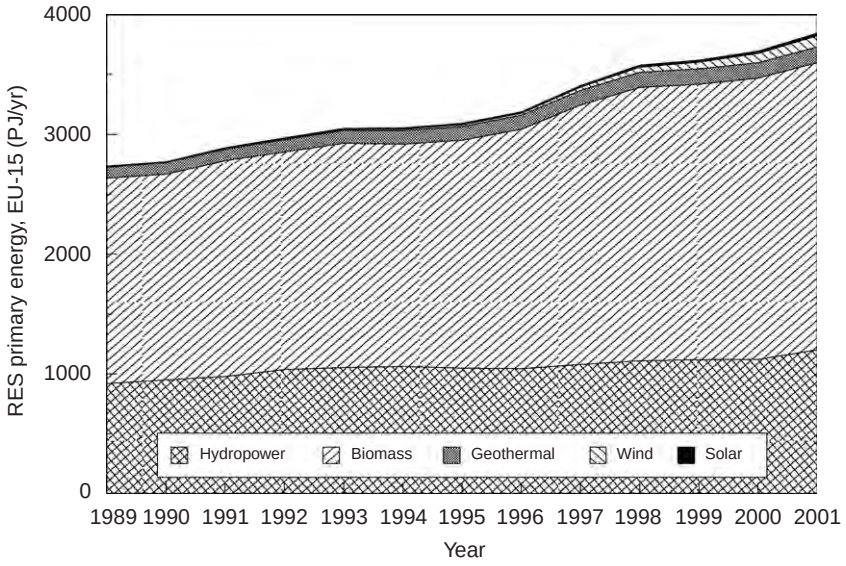


FIGURE 1 Primary energy production using different renewable energy sources since 1989 in the member countries of the European Union.

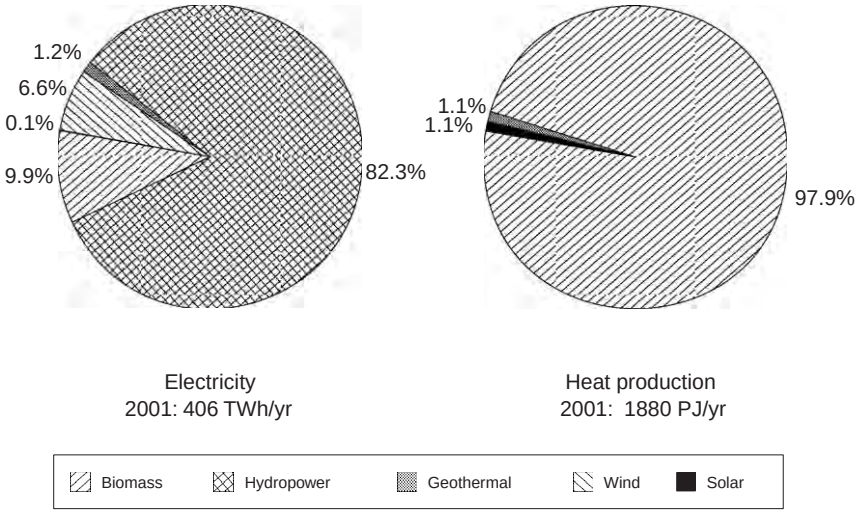


FIGURE 2 Contributions of renewable energy sources to final energy production in the EU-15 for the year 2001.

different proportions (Figs. 3 and 4). This is to be expected because the actual RES contribution is determined by a number of different parameters:

- Country-specific potential for each type of RES.
- Existence of other low-cost and abundant domestic energy resources (e.g., coal, natural gas).
- Structural boundary conditions, such as population density, structure of settlements, ratio of rural to urban areas, extension of electricity grids, and district heating networks.
- Economic conditions, such as national energy prices and taxes.
- Level of economic wealth and of total and per capita energy consumption.
- Conditions of energy policy, such as incentives and instruments to promote RES technology and national energy goals.

Renewable energy resources contribute to meeting a relatively high proportion of energy requirements in countries with large land areas, correspondingly low population density, and abundant hydropower.

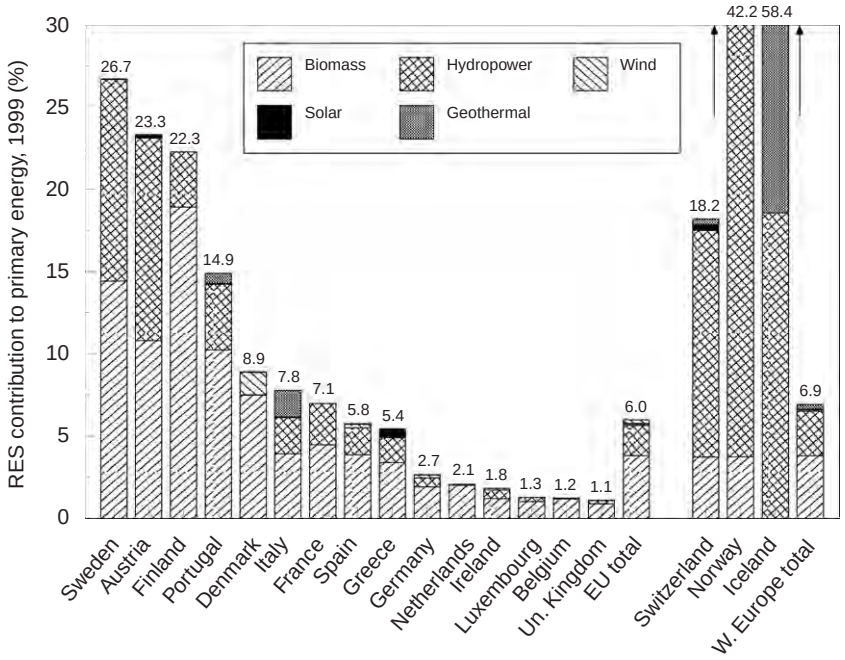


FIGURE 3 Contributions of renewable energy sources to primary energy consumption in the EU member countries and Switzerland, Norway, and Iceland in 1999.

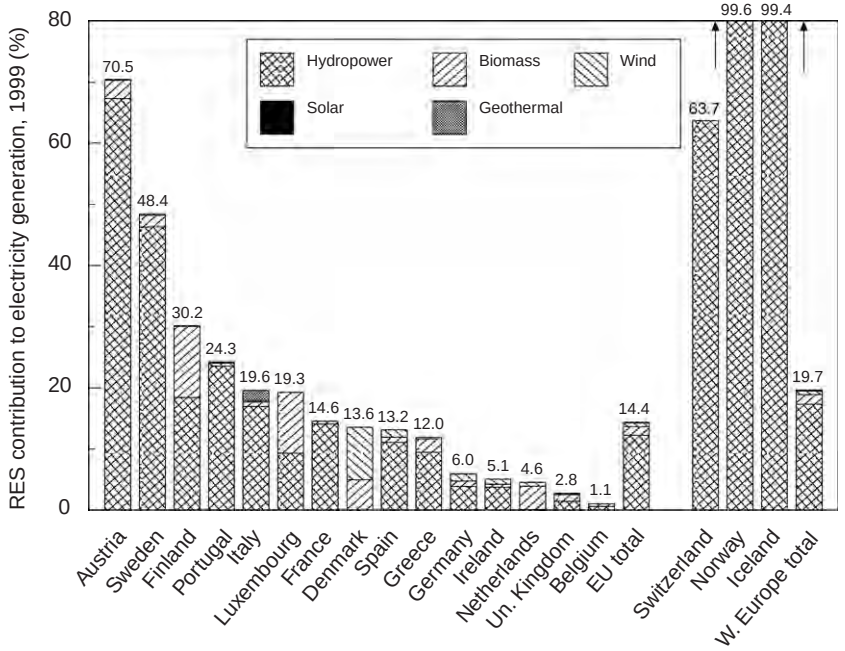


FIGURE 4 Contributions of renewable energy sources to electricity generation in the EU member countries and Switzerland, Norway, and Iceland in 1999.

Sweden, Finland, and Austria fit this profile, with RESs contributing more than 20% to primary energy consumption and electricity generation. In densely populated countries, RESs contribute less than 10%—2.7% in Germany, 2.1% in The Netherlands, and only 1.1% in the United Kingdom. On the other hand, there are strong disparities among countries with similar RES potentials and comparable

infrastructures. For example, in densely populated Denmark, RESs contribute 9% to primary energy consumption and 14% to electricity generation, showing the importance of well-defined political goals and ambitious promotion measures. Such goals and measures, however, primarily affect development of the "new" RES technologies (wind energy in Denmark). RES market development has seen little progress in the United Kingdom.

The contribution of hydropower and geothermal energy increases considerably with the inclusion of Switzerland, Norway, and Iceland, which are geographically, economically, and culturally part of Western Europe (Figs. 3 and 4). RESs in Norway and Iceland provide about 50% of the primary consumption and 100% of the electricity; in Switzerland, these values are 18 and 64%, respectively. Thus RESs accounted for 7% of primary energy consumption and 20% of electricity generation in 1999 in Western Europe. The estimated increase for 2001 is 0.5%. The five countries with the highest amount of RES primary energy are France, Italy, Sweden, Norway, and Germany, which together account for 62% of total RES primary energy production. The highest amount of RES electricity is generated in Norway, followed by France, Sweden, Italy, and Austria; these five countries account for 68% of total RES electricity production.

2.3 Technologies

The development of the "new" wind and solar technologies is of great importance for the future contribution of RESs to energy supply. Although the present wind and solar technology contribution of 0.4% to total primary energy consumption is still very small, the growth of these industries has been considerable in the past 6 to 8 years. Today they provide 10 times the energy of 10 years ago. Wind energy shows the most remarkable growth dynamics; its contribution is now reaching energetically relevant dimensions (Fig. 5). Since 1992, the average growth rate has been 35%/year. At the end of 2001, the installed capacity in Western Europe reached 17,250 MW, about 70% of the total global wind power; 1% of the Western European electricity demand is generated in this way. With 8750 MW, Germany alone accounts for 50% of the installed capacity; Spain, with 3300 MW, follows and is rapidly increasing wind installations. Spain has surpassed the pioneering wind country, Denmark, where growth decreased strongly because of a recent change in energy policy priorities. Yet, with 9%, Denmark still has the highest share of wind power for total electricity generation.

Photovoltaics also showed similar high growth rates in the past 8 years (Fig. 6), but with an installed power of 300 MW in Western Europe (280 MW in EU-15), the contribution of photovoltaics to electricity generation (0.01%) is still negligible. Even more noticeable than in the case of wind energy, the growth rate of photovoltaics is dominated by development in Germany, which accounts for 60% of the total installed power in Europe. This is the result of the very favorable conditions set up within the framework of the German Renewable Energy Resources Act and the additional "100,000 PV-installations" program. Considerable growth in photovoltaics has also occurred in The Netherlands, which has reached

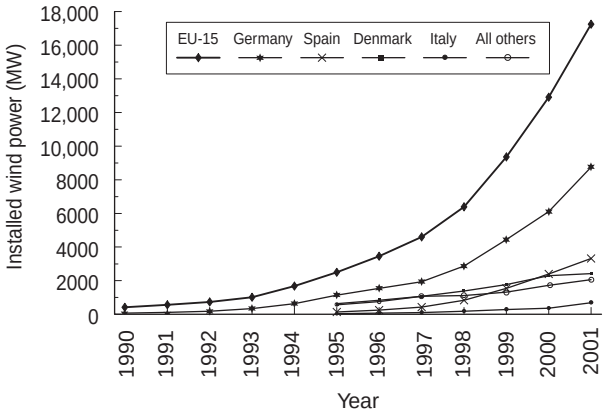


FIGURE 5 Growth of installed wind power output in selected EU member countries and the EU-15 from 1990 to 2001.

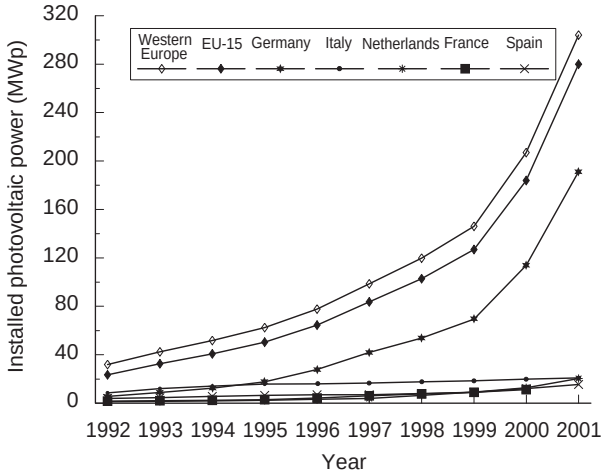


FIGURE 6 Growth of installed photovoltaic power output in selected EU member countries, the EU-15, and Western Europe from 1992 to 2001.

the power level attained by Italy (20 MW). Nevertheless, related to population, Switzerland has the highest density of photovoltaic installations, with 2 W per capita.

There has also been a solid increase in installed solar-thermal collector area, but growth has not occurred as rapidly as has growth of wind power or photovoltaics. In Western Europe, an area of $12.8 \times 10^6\,m^2$ (EU-15 $= 11.9 \times 10^6\,m^2$) supplies 0.01% of the total heat demand (Fig. 7). In Germany, Austria, and Switzerland, where climatic conditions are far less favorable than in southern European countries, the installed solar-thermal collector area is surprisingly high, primarily due to favorable incentives and promotions. In addition, nature conservation and resource protection rank much higher in public opinion in these countries, driving favorable development policy. The reverse of this, public disinterest, may be the reason that, excluding Greece, the contribution of solar-thermal collectors is currently very low in southern European countries, despite the high technical potential. In Greece, solar collectors for water heating are cost competitive against electrical water heaters.

A discussion of promising RES technologies is not complete without mentioning improved biomass conversion technologies such as advanced gasification combined with internal combustion engines, Stirling engines, or fuel cells. These technologies have future potential to convert the abundant supply of biomass more efficiently into electricity and heat than is the case today. Also, improved technologies for digestion of agricultural and municipal waste residues will be of growing importance.

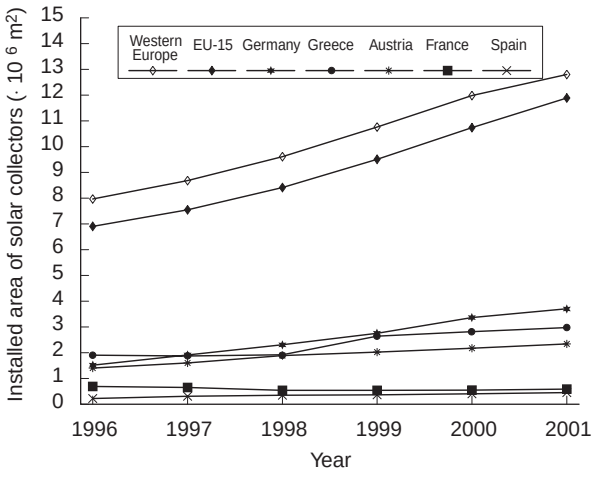


FIGURE 7 Growth of solar-thermal collector area in selected EU member countries, the EU-15, and Western Europe.

3. POLICIES TO FOSTER RENEWABLE ENERGIES IN EUROPE

An important guideline for a well-coordinated RES policy strategy involves setting definite targets, within specific time frames, for RESs to share in meeting the needs of total energy supply. Certainty in policy is a positive signal to potential investors, and attracting private investments is particularly crucial for RES; an accelerated deployment of RES technologies requires considerable funds, usually in excess of funding specified in public budgets. Different approaches are needed to support deployment of RES programs. A comprehensive strategy with a variety of different instruments, rather than a single instrument, is required to address RES development for a range of applications. Moreover, RES policies must address noneconomic barriers, research and development needs, and the cost gap between RES providers and conventional energy providers. Overcoming the renewable/conventional cost gap is a central building block of any RES policy. For this purpose, energy or emission taxes on conventional energy providers can be raised, reflecting the external costs of such fuels. The costs of a RES to the user can be reduced by providing grants, soft loans, or tax benefits. The cost gap can also be overcome by implementing the renewable portfolio standard, i.e., creating a demand for RESs through obliging electricity suppliers to generate a certain percentage of power from renewable sources and offering homeowners and businesses incentives to purchase renewable power. The voluntary purchase of "green" electricity (electricity from RESs), although widely discussed, has not met expectations in Europe so far, and can only supplement, but not substitute for, public support policies.

Establishing RESs via political and regulatory structures is a further important element for a successful RES policy. The present energy supply system and the political and regulatory framework conditions have evolved in a mutual adaptation process. Certain characteristics of RESs, such as their decentralized, dispersed nature, are disadvantaged in many cases by the present regulatory framework . A successful RES policy needs to address such noneconomic barriers for RES. This includes fair access to the electricity grid, adapted building permit procedures for wind power plants, and consideration of solar-thermal collectors in building codes. Workmen, tradesmen, and engineers need to be educated comprehensively about RES technologies and their applications. Schools and universities

need to incorporate RES studies in their curricula. Last but not least, there is an ongoing demand for research and development to exploit the cost reduction potential through technical progress.

Energy policy happens on different levels within the EU, with its institutions, the European Parliament, the European Commission, and the European Council. The promotion of renewable energy sources at every level of the European Union is gaining more and more importance. On the one hand, the general framework conditions for European energy markets are very much targeted toward creating a common market for energy, with equal conditions for all market players across the EU. As an example, the EU issued a directive, "Common Rules for the Internal Market in Electricity," which has created fairer access to the electricity grid for independent power producers using RESs and green electricity suppliers. The European Community Guidelines on State Aid for Environmental Protection allow member states to grant operating support to new RES power plants of up to 5¢/kWh, referring to external costs of conventionally fueled power plants. The European Union is also an active player in promoting RESs directly. In 1997, the European Commission issued a White Paper, "Energy for the Future: Renewable Sources of Energy," setting a goal of doubling the share of RESs for primary energy supply by 2010 and describing scenarios and policy strategies to reach this goal. As a follow-up, a "Campaign to Take-Off" was launched in 1998. The EU's own financial means allocated to this campaign have been rather limited, with $74 million (U.S. dollars) allocated over 5 years compared to the total required investment of $20 billion (U.S.). Another European directive ("On the Promotion of Electricity Produced from RES in the Internal Electricity Market") has set goals for the EU member states regarding the share of RESs for electricity production by 2010. Although the directive failed to establish a common European instrument to foster RESs for electricity generation, it has created some momentum to establish support policies on the national level, as seen recently in the United Kingdom and in Sweden. A political review of all national RES policies is scheduled for the end of 2005, with the goal of creating a basis for a common European support instrument for electricity from RESs.

National support mechanisms remain the most important means to foster development of RES technologies. For example, national minimum price standards (also referred to as fixed feed-in tariffs) have, in particular, promoted wind power in Denmark, Germany, and Spain. Favorable conditions have been created for biomass-fueled district heating in Sweden by imposing high taxes on conventional energy providers and a via CO_2 tax refund. Soft loans and direct investment grants have caused the demand for solar collectors to soar in Germany. Examples of regional and municipal RES policies include the solar ordinance in Barcelona, requiring real estate developers to install solar water heaters, and the green power purchase policy in some Dutch municipalities.

Generally, RES policies have been focused on the electricity sector rather than on transport or heating, mainly because there is a long tradition of state intervention in the electricity sector. Minimum price standards, bidding schemes, and renewable portfolio standards, also referred to as RES quotas or green certificates, have been the major way to support electricity generation from RESs on the national level. Minimum price standards require the grid operator or the default electricity supplier to purchase electricity from RES generators at fixed premium prices. It should be noted that a minimum price standard not only regulates the price, but also regulates grid access and power purchase. Within bidding schemes, RES capacity is publicly tendered periodically and power purchase contracts are awarded to the winning bids. With renewable portfolio standards, electricity suppliers are obliged to generate a certain share of their electricity supply with RESs. The engaged parties comply with the obligation by presenting tradable "green certificates," certifying the generation of a certain amount of electricity. Thus, these certificates have an economic value, generating extra income for RES electricity producers.

Countries with minimum price standards (e.g., Germany since 1991, France since 2001, Spain since 2000, and Denmark until 2000) have seen the largest growth of RES electricity. This particularly applies to wind power. At the same time, a viable RES manufacturing industry has been established in these countries. To organize political support and create local acceptance, it has been proved successful to spread ownership among many, preferably local, people. Even though it is not appropriate to attribute the success in RES deployment solely to a single policy instrument, it is clear that a well-designed minimum price standard together with supplementing policies (such as simplified building permit procedures) seems to be the most effective way to support the introduction of RES electricity. Nevertheless, as indicated by experiences in different countries, proper design of a very specific

support policy is even more crucial than the type of policy.

England and Wales introduced a bidding scheme called Non-Fossil-Fuel Obligation (NFFO) in 1990. In five bidding rounds, between 1990 and 1998, developers of RES plants could bid in different technology slots (e.g., wind power, waste to power, hydropower). The winners with the lowest bid generation costs were awarded with a 15-year power purchase agreement. The bid prices ranged between 45% (hydropower) and 70% (wind power) between the first and the last bidding rounds. Due to different conditions in the bidding procedures and the awarded power purchase agreements, the bids were not directly comparable to each other. Moreover, the large wind project put forward in the last bid round in 1998 has not been commissioned at the low average bid price of 4.5¢/kWh. Apparently, the bid prices were not economically feasible. Although the United Kingdom has the largest wind power potential in Europe, the NFFO failed to foster wind-generated and other RES electricity, perhaps due to inherent weaknesses of the NFFO and also to a lack of appropriate building permit procedures. The wind project has been abandoned.

Renewable portfolio standards have been discussed widely and have been introduced in Austria, Belgium, Italy, Sweden, and the United Kingdom. Although such mechanisms are promising in theory, practical experience has been limited and rather mixed in Europe. Larger providers are more likely than small generators to take the risk of selling electricity and certificates under uncertain conditions. Instead of a wide range of different RES technologies, only the most cost-effective technology will likely be supported at a given time. Long-term contracts rather than spot markets will govern transactions between RES providers and the obliged parties, thus undermining competition. Actually, the different designs of the national renewable portfolio standards hinder rather than enable the free trade of certificates between different countries.

Increasing prices for conventional fuels have been an effective method to deploy biomass and other RES technologies in the heating sector in Northern Europe. Widespread district heating grids support the application of RES technologies in these countries. Household solar collectors have been successfully promoted by tax benefits and direct investment grants. Building regulations allowing only a certain maximum fossil-based heat demand for new building are another effective way to promote RESs. Austria has been extremely successful in deploying solar collectors via grass-roots do-it-yourself construction groups. In some countries, the RES use in the transportation sector is supported by exempting biomass-based car fuels from taxes; this has led to a growth of demand in Germany since 2001.

An overview of promotional systems for RES technology and use in the EU-15 is given in Table I. It is apparent that most countries are using either the feed-in tariff model (minimum price standards) or the certificate trading model (the quota model). Bidding schemes, originally introduced in the United Kingdom, are used only in Ireland. The feed-in tariff model has turned out to be the most successful instrument in terms of installed RES capacity, but an increasing number of countries are considering the certificate trading model as the future winner. Possibly a mixture of both will be used in the future, because green certificates also can be combined with feed-in models.

TABLE I

Overview of Promotional Instruments for RESS in EU-15 Countries (2002)

	Country[b]														
Instrument[a]	**AU**	**BE**	**DK**	**FI**	**FR**	**GE**	**GR**	**IR**	**IT**	**LU**	**NL**	**PO**	**SP**	**SW**	**UK**
FIT	X		(X)		X	X	X			X		X	X		
BID								X							
SUB			(X)	X		(X)	(X)	(X)		(X)	X			(X)	
CTM	(X)	X	Xp						X		(X)			Xp	X

[a] FIT, feed-in tariffs; BID, bidding system; SUB, subsidies, tax reliefs; CTM, certificate trading model; X, main instrument; (X), additional instrument or combination with main instrument; p, proposed.

[b] AU, Austria; BE, Belgium; DK, Denmark; FI, Finland; FR, France; GE, Germany; GR, Greece; IR, Ireland; IT, Italy; LU, Luxembourg; NL, The Netherlands; PO, Portugal; SP, Spain; SW, Sweden; UK, United Kingdom.

4. POTENTIALS, COSTS, AND ENVIRONMENTAL CHARACTERISTICS OF REPRESENTATIVE TECHNOLOGIES

4.1 Technical Potentials

To provide energy for mankind, ambient energy, in terms of annual increases in biomass and the radiation, wind, water, and geothermal flows, has to be transformed into usable forms of energy such as electricity, process heat, or fuels by means of technical installations and processes. Determination of the potential energy and productive capacity of an energy technology requires consideration of a great number of physical, technical, and structural restrictions. To achieve a realistic, practical picture of the possible applications of a renewable energy source, it is necessary to add further boundary conditions, mainly economic and ecologic. In view of the huge amounts of ambient energy flows—e.g., the solar radiation reaching Earth's surface during 3 hours corresponds physically to the total annual global energy consumption—to meet future energy demands, it is sufficient to produce energy carriers only in per million shares (in maximum percent shares).

Numerous analyses of current RES technologies provide a detailed and sophisticated insight into the extent and structure of the technical potentials of RESs. Table II presents these data for Western Europe. The values are explicitly downsized according to infrastructure specifications and ecological restrictions. The comments in the table point out the most important assumptions in this respect. It can be assumed, therefore, that these potentials can be tapped completely as soon as they become economically viable. Compared with present energy costs, the majority of these technical potentials cannot be exploited without public subsidies. The total RES potential for electricity generation is almost double the present electricity generation of Western Europe and the total RES potential for heat production meets 75% of the present heat demand. Altogether, based on these calculated RES technical potentials, 62% of the present primary energy consumption could be met. The different RES technologies are exploited very unequally. Hydropower potential is exploited to a large extent, if the construction of new plants on nonregulated rivers is excluded. Biomass residue utilization also already takes place to a large extent. All other RES technologies are currently exploited to a very small degree. The potentials of RESs are 10 times the present amount of energy that is provided from RESs in Western Europe.

In the long-term, if markets for RES technologies are established, far more renewable energy sources can be made accessible. The following examples of potential market development are from the point of view of Western Europe:

- An extended application of offshore wind converters on European coasts, with an additional electricity generation potential of about 2000 TWh/year.
- Cultivation of energy crops on unused agricultural areas in connection with the planned enlargement of the EU to Eastern European countries, adding an additional 30 million ha (i.e., 3500 PJ/year primary energy equivalent).
- Further exploitation of geothermal resources, yielding additional electricity generation of up to 1700 TWh/year.
- Import of electricity from solar-thermal power plants in North Africa to Western European countries by means of an European–Mediterranean electricity grid, allowing, in principle, the additional mobilization of some 10,000 TWh/year of electricity.

It is obvious that these additional RES potentials are available mainly as electricity. There is no sense in further exploiting RES heat potentials, because heat supply at low temperature (up to 150°C) has to be adjusted to the demand for space heating, warm water, and low-temperature process heat. In contrast, surplus RES electricity can be transformed by electrolysis to hydrogen, which can be used as a fuel for transportation as well as for high-temperature process heat in industry. The above-mentioned additional technical RES potentials add up to more than 80,000 PJ/year of primary energy. Thus, even an increasing energy demand in Western Europe can, in principle, be met in the long term solely by RESs.

4.1.1 Concentrating Solar-Thermal Power Plants: A Representative Innovative RES Technology with Large Potential

Concentrating solar-thermal power plants (CSPs) are able to supply very large amounts of electricity and process heat in the earth's sunbelt. Like any conventional power plant, their integration into the grid does not require new measures for backup capacities. On the contrary, CSPs allow for a smooth transition from today's fossil-based power technology to a future energy economy based essentially on renewables. The history of the solar electricity generating

TABLE II

Technical Potentials of Renewable Energy Sources in Western Europe

Renewable source	Electricity generation (TWh/yr)[a]	Heat production (PJ/yr)[b]	Primary energy (PJ/yr)[b]	Used in 2001 (%)	Comments
Hydropower	640		2270	79	New plants only on regulated rivers; includes reactivation and modernization of existing plants
Wind energy	1580		5680	1.8	Restricted potentials onshore (630 TWh/yr); offshore = 950 TWh/yr; acceptance is high and ecological interaction is as low as possible
Photovoltaics	585		2105	0.03	The assumed 2300-km^2 module area = 25% of suitable roofs and 100% of suitable facades (1000 km^2)
Biomass residues	185	2350	3350	52	Residues from forestry, agriculture, and the organic fraction of municipal waste; for energy crops, the present excess agricultural area in the EU-15 (9 million ha) is assumed; 50% use in cogeneration plants, 50% direct heat production, no fuel production assumed
Energy crops and short-rotation forestry	85	1050	1500		
Thermal collectors		6570	7300	0.2	The assumed 4000-km^2 collector area = 75% of suitable roofs; contains also large collector fields with seasonal heat storage
Geothermal energy	350	5000	9155	1.4	Only known resources (lowest value of a variety of estimates)
Solar thermal power plants	2000		7200	0	Suitable sites for 100 to 200-MW solar thermal power plants in Southern Europe (Spain, Italy, and Greece, including Mediterranean islands)
Total RES potential	5415	14,970	38,560	12	
Consumption, 2001	2800	20,000	62,500		
Share of RES potential (%)	193	75	62		

[a] TWh, Terawatt-hour; 1 TWh = 1×10^{15} kWh.
[b] PJ, Petajoule; 1 PJ = 10^{15} J.

systems (SEGSs) in California shows impressive results. Steam cycle plants with up to 80 MW capacity using parabolic trough collectors have been in commercial operation for more than 15 years. Nine plants with a total of 354 MW of installed power are feeding the California electric grid with 800 GWh/year. The plants have a proved maximum efficiency of 21% for the conversion of direct solar

radiation into grid electricity. Considerable cost reductions have been achieved, with electricity costs ranging today between 10 and 15¢kWh in solar-only mode. Advanced technologies, mass production, economies of scale, and improved operation will allow reductions in solar electricity costs to a competitive level within the next 10 to 15 years, or as soon as a worldwide capacity of 5000 MW is achieved. Hybrid solar and fossil fuel plants, making use of special financial schemes, can already deliver competitively priced electricity in favorable locations.

In sunbelt countries, CSP is able to reduce the consumption of fossil energy resources and the need for energy imports, enabling diversification with a resource that is very accessible. Moreover, process heat from combined generation can be used for seawater desalination, helping to address the challenge of a growing water scarcity in many arid regions. In the medium term, large electricity grids such as the Euro-Mediterranean Power Pool via high-voltage direct current transmission will allow for intercontinental transport of RES electricity (Fig. 8), thus connecting sites with large renewable electricity resources in Europe (wind, hydro, geothermal) and North Africa (solar). This concept may be of outstanding importance for the energy future of Europeans, because it can help to stabilize political and economic relations between the countries of an enlarging European Union, as well as between the EU and the North African countries. Furthermore, it can help by contributing to a considerable reduction of the growing dependence on Middle East oil and gas resources. Primarily, CSP, in combination with other powerful RES technologies, will effectively reduce the risk of conflicts related to energy and water scarcity and to climate change.

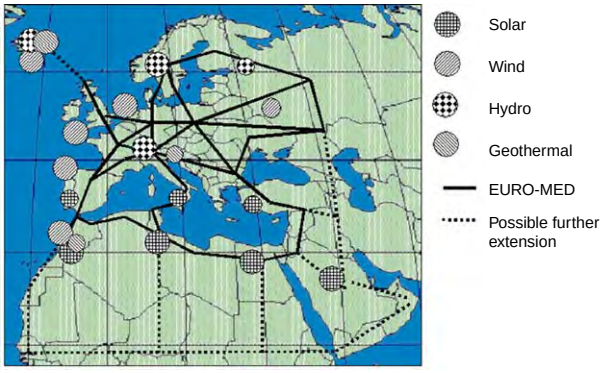


FIGURE 8 Planned extension of European high-voltage electricity grids to a European–Mediterranean (EURO-MED) pool and possible further extension in order to use renewable energy sources widely.

4.2 Costs of RES Technologies

In order to evaluate the chances of a significant commercial RES utilization in the medium term, the costs of energy production, compared to conventional energy supply, are of outstanding importance. The outcome defines which of the RES technologies are economically exploitable under the existing circumstances, and what measures are required to achieve cost competitiveness in a given time frame. It can be assumed that the cost differences between RESs and conventional energies will decrease because of further technical development and growing production rates of RES technologies, and because of an increase in prices of conventional energy sources in the medium term. Thus, a growing share of the technical RES potentials can be exploited economically in the long term. By promoting investments in RESs for a certain time or by influencing the price of energy through taxes or comparable instruments, this process can be accelerated. If future shortages of oil and natural gas or CO_2 sequestration from coal will substantially raise energy price levels, RES providers can offer even lower prices, compared to energy from fossil sources.

A survey of energy generation costs of typical RES technologies in Western European countries, representing the present status and the likely achievable cost reductions in the future, shows a very large spread (Figs. 9 and 10). This results from the great variety of energy sources and their regional differences, the broad spectrum of conversion technologies, the very large range of capacity (from about 1 kW to some 100 MW), as well as different prices for biofuels and for credits in the case of cogeneration plants.

Excluding photovoltaics, for which actual costs are typically 70¢/kWh in Central Europe and 40¢/kWh in Southern Europe, the representative cost of RESs under Western European conditions ranges from 2.5 to 15¢/kWh (Fig. 9). Only a few RES technologies generate competitive electricity today, e.g., older depreciated hydropower plants or cogeneration plants using cheap biomass residues. For that reason, these sources are already widely exploited. A large percentage of RES technologies offer electricity in the range of 7–8¢/kWh, including wind plants in coastal regions, hybrid solar-thermal power plants, and biogas and biomass gasification plants in cogeneration mode. Hence, these RES technologies require further financial subsidies for some time. If market growth of RESs takes place on a large scale, the electricity generation costs of most RES technologies will decline to 4–5¢/kWh. The newest fossil power

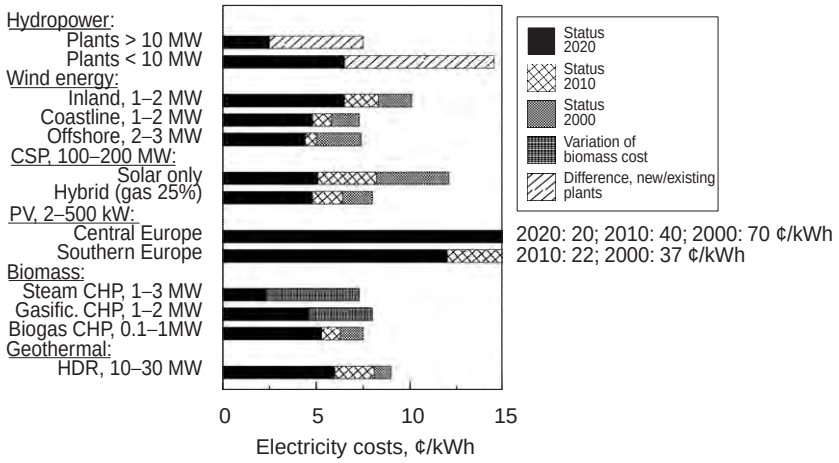


FIGURE 9 Survey of current electricity generation costs (total length of bars) of representative renewable energy source technologies and possible cost reduction until 2010 and 2020 (CSP, concentrating solar power; PV, photovoltaics; CHP, combined heat power; HDR, hot dry rock).

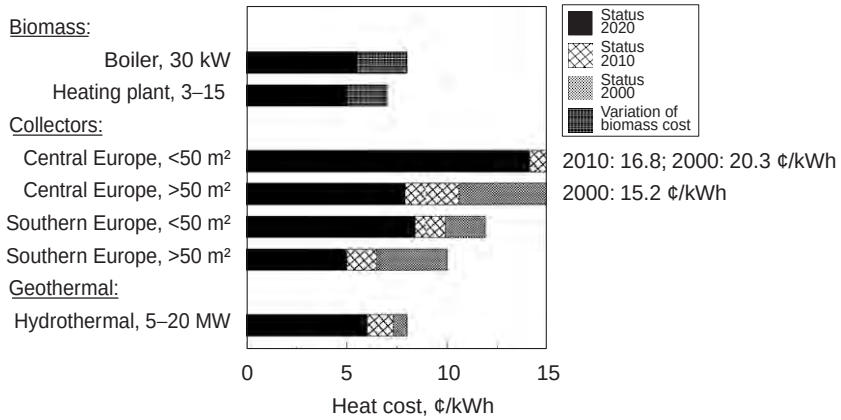


FIGURE 10 Survey of current heat production costs (total length of bars) of representative renewable energy source technologies and possible cost reduction until 2010 and 2020.

plants may generate power within the same cost range because of increasing fuel prices or, in the case of coal, because of costly measures for CO_2 separation. Under such conditions, the many RES technologies could be competitive economically. Photovoltaics, however, at 13¢/kWh in Southern Europe and 20¢/kWh in Central Europe, will still be too expensive to be used to a large extent for general electricity supply. The cost of biomass technologies will strongly depend on the price of the available biomass. Cost reductions by means of technological development or mass production will be small, because of stricter emission standards. If biofuels are to be used to a larger extent, the cost of electricity will not likely decline below a level of about 7–8¢/kWh. Therefore, it has been proposed to use biofuels preferably for the production of transportation fuels, although this will be more expensive than using gasoline or diesel in the foreseeable future. Because of relatively high taxes on fossil fuels in Western Europe, biofuels, with production costs of about 2.5–3¢/MJ, are competitive, if they are not taxed. This approach has been pursued in Germany since 2001 and has resulted in a substantial increase in sales.

Costs of RES technologies for heat production can be assessed similarly. RESs can provide heat with stand-alone systems (e.g., wood-fired boilers, small solar-thermal collectors) and by means of district heating networks. District heating networks are of outstanding importance for RESs to achieve a breakthrough in the heat market. In most cases, only district heating needs support the use of RESs, e.g., geothermal heat plants, large solar collector fields for space heating, and biomass-fired cogeneration plants.

Costs of heat distribution in typical Western European settlements are in the range of 23¢/kWh. Medium-sized heating plants and large collector fields require lower specific investments compared to small boilers or small collectors. Heating costs for the end user of both configurations are therefore comparable (Fig. 10). RES heating costs show large differences. Water heated by typical solar collectors costs about 20¢/kWh in Central Europe; larger collector fields are able to reduce the costs to 15¢/kWh, but this is still four times the price of warm water from an oil- or gas-fired boiler (actual cost of 5–6.5¢/kWh. Only biomass heating is competitive today, if wood is available at a very low cost. Warm water solar collectors in Southern Europe are less expensive (10–12¢/kWh), making them competitive with electrical boilers. Future collector systems will be able to reduce heat costs considerably. Particularly. large collector fields, covering 100 to more than 1000 m^2, will allow future costs to be about 7¢/kWh. The direct use of geothermal energy will result in future costs of about 6¢/kWh. In view of the likelihood of rising prices of fuel oil and natural gas in the future, biomass plants, geothermal units, and solar collector systems will be able to supply end users with competitive thermal energy.

4.3 Environmental Characteristics of Selected RES Technologies

The environmental problems arising from the use of fossil and nuclear energy sources are well acknowledged today; these include global climate change, acidification of ecosystems, risks from nuclear accidents, problems with long-term storage of radioactive waste, and public health effects from air pollution. Because RES technologies largely rely on natural energy and material flow cycles, they offer significant promise in reducing the environmental impact of energy use. Although for most of the RESs the energy conversion process is practically emission free, there may be environmental impacts from the development of RES technologies. The following environmental characterizations of current RES technologies for electricity generation, and their comparison against a modern natural gas-fired combined cycle (CC) power plant, are thus based on a life cycle perspective, taking into account emissions from all of the upstream and downstream processes related to energy conversion (Fig. 11).

It is obvious that for most of the RES technologies, consumption of nonrenewable energy resources and

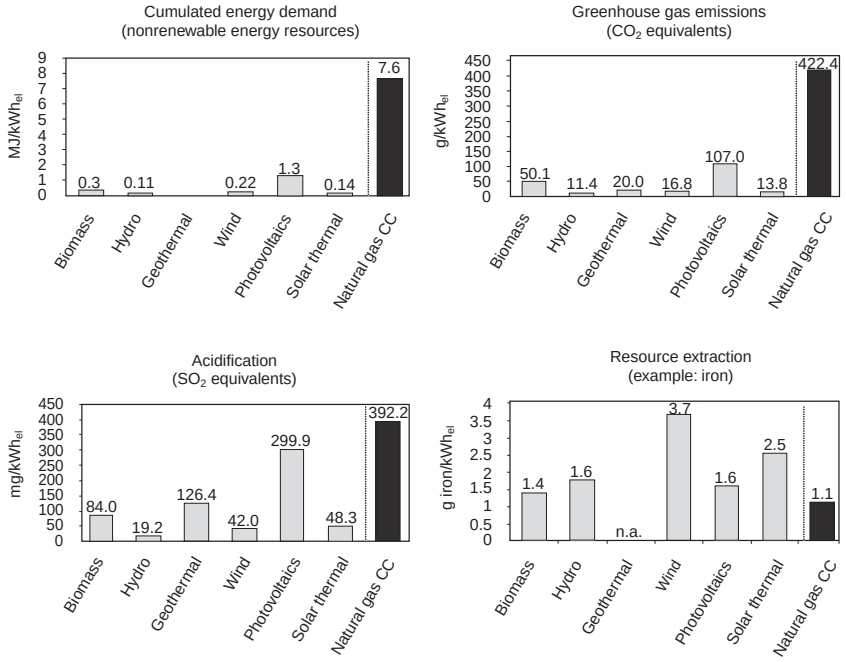


FIGURE 11 Environmental characteristics (cumulated energy demand, greenhouse gas emissions, acidification, and resource extraction) of renewable energy source technologies for electricity generation. Biomass (forest residuals), 20-MW steam turbine; hydro, 3-MW run-of-river plant; geothermal, 900-kW organic Rankine cycle cogeneration plant; wind, 1.5-MW onshore plant; photovoltaic, 3-kW p-type Si roof application (Central Europe); solar thermal, 80-MW parabolic trough (Southern Europe); natural gas CC (combined cycle), 58% efficiency.

greenhouse gas emissions are more than an order of magnitude lower than for the highly efficient gas-fired power plant; the difference is even considerably larger compared to less efficient coal-fired power plants. This fact underlines the important role of RESs in climate protection. Electricity generation from hydropower, wind, and solar-thermal power plants ranks particularly high in this regard. For geothermal energy, the combined production of heat and electricity is required to achieve similar values. The harvesting and processing of biomass fuels requires substantial combustion of fossil fuels, thus biomass electricity generation ranks slightly worse than the processes not requiring combustion. Despite significant improvements in recent years, the manufacturing of photovoltaic cells still requires a quite high material and energy input, leading to relatively high life cycle emissions, which, in the case of SO_2, are even in the same order of magnitude as emissions from gas-fired power plants. The example of iron ore indicates that for the RES technologies, the demand for nonenergy resources is higher, compared to the fossil option, a fact that is sometimes considered a critical drawback of renewable energy sources. However, the expected increase in recycling rates, which in most cases goes along with a further reduction of environmental impacts, compared to the extraction of primary raw materials, will eventually lead to a significant reduction of primary resource consumption.

The dominant environmental impacts of RES technologies are related to emissions from energy conversion in upstream processes, such as component manufacturing or transportation. The data thus primarily characterize the resource efficiency of the underlying economy rather than the performance of a specific energy conversion technology. Changes in the national energy mix, for instance, have a direct impact on the life cycle emissions of RES technologies. The evaluation of emerging technologies should thus be based on conditions that are representative for the time of their market entry, rather than associating them with the environmental load from technologies that they are expected to replace. Figure 12 illustrates the large potential for reduction of life cycle CO_2 emissions from a photovoltaic roof application due to key technological developments (use of solar-grade silicon, increase in recycling rates, and electricity generation with a high share of RESs).

Environmental impacts from energy generation include public health effects, yield losses of crops, and material damage, posing on society a significant economic burden that is not reflected in current energy prices (external costs). "Getting prices right" means imposing a surcharge on the energy price that accounts for environmental damages from power plant emissions. Although quantification of environmental impacts and the subsequent monetary valuation are extremely complex and a matter of

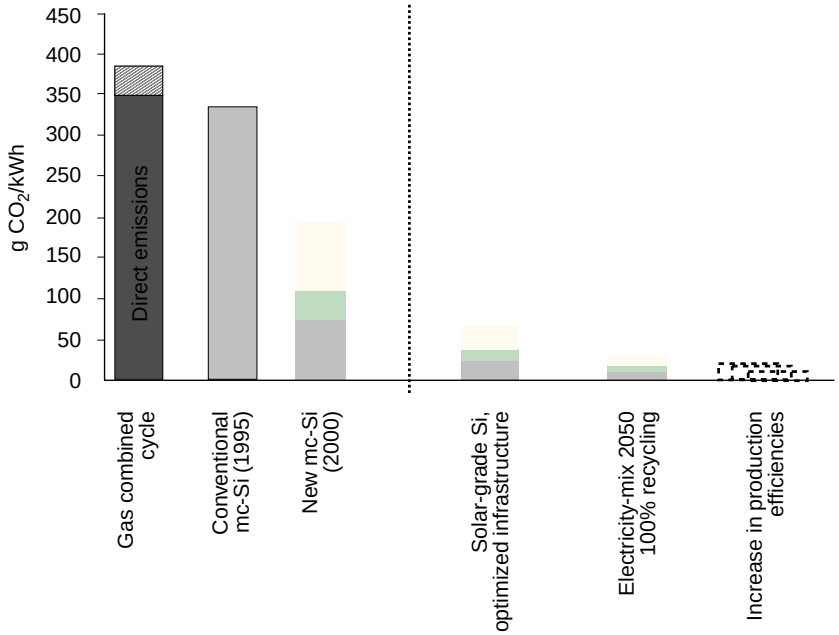


FIGURE 12 Life cycle emissions of CO_2 from a 3-kW p-type Si roof application as a function of technological development and change in the energy mix of an economy. mc-Si, Multicrystalline silicon.

many uncertainties, the European Commission, based on a thorough research effort, decided to allow member states to grant operating aid of up to 5¢/kWh, on the basis of external costs avoided, to new plants producing renewable energy. This measure for internalizing environmental damage costs underlines the environmental advantages of RESs and helps to further increase the competitiveness of RES technologies based on environmental considerations.

5. PROSPECTS OF RESs IN EUROPE

In Europe, several goals have been set to increase the use of renewable energy sources:

- In December 1997, the European Commission adopted the White Paper for a community strategy and action plan, "Energy for the Future: Renewable Sources of Energy". The objective is to increase RESs to an amount equal to 12% of the EU gross inland energy consumption by 2010. In 2001, this action plan was supplemented by a directive of the European Parliament on the promotion of electricity from RESs. The target is to increase the share of RES electricity generation to 22% of total consumption in 2010. The directive contains specific targets for the individual share of RES electricity per EU member state.
- In 1999, the European Commission started a Campaign for Take-Off with the intention to start the implementation strategy set out in the White Paper with specific goals for the period 1999–2003. In 2001, an additional draft directive on biofuels was proposed. The aim was to increase the consumption of biofuels to 2% of the consumption of diesel and gasoline in 2005.

In fact, RES production and use have accelerated clearly since 1995. Between 1990 and 1995, the average growth rate of RES primary energy production in the EU-15 was 2.2%/year; in the following 5-year period, it rose to 3.6%/year and the current growth rate amounts to 4%/year. The average growth rate of the past decade was 2.9%/year. However, the targets defined in the White Paper demand an average growth rate of 7.3%/year until 2010, 2.5 times the value of the past decade (Fig. 13); a somewhat lower increase is necessary for electricity generation (Fig. 14).

A recent investigation, having monitored the progress of RES goals, indicates that the White Paper targets will not be reached with the present

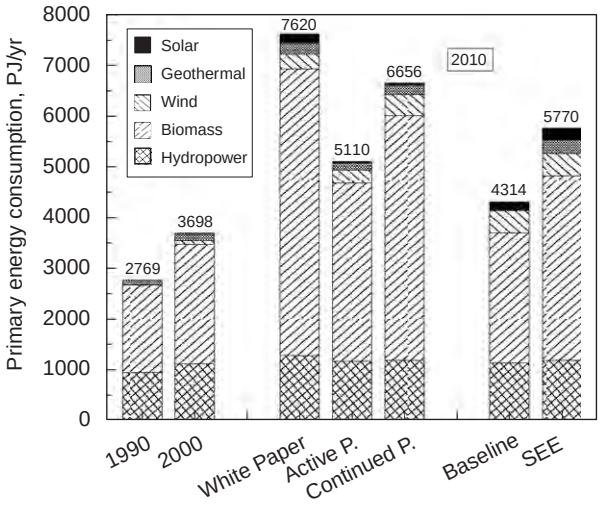


FIGURE 13 Contributions of renewable energy sources to primary energy consumption in the EU-15 in the past (1990, 2000) and to meet the goal for 2010 (from the EU Commission 1997 White Paper), and possible contributions in 2010 for different scenarios [active/continued P (active/continued policy); SEE, solar energy economy].

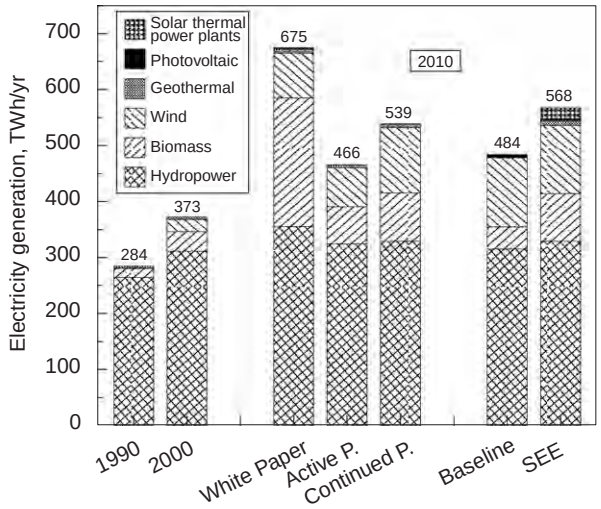


FIGURE 14 Contributions of renewable energy sources to electricity generation in the EU-15 in the past (1990, 2000) and to the goal for 2010 (from the EU Commission 1997 White Paper), and possible contributions in 2010 for different scenarios [active/continued P (active/continued policy); SEE, solar energy economy].

policy instruments and measures. The data on active policies (Figs. 13 and 14) for the year 2010 point out that only an estimated two-thirds of the target values will be achieved in time. All approved policy instruments that have passed in the parliaments of the different EU member states before September 1,

2001 have been considered. Assuming that currently existing incentives will be active in a foreseeable time (continued policy; Figs. 13 and 14), probably 80–85% of the target value will be achieved. Only the target value for wind energy will clearly be exceeded, with 60 GW of installed power in 2010 (White Paper goal: 40 GW); all other RESs will remain below the original objectives. The most severe discrepancies are expected with objectives for photovoltaics (only 20% achievement) and biomass electricity generation (35% achievement).

Substantial structural changes in an energy supply system always require long time periods. Therefore, it is of crucial importance to take into account long-term considerations concerning the transformation of the EU's energy supply structure in order to describe the prospects of RESs appropriately. The November 2000 Green Paper of the EU-Commission, "Towards a European Strategy for the Security of Energy Supply," is based on a projection of the European energy system up to 2030. Hence it can be used as a guideline for investigating the future role of RESs in Europe. The reference scenario baseline uses a business-as-usual development of the demographic and macroeconomic assumptions and is based on the EU energy policies currently in place. The gross domestic product of the EU will be 80% higher in 2030; population will rise only modestly, by 12 million people, until 2010 and will be stable thereafter, with 384 million people. Price increases for energy are assumed to be moderate until 2030, as exemplified by the crude oil price of $30/bbl (U.S.) in 2030. Policy decisions on nuclear energy remain unchanged throughout the projection period. This results in a decrease of the contribution of nuclear energy to half of the present value until 2030. These basic conditions lead to an increase in primary energy demand of 11%, to a growing dependence on energy imports from the present 49% up to 70%, and (because of growing coal consumption) to an increase in CO_2 emissions of 22%, compared to the 1990 reference value of 3068 Mt/year. Apparently, these assumptions also include a weak growth of RESs. Until 2010, RES growth roughly corresponds to the data derived from "active policy" (Figs. 13 and 14). Considering the current dynamics of wind energy growth, only the electricity generation is somewhat higher (baseline, Fig. 14), but the growth of the primary energy contribution of biomass is estimated to be clearly below the value in active policy data (baseline, Fig. 13).

Against this background, it has become clear that the targets of the White Paper for the RES contribution in 2010 cannot be met in time if only the present energy policy instruments (described in Section III) are taken into consideration. Also, after 2010, the growth of RESs will not gain substantial momentum. Hence the European energy policy has to undertake further supportive activities in the next few years to accelerate the market penetration of RES technologies. From a current point of view, the solar energy economy (SEE) scenario (Figs. 13 and 14) gives a realistic assessment of the RES contributions that can be mobilized within the next 8 years. With the strategy "continued policy" as a starting point, it is assumed that the existing energy policy instruments of the EU member states as well as current incentives will be strengthened and coordinated until 2005 in order to effect accelerated growth of RESs before 2010. In this way, almost all goals of the White Paper can be reached until 2010, except for the very high rate of biomass electricity generation assumed in the White Paper for 2010 (230 TWh/year compared to 35 TWh/year in 2000) and the somewhat lower growth rate of solar-thermal collectors. As a complement, electricity generation by concentrating solar-thermal power plants in Southern European countries is added in the SEE scenario; this has not been considered in any other scenario up to now. It is assumed that CSPs will generate 20 TWh/year electricity in 2010, corresponding to 5000 MW of installed power. This takes into consideration the legally mandated improved conditions in the Spanish electricity feed, which allows economically feasible operation of CSP plants by means of a guaranteed payment of 16¢/kWh. Several projects are in preparation now. In the SEE scenario, total electricity generation by RESs will equal 568 TWh/year in 2010, a share of 20% (Fig. 14). The contribution to gross primary energy demand increases to 5770 PJ/year, which is equivalent to 9.5% (Fig. 13).

The long-term development potential of RESs and their significance for a future energy supply system depend decisively on the status and the growth dynamics that have been reached by 2010. Comparing different long-term scenarios for the energy supply of the EU, this is quite obvious. It is noteworthy that the already discussed reference scenario "baseline" shows only a small growth of RESs over several decades (Figs. 15 and 16). The share of RESs for electricity generation obtains the 2010 target of the White Paper in 2030, and the growth in the heat and fuel market almost stagnates, so that the RES contribution to gross primary energy consumption rises only marginally. Extrapolating this scenario to 2050, RESs will fulfill at best a share

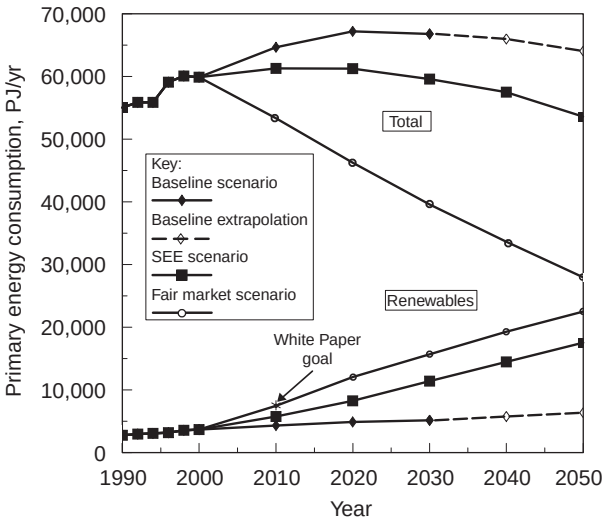


FIGURE 15 Future primary energy consumption in the EU-15 until 2050 and possible contributions of renewable energy sources for different scenarios. SEE, Solar energy economy.

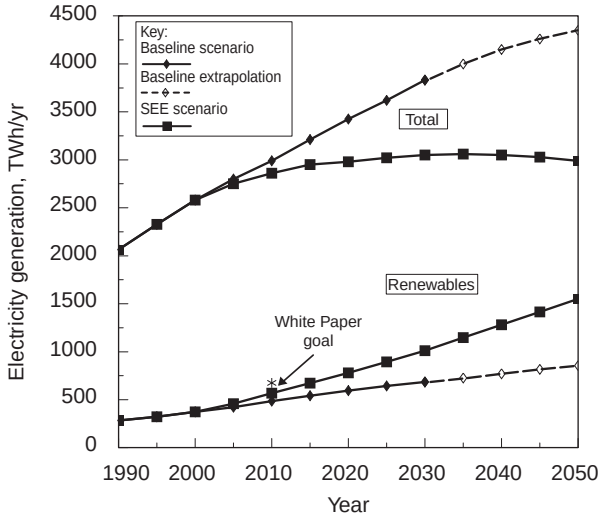


FIGURE 16 Future electricity generation in the EU-15 until 2050 and possible contributions of renewable energy sources for different scenarios. SEE, Solar energy economy.

of 10% of primary energy consumption (6400 PJ/year) and would remain a nearly negligible energy source. Such business-as-usual development would offer no chance for RESs to contribute significantly to climate change mitigation or prolongation of oil and gas resources.

A totally different perspective is offered by the fair market scenario (Fig. 15) that was developed by a consortium of five research teams from Germany, France, Belgium, and Denmark in 1997. The main approach of this scenario is the inclusion of externalities of energy consumption in energy prices to determine market behavior. Once the magnitude of external costs has been fixed, e.g., by a political consensus on taxation, the market can act on prices as usual. Adding the known externalities to market prices of energy leads to considerably strong increases and results in price levels that are twice to three times as high as the energy prices used in the baseline scenario. In parallel, learning curves have been used to calculate the decreasing future cost of RES technologies. With these conditions, a strong market growth of RESs as well as of energy-saving technologies can be expected. The transformation of the energy system will be determined mainly by structural limits, e.g., the decommissioning of power plants and boilers for heating, the reconstruction of buildings, the construction of new production sites for RES technologies, or necessary changes of infrastructure for new fuels. These very favorable conditions for RES technologies result in average growth rates for all technologies of 4.5%/year during the next three decades, starting with about 7%/year in the next 10 years. Growth is assumed to continue until 2050, with slowly declining rates, resulting in a RES contribution of 22,400 PJ/year in 2050 to primary energy consumption (Fig. 15), which corresponds to six times the present value. Energy demand declines simultaneously because many energy-saving technologies are cost-effective at the assumed prices. The fair market scenario shows that, in 2050, the demand for energy services in the EU-15 can be met by half of the present energy production. Therefore, RESs are able to contribute 80% of the remaining primary energy consumption. As a result, CO_2 emissions are negligible compared to the present value, and the dependency on external energy resources is diminished to 15% of the actual imports of crude oil and natural gas. This scenario describes a development of the European energy system that could occur under very favorable political conditions. It would considerably reduce the deficits of the current energy system with respect to sustainability. In fact, this scenario is rather unlikely from the present point of view, but it shows, in principle, the unused potentials of RESs and rational approach of modern economies.

The third scenario, SEE, tries to balance the challenges of reaching sustainability in energy supply and the well-known slowness in removing the numerous barriers and restraints, which are opposed to accelerated structural changes in the European energy system. The SEE scenario outlines the chances

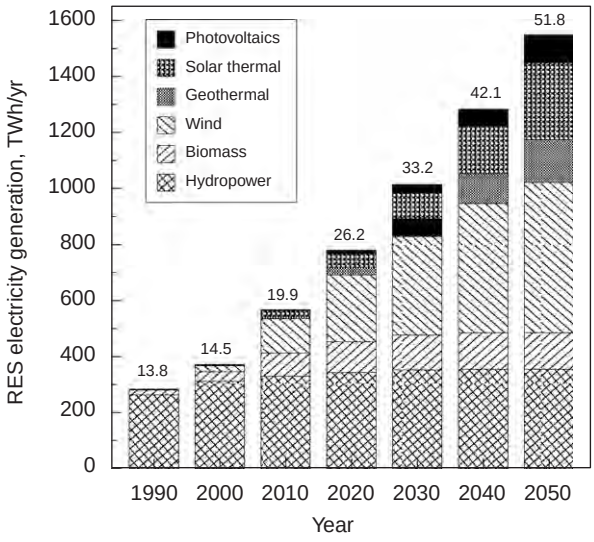


FIGURE 17 Growth of renewable energy source electricity generation in the solar energy economy scenario in the EU-15 until 2050, by individual sources. Numbers over bars indicate the percentage of total generation.

for a more independent, secure, and environmentally sound energy system in Europe with "careful optimism." This includes an efficient and quick advancement and coordination of support mechanisms for RESs (and for energy conservation) in the EU, but, on the other hand, no excessive demands by society and economy. Starting from the interim values for 2010 shown in Figs. 13 and 14, the contribution of RESs grows continuously and meets in 2050 one-third of total primary energy consumption (17,500 PJ/year). Also, reinforcement of energy conservation measures has an impact, resulting in a total energy productivity in 2050 of twice the present value. In particular, it is assumed that the steady growth in electricity demand will be slowed down by adequate energy conservation efforts. In combination with increasing RES electricity generation, about 50% of electricity demand in 2050 can be met by RESs (Fig. 16). RES contributions to the heat and fuel market are growing somewhat more slowly.

To achieve an economically attractive growth of RES production and use, a well-balanced and well-timed mobilization of all RES technologies is of paramount importance. This mobilization depends on the technical potentials of the technologies, their actual costs, their cost reduction potentials, and their technological maturity. In Fig. 17, this is demonstrated with the RES electricity generation of the SEE scenario. Until 2010, cost-effective hydropower, biomass power plants, and wind energy converters are the main contributors to RES growth. Hydropower and biomass will then reach their limits (biomass is increasingly used for fuel production). After 2010, the continuously growing use of wind energy is complemented by electricity from solar-thermal power plants and from geothermal energy. Relevant contributions from photovoltaics will not emerge until 2030 because of relatively high costs. The activities within the SEE scenario correspond widely to the minimum level of technically feasible changes. CO_2 emissions in 2050 are reduced to 2100 Mt/year, or 67% of the current value. This is only a first approach to an effective CO_2 abatement policy. Also, the dependency on imports of crude oil and natural gas, at 65% of the present value, is still too high. Altogether, the scenario represents the minimum requirements for European energy policy, if a relevant role of RESs in future energy supply is one of the substantial goals.

SEE ALSO THE FOLLOWING ARTICLES

Biomass for Renewable Energy and Fuels • Computer Modeling of Renewable Power Systems • European Union Energy Policy • Public Reaction to Renewable Energy Sources and Systems • Renewable Energy and the City • Renewable Energy in Southern Africa • Renewable Energy in the United States • Renewable Energy Policies and Barriers • Renewable Energy, Taxonomic Overview • Renewable Portfolio Standard

Further Reading

Capros, P., *et al.* (2000). "Scenarios Related to the Security of Supply of the European Union." Report for EU15. National Technical University of Athens, Greece.

Commission of the European Communities. (1997). White Paper. "Energy for the Future: Renewable Sources of Energy." EU-Commission COM (97) 599, final. European Commission, Brussels, Belgium.

Commission of the European Communities. (2000). Green Paper. "Towards a European Strategy for the Security of Energy Supply." EU-Commission COM (2000) 769. European Commission, Brussels, Belgium.

ENER Forum. (2002). "Successfully Promoting Renewable Energy Sources in Europe." The European Network for Energy Economics Research. Conference Proceedings, Budapest, Hungary, 6–7 June, 2002.

European Commission. (1999). "ExternE—Externalities of Energy, Vol. 7: Methodology 1998 Update (EUR 19083). EU-DG XII: Science, Research and Development."Office for Official Publications of the European Communities, Luxembourg.

Eurostat (Statistical Office for the European Communities). (2001). "Renewable Energy Sources Statistics in the European Union. Eurostat, Theme 8: Environment and Energy, Edition 2001." Eurostat, Luxembourg.

Hamelink, M., *et al.* (2002). Implementation of renewable energies in the European Union until 2010. "Study within the Project PETIR (Progress of Renewable Energy: Target Setting, Implementation and Realisation). Summary." Ecofys, Utrecht, Netherlands.

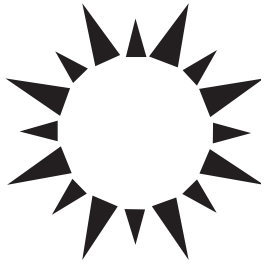
LTI-Research Group (ed.). (1998). "Long-Term Integration of Renewable Energy Sources into the European Energy System." Physica-Verlag, Heidelberg and New York.

Nitsch, J., and Staiss, F. (2003). Strategies for introducing renewable energies and the contribution of photovoltaics. *In* "Photovoltaics Guidebook for Decision Makers" (A. Bubenzer and J. Luther, Eds.). Springer, Berlin and New York.

Nitsch, J., *et al.* (1999). "Climate Protection Using Renewable Energies" [in German; English summary]. Study commissioned for the Umweltbundesamt Berlin, UBA-Report 2/00. Erich Schmidt Verlag, Berlin.

Orthen, S., and Urbschat, C. (2002). "Sunrise 2002—The European Markets for Solar Thermal Collectors and Photovoltaics" [in German]. German Energy Agency, Berlin.

Renewable Energy in Southern Africa



DIETER HOLM
Sustainable Energy Society of Southern Africa
La Montagne, South Africa

1. **Impact of Renewable Energy in Southern Africa**
2. **The Southern African Development Community**
3. **Angola**
4. **Botswana**
5. **Democratic Republic of Congo**
6. **Lesotho**
7. **Malawi**
8. **Mauritius**
9. **Mozambique**
10. **Namibia**
11. **Seychelles**
12. **South Africa**
13. **Swaziland**
14. **Tanzania**
15. **Zambia**
16. **Zimbabwe**

Glossary

biomass Biological material (in a given area).
capacity The rating of a generator, transmission, or system.
desalination Salt removal from brackish or seawater.
geothermal Pertaining to the heat of the earth's subsurface.
photovoltaic (PV) A device converting sunlight into electricity.
resources Estimated total quantity such as minerals in the ground.

Renewable energies (REs) are those naturally occurring nondepleting energies that are derived directly and indirectly from the sun, such as biomass, geothermal, hydropower, ocean currents, solar radiation, tidal wave, and wind energy. These can be harnessed through renewable energy technologies (RETs), such as photovoltaic, solar water heaters, wind and water turbines, and solar passive design. RETs deliver energy services such as mechanical work, heating, cooling, lighting, crop drying, telecommunication, transport, water pumping, desalination, and water treatment. REs are contrasted with finite nonreplenishing fossil fuels such as coal, petroleum, and uranium. REs and RETs are closely linked to new technologies such as fuel cells and the hydrogen economy as well as to energy efficiency. Africa relies heavily on traditional renewable energy in the form of fuelwood (70–80%) for pest control, water heating, space heating, food preparation, and beer brewing for the men. Beer brewing requires 25–50% of energy in villages. Women and children collect the fuelwood over ever-increasing distances, following the frontiers of desertification. The current and future impact of RE in southern Africa is a major building block toward sustainable development.

1. IMPACT OF RENEWABLE ENERGY IN SOUTHERN AFRICA

Solar cooking and solar water heaters (SWHs) can eliminate the need for collecting fuelwood and inhaling unhealthy smoke. This contributes to social sustainability because women and children who currently spend excessive time and energy collecting wood have time for education. Education and training contribute to improved work opportunities and a lower fertility rate, thereby advancing economic sustainability. The reduced population and renewable energy technologies (RETs) eliminate deforestation and desertification, thereby fostering environmental sustainability.

Solar passive buildings in urban and rural areas can totally obviate the need for artificial space

 © 2004 Elsevier Inc. All rights reserved.

heating and cooling by kerosene, which often is a cause of death by fires or accidental swallowing. It also eliminates the use of expensive imported electricity and its environmental impact where this is generated by coal, nuclear, or hydropower schemes larger than 10 MW.

Modern RETs contribute to energy supply diversity, minimizing cost exposure to imported fossil fuels. Most African countries are importers of petroleum.

REs have a much lower cradle-to-grave life cycle impact on the environment than conventional fossil fuels. The environmental costs (social costs) of conventional energies are those costs caused by the generator of energy but paid for or carried by the local, regional, or global community, including 2 billion people who have no access to modern energies.

REs can be generated locally, feeding the electrical grid and reducing network losses. This is significant in southern Africa because power generation is centralized and power is transmitted over large distances.

Distributed RE generation also democratizes power, making it less likely that politicians can manipulate energy. Localized RETs enhance ownership of local communities that receive the local benefits and revenues. RETs are a new technology creating new export potential.

RETs create many more jobs than conventional energies. Job creation is a major instrument in poverty alleviation. The European Union plans to create 900,000 new jobs by 2020. Southern Africa may create approximately 1 million new work opportunities.

Shell International foresees that REs will provide more than 50% of the world's energy. By employing modern RETs, southern Africa's RE share may be 20% by 2020 and 50% by 2050.

2. THE SOUTHERN AFRICAN DEVELOPMENT COMMUNITY

This section discusses salient characteristic communalities under the headings Natural/Physical, Institutional/Political, Social, Economic, and RE Technological. The ensuing country-specific sections are structured in the same manner to simplify cross-referencing. Although every effort was made to collate recent and credible information, such as that supplied by Dr. P. Ashton (personal communication, 2003), it should be borne in mind that statistics for the Third World generally tend to be dated and unreliable.

2.1 Natural/Physical

The Southern African Development Community (SADC) is a group of 14 countries consisting of Angola, Botswana, the Democratic Republic of Congo (DRC), Lesotho, Malawi, Mauritius, Mozambique, Namibia, Seychelles, South Africa, Swaziland, Tanzania, Zambia, and Zimbabwe.

The country with the largest area is the war-torn DRC (2,345,410 km^2), which also has the largest population (52,046 million). The SADC countries extend over sub-Saharan Africa from 5° north to 34° south latitude, from the equatorial tropics over the arid Horse Latitudes to the Mediterranean climate zone of the Cape of Good Hope. Although the Southern Hemisphere receives intensive sunshine during summer (perihelion), temperatures are tempered by altitude and cloud cover. 59% of the world's most intense winter sunshine (more than 6 kWh/m^2/day) falls on southern Africa (for comparison, London receives 1 kWh/m^2/day). The climate is characterized by steady hot and humid conditions in the Congo Basin and major temperature and humidity changes in the remaining inland zones, which are not tempered by the sea. Snow only occurs in the eastern and southern mountain ranges. The warm, low-altitude areas are malaria infested.

2.2 Institutional/Political

The forerunner of SADC was founded to destabilize South Africa during the Apartheid era. Today, SADC includes South Africa and has the aim of regional cooperation and economic development. South Africa's dominant technoeconomic power understandably makes for very unequal partners.

The New Partnership for African Development (NEPAD) proposed by South Africa's president Thabo Mbeki offers a promise that African rulers will in the future govern their countries according to rule of law in return for $64 billion in foreign investments and donations. However, potential international long-term investors and donors have taken a wait-and-see position in view of the situation in Zimbabwe and Mbeki's efforts at reconciling the demands of the poor with those of development.

The NEPAD office identified six conventional electricity projects, two of which are in SADC: a 1300-MW hydropower scheme downstream from the Cahora Bassa dam in Mozambique and an interconnector between Malawi and Mozambique. Ironically, modern renewable energy projects are not

part of the NEPAD wish list, although Africa enjoys 96% of the world's best sunshine.

An African power pool has also been proposed by South Africa's state monopoly utility. The World Bank, the African Development Bank, the West African Development Bank, and others have shown interest in a proposed West African power pool, which would involve several major electrical power stations and interconnectors at a total cost of $13 billion. However, international investors remain lukewarm because of the region's tradition of military coups.

The often quoted nominal 39,000 MW Inga hydropower, which actually could produce a maximum of 26,000 MW, has been criticized with regard to cost, time, and technical considerations, e.g., untested long-distance high-tension transmission technology. In addition, there are grave concerns about the risks in the conflict-ridden DRC.

The previously discussed schemes tend to have common traits:

1. With a view to attracting international finance with associated high transaction costs, the proposed schemes are very large by African standards. This means that African contractors are out of the game, except for performing menial tasks. At best, South African contractors would participate in worthwhile work, thereby enlarging the existing continental discrepancy.
2. All schemes are based on the dated concept of supply side-driven interventions typically promoted by large international finance, engineering, and Enron-type utility interests combined with local political positioning. The modern concept of need-driven integrated resource planning with its concomitant developmental benefits is conspicuously absent.
3. The schemes are of very large scale, tying up huge amounts of money and other resources that are far beyond Africa's ability to absorb the electricity in productive enterprises. A system of smaller, localized, modular, and incremental renewable energy programs would be much more appropriate due to Africa's immense distances and abundant renewable resources.
4. The schemes blithely ignore Africa's state of political instability, which renders centralized electricity generation systems with long-distance transmission lines extremely vulnerable, as demonstrated by the Cahora Bassa hydropower scheme.
5. It is a tacit assumption of all schemes that large-scale electrification leads to development. Numerous projects throughout the world have shown the opposite to be true: Electrification follows development. Gunter Schramm of the World Bank states that "electricity by itself does not provide new income sources. As has been shown again and again, electricity follows, rather than leads economic development."
6. Although megaelectricity schemes are important to some rulers they often lead to the removal of national protectionism thereby opening local markets to overwhelming global competition. This global competition has often grown by a combination of national subsidies like the U.S. and EU agribusiness, and by national protection like the current U.S. steel industry. The developing world perceives the World Bank and International Monetary Fund as pontificating: "Do as we say, not as we did and do."

2.3 Social

As noted by Ashton, the total population within SADC is approximately 201 million. The DRC has the largest population (52 million). Of the SADC total population, 34% is urbanized, 71% literate, 43.5% malnourished, and 9% of adults are HIV/AIDS infected. A total of 15.7% have malaria, leading to 300,000 deaths each year. Approximately 500,000 have tuberculosis, causing approximately 100,000 deaths per year. Of the two-thirds majority living in rural areas, 51.1% have no access to improved water and 52.1% no sanitation, whereas in urban areas these percentages are 10.8% and 21.9%, respectively.

There are approximately 400 languages, not dialects. South Africa has 11 official languages. The number of religious denominations is estimated to be 10,000, with 6000 in South Africa.

The cultural patterns range from prehistoric hunter–gatherers, represented by the San, to iron-age herdsmen, burn-and-slash migrant peasants, sedentary subsistence and commercial farmers, industrialists, and contemporary postindustrial cosmopolites.

Not surprisingly, the Human Development Index (HDI) varies widely from approximately 0.3 to approximately 0.7. Likewise, the HDI ranking is between 120 and 170.

2.4 Economic

SADC's total per capita income was $570 billion in 2000, yielding an annual gross domestic product (GDP) per capita of $2833. This ranges from approximately $7600 in Botswana to an unknown

figure in Zimbabwe, where the inflation rate is 400%. Two-thirds of the population lives below the poverty line of $2 per day, ranging from 40% in the Kingdom of Swaziland to more than 80% in the Republic of Zimbabwe. The GINI coefficient (the gap between highest and lowest 10% income) is 0.58, with a range of 0.38–0.70. The sub-Saharan power sector received $363 million during 1990–1998, representing a mere 0.8% of the total $46.39 billion foreign direct investment flow to the entire developing world.

Most economic indicators have been declining. A limited number of scientific publications, patents, and inventions of SADC mostly stem from South Africa. The mineral resources of SADC are in stark contrast with the economic output. Gold, diamonds, copper, zinc, iron, gems, and rare metals are mined on a large scale.

Gas has been found along the western, southern, and eastern coasts as well as in inland Botswana. From Angola northward, oil is being extracted. Open cast coal mines are found in large tracts delivering the feedstock for more than half of Africa's electricity and to Sasol. In addition, cost-effective renewable energy resources by far exceed current and projected needs.

Despite the abundance of raw materials and sustainable energy resources, SADC has not learned to beneficiate the resources, remaining largely dependent on expensive imported value-added products and know-how.

2.5 RE Technological

There are 33 telephone lines, 39 cell phones, 39 TVs, 253 radios, and only 18 Internet users per 1000 people in SADC. The number of Internet users range from 0.1 (DRC) to 71 (South Africa) per 1000. The paved road length per person ranges from 3.09 km (Namibia) to 0.02 km (DRC), with an average of 0. 9 km.Whereas Botswana has 0.54 km railway length per 1000 inhabitants, Lesotho has zero. The SADC average is 0.34 km. It must be borne in mind that after waterways, railways provide the most energy-efficient transport mode. Road transport is relatively inefficient at the higher altitudes of southern Africa, and petroleum products are often imported.

It appears that SADC countries, with the exception of South Africa, have invested relative little in conventional energy-intensive infrastructure such as centralized coal-fired or nuclear power stations with transmission networks. Furthermore, two-thirds of the population lives in widely dispersed rural areas and has low energy consumption needs and very low income levels, which do not justify grid extension. Extending the national grid into these areas would place an unrealistic burden on both the national economy and the recipients who are supposed to be beneficiaries.

Given the existing low investment in old fossil-fueled infrastructure, the very high cost and risk of extending this into thinly populated, low-consumption rural areas, and the abundance of renewable sources such as sun, wind, biofuel, and microhydro power, it seems obvious to change to clean modern and environmentally friendly technologies. However, a number of donor-driven projects introducing 50 W_p photovoltaic solar home systems (SHSs) in remote rural areas have led to disappointing results. The reasons are not well publicized:

1. SADC governments tend to have a centralist approach and try to channel all initiatives through government departments. Even in the developed world, this is not necessarily the most efficient and transparent process.
2. Localized government interventions with subsidized SHSs destroy the private enterprise market and cause suspicion of corruption.
3. Inferior products and underdeveloped technologies were introduced because there was no willingness or capacity to control quality.
4. Installation and especially maintenance proved to be extremely difficult in remote rural areas, where telecommunication and transport infrastructure is problematic, if not totally absent.
5. Existing utilities viewed SHSs as a threat and prohibited SHS developments near their power lines or suddenly brought the grid to areas where SHSs had been established. They also created the expectation, supported by politicians, that universal grid-connected power is a reality in the very near future. This established a reluctance to invest in SHSs.
6. In hindsight, it appears that the introduction of a novel technology into an impoverished market under the most trying physical conditions was not a good idea.
7. As a result of insufficient or misleading information, SHS users had unrealistic performance expectations, which led to the stigmatization of SHSs as the poor man's power. The rich role models were not seen to be using SHSs. It seemed that an experiment was being conducted on the poor, who could least afford it.
8. Experience in Europe and Japan demonstrated that the major growth in RE technologies such as

SHSs is in grid-connected applications within the framework of a RE grid-feeder law. This law encourages reverse metering of RE-generated electricity fed into the grid at higher tariffs than the fossil fuel grid tariffs. The difference is shared by all grid users and is negligible. Batteries—normally the Achilles' heel of SHSs—are not required because the grid functions as the storage medium. This reduces capital and maintenance costs. Additionally, maintenance is cheaper due to easier access and communication. The well-to-do more readily introduce new technologies and demand service. All this contributes to a growing acceptance and desirability of photovoltaic (PV) systems as a status symbol of success and progress. So the mass market becomes established, leading to products and services that are acceptable and affordable to the less fortunate. After all, electronic watches, radios, and cell phones were not initially introduced into the low-cost end of the market. Learning from their own experience and from abroad, SADC countries can alleviate poverty and initiate sustainable development by initially introducing RE technologies using the successful and proven grid-feeder tariff structure. This will establish indigenous business, create jobs, and transform the market rapidly and sustainably. In this way, governments do not delay progress by interfering with implementation but, rather, create a conducive framework.

Most RETs are imported because of a lack of awareness on the part of decision makers and the public. R&D investment in RETs, energy efficiency, and government incentives have been neglected. The potential of clean development mechanisms has yet to be realized.

3. ANGOLA

3.1 Natural/Physical

Angola is rich in oil, gas, and mineral resources. It also has considerable solar potential (5.4 kWh/m^2/day in midwinter) in the southern region, except the coastal areas. The potential of small hydro is noteworthy. Wind and ocean energy resources remain to be assessed.

3.2 Institutional/Political

Governance institutions and infrastructure have suffered severely during the prolonged war, which, strangely, did not disrupt petroleum production. This industry accounts for 54% of the GDP and 83.5% of the government tax revenue. Rural areas are not linked to the electricity grid, which has been sabotaged extensively. There is strong government control over tariffs.

3.3 Social

The life expectancy in Angola is the lowest of all SADC countries (due to war and HIV/AIDS). Literacy is also the lowest.

3.4 Economic

Foreign direct investment (FDI) has been twice as much as in Tanzania and far more than that for all other SADC countries. It has gone into the oil business; the per capita GDP is the fourth lowest. Both the African competitiveness ranking and HDI are the lowest. No concerted effort toward RE or energy efficiency has been undertaken.

3.5 Technological

There is no awareness of modern RETs, nor is there any local production. The high illiteracy rate is a barrier to employment. However, the need for PV, SWH, crop drying, and small hydropower is staggering.

4. BOTSWANA

4.1 Natural/Physical

The landlocked country has diamonds, copper, nickel, and coal of high rock and sulfur content, and the winter sunshine in excess of 6.5 kWh/m^2/day places it in the core area of the world record sunshine. This enables all direct thermal applications, such as solar chimneys, solar ponds, solar concentrating, SWH, and PV.

4.2 Institutional/Political

A yearly National Development Plan has resulted in clearly defined targets and budgets. The Ministry of Mineral, Energy & Water Affairs (MMEWA) formulates, directs, and coordinates policy through the Energy Affairs Division, whereas the Botswana Power Cooperation, a parastatal under MMEWA, is a vertically integrated electricity utility operating the coal-fired power station at Murule that delivers 58%

of the national grid electricity. The remainder is imported.

Botswana has enjoyed long-term political stability, applies the death penalty, has a low crime rate, and has established investor-friendly policies. The Botswana Technology Center, the Rural Industries Innovation Center, and nongovernmental organizations are active in implementing RETs.

4.3 Social

Botswana has the highest regional urbanization rate of 46% and third highest life expectancy of 67 years, although HIV/AIDS is prominent. It is ranked third in competitiveness, with a comparatively low literacy rate of 67%.

4.4 Economic

The per capita GDP is the third highest and has been consistently so for a number of years. Its HDI is 57.2%, ranking lower than that of the Seychelles, Mauritius, South Africa, and Swaziland. Grid electricity is relatively expensive, which encourages the use of capital-intensive RETs.

4.5 Technological

Botswana is the region's leader in the application of SWHs and has its own manufacturing industry. Problems of scaling and high salinity experienced with direct thermosiphon systems led to the use of indirect systems. These are installed mainly in government and institutional buildings. However, the market is booming, with 8000–10,000 m^2 installations per year ($2.5 million/year) and an installed capacity of 50,000 m^2. This amounts to 0.053 m^2 per capita, which is low by world standards.

PV installations are in use in the telecommunications sector. Government initiatives drive the PV market since grid extension is the costlier option for smaller structures such as housing for teachers and the defense force, prisons, public buildings, and government buildings in rural towns. Rural energization by SHSs is gaining momentum. The current market is estimated to be 13.4 kWh/year ($70,000/year), with an approximate growth rate of 12% per year.

There is some local manufacture of balance of system components. The Botswana Technology Center is active in energy efficient/solar passive buildings. This institution has also performed a wind mapping project.

5. DEMOCRATIC REPUBLIC OF CONGO

5.1 Natural/Physical

This tropical rain forest straddles the equator and is the most northerly SADC country. It is well endowed with mineral resources, oil, and gas, and the Congo River has considerable potential for hydropower at Inga (26,000 MW).

5.2 Institutional/Political

Unfortunately, since liberation the DRC has experienced conflicting foreign and local interests with regard to its oil and mineral resources and seemingly unending wars. This has destabilized the area, with ripple effects on the whole of Africa. Governance is in a survival mode. Statistics are not reliable.

5.3 Social

The DRC has the largest population in the SADC, with a life expectancy of 51 years and HDI of 43.1%.

5.4 Economic

GDP was $230/capita in 1998, which is the second lowest in the region; there appear to be no recent statistics. The FDI in 1995 was estimated to be $2 million. There is neither a global nor an Africa competitiveness rating.

5.5 Technological

The DRC does not publish any official documentation on RE. The often mentioned hydropower potential of the Congo River seems fraught with political uncertainty. In addition, the major consumer in the near term would be South Africa, which would require more than 3000 km of new power lines, in part over land-mined areas.

6. LESOTHO

6.1 Natural/Physical

The mountain kingdom of Lesotho is an enclave within South Africa and Mozambique. Owing to its altitude (3096 m), it is much colder than its latitude (30°S) would indicate. Snow is not exceptional. The existing Oxbow hydro scheme exports to South

Africa (with a planned additional capacity of 200 MW_p). The inhabitants' use of firewood has practically denuded the land of trees, although reforestation conferences are popular. There is ample microhydro potential. Daily winter sunshine is 5.5 kWh/m^2/day.

6.2 Institutional/Political

The Lesotho Energy Master Plan of 1988 was revised in 1999, with the Lesotho Electricity Master Plan Strategy of 1996 as a subcomponent, updated in 1998. Owing to the sparse population and the mountainous nature, only approximately 3% of the population has access to grid electricity, which is planned to increase to 5% by 2003. The Lesotho Electricity Corporation is being privatized.

6.3 Social

Of the total population, 17% is urbanized, but there is a relatively high literacy rate of 78%. The life expectancy is 59.4 years, whereas the HDI is 53.5%.

6.4 Economic

The GDP per capita is $920, and the Africa competitiveness ranking is 13.

6.5 RE Technological

As a result of the low grid electrification, dispersed population, high solar radiation, and deforestation, Lesotho is a good country for solar applications. This is limited by the low disposable income of the population and the lack of trained installers.

The high altitude causes frost conditions, necessitating special precautions with SWHs. Four small hydro systems exist (3268 MW), and there is potential for more. Approximately 60 kW_p of PV has been installed, with limited growth prospects.

7. MALAWI

7.1 Natural/Physical

The country on the shores of Lake Malawi (706 m altitude) enjoys more than 5 kWh/m^2/day of sunshine during the winter. The vegetation varies between savannah and tropical rain forest.

7.2 Institutional/Political

The Malawi legal system is based on English and local traditions. The energy policy stems from the Vision 2020 development plan, and the Electricity Council fulfills a regulatory function. A National Sustainable Renewable Energy Program was in place before the 2002 World Summit on Sustainable Development.

7.3 Social

Malaria, tuberculosis, HIV/AIDS, and continued malnutrition are the most pressing health problems. The literacy rate is 40% and secondary school enrollment 5%. The HDI is 40%, with a life expectancy of 52 years.

7.4 Economic

In African competitiveness, Malawi ranks 19th, with a GNP per capita of $200. Foreign investment is approximately $1.8 billion.

7.5 RE Technological

A total of 97.6% of Malawi's 214-MW grid electricity is hydro generated, and the government has borrowed $8 million from the Development Bank of Southern Africa for extended hydroelectric power. Only 3% of rural households have access to grid electricity. Wood fuel continues to dominate the fuel economy, where households consume 70% of noncommercial wood fuel. The United Nations Educational, Scientific, and Cultural Organization (UNESCO) has assisted with a solar energy promotion program.

The installed capacity of PV is approximately 50 kW_p, increasing at approximately 4 kW_p annually. The potential is considerable. The SWH market is not very active. Approximately 3000 m^2 has been installed, amounting to a negligible market penetration.

8. MAURITIUS

8.1 Natural/Physical

This 2040 km^2 island is situated in the southerly warm Indian Ocean Current at approximately 20°S latitude, receiving 4.75 kWh/m^2/day of sunshine in winter. The climate is tropical during summer (22–29° C) and subtropical in winter (18–23° C). This is

where the ebony forests and the Dodo became extinct. Rainfall varies from 1150 mm in the west to 2300 mm on the 580-m altitude plateau. Cyclones are not unusual.

8.2 Institutional/Political

The legal system is based on French and English law. Creole is widely spoken, with English as official language. In the absence of natural fossil resources, energy legislation is guided by the development of local RE resources and the involvement of the private sector, which utilizes hydropower and bagasse (sugarcane residues). However, the bulk is imported fossils. The energy demand is growing at 5% per year.

8.3 Social

The population of 1.1 million has a relatively high literacy rate of 81%, a life expectancy of 70 years, and the highest SADC urbanization of 43%. Consequently, the HDI is 77.2%, which is by far the highest in southern Africa.

8.4 Economic

Mauritius ranks second out of 24 in African competitiveness, and it is one of the few with a global competitiveness rating on record (32nd out of 75). The GNP per capita was $ 3700 in 1998. Mauritius has high credit worthiness and was ranked first in Africa in 1996. Manufacturing, tourism, and financial/professional services lead the economy.

8.5 RE Technological

Diesel engines generate 61.7% (176 MW) of the electricity. A seasonal contribution of 54.2 MW comes from hydro plants and 90.5 MW from independent power producers, mainly bagasse plants. The entire population has electrical grid access. Mauritius is the SADC pioneer of grid-connected RE, including PV. Wind speeds of 8 m/s make wind turbines an attractive option. A total of 20,000 SWHs have been installed, representing a market penetration of 8% in the household market. The growth is stable at 2000 m^2 ($1 million) per year.

More growth in grid-connected and stand-alone (on-island) PV is envisaged. Mauritius demonstrates that electrification follows development and that cheap energy is not a competitive advantage.

9. MOZAMBIQUE

9.1 Natural/Physical

Mozambique is a mostly low-lying country on the warm Agulhas Current of the West Indian Ocean, extending from 11 to 27°S latitude in a tropical to subtropical climate. Winter insolation is 5 kWh/m^2/day (approximately twice the sunshine of Bilbao). It features an area of 799,380 km^2 and a coastline of 2515 km with two major African rivers—the Zambezi and Limpopo—traversing. The huge Cahora Bassa hydroelectric dam, which exports electricity, is on the Zambezi River. The Pande field northeast of Maputo will soon export natural gas to the petrochemical plants of Sasol in South Africa. There are also extensive coalfields at Metangula, Itule, Tete Msanghansuvu, and Chiomo.

9.2 Institutional/Political

After decolonization in 1975, a prolonged civil war began that ended in the 1999 general elections, leading to economic development from a low base. Mozambique ratified the SADC trade protocol and has the lowest import tariffs, along with Malawi. An energy-intensive aluminum smelter is under construction. The RE resources have not been assessed.

9.3 Social

Of the approximately 16.8 million people, 23% are urbanized. The life expectancy is 46 years, and literacy is 28%—the lowest in SADC. Portuguese is the official language. The HDI recently increased from 28.1 to 32.2%, whereas the African competitiveness rating is 18 of 24.

9.4 Economic

The joint Maputo corridor project linking South Africa's infrastructure to upgraded harbor facilities will boost the economy. Despite climate disasters, the Mozambique economy has grown remarkably. Foreign reserves increased from $195.25 million in 1995 to $652 million in 1999. FDI was $4.07 billion in 1999.

Mozambique will own the Cahora Bassa scheme by 2014, producing 4000 MW upon completion of phase II.

9.5 RE Technological

Despite favorable solar conditions and only 7% grid connection, PV penetration is only approximately 175 kW_p, including 150 clinics. A study on PV was completed by the World Bank, but the document was not accessible. SWH is mostly limited to holiday resorts. Wind and small hydro IPP potential is estimated to be considerable, but it is unexplored.

Poverty and low energy demand resulting from civil war are barriers to short-term RE development. The joining of the Trans-Boundary Game Reserve with the existing South African Kruger National Park should offer RE opportunities.

10. NAMIBIA

10.1 Natural/Physical

Namibia, which has an area of 842,269 km^2, is bordered by the Atlantic Ocean's cold Benguella Current and has a mostly arid climate with semiarid and subtropical regions. Winter sunshine reaches 6.8 kWh/m^2/day, the world's highest. The coastal strip is often misty during the night. Extreme daily temperature swings are typical. Namibia is very rich in diamonds (Sperrgebiet), other precious and semi-precious stones, ores, and uranium (Rössing). The perennial Orange and Cunene Rivers are the north and south boundaries. The Kudu gas field is situated off the coast between Oranjemund and Lüderitz.

10.2 Institutional/Political

The former German colony was taken by South African troops during World War II and became independent in 1990. The 1998 "White Paper on Energy Policy" refers to RE as an integral and equal component, with special reference to rural water supply and electrification. The restructured utility, Nampower, retains its generation and transmission asset, whereas IPPs provide new generation capacity. Similar to the proposed South African scheme, regional electricity distributors will be responsible for distribution. Namibia imports energy.

10.3 Social

The population of approximately 1.7 million amounts to only 1.7 persons/km^2, the lowest density of SADC. Approximately 27% of the population is urbanized, the literacy rate is 76%, life expectancy is 61 years, and there is a low HDI.

10.4 Economic

Namibia has a GDP/capita of $1370 and is ranked fourth in the African competitiveness rating—the third highest in SADC. Its foreign reserves increased from $221 million in 1995 to $305 million in 1999. The main GDP shares are government services (22.4%) followed by mining (16.4%) and manufacturing (11.8%), indicating an overweight government.

10.5 RE Technological

Namibia has ideal climatic conditions for RE. Domestic SWH installations are in use in government buildings. However, the 24,000 m^2 installed capacity yields a very small market penetration of 0.001 m^2/person and there remains great potential.

PV systems are applied in telecommunication, water pumping, and SHSs. Grid-feeder applications are not known. The total installed capacity is approximately 1 MW_p, with 446 kW_p in the household sector, which is expected to grow at approximately 15 kW_p/year ($100,000/year).

Wind resources of 5–8 m/s along the coast led to advanced feasibility studies for a 5- to 20-MW grid-connected wind park at Lüderitz. This was turned down by Nampower. Traditional windmills for pumping water are gradually being replaced by PV.

11. SEYCHELLES

11.1 Natural/Physical

This archipelago of 115 islands totaling 455.3 km^2 is located 4° south of the equator in the convergence area of the warm monsoon drift and the warm southern trade wind current. The southeast trade winds blow from May to September. During the rest of the year, the northwest monsoon brings tropical rainfall, higher humidity, and temperatures up to 32°C. The area lies outside the cyclone belt.

The world heritage site Vallée de Mai is located on the island of Praslin and is the home of the world-renowned coco-de-mer and the endangered black parrot.

The annual rainfall of 2238 mm is high, but 98% is lost because of the steep topography and evapotranspiration.

11.2 Institutional/Political

Seychelles law is a hybrid of civil and English common law. An energy policy developed by the Ministry of Industry and International Business: Energy Affairs Division guides development, whereas the Public Utilities Corporation: Electricity Division is responsible for electricity generation, transmission, and distribution according to the traditional vertically integrated model.

11.3 Social

Most Seychellois speak Créole, French, and English, reflecting a colonial past, including Arabs and Portuguese taking up victuals and even a period of French and English pirates from 1685 to 1730. At 87.9% (1994), the Seychellois have the highest literacy within SADC and share the highest life expectancy of 70 years with Mauritius. The HDI is 81%, and there is no competitiveness rating.

11.4 Economic

Transport, distribution, and communications (28.3%) dominate the GDP, followed by manufacturing and handicraft (14.5%), government services (13.5%), and the growing finance and business services (11.8%). Fisheries and tourism reflect the traditional income base, which is changing from centralist control to greater private sector participation. The GDP per capita is $ 7384 (1999)—the highest in the SADC region and 40 times higher than that of Malawi.

11.5 RE Technological

A total of 95% of the population has access to expensive grid electricity generated by heavy fuel oil instead of diesel. A 50-MW ($4.8 million) power plant was built at Roche Caiman, approximately 2 km from Victoria.

Seychelles Petroleum provides oil and liquid petroleum gas (LPG). There is a government campaign to phase out kerosene for household cooking in favor of LPG. The search for hydrocarbon in the Exclusive Economic Zone continues under the Seychelles National Oil Company. High electricity costs led to the installation of more than 2400 m^2 SWHs yielding a market penetration of 0.03 m^2/capita and growing by 400 m^2 ($120,000) annually.

Steady wind regimes and high fossil fuel imports should lead to the use of offshore wind turbines. Recent international experience demonstrated that grid-connected turbines feed 50% of the capacity into the grid. This translates into a wind energy potential of 25 MW.

PV SHSs are used on the islands. Biogas and biofuel are not being exploited because of a lack of awareness.

12. SOUTH AFRICA

12.1 Natural/Physical

This 1,221,000-km^2 country occupies the southern tip of the African continent between 22 and 34°S latitude. Its 2968-km coastline borders the Indian and Atlantic Oceans. Prince Edward and Marion Islands 1920 km south of Cape Town are also part of the territory. A mountain range along the southeastern coast limits precipitation in the interior, making the country semiarid with 464 mm of annual rainfall compared to the global average of 857 mm.Winter sunshine varies from 4.5 kWh/m^2/day along the south and eastern coast to 6.8 kWh/m^2/day in the northwestern inland areas. Good wind potential exists along the coasts and mountain areas.

12.2 Institutional/Political

The "Energy White Paper" was published by the Department of Minerals and Energy in 1999, and the "Draft White Paper on the Promotion of Renewable Energies Part 1" (begun in 1998 and published in Dec. 2003) while "Energy Efficiency Part 2" (begun in 1998) is expected to be published in the near future. The 1999 "White Paper" mandates breaking up of the vertically integrated electricity monopoly Eskom, which achieved the electrification of more than 2 million households since 1992. This restructuring of Eskom into the envisaged regional electricity distributors is imminent. An independent national energy regulator oversees energy licensing and tariffs, including electricity. A large rural off-grid energization drive using private sector monopolies was launched in 1999, and implementation is imminent. South Africa acceded to the Kyoto Protocol as a developing country.

12.3 Social

The population of 43 million is 30% urbanized, 82.2% (1991) literate, had a life expectancy of 66 years in 1991 that has been reduced to 47.7 years,

and has a HDI of 6.5%. Afrikaans, English, and nine official native languages are recognized as official languages. In addition, Portuguese and other European, African, and Asian languages are spoken.

12.4 Economic

Strict monetary policy is being applied in order to reduce the inflation rate. The exchange rate of the local currency (Rand) against the dollar decreased by 39.8% from 1995 to 2002, whereas the GDP per head increased by 6.4% at constant 1995 prices. Weakening of the U.S. dollar has resulted in an improved rate of 7.20 Rand per dollar. The international credit rating is 40, the global competitiveness rating is 34 out of 75, and the Africa competitiveness rating is 7 out of 24.

South Africa's energy intensity is 15%, which is nearly double the global 8%. Eskom is the world's fifth largest electricity utility (39,870 MW installed capacity), supplying 95% of the country's electricity, which is 95% coal generated.

South Africa is the world's third largest per capita CO_2 emitter, after the United States and Canada, contributing 1.7% to global emissions but only 0.4% to the global GDP.

12.5 RE Technological

RE energy has been neglected with regard to R&D and implementation. A national wind turbine demonstration site is being developed at Darling in the Cape and another test site has been started at Klipheuwel by Eskom.

A 200-MW agricultural upwind chimney delivering 24-h load-following electricity as well as agricultural produce is being considered in the Northern Cape Province. A feasibility study for a 100-MW solar thermal plant is under way by Eskom, which also erected an experimental dish sterling machine of 25 kW. That was inaugurated during the World Summit on Sustainable Development at the Development Bank of Southern Africa. The CSIR in Pretoria is investigating wave energy generators.

Currently, only the on-shore wind potential of the approximately 3000-km coastline is being assessed. If the wind energy contribution is capped at 50% of the present capacity, it amounts to approximately 20 GW, increasing to 25 GW in 2012.

The World Bank 10-year Global Environment and Local Development project is summarized in Table I.

The project includes the energy-efficient (solar-passive) design and construction of 300,000 low-income houses. Currently, there are 184,000 m^2 SWHs, with a potential for 43 million m^2 by 2020, including domestic and commercial glazed and unglazed SWHs. There are no subsidies.

TABLE I

World Bank Renewable Energy Initiative in South Africa, 2003

	Installed/saved (MW)	Annual generation (GWh)	Estimated capital cost ($M)
Commercial solar water heating	NA	690	510
Biomass cogeneration	220	1320	172
Paper and pulp waste	195	880	174
Municipal solid waste/landfill gas	125	825	145
Wind/other	60	247	56
Total	600	3962	1056

The announcement of the rural government-subsidized SHS program in 1999 comprising 250,000 units of 50 W to be built during a 5-year period created great expectations, but in early 2003 the Department of Minerals and Energy still had implementation problems. The total national unsubsidized PV capacity in 2003 was more than 8 $MWh/year_p$.

The bioenergy potential is given in Table II. The hydropower potential is 8,360 MW or 73,230 GWh/year, whereas the installed capacity was 68 MW in 2000. The potential wave energy is 56,800 MW. There is no installed capacity. The 2020 potential savings in energy-efficient buildings is 36,000 GWh/year.

Job creation is a major motivation for renewable energy application. Biodiesel can create more than 250,000 new jobs, energy-efficient buildings 250,000, PV 125,000, one solar chimney 43,000, solar water heating more than 15,000, and solar cookers 10,000. This totals more than 693,000 permanent work opportunities in a country with an estimated unemployment rate higher than 40%.

13. SWAZILAND

13.1 Natural/Physical

The Kingdom of Swaziland covers an area of 17,364 km^2 and is bounded by South Africa and Mozambique. It lies at the 26°S latitude on the

TABLE II
Bioenergy Potential[a]

Bioenergy	2003 potential (GWh/a)
Bagasse	1400
Sawmills	7639
Pulp mills	4528
Agricultural wastes, manure, litter, grass	5615
Biodiesel	3,500,000
Bioethanol	1200
Total	3,520,382

[a] From Sustainable Energy Society of Southern Africa (2002).

eastern side of the escarpment ranging from 98 to 1800 m in altitude. This causes orographic summer rains. The average annual rainfall is between 1016 and 2286 mm, which is high by southern Africa standards. Winter sunshine is more than 5.0 kWh/m^2/day, with temperatures between 5 and 19°C, whereas summer temperatures vary between 15 and 25°C.

13.2 Institutional/Political

Swaziland's legal system is based on traditional Swazi national courts and constitutional courts derived from Roman–Dutch law. A national energy policy formulation project was completed by the energy section of the Ministry of Natural Resources and Energy assisted by Danish Cooperation for the Environment and Development.

The Swaziland Electricity Board (SEB) is a government-owned utility and provides 70% of the final commercial electricity, of which 30% is generated by SEB's hydro and diesel plants. The remainder is imported from neighboring South Africa's Eskom. Sugar and pulp cogeneration plants provide the 30% of the remaining national electricity supply not delivered by SEB.

A rural electrification policy envisages "that the access to energy is made available for all by 2010" and that "access to electricity is made available for all by 2020." This includes off-grid electrification. The primary focus is on schools, health care facilities, and other essential public facilities.

13.3 Social

English is the official language, with siSwazi spoken by the majority of the 1.02 million (1999) people. The literacy rate is 78%, urbanization 34%, HDI 57.7%, and life expectancy 57 years.

13.4 Economic

The GDP per capita is $1268 (1999), with the major contributors being manufacturing (36.9%); government services (15.3%); wholesale, retail, hotels, and restaurants (10%); and agriculture (9.8%). More than 80% of commodities, including petroleum products, come from South Africa. The African competitiveness ranking is 10 out of 24.

13.5 RE Technological

Renewable energy from bagasse and pulp cogeneration comprises a significant 30% of delivered energy. SWH is underdeveloped. PV systems are applied in an estimated total capacity of 50 kW_p, with approximately 400 domestic lighting systems (16 kW_p) and a growth of approximately 12 kW_p per year. With a 10% market penetration of stand-alone systems, the potential market is 600 kW_p. Two small hydro systems are in operation.

14. TANZANIA

14.1 Natural/Physical

This country is a merger of the former Tanganyika and Zanzibar, which comprises the Ungiya and Pemba Islands in the Indian Ocean. In the west, it extends to Lakes Nyasa and Tanganyika and in the north to Lake Victoria. The world's largest crater, the Ngorongoro, rises to 3647 m.

The land lies between 1 and 11°S of the Equator and covers an area of 945,200 km^2. The floor of Lake Tanganyika is part of the Great Rift Valley and is Africa's lowest point at 358 m below sea level.

There are four main climatic zones: the hot and humid coastal plains; the hot and arid central plateau; the high, moist lake regions; and the temperate highlands. The northeast monsoon brings the hottest temperatures from December to March. Shifting south in March to May, it carries heavy intermittent rains. The southwest monsoon brings cooler, dry weather from June to September. The short wet season is interspersed in November and December. Annual rainfall is 1250 mm on the southeast slopes of the volcanoes and less than 600 mm in the central plateau.

14.2 Institutional/Political

Tanzania gained independence from British rule in 1961. The United Nations Development Program helped to develop the "National Framework for Renewable Energy Development" as part of the National Program for Renewable Energy. Tanzania is one of the few SADC countries with a national electricity regulator and an IPP structure.

14.3 Social

The estimated population is 32 million (1999), of which 30% is urbanized and 84% literate. The life expectancy is 50 years, with a HDI of 44%. KiSwahili is the official language. English is the business language mostly spoken by the 3% of urban Asians. Religions are Muslim, Christian, and traditional in approximately equal percentages. HIV/AIDS is a concern.

14.4 Economic

Tanzania has an old tradition in trading predating Arab slave, gold, ivory, and copper trading, which was eventually taken over by the British. Currently, agriculture accounts for 57% of the GDP, grown on arable land limited to 4% of the total area. The GDP per capita is $242 (1999), and the Africa competitiveness rank is 7 out of 24.

14.5 RE Technological

There are gas fields (33 billion m^3) at Songo-Songo and 1200 million tons of coal at Ufipa, Galula, Songwe, Ruhuhu, Songea, and Mkusi. Informal fuelwood provides 90% of all energy needs, petroleum provides 9%, and hydroelectricity and coal provide the balance. The SWH market is underdeveloped.

Approximately 500 kW_p of PV has been installed in clinics and schools in rural and urban areas. Approximately 60 small hydro plants (30 MW_p) have been installed, with an additional 50 identified. Approximately 1000 biogas plants have been installed, representing the most ambitious program in SADC. The follow-up is uncertain.

15. ZAMBIA

15.1 Natural/Physical

Zambia is landlocked in a tropical belt on a plateau at 1300 m covering 752,612 km^2. Summer temperatures range from 27 to 32°C, whereas the winter range is 16–27°C. Solar radiation in winter is more than 5.0 kWh/m^2/day, rendering it suitable for solar applications. The mighty Zambezi and Lake Kariba form the southern boundary, where it can be very hot during October. Annual rainfall during the warm, humid season from November to April is 508–1270 mm.

15.2 Institutional/Political

Zambia's borders are the result of Cecil Rhodes imperialism, which fell apart in 1964. In 1972, it became a one-party state until general elections were held in late 1991. The 1994 energy policy enjoys political support. There is a national electricity regulator and a reformed electricity system. A grid and non-grid electrification program is in place. Zambia has by far the best energy-efficiency organization in the SADC.

15.3 Social

Zambia comprises approximately 10 million people (1999) of the most diverse cultures who speak more than 70 dialects and have diverse lifestyles, including fishing, cultivating, and ranching. Christians comprise 80% of the population, and the rest believe in Traditional, Islam, Hinduism, and Buddhism. The literacy rate is 79%, and 38% of the population is urbanized, representing the highest percentage in the continental SADC. The life expectancy is a 46 years, and the HDI is 43.3%.

15.4 Economic

Zambia is the world's fourth largest copper producer and owns 6% of the known global reserves. Copper accounted for 50% of the country's export earnings in 1998. Zambia also produces more than 20% of the world's emeralds. Although the country received debt relief up to 67%, the debt burden remained unsustainable. The GDP per capita was $302 in 1999, and the African competitiveness ranking is 12 out of 24.

15.5 RE Technological

Zambia produces 99% of its electricity by cheap hydropower used mainly in the mines, whereas only 20% of the population has access to grid electricity. Proven coal reserves are 30 million tons. Petroleum is wholly imported, draining approximately 10% of foreign exchange.

SWH has a minor role. Small-scale, stand-alone PV systems are in use with an estimated capacity of 1 MW. More than 20 small hydro systems have been installed and their number is increasing. IPPs offer good opportunities.

16. ZIMBABWE

16.1 Natural/Physical

Zimbabwe (390,575 km^2) shares Lake Kariba with Zambia and is situated between the Limpopo and Zambezi Rivers, wholly within the tropics between 15 and 22°S at a high altitude of 800–1500 m above sea level. The average rainfall is between 196 mm in summer and 10 mm in winter. Most of Zimbabwe receives more than 5 kWh/m^2/day in winter, representing a very good climate for solar applications.

16.2 Institutional/Political

Comrade President Mugabe is a prominent proponent of the World Solar Summit initiated by UNESCO. A major Global Environmental Fund PV program was launched in Zimbabwe. A rural electrification initiative is cross-subsidized by a levy on all electricity sales. Electricity is imported on a large scale from neighboring countries.

16.3 Social

Zimbabwe estimated its literacy rate to be 80% and its urbanization to be 31%. Life expectancy is 50 years, whereas the HDI is 55.1%. There is widespread starvation.

16.4 Economic

The GDP per capita is $505 (1999), whereas the Africa competitiveness rating is 23 out of 24, and it was ranked 75 out of 75 according to the Global Competitiveness Report (2001/2002). The 2003 inflation rate was 400%.

16.5 RE Technological

Most of the 0.75-MW_p PV installations were donor funded, an activity that destroyed private initiative and was reduced for political reasons. Domestic SWHs used to be installed at a rate of 1000 m^2 ($220,000) per year. No current statistics are available. Zimbabwe's political situation impacts the whole SADC.

SEE ALSO THE FOLLOWING ARTICLES

Biomass for Renewable Energy and Fuels • Computer Modeling of Renewable Power Systems • Development and Energy, Overview • Energy Ladder in Developing Nations • Potential for Energy Efficiency: Developing Nations • Renewable Energy and the City • Renewable Energy in Europe • Renewable Energy in the United States • Renewable Energy Policies and Barriers • Renewable Energy, Taxonomic Overview • Renewable Portfolio Standard

Further Reading

Development Bank of Southern Africa/International Solar Energy Society (2000). "Renewable Energy Technologies in Southern Africa—A Guide for Investors." Midrand, Freiburg, Germany.

Makenzie-Hoy, T. (2003). The DRC hydropower dream deconstructed. *Eng. News*, August 1–7, 56.

Mohamed, S. (2003). Nepad and global governance. *Eng. News,* August 1–7, 54.

Powering Nepad (2003). *Eng. News*. August. Available: http://www.engineerignews.co.za/eng/sector/energy/?show = 38166.

Schramm, G. (1998). Electrification programmes and the role of international development banks. *In* "ISES Utility Initiative for Africa" (D. Holm and W. Berger, Eds.), p. 107.

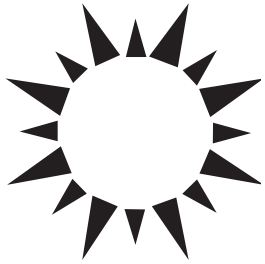
Shell International (2002). "Shell Scenario." Shell International, London.

Solarex Corporation. (1992). World design insolation [map]. Solarex Corp., Frederick, MD.

Southern African Development Community (2002). SADC review. Available: http://www.sadcreview.com.

Sustainable Energy Society of Southern Africa (2002). "Comments to the Draft White Paper on the Promotion of Renewable Energy and Clean Development." Sustainable Energy Society of Southern Africa, La Montagne, South Africa.

Renewable Energy in the United States



JOHN CARLIN
United States Department of Energy
Washington, D.C., United States

1. **Introduction**
2. **Overview of Renewable Energy**
3. **Biomass Energy**
4. **Hydroelectric Energy (Hydropower)**
5. **Geothermal Energy**
6. **Wind Energy**
7. **Solar Energy**

Glossary

biomass integrated gasification combined cycle A composite system that converts biomass feedstock into gas fuel for an electric generating unit that consists of one or more gas turbines with a portion of the required energy input provided by exhaust heat of the turbine to increase efficiency.

closed-loop biomass A process in which biomass feedstocks are grown specifically for the purpose of energy production. During the growing process, the feedstock absorbs CO_2, offsetting the forthcoming CO_2 emissions to be realized during energy use.

cofiring generation The use of a supplemental fuel in a boiler in addition to the primary fuel the boiler was originally designed to use.

criteria pollutant A pollutant determined to be hazardous to human health and regulated under the Environmental Protection Agency's (EPA) National Ambient Air Quality Standards. The 1970 amendments to the Clean Air Act require the EPA to describe the health and welfare impacts of a pollutant as the "criteria" for inclusion in the regulatory regime.

district heat Steam or hot water from an outside source used as an energy source in a building. The steam or hot water is produced in a central plant and piped into the building. The district heat may be purchased from a utility or provided by a physical plant in a separate building that is part of the same facility (e.g., a hospital complex or university).

greenhouse gases Those gases, such as water vapor, carbon dioxide, nitrous oxide, methane, hydrofluorocarbons, perfluorocarbons, and sulfur hexafluoride, that are transparent to solar (short-wave) radiation but opaque to long-wave radiation, thus preventing long-wave radiant energy from leaving the earth's atmosphere. The net effect is a trapping of absorbed radiation and a tendency to warm the planet's surface.

hot dry rock Heat energy residing in impermeable, crystalline rock. Hydraulic fracturing may be used to create permeability to enable circulation of water and removal of the heat.

landfill gas Natural gas (primarily methane gas) produced as a result of decaying waste at municipal landfills.

one sun Natural solar insulation falling on an object without concentration or diffusion of the solar rays.

renewable energy Resources that are naturally replenishing but flow limited. They are virtually inexhaustible in duration but limited in the amount of energy that is available per unit of time. Renewable energy resources include biomass, hydroelectric, geothermal, wind, and solar. Alternate transportation fuels include liquefied petroleum gas, ethanol, methanol, natural gas, electricity, and biodiesel.

run-of-river hydroelectric plant An electric generating plant using the power of moving water as it occurs and having little or no reservoir capacity for storage.

Interest in the use of renewable energy in the United States has increased during the past few decades beginning with the oil embargo of the early 1970s. Renewable energy is used in the residential, commercial, industrial, and transportation sectors. It is used to produce electricity at central grid-connected generators and at dispersed locations; cogeneration primarily in the industrial sector; heat in the industrial, residential, and commercial sectors; and ethanol for vehicle propulsion. During this time period, the U.S. energy policy has evolved around

two central themes: (i) energy/economic security and (ii) resource and environmental sustainability.

1. INTRODUCTION

Energy/economic security refers to the reliability of energy supplies and the stability of energy prices, which significantly influence individual and business expenditures and investment decisions. In the past, disruptive energy supplies and volatile energy prices have detrimentally affected economic growth and made decisions to invest in additional energy supplies and new energy technologies more difficult and risky. Renewable energy has the potential to mitigate these effects. Its prices are not volatile, and once the capital is invested, the fuel prices are stable and in most cases negligible. Any reduction in the demand for conventional fuels reduces the price of that conventional fuel for all units consumed. Even if a new technology does not actually penetrate the market, its availability, on the shelf, limits price levels for conventional fuels. Renewable energy also enhances energy diversification, reduces capital outflows, and utilizes domestic labor and resources.

Renewable energy also plays an important role in reducing emissions from fossil fuels when used in their place. These emissions include greenhouse gases and criteria pollutants such as sulfur oxide (SO_x) and nitrous oxide (NO_x). Renewable fuels, with the exception of biomass, have little or no emissions. The use of closed-loop biomass reduces carbon emissions, a very important component of greenhouse gases. Cofiring of biomass and coal to produce electricity may also reduce the emission of SO_x and NO_x.

The use of renewable energy also has some disadvantages, principally related to cost. Renewable technologies are usually more expensive than conventional technologies and frequently require larger up-front capital investments. Also, there are land-use and other environmental issues related to their use.

The following are the most important renewable fuels in use today:

Hydroelectric power (hydropower): The use of flowing water to produce electrical energy.

Biomass/biofuels: Organic matter that includes forest thinnings, forest and mill residues, agricultural by-products and crops, derived oils and wastes, wood and wood wastes, animal wastes, livestock operation residues, aquatic plants, fast-growing trees and plants, and municipal and industrial wastes. Biofuels refers to liquid fuels derived from biomass feedstocks.

Geothermal energy: Hot water or steam extracted from geothermal reservoirs in the earth's crust that can be used for geothermal heat pumps, space and district heating, water heating, or electricity generation.

Wind energy: Kinetic energy present in wind motion that can be converted to mechanical energy for driving pumps, mills, and electric power generators.

Solar energy: The radiant energy of the sun, which can be converted into other forms of energy, such as heat or electricity.

This article presents basic information on renewable energy including a cursory discussion of its principal elements. The focus is on providing data for 2001 while explaining the role of renewable technologies in the U.S. energy marketplace.

2. OVERVIEW OF RENEWABLE ENERGY

The use of renewable energy in the United States peaked during the mid-1990s at approximately 7.5 quadrillion Btus (quads), increasing from slightly more than 4 quads in 1970 (Fig. 1). Almost half of this increase occurred in the electric power sector and another one-fourth occurred in the industrial sector. The remaining growth occurred in the residential, commercial, and transportation sectors and net imports of renewable energy. Hydropower, biomass, and geothermal energy account for most of the growth.

Since the mid-1990s, however, renewable energy consumption has declined approximately 25% due to the reduced water availability for hydropower. In 2001, renewable energy consumption amounted to 5.668 quadrillion Btus (Fig. 2), which represents 6% of total U.S. energy consumption. Even with the decline in the availability of hydropower, biomass and hydropower still accounted for 92% of total renewable energy consumption.

In 2001, more than 60% of total renewable energy consumed (3.5 quads) was used to produce electricity (Table I). In addition, the industrial sector used 1.4 quads for nonelectric purposes (process heat and steam). The renewable energy use in the industrial sector was primarily a wood or wood waste feedstock. The transportation sector consumed only 0.1 quads of renewable energy

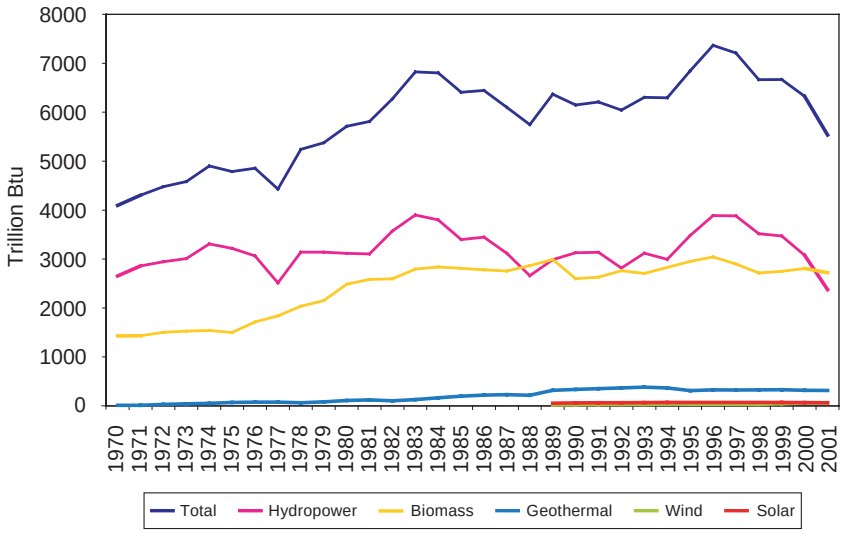


FIGURE 1 Growth in renewable energy consumption by source. From Energy Information Administration (2002). Annual energy review, 2001. Energy Information Administration, Washington, DC.

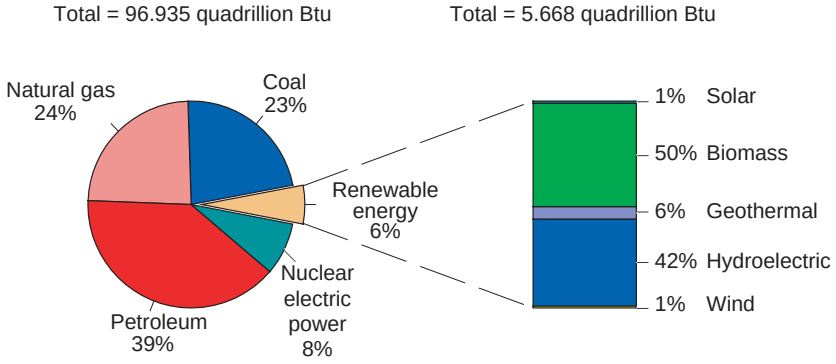


FIGURE 2 The role of renewable energy consumption in the nation's energy supply, 2001. From Energy Information Administration (2002, Fig. H1).

(ethanol). Although the use of ethanol in gasoline has doubled during the past decade, it represents only approximately 0.5% of total transportation fuel consumption.

The increased use of renewable technologies, other than hydropower, during the latter quarter of the 20th century is largely the result of government incentives and regulations (Table II). The federal government provided financial incentives and funds for research and development (R&D) during this period. By 1982, most renewable energy projects were eligible for a 10% investment tax credit, a 15% business renewable energy investment tax credit, a 40% residential tax credit, and a 5-year accelerated depreciation schedule. However, the most significant impetus to the evolution of the renewable industry was the passage of the Public Utility Regulatory Policies Act of 1978 [Public Law (P.L.) 96-917].

3. BIOMASS ENERGY

The Energy Information Administration's (EIA) current data accounts divide biomass/biofuels into four categories: wood and wood waste, municipal solid waste, other biomass waste, and corn ethanol.

Wood and wood waste come primarily from U.S. forestlands or timberlands. Other sources include private land clearing and silvaculture, and urban tree and landscape residues. Wood is almost entirely derived from U.S. timberlands, which contain round-wood, timber stock consisting of 92% sound trees

TABLE I

U.S. Renewable Energy Balances, 2001[a]

Sector and source	Energy consumption (quadrillion Btus)			Electricity net generation (thousand kWh)	Electricity net summer capacity (MW)
	Electric	Nonelectric	Total		
Total	3.541	2.127	5.668	312,767,216	96,741
Biomass	0.808	2.046	2.854	59,640,051	10,120
Geothermal	0.292	0.022	0.315	13,901,229	2793
Conventional hydroelectric	2.376	—	2.376	232,949,965	79,379
Solar	0.005	0.059	0.064	494,158	387
Wind	0.059	—	0.059	5,781,813	4062
Residential		0.475	0.475		
Biomass		0.407	0.407		
Geothermal		0.009	0.009		
Solar		0.059	0.059		
Commercial	0.027	0.072	0.098	2,063,254	433
Biomass	0.026	0.063	0.089	1,963,505	399
Geothermal		0.008	0.008		
Conventional hydroelectric	0.001		0.001	99,749	35
Industrial	0.402	1.414	1.816	32,361,740	5667
Biomass	0.365	1.409	1.774	28,739,925	4588
Geothermal		0.005	0.005		—
Conventional hydroelectric	0.037		0.037	3,621,815	1079
Transportation					
Biomass (alcohol fuels)		0.133	0.133		
Electric power	2.953	0.034	2.987	262,853,221	90,640
Biomass	0.417	0.034	0.451	28,936,621	5134
Geothermal	0.290		0.290	13,812,908	2793
Conventional hydroelectric	2.181		2.181	213,827,721	78,265
Solar	0.005		0.005	494,158	387
Wind	0.059		0.059	5,781,813	4062
Net renewable energy imports	0.159		0.159	15,489,001	
Geothermal	0.002		0.002	88,321	
Conventional hydroelectric	0.241		0.241	23,610,560	
Conventional hydroelectric (exports)	0.084		0.084	8,209,880	

[a] *Source*. Energy Information Administration. (2002).

(called growing stock) and 8% rotten, cull, or "salvable" dead trees.

Pure wood for fuel (known as fuelwood) actually accounts for a very small percentage of biomass energy consumption. The majority is wood waste. Waste wood includes principally manufacturing and wood processing wastes from wood mills. Primary wood mills include sawmills, veneer mills, and pulp mills. Secondary mills include manufacturers of dimension lumber, trusses and building components, flooring, windows and doors, cabinets, pallets, poles and fencing, barrels, boats, highway transport trailers, manufactured homes, and musical instruments. Mill waste residues are used primarily as fuel,

TABLE II

Key Milestones for Renewable Fuels

Year	Milestone	Description
1973	OPEC oil embargo	OPEC disrupted oil supplies and raised crude oil prices by 70%.
1974	Oil embargo ends	Project independence was Initiated to review strategic energy options.
1978	Public Utility Regulatory Policies Act of 1978 (PURPA)	PURPA mandated the purchase of electricity at the utilities "avoided cost" from qualifying facilities (QFs) meeting certain technical standards regarding energy source and efficiency. QFs were also exempted from both state and federal regulation under the Federal Power Act and the Public Utility Holding Company Act.
1978	Energy tax credit	A 15% energy tax credit was added to an existing 10% investment tax credit, providing incentive for capital investment in photovoltaic applications and solar thermal generation facilities for independent power producers.
1978	Energy tax act	The Energy Tax Act of 1978 (Public Law 95-618) gave a ¢ per gallon exemption from federal excise taxes to motor fuels blended with ethanol (minimum 10% ethanol) and granted a 10% energy investment tax credit for biomass–ethanol conversion equipment (in addition to the 10% investment tax credit available).
1980	Crude Oil Windfall Profits Act	The act (Public Law 96-223) extended the per gallon federal excise tax exemption for ethanol blends to December 31, 1992, and extended the energy investment tax credit to December 31, 1985. An income tax credit was also provided to alcohol fuel blenders—40 ¢ per gallon for 190-proof alcohol and 30 ¢ per gallon for 150–190 proof. The total federal tax credit for a turbine was increased to 25%. The act also raised the residential tax credit to 40% of the first $10,000 investment in photovoltaic applications.
1980	Energy Security Act	The act (Public Law 96-294) offered insured loans for small ethanol producers (less than 1 million gallons per year), loan guarantees that covered up to 90% of construction costs on ethanol plants, price guarantees for biomass energy projects, and purchase agreements for biomass energy used by federal agencies. It also established the Department of Energy (DEO) Office of Alcohol Fuels and authorized $600 million for both the U.S. Department of Agriculture (USDA) and the DOE for biomass research. Subsequent recisions altered this allocation to $20 million for the USDA and $800 million for the DOE to use for alcohol fuel loans. The Consolidated Farm and Rural Development Act of 1980, which rescinded the $505 million allocated to the USDA, appropriated $250 million for alcohol loan guarantees that were used to support 12 firms.
1981	California State energy tax credit	The state of California enacted a 25% tax credit for the capital costs of renewable energy systems.
1983	California Standard Offer Contracts	California's Standard Offer Contract system for QFs provided renewable electric energy systems a relatively firm and stable market for their output.
1984	Tax Reform Act	The act (Public Law 97-424) raised the gasohol exemption from 5 to 6 ¢ per gallon, with the overall tax unchanged at 9 ¢ per gallon of retail fuel. The blender's income tax credit was increased to 60 ¢ per gallon for the 190-proof alcohol and 45 ¢ for the 150–190 proof.
1990	Omnibus Budget Reconciliation Act	The act (Public Law 101-508) decreased the gasohol tax exemption from 6 to 5.4 ¢ per gallon. Tax credits for neat

continues

Table II continued

Year	Milestone	Description
		ethanol sales remained unchanged at 6¢ per gallon. The expiration date was extended to 2002.
1990	Clean Air Act Amendments	The amendments (Public Law 101-549) mandated the winter use of oxygenated fuels in 39 major carbon monoxide nonattainment areas (areas where Environmental Protection Agency emissions standards for carbon dioxide had not been met) and required year-round use of oxygenates in 9 severe ozone nonattainment areas in 1995. The 1990 amendments also establish the Acid Rain Program, which is designed to reduce the adverse effects of acid deposition. This improvement will be achieved primarily through reductions of SO_2 and NO_x emissions by electricity producers.
1992	Energy Policy Act	The act (EPACT) (Public Law 102–486) established a permanent 10% tax credit for investments in solar and geothermal equipment. Established a 10-year, 1.5¢ per kilowatthour (kWh) production tax credit for privately owned as well as investor-owned wind projects and biomass plants using dedicated crops (closed loop) brought on line between 1994 and 1993, respectively, and June 30, 1999. The act instituted the Renewable Energy Production Incentive, which provides a 1.5¢/kWh incentive, subject to annual congressional appropriations (section 1212), for generation from biomass (except municipal solid waste), geothermal (except dry steam), wind, and solar from tax exempt publicly owned utilities and rural cooperatives. Indefinitely extended the 10% business energy tax credit for solar and geothermal projects. It provides a tax credit (variable by gross vehicle weight) for dedicated alcohol-fueled vehicles, a limited tax credit for alcohol dual-fueled vehicles, and, a tax deduction for alcohol fuel dispensing equipment.
1995	Federal Energy Regulatory Commission (FERC) prohibition on QF contracts above avoided cost	In a ruling against the California Public Utility Commission, FERC refused to allow a bidding procedure that would have the effect of allowing rates above avoided cost from renewable QFs.
Mid-1990s	ISO4 contract Rollover in California at lower rates	Ten-year QF contracts written during the mid-1980s at rates of 6¢/kWh and higher began rolling over at mid-1990s avoided costs of approximately 3¢/kWh. This "11th-year cliff" creates financial hardship for most QFs on ISO4 contracts.

with the next most important application being use as pulp and fiber for making paper products. The paper production process, in turn, yields a by-product known as black liquor, which is the single largest source of biomass energy. Pulp and paper mills use large quantities of wood bark, edgings, and residues from their own log-stripping operations and also buy them from facilities such as sawmills.

Waste wood also includes construction and demolition debris, wood from pallets and containers, and wood pellets. Waste wood is also salvaged for fuel by private owners from woodlots, farm fence rows, cropland clearing, orchards, and operations of private urban silvaculture. Urban tree and landscape residues consist of tree limbs, tops, brush, leaves, stumps, and grass clippings. They are generated by commercial tree care firms, municipal tree trimming operations, electric utility power line maintenance departments, orchards, and landscapers.

Municipal solid waste (MSW) includes durable and nondurable goods; containers and packaging; food wastes; yard wastes; and miscellaneous waste from residential, commercial, institutional, and some industrial sources. Examples of MSW include yard trimmings (lawn clippings and branches), newspapers and magazines, office paper, boxes and packaging, glass, metal, plastics, rubber, clothing, and food scraps and cafeteria wastes. MSW also includes such large and durable items as furniture, wood pallets, and appliances. MSW management begins with methods of reducing the quantities of trash generated in the first place (source

reduction). However, once the trash has been produced, MSW management involves four components: composting, recycling, waste-to-energy via combustion, and landfilling.

The last two categories of biomass/biofuels are relatively small. Other biomass waste is a catch-all category that includes agricultural by-products and crops; sludge waste; tires; and other biomass solids, liquids, and gases. Biofuels consists of liquid biomass (primarily ethanol from corn) used in the transportation sector.

3.1 Current Data

In 2001, biomass/biofuels contributed more than 2.8 quads to U.S. energy accounts (Table III). More than 2 quads of this energy was used to produce steam or process heat for the industrial sector. The remaining 0.8 quads was used to produce approximately 60 billion kilowatt-hours (kWh) of electricity generation. More than half of the electricity generation came from the electric power sector, with most of the remainder resulting from cogeneration in the industrial sector. There was a small amount of cofiring with coal in the electric power sector.

The largest fuel category was wood/wood waste, comprising more than 75% of the total account balance. Nine products represent the wood/wood waste category: black liquor, red liquor, sludge wood, spent sulfite liquor, peat, paper pellets, railroad ties, utility poles, and wood/wood waste. The second largest contributor was the use of MSW at waste-to-energy combusting facilities and landfills. Other biomass, which consists of agricultural by-products/crops, straw, sludge waste, tires, medical wastes, solid by-products, fish oil, liquid acetonitrite waste, tall oil, waste alcohol, digester gas, and methane, amounted to 117 trillion Btus. The transportation sector, consisting of ethanol from corn, contributed only 133 trillion Btus to the data accounts.

TABLE III

U.S. Biomass/Biofuels Energy Balances, 2001[a]

Sector and source	Energy consumption (quadrillion Btus)			Electricity net generation (thousand kWh)	Electricity net summer capacity (MW)
	Electric	Nonelectric	Total		
Biomass	0.808	2.046	2.854	59,640,051	10,120
Alcohol fuels	—	0.133	0.133		
Wood/wood waste	0.483	1.687	2.170	36,871,734	6230
MSW/landfill gas	0.292	0.142	0.434	20,018,830	3387
Other biomass	0.033	0.084	0.117	2,749,487	504
Residential					
Biomass		0.407	0.407		
Wood		0.407	0.407		
Commercial					
Biomass	0.026	0.063	0.089	1,963,505	399
Wood/wood waste	0.000	0.043	0.043	19,523	7
MSW/landfill gas	0.021	0.019	0.040	1,539,085	299
Other biomass	0.005	0.001	0.006	404,897	96
Industrial					
Biomass	0.365	1.409	1.774	28,739,925	4588
Wood/wood waste	0.354	1.227	1.580	27,735,132	4388
MSW/landfill gas	0.001	0.103	0.104	61,286	21
Other biomass	0.011	0.079	0.090	943,507	179
Transportation					
Biomass (alcohol fuels)		0.133	0.133		
Biomass	0.417	0.034	0.451	28,936,621	5134
Wood/wood waste	0.130	0.011	0.140	9,117,079	1835
MSW/landfill gas	0.271	0.020	0.290	18,418,459	3070
Other biomass	0.017	0.004	0.021	1,401,083	229

[a] *Source*. Energy Information Administration (2002).

3.2 Technology and Applications

Biomass combustion in the electric power, industrial, and residential sectors is the principal energy use of biomass today. To a much lesser extent, biomass can be transformed into a gas such as landfill gas, and biofuels can be created for use in the transportation sector. Biomass-fired combustion systems are commercially available to produce steam, electricity, and heat in various applications in the industrial, electric power, and residential sectors. Direct combustion systems in the industrial and electric power sectors are usually classified according to the way the biomass is burned: pile, grate, suspension, or fluidized bed.

A Dutch oven is an example of a pile-burning system. Two chambers are used for combustion: The first chamber dries and partially burns the fuel, and the second chamber completes the burning process. Some versions of the Dutch oven design include grates that facilitate drying and transport of the fuel and cofiring capability. The spreader-stoker is an example of a grate-burning system. After the fuel is spread on a grate, it is blown into a combustion chamber and burnt in suspension. Any unburnt fuel falls onto a grate, where combustion is completed. In suspension burners the fuel is burnt in an air stream. Examples of suspension burners are cyclonic burners and solid fuel injection burners. A heated bed of sand kept in constant motion is the surface for fuel combustion in a fluidized-bed combustion system.

An advanced biomass combustion technology is the biomass integrated gasification combined-cycle generating plant. It is assumed that hot gas filtration will be used for gas cleanup purposes in this technology. The advantages of the newer hot gas cleanup technology are that it generates less wastewater and it allows the process to be operated at higher efficiencies than those of the cold gas cleanup alternative. The disadvantages are limited operational experience, higher costs, and greater complexity.

MSW is combusted into energy throughout the country, producing steam, electricity, or both. Two basic approaches are used: mass burning and refuse-derived fuel (RDF). The two most common types of mass-burn facilities are waterwall incinerators and the smaller modular combustion units. At mass-burn facilities, residential and commercial waste is deposited into a pit and moved to the furnaces with overhead cranes with little sizing or separation. The wastes are usually burned on a grate and the heat is used to create steam, which can be sold directly to industrial or commercial customers or used to power a turbine to generate electricity. Waterwall furnace designs burn waste, heating the water that circulates through the walls of the combustion chamber in steel tubes, resulting in the production of steam. Modular mass-burn facilities are small-scale combustion units that can be fabricated at the factory and shipped fully assembled or shipped in modules and assembled at the construction site.

The first stage of a RDF system is the preparation of the fuel. Raw MSW is processed to form a more homogeneous product by removing noncombustible materials, such as metal, glass, grit, and sand. Sometimes, the remaining RDF is pelletized. It may be sold to outside parties or burned on-site in a furnace. In some cases, RDF is cofired with coal.

The residential/commercial sector uses five types of wood-burning equipment: free-standing stoves, fireplaces, fireplace inserts, central heating equipment for houses, and large boilers for apartment buildings and commercial buildings. Each type is produced in different sizes, has different performance parameters, and is made from a variety of construction materials (steel, iron, and alloys). Fireplaces are the most common type but are rarely used for heating an entire house because of their low efficiency and rapid consumption of wood. Many fireplaces result in a net heat loss because they require a large amount of air from the house that is replaced with cold outside air. Improvements to fireplaces can be made by adding enclosures, heat blowers, an outside air source, or fireplace inserts. Free-standing stoves are the main type of wood-burning equipment used for home heating. The two basic woodstove designs are the radiating and the circulating stoves.

Biomass (i.e., MSW in landfills) is also a source of landfill gas (LFG) that can be recovered and the energy can be used for direct heating, electricity generation, chemical feedstock, purification to pipeline-quality gas, and heat recovery. Each of these methods has a variety of LFG applications:

1. Direct heating applications
   - Industrial boilers
   - Space heating and cooling
   - Industrial heating/cofiring

2. Electricity generation applications
   - Processing and use in reciprocating internal combustion engines
   - Processing and use in gas and steam turbines
   - Processing and use of fuel cells

3. Feedstock in chemical manufacturing processes
   - Conversion to methanol (and optional subsequent industrial or vehicular fuel use)
   - Conversion to diesel fuel (and subsequent use as vehicular fuel)

4. Purification to pipeline-quality gas
   - Utilization as vehicular fuel
   - Incorporation into local natural gas network

5. Heat recovery from landfill flares

The primary approach for converting biomass to alcohol fuels is the fermentation of corn into ethanol. The largest potential source of ethanol in the United States is the biochemical conversion of cellulosic biomass into ethanol.

3.3 Niche Markets and Outlook

The conversion of biomass into energy is in large part a waste disposal option. The waste is already available, close by, and must be disposed of. In these cases, the amount of waste available is a constraint to growth in the biomass energy industry. In addition, there are several existing and potential niche markets for biomass energy as a means of alleviating environmental and security concerns associated with the use of fossil energy.

There are approximately 30 facilities that cofire biomass and coal in existing coal-fired boilers. Coal-fired boilers can combust a premixed combination of biomass and coal fed through an existing coal-feed system or retrofitted with a separate feed system for biomass. Generally, the proportion of biomass consumed ranges from 1 to 10%. The substitution of biomass for coal at electric power plants reduces emissions of SO_2 and NO_x.

A second niche market that is growing very rapidly is the conversion of LFG into energy. Several factors have contributed to this growth: U.S. Supreme Court rulings that eliminated flow control contracts that previously directed the waste to in-state waste-to-energy combustion facilities, federal tax credits, and environmental benefits. The environmental advantages of energy recovery include decreased emissions of methane, nonmethane organic compounds, and toxics (e.g., benzene, carbon tetrachloride, and chloroform). Although carbon dioxide (CO_2) emissions increase with the energy recovery option, the net atmospheric balance is a positive one because CO_2 emissions are significantly less radiative (i.e., the alleged greenhouse effect is less) than methane emissions.

Another possible niche market application for biomass is in the transportation sector. As mentioned previously, ethanol is the only renewable fuel used in substantial quantities in the transportation sector. Ethanol is made from corn. During the past two decades, it has received an excise tax exemption, generally 5 or 6¢ per gallon on a 10% ethanol/90% gasoline blend. Even with a subsidy, methyl tertiary-butyl ether was viewed until recently as the more economic and practical choice as an oxygenate additive to gasoline or as a component of RFG.

Ethanol can also be the primary component in a transportation fuel, such as E85 (85% ethanol/15% gasoline). Despite the fact that there are approximately 3 million vehicles on the road that can use any combination of E85 and regular gasoline, only 0.7% of all ethanol is E85. This reflects the difficulty in developing an E85 refueling infrastructure, and the fact that E85 on a heating value basis has generally been at least as costly as gasoline. Should a less costly source/method of ethanol become available (e.g., lignocellosic biomass conversion and biomass gasification), fleets capable of centralized refueling would provide a major entry point for E85 into the transportation sector.

For biomass to provide large amounts of energy, dedicated crops must be developed and used for electricity generation and as an alternative to gasoline in the transportation sector. To the extent that biomass is closed-loop, its use as a substitute for fossil fuel results in a reduction in CO_2 emissions, the most significant and most difficult to harness of the greenhouse gases. The use of biomass as a substitute for gasoline could potentially mitigate energy/economic concerns associated with the import of oil from the Organization of Petroleum Exporting Countries.

4. HYDROELECTRIC ENERGY (HYDROPOWER)

Water resources that are used for power production are found in 48 of the 50 states. Generating electricity using water has several advantages. The major advantage is that water is a source of relatively inexpensive power. In addition, because there is no fuel combustion, there is little air pollution in comparison with fossil fuel plants and limited thermal pollution compared with nuclear plants. Like other energy sources, the use of water for

generation has limitations, including environmental impacts caused by damming rivers and streams, which affects the habitats of the local plant, fish, and animal life.

4.1 Current Data

As of 2001, conventional hydroelectric net summer capacity totaled 79,379 MW, which generated 232,949,965 thousand net kWh of electricity. Of that total, 92% was generated by electric utilities and independent power producers in the electric power sector, followed by generation in the industrial and commercial sectors, respectively.

4.2 Technology and Applications

There are two types of conventional hydroelectric plants: storage and run-of-river. Storage plants consist of a dam, which impedes water and regulates its flow. These plants store up to a season's supply of water. Run-of-river plants have very small impoundment, storing a day's supply of water. Source water may be from free-flowing rivers, streams, or canals. In both types of plants, the water is guided to a turbine by a penstock (tube), which in turn drives a generator to produce electricity. The amount of generation is dependent on precipitation. This is especially true for run-of-river plants, which have very small impoundment. Hydroelectric facilities can be used when demand is high because hydro units can respond to rapid load changes, which is not possible with steam-powered thermal facilities.

Challenges faced by the industry include mandatory relicensing for investor-owned facilities, improving operating efficiency, and mitigating adverse environmental effects. The U.S. Department of Energy, through its Advanced Hydropower Turbine System program, is working to improve operating efficiency and will assist in addressing some of the environmental issues. The program is expected to be completed by 2010.

4.3 Niche Markets and Outlook

The relicensing process is placing new conditions on hydroelectric operations and, therefore, productivity. Generally, relicensed plants (i.e., the current stock) will have slightly decreased generation. Although most good sites have already been developed, it was estimated in 1998 that there were 5677 sites with a total undeveloped capacity of approximately 30,000 MW available for power production, after accounting for environmental, legal, and institutional constraints. This total includes approximately 8500 MW where no facility exists, approximately 17,000 MW where a dam exists but no hydroelectricity is generated, and approximately 4300 MW of incremental hydropower potential at existing sites from capacity additions or improved efficiencies. Finally, there is some, as yet not quantified, amount of capacity that could be added by refurbishing existing plants.

5. GEOTHERMAL ENERGY

Geothermal energy is heat in the earth's crust brought to the surface from underground reservoirs of steam or hot water. As used at electric generating facilities, hot water or steam (hydrothermal resources) extracted from the earth's crust supplies steam to turbines that drive generators to produce electricity at electric power plants. Eventually, it is hoped that continued development of extraction methods will allow geothermal power production from so-called "hot dry rock" formations, which would open significantly more resources than currently utilized. Most of the known high-temperature hydrothermal resource areas in the United States are in the western part of the country. Electricity from geothermal resources is currently produced in California, Nevada, Utah, Hawaii, and New Mexico.

5.1 Current Data

In 2001, there were 43 operational geothermal power plants in the United States. Those with a capacity of 1 MW or more totaled 2793 MW, which generated 13,813 thousand MWh of electricity, the third most of all renewables after hydroelectricity and biomass. Direct use applications included space and district heating; industrial, such as food dehydration (the drying of vegetable and food products); greenhouse; aquaculture; spas and pools; and snow melting. These contributed 9 trillion Btus to U.S. energy consumption in 2001, whereas geothermal heat pumps contributed 13.5 trillion Btus. The energy consumption of each roughly doubled since 1990. Shipments of geothermal heat pumps by manufacturers have averaged 36,869 per year during the period 1996–2000.

With regard to environmental considerations, geothermal energy does not require fuel burning to produce heat or electricity. Although geothermal steam and flash plants emit mostly water vapor, some

harmful compounds, such as sulfur oxides, nitrogen oxides, hydrogen sulfide, and particulates, are also released. These plants use available emission control systems to meet the most stringent clean air standards. Binary cycle power plants run on a closed-loop system, so no gases are emitted. Whereas binary cycle plants are typically air cooled, the other plant types require a source of cooling water.

Other land uses can coexist with geothermal plants with little interference or fear of accidents. At some sites, however, there are important cultural objections to geothermal development from native populations.

Finally, a typical geothermal plant requires several wells. Although drilling these wells has an impact on the land, using advanced directional or slant drilling minimizes the impact. Several wells can be drilled from one pad, so less land is needed for access roads and fluid piping.

5.2 Technology and Applications

5.2.1 Power Plants

Geothermal power plants are reliable and productive, with a secure fuel supply. They provide baseload power and operate at capacity factors of 85–90% or more. Plants can be built in a modular fashion. This modularity provides flexibility; additional units can be added as needed.

There are three types of geothermal electricity-generating technologies in operation: dry steam, flash steam, and binary cycle. Dry steam plants use steam directly to turn a turbine, which drives a generator that produces electricity. In the United States, these plants are found exclusively at geysers, a high-quality steam-dominated resource.

Hydrothermal fluids above approximately 338°F (170°C) can be used in flash steam plants to make electricity. For these power plants, hot geothermal fluids come out of the ground under pressure and above boiling, and are sprayed into a tank held at a much lower pressure than the fluid, causing some of the fluid to rapidly vaporize or "flash" to steam. The steam then drives a turbine and generator. A second-stage, even lower pressure phase is employed in dual-flash plants to extract additional energy from fluids unflashed in the first phase.

Most geothermal areas contain moderate-temperature water (<338°F). Energy is extracted from these fluids in binary cycle power plants. Hot geothermal fluid and a secondary (hence, binary) fluid with a much lower boiling point than that of water pass through a heat exchanger. Heat from the geothermal fluid causes the secondary fluid to vaporize and drive the turbines.

5.2.2 Direct Use Applications

Direct use systems use lower temperature water [68–302°F (20–150°C)] to provide direct heat for residential, industrial, and commercial uses. These uses of heat from the earth eliminate the need to burn fuel, preventing thousands of tons of greenhouse gases from being emitted.

Direct use systems are typically composed of three components: a production facility, usually a well, to bring the hot water to the surface; a mechanical system—piping, heat exchanger, and controls—to deliver the heat to the space or process; and a disposal system—injection well or storage pond—to receive the cooled geothermal fluid.

The geothermal heat pump is an electric heat pump that uses the earth as a heat source for heating or as a heat sink for cooling. A water and antifreeze mixture circulates through pipes buried in the ground or in subsurface waters (a vertical or horizontal closed loop) and transfers thermal energy to a heat exchanger in the heat pump. The heat exchanger works through a water-to-refrigerant loop. Another variation is the open-loop system that uses groundwater pumped to the heat exchanger in the heat pump. Geothermal heat pumps offer a distinct advantage over the use of air as a source or sink because the ground is at a more favorable temperature. Compared to air, the ground is warmer in winter and cooler in summer. Therefore, geothermal heat pumps demonstrate better performance over air-source heat pumps.

5.3 Niche Markets and Outlook

Challenges faced by the geothermal power industry include the significant risk and uncertainty in locating and tapping commercially useful resources, even in land areas known to have geothermal resources; the high cost of drilling wells; the lack of advanced technology to find and characterize geothermal resources without expensive drilling; and the need to develop technology to allow commercial production from beyond the limited natural hydrothermal resource base.

Current geothermal energy research is addressing these challenges by focusing on resource exploration, drilling technologies, reservoir engineering, and power plant operating efficiencies. The objective of this research is to find and utilize available resources and lower the cost of producing electric power.

6. WIND ENERGY

Winds are created by uneven heating of the atmosphere by the sun, irregularities in the earth's surface, and the rotation of the earth. As a result, winds are strongly influenced and modified by local terrain, bodies of water, weather patterns, vegetative cover, and other such factors. Wind turbines work because they slow down the speed of the wind. The wind flows over the airfoil-shaped blades creating lift, similar to the effect on airplane wings, causing them to turn. The blades are connected to a driveshaft that turns an electric generator to produce electricity.

Although wind energy currently provides less than 2% of U.S. renewable electricity generation, the United States is rich in wind resources. However, the small fraction of this resource that can be economically exploited and the quantity of electricity that can be accommodated on the transmission grid remain highly uncertain in the evaluation of the total resource potential. Some states with good wind resources have difficult terrain, large areas excluded from development for environmental or other land use reasons, or, as in the case of North and South Dakota, are far from demand centers.

Wind power development in the United States began in the early 1980s and was almost exclusively in California due to the favorable regulatory environment and incentives for renewable energy development. It was not until 1999 that significant wind capacity appeared in other states, such as Iowa, Minnesota, and Texas. Development in these states was due in part to state mandates and renewable portfolio standards. In 2001, 20 states reported wind capacity, for a U.S. total of more than 4000 MW. The current expansion is possible because the levelized cost of wind power is becoming competitive with that of power from fossil fuels particularly when the federal production tax credit is factored in, coupled with a wide range of state incentives that benefit wind development. Excluding California, whose wind plants were developed mainly in the 1980s and 1990s, approximately 60% of U.S. wind capacity added during the 3-year period 1999–2001 was in states that currently have renewable portfolio standards.

6.1 Technology and Applications

The most common turbine used in the United States is the three-bladed horizontal axis upwind turbine, but two-bladed horizontal axis downwind turbines and vertical axis turbines also exist. Lattice- or tubular-type towers support the rotor blades and nacelle (an enclosure that houses key components of the turbine, including the gearbox and generator). In 2000, turbines with 600- to 900-kW capacity were popular at new wind farms. However, larger wind farms have been developed since using models with 1300- to 1800-kW capacity. The larger turbines can usually take advantage of some economies of scale and faster wind speeds at the higher heights of their towers. Rotors for the 1500-kW turbines are typically 70–80 m in diameter and are mounted on towers 60–80 + m tall. In the future, multi-megawatt turbines may be possible. For example, GE Wind Energy started testing a prototype 3600-kW turbine in 2002. This model has been selected for the proposed Cape Cod, Massachusetts, project, which may be one of the first offshore projects in the United States.

A handful of manufacturers supply most of the U.S. market for utility-scale turbines. Of these, only one, GE Wind, is domestic; the remainder are European owned. GE Wind came into being in mid-2002 after the acquisition of some of Enron Wind's assets. It is a vertically integrated company providing a wide array of services from planning to operations and maintenance. It has turbine-manufacturing facilities in Tehachapi, California, as well as in Spain and Germany.

Vestas is a leading manufacturer of wind turbines worldwide and supplies U.S. markets with nacelles and blades manufactured in its home country, Denmark. Towers for U.S. projects are procured from suppliers in the United States. There are plans for a future Vestas manufacturing facility in Portland, Oregon. NEG Micon is another Danish turbine manufacturer supplying the U.S. market. It had a manufacturing facility in Champaign, Illinois, but this is currently operating only as a service and repair facility. Its nacelles come from Denmark, whereas blades and towers come from U.S. suppliers. Bonus Energy also supplies U.S. markets with turbines, nacelles, and blades manufactured in Denmark; some towers used in Bonus projects are manufactured in the United States. Nordex is a German manufacturer of wind turbines with interests in the United States. Other domestic companies supply small wind turbines, such as Atlantic Orient Corporation (in the 50-kW range) and Bergey Wind (<10 kW).

6.2 Current Data

By the end of 2001, there were utility-scale wind farms in 20 states. Six states accounted for more than 88% of U.S. capacity (Table IV). In 2001, these wind farms

TABLE IV

U.S. Wind Nameplate Capacity (MW), 2001[a]

Total	4079
California	1608
Texas	1060
Minnesota	300
Iowa	277
Oregon	173
Washington	166

[a] *Source*. U.S. Department of Energy, Energy Information Administration, unpublished data.

produced approximately 6 billion kWh of electricity or approximately 2% of total renewable generation. Wind farms were operating in 10 more states by the end of 2002, with more planned for 2003.

6.3 Niche Markets and Outlook

The nature of wind as an energy source dictates that wind farms have to be located where the wind is occurring. This may be distant from load centers. Furthermore, the location for optimal gathering of wind may be in a terrain in which it is very difficult to erect a turbine tower. This has led to the development of semi-self-erecting towers. Also, bird kills is an issue, especially when towers were constructed with trestles that inadvertently served as nesting places. Newer, sleek, tubular towers do not provide such spaces.

In some areas of the country, the transmission grid is seriously congested and it is difficult and costly to construct new lines. Wind is intermittent and commonly operates with capacity factors under 35%, so backup power must be available when the wind is not blowing. Sometimes, this is accomplished relatively easily by coupling wind power with hydropower, as it is frequently done in California. However, sometimes the situation is more complicated. Also, most grid tariffs and operating procedures were not designed to consider intermittent generation and may not adequately accommodate its unique characteristics. Although new procedures for rate determination are the subject of debate (e.g., the Federal Energy Regulatory Commission's Standard Market Design), the ability to develop technology that could more reliably forecast future wind conditions is an issue. Also, economic large-scale storage is not commercially available.

Federal support for wind power has had gaps. Twice in recent years, the Federal Production Tax Credit (PTC) has expired. On both occasions, there was a period of uncertainty, and project plans were halted or delayed for a number of months until the credit was renewed. The PTC will expire again at the end of 2003.

The EIA's forecast to 2025 projects that wind power will increase in the United States. However, the rate of growth slows in the latter years of the forecast as current incentives and mandates expire and wind economics remain somewhat marginal. The EIA projects that total U.S. wind capacity will expand to 12,000 MW by 2025.

7. SOLAR ENERGY

Using the sun's energy to produce heat or electricity is not a new idea. Technologies that capture the radiation of the sun and use it to produce energy have been available in one form or another for centuries. They continue to be defined and developed today. As of 2003, solar thermal devices were used in a variety of ways, from heating swimming pools to creating steam for electricity generation, whereas photovoltaic (PV) devices use semiconducting materials to convert sunlight directly into electricity. Passive solar systems make use of spontaneous movement of heat by conduction, convection, radiation, or evaporation. (Passive solar systems are not further discussed in this article due to a lack of data.)

Solar energy technologies have important advantages for electric utility and other users. First, the solar resource has no "fuel" cost and is abundantly available. In many areas of the country, the sun shines intensely on a daily basis, providing a nearly unlimited power resource for devices that convert the sun's insolation into useful energy. Second, solar daily output often matches air-conditioning loads and thus provides high value. Third, most solar equipment is modular, allowing supply to be adjusted to demand. Fourth, most PV technologies can easily be operated in remote off-grid areas without transmission and distribution costs. Finally, solar technologies are environmentally clean, with no operating wastes, air pollution, or noise.

Solar energy technologies have some limitations for widespread use. Although costs have declined substantially during the past three decades, they are still very expensive relative to conventional technologies. Also, sunlight is intermittent, resulting in storage and load-matching problems.

7.1 Current Data

Total shipments of solar thermal collectors were 11.2 million square feet in 2001. This represented an

increase of 34% from the 2000 total of 8.4 million square feet. There were 26 companies shipping solar collectors in 2001, the same as in 2000. Import shipments totaled 3.5 million square feet, whereas export shipments totaled 0.8 million square feet. PV cell and module shipments reached 97.7 peak MW in 2001. For the first time in more than 8 years, exports declined, decreasing to 61.4 peak MW. Exports accounted for 63% of total shipments in 2001 compared to 78% in 2000.

7.2 Technology and Applications

Solar energy technologies are classified into two major categories by the type of energy used: solar thermal devices, which use the sun's heat energy, and PV devices, which use the energy inherent in the solar photons and convert it directly into electricity. A solar thermal collector usually consists of a frame, glazing, and an absorber, together with the appropriate insulation. The heat collected by the collector may be used immediately or stored for later use. A PV cell is an electric device consisting of layers of semiconductor materials fabricated to form a junction (adjacent layers of materials with different electric characteristics) and electrical contacts, being capable of converting incident light directly into electricity (direct current). A PV module is two or more cells attached and designed to be used for PV power systems.

7.2.1 Types of Solar Technologies

Data are collected by the EIA on solar–thermal collectors and PV cells and modules. These data provide information on solar energy applications.

7.2.1.1 Solar thermal collector types Solar thermal collectors are classified as low-, medium-, and high-temperature collectors. Low-temperature collectors provide heat up to 110°F through either metallic or nonmetallic absorbers and are used for applications such as swimming pool heating, which has dominated the industry since the rescission of the energy investment tax credits in 1985. Medium-temperature collectors provide heat higher than 110°F (usually 140–180°F) through either glazed flat-plate collectors that use air or liquid as the heat-transfer medium or concentrator collectors that concentrate the heat of incident insulation to greater than one sun. Before the expiration of the energy tax credits at the end of 1985, medium-temperature collector manufacturers dominated the market during the 1970s and early 1980s. Medium-temperature collectors are used for water, space, and process heating. Evacuated-tube collectors are included in this category. High-temperature collectors are parabolic dish and trough collectors used primarily by utilities and nonutility power producers in the generation of electricity for the grid. High-temperature solar thermal collectors, which operate at temperatures higher than 180°F, are used primarily for R&D projects.

7.2.1.2 Photovoltaic cell and module types PV cells and modules are classified into three categories by product type: crystalline silicon cells and modules (including single-crystal, cast silicon, and ribbon silicon), thin-film cells and modules (made from a number of layers of photosensitive materials such as amorphous silicon), and concentrator cells and modules (in which a lens is used to gather and converge sunlight onto the cells or module surface). Crystalline silicon cells and modules dominate the PV industry.

7.2.2 Uses of Solar Technologies

7.2.2.1 Solar thermal collectors Since 1986, the largest end use for solar thermal collectors has been for heating swimming pools. In 2001, solar swimming pool heaters represented 97% of solar thermal collector shipments (Table V). This application usually involved the use of low-temperature collectors. A common low-temperature pool-heating solar collector is a black plastic or rubber-like sheet with tubing through which water is circulated. The heat of the sun is transferred directly from the black absorbing material to the water circulating through the tubing to supply heat to the pool.

The second largest end use was for domestic hot water systems, which represented 2% of solar collector shipments (Table V). Typical solar water-heating systems feature flat-plate collectors or collectors installed in an internal collector storage or thermosiphon system. Unlike pool-heating systems, domestic solar water-heating systems almost always have a conventional backup (i.e., gas or electric). Medium-temperature collectors also were shipped for space heating and for installation into systems that provide both space and water heating and process heating. High-temperature parabolic dish and trough collectors were shipped for electricity generation and hot water in the past when electricity and natural gas prices were expected to increase very rapidly.

TABLE V

Shipments of Solar Collectors by Market Sector, End Use, and Type, 2000 and 2001 (in Thousands of Square Feet)[a]

Type	Low-temperature liquid/air metallic and nonmetallic	Medium-temperature liquid					High-temperature parabolic dish/ trough	2001 total	2000 total
		Air	Integral collector storage/thermosiphon	Flat-plate (pumped)	Evacuated tube	Concentrator			
Market sector									
Residential	9885	3	80	154	1	[b]	0	10,125	7473
Commercial	987	1	[b]	22	[b]	0	1	1012	810
Industrial	12	0	0	5	0	0	0	17	57
Utility	0	0	0	0	0	0	1	1	5
Other	34	0	1	0	0	0	0	35	10
Total	10,919	4	81	181	2	[b]	2	11,189	8354
End use									
Pool heating	10,782	0	0	16	0	0	0	10,797	7863
Hot water	42	0	81	149	1	[b]	0	274	367
Space heating	61	4	0	5	[b]	0	0	70	99
Space cooling	0	0	0	0	0	0	0	0	0
Combined space and water	0	0	0	12	0	0	0	12	2
Process heating	34	0	0	0	0	0	0	34	20
Electricity generation	0	0	0	0	0	0	2	2	3
Other	0	0	0	0	0	0	0	0	0
Total	10,919	4	81	181	2	[b]	2	11,189	8354

[a] *Source*. Energy Information Administration, Form EIA-63A, "Annual Solar Thermal Collector Manufacturers Survey."
[b] Less than 500 ft^2.

TABLE VI

Shipments of Photovoltaic Cells and Modules by Market Sector, End Use, and Type, 2000 and 2001 (in Peak Kilowatts)[a]

Sector and end use	Crystalline silicon	Thin-film silicon	Concentrator silicon	Other	2001 total	2000 total
Market						
Industrial	24,754	3135	174	0	28,063	28,808
Residential	28,307	4955	[b]	0	33,262	24,814
Commercial	13,440	2270	0	0	15,710	13,692
Transportation	7525	961	0	0	8486	5502
Utility	4799	747	300	0	5846	6298
Government	5375	353	0	0	5728	4417
Other	451	120	0	0	571	4690
Total	84,651	12541	474	0	97,666	88,221
End use						
Electrical generation						
Grid interactive	22,444	4482	300	0	27,226	21,713
Remote	18,772	2501	174	0	21,447	14,997
Communications	12,974	1769	0	0	14,743	12,269
Consumer goods	3664	395	0	0	4059	2870
Transportation	10,803	1833	0	0	12,636	12,804
Water pumping	6737	708	0	0	7444	5644
Cells/modules to OEM	5778	490	0	0	6268	12,153
Health	2931	272	0	0	3203	2742
Other	549	92	0	0	641	3028
Total	84,651	12,541	474	0	97,666	88,221

[a] *Source*. Energy Information Administration, Form EIA-63B, "Annual Photovoltaic Module/Cell Manufacturers Survey."
[b] Less than 500 ft^2.

7.2.2.2 Uses of photovoltaic cells and modules The largest end-use application of PV cells and modules is for electricity generation (combined grid-interactive and remote). Grid-interactive and remote (i.e., stand-alone) power generation includes applications for grid distribution and general remote uses, such as residential power and power for mobile homes. In 2001, electric generation accounted for 50% of total PV shipments (Table VI).

7.3 Niche Markets and Outlook

Solar technologies require either significant cost reductions or a combination of cost reductions and an increase in natural gas prices to become cost competitive in most grid applications. Some obstacles slowing commercialization are technology specific, whereas others are more general. The primary obstacle is that solar technologies cannot currently compete with conventional fossil-fueled technologies in grid-connected applications. On the other hand, some niche applications, such as photovoltaics for transmission and distribution support at the end of a fully loaded distribution line, have value above that of a new generating plant. Such applications may permit the technology to establish itself and benefit from economies of scale and learning effects.

Acknowledgments

I thank Mark Gielecki, Louise Guey-Lee, and James Holihan for their contributions to this article.

SEE ALSO THE FOLLOWING ARTICLES

Biomass for Renewable Energy and Fuels • Computer Modeling of Renewable Power Systems • Energy Development on Public Land in the United States • National Energy Policy: United States • Public Reaction to Renewable Energy Sources and Systems • Renewable Energy and the City • Renewable Energy in Europe • Renewable Energy in Southern Africa • Renewable Energy Policies and Barriers • Renewable Energy, Taxonomic Overview • Renewable Portfolio Standard

Further Reading

Energy Information Administration (2001). Renewable energy issues & trends 2000, Report No. DOE/EIA-0628. Energy Information Administration, Washington, DC.

Energy Information Administration (2002). Renewable energy annual 2001, Report No. DOE/EIA-0603. Energy Information Administration, Washington, DC.

Energy Information Administration (2003). Annual energy outlook 2003, Report No. DOE/EIA-0383. Energy Information Administration, Washington, DC.

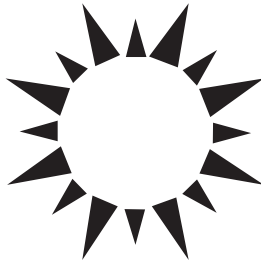
Federal Energy Regulatory Commission hydropower homepage (2003). http://www.ferc.gov.

Geo-Heat Center (2003). http://geoheat.oit.edu.

U.S. Department of Energy Geothermal Energy Program (2003). http://www.eere.energy.gov.

U.S. Department of Energy Hydropower Program (2003). http://hydropower.inel.gov.

Renewable Energy Policies and Barriers



FREDRIC BECK
Renewable Energy Policy Project
Washington, D.C., United States

ERIC MARTINOT
Global Environment Facility
Washington D.C., United States

1. **Barriers to Renewable Energy**
2. **Renewable Energy Promotion Policies**
3. **Transport Biofuels Policies**
4. **Emissions Reduction Policies**
5. **Power-Sector Restructuring Policies**
6. **Distributed Generation Policies**
7. **Rural Electrification Policies**
8. **Summary**

Glossary

biofuels Liquid fuels used in transportation that are derived from renewable biomass resources. Not limited to transportation, these fuels can also be used for heating and electricity generation. The main fuels promoted are ethanol, gasohol, biodiesel, bio-oil, and methanol.

capital costs The initial investment costs associated with a renewable energy installation, as distinct from operating and maintenance (all technologies) or fuel costs (in the case of biomass).

distributed generation Generation of electricity at or near the point of end use, often in small quantities, rather than remotely in a large centralized power plants.

independent power producer An organization that produces electric power (i.e., from a wind farm) and sells that power under contract (called a power purchase agreement) to the electric utility or to a third party. If to a third party, the electricity must be transferred over the utility transmission system, usually for a fee.

microcredit Small amounts of credit made available to households or small businesses for investments in high-capital-cost renewable energy supplies or services, usually within a rural development context.

power sector restructuring A political/policy process causing changes in the institutions, ownership, rates and tariffs, power market rules, competition, or financing in the electric power industry (primarily among power generators and electric utilities).

The 1990s saw an explosion of energy policy changes around the globe. Driven by economic, environmental, security, and social concerns, energy regulation has been in great flux. Many of the changes are having a profound influence on renewable energy, both from policies explicitly designed to promote renewable energy and from other policies that indirectly influence incentives and barriers for renewable energy. This article considers six different types of policies that affect renewable energy development, both directly or indirectly: renewable energy promotion policies, transport biofuels policies, emissions reduction policies, electric power restructuring policies, distributed generation policies, and rural electrification policies. Each policy reduces one or more key barriers that impede the development of renewable energy. These barriers are discussed first. In general, most renewable energy policies address cost-related barriers in some manner. Many policies address the requirements for utilities to purchase renewable energy from power producers. Most policies also address the perceived risks of renewable energy in one form or another (i.e., technical, financial, legal). Still others primarily address regulatory and institutional barriers. Some related policies may heighten barriers to renewable energy rather than reduce them.

 © 2004 Elsevier Inc. All rights reserved.

1. BARRIERS TO RENEWABLE ENERGY

The need for enacting policies to support renewable energy is often attributed to a variety of barriers, or conditions that prevent investments from occurring. Often the result of barriers is to put renewable energy at an economic, regulatory, or institutional disadvantage relative to other forms of energy supply. Barriers include subsidies for conventional forms of energy, high initial capital costs coupled with lack of fuel-price risk assessment, imperfect capital markets, lack of skills or information, poor market acceptance, technology prejudice, financing risks and uncertainties, high transactions costs, and a variety of regulatory and institutional factors. Many of these barriers could be considered market distortions that unfairly discriminate against renewable energy, while others have the effect of increasing the costs of renewable energy relative to the alternatives. Barriers are often quite situation specific in any given region or country; nevertheless, three broad categories of barriers are discussed in this section. Table I summarizes the key renewable energy policies and barriers presented.

1.1 Costs and Pricing

Many argue that renewable energy costs more than other energy sources, resulting in cost-driven decisions and policies that avoid renewable energy. In practice, a variety of factors can distort the comparison. For example, public subsidies may lower the costs of competing fuels. Although it is true that initial capital costs for renewable energy technologies are often higher on a cost-per-unit basis (i.e., $/kW), it is widely accepted that a true comparison must be made on the basis of total life-cycle costs. Life-cycle costs account for initial capital costs, future fuel costs, future operation and maintenance costs, decommissioning costs, and equipment lifetime. Here lies part of the problem in making comparisons: What are fuel costs going to be in the future? How should future costs be discounted to allow comparison with present costs based on expected interest rates? The uncertainties inherent in these questions affect cost comparisons. Existing analytical tools for calculating and comparing costs can discriminate against renewable energy if they do not account for future uncertainties or make unrealistic assumptions.

Many policies attempt to compensate for cost-related barriers by providing additional subsidies for renewable energy in the form of tax credits or incentives, by establishing special pricing and power-purchasing rules, and by lowering transaction costs. Despite many calls for reducing subsidies for fossil fuels and nuclear power, in practice this proves politically difficult. Thus, practical policies have tended to focus on increasing subsidies for renewable energy rather than reducing subsidies for fossil fuels and nuclear power.

1.1.1 Subsidies for Competing Fuels

Large public subsidies, both implicit and explicit, are channeled in varying amounts to all forms of energy, which can distort investment cost decisions. Organizations such as the World Bank and International Energy Agency put global annual subsidies for fossil fuels in the range of $100 billion to $200 billion, although such figures are very difficult to estimate (for comparison, the world spends some $1 trillion annually on purchases of fossil fuels). Public subsidies can take many forms: direct budgetary transfers, tax incentives, R&D spending, liability insurance, leases, land rights-of-way, waste disposal, and guarantees to mitigate project financing or fuel price risks. Large subsidies for fossil fuels can significantly lower final energy prices, putting renewable energy at a competitive disadvantage if it does not enjoy equally large subsidies.

1.1.2 High Initial Capital Costs

Even though lower fuel and operating costs may make renewable energy cost competitive on a life-cycle basis, higher initial capital costs can mean that renewable energy provides less installed capacity per initial dollar invested than conventional energy sources. Thus, renewable energy investments generally require higher amounts of financing for the same capacity. Depending on the circumstances, capital markets may demand a premium in lending rates for financing renewable energy projects because more capital is being risked up front than in conventional energy projects. Renewable energy technologies may also face high taxes and import duties. These duties may exacerbate the high first-cost considerations relative to other technologies and fuels.

1.1.3 Difficulty of Fuel Price Risk Assessment

Risks associated with fluctuations in future fossil fuel prices may not be quantitatively considered in decisions about new power generation capacity because these risks are inherently difficult to assess. Historically, future fuel price risk has not been considered an important factor because future fossil

fuel prices have been assumed to be relatively stable or moderately increasing. Thus, risks of severe fluctuations are often ignored. However, with greater geopolitical uncertainties and energy market deregulation has come new awareness about future fuel price risks. Renewable energy technologies avoid fuel costs (with the exception of biomass) and so avoid fuel price risk. However, this benefit, or risk-reduction premium, is often missing from economic comparisons and analytical tools because it is difficult to quantify. Further, for some regulated utilities, fuel costs are factored into regulated power rates so that consumers rather than utilities bear the burden of fuel price risks, and utility investment decisions are made without considering fuel price risk.

1.1.4 Unfavorable Power Pricing Rules

Renewable energy sources feeding into an electric power grid may not receive full credit for the value of their power. Two factors are at work. First, renewable energy generated on distribution networks near final consumers rather than at centralized generation facilities may not require transmission and distribution (i.e., would displace power coming from a transmission line into a node of a distribution network). But utilities may only pay wholesale rates for the power, as if the generation was located far from final consumers and required transmission and distribution. Thus, the locational value of the power is not captured by the producer. Second, renewable energy is often an intermittent source whose output level depends on the resource (i.e., wind and sun) and cannot be entirely controlled. Utilities cannot count on the power at any given time and may lower prices for it. Lower prices take two common forms: (1) a zero price for the capacity value of the generation (utility only pays for the energy value) and (2) an average price paid at peak times (when power is more valuable), which is lower than the value of the power to the utility—even though the renewable energy output may directly correspond with peak demand times and thus should be valued at peak prices.

1.1.5 Transaction Costs

Renewable energy projects are typically smaller than conventional energy projects. Projects may require additional information not readily available, or they may require additional time or attention to financing or permitting because of unfamiliarity with the technologies or uncertainties over performance. For these reasons, the transaction costs of renewable energy projects—including resource assessment, siting, permitting, planning, developing project proposals, assembling financing packages, and negotiating power-purchase contracts with utilities—may be much larger on a per kilowatt (kW) capacity basis than for conventional power plants. Higher transaction costs are not necessarily an economic distortion in the same way as some other barriers, but simply make renewables more expensive. However, in practice some transaction costs may be unnecessarily high, for example, overly burdensome utility interconnection requirements and high utility fees for engineering reviews and inspection.

1.1.6 Environmental Externalities

The environmental impacts of fossil fuels often result in real costs to society, in terms of human health (i.e., loss of work days, health care costs), infrastructure decay (i.e., from acid rain), declines in forests and fisheries, and perhaps ultimately, the costs associated with climate change. Dollar costs of environmental externalities are difficult to evaluate and depend on assumptions that can be subject to wide interpretation and discretion. Although environmental impacts and associated dollar costs are often included in economic comparisons between renewable and conventional energy, investors rarely include such environmental costs in the bottom line used to make decisions.

1.2 Legal and Regulatory

1.2.1 Lack of Legal Framework for Independent Power Producers

In many countries, power utilities still control a monopoly on electricity production and distribution. In these circumstances, in the absence of a legal framework, independent power producers may not be able to invest in renewable energy facilities and sell power to the utility or to third parties under so-called power purchase agreements. Or utilities may negotiate power purchase agreements on an individual ad hoc basis, making it difficult for project developers to plan and finance projects on the basis of known and consistent rules.

1.2.2 Restrictions on Siting and Construction

Wind turbines, rooftop solar hot-water heaters, photovoltaic installations, and biomass combustion facilities may all face building restrictions based on height, aesthetics, noise, or safety, particularly in urban areas. Wind turbines have faced specific

TABLE I

Summary of Renewable Energy Policies and Barriers

Policies	Description	Key barriers addressed
Renewable Energy Promotion Policies		
Price-setting and quantity-forcing policies	Mandates prices to be paid for renewable energy or requires a fixed amount or share of generation to be renewable	High costs, unfavorable power pricing rules, perceived risks
Cost-reduction policies	Reduces investment costs through subsidies, rebates, tax relief, loans and grants	High costs, perceived risks
Public investments and market facilitation activities	Provides public funds for direct investments or for guarantees, information, training, etc. to facilitate investments	Transaction costs, perceived risks, lack of access to credit, information, and skills
Power grid access policies	Gives renewable energy equal or favorable treatment for access to power grids and transmission systems	Independent power producer frameworks, transmission access, interconnection requirements
Transport Biofuels Policies		
Biofuels mandates	Mandates specific shares of transport fuel consumption from biofuels	Lack of fuel production or delivery infrastructure
Biofuel tax policies	Provides tax relief for biofuels	High costs
Emissions Reduction Policies		
Renewable energy set-asides	Allocates, or sets aside, a percentage of mandated environmental emissions reductions to be met by renewable energy	Environmental externalities
Emissions cap and trade policies	Allows renewables to receive monetary credit for local pollutant emissions reductions	Environmental externalities
Greenhouse gas mitigation policies	Allows renewables to receive monetary credit for greenhouse-gas emissions reductions	Environmental externalities
Power-Sector Restructuring Policies		
Competitive wholesale power markets	Allows competition in supplying wholesale generation to the utility network and eliminates wholesale pricing restrictions.	May heighten barriers of high costs, lack of fuel price risk assessment, unfavorable power pricing rules
Self-generation by end users	Allows end users to generate their own electricity and either sell surplus power back to the grid or partly offset purchased power	May reduce barrier of interconnection requirements, but heighten barriers of high costs, lack of fuel price risk assessment
Privatization or commercialization of utilities	Changes government-owned and operated utilities into private or commercial entities	May reduce barrier of subsidies, but heighten barriers of high capital costs and perceived risks
Unbundling of generation, transmission and distribution	Eliminates monopolies so that separate entities provide generation, transmission, and distribution	May provide greater incentives to self-generate, including with renewable energy
Competitive retail power markets	Provides competition at the retail level for power sales, including "green power" sales	May reduce barriers of high costs, lack of information, transaction costs
Distributed Generation Policies		
Net metering	Values renewable energy production at the point of end use and allows utility networks to provide "energy storage" for small users	Unfavorable power pricing rules
Real-time pricing	Values renewable energy production at the actual cost of avoided fossil fuel generation at any given time of the day	Unfavorable power pricing rules
Capacity credit	Provides credit for the value of standing renewable energy capacity, not just energy production	Unfavorable power pricing rules

continues

Table I continued

Policies	Description	Key barriers addressed
Interconnection regulations	Creates consistent and transparent rules, norms, and standards for interconnection	Interconnection requirements, transaction costs
Rural Electrification Policies		
Rural electrification policy and energy service concessions	Makes renewable energy part of rural electrification policy and planning and creates regulated businesses to serve rural customers	Subsidies for competing fuels, lack of skills and information, high costs, lack of access to credit
Rural business development and microcredit	Supports private entrepreneurs to provide renewable energy products and services to end users and offer consumer credit for purchases	Lack of skills, lack of access to credit
Comparative line extension analyses	Analyzes the relative costs of renewable energy with conventional fuels and power delivery	Subsidies for competing fuels, lack of information

environmental concerns related to siting along migratory bird paths and coastal areas. Urban planning departments or building inspectors may be unfamiliar with renewable energy technologies and may not have established procedures for dealing with siting and permitting. Competition for land use with agricultural, recreational, scenic, or development interests can also occur.

1.2.3 Transmission Access

Utilities may not allow favorable transmission access to renewable energy producers or may charge high prices for transmission access. Transmission access is necessary because some renewable energy resources like windy sites and biomass fuels may be located far from population centers. Transmission or distribution access is also necessary for direct third-party sales between the renewable energy producer and a final consumer. New transmission access to remote renewable energy sites may be blocked by transmission-access rulings or right-of-way disputes.

1.2.4 Utility Interconnection Requirements

Individual home or commercial systems connected to utility grids can face burdensome, inconsistent, or unclear utility interconnection requirements. Lack of uniform requirements can add to transaction costs. Safety and power-quality risk from nonutility generation is a legitimate concern of utilities, but a utility may tend to set interconnection requirements that go beyond what is necessary or practical for small producers in the absence of any incentive to set more reasonable but still technically sound requirements. In turn, the transaction costs of hiring legal and technical experts to understand and comply with interconnection requirements may be significant. Policies that create sound and uniform interconnection standards can reduce interconnection hurdles and costs.

1.2.5 Liability Insurance Requirements

Small power generators (particularly home photovoltaic systems feeding into the utility grid under so-called net metering provisions) may face excessive requirements for liability insurance. The phenomenon of islanding, which occurs when a self-generator continues to feed power into the grid when power flow from the central utility source has been interrupted, can result in serious injury or death to utility repair crews. Although proper equipment standards can prevent islanding, liability is still an issue. Several U.S. states have prohibited utilities from requiring additional insurance beyond normal homeowner liability coverage as part of net metering statutes.

1.3 Market Performance

1.3.1 Lack of Access to Credit

Consumers or project developers may lack access to credit to purchase or invest in renewable energy because of lack of collateral, poor creditworthiness, or distorted capital markets. In rural areas, microcredit lending for household-scale renewable energy systems may not exist. Available loan terms may be too short relative to the equipment or investment lifetime. In some countries, power project developers have difficulty obtaining bank financing because of uncertainty as to whether utilities will continue to honor long-term power purchase agreements to buy the power.

1.3.2 Perceived Technology Performance Uncertainty and Risk

Proven, cost-effective technologies may still be perceived as risky if there is little experience with them in a new application or region. The lack of visible installations and familiarity with renewable

energy technologies can lead to perceptions of greater technical risk than for conventional energy sources. These perceptions may increase required rates of return, result in less capital availability, or place more stringent requirements on technology selection and resource assessment. "Lack of utility acceptance" is a phrase used to describe the historical biases and prejudices on the part of traditional electric power utilities. Utilities may be hesitant to develop, acquire, and maintain unfamiliar technologies, or give them proper attention in planning frameworks. Finally, prejudice may exist because of poor past performance that is out of step with current performance norms.

1.3.3 Lack of Technical or Commercial Skills and Information

Markets function best when everyone has low-cost access to good information and the requisite skills. But in specific markets, skilled personnel who can install, operate, and maintain renewable energy technologies may not exist in large numbers. Project developers may lack sufficient technical, financial, and business development skills. Consumers, managers, engineers, architects, lenders, or planners may lack information about renewable energy technology characteristics, economic and financial costs and benefits, geographical resources, operating experience, maintenance requirements, sources of finance, and installation services. The lack of skills and information may increase perceived uncertainties and block decisions.

2. RENEWABLE ENERGY PROMOTION POLICIES

Policies whose specific goal is to promote renewable energy fall into three main categories: (1) price-setting and quantity-forcing policies, which mandate prices or quantities; (2) investment cost reduction policies, which provide incentives in the form of lower investment costs; and (3) public investments and market facilitation activities, which offer a wide range of public policies that reduce market barriers and facilitate or accelerate renewable energy markets. Historically, governments have enacted these policies in a rather ad hoc manner. More recently, national renewable energy targets (also referred to as goals) have emerged as a political context for promoting specific combinations of policies from all three categories. Such targets focus on the aggregate energy production of an entire country or group of countries. Targets may specify total primary energy from renewables or minimum renewable energy shares of electricity generation.

Several countries have adopted or are proposing national renewable energy targets. The European Union collectively has adopted a target of 22% of total electricity generation from renewables by 2010, with individual member states having individual targets above or below that amount. Japan has adopted a target of 3% of total primary energy by 2010. Recent legislative proposals in the United States would require 10% of electricity generation from renewables by 2020. China and India are the first developing countries to propose renewable energy targets. India has proposed that by 2012, 10% of annual additions to power generation would be from renewable energy; China has a similar goal of 5% by 2010. Other countries with existing or proposed targets are Australia, Brazil, Malaysia, and Thailand. In addition, countries from around the world placed increased attention on renewable energy targets at the United Nations World Summit for Sustainable Development in 2002.

2.1 Price-Setting and Quantity-Forcing Policies

Price-setting policies reduce cost- and pricing-related barriers by establishing favorable pricing regimes for renewable energy relative to other sources of power generation. The quantity of investment obtained under such regimes is unspecified, but prices are known in advance. Quantity-forcing policies do the opposite; they mandate a certain percentage or absolute quantity of generation to be supplied from renewable energy, at unspecified prices. Often price-setting or quantity-forcing policies occur in parallel with other policies, such as investment cost–reduction policies.

The two main price-setting policies seen to date are the PURPA legislation in the United States and electricity feed-in laws in Europe. The two main quantity-forcing policies seen to date are competitively bid renewable-resource obligations and renewable portfolio standards.

2.1.1 U.S. Public Utility Regulatory Policies Act (PURPA)

PURPA was enacted in 1978 in part to encourage electric power production by small power producers using renewable resources to reduce U.S. dependence

on foreign oil. The policy required utilities to purchase power from small renewable generators and cogenerators, known as qualifying facilities, through long-term (10-year) contracts at prices approximating the avoided costs of the utilities. These avoided costs represented the marginal costs to the utilities of building new generation facilities, which could be avoided by purchasing power from the qualifying facilities instead. Avoided cost calculations typically assumed an aggressive schedule of escalating future energy prices, making contract prices to qualifying facilities quite attractive.

For example, standard offer contracts under PURPA in California in the 1980s set the avoided cost of generation in the range 6 to 10 cents/kWh—very favorable rates, which initially spurred many investments by renewable energy producers. However, by the 1990s energy prices had not risen as originally expected and a large number of natural-gas-fired independent generation came online in California. Power surpluses emerged, wholesale power prices declined, and declining standard offer rates led to reduced competitiveness of renewable energy and a significant slowdown in construction of new capacity.

2.1.2 Electricity Feed-in Laws

The electricity feed-in laws in Germany, and similar policies in other European countries in the 1990s, set a fixed price for utility purchases of renewable energy. For example, in Germany starting in 1991, renewable energy producers could sell their power to utilities at 90% of the retail market price. The utilities were obligated to purchase the power. The German feed-in law led to a rapid increase in installed capacity and development of commercial renewable energy markets. Wind power purchase prices were highly favorable, amounting to about DM 0.17/kWh (US 10 cents/kWh), and applied over the entire life of the plant. Total wind power installed went from near zero in the early 1990s to over 8500 MW by 2001, making Germany the global leader in renewable energy investment.

Partly because retail electricity prices declined with increasing competition due to electricity deregulation, which made producers and financiers wary of new investments, a new German Renewable Energy Law of 2000 changed electricity feed-in pricing. Pricing became based on fixed norms unique to each technology, which in turn were based on estimates of power production costs and expectations of declines in those costs over time. For example, wind power prices remained at the previous level of DM 0.17/kWh for plants commissioned in 2001, but only for the first 5 years of operation, after which prices paid declined. Solar PV prices were set initially at DM 0.99/kWh. All prices had built-in declines over time (i.e., 1.5% annual decreases in starting tariffs paid for wind power plants commissioned in subsequent years). This provision addressed one of the historical criticisms of feed-in approaches, which was that they did not encourage technology cost reductions or innovation. The new law's provisions for regular adjustments to prices addressed technological and market developments. The law also distributed the costs of the policy (i.e., the additional costs of wind power over conventional power) among all utility customers in the country. This issue of burden sharing had become a significant political factor in Germany by 2000 because the old law placed a disproportionate burden on utility customers in specific regions where wind power development was heaviest.

Other countries in Europe with renewable electricity feed-in laws include Denmark, France, Greece, Italy, Portugal, Spain, and Sweden. A combination of feed-in tariffs, production subsidies of DK 0.10/kWh, and a strong domestic market helped the Danish wind industry maintain a 50% market share of global wind turbine production for a number of years.

2.1.3 Competitively-Bid Renewable-Resource Obligations

The United Kingdom tried competitive bidding for renewable-energy-resource obligations during the 1990s under its Non-Fossil-Fuel Obligation (NFFO) policy. Under the NFFO, power producers bid on providing a fixed quantity of renewable power, with the lowest-price bidder winning the contact. With each successive bidding round (there were four total), bidders reduced prices relative to the last round. For example, wind power contract prices declined from 10 p/kWh in 1990 under NFFO-1 to 4.5 p/kWh in 1997 under NFFO-4. One of the lessons some have drawn from the UK experience is that competitively determined subsidies can lead to rapidly declining prices for renewable energy. However, there has also been criticism that the NFFO process encouraged competing projects to bid below cost in order to capture contracts, with the result that successful bidders were unable to meet the terms of the bid or ended up insolvent. This criticism proved valid in practice; contracts awarded to low bidders did not always translate into projects on the ground. The UK abandoned the NFFO approach after the fourth

round of bidding in 1997. Other countries with similar competitively bid renewable resource mechanisms have included Ireland (under the Alternative Energy Requirement program), France (under the EOLE program), and Australia (under the Renewable Energy Commercialisation Program program).

2.1.4 Renewable Energy Portfolio Standards (RPS)

An RPS requires that a minimum percentage of generation sold or capacity installed be provided by renewable energy. Obligated utilities are required to ensure that the target is met, either through their own generation, power purchases from other producers, or direct sales from third parties to the utility's customers. Typically, RPS obligations are placed on the final retailers of power, who must purchase either a portion of renewable power or the equivalent amount of green certificates (see the next section). Two types of standards have emerged: capacity-based standards set a fixed amount of capacity by a given date, while generation-based standards mandate a given percentage of electricity generation that must come from renewable energy.

In the United States, many RPS policies have occurred as part of utility restructuring legislation. These are typically generation-based standards with phased implementation to allow utilities to reach incrementally increasing targets over a number of years. At least 12 U.S. states have enacted an RPS, ranging from 1 to 30% of electricity generation. However, the amount of new and additional generation expected from these standards varies widely depending on existing renewable energy capacity. For example, Maine historically generates over 40% of its power from renewable resources, so its 30% standard is unlikely to result in any new renewable generation. In contrast, California's requirement to increase renewable sales from 10.5% in 2001 to 20% by 2017 will likely result in a significant amount of new in-state renewable energy generation. Texas implemented an RPS in 2000 requiring 2000 MW of new renewable capacity by 2008. Partly due to federal production tax credits, Texas has been substantially ahead of schedule, with half of the targeted capacity in place by 2002.

In Europe, the Netherlands has been a leader among RPS initiatives. Dutch utilities have adopted an RPS voluntarily, based on targets of 5% of electricity generation by 2010, increasing to 17% by 2020. Other countries with regulatory requirements for utilities or electricity retailers to purchase a percentage of renewable power include Australia, Brazil, Belgium, Denmark, France, Japan, Spain, Sweden, and the United Kingdom. The UK's Renewables Obligation on suppliers will rise in annual steps from 3% in 2003 to 10% in 2010. In Denmark, legislation obliges end users to purchase 20% of their electricity from renewable sources by 2003.

In Brazil, a policy enacted during the electricity crisis of 2001 requires national utilities to purchase more than 3000 MW of renewable energy capacity by 2016. Purchase prices are set by the government at 80% of the national average electricity retail price. Thus, in contrast to most national policies elsewhere, Brazil's policy is effectively both price setting and quantity forcing.

2.1.5 Renewable Energy (Green) Certificates

Renewable energy (green) certificates are emerging as a way for utilities and customers to trade renewable energy production or consumption credits in order to meet obligations under RPS and similar policies. Standardized certificates provide evidence of renewable energy production and are coupled with institutions and rules for trading that separate renewable attributes from the associated physical energy. This enables a paper market for renewable energy to be created independent of actual electricity sales and flows. Green certificate markets are emerging in several countries, allowing producers or purchasers of renewable energy who earn green certificates to sell those certificates to those who need to meet obligations but have not generated or purchased the renewable power themselves. Those without obligations but who wish to voluntarily support green power for philosophical or public relations reasons may also purchase certificates.

Public and private institutions are emerging that keep track of renewable energy generation, assign certificates to generators, and register trades and sales of certificates. Green certificate trading is gaining ground in the UK, Belgium, Denmark, Australia, and the United States. Europe embarked on a test phase of an EU-wide renewable energy certificate trading system during 2001 and 2002. This Renewable Energy Certificate System (RECS) was established as a formal association in December 2002 with over 100 members in 14 different countries. Certificates for over 22,500 GWh were issued through 2003.

2.2 Cost-Reduction Policies

A number of policies are designed to provide incentives for voluntary investments in renewable energy by reducing the costs of such investments.

These policies can be characterized as falling in five broad categories: policies that (1) reduce capital costs up front (via subsidies and rebates), (2) reduce capital costs after purchase (via tax relief), (3) offset costs through a stream of payments based on power production (via production tax credits), (4) provide concessionary loans and other financial assistance, and (5) reduce capital and installation costs through economies of bulk procurement.

2.2.1 Subsidies and Rebates

Reduction in the initial capital outlay by consumers for renewable energy systems is accomplished through direct subsidies, or rebates. These subsidies are used to buy down the initial capital cost of the system so that the consumer sees a lower price. In the United States, at least nineteen states offer rebate programs at the state, local, or utility level to promote the installation of renewable energy equipment. The majority of the programs are available from state agencies and municipally owned utilities and support solar water heating or photovoltaic systems, though some include geothermal heat pumps, small wind generators, passive solar, biomass, and fuel cells. Homes and businesses are usually eligible, although some programs target industry and public institutions as well. In some cases, rebate programs are combined with low- or no-interest loans.

Sustained efforts to increase the use of renewables have been made via coordinated, multiyear, multipolicy initiatives. For example, Japan, Germany, and the United States subsidize capital costs of solar PV as part of their market transformation program.

- Japan's Sunshine Program provides capital subsidies and net metering for rooftop PV systems. From 1994 to 2000, the government invested 86 billion yen ($725 million USD), resulting in 58,000 system installations and over 220 MW of PV capacity. Subsidies began at 900,000 yen/peak-kW (US$5/peak-watt) in 1994, and were gradually reduced to 120,000 yen/peak-kW (US$1/peak-watt) in 2001 as PV prices fell.
- Germany began a 1000 solar roofs program in 1991 that offered subsidies for individual household purchases of solar PV of up to 60% of capital system costs. The program was expanded in 1999 to 100,000 roofs over 5 years, providing 10-year low-interest loans to households and businesses. As a result of favorable feed-in tariffs and low-interest loans, the program was expected to provide 300 peak-MW of PV capacity.
- The United States launched a million solar roofs initiative in 1997 to install solar PV systems and solar thermal systems on 1 million buildings by 2010. The program includes long-term, low-interest customer financing, government procurement for federal buildings, commercialization programs, and production incentives. Individual states also have capital subsidy programs for PV, with the California Energy Commission offering rebates of up to $4.50/peak-watt or 50% off the system purchase price, New York and New Jersey offering up to $5/peak-watt subsidies, and New York rebating as much as 70% of the cost of eligible equipment as of 2002.

Subsidy programs also exist for wind power, such as Denmark's DK 0.10/kWh (US 1.5 cents/kWh) production subsidy paid to utilities. Among developing countries, Thailand provided subsidies for small renewable energy power producers starting in 2000, soliciting bids for 300 MW of small renewable power and providing production subsidies above standard power purchase rates for at least the first 5 years of operation of each facility.

2.2.2 Tax Relief

Tax relief policies to promote renewable energy have been employed in the United States, Europe, Japan, and India. Tax relief has been especially popular in the United States, where a host of federal and state tax policies address energy production, property investments, accelerated depreciation, and renewable fuels. State policies vary widely in scope and implementation. At least 17 states have personal tax incentives, 21 states have corporate tax incentives, 16 states have sales tax incentives, and 24 states have property tax incentives.

2.2.2.1 Investment Tax Credits Investment tax credits for renewable energy have been offered for businesses and residences. In the United States, businesses receive a 10% tax credit for purchases of solar and geothermal renewable energy property, subject to certain limitations. Some U.S. states have investment tax credits of up to 35%.

2.2.2.2 Accelerated Depreciation Accelerated depreciation allows renewable energy investors to receive the tax benefits sooner than under standard depreciation rules. The effect of accelerated depreciation is similar to that of investment tax credits. In the United States, businesses can recover investments in solar, wind, and geothermal property by depreciating them over a period of 5 years, rather than the

15- to 20-year depreciation lives of conventional power investments.

India's accelerated depreciation policy allowed 100% depreciation in the first year of operation, helping spur the largest wind power industry among developing countries. However, this policy led to large investments without sufficient regard to long-term operating performance and maintenance, resulting in capacity factors lower than for wind power installations elsewhere. This led many to conclude that production-based incentives are preferable to investment tax credits and accelerated depreciation, although Germany's investment tax credits accompanied by wind turbine technical standards and certification requirements avoided the problems found in India.

2.2.2.3 Production Tax Credits A production tax credit provides the investor or owner of qualifying property with an annual tax credit based on the amount of electricity generated by that facility. By rewarding production, these tax credits encourage improved operating performance. A production tax credit in Denmark provides DK 0.10/kWh (US 1.5 cents/kWh) for wind power, but few other countries have adopted similar credits. In the United States the Renewable Electricity Production Credit (PTC) provides a per kWh tax credit for electricity generated by qualified wind, closed-loop biomass, or poultry waste resources. Federal tax credits of 1.5 cents/kWh (adjusted annually for inflation) are provided for the first ten years of operation for all qualifying plants that entered service from 1992 through mid-1999, later extended to 2001 and then to 2003.

At least five U.S. states have state or local production incentives for distributed electrical generation, renewable fuels, or both. These policies are similar to the federal PTC, with specific limits on technologies, dates-in-service, and maximum payout per provider and per year. Funds to support the incentives are obtained from a mixture of sources, including general funds, public benefit or environmental funds, and green electricity sales (so-called green tags).

2.2.2.4 Property Tax Incentives At least 24 U.S. states have property tax incentives for renewable energy. These incentives are implemented on many scales—state, county, city, town, and municipality. These are generally implemented in one of three ways: (1) renewable energy property is partially or fully excluded from property tax assessment, (2) renewable energy property value is capped at the value of an equivalent conventional energy system providing the same service, and (3) tax credits are awarded to offset property taxes.

2.2.2.5 Personal Income Tax Incentives Credits against personal state income taxes are available in many U.S. states for purchase of or conversion to eligible renewable energy systems and renewable fuels. In some cases, taxpayers can deduct the interest paid on loans for renewable energy equipment.

2.2.2.6 Sales Tax Incentives At least 16 U.S. states have policies that provide retail sales tax exemptions for eligible renewable energy systems and renewable fuels. Most exempt 100% of the sales tax for capital expenses and provide specific cents-per-gallon exemptions for renewable fuels. Some policies specify maximum or minimum sizes for eligible systems.

2.2.2.7 Pollution Tax Exemptions The Netherlands is an example where so-called green power is exempt from a new and rising fossil-fuel tax on electricity generation that is paid by end users. Starting in 2001, that fossil-fuel tax rose to the equivalent of US 5 cents/kWh, providing a large tax incentive for Dutch consumers.

2.2.2.8 Other Tax Policies A variety of other tax policies exist, such as income tax exemptions on income from renewable power production, excise duty and sales tax exemptions on equipment purchased, and reduced or zero import tax duties on assembled renewable energy equipment or on components. India, for example, has allowed 5-year tax exemptions on income from sales of wind power.

2.2.3 Grants

Many countries have offered grants for renewable energy purchases. For example, beginning in 1979, Denmark provided rebates of up to 30% of capital costs for wind and other renewable energy technologies. These rebates declined over time. In the United States, county and state governments and utilities provide grants for renewable energy ranging in size from hundreds to millions of dollars.

2.2.4 Loans

Loan programs offer financing for the purchase of renewable energy equipment. Loans can be market rate, low interest (below-market rate), or forgivable.

In many U.S. states, loans are available to virtually all sectors—residential, commercial, industrial, transportation, public, and nonprofit. Repayment schedules vary, with terms of up to 10 years common. Interest rates for renewable energy investments can often be 1% or more higher than those for conventional power projects because of the higher perceived risks involved, so government-subsidized loans that offer below-market interest rates are also common.

Renewable energy loans can take many forms. Residential loans may range from $500 to $10,000 or more, while commercial and industrial loans may extend to the millions. Funding comes from a variety of sources, including municipal bonds, system benefit funds, revolving funds, and utility penalty or overcharge funds. Financing may be for a fraction to 100% of a project. Some loan programs have minimum or maximum limits, while others are open ended. Loan terms range from 3 years to the life of a project. Some loans are contractor driven and may include service contracts in the loan amount. Sometimes grants and loans are combined; for example, Iowa provides a 20% forgivable loan combined with an 80% loan at prime rate for renewable fuels projects.

In some developing countries, notably India, China, and Sri Lanka, multilateral loans by lenders such as the World Bank have provided financing for renewable energy, usually in conjunction with commercial lending. One of the most prominent examples is the India Renewable Energy Development Agency (IREDA), which was formed in 1987 to provide assistance in obtaining international multilateral agency loans and in helping private power investors obtain commercial loans. By 2001, IREDA had disbursed the equivalent of more than US$400 million in loans for renewable energy projects in India, resulting in more than 1600 MW of renewable power generation.

2.3 Public Investments and Market Facilitation Activities

2.3.1 Public Benefit Funds

In the United States, public funds for renewable energy development are raised through a System Benefits Charge (SBC), which is a per-kWh levy on electric power consumption. Some analysts suggest that state clean energy funds seem to be one of the more effective policies in promoting renewable energy development to result from electricity restructuring. It is estimated that 14 U.S. states will collect $3.5 billion through 2011 in system benefits charges. Similar levies exist in some European countries for fossil-fuel-based generation. In general, the funds serve a variety of purposes, such as paying for the difference between the cost of renewables and traditional generating facilities, reducing the cost of loans for renewable facilities, providing energy efficiency services, funding public education on energy-related issues, providing low-income energy assistance, and supporting research and development.

2.3.2 Infrastructure Policies

Market facilitation supports market institutions, participants, and rules to encourage renewable energy technology deployment. A variety of policies are used to build and maintain this market infrastructure, including policies for design standards, accelerated siting and permitting, equipment standards, and contractor education and licensing. Additionally, policies to induce renewable technology manufactures to site locally and direct sales of renewable systems to customers at concessionary rates facilitate market development.

2.3.2.1 Construction and Design Policies Construction and design standards include building-code standards for PV installations, design standards evaluated on life-cycle cost basis, and performance requirements. Policy examples include Tucson, Arizona, which requires that commercial facilities achieve a 50% reduction in energy usage over 1995 Model Energy Code, and Florida, which requires that all new educational facilities include passive solar design.

2.3.2.2 Site Prospecting, Review and Permitting Federal and state programs reduce barriers to renewable energy development through resource, transmission, zoning, and permitting assessments. This particularly helped early promotion of wind energy projects in California. On a national scale the Utility Wind Resource Assessment Program funds a number of supporting activities, including up to 50% of the cost of wind resource assessments. India also has a large wind assessment program, with more than 600 stations in 25 states providing information to project developers on the best sites for development.

2.3.2.3 Equipment Standards and Contractor Certification A variety of equipment-related standards

and certification measures have been applied to ensure uniform quality of equipment and installation, increasing the likelihood of positive returns from renewable energy installations. Contractor licensing requirements ensure that contractors have the necessary experience and knowledge to properly install systems. Equipment certifications ensure that equipment meets certain minimum standards of performance or safety.

2.3.2.4 Industrial Recruitment Industrial recruitment policies use financial incentives such as tax credits, grants, and government procurement commitments to attract renewable energy equipment manufacturers to locate in a particular area. These incentives are designed to create local jobs, strengthening the local economy and tax base, and improving the economics of local renewable development initiatives.

2.3.2.5 Direct Equipment Sales These programs allow the consumer to buy or lease renewable energy systems directly from electric provider at below-retail rates. Some programs provide a capital buydown. Examples include Arizona, which provides a buydown of $2/peak-watt for PV, and California's Sacramento Municipal Utility District, which offers a 50% buydown plus 10-year financed loans and net metering.

2.3.3 Government Procurement

Government procurement policies aim to promote sustained and orderly commercial development of renewable energy. Governmental purchase agreements can reduce uncertainty and spur market development through long-term contracts, preapproved purchasing agreements, and volume purchases. Government purchases of renewable energy technologies in early market stages can help overcome institutional barriers to commercialization, encourage the development of appropriate infrastructure, and provide a market path for technologies that require integrated technical, infrastructure, and regulatory changes.

2.3.4 Customer Education and Mandated Generation Disclosure Information

U.S. restructuring and deregulation policies mandate that information be provided to customers about choice of electricity providers and characteristics of electricity being provided (such as emissions and fuel types). In many states, general education to raise customer awareness about renewable energy and the environmental impacts of energy generation is required, typically via websites and printed materials.

2.3.5 Solar and Wind Access Laws

Renewable access laws address access, easements, and covenants. Access laws provide a property owner the right to continued access to a renewable resource. Easements provide a privilege to have continued access to wind or sunlight, even though development or features of another person's property could reduce that access. Easements are often voluntary contracts and may be transferred with the property title. Covenant laws prohibit neighborhood covenants from explicitly restricting the installation or use of renewable equipment. Policy mechanisms include access ordinances, development guidelines addressing street orientation, zoning ordinances with building height restrictions, and renewable permits.

3. TRANSPORT BIOFUELS POLICIES

Biofuels mandates and tax policies in Brazil, the United States, and Europe have supported accelerating development of biofuels. Biofuels mandates require that a certain percentage of all liquid transport fuels be derived from renewable resources. Tax policies may provide tax credits or exemptions for production or purchase of biofuels.

Brazil has been quite successful with biofuels mandates under its ProAlcool program, which has promoted the use ethanol for transportation fuel since the 1980s. In addition to a variety of economic incentives and subsidies, Brazil has mandated that ethanol be blended with all gasoline sold in the country. Brazil has also required that all gas stations sell pure ethanol. This last requirement made it commercially viable for the automotive industry to produce ethanol-only (neat ethanol) cars. However, the share of ethanol-only cars purchased annually, after rising to 95% by 1985, subsequently declined for a number of reasons. Most cars now use the "gasohol" ethanol/gasoline blend, and more than 60% of Brazil's sugar cane production goes to produce 18 billion gallons of ethanol each year, representing 90% of global ethanol production. In 2000, over 40% of automobile fuel consumption in Brazil was ethanol.

The United States, the world's second largest ethanol producer after Brazil, has a number of biofuels tax policies and mandates. The Energy Security Act of 1979 created a federal ethanol tax credit of up to 60 cents per gallon for businesses that sell or use alcohol as a fuel. Gasoline refiners and distributors may also receive an excise tax exemption

of up to 5 cents/gallon for blending their fuels with ethanol. State-level ethanol policies also exist, whose origins in the 1980s can be traced back to initiatives by Iowa to use its corn crop for energy. Several policies in Iowa were established to encourage ethanol consumption, including a mandate for government vehicles to use ethanol-blended fuel, and a 1-cent-per-gallon fuel sales tax exemption for ethanol-blended fuels. In 1998, both the federal government and the state of Iowa extended their ethanol tax exemptions until the year 2007. In part due to ethanol incentives, more than 60 ethanol production facilities have become operational in the United States since 1976, with a production capacity of more than 2.4 billion gallons per year. A recent U.S. legislative proposal, called a renewable fuels standard, would triple biofuel use within 10 years.

In Europe, Germany is the largest user of biodiesel and provides tax incentives for 100% pure biodiesel. These incentives have had a large effect; German consumption of biodiesel went from 200 million gallons in 1991 to 750 million gallons in 2002. Other EU members provide tax incentives for 2 to 5% biodiesel blends. Germany, Austria, and Sweden use 100% pure biodiesel in specially adapted vehicles, and biodiesel is mandated in environmentally sensitive areas in Germany and Austria. Other countries producing biodiesel are Belgium, France, and Italy, the later two also providing tax incentives. The European Commission has recently proposed a biofuel directive with targets or mandates for biofuels up to 6% of all transportation fuel sold.

Among other developing countries, Thailand is considering tax policies for biofuels, including excise tax exemptions for ethanol and income tax waivers for investments in biofuels facilities. Malaysia and Indonesia also utilize biodiesel.

4. EMISSIONS REDUCTION POLICIES

Policies to reduce power plant emissions, including NO_x, SO_x, and CO_2, have the potential to affect renewable energy development. Many emissions-reduction policies create allowances for certain emissions (representing the right to emit a certain amount of that pollutant). Credits available to renewable energy generation can offset these allowed emissions. Such credits have market value and are often traded to allow electricity generators to comply with emission regulations at least cost. Three innovative examples of U.S. air quality standards, acid rain prevention programs, and state-level greenhouse-gas reduction initiatives illustrate the potential of emissions reductions policies to affect renewable energy.

4.1 Renewable Energy Set-Asides

To meet National Ambient Air Quality Standards, the U.S. Environmental Protection Agency requires 22 U.S. states and the District of Columbia to reduce NO_x emissions significantly by 2007. States can meet emission reduction targets through actual emission reductions or purchase of emission reduction credits from other states participating in a region-wide NO_x trading program. States can allocate, or set-aside, a percentage of the total state NO_x allowances to energy efficiency and renewable energy. Eligible renewable energy producers receive these set-aside allowances and can sell them to fossil-fuel-based electricity generators to enable those generators to stay within their NO_x allocation. The additional revenue from sales of these set-aside allowances can potentially provide stimulus for renewable energy development, although to date few states have implemented renewable energy set-asides.

4.2 Emissions Cap-and-Trade Policies

Under the 1990 Clean Air Act the United States instituted a cap-and-trade mechanism to reduce SO_2 emissions and facilitate least-cost compliance. Under the Acid Rain subprogram, 300,000 SO_2 emissions allowances (rights to emit SO_2) were set aside for utilities that employed renewable energy or energy-efficiency measures. Allowances were to be earned from 1992 through 1999, allocated at a rate of one allowance (one ton of SO_2 avoided) per 500 MWh of generation produced by renewable energy or avoided through increased energy efficiency. This program was not particularly effective, as only one-tenth of the 300,000 allowances were allocated to energy efficiency or renewable energy. Analysis suggests that allowance prices set by the market were not high enough to justify renewable energy investments. In addition, the program restricted participation to utilities, excluding other power generators, which also contributed to under-subscription.

4.3 Greenhouse Gas Mitigation Policies

The New Jersey Sustainability Greenhouse Gas (GHG) Action Plan was designed to promote the capture of landfill methane gas for power generation, thus avoiding methane emissions (methane has a

high global warming potential). A loan fund provides New Jersey companies with low-interest loans to support initiatives including development of biomass/landfill methane energy resources, and emission offsets from landfill methane and other renewable power generation could potentially be included in GHG trading systems.

5. POWER-SECTOR RESTRUCTURING POLICIES

Power sector restructuring is having a profound effect on electric power technologies, costs, prices, institutions, and regulatory frameworks. Restructuring trends are changing the traditional mission and mandates of electric utilities in complex ways, and affecting environmental, social, and political conditions. There are five key trends underway that continue to influence renewable energy development, both positively and negatively, as discussed next.

5.1 Competitive Wholesale Power Markets and Removal of Price Regulation on Generation

Power generation is usually one of the first aspects of utility systems to be deregulated. The trend is away from utilities monopolies toward open competition, where power contracts are signed between buyers and sellers in wholesale power markets. Distribution utilities and industrial customers gain more choices in obtaining wholesale power. Such markets may often begin with independent-power-producer (IPP) frameworks. As wholesale electricity becomes more of a competitive market commodity, price becomes relatively more important than other factors in determining a buyer's choice of electricity supplier.

The potential effects of competitive wholesale markets and IPPs on renewable energy are significant. Wholesale power markets allow IPPs to bypass the biases against renewables that traditional utility monopolies have had. Indeed, one of the very first major markets for renewable energy in the 1980s was in California, spurred by the PURPA legislation discussed earlier. In some countries, IPP frameworks have been explicitly enacted to support renewable energy. Examples are Sri Lanka and Thailand, where utility monopolies were broken and renewable energy IPPs can sell power to the grid. However, other effects of wholesale competition may stifle renewable energy development. As low-cost combined-cycle gas turbines begin to dominate new generation, renewable energy has difficulty competing on the basis of price alone. In addition, the emergence of short-term power contracts and spot markets favor generation technologies with higher variable costs and lower capital costs, like fossil fuels, rather than capital-intensive but low-operating-cost technologies like renewables.

5.2 Self-Generation by End Users and Distributed Generation Technologies

Independent power producers may be the end users themselves rather than just dedicated generation companies. With the advent of IPP frameworks, utility buyback schemes (including net metering), and cogeneration technology options, more and more end users, from large industrial customers to small residential users, are generating their own electricity. Their self-generation offsets purchased power and they may even sell surplus power back to the grid. Traditionally, regulated monopoly utilities have enjoyed economic advantages from large power plants and increasing economies of scale. These advantages are eroding due to new distributed generation technologies that are cost competitive and even more efficient at increasingly smaller scales. In fact, newer technologies reduce investment risks and costs at smaller scales by providing modular and rapid capacity increments.

Renewable energy is well suited to self-generation but faces competition from other distributed generation technologies, especially those based on natural gas. Gas has become the fuel of choice for small self-producers because of short construction lead times, low fuel and maintenance costs, and modular small-scale technology. However, with restructuring, a host of distributed generation policies, including net metering, become possible (see the section on distributed generation policies). These policies often spur renewable energy investments. On the other hand, self-generators may be penalized by utility-wide surcharges that accompany restructuring, such as those for stranded generation assets (called nonbypassable competitive transition charges in the United States). Self-generators who use renewable energy must still pay these charges, based on the amount of electricity they *would* have purchased from the grid, even if actual grid consumption is small.

5.3 Privatization or Commercialization of Utilities

In many countries, utilities, historically government owned and operated, are becoming private for-profit entities that must act like commercial corporations. Even if utilities remain state owned, they are becoming commercialized—losing state subsidies and becoming subject to the same tax laws and accounting rules as private firms. In both cases, staffing may be reduced and management must make independent decisions on the basis of profitability.

The effects of privatization and commercialization on renewable energy are difficult to judge. The environmental effects of privatization can be positive or negative, depending on such factors as the strength of the regulatory body and the political and environmental policy situation in a country. Private utilities are more likely to focus more on costs and less on public benefits, unless specific public mandates exist. On the positive side, privatization may promote capital-intensive renewable energy by providing a new source of finance—capital from private debt and equity markets. However, the transition from public to private may shorten time horizons, increase borrowing costs, and increase requirements for high rates of return. All of these factors would limit investments in capital-intensive renewable energy projects in favor of lower-capital-cost, higher-operating-cost fossil-fuel technologies.

5.4 Unbundling of Generation, Transmission, and Distribution

Utilities have traditionally been vertically integrated, including generation, transmission, and distribution functions. Under some restructuring programs, each of these functions is being unbundled into different commercial entities, some retaining a regulated monopoly status (particularly distribution utilities) and others starting to face competition (particularly generators).

Unbundling can provide greater consumer incentives to self-generate using renewable energy. If retail tariffs are unbundled as well, so that generation, transmission, and distribution costs are separated, customers have more incentive to self-generate, thereby avoiding transmission and distribution charges. In addition, open-access transmission policies that go along with unbundling have been explicitly targeted to promote renewable energy in some countries. In India, open-access policies helped catalyze the wind industry there by allowing firms to produce wind power in remote regions with good wind resources and then "wheel" the power over the transmission system to their own facilities or to third parties. Brazil enacted a 50% reduction of transmission wheeling fees for renewable energy producers, which has been credited with promoting a booming small hydro industry there. However, unbundling can also penalize inherently intermittent renewable energy if producers have to pay transmission charges on a per-capacity basis. That is, even when the transmission capacity is not being used (say the wind is not blowing), transmission charges must be paid, resulting in high average transmission costs per kWh.

5.5 Competitive Retail Power Markets and Green Power Sales

Competition at the retail level, the newest phenomena in power sector restructuring, means that individual consumers are free to select their power supplier from among all those operating in a given market. Competitive retail power markets have allowed the emergence of so-called green power suppliers who offer to sell renewable energy, usually at a premium. As green power sales grow, these suppliers are forced to investment in new renewable energy capacity to meet demand or buy power from other renewable energy producers. Green power markets have begun to flourish where retail competition is allowed, but often only in conjunction with other renewable energy promotion policies.

The Netherlands is perhaps the best-known example. Following restructuring in 2001, 1 million green power customers signed up within the first year. However, incentives played a role; a large tax on fossil-fuel generated electricity, from which green power sales were exempt, made green power economically competitive with conventional power. In the United States, green power markets are emerging in several states in response to state incentives and aggressive marketing campaigns by green power suppliers. At least 30 U.S. states have green pricing programs. Four states have mandatory green power policies that require utilities to offer customers opportunities to support renewable energy. California became one of the largest markets, with more than 200,000 customers, but this was aided by a 1 cent/kWh subsidy to green power, paid for by a system benefits charge (see the section on renewable energy promotion policies).

6. DISTRIBUTED GENERATION POLICIES

Distributed generation avoids some of the costs of transmission and distribution infrastructure and power losses, which together can total up to half of delivered power costs. Policies to promote distributed generation—including net metering, real-time pricing, and interconnection regulations—do not apply only to renewable energy but nevertheless can strongly influence renewable energy investments.

6.1 Net Metering

Net metering allows a two-way flow of electricity between the electricity distribution grid and customers with their own generation. When a customer consumes more power than it generates, power flows from the grid and the meter runs forward. When a customer installation generates more power than it consumes, power flows into the grid and the meter runs backward. The customer pays only for the net amount of electricity used in each billing period and is sometimes allowed to carryover net electricity generated from month to month. Net metering allows customers to receive retail prices for the excess electricity they generate at any given time. This encourages customers to invest in renewable energy because the retail price received for power is usually much greater than it would be if net metering were not allowed and customers had to sell excess power to the utility at wholesale rates or avoided costs. Electricity providers may also benefit from net metering programs, particularly with customer-sited PV, which produces electricity during peak periods. Such peak power can offset the need for new central generation and improve system load factors.

At least 38 U.S. states now have net metering laws. Size limits on net metered systems typically range from 10 kW to 100 kW, with the exception of a few states that do not limit system size or overall enrollment. Net metering is common in parts of Germany, Switzerland, and the Netherlands and allowed by at least one utility in the UK. Thailand is one of the few developing countries to have enacted net metering laws, following a pilot project for net-metered rooftop PV in the 1990s.

6.2 Real-Time Pricing

Real-time pricing, also known as dynamic pricing, is a utility rate structure in which the per-kWh charge varies each hour based on the utility's real-time production costs. Because peaking plants are more expensive to run than baseload plants, retail electricity rates are higher during peak times than during shoulder and off-peak times under real-time pricing. When used in conjunction with net metering, customers receive higher peak rates when selling power into the grid at peak times. At off-peak times the customer is likely purchasing power from the grid, but at the lower off-peak rate. Photovoltaic power is often a good candidate for real-time pricing, especially if maximum solar radiation occurs at peak-demand times of day when power purchase prices are higher. Real-time metering equipment is necessary, which adds complexity and expense to metering hardware and administration.

Real-time pricing has been used with some large power consumers for decades. For example, power companies in Nova Scotia and New York state offer real time pricing rates for large commercial and industrial customers that vary hourly according to the varying cost of generation. In a recent pilot project, California installed 23,000 real-time meters for large customers at a cost of $35 million. In response, summer peak demand by those customers dropped by 500 MW under time-of-use pricing, which would allow the utility to avoid $250 million to $300 million in capacity additions. Although real-time pricing has not become widespread, with favorable rate structures it has the potential to provide significant incentives for grid-connected renewable development.

6.3 Interconnection Regulations

Nondiscriminatory interconnection laws and regulations are needed to address a number of crucial barriers to interconnection of renewable energy with the grid. Interconnection regulations often apply to both distributed generation and remote generation with renewable energy that requires transmission access, such as wind power.

6.3.1 Legal Access

The ability to legally connect a renewable energy system to a grid depends on federal, state, and local government rules and regulations. These policies both allow connection and determine how physical connection is achieved. In the United States, the legal right to connect to the grid is provided for in federal laws such as the Public Utilities Regulatory Policies Act (PURPA) of 1978 and by state net metering statutes.

6.3.2 Dynamic Generation and Transmission Scheduling

Historically, transmission policies have often imposed severe penalties on unscheduled deviations from projected (advance-scheduled) power generation. These penalty structures render intermittent generation, such as wind or PV, uneconomic. Real-time accounting of power transfer deviations that provides charges or credits to producers based on the value of energy at the time of the deviation, as well as elimination of discriminatory deviation penalties, allows intermittent renewable energy to compete more equitably with traditional generation. Policies that allow near-time or real-time scheduling of the output levels of intermittent resources can further reduce deviation costs. For example, wind farms are able to predict their output much more accurately up to an hour in advance of generation and thus can be better scheduled hour by hour rather than a day ahead.

6.3.3 Elimination of Rate Pancaking

Because distributed renewables, such as wind, are often remotely located, they can incur high transmission fees as power crosses multiple jurisdictions to get to the customer. Such cumulative addition of transmission fees is known as rate pancaking. Elimination of access rate pancaking, either by consolidation of long-distance tariffs under a regional transmission organization or by creating access waiver agreements between multiple owner/operators can reduce discrimination against wind and other remote distributed renewables.

6.3.4 Capacity Allocation

When demand for a transmission path exceeds its reliable capacity, transmission congestion occurs. In such circumstances, system operators must allocate available capacity among competing users. Traditional utility policies often favor early market entrants, grandfathering them into capacity allocation rules. Wind power is particularly susceptible to transmission constraints, as it is generally located far from load. Elimination of grandfathering would allow transmission users to bid for congested capacity on an equal footing. Allowing wind to bid for congested capacity closer to the operating hour and reducing congestion through transmission line upgrades would also reduce barriers to wind energy development.

6.3.5 Standard Interconnection Agreements

Utilities may require the same interconnection procedures for small systems as are required for large independent power-production facilities. The process of negotiating a power purchase/sale contract with the utility can be very expensive, and utilities can charge miscellaneous fees that greatly reduce the financial feasibility of small grid-connected renewable installations. Standardized interconnection agreements can expedite this process. Some believe that Texas provides a good model for renewable interconnection. Under a standard agreement, renewable developers pay only for the direct costs of connecting the plant to the local system, but not for upgrades to the grid necessary to carry additional capacity. This allows generators to compete more equally.

7. RURAL ELECTRIFICATION POLICIES

Historically, renewable energy in developing countries has come from direct donor assistance and grants for equipment purchases and demonstrations. In recent years a number of new approaches have emerged for promoting renewable energy in off-grid rural areas, including energy service concessions, private entrepreneurship, microcredit, and comparative line extension analysis.

7.1 Rural Electrification Policy and Energy Service Concessions

Many developing countries have explicit policies to extend electric networks to large shares of rural populations that remain unconnected to power grids (globally, an estimated 1.7 billion people). However, in many areas, full grid extension is too costly and unrealistic. Policies and rural electrification planning frameworks have recently started to emerge that designate certain geographic areas as targets for off-grid renewable energy development. These policies may also provide explicit government financial support for renewable energy in these areas. Such financial support is starting to be recognized as a competitive alternative to government subsidies for conventional grid extensions. Countries taking the lead with such policies include Argentina, China, India, Morocco, the Philippines, South Africa, and Sri Lanka.

One form this government support can take is so-called energy service concessions. With a concession, the government selects one company to exclusively serve a specific geographic region, with an obligation to serve all customers who request service. The

government also provides subsidies and regulates the fees and operations of the concession. Rural energy-service concessions may employ a mixture of energy sources to serve customers, including diesel generators, minihydro, photovoltaic, wind, and biomass. Argentina, Morocco, and South Africa have initiated policies to develop rural concessions, with ambitious targets of 200,000 rural households in South Africa and 60,000 in Argentina. But the actual experience with this approach has been limited so far to just a few thousand installations.

7.2 Rural Business Development and Microcredit

Private entrepreneurship is increasingly recognized as an important strategy to fulfill rural energy goals. Thus, rural electrification policies have begun to promote entrepreneurship. Promising approaches are emerging that support rural entrepreneurs with training, marketing, feasibility studies, business planning, management, financing, and connections to banks and community organizations. These approaches include bundling renewable energy with existing products. Bundling can reduce costs if vendors of existing products and services add renewable energy to their activities—and use their existing networks of sales outlets, dealers, and service personnel. Dealers of farm machinery, fertilizers, pumps, generators, batteries, kerosene, LPG, water, electronics, telecommunications, and other rural services can bundle renewable energy with these services.

In conjunction with entrepreneurship, consumer microcredit has emerged as an important tool for facilitating individual household purchases of renewable energy systems like solar home systems. Credit may be provided either by the system vendors themselves, by rural development banks, or by dedicated microcredit organizations. Notable examples of consumer microcredit for solar home systems have emerged in five developing countries. In Bangladesh, Grameen Shakti, a nonprofit vendor, has offered consumer credit for terms up to 3 years. The Vietnam Women's Union offered similar credit terms for systems sold by private vendors. In Sri Lanka, Sarvodaya, a national microfinance organization, has offered credit on terms up to 5 years. In Zimbabwe, vendors sold several thousand systems on credit provided by the Agricultural Finance Corporation. And in India, new forms of rural microcredit have started to emerge. By 2002, the cumulative number of solar home system purchases made with credit in these countries had exceeded 50,000, but this was still a small fraction of the total number of solar home systems worldwide, estimated at 1.2 million.

7.3 Comparative Line Extension Analyses

Economic comparisons of line extension versus distributed renewable energy investment are also emerging in developed countries. At least four U.S. states have power line extension policies requiring that, in cases where utility customers must pay a portion of construction costs for utility power line extension to a remote location, the utility must provide information about onsite renewable energy technology options. Some of these policies require the utility to perform a cost-benefit analysis comparing line extension with off-grid renewable energy. Renewable energy options may be less expensive for rural customers, but without line extension policies, many customers would not be aware of this.

8. SUMMARY

Public support for renewable energy expanded rapidly in the late 1990s and early 2000s. A wide variety of policies are designed explicitly to promote renewable energy, while other policies focus on power sector restructuring or environmental issues more broadly and have more indirectly affected renewable energy. Experience with renewable energy policies around the world is still emerging, and more understanding is needed of the impacts of various policies. Thus, many policies could still be considered experimental in nature. Of all the policies surveyed in this article, the ones that appear to have contributed the most to renewable energy development during the 1990s and early 2000s are (1) direct equipment subsidies and rebates, net metering laws, and technical interconnection standards in the case of solar PV; (2) investment tax credits, production tax credits, and European electricity feed-in laws in the case of wind; (3) grid-access and wheeling policies supporting independent power producers and third-party sales in the case of biomass and small hydro power. Many of the trends toward restructuring of power-sector institutions and regulation that were underway throughout the 1990s have had both positive and negative influences on renewable energy. Policymakers and policy advocates have many options from which to choose and a slowly emerging body of experience and results to guide those choices.

SEE ALSO THE FOLLOWING ARTICLES

Computer Modeling of Renewable Power Systems • International Energy Law and Policy • Nongovernmental Organizations (NGOs) and Energy • Public Reaction to Renewable Energy Sources and Systems • Renewable Energy and the City • Renewable Energy in Europe • Renewable Energy in Southern Africa • Renewable Energy in the United States • Renewable Portfolio Standard

Further Reading

Bolinger, M., Wiser, R., Milford, L., Stoddard, M., and Porter, K. (2001). States emerge as clean energy investors: A review of state support for renewable energy. *Electricity J.* (Nov.), 82–95.

Dunn, S. (2000). "Micropower: The Next Electrical Era." Worldwatch Institute, Washington, DC.

Hirsh, R. F., and Serchuk, A. H. (1999). Power switch: Will the restructured electric utility system help the environment? *Environment* **41,** 32–39. 4–9.

International Energy Agency. (2003). "Renewable Energy Policies and Measures in IEA Countries." Paris.

Martinot, E., Chaurey, A., Moreira, J., Lew, D., and Wamukonya, N. (2002). Renewable energy markets in developing countries. *Ann. Rev. Energy Environment* **27,** 309–348.

Rader, N., and Hempling, S. (2001). "The Renewables Portfolio Standard: A Practical Guide." National Association of Regulatory Utility Commissioners, Washington, DC.

Reiche, D. (2002). "Handbook of Renewable Energy in the European Union: Case Studies of All Member States." Verlag Peter Lang, Bern.

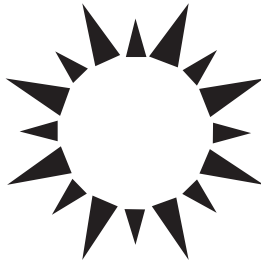
Sawin, J. L. (2001). "The Role of Government in the Development and Diffusion of Renewable Energy Technologies: Wind Power in the United States, California, Denmark, and Germany, 1970–2000." Ph. D. thesis, Tufts University, Fletcher School of Law and Diplomacy. UMI, Ann Arbor, MI.

United Nations Environment Programme (2000). "Natural Selection: Evolving Choices for Renewable Energy Technology and Policy." Paris.

U.S. Energy Information Administration (2001). "Renewable Energy Annual." U.S. Government Printing Office. Washington, DC.

Wiser, R., Bolinger, M., Milford, L., Porter, K., and Clark, R. (2002). "Innovation, Renewable Energy, and State Investment: Case Studies of Leading Clean Energy Funds." Lawrence Berkeley National Laboratory, Berkeley, CA.

Renewable Energy, Taxonomic Overview



DANIEL M. KAMMEN
University of California, Berkeley
Berkeley, California, United States

1. Introduction
2. Biomass Energy
3. Wind Energy
4. Solar-Photovoltaic and Solar-Thermal Technologies
5. Hydropower
6. Geothermal Energy
7. Renewable Energy System Cost and Performance
8. Conclusions

Glossary

bioenergy The energy derived from any cellulosic plant or agricultural or animal material that is used a fuel.
hydropower The energy derived, at a variety of scales, from water pressure; used to drive turbines.
learning curve A mathematical description of the relationship between the average price of a manufactured commodity and the market price. For a wide variety of materials, the decrease in market price is roughly 20% for each doubling of the total quantity of the item manufactured.
photovoltaics Semiconductor technology involving photon bombardment, which permits the flow of electrons.
renewable energy The energy derived from solar, wind, biomass, geothermal, or hydro sources. There is some debate about hydropower, of which the different sources are considered either to all fall into the renewable category, or are considered renewable only when involving either a small system or a "run of the river" technology.
wind energy The energy contained in the movement of air masses, generally captured by the use of mechanical blades driven by wind energy.

The potential of renewable energy sources is enormous, because, in principle, they can meet many times the world's energy demand. Renewable energy sources such as biomass, wind, solar, hydropower, and geothermal can provide sustainable energy services, based on the use of routinely available, indigenous resources. As the costs of renewable sources decline and the price of oil and gas continue to fluctuate, a transition to renewables-based energy systems is looks increasingly likely. In the past 30 years, solar and wind power systems have experienced rapid sales growth, declining capital costs, declining costs of electricity generated, and continued improvement in performance characteristics. In fact, fossil fuel and renewable energy prices, and the social and environmental costs of each, are heading in opposite directions, and the economic and policy mechanisms needed to support the widespread dissemination and sustainable markets for renewable energy systems are rapidly evolving. It is becoming clear that future growth in the energy sector will be primarily in the new regime of renewable energy, and to some extent natural gas-based systems, not in conventional oil and coal sources. Because of these developments, market opportunity now exists to both innovate and to take advantage of emerging markets to promote renewable energy technologies, with the additional assistance of governmental and popular sentiment. The development and use of renewable energy sources can enhance diversity in energy supply markets, contribute to securing long-term sustainable energy supplies, help reduce local and global atmospheric emissions, and provide commercially attractive options to meet specific energy service needs, particularly in developing countries and rural areas, creating new employment opportunities.

1. INTRODUCTION

Conventional energy sources based on oil, coal, and natural gas have proved to be highly effective drivers

 © 2004 Elsevier Inc. All rights reserved.

of economic progress, but at the same time damaging to the environment and to human health. Furthermore, fossil fuel-driven economies tend to be cyclical in nature, due to the effects of oligopoly in production and distribution. The traditional fossil fuel-based energy sources are facing increasing pressure on a host of environmental fronts, with perhaps the most serious challenge confronting the future use of coal being the Kyoto Protocol greenhouse gas (GHG) reduction targets. It is now clear that any effort to maintain atmospheric levels of CO_2 below even 550 ppm cannot be based fundamentally on an oil- and coal-powered global economy, barring radical carbon sequestration efforts.

The potential of renewable energy sources is enormous; they have the potential to meet many times the world's energy demand. Renewable energy sources such as biomass, wind, solar, hydropower, and geothermal can provide sustainable energy services, based on the use of routinely available, indigenous resources. A transition to renewables-intensive energy future is looking increasingly likely. The costs of solar and wind power systems have dropped substantially in the past 30 years, and continue to decline, whereas the price of oil and gas continue to fluctuate. Renewable energy prices and the social and environmental costs of using fossil fuels are heading in opposite directions. Furthermore, the economic and policy mechanisms needed to support the widespread dissemination and sustainable markets for renewable energy systems have also rapidly evolved. It is becoming clear that future growth in the energy sector is primarily in the new regime of renewable (and to some extent natural gas-based) systems, and not in conventional oil and coal sources. Financial markets are awakening to the future growth potential of renewable and other new energy technologies, and this is a likely harbinger of the economic reality of truly competitive renewable energy systems.

Renewable energy systems are usually founded on a small-scale, decentralized paradigm that is inherently conducive to, rather than at odds with, many electricity distribution, cogeneration (combined heat and power), environmental, and capital cost issues. As an alternative to custom, onsite construction of centralized power plants, renewable systems based on photovoltaic (PV) arrays, windmills, biomass, or small hydropower setups can be viewed as mass-produced "energy appliances" capable of being manufactured at low cost and tailored to meet specific energy loads and service conditions. These systems can have dramatically reduced as well as widely dispersed environmental impacts, rather than larger, more centralized impacts that in some cases are serious contributors to ambient air pollution, acid rain, and global climate change.

Renewable energy sources currently supply somewhere between 15 and 20% of the world's total energy demand. The supply is dominated by traditional biomass, mostly fuel wood used for cooking and heating, especially in developing countries in Africa, Asia, and Latin America. A major contribution is also obtained from the use of large hydropower systems with nearly 20% of the global electricity supply being provided by this source. New renewable energy sources (solar energy, wind energy, modern bioenergy, geothermal energy, and small hydropower) are currently contributing about 2%. A number of scenario studies have investigated the potential contribution of renewables to global energy supplies, indicating that in the second half of the 21st century their contribution might range from the present figure of nearly 20% to more than 50%, with the right policies in place.

2. BIOMASS ENERGY

2.1 Introduction

Biomass is the term used for all organic material originating from plants (including algae), trees, and crops and is essentially the collection and storage of the sun's energy through photosynthesis. Biomass energy, or bioenergy, is the conversion of biomass into useful forms of energy such as heat, electricity, and liquid fuels.

Biomass for bioenergy comes either directly from the land, as dedicated energy crops, or from residues generated in the processing of crops for food or other products, such as pulp and paper from the wood industry. Another important contribution is from postconsumer residue streams, including construction and demolition wood, pallets used in transportation, and the clean fraction of municipal solid waste (MSW). The process that converts biomass to bioenergy can be considered as management of the flow of solar-generated materials, food, and fiber in our society. These interrelationships are shown in Fig. 1, which presents the various resource types and applications, showing the flow of their harvest and residues to bioenergy applications. Not all biomass is directly used to produce energy, but rather can be converted into intermediate energy carriers called biofuels. This includes charcoal (higher energy

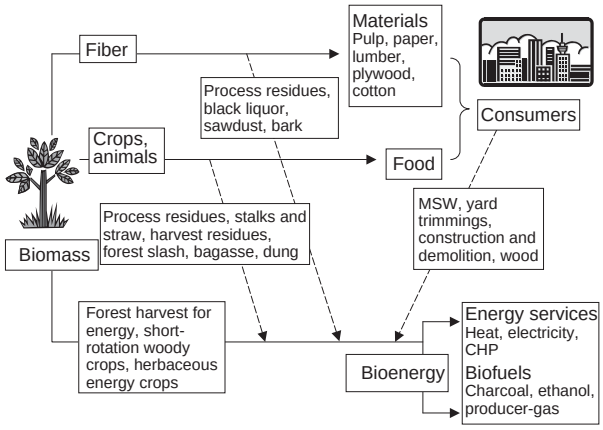


FIGURE 1 Biomass and bioenergy flowchart. MSW, Municipal solid waste; CHP, combined heat and power. From Overend (2000).

density solid fuel), ethanol (liquid fuel), or producer gas (from gasification of biomass).

Biomass was the first energy source harnessed by humans, and for nearly all of human history, wood has been our dominant energy source. Only during the past century, with the development of efficient techniques to extract and burn fossil fuels, have coal, oil, and natural gas replaced wood as the industrialized world's primary fuel. Today, some 40 to 55 exajoules (EJ $= 10^{18}$ joules) per year of biomass is used for energy, out of about 450 EJ per year of total energy use, or an estimated 10–14%, making it the fourth largest source of energy behind oil (33%), coal (21%), and natural gas (19%). The precise amount is uncertain because the majority is used noncommercially in developing countries, where biomass accounts for an estimated one-third of primary energy use. In the poorest countries, up to 90% of all energy is supplied by biomass. Over 2 billion people cook by direct combustion of biomass, and such traditional uses typically involve the inefficient use of biomass fuels, largely from low-cost sources such as natural forests, which can further contribute to human health impacts, deforestation, and environmental degradation.

The picture of biomass utilization in developing countries is sharply contrasted by that in industrialized countries. On average, biomass accounts for only 3–4% of total energy use in the latter, although where policies supportive of biomass use are in place, e.g., in Austria, Sweden, and Finland, the biomass contribution reaches 12, 18, and 23%, respectively. Most biomass in industrialized countries is converted into electricity and process heat in cogeneration systems (combined heat and power production) at industrial sites or at municipal district heating facilities. This enables a greater variety of energy services to be derived from the biomass sources, which are much cleaner than fossil fuels, and more efficient use of the available biomass resources than is typical in developing countries.

Biomass energy has the potential to be "modernized" worldwide—that is, produced and converted efficiently and cost-competitively into more convenient forms such as gases, liquids, or electricity. A variety of technologies can convert solid biomass into clean, convenient energy carriers over a range of scales, from household/village to large industrial. Some of these technologies are commercially available today whereas others are still in the development and demonstration stages. If widely implemented, such technologies could enable biomass energy to play a much more significant role in the future than it does today, especially in developing countries.

2.2 Future Role of Biomass

Modernized biomass energy is projected to play a major role in the future global energy supply. This is being driven not so much by the depletion of fossil fuels, which has ceased to be a defining issue with the discovery of new oil and gas reserves and the large existing coal resources, but rather by the recognized threat of global climate change, caused largely by the burning of fossil fuels. The carbon neutrality of biomass energy (when produced sustainably) and its relatively even geographical distribution, coupled with the expected growth in energy demand in developing countries, where affordable alternatives are not often available, make it a promising energy source in many regions of the world for the 21st century.

Most households in developing countries that use biomass fuels today do so either because it is available at low (or zero) financial cost or because they lack access to or cannot afford higher quality fuels. As incomes rise, preferences tend to shift away from biomass. For example, in the case of cooking, consumer preferences shift with increasing income from dung to crop residues, fuel wood, coal, charcoal, kerosene, liquefied petroleum gas, natural gas, and electricity (the well-known household energy ladder). This shift away from biomass energy as incomes rise is associated with the quality of the energy carrier used, rather than with the primary energy source. If biomass energy is instead modernized, then wider use is conceivable, along with

TABLE I

Role of Biomass in Future Global Energy Use[a]

Source	Time frame (year)	Total projected global energy demand (EJ/year)	Contribution of biomass to energy demand [EJ/year (percentage of total)]	Remarks
Shell Petroleum (1996)	2060	940	207 (22%)	Sustained growth[b]
	2060	1500	221 (14%)	Dematerialization[c]
Intergovernmental Panel on Climate Change (1996)	2050	560	180 (32%)	Biomass-intensive energy system development
	2100	710	325 (46%)	
Greenpeace (1993)	2050	610	114 (19%)	Fossil fuels are phased out during the 21st century
	2100	986	181 (18%)	
Johansson *et al.* (1993)	2025	395	145 (37%)	RIGES model calculation
	2050	561	206 (37%)	
World Energy Council (1993)	2050	671	94 (14%)	Outcome of one scenario
	2100	895	132 (15%)	

[a] Based on five different studies. *Sources*. Hall (1998) and WEA (2000).
[b] Business-as-usual scenario.
[c] Energy conservation scenario.

benefits such as reduced indoor air pollution. For example, in household cooking, compared to solid fuels, gaseous or liquid cooking fuels can be used far more efficiently and conveniently, reaching many more families and emitting far fewer toxic pollutants.

Estimates of the technical potential of biomass energy are much larger than the present world energy consumption. If agriculture is modernized up to reasonable standards in various regions of the world, several billions of hectares may be available for biomass energy production well into this century. This land would comprise degraded and unproductive lands or excess cropland, preserving the world's nature areas and quality cropland. Table I gives a summary of the potential contribution of biomass to the world's energy supply according to a number of studies and influential organizations. Although the percentile contribution of biomass varies considerably, depending on the expected future energy demand, the absolute potential contribution of biomass in the long term is high, from about 100 to 300 EJ per year.

An Intergovernmental Panel on Climate Change (IPCC) study has explored five energy supply scenarios for satisfying the world's growing demand for energy services in the 21st century while limiting cumulative CO_2 emissions between 1990 and 2100 to fewer than 500 gigatonnes (Gt) of carbon. In all scenarios, a substantial contribution from carbon-neutral biomass energy as a fossil fuel substitute is included to help meet the CO_2 emissions targets. When biomass is grown at the same average rate as it is harvested for energy, it is approximately carbon neutral: carbon dioxide extracted from the atmosphere during growth is released back to the atmosphere during conversion to energy. Figure 2 shows the results for the IPCC's most biomass-intensive scenario, whereby biomass energy contributes

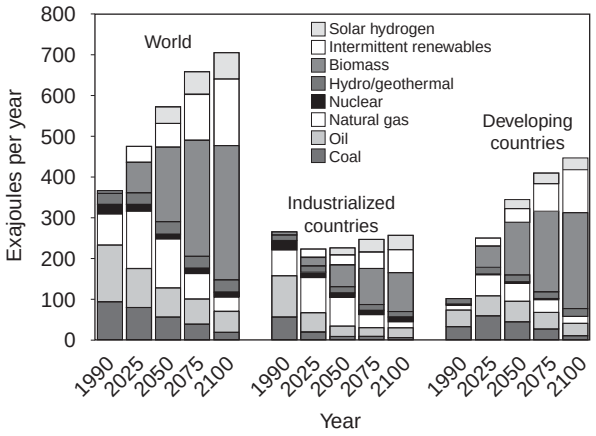


FIGURE 2 Primary commercial energy use by source for the biomass-intensive variant of the Intergovernmental Panel on Climate Change 1996 model, shown for the world, for industrialized countries, and for developing countries. From Sivan and Larson (2000).

180 EJ/year to global energy supply by 2050, satisfying about one-third of the total global energy demand and about half of total energy demand in developing countries. Roughly two-thirds of the global biomass supply in 2050 is assumed to be produced on high-yield energy plantations covering nearly 400 million hectares, or an area equivalent to one-quarter of present planted agricultural area. The other one-third comes from residues produced by agricultural and industrial activities. On the other hand, the large role biomass is expected to play in future energy supplies can be explained by several considerations. First, biomass fuels can substitute more or less directly for fossil fuels in the existing energy supply infrastructure. Intermittent renewables such as wind and solar energy are more challenging to the ways we distribute and consume energy. Second, the potential resource is large. Third, in developing countries, demand for energy is rising rapidly due to population increases, urbanization, and rising living standards. Although some fuel switching occurs in the process, the total demand for biomass will also tend to increase, as is currently seen for charcoal. Consequently, there is a growing consensus that energy policies will need to be concerned about the supply and use of biofuels while supporting ways to use these fuels more efficiently and sustainably.

2.3 Biomass Energy Conversion Technologies and Applications

Generating modern energy carriers (electricity, gas, and liquid fuels) from biomass, for use at the household ($\sim$10 kW), community ($\sim$100 kW), or industrial ($\sim$MW) scale, can be accomplished by a variety of technologies. The different technologies tend to be classed in terms of either the conversion process they use or the end product produced.

2.3.1 Combustion

The predominant technology in the world today for electricity generation from biomass, at scales above 1 MW, is the steam–Rankine cycle. This consists of direct combustion of biomass in a boiler to raise steam, which is then expanded through a turbine. The steam–Rankine technology is a mature technology introduced into commercial use about 100 years ago. The typical capacity of existing biomass power plants ranges from 1 to 50 MWe (megawatts, electricity), with an average around 20 MWe. Energy conversion efficiencies are relatively low (15–25%), due to their small size, although technologies and processes to increase these efficiencies are being developed. Steam cycle plants are often located at industrial sites, where the waste heat from the steam turbine is recovered and used for meeting industrial process heat needs. Such combined heat and power (CHP), or cogeneration, systems provide greater levels of energy services per unit of biomass consumed than do systems that generate only power, and can reach overall efficiencies of greater than 80%.

Biomass power generating capacity grew rapidly in the United States in the 1980s largely as the result of incentives provided by the Public Utilities Regulatory Policies Act of 1978 (PURPA), which required utilities to purchase electricity from cogenerators and other qualifying independent power producers at a price equal to the utilities' avoided costs. Currently in the United States, the installed biomass-electric generating capacity is about 7 GW (not including generating capacity of $\sim$2.5 GW from MSW and $\sim$0.5 GW from landfill gas). The majority of this capacity is located at pulp and paper mills, where biomass fuels are available as by-products of processing. There are also a substantial number of biomass power plants in California that use agricultural processing wastes as fuel. A significant number of biomass power plants are also found in Scandinavia, especially Sweden, where carbon taxes have encouraged recent expanded use of such systems for combined district heating and power production. By comparison to the steam–Rankine power-generating capacity installed in Organization for Economic Cooperation and Development (OECD) countries, there is relatively little capacity installed in developing countries. The most significant installation of steam–Rankine capacity in developing countries is at factories making sugar and/or ethanol from sugarcane, using bagasse, the fiber residue that remains after juice extraction from sugarcane.

The costs of steam–Rankine systems vary widely, depending on the type of turbine, the type of boiler, the pressure and temperature of the steam, and other factors. An important characteristic of steam turbines and boilers is that their capital costs (per unit of capacity) are scale sensitive. This, together with the fact that biomass steam–Rankine systems are constrained to relatively small scales (due to biomass fuel transport cost limitations), typically leads to biomass steam–Rankine systems that are designed to reduce capital costs at the expense of efficiency. For example, biomass-fired systems are typically designed with much more modest steam pressure and temperature than is technologically feasible, enabling

lower grade steels to be used in boiler tubes. This lowers both the costs and the efficiency. Because there are untapped supplies of low-cost biomass feedstocks available in many regions of the world, the economics of steam–Rankine systems are probably reasonable. For example, sugarcane-processing industries and sawmills present major opportunities for steam–Rankine-based combined heat and power generation from biomass.

An alternative to the direct-fired biomass combustion technologies, and considered the nearest term low-cost option, is biomass cocombustion with fossil fuels in existing boilers. Successful demonstrations using biomass as a supplementary energy source in large high-efficiency boilers have been carried out, showing that effective biomass fuel substitution can be made in the range of 10–15% of the total energy input with minimal plant modifications and no impact on the plant efficiency and operation. This strategy is economical when the biomass fuels cost less than the fossil fuels used. For fossil fuel plant capacities greater than 100 MWe, this can mean a substantial amount of biomass and related carbon savings and emissions reductions, particularly for coal substitution.

2.3.2 Gasification

Combustible gas can be produced from biomass through a high-temperature thermochemical process. The term "gasification" commonly refers to this high-temperature thermochemical conversion and involves burning biomass without sufficient air for full combustion, but with enough air to convert the solid biomass into a gaseous fuel. The product gas, called producer-gas, consists primarily of carbon monoxide, hydrogen, carbon dioxide, and nitrogen, and has a heating value of 4 to 6 MJ per normal cubic meter (Nm^3), or 10–15% of the heating value of natural gas. The intended use of the gas and the characteristics of the particular biomass (size, texture, moisture content, etc.) determine the design and operating characteristics of the gasifier and associated equipment. After appropriate treatment, the resulting gases can be burned directly for cooking or heat supply, or can be used in secondary conversion devices such as internal combustion engines or gas turbines for producing electricity or shaft work. The systems used can scale from small to medium (5–100 kW), suitable for the cooking or lighting needs of a single family or community, up to large grid-connected power generation facilities consuming several hundred of kilograms of woody biomass per hour and producing 10–100 MW of electricity.

One technology that has generated wide interest is the biomass integrated gasification combined cycle (IGCC) technology for larger scale power and combined heat and power generation in the range of 5 to 100 MWe. An IGCC system involves sizing and drying of the feedstock, followed by thermochemical gasification to produce a combustible gas, then cooling and cleaning of the gas and combustion in a gas turbine. Steam is generated using the hot exhaust of the gas turbine to drive a steam turbine, producing additional power and/or delivering lower pressure steam for heating purposes. The cascading of a gas turbine and a steam turbine in this manner is commonly called a combined cycle. In approximate terms, the IGCC technology will enable electricity to be made at double or more the efficiency of the steam cycle, and the capital costs per installed kilowatt for commercially mature IGCC units are expected to be lower than for comparably sized steam cycles. Thus, the overall economics of biomass-based power generation are expected to be considerably better with an IGCC system than with a steam–Rankine system, especially in situations in which biomass fuel is relatively expensive.

IGCC technology is expected to be commercially available within a few years, based on current demonstration and commercialization efforts worldwide. Several of the most advanced demonstration projects are in Sweden, the United Kingdom, Italy, and Brazil. In Varnamo, Sweden, the first complete biomass-fueled IGCC system has been operating for over 1500 hours on forest residues, generating 6 MW of electricity and 9 MW of heat for the local district heating system. In Yorkshire, England, construction of the Arable Biomass Renewable Energy (ARBRE) project, an IGCC facility that will generate about 10 MW of electricity from short-rotation biomass plantations, is nearly complete. It is supported under both the THERMIE program and the United Kingdom's Non-Fossil Fuel Obligation-3. The Bioelettrica Project near Pisa, Italy, an IGCC, is also under construction and will use poplar and robina wood chips, olive residues, grape residues, and sawdust to produce 12 MWe. At a site in Bahia, Brazil, construction began in 2000 on a Global Environment Facility (GEF)/World Bank-supported demonstration project of a 32 MW IGCC power plant using plantation-grown eucalyptus for fuel. The facility will also test the use of sugarcane bagasse. It has been estimated that if IGCC technology was applied worldwide, around 25% of all current electricity generation from sugarcane-producing countries could be produced just from their available sugarcane bagasse resource.

At the intermediate scale, producer-gas from biomass gasification can be used in modified internal combustion engines (typically diesel engines), in which it can replace 70–80% of the conventional fuel required by the engine. These smaller scale biomass gasifiers, coupled to diesel/gas internal combustion engines, operate in the 100- to 200-kWe range with efficiencies on the order of 15–25%, and have been made available commercially.

The greatest technical challenge for electricity-generating gasifier systems, at all scales, continues to be cleaning the tars and oils from the producer-gas such that the system operates efficiently, is economical, and has minimal toxic by-products and air emissions.

2.3.3 Anaerobic Digestion

Combustible gas can also be produced from biomass through low-temperature biological processes called anaerobic (without air) digestion. Biogas is the common name for the gas produced either in specifically designed anaerobic digesters or in landfills, by capturing the naturally produced methane. Biogas is typically about 60% methane and 40% carbon dioxide, with a heating value of about 55% that of natural gas. Almost any biomass except lignin (a major component of wood) can be converted to biogas; animal and human wastes, sewage sludge, crop residues, carbon-laden industrial processing by-products, and landfill material have all been widely used.

Anaerobic digesters generally consist of an inlet, where the organic residues and other wastes are fed into the digester tank; a tank, in which the biomass is typically heated to increase its decomposition rate and is partially converted by bacteria into biogas; and an outlet, where the biomass of the bacteria that carried out the process and the undigested material remain as sludge and can be removed. The biogas produced can be burned to provide energy for cooking and space heating or to generate electricity. Digestion has a low overall electrical efficiency (roughly 10–15%, strongly dependent on the feedstock) and is particularly suited for wet biomass materials. Direct nonenergy benefits are especially significant in this process. The effluent sludge from the digester is a concentrated nitrogen fertilizer and the pathogens in the waste are reduced or eliminated by the warm temperatures in the digester tank.

Anaerobic digestion of biomass has been demonstrated and applied commercially with success in a multitude of situations and countries. In India, biogas production from manure and wastes is applied widely in many villages and is used for cooking and power generation. Small-scale digesters have been used most extensively in India and China. Over 1.85 million cattle-dung digesters were installed in India by the mid-1990s, but about one-third of these are not operating for a variety of reasons, primarily insufficient dung supply and difficulties with the organization of dung deliveries. A mass popularization effort in China in the 1970s led to some 7 million household-scale digesters being installed, using pig manure and human waste as feed material. Many failed to work, however, due to insufficient or improper feed characteristics or poor construction and repair techniques. Estimates were that some 3 million to 4.5 million digesters were operating in the early 1980s. Since then, research, development, and dissemination activities have focused greater attention on proper construction, operation, and maintenance of digesters. One estimate is that there were some 5 million household digesters in working condition in China as of the mid-1990s. There are in addition some 500 large-scale digesters operating at large pig farms and other agroindustrial sites, and some 24,000 digesters at urban sewage treatment plants.

Several thousand biogas digesters are also operating in other developing countries, most notably South Korea, Brazil, Thailand, and Nepal. In addition, there are an estimated 5000 digesters installed in industrialized countries, primarily at large livestock processing facilities (stockyards) and municipal sewage treatment plants. An increasing number of digesters are located at food-processing plants and other industrial facilities. Most industrial and municipal digesters are used predominantly for the environmental benefits they provide, rather than for fuel production.

2.3.4 Liquid Biofuels

Biofuels are produced in processes that convert biomass into more useful intermediate forms of energy. There is particular interest in converting solid biomass into liquids, which have the potential to replace petroleum-based fuels used in the transportation sector. However, adapting liquid biofuels to our present-day fuel infrastructure and engine technology has proved to be nontrivial. Only oil-producing plants, such as soybeans, palm oil trees, and oilseeds such as rapeseed, can produce compounds similar to hydrocarbon petroleum products, and these have been used to replace small amounts of biodiesel. This biodiesel has been marketed in Europe and to a lesser extent in the United States,

but it requires substantial subsidies to compete with diesel. Another family of petroleum-like liquid fuels that is produced from gasified biomass is a class of synthesized hydrocarbons called Fischer–Tropsch (F–T) liquids. The process synthesizes hydrocarbon fuels [C_{10} to C_{12} hydrocarbons (kerosene) or C_3 and C_4 hydrocarbons (liquid petroleum gas)] from carbon monoxide and hydrogen gas over iron or cobalt catalysts. F–T liquids can be used as a sulfur-free diesel or can be blended with existing diesel to reduce emissions, an environmental advantage, but have yet to be produced efficiently and economically on a large scale; research and development efforts are ongoing. In addition to use as automotive fuels, F–T liquids can potentially be used as a more efficient, cleaner cooking fuel, compared to the traditional wood fuels from which it is synthesized.

Other biofuel alternatives to petroleum-based fuels are biomass-produced alcohols, which can replace gasoline or kerosene. The most widely produced bioalcohol today is ethanol, derived from the fermentation of biomass. In industrialized countries, ethanol is most commonly produced from food crops such as corn; in the developing world, it is produced from sugarcane. Its most prevalent use is as a gasoline fuel additive to boost octane levels or to reduce dependence on imported fossil fuels. In the United States and Europe, ethanol production is still far from competitive when compared to gasoline and diesel prices, and the overall energy balance of such systems has not been very favorable. The Brazilian Proalcohol ethanol program, initiated in 1975, has been successful due to the high productivity of sugarcane, although subsidies are still required. Two other potential transportation biofuels are methanol and hydrogen. They are both produced via biomass gasification and may be used in future fuel cells.

Although ethanol production from maize and sugarcane, both agricultural crops, has become widespread and occasionally successful, it can suffer from commodity price fluctuation relative to the fuels market. Consequently, the production of ethanol from lignocellulosic biomass (such as wood, straw, and grasses) is being given serious attention. In particular, it is thought that enzymatic hydrolysis of lignocellulosic biomass will open the way to low-cost and efficient production of ethanol. The development of various hydrolysis techniques has gained attention in recent years, particularly in Sweden and the United States, but cheap and efficient hydrolysis processes are still under development and some fundamental issues need to be resolved. Once such technical barriers are surmounted and ethanol production can be combined with efficient electricity production from unconverted wood fractions (such as the lignin), ethanol costs could come close to current gasoline prices and overall system efficiencies could go up to about 70% (low heating value). Though the technology to make this an economically viable option still does not exist, promising technologies are in the works and there are currently a number of pilot and demonstration projects.

2.4 Implementation of Biomass Energy Systems

Raw biomass has several disadvantages as an energy source. It is bulky, has a low energy density, and direct combustion is generally highly inefficient (other than advanced domestic heaters), producing high levels of indoor and outdoor air pollution. The goal of modernized biomass energy is to increase the fuel's energy density while decreasing its emissions during production and use. Modernizing biomass energy production, however, faces a variety of challenges that must be adequately addressed and dealt with before the widespread implementation of bioenergy systems can occur. These issues include technical problems (just discussed), resource availability, environmental impacts, and economic feasibility.

2.4.1 Biomass Resources

Biomass resources are potentially the largest renewable global energy source; there is an annual primary production of around 4500 EJ with a bioenergy potential on the order of 2900 EJ, of which about 270 EJ could be considered available on a sustainable basis. The challenge is not the availability so much as the sustainable management and conversion and delivery to the consumer in the form of modern and affordable energy services. Most of the biomass used today is either a residue in a bioprocessing industry or is an opportunity fuel that is used in households for daily living needs. It is argued that if biomass is to become a major fuel in the world, as is being proposed in future energy scenarios, then residues will not suffice and energy plantations may need to supply up to 80% of the future feedstock.

The solar energy conversion efficiency of plants is low, often less than 1%. Consequently, relatively large land surfaces are required to produce a substantial amount of energy. Moreover, biomass has a low energy density (only 8 GJ/tonne of wood, 50% moisture content). For comparison, other fuels have the following energy densities: coal, 28 GJ/tonne; mineral

oil, 42 GJ/tonne; and liquefied natural gas, 52 GJ/tonne. Consequently, transportation becomes an essential element of biomass energy systems, with transportation distances becoming a limiting factor, both from an economic and energetic point of view. Although generally it has been found that for woody biomass the energy output is 10–30 times greater than the energy input necessary for fuel production and transport, the issue is less clear for the production of liquid fuels, except for ethanol from sugarcane, which does have high net energy yields.

At present, the production of biomass residues and wastes globally, including by-products of food, fiber, and forest production, exceeds 110 EJ/year, perhaps 10% of which is used for energy. Residues concentrated at industrial sites are currently the largest commercially used biomass source. Residues are not, however, always accessible for energy use. In some cases, collection and transport costs are prohibitive; in other cases, agronomic considerations dictate that residues be recycled to the land. In still other cases, there are competing nonenergy uses for residues (e.g., fodder, construction material, or industrial feedstock).

Residues are an especially important potential biomass energy source in densely populated regions, where much of the land is used for food production. In fact, biomass residues play important roles in such regions precisely because the regions produce so much food: crop production can generate large quantities of by-product residues. For example, in 1996, China generated crop residues in the field (mostly corn stover, rice straw, and wheat straw) plus agricultural processing residues (mostly rice husks, corncobs, and bagasse) totaling about 790 million tonnes, with a corresponding energy content of about 11 EJ. To put this in perspective, if half of this resource were to be used for generating electricity at an efficiency of 25% (achievable at small scales today), the resulting electricity generation would be about half of the total electricity generated from coal in China in 1996. There is also a significant potential for providing biomass for energy by growing crops specifically for that purpose. The IPCC's biomass-intensive future energy supply scenario discussed previously includes 385 million hectares of biomass energy plantations globally in 2050 (equivalent to about one-quarter of present planted agricultural area), with three-quarters of this area established in developing countries. Such levels of land use for bioenergy raise the issue of intensified competition with other important land uses, especially food production.

There are two approaches to producing energy crops. These include devoting an area exclusively to production of such crops, and commingling the production of energy and nonenergy crops, either on the same piece of land (agroforestry) or on adjacent pieces of land (farm forestry). Because energy crops typically require several years to grow before the first harvest, the second approach has the benefit of providing the energy-crop farmer with revenue from the land between harvests of energy crops. In Sweden, productive heat power generation from willow plantations has been successful, and there has also been experience in small-scale fuel wood production in India, China, and elsewhere. In Brazil, farm forestry activities have involved small farmers in the high-yield production of biomass feedstocks.

2.4.2 Environmental Impacts and Benefits

In general, renewable forms of energy are considered "green" because they cause little depletion of Earth's resources, have beneficial environmental impacts, and cause negligible emissions during power generation. Yet, although biomass is in principle renewable and can have positive environmental impacts if managed properly, it also shares many characteristics with fossil fuels, both good and bad. Biomass can be transported and stored, allowing for heat and power generation on demand, but modernized bioenergy systems can also have negative environmental impacts associated both with the growing of the biomass and with its conversion to energy.

Many bioenergy conversion technologies offer flexibility in choice of feedstock and the manner in which it is produced. In contrast, most agricultural products are subject to rigorous consumer demands in terms of taste, nutritional content, uniformity, etc. The flexibility afforded by biomass technologies makes it easier to meet the simultaneous challenges of producing biomass energy feedstocks and meeting environmental objectives. For example, unlike the case with food crops, there are good possibilities for bioenergy crops to be used to revegetate barren land, to reclaim water-logged or salinated soils, and to stabilize erosion-prone land. Biomass energy feedstocks, when properly managed, can both provide habitat and improve biodiversity on previously degraded land.

Erosion and removal of soil nutrients are problems related to the cultivation of annual crops in many regions of the world. Although, relative to a healthy natural ecosystem, bioenergy systems may increase erosion and deplete soil nutrients and quality, bioenergy production on degraded or

erosion-prone lands can instead help stabilize soils, improve their fertility, and reduce erosion. Perennial energy crops (unlike food crops) improve land cover and form an extensive root system, adding to the organic matter content of the soil. Also, removal of soil during energy crop harvest can be kept to a minimum because roots are left in place, and twigs and leaves can be left to decompose in the field, enhancing the soil's nutrient store. This helps prevent diseases and improves soil fertility and quality. Environmental benefits of biomass crops, for carbon sequestration, biodiversity, landscape enhancement, and soil stabilization, can be particularly significant if plantations are established on intensively managed agricultural land. Although energy crops can be harvested by coppicing every 3 or 4 years, the stools (rootstocks) can survive for many decades, or even centuries, becoming significant carbon sinks. In addition, there are considerable benefits to both the landscape and to biodiversity when native species are used. For example, in Europe it would be preferable to grow willows and poplars rather than eucalyptus. Willows in particular support a high biomass and diversity of phytophagous insects, which in turn can support an important food web with many bird species. Also, when feasible, the recycling of ashes from the biomass combustion can return crucial trace elements and phosphates to the soil. This is already common practice in countries such as Sweden and Austria, where some of the ashes are returned to the forest floors, and in Brazil, where stillage, a nutrient-rich remainder of sugar cane fermentation, is returned to sugar cane plantations.

Another important potential impact from bioenergy feedstock production is the introduction of agricultural inputs such as fertilizers and pesticides into the environment. Fertilizers and pesticides can adversely affect the human health, water quality, and plant and animal life. Specific effects strongly depend on the type of chemical, the quantities used, and the method of application. Current experience with perennial crops (e.g., willow, poplar, or eucalyptus) for feedstock production suggests that necessary agricultural practices meet very strict environmental standards. Compared to food crops such as cereals, application rates of agrochemicals per hectare are a factor of 5 to 20 times lower for perennial energy crops. The abundant use of fertilizers and manure in agriculture has led to considerable environmental problems in various regions in the world: nitrification of groundwater, saturation of soils with phosphates (leading to eutrophication), and problems in meeting drinking water standards. Also, the application of phosphates has led to increased heavy metal flux to the soil. Compared to conventional agriculture, energy farming with short-rotation forestry and perennial grasses requires less fertilizer. With perennials, better recycling of nutrients is obtained. The leaching of nitrogen relating to willow cultivation can be about a factor of 2 to 10 times less than for food crops and is able to meet stringent standards for groundwater protection.

Possibly the biggest concern, and often considered the most limiting factor to the spread of bioenergy crops, is the demand on available water supplies, particularly in arid or semiarid regions. The choice of a certain energy crop can have a considerable effect on its water-use efficiency. Certain eucalyptus species, for example, have very good water-use efficiency when the amount of water needed per tonne of biomass produced is considered. However, a eucalyptus plantation on a large area could increase the local demand on groundwater and thus affect the groundwater level. On the other hand, energy crops on previously degraded land will improve land cover, which generally has positive effects on water retention and microclimate conditions. Impacts on the hydrological situation, therefore, always need to be evaluated on local level.

Finally, there is the issue of biodiversity and landscape. Biomass plantations are frequently criticized because the range of biological species they support is much narrower than is found in natural ecosystems. Although generally true, this is not always relevant. It would be relevant if a virgin forest were to be replaced by a biomass plantation—a situation that would be undesirable. However, when plantations are established on degraded lands or on excess agricultural lands, as is intended to be the case, the restored lands are very likely to support a more diverse ecology compared to the original situation. The restoration of such land is generally desirable for purposes of water retention, erosion prevention, and (micro) climate control. Furthermore, a good plantation design, including areas set aside for native flora and fauna, fitting into the landscape in a natural way, can avoid the problems normally associated with monocultures. The presence of natural predators (e.g., insects) can prevent the outbreak of pests and diseases. This issue needs more research, taking into account specific local conditions, species, and cultural aspects.

2.4.3 Economic and Production Issues

A number of key areas can be identified as being essential for the successful development and

implementation of sustainable, economically competitive bioenergy systems. The main consideration—a possible barrier—is whether the energy carriers produced are competitive. This is particularly true when specially produced biomass is used. In many situations, cheap or negative-cost biomass wastes and residues are available, thus the utilization of biomass is or could be competitive and future technology development should help further reduce the costs of bioenergy. In Sweden and Denmark, where a carbon and energy tax has been introduced, more expensive wood fuels and straw are now being used on a large scale. However, on a worldwide basis, the commercial production of energy crops is almost nonexistent. A major exception is Brazil, where subsidies have been introduced to make ethanol from sugarcane competitive with gasoline.

Closely related to the cost issue are the availability and the full-scale demonstration of advanced conversion technology that combines a high-efficiency and an environmentally sound performance with low investment costs. This is essential for competition with fossil fuels when relatively high-cost energy crops are used as energy sources. Advances in the combustion and cocombustion of biomass can considerably increase the attractiveness of combustion as a conversion technology. However, the development and the application of the IGCC technology have the potential to attain higher conversion efficiency at lower costs. Demonstration and commercialization of this technology are therefore important.

Experience with dedicated fuel supply systems based on "new" energy crops such as perennial grasses and short-rotation crops (SRCs) are very limited compared to the experience of traditional food crop cultivation and forestry techniques. Improvement of yields, increased pest resistance, management techniques, reduction of inputs, and further development of machinery are all necessary to lower costs and raise productivity. The same is true for harvesting, storage, and supply logistics. Bioenergy systems are complex in terms of organization and the number of actors that can be involved in a total energy system. The biomass is most likely to be produced by farmers or foresters whereas transport and storage are likely to be the responsibility of another party, and utilities may be responsible for the energy production. The combination of the utilities on the one hand and the agricultural system on the other will create a number of nontechnical barriers that have to be dealt with for any future system to work.

2.5 Conclusions

Biomass is one of the renewable energy sources capable of making a large contribution to the world's future energy supply. Land availability for biomass production should not be a bottleneck, provided it is combined with modernization of conventional agricultural production. Recent evaluations indicate that even if land surfaces of 400 million–700 million hectares were used for biomass production for energy about halfway the next century, this could be done without conflicting with other land-use functions and nature preservation. This can be obtained in part by better agricultural practices and in part by making use of huge areas of unproductive degraded lands. Latin America, Africa, Asia, and to a lesser extent Eastern Europe and North America represent a large potential for biomass production.

The types of biomass that can be used for energy are diverse, and optimal resources, technologies, and entire systems will be shaped by local conditions, both physical and socioeconomic. Perennial crops in particular may offer cheap and productive biomass production systems with low or positive environmental impacts. Technical improvements and optimized production systems along with multifunctional land use could bring biomass close to the costs of fossil fuels.

A key issue for bioenergy is that its use must be modernized to fit into a sustainable development. Conversion of biomass to energy carriers such as electricity and transportation fuels will give biomass a commercial value and will provide income for local rural economies. In order to obtain such a situation, it is essential that biomass markets and necessary infrastructure are developed, that key conversion technologies such as IGCC technology and advanced fuel production systems for methanol, hydrogen, and ethanol are demonstrated and commercialized, and that much more experience is gained with biomass production systems in a wide variety of contexts. Although the actual role of bioenergy will depend on its competitiveness versus fossil fuels and agricultural policies, it seems realistic to expect that the current contribution of bioenergy will increase during this century.

3. WIND ENERGY

3.1 Introduction

Wind has considerable potential as a global clean energy source, being both widely available, though

diffuse, and producing no pollution during power generation. Wind energy has been one of humanity's primary energy sources for transporting goods, milling grain, and pumping water for several millennia. From windmills used in China, India, and Persia over 2000 years ago to the generation of electricity in the early 20th century in Europe and North America, wind energy has played an important part in our recorded history. As industrialization took place in Europe and then in North America, wind power generation declined, first gradually as the use of petroleum and coal, both cheaper and more reliable energy sources, became widespread, and then more sharply as power transmission lines were extended into most rural areas of industrialized countries. The oil crises of the 1970s, however, triggered renewed interest in wind energy technology for grid-connected electricity production, water pumping, and power supply in remote areas, promoting the industry's rebirth.

This impetus prompted countries; notably Denmark and the United States, to establish government research and development programs to improve wind turbine technology. In conjunction with private industry research, this lead to a reemergence in the 1980s of wind energy in the United States and Europe, when the first modern grid-connected wind turbines were installed. In the 1990s, this development accelerated, with wind becoming the fastest growing energy technology in the world, developing into a commercially competitive global power generation industry. In 1990, only about 2000 MW of grid-connected wind power was in operation worldwide, whereas by 1999, this figure had surpassed 10,000 MW, not including the over 1 million water-pumping wind turbines located in remote areas. Since 1990, the average annual growth rate in world wind-generating capacity has been over 24%/year, with rates of over 30% in the past 2 years. Today, there is more than 31,000 MW of installed wind power, double the capacity that was in place just 3 years earlier (Fig. 3). More than 6000 MW of new wind energy capacity was installed in 2002. This dramatic growth rate in wind power has created one of the most rapidly expanding industries in the world, with sales of roughly $7 billion in 2002, and predictions of 10-fold growth over the next decade. Most 2000 forecasts for installed capacity were quickly eclipsed, and wind power passed the 10,000-MW mark in early 1999 and is now three times that level.

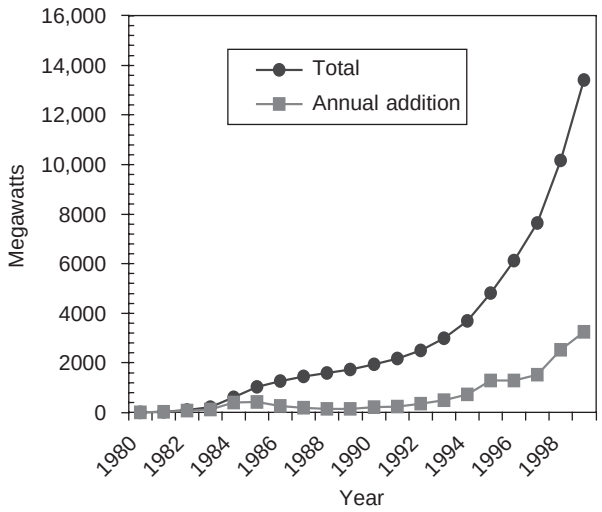


FIGURE 3 World wind-generating capacity, total and annual addition. From American Wind Energy Association (2003).

3.2 Economics of Wind Energy

Larger turbines, more efficient manufacturing, and careful siting of wind machines have brought wind power costs down precipitously, from $2600/kW in 1981 to $800/kW in 1998 and still lower today. New wind farms in some areas have now reached economic parity, or have beaten, new coal-based power plants, and as the technology continues to improve, further cost declines are projected, which could make wind power the most economical source of electricity in some countries. Market growth, particularly in Europe, has been stimulated by a combination of favorable governmental policies, lower costs, improved technology, and concern over environmental impacts of energy use. Compared to wind turbines built in 1981, modern turbines generate 56 times the energy at only nine times the cost.

Wind energy is currently one of the most cost-competitive renewable energy technologies. Worldwide, the cost of generating electricity from wind has fallen by more than 80%, from about 38¢/kWh (U.S.) in the early 1980s to a current range of 3–6¢/kWh (U.S.) levelized over a plant's lifetime, and analysts forecast that costs will drop an additional 20–30% in the next 5 years. Consequently, in the not-too-distant future, analysts believe that wind energy costs could fall lower than most conventional fossil fuel generators, reaching a cost of 2.5¢/kWh (U.S.) (Fig. 4). Wind technology does not have fuel requirements, as do coal, gas, and petroleum technologies. However, both the equipment costs and the costs of accommodating special characteristics such as intermittence, resource variability,

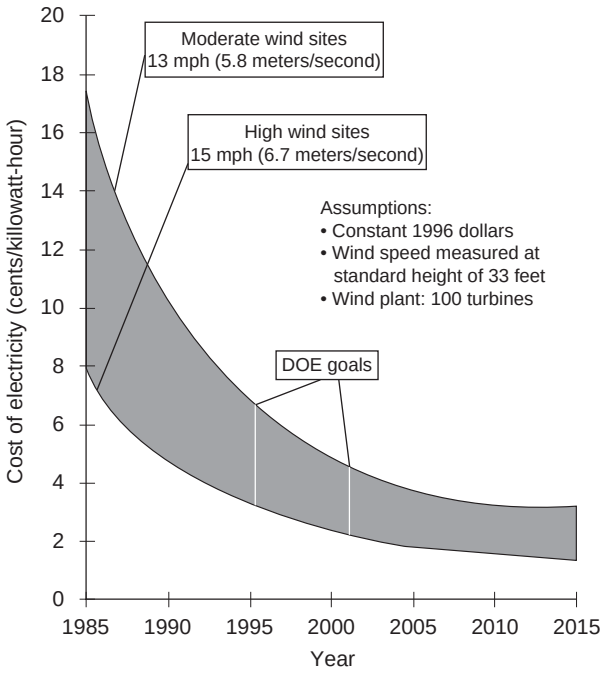


FIGURE 4 Trends in wind energy costs. From the U.S. Department of Energy (1998), (see the DOE Web site at http://www.eren.doe.gov).

competing demands for land use, and transmission and distribution availability can add substantially to the costs of generating electricity from wind.

3.3 Potential for Wind Energy: Technical, Resource, and Environmental Issues

The main technical parameter determining the economic success of a wind turbine system is its annual energy output, which in turn is determined by parameters such as average wind speed, statistical wind speed distribution, distribution of occurring wind directions, turbulence intensities, and roughness of the surrounding terrain. Of these, the most important and sensitive parameter is the wind speed (where the power in the wind is proportional to the third power of the momentary wind speed), which increases with height above the ground. As a result, vertical-axis wind turbines have mostly been abandoned in favor of the taller traditional horizontal-axis configuration. As accurate meteorological measurements and wind energy maps become more commonly available, wind project developers are able to more reliably assess the long-term economic performance of wind farms.

Some of the problems with wind power involve siting wind turbines. In densely populated countries where the best sites on land are occupied, there is increasing public resistance, making it impossible to realize projects at acceptable cost. This is one of the main reasons that countries such as Denmark and The Netherlands are concentrating on offshore projects, despite the fact that technically and economically they are expected to be less favorable, compared to good land sites. On the other hand, in countries such as the United Kingdom and Sweden, offshore projects are being planned not due to scarcity of suitable land sites, but because preserving the landscape is such an important national value. Another obstacle can be that the best wind site locations are not in close proximity to populations with the greatest energy needs, as in the U.S. midwest, making such sites impractical due to the high cost of transmission of power over long distances.

There has been a gradual growth of the unit size of commercial machines since the mid-1970s, when the typical size of a wind turbine was 30 kW installed power. By 1998, the largest units installed had a capacity of 1.65 MW, and turbines with an installed power of 2 MW are now being introduced on the market, with 3.6-MW machines now slated for construction. The trend toward larger machines is driven by the demand side of the market to utilize economy of scale, to reduce the visual impact on the landscape per unit of installed power, and to support the expectation that the offshore potential will be developed soon. Recent technical advances have also made wind turbines more controllable and grid compatible and have reduced their number of components, making them more reliable and robust.

Wind energy, although considered an environmentally sound energy option, does have several negative environmental aspects connected to its use. These include acoustic noise emission, visual impact on the landscape, impact on bird life, shadow caused by the rotor, and electromagnetic interference (influencing the reception of radio, television, and radar signals). In practice, the noise and visual impacts appear to cause the most problems for siting projects. Noise issues have been reduced by progress in aeroacoustic research, providing design tools and blade configurations that have successfully made blades considerably quieter. The impact on bird life appears to be a relatively minor problem. For instance, a research project in The Netherlands showed that bird casualties as a result of collisions with rotating rotor blades for a wind farm of 1000 MW represent only a very small fraction of birds killed from hunting, high-voltage lines, and traffic; the estimates predict a maximum level of six to seven bird

TABLE II

Estimates of Annual Total Human-Related Bird Mortality in The Netherlands[a]

Causes of death	Bird victims (× 100,000)
Road kills	20–80
Power lines	10–20
Hunting	6.5
Per 1000 MW of wind power	2.1–4.6

[a] *Source*. National Wind Coordinating Committee, Proceedings of National Avian–Wind Power Planning Meeting (1994).

collisions/turbine/year (see Table II). Avoiding endangered species habitats and major migration routes in the siting of wind farms can, for the most, part eliminate this problem.

In addition to being cost-competitive and environmentally sound, wind energy has several additional advantages over conventional fossil fuel power plants and even other renewable energy sources. First, it is modular: that is, the generating capacity of wind farms can easily be expanded because new turbines can be quickly manufactured and installed, not true for either coal-fired or nuclear power plants. Furthermore, a repair to one wind turbine does not affect the power production of all the others. Second, the energy generated by wind turbines can pay for the materials used to make them in as short as 3–4 months for good wind sites. Third, during normal operation, they produce no emissions. One estimate of wind energy potential to reduce CO_2 emissions predicts that a 10% contribution of wind energy to the world's electricity demand by 2025 would prevent the emission of 1.4 Gton/year of CO_2. Finally, there is also a strong and growing market for small wind turbines (under 100 kW), of which the United States is a leading manufacturer. Four very active U.S. manufacturers are estimated to cover a 30% market share worldwide. Small-scale turbines especially can play a significant role in rural and remote locations, particularly in developing countries, where access to the grid is either unlikely or extremely expensive.

3.4 Selected Country Profiles and Government Incentives to Promote Wind Energy

Incentives have long been viewed as a means of supporting technological developments until a new technology becomes cost-competitive. Wind-based electricity is not yet generally competitive with alternate sources of electricity such as fossil fuels. Thus, it is still dependent on nonmarket support for development to take place. Where sufficient support has been made available, wind capacity has expanded. When support has been withdrawn, or is uncertain, wind energy development has substantially slowed. The countries that have been fueling wind energy's growth throughout this decade have mainly been in the Northern Hemisphere, in particular Europe, where issues regarding the environment, fuel security, and electricity-generating diversity are a priority. Of the 10 countries with the highest installed capacity at the end of 2000, only the United States, India, and China lie outside Europe (Table III).

The stimulus for European wind energy growth continues to be regional and national policy incentives, which have facilitated the birth and development of the wind energy industry. Collectively, high energy prices, renewable energy subsidies that include investment and other fiscal incentives, and mandated purchases that include price support systems are some of the most important parameters that have encouraged wind energy development in the European Union (EU). The heightened sensitivity that Europeans have for the environment has been a determining factor in setting such policies. Another important factor, although it is less frequently mentioned, is fuel security. Developing an indigenous resource helps diminish the dependence on imported energy sources, which is of strategic importance to many countries in the EU. According to a 2003 American Wind Energy Association (AWEA) report, Europe continued to drive the engine of global wind power growth. A total of 5871 MW (worth $6.3 billion) was installed in the EU countries in 2002. Regional wind power capacity grew 33% to 23,056 MW. This volume of production offsets 20 million tons of coal burned annually in a conventional power plant, according to the European Wind Energy Association. Germany, Spain, and Denmark alone accounted for 89% of the wind power capacity installed in Europe in 2002. Non-EU European countries (such as Norway and Poland) contributed an additional 112 MW to the region's wind power growth.

3.4.1 United States

In the early 1980s, the United States accounted for 95% of the world's installed wind energy capacity;

TABLE III

Installed Wind Power in 1997 and 1998[a]

Region	Installed MW (1997)	Cumulative MW (1997)	Installed MW (1998)	Cumulative MW (1998)
United States	29	1611	577	2141
Canada	4	26	57	83
Mexico	0	2	0	2
South and Central America	10	42	24	66
Total America	43	1681	658	2292
Denmark	285	1116	310	1420
Finland	5	12	6	18
France	8	13	8	21
Germany	533	2081	793	2874
Greece	0	29	26	55
Ireland	42	53	11	64
Italy	33	103	94	197
Netherlands	44	329	50	379
Portugal	20	39	13	51
Spain	262	512	368	880
Sweden	19	122	54	176
United Kingdom	55	328	10	338
Other European countries, including Belgium, Czech Republic Slovakia, Norway, Austria, Turkey, Switzerland, and East European countries	13	57	23	80
Total Europe	1318	4793	1.766	6553
People's Republic of China	67	146	54	200
India	120	940	82	992
Remaining Asian countries, including Korea, Japan, Malaysia, Indonesia, Thailand, and Vietnam	9	22	11	33
Total Asia	196	1108	147	1224
Australia and New Zealand	2	8	26	34
Pacific Islands	0	3	0	3
North Africa, including Egypt, Ethiopia, Libya, Tunis, Algeria, Cape Verde, and Morocco	0	9	0	9
Middle East, including Jordan, Syria, Israel, Saudi Arabia, Iran, and Iraq (excluding Egypt)	8	18	0	18
Former Soviet Union countries	1	19	1	19
Total other continents and areas	11	57	27	83
Annual MW installed capacity	1568		2597	
Cumulative World MW installed		7639		10,153

[a] *Source*. WEA (2000). *Note*. The cumulative installed capacity by the end of 1998 is not always equal to the 1997 data plus installed capacity during 1998. Adjustments have been made by considering the amount of decommissioned/dismantled capacity.

the U.S. share dropped to 22% in 2002 and has essentially stagnated while other countries have increased their capacity. In 2001, 1600 MW of new capacity was installed in the United States. In 2002, that figure dropped to just over 4000 MW, although there are signs of steady growth resulting from the federal production tax credit [currently 1.7¢/kWh (U.S.) for utility-scale projects].

3.4.2 Germany

In contrast to the United States, Germany has increased its wind energy production consistently and dramatically throughout the past decade. In 2002, Germany installed over 3200 MW, or 55% of the total global new capacity. Germany overtook the United States in 1997 as the country with the highest installed wind capacity in the world, and Spain

followed suit in 2001 with a surge of new installations. A mixture of government incentives, including the Electric Feed-in Law, guaranteeing that renewable energy producers are paid up to 90% of the current domestic retail electricity price, and low interest loans in several regions, have helped make Germany the strongest wind market in the world. Germany passed another law that aims to double renewable energy's share of electricity to 10% in the next decade. This took effect in April 2000, replacing the Electric Feed-in law, and it provides a fixed level of support for all green power.

3.4.3 Denmark

Denmark ranks as the world's largest manufacturer and exporter of wind turbines, manufacturing over 60% of the world's turbines. They are 5 years ahead of their goal to produce 10% of their electricity from wind, and their new goal is to supply 50% of electric generation from wind by 2030, in part through a 4000-MW offshore wind initiative. The Danish government has been a consistent, strong supporter of its wind industry and, similar to the German legislation, their Windmill Law requires electric utilities to purchase output from private wind turbines at 85% of the consumer price of electricity, plus giving an ecotax break of 9¢/kWh (U.S.). Due to policies encouraging ownership of equipment, privately owned wind turbines represent 80% of the market. In addition, the electric utilities also receive a 1.5¢/kWh (U.S.) production subsidy for wind power generation. As of 2000, the wind energy market has been transformed from a fixed-price system to a renewable portfolio standard (RPS) such that by 2003, 20% of the country's electricity consumption will have to come from renewable sources; the current amount is between 12 and 14%. The plan is to achieve this through the trading of "green certificates" issued to renewable energy generators in proportion to their energy production. The United States is adopting RPS legislation on a state-by-state basis; the most aggressive standard (to date) is in Nevada, with a 20% RPS plan for 2015.

3.4.4 Developing Countries

Finally, among the developing countries, India continues to be the leader, followed by China. India had one of the fastest growing wind markets in the mid-1990s; growth has since slowed substantially due to underperformance of projects (often due to poor siting), transmission problems, and political and economic instability, all of which have affected investments in new projects. Extensive use of investment tax credits has also led to unsustainable revenue losses for the government. During the 1990s, electricity consumption more than doubled in India, and despite huge expenditures in the power industry, the country still has a 12% electricity deficit and 20% peak power shortage. Due to these conditions and the continued fast rate of growth in consumption, India continues to be considered a key market for wind developers. China remains mostly an untapped market with a huge future potential for wind energy; to date, relatively little money has been invested into projects, which remain small and donor based. It is expected that major commercial development, possibly to take place in the next decade, will be needed for strong wind energy development to occur in China.

3.5 Conclusions

For the first time, we are seeing one of the emerging renewable energy generating options—wind power—in a position to compete with the generation technologies of the past century. A variety of players are engaged in pushing forward wind projects worldwide. Enron Wind Corporation acquired German turbine manufacturer Tacke, NEG Micon of Denmark has built manufacturing facilities in the United States, Vestas of Denmark has built factories in Spain and India, and many manufacturers have developed joint ventures in various countries around the world. This globalization trend is likely to continue because financial institutions are beginning to view the wind industry as a promising investment opportunity. As more countries are added to the wind energy roster, uneven development focusing on a half-dozen key markets will most likely be replaced by more balanced growth. During the next couple of years, large-scale projects are expected to be developed in Egypt, Nicaragua, Costa Rica, Brazil, Turkey, Philippines, and several other countries, totaling thousands of megawatts of newly installed capacity, and expanding the number of countries using their wind resources.

4. SOLAR-PHOTOVOLTAIC AND SOLAR-THERMAL TECHNOLOGIES

There are two basic categories of technologies that convert sunlight into useful forms of energy, aside from biomass-based systems, which do this in a broader sense by using photosynthesis from plants as

an intermediate step. First, solar-photovoltaic (PV) modules convert sunlight directly into electricity. Second, solar-thermal power systems use focused solar radiation to produce steam, which is then used to turn a turbine, producing electricity. The following discussion provides a brief overview of these technologies, along with their current commercial status.

4.1 Solar Photovoltaics

At present, annual global PV module production is over 400 MW, which translates into a more than $2 billion/year business. In addition to the ongoing use of PV technologies in space, their present-day cost and performance also make them suitable for many grid-isolated and even grid-connected applications in both developed and developing parts of the world. PV technologies are potentially so useful that, as their comparatively high initial cost is brought down another order of magnitude, it is very easy to imagine them becoming nearly ubiquitous late in the 21st century. PV systems would then likely be employed on many scales in vastly differing environments, from microscopic cells to 100-MW or larger "central station" generating plants covering many square kilometers on Earth and in space. The technical and economic driving forces that favor the use of PV technologies in these widely diverse applications will be equally diverse. However, common among them will be the durability, high efficiency, quiet operation, and lack of moving parts that PV systems offer, and the fact that these attributes combine to provide a power source with minimum maintenance and unmatched reliability. PV system cost and performance have been steadily improving in recent years. PV manufacturing costs have fallen from about $30/W in 1976 to $5.00–8.00/W today, depending on the level of solar insolation at the site and other factors. These installed system costs are expected by some analysts to reach a range of $2.00–5.00/W by 2010, and if this is achieved, PV systems could achieve a sales level of over 1600 MW/year by that time (see Fig. 5).

PV efficiency has also increased markedly over the past few decades, as shown in Fig. 6. The best thin film cells tested in laboratories in 1980 achieved an efficiency level of about 10%. This was improved to about 13% by 1990 and to over 17% in recent years, for the best thin film cells made from copper indium diselenide ($CuInSe_2$) or cadmium telluride (CdTe). However, the modules that are currently being manufactured have somewhat lower efficiencies compared to the recent test cells, with typical values

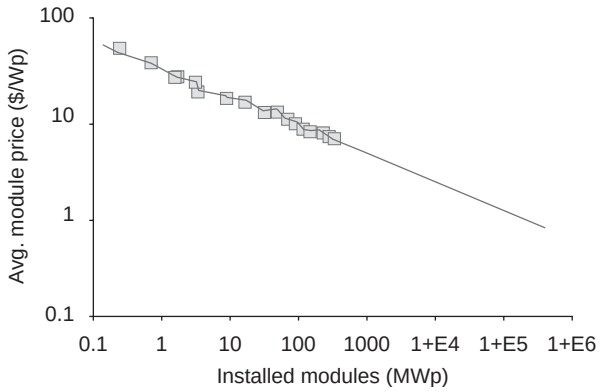


FIGURE 5 Learning curve for photovoltaics, 1976–2000. The average module price is in 1992 dollars (per watt, photovoltaic; Wp).

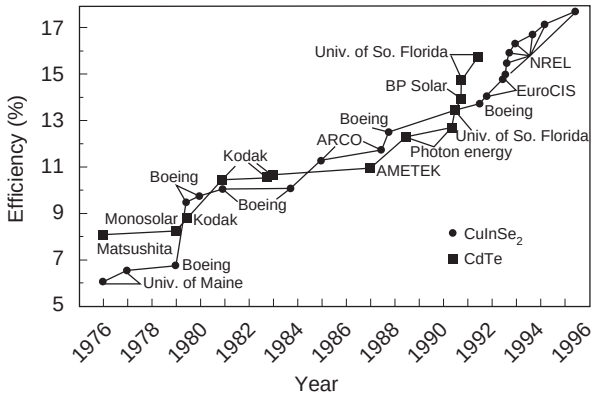


FIGURE 6 Progress in polycrystalline thin-film laboratory cell efficiencies. From the U.S. Department of Energy (1997).

in the range of 9–11% for $CuInSe_2$ and CdTe modules, and 8–10% for the more common amorphous silicon modules. The efficiency gains made over the past decade at the cell level will slowly translate into higher efficiency manufactured products, and this should combine with cell and module manufacturing cost reductions to aid in producing the reductions in per-watt installed system costs.

Figure 7 shows PV shipments by region and for the entire world, from 1976 to 1998. As can be seen, PV sales are growing rapidly; in fact, the sales growth rate throughout the 1990s was approximately 30%/year. Thus, PV systems are declining in cost, improving in efficiency, and increasing rapidly in sales. The cost of electricity produced from PV systems is still higher than it is from most other competing technologies, at about 30¢/kWh, but these costs are expected to continue to decline steadily. In fact, the U.S. Department of Energy projects that the cost of generating electricity from

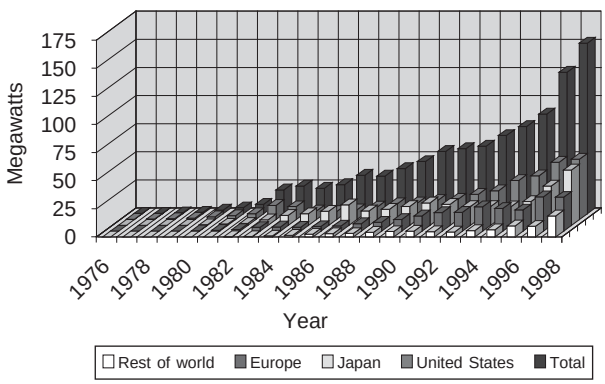


FIGURE 7 Regional and global photovoltaic shipments to different regions worldwide, 1976–2000. In 2002 (not shown), shipments reached 400 MW. From Maycock (1998).

utility-scale PV systems could be below 10¢/kWh (U.S.) by 2010, perhaps falling to about 6¢/kWh by 2020. Meanwhile, costs of electricity from residential PV systems could reach about 18¢/kWh by 2010 and 10¢/kWh by 2020. These potential reductions in cost, combined with the simplicity, versatility, reliability, and low environmental impact of PV systems, should help them to become increasingly important sources of economical premium-quality power over the next 20 to 30 years.

4.2 Solar-Thermal Systems

Solar-thermal power systems use various techniques to focus sunlight to heat an intermediary fluid, known as heat transfer fluid, which then is used to generate steam. The steam is used in a conventional steam turbine to generate electricity. At present, there are three solar-thermal power systems currently being developed: parabolic troughs, power towers, and dish/engine systems. Because these technologies involve a thermal intermediary, they can be readily hybridized with fossil fuels and in some cases adapted to utilize thermal storage. The primary advantage of hybridization and thermal storage is that the technologies can provide dispatchable power and operate during periods when solar energy is not available. Hybridization and thermal storage can enhance the economic value of the electricity produced and reduce its average cost.

Parabolic trough solar-thermal systems are commercially available. These systems use parabolic trough-shaped mirrors to focus sunlight on thermally efficient receiver tubes that contain a heat transfer fluid. This fluid is heated to about 390°C (734°F) and pumped through a series of heat exchangers to produce superheated steam that powers a conventional turbine generator, producing electricity. Nine of these parabolic trough systems, built in the 1980s, are currently generating 354 MW in Southern California. These systems, sized between 14 and 80 MW, are hybridized with up to 25% natural gas in order to provide dispatchable power when solar energy is not available.

Power tower solar-thermal systems are in the demonstration and scale-up phase. They use a circular array of heliostats (large, individually tracking mirrors) to focus sunlight onto a central receiver mounted on top of a tower. The first power tower, Solar One, was built in Southern California and operated in the mid-1980s. This initial plant used a water/steam system to generate 10 MW of power. In 1992, a consortium of U.S. utilities joined together to retrofit Solar One to demonstrate a molten-salt receiver and thermal storage system. The addition of this thermal storage capability makes power towers unique among solar technologies by allowing dispatchable power to be provided at load factors of up to 65%. In this system, molten salt is pumped from a "cold" tank at 288°C (550°F) and then cycled through the receiver, where it is heated to 565°C (1049°F) and finally returned to a "hot" tank. The hot salt can then be used to generate electricity when needed. Current designs allow storage ranging from 3 to 13 hours.

5. HYDROPOWER

5.1 Introduction

Hydropower is currently the largest renewable resource used for electricity. It plays an essential role in many regions of the world, with more than 150 countries generating hydroelectric power. A survey in 1997 by *The International Journal on Hydropower & Dams* found that hydro supplies at least 50% of national electricity production in 63 countries and at least 90% in 23 countries. About 10 countries obtain essentially all their commercial electricity from hydro, including Norway, several African nations, Bhutan, and Paraguay.

There is about 700 GW of hydro capacity in operation worldwide, generating 2600 TWh/year (about 19% of the world's electricity production). About half of this capacity and generation is in Europe and North America, with Europe the largest, at 32% of total hydro use, and North America at 23% of the total. However, this proportion is

declining as Asia and Latin America commission large amounts of new hydro capacity. Small, mini, and micro hydro plants (usually defined as plants generating less than 10 MW, 2 MW, and 100 kW, respectively) also play a key role in rural electrification in many countries. An estimated 300 million people in China, for example, depend on small hydro.

5.2 Capacity and Potential

There is vast unexploited potential worldwide for new hydro plants, particularly in the developing countries of Asia, Latin America, and Africa, but most of the best sites have already been developed in Europe and North America. There is also upgrading potential at existing schemes, though any future hydro projects will, in general, have to satisfy stricter requirements both environmentally and economically than they have in the past.

As shown in Table IV, the world's gross theoretical hydropower potential is about 40,000 TWh/year, of which about 14,000 TWh/year is technically feasible for development and about 7000 TWh/year is currently economically feasible. The last figure fluctuates most being influenced not only by hydro technology, but also by the changing competitiveness of other energy/electricity options, the status of various laws, costs of imported energy/electricity, etc. The biggest growth in hydro generation is expected in the developing countries, where there is still a large potential for hydro development, but relatively little growth is expected in most OECD countries, where more than 65% of the economic potential is already in use. Thus, at least in the near future, hydro will likely remain the main source of electricity generation in developing countries that possess adequate water resources.

Until recent years, there has been less than 100 GW (about 350 TWh/year) of new hydro capacity under construction at any one time, equivalent to less than 15% of the capacity in operation. The figure has now risen, reflecting China's vast construction program, which includes the 18.2-GW Three Gorges Project, now in its second phase of construction. Most new hydro capacity is under

TABLE IV

Potentials of Hydroelectric Installed Capacities under Construction in 1997[a]

Region	Gross theoretical potential (TWh/year)	Technically feasible potential (TWh/year)	Economically feasible potential (TWh/year)	Installed hydro capacity (GWe)	Hydro power production (TWh/year)	Hydro capacity under construction (MWe)
North America	5817	1510	912	141	697	882
Latin America and the Caribbean	7533	2868	1199	114	519	18,331
Sub-Saharan Africa	3294	1822	809	16	48	2464
Middle East and North Africa	195	216	128	9	27	7749
Western Europe	3258	1235	770	147	498	6707
Central and Eastern Europe	304	171	128	21	66	1211
Newly independent states of the former Soviet Union	3583	1992	1288	66	225	16,613
Centrally planned Asia and China	6511	2159	1302	64	226	51,672
South Asia	3635	948	103	28	105	13,003
Other Pacific areas	5520	814	142	14	41	4688
Pacific OECD countries[c]	1134	211	184	34	129	841
World total[b]	40,784	13,945	6964	655	2582	124,161

[a] *Source*. WEA (2000).
[b] Totals rounded.
[c] OECD, Organization for Economic Cooperation and Development.

construction in Asia and South America. China has by far the most, with about 50 GW underway, and India, Brazil, and Russia each have more than 9 GW under construction. Brazil has the largest resources in world (800,000 GWh/year) of economically exploitable capacity and Norway depends almost entirely hydro for its electricity needs.

Hydropower continues to be the most energy-efficient way to generate electricity. Modern hydro turbines can convert as much as 90% of the available energy into electricity. The best fossil fuel plants are only about 50% efficient. In the United States, hydropower is produced for an average of 0.7¢/kWh. This is about one-third the cost of using fossil fuel or nuclear energy and one-sixth the cost of using natural gas. Hydro resources are also widely distributed compared with fossil and nuclear fuels and can help provide energy independence for countries without fossil fuel resources. There is also significant, widespread activity in developing small, mini, and micro hydro plants. At least 40 countries, particularly in Asia and Europe, have plants under construction and even more have plants planned. China, Brazil, Canada, Turkey, Italy, Japan, and Spain all have plans for more than 100 MW of new capacity.

5.3 Small Hydro

Small-scale hydro is mainly "run of river," so does not involve the construction of large dams and reservoirs. It also has the capacity to make a more immediate impact on the replacement of fossil fuels because, unlike other sources of renewable energy, it can generally produce some electricity on demand (at least at times of the year when an adequate flow of water is available), with no need for storage or backup systems. It is also in many cases cost-competitive with fossil fuel power stations or (for remote rural areas) diesel-generated power.

Small hydro has a large, and as yet untapped, potential in many parts of the world. It depends largely on already proved and developed technology with scope for further development and optimization. Least-cost hydro is generally "high-head" hydro because the higher the head, the less the flow of water required for a given power level, and so smaller and less costly equipment is needed. Although this makes mountainous regions very attractive sites, they also tend to be in areas of low population density and thus low electricity demand, and long transmission distances often nullify the low-cost advantage. Low-head hydro, on the other hand, is relatively common, and also tends to be found in or near concentrations of population, where there is a demand for electricity. Unfortunately, the economics also tend to be less attractive unless there are policy incentives in place to encourage their development.

5.4 Environmental and Social Impacts

Although hydroelectricity is generally considered a clean energy source, it is not totally devoid of greenhouse gas emissions and it can often have significant adverse socioeconomic impacts. There are arguments now that large-scale dams actually do not reduce overall GHG emissions when compared to fossil fuel power plants. To build a dam, significant amounts of land need to be flooded, often in densely inhabited rural area, involving large displacements of usually poor, indigenous peoples. Mitigating such social impacts represents a significant cost to the project, which, if it is even taken into consideration, often not done in the past, can make the project economically and socially unviable.

Environmental concerns are also quite significant, as past experience has shown. This includes reductions in biodiversity and fish populations; sedimentation, which can greatly reduce dam efficiency and destroy the river habitat; negative impacts on water quality; and contributions to the spread of water-related diseases. In fact, in the United States, several large power production dams are being decommissioned due to their negative environmental impacts. Properly addressing these issues would result in an enormous escalation of the overall costs for producing hydropower, making it far less competitive than is usually stated. As many countries move toward an open electricity market, this fact will come into play when decisions regarding investments in new energy sources are being made. If the large hydro industry is to survive, it needs to come to grips with its poor record of both cost estimation and project implementation.

5.5 Conclusions

Hydropower is a significant source of electricity worldwide and will likely continue to grow, especially in the developing countries. Although large dams have become much riskier investments, there still remains much unexploited potential for small hydro projects around the world. It is expected that growth of hydroelectricity will continue but at a slower rate than that of the 1970s and 1980s. Thus, the fraction of hydroelectricity in the portfolio of

primary sources of energy, which is today at 19%, is expected to decrease in the future. Improvements and efficiency measures are needed in dam structures, turbines, generators, substations, transmission lines, and environmental mitigation technology if the role of hydropower as a clean renewable energy source is to continue to be supported.

6. GEOTHERMAL ENERGY

6.1 Introduction

Geothermal energy, the natural heat within the earth, arises from the ancient heat remaining in Earth's core, from friction where continental plates slide beneath each other and from the decay of radioactive elements that occur naturally in small amounts in all rocks. The amount of geothermal energy is enormous. Scientists estimate that just 1% of the heat contained in the uppermost 10 km of Earth's crust is equivalent to 500 times the energy contained in all of Earth's oil and gas resources. Yet, despite the fact that this heat is present in practically inexhaustible quantities, it is unevenly distributed, seldom concentrated, and often at depths too great to be exploited industrially and economically.

Geothermal energy has been produced commercially for 70 years for both electricity generation and direct use; its use has increased rapidly during the past three decades. From 1975 to 1995, the growth rates worldwide for electricity generation and direct use of geothermal energy were about 9%/year and about 6%/year, respectively. In 1997, geothermal resources had been identified in over 80 countries and there were quantified records of geothermal utilization in at least 46 countries.

6.2 Capacity and Potential

The worldwide use of geothermal energy amounts to about 44,000 GWh/year of electricity and 38,000 GWh/year for direct use, as shown in Table V. Although geothermal accounts for only about 0.3% of the total electrical power generated worldwide, among renewable sources (excluding hydro) it ranks first, at around 80% of the total. It is currently estimated that the geothermal potential of the world for electricity production is 12,000 TWh/year, and for direct use it is estimated to be even larger, at 600,000 EJ. A very small fraction of the geothermal potential has therefore been developed so far, and an accelerated use of geothermal energy in the near future is certainly feasible. In addition, the technology for both electricity generation and direct application is mature and accessible. Thus, the question of future geothermal development depends on whether it is economically competitive with other energy sources in the different markets around the world. Currently, the electric generation cost is variable, but averages around 4¢/kWh (U.S.); direct utilization production costs, though highly variable, are commonly under 2¢/kWh.

TABLE V

Geothermal Electricity Generation and Direct Use in 1997[a]

	Electricity generation			Direct use		
	Installed capacity	Total production		Installed capacity	Total production	
Country	(MWe)	GWh/year	%	(MWt)	GWh/year	%
European Union	754	3,832		1031	3719	
Europe, other	112	471		4089	16,058	
Europe, total	866	4303	10	5120	19,777	52
North America	2849	16,249		1908	3984	
Central and South America	959	6869				
America, total	3808	23,118	53	1908	3984	10
Asia	2937	13,045	30	3075	12,225	32
Oceania	365	2901	6	264	1837	5
Africa	45	390	1	71	355	1
World total	8021	43,756		10,438	38,178	

[a] *Source*. WEA (2000).

TABLE VI

Installed Geothermal Electricity Generation Capacity[a]

Country	1990 MWe	1995 MWe	2000 MWe
Argentina	0.67	0.67	0
Australia	0	0.17	0.17
China	19.2	28.78	29.17
Costa Rica	0	55	142.5
El Salvador	95	105	161
Ethiopia	0	0	8.52
France (Guadeloupe)	4.2	4.2	4.2
Guatemala	0	33.4	33.4
Iceland	44.6	50	170
Indonesia	144.75	309.75	589.5
Italy	545	631.7	785
Japan	214.6	413.71	546.9
Kenya	45	45	45
Mexico	700	753	755
New Zealand	283.2	286	437
Nicaragua	35	70	70
Philippines	891	1227	1909
Portugal (The Azores)	3	5	16
Russia (Kamchatka)	11	11	23
Thailand	0.3	0.3	0.3
Turkey	20.6	20.4	20.4
United States	2774.6	2816.7	2228
Total	5831.72	6833.38	7974.06

[a] Data From the International Geothermal Association.

Table VI shows the installed geothermal electricity generation capacity in 1990, 1995, and 1998 for a selection of countries. The growth of the total generation capacity from 1990 to 1995 was about 16% and from 1995 to 1998, about 17%. The largest additions in generation capacity from 1990 to 1998 (in megawatts of electricity, MWe) were in the Philippines (957), Indonesia (445), Italy (224), Japan (315), Costa Rica (120), Iceland (95), United States (75), New Zealand (62), and Mexico (43). The most progressive of these countries, the Philippines, plans to add some 580 MWe to their installed capacity through 2008.

In the industrialized countries where installed electrical capacity is already very high, geothermal energy is unlikely in the near-term future to account for more than 1% of the total. On the other hand, in developing countries, with still relatively limited electrical consumption, but with good geothermal prospects, electrical generation from geothermal energy could make a significant contribution to the total installed electrical capacity. For example, 18% of the electricity in the Philippines already comes from geothermal sources; geothermal production is currently 17% in Nicaragua, 12% in El Salvador, 11% in Costa Rica, and 6% in Kenya.

TABLE VII

Top 10 Countries in Direct Utilization of Geothermal Energy[a]

Country	Installed (MWt)	Production (GWh/year)
Japan	1159	7500
Iceland	1443	5878
China	1914	4717
United States	1905	3971
Hungary	750	3286
Turkey	635	2500
New Zealand	264	1837
France	309	1359
Italy	314	1026
Russia	210	673

[a] *Source*. WEA (2000).

Direct application of geothermal energy can involve a wide variety of end uses employing mostly existing technology. The technology, reliability, economics, and environmental acceptability of direct use of geothermal energy have been demonstrated throughout the world. In comparison with electricity production from geothermal energy, direct utilization has several advantages. It has higher energy efficiency (50–70% as opposed to 5–20% for conventional geothermal electric plants), generally a much shorter development time, and normally less capital investment is needed. In addition, direct application can use both high- and low-temperature geothermal resources and is therefore much more widely available in the world. Direct application is, however, much more site specific for the market, because steam and hot water are rarely transported long distances from the geothermal site. The longest geothermal hot water pipeline in the world, 63 km, is in Iceland. Space heating is the dominant (33%) type of direct geothermal use in the world; the various other types of direct use of geothermal energy include bathing/swimming/balneology (19%), greenhouses (14%), heat pumps (12%) for air cooling and heating, fish farming (11%), and industry (10%).

Table VII shows installed geothermal capacity and production in the top 10 direct-use countries in the world. It is worth noting that the two countries with the highest energy production, Japan and Iceland, are not the same as the two with the highest installed

capacities, China and the United States. The reason for this is the variety in the load factors for the different types of use.

6.3 Environmental Impacts

Geothermal fluids contain variable concentrations of gases, largely nitrogen and carbon dioxide, with some hydrogen sulfide and smaller proportions of ammonia, mercury, radon, and boron. Most of these chemicals are concentrated in the disposal water that is usually reinjected back into the drill holes, so that there is minimal release into the environment. The concentrations of the gases are usually low enough not to be harmful and the abatement of toxic gases can be managed with current technology. Carbon dioxide is the major component of the noncondensible gases in the steam, but its emission into the atmosphere per kilowatt-hour is well below the figures for power plants fired by natural gas, oil, or coal. Hydrogen sulfide is the pollutant of greatest concern in geothermal plants, yet even the sulfur emitted, with no controls, is only half of what is emitted from a coal-fired plant. Overall, with present technology able to control the environmental impact of geothermal energy development, it is considered to be a relatively benign source of energy.

6.4 Conclusions

Geothermal energy is generally cost-competitive with conventional energy sources and is produced by well-proved conventional technology. It is reliable and has been used for the more than half of the past century to heat large municipal districts, as well as to feed power plants generating hundreds of megawatts of electricity. It has strong potential to continue to expand, especially in the developing countries, and is a clean energy source that can help contribute to reducing our greenhouse gas emissions. It is felt that in the near term, the development of geothermal energy could help fulfill a bridging function during the next few decades, as other more modern clean fuel technologies and renewables mature enough to provide a meaningful share of the world energy supply.

7. RENEWABLE ENERGY SYSTEM COST AND PERFORMANCE

7.1 Recent Progress

There has been significant progress in cost reduction made by wind and PV systems, and biomass, geothermal, and solar-thermal technologies are also experiencing cost reductions that are forecast to continue. Figure 8 presents forecasts made by the U.S. Department of Energy (DOE) for the capital costs of these technologies, from 1997 to 2030. Of course, capital costs are only one component of the total cost of generating electricity, which also includes fuel costs and operation and maintenance costs. In general, renewable energy systems are characterized by low or no fuel costs, although operation and maintenance costs can be considerable. It is important to note, however, that operation and maintenance costs for all new technologies are generally high, and can fall rapidly with increasing familiarity and operational experience. Renewable energy systems such as photovoltaics contain far fewer mechanically active parts than do comparable fossil fuel combustion systems, and therefore are likely in the long term to be less costly to maintain. Figure 9 presents U.S. DOE projections for the levelized costs of electricity production from these same renewable energy technologies, from 1997 to 2030. Given these likely capital and levelized system cost reductions, recent analyses have shown that additional generating capacity from wind and solar energy can be added at low incremental costs relative to additions of fossil fuel-based generation. These incremental costs would be further offset by environmental and human health benefits. Furthermore, a U.S. National Renewable Energy Laboratory (NREL) analysis shows that geothermal and wind energy could actually become more economic than coal in the next 15 years.

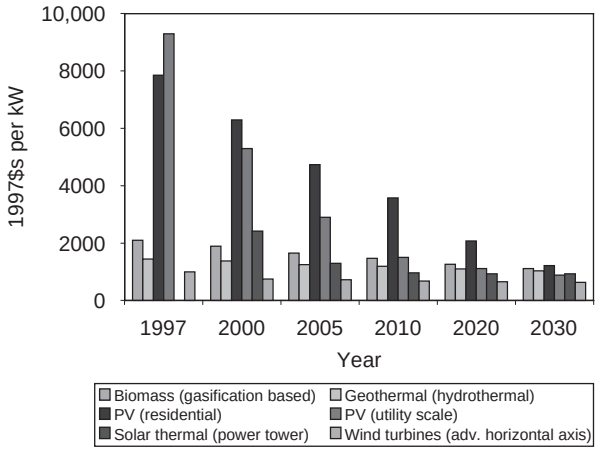


FIGURE 8 Capital cost forecasts for renewable energy technologies. PV, Photovoltaic. From the U.S. Department of Energy (1997).

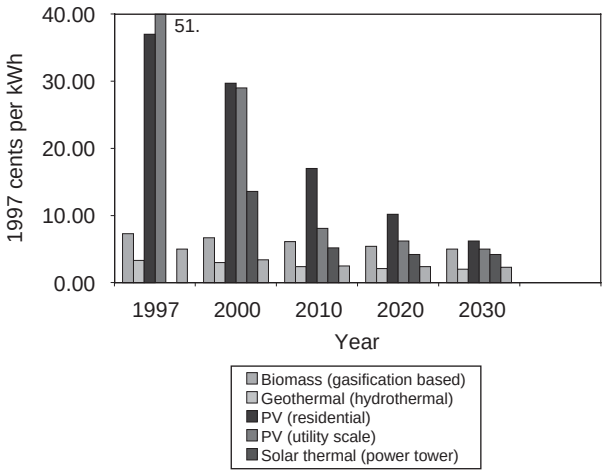


FIGURE 9 Levelized cost of electricity forecast for renewable energy technologies. PV, Photovoltaic. From the U.S. Department of Energy (1997).

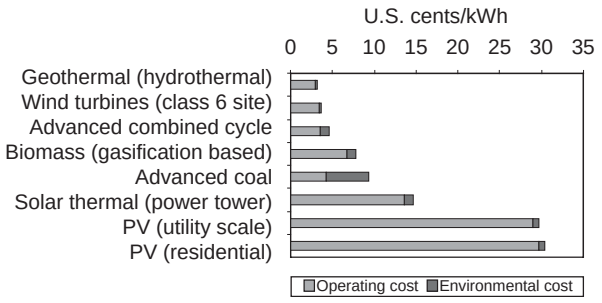


FIGURE 10 Actual electricity costs in 2000. PV, Photovoltaic. Data from the U.S. Department of Energy (1997, 2000).

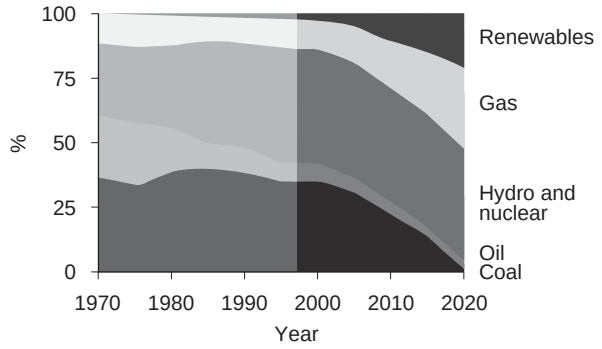


FIGURE 11 Electricity mix in Organization for Economic Cooperation and Development countries. From Shell Petroleum (2000).

Another analysis conducted by the Renewable Energy Policy Project (REPP) shows that adding 3050 MW of wind energy production in Texas, over a 10-year period, would entail only modest additional costs to residential customers. REPP estimates these additional costs to be about 75¢/month for a household using 1000 kWh/per month, or about $9 annually. The economic case for renewables looks even better when environmental costs are considered along with capital and operating costs. As shown in Fig. 10, geothermal and wind can be competitive with modern combined-cycle power plants, and geothermal, wind, and biomass all have lower total costs compared to advanced coal-fired plants, once approximate environmental costs are also included. Shell Petroleum has made one of the highest profile projections of future renewables growth. As shown in Fig. 11, Shell projects that renewables could constitute about 15% of the OECD's energy production by 2020, and that renewables and natural gas combined could account for about 50% of total production.

As already noted, the remarkable difference between the setting for renewable energy today, relative to the past 30 years, is that renewable and other clean energy technologies are now becoming economically competitive, and the push to develop them is no longer being driven solely by environmental concerns. With regard to prospects for investing in companies developing clean energy resources, Merrill Lynch's Robin Batchelor recently stated that "This is not an ethical investment opportunity, it's a straightforward business opportunity." Batchelor also noted that the traditional energy sector has lacked appeal to investors in recent years because of heavy regulation, low growth, and a tendency to be highly cyclical. He has identified 300 companies worldwide whose aim is to develop wind, solar, and wave power technologies and to advance capabilities in energy storage, conservation, and on-site power generation. Further evidence of the impending transition to renewable energy systems can be found in recent corporate reorganization among the world's largest oil companies. Corporate giants such as Shell and BP/Amoco, which would now like to be known as "energy companies" rather than "oil companies," have recently reorganized into a broader array of business units, including those exclusively focused on renewables. Such subsidiary units now include "Shell Renewables," and "BP Solar."

7.2 Lessons Learned in Developing Countries

In developing nations, renewable energy technologies are increasingly used to address energy shortages and to expand the range of services in both rural and urban areas. In Kenya, over 120,000 small

(20–100 Wp) (Wp, photovoltaic-derived watts) solar-PV systems have been commercially financed and installed in homes, battery charging stations, and other small enterprises. Meanwhile, a government program in Mexico has disseminated over 40,000 such systems. In the Inner Mongolian autonomous region of China, over 130,000 portable windmills provide electricity to about one-third of the non-grid-connected households in this region. These examples demonstrate that the combination of sound national and international policies and genuinely competitive markets—the so-called level playing field—can be used to generate sustainable markets for clean energy systems. They also demonstrate that renewable energy systems can penetrate markets in the developing world, even where resources are scarce, and that growth in the renewables sector need not be limited to applications in the developed world. Just as some developing countries are bypassing construction of telephone wires by leaping directly to cellular-based systems, so too might they avoid building large, centralized power plants and instead develop decentralized systems. In addition, to help mitigating the environmental costs of electrification, this strategy can also reduce the need for the construction of large power grids.

In the absence of policy interventions, we may remain locked-in to existing technologies, even if the benefits of technology switching overwhelm the costs. There are numerous examples, however, of an entrenched or locked-in technology being first challenged and ultimately replaced by a competing technology. This process is generally enabled by a new wave of technology, and it is sometimes achieved through a process of hybridization of the old and the new. Technological "leapfrogging" is another possibility, but this seems to occur relatively rarely. A prime example of the hybridization concept is in the case of the competition between gas- and steam-powered generators, dating back to the beginning of the century. From about 1910 to 1980, the success of steam turbines led to a case of technological lock-in, and to the virtual abandonment of gas turbine research and development. However, partly with the aid of "spillover" effects from the use of gas turbines in aviation, the gas turbine was able to escape the lock-in to steam turbine technology. First, gas turbines were used as auxiliary devices to improve steam turbine performance, and then they slowly became the main component of a hybridized, "combined-cycle" system. In recent years, orders for thermal power stations based primarily on gas turbines have increased to more than 50% of the world market, up from just 15–18% in 1985.

7.3 Leveling the Playing Field

As was shown in Fig. 10, renewable energy technologies tend to be characterized by low environmental costs. In an ideal world, this would aid them in competing with conventional technologies, but of course many of these environmental costs are "externalities" that are not priced in the market. Only in certain areas and for certain pollutants do these environmental costs enter the picture, and clearly further internalizing these costs would benefit the spread of renewables. The international effort to limit the growth of greenhouse emissions through the Kyoto Protocol may lead to some form of carbon-based tax, and this could prove to be an enormous boon to renewable energy industries. However, support for the Kyoto Protocol among industrialized countries remains relatively weak, particularly in the United States, and any proposed carbon-based taxation scheme will surely face stiff opposition.

Perhaps more likely, concern about particulate matter emission and formation from fossil fuel power plants will lead to expensive mitigation efforts, and this would help to tip the balance toward cleaner renewable systems. In a relatively controversial move, the U.S. Environmental Protection Agency (EPA) has recently proposed new ozone and particulate matter (PM) standards that are even more stringent than the current standards that remain unattained in some U.S. urban areas. The EPA has justified these new regulations with analysis that shows that new standards are necessary to provide increased protection against a wide range of potential health impacts. For example, the EPA estimates that even if Los Angeles County (California) were to meet the existing PM standards, 400 to 1000 deaths per year would still occur as a result of exposure to very fine PM (under 2.5 μm in diameter), which presently is not regulated. The combination of increased pressure to attain ozone and PM standards will further complicate siting of new fossil-fueled power plants in some areas of the United States, particularly where it is now required to find "offsets" for pollution impacts. This will indirectly but surely benefit renewable energy technologies, which do not typically face these difficulties in obtaining siting permits.

7.3.1 Public and Private Sector Investment Issues

A fundamental problem with any new technology is that, by definition, it does not have the track record

of performance that exists for older, more established systems. Proponents of existing technologies in mistaken arguments against technological change often cite this fact. New technologies and operational procedures do present greater risks, but at the same time present greater opportunities for innovation and profit. Emerging energy systems were long seen as an area of risky investments, with the history of renewable energy systems seen as the primary illustration of that "fact." It has been argued that this pattern is not only illusory, but is largely a self-fulfilling prophecy. Larger gains, both to individual companies and to society, typically stem from carefully targeted but consistently pursued avenues of research, innovation, and implementation. Renewable energy systems offer this same combination of increased uncertainty, great promise, and the potential for significant innovations and profits.

There are two principal rationales for government support of research and development to develop clean energy technologies (CETs). First, conventional energy prices generally do not reflect the social cost of pollution. This provides a rationale based on a well-accepted economic argument, for subsidizing research and development for CETs as potential alternatives to polluting fossil fuels. Second, private firms are generally unable to appropriate all the benefits of their research and development investments. Consequently, the social rate of return for research and development exceeds available private returns, and firms therefore do not invest enough in research and development to maximize social welfare. Thus, innovation "spillover" among CET firms is a form of positive externality that justifies public research and development investment. The conventional wisdom is that government should restrict its support to research and development and let the private sector commercialize new technologies. Failed CET commercialization subsidies (e.g., the U.S. corn ethanol and synfuels programs) bolster this view. Nonetheless, there are compelling arguments for public funding of market transformation programs (MTPs) that subsidize demand for some CETs in order to help commercialize them. Further, the argument that it may not be worthwhile for firms to invest in new technologies because of the spillover effects is generally false as well. Early investment in new technologies in promising market sectors has proved to be the best strategy for firms interested in long-term (and not simply short-term) profitability.

A principal motivation for considering MTPs is inherent in the production process. When a new technology is first introduced, it is invariably more expensive than established substitutes. There is, however, a clear tendency for the unit cost of manufactured goods to fall as a function of cumulative production experience. Cost reductions are typically very rapid at first, but taper off as the industry matures. This relationship is called an "experience curve" when it accounts for all production costs, and it can be described by a progress ratio (PR), where unit costs fall by $100 \times (1-\mathrm{PR})$ percent with every doubling of cumulative production. Typical PR values range from 0.7 to 0.9 and are widely applicable to technologies such as toasters, microwave ovens, solar panels, windmills, and essentially any good that can be manufactured in quantity. For example, Fig. 12 presents PRs for photovoltaics, windmills, and gas turbines. All three have initial PRs of approximately 0.8, which is a typical value observed for many products. After 1963, the gas turbine PR increased substantially, however, indicating an attenuation of experience effects.

The benefits of production experience may accrue primarily to the producing firm, or they may spill over to its competitors. Among other channels, "experience spillovers" could result from hiring competitors' employees, reverse engineering rivals' products, informal contacts among employees of rival firms, or even industrial espionage. If firms retain the benefits of their own production experience, they have an incentive to consider experience effects when deciding how much to produce. Consequently, they will "forward price," producing at a loss initially to bring down their costs and thereby maximize profit over the entire production period. It has been shown that, in each period, firms maximize profit by setting marginal revenue equal to their true marginal cost (TMC). For a firm with a discount rate of zero, TMC is defined as the cost of the last unit the firm plans to produce. If the firm has a positive discount rate, its TMC will be somewhat higher. However, due to experience effects, the TMC is always lower than current marginal cost (except

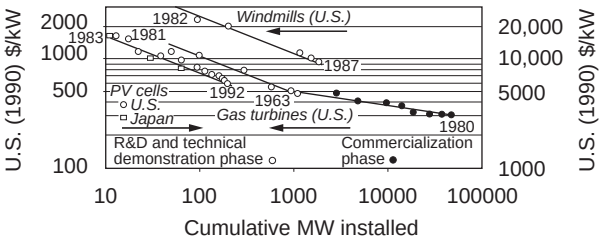


FIGURE 12 Progress ratios for photovoltaics (PV), windmills, and gas turbines. R&D, Research and development. From IIASA/WEC (1995).

for the very last unit produced). In addition, unless there is 100% spillover, the first-best social optimum requires regulating a single firm such that it produces in every period up to the level where price equals true marginal cost. This policy would require subsidizing the monopolist because it would be selling units for less than their current marginal cost in every period except the last. Absent regulation, when experience spillover is low, cumulative production experience gives incumbent firms widening cost advantages over potential entrants. Thus, early entrants can cut back their output and raise prices above the competitive level without prompting other firms to enter. The incentive to forward price partially mitigates the problem, but output remains below the social optimum. When experience does spill over substantially, firms still do not produce enough to maximize social efficiency. This follows because experience becomes a positive externality, analogous to the research and development appropriability problem. Thus, firms insufficiently forward price because they do not value the portion of their experience benefits that accrue to other firms.

To summarize, regardless of the level of spillover, strong experience effects imply that output is less than the socially efficient level. That is, some consumers would be willing to pay more than the cost of additional production, but the market fails to facilitate these mutually beneficial trades. Market transformation programs can improve social welfare by correcting the output shortfall associated with these experience effects. Moreover, as with research and development, MTPs also help to promote the use of clean energy technologies as alternatives to polluting fossil fuel technologies, and thereby reduce the social costs of pollution. When politically possible, the first-best policy is to internalize pollution costs fully (e.g., through pollution taxes set at the marginal social cost of the pollution externality or through tradable emissions permits set at the socially optimal pollution level). Governments chronically fail to achieve this, however, providing another clear rationale to support market transformation programs.

When evaluating MTPs, it is essential to account for positive feedback between the demand response and experience effects. An MTP increases the quantity produced in the first year and, due to experience effects, year 2 unit costs are lower than they would have been without the additional production from the MTP. These lower costs, in turn, imply that the quantity demanded in year 2 is higher. This "indirect demand effect," in turn, adds to cumulative production experience and further lowers unit costs in future years. This process continues indefinitely, though it gradually dissipates once the MTP is discontinued. Accounting for these indirect demand effects substantially raises the benefit–cost ratio (BCR) of typical MTPs. Even without accounting for environmental benefits, the case studies of MTPs targeting photovoltaics and efficient lighting show positive BCRs of 1.05 and 1.54, respectively.

These results suggest a role for MTPs in national and international technology policies; however, the costs of poor program design, inefficient implementation, or simply choosing the "wrong" technologies can easily outweigh cost reduction benefits. This suggests that MTPs be limited to emergent CETs with a steep industry experience curve, a high probability of major long-term market penetration once subsidies are removed, and a price elasticity of demand of approximately unity or greater. The condition that they be clean technologies mitigates the risk of poor MTP performance by adding the value of displaced environmental externalities. The other conditions ensure a strong indirect demand effect. Finally, as with energy research and development policy, public agencies should invest in a portfolio of new clean energy technologies in order to reduce overall market transformation program performance risk through diversification.

8. CONCLUSIONS

In conclusion, it is clear that the promise of renewable energy has now become a reality. Both solar photovoltaics and wind energy are experiencing rapid sales growth, declining capital costs and costs of electricity generated, and continued performance increases. Because of these developments, market opportunity exists now both to innovate and to take advantage of emerging markets, with the additional assistance of governmental and popular sentiment. The development and use of these sources can enhance diversity in energy supply markets, contribute to securing long-term sustainable energy supplies, make a contribution to the reduction of local and global atmospheric emissions, provide commercially attractive options to meet specific needs for energy services (particularly in developing countries and rural areas), and create new employment opportunities.

Acknowledgments

It is a pleasure to thank the Energy Foundation for support, as well as the University of California Energy Institute. This work draws on previous work done by the author in conjunction with two of his past postdoctoral fellows, Dr. Antonia Herzog, now of the Natural Resources Defense Council, and Dr. Timothy Lipman, now a research fellow at both the Berkeley and the Davis campuses of the University of California.

SEE ALSO THE FOLLOWING ARTICLES

Biodiesel Fuels • Biomass for Renewable Energy and Fuels • Computer Modeling of Renewable Power Systems • Energy Efficiency, Taxonomic Overview • Public Reaction to Renewable Energy Sources and Systems • Renewable Energy and the City • Renewable Energy in Europe • Renewable Energy in Southern Africa • Renewable Energy in the United States • Renewable Energy Policies and Barriers • Renewable Portfolio Standard

Further Reading

American Wind Energy Association. (2003). "Global Wind Energy Market Report." Available on the Internet at http://www.awea.org.

American Wind Energy Association. (2003). "Wind Energy Fact Sheet: Comparative Costs of Wind and Other Fuels." Available on the Internet at http://www.awea.org.

Duke, R. D., and Kammen, D. M. (1999). The economics of energy market transformation initiatives. *Energy J.* **20**(4), 15–64.

Ezzati, M., and Kammen, D. (2001). Indoor air pollution from biomass combustion and acute respiratory infections in Kenya: An Exposure–response study. *Lancet* **358**, 619–624.

Hall, D. O., and Scrase, J. I. (1998). Will biomass be the environmentally friendly fuel of the future? *Biomass Bioenergy* **15**(4/5), 357–367.

Intergovernmental Panel on Climate Change (IPCC). (1996). "Climate Change 1995: Scientific–Technical Analyses of Impacts, Adaptations, and Mitigation of Climate Change" (Watson, Zinyowera, and Moss, Eds.), Chap. 19. Cambridge University Press, Cambridge.

International Institute for Applied Systems Analysis/World Energy Council (IIASA/WEC). (1995). "Global Energy Perspectives to 2050 and Beyond" (N. Nakicenovic and M. Jefferson, Eds). IIASA/WEC, Laxenburg, Austria and London.

Johansson, T. B., Kelly, H., Reddy, A. K. N., and Williams, R. H. (eds.). (1993). "Renewable Energy: sources for Fuels and Electricity." Island Press, Washington, D.C.

Kammen, D. M., Bailis, R., and Herzog, A. V. (2001). "Clean Energy for Development and Economic Growth: Biomass and Other Renewable Energy Options to Meet Energy and Development Needs in Poor Nations." United Nations Development Programme (UNDP), Report for the 7th Conference of the Parties to the UN Framework Convention on Climate Change (COP7-UNFCCC). Marrakech, Morocco, October 29–November 9. UNDP, New York.

Margolis, R., and Kammen, D. M. (1999). Underinvestment: the energy technology and R&D policy challenge. *Science* **285**, 690–692.

Maycock, P. (1998). *PV News*.

Maycock, P. (2000). The world PV market 2000: Shifting from subsidy to 'fully economic'? *Renewable Energy World* **3**(4), 59–73.

Overend, R. P. (2000). National Renewable Energy Laboratory, Washington, D.C.

President's Committee of Advisors on Science and Technology (PCAST) Energy R&D Panel. (1997). "Federal Energy Research and Development for the Challenges of the Twenty-First Century." Report of the Energy Research and Development Panel of the President's Committee of Advisors on Science and Technology, November. U.S. Dept. of Energy, Washington, D.C.

Shell Petroleum. (2000). "Global Scenarios: 1998–2020." Summary brochure. Group Planning, Shell International, London.

Sivan, K., and Larson, E. D. (2000). "Bioenergy Primer: Modernised Biomass Energy for Sustainable Development." United Nations Development Programme (UNDP), Energy and Atmosphere Programme. UNDP, New York.

U.S. Department of Energy (DOE). (1997). "Renewable Energy Technology Characterizations." Topical report, prepared by U.S. DOE Office of Utility Technologies and EPRI, TR-109496, December. U.S. Dept. of Energy, Washington, D.C.

U.S. Department of Energy (DOE). (2000). "Annual Energy Outlook 2000." DOE/EIA-0383(00), December. Energy Information Administration, Washington, D.C.

World Energy Assessment (WEA). (2000). "Energy and the Challenge of Sustainability," Chaps. 5 and 7. Report, prepared by United Nations Development Programme, United Nations Department of Economic and Social Affairs and World Energy Council, September. United Nations Publications, New York.

Worldwatch Institute. (1999). "Vital Signs 1999. The Environmental Trends That Are Shaping Our Future." W. W. Norton, London.

Renewable Portfolio Standard

MARK JACCARD
Simon Fraser University
Vancouver, British Columbia, Canada

1. Rationale
2. Design Considerations
3. Early Adopters
4. Preliminary Assessment

Glossary

fixed feed-in tariff An offer by government or a utility to purchase electricity from renewables producers at a fixed price, regardless of the producers' costs of production.

green (renewables) certificate A certificate, provided by a regulator to renewables producers for specified units of their electricity output; can be traded to other electricity producers who need it to meet their obligation to produce a minimum percentage of renewable electricity under the renewable portfolio standard.

quantity-based policy A policy that specifies a quantity outcome (as opposed to a price-based policy such as an environmental tax or subsidy); an example is the renewable portfolio standard policy, which requires production of a specific quantity of electricity.

renewable energy Forms of energy (sunlight, wind, geothermal heat, wave power, tidal energy, hydropower, and organic matter) that flow through Earth's biosphere, available for human use indefinitely, provided that the physical basis for their flow is not destroyed.

The renewable portfolio standard (RPS) is a policy instrument to force greater production of electricity from energy sources (usually renewable) that have desirable social and environmental characteristics. By requiring market participants to deliver a minimum amount of electricity, typically as a percentage of total sales, from specific fuels and/or technologies, the RPS is a "quantity-based policy." In Europe, the policy is often referred to as "green certificate" trading. Renewable energy sources are normally defined as sunlight, wind, geothermal heat, wave power, tidal energy, hydropower, and organic matter, but the definition used in a given jurisdiction's RPS may be narrower or broader, depending on local views of the social and environmental characteristics of each technology and fuel.

1. RATIONALE

Renewables appear to have lower social and environmental impacts and risks when compared with electricity derived from conventional sources, such as fossil fuels and nuclear power. Environmental benefits may include reduced emissions contributing to smog or producing acid rain and greenhouse gases relative to a total reliance on fossil fuels, and zero risk of radiation exposure or nuclear weapons proliferation relative to nuclear power. Greater use of renewable sources may also enhance energy security by increasing supply diversity and the use of indigenous fuels, thereby reducing vulnerability to supply disruptions and volatile fuel prices. Renewables can also support economic development through reducing payments for imported fuels (increasing local multiplier effects) and fostering the development of local technical expertise.

Renewables are usually more expensive than conventional electricity sources when compared on a financial cost basis. Because of this, both monopoly and competitive electricity producers have concentrated their investment on conventional electricity technologies, with renewables usually accounting for only a small percentage of the generating stock. The only countries dependent on renewables for electricity generation are those with substantial, low-cost hydropower endowments, such as Brazil, Norway, New Zealand, and Canada. But because many large hydropower facilities disrupt ecosystems, sometimes dramatically, hydro plants exceeding a threshold of, say, 20 MW may be excluded from RPS eligibility.

 © 2004 Elsevier Inc. All rights reserved.

There are three arguments for why the financial comparison of renewables and conventional generation sources does not reflect the full costs and benefits to society of the alternative sources:

1. Some jurisdictions provide greater subsidies to conventional generation fuels and technologies.
2. The full costs of pollution (externalities) are not included in the financial costs of conventional electricity sources, and are usually much greater than pollution costs for renewables.
3. Although humans have used some renewables (such as wind) for millennia, renewables are today associated with new technologies, the relative costs of which will fall with widespread commercialization because of economies of learning and economies of scale in manufacture.

Thus, although from a strict financial cost perspective (Fig. 1, left side) most renewables are more expensive compared to conventional electricity, the situation may be reversed when these three additional factors are taken into account (Fig. 1, right side). Inclusion of these three factors is referred to as the "long-run, social cost perspective." Whereas it is not certain that renewables are always less expensive on a social cost basis, they are nonetheless recognized as desirable sources of electricity that will have difficulty penetrating the market given the pressure on electricity investors to focus on financial considerations. This constraint is exacerbated if electricity markets continue the current trend toward independent electricity production, mostly by privately owned companies.

In recognition of the possible advantages of renewables on a social cost basis, governments have developed policies to improve the financial competitiveness of renewables. There are four broad categories of measures to support renewables:

1. Assist the commercialization of renewables with indirect support. This may include funding of demonstration projects, audits, evaluations, resource assessments, research and development support, and training.
2. Provide direct financial support to renewables. This may include capital grants, preferential purchase prices, tax advantages, or low-interest loans. Energy price subsidies can be in the form of preferential fixed-price tariffs (feed-in tariffs), a fixed premium on the prevailing electricity price, or a premium payment to renewable electricity producers determined by a "competitive bidding" process.
3. Increase the cost of polluting sources of supply. This may include the reduction or elimination of subsidies to conventional sources, higher energy or pollution taxes (externality taxes), stronger technology regulations, or emission cap regulations.
4. Foster voluntary or mandatory minimum market shares for renewables. This may include voluntary agreements with producers, green tariffs that allow consumers to pay extra for renewables, and the RPS.

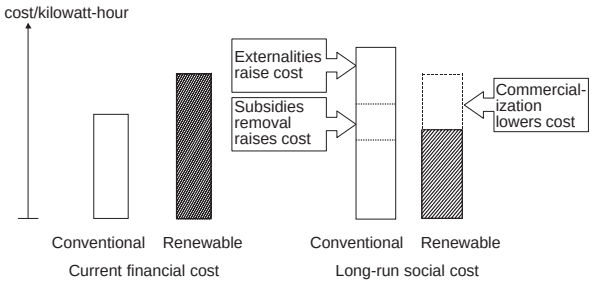


FIGURE 1 Conventional versus renewable electricity: financial and social cost comparison.

Since the energy crisis of the 1970s, governments have especially relied on the first two mechanisms to support renewables. Various indirect support policies (e.g., research and development subsidies) have been in place throughout the past 30 years. But the dominant policy in terms of impact has been direct financial support. Many governments have provided tax credits for renewables. Electric utilities and governments have offered above-market, fixed-tariff prices to acquire electricity from renewables producers. In the United States, the Public Utility Regulatory Policy Act of 1978 required utilities to purchase renewables and other independent electricity at prices that reflected the long-run cost of new, high-cost nuclear and fossil fuel plants. In Europe, fixed-tariff price support was implemented through direct government subsidy (e.g., in Denmark) or cross-subsidy from other utility customers (e.g., in Germany). In the 1990s, however, interest in the third and fourth approaches has increased as governments have explored policies that raise the cost of polluting sources of supply. Europe has focused on pollution regulations and externality taxes, and the United States has focused on regulations and systems that involve emissions caps and tradable permits. Industrialized countries, and increasingly developing countries (e.g., China), have also reduced subsidies to conventional fossil fuel and nuclear energy producers. Voluntary green tariffs and the RPS

emerged in the late 1990s in Europe, the United States, and Australia.

Three characteristics help explain the emerging interest in the RPS: First, an RPS maintains continuous incentives for renewable producers to seek cost reductions (economic efficiency) and can be designed to ensure that these cost reductions are passed on to consumers (equity). This is achieved by mechanisms that establish continuous cost competition among renewable producers for their share of the RPS. Second, because an RPS ensures the attainment of a specific market share, it can be directly linked to government policy objectives. Governments can set an RPS target as part of a package of policies to meet an environmental target such as CO_2 reduction. Third, an RPS minimizes government involvement relative to other measures. The government's budget is not implicated because customers pay producers directly the extra financial cost of renewables, and the selection of winning bids can be left to market forces (competitive bidding) instead of government evaluation. Because the RPS has only recently emerged as a policy tool, there is insufficient experience to provide an empirical evaluation of its performance relative to key alternatives, such as subsidized feed-in tariffs and competitive bidding for subsidies. However, critical issues in its design have already emerged.

2. DESIGN CONSIDERATIONS

Many considerations influence the design of an RPS. Major categories include selection of the target, determining resource eligibility, determining where and to whom the policy is applied, designing flexibility mechanisms, and setting administrative responsibilities. Although these are presented individually here, they cannot be considered in isolation from each other in the design process. A clear set of policy objectives is required for navigating these considerations. These include standard policy objectives, such as effectiveness, economic efficiency, equity, political acceptability, and administrative feasibility, but also the specific environmental and perhaps social goals that are the motivation for the RPS in the first place. For example, the RPS has been linked to the greenhouse gas reduction goals of several jurisdictions. The following discussion of design considerations assumes that the policy objectives include the standard list, as well as those environmental objectives that tend to favor renewables relative to fossil fuel- and nuclear-generated electricity.

2.1 Target Selection

The main challenge of the RPS is determination of the binding target or quota for renewable energy production. There are several aspects to this determination.

2.1.1 Target Size

Setting target size requires consideration of impacts on different objectives, including anticipated costs. These impacts will depend in part on the local cost and availability of renewables and the price of conventional electricity sources. From a political acceptability perspective, the target would be large enough to move the industry toward the environmental objective, but not so large that it results in a dramatic increase in electricity prices. The size of the target can also affect the level of competition among fuels or technologies. For example, if targets could be achieved with a relatively small number of large projects, there may be less competition and little pressure to innovate. Some knowledge of the potential costs and quantities of renewables is therefore important in establishing targets.

2.1.2 Target Timing

Another consideration in selecting the target is timing. Some lead time may be required to permit cost-effective responses to the RPS target. The length of the lead time will depend on the magnitude of the target and the characteristics of eligible resources. If existing facilities are not eligible for meeting targets, sufficient lead time will be required for investments in new facilities. A phase-in period would enable the government to conduct interim evaluations in order to assess impacts and make target adjustments where necessary, thereby maintaining fairly stable compliance costs. However, duration and stability of targets can also be important so that buyers and sellers can more easily negotiate long-term supply contracts. Long-term commitments that guarantee a revenue stream are important for securing low-cost project financing, especially for capital-intensive technologies such as renewables.

2.1.3 One or Multiple Renewables Targets

An issue in target setting is whether there should be one target for which all renewables compete, or a separate target for different classes of renewables. Under the first approach, the lowest cost options would be developed to meet the aggregate target. This approach should advance environmental goals more cost-effectively, but it might do less well in

furthering other objectives, such as greater supply diversity or increased support for newly emerging, but still expensive, types of renewables. Perhaps one renewable (e.g., wind) will capture the entire RPS market even though other renewables would eventually become competitive were they assisted in achieving commercialization and production thresholds. The state of Connecticut designed its RPS with separate targets for Class 1 renewables (solar, wind, landfill gas, and sustainable biomass) and Class 2 renewables (hydropower and municipal solid waste), the intent being to limit the share of hydropower and municipal solid waste within the RPS.

2.1.4 Target Adjustment and Other Cost Cap Measures

If the RPS is set too high, it could substantially increase the price of electricity, with perhaps significant political acceptability and economic performance impacts. There are several ways in which the government (or RPS administrator) could deal with this challenge. First, the RPS target could be voluntary (as in The Netherlands in its initial policy), or converted to voluntary status as higher-than-anticipated costs emerge. Second, the government could adjust the target downward as evidence of high cost emerged. Third, the government could set a modest penalty that participants would pay to the extent that they failed to comply completely with the RPS. This penalty provides assurance to participants that costs will not rise above the penalty level. Fourth, if the RPS includes a system of tradable renewable certificates (green certificates in Europe), the government could offer to sell an unlimited number of these at a modest price, in effect providing a price cap (e.g., The Netherlands). These measures all ensure that costs will not exceed a given level, but at the possible expense of not achieving the RPS target, and the corresponding environmental target associated with it. On the negative side, adjusting targets could jeopardize the credibility and predictability of the program, with serious consequences for its long-term success.

2.2 Eligible Resources

In addition to specific targets, the RPS must establish resource eligibility. This will depend on the objectives for the RPS and the local viability of different types of resources.

2.2.1 Renewables vs Other Desired Technologies

A key eligibility decision is whether the RPS should be limited to renewable sources of electricity or can include other energy/technology combinations. This depends on the objectives and, in some sense, on the comfort level with using the term "renewables" to symbolize the broader concept of environmentally desirable technologies. In Italy, only truly renewable forms of energy are included in the RPS definition. In some jurisdictions, however, fossil fuel-driven cogeneration or fuel cell technology is included because of the higher energy efficiency and lower local emissions of these mechanisms. In proposing its RPS, Australia included solar water heaters that displace electric water heaters that use electricity from fossil fuel plants.

2.2.2 Existing Renewables vs New Investments

Another eligibility consideration is whether the RPS should apply to all renewable resources or only to new investments in renewable resources. Applying it to new investment helps advance the new technology commercialization intent of the RPS, as adopted in the Italian RPS. However, this may increase administrative complexity by requiring ongoing monitoring and regulatory rulings as to whether a given expenditure is just maintenance and refurbishment of an existing facility or a truly new investment (a problem experienced in a different context with the U.S. Clean Air Act). Is a complete reconstruction of a small hydropower facility a new or old investment? If new, what if only the turbines are replaced? If, however, the policy goal is simply to expand output from renewables at the lowest cost, the government may be indifferent regarding investments for reconditioning or for totally new facilities.

2.2.3 Grid-Connected Renewables vs All Renewables

The RPS may apply only to grid-connected facilities or to both grid-connected and off-grid facilities. Although the simplest approach is to focus the policy on changing the grid-transmitted market share of renewables, some would argue that environmental objectives are equally served by the development of renewables anywhere, including off grid. However, administrative realities must be considered. It would be difficult to verify the annual production of a myriad of small off-grid generators, leading to high administrative costs.

2.2.4 Facility Size

The size of resources may also be part of the RPS eligibility determination. Size limits can affect the number of potential projects required to meet the

overall target. They can also indirectly affect objectives such as diversity and the location of projects. The RPS of the state of Maine in the United States had an upper limit of 100 MW, which applied to both existing and new facilities. A limit of 10 or 20 MW is sometimes suggested as the maximum size that should be allowed for hydropower projects to be included in the RPS.

2.2.5 Imports vs Domestic Renewables

External, renewable electricity producers may wish to sell into a jurisdiction with an RPS standard. Again, determination of eligibility depends on the policy objectives. If the expansion of domestic resources is key, imports of renewables may be limited or excluded from the RPS, which does not prevent them from competing in the non-RPS market against conventional generation facilities. In many jurisdictions, external participation in the RPS will require some consideration of existing trade agreements and commercial law, which may limit policymakers' ability to discriminate between domestic supplies and imports. Both Europe and North America are working toward electricity trade rules that should make external participation easier.

2.3 Applicability

The applicability of the RPS target is another important design consideration. Key issues include the geographic coverage of the standard, the applicability of the standard to specific market participants, and whether it should be applied to capacity or energy.

2.3.1 Geographic Coverage

In terms of geographic coverage, the RPS may be most effective if the breadth of its application is matched to the breadth of environmental and other objectives. This explains why the European Union (EU), which sees renewables as a component of its greenhouse gas emission reduction objective, is seeking to achieve an EU-wide RPS. However, the physical limitations to trade must be considered. In some countries, the lack of grid interconnection makes implementation of even a national RPS problematic, unless some form of credit trading is permitted. In these cases, it may be easier initially to set an RPS for each major grid area.

2.3.2 Market Participants Regulated by the RPS

The RPS must specify the market participants to which it applies. Possible participants include producers, wholesale customers or brokers, retail suppliers, distribution utilities, and consumers. Some jurisdictions want the RPS to be associated with final consumption, hoping that the public will better understand its link to environmental or other policy objectives. Denmark's proposed RPS applies to consumers, although distribution utilities or retail suppliers would presumably represent the interests of small consumers. Others believe that the administrative ease of regulating producers is important. Italy has size restrictions so that the standard applies to only about 100 producers nationwide.

2.3.3 Energy Production vs Installed Capacity

The RPS may apply to energy production or installed capacity. Capacity may be easier to verify, but energy production is directly connected to environmental objectives. An energy-determined RPS provides the incentive for producers to maximize production from individual projects, and thereby lower costs.

2.4 Flexibility Mechanisms

An RPS can be less costly to implement, yet still achieve the same target, if there is flexibility in how it is applied both to individual producers and across the entire market.

2.4.1 Account-Balancing Mechanisms for Individual Producers

One way to increase flexibility is to allow those to whom the policy is applied to have some flexibility in the choice of calendar year and in the period for achieving their annual RPS. This is known as an "account-balancing mechanism." For example, the inherent output variability of some types of renewables, such as wind, hinders accurate fulfillment of the RPS target each year. The electricity provider can be given extra months to meet the previous year's commitment as long as the full RPS requirement is achieved on average over a period of years. The Australian RPS has a 3-year balancing mechanism for electricity providers.

2.4.2 Trading Mechanisms between Producers

Another way to increase flexibility is to allow trade among electricity providers so that the RPS target is met in aggregate, even if some providers exceed the RPS and others fail to achieve it. This approach can significantly decrease the total cost of the

RPS because it allows the maximum electricity production from those renewables generators with the lowest costs. This trading can be facilitated by the creation of tradable certificates, hence the European term "green certificate" trading for the RPS. As of 2003, The Netherlands, Italy, England, Australia, and some U.S. states (Texas) had such certificates. The creation of certificates is not essential, however, for trading to occur. The RPS administrator could allow electricity providers to use a variety of instruments to demonstrate achievement of the RPS target in aggregate while they trade among themselves. Contracts between providers could be filed with the administrator and these would be verified against actual inputs to the grid.

2.5 Administrative Responsibilities

There are several administrative responsibilities in implementing a RPS. These could all be handled by one RPS administrator or delegated to specialized agencies.

2.5.1 Setting the RPS Target

The RPS target would likely be set at a senior level of government because it involves critical trade-offs between environmental targets, economic development goals, and concerns about electricity prices.

2.5.2 Certification of Renewables

Certification of renewables could be delegated to the operator of the grid, subject to periodic review by the government agency responsible for administering the RPS. Environmental regulators could help with the definition of what is a renewable resource, although, again, this may be a high-level policy decision.

2.5.3 Compliance Monitoring

Compliance monitoring could be delegated to the operator of the grid, because this entity will already have information on production of connected facilities. If nongrid production is also eligible for the RPS, compliance verification will be much more difficult and may require some other entity.

2.5.4 Setting and Collecting Penalties for Noncompliance

Penalties for noncompliance may be set at senior levels of government, for consistency with other public policies, but could then be administered by the grid operator or environmental agency.

3. EARLY ADOPTERS

The renewable portfolio standard is a new policy tool. Many jurisdictions (national, regional) are studying it, and several have RPS legislation under consideration. But as of 2003, only a few jurisdictions (and only in industrialized countries) have enacted RPS legislation. Although it is too early to assess the effectiveness of the RPS relative to alternative policies, a survey of early adopters provides an initial indication of the preferences of policymakers when trading-off implementation objectives and constraints. Because the RPS legislative situation is so dynamic, this survey is limited to general policy developments, and even some of these may soon change.

3.1 United States

The United States Congress has considered several RPS legislative proposals. In the spring of 2003, the U.S. Senate passed a bill requiring new renewables to provide an additional 1% of U.S. electricity by 2005 and 10% by 2020. The RPS would include a credit trading mechanism between producers to reduce the cost of compliance. The bill also sustains a production tax credit for renewable generators. In 2003, just over 10 U.S. states had enacted an RPS, but many additional states were considering legislation. States with an RPS enacted or in final consideration included Maine, Massachusetts, Connecticut, Pennsylvania, New Jersey, Minnesota, Wisconsin, Iowa, Texas, New Mexico, Arizona, Nevada, and California. Most RPS legislation has been associated with electricity sector reform toward greater generation competition, with RPS targeting electricity retailers. Almost all programs envision flexibility mechanisms (tradable credits), although these are mostly in the development stage. State-level electric utility regulators are usually charged with the responsibility for program administration. Urban waste and landfill gas are often included as eligible resources, and sometimes technologies such as fuel cells and cogeneration.

Variation in the RPS from one state to another reflects the unique circumstances and objectives of each state. At 30%, Maine's RPS target appears ambitious, but the RPS includes all currently operating renewables and the state already generates 50% of its electricity from eligible hydro and biomass resources. RPS advocates hail the Texas RPS as the best-designed and most ambitious policy in the

country. It envisioned the development of an additional 2000 MW of mostly wind power by 2009, which, in combination with existing renewables, would provide about 3% of total state electricity provision. Wind power has expanded rapidly since the inception of the RPS.

3.2 European Union

As part of its effort to establish a liberalized internal electricity market, the EU has been developing harmonized rules on the treatment of renewables in order to avoid distortion of trade and competition in electricity from renewable sources. In 1997, the EU issued a White Paper setting a goal for a minimum 12% share of renewable electricity supply by 2010. In 2000, the EU issued a directive proposal for a community-wide policy. To date, the directive proposal is not binding, nor does it specify the policy that should be used. Nonetheless, interest in the RPS at the EU-wide level is strong, and several countries have implemented the policy and several are close to implementation.

In Italy's RPS, operators who produce and import more than 100 GWh are obliged, from 2001, to provide at least 2% of the electricity produced and imported (net of cogeneration, exports, and self-consumption) from new, renewable generation. Renewables generators earn green certificates, which they can use or trade to nonrenewables producers in a tradable certificate market. The goal of the RPS is to triple electricity production from renewable sources by 2010.

The government of The Netherlands initially negotiated a voluntary RPS with the country's major utilities, but participants behave as if it were mandatory because meeting their RPS quota allows them to avoid an environmental tax. Producers of renewable electricity (utilities and independent power producers) receive for each kilowatt-hour provided to the grid a subsidy, which is financed by an environmental tax throughout the country on conventional energy consumption. Producers and utilities can also trade green energy certificates among themselves to meet their obligation under the RPS.

England has recently implemented an RPS with tradable certificates (called the Renewables Obligation). This replaces its system of the past decade, which involved competitive bidding by potential renewables generators for funds collected from all customers in a tariff charge called the Non-Fossil Fuel Obligation. That system did not stimulate as much renewables production as did the feed-in tariffs for producers of Denmark, Germany, and Spain, but it did lead to declining bid prices for electricity offers to the grid from renewables producers. Sweden, Belgium, and Austria implemented RPS policies in 2003 and several other European countries are close to implementation. Denmark has been known for its generous feed-in tariffs that have fostered the dramatic growth in wind power in that country. But the subsidies are a substantial drain on public resources (100 million Euros in 1998), and the Danish parliament recently developed legislation to create an RPS, slated for implementation in the 2004–2005 period.

3.3 Australia

The Australian federal government enacted a nationwide RPS in 2001 to operate within its competitive electricity market. The policy includes a certificate trading scheme.

4. PRELIMINARY ASSESSMENT

The RPS is a policy tool to expand production of renewable-generated electricity to reflect what its role would be if long-run social costs instead of short-run financial costs were the determinant of electricity generation investment. By establishing a guaranteed market share for renewables producers, the RPS provides a strong incentive for new renewable investment, an ongoing motivation for current and new renewable producers to reduce costs, and a relatively sure means of attaining environmental targets while reducing the requirement for government financial and managerial involvement. The RPS can be applied both to electric utilities that are regulated monopolies and to competitive independent producers, but it is generally associated with jurisdictions that have reformed their electricity sector to increase competition in electricity generation. Several European countries, several U.S. states, and Australia have thus far implemented the RPS, but it is under active consideration in many other jurisdictions. Legislation and implementation are underway for an EU-wide RPS in Europe and a federal-level RPS in the United States.

It is too early to provide an empirical evaluation of the RPS relative to alternative mechanisms. However, the survey reported in here provides

information on the critical considerations in designing an RPS and a general sense of the choices being made by different jurisdictions. A few generalizations can be offered at this early stage in the development of the RPS.

The selection of the RPS target involves a three-way trade-off between environmental improvement, long-term technology strategy, and short-run cost. Jurisdictions that are especially focused on getting the greatest environmental improvement at the least cost will have only one standard for all renewables and may even include technologies that are not renewables. Jurisdictions that are interested in long-term development of advanced renewables technologies will tend to have categories of renewables (and perhaps environmentally desirable technologies that use fossil fuels), ensuring a more diverse mix of renewables and associated clean technologies. Most jurisdictions are concerned with the uncertainty of cost, so cost caps or equivalent mechanisms are common.

In terms of eligibility, most jurisdictions are starting out conservatively, limiting the scope to local renewables, excluding off-grid renewables, and trying to make sure that truly new renewable resources result from the RPS. Geographic application of policies is usually limited to individual electricity jurisdictions. However, with Australia's federal system as an example, there are efforts at the U.S. federal level and the EU level for multi-jurisdictional application. The general tendency is to apply the standard to providers rather than to consumers of electricity. The standard is almost universally applied to energy production rather than to capacity, which is consistent with environmental objectives.

Flexibility mechanisms are widely accepted as desirable. Their implementation is slowed by the administrative complexity that they can entail. Administration of the RPS in the United States and Australia is conducted through a combination of government and delegation to electricity regulators. In Europe, it is more in the domain of government at this point, but various institutional reforms are being undertaken. Thus far, all jurisdictions implementing a RPS have continued with at least some of the other policies that support renewables, whether these be research and development and demonstration projects, tax credits, or subsidized feed-in tariffs. This final point reintroduces the more fundamental concern about how the RPS compares to alternative policies, given the public objectives it was meant to serve. Critiques have argued that the RPS is inevitably flawed because it forces the adoption of particular forms of energy, even though these are but indirect means to an end, such as environmental improvement. In so doing, the RPS may delay or prevent the development and commercialization of technology innovations that would more economically achieve the same objective. If, for example, a key motive for the RPS is to reduce greenhouse gas emissions, a policy focused directly on these emissions—such as a greenhouse gas tax or emission cap and tradable permit system—would open the door to technologies such as the production of hydrogen from coal with capture and permanent storage of all potentially harmful by-products in geological formations. Conceivably, this or other alternatives to renewables could meet all of the public objectives behind the creation of the RPS at a much lower cost, freeing resources for other social objectives.

Advocates of renewable energy have countered that nonrenewable resources such as "clean coal" cannot possibly outperform renewables in providing affordable energy while also satisfying long-term environmental, security, and social objectives. A less exclusionary argument in favor of the RPS is that renewables are likely nonetheless to be part of a future energy mix that might one day also include cleaner fossil fuel and nuclear technologies, but that renewables need special help at this stage to pass critical commercialization thresholds.

In particular, the RPS has several attractive features for driving long-run technological change in renewables. It can provide an incentive that influences the contribution of renewables to new electricity generation investment without changing the costs and economic viability of existing facilities; thus, it produces technological change at a pace consistent with the natural rate of equipment stock turnover and does not force economic obsolescence on existing equipment. For consumer prices, the higher costs of new renewables are blended with the lower costs of existing, conventional generation; price changes are therefore negligible and the likelihood of political acceptability is greater. The RPS ensures that renewables attain critical commercialization thresholds that should enable them to achieve economies of learning and economies of scale, with their costs of producing electricity falling accordingly. If the RPS includes tradable certificates and other flexibility mechanisms, it reduces the policy cost and again increases the likelihood of political acceptability. The RPS can be closely linked to environmental objectives because it provides a higher

level of confidence of the environmental outcome than do price-based policies. The RPS does not require government subsidies, and can even allow a reduced role for government in the policy's administration, because this role is delegated to electricity regulators. Finally, RPS targets can be set and adjusted to ensure that the costs of renewables expansion are not significantly greater than those of alternative means of achieving the slate of policy objectives they are intended to satisfy.

SEE ALSO THE FOLLOWING ARTICLES

Biomass for Renewable Energy and Fuels • *Computer Modeling of Renewable Power Systems* • *Public Reaction to Renewable Energy Sources and Systems* • *Renewable Energy and the City* • *Renewable Energy in Europe* • *Renewable Energy in Southern Africa* • *Renewable Energy in the United States* • *Renewable Energy Policies and Barriers* • *Renewable Energy, Taxonomic Overview*

Further Reading

Berry, T., and Jaccard, M. (2001). The renewable portfolio standard: Design considerations and an implementation survey. *Energy Policy* **29,** 263–277.

Loiter, J. M., and Norberg-Bohm, V. (1999). Technology policy and renewable energy: Public roles in the development of new energy technologies. *Energy Policy* **27,** 85–97.

Menanteau, P., Finon, D., and Lamy, M.-L. (2003). Prices versus quantities: Choosing policies for promoting the development of renewable energy. *Energy Policy* **31,** 799–812.

Voogt, M., Boots, M., Schaeffer, G., and Martens, J. (2000). Renewable electricity in a liberalized market—The concept of green certificates. *Energy Environ.* **11,** 65–80.

Wiser, R., and Langniss, O. (2001). The renewable portfolio standard in Texas: An early assessment. Report No. LBNL-49107. Lawrence Berkeley National Laboratory, Berkeley, California.

Reproduction, Energy Costs of

THOMAS H. KUNZ
Boston University
Boston, Massachusetts, United States
KIMBERLY S. ORRELL
Virginia Polytechnic Institute and State University
Blacksburg, Virginia, United States

1. **Introduction**
2. **How Energy Is Expended for Reproduction**
3. **Techniques Used to Study Reproductive Energetics**
4. **Reproductive Energetics of Reptiles**
5. **Reproductive Energetics of Birds**
6. **Reproductive Energetics of Mammals**
7. **Concluding Remarks**

Glossary

altricial When offspring are hatched or born at an early stage of development.
basal metabolic rate (BMR) Resting metabolic rate of a homeotherm measured at thermoneutrality.
energy budget Partitioning of energy expenditure into the competing demands of maintenance, growth, reproduction, and storage.
gestation Period of embryonic development in live-bearing vertebrates.
iteroparous More than one reproductive event per lifetime.
lactation Period of milk production in mammals.
life history trade-offs Constraining relationships among life-history traits (e.g., current reproduction and survival, current reproduction and future reproduction, number and size of offspring).
life history traits Traits that define the life cycle of an organism (e.g., size at birth, growth pattern, mode of reproduction, age and size at reproductive maturity, size of offspring, fecundity, reproductive life span, longevity).
oviparous Egg laying; offspring develop within eggs outside a female's body.
precocial When offspring are hatched or born at an advanced stage of development.
production energy Energy allocated to growth (e.g., lean somatic tissue), storage (e.g., lipid deposition), and reproduction (e.g., gonadal recrudescence, gametogenesis, egg production).
reproductive effort Proportion of total energy devoted to reproduction within a biologically meaningful time interval.
reproductive success Number of offspring produced by an individual.
semelparous One reproductive event per lifetime.
viviparous Live-bearing; offspring develop fully or partially within the female reproductive tract.

The amount of energy that an animal can devote to reproduction is limited. The habitat in which an animal lives may contain limited food and/or water resources (e.g., deserts, high altitudes), environmental conditions may limit energy acquisition by reduced foraging activity (e.g., predation, daily or seasonal weather extremes), and there are morphological and physiological constraints on the rates at which food can be obtained and processed. Within an individual, energy must also be allocated to physiological processes and behavioral activities associated with survival and growth (e.g., foraging, thermoregulation, predator evasion). Thus, the total energy that an animal obtains must be allocated in such a way that balances maintenance, growth, reproduction, and survival.

1. INTRODUCTION

Life history traits largely determine how the energy intake of an animal is allocated to reproduction, growth, and survival. Some life history traits include the size of an animal at birth (or at hatching), the age and size at which an animal reaches reproductive maturity, how it reproduces (oviparous or viviparous), how often it can reproduce (semelparous or iteroparous), the length of time it is reproductively

 © 2004 Elsevier Inc. All rights reserved.

active, the number and size of offspring produced, and its expected age-specific mortality and longevity. Constraints or trade-offs exist among life history traits, many of which are physiological and relate to energy management. A classic example of such a trade-off is size versus number of offspring. The allocation of energy by an adult female to produce larger offspring may improve the offspring's chances of survival, but it also can reduce the number of offspring that the female can produce. Another trade-off is between the energy allocated to current reproduction and future survival. If an animal allocates too much energy to reproduction, not enough energy will be available for maintenance or vice versa. After fat reserves are depleted, other body tissues (muscle mass) are metabolized for energy and body condition declines. Deterioration in body condition and immune function contributes to susceptibility to predators, parasites, and disease, eventually leading to death. In semelparous species (e.g., salmon), this is exactly what happens; an individual allocates all available energy to a single burst of reproduction and then dies (termed "big bang" reproduction). Iteroparous species are those that have a long enough life span that they have a high probability of surviving to reproduce more than once, and their life history traits tend to balance energy allocated to current reproduction and maintenance. Iteroparous species also illustrate another important life history trade-off: current reproduction versus future reproduction.

Longevity, age-specific mortality schedules, and reproductive life span also have a strong influence on the patterns of energy allocation to reproduction as individuals age. For example, some species invest less energy in reproduction when they are young and relatively small in body size than when they are older and larger. Smaller individuals have fewer energy reserves than do larger individuals, they have less space in their body cavity (limiting the number of eggs or offspring they can produce), and they have a narrower pelvic canal that limits egg or offspring size. Thus, smaller individuals are constrained both energetically and anatomically and, thus, produce fewer and/or smaller offspring rather than more and/or larger individuals. By allocating less energy to reproduction, smaller individuals can allocate more energy to growth—building up body condition and attaining larger body size more quickly than if they allocated less energy to growth. When individuals are older and larger, they can better afford to allocate a greater proportion of their energy to reproduction and do so when it enhances their reproductive success. The question arises as to why young and small iteroparous animals should reproduce at all. Why not allocate energy to growth and delay reproduction until they are larger? In fact, many species do delay reproduction when they are young, even though they are physiologically mature and able to reproduce. The age-related schedule of energy allocation to reproduction in a given species is part of a suite of life history traits that coevolve to optimize the balance of energy allocated to maintenance, growth, reproduction, and survival.

Reproductive energetics is also a central aspect of theories that explain sexual selection and the evolution of mating systems. Bateman's concept of anisogamy (differential investment by each sex in gametes) and Triver's concept of parental investment are at the root of the relationship among reproductive energetics, sexual selection, and the evolution of mating systems. In general, females allocate energy to relatively few, but energetically expensive, eggs. In some species, females make a substantial energy investment in gestation, lactation, and other forms of parental care. Thus, the reproductive output of females is limited primarily by energy resources. Because variance in female reproductive success is relatively low, the influence of sexual selection on females is relatively weak. In contrast to females, males allocate minimal energy for the production of sperm; thus, male reproductive success is not usually limited by the amount of energy available for gamete production. Instead, male reproductive success is limited by the ability to compete for and successfully mate with females. As a consequence, males often invest a considerable amount of energy in activities and traits associated with mating success (e.g., advertisement displays, territorial defense, agonistic encounters, courtship, large body size, structures used for courtship or intermale combat). In addition, males that are emancipated from parental care typically invest time and energy in defending or pursuing multiple mates, contributing to a polygynous mating system, and engaging in intense intermale competition for mates. For other species, both males and females (or, in rare cases, males alone) are required to invest time and energy in caring for offspring. Caregiving males rarely have the time and energy to pursue multiple mates; thus, a monogamous mating system usually results. Although monogamous males also expend energy to compete for mates, their reproductive success does not vary as much as that of polygynous males; thus, intermale competition and the influence of sexual selection on males are not as intense as in polygynous systems.

2. HOW ENERGY IS EXPENDED FOR REPRODUCTION

Because of wide variation in life history traits, reproductive biology, and mating strategies, the ways in which energy is expended for reproduction differ greatly among species and between the sexes. Because most vertebrate species breed seasonally, they do not typically produce gametes year-round, and many species maintain nonreproductive gonadal tissues in a reduced state. Thus, the initial investment of energy into reproduction actually begins prior to mating when energy is expended for gonadal recrudescence. Once hormones are produced, they stimulate the development or enlargement of the testes, ovaries, and oviducts or uterus. Once the reproductive organs are ready, energy is expended to produce gametes. For males, this includes spermatogenesis and production of seminal fluids. For females, egg production starts with oogenesis but goes beyond the simple production of sex cells. Oviparous reptiles and birds allocate a considerable amount of energy to egg production; eggs are provisioned with a large mass of energy-rich lipids or yolk, egg white (water and albumen), and a calcified eggshell. In comparison, viviparous (live-bearing) mammals produce eggs that contain little stored energy but allocate a large amount of energy to carry and nourish developing embryos—and for lactation to nourish offspring for all or part of the period of postnatal growth.

In addition to physiological investment in reproduction, a large amount of energy is expended for reproductive activities. Examples of reproductive activities include finding and defending a suitable territory or nest site, fighting or repelling competitors, collecting nest materials and building a nest or excavating a burrow, and finding, choosing, and courting a mate. The act of copulation is a reproductive activity, as is the laying of eggs, giving birth to live offspring, and guarding, burying, incubating, or carrying eggs. Males of some species actively guard their mates prior to and/or after copulation to ensure paternity. For crocodilians as well as many bird and mammal species, reproductive activities in the form of parental care may continue long after offspring are hatched or born. These activities include feeding the offspring while they are unable to forage for themselves, guarding them from predators, and tutoring them in social and survival skills. In some cases, it may be difficult to draw the line as to whether a particular activity should be classified as reproductive. Nonbreeding types of activities that occur during the breeding season may occur more frequently to support reproduction. For example, some animals might need to forage and feed more often prior to or during the breeding season to support reproduction, and males might forage to provide nuptial gifts to impress prospective mates or to feed nest-bound mates that incubate eggs or guard offspring. Many animals (mostly birds and mammals) exert some degree of thermoregulation to survive, but reproductive females might need to maintain higher than normal body temperatures and basal metabolic rates (BMRs) to promote the optimal development of offspring or to increase the rate of digestion, assimilation, and egg or milk production.

3. TECHNIQUES USED TO STUDY REPRODUCTIVE ENERGETICS

Because energy is expended for reproduction in so many different ways, quantifying this energy is often complicated and difficult. As a result, many studies examine only part of the energy expended for reproduction. One of the easiest components of reproduction to determine is the energy allocated to the storage of fat and the production of eggs. Lipid extraction is used to separate body fat from lean tissue and to quantify the energy contained in body fat. Bomb calorimetry is used to determine the energy content of eggs and lean body tissues (Table I). Changes in body mass and body composition observed in a cohort of individuals over time, and the energy content of fat and lean tissue, can be used to calculate the amount of energy allocated to storage and/or growth (Fig. 1).

Another technique used to examine reproductive energetics is laboratory respirometry (Table I). Respirometry is used to measure metabolic rates (in milliliters O_2 consumed per hour or CO_2 produced per hour) of inactive animals held at species-typical body temperature(s). BMRs (in mammals or birds) and standard metabolic rates (SMRs) (in reptiles) of fasted sleeping animals are the lowest rates of energy metabolism and represent the amount of energy expended for basic body functions. Resting or maintenance metabolic rates (RMRs) are determined in inactive or sleeping animals digesting food and so include the cost of digesting and assimilating food (specific dynamic action). Some studies have determined metabolic rates of animals engaged in various activities. Laboratory metabolic rates have frequently been used to extrapolate the amount of total energy expenditure, with extrapolated values then being used

TABLE I

Analytical Techniques Used to Study Reproductive Energetics of Terrestrial Vertebrates

Activity profiles: Time budgets detailing the amount of time spent in various activities. They may be combined with respirometry to calculate the amount of energy allocated to activity and maintenance.

Lab respirometry: Measures the metabolism (in milliliters O_2 consumed and/or CO_2 produced per hour, converted to joules of energy) of inactive animals under laboratory conditions.

Doubly labeled water: Determines the daily energy expended or consumed (in milliliters CO_2 per day, converted to joules of energy) by free-ranging animals. Isotopes of oxygen (O–18) and hydrogen (deuterium or tritium) are injected into an animal, blood samples are taken over two or more intervals of time, and then the turnover rate in isotopes is used to calculate the amount of energy consumed during that time by oxidative metabolism.

Lipid extraction: Used to separate and quantify the energy contained in whole body lipids. An animal carcass is dried to a constant mass and macerated, and solvents are used to extract the lipids for further analysis.

Bomb calorimetry: Determines the energy content of organic substances (e.g., body tissues, eggs). Dried material is incinerated in high-pressure oxygen in a stainless steel container at a constant volume. The heat produced by the reaction is absorbed by water surrounding the container, and the change in water temperature indicates the amount of energy released.

Energy budgets: Where data from several different techniques are combined to partition total metabolizable energy into that allocated to competing categories of resting maintenance, activity maintenance, and production.

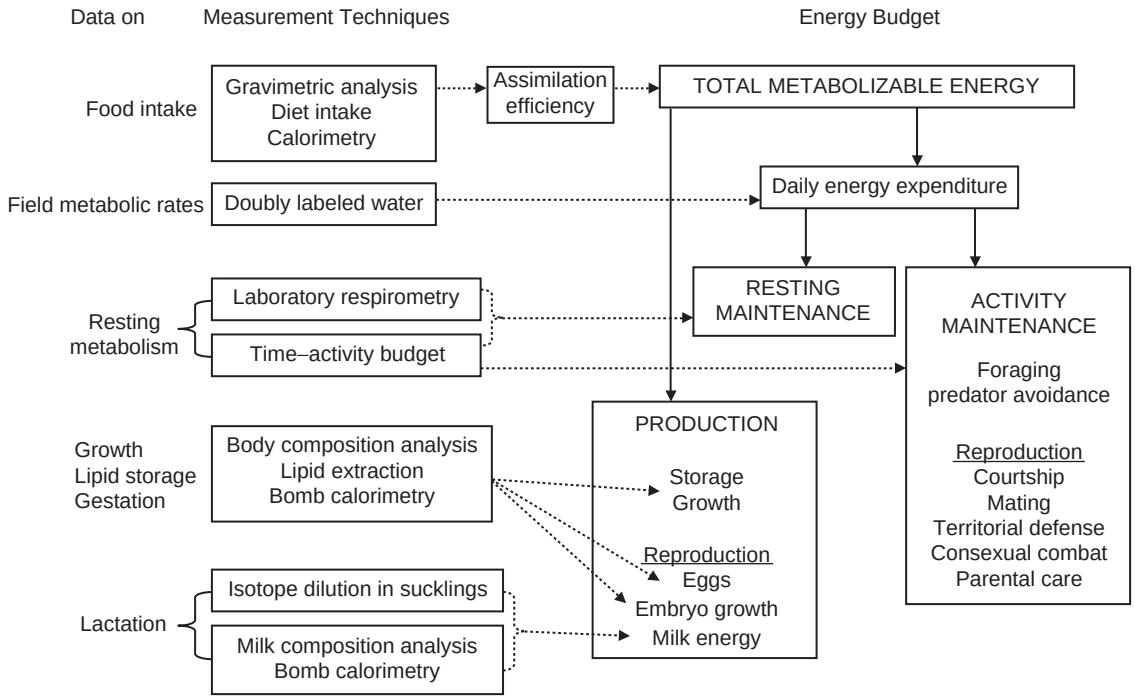


FIGURE 1 Schematic diagram of a generalized energy budget used to examine reproductive energetics, including the type of data used, measurement techniques and definitions of energy budget components, and methods used to construct an energy budget. Modified from Kunz and Nagy (1988)

to construct energy budgets (Fig. 1). Similarly, RMRs combined with time–activity budgets can be used to estimate the energy allocated to maintenance or resting metabolism for a fixed interval of time.

The doubly-labeled water (DLW) method (Table I) is used to determine the total daily energy expenditure (DEE) (in kilojoules per day) of free-ranging animals under field conditions. The DLW method is the only energy measurement technique that includes the energy allocated to activity under natural conditions. Thus, the DLW technique can be used in combination with other techniques (Fig. 1) to examine the energy allocated to reproductive activities.

Energy budgets combine several techniques to partition total energy expenditure into compartments of maintenance, growth, reproduction, storage, and export (as milk). The construction of an energy budget requires several different techniques (Table I): (1) the DLW method to determine daily energy expenditure, (2) laboratory respirometry and time–activity budgets to calculate energy expended for maintenance, (3) bomb calorimetry to measure the energy content of eggs (for birds and oviparous reptiles), (4) quantification of energy accumulated prior to breeding (fat), and (5) the amount of energy exported as milk by mammals. If one can quantify the total energy expended by an animal, the energy expended for maintenance metabolism, the energy allocated to egg production in reptiles and birds and to embryonic growth and milk production in mammals, and energy allocated to growth, the energy that is allocated to activity can be estimated by subtraction. The energy expended for reproductive activities can then be estimated based on time–activity profiles and the amount of energy stored in fat or exported in milk.

A difficult part of examining reproductive energetics is determining the appropriate temporal schedule of reproduction. Although a few species are semelparous and reproduce only once before they die (e.g., salmon, spiders, many insects), this is not a typical reproductive pattern in vertebrate animals. Most species are iteroparous (reproducing more than once) and may reproduce once per year for several years in succession or even multiple times per year, whereas others may skip years. A reproductive period or season may be relatively short for some species (a few weeks or months), whereas other species may spend several months or years providing parental care to ensure the survival of their offspring (e.g., alligators, primates, elephants). In some iteroparous species, the amount of energy allocated to reproduction varies greatly with age. Thus, for some species, it may be necessary to investigate the reproductive energetics of a cohort of individuals over their entire life span.

4. REPRODUCTIVE ENERGETICS OF REPTILES

4.1 Territorial and Courtship Activities

There are many different types of mating strategies exhibited by reptiles, but most fall within two basic patterns of behavior: territorial (most lizards and crocodilians) and nonterritorial (most snakes and turtles). In territorial species, one or both sexes (usually males) actively defend a fixed area to monopolize the resources (food, shelter, or nest sites) and/or the mates (usually females) residing within it and to prevent intrusion or takeover by consexual competitors. Territoriality should evolve when the benefits (e.g., access to resources such as space, food, and females) exceed the costs incurred (e.g., energy expenditure, risk of injury). A large amount of energy is expended for territorial activities, including visual and/or chemical signals used to advertise the territory holder's presence and territorial boundaries, aggressive interactions and/or combat with consexuals, and territorial patrol to detect and repel intruders.

Nonterritorial reptiles may stay within a well-defined or loosely defined home range that is not defended, or they may roam widely in search of mates. Although nonterritorial reptiles often use chemical signals to find mates, they do not typically invest energy in the conspicuous and frequent visual signaling and patrol activities characteristic of territorial males. However, it is typical of both territorial and nonterritorial reptiles to invest energy in aggressive conflicts and combat with consexuals, especially if males are competing for a territory or a female.

During agonistic encounters, male lizards of many species typically perform head-bob or push-up displays accompanied by size-enhancing displays and body postures, mouth gapes that convey the threat to bite, and ritualized circling and jaw sparring with opponents. Males in some snake species also exhibit ritualized combat. Male rattlesnakes wrap around each other and raise their upper bodies off the ground, and then each attempts to wrestle and pin its opponent's head to the ground. Even some male turtles (e.g., snapping turtles, wood turtles, gopher tortoise) inhabiting the same area may fight each other to establish their position in a dominance hierarchy to gain the privilege of mating with females.

Most reptile species invest energy in some type of courtship behavior. Because females are typically choosier than males about their mating partners, it is usually the males of the species that do most of the courting in an attempt to persuade females to mate with them. Courtship in reptiles typically consists of species-specific visual displays (e.g., lizard head-bob displays, display of colored skin patches), acceptance or rejection behaviors (usually given by females), and/or tactile behaviors (e.g., licking, stroking, biting, chin rubbing) exchanged between the sexes. Reptiles do not exhibit the elaborate courtship rituals often observed in birds.

The amount of energy that reptiles allocate to copulatory activity varies greatly with the type of mating system. Some reptile species copulate only once or a few times during a breeding season, whereas others copulate repeatedly, especially polygynous males. For example, males of the green anole (*Anolis carolinensis*) expend energy to court and mate with an average of three females residing within their territory, mating repeatedly with each female roughly once each week for the duration of a 4-month breeding season. Because the purpose of territorial behavior is to exclude male competitors, territorial males do not usually guard their females directly after mating. Nonterritorial species may be polygynous, promiscuous (both sexes mate with multiple individuals), or even monogamous. In some nonterritorial species, males expend energy to guard their mates following copulation to prevent them from mating with other males. Guarding their mates may also preclude males from mating with other females (hence, they are monogamous), and a few species form long-term pair bonds that span multiple reproductive seasons (e.g., snow skinks [*Niveoscincus microlepidotus*], sleepy lizards [*Tiliqua rugosa*]). The duration of copulations, and thus the amount of energy allocated to copulation, also varies widely, even among closely related species. Some *Anolis* lizards have copulations lasting for several seconds, whereas males of *A. carolinensis* average approximately 25 to 30 min for a single copulation. In many vertebrate taxa, the duration and frequency of copulations are related to the intensity of sperm competition in a species, although studies addressing this issue in reptiles are currently lacking.

Four studies on lizards have quantified the energy expended for activities during the breeding season, and all of the species examined are territorial and polygynous (Table II). For a single reproductive season, the proportion of energy expended for activities during the breeding season ranged from 23 to 57% of total energy expended during that time. For males of *A. carolinensis,* nearly all of their activity is devoted to reproduction (e.g., only 1–2% of their day is spent foraging) and nearly half of their total energy expenditure is allocated to reproductive activity (Table II).

4.2 Gamete Production

Reptiles are similar to most vertebrate species in that males typically allocate little energy to producing sperm. The amount of energy that gonadal recrudescence and sperm production requires has not been

TABLE II

Comparison of the Energy Expended for a Single Reproductive Season in Several Lizard Species (All Figures in Kilojoules per Season Unless Otherwise Noted)

Species	Mating system[a]	Sex	Body mass (g)	Daily energy expenditure (kJ)	Eggs (kJ)	Reproduction (kJ)[b]	TME (kJ)[c]	TME(kJ/g$^{0.8}$)[d]
Anolis carolinensis	T, P	M	5.7	154	—	74	154	38.3
		F	3.3	105	59	109	164	63.1
Callisaurus draconoides	T, P	M	10.5	168	—	Not examined	168	25.6
		F	8.0	132	53	Not examined	185	35.1
Cnemidophorus tigris	NT, PR	M	18.8	507	—	Not examined	507	48.5
		F	15.8	361	52	Not examined	413	45.4
Iguana iguana	T, P	M	713	3320	—	1892	3320	17.3
		F	1004	4331	3847	6317	8178	32.5
Sceloporus jarrovi	T, P	M	23.9	174	—	40	174	13.7
		F	17.4	87	43	36	130	13.2
Sceloporus occidentalis	T, P	M	11.9	186	—	105	186	27.9
		F	11.9	200	32	151	232	32.3
Uta stansburiana	T, P	M	3.7	64	—	22	64	22.5
		F	2.4	32	18	50	45	24.8

Source. Modified from Orrell *et al*. (2004).

[a] Key to mating system: T, territorial; NT, nonterritorial; P, polygynous; PR, promiscuous.

[b] Reproduction: energy allocated to breeding season activity by males, and to breeding season activity + eggs for females.

[c] TME: Total metabolizable energy, calculated as daily energy expenditure during the breeding season + energy allocated to eggs.

[d] TME/body mass$^{0.8}$: TME adjusted for interspecific differences in body size, to facilitate interspecies comparisons.

quantified, but it is likely to be a very small amount relative to the energy expended for other aspects of reproduction. In contrast, nearly all female reptiles allocate a considerable amount of energy to producing large eggs that contain energy-rich yolk. The size and number of eggs produced by different species varies considerably. However, the amount of energy contained in egg materials is very similar among reptile species, averaging approximately 6 calories per milligram of dry egg mass. The amount of energy allocated to eggs by lizards ranges from 13 to 48% of total energy expended during the breeding season (Table II) and from 7 to 13% of total annual energy expenditure. Similarly, the painted turtle (*Chrysemys picta*) allocates approximately 14% of its annual energy to eggs.

4.3 Nest-Related Activities, Incubation, and Parental Care

Because most reptiles produce precocial and independent offspring that develop and hatch from shelled eggs (ovipary), reptiles rarely allocate energy to parental care activities. Instead, reptile species provide a form of nutritional parental care by provisioning their eggs with "extra" yolk that nourishes the hatchling for days or weeks after emergence from the egg. Female reptiles may also allocate energy to digging holes or nest cavities in which they lay eggs and then cover them with substrate (e.g., most turtles, many lizards). In crocodilians and some snake species, females construct a nest mound made from soil or vegetation in which to lay eggs, where decaying organic matter provides heat for optimal incubation temperatures. In species with relatively small eggs and/or clutches (e.g., *Anolis* lizards lay single-egg clutches), burying eggs may require a relatively minor amount of energy relative to other reproductive activities. In species with large eggs, large clutch sizes, and/or long travel distances to find nest sites (e.g., sea turtles travel many miles to lay eggs on natal beaches), the energy expended for nest-related activities may constitute a substantial energy investment.

In addition to nest construction, some reptile species guard their nest of eggs for all or part of their development (e.g., king cobras, crocodilians, pythons, some lizards). Female alligators construct and guard their nest of eggs and then protect their young for up to 2 years. Female Burmese pythons (*Python molurus*) actually incubate their eggs by coiling around them and rhythmically contracting their body muscles, generating metabolic heat to maintain the optimal incubation temperature of 30°C. The increase in energy expenditure associated with incubation for female pythons is 2 to 20 times higher than that for females that are not incubating, with an inverse relationship between ambient temperature and the rate of energy expended; that is, it takes more energy to keep the eggs warm at cooler ambient temperature.

4.4 Egg Retention and Gestation

Most reptiles (chelonians, crocodilians, tuataras, and most lizards and snakes) are oviparous. They lay eggs with relatively undeveloped embryos and a large yolk mass containing enough energy to support embryonic development. For the most part, oviparity excludes the energetic costs of gestation. However, as many as one-fifth of squamate species (snakes, lizards, and amphisbaenids) exhibit various degrees of viviparity that require some amount of energy expended for gestation. At one end of the ovipary/vivipary continuum, there are many oviparous squamates that retain developing eggs in the oviducts for some length of time. Further along this continuum, ovoviviparous females provide eggs with yolk for embryo development, but eggs are enclosed by a noncalcified shell or membrane and remain in the oviduct until completely developed (e.g., *Boa constrictor*). Various intermediate states include some species providing both a yolk mass and some type of placental nutrition of embryos (e.g., garter snake [*Thamnophis sirtalis*]) or a highly vascularized oviduct (e.g., common lizard [*Lacerta vivipara*]). At the other end of the continuum, a few species exhibit extremely reduced yolk mass, and offspring are attached and nourished by a placenta formed from uterine tissue and extra-embryonic membranes (euvivipary) (e.g., skinks [*Mabuya heathi* and *Chalcides chalcides*]).

In addition to nutrition that may be provided directly to offspring, there are a number of energetic costs for egg retention and embryo gestation that do not occur with strictly oviparous species. Egg-retaining/Gestating females must expend energy to behaviorally thermoregulate and maintain optimal body temperatures for developing embryos. In addition, if egg-retaining/gestating females maintain higher than normal body temperatures (and studies show that many do), this in turn increases their own temperature-dependent maintenance metabolism. Energy is also expended to carry around the bulk and mass of the eggs and/or developing embryos. Other types of costs for egg retention and embryo gestation include decreased mobility and greater

susceptibility to predators, smaller offspring and/or clutch sizes, and lower rate of offspring production. Given all of the potential costs, there must be tremendous selective pressure on some species to evolve toward some degree of viviparity. The potential benefits of egg retention or gestation include optimal temperatures for embryonic development, delayed oviposition (or birth of live offspring) when environmental conditions (e.g., moisture, temperature) are unfavorable to embryonic development within eggs, and protection of eggs from predators, bacteria, and fungi. There is a strong trend for egg retention and viviparity to occur in squamates at high elevations, where extremes in temperature, humidity, and low atmospheric oxygen concentration inhibit or preclude embryonic development if eggs were subject to these conditions. Clearly, egg retention and various forms of gestation have both costs and selective benefits in terms of energy allocation and mortality risk versus increased offspring survival and reproductive success, making the evolution of egg retention and viviparity an example of another life history trade-off.

5. REPRODUCTIVE ENERGETICS OF BIRDS

5.1 Territoriality, Courtship, and Mating

The vast majority of bird species (>90%) have a monogamous mating system in which individuals (typically males) compete with each other for access to mates and resources critical to mating. When mates or resources are economically defendable, individuals monopolize mates and/or resources through territorial behavior. The area defended may be large and contain mates or resources important to reproduction (e.g., food, mating, nest sites), or it may be small and encompass only the space immediately surrounding the nest to protect the offspring, nest, nest site, and/or nesting materials (e.g., colonial nesting species). Song is one of the most conspicuous and energetically demanding aspects of territorial behavior of birds. Birds (mostly males) use species-specific songs to advertise their presence in a territory, to repel potential intruders, to find and attract potential mates, and to stimulate female reproductive physiology and behavior. Studies have shown that normal males rendered mute experimentally are far less effective at repelling intruders and/or maintaining their territory and mates than are vocal males. Laboratory respirometry was used to determine the energy expended for singing in captive male Carolina wrens (*Thryothorus ludovicianus*). Their metabolic rates while singing were 2.7 to 8.7 times higher than their RMRs. Another example of energetically demanding territorial behavior is the rapid and loud drumming of woodpeckers (males rapidly tap their beaks on hollow trees or other resonant structures) and grouse (males beat their wings in the air or on their breasts while standing on logs, boulders, or tree roots).

In addition to advertising a male's presence in a territory, the frequency and/or intensity of bird song or drumming may indicate male body condition and/or energy reserves and may be a signal used by males to assess competitive ability or used by females to make mate choice decisions. In addition to singing and drumming, some birds expend energy for territorial patrol of their territory to detect and repel conspecific and/or heterospecific intruders, using visual threat displays and/or chasing to protect resources and/or offspring. When threat displays and chasing fail to oust intruders, escalated aggression can involve direct combat and the risk of physical injury. It has been estimated that some males expend up to a third of their daily energy expenditure, or up to 2.4 times the cost of BMR (in house martins [*Delichon urbica*]), for territorial activity. Perhaps because territorial advertisement is energetically costly and attracts the attention of predators, females of many bird species do not sing or call for territorial reasons or they do so much less frequently than do conspecific males. Although females do not advertise their territory, they too may chase off conspecific females to protect their offspring or resources important to reproduction.

As a result of competition among males for mates, sexual selection results in male birds expending energy for conspicuous traits and behaviors used to influence female mate choice. Males of many bird species exhibit colorful and/or modified feathers and/or structures (e.g., "red-wing" blackbirds, yellow- or orange-crested cockatoos, multicolored macaws, greatly elongated and colorful tail feathers of peacocks, wattles and combs of roosters). The conspicuous ornaments of male birds may require a minor amount of energy to produce relative to total energy expended for reproduction, but these traits may also be costly to males because they attract the attention of predators or reduce a male's ability to evade predators (e.g., long tail feathers). In addition, the quality and/or size of these traits may reflect the energy stores, body conditions, and general health of individuals and could be an important signal to male competitors and

potential mates. For instance, bright and intensely colored traits could indicate that males are free of parasites and/or disease that could otherwise compromise their ability to defend their territory or that could be transmitted to mates or offspring. Quality of male ornaments may also indicate the genetic quality of individuals, that is, parasite and/or disease resistance that will be passed to offspring or inherited traits preferred by females that confer reproductive success to those males and their future offspring.

Males expend a considerable amount of energy to show off their conspicuous ornaments, using complex and ritualized behaviors or activities to court females. Courtship displays may consist of aerial dances or ritualized sequences of behaviors performed on the ground, in trees, or on the water (e.g., head bows; fluffed feathers; erected crests; fanned, waved, or flapped wings; choreographed hops or steps), and all of these displays may be combined with vocalizations. Male bowerbirds build elaborate bowers out of vegetation, often decorating them with feathers, paper, string, or other colored objects, and then use the bowers as performance stages or backgrounds for their ritualized courtship dances. In some bird species, several males congregate in "leks" and simultaneously display and exhibit their ornaments to observing females, who evaluate them and their performances before choosing their mates. Male jays, cardinals, terns, and other species also expend energy to forage for food items presented to prospective or current mates as nuptial gifts. Courtship displays and activities often require a large investment of energy to perform, especially if performed repeatedly. The daily energy expended by male sage grouse (*Centrocercus urophasianus*) performing strut displays toward females may be twice that of nondisplaying males and positively correlated with the number of struts performed per day. In many species, the frequency or intensity of courtship behaviors is correlated with the displayer's condition (i.e., sufficient energy intake to support the activity and maintain body condition) and/or with female mating preferences. Researchers have found that female sage grouse that choose among displaying males on a lek preferred those with the highest strut rates, and a positive correlation between male foraging distance and strut rate suggests that display vigor indicates male endurance and foraging skill.

Female birds rarely exhibit the bright colors, elaborate structures, or ritualized courtship behaviors and activities observed in conspecific males, possibly because these traits are selected against in females and are costly in terms of energy expenditure and susceptibility to predators. However, in some cases, females participate in male courtship rituals, especially when they are receptive to the courting males.

Although a single copulation is generally very brief in birds, the energy expended for copulatory behavior varies considerably. Some species copulate only a few times during the breeding season, whereas other species copulate hundreds of times. The frequency of copulations and the energy expended in this activity are related to sperm competition and paternity. Although the vast majority of bird species are socially monogamous, extra-pair copulations are common in most species. In general, the less assured a male is of paternity, the more frequently he will copulate with a female. For species in which a male guards the female before and after copulation and until her eggs are laid, the male has a reasonable assurance that the female has not mated with other males and that only one or a few copulations are required for him to fertilize her eggs. For a male that leaves his mate for extended periods of time, either to forage for food or to copulate with other females (a polygynous male or a monogamous male seeking his own extra-pair copulations), it is possible (or in some cases probable) that his female has copulated with one or more other males. Because males in many bird species invest a large amount of energy in parental care to raise offspring, it is crucial to a male's reproductive success to ensure that his sperm fertilizes his mate's eggs. Frequent copulations with a female may increase a male's paternity by inseminating a greater quantity of sperm, and/or by providing fresh sperm, to compete with sperm inseminated by other males. Studies using DNA markers have revealed that paternity in birds is influenced by copulation frequency, sperm number, fertilizing capacity of sperm, and the timing of insemination relative to other inseminations and to oviposition. Frequent copulations require greater sperm production, and many studies on birds have documented that large testes produce greater numbers of sperm than do smaller testes. Therefore, it is not surprising that copulation frequency is positively correlated with testis size in many vertebrate taxa, including birds. Because large testes may evolve in response to sperm competition, relative testis size provides a measure of the intensity of sperm competition in various species.

5.2 Gamete Production

Because most birds are seasonal breeders that produce gametes only during the reproductive season, energy

is allocated to gonadal recrudescence prior to the breeding season for both sexes. However, the increase in testes size equates to roughly 2% of the cost of male BMR, and the increase in ovary and oviduct mass constitute 2 to 9% of the cost of female BMR. Thus, females expend more energy than males for gonadal recrudescence, but for both sexes it is a relatively trivial amount compared with the amount of energy expended for other aspects of reproduction.

Although the amount of energy allocated to sperm production varies depending on copulation frequency, males still allocate a very small amount of energy to sperm and ejaculate production. For example, the energy that male chickens require to produce semen accounts for less than 1% of the amount expended for BMR. In contrast, females allocate a tremendous amount of energy to egg production. The total amount of energy allocated to egg production in a reproductive season depends on the number of eggs produced, egg size, and the amount of energy contained in the eggs, and these factors are influenced by female body size, energy availability, and the particular life history traits of the species. Although the total amount of energy allocated to egg production varies among individuals and species, the amount of energy contained in the eggs themselves generally is proportional to the developmental stage of the offspring at hatching (Fig. 2). The eggs of birds that produce altricial offspring tend to contain a relatively small proportion of yolk (24% of wet egg mass) and less energy (4.7 kJ/g of wet mass) in comparison with most precocial species (40% of wet egg mass). The eggs of extremely precocial species (e.g., Australian megapodes such as the brush turkey [*Alectura lathami*]) contain the greatest proportion of yolk (up to 65% of wet mass) and the greatest amount of energy (12.3 kJ/g of wet mass). Altricial offspring also tend to have the least amount of energy remaining in yolk sacs after hatching (used for development and growth [average 8% of hatchling body mass]), whereas precocial species tend to have a greater amount of energy left in yolk sacs after hatching (average 15% of hatchling body mass). The total energy allocated to egg production in various bird groups has been estimated at 40 to 170% of the energy expended for BMR.

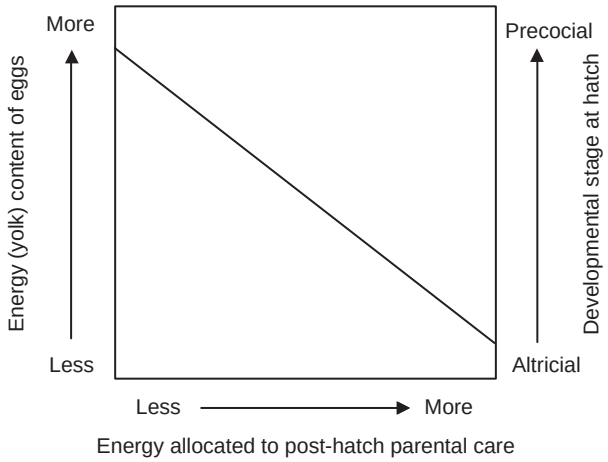


FIGURE 2 General trends in the relationships between the energy content of eggs, the energy allocated to post-hatch parental care, and the developmental stage of avian offspring at hatch.

5.3 Nest Building and Incubation

Because all birds are oviparous, nearly all species expend energy to collect nest materials, construct or excavate nests, and incubate eggs. Exceptions to this generalization include several species of parasitic birds (e.g., cowbirds) that lay their eggs in the nests of other species (or, more rarely, in the nests of conspecifics), leaving eggs to be incubated and cared for by the nest owners. Although birds invest a substantial amount of time and effort in constructing nests, the energetic cost of this effort has not been examined.

Egg incubation in birds consists of maintaining a clutch of eggs at the optimal temperature for embryonic development until they are hatched. Most birds use their own body heat to maintain egg temperatures (except megapodes, which bury and tend eggs in mounds of organic matter and depend on external sources of heat such as solar radiation, plant decomposition, and geothermal heat). The transfer of body heat to eggs and the monitoring of egg temperature are facilitated by a highly vascularized and bare patch of skin called a brood patch. In many monogamous species, each sex takes its turn incubating the eggs while the other mate forages for food. In some species, the female continually incubates the eggs while her mate provides her with food. Egg incubation may also be performed by only one parent (usually the females, although a polyandrous male may also incubate), but that parent must either forgo feeding during incubation or occasionally leave the eggs to find food. When environmental temperatures are near the thermoneutral zone of a species, the eggs may require little incubation time or metabolic heat to maintain their optimal temperature. When environmental

temperatures drop below the incubating bird's thermoneutral zone, birds may have to raise their metabolism to provide enough heat to maintain egg temperature, thereby expending energy specifically for incubation. Studies indicate that incubating birds have RMRs that are 19 to 50% higher than those of nonincubating birds at the same environmental temperatures.

5.4 Parental Care

Birds invest a considerable amount of time and energy in caring for their offspring and do so in a number of different ways. Hatchlings must typically be fed, so one or both parents must find or capture enough food to support the maintenance and growth of their offspring and to support their own maintenance and reproductive activities. Parents may also process food prior to feeding it to offspring; many birds partially digest and regurgitate food, and pigeons and doves of both sexes produce crop milk secretions to nourish hatchlings. Parental care also includes brooding offspring, that is, protecting them from extremes in temperature or other types of inclement weather. Brooding is essential to offspring survival because most chicks are poikilothermic (body temperature fluctuates with environmental temperatures) for days or weeks after they hatch because they do not have the muscle mass to produce sufficient metabolic heat for homeothermy (constant body temperature) and/or because they lack insulating down or feathers to retain body heat that is produced. Thus, species that reproduce in cold regions (e.g., penguins) invest more time and energy in brooding their young than do those that inhabit warmer tropical regions. Birds often protect offspring from predators, for example, by using alarm calls to warn offspring about the presence of predators, attacking and mobbing predators to chase them away, or luring predators away by faking injuries (e.g., the broken wing act used by killdeer [*Charadrius vociferus*]). Some bird species also teach their offspring survival skills such as leading them to food sites and demonstrating how to find and capture prey and how to evade predators.

The amount of time and energy that birds expend for parental care depends on many factors, including (but not limited to) the developmental stage of hatchlings, the number of offspring in a clutch, and the abundance and distribution of food resources. To some extent, the amount of parental care provided is inversely proportional to the hatchling developmental stage; the more developed the offspring are at hatching, the less time and energy are required for parental care. Thus, the amount of time and energy required for parental care tends to be counterbalanced by the energy invested in eggs (Fig. 2). The amount of time and energy required to raise offspring is a strong limiting factor for the number of offspring that can be raised successfully. Experiments have found that artificially increasing brood size in house martins (*D. urbica*) decreases both the size and number of offspring fledged and proportionally increases the daily energy expenditure of savannah sparrows (*Passerculus ssnwichensis*). An increase in energy expenditure in savannah sparrows was correlated with both brood size and the number of feeding visits to the nest. Similar experiments that increased brood size in Eurasian kestrels (*Falco tinnunculus*) found that males (which provide most of the food to offspring) were able to increase the number of voles caught to support the entire brood. However, when prey abundance was artificially reduced, male kestrels compensated by expending more energy to bring back food. When prey abundance was decreased and brood size was increased at the same time, kestrel offspring grew more slowly and had higher mortality rates, and the parents lost body mass and had increased mortality. The amount of energy that birds expend while raising broods has been determined in several species to be about three times their basal rate of metabolism.

In some species of birds, the energetic expense associated with parental care may be alleviated in part by conspecific "helpers" that provide alloparental care or assistance in raising and/or protecting the offspring of other birds. Helpers are usually related to the breeding birds. Most often, young adults from previous broods help to raise their younger siblings. Helpers usually delay their own reproduction due to limited breeding resources (e.g., territory space, nest sites, food). The breeding birds benefit from the assistance of helpers by successfully producing a greater number of offspring than they could produce on their own. In one study, the mortality of offspring in the pied kingfisher (*Ceryle rudis*) was 61% when parents were unassisted, decreased to 22% when there was one helper, and decreased to nearly 0% when there were two helpers. Helpers also benefit by gaining skill and experience in providing parental care, making them more successful at raising their own offspring later. In addition, by helping to raise related individuals, they increase their genetic contribution to the next generation.

6. REPRODUCTIVE ENERGETICS OF MAMMALS

6.1 Territoriality, Courtship, and Mating

Most mammal species (~95%) have a polygynous mating system in which males typically compete with each other for access to mates and resources. Males of polygynous species defend territories (or female groups) so as to acquire multiple mates, and the energy investment in this activity can be substantial. Males may spend a considerable amount of time patrolling boundaries of their space, vocalizing, scent marking, and making visual displays. In some instances, the displays require a relatively small investment in energy because once a territory is established, neighbors are conditioned and need only occasional reminders to keep out. Still, economic defendability is the most important determinant of when a territory or group of females should be defended. For example, male arctic ground squirrels (*Spermophilus parryii*) defend territories against other reproductively active males, and paternity analysis has revealed that the first male to mate with an estrous female sires her offspring. Male short-nosed fruit bats (*Cynopterus sphinx*) construct tents in which they roost by modifying leaves and other plant parts and then defend these roosts against intrusions of conspecific males. Different numbers of females join these males, but it is unknown whether females choose the males based on quality of the resource (tent) or on some quality of the males. Still, paternity analysis has revealed that females sometimes move among adjacent tents; thus, the resident males might not always sire the pups produced by the female occupants of their tents. If males defend females instead of some habitat, this is referred to as female defense polygyny. However, as in other mammal species, it is not always clear whether a male defends a territory or defends the female that occupies it.

Three general types of polygynous mating systems occur in mammals: resource defense polygyny, female defense polygyny, and leks. Resource defense polygyny occurs when males defend areas that contain the feeding, nesting, or roosting sites that are critical for reproduction. Female choice is generally influenced by the quality of the male and the area that he defends. Female defense polygyny commonly occurs in situations where females are gregarious for reasons unrelated to reproduction, for example, when females form herds for protection from predators (e.g., musk oxen [*Ovibos moschatus*]) or when they assemble near food, nesting, or roosting sites (e.g., greater spear-nosed bat [*Phyllostomus hastatus*]). In the latter situation, females become gregarious because roost sites in caves are limited and males then monopolize the females for breeding. Polygynous mating systems usually lead to intense competition among males, and this can result in marked sexual dimorphism and large variations in reproductive success. In the greater spear-nosed bat, resident males sire from 60 to 90% of the pups born to the females in a harem.

The sac-winged bat, *Saccopteryx bilineata,* is characterized by a female defense mating system in which males defend females in a harem territory. Courtship and harem defense in this species includes energetically costly flight maneuvers and hovering displays. A doubly labeled water experiment revealed that the mass-specific field metabolic rate of harem males increased with harem size. Males typically scent mark the boundaries of their territories and defend these areas against intrusions by other males. Adult males attempt to attract females into their harem territories using acoustic, visual, and olfactory displays. In daytime roosts, which are often located in well-illuminated buttress cavities of large trees around buildings in the Neotropics, females are aggressive toward one another as well as toward males. One of the most distinguishing displays of this species is a hovering flight during which odor is fanned toward roosting females. Males perform hovering displays throughout the year, although copulations occur only during 2 months of the year. Males that roost on the periphery of harems father some of the offspring, but harem males on average are the most successful. Thus, although harem males are not the exclusive fathers of the harem offspring, male reproductive success is on average correlated with harem size.

Intense competition among males can also be observed in a lek mating system, where males congregate and display communally to attract females. The only resource that a male defends on a lek is the space where courtship and mating occurs. Males arrive early in the breeding season and, through highly ritualized combat behaviors, compete for and establish their territories. Males of lekking mammals have evolved elaborate ornaments such as horns, antlers, and vocalizations that they use in intermale competition to attract females and to influence female choice. Females move through the arena (lek) in which males are displaying and mate with one or more males based on the quality of their leks (and resources within it) or on the relative qualities of the

males. Lek mating systems have been observed in at least five species of African ungulates, including the Uganda kob (*Kobus kob*) and white-eared kob (*Kobus kob leucotis*), and four species of fruit bats, including the hammerheaded bat (*Hypsignathus monstrosus*) and Wahlberg's epauletted bat (*Epomophorus wahlbergi*). These and other lek mating species occur when males are unable to economically defend either the females or the resources they require. Such a situation can arise when females exploit widely dispersed resources such as grasses (ungulates) and fruit (bats). At high female density, males of the Uganda kob form leks and display to females, but they defend harems (resource-based territories) at low density when defense is presumably economically feasible. Male hammerheaded bats display from calling roosts located in gallery forests along rivers. They emit series of sharp "honks," and when females approach, they vigorously shake their wings and increase the rate of vocalizations. Once a female chooses a particular male, they mate and the male continues to call. As in other polygynous mating systems, some males have greater success in attracting females than do others. In one study, 6% of male hammerheaded bats achieved approximately 79% of the matings in a single year. Courtship displays in species that form leks are fueled by body reserves that allow high instantaneous increases in metabolic rates (up to 20 times BMR). The short-term increase in metabolism that males incur during displays contrasts with the long-term or sustainable metabolic scope that reflects the upper limit of metabolism over a longer period of time.

Polygyny also is common among carnivores where dominant males typically roam over larger areas and seek females in estrous (e.g., felids). Although energetically costly, this strategy appears to contribute to their high reproductive success. A roaming strategy is energetically more profitable for a dominant male because the probability of taking over other females is related to his social status. When females are concentrated in patches, a dominant male can increase the number of potential matings by including a second patch within his range.

Mating systems among canids vary in relation to body size, but the pervasive mating system is monogamy. Smaller species such as red foxes (*Vulpes vulpes*) and bat-eared foxes (*Otocyon megalotis*) are usually monogamous but are sometimes polygynous. The largest canids, such as the African hunting dog (*Lycaon pictus*), are generally monogamous but with a tendency toward polyandry (several females monopolize a single male). In some situations, a polygynous mating system may change depending on the density and dispersion of females, or it may change from polygamy to monogamy depending on the abundance of (and ability to monopolize) females. When mating is restricted to only one season of the year, the preferred resources for males may differ during the mating and nonmating seasons because the two resources (receptive females and available food) tend to have different characteristics. Where males cannot monopolize feeding, nesting, or roosting resources, monogamy becomes possible. Where habitats are characterized by having clumped resources that can be defended economically, polygyny tends to be favored.

In general, variation in mating behavior in mammals can be attributed to four characteristics of females: (1) the extent to which female reproductive rate can be increased by assistance with offspring, (2) the size of female ranges, (3) the size and stability of female groups, and (4) the density and distribution of females in space. A male is better able to monopolize several females when male care of offspring is not needed and when females have small ranges or when they occur in small stable groups. Alternatively, when females are solitary in large ranges or occur in large unstable groups, or when female density is high, it is uneconomical for males to defend several females at once.

6.2 Gamete Production

As in most other vertebrates, mammals are seasonal breeders and produce gametes only during the reproductive season. Energy is generally allocated to gonadal development prior to the breeding season in both sexes. In comparison with female mammals, males allocate a relatively small amount of energy to the production of sperm and ejaculate. The size of testes, expressed relative to body mass, is a standard method for comparing the investment among different species of mammals and birds. The variation in relative testes size in mammals ranges from 0.02 to 0.75% in primates, from 0.15 to 8.40% in rodents, and from 0.12 to 8.40% in bats. Among terrestrial mammal species that have been studied, bats have the widest range of variation in relative testes size. In fact, the relative testes size of the insectivorous Rafinesque's long-eared bat, *Corynorhinus rafinesquii,* exceeds that of all other mammal species. Still, the investment made by both sexes is a relatively trivial energy investment compared with the energy allocation to other aspects of reproduction and to the

investment that oviparous (egg-laying) vertebrates make in reproduction.

Sexual selection by sperm competition has been postulated as the mechanism that contributes to the wide range and maximal size of testes in mammals. Species in which multiple males mate with multiple females (promiscuity) have the largest testes, whereas the smallest testes occur in species where single males mate with single females (monogamy) or multiple females (polygyny). Thus, the mating system appears to be an important variable that explains testes size in mammals. In bats, sperm-storing species also have larger relative testes size than do non-sperm-storing storing species. Among rodents that breed two or more times per year, males allocate more resources to sperm production.

6.3 Nest Construction/Excavation and Nest Guarding

Nest construction varies widely among terrestrial mammals, with females in colder climates commonly investing more in nest construction than those in moderate to warm regions. Females of many small terrestrial mammals (e.g., shrews, rodents, lagomorphs) frequently excavate burrows or runways near the surface or create deeper cavities in the ground and construct nests made of vegetation (leaves, grass, and tree branches) where young are cared for until they are weaned. The excavated burrows and nest cavities provide protection from many would-be predators and provide insulation for the occupants, especially altricial young and their mothers. Insulation provided by nest material helps to reduce maintenance costs (for thermoregulation) of both the mother and her young, making it possible for the female to allocate more energy to reproduction. Medium-size terrestrial mammals (e.g., foxes, woodchucks, badgers) also construct burrows and nest spaces below the ground, but because they have a greater body mass and thus thermal inertia, they do not need to invest as much time and energy in actual nest construction. The Canadian beaver (*Castor canadensis*), a medium-size monogamous mammal, invests extensively in nest construction, which typically involves constructing lodges from vegetation (tree branches and grasses). Both males and females engage in nest-building activity. Before construction of a lodge, which includes the natal nest, beavers may first construct dams to increase the depth of the water so that the only access to the lodge is from an opening beneath the water. Such lodges, with additional nest material, provide not only insulation but also protection from predators.

With the exception of large arboreal primates (e.g., chimpanzees, orangutans), most large mammals and bats do not construct nests. Instead, large mammals seek shelter from inclement weather by forming tightly spaced social groups (e.g., musk oxen, elephants) or seek shelter beneath trees (e.g., many bovids and cervids) or under rock overhangs in treeless habitats (e.g., asses, mountain sheep). Because most large mammals produce precocial offspring, extensive natal nests are not needed. Although bat species are altricial at birth, the parents do not construct shelters or nests but rather use existing structures such as tree cavities, caves, mines, buildings and similar spaces, and roost beneath foliage. In contrast to some mammals (mostly rodents) that seek shelter in similar places and use nesting material from grass and leaves, bats do not use any such material. Sites selected by female bats for rearing young are often physically confined spaces (cavities in trees and caves) that are warmer than alternative locations and, thus, provide conditions that reduce maintenance costs and promote milk production by females and growth of their young. Because females often return to these same roosts year after year, little energy is invested in finding these sites. Energetic benefits accrue to both the mothers and their young by occupying shelters in which metabolic heat can be trapped, thereby reducing heat loss to an otherwise cool environment. However, a few species of bats construct shelters by modifying leaves, other plant parts (roosts and vines), bird nests, and nests of ants and termites that, in turn, often become occupied by females to give birth to young. These so-called "tents" are constructed by single males that recruit females with whom they have a high probability of mating. Because these shelters are constructed by males that are chosen by females to rear their young, tent construction can be considered as a form of paternal investment.

Natal nests of many mammals are often strongly defended from predators and conspecific intruders by both males and females. Many colonial species (e.g., prairie dogs, marmots, bats) use a variety of signals to alert nest or group members to threats from such predators, and these may include vocalizations (e.g., chirps, howls), visual signals (e.g., flashes of white or black tails by prairie dogs, wing displays by bats), and olfactory signals (e.g., production of scent from specialized glands, secretion into the urine). Some rodents secrete substances into their urine when they are stressed, and this acts as a pheromone that may cause others to escape or avoid an area. Skunks and

other mustelids are well known for the use of visual and scent displays that they use when threatened by conspecifics and would-be predators.

6.4 Gestation and Size at Birth

Gestation in mammals occurs in the uterus; however, the development of the uterus and its role in gestation differs considerably among different mammalian taxa. In placental (eutherian) mammals, an intimate physical, nutritive, and endocrine association develops between the fetus and the uterus via a chorioallantoic placenta, which nourishes the fetus for the duration of gestation and helps to regulate the duration of gestation. In contrast, the uterus of marsupial (metatherian) mammals and associated choriovitelline (yolk sac) placenta plays a relatively minor role in embryonic development. No intimate physical or endocrine association is established, and the young are born in an extremely altricial state. However, in either case, the requirements for gestation depend on the amount of energy and matter retained by the gravid uterus and the enlarging mammary glands and the increased maintenance costs of these structures. Altricial marsupials are instead nourished throughout most of their "fetal" period from highly specialized mammary glands and nipples.

Embryonic development incurs the greatest cost to the female during gestation. In eutherian mammals, fetal growth is curvilinear, with most of the increase in embryonic mass occurring after 50 to 60% of the gestation period has elapsed. In general, the length of the gestation period and the birth mass of eutherian neonates increase as a power of maternal mass, with smaller species having shorter gestation periods than larger ones and with smaller species having smaller offspring than larger ones. However, carnivores, primates, and most bats have longer gestation periods than can be predicted from maternal body mass, whereas marsupials and red and giant pandas have shorter gestation periods than can be predicted from maternal mass. Exceptions to predictions in length of gestation occur predominantly among marsupials that have a choriovitalline placenta and have unusually short gestation periods and among eutherian mammals that undergo delayed embryonic development (e.g., mustelids, pinnipeds, some bats). For example, the gestation length of an 18-g rodent is approximately 21 days, whereas the gestation period of comparable size bat species can range upward from 2 to 6 months. Bats and bears best exemplify exceptions to the predicted trend in size at birth among eutherian mammals. Newborn little brown bats (*Myotis lucifugus*) are larger at birth than predicted, weighing about 2 g and representing nearly one-fourth of the mother's body mass. By contrast, newborn black bears (*Ursus americanus*) are smaller than predicted, ranging in mass from 200 to 450 g and averaging 1/250th of the mother's mass.

In general, birth mass is a reflection of the metabolic body mass of the mother, with larger species producing offspring that represent a smaller proportion of adult mass. The energy cost of gestation can be estimated using birth mass to maternal mass relationships, a maternal mass to gestation period regression, the composition and growth characteristics of the gravid uterus, and an estimate of the oxygen consumption of the gravid uterus per kilogram of tissue. Several studies have shown that the maintenance cost of the gravid uterus is the major cost of gestation in eutherian mammals. In general, gestation is far less costly and stressful in larger mammals than in smaller mammals because the relatively smaller neonate of larger species is produced over a much longer period. However, when one compares the total cost (maintenance and production) of a gravid uterus with that of the gestation period, the cost is proportional to maternal metabolic body mass ($M^{0.77}$). In eutherians, protein accumulates curvilinearly in the gravid uterus and mammary glands during gestation. In general, because of longer gestation periods in larger species, protein accumulates at a decreasing rate per kilogram of fetus. Changes in fetal size of gestation length are far more complex than simply comparing relative growth rates. Pregnant females can benefit by lengthening the gestation period if protein (or minerals) is limiting, but they gain nothing by prolonging the gestation period if energy is limiting. Some mammals ingest nutrients in excess of the actual costs of gestation, and because lactation is more costly than gestation, females of larger species (but not small species) can accumulate reserves during gestation when food is available and can mobilize it later for milk synthesis. Mammary glands begin to enlarge during the final trimester of gestation and, thus, can be considered as a cost of pregnancy and lactation.

6.5 Lactation

The production and secretion of milk is a trait that distinguishes mammals from all other vertebrates, although some birds (e.g., pigeons, doves, penguins, flamingos) feed crop secretions, or "crop milk," to

hatchlings during their first few days of life. Among mammals, the milk produced by a postpartum female provides an important dietary buffer between the dependent fetus and the weaned and nutritionally dependent young. Milk is produced by mammary glands that vary in position from the thoracic to the inguinal region of the body. Mammary glands generally vary from a single pectoral pair in primates, elephants, and bats up to 12 pairs located in the abdominal and inguinal regions in some rodents and marsupials. The young of placental and marsupial mammals obtain milk from nipples by a sucking action that stimulates milk letdown (the neural–hormonal release of milk from the mammary glands). Nipples are absent in monotremes (egg-laying mammals), and milk is secreted from mammary glands directly onto the ventral surface of the skin of females so that the beaked young can lap up the milk with their tongues.

Female mammals typically provision their offspring with milk for varying amounts of time. Factors known to influence the length of lactation is adult body size relative maturity of their young at birth (altricial or precocial) as well as the quantity and quality of milk. The quantity and quality (energy and nutrient content) of milk can vary according to a particular species, the type of food that is ingested by the mother, the availability of water, and the type of habitat in which they live. Terrestrial mammals that have a high-fat diet (those that live in arid regions) often produce milk with high lipid content. In contrast, mammals that feed on fruit and nectar or fleshy aquatic vegetation generally produce milk that is high in water content and carbohydrates but low in protein. Carnivorous and insectivorous species more commonly produce milk that is high in protein and fat but low in carbohydrates. The milk composition may also change over the course of lactation in many mammals. The first milk produced during a lactation cycle is called colostrum. Colostrum has a high concentration of antibodies and active phagocytic cells, each of which are thought to be important for the development of an effective immune system in neonates.

The major nutrient constituents of milk are water, minerals, protein (casein and whey), fat, and carbohydrates. Protein concentration ranges from less than 2% in primates to nearly 16% in the eastern cottontail (*Sylvilagus floridanus*). Fat content in the milk of terrestrial mammals also varies among species, ranging from a mean of 2.3% in lemurs (*Lemur* sp.) to 18.5% in the brown bear (*Ursus arctos*). The milk composition of bats also appears to be influenced by diet. For example, the milk composition of insectivorous species generally has a greater amount of dry matter, fat, and protein content than that of omnivorous and frugivorous species. The relatively low fat, protein, and dry matter content of milk of frugivorous species is consistent with the relatively low fat and protein content of available fruit. Insectivorous species produce milk with percentages of fat and protein that often exceed values reported for other small mammals, an observation that is consistent with evidence that small mammals produce milk of higher energy content than do most large mammals.

Some females suckle their young on demand (e.g., marsupials, primates, some ungulates, some bats), whereas others suckle their young on a schedule (e.g., lagomorphs, most rodents, most carnivores, many artiodactyls, some bats). As their young increase in size, so do the energy and nutrient demands on the mothers to produce more milk. Thus, the composition of milk often changes over the course of lactation. For example, in the Brazilian free-tailed bat (*Tadarida brasiliensis*), fat content of milk changes from a mean of 17.3 to 25.8% from early to peak lactation. Large mammals such as carnivores, bovids, and cervids also deposit fat reserves during pregnancy and mobilize them during lactation to either support the production of milk or support the increased demands for energy of the mother to secure food. During peak lactation, when energy demands are the highest, females may mobilize their other body reserves to supplement daily food intake. This is especially the case during periods of food shortages but also in small mammals that do not have the capacity to store large fat reserves.

As energy demands increase due to lactation, maintenance costs also increase because more time is invested in locomotor activity to secure food. The amount of time and energy that mammals allocate to foraging can increase more than twofold compared with that allocated to pregnancy. For example, female little brown bats (*M. lucifugus*) and Brazilian free-tailed bats (*T. brasiliensis*) more than double their nightly intake of insects from early to peak lactation. As lactation progresses, the energy cost of milk synthesis increases, corresponding to the increased energy demands of the growing young. Milk production generally increases until peak lactation and then begins to decrease as infants become more independent and begin to feed on other foods.

Estimates of milk intake by maternally raised young, when combined with estimates of milk composition, provide data necessary for determining

requirements for maternal production and requirements for neonatal growth. Methods used for estimating milk intake by suckling young include weighing the neonate before and after a controlled nursing period, hand milking a lactating female after a controlled period of neonate removal, and using the dilution of isotopes (usually deuterium or tritium oxide) to quantify the dilution of the body water pool by ingested milk. The isotope dilution method offers the only available method for estimating milk intake in free-ranging mammals. Thus, milk intake can be directly estimated from measurements of neonatal water kinetics. Because milk is the only source of water to a suckling before it takes solid food, measuring the daily water flux of such a neonate using isotope dilution and estimating the proximate composition of milk (energy content of the dry matter) makes it possible to estimate the amount of energy that a mother invests in her offspring in a given 24-h period or over the course of lactation. The energy expenditure of free-ranging female mammals during lactation can be quantified using a combination of time budget analysis and laboratory respirometry or from estimates of energy expenditure using the doubly labeled water technique.

Energy requirements for pregnancy and lactation are proportional to the productive process involved—synthesis of fetal, placental, uterine, and mammary tissue during pregnancy and milk production during lactation. In one study, metabolizable energy requirements for domestic ruminants and horses increased by 28.2 and 23.2% above maintenance levels, respectively, during late pregnancy and by 37.2 and 41.5% from late pregnancy to early lactation. In free-ranging golden mantled ground squirrels (*Spermophilus lateralis*), the amount of metabolized energy allocated to reproduction by females into newborn young and milk export increased by 17.7% over nonreproductive levels in one study. The total energy budget of a lactating female includes loss or storage of body reserves and the amount of energy that is produced and exported as milk (Fig. 1). In free-ranging little brown bats (*M. lucifugus*), average daily energy requirements during lactation (litter size of one) exceeded requirements during late pregnancy by a factor of 1.2 in one study. However, when milk output was included, energy required by females during lactation nearly doubled from early (33.8 kJ/day) to peak lactation (60.3 kJ/day).

Among terrestrial mammals, most young are weaned by the time they have reached approximately 40% of adult size. In contrast, lactating female bats typically nurse their young until they have nearly reached adult size. Because bats cannot fly until they are nearly adult size, mothers must continue to suckle their young until such time that they have achieved approximately 90% of adult wing dimensions and can secure food on the wing. The lactation period for most small (<60 g) insectivorous bats is 6 to 8 weeks. Larger mammals (>10 kg) have disproportionately longer lactation periods (several months) and distribute their total costs over a far greater period of time. Because larger species tend to produce relatively low levels of milk at peak lactation and during the prolonged end of the lactation cycle, the average daily cost as a function of BMR is lower in larger species. Thus, lactation costs are high but are of relatively short duration in small mammals, as compared with a relatively lower cost but longer duration in large mammals.

Because most mammals are polygynous, males assume little or no role in the care of the young they have fathered. However, among monogamous species, females suckle their young but may be assisted by fathers who provision their young with food (e.g., many carnivores). In general, male mammals do not lactate; however, recent observations have shown that some male mammals do produce milk, perhaps as a trait in some monogamous species or possibly as a consequence of eating leaves high in phytoestrogens or eating food that has been exposed to pesticides that have estrogenic properties. Males of at least two species of Old World fruit bats (*Dyacopterus spadiceus* and *Pteropus capistratus*) are known to produce milk, although it has not been determined that they actually suckle their offspring. Because the mating system of these two species is not known, any relationship between a monogamous mating system and male lactation remains unclear.

In addition to productive processes (fetal growth and lactation) associated with pregnancy and lactation in mammals, other energy costs include the establishment and maintenance of territories associated with courtship and mating, gamete production (relatively small in mammals), the increased locomotor activity associated with food acquisition, nest construction and guarding, and other forms of parental care.

6.6 Parental Care

Because of constraints imposed by polygynous mating systems and female lactation, parental care in mammals generally is largely the responsibility of females. However, in some species that breed cooperatively, offspring from previous litters help

care for members of the subsequent litter. Events and activities associated with parental care in mammals cover gestation, parturition (birth), lactation, grooming, feeding, and tutoring of young. During or before the gestation period, nest construction and nest defense are the primary energy-demanding activities that extend above maintenance costs. Parturition is an energy-demanding activity, but because this period is relatively short (minutes to hours), the overall impact on the energy budget of a female is trivial. During lactation, thermoregulation and grooming are the primary activities. Among these, lactation is the most energetically demanding. The energy demands associated with lactation will depend on the duration of lactation, the number and size of offspring, the body size of the mother, the local microclimate, and the availability of food resources. For species in which there is a prolonged postnatal period of growth and development (e.g., primates, elephants), tutoring exacts additional energy costs.

Among monogamous mammals, males invest in their offspring both indirectly and directly. For example, a dominant male gray wolf (*Canis lupus*) or dominant male coyote (*Canis latrans*) will bring food to a lactating female in its den, and when the young approach the age of weaning, he will offer them partially chewed food. Males may defend burrows that contain neonates from potential predators, either by giving alarm calls, signaling with visual displays, or distracting potential predators away from nest area. Paternal care that is most often associated with monogamy is relatively rare in placental mammals, averaging less than 5% of all species that have been studied.

Alloparental care has also been observed in several species of mammals. Among these are highly social and cooperative breeding species. For example, alpine marmots (*Marmota marmota*) live in colonies that include a territorial pair, subordinates, yearlings, and young of the year. Subordinates cluster with newborns, groom them, and cover them with grass, behaviors that improve the survival of offspring. A special case of alloparental behavior has been observed in eucocial mammals, including the naked mole rat (*Heterocephalus glaber*) and Damara mole rat (*Cryptomys damarensis*). Colonies of naked mole rats consist of one reproductive female, several males; and several nonreproductive female workers. Reproductive maturity of subordinate females is suppressed pheromonally by the presence of a dominant "queen" who mates with and gives birth to offspring fathered by a few resident males. The reproductively suppressed females construct burrows and nest chambers and assist the queen by bringing her food and grooming her. The dominant female suckles all of the offspring. Alloparental behavior also extends to assistance during the birthing period. Female Rodrigues fruit bats (*Pteropus rodrigues*) have been observed assisting other females during the birthing period. In one instance, when a female was having a difficult (breech birth), another female approached the expecting mother and groomed her pup as it was emerging from the vagina. She also tutored the mother in the correct birthing position and wrapped her wings around the mother and pup during the birth process. Similar observations of helper-assisted births have been observed in the common marmoset (*Callithrix jacchus*), African elephant (*Elephas maximus*), Afghan hunting dog (*Canis familiaris*), and spiny mouse (*Acomys cahirinus*). Sharing of a blood meal with roostmates, including young, has been observed in the common vampire bat (*Desmodus rotundus*). This behavior appears to be practiced mostly by mothers who provision their own young, but it has also been observed between adults as well as between adults and related but nonfilial young. This altruistic behavior appears to be based on reciprocity and kin selection in this highly social and communally breeding species. Parental care in mammals, especially in highly social species (e.g., primates, some rodents, bats), also involves social grooming and protecting young from extremes in temperature or inclement weather. The latter may also involve clustering and huddling to reduce energy expenditure.

7. CONCLUDING REMARKS

This article has provided a brief overview and a few glimpses of how various factors influence the reproductive energetics of terrestrial reptiles, birds, and mammals. The amount of energy allocated to various phases of reproduction (courtship and mating, incubation, gestation, lactation, parental care, and the pre- and postfledging or postlactational periods) is, in part, influenced by the type of reproduction (oviparity vs. vivaparity), different types of mating system (monogamy, polygyny, and polyandry), rates of posthatching and postnatal growth, competitive interactions with members of the same or different species, and extrinsic factors such as habitat quality, food availability, and climate. In general, costs of courtship and mating are higher for males that typically compete for females by developing elaborate structures or behaviors used in

courtship display to show females that they are fit and by defending either mating territories or females. The energy investment that male vertebrates make in the production of gametes is relatively small compared with that made by females, and female reptiles and birds typically invest more in egg production (both size and number) than do female mammals.

Apart from the production of a relatively large number of eggs, few reptiles exhibit maternal care once the young hatch. In contrast, birds and mammals often invest extensively in their young once they hatch or are born. Female birds allocate large amounts of energy to brooding and feeding their young until they have reached or exceed adult size. Some species invest further in their offspring by tutoring them once they have fledged. In contrast, mammals typically produce a relatively small number of eggs that are fertilized internally, the fetus develops in the uterus that is protected from environmental extremes, young are born alive, and females nurse their young with milk for variable periods of time and in many species may provide extended care to their offspring. Among males, courtship and mating are the most energetically expensive activities in reptiles, birds, and mammals. The production of nutrient- and energy-rich eggs is the most energetically expensive phase of reproduction in female reptiles. Production and incubation of eggs and feeding of nestlings comprise the most energetically expensive phase of reproduction in female birds. Lactation is the most energetically expensive period of reproduction in female mammals.

SEE ALSO THE FOLLOWING ARTICLES

Conversion of Energy: People and Animals • Food Capture, Energy Costs of • Human Energetics • Migration, Energy Costs of • Thermoregulation

Further Reading

Blem, C. R. (2000). Energy balance. *In* "Sturkie's Avian Physiology" (G. C. Whittow, Ed.), 5th ed., pp. 327–341. Academic Press, San Diego.

Bradbury, J. W., and Vehrencamp, S. L. (1998). "Principles of Animal Communication." Sinauer Associates, Sunderland, MA.

Carey, C. (ed.). (1996). "Avian Energetics and Nutritional Ecology." Chapman & Hall, London.

Clutton-Brock, T. H. (1991). "The Evolution of Parental Care." Princeton University Press, Princeton, NJ.

Clutton-Brock, T. H., Guinnesss, F. E., and Albon, S. D. (1982). "Red Deer: Behavior and Ecology of Two Sexes." University of Chicago Press, Chicago.

Congdon, J. D., Dunham, A. E., and Tinkle, D. W. (1982). Energy budgets and life histories of reptiles. *In* "Biology of the Reptilia" (C. Gans and F. H. Pough, Eds.), Vol. 13, pp. 233–271. Academic Press, New York.

Cummins, J. M., and Woodall, P. F. (1985). On mammalian sperm dimensions. *J. Reprod. Fertil.* **75,** 153–175.

Eisenberg, J. F. (1981). "Mammalian Radiations." University of Chicago Press, Chicago.

Feldhamer, G. A., Drickamer, L. C., Vessey, S. H., and Merritt, J. F. (2003). "Mammalogy: Adaptation, Diversity, Ecology." McGraw–Hill, New York.

Gittleman, J. L. (ed.). (1989). "Carnivore Behavior, Ecology, and Evolution." Cornell University Press, Ithaca, NY.

Gittleman, J. L., and Thompson, S. D. (1988). Energy allocation in mammalian reproduction. *Am. Zool.* **28,** 863–875.

Hosken, D. J. (1997). Sperm competition in bats. *Proc. Roy. Society London B* **264,** 385–392.

Kunz, T. H., and Fenton, M. B. (eds.). (2003). "Bat Ecology." University of Chicago Press, Chicago.

Kunz, T. H., and Hood, W. H. (2000). Parental care and postnatal growth in the *Chiroptera*. *In* "Reproductive Biology of Bats" (E. G. Crichton and P. H. Krutzsch, Eds.), pp. 415–468. Academic Press, San Diego.

Kunz, T. H., and Lumsden, L. F. (2003). Ecology of cavity and foliage roosting bats. *In* "Bat Ecology" (T. H. Kunz and M. B. Fenton, Eds.), pp. 3–89. University of Chicago Press, Chicago.

Kunz, T. H., and Nagy, K. A. (1988). Methods of energy budget analysis. *In* "Ecological and Behavioral Methods for the Study of Bats," (T. H. Kunz, Ed.), pp. 277–302. Smithsonian Institution, Washington, DC.

Lacey, E. A., Patton, J. L., and Cameron, G. N. (eds.). (2000). "Life Underground: The Biology of Subterranean Rodents." University of Chicago Press, Chicago.

Lacey, E. A., and Wieczorek, J. R. (2001). Territoriality and male reproductive success in arctic ground squirrels. *Behav. Ecol.* **12,** 626–632.

Loudon, A. S. I., and Racey, P. A. (eds.). (1987). "Reproductive Energetics in Mammals." Oxford University Press, Oxford, UK.

McCracken, G. F., and Wilkinson, G. S. (2000). Bat mating systems. *In* "Reproductive Biology of Bats" (E. G. Crichton and P. H. Krutzsch, Eds.), pp. 321–362. Academic Press, San Diego.

McNab, B. K. (2002). "The Physiological Ecology of Vertebrates: A View from Energetics." Cornell University Press, Ithaca, NY.

Nagy, K. A. (2001). Food requirements of wild animals: Predictive Equations for free-living mammals, reptiles, and birds. *Nutr. Abstr. Rev.* **71,** 21R–32R.

Oftedal, O. T. (1985). Pregnancy and lactation. *In* "Bioenergetics of Wild Herbivores" (R. J. Hudson and R. G. White, Eds.), pp. 215–238. CRC Press, Boca Raton, FL.

Orrell, K. S., Congdon, J. D., Jenssen, T. A., Michener, R. H., and Kunz, T. H. (2003). Comparison of energy expenditure by male and female *Anolis carolinensis* during breeding and postbreeding seasons. *Physiol. Biochem. Zool.,* in press.

Orrell, K. S., Congdon, J. D., Jenssen, T. A., Michener, R. H., and Kunz, T. H. (2004). Intersexual differences in energy expenditure of *Anolis carolinensis* lizards during breeding and postbreeding seasons. *Physiol. Biochem. Zool.* **77.** In press.

Peaker, M., Vernon, R. G., and Knight, C. H. (eds.). (1984). "Physiological Strategies in Lactation." Symposia of the Zoological Society of London No. 51 Academic Press, London.

Ricklefs, R. E. (1974). Energetics of reproduction in birds. *In* "Avian Energetics" (R. A. Paynter, Jr., Ed.), pp. 152–297. Nuttall Ornithological Club, Cambridge, MA.

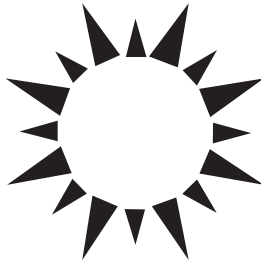
Robbins, C. T. (1993). "Wildlife Feeding and Nutrition." Academic Press, San Diego.

Stearns, S. (1992). "The Evolution of Life Histories." Oxford University Press, New York.

Voigt, C. C., von Helversen, O., Michener, R., and Kunz, T. H. (2001). The economics of harem maintenance in the sacwinged bat, *Saccopteryx bilineata*. *Behav. Ecol. Sociobiol.* **50,** 31–36.

Wilkinson, G. S., and McCracken, G. F. (2003). Bats and balls: Sexual selection and sperm competition in the *chiroptera*. *In* "Bat Ecology" (T. H. Kunz and M. B. Fenton, Eds.), pp. 128–155. University of Chicago Press, Chicago.

Research and Development Trends for Energy



PAUL J. RUNCI and JAMES J. DOOLEY
Joint Global Change Research Institute
College Park, Maryland, United States

1. **The Importance of Energy R&D**
2. **Energy R&D Performers and Sponsors**
3. **Motivations for Energy R&D**
4. **Trends in Energy R&D Investments**
5. **Assessing R&D Trends: Methodological Challenges**
6. **Conclusion**

Glossary

development A stage of the innovation process during which basic research findings are applied to the development of new materials, devices, or processes for practical use.

energy restructuring Often referred to as energy deregulation or liberalization, the process of opening the energy industries (e.g., electricity, gas, oil) to market forces and competition; in many countries, the process involves the dismantling of large, integrated state monopoly energy companies and the privatization of all, or of large parts of, the industries.

energy technologies Materials, devices, and processes for the location, recovery, preparation, storage, transport, conversion, and use of energy resources and for the treatment of associated waste products.

innovation The systematic practice of research and development and the creation of new technologies, usually through a linked network of scientists, engineers, and research sponsors at university, private sector, and government-owned research facilities.

research The initial stage of the innovation process where new knowledge of the natural world is sought through experiment and observation without specific motivation for the application of findings to particular practical problems or to the creation of new technologies.

Energy research and development (R&D) is often defined in a manner similar to that in which a U.S. Supreme Court justice defined pornography: "I know it when I see it." That is, most of those concerned with energy R&D believe that there is a common tacit definition, yet attempts at explicit definition usually reveal significant differences in understanding among researchers, analysts, and those who support energy innovation. Accepting that there are few explicit definitions of energy R&D provided in the literature, and that no extant definition is regarded as constituting a standard one, this article adopts the following definition based on that of the U.S. President's Committee of Advisors on Science and Technology:

> *Energy R&D refers to a series of linked processes by which technologies for energy supply, end use, or carbon management move from theoretical conceptualization to feasibility testing and, ultimately, to small-scale deployment. Energy R&D encompasses both basic and applied research, technology development, and demonstration associated with each phase of the energy life cycle, including production (e.g., mining, drilling), energy conversion and power generation (e.g., nuclear fission and fusion, fossil and renewable energy systems, bioenergy, hydrogen production), transmission, distribution, energy storage, end use and energy efficiency, and carbon management. Carbon management technologies aim to manage anthropogenic releases of greenhouse gases, such as those associated with the combustion of fossil fuel use, in an effort to mitigate the potential impacts of these emissions on climate systems. Carbon management technologies include advanced agroforestry practices aiming to enhance the capacity of soils, trees, and other standing biomass to absorb atmospheric carbon dioxide; engineered pre- and postcombustion carbon capture systems; and technologies for geologic and ocean carbon sequestration.*

This definition is an extension of one first put forward by the U.S. President's Council of Advisors on Science and Technology in 1999. Although this constitutes a useful and comprehensive working definition of energy R&D, it is important to note that it is descriptive rather than prescriptive in

 © 2004 Elsevier Inc. All rights reserved.

nature; no government agency or private organization uses such a definition as a guide in planning its energy R&D investments. Although many commercially available technologies, such as low-emissivity windows, are the result of established energy R&D programs, other projects and technologies are often classified as the fruit of energy R&D investments *ex post facto*. For example, the use of natural gas turbines for electric power production has been one of the most important energy technology advances of the past few decades, yet the investments that led to this breakthrough were largely made by the U.S. Department of Defense during the 1960s and 1970s under the rubric of aircraft engine research. That this work led in the long run to the technology's application in the electric power industry raises taxonomical questions about its retroactive classification as "energy R&D." Thus, defining energy R&D remains a malleable enterprise.

1. THE IMPORTANCE OF ENERGY R&D

Energy technologies are the primary determinants of energy availability, fuel choice, end use efficiency, and the degree and nature of by-product emissions, and energy research and development (R&D) is the vehicle by which new technologies become available. To the extent that energy technologies augment or diminish society's range of energy choices, energy R&D is a matter of high importance. Over the past half-century or so, awareness of energy use and its implications for economic growth, national security, and environmental quality have at times held prominence to an unprecedented degree. To a large degree, these three societal goals have formed the underlying rationale for investments in energy R&D across the globe over this period. The relative emphasis placed on each of these goals has varied from country to country and even within countries over time.

For example, two of the three founding treaties of the European Union focused on the coordinated management of energy resources and energy R&D for reconstruction and economic growth. Moreover, virtually all industrialized countries created domestic agencies for energy technology development during the post-World War II period. For energy resource-poor countries such as Japan, the fact that energy R&D continually receives a sizable percentage of government's total outlays for all kinds of R&D (more than 10% of all government R&D expenditures consistently over the past 25 years or so) is a testament to the importance of this activity. At other times and in countries such as the United States, Canada, and most of the European Union, the importance of energy and energy R&D has received relatively little attention, and as a result, public and private sector investments in energy R&D have suffered. Energy R&D tends to be a near- to midterm beneficiary of energy crises.

Energy considerations are central factors in several policy areas, including environmental, economic, and defense policy. Energy resources have served both as a cause for and as a weapon of war, and the consequences of energy use have manifest themselves in a host of environmental problems, ranging from local air pollution to global climate change. Consequently, energy R&D and the technologies that the R&D process yields broaden or narrow the portfolio of choices available to policymakers on a number of important issues.

2. ENERGY R&D PERFORMERS AND SPONSORS

Energy R&D has historically been the province of large industrialized countries with limited domestic energy resources and high levels of energy imports. Most of the world's energy R&D is funded and performed by a small group of industrialized countries: the United States, Canada, Japan, Germany, France, Italy, The Netherlands, Switzerland, the United Kingdom, and countries comprising the European Union. According to the International Energy Agency (IEA), these countries alone account for more than 90% of the industrialized world's energy R&D. Although these countries are clearly the "energy R&D-intensive" leaders, only in the case of Japan does energy R&D account for more than 1.5% of a nation's overall R&D enterprise.

The performance of energy R&D in the private sector is similarly concentrated in large industrial firms. For example, in the United States in 2000, firms that had more than 10,000 employees performed 89% of industrial energy R&D. Although it is clear that these companies are the leading performers of energy R&D, it is also clear that the performance of R&D is not a significant aspect of their business. The energy industries are among the least R&D intensive in the industrialized world. In comparison, the most R&D-intensive industries, such as medical equipment and information technology, invest 12 and 8% of annual

sales, respectively, in R&D. Although a few specialized oil and gas exploration service companies have R&D/sales ratios in excess of 7%, this level of investment is the exception within the energy industries that typically invest less than 1% of sales in R&D.

Neither of these observations is meant to exclude the possibility (indeed the cerainty) that smaller countries and smaller firms are not conducting important energy R&D. Rather, it seems clear that the performance of energy R&D is limited to this select set of large industrialized countries and large firms by the often times significant research infrastructure requirements. (It is hard to imagine a very small firm conducting all of the research needed to bring about a new fission reactor from conception to pilot scale testing.) The successful performance of energy R&D often depends on the existence of a large cadre of scientists and engineers, the availability of specialized laboratories and user facilities, funding mechanisms for sustained research support, and programs for technology commercialization and adoption. A relatively small set of countries and large firms have these capabilities; therefore, it is these entities that constitute the bulk of the global energy R&D enterprise.

Governments have historically been the principal sponsors of energy R&D, whereas the private sector has played an important but secondary role from the perspective of overall investment levels. Governments have historically been the main supporters of long-term, higher risk R&D in particular, whereas private sector investments have focused more on short-term research and technology commercialization efforts. In several industrialized countries, including the United States, Germany, and Japan, the gap between public and private sector investments has widened over the past 15 years or so. For example, whereas government and industry investments were roughly equal in 1990, government now invests approximately 50% more than the private sector in the United States. Similarly, whereas the German government invested 30% more than did industry in 1990, by 2000 its energy R&D investment was more than double that of the private sector.

3. MOTIVATIONS FOR ENERGY R&D

As noted previously, in most countries, public sector energy R&D has been closely tied to domestic policy priorities such as efforts to bolster national security, to enhance economic productivity, and to remedy specific local- to regional-scale environmental problems. Because policy priorities change continually, so too have the focus and intensity of R&D efforts in performing countries over the past half-century. For example, the oil supply shocks of the 1970s and the related economic crises were among the strongest drivers of energy R&D efforts in industrialized countries. Between 1973 and 1981, government investments in energy R&D rose to their highest levels as new programs and new institutions formed in response to oil supply interruptions initiated by the Organization of Petroleum Exporting Countries (OPEC). Domestic policy programs, such as Project Independence in the United States, and agencies with significant energy R&D missions, such as the U.S. Department of Energy (DOE), Britain's Department of Energy, and the Organization for Economic Cooperation and Development's (OECD) IEA, grew out of Western policymakers' ambitions to reduce exposure to future energy supply interruptions. Nuclear fission and fusion research programs, energy efficiency, and renewable energy research all received substantial funding increases in the interest of reducing the energy security risks associated with high levels of fossil fuel import dependence. In real inflation-adjusted terms, energy R&D investments hit their high-water marks during this period and have been on a steady decline, for the most part, since the late 1970s and early 1980s. For example, U.S. energy R&D expenditures in 1980 were nearly three times what they are at the turn of the 21st century.

These investments in energy R&D have yielded significant economic benefits. For example, over the 20-period between 1973 and 1993, energy productivity—the amount of energy required to produce $1 of gross domestic product (GDP)—rose by 1.7% annually on average in the United States and by as much as twice that rate in Japan and Europe. The decoupling of energy consumption from economic growth also dismissed a widely held belief regarding the tight interdependence of these two factors. In addition, overall levels of energy import dependence in industrialized countries declined for more than a decade, although they have since rebounded to historically high levels.

Although improvements in energy productivity over this period cannot be attributed exclusively to investments in energy R&D, these investments were an important factor. As the National Research Council found in its 2001 study of DOE-sponsored energy efficiency and fossil fuel R&D programs, real economic benefits were realized from DOE and private sector investments between 1978 and 2000.

Although other factors, including structural changes in the U.S. economy, fuel prices, and changes in the fuel mix, also played important roles, the National Research Council panel found that a small group of government programs exercised tremendous leverage in the U.S. economy. For example, government-sponsored programs addressing building efficiency and fossil fuel-related NO_x emissions realized benefits in the form of avoided energy and environmental costs that were orders of magnitude larger than the associated R&D investments.

In the private sector, profitability and the need to meet ever tighter environmental regulations have been the prime motivators for investment in energy technology development. Private energy R&D efforts intensified during the 1970s and 1980s in response to the higher market prices that made energy production more profitable for those firms that were energy producers and provided a significant motivation for energy users to find ways in which to conserve energy as a cost-cutting measure. For example, for oil and gas companies, the prospect of higher profits prompted major new investments in exploration and production technology aimed at bringing to market energy resources that had been considered uneconomical at lower prices. Important technological advances, such as three-dimensional seismic imaging and directional drilling, occurred during this period and have continued to buoy the profitability of energy companies through subsequent and prolonged periods of low energy prices.

For gas and electric utilities in many countries, the motivation for energy R&D stemmed from energy providers' "regulatory compact" with government. Typically, regulated utilities were guaranteed a geographically isolated market and a fixed rate of return on their capital and R&D investments, minimizing the risks associated with those investments. Moreover, in some cases (e.g., the United States), utilities were required to invest a specified portion of annual revenues in common associations that served as industry-wide R&D pools. The Electric Power Research Institute and the Gas Research Institute derived their existence largely from the mandated "wires and pipeline charges" paid by power and gas suppliers in the regulated energy marketplace. However, the process of energy deregulation has usually included the repeal of government-mandated R&D contributions. As a result, collective R&D institutions have seen their annual revenues fall by more than 50%.

The deregulation of the energy industries raises lingering questions about the current and future motivations for energy R&D investments. Although major problems such as climate change and energy security are indicative of the significant energy challenges that lie ahead, the structure of market and policy incentives have shifted in ways that motivate public and private sector decision makers to forgo higher overall levels of energy R&D investment. Arguably, the most pressing problem facing energy R&D sponsors and performers today is the sharp dislocation between current incentives for energy R&D investment and the likely future societal demands for new generations of cleaner and more efficient energy technologies.

4. TRENDS IN ENERGY R&D INVESTMENTS

According to the IEA, the industrialized world's overall investment in energy R&D declined by approximately 17% during the 1990s (Fig. 1). The IEA's 1999 aggregate estimate of world energy R&D investment, $7.4 billion, reflects a decrease of 50% from the high-water mark of $14.7 billion in 1980.

Of the group of nine countries that constitute the world's leading energy R&D sponsors (the United States, Japan, France, Germany, Italy, Canada, The Netherlands, Switzerland, and the United Kingdom), seven have reduced their expenditures over the past decade or more. In each of these countries, overall R&D spending has increased significantly in real terms. Some areas, such as health and biomedical research and information technology, have experienced large and sustained R&D investment increases, reflecting the major advances in these fields over the past two decades or so and the widely held belief that the private and public returns on additional investments in these areas will be significant. The absence of acute energy crises and perceptions of shortage, such as those that characterized the 1970s and early 1980s, have discounted the value of government-sponsored energy R&D in the eyes of policymakers and the general public. Thus, as a research priority, energy R&D has lost ground in both relative and absolute terms as an element of the broader research portfolios of most industrialized countries. Only France and Japan have increased their energy R&D investments in real terms since 1991 (Table I).

Although the aggregate trend data reveal steadily waning investments, individual energy technology areas are flourishing in some countries. For example, because hydrogen and fuel cells are widely regarded

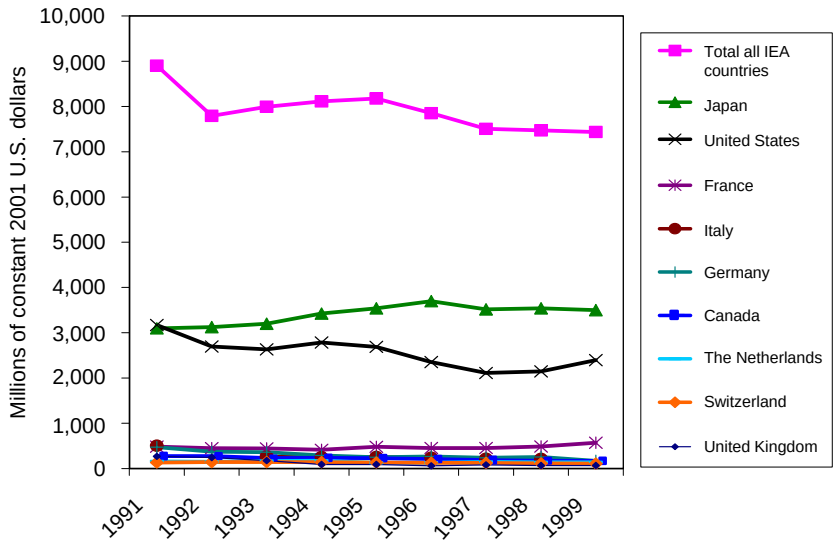


FIGURE 1 Energy R&D investment trends in major performing countries during the 1990s, as reported to the International Energy Agency (IEA). Note that currency conversions in the figure are based on market exchange rates rather than on purchasing power parities. Because energy R&D is largely a nontraded good (i.e., funded and produced within national borders rather than across them), the use of market exchange rates in converting local currencies to constant U.S. dollars may distort the real values of countries' R&D investments.

TABLE I

Real Growth in Energy R&D Investments, 1991–2001 (Percentages)

Country	Change in energy R&D investment, 1991–2001
United States	−13
Japan	+15
France	+17.8
United Kingdom	−71
Italy	−50
Germany	−46.5
Canada	−43
The Netherlands	−14
Switzerland	−11

Note. The majority of major performing countries made large cuts in energy R&D support during the 1990s. As the table shows, restructuring and deregulation of the energy industries have been strongly correlated with energy R&D funding cuts. France and Japan are the only major energy R&D-performing countries that have not yet restructured their energy industries.

as holding significant promise as a future alternative to fossil fuel-based energy systems, investment in these areas has risen sharply during recent years. In Germany, fuel cell research has grown by 400% since 1998. In the United States and Canada, hydrogen programs have nearly doubled over the same period. The European Union and Japan also sponsor major public and private sector hydrogen fuel cell programs. Similarly, government-sponsored biomass/biofuels, energy efficiency, and carbon sequestration programs have grown moderately in the United States, Japan, Germany, and Australia. Yet even with significant growth in these selected technology areas, emerging energy technology programs still constitute just a small share of energy R&D budgets.

Assessing private sector energy R&D trends has always presented difficult challenges because most energy R&D-performing firms consider these investment data sensitive and proprietary. The private sector data that are available are usually gathered via a brief survey instrument by industry associations or government agencies and are published only at an industry aggregate level. Trend data from these sources reveal that private energy R&D investments have declined even more precipitously over the past decade or so than have public sector energy R&D investments in several countries, including the United States (−63%), Germany (−70%), the United Kingdom (−90%), and Canada (−40%).

The energy R&D sponsored by the private sector is largely short term focused, aiming to develop new products or improve existing ones so as to bring them to the marketplace quickly. Long-term energy R&D, which entails larger financial risks and uncertain time horizons, has generally been supported by government agencies. In the United States, government currently sponsors more than 60% of

long-term energy R&D, whereas industry supports more than 80% of short-term research.

5. ASSESSING R&D TRENDS: METHODOLOGICAL CHALLENGES

It is important to note that all of the preceding data are based on surveys that use the term "energy R&D" in a conventional and narrow sense. The term is used to describe R&D activities that were explicitly carried out by governments or industry primarily to increase or create new energy supplies or to use energy more efficiently and cleanly. This means that these data sets will exclude R&D whose main and *a priori* purpose was to create something like a new military system, even though at times strictly defined "military R&D" has spilled over to benefit the energy industry. The most obvious example of this spillover phenomenon is the aforementioned combined cycle gas turbines, which have had a profound impact on the energy industry and are largely derivative of military jet engines.

The exclusion of these ancillary R&D efforts is driven by practicality and a desire to keep some temporal integrity to the data set being examined here. That is, only in an *ex post facto* sense can one identify these technologies that spill over into "energy R&D," and the selection of these technologies is an inherently subjective exercise. It is more feasible to track R&D expenditures that were initiated due to their relevance to energy. Analysts and decision makers who rely on energy R&D statistics need to understand this important nuance in the data on which they are relying.

Another important methodological issue that analysts must consider in assessing energy R&D data across countries concerns the exchange rate mechanism used to convert the data to a common (typically U.S. dollars) currency. Many publicly available, multicountry energy R&D data sets use market exchange rates rather than purchasing power parities for their conversion to U.S. dollars. This method can introduce significant distortions into the data, often inflating actual energy R&D investments. Given that energy R&D is nearly exclusively funded and performed within—rather than across—national borders, exchange rates based on purchasing power parity yield more accurate estimates than do market exchange rates, which are more appropriately used for internationally traded goods and services.

It is also important to note that the use of brief survey instruments, by the IEA and others, as a primary mechanism for energy R&D data gathering should instill a measure of caution in analysts working with data sets compiled in this manner. Such survey instruments typically contain questions regarding a country's or firm's overall spending levels on energy R&D in selected program areas but offer little detail below aggregate levels. The use of survey instruments in this manner also raises questions regarding data accuracy and reliability given that survey participants are usually given few if any guidelines regarding their responses. Consequently, investments in fossil energy efficiency could, for example, be classified as "fossil energy R&D" by one participant and as "energy efficiency R&D" by another. Finally, verification of energy R&D statistics provided to international agencies via surveys is not possible. Thus, although these statistical data sets may be some of the best available, they must nonetheless be used and interpreted with caution.

6. CONCLUSION

Energy R&D is an important vehicle for the achievement of a wide variety of societal and policy goals. Because energy technology and fuel choices are critical determinants of economic prosperity, environmental quality, and national security, energy R&D is an important public policy vehicle. Moreover, the relative strength of energy R&D support is also a useful metric for the assessment of a number of public goods. The extent to which firms and government agencies view energy R&D as a priority activity and fund it accordingly may be regarded as a proxy indicator of political interest in selected policy areas such as climate change mitigation, national energy security, and overall environmental stewardship.

In this regard, it is important to note that available energy R&D data sets provide an important, but limited, measurement of inputs into energy R&D and offer little insight into R&D outputs—new technologies, their deployment, and their impacts on the environment as well as the economy and other policy concerns. Other common proxy measures of R&D efforts, including statistics on patent grants and applications, doctoral degrees awarded, and trends in energy intensity, can serve as useful pieces of the broad mosaic of energy R&D data but can be difficult to interpret outside a broader analytical context. Like statistics on energy R&D expenditures, these data are used as proxy measures because energy

R&D statistics are relatively easy to collect and have the appearance of being easy to interpret. In fact, output measures such as energy intensity (i.e., energy input per unit of GDP) are notoriously difficult to interpret because the number of variables influencing observed trends is potentially large and the isolation of individual variables is often impossible. For example, given data limitations, it might not be possible to determine definitively whether improvements in industrial energy intensity are due to the introduction of more fuel-efficient energy technologies, to broader changes in production methods, to structural changes in the economy, or to any number of other factors.

Most recent studies of trends in energy R&D have documented significant reductions in the overall level of support and a concomitant redistribution of remaining energy R&D funds in favor of less risky, short-term evolutionary R&D with clear commercial applications rather than longer term, higher risk projects with potentially revolutionary impacts. Many analysts have interpreted this trend as inherently dangerous given the likelihood that such funding patterns are likely to delay or prevent the attainment of many energy policy goals such as climate change mitigation. On the other hand, some analysts have pointed out that a reduction in the amount spent on energy R&D in itself cannot always be interpreted as a step backward for a firm or for society at large. For example, it is possible that energy R&D-performing laboratories might have become more efficient in their delivery of energy R&D results due to the increasing use of computer modeling or due to mergers between firms with large and redundant R&D capabilities that are subsequently consolidated—a frequent occurrence in the oil and gas industries. In either case, energy R&D statistics would reflect the fact that smaller amounts are invested in energy R&D, yet the implications of these events for energy technology innovation would remain indeterminate.

A final caution concerns forward-looking statements regarding future patterns of energy R&D investment and performance. As all investors are warned, past performance cannot be relied on as a predictor of the future. Many factors, including economic downturns, unforeseen energy crises, or environmental surprises can and have affected energy R&D investment in significant and unpredictable ways. Likewise, policy rhetoric regarding energy R&D priorities often remains disjointed from actual R&D expenditures. For example, during the 1990s, many governments issued numerous white papers, policy briefs, and press releases about the importance of reorganizing their energy R&D portfolios to focus on climate change or to put more emphasis on renewable energy R&D. However, the historical budget data that detail how these governments actually allocated their energy R&D funds reveal modest changes in overall energy R&D emphasis. Thus, for the energy R&D policy analyst, *caveat emptor* remains sound advice. Analysts are advised to seek out multiple sources of qualitative and quantitative data on energy policy and energy R&D before drawing conclusions about the health and direction of the energy R&D enterprise.

SEE ALSO THE FOLLOWING ARTICLES

European Union Energy Policy • Experience Curves for Energy Technologies • Innovation and Energy Prices • Microtechnology, Energy Applications of • Technology Innovation and Energy

Further Reading

Dooley, J. J. (1998). Unintended consequences: Energy R&D in a deregulated energy market. *Energy Policy* **26**, 547–555.

Dooley, J. J., and Runci, P. J. (2000). Developing nations, energy R&D, and the provision of a planetary public good: A long-term strategy for addressing climate change. *J. Environ. Dev.* **9**, 215–239.

Electric Power Research Institute. (1996). "The Impact of Electricity Restructuring on Research for the Public Good," BR-107160. EPRI, Palo Alto, CA.

Lichtenberg, F. (1987). Changing market opportunities and the structure of R&D investment: The case of energy. *Energy Econ.* **9**, 154–158.

Mansfield, E., and Switzer, L. (1984). Effects of federal support on company-financed R and D: The case of energy. *Mgmt. Sci.* **20**, 562–571.

Margolis, R. M., and Kamen, D. M. (1999). Underinvestment: The energy technology and R&D policy challenge. *Science* **285**, 690–692.

Pacific Northwest National Laboratory. (2000). "Energy Research and Development: Global Trends in Policy and Investment." http://energytrends.pnl.gov.

President's Council of Advisors on Science and Technology. (1996). "Report to the President on Federal Energy Research and Development for the Challenges of the Twenty-First Century." PCAST, Washington, DC.

President's Council of Advisors on Science and Technology. (1999). "Powerful Partnerships: The Federal Role in International Cooperation on Energy Innovation." PCAST, Washington, DC.

Resource Curse and Investment in Energy Industries

PAUL STEVENS
University of Dundee
Dundee, United Kingdom

1. **A History of Interest in Resource Impact**
2. **The Presence of Resource Curse**
3. **Transmission Mechanism between Large Resource Revenues and Poor Performance**
4. **How Resource Curse May Be Avoided: The Theory and the Practice**
5. **Conclusion**

Glossary

crowding out Where one sector manages to secure first call on resources, preventing access to those resources for other sectors.

developmental state A country in which the government espouses an ideology of development and tries to create the capacity to implement policies to promote development.

Dutch disease As a result of large oil or gas revenues, the real exchange rate appreciates, causing imports to rise and non-oil or gas exports to fall. The result is a contraction in the non-oil or gas tradable sector.

oil fund An institution whose function is to receive the revenues from oil or gas projects and to control their use in terms of both investing and spending.

predatory state The opposite of a developmental state; the government is a kleptocracy in which the ruling elite is out to plunder the wealth of the country for its own selfish purposes.

rent seeking Where managers in an enterprise—state or privately owned—seek to absorb resources to improve their own working environment and lifestyle. This is not the same as corruption since rent seeking is viewed as normal and legal behavior.

terms of trade The amount of goods or services that have to be exchanged for other goods or services; for example, the number of agricultural products that must be exchanged for an amount of manufactured goods.

Large revenues from natural resources should generate wealth, promote economic progress, and reduce poverty. The logic behind this may be based on common sense (more money may be better than less) or economic development theories, including capital fundamentalism, dual-gap analysis, or the big push argument to break out of a self-feeding circle of poverty. Thus, large windfalls from oil or gas projects should overcome capital and foreign exchange constraints and generate the required "push." However, for many years casual empiricism and more formal investigation suggest the opposite. Many resource-rich countries have experienced a worse performance in economic progress and poverty reduction than countries without such apparent benefits. In the economics literature, this has become known as the resource curse thesis. However, it should be noted that although many suffered a "curse" from the influx of large natural resource revenues, others did not. Thus, resource curse is by no means the default situation.

1. A HISTORY OF INTEREST IN RESOURCE IMPACT

Concern regarding the impact of natural resources dates back many centuries, most obviously with Spain's wealth from the New World. Concern emerged among development economists such as Raul Prebisch and Hans Singer in the 1950s and 1960s. They argued that primary product exporters (the periphery) would find themselves disadvantaged in trading with the industrialized countries (the center) because of deteriorating terms of trade.

Concern was focused on primary products and natural resources generally. However, in the 1970s, attention focused on the oil exporters following the oil

 © 2004 Elsevier Inc. All rights reserved.

shocks. Their aftermath caused speculation that large-scale revenues might be bad for their development prospects. The experience of The Netherlands following the discovery of the Groningen gas field gave rise to concern regarding Dutch disease. This was defined explicitly as the contraction in the nonhydrocarbon-traded sector following a real appreciation of the exchange rate. For The Netherlands, this meant a decline in manufacturing. However, economists also began to observe that for oil-exporting developing countries, it was agriculture that took the hit.

In the 1990s, it was the impact on government behavior that attracted attention. Large windfall revenues from oil or gas projects appeared to change government behavior, damaging growth and development. This ran parallel to the wider developments in economic theory on the economic theory of politics, theories of public choice, and principal-agent analysis that laid the ideological basis that government involvement in the allocation of resources would lead to misallocation.

Recently, several factors revived interest in the subject and also brought resource curse to a wider audience. For some time, nongovernmental organizations (NGOs) have expressed concern that oil or gas projects in poor developing countries have negative effects. The World Bank Group set up an "Extractive Industry Review" to consider the bank's role in such projects. The larger, more responsible oil and mining companies are also considering their impact on host countries and how to mitigate negative consequences. This is driven by concern for reputation in a world in which ethical considerations of investment are becoming paramount. It is also driven by a view that negative effects could threaten the economics of the project.

Another cause of this revival of interest is that a number of countries—Azerbaijan, Angola, Chad, Kazakhstan, Papua New Guinea, and numerous others—will soon receive large-scale oil or gas revenues. There is general concern about the impact and use of these revenues and what should be done to maximize the benefits and minimize negative effects.

Finally, many institutions financing such extractive projects realize that the potential for resource curse from the projects threatens the viability of the projects and, hence, ultimately the value of their investments.

2. THE PRESENCE OF RESOURCE CURSE

A large body of work has attempted to establish a negative relationship between the abundance of natural resources and poor gross domestic product (GDP) performance. For the most part, the evidence appears to support the negative link. However, such work is sensitive to the period chosen and there is evidence that before the 1970s resource-abundant countries grew faster than resource-deficient ones. Also, because of price fluctuations certain periods are likely to distort results.

There are a few dissenting voices to this empirical work based on methodology rather than the overall conclusions, although some challenge both methodology and conclusions, arguing that there is little long-term evidence that natural resource-abundant countries generally underperform. Rather, natural resources have played an integral role in the success of many industrialized countries. Oil exporters have also provided support for resource curse. For example, Mikesell identified Saudi Arabia, Mexico, and Venezuela as having lower than average annual growth rates of GDP.

Apart from a negative growth impact, another aspect is the prevalence of poverty and the observation that oil- or gas-rich economies have a poor record in poverty alleviation. This is a central theme of much of the literature. Another dimension is that oil or gas projects are associated with greater conflict in a society. Large-scale resource revenues create a pot that is worth fighting for. Whoever is in power is better able to plunder that pot. Such revenues also generate higher levels of military spending. The projects often alienate local populations, especially if separatist tendencies already exist. Conflict is also linked to poverty. War and strife (e.g., inflation) are regressive since they hurt the poor more than the rich because the poor cannot mitigate against the effects. Fighting or expectations of fighting also absorb resources that could go to improving economic performance and alleviating poverty.

Another characteristic linked to resource curse is that natural resource abundance slows political change and entrenches regimes. Thus, as Auty noted, an abundance of resources significantly weakened nascent democratic institutions, repressing political parties so that power is weakly contested, public finances are opaque, and corruption by both the elite and the bureaucracy is rampant.

A final dimension is the regional impact of the projects. Although the effect at a national level might be debated, because of the local impact, damage is often done especially in terms of both the environment and human rights. Meanwhile, the benefits appear to flow to central rather than regional authority.

However, although evidence of the presence of resource curse appears convincing, there are examples of countries for which there appears to have been a "blessing" rather than a curse, including Botswana, Chile, Malaysia, Norway, Indonesia before 1997, Australia, and Canada. A study by Stevens provides strong support for the fact that these countries avoided the curse. Thus, there is a case for dropping the automatic use of the term resource curse and instead using the term resource impact and considering whether the outcome was a curse or a blessing.

3. TRANSMISSION MECHANISM BETWEEN LARGE RESOURCE REVENUES AND POOR PERFORMANCE

There is a huge literature on why countries might suffer a curse rather than a blessing following a large inflow of oil or gas revenues.

3.1 Long-Term Decline in Terms of Trade

This argues that over time, the revenue would buy less imports of capital goods, thereby inhibiting development-creating investment. However, this is controversial and the empirical and theoretical arguments have been much contested. Also, it is difficult to understand how a slow decline in prices would explain the sort of deterioration in economic performance associated with resource curse. Economies constantly face changing relative prices. If they take place gradually, economies might be expected to adjust unless other factors are at work.

3.2 Revenue Volatility

Oil or gas revenues are very volatile, especially driven by wide fluctuations in prices over relatively short periods of time. Certainly, there is plenty of empirical support for the existence of such volatility. Thus, Mikesell found that between 1972 and 1992, regions with high primary export shares experienced periods of trade volatility two or three times greater than those of industrial countries during the same period. Fluctuating revenue profiles make it difficult to pursue a prudent fiscal policy. This creates problems ranging from aggravating investor uncertainty to "stop go" spending policies. There is also concern that windfall revenues would be consumed rather than invested, although empirical evidence throws doubt on this since the permanent income hypothesis argues that windfall gains would be more likely to be saved and invested than consumed. However, intuitively it is easy to imagine fluctuating revenues, in the absence of effective stabilizing measures, creating problems for government fiscal policy and macroeconomic management more generally.

3.3 Dutch Disease

Originally, this referred to the appreciation of the real exchange rate as a result of inflation from spending the revenues plus an appreciation of the nominal exchange rate as the domestic currency attracted higher demand. The result was a contraction in the non-oil or gas sector. However, over time, the meaning has become wider to encompass all negative macroeconomic effects associated with resource curse.

These changes of meaning are reflected in the differing theoretical approaches. Corden and Neary split the impact of an oil boom into a resource movement effect and a spending effect. In the former, a higher marginal product in the booming resource sector draws resources from other sectors. Thus, factors move into oil, bidding up wages and causing other sectors to contract. The spending effect occurs when, as a result of the windfall, demand rises in both tradables and nontradables. Since prices of tradables are determined by the international market, greater demand is met by higher imports. However, prices of nontradables increase relative to tradables and so resources shift from tradables to nontradables. This approach forms the basis for much of the subsequent analysis in the literature.

Over time, other dimensions of Dutch disease—defined simply as a contraction of the tradable sector—have emerged. Thus, one argument is that subsidies used to protect nonresource tradables that are weakened by the boom aggravate the sector's problems and eventually become unsustainable. Another is the "leapfrog effect," in which governments tend to miss the labor-intensive phase of industrialization and move straight to the heavy, capital-intensive phase with negative effects for the tradable sector. It has also been suggested that resource-poor countries accumulate social capital faster than resource-rich countries. One explanation for this is that limited natural resources promote early industrialization forcing earlier urbanization.

In addition to the controversy regarding the exact nature of the theory of Dutch disease, there is also

controversy regarding whether empirical evidence supports its existence. A number of country studies provide support, including studies on Bolivia, Venezuela, Mexico, Brazil, Peru, Algeria, Ecuador, Indonesia, Nigeria, and the United Kingdom. However, often the situation is complicated by extraneous factors. For example, in Botswana a Dutch disease effect on agriculture was complicated by two serious droughts in the 1980s. Others conclude that Dutch disease was not the major factor explaining the growth pattern and that some deindustrialization may be the result of adjustment toward a new equilibrium. Furthermore, for oil-exporting countries, after the oil boom of the 1970s, although agriculture did contract, manufacturing tradables actually expanded in most cases.

3.4 Crowding Out Effects

This is a variation of the resource movement effect of Dutch disease in which the project, which is large relative to the rest of the economy, stakes first claim on scarce resources. The rest of the economy finds it difficult to secure the factors needed to develop. This is especially relevant in smaller countries when the size of the investment project is large or where there are strong regional dimensions to an economy and the project is focused in one region. It is also important in the transition economies, in which transition has destroyed the old productive base and a new and private tradable sector is struggling to emerge.

3.5 Increasing the Role of the State

If the "Washington consensus" is believed, then greater government intervention in an economy is a bad thing. Given that in most legal regimes the revenues accrue to the government, this inevitably invites greater government intervention. Many of those writing on resource curse view a major part of the explanation of the phenomenon as relating to the role of the government. Also, many argue that it was good government decisions that enabled a few countries to avoid the worst excesses of resources curse. As with crowding out, the issue is especially sensitive in the transition countries simply because the objective of transition is to reduce government intervention in the economy.

However, although many would agree with these sentiments, why oil- or gas-exporting countries appear more vulnerable to policy failure requires further explanation. Several strands of argument are possible. The first argues that large windfalls lead to poor general decision-making by governments. The development of such projects raises expectations. This pressures government to "do something" that encourages speedy responses. Often, quick, ill-coordinated decisions are bad decisions. Also, spending revenues too quickly is more likely to introduce distortions into the economy, if only because there is less chance for the economy to adjust naturally. Having more money to play with weakens prudence and normal procedures of due diligence. Thus, making the "right choices" seems somehow less important. Of particular relevance is when governments decide on capital spending without consideration of the recurrent spending implications. Finally, because in the first instance the revenues accrue to government, decision-making is concentrated in fewer hands than if a greater number of economic agents are involved in decisions regarding how to spend any windfalls.

Another strand of explanation concerns corruption and rent seeking. Having large amounts of money to play with increases the temptation for corruption and rent seeking on the part of the decision makers. However, this is a complex and controversial area. Some argue that corruption evolves from a clash between traditional values and foreign norms. Thus, in Africa, bureaucrats are under an obligation to share the benefits of their office with their kin, a characteristic common to a great many other societies. If this is the case, then because the oil or gas projects in developing countries almost invariably involve foreign companies, this alone helps explain why resource-rich countries may experience greater levels of corruption.

Rent seeking concerns how people compete for "artificially contrived transfers." The result is that the expenditure creates no social value and, arguably more important, distorts markets and hence the way the economy operates. There are several explanations for why rent seeking is greater in countries with large oil or gas revenues. Most obvious, the larger the public purse, the less noticeable the leakage to interest groups, although this may not be true if the result is a "feeding frenzy." Also, rent seeking is greater because wealth is concentrated in the public sector or possibly a small number of companies. Thus, most of the rents created are channeled by bureaucrats, most of whom are members of the politically dominant group. There is general agreement that rent-seeking behavior produces undesirable results for the economy and imposes significant losses on many economies. Such behavior distracts attention from goals of long-term development

toward maximizing rent creation and capture. Also, rent seeking creates powerful lobby groups that block much needed economic reforms. Rent seeking also makes it more difficult for governments to adjust spending when faced with revenue fluctuations.

However, such issues are far from clear-cut. An important aspect is what happens to the rewards of corruption and rent seeking once they have been secured. If used for productive investment within the economy, the outcome will be different than if used for conspicuous consumption or to enhance foreign bank accounts. After all, at the risk of being simplistic, the whole point of capitalism is for a few greedy, selfish, and unscrupulous individuals to "grab" resources and put them to productive use to secure even more resources. The "robber barons" who developed the U.S. economy in the 19th century (very resource rich by any definition) indulged in some very dubious business practices, not least in the oil sector.

Often, the nature of investment decisions following large windfall revenues creates problems for economic progress. This is in part explained by the bad decision-making previously discussed. Also, often investments made by resource-rich governments fail to develop the productive base of the economy. Furthermore, if borrowing funds these investments, this compounds the problem since the loan must be serviced and repaid. It has been argued that much of the link between resource abundance and negative growth may be attributed to the aftereffects of a borrowing spree by natural resource exporters in the 1970s.

A specific area that combines poor decision-making and bad investment decisions in many resource-rich countries concerns the industrial policy adopted. Resource-rich countries tend not to promote a competitive manufacturing sector. Rather, they adopt an industrial policy based on import substitution involving subsidy and protectionism. Often, resource revenues prompt the introduction of heavy subsidies, usually based on an infant industry argument. These become unsustainable when revenues decline. However, their creation and the subsequent rent seeking create powerful interest groups that oppose their removal. Regarding protectionism, there is a positive correlation between dependency on primary products and a closed trade regime. Once these elements of policy—subsidy and protection—are in place, continuing resource revenues reduce the incentive to create competitive manufacturing industries. Since many development economists regard competitive manufacturing as a principal source of technological progress, this has serious implications for economic progress.

3.6 Sociocultural and Political Impacts

The sociopolitical dimension is central to the discussion, although its role is much disputed. It has been suggested that resource-rich countries have weaker institutional capacities, and that the endogenously determined institutional capacities are key determinants of growth. In contrast, others argue that although differences in income can be explained by the effectiveness of institutions, this is a legacy of colonialism and there is no discernable impact from natural resource abundance.

Previous explanations related to the macroeconomic causes of the curse. Countries that avoided the curse did so by the adoption of specific policies to minimize damage. Of key interest is why such policies were or were not adopted and why the various political elites allowed them. The answer lies in the nature of the society and its politics and, as such, is a crucial part of the explanation of the transmission mechanism.

Some of the discussion revolves around rentier societies in which there is no connection between reward and effort. Thus, easy access to riches bred idleness and provided no incentive to accumulate or innovate. Although this explanation seems attractive, it is not conducive to analytical rigor and smacks slightly of stereotyping that, apart from any other potential problems, condemns the recipients to a life with no hope of redemption.

A particularly fruitful area of sociopolitical enquiry concerns the categorization of states as developmental or predatory. A developmental state has two elements—an ideological and a structural component. The ideological component involves the ruling elite adopting developmentalism as its prime objective. Legitimacy is derived from the ability to deliver development, implying both growth and poverty reduction. This elite establishes a form of ideological hegemony—via the ballot box or less desirable means—over the rest of the society. The structural component involves the capacity to implement policies "sagaciously and effectively" that will deliver development. Apart from technical capabilities, this also requires a strong ruling elite that can resist pressure from powerful, short-sighted private interests and/or some form of "social anchor" that restrains any temptation to use its autonomy in a predatory manner. A key element is the realization that developmental states can still fail. Despite the

"right" ideology and the limits of predation in place, the capacity to implement effective policies might not be enough to manage certain problems, especially if these are exogenous. Such problems may be driven by external shocks, mistakes, or bad luck.

In contrast, in the predatory state, the ruling elite only have interest in feathering their own nests. They are "takers," not "makers," who have little or no interest in promoting economic development. Thus, their objective is to plunder the state's resources to benefit the few, as opposed to a developmental state in which the objective is to benefit the many. There are no checks and balances to inhibit plundering the national wealth. Indeed, the less constraints on the ruling elite, the more they can plunder. This process becomes self-reinforcing. The greater the wealth accumulated, the more resources they have to inhibit, buy off, or control any potential checks and balances.

Large-scale oil revenues encourage predatory states by a number of mechanisms. As with rent seeking and corruption, they make the pot larger and therefore more attractive to take. The more there is to steal, the greater the temptation to do so. Revenues generate resources to suppress opposition to the predation either by the use of force or by simply buying off potential problems. Finally, the major powers, dependent on oil, may provide support and legitimacy for predatory regimes to ensure the continued flow of oil, thereby protecting the regime both from internal problems and from threats from neighbors. Iraq is a classic example.

It is tempting to attribute poor performance by resource-rich countries to a lack of democracy. However, the argument is debatable. Thus, as Mikesell suggested, democracy does not ensure good government, nor are all oligarchies poorly governed. Furthermore, the democratic credentials of countries that avoided the curse—Malaysia, Indonesia, Pinochet's Chile, and indeed Botswana—are highly debatable.

This lack of clarity on the transmission mechanisms presents a serious challenge. If it is not clear what causes the problem, then treatment is likely to prove equally elusive. However, this is important if investment in oil and gas projects under consideration is to go forward.

4. HOW RESOURCE CURSE MAY BE AVOIDED: THE THEORY AND THE PRACTICE

The literature presents a number of options for avoiding resource curse.

4.1 Leave It in the Ground

Although this cannot be regarded as a serious option since a less ethical company will always be willing to make money, irrespective of damage or reputation, it does raise issues about the speed of the project and hence the consequent production/revenue profile.

Conventional project economics argues that from a commercial standpoint, the faster the development, the better. The slower the production, the lower the total present value of the revenue stream. However, common sense argues that a slow, steady trickle of revenues should be easier to handle than a sudden surge. If revenue management problems are easier to deal with, the resource movement effect and crowding out are likely to be less severe. Furthermore, slow development allows the development of a service industry to maximize backward linkages from the project, whereas swift development must be based on imports. The problem is persuading investors to slow development. Although this goes against conventional project economics, if rapid development generates a curse and if this threatens the project because of the political fallout, slower development may well be justified on purely commercial grounds. There is a major gap in the literature on resource curse regarding how the phenomenon might affect the economics of a project. One obvious connection is that insofar as the curse embodies increased conflict within a country, this may directly physically threaten the project's equipment and infrastructure. This would be especially relevant where the curse had a strong regional dimension as a result of the local impact of the project.

A variant on this theme is the nature of the revenue flow. Differing forms of extractive agreements—concession based on royalty and taxes, joint ventures, production-sharing agreements, service agreements, etc.—will create differing profiles of revenue flow. Thus, although development of the project may still be rapid, the revenue flow could be controlled by the agreement rather than leaving it to the vagaries of some sort of stabilization fund. An example is the case in which a project is expected to create a short-lived revenue peak for the government. In this case, the project may flatten the peak by providing revenue upfront before the project is fully operational and afterwards when the project is finished. Of course, this implies that it is the company that controls the revenue flow rather than the government, which raises issues of sovereignty.

4.2 Diversification

One obvious solution is to reduce the importance of the oil or gas sector in the economy by diversifying the economy. Since the early 1970s, oil-exporting countries have paid lip service to diversification away from dependence on crude oil exports. However, the record in general has been very poor, with huge amounts of public money being poured into inefficient and uncompetitive industries.

Several factors explain these failures. Dutch disease and crowding out cause a contraction of an existing non-oil or gas tradable sector. It is not surprising if the mechanisms that destroy do not also inhibit the rise of a new tradable sector. Another explanation favored by countries is that trade restrictions imposed by the Organization for Economic Cooperation and Development exclude their outputs. Although both arguments have validity, there is a third explanation. In most cases, the diversification policy has consisted of governments trying to pick winners. Thus, much of the new economic activity has been government-owned enterprises. Governments are generally very bad at picking winners, although there have been exceptions.

The only effective diversification derives from private sector investment, although governments can play an important role in this process, as can be seen from the experience of the Asian tigers. Thus, one solution to the problem of diversification is to maximize the revenue flow to the private rather than the public sector (discussed later). This was certainly an important part of the explanation for why Australia, Canada, and the United States avoided the curse. Of course, this raises important issues of income distribution and private economic power leading to concentrations of political power. This in turn is associated with the complex and controversial issue of how much political repression is worth accepting for an improved standard of living—a constant theme of more general literature for centuries.

Another means to encourage successful diversification relates to trade policy. Countries that avoided the curse pursued a policy of trade openness with the rest of the world coupled with a deliberate policy of exchange rate depreciation to ensure the competitiveness of non-oil or gas exports.

4.3 Revenue Sterilization

A key factor is the macroeconomic policy adopted by the government, specifically the policy adopted to try to neutralize the impact of the large windfall revenue inflow on the rest of the economy. This has several dimensions. The first is the need for fiscal prudence to prevent the revenues from translating immediately into greater aggregate demand and inflation. This requires the government to resist spending pressures and either accumulate budget surpluses or channel the revenues into some form of fund. An important subissue is that the government should avoid using promises of more revenue to increase borrowing.

Another important strand to macropolicy is to avoid exchange rate appreciation. Several options are available. The first is to avoid a fixed exchange rate policy when domestic inflation of any magnitude exists. The second is to encourage currency devaluation. However, this has no effect on the real exchange rate without appropriate demand management policies. One option is for the central bank to purchase foreign exchange to hold down the nominal rate. One common factor in the case of countries that recently avoided a curse is that they all experienced significant depreciation of the real exchange rate as a result of explicit policy choices.

4.4 Stabilization and Oil Funds

There is general agreement that trying to stabilize spending to ensure stable and moderate growth is important. This can be done in several ways. Trying to stabilize the international price is no longer a serious option and probably never was despite a large literature on commodity stabilization schemes. One possible solution is to stabilize prices by using futures markets to lock in price stability. However, there are problems. For very large producers, the danger is that their involvement would simply swamp the markets. Also, such insurance costs. Although having guaranteed prices in a falling market would result in praise for the decision makers, in a rising market there is bound to be criticism. Another solution might be to use the extractive agreement to smooth revenue flows, as previously suggested.

Another mechanism is to create an oil fund. These can fulfill three functions. They can be used to insulate the economy from large revenue windfalls by investing them outside the domestic economy. The fund can also be used to stabilize revenues by setting a price assumption for budgetary purposes. If prices exceed this, then the fund absorbs the windfall. If prices are lower, then the fund's assets are used to meet the budget. Finally, such funds can be used to put assets aside for future generations. Thus, the

revenue is accumulated and only the earned income is spent, leaving the capital sum intact to allow future generations to benefit from a depletable resource.

In recent years, the role of such funds has become extremely controversial, especially within and between the World Bank and the International Monetary Fund. There are two schools of thought. One argues that the conditions for such funds to work are such that if they are present, the revenue windfall can be managed within the normal budgetary process. Hence, the funds are not needed. The other argues that even if conditions are not ideal, the creation of such funds makes some contribution to avoiding the worst excesses of resource curse if only because of the role of the fund in dampening spending expectations.

4.5 Investment Policy

The investment policy adopted by the government for the revenues will play a crucial role both in avoiding many of the macroeconomic pitfalls described previously and in encouraging the process of economic diversification by creating alternative sources of value added to the depleting resource sector. Ideally, investment is best carried out by the private sector. Although, as previously noted, the experience of the robber barons in the United States in the 19th century illustrates that this does not guarantee everyone lives happily ever after. However, private sector investment is more likely to be sustainable than when the government tries to pick winners.

4.6 Political Reforms Needed to Carry Out the Corrective Politics

A key component of any solution is the need to remove corruption and contain rent seeking. However, this is far from straightforward. For example, any cleanup campaign must start with an agreement on a common standard of morality. Options to solve problems of corruption include incorporating controls in legal codes and encouraging independent (and fearless) enquiries. Another solution is to reduce rent-seeking opportunities by depoliticizing resource allocation. Thus, deregulation allows free markets, taking politicians and bureaucrats out of the resource-allocation process. The theory with empirical support is that competition squeezes out illicit margins. Although this is probably true, it still requires the creation of competition, and often simple deregulation and privatization is not sufficient to achieve this.

Much is also made of the need for transparency as an aide to reduce corruption. Although this can have a role, there is a danger that its effectiveness can be overstated. For example, many have called for the oil companies to disclose how much money they provide to the government so the public can see how much money is being "diverted." This neglects the fact that often very large sums accrue to the national oil company as a result of the agreements, especially joint ventures and production sharing agreements. It also bestows upon both the public and the local media a degree of sophistication that may be misplaced. For example, a major problem is that because the cash flow associated with projects is uneven over time, considering only a few years could give the impression that the oil companies are taking a huge share of the proceeds, ignoring the earlier years of negative cash flow. This could revive the "obsolescing bargain" and lead to unjustified pressures to renegotiate terms. This is not to argue that there is no place for transparency, just that it is far from a sufficient condition to reduce corruption.

There is also the issue discussed previously that from an economic standpoint, it is not so much the presence of bribes and rent that is the problem. Rather, it is how they are used. Leaving aside moral issues and issues of income distribution, bribes and rent productively invested result in a successful capitalist economy. Economic problems arise when these are diverted into conspicuous consumption or foreign bank accounts.

Following from previous discussion, a key issue relates to securing a developmental rather than a predatory state as a way of minimizing the risk of suffering a curse. Developmental states require two components: a strong elite with an ideology that is conducive to economic development and whose interests are aligned with the population and also the institutional capacity to implement the necessary policies without being captured by narrow private interests. In contrast, a predatory state lacks the linkages to the population or any other constraint that prevents the elite from ransacking the economy.

Securing an alignment of interests is crucial. One route to achieve this alignment is democracy in which the government must satisfy public interests or be voted out. However, although this may seem to be the optimal route morally, as argued previously, it is not the only route. Thus, for example, Chile under Pinochet and Malaysia, Indonesia, and indeed Botswana as a "one-party democracy" managed the alignment quite well.

Of course, what constitutes an ideology that is conducive to economic development is also a moot point. Clearly, simple mantras based on "leave it all to the market" are inappropriate. However, the experience of those countries that avoided a curse suggests that although their governments did intervene extensively, they did so to push the economy toward a market-based system linked into the global economy and toward the encouragement of a vibrant private sector.

The ability of the elite to translate this ideology into effective policies is a function of institutional capacity. Clearly, both international financial institutions (IFIs) and investing companies have a role to play in training those elements of the bureaucracy that could play a key role in policy making. This can be done using a combination of seconding expertise (assuming acceptance by the governments) and training nationals using short professional training courses and more long-term basic provision of university postgraduate training.

All this is feasible and can be made to work such that the economy progresses. The real problem arises when the ruling elite lacks any social anchor and alignment with the public and the state has become predatory. Insofar as the state may be dependent on IFI good will for financing and debt management, the IFIs can be in a position to provide the constraints required to reign in predatory behavior, although this obviously raises sensitive issues of sovereignty. However, such an outcome has the seeds of its own destruction since once the resource revenues begin to flow, the IFI leverage seeps away and with it the constraint. One possibility is that the strengthening of civil society might begin to provide both the constraints and the alignment necessary to switch the state from being predatory to developmental. Again, the IFIs and companies can help strengthen civil society, but this means more than just giving the NGOs more money, although what precisely it does mean is uncertain and needs further thought and analysis. Also, it assumes the predatory state will tolerate the growing influence of civil society, which for obvious reasons of its own survival it may not. Another option is for the oil company to exert pressure on the government to curb predation, but this would only be effective if all the main oil companies were willing to take the same stance.

Of course, one solution is for the international community to refuse to invest in predatory states. However, this raises the prospect of an unethical portion of the industry investing to add further to the plunder of an economy by its own elite. It also raises the prospects of creating ever more failed states harboring resentment against the West.

5. CONCLUSION

Clearly, the issue of what causes a curse rather than a blessing and how to enjoy the latter and avoid the former is extremely complex. The literature is divided on both causes and cures, although there is a growing consensus that essentially it has something to do with governance and the answer lies more in political economy than macroeconomic analysis. Hence, the key question is not what was done in terms of policy decisions but, rather, why was it done? The outcome of the debate will be of crucial importance in terms of determining the future magnitude, type, and direction of investment in the oil and gas sector.

SEE ALSO THE FOLLOWING ARTICLES

Development and Energy, Overview • *Experience Curves for Energy Technologies* • *Investment in Fossil Fuels Industries* • *Oil-Led Development: Social, Political, and Economic Consequences* • *Oil Price Volatility* • *OPEC, History of*

Further Reading

Auty, R. (ed.). (2002). "Resource Abundance and Economic Development." Oxford Univ. Press, Oxford, UK.

Corden, W. M., and Neary, J. P. (1982). Booming sector and Dutch disease economics: A survey. *Econ. J.* **92**.

Meyer, J., Chambers, B., and Farooq, A. (eds.). (1999). "Development Policies in Natural Resource Economies." Elgar, Cheltenham, UK.

Mikesell, R. (1997). Explaining the resource curse, with special reference to mineral-exporting countries. *Resour. Policy* **23**(4).

Mkandawire, T. (2001). Thinking about developmental states in Africa. *Cambridge J. Econ.* **25**.

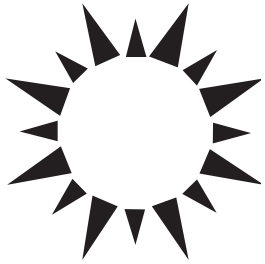
Sachs, J. D., and Warner, A. M. (1999). The big push, natural resource booms and growth. *J. Dev. Econ.* **59**.

Stevens, P. (2003). Resource impact: Curse or blessing? A literature survey. *J. Energy Literature* **9**(1).

Stevens, P. (2003). Resource impact: A curse or a blessing? *Electronic J.*, www.cepmlp.org.

Reuse and Energy

MARKO P. HEKKERT
Utrecht University
Utrecht, The Netherlands



1. **Relationship between Energy and Materials**
2. **Product Reuse in Perspective**
3. **Examples of Product Reuse**
4. **Environmental Benefits of Product Reuse**
5. **Implementation Barriers of Product Reuse**
6. **Conclusion**

Glossary

improved material management Intentional actions that lead to a reduced amount of material that is used to fulfill the same function. Since materials require a large amount of energy in production, this often leads to energy savings.

material cascading Improved material management measure. It is a form of material recycling but differs due to the reduction in product quality that can be produced with the secondary material. An example of material cascade is paper used for printing and writing that is recycled for newsprint.

material recycling Improved material management measure; reuse of materials. The original product is destroyed in order to reuse the materials for new products. This can be the same type of product or a new type of product. Energy savings often occur since the manufacturing of virgin materials often requires more energy than producing recycled materials.

product reuse Improved material management measure. This is a type of material recycling, but it leaves the original product intact. The major advantage is that energy used in the construction of a product is not lost due to destruction of the product. It is the material management measure with the highest carbon reduction potential in the packaging sector.

Since the use of fossil fuels is an important cause of increased greenhouse gas concentrations, many strategies for reducing human greenhouse gas emissions focus on a reduced use of fossil fuels. To reduce the use of fossil fuels, a wide range of options can be applied: (i) improvement of energy efficiency, leading to a reduction of the energy consumption per unit of activity or product; (ii) development and application of renewable energy sources, such as biomass, hydropower, solar, and wind energy; (iii) a shift in the use of fuels from resources with a high carbon content (coal) to resources with lower or even no carbon content (natural gas and uranium); (iv) decarbonization of fuels and flue gases, for example, by recovering carbon dioxide from energy conversion processes and storing it outside the atmosphere; and (v) improvement of material management.

This article focuses on reuse of materials, which can be categorized as an improvement of material management.

1. RELATIONSHIP BETWEEN ENERGY AND MATERIALS

Improvement of material management is not commonly incorporated in greenhouse gas reduction policies. This is strange since several studies indicate that this option has a large potential and is often economically attractive. Before the results of these studies are described in more detail, the reason why material efficiency may lead to reduced greenhouse gas emissions is discussed.

Materials are produced in the industrial sector and a major portion of fossil energy consumption is used in this sector. The industrial sector (excluding refineries) consumed approximately 33% of the total primary energy use in 1995 for developed countries. When materials are used more efficiently, less materials need to be produced and, therefore, less energy is needed in the production stage. Consequently, more efficient material management is likely to lead to reduced emissions of greenhouse gases.

A limited number of studies have been performed on the subject of improved material management and climate change. A few case studies from Utrecht

 © 2004 Elsevier Inc. All rights reserved.

University on plastic packaging and fertilizers show that a significant greenhouse gas emission reduction is possible by more efficient material use. Patel studied a limited amount of material management options for the plastics sector in Germany. He showed that these options may lead to a 24% reduction of CO_2 emissions from the German plastics-producing industries. Furthermore, the U.S. Environmental Protection Agency showed that management of waste materials presents many opportunities for greenhouse gas emission reduction. Elliott showed that recycling, an improved material management option, may lead to major savings in industrial energy use. In The Netherlands, the MATTER project, involving the Dutch Energy Research Centre and several universities, was started to model the Western European energy and materials system and determine, based on technological improvement options, the greenhouse gas emission potential in these systems in the period 1990–2050. Gielen concluded, based on these model results, that "the potential for emission reduction in the materials system seems to be of a similar magnitude as the emission reduction potential in the energy system." The importance of these studies for understanding the potentials for greenhouse gas emission reduction is stated by the United Nations Department of Policy Coordination and Sustainable Development. Based on these studies, one may conclude that the potential of improved material management on greenhouse gas emission reduction seems to be significant, and obtaining more insight into the exact potential and the ways to utilize this potential is important.

2. PRODUCT REUSE IN PERSPECTIVE

The environmental pressure that is related to a product or a material is the sum of the environmental pressures related to the production of that material or product as well as the other stages in the life cycle of the product, such as production of raw material and waste management.

Climate policy should therefore focus on greenhouse gas emissions at all stages of the product life cycle: raw material extraction, material production, product manufacturing, product use, and disposal of products and waste management.

Changes in any of these stages can lead to reduced product-related greenhouse gas emissions. However, it is important to realize that measures that lead to reduced emissions in one part of the product life cycle may lead to increased emissions elsewhere: In order to reduce product-related greenhouse gas emissions, the whole life cycle needs to be taken into account.

Figure 1 gives an overview of the different life-cycle stages and the measures that can be taken to improve the efficiency of material production and use.

Figure 1 shows that in all stages of the life cycle, measures can be taken to increase material efficiency. The improvements can be categorized in many different ways. Here, the following classification is used:

1. Technical measures in the individual life-cycle stages
2. Management of the entire material product chain

Technical measures in the individual life-cycle stages include measures such as increasing the efficiency of processes (where innovations in technological development and process design can be used to reduce the amount of materials used in the process), improved product design [where using innovative production methods facilitates the use of thinner material (layers) and using other materials in the product manufacturing stage may lead to reduced emissions; important substitution criteria are the use

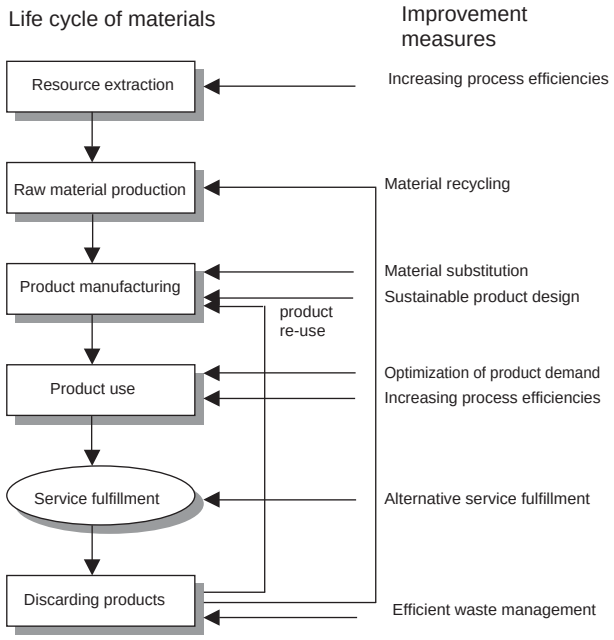


FIGURE 1 Material life cycle and improvement measures.

of energy-extensive materials, materials with a long lifetime, and materials that are suitable for reuse], and more efficient waste management (where there is optimal energy recovery from waste materials).

Management of the entire material product chain covers measures such as material recycling, material cascading (using discarded materials for the highest possible functions), extension of the useful life of products (modular design and increased fitness for repair and maintenance), sustainable product service systems (measures that lead to more efficient service fulfillment due to the substitution of products by services or combinations of products and services), and reuse (e.g., refillable plastic or glass bottles).

Reuse involves the collection and cleaning of products in such a way that they can be used again in the same form and shape as they were used in the previous life cycle. In such a way, both the structural properties of the material and the energy required for shaping the product are preserved. Therefore, to make a clear distinction between reuse and recycling, it is more appropriate to speak of product reuse and material recycling. In the case of material recycling, only the structural properties of materials are used again in new product life cycles. In this case, it is not necessary to use the recycled material for the original product. For example, polyethylene theraphtalate (PET) bottles can be collected and recycled in two different ways. The first is to clean the PET bottle and reuse it as a bottle (product reuse). The second is to shred the bottle and recycle the PET material for a different product, such as crates or fleece jackets (material recycling). Material reuse and material recycling differ from material cascading by the fact that in a material cascade the product that is made out of recycled material is always lower in quality than that of the original product. Optimal material cascading policy tries to optimize the required material characteristics for a product and the factual material characteristics. An example of a material cascade is to produce a solid wooden beam for construction purposes from a tree, then recycle the material after its product lifetime as wooden floorboards or window frames, and recycle the material again after this life cycle as fiberboard. In this case, the required wood quality of the products declines over the consecutive life cycles. Another example is the use of virgin wood pulp for the production of high-quality printing and writing paper, reuse of the wood fibers for lower quality printing and writing paper, and finally use of the fibers for newsprint production.

3. EXAMPLES OF PRODUCT REUSE

To make product reuse possible, a good working collection system is necessary. Traditional waste collection systems are not sufficient since the original products are damaged and contaminated. Therefore, separate collection systems are necessary. There are a few examples in which these systems have been set up successfully.

The first example is the reuse of beverage bottles in several European countries. In The Netherlands, several types of beverage bottles are separately collected, cleaned, and reused. Examples are the reuse of 0.33-liter beer bottles (88% of the beer volume packed), 1- and 1.5-liter soda bottles (56% of the volume packed), and 1-liter milk bottles (10% of milk volume packed). The beer bottles are made out of glass, the soda bottles are made out of PET, and milk bottles are produced from polycarbonate (PC). In packaging terminology, every time a bottle is reused it is called a trip. Thus, the trip number is the number of times a bottle is reused. On average, these trip numbers are approximately 20 for beverage bottles, implying that 20 times less material is needed to fulfill the same packaging function.

The second example is the reuse of transport packaging, also popular in many European countries. Transport packaging is defined as packaging that is used to bundle and protect multiple products for transport purposes. Examples are corrugated transport boxes, crates, and pallets. The latter two are often reused.

Crates and boxes are normally used to pack loose products, such as fruit and vegetables, meat, and product parts. (Crates are also used for transportation of bottles. No improvements for these crates have been modeled.) The crates and boxes can be used for one-way shipping or function as a returnable (reuse) package. The one-way crate is mostly made out of wood or corrugated board and is normally used to pack fruit and vegetables. One-way boxes are always made out of corrugated board. This box has an average volume of 40 liters and weighs 0.8 kg. Multiple-trip plastic crates can be used instead of one-way crates and boxes. The plastic multiple-use crate is made from high-density polyethylene (HDPE), weighs 2 kg, and has a volume of 40 liters. Using plastic returnable crates requires a closed-loop transport system. A third party often manages the logistics of this transport system. The crates currently used have a lifetime of 5–10 years and the average trip rate is 25/year.

Pallets are also used as single-trip packaging and in reuse systems. A vast majority of the pallets used in

Europe are made out of wood. Wooden pallets are popular because they are cheap, have a large carrying capacity, and are easy to repair. Moreover, they are very suitable for production of small series with deviating sizes. Approximately one-third of the pallets used in Europe are returnable. A single-use pallet weighs approximately 17 kg and a multiple-use pallet weighs approximately 25 kg. On average, returnable wooden pallets make 20 trips. Obviously, much material saving occurs when returnable pallets are used instead of single-use pallets.

In addition to wooden pallets, plastics pallets are also used. They are often used in the food industry because they are easy to clean due to their smooth surface. The most common material for plastic pallet production is PE, but sometimes PC is used. Pallets made out of PC are stronger than PE pallets. Plastic pallets are especially suitable as multiple-use pallets. They weigh approximately 30 kg. There is no consensus about the number of trips that can be made with a multiple-trip plastic pallet; it varies between 30 and 100 trips.

The shift from one-way pallets to returnable pallets requires major changes in pallet administration and management. The use of multiple-trip pallets requires a pallet pool. A pallet pool is an organization that manages the transit of the pallets between the various users. Standard sizes for pallets are introduced to make the pallet applicable for many users. In Europe, several organizations are active in the management of pallet pools, of which Europool and Chep have the largest market shares.

A third example is the reuse of product components of machinery such as televisions and copy machines. When the product has reached the end of its lifetime, it is collected and taken to a deconstruction facility, where the reusable parts are separated from the other components. The reusable parts are transported to the production facility and can be reused in new televisions and other equipment.

Two types of collection systems are generally used. The first system is completely organized by the producer. In order to make such a system work, the product is sometimes leased and not sold. The advantage is that the producer remains the owner of the product. When the product is not desired anymore by the customer, the producer (and owner) collects the product. This arrangement is typical for large, expensive products such as copy machines. In the second system, the collection is organized by either municipalities that collect electrotechnical products or retail traders that collect old products when selling new products. In The Netherlands, the latter system exists. When buying new products, a tax is paid to cover the expenses of dismantling the old products.

The success of both systems in terms of product reuse is strongly dependent on the efforts made by the producer to take back old products and reuse product parts. When this responsibility is low, the chances that product parts are actually reused are slim. In these cases, the maximum benefit is that the collection of these products offers potential for material recycling (plastic and metal recycling).

Producers that take product responsibility very seriously include Xerox, which takes back and dismantles its copy machines, and Bosch, which is making efforts to reuse the motors from returned machinery. The lifetime of a motor in electrical equipment is generally much longer that the lifetime of the total product. Reuse of motors is therefore a very good option.

4. ENVIRONMENTAL BENEFITS OF PRODUCT REUSE

The environmental benefits of reuse can be calculated. To do so, the original life cycle needs to be compared with the reuse life cycle. Specifically, all extra processes necessary to make reuse possible need to be analyzed (e.g., separation, cleaning, and transport of used products) and compared to all the processes that do not take place any longer in the reuse life cycle (e.g., material production, material transport, and product manufacturing). The environmental benefit can then be calculated by subtracting the environmental impacts of the obsolete processes from the environmental impact of the additional processes. The environmental benefits of reuse are often quite large since the environmental impact of the separation, cleaning, and transport of used products is often much smaller than the impact of material production, material transport, and product manufacturing. This is emphasized by Table I, which gives an estimate of energy use data for typical processes that occur in the production and use of PET bottles used for packaging of soft drinks. Here, it is assumed that the environmental effects of energy use are a good indicator of the total environmental effects. Table I shows that the dominant processes in terms of energy use are material production and product manufacturing. Since, product reuse leads to reduced material production and product manufacturing, the environmental benefits of reuse are easily explained.

Hekkert *et al.* analyzed a number of improvement options to reduce the environmental impacts of packaging materials. Product reuse was taken into account as an important improvement option. The results of their analyses indicate the potential environmental benefits of product reuse in relation to other options to improve a material cycle, such as material recycling, material substitution, and more material-efficient product design. In these analyses, the scope of studying environmental benefits is reduced to studying carbon emissions. In Fig. 2, the results of the analysis for packaging products are depicted.

Figure 2 needs to be read as follows. Each horizontal line represents one specific improvement option. Specifically, it represents the carbon emissions that are saved by replacing a traditional package with an improved packaging concept. Every horizontal line is numbered and the numbers relate to Table II, in which the traditional packaging concept and the improved packaging concept are listed. The length of each line indicates the size of the reduced carbon emissions, and the position of the line in relation to the vertical axis gives information on the life cycle costs associated with this improvement option. Figure 2 shows that the costs of many options have a negative value, which means that over the complete life cycle, costs are saved when the option is implemented.

TABLE I

Use of Energy for Typical Processes in the Production and Use of PET Bottles

Process	Energy use (GJ/1000 bottles)
Material production	8.50
Product manufacturing	1.22
Filling of bottle	0.20
Cleaning of bottle	0.45
Transport of bottle	0.10

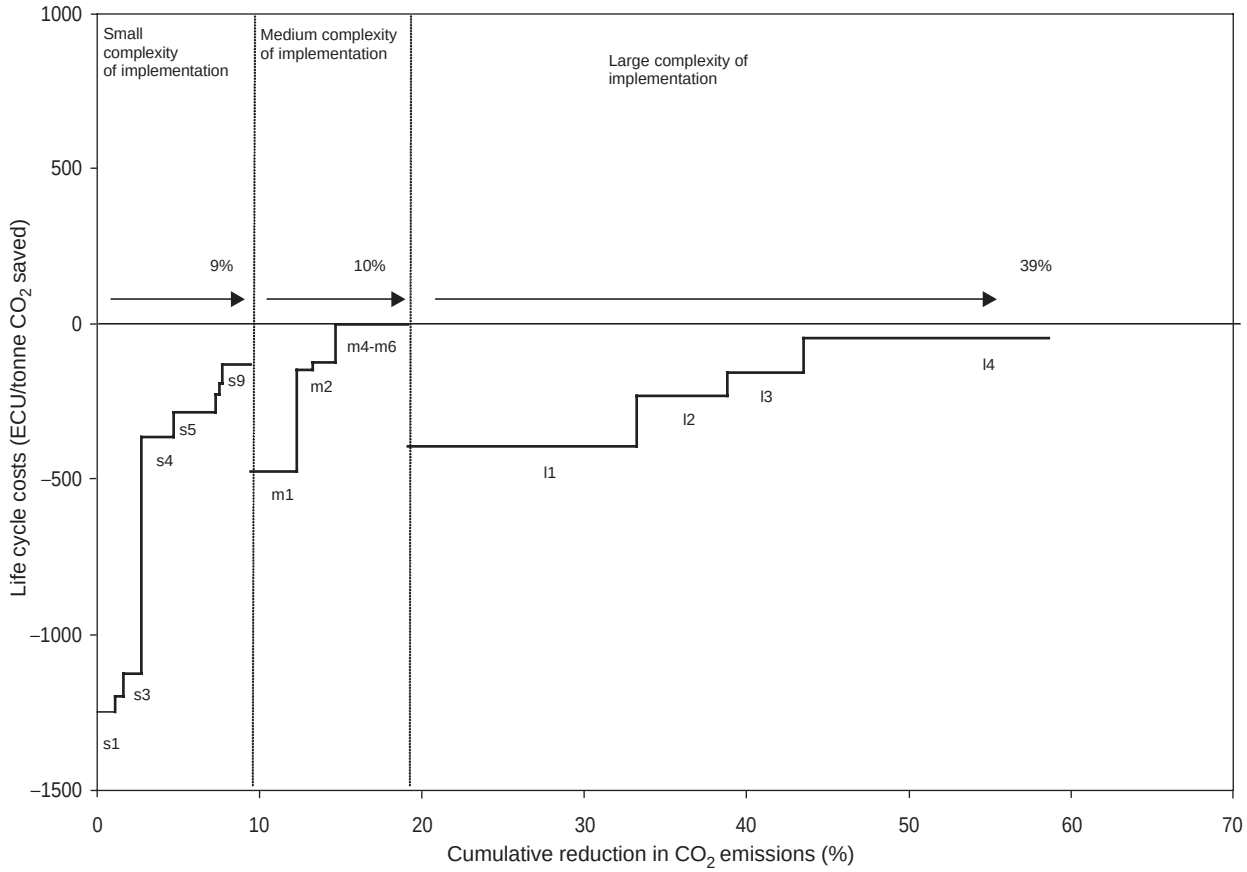


FIGURE 2 A supply curve for the reduction of CO_2 emissions by improved use of materials for primary packaging in Europe. The data refer to Table II.

TABLE II

Potential CO_2 Emission Reduction and Costs of Improved Material Management for Packaging in Western Europe[a]

No.	New packaging concept	Old packaging concept	CO_2 emission reduction (%)	Cost (ECU/tonne CO_2 saved)
S1	PP film thin	PP film	1.1	−1200
S2	Cardboard box, light	Cardboard box	0.5	−1200
S3	LDPE film thin	LDPE film	1.1	−1100
S4	Honeycomb food can	Steel food can	2.0	−360
S5	Light glass bottle, large	Glass bottle, large	2.4	−280
S6	Glass jar, light	Glass jar	0.2	−280
S7	Steel beverage can, light	Steel beverage can	0.2	−230
S8	Aluminum beverage can, light	Aluminum beverage can	0.2	−190
S9	Light HDPE bottle	HDPE bottle	1.8	−130
M1	PET bottle, one way	Glass bottle, large	5.2	−470
M2	Steel beverage can, light	Aluminum beverage can	1.1	−150
M3	PP cup	PS cup	1.4	−120
M4	PET bottle to be recycled	PET bottle, one way	1.0	0
M5	Recycled HDPE bottle	Light HDPE bottle	2.9	0
M6	Cardboard blister	PVC blister	2.2	0
L1	Pouch	Liquid board	14.1	−390
L2	Glass bottle, small refillable	Glass bottle, small	5.6	−230
L3	Pouch	HDPE bottle	4.7	−160
L4	PET bottle reuse, recycled	PET bottle, one way + PET bottle to be recycled	15.1	−40

[a] Abbreviations used: HDPE, high-density polyethylene; LDPE, low-density polyethylene; PET, polyethylene theraphtalate; PP, polypropylene; PS, polystyrene; PVC, polyvinyl chloride.

Figure 2 is divided into three parts. The first part indicates the options that can categorized as easy to implement; it contains mostly options that use a different product design to use lighter or less materials in packaging design. The second part contains options that are more difficult to implement, such as material substitution options. The final part is dedicated to reuse options and product substitution, which are the most difficult options to implement. The figure clearly shows that the total carbon emissions that can be expected from reuse (measure l2 and l4) are very large compared to the other measures. Thus, these results indicate that product reuse can be a very efficient option to reduce the environmental impacts of material life cycles compared to competing measures.

5. IMPLEMENTATION BARRIERS OF PRODUCT REUSE

It was previously mentioned that product reuse can be categorized as a rather difficult measure to improve a material life cycle. The reason for this is the fact that product reuse requires an additional infrastructure to return the products to the manufacturer or distributor. In this respect, product reuse can be seen as an ultimate form of extended producer responsibility. However, since manufacturers often strive to distinguish themselves from competitors, in every product group a whole range of different product designs can be found. This requires an extremely intense separation and returning system in which all individual products can be traced back to a certain manufacturer. In practice, this is likely not feasible.

Two types of systems do seem feasible. The first system was discussed previously. Here, products are not sold but leased and the manufacturer remains the owner of the product. In this case, the manufacturer knows the location of the product at all times and can take the product back at the end of its lifetime. Typical products for such a system are large, expensive (electronic) equipment for the professional market, such as copy machines. An important drawback of this system is that the speed of technological change in these product types is very high. This

results in a situation in which recycled product parts are often outdated. In these cases, material recycling is the second best option. However, one might question whether for material recycling, returning the product to the owner is preferable to a general collection system. The latter is much easier to organize and can benefit from scale issues.

The second system uses the concept of standardization to create an efficient returning system. In this case, several manufacturers and distributors agree on a standardized product that is manufactured by these parties. Examples are returnable beer and soda bottles in several European countries that do not differ in size, material, or design but only differ in the type of labels used. This system makes it possible for all bottles to be distributed over all manufacturers. This simplifies the setup of the returning system, but it requires the commitment of all major producers and distributors and it decreases the capabilities of the individual producers to develop new competitive product designs. The difficultly in developing this type of returning system is that it requires a large amount of companies (retail organizations, distributors, and manufacturers), the details of the system need to be agreed on by all parties, and much administration is necessary to keep the system going. This leads to transaction costs, which companies prefer to keep as low as possible.

One way to decrease the transaction costs is to make one party responsible for the complete setup and maintenance of the system. Other parties only have to join the system. Such an example is the use of returnable pallets in Europe. In this case, supranational standards have led to fixed pallet sizes, which makes it possible to rotate pallets over distributors throughout Europe. The pallet pools are organized by pool companies that own and maintain the pallets. Detailed tracking systems have been set up to calculate the pallet costs for each company in the pallet pool. The advantage for the distributors is that they do not need to own a large amount of pallets, which saves investment costs and space.

6. CONCLUSION

This article has shown that product reuse is an important method for improved material management, in addition to other methods such as material recycling, improved product design, and improved waste management. Important examples of current reuse practices can be found in the packaging and electronics sectors. The environmental benefit of product reuse seems quite large due to the fact that major environmental impacts from material production and product manufacturing are largely reduced. The problem with product reuse is that it requires a new and often additional distribution system that requires the involvement of many parties that need to align their actions. Transaction costs are regarded to be high, which can be seen as an important barrier for implementation.

SEE ALSO THE FOLLOWING ARTICLES

Corporate Environmental Strategy • Energy Efficiency and Climate Change • Energy Efficiency, Taxonomic Overview • Industrial Energy Efficiency • Material Efficiency and Energy Use • Recycling of Metals • Recycling of Paper

Further Reading

Belkom, A. V. (1994). Kunststof pallets in opmars. *Kunststof Rubber* 12.

Chappin, M. (2003). "Decomposition Analysis of the Dutch Beverage Packaging Waste." Utrecht Univ. Press, Utrecht, The Netherlands.

Duin, R. V., Hekkert, M. P., de Beer, J., and Kerssemeeckers, M. (1999). "Inventarisatie en Classificatie van EZP opties." Bureau B&G, Emst, The Netherlands.

Elliott, R. N. (1994). "Carbon Reduction Potential from Recycling in Primary Materials Manufacturing." American Council for an Energy-Efficient Economy, Washington. DC.

Gielen, D. J. (1999). "Materialising Dematerialisation." Delft University of Technology, Delft, The Netherlands.

Hekkert, M. P., Joosten, L. A. J., Worrel, E., and Turkenburg, W. C. (2000a). Reduction of CO_2 emissions by management of material and product use: The case of primary packaging. *Resour. Conserv. Recycling* **29,** 3–64.

Hekkert, M. P., Joosten, L. A. J., and Worrel, E. (2000b). Reduction of CO_2 emissions by management of material and product use: The case of transport packaging. *Resour. Conserv. Recycling* **30,** 1–27.

Patel, M. (1999). "Closing Carbon Cycles: Carbon Use of Materials in the Context of Resource Efficiency and Climate Change." Utrecht Univ. Press, Utrecht, The Netherlands.

Price, L., *et al.* (1998). Sectoral trends and driving forces of global energy use and greenhouse gas emissions. *Mitigation Adaption Strategies Global Change* **3,** 263–391.

Renia, H. M., and Sikkema, R. (1991). "Houtbijprodukten in Nederland: Een Studie naar de Omvang, Samenstelling, Herkomst en Bestemming van Resthout, Houtafval, oud Hout en Niet-Geoogst Hout." Stichting Bos en Hout, Wageningen, The Netherlands.

Turkenburg, W. C. (1995). Energy demand and supply options to mitigate greenhouse gas emissions. *In* "Climate Change Research: Evaluation and Policy Implications." Elsevier, Amsterdam.

U.S. Environmental Protection Agency (1998). "Greenhouse Gas Emissions from Management of Selected Materials in Municipal Solid Waste." U.S. Environmental Protection Agency, Washington, DC.

Worrell, E., *et al.* (1995a). An approach for analysing the potential for material efficiency improvement. *Resour. Conserv. Recycling* **13,** 215–232.

Worrell, E., Meuleman, B., and Blok, K. (1995). Energy savings by efficient application of fertilizer. *Resour. Conserv. Recycling* **13,** 233–250.

Worrell, E., *et al.* (1997). "Potentials ands Policy Implications of Energy and Material Efficiency Improvement." United Nations, Department for Policy Coordination and Sustainable Development, New York.

Risk Analysis Applied to Energy Systems

HERBERT INHABER
Risk Concepts
Las Vegas, Nevada, United States

1. History of Risk Assessment
2. Risk Comparisons
3. Voluntary versus Involuntary Energy Risks
4. Early and Late Fatalities
5. Dimensions of Risk Effects
6. Occupational versus Public Risk
7. Risk Perception
8. The Zero–Infinity Dilemma
9. Storage and Backup
10. Other Forms of Risk: Climate Change
11. Air Pollution Risk
12. Risk Aversion
13. Uncertainty
14. Catastrophic versus Noncatastrophic Risk
15. Comparative Energy Risk
16. Conclusion

Glossary

aversion A feeling of repugnance toward something with a desire to avoid or turn from it. Risk aversion plays a part in setting environmental standards.

perception Quick, acute, and intuitive cognition. In terms of risk analysis, it has been noted that perception often varies from objective studies.

pollution Environmental contamination with man-made waste. Air pollution is a major source of energy-related risk; water pollution is less so.

radiation Energy radiated in the form of waves or particles. This is a prime concern in terms of nuclear power risk.

risk Possibility of loss or injury.

risk per unit energy or electricity production To put different energy or electricity forms on an equal footing in terms of risk, analysts often divide the risks by a unit energy output, such as a megawatt-year. Other units can and are used. A large fossil fuel or nuclear plant might have a capacity of 1000 MW. If it operated without interruption for 1 year (highly unlikely), it would produce 1000 MW-years during that time.

uncertainty This may range from a falling short of certainty to an almost complete lack of conviction or knowledge. Risk analyses often contain more uncertainty than policymakers desire.

The production of electricity from any form of primary energy has some, in certain cases substantial, risks to human health and environmental effects. A balanced assessment of different forms of energy requires comparison of their environmental effects with those of their alternatives.

1. HISTORY OF RISK ASSESSMENT

Clearly, our prehistoric ancestors had some appreciation of risk, even in terms of energy. Fire, the main energy source of the time, could destroy unless it was carefully watched and controlled. In a broader context, staying far away from a saber-tooth tiger was safer than having it close at hand. However, until some numerical measure was used, what our ancestors experienced was not what we would call risk analysis. Thus, the phrase implies a type of calculation, from the simplest to the Monte Carlo probabilistic techniques of advanced risk studies.

The Babylonians were apparently among the first to use numbers extensively, often in conjunction with astronomical events. Consequently, it is little surprise that in approximately 3200 BC they apparently had the first risk assessors. These assessors did not deal with health or environmental issues. Although periodic flooding by rivers is now regarded as environmental, the Babylonians believed that these events were created by the gods and not amenable to any

 © 2004 Elsevier Inc. All rights reserved.

type of quantitative analysis. Issues such as health, death, food supply, and the like were planned by the supernatural, not to be interfered with by humans.

Rather, these forerunners of modern consultants offered advice on other risky decisions, such as deciding where to locate important buildings and when and whom to marry. These areas were clearly not energy related but have some resonance with modern life. Newspaper horoscopes, claimed by their developers to be based on mathematical formulae, are still read by tens of millions of people. In many areas of the world, the dates of major events such as marriages or contract signings are regulated by mathematics. In the United States, it has been claimed that President Ronald Reagan had the dates of some of his decisions set by the same formulae. The purpose of these calculations is to avoid risk—marrying the wrong person or losing money on a business decision.

In the Western world, the industrial revolution generated new diseases or caused the increased spread of old ones. It was realized that some diseases could be prevented or alleviated by human intervention. The scientific study of scrotum cancer in chimney sweeps by Percival Potts in Britain in the late 18th century was perhaps the first epidemiological study. This study has been cited in the first chapter of many books on risk. Since fireplaces were the main source of home heating in Britain at the time, the study was obviously energy related and perhaps the first scientific risk study dealing with energy. Epidemiology has undergirded many energy risk studies since then.

Curiously, however, although detailed in gruesome detail in books by Charles Dickens and others, the terrible diseases and accidents suffered by coal miners of the early 19th century, often children, apparently did not receive the same scientific scrutiny despite the fact that the total loss of health and life by coal miners was much greater than that of chimney sweeps. Chimney sweeps visited the homes of the wealthy, who never saw the suffering coal miners.

This omission still holds a lesson for modern energy risk analysis: The largest sources of risk for a given energy technology are sometimes ignored or glossed over as "common knowledge." This point will be illustrated later.

By the late 19th century, industrial hygiene, epidemiology, and toxicology were established as scientific areas. However, general application of this research to factories and industrial enterprises had to wait a number of decades as social consciousness grew. By the 1930s, considerable evidence on the relationship between occupational toxins and human health had accumulated. However, the Great Depression prevented widespread use of these results in regulation. Only certain large firms, such as DuPont, whose social consciousness dated back to its gunpowder manufacturing days, made use of these results.

At that time, it was known that air pollution could cause health problems. Pittsburgh, Pennsylvania, was known as the "smoke city" for the heavy pall it often endured. However, smoke was associated with industrial prowess and jobs, so little was done. During a visit to India in the 1970s, I was given a matchbook showing a factory belching smoke. It is unlikely that such a matchbook would be printed today.

The next era in risk assessment occurred indirectly with the advent of the atomic bomb in 1945. It was quickly recognized that the power of destruction could be harnessed for peaceful purposes, and major governments adopted nuclear power programs. However, the risks associated with producing this form of energy were unlike those calculated in the past. This in turn led to vast research on nuclear risks, much more than for other energy technologies. To date, more has been spent on analysis of nuclear risks than on all other energy sources combined—probably at least an order of magnitude greater.

It was quickly recognized that a major accident at a nuclear reactor could have significant consequences for health, causing cancers and perhaps other diseases. A 1957 study by Brookhaven National Laboratory for an extreme case predicted hundreds of immediate and long-term deaths from radiation. However, defenders of the nuclear industry pointed out that these large numbers were not representative of likely operation.

In the 1970s, drawing on experience from the National Aeronautics and Space Administration, the concept of probabilistic risk analysis (PRA) for complex systems such as nuclear power plants was developed. The methodology was incorporated in the so-called Rasmussen Report of 1974–1975, which showed that the average risk to health from nuclear reactors was extremely small. As in many phases of life, the average is considerably different from extreme events or conditions.

Since that time, dozens of PRAs have been performed for nuclear reactors (as well as for chemical and other large industrial facilities). In addition, because energy has assumed such a major role in modern life—the famed economist Julian Simon once called it the "master resource"—thousands of studies on various aspects of energy risk

have been completed. It would take far too much space to even list, let alone discuss, them here.

2. RISK COMPARISONS

Risk comparisons are used extensively in risk analyses for a simple reason: Lay audiences often have difficulty understanding the complex terminology and mathematics of many energy risk studies. This has been shown in both psychological studies and public meetings on energy. In this section, the reasons for risk comparisons are discussed in a general way. Later, specific comparisons are evaluated.

Regulations in the United States and many other industrialized countries require public meetings to be held to discuss new energy projects. When the subject of risk arises, observers have noted that there is little, if any, substantive contribution from the public. When roentgens, rems, and parts per billion are mentioned by speakers, the relatively low level of scientific understanding by the public becomes apparent. To remedy this, risk comparisons are often used. These comparisons can be divided into two categories: (i) comparisons between different substances and activities and (ii) comparisons between different technologies.

The first category cuts across different aspects of life and exposure. For example, a widely used comparison devised by Richard Wilson of Harvard University states that each of the following activities presents the same risk, measured as the increased risk of death within 1 year by 1 in 1 million. This means, for example, that cycling the specified distance will increase that person's chances of death within a year by one in a million, a very small chance:

- Smoking 1.4 cigarettes
- Riding 10 miles by bicycle
- Eating 40 tablespoons of peanut butter
- Drinking 30 12-ounce cans of diet soda containing saccharine
- Living within 5 miles of a nuclear reactor for 50 years

It is evident that calculating each of these risks contains a wide variety of data and assumptions. For example, nonsmokers (approximately three-fourths of the population) who smoke one or two cigarettes almost certainly will not develop lung cancer as a result. The estimate of one chance in a million of death is based on smokers. Peanut butter by itself will not cause death; any aflatoxin it contains might.

Some risk experts have objected to these types of comparisons, claiming that the activities specified have little or nothing to do with each other. Peanut butter has no known connection to nuclear power. They state that comparing the risks of different forms of energy is more illuminating to the public than contrasting the risks of smoking with nuclear power. The problem is reduced to ensuring that the entire fuel or energy cycle for each energy technology is used to allow for a fair and reasonable comparison.

3. VOLUNTARY VERSUS INVOLUNTARY ENERGY RISKS

Chauncey Starr was one of the first to quantify the difference between voluntary and involuntary risks. It is self-evident that risk, at least per hour of activity, is greater for hang-gliding and mountain climbing than for driving an automobile or sitting behind a desk. Starr estimated that voluntary activities were between two and three orders of magnitude more risky than involuntary ones, depending on the precise activities considered.

How does this translate to energy? With some difficulty. The problem is that the line between voluntary and involuntary activities, which is clear in other contexts, is often unclear in the energy situation. For example, consider a coal miner in Wales or China. He may be descended from a long line of miners and is expected to go into the mines when he reaches working age. In principle, coal mining, with its known high risks, should be a voluntary activity. Those who do not wish to accept its risks are able to take up other less risky occupations. However, in the example cited, it is not completely clear that it is voluntary.

In terms of public risk, again the line is often indistinct. Certain areas of the nation have higher levels of air pollution, mostly due to energy activities, than others. Los Angeles has its infamous smog. Again, in principle, residents who know about these air pollution risks and are concerned about them can move to other more bucolic areas of the nation, making this air pollution exposure voluntary. Some do. However, many are confined by employment and family ties to the polluted region, suggesting that the risk is more involuntary.

The question becomes more acute with regard to nuclear power, clearly the most highly publicized form of energy risk. When a new nuclear reactor is to be built, residents have the option of moving away if

their concerns about risks are great. No doubt some have done so in the past, although there do not seem to be sociological studies indicating the percentage of those who depart. It is likely to be small for the reasons mentioned previously as well as others, such as difficulty in selling a home and sheer inertia.

Thus, although using the dichotomy of voluntary versus involuntary risks can be useful in other aspects of risk analysis, it is difficult to employ in the energy context.

4. EARLY AND LATE FATALITIES

Deaths from energy technologies can be characterized as early or late. The division applies primarily to nuclear power. For other energy technologies, such as coal- or gas-fired electricity, deaths occur almost immediately after an accident or air pollution incident. For example, construction and mining accidents, which are shown elsewhere in this article to be a substantial part of some energy technology risks, produce deaths either immediately or within a few days. The effects of air pollution from fossil fuels are often calculated by the increase in deaths after a pollution alert. When a dam fails, deaths are immediate.

Nuclear power is different in the sense that if a major accident occurs, the release of radiation could cause deaths from cancers in three or four decades. The Three Mile Island (TMI) accident in 1979 did not produce enough radiation to generate this effect. Although the radioactive release from Chernobyl in 1986 was substantially greater than that of TMI, the World Health Organization and International Atomic Energy Agency calculate that the late fatalities will be small. However, they will be nonzero.

It is evident that both early and late fatalities from a nuclear reactor accident should be combined. The Rasmussen Report was criticized for not including late fatalities with its estimates of early deaths from a reactor accident.

5. DIMENSIONS OF RISK EFFECTS

Decision makers want one or two numbers in a risk comparison. Scientists would prefer to have many dimensions of a subject quantified. The tug-of-war between the two viewpoints has persisted in the past and is likely to continue indefinitely.

These generalities can be made concrete by specific examples. Inhaber determined that the risks of the energy systems he considered could be divided into two categories: public and occupational. This division is described by some as involuntary and voluntary, respectively, As noted earlier, the division is not so clear-cut. He then noted that some risk estimates were in deaths and others were in nonfatal diseases and accidents. A number of studies estimated the loss of life expectancy for fatalities, accidents, and diseases due to energy production. The data for both occupational and public effects could then be combined into man-days lost.

The next step was to provide a common denominator for the amount of energy or electricity produced by the 11 systems he considered. He chose a megawatt-year, but in principle any multiple of this could have been used.

Since he was aiming his research at decision makers, he combined the occupational and public risks, both in man-days lost per megawatt-year, into a total risk. For this, he was criticized by some for oversimplification and combining two dissimilar entities.

One of the valid criticisms of risk analyses of that era was that they did not take into account the ages of victims of accidents or diseases. Clearly, someone of age 80, perhaps already in poor health, has fewer days lost if felled by an air pollution spike than a vigorous 21-year-old killed in a mine accident. Because governments have an interest in the efficacy of health interventions, health scientists have devised the "quality days of life lost," which takes into account the previous points concerning relative ages of victims of energy risk as well as their state of health. This measure provides a more accurate view of energy risk effects.

Researchers at Clark University in Massachusetts took another approach. They attempted to characterize technologies on a scale that encompassed 12 descriptors, including social, physical, and biological aspects of risk: intentionality, spatial extent, concentration, persistence, recurrence time, population at risk, delay of consequences, annual mortality, maximum potentially killed, effects across generations, maximum potential nonhuman mortality, and experienced nonhuman mortality. Although some of the descriptors are self-explanatory, others would require a lengthy discussion of their assumptions.

However, the problem still remains: Is there a rational and defensible method of combining 2 or 12 measures, or should they be presented separately? For example, if technology A has a lower risk than technology B on 3 of 12 measures but also has a higher risk than technology C on 5, how do we rank

A, B, and C? It is unlikely that decision makers would have the skill or patience to sort through all the possible permutations and combinations. For this reason, at least some risk researchers will continue to combine different measures of risk into one or two easily understood numbers.

6. OCCUPATIONAL VERSUS PUBLIC RISK

In the energy risk studies discussed elsewhere in this article, risk to workers tends to be higher than risk to the public. In a completely rational world, emphasis should be on reducing worker risk, with less concern about public risk reduction.

Al Hoskin of the U.S. National Safety Council evaluated the relative risks to workers and the public in Superfund (abandoned waste) sites. He found that the risk to workers for the cleanup was higher than the existing risk to the public.

In reality, developed nations place the emphasis on public risk. As noted later, this may be due to the media's concentration on disasters, real or potential, affecting the public. Occupational risk to individual workers in the United States tends to receive little attention, except by the agencies devoted to the problem or when a disaster occurs, as in the case of the Pennsylvania coal miners trapped below the surface in water in 2002. The miners were rescued with no loss of life. At the same time, news came of major coal mine disasters in China, where scores were killed. Little attention was paid to these incidents.

7. RISK PERCEPTION

The following is one of the oldest questions in philosophy: If a tree falls in a forest and nobody hears it, does it make a sound? Like many philosophical issues, there is no simple answer. The analogous question in risk analysis is, If a study shows that chemical or toxin has a high (or low) risk, and this study is either unknown to the public or ignored by it, what then is the risk?

As Slovic, Fischhoff, and Lichtenstein (SFL), three of the founders of risk perception studies, have noted, subjective judgments, by both experts and lay people, are a major component of any risk assessment. Expert judgments are often obscured in impenetrable mathematics. Judgments of lay people are considerably easier to detect.

In a classic study that went from the social science literature to popular magazines, SFL noted the divergent risk estimates of four groups: members of the League of Women Voters (LWV), college students in Oregon, a business and professional club, and risk experts. SFL considered 30 sources of risk, from nuclear power to lawn mowers and vaccinations. Many of these activities rarely appear on risk lists, but they all produce some hazards. Only a few of the categories related to energy risk.

The two largest divergences between estimates of the lay group and experts were in nuclear power and X-rays. The experts rated nuclear energy as 20th riskiest of the 30 activities, the LWV and college students rated it highest, and the business and professional people rated it 8th. X-rays were rated approximately 20th in riskiness by the laypeople but 7th by the experts. The rankings of many of the other risk sources were approximately the same for laypeople and experts.

Although there are many possible explanations for why the estimates varied so much, the variation results from the benefits received from the two activities. X-rays are viewed as having considerable benefits in medical diagnosis that outweigh the risks of radiation. The benefit of nuclear power is electricity generation, but there are other ways of producing this energy form, such as coal, gas, and oil. In the United States, approximately 80% of electricity is nonnuclear. (Admittedly, there are methods other of detecting internal diseases besides X-rays, such as nuclear magnetic resonance, but these were not included on the SFL list.) Thus, the benefits and risks of nuclear energy must be compared not to those of swimming or lawn mowers but to those of other ways of generating electricity.

One aspect of the SFL study that deserves mention is the ranking of technologies as opposed to concentration on absolute values of risk, either real or perceived. Because the values in their studies ranged over many orders of magnitude, it was clearer to rank technologies rather than to discuss the absolute values.

Experts measure risk as *X* deaths or diseases per activity, but risk is perceived according to SFL by whether it is voluntary, controllable, certain to be fatal, catastrophic, immediately manifested, and unfamiliar. Some of these are discussed elsewhere in this article.

Most of the perception aspects of risk can be described by two factors: dread (related to aspects

such as certain to be fatal and catastrophic) and how unknown the risk was (related to involuntariness and newness). On these scales, nuclear power had relatively high scores for both factors, even among the experts.

How are risk perceptions formed? The media clearly plays a major role. For example, pesticides ranked approximately eighth in the list of 30 activities noted previously. Some years ago, concerns about the pesticide Alar were expressed on a widely watched television show. Would a repeat of the poll have shown higher perceived risk for pesticides? Risk assessors sometimes talk about "risk of the week" in reference to the media publicizing a seemingly new hazard on a regular basis. On the other hand, editors of newspapers and television news shows would respond that they are merely reporting on what interests the public.

8. THE ZERO–INFINITY DILEMMA

Students of mathematics know that multiplying zero by infinity produces an indeterminate value—that is, no number that can be identified. What relationship does this have to energy risk?

The problem arises in only one energy technology—nuclear. It is acknowledged by the nuclear industry that in the most extreme conditions, a reactor accident can emit considerable radiation, thus producing deaths from disease. The number of deaths is clearly not infinite but can be high.

The industry points out that, at least according to probabilistic risk analyses, the chance of such an accident occurring is extremely small, perhaps one in tens or hundreds of thousands of years of reactor operation. The chance is not zero, but very small. Thus the zero–infinity question: If one multiplies a very small number by an enormous one, what is the product?

Simple calculation shows that the result is a small risk. However, the estimate—and both the "zero" and "infinity" are no more than estimates—is conditioned more by risk perception than by mathematical analysis. Many people think of the risks of nuclear power primarily in terms of the quasi-infinite number. They tend to ignore or disregard the other factor in the product—the estimate of likelihood. Thus, the risks of nuclear power may be, in some minds at least, overestimated.

Curiously, the catastrophic risks of a renewable resource, hydroelectricity, do not seem to be weighed in the same scale. The vast majority of hydro dams are as safe as any structure on Earth. However, there have been major failures over the years of some dams, in some cases causing hundreds of deaths in the valleys below. The zero–infinity dilemma does not seem to be applied to this energy form, although the deaths it has caused are real, not hypothetical.

9. STORAGE AND BACKUP

No energy system operates at 100% capacity all the time. Even in the best run systems, parts have to be adjusted and replaced, and unforeseen accidents can occur. On average, nuclear and fossil fuel systems generally operate in the 80–90% capacity range. That is, approximately 10–20% of the time a given plant in this group is unavailable for refueling, maintenance, or other reasons.

The vast majority of these shutdowns are planned. If fossil fuel and nuclear plants generally stopped on a random or arbitrary basis (these types of shutdowns can occur but are comparatively rare), it would be extremely difficult to have a reliable electricity grid.

The situation is different for at least some renewables. They require some combination of energy storage and backup—the supply of conventional energy when the sun does not shine and the wind does not blow. Consumers generally demand highly reliable electricity service. The proof of this is the torrent of phone calls that a utility receives from irate customers when their power is temporarily cut off due to weather-related accidents.

For hydroelectricity, categorized as a renewable energy source by the U.S. Energy Information Administration and other agencies, the entire system is storage. Thus, no further storage or backup is necessary. Using wood to produce electricity also does not require backup since enough wood can be gathered to provide as reliable fuel supply as fossil fuels. Transportation using corn-based fuels also does not require backup since the sun's energy is stored in the corn.

However, the variability of the sun and wind require some combination of storage and/or backup. Some quotations from the literature on solar and wind make the point:

"Natural gas serves as the backup heat when the solar energy is unavailable or insufficient."

"A global solar-energy system without fossil or nuclear backup would also be dangerously vulnerable to drops in solar radiation from volcanic events such as the 1883 eruption of Krakatoa, which caused

widespread crop failure during the 'year without a summer' that followed."

"Luz (solar thermal) units are hybrid systems with natural gas firing capacity added as backup. … PV systems (and wind systems as well) can be used in hybrid configurations with natural gas-based backup power, as in the case of solar thermal units."

"A wind-driven generator to convert wind energy to electrical energy, backup diesel generators (for when there is no wind). …"

"The principal disadvantage of solar is the dependence on variable sunshine hours and the resulting short capacity utilization as well as the need for energy storage and backup systems."

Thus, the question is, what fraction of backup and/or storage should be required for an intermittent system? Estimates have varied up to 20% of the original system for backup. This implies that up to 20% of the risk associated with conventional energy systems is attributable to solar or wind systems. Other estimates of storage and backup are lower.

Electricity cannot be stored, except in expensive batteries impractical for an electrical grid. It is true that over a sufficiently large area, the wind is always blowing somewhere, suggesting that wind power is continuous, not intermittent. However, no technology can instantaneously transfer wind power from North Dakota to Florida, for example. If and when such technology is developed, the need for backup and storage will disappear.

10. OTHER FORMS OF RISK: CLIMATE CHANGE

The risks discussed in this article are primarily to health, in the form of accidents or diseases. However, there is another less apparent form of risk—that of climate change. The greenhouse effect refers to the effect of certain trace gases in the earth's atmosphere so that long-wave radiation, such as heat from the earth's surface, is trapped. Much of the greenhouse effect is due to carbon dioxide. An increase from 280 to 350 parts per million has occurred since the beginning of the industrial revolution.

If the connection between increased concentrations of carbon dioxide and rising world temperatures is valid, it has been stated by many climatologists that the result will be more hurricanes, tornadoes, droughts, and other extreme events. In this sense, risks to humans will be greater, if the chain of assumptions is valid.

Renewables and nuclear electricity produce no carbon dioxide directly. However, there are indirect effects, albeit much less than the direct combustion of fossil fuels to produce electricity. The electric car is sometimes held up as an example of a system emitting no carbon dioxide, which is correct in a direct sense. However, approximately 70% of the electricity used to charge its batteries in the United States derives from fossil fuels. The carbon dioxide is emitted at the plants, not on the street.

The global climatic effect of increasing CO_2 levels is a factor in the comparison of different forms of energy. Worldwide emissions of CO_2 from burning fossil fuels total approximately 25 billion tonnes per year. Approximately 38% is from coal and approximately 43% from oil. Every 1000-MWe power station running on black coal produces CO_2 emissions of approximately 7 million tonnes per year. If brown coal (a less energy-intensive type) is used, the amount is much greater.

Figure 1 shows the estimated production of CO_2, both direct and indirect, for different energy sources. As noted elsewhere in this article, there are often wide variations in estimates of energy risk and their subcomponents. Also, the significant figures listed later for some of the data points should be treated with caution.

Coal has by far the highest values, and almost all reports on the subject confirm this result. The maximum direct production of CO_2 for coal is approximately 1000 g/kWh. Indirect production (coal mining, equipment used in mining and electricity production, etc.) is approximately 300 g/kWh, for a total of 1300 g/kWh. The minimum of the range is approximately 1000 g/kWh.

The range for gas is approximately half that of coal. The other four energy sources do not have

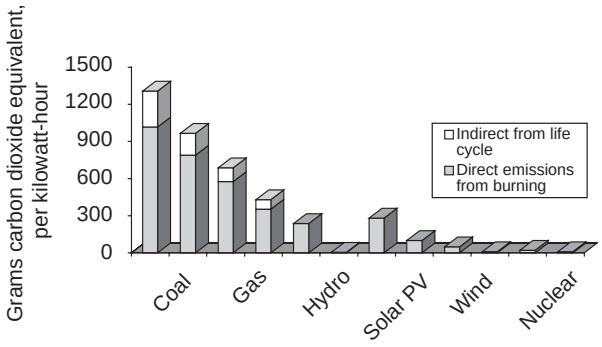


FIGURE 1 Greenhouse gas emissions from electricity production. Modified from the International Atomic Energy Agency (2000).

direct emissions. Hydroelectricity has values ranging from 4 to 240 g/kWh. One reason for the wide range of almost two orders of magnitude may be that the lower values refer to solid-earth hydro dams. Many hydro dams in developed countries are exemplified by the Hoover Dam, built in the 1930s near Las Vegas, Nevada. Much of the dam is concrete and steel, the making of which generates CO_2.

Earthen dams require soil and rock from the surrounding countryside, the transport of which generates only small amounts of CO_2. To give a representative value for hydro, indirect CO_2 production from dams built during approximately the past generation would have to be averaged.

Solar photovoltaic systems have values ranging from 100 to 280 g/kWh primarily because of the large amount of materials used per unit energy output in their construction. The values are substantially higher than those for hydro, by far the largest renewable energy system used throughout the world to generate electricity.

Wind energy has values from 10 to 50 g/kWh. Its lower range is thus higher than that of hydro, but its upper range is lower.

Nuclear has a range of 9–20 g/kWh. It is possible that the variability in nuclear data is lower than that of most other energy systems because so many studies have been performed in this area. Although nuclear, like the three other renewable systems described in Fig. 1, emits no direct CO_2, substantial amounts of metals such as steel and zirconium are used in construction of reactors. Also, given the extremely thick walls of most Western-built reactors, large volumes of concrete are used. In many countries (excluding those using the Canadian CANDU system), enrichment of the uranium for fuel rods employs fossil fuels for the required electricity.

These sources of CO_2 production are divided by the large amounts of electricity produced by typical reactors. Thus, nuclear has CO_2 values comparable to those of wind and considerably better than those of the other systems, both renewable and fossil fuel, considered in Fig. 1.

In summary, coal and gas have substantially higher emissions of CO_2 compared to renewables and nuclear. Of the renewables, solar photovoltaic seems to be highest and wind the lowest. Nuclear has values comparable to those of wind. There are considerable uncertainties regarding all the values.

11. AIR POLLUTION RISK

The largest source of risk to the public associated with fossil fuels has been air pollution effects. These effects can range from death to asthma attacks and other breathing difficulties.

However, a report in a 2002 issue of the Canadian *National Post*, quoting an article in the journal *Nature*, suggests that the estimate of 5000 Canadians (mostly elderly) dying each year from air pollution is too high by at least 1200 deaths and very likely by 3000–4000. This implies that air pollution deaths may be overestimated by approximately 30% and possibly by 150–400%. It is clearly too soon to determine the validity of these claims. Most comparative risk studies suggest that fossil fuels have higher risks per unit energy output than renewables and nuclear. If the study is validated, it would make fossil fuels closer on an absolute scale to renewables and nuclear per unit energy output. However, the relative position of these former fuels compared to their competitors probably would not change.

12. RISK AVERSION

There is a wide range of attitudes toward risk, as shown by surveys and other indirect devices. Risk-seeking individuals drink heavily, drive fast sports cars with unbuckled seat belts, smoke, sky-dive, climb mountains, and engage in other well-known risky behaviors. Risk-averse individuals read the newspapers avidly for the next food or activity to avoid, try vegetarian diets, wash their vegetables and fruits twice, bundle up heavily for cold weather, eat an apple every day, take their medicine when prescribed, move away from nuclear reactors if one is built nearby, stay away from people with colds, and try to elude all risks.

Some individuals are risk seeking for certain hazards and risk averse for others (e.g., smoking while practicing vegetarianism). Most people are somewhere in the middle, steering clear of the obvious risks but not having enough time to study and avoid all hazards.

Policymakers must account for both ends of this spectrum. If most people were risk seeking, there would be little need for any government regulation of energy or any other activity. If most were risk averse, virtually all economic activity would halt. Ironically, this would probably increase societal risk dramatically as society retreated to preindustrial age conditions.

Although many risk calculations are publicly available, there is little evidence that those who are risk averse and risk seeking make use of these data. Risk-averse people are unlikely to take more risks if they are told that they are concerned about infinitesimal hazards. Conversely, risk seekers know that climbing mountains is risky without information from a study. The inclination to join either group is a function of personality, not the results of risk assessments that people have read.

The risk averse and risk seekers seem to coexist happily in most situations. One situation in which they come together is when a waste site is being proposed nearby. Although risk estimates may indicate that the hazards of these sites are extremely small when regulations are satisfied, risk-averse and even risk-seeking citizens generally show strong concerns at siting meetings. As a result, most of these proposed sites are either abandoned or take many years to complete.

13. UNCERTAINTY

To paraphrase the poet, risk analysis, thy name is uncertainty. From the first energy risk analysis, uncertainty has made the results less convincing than they might have been. Some energy risk analyses have not discussed their uncertainties at all. Most others have done so, but sometimes the final results do not reflect the detailed uncertainty calculations in appendices. For example, in the Rasmussen Report, uncertainties were discussed in great detail in the body of the work. However, in the executive summary, which was read by far more people than were the elaborate calculations, uncertainties were downplayed. This led to criticisms.

On the other hand, the Slovic study on risk perception apparently did not show uncertainties at all. That is, if one or another group were ranking nuclear power or lawn mowers in terms of their risk, there would certainly be a statistical distribution. These distributions were never shown in the many reprints and reworkings of the Slovic study. Curiously, the work was not criticized for this omission, although it dealt with many of the risk issues that the Rasmussen Report and many other analyses dealt with.

It would take considerable space to discuss all the uncertainties associated with energy risk analyses, from data to statistical distributions and variations in assumptions. It is sufficient to state here that one of the main charges leveled at risk analyses is that the uncertainties are sometimes glossed over or not discussed at all.

14. CATASTROPHIC VERSUS NONCATASTROPHIC RISK

Many members of the public tend to think of energy risk in terms of catastrophes, those that kill or injure many at a time. However, at least for industrialized countries, most energy risk is noncatastrophic—that is, occurring to perhaps less than five people at a time.

Nonetheless, there have been catastrophes associated with energy production. Some of the major ones from 1977 to 1993 are shown in Table I. (A full listing to the present would take too much space.) Even for risk experts, many of the listings are likely to be unknown or, at best, barely recalled. In terms of coverage by Western news media, the reports tend to appear somewhere on page 13 of newspapers, with perhaps a 5-s news clip on television. Because many of them occur in Third World countries, the disasters are rarely followed up by stories explaining why they took place.

A number of comments can be made about Table I. First, many major accidents occurred prior to 1977. For example, the Vajont dam accident in Italy killed 2600–3000; The St. Francis dam failure in the United States in 1926 killed more than 400. A complete table would occupy many pages.

Second, many of the accidents took place in Third World countries, as noted previously. Because safety and environmental controls are usually better in industrialized nations, this is to be expected.

Third, although the Chernobyl nuclear accident undoubtedly received more media coverage than all the other accidents combined, it is clear from the table that the vast majority of catastrophic deaths were due to two major energy systems: fossil fuels (mostly gas and coal) and hydroelectricity. Because of their small size, renewables, with the notable exception of hydroelectricity, are unlikely to have catastrophic accidents. The largest recorded in the table was the Indian Machhu II dam failure in 1979.

Fourth, the question arises, can catastrophic and noncatastrophic risk be added in a reasonable way to produce an overall risk value for each technology? There is no simple answer to this question. However, it should be noted that catastrophic risk is small compared to ordinary risks for virtually all energy technologies. Thus, whether or not the catastrophic aspects are added is unlikely to change the comparative energy risk rankings.

TABLE I

Energy-Related Accidents, 1977–1993[a]

Location	Year	No. killed	Comment
Machhu II, India	1979	2500	Hydroelectric dam failure
Hirakud, India	1980	1000	Hydroelectric dam failure
Ortuella, Spain	1980	70	Gas explosion
Donbass, Ukraine	1980	68	Coal mine methane explosion
Israel	1982	89	Gas explosion
Guavio, Colombia	1983	160	Hydroelectric dam failure
Nile River, Egypt	1983	317	LPG explosion
Cubatao, Brazil	1984	508	Oil fire
Mexico City	1984	498	LPG explosion
Tbilisi, Russia	1984	100	Gas explosion
Northern Taiwan	1984	314	Three coal mine accidents
Chernobyl, Ukraine	1986	31 +	Nuclear reactor accident
Piper Alpha, North Sea	1988	167	Explosion of offshore oil platform
Asha-ufa, Siberia	1989	600	LPG pipeline leak and fire
Dobrnja, Yugoslavia	1990	178	Coal mine
Hongton, Shanxi, China	1991	147	Coal mine
Belci, Romania	1991	116	Hydroelectric dam failure
Kozlu, Turkey	1992	272	Coal mine methane explosion
Cuenca, Equador	1993	200	Coal mine

[a] From the Uranium Information Centre (2000).

15. COMPARATIVE ENERGY RISK

15.1 Can We Learn from Comparative Energy Risk?

In principle, comparing the risks of different forms of energy can guide both the public and decision makers to more valid decisions and attitudes. In practice, the conclusions are not so simple to draw.

Some observers, such as Andrew Stirling, have suggested that there is so much overlap between the results of energy risk analyses that few if any conclusions can be drawn. Figure 2 shows the results of approximately 30 energy risk analyses that he compiled. Certainly, the wide overlap between each technology suggests considerable uncertainty in conclusions.

Nonetheless, some crude inferences can be drawn even from Fig. 2. The fossil fuels generally have maximum values substantially higher than those of nuclear and the three renewables shown. Hydroelectricity is in the middle range.

In 1996, the Health Physics Society suggested that below a lifetime dose of 100 mSv (10 rem) (a unit of radioactive dose), "risks of health effects are either too small to be observed or are nonexistent." If this is correct, it is likely that the maximum value assigned to nuclear fission in Fig. 2 would be substantially lower than the corresponding value for hydroelectricity, making it comparable to the renewables.

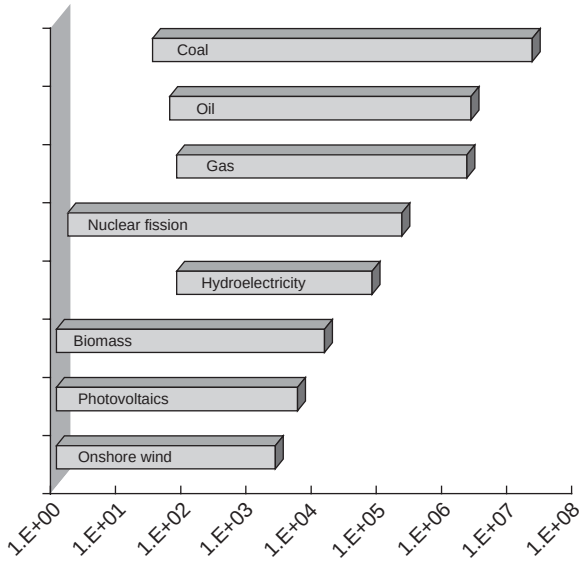


FIGURE 2 Total risk costs for selected energy systems in 1995 U.S. dollars per gigawatt-hour. Modified from Stirling (1999), illustrating ambiguities in the environmental ranking of contending options.

The problem with figures of this type is that they compare vastly different studies, with often strongly disparate assumptions. If similar studies are compared, the overlap may be substantially less. An analogy may be drawn about the heights of men and women. On average, men are taller than women. However, in a random group of perhaps 100 of each gender, the tallest woman will almost certainly be taller than the shortest man. The overall conclusion that men are taller than women is not disproved by this observation. Applying this to Fig. 2, the overlap may not be as great as it appears if like studies are compared.

15.2 Being Comprehensive

If comparisons are to be made between energy systems, great care must be taken to include all the sources of risk, from mining of the raw materials to disposition of waste, if any. This generalization can be difficult to implement in practice. Should the risks associated with building the machinery to do the mining be included? Should storage and backup, if necessary, be included as well? There has been considerable debate in the literature on these and other topics of comprehensiveness, without final resolution.

One way of solving at least some of these difficulties is the use economic input–output (I–O) tables, which show how different inputs to a system result in outputs. The risks associated with the inputs and outputs can, in principle, be calculated. Although the I–O system has been used by a number of investigators, it has not been generally adopted in the risk arena. Most investigators use a system approximately as shown in Fig. 3.

The top of Fig. 3 refers to occupational risks and the bottom to public risks. Clearly, there is considerable simplification in such a figure. For example, waste disposition, one of the greatest public concerns about nuclear power generation, is not shown. At least a rough estimate, given the fact that nuclear waste can pose some risks for thousands of years, is included in many studies of this technology.

Nonetheless, a figure of this type, with its defects, is useful. It shows decision makers, who rarely have time to sort through risk analyses of many volumes, the conceptual basis of comparative risk analysis.

15.3 The Hauptmanns Study

In 1991, Hauptmanns in Germany attempted to summarize more than a decade of major energy risk studies. He considered a wide variety of sources. Rather than develop a statistical distribution, he tabulated the maximum and minimum for 15 energy and electricity sources. Some of them, such as nuclear fusion, are mostly speculative and are not shown in Fig. 4.

In the interests of brevity, only occupational risks are discussed. Hauptmanns also estimated public risks for these technologies. Only fatalities are indicated in Fig. 4. As noted elsewhere in this article, an overall measure of occupational or public risk would also include injuries and diseases attributable to energy systems.

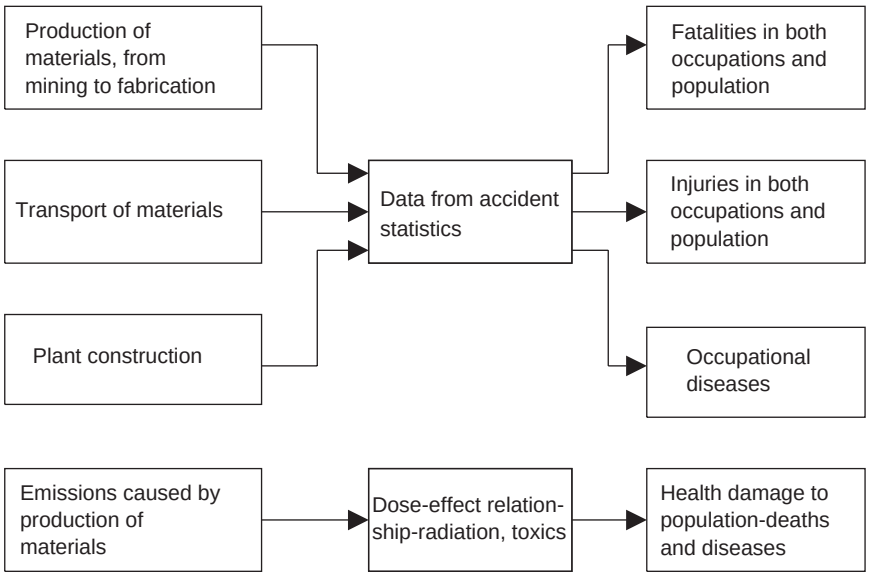


FIGURE 3 Sources of overall risk for energy systems. Modified from Hauptmanns and Werner (1991).

Although there is clearly overlap for many of the ranges shown, it is somewhat less than indicated by Stirling. The highest value seems to be for methanol, beyond the range of the 10 other systems. The fast breeder reactor apparently has the lowest value. However, the upper part of its range overlaps with that of a number of other systems, such as coal, oil, solar photovoltaic, and wind.

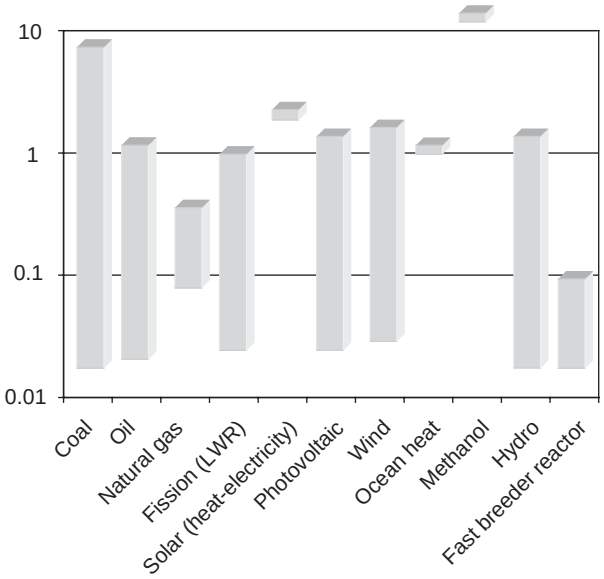


FIGURE 4 Employee (occupational) fatalities per 2.2×10^{16} J for 11 energy systems. A similar graph can be prepared for public risks. Modified from Hauptmanns and Werner (1991).

The next highest value is for coal, at least judging from its maximum. Its minimum, however, is among the lowest. This may be attributable to assuming so-called strip mining, for which occupational risks are substantially lower than those for underground mining. A fair estimate of coal occupational risks would allocate appropriate percentages to underground and surface mining.

15.4 The ExternE Project

After the Hauptmanns effort, the European Union, through the ExternE project, and the International Atomic Energy Agency, through its DECADES project, supported the largest schemes ever to estimate the relative risk of different energy and electricity systems. Taken as a whole, they are the most comprehensive set of energy risk studies to date.

Results are mostly applicable to Western Europe and, to some extent, the United States. Although evaluating the risks of energy in Third World countries would be useful as they industrialize,

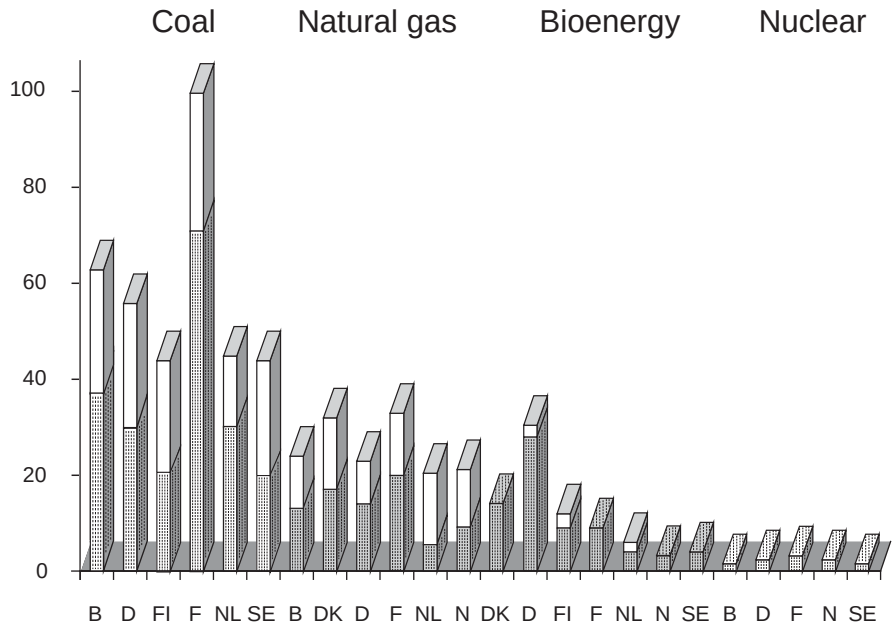


FIGURE 5 External costs of the entire fuel cycle, including global warming, in mECU per kilowatt-hour. B, Belgium; D, Germany; A, Austria; F, France; NL, The Netherlands; SE, Sweden; DK, Denmark; FI, Finland; N, Norway. The shaded portions of each column show the total environmental costs, including global warming, at a relatively low cost per ton of carbon dioxide emitted; the unshaded portions show the total costs at a higher value per ton of carbon dioxide emitted. Modified from Starfelt (1998).

there has been comparatively little written on the subject.

Starfelt attempted to draw general conclusions for the ExternE work. Results are shown in Fig. 5. He considered four systems: coal, natural gas, bioenergy, and nuclear. Different studies were performed for a variety of European Union countries. All costs, risks to human health and the environment, were put on a fiscal scale. The European Community Unit (ECU, commonly known as the euro) was at a level of approximately $1.20 in December 2003.

Note that this is similar to the calculations of Stirling noted previously. In effect, human risks per assigned a monetary value, requiring an estimate of how much a human life is worth. Although this estimate is required for the ExternE calculations, it is so buried in the voluminous calculations that no easy extraction is available. The reduction of all costs to money also implies estimates of the financial effects of climate change, which are not completely known. Thus, the ExternE results require more assumptions than the standard risk calculations for energy, which have already been criticized for having too many assumptions.

In Fig. 5, the shaded areas show the external costs, including global warming at an estimated cost of 19 ECU per ton of CO_2. Note that this is not a direct measure of potential economic costs of global warming but rather an estimate of the price that firms might pay to emit one extra ton of CO_2 in the future. The columns including the blank areas assume a cost of 46 ECU per ton of CO_2.

Despite substantial variations, coal had by far the greatest external costs, regardless of the estimated cost per ton of CO_2. French coal had the highest values, perhaps because coal from this nation is riskier to mine than elsewhere. The added cost of 46 ECU/ton made coal substantially higher than the other three systems, but it would have retained its most costly rank even if the lower value had been used.

Second was natural gas, with not as much variation as in the coal case. Values were approximately half those of coal. Using the higher value of cost per ton of CO_2 approximately doubled the values of natural gas.

Bioenergy was third, with average values approximately half those of natural gas. At approximately 33 mECU/kWh, German bioenergy was substantially higher than that of the other six nations in the survey. Lowest was that of The Netherlands, at approximately 5 mECU/kWh.

Nuclear was the lowest of the four. Most values were approximately 2 mECU/kWh. France seemed to have the highest value. The effect of valuing CO_2 at different costs played little role with regard to the relative external costs of bioenergy and nuclear, although all energy systems generate CO_2 either directly or indirectly.

16. CONCLUSION

Many conclusions can be drawn from the hundreds of studies on most aspects of energy risk. Perhaps the most remarkable is the divergence between public perception of energy risk, often centered on the hazards of nuclear power technology, and its often low relative rank in terms of comparative analysis.

Energy risk analysis is highly complex. The subjects discussed in this article represent only part of the field. Energy risk analysis is also somewhat different from other subjects in the field of energy in that much of it is highly contentious, with public debates over assumptions and results. These debates indicate the importance of the field to both policymakers and the public.

SEE ALSO THE FOLLOWING ARTICLES

Climate Change and Public Health: Emerging Infectious Diseases • Corporate Environmental Strategy • Ecological Risk Assessment Applied to Energy Development • Nuclear Power: Risk Analysis • Occupational Health Risks in Crude Oil and Natural Gas Extraction • Occupational Health Risks in Nuclear Power • Oil and Natural Gas Exploration • Radiation, Risks and Health Impacts of

Further Reading

American Chemical Society/Resources for the Future (1998). "Understanding Risk Analysis." American Chemical Society/Resources for the Future, Washington, DC.

Cohen, B. (1990). "The Nuclear Energy Option: An Alternative for the 90 s." Plenum, New York.

Douglas, M. (1985). "Risk Acceptability According to the Social Sciences." Russell Sage Foundation, New York.

Dyer, J. S., and Sarin, R. K. (1981). "Relative Risk Aversion." Univ. of Texas Press, Austin, TX.

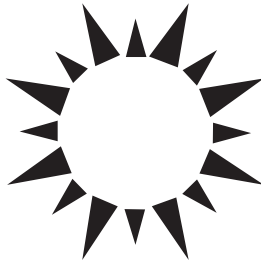
European Commission (1998). "Externalities of Fuel Cycles: Methodology," 2nd ed., DG XII, Science Research and Development. European Commission, Brussels.

European Science and Technology Observatory (1999). "On Science and Precaution in the Management of Technological Risk." European Science and Technology Observatory, Seville, Spain.

Franklin, J. (ed.). (1998). "The Politics of Risk Society." Polity, Cambridge, MA.
Hammond, B., and Coppock, R. (eds.). (1990). "Valuing Health Risks, Costs, and Benefits for Environmental Decision Making." National Academy Press, Washington, DC.
Hauptmanns, U., and Werner, W. (1991). "Technische Risiken ("Engineering Risks: Evaluation and Valuation")." Springer-Verlag, Berlin.
Heyman, B. (ed.). (1998). "Risk, Health and Health Care: A Qualitative Approach." Arnold, London.
Inhaber, H. (1982). "Energy Risk Assessment." Gordon & Breach, New York.
Leviton, L. C., Needleman, C. E., and Shapiro, M. A. (1998). "Confronting Public Health Risks: A Decision Maker's Guide." Sage, Thousand Oaks, CA.
Louvar, J. F., and Louvar, B. D. (1998). "Health and Environmental Risk Analysis: Fundamentals with Applications." Prentice Hall, Upper Saddle River, NJ.
Mossman, K L., *et al.* (1996). "Radiation Risk in Perspective." Health Physics Society, McLean, VA.
Paustenbach, D. J. (1995). The practice of health risk assessment in the united states (1975–1995): How the U.S. and other countries can benefit from that experience. *Hum. Ecol. Risk Assessment* **1,** 29–79.
Pochin, E. E. (1983). "Nuclear Radiation: Risks and Benefits." Clarendon, Oxford, UK.
Rabl, A., and Dreicer, M. (1999). An eye to the future. Long-term health and environmental impacts of energy systems. *Int. Atomic Energy Agency Bull. (Austria)* **41,** 31–34.
Rogers, J. T., and Bates, D. V. (eds.). (1982). "Symposium on the Assessment and Perception of Risk to Human Health in Canada." Ontario Science Centre, Toronto, Ontario, Canada.
Rück, B. (ed.). (1993). "Risiko ist ein Konstrukt ("Risk Is a Construct: Perceptions of Risk Perception")." Knesebeck, Münich.
Sagoff, M. (1985). "Risk–Benefit Analysis in Decisions Concerning Public Safety and Health." Kendall/Hunt, Dubuque, IA.
Schwing, R., and Albers, W. A., Jr. (eds.). (1980). "Societal Risk Assessment: How Safe Is Safe Enough?" Plenum, New York.
Sharma, H. D. (ed.). (1990). "Energy Alternatives: Benefits and Risks." Univ. of Waterloo Press, Waterloo, Ontario, Canada.
Starfelt, N. (1998). How ExternE results can be used for environmental impact comparisons: Sweden as an example. *Nuclear Europe Worldscan* (11–12), 57–58.
Starr, C. (1969). Social benefit versus technological risk. *Science* **165,** 1232–1238.
Stirling, A. (1999). "On Science and Precaution." Science Policy Research Unit, University of Sussex, UK.
Uranium Information Centre (2000). "Nuclear Electricity," 6th ed. Uranium Information Centre, Melbourne, Australia. [e-mail: uic@mpx.com.au]
U.S. Congress, House, Committee on Science, Space, and Technology (1994). "How Safe Is Safe Enough? Risk Assessment and the Regulatory Process, July 27, 1993." U.S. Government Printing Office, Washington, DC.
Warner, F. (1981). "Assessment and Perception of Risk: A Royal Society Discussion." Royal Society, London.
Whipple, C. (ed.). (1987). "De Minimis Risk." Plenum, New York.
Wildavsky, A. B. (1988). "Searching for Safety." Transaction Books, New Brunswick, NJ.
Wildavsky, A. (1995). "But Is It True? A Citizen's Guide to Environmental Health and Safety Issues." Harvard Univ. Press, Cambridge, MA.

Rocket Engines

WILLIAM ANDERSON
Purdue University
West Lafayette, Indiana, United States



1. **Introduction to Rocket Propulsion Systems**
2. **Fundamental Principles of Chemical Rocket Propulsion**
3. **Rocket Propellants**
4. **Rocket Engine Cycles**

Glossary

bipropellant Propellant system consisting of both an oxidizer and a fuel.

characteristic exhaust velocity A reference term relating achievable chamber pressure to the propellant type and nozzle throat diameter. It is a figure of merit for the propellant and efficiency of combustion.

impulse The total force imparted to a vehicle over the course of the burn time, with units of N-s. The impulse per unit mass is the specific impulse.

nozzle Thrust chamber component comprising converging and diverging sections. The minimal nozzle area is the throat, at which point the speed of the combustion gas is approximately equal to the speed of sound in the gas. The ratio of the nozzle exit area to the throat area is the expansion ratio.

propellant The working fluid in a rocket. The chemical energy contained in the propellant is converted first to thermal energy during the combustion process and then to kinetic energy during expansion in the nozzle.

thrust The reaction force imparted to a rocket primarily due to the change in momentum of propellants as they expand in the thrust chamber.

thrust coefficient The ratio between the thrust and the product of chamber pressure and throat area. It is a figure of merit for the nozzle design.

velocity increment The change in velocity of a vehicle due to its acceleration by thrust (also called ΔV).

Rockets are devices that contain all the elements necessary for propulsion within themselves. They are most useful for space travel and when high thrust rapid acceleration is required. Applications include boosting payloads to low-earth orbit, missiles, satellite station-keeping and orbit transfers, and interplanetary missions. Rockets are typified by the high velocity of the gas that is accelerated through a supersonic nozzle to generate thrust. They can be further classified according to the propellant state, the thrust level, and the type of engine cycle that is used.

1. INTRODUCTION TO ROCKET PROPULSION SYSTEMS

1.1 Rocket Propulsion

Propulsion is the means by which a force is imparted to change the velocity of a vehicle, affecting both its speed and its direction. In aerospace vehicles, this force is typically called thrust. The thrust applied to a rocket is due to a reaction from the momentum of a high-speed jet directed opposite the direction in which the vehicle is being accelerated. The thrust is proportional to the product of the mass flow rate, $\dot{m}$, and velocity, V, of the jet (i.e., $F = \dot{m}V$). These propulsive forces are used to overcome gravity and drag and to steer the vehicle.

Both rocket and jet engines use jet propulsion to develop thrust. Rocket propulsion differs from its air-breathing cousin mainly in that rockets have self-contained propellants. This feature allows operation at high altitudes and in space. Rockets are typically classified according to the energy source that is used to accelerate the working fluid. Energy sources include pressurized gas, chemical, electric, and nuclear. More exotic types, such as nuclear, solar sails, or beamed-energy approaches, may be feasible in the future, but their consideration is beyond the scope of this article.

Cold gas propulsion comprises a controlled, pressurized gas source and a nozzle and represents the simplest form of rocket engine. Cold gas has many applications in which simplicity is more important than high performance, for example, the

 © 2004 Elsevier Inc. All rights reserved.

Manned Maneuvering Unit used by astronauts. Nuclear and electric propulsion can be separated into two general categories: thermal and electrostatic/magnetic. Thermal systems work by a transfer of heat generated by the nuclear or electric energy source to a working fluid. Electrostatic/magnetic systems work by accelerating ions through an electrostatic or combined electromagnetic potential. Although some development has occurred, there are no operational nuclear rocket systems. There are operational systems that use electrothermal and electrostatic propulsion.

1.2 Chemical Rockets

The main focus of this article is chemical rockets, whereby the energy that is stored in the chemical bonds of the propellants is first converted to thermal energy through the process of combustion and then converted to kinetic energy by expanding the high-temperature combustion products through a supersonic nozzle. Since the attainable thrust is proportional to the exhaust velocity, and the exhaust velocity is proportional to the square root of the kinetic energy of the exhaust gas, it is readily seen that the chemical energy content of the propellants that can be converted to thermal energy provides a fundamental limit on the thrust that can be achieved.

Chemical rocket propulsion can be categorized according to the original state of the propellant. Rockets that use solid propellants are called solid rocket motors (SRMs), and those that use liquid propellants are called liquid rocket engines (LREs). A hybrid rocket uses liquid oxidizer and solid fuel. LREs and SRMs are the focus of this article.

A general schematic of a pump-driven LRE is shown in Fig. 1. A single gas generator (GG) is used to provide hot gas to turbines that drive the fuel and oxidizer pumps. The pressurized propellants are sent to the GG and the thrust chamber assembly (TCA), which comprises an injector, the main combustion chamber, and the supersonic nozzle. The effluent from the GG is shown to be expanded through separate nozzles; in some configurations, it is injected into the TCA nozzle at an axial location where the gas is the same pressure as that of the GG effluent. Figure 1 depicts a specific LRE cycle—the GG cycle. Other types of cycles are discussed later.

The SRM is one of the most common types of rocket propulsion systems and has a wide variety of applications. An SRM is conceptually a very simple device with very few moving parts. The entire propellant charge, the propellant grain, is contained within a pressure vessel generally called the motor case. The case is protected from hot combustion gases by internal insulation. The nozzle serves to convert the thermal energy of the combustion gases to thrust and may be moved (vectored) to provide a directional force for steering. The polar boss provides a means of connecting the nozzle assembly to the motor case, whereas the thrust skirt provides a mechanism to transmit thrust loads to an upper stage or payload.

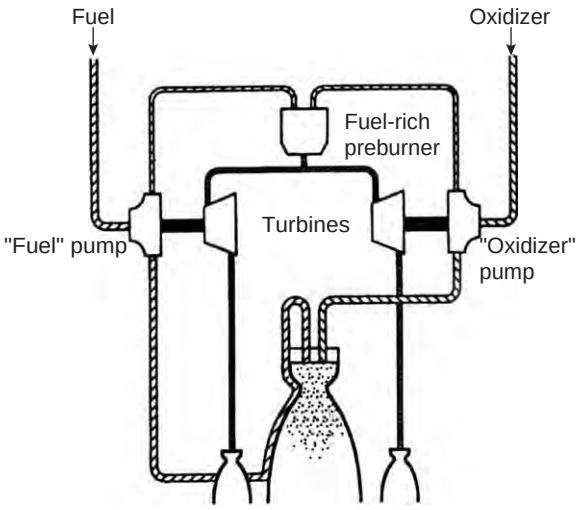


FIGURE 1 Liquid rocket engine schematic. The pump-fed engine shown uses a fuel-rich gas generator to produce turbine drive gases. The turbines provide power to pump the oxidizer and the fuel to high pressures. The high-pressure propellants enter the combustion chamber through the injector and react to form a high-temperature gas. The high-temperature, high-pressure gas is expanded through a supersonic nozzle, in which it reaches a very high velocity.

Other distinguishing features of chemical rocket propulsion include high operating pressures and high rates of propellant consumption. Operating pressures can be up to 50 MPa for LREs, and flow rates may be >4000 kg/s for high-thrust SRBs. A pump on the Space Shuttle main engine (SSME) could drain an average backyard swimming pool in 25 s, and the energy released by the three SSMEs is equivalent to the output of 23 Hoover Dams. The high energy density of the rocket provides a 100-fold increase in thrust over air-breathing engines of similar size.

1.3 Chemical Rocket Requirements

The energy requirements of a rocket are typically stated in terms of velocity increment or ΔV. The flight velocity required to escape from Earth's gravitational

field (the escape velocity, u_{esc}), neglecting the rotation of the earth, frictional drag, and the attraction of other celestial bodies, can be calculated as $u_{esc} = \sqrt{2g_e r_e} = 11.2$ km/s, where g_e is the acceleration due to gravity at the earth's surface ($g_e = 9.81$ m/s^2), and r_e is the radius of the earth ($r_e = 6.38 \times 10^3$ km). Launching an object into space at this initial velocity would not be practical because the object could not withstand the severe aerodynamic heating and dynamic pressures that it would experience in the earth's atmosphere.

For a rocket launched from the surface of the earth where the goal is to put an object into an orbit, the overall velocity increment is the difference between the velocity at the surface and the velocity required for a stable orbit, whereby the centrifugal force of the satellite balances the earth's gravitational field. A circular orbital velocity can be calculated by $u_c = \sqrt{\mu/r_c}$, where μ is the gravitational constant (3.99×10^5 km^3/s^2 for Earth), and r_c is the distance between the centers of the two bodies. The altitude above the earth surface corresponding to low Earth orbit (LEO), which serves as a stable orbit for many satellites as well as a staging point for acquisition to higher orbits, is on the order of 180 km or approximately 100 nautical miles. A typical ideal ΔV requirement for placing an object in LEO, neglecting the earth's rotation and also gravitational, steering, and drag losses, is approximately 8 km/s. Adding these real effects increases the ΔV requirement to approximately 9 km/s. Launching a rocket into an eastward trajectory near the Equator provides the maximum initial velocity and reduces the ΔV requirement. The overall flight velocity budget for a typical Space Shuttle LEO mission launching from Cape Kennedy is shown in Table I. Higher orbits require more velocity increment. Round-trips to the moon and Mars require a ΔV of approximately 18 and 27 km/s, respectively.

TABLE I

Typical Velocity Increment Budget for Space Shuttle

Circular orbit velocity at 110-km altitude	7790 m/s
ΔV gravity loss	1220 m/s
ΔV turning loss	360 m/s
ΔV aerodynamic drag loss	118 m/s
Orbit injection	145 m/s
Deorbit maneuver for reentry and aerobraking	60 m/s
Correction maneuvers	62 m/s
Initial velocity provided by Earth's rotation at 28.5° latitude	−408 m/s
Total ΔV required	9347 m/s

Other applications for chemical rockets include missiles and orbit transfer vehicles. Ballistic missiles spend most of their flight time in a ballistic trajectory in which only gravity forces are active. Ballistic flight is defined as a "zero-lift" trajectory in which the missile always flies at a zero angle of attack (i.e., the axis of the missile is aligned with its velocity vector). The ΔV for a ballistic missile is determined by its range requirement. Other missile applications include interceptors (e.g., air-to-air and surface-to-air missiles) that must close in on a moving target with high initial acceleration rates and therefore high thrust-to-weight requirements. Orbit transfer propulsion provides transfers from LEO to higher orbits [e.g., geosynchronous orbits (GEOs)] and inclination changes. The Hohmann transfer ellipse is a well-known maneuver between two circular orbits that provides a minimum velocity increment requirement. A LEO-to-GEO maneuver using the Hohmann transfer requires a ΔV of approximately 4 km/s.

2. FUNDAMENTAL PRINCIPLES OF CHEMICAL ROCKET PROPULSION

2.1 The Thrust Equation

The performance of a chemical rocket is governed by the thermal energy contained within the combustion products and the efficiency of the nozzle in converting this thermal energy into kinetic energy and thrust. Figure 2 shows a simple schematic of a chemical rocket. The thrust, F, is the reactive force applied to the rocket and is the sum of the momentum thrust and the pressure thrust:

$$F = \dot{m} V_e + A_e [P_e - P_\infty],$$

where A_e is nozzle exit area, P_e is the gas pressure at the nozzle exit, and P_∞ is the ambient pressure. The momentum thrust is much larger than the pressure thrust. When the pressure in the chamber is sufficiently higher (at least twice) than the ambient pressure, supersonic flow can be achieved in a properly designed nozzle. In a supersonic flow, the velocity increases and the pressure decreases with increasing nozzle cross-sectional area. This process of expansion is the means through which high V_e is obtained.

Although the thrust equation suggests that additional thrust can be obtained when the exit pressure

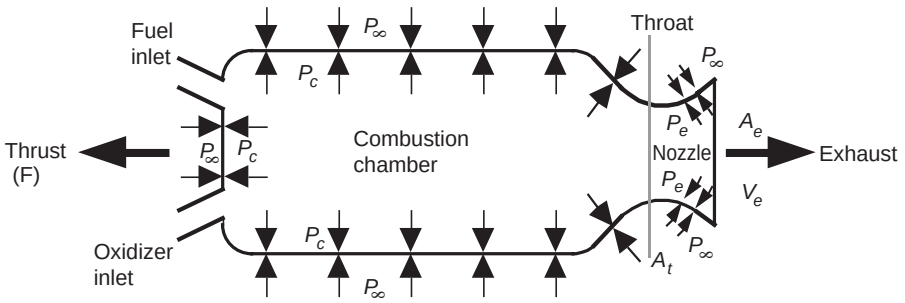


FIGURE 2 Simplified diagram of a chemical rocket. Reaction between the fuel and oxidizer in the combustion chamber produces hot gases, which are then expanded in a supersonic nozzle to high velocity.

TABLE II

Applications and Typical Performance

Propulsion technology	Launch vehicles	Orbit insertion		Orbit maintenance and maneuvering	Attitude control	Typical I_{sp} (s)
		Perigee	Apogee			
Cold gas				X	X	30–70
Solid	X	X	X			280–300
Liquid						
Monopropellant				X	X	220–240
Bipropellant		X	X	X	X	305–325
Dual mode		X	X	X	X	313–330
Hybrid		X		X		250–340
LO_2/kerosene	X	X				320–340
H_2O_2/kerosene	X	X				310–330
LO_2/H_2	X	X				405–460

is greater than the ambient pressure, the loss in the momentum thrust term is greater than the gain in the pressure thrust term. Rocket exhaust nozzles are often designed with an exit pressure equal to the ambient pressure; when P_e is equal to P_∞, the nozzle is "optimally expanded." The highest thrust for a given flow rate is achieved when the nozzle flow is expanded to pressures approaching zero in a vacuum environment. Some space engines have nozzle expansion ratios (ε, the ratio of nozzle exit area to throat area) exceeding 100. Practically, the expansion ratio is increased until the additional weight of the longer nozzle and the increase in internal nozzle drag losses result in an overall penalty. The physical envelope provided by the launch vehicle also sets limitations on the space engine nozzle expansion ratio.

If the ambient pressure (P_∞) is greater than the pressure at the exit of the nozzle, the pressure thrust term is negative. This loss in thrust is most severe at low altitudes for first-stage engines with exit nozzles that are configured to produce a maximum amount of thrust over the entire launch profile.

2.2 Rocket Performance Parameters

Mission requirements are often stated in terms of total impulse, I_t, which is a measure of the thrust imparted to the vehicle over the thrusting interval, $I_t = \int F dt$, and has the dimensional units of force-time. Performance is the figure of merit for a specific propellant combination and thrust chamber configuration and is usually related in terms of effective exhaust velocity or specific impulse (Table II). The effective exhaust velocity, c, is defined as $c = V_e + A_e[P_e - P_\infty]/\dot{m}$ and is equivalent to the actual exhaust velocity at optimally expanded conditions. The specific impulse, I_{sp}, with dimensional units of seconds, is the total impulse divided by the amount of propellant that produced that amount of impulse. It is also the ratio of the thrust to the propellant flow rate, $I_{sp} = F/\dot{m}$. The first definition is useful for

SRMs and the second definition is useful for LREs because for SRMs only the total amount of propellant that was burned is typically known, whereas for LREs precise measurements of propellant flow rate at any given time can be made.

Specific impulse is a measure of how efficiently the energy content of the propellants is converted into thrust and can also be defined as the product of the characteristic exhaust velocity, C^*, and the thrust coefficient, $C_f : I_{sp} = C^* C_f$. Definition in this way separates the physical effects occurring in the combustion chamber (C^*) and the exhaust nozzle (C_f). Characteristic exhaust velocity is a parameter primarily dependent on the propellant combination and how effectively the chemical energy contained in the propellant bonds is converted to thermal energy, whereas the thrust coefficient is a measure of how efficiently the thermal energy produced by the combustion reaction is converted to kinetic energy.

Characteristic velocity can be calculated as $C^* = f(\gamma)\sqrt{T/\mathrm{MW}}$, where T and MW are the temperature and molecular weight of the combustion products, respectively, and $f(\gamma)$ is a function dependent on the specific heat ratio, γ. The thrust coefficient is an expression that relates the magnification of the momentum thrust due to expansion in the supersonic nozzle, which can be written as $C_f = F/P_c A_t$ and which is dependent only on the specific heat ratio, the ratio of chamber pressure to nozzle exit pressure, P_c/P_e, and the expansion ratio, ε, of the nozzle exit area to nozzle throat area, A_e/A_t. These relationships show that high specific impulse is partially attainable by high-combustion gas temperatures and low-combustion gas molecular weights. Further examination of the fundamental thrust coefficient equation would show that higher values of pressure ratio coupled with higher values of ε, along with lower values of γ, lead to higher values of C_f.

2.3 The Rocket Equation

For most rocket propulsion applications, the primary requirement is the velocity increment, ΔV, that must be achieved, as discussed previously. The relationship between propulsion system performance parameters and mass fraction is the so-called rocket equation:

$$\Delta V = gI_{\mathrm{sp}} \ln\left(\frac{m_0}{m_0 - m_p}\right) = gI_{\mathrm{sp}} \ln(m_0/m_f) = gI_{sp} \ln(1/MR),$$

where $m_f = m_0 - m_p$ is the final vehicle mass, m_0 is the initial vehicle mass, m_p is the mass of the propellant consumed, and $\mathrm{MR} = m_f/m_0$ is the mass ratio. This equation allows one to calculate the mass of propellant required to reach a certain ΔV based on engine performance and the final mass of the rocket, which is very close to the dry weight of the rocket.

3. ROCKET PROPELLANTS

3.1 Thermodynamics

In a chemical rocket, the propellant serves the dual functions of energy source and the working fluid that provides the momentum thrust. The energy source is the exothermic reaction that occurs between fuel and oxidizer or during the decomposition of a monopropellant. In an exothermic reaction, the products have lower heats of formation than do the reactants, and the amount of energy that is released, the enthalpy of reaction, ΔH_{rxn}, is proportional to their difference,

$$\Delta H_{\mathrm{rxn}} = \sum_{\mathrm{products},i} n_i \Delta H_i^0 - \sum_{\mathrm{reactants},j} n_j \Delta H_j^0,$$

where ΔH_i^o is the molar heat of formation of the ith component, and n_i is the number of moles of the ith component. The heats of formation of elemental molecular species such as oxygen and hydrogen are zero by convention, the heats of formation of typical products such as steam and carbon dioxide are negative, and the heats of formation of unstable molecules such as hydrogen peroxide and hydrazine are positive.

The liberated chemical energy takes its sensible form by heating up the products of combustion. An energy balance between the reactants and products provides the adiabatic flame temperature, T_{ad}; this is the temperature that the combustion products would reach according to

$$\Delta H_{\mathrm{rxn}} = \sum_{\mathrm{products},i} \int_{T_{\mathrm{ref}}}^{T_{\mathrm{ad}}} n_i C_{p,i} dT,$$

where T_{ref} is a reference temperature (typically 298.2 K) and equal to the reference temperature for the heat of formation data. It can be seen that the desirable propellant has a high positive heat of formation and reacts with the opposite propellant to form low-molecular-weight and low-heat-capacity products of reaction. Table III summarizes thermodynamic properties of some common propellants and products of combustion.

The equation of conservation of energy may be written for an ideal flow as $h_o = h + V^2/2$, where h is

TABLE III

Chemical Thermodynamic Properties of Common Propellants and Products of Combustion

Substance	Phase	MW (g/mol)	$\Delta H_{form}{}^{o}$ (kJ/mol)	C_p (J/mol K)
Al_2O_3	L	102.0	−1620.6	79.0
CO_2	G	44.0	−394.4	37.1
H_2O	G	18.0	−228.6	33.6
N_2H_4	L	32.0	50.6	98.8
N_2O_4	L	92.0	−19.6	142.5
NH_4ClO_4	S	117.5	−295.8	128.1

the specific enthalpy, and h_o is the stagnation enthalpy, which is equal to the sum of the heat of reaction and the sensible enthalpy carried in with the propellants. As the flow is accelerated through the supersonic nozzle, the static temperature decreases. As the temperature approaches zero, h approaches zero. Thus, the stagnation enthalpy sets the limit on the velocity that can be achieved by an infinite expansion through a supersonic nozzle.

3.2 Liquid Propellants

Liquid propellants can be categorized as oxidizers and fuels and as monopropellants and bipropellants. An oxidizer plus fuel is a bipropellant system. Some oxidizers and fuels can act either as part of a bipropellant system or as a monopropellant. Monopropellants generate propulsive energy through the exothermic decomposition of a single molecule or formulated mixture of fuel and oxidizer. Monopropellants allow simple systems but offer low performance. Only three monopropellants have been used in actual flight vehicles: hydrazine, hydrogen peroxide, and propyl nitrate. Monopropellant thrusters are used primarily for attitude control devices and GGs for turbine drive.

The most commonly used monopropellant is hydrazine (N_2H_4). Hydrazine is also a bipropellant fuel hypergolic with nitric acid, nitrogen tetroxide, and hydrogen peroxide. Hydrazine derivatives include unsymmetrical dimethylhydrazine (UDMH) and monomethylhydrazine (MMH), which are more stable than hydrazine with a wider liquid range but slightly lower specific impulse values. MMH is extensively used as a fuel in small attitude control engines. Hydrogen peroxide can also be used as a bipropellant. As a monopropellant, it can be decomposed to form hot water and oxygen.

A bipropellant engine introduces the complexity of a propellant feed system but allows higher performance. Oxygen and fluorine are the strongest oxidizers but must be stored in a cryogenic state. Oxygen is the most commonly used oxidizer in liquid bipropellant rocket propulsion systems. It has a normal boiling point of 90 K with a density of 1.32 g/cm^3. Its critical temperature is 155 K and critical pressure is 5.0 MPa. Nitrogen tetroxide is a common storable oxidizer that is hypergolic with many fuels. The Occupational Safety and Health Administration lists a ceiling exposure limit of 5 parts per million with possible health hazards primarily to the lungs and irritation to the eyes, nose, and throat.

The most commonly used storable liquid fuels are RP-1 and hydrazine and its derivatives. RP-1 is a petroleum derivative with a H/C ratio of 1.953, a specific gravity of 0.8, and a MW that ranges from 165 to 195 g/mol. It is typically used with liquid oxygen (LOX) in launch stages because of its combination of high density, low toxicity, and reasonable performance.

Hydrogen is the most common cryogenic fuel. It provides high performance and is an excellent regenerative coolant. However, at its normal boiling point of 20 K, its specific gravity is approximately 0.07; thus, it requires heavy insulation and large tanks. It is especially favored for high-performance, expander cycle upper-stage engines.

In addition to performance characteristics (Table IV), a number of other factors enter into the selection of a propellant for a specific propulsion application. Most of these considerations are highly mission dependent and include propellant density, handling, stability, and transport properties.

3.3 Solid Propellants

The propellant in an SRM must burn at the proper rate to maintain the desired thrust profile and must

TABLE IV

Characteristics of Liquid Propellant Combinations[a]

Propellant	O/F	Bulk density (kg/m^3)	C^* (m/s)	T (K)	I_{sp} (s)
LO_2/LH_2	4.83	320	2385	3250	455.3
LO_2/RP-1	2.77	1030	1780	3700	358.2
N_2O_4/MMH	2.37	1200	1720	4100	341.5
N_2O_4/N_2H_4	1.42	1220	1770	3260	343.8

[a] $P_c = 6.9$ MPa; $\varepsilon = 40$; $P_\infty = 0$.

have structural continuity to withstand the pressure and acceleration loads introduced during flight. Most physical characteristics cannot be predicted analytically, so heavy use of empirical data is required. Solid propellants are classified into two general categories. In double-base propellants, fuel and oxidizer are mixed on a molecular level. These propellants generally contain a nitrocellulose type of gun powder dissolved in nitroglycerine with additional amounts of minor additives. Both these primary ingredients are explosives containing fuel and oxidizer within their molecular structure. Double-base propellants have been used most frequently in military applications (primarily in ballistic missiles), but their use has declined due to an increased emphasis on safety.

The other general type of solid propellant is called composite propellant to highlight the fact that it is a heterogeneous mixture of fuel, oxidizer, and binder. Fuels for composite propellants are generally metallic powders, with aluminum being the most common. Oxidizers are usually crystalline materials; ammonium perchlorate (NH_4ClO_4), also known as AP, is the most popular. The binder is a rubber-based material that serves to hold the powder/crystal mixture together in a cohesive grain. Current binders include hydroxyl-terminated polybutadiene (HTPB) and polybutadiene nitrile (PBAN). Occasionally, small amounts of very energetic ingredients such as "Her Majesty's Explosive" (HMX) are added to improve performance.

Solid propellant burning rates are a function of chamber pressure. Most designers assume a burning rate dependence according to St. Robert's law, $r_b = ap_c^n$, where r_b is the burn rate (cm/s), a is the burn rate coefficient, and n is the burn rate exponent. The constants a and n are determined empirically. Table V summarizes characteristics of some propellant formulations.

TABLE V

Characteristics of Composite Propellants Used in Solid Rocket Motors

Combination[a]	Density (kg/m^3)	C^* (m/s)	T (K)	P_c range (MPa)	a	n
A	1840	1541	3636	2.75–7.0	0.415	0.31
B	1800	1527	3396	2.75–5.0	0.399	0.30
C	1760	1568	3392	2.75–7.0	0.561	0.35

[a] A: TP-H-1202, 21% Al, 57% AP, 12% HMX, 10% HTPB; Star 63D; B: TP-H-1202, 18% Al, 71% AP, 11% HTPB—Star 48, Star 37; C: TP-H-1202, 16% Al, 70% AP, 14% PBAN—RSRM, Titan IV.

4. ROCKET ENGINE CYCLES

The liquid propellant must be transported to the LRE TCA by some means of force. The more simple method relies on gas pressure to force the rocket propellant through one or more main propellant tank feed valves or directly into the TCA (as in the case of solid rocket motors). The other method is to use engine-mounted hot gas-powered turbopumps to feed the propellant from its tank into the TCA. All large liquid booster engines for all launch vehicles, worldwide, that have ever flown or are flying today are turbopump driven, whereas pressure-fed systems are typically used for smaller propulsion systems, such as spacecraft attitude and velocity control propulsion systems and smaller upper stages.

Pressure-fed systems can be either blowdown or regulated. In the simple blowdown configuration, the propellant tanks are initially pressurized with a finite amount of gas, and the tank pressure decreases as the propellant is consumed. In regulated systems, a constant tank pressure is maintained by controlling the pressurant flow from separate high-pressure gas tanks.

Pressure-fed systems are inherently closed cycles but are limited in performance because of the achievable chamber pressures and hence the velocity to which nozzle flow can be accelerated. The pressurant supply typically ranges from approximately 1 to 4 MPa. The low-pressure range corresponds roughly to the low-pressure limitation of minimum-gauge wall thickness tanks. The high-pressure limit corresponds roughly to the point at which the weight penalty of high-pressure gas supply tanks overcomes the added complexity and weight of a pump-driven system.

Engine power cycles for pump-fed systems include the GG cycle, the expander cycle, and the staged combustion cycle. Generally, the pump-fed cycles require that the work required by the pump be provided by the turbine. The required pump work is related to the flow rate, $\dot{m}_p$, the liquid propellant density, ρ, the required pressure rise, Δp, and the pump efficiency, η_p:

$$W_p = \frac{\dot{m}_p \Delta p}{\rho \eta_p}.$$

The delivered turbine work is related to the turbine flow rate, $\dot{m}_t$, the turbine pressure ratio, TPR, the turbine inlet temperature, T_{ti}, turbine drive gas properties R and γ, and the turbine

efficiency, η_t:

$$W_t = \dot{m}_t \left[\frac{\gamma}{\gamma - 1}\right] RT \left[1 - \left(\frac{1}{TPR}\right)^{\frac{\gamma-1}{\gamma}}\right] \eta_t.$$

Balancing the work flows with the propellant flows can be quite challenging, particularly during engine transients, and requires a significant amount of preliminary analysis and engine system testing.

The GG cycle (Fig. 1) is the most common type of rocket engine used for main propulsion. It is classified as an open-cycle engine and in its simplest form it is composed of a set of turbopumps for each propellant, a GG to provide power to drive the turbopumps, a set of control valves, and the main chamber and nozzle. The gas from the GG passes through the turbines that drive the turbopumps and is then expelled separate from the main chamber. Typically, 4 or 5% of the total propellant flow goes through the GG, which is usually operated at a fuel-rich mixture according to the limitations on turbine inlet temperature. This overboard turbine flow is a loss to the potential performance of the engine. Sometimes the GG flow is used to provide control capability to the vehicle, and sometimes it can be used as film cooling (also known as dump cooling) on the aft end of the nozzle.

The expander cycle is the simplest of the pump-fed engines and can be configured as either an open-cycle or a closed-cycle engine. In either case, the energy to pump the propellants is obtained not from burning propellants in a gas generator but rather from a transfer of heat from the main chamber. Fuel, typically hydrogen, is used as a regenerative coolant before it is used as the turbine drive gas. The thrust chamber and nozzle heat transfer limits the energy available for pumping the propellants and thus the achievable chamber pressure. Furthermore, the fuel must have a high heat capacity and good heat transfer properties and must vaporize easily without passing through a two-phase condition.

The expander cycle is inherently easy to start and control since the dynamics and stability concerns of preburner or GG combustion are avoided. An expander engine starts by using the latent heat capacity of the chamber and nozzle to initially accelerate the turbopumps, prior to main chamber ignition. The expander cycle is ideally suited for upper stages and space transfer applications, in which ease of starting and multiple restarts are required. Operational expander cycle engines used for upper stages include the RL-10-A5 and the LE-5. Thrust levels range from approximately 20 to 40 kN.

The staged combustion (SC) cycle engine (Fig. 3) is a closed-cycle engine that presents the highest performing and most complicated cycle in use. In SC cycles, part or all of the propellants go from the turbopump discharge to a high-pressure preburner, which produces turbine drive gas. The turbine effluent is mixed with the remaining propellants and combusted in the main combustion chamber. All the combustion gases are expanded from a high pressure, so the I_{sp} is high. However, the cycle demands high preburner pressures, which stress the turbomachinery.

The earliest identified SC cycle engine, the Topping Turbine Turborocket, was built by Reaction Motors and produced 5000 pounds of thrust. The turbopump was powered by a turbine located inside the combustion chamber. The drive shaft powering the pumps exited the chamber through the center of the injector. The first SC cycle engine to enter into production was the Soviet NTO/UDMH RD-253 engine, which was used on the first stage of the Proton booster, first launched in 1965. The Soviet Union also utilized the SC cycle on all of the LOX/kerosene NK series engines used for its N1 moon rocket program in the mid-1960s. A well-known SC cycle engine is the SSME, which is probably the most advanced engine in operation today.

The SC cycle can be categorized according to whether the turbine drive gas is either fuel rich or

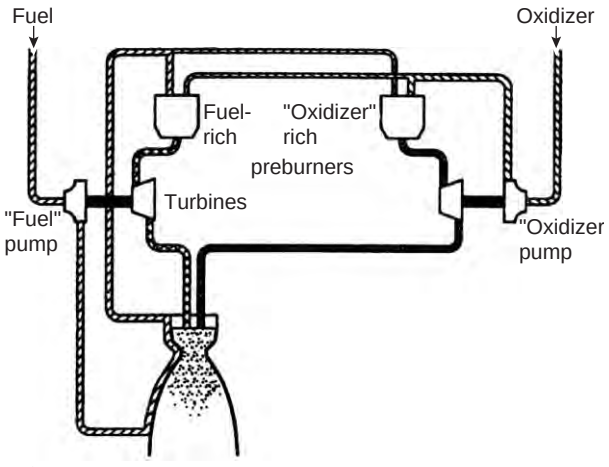


FIGURE 3 Full-flow staged combustion cycle rocket engine. Fuel and oxidizer are pumped to very high pressures before entering the fuel-rich and oxidizer-rich preburners. Prior to entering the preburner, the fuel is used to regeneratively cool the main combustion chamber and nozzle. Warm turbine exhaust gas enters the main combustion chamber, where it reacts further to produce a high-temperature and high-pressure gas that is expanded to very high velocities in a supersonic nozzle.

oxidizer rich. The fuel-rich SC (FRSC) cycle is typically applied to LOX/LH_2 engines (i.e., SSME, RD-0120, and LE-7). The oxidizer-rich SC (ORSC) cycle is used with hydrocarbon fuel-based engines and has been demonstrated in several Russian engines (i.e., RD-170, NK-33, and RD-253) using LOX/kerosene or NTO/UDMH. Another variant of the SC cycle is the full-flow SC (FFSC) cycle, which merges some of the virtues of the FRSC and ORSC cycles but has not been demonstrated in any engine system. The FFSC cycle engine requires two preburners, one producing fuel-rich hot gas to power the fuel turbopump and the other producing oxidizer-rich hot gas to power the oxidizer turbopump.

Although rocket propulsion is a fairly mature field, there is substantial room for improvement. Advancements in the following areas are required: system reliability and long-lived components; engine health diagnostics; high-energy density and storable propellants; lighter, higher temperature materials; nozzles that automatically adjust to ambient pressure conditions; and advanced thermodynamic cycles.

SEE ALSO THE FOLLOWING ARTICLES

Aircraft and Energy Use • Combustion and Thermochemistry • Internal Combustion (Gasoline and Diesel) Engines

Further Reading

Hill, P., and Peterson, C. (1992). "Mechanics and Thermodynamics of Propulsion." 2nd ed. Addison-Wesley, Reading, MA.

Huzel, D. K., and Huang, D. H. (1992). "Design of Liquid Propellant Rocket Engines." American Institute of Aeronautics and Astronautics, Washington, DC.

Manski, D., Goertz, C., Sassnick, H. D., Hulka, J. R., Goracke, B. D., and Levack, D. J. H. (1998). Cycles for Earth-to-orbit propulsion. *J. Propulsion Power* **14,** 588–604.

Sutton, G. P., and Biblarz, O. (2001). "Rocket Propulsion Elements." 7th ed. Wiley, New York.

Rural Energy in China

ZHENG LUO
United Nations Development Programme
Pathumthani, Thailand

1. Overview of Rural Energy Development
2. Energy and Rural Environment
3. Energy and Rural Economy
4. Energy and Rural Society
5. Looking Forward

Glossary

chronic obstructive pulmonary disease (**COPD**) A disease characterized by a progressive airflow limitations caused by an abnormal inflammatory reaction to the chronic inhalation of particles.

global warming potential (**GWP**) The GWP of one molecule of CO_2 is set = 1, to which the GWP of other gases are placed in relation. For instance, in a period of 20 years, one molecule of CH_4 causes a direct warming effect, which corresponds to that of 35 molecules of CO_2.

ren min bi (**RMB**) The official currency of the People's Republic of China. One U.S. dollar is about 8.3 RMB.

small hydro In China, capacity less than 25,000 KW is categorized as small hydro.

Substantial changes have taken place in China. Energy development, as both a driving force and a consequence of such tremendous changes, has had profound impact on economic, social, and environmental development. Rural energy has always been a critical issue due to years of energy shortage for both households and industries. Biomass, for long time, has been the only available fuel in many rural areas. A series of policies and strategies have gradually been put in place to encourage alternative fuel exploration and utilization as well as area-based integrated energy systems. As a result, unelectrified rural population dropped from 60% in the 1950s to 5%, while the share of commercial fuel use increased to 43%. The importance of renewable energy has been recognized, not only because of its potential in rural and remote areas, but also as means of improving the environment and generating economic opportunities.

1. OVERVIEW OF RURAL ENERGY DEVELOPMENT

This article does not pretend to provide a complete in-depth assessment of rural energy in China. However, it is hoped that through a historic summary of the development in past 50 years and analysis of some relevant aspects, this discussion will contribute to a better understanding of the current situation and future challenges.

1.1 Rural Energy Path

Generally, the development of rural energy and rural energy policy in China can be divided into three stages.

1.1.1 Energy Shortage and the Development of Biogas, Small Hydro, and Local Coal Mines (1950–1979)

Despite the high level acknowledgement of the situation, rural energy shortage was severe throughout these three decades. By the end of 1979, the energy consumption per capita in rural areas was 19% of that in urban cities. Seventy percent of rural households lacked cooking fuel, among which 47% had no fuel for more than 3 months per year. Promotion of biogas for cooking and lighting and utilizing local coal resources were strategic decisions from the top since the 1950s; however, not until the late 1970s were specific institutional arrangements and policy measures clarified for a biogas program, with allocation of financial and human resources. Although large-scale biogas projects have been carried out, most digesters failed to perform soon after installation (1 to 3 years of average life time)

 © 2004 Elsevier Inc. All rights reserved.

due to insufficient technical, managerial, and financial support. Because of this long period of lacking of alternative cooking fuels, regional deforestation, and excessive utilization of agricultural residues, among other factors, have contributed to serious degradation of rural ecological environment, which, in turn, has hampered the rural economic development. The impacts were felt well into the next stage.

1.1.2 Satisfaction of Energy Needs and Development of Policies and Regulations (1980–1990)

After the first national seminar on rural energy in 1980, series of energy policies and regulations have been announced focusing on reforestation, improvement of energy efficiency in both rural industries and households, and the development of renewable energy. The principle of integrated rural energy strategy was defined as developing all possible local alternative resources of energy while rationally utilizing the current sources. Nation wide programs have been carried out in this context, and most of them continuously constitute the current rural energy activities. Thus, it is worth while to have a detailed discussion on the following main results from this period:

Fuelwood Forest. As one of the five forest categories, the main products of the fuelwood forest are wood fuels. Reforestation plans were conceived according to local surveys on fuelwood supply and consumption. From 1981 to 1995, approximately 5 million hectares of forest have been established as fuelwood forest, which, in total, reached 6 million hectares and generated 20 to 25 million tons of wood as fuel per year. However, due to the decline of fuelwood demand, it is expected that the financial viability of these reforestation projects will be highly challenged. Technology optimization and standardization ensuring a diversified and reliable range of products are essential to improve the economic performance. Further more, flexible management and involvement of the private sector are crucial to gradually moving away from subsidies to a market-based operation.

Improved Cook Stoves and Energy Efficiencies in Rural Industries. From 1983, large-scale promotion of improved cooking and heating stoves has been carried out all over the country. By the end of 1999, it was reported that approximately 189 million rural households had been equipped with improved stoves. Generally the efficiency of these stoves, using coal and solid biomass as fuel, has improved by 20 to 40%, which enables energy saving of more than 30 million TCE (ton of coal equivalent) per year. However, to what extend these stoves have been used is not known. Rural industries have experienced a similar process. Improved boilers and kilns, mainly used in tea processing, tobacco drying, and brick and cement producing industries, could have brought about 2 million TCE per year energy saving, as well as significant quality improvement and control in production.

Rural Coal Mining. By the end of the 1970s, coal production from the mines developed at town and county level had already taken a significant role in the energy supply in China. The production of coal from these mines shared 18% of the national total in 1980. The continuous encouragement in collective and private coal mining has tripled their production to nearly 390 million tons and shared 36% of the coal production. Such increase climaxed in the year 1995–1996 when the production shared 45% of national total coal output. Despite low-efficient equipment and technologies and a poor safety record, the rural coal mines have contributed to the rural economy in three ways: easing the energy shortage, ensuring rural productivity, and improving rural employment. While majority of the coal produced from these mines has served market demand locally and from neighboring provinces, a small portion of the product, usually of better quality, has been transferred to the eastern coastal region. To a certain extent, replacing fuelwood with coal has also contributed to forest conservation. The low production and combustion efficiency of coal, however, has had adverse impact on rural environment and health. These effects have also been aggravated by the poor quality of coal in China and the fact that only a small amount has been processed. A case study on Xuan Wei County, Yunnan Province, showed that indoor air pollutants caused by coal combustion was the main health risk factor responsible for the high rate of lung cancer cases in the county.

Rural Electrification. Development of small hydropower has been the dominant strategy for rural electrification in China. In the late 1970s, several policy measures, including subsidy schemes, were adopted to promote small hydro, which resulted in a large increase in generation capacity from an average of 58,000 KW per year in the 1960s to more than 1 million KW in 1979. In the 1980s, local grids and local electricity companies have emerged, which has helped to overcome the disadvantages of standalone systems and also enhanced the rational utilization and distribution of electricity at county level. The

improvement of grid connection continues in the 1990s while the expansion moves toward cross-county connection. Establishment of hydroelectricity enterprises are further encouraged by introducing, among others, a private stockholder scheme. Per capita rural electricity consumption increased 7 times in 40 years and by the end of the year 2000, 28 million people (around 5 million households) in rural areas had no access to electricity, comparing to 450 million in 1978 (nearly half of the total population and 57% of rural population at that time).

1.1.3 Sustainable Energy Development and Market-Oriented Policy and Strategies (1990–2000)

Since the beginning of the 1990s, environmental policy has become one of the principle national development policies in China. As in many other developing and developed countries, renewable energy and energy efficiency have become more prominent in the energy agenda. Several rural energy policies have been published to address renewable energy and sustainable development, energy and environment, and rural energy technology in poverty alleviation. With the deepening rural economic reform since the mid-1980s, energy has been advocated as an integrated component of the rural development and rural market economy. The need for commercialization of conservation technologies, utilization of new and renewable energy resources, and the provision of energy services have been given great emphasis. Progresses in some of these aspects mentioned earlier are detailed in the following discussion:

Renewable Energy (RE). Direct combustion of biomass has gradually decreased both in share and absolute terms, while the utilization of solar, biogas, small and pico hydro, and biomass gasification have increased. Energy policies in addressing environmental problems in the country have set favorable conditions to RE. In particular, several financial measures, such as subsidies, low- or zero-interest loans, taxes, and so on have been essential in renewable energy development. Since 1987, the state council has financially supported more than 500 RE projects in the form of low-or zero-interest loans (more than 400 million RMB, e.g., approximately US$50 million). From 1990 to 1996, more than 1 million US$ subsidy was given to training and demonstration of RE technologies, while between 1996 and 2000, subsidies of about 7 million U.S. dollars were given to research and development of RE technologies and implementation. As a result, in 1998, renewable energy utilization, excluding traditional biomass, was 35 million TCE, accounting to more than 2% of the China's total energy consumption (including commercial and noncommercial sources).

Rural Energy Enterprises. Rural energy enterprise emerged with the promotion of biogas in the late 1970s to provide technical services and after build services. Under the strong support from the government on efficiency technologies, various types of devices and equipment have become commercialized. Into the 1990s, different renewable energy technologies have been demonstrated and promoted and, along with it, a more diverse rural energy industry has emerged. By the end of 1998, there were about 4300 rural energy enterprises, 1660 of which were in the industrial sector, manufacturing energy devices for rural users, such as solar water heater, PV, biomass gasifiers, and so on, and more than 2630 were in the service sector, providing technical, marketing, sales, and other services. Compared to 1997, the number of both types of enterprises increased by 8% and 18%, respectively, although the number of total employment has stayed roughly the same (approximately 42,000 to 43,000). Until 2000, such yearly increases have been consistent. However, data show that before 1996, enterprises and employment had been declining in the rural energy sector, and a relatively sharp decrease occurred to the collectively and state-owned enterprises. This short period of decline reflects the further market opening up in rural China, making many inefficient and less flexible state enterprises go bankrupt. Nonetheless, the rural energy industry has yet to mature and favorable policy environment is still essential, since its benefits to overall rural sustainable development have not been integrated and valued fairly (see also Section 3).

The development of rural energy, particularly in the past 20 years, has brought significant changes in the sector in terms of policy, structure, management, technology, and economy. Rural energy consumption increases at a steady rate of 4% per year from 330 million TCE in 1980 to 670 million TCE in 1998. Several comparisons of changes are listed in Table I, while the changes of rural energy structure are presented in Fig. 1.

The fall of rural energy consumption in 1996 is mainly due to the large decrease in the use of noncommercial energy sources (e.g., fuelwood and straw), while the use of coal and oil increased with the latter nearly doubled comparing to 1995.

TABLE I

1980–1998 Rural Energy Development in China

	1980	1998
Rural population (million)	796	869
Average rural income per person (RMB)	191	2162
Total energy consumption (million TCE)	330	670
Average energy consumption per capita (Kg CE)	420	770
Total electricity, consumption, including household and rural industry (billion KWh)	13	129
Average household electricity consumption per capita (KWh)	8	68
Population without electricity (million)	450	60
Clean fuel for cooking (electricity, LPG, coal gas, biogas, producer gas, solar, etc.) (million households)	3.7	56.3

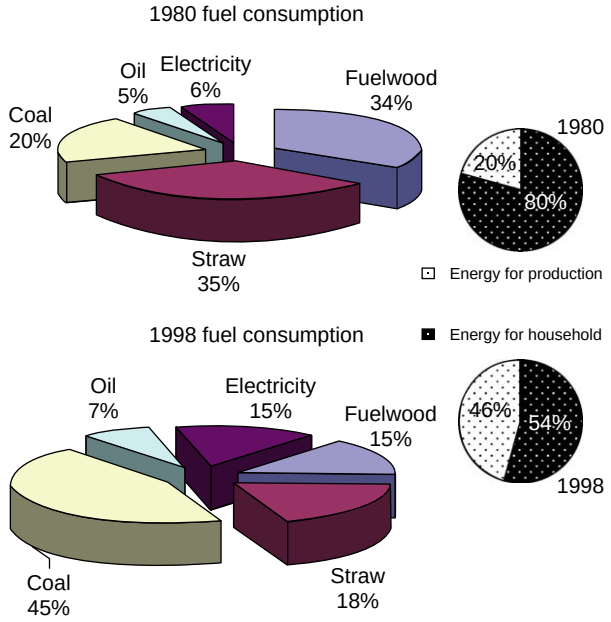


FIGURE 1 Fuel consumption structure changes in rural China 1980–1998.

Decrease in residential electricity consumed in 1996 was uncertain. Official figures from the Ministry of Agriculture show that the consumption was 29 million TCE, which was two-thirds of that in 1995 and, although increasing in the following years, it has not regained the level of 1995. However, according to the statistics from Department of Rural Electricity of the National Electricity Power Bureau, no such decrease has been observed. Nevertheless, energy

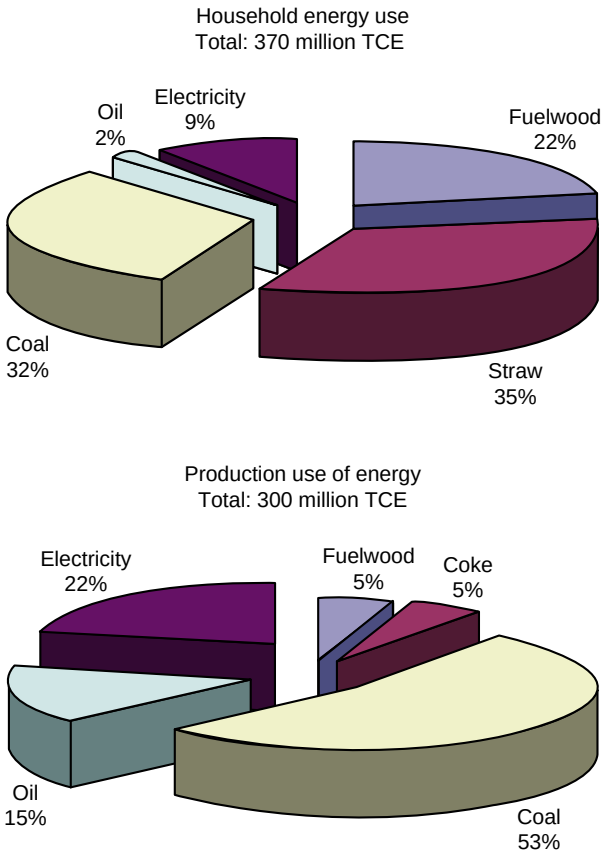


FIGURE 2 Rural energy consumption, 2000.

consumption in rural households has been grown in slow pace since 1995, if not decreased, mainly due to the closure of small coal mines and power producers. In the same period, rural productivity use of energy grew about 16% despite the successful efficiency program in rural industries nationwide. This has significantly contributed to higher income and improved livelihood in many rural areas in China.

1.2 Rural Energy Today

Since 2000, rural energy has accounted to approximately 45% of the total energy consumption and 35% of total commercial energy consumed (e.g., excluding fuelwood and straw). The total rural energy consumption was 670 million TCE in 2000 (see Fig. 2). The majority of rural households still use solid fuel for cooking and heating, while 84% of urban households are using gaseous fuels (LPG, natural gas, and coal gas). China has undergone series of reforms in the energy sector in the past 10 years, and the reform is being extended into rural power grid. The long-term goal of the power sector is

the establishment of a unified national grid. It would require a modern power market in which small and large plants are able to sell electricity to the grid at market-determined price. It would also require a well-defined regulatory role of the government. In the short to medium term, many rural areas will stay with local grid or isolated (village) grid, where renewable energy sources are most promising.

1.2.1 Energy Efficiency

As already mentioned, energy efficiency has been improved tremendously in the past years. However, comparing to national and international averages, rural industries still consume more energy per product unit as shown in Table II, and a significant difference exists between the western and the eastern regions, where efficiency is in general much higher. Because the out-of-date technologies and machinery still prevail in agricultural activities, energy consumption for mechanical power is very high, accounting for nearly half of the diesel used in the country. The current energy efficiency program, focusing on the modification of old equipment and technologies, results in diesel saving of 300,000 tons per year. However, the demand and consumption of diesel in agricultural sector will continuously increase since mechanical power is only responsible for approximately 50% of the productivity.

TABLE II

Energy Efficiency of Selected Rural Industries

Type of product (energy consumption/ unit)	Average of rural industries	National average	International average
Brick (kgce/10,000 pieces)	1093	780	780
Cement (kgce/t)	161	133	94
Coke (tce/t)	1.56	1.1	1.1
Pulp and paper (tce/t)	1.55 ~ 1.7	1.25	0.51
Electricity (from coal, gce/kWh)	450 ~ 550 (< 50 MW)	320 ~ 350 (300 MW)	329 (Japan)

1.2.2 Clean Energy

With the stabilized balance between demand and supply of energy, China has been moving forward to clean and modern energy and technologies. The demand for clean fuel (electricity, LPG, coal gas, and modern renewable fuels) from rural households has five-fold in the past 10 years, and this trend will continue. Currently, renewable energy and modern renewable energy (including small hydroelectricity) account for 19.8% and 2.3%, respectively, in the total national consumption, as shown in Fig. 3.

A detailed breakdown of the use of renewable energy resources in 1998 is presented in Table III, which shows that all new renewable energy (excluding traditional biomass and all hydroelectricity) constitutes 0.3% of total national energy consumption (i.e., 5.4 Mtce). According to national planning on renewable energy development, this figure is expected to reach 2% by the year 2015, which indicates that total consumption of new renewable energy would have to reach approximately 60 Mtce in 15 years, which may require hundreds of millions of U.S. dollars, as subsidies from central government. Thus, the commercialization of renewable energy and technologies, with a well-performing rural energy service sector, would be crucial. The existing rural energy industry covers almost all renewable energy technologies; however, their stage of development and commercialization varies greatly and are presented in detail in Section 3.

Although domestic coal meeting the demand will still be the strategy in rural energy development in the foreseeable future, coal combustion, particularly at the small scale, has been and will be unavoidably having a negative impact on rural environment and health. Modern renewable fuels, such as solar water

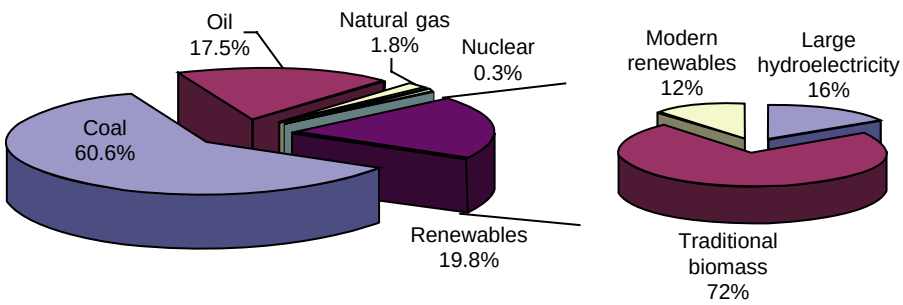


FIGURE 3 Renewable energy consumption in China.

TABLE III

Utilization of Renewable Energy Resources in China (1998)

Resources	Utilization	Mtce
Hydro electricity — small hydro[a]	208 billion kWh, 80 billion kWh	77.6, 29.84
Biomass	221.98 Mtce	221.98
Biogas	2.36 billion M^3	1.73
Bagasse for power	800 MW/2000 GWh	0.75
Geothermal for heat	0.4 Mtce	0.4
Geothermal for power	27.8 MW/110 GWh	0.05
Solar water heater	15 million M^2	1.8
Solar homes	15 million M^2	0.37
Solar cooker	244,000	0.04
Solar PV	13 MW/28.7 GWh	0.01
Wind	242 MW/641 GWh	0.24
Total		305

[a] Note: Micro hydro not included.

heaters, biogas, producer gas from agricultural residues, and electricity from wind and microhydro, could provide some viable alternatives for rural and remote areas. They have the potential in competing with the high cost of LPG and coal gas and meanwhile improving indoor air quality (see also Section 2).

A number of new and renewable energy projects, under the Tenth Five-Year Plan, are currently being carried out at nation wide. Since these projects are and will be of significance to the long-term development of the rural energy sector, it would deserve some discussions.

1.2.2.1 The "Brightness" Project By the end of the last century, about 7 million households, accounting for 3.55% of total rural households, have yet to be electrified. Most of these households are located in the western region and the majority of them are of minority origins. To address this situation, the "Light" project was initiated by the national Planning Committee and supported by several other countries [e.g., Japan, Germany, the Netherlands, the United States, and the United Nations Development Programme/Global Environment Facility (UNDP/GEF).] The 10-year target of the project is to provide electricity, in average 100W/person, to 23 million people in remote areas with renewable technologies such as wind and solar PV. The first period of the project (from 2000 to 2005) will install 1.78 million sets of household systems and 2000 sets of village systems, which will require, approximately, a total of US$1.2 billion. It is expected that the majority of the investment will come from the users themselves, while only 6.7% will be provided by the central and local government by means of subsidies, especially for the village systems and service networks. The project, in its first period, will focus on the development of domestic industries of wind generators and solar PV; the establishment of sales and a after sale service network; technical training on installation, optimization, repair and maintenance; standardization; and quality control mechanisms.

1.2.2.2 Biomass Gasification Demonstration Project With 65 million rural Chinese without electricity, 70 million are in shortage of cooking fuel. As shown in Fig. 2, agricultural residues, mainly straw, account for 35% of rural household energy use. The average efficiency of straw directly used for cooking is just 5 to 8%; however, after gasification, the overall efficiency reaches about 35 to 45%. Most of all, converting solid fuel to gaseous fuel could bring huge environmental, social, and economic benefits, such as improved indoor air quality and health, reduced of load of house work for women, the sense of modernization of rural communities, and reduced greenhouse gas emissions. However, these externalities are not internalized financially; thus, in average the daily expense on fuel in terms of producer gas is 7 to 8% higher than that in terms of coal gas. One of the first demonstration projects was initiated in Shangdong province with a gasification system and a village gas pipe line installed in 50 villages. It brought significant benefits to those villages in terms of reducing smoke and liberalizing women from the drudgeries of fuel handling. From 1999 to 2000, the national planning committee budgeted 12.5 million RMB ($\sim$US$1.5 million) for wider dissemination of the technology in seven other provinces. It is expected that 53,900 households would be using producer gas for cooking. Standardization and commercialization of gasifiers and gas stoves are key issues to be addressed, along with the demonstration projects (see Fig. 4; see also Section 2).

Modern energy is the main engineer of economic growth. Meanwhile, national policies and measures to promote such growth have a profound impact on sustainable energy development. One of the strategic developments focuses is the current Western Development Plan. Through rational and efficient exploration, distribution, and utilization of resources, the development plan aims at stimulating

FIGURE 4 Biomass gasifiers manufactured and used in Yunnan, China.

economic growth in the West and minimizes the social and economic disparities between the western and eastern regions. Western China is rich with both renewable and conventional energy resources. However, the demand of energy mainly comes form the Eastern coastal areas. Two major programs under the development plan are to transfer both natural gas and electricity (generated from coal) from the western to the eastern region. Mostly energy resources are located in rural areas and the pipe line and network will be most likely constructed along remote villages and rural towns, thus, it is hoped that these programs would enable those rural households to access clean energy services. These programs would also encourage the growth and development of energy service industries in the western region, which could provide employment and income generation opportunities.

The previously discussed rural energy development in China has had a profound impact on the sustainable development in all of its three dimensions (i.e., environment, economy, and society). The impacts are complex and interrelated. The following sections exam several characteristic issues in each of the three dimensions, and through that, the challenges and possibilities in advancing rural energy development are discussed.

2. ENERGY AND RURAL ENVIRONMENT

Small-scale coal mining has caused serious environmental problems in rural China. As mentioned in the previous section, these mines are mostly collectively or privately owned by the counties; little capital was invested in technology and equipment. The average efficiency of coal production (e.g., resource recovery)

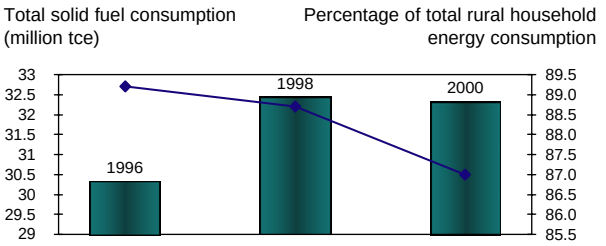


FIGURE 5 Reduced solid fuel consumption in rural household.

is much lower than 50%. Most of the mines are very labor intensive and some have the productivity as low as 1t/labor. Few environmental control measures have been adopted in these mines, and only a small portion of coal produced has been further processed with, for instance, de-sulfur technology. From the end user side, rural households and industries are the main consumers of the low-quality coal from the small mines. Rural industries are known to have outdated equipment with low efficiency (as presented in Table III). This situation has worsened as many high pollutant and energy intensive industries have moved to rural areas. In average, the brick, cement, paper, coking, and casting industries in rural areas shares 8% of the total production of the country, while they account for more than 50% of the total energy consumption in these industries. For rural households, since the late 1990s, the consumption of solid fuels (e.g., biomass and coal) has increased only marginally and their share in the total household energy use started decreasing, as shown in Fig. 5. This indicates that with the promotion of improved technologies, efficient stoves, coal, and biomass energy intensity decreased and resulted in energy saving and, consequently, emission reduction both indoors and outdoors. However, it is difficult to assess the actual changes and effects due to the lack of detailed data.

Overexploration of fuelwood still persist in some rural areas despite the relatively successful reforestation program since the 1970s. From the 1990s, fuelwood supply and consumption has been roughly kept in balance due to mainly improved energy efficiency and promotion of coal replacing fuelwood; however, statistics show that fuelwood consumption is higher than the sustainable supply at various degrees in more than ten provinces. For example, Yunnan Province, located in the Southwest of China, has relatively successful solar heating and biogas programs. However, data in 1999 show that fuelwood consumption was about 32 million M^3 while the total sustainable supply from all type of forest was only about 14 million M^3. Several counties in Yunnan

province are located upstream of, among others, the Yangtze and Mekong River. Years of over exploration of the forest for timber and fuel in these areas have resulted in soil erosion and degradation.

In 1998, agricultural plantation generated approximately 600 million tons of residues, of which 27% was utilized mainly as animal feed and for industries, 15% was returned to the field as natural nutrients, and 47% was used as fuel. On one hand, the declined consumption of agricultural residues as fuel has made the returning of residues back to the field possible. On the other hand, this has led to about 65 million tons of residues left for disposal. Since decomposing stem residues is a rather long process, most of it is burned openly to prepare the field for the next crop. In many areas of central China, this has become a serious seasonal environmental problem.

2.1 Rural Energy and Emission Reduction: The Potential of Biomass Gasification

Analysis has confirmed that China actually has reduced its greenhouse gas emissions, even as its economy has grown rapidly (as shown in Fig. 6). However, due to lack of data from monitoring and evaluation, it is difficult to analyze the share of the emissions from rural energy consumption. Estimates of all energy efficiency projects claim more than 32 million tce energy saving capacity (30 million tce from the improved fuel/stove projects in rural homes) per year, which would be responsible for about 70 million tons CO_2 avoided, that is, two-thirds of the CO_2 reduction from coal in 1998–1999 and twice of the total CO_2 reduced of the year. The steeper CO_2 reduction from coal indicates higher coal efficiency (comparing with coal consumption) and the increases of CO_2 emission from other sources.

As discussed previously, biomass gasification provides a promising alternative in replacing solid fuels for cooking and heating in rural areas. The amount of agro-residue available for energy use in 1998 was about 350 million tons (possible losses already taken into account), 58% of the total residues generated. Since 285 millions tons of the residues are already used as fuel, it could be assumed that all the available residues could be gathered and transported without imposing any major changes to practices and workload. Assuming 350 million tons of residues could be converted to gashouse fuel for rural cooking and heating, their contribution to emission reduction can be estimated. The average emission factors of coal and biomass fuel for household applications are shown in Table IV.

Based on technical specifications of producer gas in China, it is assumed that gas production rate is 2 M^3/kg residues and 1 M^3 of producer gas would replace 0.25 kgce; thus, 350 million tons of residues would generate 700 billion M3 of gas and equal to 175 million tce.

When producer gas replaces all the residues used for fuel (285 million tons) and partially replaces fuelwood (i.e., 32 million tce), it could avoid, in total, 109 million tons of CO_2; this is about the total emission reduction from coal in 1998–1999. When producer gas replaces all the residues (285 million tons) and partially replaces coal (32 million tce), it could avoid, in total, 167 million tons of CO_2.

2.2 Rural Energy and Health: A Possible Win-Win Strategy

The simple household solid-fuel stoves commonly used in rural areas do not obtain high combustion efficiency. They emit a substantial amount of fuel carbon as byproduct of incomplete combustion (PIC)—such as carbon monoxide (CO), methane (CH_4), and total non-methane-hydrocarbons (TNMHC)—as well as sulfur dioxide (SO_2) and

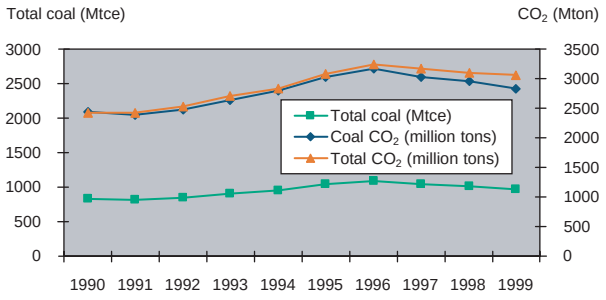


FIGURE 6 China CO_2 trends.

TABLE IV

Average Missions Factors of Household Fuels in China

Fuel	Overall efficiency (%)	g CO_2/kg fuel	g CH_4/kg fuel	Total g CO_2/kg fuel[a]
Crop residues	14	1130	5	311
Wood	23	1520	5	311
Coal	35	2280	3	2466

[a] Note that wood and residues are assumed to be CO_2 neutral. And in a 20-year time horizon, the Global Warming Potential (GWP) of one CH_4 molecule is 22.6 time higher than that as CO_2.

particulates. Some of these PICs are hazardous to health when inhaled under the concentrations commonly found in homes using unvented biomass or coal stoves. As discussed, China has carried out, in the past 20 years, efficient cooking stove dissemination programs, which claims that the various technologies have benefited 184 million households by the end of 1998, and this figure is increasing at 2.2% per year. However, again, no monitoring and evaluation data are available to exam the true effect of the stove program, particularly in terms of health improvement. With improved efficiency, carbon emissions could be reduced as discussed in the previous section. It could also result in fuel saving, if more energy is actually delivered in less PICs or if more carbon is converted to CO_2. Table V presents data from site analysis of Chinese stoves. The data show that improved efficiency does not necessarily accomplish emission reduction. However, processed and high-quality fuel is always preferable.

A study of the health benefits from emissions reduction in China has given some further interesting results. It was found that greenhouse gases (GHG) reductions resulting from changes in energy use are generally accompanied by substantial short-term health benefits. However, the degree of health benefits varies greatly with the choice of energy technologies and sectors. The variation appears to be larger when shifting between sectors rather than between technologies, and the largest benefit can be obtained in the household sector. Even conservative estimates show, for example, that the health benefits of one tonne reduction in particulates emission from household stoves are at least 40 times larger than those from coal-power plants. This study also shows that although GHG reduction itself is usually accompanied by increasing costs, when the economic benefits of improved health are counted, there is a net economic benefit for those household energy options, involving a shift from traditional use of coal and biomass. This strongly suggests that emission reduction in the household sector can be a real win-win strategy for China.

TABLE V

Health Hazardous Emissions — Improved Fuel/Stove

Fuel/stove	Efficiency	g PIC/MJ delivered	g NO_x/MJ delivered	g SO_2/MJ delivered
Unprocessed coal	37%	4	0.2	0.02
Coal briquette	14%	24	0.04	0.04
Wood/ improved	24%	20	0.1	—
Wood/brick	13%	12	0.2	—

PIC = CO + CH_4 + TNMHC-C (carbon in total none-methane hydrocarbons) + TSP (total suspended particulates).

3. ENERGY AND RURAL ECONOMY

As essential to industries, modern energy services are prerequisites for rural economic development. Most rural industries and agro-product processes have been limited to solid fuel (biomass and coal) with outdated equipment as main energy input. This, as already discussed, has resulted in unnecessary waste of energy and constraints to productivity improvement and products diversification. Figure 7 is a simplified diagram showing the inter linkages of lack of energy services and its impact on agriculture and local economy, which is complex and dynamic.

By no means, this section could provide a comprehensive overview of such complexities in rural China. Rural energy production, supply, and services as part of the rural economy have also undergone tremendous changes in terms of market potential and development, and the following will exam this aspect in detail.

3.1 Rural Energy Business

As discussed before, since 1996, the number of rural energy enterprises, as well as their market value and employment, has been increasing consistently. The current rural energy industry covers almost all renewable energy technologies; however, their stage of development and commercialization varies greatly. According to China Rural Energy Industry Association, the technologies and products are grouped in six categories and are under the mandate of six committees in the association. The market situation of these technologies/products is resented with some details.

Solar heating—including Solar Water Heater (SWH), Solar Cooker, Green House and Solar Homes—has been developed rapidly in the past 20 years in China and has become the main product of the sector. There are two main types of SWH on the market (flat-panel and vacuum tube), with an average heat production efficiency of 100 ~ 150 kgce/M^2. The total sale of the two types was 5 million M^2 in 1999, which was an increase of 25% from 1998. Total installed capacity of SWH has reached

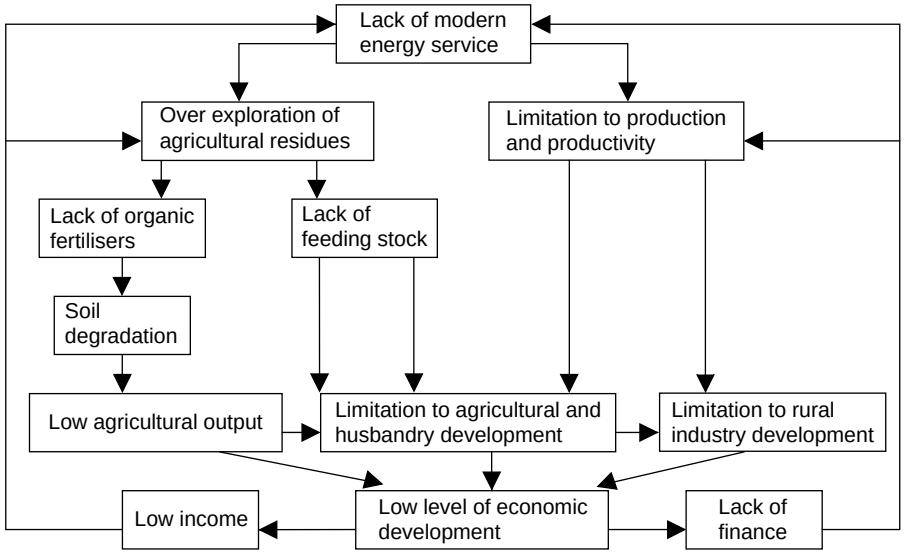


FIGURE 7 Example of energy shortage and its impact on rural economy.

20 million M^2, of which more than 50% is in rural areas. There are nearly 1000 factories with annual production capacity of 6 million M^2, which values more than US$600 million. In total, 20,000 people are employed by these enterprises. By the end of 1998, solar home system (SHS) construction reached 15 million M^2 in total, from which 6.5 million M^2 is located in rural areas. The rate of growth for SHS was 54% per year in the past 10 years. On average, SHS could accomplish 75 to 90% fuel savings on heating. Figure 8 shows a flat-panel solar water heater.

Although there is wide range of products in the energy-efficient cooking devices sector, only two have achieved a large market share. One is the drinking water heater, with a yearly production and sale of more than 300,000 and a value of more than US$12 million. There are more than 1000 factories, mainly small- and medium-sized enterprises (SMEs). The other one is coal-saving stoves. There are five factories reaching an annual production of more than 10,000 stoves.

There are more than 2600 biogas enterprises all over the country with nearly 18,000 employees. The service provided includes digesters (both large scale and household scale) and its design, necessaries, and end use devices, technologies, construction, consultation, repair maintenance, and so on. In total, about 7.6 million digesters have been installed in rural China. Still in technology demonstration stage, there are 740 large and 49,000 medium digesters installed in large farms, industries, and small cities. Although

FIGURE 8 Flat-panel solar water heater, Yunnan province.

small in size, the market for household biogas systems could be promising. Every year 300,000 new digester are being built, and another 100,000 need to be replaced, which implies approximately US$70 million per year. However, one of the key issues in commercializing small digesters is standardization. Semiconstructed digesters have been developed to secure consistent quality as well as reduce the time needed for construction.

Another technology is solid biomass conversion equipment, including briquetting, direct high efficiency combustion, gasification, liquadization, and carbonization equipment and facilities. Although these products are still in the stage of research and development and demonstration, some 50 factories have already been in production and their products are sold in several provinces. A study of the market potential of a village scale biomass gasification system shows that the average income level of a village

affects the current market of the technology. In the case study, it is shown that when yearly household income reaches 10,000 RMB (about US$1200), residues as fuel are abandoned and the financial implication of gasification is more acceptable.

A new type of liquid fuel/stove presents yet another technology. The liquid fuel in this category is mainly referring to methanol and ethanol from the waste of oil and chemical industries. There are about 600 research and production enterprises in this sector, among which several are more than $1 million enterprises. Although the technology is yet to be standardized and further improved, it has already reached more than 300,000 households.

Small-scale electricity production makes up another category. This sector includes small wind, solar, and hydropower industries. More than 50 enterprises are engaged in small wind power research and development and equipment production. Currently, the production of small wind turbines is about 20,000 units per year. Products range from 100W to 5 KW, of which the 100 to 300W types are most in demand. The technology has matured in the past 5 years and production quality has been stable. Solar PV industry has been commercially established. There are more than ten solar PV production lines in China, each with 4.5 MW per year capacity. The annual production has been around 2 MW. In total, 12 MW has been installed, of which 5 MW in rural areas. In average, the system cost is around US$10 to $12 per watt at peak and electricity could be generated at 30 U.S. cents/KWh. Other products associated with the systems include lighting appliances, batteries, and pumps. More than 100 enterprises are engaging in their production, mainly located in Gansu and Qinghai provinces. A small hydro industry has been developed to a relatively large scale with approximately 1 million employees. Meanwhile, the mini hydro industry is in expansion with annual sales of 15,000 systems, mainly in the Southwest provinces.

The development of these industries depends largely on the economic viability of these technologies in reality, which is influenced by several factors: the abundance of the energy resources, the demand and willingness to pay by end users, system configuration and performance, system costs, and financial and policy scenarios. Under these factors, the following specific and direct limitations have been identified based on past project experiences in rural China: There are no unified technical standards. Thus, it is difficult to guarantee the product quality with different standards in different regions. In some cases, markets and consumers are confused with bogus products of low quality; most of the companies are inexperienced in getting loans from banks. Therefore, the shortage of capital is a major barrier for companies to expand their sales and improve after-sales services. With China joining the WTO, increasing domestic and international competition is expected, especially for the PV and wind sectors due to their relatively low level of production technology. Markets need to be closely examined and regulated with proper financial means in order to sustain and develop rural energy enterprises, as well as to reflect their true economic benefits.

3.2 Rural Energy and Income Generation: The Biogas and Eco-Farm Model

As mentioned, biogas technology, for providing cooking and lighting fuel for rural households, was disseminated in the 1950s. However, due to the immaturity of the technology and lack of management and maintenance skills, most of the digesters failed to perform soon after being installed. Similar failures were experienced in the 1970s. More than 7 million units were built in just a few years. Despite the potential environment and health benefits of biogas technology, its economic performance was recognized as a key to success. Two variations of biogas system in combination with eco-farming model have been gradually developed in the past 20 years, the Southern model and the Northern model. The Southern model focuses on the utilization of the sludge from the digester as fertilizer, thus, animal stable, a biogas digester, and fruit/vegetable plantation have become components of one "eco farm." On average, production cost per hectare was reduced by US$360. Financial payback of the system is about 4 years. Due to low temperature, biogas use in the north had been limited. In the Northern model, animal stable, toilet, biogas digester, and green house are built on one spot to ensure sufficient and better temperature for animal growth and anaerobic reaction. The initial investment of the Northern model is about US$480 and the payback time is usually within 2 years. According to survey, households installing the Northern model were benefited with US$380 increased income per year. Currently, 210,000 and 81,300 households have adopted the Northern and Southern models, respectively. They produced more than 4 million tons of fruit and vegetable without using chemicals in 1998. By the end of 1999, 6.7 million households were provided with digesters, and their total gas production is around 2 billion M^3 annually.

4. ENERGY AND RURAL SOCIETY

Economic reform is changing the social structure and lifestyle of rural population. As part of the overall development in rural communities, the increasing availability of improved energy services has contributed to improved living and working conditions, better education facilities, saved time for other activities and leisure.

4.1 Rural Energy and Gender Equity

Based on the law of China, women and men are equal. While this has greatly increased the social significance of women in the cities, such development in rural areas has been lagging behind. Rural women have been, for centuries, taking care of household duties. They are not only responsible for the well-being of children and other members of the family, but also for the basic daily livelihood, such as food, cooking, feeding, and washing, in addition to helping in the field. With the broadening of economic reform in rural China, mobility and business opportunities largely increased, which led to more and more agricultural laborers, particularly men, transferring to nonagricultural sectors and working in peri-urban and urban factories. As a result, more responsibilities fell upon women. Thus, unless the household burden is alleviated, there will be little chance for true gender equality in rural society. For many rural women, fuel for cooking and heating has been drudgery in terms of time and labor spend on fuel collecting and smoke irritation from fuel burning. A typical rural family of five would require 3000 kg of firewood per year, which would need 80 days to collect, an average of 5 hours per day. Plus another few hours to cook food and feed, many have to work at least 15 hours a day. Women as users have deeply realized the benefits of improving their fuel situation and have been actively initiating and participating in rural energy projects.

The Forest Research Institute of Yunnan province in Southwest China has carried out a comprehensive research on women and rural energy based on case studies in nine different counties in the province. (In Yunnan province, the Forest Department of the provincial government is in charge of rural energy development.) The findings of the study show that although the form and type of technologies vary from project and region, several common achievements can be identified as follows:

- The energy project has brought great benefits to women in terms of health and living environment. The physical workload has been reduced, and in many cases less fuel consumption resulted in savings and more time available for increased opportunities for other activities.
- Decision making on fuel and household expenditures in rural China mostly relies on women. Thus, knowledge about energy technologies and related issues is as essential as active participation in energy projects. These energy projects have provided opportunities of training and self-improvement for many women, which in turn have ensured the success of project/technology promotion and implementation.
- Engaging women in energy projects has been seen as one of the means to increase their self-confidence in taking part in community development and getting involved thereafter in all other fields of rural development, some even becoming a leading force in rural economic reform.

4.2 Rural Energy and Community Development

The lifestyle of rural communities links closely to energy production and use. Issues discussed in previous sections all have a profound and complex impact on rural communities. The availability of modern energy services, such as electricity, heated water, and liquid cooking fuel, has affected rural life as it has in urban areas. It has contributed to the modernized agriculture and rural industries, improved the indoor air quality, enriched life with new information and opportunities, increased the speed of daily rhythm, and changed traditional habits and ideas. Closure of rural coal mines unavoidably seriously affected employment and income generation, directly and indirectly, since these coal mines employed more than 3 million rural laborers and were the main source of fuel for most rural industries where more than 20 million were employed. Moreover, coal has been the economic foundation of half of the coal-producing counties. The communities in these regions are in transition. Social security and training for reemployment are main issues of concern, among others.

Meanwhile, evolution of rural society, such as reduced family size and improved living standards, has in turn challenged the energy supply, technology, and policy. Stove programs in the past have helped millions of households to build stoves for large multigeneration families. Many of them are not used regularly, particularly when gaseous fuels became available and affordable. It is often observed that

various types of fuel are used in rural household due to mainly economic and security reasons. More and more agro-residues have been left unused when better quality fuel was available and higher living standard was achieved. The ban of open burning of residues would have to be combined with control and enforcement, which is rather costly. Also it does not provide alternatives to utilize the materials. As already discussed, gasification technology has been demonstrated to be a viable solution. One of the key issues on village scale application is the management of a gasification system. Village communities should be mobilized to either direct management or cooperation with service providers. Fuel supply policies, financial incentives, and energy market should be designed with these issues in mind.

5. LOOKING FORWARD

China's Tenth Five-Year Plan (2000–2005) has set several targets in relation to rural energy. By the year 2005, annual modern renewable energy utilization is expected to reach 13 Mtce (excluding hydroelectricity and traditional biomass). This would avoid 10 million tons of CO_2 and 600,000 tons of SO_2 and particulates emissions. It would also electrify 5 million to 6 million people and provide employment for 200,000 people. Annual production capacity and total installation of solar water heaters are expected to be 11 million M^2 and 64 million M^2, respectively, and for PV, 15 MW, and 53 MW, respectively. Wind power installation is expected to reach 1.2 million KW. Annual production of equipment would be equivalent to 150 ~ 200 MW capacity. Finally, large-scale biogas application in industries and husbandry and biomass gasification are expected to provide 2 billion M^3 fuel gas.

Compared to energy use in 1998, modern renewable energy utilization will more than double. Although consumption of modern renewables has been increasing fast, financial support remains crucial given the current immature development of the market and business environment. Experiences show that government plays an essential rule in this respect as an enabler rather than a controller. Financial support from all levels of government offices should be rightly focused, catalytic in nature with no distortion of the market, and costs and benefits truly reflected in economic values.

Traditional energy resources in many rural areas have been considered free. This concept does not only undermine the importance of energy efficiency, but also hinders the promotion of clean fuels, such as biogas and producer gas from gasification. Commercial utilization of biomass resources will be an important step toward better resource management, which is essential for sustainable rural development.

With the increasing gap in income and living standards among different rural regions, energy demand varies greater and greater from area to area. Income for farmers without any other source of income, accounting for 68% of total rural households, has been decreasing in 1999 and 2000, while the national average of rural income has increased 2%. Lower income households, most likely, will still rely on traditional biomass fuel, whereas, higher income farmers will demand modern and clean energy services. Rural energy policy and technology will need to be diverse and locally customized.

SEE ALSO THE FOLLOWING ARTICLES

Biomass Gasification • *Biomass: Impact on Carbon Cycle and Greenhouse Gas Emissions* • *Coal, Fuel and Non-Fuel Uses* • *Development and Energy, Overview* • *Indoor Air Quality in Developing Nations* • *National Energy Policy: China* • *Rural Energy in India* • *Sustainable Development: Basic Concepts and Application to Energy* • *Women and Energy: Issues in Developing Nations* • *Wood in Household Energy Use*

Further Reading

Byrne, J., Chen, B., and Wallace, W. (1998). Economics of sustainable energy for rural development: A study of renewable energy in rural China. *Energy Policy* **26**(1), 45–54.

China Environment and Development Review (2001).China New and Renewable Energy Network, www/newenergy.org.cn/. China Rural Energy and Its Market Potentials, www.scncny.-net/new_page_70.htm. (Accessed in 2002)

China Rural Energy Statistics, Ministry of Agriculture, 1997 (1998–1999).

Commercialization of Solar PV Systems in China, Center for Renewable Energy Development, Energy Research Institute and the U.S. Department of Energy (1996).

DOE/EIA and OECD data tables, www.geocities.com/~combusem.EIA China Country Analysis Brief, www.eia.doe.gov/emeu/cabs/china.htm.

McCreary, E. I., Gu, A. Y., Loose, V. W., and Joseph M. Roop, J. M. (1996). China's Energy—A forecast to 2015, Los Alamos National Laboratory.

Rural Energy Conservation Technologies, Bailiang Zhang (1999). Statistical Abstract. www.stats.gov.cn/ndsj/zgnj/2000/G01c.htm (-G13c), 2000.

Second Analysis Confirms Greenhouse Gas Reductions in China, www.nrdc.org/globalWarming/achinagg.asp.

Wang, J., and Zhao, Z. (2001). "China Energy Development Report."
Wang, X.D., and Smith, K. R. (1998). Near-term Health Benefits of Greenhouse Gas Reductions: A Proposed Method for Assessment and Application to China, WHO/EHG/98.12.
Wood Energy Basics, Regional Wood Energy Development Programme (RWEDP) of FAO/UN, www.rwedp.org, 2000.
Wood Energy, Health and Environment, RWEDP Field Document 58, 2000.
World Energy Assessment, UNDP (2000).
Energy for the poor, DIFD (2002).
Zhang, J., *et al*. (1999). "Greenhouse Gases and Other Airborne Pollutants from Household Stoves in China: A Database for Emission Factors, Atmospheric Environment."
Zhou, F., and Wang, Q. (2000). "China Energy 50 Years."

Rural Energy in India

IBRAHIM HAFEEZUR REHMAN
The Energy and Resources Institute
New Delhi, India

1. **Energy Consumption**
2. **Rural Electrification**
3. **Subsidy Regime for Facilitating Access and Equity**
4. **Role of Renewables in Facilitating Transition and Decentralization for Remote Areas**
5. **Conclusion**

Glossary

administered price mechanism A government-run system of price subsidies aimed at making household energy more affordable.

commercial energy Typically this refers to fossil fuel-derived fuels, such as liquefied petroleum gas, kerosene, natural gas, and coal.

rural electric cooperatives (RECOs) Public–private institutions whose mission is to extend the network and provide services in villages.

traditional energy Typically this refers to biomass-based energy sources, such as fuelwood and animal manure.

India ranks sixth in the world in terms of energy demand, accounting for 3.5% of global commercial energy demand in 2001. India's energy sector is administered at the apex level by five ministries (including the Department of Atomic Energy), whereas the Planning Commission carries out the energy planning exercise for different sectors in the form of 5-Year Plans. Such plans have advocated a strategy focusing on rationalizing tariff, giving a greater role to market forces, enhancing private sector participation, and promoting renewable sources, among other strategies. The Ninth 5-Year Plan (1997–2002) stressed the need for research and development in the fields of energy conservation and demand-side management. It also focused on enhancing the pace of exploration and development in the hydrocarbons sector to reduce dependence on imports. A major thrust was placed on infrastructure development, particularly energy and power, with plans for enhancing related public expenditure. By the end of the Ninth Plan, the share of commercial energy was expected to increase from 66 to 75%, and the share of noncommercial energy was expected to decline from 34 to 25%. The main elements of power sector reforms in the Tenth Plan (2002–2007) include attempts at rationalization of power tariffs, unbundling of state electricity boards (SEBs), gradual privatization of distribution, and 100% metering of all consumers. In addition, it focuses on hydroelectricity and the possibility of expanding nuclear energy for electricity generation. The implementation of these plans is the responsibility of respective ministries that coordinate with the state governments; the National Development Council provides a forum for coordination at the highest levels of government. In terms of rural access, the more than 700 million rural poor in India have limited access to energy sources, such as electricity, and nontraditional cooking fuels, such as liquefied petroleum gas (LPG). Consequently, the rural energy scenario in India is dependent more on nonmonetized, biomass resources than on monetized or commercial forms of energy. This is amply borne out by the fact that over 72.3% of Indian households depend on traditional sources of energy, such as fuelwood, animal dung, and crop residues, for meeting their cooking and heating requirements. In this regard, over the years, the government, through its various committees and groups, such as the Fuel Policy Committee, the Working Group on Energy Policy, the Advisory Board on Energy, and the Energy Demand Screening Group, has formulated programs aimed at rural energy provision. The efforts have been primarily in the form of national programs for rural electrification, biogas development, improved cookstove dissemination, and promotion of renewable energy technologies (RETs), such as solar cookers and lighting devices. However, despite the existence of these programs for nearly two decades, their impact on the rural energy scenario has been limited.

 © 2004 Elsevier Inc. All rights reserved.

1. ENERGY CONSUMPTION

In the context of energy consumption, considerations of equity have several dimensions. The most important of these is that adequate energy of good quality is available and accessible to all, such that the poor are able to meet their needs, particularly their needs for cooking fuel and basic lighting. In developing countries such as India, the aspect of quality is important because the use of biomass fuels with low calorific value has serious health implications on account of resultant high levels of indoor air pollution. Clean and good-quality fuels, therefore, need to replace the large-scale use of inferior fuels that are prevalent in India.

The per capita energy consumption in India stands at 479 kg of oil equivalent, which is only 20% of the global average in 1999. One of the main reasons for this is the fact that existing rural energy supply programs often bypass the rural poor because they are unable to interpret their needs and priorities. Even programs specifically meant for the rural poor—supply of kerosene through the public distribution system (PDS) or improved cookstove programs—do not reach them properly or are rejected by them because the initial design does not take into account their needs and priorities.

1.1 Electricity Consumption

With regard to electricity, the consumption pattern was 32% for agriculture, 31% for industry, and 20% for domestic use, and the remaining uses, including commercial and street lighting, make up the final 17%. At the national level, electricity and kerosene are the primary sources for lighting, in both rural and urban areas, and accounted for 99% of households. There has been an increase of 6.5% (from 37 to 43.5%) in the proportion of rural households using electricity as the major source of lighting since 1993–1994. On the other hand, there was a decrease in the percentage of households using kerosene as the primary source of energy for lighting, from 62 to 55.6% in rural India (Fig. 1).

1.2 Consumption of Biomass Fuels

Biomass fuels provide 85–90% of domestic energy and 75% of all rural energy (Table I). Of the total rural energy consumption, approximately 52.5% is met from fuelwood, and fuelwood consumption in 2001–2002 was estimated at 223 million tons (MT).

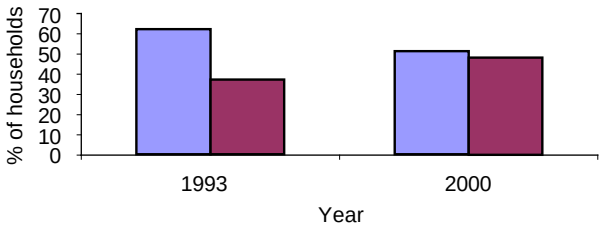


FIGURE 1 Percentage of rural households in India using electricity and kerosene for lighting.

TABLE I

Consumption of Traditional Fuels

- Biomass fuels provide ~90% of the energy in the rural domestic sector and 65% of all rural energy
- Fuelwood is the largest contributor to energy use in rural areas, followed by animal dung and crop residues
- Women and children collect most of the fuelwood, in the form of twigs and branches
- Inefficient use of biomass in traditional devices has serious global and local environmental effects
- Traditional fuels are responsible for adverse impacts on health, especially the health of women and children

Source. Reprinted from National Sample Survey Organization (2001).

Of this, 180 MT was for rural households and the balance was for cottage industries, large hotels, etc. However, against this demand the annual sustainable yield of fuelwood from different land sources is only approximately 86 MT. The consumption of animal dung and agro-waste is estimated at 130 MT; this does not include wet dung used for biogas plants.

The 55th Round of the National Sample Survey Organization survey indicated that firewood and chips were the most important source of energy used for cooking in rural India, with 76% of households dependent on it in 1999–2000. The survey also revealed that only 3% of households have switched from this source to other sources of energy since 1993–1994. The Rural Energy Database compiled by The Energy and Resources Institute has also indicated the highest per capita per day consumption for fuelwood (Fig. 2).

Among rural households, the use of fuelwood and chips declines and the use of kerosene and LPG increases with rising incomes. However, the transition across income classes is not significant, with the result that even in the highest income category, over 70% of rural households depend on fuelwood chips or dung cakes for cooking.

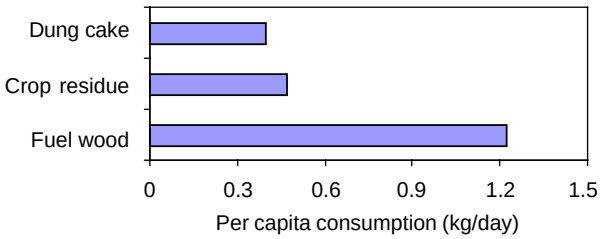


FIGURE 2 Daily per capita consumption of biofuels in rural households in India. Reprinted from TERI (2000), with permission.

There are regional variations in biomass consumption, borne out by the fact that the consumption of fuelwood ranges from 2.04 kg per household per day in Haryana to 14.14 kg per household per day in Orissa. The fuel mix also varies from region to region depending on the resource endowments. Fuelwood consumption is high in states that have considerable forest cover. For instance, in Arunachal Pradesh, it is 10.4 kg per household per day, and in Manipur, it is 10.34 kg per household per day. On the other hand, consumption of dung cakes is high in states such as Uttar Pradesh (4.95 kg per household per day) and Rajasthan (3.41 kg per household per day), which have little biomass cover. Crop residues, being inferior in quality, are used in most areas as backup fuels when there is a shortage of other fuels.

1.3 Commercial Fuel Consumption

Over the years, the dependence of rural households on dung cakes, coke, and coal has fallen and dependence on kerosene and gas has risen. However, in rural areas, the overall shift to commercial fuels has been marginal. Among commercial fuels, kerosene is used predominantly for lighting. Estimates put total kerosene consumption for rural lighting at close to 7 MT. Subsidies on petroleum products, such as kerosene, have contributed to an increase in the demand for kerosene. However, despite a significant increase in the supply of commercial energy, the consumption of commercial fuels, such as LPG, is still negligible in rural areas, with less than 5% of the households using the fuel for cooking.

1.3.1 Kerosene

Kerosene is the main source of lighting in rural India. The total kerosene consumption in India in 2000–2001 was estimated at 10.5 MT, over 50% of which was in the rural areas. In India, there are two systems for the supply of kerosene to households. One is through the PDS at subsidized rates, forming the main source of supply for rural consumers. The other is the parallel marketing system at market-determined prices. Traditional kerosene lighting devices are used by 74.62 million households throughout India. As is well known, the fuel consumption of these devices is much too high and the quality of light is very poor. Furthermore, their use has adverse impacts on the health of people. Hence, nonelectrified rural households and a majority of electrified households continue to use kerosene for lighting.

The per capita allocation (PCA) of kerosene for states ranges between 10 and 24 kg/year. In 1999–2000, the maximum PCA of kerosene had been frozen at 24 kg/year. Moreover, a private party can import kerosene after obtaining a rating certificate from an approved rating agency. The kerosene is to be sold at market-determined prices by the private parties through their own distribution networks. The quantity of kerosene marketed under the parallel marketing scheme in 1999–2000 accounted for approximately 10% of the total kerosene consumption in the country. The contribution of kerosene through parallel marketing in rural areas, therefore, remains very small.

1.3.2 Liquefied Petroleum Gas

Four public sector undertakings—Bharat Petroleum Corporation Limited, Hindustan Petroleum Corporation Limited, IBP Company Limited, and Indian Oil Corporation Limited—are engaged in the marketing of LPG in the country. With the increased availability of LPG, the number of customers enrolled by these companies has also been increasing. Consumption has grown from 100 MT in 1985 to over 6000 MT in 1999. The number of LPG customers served by the four oil companies as of January 1, 2000 was 43.6 million. However, the rural penetration during this period was just over 1% of the total households.

It is also estimated that in the next decade there will be a rapid increase in the consumption of LPG. Given the near-saturated scenario of the urban market, a large portion of this increase in estimated consumption will occur in rural areas. This is due to the fact that with the increasing nonavailability of traditional fuels, and LPG being a convenient and safe cooking fuel, the government has decided to commence marketing of LPG in rural areas. Furthermore, with a view to penetrate the rural areas, the Ministry of Petroleum and Natural Gas approved a special scheme in July 1999 for the release of new LPG connections in the rural sector against the

surrender of corresponding kerosene quota through the following strategies:

- Opening of extension counters in rural areas within a 15 km radius of the normal trading area of existing distributors operating in the urban areas.
- Award of interim LPG distributorships to the state civil supplies corporation or the concerned state government body until a regular distributor is approved.

In November 1997, the government decided to dismantle the administered price mechanism in a phased manner. As a result, the subsidy on domestic LPG was decreased to 15% in 2000–2001. Moreover, in order to increase LPG availability, the private sector has been allowed to participate in the schemes of parallel marketing of LPG since April 1993. Under the scheme, a private party can undertake the import of LPG and the procedure is the same as that for kerosene. As of January 1, 2000, 695.1 MT of LPG had been imported by the private sector under the scheme.

Despite the fact that LPG has been available for a number of years, it has not penetrated the rural areas of the country. The continued subsidy on LPG has basically hampered the growth of the parallel market and, thereby, adversely affected the penetration in rural areas. On the other hand, the low-income levels in rural areas are responsible for low buying capacities of the rural communities. Moreover, people in the rural areas are not shifting to LPG, as fuelwood is cheap and easily available. For LPG, delivery is not to their homes and means of transportation are not available. Also, there is a fear that LPG is a dangerous fuel. In most companies, the policy is that cylinders are refilled after 21 days. However, owing to logistic and economic reasons, rural customers usually take 3 months to get the cylinders refilled.

1.3.3 Consumption for Agriculture

The agriculture sector is the second largest energy consumer in rural India. Here, the animate energy (human and draught power) accounts for a substantial portion of the total energy consumed. However, the share of human labor and animal power in total energy consumption for agriculture has decreased over the years, indicating progress in farm mechanization. Agriculture accounts for approximately 32% of the total electricity and 95% of the total water consumption in India. Diesel oil and electricity are the major sources of energy used for irrigation. Very small numbers of solar pumps and wind-energy-based pumps are also used.

With the introduction of Green Revolution technologies, the demand for irrigation increased. To meet this increased water requirement, rapid farm mechanization programs along with policy measures such as underpricing of power and irrigation were initiated. As a result, the number of pump-sets has increased and was reported to be close to 20 million, of which just over 12 million had been energized. In 1999–2000, the average farm power availability was 1.15 kW/ha, whereas for desired levels of productivity and appropriate postharvest technologies, a minimum essential estimate is 2 kW/ha.

As mentioned earlier, irrigation is the most important end use, but owing to subsidies, the marginal cost of pumping water is zero. This leads to energy wastage, overpumping, and inefficient crop selection. Flat-rate pumping also masks the true cost of power to farmers. One study indicated that significant energy losses are associated with the distribution of electricity and in the poor selection, installation, maintenance, and operation of electrical motor pump systems.

Suboptimal allocation of resources toward the distribution system places a further burden on the already vulnerable system, causing deterioration in the quality and reliability of supply, leading to further burnout of motors and problems of overloading. The situation is further compounded by the fact that there are a number of illegal connections and cases of cheating in terms of practice of "motor nameplate switching," whereby higher capacity motors have their labels altered to indicate lower ratings. In a flat-rate tariff system, this further robs the utility of its legitimate revenue.

2. RURAL ELECTRIFICATION

Rural electrification received attention only after the formation of the SEBs in 1948. The Rural Electricity Corporation (REC) was formed in 1969 to administer the Central Plan outlay and provide loans to the SEBs and rural electric cooperatives (RECOs) for rural electrification. The RECOs were entrusted with the task of extending the network and providing services in villages. Thus far, 41 RECOs spread throughout 12 states have been sanctioned by the REC, of which 33 are still in operation and 8 have been taken over by SEBs. The important players in the power infrastructure scenario in India are the SEBs, Power Grid Corporation, Power Finance Corporation, public and private power-producing utilities such as the National Thermal Power

Corporation, Central Electricity Authority, and electricity regulatory commissions. All of the above entities play a very important role in the electrification of rural areas.

Of the 587,258 villages in the country (1991 census), 507,451 (86.4%) were declared electrified by December 31, 2000. Of the over 80,000 nonelectrified villages, most (99.14%) are in Uttar Pradesh, Bihar, Orissa, West Bengal, Meghalaya, Arunachal Pradesh, Tripura, and Rajasthan. The Tenth Plan envisages covering approximately 62,000 of the said villages and the remaining 18,000 remote villages are proposed to be covered by renewable and decentralized sources of energy.

On the other hand, although the per capita consumption of electricity in India was 360 kWh in 1998, the demand for electricity has been increasing at a faster rate, resulting in energy shortages. In order to further the efforts of extending the grid to rural areas, the Government of India launched special programs for rural electrification. One such program was the Minimum Needs Programme, introduced in 1974 with the objective of establishing a network of basic services and facilities in all areas within a given time frame. One of the basic services identified was rural electrification. Earlier, a 100% loan was released to the SEBs under this scheme, but since 2000–2001 it has been released to the state governments. Also, approximately 3.8 million households have been provided single-point light connections under the Kutir Jyoti scheme (1988–1989) for poor rural households.

A 7-year plan is being implemented for rural electrification and has been made part of the minimum needs program; funds are available under the Pradhanmantri Gramodaya Yojana. Under this scheme, 50% of the money is available in the form of grants. In the 2002–2003 budget, interest subsidy has also been provided to states borrowing under this scheme. Money can also be drawn through the Rural Infrastructure Development Fund. Moreover, the REC is also working on a scheme for electrification of 0.4 million hamlets. In addition, discussions are under way with the Ministry of Nonconventional Energy Sources regarding the electrification of 18,000 villages, which cannot be reached through the grid. These villages would be provided electricity through renewables.

However, a disturbing trend is that the rate of electrification of villages has been steadily declining after the Seventh Plan. State agencies are not keen on investing in areas of village electrification and pumpset energization. They do not consider them remunerative. Also, with increasing subsidies, the returns on the investment affected the loan-repaying capacity of the SEBs.

As mentioned earlier, 86% of the villages have been electrified. The remaining villages are in six states. However, there has been an 85% reduction in the outlay for rural electrification in the Ninth Plan as compared to the Seventh Plan. The situation is further compounded by the definition of village electrification as adopted by the Government of India: "A village will be deemed to be electrified if electricity is used in the inhabited locality within the revenue boundary of the village for any purpose whatsoever." Therefore, the electrification of a village does not imply that all the houses in a village have electricity. Moreover, long and cumbersome procedures for obtaining connections and paying bills and the distance to repair facilities were some of the factors that affected acceptability. It has also been observed that, starting from the electrification of a village to the procurement of a connection for a household, the time factor can vary from a couple of months to a few years.

Finally, rural India is characterized by small human settlements and given the fact that the energy requirements of these settlements are much lower than in urban and industrial centers, there is high cost of transmission along with severe transmission and distribution losses. Also, although close clusters characterize rural settlements in the plains, settlements in the hilly and forest regions are highly dispersed. Given such a sparse nature of the demand, it becomes economically infeasible to provide power transmission lines to a few houses with a load of a fraction of a kilowatt. Consequently, rural electrification programs have proved expensive and, with the existing constraints on incomes, customers have just not been able to pay for electricity supplies received by them.

3. SUBSIDY REGIME FOR FACILITATING ACCESS AND EQUITY

One of the fundamental causes of rural poverty is inaccessibility to the minimum amount of energy to meet people's basic needs. Hence, the decision makers and planners must assign a high priority to energy provision, irrespective of communities' ability to bear the costs of supply and delivery. Perhaps one of the most critical issues in rural energy development is the nonmonetized nature of many aspects of the rural economy—in particular, the bulk of energy supplies in the form of wood and other biomass fuels. Owing

to availability of "free" sources of energy, it has not been a priority in terms of expenditure on basic amenities. However, in order to move beyond satisfaction of basic needs toward higher levels of economic productivity does mean the introduction into rural societies of modern energy sources.

Recognizing the need to provide access to energy for the poor, the government has provided subsidized energy through either the PDS or the State Electricity Boards. Typically, these subsidies were provided to the residential and agricultural sectors in the form of LPG and kerosene for the former and electricity for the latter. However, the merits of government subsidies designed to accelerate substitution are debatable. Owing to the price distortions that they produce, it is inevitable that most of the subsidies benefit higher income strata and commercial establishments rather than disadvantaged households.

3.1 Subsidy for Irrigation

The total amount of the subsidy on electricity for agriculture and the rural domestic sector has been increased from 202,099 million rupees to 330,814 million rupees in 1999–2000. Of this, just over 70% is provided to the agricultural sector for meeting its irrigation requirements. In contrast, at the all-India level, only 10–12% of the cost of power supply for irrigation purposes is recovered. Moreover, Gulati *et al*. have estimated that the work of each unit of electricity going to agriculture in terms of its contribution to crop output turns out to be 9 rupees per kilowatt-hour for net output at 1995–1996 prices. The farmer, on the other hand, on an average, pays only 0.2 rupees per kilowatt-hour for power. Consequently, over the years, the impact of the subsidies has been twofold: (1) it has led to financial and technical starvation of SEBs and (2) highly subsidized power has led to the overexploitation and misuse of groundwater.

3.2 Deterioration in the Condition of State Electricity Boards

Owing to poor or negligible revenue, the operating as well as financial condition of most of the SEBs has deteriorated over time. The main cause is their inability to recover costs from different consumer segments. For SEBs to break even, state governments need to make budgetary provisions in excess of 5% of their total revenues. If the SEBs are to earn a commercial rate of return (15–16%) on their assets, the subsidies must be five times more than those needed to break even, so that the required subsidy, in 1999, would equal 25% of the state government revenues, or 12 billion dollars.

3.3 High Rate of Groundwater Withdrawal

The rate of groundwater withdrawal is closely linked to energy prices. In most states, power to agriculture is supplied at a flat rate based on the capacity (horsepower) of the pump rather than on the actual quantity of electricity consumed to pump the water. Thus, a high implicit subsidy is built into the system. As a result, many regions in India are witnessing shortages in water supply, with groundwater levels having fallen as much as 1 to 3 m per annum to levels more than 70 m or more below their levels 30 years ago. Approximately 12% of all aquifers are severely overdrawn. Overexploitation of groundwater by farmers has lowered water tables. Overexploitation is also a significant contributor to India's growing carbon emissions, as considerable additional pumping energy is required to extract deeper water supplies.

3.4 Subsidy on Liquefied Petroleum Gas and Kerosene

Even after the dismantling of the APM, it was decided that the government would continue to subsidize kerosene and cooking gas. Consequently, in the 2002–2003 union government budget, the government has set aside 64.95 billion rupees as a subsidy on kerosene and cooking gas for fiscal year 2002–2003.

4. ROLE OF RENEWABLES IN FACILITATING TRANSITION AND DECENTRALIZATION FOR REMOTE AREAS

India is endowed with abundant natural and renewable resources of energy. The country has been able to achieve significant capacity addition of 1367 MW through wind farms and ranks fifth in the world. Overall, the renewable energy sector contributes over 3400 MW, supplementing and complementing the conventional energy sector in meeting the country's growing energy needs. Moreover, it also helps in meeting the global demand for limiting the emissions of greenhouse gases into the atmosphere. The status of the installation of RETs is indicated in Table II.

TABLE II

Status of Installation of Renewable Energy Technologies

Technology	Potential	Achievement	India's position (in the world)
Biogas plants (units)	12 million	3.3 million	2nd
Improved cookstoves (units)	120 million	34 million	2nd
Wind (MW)	45,000	1702	5th
Small hydro (MW)	15,000	1463	10th
Biomass power/cogeneration (MW)	19,500	468	4th
Biomass gasification (MW)		53	1st
Solar photovoltaic (MW/km^3)	20	107	3rd
Waste to energy (MW)	2,500	25	—
Solar water heating	140 million km^2 collector area	0.68 million km^2 collector area	—

Source. Reprinted from Ministry of Nonconventional Energy Sources (2003).

By their very nature, rural areas lend themselves amenably to the use of renewable energy. RETs generally provide decentralized and modular ways of meeting energy demands. Rural areas in India have an abundance of solar energy that can be exploited for both thermal and electrical applications. Biomass energy is conventionally referred to as traditional energy but, when used in modern technology applications, also qualifies as renewable energy. The dissemination of RETs in the country has been mainly through government programs that have primarily helped in introducing rural people to the technologies, but have not been successful in wide-scale dissemination and penetration. Given the nonviability of extending the power grid to remote areas, RETs can provide a means by which a decentralized supply of energy for such areas becomes feasible and close to being economically viable. In this regard, the plans of the concerned ministry are indicated in Table III.

Clearly, most of these initiatives are geared to meet rural energy requirements. In essence, India is on the edge of a new revolution in the management of rural energy problems and must, therefore, devise policies that move away from the mindsets of the past and build on the technologies for the future.

5. CONCLUSION

Since a major part of the planning process pertaining to rural energy has been left to consumers themselves, the cost of the fuel, economic status of the community, and easy availability are the determinants for the type of fuel and technology used for meeting energy needs. One of the main reasons for the absence of a modern energy supply in rural India is the low purchasing power of the people. Given the fact that a vast majority of rural people have limited cash, they find it difficult to afford the initial costs of acquiring more efficient and cleaner fuels and technologies. Consequently, the rural poor are drawn into a cyclic pattern of energy consumption that adversely affects the resource base as well as the individuals.

With the emphasis on electricity reforms, accelerated rural electrification, and increasing the share of renewables in the overall energy consumption, the scenario is changing. The country is on the threshold of a major change in the energy supply-and-demand scenario and the right policies and initiatives will facilitate the realization of the objective of energy for all.

TABLE III

Renewable Energy Plan 2012 of the Ministry of Nonconventional Energy Sources, Government of India

- Coverage of 30 million households through improved cookstoves
- Setting up of an additional 3 million biogas plants
- Deployment of 5 million solar lanterns and 2 million solar home systems
- Deployment of solar water heaters in 1 million homes
- Addition of 10,000 MW of renewables-based power
- Electrification of 18,000 remote villages

Source. Reprinted from The Energy and Resources Institute (2002).

SEE ALSO THE FOLLOWING ARTICLES

Biomass Gasification • Biomass: Impact on Carbon Cycle and Greenhouse Gas Emissions • Development and Energy, Overview • Indoor Air Quality in Developing Nations • National Energy Policy: India • Rural Energy in China • Sustainable Development: Basic Concepts and Application to Energy • Women and Energy: Issues in Developing Nations • Wood in Household Energy Use

Further Reading

Commissioner, Registrar General of the Census. (1991). "Census of India, 1991, Table: Household Amenities and Assets." Registrar General of the Census, Commissioner, New Delhi, India.

Ministry of Petroleum and Natural Gas. (2000). "Annual Report 2000–2001." Ministry of Petroleum and Natural Gas, Government of India, New Delhi, India.

National Sample Survey Organization 55th Round. (2001). "Energy Used by Indian Households 1999–2000," Report 464(55/1.0/6). National Sample Survey Organization, Ministry of Statistics and Programme Implementation, Government of India, New Delhi, India.

Natrajan, I. (1985). "Domestic Fuel Survey with Special Reference to Kerosene," Vol. 1. National Council for Applied Economic Research, New Delhi, India.

Padmanaban, S., and Sarkar, A. (2001). "Electricity Demand Side Management (DSM) in India: A Strategic and Policy Perspective." Office of Energy, Environmental, and Enterprise, USAID, India. Presented at the International Conference on Distribution Reforms, October 12, 2001.

Planning Commission. (2001). "Annual Report on the Working of the State Electricity Boards and Electricity Departments: Power and Energy Division." Planning Commission, Government of India, New Delhi, India.

Planning Commission. (2002). "Tenth Five-Year Plan: 2002–2007, Vol. II, Thematic Issues and Sectoral Programmes." Planning Commission, Government of India, New Delhi, India.

Ravindranath, N. H., and Hall, D. O. (1995). "Biomass, Energy and Environment: A Developing Country Perspective from India." Oxford University Press, New York.

The Energy and Resources Institute (TERI). (1995). "Environmental Considerations and Options in Managing India's Long Term Energy Strategy (Eco-miles)." TERI, New Delhi, India.

TERI. (2000). "Rural Energy Matters: The Dhanwas Experience," p. 184. TERI, New Delhi, India.

TERI. (2001). "Directions, Innovations, and Strategies for Harnessing Action (DISHA)" (R. K. Pachauri and R. K. Batra, Eds.), p. 310. TERI, New Delhi, India.

TERI. (2002). "Defining an Integrated Energy Strategy for India: A Document for Discussion, Debate, and Action." TERI, New Delhi, India.

Service and Commerce Sector, Energy Use in

BERND GEIGER and PETER TZSCHEUTSCHLER
Munich University of Technology
Munich, Germany

1. **Commerce/Services: Element of a Final Energy Balance**
2. **Demarcation of Residential and Commerce/Services**
3. **Commerce/Services/Handcraft/Agriculture**
4. **Structure Analysis of the Sector Commerce/Services in Germany**
5. **Summary**

Glossary

commerce/services sector Contains commerce, handcraft, construction, public and private services, trade, agriculture, and others.
final energy Supplied to the consumer to be converted into useful energy (e.g., electricity at the wall outlet).
residential sector Comprises private households.

The world has to deal with an increase of the global energy consumption. This affects each of the classical consumer sectors: industry, transportation, residential, and commerce/services. To a large extent, the rising energy demand will be met by extending the use of conventional energy carriers such as oil, natural gas, and coal. This is a global trend, even though the characteristics of this development may vary from continent to continent and from country to country. To detect the reasons for energy use, inventories of energy consumption and analyses of energy application were always practical. Across-the-board comparisons of continents or countries are of little help. Taking into account parametrically effective factors on the energy consumption brings some transparency and understanding of what would otherwise be rated and commentated falsely. This applies in particular to the commerce/services sector, which had occupied a shadow existence in the national energy balances for a long period. Fortunately, the provision of data to this sector has improved substantially, and many gaps can be closed. Creating a first country-wise data framework for energy consumption, which can be consulted for comprehensive country-related analyses, is a large merit of international statistics. However, this may not lead to a blind belief in the reliability of energy consumption data. Critical notes are therefore always helpful and appropriate. This article deals with the energy consumption structure of the commerce/services sector, indicates parameters of influence and supplies results and perceptions that apply worldwide, for chosen countries, and in Germany.

1. COMMERCE/SERVICES: ELEMENT OF A FINAL ENERGY BALANCE

This article examines the significance and importance of the commerce/services sector by consulting the energy consumption data for 1995, as published in 1998 by the World Energy Council. Based on the six continents—North America, Central/South America, Europe, Africa, Asia, and Oceania—continental energy balances were set up as shown in Fig. 1. It appears that the proportion of the industry is situated between 18% and 40%, the transportation sector accounts for 22% to 45%, private households account for between 13% and 58%, and the proportion of the commerce/services sector varies from 3% to 12%.

By differentiating between continents, it is possible to mark out climatic influences as well as population density and economic prosperity, as shown in Figs. 2 and 3. As a consequence, a large bandwidth of the specific final energy consumption can be noted. In terms of final energy demand, North America consumes 155 GJ per capita and year, while Africa consumes 16 GJ per capita and year only.

 © 2004 Elsevier Inc. All rights reserved.

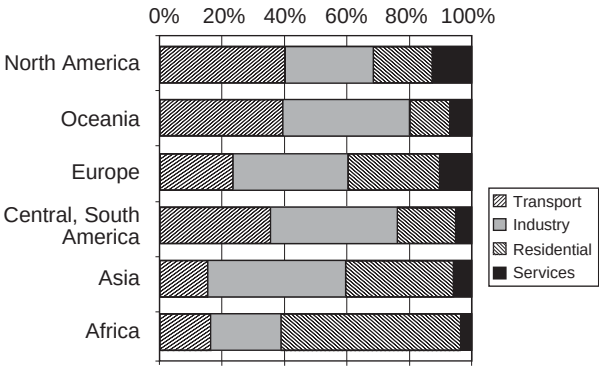


FIGURE 1 Contributions of the different sectors to the final energy demand in 1995; 100% represents the total demand of final energy.

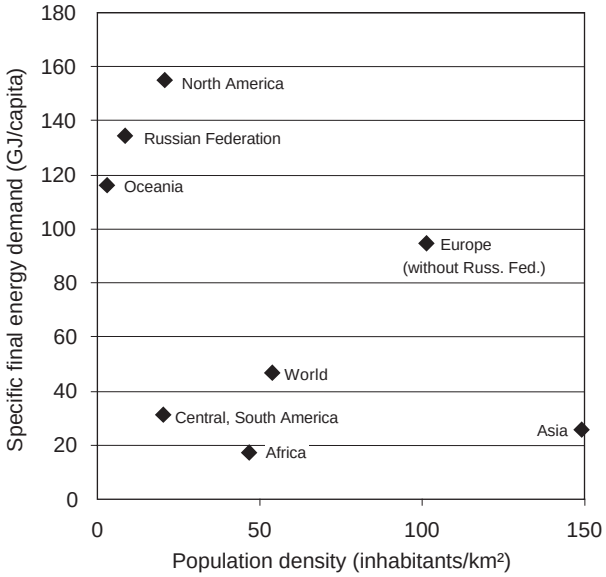


FIGURE 2 Specific final energy demand in relation to the continental population density in 1995.

It can be postulated that an increasing population density reduces the expenditure for transportation and thus its proportion of the final energy consumption. Shorter distances of transport and communication as well as shorter commuter distances result in lower consumption values in the traffic sector (Asia, Europe, Africa). On the other hand, a low population density causes substantial consumption impulses for the traffic sector. This does not remain without effect on the energy consumption in the sector commerce/services.

Each society requires products and services, whereby substantial levelings are to be assumed—not only depending on the prosperity and the gross domestic product, but also on the capability for exports. A

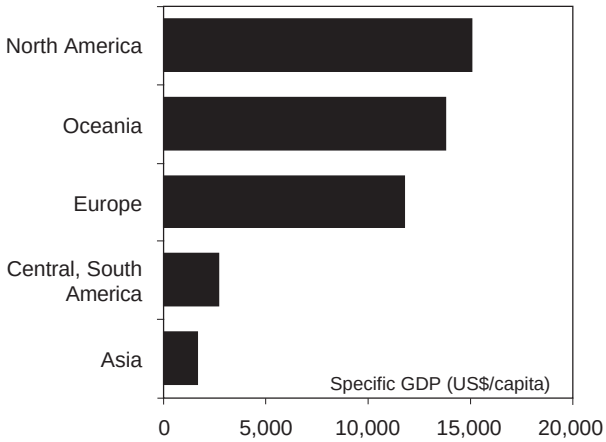


FIGURE 3 Specific GDP of the different continents in 1995.

sharp demarcation of industry and handcraft, respectively, of industrial structures and small trade activities regarding the associated energy consumption causes some problems. Thus, the proportions for industry and commerce/services in Fig 1 should be considered as a reference value instead of as a definite mirror image of accurate allocation.

The common life of each society, whether it is shaped industrially, agriculturally, or mixed, requires services and the supply of everyday products such as nourishments. In the course of social restructuring with increasing prosperity, the tendency toward division of labor is enforced. This means that while the national income is still low, handcraft, small trade, and agriculture are, to a large extent, connected to the residential sector. With increasing prosperity, a separate services sector emerges and organizes. These circumstances are also reflected in Fig 1: while in industrialized continents with above-average per capita income (North America, Oceania, Europe) the activities of commerce and public services are "appropriate," verifiable, and delimitable to other consumption sectors, this is not always possible in Asia and Africa, where these activities are partly contained in the residential sector.

We are still a long way from an accurate image and sharp delimitations, even though large progress in the collection of international consumption structure data has been made.

2. DEMARCATION OF RESIDENTIAL AND COMMERCE/SERVICES

The boundaries between the residential and commerce/services or handcraft and agriculture are often

diffuse as far as the associated energy consumption is concerned. This becomes relevant whenever the same building is used for residential and occupational purposes, when a household cannot be separated from an agricultural supplementary income, or if members of a family are temporarily bound in a business.

Statistically proven consumption delimitations have to be considered under these aspects too: what is separately accounted for as commerce/services in one country can be accounted for under the category of residential in another country. Therefore, careful handling of sector-specific consumption data is inevitable if these are to be analyzed or interpreted or if a landspreading transverse comparison is to be carried out, as is demonstrated next.

Figures 4 and 5 show the per capita consumption of electricity and fuels for households and commerce/services, again aggregated by continents. It is remarkable that the electricity consumption of commerce/services is generally higher than the electricity consumption of the residential sector; for residential it ranges from 0.3 GJ/cap in Africa to 10 GJ/cap in North America. The fuel consumption for residential is significantly higher than it is for commerce/services; the values of the latter spread between 0.6 GJ/cap in Africa and 12 GJ/cap in North America.

For a better understanding of these specific consumption values, one should know that from 20 to 50% of the population is involved in the sector "commerce/services/handcraft/agriculture" because of employment, occupation, or simply for covering vital needs.

This note seems necessary, if energy consumption is expressed as giga joule per capita and is not allocated to its origin (e.g., earner or employee of a consumption sector). If earner or employee is chosen as base factor, the specific consumption is between two and five times higher according to the relation of inhabitants per earner/employee.

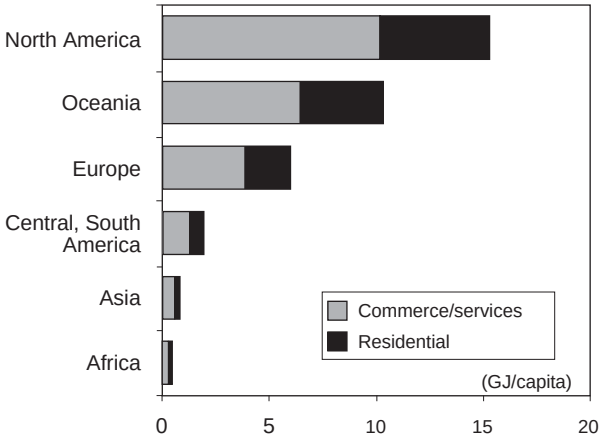


FIGURE 4 Specific electricity consumption of the commerce/services and residential sectors in 1995.

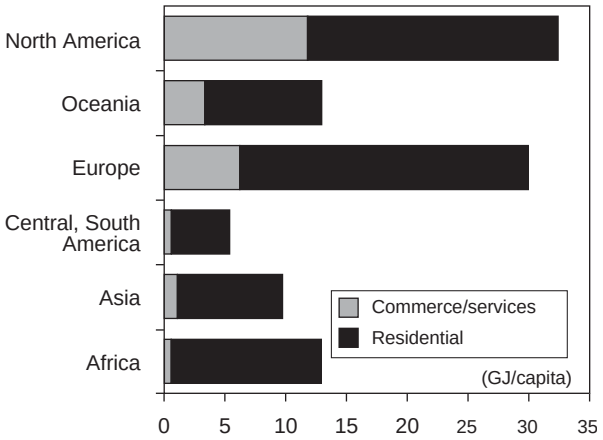


FIGURE 5 Specific fuel consumption of the commerce/services and residential sectors in 1995.

3. COMMERCE/SERVICES/ HANDCRAFT/AGRICULTURE

Figure 6 shows the final energy demand per capita for the sector commerce/services (handcraft and agriculture might also be avcontained). It is differentiated by continents and therefore also by climates and by consumption of fuel and electricity.

For the fuel consumption, mean values between 10 GJ/cap and 0.15 GJ/cap occur in the following order: North America, 10 GJ/cap; Europe, 5.6 GJ/cap; Oceania, 2.6 GJ/cap; Asia, 0.9 GJ/cap; Africa,

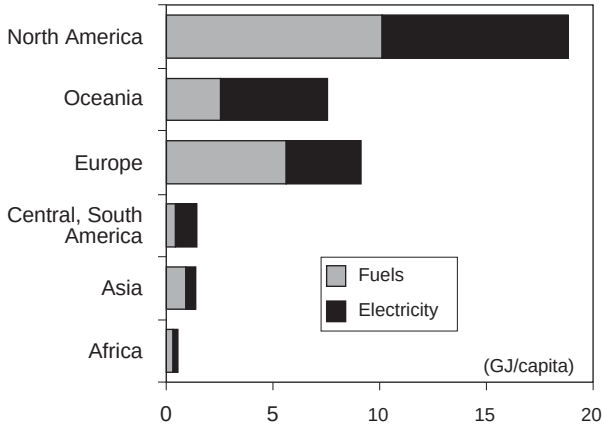


FIGURE 6 Specific final energy demand of the commerce/services sector in 1995.

0.25 GJ/cap; and Central, South America, 0.15 GJ/cap. The energy consumption values are mainly due to heating, conditioning of the work space, and the supply of warm water. The required area per employee at his or her workplace must be considered.

Based on temperature conditions as they occur in North America and Europe, an estimated area requirement of 40 m^2 per wage earner and a heating requirement of 200 kWh per m^2 and year and a specific consumption of 8000 kWh or 28.8 GJ per wage earner and year can be calculated. Rated by 0.3 wage earners per capita, an energy consumption for space heating in the sector commerce/services of 8.6 GJ per capita and year can be assumed.

This is a standard consumption level for heating purposes in many prospering industrial societies.

Another standard consumption level is represented by the need for heated water (for washing, bathing, showering), which accounts for 300 kWh per wage earner and year, respectively, 1.08 GJ per wage earner and year. This corresponds to 0.3 GJ per capita and year. Thus, with only a few characteristic values, the specific final energy demand of 10 GJ per capita and year can be explained and analyzed to a large extent.

Finally, area requirements of 40 m^2 per wage earner correspond to empirical average values, which may vary significantly in individual cases. Employment structures with areas per employee of 100 m^2 are also not unusual, if attractiveness within the service area and customer demands require this.

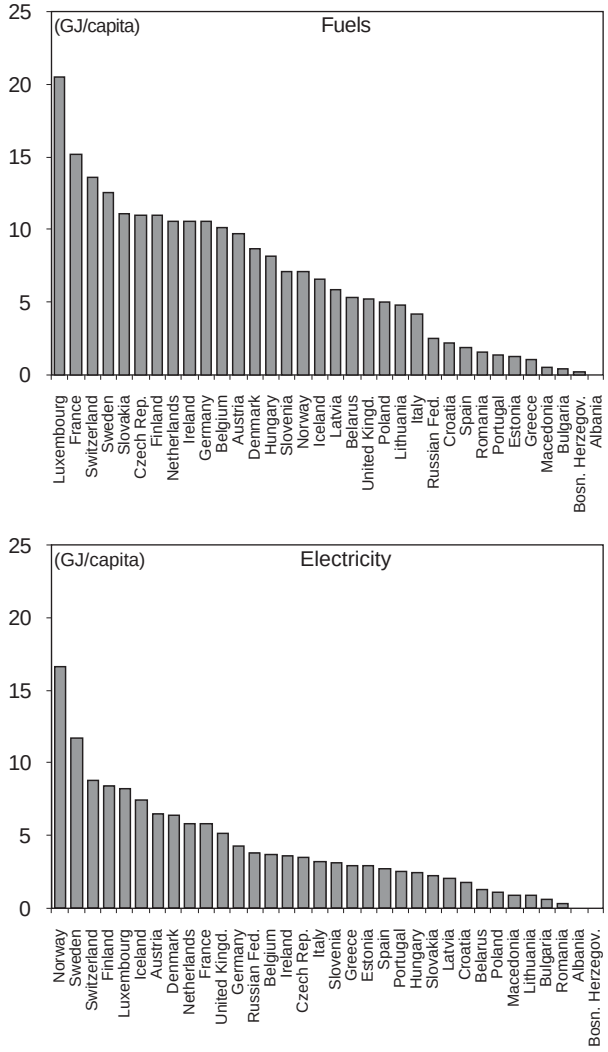


FIGURE 7 Final energy demand of the commerce/services sector in Europe in 1995.

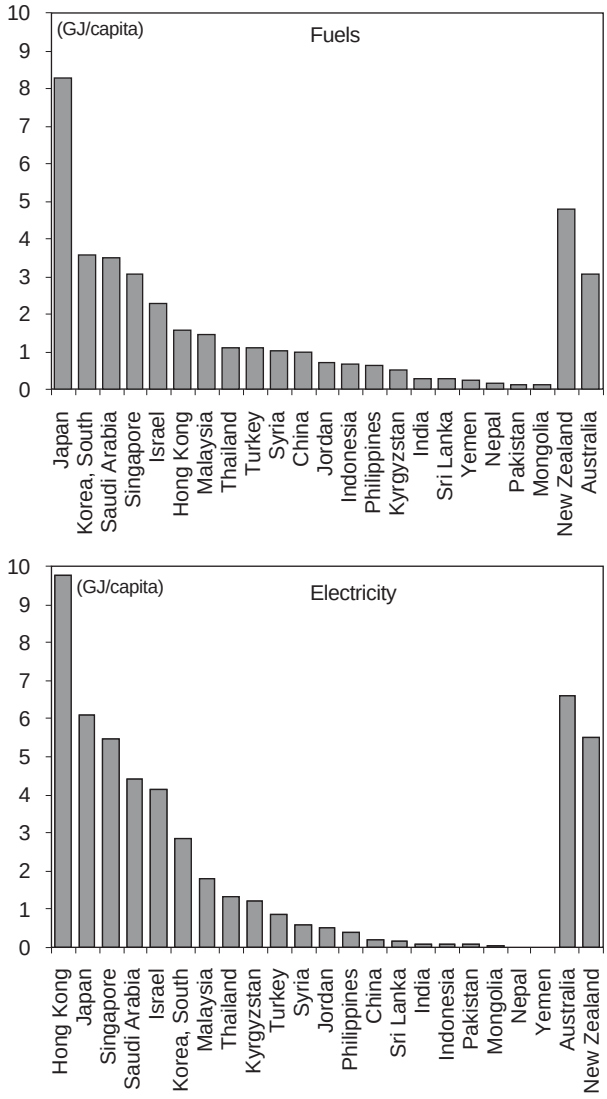


FIGURE 8 Final energy demand of the commerce/services sector in the United States in 1995.

Energy consumption values for room conditioning (heating, ventilation, air conditioning) of 200 kWh/$m^2 * a$ are also average values and result in a large bandwidth. On the one hand, thermal protection and heat recovery systems supply appropriate solutions for the reduction of consumption without a loss of comfort. On the other hand, objects with a specific consumption of more than 1000 kWh/$m^2 * a$—no rarity, by the way—are an unnecessary waste of energy.

More details can be derived from Figs. 7 to 10. The European countries indicate a large bandwidth in the specific electricity and fuel consumption, which is to due to climatic influences and different levels of prosperity (Fig. 7). This also applies to the countries in North and South America. Cool Canada and warm South America, rich North America and poor South America, compose basic conditions that influence the specific requirements of electricity and fuel (Fig. 8). The countries of Asia and Oceania are generally situated in warm and hot climates, which is reflected in low specific fuel consumption. While this temperature dependency applies to most of these countries, Japan and New Zealand show an unexpected behavior. In all remaining countries, this factor is temperature dependent as far as possible. The specific electricity consumption represents the prosperity of Hong Kong, Japan, Australia, and New Zealand (Fig. 9). Warm to hot climate zones in Africa lead inevitably to small specific fuel consumption. The economic position of Africa (with a small

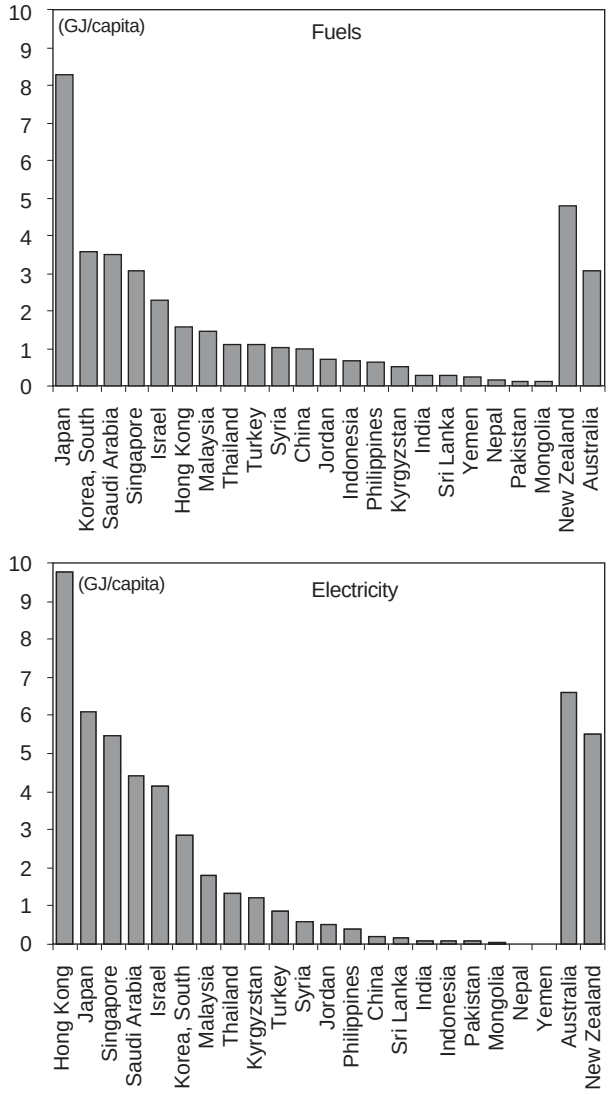


FIGURE 9 Final energy demand of the commerce/services sector in Asia and Oceania in 1995.

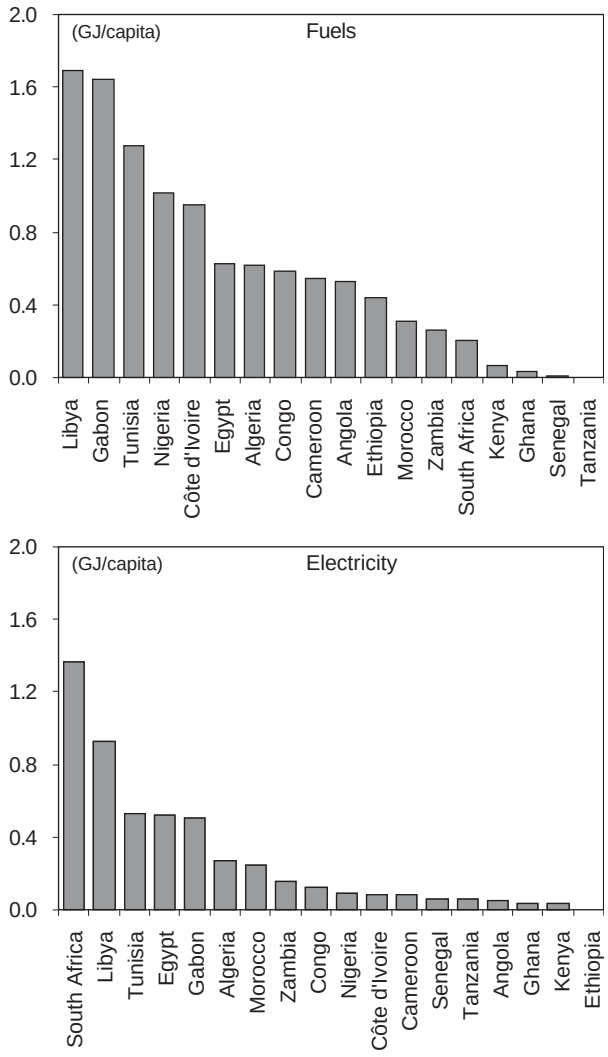


FIGURE 10 Final energy demand of the commerce/services sector in Africa in 1995.

TABLE I

Sociodemographic and Energetic Data of Chosen Countries in 1995

Country	Inhabitants [Mio.]	Average temperature [°C]	Specific GDP [U.S. $/capita]	Commerce/services	
				Fuels [GJ/capita]	Electricity [GJ/capita]
Africa					
Algeria	28.1	18.5	2170	0.62	0.26
Angola	11	24	332	0.53	0.05
Congo (Brazzaville)	2.6	25	1076	0.59	0.12
Cameroon	13.3	26	768	0.54	0.08
Egypt	57.5	21.5	884	0.62	0.52
Ethiopia	54.7	16.5	140	0.44	0.00
Gabon	1.1	25.5	6251	1.65	0.5
Ghana	17.7	27	436	0.03	0.03
Côte dl'voire	14.2	26.5	833	0.95	0.08
Kenya	30.5	20.5	303	0.06	0.03
Libya	5.4	20	5282	1.69	0.92
Morocco	26.4	19	1037	0.31	0.24
Nigeria	99	27	329	1.02	0.09
South Africa	39.5	18.5	2814	0.2	1.36
Senegal	8.6	25.5	711	0.01	0.05
Tunisia	9	18	1694	1.28	0.52
Tanzania	28.3	25	144	-	0.05
Zambia	9.1	21	332	0.26	0.15
North and South America					
Canada	29.4	6	12061	23.18	14.38
Jamaica	2.5	27	1799	0.47	1.58
Mexico	90.5	18.5	3132	0.88	0.59
Trinidad and Tobago	1.3	26	4245	0.9	1.02
United States	263	15	23368	14.42	13.06
Argentina	34.8	16.5	5217	0.83	1.15
Bolivia	7.4	13	733	0.27	0.21
Brazil	155.8	22.5	3215	0.35	1.28
Chile	14.2	14.5	3060	0.77	-
Paraguay	4.8	24	1283	0.6	0.53
Peru	23.5	16.5	1885	0.55	0.39
Uruguay	3.2	16	3102	0.61	1.53
Venezuela	21.8	24.5	2611	1.44	2.53
Asia					
Sri Lanka	18.1	27.5	551	0.26	0.16
China	1220.5	13	519	0.97	0.18
China, Hong Kong	6.2	15	15881	1.56	9.77
Indonesia	194.8	27.5	913	0.68	0.1
India	922	27.5	415	0.29	0.1
Israel	5.5	20	12634	2.26	4.15
Japan	125.5	15	25696	8.27	6.1
Jordan	5.7	17.5	1007	0.72	0.5
Kyrgyzstan	4.6	4	311	0.52	1.18
Korea, South	45.1	11.5	8069	3.58	2.87
Mongolia	2.4	-2.5	-	0.11	0.05

continues

Table I continued

Country	Inhabitants [Mio.]	Average temperature [°C]	Specific GDP [U.S. $/capita]	Commerce/ services	
				Fuels [GJ/capita]	Electricity [GJ/capita]
Malaysia	20.7	27.5	3149	1.47	1.81
Nepal	20.3	18.5	229	0.16	0.01
Pakistan	130.3	25	348	0.13	0.07
Philippines	70.3	27.5	706	0.61	0.38
Saudi Arabia	18.3	27.5	6412	3.51	4.4
Singapore	3.5	27.5	16328	3.07	5.45
Syria	14.2	18.5	1235	1.04	0.58
Thailand	59.4	27.5	1681	1.1	1.34
Turkey	60.6	15	2912	1.09	0.86
Yemen	15.4	29.5	511	0.22	-
Oceania					
Australia	18.1	18	18705	3.07	6.6
New Zealand	3.7	12	13610	4.77	5.49
Europe					
Albania	3.6	16.5	510	-	0.03
Austria	8.1	9.5	21846	9.71	6.44
Belgium	10.1	10	20310	10.15	3.65
Bosnia and Herzegov	4.2	10	309	0.23	-
Belarus	10.3	5	2139	5.32	1.24
Bulgaria	8.4	10.5	2016	0.42	0.62
Denmark	5.2	8.5	31281	8.73	6.31
Ireland	3.6	10	16815	10.57	3.6
Estonia	1.5	5	3033	1.29	2.88
Czech Republic	10.3	8	6531	11	3.5
Finland	5.1	4.5	24646	10.97	8.38
France	58.1	11.5	21678	15.21	5.75
Germany	81.7	9	25753	10.54	4.21
Greece	10.5	17.5	8586	1.04	2.88
Croatia	4.7	15	3846	2.15	1.72
Hungary	10.2	11	2989	8.21	2.44
Iceland	0.3	5	24148	6.56	7.37
Italy	57.3	15	20187	4.21	3.18
Latvia	2.5	6	2570	5.84	2.06
Lithuania	3.7	5	2414	4.84	0.88
Slovakia	5.4	10	2102	11.07	2.25
Luxembourg	0.4	9.5	32898	20.52	8.2
Macedonia	2	15	752	0.53	0.89
Netherlands	15.5	9.5	20364	10.57	5.8
Norway	4.4	5.5	31648	7.11	16.51
Poland	38.6	7.5	1705	4.99	1.09
Portugal	9.9	16	7294	1.31	2.51
Romania	22.7	10.5	1473	1.54	0.27
Russian Federation	148.1	3.5	2428	2.47	3.76
Slovenia	2	10	8492	7.16	3.11
Spain	39.2	15.5	13409	1.91	2.73
Sweden	8.8	6	26544	12.57	11.66
Switzerland	7	10	32386	13.64	8.74
United Kingdom	58.6	10	17640	5.2	5.05

specific national income) is again reflected in the low specific electricity consumption (Fig. 10).

Is it possible to deduce overall relations from the facts, like the influence of climate zones and prosperity on the electricity and fuel consumption? To explore this question, we gathered 93 countries in Table I and used these data as a basis for further analyses. Table I contains data for electricity and fuel consumption, GDP (each referred to the number of inhabitants), as well as climatical information.

Generally the application of electrical energy is favored if a high proportion of electricity is generated by hydropower. Therefore, the following countries are excluded from further analyses: Norway, Sweden, Finland, Iceland, Canada, Switzerland, and France.

In Fig. 11 depicts the relationships between the annual average temperature and the specific electricity and fuel consumption. The higher the annual average temperature, the smaller the influence of fuel consumption. The lower the annual average

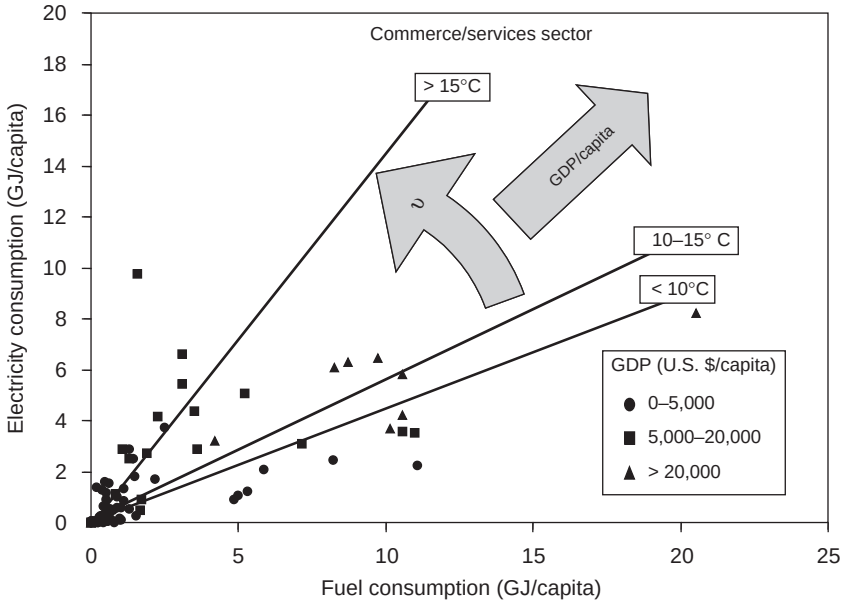


FIGURE 11 Energy consumption in relation to climate and prosperity in 1995.

TABLE II

Consumption Data for Space Heating in Commerce/Services

Country	Floor area in commerce, services per capita m^2/capita	GDP commerce/services US $/capita	Service sector fuel/heat 1995 GJ/(m^2*a)	Mean temperature °C
Italy	6.5	10,300	0.72	15.0
Japan	10.5	11,000	0.70	15.0
Norway	16.0	11,300	0.58	5.5
Australia	17.5	11,600	0.20	18.0
Denmark	17.5	10,000	0.59	8.5
USA	25.0	16,400	0.60	15.0
Sweden	15.5	10,500	0.81	6.0
Canada	16.0	10,500	0.99	6.0
New Zealand	14.0	9,300	0.35	12.0
Germany	17.8[a]		0.50	9.0
France			0.20	11.5
Finland			1.05	4.5

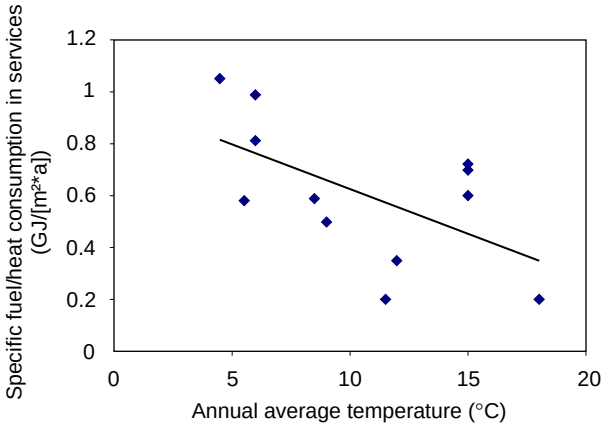


FIGURE 12 Specific fuel and heat consumption of commerce/services in relation to the annual average temperature in 1995.

temperature of a country, the higher the fuel consumption per capita, since space heating gains more and more importance. This behavior sets in from an annual average temperature of 12.5°C. With annual average temperatures between 10 and 0°C, the fuel consumption caused by space heating increases distinctly. Rising prosperity in general leads to a higher specific consumption of electricity and fuel due to more and better electric equipment and heating of larger and well-conditioned rooms.

In Table II presents interesting data for a cross-national comparison (as completed in 2000 by Geiger and Harrison). The average available floor area per capita in the commerce/services sector, GDP, the annual average temperature, and the resulting energy demand for space heating per area and year or per capita and year are included. In Fig. 12 the coherence between cold climate and rising energy consumption becomes obvious too.

4. STRUCTURE ANALYSIS OF THE SECTOR COMMERCE/SERVICES IN GERMANY

Much can be learned from observing the commerce/services sector in Germany. Figure 13 shows the energy consumption broken down into energy carriers and applications such as process heat, space heating, mechanical energy, lighting, and communication for a period of 10 years. By collecting energy requirement statistics, investigating energy consumption, and analyzing changes from year to year, we have come to the conclusion that short-term changes are, in most cases, caused by climatical or cyclical economical reasons fluctuations. Tendentious changes

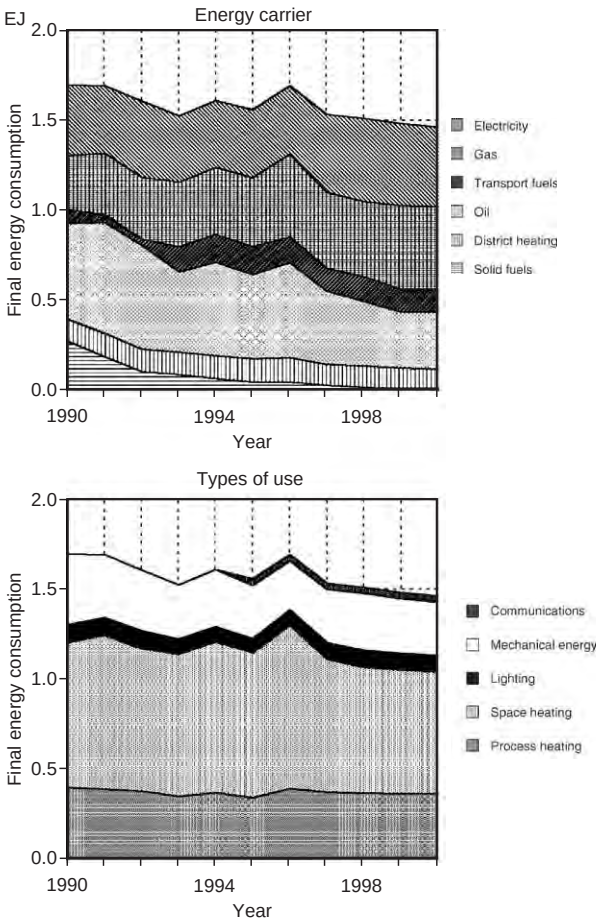


FIGURE 13 Final energy consumption of the commerce/services sector in Germany.

during larger periods result from changes in effective areas, technical equipment, and user requirements.

Beyond this, the question of who needs how much energy and for which purpose is gaining interest. "Who" refers to typical consumer groups from the commerce/services sector. By investigating how much, it is not only of interest to analyze the particular consumption proportions of different consumer groups, but also which consumption per person employed or per m^2 net area is typical. We select "persons employed" and "net area" as base factors, because these are relatively available and also are usually known. "For which purpose" specifies how the energy under investigation is being applied, such as for lighting, electrical drives, communication, space heating, or process heat.

On the basis of inspections, a metering campaign, and a survey of about 2800 companies and facilities during the 1990s, detailed information on the energy consumption and its key indicators could be sampled,

detected, and measured. Twenty-three consumer groups could be distinguished, and data for an additional nine consumer groups could be retrieved from secondary statistical material, with sufficient information quality. Within the scope of the survey, information on electricity and fuel consumption for heating and transportation could be collected. On the basis of equipment size, process load, and user profiles, the electricity and fuel consumption could be analyzed and differentiated according to the purpose of use.

Table III shows the calculated overall consumption by small consumers in 1994; additional

TABLE III

Final Energy Consumption: Consumer Groups of Commerce/Services and other Sector, Germany, 1994

Unit	Employees [Mio.]	Electricity [TWh/a]	Fuels and district heating [TWh/a]	Transp. fuels in the sector: Commerce/ services [TWh/a]	Transp. fuels in the sector: Transport [TWh/a]	Operating area [Mio. m^2]
Metal Working Industry[a]	0.96	2.33	8.50		10.92	53
Automobile trade and commerce	0.14	0.46	1.96		0.99	18
Timber industry[a]	0.21	0.68	4.25		2.36	22
Printing and paper industry[a]	0.11	0.43	1.35		0.68	6
Bakeries	0.19	1.18	5.99		1.28	7
Butcheries	0.15	1.03	2.13		0.87	6
Other food trade and commerce	0.03	0.09	0.39		0.37	2
Laundries	0.09	0.60	3.37		0.50	3
Horticulture	0.27	0.60	6.90		1.02	35
Retail trade	3.07	12.99	35.50		14.54	205
Wholesale trade	1.57	7.04	16.25		18.57	184
Credit institutes and insurances	1.04	2.80	7.22		4.27	36
Accomodations	0.35	3.25	10.91		1.02	51
Catering	0.86	5.01	11.98		4.32	45
Publishing companies	0.18	0.48	0.70		1.07	6
Other services	4.01	5.44	23.63		31.99	106
Nonprofitable organizations	1.68	4.27	24.48		2.88	152
Public corporations and social security	4.16	5.92	32.07		2.94	153
Hospitals	1.06	4.64	15.36		0.13	47
Schools	1.27	4.94	27.88		0.05	130
Baths	0.07	1.70	5.56		0.00	6
Clothing trade and commerce	0.33	1.97	10.20			
Trading agencies	0.25	1.12	2.59			29
Postal services	0.35	0.57	1.19			6
Telecommunications	0.23	2.98	0.63			
Postbank AG	0.02	0.05	0.14			1
Railway (Deutsche Bahn AG)	0.22	1.78	7.98			
Inland navigation	0.04	0.00	0.00			
Airports	0.06	0.61	0.60		0.02	
Shipping and store house	0.15	0.07	0.15			
Street lighting	0.00	3.33	0.00			
Milit.departments	-	0.81	3.25	11.39		
Main construction trade	3.00	2.47	16.87		32.45	136
Agriculture	0.79	4.90	20.30	25.20		624
Total	26.89	86.55	310.27	36.59	133.22	2.070

[a] Less than 20 employees.

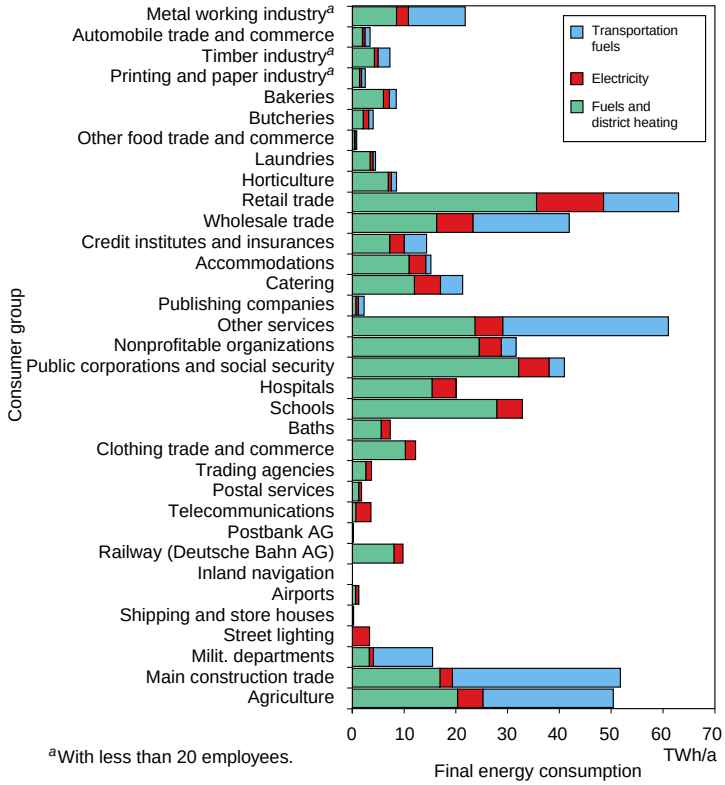


FIGURE 14 Final energy consumption: consumer groups of commerce/services and other sectors; Germany, 1994.

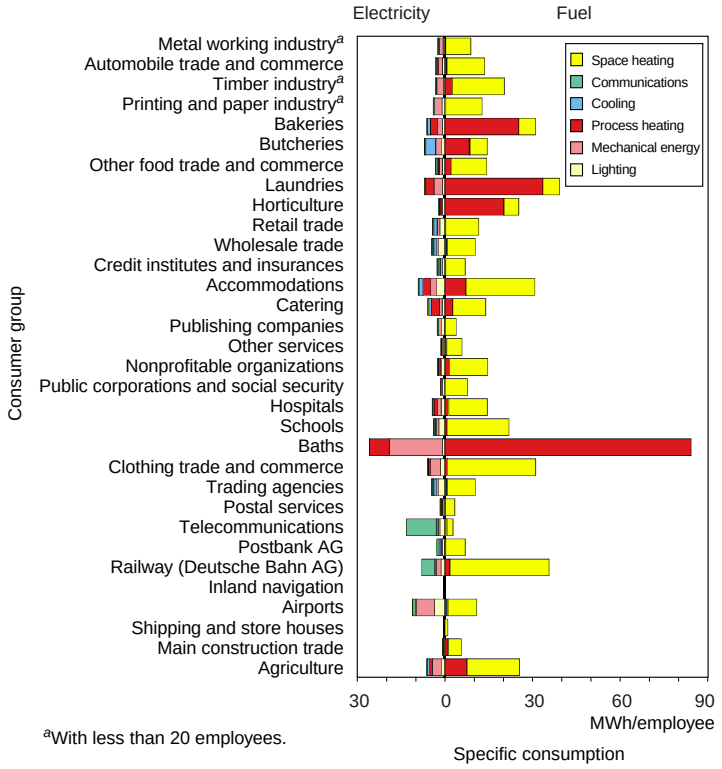


FIGURE 15 Electricity and fuel consumption related to employees; Germany, 1994.

information on the number of employees for each examined branch and the operating area is provided. This background knowledge suggests what sort of energy is consumed and where it is used.

The results for fuels (heating and transportation) and electricity are presented in Fig. 14. Table III and Fig. 14 contain all of the consumer groups, which were classified as small consumer in former times and which are now to a large extent accounted for in the commerce/services sector. These consumer groups also cover agriculture and main construction trade, which are often separated in the international consumer statistics or declared as industry.

The specifications concerning final energy consumption in Fig. 14 (electricity, fuels, space heating, and transportation fuels in the sector commerce/services) represent components of the German energy balance for the commerce/services sector. The given consumption of transportation fuels appears within the transportation sector in the German energy balance.

Figure 14 presents an overview of where the biggest proportions of energy onsumption (without transportation fuels) in Germany are situated. The most significant branches are agriculture, retail trade, public corporations and social security, schools, nonprofit organizations, and other services. These alone are responsible for about 50% of the final energy consumption in the commerce/services sector.

Figure 15 presents information on the specific electricity and fuel consumption, differentiated by purpose of use and based on the number of employees. It is obvious that the specific consumption within in the different consumer groups is very heterogeneous.

The main proportions constitute energy consumption due to baths, laundries, accommodations, bakeries, the clothing trade, and commerce. With regard

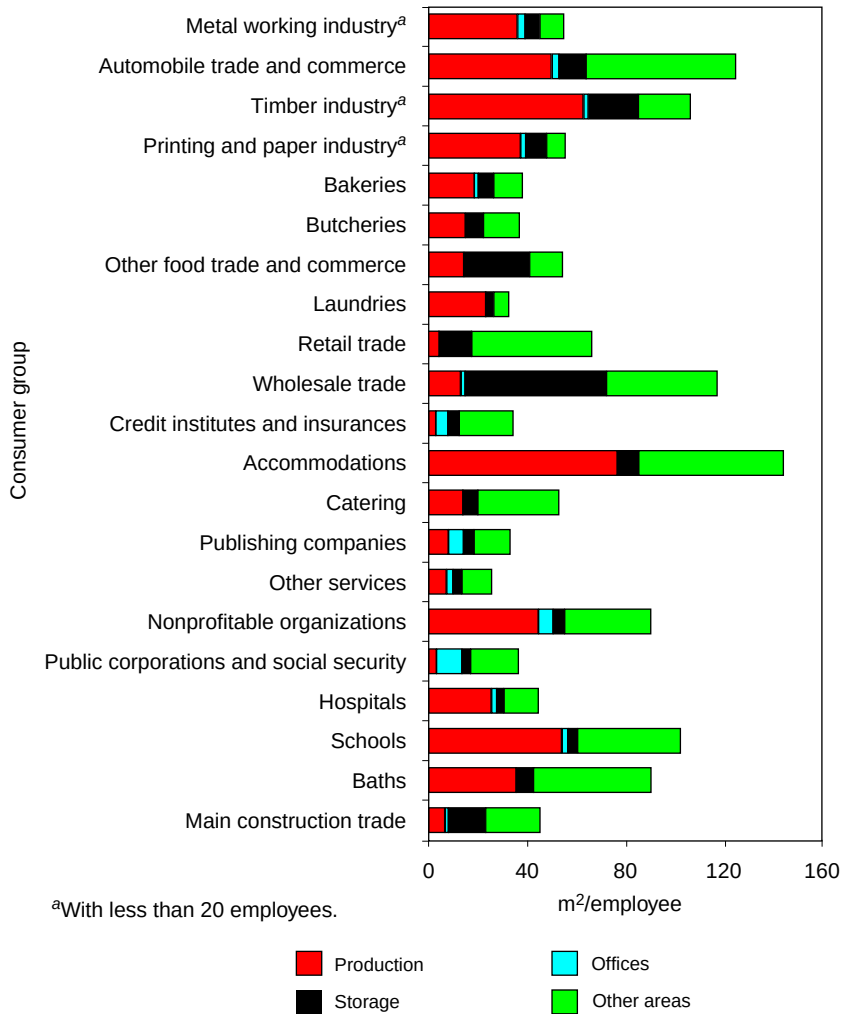


FIGURE 16 Operating area and types of use in commerce/services; Germany, 1994.

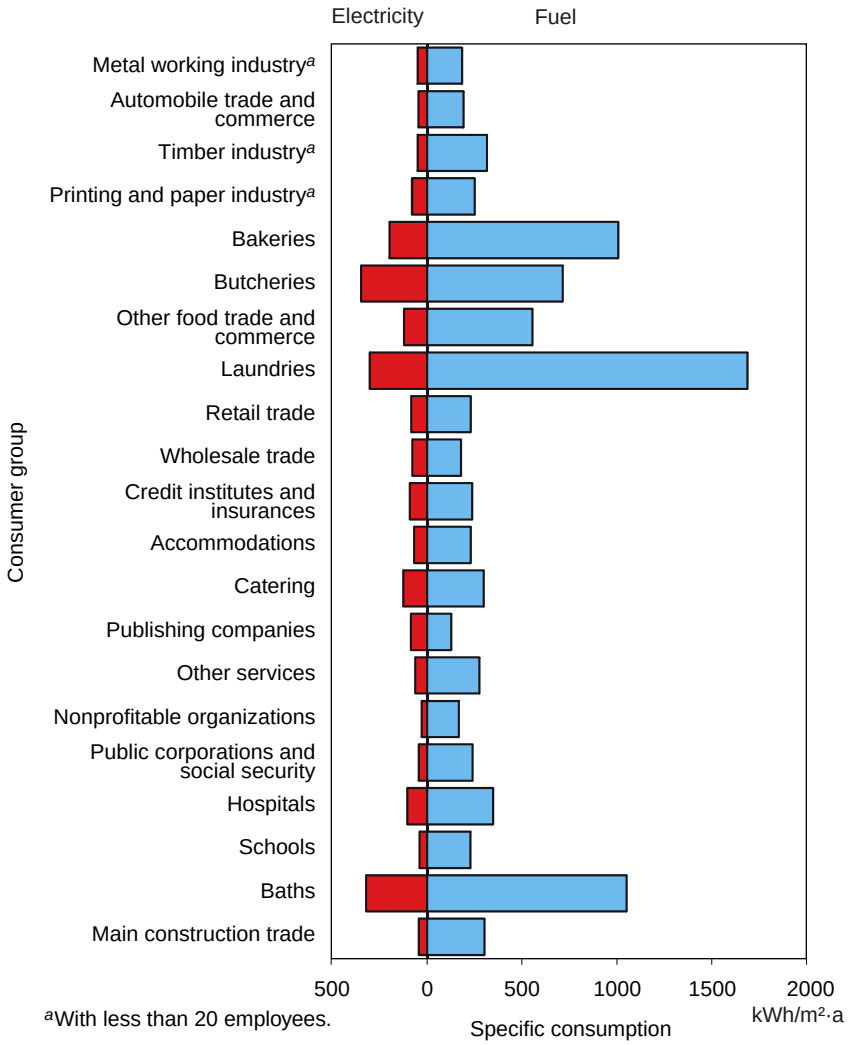


FIGURE 17 Electricity and fuel consumption related to operating area; Germany, 1994.

to the purpose of energy consumption, a differentiation is carried out by space heating, communication, cooling, process heat, mechanical energy, and lighting. Thus, the average proportions of energy consumption for certain applications are known for each consumer group. Note that the operationally effective area has a major influence on the space heating energy consumption. Production and office areas are always heated; warehouse and remaining areas are partly heated or kept at a moderate temperature.

The operating area per employee ranges from 30 m^2/employee (other services) to 150 m^2/employee (accommodation). By applying this information, the electricity and fuel consumption can be quoted by square meter operating area (see Fig. 16).

Figure 17 suggests that many branches with a very low or negligible process heat demand have a similar level of electricity and fuel consumption. Seventy-five percent of the consumer groups show a specific consumption in the range of 150 kWh/m^2 to 250 kWh/m^2, suggesting that these branches have a demand for space heating and hot water only.

Bakeries, butcheries, other food trade, and commerce, laundries, hospitals, and baths have a disproportionately high demand for fuels that process heat. Beyond this, these consumer groups are characterized by an extremely high specific electricity demand.

5. SUMMARY

This article explores the energy consumption of the commerce/services sector worldwide and particularly

in Germany. When comparing cross-national data, it has to be taken into account, that an accurate demarcation of energy consumption sectors is not always given. By analyzing surveys and data collected onsite, a significant analysis of the electricity and fuel consumption of the commerce/services sector is possible, as the example shows for Germany.

Energy consumption in the commerce/services sector mainly depends on economic prosperity, climate, and the technological status of development. One major influence on energy consumption is the conditioning of workspaces. Lower temperatures and larger workspaces per jobholder—as in the northern industrialized countries—cause higher fuel amounts of consumption for space heating, just as air conditioning in hot climates causes a higher demand for electricity. With economical development, technical standards and instrumentation improve and cause rising electricity consumption.

The analyses presented here are preconditions for better individual consumption consultation, strategic programs toward energy conservation, and an emphasis on technical development.

SEE ALSO THE FOLLOWING ARTICLES

Commercial Sector and Energy Use • Consumption, Energy, and the Environment • Corporate Environmental Strategy • Energy Services Industry • Global Energy Use: Status and Trends • Industrial Energy Efficiency • Industrial Energy Use, Status and Trends • Industrial Symbiosis • International Comparisons of Energy End Use: Benefits and Risks

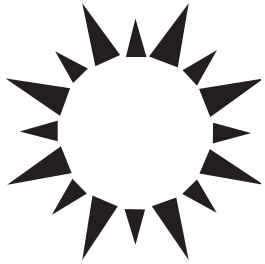
Further Reading

Geiger, B., Gruber, E., and Megele, W. (1999). Energieverbrauch und Energieeinsparung. *In* "Handel, Gewerbe, und Dienstleistung." Physica-Verlag.

Geiger, B., Harrison, A. (2000). "The State of Knowledge Leaders." IEA Workshop on Indicators of Commercial/Service Sector Activity, Energy and CO_2 Emissions, and Policy Impacts, Karlsruhe.

World Energy Council. (1998). International Energy Data 1998. World Energy Council, London.

Sociopolitical Collapse, Energy and



JOSEPH A. TAINTER
Rocky Mountain Research Station
Albuquerque, New Mexico

1. **Nature of Collapse**
2. **Energy Basis of Ancient Societies**
3. **Theories of Energy Flow and Collapse**
4. **Examples of Energy Flow in Sociopolitical Collapse**
5. **Energy Flow, Collapse, and Resiliency**
6. **Energy Flow in the Aftermath of Collapse**
7. **Synthesis and Implications**

Glossary

collapse Rapid loss of an established level of social, political, or economic complexity.
complexity A measure of the differentiation in structure and organization of a system.
marginal return In an economic process, the amount of extra output per additional unit of input. Marginal returns may increase, remain constant, or diminish.
problem solving The process of responding to the challenges that confront a society or institution.
resources The inputs of matter, energy, or information that support continuity and problem solving.
sustainability Maintaining, or fostering the development of, the systemic contexts that produce the goods, services, and amenities that people need or value, at an acceptable cost, for as long as they are needed or valued.

Collapse is the rapid simplification of a society. It is the sudden, pronounced loss of an established level of social, political, or economic complexity. Widely known examples include the collapses of Mesopotamia's Third Dynasty of Ur (ca. 2100–2000 BC), Mycenaean society of Greece (ca. 1650–1050 BC), the Western Roman Empire (last emperor deposed 476 AD), and Maya civilization of the lowlands of Guatemala (ca. 250–800 AD). There are at least two dozen cases of collapse that are known from history, archaeology, or both. States and empires are not the only types of institutions that may rapidly simplify. The entire spectrum of societies, from simple foragers to extensive empires, yields examples of collapse. Since all but a few human societies existed before the development of writing, there may be dozens or even hundreds of cases that are not yet recognized archaeologically. Collapse is therefore a recurrent process, and perhaps no society is invulnerable to it.

1. NATURE OF COLLAPSE

Many scholars have argued that collapse develops from adverse changes in the energy flows on which a society depends. In this school, collapse is thought to follow from such misfortunes as pronounced droughts, depletion of resources, or natural cataclysms. Others have searched for causes internal to societies: poor leadership, mismanagement, peasant rebellion, or even intangible factors such as changes in morality. A few scholars argue that collapses occur because of chance concatenations of misfortunes and cannot be understood as a general process. Despite this diversity of views, collapses inevitably alter energy flows. Postcollapse societies are by definition simpler, so much so that colloquially they are sometimes referred to as "dark ages." The simpler societies that follow a collapse require less energy per capita. Collapses often involve massive loss of population so that human energy extraction is greatly reduced within the region affected.

Collapse affects societies proportionate to their level of complexity. Many scholars identify collapses by the loss of such cultural elements as great traditions of art, architecture, and literature (with the decline in the last leading to the term dark age).

 © 2004 Elsevier Inc. All rights reserved.

Recognizing that collapse can affect societies organized with any degree of complexity, including simple foragers, a conception of collapse based on great cultural traditions is clearly too restrictive. It would not apply, for example, to the collapse of Chacoan society of the U.S. Southwest (ca. 725–1150 AD). Although it is known for remarkable art and architecture, Chacoan society produced no literature.

A more encompassing approach focuses on complexity. Complexity refers to differentiation in structure and to degree of organization. Societies grow more complex as they develop more parts (institutions, levels of hierarchy, and social or occupational roles) and as they become increasingly organized (i.e., as they increasingly constrain the behavior of a society's parts). One of history's clearest trends is that human societies during the past 12,000 years have shown a seemingly inexorable tendency to increase in complexity. This process still continues, more rapidly than ever.

Collapse is the rapid, substantial loss of an established degree of social, political, or economic complexity. Collapses contravene the trend of social history toward greater complexity. In a society that has collapsed there are fewer institutions, occupations, social roles, and hierarchical levels. There is less flow of information and a lower capacity for organization. People know less about the world beyond their own community. The scale of society shrinks. The Roman Empire, for example, left in its aftermath a multitude of smaller polities, none of which could match Rome's capabilities in organization, engineering, energy extraction, or scale of information flow.

2. ENERGY BASIS OF ANCIENT SOCIETIES

Before the development of fossil fuels, societies were powered largely by solar energy. Through photosynthesis, solar energy produced fuels such as wood and the agricultural produce on which all else depended. Solar energy drove the weather patterns to which forests and farms were adapted. These weather patterns made sailing possible and thus overseas commerce. The products of photosynthesis enabled people to have animals for farm, transport, and military use and animal products on one's table. Photosynthesis supported peasant populations that provided food for the cities, sons for the army, and taxes for the state. The combination of peasant population, weather patterns, and favorable terrain made irrigation possible in some places. A society's natural endowment of raw materials could be put to use only with solar energy. Thus, photosynthesis, transformed into slaves and charcoal, made it possible to mine and smelt metal-bearing ores. Metals were in turn used to obtain more energy or products of energy. Works of art, architecture, and literature were possible in large part because of the solar energy embedded in artists, architects, and writers; in the production of the raw materials in which artists and architects worked; and in the precious metals with which they were remunerated. Ancient states and empires relied fundamentally on agricultural taxes. Although we marvel today at the monuments and artistic works of ancient societies, these were never more than a small part of ancient economies. A. H. M. Jones, the noted historian of the late Roman Empire, estimated that the Roman government's income was based 90% on agricultural taxes. Trade and industry were comparatively undeveloped.

The persistent weakness of ancient states and empires was that the peasant producers, whose taxes funded government and the military, lived on very small margins of production. Subsistence farmers generate little surplus per capita and often none at all. Taxes were a way to compel peasants to produce more, but excessive taxation could undermine peasants' well-being and ultimately the state. Photosynthesis produces little energy per unit of land (maximum of approximately 170 kcal/m^2/year), and it takes much labor and other inputs to appropriate increasingly more of that for human use.

In ancient societies in general, agriculture could support only approximately one person per arable hectare. Vaclav Smil estimates that in Roman or early medieval times, wheat farming could yield approximately a 40-fold net energy gain. This is before subtracting seed requirements and storage losses. (The figure for 1900 AD is approximately 500-fold.) In many areas, this harvest was gained by supplementing human labor with draft animals (which require farmland and labor) and with extensive manuring. Between 180 and 250 h of human labor were combined with approximately 200 h of animal labor to produce approximately one-half metric ton of grain per hectare. In the Middle Ages, European farmers had to set aside one-third to one-half of a crop for the next year's seed. Roman-era farmers in Europe would have faced the same constraint. As discussed later, a late Roman peasant owed between one-fourth and one-third of the gross yield in taxes.

After adding seed needed for sowing to that sold for taxes, it is clear that Roman farmers and their families subsisted on but a small fraction of what they produced. Tax rates were fixed, so when a harvest was low, as frequently it would be, the margin available to support the farmer's family could be inadequate. Tax rates could harm an ancient society by making it difficult for peasants to raise families large enough to sustain the population. The consequence would be a decline in future taxes and in military manpower.

This was the characteristic pattern in ancient states and empires, which were funded by aggregating the small surpluses of many thousands to millions of impoverished subsistence farmers. Thus, the dilemma faced by ancient states was that solar energy severely limited the revenues that they could raise. Often, the direction that ancient rulers chose was to raise revenues in the short term, not realizing that this would undermine the capacity of the productive system to generate future revenue. When confronted with fiscal limits, ancient governments were perpetually tempted to expand spatially, thereby capturing more of the earth's surface where solar energy falls.

3. THEORIES OF ENERGY FLOW AND COLLAPSE

There are many theories of why societies collapse. As emphasized in the following discussion, however, most of these fail in one way or another to explain collapse fully. After describing these theories and their limitations, a newer approach is introduced.

3.1 Approaches to Energy Flow and Collapse

Students of collapse frequently blame interruptions in the resources on which a society depends. Without sufficient energy flow to sustain people and their institutions, collapse must follow. This may occur either because a resource is used to depletion or because external events, such as a change in climate, cause a once-dependable resource to become insufficient.

Such theories have been developed for a number of collapses. J. Donald Hughes, for example, argued that the ultimate cause of Rome's collapse was environmental deterioration, especially deforestation. The American geographer Ellsworth Huntington believed that climate change caused the barbarian migrations that the Roman Empire could not survive. Karl Butzer, a noted scholar of paleoenvironments, argued that insufficient Nile floods caused the collapse of Egypt's Old Kingdom (ca. 3100–2181 BC) and exacerbated other political upheavals in Egyptian history. The spectacular collapse of the Southern Lowland Classic Maya has prompted scholars to search for environmental factors, such as erosion, soil nutrient depletion, or conversion of forests to savanna grasslands. Rhys Carpenter, an accomplished student of Mycenaean civilization, argued that the Mycenaean collapse came from drought, which caused famine, depopulation, and migration. Barbara Bell argued that two so-called dark ages encompassing the eastern Mediterranean and Near East, ca. 2200–2000 and 1200–900 BC, can both be explained by widespread droughts that lasted several centuries.

The arid lands of Mesopotamia (modern Iraq, eastern Syria, and southeastern Turkey) provide some of the most compelling examples of resource depletion and collapse. There were two catastrophic collapses here: those of the Third Dynasty of Ur (ca. 2100–2000 BC) and the Abbasid Caliphate (749–1258 AD). In each case, the collapse has been explained as resulting from overirrigation of desert soils, which caused soil salinization that destroyed agricultural productivity. Sociopolitical collapse was accompanied by depopulation and environmental damage that in each case lasted centuries. Working in upper Mesopotamia at the site of Tell Leilan, Syria, Harvey Weiss found evidence that led him to suggest that collapse ca. 2200 BC in this region was caused by an abrupt change in climate that affected hundreds of thousands of both agricultural and pastoral peoples.

Some scholars focus more on human than natural resources to understand collapse. Students of the Maya periodically reformulate the idea that the stresses of supporting the Maya ruling classes, which involved mandatory labor to build lavish limestone monuments, sparked a peasant revolt that toppled the elites and their cities. Similar ideas have been advanced for collapses in Mesopotamia, Highland Mesoamerica, Peru, China, and elsewhere. In regard to the Roman Empire, Boak and Sinnigen argued that "Rome failed to develop an economic system that could give to the working classes of the Empire living conditions sufficiently advantageous to encourage them to support it devotedly and to reproduce in adequate numbers." Even in the fourth-century AD, Ammianus Marcellinus attributed the problems of the later Roman Empire to the growth of bureaucracy and excessive taxation.

A few historians, archaeologists, and other writers indict unpredictable natural cataclysms in collapses. These alter energy flows so wrenchingly that continuity or recovery are impossible. Hurricanes, earthquakes, and epidemics have all been implicated in the Maya collapse. As early as 1939, Marinatos suggested that the collapse of Minoan society on Crete (ca. 2000–1200 BC) was caused by the great eruption of the Aegean island of Thera. Popular attempts to link Plato's parable of Atlantis to an actual place have kept this idea alive. Recent writers have elaborated on the theme, even extending the effects of the devastation to the entire eastern Mediterranean.

3.2 Assessment of Approaches to Energy Flow and Collapse

Vaclav Smil observed that many collapses occurred without any evidence of weakened energy bases. Thus, energy flow alone cannot explain collapse as a recurrent historical pattern. Every society has arrangements to ameliorate changes in energy flow. Indeed, societies have withstood even pronounced constrictions in energy flow, provided that they prove temporary. Nineteenth-century examples of pronounced disruption in energy flow include the Irish potato famine and the great eruption of Krakatoa in the South Pacific. These are sometimes compared to the disasters of antiquity (e.g., the eruption of Thera), but neither caused a collapse. Catastrophes fail to explain collapse because societies withstand disasters all the time without collapsing. Moreover, sometimes the date of the catastrophe fails to match the date of the collapse.

Changes in climate also seem inadequate to account for the matter. In studies of climate and collapse there are problems of logic. In the cases of Tell Leilan, Syria, and the Hohokam of the U.S. Southwest, archaeologists have postulated that deterioration in energy flows caused by changing climate caused, at different times, both collapse of complexity and its increase. These scholars do not explain how climate change can have such opposite effects, leaving an apparent contradiction in their arguments. In the case of drought and the Mycenaean collapse, the part of Greece with the least rainfall, Attica, was the last to collapse. If deforestation contributed to the Roman collapse, why did the same problem, later in European history, help spur economic development, as Wilkinson has shown? A change in resources can by itself rarely explain collapse.

There is no doubt that energy flow is almost always involved in collapse. Often, energy flow changes in the aftermath of simplification and depopulation. Rarely, if ever, has a change in energy flow been the sole cause of a collapse. To understand the role of energy in collapse, it must be considered in the context of social, political, and economic changes.

3.3 Embedding Energy Flow and Collapse within Economics and Problem Solving

A new approach embeds energy in the social, political, and economic changes that a society undertakes to solve problems. Problem solving among humans (and even among animals) shows a general tendency to move from simple, general, and inexpensive (known colloquially as plucking the lowest fruit first) to complex, specialized, and costly. Sustainability or collapse follow from the long-term success or failure of problem-solving institutions. Throughout human history, a primary problem-solving strategy has been to increase complexity. Often, we respond to problems by creating new and more sophisticated technologies, new institutions, new social roles, or new hierarchical levels or by processing more information. Each of these strategies increases the complexity of a society. Problem solving has been much of the force behind the increasing complexity of human societies, especially during the past 12,000 years.

Changes in complexity can often be enacted quickly, which is one reason why we resort to complexity to solve problems. This is illustrated by the response to the attacks on the United States on September 11, 2001. Much of the immediate response to the problem of preventing future attacks involved increasing the complexity of public institutions by establishing new agencies, absorbing existing ones into the federal government, and expanding the scale and degree of control over realms of behavior from which a threat might arise. Such changes can be implemented more rapidly than can changes in military technology, in the technology and other techniques of espionage, or in the technology to detect explosives in air travel baggage.

As a problem-solving strategy, complexity often succeeds. The difficulty is that complexity costs. In any living system each increase in complexity has a cost, which may be measured in energy or other resources or in transformations of these, such as money, labor, or time. Before the development of fossil fuels, increasing the complexity of a society typically meant that people worked harder. This is still often the case. In the response to the September 11, 2001, attacks, anyone who travels by air understands that doing so has become more complex, and

the personal cost has risen. The cost is reflected not just in the price of travel (which is controlled by competition) but also in such intangibles as time spent standing in lines, the indignity of personal searches, and greater care regarding what personal items one travels with. These are examples of the subtle costs of complexity in problem solving.

If an increase in complexity resolves a problem, it is likely to be adopted for its short-term value. The costs of complexity are cumulative and long term. As simple, inexpensive solutions are adopted for problems in such areas as organization, information flow, or technology, the strategies that remain to be deployed are increasingly complex and costly. Complexity is an economic function and can be evaluated by returns on effort. Early efforts at problem solving often produce simple, inexpensive solutions that may yield positive returns. Yet as the highest-return ways to produce resources, process information, and organize a society are progressively implemented, problems must be addressed by more costly and less effective responses. As the costs of solutions increase, the marginal return to complexity in problem solving begins to decline. Carried far enough, diminishing returns to problem solving cause economic stagnation and disaffection of the support population. As illustrated later, a prolonged period of diminishing returns to complexity in problem solving makes a society vulnerable to collapse.

4. EXAMPLES OF ENERGY FLOW IN SOCIOPOLITICAL COLLAPSE

Energy, leadership, political organization, taxation, and competition are all involved in collapse. To illustrate this, I describe five cases in detail, including two collapses in Mesopotamia in which energy and collapse are clearly linked, although not as simply as sometimes described. The case of the Lowland Classic Maya exemplifies how competition and pressures to produce great monuments create stress on peasantry that undermines the production system. The Roman and Byzantine empires illustrate how complexity in problem solving makes a society vulnerable to collapse. In the Roman case, the collapse lasted centuries. In the Byzantine example, the collapse actually provided a basis for recovery.

4.1 Mesopotamia

Ancient Mesopotamia produced two cases that illustrate the relationship of collapse to energy—the collapses of the Third Dynasty of Ur (ca. 2100–2000 BC) and the Abbasid Caliphate (749–1258 AD).

In southern Mesopotamia, intensive irrigation initially increases agricultural yields that support growing prosperity, security, and stability. The Third Dynasty of Ur pursued this course. It expanded the irrigation system and encouraged growth of population and settlement. To capture this enhanced energy flow, it established a vast bureaucracy to collect taxes. This strategy plants the seed of its own demise, however. After a few years of overirrigating, saline groundwaters rise and destroy the basis of agricultural productivity. The political system loses its resource base and is destabilized. Large irrigation systems that require central management are useless once the state lacks the resources to maintain them.

A few centuries earlier, in the Early Dynastic period (ca. 2900–2300 BC), crop yields averaged approximately 2030 liters per hectare. Under the Third Dynasty of Ur this declined to 1134 liters. At the same time, Third Dynasty of Ur farmers had to plant their fields at an average rate of 55.1 liters per hectare, more than double the previous rate. Badly salinized lands go out of production almost indefinitely, so the pressure intensifies to get the most from the remaining fields. As yields declined and costs rose, farmers had to support an elaborate state structure. It was a system that took a high toll in labor. The Third Dynasty of Ur was following a strategy of costly intensification that clearly yielded diminishing returns.

The Third Dynasty of Ur hung on through five kings and then collapsed, with catastrophic consequences for the larger population. By 1700 BC, yields were down to approximately 718 liters per hectare. Of the fields still in production, more than one-fourth yielded on average only approximately 370 liters per hectare. The labor demands of farming a hectare of land were inelastic, so for equal efforts cultivators took in harvests approximately one-third the size of those a millennium earlier. Soon the region was extensively abandoned. By a millennium or so after the Third Dynasty of Ur, the number of settlements had decreased 40% and the settled area contracted by 77%. Population densities did not rebound to those of the Third Dynasty of Ur until the first few centuries AD, a hiatus of 2500 years.

Rulers of Mesopotamia, from the last few centuries BC into the Islamic period, repeated the strategy tried 2000 years earlier by the Third Dynasty of Ur. Population increased fivefold in the last few centuries BC and the first few centuries AD. The number of urban sites grew 900%. These trends

continued through the Sassanian period (226–642 AD), when population densities came to exceed significantly those of the Third Dynasty of Ur. At its height, the area of settlement was 61% greater than during the Third Dynasty of Ur.

Under Khosrau I (AD 531–579) the Sassanian dynasty reached its height, but the needs of the state took precedence over peasants' ability to pay. Taxes were no longer remitted for crop failure. Because the tax was fixed whatever the yield, peasants were forced to cultivate intensively. State income rose sharply under Khosrau II (590–628). This level of income would have been needed for the perpetual wars in which he engaged with the Byzantine Empire. Under the Abbasid Caliphate, an Islamic dynasty, tax assessments increased in every category. Fifty percent of a harvest was owed under the Caliph Mahdi (775–785), with many supplemental payments. Sometimes, taxes were demanded before a harvest, even before the next year's harvest.

Under the Abbasids there was unprecedented urban growth. Baghdad grew to five times the size of 10th-century Constantinople and 13 times the size of the Sassanian capital, Ctesiphon. The capital was moved often and each time built anew on a gigantic scale. The Caliph al-Mutasim (833–842) built a new capital at Samarra, 120 km upstream from Baghdad. In 46 years, he and his successors built a city that stretched along the Tigris for 35 km. It would have dwarfed imperial Rome.

As the state undertook these constructions, however, it did not always fulfill its irrigation responsibilities. As the irrigation system grew in size and complexity, maintenance that had once been within the capacity of local communities was no longer so. Communities came to depend on the imperial superstructure, which in turn became increasingly unstable. Peasants had no margins of reserve, and revolts were inevitable. Civil war and rebellion meant that the hierarchy could not manage the irrigation system. Mesopotamia experienced an unprecedented collapse. In the period from 788 to 915 revenues fell 55%. The Sawad region, at the center of the empire, had supplied 50% of the government's revenues. This dropped within a few decades to 10%. Most of this loss occurred between the years 845 and 915. In many once-prosperous areas, revenue declined 90% within a lifetime. The perimeter of state control drew inward, which diminished any chance to resolve the agricultural problems. By the early 10th century irrigation weirs were listed only in the vicinity of Baghdad. In portions of Mesopotamia the occupied area shrank by 94% by the 11th century. The population declined to the lowest level in five millennia. Urban life in 10,000 square kilometers of the Mesopotamian heartland was eliminated for centuries.

Collapses in Mesopotamia present clear evidence of the role of energy flow disruptions. When the resource base was destroyed by excessive irrigation, population levels and political complexity could not be maintained. Yet the clarity of these cases is deceptive because it obscures a fundamental question: Why did Mesopotamian rulers push their support systems to the point at which large populations and their own political systems were endangered? Such a question cannot be answered here. Rather, it underscores the point made previously that energy flows, although vital in understanding collapse, do not fully explain it. This will be clear in the next two cases.

4.2 The Southern Lowland Classic Maya

The Southern Lowland Classic Maya (ca. 250–800 AD) present one of history's most enigmatic collapses. The Maya Lowlands (Fig. 1) were one of the most densely populated areas of the preindustrial world, with an average of 200 persons per square kilometer. Tikal, one of the major Maya cities, had a population estimated at nearly 50,000, which is roughly the same as ancient Sumerian cities. In the southern part of the area known today as Quintana Roo, there are stretches of terrain 40–50 km long in which there is no gap of more than 100 m between Maya structures. The Maya are best known for their art, their cities and temples, and, as Robert Netting phrased it, their "abrupt departure from the stage of world history."

To support so many people the southern Lowland Maya transformed their landscape. Over much of the area the natural rainforest was cleared. One technique they employed was to channel and store water. Although the area today is a high tropical forest, precipitation varies seasonally, and many areas have little surface water. The Maya accordingly built canals, dams, reservoirs, and small wells and in Yucatán modified the large limestone sinkholes known as *cenotes*. There are 180 km of canals along the Rio Candelaria. They represent at least 500,000 days of labor to move (entirely with human workers) 10 million cubic meters of earth.

Hillsides across the Lowlands were often terraced for planting. Terracing directs water flow, traps silt, and creates cultivable land. Terraces are spread across more than 11,000 square kilometers. Much of the southern Lowlands (the area known as the

FIGURE 1 The Maya Lowlands and major archaeological sites.

Petén) also contains walled fields, a type of feature that indicates intensive, sustained cultivation.

Perhaps the most interesting agricultural features of the Lowlands are systems of raised fields and associated canals. These were built to supply dry, cultivable land in areas otherwise inundated. Raised fields were built within lakes and lagoons and at river margins, but they were primarily built within swamps, thereby reclaiming land that could be highly productive. Modified and intensively cultivated swamps may have become the most valuable land. Not surprisingly, the larger cities tend to be located near swamps. By building raised fields in such areas, the Maya could have maintained a moist, root-level environment; fertilized fields with rich, organic muck from canal bottoms; and perhaps raised fish as well. Tended carefully, such fields could have supported continuous cropping.

By the use of such facilities the energy of the Lowlands was channeled to produce a simplified

biota consisting of maize, squash, avocado, cacao, cotton, root crops, and fruit-bearing trees. Simplified food webs became the foundation for a diverse and complex society. Ranks of bureaucrats, artisans, elites, and paramount rulers held sway in cities whose centers were built of gleaming limestone. Their sense of time was based on astronomical observations so precise that they were not equaled in Europe or China for centuries to come. The Maya developed a system of writing that has only recently been deciphered. Their art is greatly admired today. All of this complexity was based on the energy flow made available from their managed landscape.

One consequence of the development of complex societies in the Maya area is that the cities were often at war. Warfare is a prominent theme in Maya art, with many scenes of domination and execution of captives. There are great earthen fortifications at such sites as Becan, Calakmul, Oxpemul, and Tikal. The defensive works at Tikal include a moat and earthwork 9.5 km long. Warfare in turn reinforced centralized political authority.

Maya complexity imposed a high cost on the support population. Great efforts were required to transform the forest, keep the land from reverting to forest, and keep the soil fertile; to support non-producing classes of bureaucrats, priests, artisans, and rulers; and to build and maintain the great cities. Over time these costs actually grew. The construction of monuments increased over the years. Across the southern Lowlands, 60% of all dated monuments were built in a period of 69 years, between 687 and 756 AD.

All this effort took its toll on the support population. At Tikal, by approximately 1 AD, the elites began to grow taller than the peasants, and within a few centuries the difference averaged 7 cm. That elites should have enjoyed better nutrition during childhood is not surprising. By the late Classic (ca. 600–800 AD), however, the stature of adult males actually dropped among both elites and peasants. The nutritional value of the Tikal diet had apparently declined for both rulers and ruled. At the site of Altar de Sacrificios there was a high incidence of malnutrition and/or parasitic infections. Childhood growth was interrupted by famine or disease. Many people suffered from anemia, pathologic lesions, and scurvy. At Altar de Sacrificios, the stature of adult males declined in the first few centuries AD, and life expectancy declined in the late Classic. Not surprisingly, the population of the southern Lowlands, which had grown during the first several centuries AD, leveled off between 600 and 830 and thereafter grew no more. Such population trends often foreshadow collapse.

The magnificence of late Maya cities clearly came at the expense of the peasant population, and ultimately at the cost of the entire system's sustainability. It was a classic case of diminishing returns to complexity. As the complexity of the political system and the cities increased through time, the costs of the system, in human labor, also increased. It is clear that the marginal return to these investments was declining by the late Classic. Ever greater investments in military preparedness, monumental construction, and agricultural intensification brought no proportionate return in the well-being of most of the population. To the contrary, as the demands on the peasant population increased, the benefits to that population actually declined. This was particularly the case in the seventh and eighth centuries, when a great increase in laborious monumental construction fell on a stressed and weakened population that was no longer growing.

The consequences were predictable. One by one, the Maya polities began to collapse and their cities were progressively abandoned (Fig. 2). At the same time, the southern Lowlands experienced one of history's major demographic disasters. In the course of approximately a century, the population declined by 1–2.5 million people and ultimately declined to a fraction of its peak. There were not enough Maya left to control the landscape, nor did those remaining need to farm as they once had. Intensive agriculture was abandoned and the forests regrew. Cultural complexity in this region was based on an engineered landscape with low species diversity, short food chains, and diversion of energy from plant to human

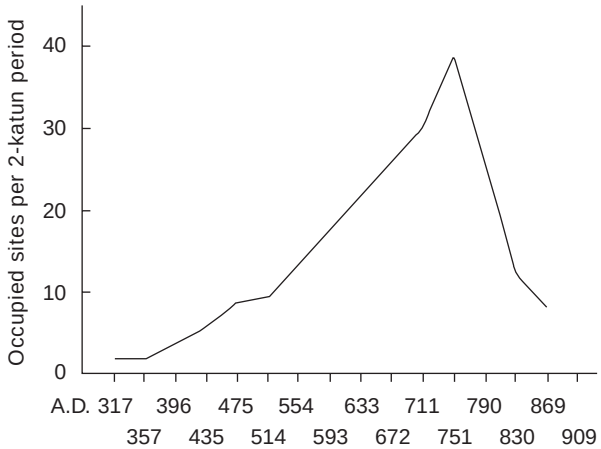


FIGURE 2 Occupation of Classic Maya centers. One katun is approximately 20 years.

biomass. In the long term, although the Lowlands landscape could seemingly support dense human populations, it could not sustain the complex political systems that large populations often require.

4.3 Collapse of the Western Roman Empire

The Western Roman Empire (Fig. 3) is one of history's most instructive collapses. The Romans' early success came from a means of expansion that was fiscally self-perpetuating. Defeated peoples gave the economic basis, and some of the manpower, for further expansion. It was a strategy with high economic returns. By 167 BC, for example, the Romans were able to eliminate taxation of themselves and still expand the empire. Yet the economics of an empire such as the Romans assembled are deceptive. The returns to any campaign of conquest are highest initially, when the accumulated surpluses of the conquered peoples are appropriated. Thereafter, the conqueror assumes the cost of administering and defending the province. These responsibilities may last centuries and are paid from yearly agricultural surpluses. The fiscal basis of such an empire shifts from accumulated surpluses (generally embodied in precious metals, works of art, and population) to current solar energy.

Once the phase of conquest was complete, and the accumulated surpluses spent, the Roman government was financed by agricultural taxes that barely sufficed for ordinary administration. When extraordinary expenses arose, typically during wars, the precious metals on hand frequently were insufficient to produce the required coinage. Facing the costs of war with Parthia and rebuilding Rome after the Great Fire, Nero began in 64 AD a policy that later emperors found irresistible. He debased the primary silver coin, the denarius, reducing the alloy from 98 to 93% silver. It was the first step down a slope that resulted two centuries later in a currency that was worthless (Fig. 4) and a government that was insolvent.

In the half century from 235 to 284, the empire nearly came to an end. There were foreign and civil

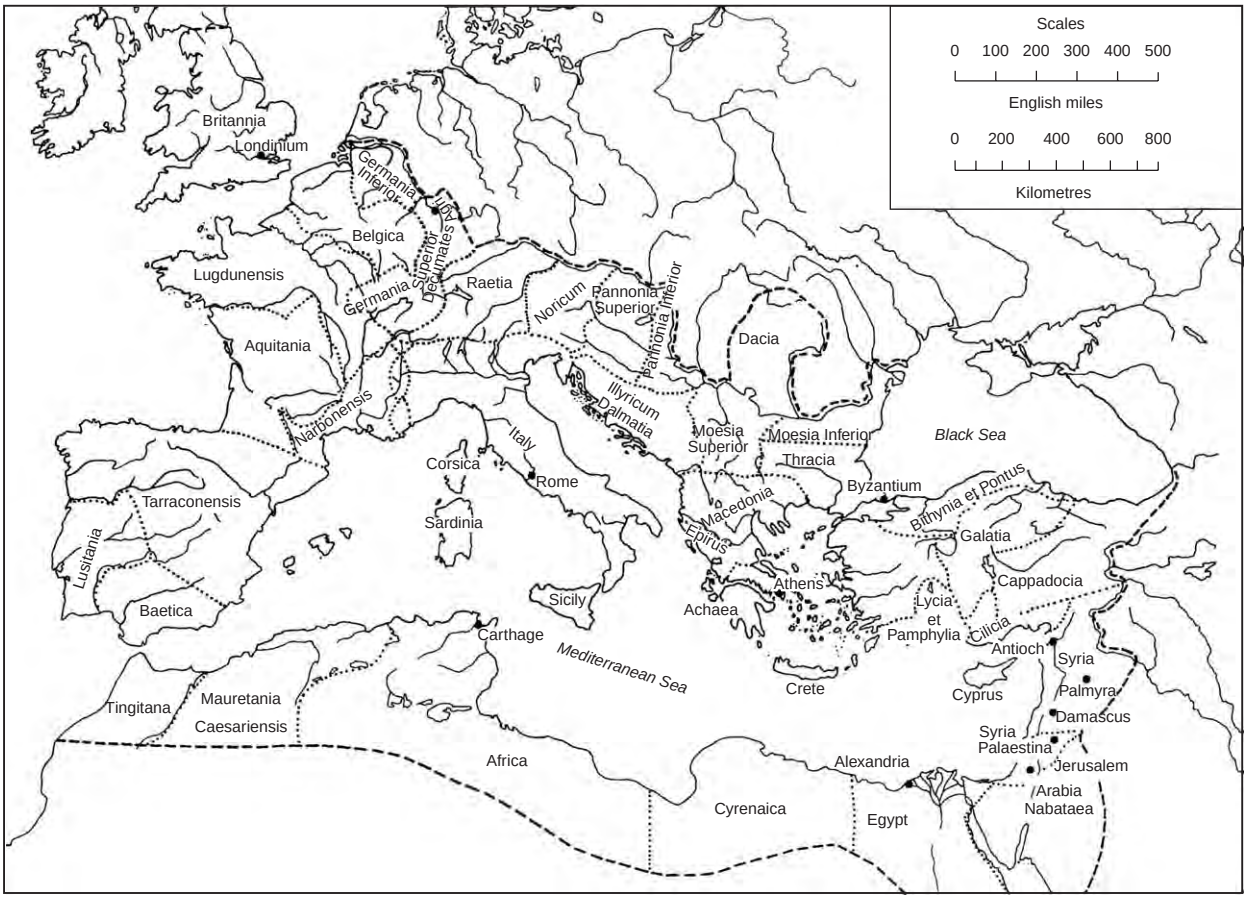


FIGURE 3 The Roman Empire in 117 AD.

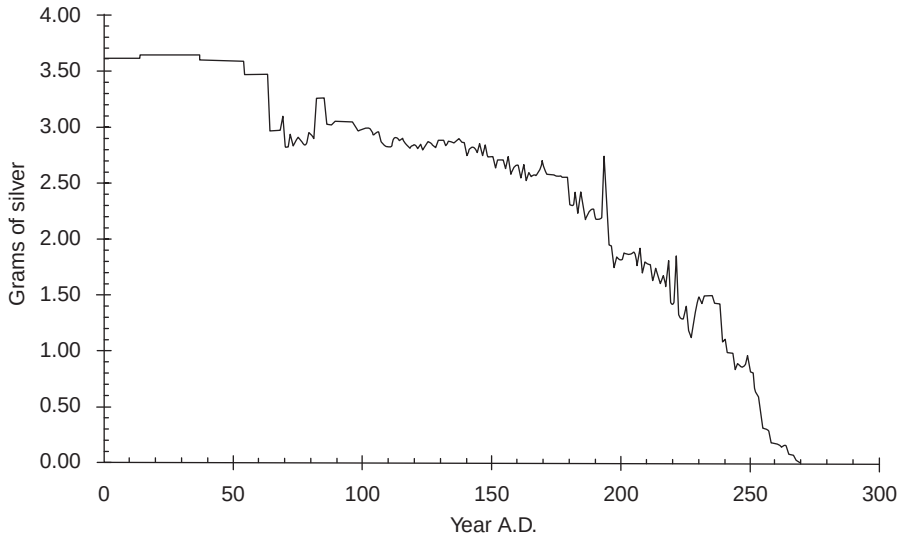


FIGURE 4 Debasement of the Roman silver currency to 269 AD.

wars almost without interruption. The period witnessed 26 legitimate emperors and perhaps 50 usurpers. Cities were sacked and frontier provinces devastated. The empire shrank in the 260s to Italy, the Balkans, and North Africa. By prodigious effort the empire survived the crisis, but it emerged at the turn of the fourth-century AD as a very different organization.

In the late third and early fourth centuries, Diocletian (284–305) and Constantine (306–337) designed a government that was larger, more complex, more highly organized, and much more costly. They doubled the size of the army, always the major part of imperial costs. To pay for this, the government taxed its citizens more heavily, conscripted their labor, and dictated their occupations. Increasingly more of the empire's energy was devoted to maintaining the state.

Diocletian established Rome's first budget, and each year a tax rate was calculated to provide the revenue. The tax was established from a master list of the empire's people and lands, tabulated down to individual households and fields. In an era when travel and communication were slow, expensive, and unreliable, it took substantial organization and personnel just to establish and administer a tax system so minutely detailed. Taxes increased, apparently doubling between 324 and 364. Villages were liable for the taxes on their members, and one village could even be held liable for another. Tax obligations were extended to widows, orphans, and dowries. Even still, to meet its needs, the government had to conscript men for the army and requisition services. Occupations were made hereditary and obligatory. Positions that had once been eagerly sought, such as in city senates, became burdensome because leading citizens were held responsible for tax deficiencies.

The tax system supporting the more complex government and larger army had unforeseen consequences. After plagues decimated the population in the second and third centuries, conditions were never favorable for recovery. There were shortages of labor in agriculture, industry, the military, and the civil service. The tax system of the late empire seems to have been to blame because the rates were so high that peasants could not accumulate reserves nor support large families. Whatever crops were brought in had to be sold for taxes, even if it meant starvation for the farmer and his family. Farmers who could not pay their taxes were jailed, sold their children into slavery, or abandoned their homes and fields. In such circumstances it became unprofitable to cultivate marginal land because too often it would not yield enough for taxes and a surplus. Despite government edicts, marginal lands went out of cultivation. In some provinces, up to one-third to one-half of arable lands came to be deserted by the late empire. Faced with taxes, peasants would abandon their lands and flee to the protection of a wealthy landowner, who was glad to have the extra labor.

On the death of the emperor Theodosius in 395 AD, the Roman Empire was divided into eastern and western parts, each of which had to raise its own revenues and most of its own troops. The east was by far the wealthier half, with fewer problems at its borders. From the late fourth century the barbarians

could no longer be kept out. They forced their way into Roman lands in western Europe and North Africa, initially causing great destruction. The government had no choice but to acknowledge them as legitimate rulers of the territories they occupied. Throughout the fifth century the western empire was in a positive feedback loop tending toward collapse. Lost or devastated provinces meant lower government income and less military strength. Lower military strength in turn meant that more areas would be lost or ravaged.

In the 20 years following the death of emperor Valentinian III (455), the western Roman army disappeared. The government came to rely almost exclusively on mercenaries from Germanic tribes. Finally, these could not be paid. They demanded one-third of the land in Italy in lieu of pay. This being refused, they revolted and deposed the last emperor in Italy, Romulus Augustulus, in 476.

The strategy of the later Roman Empire was to respond to a near-fatal challenge in the third century by increasing the size, complexity, power, and costliness of the primary problem-solving system—the government and its army. The higher costs were undertaken not to expand the empire or to acquire new wealth but to sustain the status quo. The benefit/cost ratio of imperial government declined as it lost both legitimacy and support. In the end, the Western Roman Empire could no longer afford the problem of its own existence.

5. ENERGY FLOW, COLLAPSE, AND RESILIENCY

It is common to think of collapse as a catastrophe, involving the loss of great traditions of art, architecture, and literature. In the case of the early Byzantine Empire (the Eastern Roman Empire, 395–1453 AD), though, the reduction of complexity that is the defining characteristic of collapse contributed to resiliency and recovery. The Byzantine case may be unique among collapses. It is especially instructive in regard to the economics of complexity in problem solving.

The Eastern Roman Empire survived the fifth-century debacle and met its end only when the Turks took Constantinople in 1453. The Byzantine Empire lost territory throughout much of its history and several times nearly fell. Yet it survived near catastrophes through one of history's most unusual strategies.

The eastern emperor Anastasius (491–518) established a sound coinage in the copper denominations on which daily life depended. He gave the army generous allowances for rations, arms, and food. Commerce was revitalized and the army again attracted native volunteers. Within a few decades these reforms had produced such results that Justinian (527–565) could attempt to recover the western provinces.

An army sent to North Africa in 532 conquered the Kingdom of the Vandals within 1 year. Almost immediately, the Byzantine general Belisarius was sent to reconquer Italy. He had captured the Ostrogothic King and nearly all of Italy when he was recalled in 540 to fight the Sassanian Persians.

In 541, when the job in Italy seemed about done, bubonic plague swept over the empire. It had not been seen before in the Mediterranean. Like any disease introduced to a population with no resistance, the effects were devastating. Just as in the 14th century, the plague of the sixth century killed one-fourth to one-third of the population. The enormous loss of taxpayers caused immediate financial problems. Army pay fell into arrears, and both Italy and North Africa were nearly lost. By debasing the currency and slashing expenditures, the emperor was able to send another army to Italy. Italy was reconquered by 554, but much of it was lost again when the Lombards invaded in 572. Throughout the rest of the sixth century, costly wars continued in Italy, North Africa, the Balkans, and with the Persians.

A rebellion of the Byzantine army in 602 rekindled the Persian war and began a crisis that lasted more than a century and nearly brought the empire to an end. The war against the Persians went badly. Sensing Byzantine weakness, Slavs and Avars overran the Balkans. North Africa and Egypt rebelled and placed Heraclius (610–641) on the Byzantine throne. The empire he took over was financially exhausted. The Persians captured Syria, Palestine, and Egypt, overran Anatolia, and besieged Constantinople from 618 to 626.

In 616, Heraclius cut the pay of troops and officials in half. He followed his predecessors by lowering the weight of the primary copper coin, the follis (Fig. 5). Heraclius' economic measures bought time for his military strategy to work. In 627, Heraclius destroyed the Persian army. The Byzantines got all their lost territory returned. The war had lasted 26 years and resulted in no more than restoration of the status quo of a generation earlier.

The empire was exhausted by the struggle. Arab forces, newly converted to Islam, defeated the Byzantine army decisively in 636. Syria and Palestine, which had taken 18 years to recover, were lost

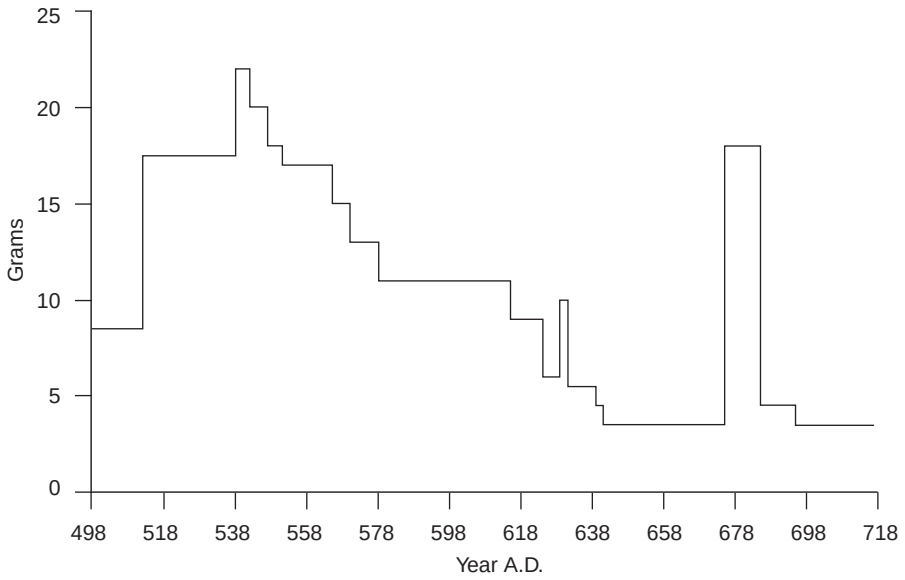


FIGURE 5 Weight of the Byzantine follis, 498–717 AD.

again. Egypt was taken in 641. The wealthiest provinces were permanently gone, and soon the empire was reduced to Anatolia, Armenia, North Africa, Sicily, and parts of Italy. The Arabs conquered the Persian Empire entirely. (In time, parts of the former Persian Empire became the center of the Abbasid Caliphate, described previously.)

Under Constans II (641–668) and throughout the seventh century, the situation continued to deteriorate. It appeared that it would be only a matter of time before the remaining Byzantine lands fell to the Arabs. The Arabs raided Asia Minor nearly every year for two centuries. Constantinople was besieged each year from 674 to 678, then continuously from 717 to 718. The Byzantines defeated both sieges, and this proved to be the turning point in a centuries-long struggle. The Arabs were never again able to mount such a threat.

During this century of crisis the social and economic life of the eastern Mediterranean was thoroughly transformed. Population declined. Monetary standards and economic institutions were destroyed. Around 659, Constans II cut military pay in half again. Army pay now stood at one-fourth its level in 615 so that the economy no longer received large influxes of new coin. By 700 most people of the eastern Mediterranean no longer used coins in everyday life. The economy developed into its medieval form, organized around self-sufficient manors.

Roman emperors of the late third and early fourth centuries had responded to a similar crisis by complexification. They increased the complexity of administration, the regimentation of the population, and the size of the army. This was paid for by levels of taxation so harmful that lands were abandoned and peasants could not replenish the population. Constans II and his successors could hardly impose more of the same exploitation on the depleted population of the shrunken empire. Instead, they adopted a strategy that is truly rare in the history of complex societies: systematic simplification, in effect a deliberate collapse.

The government had lost so much revenue that even at one-fourth the previous rate it could not pay its troops. Constans devised a way for the army to support itself. Available lands were divided among the troops. Soldiers (and later sailors) were given grants of land on condition of hereditary military service. Constans expected the troops to provide their own livelihood through farming, with a small monetary supplement. Correspondingly, the Byzantine fiscal administration was greatly simplified.

The transformation ramified throughout Byzantine society. Both central and provincial government were simplified. In the provinces, the civil administration was merged into the military. Cities throughout Anatolia contracted to fortified hilltops. Aristocratic life focused on the imperial court. There was little education beyond basic literacy and numeracy, and literature consisted of little more than lives of saints. The period is sometimes called the Byzantine dark age.

The simplification rejuvenated Byzantium. The new farmer–soldiers became producers rather than

consumers of the empire's wealth. From this new class of farmers came the force that sustained the empire. By lowering the cost of military defense the Byzantines secured a better return on their most important investment. The empire began to lose land at a much slower rate. The Arabs continued to raid Anatolia but were unable to hold any of it for long. Soldiers were always nearby. Fighting as they were for their own lands and families, they had much greater incentive and performed better.

Once the threat to the empire's existence was over, campaigns against the Bulgars and Slavs gradually extended the empire in the Balkans. Greece was recaptured. In the 10th century the Byzantines reconquered parts of coastal Syria. Overall after 840 the size of the empire was nearly doubled. The process culminated when Basil II (963–1025) conquered the Bulgars and extended the empire's boundaries again to the Danube. In two centuries the Byzantines went from near disintegration to being the premier power in Europe and the Near East, an accomplishment won by a deliberate collapse in which they systematically simplified, and thereby reduced the costs of, their army, government, and society.

6. ENERGY FLOW IN THE AFTERMATH OF COLLAPSE

Since collapse is the loss of complexity, and higher levels of complexity depend on greater flows of energy, a primary characteristic of postcollapse societies is that they require and produce less energy. This is a consequence of collapse that occurs because there are fewer elite ranks to extract resources from the populace, because expensive sectors of investment (such as the Roman army) disappear, or because a lower degree of social and economic specialization means that peasant households no longer have to support specialists. A late Roman peasant, for example, might have paid one-fourth to one-third of gross yields in taxes if he owned his own land. If he rented land, taxes and rent together took one-half to two-thirds of a crop. Thus, Roman peasants were perpetually impoverished, often undernourished, and sometimes forced to sell children into slavery when they could not be fed. This undermined the productive capacity of the empire and its sustainability. After the plagues of the second and third centuries AD, peasants were unable to replenish the population, and the empire lost much of its energy-producing capacity. In post-Roman Gaul, in contrast, taxes declined to as low as 10%. Early medieval peasants may have been better off and seen more of their children survive.

The Byzantines responded to lower energy flow by simplifying their social, political, and economic systems and by shortening energy flow networks. In the sixth-century AD, energy would typically have flowed from the sun to farms; from farms to peasants; from peasants to purchasers of grain, thereby converting the grain to coinage; from peasants to tax officials; from tax officials to the government; and from the government to the army. At each step some energy was lost to transaction costs. Just as in a trophic pyramid, energy was lost each time it was transformed or passed to another level. After the mid-seventh century, the energy on which the empire depended flowed through fewer transactions and levels. Energy passed from the sun to farms. From there part of it was harvested and used directly by soldiers, eliminating many intermediate steps. Other peasants still paid taxes, but since the government spent less of this revenue on the army there was more to allocate to other needs. Overall much less energy went to transaction costs, so net energy increased as a proportion of gross production. The empire revived and went on to expand.

In general, postcollapse societies are smaller, with simpler hierarchies and less socioeconomic specialization. They produce less monumental architecture or none at all. Population often declines, sometimes dramatically as in the Maya collapse. Many settlements are abandoned, and those that remain may contract. Long-distance trade declines. The known world shrinks, and over the horizon lies the unknown. There are often declines in literacy and numeracy because the simpler society requires less of these. Since there are fewer elites for whom to produce art, and fewer artisans, postcollapse societies produce less art. All these changes involve reductions in energy extraction and flow, either as cause or as consequence. In some cases, energy flow is so greatly reduced that postcollapse societies are nearly impossible for archaeologists to detect.

7. SYNTHESIS AND IMPLICATIONS

Collapse is a recurrent historical process. It affects the entire range of societies, from those that are least complex to the most complex. Perhaps no society is immune to the possibility of collapse, including the very complex societies of today. The misfortune of

some previous societies is potentially our good fortune. By understanding the processes that cause collapse we can develop tools to assess our own vulnerability.

Earlier research often pinpointed diminished energy flow as the cause of collapse. In many cases, though, diminished energy flow was the product of human management. This is evident in the examples of the Third Dynasty of Ur and the Abbasid Caliphate. In both cases, excessive irrigation caused salinization that destroyed productive soils. In these cases energy does not suffice to explain collapse. Other factors were involved, particularly the tendency of rulers to demand more than the land could sustainably produce. When we ask why the rulers did this, we realize that energy is but a small part of explaining these collapses. The Maya drastically altered their environment, shortening food chains and diverting large parts of the region's productivity to support dense human populations and their complex societies. Yet demands to support complexity so stressed the population that the system collapsed, the population plummeted, and the region reverted to rainforest. Again we see the interplay of energy with politics and human strategies of survival and competition.

Problem solving links these examples to a more general understanding of collapse. Moreover, problem solving is the key to using past collapses to understand our situation today. As noted previously, sustainability is a function of problem solving, as is its opposite, collapse. As the problems that a society or institution faces grow in size and complexity, problem solving grows in complexity and costliness. Problem solving is an economic process, and its success depends on the ratio of the benefits of problem solving to its costs. Problem solving tends to develop from solutions that are simple and inexpensive to those that are increasingly complex and costly. Each increase in complexity has costs, which may be measured in labor, time, money, or energy. Growth in complexity, like any investment, ultimately reaches diminishing returns, at which point increments to complexity no longer yield proportionate returns. When this point is reached, a society or other institution begins to experience economic weakness and disaffection of the populace. A fundamental element of a sustainable society is that it must have sustainable institutions of problem solving. These will be institutions that give stable or increasing returns (e.g., by minimizing costs) or diminishing returns that can be subsidized by energy supplies of assured quality, quantity, and cost. Ancient societies, constrained by photosynthesis, lacked the energy resources to pay for ever-increasing complexity. The problem persists among today's complex societies, which can pay for complex problem solving only as long as energy is abundant and inexpensive.

The role of problem solving is especially clear in the collapse of the Roman Empire. Confronted in the third century AD with severe challenges to its existence, the empire responded by greatly expanding the complexity and size of its primary problem-solving institutions: the army and government. This was costly and had to be paid for by greatly increasing taxation. The costs were so high that the empire survived by consuming its capital resources: producing lands and peasant population. This was an unsustainable strategy and collapse was inevitable. All that the empire could accomplish was to postpone it. Collapse resulted from the economic consequences of diminishing returns to complexity in problem solving. Ironically, though, had the empire not taken the steps it did, it would have collapsed even sooner.

Problem solving is the element missing in earlier collapse studies that focused unsatisfactorily on energy alone. A prolonged period of diminishing returns to complexity in problem solving renders a society fiscally weak (Figs. 4 and 5) and causes disaffection of the support population. Both factors render a society more vulnerable to stresses than would otherwise be the case. Thus, a fluctuation in energy flow that a fiscally prosperous society would survive can cause the collapse of a society weakened by the cost of great complexity in problem solving. This helps to clarify why changes in climate can have different effects in different circumstances. A society weakened by the costs of problem solving may be vulnerable to a change in climate that in other circumstances may be survivable.

Problem solving links the experiences of the past to understanding our own circumstances. The economics of problem solving, from less to more complex and from increasing to diminishing returns, constrain all institutions, including those of today. An institution that experiences a prolonged period of diminishing returns to complexity in problem solving will find itself in fiscal distress and ineffective at solving problems. If such an institution is part of a larger system that provides its funding (such as a government agency funded by appropriations), it may be terminated or require ever-larger subsidies. If an entire society finds itself in this position, as ancient societies sometimes did, the well-being of its

populace will decline and it may be unable to surmount a major challenge. Either weakness can make a society vulnerable to collapse.

How might this dilemma make itself felt today? As an example, consider the relationship of global climate change to complexity in problem solving. Nothing before in history equals contemporary global change in magnitude, speed, and numbers of people affected. Our response will involve changes in the complexity of technology, research, and government. We need much research to comprehend climate, the economic dislocations that it will entail, and the new technologies that we will need. We will experience increasing government centralization and regulation. In research, technology, and regulation, and in resettlement of people living in low-lying areas, global change will require societal expenditures that are currently beyond calculation. As discussed previously, past societies were made vulnerable to climate change when such a perturbation came after a period of diminishing returns to investment in complexity. If this is so, it would be wise for us to understand the historical position of our own problem-solving efforts. This is not to suggest that climate change will cause today's complex societies to collapse. What it may do is combine with other challenges to increase the complexity and cost of our problem-solving efforts. Barring dramatic developments that increase the availability and decrease the cost of energy, such developments would reduce fiscal resiliency and living standards and make it more difficult to resolve our problems.

The factors that lead to collapse or sustainability develop over decades, generations, or centuries. We need historical research to understand our position in a problem-solving trajectory. In medicine, for example, we have long been in a phase of diminishing returns in which ever-greater expenditures yield smaller and smaller increments to life expectancy. Once we understand this, it is clear that the costs of health care cannot be contained as long as we demand health and long lives, which we will certainly continue to do. Conversely, our investments in microelectronics appear to be in a phase of increasing returns. This happy state, of course, cannot last forever. A primary element to understanding the sustainability of a society, an institution, or a sector of investment is to understand long-term changes in the complexity and costliness of problem solving.

Our citizens and policy makers have such confidence in science that a problem such as global climate change can generate complacency. We assume that a technological solution will emerge. Few of us are aware that problem solving can reach the point of diminishing returns and become ineffective. One of our great advantages over societies of the past is that it is possible for us to learn where we are in the long-term historical patterns that lead to sustainability or collapse. The historical sciences are an essential aspect of being a sustainable society.

SEE ALSO THE FOLLOWING ARTICLES

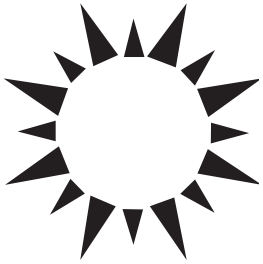
Cultural Evolution and Energy • Easter Island: Resource Depletion and Collapse • Economic Thought, History of Energy in • Transitions in Energy Use • War and Energy • World History and Energy

Further Reading

Adams, R. McC. (1981). "Heartland of Cities." Aldine, Chicago.
Allen, T. F. H., Tainter, J. A., and Hoekstra, T. W. (2003). "Supply-Side Sustainability." Columbia Univ. Press, New York.
Bell, B. (1971). The dark ages in ancient history: 1. The first dark age in Egypt. *Am. J. Archaeol.* **75**, 1–26.
Culbert, T. P. (ed.). (1973). "The Classic Maya Collapse." Univ. of New Mexico Press, Albuquerque.
Jones, A. H. M. (1964). "The Later Roman Empire, 284–602: A Social, Economic and Administrative Survey." Univ. of Oklahoma Press, Norman.
Renfrew, C. (1979). Systems collapse as social transformation: Catastrophe and anastrophe in early state societies. *In* "Transformations: Mathematical Approaches to Culture Change" (C. Renfrew, and K. L. Cooke, Eds.), pp. 481–506. Academic Press, New York.
Tainter, J. A. (1988). "The Collapse of Complex Societies." Cambridge Univ. Press, Cambridge, UK.
Tainter, J. A. (1999). Post-collapse societies. *In* "Companion Encyclopedia of Archaeology" (G. Barker, Ed.), pp. 988–1039. Routledge, London.
Tainter, J. A. (2000). Problem solving: Complexity, history, sustainability. *Population Environ.* **22**, 3–41.
Yoffee, N., and Cowgill, G. L. (eds.). (1988). "The Collapse of Ancient States and Civilizations." Univ. of Arizona Press, Tucson.

Solar Cells

RICHARD CORKISH
University of New South Wales
Sydney, New South Wales, Australia



1. Introduction
2. How Solar Cells Work
3. Efficiency
4. Types of Solar Cells
5. Conclusions

Glossary

diffuse insolation Solar radiation received from the sky, excluding the solar disk.

direct (beam) insolation Solar radiation received directly from the visible solar disk.

efficiency Fraction of incident light power converted to electrical power by a solar cell or module.

global insolation The sum of direct and diffuse insolation components.

insolation Flux of radiant energy from the sun, measured as power per unit area, whose intensity and spectral content varies at the earth's surface due to time of day, season, cloud cover, moisture content of the air, and other factors, but is constant outside the atmosphere at 1353 W m^{-2}.

peak watt (Wp) Rating for a cell, module, or system that produces 1 W electrical output under standard peak insolation conditions and at a standard cell or module temperature, usually 1000 W m^{-2} with global AM1.5 spectrum at 25°C for terrestrial applications, intended to simulate midday conditions at middle latitudes.

photovoltaic effect Generation of an electromotive force (voltage) due to incident photon radiation.

solar module Protectively packaged set of solar cells, electrically interconnected (usually several cells connected in series) to provide useful voltage and current output.

solar spectra Sunlight power content as a function of light color (of radiation frequency) and for which there exist several "standard" representations to aid calculations and comparisons, particularly the air-mass 1.5 (AM1.5) spectrum, which models transmittance through 1.5 atmospheric thicknesses for terrestrial modeling, air-mass zero (AM0), which models extraterrestrial sunlight, and ideal black-body spectra that model the sun as a thermodynamic black body at temperatures of 6000 K or less.

Solar cells are semiconductor devices that exploit the photovoltaic effect to directly create electric current and voltage from the collection of photons (quanta of light). They convert sunlight to electricity silently and without moving parts, require little maintenance, and are reliable. They are sold with warranties of up to 30 years, generate no greenhouse gases in operation, and are modular, rapidly deployable, and particularly suited to urban rooftops and façades. Solar cells power anything from lanterns in remote villages; to watches, calculators, solar home systems, and grid-connected rooftop arrays; to satellites and exploration vehicles on Mars.

1. INTRODUCTION

According to the review of the field edited by Archer and Hill, solar cells of 15% efficiency covering an area equivalent to just 0.25% of the global area under crops and permanent pasture could meet all the world's primary energy requirements, when most or all of that area would be otherwise alienated land. However, solar cells remain an expensive option for most power-generation requirements relative to fossil and nuclear sources, especially when the natural environment is attributed little or no value, and to some other sustainable options such as wind energy or energy efficiency enhancement.

There is a wide range of solar cells materials and approaches in production, under development, or having been proposed. One way to understand the range is to see the field as comprising three generations of technology. The first includes "bulk" or "wafer" cells, made from layers of semiconductor material thick enough to be self-supporting, usually up to 0.5 mm. This generation, which dominates

 © 2004 Elsevier Inc. All rights reserved.

markets, is exemplified by single and multicrystalline silicon cells and commercial production cells, which currently have efficiencies of 10 to 15%. Cost reductions for such cells will eventually be limited by the costs of the wafers. The second generation aims for lower cost at the expense of efficiency by using thin layers of active material. A rigid substrate or superstrate provides mechanical support for the semiconductor, which is deposited rapidly over large areas. It is hoped that eventually the relatively low material cost of the substrate or superstrate will dominate the cost of the complete device. Unlike the wafer-based cells described earlier, which are mechanically and electrically connected together to form weatherproof solar modules that produce convenient voltages and currents, the thin film materials are deposited on large areas of sub/superstrate and subsequently patterned to form multiple cells. Hence, the basic production unit is the module, rather than the individual cell. The four important thin-film technologies are amorphous silicon, polycrystalline silicon, cadmium telluride, and copper indium diselenide. Current production efficiencies are 4 to 9%. Thin-film cells have been expected to overtake the first-generation cells' dominance for decades, but this has not yet occurred. In fact, the evolution may be faltering since one major cell manufacturer ceased production of its thin-film products late in 2002 and others struggle for financial viability. Third-generation approaches intend to eventually combine high efficiency and low cost. One type of third-generation cells, tandem cells, are already commercially available as both low-cost, low-efficiency amorphous silicon-germanium and, at the other performance extreme, ultra-high efficiency cells made from elements of Groups III and V of the Periodic Table (known simply as III-V cells).

Market surveys estimated world production to be around 560 MW_p in 2002, an estimated 39% increase over the previous year. Ten companies accounted for 84% of global production in 2002. The three largest were Sharp, BP Solar, and Kyocera, in order of production volume. Sharp's rapid expansion during 2002 gave it nearly twice the capacity of its next largest rival. Figures 1 and 2 indicate the history of regional and global production levels and the end uses.

One figure of merit on which solar cells and modules are compared is their efficiency. Record efficiencies for research cells and modules using different technologies are published regularly by the journal *Progress in Photovoltaics*. Another more commercially important but less precisely defined

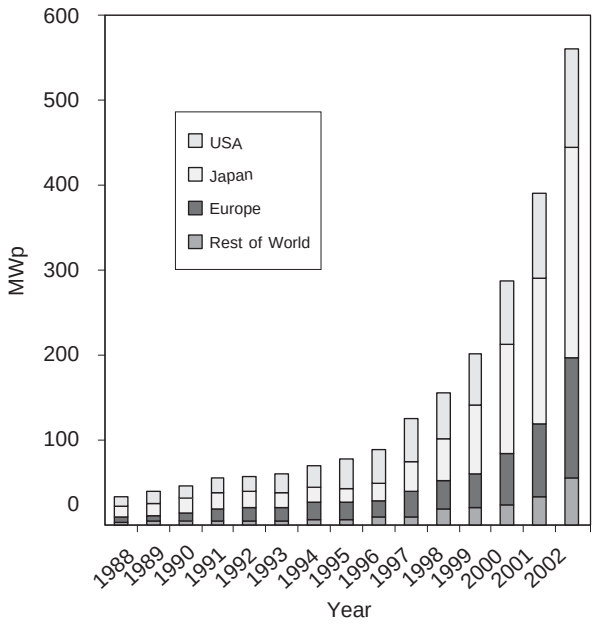


FIGURE 1 World solar cell production, 1988–2002, for the three main production zones and for the rest of the world (ROW). Data from *Photon International* and *Renewable Energy World*.

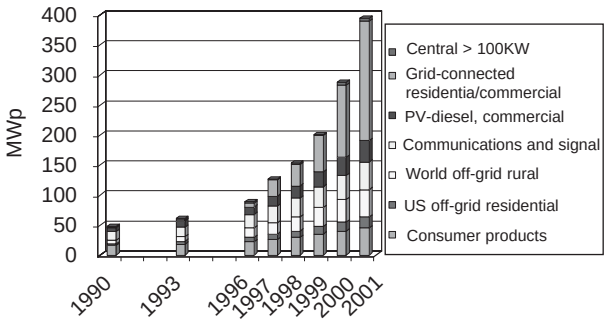


FIGURE 2 World PV market by application, 1990–2001. Data from Renewable Energy World.

figure of merit is the cost per W_p. This recognizes that efficiency is only of prime importance in some applications, and for many potential users, the capital cost and the cost of the resultant electricity are more useful measures on which to base decisions. Hence, technologies that are less efficient but cheaper per unit area are able to compete in the market.

2. HOW SOLAR CELLS WORK

Solar cells use semiconducting materials to achieve two basic tasks: (1) the absorption of light energy to generate free charge carriers within the material and (2) the separation of the negative and positive charge

carriers to produce unidirectional electrical current through terminals that have a voltage difference between them. The separation function is usually achieved by a p-n junction, the interface of regions of material that have been "doped" with different impurities to give an excess of free electrons (n-type) on one side of the junction and a dearth of them (p-type) on the other. The interface region therefore has a built-in electrostatic field that sweeps electrons one way and holes the other. In other words, a solar cell is an illuminated, large-area diode.

Figure 3 indicates the main processes. Photons enter the cell volume through the front surface. High-energy (blue or violet) photons are absorbed strongly close to the cell surface, but those from the other end of the visible spectrum are weakly absorbed and penetrate more deeply. Each absorption of a photon annihilates the photon and transfers its energy to an electron in the semiconductor, freeing the electron from its parent atom and leaving behind a positively charged vacancy, or "hole." In some materials, known as direct bandgap semiconductors, the absorption involves just the photon and the created carrier pair, but in others, indirect bandgap semiconductors, the process also needs one or more phonons (quanta of crystal atomic vibration) to take part. Absorption is weaker in the latter semiconductors. Gallium arsenide and silicon are examples of direct and indirect bandgap semiconductors. The electrons and holes are mobile within the cell and will move in response to an electric field ("drift") or by diffusion to regions of lower concentration. They are also prone to recombination with each other. Cell theory usually concentrates on minority carriers (electrons in p-type material and holes in n-type) since their flows (currents) principally determine the performance. Electron-hole pairs generated near the junction are split apart by the strong electric field there. Minority carriers are swept across the junction to become majority carriers and each crossing of the junction is a contribution to the cell's output current. Minority carriers that are generated too far from the junction to be immediately affected by the junction field can be transported to the junction by diffusion due to the minority carrier population being reduced near the junction by the field action there. Metal contacts at the front and rear of the cell allow connection of the generated current to a load. The front contact is normally in the form of a fine metallic grid to reduce blockage of the light's access to the semiconductor.

However, survival of a minority carrier is tenuous since it is always prone to recombination with one of the surrounding majority carriers. Recombination can be radiative, the inverse of the optical generation process that produced the carrier pair, in which the carrier energy is lost in the production of a new photon, or nonradiative, in which the energy is dissipated as heat in the cell. Crystal defects, accidental impurities, and surfaces are all strong nonradiative recombination sites. Most nonradiative recombination processes are, in principle, avoidable, but radiative recombination is fundamental.

The result of these processes is a light-generated flow of current in what is normally thought of as the

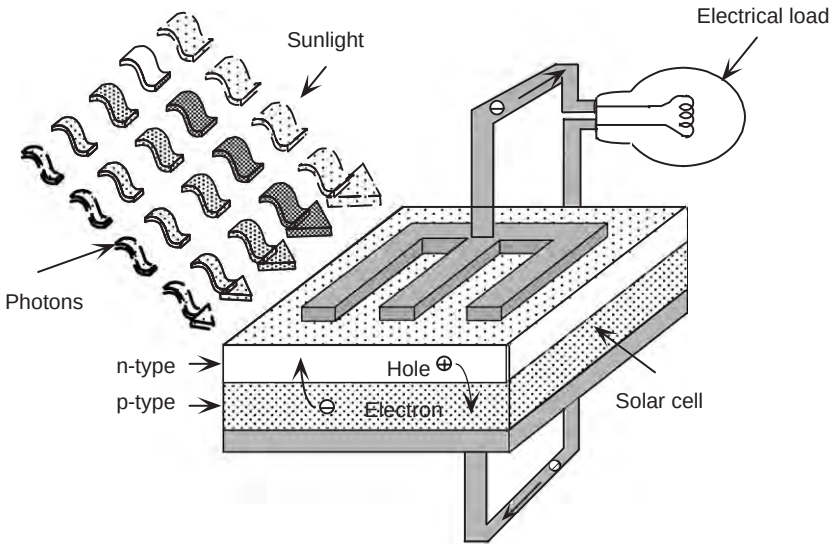


FIGURE 3 Solar cell operation. Adapted from M. A. Green and J. Hansen (2003). *Catalog of Solar Cell Drawings*. University of New South Wales.

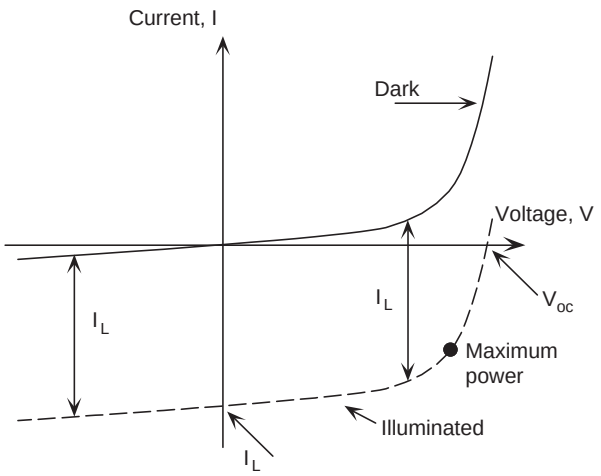


FIGURE 4 Typical current-voltage characteristic. Adapted from M. A. Green and J. Hansen (2003). *Catalog of Solar Cell Drawings*. University of New South Wales.

reverse direction in diode theory (i.e., electrons flow out of the cell into the circuit from the n-type contact and back into the cell through the p-type contact) even though the voltage across it is in the forward bias direction (i.e., positive at the p-type contact). The light-generated current is approximately independent of the voltage across the cell. Unfortunately, the normal diode current, which would remain even in the dark at the same contact voltage, may be thought of as simultaneously flowing in the opposite direction to the light-generated current. That dark current (solid line in Fig. 4) increases strongly with the voltage and at some voltage, known as the open circuit voltage (V_{oc}), completely cancels the light-generated current. As a result, we see the typical current-voltage characteristic shown as a dashed line in Fig. 4. The (forward) dark current is added to the (reverse) light-generated current to make the resultant current negative for a range of positive voltage. In that range, the cell generates power. Power is zero at zero volts and at V_{oc} with maximum power being produced near the so-called knee of the curve. The electrical load connected to the cell should be chosen to keep the operating point close to the optimum knee point during normal operation.

3. EFFICIENCY

In that brief description of cell operation we considered the main processes occurring after photons enter into the semiconductor volume. However, reflection from the surface must first be reduced to allow efficient absorption. A polished semiconductor surface will reflect a significant fraction of incident photons from the sun, around 30% for silicon. Additional antireflection layers of other materials can be applied to the front surface. In addition, texturing of the front surface can assist with reflection reduction by deflecting rays of light so that they strike the semiconductor surface again. Texturing also redirects photons into oblique paths within the cell, giving longer paths for weakly absorbed (red) photons and more opportunity for absorption (light trapping).

The bandgap energy of the semiconductor from which a cell is made sets the fundamental upper limit of the cell's conversion efficiency. The choice of a low bandgap material allows the threshold energy to be exceeded by a large fraction of the photons in sunlight, allowing a potentially high current. However, a solar cell can extract from each photon only an amount of energy slightly smaller than the bandgap energy, with the rest lost as heat. On the other hand, a semiconductor is transparent to photons with energy less than its bandgap and cannot capture their energy. These, thermalization and transparency, are two of the largest loss mechanisms in conventional cells. The best solar cell materials have bandgaps giving a compromise between these two effects. Solar cell performance is limited by the laws of thermodynamics. Thermodynamic methods and the related principle of detailed balance can be used to calculate the limits. Shockley and Queisser showed in 1961, by use of the balance between incident and escaping photons and extracted electrons, that the efficiency limit for a single-material cell is around 31% for an optimal bandgap of around 1.1 eV. That work assumed that the only unavoidable losses from the cell are the emission of photons produced by radiative recombination.

For such limit calculations, the sun is commonly assumed to behave as a thermodynamic "black body," or perfect radiator at 6000 K, although this is only a rough approximation to the solar spectrum at the earth. The cell is usually assumed to be another perfect radiator at 300 K (27°C). Other losses are assumed to be, in principle, avoidable or controllable and are neglected. Within that framework, a range of different solar cell concepts produces a range of efficiency limits (Table I) from 31% to 86.8% for maximum concentration, which refers to focusing only the solar disk onto the cell, with a concentration of around 46,000 times.

The Carnot limit for solar conversion under the preceding temperatures is 95%, but this could be achieved only if the cell were to produce infinitesimal

TABLE I

Efficiency Limits for Single and Multiple Bandgap Solar Cells for Unconcentrated Sunlight and for Sunlight Concentrated Maximally[a]

Number of bandgaps	Unconcentrated light	Maximal concentration
1	31.0%	40.8%
2	42.9%	55.7
3	49.3%	63.9%
∞	68.2%	86.8%

[a] Only the solar disk is seen by the cell and all emissions from the cell are directed back toward the solar disk. The sun is assumed to be a thermodynamic black body radiator with absolute temperature 20 times that of the cell (e.g., cell and sun temperatures of 300 K and 6000 K). From De Vos (1992).

electrical output since the cell would need to emit light at the same spectrum and intensity as the sun in order to avoid entropy production in the energy transfer. The Landsberg efficiency for direct sunlight conversion (93.3%) is the ultimate solar conversion efficiency and results from the assumption that the cell generates no entropy during absorption and emission of light but is lower than the Carnot limit because the light emitted by a cell is considered lost, not returned to the sun for later use. Inclusion of those additional entropy production mechanisms yields, for a multibandgap converter with an infinite number of bandgaps, the 86.8% value in Table I.

4. TYPES OF SOLAR CELLS

4.1 Silicon Cells

Silicon enjoys the advantages of being nontoxic and of great abundance, and its native oxide, silicon dioxide, is easily able to passivate the surfaces to improve the electronic properties there. Silicon solar cell research takes advantage of the vast body of research carried out for the silicon-based electronics industry and, until recently, off-specification silicon rejected by the electronics industry supplied most of the feedstock for the solar cell industry. In addition, useful chemical techniques, such as directionally selective etches, are available for monocrystalline silicon processing. As a result, silicon has dominated the solar cell industry since 1954 and continues to do so. Silicon is used in monocrystalline, multicrystalline, polycrystalline thin-film, and amorphous forms for different types of cells, with the first two supplying most of the market at present.

4.1.1 Silicon Wafer Cells

4.1.1.1 Monocrystalline Silicon Large (up to 300 mm diameter), cylindrical ingots of extremely pure, single-crystal silicon are grown from molten silicon. The entire ingot is doped, usually p-type with boron, during the melt phase and is sawn into circular wafers less than 0.5 mm thick, from which solar cells are made. According to *Photon International*, such monocrystalline silicon cells supplied 36% of the terrestrial PV market in 2002.

Different manufacturers use variations of the same basic process for making screen-printed cells (Fig. 5). Wafer surfaces are etched to form textured surfaces for antireflection and light trapping. Light trapping is implemented to increase the path length within the cell volume for weakly absorbed light at the red end of the spectrum. An n-type dopant, usually phosphorus, is then diffused into the surface to form the p-n junction, producing the in-built electric field to separate the charge carriers (Fig. 3). Heavily doped layers are formed at each surface to aid electrical contact and the metal contacts are formed by the screen printing of a metallic paste before firing. Optical coatings may be added to the front surface to further reduce reflection.

Production cells, using the normal boron doped, solar-grade silicon wafers grown by the Czochralski technique, have efficiencies of typically 12 to 15% although elevated temperatures during operation lead to reduced output power, characterized by a temperature coefficient in the range -0.4 to $-0.5\%/^{\circ}C$. Much higher efficiency, up to 24.7%, has been achieved in laboratory cells that use expensive photolithographic processes and purer float-zone wafers in their production, but such cells are only commercially viable for high-value applications.

4.1.1.2 Buried and Point Contact Cells An alternative production sequence, used commercially by BP Solar at their Spanish factory, is the buried contact solar cell (Fig. 6). Front contacts are plated into grooves in the cell formed by laser or mechanical grooving, allowing lower metallic shading of the surface while still permitting sufficient metal cross section to have low electrical resistance. BP Solar have reported that processing costs per unit area are only slightly higher than those of the screen-printed cell, but efficiency is significantly higher. These cells may be made from either monocrystalline or multicrystalline silicon.

Sunpower (USA) produce point contact silicon cells with front surfaces completely free from metal

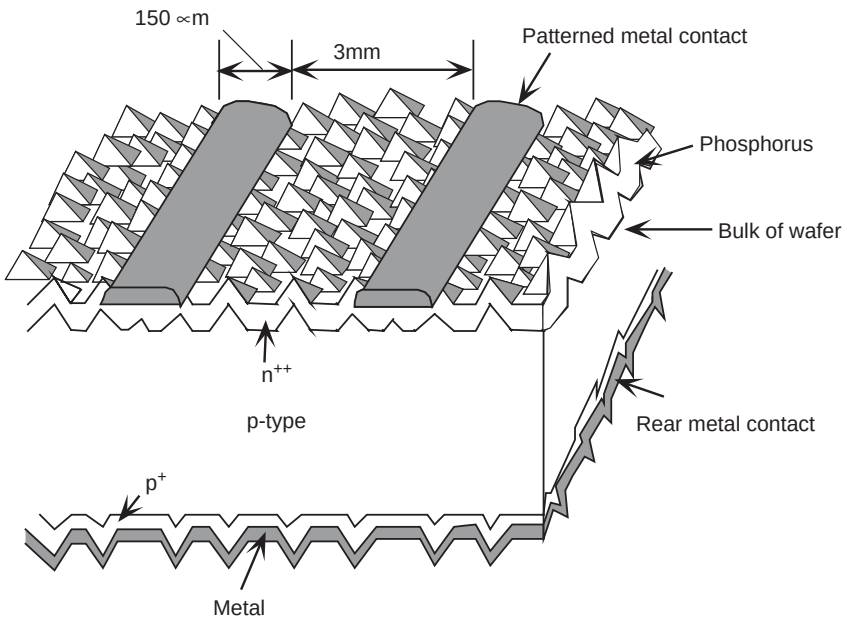


FIGURE 5 Structure of a typical commercial cell with textured surface and screen-printed contacts. From M. A. Green and J. Hansen (2003). *Catalog of Solar Cell Drawings.* University of New South Wales.

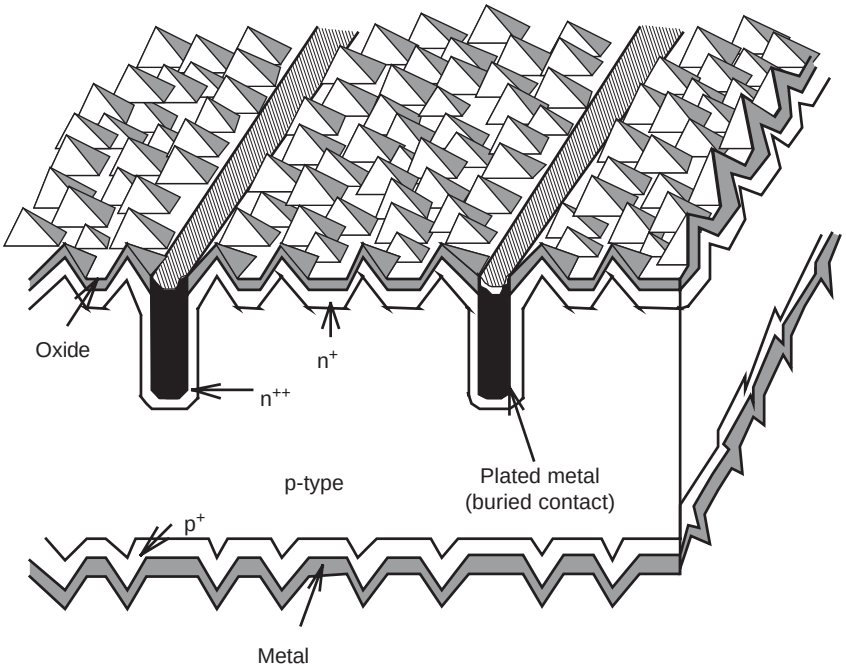


FIGURE 6 Structure of buried-contact silicon cell. From M. A. Green and J. Hansen (2003). *Catalog of Solar Cell Drawings*. University of New South Wales.

contacts. Both the n- and p-type contacts are at the rear. These cells were originally developed for concentrator applications but are predicted to be available in flat plate modules of nominally 20% efficiency from 2004.

4.1.1.3 Semitransparent Monocrystalline Silicon Different approaches have been taken toward making photovoltaic products that are aesthetically attractive for architectural applications. Semitransparency has been achieved with monocrystalline

silicon by, for example, Sunways in Germany. The cells are laser scribed through half the thickness in multiple parallel lines from the front and scribed again orthogonally from the rear. A mesh of holes through the cell is thereby created at the intersections of the lines. The holes and the additional surface area reduce their efficiency, but 9.0% cell and 7.4% module efficiencies are still achieved. Australian National University announced late in 2002 the development and imminent commercialization of a new monocrystalline silicon cell process that uses tiny interconnected cells about 1 mm wide, cut from 1-mm-thick wafers, with light trapping between lower and upper glass sheets.

4.1.1.4 HIT Sanyo's HIT (heterojunction with intrinsic thin-layer) cell, shown schematically in Fig. 7, combines crystalline and amorphous silicon (amorphous silicon is discussed below). n-type crystalline silicon wafers have layers of intrinsic and doped amorphous silicon applied to the front and rear. The junction has intrinsic amorphous layers between the n-type crystalline and p-type amorphous silicon regions to reduce recombination. Contacts are made by metal electrodes and layers of transparent conducting oxide (TCO). Unwelcome absorption in the TCO and electrically inactive doped amorphous layers is compensated by the excellent passivation of the crystalline silicon by the amorphous. Their negative temperature coefficient ($-0.33\%/^{\circ}C$) is smaller than that of conventional silicon cells. Large-area laboratory cell efficiency of 21.3% has been reported, and Sanyo's nominal cell production efficiency is up to 19.5%. Sanyo's HIT200 module has nominal cell and module efficiencies of 19.5% and 17%, respectively. Sanyo produced about 16 MW_p during 2001 and reportedly intends to increase production from to 120 MW_p per annum by 2005.

4.1.1.5 Cast Multicrystalline Silicon Cheaper and less energy-intensive wafers are produced by casting (cooling molten silicon in a mould) and sawing. These materials are produced specifically for PV production and have the advantage that the usually square wafers make cells that pack together more closely in modules, leaving less unproductive space between. The resulting material is of lower quality due to the poor electronic qualities of the boundaries between the randomly oriented silicon grains (crystals). Cast and deposited silicon is denoted as multicrystalline, polycrystalline, microcrystalline or nanocrystalline, depending on the average sizes of the crystals. An additional cause of reduced cell performance is the difficulty in texturing such material, and commercial multicrystalline cells are commonly untextured. The use of a passivating/antireflection coating such as silicon nitride is therefore common. Otherwise, processing is similar to that described earlier for monocrystalline silicon cells, and multicrytalline material is used for both screen-printed and buried-contact cells. In 2002, 52% of global production was of cast multicrystalline cells.

4.1.1.6 Ribbon and Sheet Silicon Various methods are used to directly produce silicon sheets of thickness suitable for production of self-supporting

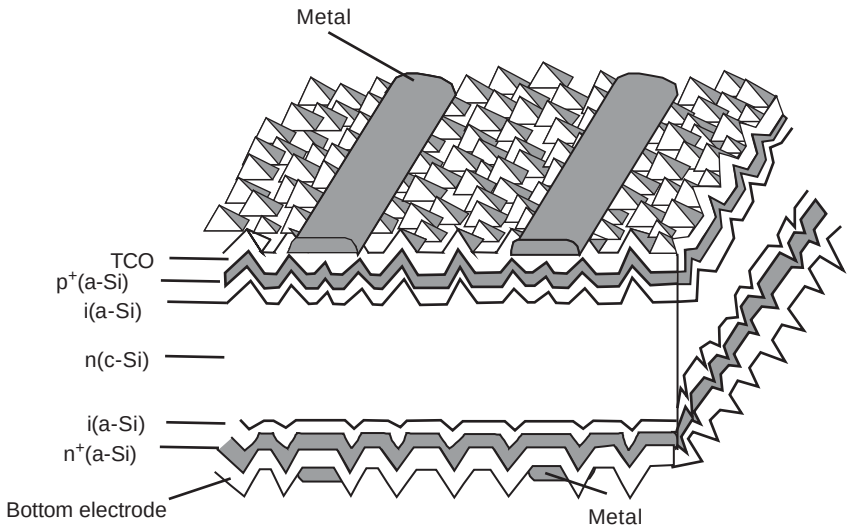


FIGURE 7 Structure of HIT cell. From M. A. Green and J. Hansen (2003). *Catalog of Solar Cell Drawings*. University of New South Wales.

cells without the wasteful sawing process. Edge-defined film growth (EFG) cells and modules are produced by RWE Schott Solar in the United States. A large, hollow nonagon of silicon is drawn through a graphite dye, and wafers are cut from each side. The material is multicrystalline with large, elongated grains and the resulting cells have efficiency exceeding 14% in production and 15 to 16% for best cases. Evergreen Solar grow cell material by a so-called string ribbon process, achieving 15 to 16% best-cell efficiency at standard temperature. Molten silicon forms a film between two carbon strings as they are withdrawn from the crucible. The company was reported in 2003 to be seeking investment to double capacity to 14 MW_p per year. Alternative processes, dendritic web and RGS silicon foil are also proceeding toward pilot production. Ribbon or sheet crystalline silicon cells accounted for 4.7% of world production in 2002, reduced from 5.6% in 2001.

4.1.2 Detached Silicon

Several approaches have been developed to grow thin crystalline silicon films on substrates from which they are detached and the substrates reused. These are aimed at greatly reducing the silicon used, saving a fraction of the necessary investment of energy and money. One, developed by Mitsubishi, involves deposition and laser-crystallization of silicon on an oxidized silicon wafer after holes have been etched through the oxide. More silicon is then grown epitaxially to increase the layer thickness prior to detachment. Another approach is to form a porous silicon layer on a wafer and use that as a crystalline template for growth of a silicon layer before detaching at the porous layer.

4.1.3 Spheral Silicon Cells

Spheral cells are currently being commercialized by Spheral Solar Power (Canada) in partnership with Photowatt (France) and by Kyosemi Corp. (Japan). Small silicon spheres, each a tiny cell of 0.5 to 2 mm diameter, are interconnected in sheets. This results in a flexible and lightweight cell structure, which uses significantly less silicon than conventional wafer cells.

4.1.4 Amorphous Thin-Film Silicon and Silicon-Germanium Cells

In amorphous silicon (a-Si), the silicon atoms are arranged more randomly than in the strict geometric order found in crystals. Many atoms have bonds that are unsatisfied by neighboring silicon atoms so hydrogen is usually incorporated into the material to reduce recombination, forming a-Si:H. Its absorption coefficient is higher than that of crystalline silicon, allowing it to absorb more strongly and to require lower film thickness ($\sim 0.4\,\mu m$). Relative to wafer-based silicon, thin-film (second-generation) amorphous cells have the potential for lower cost per unit area but also lower efficiency.

However, a light-induced degradation effect has severely restricted the early promise of this technology. Degradation has been minimized by making the cells as thin as possible, and textured TCO surfaces have been used to recover the absorptivity lost by thinning. Germanium is sometimes alloyed with the silicon to modify the bandgap energy since amorphous silicon's bandgap of around 1.7 eV is higher than optimal.

a-Si:H is usually applied by plasma-enhanced chemical vapor deposition. Amorphous cells use a p-i-n structure, with a layer of intrinsic (undoped) material between the n- and p-doped layers to take advantage of the better properties of intrinsic material. In the p-i-n design the junction electric field extends across the intrinsic region, permitting more thickness for generation and collection within the field region. Up to three junctions can be stacked in series for higher output, and triple-junction tandem modules (discussed later) are available commercially. There are two main motivations for this structure: first, the very thin layers required by stability considerations are too thin to absorb all the sunlight; second, it allows the opportunity to have a different bandgap in each junction by alloying a different fraction of germanium in each, giving a better match to the solar spectrum.

Amorphous modules have only a small negative temperature coefficient of power ($-0.1\%/^\circ C$), making them less subject to output reduction due to heat, and they are more tolerant of reduced intensity and diffuse light. The relatively low-temperature deposition technology allows plastics to be used as substrates in flexible modules. They also lend themselves to use in semitransparent building products and car sunroofs. Amorphous silicon's share of production has been steadily declining, falling to 6.4% in 2002. Best reported stabilized efficiency for an amorphous module is 10.4% for a triple-junction alloy device, and commercial triple-junction modules are around 6.3% efficient.

4.1.5 Micro- and Polycrystalline Thin-Film Silicon Cells

Various methods are being investigated to develop thin-film silicon cells deposited on foreign substrates.

If the ratio of hydrogen to silane in the gas from which amorphous silicon is deposited is increased, the resulting material becomes microcrystalline, with columns of crystallites separated by amorphous regions. The optical and electronic properties are similar to those of bulk silicon. Such material has been used as an alternative to silicon/germanium alloys in hybrid structures with amorphous silicon by Kaneka (Japan), who report stable 10% efficiency in a 39W module and who also manufacture amorphous silicon modules. The hybrid modules are suggested to be more stable than amorphous, but that has not yet been independently verified. Particular measures are necessary to allow the amorphous layers to be kept thin enough to avoid light-induced degradation while producing similar current to the microcrystalline cell(s) in series. The University of Neuchâtel, Switzerland, has developed a microcrystalline/amorphous tandem design with around 11% efficiency on a laboratory scale.

Astropower was first to commercially produce thin-films cells with millimetre-sized grains on foreign substrates. Small area cell efficiency of 16.6% has been demonstrated on an unspecified substrate and 9% on a proprietary ceramic. However, the product has evolved to films deposited on a silicon buffer layer, now resembling silicon ribbon cells. Pacific Solar is approaching commercial production with a process in which a thin-film silicon cell is formed on a textured glass superstrate. A laser is used to form craters through the active material to contact the n-type layer closest to the glass. Low-quality material is deposited then improved by subsequent thermal steps. The best efficiency achieved at the time of writing is 8.2% for 30 cm by 40 cm modules, and work is in progress to achieve the same performance from larger modules.

4.2 Thin-Film Cadmium-Telluride Cells

Cadmium telluride (CdTe) thin-film cell efficiency of 16.5% has been achieved, and the record module efficiency currently stands at 10.6%. CdTe comprised 0.7% of global cell production in 2002, and most of that was for indoor use in consumer products. Commercial modules are generally less than 7% efficient. The basic structure has a glass superstrate and a layer of transparent conducting oxide (TCO) as front contact, near-transparent *n*-type cadmium sulphide (CdS) window layer, *p*-type CdTe, and metallic rear contact. The contact films are scribed to divide the module into cells, which are series-connected to produce the required voltage.

Potential restrictions on market growth are the toxicity of cadmium and limited global supply of tellurium. Unencapsulated cells are sensitive to moisture. The toxicity issue is controversial and affects public acceptance of the technology. Relatively small-scale production has begun in Japan, the United States, and Germany but production by major players in Germany and the United States ceased in late 2002.

4.3 Thin-Film Chalcopyrite Cells

Copper indium gallium diselenide ($Cu(In,Ga)Se_2$, abbreviated to CIGS, or to CIS if the gallium is absent, is another promising thin-film PV material, although it is still only in small-scale production. Laboratory efficiencies have been difficult to transfer to reliable mass production of high-efficiency large modules. Large grains can be produced on foreign substrates, grain boundaries can be passivated, and the material is tolerant of deviations from exact ratios of the constituent elements. Problems include the need to develop specific deposition equipment, the great complexity of co-depositing so many elements, process control, inhomogeneity across large areas, and moisture sensitivity of unencapsulated cells. Global supplies of indium may be a restriction on widespread use in the long term. Toxicity of the cadmium in the window layer may possibly be avoided by use of alternative compounds such as In(OH, S), although CdS appears to be most commonly used.

A molybdenum rear contact is deposited on a soda-lime glass substrate and patterned by laser to divide into cells. The elements of the CIGS absorber layer are evaporated simultaneously before a heterojunction is formed with a cadmium sulphide layer. A TCO front electrode, mechanical scribing, contact formation, and encapsulation complete the module. One manufacturer expects to achieve better than 8% efficiency in mass production of modules.

Confirmed best cell and module efficiencies are currently 18.4% and 13.4%, respectively. Commercial production in 2002 amounted to just 0.2% of world solar cells. Flexible modules for portable, personal power provision is one high-value market being targeted by CIGS manufacturers.

4.4 Group III-V Cells

Compounds of elements from Groups III and V of the Periodic Table, exemplified by gallium arsenide (GaAs), are used extensively in the optoelectronics

and electronics industries and may be used to make excellent but expensive solar cells

4.4.1 Single Junction III-V Cells

High-efficiency solar cells are produced almost entirely for space applications where low weight, high efficiency, and hardness to cosmic radiation are more important than price. Active semiconductor layers are deposited by a variety of techniques on inactive substrates. High-efficiency single-junction cells have been made from GaAs, GaInAsP, and InP. High-bandgap AlGaAs windows are commonly used to reduce surface recombination.

Gallium and indium are not abundant elements, implying high cost and a limited ability to supply a large fraction of the global energy needs in terrestrial applications, except possibly with high concentration. Best terrestrial cell efficiency is currently 27.6% for a GaAs cell under 255 times concentration.

4.4.2 Tandem and Multijunction III-V Cells

Tandem cells are one approach to exceeding the efficiency limits of single-material cells. They reduce the two main losses mentioned earlier, the thermalization of the excess energy of high-energy photons and transparency to low energy photons. Tandem cells are stacks of *p-n* junctions, each of which is formed from a semiconductor of different bandgap energy. Each responds to a different section of the solar spectrum, yielding higher overall efficiency. By stacking the component cells in order of decreasing bandgap, as shown in Fig. 8, the light is automatically filtered as each cell extracts photons which exceed its bandgap.

The technique is applied commercially in triple-junction amorphous cells (noted earlier) and in crystalline III-V cells, but using the method with crystalline materials in high-efficiency devices presents several challenges. The materials used in the various junctions must be compatible with each other in several ways over a wide temperature range (defined by the fabrication temperatures), and this severely restricts the options. In particular, the thermal expansion coefficients should be similar and the bandgaps should be a good match to the spectrum. For monolithic devices, the crystal lattice constants must be similar to permit low-defect epitaxial growth.

Tandem cells with up to three bandgaps are commercially available, but they are used almost exclusively for space applications, especially the powering of satellites. There have long been plans for extensive terrestrial use in solar concentrators

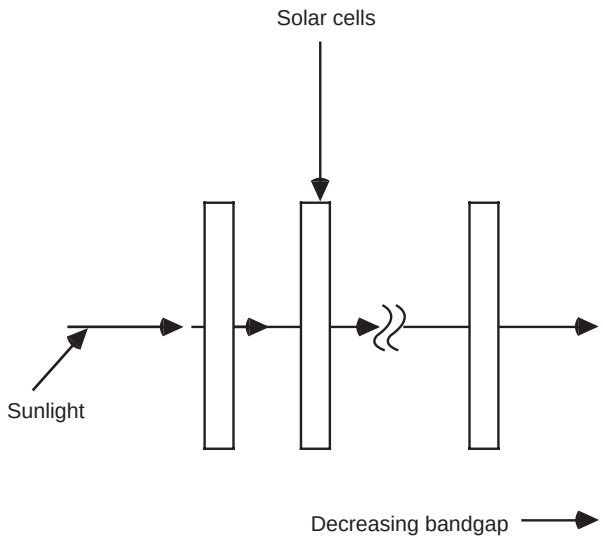


FIGURE 8 Arrangement of bandgaps in a tandem cell. From M. A. Green and J. Hansen (2003). *Catalog of Solar Cell Drawings.* University of New South Wales.

(discussed later), where the concentration of the light from a large collection area onto a cell may be able to justify such expense, but this has not yet occurred to a significant extent. The overall record cell efficiency of 35.2% is for a triple stacked GaInP/GaAs/Ge tandem cell operating under 66-times concentration. Work in Germany during 2003 achieved the first five-junction tandem cell.

4.5 Dye-Sensitized Cells

Dye-sensitized nanocrystalline cells are fundamentally different from the others discussed here in that they do not rely on semiconductor *p-n* junctions. Instead, they are electrochemical devices in which the optical absorption and carrier-collection processes are separate. A porous film of a wide-bandgap semiconductor, usually titanium dioxide (TiO_2), is coated with a redox charge-transfer dye (Fig. 9). The dye is excited by the absorption of a photon and an electron is injected into excited energy level of the oxide, and the dye is regenerated by a liquid electrolyte. The excited electron in the oxide reaches an electrode and is then able to migrate through an external circuit, do work in the load, and pass to the other electrode, where it regenerates the electrolyte. The front contact is formed from a transparent conducting oxide.

The best-demonstrated efficiency for a laboratory device is 8.2%, and reasonable stability under illumination has been demonstrated. Dye-sensitized cells coming into small-scale commercial production

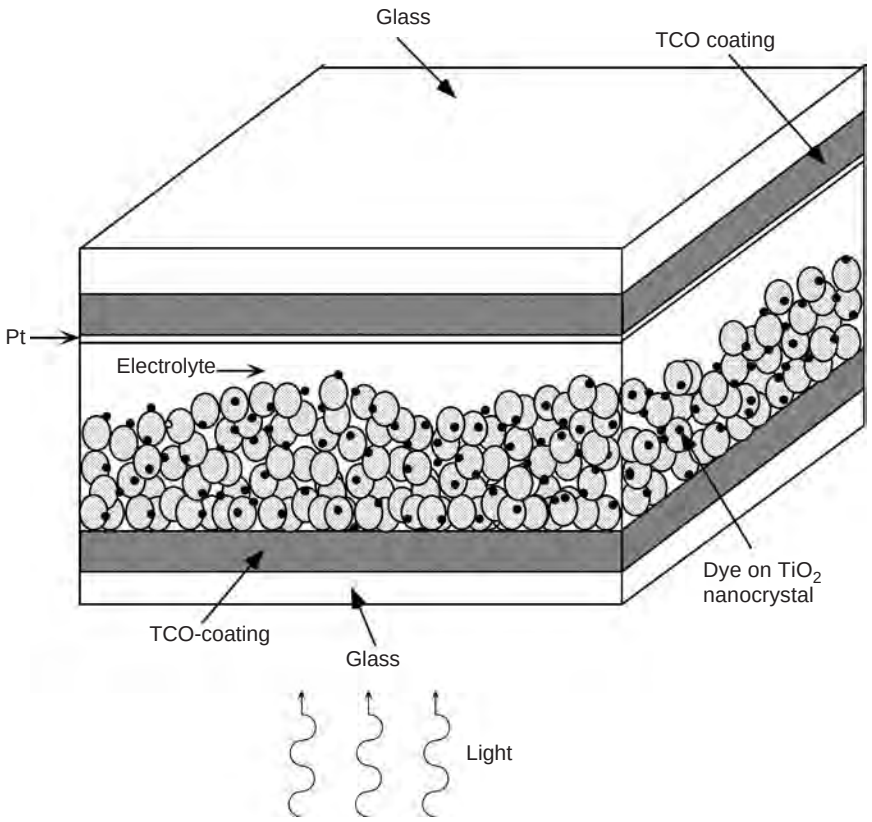


FIGURE 9 Dye-sensitized nanocrystalline cell structure. From M. A. Green and J. Hansen (2003). *Catalog of Solar Cell Drawings*. University of New South Wales.

in Australia, and the installation of the first wall panel has been announced. Dye-sensitized cells exhibit promise as potentially low-cost cells, in terms of finance and energy. Their temperature coefficient can be positive, giving increased power at elevated operating temperatures. Main concerns are to improve high temperature stability and to scale up production. A solid electrolyte would be preferable.

Tandem dye-sensitized cells are aimed particularly at the direct splitting of water for hydrogen production rather than at electricity production. In one experimental system, yielding 4.5% conversion efficiency from sunlight to chemical energy, two semiconductors, TiO_2 and WO_3 or Fe_2O_3, are used with the water being the electrolyte.

4.6 Concentrator Cells

Sunlight may be optically concentrated to reduce the cell area and reduce costs. Efficiency increases with concentration but at the expense of acceptance of only the direct fraction of the sunlight. Hence, the solar resource is reduced, especially in cloudy or coastal regions. In addition, the concentrating optics are always less than 100% efficient. Concentrators, other than static concentrators, must track the sun across the sky. They often take the form of Fresnel lenses (point or line focus) or paraboloidal dish or trough reflectors.

Cells can be of either silicon or III-V materials and several commercial products exist, although their undoubted promise has not yet been fulfilled for a variety of reasons. Only a tiny fraction of global production of cells is specifically for concentrators, and one prominent producer of silicon concentrator cells began in 2003 to manufacture flat-plate modules instead. Special care must be taken in the design of concentrator cells, particularly with respect to optics, heating, and high currents. The additional complexity is justified by the reduced cost fraction of the cells in a concentrating system. The heat production in concentrator cells becomes a benefit if it can be economically extracted and there exists a nearby market for low-grade thermal energy. A few combined photovoltaic/thermal systems have been developed.

Few modifications are needed for standard screen-printed silicon cells to operate at low concentration

up to about 10 times. Buried contact silicon cells, for example, operate satisfactorily up to around 30 times concentration, but special cells are required for higher concentration.

4.7 Innovative Solar Cells

In addition to the commercial or near commercial solar cell technologies described earlier, there is a wide array of further ideas for innovative types of cells, aimed at increasing efficiencies or reducing costs or both. Only some of the most prominent, none of which are fully developed and some of which are as yet only ideas, will be mentioned here.

Quantum well cells remain controversial since their proposal in the early 1990s. Multilayer sandwiches of extremely thin alternating layers of high- and low-bandgap semiconductors are held in the static electric field between p and n regions on either side. The layers are thin enough that quantum effects become important, modifying the material properties. The basic intention is that carrier pairs created in the low-bandgap regions will gain sufficient energy optically or thermally to permit their extraction at the higher voltage that is characteristic of the high-bandgap material. Quantum dot cells, in which the quantum effects are evident in three spatial directions instead of one, are a development that promises reduced opportunities for carriers to dissipate their additional energy as heat.

Attempts are being made to apply the impurity photovoltaic and impurity band effects in devices that would be able to utilize two or more photons in multistep absorptions via impurities that introduce energy states within the normally forbidden bandgap. These states also introduce additional recombination pathways, but calculations have shown that in the optimal situation, very high efficiencies would be possible. Making an electrically separate optical upconverter or downconverter by putting the impurities in a separate pieces of material instead of in the cell itself would avoid those recombination paths for some carriers. An all-silicon tandem cell is being pursued through the modification of the silicon bandgap in the upper subcell by forming superlattices of alternating thin silicon layers with layers of silicon dioxide or nitride.

If carriers due to high-energy photons could be collected while still energetic or "hot," the thermalization loss could be avoided and that is the aim of researchers investigating hot carrier cells. Carriers either have to reach the contacts rapidly or thermalization must be slowed. In addition, development of contacts that select the energy of the collected carriers is required.

All the approaches mentioned so far in this section are intended to produce very high efficiencies. Organic semiconductor solar cells, on the other hand, are not expected to be as efficient as present commercial cells but are expected to be produced extremely cheaply, at relatively low temperatures over large areas of cheap substrates. These cells take several forms, including interpenetrating networks of two polymers that are chosen to form a heterojunction that separates electrons and holes from each other.

5. CONCLUSIONS

Solar cells technology is evolving through three generations, with the first, comprising cells made from relatively thick (100s of μm thick) wafers of monocrystalline or multicrystalline silicon, continuing to dominate the industry. Silicon has several advantages over competing materials, but ultra-pure semiconductor wafers are inherently expensive in terms of both finance and energy. The second-generation cells—based on thin-films of amorphous, nanocrystalline, or polycrystalline semiconductors, deposited on cheap substrates or superstrates for mechanical support—are entering the market, or, as in the case of a-Si, are already established with a small market share. Second-generation cells may eventually be available at significantly lower cost per unit of delivered energy than first-generation cells, but a range of scientific, materials, production, and environmental issues have hampered the market penetration of the thin-film cells. When their market matures, it is expected that material input costs will eventually dominate their price. However, the production share of thin-film cells is currently declining, while the established first-generation approaches continue to be produced more cheaply due to enhanced manufacturing scale and technical improvements. Silicon may come to dominate the second generation of solar cell production too. Research is already under way into a wide range of attempts to combine the attributes of very high efficiency and low-cost production.

The global market, and particularly in Japan and Europe, is growing rapidly, fuelled by energy security and environmental concerns. Solar cells are complementing wind and waste-derived biomass energy in satisfying the growing demand for energy with low environmental impact. After several false starts, solar cells are finally coming of age.

Acknowledgments

The Centre of Excellence for Advanced Silicon Photovoltaics and Photonics is supported by the Australian Research Council's Centres of Excellence Scheme. The author thanks M. Green and J. Hansen for permission to use Figs. 2 through 9 and to M. Green for helpful comments.

SEE ALSO THE FOLLOWING ARTICLES

Materials for Solar Energy • Photovoltaic Conversion: Space Applications • Photovoltaic Energy: Stand-Alone and Grid-Connected Systems • Photovoltaics, Environmental Impact of • Solar Cookers • Solar Cooling, Dehumidification, and Air-Conditioning • Solar Detoxification and Disinfection • Solar Distillation and Drying • Solar Energy, History of • Solar Fuels and Materials • Solar Heat Pumps • Solar Ponds • Solar Thermal Energy, Industrial Heat Applications • Solar Thermal Power Generation

Further Reading

Archer, M.D., and Hill, R. (eds.). (2001). Clean Electricity from Photovoltaics. Imperial College Press, London.

Brendel, R. (2003). "Thin-Film Crystalline Silicon Solar Cells: Physics and Technology." Wiley-VCH, Berlin.

De Vos, A. (1992). "Endoreversible Thermodynamics of Solar Energy Conversion." Oxford Science, Oxford.

Grätzel, M. (2001). Photoelectrochemical cells. *Nature* **414,** 338–344.

Green, M. A. (1986). "Solar Cells: Operating Principles, Technology and System Applications." University of New South Wales, Sydney.

Green, M. A. (1995). "Silicon Solar Cells: Advanced Principles and Practice, Centre for Photovoltaic Devices and Systems." University of New South Wales, Sydney.

Green, M. A. (2003). "Third Generation Photovoltaics: Ultra-High Efficiency at Low Cost." Springer-Verlag, New York.

Green, M. A. (in press). "Thin-Film Photovoltaics: Advances in Solar Energy" (K. W. Böer, Ed.). Amer. Solar Energy Soc., Boulder, CO.

Luque, A., and Hegedus, S. (eds.). (2003). "Handbook of Photovoltaic Science and Engineering." Wiley, London.

Markvart, T., and Castañer, L. (eds.). (2003). "Practical Handbook of Photovoltaics". Elsevier, Oxford.

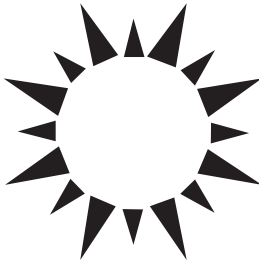
Shah, A. V., Meier, J., Vallat-Sauvain, E., Wyrsch, N., Kroll, U., Droz, C., and Graf, U. (2003). Material and solar cell research in microcrystalline silicon. *Solar Energy Mat. Solar Cells* **78,** 469–491.

Tawada, Y., Yamagishi, H., and Yamamoto, Y. (2003). Mass production of thin film silicon PV modules. *Solar Energy Mat. Solar Cells* **78,** 647–662.

Zweibel, K. (1990). "Harnessing Solar Power: The Photovoltaics Challenge." Plenum Press, New York.

Solar Cookers

S. K. SHARMA
Energy Research Centre
Chandigarh, India



1. **Introduction**
2. **Classification of Solar Cookers**
3. **Parabolic Concentrator Cookers**
4. **Fresnel Reflector Cookers**
5. **Plane Mirror Concentrating Cookers**
6. **Transmitting Concentrator Cookers**
7. **Fresnel Lens Cookers**
8. **Principles of Paraboloid Cookers**
9. **Solar Box Cookers**
10. **Multiple Glass Box Cookers**
11. **Indirect Solar Cookers**
12. **Scheffler Community Cooker**
13. **Rating of Solar Cookers**

Glossary

chemical energy Energy stored in a substance and released when the substance changes form or combines with another substance.
compound parabolic concentrator (CPC) A device that consists of two curved reflecting segments, which are part of parabolas.
concave Surface curved like the interior of a circle or a sphere.
energy Capacity, or ability, to do work.
flux concentration Density of radiation falling on and/or received by a surface or body.
parabola Plane curve formed by the intersection of a cone with a plane that runs parallel to its side.
paraboloid A solid, some of whose plane sections are parabolas.
pyranometer An instrument that measures solar radiation.
radiant energy Energy received in the form of photons or electromagnetic radiation.
rim angle An angle measured at the focus from the axis to the rim or a point where a paraboloid has been truncated.
solar concentrator A device that focuses the solar radiation received on the receiver.
solarimeter An instrument that measures solar radiation.
solar insolation Intensity of solar radiation.
solar radiation A form of electromagnetic radiation.
specific heat capacity The amount of heat energy absorbed or lost by a substance.
thermal energy Also known as heat; the kinetic energy associated with rapid random motion.

1. INTRODUCTION

Cooking is a very important and essential chore in every household. Unfortunately, this activity is also a major cause of deterioration of forest cover and soil fertility. It is also responsible for drudgery and ill health among women in developing countries. The main reason is the use of traditional biomass fuels, such as wood, cow dung, and agricultural residues, in inefficient cookstoves, in poorly ventilated kitchens. It has been estimated that there are 180 polar, 75 aliphatic, and 225 aromatic hydrocarbon compounds in smoke produced from biomass. The exposure of women and children to these carcinogenic, ciliotoxic, and mucus-coagulating agents results in numerous diseases related to the respiratory system, such as colds in adults and acute respiratory infections in infants and young children.

It has been estimated that the use of solar cookers by 2000 families for 10 years would mean a savings of $0.1 million in terms of reduced firewood consumption, $2 million worth of coal, or $0.25 million worth of kerosene oil.

1.1 History

Cooking with energy from the sun is an old concept. There are many landmarks in the history of the design of solar cookers.

 © 2004 Elsevier Inc. All rights reserved.

The first attempt to cook fruits was made by Swiss scientist Horace de Saussure in 1767. He achieved 88°C in a hot box-type cooker. In 1837, astronomer Sir John Frederick Herschel achieved 116°C in his solar cooker and cooked meat and vegetables. Muguste Mouchot designed a community cooker for baking bread and potatoes for French troops in 1870. The first manual on solar cookers, entitled "Solar Heat: A Substitute for Fuel in Tropical Countries," was written by William Adam in 1878 in India.

Maria Telkes and George O. G. Lof performed extensive studies on solar cooking at the Massachusetts Institute of Technology.

A significant attempt to introduce nearly 200 of the concentrator-type solar cookers developed at the University of Wisconsin was made in the states of Sonora, Coahuila, and Oaxaca, Mexico, with the help of the Rockefeller Foundation. After initial problems, the cooker was redesigned and the modified version of the cookers was used for many years.

One of the largest attempts to introduce solar cookers was made in India in 1953, when 200,000 paraboloid concentrator-type solar cookers developed by M. L. Gai of NPL New Delhi were brought in. The cooker consisted of a paraboloid reflector of equation $Y^2 = 180X$, with a focal length of 45 cm; the reflector was spun from an aluminum sheet into the desired shape and then anodized to protect it from the weather and to maintain reflectivity. The first commercial version of the hot box-type cooker was designed by M. K. Ghosh in India in 1945.

1.2 Cooking Methods and Solar Cookers

The principal methods of cooking food can be divided into four categories based on cooking temperature and cooking media. These are boiling, frying, baking, and roasting.

1.2.1 Boiling

Water is used as a medium in this type of cooking. This sets the upper temperature limit. In this mode of cooking, food is cooked at only 100°C. In cooking by boiling, the mixture of food and water is brought to a boil and allowed to simmer until completion of the chemical reaction of the food. Due to the presence of a large amount of water in this type of cooking, the specific heat of the water and food mixture is assumed to be equal to that of water, i.e., 4 kJ/kg °C. Thus, once the boiling point is reached, not much heat is required except that the heating rate should be equal to the rate of heat loss. The thermal losses from the cooking vessel area include the following: loss due to vaporization of water (35%), heating of the food–water mixture to the boiling point (20%), and convection losses from the vessel (45%). Box-type solar cookers are suitable for this type of cooking.

1.2.2 Frying

Oil is used as a medium in this type of cooking. The upper temperature limit of this mode is dependent on the characteristics of the oil used. Cooking is dependent on the boiling temperature of oil, which is generally between 200 and 300°C. Frying is a high-power cooking process and cooking is normally completed in a short time. Otherwise, food may become burned. Concentrator cookers are suitable for this type of cooking.

1.2.3 Roasting/Grilling

In this mode of cooking, heat is transferred to the food primarily through radiation and to some extent through convection. It is also a very high power process. Concentrating cookers are suitable for this type of cooking.

1.2.4 Baking

Baking is also a high-power cooking process in which heat is transferred from the oven wall by convection and radiation. Solar ovens are suitable for this type of cooking.

2. CLASSIFICATION OF SOLAR COOKERS

Solar cookers can be classified into four main categories: (1) concentrator cookers; (2) solar ovens; (3) box cookers; and (4) indirect (combined) solar cookers. These main categories can be further subdivided into different categories, as shown in Fig. 1.

3. PARABOLIC CONCENTRATOR COOKERS

A parabolic solar cooker is a concave bowl-shaped dish, whose inner surface is made of reflective material. Sunlight falling on the inner surface is focused onto a dark cooking pot that is suspended or set on a stand in front of the cooker.

The power output of the cooker depends on the size of the dish, the intensity of the solar radiation, the reflectivity of the inner lining, and the perfection

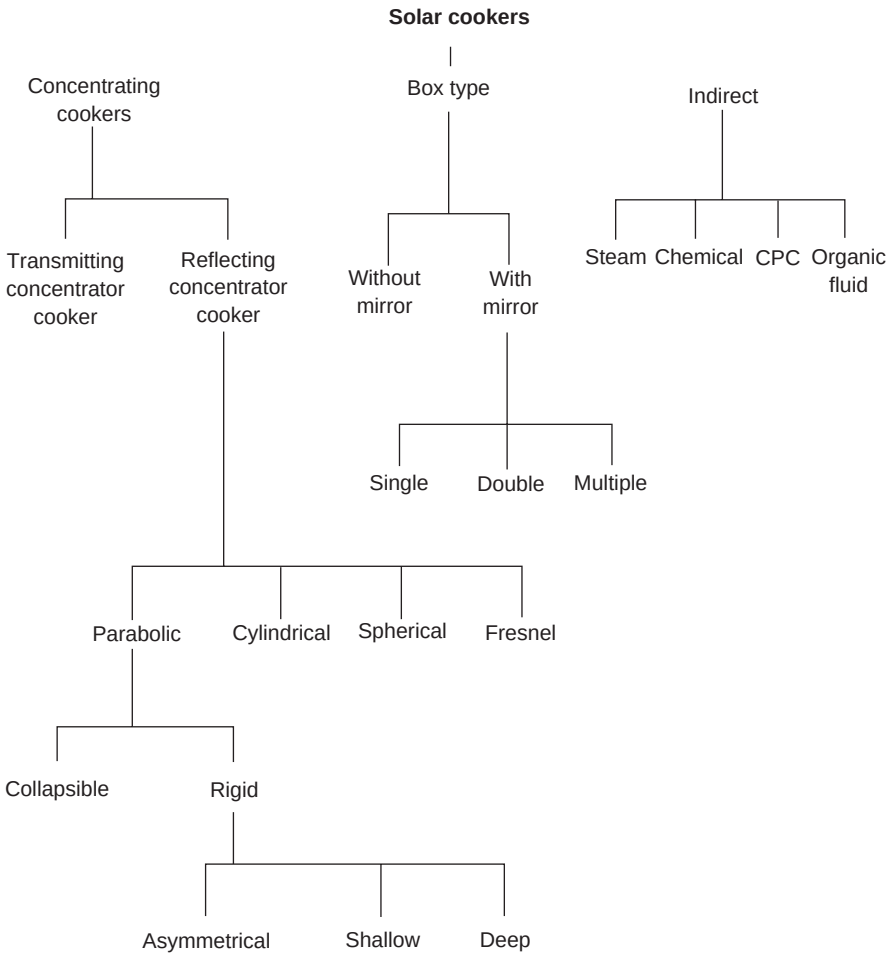


FIGURE 1 Types of solar cookers.

of the geometry. A concentrator cooker having 400 to 600 W output requires an area of 1 m^2. Concentrator cookers work only in direct sunlight; they work poorly under cloudy conditions. They require constant tracking. They also become unstable under windy conditions.

Maximum numbers of models of this category of concentrating cooker have been developed, as focusing of solar radiation is achieved much better using this geometry. However, constant tracking is required for this type of cooker as these cookers are sensitive to small changes in the relative position of the sun. Fabrication of parabolic geometry is also difficult compared to spherical geometry. In view of this, large sizes are fabricated with small strips and careful calculations are required to arrive at a suitable geometry. Some simpler methods, such as wire-reinforced concrete shells, soil cement depressions in the ground, and papier-maché shells, have been suggested. Parabolic concentrators can be classified into two categories: rigid parabolas and folding parabolas.

The majority of models that have been developed are based on the rigid parabola design. This category can be further subdivided into the following three subcategories: shallow parabolic cookers, asymmetrical parabolic cookers, and deep parabolic cookers.

3.1 Shallow Parabolic Cookers

In this type of cooker, the focus is outside the rim of the reflector.

A separate cradle-type stand is used to hold the cooking vessel as well as the reflector. In some designs, only the reflector is moved to track the sun.

3.2 Asymmetrical Parabolic Cookers

Asymmetrical parabolas are partial parabolas. This configuration enables the cooking to be done as close

as possible to the vessel. In the former design, there is a hole in the center of the reflector, where a stand is inserted to support the cooking vessel.

In the deep asymmetrical parabolic configuration, several small parabolic reflectors are arranged in an asymmetrical parabolic configuration. See Figs. 2, 3, and 4.

3.3 Deep Parabolic Cookers

In these cookers, the focus is within the rim of the reflector. The cooking vessel is hung from a stand inside the deep parabolic reflectors. Focusing is essential once every 20–30 min. These cookers were not very successful, due to bright sunlight hitting the eyes during cooking, the need to handle hot cooking vessels, and the need for periodic tracking. See Figs. 5 and 6.

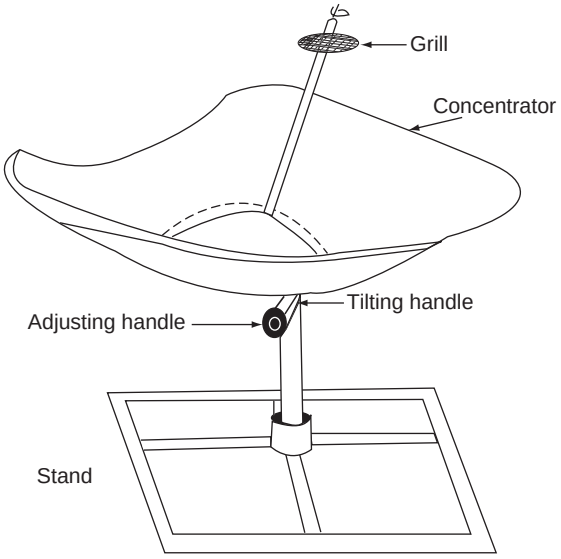


FIGURE 2 Asymmetrical parabolic cooker.

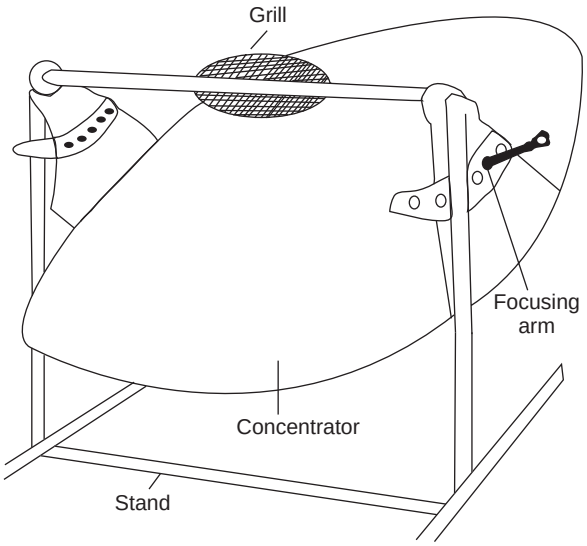


FIGURE 3 Asymmetrical parabolic cooker.

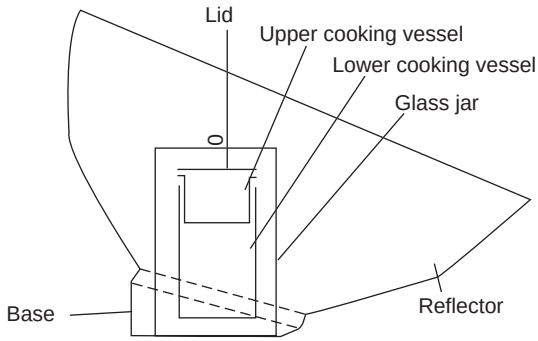


FIGURE 4 Asymmetrical parabolic cooker.

FIGURE 5 Deep parabolic cooker.

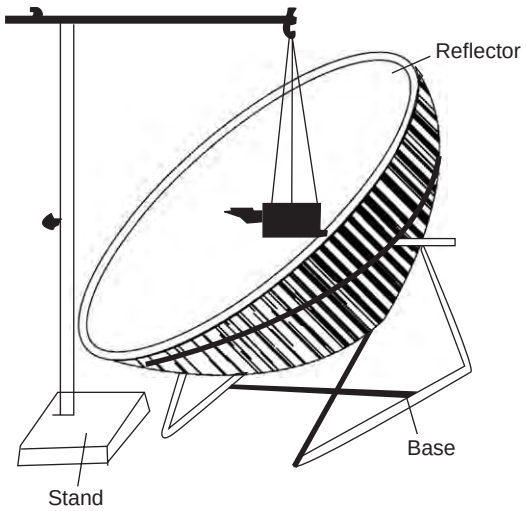


FIGURE 6 Deep parabolic cooker.

3.4 Collapsible Parabolic Cookers

Parabolic cookers are difficult to transport due to their bulky size. To facilitate easy transportation, collapsible parabolic cookers were designed. Some of the designs open like an umbrella. In others, inflatable plastic configurations have been used. Fluttering focus and unstable reflectors were significant impediments in the large-scale propagation of these cookers. See Figs. 7, 8, and 9.

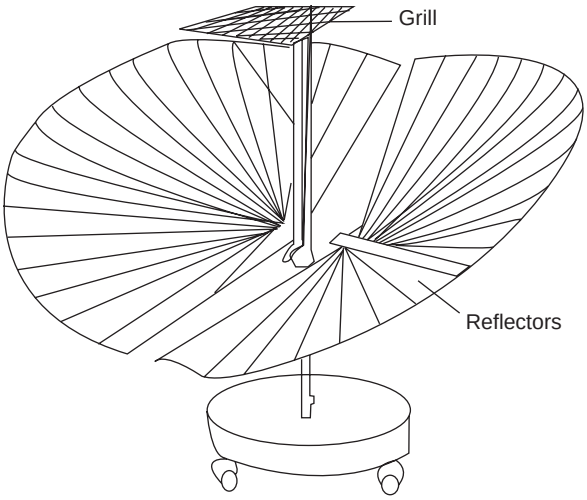


FIGURE 8 Collapsible parabolic cooker.

3.5 Cylindro-Parabolic Cookers

In cylindro-parabolic cookers, rays are focused in an insulated box in which the cooking pots are placed. One or more cylindro-parabolic mirrors can be used. The tracking arrangement of these mirrors is such that they focus solar radiation into the cavity of the insulated box. See Figs. 10 and 11.

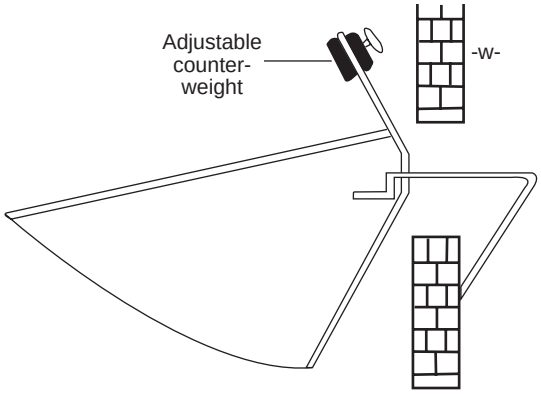


FIGURE 7 Collapsible parabolic cooker.

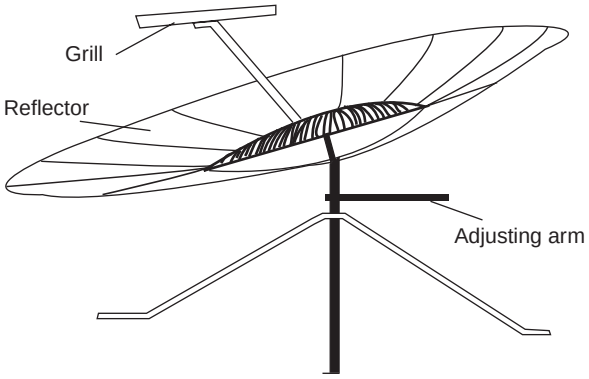


FIGURE 9 Collapsible parabolic cooker.

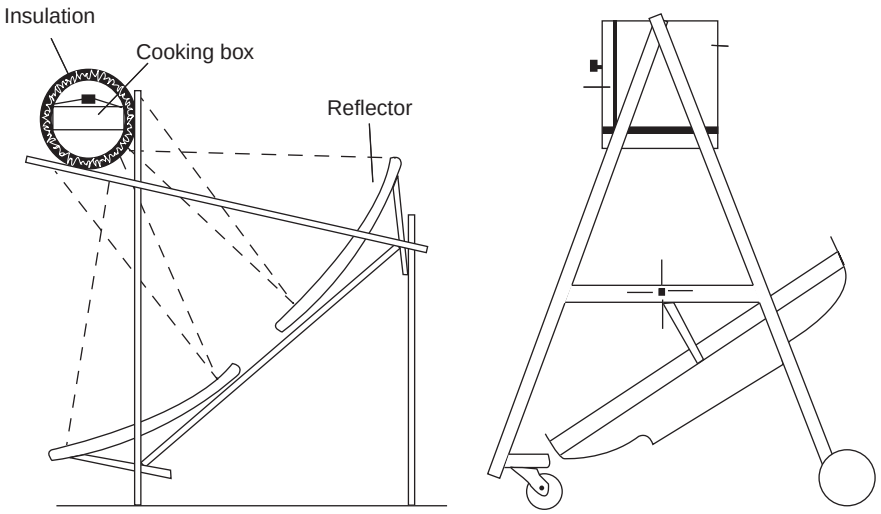


FIGURE 10 Cylindro-parabolic cooker.

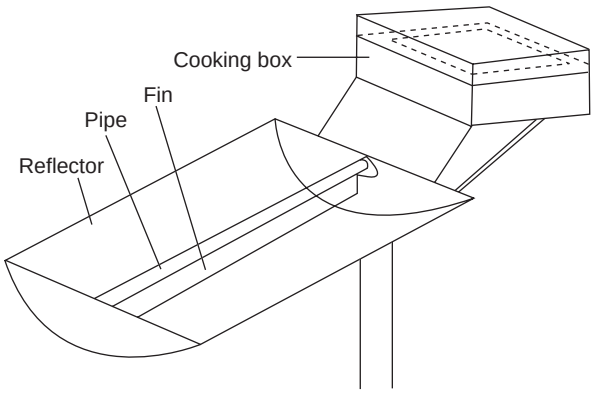


FIGURE 11 Cylindro-parabolic cooker.

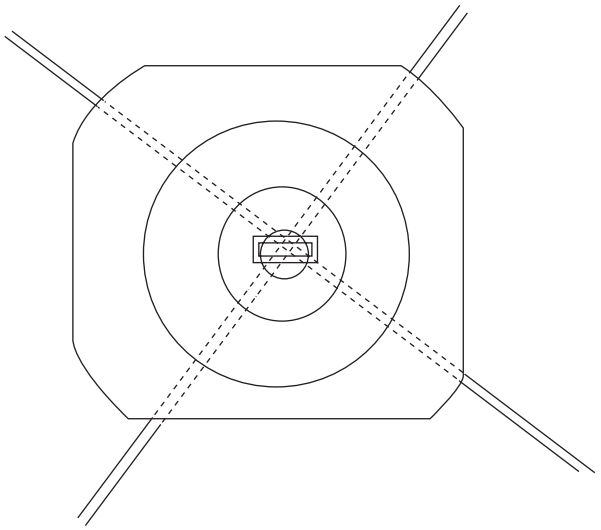


FIGURE 12 Fresnel reflector cooker.

4. FRESNEL REFLECTOR COOKERS

The reflectors for these types of cookers are made by using reflecting rings in a concentric Fresnel geometry. Any size of solar cooker can be made using this technique. However, designing such cookers requires considerable skill. See Fig. 12.

5. PLANE MIRROR CONCENTRATING COOKERS

Plane mirrors have been used in different configurations, such as cones or folding plane stands, such that these mirrors focus rays onto the cooking pot. Provisions are made to tilt the mirror assembly to focus the sun's rays onto the cooking pot during cooking. However, these designs did not become popular due to the large size of the stands. See Fig. 13.

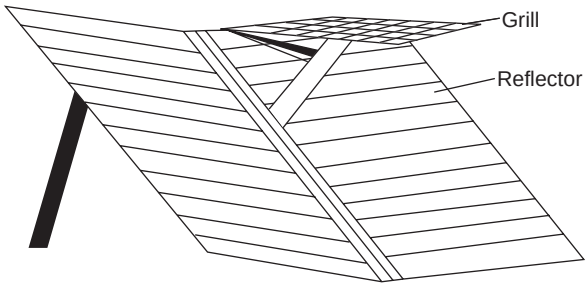


FIGURE 13 Plane mirror concentrating cooker.

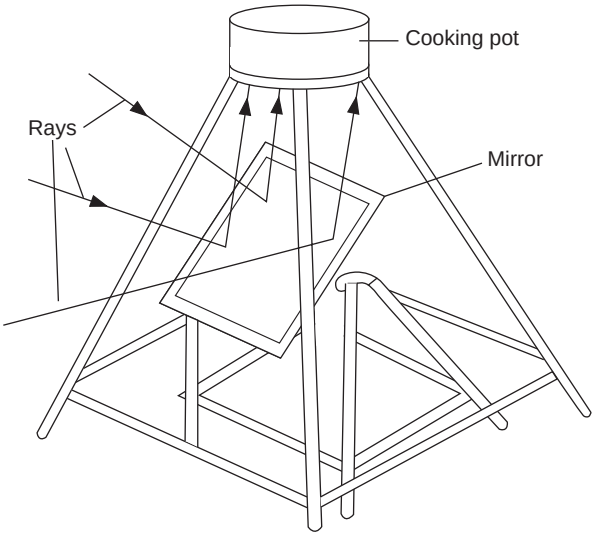


FIGURE 14 Water lens cooker.

6. TRANSMITTING CONCENTRATOR COOKERS

These cookers can be divided into three categories: water lens, dome lens, and Fresnel lens cookers.

6.1 Water Lens Cookers

These types of cookers concentrate light from above. The concentrator in such cookers consists of either plastic or glass concave sheets that are filled with water or alcohol. In some designs, an additional lens of small diameter is used to form a sharp focus. In a modified version, a deflector is used to focus the light onto the bottom of the pot. See Figs. 14 and 15.

6.2 Dome Lens Cookers

In such cookers, a dome comprising multiple cookers is used. However, fabrication of an assembly of lenses having a long focus is difficult.

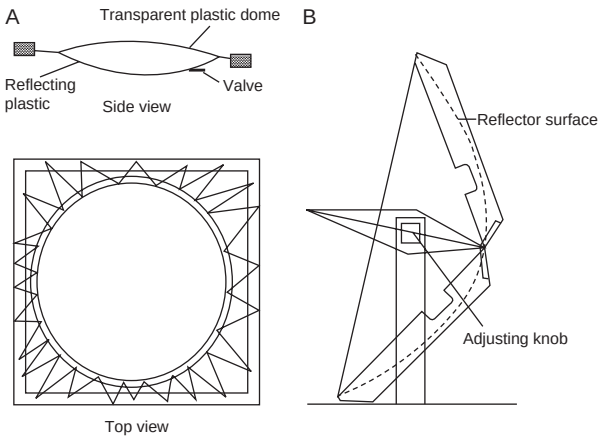


FIGURE 15 Water lens cooker.

7. FRESNEL LENS COOKERS

The use of transparent Fresnel lenses has been suggested for cooking applications. Two configurations of this design have been suggested. In one, the vessel is tracked, and in the other, the lens assembly is tracked. These designs did not become popular due to the high cost of Fresnel lenses. Also, plastic lenses deteriorate very fast. With the availability of ultraviolet-resistant polymers, such cookers may become a feasibility.

8. PRINCIPLES OF PARABOLOID COOKERS

Only the beam component of solar radiation is normally collected by paraboloid concentrating solar cookers, which are reflected to the focal point. Thus, for satisfactory performance, predominantly sunny skies are required. Thermal energy input to the cooking vessel is immediately cut whenever there is an obstruction between the sun and the cooker. The length of time spent in the shade and other factors, such as wind and the size of the cooking vessel, determine the recovery time.

8.1 Theoretical Consideration in the Design of Concentrator Cookers

The basic function of a concentrator is to collect radiation over a large area and concentrate it onto a smaller area so as to attain higher temperatures. Concentration can be defined by the following two methods: geometric concentration and flux concentration.

Geometric concentration is defined as the ratio of aperture area to receiver area

$$C_{\text{Area}} = A_{\text{A}}/A_{\text{R}}. \tag{1}$$

and flux concentration is defined as the ratio of intensity at the aperture to intensity at the receiver

$$C_{\text{Flux}} = I_{\text{A}}/I_{\text{R}}. \tag{2}$$

This involves an absorption effect in addition to geometry.

In concentrator solar cookers, geometric flux is the most widely used method.

According to the second law of thermodynamics, the maximum possible concentration for a given acceptance half-angle, θ_{c}, for a line-focusing concentrator collector is given by

$$C_{\text{Ideal 2D}} = 1/\sin\theta_{\text{c}}. \tag{3}$$

For point-focusing concentrators (cones, dishes, pyramids), the ideal concentration is given by

$$C_{\text{Ideal 3D}} = 1/\sin^2\theta_{\text{c}}, \tag{4}$$

where θ_{c} is the acceptance angle, which is defined as the angular range over which all or almost all rays are accepted without the collector being moved.

8.2 Concentration Limit

The angular radius of the sun $\Delta s = 0.25^\circ$.

Thus, the thermodynamic limit of a tracking two-dimensional solar cooker is 200 and that for a three-dimensional solar cooker is 40,000. However, the actual values achieved are much lower due to the nonoptimized design of the concentrator, tracking errors, nonspecular behavior of the optical system, and atmospheric scattering.

Focal length is an important parameter in the design of concentrator cookers. For spherical concentrator cookers, focal length changes with the length of the curve in a spherical mirror, as described by

$$B = (1/2)r^2/(r^2 - y^2)^{1/2}, \tag{5}$$

where r is radius and y is height.

In parabolic cookers, focal length depends only on the nature of the curve; it does not change with the length of the curve or the point on the curve, as is the case with spherical cookers.

8.3 Performance of Parabolic Concentrators

Parabolic concentrator performance depends mainly on two parameters, namely, the optical efficiency factor, $F\eta_o$, and the heat loss factor, FU_L. The optical efficiency factor depicts the perfection of the reflector surface area, its reflectance, and the absorption of the outer surface of the cooking pot. Thus, it defines the theoretical upper limit of the overall efficiency of a concentrator solar cooker.

The heat loss factor FU_L depends on the temperature of the water in the pot, the wind speed, the surface area of the cooking pot, and the orientation of the reflector. The heat loss factor can be evaluated by studying the cooling curve and optical efficiency can be evaluated by studying the heating curve.

8.4 Heat Loss Factor

8.4.1 Theory

The infinitesimal time interval dt for a fall of infinitesimal temperature of dT_w can be depicted by the following equations:

$$dt = \frac{(MC)\S_w dT_w}{Q_L}. \tag{6}$$

$$= \frac{(MC)\S_w dT_w}{A_{t'} F\S U_L (T_w - T_a)}. \tag{7}$$

Assuming that the heat loss factor U_L and the ambient temperature T_a are constant during the cooling test, the above equations can be integrated over the time interval t during which water temperature falls from $T_{wo}-T_w$.

$$t = \frac{(MC)\S_w}{A_t F\S U_L} \ln\left[\frac{(T_w - T_a)}{(T_{wo} - T_a)}\right]. \tag{8}$$

$$(T_w - T_a) = (T_{wo} - T_a) e^{t/t_0}, \tag{9}$$

It can be inferred from the above equations that t_o is time constant, as after time $t = t_o$, the temperature difference $T_w - T_a$ falls to $1/e$ of the initial value.

8.4.2 Optical Efficiency

Optical efficiency can be evaluated by the heating curve method. In this method, the time taken, dt, for a rise in temperature, dT, of water during heating is measured.

$$dt = \frac{(MC)\S_w dT_w}{Q_u}. \tag{10}$$

$$dt = -\frac{(MC)\S_w dT_w}{F\S[Ap\eta_o I_b - A_t U_L (T_w - T_a)]}. \tag{11}$$

If the insulation and ambient temperature are constant over a certain interval of time t, the above equations can be integrated over this time interval when temperature rises from T_{w1} to T_{w2} as

$$t = -t_o \ln\left\{\frac{f\S\eta_0 - \frac{F\S U_L}{C}\left[\frac{T_{w_2} - T_a}{I_b}\right]}{F\S\eta_o - \frac{F' U_L}{C}\left[\frac{(T_{w_1} - T_a)}{I_b}\right]}\right\}, \tag{12}$$

where t_o is the time constant obtained from the cooling curve and C is the ratio of aperture area of the cooker to the area of the pot surface (A_p/A_t).

The above equation can be rewritten as

$$F\S\eta_o = \frac{F\S U_L}{C}\left\{\frac{[(T_{w_2} - T_a)/I_b][(T_{w_1} - T_a)/I_b]}{1 - e^{-t/t_o}}\right\}^{e^{-t/t_o}}, \tag{13}$$

where $F\eta_o$ is the optical efficiency.

Time to boil can be calculated by writing the equation

$$t_{boil} = t_o \ln\left\{\frac{1}{1 - [(F\S U_L)/(F\S\eta_o)](1/C)[(100 - T_a)/I_b]}\right\}. \tag{14}$$

8.4.3 Procedure

The dimensions of the parabolic reflector cooker, such as aperture diameter and rim angle, are measured. The height, diameter, and weight of the cooking vessels are also measured. The cooking vessels are loaded with water and the parabolic concentrator solar cooker is pointed toward the sun in such a manner that the bright spot is positioned on the cooking vessel. A heating curve is obtained until the water temperature reaches 90–95°C. The concentrator is then shaded and a cooling curve is obtained until the temperature is close to ambient.

A semilogarithmic graph between $\ln(T_w - T_a)$°C and length of time of cooling is plotted. The slope of the line is equal to $(-1/t_o)$.

The heat loss factor is calculated using the following equation:

$$F' U_L = \frac{(MC)\S_w}{t_o A_t}. \tag{15}$$

Using this heat loss factor, the optical efficiency is calculated using Eq. (13).

9. SOLAR BOX COOKERS

Box-type solar cookers/ovens use the greenhouse effect to cook food. A transparent glass or plastic cover over the insulated box allows short-wave radiation to pass through. The black-coated inner box absorbs this radiation. With the increase in the temperature of the inner tray, energy is reradiated at longer wavelengths. The glass cover does not allow long-wave radiation to pass through it. The temperature of the box increases and an equilibrium temperature is reached, where the input of solar energy is balanced by the heat losses. In order to enhance the performance of this cooker, a plane reflector is attached to the box. This type of cooker can hold a number of pots of food. Food does not burn, as it is cooked at low temperatures. Due to its light weight, it is easy to carry. The cost of this type of cooker is less than that of concentrating cookers.

The most popular box solar cooker was designed by M. K. Ghosh, in 1945. It consists of a rectangular aluminum tray. This tray is placed in a box made of fiberglass-reinforced epoxy resin, wood, aluminum, and a galvanized iron sheet. The space between the aluminum tray and the outer box is filled with insulation. The tray is covered with a movable double-glass cover, attached by a hinge to one side of the outer box. A flat glass mirror encased in a fiberglass-reinforced epoxy resin shell is fixed to serve as a reflector and also as a cover when the cooker is closed. A hinge arrangement is provided to tilt the reflector at different angles to focus solar radiation into the tray. Cylindrical tight-fitting vessels with flat bottoms and top lids are used for cooking food.

The aluminum tray, the outside of the cooking vessels, and the lids of the vessels are painted dull black. A temperature of 120°C can be attained in this type of cooker, which is sufficient for boiling-type cooking. Keeping the general configuration intact, a number of alternate materials have been used by different designers to reduce costs and to use local materials. Aluminized Mylar film/anodized aluminum sheets have been used for the reflector. Agricultural residues, crushed newspapers, and other materials have been used for insulation. Cardboard has been used for the outer box and top lid. Plastic sheets have been used in place of glass. Although these cookers are cheap, food can be cooked properly in these configurations at only climatically favorable places. The lifetime of these cookers is shorter than that of custom-made cookers. See Fig. 16.

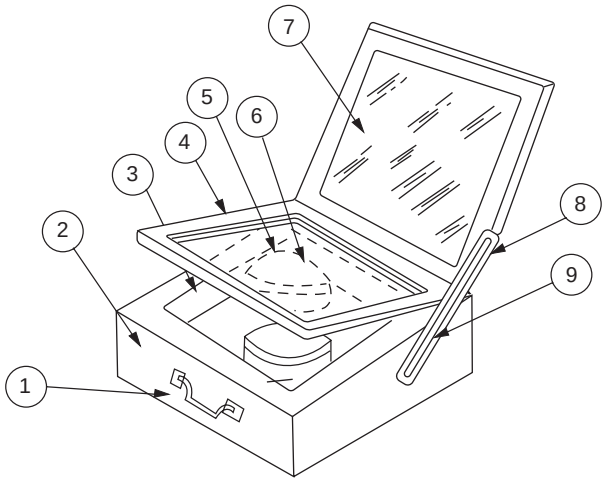


FIGURE 16 Box-type solar cooker. (1) Handle; (2) outer box; (3) insulation material (glass wall); (4) cover; (5) glass sheet; (6) cooking container; (7) plane mirror; (8) mirror support; and (9) hinged adjustor and guide.

9.1 Specifications of the Components (Bureau of Indian Standards)

9.1.1 Cover Plate

The inner cover plate is constructed of toughened glass 3 mm in thickness; the outer cover plate is constructed of plane glass 4 mm in thickness. Spacing is 10×2 mm. Solar transmittance should be 80% at a minimum. The aperture area is 500×500 mm. Tolerance is ± 5 mm.

9.1.2 Cooking Tray

The size should be $500 \pm 5\,\text{mm} \times 500 \pm 5\,\text{mm} \times 82 \pm 2$ mm. The construction material is aluminum. The sheet should be 0.56 ± 0.05 mm. The tray should be painted in a matte black finish to withstand temperatures up to 175°C. A scratch test should be carried out on a sample 10×10 mm in area.

9.1.3 Cooker Box

The material thickness for the galvanized iron sheet should be 48 ± 0.05 mm and that for the aluminum sheet should be 56 ± 0.05 mm. The FRP should be 2 mm, at a minimum. The size of the cooker box is $600 \pm 5 \times 600 \pm 5 \times 500 \pm 3$ mm.

9.1.4 Gasket

The gasket should be constructed of neoprene or silicon rubber and have a thickness of 3 mm.

9.1.5 Insulation

Insulation should consist of glass wool or mineral wool in 50 mm thick uniform pads. The insulation should be free of volatile material at 200°C.

9.1.6 Reflecting Mirror

The reflecting mirror should be 540 ± 5 mm $\times$ 540 ± 5 mm in size. The thickness should be 3 mm. Solar reflection should be equal to 65%.

9.1.7 Cooking pots

The construction material for the cooking pots should be aluminum or stainless steel. The depth of the pots should be 63 ± 2 mm. For aluminum pots of 200 ± 5 mm or 150 ± 5 mm in diameter, the thickness should be 1 mm. There should be three pots. For stainless steel pots of 200 ± 5 mm or 150 ± 5 mm in diameter, the thickness of the pots should be 50 mm. Two pots are needed.

9.1.8 Wheels

The wheels should be 25 mm in diameter, with ball bearings.

9.2 Performance Evaluation of Box-Type Cookers

Energy requirements are at a maximum during the heating period in boiling-type cooking and only a small amount of heat is required for the physical and chemical changes involved in the cooking process. Once the cooking temperature is attained, energy is required only to compensate for thermal losses and that required for physical and chemical changes and so the cooking is practically independent of the heat input rate. Thus, the time required to cook the same food in different cookers will depend on the heating period.

Two parameters have been recommended to evaluate the performance of box-type solar cookers. These are the first figure of merit, F_1, and the second figure of merit, F_2. The first figure of merit depicts optical efficiency and is the ratio of optical efficiency to heat loss factor; it is evaluated by a stagnation test without load. The second figure of merit, F_2, gives an indication of the heat-exchange efficiency factor (F') and involves heating of a full load of water.

In a box-type solar cooker, there is no control over its temperature and the operation is transient, in contrast to a solar water heating system, where a steady state is obtained by circulating water at different temperatures through the tubes. A quasi-steady state is achieved when a stagnation temperature is obtained. The heat loss factor at stagnation, UL_s, can be calculated.

Only cookers with good optical transmission of glass covers and a low overall heat loss factor will have a high value for the first figure of merit (F_1). However, good heat transfer to the vessel and its contents will be ensured by the second figure of merit, F_2. Since there is no arrangement for the flow of liquid to obtain a steady-state withdrawal of energy from the cooker, as is the case with solar water heaters, it is difficult to measure the heat-exchange efficiency factor F' between the cooker plate and the contents of the pot. A factor, F_2, is proposed for this purpose; it includes F' as well as U_{LS} and CR for the cooker utensil combination.

During both tests, the mirror booster is not used and the measurements are started approximately 2 h before solar noon.

To estimate the first figure of merit, F_1, a solar cooker with utensils is kept in the sunshine and the rise in plate temperature, the solar radiation intensity, and the plate temperature are recorded. When the stagnation temperature is reached, the ambient temperature and solar insulation are measured. The figure of merit, F_1, is calculated by

$$F_1 = \frac{\eta_o}{U_{LS}} = \frac{T_{ps} - T_a}{I_{Gs}}, \tag{16}$$

where η_o is the optical efficiency; I_{Gs} is the insulation on the horizontal surface at the time the stagnation temperature is reached (in watts per square meter); U_{LS} is the heat loss factor at stagnation; T_{ps} is the stagnation plate temperature (in degrees centigrade); and T_a is the plate temperature (in degrees centigrade).

To evaluate the second figure of merit, F_2, the solar cooker without the reflector but with a full load of water and utensils is kept in the sun. Initial water is kept at an average temperature between ambient and the boiling point. Temperature and solar radiation are measured until the temperature reaches 95°C.

The second figure of merit, F_2, is calculated using the equation

$$F_2 = \frac{F_1(MC)_w}{A(t_2 - t_1)} \times \ln\left\{\frac{1 - (1/F_1)[(T_{w2} - T_a)/I_G]}{1 - (1/F_1)[(T_{w1} - T_a)/I_G]}\right\}, \tag{17}$$

where F_1 is the first figure of merit from the stagnation test; $(MC)_w$ is the product of the mass of water and specific heat (in joules per degrees centigrade); A is the aperture area of the cooker of the cover plate (in square meters); $t_2 - t_1$ is the time taken for heating from T_{w1} and T_{w2} (in seconds); T_a is the average air temperature during the time period $t_2 - t_1$ (in watts per square meter); and I_G is the radiation during the time period $t_2 - t_1$ (in watts per square meter).

It has been assumed that the global radiation, I_G, remains constant during the test. This assumption introduces some error into the data.

9.3 Testing Methods for Box-Type Solar Cookers (Bureau of Indian Standards)

The following tests have been recommended by the Bureau of Indian Standards (BIS) for evaluating the durability and thermal performance of box-type solar cookers. These tests include leakage tests, breakage tests, reflectivity tests, exposure tests, and thermal performance tests.

9.3.1 Leakage Tests

9.3.1.1 Cooking Tray Leakage Test

1. The cooking tray should be dismantled from the main body.
2. The cooking tray is filled with water; after 1 h, the joints of the tray should be examined for any signs of leakage.

9.3.1.2 Rubber Gasket Leakage Test

1. A piece of paper is inserted between the gasket and the cover plate in at least four positions along each side of the cooker.
2. The paper used should be 50 mm wide and 0.01 mm thick.
3. The cover plate should be properly tightened.
4. The paper should exhibit a firm resistance to withdrawal by hand at all points tested.

9.3.1.3 Cover Plate Leakage Test Leakage from the cover plate may occur from the upper and lower sides. Therefore, the cooker should be tested using the following two tests:

i. Leakage test for the upper side of the cover plate:

1. The cover plate should be properly tightened.
2. A thin film of water should be poured onto the cover plate.
3. After 1 h, the cover plate should be examined for any signs of water having entered between the two glass sheets.

ii. Leakage test for the lower side of the cover plate:

1. The cooking pots are filled with water and kept in the cooking tray.
2. The cover plate is tightened.
3. The cooker is placed out in the open for exposure to the sun's rays for 4–5 h.
4. Then, the cooker should be placed in the shade for 15 min to allow any vapor to condense.
5. The cover plate is examined for any signs of water vapor entry between the inside and outside of the cover plate.

9.3.2 Rain Penetration Test

9.3.2.1 Apparatus The basic apparatus is a 5 mm spray nozzle and a balance.

9.3.2.2 Procedure

1. The closed cooker is sprayed with 10 liters of water on all sides using a spray nozzle at a pressure of 0.1 MPa.
2. Spray nozzles are directed downward from the cooker top and also toward the four corners of the cooker.
3. The cooker should be weighed before the test.
4. After the test, the external surfaces of the cooker are wiped dry and the cooker should be reweighed.
5. During wiping, transport, and placement on the weighing machine, the cooker should remain in a horizontal position.
6. The balance used to weigh the cooker should have an accuracy of at least 20 g.
7. The permissible gain in weight should not be more than 30 g.

9.3.3 Slam Tests

This test is performed to ensure that the mirror or cover plate will not be damaged when allowed to fall from the fully opened position as given below:

9.3.3.1 Cover Plate Slam Test With the lid open, the cover plate is lifted as high as possible and then it is allowed to fall to a closed position. There should be no damage to the glass sheets.

9.3.3.2 Slam Test Mirror The lid should be held at near vertical and is allowed to fall to a closed position. There should be no damage to the mirror, the cover plate, or any other part of the cooker.

9.3.4 Mirror Reflectivity Test

9.3.4.1 Apparatus The basic apparatus consists of a photovoltaic-based solarimeter and a stand for mounting a sample piece of the reflecting mirror. The

stand should have a vertical pointer of 15 cm in height and an arrangement for mounting the solarimeter in a plane parallel to the plane of stand and at a height of 30 cm.

9.3.4.2 Procedure

1. The stand is adjusted for normal incidence such that there is no shadow of the vertical pointer.
2. This sample of reflecting mirror (540 mm × 540 mm) is placed on the stand in an open space. The open space should be free of any shadows and it should not receive any reflecting radiation from the surroundings.
3. The solarimeter should be placed in the stand provided for it in such a way that the sensor is near normal ($\pm 10^{\circ}$) to the sun's rays. The solarimeter (R_1) is then read.
4. The position of the solarimeter is changed in such a way that the sensor now faces the mirror at a near normal angle. The solarimeter (R_2) is then read.
5. The reflectivity of the mirror is calculated as follows: $R = R_2/R_1$.
6. The test should be repeated six times at 10 min intervals.
7. The average of the six values is calculated. This value will represent the reflectivity of the mirror.
8. The permitted conditions of solar radiation during the test are that it should be greater than 600 W/m^2.

9.3.5 Exposure Test Apparatus

The basic apparatus consists of a solar pyranometer, along with recording devices.

9.3.5.1 Procedure

1. The solar cooker is left to stagnate, which may lead to the following types of possible degradation:
   - Breakdown of rubber or plastic material;
   - Outgassing from the insulating material;
   - Discoloration or peeling of black paint on the cooking pots and cooking tray; or
   - Deposition of water vapor, dust, or any other material inside the double-glass lid.
2. The solar cooker is left in an unshaded area for at least 30 days having irradiation levels of at least 14 MJ/m^2 on the horizontal surface.
3. These days need not be consecutive.
4. The cooking pots inside the cooker should be empty.
5. The mirror is placed vertically and the cooker should be oriented to face south.

9.3.6 Cooking Tray Paint Peeling Test

A 10 mm × 10 mm area of painted surface is selected from squares of I mm × I mm by scratching horizontal and vertical lines with a pointed tool. The surface portion is covered with adhesive cellophane tape and then the tape is removed with a jerk. The test surface should be inspected and there should be no peeling of paint.

9.3.7 Transmittance Test for Cover Plate

The solar radiation is measured with a pyranometer/photovoltaic-based solarimeter directly (without the glass) and then measured with the cover plate above the pyranometer and solarimeter. The ratio of the two readings should be at least 0.8.

9.3.8 Thermal Shock Test for Gasket/Sealant

The gasket/sealant of 10 cm in length should be kept in an electric oven at a temperature of 150°C for 4 h. It should be cooled in air for 2 h, again placed in the oven at 150°C for 4 h, and then again cooled in air. After the test, the gasket and sealant should exhibit no cracking or brittleness.

9.4 Thermal Performance Test

The thermal performance test should be conducted under the conditions given below; values of $F_1$and F_2 are then calculated.

9.4.1 Testing of Box-Type Cookers

9.4.1.1 Stagnation Temperature Condition Apparatus The apparatus consists of a pyranometer and a platinum resistance thermometer along with a recording device.

9.4.1.2 Procedure The test for evaluating F_1 should be carried out as follows in the morning hours before 10:00 A.M. so that the stagnation temperature is achieved near solar noon.

1. The solar cooker is placed in full sun.
2. The reflector of the solar cooker is covered with a black cloth.
3. The temperature inside the solar cooker is measured using RTD for a continuous 5 min interval.
4. When the cooker temperature has reached a steady state, the final temperature inside the cooker (T_{ps}), the corresponding ambient air temperature (T_a), and solar insolation are noted.

The steady-state conditions are defined as a 10 min period when:

1. The cooker temperature is constant within 1°C;

2. Solar radiation is constant within 20 W/m^2;
3. Ambient temperature is constant within 0.2°C; and
4. Solar radiation is greater than 600 W/m^2.

The performance factor F_1 is calculated using Eq. (16).

9.4.2 Heat-Up Condition

The figure of merit F_2 should be estimated as follows.

9.4.2.1 Apparatus The apparatus consists of a pyranometer and a platinum resistance thermometer, along with a recording device.

9.4.2.2 Procedure The empty cooking pots are weighed and then filled with water at 50°F (15°C). They are reweighed and the mass of water is calculated. The pots are placed in the cooker from which the mirror has been removed or covered. A temperature probe of RTD is placed in the largest of the cooking pots with the measuring tip submerged in the water. The temperature probe lead should be sealed where it leaves the cooking pots and the cooker.

The ambient temperature and wind speed are measured throughout the test. The test should start in the morning between 11:00 and 11:30 A.M. of local solar time. If radiation and temperature are measured by spot checks, these should be no more than 5 min apart.

Constant monitoring at 30 s intervals or less is desirable, with averages of radiation recorded over 2 min intervals. The following measurements should be taken:

1. Water temperature is measured along with the exact time that that measurement was recorded.
2. The data recording is continued until the water temperature exceeds 95°C.
3. Initial and final temperature/time data pairs are chosen. The initial temperature should be between 60 and 65°C and the final temperature should be between 90 and 95°C. These are denoted T_{w1}, and T_{w2}, respectively, and the corresponding times are t_1 and t_2, respectively.
4. The average air temperature (T_a) between times t_1 and t_2 is calculated.
5. The radiation recorded between the two points should not have varied by more than ± 10 W/m^2 and should always have exceeded 600 W/m^2.
6. The average radiation (I_g) over the time t_1 and t_2 is calculated.
7. The second figure of merit, F_2, is calculated as per Eq. (17).
8. As per the BIS standard, the value of F_1 should be >0.12 for grade A cookers and >0.11 for grade B cookers.

Theoretical analysis as well as experimental evaluation shows that it not possible to achieve these values with the BIS-proposed specifications. However, these values can be achieved by increasing optical efficiency and decreasing the overall heat loss coefficient.

10. MULTIPLE GLASS BOX COOKERS

It is difficult to cook food in a single reflector box-type solar cooker at higher latitudes during the winter season due to higher losses resulting from low ambient temperatures. To overcome this problem, folding two-step asymmetrical reflector box solar cookers have been designed. These consist of a hot box and two plane mirrors fixed at an angle of 13.5°. The hot box has double glazing at the top and insulation on the bottom and sides. At the end of cooking, mirrors can be folded onto the top of the hot box. See Figs 17 and 18.

Solar radiation through the aperture, AL_1, AL_2, and AL_3, can be calculated by the following equations:

$$AL_1 = B\sin(\phi). \tag{18}$$

$$AL_2 = B\sin(\theta_1 - \phi). \tag{19}$$

$$AL_3 = \sin(\theta_2 - \phi). \tag{20}$$

$$\theta_1 = \pi/3 + 2/3\phi. \tag{21}$$

$$\theta_2 = \theta_1 - 13.5. \tag{22}$$

See Fig. 19.

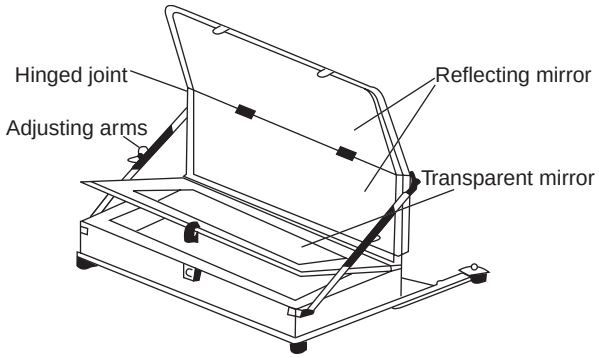


FIGURE 17 Multiple glass box cooker.

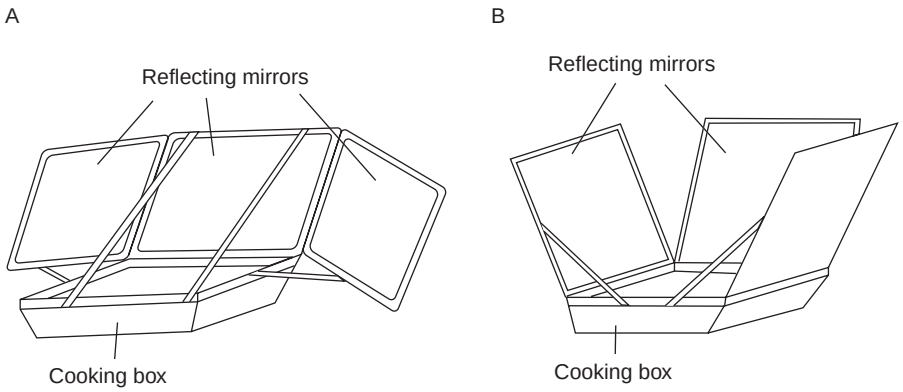


FIGURE 18 Multiple glass box cooker.

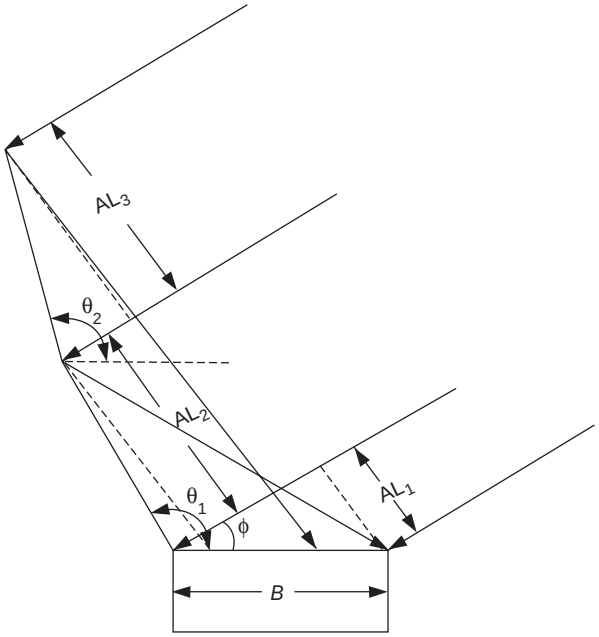


FIGURE 19 Depiction of the geometry used to determine the solar radiation through the aperture in multiple glass box cookers.

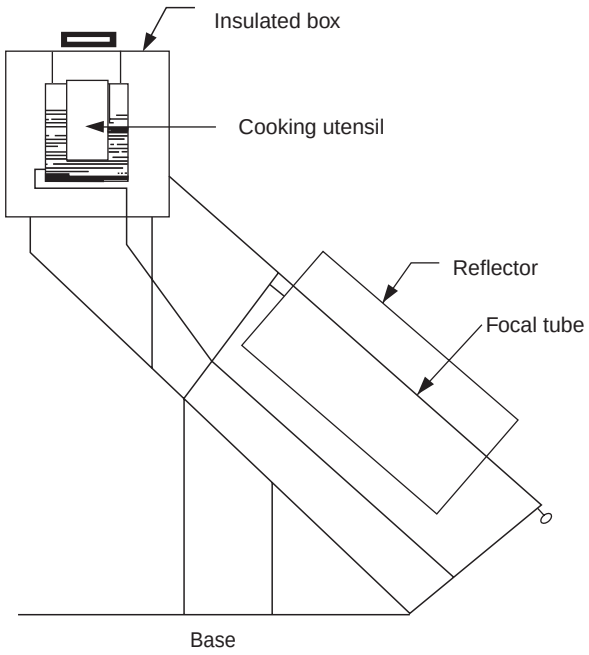


FIGURE 20 A typical configuration of an organic fluid cooker.

11. INDIRECT SOLAR COOKERS

There are no provisions in the conventional box-type or concentrator-type cookers for cooking inside the kitchen or during the period when solar radiation is not available. A number of strategies have been used for special types of indirect solar cookers.

11.1 Organic Fluid Cookers

In this type of cooker, organic fluid is heated using a cylindrical parabolic reflector/CPC/flat-plate collector outside the kitchen. The hot fluid is stored inside the kitchen using the thermosphon principle for cooking applications.

A typical configuration of this type of cooker is shown in Fig. 20.

11.2 Steam Cookers

This type of cooker is similar to the organic fluid cooker, except that low-pressure steam is used as the working fluid. Using flat-plate or concentrating reflectors, low-pressure steam is produced. Food is cooked in a double-jacketed vessel.

11.3 Chemical Cookers

These cookers are based on a reversible chemical reaction between a working fluid (gas or vapors) and

FIGURE 21 A typical configuration of a Scheffler community cooker.

a chemical compound. The reaction is endothermic in one direction and exothermic in the reverse direction. Working fluid becomes desorbed by solar energy and is stored for later use. When heat is required, the desorbed fluid is again allowed to adsorb in the chemical. The liberated heat is used for cooking.

12. SCHEFFLER COMMUNITY COOKER

This cooker has a 6.5 m^2 parabolic reflector dish, which is placed nearly 3 m away from the north-facing wall. The reflector reflects and concentrates solar radiation onto a secondary reflector through an opening in the wall. The secondary reflector deflects the radiation to the bottom of the cooking vessel. Temperatures of up to 400°C can be attained. Cooking can be carried out in a conventional manner in the kitchen. The reflector is tracked with a clockwork mechanism to ensure continuous cooking. There is an arrangement for seasonal adjustment of the reflector. Figure 21 shows the configuration of the cooker.

13. RATING OF SOLAR COOKERS

At the Third International Conference on Solar Cooker Use and Technology, the question of rating solar cookers based on field tests for comparing different models and types of solar cookers was discussed. It was proposed that apart from thermal performance, there are other important parameters that need to be considered from the user's point of view. The selection process in the field should take into consideration the particulars of the sites, conditions, and consumers. The rating criteria should include, in addition to thermal performance, critical factors such as the following: (1) cost, (2) safety, (3) heating and cooking capacity, (4) convenience, (5) durability, (6) ease of maintenance, (7) stability in the wind, and (8) operating instructions on the use of the cooker.

For wider acceptability of the solar cooker, the cost should be reasonable so that ordinary users in developing countries can afford it. It should meet the cooking needs of an average family. It should be convenient to cook food in the cooker, as women do not prefer a gadget that is cumbersome to use. The equipment should be safe to work with and should not cause burn injuries while one is cooking or removing cooked food from the apparatus. As solar cooking is done outdoors, the construction material should be durable so that the cooker may last for a minimum of 8–10 years. Cookers should be easy to maintain as there is a lack of availability of technically trained personnel for maintenance in developing countries and there is a high cost of maintenance in developed countries. A maintenance manual should be provided. Under windy conditions, the cooker can be blown off or fall, resulting in breakage of the glass components and spoiling of the food. There should be provisions for holding the cooker steady under windy conditions. A complete operating manual should be provided in written text as well as in pictorial form to overcome language barriers.

SEE ALSO THE FOLLOWING ARTICLES

Solar Cells • Solar Cooling, Dehumidification, and Air-Conditioning • Solar Detoxification and Disinfection • Solar Distillation and Drying • Solar Energy, History of • Solar Fuels and Materials • Solar Heat Pumps • Solar Ponds • Solar Thermal Energy, Industrial Heat Applications • Solar Thermal Power Generation • Women and Energy: Issues in Developing Nations • Wood in Household Energy Use

Further Reading

Blum, B. (ed.). (1989). "The Solar Box Cooker Handbook." Solar Box Cooker International, Sacramento, CA.

Bureau of Indian Standards (1992). Indian Standard No. IS 13429, Solar Cooker, Parts 1, 2, and 3.

Duffie, J. A., Log, G. I. G., and Back, B. (1953). Laboratory and field studies in plastic reflector solar cookers. *In* "Proceedings of the United Nations Conference on New Sources of Energy E 35-587," Vol. 5, pp. 339–346.

Garg, H. P., Bandyopadhyay, and Datta, G. (1985). Mathematical modelling of the performance of a solar cooker. *Appl. Energy* **14**, 233–239.

Garg, H. P., and Kandpal, T. C. (1999). "Laboratory Manual of Solar Thermal Experimentation." Narora.

Ghai, M. L. (1953). Solar heat for cooking. *J. Sci. Industr. Res.* **12A,** 117–124.

Ghosh, M. K. (1945). Solar cooker. *Bull. Inst. Eng.* **14,** 1–7.

Habeebullah, M. B., Khalifa, A. M., and Olwi, I. (1995). The oven receiver: An approach toward the revival of concentrating solar cookers. *Solar Energy* **54,** 227–237.

Kundapur, A. (1998). Review of solar cooker designs. *TIDE* **8**.

Lof, G. O. G. (1963). Recent investigation in the use of solar energy for cooking. *Solar Energy* **7,** 125–132.

Mullick, S. C., Khandpal, T. C., and Saxena, A. K. (1987). Thermal test procedure for box-type solar cookers. *Solar Energy* **39,** 353–360.

Shukla, J. K., and Patel R. C. (1990). Study of connective heat losses in solar cookers. *In* "Proceedings of the National Solar Energy Convention" (Udaipur, A. N. Mathur, and N. S. Rathore, Eds.). Himanshu Publications.

Solar Cooling, Dehumidification, and Air-Conditioning

GERSHON GROSSMAN
Technion—Israel Institute of Technology
Haifa, Israel

1. **Introduction to Solar Cooling**
2. **Thermodynamic Principles**
3. **Closed-Cycle Systems**
4. **Open-Cycle Systems**
5. **Conclusion**

Glossary

absorption A thermodynamic cycle powered by heat and employing a working fluid pair (absorbent/refrigerant) to produce cooling.

closed-cycle system Cooling/dehumidification device in which the working fluid is enclosed in the machine and interacts with the cooled medium by heat transfer through solid walls.

coefficient of performance (COP) Ratio of the cooling produced to the heat/work supplied.

desiccant Hygroscopic material serving to attract and retain humidity from surrounding air.

heat engine Cyclic device receiving heat from a high-temperature heat reservoir, rejecting part of it to a low-temperature heat reservoir, and converting the difference to work.

heat pump Cyclic device extracting heat from a low-temperature heat reservoir and rejecting it to a high-temperature heat reservoir, powered by work.

heat reservoir A source/sink of heat at constant temperature.

open-cycle system Cooling/dehumidification device in which the working fluid (water and/or air) forms part of the cooled medium.

The possibility of providing cooling and air-conditioning by means of energy from the sun has attracted the attention of humans since the early development of solar technology. The necessity of air-conditioning for thermal comfort in hot areas of the world and the abundance of sunshine in these areas have intrigued researchers with regard to how to combine the two for people's benefit. Furthermore, in contrast to other solar applications such as heating, the greatest demand for cooling occurs when the solar radiation is most intense, thus making its use for this application all the more attractive.

1. INTRODUCTION TO SOLAR COOLING

Among the various thermal applications of solar energy, cooling is one of the more complex, both in concept and in construction. This is no doubt one of the reasons why its utilization is not as widespread as space or water heating. It is not sufficient to collect, store, and distribute the solar heat. The energy must be converted to cold by means of a heat-actuated device capable of extracting heat at a low temperature from the conditioned space and rejecting it into the higher temperature of the outside air. In thermodynamic terms, a heat-pumping operation is required.

A schematic diagram of a complete solar air-conditioning system is presented in Fig. 1 showing the different components common to most such systems. A heat transfer fluid is heated in the solar collectors to a temperature well above ambient and used to power a cooling device—a type of heat-actuated heat pump. The heat transfer fluid may be air, water, or another fluid; it can be stored in a hot state for use during times of no sunshine. Heat extracted by the cooling device from the conditioned space and from the solar energy source is rejected to the environment using ambient air or water from a cooling tower. The cooling device may produce the cooling effect in a number of different ways. It can

 © 2004 Elsevier Inc. All rights reserved.

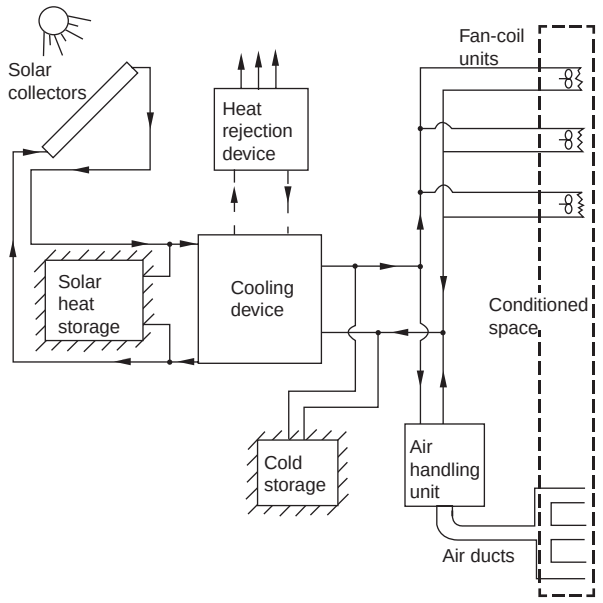


FIGURE 1 Schematic diagram of a solar cooling system.

generate chilled water that may be distributed through fan coil units or used to cool air, which is then distributed through air ducts; the chilled water may be stored for use when the cooling device is not in operation. Alternatively, the machine can cool the conditioned air directly.

Solar collectors are a major part of every solar cooling system. They convert solar radiation into heat at a temperature suitable for powering the cooling device. The collectors vary in type from low-temperature flat plates to more sophisticated high-temperature evacuated tubes or concentrators.

Growing demand for air-conditioning in recent years, particularly in hot and humid climates, has imposed a significant increase in demand for primary energy resources. Electric utilities experience peak loads on hot summer days and are often faced with brown-out situations, barely capable of meeting the demand. With suitable technology, solar cooling can help alleviate, if not eliminate, the problem. This article describes trends in solar cooling, dehumidification, and air-conditioning. It is divided into two main parts and considers closed-cycle and open-cycle alternatives, each with its own advantages and limitations.

2. THERMODYNAMIC PRINCIPLES

In considering the various options for heat pumping systems that may be employed for solar cooling, one must take into account the thermodynamic principles of a heat-powered cooling device. Typically, the system operates between heat reservoirs at three temperature levels: T_L, the low temperature of the cooled medium or the conditioned space; T_H, the high temperature supplied by the solar heat source; and T_A, the intermediate temperature of the environment serving as a sink for the reject heat. The cooling device may be described in basic thermodynamic terms as a combination of a heat engine and a heat pump, as illustrated in Fig. 2. The former receives heat Q_H from the high (solar) heat reservoir at T_H, rejects part of it to the heat sink at T_A, and converts the rest to mechanical work W. The latter employs this work to pump heat Q_L from the low-temperature reservoir at T_L to the heat sink at T_A. According to the second law of thermodynamics, best results are obtained from this machine when both cyclic devices are reversible and operating at the Carnot limit. Under these optimum conditions, the overall coefficient of performance (COP) of the cooling device is found to be

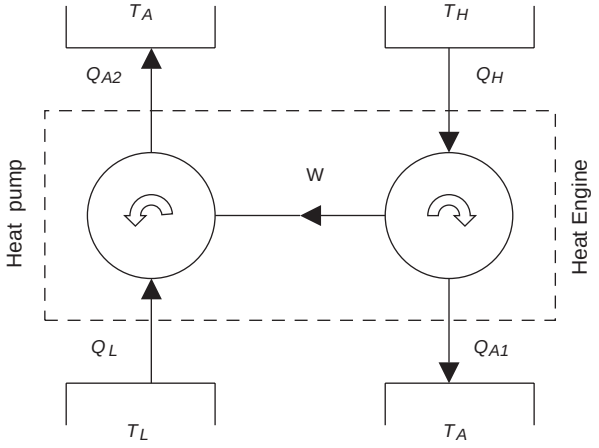


FIGURE 2 Thermodynamic description of a heat-powered cooling device.

$$\mathrm{COP} = \frac{Q_L}{Q_H} = \frac{T_L}{T_H}\frac{(T_H - T_A)}{(T_A - T_L)}. \qquad (1)$$

As evident from Eq. (1), the performance of the heat-powered cooling device is highly dependent on the temperature of the heat source, with better results obtained with higher T_H. The type of solar collector employed to supply the heat thus imposes a limitation on the COP. With common flat-plate collectors, the heat is obtained at relatively low temperatures. Concentrating collectors reach higher temperatures but are more expensive and complex to operate due to their need to track the sun. Another conclusion

from Eq. (1) is that performance improves by lowering the heat rejection temperature T_A. From this standpoint, heat rejection by water cooling is preferred to that by air cooling.

3. CLOSED-CYCLE SYSTEMS

In practice, there need not be a clear distinction between the heat engine and the heat pump parts of the machine. Cycles such as absorption or ejector (jet) have been proposed in which the work produced by part of the machine is "invisible" and goes directly to power the other. Nevertheless, the fundamental thermodynamic principles of the cooling device described previously still hold and set the Carnot upper limit for the COP as given by Eq. (1).

During the great surge in solar energy research and development following the energy crisis of the 1970s, there have been several attempts to build solar-driven cooling systems by means of a solar heat engine driving a heat pump, according to Fig. 2. The former varied from Stirling engines to Rankine engines employing organic fluids; the latter were conventional vapor-compression cooling devices with common refrigerants. These efforts have generally not gone beyond the laboratory. Experience has shown that closed-cycle systems most suitable for solar cooling are based on absorption cycles, which totally eliminate the mechanical interaction between the two parts of the machine. Absorption systems have many advantages: operation at wide ranges of heat source and heat sink temperatures; multistaging for improving COP with high-temperature heat sources; quiet operation with no moving parts; and environmentally safe working fluids. Moreover, absorption systems are amenable to combining available solar heat with backup heat for meeting the cooling demand.

Figure 3 is a schematic diagram of a typical absorption chiller in the simplest, single-effect configuration. The system employs the working fluid pair lithium bromide (LiBr)–water, in which the water serves as the refrigerant and the LiBr as the absorbent. The diagram is plotted in a system of coordinates showing the temperature on the horizontal axis and the pressure prevailing in the system's exchange units on the vertical axis.

The operation of single-effect absorption chillers is well documented in the literature and is described only briefly here. Liquid refrigerant (water) leaving the condenser at state 12 in Fig. 3 expands through the throttle valve (6) into the evaporator (1), where it evaporates from state 13 to 9, taking its heat of evaporation from the stream of chilled water and cooling it from state 1 to 2. The vapor leaving at

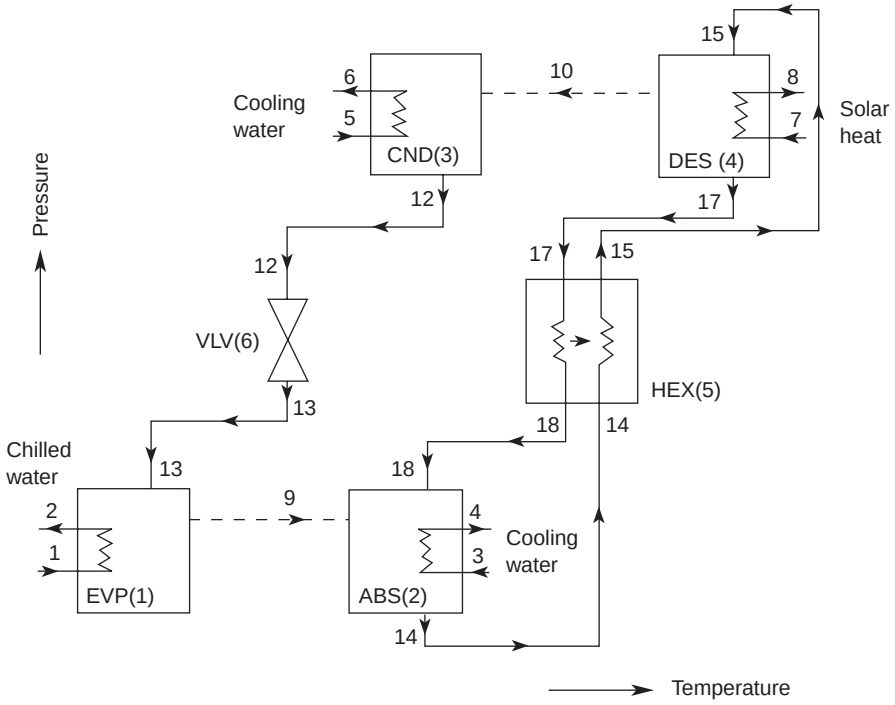


FIGURE 3 Schematic description of a single-effect LiBr–water absorption chiller.

state 9 is absorbed in the absorber (2) by a hygroscopic solution of LiBr entering concentrated (strong) at state 18 and leaving diluted (weak) at state 14. The exiting solution is pumped via heat exchanger (5) to desorber (4) by means of a small pump (not shown), requiring a small and negligible amount of work. The desorber serves to regenerate the solution from a weak to a strong state by applying to it heat from the solar-heated water stream 7–8, causing the desorption of water at state 10. The solution entering the desorber (4) at the weak state 15 is thus concentrated to state 17 and returns to the absorber (2) through the recuperative heat exchanger (5). The desorbed vapor at state 10 condenses in the condenser (3) to liquid state 12, then expands into the evaporator. The absorber and condenser are cooled by streams of cooling water 3–4 and 5–6 to reject the heats of absorption and condensation, respectively. As evident, the system operates between two pressure levels and interacts with heat sources/sinks at three temperature levels: low-temperature cooling in the evaporator, intermediate-temperature heat rejection in the absorber and condenser, and high-temperature (solar) heat supply in the desorber. These correspond to T_L, T_A, and T_H, respectively, in Fig. 2 and in the Carnot equivalent of Eq. (1).

From the previous description, it is clear that an absorption system employs a working fluid pair (absorbent/refrigerant), unlike the vapor compression heat pump, which employs a single fluid. The previously mentioned LiBr–water is the most common fluid pair in solar as well as other heat-powered absorption air-conditioning systems. Another fluid pair is water–ammonia, which generally requires higher firing temperatures and yields lower performance but can produce below-freezing temperatures and hence is suitable for refrigeration applications. Other working fluid pairs have been proposed but are rarely found in solar applications.

A great deal of experience is available on solar-powered absorption systems from demonstration projects carried out in the past three decades, most of which have employed single-effect LiBr–water chillers with low-temperature solar collectors. The first projects employed conventional machines originally designed for gas- or oil-fired operation. The heat to the desorber of these units was supplied by hot water from a solar heat source (e.g., flat plate, evacuated tube, or concentrating collectors). However, lowering the desorber temperature with respect to the original fuel-fired design—without making proper alterations in the machine—resulted in considerable downgrading of the entire device. Attempts were made to redesign the units and modify the desorber in particular, such that they could operate efficiently with solar-heated water. The first commercial absorption air conditioner built specifically for solar operation was the 3-TR Solaire unit by Arkla Industries based on an earlier gas-fired model. This unit is well-known among solar cooling engineers because it was used in a large number of demonstration projects. Its cost, however, was quite high for its small size. Soon thereafter, other manufacturers brought to the market new and improved absorption chillers based on LiBr–water and designed specifically for solar operation.

Two basic requirements must be considered by the designer of a solar absorption system that distinguish it from conventional fuel-fired systems. The first is the need to operate at a low temperature of the heat source: The temperatures available from flat-plate collectors are usually limited to below 90°C. The other is the need for a backup system that can supply the cooling load when insufficient or no energy is available from the sun. For the former requirement, a desorber design based on falling film instead of the conventional pool type has proven effective. Also, a solution preheater placed between the heat exchanger outlet and the desorber inlet was found to be effective in reducing the size and cost of the desorber and improving performance.

The problem of backup is more complicated. A good design must not only supply the full cooling demand at times of low or no sunshine but also make use of whatever solar radiation is available and take only the rest from primary sources. One possible approach is to heat the desorber of the same chiller both from the solar collectors and from an additional source of hot water (e.g., a fuel-fired boiler). However, the hot water coming from the boiler will normally interfere with the relatively colder water available from the collectors and hence drastically reduce the use of available solar radiation. Simulation results show that under conditions of low sunshine, the collectors connected in series with the backup boiler can turn into a heat sink instead of a heat source. Another possibility is to operate a second cooling machine powered by primary energy alongside the solar machine powered only from the sun, with the former supplying whatever load the latter cannot deliver. This requires installation of a complete additional unit, which would be idle a significant portion of the time. To solve the problem effectively, it has been proposed to incorporate an auxiliary desorber in the single-effect absorption

cycle. The auxiliary desorber is powered by the backup, whereas the original desorber is solar powered. The weak solution from the absorber goes first to the solar-powered desorber, where it is concentrated as much as possible with the available solar heat; it then proceeds to the auxiliary desorber and is concentrated further by means of heat from the backup source. The vapor from both desorbers is sent to the condenser. Other variations of this cycle have been proposed.

Experience from the numerous solar-powered air-conditioning demonstration projects with single-effect LiBr–water absorption systems has shown their main limitation to be cost, dominated by the solar part of the system. Single-effect absorption systems are limited in COP to approximately 0.7 and hence require a large collector area to supply the solar heat needed for their operation. This area could be reduced by employing systems with improved COP, which may be achieved with a higher temperature of the heat source, as indicated by Eq. (1). A single-stage absorption system is not suited to utilize a heat source at a temperature higher than approximately 100°C; unlike other heat-activated heat pumps that follow the Carnot trend, the absorption COP levels off with the increase in desorber temperature due to increased mixing and circulation losses. To take advantage of a high-temperature heat source, absorption systems must be configured in stages. The principle is to utilize the heat rejected from the condenser to power additional desorbers, thereby approximately doubling or tripling the amount of refrigerant extracted out of solution with no extra solar heat spent.

The principle of multistaging may be explained by referring to Fig. 4, which depicts a double-stage LiBr–water absorption chiller comprising two desorbers and two condensers serving one absorber/evaporator combination. The system may be regarded as an extension of the single-stage chiller of Fig. 3, with an additional condenser (7), desorber (8), heat exchanger (9), and expansion valve (10). As

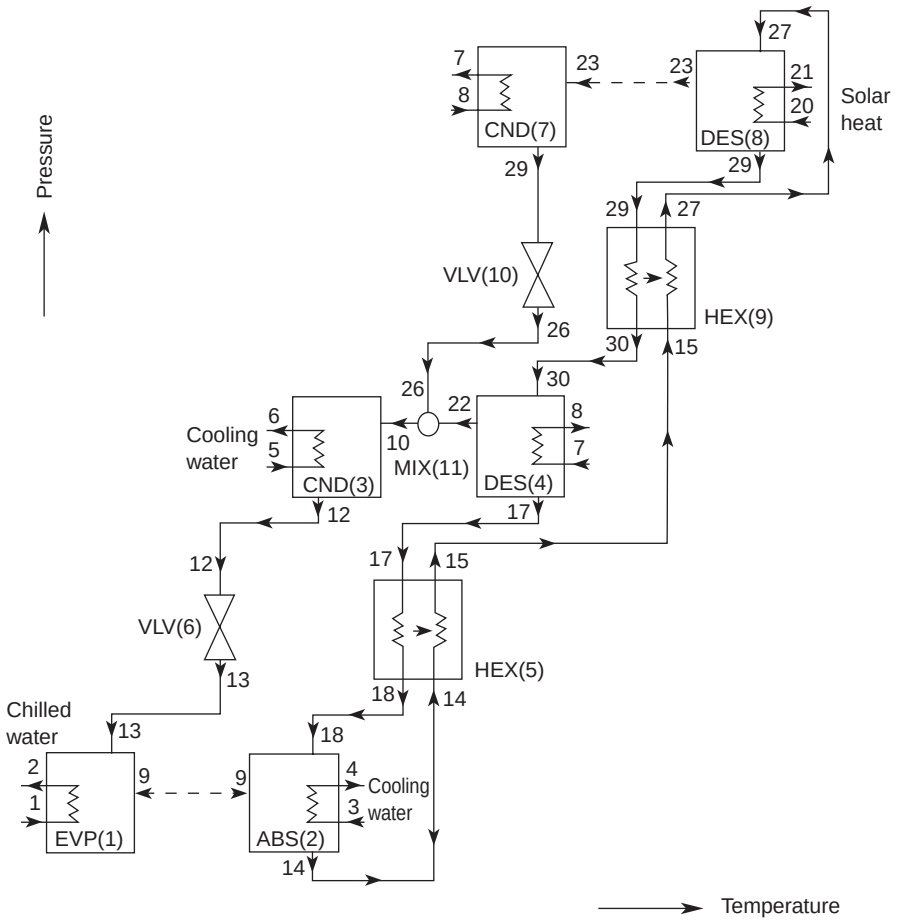


FIGURE 4 Schematic description of a double-effect LiBr–water absorption chiller.

evident, the system operates at three pressure levels and four temperature levels. The weak solution leaving the absorber (2) at state 14 is circulated to the two desorbers, connected in series, and regenerated in two stages, with each desorber doing part of the regeneration task. High-temperature solar heat is applied to the high-temperature desorber (8), whereas the low-temperature desorber (4) receives the reject heat from the high-temperature condenser (7) through the hidden loop 7–8. The condensate from both condensers expands into the evaporator and yields approximately double the cooling effect for the single-stage system of Fig. 3. The system is thus termed double-effect. The same principle may be extended with three desorbers and three condensers to give a triple-effect system when sufficiently high temperature of the heat source is available.

As mentioned previously, most solar-powered absorption cooling projects have utilized single-effect systems with low-temperature solar collectors. Due to developments in gas-fired absorption systems in recent years, mainly in the United States and Japan, higher COP LiBr–water chillers have become available: double-effect systems, available on the market, with COP in the range 1.0–1.2, and triple-effect systems, still under development but close to market, with estimated COP of approximately 1.7. These systems may be adapted to and employed in a solar-powered installation with high-temperature solar collectors. Figure 5 compares the performance of several multieffect chillers, showing the COP as a function of the solar heat supply temperature for typical single-, double-, and triple-effect chillers with the same size components under the same operating conditions: cooling water inlet at 30°C and chilled water outlet at 7°C. The corresponding Carnot performance curve, according to Eq. (1), is plotted for comparison. It is evident that for each absorption system, there is a minimum value of the heat supply temperature below which it does not perform at all. As this temperature increases, the COP increases sharply, then levels off to some asymptotic value. The reasons for this behavior are well understood and documented but are beyond the scope of this article. The single-effect system gives best results in the temperature range 80–100°C; above these temperatures, it is worthwhile to switch to the double-effect system, up to approximately 160°C, and then to the triple-effect system. Note, however, that with a double-effect system, if the solar-supplied temperature should drop below 100°C, the performance will drop sharply—below that of a single-effect system. It is therefore of utmost importance to incorporate a backup in a way that will ensure the supply of heat above the required minimum temperature.

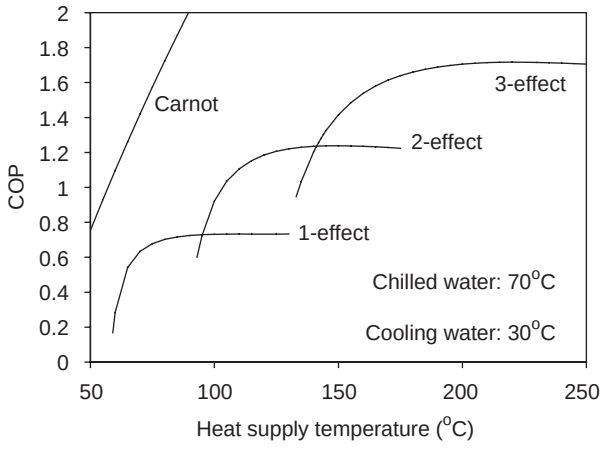


FIGURE 5 Coefficient of performance (COP) as a function of solar heat supply temperature for single-, double-, and triple-effect LiBr–water absorption chillers.

With the availability of high-temperature gas-fired systems, the following question arises: Due to their high COP, are multieffect systems more suitable for solar-powered cooling than their single-effect counterparts? It is attempted to answer this question by conducting a comparison between single-, double-, and triple-effect systems. The results of the comparison clearly indicate that the economics of solar-powered air-conditioning is dominated by the cost of the solar part of the system. This seems to show preference for the multieffect absorption systems, whose higher COP results in less collector area per kilowatt of cooling. Despite this fact, the higher cost of the high-temperature collectors often makes the total system cost higher than that of the low-temperature systems. The challenge before the solar industry is to reduce the price of high-temperature collectors, which will pave the way to employ the high-COP absorption cooling machines.

4. OPEN-CYCLE SYSTEMS

Open-cycle desiccant-type cooling systems represent a potentially promising alternative to other methods of solar-powered air-conditioning. Although the main thrust in research on novel closed-cycle absorption systems has been toward increasing the operating temperature in order to improve efficiency through multistaging, open-cycle absorption and desiccant systems have been developed for use with low-temperature heat sources such as flat-plate solar collectors. Relatively little work has been done in this

area in comparison with studies on the high-temperature alternative, despite the high cost of the latter and its various limitations: the need for sun tracking with concentrating solar collectors, their ability to use only the direct component of the solar radiation, and difficulty in storing high-temperature heat.

Desiccant systems are essentially open absorption cycles, utilizing water as the refrigerant in direct contact with air. The desiccant (sorbent) can be either solid or liquid and is used to facilitate the exchange of sensible and latent heat of the conditioned airstream. The term open is used to indicate that the refrigerant is discarded from the system after providing the cooling effect and new refrigerant is supplied in its place in an open-ended loop. In both liquid and solid desiccant types of systems, the process air is treated in a dehumidifier and goes through several additional stages before being supplied to the conditioned space. The sorbent is regenerated with ambient or exhaust air heated to the required temperature by the solar heat source.

The operation of the desiccant cycle is made possible by the property of desiccants of attracting and holding water vapor from the surrounding air at low temperatures and releasing water to air at high temperatures. The exchange of water vapor between air and desiccant depends on the relative magnitudes of water vapor pressure in the air and on the surface of the desiccant, respectively. The vapor pressure in the air is a function of the water mole fraction, per Dalton's law, and independent of temperature. The vapor pressure in the desiccant is a function of its temperature and water content: Vapor pressure increases with both temperature and water content. Thus, at low desiccant temperatures dehumidification of air will take place in a sorption process, whereas at high desiccant temperatures the same air could be used to regenerate the desiccant—that is, reduce its water content in a desorption process. A sorption process is always exothermic; that is, it generates heat due mainly to the condensation of vapor on the surface of the desiccant as well as to the heat of mixing (in the case of liquids) or heat of wetting (in the case of solids). Conversely, desorption is endothermic; that is, it requires heat input to separate and evaporate water from the desiccant.

A desiccant air-conditioning system in its simplest form consists of conditioning and regenerating cycles, as illustrated in Fig. 6. The conditioning cycle consists of a dehumidifying stage with simultaneous or subsequent removal of the heat of sorption. Warm

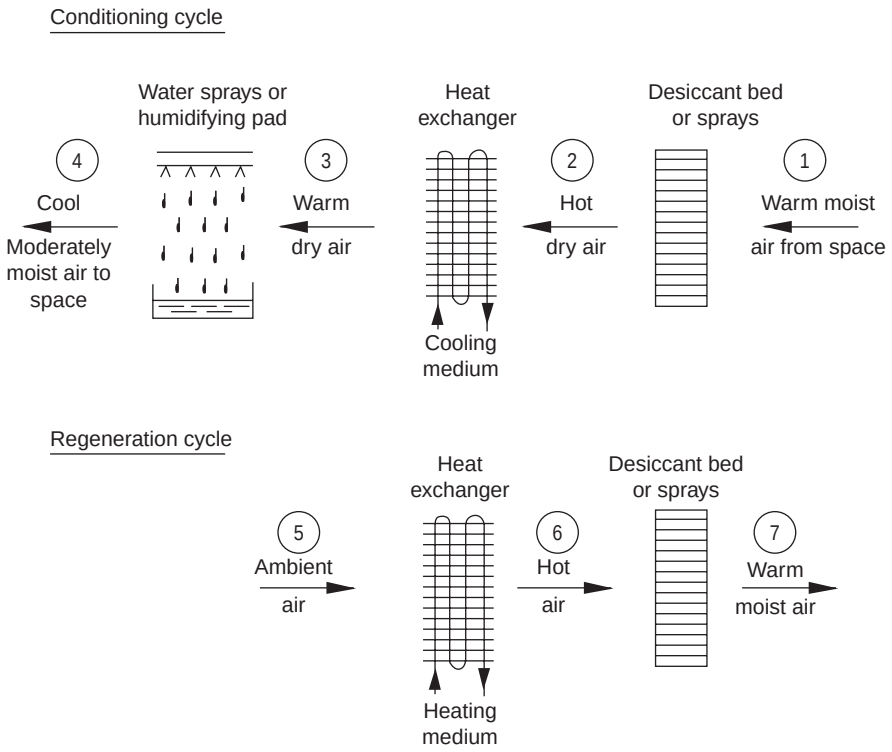


FIGURE 6 Principle of open-cycle desiccant air-conditioning.

and moist process air from the conditioned space or from the outside (state 1) is first brought into contact with an active desiccant. Water vapor is removed from the air and absorbed (in a liquid desiccant) or adsorbed (on a solid desiccant). This exothermic process causes an increase in the temperature of the air, whereas its humidity is reduced (state 2). The hot and dry air is then cooled to near ambient temperature, without change in water content (state 3), in a heat exchanger using a moderately cool medium (cooling tower water, exhaust air, or evaporatively cooled ambient air). If necessary, a humidifying stage 3–4 is added in which evaporative cooling is carried out. Sensible heat is changed into latent heat; that is, the air temperature is reduced while its water content is increased to slightly below the original value at state 1. This air thus has both sensible and latent cooling capabilities when released into the conditioned space.

The capacity of the desiccant to attract water vapor decreases with the increase in its water content, and at a certain point it has to be regenerated. In the regeneration cycle illustrated in Fig. 6, ambient air at state 5 is heated in a heat exchanger with solar-heated water (or equivalent) to state 6 and is brought into contact with the desiccant. Water vapor is transferred from the desiccant to the air in the desorption process 6–7. Warm and humid air at state 7 is exhausted and the dry desiccant is ready to be used again in the conditioning cycle.

Desiccant systems possess several advantages relative to their closed-cycle counterparts: They operate at ambient pressure and not in a vacuum or at an elevated pressure; heat and mass transfer between the air and the desiccant take place in direct contact; and both cooling and dehumidification of the conditioned air may be provided, in variable quantities, to fit the load in the conditioned space. The latent load is handled without the necessity to cool the air below its dew point; consequently, reheat is never required. The following are disadvantages: They have a low COP due to inherently inefficient regeneration; relatively large air volumes must be pumped, leading to potentially high parasitic losses; and contamination of the desiccant by dirt and dust contained in the air may require its replacement after some period of operation.

Solid desiccant systems employ adsorbents (rather than absorbents) such as silica gel or various types of molecular sieves. The humidity is adsorbed on the surface of the desiccant, which is highly porous. Since the solid desiccant cannot be circulated by pumping, these systems usually employ a rotary bed carrying the sorbent material, referred to as a desiccant wheel. The wheel features continuous operation of the system by rotating between the process and regenerating airstream. A typical desiccant wheel is shown in Fig. 7. It consists of a rotary bed or drum made of an inert material, corrugated or extruded, forming axial flow channels for the air. The wheel is packed or impregnated with the desiccant material, covering the walls of the channels. The wheel in Fig. 7 forms part of a natural gas-boosted solar air-conditioning system, employing molecular sieves as desiccant. The wheel is divided into four sections: the sorption section, through which the process airstream flows; the low-stage regeneration section, in which desorption of the desiccant by means of solar-heated air takes place; the high-stage regeneration section, in which desorption is boosted by flue gas from a burner; and a purge section in which cooling of the matrix, as it leaves the regeneration section, is carried out. The purge section ensures that the desiccant entering the process airstream is cooled sufficiently for dehumidification to take place, thus reducing the average process airstream humidity and temperature.

Solid desiccants usually allow a higher degree of dehumidification than the liquids but require a higher temperature for their regeneration. They usually operate according to the cycle depicted in Fig. 6. The past decade witnessed a significant increase in solid desiccant systems on the market, usually for gas-fired operation in applications requiring dehumidification in addition to conventional cooling, such as in supermarkets. Solar-powered operation with solid desiccants normally requires higher temperatures than required for liquids,

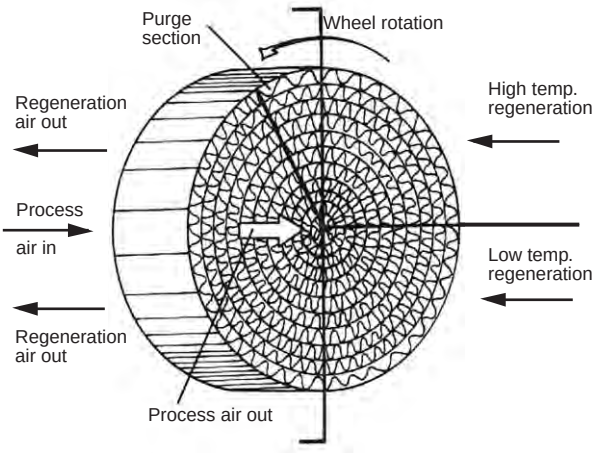


FIGURE 7 Diagram of a typical desiccant wheel.

possibly beyond the range of flat-plate collectors. The regeneration air may be heated directly in air-heating collectors or indirectly using water as a heat-carrying medium.

Liquid absorbents of water comprising aqueous solutions of lithium bromide, lithium chloride, or calcium chloride are employed in the open systems using a cycle similar to that of a closed absorption system. In comparison to the system shown in Fig. 3, the evaporator (1) and condenser (3) are eliminated. The absorber is replaced by a dehumidifier, which is designed differently to bring into contact the conditioned air with the sorbent solution in order to absorb the humidity directly from the airstream. Similarly, the desorber is replaced by a regenerator, which again brings the sorbent solution to be regenerated into direct contact with ambient or exhaust air. The net result of the operation of this device is the dehumidification of the conditioned airstream, which handles the latent load. To handle the sensible load, the temperature of the conditioned air must be lowered either by an external cooling device or by evaporating water into the now dry airstream using an evaporative cooler in place of the evaporator (1) of Fig. 3.

The basic solar desiccant system as depicted in Fig. 6 can be improved when using liquid desiccants by the addition of a number of heat exchangers for better energy economy. Such a system is shown schematically in Fig. 8. The two main components, the dehumidifier and the regenerator, have a similar design, suitable for contacting the sorbent solution with air. Both are vertical, packed towers, with the solution sprayed over the packing and flowing down and air brought in at the bottom and flowing up, in counterflow to the solution. A cooling coil inserted in the dehumidifier and employing cooling tower water serves to remove the heat of sorption during dehumidification. Similarly, a heating coil in the regenerator supplies the heat of desorption during regeneration. This is a deviation from the schematic shown in Fig. 6, in which the dehumidifier and regenerator are adiabatic; one of the advantages of a liquid desiccant system is that it allows nonadiabatic dehumidification and regeneration processes, as in closed absorption systems, which makes these processes more efficient. In the system of Fig. 8, outside (fresh) air is dehumidified in the dehumidifier and then sent to the conditioned space. Building exhaust air is preheated in the air–air heat exchanger by regenerator exhaust air and used for solution regeneration. Building exhaust air is preferred to outside air for this purpose since it is dryer. Figure 8 also shows a liquid–liquid heat exchanger between the weak and cool solution leaving the dehumidifier and the strong and hot solution leaving the regenerator for recovering heat from the latter to the former; this is again similar to the recuperative heat exchanger of a closed absorption system. Figure 8 further shows the solar part of the system, consisting of collectors and storage, for supplying heat to the regenerator.

Liquid desiccant systems have many of the same advantages of closed-cycle absorption: They allow

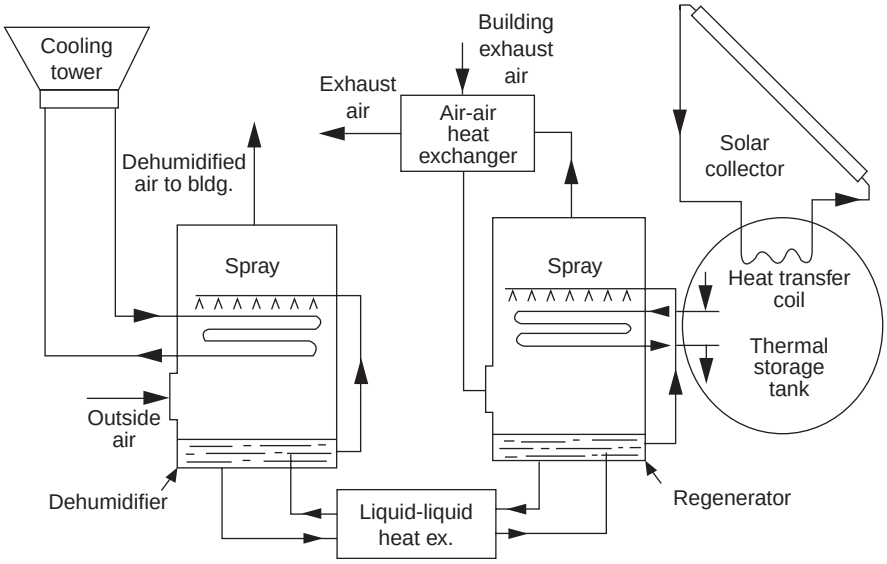


FIGURE 8 Diagram of a liquid desiccant system for dehumidification.

internal heat recovery from the strong and hot solution to the weak and cool solution by means of a recuperative heat exchanger; circulating the sorbent from one unit in the system to the next is easily done by pumping the liquid; and pressure drop in the liquid–air contactors, dehumidifier and regenerator, is relatively low, reducing parasitic losses. The liquid desiccant may be filtered to remove dust and dirt picked up from the air. For solar-powered operation, desorption is possible in a special regenerating collector exposing the desiccant directly to the sun, which can make the regeneration process simpler and more efficient. It also makes it possible to store cooling capacity by storing concentrated liquid desiccant. One disadvantage is the potential carryover of sorbent in the air, which may lead to sorbent loss and contamination of the conditioned space. There are relatively few liquid desiccant systems available on the market. A limited number of solar-powered air-conditioning systems employing liquid desiccants, mostly experimental, have been constructed during the past three decades.

A promising open-absorption heat pump cycle makes it possible to utilize the solar heat at relatively low temperatures to produce both chilled water and cold and dehumidified air in variable proportions, as required by the load. The cycle, referred to as dehumidifier–evaporator–regenerator (DER), employs environmentally benign liquid desiccants in direct contact with the air. The system typically treats the make-up supply air to handle the latent load and supplies chilled water to handle the sensible load independently. A schematic of the cycle is shown in Fig. 9. It consists of six major components, indicated by roman numerals: an indirect contact evaporative cooler (I), an air dehumidifier or absorber (II), a solution regenerator or desorber (III), two air–air heat exchangers (IV, V), and a solution–solution heat exchanger (VI). Ambient air entering the absorber at state 7 is brought into contact with a strong (concentrated) absorbent solution entering the unit at state 16 and leaving it at state 17. Water vapor is removed from the airstream by being absorbed into the solution stream. The heat of condensation and dilution Q_A released during this absorption process is rejected to a stream of cooling water or ambient air (states 3 and 4). Dehumidified air leaving the absorber at state 8 is drawn through an air–air heat exchanger (IV) into the evaporative cooler. By evaporating externally supplied water into the airstream, the heat Q_E is removed from the chilled medium (states 1 and 2) at a low temperature. The air leaves the evaporative cooler at state 10 humid,

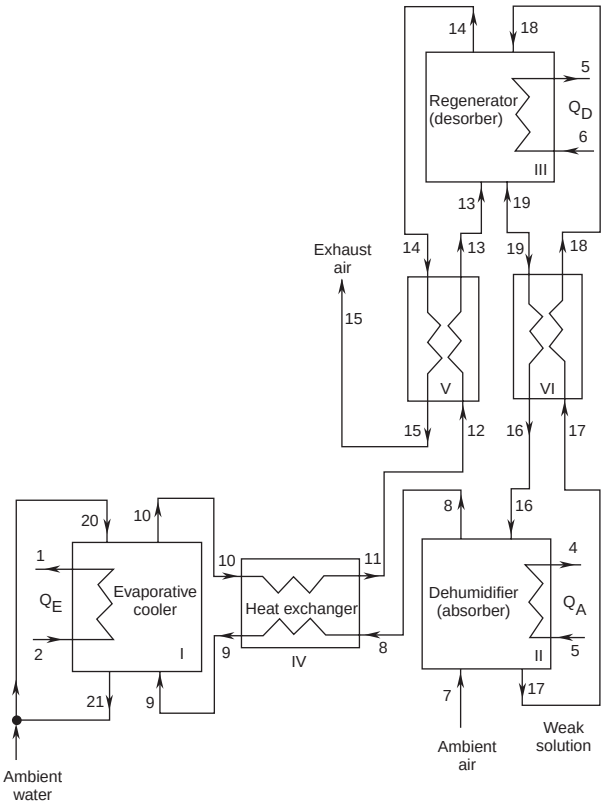


FIGURE 9 Schematic description of the DER open-absorption chiller.

but at a lower temperature than at its inlet, and passes again through the heat exchanger (IV), where it provides precooling for the airstream 8–9 in order to allow for a low cooling temperature. After being preheated in the second air–air heat exchanger (V) by exhaust air, the airstream enters the regenerator at state 13, where it serves to reconcentrate the weak absorbent solution (state 18). The heat of desorption Q_D required to remove water vapor from the solution is provided by an external (solar) heat source (states 5 and 6). The exhaust air leaves the regenerator (state 14), preheats the airstream 12–13 in the air–air heat exchanger (V), and is rejected into the environment. The solution–solution heat exchanger (VI) facilitates preheating of the weak solution (states 17 and 18) and precooling of the strong solution (states 16 and 19).

The advantages of open-cycle DER absorption chillers compared to closed-cycle systems are evident from the previous description. First, the number of main components is reduced by one by transferring condensation of the refrigerant from a condenser to the environment. Second, capital-intensive pressure-sealed units are avoided because the whole system

can be operated at atmospheric pressure. Third, efficient utilization of very low heat source temperatures is possible. The open-cycle DER chiller can be operated with any heat source temperature that is only slightly above the heat rejection temperature. This is particularly favorable for the efficient utilization of heat sources with varying temperatures, such as solar heat from flat-plate collectors. Chilled water at state 21 and/or dehumidified and cooled air at state 9 are the products of the system and may be extracted at varying quantities as required by the air-conditioning application.

5. CONCLUSION

Current trends in solar air-conditioning show a clear preference for absorption cycles over other heat-powered cycles, in two alternative directions: closed-cycle and open-cycle systems. Whereas the main thrust in research on closed-cycle absorption systems has been toward increasing the operating temperature in order to improve efficiency through multi-staging, open-cycle absorption (desiccant) systems have been developed for use with low-temperature heat sources such as flat-plate solar collectors. With both alternatives, the economics of solar cooling are dominated by the solar part of the system—collectors and storage.

Open-cycle systems hold promise for reducing the cost of the solar part of the system and for providing quality air-conditioning with independent temperature and humidity control. Both liquid and solid desiccants are already penetrating the market. Systems such as the DER open-absorption heat pump cycle form the basis for promising cooling systems capable of efficiently exploiting low-temperature solar heat as well as other, similar sources of low-grade heat.

SEE ALSO THE FOLLOWING ARTICLES

Materials for Solar Energy • Refrigeration and Air-Conditioning • Solar Cells • Solar Cookers • Solar Distillation and Drying • Solar Energy, History of • Solar Fuels and Materials • Solar Heat Pumps • Solar Ponds • Solar Thermal Energy, Industrial Heat Applications • Solar Thermal Power Generation • Solar Water Desalination

Further Reading

Alefeld, G. (1982). Regeln für den entwurf von mehrstufigen absorbermaschinen (Rules for the design of multistage absorption machines). *Brennst-Warme-Kraft* **34,** 64–73.

American Society of Heating, Refrigeration and Air-Conditioning Engineers (1998). Absorption cooling, heating and refrigeration equipment. *In* "ASHRAE Handbook," refrigeration volume. American Society of Heating, Refrigeration and Air-Conditioning Engineers, Atlanta, GA.

American Society of Heating, Refrigeration and Air-Conditioning Engineers (2000). Solar energy equipment. *In* "ASHRAE Handbook," HVAC systems and equipment volume. American Society of Heating, Refrigeration and Air-Conditioning Engineers, Atlanta, GA.

Grossman, G. (2002). Solar-powered systems for cooling, dehumidification and air-conditioning. *Solar Energy* **72,** 53–62.

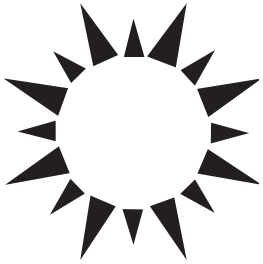
Henning, H. (ed.). (2003). "Solar Assisted Air Conditioning in Buildings—Handbook for Planners." Springer-Verlag, Vienna.

Hellmann, H.-M., and Grossman, G. (1995). Simulation and analysis of an open-cycle dehumidifier–evaporator–regenerator (DER) absorption chiller for low-grade heat utilization. *Int. J. Refrigeration* **18,** 177–189.

Herold, K. E., Radermacher, R., and Klein, S. A. (1996). "Absorption Chillers and Heat Pumps." CRC Press, New York.

Kreith, F., and Kreider, J. F. (1978). "Principles of Solar Engineering." McGraw-Hill, New York.

Solar Detoxification and Disinfection



S. MALATO-RODRÍGUEZ
Plataforma Solar de Almería (CIEMAT)
Tabernas, Almería, Spain

1. **Solar Photochemistry**
2. **Photodegradation Principles: Photocatalysis**
3. **Fundamental Parameters in Photocatalysis**
4. **Solar Detoxification and Disinfection Technology**
5. **Solar Detoxification and Disinfection Applications**
6. **Conclusion**

Glossary

biodegradable compounds Compounds that can be metabolized by microorganisms.

compound parabolic concentrator (CPC) Static collectors with a reflective surface following an involute around a cylindrical reactor tube.

conduction band A vacant or only partially occupied set of many closely spaced electronic levels resulting from an array of a large number of atoms forming a system in which the electrons can move freely or nearly so. This term is usually used to describe the properties of semiconductors.

electron/hole Negative and positive charge generated in an illuminated semiconductor.

mineralization Conversion of organic compounds to inorganic ones. Inorganic compounds are usually carbon dioxide, water, and acids (nontoxic substances).

photocatalysis Acceleration of a photoreaction by the presence of a catalyst.

photoreactor Reactor where a photochemical reaction takes place.

semiconductor Material with an appropriate energetic separation between its valence and conduction bands that can be surpassed applying external energy. Metal oxides and sulfides represent a large class of semiconductor materials suitable for photocatalytic purposes.

solar detoxification Photochemical processes driven by solar photons able to destroy toxic substances. The process is based on the generation of hydroxyl radicals in water.

total organic carbon Quantity of organic compounds dissolved in water measured as pure carbon.

The solar heterogeneous photocatalytic detoxification process consists of utilizing the near-ultraviolet (UV) part of the solar spectrum to photoexcite a semiconductor catalyst in the presence of oxygen. In these circumstances, oxidizing species, either bound hydroxyl radical ($^{\bullet}OH$) or free holes, which attack oxidizable contaminants, are generated, thus producing a progressive breaking of molecules yielding CO_2, H_2O, and dilute inorganic acids. The most commonly used catalyst is the semiconductor TiO_2 (a cheap, nontoxic, and abundant product). The solar homogeneous photocatalytic detoxification process (photo-Fenton) is based on the production of $^{\bullet}OH$ radicals by Fenton reagent. This is an extension of the Fenton process, which takes advantage of UV-visible light irradiation at wavelength values higher than 300 nm. During the past decade, thousands of studies and patents related to the heterogeneous photocatalytic removal of toxic and hazardous compounds from water and air have been published, and applications and target compounds are numerous. Where medium to high solar radiation is available, solar detoxification is useful for treating water contaminants with several hundreds of milligrams per liter of nonbiodegradable contaminants and for disinfecting water. This process has proven promising for the treatment of contaminated water, air, and soil. Nevertheless, the specific solar technology employed is not commercially available. This article reviews the use of sunlight to produce $^{\bullet}OH$ radicals. The experimental systems necessary for performing pilot plant-scale solar photocatalytic experiments are described. The basic components of these pilot plants and the fundamental parameters related to solar photocatalysis reactions are outlined. A description is given of how solar photocatalysis may become a significant segment of the wastewater treatment technologies related to disinfection and the degradation of very persistent toxic compounds.

 © 2004 Elsevier Inc. All rights reserved.

1. SOLAR PHOTOCHEMISTRY

A specific characteristic of solar photochemical processes is that the photons from the solar spectrum must be absorbed by some component(s) of the reacting system, transferring their energy to the chemical system. Solar photons can be directly absorbed by reactants (direct photochemistry) and/or by a catalyst or sensitizer (photocatalytic processes). Solar detoxification and disinfection are included within photocatalysis.

Photochemical processes use the intrinsic energy of photons (i.e., wavelength) from the solar irradiation to provoke specific reactions. The energy of a photon is a function of its wavelength, according to Eq. (1):

$$E = \frac{hc}{\lambda}, \tag{1}$$

where h is Planck's constant (6.626×10^{-34} Js), c is the speed of light, and λ is the wavelength. Approximately 4 or 5% of the sunlight at the earth's surface is in the near-ultraviolet (UV) region (wavelengths from 300 to 400 nm), approximately 45% is in the visible range (400–760 nm), and the rest is in the near-infrared (IR) and IR regions. Solar photons from the visible and near-UV spectrum are energetic enough to be used in photochemical processes. Infrared photons are normally useless in photochemistry and can be considered as waste heat in chemical processes.

2. PHOTODEGRADATION PRINCIPLES: PHOTOCATALYSIS

A wide range of contaminants to a varying degree resist photolytic, biological, and chemical degradation. In such cases, it is necessary to adopt much more effective systems: air stripping, adsorption on granulated activated carbon, incineration, ozone, and oxidation. Obviously, methods based on chemical destruction, when properly developed, provide complete pollutant abatement, whereas those that only perform phase separation leave a consequent problem of final disposal. Air stripping and adsorption on granulated activated carbon, merely transferring toxic materials from one medium to another, is not a long-term solution to the problem of loading hazardous waste on the environment. Incineration is capable of converting toxic compounds to carbon dioxide, water, and inorganic acids, but negative public perception has slowed and very often prevented its implementation. Concerns over emissions of undestroyed components and organic products of incomplete combustion have caused incineration to be a source of continuing controversy. Ozone is not very often able to completely destroy pesticides to harmless compounds.

Chemical oxidation aims at the mineralization of the contaminants into carbon dioxide, water, and inorganics. Much research has addressed this goal in the past decade, indicating the prominent role of a special class of oxidation techniques defined as advanced oxidation processes (AOPs), which usually operate at or near ambient temperature and pressure. AOPs, although making use of different reaction systems, are all characterized by the same chemical feature: production of OH radicals ($^{\bullet}OH$). This radical is strongly oxidative, much more so than other traditional oxidants (e.g., ozone, hydrogen peroxide, chloride dioxide, and chlorine). These can then be used to chemically decompose a pollutant into harmless end products. For example, in aromatic compounds, the aromatic part is hydroxylated and successive steps in oxidation/addition lead to ring opening. The resulting aldehydes and carboxylic acids are decarboxylated and finally produce CO_2. Methods based on UV, H_2O_2/UV, O_3/UV, and H_2O_2/O_3/UV combinations utilize photolysis of H_2O_2 and ozone to produce the hydroxyl radicals. In addition, other methods, such as heterogeneous photocatalysis or homogeneous photo-Fenton, are based on the use of a wide band gap semiconductor or the addition of H_2O_2 to Fe^{2+} salts and by irradiation with UV-visible light. Both processes are of special interest since sunlight can be used. The main problem with the AOPs is their high cost due to the use of expensive reactants (i.e., H_2O_2) and energy consumption (generation of UV radiation). Therefore, the future applications of these processes may be improved diminishing reactants and energy consumption through the use of catalysis and solar energy.

2.1 Definitions

The term photocatalysis implies the combination of photochemistry with catalysis. Both light and catalyst are necessary to achieve or to accelerate a chemical reaction. Photocatalysis may be defined as the acceleration of a photoreaction by the action of a catalyst. It also refers to a general label to indicate that light and a substance (the catalyst or initiator) are necessary entities for influencing a reaction. Heterogeneous processes employ semiconductor slurries for catalysis, whereas homogeneous photochemistry is used in a single-phase system.

In the case of homogeneous photocatalytic processes, the interaction of a photon-absorbing species (transition metal complexes, organic dyes, or metalloporphyrines), a substrate (e.g., the contaminant), and light can lead to a chemical modification of the substrate. The photon-absorbing species is activated and accelerates the process by interacting through a state of excitation. In the case of heterogeneous photocatalysis, the interaction of a photon produces the appearance of electron/hole pairs, the catalyst being a semiconductor (e.g., TiO_2 and ZnO). In this case, the excited electrons are transferred to the reducible specimen at the same time that the catalyst accepts electrons from the oxidizable specimen that occupies the holes. In both directions, the net flow of electrons is null and the catalyst remains unaltered.

2.2 Heterogeneous Photocatalysis

When a semiconductor is in contact with a liquid electrolyte solution containing a redox couple, charge transfer occurs across the interface to balance the potentials of the two phases. An electric field is formed at the surface of the semiconductor. This means that the two charge carriers should react at the semiconductor/electrolyte interface with the species in solution. The semiconductor-mediated redox processes involve electron transfer across the interface. When electron/hole pairs are generated in a semiconductor particle, the electron moves away from the surface to the bulk of the semiconductor as the hole migrates toward the surface. If these charge carriers are separated fast enough, they can be used for chemical reactions at the surface of the photocatalyst (i.e., for the oxidation or reduction of pollutants).

Metal oxides and sulfides represent a large class of semiconductor materials suitable for photocatalytic purposes. Table I lists selected semiconductor materials, which have been used for photocatalytic reactions, together with band gap energy required to activate the catalyst. The table also indicates the wavelength of radiation required to activate the catalysts.

Whenever different semiconductor materials have been tested under comparable conditions for the degradation of the same compounds, TiO_2 has generally been demonstrated to be the most active. TiO_2's strong resistance to chemical and photocorrosion, its safety, and its low cost limit the choice of convenient alternatives. Furthermore, TiO_2 is of special interest since it can use natural (solar) UV (Table I). Other semiconductors are degraded during the repeated catalytic cycles involved in heterogeneous photocatalysis:

TABLE I

Selected Properties of Several Semiconductors

Material	Band gap (eV)	Wavelength corresponding to band gap (nm)
SnO_2	3.9	318
ZnS	3.7	336
$SrTiO_3$	3.4	365
$BaTiO_3$	3.3	375
ZnO	3.2	380
TiO_2	3.0	390
WO_3	2.8	443
CdS	2.5	497
GaP	2.3	540
Fe_2O_3	2.2	565
CdO	2.1	590
CdSe	1.7	730
GaAs	1.4	887

$$TiO_2 \xrightarrow{h\nu} e^- + h^+ + TiO_2 \quad (2)$$

$$e^- + h^+ + TiO_2 \rightarrow TiO_2 + \text{heat and/or } h\nu' \quad (3)$$

$$(TiO^{IV} - O^{2-} - Ti^{IV}) - OH_2 + h^+_{BV} \rightarrow (TiO^{IV} - O^{2-} - Ti^{IV}) - OH^{\bullet} + H^+ \quad (4)$$

$$O_{2(ads)} + e^-_{BC} \rightarrow O^{-\bullet}_{2(ads)}. \quad (5)$$

The first effect, after absorption of near-UV radiation ($\lambda < 390$ nm), is the generation of electron/hole pairs, which are separated between the conduction and valence bands (Eq. 2). In order to avoid recombination of the pairs generated (Eq. 3), if the solvent is oxidoreductively active (water) it also acts as a donor and acceptor of electrons. Thus, on a hydrated and hydroxylated TiO_2 surface, the holes trap $^{\bullet}OH$ radicals linked to the surface (Eq. 4). In any case, it should be emphasized that even trapped electrons and holes can rapidly recombine on the surface of a particle (Eq. 3). This can be partially avoided through the capture of the electron by preadsorbed molecular oxygen, forming a superoxide radical (Eq. 5). The oxidative species formed (particularly the hydroxyl radicals) react with the majority of organic substances.

2.3 Homogeneous Photocatalysis

Although the oxidizing power of the combination of Fe^{2+} and H_2O_2 (i.e., the Fenton reaction; Eq. 6) has

been known for more than 100 years, the fact that the reaction can be enhanced by UV-visible irradiation ($\lambda < 580$ nm) was determined only recently. The so-called photo-Fenton reaction (Eq. 7) produces additional hydroxyl radicals and leads to a reduction of the photocatalyst by the irradiated light. Under these conditions, the photolysis of Fe^{3+} complexes allows Fe^{2+} regeneration, and the occurrence of Fenton reactions due to the presence of H_2O_2 and iron may be considered a true catalyst. The main advantage of the photo-Fenton process is the light sensitivity up to a wavelength of 580 nm. The depth of light penetration is high and the contact between pollutant and oxidizing agent is close since a homogeneous solution is used.

$$Fe^{2+} + H_2O_2 \rightarrow Fe^{3+} + OH^- + {}^{\bullet}OH \qquad (6)$$

$$Fe^{3+} + H_2O + h\nu \rightarrow Fe^{2+} + H^+ + {}^{\bullet}OH \qquad (7)$$

$$[Fe^{3+}L_n] + h\nu \rightarrow [Fe^{2+}L_{n-1}] + L^{\bullet}. \qquad (8)$$

Only a very simplified description is given in Eqs. (6–8). Note that reaction 8, which involves organic ligands L (e.g., RCOO−, RO−, and RNH+), is more efficient than the oxidation of organic substrates via $^{\bullet}OH$. Therefore, usually contaminants degradation is getting faster during the reaction since the more inert hydrocarbons are first oxidized by $^{\bullet}OH$ to yield substances containing several hydroxyl or carboxyl groups that are able to participate in reactions thereafter.

3. FUNDAMENTAL PARAMETERS IN PHOTOCATALYSIS

3.1 Influence of Oxygen and pH

Whatever the formation pathway, it is well-known that O_2 and water are essential for detoxification. There is no degradation in the absence of either. The pollutants are usually organic; therefore, Eq. (9) summarizes the overall process. Given the reaction stoichiometry of this equation, there is no photomineralization unless O_2 is present. Oxygen is necessary for complete mineralization and does not seem to be competitive with other reactives during the adsorption on TiO_2 since the places where oxidation takes place are different from those of reduction. The concentration of oxygen also affects the reaction rate, but the difference between using air ($pO_2 = 0.21$ atm) or pure oxygen ($pO_2 = 1$ atm) is not drastic. Oxygen also intervenes, governing the efficiency of a photocatalytic oxidative degradation, thus minimizing electron/hole recombination by capturing the photogenerated electron (see Eq. 5).

$$\text{Organic pollutant} + O_2 \xrightarrow{\text{photocatalysis}} CO_2 + H_2O + \text{mineral acids.} \qquad (9)$$

The oxide/electrolyte interface has an electrical surface charge that strongly depends on the pH of the medium. The pH of the aqueous solution significantly affects TiO_2, including the charge of the particle and the size of the aggregates it forms. The zero surface charge (PZC) yields zero electrostatic surface potentials that cannot produce the interactive rejection necessary to separate the particles within the liquid. This induces a phenomenon of aggregation and TiO_2 clusters become larger. This effect is clearly related to the capability of the suspension to transmit and/or absorb light. Furthermore, larger clusters sediment more quickly than small particles; thus, the agitation necessary to maintain perfect homogeneity must be more vigorous. The PZC for TiO_2 is approximately 7. Above and below this value, the catalyst is negatively or positively charged according to $TiOH \geq 80\%$ when $3 < pH < 10$, $TiO^- \geq 20\%$ when $pH > 10$, and $TiOH_2^+ \geq 20\%$ when $pH < 3$. Under these conditions, the photocatalytic degradation of ionizable organic compounds is affected by the pH. At first sight, a very acidic solution appears to be detrimental and a very basic solution to be favorable since the variations are modest or nonexistent around neutrality. In many cases, a very important feature of photocatalysis is not taken into account when it is to be used for decontamination of water: That is, during the reaction, a multitude of intermediate products are produced that may behave differently depending on the pH of the solution. To use only the rate of decomposition of the original substrate could yield an erroneous pH as the best for contaminant degradation. Therefore, a detailed analysis of the best pH conditions should include not only the initial substrate but also the rest of the compounds produced during the process.

3.2 Catalyst Concentration Influence

Incident radiation on the reactor and length of path of light inside the reactor are fundamental in determining optimum catalyst concentration. If the radiation comes from a source outside the photoreactor (as in a reactor illuminated by solar radiation) and the path length is short (1–2 cm maximum), maximum reaction rate is obtained with 1 or 2 g/liter

of TiO_2. If the path length is several centimeters long, the appropriate catalyst concentration is several hundred milligrams per liter. The major difference between considering the TiO_2 from dry powder and TiO_2 particles suspended in an aqueous medium involves the mean size of the particles. If two samples with the same catalyst concentration but that are dispersed with different protocols (ultrasounds, stirring, etc.) are compared; sonicated samples are found to be more efficient. Although in stirred samples there are fewer TiO_2 clusters, larger clusters screen light better than small ones. It has been demonstrated that efficiencies are better with small particles than with larger particles only when reactor diameter and catalyst concentration are optimized to allow UV photons to penetrate along the entire photoreactor path length. In this case, it is not possible to talk about particle radius because TiO_2 powder is irregularly shaped, and when it is dispersed in an aqueous medium the "particles" (300–600 nm) are clusters of primary particles (20–40 nm), not spherical or monodisperse. This leads to the conclusion that light extinction in colloidal suspensions is a determining parameter for solar photoreactor design. It should also be recalled that small particle sizes cause additional problems for catalyst separation after photocatalytic treatment. The best inner reactor diameter for solar photocatalytic applications is in the range of a few centimeters with a few hundred milligrams of TiO_2 per liter. The area:volume ratio is also a crucial parameter, and if this ratio is optimized the reactor efficiency would also be increased.

Because iron acts as a catalyst in the photo-Fenton reaction, the degradation rate should increase with the increasing amount of iron added, but it reaches a maximum at the optimum iron concentration. Further addition of iron does not increase the degradation rate. This threshold is dependent on the light source, the thickness of the irradiated layer, and the type of wastewater because internal light filtering may occur. For solar reactors, it is therefore necessary to determine the optimum iron concentration experimentally—that is, the minimum concentration at which the maximum reaction rate is obtained. When catalyst concentration is very high, after traveling a certain distance on an optical path length, turbidity impedes further penetration of light in the reactor. The percentage of photons absorbed by the suspension and the percentage of photons scattered is a very complex problem that cannot be solved experimentally but must be experimentally estimated.

3.3 Influence of Initial Contaminant Concentration

The degradation rate observed for an organic substrate follows saturation behavior. After a certain concentration is achieved, the rate increases very little and in some cases a decrease is observed. If possible, the optimum contaminant concentration in water before the photocatalytic treatment should permit the maximum reaction rate. Due to the kinetics of the photocatalytic reaction, the rate decreases considerably when contaminant concentration is low. The photocatalytic disappearance of contaminants with TiO_2 follows apparent first-order kinetics when the initial concentration is low enough and no catalyst saturation occurs. It should be emphasized that contaminants decompose, giving rise to intermediates, which could also be competitive on the surface of the TiO_2. Their concentration varies throughout the reaction until their mineralization, and thus Eq. (10) (based on the Langmuir–Hiselwwod model) may describe the kinetics:

$$r = \frac{dC}{dt} = \frac{krKC}{1 + KC + \sum_{i=1}^{n} K_i C_i (i = 1,\ n)}, \qquad (10)$$

where k_r is the reaction rate constant, K is the reactant adsorption constant, C is the concentration at any time, K_i is the adsorption constant of the intermediates, and C_i is their concentration at any time. When C_0 is low enough, Eq. (10) can be simplified ($1 + KC + \Sigma \ldots = 1$) to an order one-reaction rate equation ($r = k_{ap}C$). Likewise, at higher concentrations Eq. (10) can be simplified by adjusting it to a zero-order one: $(C_0 - C) = k_r t$.

During the treatment of real wastewaters, since the reaction is not expected to follow simple models such as first- or zero-order kinetics, very often overall reaction rate constants cannot be calculated. Total organic carbon (TOC) analysis is highly recommended for this because of its versatility and, at the same time, its suitability for "on-line" treatment plant installations. Mineralization (TOC disappearance) only occurs when the last step of the oxidation process yields CO_2. The maximum gradient of the mineralization curve, $TOC = f(t)$, which is the gradient of the tangent at the inflection point ($r_{TOC,\ 0}$), is employed to obtain a practical point of comparison for different experiments since it has a zero-order rate constant (mg/liter/min) and therefore appears easy to handle. This parameter ($r_{TOC,\ 0}$) is referred to as the maximum rate.

Once the optimum initial concentration for degradation is known, a model for predicting plant

behavior is necessary. This model should enable one to calculate the solar collector area required for treating water contaminated with different amounts of pesticides. A developed model, which has the analytic form of a Langmuir–Hinshelwood equation, has been preferred for fitting experimental data in large solar photocatalytic plants by approximate solution of the general photocatalytic kinetic system. With these considerations, the rate of TOC disappearance is given by

$$r_{TOC} = \frac{\beta_1[\mathrm{TOC}]}{\beta_2 + \beta_3[\mathrm{TOC}]}. \tag{11}$$

Experimental results can be used to calculate the constants (β_i). Equation (12) enables one to predict TOC degradation as a function of the initial TOC and of the available time (or radiation energy) and, reciprocally, to predict the incident energy on the reactor necessary to reach a specific degree of mineralization. Nevertheless, UV radiation data for the final plant location must be available.

$$\frac{1}{\beta_1}\left\{\beta_2 \ln\left(\frac{[\mathrm{TOC}]_0}{[\mathrm{TOC}]}\right) + \beta_3([\mathrm{TOC}]_0 - [\mathrm{TOC}])\right\} = t. \tag{12}$$

3.4 Influence of Radiant Flux

Since 1990, there has been clarification of the kind of solar technology that should be involved in detoxification and disinfection. The question was whether concentrating the radiation is really necessary for the photocatalysis technology and whether a nonconcentrating collector can be as efficient as concentrating ones. Initially, it was thought that concentrating collectors were the ideal alternative and, in fact, the first large pilot plants use them. However, their high cost and the fact that they can only operate with direct solar radiation (implying their location in highly insolated areas) led to consideration of static nonconcentrating collectors. Early engineering-scale tests with solar systems were based on parabolic troughs with concentrations in the range of $10\times$, but a second generation of nonconcentrating collectors soon began to be tested, including thin films, flat plates, and compound parabolic collectors (CPCs). Use of one-sun systems for water treatment is based on two factors: the high percentage of UV photons in the diffuse component of solar radiation and the dependence of rates on light intensity.

It was shown experimentally that above a certain flux of UV photons, the reaction rate changes from one to a half order depending on the intensity. This modification does not seem to happen at a specific radiation intensity because different researchers have obtain different results. Presumably, the experimental conditions have a significant effect. Most authors attribute the transition of $r = f(I^{1.0})$ to $r = f(I^{0.5})$ to excess photogenerated species (e^-, h^+, and $^\bullet$OH). Because the reaction rate depends on the amount of hydroxyl radicals present, and these are generated in the holes (see Eq. 4), $r \propto I^{0.5}$ when I is high. At higher radiation intensities, another transition from $r = f(I^{0.5})$ to $r = f(I^0)$ occurs. At this moment, the photocatalytic reaction ceases to depend on the received radiation and instead depends only on the mass transfer within the reaction. Thus, the rate is constant even though the radiation increases. This effect can be attributed to different causes, such as the lack of electron scavengers (i.e., O_2) and organic molecules in the proximity of the TiO_2 surface and/or excess products occupying active centers of the catalyst. These phenomena occur more frequently when working with supported catalyst and/or at a low agitation level. This implies a low catalyst surface in contact with the liquid and smaller turbulence. This does not favor the contact of reactants with the catalyst and the diffusion of products from the proximity of the catalyst to the liquid.

These effects may be appreciably attenuated if a product that reduces the importance of the electron/hole recombination is added. When the electrons are trapped, recombination of e^-/h^+ is impeded. Either way, the addition of oxidants can improve the efficiency of the process at high illumination intensities. Moreover, oxidants can increase the quantum yield even at low irradiation levels due to their strong oxidizing character. The use of inorganic peroxides has been demonstrated to enhance the rate of degradation of different organic contaminants because they trap the photogenerated electrons more efficiently than O_2. The increased photocatalytic reaction rate with these additives would decrease photoreactor dimensions proportionally and dramatically decrease overall costs. The oxidizing substance should not generate any toxic by-product, so hydrogen peroxide is the obvious candidate. The effect depends on H_2O_2 concentration, generally showing an optimum range of concentration. Peroxydisulfate can also be a beneficial oxidizing agent in detoxification because it reacts with the semiconductor photogenerated electrons (e^-_{CB}). The peroxydisulfate ion accepts an electron and dissociates (Eq. 13). The sulfate radical proceeds through the reaction shown in Eq. (14). In addition, the strongly oxidizing $SO_4^{-\bullet}$ ($E^o = 2.6$ V) can

directly participate in the degradation processes.

$$S_2O_8^{2-} + e_{CB}^- \rightarrow SO_4^{-\bullet} + SO_4^{2-}. \quad (13)$$

$$SO_4^{-\bullet} + e_{CB}^- \rightarrow SO_4^{2-}. \quad (14)$$

4. SOLAR DETOXIFICATION AND DISINFECTION TECHNOLOGY

4.1 Collectors for Solar Water Detoxification

CPCs are a good option for solar photochemical applications. CPCs are static collectors with an involute reflective surface around a cylindrical reactor tube. They have been found to provide the best optics for systems with a low solar radiation concentration and can be designed with a concentrating ratio close to 1 for capturing both direct and diffuse UV sunlight. The reflector design enables almost all the UV radiation arriving at the CPC aperture to be collected and available for processing in the reactor. The light reflected by the CPC is distributed around the back of the tubular photoreactor so that most of the reactor tube circumference is illuminated. It is easy to pipe water in and to distribute it. Most of the components of solar photochemical systems are made of standard materials with no special requirements except for the reactor, the reflective surface, and the catalyst.

4.2 Reflective Surfaces

The optical quality requirements of reflective surfaces for solar applications are usually related to the concentration required by the particular application under consideration. The higher the concentration desired, the stricter the requirements for quality of parameters. In the case of solar photochemical applications, the strictest requirements are for parabolic trough collectors (PTC); for example, UV-mirror materials must have a specular reflectance between 300 and 400 nm to achieve concentration ratios from 1 to 20. The larger the errors, particularly the reflective surface errors, the lower the effective concentration ratio. The reverse is also true: The lower the effective concentration ratio, the higher the optical errors may be, and therefore the lower the quality of reflective surface required. This is an important characteristic favoring low or nonconcentrating systems since these lower quality requirements (lower specular reflectance) directly translate into lower manufacturing cost; the reflector element can represent a considerable fraction of collector cost. With regard to the reflector/concentrator, another important factor is the reflective base material. The majority of solar photochemical processes require a highly reflective material for UV radiation. Reflectivity (reflected radiation/incident radiation) between 300 and 400 nm for traditional silver-coated mirrors is very low, so aluminum-coated mirrors are the best option. The ideal reflective surface for solar photochemical applications must have high reflectivity in the UV range, acceptable durability under outdoors conditions for extended service lifetimes, and a reasonable cost. The surfaces that best fit these requirements are electropolished-anodized aluminum (electrolytically formed aluminum oxide outer layer) and organic plastic films with an aluminum coating.

4.3 Photocatalytic Reactor

The photochemical reactor must contain the working fluid, including the catalyst or sensitizer, and must transmit solar UV light efficiently with minimal pressure drop across the system. When a large array of solar collectors are going to be used, such as in the case of treatment of water contaminants by solar photocatalysis, the reactor must be hard enough to work under usable water pressure. Tube configurations are clearly most appropriate for fluid containment and pumping when large volumes are to be processed. The choice of materials that are both transmissive to UV light and resistant to its destructive effects is limited. Photochemical reactors must be able to withstand summer temperatures of approximately 70–80°C. In addition, reactor material must be inert with regard to the chemicals that must be contained. Common materials that meet these requirements are fluoropolymers, acrylic polymers, and several types of glass. Quartz has excellent UV transmission, but the advantage in transmission in the terrestrial solar spectrum compared to other materials does not justify its high cost. Fluoropolymers are a good choice for photoreactors due to their good UV transmittance, but one of their major disadvantages is that in order to achieve a desired minimum pressure rating, the wall thickness of a fluoropolymer tube may have to be increased, which in turn will lower its UV transmittance. Glass is a natural alternative for photoreactors. Standard glass, used as a protective surface, is not satisfactory because it absorbs part of the UV radiation that reaches it due to its iron content. Borosilicate glass has good transmissive properties in the solar range with a cutoff

of approximately 285 nm. Therefore, such a low-iron content glass would seem to be the most adequate. Because both fluoropolymers and glass are valid photoreactor materials, cost is an important issue.

5. SOLAR DETOXIFICATION AND DISINFECTION APPLICATIONS

5.1 Wastewater Treatment

The hydroxyl radical ($^{\bullet}OH$) is the major intermediate reactive responsible for organic substrate oxidation. The free radical $HO_2^{\bullet}$ and its conjugate $O_2^{\bullet-}$ are also involved in degradation processes but are much less reactive than free hydroxyl radicals. These radicals react strongly with most organic substances by hydrogen abstraction or electrophilic addition to double bonds. Free radicals further react with molecular oxygen to give a peroxy radical, initiating a sequence of oxidative degradation reactions that may lead to complete mineralization of the contaminant. In addition, hydroxyl radicals may attack aromatic rings at positions occupied by a halogen, generating a phenol homologue. The electrons of the conduction band can also degrade organic compounds by reductive pathways.

In general, the types of compounds that can been degraded include alkanes, haloalkanes, aliphatic alcohols, carboxylic acids, alkenes, aromatics, haloaromatics, polymers, surfactants, pesticides, and dyes. Equation (15) generally holds true for a halogenated organic compound of the general formula $C_nH_mO_pX_q$:

$$C_nH_mO_pX_q + \left(\frac{(m-2p)}{4} + n\right)O_2 \rightarrow nCO_2 + \frac{m-q}{2}H_2O + qHX. \quad (15)$$

In photodegradation, transformation of the parent organic compound is desirable in order to eliminate its toxicity and persistence, but the principal objective is to mineralize all pollutants. Furthermore, the stoichiometry proposed for the general reaction (Eq. 15) has to be demonstrated in each case by a correct mass balance. Reactives and products might be lost, producing unreliable results. Mineralization rate is determined by monitoring inorganic compounds, such as CO_2, Cl^-, SO_4^{2-}, NO_3^-, and PO_4^{3-}. When organics decompose, a stoichiometric increase in the concentration of inorganic anions is produced in the treated water, and very often there is an increase in the concentration of hydrogen ions (decrease in pH). In order to demonstrate that there are no product losses, the molar ratio must be in accordance with the organic substrate structure. The absence of total mineralization has been observed only in the case of *s*-triazine herbicides, for which the final product obtained is essentially 1,2,5-triazine-2,4,6-trihydroxy (cyanuric acid), which is not toxic. This is due to the strong stability of the triazine nucleus, which resists most methods of oxidation. Nitrogen-containing molecules are mineralized into NH_4^+ and NO_3^-. Ammonium ions are relatively stable and the proportion depends mainly on the initial oxidation degree of nitrogen and on the irradiation time. Analyses of aliphatic fragments resulting from the degradation of the aromatic ring have mainly revealed formate and acetate ions. Other aliphatics (presumably acids, diacids, and hydroxylated compounds) have also been found. Formate and acetate ions are rather stable, which in part explains why total mineralization takes much longer than dearomatization.

Solar photocatalytic mineralization of organic pollutants in water, using the interaction of UV radiation and semiconductor catalysts, has major potential in the industrial destruction of toxic organics in water, and the number of applications and target compounds are numerous. In areas with medium to high solar radiation, solar detoxification is useful for treating water with a maximum organic concentration of several hundred milligrams per liter, nonbiodegradable contaminants, and hazardous contaminants present within complex mixtures of organics. As a result of intense research carried out by groups all over the world during the past 20 years, solar detoxification is a good solution with many interesting applications. During the past decade, there have been thousands of publications and patents on heterogeneous photocatalytic removal of toxic and hazardous compounds from water and air. Some examples are phenols and chlorophenols, chlorinated hydrocarbons, pharmaceutical compounds (antibiotics, antineoplastics, and other hazardous pharmaceutical waste), agrochemical waste (pesticides), and cyanides. They are produced in industrial quantities in processes such as the manufacture of phenol-formaldehyde resins, volatile organic compound (VOC) scrubbers, PVC production, wood-preserving waste, cleaning of marine tank terminals and pipelines, and manufacture of pesticides and rinse water from the equipment used for their application.

Here, pesticides are used as an example for demonstrating the practical application of solar photocatalysis. Although pesticides, and agrochemical compounds in general, have been detected in water since the 1950s and 1960s, in the past 15 years their use has increased dramatically throughout the

world and has nearly doubled every 5 years since 1975. The United Nations estimates that of all pesticides used in agriculture, less than 1% actually reach the crops. The remainder ends up contaminating land, air, and especially water. Therefore, pesticide pollution of water is one of today's greatest environmental problems, with widespread ecological consequences. In areas where agriculture is intensive, the following are major sources of water pollution:

Pesticide treatment as a routine agricultural practice: Approximately 0.2 tons/ha are usually applied in intensive agriculture by classical spraying or in the irrigation pipeline.

Rinse water polluted with pesticides from containers and spray equipment: Small amounts of pesticide remaining in empty containers (approximately 70 units/ha) are typically spilled out into the environment.

Wastewater from the agricultural industry: Water from cleaning or postharvest treatment of fruits and vegetables before packing produces approximately 1 m^3/day/ha of wastewater.

Plant waste contaminated with pesticides: It is estimated that 30 tons/ha/year of untreated plant waste is normally disposed of in landfills.

One of the major problems is the uncontrolled dumping of empty pesticide containers. Although the amount of pesticide remaining in each container is very small, millions of them are dumped every year, making it one of the most important sources of pollution in areas with this type of agriculture. A proposed solution is their selective collection and transportation to a recycling plant, where they are rinsed for later reuse. The resulting rinse water, contaminated by the different mixtures of pesticides, must be treated. Simple, inexpensive, and accessible technologies for the *in situ* treatment of such water are therefore necessary. Herbicides are traditionally removed from water using granular or powdered-activated carbon, nanofiltration, ozonation, etc., but these processes have inherent limitations in applicability, effectiveness, and costs. On the other hand, the problems associated with their disposal are quite well suited to solar photocatalytic treatment for the following reasons: (i) The initial pesticide concentration can be controlled as a function of the number of containers washed with the same water, so the most appropriate concentration for optimum photocatalytic efficiency can be chosen; (ii) toxicity is extreme, low volume, and in a well-defined location; (iii) such point sources of pollution may be ideally treated in small-scale treatment units; and (iv) intensive agriculture in greenhouses is usually concentrated in sunny countries.

Solar detoxification has also demonstrated efficiency in the degradation of the halogenated solvents found in much of pharmaceutical industry waste and VOC emissions. Environmental regulations and directives are pushing industrial managers to control VOC emissions. One of the VOC emission control methods is wet scrubbing, but the contaminated water from the scrubbers has to be treated. This could easily be done by solar detoxification. Another source of halogenated wastes is halocompound fabrication. The waste can be estimated as a low percentage of total production, dissolved at 100–200 mg/liter. PVC, for example, produces 2.5 m^3 of wastewater effluent contaminated with short-chain polymers or PVC monomer for each ton of PVC. A feasibility study determined that a 6-m^2 photoreactor can completely mineralize 100 mg/liter of chloroform in 250 liters of wastewater in 8 h of sunlight.

Decontamination of water effluents is one of the most successful photochemical applications of solar photons. The technology has been validated with the construction of large pilot plants. Solar photocatalysis promises to be an important breakthrough in the implementation of solar processes as a specific solar technology for water treatment becomes commercially available. Market analyses show that this environmentally friendly technology, which can destroy many of the most problematic persistent organic pollutants, has numerous potential applications. There can be no doubt that photocatalytic water detoxification is full of promise, as demonstrated by the progress made during the past 10 years.

However, an analysis of the historical evolution of solar photocatalysis clearly identifies three different stages of development. Initial efforts of solar-conscious researchers focused on transferring laboratory research to solar-engineered testing with existing technology. These first results produced excitement in the photochemical research community. Their extrapolation to practical situations assumed an ability to degrade almost any organic contaminant as well as some metals. Later, more appropriate collectors and designs were developed, but the need to know the fundamentals of certain aspects of the reaction led to an increasing number of studies on kinetics, mechanisms, performance of mixtures, and operating parameters with mixed results. It was a period of both promising and discouraging results. In the third stage, which seems to be under way, the boundary conditions of applications are being determined, and the technology is focusing on a specific initial applications, with the particularity that early development and unsolved questions

coexist with near-commercial and industrial applications of the technology. As a result, the environmental market, although very receptive to clean energy resources, is reluctant to accept such a "risk-assuming" initiative.

5.2 Water Disinfection

In many areas of developing countries, cross-contamination of the water supply with both chemical and microbiological contaminants is a major problem; water decontamination and disinfection is recognized by the World Health Organization as one of the most important barriers for human health protection. Solving the problem of trace levels of hazardous nonbiodegradable contaminants is very difficult in developing countries. However, microbial contamination of drinking water due to the lack of treatment sewage or to inadequate working conditions remains the principal water-quality problem in these countries. Access to clean water is restricted by the contamination of water resources, the lack of adequate storage facilities, and the absence of effective delivery systems. Chlorine is widely used as a main disinfectant, but reports show that overall water service coverage in rural areas is low because of many factors (technological, cost, maintenance, social, cultural, logistic, education, etc.). In addition, there is risk caused by the chlorinated by-products, such as trihalomethane and other total organohalides. Chlorine, the most common treatment, is best suited for broad municipal water supply projects for which it is relatively easy to control and monitor the dosage level of disinfectant. However, in the smaller rural and marginalized communities, chlorine use can be impractical or even dangerous. Without proper dosage regulation, chlorine overdoses can cause carcinogenic by-products and strong odors and tastes. Other existing technologies, such as ozone and UV disinfection, are clearly very difficult to implement in rural areas. UV light and ozonation are particularly unsuitable for isolated areas because of the amount of complicated equipment necessary and, in the case of ozone, the danger of toxic waste leakage.

The use of solar energy as a primary alternative to chlorination could prove to be economically viable in countries with a high degree of sunlight. Solar treatment could also be an efficient way to degrade organic matter, making chlorinated drinking water harmless for human consumption. Photocatalytic oxidation is a promising technology for the disinfection of water and wastewater because the $^{\bullet}OH$ radical is highly toxic toward microorganisms.

6. CONCLUSION

Although its development presents difficulties, treatment of industrial wastewater seems to be one of the most promising fields of application of solar detoxification. The only general rule is that there is no general rule; each case is completely different from any other. Consequently, preliminary research is almost always required on a case-by-case basis to assess and optimize potential pollutant treatment for any problem.

SEE ALSO THE FOLLOWING ARTICLES

Solar Distillation and Drying • Solar Fuels and Materials • Solar Heat Pumps • Solar Thermal Energy, Industrial Heat Applications • Solar Thermal Power Generation • Solar Water Desalination • Sun, Energy from

Further Reading

Ajona, J. A., and Vidal, A. (2000). The use of CPC collectors for detoxification of contaminated water: Design, construction and preliminary results. *Solar Energy* **68**(1), 109–120.

Blake, D. M. (1999). Application of the photocatalytic chemistry of titanium dioxide to disinfection and the killing of cancer cells. *Sep. Purf. Methods* **28**(1), 1–50.

Cassano, A. E. (2000). Reaction engineering of suspended solids heterogeneous photocatalytic reactors. *Catalysis Today* **58**, 167–197.

Chiron, S., Fernández-Alba, A., Rodríguez, A., and García-Calvo, E. (2000). Pesticide chemical oxidation: State of the Art. *Water Research* **34**(2), 366–377.

Helz, G. R., Zepp, R. G., and Crosby, D. G. (eds.). (1994). "Aquatic and Surface Photochemistry." Lewis, Boca Raton, FL.

Hoffmann, M. R., Martin, S. T., Choi, W., and Bahnemann, D. (1995). Environmental applications of semiconductor photocatalysis. *Chem. Rev.* **95**, 69–96.

Legrini, O., Oliveros, E., and Braun, A. M. (1993). Photochemical processes for water treatment. *Chem. Rev.* **93**, 671–698.

Malato, S., Blanco, J., and Herrmann, J. M. (eds.) (1999). Solar catalysis for water decontamination. *Catalysis Today* **54**(2–3).

Malato, S., Blanco, J., Vidal, A., and Richter, C. (2002). Photocatalysis with solar energy at a pilot-plant scale: An Overview. *Appl. Catal. B Environ.* **37**, 1–15.

Safarzadeh-Amiri, A., Bolton, J. R., and Cater, S. R. (1996). The use of iron in advanced oxidation processes. *J. Adv. Oxid. Technol.* **1**(1), 18–26.

Serpone, N., and Emeline, A. V. (2002). Suggested terms and definitions in photocatalysis and radiocatalysis. *Int. J. Photoenergy* **4**, 91–131.

Serpone, N., and Pelizzetti, E. (eds.). (1989). "Photocatalysis: Fundamental and Applications." Wiley, New York.

Solar Distillation and Drying

SALEH AL-KHARABSHEH
Hashemite University
Zarqa, Jordan

D. YOGI GOSWAMI
University of Florida
Gainesville, Florida, United States

1. **Solar Distillation Background**
2. **Solar Distillation Processes**
3. **Solar Still**
4. **Still Performance**
5. **Operation and Maintenance**
6. **Solar Drying: Definition and Background**
7. **Drying Processes**
8. **Solar Drying Processes**
9. **Solar Dryers**
10. **Dryer Performance**

Glossary

absorptivity A property of a surface, defined as the fraction of incident energy absorbed on the surface.
convection A mode of heat transfer in which heat is transferred due to the motion of the fluid.
dessicant A material that absorbs moisture from air.
dew point The temperature at which water vapor in air starts to condense when cooled.
drying capacity The quantity of products that can be dried in a single batch of loading.
drying potential The ability of air to evaporate water from the material to be dried, which depends on the temperature and relative humidity of the air.
drying ratio The ratio of the weight of wet products entering the dryer to the weight of the products leaving the dryer.
humidity ratio The ratio of the mass of water vapor in the air to the mass of the dry air.
liner Any natural black or painted material that can withstand the operating temperature and saline water.
natural convection The motion of the fluid due to buoyancy forces, which arise from density differences, caused by temperature variations in the fluid.
regeneration The process of removal of water from the desiccant.
relative humidity The ratio of the mass of water vapor in a given volume of air to the maximum amount of vapor that the same volume of air can hold at the same temperature and pressure.
saturation The condition at which a fluid holding another substance cannot hold any additional amount at the given conditions.
scale The deposition of minerals on any solid surface.
solubility limit The maximum amount of solid that can be dissolved in a liquid at a specified temperature.
sublimation The phase change of a material directly from solid to gas without passing through a liquid phase.
vapor pressure The actual pressure exerted by vapor.

Solar distillation is the process in which solar heat is used to purify water from an impure water source by evaporation and condensation. When solar distillation is used to purify water from saline water, the process is also called solar desalination. Desalination converts saline water with high salt content, about 3.5% by weight in seawater and about 0.6% in brackish water, into fresh water suitable for drinking and other purposes. Solar energy can be used for desalination, either as thermal energy through the use of solar thermal collectors or solar ponds, or as electricity. mainly through the use of photovoltaic cells.

1. SOLAR DISTILLATION BACKGROUND

Solar distillation has been practiced for a long time. The first large-scale solar distillation plant was constructed in 1872 at Las Salinas, Chile, and

 © 2004 Elsevier Inc. All rights reserved.

consisted of solar stills of about 4700 m^2 total area, producing about 23,000 l/day. Solar distillation gained more attention after World War I, when many devices were developed. Solar distillation is suitable for remote, arid, and semiarid areas, where drinking water shortage is a major problem and solar radiation is high. The drawbacks to using solar energy for distillation are the high initial cost and the intermittent nature of the sun. Due to these limitations, the present capacity of solar desalination systems worldwide is about 0.01% of the existing large-scale conventional distillation plants. Those plants are small- to medium-capacity installations serving a small number of people and poor communities.

2. SOLAR DISTILLATION PROCESSES

Solar energy can be used to supply the required energy for a desalination process either in the form of thermal energy needed for distillation or as electricity that can be used to drive processes requiring electricity. Solar desalination processes can be classified into two categories:

1. *Indirect systems*. These are usually of small or medium capacity and involve two separate subsystems, a solar energy system and a conventional plant for using the collected solar energy to produce fresh water.
2. *Direct systems*. Here the heat collection and distillation processes take place in the same equipment. In such systems, solar energy is used to produce the distillate directly in the solar still. The still acts as a trap for solar radiation (the greenhouse effect). As solar radiation passes through the transparent cover, it gets absorbed by the water and the absorber surface. As a result the water temperature, thus the vapor pressure, increases and becomes greater than that of the cover and the water evaporates. The vapor rises to the cover by natural convection where it is condensed on the inner side of the cover. A gravity feed funnels the condensate into the collection tank.

3. SOLAR STILL

The most common solar still is a greenhouse type. Greenhouse solar stills are simple to construct and operate with a transparent cover, which may be symmetrical (Fig. 1) or asymmetrical. The main parts of a greenhouse solar still are described next:

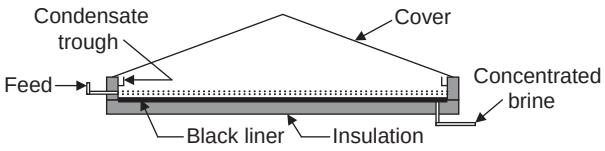


FIGURE 1 Schematic of a symmetrical greenhouse solar still.

Basin. Saline water is kept for distillation in the basin. The basin may be insulated from the bottom to minimize heat loss to the ambient. It is painted black or lined with a black sheet to maximize its absorptivity. The basin may range from 10 to 20 mm deep (shallow basin) to 100 mm or more (deep basin). The basin floor must be inclined slightly, around 1°, to allow for easy drainage of the concentrated brine. The materials needed to construct a still must be resistant to saline water corrosion, and preferably be inexpensive, durable, and available locally. Aluminum, concrete, wood impregnated with a resin, iron painted to resist corrosion, or plastic may be used.

Transparent cover. The cover acts as the condenser, so that the vapor condenses on its inside surface and collects at the bottom. It also reduces heat losses from the hot water by keeping the wind away from it. The slope of the cover must be steep enough to allow the water to flow all the way down to the collection troughs, without falling back into the basin. The cover is usually made of glass or plastic, though glass is preferred because of its high transparency, rigidity, and comparatively low cost. One disadvantage of glass is that it breaks on impact. Plastics, on the other hand, are lightweight, inexpensive, less likely to break, and easy to handle. Their main disadvantages are discoloration and deterioration due to ultraviolet (UV) radiation in the sunlight.

Collection trough. The collection troughs are placed at the lower edges of the cover to collect the distillate. Troughs are pitched along its length so that the distillate flows to the lower end of the still, from where it is collected.

The design of a solar still requires optimization of many factors: brine depth, thermal insulation, cover slope, shape, and material for the still. A well-designed still must also have adequate provision to collect the rainwater that falls on it, so that the surrounding ground is not eroded or flooded.

A variety of basin type solar still designs are used in practice. They differ from each other in the type of

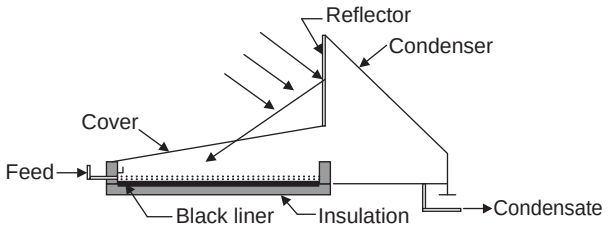


FIGURE 2 Schematic of a greenhouse solar still with a reflector.

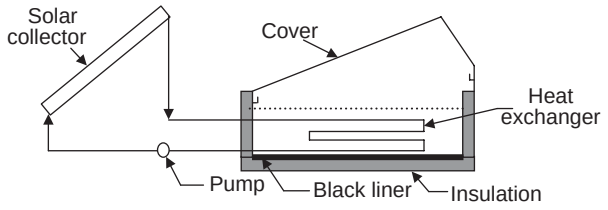


FIGURE 3 Schematic of a greenhouse solar still with a solar collector.

materials used, geometry, the method of supporting the transparent cover, and the arrangements for supply and discharge water. The incident solar radiation on the still can be increased by using a reflector (Fig. 2), which focuses it, so that the still receives both the reflected as well as incident radiation. The temperature of the saline water inside the still may be increased by connecting the still to a solar collector (Fig. 3), or the saline water may be preheated in the collector before entering the still. The latent heat of condensation may be partially utilized by means of a multieffect solar still (Fig. 4). As the vapor from one stage condenses at the lower surface of the next stage, it gives up its latent heat at that stage thus evaporating a part of the water there. The condensate flows toward the trough and is collected as the output. Thin film evaporation creates a higher output than pool evaporation. This idea may be used in a tilted solar still, as shown in Fig. 5. The still is tilted to face the sun and the feed water flows from the top inlet toward the bottom. The flow rate needs to be properly controlled to generate the optimum output. Other designs include a tilted wick, which looks like a tilted solar still, an inflated solar still where a plastic sheet is used as the transparent cover, and a still with an internal condenser to enhance the condensation process and to preheat the feed water. Stills based on the humidification-dehumidification principle have also been developed. In a humidification-dehumidification still, saline water is heated in a collection system and pumped to the top of a humidification column. Air is

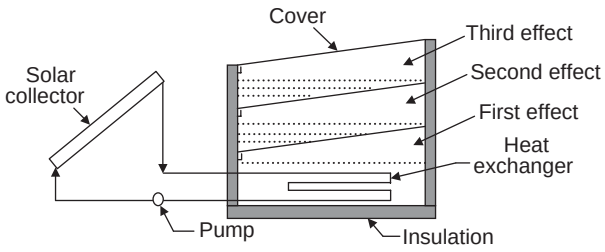


FIGURE 4 Schematic of a multieffect solar still.

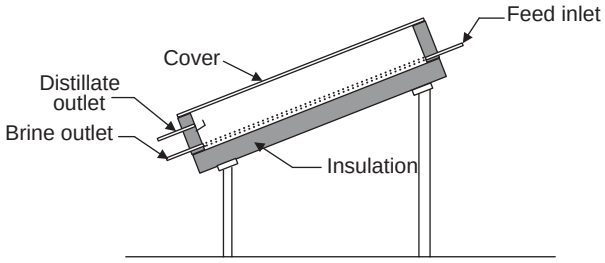


FIGURE 5 Schematic of a tilted solar still.

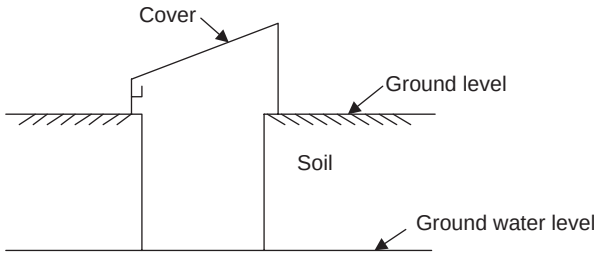


FIGURE 6 Humidity solar still.

circulated in the opposite direction and becomes humid. The humid air goes to the condenser, where the condensing vapor creates fresh water. In humid areas where there is no fresh or saline water available, fresh water can be obtained from the atmosphere. This can be done by using desiccant materials, which absorb water vapor from the atmosphere. The cooling coil of an air conditioner may be used to produce fresh water, but its temperature must be lower than the dew point of the humid air. As the humid air comes in contact with the coil, the water vapor temperature reduces below the dew point, which condenses as fresh water. For locations with very low humidity and no water resources, fresh water may be obtained from the soil moisture. At a certain depth below the earth surface, the soil remains moist. If that layer of moist soil is heated, vapor will be formed that can be condensed and collected as fresh water. Figure 6 shows a schematic of a possible arrangement.

4. STILL PERFORMANCE

Although the construction and operation of a solar still is simple, the theoretical analysis is complicated and mostly depends on experimental observations. Basic concepts of a solar still were set forth by R. V. Dunkle in 1961.

The mass transfer rate $\dot{m}$, in ($Kg/m^2.s$), can be written as

$$\dot{m} = 9.15 * 10^{-7} h_c (p_{wb} - p_{wg}).$$

The heat transfer by evaporation-condensation is

$$q_e = 9.15 * 10^{-7} h_c (p_{wb} - p_{wg}) * h_{fg},$$

where, h_c is the convection coefficient in the still, given by

$$h_c = 0.884 \left[T_b - T_g + \left(\frac{p_{wb} - p_{wg}}{2016 - p_{wb}} \right) T_b \right]^{1/3},$$

where, p_{wb} and p_{wg} are the vapor pressures, in mm Hg, of the solution in the basin at the temperature of the basin (T_b) and of water at the cover temperature (T_g). Temperatures are in degrees Kelvin and h_{fg} is the latent heat of water (J/kg).

Still efficiency, η, is defined as the ratio of the heat transfer, q_e, in the still by evaporation-condensation to the incident solar radiation, I, on the still, as given by

$$\eta = \frac{q_e}{I}.$$

Still efficiency is usually low and rarely exceeds 50%, with an average of 30 to 40%. The daily production from a solar still is about 3 to 4 l/m^2.

5. OPERATION AND MAINTENANCE

As the water inside the still is almost stagnant, suspended solids in the feed water may settle inside the still. To prevent this, suspended material should be removed from the feed before it enters the still. Growth of algae and other organisms may occur in the still. This can be minimized through the addition of algicides such as copper sulfate to the supply. The presence of scale leads to operating difficulties or loss of efficiency by reducing the heat transfer through the affected surfaces. Therefore, scale prevention is an integral part of the design, operation, and cost of the process. The scale is formed by the dissolved solids in raw water and from corrosion products. Scale usually contains chemical compounds of magnesium oxide, calcium carbonate, and calcium sulfate. The formation of scale can occur where the solubility limits of these compounds are exceeded. When pure water is separated from saline water, the concentration of dissolved solids increases and eventually reaches the saturation limit of one or more salts. Although supersaturated solutions are unstable, solid salts do not necessarily deposit immediately from these solutions. When these solutions come into contact with a solid crystal, the excess dissolved solid in the supersaturated solution precipitates. Crystals of the salt in supersaturated solution and other solids particles can act as centers of crystallization. Therefore, the best way to eliminate the scale formation problem is to flush the still before it reaches saturation.

Stills are designed to operate with water in them, if they are left dry for a period of time they may suffer serious damage. It is also necessary to clean the transparent cover periodically to maintain high transmittance of solar radiation.

6. SOLAR DRYING: DEFINITION AND BACKGROUND

Drying is one of the most important post harvest steps. It enhances the storage life of the crop products, minimizes losses during storage, and saves shipping costs. The drying process is the removal of water from the wet surface of the food. In this process, heat is transferred by convection and radiation to the surface of the produce. This heat raises the temperature and evaporates the moisture from the exterior of the agricultural products, diffusing the interior moisture to the surface and replenishing the evaporated surface moisture. Vapor in the surrounding air is removed by diffusion and movement of air. This process continues until the required drying ratio is reached. So the drying process requires a mechanism for heating the ambient air, a drying unit where moisture removal takes place, and an air handling unit (in some dryers).

The ability of the air to evaporate water from the produce is known as the drying potential. The drying potential depends on both air temperature and relative humidity. Besides the air drying potential, the airflow rate is another major factor determining the drying time. To reduce the drying time, the airflow rate and temperature need to be increased while and air's relative humidity needs to be decreased.

The food type of product to be dried determines the rate of moisture transfer from the interior to the

surface. Nonhygroscopic materials can be dried to zero moisture content; however, hygroscopic materials will always have residually bound moisture. The moisture content present in the agricultural product at equilibrium conditions (when it loses moisture to the environment at the same rate at which it absorbs moisture from the ambient) is known as the equilibrium moisture content or hygroscopic equilibrium. In such a case, increasing the temperature of the drying air will be more effective than increasing the flow rate. It is clear that the process of drying is controlled by the air properties, mainly temperature and relative humidity, and by the agricultural product properties, mainly moisture content.

The main objective of drying is to supply the required heat in the most effective way to yield the best result with a minimum overall expenditure of energy. For industrial drying hot combustion gases may be allowed to pass directly through the agricultural products, but this method is not preferred because of possible contamination by the unburned fuel. Electrical air heating is preferred, but expensive. In many cases, solar energy may be the most promising source for drying, as some agricultural products, like rice, may be better suited to drying by solar than by fossil fuel systems because case hardening and other damage are less likely at low temperatures.

7. DRYING PROCESSES

Drying processes range from simple natural sun drying to large-scale industrial drying. Based on the temperature required for drying, the processes may be classified as follows:

1. *Low-temperature drying*. In this slow-drying process (usually performed at temperatures between 15 to 50°C), the air is heated to about 1 to 6°C above the ambient. The main advantages of low-temperature drying include decreased chances of overheating and low capital and running costs, but the process is slow and does not work well for some climates. For low-temperature drying, low-cost solar collectors may be used to heat the air before it enters the dryer.
2. *High-temperature drying*. In this method, air at temperatures above 50°C is used to dry the agricultural products at a fast rate. In such systems, exhaust air from the dryer may be too hot and unsaturated to be wasted, so circulation becomes necessary to improve effectiveness and reduce costs. For this kind of drying, simple solar collectors may not be able to provide the required temperature. Concentrating collectors could be used to achieve the required temperature; however, it may be more economical to use simple solar collectors to preheat the air, and then increase the temperature of the air further using conventional fuel.
3. *Freeze drying*. In such a process, the agricultural product to be dried is first frozen, then placed inside a vacuum chamber connected to a condenser at a low temperature or a desiccant. Heat is applied to the products allowing the water to sublimate, which condenses in the condenser or is absorbed by the desiccant.
4. *Osmovac drying*. This process is normally used for drying fruit juices to make a concentrated powder. It involves two processes, an osmotic process followed by vaporization under vacuum.
5. *Desiccant drying*. In this process, a desiccant is used to remove moisture from the air to increase its drying potential. The desiccant is then regenerated by solar energy. Sometimes a portion of the product itself (especially grain) might be used as the desiccant. If solar energy is used for drying, grain could be overdried during the summer in a desiccant system. At the time of harvest, a part of the overdried grain is removed from the desiccant bin and blended with wet grain to produce a mixture with the average moisture content around the required value. Low-temperature drying may be used to achieve a uniform value for this mixture (i.e., moving moisture from wet to dry agricultural products to reach the required value).

8. SOLAR DRYING PROCESSES

Using the sun's energy to dry agricultural products (known as sun drying) is the most widely used form of drying because it is cheap, easy, and convenient. Not long ago, sun drying was the only form of agricultural drying, and it is still the most common agricultural process industry. The easiest and simplest way of sun drying is to spread the products such as fruits, vegetables, cereals, tobacco, and so forth on a suitable surface under the sun and turn them regularly until sufficiently dried so that they can be stored safely. Another important application of sun drying is the evaporation of seawater to produce salt. It is obvious that sun drying would work best in a hot and dry climate with gentle breezes but would be difficult in humid areas. Although sun drying requires little capital or expertise, it has many limitations: irregular and slow drying rate, high final

moisture content especially in humid areas, and variable product quality where some of the products may be overdried and others underdried since there is no control on the drying process. Other problems include contamination with dust and insect infestation, and sudden rain storms, which can soak the products. Also, large areas are required because sunlight is not concentrated. Problems associated with sun drying can be reduced by using mechanical dryers and conventional energy sources like electricity, coal, and fossil fuels. Due to the high cost of conventional energy sources and the associated environmental problems, solar agricultural dryers have gained popularity, particularly in developing countries. Many of these countries are located in areas that receive a good amount of solar radiation.

9. SOLAR DRYERS

Solar dryers are special structures that enhance the drying power of the sun and protect the agricultural product from dust, dirt, and insects. Figure 7 shows a block diagram of an active solar drying system, consisting of solar air collectors, a drying unit or chamber, and an air handling unit. An active solar drying system can accommodate the use of a backup heat source when there in not enough solar heat available.

Solar dryers usually employ high airflow rates at low temperatures over long periods of time, whereas fossil fuel dryers usually employ lower airflow rates and higher temperatures for fast drying. Slow drying of fruits and vegetables preserves the flavor and quality of the products much better, which makes solar drying more attractive. Therefore, many commercial solar dryers have already been marketed.

Drying capacity depends on the size of the dryer and varies with the type of the produce and the amount of moisture to be removed. Placing agricultural products on racks above each other limits the area of exposure of the food surfaces to the drying air, which might result in poor drying. Underloading will result in low efficiency because of not fully using the available capacity. The loading density can be assessed based on the following rules of thumb, which were suggested by M. A. Leon, S. Kumar, and S. C. Bhattacharya in 2002.

1. Average dryer load is about 4 kg of fresh produce per square meter of tray area.
2. Solar collector area is about 0.75 times total tray area.
3. Airflow rate is about 0.75 cubic meter per minute per meter square tray area.

There are different types of solar dryers; most of them were developed for specific products or class of products. Selection of a solar dryer represents a compromise between dryer cost, product quality, safety consideration, and installation convenience. Solar dryers may be classified into many categories depending on the mode of heating or operation. According to their operating modes, these can be classified into the following categories:

1. Passive systems: Passive systems use natural forces such as buoyancy and breeze for air flow, therefore, these systems have no external energy input except solar radiation.

a. Direct solar dryers: In such dryers, the agricultural products receive the heat directly from the sun. There are many designs for direct solar dryers, the most important of which are the following:

i. Box solar dryers: These are suitable for arid and semi-arid areas, where the relative humidity is low and the amount of incident solar radiation is high. The low initial cost and easy maintenance are the main advantages of the box dryer. Although box dryers cut the drying period by half in comparison to open sun drying, this period is still very long. Box dryers are usually rectangular, painted black on the inside, and have doors to load and unload the produce. The bottom and sides of the box may be insulated. It has transparent cover on its top. Holes are made at the base and the top of the rear vertical

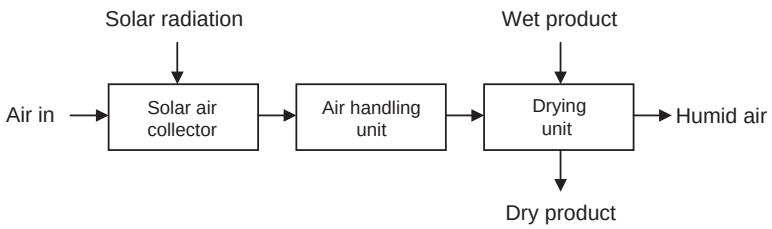


FIGURE 7 Block diagram of a solar dryer.

wall (a few centimeters below the cover) to permit movement of air by natural convection. Ambient air enters through the base holes and leaves through the top holes carrying moisture from the products. The agricultural products to be dried are kept in trays, with perforations allowing air flow through them. The trays are placed so that the product receives solar radiation directly. Heat is absorbed directly by the products themselves as well as by the internal surfaces of the drying chamber. The absorbed heat evaporates moisture from the food and induces natural convection, assisting the drying by air circulation. Figure 8 shows a schematic of a possible arrangement known as the solar box dryer. The box's overall dimensions recommend that the length is at least three times the width to minimize the shading effect. Box solar dryers can be made portable or stationary.

ii. Greenhouse solar system: A greenhouse dryer is usually made from a metallic frame covered by transparent plastic or glass with removable inner chambers, which acts as the absorber. Insulating panels may be used over the glazing at night to reduce heat loss, and a means of storing excess daytime heat is often provided. The system as a whole represents a combination of a solar collector and an integrated energy storage device. Figure 9 shows a greenhouse solar system with tobacco curing and grain drying modes (at the left and right, respectively), with an illustration of typical air flow patterns for solar drying and furnace drying. The incident solar radiation gets absorbed on the moveable black absorber. The useful heat is continuously removed by the slow rotation of the absorber and transferred to the agricultural products. Under such conditions, the dryer will most likely be an active one. For nongrain products, the absorber remains stationary. Such systems are usually of medium capacity.

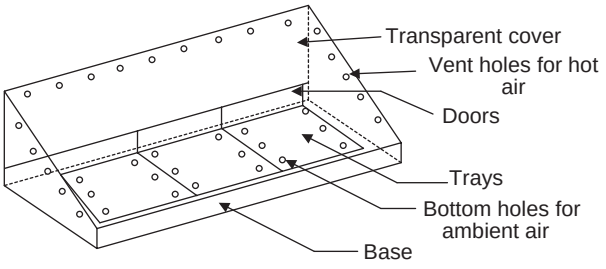


FIGURE 8 Schematic of a box type solar dryer.

iii. Rack type solar dryers: These dryers are mostly used for drying grapes. The dryer usually consists of 6 to 10 racks, each about 50 to 100 m long, 2.5 m high, and 1.5 m wide, with a north-south orientation lengthwise. Each rack is fitted with galvanized wire netting, spaced at about 25 cm from each other. The outside frame can be made of steel or wood. Products are placed or hung from the wire mesh. The whole structure is normally covered with iron sheets, and side curtains are sometimes used for protection from dust, rain, and excessive sunshine. Racks are usually constructed 9 to 12 m apart to ensure that the agricultural products are exposed to sunlight during the morning and afternoon hours. The racks are preferably placed in an exposed area, clear of any obstruction to permit free air flow.

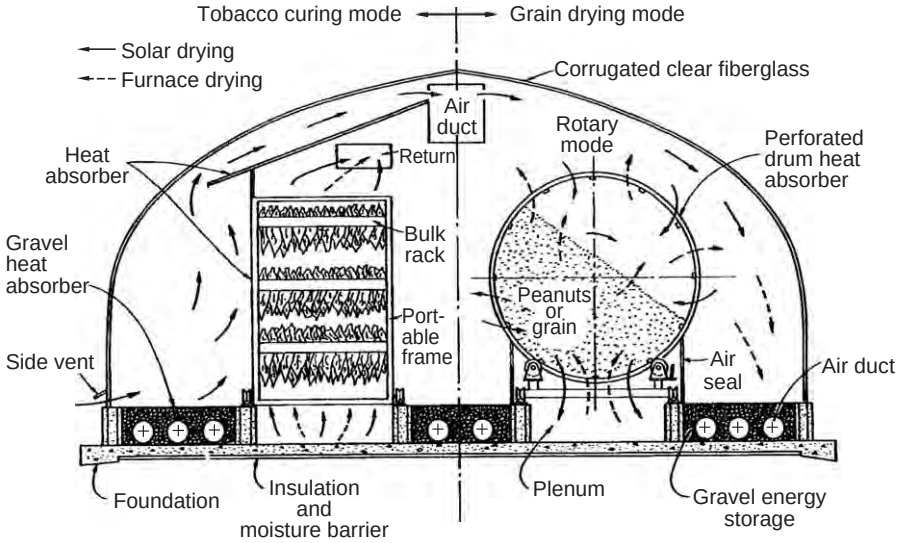


FIGURE 9 Greenhouse solar system.

b. Indirect solar dryers: This mode of drying is important especially when the products must be dried under certain specifications of temperature, humidity, and drying rate. In such dryers, the products are placed inside an opaque drying chamber, and solar collectors, usually air type, provide the hot air. Dryers of this type consist of solar air collectors and a drying chamber. The design of such dryers is the same as the conventional fuel type—the only difference is that the heat source is completely or partially solar energy. Some indirect solar dryers are described next:

i. Solar cabinet dryers: The air is heated in the collector and moves upward toward the bed due to buoyancy forces. The products to be dried are laid on the bed, which usually consists of many trays, placed close to the exit section of the collector. As the air moves across the bed, it picks up moisture from the products. Figure 10 shows one possible design of such systems.

ii. Geodesic dome solar fruit dryers: The dome shape (Fig. 11) of these greenhouse dryers makes them perform better than the conventional gable and semicylindrical shapes because they are able to collect more energy from the sun. The dryers are formed from a dome structure covered from the outside by a transparent cover and from the inside by a black plastic absorber. The transparent cover and the black absorber are separated by a suitable distance, about 1 foot. The drying trays are placed inside the inner dome. Air between the transparent cover and the black absorber is heated up by solar radiation and passed through the drying trays before it leaves the dome. The air flow can be by forced convection (Fig. 12A) or by natural convection (Fig. 12B).

c. Mixed type dryers: In mixed type of dryers, the products are heated by the hot air from the solar collectors and by direct solar radiation absorbed by them. The sides and the top of the drying chamber can be made of a transparent material, where the products absorb the solar radiation directly. Figure 10 is one possible arrangement showing the front and side walls of drying chamber as transparent.

2. Active systems: An electricity-driven blower or fan is used to circulate the air in active systems. These are further classified as follows:

a. Direct solar dryers: These consist of a drying chamber with a transparent cover and a blower. The solar radiation passing through the transparent cover heats the air. The blower forces the air to pass through the products where it picks up the moisture. Figure 8 shows one possible arrangement with a blower or a fan fixed at one side to circulate the air.

b. Indirect solar dryers: In such dryers, the air heated in the solar collectors is ducted to the drying chamber, where it evaporates and removes the moisture. Some of the most important designs of

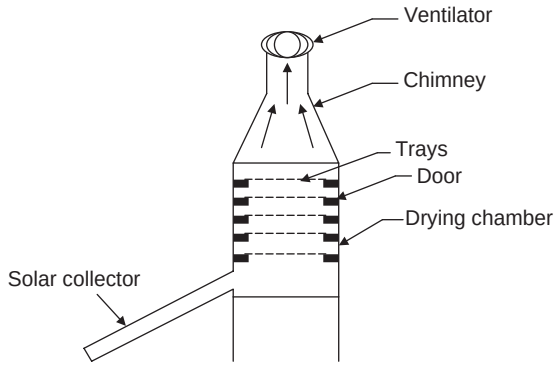


FIGURE 10 Schematic of a cabinet type solar dryer.

FIGURE 11 Structure of a geodesic dome dryer.

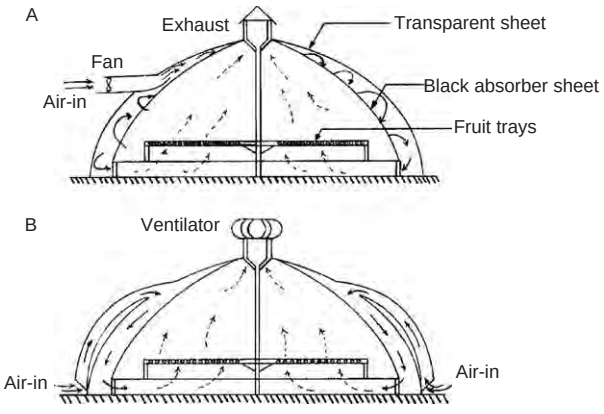


FIGURE 12 Modes of airflow in a geodesic dome dryer. (A) Forced air configuration. (B) Natural air configuration.

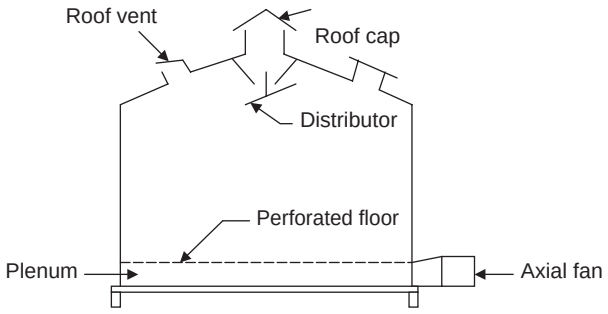


FIGURE 13 Schematic of a bin dryer.

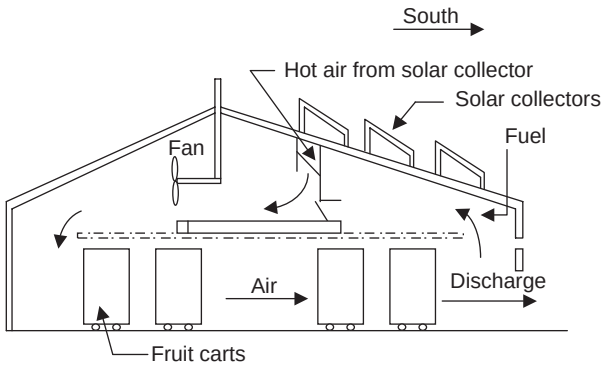


FIGURE 14 Schematic of an industrial dryer.

this type are the following:

i. Bin-type grain dryers: In these dryers, the bin is filled to the top and the hot air from the solar collector is blown from the bottom. Figure 13 shows a schematic diagram of this type.

ii. Shelf-type dryers: The drying chamber consists of many holding trays to increase the capacity of the dryer. The dryer design is similar to that in Fig. 10, but the air is circulated with the help of a blower.

iii. Hybrid or industrial dryers: Such systems use a backup source of conventional energy when solar energy is not sufficient or during off-sun hours. These are usually of a continuous process type and have large capacities. Figure 14 shows a schematic of such dryers.

c. Mixed-type dryers: These are similar to passive dryers, but air is circulated with the help of blowers or fans.

10. DRYER PERFORMANCE

The amount of heat utilized in evaporating the water, q_e, can be calculated as

$$q_e = \dot{m}_e h_{fg},$$

where h_{fg} is the latent heat of evaporation, and $\dot{m}_e$ is the amount of water evaporated in the dryer per unit area perpendicular to the direction of the flow:

$$\begin{aligned}\text{Dryer overall thermal efficiency, } \eta_d(\%) &= \frac{\text{output}}{\text{net input}} * 100 \\ &= \frac{q_e}{q_{input}}.\end{aligned}$$

If a solar collector of efficiency η_c is used to supply the hot air, the overall efficiency of the system will be

$$\eta_{ovetall} = \eta_d \eta_c.$$

In the case of active solar dryers, the energy consumed by the blower or fan is accounted for in the input. Also in the case of hybrid dryers, the amount of conventional energy consumed is added in the solar input.

SEE ALSO THE FOLLOWING ARTICLES

Solar Cookers • Solar Cooling, Dehumidification, and Air-Conditioning • Solar Detoxification and Disinfection • Solar Fuels and Materials • Solar Heat Pumps • Solar Thermal Energy, Industrial Heat Applications • Solar Thermal Power Generation • Solar Water Desalination

Further Reading

Brace Research Institute (1975). A survey of agricultural dryers, Technical report T99. Faculty of Engineering, McGill University, Quebec, Canada.

Delyannis, E., and Belessiotis, V. (2001). Solar energy and desalination. *In* "Advances in Solar Energy: An Annual Review of Research and Development." (D. Y. Goswami, Ed.), Vol. 14, pp. 287–330. American Solar Energy Society, Boulder, CO.

Dunkle, R. V. (1961). Solar water distillation: The roof type still and a multiple effect diffusion still. International Conference of Heat Transfer, University of Colorado, Publication No. 108, pp. 895–902.

Ekechukwu, O. V., and Norton, B. (1999). Review of solar-energy drying systems III: Low temperature air-heating solar collectors for crop drying applications. *Energy Conversion Management* **40,** 657–667.

Goswami, D. Y., Kreith, F., and Kreider, J. F. (2000). "Principles of Solar Engineering." 2nd ed. Taylor & Francis, Philadelphia.

Kalogirou, S. (1997). Survey of solar desalination systems and system selection. *Energy* **22**(1), 69–81.

Leon, M. A., Kumar, S., and Bhattacharya, S. C. (2002). A comprehensive procedure for performance evaluation of solar food dryers. *Renewable Sustainable Energy Rev.* **6,** 367–393.

Lior, N. (1986). "Desalination and Water Purification, 2. Measurements and Control in Water Desalination." Elsevier, Amsterdam.
Pangavhane, D. R., and Sawhney, R. L. (2002). Review of research and development work on solar dryers for grape drying. *Energy Conversion and Management* **43,** 45–61.
Sodha, M. S., Bansal, N. K., Kumar, A., Bansal, P. K., and Malik, M. A. S. (1987). "Solar Crop Drying." CRC Press, Boca Raton, FL.
Sodha, M. S., and Chandra, R. (1994). Solar drying systems and their testing procedures: A Review. *Energy Conversion and Management* **35**(3), 219–267.
Spiegler, K. S. (1977). "Salt Water Purification." 2nd ed. Plenum Press, New York.
United Nations (1970). "Solar Distillation as a Means of Meeting Small-Scale Water Demands." New York.

Solar Energy, History of

JOHN PERLIN
Rahus Institute
Martinez, California, United States

1. **Solar Architecture and Greenhouses**
2. **Solar Thermal Power**
3. **Solar Hot Water Heating**
4. **Photovoltaics**

Glossary

solar architecture The conception, design, and construction of buildings and communities so as to utilize incoming solar radiation.

solar cell A semiconductor material often made of specially treated silicon that converts sunlight directly into electricity.

solar greenhouses Structures used to cultivate plants that are designed and built to utilize solar energy for both heating and lighting.

solar thermal power Systems that capture energy from solar radiation, transform it into heat, and then use it for a variety of applications.

The ancient Greeks began using solar energy 2500 years ago. Since then, people have used the Sun for their benefit.

1. SOLAR ARCHITECTURE AND GREENHOUSES

1.1 Ancient Greece

During the fifth-century BC, the Greeks faced severe fuel shortages. Fortunately, an alternative source of energy was available—the sun (Fig. 1). Archaeological evidence shows that a standard house plan evolved during the fifth century so that every house, whether rural or urban, could make maximum use of the sun's warm rays during winter (Fig. 2). Those living in ancient Greece confirm what archaeologists have found. Aristotle noted that builders made sure to shelter the north side of the house to keep out the cold winter winds. Socrates, who lived in a solar-heated house, observed, "In houses that look toward the south, the sun penetrates the portico in winter," which keeps the house heated. The great playwright Aeschylus went so far as to assert that only primitives and barbarians "lacked knowledge of houses turned to face the winter sun, dwelling beneath the ground like swarming ants in sunless caves." The ancient Greeks planned whole cities in Greece and Asia Minor, such as Colophon, Olynthus, and Priene, to allow every homeowner access to sunlight during winter to warm their homes. By designing the streets in a checkerboard pattern running east–west and north–south, every home faced south, permitting the winter sun to flow into all houses.

1.2 Ancient Rome

Fuel consumption in ancient Rome was even more profligate than in classical Greece. In architecture, the Romans remedied the problem in the same fashion as did the Greeks. Vitruvius, the preeminent Roman architectural writer of the first century BC, advised builders in the Italian peninsula, "Buildings should be thoroughly shut in rather than exposed toward the north, and the main portion should face the warmer south side." Varro, a contemporary of Vitruvius, verified that most houses of at least the Roman upper class followed Vitruvius's advice, stating, "What men of our day aim at is to have their winter rooms face the falling sun [southwest]." The Romans improved on Greek solar architecture by covering south-facing windows with clear materials such as mica or glass.

Such clear coverings let in the light. When the light hit the masonry it converted into heat. The longer heat waves could not easily escape through the glass and the area inside heated to higher temperatures than those of previously unglazed structures. The Romans called these glass-covered structures *heliocamini*,

Encyclopedia of Energy, Volume 5. Text copyright © 2004 by Elsevier Inc. All rights reserved.
Figures copyright © John Perlin.

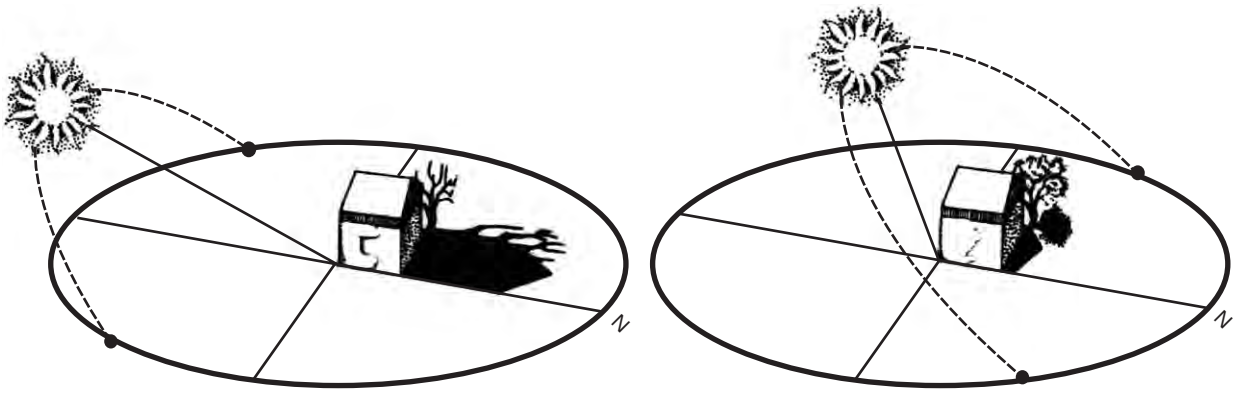


FIGURE 1 In the Northern Hemisphere, the low winter sun rises relative to Earth in the southeast, by noon it is in the south, and it sets in the southwest. During summer, the sun rises in the northeast, by noon it is directly south and very high in the sky, and it sets in the northwest. From Perlin/Butti Solar Archives.

which means "sun furnaces." The Romans used the term to describe their south-facing rooms because they became much hotter in winter than similarly oriented Greek homes, which lacked such coverings.

1.3 Ancient Roman and Renaissance Solar Greenhouses

The ancient Romans not only used window coverings to hold in solar heat for their homes but also relied on such solar heat traps for horticulture. Plants would mature quicker and produce fruits and vegetables out of season and exotic plants from hotter climates could be cultivated at home. With the fall of the Roman Empire, so too came the collapse of glass for either buildings or greenhouses. Only with the revival of trade during the 16th century was there renewed interest in growing in solar-heated greenhouses exotics brought back from the newly discovered lands of the East Indies and the Americas. Trade also created expendable incomes that allowed the freedom to take up such genteel pursuits as horticulture once more. Nineteenth-century architects such as Humphrey Repton brought the sunlit ambiance of the greenhouse into the home by attaching it onto the south side of a living room or library. On sunny winter days, the doors separating the greenhouse and the house were opened to allow the moist, sun-warmed air to freely circulate in the otherwise gloomy, chilly rooms. At night, the doors were shut to keep in as much solar heat as possible. By the late 1800s, the country gentry had become so enamored of attached greenhouses that they became an important architectural fixture of rural estates.

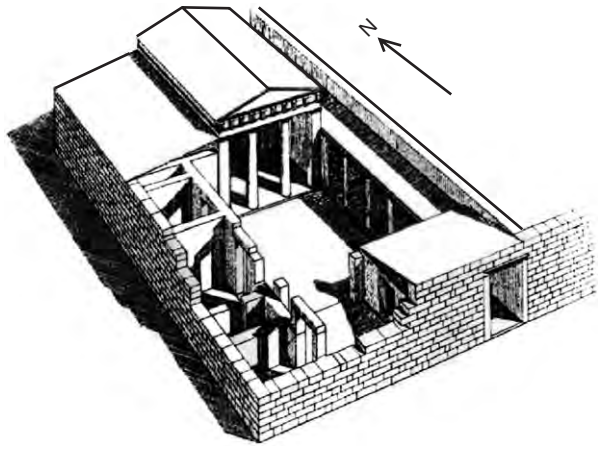


FIGURE 2 The Greeks took advantage of their knowledge of the location of the sun during the year to heat and cool their buildings. They opened their buildings to the south and protected them with a covered porch. The strategy allowed winter sunlight to pour into the house, whereas the roof shaded the house from the high summer sun. From Perlin/Butti Solar Archives.

1.4 Indigenous America

American solar architecture began with its indigenous heritage. Acoma, built by the Pueblo Indians in the 12th century and continuously inhabited since then, serves as an excellent example of their sensitivity to building with the climate in mind. The "sky city," built atop a plateau, consists of three long rows of dwelling units running east to west. Each dwelling unit has two or three tiers stacked one behind the other to allow each one full exposure to the winter sun. Most doors and windows open to the south to catch the warm solar rays of winter. The walls are built of adobe. The sun strikes these heat-absorbing south walls more directly in winter than during summer. The sun's heat travels through the adobe slow enough to reach the interior as night

falls, heating the house through the night. Insulating the ceiling with straw keeps the horizontal roof, vulnerable to the rays of the high summer sun, from allowing too much heat to enter the house.

1.5 Early Euro-Americans

Spaniards who settled in the American southwest often built to take advantage of the winter sun: They aligned their homes east to west so the main portion of the house faced south. Shutters outside the windows were closed at night to help keep the solar heat that had flowed in all day from escaping during wintertime.

Settlers in New England considered the climate when they built their homes. They often chose "saltbox" houses that faced toward the winter sun and away from the cold winds of winter. These structures had two south-facing windowed stories in front where most of the rooms were placed and only one story at the rear of the building. The long roof sloped steeply down from the high front to the lower back side, providing protection from the winter winds. Many saltbox houses had a lattice overhang protruding from the south facade above the doors and windows. Deciduous vines growing over the overhang afforded shade in summer but dropped their leaves in winter, allowing sunlight to pass through and penetrate the house.

1.6 Post-World War I Europe and America

Ameliorating the terrible slums that blighted European cities during the industrial revolution sparked a renewed interest in building with the sun. As one English city planner bent on housing reform urged in the first decade of the 20th century, "Every house should have its face turned to the sun, whence comes light, sweetness, and health." Young architects on the Continent agreed, declaring that their profession should embrace functionalism over anesthetics. To many, functionality meant designing houses that satisfied the basic needs of those living in them. In the relatively chilly European climate, this dictum meant providing dwellers with a warm haven. Therefore, Hannes Meyer, the leader of functional architecture in Germany, argued that an architect wishing to build a useful house must consider its body "to be a storage cell for the heat of the sun." Hence, "maximal use of the sun," as far as Meyer was concerned, was "the best plan for construction." Meyer, who became the director of the highly famous and influential Bauhaus architectural school, told fellow architects that to achieve optimum sunshine in buildings, they must conduct or at least be privy to scientific research into the sun's year-round movement relative to the earth. Germany led a renaissance in solar building during the 1930s that spread throughout Europe only to be stamped out by the Nazis, who condemned functional architecture as Jewish. When they came to power, a good number of German architects who designed solar buildings fled, many to the United States. George Fred Keck, a Chicago architect, befriended some of these expatriates and through their influence began designing homes in the Chicago area according solar building principles: expansive south-facing glass to trap the winter sun; long overhangs to shade the house in summer; minimal east–west exposure to prevent overheating in summer and fall; and the placement of secondary rooms, garages, and storage corridors on the north side to help insulate the living quarters from the cold north winds. Keck had a knack for publicity and called the houses he designed "solar homes." By the mid-1940s, Keck's work caught the attention of the national media. *House Beautiful*, *Reader's Digest*, and *Ladies Home Journal* featured his work. Fuel rationing during the war inclined the American public toward valuing the energy-saving features of solar homes. When war ended, the building market exploded. With the wartime-conservation ethic still imbued in most people's minds, many manufacturers in the prefabricated home industry adopted solar design features for leverage in this highly competitive market.

1.7 Post-World War II America

Studies of houses with large amounts of glass on the south side demonstrated that they experience much greater temperature swings than ordinary homes, warming quickly during the day due to accumulating the sun's heat and cooling off rapidly at night due to large heat losses through the windows. Tucson Architect Arthur Brown eliminated the unwanted fluctuations in temperature by running a thick black wall between the sun porch next to the south-facing glass and the living quarters. During sunny winter days, sunlight struck the black wall and the concrete absorbed the solar heat. Estimating that heat moves through the wall at 1 in. per hour, Brown made the wall 8 in. thick. As evening approached, the solar heat soaked up by the wall began to radiate into the rooms and continued to do so during the night. By

morning, all the solar heat in the wall had dissipated into the living areas, leaving the wall totally cooled down and ready for another heating cycle.

Brown also designed the world's first solar-heated public building in 1948. Since students attended classes between 9 AM and 3 PM, Brown did not have to worry about storing solar energy since no one needed heat at night or in the very early morning. To keep costs down, he used the roofing supports to separate the space between the ceiling and the roof into channels. In this way, the roof served as the solar collector. The school's fan, which had forced gas-heated air through the building, now pushed air through the channels. The sun's energy heated the air and a second fan distributed that heat into the classrooms when needed. The solar system provided the school with 86% of its heat. However, in 1958, when the school district decided to expand the campus, the authorities chose a gas-fired furnace due to cheap energy costs. Cheap energy running highly reliable and easy to use heating and cooling systems led to an almost universal disinterest in solar design. It was only in the 1970s, with the Arab oil embargo and skyrocketing fuel prices, that people became interested in solar architecture again.

2. SOLAR THERMAL POWER

2.1 Ancient Chinese, Greeks, and Romans

Archaeological finds and ancient literature show that solar concave reflectors and convex lenses have a long history. Convex lenses predate concave reflectors by more than 1000 years. People crafted the first convex lenses out of rock crystal. Ancient Chinese and Roman doctors relied on convex lenses to cauterize wounds. The first mention of solar concave reflectors appeared in the third-century BC writings of Euclid. The ancient Chinese and Greeks used solar concave reflectors to light fires for domestic cooking.

2.2 Medieval and Renaissance Europe

Latin translation of Arab writer Al-Hazan introduced Europeans to ancient Greek works on solar concave reflectors and convex lenses. Between the 13th and 16th centuries, many Europeans suggested burning whole armies with huge solar reflectors, but no one could build reflectors or lenses large enough for martial purposes. However, Europeans did use reflectors to ignite fires for cooking meals, to distill perfumes, and to weld metal statuary.

2.3 19th- and Early 20th-Century America, Europe, and Africa

Many in 19th-century Europe worried that the prodigious consumption of coal by industry would eventually lead to coal shortages. Among the concerned, Augustine Mouchot thought that the sun's heat could replace coal to run Europe's many steam engines. Mouchot's work with mirror technology led him to develop the world's first sun motor in the 1870s, which produced sufficient steam to drive machinery (Fig. 3). One journalist described it as a "mammoth lamp shade, with its concavity directed skyward." After 1 year of testing Mouchot's sun machine, the French government concluded,

> *In France, as well as in other temperate regions, the amount of solar radiation is too weak for us to hope to apply it for industrial purpose. Though the results obtained in our temperate and variable climate are not very encouraging, they could be much more so in dry and hot regions where the difficulty of obtaining other fuel adds to the value of solar technologies. [Hence] in certain cases, these solar apparatuses could be called upon to provide useful work.*

An American inventor, Frank Shuman, took his cue from the French report. In 1911, he chose to locate his commercial solar motor in Egypt, where land and labor were cheap, sun was plentiful, and coal was costly. To increase the amount of heat generated by the solar motor, glass-covered pipes were cradled at the focus of a low-lying trough-like reflector. A field of five rows of reflectors was laid out in the Egyptian desert. The solar plant also boasted a storage system that collected excess sun-warmed water in a large insulated tank for use at night and during inclement weather, in contrast to the solar motor designed by Mouchout, which went "into operation only when the sun [came] out from behind a cloud and [went] out of action the instant the sun disappears again." For industrialists, Shuman's solution eliminated a major obstacle to solar's appeal. Shuman's solar plant proved in Egypt to be more economical than a coal-fired plant. Engineers recognized that this plant demonstrated that "solar power was quite within the range of practical matters." However, the hopes and dreams of solar engineering disintegrated with the outbreak of World War I. The staff of the Egyptian plant had to leave for war-related work in their respective homelands. Shuman, the driving force behind large-scale solar developments, died before the war ended. After the war, the world turned to oil to replace coal. Oil and gas reserves were found in sunny and previously

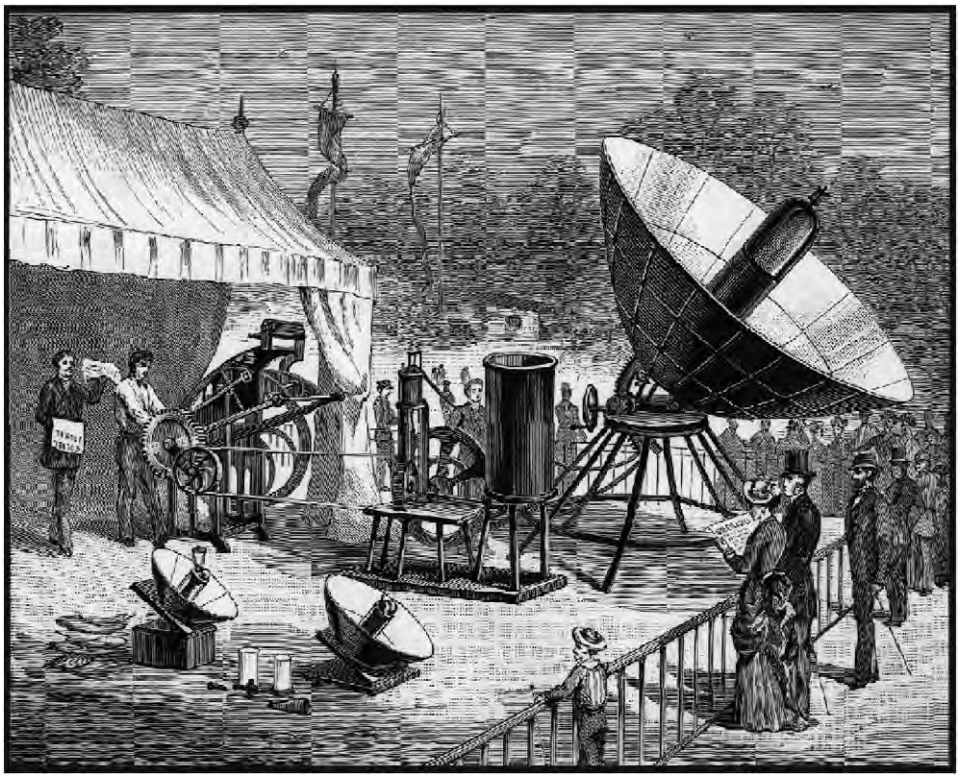


FIGURE 3 A solar-powered machine built by Mouchot's protégé, Pifre, runs a printing press. From Perlin/Butti Solar Archives.

sun-short places that had been targeted as prime locations for solar plats. With oil and gas selling at near-giveaway prices, scientists, government officials, and business people became complacent about the world's energy situation and interest in sun power came to an abrupt end.

3. SOLAR HOT WATER HEATING

3.1 Solar Hot Box

Horace de Saussure, a noted Swiss naturalist, observed in the 1760s, "It is a known fact, and a fact that has probably been known for a long time, that a room, a carriage, or any other place is hotter when the rays of the sun pass through glass." To determine the effectiveness of trapping heat with glass covers, de Saussure built a rectangular box out of half-inch pine, insulated the inside, covered the top with glass, and placed two smaller glass-covered boxes inside. When exposed to the sun, the bottom box heated to 228°F (109°C), or 16°F (9°C) above the boiling point of water. de Saussure was unsure how the sun heated the glass boxes. Today, we can better explain what happened. Sunshine penetrated the glass covers. The black inner lining absorbed the sunlight and converted it into heat. Although clear glass allows the rays of the sun to easily enter through it, it prevents heat from doing the same. As the glass trapped the solar heat in the box, it heated up. Its inventor realized that someday the hot box might have important practical applications because "it is quite small, inexpensive, and easy to make." Indeed, the hot box became the prototype for the solar collectors that have provided sun-heated water to millions since 1891.

3.2 The First Solar Water Heaters

In the 19th century, there was no easy way to heat water. People generally used a cook stove for this purpose. Wood had to be chopped or heavy hods of coal lifted, and then the fuel had to be kindled and the fire periodically stoked. In cities, the wealthier heated their water with gas manufactured from coal. However, the fuel did not burn clean and the heater had to be lit each time someone wanted to heat water. If someone forgot to extinguish the flame, the

tank would blow up. To add to the problem of heating water, in many areas wood or coal or coal gas cost a lot and many times could not be easily obtained. To circumvent these problems, many handy farmers, prospectors, or other outdoorsmen devised a much safer, easier, and cheaper way to heat water: They placed a metal water tank painted black into sunlight to absorb as much solar energy as possible. These were the first solar water heaters on record. The downside was that even on clear, hot days it usually took from morning to early afternoon for the water to get hot. As soon as the sun went down, the tanks rapidly lost heat because they had no protection from the night air.

3.3 The First Commercial Solar Water Heater

The shortcomings of the bare tank solar water heaters came to the attention of Clarence Kemp, who sold the latest home heating equipment in Baltimore, Maryland. In 1891, Kemp patented a way to combine the old practice of exposing metal tanks to the sun with the scientific principle of the hot box, thereby increasing the tanks' capability to collect and retain solar heat. He called his new solar water heater the Climax, the world's first commercial solar water heater (Fig. 4). Kemp originally marketed his invention to eastern gentlemen whose wives had left with their maids to summer resorts, leaving their husbands to fend for themselves. The solar water heater, Kemp advertised, would simplify housekeeping duties for this class of men already burdened by their wives and domestic staffs' absence and unaccustomed to such work as lighting the gas furnace or stove to heat water. In California and other temperate states, having greater amounts of sunshine throughout the year but also higher fuel costs (compared to states such as Maryland) made it essential for residents to take their solar assets seriously and not waste them (Fig. 5). The Climax sold well in such areas. Sixteen hundred had been installed in homes throughout southern California by 1900, including those installed in Pasadena, where one-third of the households heated their water with the sun by 1897.

3.4 The Flat Plate Collector

From the turn of the century to 1911, more than a dozen inventors filed patents that improved on the Climax. However, none changed the fact that the heating unit and the storage unit were one and the same and both laid exposed to the weather and

FIGURE 4 Advertisement for the world's first commercial solar water heater. From Perlin/Butti Solar Archives.

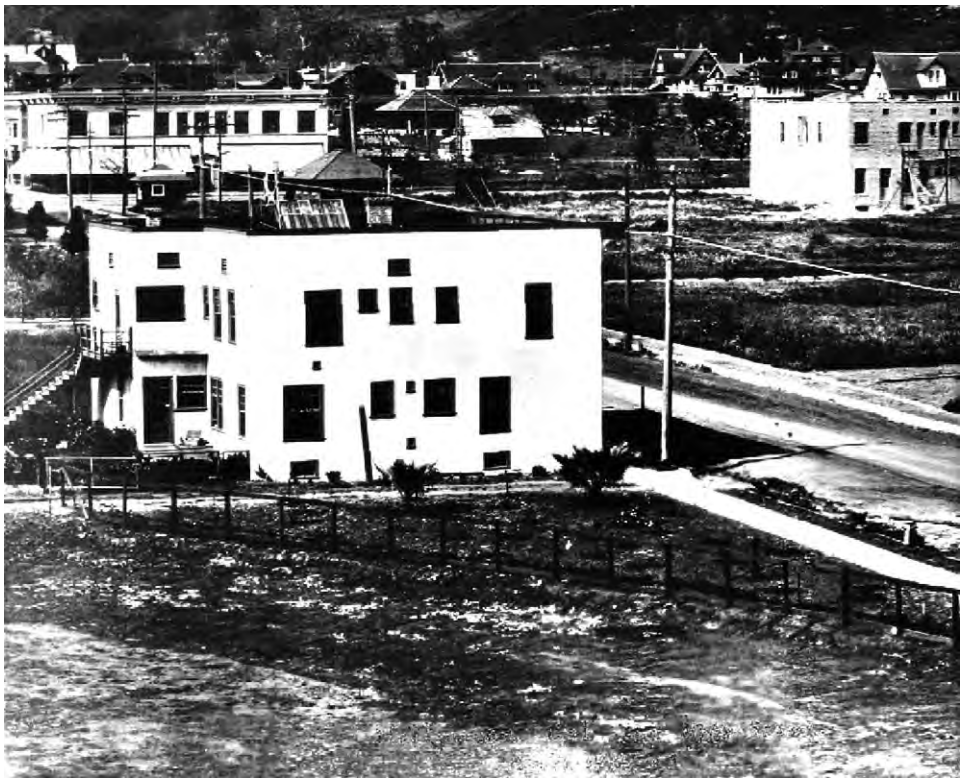


FIGURE 5 Installation of tanks in glass-covered box, Hollywood, California. From Perlin/Butti Solar Archives.

the cold night air. Hence, water heated by the sun the night before never stayed hot enough to do the wash the next morning or to heat the bath. In 1909, William J. Bailey patented a solar water heater that revolutionized the business. He separated the solar water heater into two parts: a heating element exposed to the sun and an insulated storage unit tucked away in the house so families could have sun-heated water day and night. The heating element consisted of pipes attached to a black-painted metal sheet placed in a glass-covered box. Because the water to be heated passed through narrow pipes rather than sitting in a large tank, Bailey reduced the volume of water exposed to the sun at any single moment and therefore the water heated faster. The heated water, which was lighter than the incoming cold water, naturally and immediately rose through the pipes to an insulated storage tank, where the water was kept warm for use both day and night. He also connected the storage tank to the home's furnace, guaranteeing hot water even after several rainy days. Providing hotter water for longer periods put Bailey's solar hot water heater, called the Day and Night (Fig. 6), at a great advantage over the competition. Soon, the Climax company went out of business. From 1909, when Bailey started his business, through 1918, his company sold more than 4000 Day and Night solar hot water heaters (Fig. 7).

3.5 The Demise of the Solar Water Heater in California and Its Success in Florida

The discovery of huge amounts of natural gas in the Los Angeles basin during the 1920s and 1930s killed the local solar water heater industry. Rather than lose money from the energy changes, Bailey applied the innovations he had made in solar to develop thermostatically controlled gas water heaters. His Day and Night gas water heater made him his second fortune. He also sold the patent rights of the Day and Night solar water heater to a Florida firm. A building boom in Florida during the 1920s had tripled the number of houses, but just as in California before the great oil strikes, people had to pay a high price to heat water. The high cost of energy combined with the tropical climate and the great growth in housing stock created a major business for those selling solar water heaters. By 1941, more than half the population heated water with the sun. Declining electric

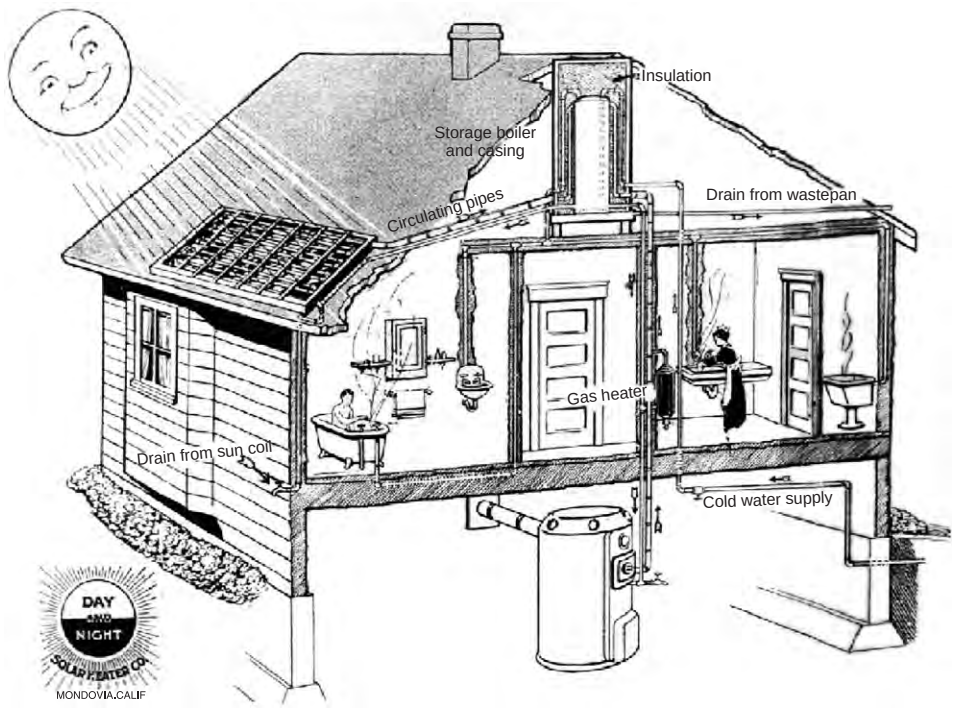


FIGURE 6 Schematic of the first flat plate solar collector with solar heat collection and solar storage separated. From Perlin/Butti Solar Archives.

FIGURE 7 An early flat plate collector installation. From Perlin/Butti Solar Archives.

rates after World War II, in association with an aggressive campaign by Florida Power and Light to increase electrical consumption by offering electric water heaters at bargain prices, brought Florida's once flourishing solar water heater industry to a screeching halt.

3.6 Solar Water Heaters in Japan

Unlike in the United States during the post-World War II years, Japan lacked cheap and abundant energy to supply hot water on demand. Rice farmers in particular yearned for a hot bath after working long hours in the hot, humid patties. To heat water, however, they had to burn rice straw, which they could have otherwise used to feed their cattle or fertilize the earth. A Japanese company began marketing a simple solar water heater consisting of a basin with its top covered by glass, and more than 100,000 were in use by the 1960s. People living in towns and cities bought either a plastic solar water heater that resembled an inflated air mattress with a clear plastic canopy or a more expensive but longer lasting model that resembled the old Climax solar water heaters—cylindrically shaped metal water tanks placed in a glass-covered box. Approximately 4 million of these solar water heaters had been sold by 1969. The advent of huge oil tankers in the 1960s allowed Japan access to new oil fields in the Middle East, supplying it with cheap, abundant fuel. As had happened in California and Florida, the solar water heater industry collapsed, but not for long. The oil embargo of 1973 and the subsequent dramatic increase in the price of petroleum revived the local solar water heater industry. Annual sales of more than 100,000 units continued to hold steady from 1973 until the second oil shock of 1979. Sales then increased to almost 500,000 that year and increased further to nearly 1 million the following year. By this time, the Japanese favored solar water heaters that closely resembled the type introduced in California in 1909 by William J. Bailey with the heating and storage units separated. As the price of oil began to stabilize in 1985 and then decline sharply in subsequent years, so did the sales of solar water heaters; however, the Japanese still purchase approximately 250,000 each year. Today, more than 10 million Japanese households heat water using the sun.

3.7 Solar Water Heating in Australia

From the 1950s to the early 1970s, a few thousand Australians relied on the sun to heat water. The numbers grew as a consequence of the major spikes in oil prices in 1973 and 1979. Interestingly, the purchase of solar water heaters during these years varied from state to state. Whereas 40–50% of those living in Australia's Northern Territory heated water using the sun, approximately 15% in Western Australia did so and less than 5% did so in the more populated eastern states. The major difference had more to do with the cost of electricity than the amount of sun available. People in the Northern Territory and Western Australia bought electricity generated by imported and increasingly costly petroleum, whereas for those in the eastern states of New South Wales, Queensland, and Victoria, electricity was produced by locally mined and very cheap coal. In the late 1980s, the Australian solar water heater market began to stagnate. Pipelines bringing newly discovered natural gas to previously fuel-short regions, such as the Northern Territory and Western Australia, eliminated growth in these once fertile markets for solar water heaters. Exports now account for more than 50% of the sales of Solahart, Australia's leading manufacturer of solar water heaters.

3.8 Solar Hot Water Heaters in Israel

Unlike the United States and much of Europe, Israel, like Japan, found itself without sufficient fuel supplies in the early 1950s. The power situation became so bleak in the early days of the Jewish state that the government had to forbid heating water between 10 PM and 6 AM. Despite mandatory domestic rationing of electricity, power shortages worsened, causing pumping stations to fail and threatening factory closures. A special committee impaneled by the government could only suggest the purchase of more centralized generators to overcome the problem. This conclusion raised the ire of Israeli engineer Levi Yissar, who complained, "How about an already existing energy source which our country has plenty of—the sun. Surely we need to change from electrical energy to solar energy, at least to heat our water." Yissar put his money where his mouth was, becoming Israel's first manufacturer of solar water heaters. His design closely resembled the type introduced in California in the early 20th century with heating and storage separated. By 1967, approximately 1 in 20 households heated water with the sun. However, cheap oil from Iran in the late 1960s as well from oil fields captured during the Six Day War drastically reduced the price of electricity and the number of people purchasing solar water heaters. Israeli success in the Yom Kippur War resulted in the infamous oil boycott of 1973. The Israelis responded by mass purchasing of solar water heaters. By 1983, 60% of the population heated water using the sun. When the price of oil dropped in the mid-1980s, the Israeli government did not want people backsliding in their energy habits as happened throughout the rest of the world. It therefore

required its inhabitants to heat water using the sun. Today, more than 90% of Israeli households own solar water heaters.

3.9 Solar Water Heating for Swimming Pools

Solar swimming pool water heaters rank as the most successful but least heralded commercial solar application. The use of solar energy for pool heating and the equipment and needs of pool owners make a perfect match. The storage unit for the solar heated water already exists—the swimming pool. The pump needed to push water through the solar collectors must be purchased irrespective of the technology used to heat the water. The pool owner merely has to purchase the solar collectors. Since those using the pool generally want the temperature of the pool to be no higher than 80°F (27°C), the solar collectors do not require a costly glass cover or expensive metal sheeting and piping. Climax solar water heaters kept pools warm in the 1890s. Later, collectors resembling those designed by William J. Bailey were put to use for this application. In the 1970s, American Freeman Ford developed low-cost plastic to act as the solar collector. Exposed to the sun, water would pass through narrow ducts in the plastic and carry enough heat to warm the pool. The outdoor swimming season harmoniously coincides with the maximum output of the solar collectors. Even with other forms of energy selling very cheaply, the pool owner who buys a solar unit will start to save money very quickly. In the United States alone, solar swimming pool heaters have produced the energy output equivalent of 10 nuclear power plants.

4. PHOTOVOLTAICS

All the methods discussed in the previous sections change solar energy into heat. In contrast, photovoltaics produces electricity directly from sunlight.

4.1 The First Solid-State Solar Cell

The direct ancestor of solar cells currently in use got its start in the late 19th century with the construction of the world's first seamless communication network—transoceanic telegraph cables. While laying them under water to permit instantaneous communications between continents, engineers experimented with selenium for detecting flaws in the wires as they were submerged. Those working with selenium discovered early on that the material's performance depended on the amount of sunlight falling on it. The influence of sunlight on selenium aroused the interest of scientists throughout Europe, including William Grylls Adams and his student Richard Evans Day. During one of their experiments with selenium, they witnessed something completely new: Light could cause a solid material to generate electricity.

However, with the science of the day not yet sure if atoms were real, scientists could not explain why selenium produced electricity when exposed to light. Most scientists of the day therefore scoffed at Adams and Day's work. It took the discovery and study of electrons and the discovery and acceptance that light contains packets of energy called photons for the field of photovoltaics to gain acceptance in the scientific community. By the mid-1920s, scientists theorized that when light hits materials such as selenium, the more powerful photons pack enough energy to knock poorly bound electrons from their orbits. When wires are attached, the liberated electrons flow through them as electricity. Many envisioned the day when banks of selenium solar cells would power factories and light homes. However, no one could build selenium solar cells efficient enough to convert more than half a percent of the sun's energy into electricity—hardly sufficient to justify their use as a power source.

4.2 The Discovery of the Silicon Solar Cell

An accidental discovery by scientists at Bell Laboratories in 1953 revolutionized solar cell technology. Gerald Pearson and Calvin Fuller led the pioneering effort that took the silicon transistor, now the principal electronic component used in all electrical equipment, from theory to working device. Fuller had devised a way to control the introduction of impurities necessary to transform silicon from a poor to a superior conductor of electricity. He gave Pearson a piece of his intentionally contaminated silicon. Among the experiments Pearson performed using the specially treated silicon was exposing it to the sun while hooked to a device that measured electrical flow. To Pearson's surprise, he observed an electric output almost five times greater than that of the best selenium produced. He ran down the hall to tell his good friend, Daryl Chapin, who had been

trying to improve selenium to provide small amounts of intermittent power for remote locations. He said, "Don't waste another moment on selenium!" and handed him the piece of silicon he had just tested. After a year of painstaking trial-and-error research and development, Bell Laboratories showed the public the first solar cells capable of producing useful amounts of power. Reporting the event on page 1, the *New York Times* stated that the work of Chapin, Fuller, and Pearson "may mark the beginning of a new era, leading eventually to the realization of one of mankind's most cherished dreams—the harnessing of the almost limitless energy of the sun for the uses of civilization."

Few inventions in the history of Bell Laboratories evoked as much media attention and public excitement as its unveiling of the silicon solar cell. Commercial success, however, failed to materialize due to the solar cells' prohibitive costs. Desperate to find marketable products run by solar cells, manufacturers used them to power novelty items such as toys and the newly developed transistor radio. With solar cells powering nothing but playthings, one of the inventors of the solar cell, Daryl Chapin, could not hide his disappointment, wondering, "What to do with our new baby?"

4.3 The First Practical Application of Silicon Solar Cells

Unbeknown to Chapin at the time, powerful backing of the silicon solar cell was developing at the Pentagon. In 1955, the U.S. government announced its intention to launch a satellite. The prototype had silicon solar cells for its power plant. Since power lines could not be strung out to space, satellites needed a reliable, long-lasting autonomous power source. Solar cells proved the perfect answer.

The launching of the Vanguard (Fig. 8), the first satellite equipped with solar cells, demonstrated their value. Preceding satellites, run by batteries, lost power after approximately 1 week, rendering millions of dollars' worth of equipment useless. In contrast, the solar-powered Vanguard continued to communicate with Earth for many years, allowing the completion of many valuable experiments. The success of the Vanguard's solar power pack broke down the existing prejudice at the time toward the use of solar cells in space. As the space race between the United States and the Soviet Union intensified, both adversaries urgently needed solar cells. The demand opened a relatively large business for companies manufacturing them. More important, for the first time in the history of solar power, the sun's energy proved indispensable to society: Without the secure, reliable electricity provided by photovoltaics, the vast majority of space applications so vital to our everyday lives would never have been realized.

FIGURE 8 An engineer holds the Vanguard 1, the first satellite to use photovoltaics. From Perlin/Butti Solar Archives.

4.4 The First Terrestrial Applications

Although things were looking up for solar cells in space in the 1960s and early 1970s, their astronomical price kept them distant from Earth. In 1968, Elliot Berman decided to quit his job as an industrial chemist to develop inexpensive solar cells to bring the technology from space to Earth. Berman prophetically saw that with a major decrease in price photovoltaics could play a significant role in supplying electrical power to locations on Earth where it is difficult to run power lines. After 18 months of searching for venture capital, Exxon executives invited Berman to join their laboratory in late 1969. Dismissing other solar scientists' obsession with efficiency, Berman concentrated on lowering costs by using lower grade and therefore cheaper silicon and less expensive materials for packaging the modules. By decreasing the price from $200 to $20 per watt, solar cells could compete with power equipment needed to generate electricity distant from utility poles.

4.5 Terrestrial Applications

Oil companies were the first major customers of solar modules. They had both the need and the money. Oil rigs in the Gulf of Mexico were required to have warning lights and horns. Most relied on huge flashlight-type batteries to run them. The batteries required heavy maintenance and also had to be replaced approximately every 9 months. This required a large boat with an onboard crane or a helicopter. In contrast, a small skiff could transport the much lighter solar module and accompanying rechargeable battery, resulting in tremendous savings. By 1980, photovoltaics had become the standard power source for warning lights and horns not only on rigs in the Gulf of Mexico but also on those throughout the world.

Oil and gas companies also need small amounts of electricity to protect well casings and pipelines from corroding. Sending a current into the ground electrochemically destroys the offending molecules that cause this problem. However, many oil and gas fields both in the United States and in other areas of the world such as the Middle East and North Africa are far away from power lines but receive plenty of sunshine. In such cases, solar modules have proven the most cost-effective way of providing the needed current to keep pipes and casings corrosion free.

The cost of changing nonrechargeable batteries on the U.S. Coast Guard's buoys exceeded the buoys' original cost. Hence, Lloyd Lomer, at the time a lieutenant commander in the Guard, believed that switching to photovoltaics as a power source for buoys made economic sense. However, his superiors, insulated from competition, balked at the proposed change. Lomer continued his crusade, winning approval to install a test system in the most challenging of environments for sun-run devices—Ketchikan, Alaska. The success of the photovoltaic-powered buoy in Alaska proved Lomer's point. Still, his boss refused to budge. Going to higher authorities in government, Lomer eventually won approval to convert all buoys to photovoltaics. Almost every coast guard service in the world has followed suit.

To solar pioneer Elliot Berman, the railroads seemed to be another natural for photovoltaics. One of his salesmen convinced the Southern Railway (now Norfolk Southern) to try photovoltaics for powering a crossing signal at Rex, Georgia. These seemingly fragile cells did not impress veteran railroad workers as capable of powering much of anything. They therefore put in a utility-tied backup. However, a strange thing happened that turned quite a few heads. That winter, the lines went down on several occasions due to heavy ice buildup on the wires, and the only electricity for miles around came from the solar array. As one skeptic remarked, "Rex, Georgia taught the Southern that solar worked!"

With the experiment at Rex a success, the Southern decided to put photovoltaics to work for the track circuitry, the railroad's equivalent of air-traffic control, keeping trains at a reasonable distance from one another to prevent head-on collisions or backenders. The presence of a train changes the rate of flow of electricity running through the track. A decoder translates that change to throw signals and switches up and down the track to ensure safe passage for all trains in the vicinity. At remote spots along the line, photovoltaics provided the needed electricity.

Other railroads followed the Southern's lead. In the old days, telegraph and then telephone poles ran parallel to most tracks. Messages sent through these wires kept stations abreast of matters paramount to the safe and smooth functioning of a rail line. However, by the mid-1970s, wireless communications could do the same tasks. The poles became a liability and maintenance expense. Railroads started to dismantle poles. Whenever they found their track circuitry devices to be too far away from utility lines, they relied on photovoltaics.

The U.S. Army Signal Corps, which pioneered the use of solar power in space by equipping the solar pack on the Vanguard, was the first to bring photovoltaics down to Earth. In June 1960, the Corps sponsored the first transcontinental radio broadcast generated by the sun's energy to celebrate its 100th anniversary. The Corps, probably photovoltaics' strongest supporter in the late 1950s and the early 1960s, envisioned the solar broadcast would lead others to use photovoltaics to help provide power to run radio and telephone networks in remote locations.

Fourteen years later, GTE John Oades realized the Corps' dream by installing the first photovoltaic-run microwave repeater in the rugged mountains above Monument Valley, Utah. By minimizing the power needs of the microwave repeater, Oades could power the repeater with a small photovoltaic panel, saving all the expenses formerly attributed to them. His invention allowed people living in towns in the rugged American west, hemmed in by mountains, to enjoy the luxury of long-distance phone service that most other Americans took for granted. Prior to the solar-powered repeater, it was too expensive for phone companies to install cable or

lines, forcing people to drive hours, sometimes through blizzards on windy roads, just to make long-distance calls.

Australia had an even more daunting challenge to bring modern telecommunication services to its rural customers. Although approximately the same size as the United States, only 22 million people lived in Australia in the early 1970s when its government mandated Telecom Australia to provide every citizen, no matter how remotely situated, with the same radio, telephone, and television service enjoyed by urban customers living in Sydney or Melbourne.

Telecom Australia tried, without success, traditional stand-alone power systems such as generators, wind machines, and nonrechargeable batteries to run autonomous telephone receivers and transmitters for its rural customers. Fortunately, by 1974, it had another option—relatively cheap solar cells manufactured by Elliot Berman's firm, Solar Power Corporation. The first photovoltaic-run telephone was installed in rural Victoria, after which hundreds of others were installed. By 1976, Telecom Australia judged photovoltaics as the preferred power source for remote telephones. In fact, the solar-powered telephones proved so successful that engineers at Telecom Australia believed they were ready to develop large photovoltaic-run telecommunication networks linking towns such as Devil's Marbles, Tea Tree, and Bullocky Bone to Australia's national telephone and television service. As a result of photovoltaics, people in these and neighboring towns could dial long distance direct instead of having to call the operator and shout into the phone to be understood. Nor did they have to wait for newsreels to be flown to their local station to view news, hours, if not days, old. There were a total of 70 solar-powered microwave repeater networks by the early 1980s, the longest spanning 1500 miles. The American and Australian successes showed the world that solar power worked and benefited hundreds of thousands of people. In fact, by 1985 a consensus in the telecommunications field found photovoltaics to be the power system of choice for remote communications.

A few years earlier, Mali, a country located south of the Sahara desert, along with its neighbors, suffered from drought so devastating that tens of thousands of people and livestock died. The Malian government knew it could not save its population without the help of people like Father Vespieren, a French priest who lived in Mali for decades and had successfully run several agricultural schools. The government asked Vespieren to form a private company to tap the vast aquifers that run underneath the Malian desert. Vespieren saw drilling as the easy part. His challenge was pumping. No power lines ran nearby and generators lay idle for lack of repairs or fuel. Then he heard about a water pump in Corsica that ran without moving parts, without fuel, and without a generating plant: It run simply on energy from the sun. Vespieren rushed to see the pioneering installation. When he saw the photovoltaic pump, he knew that only this technology could save the people of Mali. In the late 1970s, Vespieren dedicated Mali's first photovoltaic-powered water pump with the following words: "What joy, what hope we experience when we see that sun which once dried up our pools now replenishes them with water!" By 1980, Mali, one of the poorest countries in the world, had more photovoltaic water pumps per capita than any other country due to Vespieren's efforts. The priest demonstrated that success required the best equipment, a highly skilled and well-equipped maintenance service, and financial participation by consumers. Investing their own money helped pay the bill for the capital equipment and ongoing maintenance but, more important, the people came to regard the panels and pumps as valued items that they would help care for. Most successful photovoltaic water pump projects in the developing world have followed Vesperien's example. When Vespieren initiated his photovoltaic water-pumping program, less than 10 photovoltaic pumps existed throughout the world. Today, tens of thousands of photovoltaic pumps provide water to people, livestock, and crops.

Despite the success of photovoltaic applications throughout the world in the 1970s and early 1980s, institutions responsible for rural electrification programs in developing countries did not consider installing solar electric panels to power villages distant from urban areas. Working from offices either in the West or in large cities in poorer countries, they only considered central power stations run by nuclear energy, oil, or coal. Not having lived "out in the bush," these "experts" did not realize the huge investment required to string wires from power plants to the multitudes residing in small villages tens of miles away. As a consequence, only large populated areas received electricity, leaving billions in the countryside without. The disparity in energy distribution helped cause the great migration to the cities in Africa, Asia, and Latin America. As a consequence, megalopolises such as Mexico City, Lagos, and Bombay arose with all the ensuing problems, such as crime,

diseases like AIDS, pollution, and poverty. The vast majority still living in the countryside do not have electricity. To have some modicum of lighting, they have had to rely on ad hoc solutions such as kerosene lamps. Those buying radios, tape cassettes, and televisions also must purchase batteries. It has become apparent that mimicking the Western approach to electrification has not worked in the developing world. Instead of trying to install wires and poles, which no utility in these regions can afford, bringing a 20- to 30-year supply of electricity contained in solar modules to consumers by animal or by vehicle makes more sense. Whereas it requires years to construct a centralized power plant and even more time for power lines to be installed to deliver electricity, it takes less than 1 day to install an individual photovoltaic system.

Ironically, the French Atomic Energy Commission pioneered electrifying remote homes in the outlying Tahitian Islands with photovoltaics. The Atomic Energy Commission believed that electrifying Tahiti would mollify the ill feeling in the region created by its testing of nuclear bombs in the South Pacific. The commission considered all stand-alone possibilities, including generators, wind machines, and biogas from the husks of coconut shells, before choosing photovoltaics. In 1983, 20% of the world's production of solar cells went to French Polynesia. By 1987, half of all the homes on these islands received their electricity from the sun.

Rural residents in Kenya believe that if they have to wait for electricity from the national utility, they will be old and gray by the time they receive it. To have electricity now, tens of thousands of Kenyans have bought photovoltaic units. Though relatively small compared to what an American would purchase, the typical modules in Kenya range from 12 to 25 W and are capable of charging a battery with enough electricity to run three low-wattage lights and a television for 3 hours after dark. Electric lighting provided by photovoltaics eliminates the noxious fumes and the threat of fire that people had to contend with when lighting their homes with kerosene lamps. Due to photovoltaics, children can study free from eye strain and tearing and from developing hacking coughs that they formerly experienced when studying by kerosene lamp. Their performance in school has soared with better lighting. Electric lighting also allows women to sew and weave at night products that generate cash. Newly gained economic power gives women more say in matters such as contraception. In fact, solar electricity has proven more effective in lowering fertility rates than have birth control campaigns.

Electricity provided by photovoltaics allows those living in rural areas to enjoy the amenities of urban areas without leaving their traditional homes for the cities. The government of Mongolia, for example, believed that to improve the lot of its nomadic citizens would require herding them into communities so the government could connect them to electricity generated by centralized power. The nomads, however, balked. Photovoltaics allowed the nomadic Mongolians to take part in the government's rural electrification program without giving up their traditional lifestyle. Whenever they move, the solar panels move too, with yurt and yaks and whatever else they value.

With the growing popularity of photovoltaics in the developing world, solar thievery has become common. Perhaps nothing better demonstrates the high value that people living in these countries place on photovoltaics than the drastic increase in solar module thefts during the past few years. Before the market for photovoltaics took off in the developing world, farmers would build an adobe fence around their photovoltaic panels to keep the livestock out. Now, razor wire curls around the top of these enclosures to bar solar outlaws.

As the price of solar cells has continued to decline, devices run by solar cells have seeped into the suburban and urban landscapes of the developed world. For example, when construction crews have to excavate to place underground electrical transmission lines, installing photovoltaics instead makes more economic sense. Economics has guided highway departments throughout the world to use solar cells for emergency call boxes along roads. In California alone, there are more than 30,000. The savings have been immediate. Anaheim, California, would have had to spend approximately $11 million to connect its 11,000 call boxes to the power grid. Choosing photovoltaics instead, the city paid $4.5 million. The city of Las Vegas found it cheaper to install photovoltaics to illuminate its new bus shelters than to have the adjacent street and sidewalk dug up to place new wires.

In the mid-1970s and early 1980s, when governments of developing countries began to fund photovoltaic projects, they copied the way electricity had been traditionally produced and delivered, favoring the construction of large fields of photovoltaic panels far away from where the electricity would be used and delivering the electricity by building miles of transmission lines. Others, however, questioned this

approach. When Charles Fritts built the first selenium photovoltaic module in the 1880s and boldly predicted that it would soon compete with Thomas Edison's generators, he envisioned "each building to have its own power plant." Since the late 1980s, most in the photovoltaic business have come around to Fritts's view. Instead of having to buy a huge amount of land—unthinkable in densely populated Europe and Japan—to place a field of panels, many began to ask why not turn each building into its own power station. Using photovoltaics as building material allows it to double as windows, roofing, skylights, facades, or any type of covering needed on a home or building, and it makes sense since the owner gets both building material and electrical generator in one package. Having the electrical production sited where the electricity will be used offers many advantages: Producing electricity on site eliminates the electricity lost in long-distance transmission, which ranges from approximately 30% in the best maintained lines to approximately 65% in rundown lines prevalent in the developing world; it reduces the danger of power lines being overloaded as demand increases by rerouting additional electricity directly to the users; and occasionally, such as on a hot August afternoon, when peak electrical demand empties the wires, excess electricity produced by photovoltaic-equipped homes and buildings can refill the lines with much needed electricity to prevent brownouts and blackouts.

4.6 The Future of Photovoltaics

Although the photovoltaics industry has experienced a phenomenal 20% annual growth rate during the past decade, it has just started to realize its potential. Although more than 1 million people in the developing world get their electricity from solar cells, more than 1 billion still have no electrical service. The continuing revolution in telecommunications is bringing a greater emphasis on the use of photovoltaics. As with electrical service, the expense of stringing telephone wires keeps most of the developing world without communication services that people in more developed countries take for granted. Photovoltaic-run satellites and cellular sites, or a combination of the two, offer the only hope to bridge the digital divide. Photovoltaics could allow everyone the freedom to dial up at or near home and, of course, hook up to the Internet.

Opportunities for photovoltaics in the developed world also continue to increase. In the United States and Western Europe, hundreds of thousands of permanent or vacation homes are too distant for utility electric service. If people live more that 250 yards from a utility pole, paying the utility to string wires to their houses cost more than supplying their power needs with photovoltaics. Fourteen thousand Swiss Alpine chalets receive electricity from solar electricity, as do tens of thousands of others from Finland to Spain to Colorado.

Many campgrounds now prohibit RV owners from running their engines to power generators for appliances inside. The exhaust gases pollute, and the noise irritates other campers, especially at night. Photovoltaic panels mounted on the roofs of RVs provide the electricity needed without bothering others.

Restricting carbon dioxide emissions to help moderate global warming could result in the flow of large amounts of money from projects in which fossil fuels are burned to photovoltaic projects. The damage wrought by the 1997–1998 El Nino is an indication of the harsher weather expected as the earth warms. The anticipated increase in natural disasters brought on by a more disastrous future climate, as well as the growing number of people living in catastrophic-prone regions, makes early warning systems essential. The ultimate early warning device may consist of pilotless photovoltaic-powered weather surveillance airplanes, of which the *Helios* serves as a prototype. The *Helios* has flown higher than any other aircraft. Solar cells make up the entire top of the aircraft, which is merely a flying wing. Successors to the *Helios* will also have fuel cells on the underside of the wing. They will get their power from the photovoltaic panels throughout the day, extracting hydrogen and oxygen from the water discharged by the fuel cells the night before. When the sun sets, the hydrogen and oxygen will power the fuel cells, generating enough electricity at night to run the aircraft. Water discharged in the process will allow the diurnal cycle to begin the next morning. The tandem use of solar cells and fuel cells will allow the aircraft to stay aloft forever, far above the turbulence, watching for and tracking hurricanes and other potentially dangerous weather and other natural catastrophes.

Revolutionary lighting elements, called light-emitting diodes (LEDs), produce the same quality of illumination as their predecessor with only a fraction of the energy. LEDs therefore significantly reduce the amount of panels and batteries necessary for running lights, making photovoltaic systems less costly and less cumbersome. They have enabled

photovoltaics to replace gasoline generators in mobile warning signs used on roadways that alert motorists of lane closures and other temporary problems that drivers should know about. The eventual replacement of household lighting by LEDs will do the same for photovoltaics in homes.

When Bell Laboratories first unveiled the silicon solar cell, their publicist made a bold prediction: "The ability of transistors to operate on very low power gives solar cells great potential and it seems inevitable that the two Bell inventions will be closely linked in many important future developments that will influence the art of living." Already, the tandem use of transistors and solar cells for running satellites, navigation aids, microwave repeaters, televisions, radios, and cassette players in the developing world and a myriad of other devices has fulfilled the Bell prediction. It takes no great leap of imagination to expect the transistor/solar cell revolution to continue until it encompasses every electrical need from space to earth.

SEE ALSO THE FOLLOWING ARTICLES

Coal Industry, History of • Electricity Use, History of • Hydrogen, History of • Hydropower, History and Technology of • Manufactured Gas, History of • Materials for Solar Energy • Natural Gas, History of • Nuclear Power, History of • Oil Industry, History of • Solar Cells • Solar Fuels and Materials • Sun, Energy from • Wind Energy, History of • Wood Energy, History of

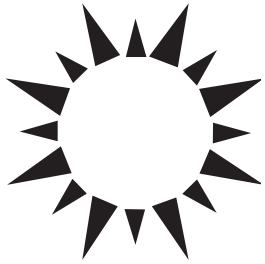
Further Reading

Beattie, D. A. (ed.). (1997). "History and Overview of Solar Heat Technologies." MIT Press, Cambridge, MA.

Butti, K., and Perlin, J. (1980). "Golden Thread: Twenty-five Hundred Years of Solar Architecture and Technology." Van Nostrand, New York.

Perlin, J. (2002). "From Space to Earth: The Story of Solar Electricity." Harvard University Press, Cambridge, Massachusetts.

Solar Fuels and Materials



ALDO STEINFELD
ETH–Swiss Federal Institute of Technology
Zurich, Switzerland

ANTON MEIER
Paul Scherrer Institute
Villigen, Switzerland

1. Introduction
2. Thermodynamics of Solar Thermochemical Conversion
3. Solar Thermochemical Processes
4. Outlook

Glossary

aperture Opening of a solar cavity receiver.
Carnot efficiency Maximum efficiency for converting heat from a high-temperature thermal reservoir at T_H into work in a cyclic process and rejecting heat to a low-temperature thermal reservoir at T_L, given by $1-T_L/T_H$.
exergy efficiency (for a solar thermochemical process) The efficiency for converting solar energy into chemical energy, given by the ratio of the maximum work that may be extracted from a solar fuel to the solar energy input for producing such a fuel.
solar cavity-receiver A well-insulated enclosure, with a small opening to let in concentrated solar energy, that approaches a black-body absorber in its ability to capture solar energy.
solar concentration ratio Nondimensional ratio of the solar flux intensity (e.g., in "suns") achieved after concentration to the incident normal beam insolation.
solar thermochemical process Any endothermic process that uses concentrated solar energy as the source of high-temperature process heat.

This article develops the underlying science of solar thermochemical processes and describes their applications for delivering clean fuels and material commodities. The thermodynamics of solar thermochemical conversion are examined, the most promising solar thermochemical processes are discussed, examples of solar chemical reactors are presented, and the latest technological developments are summarized.

1. INTRODUCTION

Using only 0.1% of the earth's land space with solar collectors that operate with a collection efficiency of merely 20%, one could gather more than enough energy to supply the current yearly energy needs of all inhabitants of the planet ($\sim 1.2 \times 10^{14}$ kWh). Furthermore, the solar energy reserve is essentially unlimited, and its use is ecologically benign. However, solar radiation is dilute (only $\sim 1\,\mathrm{kW/m^2}$), intermittent (available only during daytime and under clear-sky conditions), and unequally distributed (mostly by the equator). These drawbacks can be overcome by converting solar energy into chemical energy carriers—solar fuels—that can be stored long term and transported long range, from the sunny and desert regions to the industrialized and populated centers of the earth, where much of the energy is needed. Solar fuels can be burned to generate heat, further processed into electrical or mechanical work, or used directly to generate electricity in fuel cells and batteries to meet customers' energy demands. Solar process heat can also assist in the processing and recycling of energy-intensive materials, thereby avoiding greenhouse gas emissions and other pollutants derived from the combustion of fossil fuels for heat and electricity generation.

There are basically three pathways for making solar fuels from solar energy:

- *Solar electrochemical path:* solar-made electricity, from photovoltaic or solar thermal systems, followed by an electrolytic process
- *Solar photochemical path:* direct use of the photon energy
- *Solar thermochemical path:* solar-made heat followed by a thermochemical process

 © 2004 Elsevier Inc. All rights reserved.

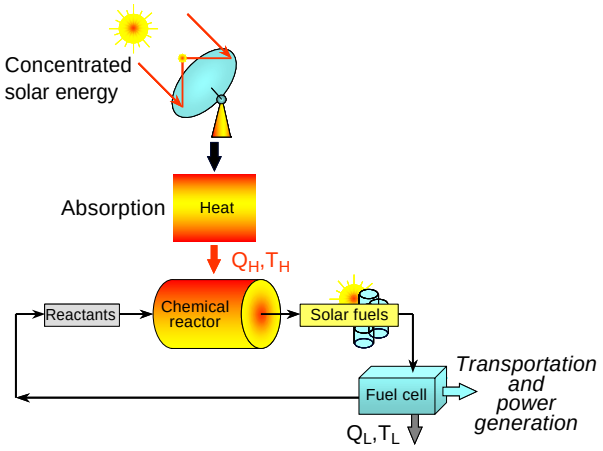


FIGURE 1 Schematic of solar energy conversion into solar fuels. Concentrated solar radiation is used as the energy source for high-temperature process heat to drive endothermic chemical reactions toward the production of storable and transportable fuels.

Combinations of these three pathways are possible, but the thermochemical path offers some intriguing thermodynamic advantages. It is this approach that is the subject of this article, whereas the other two paths are beyond the scope of the article.

Figure 1 illustrates the basic idea of the solar thermochemical path. By concentrating the diluted sunlight over a small area with help of parabolic mirrors and then capturing the radiant energy with help of suitable receivers and reactors, we can obtain heat at high temperatures for carrying out an endothermic chemical transformation and producing storable and transportable fuels. These solar fuels ultimately store solar energy within their chemical bonds. Regardless of the fuel, the higher the temperature of the reaction process, the higher the efficiency of energy conversion. However, higher temperatures also lead to greater losses by reradiation from the solar receiver. The thermodynamic implications of such phenomena are examined in the following section.

2. THERMODYNAMICS OF SOLAR THERMOCHEMICAL CONVERSION

The state-of-the-art technology for large-scale collection and concentration of solar energy is based on three main optical configurations using parabolic-shaped mirrors. These are the trough, tower, and dish systems shown schematically in Fig. 2. Trough

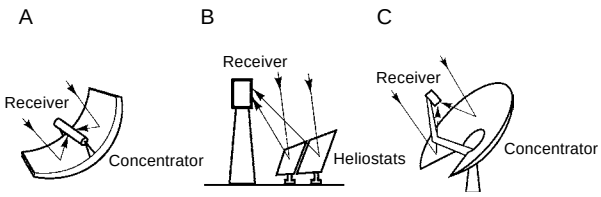


FIGURE 2 Schematic of the three main optical configurations for large-scale collection and concentration of solar energy: (A) the trough system, (B) the tower system, and (C) the dish system.

systems use linear, two-dimensional, parabolic mirrors to focus sunlight onto a solar tubular receiver positioned along their focal line. Tower systems use a field of heliostats (two-axis tracking parabolic mirrors) that focus the sunrays onto a solar receiver mounted on top of a centrally located tower. Dish systems use paraboloidal mirrors to focus sunlight on a solar receiver positioned at their focus. A recently developed Cassegrain optical configuration for the tower system makes use of a hyperboloidal reflector at the top of the tower to redirect sunlight to a receiver located on the ground level. Solar furnaces are concentrating facilities in which high-flux solar intensities are usually obtained at a fixed location inside a housed laboratory. The traditional design consists of using a sun-tracking, flat heliostat on-axis with a stationary primary paraboloidal concentrator; off-axis configurations have also been designed.

The capability of these collection systems to concentrate solar energy is often expressed in terms of their mean flux concentration ratio $\tilde{C}$ over a targeted area A at the focal plane, normalized with respect to the incident normal beam insolation I as follows:

$$\tilde{C} = \frac{Q_{solar}}{I \cdot A}, \tag{1}$$

where Q_{solar} is the solar power input into the target. $\tilde{C}$ is often expressed in units of "suns" when normalized to $1\,kW/m^2$. The solar flux concentration ratio typically obtained at the focal plane varies between 30 and 100 suns for trough systems, between 500 and 5,000 suns for tower systems, and between 1,000 and 10,000 suns for dish systems. Higher concentration ratios imply lower heat losses from smaller receivers and, consequently, higher attainable temperatures at the receiver. To some extent, the flux concentration can be further augmented with the help of a nonimaging secondary concentrator (e.g., a compound parabolic concentrator [CPC]) that is positioned in tandem with the primary concentrating system. The aforementioned

solar concentrating systems have proved to be technically feasible in large-scale experimental demonstrations aimed mainly at the production of solar thermal electricity in which a working fluid (typically air, water, helium, sodium, or molten salt) is solar heated and further used in traditional Rankine, Brayton, and Stirling cycles. Solar thermochemical applications, although not developed to the extent that solar thermal electricity generation has been, employ the same solar concentrating technology.

Solar chemical reactors for highly concentrated solar systems usually feature the use of a cavity receiver-type configuration, that is, a well-insulated enclosure with a small opening (the aperture) to let in concentrated solar radiation. Because of multiple internal reflections, the fraction of the incoming energy absorbed by the cavity greatly exceeds the simple surface absorptance of the inner walls. The larger the ratio of the cavity's characteristic length to the aperture diameter, the closer the cavity receiver approaches a black-body absorber. Smaller apertures also serve to reduce reradiation losses. However, they intercept a reduced fraction of the sunlight reflected from the concentrators. Consequently, the optimum aperture size is a compromise between maximizing radiation capture and minimizing radiation losses.

The solar energy absorption efficiency of a solar reactor, $\eta_{absorption}$, is defined as the net rate at which energy is being absorbed divided by the solar power coming from the concentrator. At temperatures above approximately 1000 K, the net power absorbed is diminished mostly by radiant losses through the aperture. For a perfectly insulated black-body cavity receiver, it is given by

$$\eta_{absorption} = 1 - \left(\frac{\sigma T^4}{I\tilde{C}}\right), \tag{2}$$

where T is the nominal cavity-receiver temperature and σ is the Stefan–Boltzmann constant. The absorbed concentrated solar radiation drives an endothermic chemical reaction. The measure of how well solar energy is converted into chemical energy for a given process is the exergy efficiency, defined as

$$\eta_{exergy} = \frac{-\dot{n}\Delta G|_{298\ K}}{Q_{solar}}, \tag{3}$$

where Q_{solar} is the solar power input, $\dot{n}$ is the products' molar flow rate, and ΔG is the maximum possible amount of work (Gibbs free energy change) that may be extracted from the products as they are transformed back to reactants at 298 K. The second law of thermodynamics is now applied to calculate the theoretical maximum exergy efficiency $\eta_{exergy,ideal}$. Because the conversion of solar process heat to chemical energy is limited by both the solar absorption efficiency and the Carnot efficiency, the maximum ideal exergy efficiency is given by

$$\eta_{exergy,ideal} = \eta_{absorption} \cdot \eta_{Carnot} = \left[1 - \left(\frac{\sigma T_H^4}{I\tilde{C}}\right)\right] \times \left[1 - \left(\frac{T_L}{T_H}\right)\right], \tag{4}$$

where T_H and T_L are the upper and lower operating temperatures of the equivalent Carnot heat engine. $\eta_{exergy,ideal}$ is plotted in Fig. 3 as a function of T_H for $T_L = 298$ K and for various solar flux concentration ratios. Because of the Carnot limit, one should try to operate thermochemical processes at the highest upper temperature possible; however, from a heat transfer perspective, the higher the temperature, the higher the reradiation losses. The highest temperature an ideal solar cavity receiver is capable of achieving, defined as the stagnation temperature $T_{stagnation}$, is calculated by setting $\eta_{exergy,ideal}$ equal to zero, yielding

$$T_{stagnation} = \left(\frac{I\tilde{C}}{\sigma}\right)^{0.25}. \tag{5}$$

At this temperature, energy is being reradiated as fast as it is absorbed. Stagnation temperatures exceeding 3000 K are attainable with solar concentration ratios above 5000 suns. However, an energy-efficient

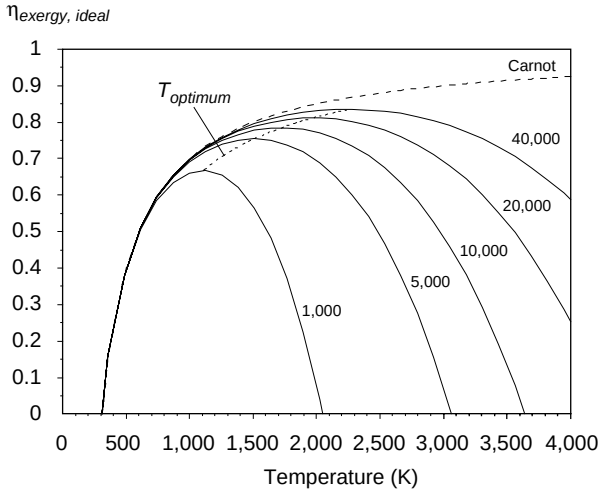


FIGURE 3 Variation of the ideal exergy efficiency $\eta_{exergy,ideal}$ as a function of the operating temperature T_H for a black-body cavity receiver converting concentrated solar energy into chemical energy. The mean solar flux concentration is the parameter: 1,000 … 40,000. Also plotted are the Carnot efficiency and the locus of the optimum cavity temperature $T_{optimum}$.

process must run at temperatures that are substantially below $T_{stagnation}$. There is an optimum temperature $T_{optimum}$ for maximum efficiency obtained by setting

$$\frac{\partial \eta_{exergy,ideal}}{\partial T} = 0. \quad (6)$$

Assuming a uniform power flux distribution, this relation yields the following implicit equation for $T_{optimum}$:

$$T_{optimum}^5 - \frac{3}{4} T_L T_{optimum}^4 - \frac{T_L I\tilde{C}}{4\sigma} = 0. \quad (7)$$

The locus of $T_{optimum}$ is shown in Fig. 3. For example, for a solar concentration ratio of 5000, the optimum temperature of a solar receiver is 1500 K, giving a maximum theoretical efficiency of 75%, that is, the portion of solar energy that could in principle be converted into the chemical energy of fuels. In practice, when considering convection and conduction losses in addition to reradiation losses, the efficiency will peak at a somewhat lower temperature.

The process modeling described in the following paragraphs establishes a base for evaluating and comparing different solar thermochemical processes in terms of their ideal exergy efficiencies. Figure 4 depicts a flow diagram of an ideal cyclic process that leads to a fuel, which uses a solar reactor, a quenching device, and a fuel cell. The complete process is carried out at constant pressure. In practice, pressure drops will occur throughout the system. However, if one assumes frictionless operating conditions, no pumping work is required. The reactants may be preheated in an adiabatic heat exchanger, where some portion of the sensible and latent heat of the products is transferred to the reactants; for simplicity, a heat exchanger has been omitted. The reactor is assumed to be a perfect blackbody cavity receiver. The reactants enter the solar reactor at T_L and are further heated to the reaction temperature T_H. Chemical equilibrium is assumed inside the reactor. The net power absorbed in the solar reactor should match the enthalpy change per unit time of the reaction,

$$Q_{reactor,net} = \dot{n}\Delta H \mid_{\text{Reactants@}T_L \rightarrow \text{Products@}T_H}. \quad (8)$$

Irreversibility in the solar reactor arises from the nonreversible chemical transformation and reradiation losses $Q_{reradiation}$ to the surroundings at T_L. It is found that

$$Irr_{reactor} = \left(\frac{-Q_{solar}}{T_H}\right) + \left(\frac{Q_{reradiation}}{T_L}\right) + \left(\dot{n}\Delta S|_{\text{Reactants@}T_L \rightarrow \text{Products@}T_H}\right). \quad (9)$$

Products exit the solar reactor at T_H and are cooled rapidly to T_L. The amount of power lost during quenching is

$$Q_{quench} = -\dot{n}\Delta H \mid_{\text{Products@}T_H \rightarrow \text{Products@}T_L} \quad (10)$$

The irreversibility associated with quenching is

$$Irr_{quench} = \left(\frac{Q_{quench}}{T_L}\right) + \left(\dot{n}\Delta S|_{Products@T_H \rightarrow Products@T_L}\right). \quad (11)$$

Introducing a reversible ideal fuel cell, in which the products recombine to form the original reactants and thereby generate electrical power, closes the cycle. The maximum amount of work that the products leaving the reactor could produce if they combined at T_L and a total pressure of 1 bar is given by

$$W_{F.C.} = -\dot{n}\Delta G|_{Products@T_L \rightarrow Reactants@T_L}. \quad (12)$$

This work value is also known as the exergy of the products at ambient temperature. The fuel cell operates isothermally; the amount of heat rejected to the surroundings is

$$Q_{F.C.} = -T_L \cdot \dot{n}\Delta S|_{Products@\,T_L \rightarrow Reactants@T_L}. \quad (13)$$

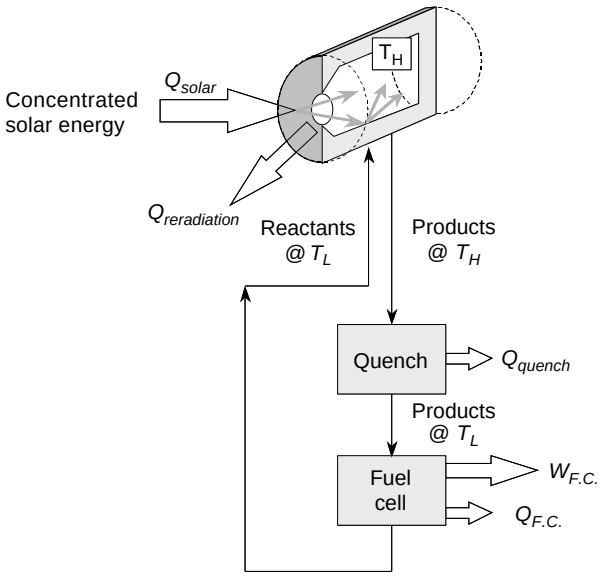


FIGURE 4 Model flow diagram of an ideal cyclic process for calculating the maximum exergy efficiency of a solar thermochemical process.

The exergy system efficiency of the closed cycle is then calculated using Eq. (3) as

$$\eta_{exergy} = \frac{W_{F.C.}}{Q_{solar}}. \quad (14)$$

This thermodynamic analysis can be verified by performing an energy balance and by evaluating the maximum achievable efficiency (Carnot efficiency) from the total available work and from the total power input. The energy balance should satisfy

$$W_{F.C.} = Q_{solar} - \left(Q_{reradiation} + Q_{quench} + Q_{F.C.}\right). \quad (15)$$

The available work is calculated as the sum of the fuel cell work plus the lost work due to the irreversibility associated with the solar reactor and with quenching. Thus,

$$\eta_{max} = \frac{W_{F.C.} + T_L \cdot \left(Irr_{reactor} + Irr_{quench}\right)}{Q_{solar}} = 1 - \frac{T_L}{T_H} = \eta_{Carnot}. \quad (16)$$

Any solar thermochemical processes can be thought of in this manner, and their exergy efficiencies can be compared as one criterion for judging their relative industrial potentials. The higher the exergy efficiency, the lower the required solar collection area for producing a given amount of solar fuel and, consequently, the lower the costs incurred for the solar concentrating system, which usually correspond to half of the total investments for the entire solar chemical plant. Thus, high-exergy efficiency implies favorable competitiveness.

3. SOLAR THERMOCHEMICAL PROCESSES

3.1 Solar Production of Fuels

3.1.1 Solar Hydrogen

Five thermochemical routes for solar hydrogen production are depicted in Fig. 5. Indicated is the chemical source of H_2:water for the solar thermolysis and solar thermochemical cycles, fossil fuels for the solar cracking, and a combination of fossil fuels and H_2O for the solar reforming and solar gasification. All of these routes make use of concentrated solar radiation as the energy source of high-temperature process heat.

3.1.1.1 H_2 from H_2O Although the single-step thermal dissociation of water, known as water thermolysis, is conceptually simple, it has been

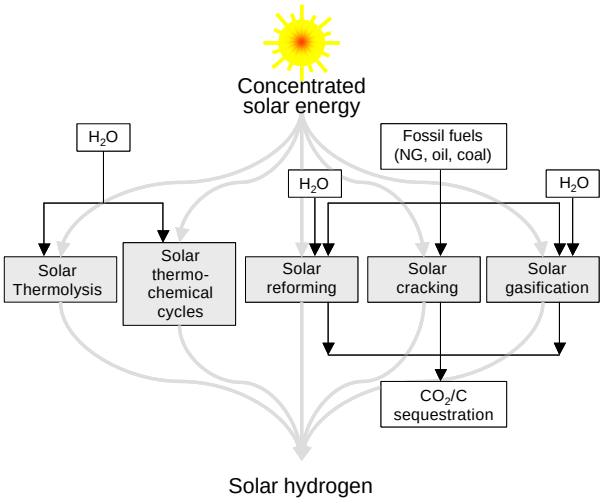


FIGURE 5 Five thermochemical routes for the production of solar hydrogen.

impeded by the need for a high-temperature heat source at above 2500 K to achieve a reasonable degree of dissociation and by the need for an effective technique to separate H_2 and O_2 to avoid recombination or ending up with an explosive mixture. Among the ideas proposed for separating H_2 from the products are effusion and electrolytic separation. Water-splitting thermochemical cycles bypass the H_2/O_2 separation problem and further allow operating at relatively moderate upper temperatures. Previous studies performed on H_2O-splitting thermochemical cycles were characterized mostly by the use of process heat at temperatures below approximately 1300 K, available from nuclear and other thermal sources. These cycles required multiple (more than two) steps and were suffering from inherent inefficiencies associated with heat transfer and product separation at each step. During recent years, significant progress has been made in the development of optical systems for large-scale collection and concentration of solar energy capable of achieving solar concentration ratios of 5000 suns or higher. Such high-radiation fluxes allow the conversion of solar energy to thermal reservoirs at 2000 K and above that are needed for the two-step thermochemical cycles using metal oxide redox reactions:

$$\text{First step (solar):} \quad M_xO_y \rightarrow xM + \frac{y}{2}O_2 \quad (17)$$

$$\text{Second step (nonsolar):} \quad xM + yH_2O \rightarrow M_xO_y + yH_2 \quad (18)$$

Here, M denotes a metal and M_xO_y denotes the corresponding metal oxide. The cycle is represented

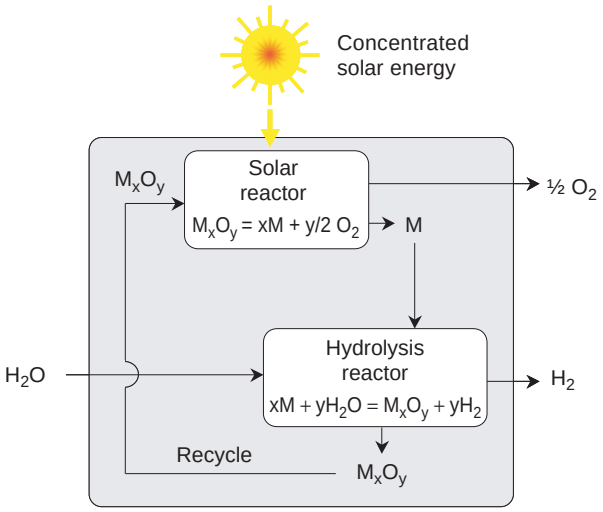


FIGURE 6 Representation of a two-step water-splitting thermochemical cycle using metal oxide redox reactions.

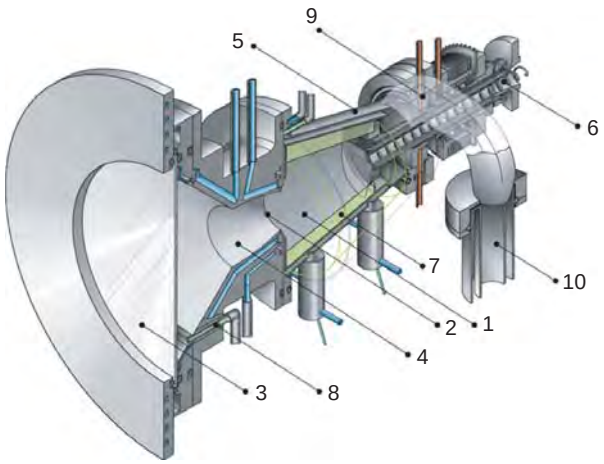


FIGURE 7 Schematic of a "rotating-cavity" solar chemical reactor for the thermal dissociation of ZnO to Zn and O_2 at 2300 K. It consists of a rotating conical cavity receiver (1) that contains an aperture (2) for access of concentrated solar radiation through a quartz window (3). The solar flux concentration is further augmented by incorporating a CPC (4) in front of the aperture. Both the window mount and the CPC are water cooled and integrated into a concentric (nonrotating) conical shell (5). ZnO particles are fed continuously by means of a screw powder feeder located at the rear of the reactor (6). The centripetal acceleration forces the ZnO powder to the wall, where it forms a thick layer of ZnO (7) that insulates and reduces the thermal load on the inner cavity walls. A purge gas flow enters the cavity receiver tangentially at the front (8) and keeps the window cool and clear of particles or condensable gases. The gaseous products Zn and O_2 exit continuously via an outlet port (9) to a quench device (10). Source: Paul Scherrer Institute, Villigen, Switzerland.

in Fig. 6. The first endothermic step is the solar thermal dissociation of the metal oxide to the metal or the lower valence metal oxide. The second, nonsolar exothermic step is the hydrolysis of the metal to form H_2 and the corresponding metal oxide. The net reaction is $H_2O = H_2 + 0.5O_2$, but because H_2 and O_2 are formed in different steps, the need for high-temperature gas separation is thereby eliminated. The most favorable candidate metal oxide redox pairs for this cycle are ZnO/Zn and Fe_3O_4/FeO. In both cases, the minimum temperature requirement for the solar step is 2300 K, and the products need to be either quenched or separated at high temperatures to prevent their recombination. Quenching introduces irreversibility and may be a factor of complexity in large-scale use.

Figure 7 shows a solar chemical reactor concept for performing the thermal dissociation of ZnO that features a windowed rotating cavity receiver lined with ZnO particles that are held by centrifugal force. With this arrangement, ZnO is directly exposed to high-flux solar irradiation and serves the functions of radiant absorber, thermal insulator, and chemical reactant simultaneously. The shell of the cavity is made from conventional steel materials and is lined with the same material as are the reactants themselves. This aspect of the design eliminates the need for using expensive and difficult-to-fabricate ceramic insulating materials for ultra-high temperatures. It also offers excellent resistance to thermal shocks that are intrinsic in short start-up solar applications. The direct absorption of concentrated solar energy by directly irradiated reactants provides efficient radiation heat transfer to the reaction site where the energy is needed, bypassing the limitations imposed by indirect heat transport via heat exchangers.

We illustrate the application of the Second-Law (exergy) analysis, Eqs. (1) to (16), for calculating the theoretical maximum exergy efficiency of the two-step water-splitting solar thermochemical cycle based on the ZnO/Zn redox reaction. Figure 8 depicts the ideal cyclic process. The solar reactor is assumed to be a cavity-receiver operating at 2300 K. The reactant ZnO(s) enters the solar reactor at 298 K and is further heated to the reaction temperature at 2300 K. The molar feed rate of ZnO to the solar reactor, $\dot{n}$, is set to 1 mol/s and is equal to that of H_2O fed to the hydrolyser reactor. After quenching, the products separate naturally (without expending work) into gaseous O_2 and condensed phase Zn. Zinc is sent to the hydrolyser reactor to react exothermally with water and form hydrogen, according to Eq. (18). The heat liberated is assumed

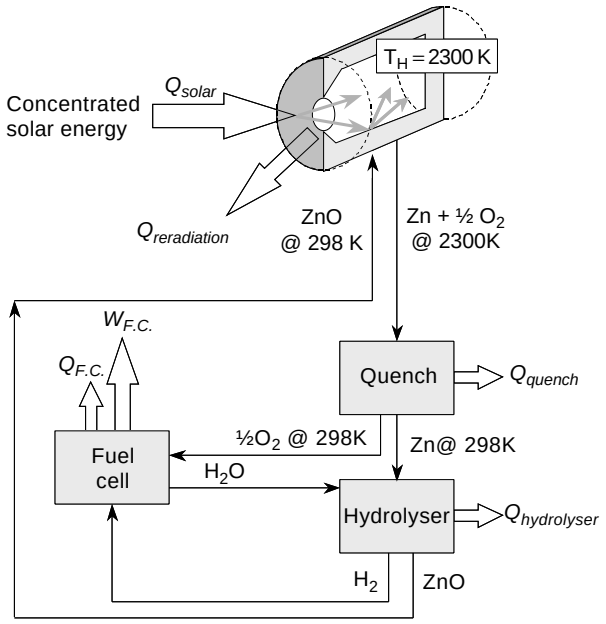


FIGURE 8 Model flow diagram of the two-step water-splitting solar thermochemical cycle used for the exergy analysis.

lost to the surroundings, as given by

$$Q_{hydrolyser} = -\dot{n}\Delta H|_{Zn+H_2O@298\ K \rightarrow ZnO+H_2@298\ K}. \quad (19)$$

Thus, the irreversibility associated with the hydrolyser is

$$Irr_{hydrolyser} = \left(\frac{Q_{hydrolyser}}{298\ K}\right) + \left(\dot{n}\Delta S|_{\ Zn+H_2O@298\ K \rightarrow ZnO+H_2@298\ K}\right). \quad (20)$$

Introducing an H_2/O_2 fuel cell, in which the products recombine to form H_2O and thereby generate electrical power, closes the cycle. $W_{F.C.}$ and $Q_{F.C.}$ are the work output and the heat rejected, respectively, given by

$$W_{F.C.} = -\eta_{F.C.}\dot{n}\Delta G|_{H_2+0.5O_2@298\ K \rightarrow H_2O@298\ K} \quad (21)$$

$$Q_{F.C.} = -T_L\ \dot{n}\Delta S|_{H_2+0.5O_2@298\ K \rightarrow H_2O@298\ K}, \quad (22)$$

where $\eta_{F.C.}$ is the efficiency of the H_2/O_2 fuel cell, assumed to be 100% when based on the maximum work output of H_2 (also known as its exergy). Finally, the exergy efficiency is calculated using Eq. (14). Table I gives the results of the calculation for two values of the solar concentration ratio: 5,000 and 10,000.

TABLE I

Exergy Analysis of the Two-Step Water-Splitting Solar Thermochemical Cycle Using the Process Modeling Depicted in Fig. 8

$\tilde{C}$	5,000	10,000
Q_{solar}	815 kW	662 kW
$Q_{reradiation}$	258 kW	105 kW
$Q_{reactor,net}$	557 kW	
$Irr_{reactor}$	0.81 kW · K^{-1}	
Q_{quench}	209 kW	
Irr_{quench}	0.52 kW · K^{-1}	
$Q_{hydrolyser}$	64 kW	
$Irr_{hydrolyser}$	0.27 kW · K^{-1}	
$Q_{F.C.}$	49 kW	
$W_{F.C.}$	237 kW	
$\eta_{absorption}$	68%	84%
η_{exergy} (without heat recovery)	29%	36%
η_{exergy} (with heat recovery)	69%	84%

Note. The ZnO and H_2O molar rates are set to 1 mol/s.

3.1.1.2 H_2 from Fossil Fuels Three solar thermochemical processes for H_2 production using fossil fuels as the chemical source are considered: cracking, reforming, and gasification. These routes are shown schematically in Fig. 9.

The solar cracking route refers to the thermal decomposition of natural gas (NG), oil, and other hydrocarbons and can be represented by the simplified net reaction:

$$C_xH_y = xC(gr) + \frac{y}{2}H_2. \quad (23)$$

Other compounds may also be formed, depending on the reaction kinetics and on the presence of impurities in the raw materials. The thermal decomposition yields a carbon-rich condensed phase and a hydrogen-rich gas phase. The carbonaceous solid product can either be sequestered without CO_2 release or be used as a material commodity under less severe CO_2 restraints. It can also be applied as a reducing agent in metallurgical processes. The hydrogen-rich gas mixture can be further processed to high-purity hydrogen that is not contaminated with oxides of carbon and, thus, can be used in proton exchange membrane (PEM) fuel cells without inhibiting platinum-made electrodes. From the point of view of carbon sequestration, it is easier to separate, handle, transport, and store solid carbon than gaseous CO_2. Furthermore, while the steam-reforming/gasification method requires additional steps for shifting CO and for separating CO_2, the thermal cracking accomplishes the removal and separation of carbon in a single step. In contrast,

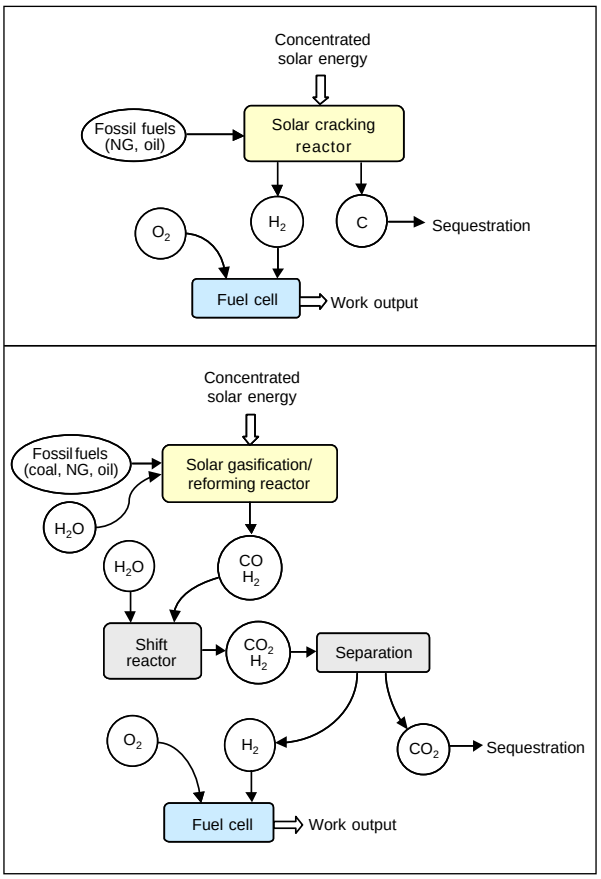


FIGURE 9 Schematic of solar thermochemical routes for H_2 production using fossil fuels and H_2O as the chemical source: solar cracking (upper box) and solar reforming and gasification (lower box).

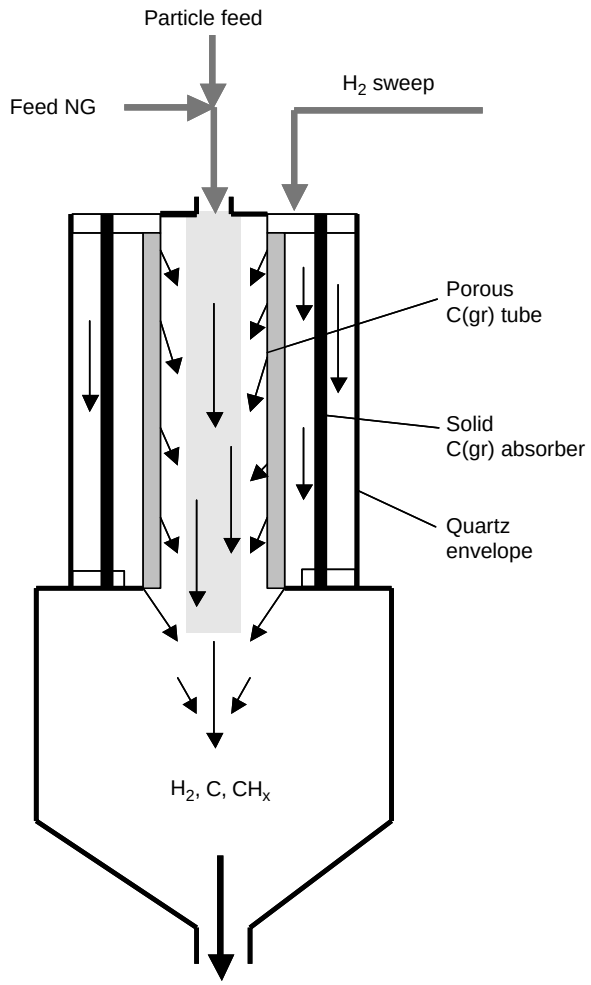


FIGURE 10 Scheme of an "aerosol" solar reactor for cracking NG. Source: National Renewable Energy Laboratory, Golden, CO.

the major drawback of the thermal decomposition method is the energy loss associated with the sequestration of carbon. Thus, the solar cracking may be the preferred option for NG and other hydrocarbons with high H_2/carbon ratios. The exergy efficiency for these routes is defined as the ratio of the work output by the fuel cell to the total thermal energy input by both the solar energy and the heating value of the reactants:

$$\eta_{exergy} = \frac{W_{F.C.}}{Q_{solar} + HHV_{reactant}}. \qquad (24)$$

where $HHV_{reactant}$ is the high heating value of the fossil fuel being processed (e.g., $\sim 890\,\mathrm{kJ \cdot mol^{-1}}$ for NG). Assuming a 65% efficient H_2/O_2 fuel cell, $\eta_{exergy} = 30\%$. This route offers zero CO_2 emissions as a result of carbon sequestration. However, the energy penalty for completely avoiding CO_2 reaches 30% of the electrical output vis-à-vis the direct use of NG for fueling a 55% efficient combined Brayton–Rankine cycle. Figure 10 shows an "aerosol" solar reactor concept for cracking NG that features two concentric graphite tubular reactors, with the outer solid tube serving as the solar absorber and the inner porous tube containing a flow of NG laden with carbon nanoparticles that serve the functions of radiant absorbers.

The steam reforming of NG, oil, and other hydrocarbons, as well as the steam gasification of coal and other carbonaceous materials, can be represented by the simplified net reaction:

$$C_xH_y + xH_2O = \left(\frac{y}{2} + x\right)H_2 + xCO. \qquad (25)$$

Other compounds may also be formed, especially with coal, but some impurities contained in the raw materials are cleaned out prior to the decarbonization process. The principal product is high-quality synthesis gas (syngas), the building block for a wide

variety of synthetic fuels, including Fischer–Tropsch-type chemicals, hydrogen, ammonia, and methanol. Its quality is determined mainly by the H_2/CO and CO_2/CO molar ratios. For example, the solar steam gasification of anthracite coal at above 1500 K yields syngas with an H_2/CO molar ratio of 1.2 and a CO_2/CO molar ratio of 0.01. The CO content in the syngas can be shifted to H_2 via the catalytic water–gas shift reaction ($CO + H_2O = H_2 + CO_2$), and the product CO_2 can be separated from H_2 using, for example, the pressure swing adsorption technique. The exergy efficiency for this route is defined by Eq. (24). Assuming a 65% efficient H_2/O_2 fuel cell and $HHV_{reactant}$ of 35,700 kJ kg^{-1} for anthracite coal, $\eta_{exergy} = 46\%$. This route offers a net gain in the electrical output by a factor varying in the range of 1.7 to 1.8 (depending on the coal type) vis-à-vis the direct use of coal for fueling a 35% efficient Rankine cycle. If CO_2 is not sequestered, specific CO_2 emissions amount to 0.53 to 0.56 kg CO_2/kWh$_e$, approximately half as much as the specific CO_2 emissions discharged by conventional coal-fired power plants.

Some of these processes are practiced at an industrial scale, with the process heat supplied by burning a significant portion of the feedstock. Internal combustion results in the contamination of the gaseous products, whereas external combustion results in a lower thermal efficiency due to the irreversibility associated with indirect heat transfer. Alternatively, using solar energy for process heat offers a threefold advantage: (1) the discharge of pollutants is avoided, (2) the gaseous products are not contaminated, and (3) the calorific value of the fuel is upgraded by adding solar energy in an amount equal to the ΔH of the reaction.

The solar reforming of NG has been studied extensively in solar furnaces with small-scale solar reactor prototypes using Rh-based catalyst. Recently, this solar process has been scaled up to power levels of 300 to 500 kW in a solar tower facility using two solar reactor concepts: indirect irradiation and direct irradiation.

Figure 11 shows a schematic of the indirect-irradiated solar reforming reactor. It consists of a pentagonal cavity receiver, insulated with ceramic fibers, containing a set of Inconel tubes arranged vertically in two banks parallel to the front-side walls. The tubes are filled with a packed bed of catalyst, usually 2% Rh on Al_2O_3 support. The tubular reactors are externally exposed to the concentrated solar radiation entering through the cavity aperture and transfer the absorbed heat to the chemical reactants. The aperture is windowless and tilted from the horizontal axis for an optimal view to the heliostat field. A matching CPC is implemented at the aperture for capturing radiation spillage, augmenting the average solar flux concentration, and providing uniform heating of the tubes. Indirect-irradiated reactors such as this one have the advantage of eliminating the need for a transparent window. The disadvantages are linked to the limitations imposed by the materials of construction of the reactor walls: limitations in the maximum operating temperature, thermal conductivity, radiant absorptance, inertness, resistance to thermal shocks, and suitability for transient operation.

Figure 12 shows a schematic of the direct-irradiated solar reforming reactor, also referred to

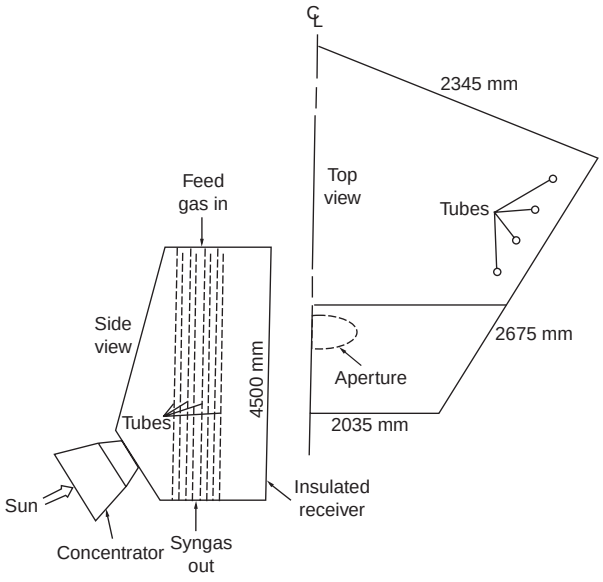


FIGURE 11 Scheme of an indirect-irradiated solar reforming chemical reactor. Source: Weizmann Institute of Science, Rehovot, Israel.

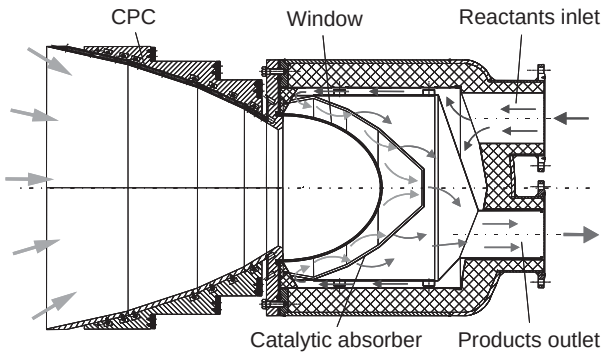


FIGURE 12 Scheme of a direct-irradiated solar reforming chemical reactor. Source: Deutsches Zentrum für Luft- und Raumfahrt, Germany.

as the volumetric reactor. The main component is the porous ceramic absorber, coated with Rh catalyst, that is directly exposed to the concentrated solar radiation. A concave quartz window, mounted at the aperture, minimizes reflection losses and permits operation at high pressures. Similar to the indirect-irradiated reactor, a CPC is implemented at the aperture. Direct-irradiated reactors such as this one have the advantage of providing efficient radiation heat transfer directly to the reaction. Furthermore, under proper conditions, direct irradiation may photochemically enhance the kinetics of the reaction. The major drawback, when working with reducing or inert atmospheres, is the requirement of a transparent window, which is a critical and troublesome component under high-pressure and severe gas environment.

The solar dry reforming of CH_4 (with CO_2) can be applied in a closed-loop chemical heat pipe (Fig. 13). The product of this reversible reaction is syngas that can be stored at ambient temperatures and transported to the site where the energy is needed. By the reverse exothermic reaction, stored solar energy is released in the form of high-temperature heat, which can be used, for example, for generating electricity via a Rankine cycle. The products of this reverse reaction are again recycled to the solar reactor, where the process is repeated. The dissociation/synthesis of NH_3 can also be applied in a chemical heat pipe for storage and transportation of solar energy.

An optional source of H_2 is H_2S, a highly toxic industrial product recovered in large quantities in the sweetening of NG and in the removal of organically

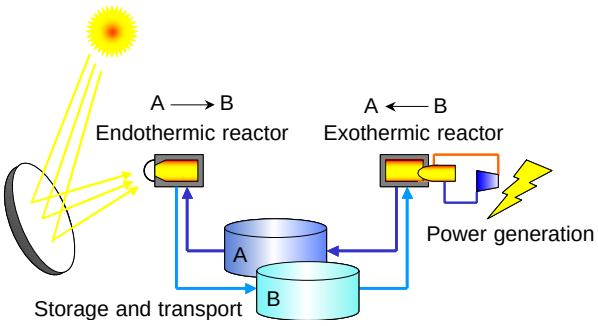


FIGURE 13 Solar chemical heat pipe for the storage and transportation of solar energy. High-temperature solar process heat is used to drive the endothermic reversible reaction $A \rightarrow B$. The product B may be stored long term and transported long range to the site where the energy is needed. At that site, the exothermic reverse reaction $B \rightarrow A$ is effected and yields high-temperature process heat in an amount equal to the stored solar energy $\Delta H_{A \rightarrow B}$. The chemical product A of the reverse reaction is returned to the solar reactor for reuse.

bound sulfur from petroleum and coal. Current industrial practice uses the Claus process to recover the sulfur from H_2S, but the process wastes H_2 to produce low-grade process heat. Alternatively, H_2S can be decomposed to H_2 and S_2 at 1800 K and 0.03 to 0.5 bar. The product gas mixture is quenched, and condensed S_2 is separated from H_2. In contrast to H_2O thermolysis, the quench is relatively easy and the reverse reaction between the products seems to be unimportant at temperatures as high as 1500 K.

3.1.2 Solar Metals

Metals are attractive candidates for storage and transport of solar energy. They may be used to generate either high-temperature heat via combustion or electricity via fuel cells and batteries. Metals can also be used to produce hydrogen from H_2O via hydrolysis, as shown in Eq. (18). The chemical products from any of these power/H_2-generating processes are metal oxides, which in turn need to be recycled. The conventional extraction of metals from their oxides by carbothermic and electrolytic processes is characterized by its high energy consumption and its concomitant environmental pollution. The extractive metallurgical industry discharges vast amounts of greenhouse gases and other pollutants to the environment, derived mainly from the combustion of fossil fuels for heat and electricity generation. These emissions can be reduced substantially, or even eliminated completely, by using concentrated solar energy as the source of high-temperature process heat.

The thermal dissociation and electrothermal reduction of metal oxides proceeds without reducing agents, while the carbothermal reduction of metal oxides uses solid carbon C(gr) or hydrocarbons (e.g., CH_4) as reducing agents. The corresponding overall chemical reactions may be represented as follows:

$$M_xO_y \rightarrow xM + \frac{y}{2}O_2 \qquad (26)$$

$$M_xO_y + yC(gr) \rightarrow xM + yCO \qquad (27)$$

$$M_xO_y + yCH_4 \rightarrow xM + y(2H_2 + CO). \qquad (28)$$

Table II lists the approximate temperatures at which the standard ΔG°_{rxn} for Reactions (26), (27), and (28) equals zero for various metal oxides of interest.

Except for the thermal dissociation of ZnO, the required temperature for effecting Reaction (26) exceeds 3000 K. At temperatures for which ΔG°_{rxn} is positive, solar process heat alone will not make the reaction proceed; some amount of high-quality energy is required in the form of work. It may be

TABLE II

Approximate Temperatures (K) for Which ΔG°_{rxn} of Reactions (26), (27), and (28) Equals Zero

Metal oxide	$\Delta G^{\circ}_{rxn\ 26} = 0$ @	$\Delta G^{\circ}_{rxn\ 27} = 0$ @	$\Delta G^{\circ}_{rxn\ 28} = 0$ @
Fe_2O_3[a]	3050	920	890
Al_2O_3	>4000	2320	1770
MgO	3700	2120	1760
ZnO	2340	1220	1110
TiO_2[a]	>4000	2040	1570
SiO_2[a]	4500	1950	1520
CaO	>4000	2440	1970

[a] Fe_2O_3, TiO_2, and SiO_2 decompose to lower valence oxides before complete dissociation to the metal.

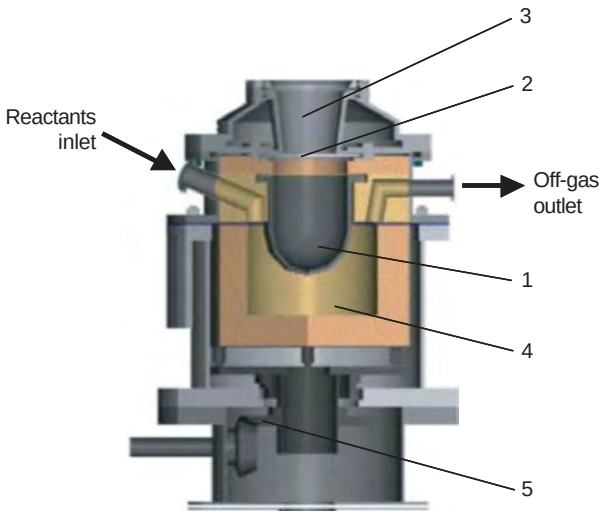


FIGURE 14 Schematic of a "two-cavity" solar chemical reactor concept for the carbothermal reduction of ZnO. It features two cavities in series, with the inner one functioning as the solar absorber and the outer one functioning as the reaction chamber. The inner cavity (1) is made of graphite and contains a windowed aperture (2) to let in concentrated solar radiation. A CPC (3) is implemented at the reactor's aperture. The outer cavity (4) is well insulated and contains the ZnO–carbon mixture that is subjected to irradiation by the graphite absorber separating the two cavities. With this arrangement, the inner cavity protects the window against particles and condensable gases coming from the reaction chamber. Uniform distribution of continuously fed reactants is achieved by rotating the outer cavity (5). The reactor is designed specifically for beam-down incident radiation, as obtained through a Cassegrain optical configuration that makes use of a hyperbolical reflector at the top of the tower to redirect sunlight to a receiver located on the ground level. Source: Paul Scherrer Institute, Villigen, Switzerland.

supplied in the form of electrical energy in electrolytic processes or in the form of chemical energy by introducing a reducing agent in thermochemical processes. An example of a solar electrothermal reduction process that has been demonstrated experimentally is the electrolysis of ZnO using an electrolytic cell housed in a solar cavity-receiver. At 1000 K, up to 30% of the total amount of energy required to produce Zn could be supplied by solar process heat. Other interesting candidates for solar high-temperature electrolysis are MgO and Al_2O_3.

If one wishes to decompose metal oxides thermally into their elements without the application of electrical work, a chemical reducing agent is necessary to lower the dissociation temperature. Coal (as coke) and NG (as methane) are preferred reducing agents in blast furnace processes due to their availability and relatively low price. In the presence of carbon, the uptake of oxygen by the formation of CO brings about a reduction of the oxides at much lower temperatures. Although Reactions (27) and (28) have favorable free energies above the temperatures indicated in Table II, a more detailed calculation of the chemical equilibrium composition shows that only the carbothermic reductions of Fe_2O_3, MgO, and ZnO will result in significant free metal formation. The carbides Al_3C_4, CaC_2, SiC, and TiC are thermodynamically stable in an inert atmosphere; the nitrides AlN, Si_3N_4, and TiN are stable in an N_2 atmosphere.

Using NG as a reducing agent, as indicated in Eq. (28), combines in a single process the reduction of metal oxides with the reforming of NG for the coproduction of metals and syngas. Thus, CH_4 is reformed in the absence of catalysts and, with proper optimizations, may be made to produce high-quality syngas with an H_2/CO molar ratio of 2, which is especially suitable for synthesizing methanol—a potential substitute for petrol. Thermal reductions of Fe_3O_4 and ZnO with C(gr) and CH_4 to produce Fe, Zn, and syngas have been demonstrated in solar furnaces using fluidized bed and vortex-type reactors. These reactions are highly endothermic and proceed to completion at temperatures above approximately 1300 K. Zinc production by solar carbothermic reduction of ZnO offers a CO_2 emission reduction by a factor of 5 vis-à-vis the conventional fossil fuel-based electrolytic or imperial smelting process. Furthermore, the use of solar energy for supplying the enthalpy of the reaction upgrades the calorific value of the initial reactants by 39%.

Two examples of solar chemical reactor concepts for producing Zn by Reactions (27) and (28) are shown in Figs. 14 and 15, respectively: the "two-cavity" solar reactor based on the indirect irradiation

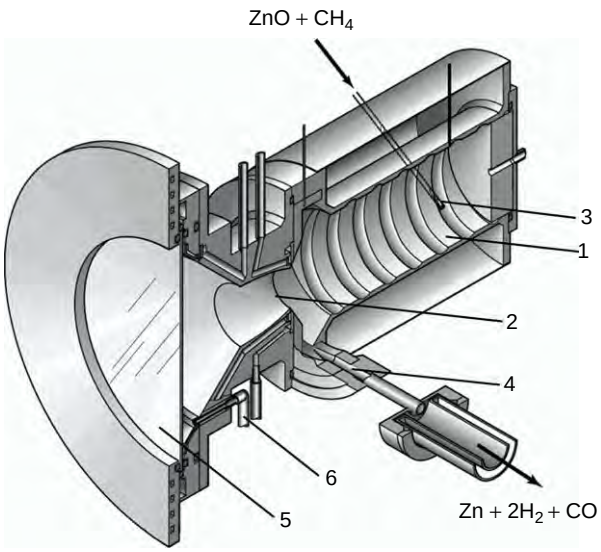


FIGURE 15 Schematic of a "vortex" solar chemical reactor concept for the combined ZnO reduction and CH_4 reforming. It consists of a cylindrical cavity (1) that contains a windowed aperture (2) to let in concentrated solar energy. Particles of ZnO, conveyed in a flow of NG, are continuously injected into the reactor's cavity via a tangential inlet port (3). Inside the reactor's cavity, the gas–particle stream forms a vortex flow that progresses toward the front following a helical path. The chemical products, Zn vapor and syngas, exit the cavity continuously via a tangential outlet port (4) located at the front of the cavity, behind the aperture. The window (5) is actively cooled and kept clear of particles by means of an auxiliary flow of gas (6) that is injected tangentially and radially at the window and aperture planes, respectively. Energy absorbed by the reactants is used to raise their temperature to above approximately 1300 K and to drive Reaction (28). Source: Paul Scherrer Institute, Villigen, Switzerland.

of ZnO + carbon and the "vortex" solar reactor based on the direct irradiation of ZnO + methane.

3.2 Solar Production of Material Commodities

3.2.1 Solar Metals, Metallic Oxides, Carbides, and Nitrides

The extractive metallurgical industry is a major consumer of process heat and electricity and is responsible for approximately 10% of global anthropogenic greenhouse gas emissions. A solar thermal process would drastically reduce these emissions. For example, solar processing of aluminum could reduce the CO_2 emissions per ton of aluminum from 44 to 5. The temperature required to reduce aluminum ores with carbon is in the range of 2400 to 2600 K, that is, at the upper limit for practical process heat addition from a combustion source alone. Preliminary tests performed in a concentrating solar facility suggest two feasible solar thermal routes. First, at approximately 2400 K, a mixture of Al_2O_3, SiO_2, and carbon results in an Al–Si alloy that can be purified after its production. Second, at above 2500 K, Al_2O_3 and carbon results in an Al–Al_3C_4 mixture of varying compositions.

As described in the previous subsection, Zn can be produced from ZnO by a solar thermal, electrothermal, or carbothermal reduction process. Other examples of solar thermal reduction processes that have been demonstrated in solar furnaces' exploratory runs are the production of Zn by dissociation of its natural ore willemite (Zn_2SiO_4), the combined ZnO reduction and biomass pyrolysis, and the production of Mg and Fe by carbothermal reduction of their oxides.

Metallic carbides and nitrides can be produced by the solar carbothermic reduction of metal oxides. The reaction is generally given by

$$M_xO_y + C(gr)\{+N_2\} \rightarrow \{M_{x'}C_{y'}, M_{x''}N_{y''}\} + CO. \quad (29)$$

Exploratory experimental studies were conducted in solar furnaces in which nitrides (AlN, Si_3N_4, TiN, and ZrN) were formed for the systems run in N_2 and carbides (Al_4C_3, CaC_2, SiC, TiC) were formed for the systems run in inert atmosphere. These ceramics are valuable materials for high-temperature applications due to their high hardness, excellent corrosion resistance, high melting points, and low coefficients of thermal expansion. CaC_2 is well known as the feedstock for the production of acetylene. The nitrides and carbides AlN, Fe_3C, and Mn_3C may also be used in cyclic processes as feedstock to produce hydrogen and hydrocarbons or may serve as intermediaries in the production of the metal. The hydrolysis of AlN yields NH_3, the hydrolysis or acidolysis of Fe_3C yields liquid hydrocarbons, and the hydrolysis of the various carbides of manganese yields H_2 and hydrocarbons in different proportions. Thus, Reaction (29) may be incorporated into cyclic processes of the type shown in Fig. 16, in which the metal oxides that result from the hydrolysis are recycled to the solar reactor, and CO from Reaction (29) may be shifted to H_2 via the water–gas shift reaction.

3.2.2 Solar Fullerenes and Carbon Nanotubes

Fullerenes and carbon nanotubes have become a major field in condensed matter physics and chemistry. However, the large-scale production of these materials with high yield and selectivity is still a

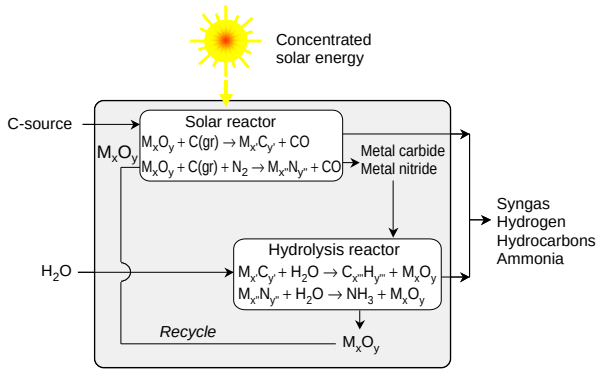


FIGURE 16 Schematic of a two-step thermochemical process for the production of syngas, hydrogen, hydrocarbons, and ammonia using metal oxides and solar energy. In the first (endothermic) step, the metal oxide is carbothermally reduced to a metal carbide or nitride using solar process heat. Subsequently, the carbide or nitride is reacted with water to produce hydrocarbons or ammonia, and CO is water–gas shifted to syngas, whereas the metal oxide is recycled to the solar step. M, metal; M_xO_y, metal oxide; $M_{x'}C_{y'}$, metal carbide; $M_{x''}N_{y''}$, metal nitride.

crucial problem. Conventional methods for the synthesis of fullerenes, such as electric arc discharge and laser ablation, fail when the process is being scaled up to higher power levels. The solar vaporization process seems to be more promising using the solar reactor shown in Fig. 17, in which a graphite rod contained under vacuum pressure behind a hemispherical quartz window is directly irradiated with peak solar concentration ratios exceeding 7000 suns and the vaporized carbon (at temperatures above 3300 K) is swept out by Ar, quenched, and collected in a filter bag. The key parameters characterizing this process are the carbon soot mass flow rate and the desired product yield. The former is a function of the target temperature, whereas the latter is a function of specific reactor variables such as fluid flow patterns, residence times, concentration of the carbon vapor in the carrier gas, target temperature, and temperature distribution in the cooling zone.

Catalytic filamentous carbon (CFC) can be produced by the solar thermal decomposition of hydrocarbons in the presence of small metal catalyst particles. Solar furnace experiments confirmed that nanotubes can be obtained using Co/MgO catalyst for CO and $CH_4 + H_2$ and that nanofibers can be obtained on Ni/Al_2O_3 catalyst for CO, CH_4, $CH_4 + H_2$, and C_4H_{10}. Figure 18 shows a solar chemical reactor used for such experimental runs. It consists of a quartz tube containing a fluidized bed of catalyst and Al_2O_3 grains. A secondary reflector, composed of a two-dimensional CPC

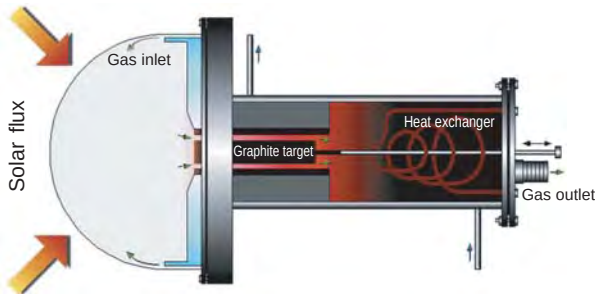


FIGURE 17 Scheme of a solar reactor for fullerene synthesis. Source: Institut de Science et de Génie des Matériaux et Prodédés, Centre National de la Recherche Scientifique, Font-Romeu, France.

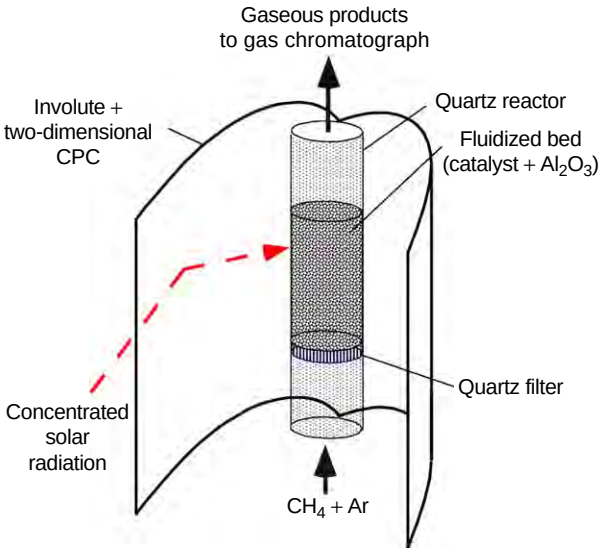


FIGURE 18 Scheme of a solar reactor for CFC production. Source: Paul Scherrer Institute, Villigen, Switzerland.

coupled to an involute, provides uniform irradiation on the tubular reactor. CFC formed typically has the following properties: surface area 100 to 170 m^2/g, pore volume 0.4 to 0.8 cm^3/g, micropore volume 0.004 to 0.008 cm^3/g, and average pore diameter 10 to 40 nm.

3.2.3 Solar Lime and Cement

The thermal decomposition of calcium carbonate (limestone) to calcium oxide (lime),

$$CaCO_3 \rightarrow CaO + CO_2, \qquad (30)$$

is the main endothermic step in the production of lime and cement at 1300 K. Substituting concentrated solar energy in place of carbonaceous fuels, as the source of high-temperature process heat, is a means to reduce the dependence on conventional

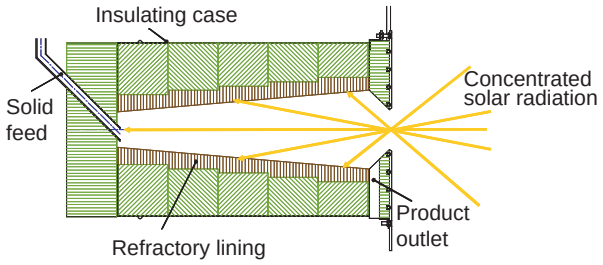


FIGURE 19 Schematic of a direct-irradiated rotary reactor for the solar production of lime. Source: Paul Scherrer Institute, Villigen, Switzerland.

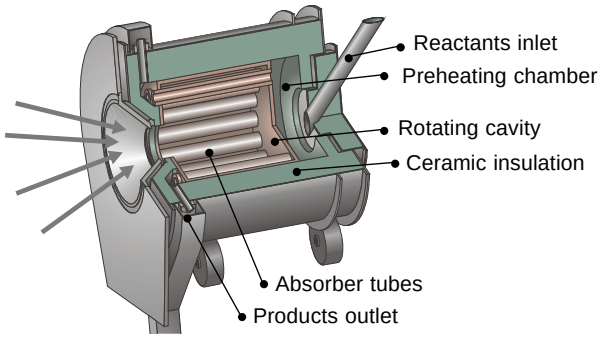


FIGURE 20 Schematic of an indirect-irradiated rotary reactor for the solar production of lime. Source: Paul Scherrer Institute, Villigen, Switzerland.

energy resources and to reduce emissions of CO_2 and other pollutants. Figures 19 and 20 show two solar reactor concepts for the solar calcination process based on direct irradiation and indirect irradiation, respectively. The direct-irradiated solar reactor (Fig. 19) consists of a refractory-lined conical rotary kiln operated in a horizontal position. Because of the rotational movement of the kiln, the reactants are transported within the conical reaction chamber from the preheating zone in the back (feeding side) to the high-temperature zone in the front (discharging side). The indirect-irradiated reactor (Fig. 20) consists of a multiple-tube absorber and a preheating chamber, both made from SiC. The well-insulated rotary reactor is tilted and works in continuous mode of operation.

3.2.4 Solar Thermal Recycling of Hazardous Waste Materials

Solid waste materials from a wide variety of sources (e.g., municipal waste incineration residuals, discharged batteries, dirty scraps, automobile shredded residue, contaminated soil, dusts, and sludge, other by-products from the metallurgical industry) contain hazardous compounds that are usually vitrified in a

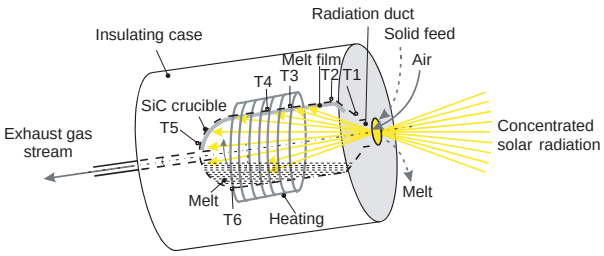


FIGURE 21 Scheme of a rotary kiln for aluminum melting. Source: Deutsches Zentrum für Luft- und Raumfahrt, Germany.

nonleaching slag and are finally disposed of at hazardous waste storage sites. However, limited storage space, increasing storage costs, and environmental regulations have led the need for developing technologies that recycle these toxic materials into useful commodities rather than deposit them in dump sites for an undetermined period of time. Thermal processes are well suited for the treatment of these complex solid waste materials. Waste materials containing carbonaceous compounds can be converted by thermal pyrolysis and gasification into syngas and hydrocarbons that can be further processed into other valuable synthetic chemicals. Those containing metal oxides may be converted by carbothermal reduction into metals, nitrides, carbides, and other metallic compounds. The commercial recycling techniques by blast, induction, arc, and plasma furnaces are major consumers of fossil fuel-based electricity and heat. Concentrated solar energy offers the possibility of converting hazardous solid waste material into valuable commodities for processes in closed and sustainable material cycles. Examples of recycling processes demonstrated in solar furnaces are the processing of electric arc furnace dust using the "two-cavity" reactor depicted in Fig. 14 and the melting of Al scrap using a SiC-lined rotary kiln shown in Fig. 21, which can be tilted for discharging molten metal and slag into molds.

Other recent applications in the area of solar thermal processing of materials that are being explored include hardening of carbon steel, production of expanded natural graphite (ENG) by exfoliation of graphite intercalated compounds, synthesis of advanced ceramics, and purification of metallurgical-grade silicon for photovoltaic applications.

3.2.5 Economics

Economic assessments indicate that solar thermochemical processes for the production of hydrogen and other solar fuels have the potential of becoming economically competitive vis-à-vis alternative paths

for producing solar fuels such as via electrolysis using solar-generated electricity. These studies further indicate that the heliostat field is responsible for approximately half of the total investment costs for the entire chemical plant, whereas the cost of the solar reactor represents 10 to 15%. However, for a fixed product throughput, the solar reactor's efficiency dictates the size of the heliostat field. Thus, reaching high solar reactor efficiencies and reducing the cost of the heliostats per unit area will have a significant impact on reducing the unit cost of the solar fuel. Application of credit for CO_2 mitigation and pollution avoidance will further enable the solar thermochemical technologies to compete favorably with fossil fuel-based processes. The weaknesses of these economic evaluations are related primarily to the uncertainties in the viable efficiencies and investment costs of the various components due to their early stage of development and their economies of scale.

4. OUTLOOK

Solar thermochemical processes have favorable long-term prospects because they avoid or reduce costs for CO_2 mitigation and pollution abatement. The products are renewable fuels and material commodities for delivering clean and sustainable energy services. Further development and large-scale demonstration are warranted.

SEE ALSO THE FOLLOWING ARTICLES

Alternative Transportation Fuels: Contemporary Case Studies • *Hydrogen Production* • *Materials for Solar Energy* • *Photosynthesis, Artificial* • *Solar Cells* • *Solar Cooling, Dehumidification, and Air-Conditioning* • *Solar Detoxification and Disinfection* • *Solar Distillation and Drying* • *Solar Heat Pumps* • *Solar Ponds* • *Solar Thermal Energy, Industrial Heat Applications* • *Solar Thermal Power Generation* • *Solar Water Desalination*

Further Reading

Fletcher, E. A. (2001). Solarthermal processing: A review. *J. Solar Energy Eng.* **123,** 63–74.

Steinfeld, A., and Palumbo, R. (2001). Solar thermochemical process technology. *In* "Encyclopedia of Physical Science and Technology." (R. A. Meyers, Ed.), vol. 15, pp. 237–256. Academic Press, San Diego.

Solar Heat Pumps

KAMIL KAYGUSUZ
Karadeniz Technical University
Trabzon, Turkey

1. Developments
2. Types and Applications
3. Performance and Economics
4. Environmental Benefits
5. Conclusion

Glossary

air conditioning The treatment of air temperature, humidity, cleanliness, and circulation so as to achieve a controlled desired result.
coefficient of performance (COP) Ratio of heating capacity to electrical input.
compressor In a vapor compression cycle, the device that increases the pressure and temperature of refrigerant vapor; it continuously draws low-pressure refrigerant vapor from the cooler, adds energy to increase the refrigerant pressure and temperature, and discharges the high-pressure vapor to the condenser.
condenser A heat exchange coil within a mechanical refrigeration system used to reject heat from the system; it is the coil where condensation takes place.
evaporator A heat exchange coil within a mechanical refrigeration system used to absorb heat into the system; it is the coil where evaporation takes place.
heat A form of energy; it is the least organized form of energy and the form of energy into which other forms deteriorate.
heat pump A comfort system in which the refrigeration cycle is reversed by using a four-way valve to supply both heating and cooling.
thermostat A device that connects or disconnects a circuit in response to a change in the ambient temperature.
vapor compression cycle A refrigeration cycle consisting of a cooler, a compressor, a condenser, and an expansion device.

The general term heat pumping technologies is used for processes in which the natural heat flow from a higher temperature level to a lower one is reversed by adding high-value energy, that is, exergy. The term heat pump is used in different ways in various regions of the world. In Japan and the United States, reversible units are called heat pumps; in addition to cooling and dehumidification, they offer the capability of providing useful heat. Most commonly, air-to-air units are used; however, ground-coupled systems have an increasing share. Chillers are more or less always called chillers, even if they are also used for heating purposes, as are heat pump chillers. In Europe, the term heat pump is used for heating-only units, with the heat sources outside air, ground, and groundwater, combined with heat distribution systems.

1. DEVELOPMENTS

Solar energy systems and heat pumps are two promising means of reducing the consumption of fossil energy sources and, potentially, the cost of delivered energy for domestic space heating and cooling with water heating. A logical extension of each is to try to combine the two to further reduce the cost of delivered energy. It is widely believed that solar heat pump combined systems will save energy, but what is not often known is the magnitude of the possible energy savings and the value of such savings relative to the additional expense. So, renewed interest is being directed at solar heat pumps by scientists, architects, engineers, manufacturers, and building owners as well as in the marketplace in general. The principal advantage of employing solar radiation as a heat pump heat source is that, when available, it provides heat at a higher temperature level than do other sources, resulting in an increase in the coefficient of performance (COP). Compared with a solar heating system without a heat pump, the collector efficiency and capacity are materially increased due to the lower collector temperature required.

The combination of a heat pump and solar energy system would appear to alleviate many of the disadvantages that each has when operating separately.

 © 2004 Elsevier Inc. All rights reserved.

During winter, the energy that could be collected by the solar system, but that would be too low in temperature to be useful for direct heating, may be used as a source for the heat pump. Because the solar collector storage system can supply energy at temperatures higher than the ambient outdoor air, the capacity and COP of the heat pump would increase over those for the heat pump alone, the peak auxiliary load requirement would be reduced, and the combined heating system would seem to operate more economically. The operation of the solar system at temperatures near or below room temperature would decrease the collector losses and allow more energy to be collected. The lower collection temperature might allow the use of collectors with one or no covers, and this would reduce the first cost from a conventional two-cover solar system. Finally, for those areas where warm temperatures occur during cloudy periods, the combined system might compensate for the reduced performance of the conventional solar system under cloudy conditions and the low capacity of the heat pump in cold weather.

2. TYPES AND APPLICATIONS

Research and development in the solar-assisted heat pump (SAHP) field has been concerned with two basic types of systems: direct and indirect. In direct systems, refrigerant evaporator tubes are embodied in a solar collector, usually of the flat plate type. Research has shown that when the collector has no glass cover plates, the same collector surface can also function to extract heat from the outdoor air. The same surface may then be employed as a condenser, using outdoor air as a heat sink for cooling.

An indirect system employs another fluid, either water or air, which is circulated through the solar collector. When air is used, an air-to-air heat pump may be employed, with the collector being added in such a way that (1) the collector can serve as an outdoor air preheater, (2) the outdoor air loop can be closed so that all source heat is derived from the sun, or (3) the collector may be disconnected and the outdoor air used as the source or sink. When water is circulated through the collector, the heat pump circuit may be of either the water-to-air or water-to-water type. Heat pumps complement solar collectors and heat storage in SAHP systems in several ways:

- Heat pumps are more efficient and can provide more heat for a given heat pump size if their evaporators can operate from a warm source. Thermal energy storage, heated by solar collectors, can provide that warming.
- Solar collectors operate more efficiently if they collect heat at lower temperatures. If the collected heat can be stored at a lower temperature because it is used to warm the evaporator of a heat pump, the collector is more efficient and, therefore, fewer collectors are needed to collect a required amount of heat.
- Heat pumps are the most efficient way in which to use electricity for backup heat for a solar heating system, even if there is no direct thermal connection of the heat pump with the solar collector and storage system.
- Heat pumps can allow sensible and latent heat storage units (e.g., water, phase change materials [PCMs]) to operate over wide temperature ranges because stored heat down to 5°C can be used in conjunction with heat pumps to heat a building.

Figure 1 shows the SAHP system linked to an energy storage tank for residential heating. In this system, solar collectors were constructed by modifying flat plate water-cooled collectors. Each solar collector had eight 1.25 cm outside diameter copper tubes spaced 10 cm apart and an effective absorber area of 1.66 m^2. The aluminum sheet was 0.55 mm thick. The 18 collectors were connected in parallel series form. The heat pump compressor was a hermetic type. It was driven by a 1490-W electric motor. The heat pump can use air and water as a heat source. The storage tank contained 1500 kg calcium chloride hexahydrate as a PCM for energy storage.

In general, SAHPs can work in three forms: parallel, series, and dual-source systems. A parallel system would use solar energy in the heat exchanger first and then use the air-to-air heat pump when the solar coil is unable to provide the necessary heat. A series system would use the solar energy in the heat exchanger first and then use the water-to-air heat pump when the solar coil is unable to provide the necessary heat. Dual-source heat pumps are currently in the development stage. A dual-source system would have the choice of either an air-heated coil when warm air is available or a water-heated coil when warm water is available. This type of capability would be advantageous in the SAHP, and in applications where well water is available as a source during the heating season. It would also alleviate the need for a cooling tower when the heat pump operates in the cooling mode.

It is not unusual for the SAHP to save up to 20% in operating costs during the heating season compared with an air-to-air heat pump. This represents a

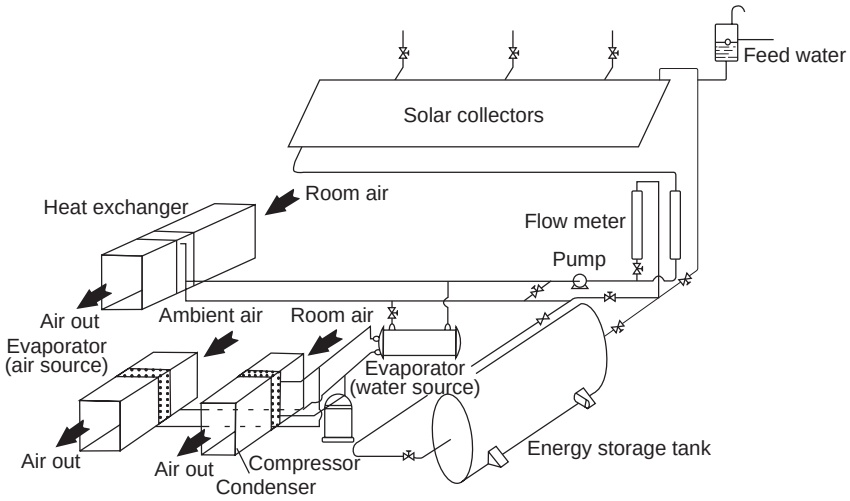


FIGURE 1 Schematic overview of experimental solar-assisted heat pump system. Reprinted from Kaygusuz *et al.* (1993).

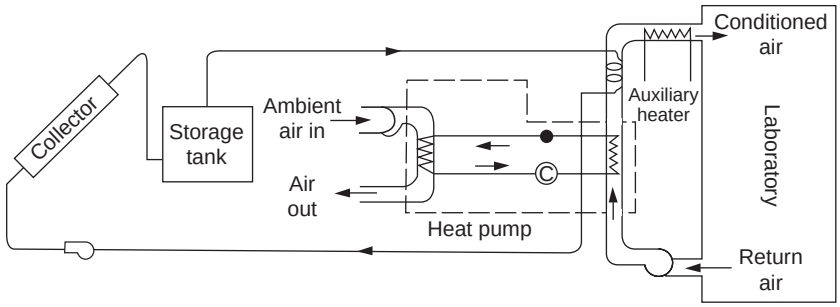


FIGURE 2 Parallel heat pump system. Reprinted from Kaygusuz *et al.* (1993).

heating COP of nearly 3.0. It has been reported that parallel systems can save up to 50% of the energy required when compared with electrical resistance heating, series systems can save up to 60% of the energy normally required, and dual-source systems can save up to 70% of the energy normally required.

A wide variety of options exist to provide air conditioning along with solar heating of buildings. The efficacy of coupling these options with a solar heating system depends strongly on which system is to be used and whether one is considering systems that can already be built reliably or ones that promise optimum performance in the future following extensive development effort. Several SAHP systems are described to illustrate their relationship with air conditioning: the simple parallel system, series system, dual-source system, and cascade system.

2.1 Parallel System

As shown in Fig. 2, the parallel system consists of a solar collector, a water-to-air heat exchanger, an air-sourced heat pump, a water-circulating pump, a storage tank, and other equipment. The parallel SAHP system is combining two main components: the solar system and the parallel heat pump system. In this system, the heat pump uses ambient air as an energy source while the water-to-air heat exchanger uses solar energy as a heat source, and they give their energies to the load one by one. Solar energy is used to meet as much of the heating requirement as possible. Thus, the total available energy of the system is the sum of the extracted energies from two different systems (the solar system and the heat pump).

2.2 Series System

Figure 3 shows a series SAHP system. The solar collector heats water, which is then stored in a tank. The tank provides heat directly to the house if the tank temperature is above 40°C, and the heat pump draws heat from the tank when the tank temperature is between 5 and 40°C. This system has the advantage that the tank can be operated at lower

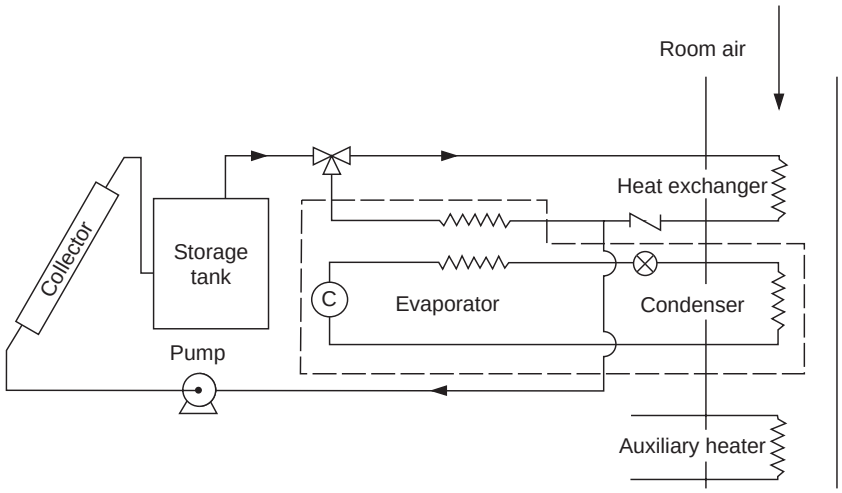


FIGURE 3 Series heat pump system. Reprinted from Kaygusuz *et al.* (1993).

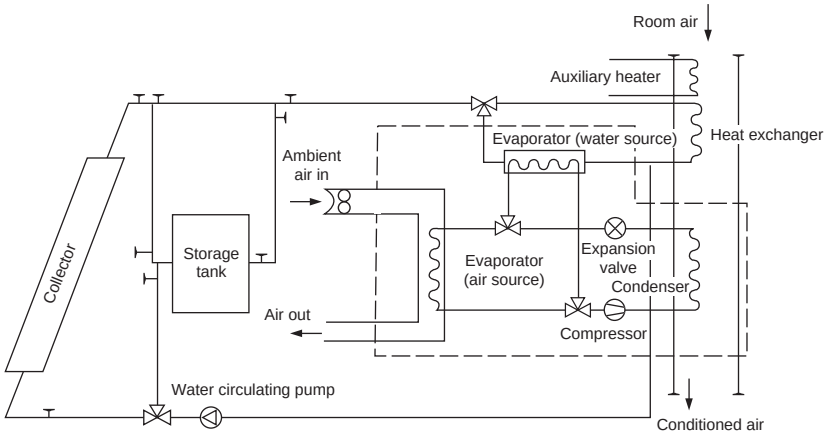


FIGURE 4 Dual-source system. Reprinted from Kaygusuz *et al.* (1993).

temperatures when required, allowing the collector to operate with high efficiency, and the tank heat storage capacity can be increased by the amount of sensible heat stored between 5 and 40°C. The disadvantage of the system is that when the tank temperature finally drops to 5°C, the heat pump cannot be used further without danger of freeze-up in the water tank. The system shown will not provide air conditioning because there is no way in which to exhaust the waste heat to outside air. It would be possible to provide air conditioning if there were another heat exchanger loop between the storage tank and outside air, but this would add to the cost of the heat exchanger and its associated circulating pump. A development effort is required on the heat pump in this system because residential heat pumps are not currently designed to operate efficiently with evaporator temperatures above 20°C, and this one would have to operate at up to 40°C.

2.3 Dual-Source System

The dual-source SAHP system, illustrated in Fig. 4, combines the advantages of the parallel and series systems and overcomes their disadvantages. The solar collector, heat storage, and building heat exchanger all continue to work as a simple solar system so long as the tank temperature remains above 40°C. At temperatures below that, the heat pump system is called on by a microprocessor controller, and a decision is made as to whether it is better to draw the heat pump's heat from the tank or from the outside air. The control strategy options are numerous for this system; it can be operated to optimize savings of electricity or to reduce peak loads. It would be the most efficient of all the systems described here. It has the disadvantage that the required heat pump is complex and will be costly to manufacture. Considerable development work will

be required to ensure that the unit can be operated in all sequences of its modes with full reliability. With this system, air conditioning is accomplished by operating the unit as a simple air-to-air heat pump in the cooling mode, using only the outside heat exchanger as a condenser. So, it is obvious that the dual-source heat pump system takes advantage of the best features of the series and parallel heat pump systems. The system is not capable of using the storage tank to reduce air-conditioning peak loads during the summer because the heat pump cannot be used to cool the tank to the outside air.

2.4 Cascade System

As Fig. 5 shows, the cascade SAHP system allows complete access to the storage tank in both the heating and air-conditioning modes. Two heat pumps are used: one between the outside air and the tank and the other between the tank and the house. The inside heat pump allows the system to operate in the series heat pump mode, providing heat to the house when the tank temperature drops below 40°C. When supplemental heat is required, the outside heat pump charges the tank with heat removed from the outside air. In air conditioning, the inside heat pump cools the house on demand, depositing the heat in storage; the outside heat pump cools the house on demand, depositing the heat in storage; and the outside heat pump then removes heat from the tank to outside air during off-peak periods. Both of the heat pumps used in the cascade system require some development effort. The outside heat pump must be optimized for operation between outside air temperatures and the cool water temperatures of the storage tank. The inside heat pump is the same as the series unit described previously except that it must include the ability to air condition as well as heat. The required development effort would probably be less than that for a three-coil unit. This system is probably the most expensive of those described here due to the cost of heat pump equipment. When operating with the outside heat pump as a supplementary heat source, it will be less efficient than the three-coil or parallel system because the heat must go through a second heat pump in heating the house. On the other hand, it is the most flexible system, allowing full heating and air conditioning as well as the use of the storage tank to store heat for off-peak use (solar or electrical utility) during the heating season and to reduce air-conditioning peak loads during the cooling season. Today, a number of solar-assisted heat pumps are commercially available either as complete systems or as component packages, as shown in Figs 6, 7, and 8.

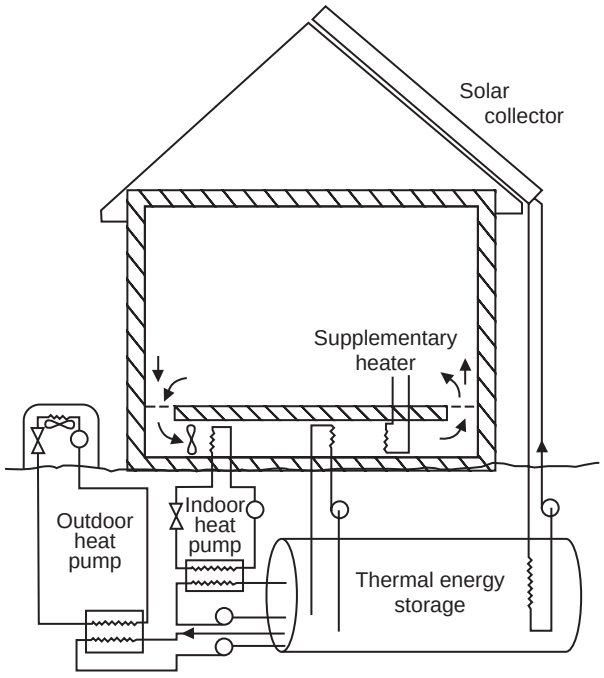


FIGURE 5 Cascade system. Courtesy of General Electric Company (reprinted from Sauer and Howell, 1991).

It is possible to view much of the work of the past few years on SAHP systems as an attempt to address a single fundamental question: what does one do when the sun is not shining and storage is depleted? Studies based on the assumption that electrical resistance is used as the backup heat source have generally concluded that the SAHP is not a good idea. However, it is now possible to identify two general paths that have been taken to avoid this difficulty: direct expansion SAHP systems and ground-coupled SAHP systems. Each type of system has its own special characteristics. These are described in what follows.

2.4.1 Direct Expansion SAHP

Refrigerant-filled, two-phase solar collectors have been used during recent years as alternatives to conventional air or water collectors. The working fluid in these collectors undergoes a liquid-to-vapor transformation, and the absorbed solar energy is manifested as the latent heat of vaporization. Two-phase collectors are usually employed in direct expansion solar-assisted heat pump (DX-SAHP) cycles. The primary objective in these cycles is to raise the evaporator temperature through solar energy input, which in turn translates into improvement in the thermal performance of the heat pump.

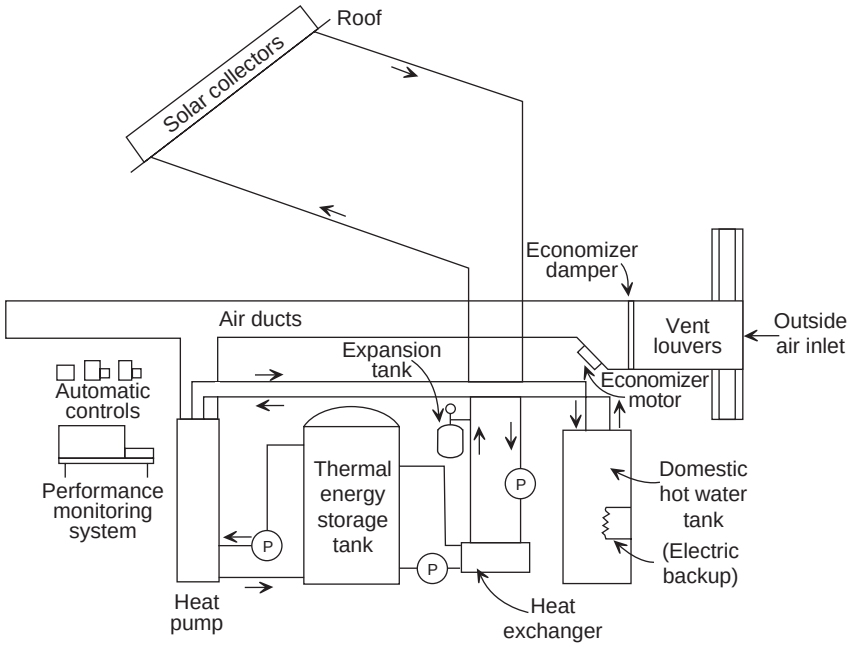


FIGURE 6 Residential dual-source heat pump system. Courtesy of Daystar Corporation–Exxon Enterprises (reprinted from Sauer and Howell, 1991).

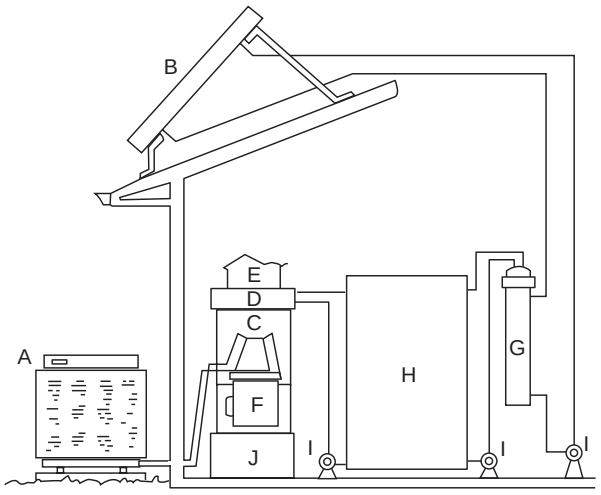


FIGURE 7 A commercially available solar-assisted heat pump based on the General Electric WeatherTron. A, WeatherTron outdoor section; B, solar heat collection panel; C, indoor heat exchanger; D, hydronic coil; E, supply duct; F, flower; G, heat exchanger; H, heat storage tank; I, pumps; J, return duct. Courtesy of General Electric Company (reprinted from Sauer and Howell, 1991).

The DX-SAHP system offers several advantages over the conventional SAHP systems. Because the collector also acts as the evaporator, the number of components in the system is reduced. For example, instead of using a collector, an evaporator, and an intermediate heat exchanger in a conventional SAHP system, only a collector panel is required in the DX-SAHP system. Circulation of refrigerant in the collector tubes also prolongs the system life because corrosion and nighttime freeze-up problems encountered in water collectors are virtually eliminated.

2.4.2 Ground-Coupled SAHP

Ground source heat pumps (GSHPs) were originally developed in the residential arena and are now being applied in the commercial sector. Many of the installation recommendations, design guides, and rules of thumb appropriate to residential design must be amended for large buildings. The term "ground source heat pump" is applied to a variety of systems that use the ground, groundwater, and surface water as a heat source and sink. Included under the general term are ground-coupled (GCHP), groundwater (GWHP), and surface water (SWHP) heat pump systems.

GCHPs are a subset of GSHPs and are often referred to as closed-loop GSHPs. A GCHP is a system that consists of a reversible vapor compression cycle linked to a closed ground heat exchanger buried in soil. The most widely used unit is the water-to-air heat pump, which circulates water or a water–antifreeze solution through a liquid-to-refrigerant heat exchanger and a buried thermoplastic piping network. A second type of GCHP is the direct expansion ground-coupled heat pump (DX-GCHP),

which uses a buried copper piping network through which refrigerant is circulated. To distinguish them from DX-GCHPs, systems using water-to-air and water-to-water heat pumps are often referred to as GCHPs with secondary solution loops.

GCHPs are further subdivided according to ground heat exchanger design: vertical and horizontal. Vertical GCHPs generally consist of two small-diameter, high-density polyethylene (PE) tubes in a vertical borehole. The tubes are thermally fused at the bottom of the bore to a close return U-bend. The nominal diameter of vertical tubes ranges from 20 to 40 mm. Bore depths range from 15 to 185 m depending on local drilling conditions and available equipment.

The advantages of vertical GCHPs are that they (1) require relatively small plots of ground, (2) are in contact with soil that varies very little in temperature and thermal properties, (3) require the smallest amount of pipe and pumping energy, and (4) can yield the most efficient GCHP system performance. The disadvantage is that they are typically higher in cost due to the limited availability of appropriate equipment and installation personnel.

Horizontal GCHPs can be divided into three subgroups: single pipe, multiple pipe, and spiral. Single-pipe horizontal GCHPs were initially placed in narrow trenches at least four deep. These designs require the greatest amount of ground area. Multiple pipes (usually two or four) placed in a single trench can reduce the required amount of ground area. With multiple-pipe GCHPs, if trench length is reduced, total pipe length must be increased to overcome thermal interference with adjacent pipes. The spiral coil is reported to further reduce the required ground area. These horizontal ground heat exchangers are made by stretching small-diameter PE tubing from the tight coil in which it is shipped into an extended coil that can be placed vertically in a narrow trench or laid flat at the bottom of a wide trench. Required trench lengths are only 20 to 30% those for single-pipe horizontal GCHPs, but pipe lengths may be double for equivalent thermal performance.

The advantages of horizontal GCHPs are that (1) they are typically less expensive than vertical GCHPs because appropriate installation equipment is widely available, (2) many residential applications have adequate ground area, and (3) trained equipment operators are more widely available. Disadvantages include (1) the larger ground area requirement; (2) greater adverse variations in performance because ground temperatures and thermal properties fluctuate with season, rainfall, and burial depth; (3) slightly higher pumping energy requirements; and (4) lower system efficiencies.

3. PERFORMANCE AND ECONOMICS

The important quantities for the evaluation of the systems are the seasonal energy flows. For space and service hot water heating, these are the sum of the house space and water heating loads (q_{load}), the heat collected by solar energy system (q_{solar}), the heat

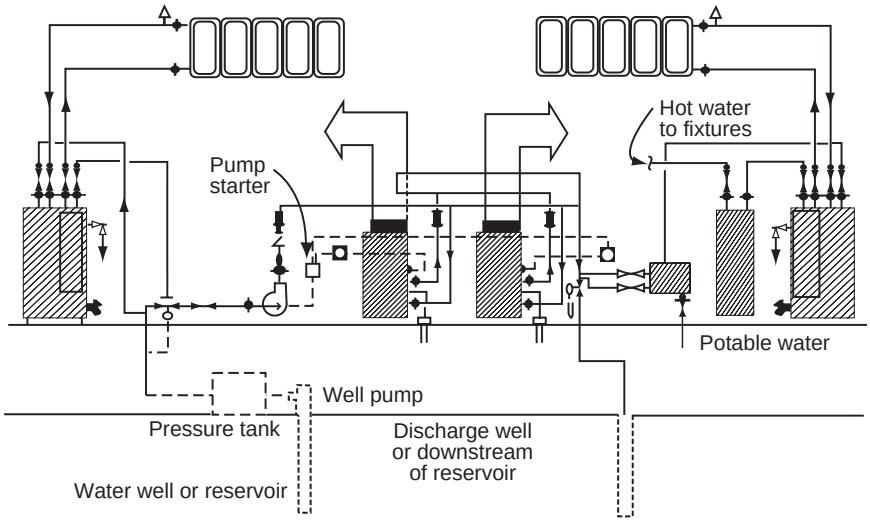


FIGURE 8 Northrup solar-assisted heat pump system and components. Courtesy of Northrup Inc. (reprinted from Sauer and Howell, 1991).

extracted by the heat pump from the ambient air (q_{air}), the electricity required by the heat pump compressor and fans (W_{hp}), and the auxiliary energy required to meet the hot water and space heating loads (E_{aux}). Here, E_{aux} is the delivered energy and is less than the purchased fuel energy due to furnace and water heater inefficiencies. Over the course of a year, there is a negligible change in the amount of energy in the storage tank, and the system energy balance is

$$q_{solar} + q_{air} + W_{hp} + E_{aux} = q_{load}.$$

The first two terms are free energy, whereas the second two terms represent purchased energy. A measure of the thermal performance of combined solar heat pump systems is the fraction (F) of the total load that is met by nonpurchased energy, defined as

$$F = (q_{air} + q_{solar})/q_{load} = 1 - (W_{hp} + E_{aux})/q_{load}.$$

The auxiliary energy may be supplied by gas, oil, electricity, or a combination, whereas the heat pump work is nearly always electricity. The thermal performance fraction does not distinguish among the different kinds and prices of fuels; rather, it reflects only energy requirements.

The fraction, F, is shown in Fig. 9 for Madison as a function of collector area for the conventional furnace, conventional heat pump, parallel, series, dual-source, and conventional single-cover solar system. For a house with neither a solar energy system nor a heat pump, F equals zero. For a house with only a heat pump, the fraction of the heating requirement supplied by nonpurchased energy is q_{air} divided by the total heating requirement and equals 36%. Because the heat pump does not contribute to the domestic hot water system, the value of F depends on the COP and the relative sizes of the space and water heating loads and is usually between 0.2 and 0.4. For a house with a conventional solar system, F depends on collector area. As collector area increases, F increases from zero asymptotically toward 1.0.

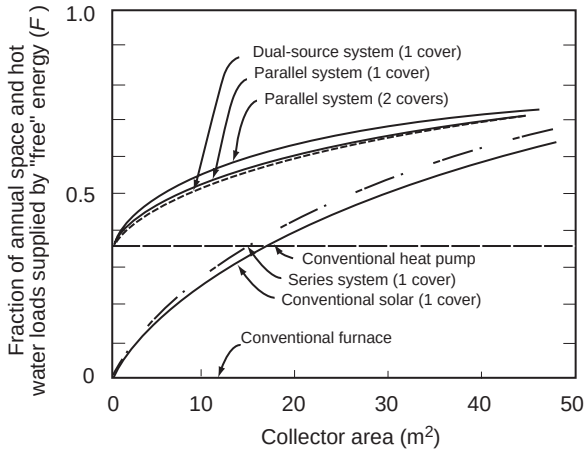


FIGURE 9 The fraction of energy from nonpurchased source (F) for series, parallel, dual-source, heat pump-only, and standard solar energy systems as a function of collector area, based on simulations of the systems on a residential building in the Madison climate. Reprinted from Freeman *et al.* (1979).

The results of the simulations done using the program TRNSYS show that with the same collector area, the parallel system is substantially better than the series system and slightly better than the dual-source system in all collector sizes in that it delivers a greater fraction of the loads from nonpurchased sources. This arises because the heat pumps in the series and dual-source systems must operate to deliver all solar energy stored below 20°C. The extra electrical energy required to deliver this energy more than compensates for the combined advantages of higher collector efficiency and higher heat pump COP. The temperature of storage in the series system (and in the dual-source system) in these simulations is rarely high enough to permit direct solar heating.

Improved collector performance and higher q_{solar} not only are the direct consequences of adding the solar source capability but also are the only benefits to system performance during the heating season. However, the improved collector performance is achieved at the expense of heat pump electrical power input that usually displaces a cheaper fuel. The annual energy balance requires that the sum of all energy supplied equals the load given as

$$q_{load} = q_{solar} + q_{air} + W_{hp} + E_{aux}.$$

The relative contributions from each of these four heat sources are shown in the bar graphs in Fig. 10. The combined height of the q_{solar} and q_{air} bars in the figure represents the percentage of the total heating requirement supplied by free energy and, therefore, is equivalent to F. The systems are ranked from left to right in order of decreasing savings of purchase energy that is decreasing F.

On the other hand, the experimental heat pump seasonal COP (i.e., ratio of rejected heat to work input) varies among the systems. The use of a solar source for the heat pump raises the seasonal COP over that of the conventional heat pump and parallel systems. As shown in Fig. 10 for Trabzon, Turkey, the annual heat pump heating COPs for the parallel, series, and dual-source systems are 3.0, 4.0, and 3.5, respectively. The relative COP values for the

conventional heat pump and the parallel system heat pump result from the fact that the heat pump is forced to operate less often but under generally more unfavorable conditions when used as an auxiliary to a solar system. As expected, the COP for the series and dual-source heat pumps are substantially higher because they use the solar-heated storage tank as a source. Table I shows the experimental performance of the SAHP system for residential heating in Trabzon.

Two configurations of series SAHPs were evaluated experimentally in TECH House I on the University of Tennessee campus in Knoxville. During the 1979–1980 heating season, a 4.2-m^3 insulated above-ground storage tank was used. During the 1980–1981 heating season, the storage system consisted of an uninsulated 7.6-m^3 steel tank buried 2.1 m in the ground. For 1979–1980, the heating season system performance factor was 2.42 and the system's maximum peak electric demand was 11.20 kW. For 1980–1981, the heating season performance factor increased to 3.16, whereas the maximum peak electric demand decreased to 4.33 kW. On the other hand, a parallel SAHP system was evaluated experimentally in TECH House IV for three heating seasons between the fall of 1979 and the spring of 1982. The seasonal performance factors (SPFs) for the 1980–1981 and 1981–1982 seasons were measured to be 2.56 and 2.11, respectively.

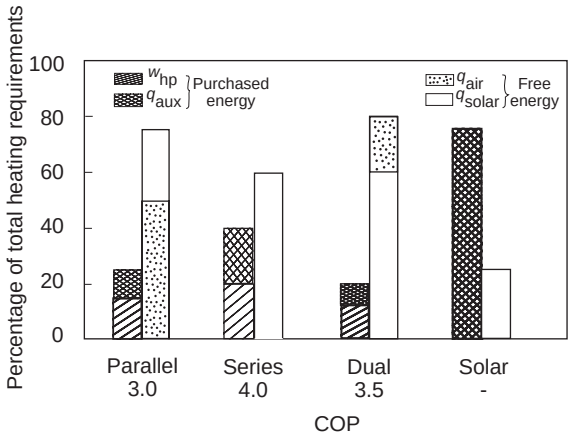


FIGURE 10 Heating contributions from all possible sources for Trabzon, Turkey. Reprinted from Kaygusuz (1995).

In the competition for commercial viability, all solar heating systems must compete for some version of life cycle economic advantage with alternative nonsolar heating systems. In going to solar heating, an economic trade-off is made between the high first cost of solar heating or cooling equipment and the cost savings realized by reduced use of fuel or electricity. One way in which to compare nonsolar and solar heating systems is to determine the cost of electricity at which they have the same overall life cycle cost. A rough comparison of this break-even cost of electricity shows that many solar-assisted systems will be more cost effective than the conventional air-to-air heat pump available in the United States today only when electricity costs become greater than 10 to 12 cents/kWh. The current national average cost of residential electricity is approximately 4 cents/kWh, with a range of up to approximately 10 cents/kWh in some major urban utility systems.

The feasibility of SAHP systems for space heating and domestic hot water preheating in Canada is

TABLE I

The Experimental Performance of the Solar-Assisted Heat Pump System for Trabzon, Turkey

	Parallel system					Series system				
Heating month	*N*	COP	D_{col} (percentage)	D_{st} (percentage)	*F* (percentage)	*N*	COP	D_{col} (percentage)	D_{st} (percentage)	*F* (percentage)
November	30	2.90	50	56	80	21	4.40	60	60	60
December	31	2.92	52	58	76	13	4.30	60	62	45
January	26	2.80	53	62	48	15	4.27	61	63	48
February	23	2.70	48	60	38	20	4.20	62	62	50
March	30	2.85	52	63	75	22	4.32	64	64	64
April	30	3.00	55	64	85	23	4.40	68	66	70
May	10	3.10	60	66	90	7	4.50	66	65	80

Source. Kaygusuz (1999a, 1999b).

Note. *N*, number of working days of the heat pump system per month; *F*, fraction of the load met by free energy; COP, coefficient of performance; n_{col}, collector efficiency; n_{st}, storage efficiency.

examined by simulating the performance of these systems on a computer using the program WATSUN. Simulations are carried out using meteorological data for seven representative Canadian cities, two different building types, and six types of system configurations. For the SAHP system, 20-year life cycle cost comparisons are made with two reference systems, a conventional resistance heating system and an air-to-air heat pump system, based on current economic parameters and projected escalation scenarios for electricity rates.

Results of the study show that the SAHP systems conserve significant amounts of energy over resistance heating and heat pump systems. On the life cycle unit cost basis, SAHP system costs are relatively insensitive to location, but the dependence on building types is substantial, with multiplex dwellings showing the least cost. Liquid-based, dual-source SAHP systems are found to be cost-effective over resistance heating at some of the locations for multiplex units. They are not cost-effective for single-family dwellings at the current time.

4. ENVIRONMENTAL BENEFITS

Energy efficiency and environmental benefit are the key drivers for heat pump application and further research and development. A simple projection of world energy consumption and dwindling energy sources during the coming decades shows that society faces a major challenge. There is a general awareness of this rapidly approaching problem, but society as a whole does not seem to be responding accordingly, assuming that there is sufficient time to find solutions. However, it is encouraging to see that there is a growing interest in applying energy-efficient technology in developing countries and economies in transition. Although costs are an obstacle for applying modern technology, the industrialized world seems increasingly prepared to look for new ways of information and technology transfer to these developing parts of the world.

How important are heat pumps in relation to the environment? Any thermodynamic specialist or heating/cooling engineer will have no problem in quickly answering this question. The numbers on COPs and SPFs, as well as on specific carbon dioxide (CO_2) emission levels, are available and can lead to only one conclusion: an increasing application of heat pumps without delay is very important. On the other hand, it is frequently argued that the refrigerants used in heat pumps today are safe and that their impact on the environment during the equipment life cycle is negligible. That may be true and significant as well, but the fact remains that a much better job can be done. There is significant scope for improved performance and initial cost reduction. Both will lead to a better environment and decreased energy consumption.

In Fig. 11, various systems for the supply with electricity and heat are compared with regard to their CO_2 emissions. The various lines in the figure represent standard systems that are already installed or at least are obtainable in Turkey. The figure shows the CO_2 emissions in kilograms for the supply with 1 kWh power plus heat (in kilowatt-hours). On the other hand, in the case of a conventional coal-fired power plant with an electrical efficiency of 0.40 and a distribution efficiency of 0.95, 0.9 kg CO_2 is emitted for each kilowatt-hour of power delivered to the user. All systems in the figure use this power plant, so all curves intersect at the same point on the ordinate. The steepest line represents a system combining this coal power plant with an oil furnace for heating. For example, for the supply with 1 kWh power from the power plant and 2 kWh heat from the furnace, we get CO_2 emissions of 1.7 kg. When we change the fuel for the furnace from oil to natural gas, we get a 12% reduction in CO_2, from 1.7 to 1.5 kg, for this example. If we further improve the boiler efficiency up to 1.0, which is the value for modern condensing furnaces, we obtain a reduction of the CO_2 emissions of 24%.

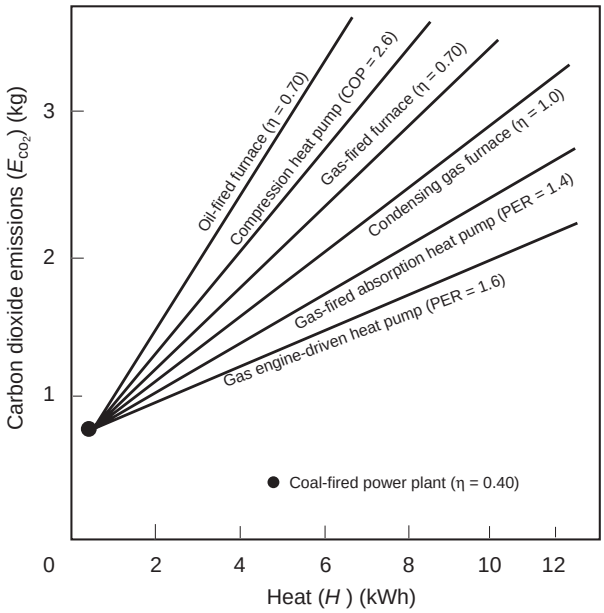


FIGURE 11 Carbon dioxide (CO_2) emissions for systems with a coal-fired power plant. Reprinted from Kaygusuz (2003).

Compression heat pumps with a COP of 2.5, operated with electric power from coal power stations, show only slightly less emissions than do oil furnaces and are even worse than standard gas furnaces and especially condensing gas furnaces. On the other hand, for a larger heating demand, only gas engine-driven compression heat pumps, with a primary energy rate (PER) of 1.5, or direct gas-fired absorption heat pumps (PER = 1.3) show a similarly high reduction in CO_2 emissions. The efficiencies assumed in Fig. 11 are not the best values for available compression heat pumps. Modern systems reach a COP of 3.0 and sometimes 4.0. Modern power stations with a combined gas-and-steam turbine cycle (CC) reach electrical efficiencies of up to 0.53.

In Fig. 12, the CO_2 emissions are plotted for these systems. The striped areas indicate the range of COPs of the advanced heat pumps. As can be clearly seen, heat pumps driven by electricity from conventional coal power stations cannot reduce the CO_2 emissions in comparison with condensing gas furnaces in spite of the COP improvement from 2.5 to 4.0. The situation changes appreciably with modern CC power plants with an efficiency of 0.53. In such a system, even heat pumps with a rather poor COP of 2.5 have lower emission ratios than do the best furnaces. Of course, much better results are achieved with COPs of up to 4.0, which are also shown in the figure.

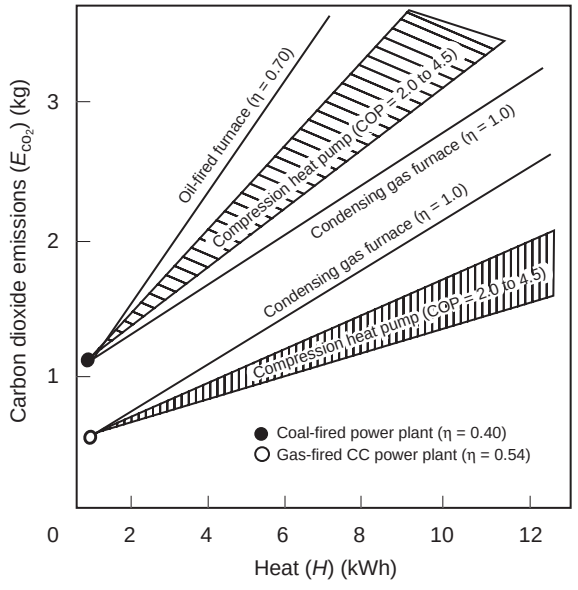


FIGURE 12 Carbon dioxide (CO_2) emissions for systems with advanced compression heat pumps. Reprinted from Kaygusuz (2003).

Worldwide, more than 100 million heat pumps are now installed. Their total annual thermal output is approximately 1300 TWh. This represents an annual equivalent CO_2 emissions reduction of 0.13 billion tons (0.6%), against the backdrop of annual global emissions of CO_2 of more than 22 billion tons. However, the current reduction potential of heat pumps is 6% of global CO_2 emissions, and improved technology could increase this to 16% in the future.

5. CONCLUSION

Heat pumps use mechanical energy to transfer thermal energy from a source at a lower temperature to a sink at a higher temperature. Electrically driven heat pump heating systems have attracted wide interest as an alternative to electric resistance heating or expensive fuels. They have two advantages: a COP greater than unity for heating (thereby saving on purchase of energy) and usefulness for air conditioning during the summer. Heat pumps may use air or water sources for energy, and dual-source machines that can use either are under development.

SAHPs for residential applications have received a great deal of attention since 1975. They combine a solar heating system with a heat pump to provide the necessary heating and cooling requirements. Both air-to-air and water-to-air heat pumps can be used in this system. The solar system basically collects the energy for the heating requirements, and the heat pump supplies this energy at the proper time and temperature and also helps to improve the efficiency of the solar collector in the water-to-air operating mode.

On the other hand, heat pump systems have other advantages for commercial buildings besides their low operating costs. They offer improved comfort due to on-demand, year-round heating or cooling. Greater flexibility in the interior layout of the building is also possible, particularly where a distributed water loop configuration is adopted. In addition, more exterior architectural freedom is possible with the adoption of GSHP systems due to the absence of any need for external equipment such as cooling towers. Because of their use of electric power, heat pumps can also provide a convenient way in which to heat a building in a remote location to which fossil fuel is not economically transportable.

In the past, heat pumps have often been a tool of government policy. For example, during the oil crises of the mid-1970s and early 1980s, governments in some cold climate countries encouraged the use of heat pumps through financial incentives to reduce

their dependence on oil and as part of their energy conservation programs. This may happen again as countries sign on to the Kyoto Protocol and as governments implement measures to reduce greenhouse gas emissions.

As a result, heat pumps are capital intensive, as are solar energy systems. Consideration of the economics of SAHP systems indicates that it may be difficult to justify the investment in a heat pump to improve solar energy system performance unless the cost of the heat pump can be justified on the basis of its use for air conditioning or unless the use of the heat pump would reduce collector costs to a small fraction of their current levels. Today, the most popular commercial heat pump configuration in the United States, Japan, and other countries such as Turkey is the single-packaged air source unit. This is followed by the packaged terminal heat pump and the water source heat pump.

SEE ALSO THE FOLLOWING ARTICLES

Ground-Source Heat Pumps • Materials for Solar Energy • Solar Cells • Solar Cookers • Solar Cooling, Dehumidification, and Air-Conditioning • Solar Detoxification and Disinfection • Solar Distillation and Drying • Solar Ponds • Solar Thermal Energy, Industrial Heat Applications • Solar Thermal Power Generation • Solar Water Desalination

Further Reading

American Society of Heating, Refrigerating and Air-Conditioning Engineers. (1995). Geothermal energy. *In* "ASHRAE Handbook: HVAC Applications," chap. 29. ASHRAE, Atlanta, GA.

Badescu, V. (2002). Model of a solar-assisted heat-pump system for space heating integrating a thermal energy storage unit. *Energy Buildings* **34,** 715–726.

Bouma, J. (2002). Heat pumps: Better by nature—Report on the Seventh IEA Heat Pump Conference. *IEA Heat Pump Centre Newsletter* **20**(2), 10–27.

Chandrashekar, M., Le, N. T., Sullivan, H. F., and Hollands, K. G. T. (1982). A comparative study of solar assisted heat pump systems for Canadian locations. *Solar Energy* **28,** 217–226.

Chaturvedi, S. K., Chen, D. T., and Kheireddine, A. (1997). Thermal performance of a variable capacity direct expansion solar-assisted heat pump. *Energy Convers. Mgmt.* **39,** 181–191.

Duffie, J. A., and Beckman, W. A. (1991). "Solar Engineering of Thermal Processes." John Wiley, New York.

Freeman, T. L., Mitchell, J. W., and Audit, T. E. (1979). Performance of combined solar-heat pump systems. *Solar Energy* **22,** 125–135.

Kavanaugh, S. (1998). Ground source heat pumps. *ASHRAE J.* **40**(12), 31–36.

Kaygusuz, K. (1995). Performance of solar-assisted heat-pump systems. *Applied Energy* **51,** 93–109.

Kaygusuz, K. (1999a). Investigation of a combined solar-heat pump system for residential heating: I Experimental results. *Intl. J. Energy Res.* **23,** 1213–1223.

Kaygusuz, K. (1999b). Investigation of a combined solar-heat pump system for residential heating: II. Simulation results. *Intl. J. Energy Res.* **23,** 1225–1237.

Kaygusuz, K. (2003). "Environmental aspects of fossil fuel combustion and reducing greenhouse gas emissions with heat pump systems in Turkey." Unpublished manuscript, Karadeniz Technical University, Trabzon, Turkey.

Kaygusuz, K., Gültekin, N., and Ayhan, T. (1993). Solar-assisted heat pump and energy storage for domestic heating in Turkey. *Energy Convers. Mgmt.* **34,** 335–346.

O'Dell, M. P., Mitchell, J. W., and Beckman, W. A. (1984). Design method and performance of heat pumps with refrigerant-filled solar collectors. *ASME J. Solar Energy Eng.* **106,** 159–164.

Reid, R. L., McGraw, B. A., and Bedinger, A. F. G. (1985). Performance of series solar-assisted heat pump systems in TECH House I. *ASME J. Solar Energy Eng.* **107,** 293–296.

Sauer, H. J., and Howell, R. H. (1991). "Heat Pump Systems." Krieger Publishing, Malabar, FL.

Solar Ponds

N. D. KAUSHIKA
Indian Institute of Technology
New Delhi, India

1. Introduction
2. Solar Pond Taxonomy
3. Salt Gradient Solar Pond
4. Saltless Nonconvective Solar Ponds
5. Applications
6. Economics
7. Environmental Impact

Glossary

aquifer A layer of soil or rock that can hold water.
double-diffusive fluid A fluid embodying the density gradients of two constituents.
nonconvective layer A fluid layer in which natural convection is suppressed.
permeable Allows liquid or gas to pass through.
solar pond Body of still water that collects solar energy and stores it as thermal energy.
thermohaline system Solar thermal system.
transparent insulation layer A layer that is transparent for solar (optical) radiations and is opaque for infrared (heat) radiations.

The solar pond represents a low-cost option among other solar systems. It possesses thermal storage capability spanning seasons and is the only viable system with 100% solar fraction. Solar pond heat can be used for a variety of applications, including generation of electricity, industrial process heating, desalination, and greenhouse heating.

1. INTRODUCTION

Solar radiant energy is characterized by low-energy density and its practical utilization involves large-area solar collectors. A solar pond is a unique large-area solar collector that uses a body of still water to collect solar energy and stores it as thermal energy. As a practical matter, all ponds, lakes, oceans, and other expanses of water in nature collect solar energy and convert it to thermal energy, but their heat retention efficiency is poor because the water near the surface is quickly cooled as the heat is rapidly dissipated to the environment. The warmer and more buoyant water in the lower region rises to the surface; this movement of water is called natural convection. Therefore, a water–pond system will be more effective in the collection and storage of solar energy if the thermal losses to the environment are reduced either by suppressing the natural convection of water in the pond or by covering the pond surface with a partially transparent insulation. The term solar pond is therefore used for a water pool with reduced heat losses. It can collect solar energy, concentrate and store it as heat up to approximately 100°C (higher temperature restricted by the boiling point of water). The solar pond involves simple technology and uses water as working material for three functions: collection of solar radiant energy and its conversion into heat, storage of heat, and transport of thermal energy out of the system.

2. SOLAR POND TAXONOMY

The first recorded use of solar energy for heating a water pool was made approximately 2500 years ago when great Roman baths were heated by the sun. Von Kalecsinsky, who reported that the bottom of Madve lagoon in Transylvania in summer had a temperature of approximately 70°C, discovered the phenomenon of the solar pond in 1902. In this lake, salt concentration was also observed to increase with depth. The downward increase in salt concentration prevents the convection and renders the upper region of the lake a partially transparent insulator trapping solar heat in the bottom region. The

 © 2004 Elsevier Inc. All rights reserved.

conceptual design of the first artificial "sun pool" was given by Tabor and Bloch in 1958. The proposed system was an artificial pond (approximately 1 m deep) with a black bottom and salt gradient with a maximum concentration at the bottom obtained by filling the pond with progressively less dense brine layers. The name salt gradient solar pond or solar pond was coined for such a system. The term solar pond has since been applied to a variety of water pool solar collector/storage systems. These may be broadly classified as shown in Fig. 1.

In nonconvective solar ponds, the basic principle is that heat loss to the environment is reduced by the suppression of natural convection in the upper layer (transparent insulation layer) of water or air medium. The most documented type of nonconvective solar pond is the salt gradient solar pond. It uses such salts as NaCl, $MgCl_2$, and bitterns. It is referred to as an unsaturated salt-stabilized pond to distinguish it from the saturated salt-stabilized pond based on the Soret effect. The latter uses salt such as KNO_3, whose solubility increases dramatically with temperature; in the pond, saturation of the salt concentration is maintained at all depths. The pond is hotter in the bottom region than at the top. Thus, a progressively larger amount of salt is dissolved toward the bottom. Owing to saturation at each level, the vertical diffusion of salt is checked and the density gradient is stable. This provides for the possibility of a maintenance-free solar pond. Alternative methods of suppressing convection have also been investigated and several solar pond systems have been proposed and tested, including the gel- or viscosity-stabilized pond, the membrane pond, and the honeycomb-stabilized solar pond. These are referred to as saltless nonconvective solar ponds.

In convective solar ponds, heat loss to the environment is reduced by covering the pond surface. The cover is transparent and the pond is often of shallow depth; it is thus referred to as a shallow solar pond. In practice, it is a water-filled plastic bag glazed at the top with a blackened bottom insulated by foam. These ponds are used for producing hot water at medium temperature (40–60°C) for industrial applications as well as for boiler preheating. Figure 2 provides a schematic of solar pond concepts.

Earliest theoretical as well as experimental studies of solar ponds predicted solar heat collection efficiencies of approximately 20% and thermal storage capability spanning seasons. Several countries have since built and operated salt gradient solar ponds. The technology of salt gradient ponds is inherently simple. However, for a solar pond energy system of good thermodynamic efficiency, a variety of factors must be considered and quantified. Therefore, the physics and engineering aspects of salt gradient solar ponds and the conceptual outline of other ponds are presented in the following sections.

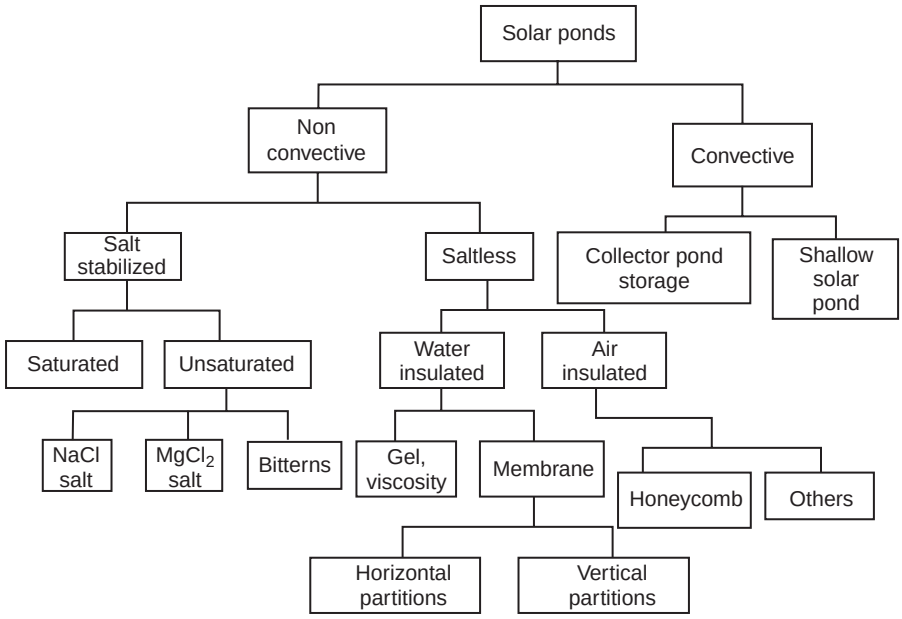


FIGURE 1 Classification of solar ponds.

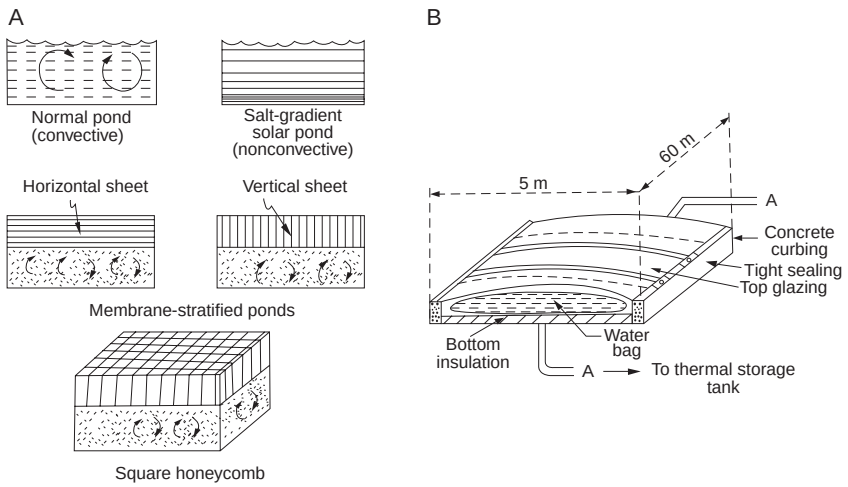


FIGURE 2 (A) Schematics of nonconvective solar pond concepts. (B) Illustration of shallow (convective) solar pond.

3. SALT GRADIENT SOLAR POND

3.1 Containment

Solar pond containment involves trade-offs between several strategies related to the location of the pond, earth excavation, the lining, insulation, and pond shape. Local geology is critically important in site selection; care must be taken to avoid underground aquifers. The pond is usually contained in earth excavated to 3 or 4 m depth. The excavation may be minimized by stacking earth to make a wall around the pond. Leakage effects from the pond can cause considerable loss of salt solution as well as stored heat. In case of permeable soil, these effects may be quite severe; care must be taken when lining the bottom and sides. A membrane liner of butynol, nylex solid vinyl, ethylene propylene diene monomer, or other suitable synthetic materials may be used. An alternative to liner is to treat the soil in such a way as to make it impermeable to hot brine solution. For example, an inexpensive liner may be buried under a layer of compacted earth.

Thermal insulation in the pond bottom and sides may be used to reduce heat losses; it may speed up pond heating but at excessive cost. Pond shape is also an important consideration. Circular surface shapes provide a minimum perimeter for a given pond surface area, involve minimum heat losses to ground, and hence are thermally the most efficient. Rectangular shapes do not have this advantage but are usually less expensive to construct. Slopping sidewalls are very effective in reducing heat loss. Sloppy walls are also desirable from the standpoint of structural stability of banks and edges; a gradient of 45° has generally been used.

3.2 Establishment and Control of Gradient

Three methods have been adopted for the establishment of initial salt density gradient in solar ponds: natural diffusion, stacking, and redistribution. The first method relies on the natural diffusion between a freshwater layer and the layer of saturated salt solution. The pond is half filled with each layer and the surface and bottom concentrations are maintained constant. Stacking involves the filling of a pond with layers of progressively changing concentrations. The redistribution method is considered to be most convenient and expedient for large-area solar ponds. In practice, the redistribution method involves filling the pond with brine of near-saturation concentration (S_c) to half of its total depth (L); the fresh water is then fed into the pond through the diffuser, which is initially placed at the bottom in the brine. The water will stir and uniformly dilute the pond brine above the diffuser. The pond surface level will start rising, and the diffuser is simultaneously raised upward in a continuous motion or in discrete steps of up to 5 cm. The timing of diffuser movement is adjusted so that the diffuser as well as the brine surface reach the final level at the same line. Consequently, in the pond region progressively diluted brine layers are generated and a constant (S_c/L) density gradient, positive downward, is established.

The mathematical expression for the rate of diffuser movement was derived by Zangrando as follows: If $x_s(t)$ and $x_d(t)$ represent the height of the pond surface and the diffuser with reference to the pond bottom at any time t, then in accordance with the proportionality criteria, we have

$$x_d(t) = b_1 x_s(t) + b_2, \tag{1}$$

where b_1 and b_2 are constants:

$$\text{At } t = 0,\ x_d(t) = 0 \text{ and } x_s(t) = 1/2 \tag{2}$$

$$\text{At } t = t_f,\ x_d(t) \text{ and } x_s(t) = 1. \tag{3}$$

Equation (1) in combination with Eqs. (2) and (3) yields

$$x_d(t) = 2x_s(t) - 1$$

or

$$\dot{x}_d(t) = 2\dot{x}_s(t). \tag{4}$$

Equation (4) implies that the rate of diffuser motion should be twice that of the surface motion.

The solar pond may be considered as a horizontal layer of incompressible fluid of depth l and infinite horizontal dimension (y,z plane); the x-axis is taken along the vertical direction and positive downward. The layer embodies the gradients of temperature and salinity corresponding to cold water at the top and hot brine at the bottom. Owing to gradients, diffusion of heat and salt occurs continuously. The solar pond may therefore be considered a thermohaline double-diffusive fluid system, and its physical behavior may be described by the following equations of fluid mechanics and salt diffusion:

1. Conservation of mass (continuity equation):

$$\frac{\partial \rho}{\partial t} + \nabla(\rho \vec{V}) = 0$$

or

$$\frac{D\rho}{Dt} + \rho \nabla \vec{V} = 0$$

2. Conservation of momentum (Navier–Strokes equation):

$$\rho \frac{D\vec{V}}{Dt} = -\nabla p + F + \mu \nabla^2 V + \left(\xi + \frac{\mu}{3}\right)\left(\nabla \vec{V}\right)$$

3. Conservation of energy (heat conduction equation):

$$\frac{DT}{Dt} = \alpha_T \nabla^2 T + Q_T$$

4. Conservation of salt:

$$\frac{DS}{Dt} = \alpha_S \nabla^2 S + Q_S,$$

where Q_T is the heat source term resulting from the absorption of solar radiation in the pond, Q_S is the salt source term resulting from the process of salt replenishment in the pond,

$$\vec{V} = \begin{bmatrix} u \\ v \\ w \end{bmatrix}$$

is the convective velocity vector, $\frac{D}{Dt}$ is the convective derivative given by

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \vec{v}\nabla,$$

μ is the coefficient of viscosity, F is the external force, and ξ is the viscosity coefficient.

Due to thermal expansion, the fluid of the bottom region tends to be lighter than the fluid at the top. Thus, the system has a top-heavy configuration, which is potentially unstable. The stability criteria in such a situation have been obtained as follows:

$$\left(\frac{\partial S}{\partial X}\right)_{\min} = \frac{\nu + \alpha_T}{\nu + \alpha_S} \times \frac{\alpha_c}{\beta_c}\left(\frac{\partial T}{\partial X}\right)$$

or

$$\left(\frac{\partial S}{\partial X}\right)_{\min} = \frac{p_r + 1}{p_r + \tau_s} \times \frac{\alpha_c}{\beta_c}\left(\frac{\partial T}{\partial X}\right), \tag{5}$$

where $p_r = \nu/\alpha_T$ is the Prandtl number, and $\tau_S = \alpha_S/\alpha_T$ is the ratio of diffusivities and referred to as the Schmidt number. Equation (5) gives the minimum salt gradient requirement corresponding to a given temperature gradient in the double-diffusive solar pond. This equation is often referred to as the hydrodynamic stability criterion in the context of static stability criteria based on a simple buoyancy model given as

$$\left(\frac{\partial S}{\partial X}\right)_{\min} = \frac{\alpha_c}{\beta_c}\left(\frac{\partial T}{\partial X}\right), \tag{6}$$

where α_c is thermal expansion coefficient, and β_c is the salt expansion coefficient.

For dilute solutions at ambient air temperature, $\nu = 7\alpha_T \approx 10^3 \alpha_S$. Therefore,

$$\frac{p_r + 1}{p_r + \tau_S} \approx 1.14.$$

This implies that a concentration gradient approximately 14% greater than the static stability is required by the hydrodynamic stability criterion.

The vertical gradient of density causes an upward diffusion of salt in the pond. For a typical sodium chloride solar pond, the amount of upward diffusion of salt is 60–80 g/m^2 day. The tendency of salt diffusion is to destroy the gradient. One gradient

control approach is to remove salt from the surface and inject it back into the bottom. For this approach, an evaporation pond is required. However, a more practical approach, which does not require the continual addition of salt, is the "falling pond" concept. A schematic of the falling pond technique is shown in Fig. 3. Hot brine is withdrawn from the bottom layer without disturbing the layers above. This is possible since in a fluid system stratified with a density gradient, selective flow of the bottom layer can be accomplished in the same manner as the withdrawal of a single card from the deck. The hot brine withdrawn from the solar pond is passed through the flash evaporator to remove some of its water. The solution, which now has a higher concentration and a smaller volume, is reinserted into the pond bottom. Consequently, the concentration gradient is maintained. The decrease in surface level is restored by the addition of fresh water to keep both the pond depth and the surface concentration constant. In the preceding process, the gradient tends to be displaced downward, hence the name falling pond technique. Following Tabor, if v_0 is the vertical velocity (measured positively downward) of the mass of the brine in the pond, V_1 and V_2 are values of volume flow rates at input and output, and A is the horizontal area of the pond (assumed to be uniform vertically), then

$$v_0 A = V_2 - V_1.$$

The resultant downward flux of salt is given by

$$v_0 S = (V_2 - V_1)S/A.$$

The condition for the maintenance of the gradient may be written as

$$(V_2 - V_1)S/A = \alpha_S(\partial S/\partial x),$$

where α_S is the coefficient of salt diffusion.

3.3 Wind and Precipitation Effects

Wind interaction with a solar pond generates waves and currents in the surface layer. Waves cascade energy through a wave spectrum that results in turbulence, which in turn depletes the salt concentration gradient and can promote complete convection in the surface region. The surface currents interact with banks and edges and create circulations in the vertical plane and cause layer mixing. The raindrops also considerably influence the dynamics of gradient erosion in the pond surface region. The wind flow over the pond surface increases the convective as well as evaporative heat losses. Suppression of wind effects is therefore very important for increasing the operating efficiency of solar ponds. Various methods of limiting the wind mixing have been tested or proposed, including floating plastic pipes, floating plastic nets, floating

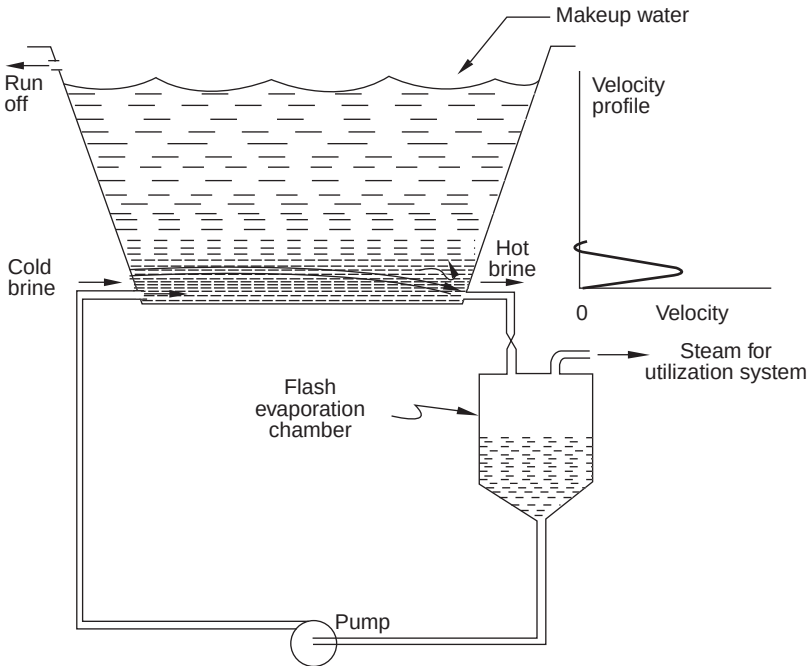


FIGURE 3 Schematic of falling pond technique of gradient control and heat removal.

plastic rings, wind baffles, and an immiscible liquid (amyl acetate and silicone oil) surface layer.

3.4 Fouling and Corrosion

Fouling and corrosion are inherent problems in solar ponds. Fouling may be caused by the falling of wind-blown material such as dirt and leaves as well as by the growth of algae or bacteria in the body of the pond. Heavy wind-blown debris normally settles to the pond bottom, where its effect on pond performance is not significant. The material of low-specific gravity is usually confined to the convective layer at the pond surface and can be removed by filtering and skimming the layer Biological growth can be checked by adding a biocide such as copper sulfate with salt at the time of filling the pond. A hot and saline environment is usually the cause of corrosion, which is often encountered in feed water pumps and heat exchangers. Careful material selection can help. Bronze, ceramic, stainless steel, and plastics (polypropylene) are well suited for saline and hot water.

3.5 Thermal Performance

In practice, a salt gradient solar pond is a three-zone configuration. The temperature gradients are relatively large in the surface region where most of the infrared portion of the solar spectrum is absorbed; the gradient induces instabilities and results in the formation of the surface convective zone. Various processes, including wind mixing, penetrative convection, and diffusive action of salt and heat, also contribute to the formation of this zone. In normal conditions, its depth ranges between 0.2 and 0.5 m. This layer is a liability on solar pond heat collection and storage since in a pond of fixed depth the growth of the mixed surface layer takes place at the expense of the thickness of the insulating gradient layer. The mixed surface layer does not provide thermal insulation to the storage layer and absorbs a considerable portion of incident solar energy. Similarly, the excessive absorption (up to 30%) of radiation by the pond floor induces thermohydrodynamic instabilities in the bottom region and makes it convective and isothermal. The bottom convective zone provides thermal storage. A convective storage layer is therefore placed in the bottom of all modern solar ponds. The minimum depth of this zone should be 0.5 m, which is sufficient to level the diurnal fluctuations in the storage zone temperature. Depths of more than 1.25 m are very effective in leveling the seasonal fluctuations, and at a depth of 10 m no seasonal variation is observed. A storage layer 1.2 m deep is usually considered appropriate. The optimum depth of the nonconvective zone is a function of collection temperature, which seems to favor a depth of approximately 1.25 m for this zone.

The following are typical thermal performance efficiencies of solar ponds:

Predicted: 20–30% at a collection temperature of 70–90°C

Achieved: 15% (Israel ponds) and 9–12% (U.S. ponds)

4. SALTLESS NONCONVECTIVE SOLAR PONDS

A saltless nonconvective solar pond was conceived primarily to overcome numerous problems encountered with conventional salt gradient solar pond. It consists of a convective water reservoir insulated on the bottom and sides. The surface insulation to the water is provided by a nonconvective layer of transparent fluid such as ethanol and silicone oil. The convective stability of the insulation layer is characterized by the stability of the fluid layer with adverse temperature gradient. In hydrodynamics, this problem is referred to as the Rayleigh–Bennard problem. The maximum stable temperature gradient across the layer is given by

$$(dT/dx)_{\max} = \frac{R_c K v}{g\beta D^4},$$

where g is acceleration due to gravity, β is the coefficient of volumetric expansion, dT/dx is the adverse temperature gradient, D is the depth of the layer, K is the thermometric conductivity or thermal diffusivity of fluid, v is the kinematic viscosity of the fluid, and R_c is the critical Rayleigh number.

The insulation layer is nonconvective if its characteristics (v, D, and R_c) are matched to the temperature difference between its bottom and top surfaces. Shaffer suggested that the insulation layer could be stabilized against convection by using viscous fluids such as silicon oil and water syrups. The control of convection by an array of horizontal membranes has been investigated by Hull. It has been shown that the separation of membranes (D) must be approximately 3–5 mm for convection to be suppressed; hence, to minimize the reflection losses at the interface, it is necessary to match the refractive index of fluid to that of the membrane. Water and ethanol are recommended as liquids for the surface

insulation zone. Teflon film seems to be an appropriate membrane. Ethanol is preferable since it can float on the hot water reservoir.

Another method of convection suppression in the insulation layer (air) involves the use of honeycomb panels. It consists of a sealed air-filled honeycomb panel floating over the hot water reservoir. It is predicted to be an attractive alternative, giving solar efficiencies of 40–60% at collection temperatures of 70–90°C. This is approximately double the efficiency of conventional salt gradient ponds.

5. APPLICATIONS

5.1 Electric Power Conversion

The Solar pond is ideal for electricity generation. One remarkable advantage is that its upper surface is cool due to evaporation and can be used for cooling the condenser. Thus, a power-generation application pond can accomplish three functions: radiant energy collection, thermal energy concentration and storage, and heat rejection. The heat available from salt gradient solar ponds is at temperatures of 50–100°C. Thus, the basic element involved in electricity generation is the low-temperature turbine. Based on the nature of the turbine, two plant types are possible: flashed steam plants and binary (organic) fluid cycle plants. Commercial turbines that operate with flashed steam cycle conditions are not available. However, binary fluid cycle turbines are commercially available. These turbines are similar to steam turbines and are based on the Rankine power cycle. Their working fluid is not steam but an organic fluid, such as freons, toluene, chlorinated hydrocarbon, or hydrocarbons with a low boiling temperature. These are high-molecular-weight fluids and so the turbine is smaller than the corresponding steam turbine. These turbines obtain 0.4–0.7 Carnot efficiency. A conceptual schematic of a binary fluid cycle-based solar pond power plant is shown in Fig. 4. Israel is the world leader in the operation of such plants. A 150-kW electric power plant was put into operation at Ein Bokek, Israel, in 1979.

5.2 Process Heating

Solar pond heat holds great promise for application for process heating. Salt production is one of the earliest applications of the solar pond. Any of the following industries and industrial processes may be supplied heat from a solar pond; salt and mineral production; drying and curing of timber; milk pasteurization; concentration and separation by evaporation; cleaning and washing in the food industry; tanning of skins; textile processing, such as wool scouring, carbonizing, and dyeing; industrial laundry; and the paper industry for preheating. Solar pond heat has been demonstrated to be applicable to the social sectors of both the developed and the developing world. Specific applications

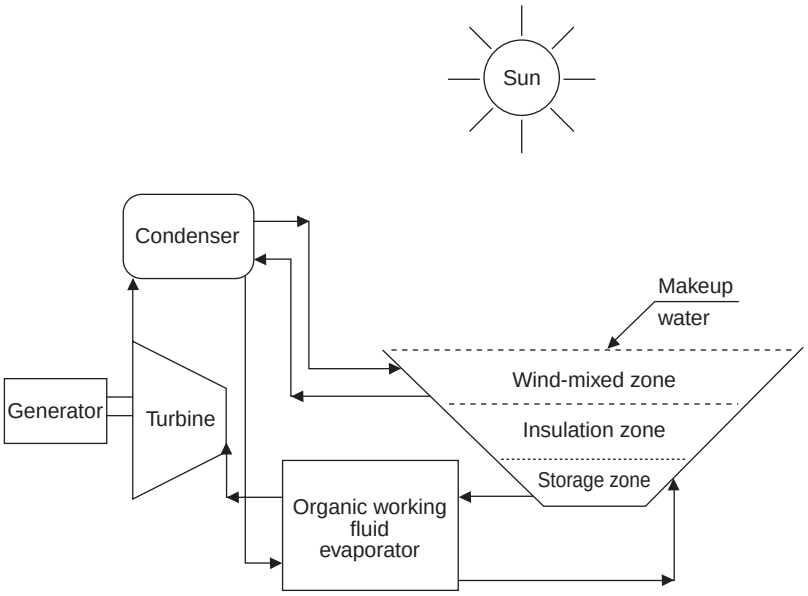


FIGURE 4 Solar pond power system.

include heating swimming pools, low-temperature cooking in southern India and western Africa, and desalinating water in outback Australia. The solar pond heat has multiple applications in the agricultural industry. At the Ohio Agricultural Research and Development Center (OARDC), a solar pond ($18.5 \times 8.5 \times 3.0$ m) has been constructed to supply the winter heat requirements of a single-family residence and partial needs of an adjacent greenhouse. In this system, hot brine is extracted from the pond and passed through shell and tube heat exchangers. If the pond temperature is higher than 40°C, the extracted heat is supplied to a water-to-air discharge heat exchanger in the green house; if the pond temperature is less than 40°C, extracted heat is first upgraded by a heat pump and then supplied for the heating of greenhouse. The method is quite suitable for this application and has good thermodynamic efficiency.

6. ECONOMICS

The major cost factors of a solar pond system are the earthwork (excavation, leveling, and compaction), the synthetic lining, the salt, the water, the land, and other miscellaneous supplies. These factors vary considerably with the location and the size of the system. Consequently, small ponds are considerably more expensive than large ponds, and there is a wide variability in the per unit area cost estimates of solar ponds operated at different locations throughout the world. Installation costs of two small-size solar ponds are provided in Table I. These estimates exclude all expenses related to the heat recovery systems.

TABLE I

Cost of Small-Size Pond Installation

Item	Australian pond ($/m^2)[a]	American pond ($/m^2)[b]
Earth work	9.81	5.00
Liner	2.79	11.00
Salt	1.91	9.70
Water	0.025	—
Land	—	—
Miscellaneous	0.52	5.90
Total	15.06	31.60

[a] Location: Alice Springs (Sheridan, 1982).
[b] Location: Miamisburg (Wittenberg and Harris, 1981).

7. ENVIRONMENTAL IMPACT

Stabilization of the greenhouse effect is the focus of global environmental concerns. The carbon dioxide emissions resulting from the use of fossil fuels must be reduced by more than 60% to stabilize the greenhouse gases at present concentrations. Stabilization cannot be achieved simply by improving fossil fuel technologies and demand management. The world, in which process heat and electricity are the major needs, has therefore pinned its hopes on the introduction of renewable energy systems. The solar pond is an inherently simple and inexpensive renewable energy means of heat and electricity. It therefore has the greatest potential for cost-effective implementation of renewable energy in the short term and long term. However, in a salt gradient pond approximately one-third of a ton of salt per unit area is used and at certain locations salt contamination may cause serious environmental hazards, such as the detrimental effect of salt in drinking water and toxicity to the vegetation surface by saline soil. The development of the honeycomb solar pond eliminates the expense of large quantities of salt and the potential problems of degrading the aquifer with salt. Fouling resulting from the falling of wind-blown debris is another inherent problem in solar ponds.

SEE ALSO THE FOLLOWING ARTICLES

Materials for Solar Energy • Solar Cells • Solar Cookers • Solar Cooling, Dehumidification, and Air-Conditioning • Solar Detoxification and Disinfection • Solar Distillation and Drying • Solar Fuels and Materials • Solar Heat Pumps • Solar Thermal Energy, Industrial Heat Applications • Solar Thermal Power Generation • Solar Water Desalination • Thermal Energy Storage

Further Reading

Clark, A. F., and Dickinson, W. C. (1980). Shallow solar ponds. *In* "Solar Energy Technology Handbook" (W. C. Dickinson and P. N. Cheremisinoff, Eds.), p. 377. Dekker, New York.

Hull, J. R. (1980). Membrane stratified solar ponds. *Solar Energy* **25**, 317.

Hull, J. R., Nielsen, C. E., and Golding, P. (1989). "Salinity-Gradient Solar Ponds." CRC Press, Boca Raton, FL.

Kaushika, N. D. (1986). Solar ponds. *Adv. Energy Systems Technol.* **5**, 75–165.

Nielsen, C. E. (1980). Non-convective salt-gradient solar ponds. *In* "Solar Energy Handbook" (W. C. Dickenson and P. N. Cheremishoff, Eds.). Dekker, New York.

Ortabasi, U., Dyksterhuis, F., and Kaushika, N. D. (1983). Honeycomb stabilized saltless solar pond. *Solar Energy* **31,** 229.

Roothmeyer, M. (1980). Soret effect and salt-gradient solar ponds. *Solar Energy* **25,** 567.

Shaffer, L. H. (1978). Viscosity stabilized solar pond. *In* "Proceedings of the ISES Congress, New Delhi, India."

Sheridan, N. R. (1982). Solar salt pond at Alice Springs, Report 1382. Queensland University, Queensland, Australia.

Tabor, H. (1980). Nonconvecting solar ponds. *Philos. Trans. R. Soc. London Ser. A* **295,** 423.

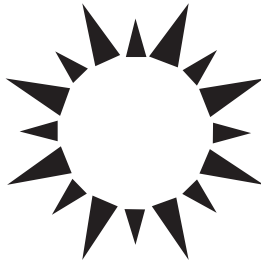
Tabor, H., and Weinberger, H. Z. (1980). Nonconvecting solar ponds. *In* "Solar Energy Handbook" (J. F. Kreider and F. Kreith, Eds.). McGraw-Hill, New York.

Weinberger, H. (1964). The physics of solar ponds. *Solar Energy* **8,** 45.

Wittenberg, L. J., and Harris, M. J. (1981). Construction and startup performance of the Miamisburg salt-gradient solar pond. *J. Solar Energy Eng.* **103,** 11.

Zangrangdo, F. (1980). A simple method to establish salt-gradient solar ponds. *Solar Energy* **25,** 467.

Solar Thermal Energy, Industrial Heat Applications



SANJAY VIJAYARAGHAVAN and D. Y. GOSWAMI
University of Florida
Gainesville, Florida, United States

1. Solar Thermal Collection Systems
2. Solar Industrial Process Heat Overview
3. Examples of SIPH Systems
4. Design and Performance of SIPH Systems
5. Economics of SIPH

Glossary

concentrator A solar collector that uses optical methods to concentrate solar radiation before absorption.
flat plate collector A solar collector that uses a flat plate geometry as the absorbing surface.
insolation A term that stands for incident solar radiation.
PTC Abbreviation for parabolic trough collector.
solar industrial process heat (SIPH) Process heat supplied using solar energy.
solar load fraction The fraction of the thermal load that is met by the solar system.
TRNSYS (Transient System Simulation) A widely used transient system simulation program developed at the University of Wisconsin.
utilizability Fraction of solar flux absorbed by the collector that is delivered to the working fluid.

Heat energy forms a large fraction of the total energy consumption in the industrial sector. For example, processes such as boiling, distillation, and polymerization that require heat input are common in chemical industries. Some other examples include cleaning and sterilization operations in the food industry, processes involving drying of materials, and the production and purification of metals and alloys. Heat is often supplied in the form of hot water and steam, or hot air, and sometimes involves the direct heating of the components in the process. All heat energy used in this fashion is referred to as industrial process heat (IPH). Most of the IPH demand today is met by burning fuels such as natural gas, oil, and coal or by using electric heating. Solar energy can be used to supply the heat input required in IPH systems. Several such systems have been designed and constructed all over the world. With the proper choice of collection system, solar systems can be built to meet any temperature level that is required by an industrial process.

1. SOLAR THERMAL COLLECTION SYSTEMS

1.1 Solar Collectors

The simplest version of a solar collector is the flat plate collector. A typical flat plate collector consists of an absorber plate that absorbs well over the range of the solar radiation spectrum. Glazing is used to reduce heat losses by thermal convection and radiation while keeping the absorber plate transparent to solar radiation. Thermal insulation is also used to reduce losses. Flat plate collectors have a working temperature of less than 70°C. To achieve higher temperatures, evacuated tube collectors, which essentially have a vacuum between the absorber and glazing tube, can be used. Because of the very reduced conduction and convection losses across the vacuum, these collectors yield higher efficiency at lower temperatures.

Concentrating collectors use optical concentration to increase the intensity of solar radiation and, thereby, achieve higher temperatures. The simplest concentrating collector can be made by using flat wall reflectors to concentrate solar radiation on a flat plate collector. Concentrating collectors can be of tracking and nontracking types. A tracking concentrator has its orientation adjusted to account for the variation in

 © 2004 Elsevier Inc. All rights reserved.

the position of the sun in the sky with time. For temperatures up to 350°C, parabolic trough collectors (PTCs) are used. These collectors focus solar radiation on a line focus where the absorber is located. A reflecting spherical or paraboloidal bowl is used when temperatures on the order of 250 to 500°C are needed. For extremely high temperatures, a large field of tracking mirrors (called heliostats) is used to concentrate solar radiation on a receiver that is located on top of a central tower.

1.2 Solar Ponds

A solar pond combines collector and energy storage in one unit. Solar radiation entering a pond travels some distance before it is absorbed in the water, increasing the temperature of the water at that depth. In a salt gradient solar pond, the density of water is increased at the lower section of the pond by the addition of a salt into the pond. Although the temperature of the lower section is increased by the absorption of solar radiation, the convection of warmer water to the surface of the pond is suppressed due to the density gradient imposed by the added salt. Salts such as magnesium chloride ($MgCl_2$) and sodium chloride (NaCl) are commonly used in solar pond applications. Bottom-layer temperatures on the order of 70 to 80°C are achievable.

1.3 Thermal Storage

Solar energy is a resource that is not available during the night and is only intermittently available on cloudy or overcast days. Many solar systems incorporate thermal storage to compensate for the intermittent nature of the solar resource. Thermal energy is normally stored as sensible heat, latent heat, or the heat of chemical reaction. In sensible heat storage, the energy is used to raise the temperature of a storage medium. Commonly used sensible heat storage materials include water, organic oils, molten salts, ceramics, and rocks. Latent heat storage takes advantage of the large amount of energy required to effect phase transformations in materials. Phase change occurs at a fixed temperature called the transition temperature or over a very small temperature range. Materials such as paraffin waxes, inorganic salts, and eutectics of organic and/or inorganic compounds are used as phase change materials. Thermal energy can also be stored as the heat of reaction in a reversible chemical reaction. The amount of energy stored depends on the heat of reaction and the extent of conversion achieved. An advantage of thermochemical energy storage is that the products of reaction can be stored at room temperature. Examples of reactions include decomposition of metal hydrides, oxides, peroxides, and ammoniated salts.

2. SOLAR INDUSTRIAL PROCESS HEAT OVERVIEW

It is estimated that solar thermal technology could theoretically supply 51% of the U.S. industrial process heating needs. Process heat solar collectors could be used throughout the year, and each system could be designed to fit the temperature level required for its specific applications (this is particularly important in the use of process heat). Table I shows the heat energy consumption by various SIC (Standard Industrial Classification) categories for selected industries in the United States. A majority

TABLE I

Summary of U.S. Industrial Heat Use by SIC Category for 1971 and 1994

	Quantity in 10^{12} kJ	
SIC group	1971	1994
20 Food and kindred products	779	1254
21 Tobacco products	14	W
22 Textile mills	268	327
23 Apparel	22	W
24 Lumber and wood products	188	518
25 Furniture	37	73
26 Paper and allied products	2006	2812
27 Printing and publishing	16	118
28 Chemicals	2536	5621
29 Petroleum products	2576	6688
30 Rubber	158	303
31 Leather	19	W
32 Stone, clay, and glass	1541	996
33 Primary metals	3468	2597
34 Fabricated metal products	295	387
35 Machinery	283	260
36 Electrical equipment	213	256
37 Transportation	310	383
38 Instruments	53	113
39 Miscellaneous	72	W
Total	14,854	22,854

Note. W, withheld to avoid disclosing data for individual establishments.

of the heat is used in mining, food, textiles, lumber, paper, chemicals, petroleum products, stone–clay–glass, and primary metals industries. In addition to the quantity of heat, the quality (i.e., temperature) is very important to match the proper collection system to the application.

Solar collection and concentrator technologies are capable of providing a broad range of temperatures to match the demand of various processes. Fig. 1 gives the distribution of the U.S. process heat requirements based on the required temperature level. Note that approximately 60% of all U.S. process heat requirements fall below 370°C. This temperature is easily achievable by using PTCs. It is estimated that energy demand in the European Union for processes at a temperature below 250°C is 1080×10^{12} kJ. Table II shows the temperatures that can be achieved with low- and medium-temperature collectors and lists a few examples of potential applications.

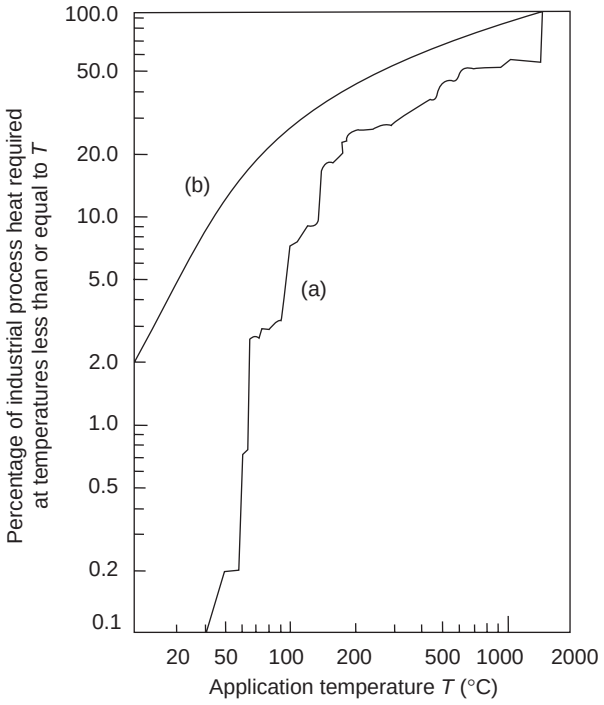


FIGURE 1 Distribution of U.S. process heat use by required temperature level: (a) heat requirements; (b) IPH requirements plus preheat from 15°C.

An example of a solar industrial process heat (SIPH) system for providing process hot water or steam is shown in Fig. 2. Solar energy is collected by circulating a heating fluid (e.g., oil, water) through the solar collectors. The choice of the collector type is strongly influenced by the temperature requirements of the process and the choices made at the time of design. For instance, in the Gould solar system (described later), parabolic troughs were used to collect solar energy, with oil circulated inside the collectors to remove the heat. Because the cost per unit collector area is very dependent on the optics used in the collection process, matching the collector characteristics to temperature requirements is very important.

The hot fluid leaving the collector is stored in a storage tank. This hot fluid can then be used to supply heat to the industrial process by exchanging heat with the process fluid. Because solar energy is an intermittent resource, most applications require a backup energy source to ensure that the requirements of the industry are met even on cloudy days and days with insufficient sunshine.

Another way in which to compensate for the intermittent nature of the solar resource is to use a storage subsystem. Using storage is particularly useful if the load is of an intermittent nature as well given

TABLE II

List of Some Low- and Medium-Temperature Solar Collector Geometries and Potential Industrial Applications

		Example applications	
Type of collector	**Typical working temperature range (°C)**	**Industry**	**Use, energy form**
Flat plate collector	≤ 70	Food industry	Washing, hot water
		Automobile manufacturing	Paint drying, hot air
		Chemicals	Borax drying, hot air
High-efficiency flat plate collector	60–120	Lumber	Log conditioning, hot water
			Kiln drying, hot air
		Dairy	Sterilization, steam
Parabolic trough collector	150–350	Textiles	Finishing, steam
		Paper and pulp	Kraft pulping, steam

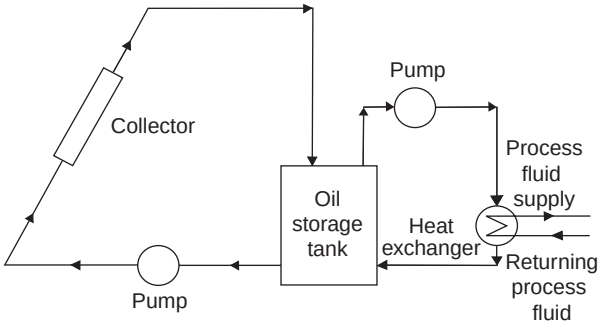


FIGURE 2 Simplified schematic of a process fluid supply system.

that the stored energy can be used to ensure that the solar system is used better. Typically, industrial process heat (IPH) systems tend not to use large storage. Instead, the solar system meets a fraction of the load, and the difference is made up using a backup source. Some SIPH systems are intentionally designed to meet only part of the load of the application. They may simply serve as the preheating arrangement so that the load on the conventional system may be reduced, thereby saving energy.

3. EXAMPLES OF SIPH SYSTEMS

Most of the IPH systems below 200°C require hot water, steam, or hot air. If the heat is needed in process air, water heating collectors along with water to air heat exchangers may be used or air heating collectors and rock storage may be used. Hot air typically is needed in agricultural drying, which may be provided by passive solar air heaters. If a large body of water is available, solar ponds may be used for low-temperature applications.

A high-temperature SIPH system may be designed to use low-, medium-, and high-temperature collectors in stages so as to minimize the cost and maximize the efficiency. A schematic of such a system designed for textile dyeing process is shown in Fig. 3.

3.1 SIPH for Textile Industries

Of all the energy used in the textile industry, 60 to 65% is used in wet processing, including dyeing, finishing, drying, and curing. The energy for wet processing is used as hot water and steam. The textile industry in the United States uses approximately 500 billion L of water per day, and approximately 25% of this water is used at an average temperature of 60°C.

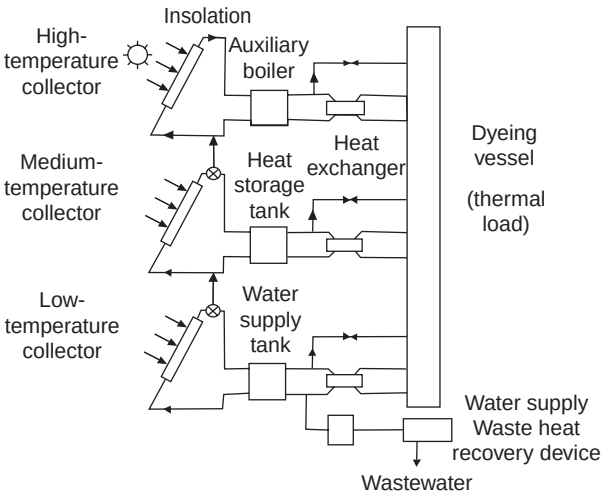


FIGURE 3 Schematic of a SIPH system for a high-temperature dyeing process.

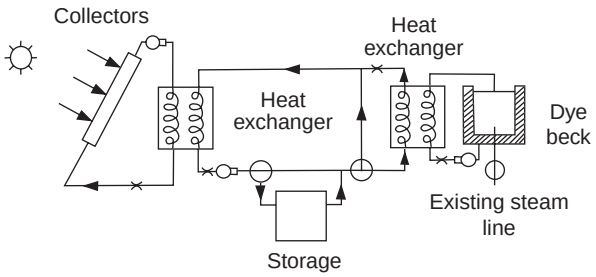


FIGURE 4 Schematic of solar energy system at Riegel Textile Corporation plant in LaFrance, South Carolina.

There are a number of examples of SIPH for textile dyeing. One of those is for a dyeing operation at the Riegel Textile Corporation plant in LaFrance, South Carolina. Figure 4 shows a schematic of the system. Because the dyeing process uses water at a temperature of 70°C, flat plate solar collectors are used in this system. The system has 621 m^2 of collector area.

The dyeing process at this plant is an atmospheric dye beck batch process at a maximum temperature of 90°C, and this is also typical of the textile industry in the United States. The batch dyeing process involves heating approximately 4500 L of inlet water between 10–25 and 90°C. The typical dye cycle is shown in Table III. The SIPH system at the Riegel Textile plant has been operating successfully. The system was simulated using the TRNSYS (Transient System Simulation) computer model. Figure 5 shows the actual performance of the system and that predicted by TRNSYS. According to measured performance of the system from August 14 to October 8, 1983, the system operated with an efficiency of 46%.

3.2 SIPH System for Milk Processing

The food system uses approximately 17% of the U.S. energy, nearly 50% of which is for food processing as hot water (<100°C) or hot air. Proctor and Morse showed that more than 40% of the energy demand in the beverage industry in Australia was in the form of hot water between 60 and 80°C. Considering the temperature and heat requirements of food processing, one would be tempted to conclude that SIPH would be ideal. However, many food processing requirements are seasonal, and this might not be economical considering the prices of conventional fuels unless the SIPH system can be used for a majority of the year. One application that is year-round is milk processing. An example of a TRNSYS simulation of a SIPH system for milk processing in the United States obtained from the literature, with the unit operations (and their respective temperatures) for this plant compatible with solar thermal energy, is as follows:

- Boiler feed makeup water (100°C)
- Pasteurizer makeup water (21°C)

TABLE III

Typical Dye Beck Heat Requirements

Dye process operation	Temperature of dye beck (°C)	Time interval of cycle (h)	Percentage of total process energy
Heat the initial load	16–32	0.5	7
Dyeing preparation	32–88	0.5	22
Dye period	88	7–9	68
Cool and reheat for dye fixation	32–43	1.0	3

- Case washer and rinsing (49°C)
- Cleanup (71°C)
- High-temperature short-time cleanup (79°C)
- Bottled water (93°C)

It was estimated that for a plant producing 170,000 kg/week of milk and 98,000 kg/week of orange juice, a total of 621.3 GJ (or 80%) of energy demand was compatible with solar energy during the summer and 724.9 GJ (or 93%) of energy demand was compatible during the winter. A 4000-m^2 flat plate collector area could provide approximately 30 to 35% solar fraction for the milk processing plant in Madison, Wisconsin; Fresno, California; or Charleston, South Carolina.

3.3 SIPH Systems for Electronic Processing

Gould Electronics, a leading U.S. electronic circuit board manufacturer, uses an SIPH system to provide the process hot water requirement in its foil division manufacturing facility in Chandler, Arizona. The SIPH system was built when Arizona had a moratorium on natural gas systems in new buildings. The SIPH system uses PTCs, through which oil is heated and stored in two 30,000-gallon (113,000-L) storage tanks. The collector field has a total aperture of 60,480 square feet (5600 m^2). The plant has a hot water requirement of 265 gallons/min (1000 L/min) at 196°F (91°C). The hot oil is used to heat up this water, with electric induction heaters serving as backup for the times when the solar system is not able to meet the full load. The field operated successfully from 1983 to 1987, but by 1990 many

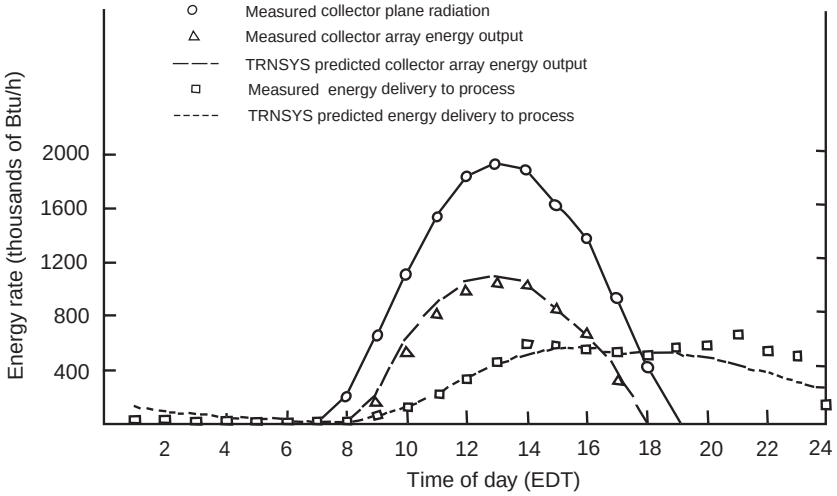


FIGURE 5 Actual and predicted system operation: August 16, 1983.

small problems had resulted in a substantial decrease in field performance. Gould engineers worked with Sandia National Laboratories to solve these problems between 1991 and 1992, restoring system availability to 90%. Based on actual operational data, it was estimated that an average saving of $7800 per month was achieved by using the solar system.

3.4 Direct Process Heat

Solar energy, with the appropriate choice of collection and concentration technology, can be used to deliver a range of temperatures, from very low to extremely high values. This allows SIPH systems to be designed to provide direct heat to a variety of processes. SIPH systems can be designed to provide temperatures that can exceed values achieved using practical combustion processes.

The use of SIPH has been proposed for various processes, including high-temperature detoxification and recycling of hazardous waste; production of fuels such as hydrogen and syngas through thermochemical cycles; production of metals such as aluminum, iron, zinc, and magnesium; and production of other materials such as silicon nitride and silicon carbide in solar furnaces. The use of concentrated solar radiation has also been proposed for the heat treatment of materials. Although several SIPH systems have been proposed for direct process heat applications, there are no known commercial systems that currently operate using these technologies. Nevertheless, research in this direction has progressed, and several laboratory and pilot plant scale demonstrations of these technologies have been completed.

4. DESIGN AND PERFORMANCE OF SIPH SYSTEMS

To assess the economic viability of any solar process, its cumulative energy delivery over its economic life on the order of years or decades must be known. It is very difficult to calculate this number accurately because (1) solar systems and their energy delivery are subject to the vagaries of local microclimates (which can change on a time scale on the order of hours) and (2) future weather cannot be predicted at this level of detail. The standard approach used to estimate future performance of a solar system is to use a typical year of past weather data and assume that it will represent the future on the average, for purposes of engineering accuracy. Two common methods used to predict long-term performance of SIPH systems are the utilizability method and TRNSYS. These methods can also be used for the design of SIPH systems. This section describes the utilizability method and the TRNSYS method.

4.1 The Utilizability Method

Utilizability, ϕ, has been used to describe the fraction of solar flux absorbed by a collector that is delivered to the working fluid. On a monthly time scale,

$$\bar{\phi} = Q_u / F\bar{\eta}_o \bar{I}_c < 1.0.$$

Here, the overbar is used to indicate monthly average values, Q_u is the monthly averaged daily total useful energy delivery, F is a heat exchanger transfer factor, $\bar{\eta}_o$ is the daily averaged optical efficiency of the collector, and I_c is the monthly averaged daily solar radiation incident on the collectors. $\bar{\phi}$ is the fraction of the absorbed solar flux that is delivered to the fluid in a collector operating at a fixed temperature T_c. The $\bar{\phi}$ concept does not apply to a system composed of collectors, storage, and other components where the value of T_c varies continuously. However, with most concentrators for concentration ratios greater than 10, the collector is relatively insensitive to a small range of operating temperatures. To check this assumption for a particular process, values of $\bar{\phi}$ at the extremes of the expected temperature excursion can be compared.

The value of $\bar{\phi}$ depends on many system and climatic parameters. Collares-Pereira and Rabl showed that only three are of the first order: the clearness index $\bar{K}_T$, the critical intensity ratio $\bar{X}$, and the ratio r_d/r_T. The first is related to insolation statistics, the second to collector parameters and operating conditions, and the third to collector tracking and solar geometry.

Empirical expressions for $\bar{\phi}$ have been developed for several collector types.

For example, for nontracking collectors,

$$\bar{\phi} = \exp\{-[\bar{X} - (0.337 - 1.76\bar{K}_T + 0.55 r_d/r_T)\bar{X}^2]\}$$

$$for\ \bar{\phi} > 0.4, \bar{K}_T = [0.3 - 0.5], \text{and } \bar{X} = [0 - 1.2].$$

Here, $\bar{X}$ is the monthly average daily critical intensity ratio, defined as the ratio of collector heat loss to absorbed solar flux at the "no net energy delivery" condition. The monthly averaged clearness index $\bar{K}_T$ is defined as the ratio of the monthly averaged daily total radiation on a horizontal surface to that on a corresponding surface outside the earth's atmosphere.

The ratio r_d/r_T is a factor that depends on solar geometry and collector tracking. Expressions are available and tabulated for various collector types.

Similar equations are available for other collector types and other clearness index values. In some cases, the collection time Δt_c might need to be determined for nontracking, low-concentration collectors.

A well-known method to calculate the long-term performance of collector systems with storage is using the "f-Chart" method developed by Klein and Beckman The f-Chart method can be used for estimating the fraction of the load that will be met by the solar thermal system. The method uses the utilizability (defined differently from the method given previously) and a few other nondimensional parameters to predict the monthly solar load fraction f_s. Predictions are based on the results of simulations for various climatic conditions by an hourly time scale computer model.

4.2 TRNSYS Computer Simulation Program

TRNSYS is a sequential–modular transient simulation program developed at the Solar Energy Laboratory of the University of Wisconsin. It is a widely used, detailed design tool involving hourly simulations of a solar energy system over an entire year. The TRNSYS computer program contains FORTRAN subroutines of nearly all the components necessary to build a solar energy system. The component models, which are either empirical or analytical, describe the component performance with algebraic and/or differential equations. A system simulation model is created by interconnecting the models of individual components. The resulting set of simultaneous algebraic and/or differential equations is solved by TRNSYS. This software is constantly being upgraded and is backed by technical support.

5. ECONOMICS OF SIPH

The commercialization of SIPH systems did not take off as was expected during the late 1970s and early 1980s. One reason why was the low price of natural gas in the market. Carwile and Hewett analyzed 15 proposed SIPH systems that were never built and the examined reasons why. It was seen that the payback period was too large when compared with industry standards for many of these projects. An additional hurdle that had to be overcome was investors' lack of confidence in this unconventional method of providing IPH. The small number of existing SIPH systems in the market made it difficult to convince investors to allocate money for the solar systems, even though the economics were favorable. Studies have been performed on the future prospects of SIPH, particularly systems using PTC for collecting solar radiation. As per the analysis, it was forecast that there was considerable potential to reduce the cost of such systems. If costs could be brought down by research and development (R&D) and by larger production volumes, a bigger market share would result in the future. Some studies estimated that with cost reductions through R&D, SIPH could have as much as 35% of the projected market by 2030.

In conclusion, SIPH systems are technically capable of meeting the IPH needs of many of today's industries. However, the higher costs of SIPH systems and the availability of cheaper alternatives have resulted in only few SIPH systems being built and still being in operation.

SEE ALSO THE FOLLOWING ARTICLES

Arid Environments, Impacts of Energy Development in • Materials for Solar Energy • Solar Cells • Solar Cookers • Solar Cooling, Dehumidification, and Air-Conditioning • Solar Detoxification and Disinfection • Solar Distillation and Drying • Solar Heat Pumps • Solar Ponds • Solar Thermal Power Generation • Solar Water Desalination • Thermal Energy Storage

Further Reading

Carwile, C., and Hewett, R. (1995). "Barriers to Solar Process Heat Projects: Fifteen Highly Promising (but cancelled) Projects," NREL/TP-472-7295. National Renewable Energy Laboratory, Golden, CO.

Collares-Pereira, M., and Rabl, A. (1979). Simple procedure for predicting long term average performance of nonconcentrating and of concentrating solar collectors. *Solar Energy* **23**, 223–234.

Duffie, J. A., and Beckman, W. A. (1991). "Solar Engineering of Thermal Processes." John Wiley, New York.

Goswami, D. Y., Kreith, F., and Kreider, J. F. (2000). "Principles of Solar Engineering." 2nd ed. Taylor & Francis, Philadelphia.

Klein, S. A., and Beckman, W. A. (1979). General design method for closed loop solar energy systems. *Solar Energy* **22**, 269–282.

Kreider, J. F. (1979). "Medium and High Temperature Solar Processes." 2nd ed. Academic Press, New York.

Procter, D., and Morse, R. N. (1975). Solar energy for Australian food processing industry. CSIRO Solar Energy Studies, East Melbourne, Victoria, Australia.

Solar Thermal Power Generation

ANDREAS LUZZI and KEITH LOVEGROVE
Australian National University
Canberra, Australia

1. **The Solar Resource**
2. **Main Solar Collector Systems**
3. **Energy Collection and Conversion**
4. **Developments of Concentrator Systems**
5. **Economics**
6. **Outlook**

Glossary

conduction band Excited electronic status of semiconductor materials, with readiness for electron transport.
heliostats Plane mirrors that continuously adjust in angle according to the sun's position, so as to reflect a beam of solar radiation to some fixed point in space.
kilowatt (kW) A unit of power equal to 1 kJ of energy applied per second. Addition of a subscript "e" indicates electrical energy, subscript "th" indicates thermal energy.
kilowatt-hour (kWh) A unit of energy equal to the power of 1 kW applied over the duration of 1 h. Addition of a subscript "e" indicates electrical energy, subscript "th" indicates thermal energy.
megawatt (MW) A unit of energy equal to 1000 kW.
nanometer A very small unit of length, equal to one one-millionth of a millimeter.
photovoltaics (PV) The science behind semiconductor-based solar energy conversion systems.
pyranometer An instrument that measures the intensity of total solar radiation.
pyrheliometer An instrument that measures the intensity of direct beam solar radiation.
receiver A device that intercepts concentrated solar radiation and transforms it into another form of energy.
secondary concentrator An arrangement of mirrors to form a light funnel to further concentrate preconcentrated solar radiation that is incident on a receiver.
stagnation temperature The equilibrium temperature a solar receiver will reach if the process that normally removes energy is turned off.
thermal loss Undesirable transfer of heat away from a solar receiver to the environment.
valence band Relaxed electronic status of semiconductor materials.

Solar thermal power generation systems capture energy from solar radiation, transform it into heat, and then use an engine cycle to generate electricity. The majority of electricity generated around the world comes from thermally driven steam-based systems. The heat for these systems is largely provided by the combustion of coal, oil, or natural gas or by nuclear fission. The concept of substituting solar radiation for these thermal inputs leads to the consideration of solar thermal power generation. The second law of thermodynamics expresses the empirically observed fact that the potential conversion efficiency from heat to electricity (work) increases with the temperature of the thermal input available. For solar systems, this naturally leads to the consideration of methods for concentrating solar radiation to achieve high temperatures. Small-scale solar thermal power generation systems were demonstrated as early as the 1860s, mainly in France and the United States. In 2003, the largest examples of solar thermal power plants are the 80 megawatt electric (MW_e) parabolic trough systems that have been operating in southern California since the late 1980s. There are nine such power plants with variable sizes that have a combined nominal capacity of 354 MW_e and generate approximately 800,000 MWh (megawatt-hour) of electricity every year. This is sufficient to service a medium-sized city of approximately 200,000 people. These power plants generate electricity on a competitive basis thanks to previous tax incentives, long-term amortization, green-power credits, some natural gas back-up, and expert operation and maintenance.

1. THE SOLAR RESOURCE

The sun is approximately 1.4 million km in diameter and approximately 150 million km from the earth. It is close to 5500°C at its surface and emits radiation

 © 2004 Elsevier Inc. All rights reserved.

at a rate of 3.8×10^{23} kW. This energy is supplied by nuclear fusion reactions near its core, which are estimated to continue for several billion years.

The major part (70%) of the solar radiation received on the earth's surface is within the visible spectrum, with a maximum intensity at approximately 500 nm. This is consistent with a "black body" radiating at 5500°C. The intensity of solar radiation reaching the surface of the earth is typically between 900 and 1000 Wm^{-2} on a clear sunny day at noon.

Only a tiny fraction (4.5×10^{-8} percent) of the total energy radiated by the sun reaches the outer surface of the earth's atmosphere, but this is still equal to approximately 1.6×10^{14} kW. Putting this in perspective, the total annual world electricity generation of approximately 1.5×10^{13} kWh could be met with a hypothetical solar thermal power generation system (using technology discussed in this article) covering an area of approximately 30,000 km^2, which is approximately equal to the size of Belgium or Taiwan, for example. The total or "global" solar radiation reaching the earth's surface is made up of "direct" and "diffuse" components. Direct radiation can be concentrated with mirrors or lenses that point directly at the sun. It is measured with a pyrheliometer. Diffuse radiation is that fraction that has been scattered by water vapor, particles, or molecules within the atmosphere. It cannot easily be concentrated and makes up approximately 15% or more of the total radiation, depending on the clarity of the sky. Diffuse radiation levels are measured with a pyranometer fitted with a shading device to exclude direct radiation. Together the diffuse and direct components combine to make up the global total and this can be measured with an unshaded pyranometer.

However, there are challenges associated with turning this solar energy source into continuous electrical power:

- The solar radiation reaching the surface of the earth is diffuse ($\sim$1000 Wm^{-2} maximum) compared to other sources of energy, such as fossil fuel flames (megawatts/m^2);
- The sun shines intermittently in a daily cycle, which in turn follows a seasonal variation and is also heavily influenced by local weather conditions;
- The position (azimuth and elevation) of the sun in the sky and hence the angle of incidence of direct beam radiation vary continuously;
- Solar radiation cannot be stored directly, but must be immediately converted into heat, electricity, or chemical energy.

Technological solutions that allow these challenges to be met do exist.

2. MAIN SOLAR COLLECTOR SYSTEMS

Various devices for collecting solar radiation thermally have been devised. At the simplest level, a flat metal plate, painted black and placed in the sun, will heat up until it reaches a temperature where the amount of heat that it radiates and loses to the air around it exactly balances the amount of energy it receives from the sun. This "stagnation temperature" occurs at approximately 80°C for a simple flat-plate solar collector. If water, for example, is passed through passages in the plate collector, then it will stabilize at a lower temperature and the water will extract some of the energy in being usefully heated up. This is the essence of solar thermal energy collection.

Greater levels of sophistication are aimed at reducing the amount of "thermal loss" from the collector surface at a given temperature. This allows energy to be collected more efficiently and at higher temperatures. The thermal efficiency of a solar collector can be defined as:

$$\text{Thermal efficiency} = \text{Energy converted/Solar energy intercepted.}$$

Starting with the flat-plate solar collector, a cover layer of glass helps by cutting down the energy lost by the circulation of cold air across the collector. If metal tubes and glass cylinders are used as collectors instead of plates, then the space can be evacuated, so that air convection losses are largely eliminated. The smart use of coating materials to produce an optically selective surface helps to reduce radiation losses. Such coatings absorb as much as possible of the high-energy (ultraviolet and visible) solar wavelengths while they emit as little thermal radiation (infrared) as possible from the plate. Various combinations of these measures are used in the design of systems for the production of solar hot water that are used around the world for domestic and industrial applications. In principle, solar collectors of this nature could be used for electricity production. However, the comparably low temperatures that are achievable limit the conversion efficiencies that are possible to low levels.

Further increasing the temperature at which energy can usefully be recovered requires some

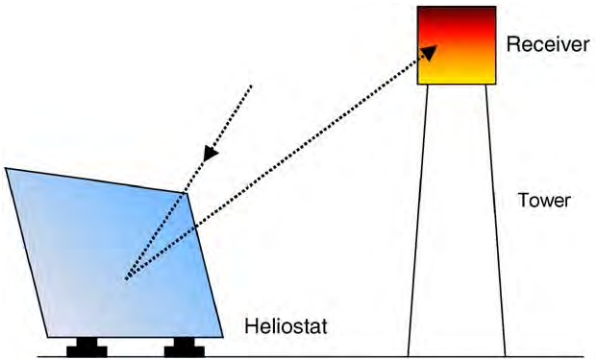


FIGURE 1 The central receiver concept; a field of plane mirror heliostats all move independently to one another to keep a beam of solar radiation focused on a single central receiver.

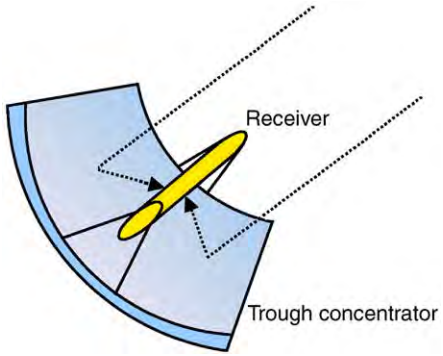


FIGURE 2 A parabolic trough concentrator focuses solar radiation onto a linear receiver when faced directly toward the sun.

method of optically concentrating the radiation so that the size of the absorbing surface and hence its thermal loss is reduced. The conceptually simplest approach is to employ a series of flat mirrors, called heliostats, that are continuously adjusted to direct solar radiation onto the absorbing surface. This is illustrated conceptually in Fig. 1. Large plants called "central receiver systems" or "power towers" have been built based on this principle and are discussed in detail in Section 4.1. The concept can also be adapted to linear absorbers and long strips of flat mirrors to create a "linear Fresnel" concentrator.

Alternatively, the mathematical properties of a parabola can be exploited for solar energy concentration. The equation for a parabola in the x–y plane is:

$$y = x^2/4f.$$

Rays of light parallel to the y axis of a mirrored parabola will all be reflected and focused at the focal point at a distance f from the vertex. As illustrated in Fig. 2, this effect can be used in a linear arrangement, where a mirrored "trough" with a parabolic cross section will concentrate solar radiation onto a line-focus when it is pointed directly at the sun. The largest solar thermal power plants thus far constructed employ this principle. They are discussed in Section 4.2. Alternatively, a mirrored dish with a parabolic cross section (a paraboloid) will focus solar radiation to a point focus (see Fig. 3). Paraboloidal dish systems are discussed in Section 4.3. Both dishes and troughs require continuous adjustment of position (or at least frequent readjustment) to maintain the focus as the sun moves through the sky.

Similar focusing effects can obviously be achieved with lenses of various kinds, but this has not been employed on the scales needed for solar thermal power systems.

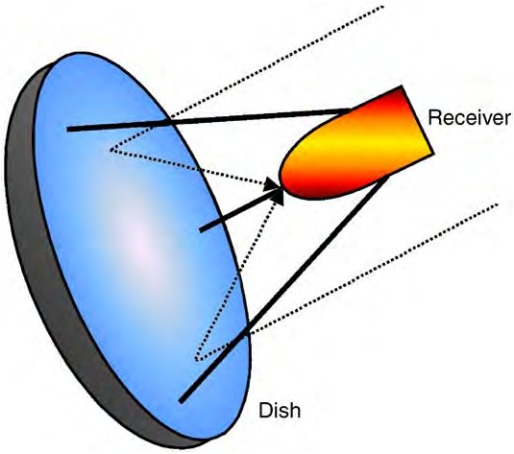


FIGURE 3 A paraboloidal dish concentrator focuses solar radiation onto a point focus receiver.

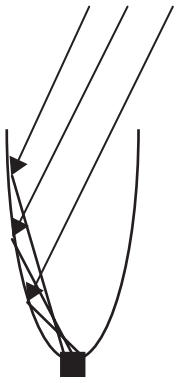


FIGURE 4 A nonimaging concentrator concentrates solar radiation without the need to track the sun.

Another alternative, which potentially avoids the need to track the sun, is to employ "nonimaging" concentration. As illustrated in Fig. 4, this involves the construction of a mirrored "light funnel" of some

TABLE I

Typical Operating Temperature and Concentration Range of the Various Solar Thermal Collector Technologies

Solar collector technology	Typical operating temperature (°C)	Concentration ratio	Tracking	Maximum conversion efficiency (Carnot)
Flat-plate collector	30–100	1	—	21%
Evacuated-tube collector	90–160	1	—	38%
Solar pond	70–90	1	—	19%
Solar chimney	20–80	1	—	17%
Linear Fresnel reflector technology	260–400	8–80	One-axis	56%
Parabolic trough collectors	260–400	8–80	One-axis	56%
Heliostat field + central receiver	500–800	600–1000	Two-axes	73%
Paraboloidal dish concentrators	500–1200	800–8000	Two-axes	80%

kind. Such a device will be able to collect rays into its aperture over a range of incidence angles and cause them to exit through a smaller aperture via multiple reflections. Nonimaging concentrators have thus far not found application as the "primary" means of concentration for solar thermal power systems but they are frequently applied as "secondary concentrators" at the focus of central receivers, dishes or troughs, where they serve to further reduce the size of the focal region.

The rays of light from the sun are not exactly parallel. This means that even a perfect optical system will produce an image of finite size, with an intensity distribution that is a maximum in the center and tapers off to zero at the edges. Imperfections in mirror shape and tracking accuracy have the effect of further spreading out the sun's image.

Each of these approaches to solar energy collection has a typical ratio of collected radiation intensity to incident solar radiation intensity, termed the "concentration ratio." Table I summarizes the options discussed and lists typical concentration ratios, the resultant operating temperatures, and the consequent thermodynamic limiting efficiency with which electricity could be produced. The limiting conversion efficiency arises from the second law of thermodynamics. The maximum efficiency for conversion of heat from a constant high temperature source is given by:

$$\text{Maximum conversion efficiency} = 1 - T_{\text{cold}}/T_{\text{hot}}.$$

This is the "Carnot limit." (Note that "absolute" temperatures must be used in the equation, usually degrees Kelvin, equal to °C + 273.15.) Real solar thermal power generation systems typically achieve approximately one-third or less of the ideal maximum Carnot efficiency.

Although higher concentration ratios give higher efficiency, they also lead to potentially higher complexity and cost. The ultimate challenge with solar thermal power systems is to produce the desired output as economically as possible. This invariably means that a tradeoff between system efficiency and capital investment drives the design process.

The three main solar thermal concentrating technologies are discussed in detail in this article as they constitute the bulk of the commercial development efforts undertaken in the area of solar thermal power generation.

3. ENERGY COLLECTION AND CONVERSION

A solar thermal power system can be presented schematically as shown in Fig. 5.

All systems begin with a concentrator, followed by a receiver that catches the concentrated radiation and converts it to another form (typically thermal energy). At this point there are two options: either the energy is further converted to the final form desired (i.e., electricity) or it is transported to another location for final conversion. The choice of the transport path provides the option of storage in the intermediate form before going to final conversion. There is also the option of designing an energy storage system after final conversion.

3.1 Receiver Types

Receivers are often built with a cavity geometry; this gives a higher effective "absorbtivity," maximizing the amount of radiation absorbed and also reducing the surface area susceptible to radiation and

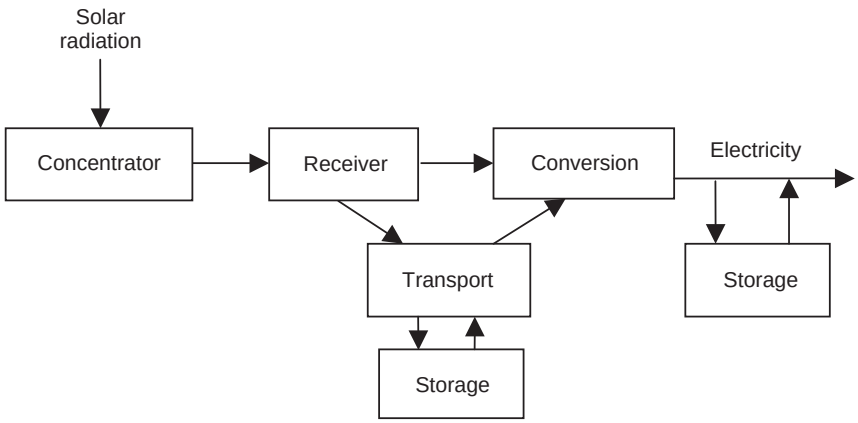


FIGURE 5 Schematic representation of the component parts of a solar thermal power system.

convection losses. Notable exceptions include the single-tube receivers with evacuated insulation used on the Californian parabolic trough power plants and the cylindrical externally irradiated receivers used on power tower systems with a completely surrounding heliostat field. There are several methods for construction. Four of the commonly used approaches are discussed here.

3.1.1 Tubular Receivers

The conceptually simplest thermal receiver type involves circulating the working fluid or heat transfer fluid through directly irradiated tubes. The tubes can be arranged to form a cavity by winding a single tube to form a helical cylinder. An alternative is arranging a number of parallel straight tubes around the outside.

Tube receivers provide a simple and reliable way of designing for high system pressures. There are issues associated with thermal cycling (particularly for steam receivers) and flow balancing in multiple parallel tube receivers.

3.1.2 Volumetric Receivers

Volumetric or direct absorption receivers aim to have the radiation absorbed by surfaces directly in contact with the working fluid. This is often achieved by having a window in front of a cavity receiver that contains an absorbing matrix, over which a fluid passes. Windows are generally made of quartz because of its superior ability to handle thermal gradients. Matrices are often ceramic (such as a foam or ceramic formed into spikes). Such receivers have been used for chemical reactors, with a catalyst impregnated in ceramic foam, and for heating air. If air is being used at atmospheric pressure, then the window is not essential.

3.1.3 Heat-Pipe Receivers

A major issue with receivers is matching the incident flux to the requirements of the energy conversion system. A very effective approach to this is the "heat pipe" receiver. The idea is to use a container with a liquid/vapor mixture. The container will be essentially all at the boiling temperature corresponding to the pressure. In regions where heat is added, liquid will evaporate, whereas it will condense where heat is extracted. The result is a "black box" of extremely high effective conductivity. Sodium (boiling point 883°C at atmospheric pressure) is favored for solar applications. This approach has been used with central receiver systems, Stirling engines, and chemical reactor receivers.

3.1.4 Solid-State Receivers

Solid-state receivers are based on the application of semiconductor materials that are designed as directly irradiated photovoltaic or thermoelectric devices. These systems produce low-voltage direct current electric power that is then transformed to higher voltage alternative current for use.

3.2 Heat Transport Mechanisms

Various methods are employed for the task of energy transport within solar thermal power plants. Five major heat transfer mechanisms for solar application are discussed here.

3.2.1 Oil

The Californian parabolic trough power plants employ a synthetic heat transfer oil. The oil is circulated through the receivers and brought back to the centrally located power plant where a heat exchanger is used to raise steam for power

generation. It has the advantage of providing stable and predictable receiver operation. Oils, however, have the disadvantages of slowly decomposing when operated at high temperatures and of environmental as well as fire hazards when leaking. The oil lines must be adequately insulated and design allowance must be made for adequate pipe expansion.

3.2.2 Steam

If the final conversion process is based on a steam Rankine cycle, then it makes sense to collect the energy as steam in the first place. The major challenge with direct steam generation, however, is the rapid thermal cycling that can occur in tubular receivers. At the point in the receiver where the tubes "dry out," the internal convective heat transfer coefficient is much less than in the regions where tubes are in contact with water. Problems arise because some sections of the tubing are exposed to rapidly alternating exposure to vapor and liquid. This causes localized thermal cycling and leads to fatigue of the metal. Insulation of pipe work and allowance for expansion is needed, as are sophisticated rotary joints to allow for solar concentrator tracking.

3.2.3 Air

Central receiver systems have been designed, among other options, to use air as heat transfer fluid either in pressurized tubular systems or in atmospherically open volumetric receivers. Systems with air temperatures of up to 1000°C have been extensively tested. Whereas tubular receivers allow for high-pressure operation with associated potential for conversion efficiency gains in gas turbines, the volumetric air receiver concept achieves very effective heat transfer from sunlight to air with minimized thermal gradients. Metallic mesh receivers have been used for temperatures of up to approximately 700°C and ceramic matrix elements for higher temperatures. Hot air from volumetric receivers can be used to heat chemical reactor tubes or boiler tubes in exactly the same way as if they were fired with fossil fuels. This provides for a uniform and controllable method for heating.

3.2.4 Molten Salt

A direct competitor to the air-based system for central receivers is molten salt. It offers the advantages of a high-specific-heat, single-phase fluid in the receiver and can also provide effective energy storage using insulated tanks. A disadvantage is that all pipes must have a provision for external heating to keep the system warm overnight and to facilitate cold starts (melting point of typical nitrate salts vary between approximately 140 and 260°C). Molten salts are corrosive and react with air and water if leaks occur. The usual nitrate salts employed, though, are environmentally benign fertilizer salts.

3.2.5 Chemical

Thermochemical cycles with fluid reactants can provide energy transport and storage. If energy is converted chemically at the receiver, then the use of counterflow heat exchangers means that reactants to and from the receiver are at ambient temperature. Hence, chemical systems not only provide an elegant way of storing energy, they also provide simple heat-loss-free energy transport. Pipes no longer need insulating, they neither expand nor contract, and much simpler mechanisms such as flexible joints are able to get through the axes of rotation.

3.3 Conversion Systems

There are three main thermo-mechanical conversion systems that are being applied with solar thermal power technologies. These are Rankine cycle, Stirling engine, and Brayton cycle systems.

3.3.1 Rankine Cycle

Steam-based Rankine cycles are responsible for the majority of electric power generation in the world. The technology is readily applicable to solar thermal systems as long as the energy collected can be transported to a central power block. Water is first compressed by a feed-water pump and then boiled and superheated (up to approximately 500°C) before being expanded through a turbine that turns an electric generator. The low-pressure steam is then condensed in a heat exchanger and fed back to the feed-water pump to be reused. Steam Rankine cycles have been, and continue to be, employed with parabolic trough, paraboloidal dish and central receiver solar thermal power plants.

Organic Rankine cycles are identical to steam-based Rankine cycles with the exception that organic fluids are employed. Organic fluids such as toluene have a lower liquid-to-gas phase change temperature than water/steam and therefore employ a lower-temperature heat input of typically between 80 and 280°C. Flat-plate, trough, dish, and solar pond systems have all been demonstrated to be suitable solar collector systems for organic Rankine cycle power generation.

3.3.2 Stirling Engines

The most extensive development of engine systems for direct solar operation has been based on the Stirling cycle, invented in 1816. The Stirling cycle employs external heating and cooling of its working fluid. The ideal cycle is made up of a constant-volume heating step from a sink temperature to a source temperature, augmented by isothermal heating, followed by a constant-volume cooling step to a sink temperature, and finished by further isothermal cooling. Dish/Stirling systems achieve very high net solar-to-electric conversion efficiencies, with the solar power world record for a precommercial 25 kilowatt electric (kW_e) unit being 29.4%.

In solar applications, Stirling engines commonly use helium or hydrogen as the working fluid. Kinematic (crankshaft-coupled) and "free-piston" engine versions are the two main concept variants developed worldwide. In a free piston engine, a floating-drive piston bounces back and forth between gas volumes at either end of a single cylinder. The piston itself serves as a linear generator. The result is a minimum number of moving parts. Solar receiver designs for Stirling engines have employed either directly irradiated fine-tubing for heating the working fluid or heat transfer heating via sodium heat pipe systems. Hybridization with fossil-fuel firing for 24 h operation has been demonstrated.

3.3.3 Brayton Cycle

The Brayton cycle is the basis of the conventional gas turbine. The cycle involves adiabatic (i.e., insulated) compression of a gas by a compressor turbine, constant pressure addition of heat, and adiabatic expansion in an expansion turbine (usually attached to the same shaft as the compressor), followed by constant pressure cooling. In a fuel-fired system, the heat addition is carried out in a combustion chamber and gases are exhausted after expansion, either with or without heat recuperation. For solar applications, heat recuperation is economically warranted for efficiency gains and gas fuel back-up advised for system control purposes as well as on-demand operation regardless of time of day.

3.3.4 Others

There are many other ways to convert thermal energy into work, many of which can potentially be applied to solar thermal power generation systems. They include, for example, the following:

- Thermoelectric converters produce electricity directly from heat. Semiconductor-based systems work in an analogous way to photovoltaic (PV) cells. Excitation of electrons from the valence band into the conduction band of the semiconductor material occurs via thermal excitation rather than individual photon absorption;
- Thermo-photovoltaics use PV cells tailored to infrared thermal radiation wavelengths to convert the radiation re-emitted from heated surfaces; and
- Magneto-hydrodynamic converters use the expansion of heated, ionized gas through a magnetic field to generate a potential difference.

4. DEVELOPMENTS OF CONCENTRATOR SYSTEMS

The three major solar thermal concentrating technologies, central receivers, parabolic troughs, and paraboloidal dishes, are discussed in more detail in this section.

4.1 Central Receiver Systems

The central receiver (or power tower) concept was first proposed by scientists in the U.S.S.R. in the mid-1950s. The first experiment was established in Sant' Ilario near Genoa, Italy, in 1965 by Professor Giovanni Francia. He installed 120 round mirrors, each the size of a "tea table," focusing on a small steam generator on top of a steel frame. The product was superheated steam (500°C, 10 MPa).

Central receivers have the advantages that large-area radiation collection to a central point occurs optically and that all the energy conversion takes place at this single fixed point. This avoids the need for energy transport networks and allows investment to improve the efficiency and sophistication of the energy conversion process to be made more cost-effectively. Associated disadvantages are that they must be built as single large systems, without the modularity benefits of distributed systems, such as troughs or dishes. The fixed position of the receiver also means that heliostats do not point directly at the sun for the majority of the time. The amount of collected solar radiation per unit area of mirror is therefore less than with the other solar concentrating technologies.

Major investigations during the past 20 years have focused on four heat transfer fluid systems. They are water/steam, sodium, molten salt, and air (see Section 3.2). An overview of the main grid-connected central receiver research facilities and demonstration

TABLE II

Summary of Grid-Connected Central Receiver Research Facilities and Demonstration Power Plants Built over the Past 20 years

	Net electric power	Total reflector area	Heat transfer fluid	Effective storage capacity (electrical or thermal)	Period of service
Eurelios (Italy)	1 MW	6216 m^2	Water/steam	0.5 MWh_e	1980–1984[a]
Sunshine (Japan)	1 MW	12,912 m^2	Water/steam	3 MWh_e	1981–1984[a]
IEA-CRS (Spain)	0.5 MW	3655 m^2	Sodium	1.0 MWh_e	1981–1985[b]
Solar One (United States)	10 MW	71,447 m^2	Water/steam	28 MWh_e	1982–1988[c]
Solar Two (United States)	10 MW	81,344 m^2	Molten salt	107 MWh_{th}	1996–1999[d]
CESA-1 (Spain)	1.2 MW	11,880 m^2	Water/steam	3.5 MWh_e	1983–1984[b]
Themis (France)	2.5 MW	10,794 m^2	Molten salt	12.5 MWh_e	1983–1986[b]
MSEE (United States)	0.75 MW	7845 m^2	Molten salt	2.5 MWh_e	1984–1985[b]
SES-5 (former Soviet Union)	5 MW	40,584 m^2	Water/steam	1.5 MWh_e	1985–1988[b]
Weizmann (Israel)	0.5 MW (since 2001)	3472 m^2	Beam-down concept	None	1988–today[b]

[a] Dismantled.
[b] Continued operation as research facility until today.
[c] Converted to Solar Two.
[d] Mothballed.

FIGURE 6 A 120 m^2 heliostat for central receiver power plants. Source: Sanlucar-120.

power plants built over the past 20 years is given in Table II.

Thus far, power generation has been via conventional steam Rankine cycles at the base of the tower. In early systems, tubular receivers were built to produce superheated steam directly. Thermal problems associated with the unsteady boundary between liquid water and steam in the tubes, however, motivated a move to secondary heat transfer fluids, such as sodium, molten salt, or air. Liquid sodium provides good heat transfer behavior, but carries the disadvantage that all the transport piping must be heated above sodium's melting point (close to 100°C). The possibility of leaks also poses a significant fire risk. The use of molten salt avoids the fire risk but is well suited to storage in tanks to allow power generation when there is no sun. As a downside, the nitrate salt mixtures commonly used have a melting point of between 140 and 260°C and corrosion challenges arise if leaks occur. Being the most environmentally friendly heat transfer fluid, air has been extensively and successfully tested at scales of up to 3 megawatt thermal (MW_{th}).

Heliostat fields can either surround the tower or be spread out on the shadow side of the tower. Whereas externally irradiated cylindrical receivers are predominantly employed for surround fields, cavity receivers are used with one-sided heliostat fields. System designers have developed optimization strategies that determine the best arrangement for a given number of heliostats. These take into account the effects of shading between heliostats during the course of the day, the spread of the field, and the optical inefficiency that increases as heliostats are further from the tower.

Two generic approaches to heliostat design have been used. The most obvious is a plane structure with rigid mirror facets mounted on it. The structure sits on top of a pedestal with a drive arrangement that allows for two-axes tracking of the sun. The other alternative is termed the "stretched-membrane" approach. As the name implies, membranes are stretched across a circular frame in a manner similar to a drum skin. Thin stainless steel sheets covered with thin-glass mirror tiles are used successfully. A variety of reflective polymer films have also been tested over the years, but face issues of limited

durability. Focusing of this heliostat concept is achieved by applying a small vacuum inside the drum. In this way heliostats can focus their own sun image onto the tower receiver rather than directing a plane beam at it. Figure 6 illustrates an example of a rigid mirror facet heliostat design. Development trends have suggested that larger heliostats might be more cost-effective. Current optima are between 100 and 150 m^2.

The largest central receiver solar thermal power plant demonstrated thus far is the "Solar Two" plant in southern California. This plant is an updated version of the previously operated "Solar One" system. Extra heliostats were added and the receiver was converted from direct steam generation to molten salt. Figure 7 shows the Solar Two plant in operation and Fig. 8 is a schematic illustrating the operating principles. When the solar field is operating, the molten nitrate salt moves from a "cold" (288°C) storage tank via the receiver at the top of the tower, where it is heated, to the hot storage tank (566°C). Independently of the solar energy collection process, salt from the hot tank is passed through a heat exchanger where heat is transferred to produce superheated steam, with the salt passing back to the cold storage tank. The steam is used in a conventional steam turbine power plant for electricity generation.

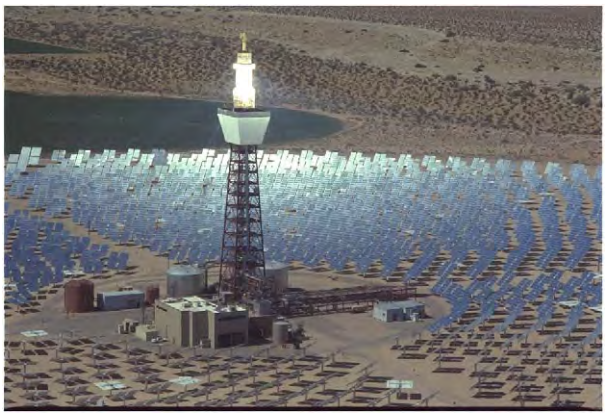


FIGURE 7 The "Solar Two" (previously "Solar One") central receiver power plant in operation. Source: Sandia National Laboratories.

The Solar Two plant has 1018 heliostats of 39.1 m^2 plus a further 108 heliostats of 95 m^2. Under nominal conditions, 48 MW_{th} is concentrated onto the receiver that sits at the top of a 91 m high tower. Steam is produced in the heat exchangers at 10 MPa and 538°C and the net electrical output is 10.4 MW_e.

One of the newest developments in the central receiver area is the "beam-down" concept proposed and tested in part by the Weizmann Institute of Science in Israel. Rather than converting the concentrated solar energy at the top of the tower in a receiver, a hyperbolically shaped secondary mirror directs the converging radiation vertically downward to a focal point at the bottom of the tower. On the ground, an additional nonimaging mirror concentrates the radiation further before the concentrated sunlight is captured by a volumetric receiver. This receiver is capable of reaching very high air temperatures to power, for instance, a Brayton cycle gas turbine.

4.2 Trough Systems

Solar thermal power in the form of mechanical energy for water pumping was established for the

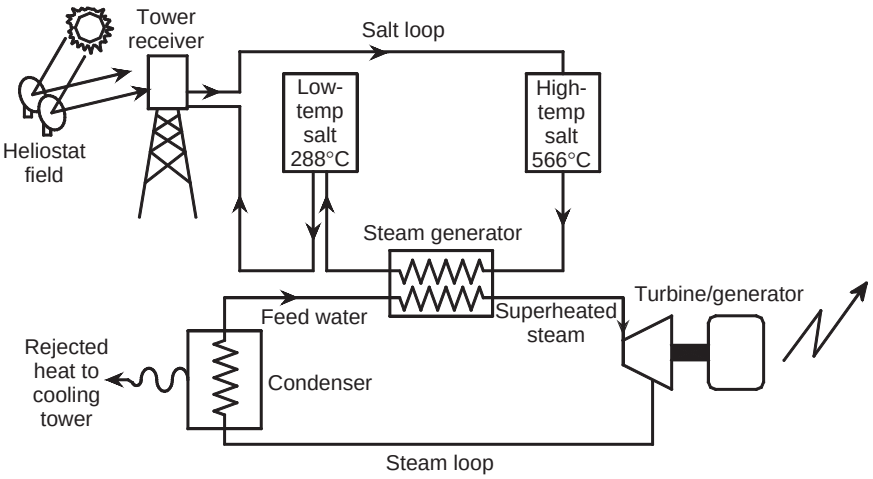


FIGURE 8 Schematic of "Solar Two" operation.

first time by Frank Shuman near Cairo in 1913 ($\sim$40 kW). It incorporated a water/steam operated parabolic trough array (five units, 4 m $\times$ 62 m each) and a low-pressure condensing steam engine. Solar electric trough development was continued in the mid-1970s by the U.S. Department of Energy. The first experimental system started operation in 1979. At the same time, a private research and development company from Israel decided to design and commercialize parabolic trough "solar electric generating systems" (SEGS) in southern California. This decision was strongly motivated by favorable power purchase agreements and tax credits offered in the state of California.

Nine SEGS power plants were built between 1984 and 1989, with a combined capacity of 354 MW_e. These investor-owned, natural gas-assisted power plants have not only been operating on a fully commercial basis ever since, but have also been continuously improved. They operate beyond their nameplate capacity, generating important solar peaking power for the southern California grid, at the world's most competitive solar power prices.

During the early 1980s, small, trough-based solar thermal demonstration power systems were constructed in the United States, Japan, Spain, and Australia. Table III lists the details of these plants. Specifications for the nine SEGS plants are given in Table IV.

Figure 9 shows one of the SEGS plants and Fig. 10 illustrates the operating principles schematically. A synthetic heat transfer oil is pumped through the trough solar collector array and heated to 400°C. This oil is then used to produce steam in heat exchangers before being circulated back to the array. The steam is used in a conventional steam turbine-based electricity-generating plant. Although some hot oil-based energy storage was provided in the first plant, the SEGS systems overall rely on natural gas firing to provide continuous operation when the sun is not available. The special tax law for these power plants limits the fossil-fuel cofiring to a maximum of 25% per annum of the total heat energy supplied to the boilers.

The SEGS troughs are built using a galvanized steel space frame. This frame supports 4 mm thick toughened glass mirror facets that are shaped by heating and molding the raw glass to match the parabolic profile, before silvering. Each mirror facet is supported at four attachment points. The most

TABLE III

Details of the Main Small, Trough-Based Solar Thermal Demonstration Power Plants

	Net electric power	Total aperture area	Heat transfer fluid	Effective storage capacity (electrical or thermal)	Duration of service
Coolidge (United States)	0.15 Mw	2140 m^2	Synthetic oil	5 MW_{th}	1980–1982
Sunshine (Japan)	1 Mw	12,856 m^2	Water/steam	3 MW_e	1981–1984
IEA-DCS (Spain)	0.5 Mw	7622 m^2	Synthetic oil	0.8 MW_e	1981–1985
STEP-100 (Australia)	0.1 Mw	920 m^2	Synthetic oil	117 MW_{th}	1982–1985

TABLE IV

Details of the Californian Parabolic Trough "SEGS" Solar Thermal Power Plants

	Net electric power	Total aperture area	Duration of service
SEGS I	13.8 MW	83,000 m^2	1984–present
SEGS II	30 MW	19,000 m^2	1985–present
SEGS III	30 MW	230,000 m^2	1986–present
SEGS IV	30 MW	230,000 m^2	1986–present
SEGS V	30 MW	251,000 m^2	1987–present
SEGS VI	30 MW	188,000 m^2	1988–present
SEGS VII	30 MW	194,000 m^2	1988–present
SEGS VIII	80 MW	464,000 m^2	1989–present
SEGS IX	80 MW	484,000 m^2	1989–present

recently constructed systems (termed LS-3) are 5.76 m wide and 95 m long, resulting in a total aperture of 545 m^2 each. Two hundred twenty-four glass mirror facets are used and a geometric concentration ratio of up to 80:1 is achieved. The trough units are lined up in north–south rows and track the sun from east to west during the course of the day. Each parabolic trough has its own positioning and local control system.

FIGURE 9 View of SEGS trough-based solar thermal power plant in southern California. Source: Kramer Junction Operating Company.

The receiver units of the SEGS troughs consist of a stainless steel tube 70 mm in diameter, covered by a Pyrex glass envelope that is sealed to the tube via metal bellows at each end. The space between the steel and glass is evacuated to minimize thermal losses. The surface of the stainless steel absorber tubes is coated with a selective surface that absorbs up to 97% of the incident solar radiation while minimizing the amount that is radiated at thermal (infrared) wavelengths. The combination of trough and receiver is capable of operating at temperatures in excess of 400°C. However, the synthetic heat transfer oil becomes chemically unstable and begins to break down at temperatures above 300°C. Approximately 350 m^3 of oil circulates in each of the 30 MW_e plants. Depending on the operating regime, this oil may need to be replaced at rates of up to 6% per year as a result of thermochemical breakdown.

Each of the latest SEGS plants employs a steam turbine with a reheat cycle and multiple extractions. With steam inlet conditions of 10 MPa and 370°C, thermal-to-electric conversion efficiencies of approximately 37% are achieved, giving overall peak solar-to-electric conversion efficiencies of up to 24%.

Regular maintenance is required for all solar thermal power plant systems. For the SEGS plants, routine cleaning and replacement of broken mirror facets and receiver modules forms a major part of the maintenance program. Major improvements have been realized over time.

Research and development on parabolic trough systems has continued since the SEGS plants were completed. A major area of investigation has targeted the replacement of the heat transfer oil with direct

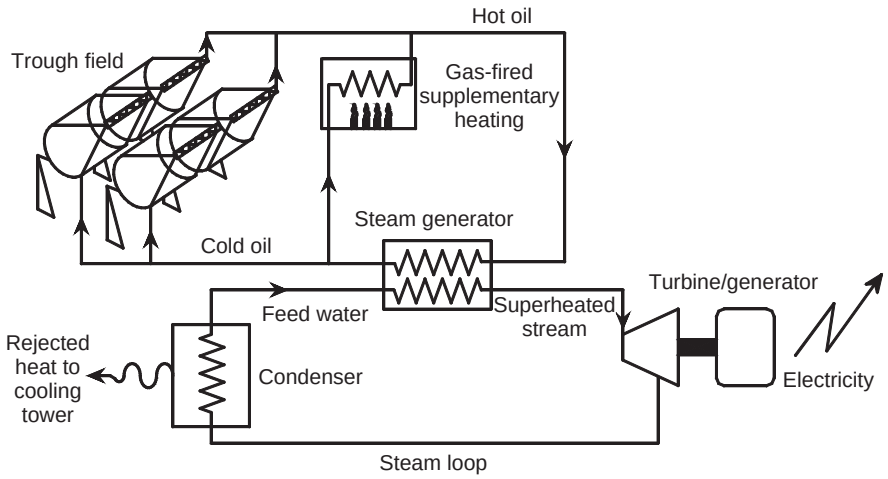


FIGURE 10 Schematic representation of SEGS plant operation.

generation of steam in the receivers. Direct steam generation allows collection of energy at higher temperatures as well as elimination of one heat-exchange step and thus improves the efficiency of the steam turbine. It also avoids the need to replace the costly oil and eliminate the inherent fire risk. When attempting once-through-to-superheat boiling, the challenge is in managing the rapidly changing thermal stresses induced in the receiver tube by the unsteady, localized movement of the liquid/vapor phase boundary when boiling is occurring.

In a variation of linear focusing technology, compact Fresnel-type concentrator systems are being developed at various places around the world. This concept dates back to the 1950s and is much like a linear version of a central receiver system. Fixed linear receivers are illuminated by a series of long, narrow mirrors that track the sun individually in a single axis. If several receiver rows are installed side by side, then the individual mirror units can switch from one receiver to another depending on the relative optical efficiencies at different times of the day. Alternatively, a secondary concentration of the reflected light near the receiver can be used to ease precision and costs of the linear primary solar concentrator.

4.3 Paraboloidal Dishes

The first realization of a solar dish concentrator/engine system design dates back to 1864. The Frenchman Augustin Mouchot built a variety of "thermo-piles" and "solar engines" using conical mirrors. There are many designs and states of development of paraboloidal dishes.

Paraboloidal dish concentrators must point directly at the sun for proper operation. This necessitates two-axes tracking mechanisms. There are two main approaches to this: "altitude/azimuth" tracking or "polar/equatorial" tracking. In the first approach, the dish is mounted with a horizontal axis pivot (to allow altitude adjustment) plus a vertical axis pivot (to allow azimuth adjustment). A polar/equatorial system has a main axis of rotation, the angle of which is adjusted on a daily basis so that it is at a right angle to the sun's noon elevation. With the polar axis correctly adjusted, tracking during the day involves following the sun from sunrise to sunset on a constant basis with a single axis of movement. Altitude/azimuth arrangements have the disadvantage that both actuators must work throughout the day and the azimuth adjustment must be relatively rapid near midday for high sun elevations. Despite this, it is sometimes used for pragmatic structural design reasons.

Dish-based solar thermal power systems can be divided into two groups: those that generate electricity with engines at the focus of each dish and those that use some mechanism to transport heat from an array of dishes to a single central power-generating block. From a research and development perspective, dish systems have two major advantages over central receivers in particular, in that a single dish prototype can be constructed and tested relatively economically. Furthermore, for the dish/engine approach, a single module on its own can demonstrate the operation of the full system. Consequently, there have been well over 40 different dish prototypes developed thus far, by both research institutes and private engineering companies.

The variations in structural design of dish concentrators are very similar to the different types of heliostats used with central receivers. Possible different approaches include the following:

- A space frame structure that supports a number of rigid mirror elements held in a paraboloidal orientation;
- A single rigid paraboloidal surface covered with mirror elements;
- A single, mirrored, stretched-membrane that is pulled into a near-paraboloidal shape on a drum-like structure; and
- A space frame supporting a series of small stretched membrane mirror facet elements.

The high concentration ratios achievable with dish concentrators allow for efficient operation at high temperatures. Stirling cycle engines are well suited to construction at the size needed for operation on single-dish systems and they function with good efficiency, with receiver operating temperatures in the range of 650 to 800°C. To achieve good power-to-weight ratios, working gas pressures in the range 5–20 MPa are employed and the use of high-conductivity gases, such as hydrogen or helium, gives improved heat transfer. Dish/Stirling units of 25 kW_e have achieved solar-to-electric conversion efficiencies of close to 30%. This represents the maximum net solar-to-electric conversion efficiency yet achieved by any nonlaboratory solar energy conversion technology.

Figure 11 shows two dish/Stirling system operating in southern Spain. The dish is fitted with a 10 kW_e two-cylinder kinematic Stirling engine. One-millimeter-thick, back-silvered glass mirror facets are employed to achieve very high focusing accuracy.

FIGURE 11 Paraboloidal dish solar concentrator with a receiver-mounted Stirling engine operating in southern Spain. Source: Schlaich, Bergermann, and Partner.

FIGURE 12 Seven hundred dishes producing steam for a central turbine system. Source: LaJet.

Including various alternative versions of the generic system, over 80,000 h of cumulative on-sun operating time have thus far been clocked with dish/Stirling systems.

The largest distributed array/central plant solar thermal power system that has yet been demonstrated is the "Solarplant-1" system built in southern California in 1983–1984. A photograph of this system is shown in Fig. 12. It consisted of 700 dishes with a total collecting area of 30,590 m^2. The dishes generated steam in their cavity receivers, with 600 of the dishes producing saturated steam at approximately 6 MPa (275°C) and the remainder taking the saturated steam to a superheat of 460°C. The steam was transported through an insulated pipe network to a central steam turbine-based generating plant that produced a nominal output of 4.9 MW_e. The dishes were constructed with multiple stretched-membrane mirror elements. The plant ceased operation in 1990.

The world's largest solar dish, with an aperture of 400 m^2, is shown in Fig. 13. It was developed for use in distributed systems similar to Solarplant-1. Two prototypes are in operation: one at the Australian National University in Canberra, Australia and the other at the Sde Boquer Desert Research Centre of the Ben Gurion University in Israel. These dishes have triangular mirror panels supported on a hexagonal aperture space frame structure. Altitude–azimuth tracking is employed, with the horizontal axis near the base of the dish so that it can be parked in a horizontal position relatively close to the ground. This helps to reduce wind resistance and thus improve storm survivability. The Canberra dish has a cavity receiver based on a single helical winding of tubing to serve as a once-through-to-superheat boiler producing superheated steam at 5 MPa and

FIGURE 13 The world's largest 400 m^2 solar dish system. Source: Australian National University.

500°C. Although intended for operation in large arrays, the prototype system is connected to a small reciprocating steam engine that is capable of generating up to 45 kW_e for the local electricity grid. The thermal efficiency from sun to steam is between 85 and 90% and, in a large steam turbine-based system, production of electricity at a rate of approximately 100 kW_e per dish would be expected.

Research and development on paraboloidal dish systems around the world strive to improve and develop dish designs to reduce the costs and to optimize the reliability and efficiency of conversion systems. A number of groups are investigating the use of small solar-driven gas turbines at the focus of dishes (dish/Brayton systems). Such systems offer the potential for high-efficiency operation and moderate maintenance requirements.

Chemical conversion systems are also investigated using paraboloidal dish concentrators. High-temperature reversible chemical reactions, such as

methane re-formation (up to 900°C) or ammonia dissociation (up to 700°C), can be used to store heat energy in chemical form. The resulting chemical products, such as hydrogen and carbon dioxide, are stored and transported at ambient temperature and thereafter recombined in an exothermic reaction to provide the heat for power generation on a continuous, 24 h basis.

5. ECONOMICS

Solar thermal power generation technologies have achieved impressive development progress over the past two decades. Some of this progress, most notably with parabolic troughs, has resulted in successful technology commercialization, delivering the lowest-cost solar power yet achieved.

Given the very many different technology routes, development status, application options, and project sizes, solar thermal power generation technologies cannot be compared, let alone summarized, on a one-line economic basis. It is, however, valuable to look at the achievements, status, and trends of capital costs, operation and maintenance costs, and levelized electricity costs. These costs are strongly dependent on the country and on site-specific labor-rate structures. Although some general observations can be made, it is not necessarily valid to translate costs experienced with a particular project in one country to operation in another by a simple exchange-rate conversion.

5.1 Capital Costs

Capital costs arise for land, site preparation, construction work, solar collectors, power conversion units, the balance of plant including controls, and infrastructure as well as services such as engineering, project management, and commissioning.

Depending on the plant capacity, its technical complexity, the chosen location (geophysics, climate, access, water, etc.), and the choice of storage and/or fuel hybridization or back-up, solar thermal power plant capital costs over the past two decades have varied greatly between approximately U.S. $3000 and U.S. $9000 per kilowatt of nominal plant capacity.

Economies-of-scale progress with the Californian SEGS parabolic trough plants has shown that a cost reduction of typically between 12 and 16% has been achieved on doubling of the plant capacity.

In 2003, 10 to 50 MW_e power plants are offered for approximately U.S. $2500–$3000 per kilowatt, whereas 0.5 MW_e to 10 MW_e power plants require investment capital of U.S. $3000–$5000 per kilowatt of nominal plant capacity. Smaller projects, such as one-off 10–25 kW_e dish/Stirling units, cost approximately U.S. $6000–$7000 per kilowatt.

Given that solar thermal power generation technologies are comparably "low-tech" offerings based primarily on structural and civil engineering skills and efforts, cost reduction curves are quite predictable. Performance improvement, manufacturing progress, economies-of-scale, and options of hybridization as well as energy storage dominate cost reductions, indicating trends toward capital costs of approximately U.S. $1500–$2500 per kilowatt.

5.2 Operation and Maintenance Costs

Operation and maintenance (O&M) costs are composed of labor and materials costs. The split between these two varies according to technology choice, remoteness of location, labor rates, and water costs, with typical ratios of approximately 1:2 for large power plants and 2:1 for small power plants.

Whereas current commercial solar thermal power plants have typical O&M costs in the range of U.S. $0.025–$0.035 per net kilowatt-hour of power generated, continuous optimization efforts indicate a strong trend toward U.S. $0.01–$0.015 per kilowatt-hour for utility-scale power stations in the range of 50–100 MW_e over the next decade.

5.3 Levelized Electricity Costs

Levelized electricity costs (LECs) are averaged electricity costs calculated on a simple basis of fixed economic parameters and a straight-line depreciation of capital cost over the duration of the project life. Technologies as well as projects are difficult to compare due to the large variety of design- and finance-specific parameters chosen.

Nevertheless, when analyzing well-demonstrated, megawatt-scale projects on a comparative basis of 30 years of project life, 8% real discount rate, 1% insurance rate, straight-line depreciation, and 0% income tax rate, typical LECs of U.S. $0.12–$0.16 per kilowatt-hour result. Small-scale, one-off projects, such as dish/Stirling systems, are likely to achieve LECs of U.S. $0.20–$0.30 per kilowatt-hour once technically matured and commercially operated. Following the ongoing efforts and predictions of major project developers, megawatt-scale solar thermal power plants are likely to ultimately generate LECs of U.S. $0.05–0.1 per kilowatt-hour.

6. OUTLOOK

The future of solar thermal power generation holds great commercial promise and a broad spectrum of interesting new opportunities. Among these possibilities are 24 h solar-only power generation thanks to thermal and/or thermochemical energy storage, the re-formation and synthesis of high-value fuels and fine chemicals, the thermal and/or thermochemical production of hydrogen, and the detoxification of hazardous waste.

SEE ALSO THE FOLLOWING ARTICLES

Arid Environments, Impacts of Energy Development in • Cogeneration • Hybrid Energy Systems • Materials for Solar Energy • Solar Cells • Solar Cooling, Dehumidification, and Air-Conditioning • Solar Energy, History of • Solar Fuels and Materials • Solar Heat Pumps • Solar Ponds • Solar Thermal Energy, Industrial Heat Applications • Solar Water Desalination • Thermal Energy Storage

Further Reading

Becker, M., Gupta, B., Meinecke, W., and Bohn, M. (1995). "Solar Energy Concentrating Systems, Applications, and Technologies." C. F. Mueller Verlag, Heidelberg, Germany.

Gordon, J. (ed.). (2001). "Solar Energy, the State of the Art, ISES Position Papers." James and James, London.

International Energy Agency's Solar PACES (Solar Power and Chemical Energy Systems) Web site: www.solarpaces.org.

Meinel, A. B., and Meinel, M. P. (1976). "Applied Solar Energy: An Introduction." Addison-Wesley, Reading, MA.

Pilkington Solar International. (1996). "Status Report on Solar Thermal Power Plants." Pilkington Solar International, Cologne, Germany.

Winter, C.-J., Sizmann, R. L., and Vant-Hull, L. L. (eds.). (1991). "Solar Power Plants: Fundamentals, Technology, Systems, Economics." Springer Verlag, New York.

Solar Water Desalination

E. DELYANNIS and V. BELESSIOTIS
National Center for Scientific Research "Demokritos"
Athens, Greece

1. **Historical Background**
2. **Solar Stills**
3. **Principles of Solar Stills**
4. **Heat and Mass Balance in the Still**
5. **Prediction of Daily Output**
6. **Materials of Construction**
7. **Multistage Solar Stills**
8. **Solar-Driven Desalination**
9. **Energy Conversion Systems**
10. **Conversion to Electricity**
11. **Sizing a Solar-Driven System**
12. **Conclusion**

Glossary

electrodialysis Desalination method using membranes to separate anions and cations from water by applying an electric field to the membranes.

evacuated tube collectors Medium-temperature solar radiation collectors (up to 250°C) using selective absorption surfaces into the vacuum tubes.

flat-plate collectors Devices that absorb solar radiation on black surfaces and transfer the thermal energy to a fluid. They are low-temperature devices up to 95°C.

focusing collectors Devices that focus solar radiation on a surface or a point. The high-temperature thermal energy is transmitted to pressurized water or it produces vapor.

photovoltaic Devices that transform solar radiation into electricity by the use of semiconductors.

renewable energies Inexhaustible natural energy resources (solar, wind, geothermal, tidal, wave, and biomass), replaceable as they are consumed. These exist in huge amounts worldwide.

reverse osmosis Pressure-driven desalination method using semipermeable membranes that permit only water to be transferred through the membrane, and the solutes are retained.

solar collectors Devices that absorb solar radiation in select black surfaces and transform it to thermal energy, which is transferred to a suitable fluid.

solar ponds Small artificial lakes with salt gradient water that absorb solar radiation and act as storage devices for hot water and as devices supplying hot water with a temperatures up to 90°C.

Solar energy desalination is a method by which the sun's energy is used to desalinate brackish or seawater to produce fresh drinkable water. There are two methods for using solar energy: directly by heating and evaporating the brackish or seawater in a solar still (this method is called solar distillation) and indirectly by capturing solar energy using one of the techniques that transform solar radiation into thermal or electrical energy to drive a conventional desalination method (the indirect method is called solar-assisted or solar-driven desalination). Indirect thermal solar desalination methods are the multistage-flash thermal distillation method, the multiple-effect thermal distillation method, and the thermal vapor compression method. Electrical power generated from the sun's energy is used to drive reverse osmosis, electrodialysis, or mechanical vapor compression desalination methods. Theoretically, solar energy can replace any other energy source, but from a practical standpoint this is not yet technically feasible, especially for large-capacity solar desalination plants. This is due to the spreading and very low concentration of solar radiation falling on the earth's surface and variations in solar incident radiation. Solar energy depends on local insolation rates, which vary depending on location, the hour of the day, the day of the month, and the month of the year. These conditions result in an unsteady state diurnal operation. On the other hand, because solar radiation is very diluted, large surface areas are needed to capture the necessary amount of energy. Storage of heat or electricity could be a solution for smooth 24-h operation, but it adds to the capital cost, which in many cases is prohibiting. Thus, solar desalination installations are of small- or medium-size capacity, ranging from a few to 100 m^3/day, and in some cases up to 300 m^3/day (26,420–79,260 gallons per day) for

 © 2004 Elsevier Inc. All rights reserved.

indirect methods. They can supply with fresh water remote, arid, or semiarid regions that have high intensity solar radiation.

1. HISTORICAL BACKGROUND

Solar-driven desalination was developed mainly in the second half of the 20th century, although solar conversion systems have been known since antiquity. It was developed mainly during the energy crisis in the 1970s. Intensive studies led to the construction of compact, small- and medium-capacity conventional desalination plants that could be driven by solar energy. However, solar distillation has been known since antiquity but only as theory or as a very primitive application. Aristotle described the "water cycle" in nature, which is an immense solar distillation procedure, concluding that this procedure may be used to produce sweet water from salt water. He explained, "as the vapors formed by the evaporation from the surface of seas and lakes rise to the atmosphere and condense again, they give sweet water." Later, there were some practical applications and references about the production of drinkable water from sea water using solar radiation. The first practical application of solar distillation was the American patent issued in the 19th century to N. M. Wheeler and W. W. Evans. The patent describes a solar still and is a complete scientific source about production of fresh water by distillation using solar energy. No significant changes have been made to this theory.

The first large solar distillation plant was built in Las Salinas, Chile, to provide fresh water from mine brackish water with 140 g/kg total dissolved solids (TDS) (140,000 ppm). The plant had a surface area of 4450 m^2 (48,000 ft^2), produced 22.7 m^3/day [$\sim$6000 gallons per day (GPD)] of fresh water for the mine and the people working in it, and was in operation for approximately 40 years until the mine was closed.

Until the early 1960s, no other large-scale application of solar distillation was reported. Nevertheless, during World War II distillation gained interest due to the lack of fresh water for troops of Africa and the islands of the Pacific region. In 1945, at the Massachusetts Institute of Technology (MIT), Maria Telkes invented individual inflatable solar stills that saved many lives. After World War II, the MIT team developed various solar still configurations, which have been studied at the office of saline water (OSW) demonstration plant in Daytona Beach, Florida. Some commercial installations have been erected in St. John, Virgin Islands, and Australia. In the 1960s, the Church World Service erected two distillation plants in the Greek Islands, where four additional solar distillation plants have been built that produce fresh water for the local communities..

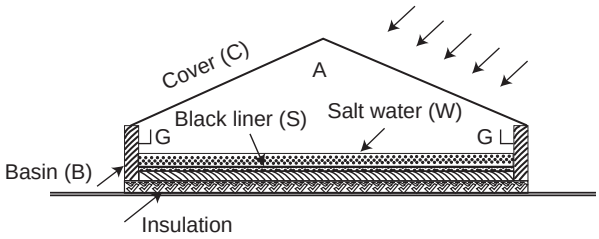


FIGURE 1 The greenhouse symmetrical-type solar still.

2. SOLAR STILLS

Since the 1980s, no large-capacity solar distillation plants have been built, mainly for the following reasons: Large installation areas are needed, solar distillation plants have low efficiencies, and solar-driven desalination has been developed, which in some cases can replace solar distillation (e.g., for medium-size communities).

A solar still is composed of simple equipment, which acts simultaneously as a converter of solar radiation into thermal energy and provides storage for the heat. There are two types of solar stills: single- and multiple-effect stills. For both types, a large variety of configurations have been studied, but most have no practical applications. Studies have focused mainly on increasing output by various means, but this leads to an increase in capital or operation costs. Thus, simple, single-effect stills are the only ones used for large-scale applications. They can be easily constructed from local materials and easily operated without skilled personnel.

There are two types of single-effect solar stills (called greenhouse solar stills): symmetrical (Fig. 1) and asymmetrical (Fig. 4). These types have many configurations depending on the geometry of the still and the type of cover (glass or plastic). Figure 2 show some of the single-effect solar still designs used to construct solar distillation plants.

3. PRINCIPLES OF SOLAR STILLS

A solar still consists of the following:

A well-insulated basin (Fig. 1B), which may be made from metal, plastic, treated wood, concrete, or

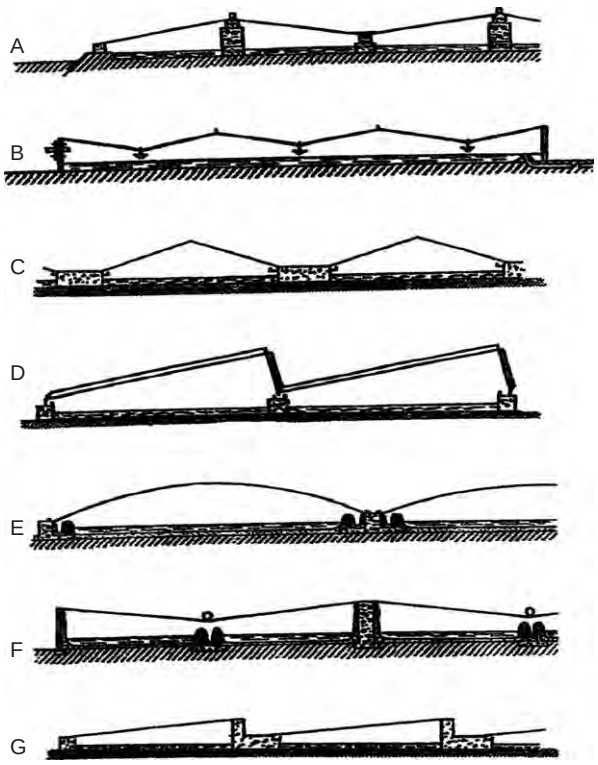


FIGURE 2 Various configurations of single-effect solar tills used to construct solar distillation plants: (A) Battele–Loef glass-covered type; (B) Howe University glass-covered type; (C) CSIRO Australian glass-covered type; (D) Technical University, Athens, glass-covered type; (E) Inflated, tedlar plastic film type; (F) V-shaped plastic cover design; and (G) stretched-inclined plastic cover type.

any material that withstands temperatures of 90–100°C, according to the local availability.

A black sheet (Fig. 1S) that covers the inside of the bottom of the basin to avoid leakage of hot water. The basin can be deep or shallow and contains the water to be heated by solar radiation. The black liner absorbs the incident radiation.

A frame that supports the transparent cover. The frame can be made of metal (e.g., aluminum), treated wood, concrete, or plastic material.

A tilted transparent cover (Fig. 1C) made from either glass or a plastic sheet that withstands ultraviolet radiation and is not easily deteriorated by heat.

Collecting canals, the gathers (Fig. 1G), placed inside the still walls to collect the condensate of fresh water, running along the transparent cover.

Various auxiliaries, such as tubes for loading the still with salt water and for unloading to drain the brine, water storage tanks, and distilled water pumps.

Solar radiation penetrates through the tilted transparent cover into the airtight space (Fig. 1A) and is partially absorbed by the salt water in the basin, which is heated to a temperature higher than that of the cover but lower than boiling point of salt water. The air–vapor mixture leaving the saltwater surface has a higher temperature and lower density than the air–vapor mixture directly beneath the cover. Thus, a convection current is formed and the saturated air–vapor mixture from the water surface moves upward, is cooled by contact with the colder cover, becomes saturated, and is partially condensed, forming a thin film that slides along the inside cover surface by gravity. The uncondensed air–vapor mixture moves back to the water surface. The process takes place in narrow layers between the two surfaces, but the bulk of the air mass does not participate because diffusion and heat conductance are low. The mechanism of the vapor flow in the space takes place by convection and the optimum space is the one that enhances conduction. Therefore, it is advantageous to keep the distance between the water surface and the cover, as well as the angle between the cover and horizontal, as small as possible to achieve good convection (normally, the lowest angle is 15°). It is obvious that the productivity of the still depends on the proper sliding of the condensate on the inner surface of the cover and the temperature difference between the saltwater surface and the cover. The larger the difference, the more intense the circulation of the air–vapor mixture becomes and consequently the vapor condensation. Thus, for a greenhouse-type solar still, it is essential to understand the convection phenomena inside the still. Flow patterns and isotherms are given in Fig. 3 for a single solar still with a 15° angle of inclination and an Ra number of 10^{10}.

4. HEAT AND MASS BALANCE IN THE STILL

Mathematical calculations and models for unsteady state conditions in a solar still are very complicated. There are numerous relations and models for heat and mass transfer coefficients to predict the output of a still—almost as many as still designs. Most of them are empirical or semiempirical and are applied only to the specific design of the still. For the single, greenhouse type, the most common is that by Dunkle. Under the assumption that the temperature of the transparent cover T_c and that of the surroundings T_a are uniform, the thermal network in the basin describing the fluxes in a

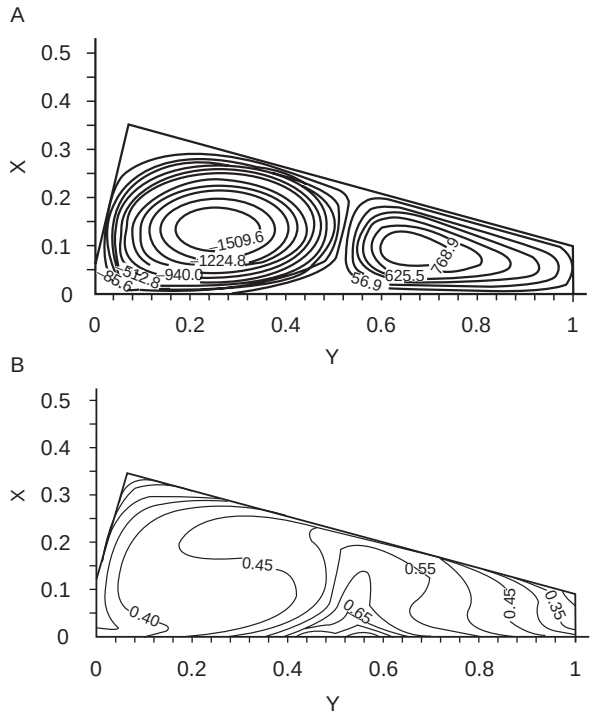


FIGURE 3 Computed streamline (A) and isotherms (B) for an asymmetrical single solar still.

greenhouse-type solar still is presented in Fig. 4. According to the figure, the energy balance for a surface area of the cover A_c and solar radiation intensity I, is

$$\text{Around the cover}: \quad A_g I + q_{rwg} + q_{ewg} = q_{cgs} + q_{rgs} \tag{1}$$

Around the bottom of the still :

$$I\tau_g A_w = q_{rwg} + q_{ewg} + q_b + \frac{C_w \times \partial T_w}{\partial t}, \tag{2}$$

where the last term is the thermal lag for unsteady state operation and is of significance only if the water in the basin is deep. Heat transfer coefficients for radiative heat transfer h_{rwg}, for convective heat transfer h_{cwg}, and for evaporative heat transfer h_{ewg} from the water surface to the cover, as well as radiative heat transfer from the cover to the sky h_{rgs}, for the previous equations are given (in W/m^{2}K) as follows:

For radiation :

$$h_{rwg} = 0.9 \cdot \sigma\left(T_w^2 + T_g^2\right) \cdot \left(T_w + T_g\right) \tag{3}$$

For convection :

$$h_{cwg} = \left[\left(T_w - T_g\right) + \left(\frac{p_w - p_g}{268.9 \times 10^3 - p_w}\right) \cdot T_w\right]^{1/3} \tag{4}$$

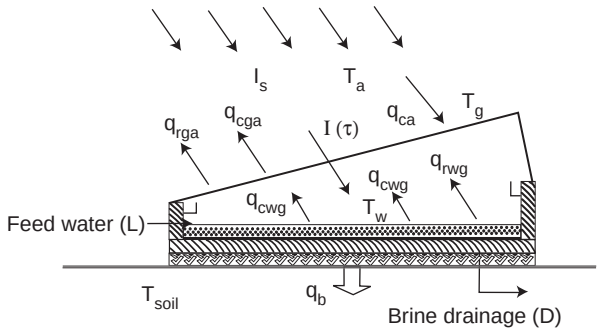


FIGURE 4 Energy balance in a solar still of the asymmetrical type.

For evaporation :

$$h_{cwg} = \frac{9.15 \times 10^{-7} \cdot h_{cwg} \cdot \left(p_w - p_g\right) \cdot \Delta h_v}{\left(T_w - T_g\right)}. \tag{5}$$

The radiative heat transfer between the cover (T_g) and the sky (T_s), which is intensive at night, is given as

$$h_{rgs} = \varepsilon_g \cdot \sigma \cdot \left(T_g^2 + T_s^2\right) \cdot \left(T_g + T_s\right), \tag{6}$$

where Δh_v is the enthalpy of evaporation, and the sky temperature T_s is given as $T_s = 0.0552 \times T_a^{1.5}$. The overall heat transfer coefficient for the upward heat flow for a solar still is given as

$$U_t = \left[(1/U_i) + (1/A_r \cdot U_o)\right]^{-1}, \tag{7}$$

where U_i is the overall heat transfer coefficient between the water surface and the cover given as

$$U_i = h_{rwg} + h_{cwg} + h_{ewg} \tag{8}$$

U_o is the overall heat transfer coefficient between the water surface and the environment given as

$$U_o = \left[h_w \cdot \left(T_g - T_a\right)/\left(T_g - T_s\right)\right] + h_{rgs} \tag{9}$$

and the surface area A_r is the ratio of cover area A_g to the free horizontal surface area of the salt water in the basin A_w:

$$A_r = A_g/A_w. \tag{10}$$

The heat balance within the solar still is thus given by the following equation:

$$\frac{C_w}{A_w} \cdot \frac{dT_w}{dt} = \theta(t) \cdot \eta_0 \cdot I(t) - U_t[T_w(t) - T_a(t)] - U_b[T_w(t) - T_a(t)] \tag{11}$$

5. PREDICTION OF DAILY OUTPUT

Operating efficiencies of solar stills in general are low and do not exceed 35%. This is due mainly to the low incident solar radiation per square meter and to the rejection to the atmosphere of the condensation

enthalpy. Consequently, single-effect solar stills have daily output ranging from approximately 2 in wintertime to 5 or 6 liters/m^{-2}/day in summertime. Double- or multiple-effect stills have efficiencies up to 46%. Many attempts have been made to increase the output of the stills, including

- Coupling the stills to solar collectors to increase feed water temperature.
- Use of a double glass cover (Fig. 5) with running water inside the cover to increase the temperature difference between the water surface and the cover.
- Trickle feeding of salt water by using wicks or using multiple-effect stills to win condensation enthalpy.
- Use of storage incorporated to the still or in connection with solar collectors to extend operation during nighttime and/or cloudy days.
- In places where rainfall occurs during wintertime, rain may be collected by outside canals. Thus, the output of distilled water may be doubled.

Nevertheless, the output by these means is not always sufficiently high to offset additional capital and operating expenses, and the single-effect, greenhouse-type solar stills remain the most simple and low cost for large-scale applications.

The daily output of a solar still is a function of solar radiation intensity, the duration of daily radiation, and heat fluxes within the still, which in turn depend on the design characteristics of the still. Solar radiation for a bright, cloudless day is given by the curve of daily incident radiation. It can be divided into three phases: phase 1, the early morning start of operation; phase 2, the main operation period called the pseudo-steady state; and phase 3, the saturation of the still at the end of the day during sunset. The mean daily output can be calculated from the following equation, which may be applied to many still configurations if parameters f_1–f_3 are known:

$$M_{m-d} = f_{1-d} \cdot H_d + f_{2-d} \cdot (\bar{T}_{w-d} - \bar{T}_{a-d}) + f_{3-d}, \quad (12)$$

where H_d is the daily solar radiation (MJ/day); T_w and T_a are the daily mean temperatures of salt water and the environment, respectively; and f_1–f_3 are characteristic parameters of the still. f_3 concerns the energy state of the still during phase 1 and represents the inertia of the still. Parameters f are functions of the characteristics of the still and of the heat fluxes. They are given as

$$f_1 = \overline{\theta} \cdot n_0 \cdot \frac{A_g \cdot A_w}{C_w} \cdot \frac{h_{ewg}}{\Delta h_v} \cdot \frac{U_t}{U_i} \cdot \Delta t \quad (13)$$

$$f_2 = (U_t - U_b) \cdot \frac{A_g \cdot A_w}{C_w} \cdot \frac{h_{ewg}}{\Delta h_v} \cdot \frac{U_t}{U_i} \cdot (\Delta t)^2 \quad (14)$$

$$f_3 = A_g \cdot \frac{h_{ewg}}{\Delta h_v} \cdot \frac{U_t}{U_i} \cdot (\Delta t)_I \cdot (T_{w-in} - T_{a-in}), \quad (15)$$

where θ is the mean radiation incident inclination angle.

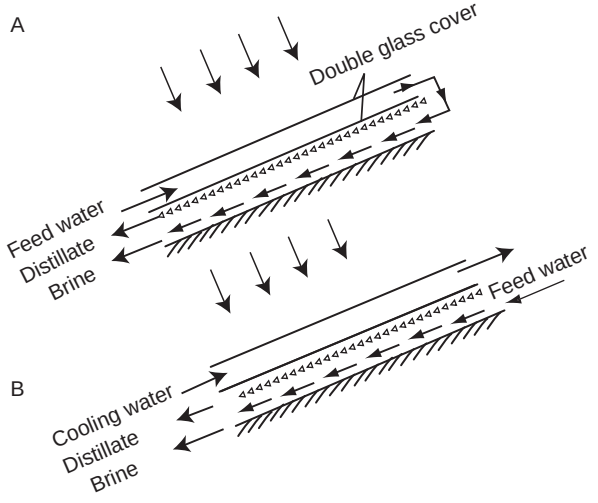


FIGURE 5 Double glass cover solar stills using (A) feedback flow and (B) counter flow.

6. MATERIALS OF CONSTRUCTION

To achieve long life and good operation conditions, the construction materials are important. Good maintenance is also important. Thus, the material used must be of good quality, low cost, and, if possible, available in local markets. Small poor communities cannot afford the high capital cost of special materials and long-distance transport expenses. Local inexpensive materials were used to construct solar distillation plants in the Caribbean Islands and India. Thus, from an economic standpoint, solar distillation plants are not comparable from place to place because cost varies according to the materials used. Nevertheless, there are some basic requirements for achieving a long life of operation and the best possible daily output:

- The parts of the still must be precut and easily assembled at the site.
- Maintenance and proper operation must be supervised continuously to achieve an effective lifetime of the plant of approximately 15–20 years.

• The system must withstand severe weather conditions, such as hale, hurricanes, and sand and wind storms. Thus, it has to be well anchored in the soil or pavement.
• Construction materials that come in contact with fresh or salt water must not be toxic and in general not contaminate the water produced.
• The still parts must be easily bound in compact packages in case transportation to the site is necessary.

The parts of the still must have the following characteristics:

1. Transparent cover, usually made from glass and, less often, from plastic sheet material:
   - The material must withstand wind, sunlight, rain, dust, hail, and/or the wind. Plastic material may be easily deteriorated by sunlight or damaged by hale or wind.
   - It must have high transmissibility (more than 90%), low absorptivity, and low reflectivity.
   - The properties of the material must not alter with time and exposure to sunlight and environmental effects and must withstand temperatures up to 90°C.
   - Especially plastic material should not have electrostatic properties. Electrostatic properties facilitate the accumulation of dust particles, thus decreasing solar radiation absorption inside the still.
2. Solar still frame and basin material, which in most cases are made from the same material. Concrete, wood impregnated with a protective coating, hard plastic material, or any metallic material (e.g., aluminum–magnesium alloy) that is not corroded by distilled or seawater may be used.
   - The material must withstand corrosion of distilled and seawater, inside and outside of the still.
   - The frames must be anchored, especially in places with strong winds.
   - They have to be easily removed or permit access for cleaning and maintenance.
3. The liner of the basin: The black sheet liner acts as an absorber surface for the solar radiation. A variety of materials have been used, including black paints resistant to seawater, butyl rubber sheets, and jute impregnated with tar. In general, the liner must have the following properties:
   - It must have absorptivity >0.95 and be impervious to water.
   - The sheet surface must be smooth. Rough surfaces facilitate scale deposition and are more difficult to clean.
   - It must not deteriorate by solar radiation and must not release toxic substances.

Some auxiliary materials are used, such as pumps, piping systems for drainage and for freshwater transportation, freshwater reservoirs, and sealants. These materials are available commercially. They must be resistant to fresh and seawater and must be cleaned often to avoid algae formation. Formation of algae is dangerous because it results in decreased efficiency and may contaminate fresh water. In stagnant conditions, even inside the still algae are deposited.

7. MULTISTAGE SOLAR STILLS

Recovery and utilization of condensation enthalpy are achieved in multiple-effect solar distillation systems. There are a variety of configurations for double- and multiple-effect solar stills. None have found practical application in large-scale solar distillation plants. Figure 6 demonstrates a two-effect still, and Fig. 7 shows the multieffect wick. A special type is the multieffect diffusion still. In this type, the gap between the cover and saltwater surface is reduced and evaporation occurs only by molecular

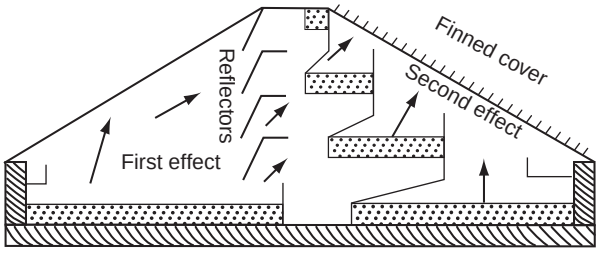


FIGURE 6 A two-effect solar desalination still.

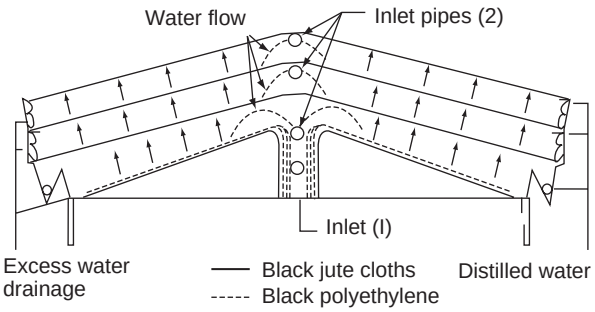


FIGURE 7 Multiple-effect wick solar still.

diffusion and not by natural convection. To determine whether to use a single- or a multieffect still, one must calculate the increase in output and the additional costs in construction and operation.

8. SOLAR-DRIVEN DESALINATION

For medium-size communities, water demand may be for households, agriculture needs, and small industries. These regions may also be remote without energy supply. In this case the solution is solar-driven desalination plant. Capacities range from 10 to 100 m^3/day (2642–26,420 GPD).

Solar-driven desalination plants combine two different technologies: thermal or electrical techniques and conventional desalination methods. Desalination methods are considered mature and commercially available, with a variety of procedures, equipment, and capacities. Solar thermal and electrical transformation methods are more or less mature and in some cases commercially available. In comparison to solar distillation plants, both are capital- and operation-intensive enterprises, but costs decrease as technology improves. In general, they are more complicated to operate than plants driven by conventional energy sources. This is due mainly to the unsteady-state solar incident radiation.

9. ENERGY CONVERSION SYSTEMS

Energy conversion systems must be chosen according to the morphology of the site, the desalination method, the availability of salty water, and the freshwater output needed. Wind energy may be included as an energy transformation source because it is an indirect solar energy source. Energy conversion systems include

- Flat-plate collectors, which supply hot water up to $\sim 95^{\circ}$C.
- Solar ponds, which produce hot water up to 90°C.
- Evacuated tube collectors, which reach temperatures up to 200°C and are more suitable for conventional distillation plants, such as multiple-effect thermal distillation (MED) and thermal vapor compression (TVC). They have been proven to be very effective in combination with MED.
- Focusing collectors, which produce high-temperature vapor and/or electricity. They are suitable for large-scale dual-desalination plants and for electricity and freshwater production. Thus, small remote regions are excluded.

Flat-plate collectors and solar ponds require large installation areas (either flat or cascading for flat-plate collectors), but if available, capital and operation costs are low. Not all solar conversion systems are suitable for each desalination method. The combined technologies have to be carefully designed to give the highest possible efficiency for the morphology and environment of the site. Flat-plate collectors, solar ponds, and evacuated tubes are coupled with MED, TVC distillation equipment, focusing collectors, and evacuated tubes for multi-stage-flash thermal distillation plants.

10. CONVERSION TO ELECTRICITY

Solar energy conversion to electricity includes wind energy, which is an indirect solar energy source. The conversion systems comprise photovoltaic devices that directly transform solar energy into electricity and wind turbines that indirectly produce electricity. The electrical energy produced is used to operate seawater and/or brackish water reverse osmosis units, electrodialysis and electrodialysis reversal units, and mechanical vapor compression distillation units. Electrical energy is also used to run auxiliary equipment, such as pumps, fans, and control devices, and to provide the site with electricity. In the case of thermal conversion systems, small photovoltaic installation may help provide the necessary electricity to the system.

Solar-driven desalination plants are unsteady-state systems. Solar energy can be stored as: sensible heat in water tanks, in phase change material, and in batteries as electrical energy to smooth diurnal operation during night and/or cloudy days.

11. SIZING A SOLAR-DRIVEN SYSTEM

For remote and arid locations, solar energy and desalination technology systems have to be as simple as possible, controlled and maintained by skilled personnel. This facilitates good operation conditions and minimal maintenance. Generally, an overcapacity design is recommended for desalination plants and freshwater storage tanks. For small- and medium-size plants, a water storage tank three to six times the daily water consumption and an overcapacity design of approximately 50% are sufficient. Conventional desalination plants are energy-intensive enterprises.

TABLE I

Indicative Typical Energy Inputs for Solar Desalination[a]

Process	PR (kg/2326 kJ)	kJ/kg of fresh water	Production rate (liters/m^2)	Form of energy	Suitable energy technology
Stills	0.53	4259	4	50–80°C	Ponds, flat-plate
MED	8	294–296	8	~71°C	Ponds, flat-plate
	15	159–161	13	~120°C	Evacuation-concentrating
MSF	8	298–300	9.5	~90°C	Ponds, evacuated
	12	199–201	14	~120°C	Concentrating
MVC	58	61	50	17 kWh m^{-3}	Solar electric, PV
SWRO	54	43	100	12 kWh m^{-3}	PV, wind
BWRO	200	12	75	3 kWh m^{-3}	PV, wind
BWED	200	12	30	3 kWh m^{-3}	PV, wind

[a] Abbreviations used: MED, multiple-effect distillation; MSF, multistage flash distillation; MVC, mechanical vapor compression distillation; SWRO, sea water reverse osmosis; BWRO, brackish water reverse osmosis; BWED, brackish water electrodialysis; PV, photovoltaic.

TABLE II

Solar Drive Desalination Plants Worldwide[a]

Plant location	Desalination process	m^3/day	Solar energy conversion plant
Abu-Dhabi, UAE	MEB, 18 effects	120	Evacuated tubes
Al-Ain, UAE	MED, 55 effects, MSF, 75 stages	500	Parabolic trough
Borkum Island, Germany	MVC	5–12.5 m^3/h	Wind, 300 kW
Cabo Verde	MSF, auto flash	300	Solar pond
El Paso, Texas	MSF	19	Solar pond
Gran Canaria, Spain	MSF	10	
Hzag area, Tunisia		0.1–0.35	Flat-plate collectors
La Desire Island, French Caribbean	MED, 14 effects	40	Evacuated tube
Lampedusa Island, Italy	RO	16	100 kWp
Margerita di Savoya, Italy	MSF	50–60	Solar ponds
Platforma Solar, Almeria, Spain	MED, 14 effects	72	Parabolic trough
Safat, Kuwait	MSF	10	Solar collectors
Takami Island, Japan	MED, 16 effects	16	Flat-plate collectors
University of Ancona, Italy	MED	30	Solar pond

[a] Abbreviations used: MEB, multiple-effect boiling; MED, multiple-effect distillation; MSF, multistage flash distillation; MVC, mechanical vapor compression distillation; RO, reverse osmosis.

Thus, energy needs are the most important factor in determining the size of a solar conversion system. For multieffect stack (MES) evaporators, which are the most preferable to combine with high-performance flat-plate and evacuated tube collectors, thermal energy requirements depend on performance and capacity:

$$Q_{ev} = M_d \cdot \frac{\Delta h_v}{R}, \qquad (16)$$

where M_d is the daily rated capacity, and R is the performance ration of the MES unit, which depends on the number of effects and is given as

$$R = -0809 + 0.93N - 0.0091N^2. \qquad (17)$$

In addition, electrical energy is provided by either diesel or photovoltaic arrays to drive the pumps. Table I gives prices of energy

consumption for solar distillation and solar-driven systems.

12. CONCLUSION

Solar distillation systems are suitable for remote, arid, or semiarid regions for small capacities. Simple designs are preferred because they are cost-effective and easier to operate and maintain. Low-cost, locally available materials can be used, taking into consideration safety and lifetime of the installation. Although some solar distillation plants with capacities of 1–40 m^3/day (in Greece, India, the Caribbean islands, and other places) are in operation, the interest currently has turned to solar-driven systems. Solar-driven desalination is suitable for small- to medium-size capacities. As these systems combine, different technologies must be carefully selected and coupled to achieve high efficiencies, simple operation, and easy maintenance. Most plants are pilot plants used to study and improve design, operation, and cost. They do not exceed 0.1% of the capacity of desalination plants driven by conventional energy sources and require further study to improve technologies and lower cost. Table II lists some of the most important solar-driven desalination plants.

SEE ALSO THE FOLLOWING ARTICLES

Desalination and Energy Use • Solar Cooling, Dehumidification, and Air-Conditioning • Solar Detoxification and Disinfection • Solar Distillation and Drying • Solar Thermal Energy, Industrial Heat Applications • Solar Thermal Power Generation

Further Reading

Anonymous. (1965). Solar desalting gain acceptance in Greece. *Chem. Eng.* **72**(26), 42–43.

Belessiotis, V., Voropoulos, K., and Delyannis, E. (1996). Experimental and theoretical method for determination of daily output of a solar still: Input–output method. *Desalination* **100,** 99–104.

Bloemer, J. W., Collins, J. A., and Eibling, H. A. (1961, 1965, 1966). Solar distillation of saline water, Research and Development Progress Reports 50, 147, and 190. OSW, Washington, DC.

Damm, W., and Geiser, P., *et al.* (1996). Wind powered MVC seawater desalination—Operating results and economic aspects. *In* "Proceedings of the Mediterranean Conference on Renewable Energy Sources for Water Production, June 10–17, Santorini," pp. 143–146. Centre for Renewable Energy Sources, Athens, Greece.

Delyannis, E., and Belessiotis, V. (1995a). Solar desalination: Is it effective? Part I. *Int. Desalination Water Reuse Q.* **4**(4), 9–14.

Delyannis, E., and Belessiotis, V. (1995b). Solar application in desalination: The Greek islands experiment. *Desalination* **100,** 27–34.

Howe, E. D., and Tleimat, B. D. (1967). Solar distiller to use in Coral islands. *Desalination* **2**(1), 109–115.

Delyannis, E., and Belessiotis, V. (1996). Solar desalination: Is it effective? Part II. *Int. Desalination Water Reuse Q.* **4**(4), 28–34.

Delyannis, E., and Belessiotis, V. (2001). Solar energy and desalination. *Adv. Solar Energy* **14,** 287–330.

Djebendjian, B., and Rayan, M. A. (2000). Theoretical investigation of the performance prediction of solar still: 4th International Water Technology Conference, Alexandria, VA, 1999. *Desalination* **128**(2), 139–145.

Duffie, J. A., and Beckman, W. A. (1991). "Solar Energy Thermal Processes." Wiley, New York.

Dunkle, R. V. (1961). Solar water distillation—The roof type still and a multieffect diffusion still. *Int. Development in Heat Transfer, ASME Proc. Int. Heat Transfer,* ***Part V****, Univ. Colorado*, 895–902, 1961.

Eckstrom, R. M. (1965). Design and construction of the Symi still. *Sun at Work* **10**(1), 7.

El-Nashar, A. M. (2000). Economics of small solar assisted multiple-effect stack distillation plants. *Desalination* **130,** 200–215.

El-Nashar, A. M., and Samad, M. (1996). A solar assisted seawater multi-effect distillation plant. 10 years of operating performance. *In* "Proceedings of the Mediterranean Conference on Renewable Energy Sources for Water Production, June 10–17, Santorini," pp. 62–72. Centre for Renewable Energy Sources, Athens, Greece.

Fath, H. E. S. (1996). High performance of a simple design, two effect solar distillation unit. *Desalination* **107,** 223–233.

Garcia-Rodrigez, L. (2002). Seawater desalination driven by renewable energies—A review. *Desalination* **143,** 103–113.

Glueckstern, P. (1999). Design and operation of medium and small-size desalination plants in remote areas. *Desalination* **122,** 123–140.

Goosen, M. F. A., Sablani, S. S., *et al.* (2000). Thermodynamic and economic considerations in solar desalination. *Desalination* **129,** 63–89.

Goswami, D. Y., Kreith, F., and Kreider, J. F. (2000). "Principles of Solar Engineering." 2nd ed. Taylor & Francis, Philadelphia.

Harding, J. (1883). Apparatus for solar distillation. *Proc. Inst. Civil Engineers London* **73,** 284–288.

Kreith, F., and Kreider, J. F. (1978). "Principles of Solar Engineering." Hemisphere, New York.

Morse, R. N., and Read, W. R. W. (1966). The development of a solar still for Australian conditions. *In* "Proceedings of the Conference on Power Production and Energy Conversion." pp. 39–40. Institution of Engineers, Melbourne, Australia.

Nebbia, G., and Nebbia-Menozzi, G. (1966). A short history of water desalination. *In* "Proceedings of the Aqua Dolce dal Mare, II Inchiesta Internationale, April 1966, Milano." pp. 129–172. Consiglio Nationale delle Richerce, Milano, Italy.

Papanicolau, E., Voropoulos, K., and Belessiotis V. (2002). Natural convection in a solar still. Accepted for publication.

Robinson, N. (1955). Solar machines. *In* "Proceedings of the World Symposium on Applied Solar Energy, Phoenix, Arizona, November, 1–5," pp. 43–61. Association of Applied Solar Energy, AZ, USA.

Sardi, L. (1996). RO desalinators powered by PV systems for small/medium Italian islands. *In* "Proceedings of the Mediterranean Conference on Renewable Energy Sources for Water Production, June 10–17, Santorini," pp. 36–44. Centre for Renewable Energy Sources, Athens, Greece.

Talbert, S. G., Eibling, J. A., and Loef, G. O. G. (1970). "Solar Distillation of Saline Water Manual." Battele Memorial Institute, Columbus, OH.

Telkes, M. (1945). Solar distiller for life rafts. U.S. Office of Technical Services Report 5225, MIT, OSRD Final Report to the National Defense Research Commission, 11.2, pp. 24–25. U.S. Office of Technical Services, Washington, DC.

Telkes, M. (1955). Solar stills. *In* "Proceedings of the World Symposium on Applied Solar Energy, Phoenix, Arizona, November, 1–5," pp. 73–79. Association of Applied Solar Energy, AZ, USA.

Steel Production and Energy

MATTHIAS RUTH
University of Maryland
College Park, Maryland, United States

1. **Background**
2. **Iron and Steel Production**
3. **Casting, Forming, and Finishing**
4. **Ancillary Processes**
5. **Drivers for Change of Energy Use in Iron and Steel Making**
6. **Conclusions**

Glossary

calcination A process of oxidation (here of lime in iron making).
charge Load of a furnace with the usual materials.
coke rate Amount of coke used per ton of iron or steel product.
energy efficiency Amount of output produced per unit of energy input.
energy intensity Amount of energy used per ton of product.
oxidation Chemical process by which oxygen is added to a reactant.
reduction Chemical process by which oxygen is removed from a reactant.

Iron and steel-based products lay the footprint for industrialization, and thus energy consumption in the transportation, commercial, industrial, and end-use sectors of an economy. Their manufacturing itself is highly energy intensive and often accounts for a significant share of a country's energy use. This article concentrates on the uses of energy for steel production.

1. BACKGROUND

Since the construction of the great Iron Bridge in England during the 1770s, the use of iron and steel-based products has come to be associated with the industrialization of economies. Iron and steel are used in such diverse applications as railroads and skyscrapers, automobiles and other durable consumer products, as well as many infrastructure systems of modern society. To this day, iron and steel are among the least expensive and most widely used metals.

As economies industrialized and grew in size, world production of steel has increased (Fig. 1). By early 1900, the United States led the world in steel production, contributing about 36% of the world's output, followed by the German Empire (23%), the United Kingdom (17%), and Russia (10%). In 2000, the three largest steel producers were China (15% of world output), Japan (13%), and the United States (12%). With the boom in its economy, China's demand for finished steel increased more than 40-fold from the early 1950s to the mid-1990s, and its production rose by more than 10% per year. However, production increases in China were less rapid than the growth of the Japanese steel industry during that time period or the Korean steel industry since the mid-1970s.

This article explores the uses of energy for steel production. The first section describes key processes for raw steel production. The second section addresses iron and steel casting, finishing and forming processes used to produce semifinished or finished products. The third section describes processes that are not always carried out by the iron and steel industry itself, such as beneficiation of ores, but that significantly influence the industry's energy use profiles. The fourth section investigates various drivers behind technology change in the industry and their influences on energy use. The chapter closes with a brief summary and conclusions.

2. IRON AND STEEL PRODUCTION

Iron is made from ores by removing oxygen and impurities from the iron oxides in ores. Chemical reduction is achieved with the help of coke, lime, or

 © 2004 Elsevier Inc. All rights reserved.

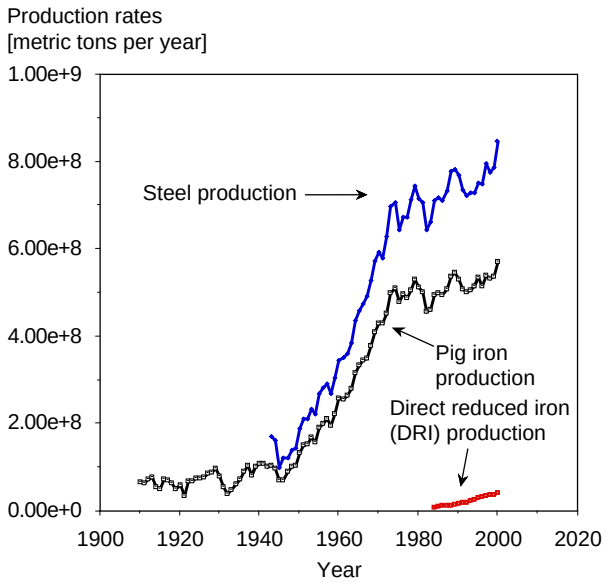


FIGURE 1 World production rates. Data from Kelly, T. D., and Fenton, M. D. (2002). Iron and steel statistics. U.S. Geological Survey, http://minerals.usgs.gov/minerals/pubs/of01-006/ironandsteelslag.html.

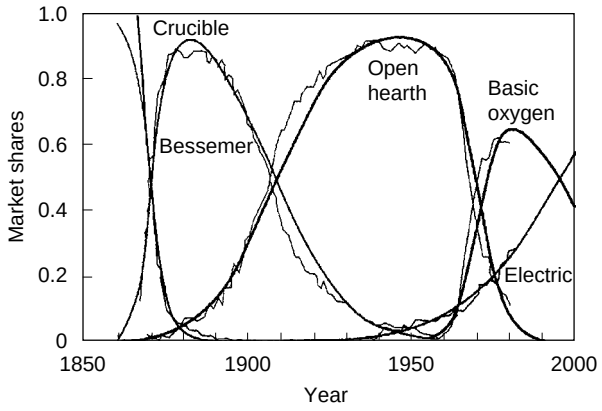


FIGURE 2 Process technology change in U.S. steel manufacturing, in fractional shares of raw steel tonnage produced. From Nakicenovic, N. (1990), Technological substitutions and long waves in the USA. In "Life Cycles and Long Waves, Lecture Notes in Economics and Mathematical Systems" (T. Vasko, R. U. Ayres, and L. Fontvieille, Eds.), No. 340, Springer-Verlag, Berlin. © Copyright Springer-Verlag. Permission granted.

other reduction agents. The molten iron typically contains from 3 to 5% carbon and other impurities. It is usually produced with blast furnaces and passed on in a molten state as input in steel-making processes.

Steel making involves reducing carbon content to about 1%. Various steel-making processes exist. The crucible, bessemer, open hearth, and basic oxygen processes start with iron as their primary material input to which in some cases iron and steel scraps are added. In contrast, the electric arc process starts with up to 100% scrap as its material input and uses electricity to melt the furnace charge. Through history, each of these processes accounted for varying contributions to overall iron and steel production (Fig. 2). By early 1900 the crucible, bessemer, and open hearth furnaces were used in the United States for 67%, 32%, and 1% of steel production, respectively. By 2000, these furnaces were replaced in the U.S. steel-making process by the basic oxygen and electric arc furnaces, which produced 53% and 46% of steel, respectively.

Basic oxygen furnaces are typically part of vertically integrated operations that produce ores, a variety of intermediate products such as coke and pig iron, and raw and finished steels. While basic oxygen furnaces were replaced gradually in the United States and many other countries as the technology matured, furnaces aged, and new technologies became available, Japanese and European competitors had to rebuild most of their production facilities after the war. Only a small fraction of the furnaces they lost were replaced by the traditional open hearth technology. Most new capacity was added in the form of basic oxygen furnaces, which reduced the time it took to produce steel from longer than 6 hours to less than 40 minutes.

No new basic oxygen furnace capacity has been added since the 1960s in vertically integrated mills in the United States, capacity fell by over 50%, and no new future capacity additions are expected, in part because of large capital requirements to build furnaces that must operate at 1.5 to 3 million tons per year in order to be economically viable. Concomitant with these trends, the share of output from electric arc furnaces has steadily increased since the 1970s (Fig. 2), and the energy use profiles of the industry as a whole have changed.

Each iron and steel-making process has characteristic energy use profiles. However, in the thermodynamic ideal only iron making requires substantial energy consumption. The theoretical minimum energy requirement to reduce ore is given by the physical and chemical properties of ore. At room temperature, the theoretical minimum energy requirement to melt Hermatite, the main ore type, is 6.9 GJ per ton of steel. This thermodynamic minimum does not change with time and establishes a lower bound on the energy necessary for steel production from ore. In practice, energy consumption for primary steel production is two to three

times the thermodynamic minimum. Consequently, energy use by the industry could, in principle, be reduced significantly.

2.1 Blast Furnace Production

Blast furnaces are the largest consumers of materials and energy in the iron and steel-making process. They are relatively flexible in their ability to use different metal charges, such as pellets, sinter, or scrap, with little change in performance. Other inputs include coke to produce carbon monoxide for the reduction of iron ore to iron, and other energy sources for various stages of the production process, such as preheating of air up to 1100°C for injection into blast furnaces. Since blast furnaces require significant quantities of coke, they are often operated in close conjunction with coke ovens, creating opportunities for joint use of byproducts, such as coke oven gas as well (Fig. 3).

The main outputs of blast furnaces includes pig iron and slag, which is formed by combination of limestone with sulfur and other impurities. Pig iron is typically tapped every 3 to 5 hours in quantities of 300 to 600 tons. After tapping, the pig iron is transported, typically in liquid form, to steel-making operations.

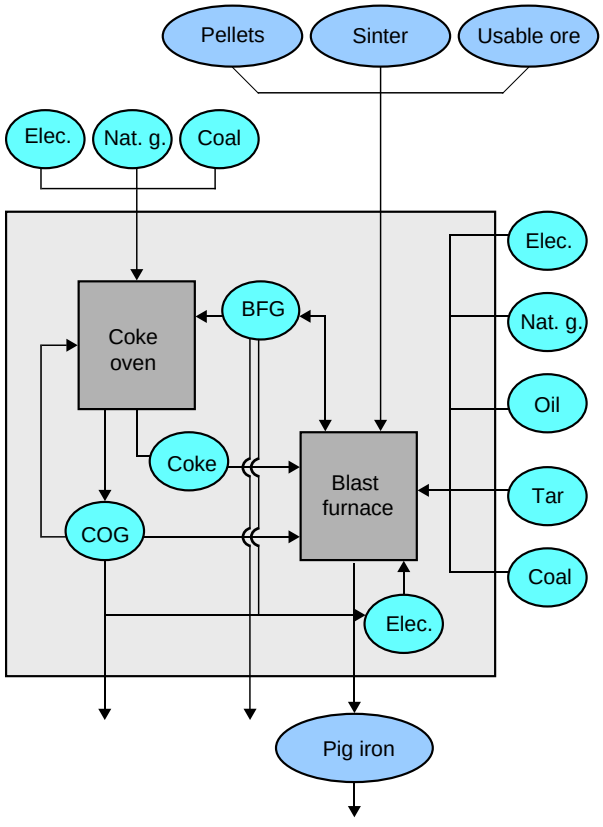


FIGURE 3 Main material and energy flows of blast furnace and coking operations. BFG = Blast furnace gas, COG, coke oven gas; Nat. G., natural gas; Elec., electricity.

Other outputs from blast furnaces include blast furnace gas, an off-gas with a heating value of 3 MJ per m^3, which is approximately one-tenth of the energy content of natural gas. This heating value has been dropping as a result of reduced coke rates and better burden preparation. The gas quality is insufficient to allow high-efficiency conversion to electricity in gas turbines. To increase heating values, blast furnace gas is often mixed with other gases such as coke oven gas and natural gas, which have higher energy content. This mixture can be converted into electricity with efficiencies of 41 to 42% in combined cycle plants. Much of that conversion takes place onsite at steel plants for electricity used in steel making and in some instances for sale into the electricity grid. A portion of the off-gas is consumed in boilers that produce compressed air via steam-powered or electric blowers. Funneling that air through hot blast stoves provides the heat required in the blast.

With the price of coke rising, steel makers have sought to reduce their dependence on coke, for example, by installing pulverized coal injection into their blast furnaces as a partial substitute for coke. A few use natural gas for the same purpose. The need for greater efficiency in blast furnaces has also focused attention on the efficiency of oxygen use so that combining coal injection and large amounts of oxygen can further diminish the need for coke.

2.2 Crucible Steel Production

The crucible process for steel making was introduced in the 1740s in England. It used iron, frequently in solid form, which was molten with the aid of coke fires (Fig. 4). Melting took place in clay crucibles to which other materials were added to achieve desired properties of the steel they produced. Production was typically small-scale (approximately 50 kg per batch) and took roughly 4 to 6 hours per batch. Both the small batch sizes and relatively long heating times meant that the crucible process was highly energy inefficient and did not yield steel at rates necessary to meet rapidly growing demand.

2.3 Bessemer Steel Production

The bessemer process reduces molten pig iron in so-called bessemer converters—egg-shaped, silica, clay,

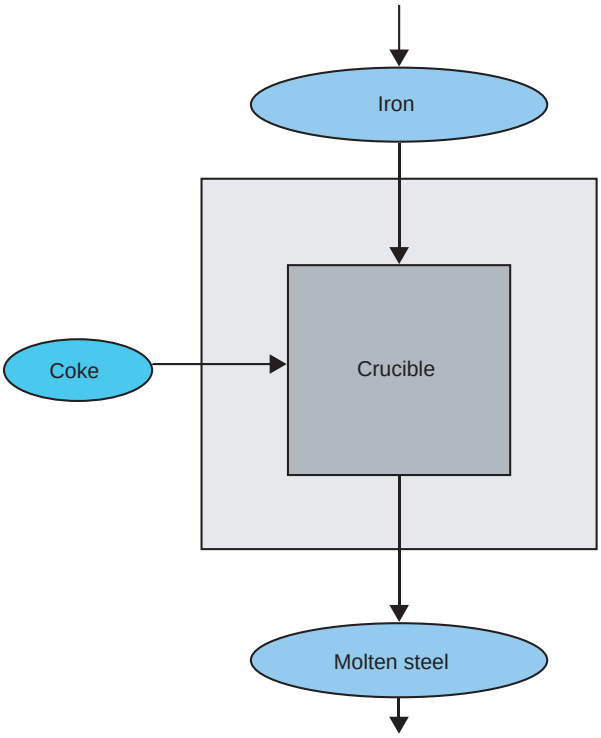


FIGURE 4 Main material and energy flows of crucible steel production.

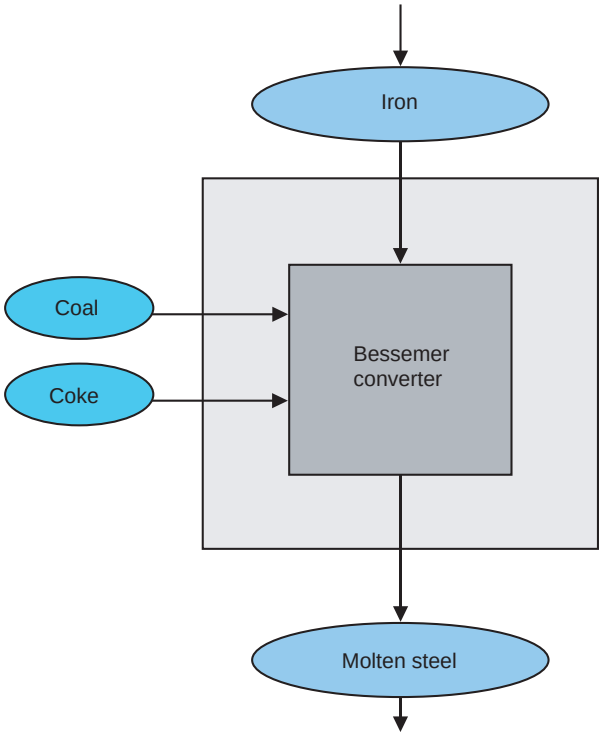


FIGURE 5 Main material and energy flows of bessemer steel production.

or dolomite-lined containers with capacities of 5 to 30 tons of molten iron. An opening at the narrow upper portion of the bessemer converter allows iron to be introduced and the finished product to be poured out. Melting of metal typically is accomplished with coal and coke fires (Fig. 5). Air is forced upward through perforations at the wide bottom-end of the converter. As the air passes upward through the molten pig iron, impurities such as silicon, manganese, and carbon oxidize. Different property steels are produced by introducing additional elements, such as spiegeleisen—an iron-carbon-manganese alloy—into the molten metal after the oxidation is completed. The converter is then emptied into ladles from which the steel is poured into molds. Since the whole process could be completed typically in well under 1 hour and for significantly larger quantities than with crucibles, bessemer steel production significantly helped meet steel demands during the onset of the industrial revolution. As steel production increased, demand for coal and coke increased in lockstep.

2.4 Open Hearth Production

Open hearth furnaces are first charged with scrap and a small amount of limestone. Limestone is used as a reduction agent. Then the fuel—typically oil, natural gas, coal, tar and pitch, or a combination of these—is ignited to melt the charge (Fig. 6). As the charge melts, proportioned amounts of molten pig iron are added to the furnace and high-purity oxygen is blown in. After various minor operations, the molten steel is trapped for approximately 8 hours to further chemically reduce the iron.

Between 1910 and 1970, open hearth furnaces were the predominant technology for steel production in much of the industrialized world. Their flexibility in using a wide range of hot metal and scrap charges added versatility to the production process. However, their long cycle times also meant significant energy use per ton of steel produced. With the onset of the energy crises of the 1970s and 1980s, the newer, more efficient technology of basic oxygen production became increasingly economically viable.

2.5 Basic Oxygen Furnace Production

Basic oxygen furnaces use pig iron and small amounts of scrap as their primary material inputs and typically melt the furnace charge with coke oven, blast furnace, and natural gas. Electricity and fuel oils are used to drive ancillary processes (Fig. 7).

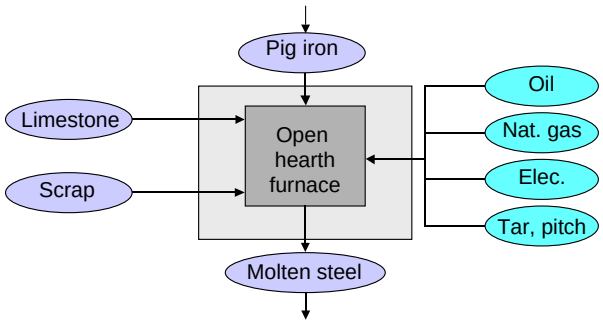


FIGURE 6 Main material and energy flows of open hearth steel production. Nat. G., natural gas; Elec., electricity.

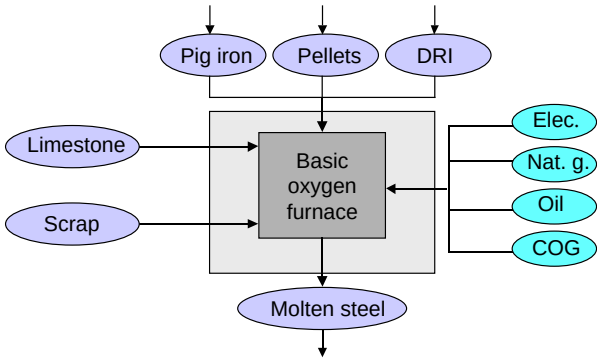


FIGURE 7 Main material and energy flows of basic oxygen production. DRI, direct reduced iron; COG, coke oven gas; Nat. G., natural gas; Elec., electricity.

Basic oxygen furnaces have the ability to cut total cycle time to less than a tenth of the cycle time of open hearth furnaces, thus enabling significant savings of energy for a given metal mix, and substantial total cost savings even though material costs themselves may be higher. However, their reduced flexibility in use of scraps, compared to the flexibility of basic oxygen furnaces, meant temporary scrap surpluses in the economy, reduced scrap prices, and opportunities for growth of yet another new technology—electric arc furnaces. At first, electric arc furnaces were predominantly used in integrated plants as the "swing" facility when scrap was scarce or plentiful. To a significant extent, this combination of versatile electric arc and efficient basic oxygen furnaces soon spelled the end for the open hearth as a major steel making technology.

2.6 Electric Arc Furnace Production

As output from integrated steel producers declined and electric arc technology improved, smaller-scale

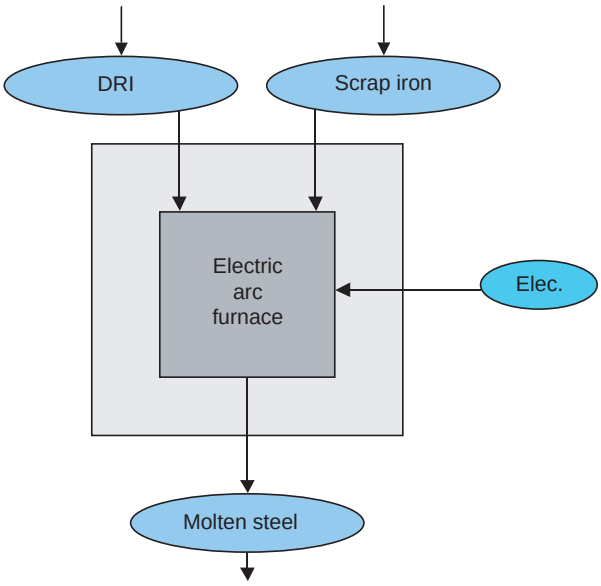


FIGURE 8 Main material and energy flows of electric arc furnace production. DRI, direct reduced iron; Elec., electricity.

electric arc furnaces quickly became an economically attractive alternative to integrated mills—compared, for example, to blast furnaces, which cost around $1 billion and usually operate at scales between 1.5 and 3 million tons per year, as noted by Crandall in 1996.

Electric arc furnaces melt metal by passing a large electric current through the charge (Fig. 8). Since it takes approximately one-third to one-half of the energy to produce steel from scrap than from ore (because it takes considerably more energy to reduce iron oxides than simply melt scrap), energy savings and onsite emissions reductions are substantial when steel made from scrap is substituted for steel made from iron ore. Their relatively smaller scale, predominant use of scrap, and cleaner production enabled electric arc furnaces to be located in large urban areas and steel fabrication centers where the supply of steel scrap is high.

However, several limitations exist to reduce energy use in steel production by expanding electric arc furnace output. First, an increase of the amount of scrap that is recycled without some decrease in the amount of hot metal production will increase, not decrease, energy use. To reduce overall energy consumption for steel production requires substitution of either scrap for hot metal in existing furnaces, or of new, scrap-based furnaces for ore-based furnaces, not just simply an increase in scrap recycling. Second, actual energy savings depend on (1) how much energy is required to collect, sort, process, and deliver scrap and (2) how much energy

is consumed by hot metal production. Considering both the direct and indirect energy requirements, some studies have found that even if all the steel were made from scrap, the total energy intensity of U.S. steel production would be reduced by only 6%, saving less than 1% of total annual U.S. energy use. However, increases in scrap use may be pursued for other reasons, such as relative rises in labor costs and relatively high ore taxes. Other constraints on increasing the market share of electric arc furnace output depend on environmental cost, scrap availability and prices, and product quality.

3. CASTING, FORMING, AND FINISHING

A series of casting, forming, and finishing processes are used to change the shape and other physical and chemical characteristics of raw steel products. Each of these requires a specific set of steps with distinct efficiencies and energy use profiles. The general pattern of energy use of casting, forming, and finishing is shown in Fig. 9. Details are provided next.

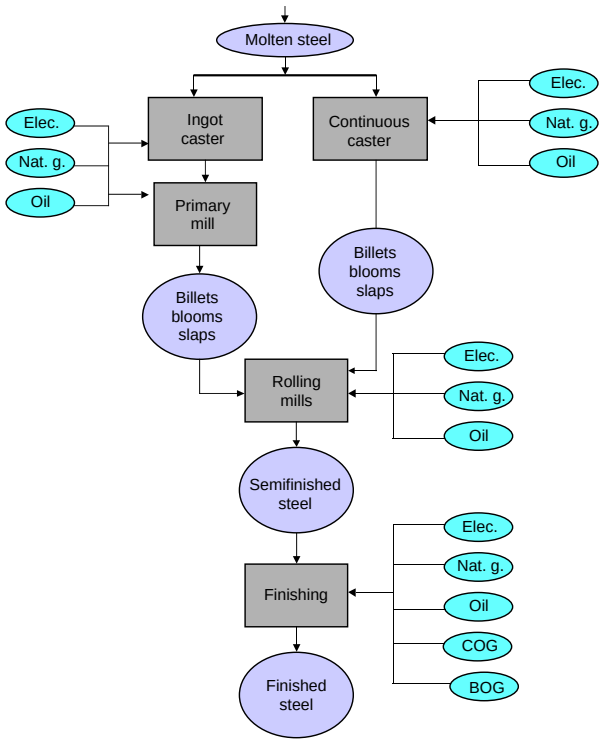


FIGURE 9 Main material and energy flows for pellet and sinter production. Nat. G., natural gas; Elec., electricity.

3.1 Casting

Most of the raw steel produced by the industry is cast either continuously or into ingots or strips. In ingot casting, which in the United States and other industrialized countries is mostly used for small batch sizes for specialty steels or for end products with certain shape specifications, the molten steel is teemed into a series of molds and allowed to solidify to form ingots. After the molds are stripped away, the ingots are heated to uniform temperature in soaking pits to prepare them for rolling. Subsequently, heated ingots are removed from the pits and rolled into slabs, blooms, or billets.

In continuous casting, the molten steel is delivered in ladles and poured into a reservoir from which it is released into a mold. The steel cools as it passes through the mold and forms a solid outer skin. As the steel proceeds onto a runout table, the center of the steel also solidifies, yielding a semifinished shape at a specified width and thickness. Depending on the type of caster used, various shapes are produced. The melting, casting, and rolling processes have been linked with each other to produce shapes that substantially conform to the desired shape of the finished product.

Continuous casting eliminates the need for teeming into ingots, mold stripping, reheating, and primary hot-rolling into semifinished shapes. As a consequence, the continuous process has higher yields compared to the ingot process and is roughly 10 times as energy efficient as continuous casting. Continuous casting also offers important benefits in product quality for most products because of the uniformity with which products are cast and the reduction of defects during solidification of the molten steel. Ultimately, the main benefit of continuous casting and direct rolling—both from an energy and overall economic perspective—may be the rapid feedback, which could eventually enable the elimination of most production errors that now result in substantial rejection of finished and semifinished products.

The choice of casting technology is closely related to existing melting technology. For example, in the United States in the 1970s, when open hearth and basic oxygen furnaces dominated steel production, less than 10% of steel was continuously cast. Similarly, in China and India where open hearth furnaces have still been widely used at the turn of the 21st century, the share of continuous casting was relatively low. In contrast, in the United States, approximately 96% of all steel produced at the turn

of the 21st century was continuously cast, the remaining 4% was ingot cast.

3.2 Forming

The slabs, billets, and blooms produced in casting operations are further processed to produce strips, sheets, plates, bars, rods, and other structural shapes through various hot forming operations, which are sometimes followed by cold forming operations. The exact sequence and temperatures of forming operations not only help generate desired shapes but also impart particular physical and chemical properties of the semifinished product. Prior to hot forming, semifinished shapes must be heated to rolling temperatures of about 1000 to 1200°C in a reheat furnace. Upon exiting the furnace, slabs may undergo a surface preparation step, which removes defects prior to entering a rolling mill.

The most common hot forming process is hot-rolling where a heated steel slab is passed between two rolls revolving in opposite direction. Each set of rolls produces an incremental reduction in the thickness of the slab. The final shape and characteristics of a hot formed piece depend on the rolling temperature, the roll profile, and the cooling process used after rolling. Rolling and reheating require considerable amounts of energy. The more the size of the cast steel resembles the size of the finished product, the less energy is required for finishing.

3.3 Finishing

Finishing processes are used to clean the surface of the semifinished hot-rolled steel products prior to cold-rolling or forming or coating operations. Rust and other oxides, oil and grease, and other impurities and contaminations of the steel are physically and chemically removed from the surface with help of solvent cleaners, pressurized water, abrasives, alkaline agents, salt baths, or acid pickling processes. The latter utilize steam generated from fuel to heat the pickling bath and preheat the incoming strip. Electricity is used to power various ancillary equipment, such as extraction fans and acid recovery. Acid recovery plants also use either steam or gas as primary energy inputs, depending on the type of acid employed.

Steel that has been hot-rolled and pickled may be cold-rolled to make a product thinner and smoother so that it is suitable for a variety of uses, such as car bodies, tin cans, pipes, and tubes. Energy requirements for rolling and finishing processes include fuels used for slab reheating and heat treatment, and electricity used by rolling mills and cooling beds. The electricity requirement for rolling is determined by the initial and final gauges and quality of the steel. For hot-rolling it typically ranges from 600 to 1000 MJ per ton steel.

Reheat furnaces are fired with a variety of fuels, including natural gas, coke oven gas, basic oxygen furnace gas, and fuel oil. Modern, efficient reheating furnaces in hot strip mills use about 1400 MJ per ton of steel from cold start to heat slab, whereas older furnaces typically require around 1800 MJ per ton steel and most efficient technologies currently have energy intensities around 1000 MJ per ton.

In some flat-rolled electric arc furnace facilities, thin slabs are sheared and fed directly from the caster to an inline tunnel furnace that acts as a buffer ahead of the hot strip mill. These tunnel furnaces have energy intensities in the range of 500 to 800 MJ per ton steel.

The energy intensity of heat treating furnaces in plate mills is in the order of 400 MJ per ton. However, since not all plate steel is typically heat treated, the specific energy consumption of treated plates can be much higher.

Both electricity and fuel are consumed in cleaning and annealing. Modern hydrogen atmosphere annealing furnaces use 600 MJ per ton of product, while the energy intensity of older furnaces may be 50% or more higher.

4. ANCILLARY PROCESSES

4.1 Mining, Beneficiation, and Agglomeration

Since extraction and some preparation of ores are carried out at the mine site, not at the steel mill, they are typically considered outside the scope of steel making. However, considerable energy is expended in ore preparation processes (Fig. 10), and use of variously prepared ores can significantly reduce energy expenditures at the steel mill.

Crude ores are sold as usable ore by mine operations after some processing and beneficiation. One such form of usable ores are pellets. Pellets usually contain between 60 and 70% iron, with the remainder being binder and substances contained in the original ore. Their surface geometry enables more efficient heat and air fluxes through blast furnaces and thus more uniform reduction of iron oxides and higher efficiency in material (iron, coke) and energy

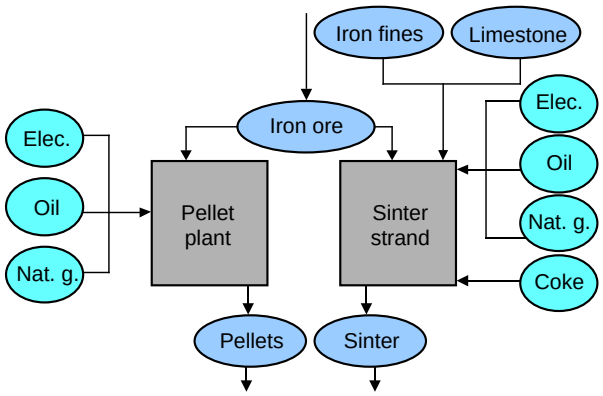


FIGURE 10 Casting, forming, and finishing processes. BOG, basic oxygen gas; COG, coke oven gas; Nat. G., natural gas; Elec., electricity.

use. Their production requires around 1000 MJ per ton of product, depending on the type of ore used, in addition to crushing, screening and washing of ores.

Other types of usable ores are distinguished, for example, as sinter and briquettes, depending on the process used. While pellet production typically occurs at the mine site or shipping port, sinter production usually occurs at the mill. A sintering plant provides a means of recycling many of the iron-bearing fines generated by iron making and steel making. In sinter strands, materials such as scale, sludge, and slag are added to ore fines and agglomerated into a suitable material. Flux, in the form of dolomite or limestone, is added to the sinter burden and calcined during the process, decreasing the energy requirements of the blast furnace process. The sinter, usually 50 to 60% iron, is easily reducible, since its high surface to volume ratio permits high contact of gases and solids. With energy intensities between 10,000 and 20,000 MJ per ton of product, sinter production is an order of magnitude more intensive than pellet production.

4.2 Coke Production

Coals suitable for coking are ground and blended before use as charge in so-called by-product coke ovens, which derive their name from the fact that during the coking process byproducts, such as light oils, tar, ammonia, and coke oven gas, are recovered for use elsewhere in the operation (see Fig. 3). Several of the byproduct fuels are mixed with combustion air for burning in combustion chambers. Under normal operating conditions, the charge is heated for 16 to 20 hours (14 or less with preheated coal—the latter makes the process approximately 10% more efficient). Then the coke is forced into waiting cars where it is water quenched to prevent combustion. After cooling, the coke is crushed and screened. Most of it is then transported to blast furnaces and the remaining coke dust is conveyed to reclamation plants and used either in sintering plants or sold. The entire process requires upward of 20,000 MJ per net ton of coke.

Form coking, a complete substitute for the conventional slot oven coke process, involves grinding and preforming the coal, which is then binded, pelletized, cured to a char, and completely devolatized to develop the necessary strength, a major problem in coke manufacture. Since briquettes are the final product, their uniform size and shape will improve blast furnace productivity. In addition, the process is enclosed, reducing emission problems. The process is not necessarily more energy efficient, although form coaking is a resource-saving technology.

Preheating of coal prior to use in the coking process may enable direct energy savings achieved in coke manufacture, improvements in the quality of the coke produced, and a consequent reduction in coke charged to the blast furnace. Alternatively, poorer quality coals may be used to produce coke equal in quality to that made from high-grade coals.

Because of high replacement costs, regionally declining supplies of low-sulfur coal and tightened emissions standards, coking capacity in the United States and many other industrialized countries has decreased significantly. Of the existing coke batteries, some are operated by iron and steel companies that produce coke mainly for their blast furnaces. The remainder is held by merchant companies that generally sell their product on the open market.

4.3 Scrap Preparation

Iron and steel scraps are used in many stages of iron and steel making. Some of the scraps (prompt scrap) accrue in production itself and are directly fed back into production, others (obsolete scrap) are returned to the industry from other industries or after use of iron and steel containing end use products. Typically, prompt scrap is of a more uniform quality and more readily available than obsolete scrap which often contains larger amounts of impurities and accrues after significant time lags between production and return to the industry.

Scrap quality is a limiting factor for increased scrap use. Certain types of scrap contain high

quantities of trace elements. High-quality steel requires a low trace element content. In the current practice, scrap is used for lower quality steel grades, such as reinforcement bars. However, electric arc furnace steel producers are gradually moving toward the production of higher quality products. In the longer run, declining steel scrap quality and increasing steel quality may pose serious challenges for recycling.

The lower the scrap quality, the lower the yield, and thus the higher the energy requirement per ton of desired output. Since prompt scrap is a waste product of the industry itself and does not replace raw material inputs, increases in prompt scrap use or quality will not noticeably improve energy efficiency. Only expansion of obsolete scrap use can lead to an overall reduction of energy requirements to the extent that obsolete scrap substitutes for steel production from ores. To increase productivity when using scrap and to replace high-cost electric energy with lower cost fuels, some basic oxygen furnaces and electric arc furnaces preheat scrap.

4.4 Direct Reduction

Direct reduced iron is iron ore in the form of lumps, fines or pellets that have had the oxygen removed by using hydrogen and carbon monoxide. Typical sources of carbon monoxide are natural gas, coal gas, and coal. Other energy inputs into the production process often include oil and electricity. Since much of the energy used is in the form of natural gas (e.g., Mexico produces virtually all its direct reduced iron with natural gas), most direct reduced iron is produced where low cost natural gas is available and where adequate capacities exist to access iron ore shipments. Use of natural gas not only reduces the need for coal and coke, but also reduces the associated emissions, such as sulfur oxides and carbon dioxide into the atmosphere.

In direct reduction the ore is heated to temperatures below its melting point. Direct energy requirements per net ton of direct reduced iron are in the order of 10,000 MJ. The product is 90 to 95% metallized and hence is suitable as a charge in blast, basic oxygen, and electric arc furnaces.

By using direct reduced iron, steel makers can dilute lower grade ores and lower cost ferrous scrap to attain acceptable levels of contamination in the mix. If direct reduced iron pellets are substituted for pelletized ores in blast furnaces, the amount of coke per ton of pig iron would be cut. However, reductions in the coke rate could result in shortages of coke byproducts, while natural gas requirements would increase considerably.

5. DRIVERS FOR CHANGE OF ENERGY USE IN IRON AND STEEL MAKING

Changes in iron- and steel-making technology, capacity utilization and production quantities all influence energy use by the steel industry. These changes are often the result of a multitude of interrelated factors, including technological changes in steel production, changes in national and international markets of iron and steel and their substitutes, changes in availability and prices of inputs, as well as environmental regulation. As the industry meets the challenges associated with technology change, product market dynamics, changes in input prices, and environmental regulation, it tends to create savings in all factors of production: labor, capital, materials, and energy.

5.1 Technology Change

Among the most notable changes that occurred in the iron and steel industry of many industrialized countries since World War II are the transition from open hearth to basic oxygen and electric arc furnaces and the use of continuous casting methods instead of ingots. Many new steel-producing countries are following these trends as their industries expand, and technologies change. Further changes in the industry are expected as product quality from electric arc furnaces continues to improve and as newer technologies such as direct reduced iron provide increasingly cost-effective alternatives to the energy intensive basic oxygen route.

The extent to which iron and steel firms adopt new technology depends in part on previous successful experience with the technology, which makes innovation less of a gamble, and on the speed at which technologies change. Firms are likely to delay adoption of technologies that are improving rapidly and tend to adopt a technology after a slowdown in technical progress, either with the furnace the firm is choosing or the competing one. Since many improvements in technology are very costly, long-lived, and once implemented determine a firm's long-term material, energy, and emissions profiles, the industry's aggregate performance changes only slowly relative to the rate at which individual technologies change.

5.2 Market Dynamics

Prior to the 1970s, the industry experienced steady growth and sustained profitability. Since furnaces with higher capacity utilization have lower heat loss per ton of furnace charge and thus higher energy efficiencies and lower emissions per ton of product, overall energy efficiencies in the industry were relatively high during years of high production and capacity utilization rates.

The recession that followed the 1970s oil price crisis has led to overcapacity and resulted in pessimistic forecasts of future demand. Nevertheless, by the beginning of the 21st century, steel makers could produce 15% more steel than the world market demanded, but they were adding 2% more capacity each year.

Since the 1970s, many traditional steel firms have shed unprofitable operations, reorganized production to eliminate waste, and focused their resources both in terms of product lines and of the geographic markets they serve. However, U.S. integrated producers and in some cases minimills were slower than their European and Japanese competitors in implementing improvements.

Integrated steel producers were more likely to close plants that had comparably small output, competed with minimills or sold to contracting industries, and plants that did not use electric arc furnaces. As a consequence, the firms that remained in the market tended to be large, and the probability of exit fell as firm size increased. As the industry consolidated, far-reaching changes in technology required larger capital investments, and as market exit of firms decreased, rates of improvements in aggregate efficiencies from factors other than economies of scale slowed down.

Other reasons for delays in technology change included a lack of profits and access to capital at sufficiently low interest rates to make the necessary large investments in new steel-making technology. The situation was exacerbated by failures to anticipate changes in product needs in the rest of the economy and underestimates of the potentials of new technologies to reduce cost and increase product quality.

5.3 Changes in Material and Energy Prices

Changes in quality of material inputs into steel making are closely related to changes in energy requirements to process furnace charge. As a consequence, the material and energy costs of steel production are related to each other as well as to other technological variables and firm experience.

For example, availability and quality of iron ores influence energy and material efficiencies. Iron ores used in China are mainly of low grade with an average iron content of 34%. Iron ores are beneficiated before they are put into blast furnaces, but because of limits to which the quality of ores can be improved through beneficiation, the Chinese steel industry historically had high coke rates and high iron-to-steel ratios. In contrast, in the 1970s and 1980s, Japanese firms have secured high-quality ore at favorable prices by engaging in long-term contracts with Brazilian and Australian ore development projects. During that time, the price of ore had dropped on the international markets to the point where Japanese firms could secure ores at prices well below the cost incurred, for example, by American firms to get ore from domestic sources, even though shipping cost for Japanese firms were substantial. The resulting cost savings helped Japanese firms to increase their market shares and profitability and thus afforded capital for investment in new capacity and improved technologies.

Improvements in competitiveness and technology also occur in response to changes in energy prices, although influences of energy prices—and for that matter any other single driver—may have be lower than often suspected. Empirical results show that even though energy is a large cost component, energy prices are not a central criterion by which integrated or minimills decide about major process changes. Other determinants include prices and quality of reduction agents. For example, for the U.S. iron and steel industry as a whole, higher coke prices are correlated with later adoption of basic oxygen furnaces and earlier adoption of electric arc furnaces.

Further influences on technology choice and energy efficiency may come from investments into a more diversified and reliable energy system, demand side management, and technology-specific investments. For example, improvements in natural gas infrastructure and technology, power recovery, and direct coal injection at the blast furnace have increased energy efficiency in the steel industry of China and its economy as a whole. Potential improvements in minimills' electricity use indicate that minimills may be a particularly good target for energy demand policies. They are large single users of electricity with a demonstrated engineering potential for efficiency improvements. Encouraging

more rapid upgrading or replacement of older capacity is one effective way to improve aggregate energy efficiency.

5.4 Environmental Regulation

Many energy intensive industries like steel are under increasing pressure to reduce their energy consumption for at least the following two key reasons. First, improving the efficiency with which energy along with other inputs is used can help realize operating cost savings and thus contribute to the industry's competitiveness. Second, reductions in energy use may occur for environmental reasons because environmental impact may reduce the appeal of steel relative to competing materials or because the need to comply with environmental regulation may increase the cost of production.

Focus on local and regional water and air quality has led to the closure of many coke ovens and smelters whose needs for cooling water and emissions of sulfur oxides resulted in significant environmental impacts. In many industrialized countries, blast furnaces continue to face closure and steel makers are increasingly reluctant to make long-term maintenance expenditures on blast furnaces and their associated technologies.

Environmental concerns have shifted to emissions of carbon dioxide and other greenhouse gases influencing global climate. With its use of large quantities of fossil fuels, the iron and steel industry accounts for about 7% of total anthropogenic carbon dioxide emissions. Efforts to reduce carbon dioxide emissions through fuel switching and efficiency improvements typically also address other inefficiencies in the industry.

A number of fuel alternatives exist for coal, the predominant source of carbon dioxide emissions by the industry. The most relevant ones from a carbon dioxide reduction perspective are natural gas, charcoal from renewable wood biomass, waste plastic, hydrogen, and electricity.

Natural gas can be injected into blast furnaces and with proper oxygen enrichment can boost hot-metal output by up to 40%, compared to about 10% in the case of coal injection. Up to 125 kg of natural gas can be injected per ton of hot metal (in energy terms equivalent to about 210 kg coal injection). The potential for carbon dioxide emission reductions amounts to 100 Mt (compared to coal injection) or about 135 Mt compared to coke use. Low carbon dioxide emissions from Mexican steel producers have in part been achieved by using a larger proportion of natural gas in the steel-making process compared to more carbon-rich fuels.

Charcoal, if produced from wood can be considered a renewable carbon source. Part of Brazilian iron production is based on the use of charcoal. Charcoal lacks the mechanical stability of coke, but it is possible to substitute blast furnace coal injection with charcoal injection. Traditional charcoal production is a very polluting activity. New, modern shaft kilns have been developed with much lower emissions compared to traditional charcoal making, but production costs are significantly higher.

Plastic wastes can be added to coke ovens to reduce coal requirements or injected into blast furnaces to substitute for coke and coal. An advantage of use in coke ovens over injection into blast furnaces lies in the possibility to generate higher value products and reduced interference with the sensitive blast furnace operation. However, byproducts from incineration of plastics, most notably gaseous emissions and generation of toxic ash, may limit overall use of plastic wastes by the steel industry.

Not all changes in fuel mix reduce overall carbon dioxide emissions. As firms in different countries undertake fuel substitution in efforts to reduce operating costs, some increases in carbon dioxide emissions may result. For example, the fuel mix in Brazil's steel industry is less carbon dioxide–intensive than in other major steel producing countries because the share of coal is low. The observed increases in net carbon dioxide emissions from Brazilian steel making have been triggered, in part, by the privatization of the steel industry in Brazil which led to a trend away from wood-based charcoal and capital-intensive hydropower generation to fossil-based electricity.

6. CONCLUSIONS

Iron and steel production is among the most energy-intensive industrial activities. Significant changes occurred in the dominant technologies used to produce steel, as well as the fuel mix of the industry and the efficiency of each technology. Several technologies widely used at the beginning of the past century—open hearth and basic oxygen furnaces—now account for only small market shares.

As technology mix in the iron and steel industry changed, upstream and downstream processes had to

adjust. Pelletization of ores and preheating of scraps, for example, are increasingly widely used methods to improve furnace performance, material and energy efficiency, and product quality. New casting and rolling technologies similarly help improve overall industry performance. For example, thin-slab casting in minimills eliminates the need for traditional hot-rolling facilities to reduce slap to hot-rolled sheet. Similarly, newly developing thin-strip casting may produce light-gauge sheet steel without several hot and cold reduction steps.

New technologies will continue to enable minimills to enter new markets. Minimills will also increase the usage of some form of iron—direct reduced iron, pig iron, hot metal—with the scrap charge to improve the quality of the steel they produce to serve even broader markets.

As the industry responds to changes in energy and material prices, changes in the cost of capital, market dynamics, and environmental regulation, the ratio of energy consumption by the steel industry to gross national product declines. For example, in constant-dollar terms, that decline was 32% in the United States between 1973 and 1984. In western Europe and Japan, input prices have risen even more sharply and even more energy conservation has been carried out.

Further improvements in technologies undoubtedly will continue to help maintain steel's competitiveness and energy and environmental performances. Many of these improvements will support the transition from aging furnaces in integrated mills to modern, smaller, more efficient operations. These changes may be most striking in China, India, and other developing countries with significant steel-production capacities and rising demand for iron- and steel-based products.

SEE ALSO THE FOLLOWING ARTICLES

Aluminum Production and Energy • Cement and Energy • Early Industrial World, Energy Flow in • Glass and Energy • Material Efficiency and Energy Use • Material Use in Automobiles • Plastics Production and Energy • Recycling of Metals • Reuse and Energy

Further Reading

Ahlbrandt, R. S., Fruehan, R. J., *et al.* (1996). "The Renaissance of American Steel." Oxford University Press, New York, Oxford.

Crandall, R. W. (1996). From competitiveness to competition: The threat of minimills to large national steel companies. *Resources Policy* **22**(1/2), 107–118.

Fenton, M. D. (2000). Iron and steel, U. S. Geological Survey.

Gielen, G., and Moriguchi, Y. (2002). CO_2 in the iron and steel industry: An analysis of Japanese emission reduction potentials. *Energy Policy* **30,** 849–863.

Hannon, B., and Brodrick, J. R. (1982). Steel recycling and energy conservation. *Science* **216,** 485–491.

Kim, Y., and Worrell, E. (2002). International comparison of CO_2 emission trends in the iron and steel industry. *Energy Policy* **30,** 827–838.

Office of Industrial Technology (OIT) (2000). Steel: Industry of the future. Washington, D.C., Office of Industrial Technology, U.S. Department of Energy, 98.

Ruth, M. (1995). Technology change in U.S. iron and steel production: Implications for material and energy use, and CO_2 emissions. *Resources Policy* **21**(3), 199–214.

Ruth, M., Amato, A., and Davidsdottir, B. (2000). Impacts of market-based climate change policy on the U.S. iron and steel industry. *Energy Sources* **22**(3), 269–280.

Sparrow, F. T. (1983). "Energy and material flows in the iron and steel industry," p. 107. Argonne National Laboratory, Argonne, IL.

Sugimoto, T. (1993). The Chinese steel industry. *Resources Policy* (December), 264–286.

U.S. Geological Survey (USGS) (2002). Mineral commodity summaries, 86–87.

Stock Markets and Energy Prices

PERRY SADORSKY
York University
Toronto, Ontario, Canada

1. **Definition and Economic Importance**
2. **Relationship between Stock Markets, Energy Prices, and the Economy**
3. **Oil Markets and Oil Prices**
4. **Futures Markets, Oil Price Volatility, and Stock Markets**
5. **Oil Prices and Currency Exchange Rate Risk**
6. **Technology Stock Prices and Energy Prices in the New Economy**
7. **Modeling Approach and Quantitative Methods**
8. **Empirical Evidence on the Relationship between Oil Price Movements and Stock Price Movements**
9. **Energy Price Risk Management**

Glossary

arbitrage pricing theory (APT) A theoretical development in corporate finance that yields a model for pricing risk. The APT model linearly relates stock returns to several risk factors.
autoregressive conditional heteroskedasticity (ARCH) A statistical modeling technique that takes into account volatility clustering.
backwardation A pricing structure in which futures prices are lower than spot prices or when the near futures price exceeds the distant futures price.
contango A pricing structure in which futures prices are higher than spot prices.
forward contract A contract for the purchase or sale of a commodity for future delivery, given by a specific delivery date, and not traded on an organized market.
futures contract Standardized contract for the purchase or sale of a commodity for future delivery, given by a specific delivery date, and traded on an organized futures markets.
heteroskedasticity The variances of a set of random variables are not constant.
ordinary least squares (OLS) A statistical modeling technique that posits a linear relationship between a set of drivers and a dependent variable.
risk A source of randomness that can be quantified.
uncertainty A source of randomness that cannot be quantified.
vector autoregression (VAR) A statistical modeling technique in which each variable of interest is explained by lagged values of itself and the lagged values of other variables in the system of equations.
volatility A tendency for prices to fluctuate widely.
volatility clustering A tendency for large movements in prices to be followed by large movements in prices in either direction.

The interaction between changes in energy prices and economic performance has been of interest for a long time. For example, in the 1890s J. D. Rockefeller publicly expressed concern over rapidly fluctuating oil prices and advocated controlling oil prices in order to stabilize oil company profits and economic activity. Conversely, the relationship between stock markets and energy prices is a relatively new and emerging field of energy economics and finance that sprung to life after the two oil price shocks of the 1970s and the development of the oil futures market in the 1980s. Serious study of the relationship between stock markets and energy prices, and energy and the macroeconomy, has only occurred during the past 20 years.

1. DEFINITION AND ECONOMIC IMPORTANCE

Higher energy prices, mostly through higher oil prices, act like an inflation tax on consumers and producers. Higher oil prices raise the cost of production for non-oil-producing companies and, in the absence of fully passing these costs on to consumers, reduce profits and dividends. Profits and dividends are included in the income side of a country's national accounts. Consequently, sharp reductions in profits and dividends reduce gross

 © 2004 Elsevier Inc. All rights reserved.

domestic product (GDP). Reduced profits dampen the future outlook for corporate earnings, a major driver of stock prices, and decrease stock returns. Steel mills, paper mills, sawmills, airlines, trucking firms, and other companies lay off workers and close plants when faced with steadily increasing fuel bills. Higher oil prices impact consumers directly via increases in the price of gasoline. This reduces the amount of disposable income consumers have left for purchasing other goods and services. Firms selling nonessential goods and services experience a decrease in business activity, which in turn results in lower profits, corporate earnings, and dividends and feeds back into the economy and the stock market.

With the U.S. Federal Reserve moving to an ever more transparent monetary policy, oil price changes have the potential to have a larger impact on stock prices than interest rate changes. This is because oil prices fluctuate more than interest rates do, and oil price changes embody more uncertainty regarding oil demand and supply imbalance, geopolitical, and institutional factors than do interest rates.

2. RELATIONSHIP BETWEEN STOCK MARKETS, ENERGY PRICES, AND THE ECONOMY

Real crude oil prices were highest in two distinct periods: 1861–1869 and 1974–1981. Real oil prices in the period 1861–1869 were very volatile and this is historically identified as the period with the highest real oil prices and the most amount of oil price volatility. Russian oil exports began in the 1880s and this helped to lower oil prices. After the discovery of oil at Spindletop, Texas, at the turn of the 20th century, real oil prices tended to stay fairly constant for long periods of time, particularly between the 1930s (at approximately the time that the East Texas oil field was discovered) and the early 1970s. Price changes did occur, but these changes were few and far between. After the oil price shocks of 1973 and 1979, the underlying structure of the world oil market changed and oil price volatility increased considerably. In fact, oil price volatility in the late 1970s was almost as high as that during the 1860s. Increased oil price volatility created a need for risk management tools and the development of forward and futures markets for crude oil followed. Futures markets for natural gas and electricity are also important, but most of the attention regarding stock markets and energy prices is focused on the crude oil futures market. Today, it is the interaction between oil futures markets and the stock market that financial managers watch closely for signs of pricing pressure on the economy. In the future, with deregulated natural gas and electricity markets, the prices of natural gas and electricity will have a much larger impact on financial markets and the economy.

Since World War II, almost every recession in the United States has been preceded by a large increase in oil prices. Generally, the negative correlation between oil prices and GDP is taken as an empirical fact. During the four quarters preceding the onset of the 1948–1949, 1953–1954, 1957–1958, 1960–1961, and 1969–1970 contractions, the real price of energy increased on average 1.5%. The average increase in real energy prices prior to the onset of the four recessions during the periods 1973–1975 (called the U-shaped recession because real GDP growth during this period resembled the letter U), 1980, 1981–1982 (the W-shaped recession), and 1990–1991 (the V-shaped recession) was 17.5%. This increase in real oil prices was much greater than in earlier recessions. The economic downturns in the United States during the 1970s and 1980s were the largest since the Great Depression. The impact of oil price shocks on the U.S. economy became more pronounced after 1970. This was the time when U.S. demand for imported oil increased and the Organization of Petroleum Exporting Countries (OPEC) cartel began to exert their influence on the world oil market. The 1970s and early 1980 were also turbulent times for world equity markets. Oil price shocks can also occur from large decreases in oil prices. Large decreases in oil prices in 1986 and 1998 are two such examples.

Until the early 1970s, the price of internationally traded oil was set by a group of major oil companies (know as the seven sisters). During this time, the major oil companies set the price of oil on a reference barrel of oil (Arabian Light 34 API) so that the price remained constant in nominal dollars. OPEC basically inherited this pricing regime with some modifications. It was not until 1987 that OPEC adopted the current pricing regime whereby the price of internationally traded oil is determined by a pricing formula based on a reference basket of crude oils.

To understand the relationship between stock markets and energy prices, it is useful to use an equity pricing model. In such a model, the price of equity at any point in time is equal to the expected present value of discounted future cash flows. A firm's cash flow can be affected by economy-wide

factors as well as firm-specific factors. Approximately 60–80% of stock price movements result from broad-based market and industry sector factors and only 20–40% are related to company-specific factors. Cash flows are affected by business cycle conditions, and business cycle conditions can be represented by several key macroeconomic variables, such as industrial production, crude oil prices, interest rates, consumer prices, and possibly currency exchange rates. The ability to manage firm-specific and economy-wide risks determines the economic evolution of a firm.

Oil prices can affect the expected present value of cash flows in several ways. Oil is a resource and essential input into the production process. The basic production function requires inputs on capital (K), labor (L), energy (E), and materials (M): This is the basis for the KLEM model. Increases in energy costs, with the prices of other inputs held constant, will increase the cost of production and, without any offsetting effects, will reduce company profits and dividends. More generally, energy price increases can be mitigated by substitution effects among the other factor inputs. The impact on stock prices depends on whether the company is a net producer or consumer of energy. Increases in energy prices lower (raise) the stock prices of energy-consuming (producing) companies.

The net effect of higher energy prices on the stock market depends on how many firms are net users of oil. Given that there are more firms using oil than producing oil, oil is an input to the global economy. As a result, increases in oil prices lead to decreases in stock prices.

Future oil price changes can also impact stock prices through the discount rate. The expected discount rate is composed of a real interest rate component and an expected inflation component. Higher oil prices put upward pressure on the general price level, thereby raising the aggregate supply curve for a country that is a net user of oil and oil-related products. Higher oil prices can lead to increases in expected inflation and higher nominal interest rates. Since interest rates are negatively correlated with stock prices, increases in interest rates depress stock prices.

Economic growth leads to feedback effects between the economy, energy prices, and the stock market. Increasing industrial production generates a stronger economy and, as a consequence, higher profits and dividends. Higher profits and dividends raise stock prices. A stronger economy also raises the demand for oil and oil-related products. An increased demand for oil and oil-related products puts upward pressure on aggregate prices and interest rates, and the reaction of the Central Bank becomes critical to the movement of the stock market. The Central Bank can either accommodate or dampen the increased economic activity via expansionary or contractionary monetary policy, respectively. Which path the Central Bank takes directly impacts stock returns.

Interest rates can impact stock prices in three ways. First, changes in interest rates are changes in the price charged for credit. The price of credit affects both firms and consumers. Firms seeking to expand existing operations or build new ones need to borrow money. Consequently, the price charged for credit is an important influence on the level of corporate profits and changes in corporate profits are a major driver of stock prices. The price of credit, through its impact on corporate profits, affects the price that investors are willing to pay for equities. Companies prefer to borrow money or issue debt when interest rates are low. In addition, a higher interest rate makes it less likely for consumers to purchase big ticket items, such as houses, automobiles, furniture, and appliances. Second, higher interest rates alter the relationship between competing financial assets. As interest rates rise, fixed-income securities become more attractive relative to equities and investors rebalance their portfolios to take this into account. Third, some stocks are purchased on margin and changes in interest rates impact the cost of carrying margin debt. Higher margin debt limits the ability of investors to participate in the stock market.

Inflation, a sustained rise in the general price level, erodes the real value of investments. This was particularly evident in the high-inflation period of the 1970s and early 1980s. As an example, consider the case of person A loaning $100 to person B for 1 year. Person A would like to earn a 3% real rate of return. Suppose that the current inflation rate is 2% per year and this is expected to be the inflation rate in the following year. Person A will then loan person B $100 at an interest rate of 5%. At the end of 1 year, person B agrees to pay person A $105. Now suppose that the inflation rate over the coming year is not 2% but instead 4%. In this case, person A's real return on his or her investment is 1% (5%−4%). Inflation (unexpected) has lowered the return on person A's investment.

Furthermore, inflationary pressures tend to lead to higher interest rates and a slowing economy, which lessens the ability of growth companies to meet revenue and earnings targets, causing stock prices to

fall. Oil price increases are one of the most likely sources of inflation and oil is a raw material used in almost everything that is made. Crude oil and natural gas are required to produce 98% of North American manufactured products. Oil price increases usually impact the transportation sector first through higher prices for fuel (gasoline, diesel, and jet fuel), with initial price hikes being absorbed by this sector. Airlines are particularly susceptible to higher oil prices because jet fuel represents a major cost of doing business in the airline industry. Persistent oil price increases lead to higher producer prices and eventually, after 6–9 months, inflation. Inflation, which erodes the real value of investments, leads to higher interest rates as contractionary monetary policy is used to bring economy-wide price increases down. Crude oil price increases in 1973, 1979, and 1990 were followed by stock market downturns.

Although oil price increases are not very good for the performance of the economy as a whole, oil price increases are very beneficial to companies engaged in the oil and gas industry. Consequently, oil price risk is a major concern for oil and gas companies. Higher oil prices lead to increased profits and spending on capital budgets, whereas lower oil prices lead to lower cash flows and reduced spending patterns of oil producers. Annualized oil price volatility is approximately 25% per year and annualized natural gas price volatility is approximately 40%. Hence, price variability is a major impact on the stock returns of oil and gas companies. There are three ways in which crude oil prices affect oil company stock prices. First, investors in oil stocks are usually only concerned about the flow of profits from their stock investment and future profits are determined by the expected average price of oil over a particular period. If investors expect the average price of oil to fall (rise), then they will sell (buy) shares in oil companies. Second, oil company stocks are more leveraged than crude oil. Oil exploration, development, and production is a capital-intensive business and oil and gas companies often borrow large amounts of money to go into production. The capital costs of exploration, development, and production depend on several factors, including whether the deposit is underground, under the ocean, or open pit; the type of plant required for oil extraction; and the location of the plant. As oil prices rise or fall, marginal oil deposits become profitable or unprofitable, which in turn changes the value of the company by more than the change in the actual price of crude oil. Third, the profits of oil companies, as for other companies, are a function of production costs and production costs are influenced by the economic business cycle, which in turn is affected by energy prices. In the usual case, rising energy prices contribute more to the revenue side of an energy company's profits than to the cost side.

3. OIL MARKETS AND OIL PRICES

The global oil and gas industry is a mature industry that has prospects for growth as long as global demand for oil and gas products continues to grow. Oil and gas product demand is difficult to predict but is generally highly correlated with worldwide industrial production. In 2000, world oil consumption was 73,905 thousand barrels of oil per day (b/d). The United States accounted for 25.6% of this world total, whereas Europe accounted for 21.4%. During the period 1990–2000, world oil consumption increased 12.9%. World demand for oil is expected to grow 2% per year during the next decade. United States energy demand sensitivities indicate that a 1% increase in real U.S. GDP raises U.S. petroleum demand by approximately 0.6% and U.S. natural gas demand by approximately 1.1%. A 10% increase in crude oil prices, assuming no price response from nonpetroleum energy sources, reduces U.S. petroleum demand by 0.3%. These data were obtained from the U.S. Energy Information Administration Short-Term Energy Outlook (www.eia.doe.gov/emeu/steo/pub/sensitivities.html). The oil and gas industry, however, is very sensitive to business cycle conditions.

In 2000, world oil production was 74,510 thousand b/d. The five largest producing countries were Saudi Arabia (12.3% of world production), the United States (9.8%), the Russian Federation (9.0%), Iran (5.2%), and Mexico (4.8%). In 2000, OPEC accounted for 41% of world production. When OPEC member countries act in unison, they command monopoly power over approximately one-third of the world's oil production. During the period 1990–2000, world oil production increased by 13.9%.

Of course, future oil production depends on where the oil reserves are located. This is where it becomes interesting because most of the world's consumption of oil (62.4% of the world total) occurs in the Organization of Economic Cooperation and Development (OECD) countries, whereas most of the world's oil reserves are located outside of OECD countries. In fact, OECD countries only account for 8.1% of the world's oil reserves. For example, the United States, the world's largest consumer of oil,

currently imports approximately 60% of its crude oil. During 2001, approximately 48% of U.S. crude oil imports came from the Western Hemisphere (19% from South America, 15% from Mexico, and 14% from Canada), whereas 30% came from the Persian Gulf region (18% from Saudi Arabia, 9% from Iraq, and 3% from Kuwait). Since 1960, U.S. petroleum imports have increased more than sixfold.

In 2000, total world proven oil reserves were estimated at 1046.4 thousand million barrels of oil. At current production rates, this implies that there are 39.9 years of oil remaining in the world. Saudi Arabia has the most proven oil reserves with 261.7 thousand million barrels of oil. This equals 25% of the world total. OPEC countries own 77.8% of the world's proven oil reserves. In comparison, the United States and Mexico have proven oil reserves numbering 29.7 (2.8% of the world total) and 28.3 (2.7%) thousand million barrels of oil, respectively. Plentiful cheap oil is found mostly in OPEC countries. The wild card is the Russian Federation, which, in the interim, can rival Saudi Arabia in terms of crude oil production.

In a competitive well-functioning market, the price of oil would be determined by the intersection of demand and supply. An increase in demand (assuming supply does not change) would increase oil prices and an increase in supply (assuming demand does not change) would decrease oil prices. The actual mechanics of the world oil market, however, are much more complicated and are heavily influenced by global politics, institutions, and futures markets.

Political factors are important determinants in oil policy in both the oil-exporting countries (historically OPEC) and the oil-importing countries. The oil-exporting countries tend to intervene in the oil market when prices are too low. Low oil prices reduce the revenue of oil-exporting countries, dampening their economic growth and prosperity. OPEC sells oil in international markets on the basis of price formulae that use as a reference the spot or futures prices of certain important crude oil prices, such as West Texas Intermediate, Brent, or Dubai. Oil-importing countries often enact oil-directed energy policies when oil prices become too high. Higher oil prices tend to be followed by inflation, which erodes the real value of investments and slows economic growth and well-being. Oil-exporting countries can also use oil as a strategic weapon in response to U.S. foreign policy. The world oil market is populated by many consumers and many suppliers, which by itself suggests a competitive market for crude oil. The reality is that the world oil market is very much influenced by institutional factors such as OPEC, strategic oil reserves in OECD countries, and the role that the large multinational oil companies play.

Generally, the difference between oil that is produced and oil that is consumed, if positive, goes into oil inventory stocks. Short-term supply disruptions can be met with a drawdown of oil stocks. Obtaining reliable estimates on oil stocks, however, is difficult because different countries use different reporting systems. Moreover, there is the issue of oil tanker traffic, which plays a major role in getting oil out of the producing countries in the Middle East and into the OECD consuming nations. Supply bottlenecks caused by oil tanker traffic can have a huge impact on gasoline prices in oil-importing countries.

Trade on the marginal barrel of oil sets the price of crude oil. The 11 members of OPEC (Algeria, Indonesia, Iran, Iraq, Kuwait, Libya, Nigeria, Qatar, Saudi Arabia, United Arab Emirates, and Venezuela) are the marginal producers. The near-term outlook for the oil market can be determined be matching data on future consumption, OPEC oil being transported by tanker, and oil inventories. The problem is that OPEC members are very secretive about how much oil they are actually exporting at any given time—so secretive, in fact, that they do not even reveal their oil exports to each other. Tracking the amount of oil in transit is difficult but essential to understanding the world oil market.

The bulk of OPEC oil moves from Quoin Island in the Persian Gulf to ports in New York, Gibraltar (via the Suez Canal), New Orleans, Montevideo, Uruguay, Fremantle, Australia, and Yokohama, Japan. It can take several days to fully load an oil tanker at Quoin Island. Once an oil tanker is loaded, it sails to its destination. Sailing time depends on the distance traveled. The shortest trips of 12 or 13 sailing days are from Quoin Island to Gibraltar and Fremantle. The longest trip takes 31 days to New Orleans. Once the crude oil arrives at its destination, it is unloaded and piped to refineries to be turned into refined products such as gasoline, heating oil, and jet fuel. The refined products are then moved in pipelines and trucks to consumer outlets. This process takes approximately 2 months. In total, it can take 3 or 4 months for crude oil from the Persian Gulf to reach gasoline stations in the United States. Consequently, knowing the amount of oil in transit is essential to knowing the amount of oil available for supply in the next 3 or 4 months. The oil futures markets react quickly to expectations about differences in oil supply.

4. FUTURES MARKETS, OIL PRICE VOLATILITY, AND STOCK MARKETS

Oil is a commodity that is traded in both the spot and the futures markets. The spot market reflects trading for oil delivered today. The futures market reflects trading for oil delivered at some specified time in the future. The futures market provides an essential ingredient to risk management. In the oil market, futures prices tend to be followed more closely than spot prices because futures prices are not as affected by the short-term supply disruptions that affect spot prices. Futures markets for petroleum products are located in New York, London, and Singapore. The three major energy futures exchanges are the New York Mercantile Exchange (NYMEX), the International Petroleum Exchange (IPE), and the Singapore International Monetary Exchange (SIMEX). The NYMEX launched the futures contract on heating oil No. 2 in November 1978. In April 1983, NYMEX introduced the popular futures contract on crude oil. A futures contract on unleaded gasoline followed a few years later. The natural gas contract has only traded since April 1990. The NYMEX natural gas contract size is 10,000 million British thermal units of natural gas deliverable to Henry Hub, Louisiana.

The deregulation of the natural gas industry began in 1978 and was completed in November 1993. As deregulation evolved, the marketplace changed from one dominated by fixed long-term contracts (provided by the pipeline companies) to a spot market, followed by a forward market, and finally a futures market. The success of the natural gas futures contract in the 1990s enabled the natural gas industry to more efficiently manage risk through hedging strategies. In turn, the natural gas futures market has become the foundation for other types of derivatives and now forms the basis for financial engineering in the natural gas industry. Since its introduction in April 1990, the NYMEX natural gas contract has increased in importance as an effective financial instrument for natural gas price discovery.

In 1996, NYMEX launched two futures contracts on electricity. Trading in electricity futures is likely to increase in importance as North America experiments with energy market deregulation. NYMEX also trades options on futures for crude oil, New York Harbor unleaded gasoline, heating oil No. 2, natural gas, and electricity. Options on futures contracts are very useful for risk management.

The IPE trades two popular petroleum futures contracts in Brent crude oil and gasoil. Gasoil can be either heating oil or diesel fuel. Gasoil is the largest single product market in Western Europe. The IPE also trades futures in natural gas. The SIMEX trades residual fuel oil and the market is relatively small compared to NYMEX and IPE.

The West Texas Intermediate (WTI) crude oil futures contract traded on NYMEX is the most closely watched oil price in the world. One contract of WTI crude oil is for 1000 barrels. The NYMEX crude oil market is the largest market (in terms of volume traded) outside of the financial futures market. The WTI crude oil futures market is clearly important to the United States. Moreover, the physical market underlying WTI crude oil is of particular importance to those interested in the demand and supply relationships that reflect the flow of oil from the Gulf of Mexico into the Chicago region. In a broader sense, the WTI crude oil futures contract has become a benchmark for the global oil industry. Reliable fundamental information on global demand and supply in the oil market is often difficult to obtain. Production estimates for some countries, for example, tend to be higher than actual production, and consumption data tend to be made available after some lag time. In contrast, reliable data on oil demand and supply in the United States are often easily accessible and very reliable. Given that United States oil consumption is approximately 25% of world consumption and the fact that the United States accounts for approximately 31% of global GDP, it is natural to use the U.S. data as a proxy for unobservable global fundamentals.

In some instances, futures prices can be used to provide a reasonably good forecast of future spot prices. In other cases, futures prices may provide poor forecasts of future spot prices. The key to determining when futures prices might reasonably be expected to forecast spot prices is to understand the role that storage plays in commodity markets.

Nonstorable commodities are those that are perishable and cannot be stored. Examples include agricultural products such as dairy products. For nonstorable commodities, futures prices reflect the relationship between demand and supply for commodities to be delivered in the future. Small changes in anticipated supply and demand conditions are immediately reflected in futures prices. In this case, futures prices can be used to forecast future spot prices.

Some commodities can be stored and have large inventories. For these commodities, futures prices are equal to the spot price plus the carrying costs associated with storage. In this case, futures prices provide little guidance to future spot prices. Examples include agricultural products such as corn and wheat and also metal products.

Some commodities, such as crude oil, can be characterized as storable commodities with modest inventory levels. In general, consumption is high relative to inventory levels. If futures prices are lower than spot prices, then the pricing structure is called backwardation. In this case, oil can be characterized as a nonstorable commodity. If oil supplies are expected to increase in the future, then the futures price would decrease relative to the spot price. This difference in price creates an apparent arbitrage possibility but low inventory levels prevent the actual collection of arbitrage profits. When commodity markets are in backwardation, futures prices are useful in predicting future spot prices.

If futures prices are higher than spot prices, then the pricing structure is called contango. In this case, oil can be characterized as a storable commodity with large inventory. If the supply of oil is expected to decrease in the future, then futures prices will be higher than spot prices. The difference between futures prices and spot prices reflects carrying costs associated with storage. When commodity markets are in contango, futures prices are of little use in predicting future spot prices.

Futures markets respond to news on four factors: demand and supply fundamentals, geopolitics, institutional arrangements, and the dynamics of the futures market. Whereas some of the spot markets for crude oil tend to be thinly traded, the futures markets (or markets for paper barrels) tend to be very liquid. Futures markets respond not just to facts about the demand and supply balance but also to rumors. In fact, some of these rumors can impact trader sentiment, which may be unrelated to the actual mechanics of the physical market for crude oil. Unanticipated changes in these factors, be it fact or rumor, can create volatility in oil futures prices.

Volatility in the oil futures market leads to uncertainty and this affects the decision making of consumers, producers, investors, and governments. Economic uncertainty creates situations in which planned spending on consumption and investment is postponed or canceled. This economic uncertainty affects the consumption, investment, and government spending categories of the expenditure side of a country's national accounts. Oil price uncertainty impacts the profits of firms, which in turn creates more uncertainty in the stock markets. Uncertainty in the stock markets usually translates into stock price declines, which further affect the economy either through reductions in consumer wealth or postponed investment by business. Volatility in oil prices makes short-term oil price forecasting very difficult. In the long term, oil price fundamentals should dominate, making it easier, at least conceptually, to forecast the price of oil. Forecasting oil prices and oil price volatility, in general, is very difficult, but progress has been made in recent years by using new statistical techniques in modeling and simulation.

5. OIL PRICES AND CURRENCY EXCHANGE RATE RISK

There is evidence that currency exchange rates impact the movement of commodity prices. Since commodities are homogeneous and traded internationally, it is very likely that they are subject to the law of one price. This means that commodities have similar prices in each country's home currency. Thus, as the U.S. dollar weakens relative to other currencies (ceteris paribus), commodity consumers outside the United States should be willing to pay more dollars for commodity inputs priced in U.S. dollars. A strengthening U.S. dollar relative to other currencies (ceteris paribus) increases the price to commodity consumers outside the United States resulting in lower demand for the commodity. Most internationally traded commodities, including crude oil, are priced in U.S. dollars. Consequently, currency exchange rate movements may be an important stimulus for energy price changes. In particular, currency exchange rate movements impact crude oil prices and crude oil prices affect the prices of derivative products, such as heating oil and unleaded gasoline. Crude oil price changes also affect the price of natural gas. Thus, changes in crude oil prices create a ripple effect throughout the pricing structure of petroleum products. The dominance of the U.S. dollar may be challenged in the future as European countries make more use of the Euro currency in their business transactions. Already, some European countries prefer to price oil in Euros and measure oil in metric tonnes. Thus, there are times when the price of crude oil changes not so much from changes in the demand and supply balance in the global oil market or changes in OPEC pricing structure but instead from changes in currency exchange rates.

6. TECHNOLOGY STOCK PRICES AND ENERGY PRICES IN THE NEW ECONOMY

There are two views on the relationship between oil price movements and technology stock prices. The

first view is that oil price changes have very little impact on technology stock prices because technology companies engage in business activities that are not very energy intensive. The second view is that oil price changes do have an impact on technology stock prices because oil price increases fuel inflation and inflation leads to economic downturns. High-technology companies are very sensitive to the overall business cycle. Both views deserve further discussion.

The basis for the first view is the fact that technology companies produce products or services that are less energy intensive than products produced in other sectors of the economy. For example, it takes less energy per unit of output to produce telecommunications equipment than it does to produce steel beams. Manufacturing, in general, has also become less energy intensive. In the United States, it takes 45% less energy to produce one dollar of GDP today than it did in 1973. Automobiles built today are 50% more fuel efficient than automobiles built in the 1970s. The combination of less energy-intensive manufacturing and energy efficiency suggests that technology stock prices are relatively unaffected by oil price movements.

The second view is that oil price movements have a major impact on technology stock prices. Technology stocks, especially prices of Internet stocks, tend to be valued at very high price earnings multiples and stock price movements are linked to expected earnings. These stocks are highly vulnerable to market cycles. In particular, technology stocks, like other growth stocks, are very susceptible to inflation because inflation reduces corporate earnings. Furthermore, inflationary pressures tend to lead to higher interest rates and a slowing economy, which lessens the ability of growth companies to meet revenue and earnings targets, causing stock prices to decline.

Oil price increases are one of the most likely sources of inflation. Oil price increases usually impact the transportation sector first, with initial price hikes being absorbed by this sector. Persistent oil price increases lead to higher producer prices and eventually, after 6–9 months, inflation. Inflation, which erodes the real value of investments, leads to higher interest rates as contractionary monetary policy is used to bring economy-wide price increases down. In the high-technology sector, oil price increases can directly impact technology stock returns by raising the costs of the transportation component of e-business. When a customer places an order for a product online, the product still needs to be shipped from the warehouse to the customer's doorstep. Higher interest rates also create competition for equity financing, and equity financing is more important than debt financing in the high-technology sector. In March 2000, for example, the debt equity ratio for NASDAQ listed stocks was 16.6%, whereas that for the rest of the U.S. stock market was 85.2%. Moreover, some large technology companies such as Cisco Systems have no debt. As interest rates rise, private investors may choose to buy fixed-income securities rather than invest in venture capital pools to finance the startup of new technology companies.

7. MODELING APPROACH AND QUANTITATIVE METHODS

There are three main ways to empirically investigate the interaction between stock prices and energy prices. The first approach uses arbitrage pricing theory (APT), the second approach uses a discounted cash flow model, and the third approach uses unrestricted vector autoregressions.

The APT approach uses an arbitrage pricing argument to relate priced risk factors to required rates of return in assets (securities):

$$R_t = \alpha + \sum_{j=1}^{k} \beta_j F_{jt} + \varepsilon_t, \qquad (1)$$

where R_t is the expected excess return on an asset or stock market index, β_j are the factor loadings or systematic risks, and F_{jt} are the risk factors. The variable ε_t is a random error term. The APT assumes that the total variability of an asset's returns can be attributed to market-wide risk factors that affect all assets and other risks that are firm specific. Market-wide risk factors are known as systematic or non-diversifiable risks, and firm-specific risk factors are known as unsystematic or diversifiable risks. A widely diversified portfolio of equities can reduce or eliminate unsystematic risks, but systematic risk cannot be diversified. Thus, an investor must be compensated for systematic risk. In APT models, assets can have exposure to a number of different risks, including market risk, inflation risk, default risk, exchange rate risk, and oil price risk. Equation (1) can be estimated by ordinary least squares (OLS) regression using a time series data set on a single asset or as a cross section for one particular point in time with many assets. Model adequacy can be checked using regression diagnostic tests. More sophisticated models may involve many assets or classes of assets.

In this case, a system of equations estimation approach would be used. There is considerable evidence that asset prices display time-varying risk premiums and volatility clustering. Time-varying risk premiums can result from time variation in the risk sensitivities (factor loadings) or time variation in the market premiums (factors). Volatility clustering means that major stock price changes follow major stock price changes but the direction of the change is not known. As a result, time-varying autoregressive conditional heteroskedasticity models are often used in the estimation of Eq. (1).

The second approach uses a discounted cash flow model to test whether stock prices react rationally to changes in oil prices:

$$RS_t = E_{t-1}(RS_t) + (E_t - E_{t-1}) \sum_{j=0}^{\infty} \rho^j \Delta C_{t+j} - (E_t - E_{t-1}) \sum_{j=1}^{\infty} RS_{t+j} + \sum_{s=0}^{k} \theta_s OIL_{t-s} + \eta_t, \quad (2)$$

where RS_t is the log real return on a stock in period t, E_t is the mathematical expectation operator, C_t is the log of real cash flow in period t, ρ is a parameter close to but not equal to 1, OIL_t is the percentage change in oil prices, and η_t is a random error term. The model can be estimated by OLS. If the stock market is rational, then the coefficients on the OIL variables should be statistically insignificant from zero. At the economy-wide level, aggregate future cash flows can be approximated by future industrial production.

The third approach uses an unrestricted vector autoregression (VAR). This approach is intentionally unrestricted and useful for determining or confirming patterns between economic variables. Many interesting dynamic relationships can be captured using multiequation regression models such as VARs. Moreover, the VAR is very useful for performing policy simulations. The basic form of the VAR treats all variables symmetrically. No reference is made to distinguish which variables are independent and which variables are dependent. Consider a congruent statistical system of unrestricted forms represented by Eq. (3):

$$\mathbf{x_t} = \sum_{\tau=1}^{p} \Pi_\tau \mathbf{x}_{t-\tau} + \varepsilon_t, \;\; \varepsilon_t \sim \text{iid}(0, \Omega), \; t = 1, \ldots, T, \quad (3)$$

where $\mathbf{x_t}$ is a (nx1) vector of variables, and the error term ε_t is independent and identically distributed with mean zero and variance Ω. Economic theory guides the choice of variables. Typical variables used in studying the relationship between stock prices and energy prices include industrial production, interest rates, stock prices, oil prices, consumer prices, and, possibly, exchange rates. The variables in the vector $\mathbf{x_t}$ might also be transformed into growth rates or may be estimates of volatility. In the latter case, the VAR would be used to investigate the dynamic relationship between oil price volatility and stock market volatility. Basic research questions include the following: What are the determinants of stock market volatility? Has stock market volatility increased over time? Has international financial integration led to faster transmission of volatility across international stock markets?

8. EMPIRICAL EVIDENCE ON THE RELATIONSHIP BETWEEN OIL PRICE MOVEMENTS AND STOCK PRICE MOVEMENTS

The empirical evidence on the relationship between oil price movements and stock price movements is mixed. Empirical models can be estimated using daily, monthly, quarterly, or annual data, and the analysis can be conducted at the firm, industry, or economy-wide level.

Empirical models using the APT model generally use monthly data to investigate the relationship between oil price movements and stock price movements across industry sectors or at the economy-wide level. Generally, the empirical results indicate that an oil price factor is an important driver behind stock price movements. At the economy-wide level, an oil price risk factor is important but it does not have the same impact on all countries' stock markets. Countries that are net importers (exporters) of oil and gas tend to have a negative (positive) correlation between oil price changes and stock returns. The Group of Seven (G7) economies (Canada, the United States, England, France, Germany, Italy, and Japan) is adversely affected by oil price increases. This effect is muted in Canada since it is a net exporter of energy products.

Industries are not all the same and different risk factors can have different effects across different industries. Empirical results suggest that factor sensitivities vary across industries. At the industry level, oil price risk factors have a positive and significant impact on oil company stock returns. For Canadian and U.S. oil and gas companies, market

betas are typically approximately 0.7 and oil price betas are typically approximately 0.3. For industries that are net users of oil and gas, such as airlines and trucking companies, oil betas have a significant negative impact on stock prices.

Empirical results from a discounted cash flow model to test whether stock prices react rationally to changes in oil prices are mixed. The reaction of Canadian and U.S. stock markets to oil price shocks can be completely accounted for by the impact of oil price shocks on real cash flows. The results for the United Kingdom and Japan are not very strong. For these countries, their respective stock markets seem to overreact to oil price shocks.

The VAR approach has been widely used in studies investigating the relationship between oil price movements and stock price movements. This approach, which is generally applied at the economy-wide level using monthly or quarterly data, has provided evidence across a variety of developed and emerging economies showing that oil price changes do impact stock returns. One of the more interesting discoveries from these studies is that oil price changes have asymmetric effects on economies and the stock market. Positive oil price changes exert a much larger impact on economies and stock markets than do negative oil price shocks. This is because positive oil price shocks are associated with an increase in uncertainty. In periods of high uncertainty, firms postpone their hiring and investment decisions. This slows GDP growth. Negative oil prices shocks tend to be viewed as temporary and unlikely to affect the investment and hiring decisions of firms. Any benefits of lower oil prices are usually seen in increased profits for energy-consuming companies.

The VAR approach applied at the industry or firm level using daily data to investigate the lead–lag relationship between oil futures prices and stock prices during the day has provided little evidence of a relationship between broad-based stock prices and oil prices. This is probably because oil price changes take several months to impact the non-oil-related sectors of an economy. As expected, there is strong evidence to support the hypothesis that daily oil company stock prices respond quickly to daily oil price movements.

The VAR approach is also useful for studying the relationship between stock market volatility and the volatility of oil prices. Empirical results indicate that the conditional volatilities of oil prices, interest rates, and U.S. inflation rates each have a significant impact on the conditional volatility of equity markets.

9. ENERGY PRICE RISK MANAGEMENT

Fluctuating energy prices create uncertainty for both consumers and producers of energy products. Energy consumers prefer low and stable energy prices, whereas energy producers prefer higher energy prices. This situation creates a demand for risk management practices. There are a number of risk management tools, including forward contracts, futures contracts, options on futures contracts, and other derivative products (some traded and others not traded).

A typical example is an electricity power generator that needs to buy a known amount of natural gas 6 months in the future. The power generator is a heavy user of energy products and is concerned that natural gas prices will increase in the future. Rising natural gas prices will increase costs and lower profits, possibly resulting in declining earnings, a cut in dividends, and declining share prices. A useful risk management strategy is to employ a hedging strategy. The power generator could buy a forward contract for delivery of the natural gas in 6 months. In this case, the power generator buys the required amount of natural gas from a seller at a price negotiated today. This locks in the price at today's forward prices. The power generator could also use the futures markets for natural gas to buy the appropriate number of natural gas futures contracts on the NYMEX for delivery of natural gas 6 months in the future. Provided that the price of natural gas increases enough in the 6-month period, the forward or futures contracts will have provided the power generator with an effective hedge against rising natural gas prices. Hedging strategies that use forward or futures contracts eliminate the upside risk by locking in the price for natural gas. This eliminates the company's exposure to rising natural gas prices but, unfortunately, also gives away all or part of the company's exposure to declining natural gas prices.

Buying options on natural gas futures contracts is another form of risk management available to the power generator. In this case, the power generator pays the premium to purchase the options. The options premium is a form of insurance against rising natural gas prices. The power generator has the right but not the obligation to purchase the underlying futures contracts. The options will be exercised if natural gas spot prices rise high enough relative to the futures prices specified in the options contract.

If spot natural gas prices decline, then the options expire worthless. The company loses the premium paid on the options but can still buy natural gas in the spot market at a lower price than would have been available under the forward or futures contract.

SEE ALSO THE FOLLOWING ARTICLES

Business Cycles and Energy Prices • Derivatives, Energy • Economics of Energy Demand • Economics of Energy Supply • Energy Futures and Options • Inflation and Energy Prices • Innovation and Energy Prices • Markets for Petroleum • Oil Price Volatility • Prices of Energy, History of • Trade in Energy and Energy Services

Further Reading

BP (2001). Statistical review of world energy. www.bp.com.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. *Econometrica* **50,** 987–1008.

Ferson, W., and Harvey, C. R. (1994). Sources of risk and expected returns in global equity markets. *J. Banking Finance* **18,** 775–803.

Fusaro, P. C. (1998). "Energy Risk Management: Hedging Strategies and Instruments for the International Energy Markets." McGraw-Hill, New York.

Hamilton, J. D. (1983). Oil and the macroeconomy since World War II. *J. Political Econ.* **91**(2), 228–248.

Huang, R. D., Masulis, R. W., and Stoll, H. R. (1996). Energy shocks and financial markets. *J. Futures Markets* **16**(1), 1–27.

Jones, C. M., and Kaul, G. (1996). Oil and the stock markets. *J. Finance* **51**(2), 463–491.

Sadorsky, P. (1999). Oil price shocks and stock market activity. *Energy Econ.* **21**(5), 449–469.

Yergin, D. H. (1991). "The Prize: The Epic Quest for Oil, Money and Power." Simon & Schuster, New York.

Storage of Energy, Overview

MARCO SEMADENI
Swiss Federal Institutes of Technology
Zurich, Switzerland

1. **Introduction**
2. **Electricity Storage**
3. **Thermal Storage Systems**
4. **Chemical Storage Systems**
5. **Conclusion**

Glossary

aquifer thermal energy storage (ATES) ATES systems store thermal energy in aquifers (natural groundwater basins or layers at different depth). The system consists of two groups of wells with hydraulic coupling, one cold and the other warm. In winter, ambient air or surface water can be preheated before use by a heat exchanger circulating water from the warm well to the cold. In the summer, excess heat, solar energy, or waste heat can be stored by reversing the process.

compressed air energy storage (CAES) Off-peak electrical energy is used to compress air into underground storage reservoirs until times of peak or intermediate electricity demand. To generate power, the compressed air is heated and then passed through turbines.

duct/borehole thermal energy storage (DTES) DTES systems store thermal energy using vertical ground heat exchangers in drilling boreholes. The thermal energy is transferred via plastic tubes in boreholes typically of 50–200 m depth. A heat carrier fluid is circulated through the tubes to transfer the energy between source and sink.

final storage The storage of final energy such as heat, which is used by the customer to sustain continuous energy services.

grid-level power storage Consists of small and high-power storage systems with the capability to improve reliability, flexibility, and quality of electricity services.

latent heat storage Storage of the energy in special materials such as phase change materials (PCMs). The energy required to change the state of a unit mass of material from solid to liquid or liquid to gas is stored without changes in temperature. This energy is then released when the material reverts from gas to liquid or from liquid to solid.

mechanical storage Energy storage by mechanical processes such as rotation.

secondary storage Storage of secondary energy.

thermal energy storage Storage of heat or cold in different medias, such as PCMs, thermo-oils, or water kept in insulated tanks, caverns, pipes, or underground. Heat storage is a secondary storage type (tertiary, if heat is generated by electricity).

underground thermal energy storage (UTES) Storage of thermal energy underground using either underground storage tubes, aquifers with reduced storage volume needs, or man-made cavern storage systems with sealed storage volumes.

uninterruptible power supply (UPS) systems Combination of high-power, high-energy-capacity storage components that work together to ascertain supply and quality for a specific application.

Energy supply is an intricate task that provides reliable energy service to consumers throughout the year. Import dependencies, seasonal differences in energy supply and use, and daily fluctuations in consumption require sophisticated management of energy resources and conversion, or energy distribution and resource intermittency, in order to guarantee continuous energy services throughout all sectors. Energy storage thus plays a critical role.

Energy is mostly stored in between conversion steps from primary to secondary energy and secondary to final energy. Often, rechargeable systems are used to refill the storage capacity when energy demand is low and energy services are not needed. Primary storage, such as large crude oil and natural gas storage tanks, is essential for the proper functioning of a country's energy system. Water stored in reservoirs behind dams is valuable for selling hydropower electricity at the right time or during seasons of increased demand. Secondary or final storage systems, such as in tanks or in batteries, are crucial for emergency situations, uninterrupted industrial production, long-distance mobility, or to secure energy services at home.

 © 2004 Elsevier Inc. All rights reserved.

Storage systems are engineered to hold adequate amounts of mechanical, thermophysical, electrochemical, or chemical energy for prolonged periods of time. Energy storage systems should be quickly chargeable and should have a large energy storage capacity, but at the same time they should also have high rates of recovery and high yields of energy regain.

Final energy in factories or households is often stored in tanks as chemical energy in the form of heating oil or natural gas or as thermophysical energy in the form of steam, hot or cold water, or thermo-oils. For special applications or for safety reasons, energy may be stored electrochemically in batteries or physically in the form of pressurized air. Other storage systems are related to electricity and apply mechanical storage in the form of spinning turbines or flywheels, physical storage in the form of water in reservoirs in highland terrains, or electrostatic storage in supercapacitors.

1. INTRODUCTION

Several different types of energy storage systems are used for primary, secondary, and final energy. Primary energy is predominantly stored chemically in oil or gas tanks or in piles of specific energy carriers, such as biomass or coal. Secondary and final storage involve silos or piles of briquettes or biomass chips or storage tanks for refined petroleum oils, fuels, and gases. The conversion of these forms of energy into useful end-user energy provides for immediate energy services. If the demand or system status changes, useful energy can also be stored using mechanical or physical storage systems, thermophysical or thermochemical systems, electrochemical and electrostatic systems, or chemical storage systems.

For proper functioning of energy supply chains, storage concepts must be well developed to appropriately incorporate the storage technology into the energy system. Imbalances in a domestic energy system for heat, gas, and electricity can result in a country's increasing dependence on foreign gas and oil companies and expensive transboundary high-voltage electricity transmission lines. The development of inexpensive storage options may enable a country to cope with imbalances between demand and capacity and variable prices in energy markets. Uninterrupted availability of affordable clean energy products for both private and industrial consumers is important for society.

Energy systems rapidly change under the influence of energy policies. The ongoing liberalization or deregulation of energy markets may make small-scale and large-scale storage capabilities increasingly beneficial because of emerging competition in the energy market. However, increasing oil prices indicate structural weaknesses, with the increasing energy dependence on oil playing a major role in fuel consumption. Strategic petroleum reserves are still an important regulatory aspect of energy supply systems and compulsory storage ratios are negotiated between importing companies and the government. Adequate storage facilities must be available not only for crude oil but also for natural gas, which is expected to largely substitute oil in the near future.

Electricity and renewable energy technologies, increase variations in the grid and must be integrated properly into the existing electricity system. This involves local storage of electricity, which may require, among other things, installation of grid-level power storage capabilities.

Industrial and commercial customers increasingly demand power quality and reliability for functioning microprocessor systems. The spread of information technology brings new requirements for the transmission and distribution of electricity.

Uncertainties in grid reliabilities shift the emphasis in energy storage to guarantee energy service from producers to distributors. Distributed power generation and storage interconnected with the distribution grid can provide the needed power quality. In addition to energy supply reliability, incentives for energy storage include the moderation of increasingly volatile energy prices. Real-time pricing in liberalized energy markets is expected to result in additional local storage within buildings and "intelligent" household appliances.

To achieve full potential, both independent and grid-based power distribution systems rely on the availability of different storage technologies. These include safe on-vehicle storage and infrastructure components to ensure fuel availability at filling stations. Other examples are hybrid systems that combine microturbines, photovoltaic panels, or fuel cells for diverse industrial, commercial, or residential applications. These custom power systems must be electronically controlled within a complex grid containing multiple generating units. They require additional storage systems, such as batteries, superconducting magnetic energy storage (SMES), and flywheel energy storage, for power quality and reliability. In both small uninterruptible power supply (UPS) systems for personal computers and

large pumped storage projects, energy storage will increase system reliability. At utility sites, however, storage capabilities that provide large-scale instantaneous response are required. In the future, underground storage technologies including SMES may be possible. Innovations for storing hydrogen, recharging batteries or fuel cells, and facilitating high-temperature superconductivity have great potential to satisfy future heat and electricity storage requirements. In addition, new regulatory standards for efficiency and environmental compatibility require faster technical or technological innovations, such as heat-, corrosion-, abrasion-, and impact-resistant materials or thermal and chemical functional materials for thermal storage, insulation, and cooling media.

In the following sections, the basics of different types of energy storage systems are discussed for electricity, thermal, and chemical storage.

2. ELECTRICITY STORAGE

Apart from the use of secondary batteries, electricity storage is predominantly in the form of mechanical energy. Large-scale electricity storage systems are suitable to balance differences between power supply and demand. Power must be deliverable in time frames of hours and days or months to balance seasonal changes. Large-scale electricity storage systems supply peak-load power using previously stored baseload electricity or make it possible to integrate renewable energy technologies characterized by intermittent resource availability (e.g., solar and wind). In addition, they can supply regulating power for proper transmission and distribution of electricity through the grid. As part of the ancillary services, minimal power must be available within seconds or minutes. Sometimes, power levels must be sustained for hours. The following storage technologies are often used for large-scale electricity storage systems: pumped storage hydropower reservoirs, compressed air energy storage, and large-scale secondary batteries and large fuel cells.

Smaller scale applications of electricity storage provide power for mobile systems or upgrade power quality and reliability (including emergency supply). Electricity storage for these applications has varied capacity and timescale characteristics. The following technologies are applicable: batteries and fuel cells, SMES, supercapacitors and ultracapacitors, and flywheel energy storage.

Batteries are often used in mobile applications and systems in which reliability with low maintenance is important. Their release of power is instantaneous. Batteries can provide relatively prolonged power compared to capacitor and flywheel energy storage, which have different discharge characteristics, releasing high power in a short time frame. Capacitors and flywheels are suitable for improving power quality, integration into emergency backup systems, and peaking in mobile systems. Although in the testing phase, fuel cells are suitable for providing power over long periods of time and for continuous operation at various power ranges.

The capital costs of both power and energy are important in evaluating the economics of storage systems. Typically, the capital cost component of power is related to the output capacity, whereas the cost of energy is related to the storage capacity of the system (Table I).

2.1 Mechanical Systems

2.1.1 Hydropower Storage

An important application of hydropower storage plants is to balance seasonal differences in electricity demand. The reservoir stores excess water from water-rich, lower demand seasons and generates electricity during water-poor, higher demand periods. Pumped-storage hydropower plants follow a different strategy. Extra water is pumped back up to the reservoir using baseload electricity when electricity demand is low, usually at night, and is used for electricity generation during peak-load periods. The

TABLE I

Estimates of Power Capacity Cost and Energy Capacity Cost for Different Storage Systems[a]

Electricity storage system	Power capacity costs ($/kW/a)	Energy capacity costs ($/kWh/a)
Compressed air energy storage in tanks	120	100–1500
Underground compressed air energy storage	90	5–100
Large-scale batteries	70	10–700
Pumped storage hydropower	50	3–90
Superconducting magnetic energy storage	40	9–150
Flywheel energy storage	30	80–1200

Annual costs assume a discount rate of 9% and a 10-year life cycle.

[a] *Source*. EESAT (1998).

increase in electricity generation from renewable energy sources, such as wind and solar, requires additional storage due to its intermittent generation.

2.1.1.1 Hydropower Storage Plants Hydropower storage plants accumulate the natural inflow of water into reservoirs (i.e., dammed lakes) in the upper reaches of a river where steep inclines favor the utilization of the water heads between the reservoir intake and the powerhouse to generate electricity. This efficient storage of potential energy allows hydropower storage schemes a broader range of energy benefits than pure run-of-river schemes. Reservoirs at the upper watershed regulate the river downstream, which typically flows more evenly throughout the year, and the run-of-river power generated downstream utilizes part of the same water used upstream for the hydropower storage plant. Reservoirs can provide a seasonal, annual, or multi-annual storage capacity. Often several hydropower plants form larger hydropower schemes to follow demand along the load curves. Hydropower produces electricity preferentially during periods of higher load. The use of large storage volumes is often controversial because of considerable environmental impact, including the issue of pumping water from distant catchment areas to enhance inflow to the reservoir.

2.1.1.2 Pumped Storage Hydropower The most widespread large-scale electricity storage technology is pumped storage hydropower. A pumped storage plant uses two reservoirs—an upper storage basin providing the head to drive the hydropower turbines and another to collect water back into the upper basin using surplus baseload electricity during off-peak hours. The stored energy is only utilized for electricity generation during daily peak load. By International Energy Agency definition, a pure pumped storage plant uses less than 5% of inflow from upper watersheds and is therefore an energy system storage component, not a renewable energy source. Using "green electricity" (e.g., from surplus wind power) for pumping and using the additional storage volume for grid regulation and balancing are interesting possibilities for a future sustainable energy system.

Hydraulic turbines have a very high efficiency of more than 95%, whereas pumps are less efficient. The result is a maximum efficiency of approximately 85% for pumped storage hydropower plants. The effective efficiency is determined by the roundtrip cycle efficiency, which calculates the cycle with the turbine operated both as a generator and as a pump. Designs of pumped storage are optimized for overall efficiency and dynamic response. The effective efficiency is 65–70%, and for each kilowatt-hour (kWh) of electricity produced, approximately 1.4 kWh of energy is necessary for pumping, assuming the same architecture for both operation modes. Pumped storage is one of the most efficient ways to store electricity.

Economic performance is highly dependent on the price of the cheapest electricity produced by other plants and the variation in daily load. A pumped storage project will provide the most efficient and cheapest operation when it can provide a high water head and volume between its two reservoirs. This will allow the greatest amount of energy to be stored in the smallest volume of water, resulting in smaller pumps and turbine, thus reducing capital costs.

Appropriate location, equipment, and construction are particularly important in view of the high costs associated with pumped storage systems. In general, pumped storage sites with high heads and short water passages are more desirable (i.e., cost-effective sites with higher water head ranges between 300 and 800 m with relatively steep topography). Shorter water passages reduce the need for surge tanks to control transient flow conditions. With higher heads, smaller volumes of water provide the same level of energy storage, and smaller size water passages can be used for the same level of power generation.

The global pumped storage hydropower capacity is estimated to be 82,800 MW. Capital costs are likely to be in line with those for a conventional project (i.e., $1000–3000/kW). The payback time depends on the difference in value of peak and off-peak power. The availability of suitable sites in the future will limit growth in pumped hydrocapacity, with initial capital costs for construction increasing due to environmental obligations. In the longer term, the role of pumped storage plants is likely to increase to cope with intermittency of renewables.

2.1.2 Flywheels

Flywheel energy storage systems provide highly reliable, high-quality, uninterruptible electric power for communications networks, computers, the Internet, industrial manufacturing, commercial facilities, and distributed generation applications. Very heavy flywheel systems have been used for many years in electricity plants and smaller test systems have been used for mobile units.

The charging of a flywheel occurs when heavy symmetrical circumferential masses are rotated. The masses can rotate at approximately 50,000 revolutions per minute almost without friction around a titanium axle using magnetic bearings. The rotating masses are used in conjunction with a dynamo to generate electricity on demand.

The higher the angular velocity of rotation (ω radians per second) and the further away from the center the flywheel masses are placed, the more energy is stored. The relationship is $1/2I\omega^2$, where I, the moment of inertia of the flywheel, is the product of the total masses M and the square of their mean distance r from the center. Since the centrifugal forces are equal to $M\omega^2 r = I\omega^2/r$, for a given flywheel diameter and thickness, higher stored energy requires stronger flywheel materials to withstand the rotational forces. For materials such as fused silica, the specific energy stored in watt-hours per kilogram (Wh/kg) is more than 20 times better than that of steel-based flywheels and approximately 40 times better than that of lead–acid electric vehicle batteries (Table II).

Modern flywheel energy storage systems make use of advanced composite materials and state-of-the-art active bearing technologies. Single flywheel units are typically 5–100 kWh and can be cascaded to achieve utility-scale power and energy ratings with much larger systems being developed in various research laboratories. A 2-kWh composite flywheel system has been introduced and flywheels with an increased energy of up to 4 kWh are available. Plans include production of an 8-kWh model, which is to be followed by a 20-kWh unit. Medium-scale storage (0.1–10.0 MWh) has some unsolved technological problems. Experiments pairing microturbines with a fuel cell or flywheel system as a complete hybrid backup power system are under way.

TABLE II

Energy Densities of Storage Media[a]

Media	Wh/kg[b]
Hydrogen	38,000
Gasoline	13,000
Flywheel, fused silica	870
Flywheel, carbon fiber	215
Flywheel, steel	48
Lead–acid battery	25

[a] *Source*. Ristinen and Kraushaar (1999).

[b] Weight of motor and generator and conversion efficiency not included.

Ceramic superconducting materials may also be applied—for example, a flywheel running on frictionless bearings by means of magnetic fields generated by the superconducting materials.

2.1.2.1 Flywheels for Vehicles Electric buses or trains could use rapidly rotating flywheels for energy storage instead of secondary batteries. Urban transit buses and military vehicles have used stored inertial energy since the 1970s. Some impractical aspects, including a short range of only approximately 10 km, have limited their applicability.

Modern flywheels can store much more electricity than contemporary metal hydride or lead–acid batteries of similar weight or volume. Flywheels are environmentally friendly and have a much longer lifetime than electrochemical systems. High-performance materials enable the design of electric vehicles with flywheel systems comparable to petroleum-powered vehicles. Similar acceleration, speed, and range per unit weight of energy stored, including regenerative braking, are achieved. Hybrid vehicles using flywheel energy storage performed similarly to electric vehicles with supercapacitor or battery systems. One goal is to develop a zero-emission vehicle combining hydrogen lean-burn spark-ignited engines with high fuel economy or hydrogen fuel cells, with both incorporating flywheels or supercapacitors as acceleration and braking energy recovery systems.

2.1.3 Compressed Air

In compressed air energy storage (CAES), the release of stored pressurized air is used to generate electricity, most efficiently in conjunction with a gas turbine. In a CAES plant, compressed air is used to drive the compressor of the gas turbine, which uses 50–60% of the total energy consumed by the gas turbine system. The most important part of the CAES plant is the storage facility for compressed air. Usually, a man-made rock, salt, or porous rock cavern, created either by water-bearing aquifers or as a result of oil and gas extraction, is used. In contrast to the other systems, aquifers can be very attractive because the compressed air will displace the water, setting up a constant pressure storage system.

CAES plants show very high availability and reliability. CAES technology in the range of 25–290 MW has been achieved throughout the industrialized world, and there are proposed projects for an installed power of up to 1050 MW. The estimated annual investment costs for a CAES are $90–120/kW/year

(i.e., uniform cash flow), depending on air storage type.

The development of large-scale CAES is limited by the availability of suitable sites, with current research focusing on the development of systems with man-made storage tanks.

2.2 Electrochemical Systems

Electrochemical storage is characterized as the ability to convert chemical binding energy directly to electricity. The process can be reversed in secondary (rechargeable) batteries or accumulators to recharge the storage media. Only recently have large-scale applications of battery storage become of interest. Utility electricity storage requires a large battery system that can be repeatedly charged and discharged. Battery systems can release power virtually instantaneously, significantly improving the stability of an electricity network. Because of slow self-discharge and the aging processes, battery cells exhibit a decreasing capacity and recharging ability, which results in electricity storage for limited periods of time only. The potential advantages of battery systems are quietness, freedom from maintenance, reduction of gaseous emissions, and their independence of the primary energy carrier. Electric batteries also store direct current (DC) and can be easily integrated into domestic DC networks. An important system parameter is the charging (or load) factor, which relates to the necessary quantity of electricity in Ampere-hours for the batteries to become fully charged to the corresponding quantity released in the previous cycle.

The secondary battery is well suited for applications involving instant, relatively low-voltage power. Battery operation must be designed to be application specific and must consider factors influencing cost, which often limit energy and power per unit weight and volume (energy and power density, specific energy, and specific power) or durability (cycle lifetime and maintenance). Costs and performance of larger systems must be competitive with those of common mechanical storage systems (Table I).

2.2.1 Energy Storage in Batteries

The performance of a battery is characterized by thermodynamic and kinetic parameters that depend on the materials and media used and on battery design. Thermodynamically, large chemical-binding energy differences between the reaction partners, represented by a difference in Gibbs reaction energy, appear as cell voltage. To achieve a high specific energy (Wh/kg), the electrochemical reaction partners or electrodes should be as light as possible. Conventional lead–(sulfuric) acid batteries store approximately 200–500 times less enthalpy than petroleum fuel. Significant improvements have been achieved with nickel–metal hydride and lithium–ion batteries compared to older lead and nickel–cadmium batteries (Table III). The main requirements are improved lifetimes and replacement of heavy metals such as lead with much lighter and nontoxic materials. In addition to the energy storage capacity of a cell, the speed and quality of recharge are often important for efficient and attractive applications of battery systems. The time required to recharge a battery is strongly dependent on operating temperatures. Based on the solubility of reaction products or possible side reactions inside the cell, the recharging process can be inhibited or stopped completely after a few recharge cycles.

2.2.1.1 High-Energy Density Batteries Electrochemical power sources can be classified as accumulators (secondary cells), which store all corresponding reactive components inside, or as fuel cells with reactants stored outside that are fed to the cell stack when power is needed. The systems are also distinguished according to their operating temperature and electrolyte. The most important high-energy density battery systems are summarized in Table IV.

High-temperature secondary batteries and fuel cells compete to claim the electric vehicle markets. In recent years, the fuel cell has received a great deal of attention since it offers good performance in lower temperature operation and allows greatly improved electric vehicle ranges.

Regenerative fuel cells can be used to store electricity in chemical form and convert the stored energy back into electricity when needed. A regenerative fuel cell plant could store more than 100 MWh of energy and provide power for hours to several thousand homes. The development of regenerative fuel cells faces many challenges, particularly regarding pole reversal and recharging cycles. A new regenerative fuel cell design, the redox-flow battery, stores electricity in liquid (dissolved) form, allowing the maximum storage capacity to be easily increased by adjusting the electrolyte volume. The energy density is 20–30 Wh/liter, which is approximately 30% of a normal lead–acid battery. Redox-flow batteries are relatively large and heavy, which makes them suitable only for stationary applications. A large-scale application with a capacity of 15 MW/120 MWh has been built based on the Regenesys

TABLE III

Battery Systems Used in Applications[a]

System	Electrolyte	Type	Storage capacity (Wh/kg)[b]	Application
Lead–acid	Aqueous electrolyte (sulfuric acid), mostly immobilized	Often sealed (valve regulated), works at low temperatures, rechargeable	20–50	Starter battery for cars, batteries for forklifts, electric vehicles, submarines, and ships (also unsealed non-maintenance-free)
Nickel–hydroxide cadmium	Aqueous electrolyte, mostly immobilized	Sealed (valve regulated) rechargeable	20–55	Battery system mostly for small equipment and some for propulsion, on-board electronics, and on-site applications
Nickel–hydroxide hydrogen gas	Aqueous electrolyte, immobilized	Sealed, pressurized, rechargeable	50–60	Battery system for many satellites
Nickel–hydroxide metal hydrides[c]	Aqueous electrolyte, immobilized	Sealed (valve regulated), relatively high load capacity, rechargeable	50–80	Can replace nickel cadmium batteries
Sodium sulfur[d]	Solid electrolyte (β-alumina) at temperatures of 300–400°C	High-temperature system, rechargeable	90–120	Electric vehicles, trams, buses
Sodium nickel–chloride[d]	Solid electrolyte (β-alumina) at temperatures of 300–400°C	High-temperature system, rechargeable	90–100	Electric vehicles, trams, buses
Lithium–ion polymer, lithium cobalt oxide, lithium manganese oxide, iron, or aluminum disulfide	Organic electrolyte at room temperature	Relatively high voltage (4V), relatively high load capacity, rechargeable	~100	Battery system for small applications, tests for some electric vehicles
Zinc bromine[e]	Aqueous electrolyte	Rechargeable	65–70	Tests for some electric vehicles
Zinc air (oxygen from air)[e]	Alkaline or neutral electrolyte	Mechanically rechargeable (i.e., exchange of zinc electrode after a few cycles)	~170	Tests for some electric vehicles

[a] *Source*. Kiehne, H. A., Berndt, D., Fischer, W., Franke, H., König, W., Köthe, K. H., Preuss, P., Sassmannshausen, G., Stahl, U.-C., Wehrle, E., Will, G., and Williams, H. (2000). "*Batterien: Grundlagen und Theorie aktueller technischer Stand und Entwicklungstendenzen*," Expert Verlag, Renningen.

[b] When applied.

[c] Some nickel alkaline electrolyte batteries also work at low temperatures.

[d] Sodium–sulfur and sodium nickel–chloride have been abandoned for traction applications.

[e] Different battery system architecture (i.e., reactive substrates stored outside the cell).

regenerative fuel cell employing the bromide/tribromide and the polysulfide/sulfide couples in aqueous solution. Technical and economic data, according to National Power PLC, are given in Table V.

2.2.1.2 Battery Systems for Different Applications Although the end use of electricity is efficient when compared to the low energy conversion efficiency of gasoline, batteries still severely limit electric vehicle range. No battery has the energy density of gasoline, and battery recharging takes longer than refueling. Nonetheless, by providing a low-emission, low-noise, and clean mobility environment, battery-powered automotive systems for indoor use are increasingly being applied in organized transport systems in automated industries.

Electric buses for public transport and electric cars are becoming more popular for environmental

TABLE IV

Properties of Secondary Batteries and Fuel Cells[a]

Class	System	Type	Operating temperature (°C)	Electrolyte	Energy density (Wh/kg)[b]	Power density (W/kg)[c]	Lifetime charging cycles/years charged
Accumulators	Lead–lead oxide system	Pb/PbO_2	<45	H_2SO_4	161/35	80/160	750/>3
	Lithium–metal sulfide (oxide) system[d]	Li/TiS_2	Ambient	Polymer	473/100	35/400	300/1
	Lithium–iron sulfide system	Li/FeS_2	400	Salt melt	650/180	30/400	1000/0.9
	Zinc–bromine system	Zn/Br_2	60	$ZnBr_2 + H_2O$	430/55	70/80	400/2
	Sodium–sulfur system	Na/S	320	Ceramics	758/110	95/150	4000/4
	Sodium–nickel chloride system	$Na/NiCl_2$	300–400	Ceramics	650/82	60/140	2000/—
					Fuel	**Maximum power (W/cm^2)**	**Maximum operation time (h)**
Fuel cells	Low-temperature fuel cells	H_2/air	<100	Polymer	H_2 or MeOH	1500	1200 (1-kW module)
	Mid-temperature fuel cells	H_2/air	>100–200	H_3PO_4	Natural gas	150	16000 (several MW)
	High-temperature fuel cells	H_2/air	600	Carbonate melt	Natural gas, gasification	150	5000 (3-kW module)
	High-temperature fuel cells	H_2/CO/air	1000	Ceramics	Natural gas, gasification	300	40000 (cell)

[a] *Source*. Kiehne, H. A., Berndt, D., Fischer, W., Franke, H., König, W., Köthe, K. H., Preuss, P., Sassmannshausen, G., Stahl, U.-C., Wehrle, E., Will, G., and Willmes, H. (2000). "*Batterien: Grundlagen und Theorie aktueller technischer Stand und Entwicklungstendenzen.*" Expert Verlag, Renningen.

[b] Theoretical/experimental: The theoretical value accounts only for the weights of the reaction substrates. The experimental value accounts for a 5-h discharging time, except for the sodium–sulfur system, which has 1 h.

[c] Continuous operation/pulse operation (30 s).

[d] Other combinations include MoS_2, NbSe, V_2O_5, and CO_2.

reasons. As Lindly and Haskew note, complete electrification of the vehicle fleet of gasoline-fueled, light-duty, internal combustion (IC) engines, including the complete fuel cycle and based on 1997 Alabama electricity generation mix, would provide a 44% decrease in nitrous oxide (NO_x) emissions, an 18% decrease in CO_2 emissions, but a 14% increase in sulfur dioxide (SO_2).

In contrast to fuel cells, research and development of electrical vehicles with lead–acid, nickel–cadmium, nickel–metal hydride, or molten salt sodium β-alumina batteries has not achieved the required reduction in weight and volume to compete with conventional IC engines. An innovative concept is the hybrid vehicle with a small, efficient IC engine to drive the vehicle after the electric motor has accelerated the vehicle and to recharge its electrochemical system when idling. Recapture braking energy adds to the recharge, and with the reserve electric motor the idle time of the IC engine can be minimized, resulting in improved fuel efficiency.

Stationary fuel cells can also be used for public mobility as centralized electrical energy supply for trams and trolleys, underground (subway), light and conventional rail, and high-speed trains if combined with high-power storage systems such as superconducting magnetic energy systems (SMES) and ultracapacitors.

For large on-site immobile battery systems, mostly lead–acid- or nickel-based batteries are used in power plants and solar installations, signal transmission (telecommunications and radio) and UPS systems, backup systems, switchgear installations, annunciator and traffic lights, safety or emergency lighting, and air-conditioning installations. Because of the nature of the applications, high reliability is required. Less than a 0.25% failure rate has been observed with lead–acid and nickel–cadmium batteries in such systems.

TABLE V

Characteristics of Regenesys[a]

Power	15-MW maximum
Storage capacity	120 MWh
Footprint	<5000 m2
Storage tanks	Two tanks, 1800 m3 each
Total installed costs	$22 million
Lifetime	15 years
O&M costs	Relatively small (unmanned facility)
Storage time	10 h
Capital costs	$185/kWh, $50/kWh (mature technology)
Overall efficiency[b]	70%

[a] *Source*. *Renewable Energy World* (2000, January).
[b] http://www.electrosynthesis.com/news/w8content.html.

For telecommunication equipment, batteries need long operating times to balance electricity supply outages of several hours. For UPS systems, batteries must be able to instantaneously release very high currents over short time periods (minutes) until the backup system responds, which is comparable to starter batteries. Safety and durability are the most important criteria. In addition, sealed valve-regulated lead–acid batteries are used because of their low maintenance requirements.

Solar thermal and photovoltaic (PV) applications are always associated with battery subsystems to balance wide fluctuations of input and output currents. To determine the optimal system dimensions for the specific storage application, the average electricity consumption and the diurnal consumption profile should be known. Other important parameters include those of the PV cells, which must be adapted to the seasonal variation of solar radiation. Use of grid-dependent PV cells reduces the intermittent nature of solar energy because the grid and small-scale electricity storage devices are employed as a backup system. Other solutions to cope with the seasonal differences combine grid, PV, and liquid natural gas (LNG) burners into a coordinated hybrid system for buildings. Table VI lists storage applications for PV supply technologies.

2.2.1.3 Batteries and the Environment The increasing use of secondary batteries, especially lead–acid or nickel–cadmium batteries, means that hundreds of tons of toxic heavy metals are discharged yearly into the environment. Potential environmental impacts from their improper disposal have been recognized as a problem worldwide. In

TABLE VI

System Power Range for PV Storage Applications[a]

Power range	Typical application	Examples
μW	Integrated transistors or chips with minimal energy consumption	Solar watches and calculators
mW	Equipment and installation with low energy demand and only periodic use	Portable receivers/transmitter, automatic devices such as ticket or vending machines, fire or security alarm systems
W	Installations for communication and measurement, small consumer households, or businesses	Devices on buoys or for TV, radio, and meteorological stations, electricity supply on boats or in vacation homes, electricity supply for heat pumps
kW	Stand-alone grids with electric equipment and installations	Remote estates, military

[a] *Source*. Kiehne, H. A., Berndt, D., Fischer, W., Franke, H., König, W., Köthe, K. H., Preuss, P., Sassmannshausen, G., Stahl, U.-C., Wehrle, E., Will, G., and Willmes, H. (2000).“*Batterien: Grundlagen und Theorie aktueller technischer Stand und Entwicklungstendenzen.*” Expert Verlag, Renningen.

Europe, the 1992 Government Action Plan for Waste and Recycling and the 1992 Waste Management Plan for Municipalities consolidated the framework of the Environmental Protection Act of 1994. In the United States, the 1996 Mercury-Containing and Rechargeable Battery Management Act (the battery act) promoted the collection and recycling or proper disposal of used nickel–cadmium and lead–acid batteries, clear labeling of batteries, and education of the public about recycling and disposal. Another goal is to phase out the use of mercury in batteries.

The use of nickel–cadmium batteries is proposed to be phased out of a wide range of household appliances or at least be restricted, for instance, to power tools requiring very rapid discharge. Another measure is to lower the threshold amounts of cadmium in batteries in a stepwise manner, which will ultimately require a change to alternative storage systems. Alternative secondary battery systems are lithium and nickel–metal hydride batteries.

The development of fuel cells for cars, trucks, and buses will reduce the environmental impact of mobility, even if natural gas supplies the fuel cells during transition stages. True "green mobility" may be achieved only if hydrogen supply and storage technology are developed using hydrogen generated from regenerative sources.

2.2.2 Ultra- or Supercapacitors

Capacitors are energy storage devices that use charge as a storage mechanism. A conventional capacitor with charge $+Q$ on one plate, $-Q$ on the other, and voltage V between them has a stored energy equal to $1/2\,QV$. Almost all capacitors use a dielectric material between the plates, which optimizes charge distributions. Dielectric capacitors can provide power densities many times more than 1 kW/kg and have very long cycle lifetimes, whereas batteries provide less than 0.2 kW/kg and have fewer cycles (generally 100–1000). However, dielectric capacitors have energy densities less than 1 Wh/kg, whereas batteries can provide more than 100 Wh/kg. The applicability of dielectric capacitors is therefore clearly limited to specific tasks requiring high power at high voltages over very short time periods (e.g., for lasers). Superconducting inductive systems have similar limitations.

Supercapacitors (scaps) or electrochemical capacitors (ECs), which have 500,000 times larger capacitance than conventional capacitors, can provide more than 1 kW/kg power output combined with more than 5 Wh/kg specific energy and a high cyclicity of thousands of recharge cycles. Scaps, also known as ultracapacitors or electrochemical double-layer capacitors, consist of carbon or other high-surface-area materials as a conductor and an electrolyte. Applications include acceleration power for electric and hybrid vehicles, electrical regenerative braking storage, power trains, starting power for fuel cells, pulse power for mobile and wireless telecommunications, and consumer and industrial electronic devices.

In contrast to batteries, scaps must be associated with a power supply system. They can be used as energy buffer banks to minimize the constraints of power supply. If coupled with batteries, they can reduce the peak power requirement, prolong battery lifetime, and reduce the energy requirement (or the size) of the battery. The batteries provide the energy, with scaps providing the instantaneous power needed. Depending on the configuration, scap systems can be used both as energy source and as buffer, or they can be applied as a load equalizer for elevators. Scaps work in PV-powered systems or stationary industrial power applications; as low-maintenance, solid-state alternatives to batteries for short-term bridging in UPS systems; or for peak-load buffering to increase efficiency and reduce the size and cost of stationary systems. As material costs decrease, low-cost scaps are expected to become the standard option of energy storage systems for consumer electronics, industrial, and transportation markets.

2.3 Electromagnetic Systems: Superconductivity

Superconductivity offers the ideal way of storing electric power directly. Theoretically, by using supercooled inductive coils, much higher storage capacities are achievable than in conventional capacitive or inductive systems.

Superconductors are being considered for SMES, in which electric energy is stored by circulating a current in a superconducting coil without resistive losses. Niobium–titanium alloys are used for storage at liquid helium temperatures (2–4 K). The use of superconducting materials with higher critical temperatures (e.g., 60–70 K) requires additional research and development. The efficiency of charging and discharging is very high because no energy conversion is involved; for this reason, SMES can respond rapidly, with a response that is limited by the time required for conversion between DC and AC power. Rapid response within seconds makes SMES attractive for load leveling and for frequency and voltage

control of the electricity grid while permitting a compact system design. Only SMES with relatively small storage capacities are commercially available. New SMES developments target storage capacities of approximately 100 kWh.

Large-scale applications depend on achieving high critical current densities (Jct) to be cost competitive with other technologies. High power and massive magnetic fields are involved, so careful consideration must be given to the consequences of an accidental loss of coolant; 10 MWh of stored energy released is equivalent to 1.18 tons of TNT.

SMES technology with 90% efficiency for storage of excess nighttime electrical energy is comparable to water-storage technology with efficiencies of 65–70%. Technological development, however, requires time and large monetary investments.

3. THERMAL STORAGE SYSTEMS

The energy content E of a thermal energy storage volume V is proportional to the temperature before (T_1) and after recharge (T_2). That is, $E = c_p(T_2 - T_1)V$, where c_p is the specific heat per unit volume—a storage media- or material-specific coefficient. Water has a high specific heat of 1 kcal/kg · °C compared to steel, which has a specific heat of 0.1 kcal/kg · °C. The transferred energy involved in temperature changes of materials is called sensible heat, whereas the energy involved in changing the phase of a material is called latent heat. Depending on restrictions of volume or pressure changes in gaseous or liquid storage media, the specific heat of constant volume C_v or of constant pressure C_p are important parameters in the design of efficient thermal energy storage systems.

Heat and cold storage systems are important for short-term and long-term (seasonal) storage. One of the benefits of short-term storage is to shift electric resistance heating to off-peak hours. Environmental benefits include the possibility of using renewable energy sources to increase the demand coverage. The benefit of long-term storage is that by balancing seasonal discrepancies in supply and demand, energy consumption can be significantly reduced.

The following systems exist for short-term storage:

- Water, which is the standard medium for heat storage
- Ice (phase change at 0°C, latent heat provides large energy density)
- Phase change materials (salt hydrates), which allow phase change at other temperatures besides 0°C are currently in the market introduction phase
- Sorption systems (e.g., H_2O/SiO_2 and silica gel), which are in the experimental stage, having higher energy densities are more suitable for long-term storage.

Other relevant parameters for thermal systems are energy losses as a percentage of the retrieved thermal energy, which depends on energy density, reflux temperatures, insulation quality, surface-to-volume ratio, and the duration of storage. The energy density (kWh/m^3) depends on the temperature of the media and its specific heat capacity. The cost of thermal systems depends on volume and energy stored (Table VII).

Water as thermal storage media has a low energy density and consequently needs larger volumes, with correspondingly large heat losses. The surface-to-volume ratio should be as small as possible, especially for long-term storage. The recovery efficiency of heat from thermal storage depends on the reflux temperature of the heat extraction system. Improvements are possible using heat pumps, although this is more costly.

Technological advances in phase change materials (PCMs) can increase the thermal mass per unit floor area of buildings and improve the efficiency of thermal systems. Adsorption systems have great potential as seasonal storage applications of solar thermal energy for single homes, for which less cooling is required. The heat coverage with solar thermal energy could technically reach 90–100%.

Long-term storage is often combined with solar thermal systems. Soil–heat exchangers integrated in pile foundation, soil, or groundwater systems (aquifers) are used. The energy used for circulating the storage media water through the system is determined by the permeability of the underground

TABLE VII

Overview of Short-Term Storage[a]

	Energy density (kWh/m^3)	Energy losses (estimate) (%)	Cost (<10 m^3) ($/m^3)	Cost (>100 m^3) ($/m^3)
Water	50–60	20	2000	700
Phase change material	50–120	12	2700	1300
Adsorption	200–250	5	6800	4500

[a] *Source*. Collinson (2001).

material and the system's depth. The corresponding net thermal energy retrieved, the size, and the temperature levels of the system are the most important characteristics. The costs are strongly determined by the required pumping capacity (m^3/h) and power capacity (kW) and less by the size of the system (Table VIII).

Underground soil or groundwater systems are very productive for cooling, when improved building construction, (automated) shading systems, highly efficient lighting systems, and electric appliances reduce cooling demand to a minimum. Combined with energy conservation strategies, cooling systems can meet the entire cooling load in most office buildings. In winter, the same system can be used to preheat the cold air drawn in for fresh air supply in office buildings, reducing heating demand by approximately 30%. Seasonal underground storage systems can even provide coverage of up to 70% space cooling in industrial buildings with cooling capacities of 40–70 W/m^2.

3.1 Water for Thermal Energy Storage

3.1.1 Thermal Storage Tanks

Centralized water thermal storage is by far the most common form of thermal energy storage. Usually, large hot-water storage tanks are buried underneath large infrastructure components such as athletic fields and parking garages. Conventionally welded steel tanks, reinforced concrete, or wire-wound concrete tank systems are used with capacities of 1 million gallons or more. Heat storage capacities range from 60 to 80 kWh/m^3.

Low heat-conducting tanks ($< 0.3\ W/m^2\ K$) have been developed in Germany using fiberglass composite materials. With volumes between 5 and 10 m^3, these innovative tanks are applicable for buffer and short-term storage in the range of hours to days. For storage in the range of weeks, volumes of 50–100 m^3, and for long-term or seasonal storage, volumes of 500–6000 m^3 can deliver thousands of ton-hours of hot water. The cost of large tanks (6000 m^3) is approximately \$100/m^3. Such tanks can also be used to store cold water.

Other thermal storage systems also used for cold-water storage are based on stratified storage underground tank systems combined with linear diffuser systems to reduce inflow turbulences. The different temperatures contained in a thermally stratified storage tank hold the density variation of warm and cold water. Thermally stratified storage tanks represent an effective technique that is widely used in energy conservation and load management applications. It is commonly applied in combination with solar energy or waste heat reuse systems. Thermal stratification improves the efficiency of storage tanks because heat at intermediate temperatures can be used to heat colder layers. System parameters include various inlet and outlet locations, velocities, temperatures, and heat loss. Heat conduction at tank walls and ambient heat loss through the cover strongly affect the stability of thermal stratification.

Long-term heat storage in combination with waste heat distribution, solar, or combined heat power (CHP) systems is also constructed using man-made water-permeable gravel pits insulated and covered with heat-damming materials, such as polypropylene foil, vacuum insulation foil, and glass granulates. The storage temperature is approximately 90°C. Heat storage capacity is 30–50 kWh/m^3.

Short-term or buffer heat storage systems are ubiquitous for generation of hot water in heaters or boilers. The size of decentralized heat storage systems depends on the water boiler output power needed to satisfy hot water and heating requirements. A minimum of 25 liters storage volume per kilowatt boiler output is needed; for more comfortable operation, 70 liters/kW is recommended. Decentralized thermal storage heater/boiler systems mostly use electrical night currents for electrical resistance heating. An important aspect of the short-term storage concept is maintaining thermal stratification. The performance of storage-type domestic electrical water heaters is thus greatly improved and energy conservation optimized. Designs are dependent on draw-off rates (usually 5–10 liters/min), with the placement of heating elements and residence time being important design parameters. Optimized ratios of tank size and draw-off rates together with better

TABLE VIII

Overview of Seasonal Storage[a]

	Pumping capacity (m^3/h)	Power capacity (kW)	Pumping energy (% of useful energy)	Cost (\$/m^3/h)
Soil heat exchangers	0.5–50	5–450	5–8	1800–4100
Groundwater system	10–500	100–4500	5–8	1600–2700

[a] *Source*. Collinson (2001).

thermal stratification inside the heater storage tank result in higher discharging efficiencies.

3.1.2 Large Underground Thermal Energy Storage Systems

Building, industrial, and agricultural sectors require much larger thermal storage dimensions. There are three forms of underground thermal energy storage systems (UTES):

- Thermal underground storage tubes (e.g., soil–heat exchanger), which require large storage volumes to improve storage efficiency, achieving 15–30 kWh/m^3
- Aquifer storage systems with reduced storage volume but site-specific requirements, achieving 30–40 kWh/m^3
- Man-made cavern storage systems with confined storage volumes, achieving 40–60 kWh/m^3.

Early adoption of energy storage in standard project designs is essential to improve energy efficiency. In nonsealed UTES systems, heat is normally stored between 15 and 40°C, whereas cold storage areas are between 5 and 25°C. Depending on the size of the area used, heat/cold storage capacities of up to 5 or 6 GWh per year are achievable. Newer concepts of UTES at approximately 70–90°C provide more economically efficient and environmentally benign energy systems with reliable long-term operation. Other types of UTES systems are aquifer thermal energy storage (ATES) and duct/borehole thermal energy storage, which can store waste heat from municipal facilities in summer and provide domestic hot water and space heating in winter.

Chemical aspects of thermal energy storage in aquifers must be investigated to solve clogging and environmental heat exchange problems. For example, at high temperature ATES consisting of vertical wells with horizontal drains, major problems can occur at the heat exchange interface where large amounts of calcite form. Developing and testing site-specific groundwater treatment methods involving chemical, microbiological, and environmental effects ensure required environmental standards. Adequate scaling of ATES improves reliability and avoids or reduces potential ecological impacts.

The implementation of heat/cold storage applications as a standard design option improves effective storage control and operating strategies. This results in increased energy efficiency and cost-effectiveness in building and industrial fields. A combined heat/cold ATES installed in Berlin in a new government building complex uses two separate aquifers at different levels, one at approximately 60 m for cold storage and the other at approximately 300 m for heat storage. The storage system is coupled with a heat pump and, in case of extra load in winter, with a CHP unit running on biofuels. At peak operation, this ATES system should provide 16-GWh heat and 3-GWh cooling capacity.

Man-made cavern storage systems are placed in leak-tight granite, gneiss, or hard sediment rock. Caverns can store heat at temperatures of approximately 90°C due to low heat-conducting properties. The cavern storage system of Lyckebo in Sweden has a storage volume of approximately 100,000 m^3 and a heat storage capacity of 5.5 GWh.

3.2 Latent Heat/Phase Change Systems

Although it is inexpensive, readily available, and possesses excellent heat transfer characteristics, water is limited as a viable storage media because of the large volumes needed to achieve adequate heat storage capacities. When space is limited, phase change systems are applied to make use of latent heat.

The oldest PCM in thermal storage is ice. Ice is used in glycol static ice systems with phase change at 0°C and an energy density of approximately 90 Wh/kg for cooling. Glycol static ice systems are costlier than cold-water systems but require less space. Heat storage at various temperatures can also be designed using other PCMs (Table IX). The overall temperature required to melt the PCM must be significantly less than the output that the heating system provides, making compounds with lower melting points more advantageous.

Organic PCMs are more expensive than inorganic salts but can be custom made to provide transition temperatures specific for a given application. Compared to inorganic salts, organic PCMs are nontoxic, noncorrosive, nonhygroscopic, more stable during repeated phase change cycles, and compatible with most building materials. PCMs have higher latent heat per unit weight, melt without leaving solid phase sediments, and, most important, show negligible supercooling tendencies. Although their potential flammability is a disadvantage, organic PCMs are used to absorb solar heat for air heating in buildings. Eutectic PCMs behave as salt hydrates but can be custom made to meet specific melting points. Costs are considerably higher than those of organic or inorganic PCMs.

PCM modules consist of a series of PCM-filled plastic pipes embedded in insulation material in order

TABLE IX

Transition Temperature and Latent Heat Energy of Common Groups of Phase Change Materials (PCMs)

Group	PCM type	Transition temperature (°C)	Latent heat (Wh/kg)	Thermal conductivity, liquid/solid (W/mk)
Inorganic salt hydrates	Calcium chloride	27–30	47–53	0.54[a]/1.09[b]
	Sodium sulfate (Glauber's salt)	32	70	—
	Zinc nitrate	36	41	0.46[c]/—
	Magnesium nitrate	89	45	0.49[d]/0.61[e]
Organics	Polyglycol E400	8	28	0.19[f]/—
	Polyglycol E600	22	35	0.19[f]/—
	Octadecane	28	68	—
	Ecosane	37	69	—
	Paraffin 116 (paraffin wax)	48	58	—
	Paraffin 6403 (paraffin wax)	62–64	48–53	0.17[g]/0.35[h]
Fatty acids	Palmatic acid	63	52	0.16[i]/—
	Capric acid	32	42	0.15[j]/—
Organic/inorganic mixes (eutectics)	Mystiric acid (mainly inorg.)	54	52	—
Aromatics	Naphtalene	80		0.13[k]/0.34[l]

[a] 38.7°C.
[b] 23°C.
[c] 39.9°C.
[d] 95°C.
[e] 37°C.
[f] 38.6°C.
[g] 63.5°C.
[h] 33.6°C.
[i] 68.4°C.
[j] 38.5°C.
[k] 83.8°C.
[l] 49.9°C.

to exchange heat with heating systems. In combination with thermal stratification tanks, flat PCM modules with different melting temperatures enhance efficiency because heat conduction at surfaces is eliminated and the stability of the stratification is improved. This method can be implemented with larger centralized thermal storage tanks.

3.3 Other Thermal Storage Systems

A different thermal storage system is based on the thermophysicochemical properties of assorted chemical compounds. For example, cooling is achieved in chemical heating/cooling pipe systems by the endothermic heat of combining solid– or liquid–liquid mixable materials. In solid–liquid pairs, urea–water and ammonium nitrate–water generate endothermic heat of 8.2 kW/kg when the materials are mixed. Similarly, isobutanol–acetonitrile and 1-butanol–acetonitrile produce 0.8 kW/kg. For regeneration, waste heat of 80°C can be used to redistill the mixtures. Evaluation of the most suitable mixing pairs in terms of cold/heat capacity must be implemented for each specific application.

Innovative latent heat storage systems have been developed for ventilation systems using molecular adsorption and desorption on chemical surfaces. With its extensive surface areas, zeolites are ideal materials that adsorb water vapor, effectively releasing a large amount of adsorption heat. This discharging process can heat inlet air from 25 to 100°C by streaming at 6000 m^3/h through 7 tons of zeolites and provides a heating power of 95 kW for 14 h per day. During low heat demand, the charging process takes in airstreams of 130°C in order to desorb water molecules from the zeolite surfaces, releasing a humidified airstream at 40°C. The storage capacity is approximately 150 kWh/m^3 of zeolite material.

4. CHEMICAL STORAGE SYSTEMS

The energy density achieved with chemical storage is approximately a factor 100 higher than that of any other storage technology discussed previously (Table II). This decisive advantage gives chemical storage systems unsurpassed competitiveness. Primary energy is predominantly stored chemically in large crude oil and natural gas storage tanks or as stockpiles or in silos of specific energy carriers, such as coal or biomass. Secondary or final chemical storage involves refined petroleum products kept in tanks of fuel stations, industries, vehicles, and households. Energy products from coal and biomass, such as briquettes, granulated coal, wood chips, or biomass pellets, are kept dry in sheds, cellars, or silos.

4.1 Coal

Coal is a combustible mineral formed over millions of years through the partial decomposition of plant material subject to increased pressure and temperature under the exclusion of air. Primary coal exists in different grades called anthracite, bituminous, and lignite, with decreasing heating values, respectively. Secondary coal is produced from primary coal into metallurgical coke, anthracite, bituminous, and lignite briquettes. World coal reserves are approximately 9.1×10^{11} tons, with annual consumption of 4.5×10^{9} tons.

In 1999, coal production was approximately 5 billion tons worldwide (Table X). Although production stagnated at this level, coal mining productivity has steadily increased during the past decade.

The key environmental drawback of coal mining is the continuous release of CO_2 due to the glow burn of coal floss. Glow burn is impossible to control because oxygen enters the vast underground coal reservoirs during mining activities, leading to self-ignition. Coal combustion releases particulates and the chemical compounds SO_2 and NO_x, which cause environmental acidification. Pollution abatement equipment enables modern plants to control these emissions. Together with other essential equipment for coal receiving, raw coal storage, coal preparation, coal transportation, and coal storage, this results in significantly higher capital investments.

Coal stocks are kept at electric utilities or at coke and power plants and other industrial facilities. At coal-handling facilities, crushers and granulators use coal delivered from different coal pits to produce coal fragments of approximately 50 mm. The raw coal for power plants or boilers is stockpiled or stored on site in coal silos. The coal is then crushed into powder in pulverizers to feed the plant's burners. Coal storage facilities consisting of coal bunkers, silos, or coal strait or stockpile slot systems allow coal from the different pits to be blended and provide short-term storage to accommodate peak shipping. Customized modeling solutions to deal with high fugitive dusting potentials, storm run-off collection, and settling ponds at ground storage facilities and railroads are necessary. Environmental quality control is maintained using air quality modeling and monitoring, including particulate source inventories, weather observation, and forecasting. In addition,

TABLE X

World Coal Production, 1999[a]

	Thousand short tons			
Region/country	Anthracite	Bituminous	Lignite	Metallurgical coke[b]
North America	4768	1,086,789	100,070	26,185
Central and South America	28	49,457	0	10,694
Western Europe	10,308	104,661	380,612	43,499
Eastern Europe and Russia	35,576	475,221	235,399	66,454
Middle East	0	1389	0	335
Africa	2881	324,294	0	5354
Asia and Oceania	298,961	1,757,120	185,329	214,601
World total	352,522	3,798,932	901,410	367,122

[a] *Source*. Energy Information Administration (http://eia.doe.gov).

the clean coal technology initiative has led to pollution control and power-generating processes using low sulfur subbituminous coal that reduces air emissions from coal-burning plants, including greenhouse gases. Newer technologies have the potential to utilize high-sulfur "dirty" coal and still achieve comparable air emissions. Enhancing power plant efficiency and capturing carbon gases further reduce CO_2 emissions.

4.2 Crude Oil

World oil reserves are estimated at 1.6×10^{14} liters (1×10^{12} barrels), with potentially comparable amounts of undiscovered oil resources. World consumption is 1.2×10^{10} liters per day. Current oil production is predicted to increase by at least 1 million barrels every year over the short term, with oil reserves remaining economically exploitable for the next 25 years. Prices are expected to increase until 2040 due to the use of costlier new exploitation technologies.

Having adequate storage capacities helps at least to balance short-term price fluctuations. In the long term, however, profitable returns on the hefty investments of oil firms may only be attainable if oil prices steadily rise and additional capital is made available at the right time. Efforts to develop a valid pricing system of as scarce a resource as crude oil seem unattainable due to many uncertainties, including unstable political scenarios. OPEC's current levels of oil production and other factors are major uncertainties impacting price stability. Storage alone provides no answer for these long-term uncertainties.

In the short term, simple market dynamics can utilize primary storage of crude oil to replenish low fuel stocks during mild winters when demand is low for heating oil, propane, and natural gas. In addition to the importance of primary storage for the functioning of the oil market, efficient secondary storage of petroleum products such as gasoline, diesel, jet fuel, other (mostly cracked) oil distillates used for heating, and other residual oil stocks (light and heavy oil) is crucial for corresponding energy services (Table XI).

Many countries regulate oil imports with the use of strategic petroleum reserves (SPRs), with compulsory storage ratios stipulated between importing companies and the government. Large consumer companies can join the contracted storage system under the obligation not to empty their stock below a given ratio. SPRs play a key role in stabilizing price fluctuations in the international energy markets.

Depending on the annual demand and capital cost of storage and the aging of oil, different amounts of stock are kept by the consumer companies. Aboveground and underground storage tanks are used for secondary storage of high- or low-sulfur distillates. The total fuel distillate ending stock in the United States was approximately 140,000 Mbbl in 2001, with approximately 40% high-sulfur and 60% low-sulfur distillates.

TABLE XI

Example of Crude Oil and Petroleum Storage: Product Stocks[a] from the United States in Mbbl (1×10^6 Barrels)[b]

Year/product	December 2001 (Mbbl)[c]
Crude oil[d]	312,000
Total motor gasoline	209,000
Reformulated	*44,900*
Oxygenated	*400*
Other finished	*116,200*
Blending components	*48,400*
Jet fuel	42,000
Distillate fuel oil[e]	144,500
$\leq$0.05% sulfur and under (low)	*82,300*
$>$0.05% sulfur (high)	*62,200*
Residual fuel oil	41,000
Unfinished oils	87,700
Other oils[f]	199,100
Total (excluding SPR)	1,036,100
Crude oil in SPR[g]	550,200
Total (including SPR)	1,586,300

[a] Product stocks include domestic and customs-cleared foreign stocks held at, or in transit to, refineries and bulk terminals and stocks in pipelines. Stocks held at natural gas processing plants are included in "other oils" and in totals. All stock levels are as of the end of the period.

[b] From the Weekly Petroleum Status Report, Energy Information Administration (http://eia.doe.gov).

[c] Data may not add up to total due to independent rounding.

[d] Crude oil stocks include domestic and customs-cleared foreign crude oil stocks held at refineries, in pipelines, in lease tanks, and in transit to refineries. These do not include those held in the Strategic Petroleum Reserve (SPR).

[e] Distillate fuel oil stocks located in the Northern Heating Oil Reserve are not included.

[f] Included are stocks of all other oils such as aviation gasoline, kerosene, natural gas liquids, and LRG's, other hydrocarbons and oxygenates, aviation gasoline blending components, naphtha and other oils for petrochemical feedstock use, special naphthas, lube oils, waxes, coke, asphalt, road oil, and miscellaneous oils.

[g] Crude oil stocks in the SPR include non-U.S. stocks held under foreign or commercial storage agreements.

4.2.1 Oil Tanks and the Environment

Oil storage tanks, some with capacities of more than 390,000 barrels (more than 16 million gallons), store crude oil, intermediate stocks (partially refined), finished products, and chemicals. To avert soil and water pollution, containment areas must be designed to hold their capacities in the event of a tank rupture. At refineries, containment systems also divert rainwater runoff to the refinery's effluent treating facilities to ensure that no traces of chemicals or hydrocarbons enter local waters. Processing crude oil can lead to methane venting, a potent greenhouse gas. To reduce methane emissions, liquid storage facilities are consolidated and centralized in the field. From an environmental perspective, all storage tanks should be impermeable to volatile organic compounds (VOCs) and should be equipped with pressure-vacuum valves combinable with a vapor recovery system. Emissions of gas built up in crude oil or produced in water storage tanks are controlled during handling by flare combusting of the VOCs or by tank battery vapor recovery systems. In order to meet air pollution standards, petroleum products are refined to sulfur concentrations as low as 5 parts per million (ppm). The gasoline additive methyl-tertiary-butyl ether must be eliminated gradually due to its persistent properties if leaked into water systems. Tens of billions of dollars are required for replacement and remediation of contaminated groundwater and soil, due to leaking petroleum hydrocarbon tanks in the United States during the early 1990s from secondary and final storage (mostly heating oil and gasoline tanks).

4.3 Natural Gas

World reserves of natural gas are estimated at $1.4 \times 10^{14}\,m^3$, with gross production of gas at $2.4 \times 10^{12}\,m^3$ per year. Natural gas hydrates from deposits in ocean floors contain vast potential, which currently cannot be exploited economically.

Seasonal fluctuation in the consumption of natural gas varies as much as 50% with changing weather conditions and heating and electricity demand. In the short term, more extensive variations can be observed. The volume of peak demand is unpredictable and the exact timing impossible without integrated storage systems. Because natural gas production does not correspond to these fluctuations, surplus gas is injected during low demand into numerous, mostly underground storage facilities for use during high-consumption periods. Underground storage systems are located in porous rocks of depleted fields or aquifers or in artificial underground caverns, such as abandoned mines or salt caverns. Well-functioning storage systems eliminate the need for expensive, additional pipeline transmission capacity that would be necessary to supply peak demand. In addition, they balance summer production levels with possible supply shortages in the winter.

A gas pipeline network connects supply and demand sides with corresponding storage sites. Storage facilities are located near consumption areas to reduce the occurrence of bottlenecks within gas transmission grids due to gaps between transportation capacities and withdrawals. Natural gas is stored in two basic ways: It is either compressed in tanks as LNG or stored in large underground storage facilities (Table XII). LNG is kept as a chilled and pressurized liquid in double-walled cryogenic vessels or tanks. For transportation by ships or trucks, LNG requires 600 times less storage space than its gaseous form, and its ignition potential is reduced. Natural gas kept in tanks as compressed natural gas is used as vehicle propellant in combustion engines. Volumetric capacity depends on the head and shell thickness of the storage vessel.

Increasing demand, changing demographics, and the deterioration of storage and pipeline systems pose significant technological challenges for the future. Natural gas usage is projected to increase at a rate of 2% per year for the next two decades due to increased substitution of petroleum, diesel, or oil. A significant shift in consumption patterns is expected as more natural gas is used for heating, cooking, and mobility along with the decentralized generation of electricity.

Although natural gas vehicles (NGVs) have penetrated niche commercial markets for buses, delivery vans, and trucks, their success is severely constrained by the cost, weight, and size of existing natural gas storage cylinders. To counter this, a state-of-the-art, safety-certified natural gas storage unit called the integrated storage system (ISS) was developed. The ISS utilizes lightweight, high-density

TABLE XII

Example of Natural Gas Storage[a]

Traditional storage	Salt caverns	Total
5927	145	6072

[a] *Source*. Primary storage in the United States for 2000 in volumes of billion cubic feet at 1 bar pressure and 15°C. Energy Information Administration, *Natural Gas Monthly* 2001 March (http://www.eia.doe.gov/oil_gas/natural_gas/nat_frame.html).

polyethylene in the construction of small-diameter cylinders (pressure cells), which are subsequently encapsulated within a high-strength fiberglass shell containing impact-absorbing foam. The total energy storage capacity of 45 liters of gasoline equivalent at a service pressure of 250 bar enables a NGV to travel more than 480 km before refueling. This is comparable to the distance traveled by vehicles powered by gasoline. ISS technology reduced the cost of NGV storage tanks by approximately 50%. Its weight savings is up to 70% compared to steel and aluminum cylinders, and ISS does not interfere with the car's crumple zone. For commercial viability, the cost of storage must be reduced.

4.4 Hydrogen

Hydrogen is mainly used as chemical feedstock in the petrochemical industry and in the food, electronics, and metallurgical processing industries. However, hydrogen is believed to have great potential as a major energy carrier for clean energy systems of the future.

Hydrogen can be used in transportation, construction, utilities, and industry as a chemical storage compound. It can be used to mitigate CO_2 emissions by upgrading carbon-rich fuels to less carbon-intensive fuels. Renewable energy resources may be used to extract hydrogen from water, which is then combusted by fuel cells, for example, reforming water and balancing the intermittent nature of renewables. Ideally, combusting hydrogen generates no atmospheric emissions other than water.

For vehicles, hydrogen would be ideal as a synthetic fuel because it is light and highly abundant. The hydrogen fuel could be stored in carbon composite tanks at pressures of up to 250 bar, providing fuel for more than 200 km. Provenzano *et al*. estimated that if a hydrogen combustion engine operates in lean burn mode, the vehicle produces no CO and only low NO_x emissions (10–100 ppm). Research is directed at developing hybrid zero-emission vehicles that use hydrogen for lean-burning, spark-ignited engines combined with flywheels, supercapacitors, or batteries as energy storage devices. Mobility based on hydrogen could become competitive with continued technological development and cost reductions. Regenerative energy storage for ultra-high-flying solar aircraft is another use for a hydrogen support system. The system was designed by NASA to store energy by converting the craft's excess solar power into hydrogen and oxygen during the day and then to use a fuel cell to "regenerate" electricity at night.

The hydrogen storage necessary to continuously power automobiles or fuel cells is a major safety problem. Hydrogen is stored as a compressed gas in vehicles or as a cryogenic liquid in physical storage systems for transport. The solid-state storage system is another, safer storage technology that could potentially store more hydrogen per unit volume. Solid-state systems bind hydrogen to a solid material such as carbon. Critical parameters for solid-state storage are hydrogen uptake capacities, controlled hydrogen release, and rapid hydrogen refilling. Potential advantages of solid-state storage are its applicability at higher temperatures and near ambient pressures. According to Schlapbach and Züttel, lightweight metal alloy hydrides are the most promising materials for hydrogen storage.

4.5 Biomass

Biomass fuels (biofuels) refer to chemical energy stored within plants by photosynthesis. Theoretically, 6.7% of the solar energy input on land could be converted to chemical energy by photosynthesis, although on average only 0.3% is stored as carbon compounds in land plants. Kheshgi *et al*. calculated net primary productivity to be 60 GtC per year on land and 50 GtC per year at sea. With different heating values, biomass varies in both its energy and its carbon content.

Primary biomass storage includes any organic matter available on a renewable basis. It comprises biomass stocks and an assorted range of waste and residue materials. The basis for secondary storage before distribution and final or end-use storage is a continuous biomass resource supply chain to produce solid and liquid biofuels or synthetic gases (syngas).

The chemical energy released as heat by combustion is mostly used for room heating or industrial processes. It can also be converted to electricity. Gaseous or liquid fuels such as methane, hydrogen, and alcohols or bio-oils can be produced from biomass instead of direct heat production. Liquid biofuels and biogas are applicable in IC engines or fuel cells as alternatives to petroleum-based fuels. The conversion of biomass to electricity is conceptually similar to the conversion of coal to electricity using either direct combustion or integrated gasification combined cycle technology.

The carbohydrates of biomass can be converted by fermentation to ethanol or by extraction and refining to plant oils. For example, corn can be converted to a fermentation feedstock, where microorganisms

including yeast and other bacteria produce ethanol, which can be refined and used for blended gasoline, an alternative transportation fuel. Another example is the reaction of plant oils (fatty acids) with methanol in the presence of a catalyst, producing biodiesel. Biodiesel is used as an alternative fuel especially in agriculture either in neat form or blended with petroleum diesel for use in diesel engines.

One major issue with regard to secondary or final storage of biofuel is the oxidative stability of the compounds. Stability test methods allow a customer to determine whether the fuel has remained stable in storage over extended periods of time. Biodiesel stored in standard diesel storage tanks interacts with the accumulated tank sediments and sludge from ordinary diesel and reduces the solvency problem of ordinary diesel. For solid fuel storage, security measures are vital for preventing (self)-ignition in the presence of heat and oxygen.

Biogas is the most promising biofuel. It is more directly accessible, has good storage stability, has better overall energetics, and can be applied more flexibly. It can also be produced by fermentation processes, pyrolysis, or gasification techniques from diverse biomass raw materials. Biogas from 1 kg of air-dried, gasified biomass stores approximately 8000 kJ of energy, which can deliver approximately 0.8 kWh of electric output. Biogas production from biomass pyrolysis takes place at high temperatures, under exclusion of oxygen in the presence of catalysts. Biogas storage is no different from the storage of natural gas. Since biogas contains hydrogen sulfide, tank materials must be H_2S resistant. The purity of biogas compared to natural gas is a controversial issue in debates on the economic feasibility of large biogas production plants, which would require hefty cost reductions and efficient conversion technologies.

The central problems in producing bioliquids are its economic cost and the negative energetics of production. Besides growing and harvesting the biomass stock, significant amounts of energy are consumed in the conversion of biomass to biofuels. Approximately 6% of the (lower) biomass heating value is used per unit production. Developments to improve the overall efficiency of biomass energy systems will be the key to making this a feasible alternative. Another distinct characteristic of biomass energy sources is the large terrain requirements for generating biomass stock. Terrestrial plants, however, are not the only option as renewable energy stock. Marine algae, including kelp, are potential renewable energy resources but are currently too expensive to utilize.

5. CONCLUSION

Although an efficiency loss is unavoidable for all storage systems, the capacity and services of energy systems, including load distribution and system reliability, must continually be optimized. Load-distribution strategies and energy end-use efficiency increase customer benefits, whereas energy storage and control devices allow consumers to time energy purchases and shift from high- to low-rate periods. Energy storage also increases system reliability and will be the key to balancing the intermittency of new renewable energy sources in the future.

Efficient electricity storage in larger quantities is only indirectly possible by using established storage techniques, such as pumped storage hydropower plants pumping water back to the reservoir at times of low electricity demand. For such large systems, the foremost limiting factor for future growth is the availability of suitable sites with low environmental impact. Similarly, the development of large-scale CAES will depend on the development of systems with man-made storage tanks.

During the past 20 years, smaller scale electricity storage technologies have advanced greatly. New energy storage systems involving flywheels are very promising, especially in combination with other storage technologies to support uninterruptible electricity supply systems, thus improving reliability and quality. Batteries remain significant for smaller appliances and applications in which grid power independence is necessary. Supercapacitors have been developed as an additional electricity storage technology to maintain grid stability and satisfy rapid power needs.

Thermal storage technologies are especially useful in combination with solar energy, CHP, and waste heat distribution to provide hot water and heat for housing. Although large thermal storage systems such as high-temperature UTES and ATES are more efficient, small-scale applications intelligently integrated into building technology have immense potential.

Chemical storage remains the definitive storage method, with its high energy density. The use of coal and oil, however, poses great problems for sustainable energy due to CO_2 emissions. The transition to natural gas, hydrogen, and biomass as prime chemical energy storage is envisaged for the future. Cost

reductions in the various chemical storage techniques will be crucial for applications in mobility, distributed electricity generation, and heat production.

Energy storage should be lean, safe, and reliable to contribute to the functioning of energy supply markets and energy systems. Some redundant capacity of energy storage is important for all sectors of the socioeconomy to cope with disruptions caused by various technical, economical, or political events.

Acknowledgments

I thank D. Spreng (CEPE) for his encouragement and advice with regard to this article. I am indebted to J. Appleby (Texas A&M University) for invaluable comments and corrections. I also thank M. Berg (Factor AG, Switzerland) for his valuable comments on the fossil fuel sections. Thanks also to D. Goldblatt for reading the manuscript. Finally, I thank Anne Semadeni-George for her extraordinary help in reducing the text to an acceptable length.

SEE ALSO THE FOLLOWING ARTICLES

Coal Storage and Transportation • Hydrogen Storage and Transportation • Natural Gas Transportation and Storage • Strategic Petroleum Reserves • Thermal Energy Storage

Further Reading

Aceves, S. M., and Smith, J. R. (1996). "Hybrid and Conventional Hydrogen Engine Vehicles That Meet EZEV Emissions." Lawrence Livermore National Laboratory/U.S. Department of Energy, Livermore, CA/Washington, DC.

Banks, F. E. (2000). "Energy Economics: A Modern Introduction." Kluwer, Boston.

Blaisdell, R. A., and Smallwood, M. E. (1993). "Evaluation of the Total Hydrocarbon Standard for Cleanup of Petroleum Contaminated Sites." Wright-Patterson Air Force Institute of Technology, OH, USA.

Chartier, P., Beenackers, A. A. C. M., and Grassi, G. (1995). "Biomass for Energy, Environment, Agriculture and Industry." Pergamon, Oxford, UK.

Collinson, A. (2001). "Daces 2050: Database Clean Energy Supply 2050." Netherlands Ministry of Housing, Spatial Planning and the Environment/Utrecht Centre for Energy Research/Department for Science, Technology and Society, Utrecht University/Netherlands Energy Research Foundation (ECN)/Ecofys, Energy and Environment/National Institute of Public Health and the Environment; Utrecht, The Netherlands.

Dinçer, I., and Rosen, M. A. (2002). "Thermal Energy Storage: Systems and Applications." Wiley, Chichester, UK.

Dresselhaus, M. S., and Thomas, I. L. (2001). Alternative energy technologies. *Nature* **414,** 332–337.

EESAT. (1998). Proceedings of the International Conference of Electrical Energy Storage Systems: Applications and Technologies. *In* "International Conference on Electrical Energy Storage Systems," Chester, UK.

Ewald, R. (1998). Requirements for advanced mobile storage systems. *Int. J. Hydrogen Energy* **23**(9), 803–814.

Howell, S. A., and Weber, J. A. (1996). "U.S. Biodiesel Overview." National Biodiesel Board, Jefferson City, MO.

International Energy Agency (1997). "Implementing Agreement on Energy Conservation through Energy Storage." International Energy Agency, Paris.

International Energy Agency (2000). "Agreement on the Production and Utilization of Hydrogen." International Energy Association/C. C. Elam National Renewable Energy Laboratory, Paris/Golden, CO.

International Energy Association (2001). "Implementing Agreement for Hydropower Technologies and Programmes—Annex III: Hydropower and the Environment: Present Context and Guidelines for Future Action." International Energy Agency, Paris.

Kheshgi, H. S., Prince, R. C., and Marland, G. (2000). The potential of biomass fuels in the context of global climate change: Focus on transportation fuels. *Annu. Rev. Energy Environ.* **25,** 199–244.

Larbalestier, D. (2001). High-Tc super-conducting materials for electric power applications. *Nature* **414,** 368–377.

Linden, D., and Reddy, T. B. (2002). "Handbook of Batteries." McGraw-Hill, New York.

Lorén, A., Hallbeck, L., Pedersen, K., and Abrahamsson, K. (2001). Determination and distribution of diesel components in igneous rock surrounding underground diesel storage facilities in Sweden. *Environ. Sci. Technol.* **35**(2), 374–378.

Masuda, M. (1988). Lecture on new technology for energy conversion and its storage. Storage of energy, part 2: Storage of electric energy by super-conductivity. *Nenryo-Kyokai-Shi J. Fuel Soc. Japan* **67**(9), 795–801.

Maxwell, R. M., and Kastenberg, W. E. (1999). A model for assessing and managing the risks of environmental lead emissions. *Stochastic Environ. Res. Risk Assessment* **13,** 231–250.

Owen, A. D. (2001). Conference Proceedings 2001: The Economics of Renewable Energy Technologies. *In* "24th IAEE International Conference. Energy Odyssey?," held in Houston, TX. Omnipress, Madison, WI.

Padró Grégoire, C. E. (2000). "Advances in Hydrogen Energy." Kluwer /Plenum, New York.

Provenzano, J., Scott, P. B., and Zweig, R. (1997). Solar hydrogen for urban trucks. *In* "8th Annual U.S. Hydrogen Meeting Proceedings," pp. 209–221. National Hydrogen Association, Washington, DC.

Ristinen, R. A., and Kraushaar, J. J. (1999). "Energy and the Environment." Wiley, New York.

Schlapbach, L., and Züttel, A. (2001). Hydrogen-storage materials for mobile applications. *Nature* **414,** 353–358.

Turner, W. C. (2001). "Energy Management Handbook." Fairmont, Lilburn, GA.

Witt, E. C., Mitchell, R. L., Holly, T. P., and Symko-Davies, M. (2001). Terrestrial photovoltaic technologies update. *Renewable Energy* **23**(3/4), 349–353.

Yeager, K., Gehl, S., Barker, B., and Knight, R. (1998). Roadmapping the technological future of electricity. *Electricity J.* **11**(10), 17–31.

Strategic Petroleum Reserves

DAVID L. WEIMER
University of Wisconsin
Madison, Wisconsin, United States

1. **Importance of Strategic Petroleum Reserves**
2. **Historical Overview**
3. **Rationale for Government Role**
4. **Economic Impacts of Stock Draw Downs**
5. **Costs of Strategic Petroleum Reserves**
6. **Time and Risk**
7. **Programmatic Approaches**
8. **Conclusions**

Glossary

consumer surplus The monetary value individuals place on consumption in excess of the amount of money they have to pay to obtain it.
draw down The release of petroleum from strategic stocks.
***in situ* storage** Holding crude oil stocks in the ground by capping wells.
marginal excess tax burden The additional cost to the economy of another dollar of taxation.
market failure A situation in which the self-interested actions of market participants does not lead to an economically efficient allocation of resources.
mmb/d Millions of barrels per day.
oil price shocks Steep increases in the price of crude oil in the world market resulting form the sudden removal of supply from the market.
price elasticity of demand The percentage change in the quantity of a good demanded in a market resulting from a 1% change in its price.
price elasticity of supply The percentage change in the quantity of a good supplied to a market resulting from a 1% change in its price.
primary stocks Petroleum held by crude oil producers and refiners. Secondary stocks are those held by wholesalers and retailers; tertiary stocks are those held by consumers.
producer surplus Payments to producers of a good in excess of the minimum amount required to produce it.
risk aversion A preference for a certain payoff over a lottery with an expected outcome equal to the certain payoff.
salt domes Caverns in underground salt formations that can be created through the injection of fresh water and the removal of the salt in solution and used to store petroleum.
speculative stocks Petroleum accumulated in expectation of profitable sales at higher future prices.
strategic stocks Petroleum stocks held either to facilitate draw downs to reduce the economic impacts of supply disruptions or to deter purposeful reductions in supply for political ends.
supply disruption The sudden removal of substantial quantities of crude oil from the world market. Major disruptions occurred in conjunction with the Yom Kipper War (October 1973 to March 1974), the Iranian Revolution (November 1978 to April 1979), the Outbreak of the Iran-Iraq War (October 1980 to January 1981), and the Gulf Crisis (August 1990 to January 1991). The largest of these in magnitude was the Iranian Revolution, which removed 5.6 mmb/d for approximately 6 months.
surge capacity The incremental supply that can be added to the market over relatively short periods of up to 6 months.
working (operational) stocks Petroleum held to facilitate normal industry operations, including stocks in pipelines, tankers, and tank bottoms, stocks to smooth tanker deliveries, and stocks to accommodate seasonal demands.

Strategic petroleum reserves, also referred to as strategic petroleum stocks, are quantities of crude oil or petroleum products held either to facilitate draw downs to reduce the economic impacts of crude oil supply disruptions or to deter purposeful reductions in crude oil supply for political ends. They reduce the vulnerability of the world economy, especially the economies of petroleum importers, to the adverse consequences of sudden and steep price increases resulting from the removal of substantial quantities of crude oil supply from the world market. They make purposeful reductions in crude oil exports for political purposes less attractive by increasing the quantities that must be withheld to inflict any threatened level of harm on petroleum importers. In these ways, strategic petroleum reserves reduce the

 © 2004 Elsevier Inc. All rights reserved.

risks to the world economy associated with reliance on geographically concentrated crude oil supplies.

1. IMPORTANCE OF STRATEGIC PETROLEUM RESERVES

A healthy world economy requires a steady flow of crude oil. In the short run, the relatively small absolute value of the price elasticity of demand for crude oil means that even modest reductions in supply can result in dramatic increases in price. Such price shocks result in direct losses to the national economies of net importing countries and produce adverse macroeconomic effects for most of the rest of the world. Indeed, economists attribute much of the poor economic performance of the world economy during the 1970s to the oil price shocks occurring in connection with the Yom Kipper War, which quadrupled world prices, and the Iranian Revolution, which doubled them. Continued dependence on petroleum as a major energy source will keep the world economy vulnerable to crude oil supply disruptions into the foreseeable future.

The risk of oil supply disruptions will remain high because of the concentration of crude oil production and reserves in the volatile Persian Gulf region. In 2002, the world consumed on average about 77 mmb/d of crude oil. Approximately 29% of this supply originated in the Persian Gulf. With more than half of the world's proved oil reserves, the Persian Gulf is likely to account for an even larger fraction of supply in the coming years. The disorder in Iraq and longer term concerns about the stability of the Saudi Arabian and Iranian regimes make it easy to imagine scenarios in which the world would lose some or all of the oil exported from the Persian Gulf. For example, a revolution in Saudi Arabia that damaged export facilities would deny the world oil market approximately 10 mmb/d and, without offsetting increments of supply, send the world price of oil to over $50/b. Less severe disruptions could originate in other parts of the world, such as the countries of the former Soviet Union. The primary purpose of strategic petroleum reserves is to provide a source for offsetting increments of supply in such situations.

2. HISTORICAL OVERVIEW

The holding of commodity stocks is one of the oldest functions of government—recall the biblical story of Joseph advising the Pharaoh to put away the surplus of the coming seven good years to be prepared for the seven lean years to follow. In the early twentieth century, oil came to top the list of critical resources as navies converted from coal to oil and economies became motorized. By the end of the century, civilian use came to overshadow military needs so that oil stocks came to be viewed as more important for economic than military security.

2.1 Naval Petroleum Reserve

The conversion of the U.S. Navy to oil-burning ships made access to petroleum an immediate national security concern. Even though the United States was then a net exporter of crude oil, beginning in 1912 the United States set aside oil-bearing lands in California, Wyoming, and Alaska as a naval petroleum reserve. The reserve was to ensure that crude oil would be available for use by the navy in the event of an extended military conflict.

The naval petroleum reserve can be thought of as *in situ* storage of strategic stocks. Rather than removing oil from the ground and placing it in storage, the oil is stored in place. Although *in situ* storage seems like a commonsense approach to creating strategic reserves, the historical experience with the naval petroleum reserve reveals several problems with storing strategic stocks in the ground.

First, unless the land area of the reserve is sufficiently large to preclude other land owners from tapping the same reservoir of oil, production must be undertaken to protect the reserve from commercial encroachment. This has been the case at one of the naval petroleum reserve sites since the 1920s, as well as intermittently at two of the other sites, including the infamous Teapot Dome. The problem was particularly severe in the U.S. context where the rule of capture gives ownership of oil and gas to anyone who brings it to the surface on their own land. Yet preserving the reserve would likely be a problem whenever the entire reservoir is not dedicated as strategic stocks.

Second, *in situ* reserves are often not located near transportation facilities with adequate capacity to handle surge production. In the case of the Alaskan Naval Petroleum Reserve, for example, a second Alaskan oil pipeline would have to be built and then left to stand idle if the oil from the reserve were to be available as additional supply to the market. Providing surge capacity places a high demand on transportation facilities.

Third, it is generally costly and difficult to keep shut-in wells in a state of readiness that allows them to achieve full levels of production rapidly. After the Six Day War in 1967, the secretary of the navy ordered one of the naval petroleum reserve sites to be prepared to produce at its maximum expected rate within 60 days. The readiness requirement was never met, and, when commercial production was begun in 1976, the maximum rates of production anticipated turned out not to be achievable.

In situ storage like that provided by the naval petroleum reserve has an even more fundamental disadvantage: the large amount of proved reserves that must be set aside to achieve substantial surge capacity. It is rarely possible to extract more than one-eighth of a reservoir within 1 year. Imagine that the United States wanted to have the capability for 3 mmb/d of surge capacity for 1 year. Reservoirs of approximately 8 billion barrels of proved reserves would have to be set aside, more than one-third of all U.S. proved reserves. Once the purpose of strategic oil stocks shifted from supporting military operations in extended conflicts to providing surge capacity for avoiding the economic losses from oil price shocks, *in situ* stocks were no longer appropriate as strategic stocks.

2.2 Stockpiling as a Response to Import Dependency

Several policies limited U.S. dependency on imported oil through the early 1960s. One was an oil import quota that attempted to limit imports directly. Initiated in 1957, it remained in effect until rising world prices and controlled domestic prices made it politically unworkable in 1973. The other was the pro-rationing of production in unitized oil fields by state regulatory agencies in Texas and Oklahoma. For example, the Texas Railroad Commission attempted to stabilize crude oil prices by raising and lowering the pro-rationing rate in response to changes in demand. By 1972, however, prorating reached 100%, removing the capacity of the state agencies to stabilize price.

The first proposals for holding strategic reserves came from economists who distinguished between import dependence and vulnerability. They argued that it was not import dependence per say that was worrisome, but rather that the imports were subject to disruption from world events. From this perspective, energy security could be achieved by reducing vulnerability even with substantial reliance on imports. These ideas began to take hold politically, but legislation to create a strategic petroleum reserve introduced in the Senate in April 1973 failed to pass.

2.3 International Energy Program

In October 1973 the Organization of Arab Petroleum Exporting Countries (OAPEC), a subset, including the most significant members, of the Organization of Petroleum Exporting Countries (OPEC), imposed an embargo on oil shipments to the United States and The Netherlands in retaliation for their support of Israel in the Yom Kipper War. At the same time, OPEC cut total production. Although the flexibility of the world oil market kept the United States and The Netherlands from suffering disproportionately from the embargo, the reductions in oil supply dramatically increased oil prices in the world market. In response, 16 of the then 24 members of the Organization for Economic Development and Co-operation (OECD) adopted the International Energy Program to protect its members from targeted embargoes and prepare for future oil supply disruptions. A formal oil sharing agreement was included to counter targeted embargoes, and requirements for national-level programs of demand restraint and the establishment of strategic petroleum reserves were included to reduce the impacts of supply disruptions.

The International Energy Program set a stock requirement for its members based on levels of imports. Specifically, each country is required to hold petroleum stocks, exclusive of military stocks and oil in transit at sea, equivalent to 90 days of net petroleum imports. (Members of the European Economic Community face a similar requirement based on consumption, with partial waivers for Britain and Denmark, its net oil exporters.) Countries may count as much as 90% of working stocks toward the requirement. In 1995, the members of the International Energy Program agreed to coordinate their draw downs of strategic stocks during oil supply disruptions, and to do so before initiating sharing arrangements.

The capability provided by the International Energy Program's stockpiling can be measured in terms of the maximum amount of petroleum that can be delivered to the world market at the onset of an oil supply disruption. Member country governments control public petroleum stocks that would enable them in an all-out coordinated draw down to release perhaps as much as 12 mmb/d of stocks during the first month of an oil supply disruption, followed by approximately 8 mmb/d for the next 2 months. In

subsequent months the volume of the draw down would decline steeply, reaching negligible levels by the eighth month. Some additional releases from commercial stocks could further augment the draw down, especially after prices peak.

3. RATIONALE FOR GOVERNMENT ROLE

Private firms routinely hold petroleum stocks. Transporters and refiners hold primary stocks to maintain the transportation system (tank bottoms and pipeline fill), to avoid production interruptions from delivery delays, and to smooth out seasonal demands. Wholesale and retail firms hold secondary stocks to ensure that they can meet customer demands. End users of petroleum products may hold tertiary stocks to reduce the consequences of delayed deliveries. In general, these are working, or operational, stocks in the sense that they provide the grease needed to keep the distribution system working smoothly. Beyond these routinely held stocks, firms may seek profits by buying speculative stocks at low prices in anticipation of future higher prices. Specifically, the holding of a barrel of speculative stocks will be profitable in an expected value sense if the percentage increase in the expected price less storage costs over the current price is greater than the rate of interest. For example, if the expected price 1 year form now were $35/b with physical storage costs to be paid at that time of $2/b, and the current price were $30/b, it would be profitable in an expected value sense to add to stocks if the interest rate facing the firm were 10% or lower. (The firm could earn $33 dollars by simply investing the $30 at the 10% interest rate instead of holding the stocks.) Does private speculation provide adequate strategic stocks from the perspective of society as a whole? In other words, are there market failures that provide a rationale for government stockpiling?

Policy analysts have identified a number of reasons why speculative stocks are likely to be too small to serve adequately as strategic stocks. First, private firms do not gain the full social benefits from draw downs. When firms draw down previously accumulated stocks to profit from higher prices, they produce an external benefit to society—lower prices for petroleum products than would have resulted in the absence of their draw down. This divergence between private and social benefits results in too little speculative stockpiling.

Second, private firms do not see the full social costs from accumulating stocks. When firms accumulate stocks, they do not take account of the social costs of the higher prices that result. Although this divergence could lead to too little speculative stockpiling, the likely timing of additions to stocks may actually add to the social costs of oil supply disruptions. Speculative stockpiling is most attractive to individual firms when they anticipate that prices will rise quickly. This is likely to be the case at the outset of supply disruptions as firms see events leading to supply reductions and attempt to add to stocks. This additional demand may make the price shock associated with the supply disruption steeper and more severe, adding to its social costs. In effect, the divergence between private and social costs is likely to lead to speculative stockpiling that increases rather than decreases social costs.

Third, the model of speculative stockpiling assumes that firms are risk neutral in the sense that they are indifferent between a certain payoff and a lottery with an expected value equal to the certain payoff. So, for example, a firm would be risk neutral if it were indifferent between a zero certain payoff and a lottery that gave a 10% chance of $9 payoff and a 90% chance of a $1 loss—although the expected payoff from the lottery is zero, 90% of the time the firm would lose money. It is quite likely that firms would exhibit some degree of risk aversion, which would lead them to choose lotteries only when their expected payoffs are substantially greater than zero. As speculative stocks are like lotteries, risk aversion would likely lead to less speculative stockpiling than would occur if the firms were risk neutral. From the social perspective, however, risk aversion calls for more stockpiling, with the small certain costs of accumulating stocks like a premium for insurance that reduces the large costs that result from oil price shocks.

Fourth, firms may anticipate that government interventions may prevent them from realizing profits from sales of stockpiled oil during price shocks. Throughout the 1970s, U.S. policies controlled the well-head prices of domestically produced crude oil and accumulated stocks. Although these price controls were fully removed in 1981, U.S. firms may anticipate the possibility of reimposition of such controls during a severe oil price shock in the future. This reduces the expected price of speculative stocks by reducing the probability of realizing extremely high prices. Of course, the prospect of the release of public stocks during price shocks also reduces the probability of realizing high prices and thus reduces

speculative stockpiling. At the same time, however, the existence of public stocks may make it less likely that the government will impose price controls and mandatory allocations, making speculative stockpiling more attractive.

Externalities, risk aversion, and fear of government expropriation make a strong case that speculative stockpiling alone will not provide sufficient strategic stocks from the social perspective. Thus, there is a strong rationale for public policy to increase the quantity of strategic petroleum reserves. Determining the appropriate policies for strategic stocks requires an assessment of their social costs and benefits.

4. ECONOMIC IMPACTS OF STOCK DRAW DOWNS

The economic benefits of strategic petroleum reserves accrue when they are drawn down during oil supply disruptions to reduce the magnitude of the resulting oil price shock. The economic impacts can be illustrated by considering hypothetical but plausible price and quantity changes in the world and the U.S. domestic crude oil markets (Fig. 1). The intersection of supply (S) and demand (D) in the world market, which yields an equilibrium price of $28/b and quantity of 76.8 mmb/d, represents a normal market period. At the world price of $28/b, the United States consumes 15.0 mmb/d and produces 5.8 mmb/d domestically. The difference between consumption and domestic production, 9.2 mmb/d, is imported. Now imagine that the 10 mmb/d produced by Saudi Arabia is removed from the market. The world supply curve shifts from S to S_d, inducing a price shock that raises price from $28/b to $64/b. In the U.S. market, the $64 price reduces the amount consumed and increases the amount domestically produced.

The net loss to the U.S. economy of the price increase is given by the area of the trapezoid inscribed by the points *cdfe*. It results from a consumer surplus loss equal to the area of trapezoid *bdfa* less producer surplus gains given by the area of trapezoid *bcea*. The total economic loss from the price shock can be divided into three components. First, consumers suffer a loss from consuming less—area *dfh*. Second, consumers transfer wealth to foreign oil producers equal to the area *cdhg*. Third, domestic producers spend real resources, equal to area *cge*, to increase production.

Now imagine that the United States, Germany, and Japan collectively draw down their national

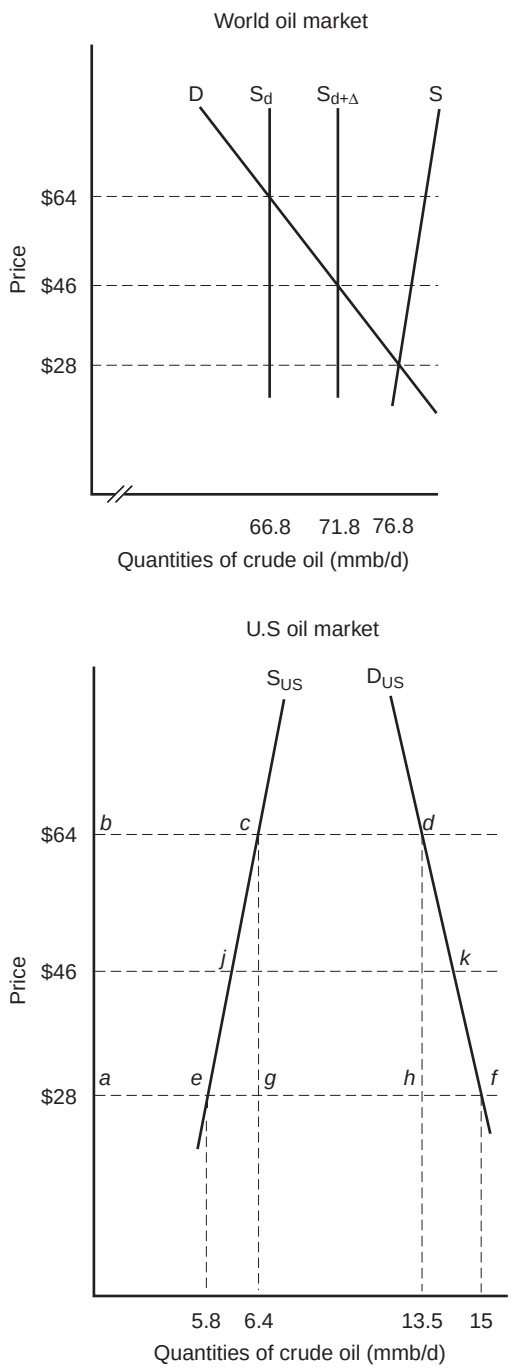


FIGURE 1 The economic impacts of strategic stock draw downs in the world and U.S. crude oil markets.

reserves at a rate of 5 mmb/d. This has the effect of shifting the world supply curve from S_d to $S_{d+\Delta}$. This additional supply in the world market drives price down to $46/b. The benefits of this draw down in the U.S. market can be seen as the difference between the surplus loss without the draw down (area *cdfe*) and the surplus loss with the draw down (area *jkfe*). The

direct net benefits of the draw down to U.S. consumers is therefore the area of *cdkj*, which equals approximately $140 million per day, or approximately $12.6 billion over 90 days.

The illustration makes clear that price is determined in the world market. Thus, draw downs are most appropriately thought of as affecting the world market. Draw downs by any country produce a benefit to petroleum consumes worldwide. Attempts to limit the benefits of draw downs domestically are doomed to failure, as relatively lower domestic prices would divert imports from the country to the higher priced world market, eventually equalizing price. The country drawing down does obtain a direct financial benefit through sales of the drawn down petroleum.

Three other points related to the illustration are relevant to the question of when strategic petroleum reserves should be used. First, efforts by private firms to add to speculative stocks as events lead them to anticipate the higher prices that will result from the supply disruption could be shown in the diagram as increases in demand. In terms of the figure, the speculative stockpiling would shift the demand curve, D, to the right, causing price to rise even higher than would be the case with only the loss of supply to the world market.

Second, the economic costs measured in the U.S. market by the net changes in consumer and producer surpluses do not include the macroeconomic costs that result from moving form low to high price equilibria. Interpretation of the surplus changes as economic effects assumes that prices of all other economic inputs and goods freely adjust to the higher price of petroleum. In most economies, however, there are price rigidities, such as labor contracts, that prevent prices from rapidly adjusting. Without free movement of prices, resources will not necessarily be used efficiently. For example, wage rigidities are likely to lead to higher rates of unemployment during oil price shocks. Uncertainty about the composition of post-disruption demand may depress investment. Overall, these additional costs are likely to grow more than proportionally with the size of the price shock, probably approximating the magnitude of the direct costs for a doubling of price.

Third, the illustration is a snapshot showing the initial impacts of the supply disruption. Over time, adjustment by producers, holders of speculative stocks, and consumers will reduce the price of petroleum even before the disrupted supply is returned to the world market. Oil producers will find it profitable to increase production from existing fields, say by using more expensive recovery methods or reopening previously closed stripper wells, which will tend to rotate the supply schedule, S, clockwise around its predisruption equilibrium. Once it appears that price has peaked, those holding speculative stocks will release them, adding to the available supply and removing any shift in demand caused by their accumulation efforts. Petroleum consumers will find ways of reducing petroleum consumption that will tend to rotate demand counterclockwise around the predisruption equilibrium. For example, during the first week of the disruption, consumers can turn down thermostats, make sure that their automobile tires are properly inflated, and take fewer trips; during the first month, they can obtain tune-ups for their automobiles, augment window insulation, and wrap water heaters in insulation; during the first 6 months, they can make more substantial changes in insulation, thermostats, appliances, and vehicles. Consequently, other things equal, the longer the disruption persists, the more market forces, both in terms of incremental supply and reduced demand, will tend to move the price of oil toward the predisruption equilibrium.

Each of these three points argue for large draw downs at the outset of supply disruptions. Immediate draw downs are likely to suppress speculative stockpiling, thus avoiding the additional price increase it causes. The disproportionate increase in macroeconomic costs with the size of the price shock argue strongly for using available strategic stocks to limit the initial magnitude of the price shock. The price mitigating market responses, as well as macroeconomic adjustments, that occur over time suggest that stocks will be relatively more valuable during major supply disruptions if used to support larger initial draw downs than to support smaller but longer draw downs.

5. COSTS OF STRATEGIC PETROLEUM RESERVES

Before strategic petroleum reserves can be drawn down, they must be accumulated and stored. In assessing costs, it is important to distinguish between budgetary costs and social costs. The former are the flows of funds from the national treasury to purchase resources for stockpiling; the latter are the economic value given up by use of the resources for stockpiling. Although budgetary cost is often politically relevant, social costs should be the basis for public policy.

Consider first government-controlled strategic stocks. Their budgetary costs include expenditures to construct (or purchase) storage facilities and maintain them, to connect storage facilities to oil distribution networks, and, most significantly in terms of magnitude, to purchase petroleum. An offsetting revenue is realized when the stocks are sold during drawn down. These budgetary costs may differ from social costs in a number of ways, however.

If petroleum purchases for strategic stocks drive up prices, then they will inflict economic losses on consumers. The economic costs of the price increase can be measured in the same way as price increases resulting from supply disruptions (Fig. 1). Generally, increments to strategic stocks are small relative to world consumption so that they generally will result in negligible, if any, price increases. The major exception occurs at the onset of supply disruptions. (Indeed, stopping ongoing purchases have the same effect as a draw down equal to the magnitude of the purchases.) If purchases are reduced or suspended when petroleum markets are experiencing upward price pressure, then the budgetary and social costs of oil purchases will closely correspond.

The taxes governments use to raise revenues create economic inefficiencies by distorting economic behavior. Economists refer to these inefficiencies as the marginal excess burden of taxation. In assessing the social costs of stockpiling, the marginal excess burden should be applied to revenue flows into and out of the national treasury. For example, if the marginal excess burden for the tax system of a country is 10%, a billion dollars of oil purchases would have a social cost of $1.1 billion and $20 billion dollars of revenue from a draw down would yield a social benefit of $22 billion.

A large discrepancy between budgetary and social costs results when the government bypasses market purchases to obtain oil. For example, the United States receives royalties of between 12.5 and 16.7% of oil obtained from leases on the federally owned outer continental shelf. Usually these royalties are collected as cash equivalents. Plans, however, call for receiving this oil in-kind and using it to fill the Strategic Petroleum Reserve. Although these royalties no longer pass through the budget, they have the same social costs as if they were purchased on the market.

A similarly large discrepancy between budgetary and social costs results when the government mandates that private firms hold strategic stocks. The budgetary costs for purchasing and storing oil will be zero, but the social costs include the purchase and storage costs borne by the firms.

Storage costs can vary greatly. Storage facilities vary greatly in their cost. Aboveground tanks typically cost between $15 and $18 per barrel to build and maintain. In contrast, storage in salt domes costs between $1.50 and $5.00 per barrel to build and maintain. Although abandoned salt mines may be available, salt dome capacity is generally created through solution mining, which involves injecting fresh water to dissolve salt that is removed in brine. Temperature differences in deep salt dome facilities provide natural circulation of stored petroleum. The abundance of salt domes on the Gulf Coast near major crude oil pipelines and tanker terminals provided the United States with opportunities for relatively low cost storage and transportation of strategic petroleum reserves.

Maintenance costs for crude oil stocks are generally low. They can be substantially higher for petroleum products, which tend to deteriorate in quality over time and therefore must be rotated out of stocks. The higher maintenance costs if petroleum product stocks has been one factor leading governments to hold mostly crude oil in government owned stocks. Another factor favoring crude oil over petroleum product stocks is the greater flexibility in product mixes afforded by crude oil stocks. For example, heating oil is likely be relatively more valuable during disruptions that occur during the Northern Hemisphere's heating season, while gasoline is likely to be relatively more valuable during its summer months. Refineries can alter the product mixes they obtain from crude oil, thus making more of the products that are relatively highly valued at the time.

6. TIME AND RISK

Strategic stocks share several characteristics with insurance. Just as the insurance premium must be paid in advance of the coverage obtained, petroleum must be accumulated prior to its use to counter the economic losses of an oil price shock. Just as events justifying claims against insurance coverage may never materialize, strategic stocks may never be drawn down. Analyses of the desirability of strategic petroleum reserves require attention to the timing of costs and benefits as well as to uncertainty about whether benefits will actually be realized.

Valuing costs and benefits that accrue at different times requires that they be put into a common metric. Economists do so by converting future costs

and benefits to their equivalent present values. The present value recognizes that consumers generally prefer to consume sooner rather than later. Although there is considerable controversy over the appropriate social discount rate that should be used to convert future costs and benefits to their present values, the general procedure is straightforward. In the case of assessing strategic stockpiling programs, current costs required to accumulate stocks receive higher weights than benefits that accrue in the future.

The more challenging analytical task in assessing stockpiling programs is how to deal with uncertainty. Will an oil supply disruption warranting use of strategic stocks ever occur? If so, when will it occur? How severe will it be? How long will it last? None of these questions can be answered with certainty.

Analysts typically replace these uncertainties with corresponding risks. For example, assume that there is some probability that a major disruption will occur each year. They can then assess the net benefits of a particular stockpiling program for different assumed values of this probability.

The earliest approaches posited a representative disruption scenario occurring, say, 10 years after the stockpiling program begins. The present value of the costs of accumulating and storing the stocks were estimated, as were the present value of benefits that would result from draw down of the stocks during the disruption scenario. The question could then be asked: How high a probability of the scenario occurring would be required to yield an expected value of net benefits of zero, so that the stockpiling program would just break even? For example, imagine that the present value of the costs of accumulating and storing the stocks were estimated to be $10 billion and the present value of the benefits were $100 billion, then the probability of the posited disruption would have to be at least 10% for the stockpiling program to break even. Decision makers could then be asked if their subjective assessment of the probability of the representative scenario were greater than this breakeven probability.

More sophisticated treatments of uncertainty require models that allow for differences in the timing, magnitude, and duration of oil accumulation, supply disruptions, and draw downs. Such models typically posit a transition matrix that specifies assumed probabilities of moving from one market state to another, say moving among normal, minor disruptions, and major disruptions. Dynamic programming or other solution techniques are then used to find the optimal stock acquisition or draw downs for any particular combination of current market state and stock size. In the United States, this approach generally supported expansion of the public reserves as well as large draw downs at the onset of major supply disruptions, even when there was a risk that the disruptions would persist beyond the point when stocks would be exhausted.

7. PROGRAMMATIC APPROACHES

OECD members have adopted a variety of approaches for meeting the International Energy Program requirement for holding 90 days of imports in stocks. All net importing members either require their domestic companies to hold strategic stocks or have standby authority to do so. Overall, company stocks account for approximately two-thirds of stocks held under the International Energy Program. In a number of countries, however, the stocks held in compliance with the International Energy program are largely working stocks that could not be fully deployed as strategic stock draw downs.

The most certain strategic stocks are those under direct government control in Japan, Germany, and the United States. Together, they could draw down 8.7 mmb/d during the first month of a major oil supply disruption. Such a rapid initiation of a coordinated draw down by these three countries at the outset of a supply disruption could substantially reduce its economic costs.

7.1 Japan National Oil Corporation

Even prior to the International Energy Program, Japan began requiring its firms to hold petroleum stocks. In 1978, it created the Japan National Oil Corporation. Currently, private firms have access to low-interest loans to cover the 25 days of imports they are required to hold above working stocks (45 days). The Japan National Oil Corporation has a goal of holding 90 days of imports. Currently, it has about 315 mmb in storage at 10 sites. Storage methods include conventional steel tanks, floating barges, and underground caverns.

The Ministry of Economy, Trade and Industry, subject to Cabinet decision, makes decisions about use of the stocks. Draw downs from company stocks are initiated by reducing the number of days of required stocks, as was the case in 1991 during the Gulf War; draw downs form the Japan National Oil Corporation would be directly ordered. Currently, the Japan National Oil Company could release 2.3 mmb/d during the first month of an all out draw

down, followed by rapid declines in succeeding months with negligible volumes by the sixth month.

7.2 Erdölbevorrantungsverband

The Oil Stockpiling Law of 1978 established the Erdölbevorrantungsverband (EBV) as an organization to hold petroleum stocks for Germany. All companies, except those in the chemical industry, that refine or import petroleum are compulsory members of the EBV. The EBV is funded solely through membership dues paid by companies in proportion to the quantities of gasoline, middle distillates, and fuel oils produced or imported. The dues are shown as itemized charges to retail customers. Thus, consumers rather than taxpayers finance the EBV. The EBV directors are appointed by an advisory board consisting of government and industry officials. Draw downs are initiated by the government under the Oil Stockpiling Law. Companies that have made higher dues payments receive higher priority in allocation.

In 1998, the government raised the stockholding requirement to 90 days of imports. Currently, the EBV stocks consist of 60% petroleum products and 40% crude oil, though the EBV has announced plans to reverse this proportion in order to increase flexibility in the products available during draw downs and reduce storage costs. Stocks are stored both regionally and in salt domes in Northern Germany. The EBV currently holds products equivalent to approximately 200 mmb of crude oil. During the first 3 months of a supply disruption, the EBV could support a draw down rate of approximately 2.3 mmb/d (0.9 mmb/d of crude oil and 1.4 mmb/d of products). A full draw down would exhaust reserves in 5 months.

7.3 The U.S. Strategic Petroleum Reserve

The Energy Policy and Conservation Act of 1975 called for the creation of a U.S. Strategic Petroleum Reserve of at least 150 mmb within 3 years and 500 mmb, then about 90 days of imports, within 7 years. Subsequent amendments increased the possible size to 1000 mmb. The Strategic Petroleum Reserve has capacity for holding 700 mmb of crude oil in salt dome caverns in Louisiana and Texas; it held 610 mmb crude oil in mid-2003. Since its permanent establishment in 2000, the Northeast Heating Oil Reserve has stored 2 mmb of heating oil in New Jersey, Connecticut, and Rhode Island.

The Strategic Petroleum Reserve provides substantial surge capacity. An all-out draw down of 4.1 mmb/d of crude oil could be sustained for approximately 3 months, with the rate declining in subsequent months as salt dome caverns are emptied. Routine draw downs occur on a regular basis to complete swaps with companies to adjust the grades of crude oil in storage. Draw downs have also been undertaken to raise revenue for deficit reduction. The president may order full-scale draw downs to counter severe energy supply interruptions that may cause major adverse impacts on national security or the national economy. Since 1990, the president has been authorized to initiate limited draw downs of up to 30 mmb in less severe circumstances.

The Strategic Petroleum Reserve was a draw down under presidential directive during the First Gulf War in cooperation with other International Energy Program countries. A week after the Iraqi invasion of Kuwait in August 1990, the secretary of energy announced a 5 mmb draw down to test distribution procedures and demonstrate resolve to use the reserve to counter an oil price shock. When the allied coalition air campaign began on January 16, 1991, the president put 30 mmb out to bid, 17 mmb of which were actually sold during the following 3 months.

The crude oil price fell almost $10/b immediately after the announcement of the draw down, but analysts disagree about whether the price decline resulted primarily from anticipation of the draw down or changing perceptions about the risks to Gulf oil supplies. Analysts also disagree about whether sooner use of the reserve would have prevented the run up in prices during the latter part of 1990's that slowed economic growth.

The history of the Strategic Petroleum Reserve often showed a divergence between congressional authorization to expand stocks and the willingness of Congress and the administration to appropriate funds to enable it. Quite frequently oil purchases for the reserve were suspended for budgetary reasons, and steps have not been taken to implement the full 1000 mmb capacity. The use of in-kind royalties from federal offshore oil leases and heightened national security concerns in the wake of 9/11, however, make it likely that, absent an oil supply disruption, the reserve will be filled to its capacity of 700 mmb over the next several years.

8. CONCLUSIONS

Strategic petroleum reserves reduce the vulnerability of the world economy to oil price shocks. Members

of the International Energy Program have accumulated substantial strategic stocks, even discounting those that are actually working stocks held by companies and unlikely to contribute to surge capacity. These stocks will become more valuable as the proportion of world supply originating in the Persian Gulf increases. International and national policies appear to be continuing to evolve in several ways that increase the chances that strategic petroleum reserves will be used effectively.

First, policy analysts and policymakers now generally realize that strategic stocks have their primary effect through the world oil market. The draw downs of national stocks have very little capability for differentially benefiting those drawing down. Domestic oil prices below world levels simply divert imports to the other countries, thereby equalizing prices. Rather than viewing strategic stocks as providing national benefits in terms of relatively lower domestic prices, they should be seen as contributing to lower world prices, which in turn reduces the negative impacts of supply disruptions on the world economy.

Second, this realization, along with increasing economic globalization, should encourage further efforts toward early and internationally coordinated use of strategic stocks. It should also make individual countries with large surge capacities, namely the United States, Germany, and Japan, less hesitant to initiate draw drowns even when agreement on coordination cannot be reached—any one country can realize large revenues from draw downs. Even if a country exhausts its strategic stocks, it will share in the benefits of later draw downs by other countries.

Third, although the language of the stockpiling agreement continues to emphasize total quantities of strategic stocks held, attention is increasingly being appropriately focused on maximum draw down rates during the first months of an oil supply disruption. This change in focus recognizes that a reduction in world supply has its most severe economic effects during the first 3 to 6 months before nondisrupted oil producers and consumers have time to change their mix of capital goods to in response to the higher prices.

Fourth, most national plans for draw downs take advantage of existing commercial networks. Allowing markets to move petroleum to its most valued economic uses helps ensure that this especially scarce resource during supply disruptions will reach its most valued uses and thereby limit economic losses to the greatest extent possible.

SEE ALSO THE FOLLOWING ARTICLES

Geopolitics of Energy • Markets for Petroleum • National Energy Policy: United States • Nationalism and Oil • National Security and Energy • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Oil Crises, Historical Perspective • Oil Price Volatility • OPEC, History of • Petroleum Property Valuation • Storage of Energy, Overview

Further Reading

Balke, N. S., *et al.* (2002). Oil price shocks and the U. S. economy: Where does the asymmetry originate? *Energy J.* **23**(1), 27–52.

Bohi, D. R., and Toman, M. A. (1995). Oil and national security: An assessment of externalities and policies. *In* "The New Global Oil Market: Understanding Energy Issues in the World Economy" (S. Shojai, Ed.), pp. 203–206. Praeger, Westport, CT.

Horwich, G., and Weimer, D. L. (1984). "Oil Price Shocks, Market Response, and Contingency Planning." American Enterprise Institute for Public Policy Research, Washington, DC.

Horwich, G., and Weimer, D. L. (eds.). (1988). "Responding to International Oil Crises." American Enterprise Institute for Public Policy Research, Washington, DC.

International Energy Agency (2001). "Oil Supply Security: The Emergency Response Potential of IEA Countries in 2000." Organization for Economic Development and Cooperation (OECD)/International Energy Agency (IEA), Paris, France.

Kuenne, R. E., *et al.* (1975). A policy to protect the U. S. against oil embargoes. *Policy Analysis* **1**(4), 571–597.

Mead, W. J., and Sorensen, P. E. (1971). A national defense petroleum reserve alternative to oil import quotas. *Land Economics* **47**(3), 211–224.

Teisberg, T. J. (1981). A dynamic programming model of the U.S. strategic petroleum reserve. *Bell J. Economics* **12**(2), 526–546.

Toman, M. A. (1993). The economics of energy security: Theory, evidence, policy. *In* "Handbook of Natural Resource and Energy Economics 3: Economics of Energy and Minerals" (A. V. Kneese and J. L. Sweeney, Eds.), pp. 1167–1218. North Holland, Amsterdam.

Verleger, P. K. (1990). "Oil Markets in Turmoil." Ballinger, Cambridge, MA.

Weimer, D. L. (1982). "The Strategic Petroleum Reserve: Planning Implementation and Analysis." Greenwood Press, Westport, CT.

Subsidies to Energy Industries

DOUG KOPLOW
Earth Track Inc.
Cambridge, Massachusetts, United States

1. Introduction
2. Subsidy Definition and Measurement
3. Subsidies through the Fuel Cycle
4. Patterns of Subsidization
5. Subsidy Reform
6. Conclusion

Glossary

consumer-subsidy equivalent Integrated metric of aggregate support provided to consumers by varied government policies; policies that act as taxes are incorporated by using the opposite sign.

cross-subsidies Policies that reduce costs to particular types of customers, products, or regions by increasing charges to other groups.

intermediation value Difference between the break-even cost of debt, insurance, or other programs to large governments and what these same services would cost if a smaller, higher risk private firm or individual had to buy them directly.

producer-subsidy equivalent Integrated metric of aggregate support provided to producers by varied government policies; policies that act as taxes are incorporated by using the opposite sign.

renewable portfolio standards Requirements mandating purchase of preset percentages of renewable electricity in particular service regions; the standards normally compete eligible supply sources against each other to minimize the per-unit subsidy.

subsidies Government-provided goods or services that would otherwise have to be purchased in the market or special exemptions from standard required payments or regulations; subsidies may be in cash but often involve shifting risks from private parties to taxpayers or the public.

Whether by intent or by accident, government interventions affect the relative prices of various energy technologies and with them the pattern of energy use and investment. Depending on their structure, interventions can act as either taxes or subsidies to particular technologies, producers, or consumers. The lines can blur; some "taxes" may be insufficient even to cover the cost of beneficial government services, leaving a residual subsidy to the "taxed" parties.

1. INTRODUCTION

Energy resources vary widely in terms of their capital intensity, reliance on centralized networks, environmental impacts, and energy security profiles. Although the policies of greatest import to a particular energy option may differ, their aggregate impact may be significant. Subsidies to conventional fuels can slow research into emerging technologies, thereby delaying their commercialization. Subsidies and exemptions to polluting fuels reduce the incentive to develop and deploy cleaner alternatives. Inadequate tracking and recovery of costs associated with protecting energy security reduces the drive toward diversification.

Justifications for energy subsidies include social welfare, protection and promotion of jobs or industries, rural development, and energy security. Existing policies often fail to achieve these aims in practice. Because subsidies can be worth millions of dollars and often require sophistication or connections to obtain, policies implemented to help poorer segments of society may end up enhancing the wealth of more powerful groups instead. Often, the objectives of subsidy programs can be achieved in a manner that is more narrowly targeted and efficient than the subsidy policies now in place. For example, decoupling subsidy payments from resource-depleting activities can greatly reduce the environmental damages associated with the transfers.

This article examines the general issue of subsidy definition and measurement and then presents central issues associated with subsidization at each stage of a

 © 2004 Elsevier Inc. All rights reserved.

generic fuel cycle. Aggregate patterns of subsidization and the challenges of subsidy reform are addressed in the subsequent two sections.

2. SUBSIDY DEFINITION AND MEASUREMENT

Government interventions encompass a wide range of regulatory, fiscal, tax, indemnification, and legal actions. By modifying the rights and responsibilities of various parties involved with the energy sector, these actions decrease (subsidize) or increase (tax) either energy prices or production costs. As discussed in what follows, differing approaches to subsidy definition and measurement have historically led to disparate subsidy analyses that are difficult to compare or compile.

2.1 Common Disagreements in Subsidy Definition

Disagreements over the proper definition of subsidy are common. Conflicts often arise over the form and timing of the transfer, the definition of the "non-subsidized" baseline, and the boundaries of analysis.

2.1.1 Form and Timing of Transfer

Energy subsidies are often viewed primarily as cash payments from a government agency to private businesses or individuals. Payments to low-income households to enable them to purchase heating oil and grants to businesses to help them develop particular energy technologies are examples. In reality, subsidies can take many different forms, and a more accurate definition must include government-provided goods or services, including risk bearing, that would otherwise have to be purchased in the market. Much market activity involves controlling and sharing the risks and rewards of economic activities, and risk-oriented subsidies are quite important. Subsidies can also be in the form of special exemptions from standard required payments such as tax breaks. Although cash payments are easily measured within a single year, more complex subsidies such as loan guarantees are best evaluated over multiple years so that patterns in losses or investment distortion can be seen more clearly.

2.1.2 Defining the Baseline

Subsidies must often be measured against a baseline. What would taxes owed have been in the absence of this special tax break? How much would industry have had to pay in interest to build that new facility if the government had not guaranteed the loan? Many disagreements over subsidy definition originate in differing views on the appropriate baseline:

2.1.2.1 Indirect Versus Direct Transfers Some argue that interventions "count" as energy subsidies only if they directly target the energy sector. For example, the U.S. Energy Information Administration did not count subsidies to energy facilities provided by tax-exempt general-purpose municipal bonds in its tallies, arguing that the bonds did not constitute an energy subsidy if hospital and road construction could also use them. Similarly, fees to use the U.S. inland waterway system have historically been insufficient to reimburse system construction and maintenance costs. Although oil and coal industries are among the largest users of the system, the fee subsidies (which allow oil and coal a lower delivery cost than would otherwise be possible) are often discounted on the grounds that many commodities use the system. Because many subsidies tilt the energy-playing field toward a particular fuel even if there are also nonenergy beneficiaries, such policies should not be ignored. Rather, any policy that has the effect of subsidizing prices or production costs should be assessed. This may include policies targeted at single sectors, multiple sectors, specific geographic areas, or specific factors of production.

2.1.2.2 Externalities Although the levels vary by fuel, most energy production and consumption generate wide-ranging externalities such as pollution. Exemptions from appropriate environmental controls (e.g., less stringent air pollution control requirements for certain old power plants in the United States) penalize cleaner energy types and are properly viewed as subsidies. However, externalities do create an analytic challenge because they are difficult to monetize. Koplow and Dernbach documented that decisions about which externalities to include and how to value them can generate very large variation in subsidy estimates, on the order of hundreds of billions of dollars per year. Segregating externality-related subsidies from fiscal subsidies can help to improve data comparability and transparency.

2.1.2.3 Intermediation Value Subsidy recipients often incorrectly claim that risk-based subsidies such as loan guarantees have no value unless there are defaults. In addition to the direct cost of the government, all of these policies have an

intermediation value as well given that the government's cost of debt or indemnification is lower than what could be attained by the recipient on its own. For particularly high-risk endeavors, such as sales of large energy assets to politically unstable countries, the intermediation value can be very large. Because some energy resources have much greater access to these government programs than do others, distortions in relative energy prices can result. For example, transactional data on subsidized international lending assembled for the World Commission on Dams show that lending has heavily favored fossil fuels over renewables.

2.1.3 Boundaries of Analysis

Energy is a primary material, an input to refined energy products (e.g., gasoline, electricity), and an input to nonenergy materials (e.g., metals, consumer goods). To provide an accurate picture of energy subsidies, analytic boundaries with respect to three areas must be addressed: calculation of net values, level of government, and subsidies to complements.

2.1.3.1 Calculation of Net Values Because interventions can act as taxes or subsidies, interventions should be treated holistically so that end values represent net, rather than gross, subsidies to energy. Policies that affect multiple sectors need to be prorated so that only the portion applicable to energy is counted. Although proration is not always possible to do precisely (some policies have joint effects), reasonable allocations based on intensity of use, share of production, or similar metrics often provide reasonable proxies. Similarly, taxes or fees levied on fuels should be counted against the gross subsidy to generate the value net of offsets.

2.1.3.2 Level of Government Interventions occur at multiple levels of government, and all levels can affect energy costs to some degree. Analysis of these policies should be internally consistent; if tax offsets at the state level are deducted from energy subsidy values, state-level subsidies should be included as well.

2.1.3.3 Subsidies to Complements Energy subsidies have many second-order effects as they flow through other activities in the economy. Although detailed assessment is beyond the scope of this article, these policies can influence basic aspects of economic structure such as materials production, recycling, and energy demand. Subsidized electricity to primary aluminum is endemic and inhibits adoption of less energy-intensive materials. Similarly, subsidies for converting wastepaper, animal waste, and landfill gas into energy reduce the viability of recycling and composting alternatives. Widespread subsidies to driving spur increased road use and demand for gasoline.

2.2 Methods of Transferring Value

Table I provides an overview of intervention types. Depending on the policy specifics, many interventions generate either a net subsidy or a net tax. Cash transfers from government to private industry may originate from a handful of the intervention types, including direct spending, government ownership, and research and development (R&D) support. However, subsidies resulting from avoided expenditures by private firms are more common. These include government provision of market-related information; access to below-market credit, insurance, or government-provided goods and services; tax breaks; and exemptions from prudent regulation on health-, environment-, or safety-related aspects of an enterprise.

Market controls, including general regulations, provisions governing access to resources, restrictions on energy-related imports or exports, and purchase requirements can act as either a tax or a subsidy depending on one's market position. For example, past restrictions on oil exports from Alaska acted as a *de facto* tax on Alaskan producers because they could not sell output to the highest bidder. However, the very same policy provided oil subsidies to consumers on the West Coast of the United States. Although driven by government policy and having important effects on energy market structure, these policies often involve transfers between various producers and consumers rather than transfers from taxpayers. Cross-subsidies follow a similar pattern, with some users paying less than they should and others paying more than they should. They commonly occur when rate structures must be approved by governments or in markets protected from competition.

Special energy taxes, as their name suggests, tend to act as taxes rather than as subsidies. However, a tax should be classified as "special" only if it is above and beyond appropriate baseline recovery of revenue. The baseline taxation of energy should (1) compensate public sector owners for the sale of valuable energy resources, (2) recover public sector costs associated with the public provision of energy-related services, (3) equal the baseline tax on

TABLE I

Common Forms of Government Interventions in Energy Markets

Intervention type	Description
Access[a]	Policies governing the terms of access to domestic onshore and offshore resources (e.g., leasing)
Cross-subsidy[a,b]	Policies that reduce costs to particular types of customers or regions by increasing charges to other customers or regions
Direct spending[b]	Direct budgetary outlays for an energy-related purpose
Government ownership[b]	Government ownership of all or a significant part of an energy enterprise or a supporting service organization
Import/Export restriction[a]	Restrictions on the free market flow of energy products and services between countries
Information[b]	Provision of market-related information that would otherwise have to be purchased by private market participants
Lending[b]	Below-market provision of loans or loan guarantees for energy-related activities
Price controls[a]	Direct regulation of wholesale or retail energy prices
Purchase requirements[a]	Required purchase of particular energy commodities, such as domestic coal, regardless of whether other choices are more economically attractive
Research and development[b]	Partial or full government funding for energy-related research and development
Regulation[a]	Government regulatory efforts that substantially alter the rights and resposibilities of various parties in energy markets or that exempt certain parties from those changes
Risk[b]	Government-provided insurance or indemnification at below-market prices
Tax[a,b]	Special tax levies or exemptions for energy-related activities

Source: Koplow, D. (1998). "Quantifying Impediments to Fossil Fuel Trade: An Overview of Major Producing and Consuming Nations." Paper prepared for the OECD Trade Directorate.

[a] Can act either as a subsidy or a tax depending on program specifics and one's position in the marketplace.

[b] Interventions included within the realm of fiscal subsidies.

other commodities, and (4) charge an appropriate levy for negative externalities associated with production and use of the resource. Many assessments of energy taxation fail to incorporate appropriate measures for baseline levels, improperly classifying as an "energy tax" or an "environmental tax" policies that in fact recover only a portion of public costs and leave residual subsidies.

Government ownership of energy-related enterprises, including power generation and transmission, oil production and refining, coal mines, and road and pipeline networks, is common around the world. Many of these activities generate large subsidies to consumers as well as depleting fiscal resources. Subsidies are multilayered. The enterprise is often fairly high risk, attains access to low-cost tax-exempt government debt, pays no taxes on net income, and is not expected to earn a return on capital. Operating losses may ensue above and beyond these cost structure subsidies due to poor controls or politicization of the rate structure. These enterprises can be complex and difficult to analyze but often contribute to significant energy market distortions.

2.3 Methods of Measuring Subsidy Magnitude

Efforts to assess subsidy magnitude have generally focused either on measuring the value transferred to market participants from particular programs (program-specific approach) or on measuring the variance between the observed and the "free market" price for an energy commodity (price gap approach). One set of methods that captures both pricing distortions (net market transfers) and transfers that do not affect end market prices (net budgetary transfers) is the producer subsidy equivalent (PSE) and consumer subsidy equivalent (CSE) metrics commonly employed in the agricultural sector. Use of these in the energy sector has so far been limited to annual assessments of PSEs for coal in a handful of countries.

These approaches differ in the amount of data required to calculate them and in the degree to which they successfully measure budget transfers plus market transfers. Program-specific transfer assessments capture the value of government programs benefiting (or taxing) a particular sector, whether these benefits end up with consumers (as lower

TABLE II

Overview of Subsidy Measurement Approaches

Approach/description	Strengths	Limitations
Program specific: Quantities value of specific government programs to particular industries: aggregates programs into overall level of support	Captures transfers whether or not they affect end market prices Can capture intermediation value (which is higher than the direct cost of government lending and insurance	Does not address questions of ultimate incidence or pricing distortions Sensitive to decisions on what programs to include Requires program-level data
Price gap: Evaluates positive or negative "gaps" between the domestic price of energy and the delivered price of comparable products from abroad	Can be estimated with relatively little data; very useful for multicountry studies Good indicator of pricing and trade distortions	Sensitive to assumptions regarding "free market" and transport prices Understates full value of supports because ignores transfers that do not affect end market prices
PSE/CSE: Systematic method to aggregate transfers plus market supports to particular industries	Integrates transfers with market supports into holistic measurement of support Separates effects on producer and consumer markets	Data intensive Little empirical PSE/CSE data for fossil fuel markets

Source: Koplow, D., and Dernbach, J. (2001). Federal fossil fuel subsidies and greenhouse gas emissions: A case study of increasing transparency for fiscal policy, *Annu. Rev. Energy Environ.* **26,** 361–389.

prices), producers (through higher revenues), or resource owners (through higher rents). Unless integrated into a macroeconomic model, this information tells little about the ultimate incidence of the subsidy programs and their effect on market prices. By definition, the price gap metric highlights observed price distortions, although it misses the often substantial fiscal supports that do not affect consumer energy prices but do affect the structure of supply. The combination of PSE and CSE data provides insights into both. Table II briefly summarizes the main approaches that have been used in both domestic and international subsidy assessments as well as their respective strengths and limitations.

3. SUBSIDIES THROUGH THE FUEL CYCLE

Because no two fuel cycles are exactly the same, examining subsidies through the context of a generic fuel cycle is instructive in providing an overall framework from which to understand how common subsidization policies work. Subsidies are grouped into preproduction (e.g., R&D, resource location), production (e.g., extraction, conversion/generation, distribution, accident risks), consumption, postproduction (e.g., decommissioning, reclamation), and externalities (e.g., energy security, environmental, health and safety).

3.1 Preproduction

Preproduction activities include research into new technologies, improving existing technologies, and market assessments to identify the location and quality of energy resources.

3.1.1 Research and Development

R&D subsidies to energy are common worldwide, generally through government-funded research or tax breaks. Proponents of R&D subsidies argue that because a portion of the financial returns from successful innovations cannot be captured by the innovator, the private sector will spend less than is appropriate given the aggregate returns to society. Empirical data assembled by Margolis and Kammen supported this claim, suggesting average social returns on R&D of 50% versus private returns of only 20 to 30%.

However, the general concept masks several potential concerns regarding energy R&D. First, ideas near commercialization have much lower spillover than does basic research, making subsidies harder to justify. Second, politics is often an important factor in R&D choices, especially regarding how the research plans are structured and the support for follow-on funding for existing projects.

Allocation bias is also a concern. Historical data on energy R&D (Table III) demonstrate that R&D spending has heavily favored nuclear and fossil energy across many countries. Although efficiency,

TABLE III

Federal Research and Development Support by Energy Type

Region	Nuclear fission	Fossil fusion	Fossil energy	Renewables, efficiency, conservation	Other[a]
United States (percentages)					
1950–1993[b]	49.2	13.1	21.5	16.2	0
1998–2003[c]	17.7	16.2	32.6	30.2	3.4
IEA members[d](percentages)					
1974–1993	49.7	10.8	14.1	14.0	11.4
1994–1998	39.3	11.0	10.6	20.8	18.4
Total spending (billions of 2001 dollars)					
1974–1998	117.3	26.5	33.0	36.8	30.6
			Coal: 23.5	Renewables: 19.8	
			Oil and gas: 9.5	Conservation: 17.0	

[a] Includes electrical conversion and distribution, energy storage, and unclassified spending.
[b] Koplow, D. (1993). "Federal Energy Subsidies: Energy, Environmental, and Fiscal Impacts: Technical Appendix." Alliance to Save Energy, Washington, DC.
[c] U.S. Department of Energy. (2003). "Budget Authority History Table by Appropriations."
[d] International Energy Agency. (2003). *Research and Development Database*.

renewables, and conservation have captured a higher share of public funds during recent years, the overall support remains skewed to a degree that may well have influenced the relative competitiveness of energy technologies. Extensive public support for energy R&D may also reduce the incentive for firms to invest themselves. U.S. company spending on R&D for the petroleum refining and extraction sector was roughly one-third the multi-industry average during the 1956–1998 period based on survey data from the U.S. National Science Foundation. For the electric, gas, and sanitary services sector, the value was one-twentieth, albeit during the more limited 1995–1998 period.

3.1.2 Resource Location

Governments frequently conduct surveys to identify the location and composition of energy resources. Although these have addressed wind or geothermal resources on occasion, they most often involve oil and gas. Plant siting is another area where public funds are used, primarily to assess risks from natural disasters such as earthquakes for large hydroelectric or nuclear installations. Survey information can be important to evaluate energy security risks and to support mineral leasing auctions, especially when bidders do not operate competitively. However, costs should be offset from lease sale revenues when evaluating the public return on these sales. Similarly, the costs of siting studies should be recovered from the beneficiary industries.

3.2 Production

Energy production includes all stages from the point of resource location through distribution to the final consumers. Specific items examined here include resource extraction, resource conversion (including electricity), the various distribution links to bring the energy resource to the point of final use, and accident risks.

3.2.1 Extraction of Energy Resources

General procedures for leasing access to energy minerals on public lands and more general subsidies for promoting energy extraction both are important areas to evaluate. Extraction-related subsidies are most common for oil and gas production, although they also support nuclear fission (due to uranium mining), geothermal, and coal.

3.2.1.1 Accessing Publicly Owned Energy Resources Terms of access for energy minerals on public lands can be a source of enormous subsidies. In

countries where leases or concessions are granted through graft rather than competitive bidding, wealth transfers worth billions of dollars can occur. Although there are not good statistics on the losses, the problem appears to be large. Oxfam America finds that states most dependent on oil tend to have very low Human Development Index (HDI) rankings. The HDI, developed by the UN Development Program, ranks states according to a combined measure of income, health, and education. Transparency International finds strong linkages between large mining and petroleum sectors as well as elevated levels of bribery and corruption. Low-cost access to energy minerals also tends to remove the incentive for careful management because profits can be had even with inefficient operation. Lease operation can also generate subsidies such as when self-reported royalties are calculated improperly. The Project on Government Oversight has documented state and federal court awards in excess of $10 billion in response to litigation in the United States over oil and gas royalty underpayments.

3.2.1.2 Promoting Extraction Activities Policies to reduce the cost of extraction are widespread and often take the form of tax or loan subsidies or royalty concessions. They are found at both the national and state levels. Particular market niches may be targeted, from geographical (e.g., deep sea recovery of oil, timbering in a particular forest), to technological (e.g., tax breaks for more advanced oil drilling or coal gasification), to life cycle related (e.g., lower royalties on idle wells that are restarted). In some cases, baseline tax policy may be applied by firms in creative ways to generate large subsidies. U.S.-based multinationals receive a tax credit for foreign taxes paid to avoid double taxation of foreign income. Yet in many oil-producing regions with low or no corporate income taxes, foreign governments have reclassified royalty payments into corporate taxes, generating a tax write-off estimated by Koplow and Martin at between $0.5 billion and $1.1 billion annually.

However, many subsidies to extraction are not restricted to particular market niches. Percentage depletion allowances in the United States allow most firms mining oil, gas, uranium, or coal to deduct more costs from their taxable income than they have actually incurred. Accelerated write-offs of extraction-related investments are also common. For example, many multiyear costs in the U.S. oil and gas industry may be deducted immediately (expensed) rather than over the useful lives of the investments. All of these special provisions tend to reduce the effective tax rate on benefiting energy industries. Data collected by the Energy Information Administration (EIA) suggest that the major U.S. energy firms paid federal taxes that were one-quarter to one-half the prevailing nominal rates between 1977 and 1995.

3.2.2 Conversion

Raw energy materials normally go through some conversion prior to consumption. Crude oil is refined into a wide range of specialized products such as gasoline and heating oil. Coal may be pulverized or cleaned prior to use. A combination of heat and machinery converts raw fuels (including wind and solar) into electricity. Common government supports to the conversion stage include capital subsidies, production tax credits or purchase requirements, and exemptions from appropriate protections for environmental quality, worker health, and accident risks. Because this third category affects multiple phases of the fuel cycle, it is addressed in a separate section.

3.2.2.1 Capital Subsidies Subsidies to capital formation, usually through accelerated depreciation or investment tax credits, are common. Although applicable to multiple economic sectors, they are often of great benefit to energy producers. This is due both to their relative capital intensity and to provisions in the tax code that grant special accelerated depreciation schedules for energy-related assets. For example, in the United States, three sectors of relevance to energy—electric light and power, gas facilities, and mining, shafts, and wells—have allowable depreciation schedules that are 28, 45, and 44% faster, respectively, than the actual economic depreciation of their assets according to data compiled by the U.S. Treasury. Capital subsidies are of greatest benefit to large-scale generation assets with long construction times (nuclear, hydro, and coal) and are of greatest detriment to energy resources that conserve capital (most prominently energy conservation).

3.2.2.2 Tax Credits/Purchase Mandates A second class of subsidies to the conversion stage are tax credits or purchase mandates for certain types of energy. These subsidies occur at both the federal and the state/provincial levels and most often support emerging power sources such as solar, wind, and biomass-based electricity. Whereas many of the subsidies to conventional power sources are expensive regardless of whether the energy investments ultimately succeed, the tax credits and purchase mandates tend to be more efficient. For example,

federal tax credits for wind energy in the United States cost taxpayers nothing unless a private investor is successful in getting a wind plant operating. If the plant goes offline, so too do the credits. Renewable portfolio standards (RPSs), a common form of purchase mandates adopted by many U.S. states, are even more efficient. In addition to providing no subsidy unless the power is delivered, RPSs often compete eligible power sources against each other, driving down the unit subsidy as technologies improve.

Despite their benefits, these approaches have run into some political problems. Specifically, as the subsidies have grown, so too has lobbying pressure to expand the range of eligible sources. Federal tax credits now include poultry waste, a great benefit to the handful of very large chicken processors. At the state level, unsustainable biomass sources are sometimes included, as are waste-to-energy and landfill gas systems. Thus, although energy diversification goals are still being met, the supply is not necessarily renewable or particularly clean.

3.2.3 Transportation and Distribution

Fuel cycles may involve multiple transportation steps, including movement of raw fuels to point of refining, refined fuels to the point of consumption, and movement of wastes to disposal sites. Relevant modes of transport include road, rail, water, pipelines, and transmission lines.

Although specific energy resources vary widely in their transport intensity and in the modes of transportation and distribution on which they rely (Table IV), there are some common themes. Government construction, maintenance, and operation of transportation infrastructure frequently give rise to subsidies when user fees do not cover costs. These subsidies are often understated because municipalities might not properly cost the resources being

TABLE IV

Impact of Transport Subsidies on the Energy Sector

Transport mode	Issues	Energy sector impacts
Water: Inland	Waterway maintenance often provided by governments; user fees may not recover costs	Reduces delivered price of bulk oil and coal
Water: Coastal and international	Coastal ports, harbors, and shipping oversight subsidized by federal and other government entities; user fees might not recover costs	Reduces delivered price of bulk oil and coal
	Fuel consumed during shipment in international waters generally tax free	
Road	Most roadways are municipally owned and operated; user fees (primarily from fuel taxes) often insufficient to cover costs	Primarily benefits refined petroleum products
	Large trucks often pay proportionately less in taxes than the damage they cause roadways	Waste products from coal combustion or waste-to-energy plants may sometimes move by truck as well
Rail	Many rail lines do not recover their full costs	Largest beneficiary is coal, with some benefits to oil
Pipeline	Rights of way, safety and security, and environmental cleanup contribute to reduced costs of pipeline ownership and operation	Primarily benefits oil and natural gas
	Depending on circumstances, government ownership may generate large subsidies to users and use government monopoly to levy high taxes on users	
Electrical Transmission Grid	Rights of way, tax breaks for municipal ownership or capital investment, and government research and development can generate subsidies to eletrical distribution	Benefits all sources of centralized electricity in proportion to their share as a prime mover in generating stations; coal, nuclear, natural gas, hydroelectricity, and oil are the main beneficiaries
	Inaccurate pricing of distance can generate cross-subsidies to rural users	

Source: Earth Track Inc., Cambridge, MA.

consumed. For example, tax exemptions on transportation bonds used to finance roads are routinely ignored, as are the free grants of rights-of-ways for rail, road, pipeline, and transmission links. So too is the opportunity cost of land covered by roadways and parking facilities. Although this space occupies 1.7, 2.1, and 3.5% of the total land area in the United States, Germany, and Japan, respectively, Todd Litman of the Victoria Transport Policy Institute noted that no property tax is paid on the vast majority of this space. This understates the direct costs of the infrastructure and the rights to use it.

Cross-subsidies between user groups may further distort relative prices. Large trucks pay less in highway fees than the damage they cause, generating an incremental subsidy to deliveries of refined fuels such as gasoline. Deep-berth ships such as large oil tankers may be the primary drivers of channel- or port-deepening projects, yet they often contribute to costs based only on volume of shipments. In the electricity sector, transmission tariffs may represent broad averages of the cost of service rather than rising as the distance traveled and density of users decline. By delivering subsidized electricity to remote users, transmission cross-subsidies mask the cost of line maintenance and new construction. This can destroy niche markets in which off-grid technologies (often renewable) would otherwise have been able to compete. Cross-subsidies between peak pricing and low demand periods are also common in electricity markets because real-time metering is not widely used at the retail level. This can dampen retail investments in demand-side management.

Power sources such as wind and solar require no shipment of input fuels or waste. Improved energy efficiency and some off-grid technologies require no transmission grid either. As a result, subsidies to energy transport can increase the barriers to renewable energy and efficiency. A major U.S. study conducted by Cone and colleagues in 1978 found that an estimated $15.2 billion in federal money subsidized transport of U.S. oil stocks between 1950 and 1977. The policies generating these subsidies have continued during the ensuing quarter-century or so.

3.2.4 Accident Risks

A handful of energy activities have the potential to cause catastrophic harm, including large oil spills, dam failures, and nuclear accidents. Many governments cap, shift, or ignore the potential liabilities from these activities. Functioning insurance markets and litigation would normally help to drive up prices for the more dangerous energy sources or particularly negligent operators. Government policies that mask these signals impede substitution to safer alternatives.

3.2.4.1 Large Oil Spills Within the United States, the Oil Pollution Act of 1990 stipulates use of commercial insurance for a first tier of insurance. A public trust fund financed by levies on oil sales provides supplemental coverage, although payments out of the fund are capped at $1 billion per incident. Based on empirical assessments of spill cleanup costs by Anderson and Talley, at least five spills over the past three decades or so would have exceeded the $1 billion cap, although most spills will be adequately covered. Internationally, the 1992 Civil Liability Convention governs liability for oil spills, also using a two-tier system. Insurance held by the vessel owner provides the first tier. Levies on cargo owners feeds the second tier, with receipts held in the International Oil Pollution Compensation Fund. The maximum compensation available from both tiers is roughly $174 million, a level shown to be insufficient by spills occurring in both 1997 and 1999. Although the caps are likely to be raised by 50%, Alfred Popp, chairman of the group working on the latest rounds of reforms, noted that concerns about liability shortfalls persist. The subsidy value of these caps is not known.

3.2.4.2 Dam Failures Many activities that would pose a very large potential risk if accident scenarios materialized rely on a system of strict liability. Strict liability focuses only on magnitude of the potential damages rather than on the intent, negligence, or degree of care of the operator. Although the failure of a large dam near a populated area can cause catastrophic loss of life, assurance for such potential liabilities is poorly characterized. Although loss of life from a dam failure will likely trigger widespread litigation, the rules of that litigation are predominantly set at the state level. Analysis by Denis Binder for the Association of State Dam Safety Officials indicates that a slight majority of states reject strict liability in dam failures. Furthermore, the piecemeal approach to coverage within the United States makes it difficult to evaluate whether existing coverages are adequate. Poor characterization of the risks extends to the international arena as well. To the extent that liability insurance is not in place or is too low, subsidies to hydroelectricity would result.

3.2.4.3 Nuclear Accidents Nuclear accidents can expose large populations to dangerous levels of radioactivity, triggering enormous liabilities for the

firm responsible. Caps on nuclear liability are common throughout the world. The United States, under the Price–Anderson Act, has a two-tier system of indemnification: a first tier of commercial insurance ($300 million per reactor) plus a second pooled tier (maximum of $83.9 million per reactor) funded by retroactive assessments on all reactors in case any reactor has an accident. Japanese nuclear operators must provide financial security of $520 million; damages above that amount will be paid by the government. In China, the limit is roughly $36 million. In Ukraine, it is roughly $70 million.

International efforts to standardize liability under the Convention on Supplementary Compensation for Nuclear Damage would establish minimum liability coverage worldwide, although for many countries this would also constitute the maximum. Under the convention, operators would directly face a first tier of liability. A country fund would provide secondary coverage. Because country payments rely on a sovereign guaranty rather than a prefunded instrument such as a trust fund, they may be at some risk.

Aggregate coverage under the U.S. system is estimated at roughly $9.2 billion per accident, although most of this is paid out over nearly 9 years by utilities, so the present value of the coverage is substantially lower. Liability levels established under the convention would provide less than $900 million per accident. Loss statistics from the Insurance Services Office and from the Disaster Insurance Information Office provide some context. Since 1950, there have been approximately 20 hurricanes with adjusted damages in excess of the convention cap, and both Hurricane Andrew and the Northridge earthquake had damages that exceeded the Price–Anderson cap even before adjusting retroactive premiums to present values.

Subsidies arise when government caps fall below expected damages from an incident and caps under both Price–Anderson and the convention are likely to do so. Damages above that level are, in effect, shifted to the state or to the affected population. Heyes estimated that the subsidy to reactors under Price–Anderson ranges between 2 and 3 cents per kilowatt-hour, a value that would roughly double the operating costs of nuclear plants. In addition, there are incremental subsidies associated with indemnification for nuclear contractors and government-owned facilities. Because other countries have lower liability caps and weaker inspection regimes, they likely have higher liability subsidies as well.

3.3. CONSUMPTION

Government support for energy consumption falls into three main categories: poverty alleviation, economy-wide below-market pricing, and targeted subsidies for certain classes of consumers.

3.3.1 Poverty Alleviation

Subsidies to heat and power for poorer citizens are common, frequently in the form of a lump-sum grant or reduced cost access to municipal resources. Often consumption oriented, these subsidies may miss opportunities to implement conservation measures among the target populations. Targeting can be a problem as well, with funds not reaching the groups most in need. According to the International Energy Agency (IEA), the poorest citizens often rely on noncommercial fuels such as dung (biomass comprises as much as 80% of the energy market in rural countries with a high reliance on subsistence agriculture) or live outside the reach of the subsidized electrical grid.

3.3.2 General Subsidies

Nations with large domestic energy industries sometimes institute policies that keep local prices well below world levels. These subsidies may protect antiquated energy-consuming industries that otherwise would be unable to compete, or they may serve as "rewards" to the electorate for supporting a particular official. For example, price gap data for Venezuela and Iran compiled by the Organization for Economic Cooperation and Development (OECD) and IEA show that these large oil producers heavily subsidize both industrial and residential use of petroleum. Subsidies are also common in many service areas close to large municipal hydroelectric generating stations. For example, rates to customers of the Power Marketing Administration dams in the United States were long heavily subsidized. Although the quantities of power or oil flowing through these regions make these subsidies seem costless, they are not. Domestic sales at subsidized rates forgo energy export revenues, increase local pollution, and contribute to a production base that is increasingly noncompetitive with that deployed elsewhere in the world.

3.3.3 Targeted Exemptions

Most OECD countries exempt coal and heavy fuel oils used in industry, as well as aviation fuels used on international flights, from the baseline levies on energy. Excise tax rates on coal used in the industrial

or power sector are often lower than those on much cleaner natural gas. The OECD noted that these exemptions "effectively mean that a large proportion of total carbon emissions in OECD countries is untaxed," generating weaker incentives to adopt even low-cost abatement options.

3.4 Postproduction Activities

Energy production and conversion require large facilities, often located in remote or pristine environments. Postoperational cleanup can be complex. Decommissioning addresses removal of physical infrastructure, whereas remediation and reclamation address problems with land and water. For markets to make accurate decisions about the relative cost of energy resources, the cost of these postproduction activities must be included in energy prices during the operating life of the facility in much the same way that the cost of an employee pension would be. Indeed, failure to accrue funds for postclosure costs during operations would make public subsidy likely given that revenues often drop to zero on plant closure.

3.4.1 Decommissioning

Decommissioning subsidies arise when infrastructure removal costs are ignored or underestimated or when accrued funds are mismanaged. Costs can be significant at large-scale energy installations such as hydroelectric dams and oil refineries. Where installations are remote (e.g., offshore oil rigs), radioactive (e.g., nuclear plants), or widely dispersed (e.g., gathering pipelines), costs per unit of capacity can be particularly high. Requirements for long-term environmental or safety monitoring (e.g., nuclear plants and some mines) can drive costs up further.

Pipelines and hydroelectric dams provide examples of costs being ignored entirely. Koplow and Martin made inquiries to many U.S. officials regarding pipeline closure. They found that although there are regulations governing proper abandonment, advance funding of closures was not required. The risks of insolvency appeared to be fairly high, especially for the smaller companies that often own older gathering pipelines. Regarding dams, the U.S. Federal Energy Regulatory Commission indicated in a 1994 policy statement that it will "not generically impose decommissioning funding requirements on licensees" but rather will stipulate them on a case-by-case basis at the time of relicensing. According to Andrew Fahlund of American Rivers, this policy has been implemented such that if a "dam owner is too poor, it is too burdensome to require them to maintain a fund, and if they are rich, they will have plenty of money available for such an eventuality."

Underestimating decommissioning requirements is of great concern with nuclear plants. IEA data indicate that the anticipated cost per unit of power capacity can vary by a factor of 10 across plants. IEA multicountry data suggest median decommissioning values of between 21 and 37% of the overnight capital cost (i.e., before financing) to build the plant. If funds are not properly accumulated during the plant's operating life, taxpayer burdens will be large. Inadequate provision for closure is also apparent in the oil and gas sector. Koplow and Martin found shortfalls in funding to plug and abandon oil wells in the United States approaching $600 million per year, of which approximately 75% represented insufficient bonding at wells still in operation.

Public bailouts can also be required if accrued funds for postclosure activities are lost through negligence, bankruptcy, or theft. If funds are retained within the firm, bankruptcy is a significant risk, especially given the 40- to 60-year time frame between fund collection and use. Increased segregation of each energy asset into its own company (now becoming the norm in the U.S. nuclear industry) greatly increases this risk. Loss through negligence is less likely where regulations preclude speculative investing. Nuclear decommissioning trusts within the United States are held outside the firm and are subject to conservative investment requirements to reduce the likelihood of loss.

3.4.2 Reclamation and Remediation

Small subsidies to site reclamation and remediation may arise through government-sponsored research into remediation technologies or through regulatory oversight of extraction activities that are not recovered via user fees. Much larger subsidies are associated with remediation of government-owned energy-related installations or where reclamation bonding has been insufficient to pay for the damage caused by private operators. James Boyd at Resources for the Future pointed to widespread inadequacy of reclamation bonding levels. For example, in the U.S. states of Indiana, Kentucky, and Tennessee, reclamation of coal mine sites is below 20%. Reclamation bond levels have generally been inadequate. Estimated liability for high priority (public health and safety concerns) coal mine remediation in the United States is $6.6 billion, according to the U.S. Office of Surface Mining Reclamation and Enforcement. Many mining regions

around the world have unreliable, incomplete, or nonexistent data on abandoned mines and their associated costs. These shortfalls may be made up by general tax revenues. However, more often, resource damage is not mitigated and continuing environmental releases are not controlled. Spending to address environmental concerns at nuclear energy-related infrastructure owned by the U.S. government has run approximately $500 million per year, much of which is paid by general taxpayers.

3.5 Energy Externalities

External costs of energy production and consumption can include pollution, land degradation, health impairments, congestion, and energy security. This article differentiates between two types of subsidies. The first involves existing government spending to address recognized problems associated with particular energy resources. Included here would be public funding to protect energy supplies and assets; public absorption of energy worker health care costs; and/or public subsidies to pollution control or abatement. Because this spending involves actual outlays, it is counted as a fiscal subsidy. A second class of policies involves loopholes in regulatory controls that allow additional damages to human health or the environment to continue without compensation. This second group is often difficult to quantify and is segregated as an externality.

3.5.1 Energy Security

Energy plays a central role in industrialized economies, and supply disruptions can trigger widespread economic dislocations. Geopolitical problems, accidents, and terrorism all are potential triggers. Lovins and Lovins identified a handful of factors that drive security concerns. These include long distribution channels, geographically concentrated delivery or production systems, interconnected systems that can spread failures, specialized labor and control systems to operate capital-intensive facilities that are very difficult to replace, and dangerous materials that can elevate the severity of any breach.

Energy security strategies include protection of energy-related assets and supply routes, stockpiling of vulnerable resources, and supply diversification. Where costs of these responses are borne by the general public rather than by the appropriate energy producers and/or consumers, the market incentive to build a more resilient, decentralized, and diversified supply system is reduced. Security subsidies tend to benefit oil the most, with particularly high transfers to imported oil from unstable regions such as the Persian Gulf. Additional beneficiaries are centralized electricity and natural gas. Off-grid power and conservation are the sources most disadvantaged. Subsidies to protecting energy installations and stockpiling are explored in detail in the following subsections.

3.5.1.1 Protection of Assets and Supply Links The larger the energy installation, the greater the target and the bigger the dislocation that an attack or accident would cause.

Defending energy-related assets is an increasing concern of governments around the world. Pipeline defense is listed as its own objective within Georgia's defense and security strategy. The United States has become involved with training the Colombian military to defend oil pipelines in that country, pushing for funding of $98 million to support the effort. Within the United States, core assets include the Trans-Alaska Pipeline System (TAPS), through which nearly 25% of total U.S. crude production flows, and nuclear plants. In response to inquiries from Koplow and Martin, Alaskan and federal officials said that no public funds were spent ensuring TAPS security. Nonetheless, the military has historically conducted training and planning exercises around the pipeline. In the nuclear sector, increased public subsidies have come through rising deployment of state-level security or National Guard troops around plants during periods of high terrorist alerts. However, surveys of nuclear plant workers by the Project on Government Oversight reveal employee concerns that training and spending levels are still insufficient. Although these anecdotes indicate that public expenditures in the area of protecting energy-related assets are likely large, data to quantify the subsidies are generally unavailable.

The costs of defending oil shipments through the Persian Gulf is an exception. As one of three central missions for the U.S. military in the region, there have been multiple efforts to value the subsidy to oil. Koplow and Martin reviewed eight historical studies of these costs and found general agreement that this presence is of great benefit to oil supply security. Disagreements centered on cost attribution. Some assessments attributed an extremely small portion of the military cost to oil, arguing that the same basic force structure would be needed for the other missions. Koplow and Martin pointed out that equivalent arguments could be made for each mission area given that the common costs of the vessels and personnel are what are most expensive.

They argued instead for treating the military presence through the lens of joint costs and allocating a reasonable portion (in this case, one-third) to the oil sector. This approach yields a subsidy to the oil sector in the range of $11.1 billion to $27.4 billion per year (roughly $1.65–$3.65/barrel originating from the region), depending on which of the detailed costing studies are used. Although funded by U.S. taxpayers, the benefits accrue to oil consumers in Europe and Japan as well. Recovering this cost via an excise fee on shipments would help to encourage increased supply diversification.

3.5.1.2 Stockpiling Petroleum has been the main focus of stockpiling efforts given its importance to world transport and military readiness. Under the terms of the IEA, oil-importing member states are required to hold stocks equal to 90 days of the previous year's net oil imports as a buffer against short-term supply disruptions. Subsidies arise if the costs of stockpiling are borne by taxpayers rather than by oil consumers. Relevant expenses include constructing and operating the stockpiles, interest costs on oil inventories and infrastructure, and any payments to third parties for nongovernmental stockpiling (two-thirds of IEA-mandated stocks are held commercially).

Buffer stocks for oil within the United States are held within the publicly owned Strategic Petroleum Reserve (SPR). The SPR has incomplete cost accounting, most prominently ignoring the interest costs associated with more than $16 billion it has spent to purchase its oil inventory since the reserve's inception. Private firms must finance all working capital, including inventory, in their operations, and cost savings from reducing inventory levels can be large. Public oil stockpiles are no different; capital tied up in the enterprise much be borrowed, at interest, through Treasury bond markets. Analysis by Koplow and Martin for 1995 estimated annual subsidies to the SPR at between $1.7 billion and $6.l billion, depending on whether unpaid interest on oil inventories is compounded. Because carrying costs are sensitive to the cost of capital, declining interest rates during recent years mean that current SPR subsidies will be lower than they were during the mid-1990s. Although the details of stockpile financing in other countries are not easy to discern (the IEA collects data only on physical flows, not on financial flows), some countries do recover at least a portion of their stockpiling costs from consumers. These include Japan, France, Germany, Korea, and Taiwan.

Subsidies to stockpiles slow transition to less vulnerable, more diversified supplies. Formal tracking of stockpile finance by the IEA, as well as the formalization of accounting rules for calculating costs, would leverage market forces for improved supply security.

3.5.2 Environmental, Health, and Safety Externalities

Externalities involve damages associated with energy production or use that are imposed on surrounding populations or ecosystems without compensation. These may include environmental damage, materials damage, human health effects, and nuisance factors such as bad smells and loud noises. Worker health is sometimes not counted as an externality under the argument that workers are compensated for the added risks of their jobs through higher wages. Such a conclusion requires that workers have some degree of choice in whether or not to accept jobs and that employers can be taken to task retroactively for gross negligence. This is not the case in many countries around the world. As a result, it is reasonable to consider as subsidies high levels of occupational illness, especially when the costs of maintaining those workers falls on the general taxpayers.

Governments are routinely involved with efforts to make certain energy-related activities safer for workers. This is most prominent regarding coal and nuclear fuel cycles, where dedicated government agencies exist to inspect and educate mines and production sites. If these costs are not paid entirely by the producers or consumers of the affected energy type, subsidies ensue. Public responsibility for workers' health care and/or pension costs also constitute subsidies. This has been quite common in the area of coal. For example, government payments to U.S. coal miners afflicted with black lung have exceeded $30 billion. Black lung levels are now rising (or are being better documented) in other countries such as Russia, Ukraine, and China. Coal mine fatalities continue at extremely high levels in many of these countries as well.

Although difficult to quantify and normalize, external costs arise in many different ways through the fuel cycle. Table V provides an overview of these costs, in large part relying on data developed by ExternE. A project of the European Commission, ExternE has more than 50 research teams trying to quantify the externalities for key fuel cycles.

External costs are highest for coal and oil, although newer conversion technologies can reduce the toll per unit of power significantly. The environmental costs of nuclear power also appear low, even

TABLE V

Overview of Externalities by Energy Resource

Fuel	Years of life lost/TWh[a]	Environmental damage costs (Euro-cents/kWh)[a]	Quantifiable external costs (Euro-cents/kWh)[b]
Coal, 1995	1065	13	2.6
Coal, >2000	113	4	3.4 (lignite)
Oil, 1995	830	10	
Oil, >2000	139	2.6	
Gas, 1995	161	2.4	1.1
Gas, >2000	29	1.7	
Nuclear<100years	1	0	0.2
Nuclear>100years	8	0.2	
Biomass, range	20–100	0–0.8	
Photovoltaic, range	4–11	0.1–0.3	0.8
Wind	4–9	0	0.09
Hydro			0.07

[a] Rabl, A., and Spadaro, J. (2001). The ExternE project: Methodology, objectives, and limitations. *In* "Externalities and Energy Policy: The Life Cycle Approach", Nuclear Energy Agency, workshop proceedings. Organization for Economic Cooperation and Development, Paris.

[b] Includes available estimates for health effects crop losses, material damage, noise, acidification/eutrophication, and global warming. Viridis, M. (2001). Energy policy and externalities: The life Cycle approach. Organization for Economic Cooperation and Development, Paris. *In* "Externalities and Energy Policy: The Life Approach". Nuclear Energy Agency workshop proceedings.

though they are quite sensitive to assumptions regarding accidents at plants. The data also indicate that renewable energy sources would become far more competitive if the market were forced to account for external costs.

4. PATTERNS OF SUBSIDIZATION

Developing an aggregate picture of energy subsidization is extremely difficult due to the scope of policies affecting the sector and to tremendous fragmentation of the data across thousands of government ministries worldwide. There has been no global multifuel estimate of energy subsidy magnitude, although two efforts come close. Myers and Kent developed a global estimate for fossil ($119 billion/year) and nuclear energy ($12 billion/year), with an additional $200 billion per year in externalities. Van Beers and de Moor estimated global subsidies to all fuels ($250 billion/year, with more than 60% supporting fossil) but did not include externalities. Both estimates were built by aggregating data from preexisting smaller assessments rather than from a comprehensive policy survey. Different definitional and valuation approaches, as well as gaps in regions covered, make aggregation difficult. Although the resultant values indicate that energy subsidies are quite large, they are unlikely to provide much precision as to the true subsidy magnitude.

Common themes are nonetheless visible across many of the smaller studies. For example, Koplow and Dernbach found that the fossil share of total subsidies exceeded 55% in four of six major U.S. multifuel energy subsidy studies conducted since 1978. In the remaining two, methodological inconsistencies largely explain the deviation. Within the fossil category, coal and/or oil normally received a larger share than did cleaner natural gas. Estimates for nuclear varied more widely. Where liability caps for accidents and capital subsidies were included, nuclear subsidies comprised 25% or more of the totals. Where they were incorrectly excluded, reported nuclear subsidies were much lower. Renewables and efficiency garnered 5% or less in most of the studies. Only when many large subsidies to other fuels were excluded did the aggregate subsidy estimates fall low enough to bring the renewables/efficiency share above 15%.

This pattern across fuels is broadly consistent with trends discussed earlier in this article. Historical spending on R&D has heavily favored fossil fuel and nuclear power, as has the use of export credit agencies. Subsidies to capital formation have supported primarily these same fuels, whereas those to energy security have favored oil and nuclear energy.

Failure to internalize pollution or health externalities generates subsidies primarily to coal and oil. Yet for many large industrial consumers, these fuels are exempt from taxation throughout the OECD. Government support for renewables, efficiency, and conservation is extremely modest. There have been recent gains in the share of public R&D supporting these areas, and some newer energy resources are eligible for tax credits or are included within purchase mandates. However, the overall pattern is one in which established energy types, often with poor environmental profiles, continue to receive most of the public support provided to the energy sector.

5. SUBSIDY REFORM

Modeling efforts, summarized by Koplow and Dernbach, suggest that subsidy reform is likely to generate environmental gains. Within the United States, modeling suggests that reforms could achieve a substantial (18% of the nonsequestration reductions) portion of the Kyoto 2010 carbon reduction targets. Modeling of international reforms projects some global reductions in greenhouse gas emissions by 2010 (0.2–8.0%), although results are sensitive to assumptions regarding subsidy values and fuel substitution options. Given the hundreds of billions of dollars involved with current subsidies, reforms also seem likely to reduce fiscal expenditures in many countries throughout the world.

Energy subsidies and subsidy reform have attracted the attention of many international governmental and nongovernmental organizations, as well as country-level ministries, over the past 15 years or so. Despite this attention, successful reforms have been few. This outcome reflects the role of political economy in subsidy creation, continuation, and reform. Who gets public resources is often an intensely political decision. Powerful groups in society are best positioned to institute policies that generate transfers to themselves. Furthermore, the theory of rent seeking indicates that groups that have received subsidies in the past will invest at least a portion of those gains to ensure that the subsidies keep coming.

Fossil, nuclear, and hydro energy sources all have been around for a long time, involve large companies and/or large government ministries, and have sufficient scale to dedicate staff to political lobbying. In addition, because the cost of many subsidies rises as the installed base eligible for them grows, the large installed base of fossil, nuclear, and hydro also contributes to these resources capturing the lion's share of subsidies.

Given the strong political opposition to subsidy reform, a transitional process to precede policy change with much increased transparency makes sense. Initial steps to qualitatively identify and describe subsidies seems simple but can greatly change the political dynamics of subsidization by making recipients more visible to their competitors and the taxpayers. Quantifying the value of these transfers is the next step and helps policymakers to prioritize which subsidies are most important to address first. Consistent approaches should be used so that estimates done for different countries, or those done by different researchers, can be compared more easily. Separating externalities from fiscal subsidies is one element of standardization, as is applying consistent rules on offsets.

Modeling the impacts of these transfers on human, environmental, and fiscal health should come next, followed by modeling the effects of potential subsidy reforms. Summarizing a series of global workshops on energy subsidies, the UN Environment Program suggested that reforms should be implemented gradually and with strong transitional policies. A gradual approach might not make sense in all cases given that political forces may be able to stop some reforms entirely. However, transitional approaches that decouple payments from practices that harm human health or the environment should be pursued. Eligibility for existing subsidies can also be made contingent on acceptable environmental practices. Finally, new or replacement subsidies should be structured to leverage competitive markets (as do the RPSs) rather than providing support whether or not there is a successful outcome.

6. CONCLUSION

Subsidies remain a large presence in energy markets throughout the world. The scope, complexity, and politics of these policies help to explain why there is no global subsidy data set. However, all indications are that these subsidies cost hundreds of billions of dollars per year, impede market penetration of cleaner and more efficient methods of providing energy services, and increase damages to human health and the environment. Efforts to overcome the inherent political resistance to subsidy reform are needed, if only to greatly improve the ability to identify, describe, and quantify subsidies to particular fuels throughout the world.

Acknowledgments

Section II.C is from Koplow and Dernbach (2001) and is reprinted with permission from the *Annual Review of Energy and the Environment,* volume 26. Copyright 2001, Annual Reviews.

SEE ALSO THE FOLLOWING ARTICLES

Corporate Environmental Strategy • Market-Based Instruments, Overview • National Energy Policy: United States • National Security and Energy • Research and Development Trends for Energy • Strategic Petroleum Reserves

Further Reading

Anderson, E., and Talley, W. (1995). The oil spill size of tanker barge accidents: Determinants and policy implications. *Land Econ.* **71,** 216–228.

Cone, B., *et al.* (1978). "An Analysis of Federal Incentives Used to Stimulate Energy Production." Rep. PNL-2410REV. Pacific Northwest Laboratory, Richland, WA.

Heyes, A. (2002–2003, Winter). Determining the price of Price–Anderson. *Regulation*, pp. 26–30.

International Energy Agency. (1999). "World Energy Outlook: Looking at Energy Subsidies—Getting the Prices Right." IEA, Paris.

International Energy Agency. (2001). "Nuclear Power in the OECD." IEA, Paris.

Koplow, D. (1993). "Federal Energy Subsidies: Energy, Environmental, and Fiscal Impacts, main report and technical appendix." Alliance to Save Energy, Washington, DC.

Koplow, D. (1996). "Energy subsidies and the environment. *In* Subsidies and Environment: Exploring the linkages." Organization for Economic Cooperation and Development, Paris.

Koplow, D., and Dernbach, J. (2001). Federal fossil fuel subsidies and greenhouse gas emissions a case study of increasing transparency for fiscal policy. *Annu. Rev. Energy Environ.* **26,** 361–389.

Koplow, D., and Martin, A. (1998). "Fueling Global Warming: federal subsidies to oil in the United States." Greenpeace, Washington, DC.

Litman, T. (2002). "Evaluating Transportation Land Use Impacts." Victoria Transport Policy Institute, Victoria, British Columbia.

Lovins, A., and Lovins, H. (1982). "Brittle Power." Brickhouse Publishing, Andover, MA.

Margolis, R., and Kammen, D. (1999). Evidence of underinvestment in energy R&D in the United States and the impact of federal policy. *Energy Policy* **27,** 575–584.

Myers, N., and Kent, J. (2001). "Perverse Subsidies: How Tax Dollars Can Undercut the Environment and the Economy." Island Press, London.

Organization for Economic Cooperation and Development. (1998). "Improving the Environment through Reducing Subsidies," part II: "Analysis and Overview of Studies." OECD, Paris.

Organization for Economic Cooperation and Development. (2001). "Environmentally Related Taxes in OECD Countries: Issues and Strategies." OECD, Paris.

Rabl, A., and Spadaro, J. (2001). The ExternE project: Methodology, objectives, and limitations. *In* "Externalities and Energy Policy: The Life Cycle Analysis Approach," Nuclear Energy Agency workshop proceedings. Organization for Economic Cooperation and Development, Paris.

UN Environment Program and International Energy Agency. (2001). "Energy Subsidy Reform and Sustainable Development: Challenges for Policymakers," synthesis report, submission to the 9th session of the UN Commission on Sustainable Development.

van Beers, C., and Moor, A. (2001). "Public Subsidies and Policy Failures: How Subsidies Distort the Natural Environment, Equity, and Trade and How to Reform Them." Edward Elgar, Northhampton, MA.

Suburbanization and Energy

DERON LOVAAS
Natural Resources Defense Council
Washington, D.C., United States

1. **Suburban Sprawl**
2. **Transportation in Suburbia**
3. **Structures in Suburbia**
4. **Land Use, Transportation, and Energy Consumption in Other Countries**
5. **Transportation Alternatives for U.S. Suburbs**

Glossary

developed land As defined by the U.S. Department of Agriculture, developed land is any land no longer a part of the rural land base of the United States. This includes large tracts of urban and built-up land, tracts of built-up land of less than 10 acres, and land outside of built-up areas that is part of a transportation system (i.e., roads, rail, and associated rights-of-way).

metropolitan statistical areas As defined by the U.S. Office of Management and Budget and used by the U.S. Census, this is the delineation of urban areas that includes the central city and any surrounding counties that are socially and economically connected to an urban center.

smart growth Alternative forms of development that use land efficiently, mix land uses (such as housing, business, and shopping), and encourage walking and public transit in addition to driving while protecting surrounding open space.

sprawl Low-density, automobile-dependent development with a strict separation of land uses.

vehicle miles traveled Surface transportation miles traveled annually by passenger and freight vehicles in the United States.

Shifting from the 19th to the 20th century, new forms of transportation reshaped U.S. urban areas. Streetcars and subway lines transformed cities such as Boston, New York, and Chicago, allowing for dense concentrations of businesses and hotels. These areas were surrounded by neighborhoods often bound by ethnicity and class.

Idealists planned expansions of the city as a whole through movements such as the City Beautiful movement and the Garden Cities movement as well as through model new suburbs such as Radburn, New Jersey, Mariemont, Ohio, and Greenbelt, Maryland. These designs were precursors for those built throughout the country to fulfill the American Dream of owning a single-family detached home and an automobile. The dream became reality for millions of Americans in the years after World War II, and construction accelerated due to incentives such as the home mortgage deduction and infrastructure investments.

Most notable in terms of its impacts on our energy consumption have been investments in highways needed to serve the form of transportation that would eclipse transit—the automobile. Additionally, the large structures connected by ribbons of asphalt in suburbia have consumed increasingly more energy. This article discusses suburbia's ascendance in the United States, the transportation and construction impacts of this trend, as well as some alternatives to sprawl being debated including "smart growth." Smart growth entails new policies and incentives to create a greater variety in the supply of modes of transportation coupled with more and better travel demand management to conserve energy and improve environmental quality.

1. SUBURBAN SPRAWL

Much of current land development takes a form called suburban sprawl. Sprawl can be roughly defined as low-density development with a strict separation of land uses connected mostly (or only) by roads and highways. Although some sprawl occurred in the form of streetcar suburbs in the first couple of decades of the 20th century, sprawl accelerated after World War II as returning G.I.s, their families, and

 © 2004 Elsevier Inc. All rights reserved.

TABLE I

Census Population Data for the 12 Largest Cities in the United States, 1950–1970[a]

	1950		1960		1970	
City	Population	Rank	Population	Rank	Population	Rank
New York	7,891,957	1	7,781,984	1	7,894,862	1
Chicago	3,620,962	2	3,550,404	2	3,366,957	2
Philadelphia	2,071,605	3	2,002,512	4	1,948,609	4
Los Angeles	1,970,358	4	2,479,015	3	2,816,061	3
Detroit	1,849,568	5	1,670,144	5	1,511,482	5
Baltimore	949,708	6	939,024	6	905,759	7
Cleveland	914,808	7	876,050	8	750,903	10
Saint Louis	856,796	8	750,026	10	622,236	18
Washington, DC	802,178	9	763,956	9	756,510	9
Boston	801,444	10	697,197	13	641,071	16
San Francisco	775,357	11	740,316	12	715,674	13
Pittsburgh	676,806	12	604,332	16	520,117	24

[a] Data from Teaford (1993).

other citizens left cities for the suburbs in increasing numbers. This led to declines in the populations of central cities during the latter half of the 20th century (Table I).

A study by Janet Rothenberg Pack for the Brookings Institution examined population trends in 277 metropolitan statistical areas and found that flight from existing urban areas occurred throughout the United States from 1960 to 1990. The findings are summarized in Table II. In addition to the city-to-suburb migration evidenced by Table II, it is also clear that there have been interregional population distribution changes. Southern and western regions experienced the most population growth during these three decades. The population of suburban areas of these regions more than doubled from 1960 to 1990 (Table II).

Evidence shows that these population trends are continuing and that land development is occurring at twice the rate of population growth. Some of the best U.S. land-use data collection is done by the U.S. Department of Agriculture under the auspices of the Natural Resources Inventory (NRI). The NRI defines developed land as land that is no longer a part of the rural land base of the United States. This includes large tracts of urban and built-up land, tracts of built-up land of less than 10 acres, and land outside of built–up areas that is part of a transportation system (i.e., roads, rail, and associated rights-of-way). This land specifically consists of cropland, pasture land, range land, forest land, land enrolled in the federal Conservation Reserve Program, and other rural land. According to the NRI, U.S. land development has accelerated from 1.4 million acres per year in the 1980s (1982–1992) to 2.1 million per year in the 1990s (1992–1997).

TABLE II

Population Data (in Thousands) for U.S. Metropolitan Statistical Areas (MSAs), Cities and Suburbs, Divided by Region, 1960–1990

	Total population				% change,
Area	1960	1970	1980	1990	1960–1990
All MSAs	127,575	149,125	164,042	182,512	43
Cities	56,529	60,568	59,279	62,114	10
Suburbs	71,046	88,557	104,764	120,398	69
Northeast	39,173	43,229	42,752	44,116	13
Cities	16,142	15,705	13,896	13,892	−14
Suburbs	23,031	27,524	28,856	30,224	31
Midwest	36,264	41,004	42,094	43,149	19
Cities	17,227	17,516	15,787	15,284	−11
Suburbs	19,036	23,489	26,307	27,865	46
South	30,625	37,278	45,463	53,349	74
Cities	14,207	16,695	17,551	18,506	30
Suburbs	16,418	20,583	27,912	34,843	112
West	21,513	27,613	33,733	41,898	95
Cities	8,952	10,653	12,045	14,431	61
Suburbs	12,561	16,961	21,689	27,467	119

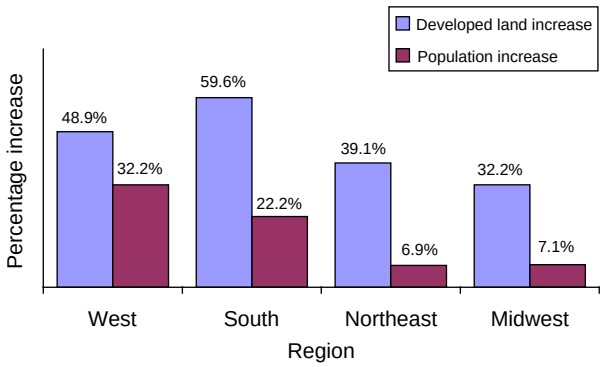


FIGURE 1 Population vs urbanized land growth by region, 1982–1997, as summarized in "Who Sprawl's Most? How Growth Patterns Differ across the U.S.," prepared in July 2001 by W. Fulton *et al.* for the Brookings Institution Center for Urban and Metropolitan Policy.

There are important regional differences in the relationship between these two factors in the sprawl equation. In the past couple of decades, the west has been growing but not sprawling as much as it has in the past (due in part to limits of developable land, whether due to topography or federal ownership), the northeast and the midwest have been sprawling but not growing much, and the south continues to grow and sprawl rapidly (Fig. 1).

2. TRANSPORTATION IN SUBURBIA

According to the U.S. Department of Transportation (DOT), as of 1995, every U.S. driver spent an average of 443 h every year behind the steering wheel, or 43 h more than in 1990. Although commuting to and from work only accounts for one-fifth of the miles that Americans drive daily, it is also notable that peak commuting hours have been increasing. In 1982, the average daily congested period was 2 or 3 h, but by 1999 this had doubled to 5 or 6 h.

There are economic impacts as well. Americans now spend more on transportation than food, clothing, or health care. As of 1998, according to the U.S. Department of Labor, 17.9¢ from every household dollar is spent on transportation, 98% of which is spent on buying, operating, and maintaining cars. This household expenditure is second only to housing (19¢). Transportation is also sapping the strength of the American economy. The Texas Transportation Institute found that congestion cost 75 metropolitan areas $67.5 billion due to lost income and increased truck operating costs as well as excess fuel consumption (5.7 billion gallons) in 2000.

During the latter half of the 20th century, U.S. planners, engineers, and decision makers relied mostly on the expansion of road system capacity to alleviate traffic congestion. As quoted in "Divided Highways," a history of the U.S. interstate highway system, in June 1969 Transportation Secretary John Volpe commented on this problematic modus operandi:

> *The federal government spends as much money on highway construction in 6 weeks as it has put into urban transit in the last 6 years. … Unless we intend to pave the entire surface of the country—and no one wants that—we have to stop this trend. We already have 1 mile of highway for every square mile of land area in the USA.*

According to the DOT, 30 years later the United States had 2 miles of highway for every square mile of territory. This country has made massive investments in building a 4-million-mile highway and road network since World War II. According to a 1997 National Research Council report, U.S. highways, streets, and associated rights-of-way totaled approximately 20 million acres, or 1% of U.S. land area. Based on this figure, the Harvard School of Design estimated roads impact one-fifth of the U.S. land base in the forms of fringe noise, air, and water pollution effects from vehicle traffic.

The environmental impacts of our growing road system are significant. Habitat loss and fragmentation reduce wildlife population diversity and abundance, contributing to species imperilment. In fact, the National Wildlife Federation reported that sprawl is the main cause of species imperilment in California. Wetlands loss can also be traced to our expanding transportation system: in the Environmental Protection Agency's (EPA) "National Water Quality Inventory: 1996 Report to Congress," 6 states reported significant wetlands loss due to highway construction and 10 reported significant loss due to residential development. This report also found that 38% of assessed U.S. estuary miles are impaired and that 46% of this impairment is due to polluted urban runoff. This is not surprising. Runoff from impervious surfaces such as roads is considerable: Runoff from a 1-acre parking lot is 16 times that of an undeveloped meadow.

Americans have made good use of increases in pavement. U.S. vehicle miles traveled (VMT) increased steadily throughout the second half of the 20th century. According to the DOT, from 1970 to 1990 VMT doubled from 1 to 2 trillion per year, and as of 1998 Americans traveled 2.6 trillion miles annually.

Driving increasingly more miles has meant turning away from other modes of transportation, such as mass transit. Using data originally from the DOT and the American Public Transportation Association and subsequently provided by the Surface Transportation Policy Project, Fig. 2 highlights this relationship for most of the 20th century. Transit boardings hit their peak in 1946 at approximately 23.5 million, whereas annual VMT decreased during World War II (1943–1945) and then continued its climb. This brief reversal of centurywide trends can be attributed to wartime policies such as the rationing of gasoline. In the 1920s, the larger trend lines had already taken shape. At century's end, however, VMT was barely growing, whereas transit stabilized and grew steadily from 1996 to 2001. The latest data available from 2002 show a decline in transit use of 1.3% and an increase in VMT of 1.7%. One overall trend is clear: Americans are driving increasingly more every year (Fig. 2).

2.1 Transportation Impacts on Energy and Air Quality

All of this driving has a major impact on U.S. petroleum consumption. According to the DOT, the transportation sector consumes more than 65% of the petroleum used in the United States and highway vehicles account for approximately 84% of this consumption. Basically, Americans are consuming increasingly more oil just to get around, despite improvements in efficiency. Stagnating fuel economy of the U.S vehicle fleet due to the popularity of sport utility vehicles and light trucks doubtless contributes to this trend. According to the EPA, U.S. fuel economy peaked in 1987–1988 and is currently 6% lower than in 1988. Using U.S. Department of Energy (DOE) data, Fig. 3 shows that oil consumption by the U.S. transportation sector closely follows VMT growth—steadily upward toward the 9 million barrels per day mark.

Unfortunately, combustion of all this oil also pumps many unhealthy pollutants into the air. In addition to carbon monoxide (CO), cars and trucks are also the source of volatile organic compounds (VOCs) and oxides of nitrogen (NO_x). When VOCs and NO_x mix while exposed to the hot sun, a chemical reaction takes place producing ozone. In recent years, ground-level ozone has been particularly problematic for large, sprawling U.S. cities such as Atlanta and Houston. Ozone is an irritant to the lungs and eyes, causing coughing and chest pains. It aggravates asthma and can have especially harmful effects on children, who breathe more air per pound of body weight than adults.

Transportation is responsible for many of these emissions, as Benfield *et al.* found in their 1999 book, "Once There Were Greenfields":

> *In particular, cars and other highway vehicles continue to emit some 60 million tons of carbon monoxide per year, about 62% of our national inventory of that pollutant; cars and other highway vehicles continue to emit some 7 million tons per year, almost 26%, of our volatile organic compounds … and they emit around 8 million tons per year, about 32%, of our nitrogen oxides.*

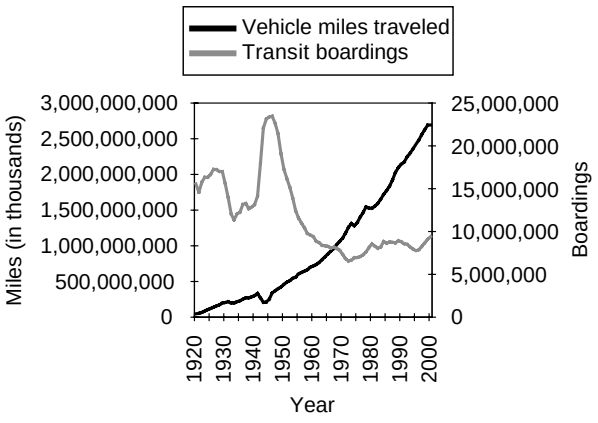


FIGURE 2 Change in transit boardings and vehicle miles traveled between 1920 and 2000. Data originally from the U.S. Department of Transportation and the American Public Transit Association as provided by the Surface Transportation Policy Project.

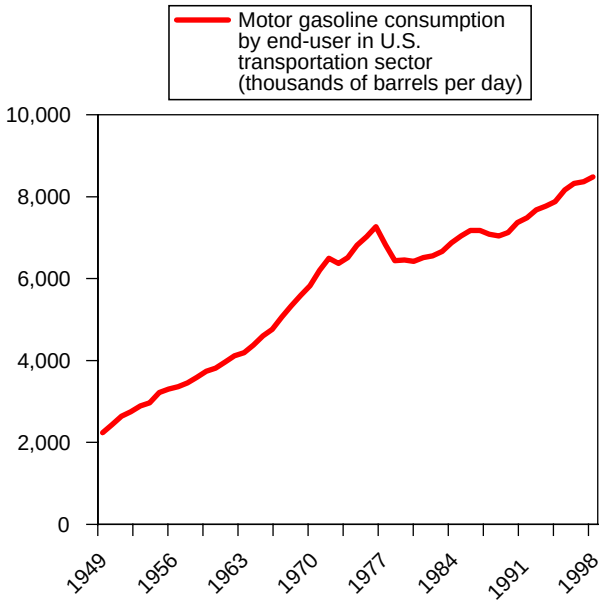


FIGURE 3 Increase in oil consumption by U.S. transportation sector, 1949–2001. Data from the U.S. Department of Energy.

The transportation sector is also a growing part of the total anthropogenic greenhouse gas emissions, especially carbon dioxide. According to the EPA, transportation sources accounted for approximately 31% of U.S. carbon emissions (460.4 million metric tons of carbon) in 1997. This sector is growing rapidly: Total emissions increased 10% between 1990 and 1997. According to the DOE, more than half of this increase can be traced to cars and trucks: As of 1999, 60% of transportation-related emissions were due to the combustion of gasoline.

3. STRUCTURES IN SUBURBIA

Although the impact of a transportation sector that underwent near-complete transformation in the 20th century has been examined most closely by experts and scholars, the low-density patterns of development have also meant an increase in the average size of houses, schools, and commercial buildings. This too has energy impacts.

3.1 Residential Buildings

According to the DOE, housing is the second largest source of carbon emissions in the United States. In 1999, residential buildings accounted for 19% of total U.S. CO_2 emissions at 290.3 million tons of carbon equivalent. According to the DOE, electricity accounted for most (63%) household carbon emissions as of 1997. Not surprisingly, the large-lot, detached-housing-only residential units that make up an increasing part of the U.S. landscape consume a great deal of electricity (as well as other types of energy) compared to more modestly sized housing types. In fact, according to the DOE, as of 1997 single-family detached housing made up by far the largest percentage of residential sector energy consumption at 73.4%.

Table III summarizes consumption levels for different U.S. housing types in 1997.

A good measurement for the increase in housing size in the United States is the number of rooms in a unit. The DOE's 1997 Residential Energy Consumption Survey showed that the size of units has grown steadily larger, resulting in greater energy use due to increased requirements for heating, air-conditioning, as well as lighting. Figure 4 shows this trend from 1978 to 1997.

Another yardstick for comparing the environmental impacts of different housing types is the ecological footprint of a given residence. Ecological footprint analysis is a tool designed by William Rees of the University of British Columbia and Mathis Wackernagel of the nonprofit Redefining Progress in the 1990s. This framework basically measures the depletion of "natural capital" by calculating how much world-average bioproductive land is needed to sustain a given population. The tool is versatile and has been used by households, schools, cities, and even nations.

This method of analyzing the impact of human activity sometimes uses the loss of "sinks" (soil,

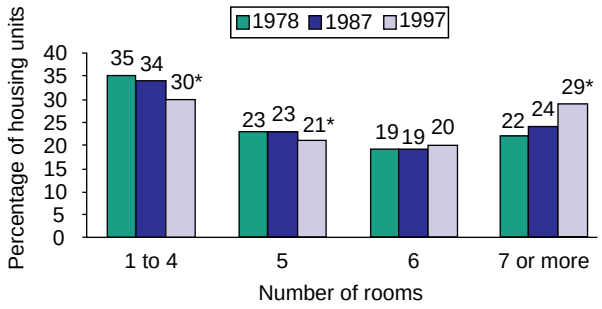


FIGURE 4 Increases in the number of rooms per household, 1978–1997. *The difference between the 1978 and 1997 estimates is statistically significant at the 95% confidence level. Data from the Energy Information Administration, Residential Energy Consumption Survey, 1978, 1987, and 1997.

TABLE III

Energy Consumption by Household Type in the United States, 1997[a]

	Type of housing unit				
	All households	Single-family, detached	Single-family, attached	Multifamily, 2–4 units	Multifamily, 5 + units
Electricity consumption per household (kWh)	10,219	11,778	8071	6505	5990

[a] Data from the Energy Information Administration.

vegetation, and water resources that absorb CO_2, removing it from the atmosphere) to account for fossil fuel consumption. It also includes the bioproductive-land-equivalent impacts of food, the built environment, and forest. Overall, it is a useful sketch tool for comparing impacts, and an analysis of housing footprints yields interesting results. Specifically, Table IV provides the results of a 1995 study of different Canadian housing types using this tool.

Suburban sprawl includes a large number of single-family detached houses. For every detached house built instead of a townhouse, ecological footprint per capita increases 17%. Also, for every detached house built instead of an apartment, a person's ecological footprint increases 40%.

As the awareness of the significant impacts of residential development on the environment has grown, a cottage "green building" industry has also grown. Several U.S. jurisdictions and groups have codes, departments, or certification programs for green building. Austin (Texas), Boulder (Colorado), San Francisco, Seattle, Dade County (Florida), Oregon, Pennsylvania, New Jersey, New York, Minnesota, Maryland, Colorado, and Texas all feature green building efforts of some kind. Among private sector groups, the National Association of Homebuilders, the American Institute of Architects, the American Lung Association, and the Energy Efficient Building Association have programs. The U.S. Green Building Council is also working with nongovernmental organizations such as the Natural Resources Defense Council (NRDC) to develop standards for residential development as part of its Leadership in Energy and Environmental Design (LEED) certification process.

This increased interest in green building is driven in part by increasing demand. For example, 89% of U.S. consumers surveyed in a 2000 *Professional Builder* magazine poll said they "were willing to pay more for green building features if they improved quality, durability, and the health of the home." More than half would pay an additional $2500–5000 and 9.5% would pay an additional $10,000 for a green home.

Considerable energy savings are possible through efficient design. In a 1998 study of the savings possible via cost-effective energy efficiency investments (in lighting, appliances, etc.) in Illinois, the American Council for an Energy Efficient Economy (ACEEE) found that 32 and 27.5% savings in electricity were achievable for existing and new single-family units, respectively. A similar 1997 ACEEE study found 37.5% (existing) and 26.5% (new) in possible savings in Pennsylvania and New York and a 51.5% (existing) and 30.0% (new) possible savings in New Jersey, all for single-family units. Therefore, possible electricity savings through energy efficiency improvements for the type of housing typically found in sprawling developments in these four states ranges from 26.5 to 51.5%.

However, few green building programs integrate criteria about neighborhood density and location efficiency, which are key components of smart growth. Changes are afoot, though. For example, Austin, Texas, has a Smart Growth Initiative complete with a matrix for calculating just how "smart" particular development proposals are, and LEED is developing new design standards that integrate "smart growth" criteria. Given the popularity of the approximately 100 New Urbanist developments built thus far, the addition of location efficiency as a criterion for determining how "green" a building is should ensure that such buildings continue to fetch a premium in the marketplace.

3.2 Educational Buildings

Buildings in the U.S. educational system comprise approximately 13% of total commercial building floorspace. According to the DOE, they are the largest at 25,100 square feet on average per building, compared to the average for all commercial buildings

TABLE IV

Footprints of Various Household Types in Acres[a]

	Category				
	Standard lot detached house	Small lot detached house	Townhouse	Walk-up apartment	High-rise apartment
Housing	1.95	1.95	1.61	1.16	1.16

[a] *Source*: Walker and Rees (1997). For more information, see http://www.ecologicalfootprint.org.

of 12,800 square feet. Newer schools use more energy per square foot than older schools, according to the Energy Information Administration. The National Trust for Historic Preservation finds that newer schools are larger and more apt to be located in sprawl locations than older ones. In 2000, the Department of Education studied the functional age of schools, or the number of years since a building had been constructed or renovated, and found that newer schools were larger:

> *On average, small schools had older functional ages than medium or large schools. Additionally, large schools were more likely than small schools to have a functional age of less than 5 years, and the small schools were more likely than large schools to have a functional age between 15 and 34 years.*

This is not surprising given requirements imposed on school construction by states and localities. For example, many states adhere to the guidelines created by the Council of Educational Facility Planners International (CEFPI) when planning new school construction:

- At least 10 acres of land plus 1 acre for every 1000 students for elementary schools
- At least 20 acres of land plus 1 acre for every 100 students for middle schools
- At least 30 acres of land plus 1 acre for every 100 students for high schools

Not surprisingly, buildings on this land are also larger than they have been in the past. In 1995, CEFPI released what it terms the "starting point" for planning the size of school buildings. Size of schools is partly determined by climatic conditions. For this reason, CEFPI lists different averages for currently existing school buildings in different regions. One region encompasses nearly the entire United States (48 states). For this region, the recommended starting point average gross square footages per student are 111.5 for elementary schools, 154.4 for middle schools, and 160.7 for high schools.

3.3 Commercial Buildings

According to the DOE, in 1999 commercial buildings consumed 16% of U.S. primary energy, and 75% of this consumption was in the form of electricity. As of 1995, the greatest energy consumers were buildings that provided or hosted food sales or services, health care-related services, lodging, public assembly, public order and safety services (i.e., police stations), and office spaces. According to the Alliance to Save Energy, carbon dioxide emissions from the commercial buildings sector have increased 30% since 1980, faster than any other sector.

The DOE sums up the scale of our stock of commercial buildings as follows:

> *In 1995, there were 4.6 (with a ±0.4 margin of error) million commercial buildings in the United States comprising 58.8 (±3.9) billion square feet of floorspace. That amount of commercial floorspace exceeds the total area of the state of Delaware and amounts to more than 200 square feet for every resident in the United States.*

In spite of the tremendous amount of land development in recent decades, this stock is surprisingly comprised mostly of older buildings. The median building age is 30.5 years, and more than half of the buildings were built before 1970. According to the DOE, site energy intensity has not decreased significantly since the beginning of the 20th century. A lack of progress in the 1990s is confirmed by a November 2000 assessment of energy consumption by the EPA, which states, "Total energy consumption in commercial buildings remained roughly constant from 1989 to 1995, as did energy use per square foot."

Highlighting the potential to conserve energy, the EPA cites studies showing that "by investing in simple profitable building upgrades, many U.S. homes and businesses could reduce their annual energy use by nearly 30%." Through both private and public sector programs, such savings have been achieved in some cases. The EPA's Energy Star program, for example, claims to have partnered with 15% of the "commercial, public, and industrial building market" and rewarded 90 energy-saving buildings with the Energy Star label. These buildings prevent 45 million metric tons of carbon equivalent worth of emissions. Also, the Rocky Mountain Institute has performed many studies that show potential energy savings for commercial buildings: In one analysis, they found potential savings of 75% for a vacant 13-story, 200,200-square foot office building outside Chicago.

4. LAND USE, TRANSPORTATION, AND ENERGY CONSUMPTION IN OTHER COUNTRIES

Urban settlement patterns in other countries have remained fairly compact. Combined with adequate transportation infrastructure, this leads to lower dependence on the automobile in countries such as those in Europe. In fact, European land-use and

transportation policies and practices are often evoked when trying to define policy alternatives for achieving more effective growth management in the United States.

For example, Pietro Nivola of the Brookings Institution finds that U.S. cities interested in improving their condition could adopt some policies used frequently in Europe, such as reducing public funding for highway construction, reducing urban crime, and providing fiscal relief for city businesses and residents through revenue-sharing arrangements such as those used by Germany and Canada. Timothy Beatley of the University of Virginia has researched best practices in European countries, including Copenhagen's 1947 land-use "finger plan" that concentrates development into a hand shape, with the city center as the palm and transit-line-centered digits extending outward with protected green spaces between the lines. Due in part to the choices offered by these innovative land-use and transportation policies, 31% of Copenhagen commuters ride transit to work and 34% bike to work. This compares poorly to U.S. cities, where transit and bike home-to-work trips make up merely 5% and 2% of the total, respectively.

European cities are faring better in terms of fuel use. In 1999, Europe was second only to the United States (at 25%), with 15% of total global carbon emissions from fossil fuels. However, according the DOE, European emissions per capita were less than half those of the United States (2.4 vs 5.6 tons).

A study of 48 global cities by Peter Newman and Jeffrey Kenworthy shows just how energy-intensive U.S. transportation patterns have become. The averages for gasoline used for private transportation are summarized in Fig. 5.

They found generally that gasoline-oriented cities are "heavy energy users while those with any significant level of electricity use [to run a rail system] in their transportation system are low energy users overall." Even after correcting for economic factors, such as U.S. income, gasoline prices, and fuel efficiencies, they found that gasoline use was still significantly higher (47–63%) in the United States than in other world cities, on average. In fact, the authors noted that "a purely economic approach to transportation matters will be inadequate, and that matters of infrastructure and urban form have direct influence on transportation patterns." One advantage to investing in more transportation options in order to reduce trips and trip lengths is a phenomenon termed transit leverage. Newman and Kenworthy cite Neff, who found that transit does not simply replace car travel on a one-to-one ratio. In fact, Neff found that 8.6–12.0 km (5.3–7.4 miles) of car travel is replaced by 1 km (0.62 miles) of transit according to data from the United States.

Interestingly, Newman and Kenworthy also found that there is little connection between transportation and economic growth. Whereas in U.S. cities gross regional products per capita were only 0.85 those of Europe, car use was 2.41 times higher in the United States.

Nonetheless, the suburban sprawl model is being adopted by Second and Third World countries as a means to prosper. Private car use in Europe and China is increasing. As Molly O'Meara Sheehan of the Worldwatch Institute wrote,

> *In the Czech Republic, car use surged and public transit use fell as the number of suburban hypermarkets ballooned from 1 to 53 between 1997 and 2000. Today, the proliferation of such developments around Prague has boosted the number of hypermarkets in the Czech Republic above that in The Netherlands.*

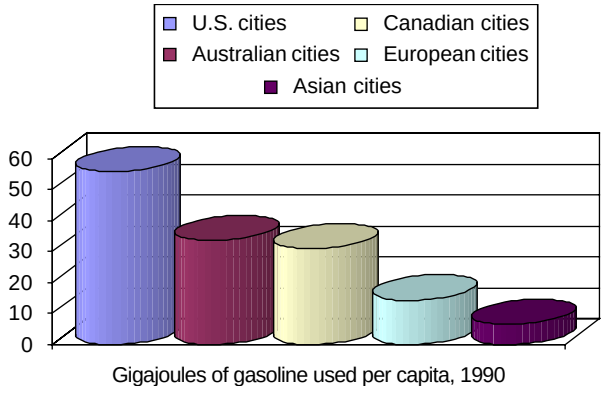


FIGURE 5 Gasoline use per capita in the United States, Canada, Europe, Australia, and Asia.

5. TRANSPORTATION ALTERNATIVES FOR U.S. SUBURBS

As discussed previously, since the mid-20th century popularization of the automobile, the usual response to demand for faster commutes and trips has been to provide a more ample supply of new highway and road capacity. However, in recent years various studies have described a phenomenon that puts in doubt the effectiveness of this tactic for reducing congestion. The phenomenon is not surprising: Greater availability of pavement spurs more and longer trips by road users. What is sobering is the extent of the phenomenon.

The influence of new road and lane construction on travel behavior is referred to as induced travel, which has been studied in all 50 states and 70 regions. The idea is simple: New road capacity acts as a magnet for drivers, inducing them to take more and longer trips than they would have otherwise. According to the growing academic literature as summarized in 2000 in a presentation by Lewison Lee Lem of the EPA, the impact can be substantial: An average increase of 10–30% of total VMT can be traced to induced travel. Short-term induced increases range from 2 to 5% for every 10% increase in road capacity. Long-term increases may result in increasing travel 5–10% for every 10% increase in capacity.

Some of this additional traffic is also a product of new land development and new users who occupy new subdivisions and shop in new malls. Transportation investments often have an impact on land development. A study performed for the Brookings Institution by Marlon Boarnett confirmed the relationship between transportation policy and land use, concluding that highways can induce changes in regional development patterns. In fact, this study found that highway investments shift economic development away from existing communities to newly built areas. Just as empty roads and traffic lanes act as magnets for new driving, highway interchanges draw new development.

Smart-growth policies and practices offer one possible way to deal with out-of-control VMT and pavement growth. These policies are being embraced by localities and states throughout the country. It is time for such a shift. Many surveys in the past few years have shown that Americans are tired of suburban sprawl and its symptoms—loss of open space, traffic congestion, and blight in older communities. For example, in a 1999 poll by the Pew Center for Civic Journalism, sprawl and its effects tied with crime as the issue of most importance at the community level.

What is smart growth? Roughly, smart growth is development that involves a mix of land uses (e.g., commercial and residential), which occurs in or adjacent to existing communities and which provides a variety of transportation choices for residents and workers. Orienting and designing new development to make it convenient for residents and workers to take transit is especially important, and there is even a name for this kind of development—transit-oriented development.

Many advocates propose taking advantage of significant infill development and redevelopment opportunities created by population declines in central cities and older suburbs and postwar leapfrog development patterns. Retrofitting U.S. suburbs so that we take advantage of vacant and abandoned lots means more compactness and convenient access to shops and potential workplaces. After decades of sprawl, there are plenty of opportunities for such retrofits. One example that is receiving much attention is "greyfields." These are the older regional malls that litter many older suburbs. Unfortunately, "litter" is an all-too-apt term since many of them are vacant, having been left behind by companies seeking to build ever-larger stores and malls. According to a study by PricewaterhouseCoopers and the Congress for New Urbanism, 140 of approximately 2000 regional malls are greyfields, and 200–250 are on their way to that status. These sites are perfect for redevelopment given their large size—35 acres on average—plus the fact that their location in older suburbs often means proximity to transit.

Redevelopment activities such as this will mean more jobs and more residents per acre, which many studies have found correlate well with decreases in VMT. A 1992 study by Cambridge Systematics found 20–25% fewer car trips per household in more compact, transit-oriented suburbs than in sprawling, auto-dependent suburbs. In a 1994 NRDC-sponsored study of several San Francisco communities, John Holtzclaw found that a doubling of residential density is accompanied by a 20–40% reduction in VMT. Also, Reid Ewing concluded in a 1997 study that when controlling for effects of variables such as income, a doubling of urban densities results in a 25–30% reduction in VMT.

It is important to clarify that when researchers say "doubling," they are usually talking about modest regional or neighborhood averages for density or compactness. Some Americans might think that the only alternative to sprawl is New York City-level density. However, this is not the case. There are a number of alternatives to suburban sprawl, including some models within suburbia. For example, Ed Risse of Synergy Planning estimated in 2001 that if the Washington, DC region were to adopt the suburb of Reston's density of 9 or 10 people per acre, the region would be able to accommodate 25 years of demand for urban development on vacant and underutilized land within a 20-mile radius of Washington's center. Reston is only compact relative to other segments of suburbia: It remains a quiet, leafy community due in part to the 1300 acres of green spaces it has permanently set aside.

Some studies of the impacts of changes in the built environment on travel demand have yielded various results. In a 2001 synthesis of the literature, Reid Ewing and Robert Cervero concluded that travel

TABLE V

Summary of Environmental Protection Agency Findings Predicting Daily Vehicle Miles Traveled (VMT) and CO_2 Reductions in Infill/Redevelopment Building Scenarios

Location	Daily VMT (%).	CO_2 reduction (%)
San Diego, CA	48 (per capita)	45
Montgomery County, MD	58 (per capita)	46
West Palm Beach, FL	61 (per capita)	50
Atlanta, GA	14.5–52	Not calculated

demand is pretty inelastic in the face of individual changes in the built environment, but that there are large cumulative effects, a combination of smart-growth characteristics, such as increases in density, mixing land uses, design changes, and wider regional accessibility of different land uses.

New, sophisticated modeling techniques have demonstrated the effectiveness of such strategies. Several EPA studies utilize four-step travel demand models to show that major reductions in VMT are achievable. These models use information about travel patterns to calculate how many trips will be generated, where trips will be distributed, and how the trips will be split between different transportation modes (i.e., walking, biking, train, and bus) and also to assign the traffic among different routes and times of day. In two studies, the EPA worked with local jurisdictions to choose likely sites for infill/redevelopment and greenfield development and used transportation models to forecast differences in miles traveled. Table V summarizes the findings.

Examining development patterns in the San Francisco, Los Angeles, and Chicago areas, NRDC researchers found strong correlations between density and urban form and VMT. In areas with smart-growth characteristics, such as small lot sizes, transit services, and walkable neighborhoods, residents find it less necessary to drive. This powerful relationship is expressed in Fig. 6.

Also, in EPA-commissioned studies of travel behavior in the Sacramento and Nashville regions, NRDC and Criterion Consultants found that VMT per capita in neighborhoods was lower in urban and older suburban neighborhoods. In Nashville, VMT per capita was 30% lower among suburbanites in an older, more compact suburb as opposed to a newer, more sprawling suburb. In Sacramento, the rate of fuel consumption was as much as 50% lower in an urban community compared to two suburban neighborhoods.

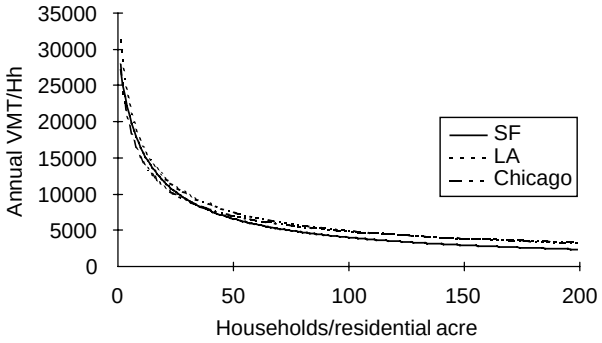


FIGURE 6 Relationship between VMT per household and residential density in the San Francisco (SF), Los Angeles (LA), and Chicago metropolitan areas.

It should be noted that there is great potential for synergy among several nascent market transformations, including smart growth. The combination of green building techniques, cleaner vehicle technologies such as hybrid engines and fuel cells, and changes in development of new suburbs and retrofits of existing communities should yield significant benefits in terms of energy conservation, climate change, and environmental quality.

5.1 The Return of Rail?

When we selected cars and trucks as our vehicles of choice, we shunned another kind of infrastructure—rail, especially intracity rail. Urban streets and even highways took the place of trolley systems that spread like webs through our cities. We had built approximately 45,000 miles of trolley lines by 1915. According to the American Public Transportation Association, we currently have a paltry 3100 miles of light and heavy rail serving our cities. As discussed previously, ridership on public transportation peaked at 23.4 billion trips in 1946 and declined steadily throughout the decades, bottoming out in 1972 at 6.5 billion.

However, there have been considerable increases in recent years and for 5 years in a row in the late 1990s the rate of transit boarding growth actually exceeded the growth rate of driving. The trend during the past 10 years is shown in Fig. 7.

In response to increasing demand, new rail starts have increased in recent years, including some starts in unexpected southern and western cities. As columnist Neal Peirce noted, applications for federal funding for new bus and rail projects leaped from 20 in 1997 to 191 in 1998. The federal program for new transit starts remains popular and hence oversubscribed. Also, the American Public Transportation Association reported that as of 2000, there

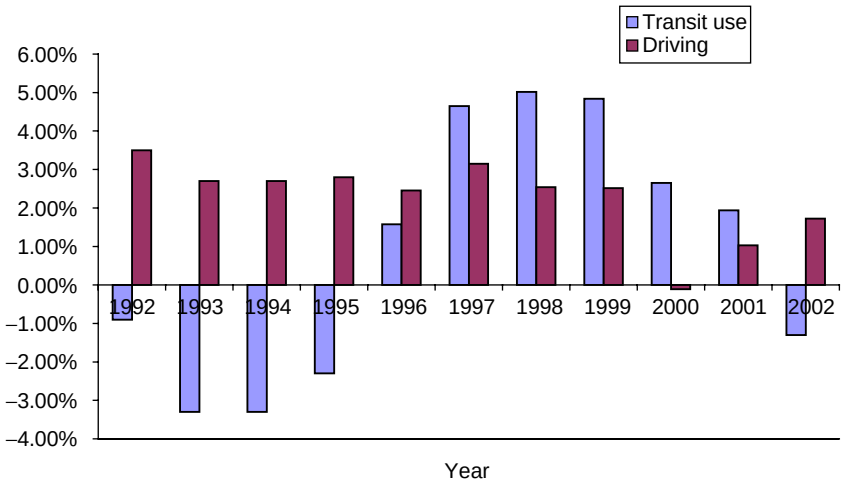


FIGURE 7 Percentage change in the growth of transit use and vehicle miles driven in the United States, 1992–2002. Data originally from the U.S. Department of Transportation and the American Public Transit Association as provided by the Surface Transportation Policy Project.

were approximately 1500 miles of new light rail and heavy rail lines in the proposal, planning, design, or construction phase.

This boom in new rail construction is a nationwide phenomenon. The following are examples from Railway Age's 2001 "Passenger Rail Planner's Guide":

- Charlotte, North Carolina, has completed a major investment study for an 11.5-mile light rail line that will share 2 miles of track with the town's trolley.
- Atlanta's Metropolitan Area Rapid Transit Authority (MARTA) is planning to extend its rail system with a 4.2-mile line, and work has begun on a 47-acre transit-oriented development on MARTA-owned land that will include BellSouth as a tenant.
- The Southwest Ohio Regional Transit Authority and the Transit Authority of Northern Kentucky are studying an 18-mile light rail segment that would connect Covington, Kentucky, and central Cincinnati.
- A 17.4-mile light rail extension in St. Louis was slated to open this year, and an additional 8.8-mile segment is planned.
- The Dallas Area Rapid Transit Authority is constructing two new lines for its successful system, which will more than double the system's current 20 miles.
- The Denver Regional Transportation District recently opened an 8.7-mile extension to its system and is planning another 19-mile line.
- Salt Lake City recently built a 15-mile light rail line that is wildly popular; several extensions are planned or under way.

Many jurisdictions devoting resources to building new transit are outside of America's Rust Belt and in the Sun Belt—the American south and west. Increasing rail investments can help to catalyze transit-oriented smart growth since concentrating development around rail nodes helps ensure their fiscal viability. In fact, the Federal Transit Administration includes land-use criteria when awarding Full Funding Grant Agreements for new transit service. New transit investments throughout the country provide evidence that smart-growth strategies are widely applicable—both for revitalizing northeastern and midwestern cities and older suburbs that have felt sprawl's impacts most keenly and for the rapidly growing southern and western United States.

In summary, the possible comeback of rail transportation in the United States is good news indeed for those concerned about 20th-century changes in development patterns that have transformed suburbanization, transportation, and energy consumption in the United States.

Acknowledgments

The author acknowledges Nancy Jakowitsch and Hannah Stutzman for their contributions to this article.

SEE ALSO THE FOLLOWING ARTICLES

Alternative Transportation Fuels: Contemporary Case Studies • City Planning and Energy Use •

Ecological Footprints and Energy • Heat Islands and Energy • Land Requirements of Energy Systems • Leisure, Energy Costs of • Lifestyles and Energy • Motor Vehicle Use, Social Costs of • Population Growth and Energy • Urbanization and Energy

Further Reading

Beatley, T. (2000). "Green Urbanism." Island Press, Washington, DC.

Benfield, F. K., Raimi, M., and Chen, D. (1999). "Once There were Greenfields." Natural Resources Defense Council and Surface Transportation Policy Project, New York.

Ewing, R., and Cervero, R. (2001). Travel and the built environment—A synthesis. *Transportation Research Record*.

Forman, R. T. T., Sperling, D., *et al.* (2003). "Road Ecology, Science and Solutions." Island Press, Washington, DC.

Gillham, O. (2002). "The Limitless City: A Primer on the Urban Sprawl Debate." Island Press, Washington, DC.

Lewis, T. (1997). "Divided Highways." Penguin, New York.

Newman, P., and Kenworthy, J. (1999). "Sustainability and Cities: Overcoming Automobile Independence." Island Press, Washington, DC.

Nivola, P. (1999). "The Laws of the Landscape: How Policies Shape Cities in Europe and America." Brookings Institution Press, Washington, DC.

O'Meara Sheehan, M. (2001). "City limits: Putting the brakes on sprawl, Worldwatch Paper 156." Worldwatch Institute, Washington, DC.

Pack, J. R. (2002). "Growth and Convergence in Metropolitan America." : Brookings Institution Press, Washington, DC.

Sierra Club (2000). "Sprawl Costs Us All: How Your Taxes Fuel Suburban Sprawl." Sierra Club, San Francisco.

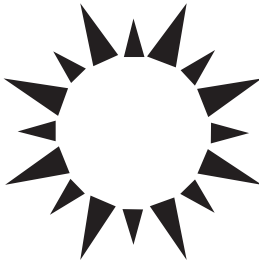
Teaford, J. C. (1993). "The Twentieth-Century American City." 2nd ed. Johns Hopkins Univ. Press, Baltimore.

U.S. Environmental Protection Agency (2001). "Our built and natural environments: A technical review of the interactions between land use, transportation, and environmental quality," EPA 231-R-01-002. Development, Community, and Environment Division. U.S. Environmental Protection Agency, Washington, DC.

Walker, L. A., and Rees, W. E. (1997). "Ecocity Dimensions: Healthy Communities, Healthy Planet" (M. Roseland, Ed.). New Society, Gabriola Island, British Columbia, Canada.

Sun, Energy from

WILLIAM LIVINGSTON
National Solar Observatory
Tucson, Arizona, United States



1. **Solar Interior and Nuclear Fusion**
2. **Magnetism and the Activity Cycle**
3. **Total Solar Irradiance**
4. **Solar Wind and Particle Flux**
5. **Orbital Modulation**
6. **Long-Term Variations in Solar Output**

Glossary

astronomical unit (**AU**) The mean distance of the earth from the sun, which is 1.495979 10^8 km.
chromosphere Layer immediately above the photosphere where many strong Fraunhofer lines are in emission at the limb of the disk.
corona The extended upper atmosphere of the sun, normally seen only at the time of a total solar eclipse or from satellites with an occulting device to cover the much brighter white light solar disk; its outer bounds reach beyond the earth as the tenuous heliosphere.
corona hole Region on the solar disk where surface magnetic fields open out to allow particle streaming and subsequent magnetic storms on the earth.
faculae Bright markings on the solar disk especially evident toward the limb; faculae give rise to positive blips in the irradiance time record.
granulation The fine mottling pattern on the white light photosphere of the solar disk.
insolation Energy input to the earth's surface that depends on the geographic latitude, atmospheric absorption, and the like.
irradiance The amount of electromagnetic power that flows across a surface (W/m^2).
limb darkening The dimming of the solar disk toward its edge; it is an indicator of a photospheric temperature gradient that diminishes with height.
photosphere The solar disk in white light; most Fraunhofer lines are formed in the upper photosphere.
radiance The power radiated from a unit area on a planar source into a unit solid angle ($W/steradian/m^2$).
sunspots Dark markings on the solar disk, usually consisting of an umbra and the surrounding less dark penumbra.
transition layer A dynamic, high-temperature layer between the chromosphere and the corona; characterized by emission from ionized lines such as Si IV, C IV, and N V, it clearly displays the structure of magnetic field lines.
ultraviolet (**UV**) Radiation shorter than visible wavelengths between 0.4 and 0.2 μm; extreme ultraviolet (EUV) is < 0.2 μm.

The sun is the primary source of energy on the earth. Most of this energy (99%) is in the form of electromagnetic radiation. Nuclear fusion in the sun's core is the origin of this energy. Space-borne radiometers have allowed the sun's radiation to be measured over the past quarter-century or so, and the result is constant to approximately 0.1%. This article considers these measurement techniques and their limitations. Ultraviolet emissions and particle flows, although relatively minute, are strongly modulated by the sun's 11-year magnetic cycle and produce noticeable effects on the earth's upper atmosphere. The long-term record of the sun's output and its magnetic cycle is indirectly revealed by historic sunspot records and from the geo-sited nuclides of ^{14}C and ^{10}Be. Other solar-type stars suggest other activity cycles that the sun might have exhibited in the past or might develop in the future.

1. SOLAR INTERIOR AND NUCLEAR FUSION

The chemical composition of the sun is 71% hydrogen, 27% helium, and 2% other elements. According to the Big Bang theory of the origin of the universe, only hydrogen and a few light elements were created. The presence of heavier elements, such as Fe, Ca, and Mg, indicates a stellar evolution that

 © 2004 Elsevier Inc. All rights reserved.

includes material processing in the distant past through a supernova explosion.

Physical conditions in the solar interior are such that perfect gas laws apply. Therefore, one can calculate that at the sun's center the density is $\rho = 150\,\text{g/cm}$ and the temperature is $15 \times 10^6\,\text{K}$. At this density and temperature, the proton–proton (p-p) reaction takes place:

$$^1\text{H} + ^1\text{H} \rightarrow ^2\text{D} + \text{e}(+) + \nu \text{ releasing } 1.44\,\text{MeV of energy } (\nu = \text{neutrino}) \quad (1)$$

$$^2\text{D} + ^1\text{H} \rightarrow ^3\text{He} + \gamma\ (5.49\text{MeV}) \quad (\gamma = \text{gamma ray radiation}) \quad (2)$$

$$^3\text{He} + ^3\text{He} \rightarrow ^4\text{He} + (2)^1\text{H}\ (12.86\text{MeV}) \quad (3)$$

Thus, H is burned to produce He, and this is the source of the sun's luminosity ($L = 3.845 \times 10^{26}$ W). The age of the sun, which is approximately the age of the earth, is estimated to be 4.5×10^9 years. In the core, 50% of H has now burned.

Fusion energy generation is confined to the inner third of the solar radius, or $0.3\,\text{R}_\odot$. Energy flows outward by radiation to $0.75\,\text{R}_\odot$, where a convective zone of transport sets in via "giant cells." At the surface, the final convection elements are the white light granulation. Bright granules of observed mean dimension 1.4 seconds of arc (range 0.25–3.50, arcsec) are buoyed to the surface and radiate into space, and the resulting cooled gas (−300 K) descends as dark intergranule lanes (Fig. 1).

It takes approximately 10 years for core radiation (gamma rays or radiation whose wavelength is $< 0.01\,\text{nm}$) to diffuse to the surface and degrade to a cooled gas (effective temperature 5777 K or visible radiation). Currently, core conditions can be sensed from the neutrino flux (in minutes), which verifies the forgoing model of the solar interior. Seismic waves, which pass through the sun in hours, are detected at the surface as small Doppler shifts of the Fraunhofer lines and likewise substantiate the solar model.

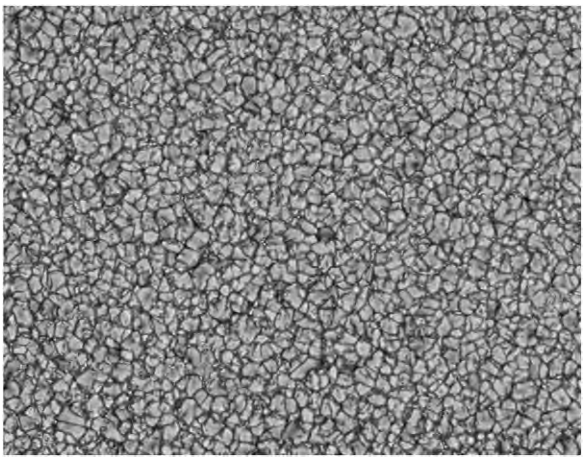


FIGURE 1 White light of solar granulation in a quiet (nonmagnetic) region. (Dutch Open Telescope, La Palma, Canary Islands, October 19, 2001)

2. MAGNETISM AND THE ACTIVITY CYCLE

The sun is a largely passive body. In fact, the sun possesses magnetic fields whose surface manifestation waxes and wanes at approximately 11-year intervals. This introduces the dynamic aspects of the sun. Unlike the earth, with its buried magnetic dipole, the sun's magnetism is observationally a surface phenomenon. Subsurface magnetism must exist, but we know little about it directly. The sun does not rotate as a solid body but rather exhibits differential rotation, with a surface angular rate at the equator greater than that at higher latitudes. Theory proposes that this differential rotation acts as a dynamo to intensify any diffuse magnetism in the sun's convective zone. Once magnetism is concentrated by this dynamo, the associated gas is less dense than nonmagnetic surroundings and is buoyed to the surface, eventually emerging as dark sunspots. Sunspots are dark because convection is inhibited inside them. (It should be noted that subsurface gas is nearly completely ionized, a plasma that is a perfect electrical conductor, and that magnetic fields are "frozen in" and virtually indestructible. A fraction of the sun's magnetism may actually be a relic from when the sun condensed out of the solar nebula.)

Sunspots typically appear in bipolar pairs having field strengths in their dark umbrae of 2000 to 3500 Gauss. It takes only a few days for a large sunspot group to form. Because sunspots are dark, the sun's total radiation diminishes proportional to their area up to 0.1%. Afterward, surface motions "nibble" away at sunspots and carry off the fields by a random walk fashion to create faculae. Faculae have fields of 500 to 1000 Gauss and are relatively bright. Thus, dark sunspots transform to bright faculae that outlive the spots themselves. In this way, over time, sunspots enhance solar luminosity by approximately 0.1%.

Figure 2 is a "magnetogram" of the sun near cycle maximum. On this picture, light areas represent positive (north) fields, and dark areas represent

negative (south) fields. The fragmented fields are cospatial with faculae. Most of the solidly light or dark areas are composed of unresolved faculae. Spots are not distinguishable here but lie within the shaded areas. Compare this figure with Fig. 3, which is the same type of picture at solar minimum. The fragmented fields in Fig. 3 are likewise faculae.

In 1843, Schwabe discovered that the number of spots on the solar disk is cyclic, with an average period of approximately 11 years. The sunspot cycle is synonymous with the magnetic or activity cycle. Besides sunspot number, other "solar indexes" include the number of sunspot groups, 10-cm radio flux, and the strength of certain spectrum lines. Fraunhofer lines, such as NaDl (596 nm) and CN (388 nm), report on the upper photosphere, CaK (393 nm) reports on the chromosphere, He (1083 nm) reports on the chromosphere and corona, and Lyman-α (121 nm) and other highly ionized lines, on the transition layer.

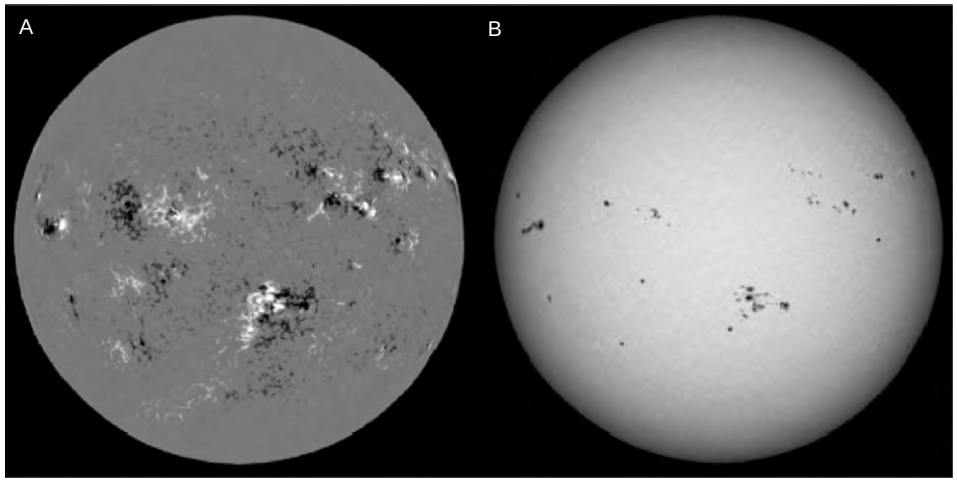


FIGURE 2 (A) A full-disk magnetogram taken near solar maximum on July 20, 2000. This is a computer-drawn picture showing polarization in the Fe (868.8-nm) line. When certain lines are formed in the presence of a magnetic field, Zeeman effect splitting occurs, with the amount being proportional to field strength, and the split components are polarized depending on magnetic polarity. A magnetograph measures these parameters. In the picture, white is north magnetic polarity, and black is south magnetic polarity. The east limb of the sun is on the left. The sun rotates east to west. The northern and southern activity belts are separated by the equator, which spans the middle. (B) A white light image for the same date.

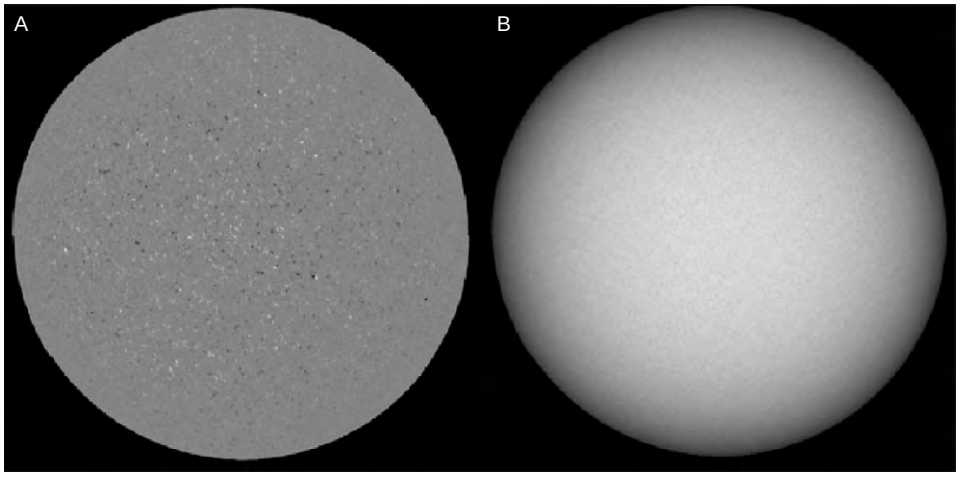


FIGURE 3 (A) A full-disk magnetogram at solar minimum on October 12, 1996. There are no sunspots or active regions at this time. Compared with Fig. 2, the sensitivity is increased to better show the weak magnetic features. These are either old facular remnants or short-lived ephemeral regions. Notice a weak grouping of white (north polarity) regions near the north heliographic pole (top) and of black (south polarity) regions at the south (bottom). These correspond to what is called polar faculae, a phenomenon dependent on the phase of cycle. (B) A white light solar image.

If spectroscopic observations are made in unimaged sunlight, in effect one is observing the sun as a star. Full-disk indexes include spectrum line strengths, or magnetic fields summed over the disk, or simply a number depending on the total count of spots on the disk. Full-disk solar indexes are found to track the activity cycle very well and can serve as surrogates for total solar irradiance (TSI). Particularly if one averages in time, solar indexes give nearly identical-looking signals. Figure 4 compares TSI, sunspot number, chromospheric He (1083 nm), chromospheric CaK, and the magnetic field obtained by averaging, without regard to polarity, the magnetic field over the whole disk. (If one pays attention to polarity, the + north-seeking fields largely cancel out the − south-polarity fields and the field average approaches zero.) Most of the differences seen in Fig. 4 arise from incompleteness of the various index time series owing to observing conditions (weather) or other factors. This means that long-term records, such as sunspot number, allow one to extrapolate recent TSI measurements, which began only in 1978, back to earlier epochs.

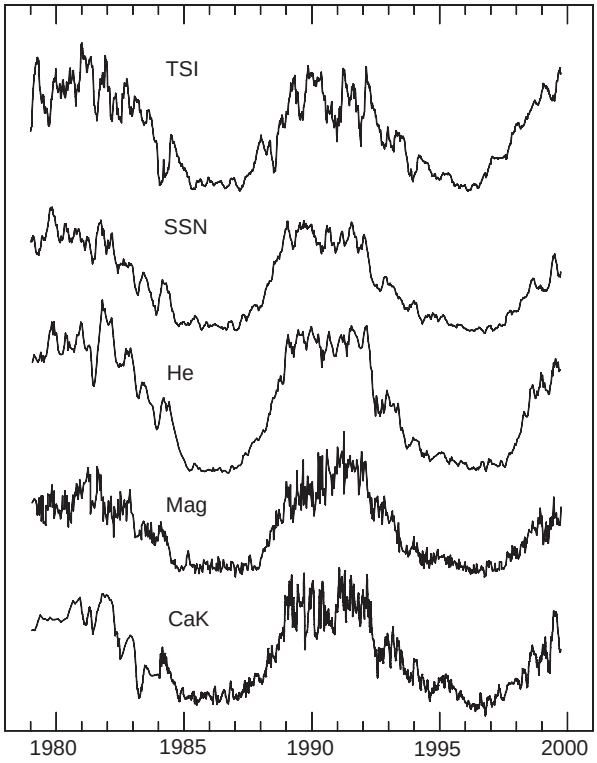


FIGURE 4 Full-disk surrogates for solar activity. TSI, total solar irradiance composite record according to Frohlich and Lean; SSN, international sunspot number; He, helium (1083-nm) line depth; Mag, magnetic field measured with the Vacuum Telescope on Kitt Peak; CaK, calcium K line central intensity measured at Kitt Peak and at Sacramento Peak. A running mean has been applied according to the completeness of the data.

Historic records of sunspot number, since the discovery of sunspots by Galileo, have shown intervals where there are markedly few sunspots. Named after their chroniclers, these epochs are called the Maunder (1645–1715) and the Dalton minima (1790–1820) (Fig. 5). Besides the 11-year Schwabe cycle, there is a weaker 90-year so-called Gleisberg cycle. Other cycles emerge from indirect indexes of radioisotopes, such as ^{10}Be and ^{14}C, in terrestrial reservoirs.

3. TOTAL SOLAR IRRADIANCE

Table I summarizes the energy output of the sun. By far, electromagnetic radiation (light) dominates. This is called TSI, which excludes a minute contribution at radio wavelengths and, due to variable transmission of the earth's atmosphere, must be measured with spacecraft radiometers. Presumably, the neutrino flux does not count because it is not absorbed on earth. The energy distribution by wavelength is given in Table II.

The modern radiometer is a precision device represented by the Active Cavity Radiometer Irradiance Monitor (ACRIM). ACRIM has been placed on-board a succession of satellites around the earth. It consists of a black body cavity into which sunlight is admitted through an exactly known aperture. When a cover is opened, the temperature rise over a time interval is compared with the comparable temperature rise due to electrical heating in an identical but closed comparison cavity. Once corrected for departure of the satellite from 1 AU, precise measurement of this electrical quantity of heat yields the TSI (in W/m^2). Both cavities are connected to heat sinks that periodically dump their accumulated heat.

Errors in TSI are determined by preflight calibration uncertainty and some deterioration of the black cavity due to outgassing and ultraviolet (UV) radiation damage in space. Two systems are always flown, with one being exposed only occasionally. Despite all precautions, it has proved to be difficult to cross-calibrate radiometers with sufficient accuracy. It has been proposed that radiometers be placed on the International Space Station. This would allow periodic recalibrations from the ground. The results from several radiometer experiments are summarized in Fig. 6, where the following is concluded:

- TSI is 1367 W/m^2 $\pm$ 0.1%.
- Accuracy (absolute calibration) is approximately 0.1%.
- Precision (scatter or noise) is 0.01%.

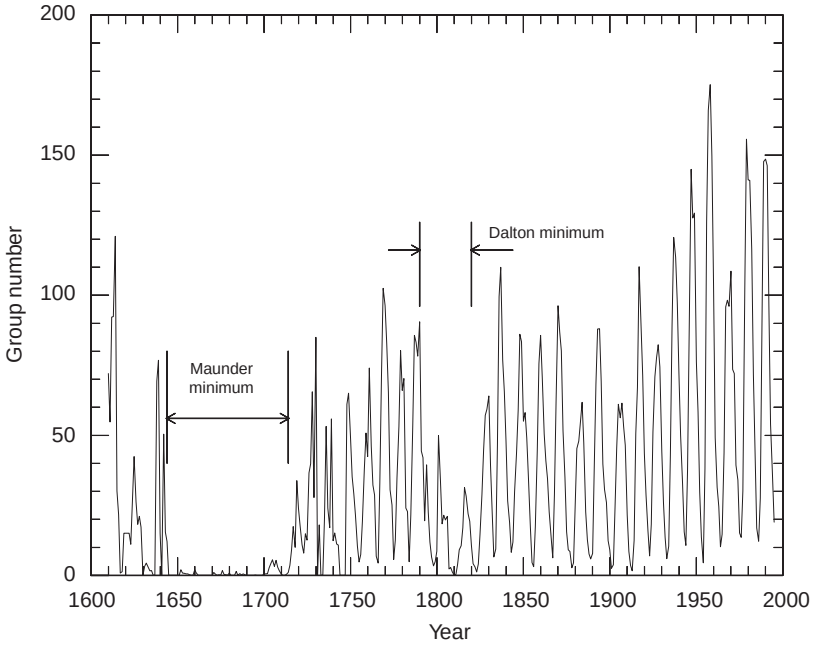


FIGURE 5 Group sunspot numbers clearly displaying the Maunder and Dalton minima.

TABLE I

Total Energy (W/m^2 at 1 AU)

Source	Energy	Activity cycle variation (percentage)
Radiation		
Irradiance	1367.10	0.08
Particle and fields		
Neutrinos	50	50 (?)
Solarwind	10^{-3}	None (?)
Magnetic field	10^{-5}	None (?)

TABLE II

Distribution of Energy (W/m^2 at 1 AU)

Wavelength	Energy	Activity cycle variation (percentage)
4 μm to ∞	0.7	—
300 nm to 4μm	1350	0.08
120 to 300 nm	16	10–50
Lyman-α	3×10^{-3}	100
30 to 120 nm	2×10^{-3}	150
10 to 30 nm	10^{-5}	—

Most of the TSI variance displayed in Fig. 6, within any given experiment, arises not from errors but rather from true solar variability imposed by dark sunspots (-0.1%) and bright faculae ($+0.1\%$). These features are seen on the solar disk in Fig. 7. Both sunspots and faculae are magnetic features; thus, the magnetic activity cycle is revealed. Careful photometry of the solar disk by astronomers at California State University, Northridge, has demonstrated that 90% of the TSI variance can be accounted for by sunspots and faculae. New results from the high-resolution Swedish Telescope indicate that perhaps tiny, not previously seen, faculae may make up part of the remaining 10%.

When experts model the radiometer experiments and attempt to connect the various experiments, two different time series emerge. One, by Claus Frohlich and Judith Lean, is seen in the upper part of Fig. 8. This record displays no trends beyond the activity cycle. The second and lower plot, by Richard C. Willson and Alex Mordvinnov, displays increased irradiance ($+0.04\%$) at the 1995–1996 minimum. Curiously, the latter finds possible supporting evidence from the behavior of Fraunhofer line strengths, as observed with the 1-m Fourier Transform Spectrometer (FTS) at Kitt Peak in Arizona (Fig. 9). Besides the FTS data, David Gray has studied the behavior of the weak temperature-sensitive C (538-nm) line, which is formed at practically the same level as the white light continuum giving rise to the TSI signal. Gray found an increase in

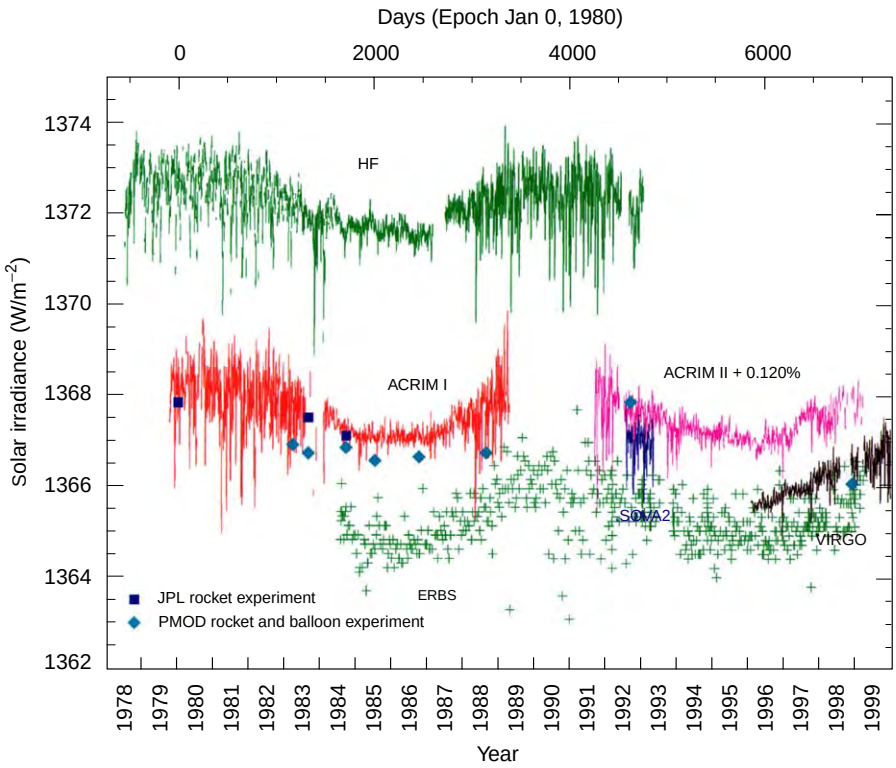


FIGURE 6 Raw total solar irradiance data, uncorrected except for adjustment to 1 AU. HF, Hickey–Frieden experiment on Nimbus 7; ACRIM, Active Cavity Radiation Irradiance Monitor on Solar Max Mission (SMM) and later on Earth Radiation Budget Satellite (ERBS); SOVA2, a solar variability experiment that has flown on several short-term flights; VIRGO, Variability of Solar Irradiance and Gravity Oscillations experiment on the SOHO spacecraft. Results from several rocket and balloon flights are included. The latter offer the advantage of postflight calibrations.

photospheric temperature of 0.014 K/year. These various, and in some ways discordant, results remain open to further study at this time.

A new set of radiometers was launched on the Solar Radiation and Climate Experiment (SORCE) early in 2003. The aim of this experiment is complete coverage of the solar output from X rays to 2.7 μm in the infrared. This data had not been released at the time of this writing. A proposed instrumental advance could be "cryogenic radiometers," but such devices have not yet been tested in space.

One might expect the quiet sun atmosphere to be heated by faculae or perhaps cooled by sunspots. Full-disk magnetograms (i.e., pictures that map the sun's magnetism) look very different at activity maximum from how they look at activity minimum (Figs. 2 and 3). At maximum, much of the solar surface seems to be covered by magnetic fields, but such pictures are misleading for two reasons: insufficient resolution and the fact that the fields expand above the photosphere, where the measurements are made, so as to appear more contiguous than they really are. Recent high-resolution pictures show that fields at the photospheric surface are in fact sparse compared with nonmagnetic surface area. Figure 10, taken in an active region near the disk center and at much higher resolution than either the facular picture in Fig. 6 or the magnetograms, illustrates this. It is not quite a picture in white light; rather, it was taken in the "G-band" (i.e., a narrow blue filter centered at 430.4 nm), which favors the middle photosphere level. This allows one to see faculae at the disk center. The small bright wiggles are really faculae that are cospatial with the magnetic fields on magnetograms. Notice that these faculae largely avoid the granules. Time-series "movies" show that the granules push around the faculae but are otherwise unaffected by them. In other words, magnetic fields thread through the solar surface without any apparent interaction with the gas there. Temperature measurements (using temperature-sensitive spectrum lines) show that the quiet sun gas is in fact the same temperature at solar minimum as at solar maximum. This means the apparent TSI rise at solar maximum represents magnetic features only and not temperature change of the global sun. (The previously cited results of Willson

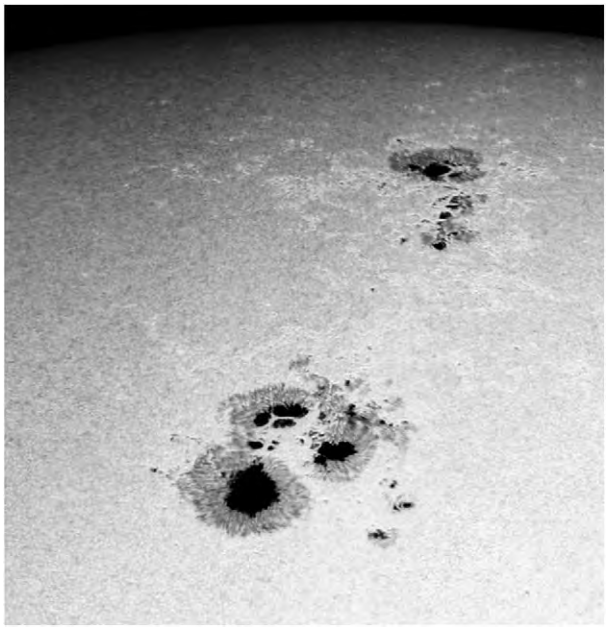


FIGURE 7 Photograph showing apparent sources of TSI variation that are evident near the solar limb. Dark sunspot umbrae, with surrounding fibral penumbrae, suppress radiation. Bright faculae, which are longer lived than sunspots, cause a net rise of TSI. Limb darkening, or the decrease in white light disk intensity as one approaches the limb, is not known to vary with time.

and Mordvinnov, and of Gray, of a temperature rise minute to minute runs counter to this statement, and the matter requires further clarification.)

Theory supports these observational results. One might expect to see a bright ring around a sunspot to balance the energy trapped at the spot umbra. Henk Spruit modeled this condition and found that the energy diffuses into the convective envelope with a time constant of 1 month to 1 year. There is no localized increase in temperature. (Reported bright rings around spots turn out to be indications of abnormal granulation that is pushed around by facular fields.)

If the sun's temperature is constant, maximum to minimum, where does the 0.1% excess energy at solar maximum come from? Eugene Parker, eminent theorist of the "solar wind" and "solar dynamo," concluded that at this time we do not understand the exact physics of a variable-output sun.

Although the sun's radiative output in the UV is very small (Table II), emission lines such as Lyman-α display a large cycle variability and affect the earth's thermosphere. During solar maximum, there can be a threefold increase in the atmospheric density at an altitude of 600 km. This induces noticeable drag to low earth orbit satellites such as the Hubble Space Telescope, so they must be reboosted from time to time.

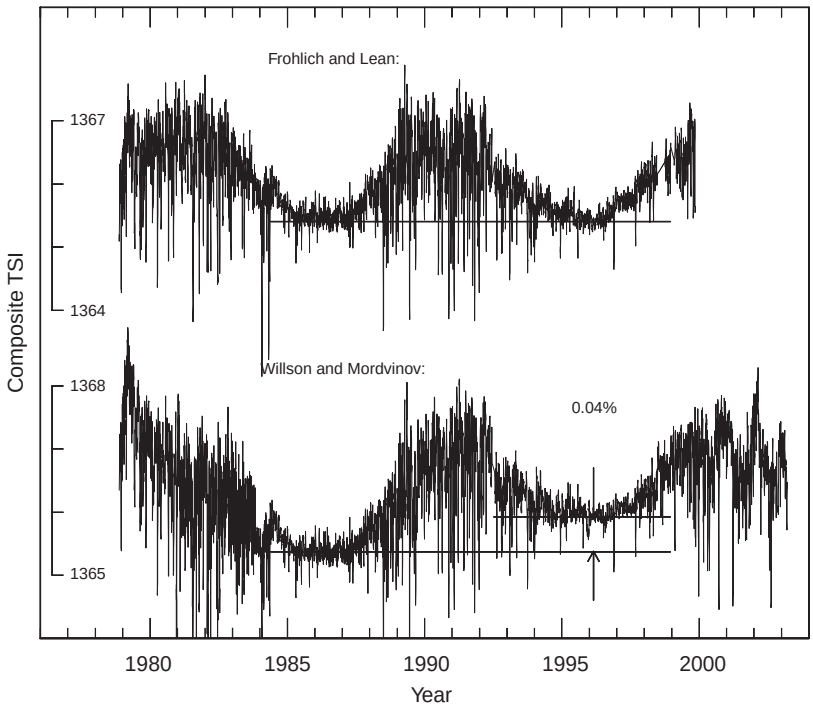


FIGURE 8 Same data as in Fig. 6 modeled in an attempt to provide a composite record capable of showing longer time trends. The Frohlich and Lean result at top shows the activity cycle with no trends. The Willson and Mordvinov construction at bottom displays a 0.04% rise between solar minima in 1985–1987 and 1995–1997.

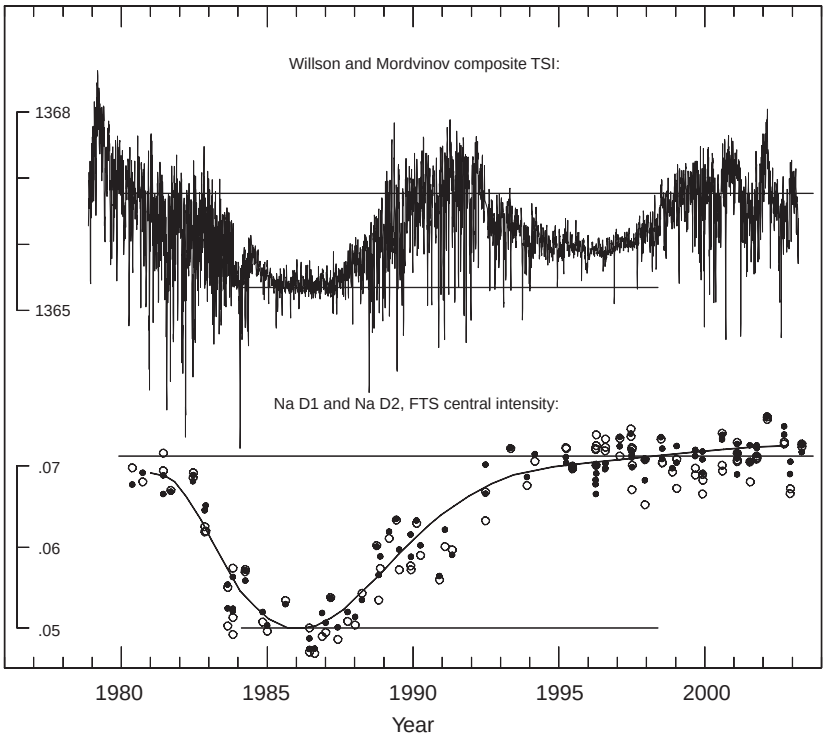


FIGURE 9 Comparison of the Willson and Mordvinov TSI solar minimum rise with the behavior of the Fraunhofer lines formed in the high photosphere. These lines failed to show the 1995–1996 minimum seen clearly in chromospheric lines. Plotted here is the central intensity of the strong sodium D lines. Other strong high photosphere lines, such as Fe and Mg, behave about the same. This indicates a temperature rise in the high photosphere at the 1996 minimum as compared with the 1986 minimum.

4. SOLAR WIND AND PARTICLE FLUX

Charged particles continually spiral outward from the sun to interact with the earth's magnetosphere. This solar wind, which is part of the extended solar atmosphere or heliosphere, traces out an Archimedes spiral path defined by the interplanetary magnetic field. Table III lists the bulk properties of the solar wind. One might expect that the chemical composition of the wind would mimic that of the photosphere, but it does not. Elements whose first ionization potential are below 10 eV are sometimes enhanced by a factor of roughly four. Some sort of selection mechanism takes place, probably in the chromosphere.

Pictures of the upper solar atmosphere, the transition region, indicate that the majority of surface fields are closed bipolar loops. However, surface evolution of magnetism may result in sizable areas where the field lines open out. These areas are called "corona holes" and are the sites of prolonged steady particle flows. When positioned favorably on the disk, coronal holes are identified with geomagnetic storms. Most easily seen in X rays and extreme ultraviolet (EUV) pictures, such as in Fig. 11, corona holes can also be mapped from the ground using the intensity of the He (1083-nm) line.

Besides the solar wind, coronal mass ejections (CMEs) and short-lived chromospheric flares can disturb the earth environment. Figure 12 shows a large CME in progress. The frequency of CMEs varies with the solar cycle, being weekly at solar minimum and increasing to once or twice daily at solar maximum. The fact that CME activity persists at solar minimum (when the TSI is constant) is evidence that there is more to the sun's output than is revealed purely by disk radiometry.

Solar flares are a result of unstable magnetic conditions, usually associated with sunspots. Observationally, a solar flare is seen in the H-α line as a brightness increase that lasts from a few minutes up to an hour. Gamma rays, hard X rays, extreme UV radiation, and microwave radio emission are released, mostly during the "flash phase" that lasts only a few minutes. This electromagnetic radiation reaches the earth at the speed of light in 8

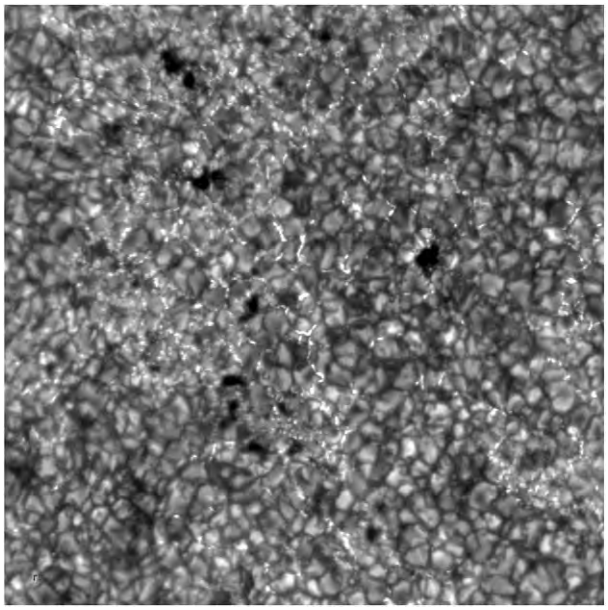


FIGURE 10 High-resolution photograph in an active region near disk center. The mottles are solar granulation similar to that in Fig. 1. Note that some granules are bright and others are not so bright. A granule is surrounded by a dark area (intergranule lane). The tiny dark features are pores, which are small sunspots without penumbrae. The bright squiggles are called filigree and represent the site of nonsunspot fields or faculae. Notice that the filigree favor intergranule lanes. Even though this is part of an active region, the relative area occupied by magnetic fields is small, and the fields thread through the granules with little or no interaction (Swedish Telescope, La Palma, Canary Islands, September 20, 1993).

TABLE III

Bulk Properties of the Solar Wind

	Low speed	High speed	Average
Particle number density N (cm^{-3})	12	4	9
Particle velocity V (km/s)	327	702	468
Particle energy (W/m^2) 2.15×10^{-3}			1.68×10^{-3} 2.05×10^{-3}
Composition (N_α/ Np)	0.038	0.048	0.047
Magnetic field B (y)	—	—	6.2 ± 2.9
Magnetic energy density $B^2/8\pi$ (W/m^2)	—	—	1.7×10^{-13}
Magnetic energy flux $V(B^2/8\pi)$ (W/m^2)	5.6×10^{-6}	12×10^{-6}	8×10^{-6}

min. Particles in the form of protons and alpha particles are also released in abundance, but like solar wind, they are confined to the Archimedes spiral of the interplanetary field and may take a day

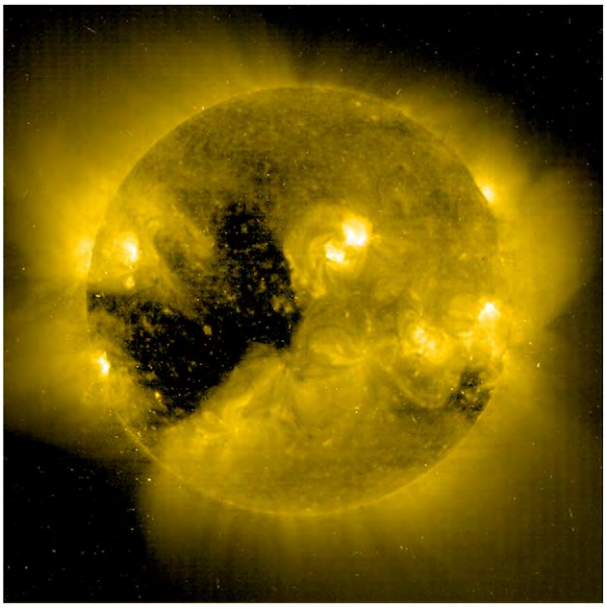


FIGURE 11 A coronal hole clearly displayed in EUV picture from the SOHO satellite on June 24, 2003. The narrow-band filter employed here is centered on Fe XV 284 A. When this hole rotates to the western side of the sun a week later, particles from the hole will spiral out to the earth and cause geomagnetic storms.

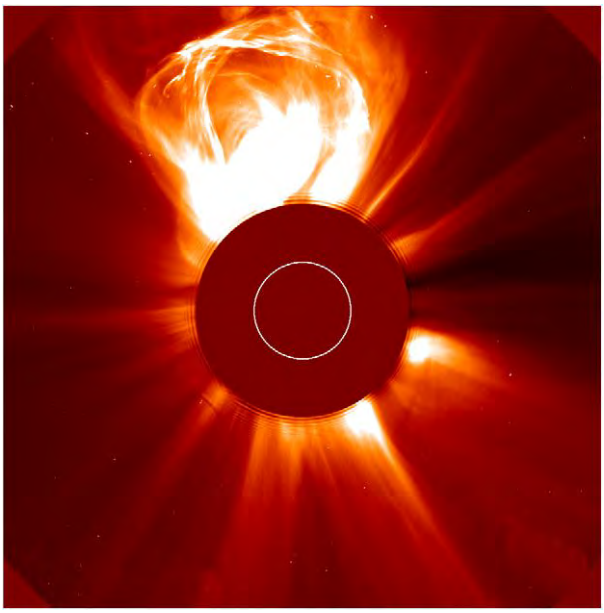


FIGURE 12 Coronal mass ejection (CME) observed on February 17, 2000, in white light by the Large Angle Spectrographic Coronagraph (LASCO) experiment on the SOHO satellite. This eruption is in the plane of the sky and not toward the earth. The solar disk is occulted with the position and size of the actual disk marked. In addition to the CME bubble, one can see the coronal streamers that are typical of solar maximum. The field of view is 6 $R_\odot$.

or more to reach the earth. Because of their spiral path, a flare site 50° west of the meridian on the solar disk is more propitious for producing a terrestrial interaction.

5. ORBITAL MODULATION

Table IV lists geometric factors related to the energy input to the earth. Apart from the 100% modulation of the night/day diurnal cycle, the first item is solar rotation. The sun rotates with a period of 27 days in a sidereal reference frame or 25 days synodic as seen from the earth. Active regions are not distributed uniformly around the sun, and as they transit the visible disk, because of solar rotation, they modulate the TSI ($\pm 0.1\%$), $\pm 1\%$ in the near UV, and $\geq 25\%$ at Lyman-α.

The second item, and a fundamental consideration, is the variable earth–sun distance owing to the earth's elliptical orbit. This induces a 365-day modulation of $\pm 1.7\%$ in apparent TSI. Maximum energy occurs around January 4 (perihelion), whereas minimum energy occurs around July 7 (aphelion). The time of perihelion precesses with a 26,000-year period.

The third effect is more subtle and is caused by the inclination of the sun's equator to the ecliptic, or plane, of the earth's orbit. The inclination is zero on January 5 and June 6. It is a maximum of $+7.2°$ on September 8 and $-7.2°$ on March 7. Its importance depends on the way in which active regions are distributed in latitude over the solar disk.

The fourth item is the Milankovich effect. In 1920, Milankovich, a Yugoslav astronomer, calculated that, owing to planetary perturbations; eccentricity of the earth's orbit varies with a period of approximately 100,000 years, obliquity oscillates with a period of 41,000 years, and the spin axis precesses with a period of 23,000 years. These factors, which are hard to quantify exactly, obviously influence solar insolation and may play a role in producing ice ages.

The fifth item, recently resurrected by astronomer Willie Soon, has to do with the wobble of the sun around the center of mass (barycenter) of the planetary system. In 1965, P. D. Jose pointed out that this motion ranges from 0.01 to $2.19\,R_\odot$, with a period of 178.7 years. Correlations with sunspot numbers (always suspect) have been pointed out, but a physical basis for such a motion affecting sunspots is lacking.

TABLE IV

Geometric Effects

Effect	Period	Modulation amplitude
Solar rotation	27 days	λ-dependent exponential decrease with λ in UV
		1% at λ 300 nm
		25% at λ 120 nm
Earth–sun distance	365 days	$\pm$ 1.7% (maximum on January 2)
		Precesses through the year, with 26,000 year period
Inclination of sun's equator to ecliptic	365 days	$B_0 = 0°$ January 5 and June 6
		$B_0 = -7.2°$ March 7
		$B_0 = +7.2°$ September 8
Milankovich effect	40,400 years	1.2% low latitudes
		2.5% high latitudes
Archimedes spiral	—	Optimum solar long: 50° west of central meridian
		Delay 1 to 4 days
Barycenter motion	178.7 years	?

6. LONG-TERM VARIATIONS IN SOLAR OUTPUT

This section first looks at what stellar evolution theory tells us about the past and future of the sun on cosmic time scales. Then, it examines the sunspot number record for evidence of solar variability in our own lifetimes and somewhat earlier. It extends the record back many millennia, using radioisotopes as tracers of activity cycles. Finally, it inquires about what other solar-type stars say about cycles in general and how this relates to the sun.

When the sun collapsed out of the solar nebulae, gravity was balanced by pressure, and possibly by magnetic fields, until the sun evolved into a stable "main sequence" star some 5×10^9 years ago. Initially, hydrogen burning was confined to the sun's extreme core, and its luminosity was then 70% of its current value. If the sun suddenly assumed this reduced output, the earth would freeze over. There is little or no geological evidence that this has happened in the past, and this early dim sun is considered a puzzle. Presumably the explanation has to do with the composition of the earth's primeval atmosphere, which was completely different from what it is today.

Currently, hydrogen burning in the p-p reaction is confined to within $0.3\,R_\odot$. As hydrogen is further depleted after another 5×10^9 years, hydrogen burning will be confined to a shell of increasing radius and

the sun will expand greatly by a factor of 30, becoming a luminous red giant. Life will no longer be possible on earth. In fact, recent modeling suggests that the earth will become uninhabitable in only 500 million years. After the hydrogen is gone, the sun will contract, burn helium, and eventually end up a white dwarf the size of the earth. That is the big picture.

Of more interest today is solar variability at the current time and in the immediate past. In 1976, Jack Eddy compared the historic sunspot number record with auroral sightings, and with terrestrial reservoirs of ^{14}C time series, to identify the Maunder minimum. Of course, this was pre-ACRIM, but his finding of a correlation between solar and terrestrial events established the validity of the exercise.

Sunspot number archives are by no means of uniform quality over time. Since the international geophysical year (1956–1957), the competition has been keen to observe and report on sunspot numbers. But before that, the efforts were mixed, and during the century following Galileo's telescopic discovery of sunspots, we have little idea how intensively the sun was observed. French astronomers, particularly scientists such as Elizabeth Ribes at Meudon, have searched obscure sources for missing sunspot data. Recently, Doug Hoyt and Ken Schatten found that by plotting sunspot group numbers rather than simply numbers of sunspots, the early record became more complete and possibly more reliable (Fig. 5). However, there is little doubt as to the reality of the Maunder minimum.

The historic sunspot record extends back roughly 400 years, whereas radioisotope tracers can, in principle, yield information on solar activity to 100,000 BP (before present). Following a recent review article by Jürg Beer, this article now examines the past millennia of solar activity using radioisotopes found in tree rings and glacial ice as tracers.

Galactic cosmic rays flow through the solar system continually. So far as is known, this flux of cosmic rays is constant. Galactic cosmic rays are made up of protons (87%), helium nuclei (12%), and heavier elements (1%). Depending on energy, these particles interact with nitrogen and oxygen in the atmosphere to produce ^{14}C and ^{10}Be. Disturbances of the earth's geomagnetic field, and the magnetic field of the heliosphere, scatter cosmic rays and reduce their terrestrial flux. In particular, coronal mass ejections, which are associated with sunspot activity, modulate the flux of the low-energy component of cosmic rays.

Naturally occurring isotopes are ^{12}C and ^{9}Be, whereas the cosmic ray products are the radioisotopes ^{14}C and ^{10}Be. ^{14}C has a lifetime of 5730 years and a cosmic ray production rate in the atmosphere of 2 atoms/cm^2/s. ^{10}Be has a lifetime of 1.5×10^6 years but a production rate of only 2% that of ^{14}C.

After creation, ^{14}C undergoes a rather complex mixing process, being circulated between the atmosphere and oceans before finally being absorbed by the biosphere (e.g., in trees). For this reason, its ability to show the 11-year Schwabe cycle is attenuated by a factor of 100 and tree rings do not show the 11-year cycle. The isotope ratio $^{14}C/^{12}C$ displays up to 10% variation over the past 10,000 years, with little latitude dependence. Long-term trends are thought to be due mostly to slow changes in the geomagnetic field, with a weak but clear solar modulation superimposed to mark the Maunder minimum.

^{10}Be has a much more direct path from its source of origin in the stratosphere until, for example, deposition in snow on a Greenland glacier. Its time resolution is roughly 1 year, so that the Schwabe cycle is nicely revealed. In common with ^{14}C, it is modulated mainly by the geomagnetic field. It shows a decided preference for polar latitudes, whereas ^{14}C is latitude independent. These differences between ^{14}C and ^{10}Be can be used to cancel the dominant signal from geomagnetism and geochemical effects. Anthropogenic change from carbon-burning fuels is a new and disturbing signal during the 19th century.

A spectral analysis by Paul Damon and colleagues of the ^{14}C tree ring records revealed the 88-year Gleissberg cycle. Similar study of ^{10}Be ice core records showed the 11-year Schwabe cycle and a 205-year cycle that is named after De Vries. Both ^{14}C and ^{10}Be records show periods of low activity that are designated Maunder, Spoerer, and Wolf minima. Although dating is a potential problem, ^{10}Be ice cores can trace the Schwabe cycle back 100,000 years.

Finally, this article turns to solar-type stars as a guide to activity cycles and what to expect. The sun is classified as a G2V (i.e., a main sequence dwarf), based mainly on the relative strengths of its spectrum lines. (Astronomers classify stars as O, B, A, F, G, K, or M, based on low dispersion spectra. The series represents a run of temperature, with O being hot stars with He II absorption and M being cool stars with strong molecular TiO bands. Each class is further subdivided by numbered category (1, 2, …9). In addition, there are luminosity classes ranging from I (supergiants) to V (main sequence dwarfs). An experienced observer can immediately classify a star simply by inspection of its spectrum. The photometric color of stars is a rough guide to spectral type and is a handy method when large numbers of stars must be examined. Of course, spectrum study requires spectra.)

During the 1960s, Olin Wilson began a program of intensive observations of the calcium H and K lines in stars judged to be similar to the sun. This was undertaken at Mount Wilson Observatory in Southern California, and many well-known astronomers, notably Sallie Baliunas and Robert Noyes, joined in the effort. The idea was an exciting one. Are there stars out there just like our sun? Of course, there are vast numbers of stars to be considered. The first criterion was blue–visual magnitude (B-V) or color (for the sun, B-V = 0.65), then spectral class as described previously. Finally, there is the question of age gleamed from the spectrum. It is known that activity cycles decline with age. Some idea about activity, and hence age, is inferred from the strength of the calcium H and K lines.

One of the first results of this search was a curious dearth of stars like the sun—solar twins. Even to this day, and after studying hundreds of candidates, no exact twin has been found. At the time of this writing, the closest solar-like star is 18 Scopii, whose spectrum is roughly the same but whose cycle activity has a greater amplitude.

On the positive side, many cycles are found that are comparable to that of the sun. Other stars, which otherwise qualify as to age and spectrum, exhibit little or no cyclic amplitude. It is conjectured that these stars may be in a state comparable to the Maunder minimum. Looking back on the millennia records from ^{14}C, one will notice that the sun has been inactive every few hundred years or roughly one-third of the time. This condition could account for stars without cycles, and it supports the hypothesis that the sun may again become inactive. On the other hand, depending on circumstances at the birth of a star, there may be cases where the nebular condensation had very low angular momentum initially. This would create stars without appreciable activity cycles, that is, stars in a perpetual state of Maunder minimum.

The question of age is a complication circumvented by a scheme devised by Mark Giampapa and Richard Radick. With help from fellow astronomers, they have been observing solar-type stars in the open-cluster M67. The advantage is that all of the stars would be expected to have the same age, thereby eliminating that variable. What has been found so far is that solar-type cycles do dominate but that there seems to be a separate class of extra-strong cycles. This project has been ongoing.

The ^{10}Be archives strongly suggest, from signals of solar-type stars and from solar activity such as CMEs (which continue to bombard the earth even at solar minimum), that there is a lot going on with the sun of which we are currently unaware. Plenty of work exists for future generations of solar physicists.

Acknowledgments

The author acknowledges help in the preparation of this article from Bernard Durney, Peter Foukal, Harry Jones, Steve Keil, Elena Malaushenko, Judit Pap, and Dick White.

SEE ALSO THE FOLLOWING ARTICLES

Earth's Energy Balance • Global Material Cycles and Energy • Lunar–Solar Power System • Materials for Solar Energy • Origin of Life and Energy • Photosynthesis and Autotrophic Energy Flows • Solar Energy, History of • Solar Fuels and Materials

Further Reading

Beer, J. (2000). Long-term indirect indices of solar variability. *Space Sci. Rev.* **94,** 53.

Cox, A. N., Livingston, W. C., and Mathews, M. S. (eds.). (1991). "Solar Interior and Atmosphere." University of Arizona Press, Tucson.

Eddy, J. A. (1976). The maunder minimum. *Science* **192,** 1189.

Foukal, P. (1990). "Solar Astrophysics." John Wiley, New York.

Frohlich, C., and Lean, J. (1998). The sun's total irradiance: Cycles, trends, and related climate change uncertainties since 1976. *Geophys. Res. Lett.* **25,** 4377.

Jose, P. D. (1965). Sun's motion and sunspots. *Astron. J.* **70,** 193.

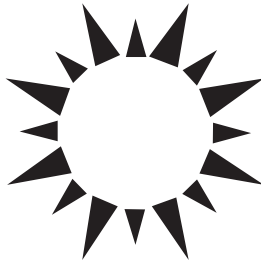
Parker, E. N. (2000, June). The physics of the sun and the gateway to the stars. *Physics Today*, p. 26.

Spruit, H. (1982). Effect of spots on a star's radius and luminosity. *Astron. Astrophys.* **108,** 348.

Stix, M. (2002). "The Sun." Springer, New York.

White, O. R., Rottman, G. J., Woods, T. N., Knapp, B. C., Keil, S. L., Livingston, W. C., Tapping, K. F., Donnelly, R. F., and Puga, L. C. (1994). Change in the radiative output of the sun in 1992 and its effect in the thermosphere. *J. Geophys. Res.* **99,** 369.

Sustainable Development: Basic Concepts and Application to Energy



MOHAN MUNASINGHE
Munasinghe Institute for Development
Colombo, Sri Lanka

1. **Basic Framework**
2. **Elements of Sustainable Development**
3. **Sustainable Energy Development**
4. **Applying the Sustainable Energy Development Framework**

Glossary

Action Impact Matrix (AIM) A sustainomics tool that helps to identify and analyze economic–environmental–social interactions and to formulate effective sustainable development policies; individual cells in the matrix help to explicitly identify key linkages, focus attention on methods of analyzing important impacts, and suggest action priorities and remedies.

durability An integrative approach focusing primarily on sustaining the quality of life, typically by meeting prudent environmental, social, and economic sustainability requirements; durable development paths permit growth but are not necessarily economically optimal.

economic sustainability Maximizing the flow of income that could be generated while at least maintaining the stock of assets (or capital) that yields these beneficial outputs; economic efficiency continues to ensure both efficient resource allocation in production and efficient consumption choices that maximize utility.

environmental sustainability Preserving the overall viability and normal functioning of natural systems, including improving system health and dynamic ability to adapt to change across a range of spatial and temporal scales; for ecological systems, sustainability is defined by a comprehensive, multiscale, dynamic hierarchical measure of resilience, vigor, and organization.

equity An ethical and people-oriented concept focusing on the basic fairness of both the processes and outcomes of decision making, including the distribution of economic, social, and environmental assets and entitlements.

optimality An approach used in economic analysis to generally maximize (or minimize) a single objective such as welfare (or costs), subject to the requirement that the stock of productive assets (or welfare itself) is non-decreasing in the long term.

social sustainability Preserving the overall viability and normal functioning of social systems.

sustainable development A process for improving the range of opportunities that will enable individual humans and communities to achieve their aspirations and full potential over a sustained period of time while maintaining the resilience of economic, social, and environmental systems.

sustainable energy development The harnessing of energy resources for human use in a manner that supports lasting development.

sustainomics A transdisciplinary, integrative, comprehensive, balanced, heuristic, and practical meta-framework for making development more sustainable.

Following the 1992 Earth Summit in Rio de Janeiro, Brazil, including the adoption of the United Nations' Agenda 21, and the 2002 World Summit on Sustainable Development in Johannesburg, South Africa, sustainable development has become well accepted worldwide. World decision makers are looking at this approach to solve many critical problems, including traditional development issues (e.g., energy use, economic stagnation, poverty, hunger, illness) as well as newer challenges (e.g., environmental degradation, globalization).

1. BASIC FRAMEWORK

Sustainable energy development involves the harnessing of energy resources for human use in a manner

 © 2004 Elsevier Inc. All rights reserved.

that supports lasting development. This article begins with a review of sustainable development itself before describing its application to energy. The World Commission on Environment and Development originally defined it as "development which meets the needs of the present without compromising the ability of future generations to meet their own needs." Among many subsequent definitions, the sustainable development triangle in Fig. 1A shows one widely accepted concept proposed by Munasinghe at the 1992 Earth Summit in Rio de Janeiro, Brazil. It encompasses three major perspectives: economic, social, and environmental. Each viewpoint corresponds to a domain (and system) that has its own distinct driving forces and objectives. The economy is geared toward improving human welfare, primarily through increases in consumption of goods and services. The environmental domain focuses on protecting the integrity and resilience of ecological systems. The social domain emphasizes enriching human relationships, achieving individual and group aspirations, and strengthening values and institutions.

Meanwhile, energy has emerged as a key resource that interacts critically with the economic, social, and environmental dimensions of sustainable development. First, it has long been perceived as a major driving force underlying economic progress, and in turn, economic growth itself further stimulates energy demand. Second, energy production and use are strongly interlinked with the environment. Third, energy is a basic human need that significantly affects social well-being. Recently, growing energy demand has also become associated with global climate change, posing an unprecedented challenge to humanity. The wide-ranging potential impacts of energy production and consumption on sustainable development, and vice versa, are reviewed in this article.

Given the lack of a specific approach or practical framework that attempts to define, analyze, and implement sustainable development, Munasinghe proposed the term "sustainomics" to describe "a transdisciplinary, integrative, comprehensive, balanced, heuristic, and practical meta-framework for making development more sustainable." Figure 1B indicates how the sustainomics framework (or the emerging "science of sustainable development"), as well as the associated transdisciplinary knowledge base, would support a comprehensive and balanced assessment of trade-offs and synergies that might exist among the economic, social, and environmental dimensions of sustainable development. Balance is also needed in the relative emphasis placed on traditional development (which is more appealing to the South) versus sustainability (which is emphasized by the North). The optimality and durability approaches (described later) play a key role in integrating economic, social, and environmental issues. Many disciplines contribute to the sustainomics framework given that sustainable development itself involves every aspect of human activity, including complex interactions among socioeconomic, ecological, and physical systems.

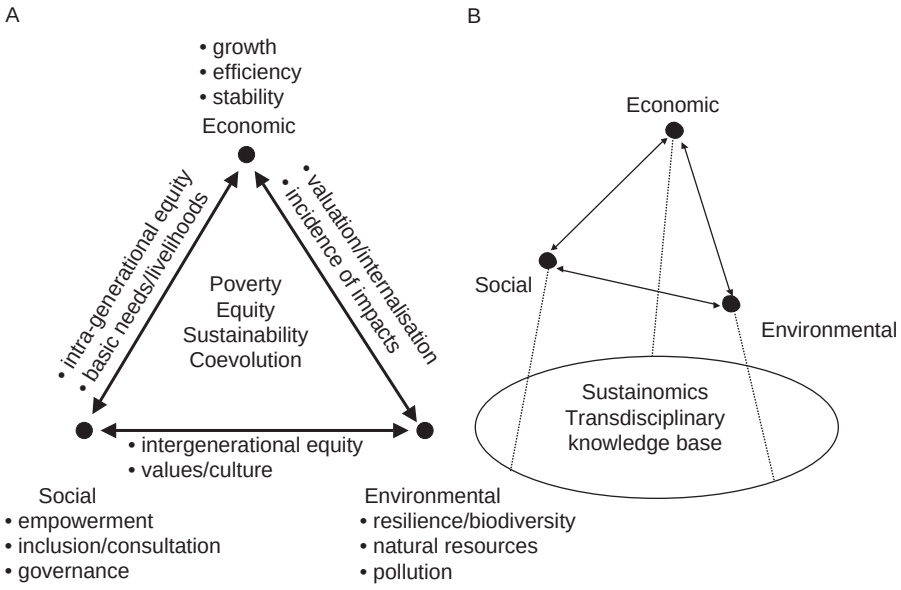


FIGURE 1 (A) Elements of sustainable development. (B) Sustainable development triangle supported by the sustainomics framework. Adapted from Munasinghe (1992).

Historically, the development of the industrialized world has focused on material production. Not surprisingly, most industrialized and developing nations pursued the economic goal of increasing output and growth during the 20th century. Thus, traditional development was strongly associated with economic growth as well as with important social dimensions (poverty and equity are discussed later). By the early 1960s, the lack of "trickle down" benefits to the growing numbers of poor in developing countries resulted in greater efforts to improve income distribution directly. The development paradigm shifted toward equitable growth, where social (distributional) objectives, especially poverty alleviation, were recognized to be as important as economic efficiency. Protection of the environment has now become the third major pillar of sustainable development. By the early 1980s, a large body of evidence had accumulated that environmental degradation was a major barrier to development, and new proactive safeguards were introduced (e.g., the environmental assessments).

In the sustainomics framework, sustainable development is described as a process for improving the range of opportunities that will enable individual humans and communities to achieve their aspirations and full potential over a sustained period of time while maintaining the resilience of economic, social, and environmental systems. Thus, sustainable development requires both increases in adaptive capacity and opportunities for improving economic, social, and ecological systems. Improving adaptive capacity increases resilience and sustainability. Expanding the set of opportunities for improvement gives rise to development. Heuristic behavior of individual organisms and systems facilitates learning, the testing of new processes, adaptation, and improvement.

The precise definition and implementation of sustainable development remains an ideal, elusive, and perhaps unreachable goal. Sustainomics proposes a less ambitious, but more focused and feasible, strategy that merely seeks to "make development more sustainable." Such an incremental (or gradient-based) method is more practical because many unsustainable activities are easier to recognize and eliminate. This approach seeks continuing improvements in the current quality of life at a lower intensity of resource use, thereby leaving behind for future generations an undiminished stock of productive assets (e.g., manufactured, natural, and social capital) that will enhance opportunities for improving their quality of life.

2. ELEMENTS OF SUSTAINABLE DEVELOPMENT

2.1 Economic Aspects

Economic progress is often evaluated in terms of welfare (or utility), measured as willingness to pay for goods and services consumed. Thus, many economic policies typically seek to enhance income and induce more efficient production and consumption of (mainly marketed) goods and services. The stability of prices and employment are among other important objectives.

Economic efficiency is measured against the ideal of Pareto optimality, which encourages actions that will improve the welfare of at least one individual without worsening the situation of anyone else. The idealized, perfectly competitive economy is an important (Pareto optimal) benchmark, where (efficient) market prices play a key role in both allocating productive resources to maximize output and ensuring optimal consumption choices that maximize consumer utility. If significant economic distortions are present, appropriate shadow prices need to be used. The well-known cost–benefit criterion accepts all projects whose net benefits are positive (i.e., aggregate benefits exceed costs). It is based on the weaker "quasi"-Pareto condition, which assumes that such net benefits could be redistributed from potential gainers to losers, leaving no one worse off than before. In more general terms, interpersonal comparisons of welfare are fraught with difficulty, both within and across nations and over time (e.g., the value of human life).

Based on the pioneering work of Lindahl and Hicks, the modern concept underlying economic sustainability seeks to maximize the flow of income that could be generated while at least maintaining the stock of assets (or capital) that yields these beneficial outputs. Economic efficiency continues to ensure both efficient resource allocation in production and efficient consumption choices that maximize utility. Problems arise in identifying the kinds of capital to be maintained (e.g., manufactured, natural, human, and social capital) and their substitutability. Often, it is difficult to value these assets and the services they provide, particularly for ecological and social resources. Even key economic assets may be overlooked, especially in informal or subsistence economies where non-market-based transactions are important. Meanwhile, the equation of welfare with monetary income and consumption has been challenged for many years. More recently, Maslow and

others have identified hierarchies of needs that provide psychic satisfaction beyond mere goods and services.

The issues of uncertainty, irreversibility, and catastrophic collapse pose additional difficulties in determining dynamically efficient development paths. Many commonly used microeconomic approaches rely on marginal analysis (e.g., comparing incremental costs and benefits of economic activities). Such methods assume smoothly changing variables and, therefore, are rather inappropriate for analyzing large changes, discontinuous phenomena, and sudden transitions among multiple equilibria. Recent work has begun to explore the behavior of large, nonlinear, dynamic, and chaotic systems and concepts such as system vulnerability and resilience.

2.2 Environmental Aspects

Development in the environmental sense is a recent concern relating to the need to manage scarce natural resources in a prudent manner because human welfare ultimately depends on ecological services. Ignoring safe ecological limits could undermine long-run prospects for development. The growing literature on the theme of environment and sustainable development includes books by Faucheux and colleagues, describing models of sustainable development, and Munasinghe and colleagues, explicitly addressing the links between growth and environment

The environmental interpretation of sustainability focuses on overall viability and normal functioning of natural systems. For ecological systems, sustainability is defined by a comprehensive, multiscale, dynamic hierarchical measure of resilience, vigor, and organization. In 1973, Holling defined resilience as the ability of an ecosystem to persist despite external shocks, that is, the amount of disruption that will cause the ecosystem to switch from one system state to another. An ecosystem state is defined by its internal structure and set of mutually reinforcing processes. Vigor is associated with the primary productivity or growth of an ecosystem. Organization depends on both complexity and structure of the system. For example, a multicellular organism such as a human is more highly organized than a single-celled amoeba. Higher states of organization imply lower levels of entropy. Thus, the second law of thermodynamics requires that sustainability of complex organisms and systems depends on the use of low-entropy energy derived from their environment that is returned as (less useful) high-entropy energy. The ultimate source of this energy is solar radiation.

In this context, natural resource degradation, pollution, and loss of biodiversity are detrimental because they increase vulnerability, undermine system health, and reduce resilience. The notions of a safe threshold and carrying capacity are important, often to avoid catastrophic ecosystem collapse. It is useful to also think of sustainability in terms of the normal functioning and longevity of a nested hierarchy of ecological and socioeconomic systems ordered according to scale; for example, a human community would consist of many individuals who are themselves composed of a large number of discrete cells. Gunderson and Holling used the term "panarchy" to denote such a hierarchy of systems and their adaptive cycles across scales. A system at a given level is able to operate in its stable (sustainable) mode because it is protected by slower and more conservative changes in the supersystem above it while being simultaneously invigorated and energized by faster changes taking place in subsystems below it.

Sustainable development is not necessarily synonymous with maintaining the ecological status quo. A coupled ecological–socioeconomic system could evolve while maintaining levels of biodiversity that guarantee resilience of ecosystems on which future human consumption and production depend.

2.3 Social Aspects

Social development usually refers to improvements in both individual well-being and the overall welfare of society resulting from increases in social capital, typically the accumulation of capacity enabling individuals and communities to work together. The institutional component of social capital involves formal laws as well as traditional or informal understandings that govern behavior, whereas the organizational component is embodied in individuals and communities operating within these institutional arrangements. The quantity and quality of social interactions underlying human existence (including levels of mutual trust and shared social norms) help to determine the stock of social capital. Thus, social capital grows with greater use and erodes through disuse, unlike economic and environmental capital, which are depreciated or depleted by use. Furthermore, some forms of social capital may be harmful (e.g., cooperation within criminal gangs).

There is also an important element of equity and poverty alleviation. Thus, the social dimension of development includes protective strategies that reduce vulnerability, improve equity, and ensure that basic needs are met. Future social development will

require sociopolitical institutions that can adapt to meet the challenges of globalization. The latter often destroys traditional coping mechanisms that have evolved in the past (especially to protect disadvantaged groups).

Social sustainability parallels environmental sustainability because human settlements, such as cities and villages, are also habitats. Reducing vulnerability and maintaining the health (e.g., resilience, vigor, organization) of sociocultural systems and their ability to withstand shocks are also important. Enhancing human capital (through education) and strengthening social values, institutions, and governance are key aspects. Many harmful changes occur slowly, and their long-term effects are often overlooked in socioeconomic analysis. Preserving cultural capital and diversity worldwide, strengthening social cohesion, and reducing destructive conflicts are integral elements of this approach. An important aspect involves empowerment and broader participation through subsidiarity, that is, decentralization of decision making to the lowest (or most local) level at which it is still effective. In summary, for both ecological and socioeconomic systems, the emphasis is on improving system health and its dynamic ability to adapt to change across a range of spatial and temporal scales rather than on conserving some "ideal" static state.

2.4 Poverty and Equity

Poverty and equity are two important issues that have social, economic, and environmental dimensions (Fig. 1A). More than 2.8 billion people (nearly half of the global population) live on less than U.S. $2 per day, and 1.2 billion barely survive on less than U.S. $1 per day. The top 20 percentile of the world's population consumes approximately 83% of total output, whereas the bottom 20 percentile consumes only 1.4%. Income disparities are worsening; the per capita ratio between the richest and the poorest 20 percentile groups was 30 to 1 in 1960 and increased to more than 80 to 1 by 1995.

Equity is an ethical and usually people-oriented concept with primarily social, and some economic and environmental, dimensions. It focuses on the basic fairness of both the processes and outcomes of decision making. The equity of any action may be assessed in terms of a number of generic approaches, including parity, proportionality, priority, utilitarianism, and Rawlsian distributive justice. Societies normally seek to achieve equity by balancing and combining several of these criteria.

Poverty alleviation, improved income distribution, and intragenerational (or spatial) equity are key aspects of economic policies seeking to increase overall human welfare. Broadly speaking, economic efficiency provides guidance on producing and consuming goods and services more efficiently but is unable to provide a means of choosing (from a social perspective) among alternative patterns of consumption that are efficient. Equity principles provide better tools for making judgments about such choices.

Social equity is also linked to sustainability because grossly unfair distributions of income and social benefits are unlikely to be lasting in the long run. Equity will be strengthened by enhancing pluralism and grassroots participation in decision making and by empowering disadvantaged groups. In the long term, intergenerational equity is vital, where the economic discount rate plays a key role with respect to both equity and efficiency aspects. The sustainomics framework outlines methods of reconciling potential conflicts between equity and economic efficiency.

Equity in the environmental sense has received recent attention due to disproportionately greater environmental damages suffered by poor groups. Thus, poverty alleviation efforts are being broadened (beyond raising monetary incomes) to address the degraded environmental and social conditions facing the poor.

In summary, both equity and poverty have not only economic but also social and environmental dimensions; therefore, they need to be assessed using a comprehensive set of indicators (rather than income distribution alone). From an economic policy perspective, emphasis needs to be placed on expanding employment and gainful opportunities for poor people through growth, improving access to markets, and increasing both assets and education. Social policies would focus on empowerment and inclusion by making institutions more responsive to the poor and removing barriers that exclude disadvantaged groups. Environmentally related measures to help poor people might seek to reduce their vulnerability to disasters, crop failures, loss of employment, sickness, economic shocks, and so on. Thus, an important objective of poverty alleviation is to provide poor people with assets (e.g., social, natural, economic) that will reduce their vulnerability. Such assets increase the capacity for both short-run coping and longer run adaptation to external shocks. The forgoing ideas blend with the sustainable livelihoods approach, which focuses on access to portfolios of

assets, capacity to withstand shocks, gainful employment, and social processes.

An even broader nonanthropocentric approach to equity involves the concept of fairness in the treatment of nonhuman forms of life or even inanimate nature. One view asserts that humans have the responsibility of prudent "stewardship" over nature, and this goes beyond mere rights of usage.

2.5 Consistent Integration of Economic, Social, and Environmental Considerations

This subsection compares the concepts of ecological, social, and economic sustainability. Maintaining the set of opportunities, as opposed to preserving value of the asset base, is important. The preservation of biodiversity maintains options and allows the system to retain resilience by protecting it from external shocks, in the same manner that preservation of the capital stock protects economic assets for future consumption. Differences emerge because under the Hicks–Lindahl income measure, a society that consumes its fixed capital without replacement is not sustainable, whereas using an ecological approach, loss of resilience implies a reduction in the self-organization of the system but not necessarily a loss in productivity. In the case of social systems, resilience depends on the capacity of human societies to adapt and continue functioning in the face of stress and shocks. Thus, linkages between sociocultural and ecological sustainability emerge through the organizational similarities between human societies and ecological systems and the parallels between biodiversity and cultural diversity. In the longer term, the concept of coevolution of social, economic, and ecological systems within a larger, more complex adaptive system provides useful insights regarding harmonious integration of the various elements of sustainable development (Fig. 1A).

It is important to integrate and reconcile the economic, social, and environmental aspects within a holistic and balanced sustainable development framework. Economic analysis has a special role in contemporary national policymaking because some of the most important decisions fall within the economic domain. Although mainstream economics that is used for practical policymaking has often ignored many crucial aspects of the environmental and social dimensions of sustainable development, there is a small but growing body of literature that seeks to address such shortcomings.

Two broad approaches, based on the concepts of optimality and durability, are relevant for integrating the economic, social, and environmental dimensions of sustainable development. Although there are overlaps between the two approaches, the main thrust is somewhat different in each case. Uncertainty often plays a key role in determining which approach would be preferred. For example, relatively steady and well-ordered conditions may encourage optimizing behavior by planners attempting to control and even fine-tune outcomes, whereas subsistence farmers facing chaotic and unpredictable circumstances might opt for more durable responses that simply enhance survival prospects.

2.6 Optimality

The optimality-based approach has been widely used in economic analysis to generally maximize welfare (or utility), subject to the requirement that the stock of productive assets (or welfare itself) is nondecreasing in the long term. This assumption is common to most sustainable economic growth models. The essence of the approach is illustrated by the simple example of maximization of the flow of aggregate welfare (W), cumulatively discounted over infinite time (t), as represented by the following expression:

$$\max \int_0^\infty W(C,Z).e^{-rt}\mathrm{d}t.$$

Here, W is a function of C (consumption) and Z (set of other relevant variables), whereas r is the discount rate. Further side constraints may be imposed to satisfy sustainability needs (e.g., nondecreasing stocks of productive assets). One typical application of optimization involves electric generation expansion planning, where analysts seek to minimize the present discounted costs of generation to meet the desired demand forecast over several decades at a given level of reliability.

Some ecological models also optimize variables related to system vigor such as energy use, nutrient flow, and biomass production. In economic models, utility is often measured in terms of net benefits of economic activities, that is, the benefits of development activities minus corresponding costs. More sophisticated economic optimization approaches include environmental and social variables (by attempting to value environmental externalities, system resilience, etc.). However, given the difficulties of quantifying and valuing many such "noneconomic" assets, the costs and benefits associated with market-based activities tend to dominate in most economic optimization models.

Basically, the optimal growth path maximizes economic output, whereas the sustainability requirement is met by ensuring nondecreasing stocks of assets. Some analysts support a "strong sustainability" constraint, which requires separate preservation of each category of critical asset (e.g., manufactured, natural, sociocultural, and human capital), assuming that they are complements rather than substitutes. Other researchers have argued in favor of "weak sustainability," which seeks to maintain the aggregate monetary value of the total stock of assets, assuming that various asset types are substitutes and may be valued.

Side constraints are often necessary because the optimization approach (including economic efficiency and valuation) might not be easily applied to ecological objectives, such as protecting biodiversity and improving resilience, or to social goals, such as promoting equity and empowerment. Such environmental and social variables cannot be easily incorporated within a single valued objective function based on cost–benefit analysis. Thus, techniques such as multicriteria analysis (MCA) may be required to facilitate trade-offs among noncommensurable variables. Moreover, the lagged price system might not anticipate irreversible environmental and social harm and nonlinear system responses that could lead to catastrophic collapse. Therefore, noneconomic indicators of environmental and social status would be helpful. The constraints on critical environmental and social indicators are proxies representing safe thresholds that help to maintain the viability of those systems. Risk and uncertainty will also necessitate the use of decision analysis tools.

2.7 Durability

The second broad integrative approach focuses primarily on sustaining the quality of life, for example, by satisfying environmental, social, and economic sustainability requirements. Such a framework favors "durable" development paths that permit growth but are not necessarily economically optimal. There is more willingness to trade off some economic optimality for the sake of greater safety (i.e., risk aversion) to stay within critical environmental, social, and technical limits. The durability approach would facilitate the design of energy systems that are more resilient and less vulnerable to disruptive shocks.

Economic system durability might require consumption levels to be maintained, that is, per capita consumption that never falls below some minimum level or is nondeclining. The environmental and social sustainability requirements may be expressed in terms of indicators of "state" that monitor the durability or normal functioning of complex ecological, social, and technoeconomic systems. For example, consider an electric power system where durability may be represented by various indexes of system reliability (e.g., loss of load probability [LOLP], frequency and duration of outages [F&D]) and system stability (especially transmission and distribution systems) in relation to external shocks (e.g., storms) that might cause the breakdown of critical system components (e.g., generators, lines).

There is the likelihood of further interaction here due to linkages among the sustainability of social, ecological, and technoeconomic systems; for example, social disruption and conflict could exacerbate damage to both ecological and technoeconomic systems and vice versa. In fact, long-standing social norms in many traditional societies have helped to protect the environment.

Durability encourages a holistic systemic viewpoint, and this is important in sustainomics analysis. The self-organizing and internal structure of complex systems often makes "the whole more durable (and valuable) than the sum of the parts." A narrow definition of efficiency based on marginal analysis of individual components may be misleading. For example, it is more difficult to value the integrated functional diversity in a forest ecosystem than to value the individual species of trees and animals. Therefore, the former is more likely to fall victim to market failure (as an externality). Furthermore, even where correct environmental shadow prices prevail, some analysts point out that cost minimization could lead to homogenization and consequent reductions in system diversity. Systems analysis also helps to identify the benefits of cooperative structures and behavior that a more partial analysis might neglect.

The possibility of many durable paths favors simulation-based methods, including consideration of alternative futures (rather than one optimal result). This approach parallels research on integrating human actors into ecological models, including multiple-agent modeling to account for heterogeneous behavior, bounded rationality leading to different perceptions and biases, and social interactions involving imitation, reciprocity, and comparison.

In the durability approach, maintaining asset stocks enhances system sustainability because various forms of capital are a bulwark that decreases vulnerability to external shocks and reduces irreversible harm rather than merely producing more

economic outputs. System vulnerability, resilience, vigor, organization, and ability to adapt will depend dynamically on the capital endowment as well as on the magnitude and rate of change of a shock; for example, in the case of an electric power system, vulnerability and risk of failure depend on the traditional reserve margin (supply minus demand) as well as on the probability of breakdown of key system components or stochastic surges in demand.

2.8 Indicators

Because asset stocks are important to both the optimal and durable approaches, the practical implementation of sustainomics principles will require the identification of relevant economic, social, and environmental indicators at various levels of aggregation ranging from the global/macro to the local/micro. Indicators must be comprehensive in scope, must be multidimensional in nature (where appropriate), and must account for spatial differences.

Measuring economic, environmental, human, and social capital also raises various problems. Manufactured capital may be estimated using conventional neoclassical economic analysis. Market prices are useful when economic distortions are relatively low, whereas shadow prices could be applied in cases where market prices are unreliable. Natural capital needs to be quantified first in terms of key physical attributes. Typically, damage to natural capital may be assessed by the level of air pollution (e.g., concentrations of suspended particulates, sulfur dioxide, greenhouse gases [GHGs]), water pollution (e.g., biochemical oxygen demand [BOD], chemical oxygen demand [COD]), and land degradation (e.g., soil erosion, deforestation). Then this physical damage could be valued using environmental and resource economics techniques. Human resource stocks are often measured by educational levels, productivity, and earning potential. Social capital is more difficult to assess. Putnam described it as "horizontal associations" among people or social networks and associated behavioral norms and values that affect the productivity of communities. Social capital may be viewed more broadly in terms of social structures that facilitate the activities of agents in society, including both horizontal and vertical associations (e.g., firms). An even wider definition is implied by the institutional approach that includes not only the mainly informal relationships implied by the earlier two views but also more formal frameworks provided by governments, political systems, and legal provisions. Recent work has sought to distinguish between social capital and political capital, that is, the networks of power and influence that link individuals and communities to the higher levels of decision making.

2.9 Complementarity and Convergence of Optimal and Durable Approaches

The two approaches are often complementary in national economic management. For example, economy-wide policies involving both fiscal and monetary measures (e.g., taxes, subsidies, interest, foreign exchange rates) might be optimized using quantitative macroeconomic models. Nevertheless, decision makers inevitably modify these economically "optimal" policies before implementing them to take into account other durable sociopolitical considerations (e.g., poverty alleviation, regional factors) that facilitate governance and stability. The determination of an appropriate target trajectory for future global GHG emissions provides another useful illustration of the interplay between durability and optimality.

In practice, the two approaches point to convergent solutions. First, wastes ought to be generated at rates within the assimilative capacity of the environment. Second, scarce renewable resources should be used at rates compatible with the natural rate of regeneration. Third, nonrenewable resource use rates should depend on the substitutability between these resources and technological progress. Both wastes and natural resource inputs might be minimized by moving from linear throughput to closed loop (or recycling) mode. Finally, inter- and intragenerational equity (especially poverty alleviation), pluralistic and consultative decision making, and enhanced social values and institutions are important additional aspects.

GHG mitigation provides an interesting example of how such an integrative framework could help to incorporate climate change policies within a national sustainable development strategy. The total GHG emissions rate (G) may be decomposed as follows:

$$G = (Q/P) \times (Y/Q) \times (G/Y) \times P,$$

where (Q/P) is quality of life per capita, (Y/Q) is material consumption required per unit of quality of life, (G/Y) is GHG emissions per unit of consumption, and P is population. A high quality of life is consistent with low total GHG emissions, provided that each of the three terms on the right-hand side could be minimized. Reducing (Y/Q) implies "social decoupling" (or "dematerialization"), whereby satisfaction becomes less dependent on material

consumption through changes in tastes, behaviors, and social values. Similarly, (*G*/*Y*) may be reduced by "technological decoupling" (or "decarbonization") that reduces the intensity of GHG emissions in consumption and production. Finally, population growth could be reduced, especially where emissions per capita are already high. The linkages between social and technological decoupling need to be explored. For example, changes in public perceptions and tastes could affect the directions of energy technology and influence the effectiveness of mitigation and adaptation policies.

Climate change researchers are currently exploring the application of large and complex integrated assessment models (IAMs). IAMs include coupled submodels representing various ecological, geophysical, and socioeconomic systems where there is scope to apply the optimality and durability approaches consistently.

2.10 Tools for Sustainable Development Analysis and Assessment

Some important tools for sustainable development analysis and assessment include the Action Impact Matrix (AIM) for prioritizing the economic, environmental, and social interactions of various macroeconomic and sectoral development policies; advanced cost–benefit analysis, including economic valuation of environmental and social impacts; and MCA, especially in cases where some impacts cannot be easily quantified in monetary terms.

Sustainable development assessment (SDA) is another important tool to ensure balanced analysis of both development and sustainability concerns in policies and projects. The economic component of SDA is based on conventional economic and financial analysis (including cost–benefit analysis). The other two key components are environmental assessment and social assessment. SDA also includes poverty assessment. Thus, SDA seeks to integrate and harmonize economic, environmental, and social analyses.

3. SUSTAINABLE ENERGY DEVELOPMENT

3.1 Linkages between Energy Use and Sustainable Development

3.1.1 Energy–Economy Linkages

The economy–energy interaction has been well analyzed in terms of energy as a driving force for economic progress, especially since the industrial revolution and the extensive use of fossil fuels. Modern economies could not function without commercial energy. Figure 2 shows that future consumption of nearly all forms of energy will

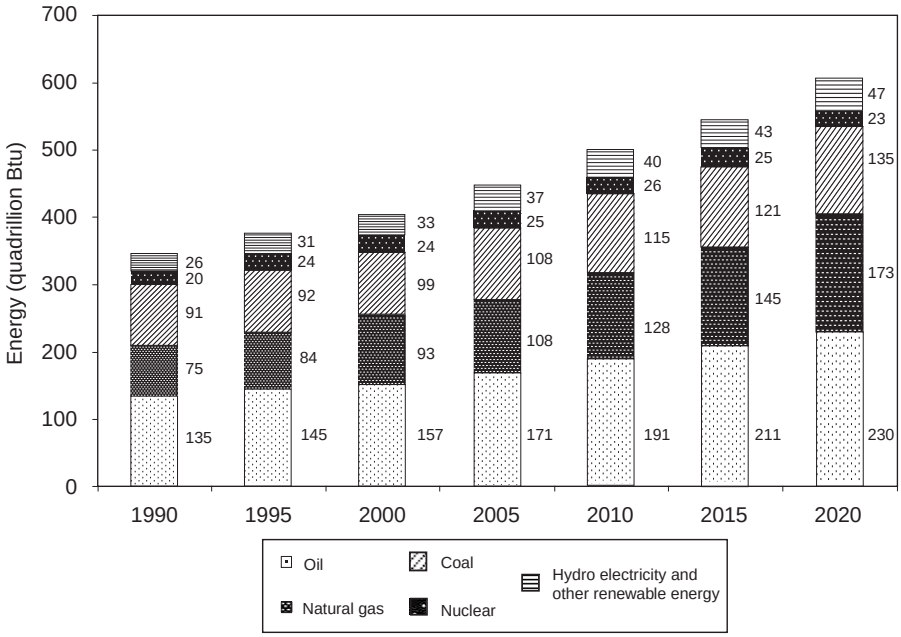


FIGURE 2 World consumption of energy by source, 1990–2020.

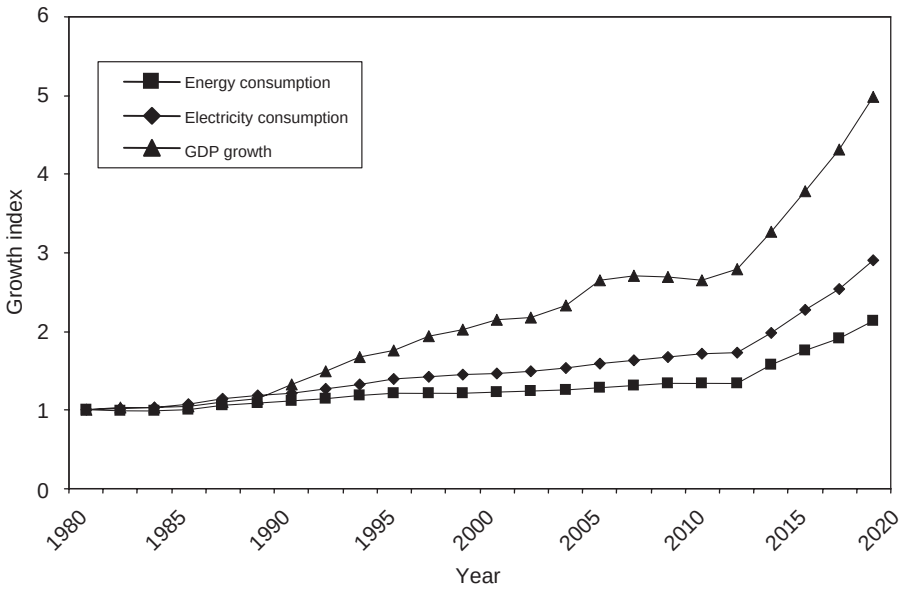


FIGURE 3 Growth of world energy and electricity consumption and gross domestic product (GDP), 1980–2020. 1980 actuals: energy, 285 quadrillion Btu; electricity, 7420 TWh; GDP, $10.8 trillion.

continue to rise, with fossil fuels such as oil, natural gas, and coal playing the dominant role.

The intimate relationship between total energy use and economic output is further clarified in Fig. 3. One hopeful trend is that technological progress and efficiency improvements have reduced the energy intensity of economic production (i.e., lower requirements of physical energy per unit of economic output). Electricity will continue to play an increasingly important role as a safe, clean, and convenient form of energy.

3.1.2 Energy–Environment–Society Linkages

The environmental and social implications of energy use have not been as well analyzed as have energy–economy linkages. Although electricity has relatively few environmental and health consequences at the point of end use, key environmental and social issues arise from power generation. However, the extent and nature of impacts differ among energy sources. Oil- and coal-fired plants not only have national impacts but also have regional and global environmental and health effects. Complete life cycle analyses involving the mining, refining, processing, transport, conversion, and transformation of various fuels such as oil, coal, and nuclear materials all give rise to significant impacts. Even new and renewable energy sources such as wind power, geothermal energy, biomass, and ocean energy, which are perceived to be "clean," have some negative social and environmental impacts.

3.1.3 Transnational Issues

Acid deposition is perhaps the most serious of the transnational issues faced today. Acid deposition is a result of oxides of sulfur and nitrogen that originate from fossil fuel combustion, falling to the ground as particulates and acid rain. Coal- and oil-fired power stations emit significant amounts of sulfur dioxide and nitrogen oxides into the atmosphere. The transport of sulfur dioxide occurs over distances of more than 1000 km, causing the deposition of emission products across national boundaries. Acid depositions caused by sulfur and nitrogen oxides result in damage to trees and crops and sometimes extend to acidification of streams and lakes, resulting in destruction of aquatic ecosystems. They also lead to the corrosion, erosion, and discoloration of buildings, monuments, and bridges. Indirect health effects are caused by the mobilization of heavy metals in acidified water and soil.

Other important transnational issues include environmental and health impacts of radiation due to severe nuclear accidents, oceanic and coastal pollution due to oil spills, downstream siltation of river water in one nation due to deforestation of water sheds and soil erosion in a neighboring country, and changes in hydrological flow and water conditions caused by dams.

3.1.4 Global Issues

The Intergovernmental Panel on Climate Change (IPCC) has identified energy use as the major

contributor to anthropogenic GHG emissions, mainly CO_2 and other gases such as N_2O, CH_4, and chlorofluorocarbons (CFCs) that will lead to climate change and undermine sustainable development prospects. First, global warming poses a significant potential threat to the future economic well-being of large numbers of humans. Second, climate change could also undermine social welfare and equity in an unprecedented manner. Third, from an environmental viewpoint, increasing anthropogenic emissions and accumulations of GHGs will significantly perturb a major global subsystem—the atmosphere. Climate change will also threaten the stability of a range of critical interlinked physical, ecological, and social systems and subsystems.

3.2 Framework for Sustainable Energy Development at the National Level

Sustainable development is the broad rationale underlying most national-level planning and policymaking. Power and energy planning must also be part of, and closely integrated with, overall sustainable development strategies to meet many specific, interrelated, and frequently conflicting national objectives. Some specific goals for sustainable energy development might include ensuring economic efficiency in energy supply and use to maximize growth, including energy efficiency-related objectives such as energy conservation; raising sufficient revenues from energy sales to finance sector development; addressing socioeconomic concerns such as meeting basic energy needs of the poor and developing special regions (particularly rural or remote areas) and priority sectors of the economy; preserving the environment; diversifying supply, reducing dependence on foreign sources, saving scarce foreign exchange, and meeting national security requirements; ensuring price stability; and so on.

3.2.1 Integrated Approach

Successful planning and implementation of national energy programs must explicitly link the energy sector to sustainable development of other parts of the economy. An integrated approach will help decision makers to formulate policies and will provide market signals and information to economic agents that encourage more efficient and sustainable energy production and use. Figure 4 depicts such an approach to decision making, within a hierarchical conceptual framework for sustainable energy development, to be implemented through a set of energy supply and demand management policies. The middle column shows the core—an integrated, multilevel analytical framework. The top level of sustainable energy development recognizes transnational linkages. Thus, individual countries are embedded in an international matrix, and global economic, social, and environmental conditions will impose exogenous inputs or constraints on national decision makers.

The second hierarchical level in Fig. 4 focuses on the multisectoral national economy of which the energy sector is a part. Thus, energy planning requires analysis of links between the energy sector and other sectors, including energy needs of user sectors (e.g., industry, transport, agriculture), input requirements of the energy sector, and impacts of energy supply and pricing policies.

The next level of sustainable energy development disaggregates the energy sector into subsectors such as electricity and petroleum products. This permits detailed analysis, with special emphasis on interactions among various energy subsectors, substitution possibilities, and resolution of policy conflicts.

The lowest hierarchical level pertains to energy analysis within each energy subsector, where line institutions (both public and private) carry out detailed energy resource evaluation, planning, and implementation of projects (including sustainability assessments).

In practice, the various levels of sustainable energy development merge and overlap considerably, requiring careful study of (inter)sectoral linkages. Energy–social–environmental interactions (represented by the vertical bar in Fig. 4) cut across all levels and provide important paths for incorporating environmental and social considerations into national energy policies.

Sustainable energy development facilitates policymaking and does not imply rigid centralized planning. The process results in the development of a flexible and constantly updated sustainable energy strategy designed to meet national goals. This strategy (of which the investment program and pricing policy are important elements) may be implemented through energy supply and demand management policies and programs that make effective use of decentralized market forces and incentives.

In particular, SED implies improvements in overall economic efficiency through better energy management. Figure 4 shows various policy instruments available to decision makers for implementing sound energy management. While formulating policy, one

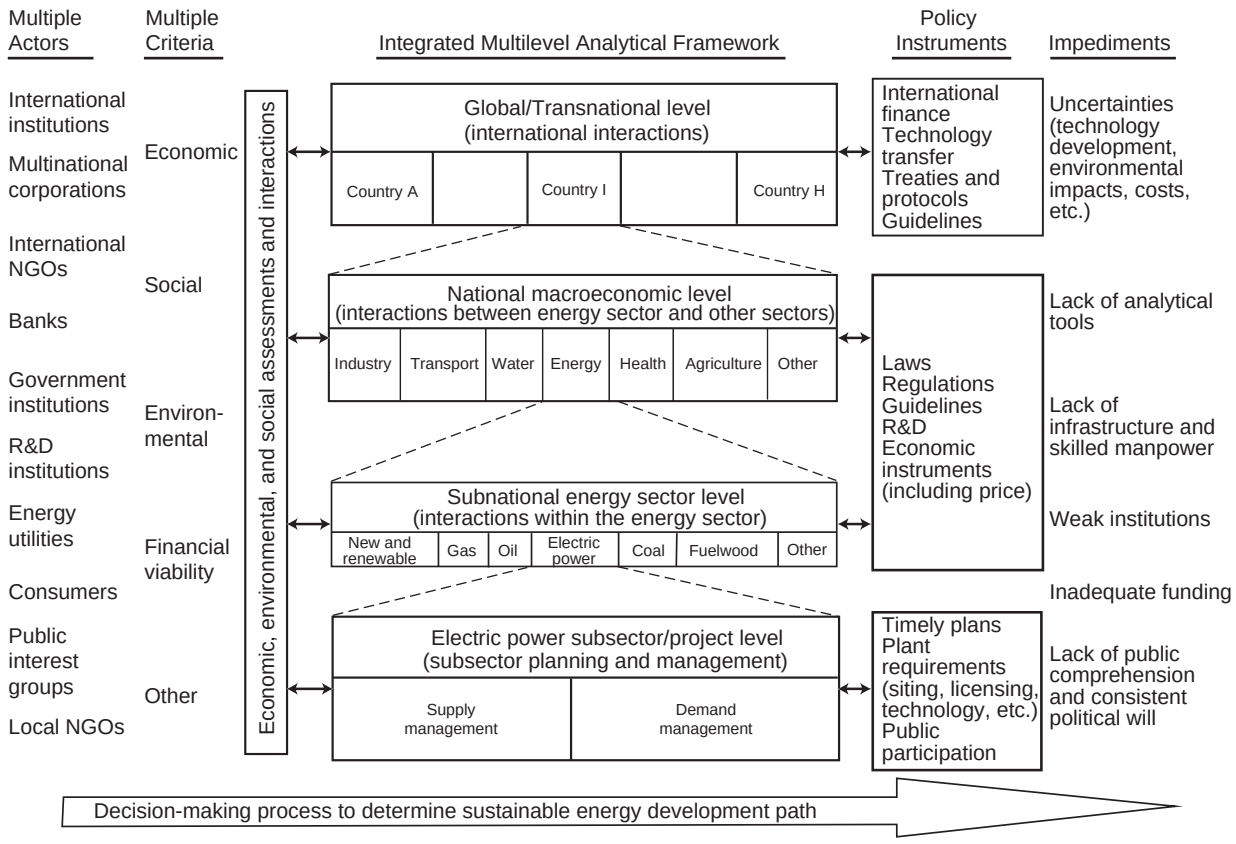


FIGURE 4 Framework for sustainable energy development. NGOs, nongovernmental organizations; R&D, research and development.

must consider the interests of multiple actors, ranging from international institutions to local energy users. The figure also indicates the most important impediments that limit the effectiveness of policies.

Investments offer a good opportunity to pursue sustainable energy development. In 10 years or so, new plants will account for more than half of the industrial output of developing countries, and in 20 years or so, new plants will account for practically all of it. Therefore, it will be possible to have a major impact by putting in place policies, legislation, mechanisms, systems, and incentives that facilitate sustainable energy development.

3.2.2 Identifying Sustainable Energy Options: "Win–Win" Options versus Trade-Offs

To identify sustainable energy options, policymakers need to consider the economic, social, and environmental aspects of sustainable development. Options that lead to improvements in all three indexes are referred to as "win–win" options. Once win–win options are realized, policymakers are able to make trade-offs among other available options.

Incorporating environmental and social externalities into energy decision making is particularly important where concerns (e.g., pollution from nuclear or fossil-fueled plants, inundation at hydro plants) have hampered project implementation. Environmental and social concerns need to be addressed early, that is, at the sectoral and regional planning stages rather than at the final stage of project SDA. Unfortunately, when one deals with energy sector issues at this aggregate planning level, the application of many project-level valuation techniques becomes extremely difficult. First, the impacts are difficult to value (e.g., health effects of pollutants from coal-fired generating stations, biodiversity loss from large-scale hydro storage, impacts of GHG emissions). Doubts raised about the valuation techniques themselves divert attention away from critical policy trade-offs. Second, many techniques appropriate at the micro level are less effective at the sector level. Thus, contingent valuation is more valid where respondents can be asked specific

questions about local impacts of a particular project to which they can relate, and it is difficult to apply at the sector level where one deals with large numbers of technology, site, and mitigation options.

In countries where inappropriate policies have encouraged wasteful and unproductive uses of some forms of energy, better energy management could lead to improvements in economic efficiency (higher value of net output produced), energy efficiency (higher value of net output per unit of energy used), energy conservation (reduced absolute amount of energy used), and environmental and social protection (reduced energy-related environmental and social costs). However, it might not be possible to satisfy all of these goals simultaneously. For example, in some developing countries where existing levels of per capita income and energy consumption are very low, affordable energy might have a high priority to meet basic energy needs.

The economic efficiency criterion, which maximizes the value of net output from all scarce resources in the economy (including energy), normally subsumes purely energy-oriented objectives such as energy efficiency and conservation. Furthermore, costs arising from energy-related adverse environmental impacts may be included in the energy economics analytical framework to determine how many other benefits society should be willing to forgo so as to avoid environmental damage.

Energy use and production in developing countries can be improved in several ways to make them more sustainable. First, energy efficiency may be increased by supply- and demand-side improvements. Second, environmentally and socially more benign technologies can be introduced, including fuel switching and renewable energy sources. Finally, price, institutional, and regulatory reforms could contribute to sustainable energy development.

3.2.3 Sustainable Energy Development Options Matrix

Table I shows typical impacts of selected energy options on the three elements of sustainable development. Although efficient supply-side options have clear economic gains in terms of savings in capital investments and environmental benefits from reductions in GHG emissions that result from decreased energy supply, they do not have obvious social benefits. Efficient end use options, as shown in the case of an efficient fuelwood stove, have benefits relating to all three elements. Although advanced technologies, such as clean coal combustion technologies, help to reduce air pollutants (e.g., CO_2, NO_x) that cause respiratory diseases and reduce productivity, many developing countries cannot afford such high-cost technologies. Likewise, renewable energy sources also provide environmental and social benefits by reducing a country's dependence on traditional fossil fuels. However, in terms of power-generating costs, renewables are often more expensive than fossil fuels.

TABLE I

Selected Sustainable Energy Development Options Matrix

	Impact		
Option	Economic	Environmental	Social
Supply efficiency	+	+	
End use efficiency	+	+	+
Advanced technologies	−	+	+
Renewables	−	+	+
Pricing policy	+	+	+/−
Privatization/decentralization	+	+/−	+/−

4. APPLYING THE SUSTAINABLE ENERGY DEVELOPMENT FRAMEWORK

This section presents practical case studies that illustrate the application of the ideas presented earlier.

4.1 Global Transnational Scale: Interplay of Optimality and Durability in Determining Global GHG Emission Target Levels

Optimization- and durability-based approaches can facilitate the determination of target GHG emission levels. Under an economic optimizing framework, the ideal solution would be, first, to estimate the long-run marginal abatement costs (MACs) and the marginal avoided damages (MADs) associated with various GHG emission profiles (see Fig. 5C where the error bars on the curves indicate measurement uncertainties). The optimal emission levels would be determined at the point where future benefits (in terms of climate change damage avoided by reducing one unit of GHG emissions) are just equal to the corresponding costs (of mitigation measures required to reduce that unit of GHG emissions), that is,

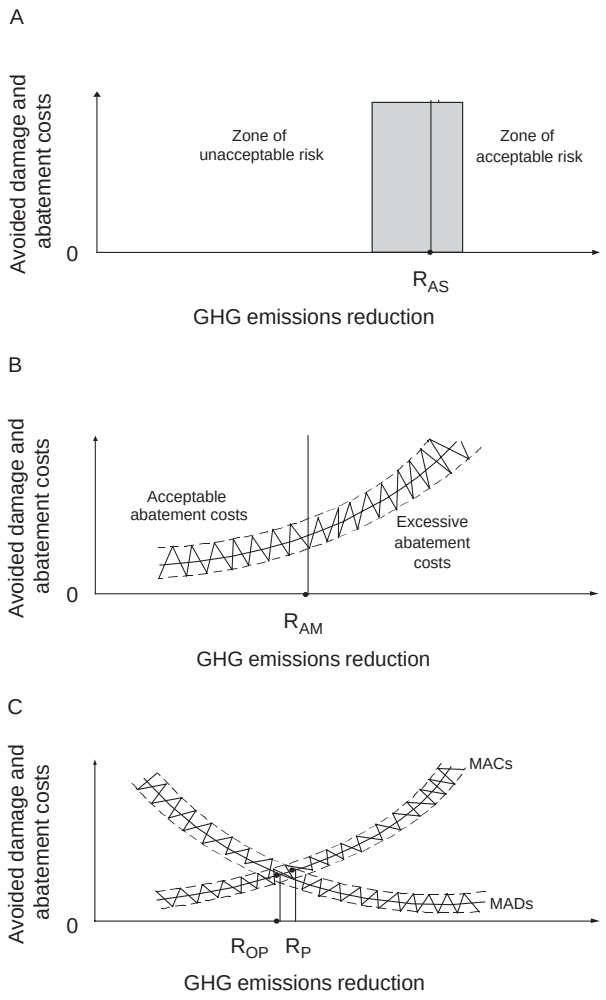


FIGURE 5 Determining abatement targets: (A) absolute standard; (B) affordable safe minimum standard; (C) cost–benefit optimum. Adapted from Intergovernmental Panel on Climate Change (1996, Fig. 5.10).

MACs = MADs at point R_{OP}. Durable strategies become more relevant when we recognize that MACs and/or MADs might be poorly quantified and uncertain. Figure 5B assumes that MACs are better defined than MADs. First, MACs are determined using technoeconomic least cost analysis, an optimizing approach. Next, the target emissions are set on the basis of the affordable safe minimum standard (at R_{AM}), which is the upper limit on costs that will still avoid unacceptable socioeconomic disruption. This is closer to the durability approach.

Finally, Fig. 5A indicates an even more uncertain world where neither MACs nor MADs are defined. Here, the emission target is established on the basis of an absolute standard (R_{AS}) or safe limit that would avoid an unacceptably high risk of damage to ecological (and/or social) systems. This last approach would be more in line with the durability concept.

4.2 National Economy-wide Scale: Using the AIM for Macro Policy Analysis

National policymakers routinely make key macro-level decisions that could have (often inadvertent) environmental and social impacts that are far more significant than the effects of local economic activities. At this macroeconomic level, recent work has focused on incorporating environmental considerations such as depletion of natural resources and pollution damage into the system of national accounts via useful new indicators and measures such as the system of environmentally adjusted environmental accounts (SEEAs), green gross national product (GNP), and genuine savings. In this context, the AIM approach is a sustainomics tool that helps to identify and analyze economic–environmental–social interactions and formulate effective sustainable development policies.

Table II shows a simplified AIM, although an actual one would be larger and more detailed. The AIM helps to promote an integrated view, meshing development decisions with priority economic, environmental, and social impacts. The left-most column lists examples of key development interventions (both policies and projects), whereas the top row indicates some typical sustainable development impacts/issues. Thus, the elements or cells in the matrix help to accomplish the following:

- Explicitly identify key linkages
- Focus attention on methods of analyzing important impacts
- Suggest action priorities and remedies

At the same time, the organization of the overall matrix facilitates the tracing of impacts as well as the coherent articulation of the links among a range of development actions, both policies and projects.

4.2.1 Screening and Problem Identification

One of the early objectives of the AIM–based process is to help in screening and problem identification by preparing a preliminary matrix that identifies broad relationships and provides a qualitative idea of the magnitudes of impacts. Thus, the preliminary AIM helps to prioritize the most important links between policies and their sustainability impacts.

One typical primary energy policy shown in the third row of Table II involves increasing energy

TABLE II

A Simplified Preliminary Action Impact Matrix

		Impacts on key sustainable development issues			
Activity or policy	Main objective	Land degradation and biodiversity loss	Air pollution GHG emissions	Resettlement and social effects	Other
Macroeconomic and sectoral policies	Macroeconomic and sectoral improvements	Positive impacts due to removal of distortions Negative impacts mainly due to remaining constraints			
Exchange rate	Improve trade balance and economic growth	(−H) (deforest open-access areas)			
Energy pricing	Improve energy use economic and efficiency		(+M) (energy efficiency)		
Others					
Complementary measures	Socioeconomic and environmental gains	Enhance positive impacts and mitigate negative impacts (above) of broader macroeconomic and sectoral policies			
Market based			(+M) (pollution tax)		
Non-market based		(+H) (property rights)	(+M) (public sector accountability)		
Investment projects	Improve effectiveness of investments	Investment decisions made more consistent with broader policy and institutional framework			
Project 1 (hydro dam)		(−H) (inundate forests)	(+M) (displace fossil fuel use)	(−M) (displace people)	
Project 2 (Reafforest and relocate)		(+H) (replant forests)		(+M) (relocate people)	
Project N					

Note. Several typical policies and projects, as well as key economic, environmental and social issues, are shown. Some illustrative impact assessments are also indicated, where + and − signify beneficial and harmful impacts, and where H and M indicate high and moderate intensity, respectively.

prices closer to marginal costs to improve energy efficiency while decreasing air pollution and GHG emissions. A complementary environmental policy measure indicated in the fourth column consists of adding pollution taxes to marginal energy costs, a measure that will further reduce air pollution and GHG emissions. Increasing public sector accountability constitutes another complementary policy that will reinforce favorable responses to these price incentives by reducing the ability of inefficient firms to pass on cost increases to consumers or to transfer their losses to the government.

Near the bottom of Table II, a major hydroelectric project is shown as having two adverse impacts (inundation of forested areas and village dwellings) as well as one positive impact (replacement of thermal power generation, thereby reducing air pollution and GHG emissions). A reafforestation project coupled with resettlement schemes helps to address the negative impacts.

Therefore, this matrix–based approach encourages the systematic articulation and coordination of policies and projects to make development more sustainable.

4.2.2 Analysis and Remediation

This process may be developed further to assist in analysis and remediation. For example, more detailed analyses and modeling may be carried out for those matrix elements in the preliminary AIM that had already been identified as representing high-priority linkages between economy-wide policies and economic, environmental, and social impacts. This, in turn, would lead to a more refined and updated

AIM that would help to quantify impacts and formulate additional policy measures to enhance positive linkages and mitigate negative ones. The types of more detailed analyses that could help to determine the final matrix would depend on planning goals and available data and resources. They may range from fairly simple methods to rather sophisticated economic, ecological, and social models, included in the sustainomics toolkit.

4.3 Subnational Scale: Energy Sector Planning

The incorporation of environmental and social externalities into decision making is particularly important in the electric power sector. Meier and Munasinghe, in their Sri Lanka study, demonstrated how externalities could be incorporated into power system planning in a systematic manner. Sri Lanka currently depends largely on hydro power for electricity generation, but over the next decade or so the main choices seem to be large coal- or oil-fired stations or hydro plants whose economic returns and environmental impacts are increasingly unfavorable. In addition, a wide range of other options (e.g., wind power, increasing use of demand-side management, system efficiency improvements) complicates decision making, even in the absence of environmental concerns.

The methodology involves the following steps: (1) definition of generation options and their analysis using sophisticated least cost system planning models; (2) selection and definition of attributes that reflect planning objectives; (3) explicit economic valuation of those impacts for which valuation techniques can be applied with confidence, with the resultant values then being added to the system costs (the main economic attribute); (4) quantification of those attributes for which explicit economic valuation is inappropriate but for which suitable quantitative impact scales can be defined; (5) translation of attribute value levels into value functions (known as "scaling"); (6) display of trade-offs to facilitate decision making; and (7) definition of options for further study, also involving discarding eliminating inferior options.

4.3.1 Main Results

The main set of sectoral policy options examined included (1) variations in the currently available mix of hydro and thermal (coal and oil) plants, (2) demand-side management (e.g., compact fluorescent lighting), (3) renewable energy options (e.g., wind generation), (4) improvements in system efficiency (using more ambitious targets for transmission and distribution losses than the base case assumption of 12% by 1997), (5) clean coal technology (e.g., pressurized fluidized bed combustion [PFBC] in a combined cycle mode), and (6) pollution control technology options (e.g., various fuel switching and pollution control options such as importing low-sulfur oil for diesels and fitting coal power plants with flue gas desulfurization [FGD] systems).

A limited number of criteria or attributes should be selected with care to reflect issues of national as well as local project level significance. CO_2 emissions were used as a proxy for the potential impact on global warming. Health impacts were measured through population-weighted increments in both fine particulates and NO_x. To capture the potential biodiversity impacts, a probabilistic index was derived. Employment creation was used as an illustrative social impact.

Figure 6A illustrates a typical trade-off curve for biodiversity. The "best" solutions lie closest to the origin. The so-called trade-off curve is defined by the set of "noninferior" solutions, representing options that are superior regardless of the weights assigned to the various objectives. For example, on this curve, the option defined as "no hydro" is better than the "wind" option in terms of both economic cost and biodiversity loss.

Although most options in Fig. 6A have index values that fall between 50 and 100, the no-hydro option has a zero value because thermal projects that constitute this option lie at sites of low biodiversity value (either close to load centers or on the coast). Meanwhile, wind plants would require large land areas, and their biodiversity indexes are higher. However, vegetation on the south coast (where wind power plants would be located) has relatively low biodiversity value; therefore, the overall biodiversity impact of this option is small. In summary, the best options (on the trade-off curve) include the no-hydro and run-of-river hydro options that require essentially zero inundation. The extreme outlier at the top right-hand corner is the Kukule hydro dam; it has a biodiversity loss index ($B = 530$) that is much larger than those for other options ($B\text{s} = 50$–70).

The trade-off curve between health impacts and average incremental cost is different (Fig. 6B). The "iresid" point on this trade-off curve (low-sulfur imported fuel oil at diesel plants) is better than the use of flue gas desulfurization systems ("FGD" point) in terms of both economic cost and environment.

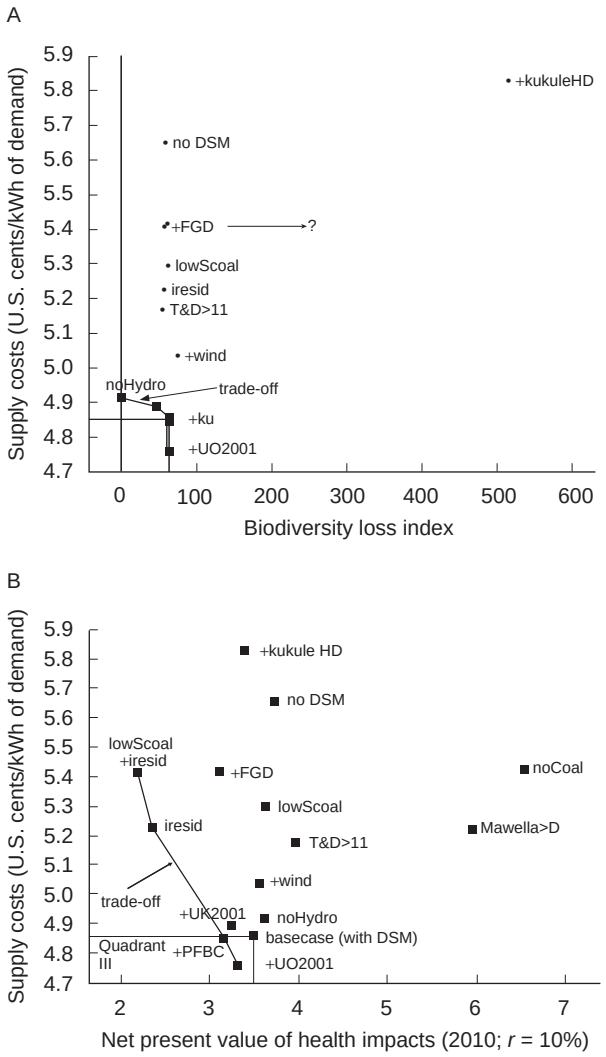


FIGURE 6 Trade-off curves between economic costs and (A) biodiversity impacts and (B) health impacts. Reproduced from Meier and Munasinghe (1994).

4.3.2 Conclusions

There are several useful conclusions. First, the results indicate that those impacts for which valuation techniques are relatively straightforward and well established (e.g., the opportunity costs of lost production from inundated land, benefits of establishing fisheries in reservoirs) are small compared with overall system costs, and so including them in the benefit–cost analysis does not materially change results. Second, even in cases where explicit valuation is difficult (e.g., mortality and morbidity effects of air pollution), implicit valuation based on analysis of trade-off curves can provide important guidance to decision makers. Third, certain options were clearly inferior or superior to other options when one examines all impacts simultaneously. For example, the high dam version of the Kukule hydro project can be safely excluded from further consideration here due to poor performance on all attribute scales. Fourth, it is possible to derive attribute scales that provide useful proxies for impacts that are difficult to value. For example, the population-weighted, incremental, ambient air pollution level was the proxy for health impacts, yielding several important conclusions independent of any economic values assigned to health effects.

Finally, with respect to the practical planning, the study identified several priority recommendations, including the need to reconsider (1) demand-side management options, especially fluorescent lighting; (2) whether the current transmission and distribution loss reduction target of 12% ought to be further reduced; (3) possibilities of PFBC technology for coal power; (4) replacement of some coal-fired power plants (on the south coast) by diesel units; and (5) cooling system options for coal plants.

4.4 Local Project Scale: Multicriteria Analysis of Small Hydroelectric Power Projects

Well-accepted environmental and social assessment procedures at the project/local level may be readily adapted to assess environmental and social effects of micro-level activities. When monetary valuation of environmental and social effects is not feasible, MCA may be used. This subsection summarizes how Morimoto and colleagues have used MCA to compare hydroelectric power schemes. The three main sustainable development issues considered are the economic costs of power generation, the ecological costs of biodiversity loss, and the social costs of resettlement.

The principal objective is to generate additional kilowatt-hours of electricity to meet growing power demand in Sri Lanka. Assuming that the benefits from each additional kilowatt-hour are the same, the analysis seeks to minimize economic, social, and environmental costs of generating one unit of electricity from various hydropower sites. Following the MCA approach, environmental and social impacts are measured in different (nonmonetary) units instead of attempting to economically value and incorporate them within the monetary-valued cost–benefit analysis framework.

4.4.1 Environmental, Social, and Economic Indicators

Sri Lanka has many varieties of endemic or endangered fauna and flora. Often, large hydro projects destroy wildlife at dam sites and in

downstream areas. Hence, a biodiversity loss index was estimated for each hydroelectric site as the main ecological indicator.

Although dam sites are usually in less densely populated rural areas, resettlement is still a serious problem. In general, people are relocated from the wet zone to the dry zone where soils are less rich; therefore, the same level of agricultural productivity cannot be maintained. In the wet zone, multiple crops, including paddy, tobacco, coconuts, mangoes, onions, and chillies, can be grown. However, these crops cannot be cultivated as successfully in the dry zone due to limited water and poor soil quality. Living standards often become worse, and several problems (e.g., malnutrition) can occur. Moreover, other social issues might arise, including erosion of community cohesion and psychological distress due to changed living conditions. Hence, minimizing the number of people resettled due to dam construction is an important social objective.

The project costs are available for each site, from which the critical economic indicator (average cost per kilowatt-hour/year) may be estimated. The annual energy generation potential at various sites ranges from approximately 11 to 210 kWh (Fig. 7). All three variables (biodiversity index, number of people resettled, and generation costs) are weighted by the inverse of the amount of electrical energy generated. This scaling removes the influence of project size and makes them more comparable.

4.4.2 Basic Results

A simple statistical analysis shows that pairwise, there is a little correlation among the quantity of electricity generated, average generation cost, number of people resettled, and biodiversity index.

Figure 7 shows that on a per kilowatt-hour/year basis, the project named AGRA003 has the highest biodiversity index, HEEN009 has the highest

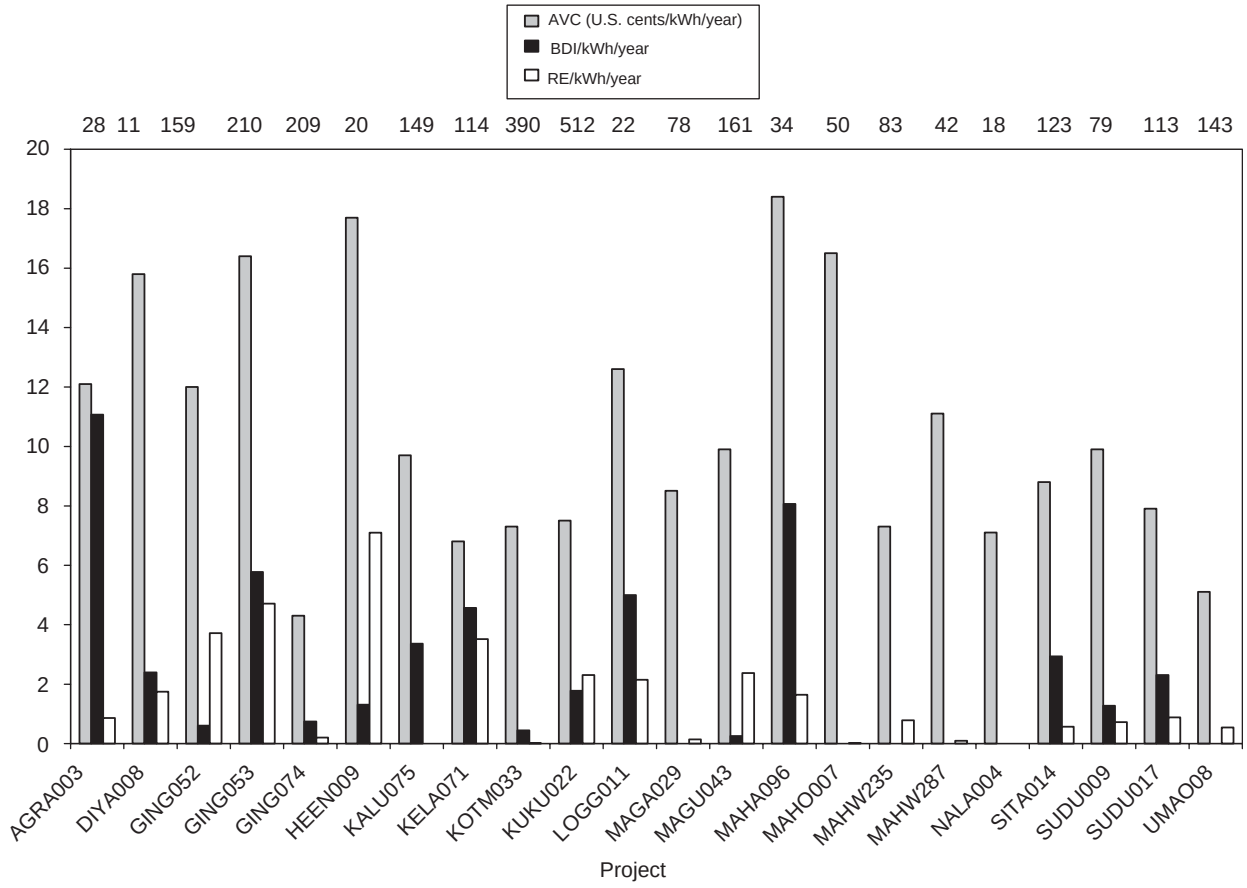


FIGURE 7 Average generation costs (AVC), biodiversity index (BDI), and number of resettled people (RE) by hydroelectric project. All indexes are per kilowatt-hour/year. Numbers of people resettled and the biodiversity index are scaled for convenience (by the multipliers 10^{-5} and 10^{-9}, respectively). The values across the top of the graph indicate the annual energy generation in gigawatt-hours. Adapted from Morimotor, Munasinghe, and Meier (2000).

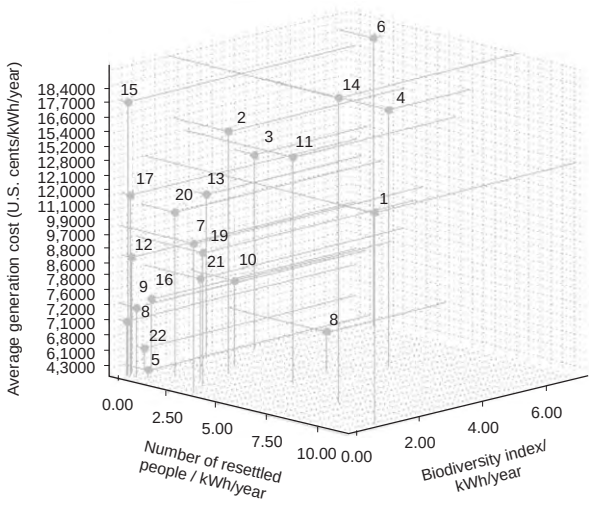


FIGURE 8 Three-dimensional MCA of sustainable development indicators for various hydropower options. Reproduced from Morimoto, Munasinghe, and Meier (2000).

number of resettled people, and MAHA096 has the highest average generation cost. Further important comparisons may be made. For example, KALU075 is a relatively large project whose costs are low, whereas MAHA096 is a smaller scheme with much higher costs with respect to all three indexes. Another observation is that a project such as KELA071 fully dominates GING053 given that the former is superior in terms of all three indicators. Similar comparisons may be made among other projects.

Figure 8 provides a three-dimensional analysis of sustainable development indicators for these hydropower sites where the respective axes represent economic, ecological, and social objectives. The closer to the origin any given coordinate point lies, the better the corresponding project is in terms of achieving these three objectives. This type of analysis gives policymakers some idea about which project is more favorable from a sustainable energy development perspective.

Suppose that all three objectives are arbitrarily given equal weight. Then, each project may be ranked according to its absolute distance from the origin. For example, rank 1 is given to the one closest to the origin, rank 2 is the second closest, and so on (Fig. 8). On this overall basis, from a sustainable energy development perspective, the most favorable project is GING074 (rank 1), whereas the least favorable is MAHA096 (rank 22).

4.4.3 Conclusions

The strength of this type of analysis is in helping policymakers to compare project alternatives more easily and effectively. The simple graphical presentations are readily comprehensible and clearly identify sustainable development characteristics of each scheme. The multidimensional analysis supplements more conventional cost–benefit analysis (based on economic analysis alone). Because each project has different features, assessing them by looking at only one aspect (e.g., generation costs, effects on biodiversity, impacts on resettlement) could be misleading.

There are some weaknesses in the MCA approach used here. First, for simplicity, each major objective is represented by only one variable, assuming that all of the other attributes are minor. In reality, there may be several additional variables that could describe the economic, social, and environmental aspects of sustainable development. Further analysis that includes other attributes might provide new insights. A second extension of this study is to include other renewable sources of energy in the analysis. Finally, more sophisticated three-dimensional graphic techniques may yield better and clearer representations.

SEE ALSO THE FOLLOWING ARTICLES

Cultural Evolution and Energy • Development and Energy, Overview • Economic Growth and Energy • Economic Growth, Liberalization, and the Environment • Energy Ladder in Developing Nations • Green Accounting and Energy • Oil-Led Development: Social, Political, and Economic Consequences • Population Growth and Energy • Potential for Energy Efficiency: Developing Nations • Women and Energy: Issues in Developing Nations

Further Reading

Arrow, K. J., Cline, W., Maler, K. G., Munasinghe, M., and Stiglitz, J. (1995). Intertemporal equity, discounting, and economic efficiency. *In* "Global Climate Change: Economic and Policy Issues" (M. Munasinghe, Ed.). World Bank, Washington, DC.

Atkinson, G., Dubourg, R., Hamilton, K., Munasinghe, M., Pearce, D. W., and Young, C. (1997). "Measuring Sustainable Development: Macroeconomics and the Environment." Edward Elgar, Cheltenham, UK.

Commission on Sustainable Development. (1998). "Indicators of Sustainable Development." CSD, New York.

Faucheux, S., Pearce, D., and Proops, J. (eds.). (1996). "Models of Sustainable Development." Edward Elgar, Cheltenham, UK.

Gunderson, L., and Holling, C. S. (2001). "Panarchy: Understanding Transformations in Human and Natural Systems." Island Press, New York.

IPCC. (1996). "Climate Change 1995: Economic and Social Dimensions," chap. 5. Cambridge University Press, London, UK.

Meier, P., and Munasinghe, M. (1994). "Incorporating Environmental Concerns into Power Sector Decision Making." World Bank, Washington, DC.

Morimoto, R., Munasinghe, M., and Meier, P. M. (2000). "Sustainable energy development: Assessing the Economic, Social, and Environmental Implications of Hydropower Projects in Sri Lanka. MIND Research–Discussion Paper NO.2." Munasinghe Institute for Development, Colombo, Sri Lanka.

Munasinghe, M. (1992). "Environmental economics and sustainable development. Paper presented at the UN Earth Summit." Rio de Janeiro, Brazil, and reprinted by World Bank, Washington, DC.

Munasinghe, M. (2002). The sustainomics trans-disciplinary meta-framework for making development more sustainable: Applications to Energy Issues. *Intl. J. Sustainable Dev.* **4**(2), 6–54.

Munasinghe, M., and Cruz, W. (1994). "Economywide Policies and the Environment." World Bank, Washington, DC.

Munasinghe, M., Sunkel, O., and de Miguel, C. (Eds). (2001). "The Sustainability of Long-Term Growth." Edward Elgar, London, UK.

North, D. (1990). "Institutions, Institutional Change, and Economic Performance." Cambridge University Press, Cambridge, UK.

Putnam, R. D. (1993). "Making Democracy Work: Civic Traditions in Modern Italy." Princeton University Press, Princeton, NJ.

Rawls, J. A. (1971). "Theory of Justice." Harvard University Press, Cambridge, MA.

Sen, A. K. (1984). "Resources, Values, and Development." Blackwell, Oxford, UK.

United Nations. (1996). "Indicators of Sustainable Development: Framework and Methodology." UN, New York.

UN Development Program. (1998). "Human Development Report." UN, New York.

World Commission on Environment and Development. (1987). "Our Common Future." Oxford University Press, Oxford, UK.

World Bank. (1998). "Environmental Assessment Operational Directive" (EAOD4.01). World Bank, Washington, DC.

System Dynamics and the Energy Industry

ANDREW FORD
Washington State University
Pullman, Washington, United States

1. System Dynamics
2. Software
3. Information Feedback
4. Models for National Energy Policy
5. Incentives for Electric Vehicles
6. Power Plant Construction in the West
7. Summary

Glossary

feebate Incentive program with fees on dirty cars and rebates for clean cars.
negative feedback A closed chain of cause and effect that acts to reduce change in a system.
positive feedback A closed chain of cause and effect that acts to magnify change in a system.
Stella Computer software for stock and flow modeling.
Vensim Computer software for stock and flow modeling.
widget A hypothetical product used in classroom examples.

System dynamics is a computer-based method for studying dynamic behavior. The method is normally applied when the dynamic behavior is viewed as a problem. The problematical behavior might take the form of highly volatile oscillations in prices, for example. System dynamics can help the analyst understand how the elements of the system interact to create the oscillations. The analyst would construct a mathematical model to represent the elements and their interactions. The model is then simulated on the computer to determine if simulated oscillations are similar to the oscillations observed in the industry. The analyst would then use the model to experiment with changes in the system that could lead to improved behavior. System dynamics has been applied to a wide variety of energy systems. The applications have dealt with challenges faced by utility managers, state regulators, and national policy makers. This article presents a description of system dynamics followed by summaries of three applications to energy systems.

1. SYSTEM DYNAMICS

The system dynamics approach was pioneered by Jay W. Forrester and colleagues at the Sloan School of Management at the Massachusetts Institute of Technology during the 1960s. In 1977, Coyle classified system dynamics as "that branch of control theory which deals with socioeconomic systems and that branch of management science which deals with problems of controllability." The method was well described by Forrester in 1961. In 2000, Sterman published the most comprehensive text.

The basic building blocks of system dynamics models are stocks and flows. Stocks represent where accumulation or storage takes place in the system. Flows represent the actions that directly influence the stocks. Figure 1 shows an example of a single stock to represent the volume of water stored in a reservoir. This stock is increased by the upstream flow and lowered by the flow of water released at the dam.

Figure 2 shows a combination of six stocks and six flows. The stocks represent the storage of petroleum products in the world oil industry. The flows show the movement of products from the oil fields of Kuwait to the fuel tanks in our personal vehicles.

Stocks and flows provide the foundation for system dynamics models. The next step is to add the variables needed to explain each flow. Figure 3 illustrates with a model of a sales company. A single stock is used to represent the number of salespersons, whereas flows are used to keep track of new hires

 © 2004 Elsevier Inc. All rights reserved.

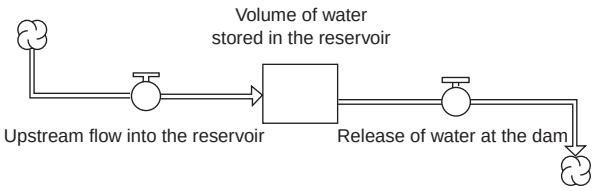


FIGURE 1 Example of a stock and flows. Adapted from A. Ford, "Modeling the Environment: An Introduction to System Dynamics Modeling of Environmental Systems." Copyright 1999 by Island Press. Reprinted by permission of Island Press.

and departures. The remaining variables are used to explain the flows. In this example, an exit rate is used to explain the departures, whereas a hiring fraction helps explain the new hires. New hires depend on the budgeted size of the sales force, which depends on the sales department budget and the average salesperson's salary. The sales department budget depends on the annual revenues and the fraction devoted to sales. Annual revenues depend on widgets sold and the widget price. Widget sales depends on the size of the sales force and the effectiveness of each salesperson.

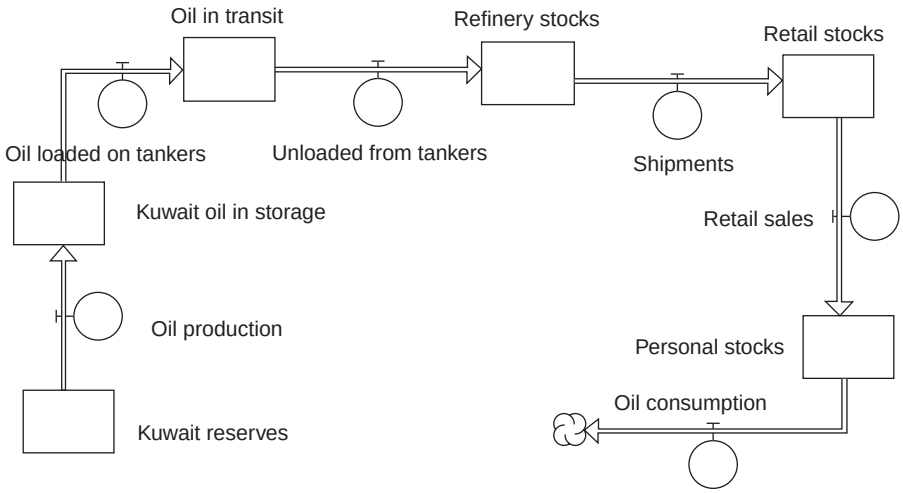


FIGURE 2 Example of stocks and flows in the petroleum industry. Adapted from Ford (1999). Reprinted by permission of Island Press.

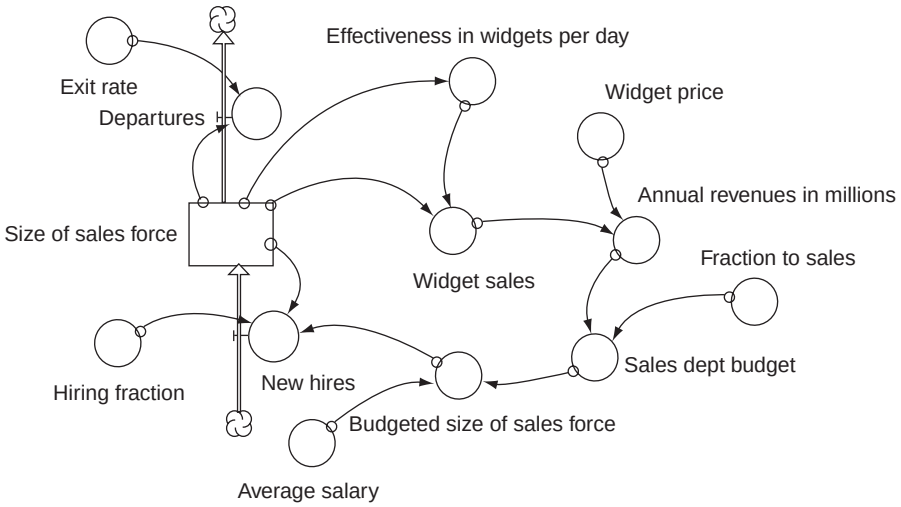


FIGURE 3 A system dynamics model of a hypothetical sales company. Adapted from Ford (1999). Reprinted by permission of Island Press.

In this classroom example, effectiveness is assumed to decline when there is a larger sales force.

2. SOFTWARE

The images in Figs. 1–3 were created with Stella software by High Performance Systems. Stella is one of several icon-based programs to aid the construction and simulation of system dynamics models. Stella diagrams are shown in many of the figures in this article, but equally useful models may be constructed with any of the stock-and-flow programs, such as Vensim or Powersim. The programs use somewhat different symbols, as shown by the Vensim diagram of the sales model in Fig. 4. Although the symbols may differ, the programs are fundamentally the same. They are all designed to facilitate the construction of stock-and-flow models and to show dynamic behavior through numerical simulation.

3. INFORMATION FEEDBACK

The sales company will grow rapidly if the personnel are highly effective at selling widgets. The rapid growth is made possible by positive feedback, a closed chain of cause and effect that acts to magnify change in a system. In this example, an increase in the sales force can lead to higher sales, higher revenues, increased budgets for the sales department, more new hires, and an even larger sales force in the future. Understanding positive feedback is crucial if we are to understand the potential for systems to grow in a rapid manner.

However, as the company grows larger, each person is assumed to be less effective at selling widgets. This assumption introduces negative feedback, a closed chain of cause and effect that acts to reduce the change in a system. In this example, an increase in the sales force can lead to lower effectiveness, lower sales, less revenues, lower budgets for the sales department, fewer new hires, and a smaller sales force in the future. Understanding negative feedback is crucial if we are to understand the limits and stability of systems.

Figure 5 shows the combination of positive and negative feedback loops at work in the sales company. This diagram uses words for each of the variables and arrows for the causal connections. The arrows are arranged to allow the eye to easily follow the closed chains of cause and effect. A plus sign in the middle of the loop denotes the positive feedback that allows the company to grow. The minus signs in the middle of the loops denote the three negative loops in the sales model.

System dynamics is based on the premise that understanding dynamic behavior requires one to understand the interplay between the feedback loops in a system. A closely related tenet is that our mental skills are not sufficient to interpret the feedback structure of systems. It appears that we are good at

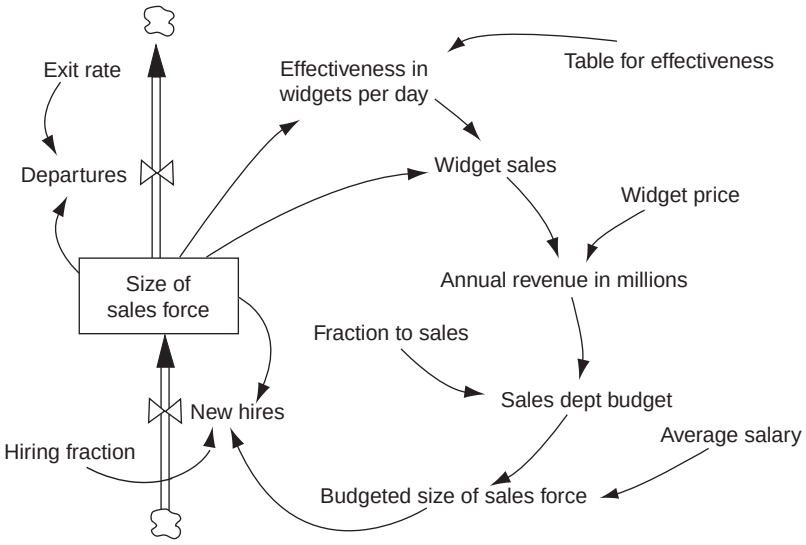


FIGURE 4 Vensim diagram of the sales company model. Adapted from Ford (1999). Reprinted by permission of Island Press.

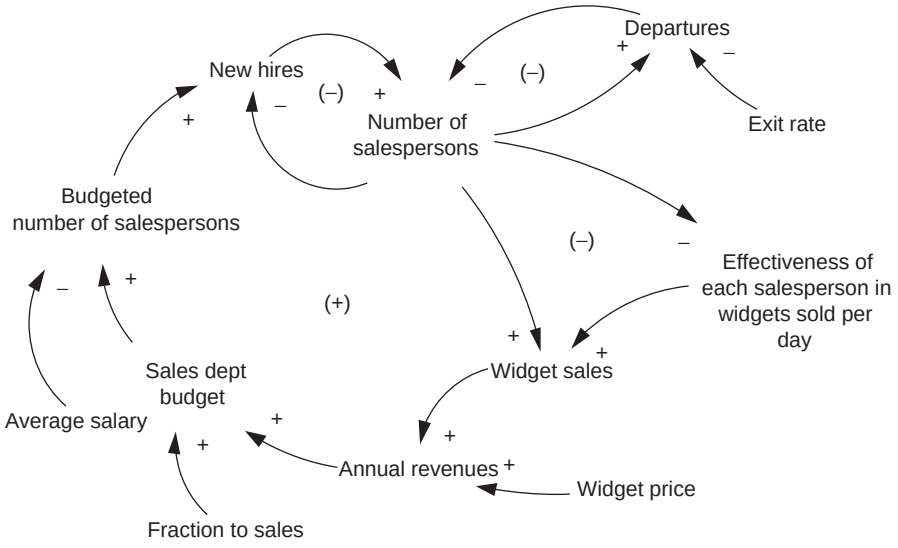


FIGURE 5 Feedback loops in the model of the sales company. Adapted from Ford (1999). Reprinted by permission of Island Press.

making lists of the many factors that might be important in a system, but we are limited in our ability to simulate the interaction between the key factors. Also, it appears that our mental skills are not sufficient to anticipate how the interactions in a system lead to growth, decay, oscillations, or other patterns of dynamic behavior.

System dynamics models are constructed with the goal of improving our understanding of dynamic behavior and anticipating the impact of policies that could lead to improved behavior. This goal has been reached in a wide range of energy models, including those that deal with privatization of electricity in the United Kingdom and deregulation of electricity in the United States. Other applications deal with variability of electricity supply in Columbia and conservation policy in the northwest region of the United States. In the following sections, three applications are described to convey a sense of the usefulness of the approach.

4. MODELS FOR NATIONAL ENERGY POLICY

One of the earliest applications to energy systems was Roger Naill's 1973 model of the exploration and production of natural gas in the United States. The model assigned one stock to represent the proven reserves of natural gas and a second stock for the unproven reserves. The model simulated the natural gas discoveries and production from 1900 to 2020 with different assumptions about the size of the resource base and different policies on price regulation. The model was used to illustrate how the resource could be depleted in a surprisingly short interval when demand is growing in an exponential fashion.

The natural gas model was expanded to represent all of the fossil fuels—oil, coal, and natural gas. The expanded model simulated end-use demands for fossil fuels as well as the intermediate demands in the electricity sector. The new model, known as FOSSIL1, was much larger than the previous model, so space does not allow for a description of the many stocks and flows. The model focused on the difficult transition from conventional fossil fuels (e.g., oil) to ultimate sources (e.g., renewables). The most important application of FOSSIL1 was a simulation analysis of the policies advocated in the first National Energy Plan issued by President Carter in April 1977. The simulations indicated that the plan would "fall short of solving the nation's energy dilemma," but it would delay world oil shortages by approximately 10 years.

Naill and colleagues were asked to expand the model for policy support at the Office of Analytical Services of the U.S. Department of Energy. The expanded model was called FOSSIL2. The new model retained the focus on the difficult transition from conventional to ultimate sources of energy, but it was greatly expanded to represent the many fuels, technologies, regulations, and policies of interest to the Department of Energy. A typical simulation covered the interval from 1900 to 2100. In 1992, Naill described the model's development and applications during both the Carter and the Reagan

administrations. He explained that system dynamics proved to be well suited for the analysis of the national energy system because of the ease of simulating stocks and flows and the importance of representing information feedback acting primarily through energy prices.

5. INCENTIVES FOR ELECTRIC VEHICLES

Another application focused on electric cars and their impact on an electric utility in California in the 1990s. Utility companies were considering incentives to promote the sale and use of electric vehicles. This is an intriguing idea but not a new one. It was first proposed by Thomas Edison, who encouraged the power companies of his day to build garages for electric vehicle owners. In the 1990s, California utilities proposed giving their customers financial incentives to reduce the purchase price of an electric vehicle. Increased sales of electric vehicles would lead to higher electricity consumption during nighttime hours when the demand could be satisfied by idle generating capacity. Utilities argued that incentives could lead to cleaner air and possibly lower electric rates. The reduction in electric rates might arise if the utility earned more in revenues from electricity sales than it would spend on fuel for the extra electricity generation. A system dynamics model was constructed to represent the vehicle purchase decisions of drivers and the accumulation of electric vehicles in the utility company's service territory. The model also simulated the operation of the company's power generators and the cash flows associated with financing the incentive program. The simulations suggested that it would be extremely difficult for the utility company to finance the incentives without having to raise its retail rates. The study concluded with the suggestion that feebates might be a more suitable method of promoting the sale of electric vehicles.

A feebate program would be run by a state or air shed regulator. The "fee" in feebates represents a fee imposed on the purchase price of a dirty vehicle. The "bate" represents a rebate to the person who purchases a clean vehicle. The idea is to finance the rebates for the clean vehicles with the fees imposed on the dirty vehicles. Proponents argue that feebates could be designed to reward cleaner vehicles, regardless of their technology or fuel. If one were thoughtful about the relative size of the fee and rebate, feebates might be designed with the cash inflow from fees balancing the cash outflow for rebates.

System dynamics was used to study the dynamic challenges faced by a program manager who would be charged with promoting the sale of cleaner vehicles while keeping the cash flows in the program fund in approximate balance. The feebate model was adapted from the model of the California utility company and designed for ease of use by students as well as industry experts. The model is sufficiently compact that the key stocks and flows may be shown here.

Figure 6 shows vehicle sales and retirements along with annual travel and annual emissions. A stock of cars is used to represent the cars in use in a hypothetical air shed. The model distinguishes between cars fueled by gasoline, methanol, electricity, and compressed natural gas, with market shares estimated with a discrete choice model based on stated preferences in a survey of Californians. The Stella diagram in Fig. 6 shows three-dimensional icons to represent variables with six different values, one for each of the six types of vehicles in the model.

Figure 7 shows the portion of the model simulating the cash accumulating in the state fund. The main cash flow into the fund is the collection of fees imposed on people who purchase dirty vehicles. The main cash outflow is rebates paid to people who elect to purchase cleaner vehicles. As in Fig. 6, three-dimensional icons are used to represent variables with six different values, one for each type of vehicle.

The feebate model has been designed as a flight simulator, a highly interactive version of a system dynamics model to promote learning through interactive experimentation. Flight simulators have proven useful in a variety of business systems, including health care, airlines, insurance, real estate, and project management. The feebate flight simulation has proven useful primarily in the classroom, where it helps students experiment with their ideas of how to control a feebate program. Figure 8 shows the flight simulator part way through a simulation in which the student has set no fees and no rebates. This "do-nothing" simulation will provide a reference case for comparison with subsequent simulations with an active feebate programs.

The students' challenge is to introduce fees and rebates to reduce emissions while keeping the balance in the state fund within reasonable limits. This is a controllability challenge, one well suited for that "branch of management science which deals with problems of controllability."

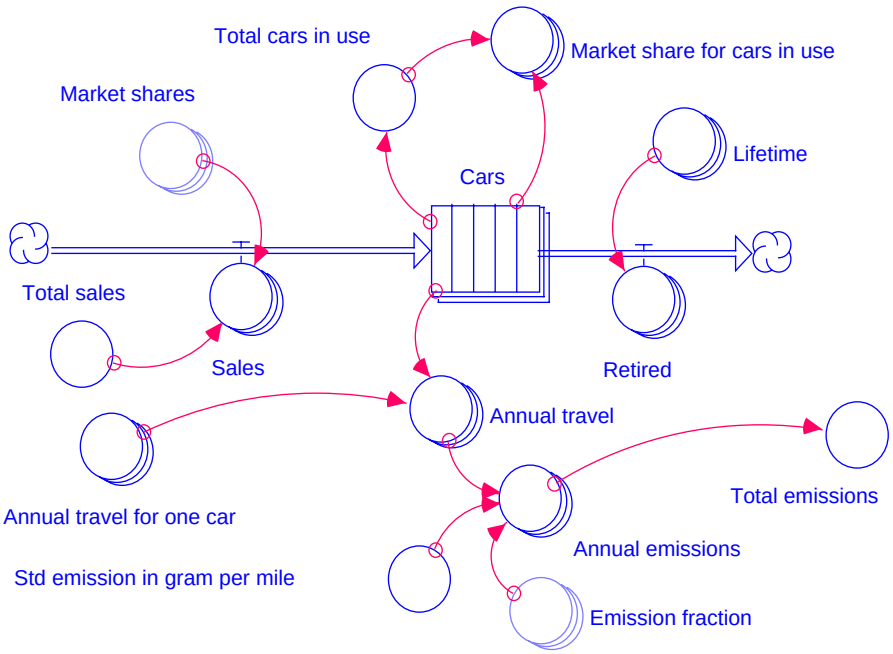


FIGURE 6 Simulating the sales and use of cars in the feebate model. Adapted from Ford (1999). Reprinted by permission of Island Press.

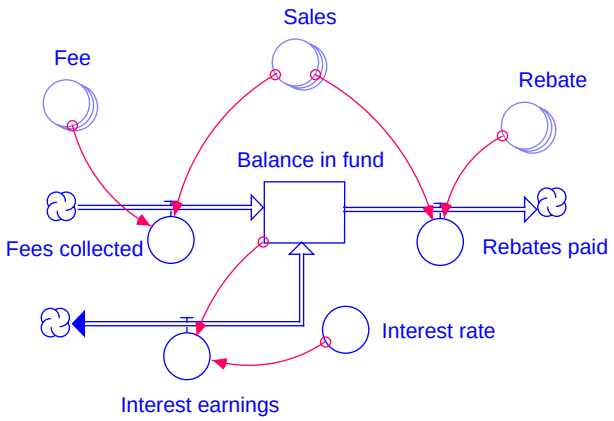


FIGURE 7 Simulating cash flows in the feebate model. Adapted from Ford (1999). Reprinted by permission of Island Press.

6. POWER PLANT CONSTRUCTION IN THE WEST

The third application deals with power plant construction in the deregulated electricity markets and is described in a special issue of the *Journal of Industry and Trade* devoted to the California electricity crisis. Many believe that competitive electricity markets are prone to the same cycles of boom and bust that appear in commodity markets and in a specialized industry such as real estate. The system dynamics study demonstrated how boom and bust might appear in the western electricity system. The model simulates wholesale electricity market prices in the western United States, as explained in Fig. 9.

Figure 10 shows the simulated construction in a business-as-usual scenario. The model represents the construction of gas-fired combined cycle (CC) power plants by merchant investors. The curves in Fig. 10 are simulation results; the small rectangles are information buttons positioned to represent actual investments in CCs. The construction curve in Fig. 10 shows that the number of CCs under construction peaks near the end of 2001. From this point forward, construction completions are greater than construction starts, and the total megawatts of CCs under construction declines. This simulation suggests that we were at the crest of a building boom in 2001. As the construction is completed during 2002 and 2003, installed CC capacity will increase to approximately 27,000 MW by the start of 2004. Figure 10 shows a second wave of construction beginning in 2006 and peaking near the end of 2007.

This simulation scenario, created in November 2001, suggests that we were in the midst of a building boom, with the actual construction well

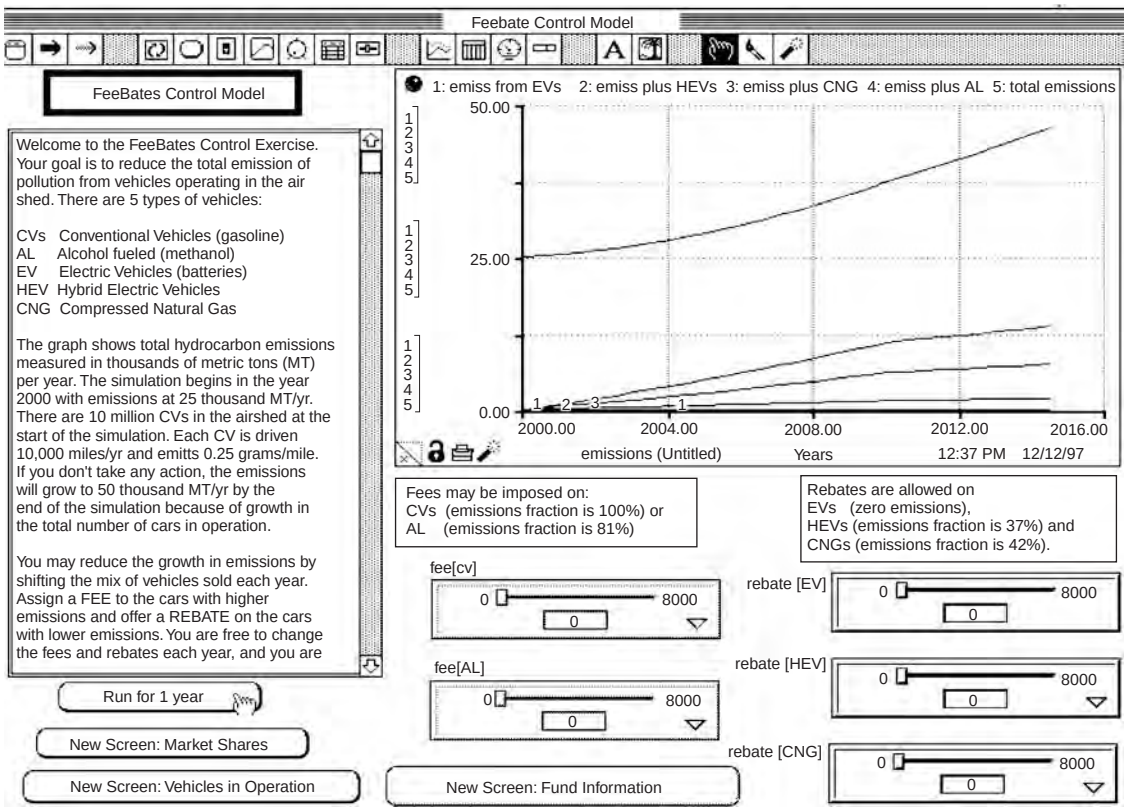


FIGURE 8 Results from a "do-nothing" simulation with the feebate model. Adapted from Ford (1999). Reprinted by permission of Island Press.

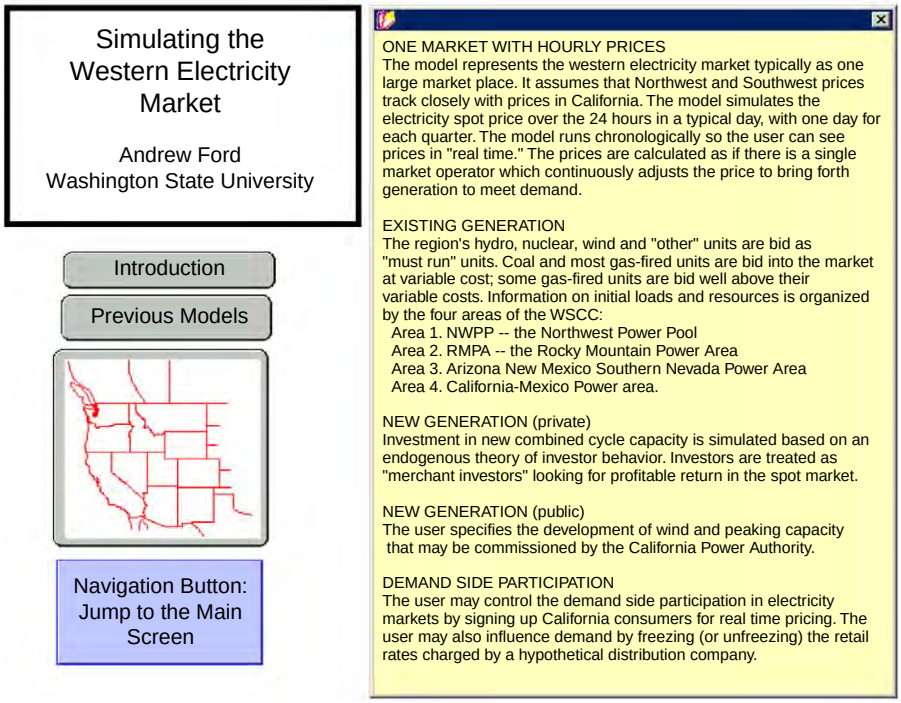


FIGURE 9 Opening screen of the western market model.

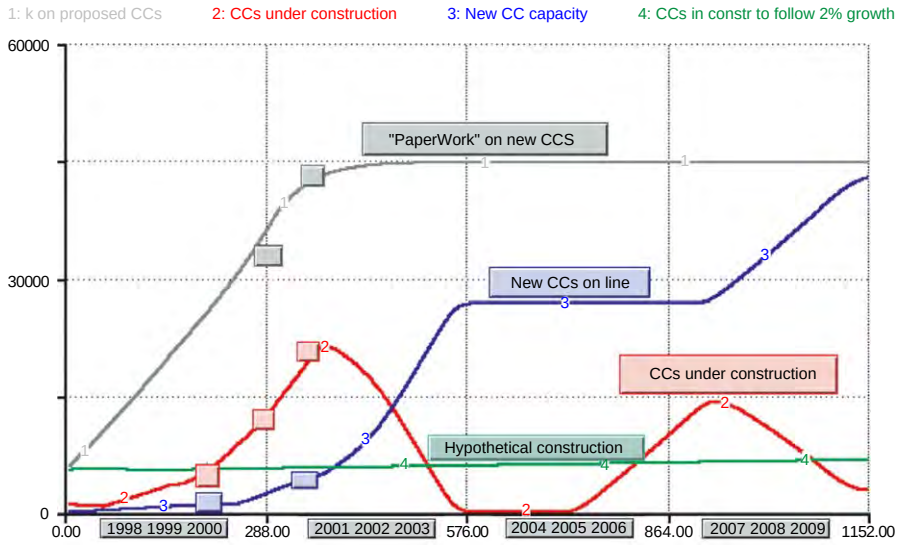


FIGURE 10 Simulated construction in the business-as-usual scenario.

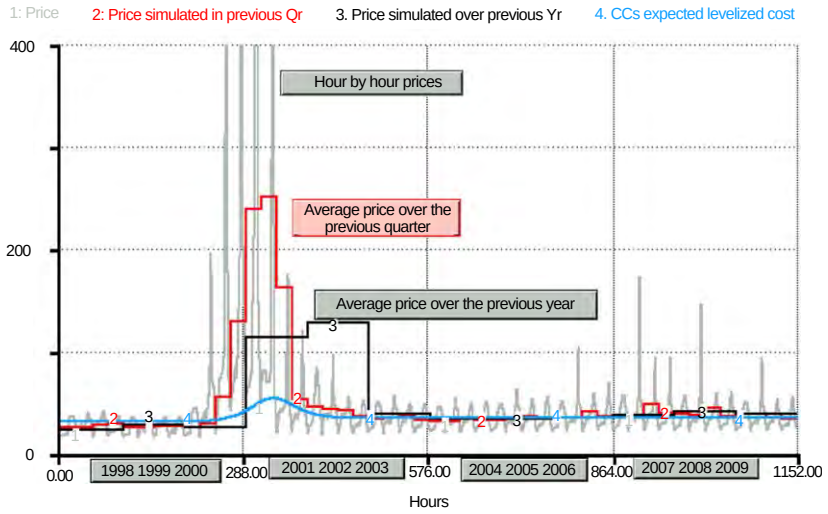


FIGURE 11 Wholesale market prices in the business-as-usual scenario.

above the hypothetical construction. This overbuilding situation continues until midway through 2003. For the next 3 years, however, the scenario envisions construction below the levels needed to keep pace with demand. During this interval, investors are reluctant to start construction, even though they hold a huge number of approved permits. Investors hold off on construction starts until 2006–2007.

Figure 11 shows market prices in the business-as-usual scenario, with the vertical scale running from \$0 to 400/MWh. The simulation indicates that hourly variations for typical days in 2002 and 2003 would be much smaller than the variations in 2001. Figure 11 shows that quarterly prices continue to decline during the interval from 2002 to 2004. The construction boom allows for much lower prices for the interval from 2002 to 2005. The figure shows that price spikes could reappear as soon as 2006 and 2007. The spikes are much less severe than the spikes in 2000 and 2001. The improved behavior arises because the scenario is based on average gas and hydro conditions. For example, gas prices do not skyrocket in 2006 like they did in 2000. Also, hydro generation does not decline in 2006 and 2007, like it did in 2000 and 2001. These assumptions mask the underlying deterioration of the

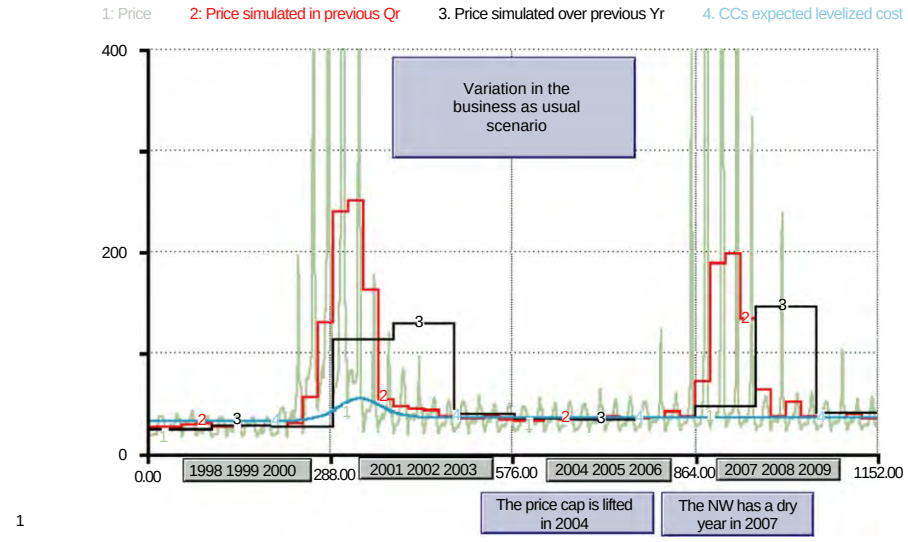


FIGURE 12 Market prices in a variation of the business-as-usual scenario.

supply–demand balance that occurs during the lull in construction.

Figure 12 shows a variation in the opening scenario to unmask the difficult conditions that could emerge within the next decade. The new simulation assumes that the northwest experiences a dry year in 2007 and that the price cap is lifted at the start of 2004. (We might envision the cap would be eliminated because those who oppose price caps would win adherents to their view during the previous 2 years—years of low prices and declining construction.) Figure 12 shows hourly prices spiking off the chart in 2007. Quarterly prices would climb to approximately $200/MWh, and the average annual price for 2007 would increase to $146/MWh, a value that exceeds the annual prices seen in either 2000 or 2001. Reserve margins (not shown here) would decline to levels lower than the simulated reserves in 2000 and 2001.

These simulations were used to promote discussion of fundamental changes in wholesale electricity markets. Without fundamental changes, the next construction boom would occur too late to prevent a decline in reserve margins and the reappearance of price spikes. If we continue with the current market structure, we run the risk of exposing the western electricity markets to another round of reliability alerts and price spikes. Alternative scenarios were generated to show the benefits of capacity payments to encourage private investment and state funding to finance public investment. On the demand side, simulation scenarios showed the benefits of implementing real-time pricing in California.

7. SUMMARY

System dynamics has been applied to a variety of energy systems at the local, state, and national level. Previous applications indicate that the method is valued for the clarity of stocks and flows and the ease of simulating the feedback loops that give rise to dynamic behavior.

SEE ALSO THE FOLLOWING ARTICLES

Complex Systems and Energy • Computer Modeling of Renewable Power Systems • National Energy Modeling Systems • Neural Network Modeling of Energy Systems

Further Reading

Coyle, G. (1977). "Management System Dynamics." Wiley, New York.

Ford, A. (1994). Electric vehicles and the electric utility company. *Energy Policy* **22**(7), 555–570.

Ford, A. (1997). System dynamics and the electric power industry. *System Dynamics Rev.* **13**(1), 57–85.

Ford, A. (1999). "Modeling the Environment: An Introduction to System Dynamics Modeling of Environmental Systems." Island Press, Washington, DC.

Ford, A. (2001). Simulation scenarios for the western electricity market, discussion paper for the California Energy Commission workshop on alternative market structures for California. Available at http://www.wsu.edu/∼forda.
Ford, A. (2002). Boom & bust in power plant construction: Lessons from the California electricity crisis. *J. Industry Competition Trade* **2**(1/2), 59–74.
Forrester, J. (1961). "Industrial Dynamics." Pegasus, Waltham, MA.
Morecroft, J., and Sterman, J. (1994). "Modeling for Learning Organizations." Pegasus, Waltham, MA.
Naill, R. (1992). A system dynamics model for national energy policy planning. *System Dynamics Rev.* **8**(1), 1–19.
Naill, R., and Backus, G. (1977). Evaluating the national energy plan. *Technol. Rev.*, July/August.
Sterman, J. (2000). "Business Dynamics." Irwin McGraw-Hill, New York.

Tanker Transportation

SHASHI N. KUMAR
Maine Maritime Academy
Castine, Maine, United States

1. **Brief History of Oil Trading and Tanker Transportation**
2. **Current Oil Movements and Tanker Demand Trends**
3. **Tanker Nomenclature and Fleet Statistics**
4. **The Economics of Tanker Markets**
5. **The Regulatory Environment**
6. **Recent Developments and Outlook for the Future**

Glossary

charterer An individual or entity, such as an oil company, that contracts the services of a ship under different chartering arrangements.
charterparty The chartering agreement or contract between a ship owner and the charterer.
deadweight (DWT) A rough measure of the cargo-carrying capacity of an oil tanker.
Panamax The largest size ship that can safely transit the Panama Canal given the limitations of the Panama Canal locks.
spot charter A ship contracted for a specific voyage by a charterer.
Suezmax A fully loaded tanker that can transit the Suez Canal.
supertankers Large and ultra large tankers.
time charter The leasing of a ship for a given period of time by a charterer.

Tanker transportation serves as a vital link in facilitating the movement of oil and petroleum products from their limited sources of origin to their innumerable destinations the world over. This particular segment of the ocean shipping industry constitutes the largest component of seaborne cargo movements today, with tanker cargoes accounting for approximately one-third of all cargoes shipped by sea. Although oil is also transported through pipelines as well as by tanker trucks and tanker rail cars, these movements are relatively minor and often confined to national or (in some limited cases) intraregional trades. Accordingly, this is a crucial transportation segment for the global community of oil consumers as well as for the oil suppliers and traders and also the owners of these ships. Aside from its vast commercial implications, tanker transportation is also becoming an epicenter of global attention for reasons ranging from environmental sensitivity to geopolitical apprehension, as evidenced through the actions and reactions resulting from a November 2002 catastrophe involving a laden tanker off the Brittany coast of Spain as well as the involvement of Venezuelan oil tankers in that South American nation's ongoing economic paralysis at the time of this writing. Although not widely recognized outside the industry, tanker transportation has always had inextricable links with oil trading and its geopolitics from its very formative years.

1. BRIEF HISTORY OF OIL TRADING AND TANKER TRANSPORTATION

1.1 The Early Years

Oil trading began during the late 19th century, with the United States being the leading source and the U.S.-based Standard Oil being the most dominant player, exporting refined oil to Europe using its fleet of purpose-built small ships. The Nobel brothers, sons of Alfred Nobel (the inventor of dynamite), built up a competing European fleet during the mid-1880s to trade oil from Russia that flourished until the Bolshevik revolution. Crude oil trading was not considered a feasible option during those years due to the relatively low value of the commodity. During the post-World War I era, short-haul crude oil trade originated, with most movements being to the United States from the nearby Latin American sources of Mexico and Venezuela. As the crude oil trade increased gradually, so did the need for new refining

 © 2004 Elsevier Inc. All rights reserved.

capacity in major consumption markets. The Middle East and its vast oil resources came under Allied control during World War II and helped to meet the wartime logistical needs of the Allied forces. Another major development during those war years was the rapid construction of a large number of T-2 tankers by U.S. shipyards to supply fuel to the Allied fleet. On completion of the war, the surplus T-2 tankers were sold off at low prices, laying the foundation for the shipping fortunes of many entrepreneurial tanker owners during subsequent years.

During the postwar era, the Middle East emerged as the focal point of the oil industry, which by this time had supplanted coal as the most dominant source of primary energy for a world economy recuperating from the ravages of two closely fought world wars. Thus, although the Gluckauf, a prototype of the modern oil tanker, was launched in 1886, the highly specialized tanker transportation sector of the current vintage did not originate until the 1940s and 1950s. The increased demand for crude oil came not only from the United States and Western Europe but also from a rapidly recovering Japan, a nation devoid of indigenous oil resources. The distances between such major consumption and production centers of crude and refined oils contributed toward increased demand for the transportation of those commodities, crude oil in particular. Accordingly, the demand for oil tankers rose, hence the meteoric growth of that specialized market in a very short span of time. As viewed by maritime economists, tanker transportation is a derived demand industry, for fluctuations in the oil volumes traded usually lead to very perceptible impacts on the trading conditions within the tanker market.

Until 1952, tankers were operated primarily by the big oil companies, with independent ship owners playing a fringe role. It was typical to find marine transportation divisions within the corporate hierarchies of the vertically integrated multinational oil majors. Traditional ship owners entered the market for a brief span in 1953, to be followed later by government-owned shipping companies. By the early 1960s, the oil majors' investment in direct ownership of oil tankers had dropped to approximately one-third of the global fleet. However, the oil majors still maintained direct control of 90% of their transportation requirements, supplementing their owned fleets with ships on long-term time charter from independent ship owners. It was important for the oil companies to maintain such control because the ocean transportation component was a factor input in the landed price of their commodity. If this were to vary depending on daily market swings, it would have an impact on the price of crude oil from various suppliers despite its being a homogenous commodity. Accordingly, the oil majors procured barely one-tenth of their transportation requirements of those years from the open market through spot charter agreements. Although they had the wherewithal to provide all of their tanker needs in-house, such an attempt would only help their competitors, especially the smaller oil companies (as postulated by Zannetos), so this was not attempted.

1.2 The Golden Years

The 1960s was a period of spectacular growth in the oil industry in general, and the tanker market in particular, due to its derived nature. The rapidly growing U.S., Western European, and Japanese economies consumed large quantities of crude and refined oils, and the United States, an oil exporter during earlier years, became dependent on foreign oil by the mid-1960s. Correspondingly, oil tankers began to increase in both number and size. The ship size limitation of 65,000 deadweight (DWT) imposed by the dimensions of the Suez Canal was no longer relevant, with the bigger vessel sizes benefiting from their economies of size and contributing toward lower unit cost of oil transportation. The construction of *Universe Apollo* of 100,000 DWT in 1959 was a major landmark, as was the construction of the first ultra large crude carrier (ULCC), *Universe Iran*, 10 years later in 1969. A study by Tusiani documented that although the first 200,000-DWT tanker was not delivered until 1966, by 1973 there were a total of 366 such very large or ULCCs in trade, with another 525 being under construction or on order at the time—ample testimony to the sanguinity of the period. Thus, with the increase in demand for crude and refined oils apparently showing no end in sight, new tankers of increasing sizes, financed by overly optimistic shipping financiers, were being ordered at unprecedented levels, characterizing the 1960s and early 1970s as golden years of the tanker transportation sector.

From the oil majors' perspective, their domination of the retail market was challenged by the evolution of the oil spot market in Rotterdam, Netherlands, during the late 1960s that allowed the entry of nontraditional retailers in the market. The role of politics in oil trading and tanker transportation had also become clearly obvious by this time, starting with a brief closure of the Suez Canal in the wake of its nationalization by Egypt in 1956, leading to an

immediate increase in freight rates for a brief span of time. A more lasting closure of the Suez Canal that began in 1967, as war broke out between Egypt and Israel, remained in effect until 1975. This increased the sailing time to destinations in Europe and North America, thereby increasing the demand for tankers and further consolidating the good fortune of some tanker owners. The Organization of Petroleum Exporting Countries (OPEC) cartel, formed in 1960, remained relatively benign during the decade and did not flex its political muscle until 1973. Thus, as discussed next, an era of turbulence and uncertainty quickly permeated the oil trade and tanker transportation market, in stark contrast to the opulent 1960s and early 1970s.

1.3 The Uncertain Years

The Yom Kippur War of 1973 put a dramatic end to the phenomenal growth that oil and tanker markets experienced during the previous two decades. This, along with the Arab embargo of oil shipments to the United States and The Netherlands and the increasing dependence of the United States on oil imports, led to the first oil price shock during which the OPEC nations raised the price of a barrel of oil fivefold on an average to $10/barrel. Furthermore, the Arab nations expropriated American and European oil investments in their countries. However, all of these were merely temporary setbacks, and the global demand for oil increased again once the embargoes were lifted. The Middle East consolidated its position as the world's largest exporter of oil by the late 1970s, supplying close to three of every five barrels of oil exported in the world. However, the increasing oil prices made oil exploration economically feasible in politically safer locations such as Alaska and the North Sea. The apparent reemergence of stability in oil market was once again disturbed by political events in Iran, hence the second oil supply shock in 1979–1980 that quadrupled oil prices. The subsequent introduction of oil conservation initiatives in major consuming nations, as well as the sudden growth of non-OPEC sources of oil, made significant inroads in the market share of the Middle East crude oil in particular. The collective effects of these developments were so remarkable that the level of global oil consumption reached in 1979 was not surpassed until 1990 and was only 13% higher even 21 years later in 2000.

The events of this period were particularly troubling for tanker owners who were left with surplus shipping capacity (built under speculative contracts signed during the waning golden years discussed in the previous section) and huge mortgage payments. Thus, the golden era of the tanker transportation sector came to an abrupt end and was replaced by a period of gloom and doom that only worsened during the 1980s as the sector began to bear the full brunt of the conservation measures enacted after the second oil supply shock. As the oil prices began to stabilize, the cohesion of the OPEC cartel itself came under threat, with member nations undercutting each other and crude oil prices dropping to as low as $10/barrel. The price eventually stabilized at approximately $15 to $18 per barrel by the late 1980s, and trade volumes resumed their increasing trend. Once again, the demand for oil tankers began to rise, as did the profitability of tanker owners, until that trend was altered by a new set of problems triggered by the grounding of the Exxon *Valdez* (discussed later). The next section discusses current trends in oil movements, an understanding of which is essential to comprehend the dynamics of the contemporary tanker transportation market.

2. CURRENT OIL MOVEMENTS AND TANKER DEMAND TRENDS

International Energy Agency statistics show that total primary energy supply through crude oil increased from 2880 million tons of energy in 1973 to 3700 million tons of energy in 2000. Table I shows the major producers of crude oil and petroleum products in 2001, and Table II shows the major exporters and importers for the year 2000. A recent study by Glen and Martin estimated that 59% of world oil produced in 2000 was traded by sea, a significant increase from the prior levels of 48% in 1990 and 56% in 1995. However, estimating demand for tanker transportation based purely on the volumes traded would be inaccurate unless the distances involved were also included. As an example, the disruption in oil movements caused by the Iraqi invasion of Kuwait induced nations such as the United States to partially substitute their dependence on Middle East oil with oil from closer, politically safer locations such as the North Sea, West Africa, Venezuela, Mexico, and Colombia. In such cases, even if the volumes imported were to stay the same, the demand in ton-miles would drop. A study by Kumar reported that a shift of 10 million barrels per day in sourcing the U.S. crude oil imports from the

TABLE I

Major Producers of Crude Oil and Petroleum Products, 2001

Crude oil			Petroleum products		
Producer	Million tons	Percentage of world total	Producer	Million tons	Percentage of world total
Saudi Arabia	421	11.8	United States	823	23.7
United States	354	9.9	Japan	207	6.0
Russia	347	9.7	People's Republic of China	196	5.6
Iran	186	5.2	Russia	174	5.0
Mexico	179	5.0	Korea	122	3.5
Venezuela	173	4.8	Germany	116	3.3
People's Republic of China	166	4.6	India	101	2.9
Norway	162	4.5	Italy	95	2.7
Canada	125	3.5	Canada	94	2.7
United Kingdom	118	3.3	France	89	2.6
Rest of world	1343	3.7	Rest of world	1458	42.0
World total	3574	100.0		3475	100.0

Source. International Energy Agency (2002).

TABLE II

Major Exporters and Importers of Crude Oil and Petroleum Products in 2000

Crude Oil				Petroleum Products			
Exporter	Million tons	Importer	Million tons	Exporter	Million tons	Importer	Million tons
Saudi Arabia	320	United States	511	The Netherlands	63	United States	74
Norway	146	Japan	214	Russia	54	Japan	51
Russia	144	Korea	123	Saudi Arabia	53	The Netherlands	45
Iran	116	Germany	104	United States	49	Germany	42
Venezuela	115	Italy	90	Singapore	41	Singapore	39
Nigeria	107	France	86	Korea	40	France	27
Iraq	102	People's Republic of China	70	Venezuela	37	People's Republic of China	24
United Kingdom	93	India	67	Kuwait	26	Korea	23
Mexico	92	The Netherlands	61	Algeria	21	Italy	20
United Arab Emirates	84	Spain	59	United Kingdom	21	Spain	20
Rest of world	656	Rest of world	642	Rest of world	352	Rest of world	319
World total	1975		2027		757		684

Source. International Energy Agency (2002).

Middle East to the North Sea would reduce tanker demand in ton-miles, making 32 very large crude carriers (VLCCs) redundant. Likewise, a shift in the sourcing of crude oil in Western Europe from the Middle East to closer sources, such as Russian and North Sea oil, would have a similar detrimental effect on tanker demand. In situations like this, unless tanker demand in other optimum long-distance routes pick up, the very large tankers will migrate into suboptimal routes that are usually served more

efficiently by smaller vessels, and this will lower the freight rate for all players in the market.

It is worth noting that the ton-mile demand for oil tankers reached its highest value in 1977, whereas the crude oil volume traded continued its upward trend until 1979. The ton-mile peak reached for crude oil movements remains unsurpassed even today. Currently, although the United States alone is responsible for one-quarter of the world crude oil consumption, much of its imports come from closer locations that include Canada, Mexico, Venezuela, West Africa, and Colombia. The Former Soviet Union is now a prominent supplier of oil to Europe. Although roughly one-half of all current world crude oil exports originate from the Middle East, and its major export markets are currently in the Asia Pacific region to countries such as Japan and South Korea. Because the Middle East–Asia Pacific trade is a relatively long distance trade optimal for the large tankers and also because the overall crude oil volumes transported by sea are on a rising trend, the redundancies made by the crude oil sourcing shifts in Western Europe and the United States have not had a sustained direct impact on the market for very large tankers. As indicated in Table III, the estimated productivity of tanker fleet in terms of cargo loaded per deadweight ton, as well as the ton-mile demand, is on a steady upward trend after its precipitous drop during the 1980s and early 1990s. However, as noted by Glen and Martin, the average tanker haul recorded in 2000 was 4917 miles, while the average in 1977 was 6651 miles. The 26% drop in ton-mile demand is too severe a shock for a forlorn tanker owner reminiscing about the golden years. The major export areas in oil trade for the year 2000, in descending order, were the Arabian Gulf, West Africa, Venezuela, Mexico, North Africa, the Black Sea, Sidi Kerir, Indonesia/East Malaysia, China, the East Mediterranean, the Baltic Sea, and the Red Sea. The major import areas for 2000, also in descending order, were Western Europe, the United States, Japan, South Korea, Canada's east coast, Taiwan, and Hong Kong.

TABLE III

Estimated Productivity of Oil Tankers for Selected Years

Year	Oil cargo carried (in million tons) by Tankers > 50,000 DWT	Tons per DWT of tankers	Demand in billion ton-miles	Ton-miles per DWT
1970	1182	8.58	6039	43.82
1980	1564	4.79	9007	27.56
1990	1427	5.96	7376	30.81
1998	1985	7.10	9465	33.86
1999	1988	7.04	9586	33.94
2000	2070	7.25	10,107	35.41

Source. UN Conference on Trade and Development (2001).

3. TANKER NOMENCLATURE AND FLEET STATISTICS

Oil tankers can be classified into two broad categories: crude oil tankers and product tankers. Crude oil tankers are typically large dedicated ships that carry solely crude oil. As an example, the *Jahre Viking*, a crude oil tanker built in 1977, is the largest floating object ever built and would require 16,500 road tanker trucks to empty its full load of cargo. Although it is theoretically possible for a crude carrier to carry petroleum products, the time and monetary costs associated with tank cleaning for such a switch would be prohibitively high so that it is rarely done these days in tanker operations. The crude carriers are classified based on their carrying capacity and suitability for a particular trade, as depicted in Fig. 1. However, there is considerable variation in the deadweight figures for a particular class. For example, a deepening of the Suez Canal would propel an upward adjustment of the capacity of a Suezmax tanker. The VLCCs and ULCCs are usually used on the long-distance hauls, whereas the medium-sized vessels, Suezmax and Aframax, are used in the medium-distance ranges (e.g., from West Africa to the United States). The Panamax vessels are the largest tanker sizes that can transit the Panama Canal.

The product carriers, which are smaller vessels than the crude carriers, transport both the clean and dirty products of the petroleum refining process. However, because of the level of tank cleaning required, they usually do not switch from carrying clean products to carrying dirty products. Contrary to the crude carriers, product carriers are built to facilitate the carriage of multiple products simultaneously. They are kept well segregated throughout the loading, loaded passage, and unloading operations using a complex and sophisticated network of pipelines and valves on-board the ships. Furthermore, whereas the crude oil movements are typically one-way trades that consist of a loaded transit between one of the few export locations in the world and a major refining center followed by a ballast voyage back to the load port, product carrier

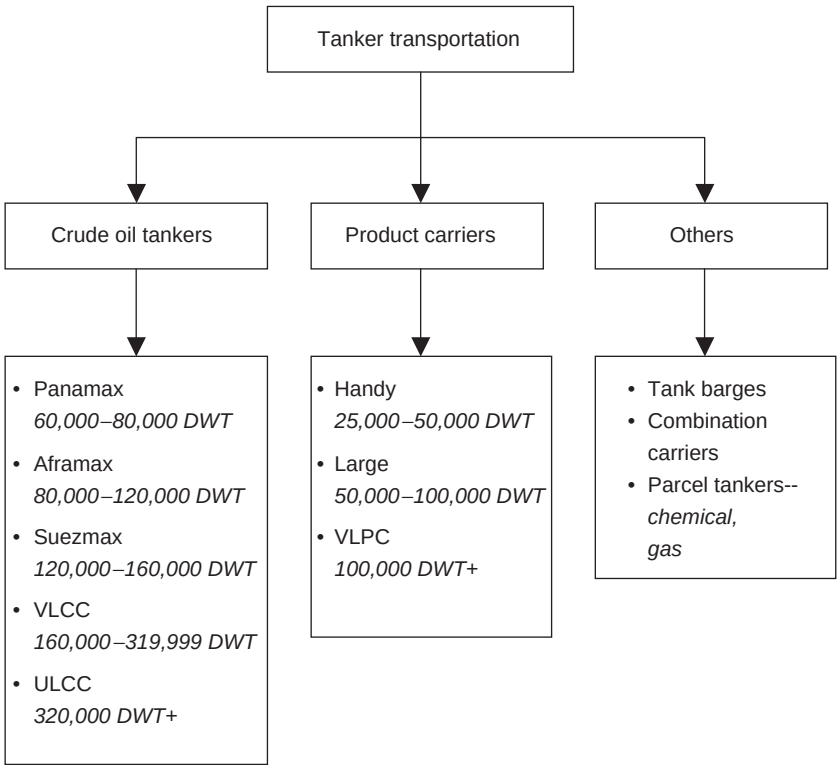


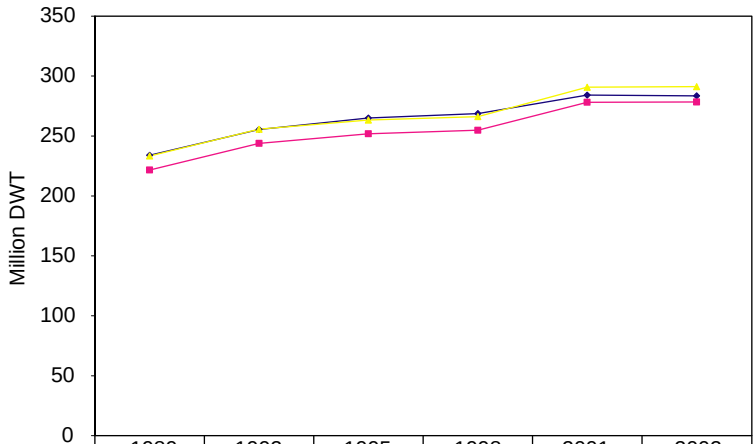
FIGURE 1 Tanker classification. VLCC, very large crude carrier; ULCC, ultra large crude carrier; VLPC, very large product carrier.

movements are more dynamic and provide better vessel use. It is not unusual to find a product carrier bringing certain products to a particular site followed by loading a different group of distillates from a refinery in that discharge port or its vicinity for discharging elsewhere.

Figure 1 also includes a third category that consists of tank barges and other ships linked to the transportation of oil and its by-products. Tank barges usually operate in coastal trade. Combination carriers, specially constructed ships that can switch from trading bulk crude oil to trading dry bulk commodities, are relatively small players in today's tanker market. The parcel tanker sector consists of highly specialized chemical tankers, liquefied petroleum gas (LPG) carriers, and liquefied natural gas (LNG) carriers. These vessels also operate outside the tanker market and are not typically included in a discussion on tanker transportation. Figure 2 provides a summary of the tanker fleet developments during the past decade or so. It shows the total tanker (gross) and net tanker fleet for each of the years, with the difference being the idle capacity. The total active fleet for each year is obtained when the combined carrier tonnage is added to the net operating fleet for each chosen year. As shown, the global tanker fleet is on a rising trend once again and the idle capacity is on the decline, showing a tightening market in general.

4. THE ECONOMICS OF TANKER MARKETS

Zannetos's extensive analysis of the tanker market found it to be rather concentrated, with 1.5% of the owners owning 35% of the fleet and the remaining 65% dispersed among the rest of the owners, none of which had a capacity share that exceeded 7% of the global fleet. Paradoxically, despite the seemingly concentrated nature of tanker ownership, Zannetos found that the market for oil tankers behaved as if it were perfectly competitive. Although a study by Kumar in 1995 based on the tanker market conditions of the post-Exxon *Valdez* years highlighted the elevation of entry barriers in tanker markets and the potential for oligopolistic behavior by incumbents, it is fair to surmise that the market has remained highly competitive throughout its life span. This section



	1989	1992	1995	1998	2001	2002
Tanker fleet	233.9	255.4	265	268.7	284.1	283.5
Net operating fleet	221.6	243.9	251.9	254.9	278.2	278.4
Active fleet	233.3	255.7	263.3	266.2	290.7	291.2

FIGURE 2 Tanker fleet statistics. Data from R. S. Platou (2002).

provides a discussion of the ownership trends, market dynamics, cost categories, and tanker freight rates.

4.1 Tanker Ownership and Market Mechanics

Although the market has maintained its competitive features, the ownership structure in the market has undergone radical changes. The oil majors, in particular, have shifted away from a policy of direct control of bulk of their tanker transportation needs to one of greater dependence on the spot market for chartering ships on an "as needed" basis. The study by Glen and Martin provided three reasons for the ongoing structural change in the composition of ownership of oil tankers: the evolution and growth of the spot market for oil and oil products, a change in the logistics strategy of oil companies, and the increasing environmental awareness of the public at large that has led to the imposition of new environmental regulations that directly affect tanker operations. Their study also found that the tanker fleet owned by the top 10 independent owners was 162% larger than the fleet size owned by the top 10 major oil companies. However, if one were to eliminate the fleet owned by government-owned oil companies from the preceding statistics, the tonnage owned by the remaining oil majors would be close to 10% of the fleet owned by the top 10 independents. Using capacity shares as a proxy for market competitiveness, the tanker market maintains a highly competitive stance. It also meets the basic behavioral assumptions of the perfect competition model within limits.

The demand for oil tankers can be satisfied either through their direct ownership or through the well-established tanker charter market that exists in commercial centers such as London and New York. The charter market brings together cargo brokers and ship brokers acting as agents of the cargo owners and ship owners, respectively. With the help of their specialized knowledge, tankers are fixed either on long-term charters or, increasingly, on spot or voyage charters. Whereas a time charter agreement allows the exclusive use of a ship and its crew for an extended period of time (e.g., 1 year), the spot or voyage charter agreement covers one specific voyage. It is also possible to acquire tankers on bareboat charter, an arrangement under which bare ships are handed over to the chartering interest, typically for long periods of time (e.g., 5 years). The charter agreement in all cases is generally referred to as a charterparty. There are a number of standardized time and voyage charterparty clauses in existence today that are widely understood by all parties involved and constitute the boilerplate clauses of all charterparties.

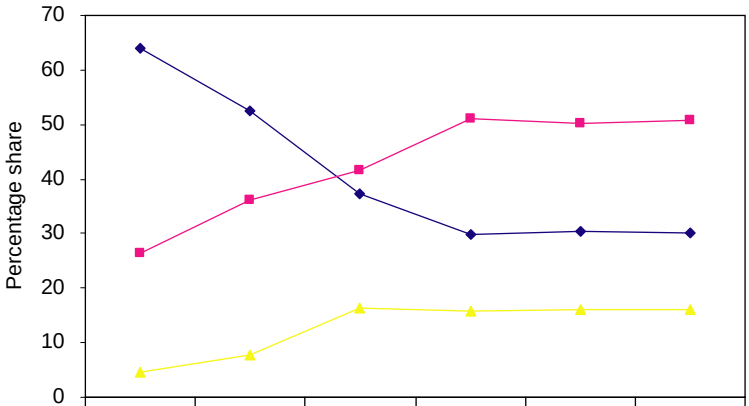
4.2 Tanker Costs and Operations Management

There are three categories of costs involved in tanker operations and management: capital cost, running cost, and voyage cost. Capital cost is the cost of acquiring the tanker and represents a fixed cost. Running cost is a semivariable cost and includes crew salaries and subsistence cost, maintenance and

dry-docking costs, stores and lubricant costs, and administrative expenses. Voyage cost varies with individual voyages depending on the distance and the route involved. Examples of this category include fuel cost, tug and pilot charges, port and harbor dues, and cost of transiting a canal(s). The preceding costs are borne by either the charterer or the ship owner under different charter agreements. With a voyage charter, the ship owner bears all three costs and any surplus from the charterhire would be the owner's profit. With time charter agreements, the ship owner bears capital cost as well as running cost, whereas with bareboat charters, the ship owner bears only the capital cost, with all other costs being borne by the charterer.

Unlike capital costs and voyage costs, running costs can be trimmed through various strategies. Accordingly, ship owners might pursue entrepreneurial options such as flagging their vessels in countries outside their own where there is a more benevolent attitude toward shipping through liberal regulatory and taxation policies. These convenient open flags of registry also offer liberal crewing options, particularly attractive to ship owners from nations such as the United States where labor costs are approximately five times more expensive than the going international market rates. As shown in Fig. 3, the tanker tonnage of developed market economy nations has been halved during the past 30 years, whereas the tonnage of the open registry fleet has doubled and that of the developing countries has quadrupled. As Kumar and Hoffman discussed in their recent study, such trends have led to the globalization of the shipping industry in general. Indeed, the tanker market exhibits very high levels of globalization, with the nations of Panama and Liberia being the largest registries of such vessels, although the beneficial ownership of those vessels still remains in traditional ship-owning nations such as Greece, the United States, and Japan. Unfortunately, such practices may also be used by unscrupulous ship owners to overcome the mandated minimum standards of safe ship operation, resulting in tarnishing the image of such liberalism in shipping.

Another strategy to bring down the operating cost of tankers, used independently or in combination with the flagging out strategy, is through outsourcing the technical management of ships to a third party that specializes in ship management. This has become particularly useful during the contemporary era because it helps the ship owners focus more on the strategic aspects of ownership while delegating the mundane technical aspects to a specialist who has the expertise as well as the economies associated with managing the operations of a larger fleet of ships. Furthermore, with the entry of unconventional ship owners such as pension funds and banks, there is no option but to use such dedicated ship managers who possess the technical know-how and expertise. The market for ship managers is also highly competitive with relatively no entry barriers, and there is considerable swapping of ships from one manager to the other as warranted by circumstances.



	1970	1980	1990	1998	1999	2000
Developed market economy countries	63.9	52.5	37.3	29.8	30.5	30
Open registry	26.4	36.2	41.6	51.2	50.2	50.8
Developing countries	4.7	7.7	16.3	15.9	16.2	16.1

FIGURE 3 Tanker ownership by country groups. Data from UN Conference on Trade and Development (2001).

4.3 Freight Rates

It is to be noted that transportation cost is a very small component in the retail price of oil. As an example, the study by Glen and Martin noted that in the United Kingdom, the transportation cost component amounts to less than 1.3% of the landed cost of petrol. The relatively minor incidence of the freight component in the landed cost makes it highly inelastic in nature, as predicted by economic theory and also as documented by various studies. However, as was made painfully clear especially after the second oil price shock, the demand for tanker transportation is directly related to economic activity in general so is elastic. Thus, it would be typical to find a global recession during which manufacturing activity is considerably reduced, negatively affecting the demand for oil tankers and with the resulting excess supply of shipping capacity suppressing the freight rates. The industry is very cyclical in nature, with no reliable forecasting mechanism to predict its vacillations. A period of high freight rates, whether caused by external stimuli (e.g., war, political boycott) or random events (e.g., an unusually strong winter in the northern latitudes), usually leads to high exuberance in the market. It is immediately reflected through increased new orders for ships, rising prices of second-hand ships, and reduced scrapping of older tonnage. The ship owners often find the circumstances to be changing precipitously in the other direction, more often for reasons beyond their control, although their own actions during the years of opulence may have indirectly contributed to the volatility and made the transition even more painful.

There are different ways of expressing tanker freight rates. The term used to quote freight rates in the spot market is in WorldScale (WS), where WS100 is a nominal index number. The WorldScale Association publishes an annual listing of WS100 rates (in dollars/metric ton) for all potential trade routes for a standard 75,000-DWT tanker operating under assumed conditions and earning a notional daily hire of $12,000. The principle behind the WS system is that regardless of the voyage undertaken, identical tankers will earn the same gross revenue per day. The WS nominal rate serves as a starting point for spot charter negotiations, and because the total earnings are proportional to the quantity of cargo carried, a vessel with a lower carrying capacity would typically bargain for a higher WS rating than would a larger vessel, assuming that both were trading on an identical trade. Any spot market fixture in which freight rates are expressed in dollars per ton can be converted into WS figures that facilitate easier comparison of market conditions. By converting the WS figures into time charter equivalent (TCE), a comparison can be made to the current rates in the time charter market. Likewise, a fixture in the time charter market can be converted into a WorldScale equivalent (WSE) for comparison with rates in the spot market. From a ship owner's perspective, it would be preferable to have the ship on long-term time charter when the spot rates are headed down, whereas spot fixtures would be more attractive in a rising market. However, tanker chartering still remains more of an art than a science given the inability to foresee the market changes with any level of comfort.

Ship owners have the option of laying up their ships when the market conditions are poor. Although this eliminates their voyage costs and reduces the running costs, there will be no respite from making capital cost payments. As the market recovers, idle capacity reenters the trade. However, as noted by R. S. Platou, a leading ship broker, so long as the capacity use remains below 80%, freight rates remain relatively stable. Glen and Martin referred to this as the elastic segment of the short-run supply curve of tankers. The slope of the curve increases beyond the 80% range and becomes fully vertical (completely inelastic) when all existing ships are used. Under such circumstances, any increase in the derived demand would cause dramatic fluctuations in the freight rate and, hence, the profitability of tanker owners. In the long run, the introduction of newer and more efficient ships would bring down the marginal cost of transporting the oil, culminating in the low incidence of ocean freight in the landed cost of oil discussed earlier. Furthermore, recent studies by Kavussanos and by Glen and Martin showed that different sizes of tankers operate in distinct markets with varying levels of risk and returns, although all of the submarkets remain interdependent in the long run. One commonality that affects all segments of the market, and that is becoming increasingly significant, is the regulatory environment in which all tankers operate.

5. THE REGULATORY ENVIRONMENT

The operations on-board an oil tanker are radically different from those on other types of ships, primarily due to the physical properties of the cargo. The entire cargo operations are highly automated and proceed with no one on-board the ship or shore seeing the cargo physically. Even minor misunderstanding of an order or a miscalculation can cause a

major spill in pristine locations. By the same token, a tanker crewed by properly trained seafarers under good management could very well be the safest ship afloat. Although most tanker voyages today are completed safely and go unreported, even a minor tanker pollution accident often gets widespread attention from the media, and the image of a polluted beach laden with dead flora and fauna is a sad and telling picture. Given the increasing environmental awareness of society at large, a tightening of the regulatory environment of oil tankers was only to be expected. Furthermore, investigations of many maritime casualties have shown repeatedly that most accidents were attributable to the human factor.

The International Maritime Organization (IMO), a multilateral governmental entity established under the auspices of the United Nations to promote safer shipping and cleaner seas, has been a key player in this field. The 1973/1978 MARPOL Convention imposed on tanker transportation greatly ameliorated the operational pollution of the seas, previously the predominant cause of oily efflux from tankers. The International Safety Management (ISM) Code Amendment to the Safety of Life at Sea Convention and the Amendments to the Standardization of Training, Certification, and Watchkeeping (STCW) Convention are two recent multilateral regulatory developments of particular significance. The ISM Code for the Safe Operation of Ships and Pollution Prevention applies a very tough set of standards to all aspects of shipboard operations and management. Furthermore, the shore office and the management of the firm also come under scrutiny, thereby expanding the scope of the regulatory reach to areas that were traditionally excluded. The Amendments to the STCW Convention have attempted to raise the threshold of seafarer education and training in general by mandating the minimum acceptable levels. The establishment of such a floor in maritime education was long overdue given the wide disparity in the knowledge, skills, and abilities of seafarers and the propensity of unscrupulous ship owners toward staffing their vessels with the cheapest crews available.

The most drastic regulatory impact on tanker transportation came from a unilateral regulation enacted by the United States in response to the grounding of the Exxon *Valdez* off Alaskan waters. The U.S. Oil Pollution Act (OPA) of 1990 is a comprehensive plan to prevent and clean up oil spills, impose greater liability on those responsible for spills, and compensate those who suffer resulting economic damage. Its provisions include liability, compensation, prevention, and removal aspects of a pollution incident. It mandated the need for double-hull tankers and established a phase-out period for the existing single-hull tankers. Above all, the law also eliminated the limitation of liability of ship owners involved in pollution incidents, and many tanker owners immediately responded by staying away from the U.S. market (at least temporarily). The IMO enacted new tanker design rules in the wake of the U.S. initiative, giving ship owners the option of building ships with double hulls or an equivalent level of protection. However, given the U.S. mandate, the industry has gravitated toward the double hull as the requisite standard bearer.

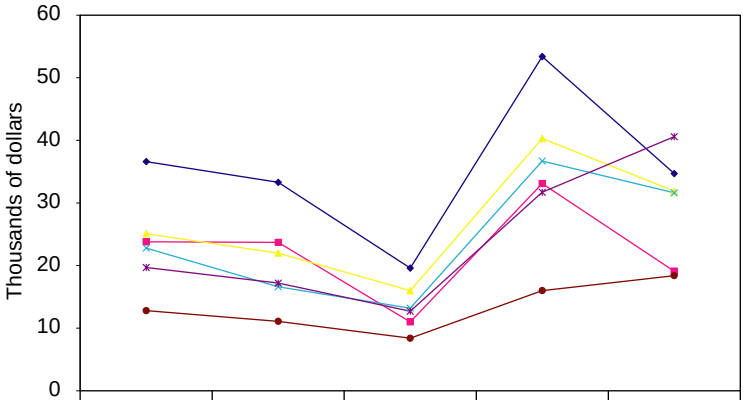
Given the diversity in ship ownership and operation, the existence of multilateral and/or unilateral regulations alone does not necessarily make the industry any safer. Investigation of past casualties has often shown some flag states to be negligent in enforcing the statutes. As a result, many nations have enacted a procedure whereby ships that visit their ports are inspected by local authorities for compliance with international standards. Such port state control inspections are now very common in most trading areas and result in rogue ships being arrested and their names being publicized through the media Despite all of the new regulations, tanker accidents still do happen, as was illustrated in the case of the ship *Erika*, a 1970s Japanese-built, Maltese-flagged, Italian-owned tanker. The ship, laden with 30,000 tons of viscous fuel oil on a voyage from Dunkirk to Leghorn (under charter to TotalFinaElf, the French oil major), encountered stormy seas and sank off the coast of France in December 1999, spilling 12,000 tons of its cargo and causing severe environmental and economic damage. The IMO, under pressure from the European Union, amended the MARPOL Convention in April 2001 to speed up the phasing out of single-hull tankers and also empowered member states to ban the entry of such tankers that are more than 25 years old. One direct consequence of the changing regulatory ambience has been the trend toward divestment of tankers by oil majors such as Exxon–Mobil. Currently, the oil majors' tanker fleet is the smallest that they have ever had during the modern tanker era.

6. RECENT DEVELOPMENTS AND OUTLOOK FOR THE FUTURE

Developments during the first 3 years of the new millennium serve as an excellent synopsis of the

unpredictability of the tanker transportation sector. Commercially, 2000 was an exceptional bull market year, with a tightening supply of ships and increasing demand. However, there was a precipitous downward dip in 2001, as depicted in Fig. 4, due to a global recession that led to lower demand for oil along with the rising security concerns resulting from the World Trade Center attacks. The relatively weak winter season in North America and the reemergence of Russia as a major player in the oil export sector did not help the tanker market either. However, the market dynamics changed dramatically in mid-October 2002, once again driven by geopolitical developments. As the possibility of a U.S. attack on Iraq became evident in October 2002, charterers began stockpiling inventory prior to the outbreak of hostilities and increasing the demand for modern larger tankers that eventually filtered down to the smaller classes of ships. Furthermore, the ongoing political crisis in Venezuela had halted oil exports from that nation. This affected the tanker market directly due to the involvement of striking government-owned Venezuelan tankers as well as to the urgency to find alternate oil supplies for the United States, a nation that depended on Venezuela for 13% of its oil imports before the political crisis began. Because the new supplies would have to come from more distant sources, this would also add to the demand in ton-miles of oil movements in 2003. The OPEC cartel, faced with the high oil prices affecting their potential market share, has been moving to increase production and attempt lowering the crude oil price to the $25 to $28 range, which is optimal for their own sustained competitiveness. The increased loading from OPEC nations will also need larger tanker tonnage and, hence, will increase their demand. The trade journals reported a tripling of the daily charterhire for modern VLCCs to $74,000 per day in early January 2003 from the 2002 average level of $23,300 per day.

The spectacular progress made in the safety of the sea and in the reduction of environmental pollution from tanker operations was documented in a recent study on oil in the sea by the U.S. National Research Council. According to that study, 62% of the petroleum hydrocarbon inputs into North American marine waters during the 1990s came from natural seeps, 32% came from the consumption of oil, and only 3.5% came from the transportation of petroleum (none of which was caused by routine operations such as tank cleaning). The corresponding worldwide figures were 46, 37, and 11.5%. The Kumar and Hoffman study also documented the decreasing loss of ships and lives at sea. However, there was yet another tanker laden with heavy fuel oil that broke and sank off the coast of Spain in November 2002, a stark reminder of the *Erika*



	1997	1998	1999	2000	2001
VLCC, modern	36.6	33.3	19.6	53.4	34.7
VLCC, old	23.8	23.7	11	33.1	19.1
Suezmax, modern	25.1	22	16	40.3	31.9
Aframax, modern	22.8	16.6	13.2	36.7	31.6
Long-range product, modern	19.7	17.2	12.7	31.7	40.6
Medium-range product, modern	12.8	11.1	8.4	16	18.4

FIGURE 4 Average single voyage freight rates. Data from R. S. Platou (2002).

tragedy. The *Prestige*, the ship in question, was a 26-year-old, Japanese-built, Bahamas-flagged single-hull tanker operated by a Greek firm, certified by a U.S.-based classification society, and on charter to a Switzerland-based Russian trading company carrying heavy fuel oil from Latvia to Singapore. This has reenergized the opposition to single-hull tankers in the European Union as well as in the United States. The European Union has been pushing for a ban on all single-hull tankers by 2010 as well as on any tanker more than 23 years old, 5 years ahead of the current mandate. It also has banned such tankers carrying the heavy fuel oil-type cargo from calling at European ports. This will affect the market significantly given that 53% of the world tanker fleet is of single-hull design and 80% of those ships are at least 15 years old.

Although the preceding analysis appears to imply that owners of good modern tankers stand to gain from the ongoing market volatility, there are notable troubling developments as well. Piracy attacks against merchant ships are on the ascent once again and currently cost the global economy $25 billion per year, according to the International Maritime Bureau. The consequences of a piracy attack on a laden oil tanker in a sensitive region could be truly catastrophic, and the economic damages could be phenomenal. Furthermore, the attack on a French tanker off the coast of Yemen in October 2002 is a sad reminder that tanker transportation could easily become a defenseless and innocent victim of terrorism. The recent Al Qaeda warning to slash the economic lifeline of the world's industrialized societies sounds ominously true given the inherent vulnerability of tanker transportation.

SEE ALSO THE FOLLOWING ARTICLES

Crude Oil Releases to the Environment: Natural Fate and Remediation Options • Crude Oil Spills, Environmental Impact of • Marine Transportation and Energy Use • Oil and Natural Gas: Offshore Operations • Oil Industry, History of • Oil Pipelines

Further Reading

Glen, D., and Martin, B. (2002). The tanker market: Current structure and economic analysis. *In* "The Handbook of Maritime Economics and Business" (C. Grammenos, Ed.), pp. 251–279. LLP, London.

Huber, M. (2001). "Tanker Operations: A Handbook for the Person-in-Charge." 4th ed. Cornell Maritime Press, Centreville, MD.

International Energy Agency. (2002). "Key World Energy Statistics 2002." IEA, Paris.

Kavussanos, M. (1996). Price risk modeling of different size vessels in the tanker industry. *Logistics Transport. Rev.* **32,** 161–176.

Kumar, S. (1995). Tanker markets in the 21st century: Competitive or oligopolistic? Paper presented at the First IAME Regional Conference, Massachusetts Institute of Technology, Cambridge, MA.

Kumar, S., and Hoffman, J. (2002). Globalization: The maritime nexus. *In* "The Handbook of Maritime Economics and Business" (C. Grammenos, Ed.), pp. 35–62. LLP, London.

Platou, R. S. (2002). Annual report, R. S. Platou, Oslo, Norway.

Tusiani, M. D. (1996a). "The Petroleum Shipping Industry: A Nontechnical Overview," vol. 1, PennWell, Tulsa, OK.

Tusiani, M. D. (1996b). "The Petroleum Shipping Industry: Operations and Practices," vol. 2, PennWell, Tulsa, OK.

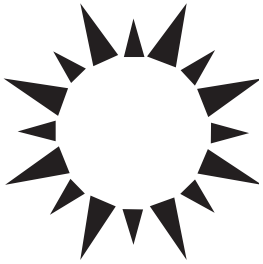
UN Conference on Trade and Development. (2001). "Review of Maritime Transport, 2001." United Nations, New York.

U.S. National Research Council. (2002). "Oil in the Sea: Inputs, Fates, and Effects." National Academy Press, Washington, DC.

Zannetos, Z. (1966). "The Theory of Oil Tankship Rates." MIT Press, Cambridge, MA.

Taxation of Energy

ROBERT BACON
The World Bank
Washington DC, United States



1. **Principles and Issues in the Taxation of Energy**
2. **Taxation of Upstream Energy Production**
3. **Taxation Of Downstream Energy**

Glossary

downstream The stage of use or consumption of energy products.
elasticity The percentage change of a response, such as the quantity demanded, caused by a 1% change in some driving force, such as the price of the good.
excise A fixed amount of tax for each unit sold of the good.
externality A negative externality is created when the actions of one producer or consumer produce an effect that has negative consequences for other producers or households not involved in the decision to take that action.
Ramsey taxes The set of taxes on goods and services designed to raise a given sum of money for the government while minimizing the welfare loss to consumers from the higher prices caused by the taxes.
rent The economic rent of a natural resource is the difference between the price it can be sold for and the full cost of production including the minimum required return on investment.
subsidy The sale of a good below the cost of production financed out of the government budget.
upstream The stage of production of oil, gas, or coal.

Taxes on energy range from those raised on electricity sales, which are generally similar to the general range of taxes on other goods and services, through to taxes on automotive fuels, which in most countries are considerably higher than taxes on other goods and contribute large amounts to government budgets, and finally to taxation on the production of oil (which occurs on a sizable scale in more than 50 countries), which can account for a very large fraction of total government revenue and of the country's national income. A number of factors determine the relative taxation on different energy products; these include factors related to income distribution and equity, to the need to attract investment into the energy sectors, and to the correction of various negative externalities, where the use of a fuel by an individual firm or household leads to adverse impacts on other firms and households. The magnitude of tax receipts from upstream energy production leads to a number of special issues. First, there is the design of the tax structure so as to yield the maximum benefit to the country not only at the present moment but also over the future through its incentives for further investment into the natural resource. Second, the tax receipts need to be transparent to avoid misappropriation, and third, their use and expenditure need to be well planned so as to avoid macroeconomic distortions following ill-founded policies and to preserve equity at the individual and regional levels. The nature of the energy sector is such that the oil sector, both upstream and downstream, has the most distinct tax issues; gas upstream followed by coal is the next most specific, and the taxation of electricity is treated virtually everywhere the same as any other good and raises few distinct issues. The relative importance of these factors is reflected in the amount of space devoted to them in this article.

1. PRINCIPLES AND ISSUES IN THE TAXATION OF ENERGY

The taxation of energy, which involves the transfer of money to the central or local government, related to the production or consumption of energy by a firm or household, is a major source of revenue for most governments in both industrial and developing countries. Taxes are levied both on the upstream (production of primary energy sources, such as crude oil or natural gas) and on the downstream, where the

 © 2004 Elsevier Inc. All rights reserved.

primary source has been transformed or transported to a point where it is consumed (electricity, gasoline). Downstream energy taxes, which are often high compared to the costs of production or the price of a unit of consumption net of tax, can have impacts throughout an economy through raising the price at which energy is sold, as well as providing a source of revenue to finance government spending. Since all consumers, apart from the poorest families in developing countries, and virtually all firms use commercial sources of energy (biomass collected and burnt by households and small farmers being an example of noncommercial and therefore nontaxed energy), this taxation, through both its universality and its magnitude, substantially affects most households directly through the price paid for energy and indirectly through the impact of the cost of energy on the price of other goods purchased.

The structure of upstream taxation on the extraction of oil, gas, and coal is extremely important for attracting investment into these sectors and hence for ensuring that the government can benefit from revenues into the future. Since natural resources often command substantial elements of rent, whereby the sale price is greater than the full cost of extraction, there is a natural tendency for governments to attempt to capture a large share of such rents, but this must be balanced against future receipts, since too high a tax rate will discourage investment and reduce the future tax base.

Taxes, in addition to providing valuable revenue for governments, alter the behavior of firms and consumers as they react to higher costs and prices. This can be used as a policy instrument for reducing certain "external effects" of the consumption of certain forms of energy, whereby damage is caused, to those not using these sources, by the emissions from their combustion.

1.1 Theoretical Issues in the Taxation of Energy

The primary purpose of all forms of taxation, whether on energy or on nonenergy items, is to raise money for the government to finance its expenditures. Taxes finance the majority of government expenditure (the rest being financed by borrowing, which will have to be repaid from future taxation) so that the general level of taxation is determined by the requirements for government spending. However, the same total revenue can be raised by many different combinations of taxes, so that the tax rates on individual items and groups of items can be adjusted to reflect other criteria. First, the balance between direct taxation of incomes and the indirect taxation of goods and services will reflect the ease of collection, with low-income developing countries relying more on indirect taxes, which are easier to collect. Second, the relative tax rates on different goods will reflect the impacts on household welfare. Principles of tax design have been developed, following the seminal work of Ramsey, which suggest that goods with higher price elasticities of demand (sensitivity to a given percentage changes in prices) should attract the lowest tax rates and those with lower price elasticities should attract higher tax rates. This principle can be elaborated to take account of the degree to which consumers will substitute between particular goods as the tax rate of one is raised relative to that on the other. A further consideration for tax design is that, under certain assumptions, it is not optimal to tax goods when they are used as an input to another good (such as the case of fuel oil used for generating electricity) because it is less damaging to welfare to tax the final good itself. The generalized Ramsey rule can then, in principle, be used to calculate a set of tax rates that raises a given sum of money at the least loss of economic welfare to households in the economy. The set of "optimal taxes" that satisfy these principles is in practice difficult to calculate precisely because of the very large informational requirements that it has and also because it rests on a number of assumptions that do not hold exactly in most countries.

Two further adjustments to the principles of determining the relative tax rates on different goods are of considerable practical importance. Because many governments are concerned with equity and income distribution, there is a common suggestion that tax rates on those goods that form a relatively higher share of the budget for low-income households should be lower and tax rates on those goods that are relatively more important for high-income groups should tend to be set relatively high. A final consideration comes from the fact that the consumption and use of various forms of energy, as well as conferring benefits on those that use them, create damage and impose costs (both indirect and direct) on those who do not use them. This consideration has led some governments to raise taxes further on such fuels in order to discourage their use and reduce the negative effects on nonusers (the "polluter pays" principle). This element is referred to as a Pigovian tax. An extension of this principle is that it is socially more beneficial to correct an externality by taxing

the offending good, rather than by subsidizing a substitute good that creates a smaller external effect.

1.2 Issues in Designing Tax Systems

Governments, in moving from general considerations for the raising of revenue to practical issues, face a number of difficulties. Often, the costs of collection of certain taxes would be so high, in relation to the revenue raised, that relying on such a tax would not be cost-effective. For example, the taxation of incomes in many countries is extremely difficult because of the lack of any apparatus to assess incomes. In particular, in developing countries with a large self-employed sector, even identifying who the income earners are would be very costly and then obtaining any independent check on the levels of incomes would be equally difficult. Hence, although higher income groups may be taxed directly, the government will typically look at indirect taxation on goods (including energy) and services as a more effective way of collecting the revenue required. Conversely, in countries with well-developed means of identifying income earners and their sources of incomes, direct taxes play a much more important role for the government's total budget.

The production of energy, as with other goods and services, is also subject to those taxes that are borne by companies. The major forms of such taxes are corporate taxes, which are a tax on profits, and dividends taxes, which are a tax on incomes received from profits. If these are applied at the same rate, irrespective of the sector involved, then the main effect is to raise revenue for the government, rather than to shift the relative supply or demand of one sector versus another. Again, such taxes are most important in countries where accounting systems are better developed.

Indirect taxation on goods and services has several forms. Tax rates can be set as a percentage of the final sales price or of value added; the amounts per unit sold collected from such taxes will rise or fall as prices rise and fall. Alternatively, taxes can be set as fixed amounts per unit (excise duty) that do not change with prices. The use of a system of value-added taxes (VAT) requires accounts to be drawn up by all firms in the production chain, which again is difficult to monitor in less advanced countries. Hence, VAT systems are found mainly in more advanced economies, whereas sales taxes, which require the identification only of final sellers, are relatively more common in less developed countries.

Downstream taxation of oil products, whether or not the country is an oil producer, tend to attract the general sales tax rate (or VAT) levied on all other goods, plus an additional excise duty, which usually varies among them. The use of this additional tax component on oil products can be justified by several arguments: there are easily identifiable points of sale; they have relatively low price elasticities; in some countries they are of larger importance for higher income groups than lower income groups; and there are associated negative externalities from their combustion. The use of an excise duty (a fixed sum per unit sold) also gives some stability to this important source of revenue, which otherwise could experience major variations because of the volatility of the sales tax per unit received caused by large fluctuations in the world oil price.

Governments raise revenue from the taxation of upstream energy, especially oil, through a number of instruments, not all of which are called taxes, but which function in largely the same way as taxes in terms of revenue raising. These include royalty payments, normal company taxes, bonus payments, profit oil, and government participation sales. The particular forms involved in the oil fiscal system, because they have different impacts on returns to the companies and to the government in varying circumstances (primarily the costs of production, the output of oil, and the price of oil), will determine the willingness of the private sector to invest in the sector and hence the government's ultimate access to revenue from this source. If the total tax rate is set very low, there will be plenty of investment but little tax revenue, whereas if the tax is set very high, there will be little investment and again little tax revenue. Good tax design attempts to ensure a fairly steady tax revenue stream, which will be high and robust against a variety of circumstances. Upstream taxes are unlike most other forms of taxation in that some elements applying to a particular project are part of a bargain struck before the investment takes place and are not usually changed during the life of that investment. They can, however, vary from project to project depending on the particular factors involved at the time of the negotiation. Upstream tax elements of general application have tended to be changed frequently as profitability has changed over time.

1.3 Taxes and the Government Budget

Energy taxes are an important source of revenue in virtually every country and, for countries that are oil and gas producers, they can be the dominant source

of government revenue. Data covering a large number of countries are not available, but information shown in Table I is indicative. For those of the lower income countries that were not oil producers and that had no upstream tax revenue, the taxes on petroleum products accounted for as much as one-quarter of all government revenue, whereas for the higher income countries, even if they were oil producers, downstream oil product tax revenue accounted for approximately 5% of total tax revenue. The United States was exceptional in its share being under 2%. This reliance on petroleum products for taxation also explains why non-oil-producing governments can be strongly affected by variations in the international price of crude oil and oil products. Since the elasticity of demand for such products is typically low—a price fall of 10% might lead to a rise in consumption of not more than 5%—a fall in oil prices would lead to a fall in government revenue if the tax rate were proportional to prices. Hence, governments usually impose a large fixed excise-duty element to ensure a more steady revenue stream. The disadvantage of excise duties is that they do not increase with a general increase in the price level (inflation) so that governments must adjust the rate periodically to ensure that tax revenue stays constant in terms of its purchasing power.

TABLE I

Tax Revenue from Domestic Consumption of Petroleum Products

Country	As % of National Income	As % of Government Revenue
Argentina	1.9	13.4
Bangladesh	0.8	9.0
Cote d' Ivoire	3.5	18.2
France	1.8	5.5
Germany	1.5	5.3
Ghana	3.3	22.2
India	1.5	10.2
Japan	0.6	4.7
Kenya	2.5	9.9
Mexico	1.5	6.8
Paraguay	0.9	7.6
Philippines	1.2	6.7
Senegal	3.6	22.7
Sri Lanka	2.1	11.2
Sweden	1.4	3.1
Uganda	2.2	29.5
United Kingdom	1.8	4.4
United States	0.4	1.8
Zambia	1.3	7.9

Source. Data derived from Gupta and Mahler (1994), with permission.

Taxes on energy products are not usually hypothecated—they are not earmarked for specific uses—but rather go into the general budget, together with all other tax revenues, for allocation to the various items of government expenditure. In some countries, taxes on transportation fuels are linked directly to government expenditures on road maintenance and construction, as a way of ensuring that those who use roads most pay directly for some of the costs of their use of the road system. This arrangement is based on a crude approximation between tax paid and the wear caused, because higher fuel consumption will be associated with heavier vehicles as well as with greater distance traveled.

Upstream sources of energy are subject to distinct forms of taxation. For countries that produce oil or gas, the importance of tax revenues from these activities can be overwhelming. For coal, where rents are altogether much smaller in most cases, the tax receipts from royalties are important but not dominant in the economy. Most of the world's large coal producers are also large economies so that the relative size of the coal sector is less dominant than is the oil sector in a number of the small economies where oil is produced.

Although details of individual upstream oil taxes (and other fiscal charges) paid are often confidential, and governments in many developing countries are reluctant to publish figures on their total oil revenues, some estimates are available. Table II shows the huge dependence of certain countries on upstream oil taxes for the government budget and also their dominant role in the macroeconomy. The availability of such large sums of money, which will last only for the life of the oil fields, poses very special problems for the government. The size of these revenues, their sensitivity to variations in the world oil price, and their temporary nature have led governments in a number of oil-producing countries to establish special "oil funds" into which part of the revenue is paid. Such funds, which may be purely a bookkeeping exercise in the general budget and not a separate fund in a true sense, have been used for budget stabilization and for savings ("future generations fund"). A stabilization fund accumulates money from oil tax receipts when the oil price is higher than some projected average value and releases funds into the budget when the oil price is lower than some value. This has the

TABLE II

Tax Revenues from Upstream Oil and Gas for Selected Countries

Country	As % of government revenue	As % of national income
Algeria	65	30
Angola	70–90	>60
Cameroon	20	<5
Congo-Brazzaville	60–80	>50
Gabon	60	40
Indonesia	25	10
Mexico	30	2
Nigeria	70	40
Norway	15	10
Venezuela	55	28

Note. Values are estimates from various sources including McPherson (2002) for the period around 2000.

effect of allowing the government expenditure to be stabilized even when oil tax revenues vary. Oil savings funds are designed to accumulate funds and invest them separately in a responsible fashion, so as to provide an income for the country at a later date when the oil starts to run out or when the country can spend the money more effectively. The amount of oil tax revenue going to the budget will depend on whether the rules for accumulation into such funds are mandatory or optional. Similarly, the rules for withdrawal are specified in different ways in different countries, although the government can always change the rules to meet pressing needs. Another way in which upstream revenues are earmarked is seen in countries where there is a formula for returning part of the revenue to the regional or local government in the area where the extraction is taking place. This is done in order to give local communities, who feel a particular ownership of the resource and are directly impacted by its extraction in their neighborhood, some direct financial benefits from the revenues collected by allowing them to have control over how it is spent in their region.

Where there is a formal federal system (e.g., Canada or Nigeria), the tax revenues can be channeled through the regional governments, which may involve significant dialogue and negotiations. Where there is no regional government, the actual channeling and use of funds according to local wishes create a need for an extra mechanism for dialogue between the central government and local communities.

1.4 Energy Taxes and Environmental Concerns

The combustion of fossil fuels and of biomass creates a number of external effects, which societies are becoming more aware of and more anxious to mitigate. The selection of different tax rates for fossil fuels is one of the most important policy instruments to achieve such an effect. The three classes of activity that have received the most attention with regard to the treatment of externalities are the use of gasoline and diesel as automotive fuels, the use of coal and fuel oil to generate electricity, and the use of biomass for cooking and heating.

Fuels used in small low-stack sources, which are primarily vehicles but also include small diesel generators, have both local and global environmental impacts. The emissions of fine-particulate matter, sulfur dioxide, and oxides of nitrogen (which are also ozone precursors) have adverse impacts on health, especially in mega-cities, where the volume and density of traffic are very high. In particular, the incidence and severity of respiratory diseases are associated with particulate emissions. Policies to improve fuels' qualities are being widely implemented in countries with serious urban air pollution, with the United States and the European Union taking the lead in pursuing the best available technology. For example, cities such as Delhi, Los Angeles, and Mexico City have suffered from deteriorating air quality and as a result have taken aggressive action to monitor and reduce air pollution. However, particularly in the European Union, the use of increased taxes on fuels, as well as raising extra revenue for governments, has been seen as a way to reduce or slow the growth in their consumption and hence to improve the levels of urban air pollution. In addition, any such reduction in fuel use will also reduce the emission of carbon dioxide, which is a greenhouse gas and is associated with global warming, which can have effects on societies far in time and place from when and where the original combustion took place. Since raising taxes on automotive fuels also reduces consumer welfare, there must be a balance between these two effects, so that tax rates reflect the relative values a society places on these factors.

Additional and important externalities of vehicle use come from the impacts on congestion, on the incidence of accidents, and on the damage to roads from their use. Greater vehicle use imposes costs on other users in terms of time taken to drive to a particular location (especially at peak times) and on

the number of accidents caused, some of which are fatal. Because the decision to drive, use another form of transport, or stay at home depends on the costs to the individual traveler, but not on the costs imposed on others, fuels taxes have been seen as a corrective to these problems also. In all these cases, there may be policies that impact more directly on the relevant factor (e.g., a tax on emissions rather than fuel used, a tax on axle weight, or a charge for driving at peak times) but the simplicity of imposing a tax on the fuel itself, which then links to all these concerns, has thus far meant that policymakers have only just begun to make use of such more finely tuned corrective schemes.

Estimates for the optimal gasoline taxes in the United States and the United Kingdom that take most of these factors into account have been made. This tax, together with other taxes, is required to raise a fixed sum of money for the government while maximizing the welfare of society as a whole. Although the calculations are based on a number of assumptions about individual parameters, such as the economic cost of a fatal accident, Table III gives some important results that are fairly robust to variations in the assumptions. These calculations indicate that the optimal U.S. excise tax rate (exclusive of federal and local sales taxes) is approximately twice the actual rate, whereas in the United Kingdom it is approximately half the actual rate. The sizes of the various components indicate that the extra amount of tax required to make the optimal adjustment for global pollution effects is small, whereas that for local pollution is substantial (approximately 15% of the optimal excise tax would be for this component). The amounts related to the direct externalities caused by congestion differ substantially between the two countries, reflecting their individual characteristics, but are large in both. The component designed to reduce the use of fuel, and hence reduce the incidence of accidents, also accounts for a large fraction of the tax. The pure Ramsey component, which relates to raising finance for the government (over and above that already raised through the component designed to discourage pollution) in such a way as to minimize the welfare losses to consumers, is approximately 25% of the total tax rate. A particularly important effect in calculating the optimal tax is shown to be the incorporation of the impacts of increased fuel taxes on the fuel efficiency of vehicle miles driven. Evidence suggests that fuel efficiency per mile does increase substantially as fuel taxes are increased, so that taxes need not be as high to combat pollution as they would be if there were no response in fuel efficiency. These calculations also suggest that gasoline taxes are much higher than the general level of taxation of goods, so that the relatively small extra component required to reduce consumption in line with damage from global warming can be easily accommodated.

TABLE III

Illustrative Calculations for Optimal Gasoline Taxes in the United States and United Kingdom (in Cents per Gallon at U.S. 2000 Prices)

Tax component	United States	United Kingdom
Pollution, fuel-related (global)	5	5
Pollution, distance-related (local)	16	17
Congestion	30	68
Accidents	24	20
Ramsey component	26	23
Optimal tax	101	134

Source. Data derived from Parry and Small (2002), with permission.

Diesel fuel, when it is used solely as an input to a further economic activity (as for the transportation of materials and as a fuel in power stations), would not need be taxed if these final products were taxed and if there were no externality associated with the fuel combustion. However, diesel does create substantial emissions that create local damage (especially from vehicles) so that considerations similar to those for determining the level of gasoline taxation apply to the taxation of automotive diesel. One particular issue that concerns policymakers in many countries is the relative tax rates for gasoline and for diesel. Gasoline taxes are usually higher, which leads, over the longer run, to a substitution away from gasoline-fueled vehicles toward diesel-fueled vehicles. Since the congestion- and accident-related effects of the two are similar, attention is then focused on their relative pollution effects and in particular on whether the local damage from diesel might not be greater than that from gasoline. More important may be the fact that diesel is also used as an intermediate good, as for commercial and industrial vehicle transport, which leads to a focus on taxing the final good rather than the intermediate good and hence to setting lower tax rates on diesel than on gasoline.

Environmental concerns also impact attitudes regarding the taxation of coal, fuel oil, and natural gas. Coal and fuel oil are consumed largely as inputs to electricity and, according to pure Ramsey-type principles, need not be taxed in that use if electricity

were taxed and there were to be no external effects from their combustion. However, both fuels create strong negative externalities in terms of their emissions, unless very substantial and expensive techniques are used for removing the pollutants. Again, taxes on these fuels can be used to encourage the switching to another less polluting fuel. Since natural gas is very much less polluting than these other fuels, the relative tax rate on its use can be adjusted downward to encourage its use in situations where it would not otherwise be used.

In some countries, such as South Africa and China, household consumption of coal is still substantial and has large negative external effects, but the desire to support the local industry has thus far kept the governments from discouraging the use of coal through higher taxation.

A final reason for lowering the taxes on those fuels (and electricity) used for cooking is that, in countries where biomass is available, poorer households often use biomass as the main fuel of choice. This creates very considerable indoor air pollution, which has large adverse health effects for women and children from inhalation, and, although this is not an externality that requires correction through a "making the polluter pay" approach, it can be seen as a demerit good where users do not fully understand the consequences of their own actions. In the general case of demerit goods, taxation is often used to discourage households from using those goods whose consumption is in fact bad for their own welfare, as in the case of tobacco. However, biomass has two distinct characteristics: first, it can be an absolute necessity for the poorest households, and second, it is often not a marketed good. In rural and periurban areas, households often collect the biomass themselves (rather than purchase it directly) since it is free in monetary terms (although it may takes much time and effort on the part of women to collect the fuel). This can lead to gradual deforestation (as in Haiti or parts of Indonesia), which in turn has adverse ecological effects. A strategy of taxing the offending good (biomass) and not reducing taxes on the substitute goods (e.g., kerosene) is not available because biomass is either not marketed at all or else is sold through very informal systems. Hence, policies to encourage a switch from the use of biomass to commercial fuels such as kerosene, through lower tax rates or even subsidies, have been justified as a way to reduce this environmental damage and to improve the health of these households. The effectiveness of such a scheme is limited in cases where households are so poor that they are unwilling to purchase any commercial fuel (at prices sustainable for the government) rather than to use nonmarketed biomass.

1.5 Taxation and Income Distribution

Any system of taxing goods and services is likely to have differential impacts on households of different income levels. In industrialized countries, the majority of households own cars, so that a tax on automotive fuels is likely to be regressive because expenditure on them will take a larger share of the household budget for lower income households. However, in developing countries, lower income households neither own cars and directly consume transportation fuels nor are connected to the electricity grid. Thus, raising taxes on these products in developing countries is likely to result in an improvement in income distribution, because the better off may pay more in both absolute and relative terms. Indeed, with such a case, equity can be seen as an extra factor to be taken into account following the analysis described in the previous section. In developing countries, this might justify tax rates on gasoline being further raised. However, the position of the taxation of other petroleum products is more sensitive to the importance of the distribution of income. For example, diesel-fueled transportation (buses and trucks) is important for poorer urban households and nonautomotive fuels (liquified petroleum gas and kerosene) are important for lighting and cooking for a wide range of the lower to middle income groups in many developing countries. Nonautomotive fuels are not associated with creating congestion and economic loss or with high levels of local pollution, so that tax rates on these fuels would be expected to be lower than that for gasoline. In addition, their greater importance in the budgets of lower income (if not the poorest) households in developing countries has led governments concerned with income distributional consequences to lower tax rates on these items, in line with a generalized Ramsey approach, which indicates that there should be lower taxes on those goods consumed relatively largely by the poor and higher taxes (so as to obtain the same overall revenue) on those goods consumed more intensively by better off households. Indeed, in many developing countries, these products are often subsidized by the state (effectively negative taxation, where the state returns some money collected from other taxes to the consumers of these goods).

Attempts to introduce substantially different tax rates on the various petroleum products, where the

costs of production are fairly similar, can lead to sufficiently large retail price differences such that fuel adulteration starts to occur. Consumers or retailers may dilute diesel fuel with kerosene, which loses the government revenue, or, more seriously, dilute gasoline with kerosene, which increases engine deposits and emissions. Similarly, the situation where fuel destined for export is not subject to taxation, yet fuel for domestic consumption incurs heavy taxation, leads to fuel designed for export to be diverted illegally to the domestic market. Large differences in tax rates between neighboring countries can lead to smuggling and illegal sales, which avoid the payment of such taxes.

1.6 Political Economy of Energy Taxation

Decisions on how to tax upstream or downstream energy supplies are taken by governments and reflect the importance of the various factors outlined above. Different groups in society attach different degrees of importance to these factors and accordingly would wish to see taxes reflect this. For downstream production (electricity, petroleum products) where the price, net of tax, would be close to the total costs of production in competitive markets, the imposition of taxes then transfers money from one group (the users) to whichever groups benefit from tax expenditures. In addition, the reduction in consumption and local pollution resulting from the higher prices may benefit a different group within the country and, through any impacts on global warming, may benefit people outside the country.

Governments that have historically subsidized or set disproportionately low taxes on energy products, as a way of supporting poorer members of society (as in many developing countries), have often later realized that the cost of such a program is huge and have wished to reduce the financial burden to the budget of the subsidies by raising prices to cover costs. Table IV indicates that for eight of the largest energy-using countries outside the industrialized world: (1) the average subsidy on energy is approximately 20% of the price of energy and (2) removing these subsidies by this amount and allowing consumer expenditure to be redistributed to other goods would increase the countries' gross domestic product by approximately 0.75%, reduce energy consumption by 13%, and reduce carbon dioxide emissions by approximately 16%. Nevertheless, the visibility of energy prices and their importance to groups in society that are politically well organized and articulate (such as motorists' organizations) has meant that governments have found it difficult to reduce subsidies or increase taxes by substantial amounts. In several developing countries, programs to increase fuel prices by reducing subsidies or increasing taxes, without elaborate public relations campaigns, have resulted in widespread demonstrations, with the result that the proposed increases have had to be substantially cut back or even canceled.

The political economy of upstream oil taxation is distinct, because of the fact that the strategy of the Organization of Petroleum Exporting Countries (OPEC) is to act as a cartel and to adjust the production of oil so as to ensure that the world price

TABLE IV

The Results of Removing Energy Subsidies

Country	Subsidy as % of price	Gain as % of GDP[a]	% Reduction in energy consumption	% Reduction in CO_2 emissions
China	10.9	0.37	9.4	13.4
Russia	32.5	1.54	18.0	17.1
India	14.2	0.34	7.2	14.2
Indonesia	27.5	0.24	7.1	11.0
Iran	80.4	2.22	47.6	49.5
South Africa	6.4	0.10	6.4	8.1
Venezuela	57.6	1.17	24.9	26.1
Kazakhstan	18.2	0.98	19.2	22.8
Average	21.2	0.73	12.8	16.0

Source. Data from International Energy Agency (1999), with permission.

[a] GDP, gross domestic product.

of crude remains well above the full cost of production, and because of the fact that the costs of supply vary so much between countries. These have resulted in a situation in many countries where there is such a large element of economic rent (the difference between the full cost of production and the selling price) that there is a natural tendency for many parties to attempt to capture it. The state, which in most cases is the owner of the unexploited oil and gas, either can develop the oil itself (as in the case of Mexico) and capture the rent directly or can sell the rights to produce the oil and capture some of the rent through this transaction and, by taxing these companies, transfer further rent through the tax system. Further bargaining within the state, as between central, regional, and local governments, over who actually receives the taxation can take place, and in countries with poor governance and lack of transparency, the rents may be captured by parties other than the government itself, through various corrupt practices.

2. TAXATION OF UPSTREAM ENERGY PRODUCTION

2.1 The Structure of Upstream Energy Taxation

The structure of taxation on the extraction of oil, gas, and coal, and the factors influencing the choice of the tax instruments, which enter contracts between the owners of the resource (the government) and the producers of the resource, can be analyzed with respect to not only the impact on the returns to the two parties but also the impacts on the risks carried by each party. In the exploration and production of oil, gas, and coal, there are some very important risks to be borne that impact strongly on the contracts formed and on the revenue (tax) flows to the government. First, there is the risk involved in finding commercially viable amounts of the resource. Exploration and production are typically very expensive, yet total costs are not known with certainty until after the contract has been signed and indeed until all after the costs have been incurred. The life and production of the well or field also carry risk, and although this is often ameliorated because of continuous improvements in technology, the total output is uncertain until the closure of the field. Equally important is the fact that, over the life of the field, the price of the resource will change and is highly uncertain, if past experience is a reasonable guide. Oil prices have seen large fluctuations, subsequent to the price hikes of the mid-1970s and early 1980s, which relate to several factors and make forecasting over a period of several years very imprecise. With these large uncertainties, each party will wish to set the tax parameters so as to obtain a good average return while also minimizing the risks that their return could be very poor. Governments need to provide sufficient incentives to the private sector to continue to invest in exploration and production to ensure that future tax revenue is secured.

2.1.1 The Taxation of Oil Production

The taxation of oil production has two broad forms, with a large number of variations and specific tax instruments. One group of countries, which are predominantly industrialized, treat the oil produced as the property of the private sector company producing the oil and use a combination of a royalty or revenue tax and profit-based taxes. A royalty or revenue tax, which is a percentage of the sale price of the oil, is simple to administer but can have strong disincentive effects for investment. As the price of oil falls or costs rise, the share of royalty payments rises as the level of profits falls and this can result in severe regressivity and even a negative posttax rate of return. As a result, some countries have reduced or even abandoned royalties and revenue-based taxes and rely solely on profits-based taxes.

Many developing countries that produce oil have entered into production-sharing agreements (PSAs), whereby the private sector companies produce the oil on behalf of the government. They sell the oil and keep that part of the proceeds that remain after taxes have been levied. PSAs tend to have four major elements, although there are many variations on the main theme. A schematic description of such arrangements is as follows: the company producing the oil sells it at the world market price. On this price, whatever the level, it pays a fixed percentage per barrel produced to the government—the royalty. This arrangement ensures that the government receives an amount that is irrespective of the costs of production, so to this extent the producing company bears the risk. To give the company some certainty with respect to recouping its costs, it is allowed to keep in the current year a fraction of the postroyalty receipts ("cost oil") as a contribution toward covering costs. This arrangement continues through the life of the project until all allowable costs have been recovered. Out of the money received from the oil sale, net of royalty and cost oil, which is termed "profit oil," there is an agreed split between

the company and the government. The company then pays income (corporation) tax on its profit oil. The government's share of receipts, net of the cost oil, when calculated on a discounted basis over the full life cycle of a project, defines its "take" on the project. The take depends not only on the value of the tax parameters, but also on the time profile of expenditures and of the allowed cost recovery. When the corporation tax rate is fixed, because of its applicability to all sectors, the factors that are variable are as follows: the percentage of royalty oil; the amount of cost oil that can be charged per barrel (until all costs have been recouped); and the share of profit oil. The average government take for oil appears to be in the range of 60–70%, but varies between a low of approximately 40% to a high of approximately 90%. These variations reflect the risks in individual countries as well as cost differences between countries.

2.1.2 The Taxation of Gas Production

Generally, gas, which is often produced jointly with oil, is taxed in a fashion similar to that for oil. Again, where gas is sold, there is usually a substantial element of rent so that the government will wish to capture its share of this via the tax system. However, one aspect of the gas market is markedly different from that of oil. Gas is usually sold on a long-term contract basis, reflecting the fact that it is not as easily transported (much being by fixed pipelines) and there is not yet a flexible international market for gas. With such contracts, much of the price uncertainty is removed, so that the contract structure can reflect the higher degree of certainty experienced by both parties. The average government take for gas is similar to that for oil, although the revenues received from oil tend to be greater than those from gas when both are being produced in a country.

2.1.3 The Taxation of Coal Production

Coal mining is usually subject to a royalty and the national corporation tax on income, net of allowable costs. The former is sometimes in the form of an excise tax of a fixed value per ton mined. Other cases are in the form of a revenue tax, which relates the tax payable to the value of sales (often approximately 10%). Both can have strong disincentive effects if the rates are set too high because, with falls in the price of coal or increases in the costs of production, the return to the producing firm can even become negative. Countries with federal systems, such as Canada, impose federal and provincial company income taxes as well as provincial royalty payments.

Coal production in some countries, notably within the European Union, has been subsidized by a number of fiscal devices in order to protect employment. This reduces the effective tax rate and the net money collected from the sector.

2.2 The Monitoring and Governance of Upstream Tax Revenues

The magnitude and individuality of tax revenues from upstream oil and gas have encouraged some governments to conceal what their total revenues from these sources are. The fact that typical tax arrangements in many developing countries are individually negotiated through contracts with producing companies makes it difficult for parties not directly involved to know what the revenues should be, unless they have full access to these contracts. When this is aggregated over a number of wells and fields, the picture can be obscured further. Unless all companies involved in a country publish their tax payments, the society as a whole may have very little idea of the magnitude of total revenue from oil and gas, unless the government chooses to publish it and is truthful in doing so. There are movements to encourage greater transparency of oil tax revenues through the publication by companies of taxes paid. Schemes that allow the aggregation of individual company payments on a credible but confidential basis, with publication of the totals, are likely to prove more acceptable than individual publication and almost as effective in promoting accountability.

The size of these payments gives incentives for misappropriation and misuse, which are facilitated by the confidentiality of the negotiations. In order to bring increased transparency to this problem, and hence to reduce the possibility of misappropriation, several countries have established or are establishing "oil funds." Rules are set for transferring some or all revenues (the amount based on a formula) to a separate fund, whose inflows and outflows are monitored separately from the overall budget and whose magnitude may be published. The purpose of these funds, whether for normal budgetary support, for budget stabilization when oil revenues fluctuate, or for saving for future generations, is also made public. Countries and regions that have established such funds include Alaska (United States), Alberta (Canada), Azerbaijan, Chad, Kazakhstan, Kuwait, Norway, Oman, and Venezuela. In some of these cases, such as Norway, the fund is actually invested and managed offshore to ensure the best possible return for the country and extensive details are

published. In other cases, such as Kuwait, the actual size of the fund is not made publicly available. Increasingly, civil society is pressurizing governments to become more transparent and the establishment of visible funds is seen as a step toward ensuring this.

2.3 The Use of Oil Tax Revenues

Where upstream tax revenues are earned that are large relative to the size of an economy, as is the case in several oil-producing countries, then experience has shown that these revenues can have very substantial and even negative effects on the rest of the economy. In large producing countries, such as Nigeria and Indonesia, despite huge tax revenues received over a number of years, the rate of economic growth has not increased and the level of poverty has stayed high. Even in industrialized countries, sudden "windfalls" following the discovery and exploitation of oil and gas have led to adverse effects. The Netherlands, following the development of large gas reserves, initially experienced contraction of other sectors (and job losses) when the increase in the exchange rate caused by the export earnings from gas made other goods uncompetitive (the "Dutch disease" phenomenon). Because oil and gas have tended to be primarily export activities, they generate little direct local employment, so that the oil tax revenues need to be the vehicle of development for economies. Hence, in many countries, even when the government has received and spent the revenues, the ensuing economic development has been very disappointing, leading to the notion of a "paradox of plenty" for oil and gas producers. For example, the average real growth rate of national income of the OPEC group of countries, in the 20 years following the first large oil price rise of the early 1970s, was lower than that in the decade preceding that price rise. This weak economic performance has been associated in some countries with a worsening of the political climate and an increase in corruption. Even when there is transparency as to the amount of the tax revenues, there is still difficulty in determining how to spend the money effectively for the future development of the economy. A spending program that boosts only current consumption, by transferring large sums to the consumers (possibly through subsidized goods and services), is not likely to raise the growth rate or to provide infrastructure for the benefit of future as well as present generations. In developing countries, the sudden discovery of oil or gas and their attendant tax revenues can provide the government with more money than can be effectively spent in the short run. Examples of "prestige" projects that do not help future development (highways that have substantial excess capacity or steel factories for whose products there is very little market) effectively waste the money. Moreover, excessive spending can drive up local prices very rapidly, which helps damage any existing industries. The planning capacity of some governments to spend such sums is also limited, so that a more gradual spending approach, while accumulating some of the tax revenue in a future generations fund, can reduce the impact of these problems.

3. TAXATION OF DOWNSTREAM ENERGY

3.1 General Levels of Downstream Energy Taxation

Information on the taxation of downstream energy use, apart from automotive fuels, is sparse except for countries within the Organisation for Economic Cooperation and Development (OECD). Table V indicates the range of taxes (both excise and VAT/sales taxes) as a percentage of final retail price including tax, charged within the OECD. Although there are often single extreme values that widen the range greatly (for household electricity, the highest tax rate is that of Denmark at 60%, whereas the next highest is Norway at 36%, followed by Austria at 28%), the overall picture is clear. Where fuels are used

TABLE V

Ranges of Taxes as a Percentage of Final Prices for Energy Products within the OECD in 2000

Fuel	Typical range
High-sulfur fuel oil for industry	5–25%
Low-sulfur fuel oil for industry	4–22%
Heavy fuel oil for electricity	5–23%
Light fuel oil for industry	3–52%
Light fuel oil for households	10–60%
Diesel for commercial use	30–70%
Premium leaded gasoline	52–72%
Premium unleaded gasoline (95 RON)	30–76%
Natural gas for industry	1–13%
Natural gas for households	5–50%
Electricity for households	5–60%

Source. Data from International Energy Agency (2001), with permission.

as inputs to other goods, tax rates tend to be lower, and where they are for final use, tax rates are higher. Fuels that have lower emissions characteristics (natural gas) also tend to have somewhat lower tax rates.

3.2 The Taxation of Automotive Fuels

There is extensive information on the worldwide taxation of automotive fuels that is obtained by surveying pump prices and then using the approximation that, net of tax, prices in all countries are the same due to international competition in the fuels market (there are, in fact, some landlocked non-oil-producing countries that experience important differences in pretax costs because of large costs of importing and transporting fuels). Differences in pump prices then represent a measure of variation in the taxation of gasoline and diesel. Table VI shows the results of a global survey. Using the approximation that the untaxed retail price of gasoline in the fourth quarter of 2000 would have been approximately 32 U.S. cents per liter, some 22 countries (mainly oil producers) actually sold below this price, with Turkmenistan being the lowest at 2 cents per liter; 25 countries had prices between the "break-even" level and the U.S. level of 47 cents per liter (tax of 15 cents per liter); 59 countries had prices between the U.S. price and the lowest of the European Union countries (the Irish Republic) at 72 cents per liter (tax of 40 cents per liter); and 52 countries had prices above the lowest EU level, with Hong Kong charging the highest price at 146 cents per liter (tax of 114 cents per liter). For diesel prices, the picture was very similar, although the spread of taxes was rather smaller, with 40 countries selling diesel below cost, and with the highest price being in the United Kingdom at 122 cents per liter. The overall average tax for diesel was approximately 20 cents per liter, whereas that for gasoline was approximately 28 cents per liter. The huge variation in tax rates, which dominate intercountry differences in pretax prices, indicates how the relative importance of the different reasons for taxation varies among countries. Governments with little need for revenue, such as certain oil producers, have very low taxation or even subsidize automotive fuels, whereas other countries, such as most of those within the European Union, are willing to impose high taxes, presumably for environmental reasons. The widespread tax difference between diesel and gasoline, averaging approximately 8 cents per liter, is large in relation to the pretax price and suggests a deliberate policy of differentiating between these automotive fuels in favor of diesel. This is most likely because of distributional criteria since the costs of the congestion- and environmental-related aspects are probably fairly similar for the two modes of transport. The tax difference may also reflect attempts to encourage the use of buses over cars and the desire not to tax commercial and industrial transport too heavily, regarding them as intermediate activities where the final good they help to produce should instead bear the taxation.

When tax rates are compared as between countries according to their level of development, a clear pattern exists. Table VII shows that tax rates in the

TABLE VI

The Distribution of Posttax Retail Gasoline and Diesel Prices in 2000 (in U.S. Cents per Liter)

Range	Number of countries in range for gasoline prices	Number of countries in range for diesel prices
Below retail cost	22 (<32¢)	40 (<30¢)
Between retail cost and U.S. price level	25 (32–47¢)	53 (30–48¢)
Above U.S. level and below minimum EU level	59 (47–71¢)	11 (48–53¢)
Above minimum EU level	52 (>71¢)	54 (>53¢)

Source. Data from Metschies (2001), with permission.
Note. Data are for the fourth quarter of 2000.

TABLE VII

Average Product Taxes and Prices in OECD and non-OECD Countries in 1999

Fuel and country group	Tax as share of final price (percentage)	Tax (U.S. cents per liter)	Net price (U.S. cents per liter)	Gross price (U.S. cents per liter)
Gasoline				
OECD	67	58.1	25.2	83.3
Non-OECD	44	22.9	26.9	49.7
Automotive diesel				
OECD	59	42.4	25.0	67.4
Non-OECD	40	16.6	22.8	39.4
Kerosene				
Non-OECD	23	5.1	12.3	17.4

Source. Data from International Energy Agency (2001) and Bacon (2001).
Note. Table is based on data from 22 OECD and 37 non-OECD countries.

countries of the OECD are substantially higher than those in non-OECD countries, with the difference being more pronounced for gasoline. Kerosene taxes are much lower than automotive fuel taxes in the non-OECD countries, both in absolute terms and in percentage terms, reflecting the desire of governments of less developed countries to keep kerosene affordable for a large range of households.

Acknowledgments

The views expressed represent those of the author and not of the World Bank. The author thanks Katherine Gratwick, Masami Kojima, and Charles McPherson for helpful comments on an earlier draft of this article.

SEE ALSO THE FOLLOWING ARTICLES

Carbon Taxes and Climate Change • Clean Air Markets • Economics of Energy Demand • Economics of Energy Supply • External Costs of Energy • Fuel Economy Initiatives: International Comparisons • Subsidies to Energy Industries

Further Reading

Bacon, R. (2001). "Petroleum Taxes: Trends in Fuel Taxes (and Subsidies) and the Implications". Viewpoint 240, Private Sector and Infrastructure Network The World Bank, Washington, DC.

Bindemann, K. (1999). "Production Sharing Agreements: An Economic Analysis, Working Paper WPM 25." Oxford Institute of Energy Studies, Oxford, UK.

Gupta, S., and Mahler, W. (1994). "Taxation of Petroleum Products: Theory and Empirical Evidence, Working Paper WP/94/32." Fiscal Affairs Department, International Monetary Fund, Washington, DC.

International Energy Agency. (1998). "Coal Licensing and Production Tax Regimes." International Energy Agency, Paris.

International Energy Agency. (1999). "Looking at Energy Subsidies: Getting the Prices Right," World Energy Outlook 1999. International Energy Agency, Paris.

International Energy Agency. (2001). "Energy Prices and Taxes: Quarterly Statistics." International Energy Agency, Paris.

McPherson, C. P. (2002). Petroleum revenue management in developing countries. *In* "Paradox of Plenty: The Management of Oil Wealth", Report 12/02 Centre for Economic Analysis, Oslo, Norway.

Metschies, G. P. (2001). "Fuel Prices and Vehicle Taxation", 2nd ed. Deutsche Gesellschaft für Technische Zusammenarbeit (GTZ) GmbH, Eschborn, Germany.

Parry, I. W. H., and Small, K. A. (2002). "Does Britain or The United States Have the Right Gasoline Tax?" Resources for The Future, Washington, DC.

Technology Innovation and Energy

AMBUJ D. SAGAR
Harvard University
Cambridge, Massachusetts, United States

1. Introduction
2. Energy Research and Development
3. Energy Technology Deployment
4. Conclusion

Glossary

energy innovation system The web of actors involved in energy technology innovation as well as the linkages, partnerships, and relationships among them.

energy technologies Technologies that help to exploit existing and new sources of energy (extraction and supply technologies), convert these to readily usable forms of energy (conversion technologies), and use these forms of energy to satisfy human and economic needs (end-use technologies).

energy technology innovation Development and widespread introduction of new or improved energy technologies; the process of innovation includes research, development, demonstration and deployment activities.

Energy technology innovation encompasses research and development on new technologies, as well as on improving existing technologies, and bringing these into widespread practical use. In fact, technological change, driven by a range of innovation efforts, has long played a central role in shaping the global energy system and its components. New and improved technologies have helped to expand energy supplies and exploit them in an economical fashion. For example, significant improvements in exploration and recovery technologies have been responsible, in part, for the increases in recoverable hydrocarbon resources over the past few decades. Energy technology innovation has provided cheaper, more efficient, and cleaner ways of converting energy into desirable end-use forms as well as a range of technologies to better satisfy consumer needs. The reductions in global energy intensity over the past decades have been driven, in part, by the development of more efficient energy conversion and end-use technologies. Technological developments, spurred by environmental regulations, have also helped to reduce the local and global environmental impacts of energy extraction, conversion, and use.

1. INTRODUCTION

Expenditures on energy account for approximately 7 to 8% of gross domestic product (GDP) worldwide. In addition, annual global investments in energy supply technologies run into the hundreds of billions of dollars, and trillions of dollars are spent annually on products that consume energy. Furthermore, the global energy system is expected to expand substantially during the next few decades, with the bulk of these increases coming in the developing world to meet unsatisfied demand in these countries. For example, a recent study by the International Institute for Applied Systems Analysis (IIASA) and the World Energy Council (WEC) estimated that the global energy consumption in the year 2020 will be 25 to 70% greater than the 1999 level of 6.75 billion tons of oil-equivalent. At the same time, the energy sector will also need to be reoriented as the demands from climate change, acid rain, local air pollution, and other environmental concerns become pressing and as fossil fuel reserves start to dwindle. With this context serving as a backdrop, the need for new and improved energy technologies is likely to continue to increase.

Concomitantly, the need to examine and better understand energy technology innovation and the

 © 2004 Elsevier Inc. All rights reserved.

underlying processes remains salient. It is only by assessing the scale, nature, and nuances of energy innovation activities can one judge their efficacy as well as their adequacy in relation to the challenges facing the energy system.

In general, energy innovation is stylized as a set of activities that include energy research, development, demonstration, and deployment, with the caveat that these activities do not occur in a purely linear and unconnected fashion. Because innovation is necessarily a recursive process requiring many iterations before a technology or product is deployed (and even subsequent to deployment, numerous incremental improvements generally continue to be made), there are invariably linkages and feedbacks among the various activities mentioned previously. For example, product design takes into account the conditions under which the technology will be deployed, data from demonstration projects reshape the technology, and product characteristics and designs are influenced by feedback on consumer needs and preferences. Still, following such a stylized model is useful in obtaining at least an initial understanding of energy innovation.

Hence, an examination of energy innovation requires an examination of energy research and development (ER&D), in terms of both the magnitude of the overall ER&D portfolio and its composition with regard to distribution of technology time horizons, risk levels, fuel source use, and end use applications. It requires a focus on the effectiveness of ER&D activities in terms of returns (e.g., social, economic) on research and development (R&D) investments. And it requires an exploration of activities influencing the eventual implementation—the deployment and eventual diffusion—of new technologies.

However, developing a detailed understanding of energy innovation is not a trivial proposition for a number of reasons. First, the notion of an energy innovation activity can be somewhat fuzzy and difficult to define clearly because a very broad range of research and technology development activities are relevant to the energy sector. In addition, many technology advances that emerge from other sectors, or are motivated by nonenergy applications, are also relevant to the energy sector. For example, advanced software packages for structural modeling, although driven initially by the space program and the aircraft industry, are of considerable use in designing advanced automobile structures, and basic research on fluid dynamics has helped to improve vehicle aerodynamics; both of these play a role in influencing the energy use characteristics of automobiles. Similarly, the advent of microprocessors has allowed substantial advances in fuel and engine management in automobiles, greatly improving their energy efficiency.

Furthermore, and to make matters more complex, numerous actors are involved in energy innovation activities of various kinds. A range of funders—governments, industry, venture capitalists, and philanthropic organizations—support the development of new energy technologies. A diverse cast of actors—government laboratories, industry, universities, research organizations and consortia, and nongovernmental organizations (NGOs)—carry out R&D activities that lead to new energy technologies. And an even larger group of organizations—government agencies, firms, bilateral and multilateral agencies, NGOs, and philanthropic groups—play a role in the deployment of energy technologies in different parts of the world. Numerous kinds of linkages, partnerships, and relationships among various actors, with flows of finances, technology, and knowledge among them, underlie the development and deployment of various energy technologies. Figure 1 illustrates the kinds of actors involved in energy technology innovation as well as the interactions among them. The web of actors and the relationships among them constitute what can be termed the "energy innovation system." This system is also shaped, and has its output influenced, by the context in which it operates (i.e., various policies [e.g., broad energy policies, tax and investment policies, environmental policies and institutions at the domestic level as well as trade, environmental, and other regimes at the international level], other macroeconomic factors, other institutional factors).

Hence, understanding of energy innovation processes may be best advanced through a broad examination of this innovation system. This requires a focus on the various activities in which relevant actors are engaged, their interests and preferences, the relationships and linkages among these actors, the flows of resources and knowledge among them, and the institutional context in which they operate. This follows the "systems" approach that has received increasing recognition within the broader community of science and technology policymakers as well as among scholars of innovation as a fruitful way in which to unpack the complexities of technological change within and across industries and nations.

Energy innovations come in many shapes and sizes. Ultimately, any energy innovation (like nearly

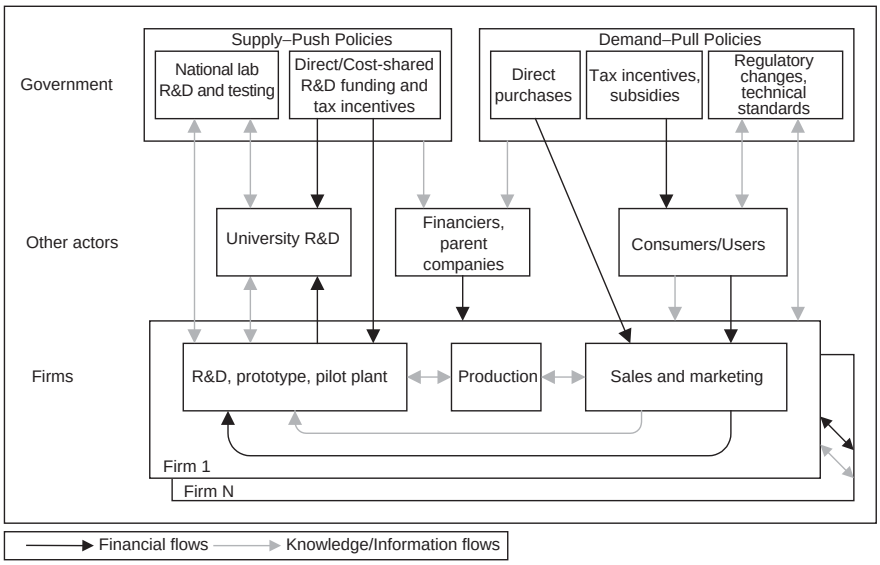


FIGURE 1 The major actors and their interactions in an energy innovation system. From Margolis, R. M. (2002). Understanding technological innovation in the energy sector: The case of photovoltaics. Ph.D. thesis, Princeton University.

any kind of technology innovation) aims to satisfy a demand or need. It aims to improve on the existing situation by providing energy services in a better fashion, expanding them, or providing new forms of service. Given that an overwhelming proportion of the world's energy services are delivered commercially, much of the demand for energy technologies emerges from, or is mediated by, the marketplace. Yet there is a significant fraction of the world's population whose members depend on noncommercial energy supplies. Estimates suggest that as many as 2 billion people in developing countries do not have access to modern forms of energy. Such differing circumstances lead to very different mechanisms for technology development as well as deployment. In the former case, consumer needs or aspirations are channeled through the market, as are resources for technology development and deployment. In the latter case, it is not feasible to rely purely on market institutions for technology development and deployment, although harnessing market mechanisms, to the extent feasible and desirable, may improve the effectiveness of any innovation efforts. Another related point is that even in a commercial energy arena, not all consumers are equal. Some consumers, such as large firms and public sector organizations, have the internal resources to assess (both technically and economically) relevant technologies in relation to energy needs. This becomes especially important for larger-scale technologies. Small and medium-size enterprises, even in industrialized countries, might not have such resources, often resulting in information or knowledge gaps that constrain uptake of new energy technologies (and, hence, the market signals to technology developers).

Hence, various kinds of innovation subsystems underlie the development and diffusion of different energy technologies. For example, the innovation subsystem underlying the development of clean coal technologies or fuel cell development does not much resemble that of improved cooking stoves. Similarly, energy innovation systems in various countries take very different shapes in terms of the kinds of actors involved, the relationships among them, and the policy contexts under which they operate. Such a perspective is necessary to develop an appropriately textured understanding of energy innovation processes underlying the development of technologies to satisfy the energy needs of varied categories of consumers.

This article presents an overview of the global energy innovation system, focusing especially on the prominent actors that constitute this system, their activities, and the policies and processes that play a role in energy innovation. It begins with a summary of the scale and scope of, and recent trends in, expenditures directed toward ER&D. It then discusses various institutions and policies that play a role in the deployment of energy technologies in different parts of the world. Finally, the interactions among the technical arena, the market, and policies that can influence the nature of energy technology innovation are highlighted through some illustrative cases.

2. ENERGY RESEARCH AND DEVELOPMENT

2.1 The Role of Governments

Not surprisingly, most ER&D takes place in industrialized countries where both the public and private sectors are involved in the development of energy technologies. Given that most energy technologies are deployed through the marketplace, the predominant locus of ER&D is within firms. Still, governments have historically played an important role in the R&D of new energy technologies through direct financial support (e.g., gas turbines) as well as through policies that promote ER&D within firms. The rationale for government involvement in ER&D is multifaceted:

- Concerns about energy security dominated the energy policy of most countries after the oil crises of the 1970s and drove the large increases in public ER&D spending seen during the late 1970s. Although the decline of oil prices by the early 1980s led to cutbacks in this spending in many countries, the energy security issue, as well as the high costs of energy imports, still motivates governments to make investments in energy technology innovation.
- The global market for energy technologies is estimated to be in the range of several hundred billion dollars per year. In this context, government involvement in energy technology development is also motivated by substantial commercial and national economic competitiveness elements.
- Although new and improved energy technologies are crucial to a robust, efficient, and clean energy system, the long time scales for energy technology development and deployment often make it unattractive to private firms to embark on many forms of ER&D efforts on their own.
- It is also recognized that affordable and reliable energy services have a "public good" nature due to their intimate linkage with various aspects of human and economic development. At the same time, reducing the environmental and other externalities of energy use is becoming paramount, especially with the expanding scale of the energy system worldwide and the emergence of problems such as climate change. Given the substantial social and environmental returns that may accrue from the widespread use of clean energy technologies, governments of most industrialized countries and many developing countries have traditionally funded appropriate ER&D efforts.

Along with public sector R&D investments (substantial in many countries), other government policies such as cost sharing, environmental regulations, and tax incentives provide stimulus for, and shape the priorities of, private sector investments in ER&D. Some policies, such as high taxes on energy and the commitment to expand and improve energy services, send a clear signal for demand for appropriate energy technologies, and this in turn creates an impetus for increased ER&D investment. Of course, government involvement in energy innovation goes well beyond R&D. Government also plays a role in promoting and assisting in the demonstration and deployment of such technologies (as discussed later).

The International Energy Agency (IEA) collects data on public sector ER&D expenditures by its member countries. Note that the IEA defines R&D expenditures in its data collection questionnaire as including funding for demonstration projects. Figure 2 shows the ER&D spending patterns over time by the seven countries that accounted for more than 90% of the energy technology development expenditures between 1990 and 2000 as reported to the IEA. Two sets of patterns emerge clearly from these data. First, public ER&D spending rose sharply after the two oil shocks of the 1970s, peaking in 1980 and then, as oil prices dropped, declining sharply for a decade or so and then declining somewhat more slowly during the 1990s. Second, the United States and Japan continue to account for a major portion of the world's ER&D. Also, Japan is the only major IEA country that has shown mostly increasing or stable ER&D expenditures over the past two decades or so.

Table I shows the distribution of total reported IEA spending for the year 2000 according to selected categories, and Fig. 3 shows recent trends in these expenditures. Some particular features should be noted:

- Nuclear fission and fusion R&D continues to account for nearly half of the total spending by IEA countries, although the proportion of total R&D devoted to this area has declined over the years, with the nuclear R&D spending peak having been reached in 1982. Since then, most countries have cut back their efforts in nuclear technologies substantially. Japan is the exception, and it now accounts for the majority of the public expenditures in this area, as nearly 70% of the Japanese government's ER&D is directed toward nuclear technologies.
- Government expenditures on conservation R&D have risen steadily over the past decade, but only two

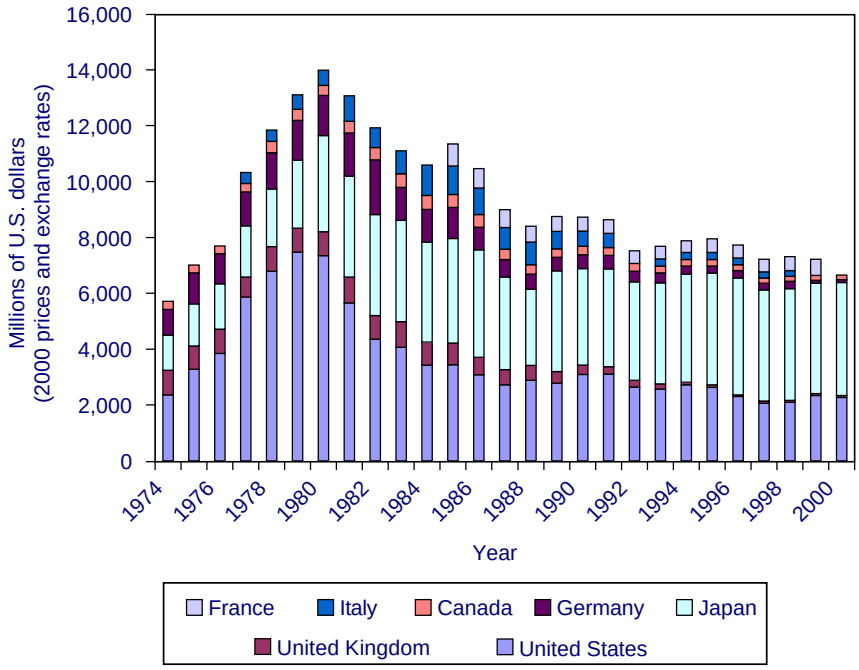


FIGURE 2 Trends in ER&D expenditures by major IEA governments. Note that Italy did not report any data for 1974–1976, 1992, 1999, and 2000. France did not report any data for 1974–1985 and 2000. From International Energy Agency R&D Database.

countries, Japan and the United States, are responsible for the bulk of expenditures in this area. Until the mid-1990s, global conservation expenditures were more or less split equally among the transportation, industrial, and residential sectors, but during the past few years, there has been a sharp rise in industrial conservation R&D efforts.

• IEA government spending on renewables R&D remained quite stable during the 1990s following a remarkable surge after 1974 that peaked in 1980 but fell steeply during the early 1980s. Nearly half of the renewables spending in 2000 was directed toward solar photovoltaic (PV) research, with R&D on biomass, wind, and geothermal sources accounting for most of the remaining amount. Small hydro R&D remains pretty much completely absent from the ER&D agendas of IEA countries. Once again, Japan and the United States lead the charge on renewables R&D, together spending nearly three-quarters of the total for IEA countries.

• Like many other areas, but perhaps even more clearly, fossil R&D has suffered a large decline during the past two decades or so. Coal R&D has been a particular casualty of these R&D cutbacks. Once again, the United States and Japan account for most of the public R&D in this area. The U.S. effort focuses particularly on enhanced oil and gas recovery as well as on coal combustion, whereas Japan accounts for most of the government coal conversion R&D expenditures among the IEA countries.

ER&D expenditures by governments are generally used to fund energy technology development efforts within government laboratories as well as in firms. Government organizations themselves may perform significant amounts of R&D. For example, in the United States, many U.S. Department of Energy (DOE) laboratories have research programs on a range of energy technologies. Relevant R&D is also carried out by other agencies such as the U.S. Department of Defense (DOD) and the National Aeronautical and Space Administration (NASA). For example, work done by DOD and NASA was instrumental in the development of fuel cells. The government's role may be even more prominent in developing countries. For example, in India, government laboratories and public sector undertakings perform much of that country's research on coal use, hydrocarbon exploration and processing, electric power generation, and atomic energy.

Governments also fund energy innovation in firms through mechanisms such as R&D contracts, cooperative research and development agreements, and grants for R&D or demonstration. Although less

TABLE I

Public Sector Energy R&D Spending by IEA Member Countries for the Year 2000 (Millions of U.S. Dollars, 2000 Prices and Exchange Rates)

Conservation	1269.4
Industry	721.3
Residential commercial	178.6
Transportation	279.2
Other conservation	70.8
Fossil fuels	426.4
Enhanced oil and gas	103.0
Refining transportation and storage	25.0
Oil shale and tar sands	9.8
Other oil and gas	49.4
Total oil and gas	187.2
Coal products preparation and transportation	16.7
Coal combustion	110.4
Coal conversion	72.2
Other coal	39.9
Total coal	239.2
Renewable energy	525.3
Solar heating and cooling	11.2
Solar photoelectric	234.5
Solar thermal–electric	22.3
Total solar	291.7
Wind	69.8
Ocean	7.7
Biomass	90.8
Geothermal	52.9
Large hydro (>10 MW)	7.2
Small hydro (<10 MW)	1.7
Total hydro	12.4
Nuclear fission/fusion	3259.3
Nuclear light water reactor	129.9
Other converter reactors	153.0
Nuclear fuel cycle	859.6
Nuclear supporting technology	1156.7
Nuclear breeder	398.4
Total nuclear fission	2709.5
Nuclear fusion	549.9
Power and storage technology	341.2
Electric power conversion	189.0
Electricity transmission and distribution	90.4
Energy storage	46.9
Other technology research	1075.0
Energy systems analysis	8.7
Other technology or research	1058.7
Total energy R&D	6965.5

Source. International Energy Agency R&D Database.

common, there are also some examples of government-funded programs on international cooperation in ER&D. A prominent example is the European

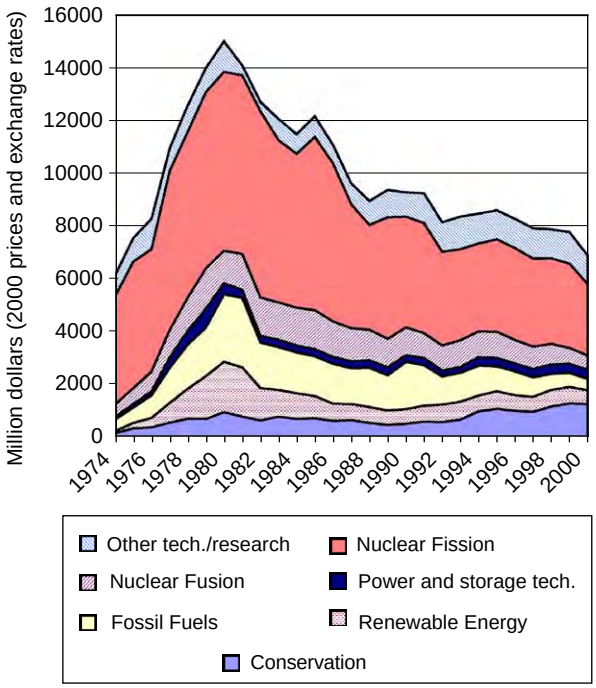


FIGURE 3 Trends in ER&D expenditures by IEA country governments, by category. From International Energy Agency R&D Database.

Union's program on research, technological development, and demonstration of energy technologies. The focus in the energy component of its Fifth Framework program (1998–2002) has been on technologies for cleaner energy as well as on economic and efficient energy, for which approximately 1 billion euros have been budgeted.

2.2 Research and Development in Firms

Industry, the main supplier of technologies for the consumer, industrial, and other applications, is the other major funder of ER&D. Its activities are targeted toward the development of new and/or improved products that help firms to compete in the marketplace so as to maintain their existing customer bases as well as expand them. Given that the market for energy technologies and related products is substantial, there is significant ER&D carried out by firms. For example, a recent study by the U.S. National Research Council estimated that between 1978 and 1999, industry accounted for roughly two-thirds of the total ER&D expenditures within the United States.

Extracting and collating data on ER&D expenditures by industry is a challenging problem for a

number of reasons. Many companies that are engaged in energy technology development have highly diversified operations, and because firms generally do not present detailed R&D data, it is nearly impossible to disaggregate industrial R&D activities to assess those that are focused on energy technologies. Even firms whose products/services are mainly in the energy sector may still engage in R&D that does not necessarily lead to energy technology development. And some firms may be focusing on technologies that have impacts across many sectors, including energy. For example, polymer composites initially designed for high-performance aircraft have been useful in producing lighter and, hence, more energy-efficient automobiles. Therefore, apportioning the underlying R&D budgets is a tricky, and possibly insurmountable, problem.

There are some attempts at surveying industrial ER&D (e.g., the U.S. National Science Foundation [NSF] does this as part of its "Research and Development in Industry" surveys), but it is unlikely that such efforts can accurately capture all of the ER&D in firms for the reasons mentioned previously. Furthermore, surveys across such a highly inhomogeneous population of firms are limited in terms of their coverage. Thus, it is likely that such data underestimate industrial ER&D. Still, such surveys could serve as a lower bound for industrial ER&D expenditures. The NSF estimated U.S. industrial ER&D spending to be $1.8 billion in 1999, whereas other studies by the Pacific Northwest National Laboratory estimated corresponding numbers to be $2.4 billion for Japan in 1997, $391 million for Italy in 1996, and $168 million for Germany in 1995. However, in addition to such data, a broad examination of the R&D expenditures and trends of firms in some particularly relevant sectors can give a direct idea of the size and nature of industrial investments in ER&D.

Table II lists some industrial categories (as well as some major firms within them) that are directly relevant to the energy sector. The primary business of firms in these categories is in the energy area; hence, a substantial fraction of their technology R&D activities have a direct and often substantial impact on the energy sector. Still, as mentioned previously, it should be understood that not all R&D activities of these firms are necessarily directed toward products and processes that may have an influence on the energy system. For example, car manufacturers can devote large amounts of development funds on designing vehicles that are aesthetically pleasing to the consumer, but many of these design features might not have a bearing on the energy characteristics of the automobiles.

TABLE II

Some Major Energy Technology Firms and Their R&D Spending for the Year 2000 (Billions of U.S. Dollars)

Transportation	
Vehicle manufacturers	
Ford Motor, United States	6.8
General Motors, United States	6.6
DaimlerChrysler, Germany	5.9
Toyota Motor, Japan	3.9
Volkswagen, Germany	3.9
Automotive systems suppliers	
Robert Bosch, Germany	1.8
Delphi Automotive Systems, United States	1.7
Aircraft engines	
Rolls-Royce, United Kingdom	0.55
Snecma, France	0.76
Oil and gas	
Integrated firms	
TotalFinaElf, France	0.64
Exxon Mobil, United States	0.56
BP Amoco (now BP), United Kingdom	0.43
Shell, United Kingdom	0.39
ENI, Italy	0.22
Oilfield services	
Schlumberger, United States	0.54
Halliburton, United States	0.23
Electric utilities	
Tokyo Electric Power, Japan	0.62
Korea Electric Power, South Korea	0.28
Enel, Italy	0.15
Taiwan Power, Taiwan	0.12
Heavy machinery	
Caterpillar, United States	0.65
Volvo, Sweden	0.52
Komatsu, Japan	0.37
Cummins, United States	0.24
Power generation/diversified industrials	
General Electric, United States	1.9
Siemens, Germany	5.2
Asea Brown Boveri, Switzerland	0.70
Alstom, France	0.47
Mitsubishi Heavy Industries, Japan	1.1
United Technologies, United States	1.3
Home appliances	
Whirlpool, United States	0.25
Electrolux, Sweden	0.14

Source. U.K. Department of Trade and Industry (2001). "2001 R&D Scoreboard". DTI, London. Exchange rate used: 1 GBP=1.49 USD.

2.2.1 Oil and Gas Exploration, Production, and Processing

Most major oil and gas firms engage in both upstream (i.e., hydrocarbon exploration and production) and downstream (i.e., hydrocarbon refining and marketing) businesses as well as related activities such as chemicals, and their R&D activities serve the needs of these businesses. During the past decade or so, the R&D spending of oil and gas firms has generally declined. According to a DOE survey, R&D spending by major energy producers within the United States declined from $3.05 billion in 1994 to $1.33 billion in 2000. Although not true of all producers internationally, the broad global trend over the past decade or so seems to be along similar lines. A series of annual surveys by the U.K. Department of Trade and Industry on R&D expenditures by major firms worldwide indicates a decline in the average R&D expenditures (as a percentage of sales) of major oil and gas firms worldwide, from approximately 0.75% in 1992, to 0.60% in 1996, to 0.34% in 2000—even as the cumulative sales of these firms rose approximately by 50% over the same period. A number of factors have likely contributed to this decline, including the recent consolidation within the industry, the decline of oil prices during the mid- and late 1990s, and the general trend toward R&D downsizing and outsourcing within firms.

Technology developed by other firms may also be useful for the oil and gas industry. For example, many of the R&D activities of an oilfield services firm such as Schlumberger are targeted toward acquiring data for better characterization of petroleum reservoirs, optimizing oil and gas production, and improving hydrocarbon recovery—all of great value to the petroleum and gas industries. To take another example, specialty chemical firms such as Akzo Nobel develop and provide catalysts useful for petrochemical processing.

2.2.2 Electricity Production and Conversion

Electric and gas utilities' R&D activities deal nearly exclusively with energy conversion, distribution, and use (at least to the extent that these firms still focus on these core businesses). Although detailed data on utility R&D are not collected systematically, it seems that R&D spending by many utilities has declined during the past decade or so. For example, a 1996 survey by the U.S. General Accounting Office (GAO) found that the combined R&D spending of the 112 largest operating utilities in the United States dropped from $708 million in 1993 to $476 million in 1996. These cutbacks were a result of overall declines in federal and state funding as well as increased competition expected from deregulation. Electric utilities' spending in Japan has also declined recently, but not as drastically. During 1993–1995, Japanese utilities' average annual expenditure was 34.1 billion yen, and by 1999–2001, this figure was 28.9 billion yen. R&D expenditure at the Electric Power Research Institute, a private research consortium funded by utilities and other firms, was approximately $463 million in 1996 and $312 million in 2001.

Much of the equipment used in electric power plants is produced by heavy engineering firms, most of which come under the category of "diversified industrials" by virtue of the fact that their products span a wide range of categories. Table III shows a list of major diversified industrials and their relevant energy technology businesses. There are also a few firms, such as ABB and Alstom, whose activities lie mostly within the energy sector (power and automation technologies for utilities and industry in ABB's case and power generation, transmission and distribution, and transport in Alstom's case). Therefore, a sizable portion of their R&D can be considered to be energy relevant.

2.2.3 Road Transport

The transport sector accounts for more than 25% of the total final consumption (TFC) of energy worldwide and more than a third of the TFC within the Organization for Economic Cooperation and Development (OECD) countries. Road transport, in turn, accounts for more than 80% of the transport energy consumption within OECD countries. Hence, R&D on automotive technologies assumes substantial importance within the general domain of ER&D.

Energy use of cars can be improved through a range of technologies (see, e.g., the fuel consumption technology matrix in Table IV) that are the focus of R&D activities of automobile manufacturers and their component suppliers. Thus, car manufacturers have research programs in areas such as innovative combustion techniques and fuel management, efficient engine designs, improved transmission technologies, and aerodynamic drag reduction to yield better fuel efficiency and lower the emissions of pollutants. Other prominent areas of research include gasoline–electric hybrid vehicles and fuel cell technology. Automobile manufacturers also carry out research in other areas such as structural mechanics and advanced materials (metal-based, ceramics, and polymers) that, by helping to lower the weight of

TABLE III

Energy Technology Segments of Some Major Diversified Industrials

General Electric	Power generation equipment and services
	Lighting
	Transportation
	Home appliances
Hitachi	Power generation, transmission, and distribution
	Industrial (electrical and electronic) equipment
	Consumer appliances
Mitsubishi Heavy Industries	Power systems
	Industrial (electric and electronic) equipment
	Industrial machinery and special vehicles
	Air conditioning
Siemens	Power generation, distribution, and transmission
	Lighting
	Industrial and building technologies
	Transportation
Toshiba	Power systems
	Industrial systems and equipment
	Transportation systems
	Consumer appliances
United Technologies	Power generation equipment
	Air-conditioning, heating, and refrigeration systems

TABLE IV

Passenger Car Efficiency Improvement Technologies (Percentages Improvement in Fuel Consumption)

Production-intent engine technology	
Engine friction reduction	1–5
Low-friction lubricants	1
Multivalve overhead camshaft	2–5
Variable valve timing	2–3
Variable valve lift and timing	1–2
Cylinder deactivation	3–6
Engine accessory improvement	1–2
Engine supercharging and downsizing	5–7
Production-intent transmission technology	
Five-speed automatic transmission	2–3
Continuously variable transmission	4–8
Automatic transmission with aggressive shift logic	1–3
Six-speed automatic transmission	1–2
Production-intent vehicle technology	
Aero drag reduction	1–2
Improved rolling resistance	1–1.5
Emerging engine technology	
Intake valve throttling	3–6
Camless valve actuation	5–10
Variable compression ratio	2–6
Emerging transmission technology	
Automatic shift/Manual transmission	3–5
Advanced continuously variable transmissions	0–2
Emerging vehicle technology	
42-V electrical systems	1–2
Integrated starter/generator (idle off–restart)	4–7
Electric power steering	1.5–2.5
Vehicle weight reduction (5%)	3–4

Note. Baseline automobile: overhead cams, four-valve, fixed timing, roller finger follower.
Source. National Research Council. (2002). "Effectiveness and Impacts of Corporate Average Fuel Economy (CAFE) Standards." National Academy Press, Washington, DC.

vehicles, can have a significant bearing on performance. R&D spending (as a percentage of sales) reported by car manufacturers also declined somewhat over the past decade or so. The average R&D intensity for the 10 largest manufacturers was approximately 4.8% in 1992 and 1996 but declined to just over 4.0% by 2000.

Component suppliers to the automotive industry, such as Bosch (which supplies gasoline and diesel injection systems), also have major R&D efforts to improve their own products. Even R&D by tire manufacturers can have an impact on the energy use of automobiles. For example, rolling resistance, which can account for 20% of the energy consumption in automobiles, can be reduced though improved tire design.

2.2.4 Residential and Commercial Sectors

The residential and commercial sectors are also major consumers of energy. Although aggregate global data are not available, these sectors account for approximately 19 and 11%, respectively, of the TFC within the OECD countries. Within the residential sector, improved technologies for space heating and cooling (the single largest category for energy use within homes), lighting, refrigerators, and smaller appliances can help to reduce energy use. Within the commercial sector, lighting and space heating and cooling are the two largest categories of energy consumption. Furthermore, building design and improved insulation and better windows can have a major influence on the energy use within residential and commercial buildings. Accordingly, the technological R&D of a range of firms, such as manufacturers of heat pumps, furnaces, water

heaters, refrigerators, air conditioners, washers, and dryers as well as those involved in building construction and materials supply, can play a role in changing the patterns of energy use in the residential and commercial sectors. Again, many of the relevant manufacturers have highly diversified operations, and residential and commercial products are only a small part of the overall operations. Still, it is possible to establish some estimate of the magnitude of R&D efforts in this area by looking at firms such as Whirlpool and Electrolux, which focus nearly exclusively on home appliances and are two of the largest home appliance manufacturers in the world, with their respective R&D budgets being $255 million and $139 million in 2000. It should also be noted that, given the relatively fragmented nature of the construction industry and little incentive for builders to explore methods and technologies for designing and constructing energy-efficient buildings, there is not much R&D effort within this industry.

2.3 Other Research and Development Issues

Although universities focus mainly on basic research rather than on technology development, their efforts are also relevant for energy innovation in many ways. Scientific research is increasingly linked to technological innovation, as indicated by the increasing citations to scientific research in patents. Thus, basic research on topics such as advanced materials may yield results that are critical to the developments of new technologies (e.g., fuel cells).

NGOs, research institutes, and universities are also involved in the development of simple energy technologies for low-income and rural groups, especially in developing countries. Such organizations rely on funding from numerous sources, including governments, multilateral and bilateral development aid agencies, and philanthropic organizations. For example, the R&D for the successful *jiko* stove (a ceramic stove for charcoal cooking) and other high-efficiency stoves that have found widespread use in Kenya and other African nations was seeded by international and local development funds. In contrast, the improved cooking stove programs in China and India were funded nearly entirely by the respective governments. As another example, the Intermediate Technologies Development Group (ITDG), an NGO that is involved in technology development and deployment in a number of developing countries, receives funds from individual donors, governments, foundations, and multilateral aid agencies.

There are no estimates of the total funds devoted to technology/product development aimed at enhancing energy services to the poor in developing countries, although it would be fair to say that this total will be only a small fraction of the ER&D expenditures discussed earlier. Beyond the obvious lack of large commercial opportunities for such applications, another reason for the limited focus on developing technologies and products for the poor in developing countries may be that much of the work in such endeavors is not cutting-edge research but rather what has been referred to as "mundane science," thereby not attracting the attention of many researchers working or interested in energy technology development. But it should also be noted that some of the ER&D expenditures in industrialized countries are also ultimately beneficial for low-income groups in developing countries. For example, advances in PV technology and "balance of system" components have allowed the development of PV-based products that provide energy services such as lighting and electricity in developing countries.

More generally, some concerns have been raised about the adequacy of ER&D spending globally. Underlying these concerns are a number of trends highlighted previously:

- Government cutbacks in energy technology research in many industrialized countries
- A change in the nature of private sector spending patterns in many technology firms, suggesting a shift to a market-driven paradigm of R&D with a focus on the shorter term as well as on greater dependence on outsourcing
- A decline in R&D spending by many major oil and gas firms, consonant with the return of petroleum prices toward historical levels after the boom prices following the oil shocks of the 1970s
- Energy sector restructuring that has affected R&D spending by utilities

But in response, it should be pointed out that a major portion of the decline in IEA government expenditures has been due to cutbacks in nuclear R&D. If nuclear R&D is excluded, the remaining total R&D expenditures seem to have leveled off after the remarkable decline of the early to mid-1980s (although there are still some fluctuations on an interannual scale). Also, although the R&D of some firms, such as petroleum majors and utilities, has declined substantially over the past decade or so, the R&D expenditures of many other energy-relevant

firms do not show a similar decline. Still, the larger point stands that as the challenges facing the energy system continue to mount, there has not been a concomitant rise in ER&D expenditures.

This also raises another issue. Much of the policy literature and analysis on energy innovation focuses on R&D budgets. Although such expenditures are a necessary component of the innovation process—clearly, there can be little innovation without ER&D investments—they are by no means sufficient in and of themselves. The eventual results of energy innovation activities depend both on the success in developing technologies through ER&D investments and on the uptake of these technologies by the eventual consumers. But performance issues of efficiency and effectiveness of innovation, important not only within the arena of energy but also more broadly, remain relatively less analyzed, although results indicate that investment in ER&D can potentially yield large benefits. For example, a recent study by the U.S. National Research Council of the DOE's energy efficiency and fossil energy R&D programs found that these yielded significant economic, environmental, and knowledge benefits. In the energy efficiency area, a $7 billion R&D investment between 1978 and 2000 was estimated to have yielded $30 billion worth of benefits, whereas the economic benefits associated with the fossil energy programs over the same period were estimated to be nearly $11.0 billion from an investment of $10.5 billion. In addition, environmental benefits resulting from these programs were estimated to be anywhere between $64 billion and $90 billion. Of course, this study also illustrates—and this leads directly to the next section—that yields from research on new or improved technologies can be realized only when these technologies find widespread use in the energy system.

3. ENERGY TECHNOLOGY DEPLOYMENT

Although investments in R&D can initiate the development of new energy technologies and help to support them through the early stages, bringing these technologies to market requires substantially more effort. Demonstration projects play an important part during this latter stage by testing in conditions approximating real-world applications so as to gain a better understanding of various performance parameters. In many cases, demonstration projects can also help to scale up technologies. This is especially critical in cases where the laboratory version of the technology is useful for testing the concept but where the eventual application of technology requires a much larger scale (e.g., in power plants). Not only do these kinds of projects help technology developers to test the feasibility of their products for eventual implementation, but they also demonstrate viability to potential buyers (often other corporations or public sector organizations) of these technologies. The data gathered from demonstration efforts can help to refine and improve technologies so as to increase the likelihood of commercialization. Energy technology demonstrations may often be cofunded by governments, given the public good aspects of energy services discussed earlier. An example of a major demonstration project is the U.S. Clean Coal Technology program that was initiated in 1986 and ultimately involved 38 projects that covered advanced electric power generation, environmental control technologies, and coal processing. The overall budget of this project was $5.3 billion, of which $1.8 billion came from the federal government.

Even if the technical feasibility of a new technology has been demonstrated, it still might not be able to compete in the market with already established technologies for a variety of reasons. The main categories of barriers include the following:

- Cost
- Infrastructure needs
- Slow capital stock turnover
- Market organization
- Information and financing options (particularly in developing countries)

Both private and public sector responses are required to overcome these barriers.

Of these, the first barrier—cost—is by far the most important given that most technologies are deployed through the market, and cost is the key variable in this arena. There are two main reasons why a new and improved technology, despite its commercial potential, might be unable to compete with an established technology in cost terms: The first reason is that consumers do not value, in financial terms, the improvement in energy service that the new technology may offer. This situation basically exists because energy prices do not reflect the full costs of energy. For example, a new energy technology may result in decreased local or global air pollution, but because such pollution is not internalized into the costs of energy use, any reductions are not reflected in market transactions. These price distortions are probably the

single largest barrier to the deployment of new and improved energy technologies.

The second reason is that costs of any technology are higher during the early stages of its production, thereby hampering the ability of this technology to compete with technologies already in widespread use. The costs of new technologies generally decline with increasing production and market experience. In fact, empirical data show that the total cost reductions of new technologies are related to their cumulative production, with the relationship between the two often referred to as "learning curves" (on a log–log plot, this appears as a linear relationship). This "learning" and concomitant reduction in cost can come from improvements in manufacturing techniques and processes as well as in product design that result from the experience gained by a firm (or an industry through spillover effects) as it engages in the production of these technologies (Fig. 4).

Eliminating or reducing market distortions so as to incorporate environmental or other externalities into market prices generally requires public policy actions. Governments can reduce these distortions by imposing taxes (e.g., pollution charges) that account for these externalities or by imposing ceilings combined with tradeable permits (e.g., pollution caps combined with trading opportunities [generally referred to as "cap and trade"]), both of which send a cost signal to the market. Such market-based instruments have gained significantly in popularity during the past decade or so, especially in the context of dealing with environmental externalities, and have been used in numerous policy experiments in different parts of the world. Another possibility is to use regulation to control the level of externalities (e.g., noise, environmental pollution, adverse health effects) that are acceptable. Both technology-based and performance-based standards have been used as regulatory policy instruments to achieve such goals. Governments can also give direct subsidies to technologies that reduce these externalities (e.g., environmentally cleaner technologies) or remove subsidies from technologies that have high externalities (e.g., polluting technologies).

Concerted private and public sector efforts are also required to overcome the second category of cost barrier by investing in continued R&D on these precompetitive technologies as well as in their production and deployment to reduce costs (through the learning effect mentioned previously) to the extent that they can overcome cost barriers and compete in the marketplace with the established conventional alternatives. This requires a learning investment as represented graphically in Fig. 5. Although ultimately much of this learning investment comes from the market (i.e., through sales of the products embodying the technology), and the developers of technology play a central role in this process, public sector policies can play an important role in assisting early deployment of technologies during their precompetitive stages.

Many governments have programs to promote the deployment of emerging energy technologies to help overcome these cost barriers. A major set of policies has involved government-mandated purchase of renewable power by utilities. In the case of the

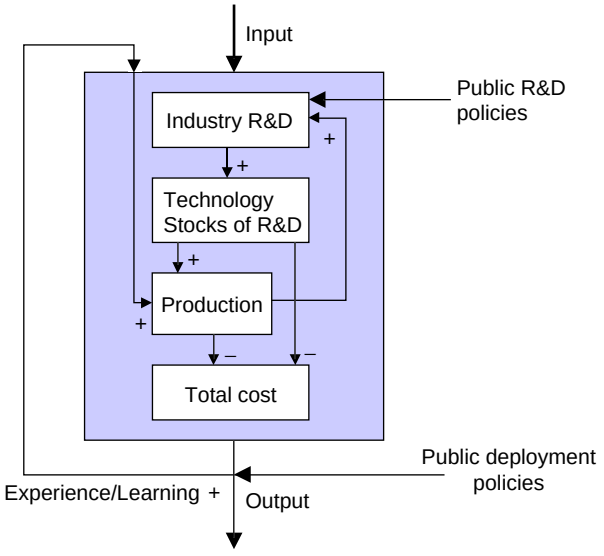


FIGURE 4 Interactions among industry R&D, production experience, and costs. © OECD/IEA, 2000, as adapted from "Experience Curves for Energy Technology."

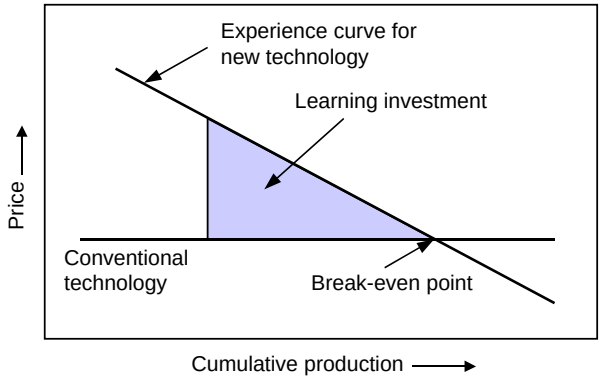


FIGURE 5 Learning investments needed for new technology to break even with conventional technologies. © OECD/IEA, 2000, as adapted from "Experience Curves for Energy Technology."

United States, this was through the Public Utility Regulatory Policies Act of 1978 (PURPA), which required utilities to buy electricity, at the utilities' "avoided cost," from qualifying facilities (which included power plants of <80 MW of capacity that used renewable energy sources). In the United Kingdom, the Non-Fossil Fuel Obligation (NFFO), a statutory obligation imposed by the British government under the Electricity Act of 1989, required the regional electricity companies that comprise the national distribution network in the United Kingdom to contract for a specified amount of generating capacity based on nonfossil fuels. A study by the Pacific Northwest National Laboratory estimated that in 1990, when the first NFFO tender was made, successful wind energy projects produced electricity at approximately 14.70 cents/kWh, whereas before the end of that decade, NFFO wind energy projects were producing at 4.85 cents/kWh. And under the German Electricity Feed Law (EFL), grid operators were obliged to purchase electricity produced from renewables within their respective supply areas at agreed and fixed prices. The value of such regulatory obligations can be enhanced if they provide a stable framework for project developers. For example, the NFFO is filled through competitive bids for long-term contracts (up to 15 years) that allow for guarantees on purchasing obligations as well as steady confidence in future markets and prices.

However, it should be noted that although these deployment programs helped to bring down the costs of renewables-based electricity generation, they were unable to ensure the long-term sustainability of renewables because the declines in cost seen in natural gas-fired generation were even greater.

Governments can also promote the deployment of improved energy technologies through the use of minimum performance standards. These kinds of regulatory standards are generally used for specified product categories (e.g., electric appliances), but they can also be set for a given percentage of products on the market, a market-wide average, or a manufacturer-based average. A commonly cited example of a standard is the U.S. Corporate Average Fuel Economy (CAFE) standards for automobiles, a manufacturer-based average. These standards, while primarily a response to the oil market disruption of 1973, have been notably successful. Greene has estimated that the total vehicle miles traveled in 1995 would have required an additional 55 billion gallons of fuel and would have cost motorists an extra $55 billion if there had been no improvement in on-road light-duty vehicle fuel efficiency levels since 1975.

Government procurement programs can also help to promote the deployment of energy technologies. For example, in the United States, the Energy Policy Act of 1992 (EPAct) directs the DOE, in association with other agencies, to "identify and designate those energy-efficient products that offer significant potential savings." The EPAct also calls for "guidelines to encourage the acquisition and use [of these products] by all federal agencies." Through its action plan to improve energy efficiency, the European Union is focusing on technology procurement that can be used to specify and develop new energy-efficient technology through both common European Union-wide procurement and coordinated procurement among its member states. In combination with public procurement policies, this will assist in both the development and diffusion of improved energy technologies.

Infrastructural barriers become important in cases where new technologies basically constitute an altogether different technological trajectory. For example, a shift to electric vehicles would require not only the necessary automotive technology but also a network of recharging stations. Similarly, a hydrogen-based energy economy would require a network for hydrogen distribution or distributed hydrogen generation. This requires the establishment of an infrastructure that may be economically infeasible for the small-scale introduction of a new technology. Often, the infrastructure requirements are so overwhelming that one set of technologies is basically "locked in." The current automobile-based transport infrastructure is a perfect example of this, where networks of roads and gasoline stations and servicing infrastructure are well established. Hence, alternative modes of transport such as light rail are at a large disadvantage because they have to compete with an existing mode of transport with a strong infrastructure. These kinds of barriers are difficult to overcome and almost certainly require some government involvement.

Improved provision of information to consumers, particularly individuals and small and medium-size enterprises (SMEs), can help to overcome information barriers to energy technology deployment. These barriers refer to the lack of information, or of the expertise to judge or analyze information, required for making appropriate choices during technology selection. To some extent, technical consultants or other experts can play a role in helping SMEs with such decisions. An approach that has found common use, especially for targeting individual consumers, is a labeling program to provide consumers with data about the performance of energy-using products and,

hence, to help them make more informed choices. Labels are generally of two kinds: those that provide information about performance on some common metrics (such programs require mandatory participation) and those that indicate the achievement (or surpassing) of a specific standard (such endorsement labels are voluntary by definition). Nearly all industrialized countries have labeling programs that cover some home appliances. The use of labels can have a significant impact on consumer behavior. A recent European study found that labels influenced the choices of a third of the purchasers surveyed. A classic and very successful example of an endorsement label program is the U.S. Energy Star program, which is a voluntary partnership among the U.S. Environmental Protection Agency (EPA), the DOE, and industry. Its basic aim is to expand the market for a range of energy-efficient products. The program covers new homes as well as products/appliances for home and office use. Webber has estimated that the program resulted in savings of 760 PJ of energy by 1999 (for reference, the U.S. energy consumption in 1999 was approximately 1×10^5 PJ).

Informational and financial barriers can be particularly problematic for individual consumers on the lower end of the economic spectrum, especially those in developing countries. Not only are low-income households particularly disadvantaged in being able to assess the relative merits of various technological options, but they are also hobbled financially in making choices. Hence, they might not have the savings to invest in an energy-efficient technology even if it is cost-effective over the lifetime of the technology, or they might not have access to credit for making choices. They also might have very high discount rates that bias any decisions toward the short term. There have been a number of approaches to tackle this barrier, including loans extended by microcredit agencies, term payment arrangements offered by service providers, and community-based saving schemes. Such arrangements can involve private sector firms, NGOs, other community-based organizations, and international aid agencies.

In fact, a number of international agencies are involved in various aspects of deployment and dissemination of improved energy technologies in developing countries. The most prominent of these is the Global Environmental Facility (GEF), which was established in 1991 as the financial mechanism for, among other agreements, the UN Framework Convention on Climate Change (UNFCCC). The GEF's climate change projects aim to mitigate the risks of global climate change while also providing energy for sustainable development. Hence, most of its climate change greenhouse gas (GHG) mitigation projects have (so far at least) focused on the energy sector. Its major goals have included promotion of the adoption of renewable energy by removing barriers, reduction of implementation costs, and reductions in the long-term costs of low GHG-emitting energy technologies. The GEF helps to overcome the incremental costs (in relation to conventional alternatives) of deploying climate-friendly technologies and, hence, plays a direct role in overcoming the cost barriers discussed earlier. From 1991 to 2001, it approved 136 climate change projects with a total grant component of more than $1 billion (and with cofinancing of more than $5 billion).

Other international aid agencies, both bilateral and multilateral, also directly support the deployment of renewable and other energy technologies in developing countries and also help to overcome barriers to such deployment. Although the private sector has taken on increasing significance in the development of the energy sector in developing countries, international aid and export credit agencies often play a critical role in the deployment and adoption of new and improved energy technologies in these countries. For example, the large increases in the deployment of renewable energy technologies in India were partly a result of the large promotion of such technologies—through financial, information, and other assistance mechanisms—by the Indian Renewable Energy Development Agency, which has received substantial international assistance from the GEF, the World Bank, and the Asian Development Bank as well as the bilateral aid agencies of Germany, The Netherlands, and Denmark.

Organizations such as NGOs, rural energy cooperatives, village committees, local vendor representatives, microcredit and other financing agencies, and energy service companies also play a major role in the deployment of energy technologies in developing countries, particularly among low-income groups. Sustainable deployment in rural areas of developing countries requires appropriate installation, service, and maintenance expertise as well as suitable financing mechanisms so as to help with deployment. For example, the ITDG, beyond its work on technology development, also assists in local manufacture and technical support, formation of producer associations, training and skills development, and enhancement of community financing (especially microfinance).

In the end, successful energy innovation requires concerted, synergistic, and sustained efforts by a

range of actors, institutions, and policies to develop technologies and bring them to the marketplace. This is clearly demonstrated by the illustrative cases of the Japanese PV development program and German wind development programs.

3.1 The Japanese Photovoltaic Program

Japan has the largest PV installed capacity in the world as a result of the sustained investments made by the country through carefully coordinated PV R&D and deployment, leading to a "virtuous cycle" among R&D, market growth, and price reduction. By the end of 2000, the total installed capacity in the country was 317 MW, of which on-grid distributed generation accounted for more than 252 MW. In 2000, Japan produced 128 MW of PV cells and 136 MW of modules, both of which were well over half of the global production.

The Japanese PV program was initiated in 1974 after the oil crisis as part of the "Sunshine Project" to develop alternative forms of energy. It has been under the direction of the New Sunshine (NSS) project since 1993. The overall target of the NSS is to install 5000 MW of PV capacity by 2010.

The NSS program is directed by the Ministry of Trade and Industry (MITI). On the R&D front, the government has been investing heavily in PV R&D; between 1991 and 1998, government spending on PV R&D exceeded $480 million. Approximately 40% of the MITI's PV R&D budget went to the eight leading Japanese PV firms, with the remainder going to national research institutes, universities, and other industries. The government support for PV R&D induced vigorous R&D expenditures in industry; by 1995, industrial expenditures on PV R&D were about twice those of the government. The MITI also established an R&D consortium for PV development that had 65 participating firms by 1996. The main focus areas of R&D were cost reduction of PV systems, silicon feedstock for PV cells, and PV deployment.

The MITI is also implementing a demonstration program that involves PV field tests for industrial use, where half of the PV installation costs are subsidized. This follows the successful completion of a demonstration program on field tests for public facilities that was completed in 1998. The aim of this program is to install trial PV systems to introduce industry to PV use, demonstrate and analyze long-term operation of PV systems, and standardize and introduce PV applications toward full-scale deployment.

The PV deployment program also has a number of components, with the main one being the residential PV dissemination program that aims to promote, through subsidies, the installation of PV systems by individuals. Other components include programs to support efforts established by regional governments and to support industrial entrepreneurs who are involved in introducing PV systems. A number of government ministries, local governments, municipalities, and utilities are also involved in this deployment effort. In addition, solar cell manufacturers and housing firms have organized the Japan Photovoltaic Energy Association to carry out public relations and dissemination in cooperation with the government. This deployment effort has led to learning in regard to both the PV modules and the balance of system components. Between 1993 and 1998, the government subsidies for this program totaled approximately $200 million but stimulated market actors to provide additional learning investments of $300 million. It is estimated that by 2007, PV commercialization can become self-sustaining. By that point, market investments in learning will be about 5.7 times the government support.

3.2 The German Wind Power Program

Germany is considered a world leader in wind energy deployment, with electricity production from wind having gone up more than eightfold—from 670 to more than 5500 GWh—between 1993 and 1999. In 1999, Germany had more than 4000 MW installed capacity. The prices also declined substantially over this period, indicating a learning rate of approximately 8%, and it is estimated that for each deutschemark (DM) spent by the government in subsidizing learning, market actors provided 2.1 DM. Furthermore, the manufacturing base for wind energy technologies, including indigenous firms as well as joint ventures, continues to expand in Germany.

A number of different actors and policies operating in tandem contributed to this success story. In 1989, when the federal government initiated a "100-MW Wind Programme," there was nearly a complete absence of a market for wind turbines. Wind plant operators could receive money from this program as well as utilities for power delivered to the grid. Investment subsidies were also provided by this program, and the Lander (states) offered additional grants. In 1991, the federal government also promulgated the EFL, which obliged grid operators to purchase electricity produced from renewables at agreed and fixed prices and, thus, provided a stable

market for investors. In the same year, the government also enlarged the 100-MW program to a 250-MW program. During the early 1990s, the government also increased its support for wind power R&D, with the budget for the 1992–1996 period being more than double that for the 1987–1991 period.

In addition, many German regions or municipalities established targets intended to promote renewable energy, including wind energy and use, and often provided financial support, including investment subsidies, to implement these renewable energy policies. Many German financial institutions also provided low-interest loans for renewable energy projects. Between 1990 and 1997, a total of 3.48 billion DM (1.78 billion euros) was disbursed for wind energy projects. Tax laws that allowed the offsetting of investments in wind farms made this an attractive option for individuals. Administrative support was provided for wind energy projects in the forms of planning guidance and land use directives that helped to assuage local concerns about these projects. At the same time, developers also often held public meetings in local communities to share information about, and raise interest in, schemes. Finally, the high level of environmental awareness as well as interest in renewables among the citizenry contributed significantly to the public support for wind energy deployment.

4. CONCLUSION

The aim of this broad sweep through the area of energy innovation, highlighting the main actors, activities, policies, institutions, and their interactions that eventually underlie energy technology development and deployment, was to present an overview of the current state of and challenges to energy innovation.

To recapitulate, some key points are as follows:

- A diverse cast of actors, with flows of finances, technology, and knowledge among them, are involved in energy technology innovation. Various policies (e.g., energy, fiscal, trade, competitiveness, environmental), as well as other macroeconomic and institutional factors, shape this "energy innovation system" and also influence its output.
- Most energy technology development takes place within firms, although other organizations (e.g., government laboratories, universities) also contribute to this process. Governments and the private sector direct large expenditures toward ER&D.
- Although most of the world's activities on energy technology innovation focus on users and consumers in the commercial energy arena, there are much smaller but parallel innovation subsystems that focus on the energy service needs of the poor in developing countries. Many of the actors in these subsystems are often quite different from the actors in the mainstream energy innovation system, although there are linkages between the two groups and their activities.
- Trends in global ER&D expenditures give some cause for concern due to the burgeoning challenges facing the energy system, although the potential for meeting these challenges may be hobbled more by ineffective deployment of suitable energy technologies than by insufficient R&D.
- For any new energy technology, the move from the precompetitive stage to deployment faces numerous barriers. Thus, successful energy innovation requires concerted, synergistic, and sustained efforts by a range of actors, institutions, and policies not only to develop technologies but also to bring them to the marketplace.
- Government policies such as moving toward full cost energy pricing are perhaps the most important drivers of energy innovation by creating a demand for new and improved energy technologies that, in turn, creates an impetus for ER&D as well as deployment.

Although there have been a number of efforts in different countries to guide the energy innovation system, much more effort will be needed to enable it to meet the energy-related challenges facing humanity. We know a fair bit about some aspects of energy technology innovation (e.g., R&D inputs, various actors and interactions among them, deployment issues), but we know much less about some others (e.g., efficiency and effectiveness of R&D, sufficiency in relation to challenges). Improving the understanding of such issues, and finding the social and political will to act on the lessons that emerge, will ultimately determine our ability to meet these daunting challenges.

SEE ALSO THE FOLLOWING ARTICLES

Consumption, Energy, and the Environment • *Development and Energy, Overview* • *Early Industrial World, Energy Flow in* • *Economic Growth,*

Liberalization, and the Environment • Energy Ladder in Developing Nations • Environmental Change and Energy • Experience Curves for Energy Technologies • Global Energy Use: Status and Trends • Information Technology and Energy Use • Innovation and Energy Prices • Microtechnology, Energy Applications of • Research and Development Trends for Energy

Further Reading

Dooley, J. J. (1998). Unintended consequences: Energy R&D in a deregulated market. *Energy Policy* **26,** 547–555.

Grubler, A., Nakicenovic, N., and Victor, D. G. (1999). Dynamics of energy technologies and global change. *Energy Policy* **27,** 247–280.

International Energy Agency. (1997). "Enhancing the Market Deployment of Energy Technology: A Survey of Eight Technologies." IEA, Paris.

International Energy Agency. (2000). "Energy Technology and Climate Change: A Call to Action." IEA, Paris.

International Energy Agency. (2000). "Experience Curves for Energy Technology Policy." IEA, Paris.

International Energy Agency. (2001). "Energy Labels and standards." IEA, Paris.

International Energy Agency. (n.d.). "IEA Energy Technology R&D Statistics." IEA, Paris. www.iea.org/stats/files/rd.htm.

Metz, B., Davidson, O. R., Martens, J-W., van Rooijen, S. N. M., and McGrory, L. V. W. (2000). "Methodological and Technological Issues in Technology Transfer." Cambridge University Press, Cambridge, UK.

National Research Council. (2001). "Energy Research at DOE: Was It Worth It?" National Academy Press, Washington, DC.

President's Council of Advisors on Science and Technology. (1997). "Federal Energy R&D for the Challenges of the 21st Century." Office of Science and Technology Policy, Washington, DC.

President's Council of Advisors on Science and Technology. (1999). "Powerful Partnerships: The Federal Role in International Cooperation and Energy Technology Innovation." Office of Science and Technology Policy, Washington, DC.

Reddy, A. K. N., Williams, R. H., and Johansson, T. B. (1997). "Energy after Rio: Prospects and Challenges." UN Development Program, New York.

Sagar, A. D. (2002). "India's Energy and Energy R&D Landscape: A Brief Overview." BCSIA Discussion Paper 2002-08, Energy Technology Innovation Project, John F. Kennedy School of Government, Harvard University.

Sagar, A. D., and Holdren, J. P. (2002). Assessing the global energy innovation system: Some key issues. *Energy Policy* **30,** 465–569.

Watanabe, C., Wakabayashi, K., and Miyazawa, T. (2000). Industrial dynamism and the creation of a "virtuous cycle" between R&D, market growth and price reduction: The case of photovoltaic power generation (PV) development in Japan. *Technovation* **20,** 299–312.

World Energy Council. (2001). Energy Technologies for the Twenty-First Century. WEC, London.

Temperature and Its Measurement

JOSEPH PRIEST
Miami University
Oxford, Ohio, United States

1. **Temperature**
2. **Thermometry**
3. **Temperature Scales**
4. **Range of Temperatures**
5. **The International Temperature Scale of 1990**
6. **Temperature Sensors**

Glossary

absolute zero The lowest temperature on the Kelvin temperature scale.

blackbody A perfect emitter and absorber of thermal radiation.

Celsius temperature scale A temperature scale based on 0 and 100°C for the ice and steam points of water, respectively; named for Swedish astronomer Anders Celsius.

emissivity The ratio of power radiated by a substance to the power radiated by a blackbody at the same temperature.

Fahrenheit temperature scale A temperature scale based on 32 and 212°F for the ice and steam points of water, respectively; named for German scientist Gabriel Daniel Fahrenheit.

infrared thermometer A device using Planck's law and measurements of thermal power in the range 0.7 to 80 μm to deduce the temperature of the source.

International Temperature Scale of 1990 (**ITS-90**) An internationally recognized standard adopted on January 1, 1990. Between 0.65 and 5 K, temperatures on the ITS-90 are defined in terms of the vapor–pressure temperature relations of helium-3 and helium-4. Above 5 K, the ITS-90 assigns temperatures to sixteen fixed points (Table I). Between the fixed points, temperatures on the ITS-90 are obtained by interpolation using a helium gas thermometer in the range 3 K to 24.5561 K, a platinum resistance thermometer in the range 13.8033 K to 1234.93 K, and an optical thermometer above 1234.93 K.

kelvin A unit of temperature defined as $\frac{1}{273.16}$ of the triple point temperature of water.

Kelvin temperature scale A temperature scale based on a temperature of 273.16 K for the triple point of water; named for physicist and mathematician Lord Kelvin, born William Thomson and knighted by Queen Victoria as the first Baron of Largs (1824-1907).

manometer A device for measuring pressure differences, usually by the difference in height of two liquid columns.

optical pyrometer A device using Planck's law and measurements of thermal power in the range 0.4 to 0.7 μm, usually approximately 0.66 μm to infer the temperature of the source.

Planck's radiation law A mathematical relationship formulated in 1900 by German physicist Max Planck(1858–1947) to explain the spectral energy distribution of radiation emitted by a blackbody.

Rankine temperature scale A temperature scale based on a temperature of 491.69°R for the triple point of water; named for Scottish engineer William Rankine (1820–1872).

resistance temperature detector (**RTD**) A sensor relating temperature to the electrical resistance of a conductor.

Seebeck effect The voltage produced when there is a temperature gradient along a conductor; named for physicist Thomas Johann Seebeck (1770–1831).

temperature At the human level, a measure of the sensation of hot and cold; at the atomic level, it is a measure of the average linear kinetic energy of an atom or a molecule in a substance.

thermal energy The average kinetic energy associated with the motion of atoms and/or molecules in a substance.

thermal equilibrium When there is no change in the thermometric properties of either system.

thermistor A temperature sensor employing the electrical resistance of a semiconductor as a thermometric property.

thermocouple A temperature sensor consisting of two dissimilar wires bonded together at one end to form a junction; a voltage related to the temperature of the junction is produced between the open ends of the two wires.

thermometer An instrument employing a thermometric property to measure temperature.

thermometric property A property of a system that changes with temperature.

 © 2004 Elsevier Inc. All rights reserved.

triple point temperature The single temperature at which liquid, vapor, and solid coexist in thermal equilibrium.

Wien's displacement law A mathematical relationship formulated by German physicist Wilhelm Wien (1864–1928) showing that the wavelength for maximum radiative power of a blackbody is inversely proportional to the Kelvin temperature.

Zeroth law of thermodynamics A relationship indicating that if two systems are separately in thermal equilibrium with a third system, they must also be in thermal equilibrium with each other.

Temperature is a measure of the sensation of hot and cold. Inanimate objects that respond to the sensation are used in a thermometer to assign a numerical value to the sensation. Thermometers employing the common Celsius and Fahrenheit scales assign 0°C and 32°F to the freezing point of water, respectively, and assign 100°C and 212°F to the boiling point of water, respectively. Temperature changes are recorded identically on the Celsius and Kelvin scales and on the Fahrenheit and Rankine scales, but the zero value differs on the respective pairs of scales. This article discusses the range of temperatures encountered in nature and the International Temperature Scale of 1990 (ITS-90). It also describes a wide variety of temperature sensors along with their operational temperature range.

1. TEMPERATURE

A person feels warm when standing near a campfire and feels cold when standing in front of an open refrigerator. Temperature is a way of recording the extent of this sensation of hotness and coldness. Inanimate things also respond to hotness and coldness. For example, the pressure of air in an automobile tire decreases when the air cools, and the length of a railroad rail increases when the rail warms. These inanimate changes are used to assign a numerical value to temperature.

A metal spoon taken from a refrigerator is cold to the touch, and water drawn from a hot water tap feels warm. If the spoon is placed in the water, the spoon warms while the water cools. Eventually, both the spoon and the water feel the same to the touch, and we say that thermal equilibrium is attained and that the two have the same temperature. This definition of temperature does not depend on any particular temperature scale. The property of temperature indicates when two objects are in thermal equilibrium.

If a spoon and a fork are placed in a refrigerator, both will come to thermal equilibrium with the environment of the refrigerator, and we would say that each has the same temperature as the refrigerator. We would also find that if the spoon and fork were placed in contact, they would be in thermal equilibrium and, therefore, at the same temperature. Although not surprising, this idea is profound and gives rise to the Zeroth law of thermodynamics: if two systems are separately in thermal equilibrium with a third system, they must also be in thermal equilibrium with each other. One of these three systems could be an instrument for measuring temperature, that is, a thermometer. If a thermometer records a certain temperature when in thermal equilibrium with one of the systems, we would expect the thermometer to record the same temperature when in thermal equilibrium with the third system.

2. THERMOMETRY

A property of a system that changes with temperature is called a thermometric property. Examples of a thermometric property include the following:

- Pressure of a gas
- Volume of a liquid
- Electrical resistance
- Color
- Electrical voltage
- Character of thermal radiation

The algebraic equation chosen to relate temperature to a thermometric property is arbitrary. However, the simplest way, and the one most generally used, is a linear equation of the form $T = mx + b$, where T is the temperature, x is the thermometric property, and m and b are experimentally determined constants. The constants m and b may be determined by assigning numerical values to two temperatures and recording the values of the thermometric property at these two temperatures. Typically, the two temperatures are the freezing point of water and the boiling point of water. These two temperatures are assigned 0 and 100°C on the Celsius scale, respectively, and are assigned 32 and 212°F on the Fahrenheit scale, respectively.

3. TEMPERATURE SCALES

3.1 Celsius and Fahrenheit Scales

To illustrate the thermometer principle and the Celsius scale of temperature, consider the "liquid-in-glass" device depicted in Fig. 1. The thermometer is a glass tube with a reservoir for the liquid at one end and a capillary (narrow cylindrical channel) feeding into the reservoir. Liquid is forced into the capillary when the temperature of the liquid in the bulb increases. Suppose that the level of the liquid is determined by a centimeter scale attached to the side of the thermometer. If the level corresponds to 1.1 cm when the bulb is at the freezing point of water (0°C) and corresponds to 10.3 cm when the bulb is at the boiling point of water (100°C), we write $0 = 1.1m + b$ and $100 = 10.3m + b$ for the two respective cases. Solutions to these two equations yield $m = 10.9$ and $b = -12.0$, and the equation relating temperature to fluid height becomes $T = 10.9x - 12.0$. An unknown temperature is then determined by measuring the fluid level (x) and calculating the temperature (T) using the equation. In practice, the 0 and 100°C temperatures are etched at the fluid level for the freezing point and boiling point of water, respectively. Then, the space between the 0 and 100°C marks is graduated into 100 equal segments and an unknown temperature is read directly from the position of the fluid level. If the thermometer were recording temperature on the Fahrenheit scale, the 0 and 100°C designations would be replaced with 32 and 212°F, respectively, and the space between the marks would be graduated into 180 equal segments. Conversion between the two scales follows from the relation $F = \frac{9}{5}C + 32$.

3.2 Kelvin and Rankine Scales

Mercury is very useful in a liquid-in-glass thermometer because it is liquid at room temperature, does not wet glass or cling to it, and has uniform volume expansion throughout its liquid range. Because mercury remains a liquid down to approximately −40°C, mercury-in-glass thermometers can record temperatures as low as −38.7°C. Mercury remains a liquid for temperatures several hundred degrees Celsius above room temperature, but other factors (e.g., softening of the glass housing) limit the highest measurable temperature to approximately 360°C. The use of a different liquid, notably ethanol, can extend the lower measurable temperature to approximately −115°C. Because of the simplicity of liquid-in-glass thermometers and the relatively large measurable temperature range, they have been widely used for roughly 300 years.

Factors that limit the reproducibility and accuracy of liquid-in-glass thermometers include the following:

- Manufacturing reasonably identical thermometers
- Recording the precise level of the liquid
- Experiencing difficulty in reproducing the ice and steam points of water

A constant volume gas thermometer (Fig. 2) using the pressure of a gas as a thermometric property avoids most of these problems. The great virtue of this thermometer is that for very low gas pressure, the temperature deduced is independent of the gas used.

The three phases of water—vapor, liquid, and solid—exist with no change in any component at a unique temperature called the triple point temperature. Gas in a container is then brought to thermal equilibrium with its surroundings, and this is maintained at the triple point temperature. The pressure of the gas (P_o) is then measured. By definition, the triple point temperature is 273.16 kelvins, symbolized as 273.16 K. When the gas is brought to equilibrium with the surroundings at a temperature other than the triple point, the fluid reservoir of the manometer is raised or lowered so that the gas volume remains constant and the gas pressure (P) is measured. Assuming that for constant volume, pressure is proportional to temperature, the ratio of two temperatures equals the ratio of the corresponding pressures. Knowing the pressure and temperature for the triple point and measuring the pressure for the unknown temperature, the tempera-

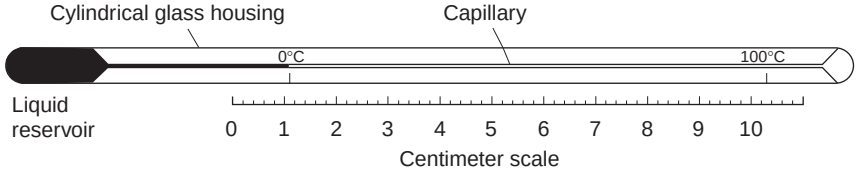


FIGURE 1 A liquid-in-glass thermometer.

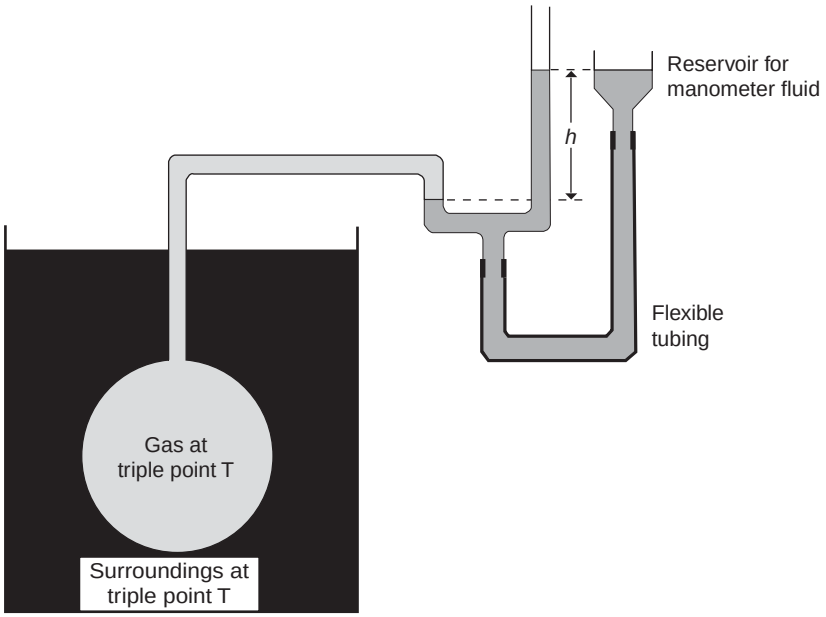


FIGURE 2 Constant volume gas thermometer principle. The apparatus illustrated here for measuring gas pressure is called a manometer. Gas pressure is related to the difference in fluid levels in the manometer. The volume of gas is held constant by raising or lowering the reservoir for the manometer fluid. There is a variety of ways in which to measure the pressure and to ensure constant gas volume.

ture is then computed from the relation

$$T = 273.16\frac{P}{P_o}. \qquad (1)$$

Temperatures deduced from Eq. (1) are expressed in kelvins, and the scale is called the Kelvin temperature scale. One kelvin is defined as $\frac{1}{273.16}$ of the triple point temperature of water. A kelvin is equal to a Celsius degree, so that changes in temperature are the same on both the Kelvin and Celsius scales. The freezing point of water on the Kelvin scale is 273.15 K, so that the conversion from kelvins to degrees Celsius becomes $K = C + 273.15$. The lowest temperature on the Kelvin scale is zero and is referred to as absolute zero. The Rankine scale of temperature is developed in the same way as the Kelvin scale by taking the triple point of water to be 491.69°R. A Rankine degree is equal to a Fahrenheit degree. The freezing point of water is 491.67°R, so that conversion between the two scales becomes $R = F + 459.67$.

The random incessant motion of atoms and/or molecules in a substance is called thermal motion. The motion of an atom or molecule always has linear components and in the case of molecules may include vibration and rotation. The average kinetic energy associated with the motion is called thermal energy. The average linear kinetic energy of an atom or molecule is directly proportional to the Kelvin temperature of the substance. At the level of atoms and molecules, it is appropriate to interpret temperature as a measure of the average linear kinetic energy of an atom or a molecule in a substance.

4. RANGE OF TEMPERATURES

Most humans would agree that a temperature of 35°C (95°F) is quite hot and that −10°C (14°F) is rather cold. Yet in 1913, humans recorded a temperature of 56.7°C (134°F) in Death Valley, California, and a temperature of −89.4°C (−129°F) in 1983 in Vostok, Antarctica. As large as these temperature extremes are by human standards, they pale in comparison with other earthbound and extraterrestrial temperatures. Liquid nitrogen at a temperature of −196°C (77 K) is routinely used in cryosurgery to destroy cancerous tumors, magnets used in magnetic resonance imaging (MRI) are cooled with liquid helium at a temperature of −269°C (4.2 K), and the temperature of the filament in an incandescent light bulb operates at approximately 2500°C (2200 K). Low-temperature research laboratories have produced temperatures in the tens of billionths of a kelvin. Extraterrestrial, we find a cosmic background temperature of 2.7 K (−270.5°C), a surface temperature of approximately 5700°C

(6000 K) for our sun, and an interior temperature of approximately 15,000,000 K for the sun. Figure 3 depicts the range of temperatures encountered in nature.

5. THE INTERNATIONAL TEMPERATURE SCALE OF 1990

The International Temperature Scale of 1990 (ITS-90) is the internationally recognized standard. Between 0.65 and 5 K, temperatures on the ITS-90 are defined in terms of the vapor pressure temperature relations of helium-3 and helium-4. Above 5 K, the ITS-90 assigns temperatures to 16 fixed points (Table I). Between the fixed points, temperatures on the ITS-90 are obtained by interpolation using a helium gas thermometer in the range of 3 to 24.5561 K, a platinum resistance thermometer in the range of 13.8033 to 1234.93 K, and an optical thermometer above 1234.93 K.

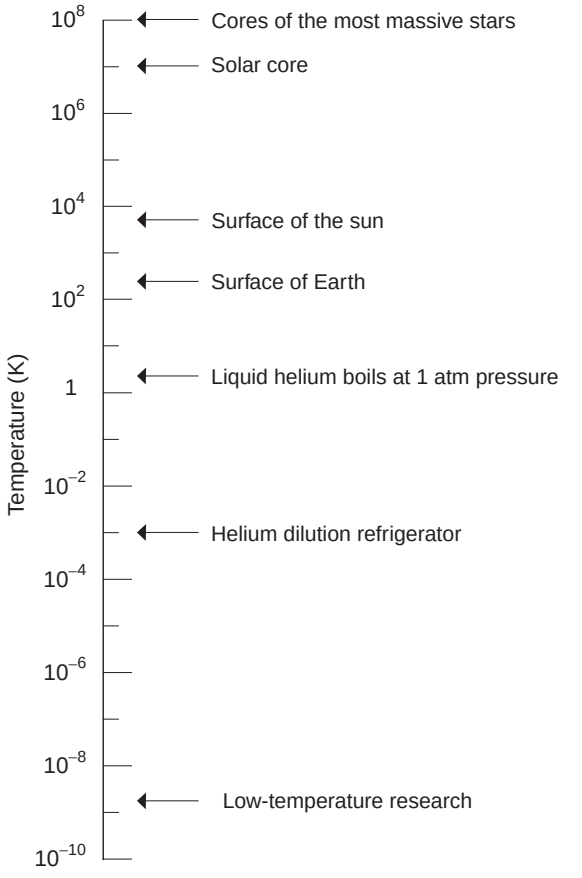


FIGURE 3 A representative few of the temperatures encountered in nature. Note that the scale is logarithmic.

6. TEMPERATURE SENSORS

6.1 Resistance Temperature Detectors

6.1.1 Platinum Resistance Temperature Detector

The resistance to the flow of electric charge in a conductor depends on temperature. For metals, the resistance increases as the temperature increases. The algebraic relation between temperature and resistance allows one to determine the temperature of the metal from a measurement of its resistance. Thermometers based on this principle are called resistance temperature detectors (RTDs). Platinum is particularly useful for an RTD because it can be made very pure, is stable and does not oxidize easily, and has a relatively high melting point of 1772°C. In addition, its resistance varies smoothly—nearly linearly—over a wide range of temperatures. Properly calibrated, a platinum resistance thermometer can achieve reproducible readings accurate to less than a thousandth of a Celsius degree.

Figure 4 shows a platinum RTD constructed from fine platinum wire approximately 0.1 mm in diameter. Typically, the sensor has a resistance of 100 ohms (Ω) at 0°C and increases approximately 0.39 Ω for each Celsius degree increase. A second, more modern, and more widely used type of platinum resistance temperature sensor is shown in Fig. 5. The platinum is deposited as a very thin narrow ribbon onto a ceramic base. Typical film

TABLE I

International Temperature Scale Calibration Temperatures

Hydrogen triple point	13.8033 K	−259.3467°C
Hydrogen vapor pressure (~33 kPa)	~17 K	−256.15°C
Hydrogen vapor pressure (~101 kPa)	~20.3 K	−252.85°C
Neon triple point	24.5561 K	−248.5939°C
Oxygen triple point	54.3584 K	−218.7916°C
Argon triple point	83.8058 K	−189.3442°C
Mercury triple point	234.3156 K	−38.8344°C
Water triple point	273.16 K	0.01°C
Gallium melting point	302.9146 K	29.7646°C
Indium freezing point	429.7485 K	156.5985°C
Tin freezing point	505.078 K	231.928°C
Zinc freezing point	692.677 K	419.527°C
Aluminum freezing point	933.473 K	660.323°C
Silver freezing point	1234.93 K	961.78°C
Gold freezing point	1337.33 K	1064.18°C
Copper freezing point	1357.77 K	1084.62°C

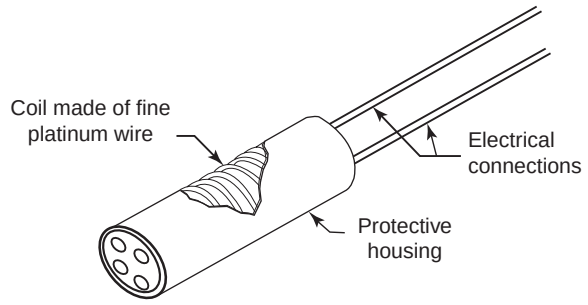


FIGURE 4 A common type of resistance temperature detector employing a fine platinum wire. The cylindrical package may be as small as 3 mm in diameter.

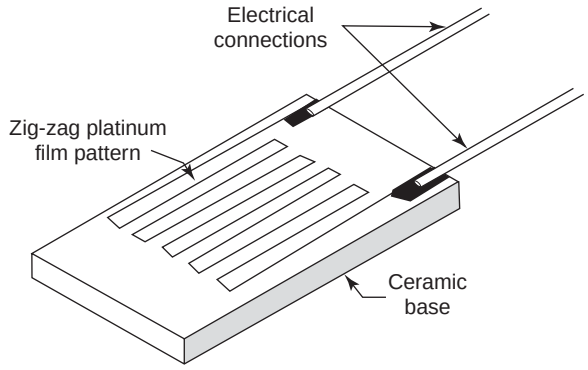


FIGURE 5 A common type of resistance temperature detector employing a thin platinum film. The overall dimensions may be as small as $2 \times 2 \times 0.5$ mm thick.

thickness ranges from 1 to 2 μm with a 10-μm thick protective coating. Precise resistances can be obtained by trimming the length of the film. The standard resistance is 100 Ω at 0°C, but by controlling the dimensions of the film (especially the thickness), a much higher resistance is achievable. A thin film platinum resistance temperature sensor can be bonded to a material to measure temperature in a very local area. Because of the small overall size and good thermal contact, the sensor responds quickly to temperature changes, making it very useful in temperature control applications. Depending on the construction, a platinum resistance temperature sensor can measure temperatures in the range of 10 to 1200 K.

6.1.2 Thermistors

To a first approximation, a semiconductor is a material having electrical properties between a conductor (e.g., copper) and an insulator (e.g., glass). Elemental silicon and germanium, as well as many oxides of metals such as nickel and copper, can have semiconducting properties. The electrical resistance of a semiconductor depends on temperature, but unlike a metal, the resistance decreases as the temperature increases and the relation between resistance and temperature is highly nonlinear. A sensor using such a semiconducting material is called a thermistor. Resistance (R) and Kelvin temperature (T) for a thermistor are generally assumed to be related by $R = R_o e^{\beta(1/T - 1/To)}$, where R_o is the resistance at temperature T_o and, depending on the material, β ranges from 3500 to 4600 K. A platinum wire having a resistance of 100 Ω at 0°C will have approximately 40% more resistance at 100°C. But the resistance of a semiconductor may change by a factor of 50 in the same temperature interval. Thermistors can be made in nearly any stable geometric shape, with the largest overall dimension being approximately 1 mm. Thermistors are relatively cheap, their response time is short, and their intrinsic accuracy is good. The fact that they are highly nonlinear is a disadvantage, as is the limited temperature range, typically −100 to 150°C. Still, thermistors find many uses in this temperature range, for example, as the sensor in a nurse's thermometer.

6.2 Thermocouples

In 1826, Thomas Seebeck discovered that a voltage is produced when there is a temperature gradient along a conductor. This phenomenon is called the Seebeck effect. The voltage produced is called the Seebeck voltage, and the variation of voltage with temperature is called the Seebeck coefficient. Two wires of different composition bonded at one end produce a thermocouple (Fig. 6). A voltage develops between the unconnected ends and is related to the temperature, making a thermocouple useful as a temperature sensor. At 25°C, a thermocouple having one wire made of pure copper and the other wire made of an alloy of copper and nickel called Constantan produces 40.6 microvolts (μV) for each Celsius degree change. The voltage produced by a thermocouple increases monotonically with temperature and is nearly linear over wide ranges of temperature. Thermocouples are widely used in industry, even though sensitive instruments are required to measure the small voltages produced.

Like an ordinary battery, there is a positive and negative polarity to the voltage produced by a thermocouple. When a meter is connected to measure a thermocouple voltage, as shown in Fig. 6, the thermocouple wires and meter connections form two unwanted thermocouple junctions. These two unwanted thermocouple junctions are of

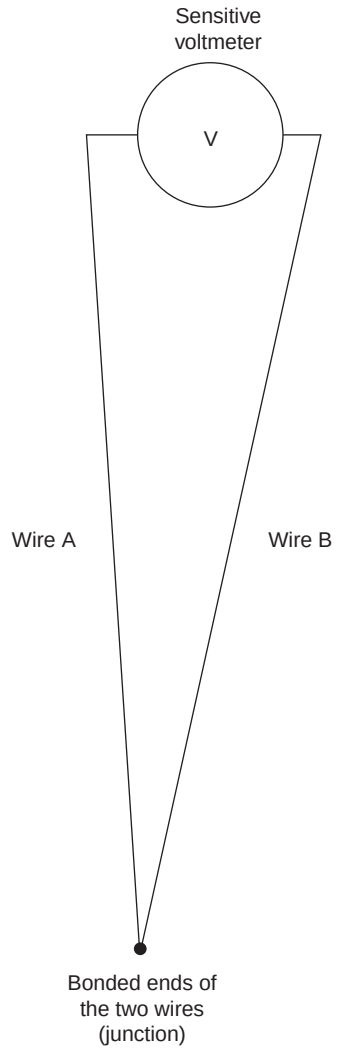


FIGURE 6 Two different metallic wires bonded at their ends constituting a thermocouple. A voltage (V) develops between the open ends when the bond (junction) is heated.

different character because the wires have different composition. The voltages produced at the unwanted thermocouple junctions add to the voltage produced at the thermocouple junction. This problem can be eliminated by using two thermocouples and a reference temperature of 0°C, as shown in Fig. 7. (Although a bath of water and ice is commonly used to produce the 0°C reference, it can be simulated using electronic techniques.) The thermocouple wires and the meter connections now form two identical thermocouple junctions whose voltages cancel each other. The two junctions produce voltages of opposite polarity, so that when the two junctions are at the temperature of the ice bath, the voltage read by the meter is zero. When

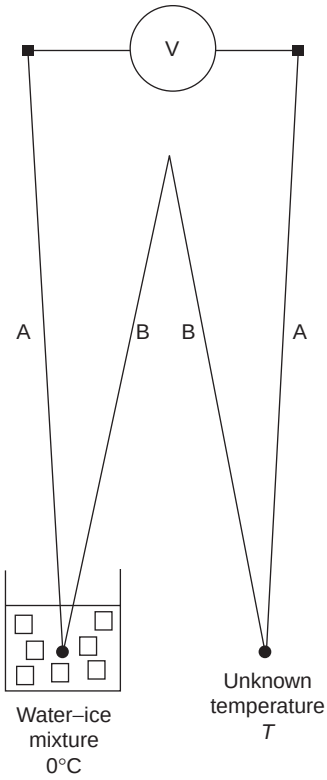


FIGURE 7 Customary manner of measuring temperature with a thermocouple.

one of the thermocouples is removed from the ice bath and is at a different temperature, the voltage read by the meter is not zero and the voltage relates to the temperature difference of the measuring thermocouple and the ice bath. This voltage–temperature relationship may be in the form of a table in a book or may be available in a format suitable for computer use. Because the voltage–temperature relationship is nearly linear, the relation between voltage and temperature can often be suitably represented by a quadratic equation of the form $T = a\mathrm{V} + b\mathrm{V}^2$, where a and b are determined from data for a particular type of thermocouple. The National Institute of Standards and Technology (NIST) maintains a database for a variety of thermocouples. Parameters for functions describing the relation between voltage and temperature are included in the NIST database. Positive features of thermocouples include the following:

- They are relatively inexpensive.
- They are rugged and not easily broken.
- They are small enough to measure temperature at a point.

TABLE II

Characteristics of Commonly Used Thermocouples

Composition A	Composition B	Commercial type	Calibration range (°C)	Sensitivity at 25°C (μV/C°)
Copper	Constantan (alloy of copper and nickel)	T	−270 to ∼400	41
Iron	Constantan	J	−210 to ∼1200	52
Chromel (alloy of nickel and chromium)	Alumel (alloy of nickel and aluminum)	K	−270 to ∼1350	41
Chromel	Constantan	E	−270 to ∼1000	61
Platinum	87% platinum 13% rhodium	R	−50 to ∼1750	6

Source. National Institute of Standards and Technology (n.d.).

- They operate over fairly wide temperature ranges.
- They have relatively short response times.

Negative aspects include the following:

- The low-voltage output is masked by electronic noise.
- Their accuracy is usually limited to approximately 0.5°C.
- They require a known temperature reference, usually 0°C, that is cumbersome unless done electronically.
- Dealing with the nonlinearity is tedious if using calibration tables.

The characteristics of commonly used thermocouples are presented in Table II.

6.3 Radiation Thermometers

A temperature sensor requiring contact with an object (e.g., a thermocouple) cannot withstand a temperature above approximately 2000 K. Thermometers measuring temperatures above that temperature are based on the character of electromagnetic radiation emitted by an object. All matter above a temperature of 0 K emits electromagnetic radiation. A radiator called a blackbody emits more radiation for all wavelength intervals than does any object having the same area and at the same temperature. The ratio of the intensity of radiation from an object to the intensity of radiation from a blackbody at the same temperature is called the emissivity. By definition, the emissivity of a blackbody is 1.0, whereas that for any other object is always less than 1.0.

Perfect blackbodies do not exist. However, they can be approximated to a high degree of accuracy and are very useful in thermometry. Blackbody radiation

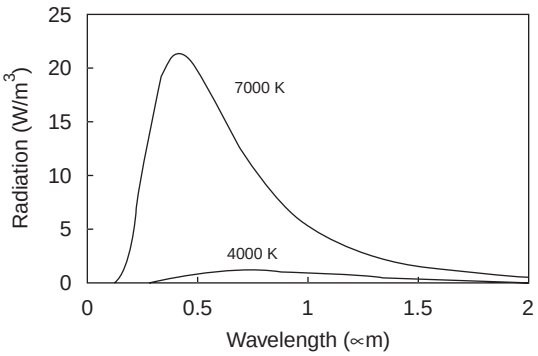


FIGURE 8 Blackbody thermal radiation distributions for temperatures of 4000 and 7000 K.

measured in watts per square meter per wavelength interval in meters as a function of wavelength (λ) is described by Planck's radiation law:

$$R(\lambda, T) = \frac{2\pi hc^2}{\lambda^5} \frac{1}{e^{hc/\lambda kT} - 1}, \tag{2}$$

where $h = 6.63 \times 10^{-34}$ J s, $c = 3.00 \times 10^8$ m/s, and $k = 1.38 \times 10^{-23}$ J/K.

The intensity distributions for blackbodies at temperatures of 7000 and 4000 K are shown in Fig. 8. Two features of blackbody radiation are of particular importance. First, the energy associated with each wavelength depends on the wavelength. Second, the temperature and the wavelength of maximum radiation (λ_m), as noted by the position of the peak in the spectrum, shifts to a longer wavelength as the temperature decreases. The wavelength of maximum radiation and Kelvin temperature (T) are related by Wien's displacement law:

$$\lambda_m T = 2898 \text{ μm kelvins}. \tag{3}$$

For temperatures between approximately 4000 and 7000 K, λ_m is between 0.4 and 0.7 μm that are visible.

For temperatures below approximately 4000 K, the radiation is mostly infrared. Two categories of noncontact thermometers are in common use. One type, called optical pyrometer, uses Planck's law and measurements of thermal power in the range of 0.4 to 0.7 μm, usually approximately 0.66 μm, to infer the temperature of the source. The second kind, called infrared thermometer, uses Planck's law and measurements of thermal power in the range of 0.7 to 80 μm to deduce the temperature of the source.

6.3.1 Optical Pyrometer

A schematic diagram of an optical pyrometer is shown in Fig. 9. The operator sights the pyrometer on an object. The optical system allows viewing of both the target and an image of the filament of a lamp. The operator adjusts the current in the filament, changing its color, until it matches the color of the object. The temperature is inferred from a calibration relating temperature to current in the filament. A second design maintains a constant current in the filament and changes the brightness of light from the object by means of a rotatable, energy-absorbing optical wedge. The object temperature is related to the amount of energy absorbed by the wedge, which is a function of its angular position. The optical pyrometer principle can be extended into the infrared region of the electromagnetic spectrum using detectors sensitive to infrared radiation. Typically, optical pyrometers measure temperature to an accuracy of 1 to 2% of the full-scale range.

6.3.2 Infrared Thermometer

An infrared thermometer uses a lens system to focus radiation onto an infrared detector that converts the energy absorbed into an electrical signal. The temperature inferred from the electrical signal is corrected for the emissivity of the source. Using optical filters, infrared thermometers may employ a very narrow range of wavelengths, whereas other systems may use a very broad range of wavelengths. In either case, the energy absorbed is related to temperature using Planck's law. Broadband infrared spectrometers are relatively inexpensive and easy to use, but their accuracy suffers from the fact that emissivity depends on wavelength, thereby making corrections difficult. Narrow-band infrared thermometers do not face this limitation because the emissivity value does not vary appreciably over the wavelengths used (Fig. 10).

Modern technology is continually improving the accuracy of infrared thermometers. For example, fiber-optic sensors allow placement of the instrument controls away from the heat of the object being studied. Recent developments in infrared thermometers have combined pulse laser technology with narrow-band, single-color infrared thermometers to automatically determine emissivity values and accurately correct for true target temperature. This technology incorporates the use of a pulse laser that emits energy within the same narrow bandwidth that the instrument measures the target thermal radiation. The instrument's infrared detector measures the amount of laser energy reflected from the target, and the microprocessor control converts this additional energy into an accurate value to compensate for emissivity. The instrument then displays true target temperatures, corrected for emissivity and accurate to within ± 3 K. Pulse laser technology is very effective with diffuse targets.

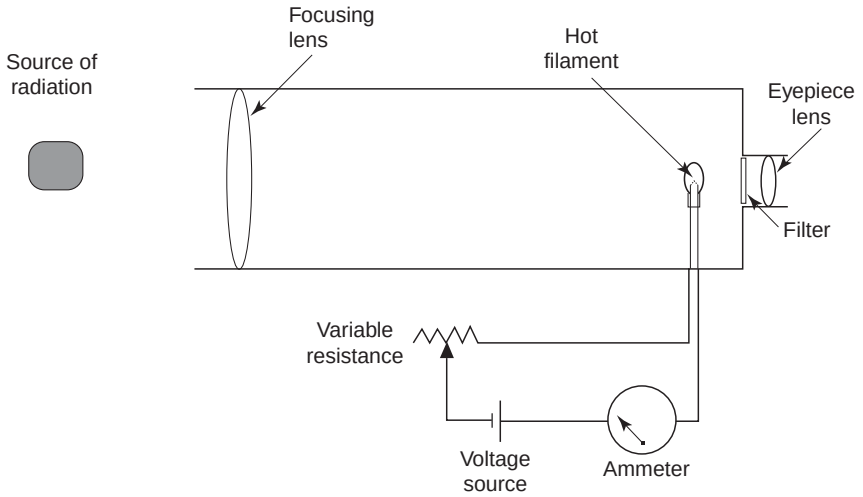


FIGURE 9 Schematic diagram of an optical pyrometer.

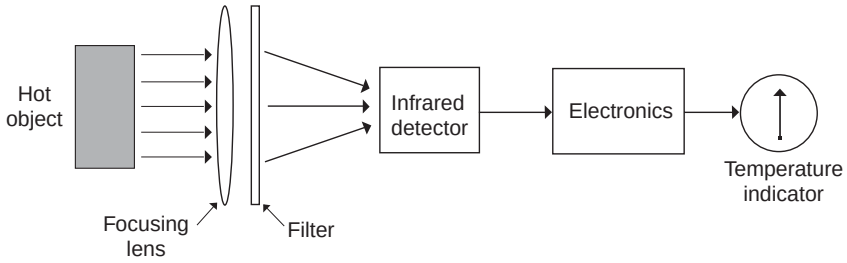


FIGURE 10 Schematic diagram of an infrared thermometer.

6.4 Very Low-Temperature Methods

Measuring temperatures below a few kelvins requires sophisticated techniques. Three of these schemes are called noise thermometry, paramagnetic thermometry, and semiconductor resistance and diode thermometry.

6.4.1 Noise Thermometry

Thermal agitation of electrons in a conductor gives rise to a randomly varying voltage. The average of the square of this voltage is directly proportional to the Kelvin temperature. Properly calibrated, a measurement of the voltage yields the Kelvin temperature. Although the technique requires sophisticated electronic instrumentation, it is one of the few alternatives to gas thermometry for precise determination of thermodynamic temperature. Noise thermometers are very useful for measuring temperatures below 1 K, but in principle they can measure temperatures in the range of a few millikelvins to above 1800 K.

6.4.2 Paramagnetic Thermometry

Many materials can be magnetized by placing them in a magnetic field. The ratio of magnetization to the applied magnetic field is called the magnetic susceptibility. For a class of materials called paramagnetic, the magnetic susceptibility is inversely proportional to the Kelvin temperature. The temperature of a paramagnetic material, such as cerium magnesium nitrate, in intimate contact with some material can be determined from a measurement of the magnetic susceptibility. The technique is useful for measuring temperatures below 4 K.

6.4.3 Semiconductor Resistance and Diode Thermometry

The electrical resistance of a semiconductor depends sensitively on temperature for temperatures below approximately 100 K. Germanium and silicon have proven to be useful for measuring temperatures as low as 0.1 K. The use of ruthenium oxide in a similar mode has pushed the lower limit of temperature measurement to approximately 0.01 K.

A p–n junction diode is a sandwich of two semiconducting materials having different electrical conduction characteristics. When connected to a voltage source, the current in a p–n junction diode depends on temperature. Thermometers using this thermometric property in silicon and gallium–arsenide p–n junction diodes are useful for measuring temperatures in the range of 1 to 300 K.

SEE ALSO THE FOLLOWING ARTICLES

Combustion and Thermochemistry • Conservation of Energy Concept, History of • Energy in the History and Philosophy of Science • Forms and Measurement of Energy • Heat Transfer • Thermodynamics, Laws of

Further Reading

Childs, P. R. N., Greenwood, J. R., and Long, C. A. (2000). Review of temperature measurement. *Rev. Sci. Inst.* **71,** 2959–2978.

Liptak, B. G. (1993). "Temperature Measurement." CRC Press, Boca Raton, FL.

McGee, T. D. (1988). "Principles of Methods of Temperature Measurement." Wiley Interscience, New York.

Michalski, L., Eckersdorf, K., and McGhee, J. (1991). "Temperature Measurement." John Wiley, New York.

National Institute of Standards and Technology. (n.d.). NIST ITS-90 thermocouple database. http://srdata.nist.gov/.

Recktenwald, G. (2001). Temperature measurements with thermocouples. www.me.pdx.edu/~gerry/class/me410epk/notes/thermocouple.pdf.

Schooley, J. F. (1992). "Temperature: Its Measurement and Control in Science and Industry." American Institute of Physics, College Park, MD.

Serway, R. A., and Beichner, R. J. (2000). "Physics for Scientists and Engineers." Brooks/Cole, Pacific Grove, CA.

Thermal Comfort

GIANFRANCO RIZZO and MARCO BECCALI
University of Palermo
Palermo, Italy
ANTONINO NUCARA
University "Mediterranea" of Reggio Calabria
Reggio Calabria, Italy

1. **The Human Body's Thermal Balance**
2. **Methods for Evaluating Thermal Comfort**
3. **Local Evaluation of Thermal Comfort in Confined Environments**
4. **Methods for Evaluating Thermal Stress Conditions**
5. **Other Issues Regarding Thermal Comfort**

Glossary

acceptable thermal environment An environment in which at least 80% of the occupants would find conditions thermally acceptable.
equivalent temperature The air temperature of an imaginary environment in which an occupant would feel the same thermal sensation as in the actual environment.
global thermal comfort A condition of mind that expresses satisfaction with the thermal environment, without mental or physical stresses.
heat (cold) stress Increase (decrease) of body temperature induced by sustained heat (cold) imbalance that cannot be fully compensated for by temperature regulation or by the activation of thermoeffector activities or other physiological regulatory systems.
local thermal discomfort Subjective dissatisfaction regarding the thermal environment, based on local conditions.
moderate thermal environment An environment in which thermal comfort conditions can be achieved with ordinary active (HVAC) or passive devices/techniques.
severe thermal environment An environment in which it is not possible to reach comfort conditions and the occupants' health may be at risk.
thermal load of the human body The difference between internal heat production and loss of heat to the actual environment for people considered to be in a condition of comfort.

Thermal comfort in built environments is an everyday issue for architects and technicians and one that has been dealt with widely in scientific literature. This article provides an overview of the concepts and methods relating to the definition of thermal comfort and the possible ways of assessing it. The ways of controlling thermal conditions and ensuring that built environments are suitable are discussed. Specific attention is given to moderate thermal environments, in which thermal comfort conditions for people are produced using natural or mechanical systems. Indeed, these systems play a major role during the process of designing built environments. Some attention is also given to severe thermal environments (in which conditions of thermal stress should be applied). Taking these aspects into consideration, a description of some rational methods for evaluating thermal comfort and thermal stress conditions is provided. These are essentially based on the human body's thermal balance equation, originally proposed by P. O. Fanger and now included in some International Organization for Standardization International Standard Certificate requirements.

1. THE HUMAN BODY'S THERMAL BALANCE

Through the use of its complex temperature regulation mechanism, the human body is able to reach a condition of thermal equilibrium with the surrounding environment when the variation of internal energy, S, at the body core level is equal to zero. The thermal exchange between a subject and the environment is equal to the difference between metabolic heat production (apart from the so-called mechanical power) and thermal losses due to respiration and exchanges of heat through the skin surface (nude and clothed areas). All these elements

 © 2004 Elsevier Inc. All rights reserved.

relating to the human body's thermal balance are described in the following sections here.

1.1 Internal Heat Production

Food and drink, when ingested, are subjected to exoenergetic chemical reactions; these constitute the metabolic heat power, M (or the metabolic rate of the subject). Through these processes, most of the food's potential chemical energy is transformed into thermal energy.

Part of this thermal energy is needed to maintain the human body's core at a normal temperature and to ensure that the body's internal organs receive everything that is necessary for their normal functioning. This rate of energy (accounting for approximately 43 W/m^2 of human body surface) is called basal metabolism, and it varies depending on the subject's height and weight. Another part of heat production is necessary for the mechanical activities performed by the subject. Through normal activities, part of the metabolic heat production is converted into mechanical power, W, which according to thermodynamic conventions is assumed as positive when it is released in the environment.

1.2 Respiratory Losses

During the activity of respiration, a certain amount of air is inhaled. This air generally has a different temperature and level of humidity with respect to the core. Moreover, a portion of the body's internal heat is transferred, in proportion to the level of activity, to the external environment by means of convective (C_{res}) and evaporative (E_{res}) exchanges. Therefore, respiratory losses depend on the level of activity performed by the subject and the temperature and humidity of the air in the environment.

1.3 Body Surface Losses

The external surface of the human body is generally characterized by thermohygrometric conditions that are generally different from those of the surrounding environment. This induces thermal exchanges between the subject and environment, that is, dry heat losses [by convection (C), radiation (R), and conduction (K)] and latent heat losses (due to the evaporation of the sweat, E).

Convective heat exchange depends on the relative air velocity, v_{ar}, and on the difference between the clothed surface temperatures, t_{cl}, defined as the mean temperature of the clothed surface, including uncovered parts, and the air temperature, t_a. Radiation heat exchange depends on the difference between the clothed surface temperature, t_{cl}, and the mean radiant temperature, $\bar{t}_r$, which is defined as the uniform temperature of an imaginary black enclosure in which an occupant would exchange the same amount of heat by radiation as in a real nonuniform enclosure. Both these heat exchanges depend on the thermal insulation of clothing, I_{cl}. Accordingly, the clothed surface temperature can be expressed as a function of the clothing's thermal insulation, I_{cl}, and of the mean skin temperature, t_{sk}.

Conductive heat exchange, caused by contact with objects, is usually neglected due to its low entity.

Finally, evaporative heat exchange through sweating depends on the clothing's evaporative resistance, $R_{e,T}$, on the skin's wetness, w (defined as the equivalent fraction of the skin's surface that can be considered fully wet), and on the difference between the water vapor pressure in the air and that in the thin liquid film that covers the skin's surface. All heat exchanges through the skin also depend on the size of the body's surface area.

1.4 Balance Equation

By applying the first law of thermodynamics to the human body, it is possible to assess its thermal balance by means of the following well-known equation:

$$S = (M - W) - (C_{res} + E_{res} + C + R + E).$$

All the terms of this equation, but the term due to the evaporation of the sweat, E, can be analytically expressed as a function of the physical and physiologic parameters reported in Table I. As a consequence, the human body's thermal balance equation can be

TABLE I

Physical Parameters Involved in the Thermal Balance Equation

Terms appearing in the thermal balance equation	Physical parameters						
	M	I_{cl}	t_a	v_{ar}	$\bar{t}_r$	p_a	t_{sk}
Metabolic power, M	✓						
Mechanical power, W	✓						
Respiratory convective heat loss, C_{res}	✓		✓				
Respiratory evaporative heat loss, E_{res}	✓					✓	
Convective heat loss, C		✓	✓	✓			✓
Radiation heat loss, R		✓			✓		✓

represented in the following form:

$$S = f(M, I_{cl}, t_a, p_a, v_{ar}, \bar{t}_r, t_{sk}, E),$$

where metabolic rate, M, and thermal insulation of clothing, I_{cl}, are closely related to the subject's activity and clothing. Although the four objective parameters (temperature, t_a, partial vapor pressure, p_a, air velocity, v_{ar}, and mean radiant temperature, $\bar{t}_r$) depend on the environment's microclimatic conditions, while the subjective parameters (mean skin temperature, t_{sk}, and evaporative heat losses through the skin, E) depend on the human body's thermoregulation mechanism.

2. METHODS FOR EVALUATING THERMAL COMFORT

For a practical assessment of a given environment's internal thermal conditions, some helpful indexes are generally used. These can be divided into two main categories: temperature indexes and sensation indexes.

Temperature indexes express the thermal response of the subject in reference to so-called equivalent temperatures, which are the temperatures of imaginary environments in which the occupants would feel the same thermal sensation as in the real environments. Sensation indexes, based on a thermal sensation scale, predict the mean value of the votes expressed by a large group of people exposed to the same microclimatic conditions. Parameters deriving from the application of the human body's thermal balance equation are known as rational indexes, whereas the others are known as empirical indexes.

2.1 The Predicted Mean Vote Method

The most widely applied index for the evaluation of indoor thermal conditions in moderate environments is the predicted mean vote (PMV). This rational sensation index, originally introduced by P. O. Fanger on the basis of a series of experimental studies, represents the mean value of the votes expressed by a large group of people exposed to the same thermal conditions. These votes are linked to the environment on the basis of the American Society of Heating, Refrigerating and Air-Conditioning Engineers thermal sensation scale reported in Table II. It thus provides information about any deviation from the comfort equation, that is, the degree of thermal discomfort experienced by people in the assigned environment.

TABLE II

American Society of Heating, Refrigerating and Air-Conditioning Engineers Thermal Sensation Scale

Vote	Sensation
+3	Hot
+2	Warm
+1	Slightly warm
0	Neutral
−1	Slightly cool
−2	Cool
−3	Cold

The main physical parameter on which the PMV index evaluation relies is the thermal load of the human body. This can be defined as the difference between internal heat production and loss of heat to the actual environment for a person hypothetically kept at the comfort condition. It is assumed that the thermal sensation at a given level of activity can be expressed as a function of the human body's thermal load.

In addition, comfort conditions may be reached only if the mean skin temperature and the thermal exchange due to the evaporation of sweat assume particular values, which depend on the activity of the subject; in this way, the number of independent variables characterizing the thermal balance equation is reduced from eight to six. The PMV index can then be functionally expressed as follows:

$$\mathrm{PMV} = f(M, I_{cl}, t_a, p_a, v_{ar}, \bar{t}_r).$$

This equation is quite difficult to solve, and it requires the support of software since some iterative techniques are also involved.

The evaluation of PMV, however, is made possible through the use of tables that contain recalculated PMV values for several typical combinations of the input parameters. An example of such tables, referring to a humidity ratio of 50% and a metabolic rate of 58 W/m^2, is shown in Table III.

Parameters for these tables are represented by operative temperatures, which are defined as the uniform temperature of an imaginary black enclosure in which an occupant would exchange the same amount of heat by radiation and convection as in the real nonuniform environment. Operative temperature is often used in moderate environments instead of the air temperature and mean radiant temperature. However, it remains a mathematical function of this parameter.

TABLE III

Example of a Table for Evaluating the PMV[a]

Thermal insulation of clothing		Operative temparature (°C)	Relative air velocity (m/s)							
Clo	m^2°C/W		<0.10	0.10	0.15	0.20	0.30	0.40	0.50	1.00
0	0	26	−1.62	−1.62	−1.96	−2.34				
		27	−1.00	−1.00	−1.36	−1.69				
		28	−0.39	−0.42	−0.76	−1.05				
		29	0.21	0.13	−0.15	−0.39				
		30	0.80	0.68	0.45	0.26				
		31	1.39	1.25	1.08	0.94				
		32	1.96	1.83	1.71	1.61				
		33	2.50	2.41	2.34	2.29				
0.25	0.039	24	−1.52	−1.52	−1.80	−2.06	−2.47			
		25	−1.05	−1.05	−1.33	−1.57	−1.94	−2.24	−2.48	
		26	−0.58	−0.61	−0.87	−1.08	−1.41	−1.67	−1.89	−2.66
		27	−0.12	−0.17	−0.40	−0.58	−0.87	−1.10	−1.29	−1.97
		28	0.34	0.27	0.07	−0.09	−0.34	−0.53	−0.70	−1.28
		29	0.80	0.71	0.54	0.41	0.20	0.04	−0.10	−0.58
		30	1.25	1.15	1.02	0.91	0.74	0.61	0.50	0.11
		31	1.71	1.61	1.51	1.43	1.30	1.20	1.12	0.83
0.50	0.078	23	−1.10	−1.10	−1.33	−1.51	−1.78	−1.99	−2.16	
		24	−0.72	−0.74	−0.95	−1.11	−1.36	−1.55	−1.70	−2.22
		25	−0.34	−0.38	−0.56	−0.71	−0.94	−1.11	−1.25	−1.71
		26	0.04	−0.01	−0.18	−0.31	−0.51	−0.66	−0.79	−1.19
		27	0.42	0.35	0.20	0.09	−0.08	−0.22	−0.33	−0.68
		28	0.80	0.72	0.59	0.49	0.34	0.23	0.14	−0.17
		29	1.17	1.08	0.98	0.90	0.77	0.68	0.60	0.34
		30	1.54	1.45	1.37	1.30	1.20	1.13	1.06	0.86
0.75	0.116	21	−1.11	−1.11	−1.30	−1.44	−1.66	−1.82	−1.95	−2.36
		22	−0.79	−0.81	−0.98	−1.11	−1.31	−1.46	−1.58	−1.95
		23	−0.47	−0.50	−0.66	−0.78	−0.96	−1.09	−1.20	−1.55
		24	−0.15	−0.19	−0.33	−0.44	−0.61	−0.73	−0.83	−1.14
		25	0.17	0.12	−0.01	−0.11	−0.26	−0.37	−0.46	−0.74
		26	0.49	0.43	0.31	0.23	0.09	0.00	−0.08	−0.33
		27	0.81	0.74	0.64	0.56	0.45	0.36	0.29	0.08
		28	1.12	1.05	0.96	0.90	0.80	0.73	0.67	0.48
1.00	0.155	20	−0.85	−0.87	−1.02	−1.13	−1.29	−1.41	−1.51	−1.81
		21	−0.57	−0.60	−0.74	−0.84	−0.99	−1.11	−1.19	−1.47
		22	−0.30	−0.33	−0.46	−0.55	−0.69	−0.80	−0.88	−1.13
		23	−0.02	−0.07	−0.18	−0.27	−0.39	−0.49	−0.56	−0.79
		24	0.26	0.20	0.10	0.02	−0.09	−0.18	−0.25	−0.46
		25	0.53	0.48	0.38	0.31	0.21	0.13	0.07	−0.12
		26	0.81	0.75	0.66	0.60	0.51	0.44	0.39	0.22
		27	1.08	1.02	0.95	0.89	0.81	0.75	0.71	0.56
1.25	0.194	16	−1.37	−1.37	−1.51	−1.62	−1.78	−1.89	−1.98	−2.26
		18	−0.89	−0.91	−1.04	−1.14	−1.28	−1.38	−1.46	−1.70
		20	−0.42	−0.46	−0.57	−0.65	−0.77	−0.86	−0.93	−1.14
		22	0.07	0.02	−0.07	−0.14	−0.25	−0.32	−0.38	−0.56
		24	0.56	0.50	0.43	0.37	0.28	0.22	0.17	0.02

continues

Table III continued

Thermal insulation of clothing		Operative temparature (°C)	Relative air velocity (m/s)							
Clo	m^2°C/W		<0.10	0.10	0.15	0.20	0.30	0.40	0.50	1.00
		26	1.04	0.99	0.93	0.88	0.81	0.76	0.72	0.61
		28	1.53	1.48	1.43	1.40	1.34	1.31	1.28	1.19
		30	2.01	1.97	1.93	1.91	1.88	1.85	1.83	1.77
1.50	0.233	14	−1.36	−1.36	−1.49	−1.58	−1.72	−1.82	−1.89	−2.12
		16	−0.94	−0.95	−1.07	−1.15	−1.27	−1.36	−1.43	−1.63
		18	−0.52	−0.54	−0.64	−0.72	−0.82	−0.90	−0.96	−1.14
		20	−0.09	−0.13	−0.22	−0.28	−0.37	−0.44	−0.49	−0.65
		22	0.35	0.30	0.23	0.18	0.10	0.04	0.00	−0.14
		24	0.79	0.74	0.68	0.63	0.57	0.52	0.49	0.37
		26	1.23	1.18	1.13	1.09	1.04	1.01	0.98	0.89
		28	1.67	1.62	1.58	1.56	1.52	1.49	1.47	1.40

[a] Metabolic rate, 58 W/m^2; humidity ratio, 50%.

2.2 The Predicted Percentage of Dissatisfied People

The PMV represents the mean value of a large group of voting participants. However, even though the mean vote would signal the achievement of thermal comfort conditions, some people may experience cold or hot sensations with respect to the thermal environment. In order to take these deviations from the mean expected conditions into account, an additional index has been introduced, the predicted percentage of dissatisfied (PPD). A dissatisfied person is defined as a subject who, when exposed to a defined thermal load, would vote +3, +2, −2, or −3 on the previously defined thermal sensation scale. This index assesses a quantitative prediction of the percentage of subjects who may declare themselves thermally dissatisfied with the indoor environment conditions. The PPD is therefore a function of the PMV; the relationship linking these two indexes is graphically shown in Fig. 1.

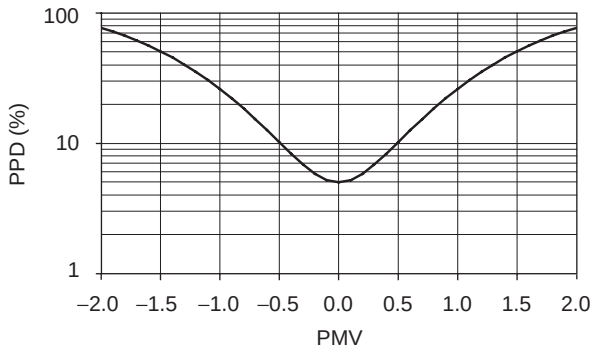


FIGURE 1 Predicted percentage of dissatisfied (PPD) as a function of the predicted mean vote (PMV).

2.3 The Adaptive Method

The PMV and PPD indexes are fairly capable of describing the thermal sensations of subjects in indoor environments equipped with conditioning and ventilation systems. The correspondence between the votes provided by the subjects and the indexes, however, are less obvious in naturally ventilated environments in which mechanical equipment is not present. It has been found that in such environments the occupants generally consider a wider range of thermal conditions acceptable and are inclined to prefer different indoor temperatures than those assumed by using the PMV index. These considerations made it necessary for researchers to develop an alternative approach to thermal comfort assessment.

Since people who are exposed to a given thermal load adapt to the changing environmental conditions, an adaptive comfort method has been proposed. According to this approach, a fixed reference value of indoor temperatures that can be used in each situation does not exist. As a consequence, only optimal values of the climatic conditions that the subjects would adapt to should be considered. The method refers to the comfort temperature, which is considered as a function of the outdoor climatic conditions.

The difficulties in applying the adaptive comfort theory are inherent to the assessment of the analytical relationship between the outdoor and indoor climatic conditions. However, the major criticism of the

method concerns the choice of the external temperature as the only parameter on which comfort would depend. Thus, people's level of activity, their clothing, and typical microclimatic parameters are not considered, even though it has been shown that they play a major role in the thermal sensation process.

3. LOCAL EVALUATION OF THERMAL COMFORT IN CONFINED ENVIRONMENTS

The PMV and PPD indexes, which make it possible to evaluate the comfort conditions of a subject in a confined environment, are useful only for so-called global comfort evaluations. Unfortunately, even if a subject experiences a sensation of thermal neutrality, some parts of his or her body may be exposed to conditions that cannot be considered acceptable. In fact, the nonuniformity of values concerning environmental variables could make people experience sensations of local discomfort.

The main causes of local discomfort are significant vertical air temperature differences, abnormal floor temperatures, significant radiant asymmetry, and air drafts.

3.1 Vertical Air Temperature Differences

Often in indoor environments, air temperature is not constant from the floor to the ceiling; it generally increases with height. If the temperature gradient is high, local discomfort conditions may occur even though the body is thermally neutral from a global standpoint.

These local discomfort conditions are evaluated from the percentage of dissatisfied people, expressed as a function of the difference between the air temperatures at head and ankle levels. Results of some experiments carried out with people in thermal neutrality are shown in Fig. 2.

3.2 Floor Temperature

Thermal sensations resulting from cool or warm floors are due to heat loss from the feet. This heat loss depends on the floor's surface temperature, its conductivity, its constituent materials' heat capacity, and the type of covering worn on the feet. However, for people wearing normal footwear, the floor's material is of little significance. Field studies have made it possible to determine the relationship, shown in Fig. 3, between floor temperature and the percentage of dissatisfied people.

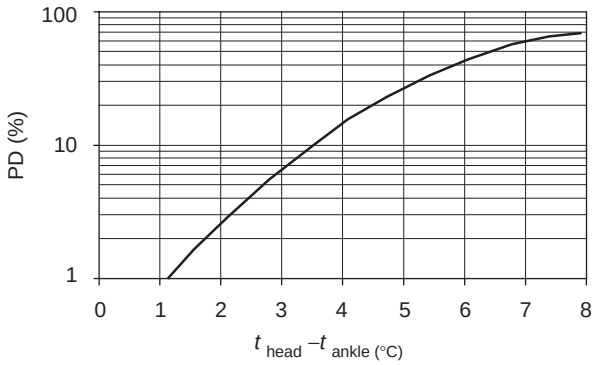


FIGURE 2 Percentage of dissatisfied people (PD) as a function of vertical air temperature differences between head and ankle levels.

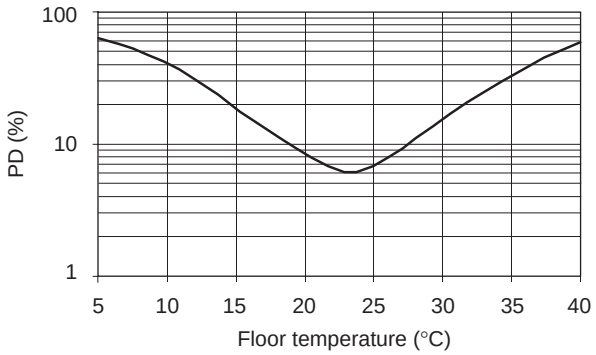


FIGURE 3 Percentage of dissatisfied people (PD) as a function of floor temperature.

3.3 Radiant Asymmetry

When a subject stands too close to a large surface that has a different temperature to that of the indoor air, he or she may feel uncomfortable thermal sensations due to the nonuniformity of thermal radiation in the environment. In order to take this phenomenon into account, the radiant temperature asymmetry parameter is used. This can be defined as the difference between the plane radiant temperatures of the two opposite sides of a small plane element. The plane radiant temperature is the uniform temperature of an enclosure in which the incident of radiation flux on one side of a small plane element is the same as that in the existing environment.

The thermal sensation induced by radiant temperature asymmetry is different for warm or cool surfaces and for vertical or horizontal surfaces. The greatest sensations of discomfort are induced by

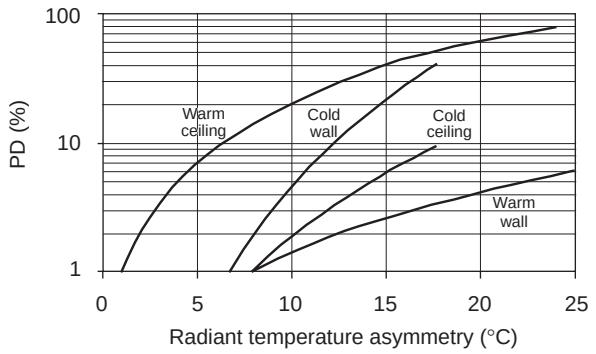


FIGURE 4 Percentage of dissatisfied people (PD) as a function of radiant temperature asymmetry.

warm ceilings and by cold walls. Cold ceilings and warm walls, on the other hand, have less influence on discomfort. Also in this case, the percentage of dissatisfied people as a function of the radiant temperature asymmetry can be used (Fig. 4).

3.4 Drafts

Since air movement in indoor environments often shows turbulent features, airflows could fluctuate significantly in different areas and at different times. Generally, this is caused by drafts, one of the most annoying factors in indoor environments.

A draft is defined as an undesired local cooling of the body caused by air movement. Generally, the sensation of discomfort caused by drafts depends on the airflow temperature (discomfort is less for higher temperatures), the difference between its value and the value of the indoor air's temperature (discomfort increases as this difference increases), the area of the body affected by the air flow (the head, neck, and ankles are the most sensitive zones), and the subject's activity (discomfort decreases with higher activity levels).

Draft rating, which is the percentage of people dissatisfied due to draft, can be evaluated by the following equation:

$$DR = (34 - t_a) \cdot (v - 0.05)^{0.62}(0.37 \cdot v \cdot \mathrm{Tu} + 3.14),$$

where t_a is the local temperature of the airflow, v is the local mean air velocity, and Tu is the local turbulence intensity defined as the ratio between the standard deviation of the local air velocity and the mean local air velocity.

4. METHODS FOR EVALUATING THERMAL STRESS CONDITIONS

4.1 Severe Hot Environments

In severe hot environments, thermoregulation mechanisms are not able to guarantee the thermal neutrality of the human body. This factor creates the need for heat storage and a consequent rise in the body core temperature, which could reach unacceptable levels. In addition, the continuous activity of sweating could lead to unbalances in the body's water and mineral equilibrium.

4.1.1 The Required Sweat Rate Index

Thermal stress conditions in hot environments can be assessed analytically by means of the required sweat rate, SW_{req}, which is a rational index representing the sweat rate that the human body should produce in order to maintain the thermal balance in equilibrium.

The required sweat rate is evaluated beginning with the following equation

$$\mathrm{SW}_{\mathrm{req}} = \frac{E_{\mathrm{req}}}{r_{\mathrm{req}}},$$

where E_{req} is the evaporation required to achieve thermal equilibrium, and r_{req} is the evaporative efficiency at the required sweat rate. The required evaporation is calculated from the thermal balance equation based on the assumption that the internal energy variation is equal to zero and the hypothesis that the mean skin temperature is a function of the microclimatic parameters, the level of activity, and clothing. The resulting equation is as follows:

$$E_{\mathrm{req}} = M - W - E_{\mathrm{res}} - C_{\mathrm{res}} - C - R.$$

Evaporative efficiency is calculated by means of the following algorithm:

$$r_{\mathrm{req}} = 1 - \frac{w^2}{2},$$

where the value of the skin wetness, w, depends on physiological considerations.

The required sweat rate is then compared with the maximum value achievable by a subject. Usually, data are provided both for acclimatized and for nonacclimatized subjects, either at work or at rest. If the resulting values for skin wetness and sweat rate are greater than a maximum threshold value, they may lead to unacceptable levels of heat storage and water loss. In this case, allowable exposure times for people are calculated.

4.1.2 The Wet Bulb Globe Temperatures Index

Evaluation of thermal conditions in severe hot environments can be calculated using a simple empirical index, the wet bulb globe temperatures (WBGT). This temperature index is in fact related to the subject's heat stress.

The evaluation of stress conditions using this index represents a compromise: The simplicity of its use does not in fact correspond to the precision provided by other rational indexes.

The WBGT index is expressed as a function of natural wet bulb temperature, t_{nw}, of the globe temperature, t_g, and of the air temperature, t_a. In particular, the natural wet bulb temperature is measured using a thermometric sensor covered with a wetted wick and naturally ventilated, whereas the globe temperature is the temperature indicated by a sensor placed in the center of a globe with standard characteristics.

Thus, the equations used for the calculation of the WBGT index are

$$\mathrm{WBGT} = 0.7t_{nw} + 0.3t_g,$$

which can be applied in indoor environments or outdoor environments without solar load, and

$$\mathrm{WBGT} = 0.7t_{nw} + 0.2t_g + 0.1t_a,$$

to be applied in outdoor environments in the case of the presence of solar load.

The WBGT value is then compared with reference values supplied for five levels of metabolic rate, both for acclimatized and nonacclimatized subjects. Also, when the WBGT exceeds the reference value, maximum exposure times for people are calculated.

4.2 Severe Cold Environments

In severe cold environments, microclimatic conditions can bring about an undesirable process of cooling in the human body since its thermoregulation mechanisms are not able to maintain the neutrality of the body core. In extreme cases, the continued activity of shivering could cause the subject to die from ventricular fibrillation.

4.2.1 The Required Thermal Insulation Index

The required thermal insulation index (IREQ) is a rational index representing the thermal insulation of a subject's clothes and which is able to maintain heat balance under specified requirements for the physiological strain. This index was defined based on the consideration that a suitable outfit of clothes could suitably protect a subject from cold environmental conditions.

On the basis of the thermal balance equation and the hypothesis that dry heat exchanges are affected by the thermal insulation of the clothing and by the skin-to-clothing surface temperature gradient, the IREQ index can be calculated using the following equations:

$$\begin{cases} \mathrm{IREQ} = \dfrac{t_{sk} - t_{cl}}{M - W - E_{res} - C_{res} - E} \\ M - W - E_{res} - C_{res} - E = R + C \end{cases}.$$

By emphasizing its dependence on the parameters that characterize the thermal balance equation, the IREQ index may be also expressed as follows:

$$\mathrm{IREQ} = f(M, t_a, p_a, v_{ar}, \bar{t}_r, t_{sk}, E).$$

Thermal equilibrium can be achieved at different levels of thermoregulatory strain, depending on the values of the mean skin temperature and sweating. Therefore, IREQ is defined at two different levels, the minimum level ($\mathrm{IREQ_{min}}$) and the neutral required insulation level ($\mathrm{IREQ_{neutral}}$). These represent the thermal insulation required to maintain the body's thermal balance at the so-called "subnormal" and "normal" levels of the mean body temperature, respectively.

The range between $\mathrm{IREQ_{min}}$ and $\mathrm{IREQ_{neutral}}$ indexes can be regarded as a clothing zone regulator, in which a subject can choose the appropriate level of protection. When insulation values are lower than those corresponding to $\mathrm{IREQ_{min}}$, there is a risk of progressive body cooling that may, however, be prevented by imposing a time limit for the subject's exposure.

4.2.2 Wind Chill and Chilling Temperature Indexes

To evaluate the effects of exposition to severe cold conditions on uncovered parts of the human body, two parameters are used: the wind chill index (WCI) and the chilling temperature (t_{ch}). The WCI is an empirical index representing the heat flow exchanged by the exposed parts of the body with the cold environment. It is a function of the air velocity, v_{ar}, and the air temperature, t_a. It is expressed by

$$\mathrm{WCI} = 1.16(10.45 + 10\sqrt{v_{ar}} - v_{ar}) \times (33 - t_a).$$

Often, technicians refer to the chilling temperature, which is a temperature index representing the air temperature of an environment with a value of air velocity equal to 1.8 m/s generating the same cooling power as in the actual environmental conditions. The

chilling temperature is evaluated as a function of the WCI by means of the following equation:

$$t_{ch} = 33 - \frac{WCI}{25.5}.$$

The values of the WCI or the chilling temperature, calculated in the assigned environment, are then compared with target values representing the effect of exposing uncovered parts of the body to cold.

5. OTHER ISSUES REGARDING THERMAL COMFORT

5.1 Energy Saving and Comfort

Energy design for buildings, the goal of which is to establish comfort conditions for the occupants, involves many activities that center around technical choices regarding the building's envelope and the characteristics of its HVAC system. The possibility of creating indoor comfort conditions is in fact closely connected to system design and conduction and to building design and construction.

All these factors will affect the energy demand required to maintain comfort conditions, along with the "quality" of the indoor comfort. It is worth noting that the optimization of these design parameters will generally improve either the amount of energy saved or the quality of the comfort.

In industrialized countries, energy consumption for indoor comfort accounts for approximately 40% of the total final energy consumption, and a significant amount of energy could be saved by "energy -conscious" building design. Possible measures include

- High-efficiency energy-conversion devices (heat pumps, condensation boilers, and cogeneration systems)
- High-efficiency equipment (variable-speed pumps, variable air-volume systems, radiant heating and cooling, heat recovery systems, and smart control systems)
- Use of renewable energy sources (sun, wind, and cold and heat from the ground or from water)
- Building shading, orientation, insulation, and thermal inertia

5.2 Categories of Thermal Environment

The International Organization for Standardization Standard 7730 introduces the concept of thermal comfort categories. Three different comfort tasks (A–C) are defined on the basis of PPD thresholds related either to global comfort or to local discomfort issues. The higher the PPD value, the greater the range of variation allowed for the microclimatic parameter. This will help building designers to tailor their work to different levels of expected performances, according to a sort of quality scale based on three steps (A–C comfort categories). Table IV shows the PPD values corresponding to each category.

5.3 Non-Steady-State Thermal Environments

The PMV–PPD method may be applied only in stationary environments, in which the physical quantities used to describe the subjects' exposure do not depend on time. However, this assumption could only be considered realistic in environments equipped with a well-dimensioned HVAC system.

Nevertheless, in most cases non-steady-state thermal conditions often occur. They are normally classified in three categories: temperature cycles (temperature changes with assigned amplitude and frequencies), temperature drift or ramps

TABLE IV

Catagories of Thermal Environment with Ranges of the Relative Indicators[a]

	Global comfort			Local discomfort		
Category	Predicted percentage of dissatisfied	Predicted mean vote	Percentage of dissatisfied due to draft	Percentage of dissatisfied due to vertical air temperature difference	Percentage of dissatisfied due to warm or cool floor	Percentage of dissatisfied due to radiant asymmetry
A	<6	−0.2–+0.2	<10	<3	<10	<5
B	<10	−0.5–+0.5	<20	<5	<10	<5
C	<15	−0.7–+0.7	<30	<10	<15	<10

[a] Metabolic rate, 58 W/m^2; humidity ratio, 50%.

(monotonic, steady, noncyclic operative temperature changes, where drift refers to passive temperature change and ramps refers to actively controlled temperature changes), and transients (sudden changes in the thermal conditions due to dramatic changes in temperature, humidity, activity, or clothing). However, even when a non-steady-state condition occurs, the relationships established for steady state may still be used in specific situations. For example, they may be used if the peak-to-peak variation of the temperature cycles is less than 1°C, if the rates of temperature change for drift or ramps are lower than 2°C per hour, and in the case of transients after increases in operative temperature. On the contrary, in case of decreasing of the operative temperature, the PMV–PPD method may not be used since it would overestimate high values; in this situation, the time required to achieve a new steady-state condition will depend on the initial conditions.

5.4 Iso-PMV Maps

An environment may be considered homogeneous from a thermal standpoint if the microclimatic parameters around the subject can be considered uniform at a given time. This condition can be easily reached for air temperature, air velocity, and air humidity, but it occurs more rarely for radiant exchanges.

When an environment's microclimate parameters are too heterogeneous, the comfort indexes can be evaluated in several locations. Moreover, taking into account the partial results obtained, the mean value has to be assessed.

Nevertheless, even if this type of analysis leads to a mean value that is used for the whole environment, it does not provide useful information about the local variation of parameters that may occur in the environment.

In this case, it is good practice to utilize iso-PMV or iso-PPD representations, which make it possible to evaluate the local variation of the thermal conditions. In this way, it is also possible to identify the zones in which thermal comfort conditions are not reached. An example of this representation is shown in Fig. 5.

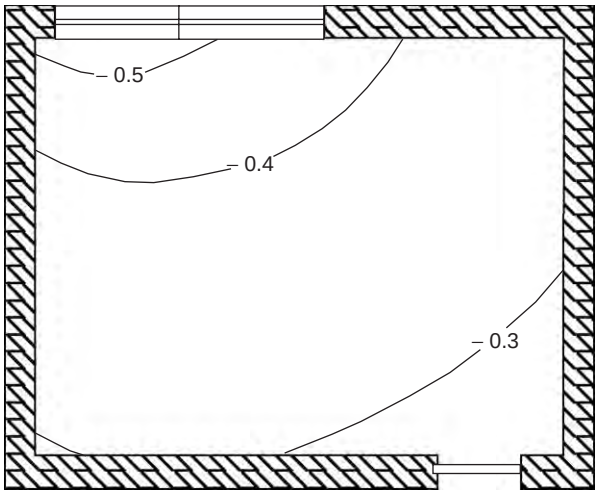


FIGURE 5 Typical example of iso-PMV maps in an environment.

SEE ALSO THE FOLLOWING ARTICLES

Commercial Sector and Energy Use • District Heating and Cooling • Human Energetics • Industrial Energy Use, Status and Trends • Refrigeration and Air-Conditioning • Temperature and Its Measurement • Thermal Energy Storage • Thermoregulation

Further Reading

Fanger, P. O. (1970). "Thermal Comfort." Danish Technical Press, Copenhagen, Denmark.

International Organization for Standardization (1989a). Hot environments—Analytical determination and interpretation of thermal stress using calculation of required sweat rate, ISO Standard 7933. International Organization for Standardization, Geneva.

International Organization for Standardization (1989b). Hot environments—Estimation of the heat stress on working man, based on the WBGT-index (wet bulb globe temperature), ISO Standard 7243. International Organization for Standardization, Geneva.

International Organization for Standardization (1990). Determination of metabolic heat production, ISO Standard 8996. International Organization for Standardization, Geneva.

International Organization for Standardization (1993). Evaluation of cold environments—Determination of required clothing insulation (IREQ), ISO Technical Report 11079. International Organization for Standardization, Geneva.

International Organization for Standardization (1994). Moderate thermal environments—Determination of the PMV and PPD indices and specification of the conditions for thermal comfort, ISO Standard 7730. International Organization for Standardization, Geneva.

International Organization for Standardization (1995). Ergonomics of the thermal environment—Estimation of the thermal insulation and evaporative resistance of a clothing ensemble, ISO Standard 9920. International Organization for Standardization, Geneva.

International Organization for Standardization (1998). Ergonomics of the thermal environment—Instruments for measuring physical quantities, ISO Standard 7726. International Organization for Standardization, Geneva.

Thermal Energy Storage

IBRAHIM DINCER
University of Ontario Institute of Technology
Oshawa, Ontario, Canada

1. **TES and Energy Savings**
2. **TES and Environmental Impact**
3. **TES and Sustainable Development**
4. **TES Methods**
5. **Technical and Operational Aspects of TES Systems**
6. **TES for Cooling Capacity**
7. **Illustrative Examples**
8. **Concluding Remarks**

Glossary

ambient air The air in the space surrounding a thermal energy storage device, or outside air.
ambient heat load The thermal load imposed on a storage device due to heat gain.
build period The operating period of a thermal storage generator during which ice is produced.
charge fluid The heat transfer fluid used to remove heat from a thermal storage device or generator during the charge or build period, or to add heat to a heat storage.
charge period The period of time during which energy (heat) is removed from a cold storage or added to a heat storage.
charge rate The rate at which energy (heat) is removed from or added to a storage device during the charge period.
discharge fluid The heat transfer fluid used to add heat to a thermal storage device.
discharge period The period of time when energy (heat) is added to a storage device.
discharge rate The rate at which energy (heat) is added to a storage device during the discharge period.
latent heat of fusion The change in energy accompanying the conversion of a unit mass of a solid to a liquid at its melting point, at constant pressure and temperature.
melt period That period of a test during which the ice produced by a thermal storage generator is melted to determine its quality.
period An interval of time during which a storage process occurs; often a period such as the total duty cycle of a thermal storage system is divided for the purpose of analysis and evaluation into 1 h time segments.
phase-change material (PCM) A substance that undergoes changes of state while absorbing or rejecting thermal energy, normally at a constant temperature.
thermal storage device Equipment that stores cooling capacity using sensible and/or latent heat. May consist solely of a storage means or be packaged with one or more components of a mechanical refrigeration package.
thermal storage equipment Any one of, or a combination of, thermal storage devices and/or generators, which may include various other components of a mechanical refrigeration package.
ton-hour A quantity of thermal energy typically used to describe the capacity of a thermal storage device, in tons, absorbed or rejected in 1 h (3.517 kW).

Thermal energy storage (TES) generally involves a temporary storage of high- or low–temperature thermal energy for later use. Examples of TES are the storage of solar energy for overnight heating, of summer heat for winter use, of winter ice for space cooling in summer, and of heat or coolness generated electrically during off-peak hours for use during subsequent peak demand hours. In this respect, TES appears to be an excellent candidate to offset this mismatch of availability and demand and is considered an advanced energy technology. Therefore, during the past decade TES has attracted the greatest interest for a wide range of thermal applications, such as space heating, hot water, cooling, and air–conditioning. Why are these established TES technologies not used more than they are? The answer is that TES finds application possibilities if there are some circumstances that require TES as a solution, such as a mismatch between energy available and demand, discontinuity in energy supply, and different electricity rates for electricity consumption during peak and off-peak hours. If such a case(s) exists, it is either encouraged by various mechanisms (e.g., incentives) or mandatorily done by regulations. Of course, TES systems and applications bring some additional costs. However, payback time is quite reasonable due to high energy and environmental

 © 2004 Elsevier Inc. All rights reserved.

costs. TES technology has been used in various forms and specific applications. Some of the more common applications have been the use of sensible heat (water, molten salt) or latent heat (ice, phase-change material) storage for heating and cooling applications. The effectiveness of sensible heat storage depends on the specific heat of the storage material and, if volume is important, on its density. Sensible storage systems commonly use rocks or water as the storage medium and the thermal energy is stored by increasing the storage-medium temperature. Latent heat storage systems store energy in phase-change materials, with the thermal energy stored when the material changes phase, usually from a solid to a liquid. The specific heat of solidification/fusion or vaporization and the temperature at which the phase change occurs are important in terms of design. Both sensible TES and latent TES also may occur in the same storage material. The main goal of this article is to introduce the fundamental aspects of TES systems and discuss their energy savings, environmental impacts, and technical and operational aspects and to highlight their importance with some illustrative and practical examples.

1. TES AND ENERGY SAVINGS

TES systems are an important element of many energy-saving programs in a variety of sectors: residential, commercial, industrial, and utility as well as the transportation sector. TES can be employed to reduce energy consumption or to transfer an energy load from one period to another. The consumption reduction can be achieved by storing excess thermal energy that would normally be released as waste, such as heat produced by equipment and appliances, by lighting, and even by occupants. Energy-load transfer can be achieved by storing energy at a given time for later use and can be applied to TES for either heating or cooling capacity.

The primary objective of most TES systems, which is to alter energy-use patterns so that financial savings occur, can be achieved in several ways:

1.1 Utilization of Waste or Surplus Energy

The consumption of purchased energy can be reduced by storing waste or surplus thermal energy available at certain times for use at other times. Useful waste or surplus thermal energy is available from many sources. Some examples include the following: (1) hot or cold water drained to a sewer; (2) hot flue gases; (3) exhaust air streams; (4) hot or cold gases or waste gases; (5) heat collected from solar panels; (6) ground source thermal energy; (7) heat rejected from the condenser of refrigeration and air-conditioning equipment; and (8) the cooling effect from the evaporator of a heat pump.

1.2 Reduction of Demand Charges

The demand of purchased electrical energy can be reduced by storing electrically produced thermal energy during off-peak periods to meet the thermal loads that occur during high-demand periods. There has been an increasing interest in the reduction of peak demand or transfer of energy loads from high- to low-consumption periods. For example, an electric chiller can be used to charge a chilled cold TES system at night to reduce the electrical demand peaks usually experienced during the day. Figure 1 shows examples of a daily load profile for a building, with and without cool storage. Figure 1A represents the case with no storage and Fig. 1B shows a full-storage case. In the latter case, the TES provides enough storage capacity to meet the peak (i.e., 9:00

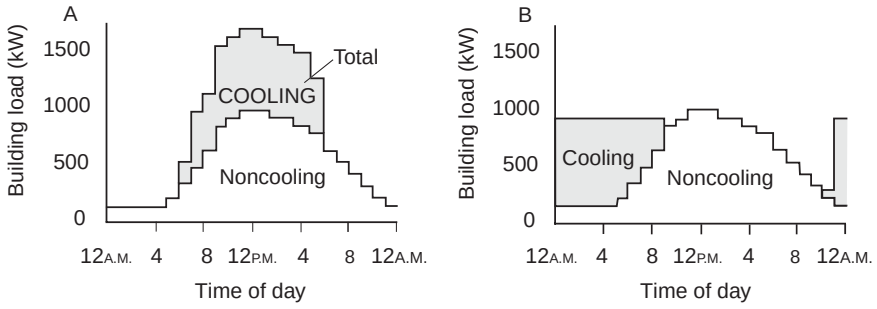


FIGURE 1 Daily load profiles of a building energy load: (A) no storage and (B) full storage. Modified from Dincer *et al.* (1997b).

A.M. to 9:00 P.M.) cooling load, shifting the entire electrical demand for cooling to off-peak hours when there is little cooling load. This particular application achieves maximal benefits in terms of demand charge reduction and use of favorable off-peak rates, both of which act to lower operating costs.

1.3 Deferring Equipment Purchases

The capacity of heating and cooling equipment is normally selected to match the part of the design-day load when the requirements for heating and cooling are close to maximum. These peak design loads occur for only short periods of time, resulting in excess capacity on average days. The use of TES can defer the need to purchase additional equipment for heating, cooling, or air-conditioning applications and reduce equipment sizing in new facilities. The relevant equipment is operated when thermal loads are low to charge the TES and energy is withdrawn from storage to help meet the thermal loads that exceed equipment capacity.

2. TES AND ENVIRONMENTAL IMPACT

TES systems can contribute significantly to meeting society's desire for more efficient, environmentally benign energy use, particularly in the areas of building heating and cooling systems and electric power generation. By reducing energy consumption, the utilization of TES systems results in two significant environmental benefits: the conservation of fossil fuels through efficiency increases and/or fuel substitution, and reductions in emissions of pollutants such as CO_2, SO_2, NO_x, and chlorofluorocarbons (CFCs).

TES can impact air emissions at building sites by reducing (1) the amount of ozone-depleting CFC and hydrochlorofluorocarbon (HCFC) refrigerants in chillers and (2) the amount of combustion emissions from fuel-fired heating and cooling equipment. TES helps reduce CFC use in two main ways. First, since cooling systems with TES require less chiller capacity than conventional systems, they use fewer or smaller chillers with less refrigerant. Second, using TES can offset the lost cooling capacity that sometimes can occur when existing chillers are converted to more benign refrigerants, making building operators more willing to switch refrigerants.

The potential aggregate air-emission reductions at power plants due to TES have been shown to be significant. For example, TES systems have been shown to reduce CO_2 emissions in the United Kingdom by 14 to 46% by shifting electric load to off-peak periods, whereas an Electric Power Research Institute (EPRI)-cosponsored analysis found that TES could reduce CO_2 emissions by 7% compared to conventional electric cooling technologies. Also, using California Energy Commission data indicating that existing gas plants produce approximately 0.06 kg of NO_x and 15 kg of CO_2 per 293,100 kWh of fuel burned, and assuming that TES installations save an average of 6% of the total cooling electricity needs, TES could possibly eliminate annual emissions of approximately 560 tons of NO_x and 260,000 tons of CO_2 statewide.

Establishing concrete methods for promoting TES technologies requires analysis and clarification about how to combine environmental objectives, social and economic systems, and technical developments to attain effective solutions. It is important to create and employ tools that encourage technological development and diffusion and to align government policies in areas such as energy and environmental standards and government purchasing.

3. TES AND SUSTAINABLE DEVELOPMENT

A secure supply of energy resources is generally agreed to be a necessary but not sufficient requirement for development within a society. Furthermore, sustainable development demands a sustainable supply of energy resources. The implications of these statements are numerous and depend on how "sustainable" is defined. One important implication of these statements is that sustainable development within a society requires a supply of energy resources that, in the long term, is readily and sustainably available at reasonable cost and can be utilized for all required tasks without causing negative societal impacts. Supplies of energy resources such as fossil fuels (coal, oil, and natural gas) and uranium are generally acknowledged to be finite; other energy sources, such as sunlight, wind, and falling water, are generally considered renewable and therefore sustainable over the relatively long-term. Wastes (convertible to useful energy forms through, for example, waste-to-energy incineration facilities) and biomass fuels are also usually viewed as sustainable energy sources. A second implication of the initial statements in this section is that sustainable development requires that energy resources be used as efficiently as

possible. In this way, society maximizes the benefits it derives from utilizing its energy resources, while minimizing the negative impacts (such as environmental damage) associated with their use. This implication acknowledges that all energy resources are, to some degree, finite, so that greater efficiency in utilization allows such resources to contribute to development over a longer period of time, i.e., to make development more sustainable. Even for energy sources that may eventually become inexpensive and widely available, increases in energy efficiency will likely remain to be sought to reduce the resource requirements (energy, material, etc.) to create and maintain systems and devices to harvest the energy and to reduce the associated environmental impacts.

The first implication, clearly being essential to sustainable development, has been and continues to be widely discussed. The second implication, which relates to the importance and role of energy efficiency in achieving sustainable development, is somewhat less discussed and understood.

Consequently, sustainable development demands a sustainable supply of energy resources that, in the long-term, is readily and sustainably available at reasonable cost and can be utilized for all required tasks without causing negative societal impacts. TES systems can contribute significantly to meeting society's desire for more efficient, environmentally benign energy use and for sustainable development of the society, particularly in the areas of building heating and cooling systems and electric power generation. By reducing energy consumption, the utilization of TES systems results in two significant benefits: (1) the conservation of fossil fuels through efficiency increases and/or fuel substitution and (2) reductions in emissions of such pollutants as CO_2, SO_2, NO_x, and CFCs.

4. TES METHODS

TES can aid in the efficient use and provision of thermal energy whenever there is a mismatch between energy generation and use. Various subsets of TES processes have been investigated and developed for building heating and cooling systems, industrial applications, and utility and space power systems. Some of the media available for sensible and the latent TES systems are presented in Table I.

TABLE I

Available Media for Sensible and Latent TES Systems

Sensible		Latent
Short–term	**Long–term (Annual)**	**Short–term**
Rock beds	Rock beds	Inorganic materials
Earth beds	Earth beds	Organic materials
Water tanks	Large water tanks	Fatty acids
	Aquifiers	Aromatics
	Solar ponds	

4.1 Sensible TES

In sensible TES, energy is stored by changing the temperature of a storage medium, such as water, air, oil, rock beds, bricks, sand, or soil. The amount of energy input to TES by a sensible heat device is proportional to the difference between the final and initial temperatures of storage and the mass of the storage medium and its heat capacity. Each medium has its own advantages and disadvantages. For example, water has approximately twice the specific heat of rock and soil. The high heat capacity of water often makes water tanks a logical choice for TES systems that operate in a temperature range needed for building heating or cooling. The relatively low heat capacity of rocks and ceramics is somewhat offset by the large temperature changes possible with these materials and their relatively high density.

A sensible TES consists of a storage medium, a container, and input/output devices. Containers must both retain the storage material and prevent losses of thermal energy. Thermal stratification, the existence of a thermal gradient across storage, is desirable. Maintaining stratification is much simpler in solid storage media than in fluids.

Sensible TES materials undergo no change in phase over the temperature range encountered in the storage process. The amount of heat stored in a mass of material can be expressed as

$$Q = mc_{\mathrm{p}}\Delta T \text{ or } Q = \rho c_{\mathrm{p}} V \Delta T,$$

where c_p is the specific heat of the storage material, ΔT is the temperature change, V is the volume of storage material, and ρ is the density of the material. Note that the ability to store sensible heat for a given material depends strongly on the value of the quantity ρc_{p}. Water has a high value and is inexpensive but, being liquid, must be contained in a better quality container than a solid.

Some common TES materials and their properties are presented in Table II. To be useful in TES applications, the material normally must be inexpensive and have a good thermal capacity. Another

TABLE II

Thermal Capacities at 20°C of Some Common TES Materials

Material	Density (kg/m^3)	Specific heat (J/kgK)	Volumetric thermal capacity ($10^6 J/m^3 K$)
Clay	1458	879	1.28
Brick	1800	837	1.51
Sandstone	2200	712	1.57
Wood	700	2390	1.67
Concrete	2000	880	1.76
Glass	2710	837	2.27
Aluminium	2710	896	2.43
Iron	7900	452	3.57
Steel	7840	465	3.68
Gravelly earth	2050	1840	3.77
Magnetite	5177	752	3.89
Water	988	4182	4.17

Source: Adapted from Norton (1992).

important parameter in sensible TES is the rate at which heat can be released and extracted.

4.2 Latent TES

The heat transfer that occurs when a substance changes from one phase to another is called the latent heat. The latent heat change usually is much higher than the sensible heat change for a given medium, which is related to its specific heat. When water turns to steam, the latent heat change is on the order of 2 MJ/kg. Most practical systems using phase-change energy storage involve solutions of salts in water. Several problems are associated with such systems, including the following:

- Supercooling of the phase-change material (PCM) may take place, rather than crystallization with heat release. This problem can be avoided partially by adding small crystals as nucleating agents.
- It is difficult to build a heat exchanger capable of dealing with the agglomeration of varying sizes of crystals that float in the liquid.
- The system operation cannot completely reversed.

Any latent heat TES system must possess at least the following three components:

- A heat storage substance that undergoes a phase transition within the desired operating temperature range and wherein the bulk of the heat added is stored as latent heat.
- A containment for the storage substance.
- A heat-exchange surface for transferring heat from the heat source to the storage substance and from the latter to the heat sink, e.g., from a solar collector to the latent TES substance to the load loop.

Latent TES is a promising storage technique since it provides high-energy storage density, second only to chemical energy storage, and can store and release heat at a constant temperature, corresponding to the phase transition temperature of the heat-storage medium.

An important material category capable of storing energy through phase change is paraffin waxes. These have the advantage of very high stability over repeated cycles of latent TES operation without degradation.

The most important criteria to be met by the storage material for a latent TES in which the material undergoes a solid–liquid or a solid–solid phase transition are as follows: high transition enthalpy per unit mass, the ability to fully reverse the transition, adequate transition temperature, chemical stability, compatibility with the container (if present), limited volume change with the transition, nontoxicity, and low cost, in relation to the foreseen application.

4.2.1 Phase-Change Materials

When a material melts or vaporizes, it absorbs heat; when it changes to a solid (crystallizes) or to a liquid (condenses), it releases this heat. This phase change is used for storing heat in PCMs. Typical PCMs are water/ice, salt hydrates, and certain polymers. Since energy densities for latent TES exceed those for sensible TES, smaller and lighter storage devices and lower storage losses normally result. Like ice and water, eutectic salts have been used as storage media for many decades. Perhaps the oldest application of a PCM for TES was the use of seat warmers for British railroad cars in the late 1800 s. During cold winter days, metal or rubber bags were filled with sodium thiosulfate pentahydrate, a PCM that melts and freezes at 44.4°C, and these were used to warm the seats. Other early applications of PCMs included "eutectic plates" used for cold storage in trucking and railroad transportation applications. Another important application of PCMs was in association with space technology, with National Aeronautics and Space Administration sponsoring a project on PCM applications for thermal control of electronic packages.

A large number of organic compounds suitable for use as storage media for heating, ventilation, and air-conditioning applications have been investigated. The most promising candidates seem to be the normal

paraffins. Their solid–liquid transitions (fusions) meet satisfactorily seven important criteria for PCMs. For example, the heat of fusion has a mean value of 35–40 kcal/kg, there are no problems in the reversing the phase change, and the transition points vary considerably with the number of carbon atoms in the chains. Also, the normal paraffins are chemically inert, nontoxic, and available at reasonably low cost. The change of volume with the transition, which is on the order of 10%, could represent a minor problem.

The zeolites are naturally occurring minerals. Their high heat of adsorption and their ability to hydrate and dehydrate while maintaining structural stability have been found to be useful in various thermal storage and solar refrigeration systems. This hygroscopic property, coupled with the rapid exothermic reaction that occurs when zeolites are taken from a dehydrated to a hydrated form (when the heat of adsorption is released), makes natural zeolites an effective storage material for solar and waste heat energy.

Application of PCMs in the temperature range 0–120°C is of interest for a variety of low-temperature applications, such as direct space heating, domestic hot water production, heat-pump-assisted space heating, greenhouse heating, and solar cooling. The development of dependable TES systems for these and other applications requires a good understanding of heat-of-fusion storage materials and heat exchangers.

Knowledge of the melting and freezing characteristics of PCMs, of their ability to undergo thermal cycling, and of their compatibility with construction materials is essential for assessing the short- and long-term performance of a latent TES. Commercial paraffins are characterized by two phase transitions (solid–liquid and solid–solid) that occur over a large temperature range depending on the paraffin involved. *n*-Paraffins are usually preferred in comparison to their iso-counterparts, as the desired solid-to-liquid phase transition is generally restricted to a narrow temperature range. Fatty acids are organic materials with excellent melting and freezing characteristics and may have a good future potential, if their costs can be reduced. Inorganic salt hydrates, on the other hand, must be carefully examined for congruent, "semicongruent," or incongruent melting substances with the aid of phase diagrams. Incongruent melting in a salt hydrate may be "modified" to overcome decomposition by adding suspension media, or extra water, or other substances that shift the peritectic point. The use of salt hydrates in hermetically sealed containers is normally recommended. Also, the employment of metallic surfaces to promote heterogeneous nucleation in a salt hydrate is seen to reduce the supercooling of most salt hydrates to a considerable extent.

In the employment of PCMs as heat storage materials in TES systems, certain desirable thermodynamic, kinetic, chemical, technical and economic characteristics are essential:

4.2.1.1 Thermodynamic Criteria

- A melting point at the desired operating temperature;
- A high latent heat of fusion per unit mass, so that less material stores a given amount of energy;
- A high density, so that less volume is occupied by the material;
- A high specific heat, so that significant sensible TES can also occur;
- A high thermal conductivity, so that only small temperature differences are needed for charging and discharging the storage;
- Congruent melting; i.e., the material should melt completely, so that the liquid and solid phases are homogenous (this avoids the difference in densities between solid and liquid that causes segregation, resulting in changes in the chemical composition of the material); and
- Small volume changes during phase transition, so that a simple containment and heat exchanger can be used.

4.2.1.2 Kinetic Criterion

- Little or no supercooling during freezing; i.e., the melt should crystallize at its freezing point. This criterion can be achieved through a high rate of nucleation and growth rate of the crystals. Supercooling may be suppressed by introducing nucleating agent or a cold trigger in the storage material.

4.2.1.3 Chemical Criteria

- Chemical stability;
- No susceptibility chemical decomposition, so that a long operation life is possible;
- Noncorrosive behavior to construction materials; and
- Nontoxic, nonflammable, and nonexplosive characteristics.

4.2.1.4 Technical Criteria

- Simplicity;
- Applicability;

- Compatibility; and
- Reliability.

4.2.1.5 Economic Criteria

- Commercial availability, and
- Low cost.

5. TECHNICAL AND OPERATIONAL ASPECTS OF TES SYSTEMS

This section presents certain technical and operational aspects of TES systems and their consequences.

5.1 Basic Principle of TES

The basic principle is the same in all TES applications. Energy is supplied to a storage system for removal and use at a later time. What mainly varies is the scale of the storage and the storage method used. Seasonal storage requires immense storage capacity. One seasonal TES method involves storing heat in underground aquifers. Another suggested method is circulating warmed air into underground caverns packed with solids to store sensible heat. The domestic version of this concept is storing heat in hot rocks in a cellar. At the opposite end of the storage-duration spectrum is the storage of heat on an hourly or daily basis. The previously mentioned use of tiles to store solar radiation is a typical example, which is often applied in passive solar design.

5.2 TES Processes

A complete storage process involves at least three steps: charging, storing, and discharging. A simple storage cycle can be illustrated as in Fig. 2, in which the three steps are shown as distinct stages. In practical systems, some of the steps may occur simultaneously (e.g., charging and storing) and each step may occur more than once in each storage cycle.

5.3 Feasibility of TES

A variety of factors are known to influence dramatically the selection, implementation, and operation of a TES system. Therefore, it is necessary to carry out a comprehensive feasibility that takes into consideration all parameters that impact the cost and other benefits of considered TES systems. However, it is not always possible to conduct all the steps in a feasibility study for an application and, in such instances, as many items as possible should be considered and studied accordingly. In such TES feasibility studies, a checklist can be helpful in ensuring that significant issues related to the project are addressed and details regarding the evaluation, selection, implementation, and operation of the system are assessed correctly. Fig. 3 provides a checklist that can be beneficial to the TES industry and analysts involved in TES projects. A checklist completed in the preliminary stages of the project guides technical staff.

5.4 Integration of TES

When considering the integration of TES into an existing thermal facility, an additional checklist (Fig. 4) can assist in conducting the feasibility study for the TES system and its incorporation into the facility. For a facility under design, the information required is generally the same as in Fig. 4, except that the database must be developed or estimated due to the lack of actual operating data.

6. TES FOR COOLING CAPACITY

Cooling capacity can be stored by either chilling or freezing water (or other materials, such as glycol and eutectic salts). Cold TES (CTES) is an innovative way of storing nighttime off-peak energy for daytime peak use. In many locations, the demand for electrical power peaks during the summer. Air-conditioning is the main reason, in some areas accounting for as much as one-half of the power

FIGURE 2 The three stages in a simple heat storage process: charging period (left), storing period (center), and discharging period (right), where the terms of Q_l are heat losses during these three periods. Reprinted from Dincer and Rosen (2002), © John Wiley & Sons Limited, with permission.

Checklist for Evaluating a General TES Project

Please check (✓) items that are available or known.

( ) 1. Management objectives
( ) 2. Economic objectives
( ) 3. Financial parameter
( ) 4. Available utility incentives
( ) 5. Status of TES system (a) New ( ) (b) Existing ( )
( ) 6. Net heating or cooling storage capacity
( ) 7. Utility rates and associated energy charges
( ) 8. Loading type of TES system (a) Full ( ) (b) Partial ( )
( ) 9. Best possible TES system options
( ) 10. Anticipated operating strategies for each TES system option
( ) 11. Space availability for TES system (e.g., tank)
( ) 12. Type of TES system (a) Open ( ) (b) Closed ( )
( ) 13. General implementation logistics of TES system under consideration
    ( ) 13.1. Status of TES system
    ( ) 13.2. TES system location
    ( ) 13.3. Structural impact
    ( ) 13.4. Heat exchanger requirements
    ( ) 13.5. Piping arrangement
    ( ) 13.6. Automatic control requirements
    ( ) 13.7. New electrical service requirements
    ( ) 13.8. Other

Signature:

Project Leader:

Date:

Project Title and Number:

FIGURE 3 Checklist for evaluating the feasibility of a TES project. Adapted from Dincer (1999).

demand during the hot midday hours when electricity is most expensive. Since, at night, utilities have electrical generating capacity to spare, electricity generated during this "off-peak" time is much cheaper. In essence, one can air-condition during the day using electricity produced at night. CTES has become one of the primary means of addressing the electrical power imbalance between high daytime demand and high nighttime abundance. If properly designed, installed, operated, and maintained, CTES systems can be used to shift peak cooling loads to off-peak periods, thereby, evenly distributing the demand for electricity and avoiding shortages usually encountered during peak periods.

CTES systems are most likely to be cost-effective in situations such as the following: where a facility's maximum cooling load is much greater than the average load; where the utility rate structure has high demand charges, high ratchet charges, or a high differential between on- and off-peak energy rates; where an existing cooling system is being expanded; where an existing tank is available; where limited on-site electric power is available; where back-up cooling capacity is desirable; and where cold-air distribution is desirable or advantageous.

It is difficult to generalize about when cool storage systems will be cost-effective, but if one or more of the above criteria are satisfied, a detailed analysis may prove worthwhile.

6.1 Operational Aspects of CTES

Several strategies are available for charging and discharging a storage to meet cooling demand during peak hours. The strategies are full storage and partial storage.

6.1.1 Full-Storage CTES

A full-storage strategy shifts the entire peak-cooling load to off-peak hours (Fig. 5a). The system is typically designed so that on the days that are anticipated to be the hottest the system operates at

Checklist for Integrating TES into an Existing Thermal Facility

Please check (✓) items that are available or known.

Database

( ) 1. Utility's incentives
( ) 2. Facility occupancy hours
( ) 3. Facility operating requirements
( ) 4. Existing physical constraints
( ) 5. Facility peak-day loads and monthly average requirements
( ) 6. Historic energy consumption rates

Analysis

( ) 1. Best possible TES system options
( ) 2. General implementation logistics of each TES system under consideration
( ) 3. Plant's yearly energy consumption with or without a TES system
( ) 4. Size utilization factor for TES
( ) 5. Projected operating cost reduction for each TES system under consideration

Conclusion

( ) 1. Financial analysis
( ) 2. System implementation recommendation
( ) 3. Other

Signature:
Project Leader:
Date:
Project Title and Number:

FIGURE 4 Checklist for integrating TES into an existing thermal facility. Adapted from Dincer (1999).

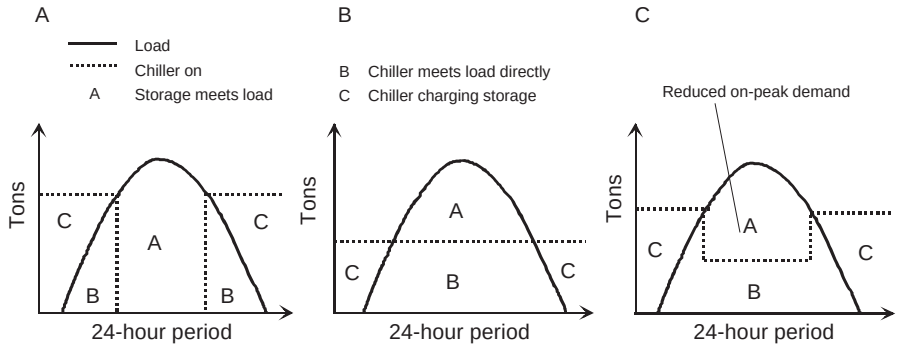


FIGURE 5 Operating strategies. (a) Full storage, (b) partial-storage load leveling, and (c) partial-storage demand-limiting. Reprinted from Dincer and Rosen (2002), © John Wiley & Sons Limited, with permission.

full capacity during all nonpeak hours in order to charge storage. This strategy is most attractive when peak demand charges are high or the peak period is short. For example, a school or business whose electrical demand drops dramatically after 5:00 P.M. in an electric utility territory where peak energy and demand charges apply between 1 P.M. and 9:00 P.M. usually can economically apply a full CTES. Cooling during the 4 h period between 1:00 P.M. and 5:00 P.M. can be full-shifted, i.e., can be met with a relatively small and cost-effective CTES system and without oversizing the chiller equipment.

6.1.2 Partial-Storage CTES

In a partial-storage method, the chiller operates to meet part of the peak-period cooling load and the rest is met by drawing from storage. The chiller is sized at a smaller capacity than the design load. Partial-storage systems may operate as load-leveling or demand-limiting operations. In a load-leveling system (Fig. 5b), the chiller is sized to run at its full capacity for 24 h on the hottest days. The strategy is most effective when the peak-cooling load is much higher than the average load. In a demand-limiting system, the chiller runs at reduced capacity during

peak hours and is often controlled to limit the facility's peak demand charge (Fig. 5c). Demand savings and equipment costs are higher than they would be for a load-leveling system and lower than for a full-storage system.

Partial storage is more often the most economic option and, therefore, represents the majority of thermal storage installations. Although partial storage does not shift as much load (on a design day) as a full-storage system, partial-storage systems can have lower initial costs, particularly if the design incorporates smaller equipment by using low-temperature water and cold-air distribution systems.

6.2 CTES Storage Media Selection and Characteristics

The main CTES storage media include chilled water, ice, and eutectic salts. Ice systems offer the densest storage capacity but have the most complex charge and discharge equipment. Water systems offer the lowest storage density and are the least complex. Eutectic salts have intermediate characteristics. Some details on each storage medium follow:

6.2.1 Chilled Water

Chilled-water systems require the largest storage tanks, but can easily interface with existing chiller systems. Chilled-water CTES systems use the sensible heat capacity of water to store cooling capacity. They operate at temperature ranges (3.3–5.5°C) compatible with standard chiller systems and are most economical for systems greater than 2000 ton-h in capacity. The most important advantage of water thermal storage is the fact that water is a well-understood and familiar medium. No danger is associated with handling water as a TES material. Basically, heating and cooling applications require only a moderate temperature range (5–45°C). Water can be used as a storage medium for both cooling and heating, allowing the same TES material to be used throughout the year. The relatively high thermal capacity of water is also attractive for TES applications. These advantages give water CTES economic advantages over other air-conditioning systems, including those using ice CTES systems.

6.2.2 Ice

Ice CTES systems can be economically advantageous and require less space than water CTES systems. This space advantage often allows heating and cooling capacity to be enlarged within the often restricted area of existing machine rooms. Ice systems use smaller tanks and offer the potential for the use of low-temperature air systems, but require more complex chiller systems. Ice CTES systems use the latent heat of fusion of water (335 kJ/kg) to store cooling capacity. To store energy at the temperature of ice requires refrigeration equipment that provides charging fluids at temperatures below the normal operating range of conventional air-conditioning equipment. Special ice-making equipment or standard chillers modified for low-temperature service are used. The low chilled-water-supply temperatures available from ice storage allow the use of cool-air distribution, the benefits of which include the ability to use smaller fans and ducts and the introduction of less humid air into occupied spaces.

6.2.3 Eutectic Salts

Eutectic salts can use existing chillers but usually operate at warmer temperatures than ice or chilled-water systems. Eutectic salts use a combination of inorganic salts, water, and other elements to create a mixture that freezes at a desired temperature. The material is encapsulated in plastic containers that are stacked in a storage tank through which water is circulated. The most commonly used mixture for thermal storage freezes at 8.3°C, which allows the use of standard chilling equipment to charge storage, but leads to higher discharge temperatures. These temperatures, in turn, limit the operating strategies that may be applied. For example, eutectic salts may be used in full-storage operation only if dehumidification requirements are low.

7. ILLUSTRATIVE EXAMPLES

This section discusses some illustrative case studies to highlight the practical importance of TES systems and show how they are performing. Further information on these case studies and some other practical examples can be found elsewhere:

7.1 Alitalia's Headquarters Building, Rome, Italy

In June 1991, Alitalia, Italy's largest airline company, officially opened its new headquarters building in the southwest area of Rome. With a total plan area of 130,000 m^2, this building contains 55,000 m^2 of office space, 72,000 m^2 of service and technical rooms, and 26 conference rooms, surrounded by a

large parking area for more than 3000 cars. For the air-conditioning and computer-cooling needs of the entire complex, ice-chiller thermal storage coils were selected to meet a total storage capacity of 65,000 kWh. This makes Alitalia the largest ice storage installation in Europe and one of the largest in the world. Due to the magnitude of this project, the designers opted to install thermal storage coils in two concrete tanks, each 34 m long by 12 m wide. During the night, the central refrigeration system builds a complete charge of ice on the coils over a 12 h off-peak period. During the daytime, the ice is melted to provide the buildings' base cooling requirements, thereby minimizing the peak demand. The refrigeration system is used during the day to provide any extra required cooling capacity upstream of the ice thermal storage coils. This results in higher refrigeration system efficiency and lower operating costs. Unlike other types of thermal storage units, the ice chiller with consistent low-temperature supply can be located downstream from the chillers for highest overall system efficiency. In addition to the thermal storage coils, four large industrial fluid coolers and six heat exchangers were also provided for this European TES project. Detailed information can be found in the 1999 publication by Baltimore Aircoil.

7.2 A Slurry–Ice Rapid Cooling System, Boston, United States

Slurry–ice is a crystallized water-based ice solution that can be pumped and offers a secondary cooling medium for TES while remaining fluid enough to pump. It flows like conventional chilled water but provides five to six times the cooling capacity.

The installed system consists of a 88 kW remote condensing slurry–ice machine and an associated 10 m^3 ice storage tank to satisfy a peak load of 180 kW (Fig. 6). The ice machine is designed to operate until the tank is full or during rapid cooling periods in order to introduce ice into the cooling circuit. The stored energy over off-peak/low-load periods is later utilized to satisfy the short and sharp peak loads. Hence, the installed refrigeration machinery is one-third of the equivalent capacity of a conventional direct cooling system. Harvested fresh vegetables are subjected to rapid cooling within the cold storage facility, whereby the energy stored during off-peak periods is recovered by circulating solution within the air-spray heat exchanger in order to provide 0–1°C air off temperatures during the rapid cooling periods.

7.2.1 Technical Benefits

- Cost-effective installation: Smaller pipework and a flexible storage tank coupled with smaller pump sizes result in lower initial installation cost.
- Reduced running cost: Reduced refrigeration machinery results in reduced maximum demand and availability charges, and coupled with nighttime low ambient temperatures and off-peak electricity prices, offers unmatched overall running cost savings.

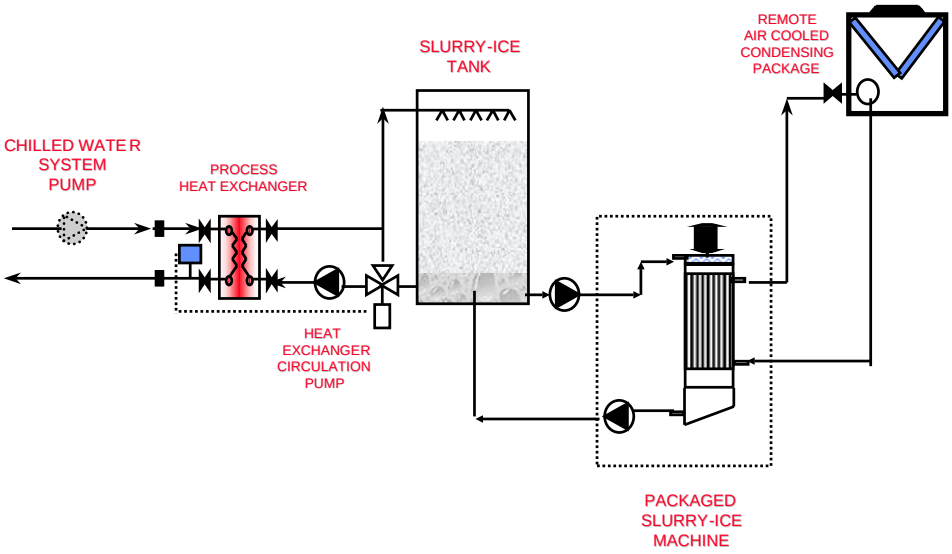


FIGURE 6 A typical slurry–ice TES system. Reprinted from Environmental Process Systems (2003), with permission.

- Quick response: Fine ice crystals offer unmatched thermal efficiency. Hence, large peak loads can be handled without affecting the system-leaving temperatures.
- Flexible system: Any future capacity or change in operational patterns can be easily handled without the need for additional refrigeration machinery.
- Full stand-by capability: The stored energy can be used to operate the process in case of breakdown or regular maintenance shutdowns.
- Green solution: Reduced refrigeration machinery leads to reduced refrigerant volume.

Further information can be obtained from the 2003 publication by Environmental Process Systems.

7.3 Harp Brewery, Dundalk, Ireland

Harp Brewery in Dundalk, Ireland (Fig. 7) produces 1,000,000 hl of lager per year and consumes 12,000,000 kWh of electricity (4,000,000 kWh for the refrigerant plant). In 1992, Harp decided to modernize its refrigeration plant. The existing system consisted of a wide range of process-cooling loads satisfied by two circuits, a high-temperature circuit at $-8^{\circ}C$ and a low-temperature circuit at $-12^{\circ}C$. During the design stage, it was decided to install a Cristopia TES System (STL) on the high-temperature circuit. The STL reduces the instantaneous peak refrigeration demand by as much as 45%, by daily shifting 10,000 kWh from day to night. The STL has reduced the maximum demand by 1 MW and the cooling consumption by 16%.

Technical data

Daily cooling energy consumption	44,000 kWh
Maximum cooling demand	4000 kW
Cooling energy stored	10,000 kWh
Storage temperature	$-10.4^{\circ}C$
Storage type	STL-N10-200 (see *Fig. 7*)
Nodules type	SN. 10
Number of tanks	2
Tank diameter	3 m
Tank height	14 m

7.3.1 Characteristics

To improve system efficiency and to reduce electrical cost, it was proposed to install a Cristopia STL-N10-200 TES system to shift electrical consumption to off-peak periods and to reduce the peak-period electrical demand. To achieve this end, a STL-N10-200 was specified, storing 10,000 kWh at $-10^{\circ}C$ during the nighttime off-peak tariff period. The STL also reduces daytime chiller power by 1000 kWh. The STL-N10-200 is composed of two 100 m^3 vertical tanks filled with SN.10 (77 mm diameter) nodules. The nodules are filled with a PCM, allowing thermal storage at $-10^{\circ}C$.

7.3.2 Technical Advantages

- Smaller chiller capacity;
- Stable flow temperature;
- Electrical supply reduced by 1000 kWh;
- Refrigerant efficiency improved;
- Servicing simplified;
- Back-up at disposal; and
- Smaller cooling towers.

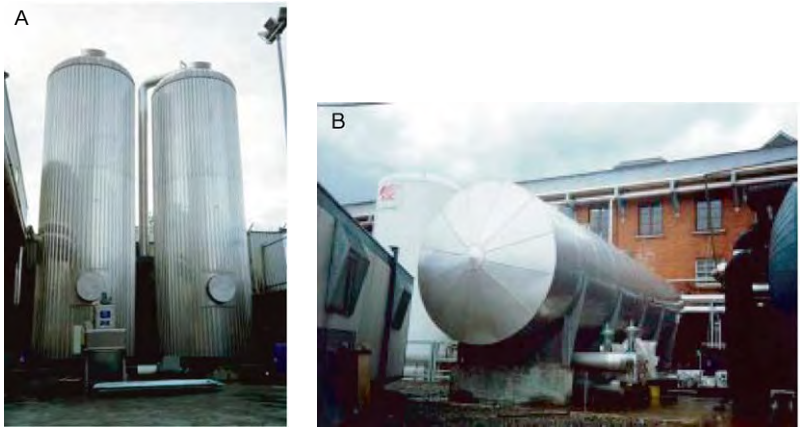


FIGURE 7 (A) Two STL-N10-100 systems and (B) Harp Brewery in Dundalk, Ireland. Reprinted from Cristopia (2001a), with permission.

7.3.3 Financial Advantages

- The system with STL costs less than a traditional system with chillers;
- Savings on maintenance costs;
- Electrical supply reduced by 1000 kWe; and
- Savings on demand charge and energy costs.

7.3.4 Environmental Advantages

- Refrigerant charge reduced (leading to ozone layer protection); and
- Emission of CO_2, SO_2, and N_2O reduced (avoiding contributions to the greenhouse effect and other environmental impacts).

Further information on this project can be obtained from the 2001a publication by Cristopia Energy Systems.

7.4 Museum of Sciences and Industry, La Villette, France

At the site of the old Paris slaughterhouse in La Villette Park, France, a modern leisure center was built in 1983 that included (1) a sciences and industry museum, (2) an audiovisual center, and (3) the GEODE 3-D cinema. The center (Fig. 8A) is one of the largest science and technology centers in the world and one of the most innovative (150,000 m^2 of displays, shows, and exhibitions). A cooling and heating plant supplies air-conditioning to a distribution network. This plant includes an STL-00-550 chiller (storing 31 MWh) and three 2700 kW chillers.

Technical data

• Main alternate current system type	Air-handling unit
• Building volume	1,350,000 m
• Daily cooling energy consumption	163,000 kWh
• Maximum cooling demand	12,500 kW
• Chiller capacity	
—Direct mode	8100 kW (6/12°C)
—Charge mode	4400 kW(−6/−2°C)
• Energy stored	31,750 kWh
• Volume	549 m^3
• Storage type	STL-00-550 (see *Fig. 8B*)
• Number of tanks	3
• Tank diameter	4 m
• Tank height	18 m

7.4.1 Characteristics

The STL is used to reduce 50% of the peak cooling demand. The storage is charged between 10:00 P.M. and 7:00 A.M. when electricity rates are low. Since 1985, significant operating cost savings have been achieved.

7.4.2 Technical Advantages

- Smaller chiller capacity;
- Smaller heat rejection plant;
- Reduced maintenance;
- System efficiency and reliability increased; and
- Increased plant lifetime expectancy.

7.4.3 Environmental Advantages

- Refrigerant charge reduced; and
- Emission of CO_2, SO_2, and N_2O reduced.

FIGURE 8 (A) The modern leisure center in La Villette Park, France, and (B) the TES tanks. Reprinted from Cristopia (2001b), with permission.

Further information on this project can be obtained from the 2001b publication by Cristopia Energy Systems.

8. CONCLUDING REMARKS

TES is considered an advanced energy technology. The use of TES systems has been attracting increasing interest in several thermal applications, e.g., active and passive solar heating, water heating, cooling, and air–conditioning. TES often is the most economical storage technology for building heating, cooling, and air–conditioning applications.

In general, TES can reduce the time or rate mismatch between energy supply and energy demand, thereby playing a vital role in improved energy management. TES use can lead to savings of premium fuels and make a system more cost-effective by reducing waste energy. TES can improve the performance of thermal systems by smoothing loads and increasing reliability. Therefore, TES systems are becoming increasingly important in many utility systems.

The selection of TES systems depends mainly on the storage period required (e.g., diurnal or seasonal), economic viability, operating conditions, etc. Some specific parameters that influence the viability of a TES system include facility thermal loads, thermal and electrical load profiles, availability of waste or excess thermal energy, electrical costs and rate structures, type of thermal generating equipment, and building type and occupancy.

The economic justification for TES systems usually requires that annual capital and operating costs be less than the costs for primary generating equipment supplying the same service loads and periods. Substantial energy savings can be realized by taking advantage of TES to facilitate using waste energy and surplus heat, reducing electrical demand charges, and avoiding heating, cooling, or air-conditioning equipment purchases.

TES can help correct the mismatch between supply and demand of energy and can contribute significantly to meeting society's needs for more efficient, environmentally benign energy use. TES plays an important role in energy conservation and can yield savings of premium fuels.

TES has enormous potential to make the use of thermal energy equipment more effective and to facilitate large–scale energy substitutions economically. A coordinated set of actions is needed in several sectors of energy systems in order to realize the maximum benefits of storage.

SEE ALSO THE FOLLOWING ARTICLES

Cogeneration • Geothermal Direct Use • Geothermal Power Generation • Ground-Source Heat Pumps • Heat Transfer • Refrigeration and Air-Conditioning • Solar Ponds • Solar Thermal Energy, Industrial Heat Applications • Solar Thermal Power Generation • Storage of Energy, Overview • Temperature and Its Measurement • Thermal Comfort

Further Reading

Abhat, A. (1989). Low temperature latent heat thermal energy storage: Heat storage materials. *Solar Energy* **30,** 313–332.

Baltimore Aircoil (1999). "Ice Keeps Alitalia Cool, Application Leaflet No." MN-92-10. Baltimore Aircoil, Baltimore, MD.

Beggs, C. (1991). The economics of ice thermal storage. *Building Res. Inform.* **19,** 342–355.

California Energy Commission. (1996). "Source Energy and Environmental Impacts of Thermal Energy Storage, Technical Report No. P500-95-005." California Energy Commission, Sacramento, CA.

Cristopia Energy Systems. (2001a). "Harp Guinness Subsidiary (Ireland)." Cristopia Energy Systems, Vence, France. Available at http://www.cristopia.com.

Cristopia Energy Systems. (2001b). "Museum of Sciences and Industry: 'La Villette' (France)." Cristopia Energy Systems, France. Available at http://www.cristopia.com.

Dincer, I. (1999). Evaluation and selection of thermal energy storage systems for solar thermal applications. *Int. J. Energy Res.* **23,** 1017–1028.

Dincer, I., and Rosen, M. A. (2002). "Thermal Energy Storage Systems and Applications." Wiley, London.

Dincer, I., Dost, S., and Li, X. (1997a). Performance analyses of sensible heat storage systems for thermal applications. *Int. J. Energy Res.* **21,** 1157–1171.

Dincer, I., Dost, S., and Li, X. (1997b). Thermal energy storage applications from an energy saving perspective. *Int. J. Global Energy Issues* **9,** 351–364.

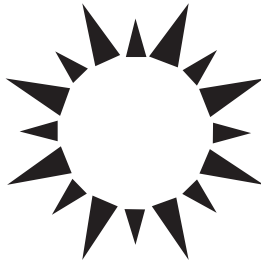
Environmental Process Systems. (2003). "Thermal Energy Storage Systems." Environmental Process Systems, Berkshire, UK. Available at http://www.epsltd.co.uk.

Norton, B. (1992). "Solar Energy Thermal Technology." Springer-Verlag, London.

Reindl, D. T. (1994). "Characterizing the Marginal Basis Source Energy Emissions Associated with Comfort Cooling Systems," Report No. TSARC 94-1. Thermal Storage Applications Research Center, Madison, Wisconsin.

Thermal Pollution

VICTOR S. KENNEDY
University of Maryland Center for Environmental Science
Cambridge, Maryland, United States



1. What is Thermal Pollution?
2. Thermal Discharges from Steam-Electric Power Plants
3. Temperature as an Environmental Factor
4. Ocean Thermal Energy Conversion
5. Cold Water Pollution

Glossary

acclimation temperature The temperature that an organism has become used to by a period of constant exposure over time.

biocide A substance like an herbicide or insecticide that kills living organisms. Chlorine and ozone have been used as biocides to prevent fouling of water pipes in power plants.

electricity units Gigawatt = 1 billion watts; megawatt = 1000 kilowatts or 1 million watts; megawatt-hour = a unit of energy equivalent to 1 megawatt of power expended for 1 hour of time.

larvae The early life history stages of aquatic animals such as fish or shellfish. Many types float in the water column as zooplankton (see Plankton).

plankton Animals and plants (usually microscopic in size) that float in aquatic systems and form part of the system's food web. They can be categorized as bacterioplankton (bacteria or microbes), phytoplankton (plants), and zooplankton (animals).

thermal efficiency Useful energy output divided by total energy input; for a steam-electric power plant, it is the proportion of the total energy that the facility consumes in relation to the amount of electric energy produced.

For the purpose of this article, thermal pollution refers to a deleterious change (increase or decrease) in the temperature of a water body as a result of human activities. However, some have declared that the term is biased in that it imputes negative effects to all thermal changes caused by humans. They have suggested that, in relation to increased temperature, more neutral terms might be thermal addition or calefaction. Where heated water provides a benefit, such as in enhancing the aquaculture of plants and animals, some have referred to that as thermal enrichment.

1. WHAT IS THERMAL POLLUTION?

Temperature changes can range from small to large and can cause negative or positive effects on organisms. Temperature in a water body can increase as the result of humans clearing vegetation from stream banks, thereby removing natural shading of the stream; of runoff of warmed water from heated land surfaces; and of disposal of heated effluents from industrial activity. Temperature in a water body can decrease as a result of inflow of cooler water from dams or of industrial activity. This article focuses on temperature increases, primarily as a byproduct of electricity generation by steam-electric power plants, but the effects of less common temperature decreases will be considered briefly.

Studies of the effects of thermal discharges from steam-electric power plants began in the 1950s in the United States and the United Kingdom, accelerating in number for about two decades thereafter. Additional studies were undertaken in Europe and the USSR from the 1960s onward. The number of studies worldwide began declining in the 1980s.

This article depends mostly on the results of those studies. It begins with a discussion of thermal discharges from steam-electric power plants, then examines temperature as a controlling factor in nature, and follows with descriptions of efforts over the years to estimate its influence on organisms in relation to heated discharges. It concludes with a brief examination of the less common problem of cold-water discharge from dams.

 © 2004 Elsevier Inc. All rights reserved.

2. THERMAL DISCHARGES FROM STEAM-ELECTRIC POWER PLANTS

2.1 Sources of Heating of Water Bodies

The areal extent of heating of water bodies caused by removal of vegetation from stream banks or by runoff from warm surfaces is minimal compared to the extent of heating caused by industrial activity. The largest user of water for industrial cooling is the steam-electric generating industry, so thermal discharges (effluents) from that industry will be the major focus of this article.

Steam-electric power plants burn fossil fuels (coal, petroleum, gas), combustible wastes, or biomass, or depend on nuclear fission to heat water in pressurized boilers to produce steam that is used to spin turbines that power electricity generators. Large quantities of additional water (coolant) are pumped from a water body through a multitude of small-diameter condenser tubes that function to condense the steam to water after it has passed through the turbines, whereupon it is recirculated into the boilers. The coolant gains heat in the process. If a power plant has a thermal efficiency of about 35% (nuclear power plants are less efficient than non-nuclear power plants), then about 65% of the heat produced by the power plant has to be disposed of. The heat in the coolant is reduced by releasing the coolant into the atmosphere or by returning it to a water body, usually the one from which it was originally extracted. The differences between the temperature of the cooling water pumped into the steam-electric power plant and its temperature as it exits the station (ΔT) can cover a wide range of values from about 5°C to 25°C.

Steam-electric power plants in industrialized countries have increased in size from an average of about 35 megawatts in the 1950s to as much as 1000 megawatts for fossil-fuel-powered plants and 1400 megawatts for nuclear-powered plants in recent decades. This has increased the demand for water for steam and for cooling activities. For example, a 1000 megawatt nuclear power plant with an open-cycle cooling system (discussed later) may use up to 5 billion liters of coolant per day. Toward the end of the 20th century, thermal power plants in the United States and the United Kingdom used about 80 to 90% of all the water used in those countries for industrial cooling. Subsequent efforts have been made to conserve and reuse cooling water, but the amount of water in industrial countries that is heated during steam-electric generation and that must be cooled remains large because of the overall increase in the number of power plants.

2.2 Cooling Systems

A variety of systems are used to reduce the heat of the coolant. Open-cycle or once-through cooling systems pump water from the water body they are located beside and can require tens of thousands of liters per megawatt-hour, releasing the heated effluent back into the water body as a surface discharge so that the heat can transfer to the atmosphere. Some heated effluents may be pumped directly into the water body (receiving water), whereas others may be released into a discharge or cooling canal to allow for some cooling to take place before the effluent flows back into the water body. This is perhaps the simplest design, but the amount of heated effluent involved means that large power plants (500 megawatts and above) must be located on very large lakes or on an estuarine or marine coast so the heated effluent has minimal effect on the receiving water as a whole. The entry of the discharge into the receiving water creates a mixing zone wherein the plume of heated water is cooled by the colder receiving water. Local hydrography influences the behavior of the plume (e.g., movement of the plume by winds or currents) and the speed of cooling.

Closed-cycle or recirculating cooling systems may incorporate ponds in which heated discharges cool sufficiently before being reused. Or they may use vertical cooling towers in which the effluent either falls from a height so that evaporative cooling occurs in air (wet cooling) or passes through a radiator in which cooling occurs with no direct contact of the effluent with air (dry cooling; this reduces evaporative loss of water). Closed-cycle systems use hundreds of liters of cooling water per megawatt-hour, and thus can be located on smaller water bodies than is the case for open-cycle cooling systems.

2.3 Growth in Electricity Use

The World Energy Council has reported that electricity represented 25% of energy use by industrial nations in the 1970s, growing to 37% by 2000. The council estimates that electricity will eventually grow to be 50% of such energy use. The Electric Power Research Institute estimates that world population growth and economic development will require an additional 10,000 gigawatts of electricity generating capacity over the next 50 years. This will

increase the amount of water that is heated to form steam to drive turbines, the amount of coolant needed for condensing the steam, and the amount of coolant that is returned to the environment.

3. TEMPERATURE AS AN ENVIRONMENTAL FACTOR

Temperature has been called an environmental master factor because of its importance in the biology of every organism. Most aquatic organisms are adapted to live within a specific temperature range, outside of which temperatures become stressful and ultimately lethal. Within the temperature range (i.e., at sublethal levels), temperature can affect an organism's metabolism, growth, and behavior, as well as reproductive timing and rates of egg and larval development. It also influences geographic distributions of species, including distributions of freshwater and marine fauna.

Temperature effects can be influenced by many other environmental factors (Fig. 1). For example, as will be demonstrated below, research has shown for some aquatic animals that age, size, life-cycle stage (egg, larva, juvenile, adult), or ambient oxygen conditions can affect the animal's temperature tolerances. Other studies have shown that diet, environmental salinity, or the presence of parasites can influence some animals' temperature tolerances. The interaction of temperature and dissolved oxygen can be very important, in that, whereas increased temperature raises the metabolism of "cold-blooded" aquatic animals, it also lowers the oxygen carrying capacity of water. Thus, at a time when the metabolic need for oxygen is increasing, the amount of available oxygen is decreasing, often causing stress or mortality to organisms in warming water (see the striped bass example described later).

Predicting the effects of temperature change on a particular species requires that the species' biology in relation to temperature, including the other factors that influence the temperature effects (Fig. 1), be well understood. The complexity of these interactions poses a challenge because of our limited knowledge of this complexity with regard to most organisms. Especially, we know little about the smallest components of aquatic ecosystems—microbes, phytoplankton, and zooplankton—that play significant roles in energy flow and nutrient flux within ecosystems. This lack of information makes it very difficult to predict the effects of increased temperature on an ecosystem, even if we have some information on the effects of temperature on some species in the ecosystem.

Finally, predicting the effects of thermal effluents from power plants can be made more difficult because biocides and other contaminants may be included in the coolant water. For example, if a biocide such as chlorine is used to reduce settlement of organisms in and subsequent blockage of condenser tubes (fouling), it may stress organisms in the vicinity of the thermal plume and cause higher mortalities or other biological changes than might occur from temperature alone.

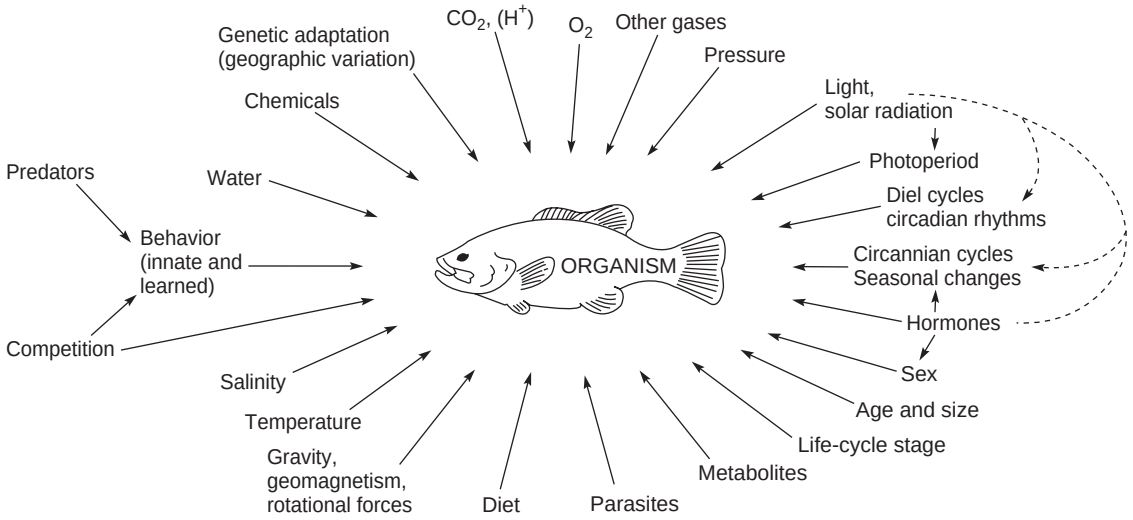


FIGURE 1 Various factors that can affect the thermal tolerances of organisms. From Hutchison (1976).

3.1 Effects of Temperature on the Biology of Organisms

3.1.1 Lethal Effects

Temperature extremes (both high and low) can be lethal. Many species are sensitive to temperatures just a few degrees higher than those they usually experience in nature. Indeed, in short-term (measured in days) temperature-tolerance experiments on estuarine invertebrates in the laboratory, a temperature increase of 1°C can raise mortalities from 0% or slightly above at the lower temperature to about 100% at the increased temperature. This is significant in terms of thermal pollution because temperature increases from heated-water discharges inevitably are greater than 1°C, as mentioned earlier.

An initial effort to predict the effects of heated water discharge on a species often involves laboratory experiments in which individuals of that species are placed in a series of water baths at high temperatures to determine the mortalities in each bath over a period of time (usually 24 to 72 hours). These relatively short-term experiments use individuals that are acclimated to the temperatures prevailing in their environment at the time of the experiments (e.g., in temperate regions, they will be acclimated to low temperatures in winter and high temperatures in summer). The results of the experiments are expressed in terms of LC50—that is, the temperature that will result in 50% of the individuals in a species dying within the time period of the test. As an example, Figure 2 presents data on the percentage of mortalities suffered by an estuarine invertebrate, the opossum shrimp *Neomysis americana,* individuals of which were acclimated in the laboratory to temperatures ranging from winter values of 1°C to summer values of 30°C before being placed in water baths holding water at the acclimation temperature or above. The sigmoid curves describing mortality over the test temperatures are steep, reflecting the slight differences between those temperatures at which mortality is limited and those that cause high mortality (see previous paragraph). In general, as acclimation temperatures increase, so do the LC50 values at each acclimation temperature. Thus, a temperature such as 18°C that would cause 100% mortality in winter obviously is not lethal to opossum shrimp that have become acclimated to summer temperatures of 25°C to 30°C. This generalization applies to other aquatic organisms exposed to temperature increases in similar laboratory experiments and means that, up to a point, heated-water discharge temperatures from steam-electric power plants may be more of a threat to survival of organisms in winter than the same temperatures might be in summer.

LC50 values can provide a reasonable estimate of the relative ability of different species to tolerate temperature increases. For example, seven species of invertebrates that are found together in Chesapeake Bay in the United States have LC50s that differ by about 10°C even though they were acclimated to the same temperature (15°C) (Fig. 3). Consequently, rules regarding temperature discharge need to account for the most thermally sensitive species, in this case the opossum shrimp.

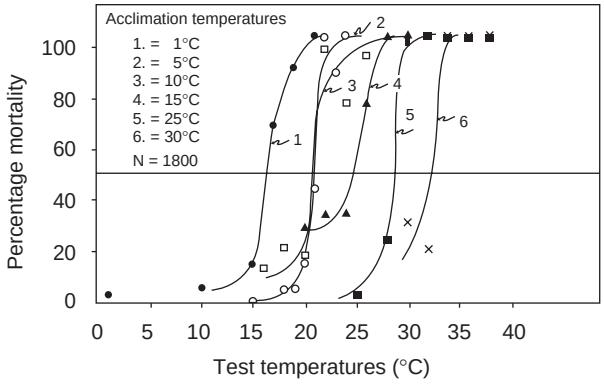


FIGURE 2 Curves of the mortalities of individuals of the estuarine opossum shrimp *Neomysis americana* acclimated to temperatures of 1°C to 30°C. The horizontal line represents the LC50 value for each acclimation temperature. Modified from Fig. 2 in J. A. Mihursky and V. S. Kennedy (1967), Water temperature criteria to protect aquatic life. *American Fisheries Society,* Special Publication No. **4**, 20–32, with permission.

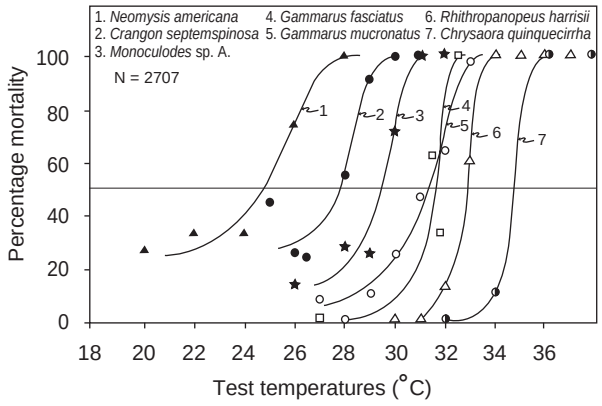


FIGURE 3 Comparative mortality curves for seven species of estuarine invertebrates from Chesapeake Bay acclimated to about 15°C. Modified from Fig. 3 in J. A. Mihursky and V. S. Kennedy (1967), Water temperature criteria to protect aquatic life. *American Fisheries Society,* Special Publication No. **4**, 20–32, with permission.

It is necessary to determine the tolerance of a species over its natural thermal range in order to set temperature discharge rules that will apply over the course of a year. This has been done for a number of clam species that live buried in the sediments on the bottom of Chesapeake Bay (Fig. 4). Here, one species, the gem clam *Gemma gemma,* is the most tolerant of temperature over its natural thermal range. Three other species that live in the same bottom habitat as *Gemma gemma* (the Baltic macoma *Macoma balthica*, the coot clam *Mulinia lateralis,* and the soft-shell clam *Mya arenaria*) have lower temperature tolerances. The youngest stages (newly settled and young *Mya arenaria* and small *Macoma balthica*) are slightly more tolerant than the adults.

These results make sense biologically. The youngest stages live on or near the sediment surface of the estuarine bottom where temperatures tend to be higher than they are deeper within the sediments where the larger animals dwell. Also, *Mya arenaria* in Chesapeake Bay is living near the southern extent of its distributional range on the western Atlantic coast, which may explain its greater thermal sensitivity compared to the other three species that range further south into warmer regions. Similarly,

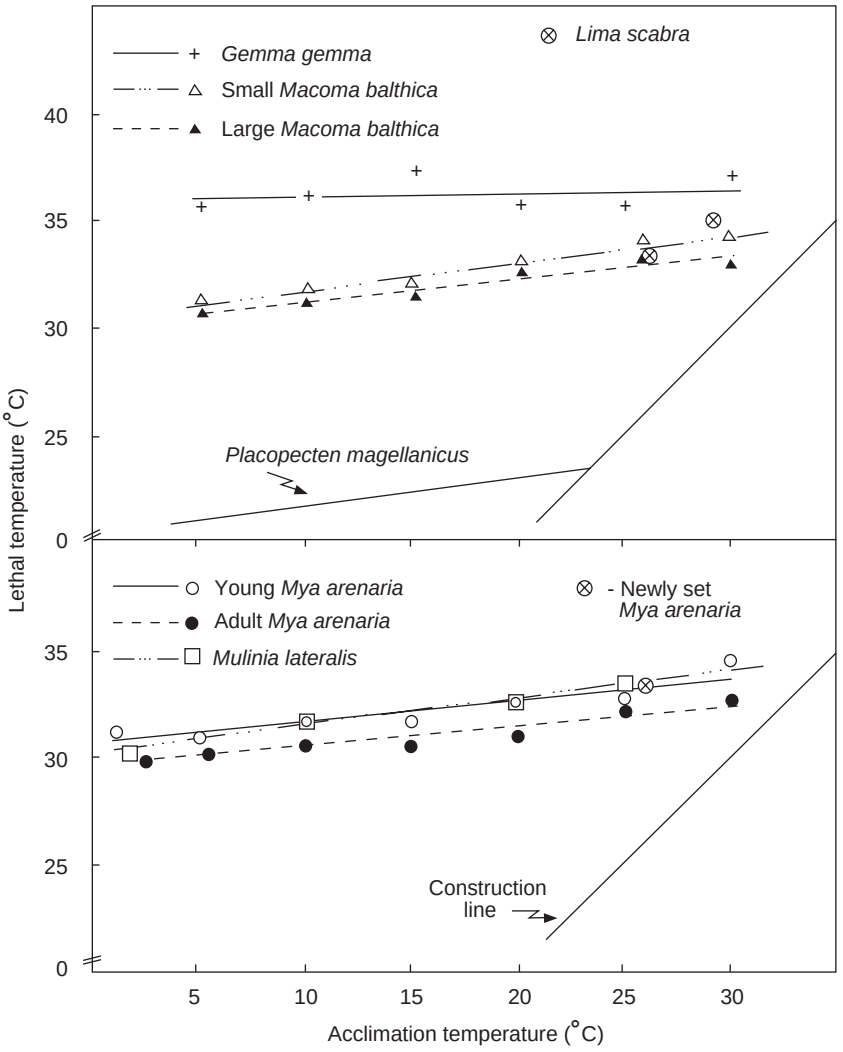


FIGURE 4 Lethal temperatures (LC50) for six species of bivalves, including newly settled, small or young, and large or adult individuals of some species. *Gemma gemma*, *Macoma balthica*, *Mulinia lateralis*, and *Mya arenaria* are clams that live buried in Chesapeake Bay sediment. *Lima scabra* lives in warm Atlantic waters, whereas *Placopecten magellanicus* lives in cold Atlantic waters. The construction line represents where the acclimation and lethal temperatures are identical. From Fig. 2 in V. S. Kennedy and J. A. Mihursky (1971), Upper temperature tolerances of some estuarine bivalves. *Chesapeake Sci.* **12**(4), 193–204. Permission granted by the Estuarine Research Federation.

the Atlantic Ocean scallop *Placopecten magellanicus* does not experience temperatures anywhere close to those experienced by the Chesapeake Bay species, which may explain its relatively low tolerances. On the other hand, the rough fileclam *Lima scabra* lives in warm Atlantic waters, which is reflected in its temperature tolerances.

Generalizing from these results, heated-water discharges into marine coastal waters where *Placopecten magellanicus* might be living would have to be much cooler than similar discharges into Chesapeake Bay or the coastal habitat of *Lima scabra*. Within Chesapeake Bay, temperatures that can be tolerated by *Gemma gemma* would be too high to be tolerated by the other clams, so the susceptibility of the less heat-tolerant *Mya arenaria* to higher temperatures should drive regulations governing thermal discharges (based on these experimental data alone).

Based on these and other research results, the state of Maryland promulgated regulations governing discharge temperatures from power plants. Figure 5 plots the allowed seasonal increase above ambient temperature of heated discharges in Chesapeake Bay (dashed lines) in relation to the thermal tolerances of the four estuarine species depicted in Fig. 4. *Gemma gemma* has a lethal thermal tolerance that is sufficiently higher than the regulated discharge temperatures so that there should be no or limited effect of the discharge on the survival of the clams even in summer (however, there is no information on how the heated-water discharges might affect growth or reproduction in this species; see the next paragraph). On the other hand, adult *Mya arenaria* (especially) and large *Macoma balthica* have lethal thermal tolerances that are very close to the upper temperature limit of allowable discharge in summer in Maryland.

Regulations elsewhere are also region specific as well as species specific. For example, in British Columbia, Canada, the maximum daily temperature of streams inhabited by temperature-sensitive bull trout *Salvelinus confluentus* or Dolly Varden *Salvelinus malma* are not allowed to exceed 15°C and the maximum temperature during spawning cannot exceed 10°C, whereas streams harboring species of Pacific salmon are not allowed to be warmed more

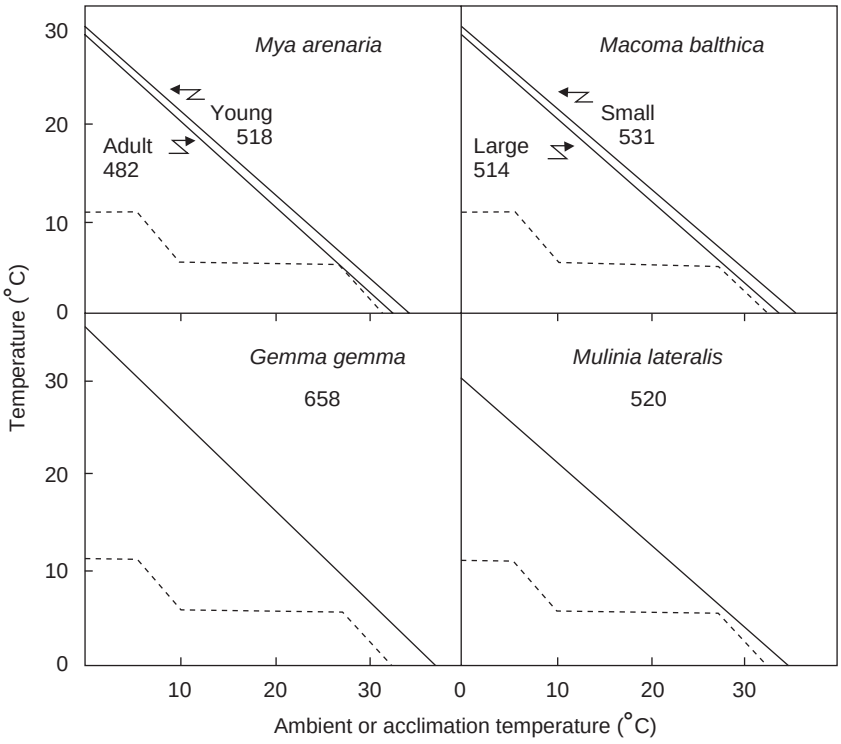


FIGURE 5 Temperature tolerance triangles of four species of estuarine bivalves from Chesapeake Bay, Maryland (solid lines developed by subtracting an LC50 value from its corresponding acclimation temperature), in relation to Maryland's thermal discharge regulations for tidal waters (dashed lines show how many degrees above ambient temperature a discharge can be). Numbers refer to the areas of the triangles in square degrees Centigrade. From Fig. 3 in V. S. Kennedy and J. A. Mihursky (1971), Upper temperature tolerances of some estuarine bivalves. *Chesapeake Sci.* **12**(4), 193–204. Permission granted by the Estuarine Research Federation.

than 1°C beyond the optimum temperature range for each life history phase of the most sensitive salmon species present.

Warm temperatures over coral reefs can cause coral bleaching, a breakdown in the symbiotic relationship between the coral animal and the unicellular algae that live within coral tissues. The algae (called zooxanthellae) photosynthetically produce food materials that the coral can use, and they are in turn sustained by the coral's waste products that serve as nutrients. Under warmer climatic conditions, coral animals have expelled the zooxanthellae, leaving the animals pale to colorless. The symbiotic relationship is thus suspended and the host is weakened until environmental conditions improve and normal zooxanthellae density is regained. Thermal discharges can have similar effects on coral reefs in their vicinity, so power plant discharges in the tropics are directed away from coral reefs.

Thermal discharges are not the only aspect of steam-electric generation that can have a lethal effect on organisms. The water that is pumped into the power station to serve as coolant harbors entrained microscopic organisms (bacterioplankton, phytoplankton, and zooplankton, including fish and invertebrate larvae in breeding seasons). This coolant experiences very rapid increases in temperature as it passes through the condensers. Such temperatures will usually be much higher than those experienced by organisms exposed to the thermal effluent from the power station, so mortalities are usually higher for entrained organisms (the threat from higher temperature is exacerbated by pressure changes or mechanical stress as the organisms are pumped through the condenser system, as well as by any biocides like chlorine that are used to keep fouling of the system's tubes to a minimum).

Experiments that have been performed to focus on the relationship between temperature increase and residence time within a condenser system have shown, unsurprisingly, that mortalities of entrained organisms can increase depending not only on the increase in temperature but also on the length of exposure to increased temperature. Such experiments have been performed on the fertilized eggs and larvae of a number of estuarine bivalves. Figure 6 shows the results on such an experiment on an early larval stage of the estuarine hard clam *Mercenaria mercenaria*. The larvae were subjected to a variety of temperatures from 18°C to 43°C for periods of time that ranged from 1 to 360 minutes. As temperature increased from about 36°C, so did mortalities. In addition, after about 10 minutes, the longer the

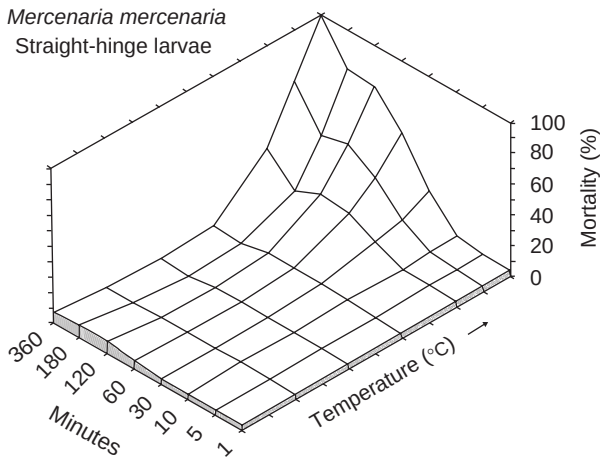


FIGURE 6 Temperature tolerances over time for the straight-hinge larval stage of the estuarine hard clam *Mercenaria mercenaria*. From Kennedy *et al*. (1974).

larvae were exposed to temperatures above about 38°C, the higher their mortality. Although differing in detail, these relationships hold for fertilized eggs and other larval stages of the hard clam, as well as for fertilized eggs and larvae of other species of estuarine bivalves that have been studied.

Based on these and other findings, the lesson for regulators is that passage of coolant through condenser systems should be as rapid as possible if mortalities of entrained organisms are to be minimized. A variety of estimators have been developed to enable regulators to predict the actual exposure times that can be allowed in order to minimize mortality.

3.1.2 Sublethal Effects

Research into thermal tolerances has determined that lethal temperatures usually are higher (or lower in the case of cold conditions) than are temperatures that influence activity and growth, which in turn are higher (or lower) than temperatures that affect reproduction. This information has been used to build a schematic representation of the effects of high and low temperatures on Pacific salmon (Fig. 7). In terms of survival, the egg and hatchling stages are less tolerant of warm temperatures than are older stages. In turn, thermal tolerances vary with season as explained above for the opossum shrimp, with tolerances to warm temperatures being higher in summer than winter (thus the undulation in the upper line on the figure). Warm temperatures that inhibit growth and reproduction are, in turn, lower than those that cause mortality. A similar scenario can be developed for freshwater

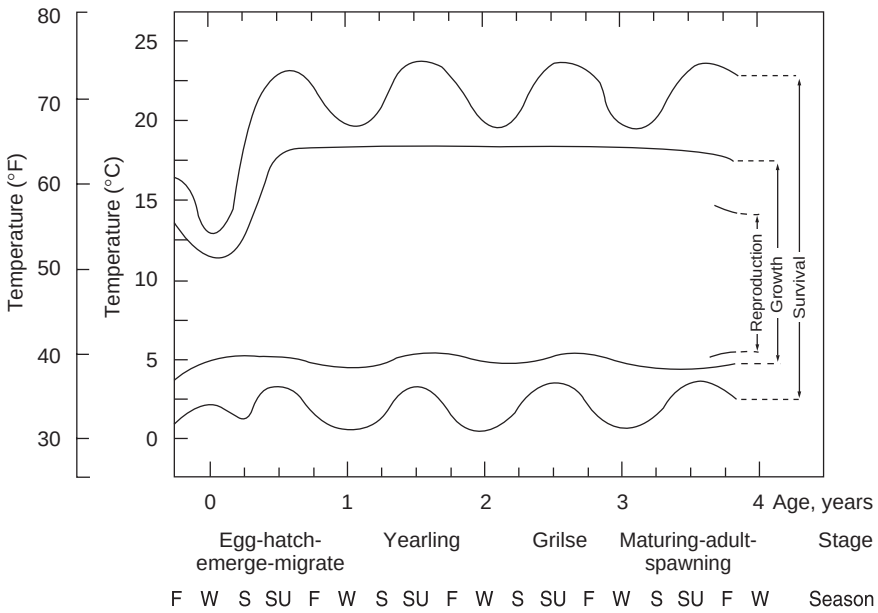


FIGURE 7 Temperature tolerances (measured and hypothetical) that affect different life cycle stages of Pacific salmon over the seasons during 4 years of life. Lethal temperatures (survival) are higher or lower than are the temperatures that inhibit growth, which in turn are higher or lower than the temperatures that inhibit reproduction (spawning). From Brett (1960).

carp and brown trout (Fig. 8) in which the temperature ranges that affect optimal growth are narrower than those influencing feeding and different from those affecting egg development, with the ranges that cause stress extending beyond these other ranges. In terms of differences between species, carp are more tolerant of higher temperatures than are brown trout.

Laboratory studies have shown that organisms under stress pay a metabolic cost in the form of a continued expenditure of energy. These findings mean that a thermal discharge that is sublethal may nevertheless have deleterious effects on growth and reproduction. Stunted animals may be more susceptible to predation and may be delayed in maturing and reproducing. And if a species cannot reproduce, eventually its population might disappear from the heated locale unless other individuals of that species move into the locale when temperatures are not deleterious. Clearly, these considerations are more important for immobile organisms (e.g., fish eggs laid among gravel in streams, oysters, barnacles, underwater grasses) that are unable to leave the vicinity of the heated discharge than they are for mobile organisms (e.g., fish, some invertebrates) that can avoid the deleterious conditions. Thus, regulations governing thermal discharges may need to be focused more on organisms that cannot avoid the discharge than on those that can.

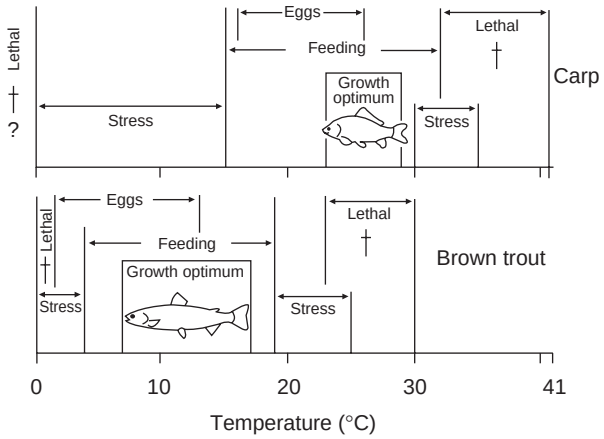


FIGURE 8 Effects of temperature on egg development, feeding, growth, and survival of freshwater carp *Cyprinus carpio* and brown trout *Salmo trutta*. From J. M. Elliott (1980), Some aspects of thermal stress in freshwater teleosts. *In* "Stress and Fish" (A. D. Pickering, Ed.), pp. 209–245. Academic Press, New York, with permission.

3.1.3 Beneficial Effects

Heated-water discharges may be a beneficial resource in human endeavors, and the high temperatures of power-station discharges can be ameliorated by diverting the discharge through a system in which the heat can be put to productive use. For example, such discharges have been used to heat buildings or to warm greenhouses and other facilities in which

commercial plants and animals are cultured. However, there is no guarantee that the discharges will not be shut off for unexpected reasons such as unscheduled maintenance or decreased electricity demand, so these eventualities have to be considered by users of the thermal resource.

3.2 Effects of Temperature on Abundances and Distributions of Organisms

Temperature changes as small as 1°C can have important effects on the abundance and distribution of species, as happened in Monterey Bay, California. There, local temperatures increased by about that amount between the early 1930s and the late 1990s as a result of a climate-warming event (not associated with industrial discharges). The abundances of intertidal species with southern (warmer water) distributions became higher than in the early 1930s and the abundances of northern (cooler water) species declined. Thus, although higher temperatures might not result in extinction of a species throughout its geographic range, the species may be eliminated from part of its range. However, with regard to industrial heated discharges, such discharges are of limited regional scope and are under governmental regulation. Thus, it would be unusual for thermal discharges to affect the distributions of organisms beyond the vicinity of the thermal discharge itself.

Mobile organisms can respond to temperature by avoiding it or becoming attracted to it. This will vary seasonally, depending on the organism's thermal tolerance or its thermal preference. In winter it is common to find mobile organisms like some invertebrates, fish, turtles, and mammals like manatees congregating in the discharge region of a power station if there is enough food available to meet their metabolic needs. Waterfowl may be able to take advantage of ice-free conditions caused by discharges to feed in aquatic systems that ordinarily would be ice covered and inaccessible. However, if the discharge is halted abruptly by maintenance needs of the utility or for some other reason, "cold-blooded" animals like invertebrates, fish, and turtles may be at risk. The water body will soon revert to ambient winter temperatures, which may lead to cold shock and mortality of those organisms that have become adapted to higher temperatures while living in the thermal effluent and thus more susceptible to cold than they would have been if they had been acclimated naturally to winter temperatures.

3.3 Effects of Temperature on Species Interactions

Temperatures do not have to be lethal to influence changes in species interactions (e.g., predator-prey or parasite-host relationships, or competition among organisms for resources) in ecosystems. Changes in local distribution patterns as noted earlier would change the mix of predators, prey, parasites, and competitors in a thermally influenced ecosystem in ways that are not yet understood but that could alter the functions of the ecosystem and the productivity of resident organisms. However, in relation to thermal pollution, these alterations will be limited to the vicinity of the thermal discharge and will have no regional effect outside the mixing zone.

3.4 Effects of Temperature on Oxygen Concentrations

Oxygen-dependent (aerobic) aquatic organisms extract the oxygen necessary for their metabolic demands from the water in which they live. Temperature influences the amount of oxygen that water can hold, with warmer water holding less oxygen than does cooler water. Thus, as warmer temperatures cause an increase in the metabolism of aerobic organisms, there is less oxygen available to support their metabolism. This can be lethal, or can stress an organism, perhaps rendering it unable to feed or to escape being fed upon by a predator more tolerant of low-oxygen conditions, or making it more susceptible to disease.

Interactions between higher temperatures and depleted oxygen could constrict the available habitat for certain aquatic species. For example, the habitat of striped bass may be "squeezed" in the vicinity of a thermal discharge. In late spring through autumn in western Atlantic coastal and estuarine waters unaffected by humans, the upper layers that are the habitat of juvenile striped bass are usually warmer than the deeper waters inhabited by older fish (Fig. 9). The squeeze can occur if a thermal discharge warms the surface waters so that the depth of warm water increases, leaving less cool water for the older stages. The warmer conditions will also speed up the decay of organisms sinking into the depths, thereby increasing the amount of oxygen-depleted (anoxic) bottom waters that the fish will avoid. Thus, the suitable habitat for the older stages will be squeezed between the increased volume of warmer upper waters and the anoxic bottom waters. However, striped bass are mobile and can swim to cooler

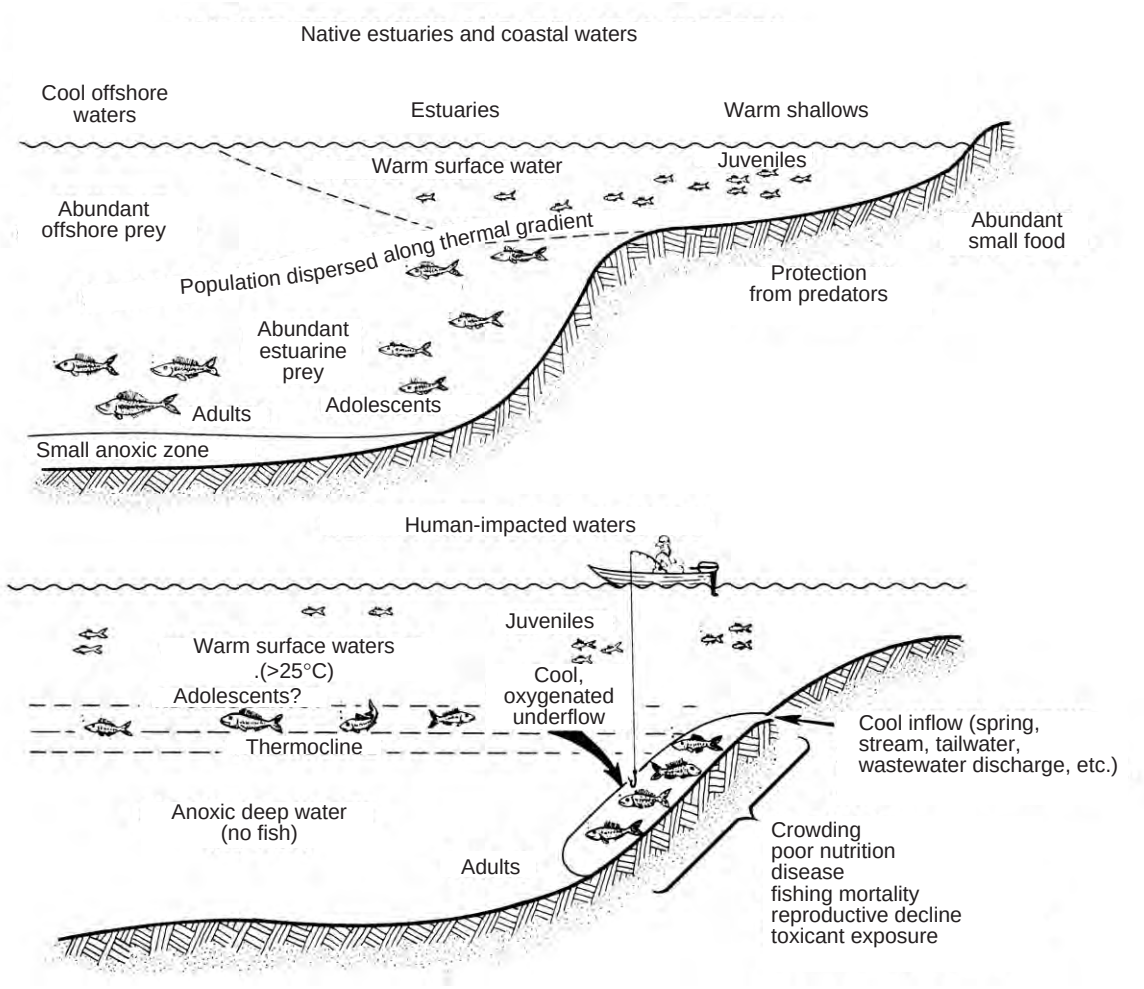


FIGURE 9 Distributions of striped bass in relation to normal (upper panel) and warmed temperature conditions. From C. C. Coutant (1987), Thermal preference: When does an asset become a liability? *Environmental Biology of Fishes* **18**, 161–172, with kind permission of Kluwer Academic Publishers.

habitat elsewhere along the coast, so it is unlikely that a thermal discharge will affect a biologically significant amount of habitat.

4. OCEAN THERMAL ENERGY CONVERSION

An alternate system of generating electricity is the Ocean Thermal Energy Conversion (OTEC) system, which has been employed in tropical ocean waters. It takes advantage of the difference in temperatures between solar-heated surface water (usually at 24°C to 30°C) and cooler, subsurface water at up to 1000 m depth (4°C to 8°C). A closed-cycle OTEC system uses heat from the warm seawater to vaporize a working fluid (e.g., ammonia), with the expanding vapor spinning a turbine that powers a generator. Cold seawater is pumped from depth through a condenser that condenses the working fluid vapor back into a liquid that is then recycled through the system. An open-cycle OTEC system vaporizes warm surface water itself (the working fluid) in a near vacuum at the temperature of the surface water to spin a low-pressure turbine. Again, cold deep water condenses the vapor to water (freshwater, as the seawater has lost its salt upon vaporization). In the process of cooling the vapor, the cold seawater is warmed to a temperature intermediate between its original temperature and that of the surface. Thus, the heated effluent is returned to the ocean at a depth (e.g., 60 m) to reduce temperature anomalies that would have adverse effects on organisms such as described earlier.

5. COLD WATER POLLUTION

An uncommon form of thermal pollution involves the release of cool or cold water from storage reservoirs into warmer water, thereby lowering the temperature of the receiving water bodies. Australia is one country in which this occurs. There, water is released from dams for irrigation in summer and autumn when river temperatures are high. If the water is released from a low point in the dam wall, then the coolest water in the dam is what is released (in summer, the dam's surface water is warmed by the sun and the lighter warm water floats on the denser cooler water). This water can be more than 10°C cooler than the water in the receiving water body. The deleterious effects of cold-water pollution can be detected in Australian rivers from 50 to 400 km from the release point.

The results of cold-water pollution can be similar to those of warm-water pollution in that organisms may die, or their physiology may be altered, and their communities may be disrupted. Temperatures may be so cold as to be lethal for a species, and sublethal effects include the lowering of body temperatures, thus slowing growth and retarding reproduction (see Figs. 7 and 8), as well as influencing the ability to avoid more cold-tolerant predators. Cold-water pollution can be prevented by adding warmer tempering water to the cold water as it is being released or by taking the water for release from near the surface of the dam where it is warmer.

SEE ALSO THE FOLLOWING ARTICLES

Air Pollution from Energy Production and Use • Air Pollution, Health Effects of • Ecosystems and Energy: History and Overview • Electricity, Environmental Impacts of • Entrainment and Impingement of Organisms in Power Plant Cooling • Ocean, Energy Flows in • Ocean Thermal Energy

Further Reading

Brett, J. R. (1960). Thermal requirements of fish–three decades of study, 1940–1970. *In* "Biological Problems in Water Pollution" (C. M. Tarzwell, Ed.), pp. 110–117. U.S. Department of Health, Education, and Welfare. Robert A. Taft Sanitary Engineering Center, Cincinnati, OH.

Coutant, C. C. (1987). Thermal preference: when does an asset become a liability? *Environ. Biol. Fishes* **18,** 161–172.

Elliott, J. M. (1980). Some aspects of thermal stress in freshwater teleosts. *In* "Stress and Fish" (A. D. Pickering, Ed.), pp. 209–245. Academic Press, New York.

Godfriaux, B. L., Eble, A. F., Farmanfarmaian, A., Guerra, C. R., and Stephens, C. A. (eds.). (1979). "Power Plant Waste Heat Utilization in Aquaculture." Allanheld, Osmun & Co., Montclair, NJ.

Hutchison, V. H. (1976). Factors influencing thermal tolerances of individual organisms. *In* "Thermal Ecology II" (G. W. Esch and R. W. McFarlane, Eds.), pp. 10–26. United States Energy Research and Development Administration, Springfield, VA.

Kennedy, V. S., and Mihursky, J. A. (1971). Upper temperature tolerances of some estuarine bivalves. *Chesapeake Sci.* **12,** 193–204.

Kennedy, V. S., Roosenburg, W. H., Castagna, M., and Mihursky, J. A. (1974). *Mercenaria mercenaria* (Mollusca: Bivalvia): Temperature-time relationships for survival of embryos and larvae. *Fishery Bull.* **72,** 1160–1166.

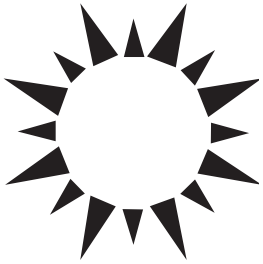
Krenkel, P. A., and Parker, F. L. (1969). "Biological Aspects of Thermal Pollution." Vanderbilt University Press, Nashville, TN.

Langford, T. E. L. (1990). "Ecological Effects of Thermal Discharges." Elsevier Applied Science, London and New York.

Mihursky, J. A., and Kennedy, V. S. (1967). Water temperature criteria to protect aquatic life. *American Fisheries Society,* Special Publication No. 4, 20–32.

Schubel, J. R., and Marcy, B. C., Jr. (eds.). (1978). "Power Plant Entrainment: A Biological Assessment." Academic Press, New York.

Thermodynamics and Economics, Overview



ROBERT U. AYRES
Center for the Management of Environmental Resources
Fontainebleau, France

1. **Work and Power**
2. **Energy (Exergy) Conversion to Useful Work**
3. **The Economy as a Materials Processor**
4. **Further Implications of the First Law of Thermodynamics**
5. **Further Implications of the Second Law of Thermodynamics**

Glossary

exergy The maximum amount of useful work that can be extracted from a material as it approaches thermodynamic equilibrium with its surroundings by reversible processes; the second law of thermodynamics implies that it is never possible to use all available or potentially useful work for practical purposes.

power The time rate at which work is performed, measured in horsepower, watts, or Btu per hour.

thermodynamics The science that studies the nature and transformation of energy. The word thermodynamics stems from the Greek words *therme* (heat) and *dynamis* (power).

work The energy transfer associated with a force acting through a distance.

The word energy is widely misused, and for the sake of precision, it is useful to introduce a different term, exergy, which is less familiar but more precise. Energy is a conserved quantity (the first law of thermodynamics), meaning that it can only change form or quality (e.g., temperature) but can never be created or destroyed. Energy and mass are interconvertible (according to Einstein's formula, $E = mc^2$), although nuclear reactions convert only infinitesimal amounts of mass into energy, whereas there are no practical processes for converting energy back into mass. For all other processes of concern to humans, both the mass and energy content of materials and energy flows are conserved independently. This means that the mass and energy of inputs (including water and air) are exactly the same as the mass and energy content of outputs, including waste products. What has changed is the availability of the energy in the inputs (solar insolation or fuel) for doing potentially useful work.

1. WORK AND POWER

Work is usually defined as a force operating over a distance, and this is scarcely helpful if force is undefined. To avoid infinite regress, the best explanation may be historical. Work was originally conceptualized during the 18th century in terms of a horse pulling a plow or a pump raising water against the force of gravity. Since the discovery of the pendulum, it has been realized that raising a bucket of water or going up a hill converts kinetic energy into potential energy (of gravitation) and that gravitational potential can be converted back into kinetic energy by reversing the process. In the absence of frictional losses, the two forms of energy are equivalent. (Of course, frictional heat becomes unavailable.) Work is also performed when a force acting on a mass increases its velocity and, hence, its kinetic energy, which is essentially mechanical exergy.

A piston compressing a gas does work by increasing the pressure of the gas, just as a gas expanding against a piston can do work by turning a wheel. Adding heat to a compressible fluid in a fixed volume (i.e., increasing its temperature) increases its pressure. This fact makes it possible to convert heat into work. However, it turns out that whereas kinetic and potential energy are interconvertible without loss (in principle), this is not true of heat and

 © 2004 Elsevier Inc. All rights reserved.

pressure. The theory of heat engines, beginning with the idealized engine devised by Sadi Carnot in 1816 and subsequently extended to other engines (e.g., Rankine, Stirling), is all about converting "thermal energy" in the form of heat into "kinetic energy" capable of doing useful work.

Later still, it was realized (independently) by Michael Faraday and Joseph Henry that electric and magnetic fields also constitute forms of potential energy analogous to the gravitational field and that kinetic energy and electromagnetic field energy are interconvertible through the phenomenon of magnetic induction. It follows that changes in electromagnetic potential, known as voltage, can also generate a force and do work. Similarly, kinetic energy of motion, such as when a conductor moves through a magnetic field, can generate a voltage and a current. Normally, there are frictional losses in both processes (known as electrical resistance), but in their absence, the two forms of energy (kinetic and electromagnetic potential) are essentially equivalent.

Finally, during the late 19th century, the notion of potential energy was generalized to chemical reactions by J. Willard Gibbs. Combustion is a process that converts chemical energy—strictly, chemical potential energy—into kinetic energy (motion) and electromagnetic radiation at the molecular level. This heat energy can, in turn, perform physical work by means of a heat engine, as mentioned previously. But there are also chemical processes that generate electrical potentials directly without producing (much) heat, as in storage batteries. Similarly, there are chemical and electrochemical processes that convert heat, chemical potential, and/or electromagnetic potentials into chemical potential, albeit with some entropic losses. Obvious examples are carbothermic reduction (smelting) (e.g., of iron) and electrolytic reduction (e.g., of aluminum). Such processes can also be considered as examples of doing (chemical) work.

It is convenient at this point to introduce the notion of "quasi-work" not involving kinetic energy of motion. This refers to driving an endothermic chemical process (e.g., metal smelting) or moving heat from one place to another across some thermal barrier (e.g., space heating, water heating). Electricity can be regarded as "pure" work because it can perform either mechanical or chemical work with very high efficiency, that is, with very small frictional losses. It is also convenient to distinguish primary and secondary work, where the latter is work done by electrical devices or machines. In all of these cases, the physical units of work are the same as the units of energy or exergy.

Potentially useful work or availability is quantifiable. By general agreement among physical scientists, this quantity is now denoted as exergy. The physical concept underlying the term exergy also goes back to Gibbs, although it has been reintroduced and elaborated several times under a variety of names, including availability, available work, and essergy. The fundamentals of the theory can be found in a number of modern texts on thermodynamics. The formal definition of exergy is the maximum amount of useful work that can be extracted from a material as it approaches thermodynamic equilibrium with its surroundings by reversible processes. The second law of thermodynamics implies that it is never possible to use all available or potentially useful work for practical purposes. There is always some waste or loss of exergy in every process. Lost exergy corresponds to increasing entropy.

There are four components of exergy, each with a corresponding type of work. The first component is kinetic energy associated with relative motion. Mechanical work associated with motion is exemplified by the action of a water wheel or windmill. The second component is potential field exergy, which is associated with gravitational or electromagnetic field differentials. Work must be done on a mass escape from a gravitational field or, on a charged particle, to overcome an electrical field. The third component is thermal exergy (from pressure or temperature differentials), where a heat engine exemplifies the conversion to work. Finally, the fourth component, chemical exergy, arises from differences in chemical composition. Exergy is only nonzero when the system under consideration is distinguishable from its surroundings—the environment—in one or more of these four dimensions.

The first three components of exergy described in the preceding paragraph are analyzed in depth for engineering purposes, especially optimization of mechanical, electrical, and chemical engineering systems, respectively. The term "thermo-economics" has been coined to cover this realm of intellectual activity. However, for purposes of economic analysis on a macro scale, a considerable simplification is possible. In considering mass flows into and out of economic (i.e., industrial) processes, the first three components of exergy can be safely neglected. Only the fourth category, chemical exergy, is important. To calculate the chemical exergy of a mass flow stream, it is only necessary to have data on the chemical composition of that stream (together with thermodynamic characteristics of the components) vis-à-vis the environmental sink into which it (eventually)

flows. Obviously, to calculate chemical exergy, the appropriate environmental reference state must also be characterized quite precisely.

2. ENERGY (EXERGY) CONVERSION TO USEFUL WORK

Reverting to the broader theme, one can say that work is whatever increases the kinetic or potential (including chemical) exergy of a subsystem (within a larger system where energy is always conserved by definition). Electricity can perform either mechanical or chemical work with very high efficiency, that is, with very small frictional losses. Hence, electric power can be regarded as (nearly) "pure" work and is so regarded hereafter. In physical terms, power is defined as work performed per unit of time. Before the industrial revolution, there were only four sources of mechanical power of any economic significance: human labor, animal labor, water power (near flowing streams), and wind power. (The advent of steam power during the early 18th century led to the first quantification of power in terms of equivalent "horsepower" by James Watt.)

It is helpful to distinguish primary and secondary work. Primary work is done by the first stage of energy conversion, for example, by means of a heat engine or hydraulic turbine. Secondary work is work done by electrical devices or machines. It is also useful to consider the notion of quasi-work done by driving an endothermic chemical process or moving heat energy from one place to another across some thermal barrier. In all of these cases, the physical units of work are the same as the units of energy or exergy Electricity can also be converted into heat with very high (100%) efficiency. However, for low-temperature heat, there is an apparent contradiction arising from the fact that resistive heating for low-temperature heat is likely to be less efficient than the use of (electric) heat pumps. For this reason (among others), it is convenient to use the term "second law of efficiency" introduced by the well-known summer study sponsored by the American Physical Society.

Prior to the 18th century, essentially the only source of chemical work (needed mainly for iron and copper smelting, ceramics and glass manufacturing, and cement, quicklime, and plaster of Paris production) was heat from charcoal-fired furnaces. Coal had entirely replaced charcoal in England before 1800 due to prior deforestation. In the United States, the substitution process took roughly a century longer. Inanimate sources of work exceeded animal work in the United States for the first time in 1870. However, not until the 20th century did the contribution from fossil fuel combustion and heat engines using fossil fuels outstrip the contribution from biomass (agriculture and forests), and this was the case only for industrial countries. In many developing countries, the contribution of agricultural and forest products to primary exergy flows and useful work is still dominant.

For purposes of empirical estimation, it is helpful to distinguish between two categories of fuel use. The first category is fuel used to do mechanical work, meaning fuel driving so-called "prime movers," including all kinds of internal and external combustion engines, from steam turbines to jet engines. (Electric motors are not included in this category because electricity is essentially equivalent to mechanical work, as noted previously. Electric power is generated mostly by a prime mover of some other sort.) The second category is fuel used to generate heat as such, either for industry (e.g., process heat, chemical energy) or for space heat and other uses (e.g., hot water for washing, cooking heat for residential and/or commercial users). Historical statistics have never been compiled by official government agencies to distinguish between these two categories of fuel use, but detailed estimates for the 1900–1998 period have been compiled and published for the United States.

It is possible to estimate human and animal contributions to mechanical work crudely on the basis of food or feed intake multiplied by a biological conversion efficiency adjusted for the fraction of time spent doing physical (muscle) work. However, because human labor is treated independently in economic analysis, and because human muscle power is no longer an important component of human labor in the industrial world as compared with eye–hand coordination and brain work, this article neglects it hereafter. (The magnitudes would be trivial in any case.) However, work done by animals, especially on farms, was still important at the beginning of the 20th century and remained significant until midcentury, when trucks and tractors displaced horses and mules. Recent estimates by several authors converge on 4% efficiency or 25 units of feed per unit of work. Higher precision is probably unnecessary for the quantitative estimates in the U.S. case because the magnitude of animal work is relatively small compared with inanimate power sources. In any case, animal conversion efficiency does not change over time.

TABLE I

Average Exergy Efficiency of Performing Work (Percentages)

Year	Electric power generation and distribution	Other mechanical work (e.g., auto transport)	High-temperature industrial heat (steel)	Medium-temperature industrial heat (steam)	Low-temperature space heat
1900	3.8	3	7	5	0.25
1910	5.7	4.4			
1920	9.2	7			
1930	17.3	8			
1940	20.8	9			
1950	24.3	9			
1960	31.3	9			
1970	32.5	8	20	14	2
1980	32.9	10.5			
1990	33.3	13.9	25	20	3

Of course, this is not true of other types of energy conversion systems. Again, official historical time series do not exist except in the case of electric power generation, but crude estimates have been prepared for the U.S. case, as shown in Table I.

3. THE ECONOMY AS A MATERIALS PROCESSOR

The economy has been interpreted as a self-organized system, far from thermodynamic equilibrium. It converts low-entropy, low-information materials into high-entropy wastes and high-information products and services. Put another way, the economy creates useful order from natural disorder and embodies this useful order (mostly) in material form, using large quantities of exergy from biomass (i.e., the sun) or from fossil fuels.

Energy (exergy) flux is transformed into an intermediate service ("useful work"), driving machines and substituting for human and animal labor. By driving down the cost of other products and, thus, increasing demand and production, this long-term substitution has been the dominant driver of economic growth during the past two centuries. In this context, exergy or exergy services (useful work) can be regarded as a factor of production, playing a role complementary to capital services and labor services.

This interpretation explains the close observed correlation between exergy input and economic output without any necessary implication that energy (exergy) content of physical products is proportional to value. It also allows the interpretation of technological progress on a macro level in terms of the efficiency of conversion of exergy inputs-to-service (=work) outputs.

From an evolutionary perspective, the economic system can be viewed as an open system that extracts and converts raw materials into products and useful services. The economy consists of a sequence of processing stages: extraction, conversion, production of finished goods and services, and final consumption (and disposal of wastes). Most of the nonstructural materials are discarded in degraded form. These conversion processes correspond to exergy flows, subject to constraints (including the laws of thermodynamics). The objective of economic activity can be interpreted as a constrained value maximization problem (or as a cost minimization problem). Value is conventionally defined in terms of preferences for consumption goods or services.

The simplest model representation of the economy consists of a single sector producing a single all-purpose product that is both a capital good and a consumption good. This simplification is widely accepted in undergraduate textbooks, despite its lack of realism, because two- or three-sector models are far more difficult to analyze mathematically yet not much more realistic. For example, one might consider a model of two sectors with a single intermediate product. The first sector would include extraction and primary processing (e.g., to finished materials). The second sector would include manufacturing and service activities. Obviously, three or more sectors would add a little more to the realism of the scheme, but the mathematics for a three-sector

model is nearly impenetrable. Of course, the more stages in the sequence, the more necessary it is to take into account feedbacks (e.g., from finished goods to extraction of primary processing sectors). The *N*-sector version would be an input–output model of the Leontief type in which the sequential structure tends to be obscured.

An adequate description of a materials processing system must include materials and energy flows as well as money flows. These flows and conversion processes are governed by the laws of thermodynamics as well as by accounting balances. At each stage, and until the final one, mass flows are split by technological means into "useful" and waste categories. Value (and information) is added to the useful flows, reducing their entropy content and increasing their exergy content per unit of mass (thanks to exogenous inputs of exergy), whereas the high-entropy wastes are returned to the environment.

From a very macroscopic perspective, the output of the economic system—viewed as a materials/exergy converter—can be decomposed into a product of successive conversion stages with corresponding efficiencies, viz.

$$GDP = R \times \frac{IO_1}{R} \times \frac{IO_1}{IO_2} \times \times \quad \ldots \quad \times \frac{GDP}{IO_n},$$
$$= R \times f_1 \times f_2 \times \quad \ldots \quad g \tag{1}$$

where f_1 is the conversion efficiency of the resource (exergy) inflow R into the first-level intermediate product, f_2 is the conversion efficiency to the second-level intermediate product, and so forth. The term g is just the ratio of output to the last intermediate product. The necessary feedbacks are implicitly taken into account in the efficiency definitions.

As a first approximation, it is convenient to assume a two-stage system with a single intermediate product, U. This intermediate product can conveniently be identified as exergy services or useful work. Then,

$$Y = Rfg = Ug, \tag{2}$$

where f is the overall technical efficiency of conversion of "raw" exergy inputs R to useful work output U.

If we define $g = Y/U$ then equation (1) is an identity, not a theory or model. However, the right-hand side of Eq. (1) might be interpreted as an aggregate production function provided that g is a suitable homogenous function of order zero whose arguments are labor L, capital K, and resource (exergy) flows R or useful work U. Effectively, there are two definitions of useful work U to be considered:

$$U_R = f_R R \tag{3a}$$

$$U_E = f_E R \tag{3b}$$

where R is a measure of the total exergy content of all natural resource flows entering the economic system and E is a measure of the commercial fuels (coal, petroleum, and gas as well as hydro power and nuclear heat). The efficiency measures f_R and f_E are defined as total useful work output divided by exergy input. The more general definition f_R takes into account animal work (and biomass inputs), whereas the narrower definition considers only fossil fuels. Multiplying historical resource (exergy) inputs by the estimated efficiencies, and summing over fuels and types, one obtains the results shown in Fig. 1.

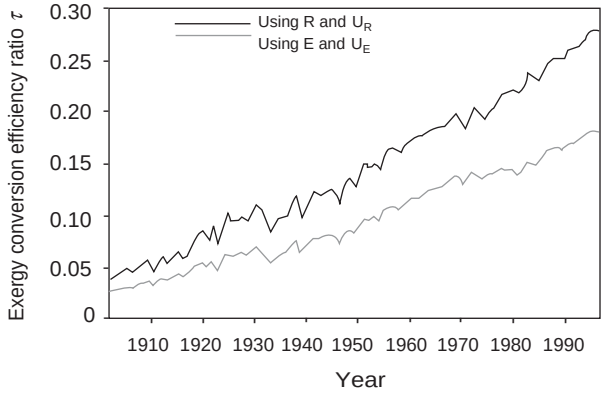


FIGURE 1 Exergy conversion efficiency for two definitions of work and exergy, United States, 1900–1998.

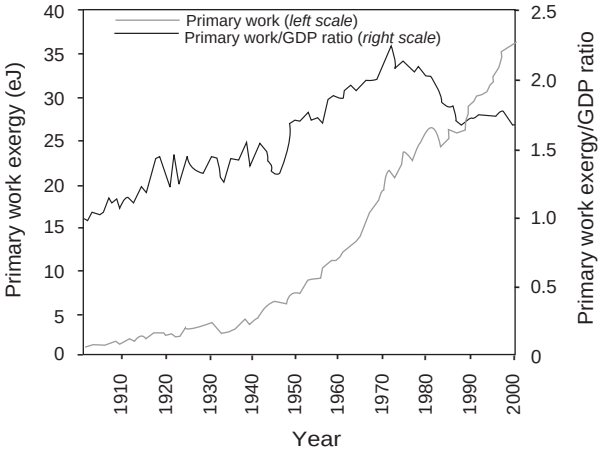


FIGURE 2 Primary work and primary work/GDP ratio, United States, 1900–1998.

The useful work (left axis) and work/GDP ratio (right axis) are also shown for the more general case (f_R) in Fig. 2. The useful work/gross domestic product (GDP) ratio exhibits a pronounced inverted "U" shape, with a clearly defined peak around the time of the Arab oil boycott and the so-called "energy crisis" during 1973–1974.

4. FURTHER IMPLICATIONS OF THE FIRST LAW OF THERMODYNAMICS

The first law of thermodynamics (conservation of mass) implies that mass outputs from any process equal mass inputs. However, useful outputs are almost invariably a fraction of total inputs—sometimes a small fraction (as in the case of refining low-grade ores). In some cases, the output mass is entirely wasted, as with combustion processes. Thus, wastes are an unavoidable by-product of physical production.

On the other hand, the law of mass conservation is far from trivial. The so-called "mass balance principle" states that mass inputs must equal mass outputs for every process (or process step) and that this must be true separately for each chemical element. All resources extracted from the environment must eventually become unwanted wastes and pollutants. Waste emissions are not exceptional phenomena that can be neglected or treated as exceptions. The standard multisector economic model of commodities produced from other commodities is misleading.

It follows, too, that virtually all products are really joint products except that wastes have no positive market value. On the contrary, in most cases, they have a negative value. A producer of wastes will need a "sink" for disposal. Options for free disposal are becoming rarer. Producers must increasingly pay to have waste residuals removed and treated, safely disposed of, or recycled. Therefore, the implication that there exists a price-determined equilibrium between supply and demand (of commodities) must be modified fundamentally.

This means, among other things, that "externalities" (market failures) associated with production and consumption of materials are actually pervasive and that they tend to grow in importance as the economy itself grows. Materials recycling can help (indeed it must), but recycling is energy (exergy) intensive and imperfect (thanks to the second law of thermodynamics), so it cannot fully compensate for a declining natural resource base. Long-term sustainability must depend, to a large extent, on dematerialization and "decoupling" of economic welfare from the natural resource base.

The mass balance condition provides powerful tools for estimating process wastes and losses for industrial processes, or even whole industries, where these cannot be determined directly. Even where other data are available, the mass balance condition offers a means of verification and interpolation to fill in gaps.

The popular notion of perfect recycling (by industrial analogues of decay organisms) with "zero emissions" can be ruled out. It violates the first law of thermodynamics. In fact, the notion that ecosystems are perfect recyclers is false. The biosphere recycles carbon and nitrogen with fairly high efficiency, although coal and petroleum are actually transformed biological wastes. Other elements needed by living systems, including phosphorus, sulfur, calcium, and iron, are not recycled biologically at all. Chalk, iron ores, and phosphate rock all are biological wastes. The ways in which materials are extracted, transformed, and used in the real industrial economy, and the environmental implications, are the substance of the new field of industrial ecology.

5. FURTHER IMPLICATIONS OF THE SECOND LAW OF THERMODYNAMICS

Many economists, as well as most physical scientists, assume that the relationship between economics and the second (entropy) law of thermodynamics concerns resource depletion and scarcity. In this belief. they are, in a sense, disciples of the late Nicolas Georgescu-Roegen, who famously said, "The entropy law is the taproot of economic scarcity" and many other words to that effect. As noted at the beginning of this article, the economy is a system that extracts low-entropy resources from the environment and rejects high-entropy wastes back into the environment. Although solar energy was the original source of fossil fuels that accumulated in the earth's crust during the Carboniferous Era several hundred million years ago, humans are dissipating those resources at a rate thousands or even millions of times faster than the resources were created.

An aspect of the depletion argument concerns recycling. One consequence of the second law of thermodynamics is that recycling can never be 100% efficient. At first sight, this would imply that scarce materials such as platinum must actually disappear from the surface of the earth, and this is not the case.

What is true is that as the quality of the resource base declines toward the average in the earth's crust, the amount of exergy required to extract and reconcentrate it increases to a maximum. In a finite planetary environment, the concentration of a scarce metal can never fall below the average. This means that recycling will become more difficult over time but will never become impossible.

The idea that economic growth must be limited by physical resource scarcity actually has quite a long history. It goes back at least to Thomas Malthus, who saw arable land as the limiting factor. During the 19th century, Jevons worried about future availability of energy from coal. Since 1919, there has been a series of predictions that petroleum reserves are about to run out, with each "crisis" being followed by new discoveries and another glut. Scarcity worries were behind the neo-Malthusian "limits to growth" thesis, propounded during the 1970s by the Club of Rome. However, an authoritative study of minerals published by Resources for the Future strongly indicated that scarcity, as indicated by price, production, and reserve trends, was not yet a problem for any exhaustible mineral resource. A follow-up in 1979 seemed to confirm that result.

The long-running debate between neo-Malthusians (who worry about scarcity) and "cornucopians" (who do not) remains unresolved. It is, in any case, beyond the scope of this article.

SEE ALSO THE FOLLOWING ARTICLES

Aggregation of Energy • Economic Systems and Energy, Conceptual Overview • Economic Thought, History of Energy in • Emergy Analysis and Environmental Accounting • Entropy and the Economic Process • Exergy Analysis of Energy Systems • Exergy: Reference States and Balance Conditions • Marx, Energy, and Social Metabolism • Net Energy Analysis: Concepts and Methods • Physics and Economics of Energy, Conceptual Overview • Thermodynamics, Laws of

Further Reading

Ayres, R. U. (1994). "Information, Entropy, and Progress." American Institute of Physics, New York.

Ayres, R. U. (1998). Eco-thermodynamics: Economics and the second law. *Ecol. Econ.* **26,** 189–209.

Ayres, R. U. (1999). The second law, the fourth law, recycling, and limits to growth. *Ecol. Econ.* **29,** 473–484.

Ayres, R. U., and Kneese, A. V. (1989). Externalities: Economics and thermodynamics. *In* "Economy and Ecology: Towards Sustainable Development" (F. Archibugi and P. Nijkamp, Eds.), pp. 89–118. Kluwer Academic, Dordrecht, Netherlands.

Ayres, R. U., and Warr, B. (2003). Exergy, power, and work in the U.S. economy, 1900–1998. *Energy* **28,** 219–273.

Barnett, H. J., and Morse, C. (1962). "Scarcity and Growth: the Economics of Resource Scarcity." Johns Hopkins University Press, Baltimore, MD.

Carnahan, W., Ford, K.W., Prosperetti, A., Rochlin, G.I., Rosenfeld, A.H., Ross, M.H., Rothberg, J.E., Seidel, G.M., Socolow, R. H. (1975). "Efficient Use of Energy: A Physics Perspective," study report 399. American Physical Society, New York.

Cleveland, C. J., Costanza, R., Hall, C. A. S., and Kaufmann, R. K. (1984). Energy and the U.S. economy: A biophysical perspective. *Science* **255,** 890–897.

Georgescu-Roegen, N. (1971). "The Entropy Law and the Economic Process." Harvard University Press, Cambridge, MA.

Gibbs, J. W. (1948). "The Collected Works of J. Willard Gibbs." Yale University Press, New Haven, CT.

Gruebler, A. (1998). "Technology and Global Change." Cambridge University Press, Cambridge, UK.

Jantsch, E. (1975). "Self-Organization and Planning in the Life of Human Systems." Braziller, New York.

Jantsch, E. (1980). "The Self-Organizing Universe: Scientific and Human Implications of the Emerging Paradigm of Evolution." Pergamon, Oxford, UK.

Malthus, T. R. (1946). An essay on the principle of population as it affects the future improvement of society. *In* "Masterworks of Economics: Digest of Ten Great Classics" (L. D. Abbott, Ed.), pp. 191–270. Doubleday, New York.

Mayumi, K. (1993). Georgescu-Roegen's fourth law of thermodynamics: The modern energetic dogma and ecological salvation. *In* "Trends in Ecological Physical Chemistry" (L. Bonati, U. Consentino, M. Lagsagni, G. Moro, D. Pitea, and A. Schiraldi, Eds.), pp. 351–363. Elsevier, Amsterdam, Netherlands.

Meadows, D. H., Meadows, D. L., Randers, J., and Behrens, W. W., III (1972). "The Limits to Growth: A Report for the Club of Rome's Project on the Predicament of Mankind." Universe Books, New York.

Prigogine, I. (1976). Order through fluctuation: self-organization and social system. *In* "Evolution and Consciousness" (E. Jantsch and C. H. Waddington, Eds.), pp. 95–133. Addison–Wesley, Reading, MA.

Smith, V. K., and Krutilla, J. (eds.). (1979). "Scarcity and Growth Revisited." Johns Hopkins University Press, Baltimore, MD.

Sraffa, P. (1960). "Production of Commodities by Means of Commodities: Prelude to a Critique of Economic theory." Cambridge University Press, Cambridge, UK.

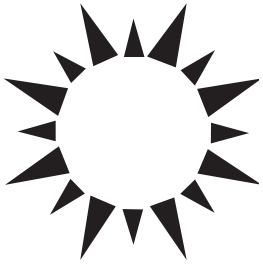
Szargut, J., Morris, D. R., and Steward, F. R. (1988). "Exergy Analysis of Thermal, Chemical, and Metallurgical Processes." Hemisphere Publishing, New York.

von Neumann, J. (1945). A model of general economic equilibrium. *Rev. Econ. Stud.* **13,** 1–9.

Walras, L. (1874). "Elements d'economie politique pure." Corbaz, Lausanne, Switzerland.

Yergin, D. (1991). "The Prize: The Epic Quest for Oil, Money, and Power." Simon & Schuster, New York.

Thermodynamic Sciences, History of



KEITH HUTCHISON
University of Melbourne
Melbourne, Australia

1. **The Carnot Memoir of 1824**
2. **Thomson's Early Enthusiasm for Carnot (Late 1840s)**
3. **The Clash with Energy Conservation**
4. **Clausius Resolves Part of the Clash (1850)**
5. **Rankine's Vortex Hypothesis (1850)**
6. **Thomson Reconciles Dissipation with Energy Conservation**
7. **Interpreting the Second Law**
8. **The Phenomenological Style of Thermodynamics**

Glossary

Boltzmann, Ludwig (1844–1906) Austrian physicist who sought (initially) a purely mechanical reduction of the second law but later argued it was only a statistical truth.

Carnot, Nicolas Léonard Sadi (1796–1832) French military engineer, author of the 1824 "Reflections," which posed the problems (and articulated the core reasoning) that later generated the discipline of thermodynamics.

Clapeyron, Benoit-Pierre Émile (1799–1864) French engineer who published an account of Carnot's theory in 1834.

Clausius, Julius Rudolf Emmanuel (1822–1888) German physicist who created the discipline of thermodynamics by reconciling the core of Carnot's memoir with the energy conservation of Joule, Helmholtz, Mayer, *et al.*

Helmholtz, Hermann von (1821–1894) German physicist/physiologist. One of the main discoverers of energy conservation, notably in his 1847 memoir, "On the Conservation of Force."

Joule, James Prescott (1818–1889) British physicist. One of the main discoverers of energy conservation. He gave it much experimental support in the 1840s.

Maxwell, James Clerk (1831–1879) British physicist. One of several to argue that the second law was only a statistical truth.

Mayer, Julius Robert (1814–1878) German physician. One of the main discoverers of energy conservation.

Rankine, William James Macquorn (1820–1872) Scottish engineer who developed a strange kinetic thermodynamics around 1850. He introduced entropy into thermodynamics but denied dissipation. He made the subject accessible to engineers.

Regnault, Henri Victor (1810–1878) French physicist who, in the mid-19th century, carried out a detailed and authoritative study of the thermal properties of steam and other gases.

Thomson, James (1822–1892) British engineer, brother of William.

Thomson, William, later Baron Kelvin of Largs (1824–1907) British physicist/engineer who took up the Carnot theory in the late 1840s. Later introduced Carnot's dissipation into thermodynamics.

The fact that thermal energy cannot be fully transformed into mechanical work has far-reaching consequences. Thermodynamics is the science that develops these consequences. It emerged in the approximately 30 years that followed Sadi Carnot's amazingly fertile 1824 study of the efficiency of steam engines when the key ideas in that study were reconciled (in the early 1850s) with the principle of energy conservation (as consolidated in the late 1840s). This article is a survey of the thermodynamical thinking of those three decades.

1. THE CARNOT MEMOIR OF 1824

1.1 The Puzzle of Steam Engine Efficiency

In England, around the end of the 18th century there was an extraordinary increase in the efficiency with which coal could be converted into useful work. This was made possible by a series of improvements, both

 © 2004 Elsevier Inc. All rights reserved.

major and minor, in the design of steam engines, the most notable of which were those introduced by James Watt (the separate condenser, double action, and expansive working), together with a growing tendency to use increasingly higher steam pressures. Progress was so rapid—from 1775 to 1825 there was a 10-fold increase in efficiency—that the question came to be asked as to whether the motive power of heat is unbounded, whether the possible improvements in steam engines have an assignable limit—a limit that the nature of things will not allow to be passed by any means whatever—or "whether, on the contrary, these improvements may be carried on indefinitely".

1.2 Carnot's Key Insight: Cold Is Needed to Produce Work

The challenge of answering this question was taken up by a young French Army engineer, Sadi Carnot, who in 1824 published a masterly memoir on the subject, "Reflections on the Motive Power of Heat." In order to produce work by thermal means, Carnot asserted (recalling Watt's separate condenser), it is not sufficient to have a source of heat: It is also necessary to have a "source" of cold—that is, a supply of some substance at a low temperature that can be used to absorb the "exhaust" heat that (he proposed) every thermal engine necessarily produces. He theorized that if the temperature of these two sources is fixed, there is a maximum to the amount of work that may be derived by the extraction of unit quantity of heat from the hot source. This maximum, he indicated, is the amount of work done by any thermal engine working reversibly between the two temperatures in question. Furthermore, Carnot was able to calculate a few rough estimates of his maxima: He could not do more because he lacked the necessary empirical data.

1.3 Water Engines and Reversibility

The reasoning Carnot used to support his conclusions is particularly important in that it was the persuasiveness of his arguments that was later to ensure their survival in the face of experimental evidence that directly contradicted the theory of heat he had used. To appreciate this reasoning, it is helpful to consider the development of the theoretical understanding of water power, a close rival to steam in the 18th century. This century was one of rapid progress in the harnessing of the power of falling water, and some of the designs produced by engineers are astonishing in their sophistication. A good example of this is the so-called column-of-water engine, whose appearance and action are both very similar to those of a high-pressure steam engine and so unlike the crude water wheel we might unreflectingly imagine to be the typical product of the age.

Such an engine (and many other water engines) could be run backwards to act as a pump, and in the ideal situation in which all friction and similar imperfections were absent, the machine would produce exactly as much work in consequence of the descent of a unit of water from its supply reservoir to its exhaust as would be required to pump that same unit of water from the exhaust back to the supply with the engine running backwards. Such could be called a (perfectly) reversible water engine.

Furthermore, no water engine could be more efficient that a reversible one. If another engine were to be capable of producing more work than a reversible one, it could be coupled to a reversible engine, with the latter used as a pump, consuming only part of the work of the former while supplying its total fuel (i.e., water flows). Therefore, the combination would be producing work without making any demands on its environment and without suffering any net internal change. Such perpetual motion, it was concluded in the 18th century, was impossible. So there is a clear limit to the efficiency of a water engine, and this limit depends on the height of the intake and exhaust reservoirs. Furthermore, the limit is the actual amount of work produced by a reversible engine acting over the same drop.

All these theories had been developed by the beginning of the 19th century, and they are important here because what Carnot did to solve his problem was to borrow (and enhance) the reasoning of the water engineers and reapply it to heat engines.

1.4 Heat Conservation and Caloric Theory

There is however one vital precondition that must be satisfied so that the analogy between heat and water engines is complete enough for the arguments evolved for water engines to retain their validity when applied to heat engines: No heat must be consumed during the production of work by heat engines. Just as a water engine works by lowering a fixed quantity of water from a high reservoir to a low one, so too must it be assumed that a heat engine operates by "lowering" an invariable quantity of heat from a region of high temperature to a region of low temperature. Under

this assumption (now known to be false), one can argue for heat engines just as has already been done for water engines, and the analogy is so close that there is no point in repeating the argument here.

To Carnot, this crucial assumption of heat conservation was a natural one to make and he adopted it explicitly. It was inherent in the caloric theory, then the most widely accepted explanation of the nature of heat and especially favored in France. According to this theory, heat was (or behaved like) an indestructible fluid ("caloric"), whose flows in and out of bodies accounted for their various thermal properties. Therefore, by adopting the caloric theory, and by borrowing the beginnings of a line of reasoning from his predecessors in hydraulic engineering, Carnot was able to devise a remarkably convincing solution to the problem he had set himself—a solution that had some interesting consequences.

It enables one to impose, seemingly a priori, a variety of restrictions on the thermophysical properties of all substances and hence to deduce some of their physical characteristics from a knowledge of apparently quite different ones. The most startling early example of this was the deduction by James Thomson (using Carnot's reasoning approximately 25 years after Carnot had set it down) of the hitherto unsuspected fact that since water expands on freezing, the melting point of ice must decrease as the pressure exerted on it increases. Or else, the Carnot logic demands, a perpetual motion machine is possible. The power of such reasoning was to prove most attractive to some of Carnot's later followers. It eventually became a standard feature of thermodynamics.

1.5 Irreversibility and Dissipation

Furthermore, the Carnot theory implies that whenever heat descends from a region of high temperature to one of low temperature without being used to produce the maximum possible work—whenever heat descends irreversibly (e.g., as it does in simple heat conduction)—there takes place concomitantly some sort of destruction of work capability: The universe is less capable of doing work after the descent than it was before the descent. Irreversibility, in other words, implies permanent waste. Carnot did not stress this consequence of his analysis, but this key thermodynamic idea was taken up by his followers.

At first, followers were few and far between because Carnot's discussion was very abstract and had little in common with the physics and theoretical engineering of his contemporaries. The first person to make something substantial out of it was another French engineer, Clapyeron, who in 1834 published a summary of the original, rephrasing much of the analysis in more mathematical form (including the PV diagrams so common in thermodynamics today). Clapyeron did make one important conceptual addition, however. He supplied something missing from the Carnot memoir—a reversible design for an ideal steam engine. (Despite the context that motivated his study, Carnot had not been able to solve this small riddle.) In any case, through Clapyeron's revision, the theory came to the attention of the young British physicist–engineer William Thomson (later Lord Kelvin) and the German physicist Clausius, both of whom were strongly attracted to it.

2. THOMSON'S EARLY ENTHUSIASM FOR CARNOT (LATE 1840S)

Thomson saw that the Carnot theory could be used to define a temperature scale with the long sought-after property of being independent of the physical characteristics of any narrow class of substances, and he published such a definition of "absolute temperature" in 1847. The principle behind it is simple: Unit temperature interval is defined as that temperature interval through which the reversible descent of unit quantity of heat produces unit quantity of work. Carnot's logic shows that this depends on nothing more than the two temperatures, exactly as required. Furthermore, Thomson, in this and a later paper, systematically extended some of Carnot's rough calculations (using data recently obtained by Regnault) to establish the numerical connection between such an abstract scale and the practical scales in everyday use. (It turned out to be approximate coincidence.) In doing so, Thomson was also able to improve Carnot's estimates so far as to tabulate systematically the quantity of work that unit quantity of heat could produce in descending reversibly from the higher to the lower of any pair of temperatures chosen from the range covered by his data, 0–230°C. A few years later, Thomson was able to verify his brother's prediction (mentioned previously) that the freezing point of water is depressed by pressure, a totally unexpected consequence of Carnot's theory.

It is not difficult to understand why Thomson found the analysis attractive. Apart from the inherent elegance of its reasoning, it (i) led to an effectively complete solution to the problem of evaluating the motive power of heat; (ii) provided a basis for

absolute thermometry; and (iii) showed itself capable of discovering new phenomena, well beyond the imagination of its creator.

3. THE CLASH WITH ENERGY CONSERVATION

Yet it was wrong, as Thomson was slowly and reluctantly forced to admit. His personal mentor was Joule, but Joule was only one of a small group of men (Mayer and Helmholtz were the other principals) who, in the 1840s, gradually persuaded the scientific world to give up the belief that heat was conserved and to replace that axiom with the principle of energy conservation, later known as the first law of thermodynamics. This new doctrine spread remarkably fast, and by 1850 it was widely accepted that heat and work were interconvertible. Thus, heat did not behave at all like the fluid caloric, and this fact utterly undermined Carnot's argumentation. Carnot had been suspicious of the caloric theory in the 1824 memoir but still managed to put it to good use. His posthumous papers show that he soon came to the conclusion that heat could be transformed into work and that he even managed to estimate the conversion factor.

Joule objected to the Carnot theory on two grounds and urged its abandonment. First, it regarded the essence of the process whereby work is generated thermally as descent of heat, whereas Joule saw it as a destruction. Second (and apparently more important for Joule and colleagues), it demanded the annihilation of motive power whenever an irreversible (thermal) process took place. To Joule, the universe was equipped with a fixed quantity of motive power, and no natural process could cause either an increase or a decrease in this total: The most that could happen was that it could alter its form.

Thomson was one of the first to appreciate the force of Joule's evidence, but he could not accept Joule's recommendation to abandon Carnot. Thomson knew that no matter how correct Joule seemed to be, there was still a lot of truth in Carnot's theory. The conflict between these two theories was a paradox for Thomson, and he looked to further experiment to resolve it.

4. CLAUSIUS RESOLVES PART OF THE CLASH (1850)

In fact, the paradox was soluble without further experiment, as Clausius was soon to reveal. In May 1850, he published a major paper, "On the Motive Power of Heat, and on the Laws which Can Be Deduced from It for the Theory of Heat," in which he discovered the new science of thermodynamics by showing "that [Joule *et al.*'s] theorem of the equivalence of heat and work, and Carnot's theorem, are not mutually exclusive, but that, by a small modification of the latter, which does not affect its principal part, they can be brought into accordance."

What was Clausius's approach? Above all, he accepted unequivocally the new energy principle and attempted to salvage as much of the Carnot theory as possible. In fact, he found that he could salvage so much that what was lost might reasonably be described as inessential. In particular, he perceived that although the Carnot theory was correct in demanding a descent of heat when work was produced thermally, it was wrong in assuming nothing but a descent: Clausius saw that in reality both a descent and a destruction took place. Furthermore, Clausius saw that the axiom of heat conservation was not really inherent in Carnot's conclusions but was merely used as a tool to establish them. Clausius realized that if this axiom were replaced by a different one, then new arguments might be devised to justify much the same conclusions as Carnot had deduced from the false premise. Clausius was able to do this, the most important example being his new deduction that a reversible engine is one of maximal efficiency.

Carnot argued that if a reversible engine absorbing heat at one temperature and releasing it at another is not the most efficient of all engines absorbing and releasing heat at the same temperature, then a perpetual motion would be possible. Clausius could only establish that if the reversible engine were not the most efficient, then a heat pump would be theoretically constructible capable of pumping heat from a region of low temperature to a region of high temperature without consuming any mechanical work (later called a perpetual motion of the second type). Such a pump he declared to be impossible because heat "always shows a tendency to equalize temperature differences and therefore to pass from hotter to colder bodies."

The conclusion that the reversible engine is that of greatest efficiency remains unchanged. Only the argument justifying the conclusion has been altered, and even then, the new argument retains the spirit of the old. Carnot's premise that heat is conserved is replaced by the premise just quoted. This is the first published version of what is now called the second law of thermodynamics.

Clausius also showed that his new approach was just as fertile as that of Carnot, and that it led to similar interrelations between the physical characteristics of substances. In particular, Clausius showed that the new theory also requires that the freezing point of water be lowered by pressure or else one of the heat pumps forbidden by the second law would be constructible. The new theory also revealed unsuspected facts: Clausius deduced from it the apparently paradoxical result that saturated steam has a negative specific heat (i.e., as steam expands under conditions of saturation it absorbs heat, whereas its temperature decreases).

5. RANKINE'S VORTEX HYPOTHESIS (1850)

Clausius was not the first to publish the result just mentioned. Three months before Clausius's publication, the Scottish engineer Rankine (later a colleague of William Thomson's at the University of Glasgow) had come to the same conclusion in one of two abstruse papers outlining a complicated mechanical theory of heat. Rankine's papers, read in February 1850, are the very first to base a significant theory of the transformability of heat into work on energy conservation. However, they are not seen as founding the science of thermodynamics because they contain virtually nothing recognizable as phenomenological thermodynamics.

In contrast, they deal with the internal mechanics of a characteristic vortex model of the atom, which Rankine had proposed to explain the thermal and radiant properties of matter. According to this model (later transformed into the Bohr atom), macroscopic matter is imagined as composed of atoms (or molecules; the distinction is irrelevant), assumed for mathematical conveniences to be spherical, with each consisting of a nucleus surrounded by a swirling gas-like atmosphere. Each portion of this subatomic atmosphere rotates in a small circle about the radius connecting it to the nucleus. Centrifugal effects are induced by the swirling, and upon elastic transmission to the exterior of the substance in question and modification by interatomic actions at a distance, they generate the external pressure exerted. Hence, Rankine was able to establish relationships between the external pressure and the kinetic energy of subatomic thermal motions. The latter he regarded as simply the heat "contained" within the substance, and the former is what does external work. Therefore, it is not too difficult to accept that such a model might lead to some sort of theory of the convertibility of heat and work.

5.1 Rankine Introduces the Entropy-Function

It is remarkable that Rankine's vortex hypothesis did not lead to some sort of theory, but (despite its complete and justifiable rejection by modern science) it led to a variety of results identical to the correct thermodynamic ones. Nor did Rankine use his model to provide an alternative derivation of such results after they had been discovered by someone else: He himself used the model to make discoveries. The negative specific heat of saturated stream is one example. Another is Rankine's discovery of the existence of entropy (an achievement often falsely attributed to Clausius). By a devious path (which is ignored here), the vortex theory led Rankine to conclude that many materials will possess a state function, typically $S(P, V)$ satisfying the fundamental equation $dH = T \times dS$, where T is absolute temperature, and dH is the heat absorbed by the material during some infinitesimal change. It can easily be seen from this equation that S is the entropy, although in 1854 Rankine called it the thermodynamic function. He introduced it as some sort of thermodynamical potential (a function that enabled him to perform important calculations) but had absolutely no idea of its connection with Carnot's important notion that irreversible thermal processes were wasteful. He did not realize that dH is only equal to $T \times dS$ for reversible changes and at one point even announced that his thermodynamic function was conserved.

6. THOMSON RECONCILES DISSIPATION WITH ENERGY CONSERVATION

In his 1850 paper, Clausius gave little attention to Carnot's notion of dissipation, but this was soon rescued by William Thomson. Approximately 1 year after Rankine's early papers, and approximately 10 months after Clausius's breakthrough, Thomson published his own solution to the paradox that had previously confronted him in the first of a long series of papers titled "On the Dynamical Theory.…" In essence, Thomson's solution is identical to that of Clausius, although Thomson recognized more clearly than Clausius that a distinct new law had been

discovered: He declared that "it is impossible, by means of inanimate material agency, to derive mechanical effect from any portion of matter by cooling it below the temperature of the coldest of the surrounding objects." (The restriction here reflects Thomson's suspicion that biological processes were exempt from the second law.)

Thomson published quite extensively on the subject during the next few years, concentrating on the rectification of his earlier work on the Carnot theory. He modified the old definition of absolute temperature into that now used today, simply to retain the convenient coincidence with everyday thermometers. Oddly, this purely pragmatic decision had the effect of giving the new scale an absolute zero. (However, such an idea was not surprising since the caloric theory had made it seem quite natural that there be some natural minimum to temperature—when all caloric had been extracted from a body.) This scale was subjected to close scrutiny in the long series of Joule–Thomson experiments, performed to remove doubt about some of the empirical parameters vital to a range of thermodynamic calculations. He also returned to Joule's earlier puzzle of reconciling the losses apparently generated by irreversibility with the conservation asserted by the first law.

In an 1852 paper titled "On a Universal Tendency in Nature to the Dissipation of Energy," Thomson resolved this paradox in much the same way that Clausius had resolved Joule's other dilemma—by simply thinking things through. Energy, Thomson affirmed, is always conserved. However, it shows a spontaneous tendency to transform (or dissipate) itself into forms that are less and less accessible to humanity. Therefore, it is the availability of energy, and not energy itself, that is destroyed by irreversibility. In consequence, added Thomson, the earth is gradually dying and will one day be unfit for human habitation.

In 1865, this idea was finally attached to Rankine's thermodynamic function when Clausius wrote a systematic review of his work of the past 15 years. Clausius had been hoping to find some parameter whose values indicated whether a transformation was allowed by the second law. He was never able to complete this task, but his research had shown that the dissipation noted by Thomson and Carnot had the effect of increasing Rankine's state function. Furthermore, this fact was substantially equivalent to earlier versions of the second law, so Rankine's function clearly captured something very profound. Accordingly, Clausius gave the function a new name, entropy, deliberately adapted (from a Greek word for "transformation") to parallel the word energy. Together, these two magnitudes seemed to summarize the whole of this new branch of physics.

7. INTERPRETING THE SECOND LAW

With these statements articulated, the foundations of the subject were complete. However, an obvious question remained: How were the newly discovered laws of thermodynamics to be understood? With the first law, there was little problem. When heat conservation was abandoned in the 1840s, it was universally taken for granted that heat was simply some form of microscopic motion. Such an idea had been around since the 17th century, and its competitor was now eliminated. Thus, the new account was referred to as the dynamical theory of heat (or "thermodynamics" in condensed form). However, it was well-known that mechanical energy was conserved in many dynamical systems, most notably those that were isolated and whose internal forces were conservative. Helmholtz had made much of this fact in his 1847 case for energy conservation.

Dynamics, however, provides no obvious clue to the second law. It is true that Rankine had seemed to derive the entropy dynamically, but he had totally ignored dissipation. In any case, his account remained tied to the details of a very speculative hypothesis about the microstructure of matter. Despite Rankine's later attempts to generalize the theory, none of his contemporaries saw Rankine as providing acceptable foundations for the new subject (although Rankine's famous textbooks distilled the subject for application by engineers). For a while, it seemed (especially to Clausius and Boltzmann in the late 1860s and early 1870s) that there might be some complicated dynamical function that tended to increase with time, but a potent objection was raised against this proposal. Conservative dynamical systems appear to be completely reversible: If a dynamical function increases during some change of state, it will have to decrease when that state change is reversed. Therefore, no purely mechanical function can always increase. In the end, it was concluded that the second law is purely statistical. As Maxwell stated, "The second law of thermodynamics has the same degree of truth as the statement that if you throw a tumblerful of water into the sea, you cannot get the same tumblerful of water out again." His "demon" was introduced (in correspondence, in the late 1860s)

to defend this view: It was a microscopic intelligence apparently capable of circumventing the second law by separating the faster molecules in a gas from the slower ones. However, this made it seem that thermal motions were not intrinsically random, just beyond human manipulation. Later, a more profoundly stochastic interpretation was adopted.

8. THE PHENOMENOLOGICAL STYLE OF THERMODYNAMICS

To pursue this question further would take us into other branches of physics, kinetic theory, and statistical mechanics because thermodynamics traditionally avoids dealing with the microstructure of matter and focuses primarily on data at the level of everyday human experience—pressures, volumes, temperatures, etc. Such a restriction was quite natural in the Carnot memoir because the subject matter did not demand any microanalysis. However, in 1850, the situation was quite different. The focus was no longer on steam engines but on energy, and both Clausius and Rankine made vital reference to the microscopic energy flows that everyone believed were occurring. However, insofar as they did this, their research was speculative. When Thomson later prepared his own version of the theory, he purged his account of all such speculation. This was not because Thomson had any methodological doubts about discussing the microworld envisioned by Clausius. Thomson's decision was purely pragmatic. He was, after all, not the pioneer here—he was rephrasing the work of the other two—and he saw that it was possible to solve the paradoxes articulated by Joule without so much reference to doubtful detail. Since it was possible to eliminate a source of uncertainty by avoiding the microdynamics, he did it (and did it well). Since then, the practice has remained.

SEE ALSO THE FOLLOWING ARTICLES

Conservation of Energy Concept, History of • *Energy in the History and Philosophy of Science* • *Entropy* • *Heat Transfer* • *Temperature and Its Measurement* • *Thermodynamics and Economics, Overview* • *Work, Power, and Energy*

Further Reading

Cardwell, D. S. L. (1971). "From Watt to Clausius: The Rise of Thermodynamics in the Early Industrial Age." Heinemann, London.

Carnot, S. (1986). "Reflexions on the Motive Power of Fire: A Critical Edition with the Surviving Manuscripts," [R. Fox, trans. (from 1824 French) and Ed.]. Manchester Univ. Press, Manchester, United Kingdom.

Carnot, S., Clapeyron, E., and Clausius, R. (1960). "Reflections on the Motive Power of Fire; and Other Papers on the Second Law of Thermodynamics" (E. Mendoza, Ed.). Dover, New York.

Clausius, R. J. E. (1867). "The Mechanical Theory of Heat" (J. Hirst, Ed.). Van Voorst, London.

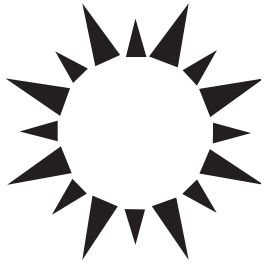
Gillispie, C. C., *et al.* (eds.). (1970–1990). "Dictionary of Scientific Biography," 18 Vols. Scribner's, New York.

Home, W. W., and Gittins, M. J. (1984). "The History of Classical Physics: A Selected, Annotated Bibliography." Garland, New York.

Rankine, W. J. M. (1881). "Miscellaneous Scientific Papers" (W. J. Millar, Ed.). Griffin, London.

Thomson, W. (Lord Kelvin) (1882–1911). "Mathematical and Physical Papers" (W. Thomson and J. Larmor, Eds.), 6 Vols. University Press, Cambridge, UK. [See Vol. 1]

Thermodynamics, Laws of



SERGIO ULGIATI and CARLO BIANCIARDI
University of Siena
Siena, Italy

1. **Introduction**
2. **The Principle of Equivalence**
3. **The Zeroth Law of Thermodynamics**
4. **The First Law of Thermodynamics**
5. **The Second Law of Thermodynamics**
6. **The Third Law of Thermodynamics (Nerst's Theorem)**
7. **How Many Laws for Thermodynamics?**
8. **Conclusion**

Glossary

emergy The amount of exergy of one form (usually solar) that is directly or indirectly required to provide a given flow or storage of exergy or matter. In so doing, inputs to a process are measured in common units. Solar emergy is measured in solar emergy joules (seJ). Emergy is a memory of the available energy (exergy) previously used up in a process and measures the global support from the biosphere to a process. Converting flows of different nature into flows of one kind requires conversion coefficients called transformities (seJ/J). Emergy accounting acknowledges that different resources, measured by their exergy content, may require a different environmental support and may show different properties.

energy Usually defined as the capacity for doing mechanical work (i.e., displacing an object from a position to another by means of the application of a force). This definition is sometimes criticized for being reductive and not always easily applicable to processes other than physical. A broader definition is "the ability to drive a system transformation," which clearly includes all kinds of physical and biological processes.

exergy The amount of work obtainable when some matter is brought to a state of thermodynamic equilibrium with the common components of the natural surroundings by means of reversible processes, involving interaction only with the previously mentioned components of nature. Exergy is measured in joules or kilocalories. Real processes are irreversible. Therefore, measures of actual work compared with theoretical value provide indicators of exergy losses due to the irreversibility of a process. In addition to being a measure of what can be in principle obtained from a resource, exergy also provides a thermodynamic assessment of potential improvements at the scale of the investigated process.

power The amount of energy transferred or work performed in the unit time (usually in 1 s).

reversible/irreversible A reversible change in a process is a change that can be reversed by an infinitesimal modification of any driving force acting on the system. The key word "infinitesimal" sharpens the everyday meaning of the word "reversible" as something that can change direction. In other words, the state of the system should always be infinitesimally close to the equilibrium state and the process be described as a succession of quasi-equilibrium states. A system is in equilibrium with its surroundings if an infinitesimal change in the conditions in opposite directions results in opposite changes in its state. It can be proved that a system does maximum work when it is working reversibly. Real processes are irreversible; therefore, there is always a loss of available energy due to chemical and physical irreversibilities.

self-organization All systems in nature self-organize—that is, use the available energy and matter inputs to achieve higher levels of structure and functioning under natural selection constraints. In so doing, they may maintain their present level of complexity or become increasingly more complex.

work The transfer of energy occurring when a force is applied to a body that is moving in such a way that the force has a component in the direction of the body motion. It is equal to the line integral of the force over the path taken by the body, $\int \mathbf{F} \times d\mathbf{s}$. Its meaning is different from the usual life work, often associated with activities not always involving visible motion. Physical work requires a body motion under the action of the force.

Thermodynamics, the so-called queen of sciences, is important because it is the physical science that accounts for the transformations of thermal energy (heat) into mechanical energy and its equivalent forms (work, other forms of energy, and self-organization of complex systems) and vice versa.

 © 2004 Elsevier Inc. All rights reserved.

All phenomena and processes that happen on Earth as well as in the universe belong to the realm of this science. Although it was relatively easy to investigate the dynamics of equilibrium systems and generate a set of laws describing the tendency of systems toward equilibrium by means of ideal reversible processes (classical thermodynamics), irreversible processes dealing with nonequilibrium systems (including self-organization of the biosphere and life processes) are not fully described within a consistent theoretical framework, but significant progress has been made (nonequilibrium thermodynamics). This article provides a historical outline of the development of thermodynamics, with reference to some of the main scientists who contributed to its development. It then provides a short description and explanation of the laws of thermodynamics, including the mathematical derivation of their main features and consequences. Finally, a series of tentative "fourth laws of thermodynamics" are discussed in order to provide a picture of the current efforts of the scientific community toward a thermodynamic description of self-organization patterns of complex and living systems.

1. INTRODUCTION

Classical thermodynamics was founded between 1850 and 1860 by W. Thomson (Lord Kelvin), R. Clausius, and J. C. Maxwell, building on the seminal work of Lazare Sadi Carnot, "Réflexions sur la Puissance Motrice du feu et sur les Machines Propres à Developper cette Puissance" (1824). They went beyond the theory of caloric that dominated the early studies about heat and its nature between 1700 and 1800 thanks to the experimental results of R. Mayer and J. P. Joule. These authors had achieved a numerical value for the quantitative equivalence between mechanical work and heat. Hence, any reason for the supposed existence of an immaterial substance called "caloric" was experimentally rejected. The findings of Mayer and Joule were also the basis for the future statements of the first law (conservation of energy).

It is worth noting that the investigation of thermal engines and the limits to the conversion of heat to mechanical work (Carnot, second law) preceded the statement of the first law of thermodynamics. This is because heat conversion to mechanical work (steam engine; second half of the 18th century) generated attention first to the existing constraints and then a clear understanding of the nature of heat and a quantitative equivalence between heat and mechanical work (mid-19th century).

In the following years, Clausius provided Carnot's "Réflexions" with a stronger mathematical basis and introduced the concept of entropy, S, developing a clear statement of the second law (1864). The entropy concept is perhaps the most complex one in physical science. It was gradually assigned many meanings belonging not only to thermodynamics but also to a variety of other fields and disciplines (e.g., information theory). A further definition of the entropy concept came from L. Boltzmann in the second half of the 19th century by means of the so-called H-theorem, in which the statistical meaning of entropy was definitely assessed. According to this theorem, any macroscopic thermodynamic state of a system—defined by measurable pressure P, volume V, and temperature T—can be generated by N microscopic behaviors of the system's molecules or microstates. Then the probability of the macrostate is proportional to N, and Boltzmann concluded that the most probable macrostate was the equilibrium state, to which the maximum number of microstates can be associated. As a consequence, the entropy of such an equilibrium state should also be a maximum, due to the Boltzmann formula

$$S = k \times \log N,$$

where k is the so-called Boltzmann constant. Any thermodynamic system tends to its equilibrium state, defined by certain state parameters P, V, and T, in an irreversible way. According to Boltzmann, we can state that this irreversibility is of statistical nature. In other words, the equilibrium state with maximum entropy is the most probable state, although a system's shift to a lower entropy state cannot be a priori excluded (with an extremely low probability).

At the beginning of the 20th century, a rich debate took place about reversibility (or symmetry with respect to time) in classical mechanics and irreversibility in thermodynamics. Fundamental contributions were given by Caratheodory, Ehrenfest, Hilbert, and Nerst. The latter proposed a third law of thermodynamics, characterized by the concept of absolute entropy (1906). He postulated that $S = 0$ for any system at $T = 0$ (where T is absolute temperature). This made possible the definition of a scale of positive absolute entropies, based on which the differences of entropy between two states could be calculated.

During the 20th century, a great number of authors applied the laws of thermodynamics to

problems different than heat-to-work conversion, such as the self-organization and growth of complex, far from equilibrium systems (biological, economic, and social). Authors such as Onsager, Prigogine, Georgescu-Roegen, Lotka, and Odum contributed to this debate and several statements of thermodynamics laws were tentatively reformulated or introduced anew.

2. THE PRINCIPLE OF EQUIVALENCE

Energy is usually referred to as the ability to do mechanical work, *W*. The latter is defined without ambiguity as the scalar product of two vectors, the force **F** and the displacement **s**:

$$W = \mathbf{F} \times \mathrm{s} = Fs\cos\theta, \qquad (1)$$

where F and s are the intensities of force and displacement vectors, respectively, and θ is the angle between them. Work, and therefore energy, can be experimentally measured (with units of Joule in the International System of Units).

Today, the link between work and energy (and the link between energy and heat) appears to be clearly assessed and quantifiable, but this was not so clear to the earlier scientists who investigated the nature of heat. Many experiments or phenomena demonstrate that some heat is generated whenever mechanical energy is used up or lost. Moreover, we know that heat can produce mechanical work, as a steam engine demostrates. Thus, we can argue that heat is some form of energy. However, the perception alone is not sufficient in physics. A quantitative and experimentally measurable equivalence between heat and work is needed. Many experiments were performed during the mid-19th century and successively by J. P. Joule and other researchers that were aimed at measuring the ratio between work (W) and heat (Q) when a system undergoes a cyclical process or when its final state coincides with its initial state.

In this regard, the most classic experiment is the so-called Joule's wheel. The original Joule's scheme of the experiment is shown in Fig. 1. The system consists of a liquid in a vessel, which is in turn contained in a calorimeter in order to measure the heat released or absorbed by the liquid. A paddle wheel is dipped into the liquid. The fall of a weight connected to the wheel via a system of pulleys makes it rotate so that some mechanical work is performed on the liquid. In 1845, Joule described the experiment to readers of the *Philosophical Magazine* as follows:

> *The paddle moved with great resistance in the can of water, so that the weights (each of four pounds) descended at the slow rate of about 1 foot per second. The height of the pulleys from the ground was 12 yards, and consequently, when the weights had descended through that distance, they had to be wound up again in order to renew the motion of the paddle. After this operation had been repeated 16 times, the increase of the temperature of the water was ascertained by means of a very sensible and accurate thermometer.*

The water, initially at a temperature T_A, warms by friction to a final temperature T_B. The mechanical work performed on the liquid is measured by the work done by the force of gravity when the weight falls by a height h ($W = mgh$, where m is the mass of the weight, g is gravity, $9.8\,\mathrm{m/s^2}$), assuming the pulley's friction to be negligible. The calorimeter measures the heat generated within the liquid during the process and then released by the liquid when it cools down to its initial temperature T_A. Joule and other researchers performed similar experiments using different liquids. Joule found a W/Q ratio equal to 4.18 J/cal using water and 4.16 J/cal using mercury. Hirn measured the W/Q ratio in a steam engine. In this case, Q was the measurable difference between the heat provided by the steam boiler and the heat released to the condenser after performing the work W. Hirn found a ratio equal to 4.05 J/cal. Joule concluded,

> *Any of your readers who are so fortunate as to reside amid the romantic scenery of Wales or Scotland could, I doubt not, confirm my experiments by trying the temperature of the water at the top and at the bottom of a cascade. If my views be correct, a fall of 817 feet will course generate 1 degree of heat, and the temperature of the river Niagara will be raised about one-fifth of a degree by its fall of 160 feet.*

A different experiment can be made by rotating an appropriately shaped iron bar within a strong magnetic field. The electric currents induced within the bar warm the iron by a measurable amount Q. In order to rotate the bar against the mechanical resistance of the magnetic force of Lorentz due to the induced currents a measurable amount of work is needed. In this case, a similar W/Q ratio can be found.

In summary, when a system undergoes a cyclical process in which a measurable amount of work W is converted into a measurable heat Q or vice versa, we obtain a ratio W/Q, which is usually determined as 4.186 J/cal, apart from random experimental errors. This result constitutes the so-called principle of equivalence between heat and mechanical work.

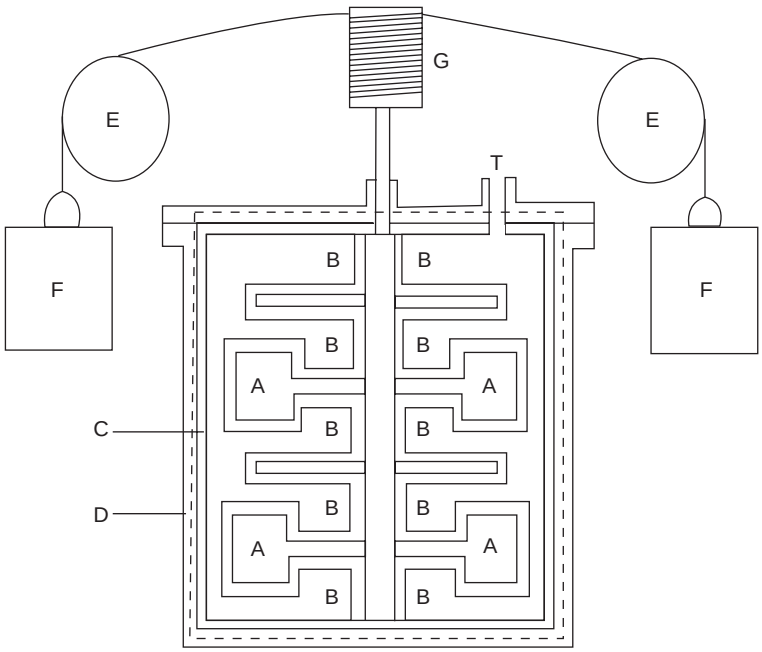


FIGURE 1 Joule's paddle wheel experiment. A, revolving blades; B, stationary vanes; C, metal framework; D, copper jacket; E, pulleys; F, weights; G, drum; T, hole for thermometer. From Beattle, J. A., and Oppenheim, I. (1979). "Principles of Thermodynamics." Elsevier, New York [originally published in *Philos. Trans. R. Soc. London* **140,** 61–82, 1840].

The ratio

$$J = W/Q = 4.186\,\mathrm{J/cal} \qquad (2)$$

is known as the mechanical equivalent of heat.

3. THE ZEROTH LAW OF THERMODYNAMICS

When two bodies are placed in thermal contact, heat flows from one to another in some cases (i.e., from the hotter to the colder one). Heat will keep flowing for awhile, but sooner or later the heat flow will stop and a thermal equilibrium (the absence of heat flows) occurs. We may make an apparently trivial statement about thermal equilibrium:

Zeroth law: If a body A is in thermal equilibrium with a body B and if B is also in thermal equilibrium with a third body C, then A is in thermal equilibrium with C.

This law justifies the concept of temperature. Suppose that B is a thermometer. If B is used to measure the temperature of A and C and if A and C are found to have the same temperature, the heat will not flow if A and C are placed in thermal contact.

4. THE FIRST LAW OF THERMODYNAMICS

The equivalence principle is the basis of the first law of thermodynamics. This law states the general principle of energy conservation when thermal processes are also accounted for. It can be formulated by stating that the energy variation ΔU of a system in a process is equal to the total energy exchange of the system with the surrounding environment. As a consequence, we can also state that the energy of an isolated system remains constant.

We have seen that in a cyclical process the involved W and Q are energetically equivalent by simply taking into account that $1\,\mathrm{cal} = 4.186\,\mathrm{J}$. Hence, we can use the same unit (Joule or calorie) to measure mechanical work and heat. In what follows, we write equalities involving W and Q without specifying the unit. We also consider $W > 0$ if work is done by the system to the environment and $W < 0$ if the opposite is true; similarly, $Q > 0$ if heat is received by the system and $Q < 0$ if heat is released. This is because historically the steam engine received heat (positive) from the environment and performed work (positive) on the environment.

Now consider a system, for example, the liquid in the vessel of Joule's wheel. We can warm the liquid from the temperature T_A to the temperature T_B by

supplying mechanical work ($-W$, according to the previous sign rule), by supplying heat from an external source ($+Q$), or by a combination of the two ($Q_i - W_i$). If we perform experiments in which W and Q assume variable values, expressed by the same unit, we always have

$$-W = Q = Q_i - W_i = \text{constant}, \tag{3}$$

where Q_i and W_i are the quantities involved in the ith experiment. That the same temperature increase can be achieved by supplying work ($-W$), heat (Q), or a combination of both ($Q_i - W_i$) and that this quantity is always the same is a very important result. In fact, it shows that if a system undergoes a process from a state A to a state B, the employed energy (mechanical and/or thermal) is the same independent of the way in which the system goes from A to B. As a consequence, it is possible to associate with each state of a system an energy value U, which is only a function of the state and the variables (pressure, volume, temperature, and concentration) that characterize the state.

This energy of the state or internal energy, U, can be viewed as the sum of the mechanical, electrical, chemical, etc. energies of all the particles that constitute the system and whose amount cannot be quantified in absolute value because it is reference dependent. Since the goal of energy conversion is the work obtainable when the system goes from A to B, absolute values are not of interest to us; instead, we are interested in gradients (i.e., the difference between the value assigned to state A and the value assigned to state B):

$$\Delta U = U_B - U_A. \tag{4}$$

In this case, the internal variation ΔU must equate the exchange of mechanical and/or thermal energy between the system and the environment, $Q - W$:

$$\Delta U = U_B - U_A = Q - W. \tag{5}$$

This is the mathematical statement of the first law of thermodynamics. In words, the energy variation of a system that undergoes a process from the state A to the state B is equal to the energy that the system exchanges with the environment and it is not dependent on the way by which the system goes from A to B. Therefore, given two states A and B, $\Delta U = U_B - U_A$ is the same, whereas Q and W may vary in such a way that $Q - W = \text{constant} = \Delta U$.

It was previously noted that the first law expresses the general principle of conservation of energy when we include thermal energy. For $Q - W = 0$ (cyclical process), we obtain the equivalence principle as a particular case of the first law.

The general principle of conservation of energy also means that a machine cannot do work without any input of heat or other energy. Such a machine, which would create energy from nothing, is called a perpetual motion machine of the first kind. Thus, another equivalent formulation of the first law is the following: It is impossible to construct a perpetual motion machine of the first kind.

If the process is infinitesimal or if the final state B is infinitely close to A, the first law of thermodynamics is expressed in differential form:

$$dU = \delta Q - \delta W, \tag{6}$$

where dU is the energy difference independent of the particular process, and the involved δQ and δW may vary with the process leading from A to B. It should be observed that dU, unlike δQ and δW, is an exact differential because it is not path dependent. Hereafter exact differentials are indicated with d and path differentials with δ.

An important and frequently met system is constituted by a fluid or gas (hereafter an ideal gas, for the sake of simplicity) contained in a cylinder of section S, within which a frictionless piston can run (Fig. 2). In this case, the external force F acting on the fluid (also including the weight of the piston, if appropriate) is, by definition of pressure,

$$F = p \times S, \tag{7}$$

where p is the external pressure. Hence, the work performed by gas or by the environment ($ds > 0$ or $ds < 0$) is calculated as the product of the force F multiplied by the infinitesimal displacement, ds,

$$\delta W = p \times S ds \tag{8}$$

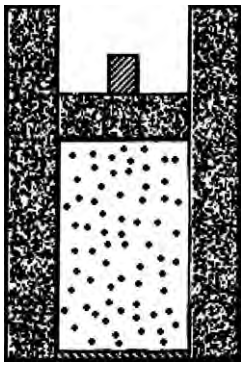


FIGURE 2 A thermodynamic fluid within a frictionless cylinder–piston system.

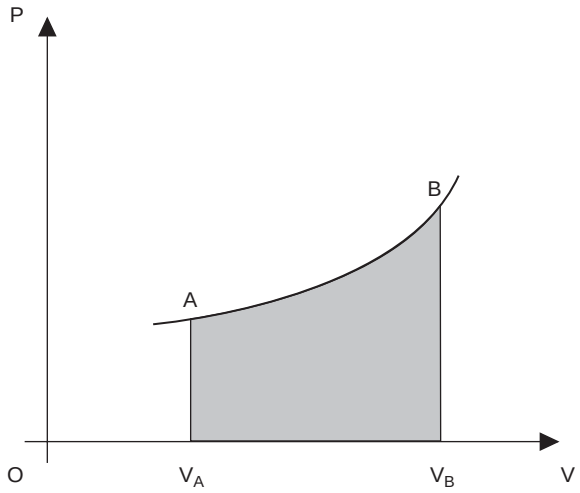


FIGURE 3 Work performed by a fluid expanding against an external force *F*. Work is indicated by the shaded area.

or as $\delta W = p \times dV$, where $S \times ds$ is the infinitesimal variation of volume dV. For a finite variation, the work is calculated as an integral between the two states

$$W = \int_{V_A}^{V_B} p \times dV \qquad (9)$$

and clearly depends on the particular process leading from A to B. This is represented by the shaded area in Fig. 3.

5. THE SECOND LAW OF THERMODYNAMICS

The equivalence principle and the first law do not deal with the problem of the possibility of transforming heat into mechanical work or vice versa. The equivalence principle only states that when this transformation takes place, the W/Q ratio must be 4.186 J/cal. In other words, on the basis of these statements only, thermal and mechanical energies should be perfectly equivalent not only from the quantitative standpoint but also from the qualitative one.

Experiments show that although it is always possible to completely transform mechanical energy into thermal energy (e.g., by friction), the opposite is not true. The transition from mechanical work to heat occurs without any restriction, but a heavy constraint is imposed on the opposite transition, namely

1. Not all the heat that a system receives from a reservoir is converted into mechanical work, but a fraction remains as heat at lower temperature.

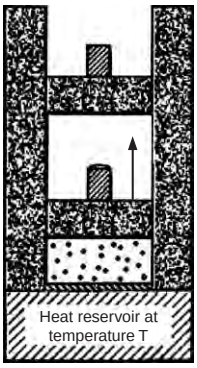


FIGURE 4 A fluid expansion within a cylinder–piston system, from volume V_1 to volume V_2. Due to the finite length of the cylinder, the fluid cannot expand forever but must be compressed for the cycle to start again.

2. If all heat were converted into mechanical work, the latter would not be the unique final result, in such a way that this work would be unusable.

The fluid contained in a frictionless cylinder–piston system connected to a heat reservoir at temperature T is an appropriate example of statement 2 (Fig. 4). If we warm the fluid and make it expand, the heat is converted into work but this is not the unique result. In fact, increased volume of the fluid is also a result. Unless the cylinder is infinitely long, we have to compress the fluid again for the system to be usable. In so doing, the work performed on the system would be greater than the work performed by the system.

On the basis of these considerations, Carnot proposed an ideal engine connected to two heat reservoirs at temperatures T_1 and T_2 ($T_1 > T_2$). The reservoir at a lower temperature serves to cool the fluid in such a way that it can return to its initial volume. Carnot showed that not all the heat can be converted into mechanical work, but a fraction is always released in the form of degraded heat at lower temperature. This result is an example of statement 1 given previously.

5.1 Reversible Transformations and Machines

It would be a mistake to think that the limits imposed by the conversion of heat to mechanical work are of a technical nature (i.e., are due to the unavailability of tools or cylinders and pistons without friction or perfectly sealed). In order to eliminate this possible misunderstanding, in the rest of this article we refer to reversible transformations and reversible machines. In these ideal transformations and machines,

it is assumed that all technical energy losses (friction, inadequate materials, etc.) have been removed. Even in this case, the impossibility of integral conversion of heat to work as a unique result is demonstrated. This means that this impossibility relies on the theory; it is a constraint imposed by Nature and not by the use of technically poor instruments.

Reversible transformations and machines were introduced to help describe real (irreversible) transformations and machines because the latter cannot be easily investigated. In fact, for the expansion or compression of a real fluid within a real cylinder–piston system, we are hardly able to assess the value of volume, pressure, and temperature in any instant of the process. Moreover, the cylinder wall will never be perfectly thermally insulated, nor will we know the exact value of its thermal conductivity coefficient. The reversibility assumption implies that the system (initially in A) can reach the final state B by means of infinitely slow transformations, in which infinitely small forces act on the fluid. In so doing, it is in principle possible to determine the exact value of its state variables in every instant. Moreover, we assume that (i) the piston runs within the cylinder with no friction and is perfectly sealed, (ii) the side walls of the cylinder are perfectly heat insulated, and (iii) the cylinder base is a perfect heat conductor. In other words, machines and containers used in a reversible process are absolutely ideal ones, as is the process. The fluid transformation from A to B can therefore be thought of as a succession of equilibrium states so that the direction of the process can be inverted in any instant by means of an infinitely small force applied against it. This is why such a transformation is called reversible and is indeed a powerful tool to investigate the relation among the parameters that describe the initial and final states of the fluid. It is clear that a reversible transformation can be exactly described, for instance, in a (V, P) plane by drawing the curve $P = P(V)$ as the succession of all the equilibrium states existing between A and B (Fig. 5).

On the other hand, it would be impossible to draw a curve describing a real transformation between the same states because of the existing indeterminacy of the actual values of pressure and volume in any given instant of the process. In such a case, the curve is not exactly defined (Fig. 6).

Reversible transformations are sometimes considered the final limit approached by the behavior of real transformations, when the latter occur so slowly that the time needed is a good approximation of the theoretically infinite time required by the corresponding reversible transformation.

Finally, it is worth noting that reversible transformations cannot occur in isolated systems (i.e., in thermodynamic systems that do not exchange any energy or matter flow with the outside) because the existing gradient that pushes the transformation toward one direction cannot be inverted unless an external force is available. In the latter case, the system would no longer be isolated.

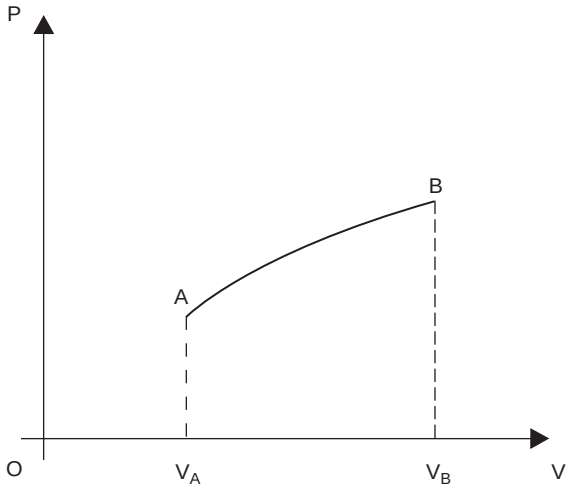


FIGURE 5 A reversible transformation is exactly described in a (V, P) plane by the curve $P = P(V)$ drawn as the succession of all the equilibrium states existing between *A* and *B*.

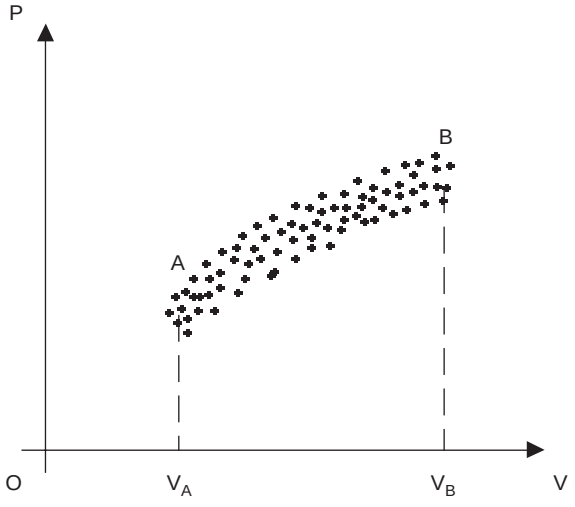


FIGURE 6 Irreversible transformations cannot be described as a succession of equilibrium states due to the existing indeterminacy about the actual values of pressure and volume in any given instant of the process. In such a case, the curve is not exactly defined.

5.2 Lord Kelvin's and Clausius's Statements of the Second Law

The observed lack of symmetry of work-to-heat transformations and vice versa is equivalent to a substantial irreversibility of real processes that cannot therefore be of a purely mechanic nature. This is the basis of the second law of thermodynamics, which can be exemplified by means of typically irreversible phenomena, such as a gas expansion (the gas will never spontaneously return to its initial volume) or the heat transfer from a body at higher temperature to a body at lower temperature (the heat will never flow spontaneously from the colder to the warmer body).

Lord Kelvin ("Dynamic Theory of Heat," 1850) pointed out the impossibility of an integral conversion of heat to work (in which work was the unique result) as the appropriate statement of the second law of thermodynamics. The following is Lord Kelvin's statement: A transformation the unique result of which is the conversion of heat from a heat source at a temperature T to mechanical (or electromagnetic) energy is impossible. The heat source at temperature T is defined as a body every point of which is at a temperature T and whose thermal capacity is so high that it can absorb heat from or release heat to other bodies without undergoing any temperature change. It is, again, an ideal object, whose approximation in real processes can be a very large boiler.

The most important part of Lord Kelvin's statement, as noted in Section 5.1, is the requirement of work as the unique result. The fluid described in that section undergoes a reversible and isothermal ($\Delta T = 0$) expansion. Its internal energy does not change because the internal energy of ideal fluids only depends on the temperature ($\Delta U = 0$). As a consequence of the first law of thermodynamics, we have $W = Q$. As already noted, the achieved conversion of heat to work is not the unique result because the volume increase makes this (although ideal) machine unable to actually deliver work. In this regard, it must be observed that the first law precludes the possibility of creating energy (perpetual motion of the first kind), while Lord Kelvin's statement of the second law precludes the possibility of the so-called perpetual motion of the second kind: It is impossible to construct a perpetual motion machine of the second kind. A perpetual motion machine of the second kind would convert heat taken from some reservoir into an equivalent amount of work. It does not violate the first law because the work energy is equivalent to the heat energy. It is this impossibility that precludes the exploitation of enormous heat sources such as the oceans to generate mechanical energy. In fact, should a ship be able to convert the ocean's heat into work for its own motion as the unique result, thus skipping Lord Kelvin's statement, this work would again be converted into heat due to the friction between the ship and the water. In so doing, the thermodynamic cycle would be closed and the ship would travel practically at no energy cost (i.e., the perpetual motion of the second kind).

The impossibility not just to make real machines, but also to even design ideal machines capable of completely converting a source heat into mechanical work as a unique result suggested to Lord Kelvin his second law statement. If only one experiment, real or even ideal, could be performed against Kelvin's statement, the second law would completely lose its significance and applicability.

Another typically irreversible phenomenon could be used to derive an alternative statement of the second law, namely the heat transfer from a body A at higher temperature T_A to a body B at lower temperature T_B. Clausius used this phenomenon to illustrate his statement of the second law while at the same time developing a mathematical formulation of it and introducing the entropy concept. For this to be correctly enunciated and understood, we must first note that an empirical temperature scale can be defined based on the direction of heat exchanges between two bodies when they are put in contact with each other. Independently from their heat capacity, body A is defined as having a higher temperature than B if, when they come into contact, some heat flows from A to B. This statement is equivalent to saying that heat flows spontaneously from a body at higher temperature to a body at lower temperature until a thermal equilibrium (same temperature) is reached. A clear example can be given by considering a small piece of iron heated to "white heat" temperature (approximately 1000°C) and a vessel containing a large amount of boiling water (100°C). Although the former contains a much smaller amount of heat than the latter, its temperature is higher. Therefore, if we dip the iron into the boiling water some heat will flow from the iron to the water (i.e., from the body at higher to the body at lower temperature) until an equilibrium is reached and the iron recovers its original gray color and solid state (this is what blacksmiths do to forge products, although they use colder water for faster results).

Therefore, the following is Clausius's statement of the second law: If heat spontaneously flows from a

body A to a body B (i.e., if A has a higher temperature than B), it is impossible to perform a transformation process the unique final result of which is a flow of heat from B to A (i.e., from the body at lower temperature to the body at higher temperature).

This statement was also suggested by the failure of all real and ideal processes to achieve the unique result of heating warmer bodies by means of colder bodies. Again, if only one experiment could be found against Clausius's statement, the second law would lose its validity. It is trivial to observe that a refrigerator does not contravene Clausius's statement. In fact, although heat flows from inside the refrigerator to the outside environment (the latter at higher temperature), this occurs at the expense of electric work and therefore is not the unique final result.

Kelvin's and Clausius's statements are equivalent, as expected. If the first one is false, the second one is also false and vice versa. Assume Kelvin's statement is false and that heat from a source at temperature T_1 can be entirely converted into work as the unique result of a process. We could therefore, by carefully designing the process, convert again this work into heat at any temperature, even higher than T_1. If so, we would obtain a process the unique result of which would be a flow of heat from a source at temperature T_1 to a body at temperature T_2, higher than T_1. As a consequence, Clausius's statement would also be false. The opposite case (if Clausius's statement is false, then Kelvin's statement is also false) will be demonstrated after discussing Carnot's cycle and theorem.

5.3 Carnot's Cycle

In his seminal "Réflexions" (1824), Carnot dealt with the problem of heat-to-work conversion by means of ideal machines in which a thermodynamic fluid was capable of performing reversible transformations, receiving heat from a source at higher temperature, and releasing heat to a sink at lower temperature. He demonstrated that the heat received from the source cannot be entirely converted into work and that the nonconverted fraction is released as heat at lower temperature, thus being depreciated from the standpoint of its ability to perform work. This conclusion translates into the fact that the efficiency of such an ideal machine is much lower that 100% (i.e., the work produced is much less than the heat supplied).

Carnot's findings constitute an early statement of the second law of thermodynamics, for a long time not completely understood in its full significance. It can be demonstrated that it is completely equivalent to those provided by Kelvin and Clausius approximately 30 years later. It can also be demonstrated that the efficiency of any real machine is lower than that of the ideal Carnot's machine.

For this purpose, assume we have an ideal Carnot's machine composed of an ideal thermodynamic fluid within a cylinder–piston system, without friction, and with the lateral walls being perfectly insulated and cylinder base being a perfect heat conductor. Figure 7A shows the different positions of the piston within the cylinder over time, whereas Fig. 7B diagrams the isothermal and adiabatic reversible patterns that the fluid must follow during the process in order to close the cycle (Carnot's cycle).

In Fig. 7A, a heat source (a boiler) at temperature T_1 supplies heat Q_1 to the fluid through the base of the cylinder. Since the piston is free to run through the cylinder, the fluid undergoes a reversible and isothermal expansion from the initial volume V_1 to a volume V_2, according to the equation $PV = nRT$ (the so-called ideal gas equation of state). A thermal insulation is then provided to the base of the cylinder and the fluid is allowed to further expand reversibly and adiabatically (no heat exchange with the outside environment) to a new volume V_3, according to the equation $PV^{\gamma} = K$ (applicable to adiabatic processes; $\gamma = C_p/C_v$, where C_p and C_v are respectively the heat capacities at constant pressure and volume, and the constant K depends on initial conditions). In this step, the fluid temperature decreases from T_1 to T_2. Now we remove the thermal insulation of the base and put the cylinder on a heat sink (a body at lower temperature than that of the heat source) at temperature T_2. If we compress the fluid (i.e., some work is performed on it), its volume V_3 will decrease reversibly and isothermally to the volume V_4, releasing a heat Q_2 to the heat sink. Finally, if we insulate the base again and compress the fluid reversibly and adiabatically until it reaches the initial volume V_1, pressure P_1, and temperature T_1, the cycle is closed and the initial state is restored. From an energy standpoint, it is $\Delta U = 0$ because the system is again in its initial state. The heat supplied to the system is Q_1, whereas the work performed over the whole cycle is

$$W = Q_1 - Q_2 \tag{10}$$

according to the first law of thermodynamics. The heat Q_2 released to the heat sink cannot be transferred again to the boiler because the latter has a temperature higher than that of the heat sink. Therefore, Q_2 is definitely unusable. The machine

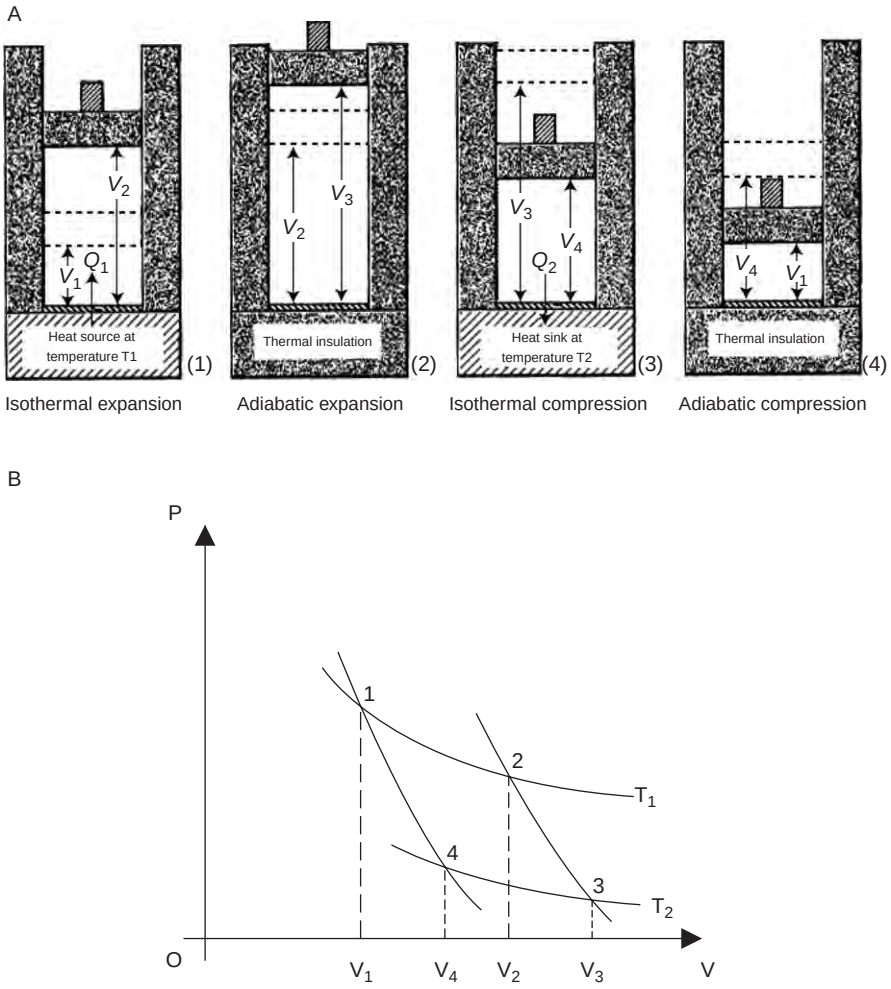


FIGURE 7 Carnot's cycle. (A) The different positions of the piston and the volumes of the fluid over time. (B) The isothermal and adiabatic reversible patterns that the fluid must follow in the process in order to close the cycle.

efficiency (work performed/heat supplied) is therefore

$$\eta = W/Q_1 = (Q_1 - Q_2)/Q_1 = 1 - Q_2/Q_1 < 100\% \tag{11}$$

This result may be assumed to be Carnot's statement of the second law, equivalent to those of Kelvin and Clausius. In fact, if the heat Q_2 could be transferred from the sink to the source without an additional energy expense, the efficiency η would be 100%, but this transfer from a colder to a warmer body would be against Clausius's statement. In addition, the immediate consequence would be the complete conversion of the heat Q_1 into work, against Kelvin's statement. The already shown equivalence between Kelvin's and Clausius's statements is also further demonstrated.

5.4 Carnot's Theorem

A fundamental theorem named after Carnot follows from the Carnot's cycle. This theorem permits the introduction of the so-called absolute temperature T and the entropy concept, thus offering a way to determine the direction of real transformation processes. This theorem can be stated as follows: The efficiency of a reversible thermal machine R operating between two temperatures is never lower than that of a whatever machine S operating between the same temperatures:

$$\eta_R \geq \eta_S, \tag{12}$$

where the equality only applies if S is also reversible.

Assume that we have two machines (the first reversible and the second nonreversible) as in Fig. 8,

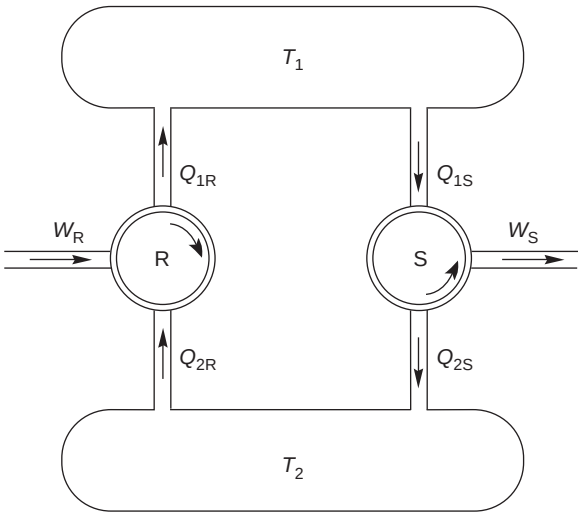


FIGURE 8 A reversible Carnot machine R working backwards between two heat reservoirs at the expense of an amount of work W_R provided from outside. R delivers heat Q_R to the reservoir at temperature T_1. This heat is transferred to the irreversible machine S, which converts it into work W_S. If the efficiency of the irreversible machine were higher than that of the reversible one, W_S would be higher than W_R, in contrast to Kelvin's statement of the second law of thermodynamics.

both operating between the same temperatures T_1 (source) and T_2 (sink). Each is able to receive a given amount of heat from the source at temperature T_1 and to deliver a given amount of work by also releasing some waste heat to the sink at lower temperature.

The theorem can be demonstrated ab absurdo assuming that, instead, it is $\eta_S > \eta_R$; that is,

$$\eta_S = (Q_{1S} - Q_{2S})/Q_{1S} > \eta_R = (Q_{1R} - Q_{2R})/Q_{1R} \tag{13}$$

Since R is reversible, it could also work backwards as shown in Fig. 8; that is, it could receive an amount of work $W_R = Q_{1R} - Q_{2R}$ from outside and release heat Q_{1R} to the source at temperature T_1. If the input work W_R is such that $Q_{1R} = Q_{1S}$, the source of heat at higher temperature T_1 becomes unimportant. It follows from Eq.(13) that

$$W_S = (Q_{1S} - Q_{2S}) > (Q_{1R} - Q_{2R}) = W_R$$

and finally that

$$W_S - W_R > 0,$$

thus creating work by extracting heat $(Q_{2R} - Q_{2S})$ from a unique source at temperature T_2. This is in clear contrast with Kelvin's statement. It must therefore be $\eta_S < \eta_R$. If S were also reversible, the second part of the theorem, $\eta_S = \eta_R$, could easily be demonstrated in a similar way.

Consequently, the efficiency of any reversible machine depends only on the two temperatures T_1 and T_2 between which the machine operates. After adequately choosing the thermometric scale in order to have $T = t + 273.14$ (t is °C), it can be demonstrated that

$$Q_2/Q_1 = T_2/T_1. \tag{14}$$

The temperature T is called absolute temperature because it does not depend on the particular thermometric substance used and therefore coincides with the one obtained by using an ideal fluid as the thermometric substance. A unit "absolute" degree (K) coincides with a unit centigrade degree. The latter is defined by dividing by 100 the temperature interval between the temperature of melting ice (0°C) and the temperature of boiling water (100°C) at sea level.

In summary, the efficiency of a reversible Carnot's cycle depends only on the absolute temperatures of the heat source and the heat sink and not on the fluid or the amount of heat exchanged:

$$\eta_R = (T_1 - T_2)/T_1. \tag{15}$$

Equation (14) can be rewritten as

$$Q_1/T_1 - Q_2/T_2 = 0, \tag{16}$$

and generalized to the case of n sources in a reversible cycle as

$$\sum Q_i/T_i = 0, \tag{17}$$

with positive sign for heat absorbed by the fluid and negative sign for heat released. Equation (17) can be further generalized to the case of infinitesimal amounts of heat δQ exchanged by each source in a reversible process with continuous temperature changes:

$$\int_{\text{cycle}} \delta Q/T = 0. \tag{18a}$$

For an irreversible process S, it would be

$$\int_{\text{cycle}} \delta Q/T < 0 \tag{18b}$$

because $\eta_S < \eta_R$.

5.5 The Entropy Concept

Now consider the most general case of a reversible cycle, such as the one shown in Fig. 9, with continuous temperature changes. It can always be

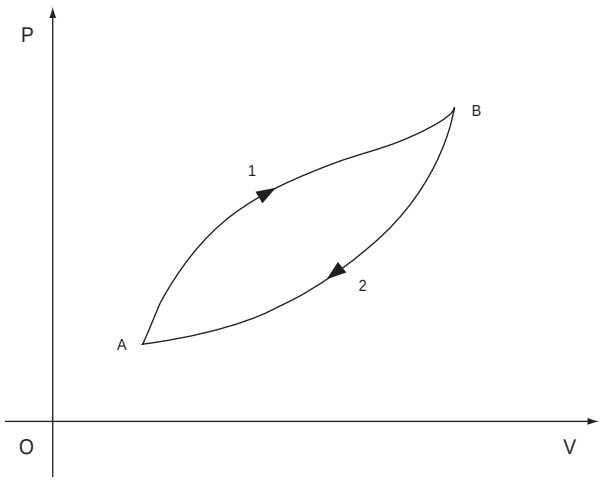


FIGURE 9 Diagram of a reversible cycle, with continuous temperature changes, simplified as the succession of two reversible transformations from *A* to *B* (along pathway 1) and vice versa (along pathway 2).

simplified as the succession of two reversible transformations from A to B (along pathway 1) and vice versa (along pathway 2).

According to Eq. (18a), we may write

$$\int_{A\to B,(1)} \delta Q/T + \int_{B\to A,(2)} \delta Q/T = 0, \qquad (19)$$

where the subscripts indicate the paths from A to B and vice versa. Due to the reversibility of the two transformations $A\to B$ and $B\to A$, we may write

$$\int_{A\to B,(1)} \delta Q/T = -\int_{B\to A,(2)} \delta Q/T = \int_{A\to B,(2)} \delta Q/T. \qquad (20)$$

This result is of paramount importance because it demonstrates that the integral does not depend on the specific reversible transformation from A to B but only on the actual states A and B. We are therefore able to define a new function depending only on the system state (such as the internal energy *U*). Clausius called this function entropy, defined as follows: Given an arbitrary initial system state O, an arbitrary final state A, and an arbitrary reversible transformation from O to A, the value of the integral

$$S(A) = \int_{O\to A} \delta Q/T + S(O) \qquad (21)$$

calculated over the reversible transformation $O\to A$ is a function only of the states O and A. This value is called entropy.

Entropy is measured in calories per degree Kelvin or Joules per Kelvin. It clearly appears that $S(A)$ is defined but for an arbitrary constant $S(O)$, which has no practical utility because we are always interested in entropy differences between states and not in absolute values. We can therefore equate this constant to zero.

It can also be demonstrated that

If an isolated system undergoes a transformation from an initial state A to a final state B, the entropy of the final state is never lower than the entropy of the initial state.

This statement is the so-called entropy law.

6. THE THIRD LAW OF THERMODYNAMICS (NERST'S THEOREM)

Sometimes, the arbitrary constant $S(O)$ of Eq. (21) becomes important and cannot be neglected. In 1906, Nerst suggested the statement of a third law of thermodynamics as well as a specific value for the arbitrary constant $S(O)$. The statement (the so-called Nerst's theorem) can be expressed as follows: The entropy of any thermodynamic system at the absolute temperature of 0 K is equal to zero, whatever is the state of the system.

The most important consequence of the third law and the Nerst's theorem is the impossibility for a system to reach the absolute zero ($T = 0$ K). In fact, it can be demonstrated that as the system approaches the absolute zero it also becomes more difficult to further decrease its temperature and that this difficulty increases faster than the actual temperature decrease. Therefore, the absolute zero cannot be reached. A surprising and unexpected result follows: The absolute scale of temperature must be considered unlimited in both directions, either toward the high temperatures and toward the absolute zero.

The impossibility of reaching the absolute zero is the basis of another perpetual motion statement: It is impossible to construct a perpetual motion machine of the third kind. A perpetual motion machine of the third kind continuously converts heat into work by operating between a hot reservoir at temperature $T > 0$ and a cold reservoir at $T = 0$. The third law states that it is impossible to achieve absolute zero temperature, thus making such a machine impossible.

Another consequence of the third law is that the specific heat $c(T)$ of a solid must tend to zero for $T \to 0$, at least according to the first power of *T*.

Otherwise, being $\delta Q = mc(T)dT$, the expression

$$S(\mathrm{A}) = \int_{\mathrm{O}\to\mathrm{A}} \delta Q/T = \int_{\mathrm{O}\to\mathrm{A}} mc(T)\, dT/T \qquad (22)$$

would diverge toward ∞ for $T \to 0$ instead of converging to a specific $S(\mathrm{A})$ value.

7. HOW MANY LAWS FOR THERMODYNAMICS?

Thermodynamics originated within the field of energy conversion and use, but its generality and the fact that it seems to apply to almost every phenomena in nature as well as in human-dominated systems generated major interest among nonspecialists. Many tried to establish links or derive thermodynamic principles capable of describing some aspects of their specific field of interest. In addition, several thermodynamics experts have also found interesting new properties or special aspects when applying thermodynamic principles to living systems. Both kinds of interest generated a multiplicity of tentative statements for new laws of thermodynamics and beyond. Some of these focused explicitly on specific aspects already covered by the second law, whereas others dealt with the redefinition of the concept of entropy in order to make it applicable to irreversible processes or with the macroscopic behavior of self-organizing systems in the biosphere seeking a general description of ecological patterns.

7.1 Georgescu-Roegen's Entropy of Matter

The economist Nicholas Georgescu-Roegen proposed a fourth law of thermodynamics concerned with the so-called entropy of matter. Georgescu-Roegen proposed that in a closed system (i.e., a system exchanging energy but not matter with the outside environment), a "material entropy" occurs and gradually reaches its maximum value (maximum disorder and mixture of matter) in such a manner that "all matter ultimately becomes unavailable." He identified the biosphere as a closed system (exchanges with outside space are too slow and therefore not significant on the scale of the economy) and pointed out that it would be impossible to completely recover the matter involved in the production of mechanical work or wasted in friction. The disappearance of any qualitative difference between materials is a sort of "material death" of the system, which is similar to the famous "thermal death" dictated by the entropy law for isolated systems. Georgescu-Roegen emphasized the importance of the behavior of matter in opposition to the so-called energy dogma for which any recovery of high-quality pure substance is possible (complete recycling) when a sufficient amount of energy and time are available. Bianciardi *et al*. demonstrated that apart from the importance of this statement from the standpoint of economic processes or even the ethics of resource use, it is false in the field of physical laws where the author intended it to stand due to its severe inconsistency with the second law. They argued that if the fourth law were true, then the second law and its entropic consequences must be false. Photosynthesis, the generation of organized from dispersed matter driven by solar energy, provides proof of this issue. Georgescu-Roegen's statement is important in the context of the second law without introducing a fourth law. In fact, to recover the tires or the coins worn away by friction requires that an ordered form of energy be invested in the process to decrease the entropy of this matter. As a consequence, the entropy of the surrounding environment (the rest of the biosphere) would increase to a larger extent, generating a net entropy increase that would not be sustainable for the biosphere.

7.2 Onsager's Reciprocal Relations

Early attempts to treat irreversible processes such as heat conduction by means of classical thermodynamics resulted in little success. Despite its name, classical thermodynamics was not suited to the treatment of dynamic processes. It is instead a perfect tool for the study of static states and chemical equilibria. In 1968, the Norwegian physical chemist Lars Onsager was awarded the Nobel Prize in chemistry for the discovery of the reciprocal relations named after him and basic to irreversible thermodynamics. Onsager presented his fundamental discovery at a Scandinavian scientific meeting in Copenhagen in 1929. The work was published in its final form in 1931 in the well-known journal *Physical Review* in two parts with the title "Reciprocal Relations in Irreversible Processes." The importance of these relations was not immediately recognized by the scientific community; thus, the Nobel Prize was attributed to Onsager only many years later. When giving his Nobel presentation speech, S. Claesson of the Royal Swedish Academy of Sciences emphasized that "the Onsager's reciprocal relations represent a further law making possible

a thermodynamic study of irreversible processes." He also pointed out that "the size of the two papers was no more than 22 and 15 pages, respectively. Judged from the number of pages this work is thus one of the smallest ever to be rewarded with a Nobel Prize."

7.3 Prigogine's Excess Entropy Production

The Onsager's reciprocal relations later became the basis of Prigogine's irreversible thermodynamics, for which Prigogine was awarded the Nobel Prize in chemistry in 1977. P. T. Landsberg suggested the status of a fourth law for the results of Prigogine and collaborators, who investigated the dissipative behavior of open systems moved away from equilibrium by the fluxes of matter and energy across their boundary. They showed that open dissipative structures can maintain themselves for a period of time away from thermodynamic equilibrium in locally reduced entropy steady state, and they developed a fluctuation and stability theory fully relying on the assumptions of a "local equilibrium." This is the so-called Landsberg "basic trick" of thermodynamics, which enables a system (which may have gradients of temperature, etc. built into it) to be considered as composed of elements to each of which classical thermodynamics applies. In so doing, each element is large enough for thermodynamic consideration to apply but small enough for the gradients to be unimportant within each element. The additional assumption that the summation of corresponding local extensive variables over the whole system (near the reference state) gives the value of the corresponding global variable enabled Prigogine and colleagues to define a nonequilibrium entropy concept on which their whole stability and fluctuation theory is based. The main result of this approach was the definition of the concept of excess entropy production, from which global stability conditions could be derived for a thermodynamic system:

$$\delta^2 S \leq 0 \quad \text{and} \quad \frac{1}{2}\frac{\partial}{\partial t}(\delta^2 S) \geq 0.$$

where $\delta^2 S$ is the second order perturbation of the S function around the equilibrium state and $\partial/\partial t$ is the partial derivative with respect to time, called "excess entropy production". It should be noted that the symbol δ has here a different meaning than in the previous sections.

Landsberg pointed out that without this assumption, Prigogine's theories of fluctuation and stability "would not be thermodynamic in character." However, he also recognized that the assumption worked sufficiently "in that the results obtained from it cohere with other theories and experiments" and suggested the definition of nonequilibrium thermodynamics to qualify as a fourth law of thermodynamics. Instead, according to Schneider and Kay, "Prigogine's linear expansion of the entropy function about equilibrium … is a severe restriction on the application of his theory and in particular precludes its formal application to living systems." These considerations led Schneider and Kay to formulate a restated second law of thermodynamics, based on previous extended formulations by Hatsopoulos and Keenan and by Kestin and analogous to Le Chatelier's principle for chemical systems.

7.4 Lotka's Maximum Power Principle

A different set of statements came from scientists working in the field of ecology, willing to explore and understand the energy and resource basis of living systems' self-organization. A fundamental contribution was provided by the mathematical biologist Alfred Lotka, who stated the so-called maximum power principle:

> *In the struggle for existence, the advantage must go to those organisms whose energy-capturing devices are most efficient in directing available energy into channels favorable to the preservation of the species. … In every instance considered, natural selection will so operate as to increase the total mass of the organic system, to increase the rate of circulation of matter through the system, and to increase the total energy flux through the system, so long as there is presented an unutilized residue of matter and available energy. This may be expressed by saying that natural selection tends to make this energy flux through the system a maximum, so far as compatible with the constraints to which the system is subject. … It appears, however, at least a priori probable that among the certainly very large (if not infinite) variety of types presented for selection, sooner or later those will occur which give the opportunity for selection to operate in the direction indicated, namely so as to increase the total mass of the system, the rate of circulation of mass through the system, and the total energy flux through the system. If this condition is satisfied, the law of selection becomes also the law of evolution.*

Systems that self-organize to increase the throughput flow of energy are rewarded by survival and selection. The principle of selection "functions … as a third law of thermodynamics (or a fourth, if the third place is given to the Nerst principle)." This is a major advancement with respect to classical thermodynamics. According to Lotka,

> *The principle of natural selection reveals itself as capable of yielding information which the first and second laws are not competent to furnish. The two fundamental laws of*

thermodynamics are, of course, insufficient to determine the course of events in a physical system. They tell us that certain things cannot happen, but they do not tell us what does happen.

7.5 Odum's Maximum Empower Principle

In 1987, Howard T. Odum was awarded (together with his brother Eugene) the Crafoord Prize for ecology by the Royal Swedish Academy of Sciences. In his acceptance speech, Odum pointed out that

The study of ecosystems suggests principles by which energy flows generate hierarchies in all systems. From these it was clear that energy laws controlling self-organization were principles of open systems quite beyond classical energetics, involving a generalization of concepts of evolution in energy terms. During the trials and errors of self-organization, species and relationships were being selectively reinforced by more energy becoming available to designs that feed products back into increased production.

In order to account for resource convergence into self-organization processes, Odum introduced the concept of emergy, defined as "the total amount of available energy of one kind (usually solar) that is directly or indirectly required to make a given product or to support a given flow." The total emergy directly supporting many items (or grams or joules) at the lowest level of a hierarchical chain also supports indirectly a smaller amount of items (grams or joules) at the next higher levels. Therefore, where several levels of energy hierarchy are involved, maximizing power according to Lotka can be misleading because it might imply that priority is given to increasing energy flow through the lower levels, which may have more energy throughput. Instead, expressing resources in emergy terms permits the restatement of Lotka's maximum power principle as a maximum empower principle, according to which the simultaneous maximization of empower (emergy throughput flow) at all system's levels is required if a self-organizing system or ecosystem is to be stable and prosperous. The following is Odum's revised statement of Lotka's fourth law: "On each scale, systems designs prevail that maximize empower, first by reinforcing inflows of available energy and second by increasing the efficiency of useful work."

In his accounting for energy-based hierarchical patterns of self-organization, Odum pointed out that evolution is driven by the emergence of quality. The systems that are able to maximize production by reinforcing their resource basis as well as by increasing their resource use efficiency (i.e., those systems capable of generating new patterns and tools for resource use—organization, information, and feedback control) are rewarded by a higher probability of being selected in the natural competition. This creative ability makes the quality of a process product generally much higher than the quality of process inputs. A mathematical formulation of the maximum empower principle under the "light" of quality has been given by Giannantoni by means of driving forces (so-called source terms, e.g., information, DNA, and feedback control from the next larger scale) supporting the generative processes as well as by means of a new kind of fractional differential calculus. The latter is based on "binary" functions capable of accounting for the nonconservative aspects of a generative process (emergy of coproducts) and the emergence of quality. It obtains the three traditional principles of thermodynamics as special cases of the maximum empower principle by simply assuming that the source terms are equal to zero (i.e., by focusing on the quantitative aspects only). The maximum empower principle becomes, in this perspective, an extension of classical thermodynamics to ecological systems.

7.6 Jorgensen's Ecological Law of Thermodynamics

Exergy, the free energy of a system relative to the surrounding environment, measures the amount of work that can be extracted from the system when it is allowed to reach the thermodynamic equilibrium with the environment. According to Szargut *et al.*, exergy is "the amount of work obtainable when some matter is brought to a state of thermodynamic equilibrium with the common components of the natural surroundings by means of reversible processes, involving interaction only with the above mentioned components of nature." Whenever a work is performed, an existing gradient must be dissipated. Exergy, the ability to do work, is strictly correlated with the existence of a gradient of some physical property. This definition suggests exergy as a measure of the information content of a system as well as of its distance from equilibrium. On this basis, Jorgensen enunciated a tentative fourth law, also called the ecological law of thermodynamics (ELT):

If a system receives a through-flow of exergy, it will utilize this exergy to move away from thermodynamic equilibrium. If the system is offered more than one pathway to move away from thermodynamic equilibrium, the one yielding most stored exergy, i.e., with the most organized

TABLE I

Comparison among Alternative Statements of the Laws of Thermodynamics

Law	Mathematical statement[a]	Traditional statement[a]	Perpetual motion statement[a]	Landsberg (1972) statement	Las Vegas statement[b]
Zeroth law		If a body A is in thermal equilibrium with a body B and if B is also in thermal equilibrium with a third body C, then A is in thermal equilibrium with C.		Empirical temperature exists.	
First law	$\Delta U = Q + W$	The total energy is conserved.	It is impossible to construct a perpetual motion machine of the first kind.	Internal energy exists.	You cannot win.
Second law	$dS = \delta q_{rev}/T$, $dS_{univ} > 0$	The total entropy increases.	It is impossible to construct a perpetual motion machine of the second kind.	Entropy and absolute temperature exist.	You cannot break even.
Third law	$T > 0$	The absolute temperature of any system can never be zero.	It is impossible to construct a perpetual motion machine of the third kind.	States with $T = 0$ do not exist.	You cannot get out of the game.

[a] This article.
[b] Ronald D. Poshusta, www.poshusta.chem.wsu.edu.

structure, or the longest distance to thermodynamic equilibrium under the prevailing conditions, will have the propensity to be selected.

The strong agreement of the ELT with Darwinian theories of evolution convinced Jorgensen that exergy is a convenient "indicator of the development of ecosystem structure and changes in species composition. Structural exergy, in particular, should be considered an appropriate measure of ecological integrity." He also noted that his tentative fourth law, like the other laws of thermodynamics, cannot be proved by means of deductive methods and can at best only be proved by inductive methods (i.e., by the results of as many cases as possible). He investigated and published a huge number of ecological systems cases, providing empirical data and simulation models in support to his statement. However, he had to admit that "it is hardly possible to make deterministic statements about the development of an ecosystem" and that "it is absolutely necessary to test the tentative law with many more case studies before it can be recommended to use it more generally."

8. CONCLUSION

We have discussed several different statements of the laws of thermodynamics in this article. Some of them are shown in Table I. Looking at the table, it can be clearly seen how many different aspects of the same law can be pointed out and how these aspects are linked to each other.

Instead, the tentative fourth laws presented in this article are not summarized in Table I due to a lack of consensus on their general validity. However, they are very interesting attempts to describe and quantify the growth of complexity, the storage of information, the distance from thermodynamic equilibrium of systems at different hierarchical levels, and the strategies of ecosystem development. In other words, all of them stem from the need to explain why self-organization processes occurring in open systems seem to violate the second law dictate and how they instead generate low-entropy complex structures. The equilibrium thermodynamics was more easily described in terms of formal and mathematical statements than nonequilibrium thermodynamics due to the nondeterministic nature of self-organizing

complex systems. Simplifying the model in order to give a formal statement of self-organization requires assumptions that limit the applicability of the statement to real systems and do not capture their complexity. The tentative fourth laws share the goal of identifying clear links between thermodynamics, ecology, and evolution.

Born as an applied science aimed at usefully converting heat into mechanical work, thermodynamics rapidly developed into a more general system of knowledge encompassing almost all branches of life sciences. There is virtually no aspect of reality that cannot be described in thermodynamic terms, and there is virtually no phenomenon that is not subject to the thermodynamic principles. Eddington stated,

> *If someone points out to you that your pet theory of the universe is in disagreement with Maxwell's equations—then so much the worse for Maxwell's equations. … But if your theory is found to be against the second law of thermodynamics I can give you no hope; there is nothing for it but to collapse in deepest humiliation.*

Its generality was unconvincing to many, whereas the complex thermodynamic interrelationships within living systems very often made the inner essence of this science difficult to capture. There is no doubt, however, that thermodynamics can be defined as the general science of energy, not only because of its ability to fully describe energy transformations in human-dominated devices (from which it started) but also because of its still developing ability to describe and explain all energy and matter transformations within any systems in the biosphere (and the universe as well).

SEE ALSO THE FOLLOWING ARTICLES

Complex Systems and Energy • Conservation of Energy Concept, History of • Entropy • Entropy and the Economic Process • Physics and Economics of Energy, Conceptual Overview • Temperature and Its Measurement • Thermodynamics and Economics, Overview

Further Reading

Bianciardi, C., Tiezzi, E., and Ulgiati, S. (1993). Complete recycling of matter in the frameworks of physics, biology and ecological economics. *Ecol. Econ.* **8,** 1–5.

Carnot, L. S. (1927). "Réflexions sur la Puissance Motrice du feu e sur les Machines Propres à Developper cette Puissance. Biographie et Manuscripts de Sadi Carnot." Académie des Sciences, Paris. [Original work published 1824; English version published by Dover, New York, 1960].

Clausius, R. G. (1850). Ueber die bewegende Kraft der Wärme. *Annal. Physik Chemie* **79**, 368–397, 500–524. [Translated and excerpted in Magie, W. F. (1935). "A Source Book in Physics." McGraw-Hill, New York].

Eddington, A. S. (1948). "The Nature of the Physical World." Macmillan, New York.

Georgescu-Roegen, N. (1971). "The Entropy Law and the Economic Process." Harvard Univ. Press, Cambridge, MA.

Giannantoni, C. (2002). "The Maximum Em-Power Principle as the Basis for Thermodynamics of Quality. SGE Società Grafica Editoriale, Padova, Italy.

Giannantoni, C. (2003). The problem of the initial conditions and their physical meaning in linear differential equations of fractional order. *Appl. Mathematics Computation* **141,** 87–102.

Glansdorff, P., and Prigogine, I. (1971). "Thermodynamic Theory of Structure, Stability and Fluctuations." Wiley, New York.

Hatsopoulos, G., and Keenan, J. (1965). "Principles of General Thermodynamics." Wiley, New York.

Jørgensen, S. E. (1992). "Integration of Ecosystem Theories: A Pattern." Kluwer, Dordrecht.

Jørgensen, S. E. (2000). The tentative fourth law of thermodynamics. *In* "Handbook of Ecosystem Theories and Management" (S. E. Jorgensen and F. Muller, Eds.), pp. 161–174. Lewis, Boca Raton, FL.

Joule, J. P. (1845). On the existence of an equivalent relation between heat and the ordinary forms of mechanical power. *Philosophical Magazine* **27**(Series 3), 205.

Kelvin (Lord, Thomson, W.) (1851, March). On the dynamical theory of heat: With numerical results deduced from Mr. Joule's equivalent of a thermal unit and Mr. Regnault's observation on steam. *Trans. R. Soc. Edinburgh*.

Kestin, J. (1966). "A Course in Thermodynamics." Blaisdell, Boston.

Landsberg, P. T. (1961). "Thermodynamics with Quantum Statistical Illustrations." Wiley–Interscience, New York.

Landsberg, P. T. (1972, July). The fourth law of thermodynamics. *Nature* **238,** 229–231.

Lotka, A. J. (1922a). Contribution to the energetics of evolution. *Proc. Natl. Acad. Sci. USA* **8,** 147–151.

Lotka, A. J. (1922b). Natural selection as a physical principle. *Proc. Natl. Acad. Sci. USA* **8,** 151–155.

Nernst, W. (1906). Ueber die Berechnung chemischer Gleichgewichte aus thermischen Messungen. *Nachr. Kgl. Ges. Wiss. Gött.* **1,** 1–40.

Nicolis, G., and Prigogine, I. (1977). "Self-Organization in Nonequilibrium Systems." Wiley–Interscience, New York.

Odum, H. T. (1988). Self-organization, transformity, and information. *Science* **242,** 1132–1139.

Odum, H. T. (1996). "Environmental Accounting: Emergy and Environmental Decision Making." Wiley, New York.

Onsager, L. (1931). Reciprocal relations in irreversible processes. *Phys. Rev.* **37,** 405–427.

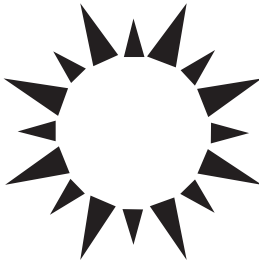
Prigogine, I. (1978). Time, structure, and fluctuations. *Science* **201,** 777–785.

Schneider, E. D., and Kay, J. J. (1994). Life as a manifestation of the second law of thermodynamics. *Mathl. Comput. Modelling* **19**(6–8), 25–48.

Szargut, J., Morris, D. R., and Steward, F. R. (1988). "Exergy Analysis of the Thermal, Chemical and Metallurgical Processes." Hemisphere, London.

Thermoregulation

JOHN R. SPEAKMAN
University of Aberdeen and Rowett Research Institute
Aberdeen, United Kingdom



1. Introduction
2. Mechanisms of Heat Exchange
3. Managing Heat Flows
4. Summary

Glossary

aestivation Reducing energy expenditure below basal levels at very high ambient temperatures.
basal metabolic rate The lowest metabolic rate reached by an animal maintaining bodily functions when at rest and not having recently eaten a meal.
behavioral thermoregulation Behaviors that alter management of heat flows from outside to allow regulation of body temperature.
endothermy Regulation of body temperature primarily using generation of heat internal to the animal.
euthermic Precise regulation of body temperature around a set point in the approximate range of 35 to 45°C.
exothermy (ectothermy) Regulation of body temperature using management of heat sources primarily external to the animal.
heliothermy Thermoregulation primarily dependent on varying the interception of incoming solar radiation.
homeothermy (homiothermy) An outmoded term describing the stability of body temperatures of endothermic animals; literally "constant temperatured."
physiological thermoregulation Physiological strategies that contribute to heat balance.
poikilothermy An outmoded term describing the variations in body temperatures of exothermic animals; literally "variable temperatured."

Thermoregulation describes the process by which animals manage the flow of energy into and out of their bodies, with the aim of maintaining a relatively constant body temperature within a range of 2 to 3°C. Most vertebrate animals attempt to regulate their body temperatures within such relatively narrow limits because enzyme functions are strongly dependent on temperature; therefore, there is only a narrow range of body temperatures within which metabolic processes can be optimized. Once body temperatures move outside these relatively narrow limits, animals start to experience profound reductions in their cellular functions and this has ramifications for their performance at the whole animal level. For many animals, exposure for even short periods to temperatures more than 10°C below or 5°C above those at which they routinely regulate their body temperatures is often fatal.

1. INTRODUCTION

The set points at which body temperatures are regulated vary tremendously among different animal groups. however, many animals regulate their body temperatures in the range of 30 to 45°C (303–318 K), and these animals are generally termed "euthermic." Body temperature can be regulated by altering the flow of heat into and out of the animal (behavioral thermoregulation or exothermy) and by generating heat internally (endothermy). Different species vary in the extent to which they use these alternative strategies for regulating their body temperatures, and in most animals the strategies are combined to varying extents. However, the physical laws that govern the amounts of energy required under various circumstances are common to all animals, independent of how they ultimately meet these demands.

2. MECHANISMS OF HEAT EXCHANGE

All animate and inanimate objects exchange energy with their surroundings by four different mechanisms: radiation, convection, conduction, and evaporation. This section reviews the essential features

 © 2004 Elsevier Inc. All rights reserved.

of these physical mechanisms in the context of their importance for animal heat balance.

2.1 Radiation

All objects emit electromagnetic radiation. The emitted radiation can be characterized by two parameters: the wavelength of emitted radiation and its intensity. Emitted radiation has a characteristic distribution of wavelengths that depends on surface temperature of the emitting object. The distribution at any particular temperature is peaked with a positive skew. The wavelength of modal intensity varies as a function of the surface temperature. As objects get hotter, the wavelength of peak intensity of radiation they emit gets shorter. This inverse relationship is generally called Wien's displacement law. Most objects on the surface of the earth have surface temperatures in the range of 253 to 313 K. The peak wavelengths of radiation emitted by objects in this temperature range are in the infrared. Because human eyes cannot detect infrared radiation, one cannot see the emitted radiation from most natural earth-bound objects directly. Glowing larva, molten metals, and hot coals in fires are notable exceptions of earth-bound objects that reach sufficiently high temperatures that the emitted radiation is directly visible by humans. Some animals have special sensory adaptations that enable them to detect infrared radiation. In particular, a group of snakes known as pit vipers can perceive emitted radiation from objects in the range of 273 to 313 K and use this to detect their prey. Because the prey thermoregulate, they are generally hotter than their surroundings and, thus, emit radiation with greater intensity and at higher wavelengths.

The other main source of radiation on earth comes from the sun. The surface temperature of the sun is approximately 6000 K. The wavelengths of radiation generated by objects at this temperature are between 400 and 700 nm. It is no coincidence that this is the range of wavelengths to which humans' eyes are sensitive and which, therefore, is defined as the visible spectrum. When one sees "an object," what one actually sees is the electromagnetic radiation that is reflected from its surface. The amount of energy that is radiated by a body depends on the surface temperature of the object and on whether it is a good or poor emitter of radiation. An object's capacity to emit radiation is defined relative to that of a perfect emitter, which is called a blackbody emitter. Capacity of a surface to emit radiation is defined by a term called the emissivity (ε). Emissivities vary between 1.0 (equal to the emissivity of a perfect emitting blackbody) and 0.0 (meaning the body would emit no radiation).

All objects emit amounts of radiation that are dependent on their surface temperatures and surface emissivities, and all objects also absorb radiation. Humans emit only in the infrared, and they absorb in both the infrared and visible spectral ranges. Objects that are good at emitting radiation at a certain wavelength are also good at absorbing it. Because the dominant sources of radiation are bimodally distributed with respect to their dominant wavelengths (i.e., in either the infrared or visible ranges), the capacity to absorb long- and shortwave radiation separately is often discussed. Longwave radiation emissivity/absorptivity tends to bear little relation to the color of the emitting surface as humans perceive it. Most animate objects have high water contents, and because water is a good emitter of longwave radiation, surface emissivities and absorptivities of most living organisms tend to be high ($\sim$0.95). However, absorption and reflection of shortwave radiation do vary with the color of the surface. Hence, dark colors have high absorptivities and light colors have low absorptivities. Capacities to absorb shortwave radiation are defined by the term "albedo." This is a value between 0.0 and 1.0 that expresses the capacity of a surface to absorb shortwave radiation relative to a perfect whitebody reflector. Hence, an object with an albedo of 1.0 is a perfect reflector and absorbs minimal amounts of incident shortwave radiation, whereas an object with an albedo of 0.0 is a good absorber. People can directly experience the impact of variations in albedo on absorbed radiation amounts very easily by changing the color of their clothing and exposing themselves to bright sunlight. Few people will fail to recognize the advantages of wearing light-colored clothing on hot summer days. This trend of wearing light-colored clothing during summer is a behavioral thermoregulatory behavior.

An interesting question that puzzled researchers for a considerable period is why races of humans that originate in tropical areas have dark skin colors compared with races that come from areas in the temperate and arctic regions. Explanations of this difference often suggested that skin color was related to both long- and shortwave radiation and that, overall, the dark skin color was advantageous in hot conditions because on balance it absorbed less radiation. However, this explanation is incorrect. In fact, dark skin is not a thermoregulatory adaptation at all but rather is a response to increase

absorption of shortwave radiation before it can penetrate the skin and cause dangerous lesions to epithelial cell DNA. This means that dark races actually do absorb more shortwave radiation than do their light-skinned counterparts. However, the benefits of this absorption are illustrated by comparing the rates of skin cancers in aboriginal Australians, where skin cancer is virtually unknown, and white Australians, who have a broadly European heritage and where skin cancers occurred in nearly 3% of the population during the early 1990s until vigorous campaigns were started to encourage safe practices when out in the sun.

The intensity of emitted radiation is very strongly dependent on surface temperature. In fact, emitted radiation is proportional to the surface emissivity multiplied by the absolute surface temperature (K) raised to the fourth power. This scaling exponent of 4 is very important because it means that as objects get hotter, the amounts of emitted radiation become disproportionately large. Hence, the surface of the sun at 6000 K is only 20 times hotter than humans are, at approximately 300 K, but each square meter of the sun's surface emits 160,000 times more electromagnetic radiation than do humans. Of course, because of the inverse square law and the fact that the earth is 150 million km from the sun, the amounts intercepting our planet are very much diminished. The consequence of this is that shortwave solar radiation input and longwave radiation outputs of our bodies are of the same order of magnitude, with both being approximately 1 kW per square meter of surface

2.2 Conduction

When any two solid objects are in direct physical contact, they exchange heat. Heat is "conducted" from the hotter object to the cooler object. The rate at which the heat flows depends on the surface area that is in contact between the bodies, properties of the material that connects them, and the path distance over which the heat is being conducted. The main property of an object that defines its capacity to conduct heat is called its thermal conductivity. Objects that conduct heat well are said to have high thermal conductivity, whereas those that conduct poorly are said to have low thermal conductivity. The inverse of conductivity is thermal insulation. Typical high-conductivity/low-insulation objects are metals and rocks, whereas typical low-conductivity/high-insulation materials are wood and fur. Among the lowest thermal conductivities of all is that of still air. This is why natural and man-made insulative materials tend to consist of structures that aim to trap air and keep it stationary (e.g., expanded polystyrene). In fact, the thermal conductivity of fur exceeds that of still air, so if the fleece of a sheep is compressed (thereby reducing the amount of air trapped in it), the conductance of heat across the fur actually increases.

One might imagine that because most animals do not have large amounts of their surfaces in contact with other solid objects, the role of conductance in the process of heat exchange would be minimal. But although this is generally true, one area where understanding the principles of conductance is important is in understanding the flow of heat from the core of a body to its surface. Because animals consist of solid objects, the heat flows from the central core to the surface via conductance. To understand the flow of heat from an animal's core to the surrounding air, imagine that heat is produced centrally at a notional single point. The heat flows first through the tissues, carried partly by blood flow through the arterial and capillary network and partly by conduction through the cells and tissues themselves. In some animals, the heat must then flow through a barrier of fur or feathers, primarily by conduction through the hairs or feathers themselves but also by conduction through the trapped air, and finally dissipates from the thin boundary air layer overlying the outer surface of the fur, primarily by convection and the net loss of thermal radiation. Both the type of heat transport and the path length along which heat must travel determine the resistance that each layer offers to the flow of heat. Heat is transferred not only by conductance but also by convection due to movement of the blood, and the process of convection transfers heat much more rapidly than does the process of conduction. Thick layers of material with low conductivity by definition offer more resistance to heat flow than do thin layers. Because each layer offers a resistance to heat flow, temperature drops progressively between the core and the skin, through the fur, and finally into the surrounding air. The extent to which the temperature drops across a given layer is determined by the proportional contribution of that layer to the total resistance offered by all layers to the flow of heat. A thick and efficient fur or feather layer, which offers a high resistance to heat flow, will exhibit a large drop in temperature between its inner and outer surfaces, and the total flow of energy across it will be reduced. Conversely, the difference between the core of the body and the area under the skin may differ

very little in temperature due to the high conductivity (assisted by convection effects).

There are very strong parallels between thermal and electrical conductivity and the concepts of thermal insulation and electrical resistance. Concepts of electrical circuitry have been widely borrowed to help explain how heat flows within animals from core to periphery. In this context, it is useful to imagine the heat as though it were electricity flowing through a number of resistances that may be arranged in parallel within a given layer and in series between layers. The value of this approach is that it allows one to apply the same formulas that predict current and voltage in electrical circuits: $I = V/R_{total}$, where I is the current, V is the voltage potential, and R_{total} is the circuit resistance. When thinking in terms of heat rather than electricity, we simply substitute heat flow (H) for I, temperature gradient (ΔT°) for V, and the sum of the resistances in the various layers for R_{total}. Thus, the overall flow of heat away from the core is calculated by $H = \Delta T^{\circ}/R_{total}$, and because $R = 1/C$, this is the same as $H = \Delta T^{\circ} \cdot C$.

There are several benefits to thinking in terms of resistances and ΔT°. One is that the overall resistance (R_{total}) can simply be found from the sum of the equivalent resistances for each layer, remembering that an equivalent resistance refers to a single resistance that could replace two or more resistances in parallel. A second benefit is that the resistance offered by a given layer (x) can be calculated from the heat flow and the temperature drop across that layer if the total heat flow is known ($R_x = \Delta T^{\circ}_x/H$). A third is that the relative importance of each layer in limiting heat flow is easy to visualize. Just as the voltage drop (ΔV_x) across a single resistor of a series can be predicted by the ratio of that resistor (R_x) to the total circuit resistance ($\Delta V_x = [R_x/R_{total}] \cdot V$), the temperature drop across a given layer is determined by the proportional contribution of the resistance in that layer to the total resistance offered by all layers ($\Delta T^{\circ}_x = [R_x/R_{total}] \cdot \Delta T^{\circ}$). Moreover, in this analogy, the total current (I) is equivalent to the total heat flux (H). Thus, factors that diminish I, such as elevated total resistance, can also be seen to parallel the effects of elevated total insulation on total heat requirement.

When peripheral blood flow is high due to dilation of the capillaries under the skin surface, considerable amounts of heat are transported to the skin surface by convective blood flow and the ratio of tissue resistance to the total resistance (R_{tissue}/R_{total}) is small. As a result, there is only a small temperature difference between core and skin, and the skin is seen as "hot." The surface temperature depends on the thickness of the overlying insulative layer and, thus, the total resistance afforded to heat flow along any particular path. A thermal image of an animal that picks up the emitted radiation clearly shows this variation in surface temperature as a function of the thickness of the insulative layer and the resultant heat loss across this layer. Figure 1 shows a thermal

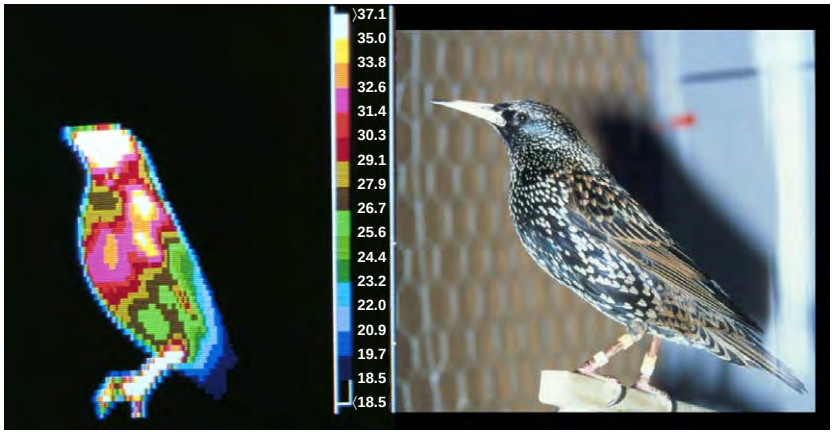


FIGURE 1 (A) Thermal image of a perching bird (a European starling, *Sturnus vulgaris*) and (B) a photograph of the same bird. In the first image (A), the camera detects the emitted longwave radiation and a computer uses the intensity of this radiation to generate, in real time, a map of the surface of the animal (see temperature scale at the side of the image). The second photographic image (B) relies on reflected shortwave radiation from the animal's surface. The surface temperature in panel A depends on the thickness of the external insulation layer. Areas where the layer is thinnest, particularly the head and legs, have the highest temperatures and the greatest flux of heat energy

image of a perching bird illustrating that the hottest areas are the legs and the face where the feathers are either absent or thin. The coolest areas are where the wings add an extra layer of insulation over the animal's lateral and ventral surfaces. Understanding this regional variation in surface temperature provides clear insight into why roosting birds often tuck their heads down under their wings and pull one leg up into their ventral plumage. By doing so, the birds conceal the hot parts of their bodies under layers of insulation and, thus, reduce the total heat flux. This is another form of behavioral thermoregulation.

Clearly, if the skin surface is hot due to dilation of the capillaries, another mechanism that animals might use to conserve heat is to reduce the flow of blood to the surface by closing down these capillaries, that is, vasoconstriction. In these conditions, convective heat transport to the skin declines toward zero. Most of the tissue resistance is now determined by the slower conduction of heat through the tissue matrix. The skin is seen as "cool" due to this increase in tissue resistance, resulting in a higher R_{tissue}/R_{total} ratio. As a result, less heat is delivered to the base of the fur, ΔT° through the fur is smaller, and heat flow through the fur is reduced.

The ability to depress skin temperature by vasoconstriction is affected by body size. By definition, small animals have short path lengths between the organs (where much of body heat is produced) and the skin surface, whereas large animals necessarily have longer path lengths. As a result, small animals have relatively low tissue resistance, and hence high skin temperatures, compared with those of large animals. This effect may be enhanced not only by constricting the blood supply to the surface but also by covering the undersurface of the skin with a layer of fat or blubber. Fat has a lower thermal conductivity than does lean tissue and, hence, acts as a barrier to heat flow. The best-known examples of this are the aquatic seals and whales that rely on a thick layer of subcutaneous blubber as insulation. Many terrestrial hibernating mammals also lay down substantial subcutaneous fat deposits. Although fat in these hibernating animals serves primarily as a means of energy storage, by analogy with aquatic mammals, one might be led to think that fat could also have a secondary role as an insulation layer for terrestrial animals. To analyze this problem, one should imagine a small lean animal with a tissue path length of 4 mm and a fur thickness of 6 mm and then consider what effect the deposition of a 1-mm thick subcutaneous fat deposit would have on heat flow. Muscle, fat, and fur offer resistances to heat flow of approximately 21.7, 46.9, and 262.9 $(W/^{\circ}C/mm)^{-1}$, respectively. For a lean animal having no fat deposit, the tissue resistance is 4 mm · 21.7 $(W/^{\circ}C/mm)^{-1}$, the fur resistance is 6 mm · 262.9 $(W/^{\circ}C/mm)^{-1}$, and the overall resistance is the sum of the two or 1664.2 $(W/^{\circ}C)^{-1}$. Adding a 1-mm subcutaneous fat layer would increase the overall resistance by 1 mm · 46.9 $(W/^{\circ}C/mm)^{-1}$, raising the overall resistance to 1711.1 $(W/^{\circ}C)^{-1}$, only 2.8% higher than the lean condition. Thus, the fat layer would reduce heat flow by only 2.8%. Therefore, we can conclude that 1 mm of fat is not an effective barrier to heat loss for small terrestrial animals and that fat storage in hibernators and other small animals probably does not serve as an insulator.

2.3 Convection

Conduction of heat occurs when two solids of different temperature are in contact. Convection occurs when a solid is in contact with a fluid (generally air or water on Earth) of different temperature. The "different temperature" part of this definition is often neglected, but it is clear that energy can only flow down a thermal gradient, so if there is no temperature differential the heat flow will similarly disappear. Using a fan increases airflow across the surface of the face, and because this is naked skin that is well perfused with blood, the increase in convective heat flow is an effective way in which to contribute behaviorally to thermoregulation by elevating surface heat losses. However, there is an increase in endogenous heat production when using a fan due to the muscular activity involved in the waving motion. Once air temperature rises to the same level as skin temperature, heat losses fall to zero no matter how hard one fans. The heat generated by fanning exceeds that from cooling under these conditions.

In fact, the heat loss by convection is a function of several factors. One is the driving temperature gradient, that is, the difference in temperature between the solid and the fluid in which it is immersed. The area over which heat is being lost is obviously important, and the shape and structure of the surface also affect how exposed it is to the fluid flow and, hence, the magnitude of the impact of the velocity of flow. The conductivity of the fluid is also an important factor. For example, water has a conductivity 23 times greater than that of air, and this greatly magnifies the cooling effect of convection in water. Although velocity is a factor in heat loss by convection, there is still some loss when an object is

stationary and there is no external flow of the fluid. This is a state called free convection. Free convection occurs because the air close to the surface of an object becomes heated by the object itself, and as this air warms up, its density declines and it rises upward. Standing humans in a completely still condition have plumes of hot air rising off the tops of their heads, and this hot air is replaced by colder air coming in from below. Therefore, convection currents are generated by the existence of a hot object within a cool liquid. But clearly, as the fluid moves as a whole, the velocity of that flow influences the convective heat loss. This is a state of forced convection.

The interaction between surface characteristics and velocity of flow on forced convective heat loss is complex. However, the relationship between airflow and heat loss generally follows a nonlinear pattern of diminishing increase with increasing velocity until at some point further elevations in velocity have only marginal effects. The exact exponent to which velocity is raised in these relationships varies somewhat with the shape and surface structures on the object, but a commonly assumed exponent is 0.5. Thus, heat loss increases as a function of the square root of fluid velocity. This is because the surface insulation is provided primarily by the thin skin of still air at the body surface known as the boundary layer. This air is gradually stripped away as velocity increases, but eventually when it has all gone, there is nothing else to lose. The manner in which the boundary layer is stripped depends on how the flow transits from laminar to turbulent flow, and characteristics of the shape and surface can influence this transitional velocity, thereby complicating the exact value of the exponent. It is clear that animals can use these properties to minimize convective heat loss by altering their shape to promote laminar flows. The example cited earlier of birds tucking their heads under their wings and standing on one leg may also function to present an aerodynamically favorable perspective in an airflow to minimize turbulent airflow and, thus, minimize convective heat loss.

2.4 Evaporation

The previous three modes of heat flow all work down temperature gradients. Heat flows from hot objects to cold objects by the processes of radiation, conduction, and convection. However, evaporation allows an object to lose heat even if the surrounding medium in which it is immersed is hotter than the object itself. This is consequently an essential thermoregulatory mechanism for many animals because it is the only method they can employ to keep cool if ambient temperatures exceed their body temperatures. Evaporation involves using energy to convert liquid water into the gaseous phase. Because water is strongly hydrogen bonded, the amount of energy required to convert liquid water to gas is high and, thus, evaporation is a very effective method for cooling down. To achieve this, many animals have specialized mechanisms such as sweating (where water oozes onto the surface of the animal and evaporates from there), panting (where a respiratory airflow is passed over the tongue to promote evaporative losses), and self-anointing (where the animal licks itself and the evaporating saliva provides the cooling). The capacity for evaporation depends on the ability of the surrounding medium to accept water vapor. The capacity of air to hold water vapor increases as a function of temperature. When the air is saturated, it is said to have a relative humidity of 100%. At this point, the air can accept no further water vapor and net evaporation from the surface of an object in that air ceases. This is why humid conditions are far more uncomfortable for humans than are dry conditions, even though in the latter case it may be hotter in absolute temperature. In dry conditions, one can thermoregulate by evaporation, but when the conditions also include very high humidity, even this avenue for heat loss is shut down.

3. MANAGING HEAT FLOWS

For very small animals (less than 1 mg), their surface areas are such that sustaining any temperature differential between the inside of the animal and the environment is nearly impossible. These animals maintain a constant internal temperature only by locating themselves in an environment where the temperature itself is constant. If the environment fluctuates in temperature, the animals have mechanisms to avoid the most extreme conditions in suspended states until the conditions improve again. Most lower invertebrates approach the problem of thermoregulation in this manner, that is, by avoidance. However, once animals become larger, their relative surface areas decline and it becomes feasible to sustain a temperature differential between the inside and outside of their bodies. This means that the animals can maintain homeostasis of their internal temperatures despite the range of external conditions and the variability in internal heat production as a by-product of activity. Most larger invertebrates have some limited capacities to regulate

their temperatures to some extent, and they combine these capacities with physiological or actual physical avoidance to survive temperature extremes.

However, it is in the vertebrates that thermoregulation has evolved to its greatest diversity and sophistication. The remainder of this article discusses the thermoregulatory strategies adopted by the major vertebrate groups.

3.1 Fish

Most fish spend most of their lives in water. This lifestyle simplifies the consideration of heat fluxes because it is not possible to engage in evaporation underwater, and beyond only a few meters of depth nearly all radiation has been absorbed. Thus, the heat budget management of fish is more simplified than that of terrestrial vertebrates. Water has a high specific heat capacity and, hence, tends to buffer environmental extremes. For example, the surface temperature of the Pacific Ocean differs by approximately 5 K over a distance of several thousand kilometers. The same temperature difference on land can be found between the shaded and exposed sides of a tree, that is, over a distance of a few meters. Therefore, fish are generally in a medium where the external temperature is relatively stable. The key problem, then, is that because of variations in activity, as well as variations in the energy required to digest food, the rates of endogenous heat production by the animals fluctuate widely. Because the external temperature is relatively constant, the driving gradient for heat loss by convection would also appear to be constant. Therefore, one might anticipate that as fish become active, they may experience thermoregulatory problems dissipating their body heat.

However, in addition to having high thermal conductivity relative to air, water has a low oxygen content relative to air. In fact, even oxygen-saturated water has only 1/40th the oxygen content of atmospheric air. Therefore, fish need extremely efficient mechanisms to extract oxygen from their environment to support their metabolism. The main mechanism they use is a countercurrent exchange mechanism in the gill arches, which have highly specialized enlarged surfaces to facilitate oxygen uptake and the removal of carbon dioxide generated by oxidative metabolic processes. What is a good mechanism for extracting oxygen from water is also a very efficient mechanism for getting rid of body heat. Indeed, because water has 23 times the cooling power of air yet only 1/40th the oxygen content, for each milliliter of oxygen extracted, the cooling power available to fish is approximately $23 \times 40 = 920$ times greater. Fortunately, elevated oxygen requirements occur when the animals are active and generating the most heat. So, there is an automatic passive feedback system that keeps the body temperatures of fish at the same level as their environment regardless of their level of activity. When activity increases accompanied by an increase in heat production, increased ventilation of the gills to get oxygen to support the activity automatically dumps the excess heat, avoiding overheating.

This is a workable strategy so long as the environmental temperature remains stable. However, temperatures vary spatially in bodies of water, particularly if the water is shallow and able to be heated by the sun in places but shaded elsewhere. Moreover, most bodies of water develop thermal stratification where the upper layers are heated by solar radiation effects but there is an abrupt thermocline separating this surface-heated water from the more stable cold water at depth. Obviously, fish might swim into such cool areas or into hot areas, with potentially serious consequences for their body temperatures. Fish accommodate such spatial temperature effects by behavioral strategies. They swim faster and turn more as they move into water that is cooler or hotter than their preferred temperature, and these simple kinetic mechanisms bring the animals back to an environment with a water temperature they prefer. However, if rapid temperature changes occur across an entire area, the animals have no mechanisms to avert disaster. Anyone who has kept tropical fish and had a heater fail in the tank in a cold temperate environment likely has observed the inability of fish to survive such an event. In the wild, these situations must be very rare. However, seasonal shifts in the temperature of the environment are commonplace, and fish have evolved a series of complex behavioral and physiological responses to overcome such seasonal effects.

In fact, a common strategy in the face of the change is simply to avoid it by migrating, and many fish species do this in the ocean. However, in freshwater lakes where the whole available environment is heating or cooling, migration is often not an option. As water cools down, the problem facing the fish is that their cellular performance, and hence metabolic and ultimately whole body performance, might be compromised. The issue then is to sustain cellular processes functioning at the lower temperature. To achieve this, fish alter the amounts of enzymes they manufacture, shift production to

isozymes that work best at lower temperatures, and modulate the medium in which the enzymes are working to maximize their activities. These have been called the quantitative, qualitative, and modulative responses, respectively, and together they comprise the physiological thermoregulatory strategy that allows fish in seasonally changing environments to survive without running away.

However, some fish do not conform to this pattern of thermoregulatory control. This group includes tuna, swordfish, and large game fish such as marlin. In these animals, there is a complex countercurrent heat exchange blood flow that reclaims heat from the blood before it passes through the gills. Hence, blood at the gills is at water temperature, and blood in the core body can be up to 15°C warmer. This allows these animals considerable flexibility to sustain performance in waters of varying temperature, and they use this capacity to swim below the thermocline into cooler waters and exploit prey populations living there.

3.2 Reptiles

Most reptiles live on land. Therefore, the problems they face include those faced by fish, of highly variable rates of endogenous heat production when the animals are physically active or digesting food. But combined with this problem is the fact that the environment in which reptiles often find themselves contains a tremendous spatial matrix of conditions where potential heat flux varies over orders of magnitude. Moreover, the modes by which heat flows between the animals and their environment entails a level of complexity way above that experienced by the average fish because radiation, evaporation, and conduction all come into play in addition to convection. Therefore, as might be anticipated, reptiles have extremely sophisticated thermoregulatory capacities to sustain their bodies at relatively constant temperatures in the face of this potential heat management nightmare.

This seems so obvious that it may come as a surprise to find that for a long time the thermoregulatory capacities of reptiles were considered extremely rudimentary. These impressions came from observations made during the 1800s of animals that had been captured in the wild. The typical reptile was captured from the wild and held overnight in a box, as was frequently the practice for animal collectors. When the reptile was removed from the box the next morning (assuming that the animal was not unduly large), it would generally have reached thermal equilibrium with its surroundings and have a body temperature similar to that of the box. If it was cold, the body temperature would also be cold and the animal would have the consequent impacts of reduced body temperature on performance, including extremely sluggish movements and a generalized inability to respond to stimuli. Therefore, it is a small wonder that the animals were termed "cold-blooded" or poikliothermic (literally "variable temperatured"), reflecting the notion that they were incapable of thermoregulation and, thus, that their body temperatures varied along with the many temperatures of the environment. One of the few physiological skills the animals appeared to have was the ability to survive body temperature fluctuations that would be lethal to most mammals.

This notion arose from the abstraction of these animals from their environment. When a pioneering group of physiologists during the 1950s started to examine animals in the wild, they uncovered the true nature of reptilian thermoregulatory capacities. In fact, during the daytime, most reptiles are very good at regulating their body temperatures within very narrow limits. They do this primarily by regulating the level of absorption of shortwave radiation. To increase body temperature, reptiles bask in the sun. They maximize uptake of absorption by exposing the maximal area they can toward the sun, adopting body angles that maximize interception of radiation. Many reptiles have specialized dark patches on their surfaces that serve to decrease surface albedo and, thus, absorb more incident shortwave radiation, not unlike little thermal windows on their surfaces. By locating themselves onto hot surfaces that have heated up in the sun, they can gain additional energy to heat themselves by conduction. To cool down, the animals turn themselves away from the sun, minimizing their interception area and turn their dark patches away so that they no longer absorb. If the problem is severe, they seek shaded areas where the incoming shortwave radiation is massively reduced and the substrate in which they are in heat exchange by conduction is cooler. Ultimately, animals may avoid the environmental extremes by burying themselves, either under rocks or in the substrate itself.

The dependence of reptiles on shortwave radiation from outside their bodies to control their energy balance, and hence their body temperatures, has been termed "heliothermy" (from *helios* or "the sun"). However, a more common term emphasizing the fact that most of the heat used to control body temperature is from external sources is "exothermy" or "ectothermy." One disadvantage of relying so

heavily on the solar radiation for energy is that once the sun goes away, the animals are deprived of their primary heat source and start to cool down. This is why reptiles left in boxes overnight generally have body temperatures in the morning that are at the same temperature as the ambient temperature and, thus, appear to be unable to thermoregulate. This thermal dependency is mitigated somewhat by size given that larger animals, by virtue of their greater surface/volume ratio, cool more slowly. For example, komodo dragons, which are the largest extant lizards, spend all day at body temperatures of roughly 36 to 38°C but overnight cool by only 5 to 8°C, even when the environmental temperature is approximately 20 to 25°C. This slow cooling is primarily a function of their large size. The heavy dependence on the sun also explains why the distribution of reptile species is primarily tropical.

Problems arise for reptiles if they require going into water. Because of the very high thermal conductivity of water, combined with the fact the animals are cut off from their usual source of external radiation, reptiles in water cool rapidly to the water temperature. Marine iguanas from the Galapagos Islands routinely feed in the splash zone at the top end of beaches and may dive to feed on algae up to 3 m deep, spending protracted periods submerged. That these animals can submerge for long periods without succumbing fatally has been known for some considerable time. Darwin, in *The Origin of Species,* remarked on a rather crude experiment performed by a sailor in which a marine iguana was tethered to rocks and submerged for an hour but was still alive on recovery. Studies of blood circulation have shown that marine iguanas have exquisite control over their circulatory systems, allowing them to vasoconstrict during diving to concentrate blood flow in their core and, thus, retard total heat loss and their cooling rate, but then to vasodilate on emergence to rapidly heat up by a combination of basking (uptake of solar radiation) and direct contact with the black lava rocks that form the beachheads (conduction). Studies of other marine reptiles such as sea turtles indicate that similar circulatory adjustments allow them to minimize cooling and/or maximize heating rates by adjusting blood flow into their front flippers as they encounter waters of varying temperatures. Leatherback turtles might even have some degree of endothermic heat production beyond the slow cooling engendered by their large size, allowing them to stray outside tropical waters, but this capacity is a matter of some debate.

However, other reptiles certainly do have the capacity to generate heat internally. Probably the best example of this is brooding pythons. These large snakes are able to generate heat internally, allowing them to transfer heat to their eggs and thereby stabilize incubation temperatures. The mechanisms involved in this internal heat production appear to include only shivering, that is, systematic contraction of the skeletal muscles not linked to gross body movement that generates heat as a by-product. However, even in these large snakes, the amounts of heat generated are insufficient to enable the animals to maintain a constant high body temperature in the face of declining ambient temperature when there is no external source of radiation.

3.3 Birds and Mammals

If a bird or mammal is placed overnight in a box, and the animal is deprived of an external source of heat but is provided with food, in the morning its body temperature will be largely independent of the ambient temperature. This contrasting observation to the response of reptiles led early researchers to classify these animals as "homeotherms" (literally "same temperatured"), reflecting the stability in the body temperatures, and to assume that these animals have much more sophisticated systems of thermoregulation. In fact, the problems faced by birds, mammals, and reptiles are very similar—enormous changes in the driving gradients combined with great variability in the endogenous heat production due to exercise.

However, the solution to the problem that has evolved in birds and mammals is radically different. This does not mean that birds and mammals do not take advantage of the ability to manage their heat flows in behavioral exothermic ways. Many birds and mammals bask in the early morning sun, using exogenous shortwave radiation to raise their body temperatures. Both groups seek shade to avoid overheating, husky dogs prostrate themselves on snow to cool down after running, and nearly all birds and mammals exhibit a large number of other minor behavioral adjustments to manage their heat budgets (some of which were mentioned previously in the context of describing the modes of heat transfer). But the main difference is that whereas nearly all of the heat used by reptiles is externally derived, for mammals and birds it is internally derived. Hence, these animals are often called "endotherms." Functionally, both birds and mammals are endothermic, but the physiological mechanisms they use to

generate the heat are radically different; birds use shivering, whereas mammals have evolved a specialized tissue (called brown adipose tissue) that serves to generate heat.

To understand the variations in the energy demands of endothermy, it is useful to consider the energy that would be required to sustain the temperature of a beaker of water that nominally represents an animal's body. If one had a beaker of warm water at, say, 38°C and it was standing in an environment of 20°C, it would cool down by the processes of free convection, net radiation losses, and conduction to the surface on which it was standing. Therefore, to keep the water hot, one would need to place a heater in it and deliver sufficient energy internally to balance these heat losses. Because radiation, convection, and conduction all are functions of temperature difference between the beaker and the environment, the amount of energy one would need to supply with the heater would be a linear function of ambient temperature, crossing the *x* axis at the temperature of water in the beaker (Fig. 2A). That is, the energy one would need to supply if the ambient temperature was 38°C would obviously be zero because the beaker at 38°C would already be in thermal equilibrium. The gradient of the relation between heat requirement and temperature is termed the "thermal conductance" of the system under consideration. The gradient of this line reflects how efficiently the body exchanges heat with its environment. Hence, one could reduce heat flow and, hence, the heat requirement by wrapping the beaker in a blanket. The more insulation layers one added, the shallower and shallower the gradient would become as thermal conductance declined.

In the imaginary system just described, the minimal heat requirement to sustain the beaker temperature is zero. However, real animals cannot have zero heat production. This is because everything they do requires energy. Even when an animal is completely stationary, all its cellular metabolic processes still generate heat as a by-product. If we cross the lines of thermal conductance with this minimal heat production, it becomes clear that these lines intersect at certain points. These are the ambient temperatures where the animal's minimal heat production exactly balances its heat requirement from its thermal conductivity. Above this temperature, the animal is producing too much heat, and below this temperature, it is not producing enough heat to maintain energy balance. This point is called the lower critical temperature for an endotherm. Lower critical temperatures are lower

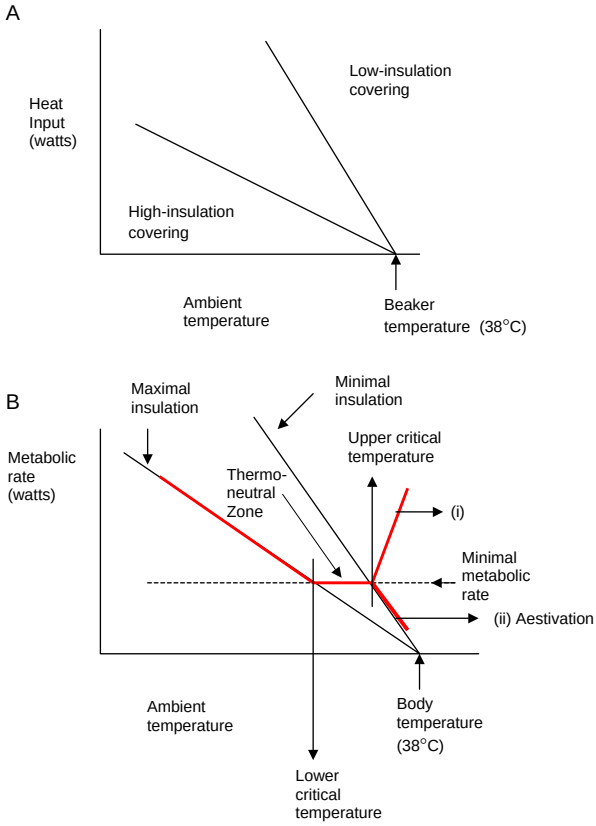


FIGURE 2 (A) Heat input required to sustain a beaker of water at a constant fixed temperature of 38°C as a function of ambient temperature and the thermal conductance of the material surrounding the beaker. Lower thermal conductance (higher insulation) leads to lower energy demands. When the environmental temperature equals 38°C, the heat requirement falls to zero. (B) Pattern of heat requirement as a function of ambient temperature and thermal conductance, with the minimal metabolic rate (= heat production) of an animal overlaid on this pattern. Where the minimal heat production crosses the minimal thermal conductance is called the lower critical temperature point. At lower ambient temperatures, an endothermic animal must supply heat internally to remain at the same body temperature. Above the lower critical temperature, metabolism remains constant (thermoneutral zone) and animals modulate their thermal conductance until they reach their maximal thermal conductance (lowest insulation) at the upper critical temperature point. Above this point, animals can recruit mechanisms to dissipate heat but paradoxically raise their heat production to do this (i), or they can start to shut down their minimal processes and aestivate (ii).

in animals with lower thermal conductances. At temperatures below the lower critical point, an animal must elevate its endogenous heat production to balance heat losses if it is to sustain its body temperatures. The exact requirements track along the line defined by the thermal conductance. Above the lower critical point, the problem an animal faces is to get rid of the surplus heat. It can do this by

making behavioral adjustments that increase thermal conductance (e.g., adopting postures that facilitate airflow, reducing radiation uptake by seeking shade). Humans are uniquely able to rapidly manipulate their external insulation levels by taking off their outer clothing.

It is possible, in making these behavioral adjustments, to elevate conductance so that the minimal heat production balances the heat loss. However, at some point, these behavioral mechanisms can achieve no greater increase in conductance. When ambient temperatures rise above this point, the animal has to do something different. This point is called the upper critical temperature. Between the lower and upper critical temperatures of an endotherm, metabolic heat production remains constant and independent of temperature. This zone is called the thermoneutral zone. It is bounded by the upper and lower limits of variation in an animal's thermal conductance. If temperature rises above the upper critical point, an animal can do one of two things. First, to balance heat production with heat requirement, it can start to shut down its basic metabolic processes and fall into a state of lethargy that is called aestivation. Second, it can spend a little energy, paradoxically increasing heat production, if the net heat loss by this procedure is increased. In general, these mechanisms involve increasing evaporative water loss. When temperatures continue to rise, this increase gets exponentially steeper as the animal brings into play increasingly desperate measures to stay cool. Obviously, at some point, water loss in these states becomes a critical factor, and eventually the animal cannot regulate temperature any longer and the body temperature starts to rise—perhaps fatally if water cannot be replenished.

This model of heat loss in endothermic animals is generally called the "Newtonian cooling model," and although it involves several simplifications of the actual situation observed in many animals, the model highlights some of the features of being endothermic and how these differ critically from those of exothermic animals. First, examination of the gradients of the curves relating thermal conductance to energy requirements clearly shows that the greater the degree of external insulation, the lower the energy demands of endothermy. Therefore, simple power demand arguments suggest that as the temperatures get cooler, endothermic animals should thicken up their external insulation layers. Indeed, many examples of this are observed in nature; Tropical endotherms generally have much less external insulation than do temperate and arctic equivalents, and temperate and arctic animals often molt their external insulation seasonally between a thick winter coat and a thin summer coat. Indeed, humans also make the same seasonal and latitudinal alterations in their external insulation and may make such modulations over much shorter time scales of minutes as the heat requirements change.

What is perhaps less obvious is that the surface/volume ratio effects mean that larger animals also have lower thermal conductances and lower requirements to balance heat loss (per gram of body tissue). Therefore, a power argument suggests that it is advantageous for animals in cold conditions to be larger than their conspecifics in warmer climates. This is known as "Bergmann's law" and has a corollary, "Allen's rule," that not only should animals get larger in colder climates, but their appendages should get smaller as well so as to minimize heat losses. There are many examples of animals conforming to Bergmann's law, but as many endotherms do not conform to it as do, and many species actually show converse trends. Moreover, many exotherms show similar trends despite the energetics arguments not applying to them. This illustrates that power arguments and surface/volume ratios are but one of many selective pressures on animal body size, making generalizable rules difficult to formulate.

As animals get very large, they start to face the problem that their thermal conductance curves are so shallow that they are nearly always below the minimal metabolism line. These animals face problems in dissipating heat; consequently, to facilitate this process, they reduce the extent of their external insulation. Indeed, the largest extant land mammals, elephants and rhinos, have dispensed with external insulative fur completely. However, for most endothermic animals, the main problem they face is keeping heat they generate internally inside their bodies, and that is why nearly all of the endothermic birds and mammals have well-developed fur or feathers. This contrasts with the situation in reptiles that have no external insulation at all. External insulation layers would clearly be detrimental to reptiles, whose major source of heat is external, because the insulation would retard the heat from getting in. An obvious and seemingly paradoxical exception to the external insulation rule are the marine mammals. This is paradoxical because if one group of animals might be expected to most need external insulation, it would be endotherms immersed in water—with their thermal conductivity 23 times greater than that of air. Why have these

animals not evolved or lost their fur? The answer is that fur is not hydrodynamic given that its major function is to trap a layer of air. By replacing the external fur with an internal blubber layer, many marine mammals have overcome the problem that the fur layer would interfere with their abilities to swim and perform in the aquatic environment. It was shown earlier that fat layers are ineffective as insulation for smaller mammals. Similarly, smaller aquatic animals, such as sea otters, mink, and water shrews, retain their fur despite its unfavorable hydrodynamic properties.

Endothermy clearly has many advantages in freeing endothermic animals from the dependency on solar radiation. Endotherms can use solar radiation if they wish (somewhat less efficiently due to their external insulation), but if the sun does not shine, they are still able to function effectively, unlike their exothermic relatives. The main disadvantage of being endothermic is that because the energy for heat generation comes from within, the animals must find and eat considerably more food to fuel that energy requirement. Estimates made on small birds and small reptiles of their total daily food requirements suggest that small endothermic birds may, in fact, need to eat 50 to 100 times more food daily than do small lizards of equal body mass.

3.5 Dinosaurs

The question of whether the dinosaurs were endotherms or had thermoregulatory strategies similar to the modern exothermic reptiles to which they were more closely related has been an issue of debate for more than 100 years and, in fact, dates back to the very first discoveries of dinosaur fossils. Wilfred Owen, who gave the group the name "Dinosauria," actually suggested they were probably warm-blooded animals, but this appears to have been more of a political statement to attempt to refute Darwinism than a truly held belief. Throughout the first half of the 20th century, it was mostly believed that dinosaurs were exotherms. However, opinion started to turn during the late 1960s and early 1970s, particularly due to the work of Robert Bakker and publication of his book, *The Dinosaur Heresies,* in 1977. By the late 1990s the public perception of dinosaurs had certainly embraced the idea that they were endotherms.

Several lines of evidence point toward this conclusion. However, many other lines of evidence paint a more confused picture. For example, although we might not expect the larger dinosaur representatives to have external insulation, we would certainly expect this for the smaller representatives. Yet, apart from an indication of some feathering on recent fossils from China, all other small dinosaurs appear to have been devoid of external insulation. In contrast, there is clear evidence that another group of reptiles, the Pterosaurs, appear to have been endothermic, supporting their capacity to fly, and also had external insulation in the form of fur on their bodies. Moreover, extant endotherms have bony structures in their nasal passages that support structures to reclaim heat from respiratory gas exchange, that is, turbinates. All extant endotherms have these bony structures, but they are absent in fossil dinosaurs. Overall, then, the evidence supporting an endothermic dinosaur thermoregulatory strategy is at best confused.

4. SUMMARY

Thermoregulation is the maintenance of a stable body temperature by managing the flow of heat into and out of the body. Although there are two major strategies, exothermy (relying on external heat sources) and endothermy (relying on internally generated heat), all animals use a combination of behavioral and physiological strategies to control their heat balances to some extent. In fish, amphibia, and reptiles, the balance is much more toward exothermic regulation, whereas in mammals and birds, the balance is shifted toward endothermy. The jury is still out on the thermoregulatory strategies used by the extinct dinosaurs, although one ancient extinct reptilian group whose members were contemporaries of dinosaurs, the Pterosaurs, were almost certainly endotherms.

SEE ALSO THE FOLLOWING ARTICLES

Food Capture, Energy Costs of • *Heterotrophic Energy Flows* • *Human Energetics* • *Migration, Energy Costs of* • *Photosynthesis and Autotrophic Energy Flows* • *Reproduction, Energy Costs of* • *Thermal Comfort*

Further Reading

Bakker, R. T. (2001). "The Dinosaur Heresies: New Theories Unlocking the Mystery of the Dinosaurs and Their Extinction." Citadel Press, Secaucus, NJ.

Blumberg, M. S. (2002). "Body Heat: Temperature and Life on Earth." Harvard University Press, Cambridge, MA.

Bowler, K., Cossins, A., and Heldmaier, G. (in press). "Temperature Biology of Animals." Cambridge University Press, Cambridge, UK

Davenport, J. (1991). "Life at Low Temperatures." Cambridge University Press, Cambridge, UK.

Heinrich, B. (1993). "The Hot-Blooded Insects: Strategies and Mechanisms of Thermoregulation." Harvard University Press, Cambridge, MA.

Johnston, I. A., and Bennett, A. F. (1996). "Animals and Temperature: Phenotypic and Evolutionary Adaptation." Cambridge University Press, Cambridge, UK.

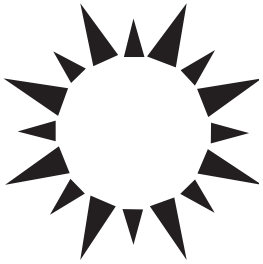
Jones, J. C. (2000). "The Principles of Thermal Sciences." CRC Press, Boca Raton, FL.

Schmidt Nielsen, K. (1997). "Animal Physiology: Adaptation and Environment." 5th ed. Cambridge University Press, Cambridge, UK.

Schonbaum, E., and Lomax, P. (1991). "Thermoregulation: Physiology and Biochemistry." Pergamon, London.

Tidal Energy

IAN G. BRYDEN
The Robert Gordon University
Aberdeen, United Kingdom



1. Overview
2. The Tidal Phenomenon
3. Tidal Barrage Methods
4. Tidal Current Systems
5. The Future of Tidal Power

Glossary

barrage An artificial barrier across an estuary or river to control the flow of water.
gigawatt (GW) The unit of power defined as 1 billion (1,000,000,000) watts.
gigawatt hour (GWhr) The unit of energy defined as that delivered by a source of one GW in 1 hour.
terawatt (TW) The unit of power defined as 1000 billion (1,000,000,000,000) watts.
terawatt hour (TWhr) The unit of energy defined as that delivered by a source of 1 TW in 1 hour.
tide The motion of the worlds seas and oceans induced by the gravitational and centrifugal influence of the earth, moon, and sun.
turbine A rotating mechanical device, which converts the kinetic energy of a flowing fluid into mechanical motion.

Tidal power is not a new idea. Mills, which used tidal flows in bays and estuaries to drive machinery to grind cereal, were widely used in medieval times. Tides have been seriously reexamined as a potential source of energy for industry and commerce. In the United Kingdom, there were numerous proposals throughout the 20th century to exploit the tidal energy potential of the Severn Estuary. None have yet been developed, primarily as a result of the anticipated costs and concern over the environmental changes that would arise from such a large-scale development.

1. OVERVIEW

The world's first serious scheme to exploit tidal energy was constructed in France, at La Rance in Brittany between 1961 and 1967, and consists of a barrage across a tidal estuary to utilize the rise and fall in sea level induced by the tides. This scheme has proven to be highly successful despite some early teething problems. Many engineers and developers now favor, however, the use of alternative technology, which will utilize the kinetic energy in flowing tidal currents.

Like any energy resource, the prospect of tidal power exploitation has stimulated opposition. Tidal barrage proposals, like La Rance, have produced the most vocal opposition. Environmental groups, although generally in favor of the exploitation of alternative energy sources, are suspicious of the likely environmental changes large estuary based schemes would produce. Political opposition has also been considerable in some cases. One politician in the United Kingdom likened the proposed creation of a barrage across the Severn Estuary to the formation of a "large stinking lake." Similar political opposition has been voiced against any development of the tidal resource in the Solway Firth between Scotland and England. It is anticipated that public and political opposition will limit the development of tidal barrage schemes in the short term.

Opposition to schemes designed to exploit the kinetic energy in tidal currents using open turbines, analogous to wind turbines, has been less voluble, although there has as yet been no commercial demonstration of such technology. If opposition remains muted, then tidal current turbines could be installed in many parts of the world in the early 21st century.

2. THE TIDAL PHENOMENON

2.1 Introduction

The tides are cyclic variations in the levels of the seas and oceans. Water currents accompany these

 © 2004 Elsevier Inc. All rights reserved.

variations in sea level, which, in some locations such as the Pentland Firth to the North of the Scottish mainland, can be extreme.

The explanation of the existence of tides represented one of the greatest challenges to early oceanographers, mathematicians, and physicists. It was not until Newton that a satisfying theory emerged to explain at least some of the properties of the tides. He formulated a theory, which has become known as the equilibrium theory of tides.

2.2 Newton's Approach: The Equilibrium Theory of Tides

The equilibrium theory of tides gives a partial description of tidal behavior for an abstract planet Earth, which is entirely and uniformly covered by water. According to the Newtonian model, the earth-moon system rotates around a common center of mass (CoMs) and the radius of this circulation is given by r, as shown in Fig. 1. The separation of the center of mass of the earth (CoM_e) from the center of mass of the moon (CoM_m) is given by R, which is also shown.

If the earth were not itself rotating, each point on, or in, the Earth would rotate about its own center of rotation, the radius of the rotation would also be given by r and the period of rotation would be equal to the rotational period of the earth-moon system. This results in acceleration toward the local center of rotation.

At the center of the earth, the apparent centrifugal acceleration resulting from the rotation exactly matches the gravitational acceleration. At all other points, there is an imbalance between gravitational and centrifugal effects. At the point B, as shown in Fig. 1, the centrifugal effects exceed the lunar gravitational attraction. In effect, at the surface of the earth, there will be a net flow of water from C & D to A & B. This effect results in the lunar tidal cycle. The equilibrium theory suggests, therefore, the establishment of tidal bulges in the fluid surrounding the earth, as shown in Fig. 2.

The earth rotates, and the two tidal bulges must maintain their positions with respect to the moon. They, therefore, have to travel around the earth at the same rate as the earth's rotation. The moon rotates around the CoM_s every 27.3 days in the same direction that the earth rotates every 24 hours. Because the rotations are in the same direction, the net effect is that the period of the earth's rotation, with respect to the earth-moon system, is 24 hours

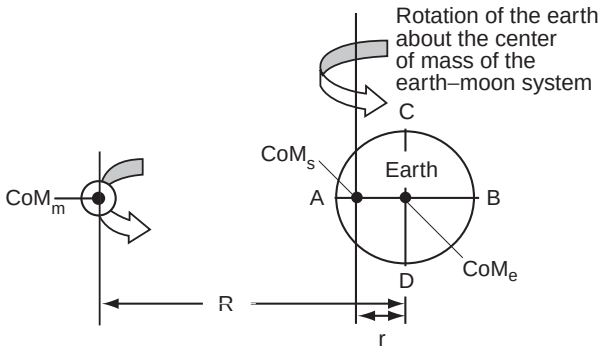


FIGURE 1 Schematic of the earth-moon system.

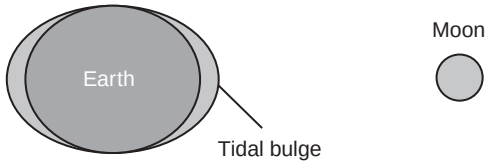


FIGURE 2 The tidal bulge.

and 50 minutes. This explains why the tides are approximately an hour later each day.

During the lunar month, which is the rotational period of the moon around the earth, there will be variations in the lunar tide influence. The lunar orbit is not circular but is elliptical in form, and the tide-producing forces vary by approximately 40% over the month. Similarly, the moon does not orbit around the earth's equator. Instead there is an angle of approximately 28° between the equator and the plane of the lunar orbit. This also results in monthly variations.

2.3 Influence of the Sun on the Tides

The earth-sun system is also elliptical but with only a 4% difference between the maximum and minimum distance from the earth to the sun. The relative positions of the earth, moon, and sun produce the most noticeable variations in the sizes of the tides.

In the configuration shown in Fig. 3, the influence of the moon and sun reinforce each other to produce the large tides known as spring tides, or long tides. A similar superposition also exists at the time of full moon.

When the sun and moon are at 90° with respect to each other, the effect is one of cancellation as shown in Fig. 4. This configuration results in neap tides, also know as short tides.

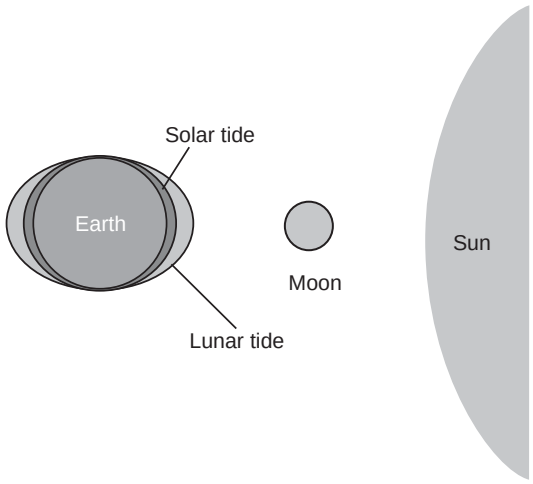


FIGURE 3 Earth, sun, and moon during spring tides.

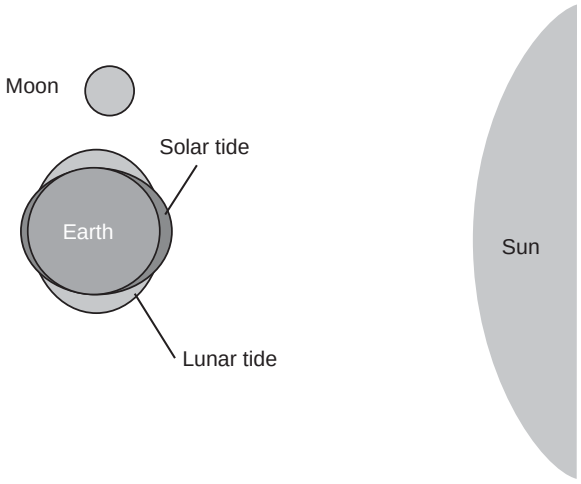


FIGURE 4 Earth, sun, and moon during neap tides.

2.4 The Presence of Land and the Resulting Tidal Dynamics

If the earth were covered entirely by water of a constant depth, the equilibrium theory of tides would give a perfectly reasonable description of water behavior. Fortunately, the oceans are not all of a constant depth and the presence of continents and islands severely influences the behavior of the oceans under tidal influences. The Coriolis force is a particularly important effect which, in the Northern Hemisphere, diverts moving objects to the right and, in the Southern Hemisphere, diverts moving objects to the left. The influence of this force, in the presence of land, can be considered, as shown in Fig. 5, by visualizing water flowing into and out of a semi-enclosed basin in the Northern Hemisphere under the influence of tidal effects. On the way into the channel, the water is diverted to the right toward the lower boundary. When the tidal forcing is reversed, the water is diverted toward the upper boundary. This results in a substantially higher tidal range at the basin boundaries than at the center.

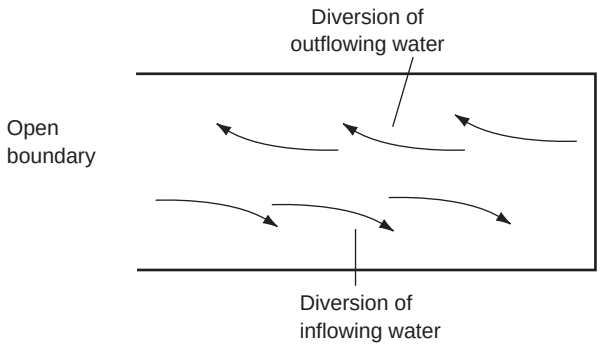


FIGURE 5 Flow of water into and out of a semi-enclosed basin in the Northern Hemisphere.

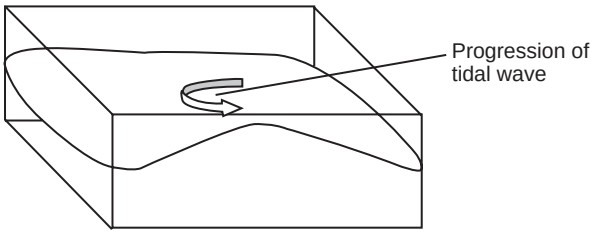


FIGURE 6 Progression of a tidal wave in a basin.

The net result of this effect is to generate a so-called tidal wave, which processes anti-clockwise around a point in the center of the "basin," as shown in Fig. 6.

In effect, the tides represent the terrestrial manifestation of the potential and kinetic energy fluxes present in the earth-moon-sun system. These fluxes are complicated by the presence of continents and other landmasses, which modify the form and phase of the tidal wave. As a result, substantially higher local fluxes occur in some regions of the world than others. The Bay of Fundy in Canada and the Bristol Channel between England and Wales are two particularly noteworthy examples of high-flux regions.

2.5 Global Flux

It has been estimated that the total energy from the tides, which is currently dissipated through friction and drag, is equivalent to 3000 GW of thermal energy worldwide. Much of this power is in inaccessible places, but up to 1000 GW is available

in relatively shallow coastal regions. Estimates of the achievable worldwide electrical power capability range from about 120 GW of rated capacity to approaching 400 GW. This is obviously a substantial energy resource, the significance of which has yet to be fully appreciated. Many enthusiasts believe these to be hopelessly pessimistic estimates. It is probably reasonable, however, to consider these estimates as representing, with an element of uncertainty, what could be exploited now using available technology.

3. TIDAL BARRAGE METHODS

3. History

There are many places in the world in which local geography results in particularly large tidal ranges. Sites of particular interest include the Bay of Fundy in Canada, which has a mean tidal range of 10 m, the Severn Estuary between England and Wales, with a mean tidal range of 8 m and Northern France with a mean range of 7 m. A tidal barrage power plant has been operating at La Rance in Brittany since 1966. This plant, which is capable of generating 240 MW, incorporates a road crossing of the estuary. Other operational barrage sites are at Annapolis Royal in Nova Scotia (18 MW), the Bay of Kislaya near Murmansk (400 kW), and at Jangxia Creek in the East China Sea (500 kW). Schemes for energy recovery, not surprisingly, have been proposed for the Bay of Fundy and for the Severn Estuary but have never been built.

3.2 Principles of Operation

3.2.1 The Role of the Barrage

The approach is essentially always the same. An estuary or bay with a large natural tidal range is identified and then artificially enclosed with a barrage. This would, typically, also provide a road or rail crossing of the gap in order to maximize the economic benefit. Electrical energy is produced by allowing water to flow from one side of the barrage through low head turbines, as shown in Fig. 7.

There are a variety of suggested modes of operation. These can be broken down initially into single basin schemes and multiple basin schemes. The simplest of these are the single basin schemes.

3.2.2 Single Basin Tidal Barrage Schemes

These schemes, as the name implies, require a single barrage across the estuary, as shown in Fig. 8. There

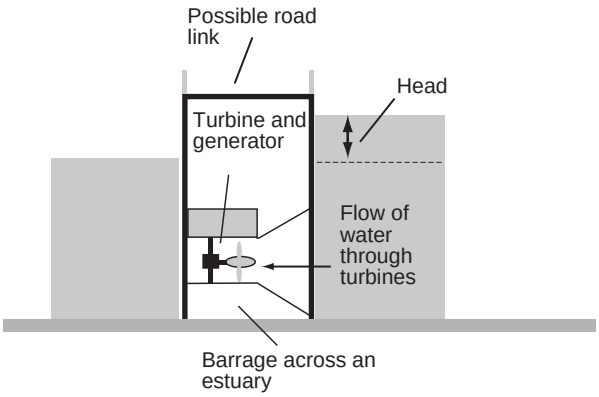


FIGURE 7 Schematic of a tidal power barrage.

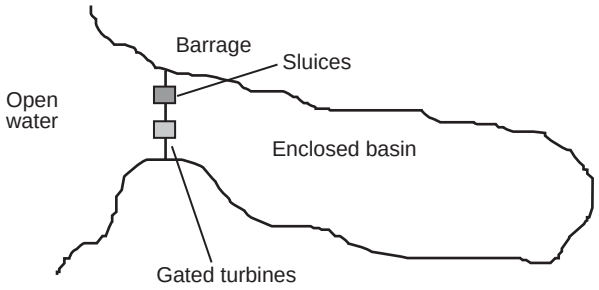


FIGURE 8 Schematic of the single basin generation scheme.

are, however, three different methods of generating electricity with a single basin. All of the options involve a combination of sluices which, when open, can allow water to flow relatively freely through the barrage and gated turbines, the gates of which can be opened to allow water to flow through the turbines to generate electricity.

3.2.3 Ebb Generation

During the flood tide, incoming water is allowed to flow freely through sluices into the barrage. At high tide, the sluices are closed and water is retained behind the barrage. When the water outside the barrage has fallen sufficiently to establish a substantial head between the basin and the open water, the basin water is allowed to flow out though low-head turbines and to generate electricity.

The system may be considered to operate in a series of steps as shown in Fig. 9.

These steps can be represented, as in Fig. 10, to show the periods of generation associated with stages in the tidal cycle.

Typically the water will only be allowed to flow through the turbines once the head is approximately

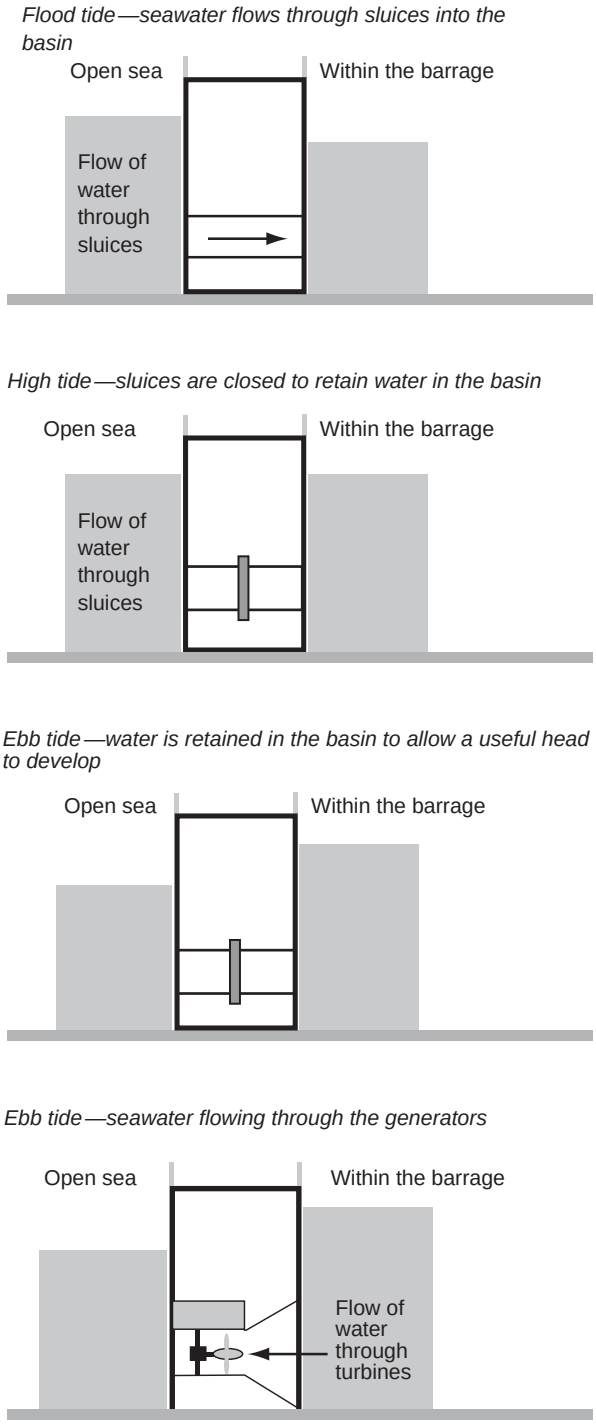


FIGURE 9 Operational steps in an ebb generation barrage scheme. (A) Ebb tide—water is retained in the basin to allow a useful head to develop. (B) Ebb tide—seawater flowing through the generators.

half the tidal range. It is likely that an ebb generation system will be able to generate electricity for, at most, 40% of the tidal cycle.

3.2.4 Flood Generation

The sluices and turbine gates are kept closed during the flood tide to allow the water level to build up outside of the barrage. As with ebb generation, once a sufficient head has been established, the turbine gates are opened and water can, in this case, flow into the basin and generate electricity, as Fig. 11 shows.

This approach is generally viewed as being less favorable than the ebb method because keeping a tidal basin at low tide for extended periods could have detrimental effects on the environment and shipping. In addition, the energy produced would be reduced as the surface area of a basin would be larger at high tide than at low tide, which would result in a rapid reductions in the head during the early stages of the generating cycle.

3.3.5 Two-Way Generation

In this mode of operation, use is made of both the flood and ebb phases of the tide. Near the end of the flood-generation period, the sluices would be opened to allow water to get behind the barrage and would then be closed. When the level on the open water side of the barrage had dropped sufficiently, water would be released through the turbines in the ebb generation mode, as shown in Fig. 12.

Unfortunately computer models do not indicate that there would be a major increase in the energy production. In addition, additional expenses would be associated with a requirement for either two-way turbines or a double set to handle the two-way flow. Advantages include, however, a reduced period with no generation and the peak power would be lower, allowing a reduction in the cost of the generators.

3.3.6 Double-Basin Systems

All single-basin systems suffer from the disadvantage that they only deliver energy during part of the tidal cycle and cannot adjust their delivery period to match the requirements of consumers. Double basin systems, such as that shown in Fig. 13, have been proposed to allow an element of storage and to give time control over power output levels.

The main basin would behave essentially like an ebb generation single basin system. A proportion of the electricity generated during the ebb phase would be used to pump water to and from the second basin to ensure that there would always by a generation capability. The cycle can be seen in Fig. 14.

It is anticipated that multiple basin systems are unlikely to become popular, as the efficiency of low-head turbines is likely to be too low to enable effective economic storage of energy. The overall

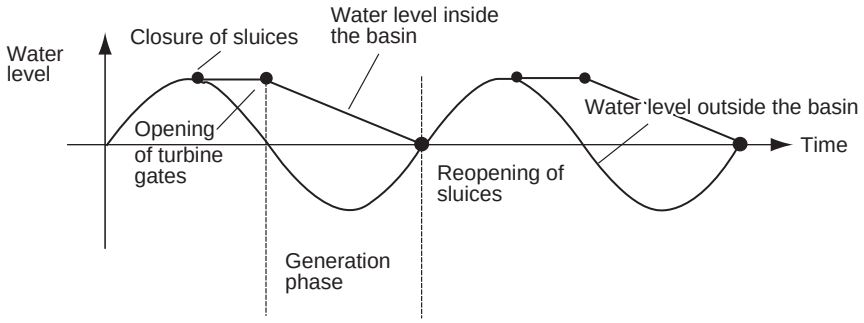


FIGURE 10 Water level inside and outside of an ebb generation barrage.

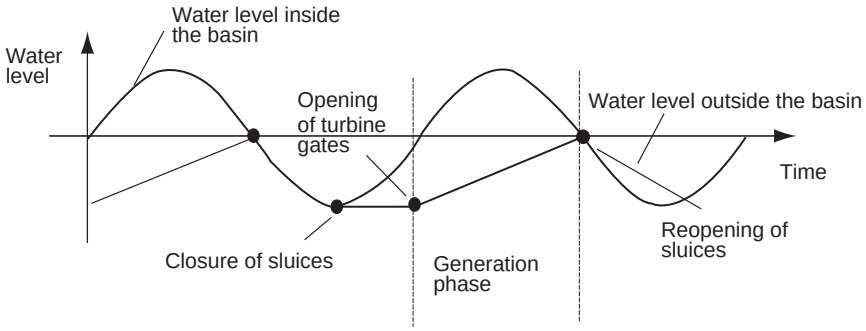


FIGURE 11 Water level inside and outside of a flood generation barrage.

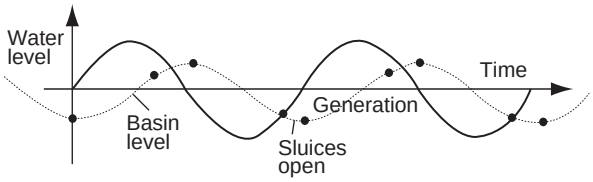


FIGURE 12 Water level inside and outside of a two-way generation barrage.

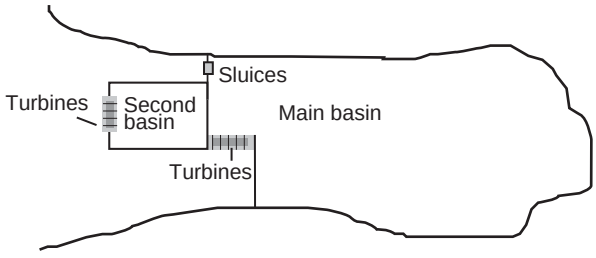


FIGURE 13 Schematic of a double basin generation system.

efficiency of such low-head storage, in terms of energy out and energy in, is unlikely to exceed 30%. It is more likely that conventional pump-storage systems will be utilized. The overall efficiencies of these systems can exceed 70%, which is, especially considering that this is a proven technology, likely to prove more financially attractive.

3.3 Optimal Design of Barrage Systems

This is not a simple procedure, as the construction of a barrage will inevitably change the nature of the tidal environment. It is theoretically possible, for example, for the tidal range to be very much less once a barrage has been constructed than it was prior to construction. To avoid an expensive mistake, it is necessary that reliable computer models be used to predict future behavior as much as possible. Such models are available and can be used with a great degree of confidence.

3.4 Locations for Tidal Barrage Developments

A considerable number of sites worldwide are technically suitable for development, although whether the resource can be developed economically is yet to be conclusively determined. Some of these are listed in Table I.

3.5 The Environmental Impact of Basin Systems

The construction of a barrage across a tidal basin will change the environment considerably. An ebb

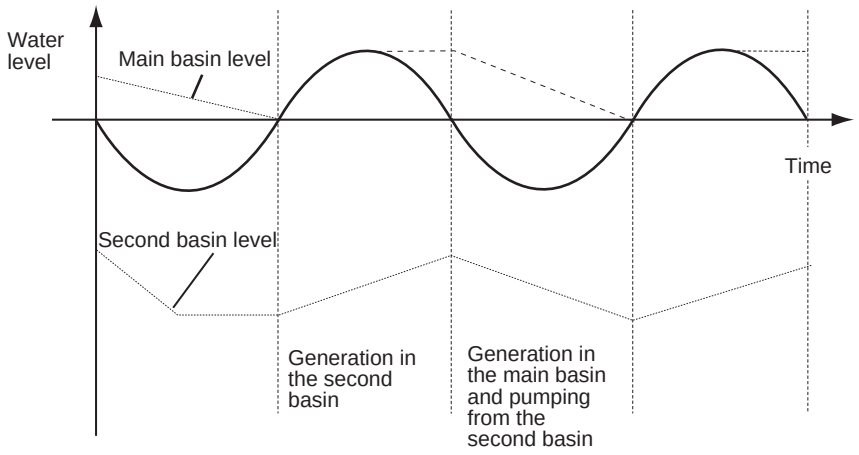


FIGURE 14 Water levels associated with a two-basin generation system.

TABLE I

Resource Estimates for Selected Sites

Site	Mean tidal range (m)	Barrage length (m)	Estimated annual energy production (GWhr)
Severn Estuary (UK)	7.0	17,000	12,900
Solway Firth (UK)	5.5	30,000	10,050
Bay of Fundy (Canada)	11.7	8000	11,700
Gulf of Khambat (India)	6.1	25,000	16,400

system will reduce the time tidal sands are uncovered. This change will have considerable influences on the lives of wading birds and other creatures. The presence of a barrage will also influence maritime traffic, and it will always be necessary to include locks to allow vessels to pass through the barrage. This will be much easier for an ebb system, where the basin is potentially kept at a higher level, than it would be with a flood-generation system, in which the basin would be kept at a lower than natural level. The environmental impact from increasing the area of exposed sediment in an estuarial environment is difficult to quantify but likely to be detrimental.

3.6 The Financial Implications of Tidal Barrage Development

Barrage schemes can produce a vast amount of energy. It has been suggested that the Severn Estuary could provide in excess of 8% of the United Kingdom's requirement for electrical energy. However, all barrage schemes are constrained by the massive engineering operations and associated financial investment that would be required. Construction would take several years. La Rance, for example, took 6 years, and no electricity was generated before the total project was completed. This is a major disincentive for commercial investment.

4. TIDAL CURRENT SYSTEMS

4.1 The Resource

The public perception of tidal power has traditionally been of large barrage schemes such as that at La Rance or proposed for the Severn Estuary. In addition to changes in the sea-surface level, however, the tides also generate water currents. In the open ocean, these currents are typically very small and are measured in cm/s at most. Local geographical effects can, however, result in quite massive local current speeds. In the Pentland Firth, for example, there is evidence of tidal currents exceeding 7 m/s. The kinetic energy in such a flow is considerable. Indeed many experts describe the Pentland Firth as the Saudi Arabia of tidal current power, which explains the considerable interest in this site, which has been expressed by the UK and Scottish governments. At present, however, there is no available technology to exploit this massively energetic but hostile channel.

Other sites, in Europe alone, with large currents, include the Channel Islands and The Straits of Messina. In addition to major sites such as the Pentland Firth, numerous local sites experience rapid currents capable of generating electricity with suitable technology.

Since the late 1980s, interest in the possibility of using tidal currents to generate electricity has developed enormously. A study prepared for the Scottish government in 2001 suggested that the UK "base case" potential could exceed 33.5 TWhr per annum at reasonable cost and using available technology. An earlier study for the European Union reported in 1996 that the European Union resource should exceed 48 TWhr. This figure appears low in comparison with the later UK figure, and it is suggested that, rather than the European resource outside of the United Kingdom representing only 14.5 TWhr per annum potential, the state of knowledge and understanding had increased dramatically in the 5 years separating these studies. There has not yet been an authoritative study of the achievable world resource but, drawing parallels with UK and European studies, it is anticipated that the achievable world resource could exceed 270 TWhr per annum using technology that is already under consideration. If problems associated with harnessing energy from currents in deep water can be overcome, then this figure might prove to be highly pessimistic, especially if technological developments make the development of sites such as the Pentland Firth feasible.

Additional studies have suggested that tidal currents could offer particular opportunities for peripheral rural regions. No one doubts the size of the worldwide potential of the tidal current resource but, unlike tidal barrage power, there has not yet been a large-scale prototype study. It is anticipated that this will soon be remedied.

4.2 Extracting Energy from Tidal Currents

The physics of the conversion of energy from tidal currents is similar, in principle, to the conversion of kinetic energy in the wind. Many of the proposed devices have, therefore, an inevitable, though superficial, resemblance to wind turbines.

There is no total agreement on the form and geometry of the conversion technology itself. Wind power systems are almost entirely horizontal axis rotating turbines, as shown schematically in Fig. 15. In these systems, the axis of rotation is parallel to the direction of the current flow.

In horizontal axis designs, the rotational axis is parallel to the direction of the water flow. Many developers favor this geometry for tidal conversion. Vertical axis systems, such as that shown schematically in Fig. 16, in which the axis of rotation is perpendicular to the direction of current flow, have not been rejected, however.

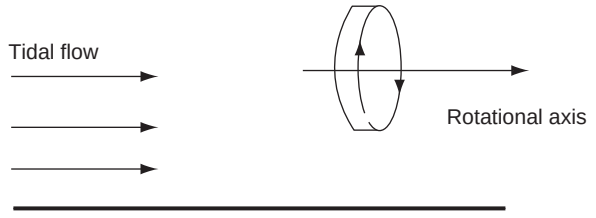


FIGURE 15 Schematic of a horizontal axis tidal current turbine.

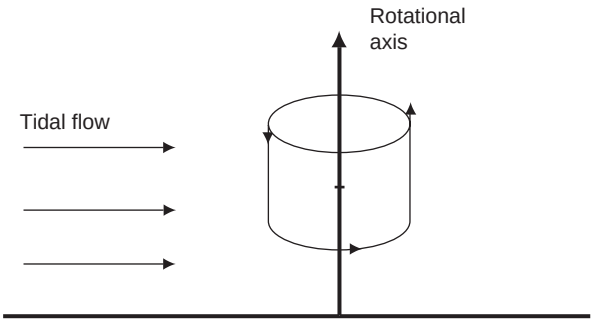


FIGURE 16 Schematic of a vertical axis tidal current turbine.

The environmental drag forces on any tidal current energy conversion system are large when compared with wind turbines of the same capacity. This poses additional challenges to the designer. Designs exist for devices that are rigidly attached to the seabed or are suspended from floating barges, as shown in Fig. 17. It is generally accepted that fixed systems will be most applicable to shallow water sites and moored systems for deep water. There may be exceptions to this, however.

The density of seawater, at approximately 1023 to 1025 kg/m^3, is substantially greater than that of air. This, combined with the rapid currents experienced in many sites across the world, suggests that very high-energy density fluxes will be encountered. Figure 18 shows the anticipated relationship between current speed and incoming energy flux density. Spring current speeds exceeding 3m/s are encountered in many locations, which suggests that the exploitable resource will be considerable. The challenge for the system designer is to ensure that the system can convert energy efficiently throughout as much of the tidal cycle as possible while being robust enough to survive the inevitable environmental loading that will be experienced. The difficulty of this task should not be overestimated. In practice, the efficiency of conversion of a practical system is likely to be limited to approximately 35%, but the potential for energy

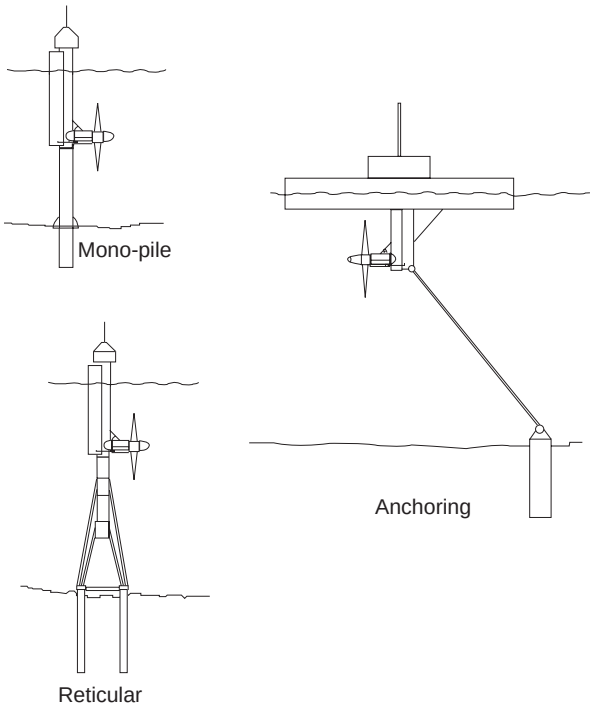


FIGURE 17 Alternative location systems for tidal current technology.

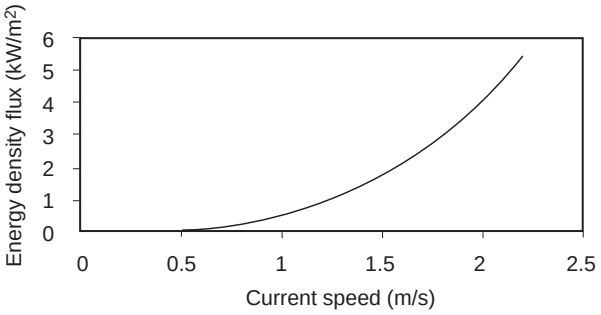


FIGURE 18 Influence of current speed on available energy flux density.

conversion in tidal current speeds, which might exceed 3 m/s in spring tides, is considerable.

The environment tidal devices will operate in is very different from that experienced by wind turbines, and there are some rather difficult problems associated with installation, survivability and maintenance that need to be solved before true commercial exploitation can be achieved. Many industrial, commercial, and public bodies have suggested that there is a high degree of synergy between the development of a tidal current generation industry and the offshore oil and gas industry. This offers the intriguing prospect of a new renewable industry developing in partnership with the petroleum industry and could, perhaps, result in accelerated development, due to the availability of expertise and technology, which would otherwise have to be developed from scratch.

Unlike the wind, tides are essentially predictable as they derive from astronomic processes discussed earlier in this article. Wind power systems are dependent on random atmospheric processes, which result in it being difficult to integrate large wind power developments into strategic electricity distribution networks. The predictability of the tides will make this integration much easier. Figure 19 shows the predicted power output for one design concept over a spring/neap tidal cycle, in which the spring peak current speed is 3 m/s and the neap peak current speed is 1.5 m/s. Although there would be some deviation from this prediction, as a result of weather related influences, the predictable nature can be clearly seen.

4.3 Testing the Technology for Tidal Current Energy Extraction

At the time of writing, there is no commercial generation of electricity from tidal currents anywhere

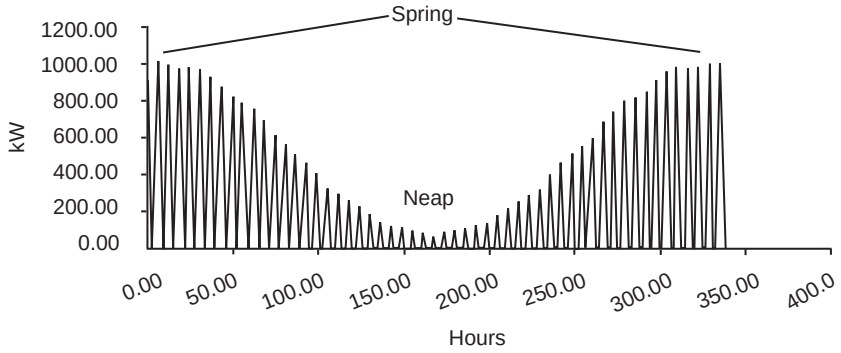


FIGURE 19 Predicted power output for a constant speed, 20-m diameter horizontal axis turbine.

FIGURE 20 Location of Lynmouth in the Bristol Channel.

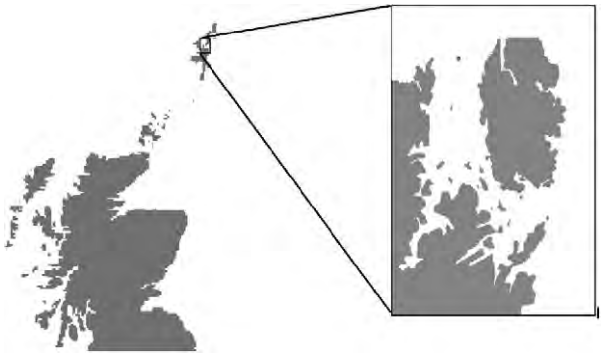


FIGURE 21 Location of Yell Sound in Shetland.

in the world. It is anticipated, however, that two large prototype systems will be installed in European waters in 2002. These devices—one of which will be located close to Lynmouth in the Bristol Channel, which lies between England and Wales, as shown in Fig. 20, and the other in Yell Sound in Shetland, as shown in Fig. 21—will be capable of producing approximately 300 kW each.

The two proposed systems are radically different in engineering concept, although both rely on the lift induced by fluid flow on an appropriately shaped lifting surface. The Lynmouth device, which is shown in Fig. 22, is a horizontal axis turbine, which resembles a traditional wind turbine.

The Yell Sound device, which is shown in Fig. 23, is a linear lift-based device, in which a reciprocating hydrofoil, which has a variable pitch angle, is used to pump oil, which then drives a hydraulic motor.

FIGURE 22 Artist's impression of the turbine being installed in the Bristol Channel. Illustration reproduced with the kind permission of Marine Current Turbines Ltd.

It is likely that, if tidal currents are to be commercially exploited, the generators will have to be mounted in clusters (tide farms?). If this is done, then, as with wind turbines, the devices will have to be sufficiently spread to ensure that the turbulence from individual devices does not interfere with others in the cluster. The nature of tidal currents, in which the most energetic flows are bidirectional in nature may make it possible to use relatively close spacing of devices perpendicular to the flow, with much larger spacing in the direction of the flow as seen in Fig. 24. This also shows one of the devices undergoing maintenance by a surface vessel.

4.4 Barriers to the Development of Tidal Current Generation

At present, there is little worldwide experience in harnessing tidal current energy. Experimental turbines have been used in Japan and the United Kingdom, but as yet they have never been used to generate electricity for commercial consumption. Scottish trials in Loch Linnhe, using a floating

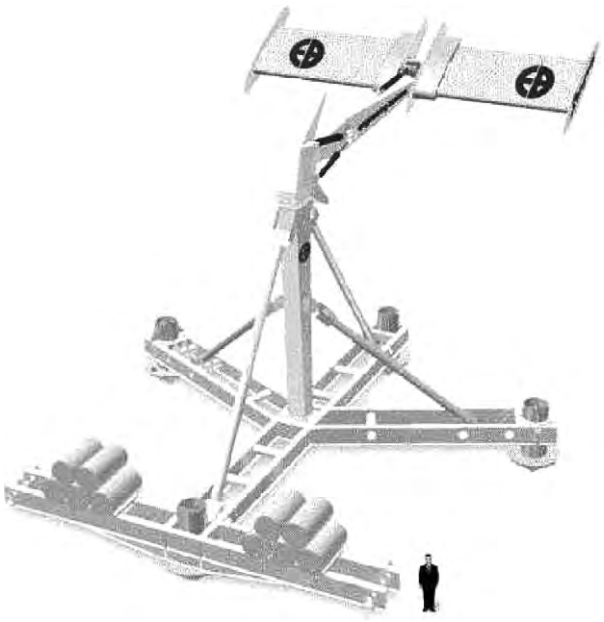


FIGURE 23 Artist's impression of the Stingray device being installed in Yell Sound. Illustration reproduced with the kind permission of The Engineering Business Ltd.

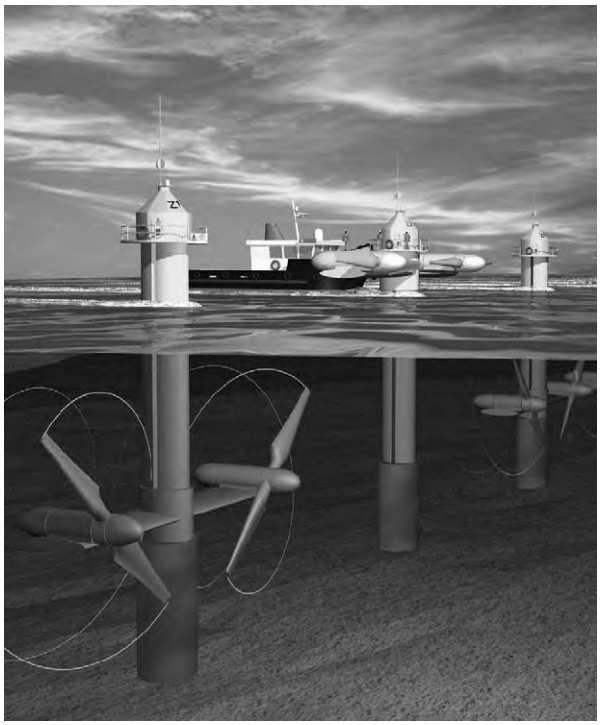


FIGURE 24 Artist's impression of a array of twin rotor tidal current devices. Illustration reproduced with the kind permission of Marine Current Turbines Ltd.

horizontal axis system, demonstrated that the technology is feasible. Experience in Japan has been largely with vertical axis systems. Until a large-scale device is commissioned and has demonstrated its performance for an extended period, there will be uncertainties about the long-term reliability and maintenance. Access to machines located in high energy tidal streams will always be problematic and could prove to require expensive maintenance procedures. Similarly, the rapid tidal currents themselves could provide difficulties in transferring energy via seabed cables to the coastline. Indeed some studies have suggested that the cable and its installation could be responsible for more than 20% of the total cost of a tidal current system. It is still too early to make reliable predictions of the cost of tidal current generation. There are still many uncertainties about the long-term reliability of systems and the difficulties associated with routine maintenance. The UK government's Department of Trade and Industry has indicated, however, that it anticipate that devices based on the Seaflow prototype should be able to generate electricity with a cost similar to onshore wind systems, assuming that problems associated with installation of multiple devices can be solved. Completion of the present program of prototype studies will allow more confidence to be given to predictions.

Environmental impacts from tidal current systems should be minimal, if projects are planned to take environmental issues into account. Energetic tidal channels do not tend to be home to many aquatic species and, although some species of marine mammals use tidal channels in their migration, the slow motion of the turbines and careful design should ensure minimal environmental impact. The most serious environmental impacts are likely to result from the extraction of kinetic energy itself, resulting in changes in the hydraulic nature of the flow environment. This could result in modification to the sediment transport systems and, of course, changes in the tidal currents, which are providing the energy. Many experts assume that up to 10% of the kinetic energy in a tidal flow can be extracted without serious flow degradation.

5. THE FUTURE OF TIDAL POWER

The high capital costs associated with tidal barrage systems are likely to restrict development of this resource in the near future. What developments

do proceed in the early 21st century will most likely be associated with road and rail crossings to maximize the economic benefit. In a future in which energy costs are likely to rise, assuming that low-cost nuclear fusion or other long term alternatives do not make an unexpectedly early arrivals, then tidal barrage schemes could prove to be a major provider of strategic energy in the late 21st century and beyond. Under some local conditions, small-scale barrages might also prove attractive. The technology for tidal barrage systems is already available, and there is no doubt, given the experience at La Rance, that the resource is substantial and available.

It is likely that tidal current systems will appear in experimental form in many places around the world. Initially these will be rated at around 300 kW, such as those planned for the Lynmouth and Yell Sound demonstrations. If these schemes prove successful, then the first truly commercial developments may appear soon. Systems may not presently have the strategic potential of barrage systems but, in the short term at least, they do offer opportunities for supplying energy in rural coastal and island communities. In the longer term, massive sites such as the Pentland Firth, which has been estimated to have the potential to support a 16-GW-capacity development, could become strategically important.

SEE ALSO THE FOLLOWING ARTICLES

Hydropower Resources • *Hydropower Technology* • *Ocean, Energy Flows in* • *Ocean Thermal Energy* • *Renewable Energy in Europe* • *Wave and Tidal Energy Conversion* • *Wind Energy Technology, Environmental Impacts of* • *Wind Resource Base*

Further Reading

Baker, C. (1991). "Tidal Power." Peter Peregrinus, Stevenage, UK.

Banal, M., and Bichon, A. (1991). "Tidal Energy in France. The Rance Tidal Power Station-Some Results After 15 Years in Operation." Proceedings of the Second International Symposium on Wave and Tidal Energy, Cambridge.

CENEX. (1993). "Tidal and Marine Currents Energy Exploitation. "European Commission, DGXII, JOU2-CT-93-0355.

Charlier, R. H. (1983). "Alternative Energy Sources for the Centralised Generation of Electricity." Adam Hilger, Bristol, UK.

Department of Energy (UK) (1989). "The Severn Barrage Project Summary: General Report." Energy Paper 57, HMSO.

Energy Technology Support Unit (ETSU) (1993). "Tidal Stream Energy Review" (ETSU T/05/00155/REP), DTI (UK).

Sodin, H. (2001). "Scotland's Renewable Resources 2001." Scottish Executive Publications, 285/GR/04/2001, Edinburgh.

Watson, W. (1992). Tidal power politics. *Int. J. Ambient Energy* **13**(3), 1–10.

Wilson, E. (1988). A tide in the affairs of men. *Int. J. Ambient Energy* **9**(3), 115–117.

Trade in Energy and Energy Services

GEERT VAN CALSTER
University of Leuven
Leuven, Belgium

1. **The World Trade Organization Agreement**
2. **The Issues Involved and GATT/WTO Membership**
3. **Dual Pricing**
4. **Tariffs**
5. **Renewable Energy, Including Tax Differentiation for Renewable Forms of Energy and the Impact of the Kyoto Protocol**
6. **Trade in Energy Services**

Glossary

border tax adjustment (BTA) A system in which traded goods are subject to a tax from the importing country but exempt from such a tax of the exporting country; e.g., a CO_2 emission tax.

dual pricing The fact of a given product having two different prices in different markets, especially a different domestic and export price.

most favored nation (MFN) A trade status extended by a country to a trading partner, allowing for the import of goods from the trading nation at the same tariff rate which the importing nation extends to its other trading partners that are part of the same agreement. In current usage the term normal trade relations (NTR) is often preferred because typically a nation will extend this status to all its trading partners except for a few rogue states.

national treatment The principle or policy that foreign products, services, and intellectual property should be given the same status or treatment they would have if they were domestic.

tariff A duty imposed on foreign goods imported into a country.

TRIPs An acronym for Trade-Related Intellectual Property Rights; i.e., the fact or policy of a nation's agreeing to provide protection for the rights to foreign-made intellectual property (innovative products of the human mind and imagination such as copyrights, trademarks, patents, industrial designs, and trade secrets).

Trade aspects of energy and energy services typically involve the following concerns: import and export promotion and/or protection, security of supply, and environmental protection. These concerns taken together obviously make for an interesting mix of often competing regulatory goals and hence do not contribute to creating a stable regulatory environment in which international trade can thrive. The World Trade Organization (WTO) is an obvious focal point for current and future liberalization in the area. However, other multilateral organizations have also had a role to play, including the North American Free Trade Association, the Asia-Pacific Economic Cooperation, the Energy Charter Treaty, and the European Union. This article focuses on the most multilateral framework of trade negotiations, the WTO, and on its impact with respect to trade in energy and energy services.

1. THE WORLD TRADE ORGANIZATION AGREEMENT

1.1 History

The World Trade Organization (WTO) has more than 130 members, accounting for more than 90% of world trade. More than 30 others are negotiating membership. The forerunner of the WTO is the 1947 General Agreement on Tariffs and Trade (GATT). GATT was the cornerstone of international trade law and trade negotiations for more than four decades. It only achieved this status after the collapse of the Agreement on an International Trade Organization (ITO). ITO had been prepared after World War II under the auspices of the United Nations (UN), particularly the UN Economic and Social Council, and the UN Conference on Trade and Development.

 © 2004 Elsevier Inc. All rights reserved.

After it became clear that the U.S. Congress would not ratify the ITO agreement, plans to have the ITO established were dropped, and GATT became the single most important international trade (law) instrument by default.

GATT's profile was at its highest at the time of negotiation of the eight successive so-called "rounds" of tariff negotiations. These each led to new "schedules" of tariff concessions, in which the contracting parties (as the states adhering to GATT were known) individually agreed to cut tariffs on specified products, according to the agreed timetable. The last of these rounds, the Uruguay Round, led to the creation of the WTO by the 1994 Marrakech Agreement. GATT 1947 was integrated into the WTO agreement and, together with the GATT 1994 agreement, forms the backbone of international trade in goods.

Decisions in the WTO are made by the entire membership. This is typically by consensus. A majority vote is also possible, but it has never been used in the WTO and was extremely rare under the WTO's predecessor, GATT. The WTO's agreements have been ratified in all members' parliaments.

1.2 WTO Institutions

The WTO's top-level decision-making body is the Ministerial Conference, which meets at least once every 2 years. Below this is the General Council (normally ambassadors and heads of delegation in Geneva, but sometimes officials sent from members' capitals), which meets several times a year at the Geneva headquarters. The General Council also meets as the Trade Policy Review Body and the Dispute Settlement Body. At the next level, the Goods Council, Services Council, and Intellectual Property [General Agreement on Trade-Related Aspects of Intellectual Property (TRIPs)] Council, report to the General Council. Numerous specialized committees, working groups, and working parties deal with the individual agreements and other areas, such as the environment, development, membership applications, and regional trade agreements.

The first Ministerial Conference in Singapore in 1996 added three new working groups to this structure. They deal with the relationship between trade and investment, the interaction between trade and competition policy, and transparency in government procurement. At the second Ministerial Conference in Geneva in 1998, ministers decided that the WTO would also study the area of electronic commerce, a task to be shared among existing councils and committees.

The WTO Secretariat's main duties are to supply technical support for the various councils and committees and the ministerial conferences, to provide technical assistance for developing countries, to analyze world trade, and to explain WTO affairs to the public and media. The Secretariat also provides some forms of legal assistance in the dispute settlement process and advises governments wishing to become members of the WTO. It is based in Geneva.

1.3 WTO Agreements

The 1986–1994 Uruguay Round negotiations, as noted previously, included a major revision of the original GATT. GATT is now the WTO's principal rulebook for trade in goods. The Uruguay Round also created new rules for dealing with trade in services [the General Agreement on Trade in Services (GATS)], relevant aspects of intellectual property (the TRIPs agreement, dispute settlement, and trade policy reviews. The complete set consists of approximately 60 agreements (including on public procurement), and separate commitments (called schedules), made by individual members in specific areas, such as lower customs duty rates and services market opening.

The Marrakech Agreement is the single most exciting development in international trade law in the past century. The creation of the agreements on intellectual property and services is the most obvious development. This article does not discuss the details of each of these agreements. One agreement is developed later, namely the Understanding on Rules and Procedures Governing the Settlement of Disputes.

The structure of the agreement establishing the WTO, corresponds with its institutional set-up. The issues involved in the liberalization of the energy sector run across a wide variety of these agreements. Hence, there is a *priori* no single WTO agreement whose relevance for the energy sector can summarily be dismissed.

Annex 1 to the agreement has three subannexes. Annex 1A reads "Multilateral agreements on trade in goods." At its core lies the GATT 1994, which comprises the unamended GATT 1947, protocols and decisions adopted under the GATT 1947 structure (typically relating to tariff concessions and waivers), and a number of "understandings" with respect to specific GATT articles and/or concepts. The GATT 1947 continues to include the basic principles of the international trade in goods: most favored nation treatment (MFN; i.e., making

no distinction, for the treatment of imported goods, between their origin), national treatment (i.e., treating imported goods like goods of national origin), and, generally, nondiscrimination. Importantly, one cannot gauge the degree of liberalization of the trade in a given category of goods without consulting the specific tariff "commitments" of each individual member of the WTO. These are the lengthy documents, attached to the WTO agreement, in which members commit to either reducing or eliminating tariffs levied on the goods specified in them.

Other agreements within Annex 1A deal with specific sectors, particularly the Agreement on Agriculture, which remains a focus for further liberalization, and the Agreement on Textiles and Clothing. It is noteworthy in this respect that agriculture and textiles are domains in which the interests of the developed world are often pitched against those of the developing countries.

Finally, Annex 1A covers a number of agreements that deal with specific issues. Reference can be made in this respect to the Agreement on the Application of Sanitary and Phytosanitary Measures, the Agreement on Technical Barriers to Trade, the Agreement on Trade-Related Investment Measures, the Agreement on Subsidies and Countervailing Measures, and the Agreement on Rules of Origin, i.e., these agreements deal with specific sectors.

Annex 1B contains the GATS. The inclusion, during the Uruguay Round negotiations, of services in international trade liberalization is widely acknowledged as being one of the most important developments brought about by the WTO. GATS' structure is not unlike the structure of Annex 1A, although there are important differences. As in Annex 1A, GATS operates on three platforms: a main text containing general principles and obligations; specialized agreements dealing with specific sectors; and the commitments of individual countries, offering specified access to their markets in specified services. Importantly, however, GATS also includes a list of exemptions, temporally allowing members not to apply MFN in specified sectors. These arrangements in particular are scheduled to be reviewed. Negotiations to this effect were initiated (at a slow pace, following the impasse of the Seattle Millennium Round) in 2000.

The GATS agreement is not limited in scope. It covers all service sectors, typically divided into four categories [Article I(2)]: cross-border supply (services supplied from the territory of one member into the territory of another member); consumption abroad (services supplied in the territory of one member to the service consumer of another member); commercial presence [services supplied by a service supplier of one member through the commercial presence in the territory of another member; in the equivalent European Community (EC) jargon, this is referred to as the freedom of establishment], and the presence of natural persons (services supplied by a service supplier of one member through the presence of natural persons of a member in the territory of any other member; in the equivalent EC jargon, this is referred to as free movement of services).

Importantly, Article II of GATS contains the MFN principle, specifying, however, that a member may maintain measures inconsistent with MFN provided they are included in the annex on MFN exemptions. GATS also contains provisions with respect to transparency, mutual recognition of domestic regulations, reasonableness of such domestic regulations, and progressive liberalization (including the requirement that further talks be initiated in 2000).

It is noteworthy that for four sectors negotiations resumed immediately after the conclusion of the Uruguay Round: basic telecommunications (for which new national commitments took effect in January 1998), maritime transport (further negotiations did not produce any more results, and these are now included in the general GATS review), movement of natural persons (only modest results were achieved over and above what was already included in members' schedules of commitment), and financial services (leading to further commitments at the end of 1997).

Annex 1C is the TRIPs agreement. This sector is influenced by the fact that there are quite a number of international treaties outside the WTO context that protect intellectual property. Consequently, the WTO has to find its place among all these agreements, apart from trying to devise something of a global approach to intellectual property issues.

The TRIPs agreement deals with four main areas:

- The General Provisions and Basic Principles emphasizes national treatment and MFN. In contrast with GATS, MFN is not in principle subject to exemptions in the TRIPs context. Among the general principles, the agreement also recognizes that intellectual property protection needs to take place within its societal context. Article 8 therefore provides that intellectual property protection must not stand in the way of in particular public health law and the protection of socioeconomic and technological development, provided a balance is struck.

• Part II of the agreement, Standards Concerning the Availability, Scope and Use of Intellectual Property Rights, details the minimum standards that members must meet in offering intellectual property protection. The starting point here was two agreements originating in the World Intellectual Property Organization, whose standards are expanded upon by the TRIPs agreement. To this end, it specifies the level of protection that has to be achieved for different categories of intellectual property rights: copyright and related rights, trademarks, geographical indications, industrial designs, patents, layout designs of integrated circuits, the protection of undisclosed (business) information, and avoiding anticompetitive effects of licenses distributed by the intellectual property holder.

• Part III of the agreement includes the standards to which members must adhere when enforcing their intellectual property rights.

• Part V provides for dispute prevention and settlement.

The trade policy review mechanism is provided for in Annex 3. It serves the goal of transparency in members' trade policies. It enables the regular and collective appreciation of the full range of members' trade policies and practices, thus facilitating progress toward the implementation of the agreements and general understanding of each other's policies. Although the review undoubtedly may lead to the identification of apparent shortcomings in WTO obligations, this is not the review mechanism's goal.

Finally, Annex 4 of the WTO agreement originally contained four but currently contains only two so-called plurilateral trade agreements (i.e., agreements that are not adhered to by all WTO members). These agreements effectively were negotiated during the Tokyo Round of international trade negotiations but, unlike most of the products of that area, they were not integrated into the WTO. Both the International Dairy Agreement and the International Bovine Meat Agreement were terminated in 1997. The Agreement on Trade in Civil Aircraft and the Agreement on Government Procurement remain in force.

2. THE ISSUES INVOLVED AND GATT/WTO MEMBERSHIP

As noted previously, there is a priori no single WTO agreement whose relevance for the energy sector can summarily be dismissed. The following issues all weigh heavily both in the negotiations leading to accession of new members and in the agenda of the WTO:

• Governmental controls on production and exports of petroleum-based products, including in the framework of the Organization of Petroleum Exporting Countries (OPEC)

• Domestic prices and pricing policy; this refers *inter alia* to the issue of dual pricing, a practice according to which petroleum-exporting countries keep energy prices at the domestic level at a much lower rate than the price at which the goods are being exported

• Export tariffs and taxation

• Operations of state trading enterprises and monopolistic practices in the sector

• Unfair trade practices, such as subsidies and dumping

• Investment

• Trade in services related to the exploration, extraction, transportation, and processing of petroleum

At the risk of oversimplifying, there is a lot of truth in the argument that in light of this extensive list of issues that need to be discussed and may lead to further liberalization, quite a few energy-producing countries (particularly petroleum-exporting countries) are attracted to the argument that it may be worth their while to stay out of the WTO (and its predecessor, GATT) so as not to have to put up as it were with WTO interference in their domestic energy policies. However, in recent years, the balance for all but a few petroleum-exporting countries seems to have tilted in favor of joining the WTO, particularly because the WTO deals with many more than merely energy issues. Tables I and II give an overview of OPEC and non-OPEC countries and their membership in GATT and WTO.

3. DUAL PRICING

Dual pricing is a generic term for measures to ensure that domestic prices are lower, or export prices higher, than would be achieved by market forces. This raises the prospect of export restrictions falling foul of either GATT Article XI, which prohibits quantitative restrictions to trade (this would be the case in the event of straightforward trade restrictions for export), or GATT Article III, which prohibits discriminatory tax practices that impact on trade.

GATT Article XI states,

> *1. No prohibitions or restrictions other than duties, taxes, or other charges, whether made effective through quotas,*

TABLE I

OPEC Member Countries and Their Status in GATT and WTO as of December 1999[a]

Country[b]	Status in GATT	Status in WTO
Algeria	Applying to GATT on a *de facto* basis since independence (July 3, 1962)	In the process of accession since June 1987
Indonesia	February 24, 1950	January 1, 1995
Iran. Islamic Republic of[c,d]	Observer	None
Iraq[c,d]	Observer	None
Kuwait[c]	May 3, 1963	January 1, 1995
Libyan Arab Jamahiriya	Observer	None
Nigeria	November 18, 1960	January 1, 1995
Qatar	April 7, 1994	January 13, 1996
Saudi Arabia[c]	Observer	In the process of accession since July 1993
United Arab Emirates	March 8, 1994	April 10, 1996
Venezuela[c]	August 31, 1990	January 1, 1995

[a] *Source*. United Nations Conference on Trade and Development (2000).

[b] Algeria is an observer government to the WTO. Algeria's working party was established on June 17, 1987, and met for the first time in April 1998; the fifth meeting of the working party took place on May 20, 2003. Saudi Arabia is an observer government to the WTO; the working party on the accession of Saudi Arabia was established July 21, 1993.

[c] Founder member of OPEC.

[d] Signed the Havana Final Act but did not become a GATT contracting party.

import or export licenses, or other measures, shall be instituted or maintained by any contracting party on the importation of any product of the territory of any other contracting party or on the exportation or sale for export of any product destined for the territory of any other contracting party.

Article XI supplements the opening of the world's market, as agreed in tariff and other concessions. Article XI applies to all measures instituted or maintained by a member prohibiting or restricting the importation, exportation, or sale for export of products, other than measures that take the form of duties, taxes, or charges (these are covered by Article III).GATT Article III states,

1. The contracting parties recognize that internal taxes and other internal charges, and laws, regulations, and requirements affecting the internal sale, offering for sale, purchase, transportation, distribution, or use of products, and internal quantitative regulations requiring the mixture, processing, or use of products in specified amounts or proportions, should not be applied to imported or domestic products so as to afford protection to domestic production.

2. The products of the territory of any contracting party imported into the territory of any other contracting party shall not be subject, directly or indirectly, to internal taxes or other internal charges of any kind in excess of those applied, directly or indirectly, to like domestic products. Moreover, no contracting party shall otherwise apply internal taxes or other internal charges to imported or domestic products in a manner contrary to the principles set forth in paragraph 1.

The interpretative note Ad III:2 adds,

A tax conforming to the requirements of the first sentence of paragraph 2 would be considered to be inconsistent with the provisions of the second sentence only in cases where competition was involved between, on the one hand, the taxed product and, on the other hand, a directly competitive or substitutable product which was not similarly taxed.

Article III:4 reads,

The products of the territory of any contracting party imported into the territory of any other contracting party shall be accorded treatment no less favorable than that accorded to like products of national origin in respect of all laws, regulations, and requirements affecting their internal sale, offering for sale, purchase, transportation, distribution, or use. The provisions of this paragraph shall not prevent the application of differential internal transportation charges which are based exclusively on the economic operation of the means of transport and not on the nationality of the product.

The national treatment requirement of Article III complements the MFN principle contained in Article I and the schedules of tariff concessions that GATT

TABLE II

Non-OPEC Petroleum-Exporting Developing Countries and Economies in Transition and Their Status in GATT and WTO as of December 1999[a]

Country[b]	Status in GATT	Status in WTO
Angola	Applying to GATT on a *de facto* basis	December 1, 1996
Argentina	October 11, 1967	January 1, 1995
Azerbaijan	None	In the process of accession since July 1997
Brazil	July 29, 1948	January 1, 1995
Brunei Darussalam	December 9, 1993	January 1, 1995
Cameroon	January 1, 1960	December 13, 1995
China	GATT contracting party from 1947 to 1949	In the process of accession since March 1987
Colombia	October 3, 1981	April 30, 1995
Congo	August 15, 1960	March 27, 1997
Ecuador	Observer	January 21, 1996
Egypt	May 9, 1970	June 30, 1995
Gabon	May 3, 1963	January 1, 1995
India	July 8, 1948	January 1, 1995
Kazakhstan	None	In the process of accession since February 1996
Malaysia	August 31, 1957	January 1, 1995
Mexico	August 24, 1986	January 1, 1995
Oman	None	In the proccess of accession since June 1996
Papua New Guinea	Applying GATT on a *de facto* basis since independence (September 16, 1975)	June 9, 1996
Peru	October 7, 1951	January 1, 1995
Russian Federation	Observer	In the process of accession since June 1993
Syrian Arab Republic	Contracting party from 1947 to 1950	None
Trinidad and Tobago	August 31, 1962	March 1, 1995
Tunisia	August 19, 1990	March 29, 1995
Uzbekistan	None	In the process of accession since December 1994
Vietnam	None	In the process of accession since January 1995
Yemen	Applying GATT on a *de facto* basis since independence (November 30, 1967)	Observer

[a] *Source*. United Nations Conference on Trade and Development (2000).

[b] Azerbaijan's working party was established on July l6, 1997; the first meeting of the working party was held on June 7, 2002. China joined the WTO on December 11, 2001. Kazakhstan's working party was established on February 6, 1996; the working party last met in December 2002. Oman became a member on November 9, 2002. The working party on the accession of the Russian Federation was established on June 16, 1993; the next working party meetings are scheduled for November 2003. The working party on the accession of Uzbekistan was established on December 21, 1994; the first meeting of the working party took place on July 17, 2002. The working party on the accession of Vietnam was established on January 31, 1995; the next meeting of the working party may take place in late 2003. Yemen's request for accession was circulated on April 14, 2000; the working party has not met.

members adhere to according to Article II. The broad and fundamental purpose of Article III is to avoid protectionism in the application of internal tax and regulatory measures.

Although theoretically, dual pricing could be dealt with under standard GATT law in particular, a number of countries, including those practicing dual pricing, but also the European Union, argue that the issue deserves a more *sui generis* approach, taking account of the various interests at stake.

4. TARIFFS

The WTO, like its predecessor the GATT, endeavors to reduce tariffs throughout the various sectors covered by the WTO agreements. Tariff-reduction negotiations continue to consume a considerable amount of time and effort. However, tariffs are the WTO's "preferred" obstacle to trade, in the sense that tariffs are a much more transparent trade instrument than other, such as technical, barriers to

trade or indeed certain types of subsidies, such as tax exemptions and tax privileges.

Tariffs on actual energy products in the main importing trade blocks are by and large limited and did not figure in the Uruguay Round negotiations (i.e., the round of trade negotiations that led to the creation of the WTO). This is understandable given the security of supply considerations by those importing nations. It is noteworthy that for petroleum products (e.g., plastics), tariffs are higher in general than those for crude oil.

Local taxation, particularly excise duties, may of course be very high, but in general it does not raise any issues under GATT Article III. Excise duties are generally applied across the board (i.e., without concern for favoring domestic production over imported energy products). To the extent that such discrimination does exist, GATT Article III would seem well equipped to deal with this issue. A development of note, however, is the policy of a number of importing countries to employ tax differentiation as a means of encouraging the development of renewable sources of energy at the expense of fossil fuels.

5. RENEWABLE ENERGY, INCLUDING TAX DIFFERENTIATION FOR RENEWABLE FORMS OF ENERGY AND THE IMPACT OF THE KYOTO PROTOCOL

Excise duties in particular, but also other types of taxes, in addition to raising revenue for the treasury are also designed to deal with externalities of trade. These are the societal costs of production and/or consumption, which are not reflected in the price of the product. By employing excise duties or other taxes, states can "internalize" these externalities in the cost of a product. A typical example in the energy sector is the use of taxation to phase out leaded gasoline.

Tax differentiation has become more of an issue in recent years with the advent of state interest in renewable forms of energy. To encourage the development and consumption of these types of energy, inevitably at the expense of fossil fuels, a number of Western European governments have enacted tax breaks and other forms of regulatory distinction. A 1991 European proposal may serve as a useful benchmark in this respect.

The proposal included a mix of taxes on products and on production processes and methods (PPMs). One of the main reasons for the EC initiative was to avoid a distortion of the single market by the existence of a variety of national ecotaxation schemes. Essentially, the proposal amounted to the following: The arrangements would have provided for the taxation of fossil sources of energy. Half of the total tax would have been based on CO_2 emissions, and the other component would have related to the calorific value of the energy sources, whether fossil or not, excluding renewables. The two-tiered structure gave it the name CO_2 and energy tax.

The tax was to be levied on all solid, viscous, or gaseous fossil fuels, either in their crude state (in which they are used as motor or heating fuel) or in their final form (where the fuel is obtained by processing the crude product). Raw materials were therefore not taxed. Importantly, rates of tax were also set for heat and electricity. Regarding the CO_2 component of the tax, heat and electricity produced from primary fossil sources of energy would not have been taxed as such but, rather, on the basis of the fossil inputs used.

The chargeable event would have been the extraction of the crude products, the production of their derivatives, or importation; in other words, the moment when the products are released for consumption. For energy-intensive firms, member states could have been authorized by the commission to grant specific graduated tax reductions. All firms could have been given tax incentives by the member states, reducing the amount of tax payable by the cost of new investments made in energy saving and CO_2 abatement.

Since the tax would clearly have taxed "like products" upon importation, and taxes would have been rebated upon export, border tax adjustment (BTA) was an essential part of the proposal.

5.1 BTA for the CO_2 and Energy Tax on the Import Side: GATT Compatible?

The very eligibility of the tax for BTA under GATT rules would have been a contested issue. Energy taxes are normally qualified as "taxes occultes," thus falling outside the scope for BTA in the classic GATT doctrine. However, energy has a dual nature because it is both a product and a fundamental part of most production. With respect to the levying of the tax on the products, BTA could have been allowed under Article II *juncto* Article III GATT. Previous panel findings (e.g., in the Superfund case) support the conclusion that the levying of the tax on

imported energy products would have passed the GATT hurdle. Both the calorific value of the products and the level of their CO_2 emissions upon combustion could be accepted as being related to physical characteristics of the products concerned. A distinction based on these two parameters would thus seem to be compatible with the GATT-approved methods of distinguishing between products. Importantly, the EC would not have distinguished between the origin of the products concerned, nor would

TABLE III

GATS, Commitments; Services Incidental to Energy Distribution[a]

Member	Sector	Market access limitations	National treatment limitations
Australia	Services incidental to energy distribution (887**) Covers consultancy services related to the transmission and distribution on a fee basis of electricity, gaseous fuels, and steam and hot water to household, industrial, commercial, and other users	1. None 2. None 3. None 4. Unbound except as indicated in the horizontal section	1. None 2. None 3. None 4. Unbound except as indicated in the horizontal section
Dominican Republic	Services incidental to energy distribution (887)	1. None 2. Unbound 3. None 4. Unbound	1. Unbound 2. Unbound 3. Unbound 4. Unbound
Gambia	Other business services (including CPC 887)	1. None 2. None 3. Unbound except as indicated in the horizontal section 4. Unbound except as indicated in the horizontal section	1. None 2. None 3. None 4. Unbound except as indicated in the horizontal section
Hungary	Consultancy services incidental to energy distribution (CPC 887)	1. None 2. None 3. None 4. Unbound except as indicated in part I	1. None 2. None 3. None 4. Unbound except as indicated in part I
Nicaragua	Services incidental to energy distribution (887)	1. None 2. Unbound 3. None 4. Unbound	1. None 2. Unbound 3. None 4. Unbound
Sierra Leone	Other business services	1. None 2. None 3. Foreign service providers are permitted to establish service companies or institutions with Sierra Leonean partners in the form of joint ventures 4. Unbound except as indicated in horizontal commitments	1. None 2. None 3. None 4. Unbound except as indicated in horizontal commitments
Slovenia	Services incidental to energy distribution—for gas only (CPC 887**)	1. None 2. None 3. None 4. Unbound except as indicated in the horizontal section	1. None 2. None 3. None 4. Unbound except as indicated in part I
United States	Services incidental to energy distribution	1. None 2. None 3. None 4. Unbound except as indicated in part I	1. None 2. None 3. None 4. None

[a] *Source*. World Trade Organization (1998).

TABLE IV

GATS Commitments; Pipeline Transportation of Fuels[a]

Member	Sector	Market Access Limitations	National Treatment Limitations
Australia	Pipeline transport—transportation of fuels (7131)	1. None 2. None 3. None 4. Unbound except as indicated in the horizontal section	1. None 2. None 3. None 4. Unbound except as indicated in the horizontal section
Hungary	Pipeline transport	1. Unbound 2. None 3. Services may be provided through a contract of concession granted by the state or the local authority 4. Unbound except as indicated in part I	1. None 2. None 3. None 4. Unbound except as indicated part I
New Zealand	Pipeline transport (713)	1. None 2. None 3. None 4. Unbound except as indicated in the horizontal section	1. None 2. None 3. None 4. Unbound except as indicated in the horizontal section

[a] *Source*. World Trade Organization (1998).

different rates have applied depending on origin. As previously noted, the use of the proceeds of the tax would have had no bearing on its GATT legality.

The proposal gave tax credits to firms which had invested in CO_2-reducing technology. The principle of nondiscrimination, embedded in GATT Article III, would have required the EC to extend any such exemption schemes to importing firms that made similar efforts. The proposal did not include any specifications to this effect, indicating that the EC did not envisage such a situation. The tax on the final products (namely heat and electricity) whose production involves the use of CO_2-releasing energy products is a classic case of a non-product-related PPM. It is uncertain how the EC would have organized BTA for these products. In any event, distinguishing between products' fiscal treatment, depending on the energy source that produced them, is precarious in light of GATT's "like product" distinction.

5.2 BTA for the CO_2 and Energy Tax on the Export Side: GATT Compatible?

GATT Article XVI suggests that where the rebate of taxes for products destined for export does not exceed the tax burden that the product has incurred, no subsidy is deemed to have been granted. Where subsidies are deemed to exist, countervailing measures are possible under GATT Article VI, provided the nation introducing them demonstrated material injury to its domestic industry as a result of the export subsidy.

Whether details of the 1991 proposal, or indeed any future EC plan for introducing energy taxes, would have to be considered export subsidies would have very much depended on the precise implementation of its provisions. The proposal provided for tax reduction and/or exemptions in the case of firms with a high energy consumption, tax reductions and refunds for environmental investments, and required the member states to ensure tax neutrality. Generally, there was a real danger that the imprecise nature of the provisions of Articles X and XI could have created preferential tax treatment for exporting industries to the detriment of the industry in the importing third countries.

Eventually, the proposed tax was dropped because no consensus could be reached in the Council of Ministers. However, with a large number of countries undertaking far-reaching commitments in the framework of the Kyoto Protocol, more countries eventually will have to resort to encouraging consumption of CO_2-neutral forms of energy. This may lead to systematic discrimination of fossil fuels, thereby distorting trade patterns and raising the stakes of energy negotiations in the framework of the WTO. This also very firmly links the issue of energy trade with the so-called trade and environment debate. Trade and environment is one of the areas in which trade blocks such as the European Union and the United States would like to see more

TABLE V

GATS Commitments: Other Energy-Related Commitments[a]

Member	Sector	Market access limitations	National treatment limitations
Canada	General construction work for civil engineering	1. None, other than: cabotage 2. None 3. None, other than: construction (Ontario)—an applicant and holder of a water power site development permit must be incorporated in Ontario 4. Unbound except as indicated in the horizontal section	1. None 2. None 3. None 4. Unbound except as indicated in the horizontal section and construction (Ontario)—an applicant and holder of a water power site development permit must be a resident of Ontario
	Power facilities and pipelines		
	Other: site preparation for mining (CPC 511*)	1. None, other than: cabotage 2. None 3. None 4. Unbound except as indicated in the horizontal section	1. None 2. None 3. None 4. Unbound except as indicated in the horizontal section
Colombia	Design, construction, operation, and maintenance of oil and gas pipelines	1. None 2. None 3. None 4. Unbound except as indicated in the horizontal section	1. None 2. None 3. None 4. Unbound except as indicated in the horizontal section
Côte d'Ivoire	Installation and assembly work	1. Unbound 2. Unbound	1. Unbound 2. Unbound
	Prospecting, mining, or processing of minerals	3. None 4. Unbound except for the entry and temporary stay of some categories of natural persons	3. Government approval, granted subject to preferential use of local services and employment and training of local executives and supervisors, required for enterprises
	Energy generation, excluding wholesale or retail services relating to equipment or parts thereof		4. None
Ecuador	Site preparation work for mining (CPC 5115)	1. None 2. None 3. None 4. Unbound except as indicated in the horizontal section	1. None 2. None 3. None 4. Unbound except as indicated in the horizontal section
Egypt	General construction work for civil engineering	1. Unbound* 2. Unbound* 3. Commercial presence only allowed for joint-venture companies; foreign capital equity shall not exceed 49% of the total capital required for the project 4. None	1. Unbound* 2. Unbound* 3. None 4. None
	Long-distance pipelines, communication and power lines (cables)		
	Installation work		
	Gas-fitting construction work		
India	Construction work for civil engineering	1. Unbound* 2. Unbound* 3. Only through incorporation with a foreign equity ceiling of 51% 4. Unbound except as indicated in the horizontal section	1. Unbound* 2. Unbound* 3. None 4. Unbound except as indicated in the horizontal section
	Construction of pipelines, power lines; constructions for mining, power plants, and coke ovens (ex CPC 513)		
	Research and development services on natural sciences	1. Unbound 2. Unbound* 3. Only through incorporation with a foreign equity ceiling of 51% 4. Unbound except as indicated in the horizontal section	1. Unbound 2. Unbound* 3. None 4. Unbound except as indicated in the horizontal section

continued

Table V continued

Member	Sector	Market access limitations	National treatment limitations
	Heat, light, electromagnetism, astronomy, but excluding atomic energy and related matters (CPC 85101)		
	Engineering and technology, including applied science and technology for casting, metal, machinery, electricity, communications, vessels, aircraft, civil engineering, construction, information, etc. (CPC 85103)		
Malaysia	Management consulting services covering advisory, guidance, and operational assistance services concerning management of the transmission of nonconventional energy (CPC8 650)	1. None 2. None 3. Only through a locally incorporated joint-venture corporation with Malaysian-controlled corporations or both and Bumiputera shareholding in the joint-venture corporation is at least 30% 4. Unbound except as indicated in 1(a) and (c) in the horizontal section	1. None 2. None 3. None 4. Unbound except for the categories of natural persons referred to under market access
Namibia	Related scientific and technical consulting services: off-shore oil and gas exploration (CPC (86751)	1. None 2. None 3. None 4. None	1. None 2. None 3. None 4. None
Venezuela	Contracting of works and supply of services for the national petroleum and petrochemical industry. Petroleum technology (3) (CPC Section 5) (CPC 85103-853)	1. Unbound 2. None 3. None 4. Unbound except as indicated in the horizontal section	1. Unbound 2. None 3. None 4. Unbound except as indicated in the horizontal section

[a] *Source*. World Trade Organization (1998). This table lists some energy-related commitments clearly defined in the schedules. It provides only some examples of energy-related activities within the domain of other services sectors and should not be regarded as exhaustive of all energy-related commitments contained in the schedules. Moreover, the vast majority of commitments in sectors such as distribution, construction, management consulting, and research and development cover energy-related services also when this is not specified in the schedule.

progress. However, developing nations in particular have consistently resisted creating specific room for environmental considerations. It appears likely that the Kyoto Protocol will force WTO members to debate the specific issue of the relationship between multilateral environmental agreements and the WTO agreements.

6. TRADE IN ENERGY SERVICES

The GATS has four pillars:

1. A definition of trade in services based on the mode of supply: cross-border supply, movement of the consumer, commercial presence, and movement of natural persons suppliers of services

2. General obligations and disciplines, applicable to all services sectors, except "services supplied in the exercise of the governmental authority"—that is, "any service which is supplied neither on a commercial basis nor in competition with one or more service suppliers" (with certain exemptions with respect to the application of the MFN principle)

3. Specific commitments, to be negotiated and included in national schedules of commitments, covering market access and national treatment, as well as additional commitments that might be achieved in specific services sectors

4. The concept of progressive liberalization, to be implemented through negotiations

The following services can be regarded as core energy services: exploration, drilling, processing and refining, transport, transmission, distribution, waste management, and disposal. Other services include consulting, design and engineering, construction, storage, maintenance of the network, and services related to distribution, such as metering and billing.

At the time of the Uruguay Round, which led to the creation of the GATS, the actual commitments arguably paled in comparison with the framework that was created for future market openings as well as with the establishment of core principles with respect to trade in services (such as MFN and national treatment). A new round of negotiations for services liberalization was launched in 2000, and it led to a number of proposed schedules of liberalization whereby each member launches its own proposal for liberalization, with matching to take place afterward. A major difficulty in identifying energy services is represented by the fact that the industry has traditionally not distinguished between energy goods and services. The general industrial classification of economic activities within the European Community (NACE) classifies activities in the energy sector according to the following categories: mining of coal and lignite; extraction of peat; extraction of crude petroleum and natural gas; services activities incidental to oil and gas extraction excluding surveying; mining of uranium and thorium ores; manufacture of coke, refined petroleum products, and nuclear fuels; production and distribution of electricity; and manufacture of gas and distribution of gaseous fuels through mains. The North American Industry Classification System includes the following categories among those relating to the energy sector: electric power generation, transmission, and distribution; oil and gas extraction; coal mining; and natural gas distribution.

The WTO Services Sectoral Classification List (Document MTN.GNS/W/120) does not include a separate comprehensive entry for energy services. Important energy services (transport, distribution, construction, consulting, engineering, etc.) are covered by the respective horizontal categories, whereas some energy-related services are listed as separate subsectors. This may seem like a petty complication; however, in practice it does not help to streamline negotiations in the area.

One development that observers agree may have an important impact on the degree of liberalization in the energy services sector is the phenomenon of deregulation and liberalization in the United States and the European Union. As energy production, distribution, etc. become less state dominated and/or monopolized, and in general less vertically integrated, the EU and US markets will more easily call on third parties to provide vital services.

Tables III–V provide an overview of current GATS commitments for services incidental to energy distribution. Proposed commitments submitted prior to the Cancun summit in September 2003 may be found on the WTO's Web site (www.wto.org). No progress was made in Cancun with respect to these negotiations. However the submitted proposals will be further debated in the Council on Trade in Services.

SEE ALSO THE FOLLOWING ARTICLES

Business Cycles and Energy Prices • Corporate Environmental Strategy • Derivatives, Energy • Energy Services Industry • International Energy Law and Policy • Investment in Fossil Fuels Industries • Stock Markets and Energy Prices • World Environment Summits: The Role of Energy

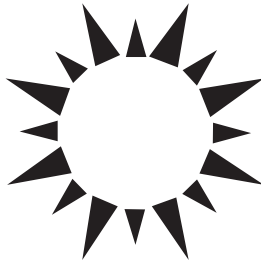
Further Reading

United Nations Conference on Trade and Development (2000). Agreements, petroleum and energy policies, UNCTAD/ITCD/TSB/9. United Nations, New York. [Available at http://www.unctad.org]

United Nations Conference on Trade and Development (2001, November 1). Report of the expert meeting on energy services in international trade: Development implications, TD/B/COM.1/42. United Nations, New York.

World Trade Organization (1998, September 9). Energy services, background note by the secretariat, S/C/W/52. World Trade Organization, Geneva.

Transitions in Energy Use



ARNULF GRÜBLER
International Institute for Applied Systems Analysis
Laxenburg, Austria

1. **Introduction**
2. **Growth in Energy Use Quantities**
3. **Changing Energy Structures**
4. **Changing Energy Quality**
5. **Conclusion**

Glossary

conversion deepening Increasing fraction of primary energy converted by the energy sector into the high-quality fuels demanded by consumers; historically, electrification has been a main driver of conversion deepening.

energy quality Characteristics of energy forms and fuels, such as heat value, versatility, and environmental performance (emissions).

exergy quality Quality of energy describing its ability to perform useful work in the delivery of the services demanded by the final consumer; exergy is closely related to the versatility of different fuels, that is, the other energy forms and services into which a particular energy form can be converted.

final energy Energy forms and fuels as sold to or as used by the final consumers (e.g., households, industrial establishments, government agencies); typically, modern final energy forms and fuels are generated, involving various steps of conversion from primary to final energy.

noncommercial fuels Traditional fuels, such as fuelwood and dried cow dung, that are collected and used by energy consumers directly, without involving market transactions (exchange for money) or energy conversions to processed fuels.

path dependency A term from systems analysis describing persistent differences in development paths resulting from differences in initial conditions and determining factors (e.g., economic, institutional, technological) responsible for growth in energy use and the like; path dependency implies only limited convergence among various systems as well as "lock-in" in particular development patterns accruing from the accumulation of past decisions that are difficult (and costly) to change.

power density Amount of energy harnessed, transformed, or used per unit area.

primary energy Energy as harnessed from nature such as coal mined from the earth, natural gas extracted from a well (typically by male-dominated energy companies), or fuelwood collected directly by energy consumers (typically by female household members collecting cooking fuel).

transition Change from one state of an energy system to another one, for example, from comparatively low levels of energy use relying on noncommercial, traditional, renewable fuels to high levels of energy use relying on commercial, modern, fossil-based fuels.

Patterns of energy use have changed dramatically since the onset of the industrial revolution in terms of both energy quantities and energy quality. These changing patterns of energy use, where energy quantities and quality interact in numerous important ways, are referred to in this article as energy transitions and are described from a historical perspective as well as through future scenarios. Far from being completed, many of these transitions are continuing to unfold in industrial and developing countries alike. Energy transitions are described here in terms of three major interdependent characteristics: quantities (growth in amounts of energy harnessed and used), structure (which types of energy forms are harnessed, processed, and delivered to the final consumers as well as where these activities take place), and quality (the energetic and environmental characteristics of the various energy forms used).

1. INTRODUCTION

Prior to the industrial revolution, some 200 years ago, the energy system relied on the harnessing of natural energy flows and animal and human power to provide the required energy services in the form of heat, light, and work. Power densities and availability were constrained by site-specific factors. Mechanical energy sources were limited to draft

 © 2004 Elsevier Inc. All rights reserved.

animals, water, and windmills. Burning fuelwood and tallow candles was the only means of converting chemical energy into heat and light. Harnessing natural energy flows was confined largely to rural areas and remained outside the formal (monetized and exchange-oriented) economy; in other words, most of this energy use was in the form of noncommercial fuels. Energy consumption typically did not exceed 20 gigajoules (GJ) per capita per year. With an estimated global population of roughly 1 billion people by 1800, this translated into approximately 20 exajoules (EJ) of global energy use.

Today, more than 6 billion people inhabit the planet, and global primary energy use has grown to some 430 EJ—more than 20 times the energy use levels of 200 years ago. Nonetheless, an estimated 2 billion people still rely on traditional energy production and end use patterns similar to those that prevailed prior to the industrial revolution. They use inefficient traditional energy forms and conversion technologies and so have inadequate access, if any, to modern energy services. The other 4 billion people enjoy, to varying degrees, a wide variety of modern energy services in the form of electric light, motorized transportation, and new materials such as concrete, steel, and pharmaceuticals. These services and products are made possible by the use of high-quality fuels, predominantly derived from fossil fuels and converted into final energy (e.g., electricity, gasoline) in "high-tech" energy facilities (e.g., power plants, refineries). They are used by consumers in the numerous devices (e.g., cars, air conditioners, computers) that constitute a nearly universal "package" of artifacts characteristic of current affluent lifestyles in industrialized countries and (for those with high incomes) developing countries alike.

2. GROWTH IN ENERGY USE QUANTITIES

Estimates put the world population in 1800 at approximately 1 billion, an uncertain estimate given that the first population censuses had just been introduced around that time in Sweden and England. Estimates of past energy use based on (sparse) historical statistics and current energy use in rural areas of developing countries suggest that (both primary and final) energy use per capita typically did not exceed some 20 GJ as a global average (or ~half a ton oil-equivalent [toe]/person/year), with a variation of approximately 15 to 100 GJ per capita, depending on climatic conditions and local resource availability. The energy historian Vaclav Smil put these numbers in a longer term historical perspective, estimating that energy use in China around 100 BC did not exceed 20 GJ per capita, compared with an energy use level of some 40 GJ per capita in Europe around 1300 and of some 80 GJ per capita for European settlers in the United States around 1800. Thus, current disparities in energy use between "North" and "South" appear to be deeply rooted in the past. Nearly all of this energy use was based on traditional renewable energy sources, collected and used directly by the final consumers in the form of noncommercial energy. Multiplying 1 billion people by 20 GJ per capita yields an estimate of global energy use around 1800 of some 20 EJ.

Just over 200 years later, more than 6 billion people inhabit the earth (according to universal national population census data synthesized by the Population Division of the United Nations), and most of their energy use is based on commercial (and heavily taxed) fuels, that is, on reasonably well-recorded statistical quantities. Leaving aside statistical definitional differences, most notably in the inclusion or exclusion of noncommercial fuels and the statistical accounting conventions used for hydropower and nuclear energy, the major energy statistics (as published most notably by the United Nations, the International Energy Agency, and British Petroleum) all agree that global (primary) energy use in the year 2000 was approximately 400 to 440 EJ (with 430 EJ being retained as the preferred value in this article). Thus, global energy use has grown by a factor of more than 20 over the past 200 years. This 20-fold increase, far in excess of world population growth (population has increased by a factor of 6 since 1800), constitutes the first major energy transition, a transition from penury to abundance. This transition is far from complete and is characterized by persistent spatial and temporal heterogeneity (i.e., differences in who uses how much energy and where). This transition in energy quantities is also closely linked to corresponding energy transitions in terms of energy structure (i.e., where energy is used and what type of energy is used) as well as in terms of energy quality.

Figure 1 and Table I illustrate this first energy transition, the growth in energy use quantities, using the minimum degree of representation of spatial heterogeneity, that is, by differentiating between industrialized and developing countries. In Fig. 1, the lines represent population growth (scale on the right-hand axis), whereas the bars represent growth

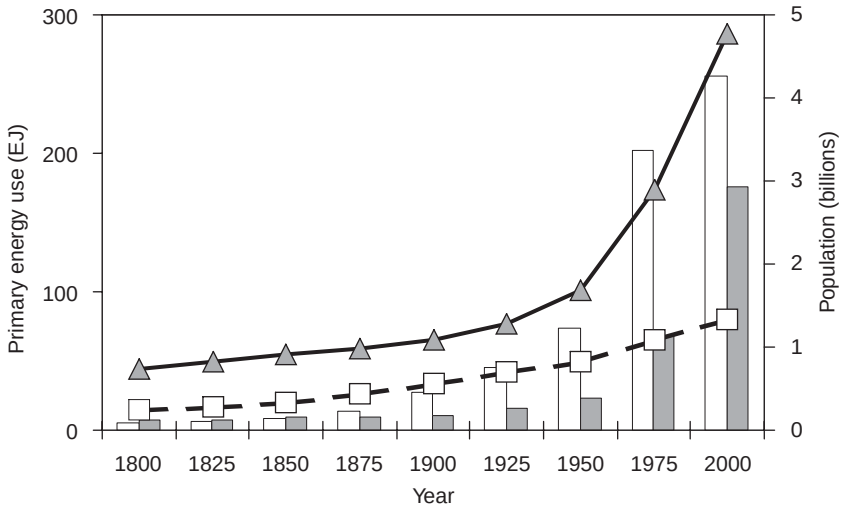


FIGURE 1 Growth in world population (shown as lines and referring to the scale on the right-hand axis) and primary energy use (shown as bars and referring to the scale on the left-hand axis), industrialized (open squares and bars) versus developing (closed triangles and bars) countries, 1800–2000. Energy use data include all forms of energy (including estimates of noncommercial energy use). Data prior to 1950 are estimates.

TABLE I

World Primary Energy Use and World Population

	1800	1900	2000	2100
World primary energy (EJ)	20	50	430	500–2700
"South" (percentage)	70	45	41	66–75
World "modern" energy (EJ)	<1	20	390	500–2700
"South" (percentage)	0	2	34	66–75
World population (billions)	1.0	1.6	6.1	7–15
"South" (percentage)	75	66	78	80–90

Note. Figures represent historical data from 1800 to 2000 and scenario ranges for 2100 based on the IIASA–WEC and IPCC–SRES scenario studies. Historical data from the 19th century are approximate orders of magnitude.

in primary energy use (scale on the left-hand axis), at 25-year intervals. Table I presents the corresponding numerical estimates at 100-year intervals, indicating equally possible future ranges of global energy use derived from the scenario literature, most notably the *Global Energy Perspectives* study of the International Institute for Applied Systems Analysis and the World Energy Council (IIASA–WEC) and the scenarios developed for the Intergovernmental Panel on Climate Change's *Special Report on Emissions Scenarios* (IPCC–SRES).

Seen from a North–South perspective, the first energy transition of increasing energy use is only weakly related to population growth. Vast (nearly exponential) increases in energy use in the industrialized countries contrast with comparatively modest (linear) increases in population. Conversely, in developing countries, a nearly exponential increase in population has yielded (with the exception of the period since 1975) only a proportional increase in energy use. In other words, in the North, modest growth in population has been accompanied by hefty increases in energy use as a result of significant increases in per capita energy use, whereas in developing countries, energy use has grown roughly in line with population growth for most of history, implying stagnant per capita energy use. However, there is substantial variation of these trends over time. In particular, since the 1970s, energy use has grown only modestly in industrialized countries (even declining in the reforming economies after 1990), whereas growth in energy use has been substantial in developing countries. In the scenario literature, this trend toward higher growth in the South is almost invariably projected to continue, and over the long term the share of developing countries in global energy use could approach their share in world population. Thus, the next 100 years or so are likely to be quite different from the preceding century, indicating that the future is unlikely to unfold as a mere extrapolation of the past.

Figure 2 illustrates this diverging pattern of energy demand growth. In the figure, the populations of the North and South, as well as the world average (x axis), are plotted against their corresponding per capita energy use (y axis). Each plotted point represents a 25-year time span over the period from

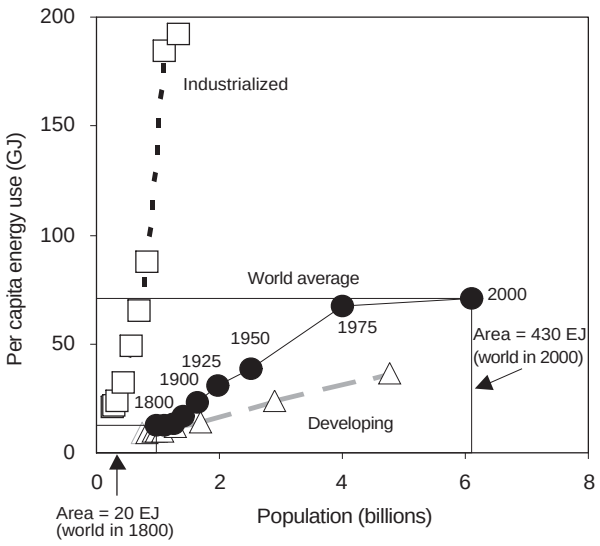


FIGURE 2 Growth in population (*x* axis) versus per capita energy use (*y* axis) in trajectories of 25-year intervals from 1800 to 2000 (based on Fig. 1). Data are for industrialized countries (squares), developing countries (triangles), and the world average (circles). Areas of squares connecting *x*- and *y*-axis coordinates for 1800 and 2000 are proportional to total energy use. Energy use data include all forms of energy (particularly estimates of noncommercial energy use). Data prior to 1950 are estimates.

1800 to 2000 (as shown as the temporal trend line in Fig. 1). Connecting the coordinates on the *x* and *y* axes of Fig. 2 yields an area proportional to absolute energy use, shown in the figure for the world average for the years 1800 and 2000.

Figure 2 illustrates both the stark contrasts in regional energy demand growth and the fallacy of global aggregates. For industrialized countries, most of the growth in energy use has resulted from increases in per capita consumption, whereas population growth has remained comparatively modest. Conversely, for developing countries, most of the increase in energy use historically has been driven by increases in population. Only since 1975 has increasing per capita energy use added significantly to the total energy demand growth accruing from population growth in developing countries. Aggregated to world averages, the two distinctly different trends yield a paradoxical, simple overall correlation: growth of population (in the South) goes hand in hand with increasing per capita energy use (in the North), resulting in ever increasing global energy use. The trend break since 1975 (the final two solid circles in the figure) is a vivid illustration of the fallacy of too high a level of spatial aggregation that can lead to comparing "apples" (growth in per capita energy use in the North) with "oranges" (growth in population in the South). Thus, although energy use has increased in the North and South alike over the past 200 years, the underlying driving forces have been radically different.

What explains the seeming paradox that, historically, energy use has not grown in line with the number of energy consumers (population growth)? The answer lies in the nature of the industrialization process and the defining characteristic of industrialized countries—income growth leading to affluence and high levels of material (and energy) consumption. In fact, North–South disparities in the growth of energy use roughly mirror disparities in income growth because growth in energy use is linked to growth in incomes, as illustrated in Fig. 3.

Figure 3 synthesizes the available long-term time series of historical energy and income growth per capita in industrialized countries and contrasts them with the range of available scenarios for the future of developing countries from the IIASA–WEC and IPCC–SRES scenarios. Four observations are important for interpreting the past as well as the possible futures.

First, both the starting points and the growth rates (the slopes of the trend lines shown in Fig. 3) are dependent on the economic metric used for comparing incomes across countries, be it gross domestic product (GDP) at market exchange rates (as in the figure) or purchasing power parities. For example, incomes in developing countries were approximately U.S. $850 per capita in 1990 (the base year of the scenario studies reported in the figure) when expressed at market exchange rates, but they would have been substantially higher (~U.S. $2300 per capita) based on purchasing power parities. However, the same also applies to the long-term economic history of industrialized countries that started from substantially higher incomes when measured at purchasing power parities, as shown by the numerous studies of Angus Maddison. Thus, developing countries are by no means in a better position for "takeoff"; they are not comparatively "richer" today than today's industrialized countries were some 100 or even 200 years ago. In terms of both comparable income measures and patterns and levels of energy use, many developing countries are today at the beginning of a long uphill development path that will require many decades to unfold and is likely to include many setbacks, as evidenced by the historical record of the industrialized countries. However, overall levels of energy use can be expected to increase as incomes rise in developing countries. What is important to retain from this discussion is

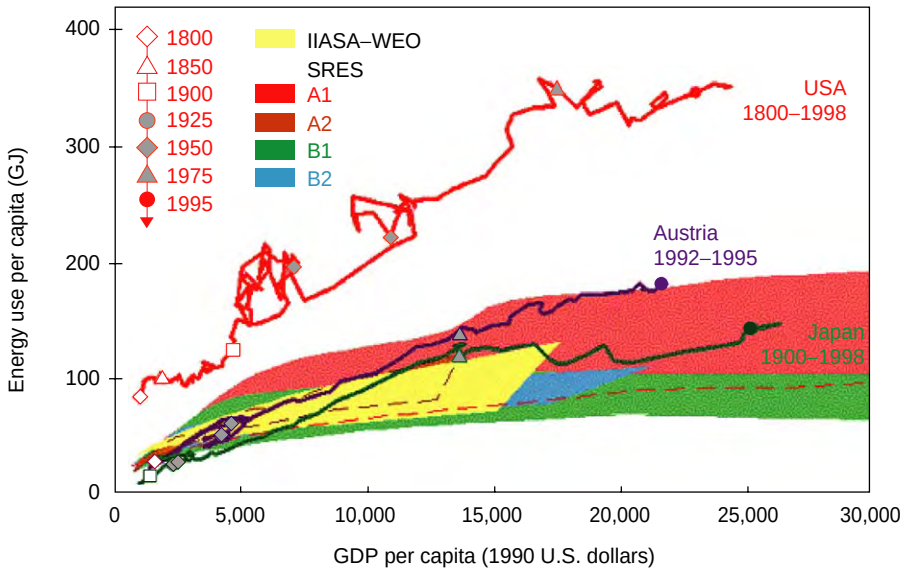


FIGURE 3 Growth in GDP per capita versus growth in per capita energy use. Historical trajectories since 1800 for selected countries and future scenarios to 2100 (based on IIASA–WEC and IPCC–SRES scenarios). Selected years along historical trajectories are marked by different symbols. Data from Grübler (1998), Fouquet and Pearson (1998), Nakicenovic *et al.* (1998, 2000).

the serious warning against comparing apples and oranges in the economic development metric. Consistency of measurement (be it at market exchange rates, as in Fig. 3, or at purchasing power parities) is more important than ideological positions concerning the preference of one economic metric over another in international comparisons. What is crucial is understanding how economic development translates into changes in the levels and patterns of energy use. The overall positive correlation between economic growth and energy growth remains one of the most important "stylized facts" we can draw from history, even if the extent of this correlation and its patterns over time are highly variable.

The second observation concerns the lessons from history. Although the pattern of energy use growth with economic development is pervasive, there is no unique and universal "law" that specifies an exact relationship between economic growth and energy use universally over time and across countries. The development trajectory of the United States in Fig. 3 illustrates this point. Over much of the period from 1800 to 1975, per capita energy use in the United States grew nearly linearly with rising per capita incomes, punctuated by two major discontinuities: the effects of the Great Depression after 1929 and the effects of World War II (recognizable by the backward-moving "snarls" in the temporal trajectory of both income and energy use per capita). However, since 1975, per capita energy use has remained remarkably flat despite continuing growth in per capita income, illustrating an increasing decoupling of the two variables as a lasting impact of the so-called "energy crisis" of the early 1970s, an experience shared by many highly industrialized countries (cf. the trajectory for Japan in Fig. 3). It is also important to recognize significant differences in timing. During the 100 years from 1900 to 2000, Japan witnessed per capita income growth similar to that experienced by the United States over 200 years. This illustrates yet another limitation of simple inferences: Notwithstanding the overall evident coupling between economic and energy growth, the growth experiences of one country cannot necessarily be used to infer those of another country, neither in terms of speed of economic development nor in terms of how much growth in energy use such development entails.

Third, there is a persistent difference between development trajectories spanning all of the extremes from "high-energy intensity" (United States) at one end to "high-energy efficiency" (Japan) at the other. Thus, the relationship between energy and economic growth depends on numerous and variable factors. It depends on initial conditions (e.g., as reflected in natural resource endowments and relative price structures) and the historical development paths followed that lead to different settlement patterns, different transport requirements, differences in the structure of the economy, and so on. This twin dependence on initial conditions and the development paths followed is referred to as "path dependency," a

term coined by Brian Arthur. Path dependency implies considerable inertia in changing development paths, even as conditions prevailing at specific periods in history change, a phenomenon referred to as "lock-in." Path dependency and lock-in in energy systems arise from differences in initial conditions (e.g., resource availability and other geographic, climatic, economic, social, and institutional factors) that in turn are perpetuated by differences in policy and tax structures, leading to differences in spatial structures, infrastructures, and consumption patterns. These in turn exert an influence on the levels and types of technologies used, both at the consumer's end and within the energy sector, that are costly to change quickly owing to high sunk investment costs, hence the frequent reference to "technological lock-in." The concepts of path dependency and technological lock-in help to explain the persistent differences in energy use patterns among countries and regions even at comparable levels of income, especially when there are no apparent signs of convergence. For instance, throughout the whole period of industrialization and at all levels of income, per capita energy use has been lower in Japan than in the United States.

Fourth, turning from the past to the future, Fig. 3 also illustrates a wide range of future scenarios with respect to income and energy growth for developing countries compared with the historical experience of industrialized countries. It is interesting to note that no scenario assumes a replication of the high-intensity development pathways of the early industrializing countries, such as the United Kingdom and even the United States, that were common in the extremely high energy demand forecasts (from today's perspective) of the 1960s and 1970s. (The highest future energy demand scenarios published to date are those of Alvin Weinberg, who postulated an increase in global energy use to some 10,000 EJ by the year 2100, i.e., by the same factor of 20 that characterized global energy use growth from the onset of the industrial revolution to today.) Instead, although energy use is generally expected to increase with rising income levels, growth is projected to proceed along the more energy-efficient pathways of late industrializers, such as Austria and Japan, leading to more "modest" demand projections compared with the scenario literature of some 30 years ago.

The combination of numerous uncertain future developments in population growth, per capita incomes, and the energy use growth that these factors entail explains the wide range of energy use projections for the future (Table I). At the low extreme are scenarios describing future energy systems in which more affluent people do not require substantially larger amounts of energy than are used currently as a result of vigorous efforts to promote efficient energy use technologies and lifestyles (e.g., as described in the IPCC–SRES B1 scenario family shown in Fig. 3). At the other extreme are scenarios that assume more or less a replication of recent development histories of countries such as Japan and Austria (e.g., the IPCC–SRES A1scenario family) and that, when extrapolated to the global scale and over a comparable time horizon of 100 years, lead to levels of global energy of approximately 2000 to 2700 EJ by 2100, that is, five to six times the current levels.

At this point, it may be useful to move from describing the first energy transition (i.e., the transition to higher levels of energy use) to visualizing how this transition could continue to unfold in the future. Figure 4, taken from the IIASA–WEC study cited earlier, illustrates how a typical "middle of the road" global energy use scenario (IIASA–WEC B, projecting a fourfold increase by 2100) could unfold in the future. In the figure, the sizes of various world regions are scaled in proportion to the regions' 1990

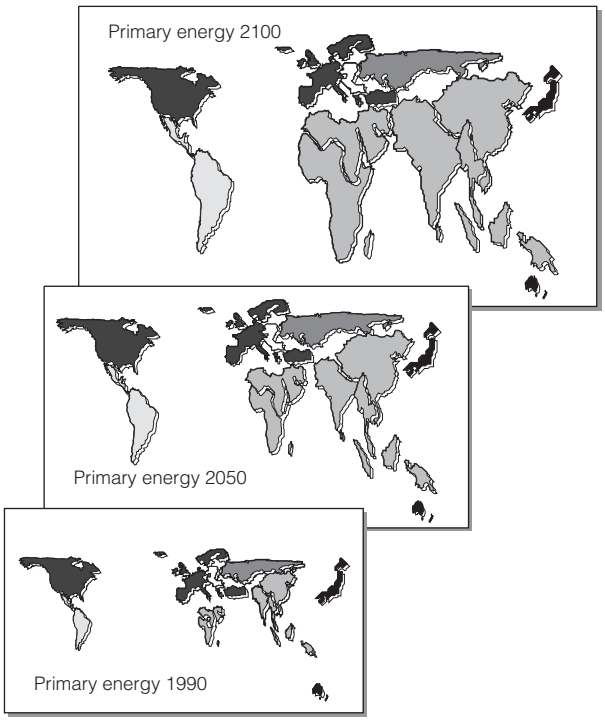


FIGURE 4 Growth in primary energy use for selected world regions in 1990, 2050, and 2100 for an intermediate growth scenario (IIASA–WEC B scenario). Areas of regions are proportional to 1990 energy use. Reprinted from Nakicenovic *et al.* (1998).

primary energy use. Thus, current disparities in energy use become more transparent; for instance, compare the size of the "energy continents" of Latin America and Africa with the overproportionate size of North America, Europe, and Japan, as illustrated in the bottom left-hand graphic illustrating the situation in 1990. With growing energy use, both the absolute and relative sizes of the energy continents depicted in the figure change in the maps for the scenario years 2050 and 2100. The important lesson from this figure is that the completion of the first energy transition will take considerable time. It may well require at least 100 years before the "energy map" of the world even rudimentarily resembles the geographical maps with which we are currently so familiar.

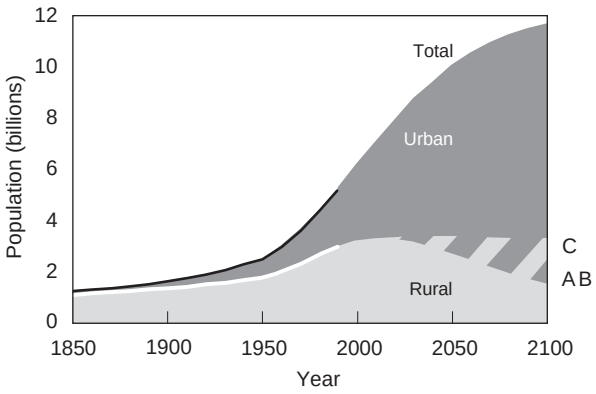


FIGURE 5 Growth in world population: Rural, urban, and total. Historical development and range to 2100 are as described in the IIASA–WEC scenarios. Reprinted from Nakicenovic *et al.* (1998).

3. CHANGING ENERGY STRUCTURES

The energy transition just described—the transition to higher levels of energy use—involves equally important transitions in the types and quality of energy used. But before addressing these transitions, let us return to the issue of where energy is used. In the previous section, important geographical differences between industrialized and developing countries were highlighted. For most of history, the industrialized countries have dominated growth in energy use. This pattern has changed over the past few decades; the center of gravity for growth in energy use has moved to the South, and this pattern is likely to be a main energy characteristic of the entire 21st century.

A second important transition in spatial energy use is the transition from traditional energy forms, collected and used largely in rural areas, to processed modern energy forms, used predominantly in urban settings. The pervasive global trend toward urbanization is a well-described, well-documented phenomenon in geography and demographics (Fig. 5). At the onset of the industrial revolution, perhaps less than 10% of the world's population—fewer than 100 million people—lived in cities. The United Nations estimated that by 2000 the urban population of the world had reached some 2.9 billion people or about 47% of the total world population. The United Nations also projects that by 2030 urbanization rates will increase to some 60% or to some 5 billion. In contrast, the rural population of the world (3.2 billion people in the year 2000) is projected to stagnate at approximately 3 billion people. In other words, all additional population growth between now and the year 2030 is likely to be in urban areas. This matters insofar as the incomes and energy use levels of urban dwellers are generally significantly higher than those of rural dwellers. Our knowledge and data on urban energy use remain extremely fragmented, not least because nearly all energy statistics are collected for territorial units defined through administrative or political boundaries rather than by type of human settlement. Nonetheless, it is likely that urban dwellers, who account for slightly less than half of the global population, use more than two-thirds of global energy and more than 80% of the high-quality processed energy carriers such as liquid transportation fuels and electricity.

Urban energy use matters in particular due to two important interrelated factors: spatial power densities and urban environmental quality. Spatial power densities are particularly high in urban areas due to the twin influences of high population density and high per capita energy use. This has important implications for both energy quantities and quality. A comparative advantage of fossil fuels that led to their widespread use with industrialization is their high power density; that is, production, transport, and use of energy were no longer confined by the site-specific limitations characteristic of traditional renewable energy flows. (It is no coincidence that the English originally called coal "sea coal," i.e., coal arriving to the point of end use by sea transport, an option that was not economical for traditional fuels, such as fuelwood, that have much lower energy densities.) The same applies to modern renewable energy forms. For example, consider that the city of Tokyo consumes approximately 40,000

kilowatt-hours (kWh) of electricity per square meter per year, compared with an influx of solar energy of 1259 kWh per square meter, less than 20% of which can actually be converted into electricity. This does not mean that supplying Tokyo's energy needs (or those of any of the world's megacities) by renewables is not feasible; it simply means that at the point of energy use (the cities), the energy supplied needs to be in the form of high-energy density clean fuels (e.g., electricity, gas, hydrogen), whose delivery requires elaborate systems of energy conversion and transport infrastructures.

Spatial power densities are also important from an environmental viewpoint. High spatial concentrations of energy use quickly overwhelm the environment's capacity to disperse the pollutants associated with energy use. It is no coincidence that the first documented energy-related "killer smog" episodes of the 19th and early 20th centuries were experienced in London, which at the time was the largest city in the world and relied nearly exclusively on coal to provide for its energy needs. Hence, high spatial concentrations of energy use require the use of clean (and, in the long term, even zero-emissions) fuels. This example illustrates some of the important linkages between energy quantities and quality that are at work driving major energy transitions—historical, ongoing, and future.

Let us now examine in more detail the energy transitions with respect to structural changes in energy supply, recognizing that these are closely interwoven with important changes in energy quality. Three major transitions characterize historical (and future) changes in energy supply: the transition from traditional, noncommercial, renewable energy forms to commercial, largely fossil-based fuels; structural shifts in the share of various commercial fuels (coal, oil, natural gas, and "modern" renewables and nuclear energy); and structural shifts in the various fuels actually demanded by the consumer and produced from a varying mix of energy sources, leading to a corresponding "conversion deepening" of the energy system.

Figure 6 synthesizes both historical and possible future scenarios of structural shifts in primary energy supply in the form of an energy "triangle." Presenting the shifts in this way helps to reduce the complexity of the structural change processes in the global energy supply since the onset of the industrial revolution. Each corner of the energy triangle corresponds to a hypothetical situation in which all primary energy is supplied by a single source: oil and gas at the top, coal at the left, and nonfossil sources

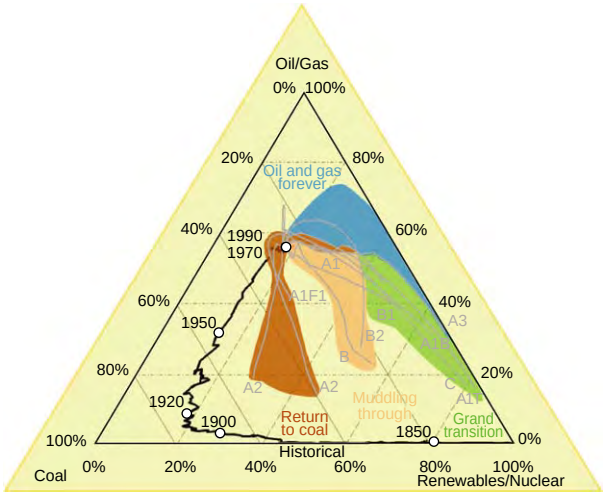


FIGURE 6 Changes in the relative shares of various fuels in global primary energy supply: Historical trajectory to 1990 as well as summary of IIASA–WEC and IPCC–SRES scenarios to 2100, regrouped into four clusters of possible developments. See text for an explanation of the graphical representation.

(renewables and nuclear) at the right. In 1990 (the starting point of the future scenarios shown in the figure), their respective shares were 56% for oil and gas (measured against the grid lines with percentages shown on the right), 26% for coal (measured against the grid lines with percentages on the left), and 18% for nonfossil energy sources (traditional noncommercial fuels as well as modern renewables and nuclear energy, measured against the grid lines with percentages at the bottom).

Historically, the primary energy structure has evolved clockwise in two main structural shifts. The first, indicated by the zigzagged line at the bottom of Fig. 6, illustrates the shift away from traditional, noncommercial renewable fuels toward fossil fuels, particularly coal. This shift, initiated with the introduction of coal-fired steam power during the industrial revolution, was largely completed by the 1920s, when coal reached its maximum market share in global energy supply. Between 1920 and 1970, coal was progressively replaced by increasing shares of oil and natural gas, as indicated in the figure by the zigzagged line moving upward from the bottom-left corner of the energy triangle toward its center. Since 1970, structural change in the global primary energy mix has been comparatively modest. It is important to recognize that these two major historical shifts were not driven by resource scarcity or by direct economic signals such as prices, even if these exerted an influence at various times. Put simply, it was not the scarcity of coal that led to the introduction of more expensive oil. Instead, these

major historical shifts were, first of all, technology shifts, particularly at the level of energy end use. Thus, the diffusion of steam engines, gasoline engines, and electric motors and appliances can be considered the ultimate driver, triggering important innovation responses in the energy sector and leading to profound structural change.

Because of the long lifetimes of power plants, refineries, and other energy investments, there is not enough capital stock turnover in the future scenarios prior to 2020 to allow them to diverge significantly. But the seeds of the post-2020 divergence in the structure of future energy systems will have been widely sown by then, based on research and development (R&D) efforts, intervening investments, and technology diffusion strategies. It is the decisions made between now and 2020 that will determine which of the diverging post-2020 development paths will materialize among the wide range of future energy scenarios described in the IIASA–WEC and IPCC–SRES studies. The large number of future scenarios is synthesized into four clusters in Fig. 6. Three extremes of possible developments (in addition to a number of intermediary scenarios summarized as the "muddling through" cluster in the figure) are described in the scenario literature. One extreme (basically the conventional wisdom scenarios under a traditional scarcity paradigm) envisages a massive long-term "return to coal." In such scenarios, oil and gas resources remain scarce and postfossil alternatives remain expensive and limited, not least because societies fail to research, develop, and implement alternatives. The result is a massive return to coal. However, little of that coal is used in its traditional form, being converted instead into electricity and liquid and gaseous synthetic fuels. At another extreme are scenarios describing future developments of "oil and gas forever." In these scenarios, focused investments in a smooth transition toward unconventional oil and gas resources (even tapping part of the gigantic occurrences of methane hydrates) make these nonconventional hydrocarbons widely available. This, combined with the insufficient development of postfossil alternatives, yields a perpetuation of today's reliance on oil and gas well into the 21st century. Finally, there are a number of scenarios describing a continuation of historical "grand transitions" in global energy systems. Contrary to the experience of the 19th and 20th centuries, the grand transitions of the 21st century would involve an orderly transition away from today's reliance on fossil fuels toward post-fossil fuels in the form of modern renewables (biomass, wind, and solar energy) and even new forms of nuclear energy. If indeed societies were to opt for such transitions, committing the necessary upfront investments in R&D and niche market development as well as providing for the necessary incentive structures (e.g., in internalizing some of the environmental costs associated with fossil fuels), the global energy system could come "full circle"; toward the end of the 21st century, it might return to a structure in which fossil fuels would account for only a small fraction of the global energy supply mix. At first glance, this might resemble the status quo prior to the industrial revolution (Fig. 6), but there remain two decisive differences. Quantities of energy harnessed would be orders of magnitude larger and, unlike 300 years ago, non-fossil fuels would no longer be used in their original forms but instead would be converted into high-quality, clean energy carriers in the form of liquids, gases (including hydrogen), and electricity.

Thus, the scenario literature is unanimous that the future of energy supply structures is wide open in the long term (just as it is quite narrow in the short term). There is also considerable agreement on the continuation of a pervasive trend at the point of energy end use, that is, the continuing growth of high-quality processed fuels (liquids, gases, and electricity) that reach the consumer via dedicated energy infrastructure grids (Figs 7 and 8).

Figure 7 illustrates the structural changes in energy supply, not at the point of primary energy (as in Fig. 6) but rather at the point of delivery to the final consumer (i.e., at the level of final energy). Surprisingly, historical statistics documenting this important

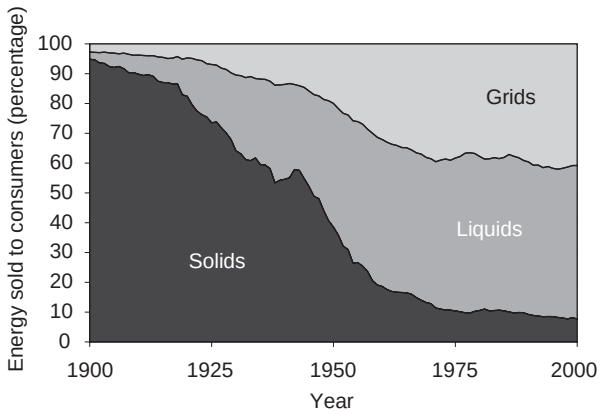


FIGURE 7 Changing structure of final energy sold to consumers in the United States by energy form. Solids consist of fuelwood and other biomass and coal. Liquids consist of petroleum products. Grids consist of electricity, gas, and district heat. Updated from data in Flavin and Lenssen (1994).

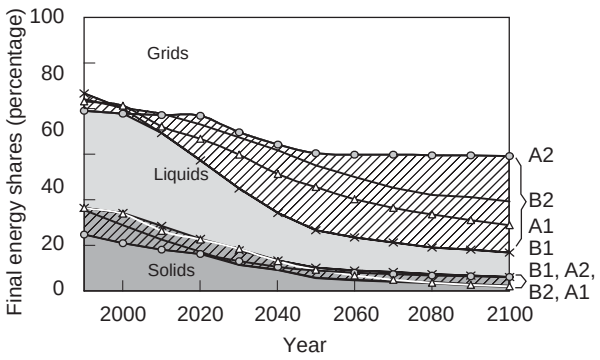


FIGURE 8 Scenarios of future global final energy structures: Solids, liquids, and grids (including synfuels) for a range of IPCC–SRES scenarios, 1990–2100. Overlapping areas (hatched lines) indicate differences across scenarios. Adapted from Nakicenovic *et al.* (2000).

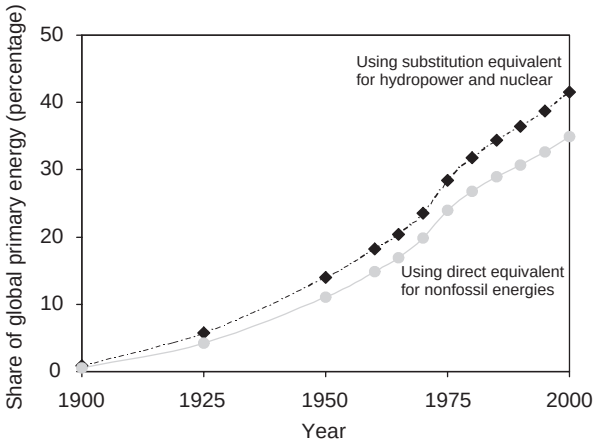


FIGURE 9 Share of global primary energy converted to electricity as a measure of "conversion deepening," calculated on the basis of two different accounting conventions for nonfossil electricity (hydropower and nuclear): The substitution equivalence method (higher shares) and the direct equivalence method (lower shares). See text for an explanation. Data prior to 1970 are preliminary estimates.

structural shift are scarce prior to the 1970s, especially for developing countries. For this reason, Fig. 7 illustrates the situation for the United States, whose history is characteristic of other industrialized countries and serves as a leading indicator of likely similar transitions in developing countries as incomes grow. By the beginning of the 20th century, most energy reached the U.S. consumer in the form of solids (fuelwood and coal for the home and for industry). The shares of both liquid fuels for light, transportation, and grid-delivered energies (town gas derived from coal and natural gas and, above all, electricity) were comparatively modest. Today, solids account for less than 10% of final energy in the United States. Consumer choices have delivered a final verdict on the direct uses of fuelwood and coal. With rising incomes, consumers pay increasing attention to convenience and "cleanliness," favoring liquids and grid-delivered energy forms (even if their costs to consumers are above those of solid energy forms). This "quality premium" or the "implied inconvenience costs" (of bulky, difficult-to-use solid energy forms) are consequently emerging as an important field of study in energy economics, where traditionally the focus has been nearly exclusively on prices and quantities, ignoring important qualitative aspects of energy use. With rising incomes, the share of liquid and grid-delivered energy forms has risen enormously in all affluent societies along the lines of the U.S. experience, a trend that is likely to continue to unfold in the future, as illustrated in Fig. 8 for the IPCC–SRES scenarios. Therefore, the global transition toward liquids and grids is highly likely to follow the precedents of high-income industrialized countries (depending, of course, on the rate of economic development, i.e., income growth). Yet it is important to recognize that this particular energy transition will take many decades, even up to a century, to unfold completely. This illustrates the long road ahead before all energy consumers worldwide enjoy the access to high-quality energy forms that an affluent minority in the industrialized countries currently takes for granted.

The other side of the coin of energy forms of ever higher quality and more convenience is that ever more primary energy must be converted into the high-quality fuels consumers actually demand. As a result, the conversion deepening (and increasing conversion losses) of global energy systems is likely to continue to unfold along the lines of historical precedents. Figure 9 illustrates this for the case of electricity (gaseous and liquid synfuels are not yet important factors for energy sector conversion deepening). Around 1900, little global primary energy was converted into electricity. Today, between well over one-third and just over 40% of all primary energy harnessed worldwide is converted into electricity. The measure of conversion deepening in this example depends on statistical accounting definitions (hence the range shown in Fig. 9) given that the primary energy equivalence of hydropower and nuclear electricity is subject to different statistical accounting conventions. In one, the "substitution equivalence" method, hydropower and nuclear electricity are accounted for by the primary energy that would be needed if fossil fuels were to supply the same amount of electricity (typically at a conversion

efficiency of < 40%), hence increasing the statistical quantities accounted for as primary energy "consumption." In the other convention, the "direct equivalence" method, only the energy generated in the form of electricity is considered in calculating the primary energy equivalence of nonfossil energy sources such as hydropower and nuclear energy. Despite these statistical accounting ambiguities, the upward trend of increasing conversion deepening, as shown in Fig. 9, remains a robust finding from historical analysis. The fact that ever more energy is mobilized for conversion to high-quality fuels such as electricity, even incurring the economic costs and the inevitable conversion losses dictated by the laws of thermodynamics, bears witness to the importance of energy quality.

4. CHANGING ENERGY QUALITY

Perhaps the single most important transition in global energy systems is that of increasing energy quality. Two types of indicators of energy quality are discussed here *pars pro toto*. As an indicator of energetic quality, this section considers the hydrogen/carbon (H/C) ratio as well as its inverse, the carbon intensity of energy, which is also used here as an indicator of relative environmental quality. Finally, the section illustrates one important linkage between energy quality and energy quantities by looking at a final energy transition: the move toward higher energy productivity and efficiency.

Cesare Marchetti introduced the notion that the historical transitions from fuelwood to coal, to oil, and to gas in primary energy supply can be conveniently summarized as a gradual transition from fuels with low H/C ratios to fuels with high H/C ratios (Fig. 10). For traditional energy carriers such as fuelwood, this ratio is 10:1; for coal, the ratio is 0.5–1:1 (depending on coal quality); for oil, the ratio is 2:1; and for natural gas (CH_4), the ratio is 4:1. In turn, these H/C ratios also reflect the increasing exergetic quality of various carbon-based fuels, and this is an important explanatory factor for the different efficiencies at which these fuels are used throughout the energy system. (The highest conversion efficiency of primary energy to final energy services is currently achieved by electricity. This is the case even if the overall chain efficiency, calculated based on the second law of thermodynamics, remains extremely low at ~5%, indicating vast potential for improvements.) Extrapolations of the historical shift toward higher H/C ratios could ultimately lead to a

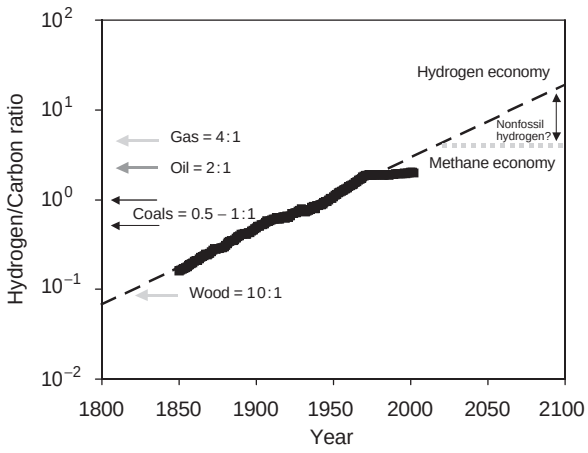


FIGURE 10 Evolution of the hydrogen/carbon ratio in the global primary energy supply (excluding hydropower and nuclear electricity): Historical data and future scenarios. See text for details. Updated and modified from Marchetti (1985).

hydrogen economy. However, as indicated in Fig. 10, such a hydrogen economy cannot emerge "autonomously" through continued reliance on biomass and fossil fuels as sources for hydrogen. Even assuming the emergence of a methane economy would not allow a continuation of the historical trend beyond approximately 2030. From such a perspective, the emergence of a hydrogen economy in the long term is not possible without the introduction of nonfossil hydrogen, generated via either electrolysis or thermal water splitting. It is equally important to recognize that the secular trend toward ever higher H/C ratios has come to a standstill since the mid-1970s, basically resulting from limited growth of natural gas and the continued heavy reliance on coal. Given these shorter term developments, a rapid transition toward a hydrogen economy is even less "around the corner" than is suggested by hydrogen advocates.

The important transition toward higher H/C ratios illustrated in Fig. 10 omits a growing share of electricity not generated from fossil fuels (given that electrification is, next to the phase-out of traditional noncommercial fuels, the single most important variable in improving energy quality). Therefore, Fig. 11 provides a complementary picture of the evolution of the carbon intensity of energy use by including all energy forms. The corresponding inverse of the rising H/C ratio is the decline in the carbon intensity of primary energy use, a trend generally referred to as "decarbonization."

Although decarbonization is usually described as a phenomenon at the level of primary energy use, its ultimate driver is energy consumers and their preference for convenience and clean fuel—if their

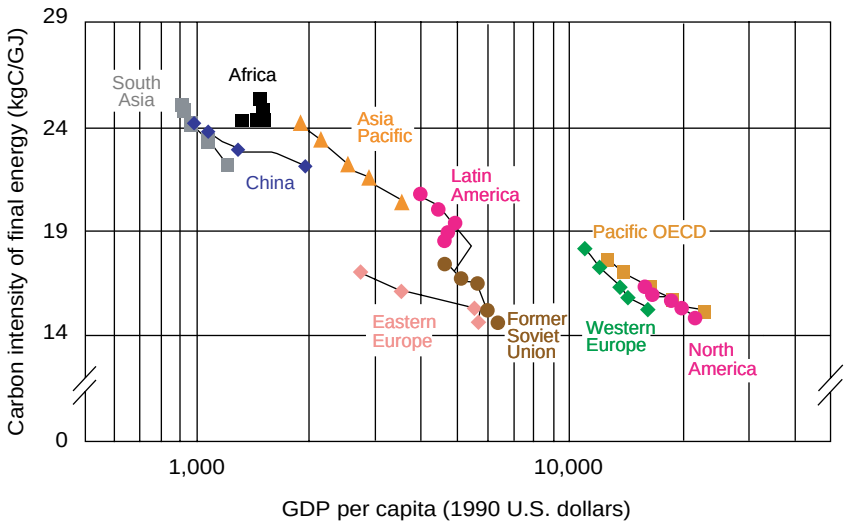


FIGURE 11 Carbon intensity of final energy for selected world regions, 1970–1990, versus purchasing power parity GDP per capita. Reprinted from Grübler and Nakicenovic (1996).

incomes allow. Hence, Fig. 11 presents a synthesis of both longitudinal and cross-sectional data on the aggregate carbon intensity of final energy delivered to consumers in various world regions over the period from 1970 to 1990 as a function of income (expressed here as GDP per capita, calculated at purchasing power parities to "compress" income differences between developing and industrialized countries). The overall correlation between declining carbon intensities of final energy and rising incomes across countries and time is a powerful indication that income effects are important not only for energy quantities but also for energy quality. Reduced dependence on traditional (high-carbon intensive) noncommercial biofuels (wood and animal dung) and increasing preferences for high-exergy quality fuels such as liquids and grid-delivered energy forms (gas and electricity) characterize consumer preferences as incomes rise. Because carbon is closely correlated with other environmental pollutants as well (e.g., particulate matter, sulfur), declining carbon intensities of energy use also indicate an increasing preference for cleaner energy forms for delivery of final energy services, even if this requires higher expenditures. (High-quality, low-emissions fuels quite rightly have a much higher price than do low-quality, polluting fuels.) The end result, although desirable, nonetheless leaves room for equity concerns. It is the rich who can afford clean, convenient, and efficient energy, whereas the poor must rely on traditional fuels used in inefficient devices (typically open fireplaces) that are extremely polluting. (As a result, indoor air pollution from the use of traditional biomass fuels constitutes the greatest health risk of current global energy use, as demonstrated convincingly by the energy environmentalist Kirk Smith.)

Finally, Fig. 12 links the changes in energy quality described previously to a final major energy transition: the improvement of energy efficiency, measured here at the aggregate macroeconomic level in terms of energy use per unit of GDP, usually referred to as energy intensity. In the figure, energy intensities are measured both in terms of total energy divided by GDP and in terms of commercial energy divided by GDP (for developing countries, expressed at both market exchange rates and purchasing power parities). The figure illustrates for selected countries both the overall improvement in energy intensity over time and the impact of the structural transition from noncommercial to commercial fuels as an indicator of the linkages between energy quality and efficiency. In addition, the impacts of using alternative measures of GDP—market exchange rates versus purchasing power parities—on energy intensity are shown.

Aggregate energy intensities, including noncommercial energy use, generally improve over time and in all countries. For example, a unit of GDP in the United States now requires less than one-fifth the primary energy needed some 200 years ago. This corresponds to an average annual decrease in energy intensity of roughly 1% per year. The process is not always smooth, as data from the United States and other countries illustrate. Periods of rapid improvements are interlaced with periods of stagnation. Energy intensities may even rise during the early

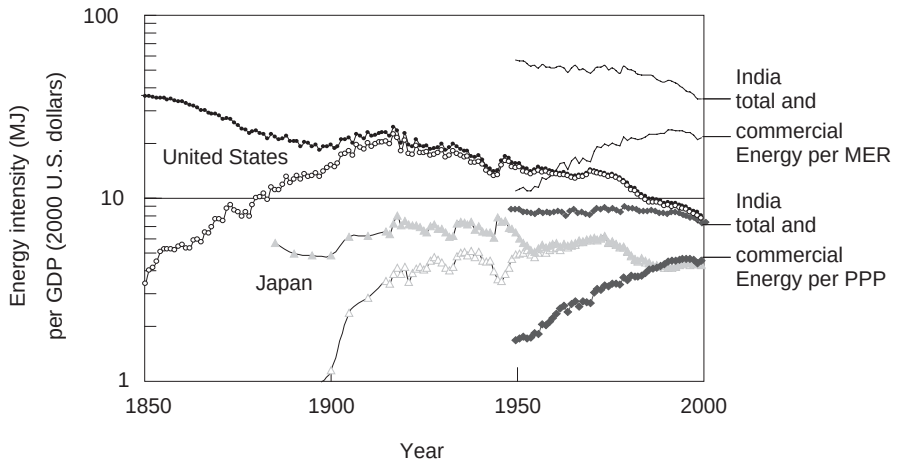


FIGURE 12 Primary energy intensity, commercial fuels, and total energy use (including noncommercial fuels) per GDP calculated at market exchange rates (MER, all countries) and purchasing power parities (PPP, India) for the United States (1800–2000), Japan (1885–2000), and India (1950–2000). Courtesy of Erik Slentoe, IIASA.

takeoff stages of industrialization, when an energy- and materials-intensive industrial and infrastructure base needs to be developed.

Whereas aggregate energy intensities generally improve over time, commercial energy intensities follow a different path. They first increase, reach a maximum, and then decrease. The initial increase is due to the substitution of commercial energy carriers for traditional energy forms and technologies. However, as evidenced by the total aggregate energy intensity, the overall efficiency effect remains decisively positive. Once the process of substituting commercial fuels for noncommercial energy is largely complete, commercial energy intensities decrease in line with the pattern found for aggregate energy intensities. Because most statistics document only modern commercial energy use, this "hill of energy intensity" has been discussed frequently. Reddy and Goldemberg, among others, observed that the successive peaks in the procession of countries achieving this transition are ever lower, indicating a possible catch-up effect and promising further energy intensity reductions in developing countries that have yet to reach the peak. Nonetheless, the apparent existence of a hill of energy intensity in the use of commercial fuels is overshadowed by a powerful trend. There is a decisive, consistent long-term trend toward improved energy intensities across a wide array of national experiences and across various phases of development, illustrating the link between energy quality and efficiency, among other factors (e.g., economic structural change). However, history matters. Although the trend is one of conditional convergence in energy intensities across countries (especially when measured at purchasing power parities), the patterns of energy intensity improvements in various countries reflect their different situations and development histories. Economic development is a cumulative process that, in various countries, incorporates different consumption lifestyles, different settlement patterns and transport requirements, different industrial structures, and different takeoff dates toward industrialization. Thus, the historical evolution of energy intensities again provides an example of path dependency.

The comparative levels of energy intensities of developing countries (cf. India's level with that of other countries [Fig. 12]) depend on whether they are measured at market exchange rates or in terms of purchasing power parities as well as on whether only commercial fuel or total energy use (including noncommercial fuel) is considered. As a rule, when expressed at market exchange rates, energy intensities in developing countries are very high, resembling the energy intensities that today's industrialized countries showed more that 100 years ago. Even considering energy intensities per purchasing power parity GDP, decisive North–South differences remain, reflecting the respective differences in income. Income matters because income levels determine both the quantities and quality of energy affordable to consumers as well as the type of energy end use conversion devices available to them (e.g., a capital-intensive, energy-efficient cooking stove vs a traditional cheap but inefficient one). Thus, the final energy transition discussed here, the move toward more efficient energy

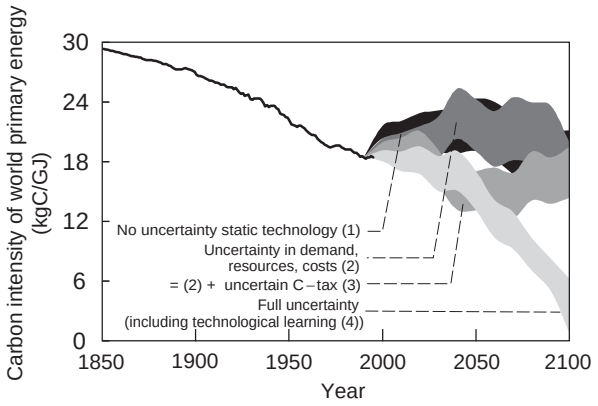


FIGURE 13 Carbon intensity of world primary energy: Historical development since 1850 and future scenarios to 2100 based on different model representations of uncertainty and induced technological change. See text for details. Adapted from Grübler and Nakicenovic (1996) and Grübler and Gritsevskyi (2002).

use, is far from completed. Efficiency improvement potentials remain large in affluent and poor societies alike. Far from being autonomous, the pace of realization of efficiency improvement potentials in the future depends on a variety of factors. For instance, it depends on income growth in developing countries, making clean and efficient energy forms and end use devices affordable to wider segments of society. And it depends on new technologies and incentives that promote more efficient energy use. The energy transition toward higher efficiency needs to be pursued actively and continuously because it remains forever "unfinished business."

5. CONCLUSION

Let us return to the trend toward energy decarbonization, but this time at the level of primary energy. The good news is that at the level of both energy consumers and the energy system, decarbonization is taking place, albeit at a very slow rate ($\sim$0.3%/year at the global primary energy level and slightly faster at the final energy level). The bad news is that the models developed to date to describe the dynamics of global energy systems cannot replicate decarbonization (or, for that matter, any other major historical transitions described in the preceding sections) in any endogenous way, regardless of whether the models embrace an economic ("top-down") or engineering ("bottom-up") perspective. But the state of the art in modeling energy–technology–environment interactions is expanding rapidly. Figure 13 illustrates how future decarbonization rates depend on the representation of technology. In the (currently dominant) view of a largely static technology base, past and future decarbonization cannot be modeled without resort to exogenous modeling "fudge factors" and constraints. Even treating the traditional energy model variables of resource availability, demand, and environmental constraints (taxes) as uncertain, justifying diversification away from current dominant fossil technologies, does not change this picture significantly. Currently, the only way in which to replicate historical decarbonization and generate scenarios of future decarbonation in an endogenous way is to use models that incorporate full uncertainty, including those of increasing returns to adoption for new energy technologies. Thus, innovation is key—even if uncertain—both for future scenarios and for explaining the transitions that have taken place in global energy use since the onset of the industrial revolution.

SEE ALSO THE FOLLOWING ARTICLES

Cultural Evolution and Energy • Development and Energy, Overview • Economic Growth, Liberalization, and the Environment • Energy Ladder in Developing Nations • Environmental Change and Energy • Global Energy Use: Status and Trends • International Comparisons of Energy End Use: Benefits and Risks • Population Growth and Energy • Sustainable Development: Basic Concepts and Application to Energy • Technology Innovation and Energy • Women and Energy: Issues in Developing Nations

Further Reading

Arthur, W. B. (1994). "Increasing Returns and Path Dependence in the Economy." University of Michigan Press, Ann Arbor, MI.

Flavin, C., and Lenssen, N. (1994). "Power Surge: Guide to the Coming Energy Revolution." Worldwatch Institute, Washington, DC.

Fouquet, R., and Pearson, P. J. G. (1998). A thousand years of energy use in the United Kingdom. *Energy J.* **19**(4), 1–41.

Grübler, A. (1998). "Technology and Global Change." Cambridge University Press, Cambridge, UK.

Grübler, A., and Gritsevskyi, A. (2002). A model of endogenous technological change through uncertain returns on innovation. *In* "Technological Change and the Environment" (A. Grübler, N. Nakicenovic, and W. D. Nordhaus, Eds.), pp. 337–402. RFF Press, Washington, DC.

Grübler, A., and Nakicenovic, N. (1996). Decarbonizing the global energy system. *Technol. Forecasting Social Change* **53**, 97–100.

Grübler, A., Nakicenovic, N., and Nordhaus, W. D. (eds.). (2002). "Technological Change and the Environment." RFF Press, Washington, DC.

Holdren, J. P., and Smith, K. (2000). Energy, the environment, and health. *In* "World Energy Assessment," pp. 61–110. UN Development Program, New York.

MacDonald, G. J. (1990). The future of methane as an energy source. *Annu. Rev. Energy* **15,** 53–83.

Maddison, A. (2003). "The World Economy: A Millennial Perspective." Overseas Press, New Dehli, India. (For Organization for Economic Cooperation and Development, Paris.)

Marchetti, C. (1985). Nuclear plants and nuclear niches. *Nucl. Sci. Eng.* **90,** 521–526.

Nakicenovic, N., Grübler, A., and McDonald, A. (eds.) (1998). "Global Energy Perspectives." Cambridge University Press, Cambridge, UK. (For International Institute for Applied Systems Analysis [IIASA] and World Energy Council [WEC])

Nakicenovic, N., Grübler, A., Ishitani, H., *et al.* (1996). Energy primer. *In* "Climate Change 1995" (R. T. Watson, M. C. Zinyowera, R. H. Moss, and D. J. Dokken, Eds.), pp. 75–92. Cambridge University Press, Cambridge, UK. (For Intergovernmental Panel on Climate Change [IPCC])

Nakicenovic, N., *et al.* (2000). "Special Report on Emissions Scenarios (SRES). Cambridge University Press, Cambridge, UK. (For Intergovernmental Panel on Climate Change [IPCC])

Population Division of the United Nations. (2002). "World Urbanization Prospects: The 2001 Revision." United Nations, New York.

Reddy, A. K. N., and Goldemberg, J. (1990). Energy for the developing world. *Sci. Am.* **263**(3), 110–118.

Smil, V. (1994). "Energy in World History." Westview, Boulder, CO.

Smith, K. R. (1993). Fuel combustion, air pollution exposure, and health. *Annu. Rev. Energy Environ.* **18,** 529–566.

Weinberg, A., and Hammond, P. (1972, March). Global effects of increased use of energy. *Bull. Atomic Sci.,* pp. 5–8, 43–44.

Transportation and Energy, Overview

DAVID L. GREENE
Oak Ridge National Laboratory
Knoxville, Tennessee, United States

1. **Introduction**
2. **Societal Concerns**
3. **History of Propulsion Technology and Energy**
4. **Patterns and Trends in Transportation Energy Use**
5. **Energy Efficiency and Energy Consumption**
6. **Energy Sources**
7. **Transportation Energy in the Future**

Glossary

alternative fuel A nonpetroleum energy source used to power transportation vehicles; most often applied to road vehicles.

energy density The ratio of the energy content of a fuel or storage capacity of a battery to its volume or mass.

energy intensity The quantity of energy used per unit of transportation service provided; most often measured in energy use per passenger-mile (or passenger-kilometer) or per ton-mile (or ton-kilometer).

fuel economy Vehicle-miles (or vehicle-kilometers) produced per unit of energy consumed, most often measured as miles per gallon of fuel consumed or as liters of fuel consumed per 100 km traveled.

hybrid vehicle A vehicle that employs two sources of propulsion, especially an internal combustion engine and an electric motor. Some definitions also require that energy be recaptured during braking and stored in a battery or other energy storage device for later use and that the internal combustion be shut off when it would otherwise be idling.

mode of transportation A particular combination of transportation vehicles and media. The most basic distinction is among highway, rail, waterborne, air, and pipeline transport modes.

Transportation is the movement of people or objects from one place to another. Energy is essential to transportation. Moving people, objects, and the vehicles that carry them requires the application of force to overcome the forces of inertia and friction. The application of force over a distance is work and energy is the ability to perform work. Modern, mobile societies use vast amounts of energy for passenger travel and freight transport; approximately one-fourth of the total energy used by humans is for transportation. Today, a century after the introduction of the internal combustion engine as a means of propulsion, approximately 95% of transportation energy use is supplied by petroleum. Petroleum's advantages include low cost, high energy density, and up to the present, abundant supply. Burning vast quantities of petroleum to power transportation does significant damage to the environment in the form of air pollution, spills on water and land, and greenhouse gas emissions. The near-total dependence of transportation systems on petroleum has also been a source of strategic and economic costs. As the world's resources of fossil petroleum are depleted, the world's ever-increasing demand for mobility raises questions about where the energy to power transportation can be obtained. Solving these problems while maintaining and expanding the mobility on which the world's economies have come to depend is the central problem of transportation and energy.

1. INTRODUCTION

Modern, mobile societies use vast amounts of energy for passenger travel and freight transport. Global transportation energy amounted to 4 EJ in the year 2000, one-fourth of the final energy used by human societies for all activities. Almost all (95%) of the energy use for transportation was derived from the

 © 2004 Elsevier Inc. All rights reserved.

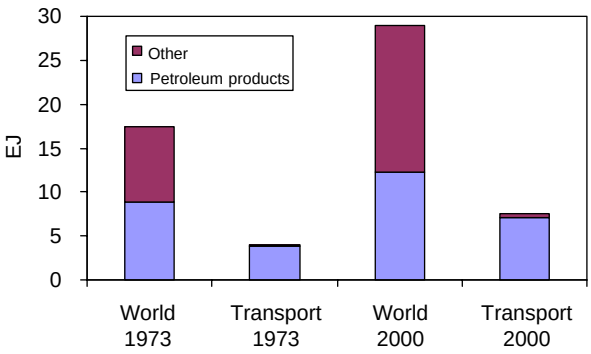


FIGURE 1 World final energy use.

combustion of petroleum (Fig. 1). Burning such vast quantities of petroleum does significant damage to the environment. For most developed economies, it also creates an economic and strategic dependence on imported oil. As supplies of conventional oil are depleted, the world's ever-increasing demand for mobility raises questions about where energy for transportation will come from in the 21st century and what the consequences of its use will be.

Energy is essential for transportation. Transportation, the movement of people or objects from one place to another, is a form of work, as that term is defined in physics. Moving transportation vehicles requires the application of force to overcome inertia and continuous force to keep vehicles moving against the opposing forces of friction from rolling resistance, aerodynamic drag, or hydraulic drag. Work is defined as the application of force through a distance and so the essence of transportation is work. Energy is the ability to do work and thus transportation cannot be accomplished without using energy.

Energy for transportation vehicles can be supplied externally (e.g., sailboats, electric trains, trolleys) or carried on board (e.g., automobiles, jet aircraft). In the 20th century, the internal combustion engine became the predominant source of power for transportation vehicles, displacing animal, wind, and steam power. Internal combustion engine vehicles must carry a supply of energy on board. When energy is carried on board a vehicle, a high value is placed on energy density because stored energy displaces payload (by weight and volume) and because storing more energy on board facilitates speed and the convenience of nonstop operation. After a brief period of experimentation early in the 20th century, liquid fuels derived from petroleum emerged as the clearly preferred form of energy for internal combustion engines. By midcentury, petroleum supplied 90% of the energy for transportation vehicles. Petroleum fuels have four key advantages: (1) they have high energy density per volume and weight; (2) they are easily handled and stored because they are liquids under ambient conditions; (3) throughout most of the 20th century, petroleum resources have been in abundant supply; and (4) they are relatively inexpensive.

2. SOCIETAL CONCERNS

Consumption of petroleum fuels has also brought problems: (1) air pollution from vehicle exhausts is harmful to human health to animals and to vegetation; (2) greenhouse gas (GHG) emissions, mostly in the form of carbon dioxide, have begun to alter the earth's climate; (3) the concentration of the world's oil resources in the Persian Gulf region has contributed to international conflict and produced oil price shocks that triggered economic recessions; and (4) crude oil and fuel spills and leaking storage tanks have polluted water bodies and groundwater.

Enormous technological progress has been made in reducing the pollution caused by the combustion of petroleum fuels in motor vehicles. A properly operating model year 2000 automobile produces approximately 1% of the pollutant emissions per kilometer of vehicle travel of a vehicle built in 1965. New, tighter emissions standards to be implemented in the next 5 years require further reductions in emissions. Still, because of the continued growth in the number of vehicles and kilometers traveled and because of old and malfunctioning vehicles, air pollution by motor vehicles remains a serious health concern, especially in the megacities of developing economies.

The overwhelming majority of GHG emissions from transportation are in the form of carbon dioxide (Fig. 2). A typical passenger car will emit several times its weight in carbon dioxide each year. Because the carbon content of petroleum fuels varies little, carbon dioxide emissions depend directly on how much transporting is done and the energy efficiency of transportation vehicles. Globally, transportation accounts for one-fifth of carbon dioxide emissions. Thus far, reducing transportation's GHG emissions has proven to be much more difficult than reducing air pollutant emissions because of the lack of economically competitive, alternative, low-carbon fuels.

Because no fuel has been able to displace petroleum, increasing energy efficiency and reducing motorized vehicle travel are the most effective

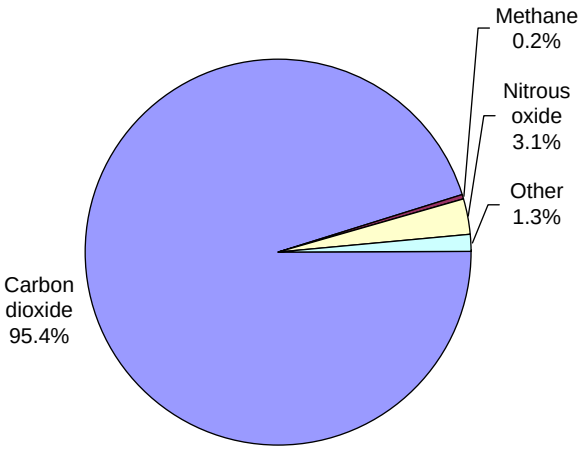


FIGURE 2 Transportation greenhouse gas emissions, by type of gas, in the United States in the year 2000.

remaining strategies for controlling transportation's GHG emissions over the next few decades. Since 1970, increasing energy efficiency has proven to be more effective than decreasing vehicle travel. From 1970 to 2000, highway vehicle travel in the United Stated increased by almost 150%. At the same time, the fuel economy of passenger cars and light trucks in use increased by 55% (from 13.0 to 20.1 miles per gallon). The increase in passenger car and light truck fuel economy in the United States from 1975 to 2000 reduced petroleum consumption by 2.8 million barrels per day (6.25 EJ per year), equivalent to 100 million metric tons of carbon emissions.

Driven by fuel economy standards and higher fuel prices, the fuel economy of new passenger cars sold in the United States doubled from 1974 to 1985 and that of light trucks increased by over 50%. U.S. fuel economy standards have not changed significantly since 1985 and fuel prices decreased. As a result, both new car and light truck fuel economies have remained approximately the same and the combined fuel economy of light-duty vehicles has declined due to the increased popularity of light trucks. It takes approximately 15 years for the total fleet of light-duty vehicles in use to reach the same level of fuel economy as new vehicles (older, less efficient vehicles must be scrapped and replaced by newer ones). As a result, total light-duty vehicle fuel economy has continued to increase slowly, reaching 19.5 miles per gallon in 1991 and 20.1 miles per gallon in 2000. Thus, it appears that in the future, carbon dioxide emissions will increase at the same rate as vehicle travel grows.

World resources of petroleum are highly geographically concentrated. Nations belonging to the Organization of Petroleum Exporting Countries (OPEC) hold almost 80% of the world's proved reserves of conventional oil. Proved reserves, however, are a stock held for future production and do not fully reflect the extent of developable oil resources. Considering estimated undiscovered resources and the potential for technological progress in discovery and production, OPEC members still hold more than one-half of all the conventional oil that remains (Fig. 3). The willingness of OPEC members to use their position of market dominance has enabled them to raise oil prices at the expense of oil-producing nations and has contributed to international conflict.

Though estimates vary, it appears that human beings have consumed approximately one-third of all the conventional petroleum that exists. Because consumption has been steadily increasing, it will take no more than a few decades to reach the point at which 50% of the earth's ultimately recoverable conventional petroleum resources have been consumed. If that point is reached, transportation may have to begin a transition to alternative sources of energy if world mobility is to continue to increase. One option is to make conventional fuels out of other fossil carbon resources. There are vast deposits of unconventional oil in the form of heavy oil in Venezuela and oil sands in Canada that are beginning to be developed to supply petroleum fuels at competitive prices. In addition, excellent conventional transportation fuels can be made at a higher cost from natural gas or coal via Fischer–Tropsch or similar conversion processes. The United States also contains enormous reserves of oil shale, which could be developed at much greater expense and with increased environmental impacts. On the other hand, given the right technological advances, the energy for 21st century transportation systems could be

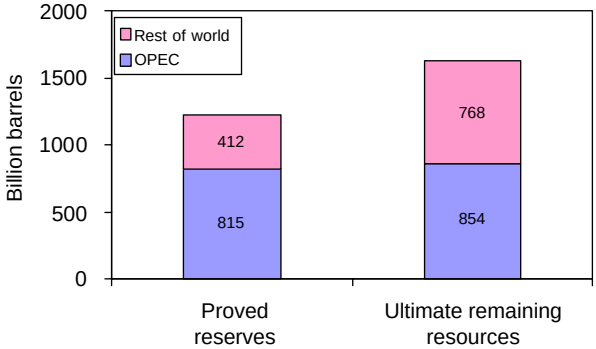


FIGURE 3 Estimated quantity and distribution of world's remaining oil resources.

supplied by very different fuels, such as hydrogen to power fuel cell vehicles or electricity for high-speed railways.

3. HISTORY OF PROPULSION TECHNOLOGY AND ENERGY

The history of transportation and energy is one of finding ways to apply greater power and to reduce the forces opposing motion, so that ever larger payloads could be moved at greater speeds. Defined as the tools devised by human beings to accomplish the transportation of persons and goods, vehicles could include everything from domesticated animals to jet aircraft. At a minimum, a vehicle consists of a structure for supporting a payload and a means of using energy to create motion. A wide variety of forms of energy have been applied to make vehicles go. The goals of increasing speed and payload have driven transportation vehicles not only to use more energy, but to use it more effectively by increasing the efficiency with which energy is converted to propulsion and by minimizing the forces of inertia and friction.

From the earliest times, human beings sought to augment the mobility that their legs provided by domesticating animals and harnessing their power. Horses, camels, oxen, donkeys, and even elephants and dogs have been employed by humans to pull or carry. Animals could be ridden or laden, but generally this did not take full advantage of their power. Larger amounts could be transported by harnessing animals to a sled or by harnessing them to two poles and dragging a load suspended between them. The Native Americans of the northern plains used such vehicles (with pointed pole ends to reduce friction) until the 19th century. The high friction produced by dragging across the ground was dramatically reduced by the invention of the wheel. Wheeled vehicles in the form of carts and chariots have been in existence since at least the second millennium BC. For the next 3000 years, animal power provided the energy for ground transportation vehicles and technological advances focused on reducing friction through the invention of rounder, lighter spoked wheels and turnable front axles and through the improvement of road surfaces and reduction of grades.

Waterborne vessels are known to have been used for transporting people and goods as long ago as 4000 BC. Motive power was provided by rowers, by the wind, and on rivers by downstream currents. To make oared boats larger and faster, two and three banks of oars were used. At the same time, hull designs were elongated and streamlined to reduce hydraulic drag. From a single square or triangular sail on a mast, more sails and more masts were added to increase the amount of wind energy that could be used. Sails were made maneuverable to allow headway to be made under a greater range of wind directions and to increase speed. Larger and faster sailing ships allowed humans to travel around the earth for the first time in the 15th century. On inland waterways, the construction of canals with towpaths allowed mules or oxen to take over from human power as a source of propulsion.

In the latter half of the 18th century, the invention of the (external combustion) steam engine began a revolution in ground and water transportation. For the first time, a vehicle's source of energy, generally coal, could be carried on board. The first steam engines were large and, because they operated at low pressures, provided relatively little power for their weight and size. Because of their low power density, the first use of steam engines in transportation was on ships. Early paddle-wheel-driven steamboats were no faster than sailing ships but had the advantages of more reliable propulsion and the ability to travel in any direction regardless of which way the wind blew. In the 19th century, improvements such as higher pressure engines using double and triple expansion to maximize the work derived from steam, the screw propeller, and steel hulls that could withstand the greater forces produced by the more powerful engines allowed the steamship to flourish. Just before the turn of the century, steam turbine engines were installed on ocean liners and proved to be somewhat faster than the reciprocating piston steam engines.

On land, early steam engines were installed on vehicles that operated on roadways and on fixed guideways. Operation on steel rails proved the most effective because of the low friction between wheel and rail. With only 1 atm of pressure, the earliest steam engines had too little power for their size to be effective propulsion systems for land transport. By the beginning of the 19th century, however, steam engines with 10 times the pressure of the early engines had been invented and the age of rail began. Rail's dominance of land transport lasted a full century until a new engine emerged with much higher power density, its energy supplied by liquid fuel.

The invention of the internal combustion engine has proven to be the single most important event in

the history of transportation. Almost immediately after the first successful internal combustion engines were developed in the 1870s, they were installed on tricycles, bicycles, and carriages; after only slightly more than 10 years, the first automobiles appeared on the roads of Europe and North America. In 1901, 425 automobiles were sold in the United States. They were chiefly toys, affordable only by the wealthiest Americans. The introduction of the Model T by Henry Ford in 1908 changed that forever. By 1929, 27 million automobiles traveled a largely inadequate U.S. road system.

New forms of propulsion that could displace the internal combustion engine in the 21st century are on the horizon. Fuel cells powering electric motors have made enormous progress as a potential propulsion system for automobiles. Fuel cells produce electricity by a chemical reaction between hydrogen (the fuel) and oxygen (usually supplied from the air). Unlike a battery, the fuel cell can supply electricity indefinitely as long as new fuel and oxygen are provided to the cell. Whereas modern gasoline internal combustion engines can achieve efficiencies of up to 30%, fuel cells coupled with electric motors can convert up to 55% of the energy in fuel to useful work. The hydrogen fuel cell is an inherently clean technology; the only product of its chemical reaction is water. In addition, hydrogen can be produced from a number of energy sources including natural gas, coal, biomass, or, via electrolysis, any source of electricity, including nuclear, wind, or solar power.

Magnetic levitation could someday revolutionize fixed guideway transportation. Magnetic forces can levitate a train above its guideway for noncontact, almost frictionless suspension, guidance, and propulsion at speeds up to 500 km/h. Successful technical demonstrations of maglev trains have been made in Japan, Europe, and the United States.

For aircraft to fly at speeds greater than mach 5 (hypersonic speed, greater than 1.6 km/s), they will need new propulsion systems and new fuels. Aerospace researchers are developing ramjet/scramjet engines to achieve these hypersonic flight speeds. Unlike conventional jet engines, which mechanically compress air prior to combustion, ramjets use the speed of the aircraft itself to compress the air that travels through the jet engine at subsonic speeds. In the scramjet, the air travels through the engine at supersonic speeds. This requires a fuel with an extremely fast rate of combustion. Petroleum-based fuels combust too slowly, so a new fuel, such as liquid hydrogen, will be needed.

4. PATTERNS AND TRENDS IN TRANSPORTATION ENERGY USE

By far, highway vehicles consume more energy than any other mode of transport. In both the developed economies belonging to the Organization for Economic Cooperation and Development and developing economies, road vehicles account for 80% or more of total transportation energy use (Fig. 4). In the Unites States, light-duty vehicles (passenger cars and light trucks) alone account for more than one-half of the transportation sector's total energy use. Air travel is the second largest and fastest growing energy user. Between the two, they consume more than 90% of transportation energy use. Other modes, rail, pipeline, and water, account for approximately 1–2% each.

Despite the oil price shocks of the 1970s and 1980s, world transportation energy use more than doubled from 1970 to 2000 (Fig. 5). Energy use in developing economies grew faster (3.5% per annum) than energy use in the developed economies (2.2% per year). If the growth in transportation activity continues at this pace, and if significant improvements in energy efficiency are not made, the world will soon be searching for alternative sources of energy to fuel the growth of its transportation systems.

5. ENERGY EFFICIENCY AND ENERGY CONSUMPTION

Increasing the energy efficiency of transportation vehicles has always been a key motivation for

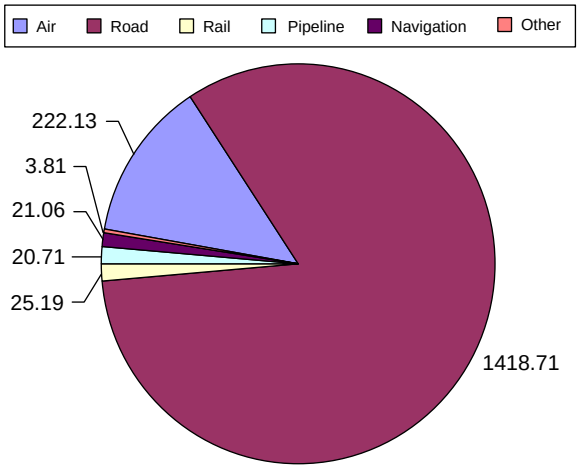


FIGURE 4 World transportation energy use by mode, in the year 2000 (million metric tons of oil equivalent).

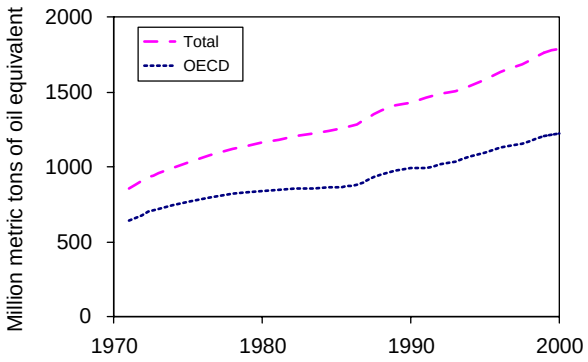


FIGURE 5 World transportation energy use. OECD, Organization for Economic Cooperation and Development.

TABLE I

Typical Engine and Transmission Efficiencies

Technology	Urban driving	Highway driving
Gasoline engine	10–15%	25%
Diesel engine	15–22%	35%
Automatic transmission	70–80%	85%
Manual transmission	85–90%	90–95%

Source. Reprinted from Massachusetts Institute of Technology and Charles River Associates, Inc. (2002), Table 2–3, with permission.

advances in energy technology. It is useful to distinguish between propulsion efficiency (the efficiency with which energy delivered to a vehicle is converted into useful work) and vehicle efficiency (the energy required from a vehicle's propulsion system relative to the transportation services it delivers). In some circumstances (such as comparing the environmental impacts of alternative technologies and fuels), it is important to consider the life cycle or "well-to-wheels" energy efficiency of a complete fuel and vehicle system.

Propulsion efficiency is the ratio of the output of an engine in useful work to the energy content of the fuel used. The second law of thermodynamics requires that this ratio be less than 1 because in any reaction entropy must increase. Many other factors prevent engines from reaching thermodynamic efficiency limits, including internal friction, limits on temperatures and pressures, and energy requirements of auxiliary systems. All of these areas constitute possibilities for technological advances to increase efficiency. The spark-ignited, gasoline, internal combustion engine of a modern automobile, in typical urban driving, converts less than 20% of the energy in gasoline into useful work at its crankshaft (see Table I). Various engine losses (especially heat rejected to the engine's cooling system and exhaust) account for more than 60% of the energy in the fuel; fuel consumed during idling, braking, and coasting accounts for almost 20%. Although idling and braking losses are not a function of the conversion efficiency of the engine, they are an important component of the overall efficiency of the propulsion system. It is possible to recapture much of a vehicle's kinetic energy normally dissipated as waste heat during braking. Hybrid propulsion systems that combine an internal combustion engine and an electric motor minimize these losses by (1) shutting off the internal combustion engine during idling, coasting, or braking and restarting it quickly with the electric motor when needed and (2) using the electric motor as a generator during braking to convert kinetic energy into electricity, which is then stored in a battery for future use by the electric motor.

For any given driving cycle (a driving cycle is typically defined by a schedule of velocity versus time at a given load and under ambient conditions), the energy required to propel a vehicle is determined by the forces that oppose its motion. The amount of inertial energy is determined by the mass of the vehicle, since for the same rate of acceleration proportionately increased force must be applied as mass increases (Force = Mass × Acceleration). To a lesser degree, this also applies to the moving parts of an engine, which must also be accelerated to increase engine speed. Reducing the mass of these moving parts reduces friction and the inertial work internal to the engine. Force to overcome inertia must also be applied when vehicles change direction.

The remaining forces opposing motion are frictional. All vehicles moving within the earth's atmosphere must overcome aerodynamic drag. Aerodynamic drag becomes increasingly important as vehicle speed increases because drag grows with the square of speed. As a result, the energy required per unit of distance to overcome aerodynamic drag increases in proportion to vehicle speed. At any given speed, drag can be reduced by means of slipperier vehicle designs with lower drag coefficients. A bluff design with its full frontal area perpendicular to the direction of motion would have a drag coefficient of 1.0. Modern passenger cars have drag coefficients on the order of 0.3, whereas large truck drag coefficients are often 0.7 or higher. Advanced passenger car designs have achieved drag coefficients of 0.15. For motor vehicles, the other

major frictional force, rolling resistance, is caused by friction in wheels, axles, transmission, brakes, and the contact between wheels or tires and roads or tracks. Rolling resistance increases directly with speed so that per unit of distance traveled it is approximately constant. In a modern automobile, most rolling resistance is due to the deformation of tires as they meet the pavement.

Losses in the transmission and other driveline components of an automobile use slightly more than 5% of the total energy in fuel. Of the 12–13% of the original energy in fuel that reaches the wheels, approximately 6% goes to overcoming inertia, 3% to overcoming aerodynamic drag, and 4% to rolling resistance. In highway driving, the distribution is quite different, reflecting the greater importance of frictional forces at highway speeds: 2% to overcoming inertia, 7% to rolling resistance, and 11% to overcoming aerodynamic drag.

Commercial aircraft have settled on the gas turbine engine fueled by kerosene as the propulsion system of choice. Unlike the engines used in automobiles that employ reciprocating pistons to compress air prior to combustion, turbines use fan blades rotating about an axis to achieve compression. Turbine engines are more efficient for the constant speed operating conditions typical of most air travel. The overall efficiency of gas turbine propulsion is determined by its thermodynamic efficiency (which increases as the temperatures and pressures that can be achieved within the engine increase) and propulsion efficiency. Early turbojets achieved thrust solely by accelerating the air that passed through the engine's combustor. By accelerating air out the back of the jet engine, an equal and opposite force that accelerates the aircraft forward is created. The turbofan engine greatly increased propulsion efficiency by also producing thrust by means of fan blades driven by the turbojet engine to accelerate air passing around the engine. By 1990, turbofans employed bypass ratios of 5–6; in 2003, high-bypass engines have ratios of 8–10. It may be possible to increase ratios to 15–20 with further gains in propulsion efficiency. The overall energy efficiency of aircraft propulsion has improved dramatically over the past 50 years. In 2003, energy use per unit of thrust produced is less than two-thirds of what was required in 1960.

Aircraft must continuously overcome the force of gravity, which is proportional to their mass. They do so by using an airfoil to create lift. The airfoil also creates a frictional force (aerodynamic drag) that opposes the forward motion of the aircraft. The ratio of the force of lift to that of drag (L/D) is the key measure of the aerodynamic efficiency of aircraft. Early jet aircraft had L/D ratios of approximately 16. Today, long-range aircraft achieve L/D ratios of 20 and short- to medium-range aircraft have ratios of approximately 18.

The overall energy efficiency of air travel has also benefited from increased aircraft size (seats per aircraft) and higher load factors (average occupancy per seat). In 1970, airlines used 6.5 MJ per passenger-kilometer of service. In 2000, only 2.5 MJ was needed to produce 1 passenger-kilometer of service, approximately the same as the average energy efficiency of travel by passenger car.

Transportation modes differ in energy efficiency, but perhaps not as much as might be expected. Measuring efficiency for modal comparisons is made difficult by differences in the services that modes provide. The private passenger car provides the ultimate in personal convenience, offering privacy, point-to-point transport, and a completely flexible schedule. However, passenger cars cannot yet cruise at 800 km/h the way jet aircraft can. Also, although railroads consume far less energy per ton of freight carried, they cannot match trucks for point-to-point service and flexibility in scheduling. Nor can either alone guarantee overnight transcontinental delivery as can aircraft combined with trucks. Such differences in service are not reflected in readily measured energy efficiency indicators, such as energy use per passenger-kilometer or ton-kilometer. Bearing these limitations in mind, data for the United States indicate that energy intensities of passenger modes have tended to decline over time and suggest a degree of convergence across modes (Fig. 6). Air travel, the newest mode in terms of technological maturity, achieved the greatest reduction, most of it prior to 1980. Passenger car and light truck energy intensities declined through 1990, whereas most other modes' energy intensities remained the same or even increased.

For freight modes, data are inadequate to calculate energy intensities for truck and waterborne freight. Energy use per ton-kilometer transported by rail in the United States was cut by one-half between 1970 and 2000. These aggregated comparisons obscure significant variations from place to place and across types of service within modes.

6. ENERGY SOURCES

Early transportation vehicles relied on animal power, wind, flowing water, and gravity. With the motorization of transportation vehicles, combustible

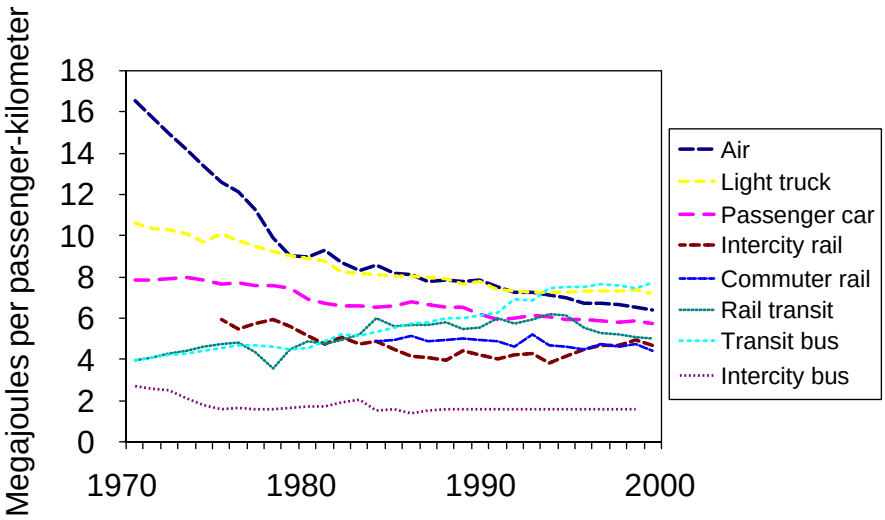


FIGURE 6 Energy intensities of passenger modes in the United States.

fuels and electricity became important sources of energy. External combustion engines, such as the steam engine (piston and turbine), played an important role in water and rail transport during much of the 19th century and early part of the 20th century. Solid fuels, such as coal and wood, are suitable for external combustion because the speed and completeness of combustion are not of great importance. The low cost and abundance of coal made it the fuel of choice for steam engines. The combination of the development of the internal combustion engine and discovery of vast resources of petroleum led to petroleum displacing nearly every other form of energy as the predominant energy source for transportation. Internal combustion engines favor volatile liquid and gaseous fuels because combustion must take place rapidly in a confined space (favoring fuels that vaporize easily or are already gaseous) and combustion products must be fully expelled in a gaseous medium.

The first electric automobile was a tricycle powered by lead–acid batteries, introduced to the world in 1881. An 1882 version is known to have carried 68 kg of batteries with an energy storage of 1.5 kWh, enough for a range of 15–40 km traveling at approximately 15 km/h. One and one-half kilowatt-hours is equivalent in energy to approximately one-sixth of a liter of gasoline. In 1900, electric automobiles had a tremendous advantage in energy efficiency over their competitors. Steam-powered automobiles were getting approximately 2.5 km/liter, gasoline-engine vehicles were getting approximately 5 km/liter, whereas electric cars could travel up to 75 km per energy equivalent of 1 liter of gasoline. Unfortunately, electric vehicles could store only approximately 30 kWh of energy on board (equivalent to approximately 3 gal of gasoline) and took hours to refuel.

The electric vehicle arrived first and, initially, competed well against the noisy and frequently unreliable gasoline-engine automobiles. But by 1912, there were 33,000 electric vehicles in use in the United States and over half a million gasoline-powered vehicles. Electric vehicles had distinct advantages in quietness, cleanness, and reliability, but gasoline vehicles had the decisive advantage: energy storage. The costs of electricity and gasoline were relatively minor factors. By either weight or volume, more than 100 times as much energy could be stored on board a vehicle in the form of gasoline versus electricity stored in batteries. In addition, batteries were, and still are, expensive storage devices in comparison to a gasoline tank. Primarily due to the cost of batteries, electric vehicles in 1913 cost $2000 to $3000, whereas gasoline-engine automobiles could be purchased for less than $1000.

Energy density remains a problem for alternative fuels. The energy densities of gaseous and liquid fuels are compared in Fig. 7. Diesel fuel has 11% more energy per volume than gasoline and thus requires only 90% of the volume to store the same amount of energy. Two alcohols that have been widely used as motor fuels, ethanol and methanol, require 50 and 100% greater volumes, respectively, to store the same amount of energy. Compressed natural gas (methane) needs 4 liters of storage space to equal the energy in a single liter of gasoline and liquefied hydrogen requires almost the same volume. Compressed hydrogen at 5000 psi needs 16 times the

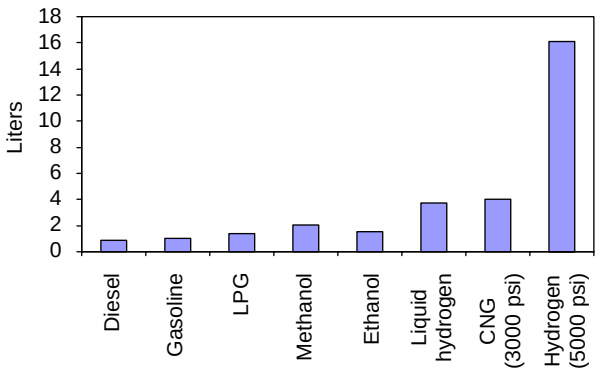


FIGURE 7 Volume required to store the energy equivalent of 1 liter of gasoline. CNG, compressed natural gas; LPG, liquefied petroleum gas.

volume. Batteries have even lower energy densities but are more often limited by weight than by volume. Three hundred kilograms of nickel metal hydride (NiMH) batteries (the kind used in hybrid vehicles today) are required to store the energy in 1 kg of gasoline. The most advanced NiMH batteries available could store that much energy in 170 kg of batteries.

Another problem faced by alternative fuels is the difficulty of displacing a fuel with an enormous, efficient infrastructure. The phrase "chicken or egg" dilemma has been used to describe the reluctance of energy companies to invest in fuel supply infrastructure when there are no vehicles that use the fuel and the reluctance of vehicle purchasers to buy, and automobile companies to manufacture, alternative fuel vehicles when there is no fuel supply infrastructure to support them. The economic barriers to displacing an established technology are real and substantial. In addition to the chicken or egg dilemma, initial problems include the lack of diversity of choice for novel technologies and large economies of scale in automobile production.

Alternative fuels without a decisive advantage over petroleum are not likely to succeed. Petroleum fuels and internal combustion engines have proven to be resilient and adaptable. In the 1980s, demands for cleaner cars seemed to favor inherently cleaner burning fuels, such as natural gas and alcohols. But the petroleum and motor vehicle industries responded by developing advanced emissions controls and a new reformulated gasoline, whose chemistry had been tailored to meet the emissions goals established by alternative fuels.

After nearly a quarter century of efforts to promote alternative fuels in the United States, there were only 456,000 alternative fuel vehicles in use out of 213 million (0.2%) in total and alternative fuels constituted only 0.2% of total fuel consumption by highway vehicles (1.3 billion out of 680 billion liters). Nonpetroleum energy has had greater success in blends with gasoline. In the United States in 2000, 11 billion gasoline energy equivalent liters of methyl tertiary butyl ether (made primarily from natural gas and butane) and 4 billion gasoline energy equivalent liters of ethanol (made from corn) were blended with gasoline.

7. TRANSPORTATION ENERGY IN THE FUTURE

What will transportation's energy future look like? Will demand for mobility continue to grow indefinitely or will a saturation point be reached and growth slow to a halt? Clearly, there are no certain answers to such questions. Yet transportation research has developed some useful principles that can help one peer into the future. Future world transportation energy use may well be determined more by what happens in the world's developing economies than by the evolution of developed economies. Not only is it likely that population growth and economic growth will be most rapid in the developing world over the next 50 years, but economic growth is likely to have the greatest impact on mobility in the developing world through its impact on the motorization of transport.

The average human being spends 1.1 h per day traveling. In addition, households not owning a personal car devote, on average, 3–5% of their expenditures to transportation. Typical car-owning households spend 10–15%. As income grows, the average time spent traveling remains approximately constant, but money expenditures grow. The increased expenditures go to faster (and more convenient) motorized modes of transport.

Whereas motorized transport is predominant in developed economies, walking, cycling, and mass transport are still the most commonly used modes in the developing economies. The average resident of Africa or China spends on the order of 10 min/day or less in a motorized form of transport. The average North American spends more than 1 h/day. The potential for growth is enormous and so are the rates of growth. In the developing world, automobile ownership rates are low: less than 50 vehicles per 1000 persons in China and Africa, 100 vehicles per 1000 persons in Latin America, and less than 150 vehicles per 1000 persons in the nations of the

former Soviet Union. By contrast, ownership rates exceed 500 vehicles per 1000 persons in Europe, are over 600 vehicles per 1000 persons in Japan, and top out at 800 vehicles per 1000 persons in the United States. In terms of motorization, China and Africa are approximately where the United States was in 1915 in terms of automobile ownership. In the developed world, travelers are switching from cars to airplanes to obtain more mobility out of their 1.1 h (average) of daily travel time. In the developing world, travelers are switching from animal and human power to motor vehicles. The impact on transportation energy use is likely to be enormous.

What energy sources will fuel the future growth of mobility? The world's fossil resources, especially coal and oil shale, are more than adequate to supply synthetic petroleum fuels for expanding mobility well beyond the end of this century. But will the world choose that path? It is also possible to manufacture hydrogen or electricity from those same fossil fuels or from almost any other source of energy. But will the technological challenges of fuel cells and hydrogen storage be solved? Will passengers and freight soon speed along at 500 km/h in magnetically levitated trains driven by linear electric motors? Or will unimagined technologies provide the increased speed, capacity, comfort, and convenience that human societies have sought ever since the first draft animals were domesticated? Regardless, an increasingly mobile world will demand more transportation, and transportation will require energy.

SEE ALSO THE FOLLOWING ARTICLES

Aircraft and Energy Use • Conservation Measures for Energy, History of • Energy Efficiency, Taxonomic Overview • Fuel Economy Initiatives: International Comparisons • Hybrid Electric Vehicles • Intelligent Transportation Systems • Internal Combustion Engine Vehicles • Lifestyles and Energy • Marine Transportation and Energy Use • Passenger Demand for Travel and Energy Use • Transportation and Energy Policy • Transportation Fuel Alternatives for Highway Vehicles

Further Reading

Ausubel, J. H., Marchetti, C., and Meyer, P. S. (1998). Toward green mobility: The evolution of transport. *Eur. Rev.* **6,** 137–156.

Davis, S.C., and Diegel, S. (2002). "Transportation Energy Data Book," 22nd ed., ORNL-6967. Oak Ridge National Laboratory, Oak Ridge, TN.

Greene, D. L. (1996). "Transportation and Energy." Eno Foundation for Transportation, Washington, DC.

Massachusetts Institute of Technology and Charles River Associates, Inc. (2002). "Mobility 2001: World Mobility at the End of the Twentieth Century and Its Sustainability." World Business Council for Sustainable Development, Geneva, Switzerland. Available at www.wbcsdmobility.org.

Schafer, A., and Victor, D. G. (2000). The future mobility of the world population. *Transport. Res. A* **34,** 171–205.

Schipper, L. (1995). Determinants of automobile use and energy consumption in OECD countries. *Annu. Rev. Energy Environ.* **20,** 325–386.

U.S. Congress, Office of Technology Assessment. (1994). "Saving Energy in U.S. Transportation," OTA-ETI-589. U.S. Government Printing Office, Washington, DC.

U.S. Department of Transportation, Bureau of Transportation Statistics. (1996). "Transportation Statistics Annual Report 1996: Transportation and the Environment." Washington, DC.

Weiss, M. A., Heywood, J. B., Drake, E. M., Schafer, A., and AuYeung, F. F. (2000). "On the Road in 2020: A Life-cycle Analysis of New Automobile Technologies," Energy Laboratory Report MIT EL 00-003. Massachusetts Institute of Technology, Cambridge, MA.

Yergin, D. (1992). "The Prize: the Epic Quest for Oil, Money and Power." Simon and Schuster, New York.

Transportation and Energy Policy

LEWIS M. FULTON
International Energy Agency
Paris, France

1. **Recent Trends and a Forecast for the Next 30 Years**
2. **Looking behind the Energy Trends**
3. **Concerns for Transport Energy Sustainability**
4. **Is an Alternative Future Possible?**
5. **Conclusions**

Glossary

ASIF matrix Basic structure of travel, providing a system for analysis; travel energy use is a function of the ASIF components: activity, structure, intensity, and fuel type.

fuel economy Fuel consumption per kilometer of vehicle travel (more commonly used in North America than elsewhere).

fuel efficiency The efficiency with which vehicles convert energy into movement—not necessarily the same as fuel economy; for example, one vehicle might have better technology and be more fuel efficient than another, but if it is larger and heavier than the other vehicle, it might still have worse fuel economy.

fuel intensity Similar to fuel economy, but where the term often describes the average fuel use per kilometer of the vehicle stock rather than of an individual vehicle.

light-duty vehicles A general term to describe cars, vans, sport utility vehicles, and even small trucks commonly used for personal transport; in the United States, more than 50% of new vehicle sales to households are no longer cars but rather other types of light-duty vehicles.

"on road" or "in use" fuel economy The fuel economy or fuel intensity of vehicles in actual use; this is typically anywhere between 10 and 25% worse than the "rated" fuel economy of vehicles, measured on test equipment using a fixed-duty cycle.

Organization for Economic Cooperation and Development (OECD) Organization that serves primarily industrialized, more developed countries, including North America, Western Europe, Japan, Korea, Australia, and New Zealand; OECD countries have become synonymous with the developed world, and non-OECD countries have become synonymous with the developing world.

transportation The movement of people and goods, providing mobility and access to desired destinations.

"well-to-wheels" greenhouse gas (or CO_2) emissions Emissions that occur not just from the vehicle but during the entire process of producing fuel and bringing it to the vehicle; this term may also include emissions during vehicle production and disposal.

Like all human activities, transportation, the movement of people and goods, is made possible by the use of energy. Current transport systems around the world rely nearly completely on a single form of energy—petroleum. As demand for petroleum fuels continues to increase, the unsustainability of this reliance has become an increasing concern. Not only is petroleum nonrenewable, but its use produces side effects such as air pollution and greenhouse gas emissions that could contribute to climate change. However, there is a range of available technologies and other options for reducing petroleum consumption in transport, mostly involving either switching to new fuels or cutting fuel use by improving transport efficiency. Significant changes in travel patterns are also possible. This article reviews current trends and outlines the options available for changing their direction so as to put transport on a more sustainable path.

1. RECENT TRENDS AND A FORECAST FOR THE NEXT 30 YEARS

In nearly every region of the world, transport energy use is increasing rapidly (Fig. 1) and is fueled nearly entirely by petroleum—more than 97% of transport fuel use worldwide. Because emissions of carbon dioxide (CO_2) are closely linked with consumption of petroleum (and other fossil fuels), these emissions have risen—and will continue to rise—in lock-step with the increases in petroleum use. Gasoline and diesel vehicles emit a variety of air pollutants, although

 © 2004 Elsevier Inc. All rights reserved.

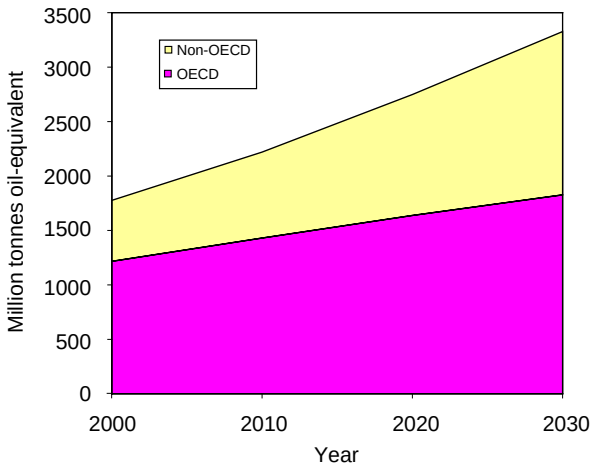


FIGURE 1 IEA forecast of transportation energy use through 2030: OECD and developing world.

tightened emissions control requirements have cut these emissions substantially during recent years, particularly in "developed" (i.e., Organization for Economic Cooperation and Development [OECD]) countries. Concerns regarding energy security, particularly "oil dependence," have become focused mainly on transport because most of the increase in global oil use over the past 30 years has been for the transport sector. In the developed world, oil use in other energy end use sectors, such as buildings and industry, has actually declined over this period.

Recent forecasts, such as the *World Energy Outlook 2002* of the International Energy Agency (IEA), suggest that most of the current trends are not likely to change over the next 30 years or so unless strong new policies and measures are adopted by governments. The IEA forecasts that between 2000 and 2030, global transport energy use (and associated CO_2 emissions) will nearly double (Fig. 1). Energy use will grow much more quickly in the developing (non-OECD) parts of the world, where it is expected to nearly triple; in OECD countries, the growth is expected to be "only" 50%. Even by 2010, global transport sector energy use and CO_2 emissions are forecast to increase by nearly 50% in the developing world, and only 15% in OECD countries, compared to 2000.

2. LOOKING BEHIND THE ENERGY TRENDS

Why is energy use continuing to increase in transport? It is helpful to consider the basic components of transport energy demand. These can be thought of in terms of the "ASIF" equation:

$$A * S_i * I_{ij} = F_{ij},$$

where A is total travel activity, often measured as person-kilometers of travel and freight ton-kilometers of travel; S_i relates to the structure of travel in terms of the types of travel modes i used and how they are used, often measured as vehicle-kilometers of travel for each mode per unit passenger-kilometer or freight ton-kilometer of travel; and I_{ij} is the energy intensity of mode i operating on fuel type j, often measured as energy use per vehicle-kilometer of travel. For any given mode i and fuel type j, this equation is an identity that determines F, that is, fuel use. Emissions of CO_2 can also be calculated with this equation simply by adding the appropriate coefficient of carbon per unit fuel use, generally fixed for any particular fuel type (although "well-to-wheel" emissions can vary considerably, as discussed later).

As this equation suggests, to affect transport energy use, either travel activity, structure, or energy intensity must be changed. This corresponds roughly to how much travel occurs, how it occurs, and how energy intensive it is. These represent the three main areas for potential policy intervention. It is also worth noting that the first of these represents largely a behavioral issue, the second is a combination of behavior and more technical issues, and the third—modal energy intensity—is largely (but not completely) a technical issue.

Travel activity is influenced by a number of factors. Most important is wealth; as incomes rise, so does travel activity, at least in terms of the average distance each person travels and the types of modes he or she chooses. The amount of time spent traveling is remarkably consistent across incomes as well as across regions, cultures, and even history (humans have been traveling roughly an hour each day, on average, throughout much of their history). But increasing wealth, along with improved technology, enables access to faster modes of travel, allowing humans to cover greater distances with the hour or so they spend traveling each day. One of the principal challenges in slowing—and perhaps eventually reversing—travel growth is encouraging people in an increasingly wealthy world to reorganize their lives so as to be closer to their destinations of choice (e.g., home, work, shopping, recreation) despite their increasing ability and apparent inclination to arrange these activities farther and farther apart. Increasing the cost of

travel is one way in which to do this; but stronger rules governing the way in which cities and regions are spatially developed is perhaps the best long-term solution.

The potential for travel growth around the world is staggering. This becomes clear when observing the current differences in travel per capita around the world, both in terms of amount and in terms of modes used (Fig. 2). Most of the world still travels far less than people do in OECD countries. But as incomes increase and people gain access to faster modes of transport, such as motorcycles and cars, their travel will increase. If the entire world traveled as much, and in the same modes, as North Americans do now, total world travel would increase by more than fivefold, and fuel use would increase by more than that (because travel would shift to more energy-intensive modes).

Travel structure, or the manner in which people travel in terms of the modes they use and the extent to which they travel together or alone, is also largely a behavioral issue. Automobiles have become the dominant mode of travel, not because they are technologically superior to buses or trains but rather because they provide a travel structure that most people prefer—flexibility in when and where travel occurs, comfort, convenience, and perhaps a benefit in terms of social status. Whether the strong trends toward car ownership and travel around the world can be reversed is unclear, but such a reversal will certainly require strong disincentives to car use, such as high fuel taxes, and significant improvements in other alternatives, such as bus and rail transit, to attract travelers back to these modes.

The energy intensity of vehicles and modes is a function of several factors. These include the technical efficiency of propulsion systems, vehicle characteristics such as weight and aerodynamic design, and the manner in which vehicles are used and maintained. For light-duty passenger vehicles (e.g., cars, minivans, sport utility vehicles [SUVs]), the actual "in use" energy intensity of vehicles can be up to 25% higher, on average, than their rated efficiency under test conditions. Human behavior enters this picture not only in terms of the manner in which people drive (which can have a significant impact on efficiency) but also in terms of the types of vehicles they choose to buy. Beyond the obvious differences between larger and smaller vehicles, even vehicles of a similar size can have quite different energy efficiencies due to differences in the factors mentioned.

There are strong differences in the energy intensities of different modes (Fig. 3). Until recently, aircraft were the most energy intensive per passenger-kilometer, but they have improved dramatically over the past 30 years or so due to strong, steady efficiency improvements in aircraft and aircraft engines as well as a steady trend toward larger and fuller aircraft that can carry many more passengers per unit fuel use. Today, light trucks (i.e., light-duty vehicles such as vans and SUVs but not cars) are the most energy-intensive vehicles. Passenger rail is by far the most efficient mode per passenger-kilometer. For freight, trucks are far more energy intensive than other modes; rail and waterborne transport are very energy efficient. Clearly, shifting some travel to more efficient modes could reduce overall transport energy use considerably.

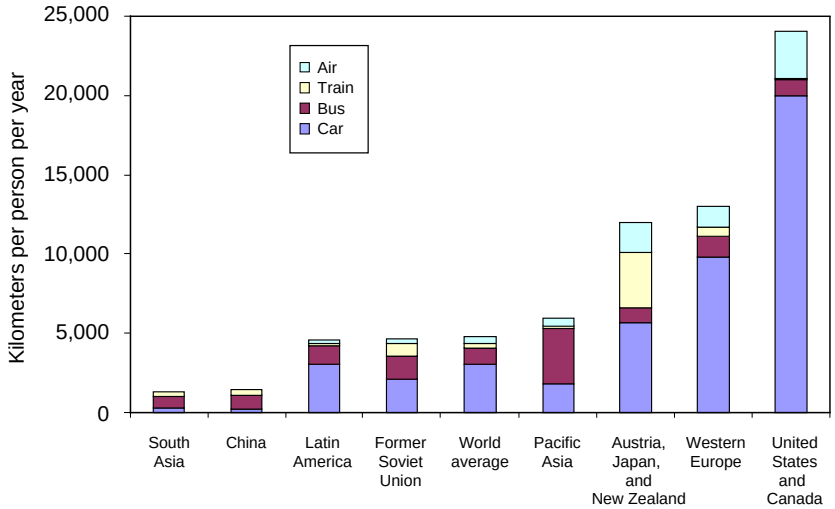


FIGURE 2 Travel by mode and region, 2000. Reprinted from World Business Council for Sustainable Development (2001).

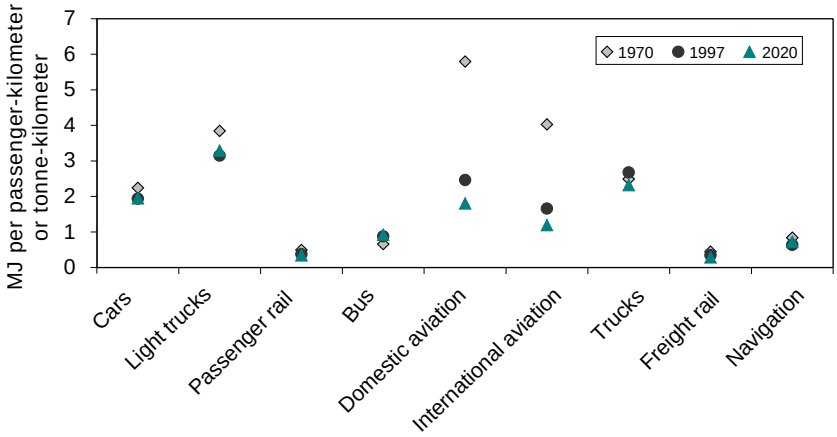


FIGURE 3 Energy intensities of various modes over time (OECD averages). From International Energy Agency (2002b).

What are the current trends regarding each of these factors? The following highlights the key trends identified in the IEA's forecast:

- In OECD countries, travel activity growth is slowing somewhat—most in OECD Pacific, less in OECD Europe, least in North America, and less for freight than for passenger travel. Europe's road freight travel is still growing significantly. Travel activity in most non-OECD regions is expected to maintain recent high growth rates for many years to come.
- Substantial fuel economy improvements will likely occur only in the European Union (EU) and Japan. These are the only world regions with aggressive policies to improve fuel economy. In other regions, improvements in vehicle technology will occur but are unlikely to translate into better new vehicle fuel economy and lower fuel intensity of the vehicle stock as consumers continue to opt for larger and faster vehicles rather than for more fuel-efficient ones. Even in Europe and Japan, improvements in fuel economy probably will not be enough to offset the expected growth in travel activity.
- In freight, rapid increases in truck travel will continue to raise average fuel use per ton-kilometer. Road freight and, increasingly, aviation freight are expected to gain shares in most regions. Small commercial trucks and vans are among the fastest growing vehicles and the most energy intensive.
- Growth in air travel is expected to account for nearly a quarter of the projected increase in fuel demand to 2030. Despite strong past improvements in fuel efficiency of aircraft, future reductions in fuel intensity are expected to offset only one-quarter of increases from expanding activity.
- Alternative fuels will continue to play a small role. Although significant increases may occur in the use of biofuels (e.g., ethanol) and, in a few places, in the use of gaseous fuels, the net effect is likely to be small without strong new policy interventions. The oil share of energy in transport is expected to remain at approximately 97%.
- CO_2 emissions from transportation will continue to grow significantly—by about the same rate as fuel demand. Globally, these will nearly double from 2000 to 2030.

Thus, the overall picture provided by the IEA forecast is one of continued strong growth in both oil consumption and CO_2 emissions from transport. The key drivers are travel demand growth and the shift from mass transit modes to personal vehicles.

3. CONCERNS FOR TRANSPORT ENERGY SUSTAINABILITY

One of the biggest long-term concerns for transport is its near total dependence on petroleum fuels. As petroleum supplies dwindle and become increasingly concentrated in the Middle East, the risk of supply disruptions and price shocks will likely increase. Management of these risks, such as through diversification of energy sources, represents a critical aspect of long-term transport energy sustainability.

A second major risk to long-term transport sustainability is its impact on climate change. CO_2 and other greenhouse gases (GHGs), including CO, N_2O, CH_4, and certain fluorocarbons, are emitted from all transportation modes and all types of vehicles to varying degrees. For alternative fuels,

such as natural gas, propane, and even renewable fuels (e.g., the heavily processed corn-derived ethanol in the United States), a well-to-wheels analysis of GHG emissions, including feedstock production, conversion, fuel distribution, and vehicle refueling, must be made to ascertain the full emissions impacts of these fuels, not just emissions occurring during driving. Recent studies show that most alternatives to gasoline and diesel have higher energy losses and GHG emissions during fuel production and transport to refueling stations. But several have lower vehicle emissions, and a few (e.g., pure hydrogen) produce no vehicle GHG emissions. For these fuels, the challenge is to produce and transport them with low GHG emissions, and this might not be simple or cheap. As discussed later, few alternatives appear to provide dramatic overall reductions. Those that do include hydrogen or electricity derived from renewable (or nuclear) sources and some types of biofuels such as biodiesel from oil crops and ethanol from cellulosic biomass.

Transport is currently running a "collision course" with a world increasingly convinced of the need to take major steps to dramatically reduce its CO_2 emissions over the next 10 years and beyond. The Kyoto accord, signed in 1997, targets a reduction in overall CO_2 emissions from annex I countries (mostly the developed countries, although not all have ratified the treaty) so that by 2010 they are returned to their previous 1990 levels on average. It may be possible to meet the Kyoto commitments without outright reductions in transport CO_2 emissions over the next 10 years or so (by more than offsetting transport increases with reductions in other sectors). But a major risk for transport is that the longer countries wait before beginning a shift toward significant decarbonization of this sector, the more difficult this shift may be. On the other hand, by waiting, it may be possible to develop and implement less expensive technologies to achieve sustainable levels of GHG emissions (which will eventually need to be far below 1990 levels).

Not all choices for our future transport systems will affect all aspects of sustainability equally or even in the same direction. Although the dominant (direct) environmental impact of energy use in transport is CO_2 emissions, the types and amount of energy used also affect emissions of other pollutants such as NO_x and particulate matter. The decisions that are made to reduce CO_2 emissions could, in some instances, increase the emissions of some other pollutants. Similarly, although energy use does not directly affect transport's impacts on the natural environment (e.g., ecosystems), humankind's willingness to allocate ever greater levels of energy to the transport system serves as an enabling force for an ever expanding transport system, encouraging sprawling urban development. This pattern might not itself be sustainable.

Another important determinant of global transport sustainability is the manner in which transport growth is handled in developing countries. Although it is only fair that all countries should eventually achieve the same level of mobility currently enjoyed by the wealthiest countries, the implications of a world with average travel and energy use levels as OECD countries not only is probably unsustainable but might also be very difficult to achieve, at least with the current mix of fuels. However, changing the direction of transport growth in developing countries may in some ways be easier than doing so in OECD countries because there is typically less of an existing entrenched transport system that must be changed. On the other hand, the amount of investment and policy effort that must be devoted to directing transport growth toward sustainable pathways in developing countries, relative to what is occurring today, may represent an even bigger shift than what will be required in IEA countries.

Transport's central economic role and its strong influence on daily life have made rapid changes in this sector difficult to achieve. Its relatively weak responsiveness to energy price movements and the slow turnover of its infrastructure render it both crucial and difficult for governments to implement strong measures to improve sustainability. In particular, twin challenges of reducing oil dependence and emissions of GHGs suggest that there are several key types of action needed:

- Improving vehicle and system energy efficiency
- Increasing the use of renewable and other low-carbon energy sources
- Expanding access around the world to clean and efficient transport energy technologies and developing and deploying advanced technologies more rapidly
- Developing politically acceptable, but aggressive, policies to shift the direction of transport toward a more sustainable path

4. IS AN ALTERNATIVE FUTURE POSSIBLE?

Having outlined one possible scenario for achieving a sustainable pathway in transport, the question arises

as to whether such a scenario can be realized. The answer depends on the following:

- What policy options are available to bend the curves of transport energy use and CO_2 emissions?
- What can specific policies accomplish in terms of oil and CO_2 reductions?
- What is the feasibility of implementing a package of policies strong enough to put transport on a new pathway?

Most countries that have tackled energy use and CO_2 emissions from transport have linked these issues directly to comprehensive transport policy reform, with an emphasis on improving transport efficiency and "getting the prices right." Other measures have focused on reducing the growth rate (or outright levels) of personal vehicle travel and improving traffic conditions. These policies typically are driven more by transport issues—congestion, safety, noise, and nuisance concerns—than by energy or CO_2 reduction concerns. It is primarily in the area of vehicle fuel efficiency and promotion of alternative fuels that transport energy policy has been the dominant driver.

Probably the most important technology-oriented policy measure now in place is the voluntary agreement for emissions reductions in new cars between the EU and the European Association of Car Manufacturers (ACEA). This EU–ACEA agreement will account for the bulk of reductions from baseline that are expected for the European countries that were studied. If successful, this measure alone could reduce emissions from cars by 15 to 20% below trend by 2010, with even greater reductions after 2010 when the more efficient new cars will have fully penetrated the vehicle stock.

Some measures to encourage modal shifts and nonmotorized transport modes, as well as measures to encourage less travel, are part of all the European programs, but it is difficult to judge what savings they have realized or will realize on their own. Improved transport services are expected from the privatization of railways, but the ultimate effect is unclear. Land use planning forms will complement transport policy reforms, but it is also difficult to predict how much this will reduce travel or induce modal shifting. In the European countries studied, infrastructure expansion and investment are increasingly tilting toward nonroad or intermodal facilities. Such efforts should at least slow the erosion of the share of low-CO_2 modes.

The following subsections provide a brief outline of potential measures for reducing travel, improving vehicle efficiency, encouraging fuel switching, and improving freight system efficiency. A listing (by no means exhaustive) of potential measures is provided in Table I.

TABLE I

Potential Measures to Reduce Transport Energy Use

Passenger travel: Measures to reduce private vehicle travel
Improving transit systems
Fuel and vehicle taxes
Travel pricing mechanisms (e.g., electronic road pricing)
Parking-related measures
Promotion of nonmotorized modes (e.g., biking, walking)
Stronger land use planning
Promoting telematics and telecommuting (substitution of communications for travel)
Measures to improve light-duty vehicle test and on-road efficiency
New vehicle fuel economy improvement
Improved vehicle maintenance
Vehicle on-board fuel use information and driver training
Speed limit reductions and increased enforcement
Roadway traffic flow improvement and capacity expansion
Vehicle scrappage programs
Alternative fuels
Incentives for marketing and purchase of advanced vehicle technologies
Incentives for and direct investment in refueling infrastructure
Removal of market barriers such as restrictions on use of specific vehicle types
Highway and surface freight movement
Technical efficiency improvement and alternative fuels in new trucks
Trucking system and operational efficiency improvements
Mode switching (e.g., from truck to rail and water)
Changes in freight demand patterns through spatial adjustments

4.1 Reductions in Growth of Motor Vehicle Travel

Measures that aim to restrain passenger travel in light-duty vehicles (among the most energy-intensive modes) and shift it to more efficient modes have begun to play an increasingly important role in the policy plans of many OECD governments. Other sustainability concerns, such as traffic congestion, urban sprawl, and "quality of life," often provide the key motivations to implement such measures. However, one problem with demand restraint policies for national governments is that many of these policies are implemented on regional or local levels.

Demand-oriented policies can encompass pricing of different transport activities and modes, investments in mass transport infrastructure and rolling stock, regulation and restriction of traffic, and urban and land use planning.

Apart from imposing extremely high (and continuously increasing) levels of fuel taxation, which in many countries may be one of the least politically feasible of available strategies, reigning in travel probably can be accomplished only through a coordinated, but ambitious, package of measures. This package must both discourage personal vehicle travel and encourage use of alternative modes and alternatives to travel altogether. Strong packages should include several of the following elements:

- Widespread electronic roadway pricing, at least throughout all limited-access highway systems. Cordon pricing in urban areas, such as has recently been implemented in London, may also be an important element.
- Increased parking costs and better enforcement in many areas.
- Much stricter land use controls. These could potentially include growth boundaries, improvements in pedestrian and nonmotorized travel (e.g., bicycle) infrastructure, restrictions on parking supply in new construction, strong incentives for mixed-use multimodal developments, and other elements that can help to end the current practice of building car-dependent communities.
- Incentives to speed the rate of adoption of telecommuting, especially 1- or 2-day-a-week home-based telecommuting. The latter may have relatively small secondary impacts on relocation and generation of increases in other types of travel. These incentives could include tax breaks to businesses and individuals for telecommuting equipment and even straight tax breaks to businesses per employee who telecommutes.
- Carefully coordinated improvements to public transit that include measures to increase average ridership and "load factors" for existing systems. New concepts in transit, embodied in the term "bus rapid transit" (BRT), are emerging. BRT typically includes features such as conversion of traffic lanes into dedicated bus lanes (and enforcing this measure), better bus stops or even bus "stations" with prepayment of fares and rapid boarding, and bus priority treatment at traffic signals. These features can increase transit capacity, improve service (faster travel times), and discourage personal vehicle travel in the same corridors, potentially yielding high load factors and significant reductions in overall system CO_2 emissions. However, simply expanding transit service by adding more routes may lead to situations with more nearly empty busses with relatively high CO_2 emissions per passenger-kilometer of travel, and net increases in CO_2.
- A strong education/publicity campaign to encourage society to collectively take actions to move toward a much less car-oriented society. The seeds of such efforts are now being planted in many places, as reflected by "car-free" days and the like.

All of these measures are already well known and often considered by planners and have been implemented in various forms in many cities around the world. But few places have taken an aggressive strategy of coordinating these measures and implementing them at levels of policy intensity sufficient to come anywhere near ending the growth in personal vehicle travel. Even in cities with good transit service and good facilities for nonmotorized travel, such as in The Netherlands, significant problems remain with automobile traffic congestion and air pollution. A relatively laissez-faire attitude about the car and levels of personal vehicle travel is still evident in nearly every city in the world.

4.2 Improving Technical Efficiency of Vehicles

Currently available technology offers a large potential for improving vehicle efficiency at a relatively low cost. But in many countries, fuel prices have not given a strong signal to consumers to demand vehicles with higher fuel economy or to producers to provide vehicles with better fuel efficiency. Trends in car size and power are moving in the opposite direction; as consumers around the world become wealthier, they tend to choose ever larger, more powerful vehicles at the expense of fuel efficiency. The net effect is fairly flat, or even deteriorating, fuel economy in most regions of the world.

Recent estimates of the penetration of various technologies in new vehicles in three OECD countries suggest that many technologies have only just begun to be used (Fig. 4). Even without using advanced technologies (e.g., fuel cells), there is still a large scope for technical improvements to vehicles. Many of these technologies would pay for themselves in fuel savings. Most of these will eventually be added to most vehicles, but at a pace that is much slower than possible with an aggressive fuel economy policy.

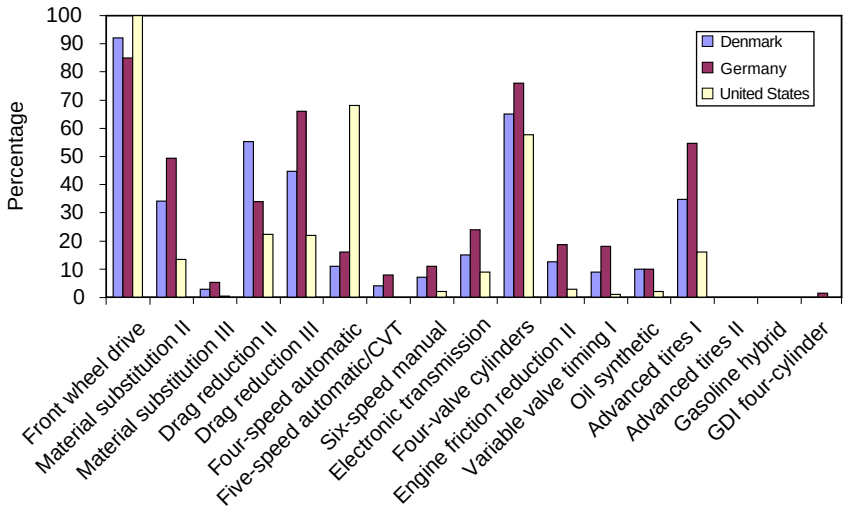


FIGURE 4 Market penetration of selected technologies in compact cars: Denmark, Germany, and the United States, 1998. Reprinted from International Energy Agency (2001).

Various recent studies estimate that there is considerable potential for cost-effective improvements in light-duty vehicle fuel economy during the next 10 years and beyond. Using just conventional technologies such as those shown in Fig. 4, studies estimate anywhere from a 20 to 50% potential reduction, albeit at varying costs. A recent IEA study estimated that low-cost reductions in new car fuel use per kilometer on the order of 25% are possible by 2010 compared with the late 1990s. This estimate is similar to the goals of the EU manufacturer association (ACEA) voluntary agreement to improve fuel economy. Increasing sales of diesel cars and avoiding purchase shifts to larger, heavier cars could contribute to these targets. After 2010, introduction of some advanced technology vehicles with radically improved engine and drivetrain efficiency (e.g., advanced hybrid electric and fuel cell vehicles) could provide significant additional fuel economy improvements.

Figure 5 shows that although implementing a full set of technologies could result in much lower fuel consumption per kilometer from new cars in the future, much of this improvement is likely to be lost to shifts in consumer purchases to larger heavier vehicles. The figure, which shows the U.S. scenario, focuses on just one size class of vehicles: compact cars. Further erosion in fuel economy improvements is likely to come from purchase shifts from this car class to larger heavier classes such as large cars, minivans, and SUVs.

Thus, although the availability of cost-effective technologies holds promise, these technologies might not be deployed to save energy and CO_2 without strong market signals in that direction. An IEA examination of automobile advertising in virtually every OECD country in 1997 confirms that CO_2 reduction is not a major marketing theme.

The existing EU voluntary agreement and the Japanese "Top Runner" fuel economy program are policies already in place that appear likely to help achieve much of the available cost-effective potential by 2010. In the EU agreement, automakers have committed to a 25% reduction in the CO_2 emissions of new vehicles between 1995 and 2008. As of 2001, they were on track to meet their target. The Japanese program identifies the most fuel-economic vehicles in each vehicle weight class and requires all new vehicles to approach this level over time. In North America, there is not currently any policy that appears to be pushing the market toward taking advantage of these new technologies given that the U.S. corporate average fuel economy (CAFÉ) standards have not changed significantly in nearly 20 years. Without additional policy efforts, much of the available technology may be used to increase vehicle size, weight, and horsepower while holding fuel economy constant, as has occurred over the past 10 years or so.

There are several types of policies that could be implemented to help maximize the fuel economy benefit of existing technology. In addition to voluntary agreements and "top runner"-style policies, implementing fuel economy-based fees and rebates (also called "feebates") and market incentives for the adoption of advanced technologies could play useful roles in North America and elsewhere.

To achieve early market penetration of advanced technologies, such as hybrid electric and fuel cell

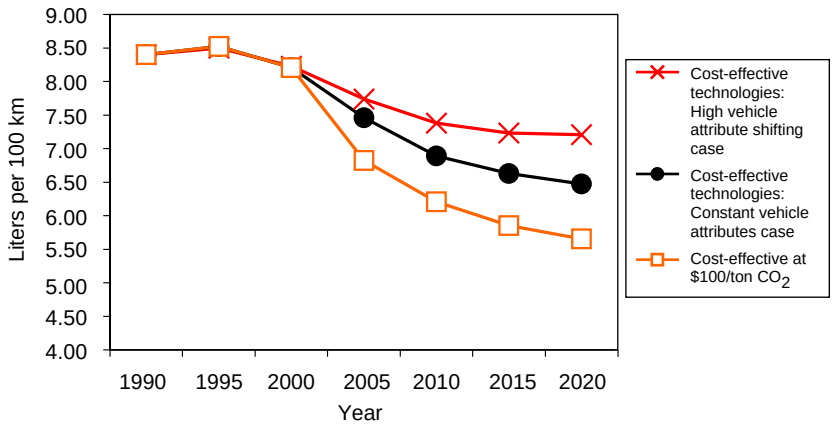


FIGURE 5 Cost-effective technologies implemented in cars over time in the United States: Three scenarios. Reprinted from International Energy Agency (2001).

propulsion systems, governments might need to provide relatively large rebates on these vehicles so as to make them attractive to consumers (and attractive to producers to build). As an alternative to a broad feebate program (which could itself provide significant rebates for very fuel-efficient cars), more targeted incentives could be used to overcome the very high initial capital costs and risks of producing such vehicles and to encourage consumers to purchase vehicles containing such technologies. Vehicle purchase incentives could be implemented in the form of price subsidies, either for vehicles possessing specific technologies or simply for vehicles meeting a certain standard for fuel economy or CO_2 emissions per mile. The latter approach has the advantage of encouraging the development of new technologies to meet societal goals. Such subsidies could also be designed to target both reductions in CO_2 and reductions in pollutant emissions.

An example of "next generation" technology price incentives is found in Japan, where there is a rebate of U.S. $3500 per vehicle for purchases of hybrid electric vehicles. Japan is the first country with significant sales of these advanced technology vehicles. A time-phased approach, which includes large initial subsidies per vehicle that decline through time as vehicle sales increase (and as vehicle costs decline), may produce a much more rapid buildup in production volumes and the rate of learning than a scenario where expensive vehicle technologies must compete on their own during the early phases.

4.3 Shifting to Alternative Fuels

Alternative fuels continue to hold promise, but for a variety of reasons none has yet significantly penetrated the vehicle market around the world. There are several fuels and propulsion systems that offer the possibility of displacing oil and lowering vehicle CO_2 emissions, including natural gas, liquefied petroleum gas (LPG), alcohol, and biodiesel for internal combustion engine vehicles; electricity for battery-powered vehicles; and hydrogen (produced from any number of different fuels) for fuel cell-powered vehicles. All of these options are making only slow progress in the marketplace due to factors such as higher costs of the vehicles and/or fuel, lack of available vehicle models, and lack of widespread retail fuel availability. Most fuels and accompanying vehicle technologies also have one or more other drawbacks such as lower driving range, longer refueling times, smaller trunk space, and poor acceleration performance. As a result, consumers have so far been reluctant to embrace any of these options. Without strong regulations or the creation of significant price advantages, alternative fuels are making little headway.

Although all alternative fuels offer an opportunity to displace petroleum fuel, many offer only relatively small CO_2 reductions on a life cycle or well-to-wheels basis (Fig. 6). Only four options clearly can provide greater than a 50% reduction in CO_2-equivalent emissions per kilometer of driving: alcohol derived from cellulosic biomass (and using biomass as the process fuel), biodiesel from oil crops, hydrogen produced from a low-GHG source (e.g., renewably powered electricity) and used to run fuel cell vehicles, and low-GHG electricity itself when used with electric vehicles. These options all are relatively expensive, and the latter two would require large investments in new types of vehicles and refueling infrastructure.

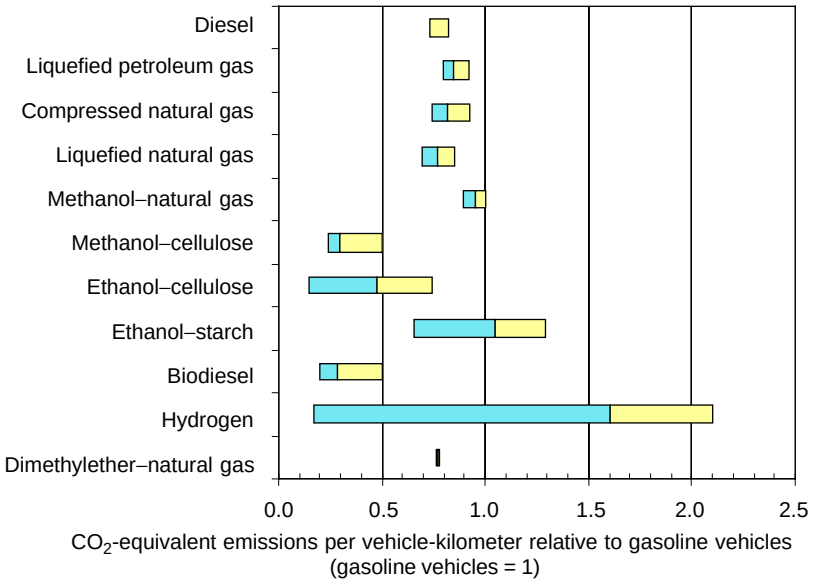


FIGURE 6 Alternative fuels: Near-term CO_2-equivalent reduction potential compared with conventional gasoline vehicles (well-to-wheel). The shaded and hatched areas represent bands of uncertainty around the middle vertical line on each bar, with the middle line being considered the "reference" case. Reprinted from International Energy Agency (2001).

As a result, few OECD countries consider alternative fuels as an important near-term option for mitigating energy security concerns and GHG emissions. Because the best alternatives would be expensive and would require substantial (and risky) investment, most countries are taking a longer view of these options—focusing on developing niche markets, funding research and development (R&D), and taking other preparatory actions.

Although not strictly an alternative fuel—or at least not a nonpetroleum fuel—diesel can provide significant potential for lower net fuel intensity and CO_2 emissions and does as well as many alternative fuels such as compressed natural gas (CNG) and LPG. New light- and medium-sized diesel cars in Europe with turbo-direct injection (TDI) offer up to 25% lower fuel consumption than do similar gasoline vehicles. On the other hand, diesel fuel is taxed much more lightly than gasoline in some countries; not surprisingly, its consumption is associated with significantly higher car travel. Diesel TDI is also appearing on heavier American-style vans in Europe, bringing fuel-driving costs for such larger vehicles down to more affordable levels. Thus, the net impact of Europe's trend toward "dieselization" is difficult to determine after its "behavioral" impacts are taken into account.

Over the longer term, perhaps the most promising option is the fuel cell vehicle running directly on hydrogen fuel. With renewably or nuclear power-derived hydrogen, this technology could provide near-zero life cycle CO_2 emissions. However, during the 2000–2010 time frame, fuel cells are not likely to be produced in large numbers due to ongoing technical hurdles, their current very high cost, and lack of hydrogen refueling infrastructure. The refueling infrastructure problem could be alleviated by using hydrogen-rich fuels (e.g., methanol, gasoline) and reforming these fuels on-board fuel cell vehicles to produce hydrogen, but this option appears to be falling out of favor with most car companies. The complexity and cost of on-board reforming appear to offset the benefits it may provide. Furthermore, vehicles with on-board reforming will not be zero emissions for either pollutants or GHGs (because reforming fuels on-board produces some emissions), significantly reducing the benefits of switching to fuel cells.

Whether or not electric vehicles ever reach a commercial stage of production (prevented so far primarily due to shortcomings in battery technology), electric drive systems (including electric transmission systems and even motors) are likely to become a feature on all vehicles in the future. Although hybrid electric vehicles now being sold in OECD countries run on gasoline, with the hybrid system used to boost fuel economy, these vehicles introduce motors and larger battery systems into vehicle propulsion design and help to pave the way for future electric or fuel cell systems. If battery

technology improves, hybrid vehicles could take on larger battery storage systems and eventually be "pluggable" (rechargeable), offering a bridge to electric vehicles.

Electric- or hydrogen-powered propulsion systems may offer attractive ways in which to eliminate internal combustion, and the emissions it produces, from vehicles, but they may simply move this process to central power plants (which produce the fuel). This would help to move pollutant emissions away from cities, but unless electricity or hydrogen is produced in a manner with low carbon emissions, the net GHG benefits may be small. The most obvious options for low-carbon production of electricity and hydrogen are renewable fuels and nuclear power. These currently account for less than 20% of electricity generation around the world.

How could a low-carbon hydrogen or electric vehicle scenario come about? One transition path is for governments to strongly encourage development and commercialization of fuel cell (or electric) vehicles, especially demonstration programs and small-scale production to speed learning. Concurrent with (or slightly preceding) the beginnings of significant vehicle sales in targeted areas, a strong push toward developing hydrogen (or electric) refueling infrastructure will be needed. Policies such as tax credits for installation of refueling infrastructure could be applied, or governments could invest directly in infrastructure. The level of investment needed over 30 years to pay for refueling infrastructure and other support systems (e.g., garages that are trained to repair the vehicles) and the incremental costs of fuel cell vehicles over this time both are likely to be quite large. Society must be prepared to help pay these costs if it is to have a reasonable chance to achieve the desired transition. Willingness to pay these costs will largely reflect the level of determination of the public to move transport onto a more sustainable course. Of course, ongoing R&D programs can help to reduce costs before a transition begins, especially in the areas of fuel cell system production cost and on-board hydrogen storage. But chances are that the initial costs of both vehicles and refueling infrastructure will be high.

Biofuels, such as ethanol from crops, holds the potential for substantial oil displacement and GHG reduction, although current production practices fall far short of this potential. An important advantage of biofuels over most other alternative fuels is that they can be blended with conventional fuels and used, to a point, in current vehicles. In the United States and the EU, a small share of transport fuel is blended up to 10% with ethanol fermented from the sugars in grain crops (e.g., corn, wheat). The processes used require substantial fossil fuels to drive them, and the net displacement of petroleum fuel is less than 30% in most cases (i.e., it takes 0.7 unit of fossil fuel to produce 1 unit of ethanol). As a result, GHG reductions are also generally no better than 30% per liter of gasoline displaced by ethanol.

However, considerable research is under way to develop new conversion processes that would allow ethanol to be produced from "cellulosic biomass," specifically the green portions of plants made of cellulose. Using advanced enzymes and processes for conversion of cellulose to alcohol, and using cellulose as the process fuel (rather than natural gas or another fossil fuel), cellulosic ethanol could bring life cycle GHG emission reductions of 80% or more compared with fossil fuels on a well-to-wheel basis. This approach is still in the R&D phase of development, and although it appears to have strong cost reduction potential, it is unclear when it will become cost-effective or even more cost-effective than current grain-to-ethanol processes (which themselves cost up to twice the cost of producing conventional gasoline).

Biodiesel fuel, typically a methyl ester produced from oil crops such as rapeseed, can be blended with conventional diesel fuel in any proportion and used in today's diesel vehicles. Biodiesel has been estimated to provide anywhere from 30 to 70% reductions in well-to-wheel GHG emissions, but it is very costly. Biodiesel in countries such as Germany and France, where it is publicly available, costs up to three times the cost of conventional diesel fuel.

Even if substantial cost reductions in biofuel production costs are achieved, a separate question is whether biomass resources are best used for transportation fuels (as opposed to using them, e.g., to produce electricity). Liquid fuel production requires substantial energy input per output unit of useful energy, and biomass feedstock might have more cost-effective use elsewhere.

4.4 The Role of Fuel Taxation

Fuel taxation plays an important role in all of the policy areas covered here. Fuel price at the pump provides an incentive to consumers to drive less, to choose more efficient cars, and (potentially) to choose cars running on alternative fuels. Increases in taxation can help to limit the "rebound" effects of other policies that reduce fuel costs, for example, by making vehicles more efficient. Differentiated fuel taxes can

be applied to support or limit the use of specific fuels. Fuel tax increases help to restrain growth in transport activity and have ramifications for the relative competitive advantages of various modes.

Yet one should not overstate the role of taxation, especially as an isolated measure, or the potential to increase its role in the future. The response to fuel tax increases tends to be very limited in the short term, although it is somewhat greater in the long term. Many studies across the OECD have estimated that a 10% increase in fuel taxes will reduce travel by only 1 to 3%. Furthermore, in many countries, fuel taxes appear to already be at levels that are not politically popular, and significant additional increases in the near future may be problematic. It may be that more selective pricing mechanisms, such as congestion charges and other road user charges, may be more viable.

4.5 Increasing the Efficiency of Freight Movement

There are four ways in which freight travel (at least truck travel), energy use, and GHG emissions could be dramatically reduced:

- Increasing investment in rail and waterborne transport modes and increasing investments in intermodal facilities around the world to allow as much use of these modes as possible, reserving truck transport for relatively short-distance trips. Pricing incentives can also play an important role in pushing such shifts to the maximum extent feasible.
- Fuel shifting in medium and heavy trucks. Just as hydrogen fuel cell propulsion systems is probably the closest thing to a "silver bullet" for light-duty vehicles, this is also true for most types of heavy-duty vehicles. Fuel cell engines are being developed for many types of heavy-duty vehicles, including trucks, buses, and trains, and applications in some market niches (e.g., urban transit buses) may become cost-effective before light-duty vehicle applications. On the other hand, diesel engines are very cost-effective and durable and are very popular with truck operators.
- Incentives and government infrastructure support to rapidly develop much more sophisticated systems and linkages between suppliers and their clients, and with each other, to significantly increase the efficiency of goods delivery in metropolitan areas. Truck load factors, the numbers of "back haul" trips where trucks run empty, and the sizes of trucks for various applications all could be improved significantly, resulting in far fewer kilometers of truck travel and much less fuel use per ton-kilometer of goods movement. Systems are now being installed in many places to improve freight logistics, but the relative fragmentation of the truck industry impedes the types of cooperation needed to maximize the potential efficiency gains. Governments could take a much more active role to change this in the future.
- Rationalization of locations of goods producers, distributors, and end users. There is some evidence to suggest that many "upstream" (nonretail) activities could be moved much closer to other steps in the chain of taking a product through various stages of production and, in some cases, delivery to the end user. However, this is a particularly difficult problem to address because shipping costs are typically such a small part of location decisions that they are not likely, even with much higher taxes placed on transport costs, to drive significant changes in many location decisions. But there may be other ways in which to assist companies in optimizing location decisions that should be pursued aggressively.

5. CONCLUSIONS

This article has sketched out some of the key issues regarding transport energy sustainability, outlined where transport energy use and GHG emissions appear to be heading over the next 30 years, and identified a variety of potential technologies and policy areas that could be used to move transport away from its current trends. The current path is far from any reasonable characterization of a sustainable path.

Among the first steps for moving in a new direction will be for governments to convince citizens of the need for aggressive action and to educate people about the risks associated with the current path and how important a role transport must play in reducing oil dependence and CO_2 emissions. Without strong public support and understanding, there will be little political will to enact policies. Under such conditions, even low-cost actions that provide clear societal benefits beyond fuel savings and CO_2 reductions can be blocked by individual stakeholders who may stand to lose. Governments also need to overcome "traditions" that no longer make sense such as company car policies, very high fixed (but low variable) taxation, and traditions of generally cheap fuel (in a few countries). Governments need to become more entrepreneurial. The

presence of good transit or walking/biking facilities does not necessarily translate into high use of such facilities; the public has to be convinced that the benefits that these modes offer—lower traffic congestion, cleaner air, and lower CO_2 emissions—are worthwhile and that they will be attainable only if people are willing to make changes to their lifestyles.

Governments would be well served to identify their own points of sustainability and sketch out sustainable pathways, such as the ones presented here, to help develop a plan of action that has any hope of yielding the desired path. Sharing such scenarios with the public—indeed, involving the public in a meaningful way and investing in widespread dissemination of such efforts—may also be a critical element in generating a social consensus on what pathway the public most supports, taking into account both the benefits of achieving such a pathway and the costs of getting there.

SEE ALSO THE FOLLOWING ARTICLES

City Planning and Energy Use • Equity and Distribution in Energy Policy • European Union Energy Policy • Fuel Economy Initiatives: International Comparisons • Global Energy Use: Status and Trends • Lifestyles and Energy • National Energy Policy: United States • Passenger Demand for Travel and Energy Use • Transportation and Energy, Overview • Transportation Fuel Alternatives for Highway Vehicles • Vehicles and Their Powerplants: Energy Use and Efficiency

Further Reading

American Council for an Energy Efficient Economy. (1998). "Transportation, Energy, and Environment: How Far Can Technology Take Us?" (J. M. DeCicco and M. Delucchi, Eds.). ACEEE, Washington, DC.

Greene, D. L., and DeCicco, J. (2000). Engineering–economic analyses of automotive fuel economy potential in the United States. *Annu. Rev. Energy Environ.* **25**, 477–536.

Greene, D., and Wegener, M. (1997). Sustainable transport. *J. Transport Geogr.* **5**, 177–190.

International Energy Agency. (2001). "Saving Oil and Reducing CO_2 Emissions from Transport: Options and Strategies." IEA, Paris.

International Energy Agency. (2002a). "Annual Energy Outlook 2002." IEA, Paris.

International Energy Agency. (2002b). "Transportation Projections in OECD Regions." IEA, Paris. [http://www.iea.org]

Schafer, A. (2000). Regularities in travel demand: An International Perspective. *J. Transport. Stat.* **3**(3), 1–31.

Schipper, L. J., Steiner, R., Figueroa, M. J., and Dolan, K. (1993). Fuel prices and economy: Factors affecting land travel. *Transport Policy* **1**(1), 6–20.

Sperling, D. (1994). "Future Drive: Electric Vehicles and Sustainable Transportation." Island Press, Washington, DC.

Sperling, D., and Delucchi, M. (1989). Transportation energy futures. *Annu. Rev. Energy* **16**.

World Business Council for Sustainable Development. (2001). "Mobility 2001." WBCSD, Geneva, Switzerland.

Transportation Fuel Alternatives for Highway Vehicles

DANILO J. SANTINI
Argonne National Laboratory
Argonne, Illinois, United States

1. **Blends versus Neat and Near-Neat Fuels: 1970s to Present**
2. **Fuel Properties and Fuel Switch Efforts**
3. **Fuel Properties and Centuries-Long Transportation Fuel Switches**
4. **Conclusion**

Glossary

biodiesel Soybean, rapeseed, and sunflower oils can be burned in compression ignition engines, but cause deposits and considerably higher maintenance requirements. Esterification with methanol or ethanol can largely eliminate these problems. The modified ester is called biodiesel. The more common form is a methyl ester. Soy-based biodiesel is used in the United States, while rapeseed based biodiesel is used in Europe.

carbon dioxide (CO_2) CO_2 is the leading greenhouse gas. Volumes of CO_2 produced as a result of combustion account for well over 99% of the volume of exhaust gases.

carbon monoxide (CO) Tailpipe pollutant that is a primary pollutant, which is worst in winter and at high altitude. Per unit of fuel burned, gasoline-fueled vehicles produce more CO than diesel-fueled vehicles.

catalytic converter Device used to convert hydrocarbons, carbon monoxide, or nitrogen oxides to innocuous chemicals such as carbon dioxide, water, oxygen, or elemental nitrogen. The need for unleaded fuel to allow catalytic converters to successfully reduce tailpipe emissions of gasoline-fueled motor vehicles is one of the major factors (aside from health damages of lead itself) causing removal of the highly effective octane enhancer tetraethyl lead from gasoline and the exclusive U.S. use of unleaded gasoline. Ethanol was one of the octane enhancers that got its start in part from this sequence of events (see E10). The next wave of improvements in fuel composition to enhance the performance of catalytic converters is reduction of fuel sulfur content. The World Fuels Charter provides recommended fuel specifications for areas of the world as a function of seriousness of need for emissions control. The greater the need for (or requirements for) emissions control, the lower the recommended lead and sulfur content of the fuel. Lead is scheduled to be completely removed.

cetane An indicator of the combustion quality of fuel for CI engines, such as diesel fuel. As cetane rises, the top speed (rpm) of compression ignition engines designed for the fuel rises. This, however, does not necessarily correspond to engine efficiency. Low-speed diesels are usually more efficient than high-speed diesels. The cetane rating of diesel fuels used for low-speed CI engines (100 to 500 rpm) is 25 to 35. For medium speed engines (300 to 1500 rpm), cetane can range from 35 to 45, while for high-speed diesels such as those found in heavy trucks and passenger cars, cetane ranges from 45 to 55.

Clean Air Act (CAA) and Clean Air Act Amendments (CAAA) The original CAA preceded the oil price shocks of the 1970s and CAFE. One of the sets of amendments occurred in 1990. The CAA and its amendments created the rules and regulations that are largely responsible for the removal of tetraethyl lead from gasoline and the introduction of alternative octane enhancers, including ethanol and, to a lesser extent, MTBE, for blending into gasoline. The unintended introduction of ethanol blends in the early 1980s followed the first CAA by many years and preceded AMFA and EPACT by about a decade.

compressed natural gas (CNG) CNG has high octane and is readily used as a fuel for spark-ignited (SI) engines, but only with diesel fuel ignition assist in compression ignition engines. Because NG displaces air (oxygen) in the cylinder, NG engines of a given technology have less peak power than gasoline or diesel-fueled versions.

compression ignition (CI) A method of initiating combustion in an internal combustion engine (ICE) in which heat of compression causes ignition. The CR of CI engines can range from 16 to 22 when diesel fuel is used. Suitable CRs for natural gas, EtOH, or MeOH fall

 © 2004 Elsevier Inc. All rights reserved.

between those for gasoline and diesel fuel. These fuels are more naturally suited to use in SI engines.

compression ratio (CR) CR is a measure of the degree of expansion of volume in a cylinder of an ICE from minimum to maximum, with the ratio being maximum/minimum volume. As a rule of thumb, the higher the CR of an engine, the greater is the efficiency.

corporate average fuel economy regulations (CAFE) These are government regulations originally adopted in the 1970s. The adoption of these regulations, along with oil price shocks in the 1970s, led to a doubling of new light-duty vehicle fuel economy from the early 1970s to mid-1980s.

E10 A blend of 90% gasoline and 10% volume ethanol, also called gasohol.

E85 A blend of 15% gasoline-like hydrocarbons and 85% volume ethanol introduced in the 1990s, after M85-fueled vehicles failed in the market.

electric vehicle (EV) Since very little of electricity is generated by use of petroleum products, electric vehicles are effectively alternatively fueled vehicles.

Energy Policy Act of 1992 (EPACT) This act "required" purchases of alternatively fueled vehicles by many classes of industry and government, as a way of starting a movement away from petroleum. Like AMFA, it "required" alternatively fueled vehicles but not use of alternative fuel. Further, it contained a broad exception that requirements need not be met if the alternative fuels and vehicles were not economically competitive. It had a goal (unrealized) of 10% replacement of gasoline in 2000, 30% by 2010.

ethanol or ethyl alcohol (EtOH) (also called E100) Produced in the United States from corn and in Brazil from sugarcane. Selected woods and grasses, urban waste, sugar beets, and rice straw are candidate feedstocks for expanded ethanol production.

Fischer-Tropsch (FT) processes Processes used to convert natural gas or coal to liquid fuels that can be used in conventional vehicles. These processes—particularly for planned facilities—primarily produce fuel suitable for use in compression ignition (diesel) engines. Due to need for much lower cost natural gas than available in North America, none of the announced plants are in North America. The diesel fuel produced (FTD) is high quality with respect to sulfur content (very low or zero) and cetane (high).

flex fuel vehicle (FFV) A vehicle that can switch between gasoline and an alcohol fuel (E85 or M85) in the same tank.

fuel cell (FC) Instead of combustion, FCs use chemical reactions to directly produce electricity. The proton exchange membrane (PEM) FC, the leading FC candidate for vehicle propulsion, reacts hydrogen and oxygen to produce electricity and water.

greenhouse gas (GHG) For purposes here, a gas specified in the Kyoto international protocol on global warming whose emissions are to be calculated and weighted to estimate the contribution of an individual signatory nation (the United States is not a signatory). Six gases are so specified. The most important by far is carbon dioxide, the predominant gas emitted as a result of combustion, and the predominant GHG. However, methane (CH_4) and nitrous oxide (N_2O) are also specified GHGs. The former is somewhat important to account for in use of natural gas fuels, the latter in ethanol fuels.

heavy oil Approximately, crude oil with American Petroleum Institute (API) gravity degrees in the low 20 s is heavy. Concentrations and most production from heavy oil are from California, Canada, and Venezuela.

hybrid electric vehicle (HEV) A hybrid is a vehicle that uses both electrical and mechanical systems to provide power to drive the vehicle. Many configurations are possible. Commercially available light-duty vehicle HEVs run on gasoline, while heavy-duty vehicle HEVs run on diesel. It is possible to configure a hybrid so that it can be plugged in to the electric grid, to provide part of the energy used to move the vehicle. When electrically operated, such a vehicle would be alternatively fueled. This type of hybrid is not presently receiving much attention, but the Electric Power Research Institute continues to evaluate it. Also, the California Air Resources Board provides incentives in its emissions regulations (see CALEV) for future plug-in hybrids.

hydrocarbons (HC) Chemical constituents of gasoline and diesel that are also a tailpipe and evaporated pollutant. Gasoline is more volatile than diesel fuel and contributes more hydrocarbons to air pollution than diesel fuel. The chemicals in gasoline (and those in alternative fuels) have a wide variation in "reactivity" in ozone formation. Many regulations of gasoline and alternative fuels are designed to change the composition of the fuel to reduce its reactivity and contribution of hydrocarbons to ozone.

hydrogen (H_2) There are many technologies and feedstocks from which hydrogen can be extracted, one of its great potential advantages. It is presently predominantly produced from natural gas. Hydrogen is a gas with very low energy per unit volume. It is costly to transport and store, and it causes significant power loss relative to gasoline or diesel fuel used in engines of comparable technology. It is the fuel that allows the fuel cells presently considered the leading candidate for use in motor vehicles to work.

liquefied natural gas (LNG) LNG is created when natural gas is cooled to -127°C. LNG tends to be used instead of CNG on heavy-duty trucks to reduce fuel storage volume. An LNG fuel storage system requires about half the volume of a CNG system. Still, LNG contains only about 50% of energy per unit volume as diesel fuel. For ocean shipping of NG, liquefaction is used in especially designed tankers.

liquefied propane gases, or liquefied petroleum gases (LPG) Despite the latter name, these gases are primarily

derived from condensing out NGLs from natural gas. The process that condenses these gases results in two streams. The lighter of these streams is primarily propane, with small amounts of butane and ethane. This is marketed as LPG. The heavier fraction is generally added to gasoline to control volatility. Some of the propane used in LPG is also a by-product of refining. Among gaseous fuels, a major advantage of LPG is that it can be liquefied with relatively little pressure. As stored in a lightly pressurized vehicle fuel tank, it has a much higher energy density than CNG or hydrogen. LPG is used exclusively in spark-ignited internal combustion engines.

M85 A blend of 15% gasoline-like hydrocarbons and 85% volume methanol introduced in the late 1980s. The hydrocarbons are included for flame luminosity, cold starting, and as a denaturant.

methane The predominant gas in natural gas. Relative to the many different hydrocarbons in gasoline, methane is nonreactive when it comes to forming ozone and is therefore not a regulated air pollutant in the United States. However, methane is a greenhouse gas that is included for control in other nations according to the international Kyoto protocol on greenhouse gases contributing to global warming.

methanol or methyl alcohol (MeOH) (also called M100) Predominantly produced from natural gas, but also producible from any hydrocarbon containing feedstock.

methyl tertiary butyl ether (MTBE) A gasoline octane enhancing and emissions reducing additive made primarily from methanol and isobutylene. Isobutylene in turn uses butane from NG processing and excess butane from petroleum refining.

natural gas (NG) A combination of gases that is predominantly methane. Larger volumes of other (condensable) gases include ethane, propane, and butane.

natural gas liquids (NGL) NGLs are condensable liquids from natural gas production. Propane is a leading NGL. Ethane is another. By the definition used here, when used by refiners to expand the volume of products, NGLs are alternative fuels.

nitrogen oxides (NO_x) Tailpipe pollutant resulting from combination of oxygen and nitrogen during combustion. Per unit of fuel, diesel fuel tends to contribute far more NO_x to air pollution than gasoline. NO_x is also a primary pollutant (itself regulated).

octane Index of ability of a fuel to resist preignition (knock) in a SI ICE. As octane rises, compression ratio can rise, and thermodynamic efficiency can also rise. There are two methods of measuring octane, the research (R) and motor (M) methods.

ozone Air pollutant for which standards exist that are typically violated in major urban areas on clear hot summer days. Ozone is created primarily by reactions of hydrocarbons and nitrogen oxides. Carbon monoxide is also a minor contributor.

reformulated gasoline (RFG) This term arose after the Clean Air Act Amendments of 1990. This law mandated changes in gasoline used within nine areas of the United States that were and are in violation of ozone air quality standards. The blending agent most commonly used in RFG is MTBE. Sulfur content of RFG was lowered, as was reid vapor pressure (RVP). In addition to summertime gasoline reformulation for ozone, wintertime reformulation to reduce carbon monoxide (CO) required 2.7% oxygen in gasoline. This was met with either MTBE or ethanol.

shale oil A much lower quality source of hydrocarbons than heavy oil or tar sands. Early 1980s estimates indicated that "high-quality" (a misnomer) shales contained the energy equivalent of about half of that of conventional crude oil. Such shale oil reserves were estimated to be concentrated in the United States (12 gigatonnes), Canada (6.5), Brazil (7), the former Soviet Union and Eastern Europe (4), and China (4).

tar (or oil) sands These have an API gravity of less than 12° and could be called very heavy oil. Venezuela was estimated in the early 1990s to have 4000×10^9 bbl of reserves; Canada, 1350; Russia, 600; and the United States, 28.

tetra-ethyl lead (TEL) The predominant gasoline octane-enhancing additive used from the late 1920s to mid-1980s in the United States. MTBE and EtOH have been the predominant octane enhancing additives used since TEL was phased out of gasoline in the 1980s.

Alternative transportation fuels are here defined as fuels produced from feedstocks other than petroleum. For petroleum, both conventional and unconventional oils are included in the umbrella term petroleum. Thus, unconventional oils such as oil sands in Canada and heavy crude oils in Venezuela are not alternative fuels, even though they can allow extension of the time that transportation can rely on refined liquid fossil fuels while continuing to use the same types of powertrains and fuel delivery infrastructure now in use. Oil shale—also an unconventional oil—is also defined here as petroleum. Oil shale is the most costly source of petroleum feedstock that would provide petroleum for refineries producing diesel or gasoline. Accordingly, oil shale is not presently a feedstock used in refined petroleum products. Since gasoline and diesel fuel can be produced from hydrocarbon feedstocks other than the sources defined here as petroleum, we use the terms petro-gasoline and petro-diesel to refer to gasoline or diesel fuel refined from conventional or unconventional oil.

The distinction here relies on replacement of refining capacity as the dividing line that defines alternative fuels. For example, in the 1970s, Mobil Corporation developed a process to produce gasoline

from methanol that was in turn produced from natural gas and built a plant in New Zealand. New Zealand has a slow turnover of its vehicle fleet and does not produce its own vehicles. Accordingly, the discovery of natural gas offshore at a time when oil prices were high (1970s) led to implementation of the gasoline from natural gas choice, rather than attempt to use methanol (an intermediate, cheaper product) directly in newly purchased vehicles designed for use of methanol. However, following the decline of world oil prices in the mid 1980s, gasoline production from this plant ceased. There is no discussion at the present time of reviving this inefficient method of producing transportation fuel. By this discussion's definition, this gasoline was an alternative fuel, though its properties as a vehicle fuel were nearly identical to gasoline refined from petroleum.

Similarly, by this discussion's definition, transportation fuels produced from coal or natural gas via the Fischer-Tropsch process are also alternative fuels. South Africa long produced a diesel fuel substitute from coal with the Fischer-Tropsch process, but this was done for political reasons, since inexpensive oil was not available to South Africa. Though an inconsequential alternative fuel at this time, Fischer-Tropsch diesel fuel from natural gas feedstock is in production and is scheduled for rapid expansion (from a small base) worldwide. This fuel is similar, though not identical, to diesel fuel processed from petroleum. Negligible sulfur content is a distinct advantage for future environmental needs.

1. BLENDS VERSUS NEAT AND NEAR-NEAT FUELS: 1970s TO PRESENT

If a fuel produced from a nonpetroleum feedstock is substantially like gasoline or diesel fuel, it is evident that it could be blended into such a fuel, and perhaps not recognized as having any difference from petro-gasoline or petro-diesel. In fact, number of chemicals other than gasoline or diesel fuel are blended into those fuels in order to alter their properties. In particular, for gasoline, properties such as octane, sulfur content, and volatility can be altered by blending chemicals—often produced from feedstocks other than petroleum—into the gasoline in relatively small percentages. For diesel fuel, the key properties that can be affected are cetane and sulfur content. Blending for improved properties also generally involves removal of some chemicals from gasoline, such as highly volatile butane.

Table I lists the leading alternative fuels in the United States, by our definition. The first two—methyl tertiary butyl ether (MTBE) and ethanol—are each separately used in considerably larger quantities as blending agents in gasoline than are all other alternative fuels combined. To supplement corn-based ethanol, the U.S. Department of Energy is developing technology to produce ethanol from other biomass feedstocks, including grasses and trees. On a percentage and total volume basis, Brazil uses more ethanol in gasoline than any nation, producing it from sugarcane. In 1989, Brazil consumed 11.9×10^9 liters of ethanol, while the United States consumed 3.6×10^9 liters in 1990. Quantities have varied in Brazil since that time, while quantities have moved relatively steadily upward since then in the United States. However, 2002 U.S. consumption is still well below the 1989 level in Brazil.

Vehicles designed to use only ethanol were the majority of SI ICE-powered vehicles produced in Brazil in the mid-1980s, but this program was reversed after ethanol shortages occurred and oil prices dropped sharply from the historical highs of the early 1980s. However, production of ethanol from sugarcane and use in motor vehicles has been estimated to allow significant reductions in greenhouse gases, and oil prices are high as this discussion is written, so interest in a second wave of expansion of ethanol use in Brazil is under way.

MTBE—a common gasoline additive to reformulated gasoline in the 1990s—is produced, on average, from additional processing of a combination of petroleum-derived and natural-gas-derived chemicals. The predominant natural gas derived chemical feedstock for conversion to MTBE is another alcohol, methanol. Early in the 20th century, methanol was produced from wood and called wood alcohol. However, now it is predominantly produced from natural gas feedstock and is an important industrial chemical. The U.S. EPA has funded past research to produce methanol from biomass, as was done early in the 20th century, but funding for this approach has been dropped. Since MTBE is produced from a combination of chemicals derived from petroleum and natural gas (methanol and isobutylene from butane), with natural gas likely contributing more than half of the available energy, it does qualify as an alternative fuel under the definition used here.

The term neat, or near-neat, refers to a fuel used in its pure form (neat) or with a small percentage of additives (near neat). Among such fuels, E95 (a blend

TABLE I

Some Key Alternative Fuel Attributes

Name	Primary feedstock (s)	Cetane/octane (motor/ research)	2002 Sulfur content	Energy/ volume (kj/L)	Pressure for storage	Vapor pressure at 38°C
Gasoline	Petroleum	80–90 (M)	300 ppm (40 in CA)	30–33	N/A	54–62 after 1992 (48 in CA)
		88–100 (R)				
Diesel No.2	Petroleum	40–55	300 ppm	35–37	N/A	<1
Methyl tertiary butyl ether (MTBE)	Natural gas, Isobutylene from natural gas or refining	100–118 (pure state)	Max. 10 ppm by weight	Not obtained	N/A	Lower than gasoline
		98–120 (M)				
		115–135 (R) (blending)				
Ethanol (EtOH)	Corn, sugarcane, or beet, switchgrass, poplar and other biomass	96–130 (pure state)	None unless included in denaturant	21.1	7–21 kPa	15.9 kPa at 38°C (pure state); causes nonlinear increase in gasoline vapor pressure in blends.
		100–104 (M)				
		132–135 (R) (blending)				
LPG (Propane data)	Natural gas liquids, refining byproducts	112/97	Not obtained	23	43.5 kPa	1303
Methane (CNG, LNG)	Processed natural gas	120	Not in methane, varies in CNG	9 at 20700 kPa	(CNG) at 17–25000 kPa	N/A
Electricity	Numerous.					
	Oil % in U.S. only 2 to 3%	Not applicable	Not applicable		N/A	N/A
Methanol (MeOH)	Natural gas	89 (M)	Near zero	15.8	7–21 kPa	32 kPa at 38°C (pure state); causes nonlinear increase in gasoline vapor pressure in blends.
		109 (R) (purestate)				
		98–100 (M)				
		132–134 (R) (blending)				
Fischer-Tropsch diesel	Natural gas		Near zero		N/A	<1
Biodiesel	Soybeans, rapeseed, waste cooking oils	52	None	32	N/A	<1
Hydrogen	Many	No rating	None	2 as gas, 7 liquid	>25000 kPa	N/A

Note. Values in this table are for illustration only. See state, federal, and national regulations and industry association and professional society recommendations before making any technical judgments about these fuels. Some cited values vary considerably, depending on local requirements and needs, or even test equipment capability.

of 95% ethanol and 5% hydrocarbons) and M100 (100% methanol) are both compression ignition engine fuels. Both may be regarded as failures, having been tried in urban buses in the early 1990s and found wanting. Neither of these fuels is currently of any importance in the United States. The alcohol-fueled urban buses were abandoned due to considerable maintenance problems emerging after several tens of thousands of miles of operation and disappointing fuel economy. Ethanol and methanol have very low cetane ratings and do not ignite well in CI engines at low loads (idling conditions). Since urban buses spend a high percentage of engine hours idling, this experiment was ill fated. Attempts to develop economically competitive cetane-enhancing additives for alcohols in CI engines were not successful.

LPG and CNG are used as neat fuels in spark ignited light-duty vehicles. CNG and LNG are used in both spark-ignited and compression-ignited engines in heavy-duty vehicles. Both CNG and LPG enjoy far greater success in 2002 than E95 and M100. M85 was used in FFVs in light-duty vehicles with spark-ignited engines in the early 1990s (but no longer), while E85 can be used (but generally is not) in the numerous E85 FFVs sold in the late 1990s and at the present time. Electricity was used for electric vehicles, especially in California, but also more aggressively than elsewhere in Georgia and Massachusetts. However, automakers are no longer producing full-function electric vehicles, after marketing several models in the late 1990s. Fischer-Tropsch diesel is blended into petro-diesel in very small quantities, if it is used at all. However, a few large Fischer-Tropsh plants are scheduled for construction in Southeast Asia and the Middle East, so this fuel may be used increasingly in the next decade. It can be used in existing compression ignition engines, especially when blended with petro-diesel fuel. The Alternative Fuels Data Center provides information on numbers and locations of five neat or nearneat alternative fuels—CNG, E85, LPG, electricity, biodiesel, and LNG. Presently the fewest stations nationally among these five types are those providing biodiesel.

For motor vehicles used worldwide, there tends to be a correspondence between fuel properties, engine types, and size (duty) class of vehicle. Gasoline is used in engines that use sparkplugs to initiate combustion (spark ignition). Such engines use fuels that have high-octane ratings. Except in Europe, where diesel fuel is taxed at a much lower rate than gasoline, gasoline tends to be the dominant fuel in light-duty vehicles used primarily for movement of persons.

Diesel fuel tends to be used in engines that do not use sparkplugs and instead use heat from compression of air to ignite injected fuel—so called compression ignition (CI) engines—which are more commonly known as diesels. Diesel-fueled CI engines are generally used in trucks (heavy-duty vehicles), locomotives, and ship engines to move freight, or in buses to move large numbers of people. Diesel fuels use cetane rating as the figure of merit rather than octane. Among candidate fuels for use in internal combustion engines (ICEs), octane and cetane tend to be inversely related.

Some analysts distinguish engines by normal operating speed, in revolutions per minute (rpm). Generally, as the size and power of an engine drops, its normal operating speed rises. For petroleum-based fuels, the quality of fuel used and degree of refining needed tends to increase as normal engine operating speed increases. Ship engines use one of the lowest quality refined petroleum products, residual oil, and operate in the hundreds of rpm. Locomotive engines also use low-quality distillate oil, while truck engines use relatively high-quality distillate oil and are capable of peak rpm of about 4000. Light-duty passenger cars and trucks generally use spark-ignited engines and gasoline fuel, which is more highly refined than diesel fuels. These engines can peak at 5500 rpm and greater. The sulfur content of refined petroleum fuel declines as its quality increases. Volatility is important for spark ignition engines. Gasoline is more volatile than diesel fuel. Volatility is an environmental problem leading to evaporative emissions, so the volatility of gasoline was required to be reduced considerably since the 1970s. It remains true, however, that the volatility of gasoline (RVP, Table I) is considerably higher than diesel fuel. The mix of engines used has actually been adapted over the years to the mix of products that can be (and is) produced via refining.

One of the problems of near neat alternative fuels is that they consistently have different properties than gasoline or diesel fuel. Gases have to be compressed for convenient storage and behave differently if they leak. Alternative fuels have different octane or cetane ratings, requiring an engine to be adapted to make best use of such fuels. Natural gas, ethanol, and methanol all have higher octane ratings than petro-gasoline and lower cetane ratings (if there is a cetane rating) than petro-diesel. Unfortunately, if the engines are adapted to make best use of the properties of the alternative fuels, then they may not be able to run on available refined petroleum products. Thus, using alternative fuels most efficiently, in engines optimized and exclusively suited for them, creates a fuel supply

problem. Dedicating the engine to a scarce alternative fuel makes it improbable that an average consumer would use vehicles designed to use only that alternative fuel because of a lack of refueling stations.

The response to this is to design what are called dual fuel or flex fuel vehicles, or to limit target markets to fleets of vehicles that use the same refueling facility. Gaseous fuels in a dual fuel vehicle that can use refined petroleum products require two separate fuel tanks. Thus, the term dual fuel is applied to these vehicles. Vehicles that can switch between near-neat alcohol fuels and gasoline in the same single tank are called flex-fuel vehicles. Many flex fuel vehicles sold are capable of using both E85 and gasoline, but there are relatively few E85 stations, and it almost always costs more to run on E85 than on gasoline. Public policy is responsible for the existence of these vehicles, since the vehicle manufacturer is allowed under the AMFA to assume that the rated fuel economy is doubled from its gasoline rating. Since fuel economy is regulated under the Corporate Average Fuel Economy (CAFE) regulation, and consumers are buying larger and less fuel-efficient vehicles, the incorrect assumption that these E85 flex fuel vehicles actually use E85 is beneficial to automakers wishing to provide larger vehicles to U.S. consumers.

On the other hand, should oil supplies become tight and oil prices rise sharply, having these vehicles on the road could allow biomass-based ethanol production facilities to be built, and a refueling infrastructure, during the approximate average 13-year lifetime of these vehicles. The prior sales of E85 FFVs could then allow successful marketing of biomass-based ethanol. Such ethanol would also reduce greenhouse gases significantly, so the existence of these vehicles does provide the United States with some degree of energy security and an additional environmental option, should it need these in the early 21st century. However, in the short run, these options are made available by selling vehicles that consume more refined petroleum products than would otherwise be the case under CAFE regulations. If significant expansion of ethanol fuel for transportation is to be achieved, biomass-based ethanol production must move from corn to woody biomass or switchgrass feedstock. There is little evidence that plans are under way to accomplish such an expansion in the United States within the lifetime of the E85 FFVs that have been placed in the fleet in the early 21st century. This contrasts with plans for expansion of FTD production from natural gas outside the United States and considerable expansion of production of Canadian oil sands.

Among nations, the United States has required more sophisticated and highly refined fuels for purposes of low emissions and pays more for suitable oil than other nations. One exception to the rule of thumb that the United States requires the most refined and cleanest fuel is Sweden, where a very clean (low-sulfur) diesel fuel is required for use in urban areas. Within the United States, cities with severe air quality problems use the most sophisticated and highly refined petroleum products for emissions control. As a result of the Clean Air Act Amendments of 1990, nine areas with severe ozone air quality violations were required to introduce reformulated gasoline to reduce the emissions of the light-duty gasoline burning fleet of vehicles in those areas.

The state with the worst air quality problems has been California. As a result of uniquely severe problems, California developed the first emissions regulations, followed by the federal government. California has since retained the legal authority to set more strict regulations. One result is that the most costly and most highly refined reformulated gasoline in the country is required in California. Prior to the Clean Air Act Amendments of 1990, California adopted regulations that included a requirement for zero-emissions vehicle sales, starting in 1998. The states containing the nine regions required to use reformulated gasoline in the Clean Air Act Amendments of 1990 were allowed to adopt the California standards if they chose to do so. New York, Massachusetts, and other New England states attempted to do so, but legal challenges by the automobile industry were fought throughout the 1990s. Nevertheless, as a result of a negotiated agreement with California, electric vehicles were produced and test-marketed by major manufacturers through the latter part of the 1990s, and electric refueling infrastructure was set up by several states. Most noteworthy, of course, was California.

During the 1990s, the additive MTBE, an alternative fuel by the definition used here, was the dominant additive in reformulated gasoline. Expanded use of reformulated gasoline led to expanded use of MTBE and, in effect, expansion of the amount of natural gas feedstock used in transportation fuels. The oil and automobile industries have exhibited and continue to exhibit a strong preference for MTBE as an octane-enhancing and emissions-reducing additive rather than ethanol. It can even be safely asserted that, absent regulation to the contrary, tetraethyl lead remains the preferred octane enhancer. However, lead removal regulations are spreading worldwide,

and only the lowest income nations still allow their refiners to take advantage of its powerful octane-enhancing effects. Footnotes in the World-Wide Fuel Charter make it clear that lead is still not gone and its content must be regulated. This publication by automakers (no oil companies are coauthors) states, "leaded gasoline should be removed as soon as possible." Recommendations for gasoline are for unleaded gasoline, not gasoline generically. For the dirtiest of the four unleaded gasoline types, lead content equal to a "maximum level of 0.005 g/l is acceptable during the transition period."

At the same oxygen concentrations, blending ethanol causes an increase in volatility and, as a result, refining severity must be greater to obtain a specified volatility for the gasoline-oxygenate blend. Further, ethanol has an affinity for water and is a stronger solvent than gasoline. As a result, it cannot be blended at the refinery and piped to customers, while a gasoline MTBE blend can. Ethanol is often splash-blended at a location close to the final customer. Ethanol and associated impurities are more likely than MTBE to have materials compatibility problems with fuel systems. Accordingly, the World Fuels Charter explicitly indicates a preference for ethers rather than ethanol. This preference is stated despite the fact that the dominant ether MTBE is being removed from gasoline in the United States because of its contribution to water pollution.

In the Midwestern United States, the political support for blending of ethanol produced from corn into gasoline was very strong starting in the 1980s, after the two oil price shocks of the 1970s. It must be remembered that, although most analysts focus almost exclusively on the politics of support for ethanol, ethanol and tetraethyl lead were the two competing additives to increase octane in 1920s gasoline. Tetraethyl lead won the first battle to be the octane enhancer of choice, but by the 1980s it was being removed from gasoline both because of its toxic side effects and its poisoning of the newly required tailpipe emissions reducing catalytic converters. After 1975, all gasoline marketers were required to provide at least one grade of unleaded gasoline. The removal of lead from gasoline meant that refiners needed other additives or additional refining investment to provide octane that was being lost as lead was removed. So one of the technical arguments that could properly be brought to bear by ethanol advocates was its ability to provide octane that was needed as tetraethyl lead was phased out. Similarly, ethanol is a domestically produced fuel, so its substitution into gasoline also was correctly argued to reduce dependence on foreign petroleum. The political advantage of these arguments in states in which corn production is a major part of the economic base is clear. As a result of this combination of factors, and one more, the E10 blend of 10% ethanol by volume and 90% gasoline—so called gasohol—has become prevalent in the Midwestern United States. EPA rules revisions in 1979 and 1981, at the peak of real oil prices since World War II, allowed gasohol to be marketed as a gasoline substantially similar to industry average gasoline. Ethanol use in gasoline has reasonably steadily expanded since that time.

The other factor that played an important role in the success of ethanol was the 1980s discovery that gasohol reduced the wintertime emissions of carbon monoxide in the front range of Colorado, helping that area deal with its wintertime air quality problems. This was also true of MTBE. The reduction occurred primarily in older vehicles using carburetors, but the timing was such that this reduction was perhaps the final winning argument in favor of ethanol expansion in Colorado and elsewhere in the cold Midwestern United States. Though the amount of reduction of carbon monoxide by using gasohol in modern fuel injected vehicles is negligible, the 1980s proponents did not wish to wait for the advanced fuel injected 1990s vehicles to reduce carbon monoxide, and they further pushed ethanol (along with MTBE) into the gasoline blend market.

Ethanol also had opponents who estimated that it required so much energy to produce that it was actually a loser in terms of energy use and greenhouse gases. However, in the 1990s, reinvestigation of the efficiency of production of corn and of ethanol led to estimates that corn ethanol produced in the 1990s and since actually reduces greenhouse gases. While it was true that ethanol of the 1980s increased greenhouse gases, the experience with production of ethanol led to improvements in plant design, and refinement of existing plants, increasing efficiency of ethanol production. In the meantime, the productivity of corn growing rose steadily, and fertilizer use per bushel produced dropped (reducing emissions of the greenhouse gas nitrous oxide), so the contribution of corn farming to the greenhouse gas emissions of ethanol also dropped significantly. Thus, ethanol's success in getting into the market and staying there has been the result of a diverse set of factors and to evolving emphasis on key attributes. For the early 21st century, it appears that the absence of negative effects of ethanol on groundwater is going to give it an advantage relative to MTBE, and it will

likely expand in use as a gasoline additive for yet another reason.

Ethanol is only blended into gasoline. However, research is ongoing on blending ethanol into diesel. Many problems have been noted (often similar in nature to those for adding ethanol to gasoline), so the World-Wide Fuels Charter recommends that more research is needed before motor vehicle and engine manufacturers and governments could acceptably agree that ethanol blends in diesel fuel are technically suitable for use.

There are multiple state and federal subsidies or tax relief programs for ethanol. Generally, in the Midwest, there are state subsidies and tax exemption programs that supplement federal subsidies/exemptions, leading to a much higher market penetration of ethanol blends into the Midwest. In addition to regulations and subsidies promoting gasohol throughout the Midwest, there is a federal oxygenate requirement for reformulated gasoline. Since both MTBE and ethanol contain oxygen, this requirement can be met with either. Since ethanol contains a higher proportion of oxygen, a lesser proportion of ethanol blended into reformulated gasoline can meet a given gasoline oxygen percentage requirement. The percentages of ethanol in reformulated gasoline, when ethanol is present, are often considerably less than the 10% portion in gasohol.

Compared to most refined petroleum products, alternative fuels are generally less complex chemicals. Refined petroleum products are a soup of chemicals tailored to match the attributes of the engines and fuel delivery system developed for them over the years. If alternative fuels cannot fit into this preexisting system with little required modification, then a significantly different system has to be developed. This is known as the infrastructure problem. Blends that allow alternative fuels to make use of existing fuel delivery infrastructure obviously represent the easiest way to get alternative fuels into the system, but they nevertheless have limitations. Generally these limitations are defined by their chemical properties in interaction with the gasoline (or diesel fuel) with which they are blended. A mixture of gasoline and an alternative fuel never has the properties of either fuel. Since the fuel delivery and maintenance systems (as well as vehicles) are designed for gasoline or diesel fuel properties, the limiting factors tend to be set by rules, regulations, and practices originally established for the existing dominant fuels. Usually, the alternative fuels are blended into gasoline or diesel fuel because they provide desired enhancement in various properties, but the extent of their addition is limited by undesirable deterioration in other properties. In particular, the octane blending benefits of tetraethyl lead, ethanol, and MTBE are highly nonlinear, and these components are almost always considerably more costly than gasoline. Thus, the amount blended into gasoline tends to be minimized to give the desired properties. In the case of gasohol, governments got around this attribute by specifying subsidies only for a 10% blend. This property also meant that it was more difficult for refiners to cope with the last small amounts of tetraethyl lead removed from gasoline than the previously removed quantities.

Another important point is that the blending octane of ethanol and MTBE tends to be higher than the octane rating of the pure compound. Accordingly, when used primarily for purposes of providing octane, ethanol and MTBE have a higher value blended into gasoline than used in near-neat form in a vehicle (see Table I).

To a far greater extent than alternative fuels that are suitable blending components, neat fuels will have properties that require significant adaptation of the system. Mechanics have to be retrained. New fuel infrastructure has to be built or existing facilities adapted. For example, diesel fuel pools on the floor if it leaks and is quite difficult to ignite. Natural gas rises in the air if it leaks and ignites readily in the presence of a spark if its concentration is high enough. Electric resistance heaters at the roof of a diesel bus maintenance garage could not ignite leaked diesel fuel, but if natural gas were to leak, an explosive fire could easily result. For this and other reasons, conversion of bus maintenance facilities to safely maintain natural gas buses can be costly. This is but one example.

2. *FUEL PROPERTIES AND FUEL SWITCH EFFORTS*

The 1980s and 1990s show that introduction of alternative fuels is quite difficult. At the time of publication, energy security was a major concern, due to high oil prices and elevated dependence on oil from politically unstable nations. Interest in alternative fuels waxes and wanes as oil prices rise and fall. When the United States Energy Policy Act (EPACT) of 1992 was passed, memory of a 1988–1990 oil price shock heightened by the Iraq-Kuwait war of 1990, followed by a recession, had caused elevated concerns with petroleum dependence by the

United States. This act nominally called for the United States to be using 10% alternative fuels in highway transportation by 2000, and 30% by 2010. The act was exhortative and had loopholes if economics did not support use of alternative fuels. For most of the 1990s, oil prices were well below the levels that induced the writing of the EPACT, and economics did not support use of alternative fuels to the extent envisioned in the EPACT. Nevertheless, one may take a glass partially full perspective toward the 1990s, since expansion of use of alternative fuels did occur, and numerous potentially valuable attempts to implement alternative fuels were made.

While it is possible to emphasize the perspective that energy security was the primary driver behind the EPACT, this would be misleading. Prior to the EPACT, the Clean Air Act Amendments of 1990 (CAAA90) were passed. At the time, alternative fuels were also regarded as cleaner fuels than gasoline. Prior to these amendments, California had developed low-emission vehicle regulations with far-reaching implications for automobile makers. The California regulations included requirements for introduction of zero-emissions vehicles, as well as other low-emissions vehicles much cleaner than the motor vehicle industry could produce. Prior to writing of the California zero-emissions vehicle requirements, General Motors had announced that it would develop a commercial electric vehicle. Zero emissions at the time implied electric vehicles. California had also been experimenting with methanol as an alternative fuel and anticipated that methanol vehicles would have considerably lower emissions than gasoline vehicles. An extremely clean experimental vehicle dedicated to the use of methanol was highlighted in research of the time. This vehicle was contrasted with a methanol FFV that had less power, slower acceleration, and higher emissions. The automobile industry had worked with California to develop versions of its standard vehicles capable of running on M85 and gasoline.

The summer of 1988 had been an unusually hot summer. As the environmentally motivated removal of the toxic but highly effective octane enhancer tetraethyl lead from gasoline was coming to an end, oil companies added volatile chemicals such as butane to keep the gasoline octane at levels previously offered to consumers. This caused greater amounts of evaporation of hydrocarbons, which tended to cause ozone air quality to deteriorate. Combined with the hot summer, this led to a reversal of years of consistent declines of summertime ozone concentrations and an alarmed response by environmentally concerned citizens. The CAAA of 1990 resulted directly from this reversal of progress. California and others thought that M85-fueled vehicles and electric vehicles might be used to address the need for even lower vehicle and refueling delivery system emissions. However, as work progressed on the CAAA, oil prices were rising, and a war between Iraq and Kuwait occurred, increasing the interest in alternative fuels for energy security reasons. EPACT followed 2 years later, and alternative fuel advocates were excessively optimistic.

The share of alternative fuels in use in highway transportation in 2000 fell far short of the 10% EPACT goal, and the probability of falling short of the 30% goal for 2010 is almost 100%. Instead of electric and M85 vehicles, the more successful alternative fuels are ethanol and MTBE blended into gasoline. Such fuels are termed replacement fuels in the Energy Policy Act of 1992, while they are referred to here as alternative fuels. The more successful neat and near neat alternative fuels are compressed natural gas (CNG) and ethanol, not electricity and methanol, the seeming leaders when EPACT was written. But the true success of ethanol and natural gas to date are not as neat or near-neat fuels, but as a constituent of gasoline. The vast majority of ethanol fuel produced in the United States is blended into gasoline. The amount of natural gas as a feedstock making its way into light-duty vehicles is considerably greater in MTBE than in CNG.

2.1 Status of Alternative Fuels and Trends Under Way

The U.S. Department of Energy's Energy Information Agency began compiling the amount of alternative fuels and alternatively fueled vehicles sold in the United States, by region, by fuel type, and by vehicle class since passage of the EPACT. Thus, there is a record of what has happened during the 1990s. The Alternative Fuels Data Center, maintained by the National Renewable Energy Laboratory, lists refueling stations by type of alternative fuel and provides technical details on alternatively fueled vehicles. The Electric Vehicle Association of America, since renamed to the Electric Drive Vehicle Association, provides data on the history of electric vehicles, by class of vehicle. The Natural Gas Vehicle Coalition provides information on natural gas vehicles and public policy (incentives) information for consumers. The Fuel Economy Guide, maintained by the U.S. EPA and Department of Energy, provides fuel

economy ratings of alternatively fueled vehicles in its published version of the guide. In the Web version, a weighted ranking from zero to ten (ten is best) is provided on "criteria pollutant" emissions (emissions for which standards exist). Information on full fuel cycle greenhouse gas emissions is also provided on the Web site.

2.1.1 Blended Alternative Fuels

According to Energy Information Agency (EIA) data, in the early years of the 21st century, the share of MTBE in gasoline was about 2.5%. The share of natural gas feedstock versus petroleum feedstock for MTBE is not available, but 50% is a good rough estimate. Thus, about 1.25% of gasoline is derived from natural gas feedstocks. Ethanol is blended into about 0.8% of gasoline. Thus, by the definition of alternative fuels used here, a little over 2% of gasoline is presently alternative fuel. Note, however, that diesel fuel, residual oil, and jet fuel are also petroleum products used in transportation, so the share of alternative fuel blended into refined petroleum products is probably a bit over 1%. However, this does not account for the use of hydrogen and natural gas liquids to produce other blendstocks used in refined petroleum products. These are not reported by EIA in its alternative fuel statistics. Hydrogen produced for such purposes is usually derived from natural gas, so the percentage of nonpetroleum feedstocks in refined petroleum products probably exceeds 1% by a significant margin.

The future for MTBE in the United States is troubled. It was adopted to reduce tailpipe emissions. The risks of its leakage and its effects if it did leak into aquifers were mistakenly judged to be small relative to the air quality benefits. In *very* small quantities it ruins the taste of water. It also is highly resistant to biodegradation and has negative health effects at small concentrations. The flaw in regulatory analyses of MTBE were that the concerns over concentrations in water were based on health effects rather than taste, which has a considerable economic value, even if the water is safe to drink. Health effects of concern are not reached until concentrations reach those well above the point where it fouls the taste of water. The effect on taste appears to be the Achilles' heel of MTBE. Due in part to leaks of underground storage tanks and in part to unburned fuel in two stroke boat motors, MTBE has tainted the water supplies of a number of localities, leading to a backlash against it and widespread concern over its long-term environmental effects. Small quantities are being found in many locations as a result of unburned MTBE emitted from motor vehicle tailpipes, so its gradual buildup, in light of public alarm over taste of MTBE contaminated water at several prominent locations, has led to widespread concern, and movements to remove MTBE from gasoline. Early in the 21st century, ethanol use as a blending component of gasoline will increase, while MTBE use will decrease. The net effect will likely be a significant tapering off of growth of alternative fuels blending into gasoline —perhaps even a decline.

2.1.2 Neat and Near-Neat Alternative Fuels

Liquefied petroleum gases (LPG) are the leading neat or near-neat alternative fuels in the United States Despite the name, a large fraction of the fuel is derived from natural gas liquids, propane in particular. LPG technically can be a mix of gases. The mix is generally butane and propane, with the percentage of butane limited. The abbreviation sometimes stands for liquefied propane gas, because propane is the primary constituent. Only two models of light-duty LPG vehicles are listed in the 2002 Fuel Economy Guide. Both are versions of the Ford F150 pickup truck, the leading selling vehicle model. Among vehicle body types, pickups are sold in higher quantities in rural areas than any other. There is also a medium-duty truck that uses LPG. In any case, LPG is predominantly an alternative fuel sold in rural areas. Vehicles are often converted to use LPG. These are generally in areas where air quality is not a problem and vehicles are not inspected for emissions. According to EIA, LPG sales in the United States for transportation have been about 250 million gallons of gasoline equivalent since about 1994. For comparison purposes, 1% of the 2001 gasoline market was 1260 gallons, about five times as much as the gasoline equivalent sales of LPG. Thus, sales of LPG, at about 0.2% of gasoline, are far less than ethanol blended into gasoline (about 0.8%) or natural gas used in MTBE blended into gasoline (about 1.25%). Further, sales are not growing. Thus, according to earlier calculations, merchant hydrogen is probably already used in refined petroleum products to an extent similar to the leading neat alternative fuel, LPG. The use of merchant hydrogen by refiners is probably rising.

Among neat and near-neat alternative fuels, natural gas is second in sales to LPG. In 2001, sales of natural gas in transportation were estimated to be about 105 million gallons of gasoline equivalent, up from about 20 in 1993. So growth of sales of natural gas is significant, but in 2001, natural gas as a neat fuel represented 0.1% of gasoline sales, far less than

the 1.25%-plus share of natural gas feedstock that works its way into refined petroleum products via MTBE and hydrogen.

After natural gas, not one of the remaining alternative fuels constitutes more than about 0.01% of gasoline sales. This includes liquefied natural gas, M85, E85, M100, E100, and electricity.

According to the 2002 Fuel Economy Guide, the mix of new 2002 light-duty alternative fuel vehicles offered to consumers was considerably out of balance with the relative success of alternative fuels to date. Twenty-seven E85 FFV models were offered, versus eight CNG models. As noted, there were two LPG models and two electrics. The number of electric vehicles offered had dropped sharply from earlier years, and the number of CNG models had dropped moderately. Numbers of E85 vehicles had expanded rapidly. No M85 vehicles were offered in 2002. Such vehicles had been available early in the decade. Methanol and electricity, at the end of a decade or so of experimentation, appear now to be abandoned as alternative fuels for vehicles designed to clean the environment or to provide energy security.

It has been pointed out that a prime motivator for several of the alternative fuels has been emissions reduction rather than energy security. MTBE has been used in reformulated gasoline for that purpose. Ethanol's ability to reduce carbon monoxide emissions in the 1980s fleet of vehicles helped it expand its market share. Electric vehicles were the hoped for zero-emissions vehicle for California and other states that wished to adopt the California low-emissions vehicle standard under rules written into the Clean Air Act amendments. Natural gas was also advocated as a clean fuel. Compressed natural gas vehicles designed to use only natural gas have proven to be very clean, on average. Figure 1 illustrates an average rating emissions index rating from the Fuel Economy Guide's 2002 models. It shows that CNG vehicles obtain almost the same rating as electric vehicles. It was stated that LPG tends to be a rural fuel, and few clean LPG vehicles are available from manufacturers.

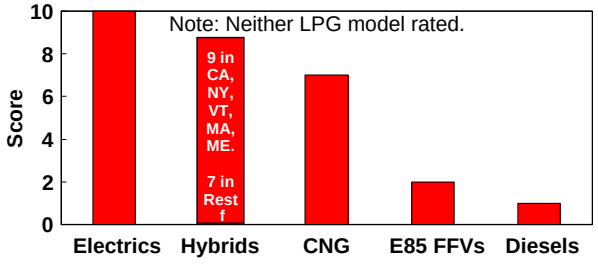


FIGURE 1 Average air pollution score of non-gasoline light-duty vehicles and hybrids on the EPA Fuel Economy Guide Web site (10 is best, 0 is worst).

2.2 Location Patterns

Since EPACT was enacted, efforts to introduce alternative fuels have included installation of refueling stations to provide the fuels. Electric refueling stations and E85 refueling stations are highly concentrated geographically, while natural gas refueling stations are not.

The cleanest of vehicles, the CNG and electrics, have had infrastructure for them preferentially installed within a 25-mile radius of large metro areas. The share of LPG stations rises sharply in the 25- to 50-mile annular ring outside metro areas. Small metro areas have a higher proportion of LPG stations than the nation as a whole, and they have a considerably larger share than within a 25-mile radius of large metro areas. The proportion of electric vehicle stations drops far more sharply in the small metro areas than does the other clean fuel, CNG. Where electric vehicle stations exist, the proportion of electric refueling stations found within the 25-mile radius of the metro area is consistently greater than for other fuels. It is fair to say that attempts to introduce electric vehicles have focused on the core of large urban metro areas with ozone air quality problems. It is also clear that the mix of fuels being sold is far different from the mix of refueling stations available to refuel them with an alternative fuel.

Various states and metro areas choose different fuels for primary emphasis and have different levels of effort at developing an alternative fuel infrastructure. The shares of electric stations in major metro areas of California are greater than elsewhere. California has the most electric vehicles in the nation. However, California is also listed as the state with the most LPG vehicles, amounting to about nine times as many as electrics. California also leads the nation in number of CNG vehicles and has about five times as many CNG vehicles as electric vehicles. In fact, California is estimated to have over twice as many methanol vehicles as electrics, as a result of methanol vehicles remaining from the methanol programs of the early 1990s and small numbers of electric vehicles sold since then.

Metro areas outside of California that also have electric vehicle refueling stations have not emphasized electric stations to the degree that California has. There, CNG is the fuel with the largest

proportion of alternative fuel stations, not electricity. The only metro area that has chosen to emphasize E85 refueling stations is Minneapolis. Minneapolis has no ozone violations, but it is in violation of the CO standard. Minnesota is the state with the most E85 FFVs actually using the fuel.

Among the metro areas examined, those not including electric vehicles also generally have fewer alternative fuel stations. Milwaukee, Chicago, and Houston have about the national proportion of LPG stations and a considerably higher proportion of CNG stations than the average. Texas is the leader in LPG vehicle totals, while Illinois is third. California, Texas, and New York lead in implementation of CNG vehicles, consistent with their populations.

Only a few states are estimated to have more alternatively fueled vehicles using ethanol, natural gas, or electricity than LPG. For Arizona, Arkansas, Delaware, Nevada, Utah, and West Virginia, natural gas vehicles are the leaders. Delaware, the District of Columbia, Maryland, New Jersey, Utah, and West Virginia have more CNG stations than any other type. The District of Columbia, Hawaii, Minnesota, and North Dakota are estimated to have more E85 vehicles than any other type. Only Minnesota has more E85 stations than any other type. Only Puerto Rico is estimated to have more electric vehicles than any other type. California, Georgia, and Puerto Rico have more electric refueling stations than any other type.

3. FUEL PROPERTIES AND CENTURIES-LONG TRANSPORTATION FUEL SWITCHES

Clearly, many detailed properties of alternative fuels are important to their success or failure. Nevertheless, it is also necessary to consider the overarching general properties and their role when taking a long view. Cost, a function of abundance of feedstock for the fuel, is, of course, a primary consideration. However, whether a fuel is a solid, liquid, or a gas is also important. Use of solid fuels in 19th century transportation (grains for horses, wood, and coal for steam locomotives) gave way to use of refined liquids from petroleum in the 20th century. Some energy forecasters hold the belief that the transportation fuel of the 21st century will be a gas—hydrogen. One argument in favor of the logic of a transition to hydrogen is that there is a natural transition from fuels with higher carbon content toward lower carbon content, with hydrogen at the end of the sequence. As far as the percentage of carbon versus hydrogen in the molecular weight of past and candidate future transportation fuels is concerned, the percentage declines as one moves from wood to coal to petroleum to natural gas (methane) to hydrogen. There are debates about a transition to hydrogen. Some argue that natural gas will be the next major alternative transportation fuel, followed by hydrogen. Others assume that natural gas will not be used in significant quantities in transportation and hydrogen will be the next major neat fuel.

The primary argument for an expectation that nonpetroleum fuels will be dominant in the 21st century is that petroleum will be depleted. Those who advocate this argument generally use a depletion model known as the Hubbert curve. By this model, petroleum will first be depleted, then natural gas depletion will follow within a very few decades. If natural gas were to be more abundant than presently estimated, then perhaps methane could be a significant worldwide transportation fuel well into the 21st century. If not, then the argument that hydrogen will be the dominant fuel can be based on the projection that *both* petroleum and natural gas will be depleted.

Hydrogen is seen as an energy carrier that can be produced via a large number of feedstocks and processes. A hydrogen delivery infrastructure for the 21st century would be analogous to the present electricity transmission and distribution system, where electricity fed into the system comes from many different primary sources. Presently, hydrogen is produced predominantly from natural gas, which is the cheapest resource for high-volume hydrogen production. In fact, the hydrogen production capacity in existence is as large in energy terms as most alternative fuels. Hydrogen production outside of refineries was estimated in 1995 to be equivalent to two days of gasoline consumption. Expressed on a percentage basis, this would be 0.55%. The leading use of hydrogen in 1988 was for ammonia production, while the second leading use was as a refinery input. Worldwide, in that time frame, 77% of hydrogen was estimated to be produced from natural gas, 18% from coal, and less than 1% from oil. In 1988, it was estimated that 37% of hydrogen was used by refineries. Thus, hydrogen production dedicated to transportation fuel may be on the order of 0.2% of gasoline production or more. Since 1988, use of hydrogen in refining has likely increased significantly. A large portion of the hydrogen

produced is actually used in petroleum refining to upgrade refined petroleum products to the tighter specifications being imposed largely for environmental reasons, and these restrictions have become tighter in the 1990s. In this sense, petroleum feedstock is already being replaced by natural gas feedstock for production of gasoline and diesel fuel in ways completely transparent (unknown) to the average consumer.

Another way that domestic natural gas use in refining has likely increased is through the expanded use of natural gas liquids. While oil production in the United States has typically declined, production of natural gas liquids has increased. One of the markets for these liquids has traditionally been in gasoline refining.

Even if the transportation system were to make a transition directly to hydrogen-fueled vehicles instead of an intermediate transition to methane-fueled vehicles, it is probable that natural gas would be the primary feedstock for the hydrogen used. Past and recent research on the costs of hydrogen production and delivery indicate that hydrogen will be significantly more costly than gasoline or diesel fuel, a clear drawback. Hydrogen is also more costly per unit of energy than natural gas or ethanol, but it has the potential advantage of being converted to tractive energy in a fuel cell vehicle more efficiently than these two fuels in spark ignited internal combustion engines. The high cost per unit of delivered hydrogen is directly related to the fact that hydrogen is a gas with a low amount of energy per unit volume. It would also be possible instead to use the natural gas directly in vehicles rather than convert it to hydrogen. Credible sources obtain differing estimates of the relative total full fuel cycle energy and greenhouse gas emissions of future transportation technologies. Depending on the source consulted and time frame examined, one can find estimates that (1) petro-diesel used in advanced hybridized conventional compression ignition engine vehicles, (2) natural gas used in advance hybridized spark ignited engine vehicles, (3) hybridized hydrogen fuel cell vehicles, or (4) electric vehicles will result in the least greenhouse gas emissions. Electric vehicles, however, have limited driving range relative to the other three types. This has proven to be a show-stopper for electric vehicle experiments of the 1990s.

When one considers a half century or full century time frame, then the issue of adequacy of natural gas supply as a feedstock for hydrogen production leads some to argue that there is no sense in attempting three transitions—first to methane and CNG, second to methane and hydrogen, and third to multiple primary sources of hydrogen. By this argument, transitions themselves are inherently difficult, so avoiding intermediate transitions and jumping to a long-term solution is desirable. Economists do not make such an argument, but some scientists, environmentalists, and motor vehicle manufacturers (for which transitions are indeed costly) do, for different reasons. Some environmentalists tend to stress the opportunities to produce hydrogen from nonfossil primary sources such as wind, hydro, and solar power. Some scientists stress production of hydrogen via nuclear power as a source of abundant supply. Some in the automobile industry want to spend scarce research dollars to develop reliable, reasonable cost hydrogen powertrain technology for vehicles before proceeding to a hydrogen-based transportation system.

An argument against hydrogen is that liquids are the ultimate fuel for transportation because of their energy density and ease of handling in a fuel system. In this regard, the advantage of a liquid is clear for internal combustion engines (ICEs). Gases tend to reduce the peak power output of an ICE of a given size and also require bulky pressurized tanks that require a lot of physical volume relative to liquids. Converting an ICE to hydrogen does offer promise of elimination of hydrocarbon and carbon monoxide emissions, but not nitrogen oxides. Further, the effective efficiency of the vehicle will decline due largely to the power loss, which will require a larger and heavier powertrain. While projections indicate that hydrogen powered ICEs could be advanced to have considerably lower greenhouse gas emissions than present ICEs, it is also true that present ICEs can also advance. At any level of technology, Ricardo Consulting Engineers estimated in 2002 that hydrogen-fueled ICEs using hydrogen from natural gas will cause higher greenhouse gas emissions than comparable technology petro-diesel vehicles. So according to Ricardo's estimate, unless there is some unequivocal great long-term advantage to immediately starting to convert to a natural gas derived hydrogen fuel infrastructure, there is no clear reason to switch to hydrogen ICEs instead of petro-diesels. Ricardo constructed estimates of two pathways to hydrogen, with the early introduction pathway taking advantage of ICE engines converted to use hydrogen and the late introduction pathway waiting until fuel cell vehicles are ready for the market, estimated by Ricardo to be about 2020. Still, by Ricardo's estimates, very advanced hydrogen ICEs could compete with hydrogen fuel cell vehicles in terms

of efficiency, so they may be worth examining on a research basis since they may be considerably cheaper than hydrogen fuel cell vehicles.

It should be pointed out that technology exists and has been demonstrated to produce gasoline and diesel fuels from unconventional oils, from shale, and from coal. However, use of these technologies in their demonstrated forms would increase greenhouse gas emissions and probably would cause other environmental difficulties. One possible solution to this problem would be to sequester carbon dioxide produced when coal or other heavy fossil fuels are converted to hydrogen. If this could be done at reasonable cost and in an otherwise environmentally acceptable manner with abundant coal, then coal might actually again provide significant quantities of transportation fuel. The effort to develop hydrogen fuel cell vehicles is, in effect, a long-term effort to reduce urban motor vehicle emissions to inconsequential amounts. Without environmental concerns over ambient air pollution and greenhouse gases to drive this process, it is likely that reliance on fossil fuels could and would continue perhaps through the entire 21st century. Given these concerns, energy sources such as nuclear, biomass, wind, hydro, and solar could provide original energy for hydrogen production for transportation instead of fossil fuels.

It is quite likely that if the clean fuel of the future is hydrogen, it will bring with it an entirely new powertrain. Just as steam and coal were matched for the 19th century and petroleum and the ICE for the 20th, the advocate of hydrogen will note that hydrogen and the fuel cell will be matched for the 21st century. The fuel cell is the prime mover (a generic term for the primary power source for a motor vehicle) for highway vehicles for which pure hydrogen is clearly the ideal fuel from a thermodynamic efficiency perspective. In particular, at this time, a type of fuel cell called a polymer electrolyte membrane (PEM) fuel cell is the candidate of the future. It is possible to obtain a jump in peak efficiency with FCs relative to ICEs—especially SI ICEs—if hydrogen is the fuel. FC powertrains using liquid fuels—methanol in particular—are also being developed, but such a powertrain's efficiency is not as high as for one equipped with PEM cells fueled by hydrogen stored on board. This is because, in addition to the fuel cell, the methanol fuel cell vehicle requires a reformer—a chemical processing device to convert methanol to hydrogen and carbon dioxide. So the perceived need to evolve toward gaseous hydrogen is promoted by the estimated high efficiency of the PEM FC using hydrogen stored on board the vehicle.

Thus, many think, with good reason, that the hydrogen-fueled PEM FC vehicle powertrain will be the prime mover and fuel combination for transportation in the 21st century. To the extent that this combination relies on natural gas feedstock, most analysts estimate that it will extend the life of natural gas resources relative to use of CNG in ICEs. As previously noted, using wind, solar, geothermal, biomass, coal, or nuclear electric power to produce hydrogen would also be technically possible. Compared to use of CNG in ICEs, the hydrogen-fueled PEM FC could significantly reduce urban emissions of carbon monoxide, hydrocarbons, and nitrogen oxides (the hydrogen ICE can also eliminate the former two and probably sharply reduce the latter). Some see the hydrogen PEM FC as the zero-emission vehicle that can succeed, now that electric vehicles have failed to garner consumer support.

Counterarguments against a switch to hydrogen essentially involve cost. Arguments that hydrogen is excessively costly imply many decades of reliance on the technologies and fuels discussed in the prior section. High overall hydrogen transportation system cost is because of the shift to an entirely new powertrain type that presently appears inherently more costly than ICEs and a shift of fuel delivery infrastructure, which is inherently more costly than for refined petroleum products or natural gas. The cost of the needed fuel infrastructure shift could be avoided to some degree if the fuel feedstock to be used in all cases is natural gas, while the powertrain could be reduced in cost by using NG directly in a natural gas-fueled internal combustion engine. In this regard, CNG can be burned in either spark-ignited engines for light-duty vehicles or in compression-ignited engines for heavy-duty vehicles. An extensive natural gas delivery infrastructure exists, but it would be completely inadequate for the significant use of natural gas in transportation as long as other users of natural gas continue to use it. Thus, the cost issue would be the cost of new natural gas infrastructure versus new hydrogen infrastructure, unless a single infrastructure compatible with either natural gas or hydrogen could be developed.

Fischer-Tropsch diesel fuel produced from natural gas can be used in compression ignition engines, which are presently dominant in the United States in heavy-duty vehicles. It is, of course, suitable for light-duty diesels as well, which are common in Europe.

Fischer-Tropsch diesel fuel is very clean with respect to sulfur and tailpipe emissions, and in present applications it is generally a premium-priced fuel stored in separate tanks from normal diesel. However, this is only necessary where emissions are problematic. Otherwise, Fischer-Tropsch diesel can be shipped and stored in the same equipment presently used for diesel fuel. According to most studies of light-duty vehicles, natural gas and Fischer-Tropsch diesel fuel pathways use more energy and cause greater greenhouse gas emissions than hydrogen in fuel cell vehicles. However, the differences are not large, and the cost of hydrogen is presently estimated to be considerably greater. A third nonhydrogen option is biomass-based ethanol (not corn-based). This fuel could be burned in the many E85 flex fuel vehicles being sold into the market. It does require a separate infrastructure for delivery of the fuel. This option can actually reduce greenhouse gases by more than hydrogen fuel cell vehicles for which the hydrogen is produced from natural gas. A comparison of a scenario heavily dependent on hydrogen and fuel cells versus one using the more conventional technologies discussed in this paragraph was one evaluated in 2003 in a joint U.S. Department of Energy and Canadian Energy Mines and Resources study.

The results of that study indicated that it is quite possible to envision an evolutionary shift of fuels for ICE engines that would make use of a combination of CNG and Fischer-Tropsch diesel fuel from natural gas feedstocks and ethanol from biomass. This shift could well provide the same level of service as if a single hydrogen with FC transition pathway were pursued, provide greenhouse gas emissions as low as for the hydrogen pathway, and use no more natural gas. The CNG would most likely be used predominantly in spark ignited ICE engines in urban areas with more severe air quality problems. As such, it would be the clean gaseous fuel used where air quality needs are greatest, instead of hydrogen. E85 would also be used in spark-ignited ICE engines, in areas where air quality was less of a concern. Fischer-Tropsch diesel fuel would be used in compression ignition engines in heavy-duty vehicles.

The two drawbacks of this latter option, which relies far more heavily on liquids than the hydrogen transition, are (1) once natural gas was depleted, a transition to hydrogen and fuel cell vehicles would probably still be necessary, and (2) there would be no zero-emissions vehicles in the mix of vehicles available to deal with air quality problems.

4. CONCLUSION

Neat and near-neat alternative fuels and alternatively fueled vehicles contribute very little to the transportation services of the United States. Where there has been a small degree of success, the states, localities, and businesses involved have generally chosen to emphasize only one or two alternative fuels. LPG is the largest neat alternative fuel in the market to date, but where cleaner natural gas or electric vehicles have been pursued, LPG seems to have been deemphasized. Natural gas and ethanol fuels have a tiny share of the market, but are still growing. Electric vehicles attained an even tinier share of the market and appear now to have been an experiment that will be abandoned.

One reason for abandoning electric vehicles is the promise that hydrogen fuel cell vehicles can also provide zero tailpipe emissions with far more consistent reductions in greenhouse gases and at a cost less than for electric vehicles, with considerably greater overall satisfaction to the consumer. Beyond this, many anticipate hydrogen to be the leading candidate fuel for light-duty vehicles for the 21st century. Nevertheless, for the next few decades the role of petroleum as the leading transportation fuel seems secure. However, projections that this cannot continue for the entire century are now conventional wisdom.

One question to ask at this time is whether it is wise to abandon the alternative fuels experiments of the 1990s and wait another decade or two until hydrogen is technically (and, it is hoped, economically) demonstrated, or should ongoing trends be reinforced, refined, and expanded via public policy or corporate investment. Another is whether improved transportation energy efficiency—largely ignored in this discussion—should instead be the primary goal for the next two decades.

SEE ALSO THE FOLLOWING ARTICLES

Biodiesel Fuels • Biomass for Renewable Energy and Fuels • Ethanol Fuel • Fuel Cell Vehicles • Fuel Cycle Analysis of Conventional and Alternative Fuel Vehicles • Fuel Economy Initiatives: International Comparisons • Internal Combustion Engine Vehicles • Oil Sands and Heavy Oil • Oil Shale • Transportation and Energy, Overview • Transportation and Energy Policy • Vehicles and Their Powerplants: Energy Use and Efficiency

Further Reading

"Alcohols and Alcohol Blends as Motor Fuels" (1986). International Energy Agency Report (Energy Information no. 580) prepared by the Swedish Motor Fuel Technology Co. (SDAB), Stockholm.

"Alternative Fuels" (2000). U.S. Energy Information Agency Web site, www.eia.doe.gov.

Alternative Fuels Data Center. U.S. Department of Energy Website, www.afdc.nrel.gov.

Bechtold, R. (1997). "Alternative Fuels Guidebook: Properties, Storage, Dispensing and Vehicle Facility Modifications." Society of Automotive Engineers, Warrendale, PA.

Electric Drive Transportation Association of America. (EDTA) Web site, http://capwiz.com/evaa/index.htm.

Franklin, P. M., *et al.* (2000). Clearing the air: Using scientific information to regulate reformulated fuels. *Environmental Science and Technology* **34**(18), 3857–3863.

"Fuel Economy Guide" (2002). U.S. EPA Web site, www.fueleconomy.gov.

General Motors Corp., *et al.* (2001). "Well-to-Wheel Energy Use and Greenhouse Gas Emissions of Advanced Fuel/Vehicle Systems—North American Analysis, Executive Summary Report." In 2003, available electronically at Argonne National Laboratory's Transportation Technology Research and Development Center Web site, www.transportation.anl.gov.

Graham, R., *et al.* (2001). "Comparing the Benefits and Impacts of Hybrid Electric Vehicle Options." Final Report, July 2000, Electric Power Research Institute, Palo Alto, CA.

Natural Gas Vehicle Coalition (NGVC) (2002). Web site, www.ngvc.org.

Owen, N., and Gordon, R. (2002). "Carbon to Hydrogen: Roadmaps for Passenger Cars: A Study for the Department for Transport and Department of Trade and Industry," RCEF.0124.31.9901. Ricardo Consulting Engineers Ltd. Department of Transport, United Kingdom. November 8.

Plotkin, S., Patterson, P., and Singh, M. (2003). "Overview of the 2050 Study." Presentation at the 82nd Annual Meeting of the Transportation Research Board. Session 687. Transportation Energy Use in the Long Term, Part 1—Insights from Long-Range Studies. Argonne National Laboratory, January 15, Washington, DC.

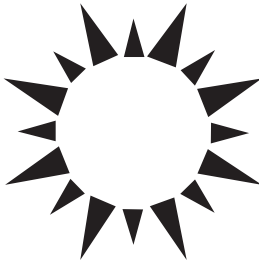
Wang, M. Q., Saricks, C. L., and Wu, M. (1999). Fuel Ethanol Produced from U.S. Midwest Corn: Help or Hindrance to the Vision of Kyoto? *J. Air Waste Management Assn.* **49**, 756–772.

Weiss, *et al.* (2000). "On the Road in 2020: A Life-Cycle Analysis of New Automobile Technologies" MIT Energy Laboratory Report No. MIT EL 00-003, Energy Laboratory, Massachusetts Institute of Technology, Cambridge, MA, October.

"World-Wide Fuel Charter." (2002). European Automobile Manufacturers Association, Brussels; Alliance of Automobile Manufacturers, Washington, DC; Engine Manufacturers Association, Chicago; Japan Automobile Manufacturers Association, Tokyo; December.

Turbines, Gas

LEE S. LANGSTON
University of Connecticut
Storrs, Connecticut, United States



1. Introduction
2. Gas Turbine Applications
3. Gas Turbine Cycles
4. Gas Turbine Components
5. Gas Turbine Usage

Glossary

Aeroderivative An aviation propulsion gas turbine (jet engine) used in a nonaviation application (e.g., an electric power plant) to provide shaft power.
Brayton Cycle Ideal gas cycle that is approximated by a gas turbine simple cycle.
Cogeneration A gas turbine operating to produce shaft power and to supply heat from its exhaust for some useful purpose.
Combined-Cycle Power Plant The combination of a gas turbine power plant and a steam turbine power plant. The gas turbine exhaust energy is used as heat input to generate steam.
Gas Generator The compressor, combustor, and turbine components of a gas turbine.
Isentropic A process that is reversible and adiabatic; having constant entropy.
Jet Engine An aviation gas turbine used for aircraft propulsion thrust power.
Rankine Cycle Ideal gas cycle that is approximated by a simple steam turbine power plant cycle.
Thermal Efficiency Useful output power divided by costly input energy rate, expressed as a percentage.
Working Fluid In the case of gas turbine, the gas (usually air) that flows through the machine to produce work or power.

The first jet engine-powered aircraft flight and the first electric power produced by a land-based gas turbine both occurred in 1939, making the gas turbine one of the youngest energy conversion devices. A gas turbine consists of a rotating assembly of a compressor to draw in a gas (usually air) and a turbine to extract power from the gas flow. Energy (usually combusted fuel) is added in a combustor between the two rotating components to heat the gas flow and power the turbine. Gas turbines are used to produce either thrust power (e.g., an aircraft jet engine) or shaft power (e.g., to turn an electric generator). The thermal efficiency and power output of a gas turbine are dependent on the compressor pressure ratio and on the gas flow temperature at the inlet to the turbine. Gas turbines range in size from hand-held units to whale-size units, weighing from a few kilograms to 300,000 kg, and producing power from tens of kilowatts up to 250 MW.

1. INTRODUCTION

In the broad sense, a turbine is any kind of spinning device that uses the action of a fluid to produce work or power (work per unit time). Typical fluids are air, water, steam, gaseous combustion products, and helium. Windmill turbines, water turbines in hydroelectric dams, and steam turbines have been used for decades to turn electrical generators to produce electrical power for both industrial and residential consumption. Simple turbines date back centuries, with the first known appearance in ancient Greece (Hero's turbine).

In the history of energy conversion, the gas turbine is a relatively new energy converter, or prime mover. The first practical gas turbine used to generate electricity ran at Neuchatel, Switzerland, in 1939 and was developed by the Brown Boveri firm. The first gas turbine-powered airplane flight also took place in 1939 in Germany using the gas turbine developed by Hans P. von Ohain. In England in the 1930s, the invention and development of the aircraft gas turbine by Frank Whittle resulted in a similar British flight in 1941.

Table I presents a list of energy converters and the year in which a first working device was successfully

 © 2004 Elsevier Inc. All rights reserved.

run. The gas turbine is by far the "youngest" of the energy converters listed. Note that the common automotive gasoline engine (Otto engine) has been around a long time (1876), as has the fuel cell (1839).

The name "gas turbine" is somewhat misleading, for it implies a simple turbine that uses gas as a working fluid. Actually, a gas turbine (as shown schematically in Fig. 1) has a compressor to draw in and compress gas (usually air), a combustor (or burner) to add combusted fuel to heat the compressed gas, and a turbine (or expander) to extract power from the hot gas flow. The gas turbine is an internal combustion (IC) engine employing a continuous combustion process, as distinct from the intermittent combustion occurring in a Diesel or Otto cycle IC engine.

TABLE I

History of Energy Converters

Energy converters	Date of first working device
Steam engine	1769
Fuel cell	1839
Otto engine	1876
Steam turbine	1884
Diesel engine	1897
Gas turbine	1939

Because the 1939 origin of the gas turbine lies both in the electric power field and in aviation, there has been a profusion of "other names" for the gas turbine. For land and marine applications, it is generally called a gas turbine, but also a combustion turbine, a turboshaft engine, and sometimes a gas turbine engine. For aviation applications, it is usually called a jet engine, and various other names (depending on the particular aviation configuration, or application), such as jet turbine engine, turbojet, turbofan, fanjet, and turboprop or prop jet (if it is used to drive a propeller) are also sometimes used. The compressor–combustor–turbine part of the gas turbine (Fig. 1) is commonly termed the gas generator.

2. GAS TURBINE APPLICATIONS

In an aircraft gas turbine, all of the turbine power is used to drive the compressor (which may also have an associated fan or propeller). The gas flow leaving the turbine is then accelerated to the atmosphere in an exhaust nozzle (Fig. 1A) to provide thrust or propulsion power. Gas turbine or jet engine thrust power is the mass flow momentum increase from engine inlet to exit, multiplied by the flight velocity.

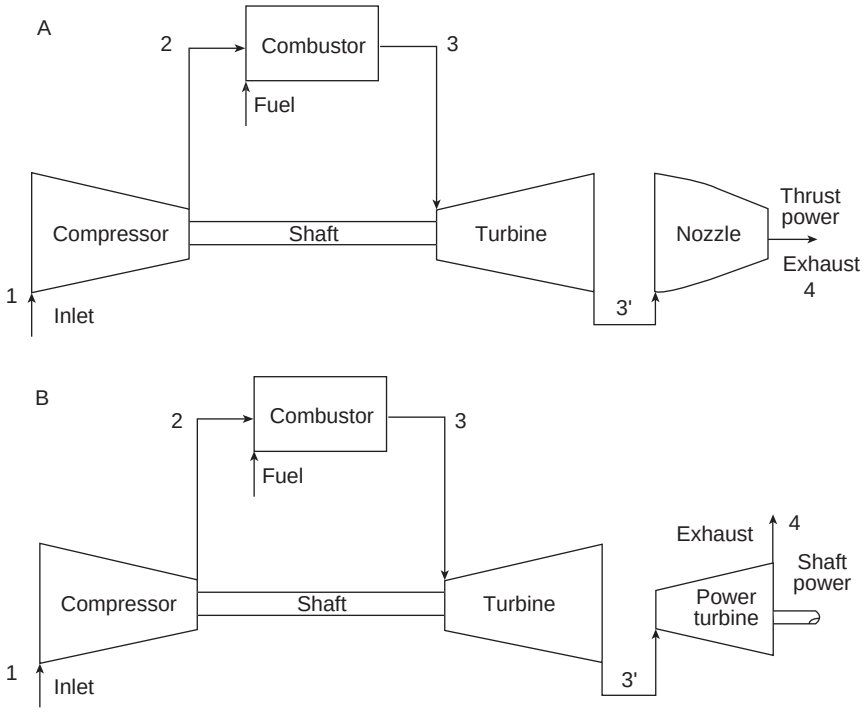


FIGURE 1 Schematic for (A) an aircraft jet engine and (B) a land-based gas turbine.

A typical jet engine is shown in Fig. 2. Such engines can range from approximately 100 lbs of thrust (445 N) to as high as 100,000 lbs of thrust (445,000 N), with dry weights ranging from approximately 30 lbs (134 N) to 20,000 lbs (89,000 N). The jet engine shown in Fig. 2 is a turbofan engine, with a larger diameter compressor-mounted fan. Thrust is generated both by air passing through the fan (bypass air) and by air passing through the gas generator itself. With a large frontal area, the turbofan generates peak thrust at low (takeoff) speeds, making it most suitable for commercial aircraft. A turbojet does not have a fan and generates all of its thrust from air that passes through the gas generator. Turbojets have smaller frontal areas and generate peak thrusts at high speeds, making them most suitable for fighter aircraft.

In nonaviation gas turbines, part of the turbine power is used to drive the compressor. The remainder, the "useful power," is used as output shaft power to turn an energy conversion device (Fig. 1B), such as an electrical generator or a ship's propeller. As shown in Fig. 3, shaft power land-based gas turbines can become very large (as high as 240 MW). The unit in Fig. 3 is called an industrial or frame machine, used mostly for electric power generation. It is constructed for heavy duty, ruggedness, and long life, where weight is not a major factor as it is with a jet engine. Typically, frame machines are designed conservatively and have made use of jet engine technical advances when it has made sense to do so.

Lighter-weight gas turbines derived from jet engines (such as in Fig. 2) and used for nonaviation applications are called aeroderivative gas turbines. Aeroderivatives are used to drive natural gas pipeline compressors, to power ships, and to produce electric power. They are particularly used to provide peaking and intermediate power for electric utility applications. [Peaking power supplements a utility's normal electric power output during high-demand periods (e.g. summer air conditioning in major cities).]

Some of the principle advantages of the gas turbine are as follows:

(1) It is capable of producing large amounts of useful power for a relatively small size and weight.

(2) Since motion of all its major components involves pure rotation (i.e., no reciprocating motion as in a piston engine), its mechanical life is long and the corresponding maintenance cost is relatively low.

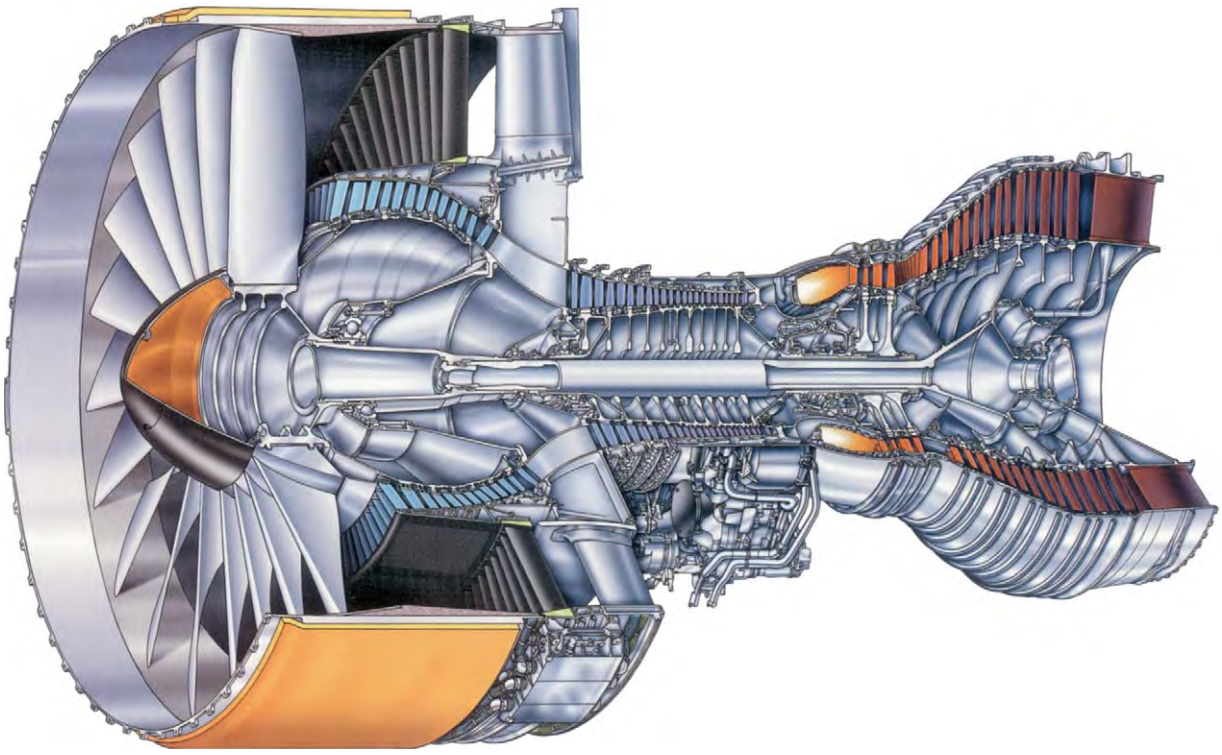


FIGURE 2 A modern jet engine used to power Boeing 777 aircraft. This is a Pratt and Whitney PW4084 turbofan, which can produce 84,000 lbs (374 kN) of thrust (~ 63 MW at 168 m/s). It has a front-mounted fan of 112 in. (2.85 m) in diameter, a flange-to-flange length of 192 in. (4.87 m), and a mass of ~ 15,000 lbs (6804 kg).

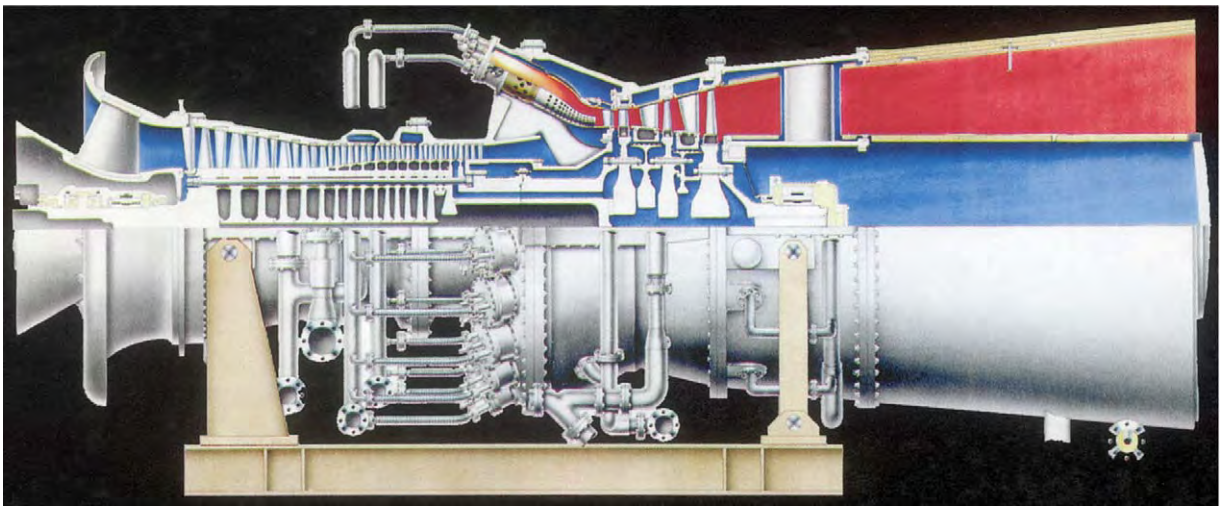


FIGURE 3 A modern land-based gas turbine used for electrical power production and for mechanical drives. This is a General Electric MS7001FA frame machine with a length of ~61 ft (18.6 m) and a mass of ~377,000 pounds (171,000 kg). It produces shaft power of 159 MW (213,140 hp) at 3600 rpm.

(During its early development, the deceiving simplicity of the gas turbine caused problems, until aspects of its fluid mechanics, heat transfer, and combustion were better understood.)

(3) Although the gas turbine must be started by some external means (a small external motor or other source, such as another gas turbine), it can be brought up to full-load (peak output) conditions in minutes, as contrasted to a steam turbine plant, whose startup time is measured in hours.

(4) A wide variety of fuels can be utilized. Natural gas is commonly used in land-based gas turbines, whereas light distillate (kerosene-like) oils power ship and aircraft gas turbines. Diesel oil or specially treated residual oils can also be used, as can combustible gases derived from blast furnaces, refineries, land fills, sewage, and the gasification of solid fuels such as coal, wood chips, and bagasse.

(5) The usual working fluid is atmospheric air. As a basic power supply, the gas turbine requires no coolant (e.g., water).

In the early days of its development, one of the major disadvantages of the gas turbine was its lower efficiency (hence higher fuel usage) when compared to other IC engines and to steam turbine power plants. However, during the past 50 years, continuous engineering development work has pushed the thermal efficiency (18% for the 1939 Neuchatel gas turbine) to approximately 40% for simple-cycle operations and up to 60% for combined-cycle operations (i.e., when combined with a heat-recovery steam-generation unit and a steam turbine). Even more fuel-efficient gas turbines are in the planning stages, with simple-cycle efficiencies predicted as high as 45–47% and combined-cycle machines in the >60% range. These projected values are significantly higher than those for other prime movers, such as steam power plants.

3. GAS TURBINE CYCLES

The ideal gas Brayton cycle (1876, and also proposed by Joule in 1851), shown in graphic form in Fig. 4A as a pressure–volume diagram, is an idealized representation of the properties of a fixed mass of gas (working fluid) as it passes through a gas turbine in operation (Fig. 4B).

A unit mass of ideal gas (e.g., air) is compressed isentropically from point 1 to point 2. This represents the effects of an ideal adiabatic (isentropic) compressor (Figs. 1 and 4B) and any isentropic gas flow deceleration for the case of an aviation gas turbine in flight (Fig. 1A). The ideal work necessary to cause the compression is represented by the area between the pressure axis and the isentropic curve 1–2.

The unit gas mass is then heated at constant pressure from 2 to 3 in Fig. 4 by the exchange of heat input, Q_{23}. This isobaric process is the idealized representation of heat addition caused by the combustion of injected fuel into the combustor in Figs. 1A, 1B, and 4B. The mass flow rate of the fuel is very much lower than that of the working fluid (approximately 1:50), so the combustion products can be neglected as a first approximation.

The unit gas mass is then isentropically expanded (lower pressure and temperature and higher volume) from 3 to 4 in Fig. 4, where $P_4 = P_1$. In Fig. 4B, this represents flow through the turbine (to point 3′) and then flow through the exit nozzle in the case of the jet engine (Fig. 1A) or flow through the power turbine in Fig. 1B (point 4). In Fig. 4, the area between the pressure axis and the isentropic curve 3–3′ represents the ideal work, $W_{33'}$, derived from the turbine. This work must be equal to the ideal compressor work W_{12} (the area bounded by curve 1–2). The ideal "useful work," $W_{3'4}$, in Fig. 4 (the area bounded by the isentropic curve 3′–4) is that which is available to cause output shaft power (Figs. 1B and 4B) or thrust power (Fig. 1A).

The Brayton cycle is completed in Fig. 4 by a constant pressure process in which the volume of the unit gas mass is decreased (temperature decrease) as heat Q_{41} is rejected. Most gas turbines operate in an open cycle mode, in which, for instance, air is taken in from the atmosphere (point 1 in Figs. 1 and 4) and discharged back into the atmosphere (point 4), with exiting working fluid mixing to reject Q_{41}. In a closed-cycle gas turbine facility, the working fluid is continuously recycled by ducting the exit flow (point 4) through a heat exchanger (shown schematically in Fig. 5) to reject heat Q_{41} at (ideally) constant pressure and back to the compressor inlet (point 1).

Because of its confined, fixed-mass working fluid, the closed-cycle gas turbine is not an internal combustion engine. The combustor (shown in Fig. 1) is replaced with an input heat exchanger, supplied by an external source of heat. Closed-cycle designs were developed in the 1960s for providing power to space satellites on deep space missions, using a radioisotope heat source to heat helium gas

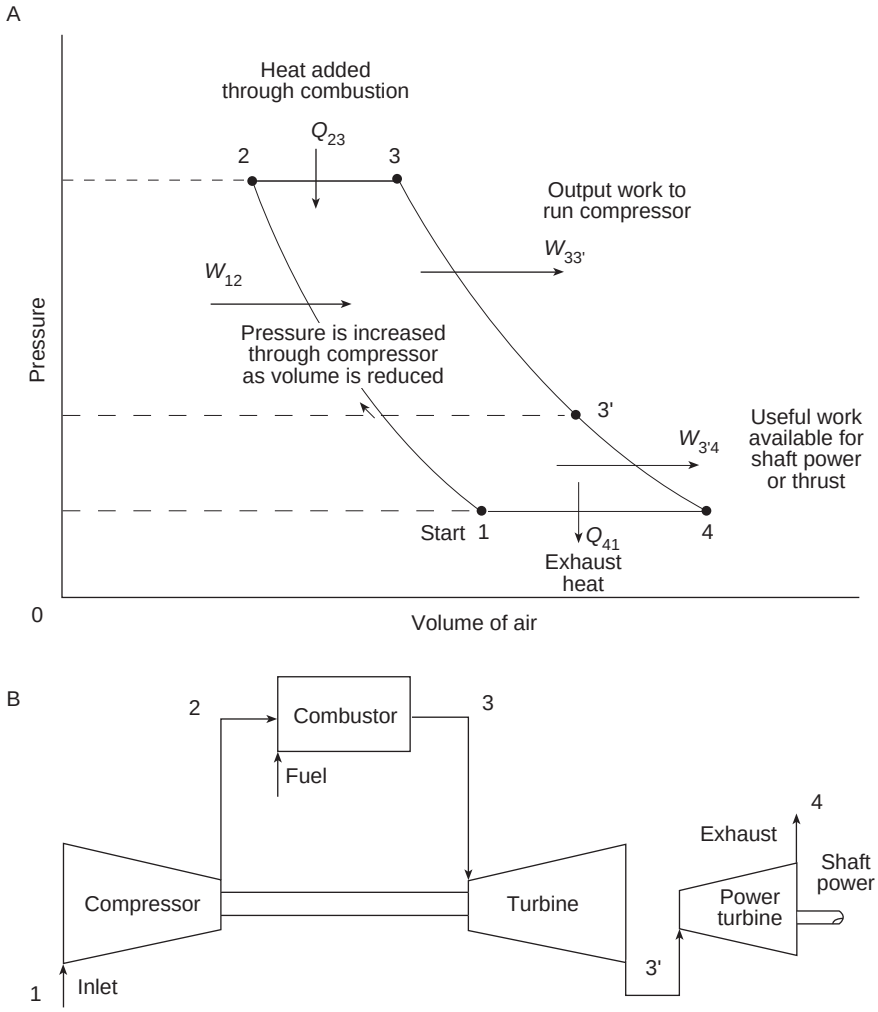


FIGURE 4 (A) Brayton cycle pressure-volume diagram for a unit mass of working fluid (e.g., air), showing work (W) and heat (Q) inputs and outputs. (B) Gas turbine schematic showing relative points from the diagram for the Brayton cycle.

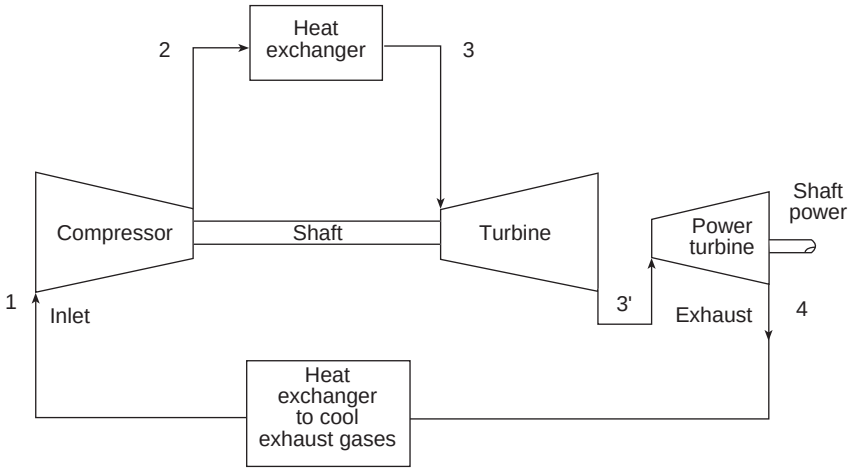


FIGURE 5 Closed-cycle system. Heat exchangers interact with external heat reservoirs.

as the gas turbine working fluid and a space radiator for cycle heat rejection. Development work to use a nuclear reactor (e.g., a pebble bed reactor) to heat helium gas in a closed-cycle gas turbine to be used for electric utility power generation in the 100 MW range is in progress.

The ideal Brayton cycle thermal efficiency, η_B, can be shown to be

$$\eta_B = 1 - \frac{T_4}{T_3} = 1 - \frac{1}{r^{(k-1)/k}}, \tag{1}$$

where the absolute temperatures T_3 and T_4 correspond to points 3 and 4 in Fig. 4, $k = c_p/c_v$ is the ratio of specific heats for the ideal gas, and r is the compressor pressure ratio,

$$r = \frac{P_2}{P_1}. \tag{2}$$

Thus, the Brayton-cycle thermal efficiency increases with both the turbine inlet temperature, T_3, and compressor pressure ratio, r. Modern gas turbines have pressure ratios as high as 30:1 and turbine inlet temperatures approaching 3000°F (1649°C).

The effect of real (nonisotropic) compressor and turbine performance can be easily seen in the expression for the net useful work $W_{net} = W_{3'4}$

$$W_{net} = \eta_T c_p T_3 \left[1 - \frac{1}{r^{(k-1)/k}}\right] - \frac{c_p T_1}{\eta_C}[r^{(k-1)/k} - 1], \tag{3}$$

where c_p is the specific heat at constant pressure, and η_T and η_C are the thermal (adiabatic) efficiencies of the turbine and compressor, respectively. This expression shows that for large W_{net} one must have high values for η_T, η_C, r, and T_3. For modern gas turbines, η_T can be as high as 0.92–0.94 and η_C can reach 0.88.

A gas turbine that is configured and operated to closely follow the ideal Brayton cycle (Figs. 4 and 5) is called a simple-cycle gas turbine. Most aircraft gas turbines operate in a simple-cycle configuration since attention must be paid to engine weight and frontal area. However, in land or marine applications, additional equipment can be added to the simple-cycle gas turbine, leading to increases in thermal efficiency and/or the net work output of a unit. Three such modifications are regeneration, intercooling, and reheating.

Regeneration involves the installation of a heat exchanger (shown in Fig. 6A) through which the turbine exhaust gases (point 4 in Fig. 6A) pass. The compressor exit flow (point 2 in Fig. 6A) is then heated in the exhaust gas heat exchanger (regenerator, and also called a recuperator) before the flow enters the combustor. If the regenerator is well designed—that is, if the heat exchanger effectiveness is high and the pressure drops are small—the cycle efficiency will be increased over the simple-cycle value. However, its relatively high cost must also be taken into account. Regenerated gas turbines can increase efficiency by 5–6 percentage points and are even more beneficial for part-load operation.

Intercooling also involves the use of a heat exchanger. An intercooler, as shown in Fig. 6B, is a heat exchanger that cools compressor gas during the compression process. For instance, if the compressor consists of a high-pressure unit and a low-pressure unit, the intercooler can be mounted between them (Fig. 6B) to cool the flow and decrease the work necessary for compression in the high-pressure

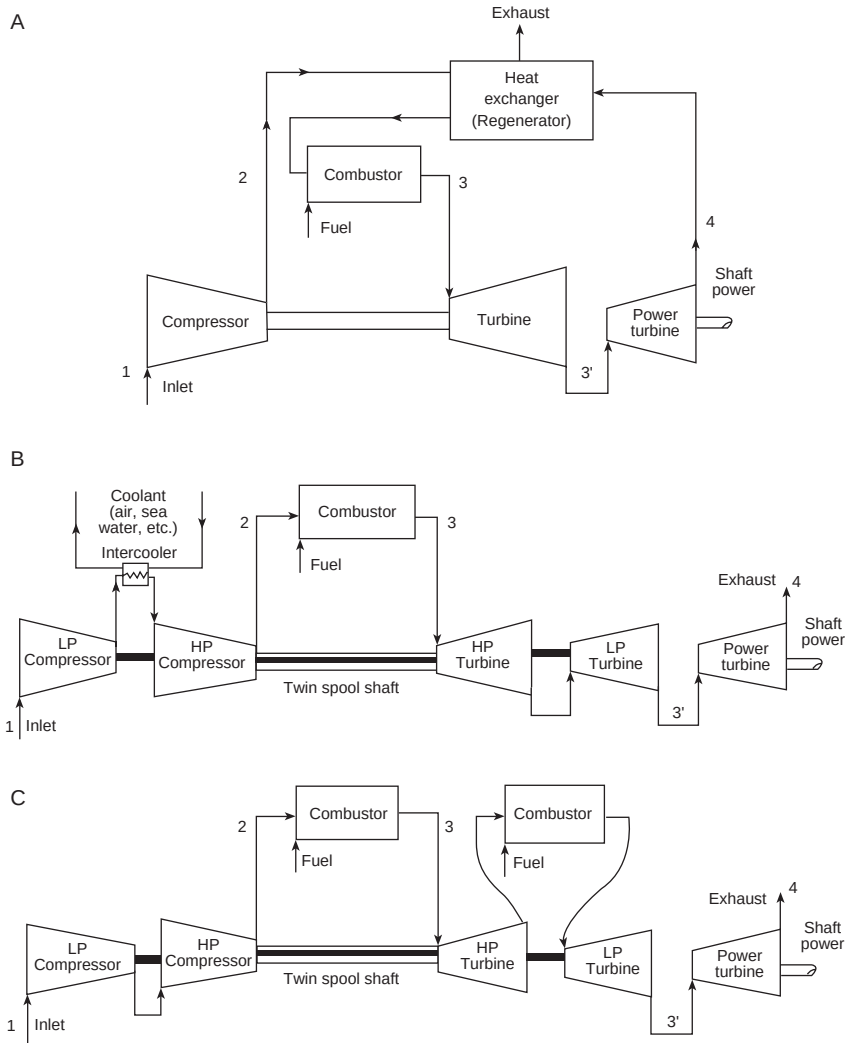


FIGURE 6 Modifications available for the simple Brayton cycle.

compressor. The cooling fluid could be atmospheric air or water (e.g., sea water in the case of a marine gas turbine). The net work output of a given gas turbine is increased with a well-designed intercooler. Studies of gas turbines equipped with extensive turbine convective and film cooling show that intercooling can also allow increases in thermal efficiency by providing cooling fluid at lower temperatures, thereby allowing increased turbine inlet temperatures [T_3 in Eq. (1)].

Reheating, as shown in Fig. 6C, occurs in the turbine and increases turbine work without changing compressor work or exceeding the material temperature limits in the turbine. If a gas turbine has a high-pressure turbine and a low-pressure turbine, a reheater (as shown in Fig. 6C, usually another combustor) can be used to "reheat" the flow between the two turbines. Reheat in a jet engine is accomplished by adding an afterburner at the turbine exhaust, thereby increasing thrust, with a greatly increased fuel consumption rate.

A combined-cycle gas turbine (CCGT) power plant, shown schematically in Fig. 7, is essentially an electrical power plant in which a gas turbine provides useful work (W_B) to drive an electrical generator. The gas turbine exhaust energy (Q_{BR}) is then used to produce steam in a heat exchanger (called a heat recovery steam generator) to supply a steam turbine whose useful work output (W_R) provides the means to generate more electricity. [If the steam is used for heat (e.g., heating buildings), the unit would be called a cogeneration plant.] Figure 7 is a simplified thermodynamic representation of a CCGT and shows it to be two heat engines (Brayton and Rankine) coupled in series. The "upper" engine is the gas turbine (represented as a Brayton-cycle heat

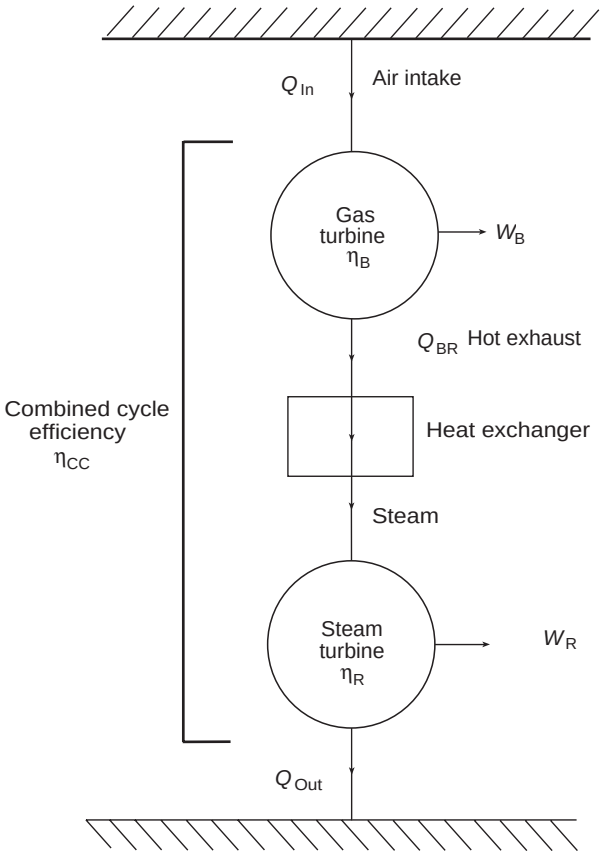


FIGURE 7 Schematic of a combined-cycle gas turbine plant. Overall efficiency (η_{cc}) is a combination of the efficiency of the Brayton gas turbine cycle (η_B) and the Rankine steam turbine cycle (η_R). Total combined-cycle efficiency $\eta_{cc} = \eta_B + \eta_R - \eta_B\eta_R$.

engine), whose energy input is Q_{In}. It rejects heat (Q_{BR}) as the input energy to the "lower" engine (the steam turbine, represented as a Rankine-cycle heat engine). The Rankine heat engine then rejects unavailable energy (heat) as Q_{Out} by means of a steam condenser. The combined thermal efficiency (η_{CC}) can be derived fairly simply and is given as

$$\eta_{CC} = \eta_B + \eta_R - \eta_B\eta_R, \tag{4}$$

where $\eta_B = W_B/Q_{In}$ and $\eta_R = W_R/Q_{BR}$ are the thermal efficiencies of the Brayton and Rankine cycles, respectively. Taking $\eta_B = 40\%$ (a good value for modern gas turbines) and $\eta_R = 30\%$ (a reasonable value under typical CCGT conditions), the sum minus the product in Eq. (4) yields $\eta_{CC} = 58\%$, a value of combined-cycle efficiency greater than either of the individual efficiencies.

This remarkable equation gives insight into why CCGTs are so efficient. The calculated value of η_{CC} represents an upper bound on an actual CCGT, since Eq. (4) does not account for efficiencies associated with transferring Q_{BR} (duct losses, irreversibilities due to heat transfer, etc.). Actual efficiency values as high as 52 to 57% have been attained with CCGT units during the past few years, thus coming close to values given by Eq. (4).

4. GAS TURBINE COMPONENTS

A greater understanding of the gas turbine and its operation can be gained by considering its three major components (Figs. 1, 2, and 3): the compressor, the combustor, and the turbine. The features and characteristics of these components are briefly examined here.

4.1 Compressors and Turbines

The turbine and compressor components are mated by a shaft, since the former powers the latter. A single-shaft gas turbine has but one shaft connecting the compressor and turbine components. A twin-spool gas turbine has two concentric shafts, a longer one connecting a low-pressure compressor to a low-pressure turbine (the low spool), which rotates inside a shorter, larger-diameter shaft (e.g., see Figs. 6B and 6C). The latter connects the high-pressure turbine with the higher pressure compressor (the high spool), which rotates at higher speeds than the low spool. A triple-spool engine would have a third, intermediate-pressure compressor-turbine spool.

Gas turbine compressors can be centrifugal, axial, or a combination of both. Centrifugal compressors (radial outflow) are robust, generally cost less, and are limited to pressure ratio of 6 or 7 to 1. They are found in early gas turbines or in modern, smaller gas turbines.

The more efficient, higher capacity axial flow compressor is used on most gas turbines (e.g., Figs. 2 and 3). An axial compressor is made up of a number of stages, with each stage consisting of a row of rotating blades (airfoils) and a row of stationary blades (called stators) configured so the gas flow is compressed (adverse or unfavorable pressure gradient) as it passes through each stage. It has been said that compressor operation can founder on a metaphoric rock, and that rock is called stall. Care must be taken in compressor operation and design to avoid the conditions that lead to blade stall, or flow separation. The collective behavior of blade separation can lead to compressor stall or surge, which manifests itself as an instability of gas flow through the entire gas turbine.

Turbines are generally easier to design and operate than compressors, since the flow is expanding in an overall favorable pressure gradient. Axial flow turbines (Figs. 2 and 3) will require fewer stages than an axial compressor for the same pressure change magnitude. There are some smaller gas turbines that utilize centrifugal turbines (radial inflow) but most utilize axial turbines (e.g., Figs. 2 and 3).

Turbine design and manufacture are complicated by the need to ensure turbine component life in the hot gas flow. The problem of ensuring durability is especially critical in the first turbine stage, where temperatures are highest. Special materials and elaborate cooling schemes must be used to allow metal alloy turbine airfoils that melt at 1800 to 1900°F (982 to 1038°C) to survive in gas flows with temperatures as high as $T_3 = 3000$°F (1649°C).

4.2 Combustors

High-pressure air leaves the compressor (Figs. 1, 2, and 3) and enters the gas turbine combustor, where, in a very short space and at high flow rates, heat is added to it by the combustion of fuel (gas or liquid). Modern gas turbine combustors are of three types: cans (individual cylindrical chambers arranged circumferentially), annular (consisting of a cylindrical single inner and single outer shroud), and can–annular (a fusion of the can and annular geometrics). Because of the high temperatures in the combustors, combustor parts are regularly inspected and will require repair and/or replacement during the lifetime of the gas turbine.

Compared to some other energy converters [such as the Diesel or Otto (gasoline) engines listed in Table I], gas turbines produce very low levels of combustion pollution, such as unburned fuel, carbon monoxide, oxides of nitrogen (NO_x), and smoke. New electric power plants using gas turbines drastically reduce area pollution when they replace older steam- or diesel-powered plants. For instance, due to new strict environmental restrictions being placed on Alaskan cruise vessels, new cruise ships are installing gas turbines, as the only means available to comply with the emission levels that have been put in place.

Gas turbine combustion systems must have very low pressure losses and be carefully designed to provide a stable flow, free of pressure pulsations and characterized by uniform exit temperatures (flow "hot spots" can damage turbine gas path parts, such as stationary vanes). New electric power gas turbines have experienced serious pressure oscillations with new "lean premixed" combustion systems designed to meet lower NO_x emission regulations. System pressure oscillations brought about by combustion mechanisms that not yet fully understood have been given descriptive names ranging from humming to hooting, screech, chub, buzz, and rumble. Manufacturers have taken a number of paths to deal with this problem (which, if left unattended, can shake a machine apart), such as the use of catalysis, resonators, fuel schedule timing, active controls, or simple cutting back on power output levels.

5. GAS TURBINE USAGE

Although they are used in an ever-increasing variety of applications, gas turbines are used extensively in two areas: aviation and electric power generation.

It is difficult to remember a time when the aviation gas turbine—the jet engine—was not part of aviation. Before jet engines, an aviation piston engine manufacturer could expect to sell 20–30 times the original cost of piston engines, in aftermarket parts. With the advent of the jet engine, this aftermarket value dropped to 3–5 times the original cost (an important reduction that made air travel affordable and reliable, and airlines profitable). Technology and airline/military demands made on jet engine manufacturers have resulted in even longer lasting engine components, dropping the aftermarket value even lower.

A well-managed airline will regularly try to keep a jet-powered plane in the air as much as 18 h per day, 365 days a year, to produce revenue. The airline expects the engines, if well-maintained, to remain in service and on the wing for up to 15,000–20,000 h of operation, depending on the number of takeoffs and landings experienced by the plane. After this period (or time between overhauls), the jet engine will be taken off of the wing and overhauled and usually hot-section (combustor and turbine) parts will be replaced as needed. (An experienced air traveler can generally believe the numbers presented here, by reflecting how few (if any) times he or she has experienced an in-flight shutdown of a jet engine while flying.)

The life of electric power gas turbines are much longer that those of their aviation counterparts, because they run at constant speeds and are not usually subjected to the wide range of power settings experienced by jet engines. Modern power plant operators expect trouble-free units to run as long as 30,000–50,000 h, depending on load conditions, maintenance procedures, and the level of turbine

inlet temperatures. Overhauls usually involve replacement or refurbishment of hot-section parts in the combustor and turbine sections. Normally, these will take place approximately three or more times during the life of a gas turbine unit.

In summary, the power spectrum of the gas turbine units produced is larger than that of most other prime movers, spanning a range that varies by a factor of 10,000. New microturbines [weighing little more than 100 pounds (45 kg)] are being produced in the range of ~25 kW to provide electrical power and heat for fast-food restaurants. Base-load electric power gas turbines as large as 250 MW [weighing as much as 330 tons (300,000 kg)] are being installed to drive electric generators and to supply heat for steam turbines in combined-cycle operations, with resultant thermal efficiencies approaching 60%. In an evolving role as the engineer's most versatile energy converter, gas turbines produce thrust power [as much as 120,000 pounds thrust (534,000 N) for the largest commercial jet engines] to propel most of the world's military and commercial aircraft. In closing, it can be noted that at the time this article is being written, the annual worldwide value of production of gas turbines for shaft power applications (electric power, ship propulsion, natural gas pipeline compression) is approximately equal to that for aviation thrust power applications.

SEE ALSO THE FOLLOWING ARTICLES

Aircraft and Energy Use • Life Cycle Analysis of Power Generation Systems • Turbines, Steam

Further Reading

Bathie, W. W. (1996). "Fundamentals of Gas Turbines." 2nd Ed. Wiley, New York.

Cohen, H., Rogers, G. F. C., and Saravanamuttoo, H. I. H. (1987). "Gas Turbine Theory." 3rd ed. Longman Scientific and Technical, Essex, England.

Horlock, J. H. (1993). "Combined Power Plants." Pergamon, Oxford, England.

Huang, F. F. (1988). "Engineering Thermodynamics." 2nd ed. Macmillan, New York.

Kerrebrock, J. L. (1980). "Aircraft Engines and Gas Turbines." MIT Press, Cambridge, MA.

Transactions of the ASME, published quarterly by the American Society of Mechanical Engineers, New York. Two ASME gas turbine journals are *Journal of Turbomachinery* and *Journal of Engineering for Gas Turbines and Power.*

Turbines, Steam

ENRICO SCIUBBA
University of Roma 1 "La Sapienza"
Rome, Italy

1. **Some Basic Concepts**
2. **Energy Transfer in a Turbine**
3. **Outline of a General Procedure for Preliminary Steam Turbine Design**
4. **Off-Design Considerations**
5. **Conclusions**

Glossary

back-pressure turbine A steam turbine that discharges the steam (usually at $p > 1$ bar) into a pressurized system.
condensing turbine A steam turbine at whose exit the steam is condensed in an air- or water-cooled condenser; the pressure in the condenser (and at the turbine exit) is usually less than 1 bar.
degree of reaction (R_ρ) Ratio of the static enthalpy drop in the rotor to the total enthalpy drop in the stage.
dimensional analysis A mathematical theory that derives dimensionless groups by proper combination of dimensional parameters, with the goal of extracting from these groups further and more general physical information.
extraction turbine A steam turbine in which a portion of the working fluid is tapped between stages and used for purposes other than generating mechanical power.
flow coefficient (φ) Ratio of the local average meridional absolute velocity component to the local peripheral velocity.
head coefficient (ψ) Ratio of the local average tangential absolute velocity component to the local peripheral velocity.
impulse stage A stage with zero degrees of reaction.
isentropic A thermodynamic transformation in which the initial and final points have the same specific entropy.
reaction stage A stage with a degree of reaction larger than zero.
secondary flows Conventionally defined as all fluid motions extraneous to the ideal "mean" flow path.
similarity The property of two "objects" (physical dimensions, components, or flow patterns) that have the same ratio of a certain number of homologous characteristic features.
solidity (σ) At a certain radius, the ratio of the blade chord to the interblade spacing.
specific diameter (d_s) A dimensionless group linking one characteristic dimension, volumetric flow rate, and stage enthalpy drop.
specific speed (n_s) A dimensionless group linking angular velocity, volumetric flow rate, and stage enthalpy drop.
supercritical A steam turbine in which the feed steam is above the critical conditions.
swirl The tangential component of the absolute or relative velocity.

Turbines are machines that convert the energy carried by a continuous flow of fluid into mechanical shaft rotational power. Their essential feature is that they do not contain parts in alternating motion; therefore, the flow of both mass and energy through them, except for secondary transitory phenomena, may be considered steady. This is in contrast to piston-based machinery, where the energy exchange is fundamentally linked to the alternating motion of the piston. Traditionally, and in spite of the fact that the basic physics of the energy exchange is not much different, hydraulic turbines, where the working fluid is incompressible, are treated separately from gas and steam turbines, where compressibility phenomena play a major role in the fluid/machine coupling. In compressible flow turbines, it makes little difference, from a theoretical point of view, whether the working fluid is a "gas" that follows an ideal gas law or a "steam" described by one of the variants of van der Waals's equation of state. Physically, the only difference is that gases are farther away from their two-phase region, so that the variation of their specific heat with temperature is less complicated. But in practice, there are enough differences in the

 © 2004 Elsevier Inc. All rights reserved.

design procedures to grant separate treatment of gas and steam turbines.

1. SOME BASIC CONCEPTS

1.1 The Flow through a Turbine

Figure 1 shows a schematic cross section of an axial (Fig. 1A) stage and a radial stage (Fig. 1B) of a modern steam turbine. High-pressure steam enters the machine at station "in" (throttle valve exit section), passes through the inlet channel, and reaches the stator inlet (station "0"). The stator accelerates the flow by expanding it by a fraction $\beta_{st}\,(1 \geq \beta_{st} > 0)$ of the total stage pressure drop and introduces a tangential velocity component (inlet swirl). The steam jets exiting the stator (station "1") impinge on the rotoric blades that deviate the flow so as to decrease this inlet swirl and, thus, absorb rotational momentum. At rotor exit (station 2), the steam is at a lower pressure and energy level than at station 0 and is discharged through the outlet scroll or diffuser channel. Station "out" is the outlet section of the stage. The purpose of the inlet channel is to guide the flow from the turbine inlet port to the statoric vanes with the minimum possible loss in total pressure. If the first stage after the admission throttle is axial, the inlet is usually shaped as a pseudo-conical transition piece between the inlet duct and the (larger in diameter) stator. In this case, no flow deviation is imposed and only a modest acceleration is attained. For intermediate axial stages, the inlet is missing and the flow from the exit of the previous stage flows directly onto the stator of the following stage. Radial turbines can be centripetal (flow inward) or centrifugal (flow outward). Because the only significant centrifugal application is the Ljüngstrøm turbine, shown in Fig. 2 and discussed at the end of this section, "radial turbine"

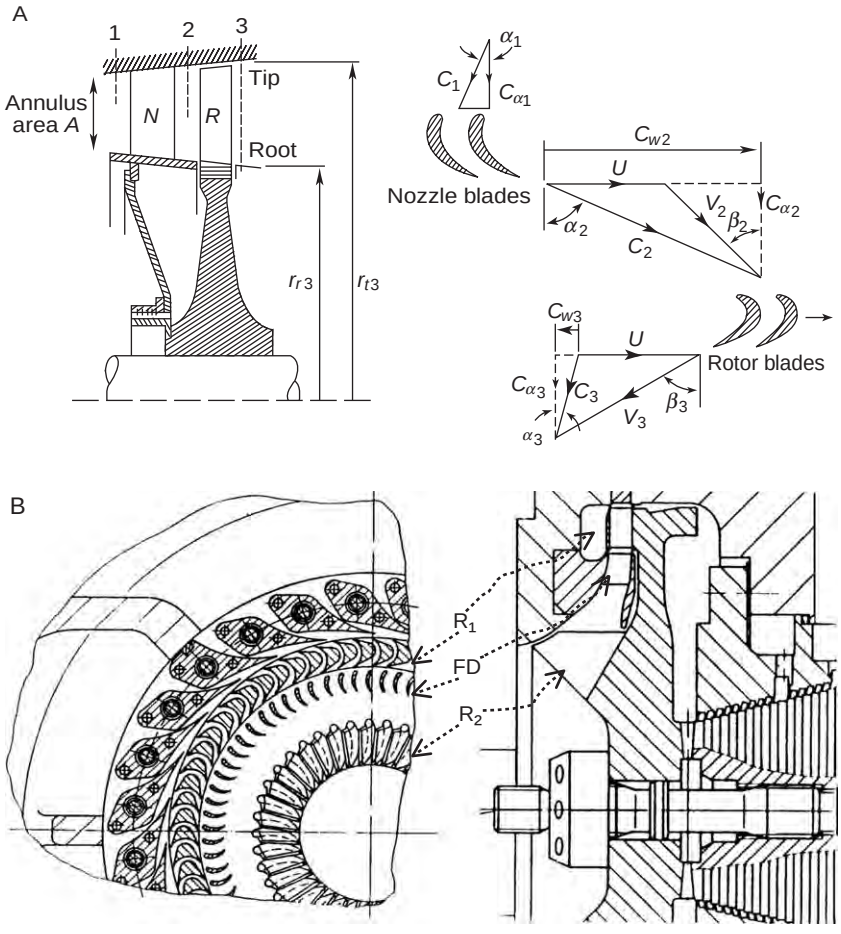


FIGURE 1 (A) Axial turbine stage. (B) Radial turbine stage. Notice the two-stage rotor (R_1 and R_2) and the flow deviator (FD).

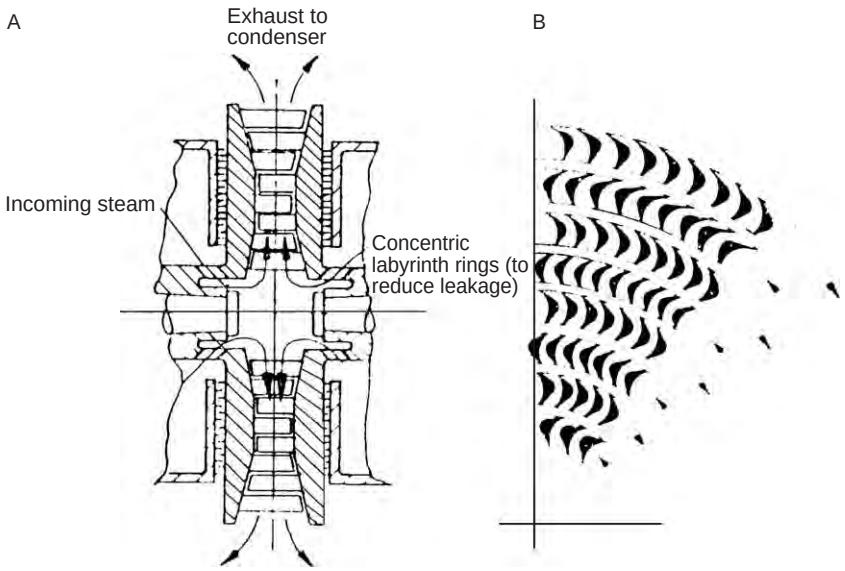


FIGURE 2 The Ljüngstrøm radial turbine. (A) Meridional section through turbine. (B) Blading arrangement and directions of rotation.

implies centripetal in the remainder of this article. Radial turbines are usually single-staged with a tangential inlet. The flow is admitted into the statoric vanes through a scroll of variable cross section that accommodates the decreasing mass flow rate in the tangential direction.

The purpose of the stator is as follows:

- To induce a tangential velocity component (swirl) in the flow
- To transform a portion β_{st} of the stage expansion ratio $\beta = p_{in}/p_{out}$ into kinetic energy
- To attain the above with the minimum possible irreversible dissipation
- To serve as an auxiliary structural support for the casing

Stators consist of a large number of rather thick, curved, airfoil-like blades that turn the oncoming flow, whose main flow component is parallel to the turbine axis, in the tangential direction. Axial and radial stators differ in shape, as shown in Fig. 3. If large variations of the steam mass flow rate are foreseen, variable geometry stators are often employed where the individual vanes may rotate about a radial (in axial turbines [Fig. 3C]) or meridional (in radial turbines [Fig. 3D]) axis.

The purpose of the rotor is as follows:

- To extract energy from the main flow by recovering its tangential velocity component
- To transfer this energy to a rotating shaft
- To attain the above with the minimum possible irreversible dissipation

Rotors consist of a large number of curved, airfoil-like blades that turn the oncoming flow so as to cause a decrease in its angular momentum, thereby generating a tangential force on each blade. The blades are solidly attached to the turbine shaft by a disk whose only purpose is to transfer the resulting torque. Axial blades resemble twisted airfoils, with their chord decreasing from hub to tip, and are attached to the disk on their hub side. Radial blades look more like paddles attached to the disk on their hub side and strongly twisted in three dimensions. Axial blades can be forged onto the disk, but more commonly they are individually inserted into proper circumferential grooves machined on the external surface of the disk and are firmly held in position by spacers and stiffeners. For smaller rotors, they may be machined together with the disk from a single, cylindrical, forged "raw" disk. Radial blades are most commonly an integral part of the rotor, being forged or machined from one initial pseudo-conical billet.

It is often convenient, for design reasons, to let the flow exit the stage with substantial kinetic energy. This lowers the stage exit pressure and increases the stage expansion ratio β_s. In addition, it allows handling the same mass flow rate with a smaller flow passage area (a smaller diameter implies a less costly machine). The kinetic portion of the total enthalpy can be partialy recovered in the diffuser, which is a passage of increasing flow area (visible in

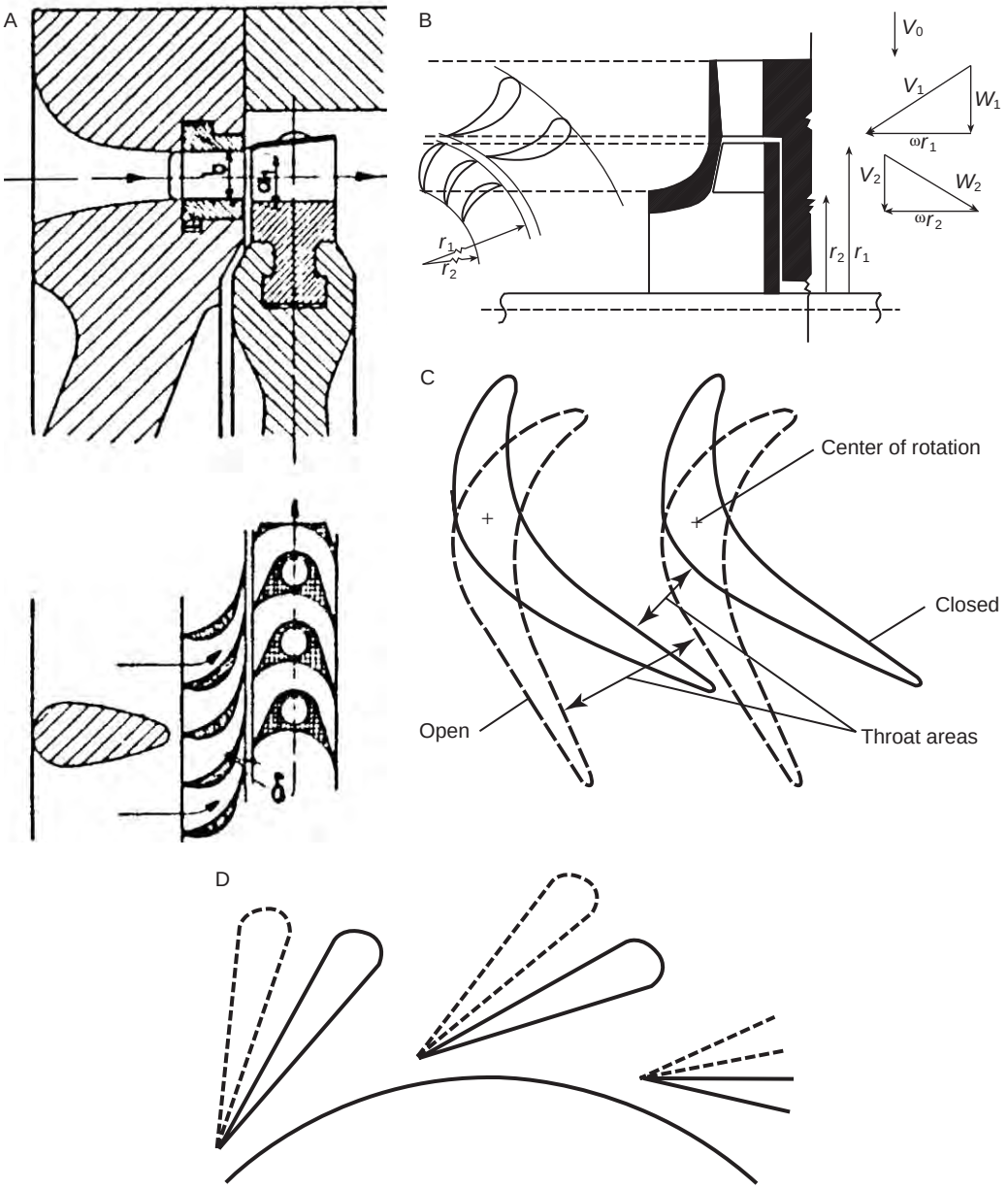


FIGURE 3 (A) Axial stage geometry. (B) Radial stage geometry. (C) Variable geometry axial stator. (D) Variable geometry radial stator.

Figs 1A and 1B). If the area increase is obtained by increasing the average radius, there is the additional advantage that the exit swirl is recovered as well.

Ljüngstrøm turbines (Fig. 2), are atypical in more than one sense. The steam enters the machine radially from slots in the casing or in the shaft and proceeds radially outward. The wing-like blades are inserted in two counterrotating disks, with the arrangement being such that each blade row rotates between two rows attached to the opposite disk. Each blade row is a rotor in the sense that it extracts energy from the flow and provides it to the shaft, and is a stator in the sense that it introduces a (relative) tangential velocity component for the next blade row. Ljüngstrøm turbines are used when axial size constraints are very stringent because their power density is higher than that of other types of turbines. However, their use is restricted to lower mass flow rates due to blade length limitations.

1.1.1 Partial Admission

If the throttle pressure is very high, the incoming steam into the first stage has a very high density;

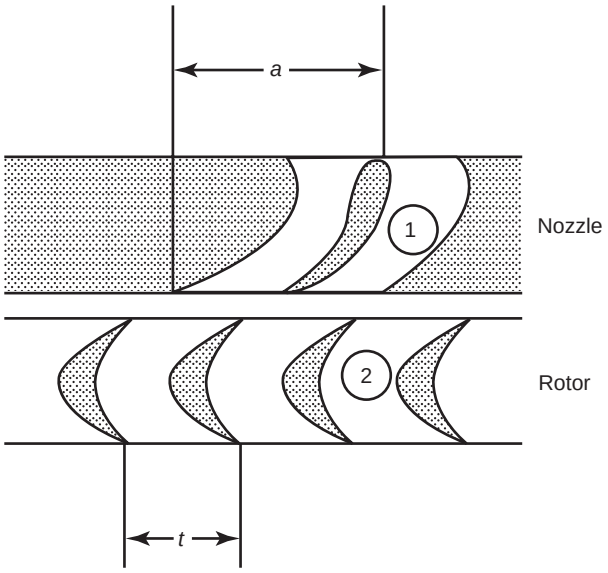


FIGURE 4 Partial admission.

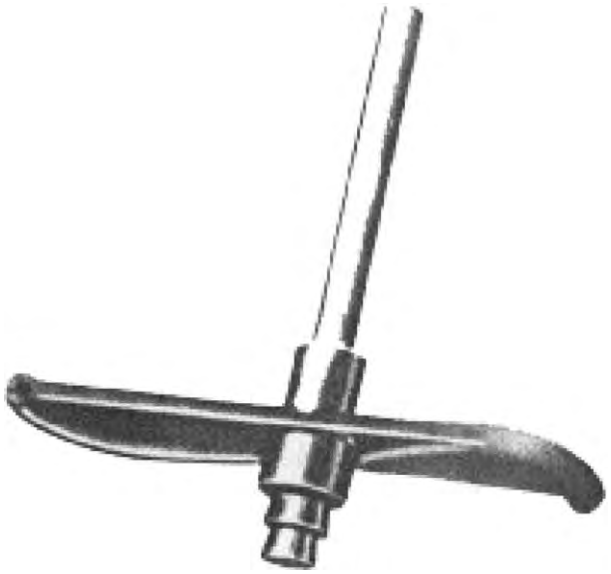


FIGURE 5 Avery's 1831 reaction turbine.

therefore, even for large flow rates, the required passage area may be relatively small. Because the external diameter (both in axial and radial stages) is selected on the basis of the best attainable efficiency and provides an optimal d_s, this may lead to extremely low blade heights, which increase the end wall losses and decrease the practical stage efficiency. To uncouple (at least to a certain extent) the blade height from the external diameter, the statoric nozzles of the first stage of class 4 and 5 turbines are so constructed as to limit the steam injection angle to a value $a < 360^\circ$ (Fig. 4). Thus, only the statoric channels within this angle a will be active, and for a given external diameter, the blade height can be increased of $1/a$ without efficiency penalties. This configuration is called partial admission and can be introduced only in pure impulse stages, where the steam in the rotoric passages is not subject to a streamwise Δp (otherwise, massive leakage into the inactive adjacent channels would destroy the regularity of the flow).

1.2 Brief Historical Background

The word "turbine" was first proposed in its modern usage by the French Claude Burdin in 1820, with reference to hydraulic machines. Although the first steam turbines in commercial service were those installed in the northern United States by William Avery in 1831 to power some sawmills (Fig. 5), the idea of exploiting the "motive power" of live steam is much older. In the first century BC, Hero of Alexandria, in his treaty "De Re Pneumatica" (literally "On Fluid Flow"), described what in modern terms would be defined as a "single-stage, pure reaction turbine" (Fig. 6). Hero, who seemed to imply that the "turbine" had been actually built prior to his description of it, provided no indication about practical uses for this device, although other sources indicate that similar machines were used to power the opening mechanisms of the heavy bronze doors of some Egyptian or Babylonian temples. (This indicates that the turbine provided ~ 0.5 kW with a steam mass flow rate of ~ 0.01 kg/s, which seems unfeasible for those times because it implies a nozzle diameter of ~ 15 cm.) Be that as it may, the technological level of those times clearly had no use for such machines, and for centuries hydraulic and wind turbines were the only turbomachines employed in power generation. Much later, there were isolated attempts to convert the motive energy of steam (now called stagnation exergy) to mechanical shaft power. Giovanni Branca's machine (in Italy in 1629 [Fig. 7A]) was actually inspired by an earlier illustration (from 1551) contained in a sort of technological almanac compiled by the Arab historian Taqi-al-Din. John Dumbell's proposal (in England in 1808 [Fig. 7B]), although in essence based on the same principle, was technologically much more advanced. However, neither machine ever produced any net power. In 1848, the Scot Robert Wilson patented a series of steam turbines, one of which

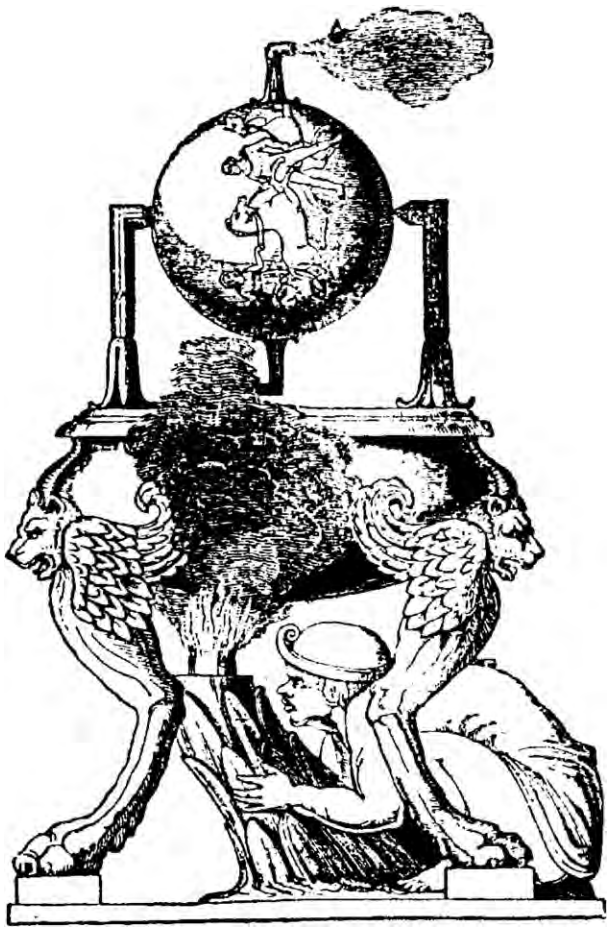


FIGURE 6 Hero's "steam turbine" (probably conceived before the first century BC) in a 16th-century rendering.

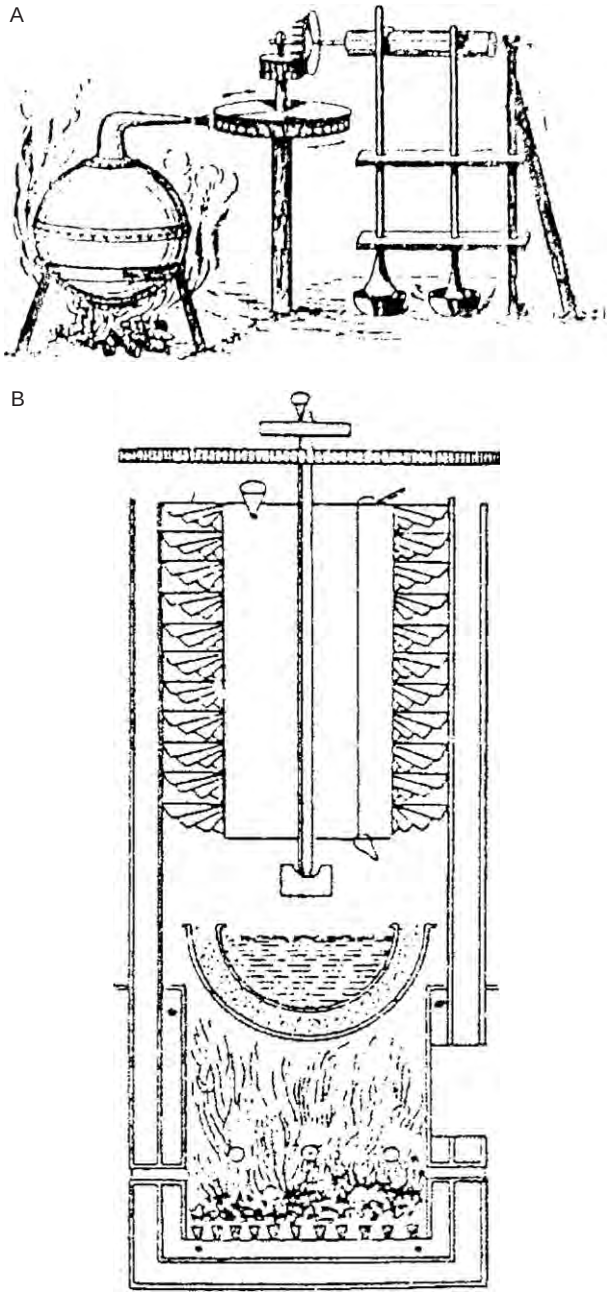


FIGURE 7 (A) Branca's 1629 steam turbine-powered mill. (B) Dumbell's 1808 steam turbine.

endowed of a radial inlet. In 1875, Osborne Reynolds designed and later built a multistage steam turbine (Fig. 8), and between 1880 and 1890, the Briton Charles Parsons and the Swede Carl-Gustav de Laval introduced the first multistage reaction turbine (Fig. 9 [this type of arrangement is still today sometimes called "Parsons"]) and the impulse turbine (Fig. 10), respectively. Between 1890 and 1897, de Laval installed some of his turbines on steam-powered ships. During the same years, the French Charles Rateau introduced the first multistage impulse turbines, and the American Charles Curtis introduced the first double-bladed impulse rotors. In 1891, Parsons introduced the condenser, a sealed tank bottoming the turbine where the steam condensed on banks of water tubes, effectively lowering the discharge pressure of the turbine and allowing for a remarkable increase of the specific work output. In 1910, the Swede Berger Ljüngstrøm presented his axial inlet centrifugal turbine. The modern development of steam turbines began after the second decade of the 20th century when large industrial groups such as General Electric, Allis–Chalmers, Westinghouse, and Brown–Boveri built turbine-powered electricity generation plants. Currently, the vast majority of fossil-fueled thermoelectrical conversion plants employ steam turbines as mechanical drivers.

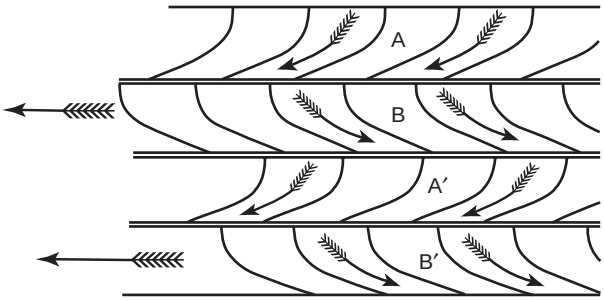


FIGURE 8 Original 1875 sketch by Reynolds of his own multistage turbine.

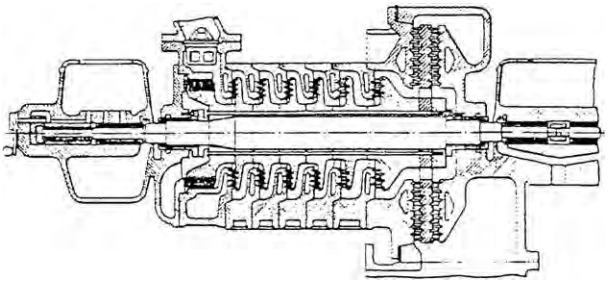


FIGURE 9 Parsons's (circa ~ 1885) multistage reaction turbine.

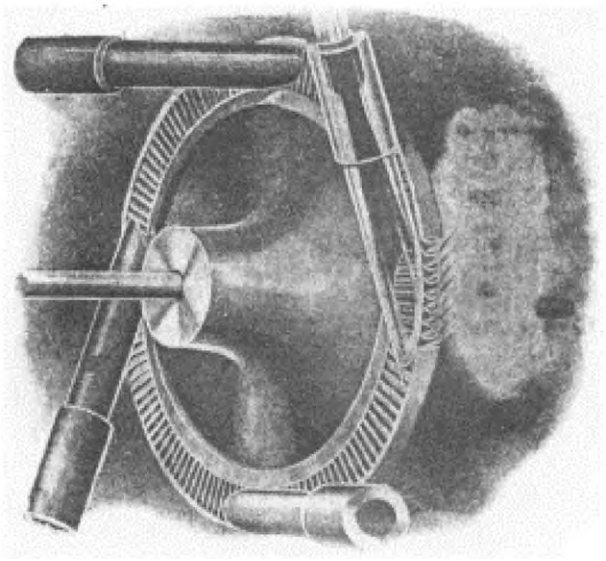


FIGURE 10 De Laval's 1890 impulse turbine.

1.3 Steam Conditions and Fundamental Classification

Modern turbines are designed for a variety of applications and use steam at different inlet conditions, as shown in Tables I and II. Figure 11 reproduces the so-called Mollier chart, where the thermodynamic properties of steam are plotted on an entropy/temperature diagram. Higher inlet steam pressures and temperatures would generally increase the efficiency of the cycle in which the turbine is inserted, but for technological and economical reasons, it is unlikely that during the next 10 to 15 years or so an increase of more than 50 K over the current top inlet temperatures of 550 to 600 K will occur. This article makes reference essentially to industrial and power generation applications. Typical values for admission steam conditions are shown in Table I together with approximate indications of the power range.

Steam turbines may be classified according to several criteria. All of these criteria have some practical or theoretical relevance, but none is universally accepted. Table II summarizes the seven most common classification criteria.

1.3.1 Supercritical Turbines

A turbine stage is said to be "supercritical" when the steam conditions at its inlet are above the critical point. Although the higher temperatures do not pose particular problems in the first stages, the high pressures increase the demand for sophisticated design solutions in the boiler, so that supercritical units are costly and used less frequently. Of course, the advantage of using a supercritical high-pressure section is a larger power output for the same mass flow rate, that is, a higher power density.

1.3.2 Condensing versus Noncondensing Turbines

A condensing steam turbine is provided with a vacuum device at its exit (the condenser) where the forced flow of an external coolant (water or air) maintains the outlet pressure at a level dependent only on the coolant inlet and outlet temperatures. In practice, the exhaust steam is condensed on a bank of "cold" tubes (Fig. 12) and the condensate is collected in the so-called cold well. Because the condenser is otherwise sealed, the pressure in the cold well (and in the whole condensing chamber) is the saturation pressure corresponding to the condensate temperature, which for current applications ($T_{coolant} = 25$–60°C) is lower than atmospheric ($p_{sat} = 0.01$–0.2 bar). The increased pressure differential across the turbine leads to a larger enthalpy drop of the steam and to a higher shaft power for the same mass flow rate. Most turbines in power generation plants are of the condensing (or C) type.

A noncondensing (or back-pressure) turbine has no condenser, and its exhaust steam is used directly for process purposes. Specific processes require different steam temperatures, but the range of exhaust steam pressures generally varies between 1.5 and 10.0 bars.

Most of the so-called cogeneration units in use by the industry are of the noncondensing (or NC) type.

For the same rated power output and inlet conditions, NC turbines are much more compact than C turbines because the steam at the former's outlet has a much lower specific volume. For large class 5 turbines, the size of the exhaust end limits the acceptable steam volumetric flow through the final stages because at the usual rotational regimes (1500/1800 or 3000/3600 rpm) "long" blades (≥ 1 m) are subject to excessive centrifugal stresses. Therefore, the largest units (especially in nuclear plants) are built in a double-flow configuration where there are two low-pressure turbine bodies for each medium-pressure one (as shown in Fig. 1A).

TABLE I

Typical Values for Steam Admission Conditions in Modern Turbines

Type of application	Throttle steam pressure (bars)	Throttle steam temperature (°C)	Net power putput per unit (MW)
Geothermal	5–15	150–250	0.5–10
Nuclear	45–70	200–300	50–250
Industrial cogeneration	40–60	400–450	1–50
Electrical power generation	120–180	510–550	50–250
Supercritical	> 220	550–600	50–100

1.3.3 Extraction Turbines

For the overall economy of the steam turbine cycle, it is often convenient to extract a portion of the steam evolving in the turbine before it reaches the turbine exhaust. This involves an extraction turbine. The extraction can be automatic or controlled. In both cases, steam is extracted through openings properly located on the turbine casing between one stage and the other. However, in the automatic extraction case (Fig. 13A), there is no means of controlling the pressure of the extracted stream, whereas in the controlled extraction case (Fig. 13B), a system of throttling valves can modify the pressure drop

TABLE II

Some General Classification Criteria for Steam Turbines

Criterion	Classification	Application field in which the standard is most used
Inlet steam Conditions	*Subcritical*: $p_{in} < 220$ bars and T_{in} usually < 550°C	Design and Process Engineering
	Supercritical: $p_{in} > 220$ bars and T_{in} usually > 550°C	Industry
Exhaust pressure	"C", *condensing* if $p_{\text{exhaust}} < p_{\text{atmospheric}}$ and the exhaust steam is condensed	Design and Process Engineering
	"NC", *noncondensing* otherwise (also called *back-pressure*)	
Main steam flow path	*Axial* if the main steam flow path is parallel to the turbine axis	Design
	Radial otherwise	
Unit power	Class 1: 15–400 kW, noncondensing	Industry
	Class 2: 450–1500 kW	
	Class 3: 2000–15,000 kW	
	Class 4: 15,000–100,000 kW	
	Class 5: > 100,000 kW, reheat and condensing	
Steam extraction	*Extraction* if a portion of the steam is extracted between stages for process purposes other than power generation	Design and Process Engineering
	Non-extraction otherwise	Industry
Inlet steam pressure	*High pressure*, above 40 bars	Design
	Intermediate pressure, 5–40 bars	Industry
	Low pressure, < 5 bars	
Blade design	*Impulse* if the stage pressure drop takes place only in the nozzles	Research
	Reaction otherwise	Design

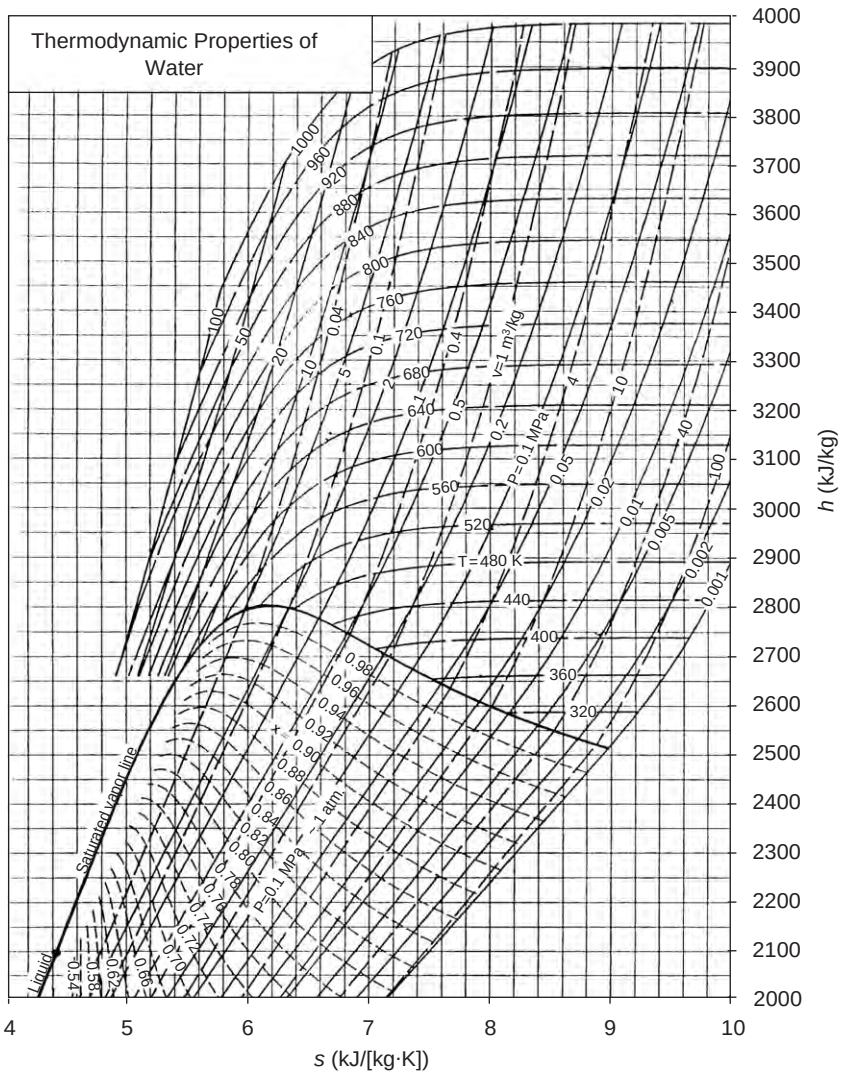


FIGURE 11 The Mollier steam chart. From Stanford University, Department of Mechanical Engineering.

between the extraction point and final use, thereby actively regulating the flow rate of the extracted steam. Extractions are commonly used to preheat the feed water in modern steam cycle-based power plants (class 5 C-type turbines) and to supply process steam in industrial plants (classes 3, 4, and 5, usually C-type turbines).

2. *ENERGY TRANSFER IN A TURBINE*

2.1 Expansion of a Fluid through a Turbine: Stage Efficiency

The flow through a turbine stage is usually modeled as an adiabatic expansion because the contact time of the fluid with the statoric and rotoric blades is very short (typically 0.1–1.0 ms), so that the amount of energy exchanged between the fluid and blade under the form of heat is entirely negligible with respect to that exchanged under the form of work. In the ideal case (no internal friction), the expansion is also isentropic (Fig. 14). At the inlet, the steam possesses a static enthalpy $h_0(p_0,T_0)$ and an absolute velocity V_0, so that its total (stagnation) enthalpy content is $h_{0,tot} = h_0 + 0.5V_0^2$. At stage exit, the static enthalpy has dropped to $h_{2i}(p_{2i},T_{2i})$ and the velocity is V_{2i}, so that the stagnation enthalpy is now $h_{2i,tot} = h_{2i} + 0.5\,V_{2i}^2$. The maximum specific energy that can be extracted from the fluid under these conditions is called the isentropic (or adiabatic) work:

$$w_{ad} = \Delta h_{ad} = h_{0,tot} - h_{2i,tot} \qquad (1)$$

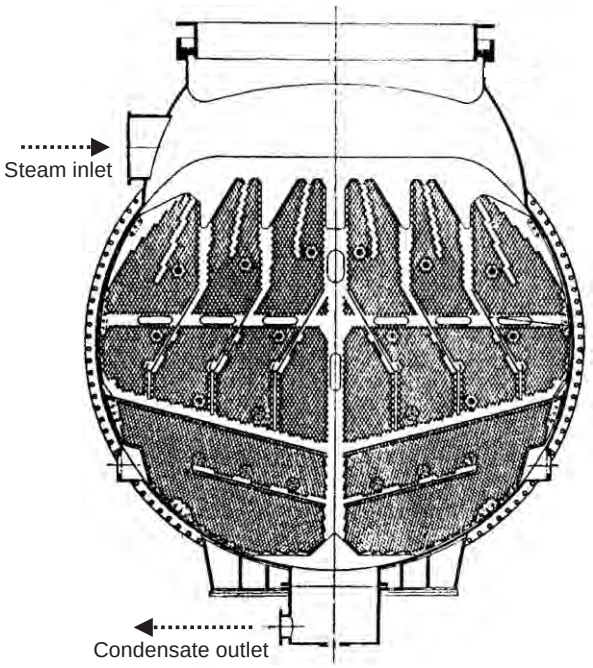


FIGURE 12 Meridional section of a steam condenser.

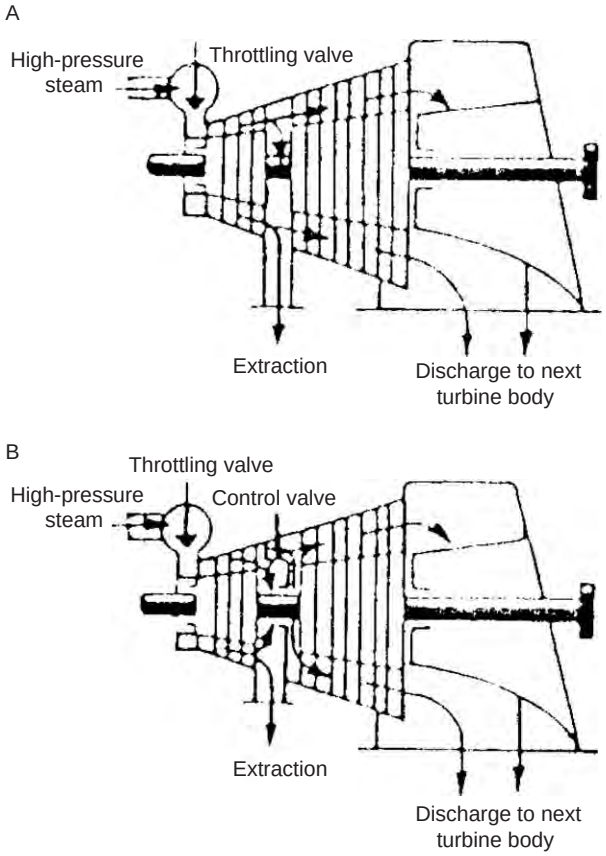


FIGURE 13 (A) Automatic extraction. (B) Controlled extraction.

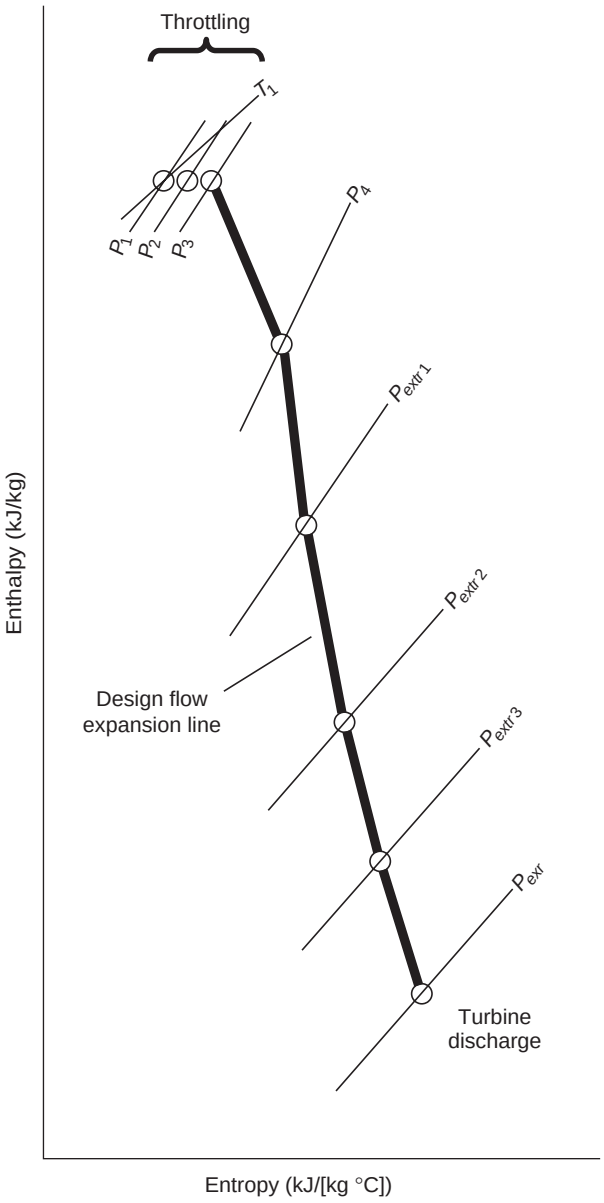


FIGURE 14 A real expansion line through a steam turbine.

The actual expansion is neither ideal nor adiabatic. Some heat flows from the hot fluid to the colder blades, creating a temperature gradient in the tangential and axial directions at each station. Boundary layers on the hub and casing walls originate velocity and temperature gradients in the radial direction. The flow is highly turbulent; neither the blade nor the disk and casing surfaces are perfectly smooth. A non-negligible amount of steam leaks through the unavoidable mechanical gaps and flows around the blades, originating "local recirculations" called secondary flows. All of these irreversible phenomena lead to the production of entropy,

degrading a portion of the steam energy and decreasing the overall amount of work provided to the blades by the fluid. Strictly speaking, a nonideal expansion cannot be represented on the Mollier chart because the "local" details of the entropy generation at each (p, T) point along the expansion line are unknown. But conventionally, the points representative of the inlet and outlet conditions can be connected with a line that has a concavity toward the right (Fig. 14) because the irreversible entropy generation increases with the stage expansion ratio. Because of the nonzero slope of the isobars, the actual total enthalpy drop is smaller than the isentropic one and the ratio of the two is referred to as the isentropic efficiency:

$$\eta_{is,tt} = \frac{h_{0,tot} - h_{2,tot}}{\Delta h_{ad}} = \frac{h_{0,tot} - h_{2,tot}}{h_{0,tot} - h_{2i,tot}}. \tag{2}$$

The quantity described by Eq. (2) is called the total-to-total efficiency as a reminder that the enthalpies appearing in its definition are stagnation values. It correctly accounts for the exit kinetic energy, which is a useful quantity because it contributes to the energy input in the next stage. Total-to-static and a static-to-static efficiencies are also used:

$$\eta_{is,ts} = \frac{h_{0,tot} - h_{2,tot}}{h_{0,tot} - h_{2i}} \tag{3}$$

$$\eta_{is,ss} = \frac{h_0 - h_2}{h_0 - h_{2i}}. \tag{4}$$

V_{2i} is always smaller than V_0 because in the absolute motion the fluid cannot accelerate; furthermore, the actual stage exit velocity V_2 is always somewhat smaller than its ideal counterpart because the same flow area must accommodate, *ceteris paribus*, a lower volumetric flow rate. It is possible to derive a relationship among the three efficiency expressions. Let

$$\begin{aligned}\Delta h_{tot} &= h_{0,tot} - h_{2,tot} \\ &= h_{0,tot} - h_{2i,tot} - T_0(s_2 - s_0) \\ &= \Delta h_{i,tot} - T_0 \Delta s\end{aligned}$$

and

$$V_2^2 = V_{2i}^2(1 - \alpha) \quad \text{where } \alpha < 1.$$

The following is obtained:

$$\eta_{is,tt} = 1 - \frac{T_0 \Delta s}{h_{0,tot} - h_{2i,tot}} \tag{5}$$

$$\eta_{is,ts} = \eta_{tt} \frac{1}{1 + \eta_{tt} \dfrac{V_{2i}^2(1 - \alpha)}{2\left(h_{0,tot} - h_{2i,tot}\right)}}. \tag{6}$$

$$\eta_{is,ss} = 1 - \frac{1 - \eta_{tt}}{1 - \dfrac{V_0^2 - V_{2i}^2}{h_{0,tot} - h_{2i,tot}}} + \frac{\alpha V_{2i}^2}{2\left(1 - \dfrac{V_0^2 - V_{2i}^2}{h_{0,tot} - h_{2i,tot}}\right)}. \tag{7}$$

The denominator in the right-hand side of Eq. (6) is greater than 1, so that $\eta_{is,ts} < \eta_{is,tt}$. With the usual values of $\Delta V_i (\cong 0.1 \Delta h_{i,tot})$ and of α ($\cong 0.05$), the values of $\eta_{is,ss}$ and $\eta_{is,tt}$ are indistinguishable for practical purposes. When in doubt, Eq. (5) is to be used.

2.2 The Steam Rate

It is still common in industrial publications to find mention of an alternate indicator of turbine performance, the so-called steam rate, defined as the mass of steam that must flow through the turbine to deliver a unit of power. Thus, this quantity ought to be expressed in kilograms per kilojoule but is commonly reported as "x kilograms per kilowatt-hour." The theoretical steam rate (TSR) is simply equal to the reciprocal of the ideal enthalpy drop,

$$TSR = \frac{1}{\Delta h_{ad}}, \tag{8}$$

whereas the actual steam rate (ASR) is the reciprocal of the actual work output,

$$ASR = \frac{m_{throttle}}{P_{out}}, \tag{9}$$

and the ratio ASR/TSR is equal to the net turbine efficiency. The steam rate is a redundant (and therefore useless) indicator that is cumbersome to use due to the necessary conversion between different systems of units, and its generality is limited by its explicit dependence on the initial and final expansion conditions. Its use ought to be discouraged.

2.3 From Zeroth-Order Approximation to Three-Dimensional Design

The thermodynamic stage efficiency constitutes a useful global description of the expansion process. However, the design of a turbine must be based on a detailed knowledge of the actual coupling between the working fluid and the machine channels. The "real" flow is fully three-dimensional, turbulent, and unsteady, and this makes a complete description

extremely difficult. As in many engineering tasks, we proceed by successive approximations:

1. First, we consider the so-called zero-dimensional problem in which the flow is steady and the fluid–machine interaction is expressed by global parameters such as mass (or volume) flow rate, expansion ratio, average values for the value and direction of velocity at the inlet and outlet of the blades, angular velocity of the rotor, and average physical properties of the fluid. We derive, via dimensional analysis, a set of general relations that constitute a basic design tool.
2. We then turn our attention to the ideal, steady, quasi-one-dimensional flow, devoid of boundary layers on either moving or fixed boundaries, and derive first-order expressions for the most important characteristic quantities. The flow rate and the specific work are expressed in terms of average velocities, and the stage performance is linked explicitly to the average flow field parameters.
3. The effects of boundary layers and other viscous and compressibility phenomena are then taken into account, and the results previously obtained are corrected to include the effects of losses, secondary flows, flow nonuniformities, sonic phenomena, and so on.
4. The fully unsteady, three-dimensional flow is studied. As of today, this is possible only under stringent simplifying assumptions and at the expense of substantial computational resources. Usually, cascade or stage simulations are performed to investigate particular characteristics of the flow field and are validated by comparing their results with experimental investigations.

2.3.1 Zeroth-Order Approximation: Dimensional Analysis

Dimensional analysis is a branch of applied mathematics that investigates the relationship between global characteristics of a physical phenomenon and a possibly small set of likewise global parameters that uniquely define it. This article adopts here a simple (but theoretically sound) approach. In turbomachinery design, dimensional analysis is essentially linked to the concept of similarity. Two machines, A and B, are said to be geometrically similar if the ratios between homologous dimensions all are equal (i.e., if $D_{2,A}/D_{2,B} = c_{tip,A}/c_{tip,B}$). The same two machines are also said to be kinematically similar if the ratios between corresponding velocities are the same (i.e., if $U_{2,tip,A}/U_{2,tip,B} = V_{1,A}/V_{1,B}$). Finally, two machines are said to be dynamically similar if the ratios between corresponding forces (or force-related quantities) are the same (i.e., if $\Delta p_{stator,A}/\Delta p_{stator,B} = \sigma_{centrifugal,tip,A}/\sigma_{centrifugal,tip,B}$). Notice that similarity considers ideal flow concepts only. Strictly speaking, two geometrically and kinematically similar machines are dynamically similar only if the dissipative features of the flow (leakages, boundary layers, and viscous and inviscid losses) are the same.

With specific reference to a turbine stage, let us consider the following question: can the global design parameters of the stage be reduced to a small set of fundamental dimensional variables that can be used to construct an even smaller set of dimensionless indexes that describe stage performance? A brief reflection indicates that a stage is described by the mass flow rate it elaborates, the work it extracts, the conditions of the fluid and the flow passage area at inlet and outlet, a proper measure of its size, and the angular velocity of its rotor. A possible set of relevant design parameters is listed in Table III. A fundamental set of dimensionless parameters that can be constructed with these quantities is reported in Table IV.

Turbines with equal Reynolds numbers, specific speeds, and specific diameters are dynamically similar (because the losses scale, albeit only approximately, with Reynolds numbers). This result, which can be derived theoretically and justified on the basis of first- and second-order considerations, has been confirmed by an extraordinarily large and comprehensive series of experimental results. It has several important consequences. First, the efficiencies of the two machines are also equal. Second, if we know the performance ($Q_{2,A}$, ω_A, and $\Delta h_{ad,A}$) and the scale ($D_{2,A}$) of the first machine, we can derive the corresponding quantities for the second machine once we fix either its dimensions ($D_{2,B}$) or one of its design specifications ($Q_{2,B}$ or $\Delta h_{ad,B}$, or ω_B). The

TABLE III

A Fundamental Set of Design Parameters for a Turbine

Parameter	Physical phenomena to which it relates
Fluid enthalpy drop (Δh_{tot})	Specific work
Fluid flow rate (m_o)	Machine size, power
Fluid density ($\rho(\rho,T)$)	Machine size, losses
Rotor and stator dimensions	Machine size, mechanical stresses, flow kinematics
Stage efficiency (η_{is})	Irreversible dissipation in blade rows
Rotor angular velocity (ω)	Mechanical stresses, flow kinematics

TABLE IV

A Possible Fundamental Set of Dimensionless Parameters for a Turbine

Parameter	Definition
Average flow Reynolds number	$Re = \frac{\rho U_{tip} D_{tip}}{\mu}$
Average flow Mach number	$Ma = \frac{C}{\sqrt{kRT}}$ (C is a local representative velocity)
Specific speed	$n_s = \frac{\omega\sqrt{Q_2}}{(\Delta h_{ad})^{3/4}}$
Specific diameter	$d_s = \frac{D_{tip}(\Delta h_{ad})^{1/4}}{\sqrt{Q_2}}$
Head coefficient	$\Delta\psi = \frac{\Delta h_{ad}}{U_{tip}^2}$
Flow coefficient	$\varphi = \frac{Q}{U_{tip}A}$ (A is local flow passage area)
Stage efficiency	$\eta_{is} = \frac{w_{real}}{w_{ad}}$
Degree of reaction	$R_\rho = \frac{h_1 - h_2}{h_{0,tot} - h_{2,tot}}$
Tip clearance/blade height ratio	$s = \frac{h_{gap}}{h_{blade}}$ (h is average height or clearance)
Rotor blade aspect ratio	$a = \frac{h_{blade}}{c_{blade}}$ (h and c are average representative values)

following relations hold:

$$\frac{\omega_A D_{2,A}}{\Delta h_{ad,A}^{1/2}} = \frac{\omega_B D_{2,B}}{\Delta h_{ad,B}^{1/2}} \tag{10}$$

$$\frac{\omega_A Q_{2,A}}{D_A \Delta h_A} = \frac{\omega_B Q_{2,B}}{D_B \Delta h_B} \tag{11}$$

$$\omega_A D_{2,A}^2 = \omega_B D_{2,B}^2, \tag{12}$$

where Eq. (10) derives from the Reynolds numbers equality, assuming that the steam viscosity at the exit of the two machines is the same (see Fig. 15 for the dependence of μ on temperature).

As can be seen in the following, the flow and head coefficient are also fundamental design parameters. They are related to n_s and d_s by

$$\Delta\psi = \frac{1}{(n_s d_s)^2} \tag{13}$$

$$\varphi = K_{shape} \frac{n_s \Delta\psi}{d_s}, \tag{14}$$

where K_{shape} is equal to $\frac{\pi\delta_b(1-\chi^2)}{2}$ and $2\pi\delta_b\left(\frac{b}{D}\right)$ for axial and radial turbines, respectively. The Reynolds number definition is clearly conventional. In most real cases, nowhere in the real flow field does the Reynolds number assume the value given by the formula in Table IV. Nevertheless, if used consistently, it allows a direct relative estimate of global

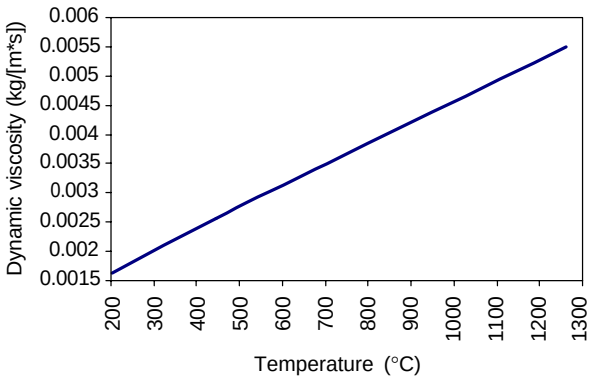


FIGURE 15 Steam viscosity versus temperature.

losses. For two turbines with the same n_s and d_s (or $\Delta\psi$ and φ, which is the same), a widely used formula, originally from Staufer in 1925, states that

$$\frac{1-\eta_{is,tt,A}}{1-\eta_{is,tt,B}} = C_1 + C_2\left(\frac{\mathrm{Re}_B}{\mathrm{Re}_A}\right)^n, \tag{15}$$

where $C_1 \approx 0.3-0.4$, $C_2 \approx 0.7-0.6$, and $n \approx 0.2-0.25$. Equation (15) implies that not all losses are of viscous origin (C_1 is the percentage of inviscid losses on the total).

On the basis of similarity analysis, universal performance charts can be compiled. If we have one verified performance map of a given turbine stage, similarity theory states (and experimental results confirm) that such a map applies to all stages with the same pair (n_s,d_s)—or (ψ,φ), which is the same—so long as their operation implies Reynolds numbers on the order of the one in the test. The admissible range is somewhat application dependent, but for turbines of commercial standard, a variation between $Re = 10^5$ and $Re = 10^7$ leads to a total-to-total efficiency change of approximately 0.05 (from 0.87 to 0.92). Two widely used performance maps, known as the Balje and Smith charts, are shown in Figs. 16 and 17, respectively.

In spite of the formal identity, there is a substantial practical difference in taking the sets (Re, n_s, d_s) or (Re, $\Delta\psi$, φ) to represent the variation of the stage efficiency at different design conditions. By definition, n_s/d_s maps are dependent on the stage outlet temperature and pressure and, therefore, do not apply strictly to similar machines working under different operating conditions. To eliminate this inconvenience, "corrected" quantities are sometime used, especially in industrial references. If for a turbine stage m_r is defined as the reference mass flow rate evolving in the stage when the inlet steam is at T_r and p_r and the rotor is rotating at a reference angular

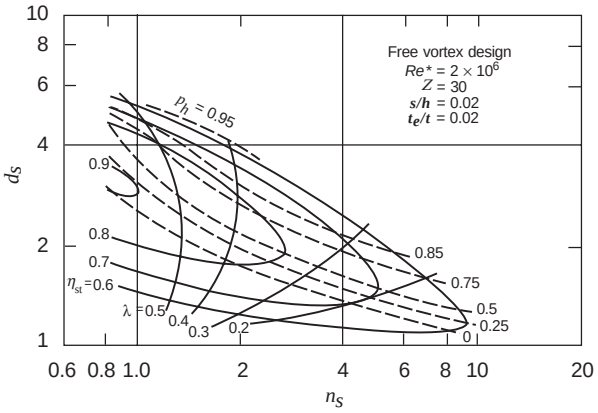


FIGURE 16 A typical Balje chart for steam turbines.

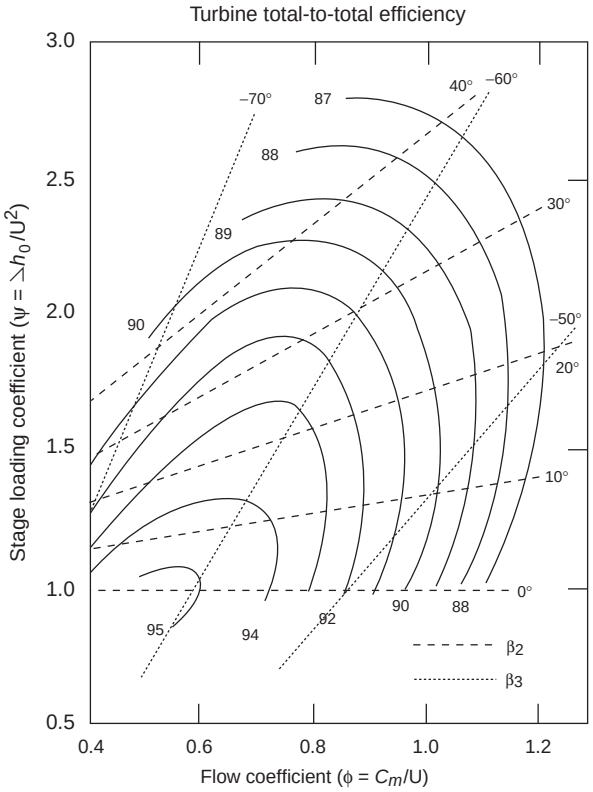


FIGURE 17 A typical Smith chart for a steam turbine with $R_\rho = 0.5$.

velocity ω_r, the following relations hold:

$$m_r = m\frac{\sqrt{T/T_r}}{p/p_r} \tag{16}$$

$$\omega_r = \frac{\omega}{\sqrt{T/T_r}}\,.$$

Of course, reference conditions are still included in the preceding definitions, but this reduces to a matter of convention. For geometrically and kinematically similar machines, corrected stage performance maps are displayed in Fig. 18.

2.3.2 First-Order Approximation: The Euler Equation and the Velocity Triangles

Consider the flow in a generalized turbine rotor (Fig. 19). The fluid enters the upstream side of the rotor with absolute velocity V_1 and exits from the downstream side with velocity V_2. Notice that suffixes 1 and 2 refer to the (virtual) geometric center of the inlet and outlet sections, so that V_1 and V_2 are proper average velocities, and that the rotor is rotating with a steady angular velocity ω. Newton's second law states that the module of the tangential force acting on the rotor is equal to the rate of change of angular momentum in the fluid, $|F_t| = \frac{d(mV_t)}{dt}$, which for steady V_t and m reduces to $F_t = F_{t1} - F_{t2} = m(V_{t1} - V_{t2})$. Therefore, the torque on the rotor shaft is

$$M = r_1F_{t1} - r_2F_{t2} = m(r_1V_{t1} - r_2V_{t2}). \tag{18}$$

The specific work (power per unit mass) is then

$$w_{ad} = P/m = M\omega/m = \omega(r_1V_{t1} - r_2V_{t2}). \tag{19}$$

Defining $U = \omega r$ as the rotational velocity, and recalling that $w_{ad} = \Delta h_{ad}$, we obtain the so-called Euler equation:

$$\Delta h_{ad} = \omega\,(r_1V_{t1} - r_2V_{t2}) = U_1V_{t1} - U_2V_{t2}. \tag{20}$$

This equation is fundamental in turbomachinery design because it relates the thermodynamic work (computed via steam charts or thermodynamic formulas) to the mechanical work expressed in terms of geometric and (average) kinematic quantities. Notice that the velocities appearing in Eq. (18) are ideal velocities (no losses included) averaged over the passage area.

The flow kinematics is conveniently represented by means of the so-called velocity triangles, which are simply a graphical display of the basic velocity vector relationship:

Absolute Velocity Vector
= System Velocity Vector
+ Relative Velocity Vector

or

$$V = U + W. \tag{21}$$

These diagrams (Figs. 1 and 19) are drawn on a single plane, tangent to the pseudo-cone that represents the machine at radius r. Three values are usually provided: at the hub (r_h), midspan (r_m), and tip (r_t). For radial stages, there is only one value at

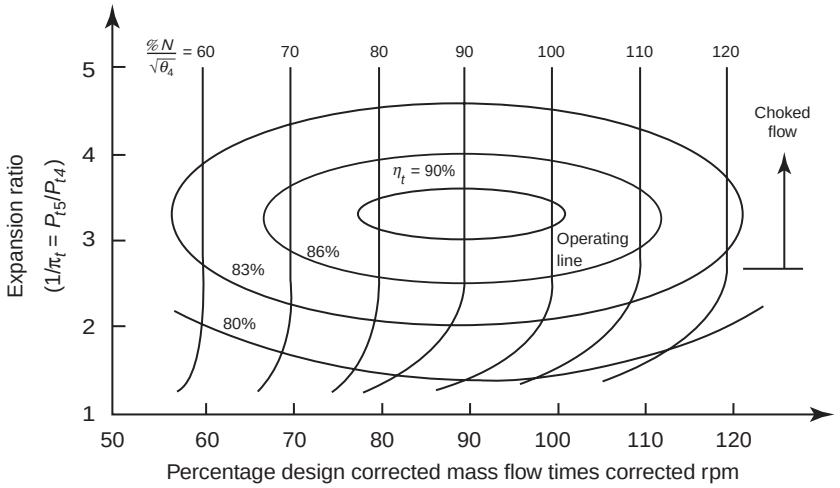


FIGURE 18 Corrected turbine performance map.

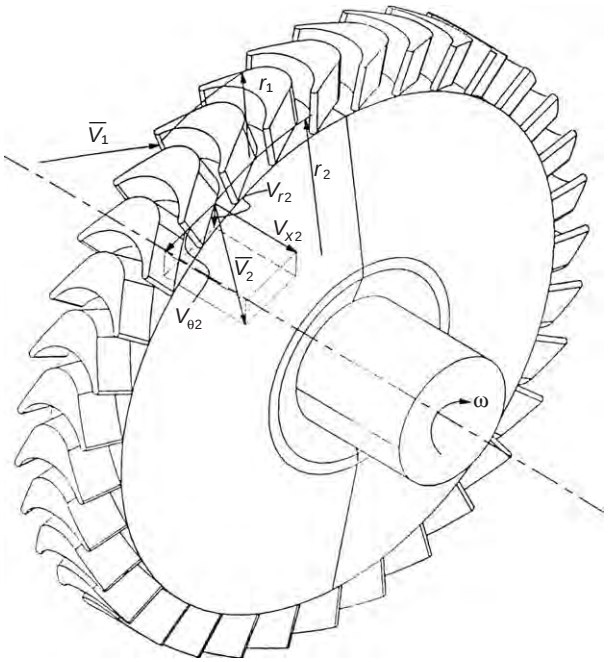


FIGURE 19 Schematics of the flow in an axial turbine rotor.

the inlet (r_1). These triangles can also be defined in terms of one velocity and two angles: the relative flow angle β and the absolute flow angle α. The cascade nomenclature provided in Fig. 1 is not universally adopted, and sometimes (especially in English monographs) the values $\beta-90^\circ$ and $90^\circ-\alpha$ are used instead. The following relations hold between the modules of the velocity vectors ("j" is a running index denoting the station to which the triangle refers):

$$V_{j,x} = V_j \sin\alpha = \varphi_j U_j \tag{22}$$

$$V_{j,t} = V_j \cos\alpha = \psi_j U_j \tag{23}$$

$$V_j^2 = V_{j,x}^2 + V_{j,t}^2 = U_j^2\left(\varphi_j^2 + \psi_j^2\right) \tag{24}$$

$$\begin{aligned} W_j^2 &= W_{j,x}^2 + W_{j,t}^2 \\ &= U_j^2\left[\varphi_j^2 + \left(1-\psi_j\right)^2\right] \end{aligned} \tag{25}$$

$$W_{j,t} = V_{j,t} - U_j \tag{26}$$

$$V_{j,x} = W_{j,x} \tag{27}$$

$$U_j V_{j,t} = \frac{\left(U_j^2 + V_j^2 - W_j^2\right)}{2}. \tag{28}$$

Relative total temperatures (T_{rel}) are defined as well:

$$\begin{aligned} T_{j,tot,rel} &= T_j + \frac{W_j^2}{2\overline{c_p}} \\ &= T_{j,tot} + \frac{\left(W_j^2 - V_j^2\right)}{2\overline{c_p}} \\ &= T_{j,tot} + \frac{U_j^2(1-2\psi)}{2\overline{c_p}}. \end{aligned} \tag{29}$$

$T_{tot,rel}$ is the stagnation temperature in the rotating frame of reference and, therefore, is the temperature to which the rotor material is subjected.

It is customary to classify turbine stages on the basis of four characteristic indexes identifying their velocity triangles: the degree of reaction R_ρ, the exit swirl ω_2, the stage loading coefficient $\Delta\psi = \psi_1 - \varepsilon^2\psi_2$, and the flow coefficient φ. The respective values must be labeled with the radius at which the index is computed (in the absence of another indication, blade midspan is to be assumed). The degree of

reaction is defined as the ratio of the static enthalpy drop across the rotor to the decrease of the total enthalpy across the stage:

$$R_\rho = \frac{h_1 - h_2}{h_{0,tot} - h_{2,tot}} = \frac{(U_1^2 - U_2^2) - (W_1^2 - W_2^2)}{(U_1^2 - U_2^2) + (V_1^2 - V_2^2) - (W_1^2 - W_2^2)}. \quad (30)$$

A stage in which $R_\rho = 0$ is said to be an impulse stage. For axial turbines, Eq. (30) implies that in such a stage there is no pressure drop across the rotor ($W_1 = W_2$); that is, the entire stage pressure drop is attained in the inlet nozzle (the stator), and the rotor absorbs only the fluid's kinetic energy. All stages for which $R_\rho > 0$ on a portion of the blade are said to be reaction stages. The value of R_ρ is, whenever possible, limited to approximately 0.5 to limit the absolute kinetic energy at stage outlet.

A zero exit swirl corresponds to the minimum absolute kinetic energy at stage exit. For a single stage and for the end stages of all machines, this is a desirable condition. For an intermediate stage, where the exit kinetic energy can be recovered by the next (bottoming) stage, work maximization often leads to a negative ψ_2.

The loading coefficient is chosen as large as possible, compatible with the desired value of the efficiency (Fig. 20). Common values for axial stages range from 1 to nearly 3. For radial stages, $\Delta\psi \cong 1$.

2.3.2.1 Impulse versus Reaction Stages: Blade Twist It is useful to provide a more detailed discussion of impulse and reaction stages. In an impulse stage, the steam experiences the entire stage pressure drop in the statoric channels and is ejected from the stator as a jet (or, better, as a number of jets—one for each statoric passage). In the rotor, these jets flow through a blade with a (nearly) constant relative velocity along the rotoric passage. Therefore, the pressure drop in the rotor is zero (Eq. 30). The stage work is produced in the rotor at the expense of the absolute velocity of the steam. Assuming zero losses in the stator,

$$h_{0,tot} - h_{2,tot} = \frac{(U_1^2 - U_2^2) + (V_1^2 - V_2^2)}{2}, \quad (31)$$

we see the following:

1. The stage work increases with the expansion ratio p_0/p_2 until V_1 becomes sonic and the stage is "choked."
2. With the stage work being maximum for $V_{2,t} = 0$, it can be easily shown that the optimal U/V_1 ratio for axial turbines is approximately 0.5. In centripetal turbines, experience shows that $(U/V_1)_{optimal} \sim 0.7/0.8$.
3. In centripetal turbines, the flow configuration is such that even if $V_2 = V_1$ (which is never realized in practice because it would lead to high rotoric losses), the stage would deliver a positive work $(U_1^2 - U_2^2)/2$.
4. Radial centrifugal turbines deliver, for the same steam conditions and peripheral velocity, a lower stage work. In fact, they have been abandoned except for the Ljüngstrøm turbine (described in the first section of the article).

Therefore, in an impulse stage, the rotor is likely to be unshrouded and devoid of the leakage seals that abound in the stator. A "typical" axial impulse stage is depicted in Fig. 21. Notice that for today's higher throttle conditions (400–560°C and 30–220 bars), the sonic velocity in the first nozzle is so high that the ratio U/V_1 would be much lower than 0.5, leading to a substantial efficiency penalty. An ingenious solution (first suggested by Curtis) is that of absorbing the absolute kinetic energy drop in two separate rotors "fed" by a single stator. The intermediate fixed blading is simply a deflector (Fig. 21B). This configuration is often adopted in the first stage(s) of large steam turbines because, for the same peripheral velocity U, the velocity V_1 is higher than $2U$ (its optimum is at $\sim 4U$); therefore, the enthalpy (and pressure) drop per stage is increased. This means not only a lower overall number of stages but also a smaller high-pressure casing, and this results in a lessening of the installation costs. The price is a somewhat lower stage efficiency.

Because the ratio of the blade height to the external diameter is small only for the very first high-pressure stages, and because the requirement of

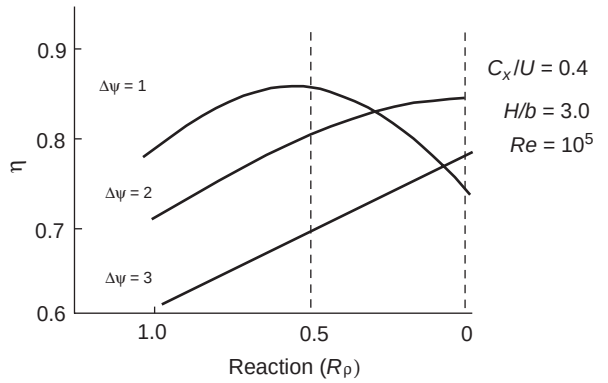


FIGURE 20 Loading coefficient versus degree of reaction and stage efficiency.

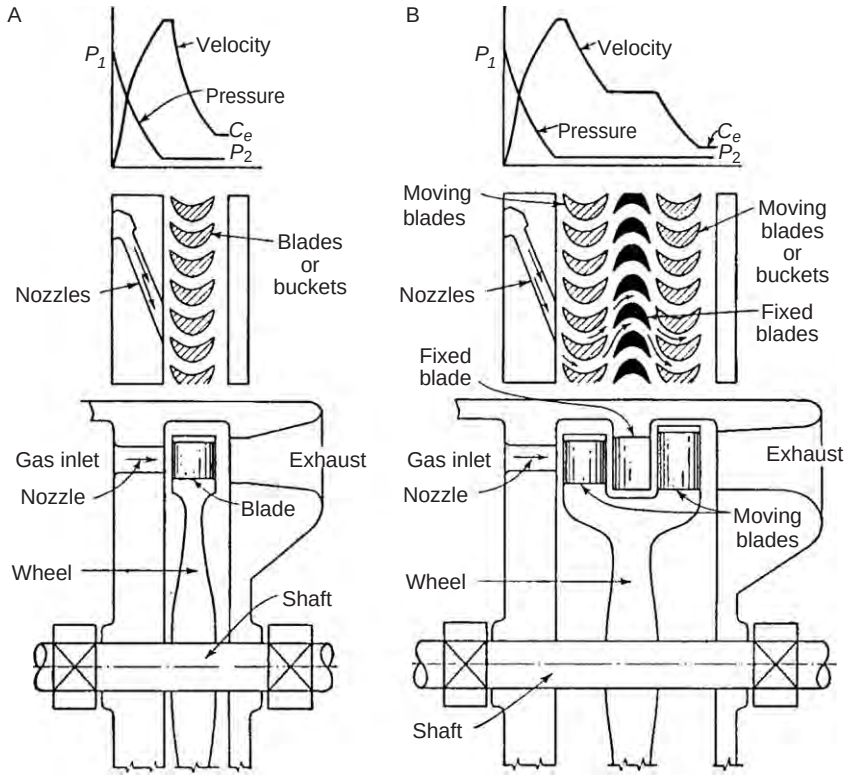


FIGURE 21 Impulse stages: (A) simple and (B) velocity compounded ("Curtis").

constant total enthalpy along the radius (to avoid excessive radial motion in the nonrotating channels) demands that the Euler work given by Eq. (20) be also constant along the radius, it follows that neither V_t nor R_ρ can be constant along the blade. Therefore, the term reaction stage implies that the degree of reaction at its midspan diameter is greater than zero. Actually, there are cases in which $R_\rho = 0$ at the hub and reaches values higher than 0.5 at the tip. For $h/D > 0.02$, the velocity triangles differ markedly at the hub, midspan, and tip, and the blade must accommodate to radially varying V_{1t} and V_{2t}. Several solutions are possible, with the one called free vortex ($V_t{}^*r = \text{constant}$) being used most frequently. The resulting rotation of the blade profile at various radii around its central radial axis is called "twisting." An example of a twisted blade is shown in Fig. 22.

2.3.3 Second-Order Approximation: Loss Correlations

The difference between the ideal turbine work Δw_{ad} and the actual shaft work is due to the losses in the turbine. It is common design practice to apportion these losses to each component and to write

$$\Delta w = \Delta w_{ad} - \Delta h_{tot,inlet} - \Delta h_{tot,stator} - \Delta h_{tot,rotor} - \Delta h_{tot,diffuser} - \Delta h_{tot,exit} - \lambda_{mechanical}. \quad (32)$$

Notice that fluid-related losses have been expressed in terms of total enthalpy drop. According to the Gouy–Stodola theorem, this may be also written

$$\Delta h_{tot,j} = T_0 \Delta s_j = \Delta e_j, \quad (33)$$

where $T_0 \Delta s_j$ is the irreversible entropy generation for the jth loss and Δe_j is the corresponding exergy destruction. In what follows, a concise and general description of how to include these losses in design calculations is provided.

- *Inlet losses*. These are due to frictional and flow turning effects (which cause flow separation and originate recirculation regions).
- *Partial admission losses*. In partial admission stages, there are two additional losses caused by the drag experienced by the blades in their inactive arc ("ventilation" losses) and by the fact that a portion of the energy of the statoric jet is perforce lost when the blade first enters its active arc and when it exits it ("sector" losses).
- *Statoric losses*. These are due to friction within the blade row, to secondary flows caused by turning effects (which cause flow separation and originate

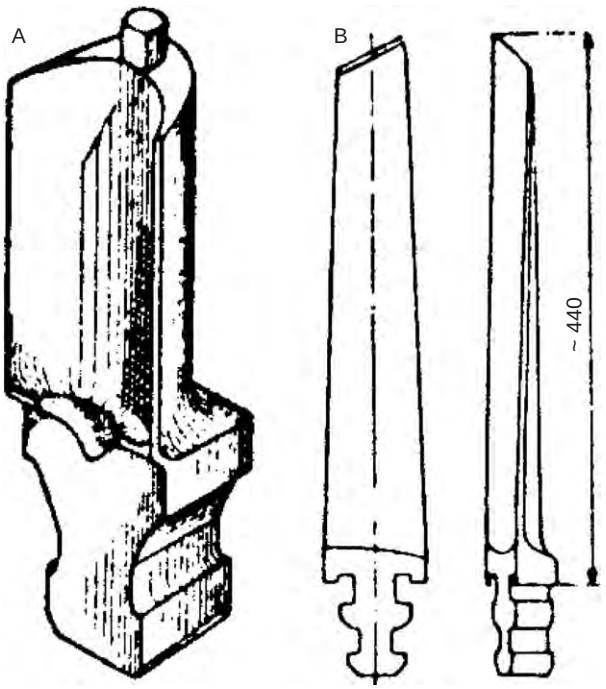


FIGURE 22 Untwisted (A) and twisted (B) steam turbine blades.

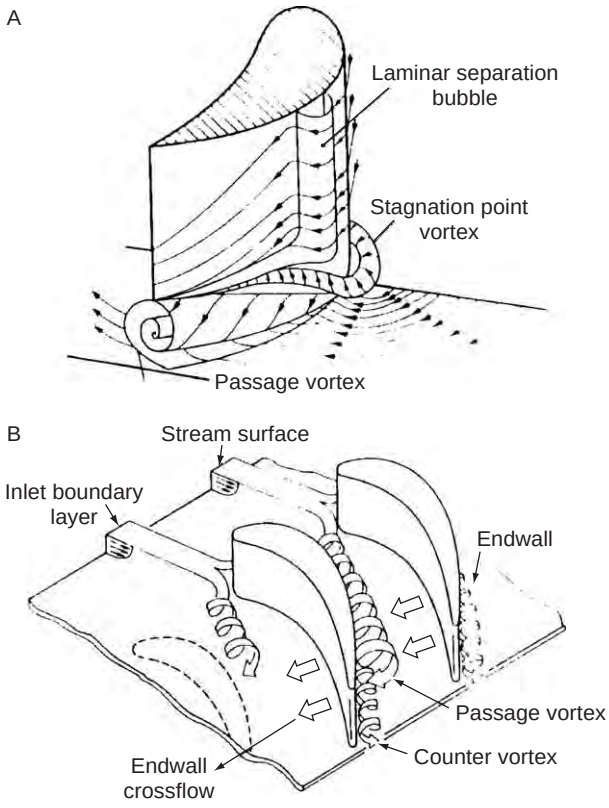


FIGURE 23 Vortical structures along the blade channel.

recirculation regions), to boundary layer growth along the blade ("blockage"), and to flow compressibility effects. These losses are usually split into their individual contributions, which are usually considered separately during design. Profile losses are originated by a complex system of vortical structures along the blade channels that strongly interact with each other and cannot be conveniently treated separately (Fig. 23). These losses are usually evaluated by computing or measuring the so-called lift and drag coefficients, defined as the ratios of the lift and drag force, respectively, to the kinetic energy of the oncoming flow (Fig. 24A). A large database exists for several blade profiles, and designers make use of "cascade performance charts" like the one depicted in Fig. 24B. If the velocity triangles are known, so is the cascade efficiency. Sometimes an additional loss, the wake loss, is included in the profile loss. Secondary and blockage losses are similarly modeled in terms of the square of the average meridional velocity, whereas shock losses, originated by sonic and supersonic shocks interacting with the boundary layers and with the wake, are computed on the basis of semi-empirical charts. Leakage flows (usually negligible in stators) are due to the unavoidable leaks around seals and gaps. They depend on the blade aspect ratio (h/l, height over chord) and gap size (g/h).

• *Rotoric losses.* These are modeled in a manner similar to that of stators. However, here the relevant velocity is the relative inlet velocity W_1:

$$\begin{aligned}\Delta h_{tot,rotor} &= \Delta h_{tot,profile} + \Delta h_{tot,secondary} + \Delta h_{tot,blockage} \\ &\quad + \Delta h_{tot,shock} + \Delta h_{tot,leakage} + \Delta h_{tot,windage} \\ &= 0.5\, K_{rotor} W_1^2, \end{aligned} \tag{34}$$

where K_{rotor} is a weight-averaged loss coefficient.

• *Diffuser losses.* These are also due to frictional and recirculation effects. The diffuser is a component in which the fluid flows against a pressure gradient (velocity is recovered into pressure); therefore, it is intrinsically affected by the highest losses (in percentage). In steam turbines, the importance of diffusers is normally negligible.

• *Mechanical losses.* These are due to frictional dissipation in bearings and seals. Customarily, they are accounted for introducing an organic efficiency:

$$\lambda_{mechanical} = \left(1 - \eta_{organic}\right)\Delta w_{ad}, \tag{35}$$

with $\eta_{organic}$ ranging from 0.95 to 0.98 for the entire shaft line (if a shaft supports N stages, the loss apportioned to each stage is $1 - \eta_{organic}/N$).

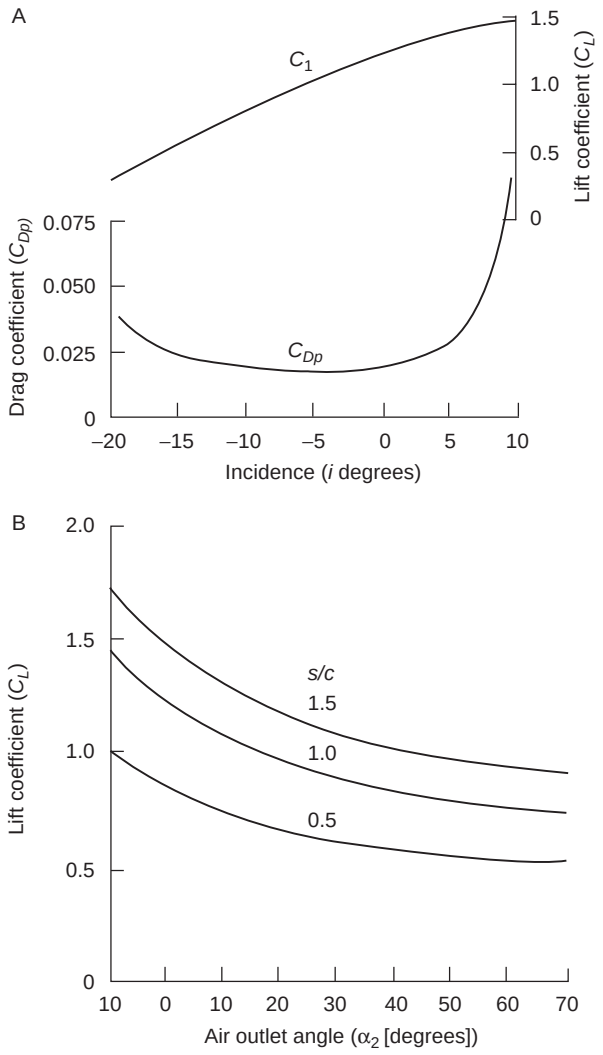


FIGURE 24 (A) Lift and drag coefficient for a cascade of fixed geothermal form. (B) Design lift coefficients.

Once the total stage losses have been computed, they are used to

1. Correct the ideal Euler work (Eq. 20). and derive the actual work delivered by each stage (Eq. 31).
2. Recompute the fluid thermodynamic conditions at the exit of each stage and draw the actual expansion line of Fig. 14.

2.3.4 Three-Dimensional Calculation of the Real Flow in a Complete Turbine Stage

The real flow in a turbine stage is highly turbulent, fully three-dimensional, and intrinsically unsteady. Although its general features are well understood (Fig. 23), it is the local features and their complex interactions that may make the difference in the stage efficiency. It is difficult to capture the individual flow phenomena through test bench experiments, and a large (and increasing) portion of turbine studies are performed by means of numerical simulations. Such simulations are complex, resource-consuming, and often not univocal (in the sense that some of the detected phenomena may be purely numerical features and not necessarily reflective of physical phenomena). Therefore, they must be carefully planned and performed, and their results are always calibrated against available experimental evidence. Basically, the flow domain (the portion of space in which the phenomena are studied) includes a portion of the interblade passage and also covers a limited portion of the channel upstream and downstream of the blade passage. Periodic tangential conditions are usually assumed (i.e., the basic features of the flow are considered to be the same in adjacent blade passages), and only recently has the coupling between statoric and rotoric blade rows been simulated in full. All of these studies strongly depend on the turbulence modeling adopted ("full" calculations, or the so-called DNSs [direct numerical simulations], are still impossible on the complex geometries of a turbine stage and in their operational high Reynolds numbers), and some of the effects can be assessed only under the simplifying assumption of inviscid flow. Nevertheless, reliable work has been performed to study both inviscid (i.e., purely kinematical) and viscous (i.e., including boundary layers and wake interactions) flow features, and the results compare well with global and local experimental measurements. A situation is now being approached where blade design can be performed on the basis of successive approximations, by iterative adjustment of several design parameters, to determine the best blade profile for a specified performance. However, at the industrial level, most of the design activity is still based on proprietary loss correlations, and fully three-dimensional simulations are still considered a helpful tool but not a completely reliable design procedure. With the advance of computer hardware (which leads to higher computational speeds and reduced simulation time), computational fluid dynamic research (which provides a better understanding of local flow features), and available software (which improves the quality of the commercially available flow models), it is clear that in the near future, inverse design problems will be solved by means of standard, well-understood, and relatively easy-to-use procedures.

3. OUTLINE OF A GENERAL PROCEDURE FOR PRELIMINARY STEAM TURBINE DESIGN

The actual design of a steam turbine is a complex task, usually carried out by a multidisciplinary design team whose members constantly interact with wind tunnel and test bench specialists. The technology is mature and extremely resource intensive, so that production is concentrated in very few locations worldwide. The independent steam turbine producers of commercial and industrial relevance (i.e., those that develop their own products, do not produce under license, and generate a significant number of units per year) amount to no more than perhaps 15 worldwide. Each has its own design procedure, formalized in detailed design manuals, usually meant for internal use only. However, it is highly likely that all of these proprietary design procedures follow the very same general logical flowchart and that the differences lie either in the formulas used for case-specific calculations (e.g., loss coefficients, blade angle corrections, material selection) or in the selection of a particular turbine shape (due to patent history and/or the company's technological background). Therefore, steam turbine design can in principle be said to follow nine general steps:

1. Calculation of the thermodynamic expansion process
2. Selection of the type of turbine (single stage vs multistage, radial vs axial)
3. Selection of the maximum tip speed
4. Calculation of n_s and selection of the proper similarity parameters
5. Tracing of the velocity triangles
6. Loss accounting and possible iteration (go back to step 3)
7. Preliminary blade design and costing
8. Numerical simulations under realistic operating conditions
9. Final design and costing

Because of space limitations, a detailed description of each of these phases cannot be provided here. The discussion is limited to a brief explanation of the most important features of each design step.

3.1 Calculation of the Thermodynamic Expansion Process

The expansion process is customarily calculated manually on *s/h* graphs such as the Mollier diagram. More recently, property packages have been developed that allow the calculation of a full expansion line on a personal computer. Notice that to trace the real expansion line, a turbine efficiency must be assumed here (Fig. 14). Clearly, this initial assumption requires an iterative verification.

3.2 Selection of the Type of Turbine

For small applications, where the design specifications require a low power output (less than a few megawatts) in conjunction with a low mass flow rate, single-stage radial steam turbines are probably most convenient. In all other cases, single- and multistage axial turbines dominate the field. The selection can be made either on the basis of specific experience on the part of the designer or of the company (there are companies that produce only a certain type of machinery) or with the help of "operating charts" (Fig. 25) that condensate decades of experience and are updated periodically. The semi-empirical Balje's charts (Fig. 16) are also a great help because they indicate the ranges in which a certain type of machine has performed well at the current stage of technology.

3.3 Selection of the Maximum Tip Speed

There are two major limitations in turbine design: excessive mechanical stresses on the blade and disk and choking of the flow in the blade passages. Both limitations constrain the maximum allowable tip velocity U_t. Very early in the design procedure, it is necessary to assign a value of U_t, and the choice is usually made on the basis of specific designers' experience, often explicitly tabulated in design manuals. Because this choice has a significant influence on the remaining design parameters, an iterative procedure must be used.

3.4 Calculation of n_s and Selection of the Proper Similarity Parameters

The specific speed is conventionally computed on the outlet section of the stage (i.e., the volumetric flow rate is computed on the basis of the exit steam density):

$$n_s = \omega \frac{\sqrt{\frac{\dot{m}_{out}}{\rho_{out}}}}{\Delta h_{tot,stage}^{3/4}}.$$

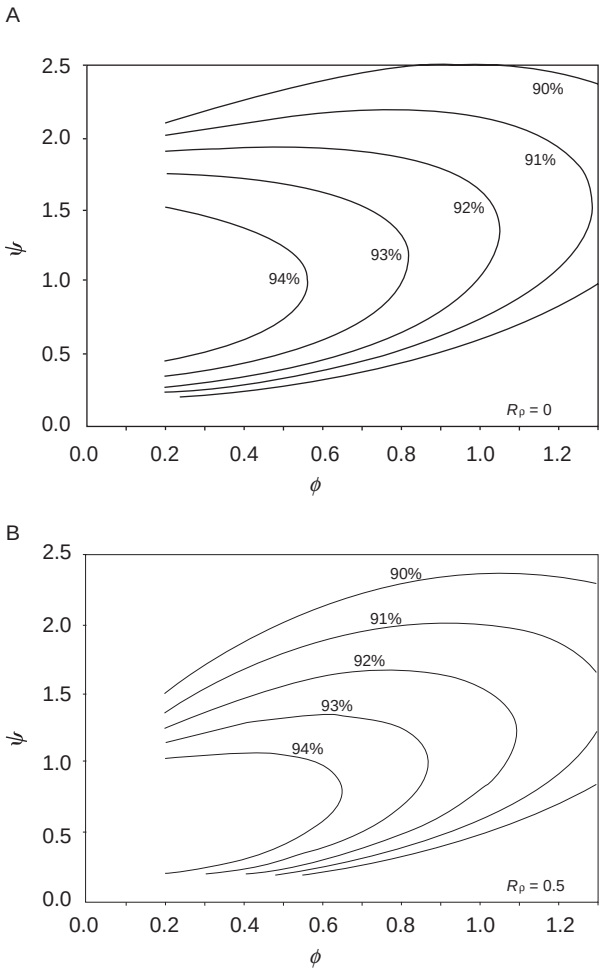


FIGURE 25 (A) Predicted total-to-total efficiencies for axial turbines with 0% reaction (impulse stages) and blade aspect ratio ($h/b = 10$). (B) Predicted total-to-total efficiencies for axial turbines with 50% reaction and blade aspect ratio ($h/b = 10$).

Confirmed experience has shown that the n^s for steam turbines varies between 0.1 and 4. If values smaller than 0.1 are obtained, a multistage configuration ought to be chosen and a new n_s should be computed using the stage enthalpy drop $\Delta h_i = \Delta h_{tot}/N_{stages}$. If, on the other extreme, n_s exceeds 4 or 5, to maintain high stage efficiency, it is mandatory to split the steam flow among M turbine bodies (parallel operation). M is determined by the condition that Q_{tot}/M is such that the n_s based on it falls within the range of 0.1 to 4. Once the specific speed is known, the remaining similarity parameters (most importantly the head coefficient ψ and the flow coefficient φ) can be determined on the basis of semi-empirical diagrams such as the Smith (for axial stages [Figs. 17 and 25]) and Chen–Baines (for radial stages [Fig. 26]) charts.

3.5 Tracing of the Velocity Triangles

The choice of the number of stages, of the n_s, and of the flow and head coefficients for each stage is far from being unique. For example, there are several reasons that suggest an uneven allocation of the Δh_{tot} among the stages (e.g., the need to maintain an optimal value of n_s, a convenient construction of the turbine bodies with each containing N_b stages). Whatever the choice is, at this point one possesses all of the information needed to draw the preliminary blade profile and can proceed as follows:

1. Introducing the chosen value of φ into the continuity equation applied at the rotor inlet section, we derive the blade height h (or the hub-to-tip ratio γ).
2. From U_t, ψ, and φ, we derive V_t, V_m, α, and β in each section of interest. For a radial stage, the inlet diameter at blade midspan is used together with both the tip (shroud) and hub exit values. For axial stages, it is customary to draw at least three triangles at the hub, midspan, and tip diameter. The ψ and φ given by the charts are tip values, and the other values must be computed properly. Usually, the assumption is that the V_m is constant along the radius and that V_t follows the so-called free vortex law: $V_t{}^* r = \text{con-}$ constant. But other choices, especially where the V_t is concerned, are possible.
3. The Euler work is computed at each section, and its average value on the entire blade is assessed. The ideal values for the relative flow angles (β) are calculated. A proper blade profile may now be chosen (or designed anew).
4. From a momentum balance in the relative flow, the pressure drop in the rotor can be calculated. Hence, the statoric pressure drop is obtained.
5. Once α_1 is known along the blade height, the statoric profile is also determined because it is reasonable to assume an entirely axial incoming flow ($V_o = V_{m,o}$, or $\alpha_o = 90^\circ$).

3.6 Loss Accounting and Possible Iteration

Now, the losses expressed by the semi-empirical relations discussed previously can be computed, and their overall effects can be considered before modifying the velocity triangles. Some of these losses affect the flow rate, and some others affect the flow section. Therefore, it is necessary to iterate to attain a reasonable compromise between good overall efficiency and a feasible design. Of course, the design specifications are imposed as strong constraints.

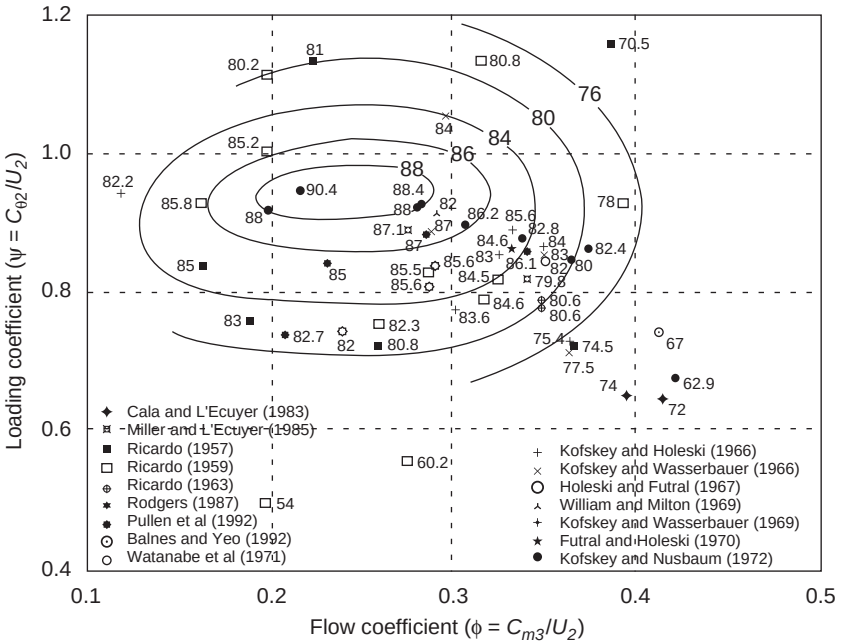


FIGURE 26 A Chen–Baines diagram for a radial turbine.

3.7 Preliminary Blade Design and Costing

Once all geometrical dimensions are known, the blades and all other components can be preliminarly sized. The choice of materials is imposed in practice by the environment within which the design activity takes place (with industrial and commercial strategies usually conflicting with purely technical considerations). A preliminary costing assessment of the machine is developed.

3.8 Numerical Simulations under Realistic Operating Conditions

As mentioned previously, the relatively simple and essentially two-dimensional procedure based on the velocity triangles can capture only the average and large-scale details of the flow in the turbine channels. Nevertheless, machines have been designed successfully with this method or similar methods for decades, attaining overall efficiencies of more than 90%. However, the progressive increase of the steam turbine unit power led to a situation where even minor efficiency improvements became attractive. Because an essentially two-dimensional design can be improved only by extensive testing, an impressive amount of experimental work was performed on models and full-scale stages alike. The advancement of computer technology made it possible to substitute physical testing with its numerical counterpart. Numerical experiments, which now constitute a substantial part of research and development (R&D) programs in the field, are attractive because, if properly conceived and performed, they share most of the properties of physical experiments but can be carried out at a fraction of the cost and possess an intrinsic degree of flexibility that makes their repetition under slightly different conditions relatively easy. To be useful for design purposes, numerical simulations must have the following requisites:

- *Physical relevance*. They must simulate phenomena that are relevant to design engineers. For instance, they must reproduce ranges of Reynolds numbers and Mach numbers that are found in commercial turbines.
- *Completeness*. They must include a realistically extended portion of the flow domain, so that the information that designers gather from the results may be immediately applicable to global design procedures. For instance, they must include at least one complete blade passage and possibly a complete stage.
- *Reliability*. Their results must not contradict known physical evidence (i.e., they must be calibrated against it). In addition, new phenomenological predictions that emerge from their results must be in accord with the established knowledge in the field.
- *Robustness*. Their results must be reproducible under slightly varied operational conditions within a

reasonable variation of the operating parameters. For instance, the limit behavior in time for all stable unsteady simulations must tend to the steady solution.

• *Physical consistency.* The models (numerical as well as physical) embedded in a numerical simulation must be in agreement with the general physics of all relevant phenomena. For instance, a turbulent model that assumes a certain relationship between the fluctuating velocity components and the local turbulent kinetic energy cannot violate any of the known energy balances established by global considerations.

The first numerical simulations (during the 1970s) were based on inviscid models of the Navier–Stokes equations and could be applied only to very simple geometries. Nevertheless, they demonstrated the capability of capturing some of the known physics. For instance, they correctly predicted the variation of the curvature of streamlines at different radii due to a Coriolis-like effect. During recent years, completely three-dimensional simulations of complex geometries have been presented where the entire stage is simulated (although a correct accounting of the relative motion of the rotoric and statoric blades is still problematic). At the current state of the art, a complete stage can be simulated only under some simplifying assumptions with regard to its geometry and the actual flow conditions. It is likely that in the short to medium run (3–5 years), fully three-dimensional simulations of realistic geometries can be carried out using complex turbulent models such as the LES (large eddy simulation) protocol or the time-dependent "RSM" (Reynolds stress model). Such extensive calculations require well over 10^6 grid points, solve 11 discretized differential equations per grid point on average, and can be performed today only on large, massively parallel computers.

3.9 Final Design and Costing

The two-dimensional design can be updated with the help of numerical simulations and generally results in minor modifications to the blade geometry (mostly, the tip region of the blade and the shape of the casing at the hub are affected). After some necessary testing, the design is finalized, is fully engineered, and goes into production. In this phase, detailed engineering cost accounting procedures are employed to provide an appropriately accurate calculation of the turbine production costs.

4. OFF-DESIGN CONSIDERATIONS

Steam turbines are ill suited to follow large ($>30\%$) load variations. A change in the thermodynamic parameters (pressure and temperature) causes a change in density and affects the fluid dynamics of the fluid–blade coupling. Inlet, profile, blockage, and secondary losses invariably increase if the operating point deviates from the design one. A mass flow rate reduction (or increase) at the entrance of any stage is reflected in a variation of pressures and densities at stage exit and results in an efficiency penalty as well. An exception is partial admission stages, where a rather large mass flow rate adjustment can be made by increasing or decreasing the number of active sectors.

Reaction stages can be regulated in a limited way by adjusting the interstage extractions. But this is not always possible, with extraction conditions often being dictated by process requirements. Thus, the load is usually controlled by throttling the steam upstream of the first stage. This practice is thermodynamically wasteful and in effect limits the load-following capabilities of these machines.

Part-load performance charts are easily drawn. The power delivered by a turbine is $P = m^* \Delta h^* \eta_t$, and it is experimentally verified that, for a fixed condenser pressure, throttling makes the Δh nearly proportional to the Δp. Because the stage pressure drop is linked to the stage mass flow rate by the turbine characteristic line, the partial load curve can be traced point by point. It can also be shown that a multistage turbine behaves as a subsonic nozzle near to its choking point, with its mass flow rate being nearly linearly proportional to the total expansion ratio (and therefore to the degree of throttling). Thus, the delivered power will be proportional to the mass flow rate. In the P/m plane, the line representing part-load performance is straight (in this approximation) and is called the Willans line. It will intersect the m axis at a nonzero point called the no-load flow rate, and it is completely defined by two experimental measurements that are easily performed on an installed machine. If extractions are present, the Willans line is shifted toward higher mass flow rates and the complete part-load performance map takes the appearance shown in Fig. 27. Notice that this "fan" of Willans lines can still be derived experimentally on the basis of two measurements for each extraction flow rate.

The Willans line represents a first-order approximation of the real behavior of a multistage turbine. At its lowest and upmost range, the power

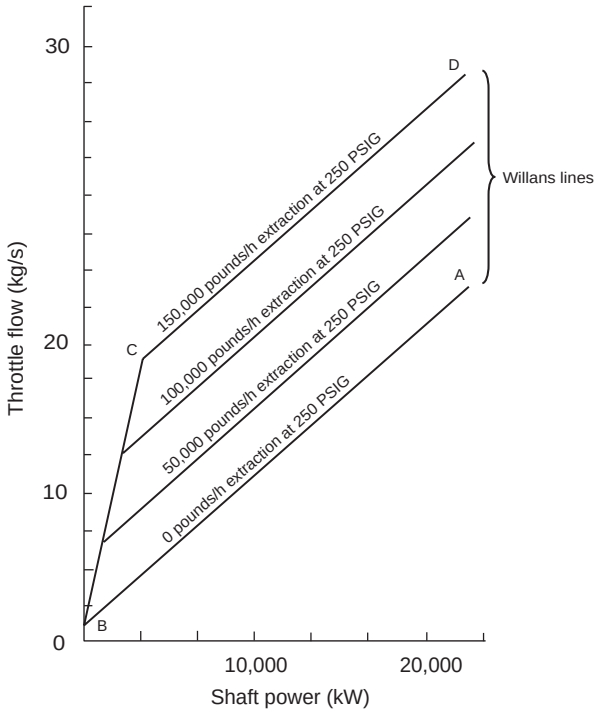


FIGURE 27 A typical extraction turbine performance map showing the Willans lines.

output/mass flow rate dependence is no longer linear.

5. CONCLUSIONS

Steam turbines are the most important machines in the energy conversion field. Their cumulative mechanical and electrical installed power exceeds that of any other type of unit by far. Their design and operational procedures are very advanced, and they can be regarded as mature technological components. They are characterized by the highest stage efficiency among all "prime movers" (with η_{stage} ranging from 0.80 to 0.95), by the highest power output per unit (up to 300 MW for a single body), by one of the highest power densities (up to 700 kW/cm^2 of active flow area), and by one of the highest mean times between failure ($>100{,}000$ h between major overhauls) and also cover an extremely wide range of applications (from a few kilowatts as auxiliary power units to more than 1000 MW in major nuclear plants). Their installed cost is relatively modest in comparison with those of their competitors in a very broad band of their operational field. They do not easily adapt to excessive load variations and, therefore, are better suited for base load plants. Steam turbine-based cogeneration plants have reached guaranteed exergetic efficiencies of more than 45%, and combined cycle units (with the steam turbine bottoming a gas turbine) nowadays exceed 50%. Although only marginal further increases in their performance can be expected to come from design improvements, a more substantial gain may be obtained by increasing the temperature and pressure of the feed steam. Steam turbines for motive applications (marine and terrestrial) are rapidly becoming obsolete because in this field they are being replaced by the lighter and more flexible diesel engines and (especially for the higher power outputs) gas turbines.

Turbine design is also a very mature engineering field. Current procedures are quite well defined and make extensive use of the automated optimal design methods. In this respect, a quantitative leap is necessary in computer speed and numerical methods (especially turbulent modeling) to attain an appreciable advantage over the current state of the art.

SEE ALSO THE FOLLOWING ARTICLES

Cogeneration • Mechanical Energy • Turbines, Gas

Further Reading

Balje, O. (1981). "Turbomachines: A Guide to Design, Selection, and Theory." John Wiley, New York.

Demeulenaere, A., and van den Braembusche, R. A. (1998). Three dimensional inverse method for turbomachinery blading design. *ASME J. Turbomachinery* **120**, 247–255.

Lakshminarayana, B. (1996). "Fluid Dynamics and Heat Transfer of Turbomachinery." John Wiley, New York.

Mathis, D. M. (1994). Fundamentals of turbine design. *In* "Handbook of Turbomachinery" (E. Logan, Jr., Ed.). Marcel Dekker, New York.

Peltier, R. V. (1994). Steam turbines. *In* "Handbook of Turbomachinery" (E. Logan, Jr., Ed.). Marcel Dekker, New York.

Stodola, A. (1945). "Steam and Gas Turbines." Peter Smith, New York.

Wislicenus, G. F. (1947). "Fluid Mechanics of Turbomachinery." McGraw–Hill, New York.

Ultralight Rail and Energy Use

JOHN A. DEARIEN
CyberTran International, Inc.
Alameda, California, United States

1. **Definition of Ultralight Rail**
2. **A Brief Discussion of Conventional Rail**
3. **Capacity and Energy Consumption of Highways**
4. **Ultralight Rail—Operational Characteristics**
5. **Energy Requirements of Ultralight Rail and Conventional Rail**
6. **Cost of Ultralight Rail Systems**
7. **Barriers to Implementation**
8. **Energy-Saving Opportunities with Ultralight Rail**

Glossary

consist An individual unit or multiple vehicles traveling in a linked set.
grade The slope, or inclination, of the guideway.
guideway The rail, support structure, power delivery conductors, and control sensors that make up the path on which ultralight vehicles move.
headway The time period between trains or individual vehicles passing a particular point on the guideway.
kilowatt-hour (kWh) Base energy unit equaling 1 kW of power used for 1 h.
passengers per hour per direction (pphpd) The measure of capacity of a rail transit system, in terms of the maximum number of passengers the system can move in one direction.

To best describe the relationship between ultralight rail and energy, it is necessary to establish a basis of understanding. This basis is best defined in terms of the energy requirements of known and familiar modes of transportation, since there are so few examples of ultralight rail in operation. This article establishes this basis in terms of conventional rail systems and, where applicable, bus, automobile, and truck systems. The basic characteristics of both conventional and ultralight rail systems are described, followed by a comparison of the two systems, with an emphasis on the energy aspects of their direct operation. The energy aspects of ultralight rail in the overall transportation system of the United States are analyzed and discussed.

1. DEFINITION OF ULTRALIGHT RAIL

Ultralight rail (ULR) is at the low end of the spectrum of weights of rail vehicles and is generally considered any vehicle with a loaded weight of less than 10,000 lbs and with a more personal type of operation. This compares with light rail transit (LRT) (www.travelportland.com/visitors/transportation.html) vehicles that can weigh over 100,000 lbs loaded and the locomotives that pull freight, commuter trains, and Amtrak trains and weigh over 300,000 lbs. The weight of a passenger vehicle on what is classified as a heavy metro rail, such as Bay Area Rapid Transit (BART) in the San Francisco Bay area (www.bart.gov) and the New York City (NYC) subway, can actually weigh less (approximately 60,000 lbs) than an LRT vehicle due to its construction and propulsion system (light refers more to total passenger capacity than individual vehicle weight).

In the ultralight range of steel-wheeled rail vehicles, there are several technologies in development but none in operation. CyberTran (www.cybertran.com), shown in Fig. 1, is an example at the top of the range (10,000 lbs, loaded). CyberTran was developed at the Idaho National Engineering and Environmental Laboratory (www.inel.gov), a Research and Development Laboratory of the United States Department of Energy, and transferred out of the government for commercialization. An Australian technology, Austrans (www.Austrans.com), is somewhat lighter, at approximately 6600 lbs, loaded. Futrex (www.futrexinc.com) is between the ultralight range and the LRT, or full-size rail vehicles, weighing approximately 20,000 lbs loaded.

 © 2004 Elsevier Inc. All rights reserved.

FIGURE 1 CyberTran vehicle.

There are several systems with vehicles in the ultralight category that utilize rubber-tired vehicles and channel-type guideways, such as Taxi 2000 (www.taxi2000.com) and the ULTra system (www.atsltd.co.uk) at the University of Bristol in Cardiff, England. The Morgantown People Mover (www.wvu.edu/~facserv/prt.htm), at the University of West Virginia, Morgantown, is just over the 10,000 lb loaded weight (9000 lbs empty with a passenger capacity of 21), but represents the best operating example of ultralight-type systems. Although these rubber-tired systems share many of the energy-saving aspects of ultralight rail, they have a larger rolling friction than steel wheel on steel rail. The reader should keep these systems in mind regarding energy discussions that are functions of size, passenger load, small infrastructure costs due to vehicle size, and vehicle movement options generated by small vehicles. An excellent coverage of ultralight systems and innovative transportation concepts under development can be reviewed at the Web site of Professor Jerry Schneider, University of Washington (http://faculty.washington.edu/jbs/itrans).

2. A BRIEF DISCUSSION OF CONVENTIONAL RAIL

Regardless of the weight of the vehicle—ultralight to heavy metro—a steel wheel on a steel rail is the most energy-efficient way to move a load over land. This efficient process was initiated by placing an iron-rimmed wagon wheel on an iron-plated beam, allowing draft animals to pull much heavier wagon loads. With the advent of steam engines to pull the wagons and flanges to keep the wagon wheels from slipping off the steel rails, the essence of the modern railway was born in the early 1800s in the coal mines of England and quickly spread throughout the rest of the world. The energy efficiency of the basic steel wheel on steel rail system was further enhanced with the advent of roller bearings on the axles and propulsion systems with electric and diesel–electric power plants.

2.1 Energy Efficiency of Steel Wheel on Steel Rail

The primary reason for the energy efficiency of steel wheel on steel rail vehicles is the extremely low rolling resistance of the wheel, even under very high loads and/or stresses. A typical loaded freight car can weigh as much as 280,000 lbs, being supported by eight wheels, with each wheel supporting over 35,000 lbs. This wheel load is supported through a contact patch approximately the size of a dime, resulting in a contact stress between the wheel and the rail of approximately 100,000 psi. Although this stress actually deforms the steel wheel and the steel rail, the energy lost in this process is small compared to the energy it takes to deform and roll a rubber tire under a heavy load. The force required to roll a standard railroad vehicle on level ground is approximately 2–4 lbs per ton of vehicle weight.

2.2 Conventional Systems: Vehicle Weights and Capacities

Conventional transit vehicles typically weigh much less than loaded freight cars, but still represent a heavy load to place on wheels and rails. A BART car weighing 80,000 lbs loaded, and supported by eight wheels, has a per wheel load of 10,000 lbs. Articulated LRT vehicles generally have three sets of four-wheeled trucks and, with loaded weights up to 120,000 lbs, this results in a load of 10,000 lbs per wheel. These vehicles roll even more efficiently than heavily loaded freight cars, but still represent a sizable load, with energy-absorbing deformation of the wheel/rail contact area. Rolling friction of typical transit vehicles is on the order of 3 lbs per ton. Figure 2 shows a typical LRT vehicle, operating at ground level in Portland, Oregon. Figure 3 shows an example of a heavy metro rail system, BART, operating on an elevated track in the San Francisco Bay area. LRT vehicles can transport over 100 passengers per vehicle and a 10-car BART train can carry over 1000 passengers.

FIGURE 2 LRT vehicle in Portland, Oregon.

FIGURE 3 BART train.

2.3 Operation of Conventional Rail Transit Systems

The operational process of picking up passengers and moving vehicles around the system is an area where ultralight systems differ significantly from conventional rail transit systems, resulting in opportunities for energy and operational cost savings.

Conventional rail transit systems generally operate as a linear system; i.e., vehicles or trains proceed down the track in sequence, with each train stopping at each station and loading/unloading passengers. The NYC subway system is a rare exception to this process, having express trains that bypass certain stations in favor of decreasing the travel time to just a few stations. The NYC system requires an additional set of tracks in order to operate in this fashion. For most conventional systems in the world, however, the trains run in single file along the system, with the average speed of the system governed by the rate at which the trains can be cycled through each station. For an average station, it takes approximately 30 s for a train to approach a station and come to a complete stop, another 30–45 s to open the doors, unload/load passengers, and close the doors, and another 30 s or so to clear the station with enough margin for the next train to come in. The smallest delay in any of these functions slows the entire line of traffic, with subsequent negative impact on the average speed of the system. Minimum headways of 1.5–3 min are the norm when safety margins are included. Ultralight systems use a different operational process that minimizes the effect of local delays, resulting in a positive effect on average system speed.

2.4 Passenger Flow Rates of Conventional Rail Systems

Capacities of conventional rail systems vary substantially, depending on the capacity of each vehicle or train, the headway between trains, and the speed of the trains. The LRT system in Portland, Oregon is made up of 2-car consists with a total capacity of approximately 320 passengers (seated and standing). At their average headway of 5 min during peak flows, or 12 vehicles per hour, this constitutes a maximum flow rate of 3840 passengers per hour per direction (pphpd). BART, utilizing 10-car consists capable of carrying over 1000 passengers and operating at less than 3 min headways, has a capacity approaching 26,000 pphpd. The Paris Metro, using 5-car consists capable of carrying 1000 people and operating at 90 s headways, has a capacity approaching 40,000 pphpd.

3. CAPACITY AND ENERGY CONSUMPTION OF HIGHWAYS

Highways can also be rated with respect to their maximum capacity and energy consumption, thus making it easier to compare them with rail as a related transportation mode, relative to cost and energy usage. A great deal of research and observation has resulted in the understanding that a multilane highway has a maximum capacity on the order of 2400 vehicles per hour per lane. Considering the fact that approximately 95% of freeway traffic consists of single-occupancy vehicles, a highway lane has a passenger capacity of approximately 2500 pphpd. One can see instances, especially on Los Angeles freeways, of groups of vehicles moving at densities of 50% or more in excess of this number, but not on a sustained basis or over an extended distance.

Considering only those vehicles that meet the CAFE (Corporate Average Fuel Economy) standards of 27.5 miles per gallon, an average vehicle capacity of four seats, and an energy equivalent of 40 kilowatt-hour (kWh) per gallon, the energy efficiency of an automobile is approximately 0.37 kWh per seat-mile. New technology vehicles such as the hybrid-powered automobiles by Honda and Nissan have decreased fuel consumption by a factor of 2 or more, for an energy efficiency of approximately 0.18 kWh per seat-mile. The ultracar technology by Dr. Amory Lovins of the Rocky Mountain Institute (www.rmi.org) promises to increase this efficiency by another 50%, or 0.1 kWh per seat-mile, by achieving a fuel efficiency of approximately 80 mpg.

4. ULTRALIGHT RAIL—OPERATIONAL CHARACTERISTICS

The most dramatic and obvious characteristic of ultralight systems is the size of the individual vehicles relative to conventional rail vehicles. Figure 1 shows the CyberTran No. 2 Test Vehicle, a 38 ft long vehicle, weighing 10,000 lbs loaded, and designed to carry 15–20 passengers. Figure 4 shows the Taxi 2000 vehicle, designed to carry 3 passengers at a loaded weight estimated to be less than 1 ton.

4.1 Computer-Controlled Systems

An important characteristic of ultralight rail system vehicles is that they are computer-controlled.

FIGURE 4 Taxi 2000 vehicle.

Although this characteristic is not unique to ultralight systems (the Vancouver, British Columbia, SkyTrain Metro system and the Paris Metro Meteor line are driverless), in the ultralight systems it is almost a necessity for any reasonable operating economics. If a driver was required in each small vehicle, the overhead costs would be intolerable, requiring trip costs typical of taxicabs without the route flexibility of taxis. It is this characteristic, however, that offers the opportunity for significant energy and cost savings due to the different operating modes possible.

4.2 Ultralight Rail Station Configuration

Two important characteristics of ultralight rail systems are the rapidity with which passengers can be dispatched toward their destination and the speed with which the passengers arrive at their destination. Both of these characteristics are due to a passenger loading/unloading process known as off-line loading and direct-to-destination routing. Off-line loading means that the loading and unloading process takes place off the main line, in a station that is reached by going through a rail switch and proceeding to a loading platform well away from the high-speed traffic of the main line. This accomplishes two important functions: (1) delays in the loading/unloading process at one station do not slow vehicles on the main line and (2) by being able to maintain a high speed between origin and destination, a vehicle has a high average speed over its route. Direct-to-destination routing means that the vehicles do not stop at intermediate stations between their origin and their destination. The ability of a system to send vehicles directly between origin and destination requires that all passengers in that vehicle are going to the same destination, a situation that can be realized only by using small vehicles in conjunction with high speed and computer-controlled direction of passengers to their proper vehicles. Computer control over the operation of each vehicle, and between individual vehicles, results in headways between vehicles of 15 s or less. Managers of the Morgantown People Mover are suggesting going to 7.5 s between vehicles on the main line, and Dr. Edward Anderson, with Taxi 2000, is proposing running at even smaller time increments (www.taxi2000.com). The effect of these short headways is to get more people to their destination quicker, with fewer energy-consuming starts and stops along the way.

4.3 Ultralight Rail Structures

An important characteristic of ultralight rail is that the structures carrying the vehicles are much smaller than the bridges, columns, foundations, etc., normally associated with conventional rail transit systems or even highway bridges. This characteristic is the one having the most impact on relative costs between conventional and ultralight rail systems. The size of structures required to support a 100,000 lb vehicle 15 ft in the air versus elevating a 10,000 lb vehicle not only has a dramatic cost effect on the construction effort, but also in the amount of system energy it takes to produce the extra material and hardware.

4.4 Power Requirements of Ultralight Systems

A hidden, yet important characteristic of ultralight systems is the low power of the electric propulsion motors required to move the vehicles around. Though rolling friction is low for all rail vehicles, the energy required to accelerate or elevate a vehicle is in direct proportion to its weight. This characteristic of having low-power propulsion motors is of most importance relative to the movement of empty or low-ridership vehicles. Conventional, large, high-capacity rail systems must move 10 s of 1000 s of pounds of hardware around the system even if they are empty or lightly loaded. Ultralight rail systems move much smaller vehicles around and then only the number of vehicles that are required to meet the passenger load demands. The advantages of this aspect of ultralight rail systems are discussed further below.

4.5 Passenger Flow Rates

Capacities of ultralight systems cannot approach the capacities of heavy metro rail such as BART, the NYC subway system, or the Paris Metro, but can provide the carrying capacity normally provided by the various LRT systems around the United States. The capacity of a system like the Morgantown People Mover or CyberTran, both having a capacity of approximately 20 passengers and operating at 15 s headways, would be 240 vehicles per hour, for a capacity of 4800 pphpd. Morgantown operators are suggesting that the next upgrade to their system will provide for operation using 7.5 s headways, a headway also evaluated for use with CyberTran. Should these headways be implemented, these two systems would provide a line capacity of 9600 pphpd.

These headways are based on the requirement to allow enough time and distance between adjacent vehicles so that should the leading vehicle come to a sudden stop, the trailing vehicle has the time and distance to stop before hitting the leading vehicle. With this requirement in mind, during high-passenger-flow conditions, it is possible to slow vehicles down and operate under headways even shorter than the proposed 7.5 s. If one reduces the headway to 5 s, a speed of 40 miles per hour (mph) represents a separation the length of a football field and the safe emergency stopping distance is less than one-half of that distance. An operational scenario like this results in a passenger flow rate of 14,400 pphpd, which rivals the advertised capacity of almost any LRT system and is far greater than any actual flow being carried by an existing LRT system. The Taxi 2000 system is proposed to operate three-passenger vehicles at a headway of 0.5–1 s for a passenger flow rate of 7200 pphpd.

5. ENERGY REQUIREMENTS OF ULTRALIGHT RAIL AND CONVENTIONAL RAIL

This section examines the major factors affecting the energy requirements of ultralight and conventional rail systems and the factors that differentiate the energy requirements of the two systems.

5.1 Basic Energy Requirements of Rail Systems

There are four primary energy requirements in the operation of both ultralight rail and conventional rail. These are acceleration, rolling friction, elevation in grade, and aerodynamic drag. Energy requirements for acceleration, rolling, and grade changes are all a direct function of the weight of the vehicle and its occupants. Aerodynamic drag is a function of size, shape, length, etc., and although it is a small force relative to conventional system acceleration and grade-climbing forces, it can be a significant energy consumer in ultralight systems where vehicle weight is one-tenth that of conventional rail vehicles.

A major expenditure of energy is required for elevation change (upward) of conventional rail vehicles and their loads and it is in this area that one of the major differences in system energy consumption can be seen between conventional systems and ultralight rail systems. It is also the

reason that conventional rail systems, especially locomotive-powered systems, are limited to rather flat grades throughout their routing. A conventional, 100,000 lb LRT traveling up a 2% grade at 60 mph requires 240 kW of power just for energy associated with the grade change, plus the energy required to overcome the rolling friction and aerodynamic drag at this speed. An ultralight vehicle loaded to 10,000 lbs would require only 24 kW to handle the same grade. In transit systems where a series of "trailer" cars are pulled by a locomotive, such as with Amtrak, all of the "grade" drag must be supplied by the traction of the locomotive. In the case of a train of 10 cars, each weighing 100,000 lbs, the locomotive pulling this train must exert a pulling force of 10,000 lbs for each degree of slope of the roadbed (0.01grade $\times$ 10 cars $\times$ 100,000 lbs), plus the adhesion required to move the locomotive (2500 lbs for a typical 250,000 lb locomotive). BART has four 110 kW electric motors per car, for a total of 4400 kW for a 10-car train. LRT vehicles typically have electric motors in the range of 400 kW per vehicle. Ultralight vehicles require motors of less than 50 kW to meet rolling, acceleration, grade, and aerodynamic energy needs.

5.2 Vehicle Weight per Passenger

For rail vehicles not pulled by a locomotive (BART, NYC subway, Paris Metro, LRT, ultralight systems), the vehicle weight per passenger is remarkably similar. LRT vehicles have an advertised capacity in the neighborhood of 160 passengers and an empty weight of approximately 80,000–90,000 lbs (depending on manufacturer and configuration), for a per passenger capacity weight of approximately 500–560 lbs. BART has a vehicle weight of approximately 60,000 lbs and a capacity of approximately 120 passengers per vehicle, or 500 lbs per passenger. CyberTran has an empty weight of approximately 7500 lbs and a seated capacity of 15–20, or a maximum per passenger weight of 375–500 lbs per passenger. The Morgantown People Mover vehicle is approximately 9000 lbs with a 21-passenger capacity, or 428 lbs per passenger.

5.3 Vehicle Weight per Passenger versus Load Factor

In addition to the similarity of weights per passenger, all of the rail systems described use very efficient electric motors in their propulsion systems, albeit of greatly different power between the heavy and the ultralight systems. Therefore, on a per passenger basis, one should not expect much difference in the energy efficiency of a heavy system and a light system when both systems are operating at or near their capacity. It is when the systems are required to operate at less than their maximum loads (which is the case for most systems having a morning and evening rush hour) that significant differences can be seen between the operation of conventional rail systems and ultralight systems.

To illustrate the effects of weight, passenger capacity, and operating process on the energy consumption of a rail transit system, assume an LRT with an empty weight of 75,000 lbs and a passenger capacity of 150 (at 166 lbs per passenger), comparing its operation with an ultralight system of 7500 lb vehicles and a capacity of 15 (to simplify the math). Table I shows the effective system weight and the weight per passenger of a conventional system and an ultralight system as the ridership drops (assuming constant frequency for the LRT and sufficient ultralight vehicles to carry the ridership).

Table I shows the dramatic effect of operating a system with small vehicles and one that allows the system control to use only the number of vehicles (capacity) required to carry the load at any one time. The LRT system is forced to provide an entire, high-capacity vehicle even when there are only a few riders, whereas the ultralight system merely provides enough vehicles to carry the riders. From Table I, it can be seen that the primary energy requirements of an ultralight system that are based on vehicle and passenger load weights range from equal at full load to only 10% as the passenger load approaches zero.

TABLE I

Effects of Load Factor on Vehicle Efficiency

Riders/ cycle	Effective system weight		Vehicle weight per passenger	
	LRT	ULR	LRT	ULR
150	100,000	100,000[a]	500	500
75	87,500	50,000[b]	1000	500
37	81,166	28,642[c]	2027	608
20	78,333	18,320	3750	750
10	76,666	9166	7500	750
1	75,166	7666	75,000	7500

[a] Assumes 10 fully loaded vehicles to carry load at 10,000 lbs per loaded vehicle.

[b] Five vehicles at 7500 pounds each plus 12,500 lbs of passengers.

[c] Requires three vehicles to carry 37 passengers.

5.4 Energy Efficiency of Off-Line Loading and Direct to Destination Routing

Another aspect of operational differences between conventional and ultralight systems that can have effects on energy consumption is the ability of ultralight systems to operate in a full-time express mode, i.e., direct to destination, as opposed to the linear operating mode of conventional rail, i.e., stopping at every station. The standard operating mode of ultralight systems is the use of off-line loading/unloading, as described in the previous section. With an appropriately designed station, an ultralight vehicle can leave any station and go directly to any other station, in either direction. A conventional linear system, however, must start at one end of the line with all of the capacity it will require to handle the heaviest traffic between any two stations on the line.

Table II illustrates the effect of this necessity for "carrying your capacity" through the entire line versus the option available with ultralight systems to place in service only the capacity needed to carry the load.

The operational efficiency of each system for this example, defined as the passenger segments required/passenger segments provided, is 46.2% for the conventional system and 93.2% for the ultralight system.

Each unit of capacity delivered carries a cost in energy as well as wear and tear on the system. With each system having a vehicle weight of approximately 500 lbs per unit of capacity, accelerating each of these extra units of capacity and carrying them through any change in elevation, plus the small amount of rolling energy, indicates that, for this example, the ultralight system is more than twice as energy efficient in carrying the required passenger load. This increased efficiency is due entirely to the operating options available with the ultralight system, since the basic weight per passenger capacity, motor efficiency, and effects that are a function of weight, such as acceleration, elevation change, and rolling, are essentially the same.

An additional effect that is not illustrated in Table II, but can be readily seen as a source of additional energy savings, is the savings associated with not having to stop at intermediate stations. When traffic becomes sufficiently large, the small vehicles can be filled at one station with passengers destined for a single other station, as opposed to conventional linear rail systems where all passengers to all stations are in the same vehicle and stop at all stops. This single-destination vehicle then proceeds directly from its originating station to its destination station, not having to expend the energy of stopping and starting at each intermediate station, and traveling between the stations at a much higher average speed. Computer simulations between direct to destination routing and linear system routing indicate an increase in average system speed of a factor of 3 (from 15.7 mph in the Portland MAX system to 51 mph in a simulated system with ultralight vehicles).

TABLE II

Conventional versus Capacity-Dictated Vehicle Dispatch

	Passenger flow				
Station	Entry	Exit	Net capacity required	LRT capacity	ULR[a] capacity
1	20	0		150	30 (2 vehicles)
2	50	10	60	150	60 (4 vehicles)
3	50	20	90	150	95 (7 vehicles)
4	70	10	150	150	150 (10 vehicles)
5	20	80	90	150	95 (7 vehicles)
6	10	50	50	150	60 (4 vehicles)
7	5	30	25	150	30 (2 vehicles)
8	0	25	0	150	
Total capacity required:			485 passenger segments[b]		
Total capacity delivered (passenger segments):				1050	520

[a] Assumes that vehicles carry multidestination passengers, stop only at stations to which passengers are going, and can reverse direction at any station.

[b] One passenger segment is one passenger transported between adjacent stations.

5.5 Energy Efficiency Potential from Intermodal Operations

As illustrated in Table I, high-capacity, heavy metro and LRT systems can be very efficient, energy-wise, when operated at high passenger loads, and conversely, can be extremely inefficient during periods of low passenger flow. Figure 5 illustrates the effect of load on energy efficiency as the

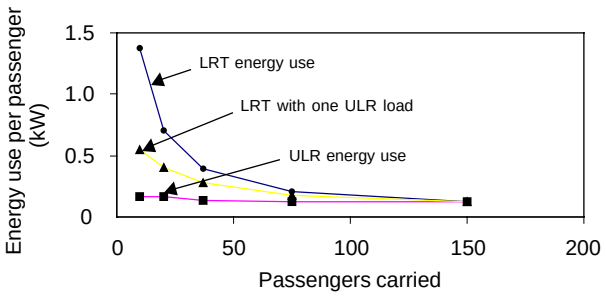


FIGURE 5 Energy use of LRT and ULR with intermodal effects.

passenger load decreases in (1) a 150-passenger LRT vehicle and (2) 10 ULR vehicles, with the ultralight vehicles being removed from service as the load decreases.

Ultralight rail systems have the potential to increase the energy efficiency of high-speed and heavy rail systems by feeding extra passengers into existing stations from outlying areas off the sides and ends of the high-capacity systems. Figure 5 provides an example of the increase in energy efficiency possible by the intermodal connection of conventional rail systems and ultralight rail, where a single ultralight passenger load is brought in from the side or end and added to the passenger load of a conventional LRT-size vehicle.

Additional passengers brought in from the side or ends of conventional systems by small ultralight vehicles not only increase the energy efficiency of the conventional systems, but reduce the need to provide additional parking in the immediate station area. Parking for the ultralight systems can be placed far from the conventional system stations in less expensive and more accessible real estate.

Conventional heavy metro and LRT systems are expensive ($50,000,000 to $100,000,000 per mile) and have a large ground footprint, two characteristics that detract from extending the systems into lower density sources of passengers. Additionally, it is not only expensive to provide parking around stations, where the land is at a premium, but 5 acres of expensive parking can hold only ~750 vehicles, approximately the passenger capacity of one heavy metro train. Therefore, the first train through the system can take all of the parking lot patrons, leaving the station isolated to many potential passengers and lowering the efficiency of the large transit system. Ultralight systems that cost $10,000,000 to $20,000,000 per mile and have a much smaller ground footprint can be extended far beyond the walking limits of conventional rail stations, bringing additional passengers to the conventional system in an energy-efficient process, making the operation of the conventional system even more efficient, since the marginal energy cost of additional passengers on a conventional system is extremely low.

5.6 Ultralight Rail and Cargo Transport

A final point in the spectrum of possibilities with ultralight rail is the transportation of cargo. Ultralight vehicles with a maximum gross vehicle weight of up to 10,000 lbs have a cargo-carrying potential of over 5000 lbs. With a rolling resistance and an aerodynamic drag of an ultralight vehicle having an average power requirement of approximately 12 kW, the energy consumption of an ultralight cargo vehicle can be on the order of 0.08 kWh per ton-mile. This value compares with a Class 8 truck that can carry 25 tons at an average diesel consumption of 6 miles per gallon, or for an energy equivalent of 40 kWh per diesel-gallon, an energy consumption of 0.266 kWh per ton-mile. Although the energy consumption of highway trucks is approximately a factor of 3 times that of an ultralight rail vehicle, the monetary economics of the process is not as disparate. For an electricity cost of $0.10 per kWh, the direct energy cost of an ultralight rail vehicle is $0.008 per ton-mile. For a diesel price of $1.50 per gallon, the direct energy cost of truck cargo is $0.01 per ton-mile, only 25% more that the ultralight vehicle cost. This cost comparison is more a reflection of the energy policy of the United States than of the energy consumption rates of the two modes of cargo transport. However, when one adds in the cost of the truck driver, approximately $0.02 per ton-mile, the monetary cost difference between the two modes of transport is comparable with the differences in energy consumption between ultralight rail and heavy trucks.

Obviously, ultralight rail systems would be best utilized in certain types of high-value cargo, such as the U.S. mail, electronics, or any item with a high time value.

6. COST OF ULTRALIGHT RAIL SYSTEMS

There are two primary areas of cost savings with ultralight systems, those being in the initial construction of the system and in the operation of the system. A further area of potential savings is in the process by which these systems are planned, designed, and constructed.

6.1 Construction Costs

The cost of any transit system is dominated by the structural infrastructure costs (~60%), followed by power systems (~20%), vehicles (~10%), and miscellaneous (~10%). Specific costs for planning, engineering, permits, etc. can be included in the individual unit costs or sometimes identified as separate unit costs.

The important item to note in the cost numbers is that the largest component by far is the structural aspect of the system, that being right-of-way real estate, support structures, bridges, rail and rail support systems, tunnels, etc. The size and cost of all of these subcomponents are driven by the size and weight of the vehicles that must be carried through the system. Not only are the size and cost of these components driven by the size and weight of the vehicles carried, but the components have a nonlinear relationship to the vehicle weight. As the weight of the vehicle to be carried increases, the size and weight of the bridge or guideway section increase due to the higher vehicle weight plus an increase in size and weight due to the increased weight of the structural section itself. This is a vicious circle that can result, in many cases, in the primary load on the structure being the structure itself. This effect can be particularly onerous in areas with high seismic activity where extra mass in elevated structures (elevated concrete guideways) can result in extremely large structures.

In addition to the nonlinearity in size caused by large structures having to support themselves, the cost of constructing these structures can have a nonlinear effect. It takes a disproportionate amount of time, labor, and construction machinery to put large structural systems in place relative to the lightweight components required for ultralight rail systems. Whereas very large components of heavy vehicle systems may have to be formed and poured in place (requiring a great deal of on-site labor and machinery), lightweight components can be mass produced off-site, trucked on-site, and lifted into place with readily available cranes and rigging.

As an example of relative weights, a 60 ft elevated steel guideway truss for CyberTran weighs approximately 3 tons, whereas the elevated concrete guideway sections typical of conventional people mover systems can weigh as much as 120 tons (prestressed section 3 ft deep and 9 ft wide). The guideway section for a Taxi 2000 vehicle would be expected to weigh even less than the CyberTran section. It is clear that a cost advantage exists in both the fabrication and erection processes between conventional systems and the ultralight systems, given that a 3-ton weight can be easily fabricated off-site, trucked to the site, and quickly lifted into place, as opposed to the same process associated with dealing with a 120-ton unit.

Structural components as low in the system as the rails also show a dramatic difference in size and costs. The Portland MAX system uses 115 lb rail (115 lbs per yard), whereas the CyberTran system is designed to use a rail weighing 12.5 lbs per yard. This approximate factor of 10:1 between the rail weights not only means a 10:1 factor in the basic material costs, but a 12 ft standard rail length weighs approximately 50 lbs and can be handled by one construction person, whereas a standard 39 ft rail section weighs approximately 4500 lbs and must be handled with machinery and at least two construction personnel.

Power delivery systems can show a similar ratio between size, and therefore cost, of conventional systems and ultralight systems. Two-car consists of conventional LRT vehicles (with 400 kW per vehicle) passing each other on a bidirectional right of way require a delivered capacity of 1600 kW at each point along the guideway. Two CyberTran vehicles (with 75 kW per vehicle) require a point capacity of 150 kW, or a factor of over 10:1 between the conventional system and the ultralight system. Lighter systems such as Taxi 2000 and Austrans would be expected to have even larger ratios.

The power ratios between ultralight systems can be even more dramatic when compared with high-speed trains, which some systems, such as CyberTran, are capable of competing with in the speed range of 125–150 mph. A French Train Grande Vitesse (TGV) has 8800 kW of motor capacity on board, which translates to the potential of 17.6 MW of power requirement at any point on the system where two trains might pass. Two high-speed versions of the CyberTran vehicles, using two 75 kW motors each, would require a point power delivery requirement of 300 kW, for a power system ratio of almost 60:1. A ratio like this results in a dramatic difference in the size and cost of power supply components such as transformers, rectifiers, cables, relays, and fuses.

Although the specific capital cost of any transit system is a function of the local conditions, it is apparent from the above comparisons that the cost of an ultralight system can be substantially less than the cost of a conventional technology system. Although a commercial ultralight system has not been placed in operating revenue service yet, a number of cost estimates have been made on these systems using conventional estimating techniques based on the use of conventional construction materials and techniques.

These estimates project a capital cost for ultralight systems in the range of $10 million to $20 million per bidirectional route mile. This compares with experience and estimates for conventional technology, LRT-type systems, in the range of $30 million to $60 million per route mile, and heavy rail systems like BART, on the order of $100 million per route mile.

6.2 Operating and Maintenance Costs

Ultralight systems have the opportunity to realize significant reductions in operating and maintenance costs due to the basic operating mode of the systems (automated system control and driverless computer-controlled vehicles) and the maintenance processes associated with small vehicles and small system components (rails, switches, etc.).

Maintenance facilities for ultralight systems can be substantially smaller and less costly than conventional technology maintenance facilities due to the much smaller size and weight of the vehicles to be handled. Buildings can be smaller and the associated lifting and handling machinery can be one-tenth the size of the machinery required to handle conventional rail equipment. The general practice in the ultralight rail concepts of keeping and storing vehicles at stations (for ready availability) results in smaller real estate needs for fleet storage, relative to the large marshaling yards typical of conventional rail technology.

6.3 Cost Effects of Construction Process

The construction procedure for conventional technology systems is a highly fractured process, even when one company takes on the responsibility for developing the complete system, such as a DBOM (design, build, operate, and maintain), because almost all of the major subcomponents (engineering, civil structure construction, vehicles, signaling, power systems, permanent way components, fare collection, etc.) are handled by different companies. This process results in excess costs for program management as well as the basic needs for profit and overhead costs by each of the subcomponent suppliers.

Significant savings in the supply of ultralight systems are possible if the developers of ultralight systems define a standardized system with standardized components available from efficient suppliers or produced internally. This process of standardizing the product and making the subcomponent supply line extremely efficient was started by Henry Ford in his automotive production process. Instead of buying many components from a variety of suppliers and using expensive custom manufacturing procedures, Ford generated many of his own components and put efficient mass production procedures in place. A very similar process was initiated by General Electric (GE) and General Motors (GM) in the production of rail locomotives. GE and GM produced a standard locomotive and the rail lines met their varied traction needs by using a varying number of locomotives, as opposed to the practice of many steam locomotive manufacturers that produced engines to the specific requirements of individual rail lines.

7. BARRIERS TO IMPLEMENTATION

There are significant barriers to implementation of any of these ultralight concepts, but none are technical in nature. The technologies associated with each of the concepts mentioned in this article are essentially off-the-shelf, with many of the subcomponents planned for use at a technical level much lower than already available and in use. The communication, computer control, and data transfer requirements of the ultralight rail systems are orders of magnitude less complex than similar functions performed continuously by the world's cell telephone operation or even a medium-sized office's computer network. The mechanics of ultralight systems are operationally similar to the systems that have been operating for 180 years and construction of the components of the system is a process within the capabilities of a broad spectrum of American and world industry.

The primary barriers to implementation are as follows:

1. The almost nonexistence of government research funds for the development and implementation of new technology
2. The reluctance (and sometimes legal barrier) to buy unproven technology or technology that is not yet in revenue service (a great Catch-22 that benefits in-place technologies)
3. The reluctance of existing industry members to take on a new technology with which they have no experience or expertise
4. The opposition by groups that view this concept as competition to their particular technology (automobiles, buses, taxis, airline, conventional rail) or profession (road construction, supplier to competing technologies, consultants for conventional technology)
5. Members of certain environmental groups who see any system that makes transportation more efficient as being a promoter of urban sprawl

Having identified all of the seemingly formidable barriers above, it is important to point out that these barriers are not new or unique to this particular technology. The diesel–electric locomotives faced similar concerns relative to the long history of steam locomotives and their technology support infrastructure. Automobiles faced opposition from those in the horse-drawn carriage trade and the manufacturers of the carriage components, as well as the legal structure concerned with public safety. Business leaders initially saw little need for telephones, considering the cost, convenience, and experience of message boys. American locomotive engineers have fought the automated control of locomotives in switching yards for years and have only just begun to accept a very efficient technology that has been used in Canada for years. One should also remember that it took George Westinghouse over 20 years—and took an act of Congress—to get air brakes on American trains, despite the obvious safety needs at the time for the system.

Dealing with these barriers to implementation is part of the development process for any new technology, and if history is any guide, an appropriate and timely technology will eventually win out.

8. ENERGY-SAVING OPPORTUNITIES WITH ULTRALIGHT RAIL

The above data and comparisons indicate that significant energy savings and cost savings are possible with the implementation of ultralight rail systems. The energy savings of ultralight rail systems relative to conventional rail systems is due primarily to the ability of ultralight systems to quickly adjust the number of vehicles, and thus capacity, to the ridership requirements. These savings are possible not only with stand-alone ultralight rail systems, but with their use as support and passenger-feed systems to high-passenger-volume conventional rail systems, such as BART, the NYC subway, and certain LRT systems.

The energy savings of ultralight rail relative to automobiles are due not only to the basic increased energy efficiency of a steel-wheeled vehicle over a rubber-tired vehicle, but also to the highly inefficient utilization of automobiles in congested commuting situations.

Table III illustrates the potential effect of ultralight rail on energy consumption in the human transportation sector of the United States economy. Table III shows the energy consumption of four transportation modes, in terms of their potential usage (kilowatt-hour per available place mile) and in terms of their actual (or best estimated) usage (kilowatt-hour per passenger mile).

TABLE III

Energy Usage of Various Transportation Modes

	LRT[a]	ULR	Bus[a]	Automobile[b]
Potential (k Wh per place mile)	0.04	0.04[c]	0.14	0.37
Actual(k Wh per place mile)	1.14	0.106[d]	4.06	1.4[e]

[a] Data from Portland MAX (1989) Special Report 221, Transportation Research Board, Washington, DC.

[b] Four passenger auto mobile getting 27 miles per gallon.

[c] Estimate based on weight and seats compared to LRT.

[d] Based on dispatching out at 60% occupancy and deadheading back empty.

[e] Based on 1.05 passenger average occupancy—units of kWhrper/passenger mile.

Table III shows that conventional rail and ultralight rail are essentially the same in terms of energy efficiency when operated at their maximum potential, but that the capability of ultralight rail to leave unneeded capacity at the station has a dramatic effect on energy efficiency at less than maximum loads. Table III indicates that in situations where ultralight rail can augment or be substituted for other transportation modes, significant energy savings can result, such as with LRT (91%), buses (97%), and personal automobiles (92%). Although there will always be a necessity for buses and automobiles, the energy consumption numbers indicate that rail in general, and ultralight rail in particular, should be implemented as widely as possible for the energy-saving potential the technology offers.

Significant energy savings are possible with the implementation of ultralight rail systems and, in addition, the energy source for these systems, electricity, can be generated from resources under domestic control (coal, hydroelectric, nuclear, natural gas) as opposed to the foreign oil base of the majority of the transportation infrastructure of the United States.

SEE ALSO THE FOLLOWING ARTICLES

Intelligent Transportation Systems • Magnetic Levitation • Transportation and Energy, Overview • Transportation and Energy Policy

Further Reading

Armstrong, J. H. (1994). "The Railroad—What It Is, What It Does—The Introduction to Railroading." 3rd Ed. Simmons-Boardman, Omaha, NE.

Charles River Associates. (2000). "Diesel Technology and the American Economy," Report D02378-00, Charles River Associates, Washington, DC.

Irving, J. H. (1978). "Fundamentals of Personal Rapid Transit." D. C. Heath and Co., Lexington, MA.

Tayler, A. (1992). "HiTech Trains." Chartwell Books, Secaucus, NJ.

Transportation Research Board. (1989). "Light Rail Transit—New System Successes at Affordable Prices," Special Report 221. Transportation Research Board, Washington, DC.

Transportation Research Board. (1994). "Highway Capacity Manual," Special Report 209. Transportation Research Board, Washington, DC.

Transportation Research Board. (1991). "In Pursuit of Speed—New Options for Intercity Passenger Transport," Special Report 233. Transportation Research Board, Washington, DC.

United Nations Energy Agreements

CLAIRE J. NORRIS
Boston University
Boston, Massachusetts, United States

1. Background to Type II Agreements
2. Analysis of Type II Energy Agreements
3. Evaluation of Type II Energy Agreements

Glossary

Commission for Sustainable Development (CSD) The United Nations entity tasked to manage the submission of Type II agreements arising from the World Summit on Sustainable Development.

nongovernmental organizations (NGOs) Played an important participatory role in the World Summit on Sustainable Development.

sustainable development Development that meets the needs of the present without compromising the ability of future generations to meet their own needs.

Type I agreements Agreements between governments to implement sustainable development.

Type II agreements Partnerships between a combination of governments, NGOs, and the private sector for the implementation of sustainable development.

World Summit on Sustainable Development (WSSD) Held in Johannesburg, South Africa, August 26 to September 4, 2002.

Sustainable development has been on the international environmental agenda since the 1980s. One of the most commonly cited definitions of sustainable development comes from the World Commission on Environment and Development (WCED), which states that "sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs." Developing the generation and usage of energy in a sustainable manner has been a primary focus of sustainable development initiatives. The involvement of private-sector energy companies has evolved during progressive global conferences, in the context of increasing globalization and a heightened focus on trade and development over pure environmental issues.

1. BACKGROUND TO TYPE II AGREEMENTS

During the United Nations Conference on the Human Environment, held in Stockholm in June 1972, sustainable development as an aim had not yet been identified, but the significant impact of business in environmental matters certainly had been. Thirty years ago, business was viewed as a problem to be solved by nation-states, certainly not as a stakeholder to be involved in any negotiation on international environmental treaties. The private sector was practically unanimously perceived as being on the side of development, encouraging increasing consumption of natural resources and therefore against the environment almost by definition.

Twenty years after UNCHE, the United Nations Conference on Environment and Development (UNCED) met in 1992 at Rio to produce fundamental principles and a program of action for achieving sustainable development, crystallized in the Rio principles and Agenda 21. In this forum, business as a driver of energy consumption was still viewed as the problem, but by this stage there existed a degree of grudging acknowledgement that to achieve the aims of sustainable development, there would have to be some level of dialogue between business and the nation-states. This acknowledgement was partially fueled by the increasing global power of large multinational corporations, some of whose individual economies were more significant than those of some states. In this context, any individual

 © 2004 Elsevier Inc. All rights reserved.

state had become unable to monitor and therefore control the activities of registered companies.

This push factor played a part in forcing nation-states to accept business to the table, albeit a table geographically removed from that of the 184 governments at Rio. This was documented in Agenda 21, which identified business and industry as one of nine major groups that can empower communities to mobilize for the environment. The North/South negotiating dynamic also increasingly played a part in involving business with sustainable development negotiations, as some countries from the South viewed their participation in potentially GDP-reducing environmental projects as contingent on support for development to alleviate poverty. Governments from the North possessed capacity to provide that support, but the private sector was vital to any implementation, which provided additional pressure to include business in the dialogue. The ecological interdependence of development and the natural environment became replicated in the negotiating dynamics of UNCED. Acknowledgement of this interdependence can be observed in the Rio Declaration on Environment and Development, which states in Principle 4 that "in order to achieve sustainable development, environmental protection shall constitute an integral part of the development process and cannot be considered in isolation from it."

After Rio, environmental issues became firmly established as global problems that require real solutions, and voices for business took the position that it is better be at the table and provide appropriate input rather than take an isolationist stance that may lead to the imposition of suboptimal legislation later down the line. With the formation of organizations such as the World Business Council for Sustainable Development (WBCSD) in 1995, the private sector began to develop a coherent voice on sustainable development issues, although the cooperation between those organizations and governments was limited. Toward the end of the 1990s, interest in business as a lever for promoting sustainable development began to increase, although environmental issues themselves fell down the list of government priorities in favor of global issues such as trade. Membership of the WBCSD grew to more than 170 and the organization began to develop sustainability reports on the energy sector, which were submitted to global conferences.

As the World Summit on Sustainable Development (WSSD) 2002 approached, the Preparatory Committee began to realize that a new type of agreement would be necessary at Johannesburg to enable implementation of sustainable development objectives. The low level of implementation since Rio was noted at the second session of the Preparatory Committee (Prepcom) for the WSSD (New York, January 2002), which also recognized that the WSSD would be unlikely to achieve specific commitments between governments in that global political climate. An outline of the nature of a second type of partnership was prepared, inviting governments, international governmental organizations (IGOs), and major nongovernmental organizations (NGOs) to put forward preliminary partnership proposals at the third (New York, March 2002) and fourth (Jakarta, May 2002) Prepcoms.

Proposals were required to be aimed at supporting practical implementation of Agenda 21 and sustainable development activities in developing countries. Additionally, proposals were to be international in scope, consist of more than one partner, and be specifically developed for the WSSD. This new type of agreement, labeled Type II, was intended as informal, voluntary commitments, generally between a combination of governments, business, NGOs, and IGOs. In contrast, Type I agreements are specific commitments between governments. The chairman's notes from the second Prepcom also stated that "partnerships and initiatives to implement Agenda 21 are expected to become one of the major outcomes of the World Summit on Sustainable Development." As these partnerships are a specified major outcome of a global summit attended by 189 governments, more than 700 NGOs, and hundreds of international corporations and quasi-governmental interests, the success of generating partnerships and whether those partnerships retain the characteristics intended by the WSSD both merit investigation.

Advantages of Type II agreements over Type I agreements cited during the later Prepcoms include the reduction of time taken to develop agreements, leverage of nongovernment resources and increased credibility, participation, and range of expertise. By the fourth Prepcom (May 2002), more than 400 stakeholders from 50 countries had collaborated to create 14 draft proposals, energy being one of the four broad issues addressed. At Johannesburg, more than 170 Type II proposals had been developed for presentation. Principle 27 of the Johannesburg Declaration on Sustainable Development stated, "We agree that in pursuit of their legitimate activities the private sector, both large and small companies, have a duty to contribute to the evolution of equitable and sustainable communities and societies." This position is more collaborative than in

either of the previous two declarations and may represent the balance point of business involvement in negotiated sustainable development. Since the WSSD, updates to both initial proposals and completely new proposals have been added to this set of agreements, and there is no submission deadline.

2. *ANALYSIS OF TYPE II ENERGY AGREEMENTS*

Some analysis of Type II agreements has been carried out by the Commission on Sustainable Development (CSD), which emerged from Agenda 21 after UNCED and is now designated to oversee progress on these agreements. As these partnerships were a last minute addition to the WSSD agenda, promoted as a significant outcome to the summit once it became clear that no new intergovernmental agreements could be achieved, it is important to establish to what extent they can add value to the quest for sustainable development. A detailed database analysis of Type II partnership agreements provides some level of insight into this question, using three CSD lists of Type II Agreements. The first was submitted as of August 16, 2002, for discussion during the conference, containing 172 agreements, and the second was published on April 24, 2003, containing 266 agreements. The CSD then reviewed this April list to ensure partnerships complied with guidelines laid out in their Implementation Track for Agenda 21 published in May and redrafted an approved partnership list of 232 agreements in June. This Implementation Track incorporated the Bali Guiding Principles and General Assembly resolution 56/76.

2.1 Partnership Aims

The majority of Type II agreements have multiple aims, and there are a number of ways to cluster the data set. The information can be assessed based on the target of the partnership (e.g., energy), the means of implementing the partnership (e.g., technology transfer and capacity building) or the geographic focus of the agreement. Africa and small island developing states (SIDS) were singled out by the CSD as geographic categories meriting primary aim status. For clarity, these aims have been hierarchically mapped into broad categories (Fig. 1). Tertiary aims of the agreements were then identified using information provided in the WSSD Type II Partnership lists and additional data provided by multiple online forms as submitted by individual partnerships.

Ninety-three agreements approved to date primarily characterize themselves as managing the natural resource base, with 88 implementation

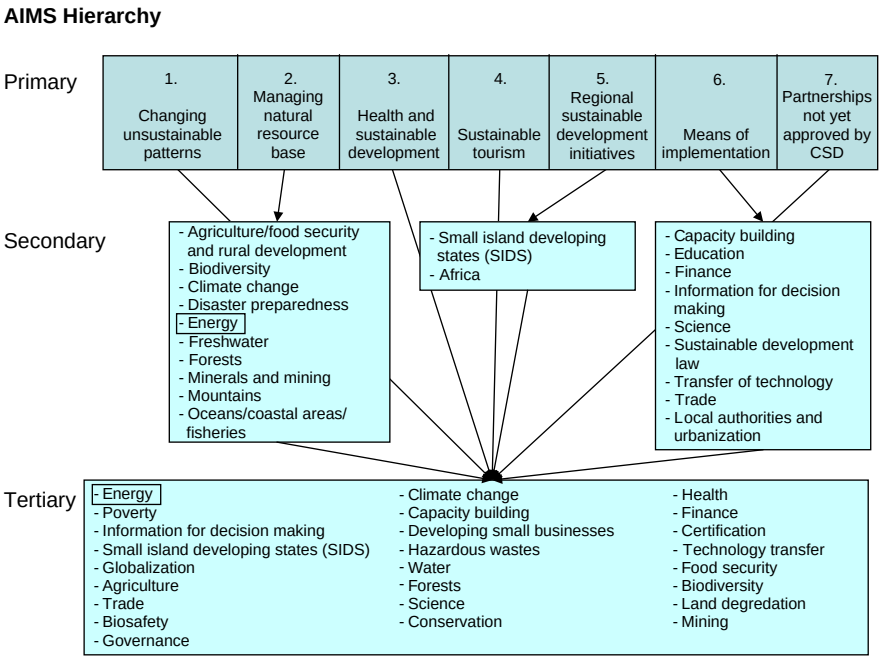


FIGURE 1 Flowchart illustrating the hierarchy of aims within which Type II agreements can be analyzed and the positioning of energy agreements within that hierarchy.

agreements as the second most frequent category (Fig. 2). Together, the categories of Managing Natural Resources and Implementation make up over 57% of the partnership proposals submitted to date and 70% of the 232 successful partnerships. Breaking these categories down into secondary aims demonstrates that energy is a major focus of natural resource management, and energy issues are explicitly addressed by 39 approved partnerships. Freshwater, agriculture, and oceans each comprise 15% of natural resource agreements. Biodiversity and climate change have 9 and 7 agreements, respectively, with disaster preparedness, forests, mountains, and mining each comprising of fewer than five agreements. The significance of African sustainable development initiatives in the third most frequent category of partnerships reflects the numerous sustainable development challenges faced by Africa. Successive United Nations Development Programme (UNDP) Human Development Reports document the challenges faced by sub-Saharan Africa, including low rates of access to electricity, low energy consumption per capita, declining income per capita, and information-poor schools and hospitals, the breadth and depth of which are unrivaled by any other continent.

Strengthening of local authorities (which includes urbanization and city planning partnerships) and provision of education are the methods by which more than half of the implementation agreements are to be enacted. Provision of information, science, trade, and capacity-building are also important implementation categories, with sustainable development law, finance, and technology transfer each comprising fewer than five agreements. The focus appears to be on implementation by empowerment, organization, and information (local authorities, education, capacity-building) rather than implementation by the application of explicit tools (finance, trade, law).

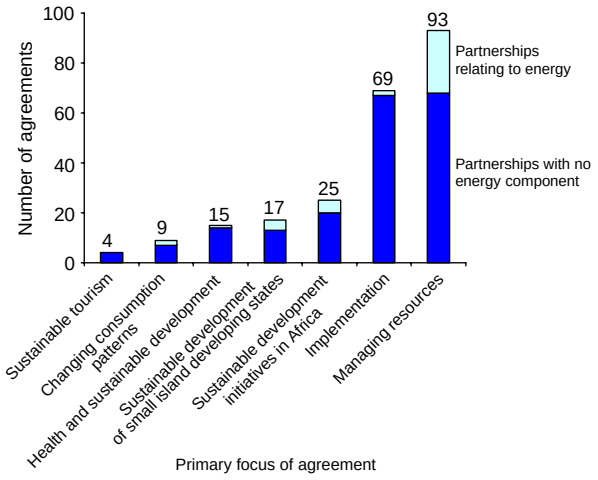


FIGURE 2 Bar chart displaying the primary focus of Type II agreements and highlighting the relative importance of energy as a sustainable development aim. Number of agreements (y axis) is plotted against primary aim of agreement (x axis).

The majority of tertiary aims appear to focus on implementation issues, the three most frequent tertiary aims being capacity building, technology transfer, and information for decision making, in that order. Poverty is a category not explicitly covered in primary or secondary aims and is the most frequent nonimplementation tertiary aim, stated by 48 (over a fifth of) Type II partnerships. Many sustainable development projects appear to have the underlying aim of poverty reduction while tackling primary natural resource management issues including energy, agriculture, freshwater, and biodiversity. Education partnerships have a large number of tertiary aims, reflecting the wide range of sustainable development topics that can be taught and the multiple benefits of education. The importance of both poverty eradication and education are reflected as two of the UN Millennium Development Goals: to reduce by half the proportion of people who live on less than a dollar a day and to ensure that all boys and girls complete a full course of primary schooling.

A number of the so-called small island developing states projects touched on similar concerns as those within the Oceans bracket, which enhances the importance of marine issues. It could be hypothesized that many of the challenges faced by SIDS stem from concerns on climate change, which is directly impacted by sustainable energy development initiatives. However, climate change does not feature highly on list of SIDS partnership tertiary aims. This may reflect the possibility that only a finite number of scientific partnerships on climate change are required, but that a host of development projects will be required to support the SIDS if climate change causes sea levels to rise significantly.

Although all partnership categories are weighted to target developing countries, local authority and climate change partnerships have the highest developed country target of approximately 28%. This does not reflect the WSSD aim that "only those partnerships/initiatives that … are aimed at sustainable development activities in developing countries … would be announced during the official partnership events." However, climate change and the interlinked energy consumption trends remain fundamentally global issues, even if initiatives to tackle the problem are based in developed countries.

2.2 Geographic Focus

Within this discussion, the North refers to developed countries and the South refers to developing countries and countries in transition. This differentiation is primarily but not solely economic, and there are more subtle geographic, linguistic, and cultural linkages that also define why the South (e.g., the G77 group of 77-plus countries) often negotiates as a caucus in the UN system.

More than half the agreements are targeted at enhancing the sustainable development of countries from the South (165 of the partnership proposals), with another third (98 of the partnership proposals) targeted globally. The global partnerships most commonly focus on health, education, and local authority development issues. Only 6% of the partnerships appear to be focused solely at countries in the North, which does not match the WSSD intentions of Type II agreements focusing on developing countries.

Over twice as many NGOs from the North submitted agreements as NGOs from the South (Fig. 3). Given the larger number of well-funded NGOs situated in the North, this is to be expected, although it may be of concern given the potentially remote nature of project management. However, all of these partnerships are open to approaches from other organizations, and they are continually evolving. Twenty partnerships already involve NGOs from the North and South working in partnership, and many North-based NGOs may have country offices in the South, staffed by people from the South.

Of the 39 approved agreements relating to energy, six have participants from the South, Indonesia being the only active lead government in an energy agreement from the South. Tanzania is working in partnership with Germany, and Mexico is working in partnership with Australia on sustainable energy projects. The Pacific Islands Energy for Sustainable Development agreement, run by the Pacific Islands Forum Secretariat, is based in Fiji, and from South Africa two energy projects have been submitted by a collective of African governments and the South African branch of the World Conservation Union (IUCN). When compared to the average level of intergovernment collaboration between North and South (4% of Type II partnerships), that for energy agreements is relatively high (8%), which is a positive indicator of cooperation within the sustainable energy sector.

2.3 Source of Partnerships

The high proportion of 1 partner only agreements (205 agreements) suggests they are in the process of evolving (Fig. 4). The number of statements

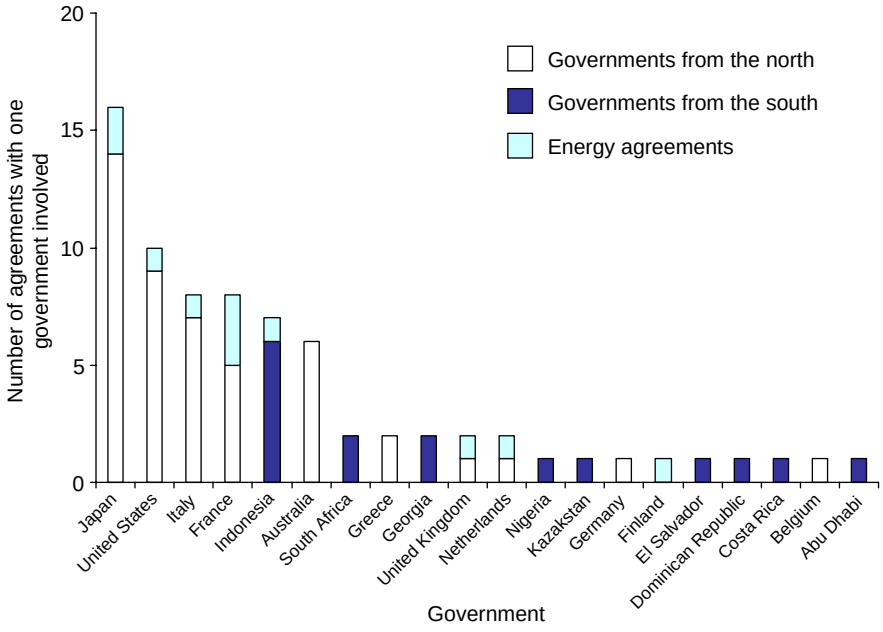


FIGURE 3 Bar chart illustrating the source of partners involved in Type II agreements, bisected into the categories North and South. A bar chart further illustrates the degree of government involvement in Type II agreements, plotting number of agreements (y axis) against government (x axis).

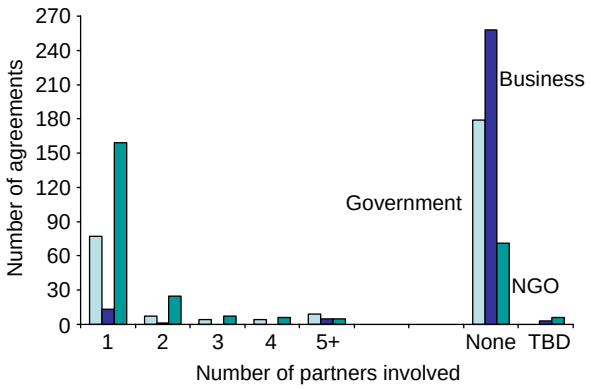


FIGURE 4 Bar chart illustrating the nature of partners involved in Type II agreements (government, business, and NGO) and highlighting the nature of partners involved in agreements relating to energy issues. Number of agreements (y axis) is plotted against number of partners involved (x axis).

suggesting that further partners are being sought (41 agreements are seeking business partners) further indicates this evolution. Nearly 70% of agreements involving NGOs do not involve other partners, and where governments are involved, 61% operate alone. Businesses are only explicitly involved in 10 agreements analyzed to date, 9 of which involve other partners. Energy enjoys the highest level of business involvement of any sector, present in 6 of the 39 approved energy partnerships. Types of business range from government energy utilities (e.g., Eskom and NamPower) to relatively small northern business interests (e.g., Velo Mondial).

At least one UN program is involved in approximately one-quarter of all agreements (66 agreements), suggesting that the IGO instigating Type II agreements is keen to provide a critical mass of initiatives. The agreements involving UN bodies are more likely to involve multiple partners (48% versus 26% overall) and appear to be more developed in other ways (target-setting, etc.). It is also evident that UN organizations provide funding for a number of NGO initiatives even when they are not explicitly involved in the Type II partnership. The number of parties involved in any agreement will also relate to the ability of the instigating parties to attract new participants.

Education initiatives have garnered the highest number of partners per agreement to date (see Fig. 5) and appear to be highly collaborative in nature. On average, Type II agreements each involve just under two partners, with agreements ranging from 1 partner only to 29 partners, as in the case of the Global Higher Education for Sustainability Partnership (GHESP). NGOs (including IGOs and other

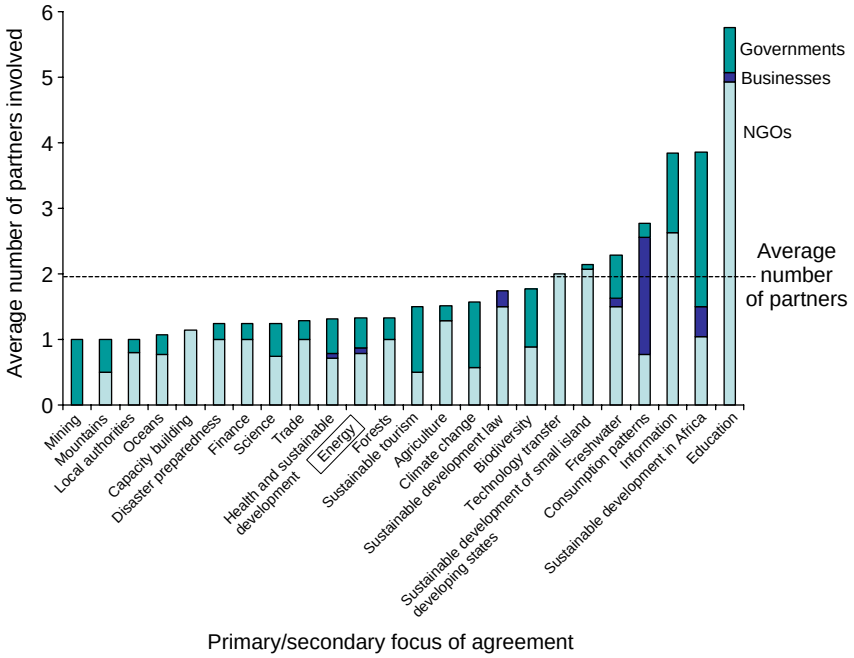


FIGURE 5 Stacked bar chart illustrating how the nature of partners involved in Type II agreements (government, business, and NGO) varies depending on the primary aim of the agreement. Number of partners involved (y axis) is plotted against primary aim of agreement (x axis).

international organizations) are more actively involved with Type II partnerships than governments, and they are far more actively involved than businesses. Capacity building and technology transfer partnerships solely involve NGOs. The majority of NGO/IGOs (153 of 181) are involved in just one agreement. The most active organizations with more than one agreement are fairly evenly divided between UN organizations, UN-influenced organizations, and organizations relatively independent from the UN. The United Nations Environment Programme (UNEP) and UNDP are the two most active organizations, and their agreements span a wide variety of sectors (>6 sectors each) and geographies. Both UNEP and UNDP have a significant amount of sectoral overlap, and they both focus on energy issues, in addition to biodiversity, freshwater, and ocean projects.

Although the three most active NGO/IGOs are UN-run or related, three other active organizations are not directly affiliated to the UN. IUCN′s 7 agreements concentrate on biodiversity and sustainable development issues in Africa, particularly regarding sustainable development law and information provision for decision making. The Caribbean Committee (CARICOM) and the Pacific Islands Secretariat each put forward seven partnerships, all of which relate to issues affecting the SIDS for their specific regions.

At a tertiary level, the aims of thse agreements address climate change and energy issues via capacity building. The Stakeholder Forum for Our Common Future has produced agreements across a wide range of sectors and is particularly active in inviting other organizations to participate.

In summary, NGO/IGOs appear to be most actively involved in forming Type II partnerships, particularly those with a UN affiliation. The international nature of these organizations makes it hard to differentiate a North/South bias, as many of these partnerships are initiated at a country office level. For example, IUCN′s South Africa office initiated the "Johannesburg Climate Legacy" partnership, although IUCN HQ is based in Switzerland, part of the North. Government involvement can be more clearly differentiated between North and South, and of the 200 cases of government partnership involvement, 91 are governments from the North and 109 from the South. Business has shown what could be considered a disappointing lack of involvement in Type II partnerships, with only 23 cases of direct business involvement. When launching the Commission on Private Sector and Development on July 25, 2003, Kofi Annan, secretary general of the UN, stated that "the birth of this Commission is yet another illustration of the rapidly growing partnership between the United Nations and the private sector … it also underscores the importance of that partnership in our work to reach the Millennium Development Goals … we cannot reach these goals without support from the private sector." In this light, it is notable that the private sector has thus far failed to support UN-instigated Type II partnerships.

2.4 Funding Arrangements

By August 16, 2002, only 5% of the Type II proposals made any mention of project funding issues, and less than 1% stated figures. Budget considerations would undoubtedly be a sensitive aspect of these proposals, and reasonable certainty would be required from potential funders before figures could be realized. Prior to the WSSD, the only two agreements containing specific funding instructions were the "Awareness Raising and Training on Sustainable Consumption and Productio partnership led by the UNDP and the Johannesburg Climate Legacy partnership led by the IUCN. These partnerships specified $2 million over 3 years and $5 million over $5\frac{1}{4}$ years, respectively, implying project burn rates of $670,000 p.a. and $950,000 p.a., respectively. These results should not be viewed as valuable or representative, given the small sample size. The funding for the first agreement came from the UNDP budget, and the second agreement planned to raise funds from WSSD participants to offset the environmental cost of Johannesburg, specifically referring to energy costs, although progress here is unclear in subsequent partnership documents. Within the partnership documents, the only specific funding information added after the WSSD date is for the United States. Water for the Poor Signature Initiative, which commits $970 million over 3 years (2003 to 2005), implies a burn rate of $323 million p.a. This is a funding level more than a factor of 100 greater than the previous two partnerships discussed. Partnership documents appear reluctant to specify exact funding terms and sources. This may be due to confidentiality issues or concern that publicizing a funding source may lead to unsolicited approaches to that source from many other partnerships.

The CSD researched the funding situation of the partnerships, and claimed that "nearly two thirds of the approved partnerships have funding, either to carry out initial phases or for the entire initiative." The analysis concedes that not all submissions

published figures of funds necessary to carry out projects, and the analysis appeared to utilize information that is not publicly available. Based on that information, the CSD estimated that $1,250 million had been committed by April 29, 2003, with a further $120 million in negotiation with potential donors. Of the $1.25 billion in secured funds, $977 million (78%) is accounted for in just three partnerships previously discussed, which implies that the majority of partnerships are yet to secure significant funds. Given the sums involved, it may be expected that this may take some time. Of the remaining funds, there is no specification as to which partnerships received what proportion of these secured funds, over what time period the funding is committed, or whether this is genuinely new funding or reallocated budget from existing organizations.

It is worth noting that during the WSSD, German Chancellor Schroeder promised €500 million over 5 years to promote renewable energy in developing countries. This new money was not classified within Type II agreement documentation.

2.5 Length of Agreements

Approximately 80% of the agreements analyzed state project length. Some agreements make qualitative statements suggesting the project may continue after a stated end date, although given the high level of uncertainty this element has not been included in the analysis. The average life span of a Type II agreement in this sample is 5.3 years, and 60 projects have no end date (Fig. 6). Of the 39 agreements concerning energy, 11 have no end dates and the average stated life span of the agreements is slightly longer than the overall figure at 5.7 years. The longest projects center on the SIDS and changes to sustainable development law. Agreements to address disaster preparedness and financing have the average shortest time span at 2.25 years.

2.6 Target Setting

Less than 10% of agreements contained quantifiable targets as of August 16, 2002. By June 2003, every partnership had stated targets, although few were easily quantifiable. This is one area which has certainly been the subject of updates within the agreements on the WSSD Web site, as project plans are gradually further developed. Targets involve the production of documents, the launch of new products, the establishment of financial funds and the training of personnel. Successful completion in many of these areas is hard to assess, although a qualitative target providing focus for a project may be preferable to no target at all. Agreements with targets may be expected to operate within a shorter and more focused time frame.

To what extent partnership targets can be quantified also appears to vary by the focus of the agreement. By examining partnership agreements, it becomes apparent that partnerships involving energy,

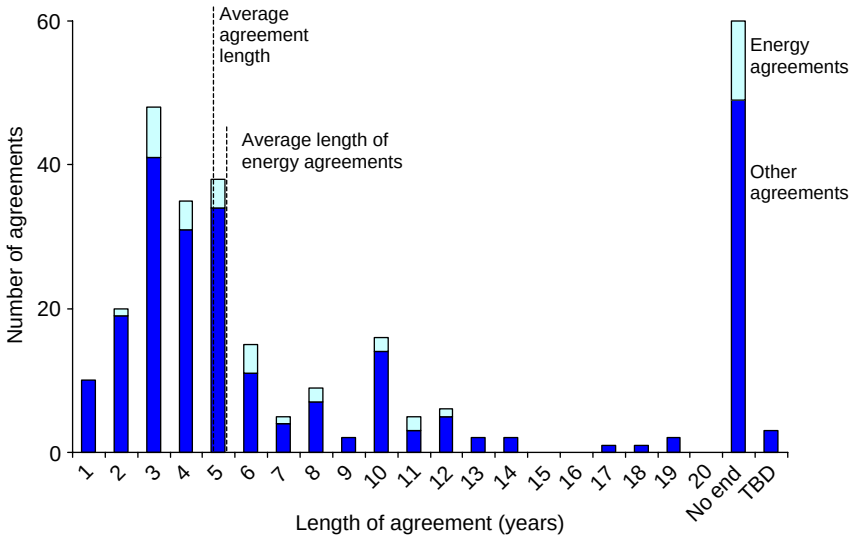


FIGURE 6 Bar chart illustrating the typical length of Type II agreements, highlighting the average length of those agreements relating to energy issues. Number of agreements (y axis) is plotted against length of agreement (x axis).

consumption patterns, and trade have more numerical targets than partnerships involving information for decision making and capacity building.

2.7 Originality

It is not possible to precisely identify recycled projects that have been rebranded as Type II agreements with no obvious additional investment. Projects run solely by government departments with similar names to the project and projects with start dates prior to 2002 provided indications that the project might be recycled. Research suggests that the large majority of Type II agreements are genuinely new partnerships, created in response to the WSSD. Approximately 217 (78%) of projects submitted are likely to be genuinely new agreements, and energy agreements appear to reflect this ratio. It is interesting to observe whether the new agreements possess any markedly different characteristics to the recycled agreements. The most striking indicator lies in project length. Also, 47% of recycled projects have no end date, whereas only 15% of new initiatives are so long term. Given that Type II agreements are a relatively new concept, statements regarding indefinite project length might seem optimistic had the project not been running prior to the commencement of Type II agreements. The sectors most prone to recycling of initiatives include mining, forests, sustainable development law, and technology transfer initiatives. Over half of the partnerships in each of these categories appear to be recycled; however, these categories do have lower than average sample sizes.

3. EVALUATION OF TYPE II ENERGY AGREEMENTS

3.1 Governments Must Continue to Pursue Type I Partnerships

Type II partnerships could only really be considered an abdication of responsibilities when their impact on the level (quality, number, etc.) of Type I agreements can be assessed, and it is too early to assess that impact. If governments are substituting effort from Type I to Type II agreements for sustainable development, then initial trends in Type II agreements may give cause for concern. As it stands, only 71 of the 189 governments attending the WSSD have any involvement in Type II agreements, and 38 of these governments are only involved in one partnership.

Only a handful of governments have shown significant leadership in creating cooperational Type II agreements. These include Japan, the United States, Italy, France, Indonesia, and Australia by order of partnership number. However, some governments appear to be using Type II to showcase a set of standalone initiatives that appear to be more prescriptive than cooperational in nature and thus do not serve to replace any type of collaborative agreement. Governments from the South appear to be participating in terms of collaborative effort, although individually they are not putting forward as many agreements as their counterparts in the North. Certain important governments (in terms of population) appear to be entirely absent from Type II agreements, including China and India. Type II agreements will gain strength from a critical mass of participants, and the noninvolvement of too many major players may serve to discredit these initiatives.

Smaller countries (in terms of economic and political clout) appear to be waiting for larger countries to take the initiative and then sign up to the outcome. This is particularly common in Africa, South America, and the states of the former Soviet Union. Countries from the South often experience resource constraint while negotiating Type I agreements, which may unfairly bias negotiation outcomes, and there may be a parallel within Type II agreements. If countries from the South wait to be offered participation in preprescribed Type II agreements, then they will lose another opportunity to lay down an agenda for sustainable development. Type II agreements may be just beginning a long process of evolution, and projects targets are far from clear. Given these concerns, work on Type I agreements should proceed as usual, until the success of Type II agreements can be assessed in the medium term.

3.2 Lack of Business Involvement

Prior to the WSSD, a number of articles were published by civil society organizations voicing concerns that the focus on Type II public-private partnerships could "leave the basic needs of billions of poor to the market forces." During the WSSD, articles in the mainstream press suggesting that Type II promises by business were unlikely to be met further discredited Type II agreements as a forum for business to contribute to sustainable development. Business is often criticized for not acting in a sustainable manner, but then accused of greenwash if it attempts to initiate sustainable development initiatives and publicize them. The bottom line for

business should be to maximize shareholder value; on some occasions this can be achieved in a sustainable manner, and on other occasions it cannot. Type II agreements could have provided a forum for business to learn about sustainable development approaches and how they can add to shareholder value, but the negative publicity associated with business involvement in Type II agreements appears to have discouraged business participation. A large number of businesses may have attended the WSSD, but very few followed this attendance through with concrete participation in sustainable development partnerships. This lack of business participation may be one of the most disappointing aspects of WSSD Type II agreements, although energy agreements provided the highest level of business involvement observed.

On July 25, 2003, Kofi Annan, Secretary General of the UN, stated that "the birth of this Commission on Private Sector and Development (CPSD) is yet another illustration of the rapidly growing partnership between the United Nations and the private sector … it also underscores the importance of that partnership in our work to reach the Millennium Development Goals … we cannot reach these goals without support from the private sector." The CPSD may help business undertake standalone sustainable development initiatives, but it does not hold the potential that Type II agreements once held for engendering collaboration and understanding between the state, corporations, NGOs, and IGOs. UN organization cannot aim to involve business in sustainable development initiatives if there is not a positive communications forum between business, government, and civil society.

This situation is in danger of becoming self-reinforcing. As business fails to find support from NGOs, it then pursues sustainable development initiatives within business-only organizations, such as the WBCSD and Business Action for Sustainable Development (BASD), which in turn are criticized for their insularity and business focus. Meaningful and mutually respectful communications and resulting collaborations between business and civil society are required. Type II agreements are missing the opportunity to achieve this and will require significant work by the CSD to recapture this opportunity. Credibility in Type II agreements must be regained to minimize the reputational risks faced by business in its involvement with collaborative sustainable development initiatives. This can be partially addressed by target setting and by tracking the progress of partnership projects.

3.3 Lack of Quantifiable Targets, Tracking and Accountability

Type II voluntary partnerships have not yet been in any way rigorous in setting out clear targets, timetables, and means to measure results consistent with sustainable development goals of economic efficiency, environmental integrity, and social equity. This could be anticipated as, early in the evolution of these agreements, it may be foolhardy to expect that project planning is sufficiently rigorous to allow for significant levels of accuracy in target assessment and therefore accountability. In this way, Type II agreements could be viewed in a similar light as conventions, which essentially act as empty shells within which greater levels of detail are gradually added. This may reduce the accountability of these agreements, but given their long-term nature (average 5.3 years), the retention of budget flexibility may be preferable to demanding specific goals at the outset. A logical compromise may be the introduction of short-term targets within these agreements (e.g., over 1 year rather than the full average 5.3 years), which can be used to demonstrate progress but would not set the direction of projects in stone. These targets could also help fundraising, as successful partnerships could showcase well-conceived and structured sustainable development projects, demonstrating progress against targets in order to pitch for additional investment from funding sources.

3.4 Transparency Is Required on Partnership Funding

Little quantitative information is available on the funding sources for Type II partnerships. In the small number of agreements that refer to funding mechanisms, the funding sources are the United States, UNDP, and IUCN offset fundraising from Johannesburg participants, totaling $977 million over 3 years. It is unclear whether the U.S. funds are restated from existing USAID budget or entirely additive. The UNDP source may take funding away from other projects, although the IUCN offset fundraising does appear to raise a separate, additive funding source. None of these sources appears likely to add to developing world debt, which was one concern cited by WSSD participants. However, the nature of these funds is not specified and information such as rates of interest, if any, is not addressed. Notably, €500 million over 5 years promised by German Chancellor Schroeder for promotion of renewable energy in

developing countries was not classified as a standalone Type II agreement.

To execute on these partnership agreements, funding will doubtless be required, and information on that funding should be made publicly available. The success of Type II partnerships will be assessed on their ability to raise additive funds, meaning funds that are not already dedicated to non-Type II sustainable development initiatives. Attention should also be paid to funds that are transferred between sustainable development-focused NGO/IGOs to fund partnerships, as this could lead to duplication when assessing the total amount of Type II funding generated. Fundraising cycles frequently take many months, if not years, to complete, therefore it is not particularly surprising that there is not a great deal of detailed information available to date. Partnerships should be encouraged to secure funds and report funding details for transparent analysis.

3.5 Were Type II Energy Agreements a Useful Outcome of the WSSD?

In terms of volume, approximately 280 partnerships have been submitted to date, involving 71 governments, 181 NGO/IGOs, and at least 23 businesses. Of these partnerships, 43 addressed energy issues and 39 of those partnerships have been approved by the CSD. Of initiatives addressing management of natural resources, energy agreements form the largest single body Type II agreements. The publication of these partnerships will, at the very least, raise awareness of sustainable development issues, which is a valuable outcome regardless of the final success of the partnerships. Additionally, the collaboration between partners will have led to the exchange of information and expertise on sustainable development issues, which should enhance the prospects for sustainable development.

After it became clear that Type I agreements were not a feasible outcome of the WSSD, the summit still had to go ahead due to the initial financial investment and the political embarrassment that could have resulted from a cancellation. In this scenario, Type II agreements were a relatively valuable fallback position, as they allowed the summit to focus media attention on the implementation of sustainable development initiatives while reducing the negative press that may have resulted from no agreements.

In addition, the shift toward Type II agreements has demonstrated that responsibility for sustainable development is collective and not limited solely to the state, civil society, or business alone. It therefore follows that implementation of sustainable development initiatives must be a collective activity, and partnerships are a valuable way in which information and understanding can be transferred between the state, civil society, and business. The promotion of these partnerships and the subsequent press attention appears to have created a sense of collective responsibility, which is beneficial regardless of the final results of partnership initiatives.

However, for these agreements to be worth the trouble as tangible sustainable development projects, rigorous target-setting and tracking of partnership performance against those targets is necessary. Credibility is vital to attract additional funding for Type II agreements, and this funding will be needed for the partnerships to implement their projects and demonstrate that Type II agreements were worthwhile. That credibility can only be built and sustained by openness, target setting, and tracking of Type II partnerships. Building the credibility of Type II agreements will maximize their potential value for sustainable development.

SEE ALSO THE FOLLOWING ARTICLES

European Union Energy Policy • Global Energy Use: Status and Trends • International Comparisons of Energy End Use: Benefits and Risks • International Energy Law and Policy • Labels and Standards for Energy • Nongovernmental Organizations (NGOs) and Energy • Sustainable Development: Basic Concepts and Application to Energy • World Environment Summits: The Role of Energy

Further Reading

Calder, F. (2002). "The Potential for Using the Multistakeholder Network Model to Develop and Deliver Partnerships for Implementation ("Type Two Outcomes") for the World Summit on Sustainable Development." Royal Institute of International Affairs, London UK.

Commission on Sustainable Development. (2003). "Partnerships for Sustainable Development—Summary Analysis." CSD.

Kara, J., and Quarless, D. (2002). "Guiding Principles for Partnerships for Sustainable Development ("Type 2 Outcomes") to Be Elaborated by Interested Parties in the Context of the World Summit on Sustainable Development (WSSD). WSSD Prepcom 4.

Lélé, S. (1991). Sustainable development: A critical review. *World Development* **19**(6), 607–612.

Monbiot, G. (2002). Trouble in the pipeline: The corporate promises being made at the earth summit are likely to prove hollow. *The Guardian*, September 3, 2002.

Najam, A. (1993). International environmental negotiation: A strategy for the South. *In* "Papers on International Environmental Negotiation". (L. Susskind, W. Moomaw and A. Najam, eds.). Vol. III. Program on Negotiation at Harvard Law School, Cambridge, MA.

World Commission on Environment and Development. (1987). "Our Common Future." Oxford University Press, Oxford UK.

World Summit on Sustainable Development. (2003). "Partnerships for Sustainable Development Summary." United Nations, June 23.

Uranium and Thorium Resource Assessment

J. STEPHEN HERRING
Idaho National Laboratory
Idaho Fall, Idaho, United States

1. **Importance of Uranium and Thorium**
2. **Historical Context of Uranium Exploration**
3. **Geology and Geochemistry of Uranium**
4. **Current Uranium Production**
5. **Conventional Uranium Resources**
6. **Known Uranium Resources**
7. **Unconventional Resources**
8. **New Technologies for Uranium Extraction**
9. **Impact of Uranium Scarcity and Higher Extraction Costs**
10. **Summary**

Glossary

cross section Probability of neutron interaction with a nucleus, expressed in terms of area, in units of barns, with 1 barn equal to 1.0×10^{-24} cm^2.

enrichment The fraction of an isotope, usually fissile ^{235}U, in a mass of uranium; enrichment is commonly quoted as the weight percentage of the particular isotope.

enrichment tails (or depleted uranium) The uranium remaining after the enrichment of natural uranium into fuel; today it is approximately 0.3% ^{235}U, whereas earlier it was 0.20 to 0.25% ^{235}U.

fractionation Crystallization from a magma in which the initial crystals are prevented from equilibrating from the parent liquid, resulting in a series of residual liquids of more extreme composition than would have resulted from continuous reaction.

highly enriched uranium (HEU) Uranium containing more that 20% ^{235}U.

J_{th} Joule (i.e., watt-second) thermal, with 1 British thermal unit (BTU) equal to 1055 J_{th}.

low enriched uranium (LEU) Uranium containing less that 20% ^{235}U.

mafic Composed chiefly of dark ferromagnesian minerals.

mega-separative work unit (MSWU) A million separative work units; a separative work unit is the separative work done to change the separation potential of 1 kg of a mixture of isotopes by 1. (See Benedict, 1981, p.6 for a more complete definition.)

mixed oxide fuel (MOX) Usually consists of a ceramic mixture of uranium dioxide and plutonium dioxide.

quad A quadrillion (i.e., 10^{15}) British thermal units, with 1 quad equal to 1.055×10^{18} J_{th}.

pegmatite An exceptionally coarse-grained igneous rock, with interlocking crystals, often found at the margins of batholiths.

placer A mineral deposit at the surface formed by a sedimentary concentration of heavy mineral particles from weathered debris.

unconformity A break or gap in the geological record such as an interruption in the normal sequence of deposition of sedimentary rocks or a break between eroded metamorphic rocks and younger sedimentary strata.

yellowcake A concentrate of uranium ore containing 80 to 90% U_3O_8.

This article discusses uranium resources in a global sense beyond relying on the official estimates of the International Atomic Energy Agency and the Organization for Economic Cooperation and Development. As an introduction to that discussion, the origins of the earth's current inventory of uranium, the geophysical and geochemical processes that serve to concentrate that uranium into economically viable deposits, and the technologies now being used for extraction and concentration of uranium ores are described. Finally, the potential impacts of technologies now under development and the overall impact of the cost of uranium on the cost of energy from nuclear fission are reviewed.

1. IMPORTANCE OF URANIUM AND THORIUM

Uranium is a widespread element that is essential, at least in the near term, to the use of nuclear fission as

 © 2004 Elsevier Inc. All rights reserved.

a source of energy. Uranium is ubiquitous in the earth due to the wide variety of minerals in which it can occur and due to the variety of geophysical and geochemical processes that have transported it since the primordial formation of the earth from stellar debris. Uranium is approximately as common in the earth's crust as is tin or beryllium, and it is a minor constituent in most rocks and in seawater. The average crustal abundance of uranium is 2.76 weight parts per million (wppm), higher than the average concentrations of economically important elements such as molybdenum (1.5 wppm), iodine (0.5 wppm), mercury (0.08 wppm), silver (0.07 wppm), and gold (0.004 wppm). Uranium occurs in ores as uraninite (UO_2, pitchblende), carnotite, autunite, uranophane, davidite, tobernite, and other minerals containing U_3O_8 (actually a stable complex oxide $U_2O_5 \cdot UO_3$).

Since the discovery of nuclear fission, uranium has been seen as a scarce resource. Uranium-235 (^{235}U) is the only naturally occurring isotope that can be made to fission with thermal neutrons. Consequently, the resources of uranium have been believed to inherently limit the sustainability of nuclear energy. There have been two periods of extensive exploration for uranium, during the 1950s and during the 1970s, and both were followed by long periods of severe contraction in the market and in exploration activity. Today, exploration activity is still at a low level for several reasons. First, deposits found during earlier exploration periods have proven to be larger than was initially estimated. Second, nuclear energy has not grown as rapidly as was forecast earlier. Third, improved nuclear fuel management techniques and materials have allowed higher burn-up and longer operating cycles. Finally, the conversion ("down-blending") of highly enriched uranium (HEU) of military origin to civilian purposes has postponed the need for large amounts of newly mined natural uranium.

Concern that uranium would soon be exhausted was one of the driving forces in the development of fast breeder reactors, in which the nonfissile isotope ^{238}U is converted into the fissile isotope ^{239}Pu. These concerns also led to the development of reactors capable of converting thorium into the fissile, but not naturally occurring, isotope ^{233}U. Thorium, consisting nearly entirely of the isotope ^{232}Th, is approximately three times more abundant than uranium and, thus, may represent a source of fissile material in the distant future.

The importance of the overall uranium and thorium resource is demonstrated by the attention given the estimates of international agencies such as the International Atomic Energy Agency (IAEA) and the Nuclear Energy Agency (NEA) of the Organization for Economic Cooperation and Development (OECD). The estimates of these organizations are backed by a great deal of research and are compiled regularly in the Red Book.

However, the estimates of uranium resources in the Red Book are based on the known reserves that can be economically extracted using current technology. Because of the wide range of uranium concentrations in various minerals, the cost of extraction serves as the independent variable against which resources are estimated. Furthermore, the sizes of reserves are categorized by the reliability of the estimates.

Despite the success of earlier exploration efforts and the continuing low prices of natural uranium, there is a persistent and widespread belief that global resources of uranium will limit the long-term sustainability of nuclear power. That belief is supported by a misinterpretation of the Red Book estimates. The World Nuclear Association (formerly the Uranium Institute) estimates the reserve based on current and prospective mining projects and concludes that there are now 1.32 million metric tons (t) of uranium of the best proven category recoverable worldwide at a marginal cost of $40 per kilogram or less. The current consumption of natural uranium by reactors worldwide is approximately 64,000 t per year; consequently, these best proven reserves could be expected to last only approximately 20 years. If this interpretation of the reserve and resource estimates were accurate, there would be no long-term resource base for even the current fleet of light water reactors (LWRs).

However, the preceding is a serious misinterpretation of the real uranium resource situation. The most recent estimates of the current resource base, which includes speculative and higher cost resources, are 14 million to 17 million t. Low market prices, the slow growth of nuclear power, and the ongoing down-blending of HEU have resulted in both a very low level of exploration and little effort in the development of advanced extraction technologies.

It is important to note at the outset that uranium today is used overwhelmingly in the LWR fuel cycle, where only approximately 1.1 to 1.5% of the ultimate energy of the mined uranium is extracted via fissioning of ^{235}U and the small amounts of ^{239}Pu bred *in situ*. The rest of the uranium remains either in the spent fuel or in the depleted uranium after enrichment. Of the 1.8 million t of uranium produced worldwide since 1945, the location of all but approximately 1500 t is known. Only that uranium used either in nuclear

weapons or in armor-piercing projectiles is not known. The inventory of spent fuel and of depleted uranium represents a very significant resource.

Although uranium is an essential input for the production of nuclear energy, the costs of natural uranium are a minor component of the overall cost. Today, at uranium prices of approximately $25 per kilogram, the uranium itself is responsible for only roughly 2% of generating costs. The fuel element cost is 15 to 20% of the generating costs, but those costs include conversion of the uranium ore to UF_6, enrichment of the natural uranium, production of the ceramic UO_2 fuel pellets, and fabrication of the fuel assemblies. A 10-fold increase in the cost of natural uranium would not be welcome but also would not fundamentally change the economics of nuclear power. A 10-fold increase in uranium prices would, at first estimate, be expected to increase the cost of nuclear electricity by 20%. However, a more detailed calculation, optimizing the ^{235}U content of the depleted uranium tails and adjusting fuel management for a higher priced resource, would result in an increase in the cost of electricity of significantly less than 20%.

1.1 Energy Content of Uranium and Thorium

Uranium has 18 known isotopes, none of which is stable and only 2 of which have half-lives longer than a million years: ^{235}U (704 million years) and ^{238}U (4.47 billion years). Only ^{235}U, which is approximately 0.720 weight percentage (wt%) of natural uranium, is fissile, that is, will fission using thermal (i.e., low-velocity) neutrons. ^{238}U, which is by far the dominant isotope at 99.2745 wt%, will fission if struck by high-energy neutrons. However, ^{235}U is the only naturally occurring isotope capable of sustaining a neutron chain reaction in a suitably designed reactor.

^{238}U can be transformed into ^{239}Pu through the absorption of a neutron and two subsequent beta decays, as shown in the following reaction:

$$^{238}\mathrm{U}(n,\gamma)^{239}\mathrm{U} \xrightarrow[\beta^- t_{1/2}=23.5\ \mathrm{min}]{}$$
$$^{239}\mathrm{Np} \xrightarrow[\beta^- t_{1/2}=2.35\ \mathrm{days}]{}$$
$$^{239}\mathrm{Pu} \xrightarrow[\sigma_{f\ thermal}=750b]{} \text{fission},$$

where β^- indicates a beta decay with electron emission and $\sigma_{f\ thermal}$ is the fission cross section in barns at a neutron energy of 0.025 eV.

Thorium has 25 isotopes, of which only the nonfissile isotope ^{232}Th is long-lived, with a half-life of 14 billion years. However, in the following reaction, ^{232}Th can be transformed into ^{233}U, a fissile isotope:

$$^{232}\mathrm{Th}(n,\gamma)^{233}\mathrm{Th} \xrightarrow[\beta^- t_{1/2}=22.3\ \mathrm{min}]{}$$
$$^{233}\mathrm{Pa} \xrightarrow[\beta^- t_{1/2}=27.0\ \mathrm{days}]{}$$
$$^{233}\mathrm{U} \xrightarrow[\sigma_{f\ thermal}=531b]{} \text{fission}.$$

The fission energy of ^{233}U is 190 MeV and that of ^{239}Pu is 200 MeV. If 1 kg of thorium is bred into ^{233}U, the fission energy available is 78.9 TJ_{th} (thermal), and the fission energy in 1 kg of uranium, bred to ^{239}Pu, is 80.4 TJ_{th}. Thus, thorium and uranium are quite similar in maximum energy content, but uranium is far more important in the near term due to the need for ^{235}U as a naturally occurring fissile isotope.

1.2 Thorium Resources

Thorium averages 7.2 parts per million (ppm) in the earth's crust and is the 39th most abundant of the 78 crustal elements. Soil commonly contains an average of 6 wppm of thorium. It is about three times more abundant than uranium. When bred to the fissile ^{233}U, thorium releases about the same energy per unit mass (79 TJ_{th}/kg) as does uranium when bred to ^{239}Pu (80.4 TJ_{th}/kg). When thorium is used as a nuclear fuel, much less plutonium and other minor transuranics (e.g., Np, Am, Cm) are produced than are produced in uranium fuel cycles. The reduced production of transuranics occurs for two reasons. First, the fissile product of ^{232}Th is ^{233}U, further in the actinide series from plutonium and the minor actinides. Second, the fission/capture ratio of ^{233}U is approximately 9, whereas that of ^{239}Pu is approximately 3, thereby resulting in lower production of the transuranics.

Thorium and its compounds have been produced primarily as a by-product of the recovery of titanium, zirconium, tin, and rare earths from monazite, which contains 6.0 to 8.5 wt% thorium oxide. Only a small portion of the thorium produced is consumed. Limited demand for thorium, relative to the demand for rare earths, has continued to create a worldwide oversupply of thorium compounds and mining residues. Most major rare earth processors have switched feed materials to thorium-free intermediate compounds to avoid the handling of radioactive

thorium. Excess thorium not designated for commercial use is either disposed of as a radioactive waste or stored for potential use as a nuclear fuel or other applications. Increased costs to comply with environmental regulations and potential legal liabilities and costs to purchase storage and waste disposal space were the principal deterrents to its commercial use. Health concerns associated with thorium's natural radioactivity have not been a significant factor in switching to alternative nonradioactive materials. According to the U.S. Geological Survey, U.S. consumption of thorium, all for nonenergy uses, has decreased from 11.4 to 5.27 t (thorium content) since 1997. In fact, 76 t of thorium was shipped to Nevada for disposal as low-level radioactive waste in 2000. The principal applications of thorium today make use of the very high melting point of ThO_2 (3300°C, the highest of all binary oxides) and the electron-emitting capability of thorium when alloyed with tungsten for use in filaments for high-powered magnetrons for radar.

In the short term, thorium is available for the cost of extraction from rare earth processing wastes. In the longer term, large resources of thorium are available in known monazite deposits in India, Brazil, China, Malaysia, and Sri Lanka.

Quoted prices for thorium dioxide in 2001 were $82.50 per kilogram for 99.9% purity and $107.25 per kilogram for 99.99% purity. Mantle-grade thorium nitrate was available for $56 per kilogram of contained thorium.

Thorium is present in seawater at only approximately 0.050 wppb, due primarily to the insoluble nature of its only oxide, ThO_2. Thus, the recovery of thorium from seawater is not a realistic option.

Because ^{232}Th is the only isotope of natural thorium, there are no enrichment plant tails from thorium nuclear fuel. Therefore, the cost of thorium in a mixed thorium–uranium LWR fuel or in a pure thorium–^{233}U fuel cycle is relatively minor. However, the cost of chemical processing of ThO_2-based fuel and the separation of ^{233}U is high.

2. *HISTORICAL CONTEXT OF URANIUM EXPLORATION*

Uranium has been fundamentally important during the past half-century, first as a military resource and later as a commercial resource now responsible for 8.1% of the primary energy in the United States and 6.7% worldwide. The perceptions of uranium resources and a growing understanding of geological processes have led to widely varying levels of exploration for uranium during that time.

2.1 Periods of Scarcity and Surplus

There have been two boom periods in uranium production; the first was driven by military demands for HEU in the wake of World War II and the onset of the cold war, and the second was to fuel rapidly growing commercial nuclear programs in the wake of the oil embargo of 1972. Driven by those concerns, there have been two periods of extensive exploration for uranium, during the 1950s and during the 1970s, and both were followed by long periods of severe contraction in the market and the near cessation of exploration activity. Overexpansion of the uranium supply infrastructure during the 1970s led to limited exploration and the closure of operating mines during the past 20 years or so. A second reason has been the much slower growth of commercial nuclear power than was originally anticipated.

2.2 Demilitarization of Highly Enriched Uranium

More recently, a third disincentive to uranium exploration has emerged: disarmament. As much as one-third of the natural uranium mined since 1945 was used for the production of HEU, primarily for weapons programs and to a lesser degree for naval propulsion and research reactors. The Strategic Arms Reduction Treaties (START-I and -II) and heightened concerns for the proliferation of nuclear weapons have led to ongoing programs for the downblending of weapons-derivative HEU. In the downblending programs, HEU is mixed with natural or depleted uranium to produce low enriched uranium (LEU) fuel for commercial LWRs. The civilian use of downblended HEU effectively eliminates the need for as much as 600,000 t of natural uranium, equivalent to 10 years of current worldwide production. Thus, the market prices for natural uranium have been further depressed, and incentives for further exploration have been very small.

Specifically, the Russian and U.S. governments agreed to a program to convert 500 t of HEU from weapons into LEU for civil reactors. These 500 t of HEU at 90% ^{235}U would be blended with 14,759 t of LEU at 1.5% ^{235}U to obtain a product of 15,259 t of LEU at 4.4% ^{235}U. This blending process would

take place over a period of 20 years. By using depleted uranium to obtain the LEU at 1.5% ^{235}U and the HEU, this agreement will replace 152,000 t of natural uranium, which could have been enriched instead to yield the 15,259 t of LEU at 4.4% (with tails at 0.3%). The fuel can be expected to generate approximately 600 GWe-years of electricity.

2.3 Higher Burn-up and Plutonium Recycle in Light Water Reactors

Recycling of uranium and plutonium adds a potential savings of 15 to 20% in natural uranium requirements for a mixed oxide fuel (MOX)-fueled reactor park. Although the worldwide impact of U/Pu recycling on uranium demand remains small today, these savings potentially could contribute to a more sustainable use of the available resources. In this context, Japan has an aggressive plan for using MOX fuel in LWRs due to the lack of uranium resources in that country. In addition to the fabrication of MOX fuel in Europe, Japan Nuclear Fuels Ltd. (JNFL) has started preparation of domestic fabrication of MOX fuel. However, in today's political and economic context, no major expansion of this recycling is envisaged for the coming years in OECD member countries.

3. GEOLOGY AND GEOCHEMISTRY OF URANIUM

Because uranium has become an economically significant element, and because decisions on future energy choices will be influenced by the accuracy of estimates and perceptions of the adequacy of uranium and thorium resources, it is incumbent on us to first recognize the origins of those elements.

3.1 Uranium Abundance in the Earth

The bulk amounts of uranium inherited by the earth during its accretion from primordial stellar debris can be predicted through the use of various astrophysics models of Alpher and Urey. In these astrophysical models, the formation of the elements is the result of a continuous building up of heavier elements through neutron absorption in a highly compressed neutron gas. The nuclear matter decayed into protons and electrons to form first protium and then deuterium, followed by the heavier elements in a rapid combination of absorption and decay. The process proceeded rapidly until arrested when the neutron gas pressure fell due to universal expansion.

Urey cited estimates by Goldschmidt of the primordial abundance of 41 weight parts per billion (wppb) for uranium and 106 wppb for thorium. Alpher's theoretical curves and Brown's observed astrophysical data show uranium approximately 6.5 orders of magnitude less abundant than silicon, resulting in a primordial abundances of 57 wppb. Deffeyes, accounting for the decay of uranium since the expansion of the primordial neutron gas, estimated global uranium abundance at 10.5 wppb. The half-lives of the isotopes of interest are ^{232}Th, 14.0×10^9 years; ^{235}U, 0.704×10^9 years; and ^{238}U, 4.47×10^9 years.

Therefore, based on these astrophysical models, it is fairly clear that the earth currently contains approximately 10 wppb uranium and 40 wppb thorium. Stated in other terms, the current global inventory is 63 Tt (63×10^{12} t) of uranium and approximately 400 Tt of thorium. Although this inventory is a vast amount of both elements, if uranium and thorium had a uniform radial distribution throughout the earth (as is assumed in the cold accretion model), concentrations of uranium and thorium would be far too small to be economically extracted.

Thermal models of the earth point to inevitable melting of the earth soon after its accretion due to gravitation energy and due to radioactive decay of uranium, thorium, and potassium. Because of its large ionic size and heating due to radioactive decay, uranium is transferred into low melting temperature fractions and out of the earth's core and mantle into the crust. These geochemical and geophysical models predict that two-thirds of the initial 63 Tt of uranium present in the earth is now concentrated in the crust, which constitutes only 0.4% of the earth's total mass. The low uranium and high iron concentrations predicted for the earth's mantle and core have been supported by concentrations in iron meteorites and in mantle issuing from oceanic spreading zones (0.1 ppm U), compared with uranium concentrations in magma and crust in subduction zones (2 ppm U).

3.2 Mechanisms for the Concentration of Uranium

Unlike other energy resources such as coal and petroleum, the resources of uranium are not fundamentally changed by geological processes. Whereas petroleum might be lost through evaporation or

combustion or a natural gas reservoir may vent into the atmosphere, uranium is lost only through radioactive decay or through the relatively rare formation of a natural reactor. Therefore, the primordial inventory of uranium, reduced by radioactive decay, is essentially entirely present somewhere in the earth. The crucial question is "where?"

The natural distribution of elements in the earth's crust is controlled by two major factors. The first is the set of ambient geological fractionating processes that leads to regions of depletion and concentration of the element. The second factor includes the overall geochemical characteristics of the element. Elements that are concentrated by a small number of fractionation processes can be expected to have a multimodal distribution, with a peak in the tonnage versus grade curve for each of the modes of geochemical concentration. For elements having a large number of applicable concentration processes, the peaks overlap and the resulting tonnage versus grade curve takes on a log–normal characteristic. For example, for the element chromium, whose distribution at high concentrations is solely governed by fractional crystallization in mafic magmas (i.e., high in magnesium and iron), one would expect a bimodal distribution of concentrations, with one peak at the average crustal abundance and the high-concentration peak at the mafic fractionation concentration. On the other hand, most elements, including uranium, can undergo a wide variety of fractionating processes, and deposits would be expected over a wide range of concentrations. In this latter case, the tonnage versus grade distribution would be expected to be log–normal. Bear in mind that geological conditions change over time and that, therefore, the distribution patterns have varied with time.

In considering uranium in particular, it is important to examine the tectonic and igneous processes that have redistributed the uranium within the crust. During the past 4 billion years, the most important processes are continental accretion and plate tectonics. During the accretion process, crust formed into masses of continental dimensions. During the second (continuing) process, the continental crust has become much richer in uranium while the oceanic crust, although larger in area, has relatively little uranium.

3.3 Igneous Processes

Igneous processes begin with the fusion of mantle rocks at depths of 60 to 200 km, followed by the migration of less dense liquids to the surface. The migration of these less dense minerals to the surface is a predominant process in the formation of the continental crust. The extruded liquid forms crust in two general locations: at mid-oceanic ridges, where the upwelling material forms new oceanic crust, and in subduction zones, where the oceanic crust plunges back into the mantle, usually passing under the edge of a continent.

The behavior of uranium in igneous processes is dominated by two characteristics of the element. In the $+4$ oxidation state, the condition expected in the earth's mantle, the U^{+4} ion has an ionic radius (97×10^{-12} m [picometers (pm)]) the same as the Na^{+1} ion (97 pm). Other ions common in the core and mantle are significantly smaller in radius: Fe^{+2}, 74 pm; Ni^{+2}, 69 pm, Mg^{+2}, 66 pm; and Al^{+3}, 51 pm. Thus, like sodium and the other large ions, uranium ions selectively enter partial melts within the mantle and are transported to the surface.

The second characteristic of uranium is its radioactivity, serving as a source of heat for melting the mantle and core. Like Th^{+4} (ionic radius 102 pm) and K^{+1} (133 pm), these heat-producing elements are readily fractionated out of the mantle and toward the surface. Deffeyes noted that the earth would be a radically different place if the heat-producing elements had small radii given that the geothermal energy source would then be located deep within the core and the convection currents driving plate tectonics would be much stronger.

The rocks forming the oceanic crust at mid-oceanic ridges are characterized by a uniform uranium concentration of approximately 0.1 wppm. Conversely, the crust formed above subduction zones is characterized by uranium concentrations of approximately 2 wppm. The wide difference in concentration is due to the differences in the source materials and to the different chemistries. The upwelling mantle at the oceanic ridge has a uranium concentration of approximately 0.005 wppm, whereas the subduction zones have as their source material oceanic crust and bits of continental crust, with an average uranium concentration of approximately 0.1 wppm. The continuous upwelling at the oceanic ridges serves as a mechanism for depleting the core and mantle of uranium and incorporating that uranium into the oceanic crust. That relatively low concentration of uranium in the oceanic crust is augmented with uranium from continental runoff that subsequently precipitates in the ocean basins. At the subduction zones, the oceanic crust is again subjected to partial melting and the uranium is again fractionated in the melt and transported to the surface.

As a result of the various igneous processes, the average concentration of uranium is highest at the surface of the continental crust and decreases approximately exponentially with depth.

The anticipated variation of uranium concentration with depth is given by the equation $U(Z) = U(Z = 0)e^{(-Z/6300\ \mathrm{m})}$, where z is the depth in meters and $U(z)$ is the concentration at depth z in weight parts per million. $U(z = 0)$ is the average continental crustal abundance of uranium at the surface (2.76 wppm). Thus, 15% of the crustal uranium inventory would be expected to be within 1000 m of the surface, and 27% would be expected to be within 2000 m.

3.4 Geochemical Beneficiation Processes

The governing characteristic in the geochemical transport of uranium is the fact that the uranium is highly soluble in oxidizing environments and essentially insoluble in reducing environments. Thus, the change in the earth's atmosphere from a reducing condition to an oxidizing condition some 1.8 billion years ago is responsible for a fundamental change in the dominant processes in uranium transport. During the earlier age, igneous processes and fractionation of uranium in partial melts due to its large ionic size were dominant. During the past 1.8 billion years, the transport of uranium by means of groundwater oxygenated at the surface has been dominant.

Thus, during the period more than 1.8 billion years ago, uranium was primarily concentrated in placer deposits as a chemically inert and physically dense phase. Because of the low solubility of uranium in reducing environments, rivers, lakes, groundwater, and the sea contained very low uranium concentrations. The placer deposits at Elliott Lake, Canada, and at Witwatersrand, South Africa, are typical of the deposits formed during this period.

With the dominance of photosynthesis during the past 1.8 billion years, the atmospheric and groundwater conditions have been oxidizing and uranium minerals have been highly soluble in the sedimentary weathering cycle. Placer deposits no longer formed and, in fact, began to dissolve. The uranium content of rivers, lakes, and groundwater increased, and gradually the uranium concentration in the oceans also increased. Nevertheless, the uranium concentration remained well below saturation.

However, in a few isolated locations, oxidation of organic-rich beds by groundwater led to locally reducing conditions. In these locations, the uranium ions or their complexes would reach supersaturation and reprecipitate. An important example of this reprecipitation is in the Mesozoic sandstones of the Colorado plateau. The uranium ores are found in organic-rich zones where the oxygen in groundwater was removed by carbon-rich debris. Precipitation of uranium has also occurred where restricted circulation in the oceans and organic-rich sediments led to anoxic conditions. Good examples are the black shales of Chattanooga shale and the phosphorite shale of the Phosphoria formation.

3.5 Specific Deposits

Before the discovery of the McArthur River deposit, the highest grade ores were obtained from igneous sedimentary deposits such as Great Bear Lake in Canada, Joachimsthal in the Czech Republic, and Katanga in the Congo. The deposits at Oklo in the Gabon Republic in Africa were of high enough concentration, and the uranium 1.7 billion years ago contained 3% ^{235}U, rather than the current 0.72%, such that several critical natural reactors operated in the deposit. The reactor zones released approximately 15,000 MW-years of fission energy over the course of some 250,000 years. These deposits were formed by the movement of hot water through fractures in blocks of rocks heated by their own uranium and thorium content.

Precambrian sandstones overlie much older Precambrian granites and metamorphic rocks. At the interface, there is a discontinuity in the age of the rocks. This type of discontinuity is termed an unconformity. Unconformity deposits, such as those in Saskatchewan and northern Australia, occur where uranium from the sandstone has formed into veins in the open spaces of the interface and has been heated to temperatures of several hundred degrees Celsius.

3.6 Roll-Front Deposits

As mentioned earlier, uranium oxide precipitates when the solution enters a reducing environment. The uranium oxide can be redissolved *in situ* by oxygenated leach solutions. In sandstone deposits, the uranium minerals have been deposited in the interstices between the sand grains. The deposits are often moving (very slowly) through the sandstone due to the flow of groundwater, much like the movement of a front through a liquid chromatography column. Oxygenated water from the surface

TABLE I

Natural Uranium Distribution

Geological unit	Mass (whole earth, 10^{19}t)	Concentrated uranium (ppm)	Mass-Uranium (whole earth, billion t)	Mass-Uranium (continental United States, billion t)
Whole earth	597.6	0.0105	63,000	
Core	193.6	0.00001	20	
Mantle	401.6	0.0005	20,000	
Crust	2.5	1.67	43,000	
Continental crust	2.09	1.93	40,000	
Mafic	1.01	1.45	14,625	980.0
Stalic	0.99	2.56	25,375	1700.16
Metamorphic	0.43	2	8000	536
Igneous	0.43	1.93		
Plutonic	0.1754			
Veins/Pegmatites		5000–10000	0.0018	0.00012
Veins/Pegmatites		>10000	0.0018	0.000012
Uranium-granites		5–15	597	40
Granodiorites	0.1750	3	5000	335
Dioritie/Gabbro	0.0004	1.63	80	5.36
Volcanic				
Uranium-rhyolite		100–300	0.104	0.007
Andesite	0.0437	2	800	53.6
Basalt	0.0510	1	4100	274.7
Sedimentary	0.128		3200	
Fold Belts	0.094		2200	
Sandstones	0.018	1.01	180	12.06
Shales	0.037	4	1500	100.59
Limestones	0.015	2.2	330	22.11
Volcanogenic	0.024	1	240	16.08
Evaporites	0.0003	0.2	6	0.4
Platforms	0.034		940	

enters the sandstone, where reducing agents, such as sulfides and organic matter, are located in the interstitial spaces. The organic carbon in one pore volume of sandstone can remove all of the oxygen dissolved in 50,000 pore volumes of oxygenated groundwater.

Therefore, the front between the oxygenated groundwater and the oxygen-free groundwater moves slowly through the sandstone. Uranium dissolved at the surface and uranium dissolved from the sandstone by the oxygenated groundwater is swept along and precipitated at the front. Upstream of the front, the uranium is present in the groundwater as the soluble hexavalent uranyl carbonate complex. As the oxygen is removed from the groundwater at the front, the soluble hexavalent uranium is reduced to the insoluble quadrivalent state.

The quadrivalent uranium precipitates in the form of the mineral uraninite (UO_2). Thus, the location of ore bodies is often associated with deposits of carbonaceous materials where the carbon, in much larger quantities compared with the uranium, has removed the oxygen from the groundwater.

Based on the various modes for the formation of uranium ore bodies, reviewing the large body of prior research, Deffeyes and MacGregor estimated the uranium content of the various crustal regimes in a report for the U.S. Department of Energy (DOE) in 1978. That information is shown as Table I.

Because of the wide variety of uranium-containing minerals and given the known quantities of those minerals, Deffeyes and MacGregor argued that worldwide uranium deposits have a log–normal distribution of mass versus grade. Current mining

Table I continued

Geological unit	Mass (whole earth, 10^{19}t)	Concentrated uranium (ppm)	Mass-Uranium (whole earth, billion t)	Mass-Uranium (continental United States, billion t)
Uranium-sandstones		600–1000		0.004
		1000–3000	0.75	0.008
Sandstones	0.008	1.01	81	5.43
Placers		100–2000	0.001	0.00007
Shales	0.017	4	680	45.56
Uranium-shales		10–40	29.1	1.95
Uranium-shales		40–200	0.09	0.006
Phosporites		120–300	1.94	0.13
Limestones	0.007	2.2	150	10.05
Volcanogenic	0.001	1	10	0.67
Evaporites	0.0007	0.2	14	0.94
Oceanic crust	0.48		2600	
Igneous/Metamorphic	0.38	0.1	380	
Sedimentary	0.1005	2	2200	
Deep ocean	0.02	1.0	200	
Calcareous oozes	0.008	0.77	60	
Siliceous oozes	0.003	0.45	14	
Ferrugenous clay	0.001	1.3	13	
Red clay	0.006	3.00	180	
Volcanogenic	0.0006	1.00	6	
Shallow ocean		1.92	1700	
Shelf	0.0125	3.02	380	
Hemipelagic	0.068	2.00	1400	
Hydrosphere				
Rivers/Lakes	0.000051	0.001	0.00005	0.0000335
Groundwater	0.00051	0.001	0.0005	0.000335
Oceans	0.141	0.003	4.2	
Ice	0.0023	<0.0001	0.0022	

Source. Deffeyes and Mac Gregor (1978).

activities are recovering uranium at market prices of $25 per kilogram from ores containing 0.05 to 20% U_3O_8. Given the log–normal distribution and the known quantities of the various uranium mineralizations, they estimated that a 10-fold increase in the price of uranium (and thus a 10-fold decrease in economically viable ore grade) would result in a 300-fold increase in the amount of uranium available. Equivalently, the World Nuclear Association estimated that a doubling of uranium prices would result in a 10-fold increase in supply.

The distribution of mass versus grade for the various types of uranium deposits is shown in Fig. 1. The three bars on the left (grade >1000 ppm U) indicate deposits of the type now mined for uranium alone. The expected log–normal distribution is shown, and the mass and grade of two Canadian mines discovered since 1978 are also indicated: (1) the McArthur River deposit, 137,000 t of proven reserves averaging 18 wt%, and (2) the Cigar Lake deposit, 90,000 t at an average grade of 17 wt%. Other known larger but lower grade deposits have not been shown.

The slope of the log–normal curve at currently mined grades is also shown in Fig. 1. This slope, roughly $+3.5$, indicates that for a doubling of the cost of mining (i.e., mining ore at half the current concentration), the economically available resources of uranium would increase by more than an order of magnitude ($2^{3.5} \approx 11$). This estimate presumes a continuation of current mining techniques and does not consider the increased use of *in situ* leaching (ISL) and recovery of uranium as a by-product in gold, copper, and phosphate mining.

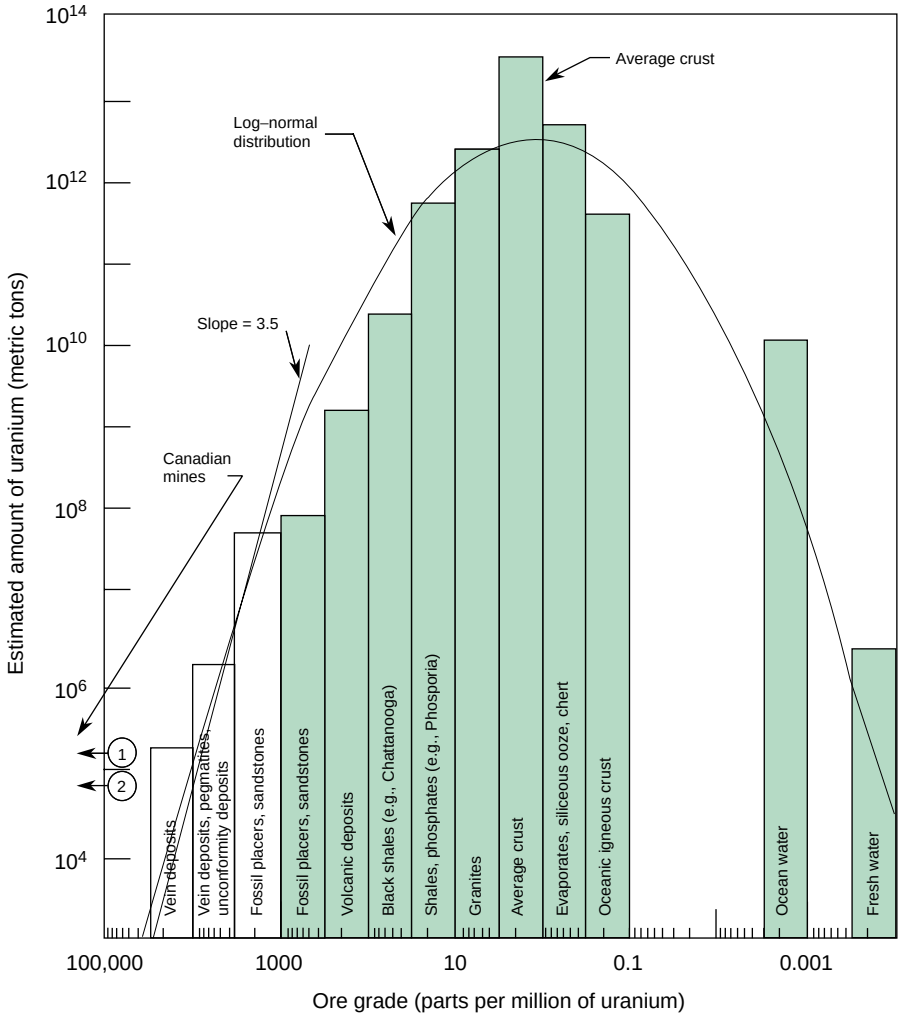


FIGURE 1 Distribution of uranium in the earth's crust. Adapted from Deffeyes and MacGregor (1978, 1980).

4. CURRENT URANIUM PRODUCTION

Canada produces the largest share of current world uranium production from mines (35%), followed by Australia (22%). Uranium production from mines in other countries is shown in Table II. Note that uranium production from mines has been depressed during the past 5 years due to the downblending of military HEU from the former Soviet Union. That demilitarization has effectively displaced the need for 10 years of natural uranium production (~600,000 t U).

4.1 Current Uranium Production Methods

As shown in Table III, the distribution of uranium production has been trending away from open pit mines and toward ISL and by-product production for the past 20 years or so. Those two methods are expected to grow in the future due to their relatively low cost and reduced tailings production.

5. CONVENTIONAL URANIUM RESOURCES

Today's uranium production comes from approximately 63 mining centers with concentrates produced at roughly 44 processing plants. Employment associated with this uranium production accounts for some 53,700 person-years annually, a reduction from some 64,000 person-years in 1994. Today, after more than 10 years of decreasing exploration expenditures, an increase of exploration expenditures has been noted, to $130 million annually worldwide.

5.1 Uranium Price History

The projected production capabilities (of existing and committed production centers) are expected to fall from approximately 45,800 t of uranium in 1997 to 33,000 t in 2015 due to the resorption of ex-military stocks and the closure of existing mines as a result of resource depletion or the lack of economics. Existing and committed capability at $40 per kilogram uranium or less is likely to meet only 40 to 60% of the requirements in 2015. However, the need for uranium will continue so long as nuclear electric generation continues and new mine production will be needed in large quantities sometime during the period to supply requirements once the excess supplies from past military efforts are exhausted. In the longer term, resources recoverable at higher costs and additional supplies would be necessary to fill the potential production shortfall indicated by some of the projections. Overall world reactor-related uranium requirements are expected to evolve from approximately 60,488 t of uranium in 1996 to levels ranging between 54,500 and 79,800 t per year by the year 2015.

As shown in Fig. 2, natural uranium prices, in the form of U_3O_8, have trended downward for the past 15 years after peaking in the wake of the oil shocks of the 1970s.

TABLE II

Worldwide Uranium Production (t uranium)

Country	2001	2000
Canada	12,520	10,682
Australia	7720	7578
Niger	3096	2895
Namibia	2239	2714
Uzbekistan	2400	2350
Russia (estimated)	2000	2000
Kazakhstan	2018	1752
United States	1000	1456
South Africa	898	878
China (estimated)	500	500
Ukraine (estimated)	500	500
Czech Republic	330	500
India (estimated)	200	200
France	124	319
Romania	115	50
Spain	30	251
Others	77	121
Total world	35,767 (42,180 t U_3O_8)	34,746 (40,976 t U_3O_8)

6. KNOWN URANIUM RESOURCES

For the reasons already noted, uranium is widely distributed in the earth's crust. Today's conventional uranium reserves are estimated at 4 million t (known conventional resources) at costs of less than $130 per

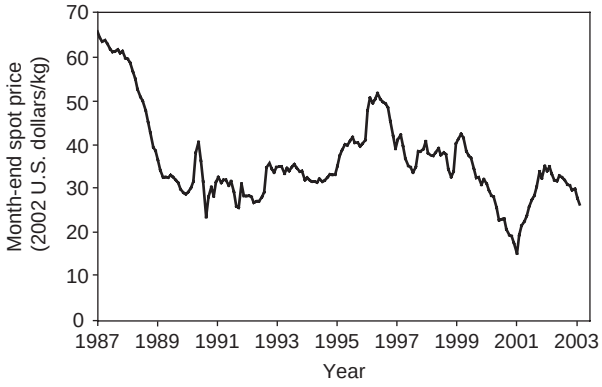


FIGURE 2 Historical uranium prices. (Compiled with data from the Ux Consulting Company, LLC and the U.S. Bureau of Labor Statistics.)

TABLE III

Distribution of Uranium Production Technology Types (Percentages)

Material source	1981	1994	1995	1996	1997	1998	2001
Open pit	68	38	37	39	49	47	29
Underground	27	41	43	40	32	34	40
ISL	5	14	14	13	13	14	16
By-product[a]		7	6	8	6	5	15

[a] Phosphate by-product, heap, and in-place leaching and mine water recovery. Olympic Dam is in the by-product category rather than the underground category.

TABLE IV
Overview of Uranium Resources

Resources reported in 1999 (thousand t)			Number of years of current nuclear electricity production
Uranium stocks		200	4
HEU and Pu		600	12
Known conventional resources	< $40/kg	>1254	
	< $80/kg	3002	
	< $130/kg	3954	80
Undiscovered conventional resources	< $80/kg	1460	
	< $30/kg	5338	
	Total	11,459	230
Uranium in phosphates		22,000	440
Uranium in seawater		4,200,000	80,000

kilogram. Reasonably assured resources, recoverable at costs of less than $130 per kilogram, account for an additional 2.3 million t. Besides conventional resources, large resources of uranium exist. Phosphates are known to contain significant amounts of uranium, and industrial production has been performed in the United States and Belgium for the past 20 years or so. Another practically inexhaustible potential source of uranium is seawater (Table IV). The conventional resources alone could sustain the nuclear option for at least the next 100 years, based on today's nuclear reactors. In principle, they would last more than 60 times longer if uranium and plutonium present in the spent fuels were systematically recycled in fast reactors, subject to technical developments and economic considerations.

6.1 The Perspective for the Next 20 Years

However, despite these formidable resources, only highly localized deposits with a concentration of approximately 500 ppm or more can be mined economically in today's market conditions, that is, at prices of less than $40 per kilogram. Uranium is actually produced using both conventional mining (open pit and underground mining) and unconventional production techniques. Unconventional techniques include ISL technology, phosphate by-product recovery, and heap leaching. ISL technology has a very low environmental impact but can be used to extract uranium only from suitable sandstone-type deposits. The only three plants using phosphate by-product recovery (two in the United States and one in Belgium, with total production of 495 t U/year) suspended their uranium production in early 1999. The distribution of production for the four technology types or material sources is shown in Table III.

In the next 20 years or so, low-cost uranium production from new projects is expected to be primarily from high-grade unconformity-type deposits, multimineral deposits, and sandstone-type deposits amenable to ISL mining technology. It is expected that the use of conventional production technology will increase, with particular emphasis on underground mining, while ISL technology could maintain its relative share.

Today, primary production amounts to only 60% of the total world demand. Four additional (secondary) uranium resources make up the balance of supply and demand: uranium stockpiles, reenrichment of depleted uranium, recycling of uranium and plutonium, and the use of former weapons-grade material (HEU).

6.2 Locations of Known Uranium Resources

Australia, Kazakhstan, Canada, and South Africa dominate the known uranium resources of the world, with roughly 28, 15, 14, and 10% of the known resources recoverable, respectively, at a cost of less than $80 per kilogram (Table V).

The estimates of resources must be used with some caution. Often, the reasonably assured resources significantly underestimate resources that will be found with further exploration. For example, although the U.S. resources are listed as approximately 104,000 t uranium, 354,000 t were produced from domestic mines from 1949 through 2001. U.S. uranium mines have been closed, not for lack of resources but rather because current prices ($29.30/kg in June 2003 according to the Ux Consulting Company) no longer support profitable operation. The NEA and IAEA assess the current uranium resource base, including more speculative and higher cost resources, at 16.2 million t.

6.3 Canadian Reserves

The Canadian reserves are a good example of the potentially much larger size of the resource base. In

TABLE V

Known Recoverable Resources of Uranium

	Metric tons	Percentage of world
Australia	863,000	28
Kazakhstan	472,000	15
Canada	437,000	14
South Africa	298,000	10
Namibia	235,000	8
Brazil	197,000	6
Russian Federation	131,000	4
United States	104,000	3
Uzbekistan	103,000	3
World total	3,107,000	

Source. Organization for Economic Cooperation and Development, National Energy Agency, and International Atomic Energy Agency (2001).

the Red Book estimates, the Canadian resources are listed as 331,000 t uranium. These vast economic concentrations are unconformity-type deposits in Proterozoic sedimentary basins. However, the McArthur River deposit, 620 km north of Saskatoon, Saskatchewan, itself contains 161,000 t uranium in ore that average 23% U_3O_8.

6.3.1 McArthur River/Key Lake

McArthur River, which is the world's highest grade and largest uranium mine, was discovered in 1988 and, after regulatory approvals were obtained, began mining operations in 1999. Between 1988 and 1991, 60 exploratory holes were drilled from the surface, of which 37 intersected uranium mineralizations. Approximately 70% of the deposit is based on seven of those holes. One hole graded 43% U_3O_8 for a distance of 25 m. The proven reserves at McArthur River are 137,000 t of uranium from ore averaging 17.8 wt%, probable reserves of 15,000 t averaging 19.5 wt%, and indicated resources of 87,000 t averaging 9.1 wt%. Like the other large Saskatchewan deposits, the McArthur River deposit occurs close to the unconformity between the flat-lying unmetamorphosed Proterozoic sandstones of the Athabasca Group and the lower Proterozoic and Archean rocks beneath. At the McArthur River site, this unconformity is at a depth of 500 to 600 m.

Remote-controlled equipment and nonentry mining methods are used at McArthur River to minimize the direct and inhalation radiation exposure to workers extracting the high-grade uranium ore. Because the groundwater, at a normal hydrostatic pressure of 5.1 MPa, has radon concentrations as high as 8.9×10^9 Bq/m^3 (0.24 Ci/m^3), special provisions must be made for adequate ventilation of the radon and for freezing of the adjacent groundwater to prevent flow into the mine.

The ore is milled 80 km south of the McArthur River deposit at Key Lake. Uranium production from the combined McArthur River–Key Lake complex in 2001 was 7225 t. Although Key Lake reserves, estimated to total 70,400 t, will be depleted over the next 5 years or so, production from McArthur River is expected to continue for several decades, according to Jamieson and COGEMA Resources.

6.3.2 Cigar Lake

Cigar Lake, 660 km north of Saskatoon, is the world's second-largest uranium deposit. Discovered in 1981 and approved as a test mine in 1987, the mine received government approval in 1998 and could begin uranium production in 2005 or later depending on market conditions. During the first 15 years of operation, the mine is expected to produce approximately 232 million pounds of U_3O (90,000 t U) at an average grade of 21%. The mine is expected to have a life span of 30 to 40 years. The overall Cigar Lake reserve is estimated at 137,000 t of uranium.

The Jasper Lake deposit contains approximately 123,000 t of uranium in ore that is 13.6% U_3O_8. The two deposits cover less than 120 square miles and were first explored some 20 years ago. Cigar Lake (137,000 t U) and Key Lake (70,400 t U) are also within 20 miles of each other.

The richness of the Canadian deposits and the relatively small area that has been explored both suggest that additional demand and exploration will result in substantial further discoveries.

6.4 Australian Reserves

Australia has emerged as a major producer of uranium since 1960. An intensive exploration effort at that time led to the discovery of the Ranger, Jabiluka, and Olympic Dam deposits as well as several others. Today, Australia has 27% of the world's proven reserves and a similar share of worldwide production. Australian deposits are roll-front deposits in sandstone or caliche, unconformity deposits, and a few veins in pegmatites. In general, the Australia deposits are lower in grade than the currently operational Canadian deposits.

The Ranger deposit, located 230 km east of Darwin, originally contained approximately 93,000 t of uranium at an average grade of 0.16 to 0.25 wt%. The mine has been producing approximately 4300 t annually for the past two decades. Roughly half of the original deposit has been exhausted.

The Jabiluka deposit, 20 km north of the Ranger deposit, contains at least 60,000 t of uranium at an average grade of 0.43 wt%.

The Olympic Dam deposit, located 265 km north of Port Augusta in South Australia, is potentially one of the world's largest uranium deposits. Estimated reserves are 266,000 t with an ore grade of 0.04 wt% and an additional indicated resource of 388,000 t with an average grade of 0.03 wt%. In addition, the deposit contains approximately 12 million t of copper and significant amounts of gold and silver. Currently, the Olympic Dam deposit is not being operated as a primary uranium producer but instead recovers uranium as a by-product of copper and gold extraction. Production capacity is being increased to 4700 t per year from the current 4000 t.

6.5 Other Ore Deposits

The United States has extensive ore deposits in the West, where many production centers were built between 1947 and the 1970s. From 1960 to 1980, the United States produced an average of 10,850 t of uranium annually. From 1949 through 2000, U.S. production was 354,421 t annually.

However, because of slow demand growth and large discoveries in Canada and Australia, many of these facilities have been closed and dismantled or put on long-term standby. Current production of 1000 t of uranium per year from an aggregate capacity of 3060 t per year uses ISL at six locations. In addition, one by-product plant is operational with a capacity of 290 t per year. Four ISL, three by-product, and six conventional mills are on standby.

Reasonably assured resources, at up to $80 per kilogram of uranium, are estimated to be 104,000 t, whereas resources in the range of $80 to $130 per kilogram are estimated to be 244,000 t. Estimated additional resources are 839,000 t at up to $80 per kilogram and 434,000 t at $80 to $130 per kilogram. Speculative resources in the same cost brackets are 504,000 and 354,000 t, respectively. Additional speculative resources, with no cost range assigned, are estimated to be 482,000 t.

The only current source of Russian uranium production is the Streltsovsky region near Krasnokamensk in eastern Siberia. Output in 2001 was approximately 2000 t of uranium, of which an estimated 90% was derived from underground mining, with the balance being derived from low-grade ore by heap or in-place leaching. The Russian Federation has reasonably assured resources, at up to $80 per kilogram of uranium, of 140,900 t. The balance of the known conventional resources consists of 36,500 t recoverable at less than $80 per kilogram. Undiscovered resources are estimated to be exceedingly large—nearly 105,000 t of uranium at up to $130 per kilogram plus 1 million t of speculative resources, of which more than half are thought to be recoverable at less that $80 per kilogram.

7. UNCONVENTIONAL RESOURCES

Unconventional resources are those outside the mining industry or those in deposits not generally mined for the uranium or thorium alone.

7.1 Enrichment Tails

An unconventional source of uranium is the reenrichment of depleted uranium by using today's overcapacity of enrichment. Because of the low prices of natural uranium during recent years, many enrichment plants have been operating with tails assays of up to 0.3% ^{235}U. Therefore, the 1.18 million t of depleted uranium currently stored at enrichment plants could supplant a few hundred thousand metric tons of natural uranium if demand required. The inventory of depleted uranium is expected to increase by approximately 51,400 t per year through at least 2010. The enrichment capacity in 1999 was reported to be 55.6 million separative work units (MWSU) per year, compared with an annual demand of 35.4 MWSU per year. The current spare capacity in enrichment plants in the world, at roughly 15 to 20 MSWU per year, theoretically represents an equivalent of 7000 t of natural uranium per year if this spare capacity was used for enrichment of depleted uranium with an assay of 0.3% and a new tails assay of 0.1%.

However, the economics of reenrichment depend on the ^{235}U assay of the depleted uranium and the relationship between the price of uranium and the cost of enrichment services. A tails assay of more than 0.3% is preferable if reenrichment of depleted uranium is to be considered a possibility. Reenrichment of depleted uranium for the production of LEU and natural uranium in the Russian Federation has taken place for several years.

7.2 Gold and Phosphate Tailings

In addition to the discovery of new resources through increased exploration, improvements in mining technology are lowering the cost of previously high-cost deposits. In particular, ISL is of growing significance and could be applied to existing gold and phosphate tailings piles. The resource base of 16.2 million t does not include uranium in gold and phosphate tailings. The phosphate deposits are estimated at 22 million t.

7.3 Recovery of Uranium from Seawater

The recovery of uranium from seawater places an upper limit on the cost of uranium. Uranium is dissolved in seawater at 3 mg/m^3 (3 wppb) and represents a well-known resource of 4 billion t, more than 300 times the known land-based resource. Estimates of recovery costs have been in the neighborhood of $200 per kilogram, although these estimates are highly speculative. The uranium content of the oceans is relatively constant, and large-scale extraction can be done without local depletion of the resource. Because only approximately 3% of the global population live in landlocked countries, extraction of uranium from seawater is truly the bounding cost for uranium.

Uranium recovery from seawater has been studied in Japan for the very long term or to face a very strong development of fission energy. In a laboratory-scale experiment performed by the Japan Atomic Energy Research Institute (JAERI), uranium is trapped by an amidoxime adsorbent that has been prepared on nonwoven polyethylene material with the aid of radiation-induced cografting. This experiment, conducted 7 km offshore from Sekine-Hama in Aomori Prefecture, Japan, produced more than 1 kg of uranium on 350 kg of nonwoven fabric during a total submersion time of 240 days, as reported by Seko. Yet, at this stage of the study, it is difficult to predict the practical application of uranium recovery from seawater. An economic assessment has been reported indicating a possible cost for this uranium process in a 1000-t/year commercial plant of approximately $600 per kilogram. Ultimately, a goal of approximately $100 per kilogram was reported to be feasible in the longer term with an improved absorbent.

The recovery of uranium from seawater is highly speculative and may never prove to be economical. For example, 1 t of seawater contains 3 mg of natural uranium, which can deliver 244 MJ_{th} in a breeder or approximately 2.5 MJ_{th} in a present-day LWR. Simple calculations show that the pumping energy needed in an extraction plant could easily consume all of the energy available, particularly in the LWR case. Thus, seawater extraction conceptual designs relying on ion exchange or adsorption have used ocean currents or wave action to move the seawater past the uranium-collecting surfaces.

As discussed later, harvesting the remaining fissile material from enrichment tails or harvesting fissile isotopes from discharged fuel may prove to be more economical routes for uranium resource extension than is seawater extraction.

However, the magnitude of the seawater resource places an upper limit on the cost of uranium for several reasons. First, seawater is available to nearly all countries of the world at virtually the same uranium concentration and without local depletion due to the extraction of uranium. Second, because no group of countries can form a cartel over the uranium supply if seawater extraction is practiced, the price of uranium is unlikely to be driven artificially high through market manipulation. Furthermore, the only current limitation on the extraction of uranium from seawater is knowledge of the technology and resins. Thus, one would expect that if conventional sources of uranium become limiting, a healthy competition in research and development would drive the price toward the more optimistic $200 per kilogram.

Therefore, a successful technology for the extraction of uranium from seawater places a cap on the cost of uranium. However, for the foreseeable future, that cap will be well above the market prices of uranium from geological deposits.

The additional cost of uranium recovered from seawater, even under somewhat pessimistic costs of $1000 per kilogram, would not add more than 30% to the cost of electricity. Thus, although the availability of uranium from conventional mines is of some interest in the development of new reactor types, limited uranium resources should not be a determining factor in the choice of a reactor type. The increase in the overall cost of electricity would be less than that seen routinely in fossil fuels (e.g., the price fluctuation of natural gas since 1998).

8. NEW TECHNOLOGIES FOR URANIUM EXTRACTION

The current low prices of uranium provide little motivation for the development of new extraction technologies. However, regulations to minimize the

TABLE VI

Impact of Uranium Costs on the Cost of Electricity

	Current uranium prices		Uranium from seawater				
			Optimistic:$200/kg		Pessimistic: $1000/kg		
	UO_2 fuel 45 MW-day/kg	UO_2 fuel 60 MW-day/kg	UO_2 fuel 45 MW-day/kg	UO_2 fuel 60 MW-day/kg	UO_2 fuel 45 MW-day/kg	UO_2 fuel 60 MW-day/kg	Units
Specific power	37.9	37.9	37.9	37.9	37.9	37.9	KWth/kg
Cycle parameters							
Total cycle length	3.8	5.1	3.8	5.1	3.8	5.1	years
Capacity factor	85%	85%	85%	85%	85%	85%	
Effective full power days	1186	1582	1186	1582	1186	1582	efpd
Burnup	45.0	60.0	45.0	60.0	45.0	60.0	MW-day/kg ihm
Feed ^{235}U content	0.72%	0.72%	0.72%	0.72%	0.72%	0.72%	atom%
Product ^{235}U enrichment	4.05%	5.40%	4.05%	5.40%	4.05%	5.40%	atom%
Tails ^{235}U content	0.30%	0.30%	0.15%	0.15%	0.05%	0.05%	atom%
Feed ^{235}U content	0.71%	0.71%	0.71%	0.71%	0.71%	0.71%	wt%
Product ^{235}U enrichment	4.00%	5.34%	4.00%	5.34%	4.00%	5.34%	wt%
Tails ^{235}U content	0.30%	0.30%	0.15%	0.15%	0.05%	0.05%	wt%
Separative work	5.277	7.852	7.554	11.053	11.663	16.829	kg-SWU/kg fuel
Natural uranium	9.009	12.257	6.845	9.216	5.973	7.990	kg/kg fuel
	0.200	0.204	0.152	0.154	0.133	0.133	kg MW-day

impact of mining on the environment and radiation exposure to workers have led to the use of technologies where uranium is extracted *in situ* or where ores previously mined for another element are processed for uranium extraction. In addition, the technologies described previously for the extraction of uranium from seawater could play a critical role in minimizing environmental impact and radiation exposure.

8.1 *In situ* Leaching

In conventional mining, the rock of the ore body is removed from the ground, transported to a mill, and treated to remove the minerals of economic value. The opening of the mine, the transport of the ore, the milling, and the disposal of remaining treated rock can create severe environmental impacts. ISL, sometimes known as solution mining, involves the use of liquids to dissolve the desired elements from the ore body without removing it from the ground. The liquid is pumped through the ore body and returned to the surface, where the desired elements are removed from the solution by precipitation, electrochemistry, or other means. The leaching liquid is then returned to the ore body and the process is repeated. ISL eliminates the need to remove large quantities of ore from the ground and to transport it to the mill, thereby minimizing surface disturbance. ISL also eliminates the need to dispose of the tailings or waste rock. However, for ISL to be effective, the ore body must be permeable to the flow of the leaching liquid. Furthermore, the ISL site must be located so as to not contaminate groundwater away from the ore body.

Today, because of its reduced surface impact and lower cost, ISL is used for 85% of U.S. uranium production. Most of the operations in Wyoming,

Table VI continued

	Current uranium prices		Uranium from seawater				
			Optimistic:$200/kg		Pessimistic: $1000/kg		
	UO_2 fuel 45 MW-day/kg	UO_2 fuel 60 MW-day/kg	UO_2 fuel 45 MW-day/kg	UO_2 fuel 60 MW-day/kg	UO_2 fuel 45 MW-day/kg	UO_2 fuel 60 MW-day/kg	Units
Rates							
Interest rate	8.0%	8.0%	8.0%	8.0%	8.0%	8.0%	per year
Natural uranium	$30.17	$30.17	$200.00	$200.00	$1000.00	$1000.00	/kg U
	$11.63	$11.63	$77.09	$77.09	$385.45	$385.45	/lb ^{308}U
Conversion ^{308}U to UF6	$5.00	$5.00	$5.00	$5.00	$5.00	$5.00	/kg natural U
Separative work	$78.00	$78.00	$78.00	$78.00	$78.00	$78.00	/kg-SWU
Costs							
Natural uranium	$271.79	$369.80	$1369.00	$1843.19	$5972.66	$7989.80	/kg fuel
Separative work	$411.58	$612.44	$589.24	$862.17	$909.72	$1312.66	/kg fuel
Conversion	$45.04	$61.29	$34.22	$46.08	$29.86	$39.95	/kg fuel
Fabrication	$300.00	$300.00	$300.00	$300.00	$300.00	$300.00	/kg fuel
Total cost	$1028.41	$1343.52	$2292.46	$3051.44	$7212.25	$9642.40	/kg fuel
Interest during use	$157.14	$273.81	$350.29	$621.88	$1102.03	$1965.12	/kg fuel
Total fuel cost	$1185.55	$1617.33	$2642.75	$3673.32	$8314.28	$11,607.52	/kg fuel
	$26.35	$26.95	$58.74	$61.22	$184.80	$193.44	/MW_{th}-day
	$0.322	$0.329	$0.717	$0.747	$2.257	$2.362	/million BTU
	$3.26	$3.33	$7.26	$7.57	$22.85	$23.92	/MW-h
COE increase due to uranium from seawater			$4.00	$4.24	$19.52	$20.58	/MW-h
			4.0	4.2	19.5	20.6	mills per kWe-h

Nebraska, and Texas have been in existence for less than 10 years. Worldwide, approximately 16% of uranium production uses ISL, including all of the production in Uzbekistan and Kazakhstan.

ISL can be used to extract uranium from deposits below the water table in permeable sand or sandstone, provided that the deposits are confined above and below by impermeable strata. Suitable candidates are often roll-front deposits, as described earlier. The uranium minerals are usually uranium oxide or uranium silicate coatings on the individual sand grains. The ISL process replicates, in a few months, the conditions that led to the formation of the roll-front deposit in the sandstone initially.

There are two types of ISL, depending on the chemistry of the sandstone and groundwater. When the ore body is limestone or gypsum (i.e., containing significant amounts of calcium), an alkaline leaching agent, such as sodium bicarbonate and CO_2, must be used. Otherwise, an acid leaching agent, such as weak sulfuric acid plus oxygen at a pH of 2.5 to 2.8 (about the same as vinegar), is preferred. ISL in Australia is primarily acid, whereas ISL in the United States is primarily alkaline.

In general, the uranium is extracted by progressively drilling wells into the deposit on a rectangular grid with 30-m spacing. The leaching fluid is pumped into four wells surrounding a central extraction well, into which a submersible pump has been lowered. The wells are cased to ensure that the fluids do not enter strata above the deposit. In the United States, the production life of an individual alkaline ISL well is typically 6 to 10 months. The most successful operations have extracted 80% of the uranium from the ore. Production life is often limited when the sandstone is plugged by mobilized clay and silt. Sometimes the blockages can be dislodged by

reversing the flow through the field or by increasing the injection pressure.

The uranium is recovered from the pregnant solution in an ion exchange or solvent extraction process. Solvent extraction is preferred if the groundwater is saline, whereas ion exchanges is most effective if the chloride content is less than 3000 ppm. With alkaline leaching, ion exchange is effective until the total dissolved solids reach 3000 ppm. The uranium is then stripped from the resin or solution for further processing.

Before the process solution is reinjected, it is reoxygenated and recharged with sulfuric acid, if necessary, to maintain a pH of 2.5 to 2.8. Approximately 1% of the process solution is bled off to maintain a pressure gradient toward the wellfield. The pressure gradient ensures that groundwater from any surrounding aquifer flows into the wellfield and that ISL mining solutions do not enter the aquifer.

9. IMPACT OF URANIUM SCARCITY AND HIGHER EXTRACTION COSTS

Table VI shows the approximate impact of increases in the price of natural uranium on the cost of electricity from an LWR. In this set of calculations, the cost of natural uranium is set at $30.17 per kilogram (the approximate average price of domestic uranium to U.S. utilities in 2000), with $200 per kilogram representing an optimistic cost of extraction from seawater and $1000 per kilogram representing a more pessimistic (or perhaps more realistic) cost of extraction from seawater. Two burn-ups are shown for uranium price: 45 and 60 MW-day (thermal) per kilogram of initial uranium. The specific power of the fuel remains constant at 37.9 kW_{th} per kilogram of initial uranium, as does the interest charged on the fuel during the fuel cycle. In the cases with higher cost uranium, the tails assay ("tails ^{235}U content") to optimize the balance between raw materials and enrichment costs.

As shown in the final two lines of Table VI, an increase in the cost of natural uranium from $30.17 to $1000 per kilogram results in an increase in the cost of electricity of approximately 20 mills ($0.02) per kilowatt-hour. An increase in the price of natural gas of $3.00 per thousand cubic feet, as occurred in 2000, also results in an increase in the cost of electricity of $0.02 per kilowatt-hour.

9.1 Uranium and Thorium Resources Compared with Future Energy Needs

A simple calculation is needed to place the magnitude of current uranium mining in perspective. If we assume that the world population reaches a steady-state level of 10 billion and that each person consumes energy at the average rate of a U.S. resident in 2001, total world consumption of energy would be approximately 3.7×10^{21} J_{th}. Although that high rate of consumption would probably not be sustainable for a variety of other reasons, the required natural uranium input to a system of fast reactors to produce 3.7×10^{21} J_{th} would be approximately 45,000 t. Average worldwide uranium production, from both mines and the downblending of HEU, now exceeds 45,000 t per year. Expected requirements for the next 20 years or so are 60,000 to 80,000 t per year.

9.2 The Need for Fast Reactors

The early development of fast breeder reactors and of thermal breeder reactors using thorium was driven in part by the apparent scarcity of fissile isotopes to fuel a rapidly growing set of nuclear reactors. Slow growth in nuclear power and large discoveries of natural uranium have reduced the urgency for breeding more fissile material.

However, the use of a fast neutron spectrum or thorium offers unique capabilities for the consumption of the long-lived actinides, particularly plutonium, neptunium, americium, and curium.

In the long term, the capability of a fast reactor to make use of both ^{238}U and ^{235}U will be critical in meeting future energy needs. However, in the coming century, fast reactors will be crucial for the management of actinides and the reduction of the long-term radiotoxicity of the nuclear fuel cycle by at least two orders of magnitude.

9.3 Comparison of Fossil Fuel and Uranium Reserves

It is interesting to compare the cited uranium known reserves and the inventory of depleted uranium with the estimated reserves of coal, oil, and natural gas. In Table VII, the reserves of fossil fuels and uranium are compared on the basis of known economically recoverable reserves.

Fossil fuel reserves are those of the World Energy Council, the *Oil & Gas Journal,* and *World Oil,* as cited by the Energy Information

TABLE VII

Comparison with Fossil Reserves

Coal	1089 billion tons
	$= 2.41E + 22$ J_{th}
	= 22,869 quad
Petroleum	1028 billion barrels
	$= 6.29E + 21$ J_{th}
	= 5,962 quad
Natural gas	5450 trillion cubic feet
	$= 5.95E + 21$ J_{th}
	= 5,641 quad
Total fossil reserves	$3.64E + 22$ J_{th}
	= 34,472 quad
Uranium	3.1 million t natural Uranium
	+ 1.3 million t depleted Uranium
	$= 1.34E + 21$ J_{th} in LWRs
	= 1268 quad in LWRs
	$= 3.57E + 23$ J_{th} in fast reactors
	= 338,462 quad in fast reactors

Note. quad, quadrillion BTU.

Administration. In using Table VII, the caveat cited previously applies: resources of both fossil fuels and uranium are undoubtedly much larger than the reserves cited. Nevertheless, it is noteworthy that the uranium resources are energetically equivalent to approximately 20% of natural gas and oil resources, even with the use of LWRs only. The 1.3 million t of depleted uranium currently in storage itself represents an energy source larger the fossil reserves.

10. SUMMARY

Uranium is ubiquitous in the continental crust and is concentrated in economically recoverable deposits by several relatively well-understood processes. Today, we are mining uranium from the richest and most convenient of the deposits, although little exploration has taken place during the past 20 years. Uranium and thorium are often being extracted as by-products of mining for other elements. It is likely that other similarly rich deposits exist in relatively unexplored regions of Asia and Africa.

Prices in the current uranium market are dominated by large discoveries during the past 20 years and by the conversion of military HEU for civilian purposes. The continued use of nuclear energy and the end of downblending can be expected to raise uranium prices, encourage exploration, and return the uranium to a slightly higher priced equilibrium.

However, because of the wide range of igneous and geochemical processes that are responsible for the formation of uranium deposits, we can expect that uranium will be found in significant quantities with renewed exploration.

Emerging technologies for the extraction of uranium, particular ISL, will make resources in sandstone and shale economically recoverable and will minimize the surface disruption due to open pit mining and the occupational radiation exposures of underground mining.

In the more distant future, the extraction of uranium from seawater will make this fuel available to virtually every nation. Although extraction from seawater is likely to be 10 to 20 times more expensive than uranium is today, the overall increase in the cost of electricity and other energy products would be minimal.

Therefore, the need for fast reactors in the near term is driven not by a scarcity of uranium but rather by a need to effectively manage the long-lived actinides in spent fuel. A fast neutron spectrum is uniquely capable of fissioning the higher actinides and reducing the long-term radiotoxicity of the nuclear fuel cycle. Likewise, although thorium is more abundant than uranium, the primary use of thorium will probably not be for the breeding of ^{233}U; rather, it will be as a host material for the transmutation of the higher actinides.

There are many challenges in the development of safe, proliferation-resistant, and economical reactors and fuel cycles. Fortunately, the uranium and thorium resources do not appear to be a near-term limitation.

SEE ALSO THE FOLLOWING ARTICLES

Nuclear Engineering • Nuclear Fission Reactors: Boiling Water and Pressurized Water Reactors • Radiation, Risks and Health Impacts of • Uranium Mining: Environmental Impact • Uranium Mining, Processing, and Enrichment

Further Reading

Alpher, R. A., Bethe, H., and Gamow, G. (1948). The origin of chemical elements. *Physical Rev.* **37**(7).

Bates, R. L., and Jackson, J. A. (eds.). (1984). "Dictionary of Geological Terms," 3rd Ed. Anchor Books, New York.

Benedict, M., Pigford, T. H., and Levi, H. W. (1981). "Nuclear Chemical Engineering." McGraw–Hill, New York.

COGEMA Resources (2002). "Uranium Joint Ventures: McArthur River/Key Lake." www.cri.ca/uranium/mcarthur.html.

Deffeyes, K. S., and MacGregor, I. D. (1978). "Uranium Distribution in Mined Deposits and in the Earth's Crust." Final report, GJBX-1(79), Department of Geological and Geophysical Sciences, Princeton University.

Deffeyes, K. S., and MacGregor, I. D. (1980). World uranium resources. *Sci. Am.* **242**, 66–76.

Energy Information Administration (2002). "Annual Energy Review 2001," DOE/EIA-0384(2001). U.S. Department of Energy, Washington, DC.

International Atomic Energy Agency (n.d.). "Analysis of Uranium Supply to 2050," IAEA-SM-362/2. IAEA, Vienna, Austria.

Jamieson, B. W., and Frost, S. E. (1997). "The McArthur River Project: High Grade Uranium Mining." www.world-nuclear.-org/sym/1997/jamieson.htm.

Kidd, S. (2001). The World Nuclear Association market report 2001. *The Nuclear Eng.* **42**(6), 191.

Organization for Economic Cooperation and Development, National Energy Agency, and International Atomic Energy Agency (2001). "Uranium 2001: Resources, Production, and Demand." OECD, Paris.

Urey, H. C. (1952). The abundances of the elements. *Physical Rev.* **88**(2).

US Bureau of Labor Statistics. (2001). Energy information administration. *Ann. Energy Review 2001,* November 2002.

US Geological Survey (2001). "Minerals Yearbook 2001." USGS, Washington, DC.

Ux Consulting Company LLC (2003). "World Spot Market Price for U_3O_8." www.uxc.com.

van den Durpel, L. (ed.). (2002). "Trends in the Nuclear Fusion Cycle: Economic, Environmental, and Social Aspects." Nuclear Energy Agency, Paris.

World Nuclear Association (2001). "Uranium Mining in Australia and Canada." www.world-nuclear.org/education/mining.htm.

World Nuclear Association (2002). "Supply of Uranium." www.world-nuclear.org/info/inf75print.htm.

Uranium Mining: Environmental Impact

DONALD R. METZLER
U.S. Department of Energy
Grand Junction, Colorado, United States

1. **Uranium and Its Characteristics**
2. **Recognition of Risks to Human Health and the Environment**
3. **Historical Review of Uranium Ore Mining**
4. **The Cold War Era**
5. **Environmental Impacts in the United States**
6. **Environmental Impacts in Germany**
7. **Remediation in the United States**
8. **Other Remediation Programs**
9. **Protecting the Environment for the Long Term**
10. **Comparison of Decommissioning and Cleanup Costs**
11. **Managing the Long-Term Risk**

Glossary

alpha particle A particle consisting of two protons and two neutrons given off by the decay of many elements, including uranium, plutonium, and radon. Alpha particles cannot penetrate a sheet of paper; however, alpha-emitting isotopes in the body can be very damaging.

atom The basic component of all matter. The atom is the smallest particle of an element that has all of the chemical properties of that element. Atoms consist of a nucleus of protons and neutrons surrounded by electrons.

beryllium The fourth lightest element. Some nuclear weapon parts are made of beryllium.

beta particle A particle emitted in the radioactive decay of many radionuclides. A beta particle is identical to an electron. It has a short range in air and a low ability to penetrate other materials.

Cold War A conflict over ideological differences between the United States and the Soviet Union and their allies lasting from the late 1940s until the early 1990s.

Curie The amount of radioactivity in 1 g of the isotope radium-226. One curie is 37 billion radioactive disintegrations per second.

daughter products Radionuclides that are produced from other radionuclides when they decay.

dose A specific amount of ionizing radiation or a toxic substance absorbed by a living being.

gamma radiation High-energy, highly penetrating electromagnetic radiation emitted in the radioactive decay of many radionuclides. Gamma rays are similar to X-rays.

half-life The time it takes for one-half of any given number of unstable atoms to decay. Each isotope has its own characteristic half-life that ranges from small fractions of a second to billions of years.

heavy metals Metallic elements with high atomic weights (e.g., mercury, chromium, cadmium, arsenic, and lead) that can damage living organisms at low concentrations and tend to accumulate in the food chain. Uranium, thorium, and plutonium are heavy metals.

highly enriched uranium Uranium with more than 20% of the uranium-235 isotope used for making nuclear weapons and as fuel for some isotope production, research, and power reactors. Weapons-grade uranium is a subset of this group.

isotopes Forms of the same chemical element that differ only by the number of neutrons in their nucleus. Most elements have more than one naturally occurring isotope. Many more isotopes have been produced in nuclear reactors and accelerators.

mill tailings The sandlike materials left over from separating uranium from its ore. More than 99% of the ore becomes tailings. Mill tailings typically contain approximately 85% of the radioactivity present in unprocessed ore.

natural uranium Uranium that has not been through the enrichment process. It consists of 99.3% uranium-238 and 0.7% uranium-235.

plume A subsurface zone that contains predominantly dissolved and sorbed contaminants that originate from a contaminant source area. A plume can extend for some distance, depending on groundwater flow and chemistry.

plutonium (Pu) A man-made fissile element. Pure plutonium is a silvery metal heavier than lead. Material rich in the plutonium-239 isotope is preferred for

 © 2004 Elsevier Inc. All rights reserved.

manufacturing nuclear weapons. The half-life of plutonium-239 is 24,000 years.

radiation Energy transferred through space or other media in the form of particles or waves. Certain radiation types are capable of breaking up atoms or molecules. The splitting, or decay, of unstable atoms emits ionizing radiation.

radioactivity The spontaneous emission of radiation from the nucleus of an atom. Radionuclides lose particles and energy through the process of radioactive decay.

radionuclide A radioactive species of an atom. For example, tritium and strontium-90 are radionuclides of elements of hydrogen and strontium, respectively.

radon A radioactive inert gas that is formed by the decay of radium. Radium, in turn, is a link in the decay chain of uranium-238. Radon, which occurs naturally in many minerals, is a main hazard of uranium mill tailings.

thorium A naturally occurring radioactive element.

uranium The basic material for nuclear technology. This element is naturally slightly radioactive and can be refined to a heavy metal more dense than lead.

uranium mill A processing plant in which uranium is separated from mined ore.

Uranium Mill Tailings Radiation Control Act (UMTRCA) of 1978 (U.S. Public Law 95-604) The act that directed the U.S. Department of Energy to provide for stabilization and control of the uranium mill tailings from inactive sites in a safe and environmentally sound manner to minimize radiation health hazards to the public. It authorized the department to undertake remedial actions at 24 designated inactive uranium ore-processing sites and at an estimated 5000 vicinity properties.

Uranium Mill Tailings Remedial Action (UMTRA) Project A program to reduce the hazards posed to the public by uranium mill tailings. The program was created by the U.S. Department of Energy in response to the Uranium Mill Tailings Radiation Control Act enacted in 1978. The U.S. Department of Energy's Office of Environmental Management is responsible for implementing the UMTRA Project.

uranium mining, milling, and refining Mining and milling involves extracting uranium ore from the earth's crust and chemically milling (processing) it to prepare uranium concentrate (U_3O_8), sometimes called uranium octaoxide or yellowcake. Uranium concentrate is refined, or chemically converted, to purify it into the form suitable as feed material for further use.

uranium-238 The heavier of the two main isotopes of uranium. Uranium-238 makes up more than 99% of the uranium ore that is mined. It has a half-life of 4.5 billion years and is not easily split by neutrons.

vicinity properties Locations away from inactive millsites where uranium mill tailings were used for construction or were transported by wind or water erosion.

yellowcake A common uranium compound, U_3O_8, named for its typical color. Uranium is sent from the uranium mill to the refinery in this form.

1. URANIUM AND ITS CHARACTERISTICS

Since the 1940s, uranium mining and milling have involved extracting ore from the earth's crust and chemically processing it to prepare uranium concentrate, sometimes called yellowcake. Processing of the uranium ore resulted in the generation of large volumes of slightly radioactive sandlike residues called mill tailings that typically contain radium, thorium, radon, and nonradioactive heavy metals in low concentrations. Usually, the tailings were deposited as slurry in unlined pondlike structures. Evaporation of the liquid content of the tailings left the fine-grained tailings materials exposed to wind and precipitation (Fig. 1).

To put the legacy waste problem into perspective, the largest residual piles of mill tailings in the United States and Canada may total 30 million tonnes of solid material. Near Nucla, Colorado, the Uravan pile contained more than 14 million tonnes of solids or mill tailings. The Helmsdorf pile near Zwickau in Saxony, Germany, consists of 50 million tonnes of tailings and the Culmitzsch pile near Seelingstädt in Thüringia, Germany, contains 86 million tonnes of solids.

Many tailings piles and ponds were left in a state of uncontrolled abandonment that harmed the environment. Windblown tailings contaminated soils, and radioactive and hazardous constituents migrated into surface waters and groundwaters. Significant efforts

FIGURE 1 In the early years of processing uranium ore in the United States, liquid waste was piped to ponds where the liquid component evaporated, leaving fine sandlike tailings.

have been implemented during the past 20 years to control and stabilize these piles and ponds and to mitigate the adverse environmental consequences. Today, more rigorous regulations and standards for mining and milling uranium ore help prevent negative effects to human health and the environment.

Uranium is a naturally occurring element and a metal that has been identified in approximately 50 minerals. Uranium minerals are not particularly rare. The main ores of uranium are uraninite (pitchblende) and secondary minerals formed from it by weathering. Another important source of uranium ore is carnotite. Carnotite is a complex uranium and vanadium mineral that occurs in sedimentary rocks, is visible as yellow streaks in sandstones, and is common in the southwestern United States. Uranium minerals are widespread in granites and pegmatites but generally not in large concentrations. The average crustal abundance of uranium varies from 2 to 4 parts per million, which is similar to the natural occurrences of arsenic, beryllium molybdenum, and tungsten.

A unique characteristic of uranium is its radioactive isotopes that are unstable and can change or "decay" to other forms. Some of uranium's descendants in its decay chain are more radioactive than uranium and emit a significant level of radiation.

Uranium is represented in the periodic table by the symbol U and the atomic number 92. It is an actinide and one of three naturally occurring isotopes that exist in nature: uranium-234, uranium-235, and uranium-238. The natural abundance (atom percent) of uranium-238 is greater than 99%. All three are long-lived radioactive isotopes with half-lives of 2.45×10^5 years for uranium-234, 7.04×10^8 years for uranium-235, and 4.46×10^9 years for uranium-238. A half-life is the time it takes for half of any sample or amount to decay.

Uranium-235, although occurring in natural uranium to the extent of only 0.71%, is so fissionable with slow neutrons that a self-sustaining fission chain reaction can be made in a reactor constructed from natural uranium and only a suitable moderator, such as heavy water or graphite. Because uranium-235 is the only uranium isotope with fissionable characteristics and because it is found in relatively low natural abundance compared to uranium-238, significant amounts of uranium ore need to be mined (extracted from the earth's crust) and milled (crushed and chemically processed) to produce uranium oxide (U_3O_8) or concentrate as the preparatory step in processing uranium for the nuclear fuel cycle. Yellowcake is the common term for uranium oxide.

Uranium oxide, the product of the milling process, needs further refinement and chemical conversion to purified forms that are suitable as feed materials. The product at this stage is uranium hexafluoride (UF_6), which requires enrichment as the succeeding step in the fuel cycle process. Enrichment is the isotope separation of the fissile material (uranium-235). The fuel cycle material (highly enriched uranium) is destined for fuel and target fabrication, reactor operations, chemical separations, component fabrication, and ultimately weapons operations. The depleted uranium (containing less than 0.711% uranium-235) is used in weapon components and in reactor targets to be irradiated for the production of plutonium-239.

2. RECOGNITION OF RISKS TO HUMAN HEALTH AND THE ENVIRONMENT

Uranium mill tailings are the sandy waste produced by the uranium ore milling process. Because only 1–5 pounds of usable uranium (depending on the ore grade) is extracted from each tonne of ore, large quantities of waste were produced during more than 40 years of processing ore for Cold War efforts, nuclear power plants, and medical uses. These tailings contain many naturally occurring hazardous substances, both radioactive and nonradioactive, and include approximately 85% of the radioactivity present in the unprocessed uranium ore. The greatest threat to public health and safety is the radioactive decay process of this material into radium and radon-222, an inert gas that may cause cancer or genetic mutations.

The dominant hazard from tailings is the radioactive decay products of uranium-238, particularly radium-226 and its short half-life decay products (Table I). Each gram of natural uranium ore contains approximately 500 pCi of uranium-238. In addition, natural uranium ore contains approximately 23 pCi of uranium-235 and 2 pCi of thorium-232. Because these isotopes occur in relatively small proportions and pose less risk to health, uranium-235 and thorium-232 and their radioactive decay products are usually ignored in evaluating the hazard of uranium mill tailings. However, thorium-230, with a half-life of 77,000 years, is a significant decay product because it is the radioactive parent of radium-226.

TABLE I

Decay Chain of Natural Radioactivity Starting with Uranium and Ending with Lead

Isotope	Name	Half-life	Main radiations emitted
U-238	Uranium	4.47×10^9 years	α
↓			
Th-234	Thorium	24.1 days	β, γ
↓			
Pa-234	Protactinium	1.17 min	β, γ
↓			
Pa-234	Protactinium	6.75 h	β, γ
↓			
U-234	Uranium	2.45×10^5 years	α
↓			
Th-230	Thorium	7.54×10^4 years	α
↓			
Ra-226	Radium	1.60×10^3 years	α
↓			
Rn-222	Radon	3.82 days	α
↓			
Po-218	Polonium	3.11 min	α
↓			
Pb-214 and At-218	Lead	26.8 min	β, γ
↓ ↓	Astatine	1.6 s	α
↓			
Bi-214	Bismuth	19.7 min	β, γ
↓			
Po-214 and Tl-210	Polonium	164 μs	α
↓ ↓	Thallium	1.30 min	β, γ
↓			
Pb-210	Lead	22.3 years	α, β
↓			
Bi-210	Bismuth	5.01 days	α
↓			
Po-210 and Tl-206	Polonium	138 days	α
↓ ↓	Thallium	4.20 min	β
↓			
Pb-206	Lead	Stable	—

Uranium tailings emit three types of radiation: alpha (α) particles, beta (β) particles, and gamma (γ) rays. The type of radiation emitted by a nucleus depends on the nature of the imbalance in the nuclear forces, but all types are forms of ionizing radiation that breaks up molecules into electrically charged fragments called ions. Ions are created when an electron is removed from one of the orbital electron shells of an atom. This process is called ionization. In biological tissues, ionization can produce harmful cellular changes. At low radiation levels usually encountered in the environment, we expect the effects of such changes to be difficult to detect. Studies show, however, that individuals exposed to ionizing radiation have a greater chance of developing cancer.

The increased chance of cancer or genetic damage after exposure to radiation cannot be predicted with precision. Health experts have based risk estimates on studies of individuals exposed at doses higher than those associated with uranium mill tailings and on the assumption that the effects will be proportionally less at lower doses. The slope function of the proportionality is not well understood.

Alpha, beta, and gamma radiation from tailings can cause cancer or genetic damage. However, the major threat comes from breathing air containing radon decay products with short half-lives, such as polonium-218, with a half-life of 3.11 min. However, if alpha emitters have been inhaled, ingested (swallowed), or absorbed into the bloodstream, sensitive living tissue can be exposed to alpha radiation. The resulting biological damage increases the risk of cancer; in particular, alpha radiation is known to cause lung cancer in humans when alpha emitters are inhaled.

Human exposure to gamma rays from radioactive material in the tailing piles is possible under uncontrolled settings. Radium is also a source of gamma rays that can travel much farther than alpha or beta particles and have enough energy to pass through the entire human body, potentially exposing all organs. A large portion of gamma radiation passes through the body without interacting with tissue. The human body is mostly empty space at the atomic level and gamma rays are massless uncharged "packets" of electromagnetic energy. In contrast, alpha and beta particles inside the body lose all their energy by colliding with tissue and causing damage.

The body's internal organs would still be exposed to radiation from radionuclides even if all the uranium mill tailings piles suddenly disappeared because radon, radium, uranium, thorium, and other radioactive elements occur naturally in air, rocks, and soils. A typical background concentration of radium is 1 pCi of radium per gram of soil; outdoor air contains a few tenths of a picocurie of radon per liter. Normal eating and breathing introduces these and other radioactive materials into the body, increasing the potential for cancer and genetic changes. Understanding the occurrence of natural radiation and the potential for everyday exposure does not justify the risk of exposure to tailings but does provide a realistic context for comparison.

Tailings also contain toxic elements that could eventually be inhaled or ingested by man or animals or absorbed by plants. Often, hazardous elements leach from the uncontrolled tailings piles and contaminate shallow groundwater.

2.1 Radon and Radon Progeny

Because the milling and extraction processes removed most of the uranium from the ore, the remaining significant radioactive members of the uranium series (Table I) are determined by the presence of thorium-230 and its decay progeny. The thorium-230 decay product, radium-226, has a 1600-year half-life. Both thorium and radium are relatively insoluble and immobile in their usual chemical forms. However, the decay product of radium-226 is radon-222, an inert radioactive gas that readily diffuses through interstitial spaces to the surface of the uncontrolled tailings piles where it becomes airborne. The half-life of radon-222 is 3.8 days, so some radon atoms can travel a long distance through the atmosphere before decaying to nonradioactive lead-210.

The radon decay process involves seven principal decay products before ending with lead, a stable nuclide (Table I). The four short half-life radioactive-decay products immediately following radon in the decay chain are the most important sources of cancer risk. These decay, for the most part, within less than 1 h. Members of the decay chain with relatively long half-lives (beginning with lead-210, which has a 22.3-year half-life) are more likely to be ingested than inhaled and represent a much smaller risk.

The principal short half-life products of radon are polonium-218, lead-214, bismuth-214, and polonium-214. Polonium-218 has a half-life of 3.11 min, which is long enough for most of the electrically charged polonium atoms to attach to microscopic airborne dust particles that are typically less than 1 millionth of a meter across. When inhaled, these small particles have a good chance of sticking to the moist epithelial lining of the bronchial tubes in the lungs.

Most of the inhaled particles are eventually cleared from the bronchi by mucus but not quickly enough to keep the bronchial epithelium from being exposed to alpha radiation from polonium-214 and polonium-218. This highly ionizing radiation passes through and delivers radiation doses to several types of lung cells.

Since the 15th century, middle-aged miners who had worked underground in the European mining district of Schneeberg, Germany, for at least 10 years often died from a pulmonary disease that the miners called *Bergkranheit* or mountain disease. In 1879, doctors diagnosed the disease as lung cancer. By the mid-1930s, industrial health experts understood that radioactivity of the ore and the radon gas in the air were the leading causes of lung cancer among the European miners.

During the boom years of underground uranium mining, many miners were exposed to increased levels of radon by the very nature of their work (Figs 2 and 3). The ore zones were small and confined, and blasting in the stope (underground excavation

FIGURE 2 Burros pulled the ore carts in uranium mines on the Colorado Plateau during the 1950s. At that time, knowledge was limited about the dangers of radon in these mines. Photograph courtesy of the U.S. Bureau of Land Management.

FIGURE 3 A miner uses a jackhammer to drill a blast hole in a uranium mine (United States). Miners often worked without safety glasses and other personal protective equipment in the 1950s. Photograph courtesy of the U.S. Bureau of Land Management.

formed by the extraction of ore) created dusty conditions and suspended radon daughters in the confined air. The ionizing radon tended to attach to solid particles such as dust and cigarette smoke that were inhaled by the miners. The miners and millers often worked in unprotected conditions because of the desire to increase production. Radon exposure was not limited to the mines and mills. Often, unprotected workers carried radon home on their clothes and ingested it as they drank and ate.

2.2 Exposure Pathways

Uncontrolled use of uranium mill tailings may lead to radiation exposure in a number of ways (Fig. 4). Sandy tailings removed from piles and used for backfill and to improve drainage around foundations or for other construction purposes pose the largest hazard by increasing indoor concentrations of radon decay products. Tailings at an uncontrolled millsite emit radon gas into the atmosphere and are a source of radioactive windblown particles and direct gamma radiation.

Because of the chemicals used in uranium processing, the wet discarded tailings contain large volumes of toxic chemicals that leach out of the unlined piles and into the underlying soils and groundwater. The contaminated groundwater becomes a contaminant pathway if humans and animals drink the water. In some cases, the contaminated groundwater discharges to streams and lakes and contaminates surface water.

Movement of contaminants leaching out of uncontrolled piles into underlying soils depends on complex chemical and physical properties of the subsurface environment and on local climatic conditions, such as precipitation and evaporation. Chemical, physical, and biological processes in the subsoil remove a portion of some of the contaminants from soil–water interactions. However, some contaminants are relatively nonreactive and move though the soil at the velocity of water movement. For example, common tailings contaminants such as nitrate, sulfate, and natural uranium are more nonreactive than vanadium, selenium, and molybdenum.

The greatest hazard from uranium mill tailings that have been removed from waste piles and used in construction is their potential to increase levels of radon decay products in buildings. The concentration of radon decay products in a building depends on the amount of radium in the tailings and the location of the tailings in, under, or adjacent to the structure.

Many of the radioactive isotopes in tailings piles and abandoned mines emit gamma radiation. Unlike alpha radiation, which must originate within the body to become hazardous, gamma radiation can travel farther and penetrate both air and tissue. Near the edge of a pile, gamma radiation can be greater than the background level (unaffected) in uncontaminated areas. The gamma radiation from a pile, however, decreases rapidly because of shielding. Data show that at an approximate distance of a few tenths of a mile from a typical uncontrolled uranium ore tailings pile, the gamma ray measurement cannot be differentiated from the normal background, which is 80–100 millirems (mrem) per year.

Levels of radiation from an uncontrolled uranium tailings pile depend on the amount of radium in the

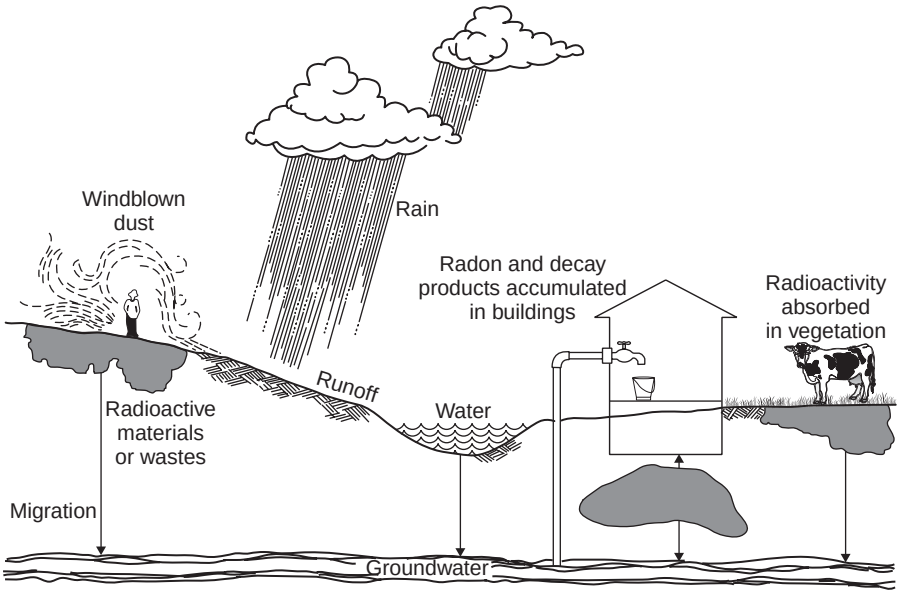


FIGURE 4 Some of the pathways by which environmentally dispersed radionuclides and wastes from uranium mill tailings can affect living organisms.

tailings. Field measurements of gamma radiation on top of tailings piles show levels up to 4000–8000 mrem per year.

3. HISTORICAL REVIEW OF URANIUM ORE MINING

Mining of uranium-bearing ores began in 16th-century mining centers such as the Erzgbirge Mountains (Ore Mountains) of Saxony and nearby Bohemia in Germany. Georgius Agricola (1494–1555), credited as the father of modern geology, artfully illustrated and scrupulously described uranium ore mining in "De Re Metallica." During this period of mining for silver, copper, and other base and precious metals, the uranium-bearing ores were discarded as waste because they had no known economic value.

At the beginning of the 20th century, there were no large-scale commercial uses for uranium. Marie Curie announced the discovery of the 88th element radium at the turn of the century. It was known that radium was associated with uranium ore, and European refiners needed radium for medical uses. Uranium ore was mined for radium, and the uranium was considered a by-product for the most part. At one time, it was thought that uranium was a viable alloy for hardening steel, but this industrial use never materialized. In the 1920s and 1930s, uranium was used in some brands of pottery (Fiesta Ware) to produce an attractive orange color. Prior to World War II, uranium production was not in high world demand because of limited commercial uses, and its scarcity kept the price of the metal too high to promote new industrial development.

Scientists' investigations to unlock the secrets of uranium evolved over 100 years. Martin Heinrich Klaproth isolated "a strange kind of half-metal" from its naturally occurring ore of pitchblende in 1789 in Germany and named it uranium after the planet Uranus. However, the uranium metal was first isolated in 1841 by Eugene-Melchoir Peligot, who reduced the anhydrous chloride uranium chloride (UCl_4) with potassium. The radioactive nature of uranium was not appreciated for another 55 years when Henri Becquerel detected its radioactivity in 1896.

Research by Ernest Rutherford led to publication of a paper in 1904 about the nature of radioactivity and the theory of radioactive decay. Physicists were aware of the potential military uses of nuclear energy by the 1930s, and Otto Hahn and Lise Meitner of Germany determined that uranium could release energy by nuclear fission in 1939. A chain reaction occurs during the nuclear fission process, starting with one nucleus that splits and releases neutrons that collide with other nuclei, causing those nuclei to split and release more neutrons, and the chain reaction continues.

Albert Einstein, a German-born scientist, sent a letter to U.S. President Franklin Roosevelt in 1939

alerting him to the importance of the nuclear research being conducted in Germany. Soon after the United States entered World War II, the U.S. government secretly established the Manhattan Project to develop an atomic bomb. Uranium for this new weapon was acquired from the Belgian Congo, Canada, and the Colorado Plateau in the United States. Deployment of two atomic bombs by the United States in 1945 ended World War II and heralded the arrival of the atomic age and the beginning of the Cold War.

4. THE COLD WAR ERA

Uranium ore mining increased dramatically following the end of World War II and during the Cold War era from 1946 to 1991. The Cold War was a result of growing tensions from the concluding resolutions of World War II. The Soviet Union and the United States were competing in an arms race to have the best, the largest, and the most nuclear weapons. Large quantities of uranium were necessary to produce those weapons.

Uranium was also needed to supply fuel for the nuclear reactors being constructed throughout the world to produce electricity and to replace diesel-powered units in submarines, aircraft carriers, and other naval vessels. Medical uses of uranium products include X-rays, radiation therapies, and radiation isotopes.

At the end of World War II, Germany was a divided country. The Allies controlled West Germany and the Soviet Union controlled East Germany. The Soviet Union knew of the large deposits of uranium ore in Saxony and Thüringia in East Germany. During the Cold War years, both East Germany and West Germany embarked on ambitious government-sponsored programs to explore, mine, mill, and refine uranium ore. The uranium would fuel the nuclear weapons programs. Most of the uranium fuel-cycle program was kept secret during the Cold War by both the Allies and the Soviet Union and its associates. Exploration for uranium ore spread from Eastern Europe and the United States to countries throughout the world. Table II presents production statistics for the top uranium producers during the 1946–1992 period.

Unfortunately, there were no international standards in the past, no best practices, and no environmental laws to protect workers, public health, and the environment and to ensure worker safety. Most of the uranium ore mining was accomplished with conventional mining methods, such as in open pits and underground excavations. In recent years, *in situ* mining methods with chemicals have been employed when the geologic conditions are appropriate. This method has the potential to reduce costs but adds complexity to the environmental consequences because of difficulties in controlling subsurface chemical excursions.

TABLE II
Uranium Production, 1946–1992

Country	Uranium (tonnes)
USA	339,290
Canada	257,692
East Germany	217,791
South Africa	143,302
Czechoslovakia	102,245
Kazakhstan	72,000
France	68,174
Namibia	56,682
Niger	54,143
Australia	25,600
Zaire	25,600
Gabon	22,226
Bulgaria	21,000
Rumania	16,850
Hungary	16,718
India	5920
Total	1,445,233

The principal known deposits of uranium ores in the United States are located in three major districts: the Wyoming Basins, the Colorado Plateau, and the Gulf Coast in Texas. The general area known as the Colorado Plateau covers more than 100,000 square miles in southeastern Utah, western Colorado, northeastern Arizona, and northwestern New Mexico.

Processing of tailings from dumps on the Colorado Plateau supplemented production of uranium from the Belgium Congo and Canada during the early 1940s to meet the military demand. Crude ore milling processes used during this period on the Colorado Plateau reflected little change from methods used 40 years earlier: Uranium was recovered in the milling process from leach solutions based on several stages of selective precipitation. Passage of the Atomic Energy Act of 1946 encouraged prospecting and mining through various forms of assistance. These incentives led directly to an increase in the known mineable uranium ore reserves in the western United States from

approximately 1 million tonnes in 1946 to 80 million tonnes in 1959.

During the peak production years from 1960 to 1962, the number of uranium ore mills in operation in the United States varied from 24 to 26, with the total annual production exceeding 13,000 tonnes of uranium per year from the treatment of 7.25 million-tonnes of ore per year. The number and production rates of the active uranium ore mills decreased from 1962 to 1976 with the cessation of government procurement contracts (Table III).

In the past, the five countries that produced the largest volumes of uranium had distinctively different mining profiles. The United States mined relatively low-grade uranium ores (average 0.03% uranium) by conventional underground methods, and the ore was transported to mills that were often located near waterways. Canada was, and still is, fortunate to possess the highest grade uranium ores in the world (Table IV). In fact, the unusually high grade of some Canadian ores requires special robotic procedures for extraction. Milling of this high-grade uranium ore produces less residual waste and, consequently, fewer adverse environmental impacts.

East Germany mined large quantities of uranium ore during the Cold War (Table II). Wismut, the German Democratic Republic/Soviet Company that mined the ore, produced more than 217,000 tonnes of uranium ore between 1946 and 1990. During peak times in following years, production exceeded 7000 tonnes per year with all uranium products delivered to the Soviet Union. Initially, the uranium was processed exclusively for nuclear weapons, but later it was also used for nuclear power plants. The ore grades were low, even less than those of the sandstone deposits of western United States. Toward the end of the Wismut mining era, the ore grades averaged 0.07%.

In South Africa, uranium ore is a secondary ore associated with gold mining. Large heaps of mill tailings from processing ore to recover gold remain visible in the Transvaal, often near residential areas. Although the tailings are regulated as radioactive, the mining and milling are primarily for other commodities and, therefore, the uranium is more a commodity credit than a stand-alone extraction and processing industry.

The Czech Republic (formerly Czechoslovakia) has mined uranium ore for many years. Diamo, a government-owned company, produced uranium primarily with *in situ* methods using acid. In the early years, Diamo mined uranium ore using conventional underground methods but later converted the mining operations to *in situ* leach methods. The *in situ* methods inflicted serious adverse effects to the environment, with challenging problems to solve. In this case, the environmental effects involve large volumes of low pH-contaminated groundwater requiring active engineered hydraulic controls to contain further migration of the contamination.

TABLE III

U.S. Uranium Production (Tonnes U in Concentrate)

Years	Conventional mining methods	Unconventional mining methods	Total
Pre-1983	281,000	10,000	291,800
1983–1989	23,800	14,250	38,050
1995–2001	Not available	Not available	13,050

TABLE IV

Canadian Uranium Ore Reserves, Grades, and Production Capacity

Mine	Proven reserves (tonnes U_3O_8)	Average ore grade (% U_3O_8)	Annual production (tonnes U_3O_8)
Key Lake	450	0.41	353
Rabbit Lake	10,150	1.30	2070
Cluff Lake	4100	2.7	1519
McClean Lake	19,500	2.40	2990
McArthur River	161,300	21.00	7830

Since the end of the Cold War, world uranium ore mining production has shifted (Table V). Now, Canada produces approximately one-third of the world's uranium. Germany is no longer one of the top 16 producing countries, and the United States has fallen to eighth place. Today, most countries mining and milling uranium ore follow international standards and best practices as provided by the International Atomic Energy Agency (IAEA), and many countries are actively engaged in the cleanup and disposal of the residual wastes.

5. ENVIRONMENTAL IMPACTS IN THE UNITED STATES

The early mills in the United States were built next to rivers within floodplains to harness the water for processing. Most of the tailings left the mill as a

TABLE V

Production from Mines (Tonnes of U_3O_8), 2002

Country	Production
Canada	11,604
Australia	6888
Niger	3075
Russia (estimated)	2900
Kazakhstan	2800
Namibia	2333
Uzbekistan	1860
USA	919
South Africa	824
Ukraine (estimated)	800
China (estimated)	730
Czech Republic	465
India (estimated)	230
Brazil	270
Romania	90
Spain	37
France	20
Others	252
Total world production	36,097

FIGURE 5 This mill, located near Grand Junction, Colorado, processed uranium ore from 1951 until 1970 (three tailings ponds are visible in the foreground). Remediation of the land and buildings began in 1989 and was completed in 1994. The Uranium Mill Tailings Remedial Action Project removed contaminated material, backfilled remediated areas with uncontaminated soil, recontoured the land to promote surface drainage, and revegetated the site.

FIGURE 6 A portion of the remediated site of the former mill near Grand Junction, Colorado, located adjacent to the Colorado River, is now owned by the City of Grand Junction. Site restoration focused on compatibility with long-term plans for development of a riverfront park and trail system. Only one building (circle) remains on the site.

semiliquid waste or slime (fine-grain sandlike material) generated during the milling process. A common practice was to dispose of the radioactive liquid mill effluent directly into the rivers after it had been minimally diluted to reduce radiation and acidic or basic strength. Some tailings were discharged directly into river tributaries. More commonly, mills stored the wet tailings behind massive earthen dikes in structures called tailings ponds (Fig. 5). These containment structures were unlined, and most leaked pore fluids into the shallow groundwater systems beneath the ponds. As the moisture evaporated or seeped out of the ponds, the ponds became generally unstable piles of sand and slimes that were easily eroded by wind and rain.

As the price of uranium declined, many of the mills ceased operations. During the 1960s, going out of business for a uranium mill generally meant abandonment. No controls were in place at the abandoned mills to ensure that the tailings would not interact with the environment and would not be used for construction purposes. Environmental damage occurred for many years until the Uranium Mill Tailings Remedial Action Project, managed by the U.S. Department of Energy, began remedial action to remove, stabilize, and control the contamination (Fig. 6). In many cases, tailings had been taken from the millsites for use as foundation material under buildings and as an additive in concrete. Radium, which decays to radon gas, and radon progeny remained in the tailings. Risks to human health were identified by how the tailings were used in habitable structures, such as in mortar. More than 5000 of these contaminated properties, called vicinity properties, were identified and remediated in the United States.

In some cases, the tailings were left in place and the condition for nonremoval of the contamination

was designated as supplemental standards. An example is tailings used as cover material around underground utilities or under public streets. The U.S. Department of Energy and the state of Colorado cleaned up more than 4000 vicinity properties within the city boundaries of Grand Junction. An "engineered hole" was left in the disposal cell constructed near Grand Junction for future tailings disposal because additional tailings will be uncovered in the community during the next 25 years, primarily around utility lines.

The problem of radiological pollution from former uranium ore mill operations was first identified in the surface waters of the Colorado River Basin. Radium was the most hazardous of the elements in the tailings and was therefore of primary concern to public health officials. From a radiological standpoint, tailings were more dangerous than unprocessed uranium ore because tailings particles are extremely small. Water could more easily dissolve the radium component in the tailings than the radium component in its natural state in uranium ore.

During the 1950s, the U.S. Public Health Service identified radium concentrations that were more than 100 times the background level in samples from rivers in which raffinate liquids had been discharged by the uranium ore-processing mills. Aquatic insects collected during a 1955 survey revealed radium accumulation as high as 60 times normal background levels; algae samples showed radium concentrations as high as 100 times normal levels. By the 1970s, all the mills in the United States had stopped discharging raffinate liquids directly to river courses. As an alternative, the mills adopted the standard practice of discharging the contaminated liquids to unlined ponds. When contained in unlined ponds, contaminated liquid would partly seep into the underlying shallow groundwater and partly evaporate. Of the 24 sites specified in the Uranium Mill Tailings Radiation Control Act, 23 had associated contaminated groundwater.

6. ENVIRONMENTAL IMPACTS IN GERMANY

Disposal practices for wastes in other countries were similar to those in the United States. In Germany, some of the most egregious environmental impacts occurred in Saxony and Thüringia. Successful prospecting for uranium in the immediate postwar period induced the Soviet Union, the occupying power in East Germany, to launch Wismut, a state-run nonferrous metal stock company based in Aue (a small mining town in Saxony) in 1947. Managed by units of the Soviet Union military, the company's only aim was the rapid exploitation of East German uranium deposits for the Soviet Union's nuclear program.

To cover reparation claims of the war, Saxon mining companies were transferred to Soviet Union ownership; land was seized arbitrarily and became a "no go" area. An enormous workforce (peak of 130,000 employees) was employed, and uranium mining operations rapidly spread throughout the Ore Mountains and into eastern Thüringia. Poor working conditions and complete disregard for the environment characterized mining of uranium ore in these densely populated areas.

The binational Soviet–German company Wismut, founded in 1954, continued uranium mining, with the German Democratic Republic sharing equity capital. Intense exploration efforts increased Wismut's reserves as new deposits were discovered. Working conditions were gradually improved, and environmental protection legislation introduced during the 1960s enhanced awareness of radiation protection requirements.

Technology developments had an increasingly positive influence on production processes, reducing the workforce to a permanent staff of approximately 45,000 in later years. Market economy conditions after German reunification put a sudden end to uranium mining on December 31, 1990. Soviet Union personnel left the mines and mills, and Wismut and its environmental legacy became the sole possession of the Federal Republic of Germany.

Following the restructuring of Wismut, the company was legally changed at the end of 1991 to a company of German corporate law under the provisions of the Wismut Act. The corporate purpose of Wismut GmBH shifted from uranium production to decommissioning of the former mining holdings and remediation of the land that had been adversely affected by decades of unrestrained mining and milling of uranium ores. This remediation project will cost 6.6 billon euros, to be provided by federal funding, and it will take 15 years to complete the large-scale environmental restoration requirements (Table VI).

Exposure levels were measured at all the Wismut sites and vicinity properties. A total of 11,885 ha, including 8223 ha of vicinity properties associated with the nearby communities, has been investigated. Characterization data show that 85% of the radioactive dose rate representing all the Wismut GmBH sites is less than 200 nanoseiverts per hour.

TABLE VI

Summary of Wismut Remedial Operation Units at Beginning of Remedial Action

Site	Subsites	Unit area (ha)	Waste piles (ha)	Tailings ponds (ha)
Aue	Aue, Pohla	569.4	342.3	10.4
Koigsteinn	Konigstein, Dresden-Gittersee	144.8	40.0	4.6
Ronneburg	Ronneburg	1670.7	604.4	9.0
Seelingstadt	Seelingstadt, Crossen	1314.8	533.1	76.7
Total		3699.7	1519.8	730.7

7. REMEDIATION IN THE UNITED STATES

The demise of the United States uranium industry began quietly in 1958, when the U.S. Atomic Energy Commission withdrew its offer to purchase uranium from ore reserves developed thereafter. This action led either to the shutdown of mills after expiration of contracts or to continued mill production through so-called stretch-out contracts that were extended through 1970. Stretch-out contracts reduced government purchases and served to ease the uranium ore mining and milling industry through the transition period when nuclear power growth had not progressed sufficiently to create a significant commercial demand. United States production totaled approximately 17,000 tonnes of uranium oxide in 1980, but less than 1000 tonnes of uranium was mined and milled in 2002 (Table V).

When the industry realized that no more nuclear power plants were being constructed and that demand for uranium was greatly reduced, mining and milling of uranium ore again drastically declined. By the end of 1984, only eight conventional mills were still operating. Many of the unconventional uranium recovery plants were also inactive. Today, only a few mines in the United States are being worked, and these are mostly unconventional operations that employ *in situ* leach methods. Canada and Australia continue to produce more than half of the world's uranium (Table V).

The United States was the first country to implement a program to assess the risks of uranium mill tailings and begin cleanup of the adverse effects to the environment caused by past uranium ore milling practices. In the United States, the hazards posed by mill tailings were not completely recognized in the early years of production by the uranium industry. The Atomic Energy Act of 1954 instituted licensing of mill operations, but the tailings remained free of controls. Although numerous studies had assessed the hazards of uranium mill tailings and several federal agencies and states had acknowledged a need for controls, a comprehensive program to address the problem was not started until the late 1970s.

In 1974, the subcommittee on raw materials of the Joint Committee on Atomic Energy conducted a hearing to discuss two congressional bills. These bills proposed that the U.S. Atomic Energy Commission (later designated the Energy Research and Development Administration and now called the U.S. Department of Energy) and the state of Utah jointly assess and act to limit public exposure to radiation from a uranium mill tailings site located in South Salt Lake City, UT.

The U.S. Environmental Protection Agency endorsed the objectives of the bill but, with the U.S. Atomic Energy Commission, recommended instead that the two agencies, in cooperation with the affected states, assemble comprehensive studies of all inactive uranium mill tailings sites. The studies would be divided into two phases. Phase I studies would establish the condition of the sites, ownership, surrounding areas, and the need, if any, for more detailed studies. Phase II studies would, as needed, evaluate the hazards and analyze disposal alternatives and their costs. The U.S. Congress accepted this proposal. Phase I studies began in 1974, followed by phase II studies in 1975. Both studies were completed by 1978.

On the basis of these studies, the Congress enacted the Uranium Mill Tailings Radiation Control Act of 1978 (Public Law 95-604, 42 United States Code 7901). Congress found the following:

> *Uranium mill tailings located at active and inactive mill operations may pose a potential and significant radiation health hazard to the public.*
>
> *Every reasonable effort should be made to provide for the stabilization, disposal, and control in a safe and environmentally sound manner of such tailings in order to prevent or minimize radon diffusion into the environment and to prevent or minimize other environmental hazards.*

The act established a 7-year completion deadline for the project and outlined the funding formula for the cleanup. Participating states were responsible for 10% of the cost of remedial action, and the federal government would pay the remaining 90%. For those sites on Native American lands, the federal government was responsible for 100% of the cost. The Department of Energy would manage cleanup of the sites and the contaminated groundwater under the Uranium Mill Tailings Remedial Action Project. The Uranium Mill Tailings Radiation Control Act originally specified 24 former uranium ore processing sites for cleanup, but one state (North Dakota) withdrew 2 sites. The act requires the Department of Energy to keep community leaders informed of plans and activities involved in the remedial actions.

Remedial actions at the sites designated in the act began in the early 1980s (Figs. 5 and 6). By 1998, 22 disposal cells had been constructed for permanent entombment of the wastes. Ten of the disposal cells were constructed on former uranium processing sites, and 12 disposal cells were sited at locations that were more advantageous. Nearly all these former uranium ore processing sites are located in the arid west. Two commonalities associated with the arid west are low annual precipitation and vast open areas with few inhabitants. This scenario contrasts with that of the German Wismut program because the mining and processing sites are subjected to high annual volumes of precipitation and few uninhabited areas are available within the country. Therefore, relocation of tailings and other contaminated material has limited potential.

Most older uranium ore processing sites have a component of relic-contaminated groundwater related to chemicals used in the recovery processes. The contaminated groundwater is generally limited to shallow unconfined aquifers. Contaminants of concern are identified by potential risk based on toxicities and pathways to receptors. Ammonia, arsenic, cadmium, gross alpha, molybdenum, nitrate, radium, selenium, sulfate, and uranium are the most common groundwater contaminants associated with older uranium processing sites. Of these, nitrate, sulfate, and uranium are the most mobile and tend to move with the average velocity of groundwater. At many of the former processing sites, the groundwater contaminant of most concern is not a radionuclide but a stable inorganic, such as nitrate or sulfate.

Contaminated groundwater is a concern at most sites, especially those sites in arid regions where groundwater is a precious and valuable resource. At some sites, such as the Tuba City, Arizona, site located on the Navajo Indian Reservation, the contaminated groundwater occurs in a sole-source aquifer (the Navajo aquifer). This aquifer provides drinking water and irrigation water for northeastern Arizona and is noted as a high-quality water resource. A pump-and-treat groundwater remedial action was required at the Tuba City site to reduce the contaminant concentrations to acceptable levels. Active engineered treatment of the contaminants in the groundwater is expected to take at least 20 years employing extraction wells and an ion exchange system, coupled with a distillation unit and reinjection into the subsurface of the treated groundwater.

Contaminated groundwater sites near rivers are good candidates for natural flushing when modeling with computer software predicts that the contaminants will decrease to acceptable levels through natural attenuation processes and dilution within 100 years. Institutional controls and verification monitoring will be required at these sites. Institutional controls are "checks and balances" that effectively protect public health and the environment by reducing exposure to or reducing health risks by preventing inappropriate intrusion into contaminated groundwater or by restricting access to or use of contaminated groundwater for unacceptable purposes (e.g., domestic household use). One form of an institution control is a deed restriction on the property.

The global experience of using active engineered methods and technologies to reduce the contaminant levels to premilling water quality demonstrates that it is extremely difficult, if not impossible, to achieve complete restoration to pristine conditions. This fact is one more reason why today the emphasis is placed on prevention rather than intervention in protecting the environment.

8. OTHER REMEDIATION PROGRAMS

Wismut has become a global leader in uranium mill tailings cleanup, along with the United States, Canada, and other countries (Fig. 7). In recent years, Germany and the United States have emphasized integrating technical and nontechnical factors into remediation decisions. One example of integration is tailings cover designs. Environmental remediation factors for cover design extend beyond geotechnical and erosion criteria and include considerations such as esthetics. Will the cover material blend into the

FIGURE 7 The worked-out open pit at the Ronneburg Wismut uranium mine at Thηrgen, Germany, was backfilled with waste rock from mining operations.

FIGURE 8 Former waste rock piles near Schlema, Germany, generated from processing of uranium ore were recontoured and covered with uncontaminated soil.

surrounding environment? Will the cover be attractive to the public and nearby residents?

One of the Wismut remediation projects in Schlema, Saxony, has been advertised as "toxic town to health spa." Schelma is a small town southwest of Chemnitz in Saxony, a region in eastern Germany that has a southerly border with the Czech Republic. In the 1920s, individuals afflicted with rheumatism used a radioactive spring near the town. After World War II, the Soviet Union began mining uranium ore, and the town and former spa were sacrificed as a mining production zone. The mines were closed in 1990, and more than 45,000 jobs were lost.

The Wismut remediation program for Schelma considered a number of options: seal off the town and surrounding contaminated areas, continue operating the mine in a noncompetitive market, or convert the town into a world-class health spa featuring the radioactive springs. In 1992, Wismut began to decommission the former mine workings and actively rehabilitate the land (Fig. 8). The Schelma radon-containing springs and spa gardens are an example of integrating the technical requirements with land use decisions and considering esthetically pleasing factors. This approach to environmental restoration is becoming the norm today rather than the exception.

In Canada, uranium mine and mill tailings constitute the largest inventory by volume of radioactive waste. Uranium has been mined since the early 1930s in Canada, and approximately 225 million tonnes of tailings has accumulated in Ontario province, Saskatchewan province, and the Northwest Territories. This volume of tailings constitutes a small fraction of all nonuranium mine and mill tailings in Canada. Uranium ore is still actively mined in Canada, but the volume of tailings is expected to increase only marginally in the future because the current higher grade of uranium ore will generate lower volumes of mine and mill tailings.

The uranium mine and mill tailings in Canada are currently managed by a few large mining companies, and the tailings are held in containment areas close to the mine sites. Because of their relatively large volumes, the tailings are generally disposed of where they are deposited during operations, typically in tailings ponds or in mined-out pits. The final decommissioning of these tailings and the associated costs are the responsibility of the uranium mining companies and will be regulated in accordance with Canadian Atomic Energy Control Board regulations.

9. PROTECTING THE ENVIRONMENT FOR THE LONG TERM

In the United States, tailings must have a disposal cover designed to control radiological hazards for a minimum of 200 years and for 1000 years to the greatest extent reasonably achievable. Disposal covers must also limit radon (radon-222) releases to 20 $pCi/m^2/s$ averaged over the disposal area. Radon-release limitation requirements apply to any portion of the tailings disposal site unless radium concentrations do not exceed 5 pCi/g in the first 15 cm below the surface and 15 pCi/g in layers more than 15 cm below the surface.

No law exists in Germany specific to the management of the uranium-mining legacy. The German

government favors the Canadian approach of site-specific decisions. The Wismut program rehabilitates the uranium tailings, heap piles, mines, and other contaminated media based on site-specific risk-based analyses that demonstrate protection of human health and the environment.

10. COMPARISON OF DECOMMISSIONING AND CLEANUP COSTS

Wismut and other remediation programs investigated remedial action costs associated with the cumulative uranium production of 14 major uranium-producing countries through 1992 (Table VII). Cumulative production resulted in approximately 1.14 million tonnes of uranium or nearly 3 billion pounds of uranium oxide (U_3O_8), corresponding to approximately 63% of the cumulative world production of uranium. This production is accompanied by approximately 1681 million tonnes of processing residues, of which 780 million tonnes is related to processing plants that produced or still produce uranium as a by-product. The accumulated and estimated costs for the decommissioning and rehabilitation of the uranium-producing facilities in this study total approximately $3.7 billion (all costs are based on 1993 U.S. dollars). The resulting specific rehabilitation costs are $1.25 per pound of U_3O_8 and $2.20 per tonne of tailings. Omitting plants that produce or produced uranium as a by-product of gold and copper production, the specific cost per tonne of milling doubles to approximately $4.00.

These values, however, are only indicative because the specific costs for different projects and countries vary widely, ranging from a minimum of $0.12 per pound of U_3O_8 (Canada) and $0.12 per tonne of tailings (South Africa) to a maximum of approximately $40.00 per pound of U_3O_8 (Sweden) and $68.00 per tonne of tailings (Uranium Mill Tailings

TABLE VII

Decommissioning and Cleanup Costs of Uranium-Producing Projects

Country	Production (tonnes of U, including 1992)	Tailings (million tonnes)	Total cost (million U.S. $, 1993)	Specific cost	
				U.S. $ per tonne of tailings	U.S. $ per pound of U_3O_8
Australia	54,225	98.7	85.10		0.60
Bulgaria	21,871	23.0	173.10	7.53	3.04
Canada	257,702	160.6	77.10	0.48	0.12
Czech Republic	101,901	48.8	433.33	8.88	1.64
France	70,038	31.4	128.45	4.09	0.71
Gabon	21,446	6.5	30.13	4.64	0.54
Germany	218,463	160.2	7893.9	49.50	13.90
Hungary	19,970	19.0	78.40	4.13	1.51
Namibia	53,074	350.0	53.20	0.15	0.39
Niger	56,845	17.2	79.87	4.64	0.54
South Africa	143,305	700.0	81.97	0.12	0.22
Spain	1145	1.2	14.82	12.35	4.98
Sweden	200	1.5	20.98	13.99	40.35
USA (total)	310,000	222.9	2428.96	10.90	3.01
Uranium Mill Trailings Radiation Control Act (UMTRCA) Title I sites (federal government)	56,000	31.3	2140.00	68.37	14.70
UMTRCA Title II sites (private industry)	254,000	191.6	288.96	1.51	0.44

Radiation Control Act Title I Program in the United States). This large range substantiates that remediation costs depend on numerous factors, such as upfront environmental studies, types of contaminants, concentration levels of contaminants, on-site or relocation disposal options, and required level of protection.

11. MANAGING THE LONG-TERM RISK

Consequences of uranium mining and milling will remain for a long time. The half-lives of uranium, radium, and thorium are long lived. Disposal considerations cover millenniums, but managing uncertainty for millenniums becomes impossible to understand with precision and accuracy. Are today's best management practices adequate environmental protection for more than 1000 years or are some countries going overboard with conservatism? This question will take time to answer. Over time, verifying the performance of disposal cells, checking cultural and economic paradigms shifts, adjusting to climate change, and understanding how technologies will change the paradigm of future generations are factors that will form the scorecard on judging our decisions for addressing the uranium mining and milling environmental legacy during the 1980s, 1990s, and early 21st century.

The IAEA, with more than 130 member states, is an independent, intergovernmental science and technology-based organization in the United Nations family that serves as the global focal point for nuclear cooperation. One of the missions of the organization is to develop nuclear waste management standards and, based on those standards, promote the achievement and maintenance of high levels of safety in applications of nuclear energy as well as the protection of human health and the environment against ionizing radiation.

Within the IAEA, the Waste Technology Section is responsible for fostering technology transfer, promoting information exchange and cooperative research, and increasing the number of member states to manage radioactive wastes from both the nuclear fuel cycle and nuclear applications. The agency routinely meets with member states for discussions concerning radioactive waste issues while emphasizing that radioactive waste is an unavoidable remnant from the use of radioactive materials and nuclear technology. IAEA representatives review practices involving industrial activities with naturally occurring radioactive material, such as the mining and milling of naturally radioactive ores.

The IAEA stresses that all radiation sources, including radioactive waste, are potentially hazardous to health and, therefore, must be managed to protect human health and the environment. The organization approaches the problem by acknowledging that a relatively small fraction of radioactive waste is routinely released into the environment in the form of discharges that need to be properly controlled. Some amounts may remain in the environment as radioactive residues, particularly after the termination of past operations and the decommissioning of installations, and may require restoration of the affected environments. The bulk of radioactive waste must be disposed of in repositories isolated from human and animal intrusions.

Technical meetings sponsored by the IAEA in Vienna, Austria, and in countries with significant uranium ore mining and milling operations share member states' experiences, lessons learned, and technical information. Today's challenges are final disposition of uranium wastes, risk reduction, long-term performance of disposal facilities, cost efficiencies, and archiving of records. The lessons learned and the experiences gained during the past 25 years of studying the effects of uranium mining and milling operations and conducting remedial actions to remedy those effects can help prevent negative impacts in the future.

SEE ALSO THE FOLLOWING ARTICLES

Nuclear Engineering • Nuclear Fission Reactors: Boiling Water and Pressurized Water Reactors • Nuclear Power: Risk Analysis • Occupational Health Risks in Nuclear Power • Radiation, Risks and Health Impacts of • Uranium and Thorium Resource Assessment • Uranium Mining, Processing, and Enrichment

Further Reading

Agricolo, G. (1950). "De Re Metallica" (H. C. Hoover and L. H. Hoover, Trans.). Dover, New York.

De Voto, R. H. (1978). "Uranium Geology and Exploration." Colorado School of Mines, Golden, CO.

Federal Ministry of Economics and Technology (2000). "Wismut, an Ecological Challenge, New Horizons through Remediation." Druckhaus, Berlin.

Ford, Bacon and Davis Utah (1976–1978). Phase II Title I, Engineering assessment of inactive uranium mill tailings, 20 contract reports for the U.S. Department of Energy. Ford, Bacon and Davis Utah, Salt Lake City, UT.

International Atomic Energy Agency (2002). "Preliminary final report on the Coordinated Research Project: The long-term stabilisation of uranium mill tailings." International Atomic Energy Agency, Vienna, Austria.

Lavore, M. (1987). "Radon, the Invisible Threat." Rodale, Emmaus, PA.

Merritt, R. C. (1971). "The Extractive Metallurgy of Uranium." Colorado School of Mines Research Institute, prepared under contract with the U.S. Atomic Energy Commission, Golden, CO.

Mogren, E. W. (2002). "Warm Sands, Uranium Mill Tailings Policy in the Atomic West." Univ. of New Mexico Press, Albuquerque, NM.

Murray, R. L. (1989). "Understanding Radioactive Waste." 3rd Ed. Battelle, Columbus, OH.

Organization for Economic Cooperation and Development and International Atomic Energy Agency (1989). "Uranium resources, production, and demand." A joint report by the Organization for Economic Cooperation and Development and International Atomic Energy Agency, Paris.

U.S. Department of Energy (1996). "Final Programmatic Environmental Impact Statement for the Uranium Mill Tailings Remedial Action Ground Water Project." U.S. Department of Energy, Albuquerque Operations Office, Albuquerque, NM.

U.S. Department of Energy (1997). "Linking Legacies, Connecting the Cold War Nuclear Weapons Production Processes to Their Environmental Consequences." U.S. Department of Energy, Washington, DC.

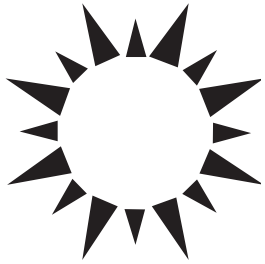
U.S. Department of Energy (1999). "Uranium Mill Tailings Remedial Action Surface Project 1979–1999, End-of-Project Report." U.S. Department of Energy, Uranium Mill Tailings Remedial Action Project, Albuquerque Operations Office, Albuquerque, NM.

U.S. Department of Energy (1999). "Uranium Mill Tailings Remedial Action Surface Project 1978–1998, Disposal Cell Design." U.S. Department of Energy, Uranium Mill Tailings Remedial Action Project, Albuquerque Operations Office, Albuquerque, NM.

U.S. Department of Health, Education, and Welfare (1966). "Disposition and Control of Uranium Mill Tailings Piles in the Colorado River Basin." U.S. Department of Health, Education, and Welfare, Washington, DC.

U.S. Environmental Protection Agency (1982). "Final Environmental Impact Statement for Remedial Action Standards for Inactive Uranium Processing Sites." U.S. Environmental Protection Agency, Washington, DC.

Uranium Mining, Processing, and Enrichment



IAN HORE-LACY
World Nuclear Association
London, United Kingdom
Uranium Information Centre
Melbourne, Australia

1. Uranium Mining, Milling, and Processing
2. Uranium Refining and Conversion
3. Uranium Enrichment

Glossary

enrichment The physical process of increasing the proportion of U-235 to U-238.
isotope An atomic form of an element having a particular number of neutrons. Different isotopes of an element have the same number of protons but different numbers of neutrons and hence different atomic mass (e.g., U-235, U-238).
mill Treatment plant producing concentrate from ore.
ore Rock containing economic concentrations of minerals.
radon An inert but radioactive gas which is a decay product of radium.
stope Cavity formed underground by drilling, blasting, and removing ore.
tailings Crushed and ground rock from which economic minerals have been recovered. Uranium tailings are radioactive due to decay products that remain.

There are many uranium mines operating around the world, in some 20 countries, though almost three-quarters of world production comes from just 10 mines. Most of the uranium ore deposits at present supporting these mines have average grades in excess of 0.10% of uranium. In the first phase of uranium mining to the 1960s, this would have been seen as a respectable grade, but today some Canadian mines have huge amounts of ore up to 20% U average grade. Some uranium is also recovered as a by-product with copper, as at Olympic Dam in Australia, or as by-product from the treatment of other ores, such as the gold-bearing ores of South Africa. In these cases, the concentration of uranium may be as low as a tenth of those in orebodies mined primarily for their uranium content. Olympic Dam quotes ore reserves down to 0.045% U.

1. URANIUM MINING, MILLING, AND PROCESSING

1.1 Overview

Generally speaking, uranium mining is no different to other kinds of mining unless the ore is of a very high grade. Where orebodies lie close to the surface, they are usually accessed by open cut mining, involving a large pit and the removal of much overburden as well as a lot of waste rock. Where orebodies are deeper, underground mining is employed, involving construction of access tunnels and shafts but with less waste rock removed and less environmental impact. In either case, the grade is usually controlled by measuring radioactivity as a surrogate for uranium concentration. (The radiometric device detects associated radioactive minerals, which are decay products of the uranium rather than the uranium itself.)

At Ranger in north Australia, Rossing in Namibia, and most of Canada's Northern Saskatchewan mines through to McClean Lake, the orebodies have been accessed by open cut mining. Other mines such as Olympic Dam in Australia, McArthur River, Rabbit Lake and Cigar Lake in Northern Saskatchewan, and Akouta in Niger are underground, up to 600 m deep. At McClean Lake, mining will be completed underground. Table I lists the largest-producing uranium mines in the West.

 © 2004 Elsevier Inc. All rights reserved.

TABLE I

The Largest-Producing Western World Uranium Mines in 2001

Mine	Country	Main owner	Type	Production (tU)	Percentage of world
McArthur R.	Canada	Cameco	Underground	6937	19.4
Olympic Dam	Australia	WMC	By-product/underground	3693	10.3
Ranger	Australia	ERA (Rio Tinto 68%)	Open pit	3564	10.0
McClean L.	Canada	Cogema	Open pit	2512	7.0
Rossing	Namibia	Rio Tinto (69%)	Open pit	2239	6.3
Akouta	Niger	Cogema/Onarem	Underground	2027	5.7
Rabbit Lake	Canada	Cameco	Underground	1755	4.3
Cluff Lake	Canada	Cogema	Open pit/underground	1306	3.6
Arlit	Niger	Cogema/Onarem	Open pit	1069	3.0
AngloGold	South Africa	Anglo American	By-product/underground	840	2.9
Top 10 total				25,942	72.5

Some orebodies lie in groundwater in porous material and may be accessed simply by oxygenating that groundwater and pumping it out—this is *in situ* leach (ISL) mining. ISL mining means that removal of the uranium minerals is accomplished without any major ground disturbance. Weakly acidified or alkaline groundwater with a lot of oxygen injected into it is circulated through an enclosed underground aquifer, which holds the uranium ore in loose sands. The leaching solution with dissolved uranium is then pumped to the surface treatment plant.

Over half of the world's uranium now comes from underground mines, about 30% from open cut mines and 16% from ISL. Conventional mines have a mill where the ore is crushed, ground, and then leached with sulfuric acid to dissolve the uranium oxides. Most of the ore, however, remains undissolved in the leaching process, and these solids or tailings are then separated from the uranium-rich solution, usually by allowing them to settle out. At the mill of a conventional mine, or the treatment plant of an ISL operation, the uranium is then separated by ion exchange before being dried and packed.

1.2 Hard Rock Mining

With open cut mining, the overburden needs to be removed to expose the orebody, the part of the rock that contains economic concentrations of the minerals being sought. This, and subsequent mining, is undertaken by drilling and blasting, followed by removal of the broken rock. At open cut uranium mines, the drill cuttings from mine development may be tested radiometrically so that ore and waste (below ore-grade material) are blasted separately. In addition, loads of mined material are often scanned radiometrically to ascertain average ore grade, and trucks are then directed to stockpiles, primary crusher, or waste dump accordingly.

At Canada's McClean Lake mine, waste is material with less than 0.025% U or less than 0.085% U if it has high acid-generating potential and requires special disposal. Anything over 0.085% U is unequivocally ore, but is stockpiled according to grade so that mill feed may be varied according to market conditions.

With underground mining, the objective is to move as little waste rock as possible in sinking shafts or declines and digging tunnels (drives and adits) accessing the orebody as fully as possible. Mining the actual orebody is then generally more selective. Various mining techniques are employed, but open stoping is common. This involves carefully drilling and then blasting a large mass of rock so that the broken material can be removed. When the void is empty, it is usually filled from above with waste rock and coarse tailings material, with some cement and flyash binder, to maintain structural support and enable extraction of adjacent ore.

At Olympic Dam, mining is by longhole open stoping, back-filled after ore extraction. This means that a block of ore is drilled from below and above, then blasted so that up to 80,000 tonnes is broken at once. Each stope will yield up to 500,000 tonnes. Ore is extracted from drawpoints at the bottom of the stope and taken to the underground primary crusher. Crushed ore is hoisted to the surface in skips.

At McArthur River in Canada, the orebody is so high-grade (up to 20% U) that remote methods need to be employed for mining it because the minerals

associated with the uranium are also so concentrated that they emit high levels of gamma radiation. It is also about 600 m deep and in wet, porous rock, so control of water under high pressure is another problem. Manned operations are confined to the barren rock surrounding the orebody, and water is controlled by freezing all around the orebody with some grouting. The orebody is than accessed remotely by raise boring, using a 3-m diameter reaming head. Remotely operated front-end loaders remove the ore cuttings that fall from this. Each load is radiometrically scanned, and waste rock or low-grade mineralized material is hoisted to the surface for disposal. While this whole exercise is high cost per tonne of ore removed, the ore is of such a grade that relatively little is involved, about 120 tonnes per day.

The low volume also means that the ore cuttings are fed straight to an underground grinding circuit, and the ground slurry is then pumped to the surface. It is then loaded into special transportation tanks and road-hauled 80 km to the mill for further treatment. Meanwhile, the void from the raise boring is filled with concrete to provide ground support for subsequent raise boring operations.

At the Cigar Lake underground mine in Canada, poor ground conditions mean that water jet boring will be employed instead of raise boring, with ground freezing. Below ore-grade material with significant mineralization from open cut mines has in the past been heaped up on an impervious pad and leached with sulfuric acid. This heap leaching produces pregnant liquor, which is fed to the mill for uranium extraction. However, the process can give rise to significant problems of rehabilitation, so it is apparently no longer used. At Australia's Nabarlek mine in the 1980s, about 50% of the uranium recovery was achieved from 150,000 tonnes of heap leach material.

1.3 *In Situ* Leach (ISL) Mining

In situ leaching (ISL), also known as solution mining, involves leaving the ore where it is in the ground and using liquids that are pumped through it to recover the minerals out of the ore by leaching. Consequently there is little surface disturbance and no tailings or waste rock generated. However, the orebody needs to be permeable to the liquids used and located so that they do not contaminate groundwater away from the orebody.

Some ISL mining in the past, notably in the eastern bloc, has occurred in broken rock with insecure containment of fluids, and this has resulted in considerable pollution. ISL mining was first tried on an experimental basis in Wyoming during the early 1960s. The first commercial mine began operating in 1974. About a dozen projects are licensed to operate in the United States (in Wyoming, Nebraska, and Texas), and most of the operating mines were less than 10 years old in the early 21st century. Most are small, but they supply some 85% of the U.S. uranium production. About 16% of world uranium production is by ISL (including all Kazakhstan and Uzbekistan output).

ISL can also be applied to other minerals such as copper and gold. Uranium deposits suitable for ISL occur in permeable sand or sandstones, confined above and below by impermeable strata and below the water table. They may either be flat, or roll front—in cross section, C-shaped deposits within a permeable sedimentary layer. Such deposits were formed by the lateral movement of groundwater bearing oxidized uranium minerals through the aquifer, with precipitation of the minerals occurring when the oxygen content decreased, along extensive oxidation-reduction interfaces. The uranium minerals are usually uraninite (oxide) or coffinite (silicate) coatings on individual sand grains. The ISL process essentially reverses this ore genesis in a much shorter time frame. There are two operating regimes for ISL, determined by the geology and groundwater. If there is significant calcium in the orebody (as limestone or gypsum), alkaline (carbonate) leaching must be used. Otherwise, acid (sulfate) leaching is generally better.

Techniques for ISL have evolved to the point where it is a controllable, safe, and environmentally benign method of mining, which can operate under strict environmental controls and which often has cost advantages. The mine consists of well fields, which are progressively established over the orebody as uranium is depleted from sections of the orebody after leaching. Well-field design is on a grid with alternating extraction and injection wells, each of identical design and typical of normal water bores. The spacing between injection wells is about 30 m with each pattern of four having a central extraction well with a submersible electric pump. A series of monitor wells are situated around each mineralized zone to detect any movement of mining fluids outside the mining area. The wells are cased to ensure that liquors only flow to and from the ore zone and do not affect any overlying aquifers. They are pressure-tested before use.

The submersible pumps initially extract native groundwater from the host aquifer prior to the addition of uranium complexing reagents (acid or alkaline) and an oxidant (hydrogen peroxide or

oxygen) before injection into the well field. The leach liquors pass through the ore to oxidize and dissolve the uranium minerals *in situ*.

While uranium production in Australia uses acid leaching of the crushed ore, ISL elsewhere normally uses alkaline leaching agents such as a combination of sodium bicarbonate and carbon dioxide. At Beverley and Honeymoon in South Australia the process is acid leaching, with weak sulfuric acid plus oxygen. The leach solution is at a pH of 2.0 to 3.0, about the same as vinegar.

The pregnant solution from the production wells is pumped to the treatment plant where the uranium is recovered (see Section 1.4). Before the process solution depleted of uranium (i.e., barren liquor) is reinjected, it is oxygenated and if necessary re-charged with sulfuric acid, or with sodium bicarbonate or carbon dioxide, to maintain its pH.

Most of the solution is returned to the injection wells, but a very small flow (about 0.5%) is bled off to maintain a pressure gradient in the well field and this, with some solutions from surface processing, is treated as waste. It contains various dissolved minerals such as radium, arsenic, and iron from the orebody and is reinjected into approved disposal wells in a depleted portion of the orebody. This bleed of process solution ensures that there is a steady flow into the well field from the surrounding aquifer and serves to restrict the flow of mining solutions away from the mining area.

In the United States, the production life of an individual ISL well pattern is usually less than 3 years, typically 6 to 10 months. Most of the uranium is recovered during the first 6 months of the operation of those wells. The most successful operations have achieved a total overall recovery of about 80% of the uranium from the ore. Over time, production flows decrease as clay and silt become trapped in the permeable sediments. These can be dislodged to some extent by using higher pressure injection or by reversing the flow between injection and production wells. At established operations in the United States, after ISL mining is completed, the quality of the remaining groundwater must be restored to a baseline standard determined before the start of the operation so that any prior uses may be resumed.

In contrast to the main U.S. operations, the water quality at the Australian sites is very low to start with, and it is quite unusable. At Beverley the groundwater in the orebody is fairly saline and orders of magnitude too high in radionuclides for any permitted use. At Honeymoon, the water is even more saline and high in sulfates and radium. When oxygen input and leaching are discontinued, the water quality soon reverts to its original condition. With ISL, no tailings are involved and very little waste is generated. ISL thus has clear environmental advantages in the places it can be applied.

1.4 Milling and Extraction Process

Milling refers to the extraction of uranium product from the mined ore. The following is a simplified account. At conventional hard rock mines, the ore must first be crushed. At Olympic Dam and some other underground mines, the first stage is done underground. Primary crushing breaks up the largest pieces followed by secondary crushing to reduce the material to small pieces of about 20 mm and less. At some mines, after primary crushing the broken ore is fed through a radiometric ore sorter so that individual pieces of low-grade material are rejected. The broken ore is then ground in water to produce a slurry of fine ore particles suspended in the water. At McArthur River, the grinding is done underground. Because that is such a high-grade ore, it has to be diluted down to 4% U_3O_8 at the mill before being treated in order to limit radiation doses to staff at the mill.

The U_3O_8 slurry, effectively a mixture of UO_3 and UO_2, is leached with sulfuric acid to dissolve the uranium oxides, this being the same process as in acid ISL:

$$UO_3 + 2\,H^+ ==> UO_2^{++} + H_2O$$
$$UO_2^{++} + 3\,SO_4^{--} ==> UO_2(SO_4)_3^{4-}.$$

With an oxidant such as hydrogen peroxide or ferric iron present, the UO_2 is oxidized to $UO_2^{+\,+}$.

With some ores, and notably with ISL in the United States, carbonate leaching is used to form a soluble uranyl tricarbonate ion: $UO_2(CO_3)_3^{4-}$. This can then be precipitated with an alkali (e.g., as sodium or magnesium diuranate). Alkaline leaching or treatment of ores is not undertaken in Australia or Canada.

The liquid containing the uranium is filtered and the uranium then separated by an ion exchange process. Two methods are used for concentration and purification of uranium: solid ion exchange (IX) with resin and liquid ion exchange, better known as solvent extraction (SX). Early operations in Australia used ammonium type resins in polystyrene beads for ion exchange, but solvent extraction with amines in kerosene is now predominant.

With pregnant solution from the ISL production wells, the choice is largely determined by the salinity of the groundwater. SX is better with high salinity, as at Honeymoon (17,000 to 20,000 ppm dissolved solids), while IX is most effective below 3000 ppm chloride, as presently at Beverley. With alkaline leaching, IX is effective to about 3000 ppm dissolved solids, but there are no commercially available resins that work at high chloride levels.

In solvent extraction, tertiary amines ("R" is an alkyl [hydrocarbon] grouping, with single covalent bond) are used in a kerosene diluent, and the phases move countercurrently:

$$2R_3N + H_2SO_4 =====> (R_3NH)_2SO_4$$
$$2(R_3NH)_2SO_4 + UO_2(SO_4)_3^{4-} =====> (R_3NH)_4UO_2(SO_4)_3 + 2SO_4^{2-}.$$

In the mill at Olympic Dam (which is primarily a copper mine; uranium is the major by-product), the ore is ground and treated in a copper sulphide flotation plant. About 80% of the uranium minerals remain in the tailings from the flotation cells, from which they are recovered by acid leaching. The copper concentrate is also processed through an acid leach to recover the other 20% of the uranium. The pregnant liquor is then separated from the barren tailings, and in the solvent extraction plant the uranium is removed using kerosene with an amine. The loaded solvents may then be treated to remove impurities. First, cations are removed at pH 1.5 using sulfuric acid, then anions are dealt with using injected gaseous ammonia.

The solvents are then stripped in a countercurrent process using ammonium sulfate solution:

$$(R_3NH)_4UO_2(SO_4)_3 + 2(NH_4)_2SO_4 ====> 4R_3N + (NH_4)_4UO_2(SO_4)_3 + 2H_2SO_4.$$

Yellow ammonium diuranate is then precipitated from the loaded strip solution by adding gaseous ammonia to raise the pH and neutralize the solution (though in earlier operations caustic soda and magnesia were used).

$$2NH_3 + 2UO_2(SO_4)_3^{4-} ===> (NH_4)_2U_2O_7 + 4SO_4^{2-}.$$

The diuranate is then dewatered in a thickener and then by filter or centrifuge before being roasted in a furnace to produce a uranium oxide concentrate, about 99% U_3O_8, which is the form in which uranium is marketed and exported. This is sometimes known as yellowcake, though it is khaki.

With IX, such as at Beverley, uranium is stripped from the ion exchange resin and precipitated chemically, usually with hydrogen peroxide. The uranium slurry is dewatered and dried to give hydrated uranium peroxide ($UO_4.2H_2O$) product. Either product is then packed into 200-liter steel drums, which are sealed for shipment. The U_3O_8 is only mildly radioactive and requires no special packaging on that account.

1.5 Wastes from Mining and Milling

In most respects, conventional mining of uranium is the same as mining any other metalliferous ore, and well-established environmental constraints apply to avoid any offsite pollution. From open cut mining, there are substantial volumes of barren rock and overburden waste. These are placed near the pit and either used in rehabilitation or shaped and revegetated where they are. At Ranger mine, the development of the first orebody involved a waste to ore ratio of 2.2:1. However, uranium minerals are always associated with more radioactive elements such as radium and radon in the ore. Therefore, although uranium itself is not very radioactive, the ore that is mined, especially if it is high-grade, is handled with some care for occupational health and safety reasons.

Mining methods, tailings and run-off management, and land rehabilitation are subject to government regulation and inspection. Mining operations are undertaken under relevant national health and radiation protection codes of practice. These set strict health standards for exposure to gamma radiation and radon gas. Standards apply to both workers and members of the public.

1.5.1 Tailings and Radon

Solid waste products from the milling operation are tailings. They constitute most of the original ore and they contain most of the radioactivity in it. In particular, they contain all the radium present in the original ore. At an underground mine, they may be first cycloned to separate the coarse fraction, which is used for underground fill. The balance is pumped as a slurry to a tailings dam, which may be a worked-out pit as at Ranger and McClean Lake.

When radium undergoes natural radioactive decay, one of the products is radon gas. Because radon and its decay products (daughters) are radioactive and because the tailings are now on the surface, measures are taken to minimize the emission of radon gas. During the operational life of a mine,

the material in the tailings dam is usually covered by water to reduce surface radioactivity and radon emission (though with lower-grade ores neither pose a hazard at these levels).

On completion of the mining operation, it is normal for the tailings dam to be covered with some 2 m of clay and topsoil to reduce radiation levels to near those normally experienced in the region of the orebody, and for a vegetation cover to be established. At Ranger and Jabiluka in North Australia, tailings will be returned underground, as was done at the now-rehabilitated Nabarlek mine. In Canada, ore treatment is often remote from the mine that the new ore comes from, and tailings are emplaced in mined out pits wherever possible and engineered dams otherwise.

The radon gas emanates from the rock and tailings as the radium or thorium decays. It then decays itself to (solid) radon daughters, which are significantly alpha radioactive. About 95% of the radioactivity in the ore is from the U-238 decay series, totaling about 450 kBq/kg in ore with 0.3% U_30_8 (e.g., from Ranger). The U-238 series has 14 radioactive isotopes in secular equilibrium, thus each represents about 32 kBq/kg (irrespective of the mass proportion). When the ore is processed, the U-238 and the very much smaller masses of U-234 and U-235 are removed. The balance becomes tailings and at this point has about 85% of its original intrinsic radioactivity. However, with the removal of most U-238, the following two short-lived decay products (Th-234 and Pa-234) soon disappear, leaving the tailings with a little over 70% of the radioactivity of the original ore after several months. The controlling long lived isotope then becomes Th-230, which decays with a half life of 77,000 years to radium-226 followed by radon-222. Radon occurs in most rocks, and traces of it are in the air we all breathe. However, at high concentrations it is a health hazard.

1.5.2 Water

Run-off from the mine stockpiles and waste liquors from the milling operation are collected in secure retention ponds for isolation and recovery of any heavy metals or other contaminants. The liquid portion is disposed of either by natural evaporation or recirculation to the milling operation. Most Australian mines adopt a zero-discharge policy for any pollutants.

Process water discharged from the mill contains traces of radium and some other metals that would be undesirable in biological systems downstream. This water is evaporated and the contained metals are retained in secure storage. During the operational phase, such water may be used to cover the tailings while they are accumulating.

With *in situ* leach (ISL) operations, the orebody stays in the ground and uranium is recovered by circulating oxygenated and acidified groundwater through it, using injection and recovery wells. The saline quality of this groundwater in Australian ISL mines makes it far from potable in the first place, and after the uranium is recovered, oxygen input and circulation are discontinued, leaving the groundwater much as it was. The main environmental consideration with ISL is avoiding pollution of any groundwater away from the orebody and leaving the immediate groundwater no less useful than it was initially.

1.5.3 Rehabilitation

At the conclusion of mining, tailings are covered permanently with enough clay and soil to reduce both gamma radiation levels and radon emanation rates to levels near those naturally occurring in the region, and enough rock to resist erosion. A vegetation cover is then established.

Mary Kathleen in Queensland was the site of Australia's first major rehabilitation project of a uranium mine. It involved the plant site, a 28-hectare tailings dam, and a 60-ha evaporation pond area. All this has now returned to being a cattle station, with unrestricted access. The rehabilitation project was completed at the end of 1985 at a cost of some $19 million and won an award for engineering excellence.

The Nabarlek uranium mine in the Northern Territory, c 270 km east of Darwin, was the first of the new generation of uranium mines to commence operations and the first to be rehabilitated. Environmental protection was stressed at Nabarlek since before mining commenced, and everything proceeded with eventual rehabilitation very much in mind. During the life of the mine, the company worked together with government agencies, the Northern Land Council (NLC), and Aboriginal landowners to ensure a high standard of environmental management, culminating in its decommissioning and successful rehabilitation.

Apart from groundwater considerations, rehabilitation of ISL mines is straightforward, making this a technique with remarkably low environmental impact. Upon decommissioning, wells are sealed or capped, process facilities removed, any evaporation pond revegetated, and the land can readily be returned to its previous uses. Experience at many mine sites is networked throughout the industry and available to present and future operators.

1.6 Health of Workers

In Australia, all uranium mining and milling operations are undertaken under the *Code of Practice on Radiation Protection in the Mining and Milling of Radioactive Ores*. This was drawn up by the Commonwealth in line with recommendations of the International Commission on Radiological Protection (ICRP), but it is administered by state health and mines departments. This Health Code, which was updated in 1995 and again in 2002–2003 (as part of a combined Health-Waste Code), sets strict health standards for radiation and radon gas exposure for both workers and members of the public.

In Canada, the Canadian Nuclear Safety Commission is responsible for regulating uranium mining as well as other aspects of the nuclear fuel cycle. In Saskatchewan, provincial regulations also apply concurrently and set strict health standards for both miners and local people. Similar standards are set in other countries.

While uranium itself is only slightly radioactive, radon, a radioactive inert gas, is released to the atmosphere in very small quantities when the ore is mined and crushed. Radon is one of the decay products of uranium and radium, and occurs naturally in most rocks—minute traces of it are present in the air which we all breathe.

Australian uranium mines have mostly been open cut and therefore naturally well ventilated. The Olympic Dam and Canadian underground mines are ventilated with powerful fans. Radon levels are kept at a low and certainly safe level in uranium mines. (Radon in non-uranium mines also may need control by ventilation.)

Gamma radiation may also be a hazard to those working close to high-grade ores. It comes principally from radium in the ore, so exposure to this is regulated as required. In particular, dust is suppressed, since this represents the main potential exposure to alpha radiation as well as a gamma radiation hazard.

At the concentrations associated with uranium (and some mineral sands) mining, radon is a potential health hazard, as is dust. Precautions taken during the mining and milling of uranium ores to protect the health of the workers include the following:

- Good forced ventilation systems in underground mines to ensure that exposure to radon gas and its radioactive daughter products is as low as possible and does not exceed established safety levels
- Efficient dust control, because the dust may contain radioactive constituents and emit radon gas
- Limiting the radiation exposure of workers in mine, mill, and tailings areas so that it is as low as possible and in any event does not exceed the allowable dose limits set by the authorities; in Canada this means that mining in very high-grade ore is undertaken solely by remote control techniques and by fully containing the high-grade ore where practicable
- The use of radiation detection equipment in all mines and plants
- Imposition of strict personal hygiene standards for workers handling uranium oxide concentrate

At any mine, designated employees (those likely to be exposed to radiation or radioactive materials) are monitored for alpha radiation contamination and personal dosimeters are worn to measure exposure to gamma radiation. Routine monitoring of air, dust, and surface contamination is undertaken. Canadian mine and mill facilities are designed to handle safely ore grades of up to 26% U.

If uranium oxide is ingested, it has a chemical toxicity similar to that of lead oxide. Similar hygiene precautions to those in a lead smelter are therefore taken when handling it in the drying and packing areas of the mill. The usual radiation safeguards are applied at an ISL mining operation, despite the fact that most of the orebody's radioactivity remains well underground and there is hence minimal increase in radon release and no ore dust.

2. URANIUM REFINING AND CONVERSION

Most nuclear reactors require uranium to be enriched from its natural isotopic composition of 0.7% U-235 (most of the rest being U-238) to 3.5 to 4% U-235. The uranium therefore needs to be in a gaseous form, and the most convenient way of achieving this is to convert the uranium oxides to uranium hexafluoride, commonly referred to as hex. Conversion plants are operating commercially in United States, Canada, France, the United Kingdom, and Russia (see Table II).

Uranium leaves the mine as the concentrate of a stable oxide known as U_3O_8 or as a peroxide. It still contains some impurities, and prior to enrichment it has to be further refined before being converted to uranium hexafluoride (UF_6). The uranium oxide

TABLE II
World Uranium Conversion Capacity in 2001

	Tonnes U/yr
Canada	10,950
France	14,000
UK	7200
USA	12,700
Russia	18,700

Source. OECD NEA (2002). Nuclear energy data. Nuclear Engineering International (2002). "World Nuclear Handbook."

concentrate received by the refinery is dissolved in nitric acid. The resulting solution of uranyl nitrate $UO_2(NO_3)_2.6H_2O$ is fed into a countercurrent solvent extraction process, using tributyl phosphate dissolved in kerosene or dodecane. The uranium is collected by the organic extractant, from which it can be washed out by dilute nitric acid solution and then concentrated by evaporation. The solution is then calcined (heated strongly) in a fluidized bed reactor to produce UO_3.

The purified uranium oxide UO_3 is then reduced in a kiln by hydrogen to UO_2:

$$UO_3 + H_2 ===> UO_2 + H_2O$$
$$\Delta H = -109\,\text{kJ/mole}.$$

This reduced oxide is then reacted in another kiln with gaseous hydrogen fluoride (HF) to form uranium tetrafluoride (UF_4), though in some places this is made with aqueous HF by a wet process:

$$UO_2 + 4HF ===> UF_4 + 2H_2O$$
$$\Delta H = -176\,\text{kJ/mole}.$$

The tetrafluoride is then fed into a fluidized bed reactor with gaseous fluorine to produce uranium hexafluoride, UF_6. Hexafluoride ("hex") is condensed and stored:

$$UF_4 + F_2 ===> UF_6.$$

Removal of impurities takes place at several of these steps.

The UF_6, particularly if moist, is highly corrosive. When warm it is a gas, suitable for use in the enrichment process. At lower temperature and under moderate pressure, the UF_6 may be liquefied and the liquid run into specially designed steel shipping cylinders, which are thick walled and weigh up to 15 tonnes when full. As it cools, the liquid UF_6 within the cylinder becomes a white crystalline solid and is shipped in this form. The siting and environmental management of a conversion plant requires no special arrangements beyond those for any chemical processing plant involving fluorine chemicals.

3. URANIUM ENRICHMENT

3.1 General

Uranium found in nature consists largely of two isotopes, U-235 and U-238. The production of energy in nuclear reactors is from the fission or splitting of the U-235 atoms, a process that releases energy in the form of heat. U-235 is the main fissile isotope of uranium.

Natural uranium contains 0.7% of the U-235 isotope. The remaining 99.3% is mostly the U-238 isotope, which does not contribute directly to the fission process (though it does so indirectly by the formation of fissile isotopes of plutonium).

Uranium-235 and U-238 are chemically identical but differ in their physical properties, particularly their mass. The nucleus of the U-235 atom contains 92 protons and 143 neutrons, giving an atomic mass of 235 units. The U-238 nucleus also has 92 protons but has 146 neutrons, three more than U-235, and therefore has a mass of 238 units. The difference in mass between U-235 and U-238 allows the isotopes to be separated and makes it possible to increase or enrich the percentage of U-235. All enrichment processes, directly or indirectly, make use of this small mass difference.

Some reactors—for example, the Canadian-designed Candu and the British Magnox reactors—use natural uranium as their fuel. Most present day reactors (Light Water Reactors or LWRs) use enriched uranium, where the proportion of the U-235 isotope has been increased from 0.7% to about 3 or up to 4%. (For comparison, uranium used for nuclear weapons would have to be enriched in plants specially designed to produce at least 90% U-235.)

A number of enrichment processes have been demonstrated in the laboratory and some on a larger scale, but only two, the gaseous diffusion process and the centrifuge process, are operating on a commercial scale. In both of these, UF_6 gas is used as the feed material. Molecules of UF_6 with U-235 atoms are about 1% lighter than the rest, and this difference in mass is the basis of both processes. Large commercial enrichment plants are in operation in the United States, the United Kingdom, France, Germany, the Netherlands, and Russia, with smaller plants elsewhere (see Table III).

TABLE III

World Uranium Enrichment Capacity in 2001

	Method	× 1000 kg SWU/year
France	Diffusion	10,800
Germany, Netherlands, United Kingdom	Centrifuge	5250
Japan	Centrifuge	1050
United States	Diffusion	18,700
Russia	Centrifuge	19,000
China	Mostly centrifuge	400–800
Pakistan	Centrifuge	5
Total		55,000 approx.

Source. OECD NEA (2002). Nuclear energy data. Nuclear Engineering International (2002). "World Nuclear Handbook."

The capacity of enrichment plants is measured in terms of separative work units, or SWU. The SWU is a complex unit that is a function of the amount of uranium processed and the degree to which it is enriched (i.e., the extent of increase in the concentration of the U-235 isotope relative to the remainder). The unit is strictly a Kilogram Separative Work Unit, and it measures the quantity of separative work performed to enrich a given amount of uranium a certain amount. It is thus indicative of energy used in enrichment when feed and product quantities are expressed in kilograms. The unit tonnes SWU is also used. For instance, to produce one kilogram of uranium enriched to 3% U-235 requires 3.8 SWU if the plant is operated at a tails assay 0.25%, or 5.0 SWU if the tails assay is 0.15% (thereby requiring only 5.1 kg instead of 6.0 kg of natural U feed).

About 100 to 120,000 SWU is required to enrich the annual fuel loading for a typical 1000 MWe light water reactor. Enrichment costs are related to electrical energy used. A gaseous diffusion plant typically demands 2500 kilowatt hours (9000 MJ) per SWU (kWh/SWU), while modern gas centrifuge plants require only about 50 kWh (180 MJ) per kg SWU. Enrichment accounts for almost half of the cost of nuclear fuel and about 5% of the total cost of the electricity generated. It can also account for the main greenhouse gas impact from the nuclear fuel cycle if the electricity used for enrichment is generated from coal. However, it still typically amounts to only 0.1% of the carbon dioxide from equivalent coal-fired electricity generation if gas centrifuge plants are used, or up to 3% in a worst case situation.

3.2 Early Processes

The electromagnetic isotope separation (EMIS) process was developed in the early 1940s in the Manhattan Project to make the highly enriched uranium used in the Hiroshima bomb, but was abandoned soon afterward. However, it reappeared as the main thrust of Iraq's clandestine uranium enrichment program for weapons discovered in 1992. EMIS uses the same principles as a mass spectrometer (albeit on a much larger scale). Ions of uranium-238 and uranium-235 are separated because they describe arcs of different radii when they move through a magnetic field.

Two aerodynamic processes were brought to demonstration stage. One is the jet nozzle process, with demonstration plant built in Brazil, and the other the Helikon vortex tube process developed in South Africa. Neither is in use now. They depend on a high-speed gas stream bearing the UF_6 being made to turn through a very small radius, causing a pressure gradient similar to that in a centrifuge. The light fraction can be extracted toward the center and the heavy fraction on the outside. Thousands of stages are required to produce enriched product for a reactor. Both processes are energy-intensive—more than 3000 kWh/SWU.

3.3 Gaseous Diffusion Process

Commercial uranium enrichment was first carried out by the diffusion process in the United States. It has since been used in Russia, the United Kingdom, France, China and Argentina as well. Only the United States and France use the process on any significant scale. The remaining large USEC plants in the United States was originally developed for weapons programs and have a capacity of some 19 million kg SWU per year. At Tricastin, in southern France, a more modern diffusion plant with a capacity of 10.8 million kg SWU per year has been operating since 1979. This plant can produce enough 3.7% enriched uranium a year to fuel some 90 1000-MWe nuclear reactors.

The gaseous diffusion process accounts for over 57% of world enrichment capacity. However, though they have proved durable and reliable, most gaseous diffusion plants are now nearing the end of their design life and the focus is on which enrichment technologies will replace them.

The process involves forcing uranium hexafluoride gas under pressure through a series of porous membranes or diaphragms. As U-235 molecules are

lighter than the U-238 molecules they move faster and have a slightly better chance of passing through the pores in the membrane. The UF_6 that diffuses through the membrane is thus slightly enriched, while the gas which did not pass through is depleted in U-235.

This process is repeated many times in a series of diffusion stages called a cascade. Each stage consists of a compressor, a diffuser, and a heat exchanger to remove the heat of compression. The enriched UF_6 product is withdrawn from one end of the cascade and the depleted UF_6 is removed at the other end. The gas must be processed through some 1400 stages to obtain a product with a concentration of 3% to 4% U-235.

3.4 Centrifuge Process

The gas centrifuge process was first demonstrated in the 1940s but was shelved in favor of the simpler diffusion process. It was then developed and brought on stream in the 1970s as the second-generation enrichment technology. It is economic on a smaller scale (e.g., less than 2 million kg SWU/year, which enables staged development of larger plants). It has been deployed at a commercial level in Russia and in Europe by Urenco, an industrial group formed by British, German and Dutch companies. Russia's three plants at Seversk, Zelenogorsk and Novouralsk account for some 30% of world capacity. Urenco operates enrichment plants in the United Kingdom, the Netherlands, and Germany and is participating in a proposal for one in the United States.

In Japan, JNC and JNFL operate small centrifuge plants, the capacity of JNFL's at Rokkasho planned to be 1.5 million SWU/year. China also has a small centrifuge plant imported from Russia at Lanzhou, which is expected to reach 0.5 million SWU/year about 2005. Another small plant at Hanzhong is operating and was to reach 0.5 million SWU/year by 2003. Brazil has a small plant that is being developed to 0.2 million SWU/year. Pakistan has developed centrifuge enrichment technology based on Urenco designs, and this appears to have been sold to North Korea. Iran has sophisticated centrifuge technology, which had not been commissioned as of early 2003.

Both France and the United States are now considering centrifuge technology to replace their aging diffusion plants, not least because they are more economical to operate. As noted, a centrifuge plant requires as little as 50 kWh/SWU power (Urenco at Capenhurst, United Kingdom, input 62.3 kWh/SWU for the whole plant in 2001–2002, including infrastructure and capital works).

Like the diffusion process, the centrifuge process uses UF_6 gas as its feed and makes use of the slight difference in mass between U-235 and U-238. The gas is fed into a series of vacuum tubes, each containing a rotor about 1 to 2 m long and 15 to 20 cm diameter. When the rotors are spun rapidly, at 50,000 to 70,000 rpm, the heavier molecules with U-238 increase in concentration toward the cylinder's outer edge. There is a corresponding increase in concentration of U-235 molecules near the center. These concentration changes are enhanced by inducing the gas to circulate axially within the cylinder. The enriched gas is drawn off and goes forward to further stages while the depleted UF_6 goes back to the previous stage.

To obtain efficient separation of the two isotopes, centrifuges rotate at very high speeds, with the outer wall of the spinning cylinder moving at between 400 and 500 m per second to give a million times the acceleration of gravity. Although the capacity of a single centrifuge is much smaller than that of a single diffusion stage, its capability to separate isotopes is much greater. Centrifuge stages normally consist of a large number of centrifuges in parallel. Such stages are then arranged in cascade similarly to those for diffusion. In the centrifuge process, however, the number of stages may only be 10 to 20 instead of a thousand or more for diffusion.

3.5 Laser Processes

Laser enrichment processes have been the focus of interest for some time. They are a possible third-generation technology promising lower energy inputs, lower capital costs, and lower tails assays, hence significant economic advantages. None of these processes is yet ready for commercial use, though one is well advanced.

Development of the Atomic Vapour Laser Isotope Separation (AVLIS and the French SILVA) began in the 1970s. In 1985, the U.S. government backed it as the new technology to replace its gaseous diffusion plants as they reached the end of their economic lives early in the 21st century. However, after some US$2 billion in R&D, it was abandoned in United States in favor of SILEX, a molecular process. French work on SILVA has now ceased.

Atomic vapor processes work on the principle of photo-ionization, whereby a powerful laser is used to ionize particular atoms present in a vapor of uranium metal. (An electron can be ejected from an atom by light of a certain frequency. The laser techniques for uranium use a frequency that is tuned to ionize a

U-235 atom but not a U-238 atom.) The positively charged U-235 ions are then attracted to a negatively charged plate and collected. Atomic laser techniques may also separate plutonium isotopes.

The main molecular processes that have been researched work on a principle of photo-dissociation of UF_6 to solid UF_5, using tuned laser radiation as noted earlier. Any process using UF_6 fits more readily within the conventional fuel cycle than the atomic process. The leading molecular laser process on the world stage is SILEX, an Australian development that also utilizes UF_6. In 1996, USEC secured the rights to evaluate and develop SILEX for uranium (it is also usable for silicon and other elements). The SILEX process is now at prototype stage in Sydney and applications to silicon and zirconium are being developed.

3.6 Enrichment of Reprocessed Uranium

In some countries, spent fuel is reprocessed to recover its uranium and plutonium and to reduce the final volume of high-level wastes. The plutonium is normally recycled promptly into mixed-oxide (MOX) fuel by mixing it with depleted uranium.

Where uranium recovered from reprocessing spent nuclear fuel is to be reused, it needs to be converted and re-enriched. This is complicated by the presence of impurities and two new isotopes in particular: U-232 and U-236, which are formed by neutron capture in the reactor. Both decay much more rapidly than U-235 and U-238, and one of the daughter products on U-232 emits very strong gamma radiation, which means that shielding is necessary in the plant. U-236 is a neutron poison, which impedes the chain reaction and means that a higher level of U-235 enrichment is required in the product to compensate. Both isotopes tend to report with the enriched (rather than depleted) output, so reprocessed uranium that is re-enriched for fuel must be segregated from enriched fresh uranium.

Both diffusion and centrifuge processes are used for re-enrichment, though contamination issues are more readily managed with the latter. A laser process would theoretically be ideal as it would ignore all but the desired U-235, but this remains to be demonstrated with reprocessed feed.

3.7 Environmental Aspects

With the minor exception of reprocessed uranium, enrichment involves only natural, long-lived radioactive materials; there is no formation of fission products or irradiation of materials, as in a reactor. Feed, product, and depleted material are all in the form of UF_6, though the depleted uranium may be stored long-term as the more stable U_3O_8.

Uranium is only weakly radioactive, and its chemical toxicity—especially as UF_6—is more significant than its radiological toxicity. The protective measures required for an enrichment plant are therefore similar to those taken by other chemical industries concerned with the production of fluorinated chemicals.

Uranium hexafluoride is a corrosive material, therefore any leakage is particularly undesirable. Hence,

- In most areas of a centrifuge plant the pressure of the UF_6 gas is maintained below atmospheric pressure and thus any leakage could only result in an inward flow.
- Double containment is provided for those few areas where higher pressures are required.
- Effluent and venting gases are collected and returned to the process.

3.8 After Enrichment

The enriched UF_6 is converted back to UO_2 and made into fuel pellets—ultimately a sintered ceramic—which are encased in metal tubes to form fuel rods, typically up to 4 m long. A number of fuel rods make up a fuel assembly, which is ready to be loaded into the nuclear reactor.

SEE ALSO THE FOLLOWING ARTICLES

Coal Mining, Design and Methods of • Nuclear Engineering • Nuclear Fission Reactors: Boiling Water and Pressurized Water Reactors • Radiation, Risks and Health Impacts of • Uranium and Thorium Resource Assessment • Uranium Mining: Environmental Impact

Further Reading

Hore-Lacy, I. (2003). "Nuclear Electricity," 7th Ed. World Nuclear Association, London and Uranium Information Centre, Melbourne. [Available at http:/ www.world-nuclear.org.]

Heriot, I. D. (1988). "Uranium Enrichment by Centrifuge," Report EUR 11486, Commission of the European Communities, Brussels.

Kehoe, R. B. (2002). "The Enriching Troika, a History of Urenco to the Year 2000." Urenco, Marlow UK.

Pollock, R., and Rowson, J. (2002). COGEMA Resources' McClean Lake and other mines. *The Nuclear Engineer* **43**(1), 9–13.

Quick, M. (2001). New uranium mines in Canada: McArthur River. *The Nuclear Engineer* **42**(6), 179–183.

Wilson, P. D. (ed.). (1996). "The Nuclear Fuel Cycle—From Ore to Wastes." Oxford University Press, Oxford, UK.

Woodcock, J. T., and Hamilton, J. K. (1993). "Australasian Mining and Metallurgy, the Sir Maurice Mawby Memorial Volume." Four papers, pp. 1160–1170. Aust IMM, Melbourne.

Urbanization and Energy

DONALD W. JONES
RCF Economic and Financial Consulting, Inc.
Chicago, Illinois, United States

1. **Urbanization, Development, Industrialization, and Energy Use**
2. **Effects of Urban Population Concentration on Energy Use**
3. **Effects of Increases in Individual Income on Energy Demands During Development**
4. **Aggregate Energy Implications of Urbanization**

Glossary

elasticity A dimensionless measure of the responsiveness of a dependent variable to a causal variable, constructed as the percentage change in the responding variable to a 1% change in the causal variable; the range of magnitude of an elasticity is from $-\infty$ to $+\infty$.

embedded (embodied) energy Energy used to produce a good; for example, the energy embedded in an automobile is that required to mine the minerals and fuels used in its production.

modern energy Fossil fuels and electricity.

traditional energy Firewood, charcoal, animal dung, human and animal traction, and stationary (nonelectric) hydropower.

Urbanization is the process of transferring a country's population and labor force from rural areas and predominantly agricultural pursuits to cities and industrial and related service occupations. A country's urbanization is largely coincidental with its industrialization and is a major feature in its economic development. Economic development is a society's transformation from predominantly local orientations, based largely on animate- or renewables-powered production and using technologies that have changed little if any over several centuries, to a more nationally and internationally integrated, larger scale society that uses more powerful and specialized technologies, usually powered by more flexible energy sources. Cities collect the concentrated labor forces required to undertake more specialized industrial production for larger markets and to facilitate the coordination between city and country, as well as among various locations in the countryside, required to support an efficient agricultural sector.

1. URBANIZATION, DEVELOPMENT, INDUSTRIALIZATION, AND ENERGY USE

1.1 Development as Urbanization

The 18th, 19th, and 20th centuries witnessed the urbanization of Europe and North America, with its Australian and New Zealand outposts, and Japan joined the group two-thirds of the way through the 20th century. These countries transferred approximately two-thirds of their labor forces from rural agricultural settings to cities over 100 to 150 years, with the gross national product (GNP) shares of agriculture and nonagriculture (industry and services) changing similarly. The increases in GNP shares coming from industry, particularly metal-producing and -using manufacturing industries, both induced and were facilitated by substitutions of coal for wood and charcoal as fuels. For financial feasibility, larger scale manufacturing required more dependable fuel supplies and found greater flexibility in delivery of coal, and later oil, than could be achieved with either of the wood renewables or with stationary hydropower. These other energy sources also proved to be capable of providing power with different qualities, such as controllability, and their energy densities permitted delivery of larger quantities of effective power.

Urbanization in the Third World outstripped late 20th-century projections, rising from about 27% urban in 1975 to nearly 40% in 2000. The 1999 revision of the United Nations' *World Urbanization*

 © 2004 Elsevier Inc. All rights reserved.

Prospects projected that the world's less developed regions' urbanization will rise to 56% by 2030 while that of the more developed regions will reach 83%. The populations of both China and India, the world's two most populous countries, are currently less than 25% urbanized.

1.2 Changes in Agriculture

A large part of the urban population comes from agriculture during earlier phases of urbanization, so the increments to urban labor forces are matched *pari passu* by decreases in agricultural labor forces. Consequently, farmers who largely fed only themselves and their own families must devise ways in which to feed a larger number of people. Most (but certainly not all) countries cannot shift increases in economy-wide food consumption per farm producer entirely to the foreign trade sector. Agricultural production must change to permit a smaller farming population to feed a larger nonfarming population. The requisite productivity increases come primarily through technological and organizational change: use of new seed varieties and chemical fertilizers, mechanization, irrigation (particularly pumped irrigation), and improvements in infrastructure such as rural credit, agricultural extension, and marketing. The use of machinery reduces labor shortage problems at peak demand times such as harvesting.

Traditional agriculture uses virtually no fuels; its power sources are human and animal labor. As with traditional fuels in cities, food production considered as a fuel source is a more spatially diffuse activity. Land requirements of fodder production have represented major shares of cropland in pre-fossil fuel societies. Mechanization increases the ability of a farm labor force to feed more consumers per worker, but it requires fossil fuels. India devoted nearly 17% of its oil imports to agricultural tractors alone during the early 1970s, and its agriculture was not highly mechanized. Replacement of draft animals releases land used for fodder for use in food production, equivalent to roughly a 20% yield increase in South Asia. Introduction of modern heating, lighting, and cooking fuels concomitant with mechanization releases more animal dung for use as fertilizer instead of fuel; however, eventual decline in the number of draft animals requires chemical fertilizers to substitute for dung, and chemical fertilizers require petroleum feedstocks in their production. Chemical fertilizer use may require pumped irrigation, which also requires fossil fuels. In India during the late 1970s, fertilizer production was the third most energy-intensive manufacturing activity, behind cement and iron and steel. As more multiple cropping is undertaken, coordination of the timely arrival of agricultural inputs requires more attention, and non-farm-produced supplies are more likely to be delivered by fueled vehicles. The same holds for transport of harvested outputs.

Evidence from post-World War II Taiwan shows a shift from 67% human-supplied and 33% animal-supplied energy in agriculture to roughly 35% human-supplied, 5% animal-supplied, and 60% mechanically supplied energy within a quarter-century. Evidence from India suggests that shifting workers from traditional agriculture to the least energy-intensive manufacturing (textiles) can quadruple the per worker energy requirements of the transferred workers.

1.3 Industrial Composition Changes and Energy Use

As a country urbanizes and develops, the composition of its employment and production systematically changes from predominantly agriculture to a much larger prominence of industry and services. Industry, even traditional industry, uses considerably more energy per worker and per unit of output than does agriculture, and the mechanization of energy, with its attendant substitution of modern energy for traditional, exacerbates this effect. The changes in industrial structure accompanying urbanization and economic development increase fuel consumption, particularly consumption of fossil fuels, per worker and per unit of output. Not all industrialization occurs within urban areas, and the energy use increases from industrialization can be separated from those increases attributable to urbanization, although the two developments are related.

Within manufacturing, the share of metals increases, partly due to increases in the array of capital goods and consumer durables produced and partly due to substitution of metal parts for wooden and leather ones in existing products. Similar substitution occurs with plastics. Both metals and plastics are more energy intensive than wood and leather parts. Plastics production ordinarily uses fossil fuels, although some metal products, particularly steel, may get by with charcoal or combustible scrap materials, at least during the early stages of industrialization.

Development involves the acquisition of market discipline among both workers and firms. Unreliable producers either improve their performances or go out of business. Fossil fuels offer greater reliability in

heating performance and often are required to produce products of required quality standards.

1.4 Characteristics of the Empirical Evidence

The structural outlines of the changes in energy use involved in urbanization are clear, but much of the empirical evidence is scattered to the point of being anecdotal. National-level data on energy use are compiled by the United Nations and the Organization for Economic Cooperation and Development (OECD), and these can be supplemented by comparably national economic and demographic data from which aggregate relationships can be estimated. Beyond cross-national studies of long-term developmental and structural trends, a number of country-specific studies, and even city-specific studies, have focused on particular issues such as energy embodied in food delivery to cities, the composition of urban and rural residential energy use, and urban transportation choices. In general, these studies are time-specific snapshots, although an occasional study tracks changes over time. Similar topics are sometimes studied in different cities but cover different time periods, and these may imply different developmental phases in the history of the societies. Some legitimate generalizations may emerge from these studies, although with circumspection and qualification.

2. EFFECTS OF URBAN POPULATION CONCENTRATION ON ENERGY USE

2.1 Scale of Production

Higher concentrations of populations in cities allow larger labor forces to be assembled for employment in larger scale, more specialized facilities. Assembling these larger labor forces, despite their greater residential density, requires reaching labor markets over greater distances, and selling the outputs of these facilities requires selling over larger markets. Although the larger scale of production often lowers unit production costs, the transportation component of total costs rises to sustain the scale increase. Data on Turkish cement production show that expansion of a market area from 6 to 190 km reduces capital and labor costs by 60% and delivered output price by 47% while increasing transportation costs 16 times.

The large scale of production drives up urban land prices and induces urban producers to conserve on their use of land for low-value uses, such as storing low-energy traditional fuels (e.g., firewood), leading in turn to substitution of higher energy per volume fossil fuels. This fuel substitution also permits production at greater distances from the fuel source, and the transportation of the fossil fuel is likely to encourage further substitution of fossil fuel for animate traction.

2.2 Food Delivery

In traditional agriculture in Third World countries, farmers consume large proportions of what they produce and also produce most of what they consume, particularly in terms of food. The separation of food consumers from food producers is a major consequence of urbanization. In low-income countries, food processing, particularly for nonexported products, remains relatively simple. Urbanization permits larger scale and more efficient food processing, but urban food processing is more likely to use fuels instead of human or animal power sources. During the mid-1960s, the United States devoted more than 7% of its energy use to food processing and another 2% to transporting food. During the mid-1970s, with considerably lower levels of urbanization, India and Pakistan devoted 2.3 and 3.6% of their national energy consumptions to food processing, respectively, and each devoted another 1.3% to food transportation. Similar percentages of processing and transportation energy were estimated for food delivered to Mexico City during the mid-1970s.

2.3 Infrastructure

The large concentrations of people in cities require the construction and operation of both public and private infrastructures. The assembly phase of construction is not particularly energy intensive, but construction materials, particularly cement and structural steel, are quite intensive. Data from India during the late 1970s indicate that construction required 2087 kcal per rupee of output, compared with 15,344 for cement and 8757 for iron and steel. These compared with 4552 for nonferrous metals, 2708 for chemicals, 2378 for electrical goods, 2004 for transport vehicles and equipment, and 625 for food and agriculture. Data from The Philippines from 1981 show a similar pattern of relative energy intensities: 5724 kcal per peso for cement, 1431 for basic metals, 399 for machinery and equipment, and 95 for agriculture and livestock. Clearly, the building

of a city—streets, bridges, private and public nonresidential buildings, sanitation systems, and power generation facilities—requires considerable expenditure of energy. Traditionally constructed residential housing in Third World cities is not particularly energy intensive, although the use of cement will involve embedded energy.

However, once the infrastructure is built, it still must be operated and maintained. The provision of urban services, such as water treatment and supply, municipal lighting, and garbage collection, can consume approximately 5 to 6% of the total private fuel use of residents in cities, judging from evidence collected in Mexico City, Nairobi, and India from the mid-1970s. These services are not urban luxuries but rather important inputs to maintaining the productivity of a high-density urban labor force. The diseases accompanying polluted water supplies and rotting garbage are well known from the histories of Western cities, and the resources devoted to these services are repaid in higher productivity.

2.4 Changes in Domestic Activity

In cities, energy use changes in the home as well as in the market. Households use 40 to 90% of all energy in Third World countries, so small percentage changes in domestic energy use can be important in the aggregate. Urban households purchase many goods and services that rural households provide domestically, with an increase in the energy used in their production. Increased specialization and division of labor accompany urbanization, and many production activities performed at home in rural areas are transferred to the market in cities. Some of these activities require fuels in the home, typically traditional fuels, but can be performed with lower fuel inputs per unit of output by specialized market providers, albeit often using modern energy forms. Such activities include baking, food preparation, and water heating for clothes washing. Other activities, such as weaving and sewing, generally use only human energy when conducted at home but use fuel when conducted in the market.

Probably the most obvious energy-using change in individual activity occurs in personal transportation. Urban life requires transportation to and from work, often over greater distances than those separating farmers from fields in agriculture. Although farmers may spend a quarter to a third of their total labor time traveling to and from fields, they almost always walk or use an animate power source. In cities, use of fuel-consuming transportation is much more the rule, even among lower income workers. Urban transportation has a large, modern, fuel-using component that is absent in rural settings, that is, automobiles and buses. Despite widespread use of bicycles and feet (walking) among low-income groups, transportation can take up to 25 to 60% of total household energy use in Third World cities. There is not an equivalent fuel use in agriculture.

2.5 Urban Density and Fuel Use Patterns

Whereas rural dwellers will send women and children into forests to collect wood, urban dwellers in many Third World cities are served by extensive markets in fuelwood transported by fossil-fueled trucks. Nonetheless, the higher residential densities of cities place a premium on the compactness and portability of domestic fuels and encourage the substitution of more compact modern energy forms for traditional fuels. Statistical analysis of Hong Kong census data from the early 1970s indicate that population density depressed the share of firewood and charcoal in household fuel consumption, with an elasticity of -0.25. (An elasticity is the percentage change in one variable, in this case the firewood and charcoal share of domestic energy use, caused by a 1% change in another variable, in this case population density.) Population density in Hong Kong is also associated with a larger share of domestic energy derived from kerosene and with a larger share of energy devoted to commercial uses as opposed to domestic uses.

Higher urban residential densities also reduce per capita transmission and distribution costs of electricity, reducing supply costs for electricity in cities by reducing hook-up costs. Data on the magnitude of this effect, holding the effects of income constant, are not available.

3. EFFECTS OF INCREASES IN INDIVIDUAL INCOME ON ENERGY DEMANDS DURING DEVELOPMENT

Real income increases with economic development, and in developing countries substantial differences exist between rural and urban real incomes. Movement to a city can be expected to raise the income of the average migrating individual or family immediately or within an acceptable period of time. These real income changes propel a number of consumption changes affecting energy.

Income elasticities of household demand for energy have been estimated to be greater than 0.5, and sometimes greater than 1.0, for a number of developing countries in East and South Asia, Africa, and Latin America. Income elasticities are higher for modern energy forms than for traditional forms. The magnitude of these demand elasticities will ensure increased consumption of energy when rural residents migrate to cities, with a shift to modern energy sources. The Hong Kong data show a strong effect of higher income on the traditional fuel share, with an elasticity of −1.17. That is, a 1% increase in income will lower the share of traditional fuels by 1.17%.

Real income increases also encourage the purchase of energy-using appliances. The energy use by these purchases is included in the energy consumption on which the income elasticities just noted are estimated, but their full energy implications are wider. In countries with domestic markets large enough to support appliance production, the increased demand for appliances associated with urbanization feeds back into industrial demand for energy to produce goods made of metals and plastics.

These income-induced increases in end use energy may be mitigated somewhat by the use of more efficient fuels and combustion technologies by urban consumers. In addition, some potential fuel substitutions induced by higher incomes may be restricted by food preparation traditions. For example, kerosene and electricity are poor substitutes for wood and charcoal in grilling food.

4. AGGREGATE ENERGY IMPLICATIONS OF URBANIZATION

4.1 Studies and Data

As noted earlier, although many of the preceding insights have been quantified, and although some have even been subjected to statistical analysis, they defy combination into a comprehensive measure of how much urbanization can be expected to contribute to a nation's energy use. However, two studies have assembled aggregate data for groups of countries and conducted statistical analyses of the relationships among urbanization, various indexes of development, and energy use. In 1989, Donald Jones analyzed a cross section of 1980 data from 59 developing countries. In 1995, Jyoti Parikh and Vibhooti Shukla conducted a somewhat different statistical analysis on a series of 78 developing and industrialized countries over the period from 1967 to 1985. Both studies yielded elasticities of energy use with respect to urbanization, that is, the percentage change in national energy use for a 1% increase in the share of the national population living in cities.

The two studies' results differ in some details, but altogether they largely corroborate each other. Jones analyzed modern, traditional, and total energy separately, whereas Parikh and Shukla analyzed only total energy. Where their quantitative results do differ somewhat, these differences can be explained largely by the differences in their measurements of income and energy and the variables they use to control for outside influences.

Jones used gross domestic product (GDP) data that had been adjusted for purchasing power parity (PPP) corrections for the prices of the nontradable components of GDP, whereas Parikh and Shukla did not use such PPP corrections in their GNP data. (The goods that a developing country produces that are traded internationally, or that compete with internationally traded goods, are priced—even domestically—at close to the international prices. However, the prices of personal services and immobile or otherwise nontraded goods, such as housing, depend on the income levels of the populations. The PPP correction raises the prices attributed to those nontraded goods and services closer to their international valuations. An example might be a lawyer in Chicago who gets a haircut for $20 before leaving for India, where he spends 6 weeks and needs another haircut, which costs him, calculating through the exchange rate [ER], $2. It is the same haircut, but it is offered by a barber with sharply lower costs of living in India. PPP corrects for such discrepancies in prices of equivalent goods that cannot be brought into parity by international trade.) Parikh and Shukla controlled for the economic structure of countries by using the share of agriculture in GDP, whereas Jones accomplished that control function with the share of GDP in industry (although Jones used a simple ER conversion, without the PPP corrections, in the share calculation, and this will yield a smaller industrial share for poorer countries). Jones controlled for nationwide population density with arable land, whereas Parikh and Shukla used all land as their control for population density. Energy use can be expected to exhibit more sensitivity to the PPP measure of GDP than to the ER measure because the adjustments in the PPP measure tend to raise estimates of GDP at earlier dates in a given country and at any particular time to raise the GDPs of lower income countries. Consequently, a given variation in energy use responds to a larger change in ER-measured GDP, leaving the response

appearing smaller with ER–GDP elasticities than with PPP–GDP elasticities. Both measures are valid, but the differences in what they measure should be kept in mind when considering the responsiveness of other variables to them.

4.2 The Pure Urbanization Effect on Energy Use

Despite these data differences, similar stories emerge. The magnitude of the urbanization elasticity of energy use depends importantly on whether one refers to all energy or just modern energy. The changes in total energy include substitutions of modern energy for traditional, and this depresses the urbanization elasticity of total energy (0.35, meaning that a 1% increase in urbanization increases total energy use per unit of GDP by 0.35%) below that of modern energy (0.47). The urbanization elasticity of traditional energy, whether measured on a per capita basis or per dollar of GDP, is not statistically different from zero. Parikh and Shukla's urbanization elasticity for total energy per capita (0.47) is virtually the same magnitude as Jones's urbanization elasticity for modern energy per dollar of GDP (0.48) and is very close to the per capita version (0.45), but it is roughly half again the magnitude of Jones's urbanization elasticities for total energy (0.35 and 0.30).

Parikh and Shukla's alternative analysis took advantage of the cross-sectional component of their data set to estimate country-specific effects for each country estimated from the time series. This estimation cut the magnitude of the urbanization elasticity in half (to 0.25), but it also cut the estimates of all other influences by the same magnitude or more, transferring much of the explanatory power of the analysis to the country effects. Approximately 40% of the country-specific effects can be explained by the percentage of a country's urban population residing in cities with populations exceeding 500,000; the larger the share of the population living in those big cities, the more energy use that is accounted for by the country-specific effect. This extra explanatory power of population concentration in larger cities may capture additional energy use in transportation congestion. The fact that another dimension of urbanization accounts for so much of the variation in per capita energy use across countries suggests that the 0.25 urbanization elasticity from the analysis relying on country-specific effects is not entirely comparable to the urbanization elasticities from the other statistical approaches, that is, that the difference between that and 0.47 depends on how the urban transportation energy is accounted.

4.3 The Effect of Income Changes, Industrialization, and Population Density on Energy Use

These urbanization elasticities were estimated while controlling for GDP per capita, the level of industrialization, and (in some cases) population density. Jones's estimates of the GDP elasticity with the PPP-adjusted GDP yielded larger values than did Parikh and Shukla's ER-converted GDP value (0.47). The GDP elasticities of modern energy of 1.10 and 0.95 are larger than the corresponding elasticities of total energy (0.77 and 0.64). The GDP elasticities of traditional energy are effectively zero. Jones's industrialization elasticity is positive, as expected, but varies considerably among modern, traditional, and total energy as well as across per dollar of GDP and per capita measures of energy. The elasticity per dollar of GDP for industrialization is 1.08 for modern energy but is effectively zero for traditional energy, and this weakens the statistical estimation of the corresponding elasticity of total energy. For energy per capita, the industrialization elasticity for modern energy is smaller than on the per product basis (0.83), but the elasticity for traditional energy per capita is nearly as large but negative (-0.67); as a country industrializes, people move away from traditional energy. Parikh and Shukla's agricultural GDP share variable depresses energy use per capita, with an elasticity of -0.69 that becomes effectively zero when controlling for country-specific effects. Parikh and Shukla's variable and Jones's industrial share variable are roughly mirrors of one another.

The control for population density is useful only in the analysis of traditional energy per dollar of GDP and modern and traditional energy per capita. It is small and of marginal statistical significance in Parikh and Shukla's analyses, but their definition, using all land rather than arable land, is less precise than the definition using arable land. Higher arable population densities reduce traditional energy use per dollar of GDP and per capita fairly sharply, and they yield small positive elasticities of modern energy use per capita. The effect on traditional energy reflects the higher land cost of supporting animals for dung and forests for wood and charcoal at higher population densities as well as the greater convenience of more compact modern fuels and the lower

hook-up costs of electricity. The opposite effect on modern energy per capita simply mirrors the substitution of modern fuels for traditional fuels prompted by higher densities.

Increases in energy prices would moderate energy consumption. Parikh and Shukla found that an index of sensitivity to energy price changes depressed the magnitudes of the country fixed effects. Jones found that two fuel-specific prices, those for oil and gasoline, depressed modern energy consumption, both per capita and per dollar of GDP, with an elasticity of approximately −0.30.

4.4 Significance of the Aggregate Findings

The United Nations' projection of urbanization in the developing world between 2000 and 2030, or roughly one generation, is a 40% increase. If GDP per capita and industrialization remained unchanged over that period, the urbanization alone, using Parikh and Shukla's elasticity obtained without the country-specific effects, would cause a 19% increase in per capita consumption of modern and traditional energy. Jones's elasticity for energy per dollar of GDP predicts the same 19% increase in modern energy alone. Some developing countries had well below 40% urbanization levels in 2000, with China and India being the most prominent at 32 and 28%, respectively. It is not out of the realm of possibility that those two countries' urbanization levels could double in two or possibly three generations. Again, without any increase in income or industrialization, the doubling of urbanization levels alone could increase energy consumption per capita or per dollar of GDP by nearly 50%.

It is reasonable to believe that these parameters would change over a two- to three-generation period of time, probably decreasing somewhat in absolute magnitude. Nonetheless, the Parikh and Shukla elasticities were estimated with the world's full range of development, income, and urbanization, and their fixed-effects elasticity of 0.28 might be a reasonable long-run lower bound. A 28% increase in energy consumption would still be worth noticing. Technological change probably would tend to mitigate such increases, and price increases would further retard growth in energy consumption. Nonetheless, urbanization can be expected to be a major contributor to growth in energy consumption over the next century.

SEE ALSO THE FOLLOWING ARTICLES

City Planning and Energy Use • Ecological Footprints and Energy • Economic Growth and Energy • Heat Islands and Energy • Land Requirements of Energy Systems • Lifestyles and Energy • Population Growth and Energy • Renewable Energy and the City • Suburbanization and Energy

Further Reading

Jones, D. W. (1989a). Energy implications of urbanization in the Third World. *In* "Spatial Energy Analysis: Models for Strategic Decisions in an Urban and Regional Context" (L. Lundqvist, L.-G. Mattsson, and E. A. Eriksson, Eds.), pp. 49–69. Avebury, Aldershot, UK.

Jones, D. W. (1989b). Urbanization and energy use in economic development. *Energy J.* **10**(1), 29–44.

Jones, D. W. (1989c). "Urbanization and Energy Use in Economic Development," ORNL-6432. Oak Ridge National Laboratory, Oak Ridge, TN.

Jones, D. W. (1991). How urbanization affects energy use in developing countries. *Energy Policy* **19**, 621–630.

Parikh, J., and Shukla, V. (1995). Urbanization, energy use, and greenhouse effects in economic development: Results from a cross-national study of developing countries. *Global Environ. Change* **5**, 87–103.

Value Theory and Energy

ROBERT COSTANZA
University of Vermont
Burlington, Vermont, United States

1. **Definitions of Value**
2. **History of Economic Concepts of Value**
3. **Ecological Concepts of Value**
4. **Cost and Price**
5. **Conclusion**

Glossary

diminishing marginal utility The desire for one additional unit declines with successive units of the good.

embodied energy The direct and indirect energy required to produce a good or service.

evolution In natural systems it has three components: (i) generation of genetic variation by random mutations or sexual recombination, (ii) natural selection by relative reproductive success, and (iii) transmission via information stored in the genes.

valuation The process of assessing the contribution of a particular object or action to meeting a particular goal, whether or not that contribution is fully perceived by the individual.

value The contribution of an object or action to specific goals, objectives, or conditions.

value systems Intrapsychic constellations of norms and precepts that guide human judgment and action. They refer to the normative and moral frameworks people use to assign importance and necessity to their beliefs and actions.

The concepts of value system, value, and valuation have many meanings and interpretations and a long history in several disciplines. This article provides a survey of some of these meanings to set the stage for a discussion of the relationship between energy and value. There is clearly not one "correct" set of concepts or techniques to address this important issue. Rather, there is a need for conceptual pluralism and thinking "outside the box." After a long and interesting history, the issue of value is now going through another period of rapid development that should help us to make better, and more sustainable, decisions, not only as individuals but also as groups, communities, and stewards of the entire planet.

1. DEFINITIONS OF VALUE

The terms value system, value, and valuation have a range of meanings in different disciplines. The definitions provided here are used in this article in order to impose some consistency and to put the range of prior and current definitions in context.

Value systems refer to intrapsychic constellations of norms and precepts that guide human judgment and action. They refer to the normative and moral frameworks people use to assign importance and necessity to their beliefs and actions. Because value systems frame how people assign importance to things and activities, they also imply internal objectives. Value systems are thus internal to individuals but are the result of complex patterns of acculturation and may be externally manipulated through, for example, advertising.

Value refers to the contribution of an object or action to specific goals, objectives, or conditions. The value of an object or action may be tightly coupled with an individual's value system because the latter determines the relative importance to the individual of an action or object relative to other actions or objects within the perceived world. However, people's perceptions are limited, they do not have perfect information, and they have limited capacity to process the information they do possess. An object or activity may therefore contribute to meeting an individual's goals without the individual being fully (or even vaguely) aware of the connection. The value of an object or action therefore needs to be assessed both from the subjective standpoint of individuals and their internal value systems and from the

 © 2004 Elsevier Inc. All rights reserved.

objective standpoint of what we may know from other sources about the connection.

Valuation is the process of assessing the contribution of a particular object or action to meeting a particular goal, whether or not that contribution is fully perceived by the individual. One can thus (and must if one hopes to be comprehensive and accurate) do valuation from multiple perspectives, using multiple methods (including both subjective and objective), against multiple goals.

Intrinsic value refers more to the goal or basis for valuation and the protection of the "rights" of these goals to exist. For example, if one says that nature has intrinsic value, one is really claiming that protecting nature is an important goal in itself. Values (as defined previously) are based on the contribution that something makes to achieving goals (directly or indirectly). One could thus talk about the value of an object or action in terms of its contribution to the goal of preserving nature but not about the intrinsic value of nature. Therefore, intrinsic value is a confusing and unnecessary term and will not be used further. One should more accurately refer to the intrinsic rights of nature to qualify as a goal against which to assess value, in addition to the more conventional economic goals.

2. HISTORY OF ECONOMIC CONCEPTS OF VALUE

The history of economic thought is replete with struggles to establish the meaning of value, both what is it and how is it measured. Aristotle first distinguished between value in use and value in exchange. The distinction between use and exchange value has been "resolved" several times, but it remains an important issue even today. For example, the diamond–water paradox observed that although water has infinite or indefinite value, being necessary for life, its exchange value is low, whereas unessential diamonds bear a high exchange value. Following this observation, there was widespread recognition of the distinction between exchange value and use value of goods. Galiani defined value to mean a relation of subjective equivalence between a quantity of one commodity and a quantity of another. He noted that this value depends on utility and scarcity (*utilita et rarita*). Two hundred years later, Adam Smith distinguished between exchange value and use value of goods by citing the diamond–water paradox but used it to dismiss use value as a basis for exchange value. Smith formulated a cost of production theory of value, whereby wages, profit, and rent are the three original sources of exchange value. In his famous beaver–deer example, he suggested a labor theory of exchange value: If it takes twice the labor to kill a beaver than to kill a deer, one beaver will sell for as much as two deer. He also suggested a labor-disutility theory of exchange value, noting that goods exchange based on the unpleasantness of the labor required to bring the goods to market. However, it is significant that Smith limited his labor theory to "that early and rude state of society which precedes both the accumulation of stock and the appropriation of land." In other words, when labor is the only scarce factor, goods will exchange based on the ratio of labor use.

In addition to formulating his hypothesis regarding the origins of exchange value, Smith sought to establish a unit of measure of value or what he termed the real measure or real price of a good. He proposed that "labor alone … never varying in its own value, is alone the ultimate and real standard" of the values of all commodities. Hence, labor could be a numeraire, and it had special properties of invariant value.

Ricardo also sought an invariant unit of measure for value. He believed that there was no commodity, including labor, whose exchange value could serve as an invariant standard to measure the variation in exchange values of other commodities. Also, it was not possible to add up commodities to measure national wealth or production with only exchange ratios. According to Ricardo, this measure must be invariant to changes in relative factor rewards (i.e., capital versus labor) and be a commodity whose capital and labor use did not vary over time (i.e., no technological change). He proposed that both wheat and gold possessed these properties. Although not creating value, they could measure value.

Although Ricardo had several followers, including J. S. Mill and Marx, labor theories of value and the pursuit of an invariant standard of value waned in the late 19th century. This was partially in response to the logic of the utilitarians, such as Menger, Gossen, Jevons, and Walras, who argued that exchange value was based on both utility and scarcity. Sraffa, a noted Ricardian scholar, sought to resurrect the classical pursuit of a theory of value independent of demand or value in use. In his 1960 book, "Production of Commodities by Means of Commodities: Prelude to a Critique of Economic Theory," Sraffa established conditions under which exchange ratios between commodities can be determined based on their use

in production (i.e., a set of commodity prices that would exhaust the total product). These exchange ratios were not based on any optimality or marginality conditions. Instead, Sraffa divided commodities into basic (goods that entered into all production processes) and nonbasic, and he showed that an invariant standard of value would be a combination of basic commodities reflecting average input proportions in production. This contrived "commodity" would then be usable as a measure of national wealth or income.

The "marginal" revolution in value theory originated with the confluence of several related streams of economic thought in the 20th century. Menger proposed there were different categories of wants or desires, such as food, shelter, and clothing, that could be ordered in terms of their subjective importance. Within each category, there is an ordered sequence of desires for successive increments of each good. He postulated that the intensity of desire for one additional unit declines with successive units of the good. Replacing the term "desire for one additional unit" with the term "marginal utility," we thus have the economic principle of diminishing marginal utility.

The idea that people have different, but ordered, categories of wants or desires raises the critical issue of whether trade-offs exist between categories. If individuals weight categories, it implies a trade-off. At one extreme, categories may be lexicographically ordered, like words in a dictionary. One level of want must be satisfied before a lower level becomes relevant in the process of valuation. There are no trade-offs between levels of wants. For example, the need for caloric intake is likely superior to that of recreational pleasure: No number of recreational opportunities will likely substitute for an insufficient diet. In the lexicographic case, individuals would use their monetary resources hierarchically, satisfying higher order wants and needs first. When a higher order want or need is at risk, the individual would take resources away from lower level ones until higher level needs were satisfied. Lexicographic preferences do not mean monetary valuation is impossible because individuals would still be able to state how much of their resources they would be willing to sacrifice for a good or service; however, this may be all their resources if a high-level need is at risk.

More problematic for valuation are instances in which basic needs cannot be satisfied by the resources at an individual's disposal (i.e., time or money). Similar to Menger, Ekins and Max-Neef suggested the universality of basic human needs, including subsistence, affection, protection, understanding, leisure, identity, and freedom. Although one can imagine needs such as affection being "purchasable" with money, or freedom being purchasable by migration, many of these needs may not be satisfied by money or time because individuals simply may not consider them to be purchasable by money or time. Thus, not only is it possible that trade-offs between needs will not be possible but also some needs may not be reducible to money or time.

Lancaster introduced the concept of consumption technology, whereby consumers consider characteristics of goods. For example, food may be evaluated on caloric, protein, or vitamin content. Different foods are substitutable depending on the composition of their characteristics. People allocate their budget across characteristics, purchasing goods that are efficient sources of desired characteristics. The technological inability to substitute characteristics may restrict the margins on which environmental goods and services can be valued. For example, although health may be valued, and individuals would be willing to pay for it, the proper mix of calories, protein, and vitamins may make marginal increases or decrements in one of these characteristics either very highly valued or of very low value.

Building on this insight, multiattribute utility theory formalizes the utility-generating technology by proposing that total utility is a function of the characteristics of goods or services. A simple example is the case in which utility, U, from food consumption is a linear function of the caloric, C, protein, P, and vitamin, V, content:

$$U = aC + bP + cV,$$

where the parameters a–c reflect the weighting of three factors in determining utility for food consumption. When utilities are measurable in monetary willingness to pay or willingness to accept compensation, these parameters represent the marginal monetary value of each characteristic. This logic forms the basis for hedonic pricing models of valuation (discussed later) whereby the value of market goods such as a house depends on the characteristics of the house and its location as well as surrounding environmental amenities or disamenities.

Gossen proposed that in order to maximize satisfaction from a good, such as labor or money, an individual must allocate that good across different uses to equate its marginal utility in each use. Hence, marginal utility would provide a basis for explaining exchange value. If we treat things such as iron, cement, fertilizer, natural agents, and labor as incomplete consumable goods, the marginal utility

of the goods they produce can be used to explain their exchange value. This logic established a theory of value. It also demonstrated that exchange values could be based on marginal use value.

While the classical theorists sought a standard physical commodity unit for measuring exchange value, neoclassical theorists substituted utility for such a commodity. Because value was assumed to be determined solely by consumption utility on the margin, and consumers were assumed to allocate money optimally across uses (in possession of perfect information, no externalities, fixed preferences, and no interpersonal effects), the marginal utility of money was the same for an individual in all its uses. Money thus became the standard unit of measure.

The general optimization model of labor/leisure and consumption/saving given time and wealth constraints would yield equivalencies of goods for money, goods for time, and time for money. Time or money can thus be used as a standard of measure of use value; how much time or money will a person willingly sacrifice to obtain commodity X?

The utility-based values of goods and services are reflected in people's willingness to pay (WTP) to attain them or their willingness to accept compensation to forego them (WTA). WTP and WTA become measures of these values. They may be based on small marginal changes in the availability of these goods and services or on larger changes, including their complete absence or presence. These valuations are reflected in Fig. 1. Let the curve D represent the WTP for each unit of the good or service, T, for an individual or group. This is a marginal WTP. The total WTP for T_0 units of T is the aggregated areas A + B. Area A may be very large for goods or services that have some utility threshold where the good becomes increasingly valuable as it becomes scarcer. This is true for many ecological goods and services, including life-support goods such as oxygen and water; the marginal value is finite but the total value is indeterminate. This is the distinction that lies behind the diamond–water paradox discussed previously.

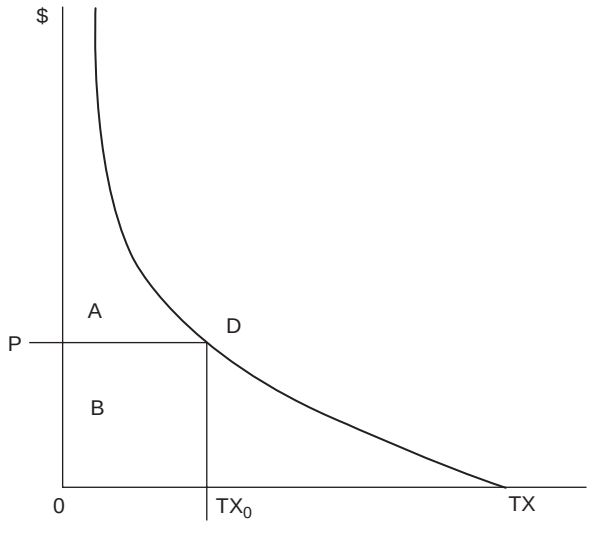


FIGURE 1 Utility- and exchange-based values of goods and services.

Exchange-based values are reflected in the prices, P, at which the goods or services are exchanged. When supply is T_0, and the item is sold competitively, a price P is determined that clears the market. These prices also reflect the marginal valuations placed on available quantities around T_0. Therefore, prices reflect marginal values when there are markets for the goods or services.The total exchange value of T_0 is $P \times T_0$. This is an observable market value when there are markets to observe. However, when there are no such markets, P must be determined indirectly, and $P \times T_0$ would represent a pseudo-market value. This would be the total exchange value of the good if there were a market with an available supply of T_0.

Measures of economic value are assumed to reflect the difference that something makes to satisfaction of human preferences. If something is attainable only at a cost, then the difference it makes to satisfy preferences is the difference between its utility and the cost of attaining it. Formal concepts of compensating and equivalent variations are used to reflect this difference. For example, suppose in Fig. 1 that T_0 is available at a cost of P. Under these terms of availability, the welfare difference made by T_0 is area A. The marginal value that alterations in availability make to welfare would be reflected by changes in area A. Using timber from trees as an example, suppose timber is harvested at a cost of P per unit of timber. The value of trees, per se, would be represented by area A, which is less than A + B.

Thus conceived, the basic notion of value that guides neoclassical economic thought is inherently anthropocentric or instrumental. Although value can generally mean the contribution to a goal, objective, desired condition, etc., the mental model used by neoclassical economists is that value is based on want satisfaction, pleasure, or utility goals. Things have value insofar as they propel individuals toward meeting pleasure and need objectives. Values of objects in the environment can be considered on the margin as well as on the whole (i.e., the value of one additional tree versus the value of all trees).

Although value relates to the utility of a thing, the actual measurement of value requires some objective measure of the degree to which the thing improves pleasure, well-being, and happiness.

In a finite world, the resources people have available to meet their personal objectives are limited. Economists have thus developed an extensive theory of how people behave in the presence of constraints on feasible activities. The working hypothesis is that people make decisions in order to optimize satisfaction, pleasure, or utility. This optimization always takes place in the presence of constraints, such as income, wealth, time, and resource supply. Optimization thus yields a set of possible decisions in most real-world situations—when constraints change, so do the decisions.

The essence of this perspective is that the economic world works largely deterministically, moving from one equilibrium to another in relatively stable fashion, and responds to changes in constraints in a predictable fashion. The determination of equilibrium is a resultant of conflicting forces, such as supply and demand, or unlimited wants and limited means. Although there are instances of instability, disequilibria, and indeterminism, these are treated as exceptions rather than the rule.

Since individuals can be observed making choices between objects in the marketplace while operating within the limits of income and time, economists have developed measures of value as imputations from these observed choices. Although monetary measures of value are not the only possible yardstick, they are convenient since many choices involve the use of money. Hence, if you are observed to pay $10 for a bottle of wine, the imputation is that you value wine to be at least $10 and are willing to make a trade-off of $10 worth of other things to obtain that bottle. The money has no intrinsic value but represents other things you could have purchased. Time is often considered another yardstick of value; if you spend 2 h golfing, the imputation is that you value the golf experience to be worth more than 2 h spent in other activities. Value is thus a resultant of the expressed tastes and preferences of persons and the limited means with which objects can be pursued. As a result, the scarcer the object of desire, the greater its value will be on the margin.

Importantly, the "technologies" of pleasure and production allow for some substitution between things. A variety of goods can induce pleasure and are thus treated conceptually as utility substitutes. A bear may substitute for an elk in consumption, hunting, and a wildlife viewing experience even though bears and elk are not substitutes in terms of ecosystem function. On the production side, inputs are also considered to be substitutable for one another. Machines and technology can substitute for people and natural inputs. Clearly, economists recognize that the relations between goods and services are often more complicated. For malnourished people, sugar is no technological substitute for protein, even though they both provide calories. As discussed previously, preferences may be lexicographic: Some things are more important than others and cannot be substituted for lower level wants or needs. On the production side, no number of lumbermen is a substitute for timber when there is no timber. Production may require certain inputs, but at the same time there may be substitutability between others. As Krutilla suggests, there may be close substitutes for conventional natural resources, such as timber and coal, but not for natural ecological systems.

The neoclassical perspective also assumes that tastes and preferences are fixed and given, and that the fundamental economic "problem" consists of optimally satisfying those preferences. Tastes and preferences usually do not change rapidly and, in the short term, this basic assumption is probably not too bad. In the longer term, however, it does not make sense to assume tastes and preferences are fixed. People's preferences do change over longer time frames, as the existence of a robust advertising industry attests. This observation is important because sustainability is an inherently long-term concept and ecosystem services are expected to continue far into the future. This fact is very disturbing for many economists because it takes away the easy definition of what is optimal. If tastes and preferences are fixed and given, then we can adopt a stance of "consumer sovereignty" and just give people what they want. We do not have to know or care why they want it; we just have to satisfy their preferences efficiently.

However, if preferences change over time and under the influence of education, advertising, changing cultural assumptions, variations in abundance and scarcity, etc., we need a different criterion for what is optimal. Moreover, we have to determine how preferences change, how they relate to this new criterion, and how they can or should be changed to satisfy the new criterion. One alternative for the new criterion is sustainability, or more completely a set of criteria: sustainable scale (size of the economic subsystem), fair distribution, and efficient allocation. This set of criteria implies a two-tiered decision

process of first coming to a social consensus on a sustainable scale and fair distribution and then using the marketplace and other social institutions, such as education and advertising, to implement these decisions. This might be called community sovereignty as opposed to consumer sovereignty. It makes most economists very uncomfortable to stray from consumer sovereignty because it raises the following question: If tastes and preferences can change, then who is going to decide how to change them? There is a real danger that a totalitarian government might be employed to manipulate preferences to conform to the desires of a select elite rather than the individuals in society.

Here, two points need to be kept in mind: Preferences are already being manipulated every day, and we can just as easily apply open democratic principles to the problem as hidden or totalitarian principles in deciding how to manipulate preferences. Viewed in this light, the aforementioned question is transformed: Do we want preferences to be manipulated unconsciously, either by a dictatorial government or by big business acting through advertising, or do we want to formulate preferences consciously based on social dialogue and consensus with a higher goal in mind? Either way, we believe that this issue that can no longer be avoided and is one that will best be handled using open democratic principles and innovative thinking. This leads us back to the role of individual preferences in determining value. If individual preferences change in response to education, advertising, and peer pressure, then value cannot solely originate with individual preferences. Values ultimately originate from within the constellation of shared goals to which a society aspires—value systems—as well as the availability of "production technologies" that transform things into satisfaction of human needs.

In addition to income and education, time places constraints on value creation. Constraints of time and intertemporal substitutabilities create temporal implications for value. Neoclassical economists presume that a present time preference exists due to limited time horizons and concerns for uncertainty in the future. This means individuals will discount values of things in the future in comparison to the same things in the present. If I have an equal endowment of apples now and a year from now, I would place a greater value on having an apple now than on having an apple 1 year from now. The ability to convert things to money in the presence of positive financial interest rates will therefore result in the "optimizing individual" discounting things in the future.

In contrast to economists' traditional assumptions of positive time preferences, or positive discount rates, psychologists suggest time preference is more complicated. For example, Lowenstein and Prelec find that in some circumstances people behave as if they have negative time preference, preferring more in the future to more now. The authors suggest this is due to dread, the anticipation of savoring better conditions in the future, and the aversion to loss. However, this negative time preference may not be operative when the time period is ambiguous. The implications of such experimental results for discounting in environmental policy settings are not clear, but they do raise serious questions about the standard practice of discounting future environmental benefits.

3. ECOLOGICAL CONCEPTS OF VALUE

Value is a term that most ecologists and other natural scientists would prefer not to use at all, except perhaps in its common usage as a reference to the magnitude of a number (e.g., "the value of parameter b is 9.32"). Using the definitions provided earlier, ecosystems and nonhuman species are presumed not to be pursuing any conscious goals, and therefore they do not have a value system. Likewise, one cannot talk about value as the degree to which an item contributes to achieving a goal in this context since there is no conscious goal being pursued. Nevertheless, some concepts of value are important in the natural sciences and are in fact quite commonly used.

If one limits the concept of value to the degree to which an item contributes to an objective or condition in a system, then we can see how natural scientists use the concept of value to talk about causal relationships between different parts of a system. For example, one could talk about the value of particular tree species in controlling soil erosion in a high slope area or the value of fires in recycling nutrients in a forest.

There are other ways in which the concept of value is used in the natural sciences. For example, a core organizing principle of biology is evolution by natural selection. Evolution in natural systems has three components: generation of genetic variation by random mutations or sexual recombination, natural selection by relative reproductive success, and transmission via information stored in the genes.

Although this process does not require conscious, goal-directed behavior on the part of any of its participants, one can still think of the overall process as being "goal-directed." The goal of survival is embedded in the objective function of natural selection. Although the process occurs without consciousness of this goal, individuals and species as a whole can be observed to behave as if they were pursuing the goal of survival. Thus, one often hears evolutionary biologists talk about the survival value of particular traits in organisms. Natural selection models, which maximize the fitness of individuals, are not only testable but also bear close similarities to economic utility maximization models.

In addition, the idea of coevolution among a whole group of interacting species raises the possibility that one species is valuable to the survival of another species. Extending this logic to the coevolution of humans and other species, we can talk of the value of natural ecosystems and their components in terms of their contribution to human survival.

Ecologists and physical scientists have also proposed an energy theory of value to either complement or replace the standard neoclassical theory of subjective utility-based value. It is based on thermodynamic principles in which solar energy is recognized to be the only primary input to the global ecosystem.

This theory of value represents a return to the classical ideas of Ricardo and Sraffa but with some important distinctions. The classical economists recognized that if they could identify a primary input to the production process then they could explain exchange values based on production relationships. The problem was that neither labor nor any other single commodity was really primary since they all require each other for their production. The traditional primary factors are really intermediate factors of production.

The classical economists were writing before the science of thermodynamics had been fully developed. Energy—or, more correctly, free or available energy defined as the ability to do work—has special characteristics that satisfy the criteria for a primary input described previously:

1. Energy is ubiquitous.
2. It is a property of all the commodities produced in economic and ecological systems.
3. It is an essential input to all production processes.
4. Although other commodities can provide alternative sources for the energy required to drive systems, the essential property of energy (the ability to do work) cannot be substituted.
5. At the global scale, the earth is essentially a closed system in thermodynamic terms (only energy crosses the boundary), so at this scale it is the only primary input.
6. Smith's three sources of exchange value (wages, profits, and rent) are intermediate inputs in this global scheme and interconvertable using the primary energy input.

Available energy is thus the only basic commodity and is ultimately the only scarce factor of production, thereby satisfying the criteria for a production-based theory that can explain exchange values.

Energy-based concepts of value must follow the basic principles of energy conversion. The first law of thermodynamics states that energy and matter are conserved. However, this law essentially refers to heat energy and mechanical work (raw energy or the bomb calorimeter energy). The ability to do work is related to the degree of organization or order of a thing relative to its environment, not its raw heat content. Heat must be organized as a temperature gradient between a high-temperature source and a low-temperature sink in order for useful work to be done. Similarly, complex manufactured goods such as cars have an ability to do work that is not related to their raw energy content but may be related to their degree of organization relative to their environment. The second law of thermodynamics states that useful energy (organization) always dissipates (entropy or disorder always increases) within an isolated system. In order to maintain organized structures (e.g., an economy) one must constantly add organized, low-entropy energy from outside the system. The earth as a whole is just such a system. In thermodynamic terms, it is essentially closed (i.e., energy, but not matter, crosses the boundaries). Of course, some matter does cross the boundaries (i.e., meteorites and spacecraft) so the earth system is at least slightly open, but these flows of matter are very small compared to the flow of solar energy in and heat energy out.

Estimating total energy consumption for an economy is not a straightforward matter because not all fuels are of the same quality; that is, they vary in their available energy, degree of organization, or ability to do work. Electricity, for example, is more versatile and cleaner in end use than petroleum, and it also requires more energy to produce. In a oil-fired power plant it takes 3–5 kcal of oil to produce 1 kcal of electricity. Thus, adding up the raw heat equivalents

of the various forms of fuel consumed by an economy without accounting for fuel quality can radically distort the picture, especially if the mix of fuel types is changing over time.

An energy theory of value thus posits that, at least at the global scale, free or available energy from the sun (plus past solar energy stored as fossil fuels and residual heat from the earth's core) are the only primary inputs to the system. Labor, manufactured capital, and natural capital are intermediate inputs. Thus, one could base a theory of value on the use in production of available energy that avoids the problems that the classical economists encountered when trying to explain exchange values in economic systems.

There have been a few attempts to empirically test this theory using both time series data and cross-sectional data. Studies that have tried to adjust for fuel quality have shown a very close relationship between available energy consumption and economic output. Cleveland *et al.* and Kaufmann have shown that almost all the changes in energy:gross national product (E:GNP) (or energy:gross domestic product) ratios in the United States and Organization for Economic Cooperation and Development countries can be explained by changes in fuel quality and the percentage of personal consumption expenditures (PCE) spent directly on fuel. The latter effect is due to the fact that PCE is a component of GNP and spending more on fuel directly will raise GNP without changing real economic output. Figure 2 is an example of the explanatory power of this relationship for the U.S. economy from 1932 to 1987. Much of the apparent gain in energy efficiency (decreasing E:GNP ratio) is due to shifts to higher quality fuels (e.g., natural gas and primary electricity) from lower quality ones (e.g., coal). Renewable energy sources are generally lower quality and shifts to them may cause significant increases in the E:GNP ratio.

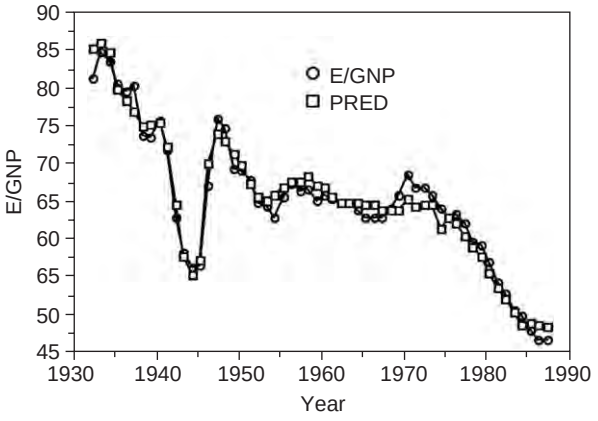


FIGURE 2 The energy:GNP ratio for the U.S. economy from 1932 to 1987. The predicted ratio (PRED) is based on a regression model with percentage of primary energy from petroleum and from electricity and percentage of personal consumption expenditures spent on fuel as independent variables ($R^2 = 0.96$). From Cleveland *et al.* (1984) and Kaufmann (1992).

Another way of looking at the relationship between available energy and economic output uses cross-sectional rather than time-series data. This avoids some of the problems associated with changes in fuel mix and distortions in GNP. For example, Costanza and Costanza and Herendeen used an 87-sector input–output model of the U.S. economy for 1963, 1967, and 1973, modified to include households and government as endogenous sectors (to include labor and government energy costs), to investigate the relationship between direct and indirect energy consumption (embodied energy) and dollar value of output. They found that dollar value of sector output was highly correlated ($R^2 = 0.85$–0.98) with embodied energy when this was calculated including the energy costs of labor, government, and environmental inputs (although not with direct energy consumption or embodied energy calculated excluding labor and government energy costs). Thus, if one makes some necessary adjustments to estimates of energy consumption in order to better assess available energy, it appears that the empirical link between available energy cost and economic value is rather strong.

Some neoclassical economists have criticized the energy theory of value as an attempt to define value independent of consumer preferences. This criticism is, on the one hand, axiomatic since a major purpose of an energy theory of value was to establish a theory of value not completely determined by individual preferences. On the other hand, techniques for calculating embodied energy utilize economic input–output tables. These tables summarize production interdependencies but they are not completely independent of consumer preferences, which helped to structure the production interdependencies over time.

In summary, the energy theory of value overcomes some of the problems of earlier production-based theories of value encountered by the classical economists and does a fairly good job of explaining exchange values empirically in the few cases in which it has been tested. Despite the controversy and ongoing debate about the validity of an energy theory of value, it seems to be the only reasonably successful attempt to operationalize a general biophysical theory of value.

4. COST AND PRICE

Energy (and earlier labor) theories of value are inherently based on relative production costs. Thus, it is more accurate to speak of energy cost or labor cost and not energy value or labor value. However, in economic systems it is well-known that cost and price will, in general, come to equilibrium. This is the essence of the basic ideas of supply and demand. If a commodity has a much higher value than its cost of production, profits will be high and more of the commodity will be produced (with increasing marginal cost) until cost just equals price and profits are 0 for the last unit of production. Likewise, if cost is much higher than price, less will be produced until they are again equal. Therefore, a method that estimates costs, although not technically estimating price or value, should in fact be a fairly good approximation to price and value in cases in which markets have reached an approximate equilibrium. Since markets are never really in equilibrium and there are many types of market failure to complicate things, we would expect there to be some divergence between cost and price in real systems. The question then becomes the following: Are the commodities overpriced or underpriced? Should the energy costs be taken as the standard and market exchange values be adjusted, or should the energy costs be ignored and the market prices taken as the standard? Given, on the one hand, the enormous data requirements to calculate energy costs accurately and, on the other hand, the pervasive market imperfections complicating market prices, there is no unambiguous correct answer. However, one can learn much by examining energy costs and market prices and comparing their degree of correspondence. Costanza did just that at the aggregate level of an 87-sector input–output model and found a fairly high degree of correspondence for those sectors in which markets were fairly functional and lacking severe externalities and a low correspondence for those sectors in which they were not (basically, the extractive sectors at this level of aggregation).

As discussed previously, embodied labor costs alone would give a much poorer approximation to exchange value than embodied energy costs since embodied labor is a less comprehensive measure of total cost than embodied energy, especially in industrial economies in which capital, direct energy, and government expenditures are significant components of total cost relative to labor.

5. CONCLUSION

A modern approach to value and valuation needs to be pluralistic, including a broad range of both subjective and objective methods. Embodied energy cost has proven to be an important element in this spectrum of approaches, and it makes the link with the classical economists' production-based theories of value. Embodied energy should more accurately be called a theory of cost rather than value, although the two are obviously related in well-functioning markets. The most productive approach is to keep these distinctions in mind while applying a range of subjective and objective methods in order to obtain a clearer picture of the complex nature of value and, ultimately, to do a better job of valuation against a much broader range of social goals.

SEE ALSO THE FOLLOWING ARTICLES

Cost–Benefit Analysis Applied to Energy • *Depletion and Valuation of Energy Resources* • *Evolutionary Economics and Energy* • *Industrial Ecology* • *Information Theory and Energy* • *Petroleum Property Valuation* • *Physics and Economics of Energy, Conceptual Overview*

Further Reading

Alexander, A., List, J. A., Margolis, M., and d'Arge, R. C. (1998). A method for valuing global ecosystem services. *Ecol. Econ.* **27**(2), 161–170.

Arrow, K., and Fisher, A. C. (1974). Environmental preservation, uncertainty and irreversibility. *Q. J. Econ.* **88**, 312–319.

Bishop, R. C. (1978). Endangered species and uncertainty: The economics of a safe minimum standard. *Am. J. Agric. Econ.* **60**(1), 10–18.

Blamey, R.K., and James, R. F. (1999). Citizens' juries—An alternative or an input to environmental cost–benefit analysis. *In* "Conference of the Australian and New Zealand Society for Ecological Economics, Brisbane, Australia, July 7." Griffith University, Australia.

Blaug, M. (1968). "Economic Theory in Retrospect." Irwin, Homewood, IL.

Brown, M. T., and Herendeen, R. A. (1996). Embodied energy analysis and energy analysis: A comparative view. *Ecol. Econ.* **19**, 219–235.

Chichilnisky, G., and Heal, G. (1998). Economic returns from the biosphere. *Nature* **391**(6668), 629–630.

Ciriacy-Wantrup, S. V. (1963). "Resource Conservation: Economics and Policies." Univ. of California Press, Berkeley.

Clark, C. W. (1973). The economics of overexploitation. *Science* **181**, 630–634.

Cleveland, C. J., Costanza, R., Hall, C. A. S., and Kaufmann, R. (1984). Energy and the U.S. economy: A biophysical perspective. *Science* **225,** 890–897.

Cobb, J. B. (1988). A Christian view of biodiversity. *In* "Biodiversity" (E. O. Wilson, Ed.). National Academy Press, Washington, DC.

Coote, A., and Lenaghan, J. (1997). "Citizen Juries: Theory into Practice." Institute for Public Policy Research, London.

Costanza, R. (1980). Embodied energy and economic valuation. *Science* **210,** 1219–1224.

Costanza, R. (2000). Social goals and the valuation of ecosystem services. *Ecosystems* **3,** 4–10.

Costanza, R., and Herendeen, R. A. (1984). Embodied energy and economic value in the United States economy: 1963, 1967, and 1972. *Resour. Energy* **6,** 129–163.

Daly, H. E. (1992). Allocation, distribution, and scale: Towards an economics that is efficient, just, and sustainable. *Ecol. Econ.* **6,** 185–193.

Daly, H. E., and Cobb, J. B. (1989). "For the Common Good: Redirecting the Economy toward Community, the Environment and a Sustainable Future." Beacon, Boston.

Desvouges, W.H., and Smith, V.K. Focus groups and risk communication: The science of listening to data. *Risk Anal.* **8,** 479–484.

Dryzek, J. S. (1987). "Rational Ecology: Environment and Political Economy." Basil Blackwell, New York.

Ehrlich, P., and Raven, P. (1964). Butterflies and plants: A study in coevolution. *Evolution* **8,** 586–608.

Ekins, P., and Max-Neef, M. (eds.). (1992). "Real-Life Economics." Routledge, London.

Fisher, I. (1930). "The Theory of Interest." Macmillan, New York.

Fishkin, J. S. (1991). "Democracy and Deliberation." Yale Univ. Press, New Haven, CT.

Freeman, M. (1993). "The Measurement of Environmental and Resource Values: Theory and Methods." Resources for the Future, Washington, DC.

Gilliland, M. W. (1975). Energy analysis and public policy. *Science* **189,** 1051–1056.

Goulder, L. H., and Kennedy, D. (1997). Valuing ecosystem services: Philosophical bases and empirical methods. *In* "Nature's Services: Societal Dependence on Natural Ecosystems" (G. C. Daily, Ed.), pp. 23–48. Island Press, Washington, DC.

Habermas, J. (1984). "The Theory of Communicative Action." Beacon, Boston.

Hall, C. A. S., Cleveland, C. J., and Kaufmann, K. (1992). "Energy and Resource Quality: The Ecology of the Economic Process." Univ. of Colorado Press, Boulder, CO.

Hannemann, W. M. (1991). Willingness to pay and willingness to accept: How much can they differ? *Am. Econ. Rev.* **81**(3), 635–647.

Harrod, R. (1961). Book review. *Econ. J.* **71,** 783–787.

Heuttner, D. A. (1976). Net energy analysis: An economic assessment. *Science* **192,** 101–104.

Hicks, J. R. (1939). "Value and Capital: An Inquiry into Some Fundamental Principles of Economic Theory." Oxford Univ. Press, London.

Jacobs, M. (1997). Environmental valuation, deliberative democracy and public decision-making. *In* "Valuing Nature: Economics, Ethics and Environment" (J. Foster, Ed.), pp. 211–231. Routledge, London.

James, R.F., and Blamey, R. K. (1999, July 4). Citizen participation—Some recent Australian developments. Paper presented at the Pacific Science Conference, Sydney, Australia.

Kaufmann, R. K. (1992). A biophysical analysis of the energy/real GDP ratio: Implications for substitution and technical change. *Ecol. Econ.* **6,** 35–56.

Kopp, R. J., and Smith, V. K. (1993). "Valuing Natural Assets: The Economics of Natural Resource Damage Assessment." Resources for the Future, Washington, DC.

Krutilla, J. V. (1967). Conservation reconsidered. *Am. Econ. Rev.* **57**(4), 777–786.

Lancaster, K. (1971). "Consumer Demand: A New Approach." Columbia Univ. Press, New York.

Leopold, A. (1949). "A Sand County Almanac." Oxford Univ. Press, New York.

Loewenstein, G., and Prelec, D. (1991). Negative time preference. *Am. Econ. Rev.* **81**(2), 347–352.

Low, B. S. (2000). "Why Sex Matters: A Darwinian Look at Human Behavior." Princeton Univ. Press, Princeton, NJ.

Mitchell, R. C., and Carson, R. T. (1989). "Using Surveys for Value Public Goods: The Contingent Valuation Method." Resources for the Future, Washington, DC.

Norton, B., Costanza, R., and Bishop, R. (1998). The evolution of preferences: Why "sovereign" preferences may not lead to sustainable policies and what to do about it. *Ecol. Econ.* **24,** 193–211.

Odum, H. T. (1971). "Environment, Power and Society." Wiley, New York.

Odum, H. T. (1983). "Systems Ecology: An introduction." Wiley, New York.

Page, T. (1977). "Conservation and Economic Efficiency." Johns Hopkins Univ. Press, Baltimore.

Reder, M. (1961). Book review. *Am. Econ. Rev.* **51**(4), 688–725.

Sagoff, M. (1998). Aggregation and deliberation in valuing environmental public goods: A look beyond contingent valuation. *Ecol. Econ.* **24,** 213–230.

Schumpeter, J. A. (1978). "History of Economic Analysis." Oxford Univ. Press, New York.

Slesser, M. (1973). Energy analysis in policy making. *New Scientist* **58,** 328–330.

Sraffa, P. (1960). "Production of Commodities by Means of Commodities: Prelude to a Critique of Economic Theory." Cambridge Univ. Press, Cambridge, UK.

Tobias, M. (ed.). (1985). "Deep Ecology." Avant, San Diego, CA.

Varian, H. R. (1992). "Microeconomic Analysis." Norton, New York.

Vehicles and Their Powerplants: Energy Use and Efficiency

PHILLIP S. MYERS and DAVID FOSTER
University of Wisconsin at Madison
Madison, Wisconsin, United States

1. **History of Ground Vehicle Power Plants**
2. **Use Classification of Ground Vehicles**
3. **Constraints**
4. **Past Trends**
5. **Energy Sources for Ground Vehicles**
6. **Prognosis for the Future**
7. **Considerations and Problems in Reducing Fuel Consumption**

Glossary

absolute pressure (**P**) Pressure measured relative to total vacuum—or zero pressure.
absolute temperature (**T**) Temperature measured in Kelvin or Rankine.
Carnot cycle An external combustion engine whose compression and expansion strokes are ideally accomplished without heat transfer and for which the heat addition and rejection processes occur at different but constant temperatures. The Carnot cycle engine achieves the maximum theoretical efficiency possible for an external combustion engine operating between the temperatures at which heat addition and heat rejection take place. Its thermal efficiency is given by $\eta = (T_{high} - T_{low})/T_{high}$. For a Carnot cycle, or any external combustion engine, to achieve an efficiency of 100% would require that either $T_{high} = \infty$ or $T_{low} = 0$ (absolute zero).
enthalpy (**H**) Defined as $H = U + PV$.
entropy (**S**) A measure of disorder.
external combustion engine An engine in which heat transfer from combustion occurring outside the engine is transferred across the boundary of the system to the working fluid of the engine (e.g., a steam engine).
Gibb's Free Energy (**G**) Defined as $G = H - TS$.
heat (**Q**) Energy transferred across the boundary of a system because of a temperature difference.
higher heating value (**HHV**) Heat transferred out of a constant pressure system in which complete combustion occurs and the products are brought back to the original temperature, with any resulting water being condensed to liquid.
internal combustion engine An engine in which the working fluid is initially an air-fuel mixture and, after combustion and until exhausted, the products of combustion. It is different from an external combustion engine. The maximum work obtainable from a given amount of fuel is the Gibbs free energy. High temperature is not required to achieve high thermal efficiency. The Carnot cycle limitation of thermal efficiency does not apply to an internal combustion engine. The ideal thermal efficiency for any internal combustion engines is 100%.
internal energy (**U**) Total energy in a defined closed system.
thermal efficiency (η) The ratio of the work out to the input fuel HHV.

Rapid transportation of persons and material is a critical component of modern society. Internal combustion engines, using petroleum as fuel, are the principle energy converter used in transportation. Consequently, transportation is a major consumer of petroleum. Because of increasing use and rapid growth of developing countries, petroleum will undoubtedly increase in cost and the supply is finite. The desirable results of rapid transportation are accompanied by undesirable air pollution, safety issues, and other problems. This article reviews the past trends of emissions, fuel economy, and safety together with customer preferences as well as techniques for minimizing the amount of petroleum needed to continue to meet society's desires for mobility.

1. HISTORY OF GROUND VEHICLE POWER PLANTS

Humans have always valued the capability of transporting themselves and possessions to desired

 © 2004 Elsevier Inc. All rights reserved.

locations. Although wind power was used extensively for water transportation, personal transportation was initially human powered. However, it was soon recognized that the augmented power available by using animals increased the distance that could be covered in a given time. During this period of development, horses became the most commonly used animal, hence the term horsepower. Fundamentally, the basic energy conversion process to develop this power was low-temperature, catalytic oxidation of combustible material (i.e., food).

As technological capability increased, it became increasingly feasible to use high-temperature oxidation (combustion) as the source of transportation energy. Suzuki shows a drawing, made by Huygens in 1673, of a primitive gunpowder internal combustion engine. Also shown is the Newcomen steam engine (1712), an external combustion engine. Each type of energy converter underwent successive development. The internal combustion engine, with its combination of high thermal efficiency and low weight per unit of power, has become the dominant choice for powering transportation vehicles. The manufacture of internal combustion engines for automobiles, boats, airplanes, trains, small power units, etc. is one of the largest industries in the world.

1.1 Thermodynamics

Thermodynamics, using ideal reversible processes, establishes the upper limit for the useful work that can be obtained from a given quantity of fuel and air. The upper limit of efficiency for an external combustion engine is temperature dependent, whereas the upper limit for an internal combustion engine is 100% (i.e., an internal combustion engine can theoretically convert all of the heating value of the fuel into work).

1.2 External and Internal Combustion Engines

The thermodynamic processes involved in external combustion engines and internal combustion engines are fundamentally different. The practical significance of this difference is not generally realized. For simplicity, most engineering thermodynamic texts do not make this distinction clear. For example, the internal combustion engine is often analyzed as if it were an air standard Otto cycle, which is an external combustion engine with heat added and rejected at constant volume. In some situations, this can aid in understanding, but it is a simplification that is fundamentally incorrect. It should also be clear that for both internal and external combustion engines, inevitable irreversible processes, such as friction (mechanical and fluid) and heat transfer to combustion chamber walls, will occur that decrease system thermal efficiency.

2. USE CLASSIFICATION OF GROUND VEHICLES

Transportation energy is responsible for 65% of all petroleum usage in the United States. Riley shows that personal transportation vehicles consume more than half of this petroleum (i.e., light-duty, personal transportation vehicles account for more than 35% of the nation's total consumption of petroleum and are the largest consumers of transportation fuel in the United States) (Fig. 1). Due to the current low price of gasoline, market forces provide little incentive to reduce fuel consumption of light-duty vehicles. However, competitive market forces provide strong incentives to minimize fuel consumption in the remaining sectors shown Fig. 1: freight trucks, air transport, marine, and rail. As discussed in the *Wall Street Journal* (May 14, 2002), definitions of a light-duty vehicle differ. Furthermore, it is not clear if small users of petroleum (i.e., agriculture, construction, lawn maintenance, recreation, etc.) are included in the data depicted in Fig. 1. This article focuses primarily on light-duty vehicles (i.e., cars and light-duty trucks).

3. CONSTRAINTS

Transportation vehicles are sold in a competitive market. However, for reasons of environmental impact, public health, and safety, government

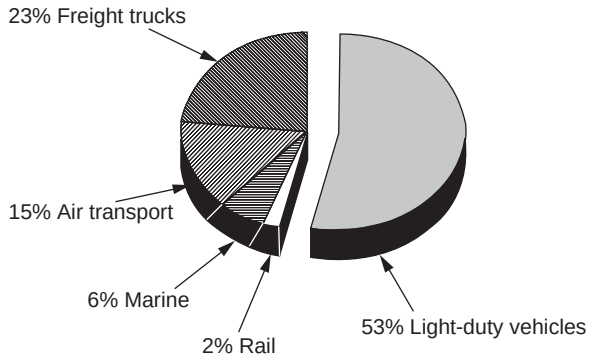


FIGURE 1 U.S. transportation energy consumption, 1990.

regulations, which must be met before vehicles can be sold, have been imposed. Often, these regulations result in attributes of the vehicle that are not part of the consumer's criteria for purchasing the vehicle.

3.1 Government Constraints

There are numerous government-imposed, regulatory constraints on the light-duty fleet. The most significant regulations embody interrelated aspects of a vehicle's operation: emissions, safety, and fuel economy.

3.1.1 Emissions

Four gaseous emissions from the engine are of primary concern for both gasoline and diesel engines: unburned or partially burned fuel, in both vapor and solid form; nitrogen–oxygen compounds, which are formed in the cylinder during high-temperature combustion; carbon monoxide (CO); and carbon dioxide (CO_2). These emissions enter the atmosphere via the vehicle exhaust. Reduction of pollutants in the engine exhaust, except for CO_2, is the primary control technique for reducing emissions. Carbon dioxide, which is directly related to the fuel consumed, is increasingly being viewed as a pollutant because of its impact as a greenhouse gas. Vehicle emissions are measured on a chassis dynamometer that is loaded, in steps, to represent different vehicle weights. The vehicle is operated over a specific driving cycle and emissions and fuel consumption are measured. Different countries have different driving cycles. For a given vehicle classification, regulations are typically expressed in mass of pollutant emitted per unit distance driven. The driving cycle requires engine operation over a variety of speeds and loads. This variation of operating conditions changes the engine's exhaust gas temperature and exhaust constituents. Confounding the situation, the variations in exhaust gas temperature and constituent composition during the driving cycle will be different for different types of engines (e.g., a compression ignition and a spark ignition engine).

3.1.2 Safety

There are many regulations on vehicle and occupant safety. Achieving vehicle occupant safety is difficult and is affected by many factors, including road characteristics; multiple- or single-vehicle collisions; vehicle weight, design, structure, and maneuverability; driver experience, mood, use of alcohol, distraction, use of safety constraints such as seat belts; solid objects within the vehicle; and occupant resiliency.

3.1.3 Fuel Economy

There are many interactive parameters that influence a vehicle's fuel economy. U.S. consumers are familiar with the term miles per gallon to express fuel economy. However, the inverse of miles per gallon, the volume of fuel consumed per unit distance, is the most useful for analysis purposes. Note that there is a nonlinear, inverse relationship between the two expressions. Europe's future vehicle fuel economy goal is 3 liters of fuel per 100 kilometers traveled.

The standard fuel consumption parameter used for engines, brake-specific fuel consumption, is the amount of fuel required to produce one brake horsepower-hour, or kilowatt-hour, of work. An analogous expression for vehicle fuel consumption of cars or light trucks would be gallons/vehicle ton/mile—that is, the gallons of fuel required to move a 1-ton vehicle 100 miles). Obviously, it will take twice as many gallons to move a 2-ton vehicle 100 miles than it will to move a 1-ton vehicle 100 miles. Figure 2 presents this data for 2000 and 2001 model year cars and light trucks. Vehicles of the same weight can have engines with major differences in power, performance, and fuel economy. Consequently, for vehicles of the same weight, significant differences in fuel consumption (i.e., vertical spread in Fig. 2) can be expected. The current system to regulate fuel consumption of light-duty transportation vehicles in the United States is the corporate average fuel economy (CAFE). Vehicles sold by a given manufacturer are divided into two groups—cars and light-duty trucks. Currently, the sales-weighted average fuel economy of all cars sold by a given manufacturer must be 27.5 miles per gallon or higher, whereas the sales-weighted average fuel economy of all light-duty trucks sold by a given manufacturer must be 20.7 miles per gallon or higher. Failure to meet these standards results in significant penalties.

3.2 Customer Preferences

To be saleable, cars and light trucks must meet customer desires. Customer preferences are obviously affected by the cost of desired amenities and by current trends. At current U.S. fuel prices, fuel economy is not a high-priority consumer item. In contrast, in Japan and Europe, where fuel prices are two to four times higher than those in the United States, fuel economy is much more important to the consumer. The trends in consumer choices of vehicles in these countries reflect this difference in priorities.

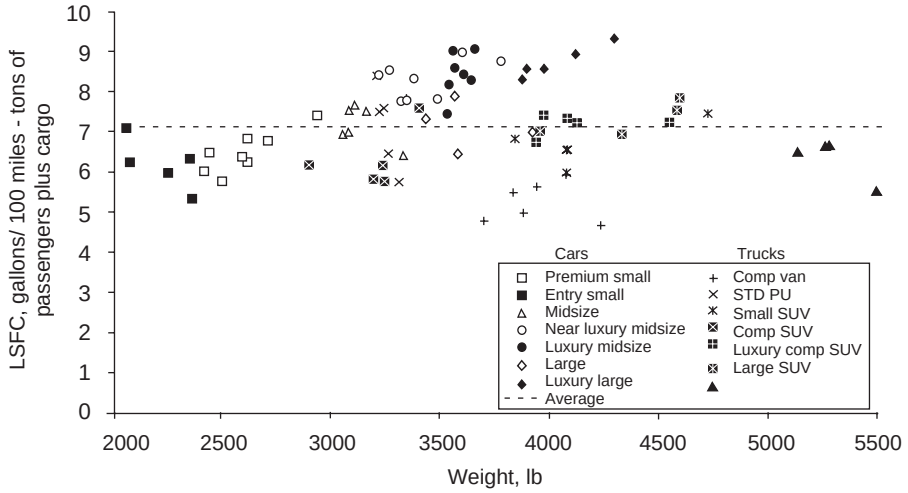


FIGURE 2 Weight-specific fuel consumption versus weight for all vehicles.

4. PAST TRENDS

4.1 Emissions

Vehicle emissions have been of concern since the early 1950s when a California researcher determined that traffic was to blame for the smoggy skies over Los Angeles. At that time, typical new cars were emitting nearly 13 g of hydrocarbons per mile, 3.6 g of nitrogen oxides per mile, and 87 g of CO per mile. Since then, the federal government has set increasingly more complex and restrictive standards to reduce these pollutants. Figure 3 shows the current standard for nitrogen oxides as well as the tier 2 standard. It is clear that dramatic reductions in emissions have been achieved from the vehicles of the early 1950s, but there is still work to be done to meet the tier 2 standards.

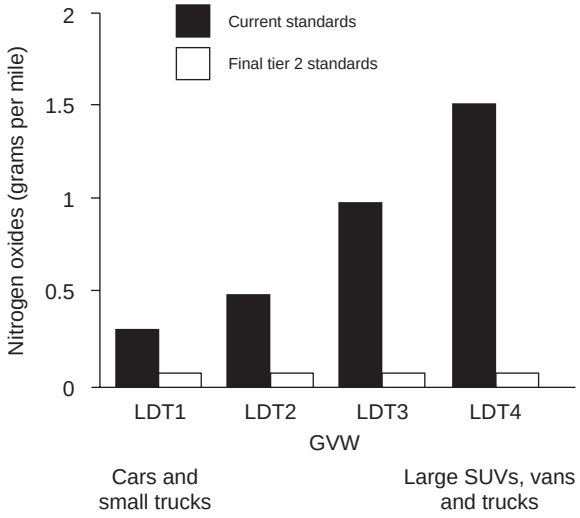


FIGURE 3 Comparison of current vehicle emission standards for nitrogen oxides and final tier 2 standards.

4.2 Safety

Numerous factors influence safety: vehicle design including mass, maneuverability, passenger restraints, driver age, driver sobriety, peripheral distractions, etc. Also significant are traffic design factors, such as traffic control lights, merging lanes, and speed limits. Figure 4 summarizes the history of death rates for vehicle crashes between 1950 and 1998 using three different normalizations. The lower curve (death rate per million population) initially exhibits a small increase followed by a continuing small decrease. The other two curves, the death rate per million vehicles and the death rate per 10 million miles traveled, show significant decreases over time. The complexity of and interaction between the factors causing traffic deaths (two-vehicle crashes, single-vehicle crashes, vehicle–pedestrian crashes, vehicle–fixed-object crashes, intoxication, speed, etc.) preclude simple conclusions and result in disagreement between authorities in the field.

4.3 Fuel Economy

From a national standpoint, fuel economy can be expressed in many different ways: per vehicle; per class of vehicle, car or light truck; a sales-weighted average for vehicles of a given classification from a given manufacturer; etc. For regulation purposes, the U.S. federal government uses the fuel consumption

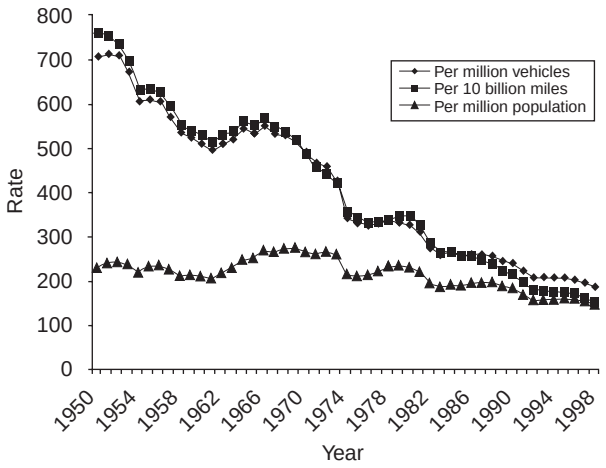


FIGURE 4 Motor vehicle crash death rate.

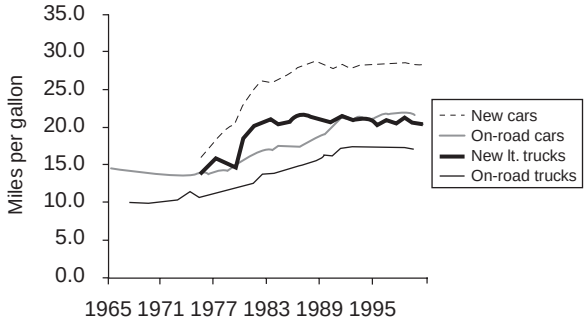


FIGURE 5 Fleet fuel economy of new and on-road passenger cars and light trucks.

results obtained from the emission-measurement test cycle. Figure 5 shows the history of fleet fuel economy for cars and light trucks as well as their on-road fuel economy. The difference between chassis-measured fuel economy and on-road fuel economy is significant. Figure 6 also shows trends in fuel economy-related characteristics of passenger cars, whereas Fig. 7 shows the fuel economy trends for light trucks. In both cases, miles per gallon has been relatively constant, probably reflecting the fact that the CAFE requirements have not changed. It is worth noting that horsepower/pound, or kilowatts/kilogram, has increased more rapidly for passenger cars than for light trucks, and weight has remained relatively constant for cars but increased slightly for light trucks. Figure 8 illustrates a different aspect of fuel economy, total fleet fuel economy. As shown in Fig. 8, the percentage of light trucks, which consume more gallons per mile (or liters per kilometer) than cars, has increased significantly in the past two decades. Consequently, the total fleet fuel economy

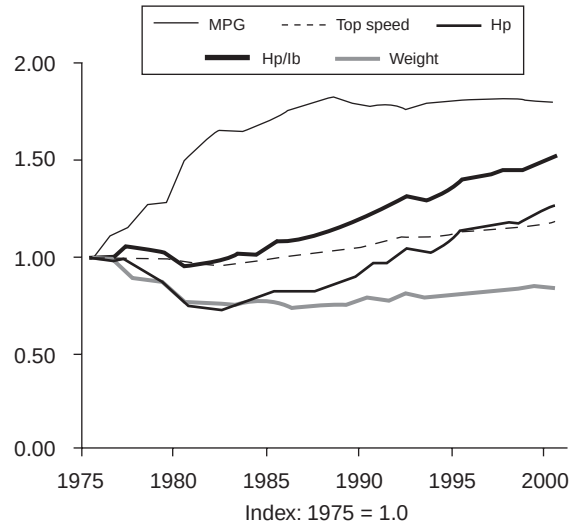


FIGURE 6 Trends in fuel economy-related characteristics of passenger cars, 1975–2000.

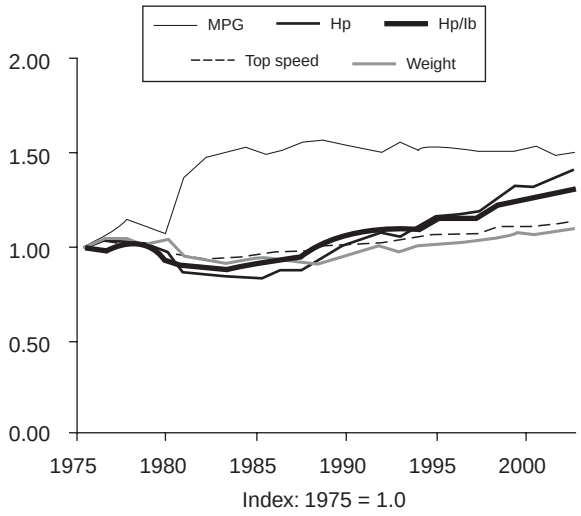


FIGURE 7 Trends in fuel economy-related characteristics of light trucks, 1975–2000.

(the "both" curve in Fig. 8) has decreased during the past decade.

4.4 Customer Preferences

Customer preferences, expressed at the time of purchase, depend on many factors and cover a range of costs. The time interval over which consumer preferences change varies from a few years to more than a decade. These preferences may be in contrast to general goals targeted by regulations. For example, consider vehicle volume. Figure 9 shows car interior volume for domestic (United States),

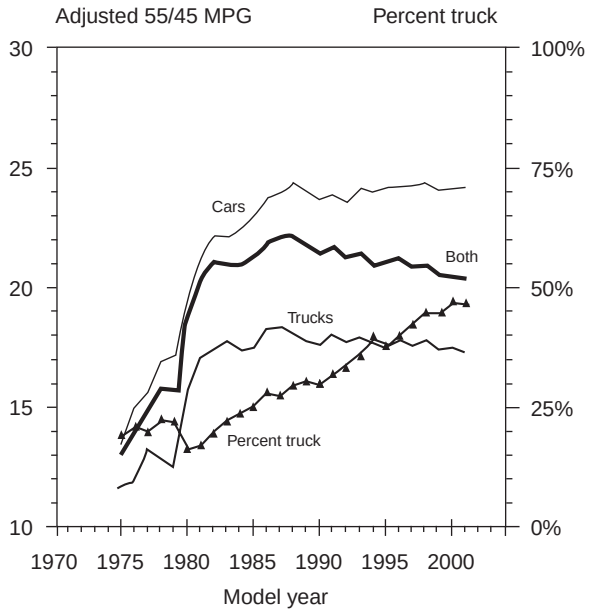


FIGURE 8 Miles per gallon versus model year of cars and light trucks and light truck percentage of sales.

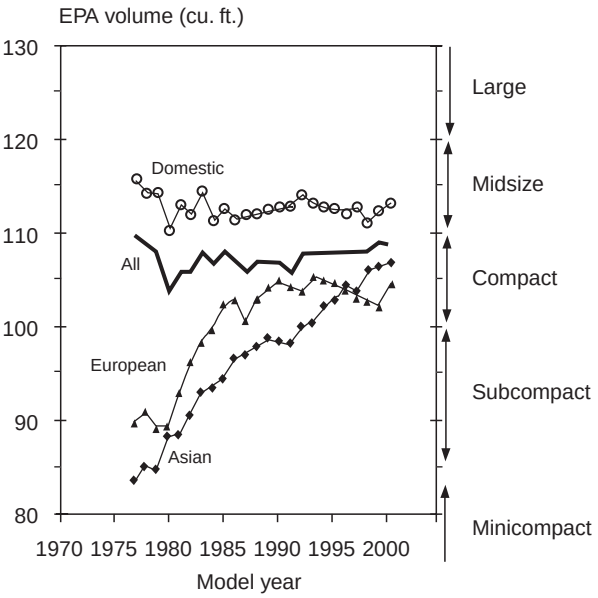


FIGURE 9 Car interior volume as a function of time.

European, Asian, and all as a function of model year. The volume of domestic cars remained essentially constant, whereas the volumes of European and Asian cars increased significantly. This increase reflects changes in vehicle sales. Initially, the foreign vehicles sold were compact or smaller vehicles; the progression in size simply reflects the strong U.S. market demand for high interior-volume vehicles. Another example is vehicle acceleration, typically

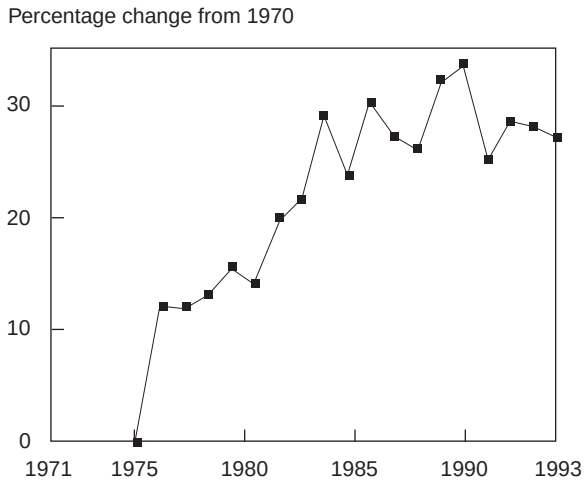


FIGURE 10 The increase in technical efficiency in new U.S. passenger cars, model years 1971–1993.

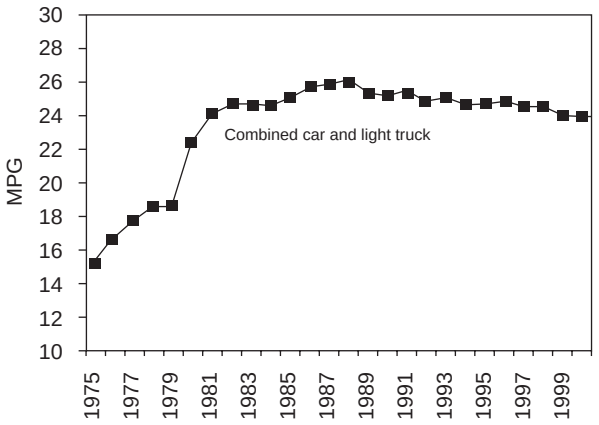


FIGURE 11 Fuel economy history of combined car and light truck.

expressed as the time, in seconds, to go from 0 to 60 miles per hour. High acceleration is of special interest when passing or merging. However, there is a trade-off between acceleration and fuel economy. There have been significant increases in technical fuel efficiency in new U.S. passenger cars (Fig. 10). However, this increase in technical efficiency has not necessarily translated into decreased fuel consumption.

To quote John German of Honda,

> *There is a misconception that vehicle manufacturers have not introduced fuel-efficient technology since the mid-1980s. This is understandable, as the car and light truck CAFE have remained constant for the past 15 years and the combined fleet has gone down due to increasing light truck market penetration as shown in Fig. 1 [Fig. 11]. However, there has been a substantial amount of fuel efficiency*

technology introduced in this time period. Some examples for the entire car and light truck fleet from EPA's 2000 "Fuel Economy Trends" are shown in Fig. 2 [Fig. 12]. However, this new technology has been employed more to respond to vehicle attributes demanded by the marketplace than to increase fuel economy. Over the past two decades consumers have insisted on such features as enhanced performance, luxury, utility, and safety, without decreasing fuel economy. Figure 3 [Fig. 13] shows the changes in vehicle weight, performance, and proportion of automatic transmissions since 1980 in the passenger car fleet. It is clear that technology has been used for vehicle attributes which consumers value more highly than fuel economy. Figure 4 [Fig. 14] compares the actual fuel economy for cars to what the fuel economy would have been if the technology were used solely for fuel economy instead of performance and other attributes.

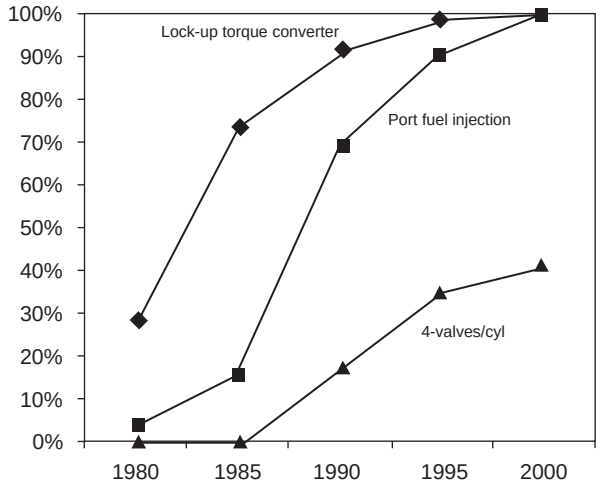


FIGURE 12 Examples of efficiency technology introduced since 1980.

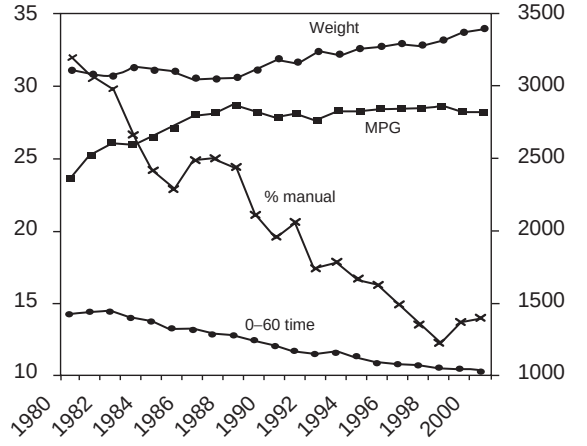


FIGURE 13 Changes in vehicle weight, fuel economy, proportion of automatic transmission, and acceleration time since 1980.

When considering vehicle trends, it is imperative to distinguish between individual and fleet trends. For example, Figs. 6 and 7 show vehicle weights remaining relatively constant with time. A different picture is obtained when the sales fraction of vehicles is included. Environmental Protection Agency data (Figs. 15 and 16) show a clear customer preference for cars and light trucks that are heavier. A similar picture is obtained when examining fuel economy; that is, a higher percentage of customers are purchasing new vehicles in the light truck category, resulting in a decrease in total fleet fuel economy even though the fuel economy standards for both cars and light trucks have remained constant.

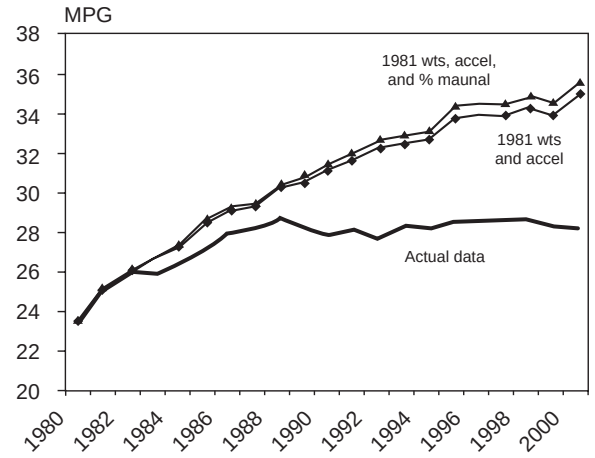


FIGURE 14 Fuel economy that could have been obtained if advanced technology had been used to increase fuel economy.

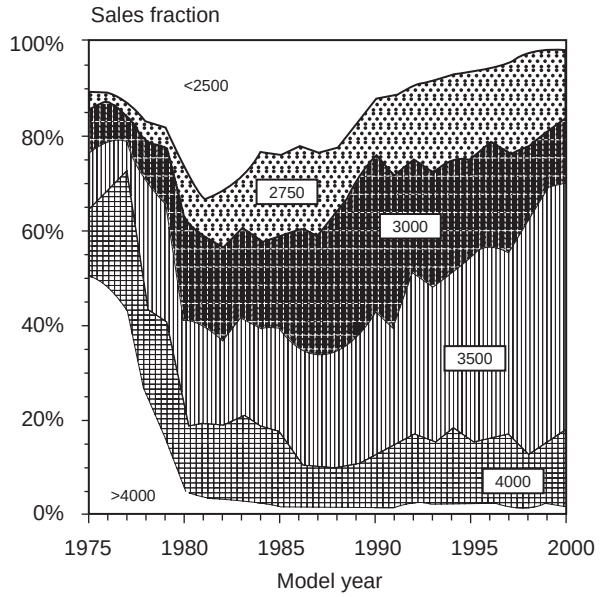


FIGURE 15 Changes in inertia weight class of cars with time.

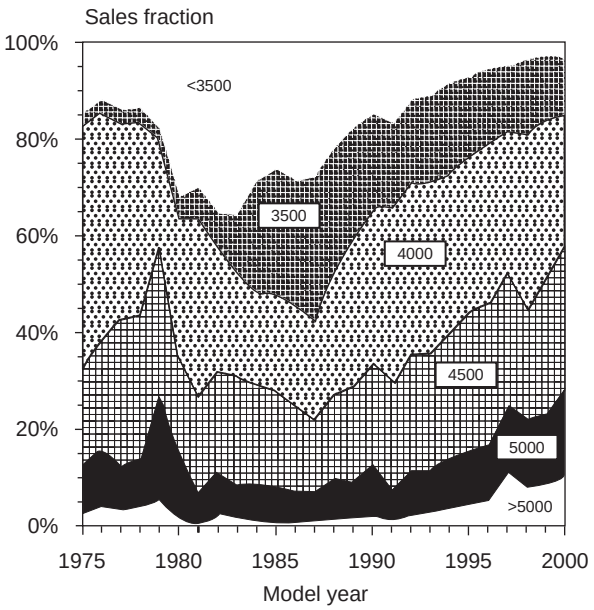


FIGURE 16 Changes in inertia weight class of light trucks with time.

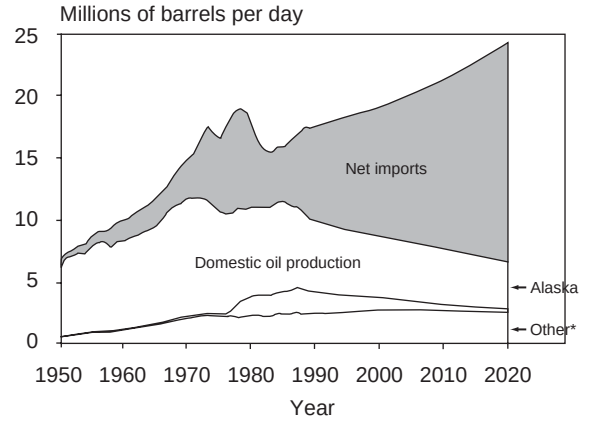


FIGURE 17 U.S. oil import trends. *Natural gas liquids and other.

4.5 Fuel Availability and Cost to Consumer

The United States has been the largest consumer of petroleum-based transportation. Prior to the past decade or two, the discovery of new petroleum resources in the United States kept pace with the increased usage. However, as shown in Fig. 17, an increasing amount of the petroleum used in the United States has been supplied by importation; it appears that this trend will increase with time. As shown in Fig. 18, the cost of gasoline to the consumer, expressed in constant dollars, has decreased with time. Note also that the average annual miles driven per car in the United States has increased slightly with time. Figure 19 suggests that there is a relationship between taxation rate and per capita consumption of petroleum.

5. ENERGY SOURCES FOR GROUND VEHICLES

Essentially all cars and light trucks have internal combustion engines using petroleum as their primary energy source. Optimum performance of a given type of internal combustion engine is achieved when fuel characteristics are optimized for that type of engine (e.g., diesel fuel, jet fuel, and gasoline).

All aspects of providing fuel to the vehicle must be considered. Because the fuel is necessarily combustible, fire and explosions during production, storage, distribution, and use are a common concern. Appropriate safeguards and distribution procedures must be followed. These are not without cost. In addition to the emissions from the vehicle, emissions during fuel procurement, modification, and use are also of concern. Consequently, to obtain a true picture of the global impact of using a particular fuel, engine, and vehicle combination, the "well-to-wheels" assessment, as described by Louis, should be used in all efficiency and emission analyses. For example, an electric vehicle is not truly a zero emission vehicle. Generating the electricity to charge the batteries is not without environmental impact.

A vehicle's range depends on both the overall operational efficiency of the vehicle and the amount of energy that can be stored onboard as fuel. To satisfy current consumer demands, energy density of the fuel used in the vehicle must be sufficiently high to yield a vehicle driving range in the hundreds of miles with an acceptable size and weight of the fuel storage tank. Figure 20 shows large differences in required tank volume and weight between different "energy carriers" (i.e., fuels). Although some progress has been made, Fig. 20 illustrates one of the reasons why vehicles using batteries for energy storage have had limited success in the marketplace. Hybrid vehicles, which accomplish battery recharging via engine and wheel generators, overcome this limitation and show significant improvement in fuel economy. They are in limited production. However, with today's gasoline prices in the United States, they are not cost-effective. As another example, solar energy falling on $1\,m^2$ is approximately equal to 1 horsepower (0.745 kW). Consequently, if solar energy is to be used extensively

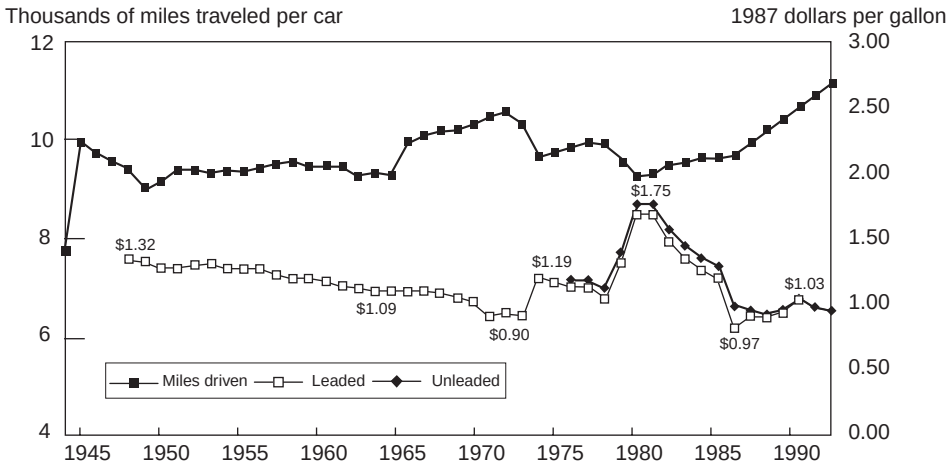


FIGURE 18 Real gasoline prices and annual miles driven per car, United States.

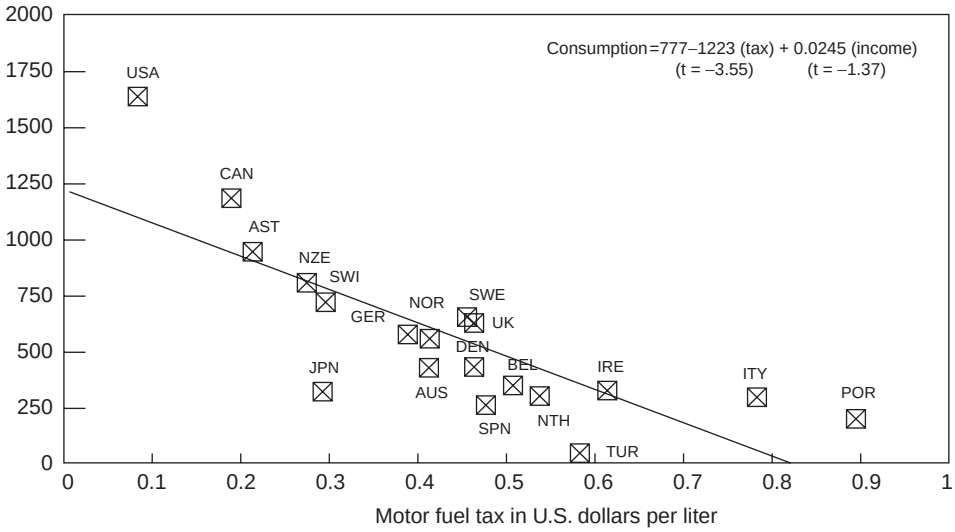


FIGURE 19 Relationship between motor fuel consumption and tax rates.

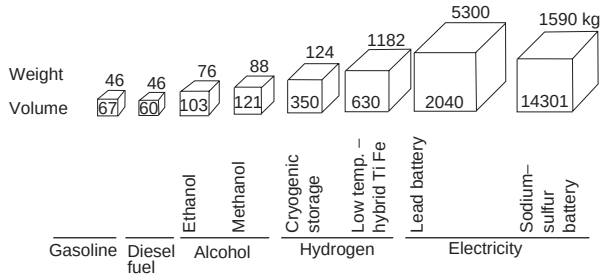


FIGURE 20 Tank volume and weight required for various energy carriers based on 55 liters of gasoline. Source: Sieffert and Walzer (1990).

for transportation, energy conversion and storage will be essential elements.

Increasing concern about greenhouse gas emissions may result in additional legislation affecting fuel economy. For example, California passed a law similar to other exhaust emission restrictions, limiting the permissible output of greenhouse gases per unit distance. Decreases in fuel consumption are the most obvious way to meet such a regulation. Whether the law will stand up in court is not known. However, the implication of the regulation in question is clear: Reduce CO_2 emissions via reduced fuel consumption.

6. PROGNOSIS FOR THE FUTURE

6.1 Maximum Fuel Energy Recoverable as Work

As indicated previously, when using an internal combustion engine the maximum fuel energy that

can be theoretically converted to work is the Gibbs free energy, which for hydrocarbon fuels is approximately equal to heating value of the fuel. In today's transportation market, the only engines used, or proposed for use, are internal combustion engines. Consequently, the thermodynamic efficiency of all transportation engines is computed as the ratio of the work obtained from the engine to the energy content of the fuel consumed to obtain that work, which is the heating value of the fuel: It has a maximum theoretical value of 100%. This is true for diesels, spark ignition engines, fuel cells, humans, animals, etc. Irreversibilities that naturally occur in all real processes decrease the thermodynamic efficiency actually obtained.

6.2 Where Fuel Energy Is Used, Reductions and Recovery

Figure 21 illustrates that fuel consumption is a result of the system efficiency and road load. The system efficiency is determined by the power plant and the driveline, which consists of the transmission and the final drive gear. Fuel is consumed and power is produced in an internal combustion engine (gasoline, diesel, fuel cell, etc.). A transmission (mechanical, electrical, hydraulic, etc.) and final drive are used, insofar as possible, to match the operation of the power plant to its optimum efficiency for each of the widely varying power requirements demanded during the driving cycle. In addition, power is also needed for consumer-demanded amenities. The road load of the vehicle includes aerodynamic drag, rolling resistance, and weight resulting in inertia. A vehicle must be moving to experience a road load. However, system power can be expended whether or not the vehicle is moving via the transmission, generator/alternator, water pump, amenities, etc. The magnitude of these expenditures is shown in Fig. 22.

Fuel consumption is determined by the integrated operation of the total system. It should be emphasized that the components of the system are interactive. How the functions of these components are achieved in future vehicles may be impacted by new technologies. The subsequent impact on fuel consumption will be dependent on how the interaction of the various components is integrated, the extent to which consumer desires continue along current trends, and how manufacturers choose to implement the capabilities of the new technology. To make projections of potential fuel economy gains requires assessing these trends and will be subject to significant engineering judgment. Comparisons of more detailed estimations are presented later.

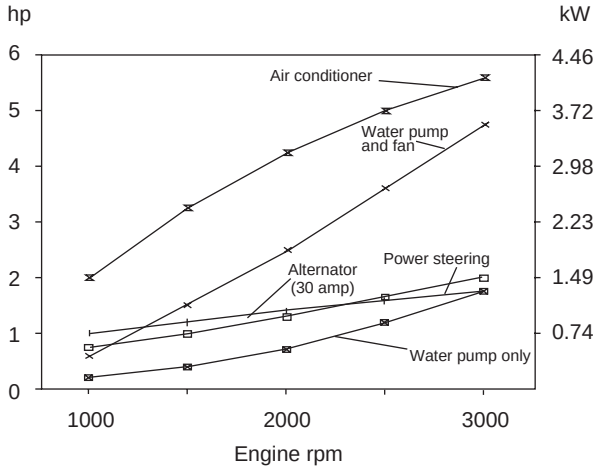


FIGURE 22 Accessory power requirements (full-size sedan).

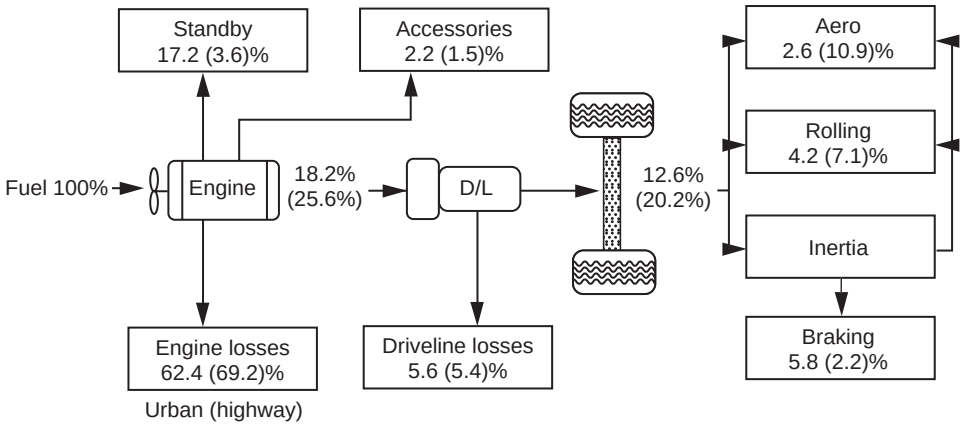


FIGURE 21 Energy flow diagram for a midsize car. Source: DOE (1994).

6.2.1 Power Plant

The power plant (engine) in Fig. 21 converts fuel energy into useful work to power the vehicle and customer-desired amenities. In a reciprocating engine, the work delivered to the top of the piston is called indicated work. The work ultimately delivered to drive the wheels is less than the indicated work because work is required to pump gases (air, fuel, and exhaust) into and out of the engine, overcome the friction of the engine, drive components, and drive amenities (Fig. 22). Figure 23 shows that for gasoline engines the maximum brake thermal efficiency achieved has increased with time. However, engines used in transportation vehicles necessarily work over a wide range of loads and speeds, with subsequent variations in brake thermal efficiency. This can be seen in an engine map as shown in Fig. 24. Diesel engines typically have higher maximum brake thermal efficiencies and different speed-torque maps than do spark ignition engines. Figure 25 shows the relationship between the equivalent of brake thermal efficiency and load for a fuel cell. Notice that the efficiency of the fuel cell declines significantly as the load increases. This is opposite the behavior of a typical spark ignition engine, whose efficiency increases with load.

During operation of the engine, fuel consumption can be decreased by increasing indicated efficiency or decreasing ancillary losses, such as rolling resistance, aerodynamic drag, or accessory power consumption. Fuel consumption can also be decreased at constant brake thermal efficiency. For example, essentially instantaneous engine starting is being introduced. This permits automatically turning off and restarting the engine during stops, coasting, etc. This does not change the efficiency of the engine when it is operating, but it does decrease fuel consumed during the driving cycle, and it also results in lower values of integrated emissions.

One of the reasons the indicated efficiency of the diesel engine is higher than that of the spark ignition engine is because its pumping losses are lower. Variable valve timing in spark ignition engines may be able to offset some of these pumping losses. Valve timing can be varied to increase indicated efficiency in a variety of ways. The engine can be run as a variable displacement engine (displacement on demand). All cylinders are active (large displacement) for high load operation, whereas some cylinders are deactivated (smaller displacement) for lighter load operation.

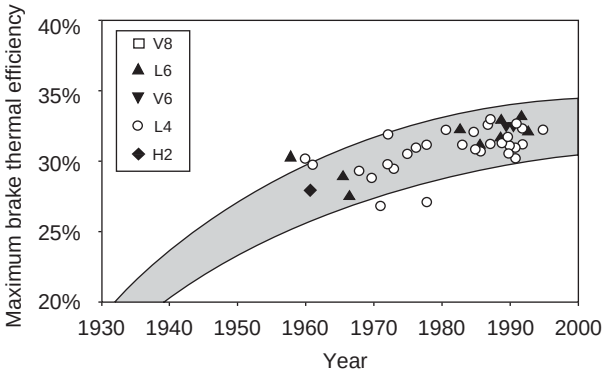


FIGURE 23 Brake thermal efficiency of gasoline engines.

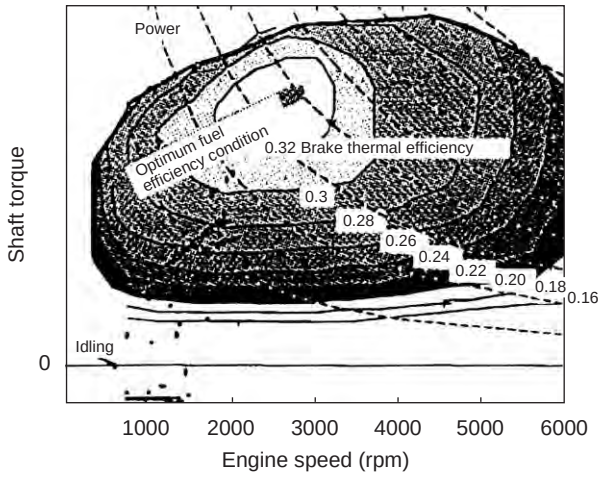


FIGURE 24 Engine brake thermal efficiency map for gasoline engines.

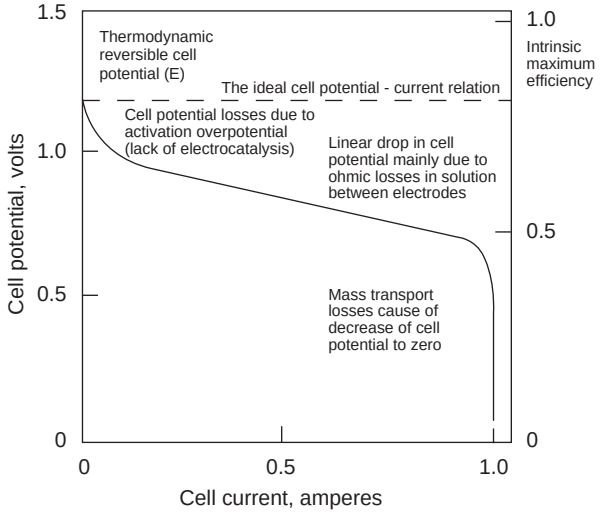


FIGURE 25 Typical plot of cell potential vs current for fuel cells.

Improved surface finishing of engine parts permits use of lower viscosity lubricants to reduce engine friction. Ringless carbon pistons are claimed to have lower weight, a lower coefficient of thermal expansion, and reduced friction. However, the introduction of new technologies can result in some detriments as well as benefits. For example, the more sophisticated valve mechanisms, which allow variable displacement operation, may result in higher frictional loads and more weight than the simpler valve trains they replace. This increased friction and weight would be a detriment and would have to be evaluated against the benefit achieved. As mentioned previously, trends in consumer demand also play a role. Accessory loads have continued to increase with time: We are demanding more amenities in our vehicles.

6.2.2 Transmission

The transmission (driveline) in Fig. 21 is intended to permit the engine to operate at a speed that will give minimum fuel consumption while still delivering the desired torque. This is accomplished by "stepping" through different ratios of engine speed to wheel speed (i.e., permitting the engine to operate as closely as possible to the minimum fuel consumption points on the engine map; Fig. 24). A large number of transmission steps provides the flexibility for a better match (i.e., lower fuel consumption) but may result in undesirable shifting patterns when the load is changed suddenly. Today's transmissions typically have three to five steps. The efficiency of manual shifting transmissions is quite high but requires active participation of the driver in determining shift points and therefore performance. The inclusion of a torque converter and automatic shifting improves consumer-demanded performance but may increase fuel consumption slightly. Continuously variable transmissions (CVTs) have, in concept, an infinite number of steps, but current designs tend to be power limited. Figure 26 illustrates how a CVT permits engine operation to minimize fuel consumption.

6.2.3 Aerodynamic Drag and Rolling Resistance

The road load is the power required to move the vehicle down the road. Figure 27 shows a representative proportioning of the road load between aerodynamic drag, rolling resistance, and inertia for both city and highway driving. The aerodynamic drag and rolling resistance portions of the road load increase with speed. Figure 28 shows the energy from the fuel needed to overcome aerodynamic drag and rolling resistance. This energy is dissipated and not

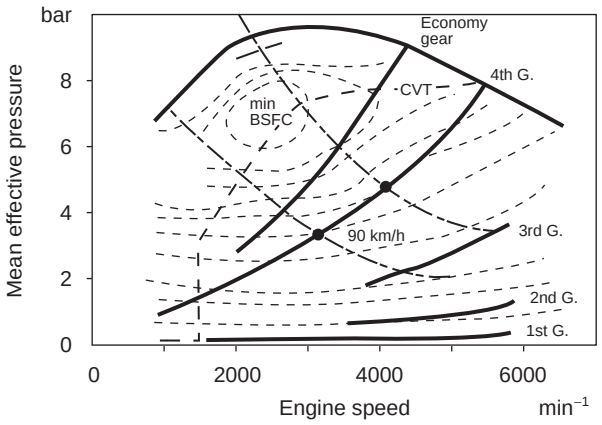


FIGURE 26 Engine map showing the line for a continuously variable transmission.

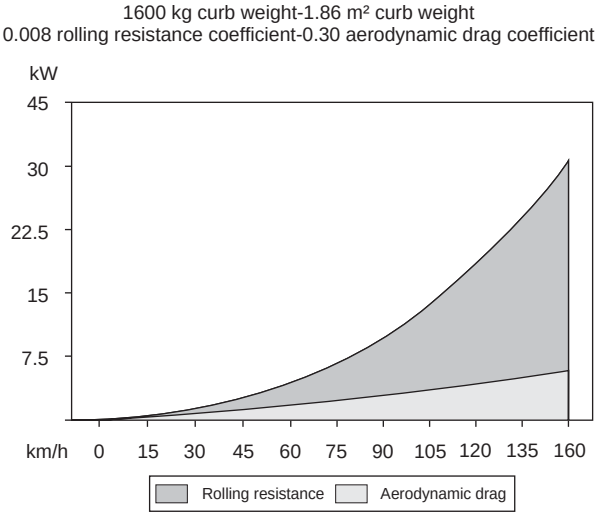


FIGURE 27 Steady-state road load of a full-size car.

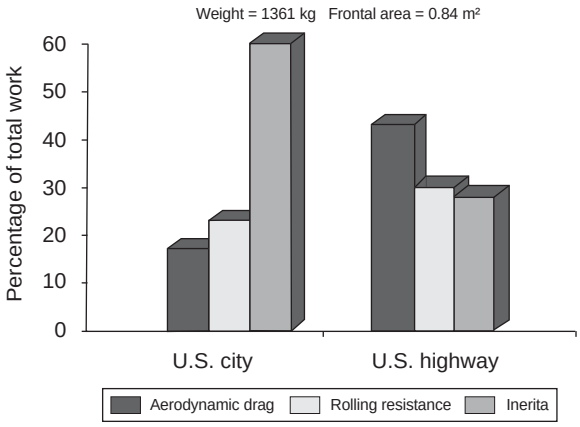


FIGURE 28 Driving cycle load comparisons.

recoverable as work. Consequently, efforts must be made to minimize aerodynamic drag and rolling resistance. Throughout the years, the rolling resistance of tires has been reduced, but there are tradeoffs: Wear, traction, ride, etc. may be compromised. Aerodynamic drag, which varies as the square of the speed, has been reduced over time in some vehicles (Fig. 29). However, in advanced vehicles aerodynamic drag seems to be close to minimum values. Considering the current price of gasoline, and that a significant fraction of public driving is at low to moderate speeds, the interaction between design characteristics of minimum drag vehicles and other desired consumer characteristics such as interior volume provides little incentive to reduce drag further.

6.2.4 Inertia Weight

Additional fuel is consumed when the kinetic or potential energy of the vehicle is increased. This occurs when accelerating or climbing a hill. Assuming reversible storage and recovery, this energy could theoretically be completely recovered and stored for future use. This would be accomplished by converting the vehicle's potential or kinetic energy into electrical energy and storing the electricity in a battery, for example. This is one of the ways in which fuel consumption is reduced in a hybrid car.

6.2.5 Onboard Energy Storage

Gasoline and diesel fuels have very high-energy storage densities (Fig. 20). For electric energy, the energy storage density is highest for advanced batteries but still not comparable with gasoline (Fig. 20). Electric hybrid vehicles try to overcome storage density limitation by retaining gasoline, or diesel fuel, as the primary energy storage media. Electricity generation, with temporary storage, occurs on the vehicle along with recapturing the vehicle's kinetic and potential energy through regenerative braking. However, the rate of charging required to capture the electricity that is being produced by regenerative braking often exceeds current battery capabilities. The noncaptured energy must be dissipated by the brakes and is lost. Exchange efficiencies are low for batteries (approximately 50%). Hydraulic storage exchange efficiencies are higher (85–90%), but energy storage densities are as much as an order of magnitude less than for batteries.

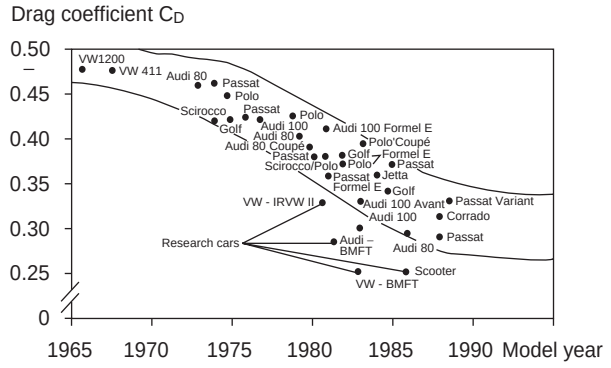


FIGURE 29 Drag coefficient improvements made over the years (VW–Audi models).

There is potential for a safety problem with hybrid vehicles, particularly if the power derived from the temporary energy storage is a large fraction of the total engine plus stored-energy power system. For example, performance in critical situations such as passing on a long hill, when the temporary energy storage is nearly depleted, could be a safety problem.

6.2.6 Estimates of the Costs of Reducing Fuel Consumption

Integrating new technologies into a vehicle to reduce fuel consumption must be cost-effective and acceptable to consumers. As seen in Fig. 21, the propulsion system is complex and must meet a number of constraints. It must operate over a driving cycle in which vehicle speed and load vary widely; meet mandated, increasingly restrictive emission and safety standards; and meet customer needs and desires (fads), etc. The ideal first step in choosing potential new technologies would be to mathematically estimate reductions in fuel consumption using these technologies when the vehicle is "driven" over the appropriate test cycle and then estimate the cost of the new or revised propulsion system. However, performance of new system components is almost inevitably not well defined, especially when projecting into the future. Production numbers and costs estimates for new systems are estimates. Consequently, it is not surprising that fuel economy and cost estimates for proposed future vehicles differ widely.

The ideal, but impractical, approach to predicting potential fuel economy gains would be to construct a complete, detailed mathematical model of the vehicle and its proposed propulsion system and drive the vehicle over the test cycle. A less precise method is to use a multiplicative technology menu approach. The literature reports many such approaches, and the results obtained span a range of predicted fuel consumption reductions. Most studies conclude that there is a clear technical potential to significantly reduce fuel consumption of light-duty cars and trucks.

TABLE I

Comparison of Performance Characteristics between the Aluminum Cow and a Stock Taurus

	Aluminum cow	Stock taurus
Fuel economy (mpg)	65	25
Weight (lbs)	2800	3400
1/8-mile acceleration event (s)	10.7	11.2

As further evidence of the technical potential to reduce fuel consumption, consider the results of a competition. In 1998 and 1999, the Society of Automotive Engineers sponsored a student competition to achieve low fuel consumption. A team of student engineers assembled the "aluminum cow" using a 1994 Ford Taurus aluminum body and a diesel–electric hybrid. Table I lists comparative performance characteristics between the aluminum cow and a stock Taurus.

However, as discussed previously, the manufacturer of a vehicle must take into account many other factors that were not part of the competition, such as incremental cost of the vehicle, environmental impact of its manufacture, sales appeal, and safety.

7. CONSIDERATIONS AND PROBLEMS IN REDUCING FUEL CONSUMPTION

The inherent advantages of current internal combustion engines (diesel, spark ignition, or fuel cell), make these engines preferable energy converters for personal mobility purposes. It is technically possible to achieve significant improvements in the fuel economy of cars and light trucks using internal combustion engines. However, there is no simple or definitive path to be followed to get more fuel-efficient vehicles into the fleet. Figure 30 illustrates the many considerations that vehicle manufacturers must take into account if for a vehicle to be marketable.

The diesel engine is a specific example of these complex considerations. A modern diesel engine is indistinguishable from a gasoline engine either by sound or exhaust color and visibility. It has significantly lower fuel consumption but higher initial cost than a comparable gasoline engine. Because of the inherent differences between the two engines, they have different exhaust emission characteristics, which present different challenges in meeting regulated emission standards. In Europe, where fuel prices are two or three times higher than

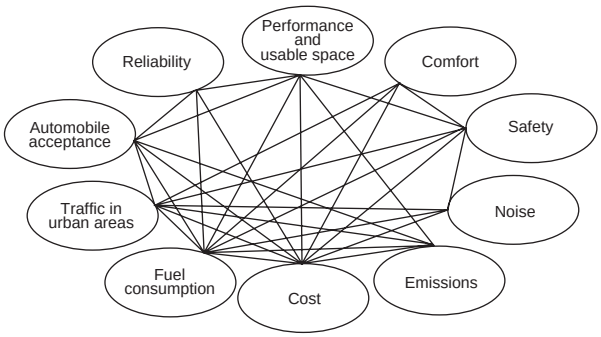


FIGURE 30 Interrelationships between requirements for an automobile.

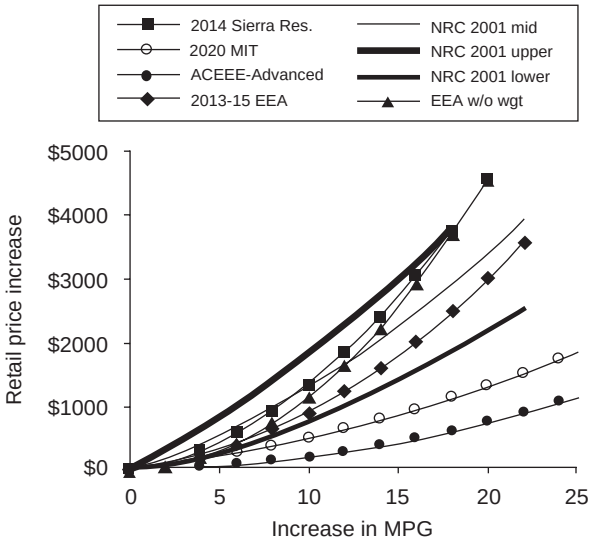


FIGURE 31 Passenger car fuel economy cost curves from selected studies.

the U.S. price, approximately half of new cars sold use diesel engines. Diesel sales in the United States are essentially zero. Proposed U.S. 2007 automotive particulate (smoke) regulations are significantly more restrictive than those for the same time period in Europe. This would mandate inclusion of even more advanced, and expensive, technology for diesel engines to be able to meet the emissions regulations. When fuel is inexpensive, the increased cost to the consumer of the more technologically advanced vehicle may exceed the savings from reduced fuel consumption. Figures 31 and 32 show the break-even price increase, obtained from different studies for passenger cars and light trucks, that would be acceptable through specific increases in miles per gallon. Finally, assuming the vehicle is marketable, it will still take a decade or two for the new technology to penetrate the fleet to the extent that it will influence fleet fuel economy (Fig. 33).

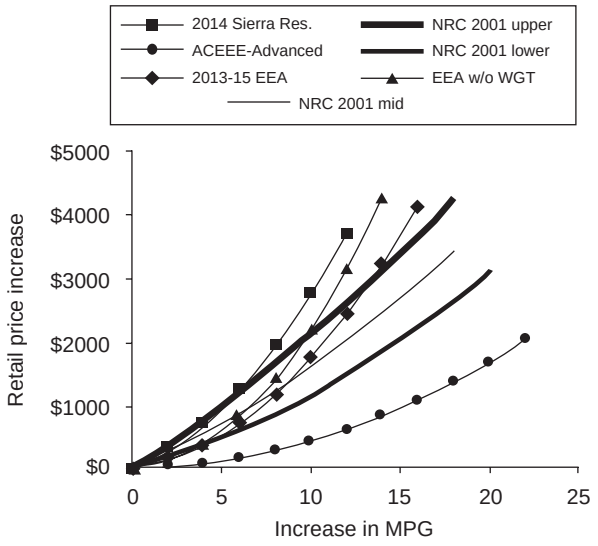


FIGURE 32 Light truck fuel economy cost curves from selected sources.

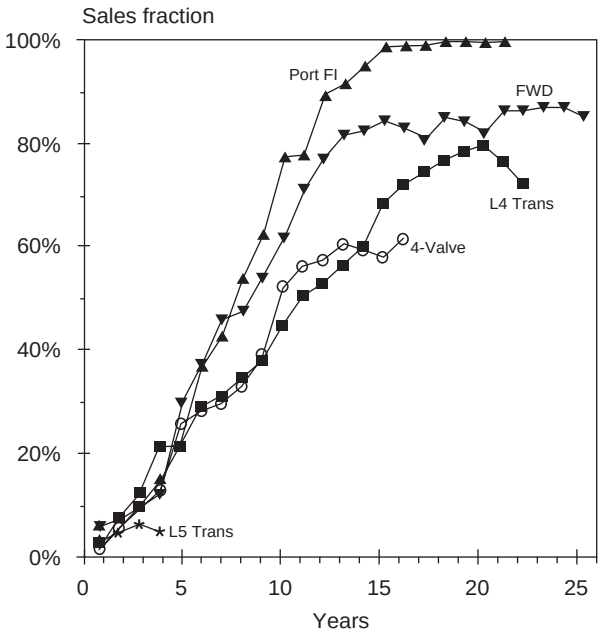


FIGURE 33 Car technology penetration: Years after first significant use.

There is no clear-cut and simple approach to achieving lower fuel consumption and more environmentally benign vehicles. However, there are many practical realities that cannot be dismissed. Because of their high energy density and ease of processing, transport, and storage, petroleum-based fuels, or more generally hydrocarbon fuels, are very well suited for mobility power plants. Potential alternative fuels or power plants all have significant shortcomings that need to be overcome. There are many pending issues. As shown in Fig. 17, the bulk of petroleum used in the United States is imported, and the percentage is increasing. Also, annual per capita travel is increasing with time. The current cost of petroleum in the United States does not explicitly include the costs of ensuring access to petroleum reserves outside of the United States or the impact of its use on health and the environment. Established petroleum reserves are typically discussed in terms of tens to hundreds of years, not thousands. The rapidly growing demand for petroleum by nations such as China and India will put additional strain on petroleum reserves. Increased use of petroleum because of an increased number of vehicles and increased demand for mobility will increase the emission of greenhouse gases into the atmosphere as well as the load on petroleum reserves.

When attempting to decrease the fuel consumption of the light-duty car and truck fleet, there are many problems to be solved. At the current price of gasoline in the United States, fuel economy is simply not a high priority for a prospective purchaser of a vehicle. Taxes on the fuel used by cars and light-duty trucks in the United States are relatively low in comparison to those of the rest of the industrial world (Fig. 19). The current U.S. fuel consumption control system, CAFE, is antiquated and works on the supply side. Consider that in the past decade U.S. fleet fuel economy has decreased significantly even though CAFE standards have remained constant. The European Union system for moderating consumption (increased taxation) works on the demand side and aligns consumer choice and use patterns with public policy objective, as evidenced by the increasing percentage of low fuel consumption vehicle technologies being introduced into the European market.

If major geopolitical catastrophes can be avoided, the increasing world population along with our increasing demand for personal mobility portend a necessary reduction in per capita fuel consumption. This will be necessary to preserve natural resources and to address the increased stress that our activities are placing on the ecosystem. This much is clear. The path by which we achieve this goal is not.

SEE ALSO THE FOLLOWING ARTICLES

Alternative Transportation Fuels: Contemporary Case Studies • Bicycling • Consumption, Energy, and the Environment • Fuel Cycle Analysis of

Conventional and Alternative Fuel Vehicles • Fuel Economy Initiatives: International Comparisons • Internal Combustion Engine Vehicles • Motor Vehicle Use, Social Costs of • Obstacles to Energy Efficiency • Passenger Demand for Travel and Energy Use • Transportation and Energy, Overview • Transportation and Energy Policy

Further Reading

DOE, USCAR, PNGV (1994, July). "Program Plan," Figure 5-1.

German, J. (2002, January 24). Statement before the Committee on Commerce, Science and Transportation, U.S. Senate.

Lauck, F., Uyehara, O. A., and Myers, P. S. (1963). "An Engineering Evaluation of Energy Conversion Devices, 1963 Transactions." Society of Automotive Engineers, Warrendale, PA.

Louis, J. J. J. (2001). Well-to-wheel energy use and greenhouse gas emissions, Paper No. 2001-01-1343. Society of Automotive Engineers, Warrendale, PA.

Morimitus, N. (1998, April). Fuel efficiency improvement of gasoline engine vehicle. *Toyota Tech. Rev.* **147**(2).

National Research Council. (1992). "Automotive Fuel Economy—How Far Should We Go?" National Academy Press, Washington, DC.

National Research Council. (2002). "Effectiveness, Impact of Corporate Average Fuel Economy (CAFE) Standards." National Academy Press, Washington, DC.

Norbeck, J. M., Heffel, J. W., Durbin, T. D., Tabbara, B., Bowden, J. M., and Montano, M. C. (1996). "Hydrogen Fuel for Surface Transportation." Society of Automotive Engineers, Warrendale, PA.

Novola, P. S., and Crandall, R. S. "The Extra Mile—Rethinking Energy Policy for Automotive Transportation." Brookings Institute, Washington, DC.

Riley, R. Q. (1990). "Alternate Cars in the 21st Century." Society of Automotive Engineers, Warrendale, PA.

Seiffert, U., and Walzer, P. (1990). "Automobile Technology of the Future." Society of Automotive Engineers, Warrendale, PA.

Sovran, G., and Bohn, M. S. (1981). Formulae for the tractive-energy requirements of vehicles driving the EPA schedules, Paper No. 81084. Society of Automotive Engineers, Warrendale, PA.

Suzuki, T. (1997). "The Romance of Engines." Society of Automotive Engineers, Warrendale, PA.

Weiss, M. A., Heywood, J. B., Drake, E. M., Schafer, A., Felix, F., and Au Yeung, F. F. (2000). "On the Road in 2020: A Life Cycle Analysis of New Automobile Technologies." MIT Press, Cambridge, MA.

War and Energy

VACLAV SMIL
University of Manitoba
Winnipeg, Manitoba, Canada

1. A Brief History of Weapons
2. Energy Costs and Consequences of War
3. Energy as the Cause of War

Glossary

explosives Compounds whose destructive effect results from virtually instantaneous release of kinetic energy that is generated either by internally oxidized chemical reactions that produce large and rapidly expanding volumes of gas (chemical explosives) or by fission of heavy nuclei or fusion of light nuclei (nuclear explosives).

intercontinental ballistic missiles (ICBMs) Missiles that can be launched from fortified silos or nuclear submarines and can carry a single warhead or multiple independently targeted reentry vehicles (MIRVs).

military spending A category of expenses whose definition varies widely among countries and is often greatly undervalued in official accounts; for example, official Soviet military spending included merely operation and maintenance costs, and the real total (including weapons research and development and production) was roughly an order of magnitude higher.

nuclear weapons Both strategic and tactical warheads, bombs, and munitions, carried by intercontinental ballistic and medium- and short-range missiles or delivered by bombers or field artillery, whose destructive power is released either through nuclear fission or, in thermonuclear weapons, through fusion.

Wars demand an extraordinary mobilization of energy resources, they represent the most concentrated and the most devastating release of destructive power, and their common consequence is a major disruption of energy supplies in regions or countries that were affected by combat or subjected to prolonged bombing. Given these obvious realities, it is inexplicable that wars have received very little attention as energy phenomena. At the same time, there is a fairly common perception—one that has been greatly reinforced by the American conquest of Iraq in 2003—that energy is often the main reason why nations go to war. This article addresses all of these concerns. Because the destructiveness of war depends largely on the weapons used, it first outlines their brief history. The energy cost of individual armed conflicts, as well as the peacetime energy cost of preparing for war, is difficult to assess, and the article presents some representative calculations and cites some recent statistics using the major 20th-century wars waged by the United States. Casualties are the most horrifying consequence of wars, and the greater destructive power led to their increase until the recent development of precision-guided munitions made it possible to minimize civilian deaths by careful targeting. Impacts of war on energy use during the conflict and in its aftermath, particularly in defeated countries that suffered a great deal of destruction, are clearly seen in consumption statistics. Finally, the article argues that the wish to be in actual physical control of energy resources has not been the sole reason, or even the primary reason, for any major modern armed conflict.

1. A BRIEF HISTORY OF WEAPONS

Weapons are the prime movers of war. They are designed to inflict damage through a sudden release of kinetic energy (all handheld weapons, projectiles, and explosives), heat, or a combination of both. Nuclear weapons kill nearly instantaneously by the combination of blast and thermal radiation, and they also cause delayed deaths and sickness due to the exposure to ionizing radiation. Classification of wars based on these destructive prime movers divides history into four distinct periods. All prehistoric, classical, and early medieval warfare was powered only by human and animal muscles. Invention of

 © 2004 Elsevier Inc. All rights reserved.

gunpowder led to a rapid diffusion of initially clumsy front- and breach-loading rifles and to much more powerful field and ship guns.

By the late 1860s Nobel's combination of nitroglycerine and diatomaceous earth produced the first practical high explosive, and other variants were soon made into munitions for newly invented machine guns as well as into large-caliber gun shells and airplane bombs. Wars of the 20th century were dominated by the use of these high explosives delivered in shells (from land and ship-borne artillery and tanks), torpedoes, bombs, and missiles. Finally, strategic thinking and the global balance of power changed after World War II with the development of nuclear weapons.

The period of muscle-powered warfare lasted nearly until the very end of the Middle Ages. The daily range of advancing armies was limited less by the stamina of walking or running troops than by the speed of their supply trains, made up of animals ranging from oxen and horses in Europe to camels and elephants in Asia. Men used axes, daggers, swords, and lances in close combat and used spears and bows and arrows for attacking unprotected enemies as far as 200 m away. Much more powerful crossbows were used since the fourth century BC in both Greece and China. Inaccurate targeting and relatively low kinetic energy of these assaults (Table I) limited the magnitude and frequency of injuries that could be inflicted by these weapons. Massed human power was also used to wind winches of catapults and to move assorted siege machines designed to overcome urban fortifications.

TABLE I

Kinetic Energy of Projectiles

Weapon	Projectile	Kinetic energy (J)
Bow and arrow	Arrow	20
Heavy crossbow	Arrow	100
Civil war musket	Bullet	1×10^3
Assault rifle (M16)	Bullet	2×10^3
Medieval cannon	Stone ball	50×10^3
18th-century cannon	Iron ball	300×10^3
World War I artillery gun	Shrapnell shell	1×10^6
World War II heavy AA gun	High–explosive shell	6×10^6
M1A1 Abrams tank	Depleted U shell	6×10^6
Unguided World War II rocket	Missile with payload	18×10^6
Boeing 767 (September 11, 2001)	Hijacked plane	4×10^9

Light horses were initially used to pull chariots, and only after the general adoption of stirrups (beginning in the third century AD) did mounted warriors become a particularly effective fighting force. Asian riders were unarmored, but they had small and extraordinarily hardy horses and could move with high speed and maneuverability. These abilities brought the Mongol invaders from the heart of Asia to the center of Europe between 1223 and 1241. In contrast, European knights relied on their heavy armor as they rode increasingly heavier animals. Their most spectacular series of long-distance forays brought them as Crusaders from many countries of Europe to the Eastern Mediterranean, where they established temporary rule (between 1096 and 1291) over fluctuating areas of the littoral. The only important nonanimate energies used in the pre-gunpowder era were various incendiary materials (sulfur, asphalt, petroleum, and quicklime were used in their preparation) that were either fastened to arrowheads or hurled across moats and walls from catapults.

Gunpowder's origins can be traced to the experiments of medieval Chinese alchemists and metallurgists. Clear directions for preparing gunpowder were published in 1040, and eventually the proportions for its mixing settled at approximately 75% saltpeter (KNO_3), 15% charcoal, and 10% sulfur. Whereas ordinary combustion must draw oxygen from the surrounding air, the ignited potassium nitrate provides it internally and gunpowder undergoes a rapid expansion equal to roughly 3000 times its volume in gas. The first true guns were cast in China before the end of the 13th century, and Europe was just a few decades behind. Gunpowder raised the destructiveness of weapons and radically changed the conduct of both land and maritime battles.

When confined and directed in barrels of rifles, gunpowder could impart to bullets kinetic energy an order of magnitude higher than that of a heavy arrow shot from a crossbow gun, and larger charges in field artillery and ship guns could propel even heavy projectiles (Table I). Increasingly accurate gunfire from far beyond the range of archers eliminated the defensive value of moats and walls and did away with lengthy sieges of cities and castles. As a new defense, the fortified structures were built as low spreading polygons with massive earthen embankments and huge water ditches. French military engineer Sebastien Vauban became the most famous designer of these fortifications that embodied large

amounts of energy. His largest project, at Longwy in northeastern France, required moving 640,000 m^3 of rock and earth (volume equivalent to about a quarter of Khufu's pyramid) and emplacing 120,000 m^3 of masonry.

The impact of guns was even greater in maritime engagements. Detailed historical accounts document how gunned ships (equipped with two other Chinese innovations, compass and good rudders, as well as with better sails) became the carriers of European technical superiority. By the beginning of the 16th century, they were the principal means of global empire building by the nations of the Atlantic Europe. By the late 17th century, the largest "men-of-war" carried up to 100 guns each. Dominance of these ships did not end until the introduction of naval steam engines during the 19th century.

The next weapons era began with the formulation of high explosives, which are prepared by the nitration of organic compounds such as cellulose, glycerine, phenol, and toluene. Ascanio Sobrero prepared nitroglycerin in 1846, but its practical use did not begin until Nobel mixed it with an inert porous substance (diatomaceous earth) to create dynamite and introduced a practical detonator (Nobel igniter) without which the new explosive would be nearly impossible to use. A single comparison conveys the explosive power of dynamite: its velocity of detonation is as much as 6800 m/s compared with less than 400 m/s for gunpowder (Table II). Slower acting, and preferably smokeless, propellants needed for weapons were produced during the 1880s: *poudre B* (gelatinized and extruded nitrocellulose) by Paul Vieille in 1884, Nobel's own ballistite (using nitroglycerine instead of ether and alcohol) in 1887, and cordite, patented in England by Frederick Abel and James Dewar in 1889.

Ammonium picrate was prepared in 1886. Trinitrotoluene (TNT), which was synthesized by Joseph Wilbrand in 1863 and which must be detonated by a high-velocity initiator, was used as an explosive by the end of the 19th century. The most powerful of all prenuclear explosives, cyclonite (cyclotrimethylenetrinitramine, commonly known as RDX [royal demolition explosive] and now a favorite of some terrorists), was first made by Hans Henning in 1899 by treating a formaldehyde derivative with nitric acid (Table II). Better propellants and inexpensive, high-quality steels increased the power and range of field of naval guns from less than 2 km during the 1860s to more than 30 km by 1900. The combination of long-range guns, heavy armor, and steam turbines (a superior new prime mover invented by Charles Parsons during the 1880s that made much faster speeds possible) resulted in new heavy battleships, with *Dreadnought,* launched in 1906, being their prototype.

TABLE II

Detonation Velocity of Common Explosives

Explosive	Density (g/cm^3)	Detonation velocity (m/s)
Gunpowder	1.0	1350
Dynamite	1.6	5000
TNT	1.6	6700
RDX	1.8	8800
ANFO	1.0	3200

Other new highly destructive weapons whose use contributed to the unprecedented casualties of World War I included machine guns, tanks, submarines, and the first military planes (e.g., light bombers). The two decades between the world wars brought rapid development of battle tanks, fighter planes, and long-range bombers and aircraft carriers, all of which were the decisive weapons of World War II. The German defeat of France in 1940 and advances of Wehrmacht in Russia in 1941 and 1942 were made possible by rapid tank-led penetrations. Japan's surprising assault on Pearl Harbor on December 7, 1941, would have been impossible without a large carrier force that launched fighter planes and dive bombers. The same types of weapons—Soviet T-42 tanks driving all the way to Berlin and American dive bombers deployed by the U.S. Navy in the Pacific—eventually led to the defeat of the Axis.

The closing months of World War II saw the deployment of two new prime movers and of an entirely new class of weapons. Gas turbines, independently invented during the 1930s by Frank Whittle and Hans Pabst von Ohain in the United Kingdom and Germany, respectively, were installed in the first jet fighters (the British Gloster Meteor and the German Messerschmitt 262) in 1944. During that same year, rocket engines were used in the German ballistic missile V-2 to terrorize England. And the first fission bomb was tested at Alamogordo, New Mexico, on July 11, 1945, with the second one destroying Hiroshima, Japan, on August 6, 1945, and the third one destroying Nagasaki, Japan, 4 days later.

Jet propulsion enabled the fastest fighter aircraft to surpass the speed of sound (in 1947) and eventually to reach maximum velocities in excess of mach 3. The postwar arms race between the United States and the Soviet Union began with the assembly of more

powerful fission bombs to be carried by strategic bombers. The first fusion bombs were tested in 1952 and 1953, and by the early 1960s the two antagonists were engaged in a spiraling accumulation of intercontinental ballistic missiles (ICBMs). But these are not technically weapons of war given that their real purpose was to deter their use by the other side. However, to achieve this objective, the superpowers did not have to amass more than 20,000 nuclear warheads. The sudden end of this expensive arms race did not usher in an era of extended peace; ironically, it initiated an era of even more acute security threats.

In a complete reversal of dominant concerns, the most common—and generally the most feared—weapons in the new war of terror are both inexpensive and easily available. A few kilograms of high explosives fastened to the bodies of suicide bombers (often spiked with metal bits) can cause dozens of deaths and gruesome injuries and can create mass psychosis among the attacked population. Simple "ANFO" car bombs are much more devastating (Table III). These devices, some of which weigh hundreds of kilograms and are able to destroy massive buildings and kill hundreds of people, are made from the mixture of two readily available materials: ammonium nitrate (a common solid fertilizer that can be purchased at, or stolen from, thousands of locations around the world) and fuel oil (which is even more widely available).

The most shocking weapons were fashioned on September 11, 2001, by 19 Islamic hijackers simply by commandeering rapidly moving massive objects, Boeings 757 and 767, and steering two of them into the World Trade Center in New York and one into the Pentagon in Washington, D.C. The World Trade Center towers were designed to absorb an impact of a slow-flying Boeing 707 lost in the fog and searching for a landing at the JFK or Newark airport. Gross weight and fuel capacity of that plane are just slightly smaller (15 and 5%, respectively) than the specifications for the Boeing 767-200, and the structures performed as intended even though the impact velocity of hijacked planes as they rammed into the buildings was more than three times higher (262 vs 80 m/s) than that of a slowly flying plane close to the landing.

As a result, the kinetic energy at impact was approximately 11 times greater than envisaged in the original design ($\sim$4.3 GJ vs 390 MJ). But because each tower had the mass of more than 2500 times that of the impacting aircraft, the enormous concentrated kinetic energy of the planes acted much like a bullet hitting a massive tree. That is, it penetrated instead of pushing; it was absorbed by bending, tearing, and distortion of structural steel and concrete; and the perimeter tube design redistributed lost loads to nearby columns. Consequently, it was not this instantaneous massive kinetic insult but rather a more gradual flux of the ignited fuel (each 767 carried more than 50 metric tons (t) of kerosene whose heat content was more than 2 TJ) that weakened the columns of structural steel.

TABLE III

Kinetic Energy of Explosives and Total Energy Released by Nuclear Weapons

Explosive device	Explosive	Kinetic energy (J)
Hand grenade	TNT	2×10^6
Suicide bomber	RDX	100×10^6
World War II gun shrapnell	TNT	600×10^6
Truck bomb (500 kg)	ANFO	2×10^9
Nuclear bomb	**Reaction**	**Total energy (J)**
Hiroshima bomb (1945)	Fission	52×10^{12}
U.S. ICBM	Fusion	1×10^{15}
Novaya Zemlya bomb (1961)	Fusion	240×10^{15}

Unfortunately, an eventuality of such a fire was not considered in the original World Trade Center design. Moreover, no fireproofing systems to control such fires were available at that time. Once the jet fuel spilled into the building, it ignited an even larger mass of combustible materials (mainly paper and plastics) inside the structures, and the fires burned with a diffuse flame with low-power densities of less than 10 W/cm^2. This left enough time for most people to leave the buildings before the thermally weakened structural steel (the fuel-rich, open-air fire could not reach the 1500°C needed to actually melt the metal), thermal gradients on outside columns, and nonuniform heating of long floor joists precipitated the staggered floor collapse that soon reached the free-fall speed as the towers fell in only approximately 10 s.

Although the preparation of the most feared viral and bacterial weapons of mass destruction requires a fair degree of scientific expertise and a high degree of technical expertise (particularly to prepare the cultured pathogens for widespread dispersal), the overall energy cost of their development is very low in comparison with their potential impact, be it measured in terms of actual casualties and ensuing population-wide fear or in terms of long-term economic losses. There is no doubt that even a largely failed large-scale smallpox or anthrax attack

on a major urban area would send property values falling and would lead to the flight of residents and businesses to less crowded locations.

2. ENERGY COSTS AND CONSEQUENCES OF WAR

Preindustrial land wars, fought largely by foot soldiers with simple weapons, did not require large amounts of energy embodied in arms and equipment, but even so, that investment was often a significant part of annual energy use in many poor subsistence societies. Large armies in transit, or during periods of prolonged sieges, almost always had a deep, and often quite ruinous, effect on regional supplies of food and fuelwood as they commandeered their provisions from the surrounding countryside. In contrast, maritime forays far from friendly shores always required careful planning of supplies that had to be carried by the ships during months of self-sufficient sailing.

Modern wars are waged with weaponry whose construction requires some of the most energy-intensive materials and whose deployment relies on incessant flows of secondary fossil fuels (e.g., gasoline, kerosene) and electricity to energize the machines that carry them and to equip and provision the troops who operate them. Production of special steels in heavy armored equipment typically needs 40 to 50 MJ/kg, and obviously, the use of depleted uranium (for armor-piercing shells and enhanced armor protection) is much more energy intensive. Aluminum, titanium, and composite fibers, the principal construction materials of modern aircraft, typically embody 170 to 250 MJ/kg, as much as 450 MJ/kg, and 100 to 150 MJ/kg, respectively.

The most powerful modern war machines are naturally designed for maximized performance and not for minimized energy consumption. For example, America's 60-t M1/A1 Abrams main battle tank, powered by a 1.1-MW AGT-1500 Honeywell gas turbine, needs (depending on mission, terrain, and weather) 400 to 800 L/100 km; in comparison, a large Mercedes S600 automobile consumes approximately 15 L/100 km, and a Honda Civic needs 8 L/100 km. Jet fuel requirements of highly maneuverable supersonic combat aircraft, such as the F-16 (Lockheed Falcon) and F-18 (McDonnell Douglas Hornet), are so high that no extended mission is possible without in-flight refueling from large tanker planes (the KC-10, KC-135, and Boeing 767).

Moreover, these highly energy-intensive weapons have been used in increasingly massive configurations; hence, the overall energy cost of a conflict can mount rapidly. The most concentrated tank attack during the final year of World War I involved nearly 600 machines, whereas nearly 8000 tanks, supported by 11,000 planes and by more than 50,000 guns and rocket launchers, took part in the final Soviet assault on Berlin in April 1945. During the Gulf War in 1991 ("Desert Storm," January–April 1991) and the months leading to it ("Desert Shield," August 1990–January 1991), some 1300 combat aircraft flew more than 116,000 sorties and the supporting transport and aerial refueling planes logged more than 18,000 deployment missions.

Another historically recurrent phenomenon is the necessity to expand the mass production of these energy-intensive machines in the shortest possible period of time. In August 1914, Britain had only 154 airplanes, but just 4 years later, the country's aircraft factories were sending out 30,000 planes per year. Similarly, when the United States declared war on Germany in April 1917, it had fewer than 300 second-rate planes, none of which could carry machine guns or bombs on a combat mission, but 3 months later Congress approved what was at that time an unprecedented appropriation ($640 million or ~$8 billion in 2000 dollars) to build 22,500 Liberty engines for new fighters. The situation was reprised during World War II. American industries delivered just 514 aircraft to the U.S. forces during the final quarter of 1940, but the total wartime production reached more than 250,000 planes, exceeding the combined output of Germany and Britain.

There are no detailed reasoned studies of energy cost of modern wars. This is not surprising given that even their financial costs cannot be accounted for with a fair degree of accuracy. This has to do primarily with deciding what to include in such accounts. When the very physical survival of a society is at stake, it becomes impossible to separate the output of such a wartime economy into easily identifiable civilian and military sectors and then to assign approximate energy costs to these activities. Available aggregates show the total U.S. expenditures on major 20th-century conflicts as approximately $250 billion for World War I, $2.75 trillion for World War II, and $450 billion for the Vietnam war (all in constant 2000 dollars). Expressing these costs in monies of the day and multiplying the totals by the adjusted averages of respective average energy intensities of the country's gross domestic product (GDP) during those periods sets the minimum energy costs of these conflicts.

Adjustments to average energy intensities are necessary because the war-related industrial production and transportation require considerably more energy per unit of output than does the rest of economic activity. Given the typical sectoral energy intensities of past GDPs, this article uses the conservative multiples of 1.5 for World War I, 2.0 for World War II, and 3.0 for the Vietnam war. Although this procedure cannot yield any accurate figures, it conveys well the most likely magnitude of the energy burden. This burden was approximately 15% of the total U.S. energy consumption during the 1917–1918 period (World War I), averaged roughly 40% during the 1941–1945 period (World War II), but was definitely less than 4% during the 1964–1972 period (main combat action of Vietnam war). Naturally, these shares could be significantly higher during the years of peak war endeavor. For example, the peak World War II spending, expressed as a share of national economic product, ranged from 54% in the United States (in 1944) to 76% in the Soviet Union (in 1942) and in Germany (in 1943).

Peacetime expenditures directly attributable to the preparation for armed conflicts can also be significant. For decades, the highly militarized Soviet economy was spending on the order of 15% of its GDP on the development, procurement, and maintenance of weapons and on its large standing army and border guard and paramilitary forces. In contrast, the U.S. defense spending reached nearly 10% of GDP during the peak year of the Vietnam war, fell to below 5% by 1978, rose to 6.5% during President Reagan's mid-1980s buildup, and has stayed below 5% since 1991. Given the enormous U.S. energy consumption (equivalent to $\sim$2.5 Gt of oil in 2000), the direct peacetime use of fuel and electricity by the U.S. military is a tiny fraction of the total. For example, the U.S. Department of Defense claimed less than 0.8% of the country's total primary energy supply in 2000.

But the comparison looks different in absolute terms. The average direct annual consumption of primary energy by the U.S. armed services was equivalent to approximately 25 Mt of oil during the 1990s. That is roughly the same amount of primary energy consumed annually by Switzerland or Austria; more remarkably, it is a total higher than the commercial energy consumption in nearly two-thirds of the world's countries. Naturally, operations such as the Gulf War, the war in Afghanistan, and the Iraq war boost the energy use by military not only due to the combat but also due to enormous long-distance logistic support. And of course, there are additional (and not easily quantifiable) expenditures of energy for the military research and development (R&D) and the procurement of weapons.

The link between energy use and success in modern war (or in preventing war) is far from being a simple matter of strong positive correlations. Of course, there is no doubt that the possession of nuclear weapons (the MAD [mutually assured destruction] concept) was the main reason why the two superpowers did not fight a thermonuclear war, but the nuclear stockpiles, and hence their energy cost, went far beyond any rational deterrent level. Development and deployment of these weapons, beginning with the separation of the fissile isotope of uranium and ending with the construction of nuclear submarines to carry them with nearly complete invulnerability during the extended submerged missions, has been highly energy intensive.

A conservative estimate might be that at least 5% of all U.S. and Soviet commercial energy that was consumed between 1950 and 1990 was claimed by developing and amassing these weapons and the means of their delivery. And the burden of these activities continues with expensive safeguarding and cleanup of contaminated production sites. Estimated costs of these operations in the United States have been steadily rising, and much greater investment would be needed to clean up the more severely contaminated nuclear weapons assembly and testing sites in Russia and Kazakhstan. Even so, given the potentially horrific toll of a thermonuclear exchange—even when limited to targeting strategic facilities rather than cities, the direct effects of blast, fire, and ionizing radiation would have caused 27 million to 59 million deaths during the late 1980s—one could argue that the overall cost/benefit ratio of the nuclear arms race has been acceptable.

Of course, there is no doubt that the rapid mobilization of America's economic might, which was energized by a 46% increase in the total use of fuels and primary electricity between 1939 and 1944, was instrumental in winning the war against Japan and Germany. In contrast, the Vietnam war was a perfect illustration of the fact that to win a war, it is not enough to use an enormous amount of explosives (the total was nearly three times as much as all bombs dropped by the U.S. Air Force on Germany and Japan during World War II) and to deploy sophisticated weapons (state-of-the-art jet fighters, bombers, helicopters, aircraft carriers, defoliants, etc.). The attacks of September 11, 2001, on the World Trade Center and the Pentagon illustrate the perils and penalties of the aptly named asymmetrical threats. In those attacks, 19 Muslim *jihadis*, at the

cost of their lives and an investment that might not have surpassed $100,000, caused approximately 3000 virtually instantaneous deaths (and the death toll could have easily surpassed 10,000 if the towers were not so structurally sound), created enormous direct and indirect economic dislocations, and led to costly deployments of military and covert power.

During the 2 years following the terrorist attacks, these costs (including the stock value declines and the lost economic output) could be estimated at nearly $2 trillion—and the count keeps rising. Airlines still operate well below their pre-September 11 capacity and are not expected to recover, even in the absence of further attacks, before the year 2007. At the time of this writing, the proposed U.S. 2004 budget envisions nearly $40 billion for the new Department of Homeland Security and approximately $100 billion in additional spending for Iraq and Afghanistan besides the increased baseline spending on defense. Unfortunately, there is no easy military solution for this predicament. Both the classical powerful weapons and the new "smart" machines are of very limited use in this new, highly asymmetric global war, as are most of the security checks and precautions that have been taken so far at the airports.

The two most important and unprecedented consequences of warfare waged by modern societies are the just noted extent of economic mobilization for major armed conflicts and the growing magnitude of destructive power. The latter trend has brought increased casualties, including larger numbers of noncombatants, and huge material losses. World War I was the unmistakable divide. Enormous resources had to be rapidly diverted to war production, and the death toll, in both absolute and relative terms, surpassed that of any previous experience. The increasing destructiveness of modern wars is best illustrated by comparing the overall conflict casualties. Battle deaths, expressed as fatalities per 1000 men of armed forces fielded at the beginning of a conflict, were less than 200 during the Crimean War of 1853–1856 and the Franco–Prussian War of 1870–1871. They surpassed 1500 during World War I and 2000 during World War II (when they were more than 4000 for Russia).

Civilian casualties of modern warfare grew even faster. During World War II, they reached approximately 40 million, more than 70% of the 55 million total. Roughly 100,000 people died during nighttime raids by B-29 bombers using incendiary bombs that leveled approximately 83 square kilometers of Japan's four principal cities between March 10 and March 20, 1945. Five months later, the explosion of two nuclear bombs, which killed at least 100,000 people, released energy of 52.5 and 92.4 TJ, respectively (Table III). At that time, only a few people envisioned how much more powerful these weapons would get. Expressed in common units of TNT equivalents (1 t TNT = 4.184 GJ), the two bombs dropped on Japan rated 12.5 and 22 kt, respectively. The most powerful thermonuclear bomb tested by the Soviet Union, over the Novaya Zemlya on October 30, 1961, rated 58 Mt, equal to 4600 Hiroshima bombs, and by 1990 the total power of U.S. and Soviet nuclear warheads surpassed 10 Gt, the equivalent of 800,000 Hiroshima bombs.

Effects of major wars on energy consumption, be it in defeated countries or in the aftermath of major civil conflicts, are usually quite pronounced. The Soviet Union after the civil war of 1918–1921 and Germany and Japan after their defeat during the late 1940s are the best documented examples. Japan's accurate historical statistics reveal the magnitude of this impact. In 1940, the country's primary energy supply was an equivalent of approximately 63 Mt of oil; by 1945, the total had fallen nearly exactly by half; and in 1946, the aggregate declined by another 10%. The 1940 level was not surpassed until 1955, 10 years after the surrender.

3. ENERGY AS THE CAUSE OF WAR

Finally, this article concludes with a few paragraphs on energy resources as the *cassus belli*. Many historians single out Japan's decision to attack the United States in December 1941 as a classic case of such causation. In January 1940, President Roosevelt's administration, determined to aid the United Kingdom that was under Nazi attack, abrogated the 1911 Treaty of Commerce and Navigation with Japan. In July 1940, Roosevelt terminated the licenses for exports of aviation gasoline and machine tools to Japan, and in September the ban was extended to scrap iron and steel. Already by July 1940, some of Japan's top military officers warned that the navy was running out of oil and that they wanted a quick decision to act. An attack on the United States was to clear the way for the assault on Southeast Asia with its Sumatran and Burmese oilfields.

Although it would be wrong to deny the proximate role that Japan's declining oil supplies might have played in launching the attack on Pearl Harbor, it is indefensible to see Japan's aggression as an energy-driven quest. The attack on Pearl Harbor was preceded by nearly a decade of expansive Japanese militarism (so clearly demonstrated by the 1933

conquest of Manchuria and by the attack on China in 1937), and Marius Jansen, one of the leading historians of modern Japan, wrote about a peculiarly self-inflicted nature of the entire confrontation with the United States. No convincing arguments can made to explain either Hitler's serial aggression—against Czechoslovakia (in 1938 and 1939), Poland (in 1939), Western Europe (beginning in 1939 and 1940), and the Soviet Union (in 1941)—or his genocidal war against Jews by concerns about energy supplies.

The same is true about the geneses of the Korean War (North Korea is the coal-rich part of the peninsula), the Vietnam war (waged by France until 1954 and by the United States after 1964), the Soviet occupation of Afghanistan (1979–1989), and the U.S. war against the Taliban (launched in October 2001) as well as about nearly all cross-border wars (e.g., Sino–Indian, Indo–Pakistani, Eritrean–Ethiopian) and civil wars (e.g., Sri Lanka, Uganda, Angola, Colombia) that have taken place (or are still unfolding) during the past two generations in Asia, Africa, and Latin America. And although it could be argued that Nigeria's war with the secessionist Biafra (1967–1970) and Sudan's endless civil war had clear oil components, both were undeniably precipitated by ethnic differences and the latter one began decades before any oil was discovered in the central Sudan.

On the other hand, there have been various indirect foreign interventions in the Middle Eastern countries—through arms sales, military training, and covert action—that aimed at either stabilizing or subverting governments in the oil-rich region. Their most obvious manifestation during the cold war was the sales (or simply transfers) of Soviet arms to Egypt, Syria, Libya, and Iraq and the concurrent American arms shipments to Iran (before 1979), Saudi Arabia, and the Gulf states. During the 1980s, these actions included strong and, as it turned out, highly regrettable Western support of Iraq during its long war with Iran (1980–1988). This brings us to the two wars where energy resources have been widely seen as the real cause of the conflicts.

By invading Kuwait in August 1990, Iraq not only doubled crude oil reserves under its control, raising them to approximately 20% of the world total, but it also directly threatened the nearby Saudi oilfields (including the four giants—Safania, Zuluf, Marjan, and Manifa—that are on- and offshore just south of Kuwait) and, hence, the very survival of the monarchy that controls 25% of the world's oil reserves. Yet even in this seemingly clear-cut case, there were other compelling reasons to roll back the Iraqi expansion. At that time, the United States was importing a much smaller share of its oil from the Middle East than were Western Europe and Japan (25% vs more than 40% and 67%, respectively, in 1990), but the Iraqi quest for nuclear and other nonconventional weapons with which the country could dominate and destabilize the entire region, implications of this shift for the security of U.S. allies, and risks of another Iraqi–Iranian or Arab–Israeli war (recall Saddam Hussein's missile attacks on Israel designed to provoke such a conflict) certainly mattered a great deal.

There is one very obvious question to ask: if the control of oil resources was the sole objective, or at least the primary objective, of the 1991 Gulf War, why was the victorious army ordered to stop its uncheckable progress as the routed Iraqi divisions were fleeing north, and why did it not occupy at least Iraq's southern oilfields? Similarly, more complex considerations were behind the decision to conquer Iraq in March 2003. Although the *ex post* perspectives may question a justification based on the Iraqi development of weapons of mass destruction, it could be argued that Hussein's past actions and the best available prewar evidence did not leave any responsible post-September 11 leadership in the United States with the luxury of waiting indefinitely for what might happen next. Such a wait-and-see attitude was already assumed once, with fatal consequences, after the World Trade Center bombing in February 1993.

Of course, a no less important reason for the Iraq war has been the grand strategic objective of toppling the Baathist regime and to eventually replace it with a representative government that might serve as a powerful and stabilizing political example in a very unsettled region. Once again, after the events of September 11, one could argue that no responsible U.S. leadership could have ignored such a risky but potentially immensely rewarding opportunity to redefine the prospects for a new Middle East. And so even this instance, which so many commentators portray so simplistically as a clear-cut case of energy-driven war, is anything but that—yet again confirming the conclusion that in modern wars, resource-related objectives have been generally determined by broader strategic aims and not vice versa.

SEE ALSO THE FOLLOWING ARTICLES

Cultural Evolution and Energy • *Global Material Cycles and Energy* • *Gulf War, Environmental Impact of* • *Nationalism and Oil* • *National Security*

and Energy • Nuclear Proliferation and Diversion • Oil Crises, Historical Perspective • OPEC Market Behavior, 1973–2003 • Sociopolitical Collapse, Energy and • World History and Energy

Further Reading

Clark, G. B. (1981). Basic properties of ammonium nitrate fuel oil explosives (ANFO). *Colorado School Mines Q.* **76,** 1–32.

Committee for the Compilation of Materials on Damage Caused by the Atomic Bombs in Hiroshima and Nagasaki. (1981). "Hiroshima and Nagasaki." Basic Books, New York.

Eagar, T. W., and Musso, C. (2001). Why did the World Trade Center collapse? Science, engineering, and speculation. *J. Metals* **53,** 8–11. www.tms.org/pubs/journals/jom/0112/eagar/eagar-0112.html.

Federation of American Scientists (1998). "Operation Desert Storm." FAS, Washington, DC. www.fas.org/man/dod-101/ops/desert_storm.htm.

Federation of American Scientists. (2003). "America's War on Terrorism." FAS, Washington, DC. www.fas.org/terrorism/index.html.

Jansen, M. B. (2000). "The Making of Modern Japan." Belknap Press, Cambridge, MA.

Lesser, I. O. (1991). "Oil, the Persian Gulf, and Grand Strategy." RAND, Santa Monica, CA.

McNeill, W. H. (1989). "The Age of Gunpowder Empires, 1450–1800." American Historical Association, Washington, DC.

MoveOn.org. (2002). "Energy and War." www.moveon.org/moveonbulletin/bulletin8.html#section-9.

Robertson, L. E. (2002). Reflections on the world trade center. *The Bridge* **32**(1), 5–10.

Singer, J. D., and Small, M. (1972). "The Wages of War 1816–1965: A Statistical Handbook." John Wiley, New York.

Smil, V. (1994). "Energy in World History." Westview, Boulder, CO.

Smil, V. (2003). "Energy at the Crossroads: Global Perspectives and Uncertainties." MIT Press, Cambridge, MA.

Urbanski, T. (1967). "Chemistry and Technology of Explosives." Pergamon, Oxford, UK.

Waste-to-Energy Technology

MICHAEL L. MURPHY
Energy Products of Idaho
Couer d'Alene, Idaho, United States

1. **Waste to Energy**
2. **History of Waste Combustion**
3. **Typical RDF Waste-to-Energy Facility Design**
4. **Conversion to Energy**
5. **Emissions Control Technology**
6. **Not in My Backyard**

Glossary

fluidized bed combustor The term used to describe a method of combustion in which a bed of sand is suspended in an upward-flowing airstream; the turbulence created by the "boiling" sand bed enhances combustion and promotes efficient heat transfer and uniform mixing.

gasification The process of partially oxidizing the fuel feedstock by introducing it into a furnace with insufficient air/oxygen to support complete combustion; the resulting exhaust gas contains hydrogen, methane, and other organic gases along with carbon monoxide, carbon dioxide, and nitrogen.

mass burn The term used to describe the practice of combusting municipal solid waste (MSW) in a furnace where the material is fed in with little or no prescreening or separation; within the furnace, the combustible content is burned away and the remaining ash, glass, and metal from the MSW is removed as a bottom ash.

municipal solid waste (MSW) A general characterization of the waste stream resulting from the accumulation of all residential, commercial, and light industrial waste normally handled by local municipalities and typically deposited in a landfill; this waste stream consists of metal, glass, dirt, vegetation, paper, plastic, and the like, whether recyclable or nonrecyclable in nature.

Rankine power cycle The process of converting heat energy from fuel into high-temperature, high-pressure steam that is then used to drive a steam turbine generator to create electricity.

refuse-derived fuel (RDF) The fraction of municipal solid waste (MSW) remaining after processing, separating, and sizing of the MSW to remove any recyclable components and noncombustible materials. This stream typically consists of paper, plastics, vegetation, and wood along with minimal glass and metal.

Every year, millions of tons of solid waste are disposed of by landfilling. At the same time, millions of therms of energy from oil, natural gas, and/or coal are burned to create electrical power. The energy content available from the solid waste residue represents a significant "alternative" energy supply to alleviate the use and dependency on fossil fuels. This article discusses the details regarding the energy content of municipal waste and the potential technology to convert and recover that energy as useful electricity in a waste-to-energy system.

1. WASTE TO ENERGY

The category of waste to energy broadly describes any of a number of processes in which a useful by-product (energy) is recovered from an otherwise unusable source. The sources of waste can vary nearly as much as the varieties of social and economic communities worldwide. By definition, the term "waste" implies a nonusable by-product of a process or production. However, the factors that make this a waste stream from one perspective do not necessarily make the material worthless or useless in some other form or fashion. In fact, in the global aspect of waste management and disposal, recycling and reuse of waste materials have become prominent means of reducing the quantities of discarded materials. Recovery of available energy from the waste is, in truth, simply another specific method of reuse. In general terms, "waste" could describe anything from liquid chemical effluent to used automobile tires or anything from municipal sewage treatment residue (sludge) to municipal solid waste (MSW), also known as garbage. With the

 © 2004 Elsevier Inc. All rights reserved.

exception of certain private disposal locations, most of the waste generated from any single source will eventually end up as a contributing component to the overall MSW disposal composition for a specific community. Depending on population density, local industrial development, and numerous other socio-economic factors, the actual makeup of the MSW stream could be significantly weighted to consumer waste products or industrial wastes. In addition, MSW composition can have significant seasonal variations, with yard wastes and overall quantities increased during the nonwinter months. Such variations in the waste characterization can have a significant impact on the value of the material as an energy source.

The value of waste as an energy source is determined by its composition and fuel calorific heating value, typically measured in British thermal units (Btu) per pound. The heating value is determined specifically from the available hydrogen and carbon content in the fuel that provide heat from combustion and result in by-products of water and carbon dioxide. Conversely, the heating value is reduced by the level of noncombustible material remaining in the waste. Glass, metal, and dirt are the most significant factors to the noncombustible portion of the waste stream.

A quick review of the general composition of MSW (Table I) provides a reference for the sources of energy available within a waste stream and provides a relative comparison of which components are most and least valuable for energy. Interestingly, many of the components within the waste stream least valuable as an energy source are most valuable as recyclable materials. Aluminum, ferrous metals (iron), plastics, newsprint, cardboard (corrugated paper), and glass are typically among the sought after recyclable components of a waste stream. With the exception of the plastics and paper components, the other recyclable components contribute nothing to the energy content of the waste stream. The most recent developments in solid waste management efforts have taken advantage of this fact by endeavoring to remove these recyclable materials from the waste stream prior to converting into a fuel.

TABLE I

Sources of Energy Available within a Waste Stream

Component	Contribution (percentage)	Heat value (Btu/pound, dry)	Moisture (percentage)
Newspapers/Magazines	18.0	8900	30
Paper/Cardboard	20.0	8900	32
Plastic	4.5	12,000	0
Dense plastic	4.2	16,700	0
Textile, leather	3.3	9000	20
Glass	5.8	0	0
Vegetable waste	20.2	8200	55
Metal—iron	5.7	0	0
Metal—aluminum	1.0	0	0
Wood/Yard waste	7.3	8400	40
Grit, dirt, misc	10.0	0	55
Rubber tires	NA	15,000	0

2. HISTORY OF WASTE COMBUSTION

The earliest waste combustion systems were simple incinerators. With no energy recover capabilities and only basic wet scrubber gas cleanup technology, this practice was able to reduce the quantity of waste ultimately routed to the landfill by 75 to 85%. These early incinerators were typically small (only 50–100 tons/day capacity) and were considered simple, mass burn, starved air, multistaged incinerators. During operation, the waste was introduced into the furnace onto the first of a series of ram-activated cascading grates. In this primary zone, the fresh fuel would contact some combustion air and a stream of recycled gases from the exhaust header. As the waste began to burn, the ram feeders would push the pile off the grate onto the next step. In the process, the material would be stirred and agitated to present a fresh combustion surface to the airstream. Likewise, the material from this second step was ram fed inward until it "cascaded' onto the next grate. This process continued until the remaining noncombustible material plus some unburned carbon fell into a water-cooled ash trough and was removed from the furnace in an ash conveyor. The combustion and volatile gases driven off from the fuel pile were mixed with combustion air in the primary chamber directly above the grate. Further combustion and quench occurred in the secondary and tertiary chambers, where additional ambient air and recycled flue gases were mixed into the combustion exhaust stream. All of the gases from the tertiary chamber were conveyed into the wet scrubber, where particulate was removed.

During the early 1970s, the surge in energy costs, spurred by the oil embargo, created a demand for energy recovery in these waste disposal facilities. At that time, although waste-to-energy technology was

still in its infancy stage, it was poised to take one of two paths to implementation: mass burn or process fuel (typically known as refuse-derived fuel [RDF]). Like the earlier technologies, mass burn technology opted to burn the waste virtually as it was received, eliminating nearly all processing of the material prior to burning. The mass burn units were typically very large furnaces, most often field erected, where the waste was literally plucked from a receiving floor with a grappling hook and ram fed into the furnace onto a cascading grate design. Following the success of other European designs beginning in the 1950s, these systems were typically of the water wall furnace design as opposed to a more "conventional" refractory furnace design. One advantage of this was the direct removal of much of the heat from the furnace by the water walls, thereby minimizing the requirement for significant volumes of excess combustion air or recycled flue gas to maintain acceptable operating temperatures in the furnace. These plants have also evolved into much larger capacity facilities—they were already field erected, so upsizing the capacity did not represent any further economic disadvantage—and have been built to capacities as large as 2500 tons per day (TPD).

The extent of preprocessing of the waste stream, and the resultant potential for recovery and recycling of a majority of the waste materials, is also a significant factor in determining what type of energy recovery technology might be considered for the waste stream. Waste-to-energy technologies can be generally categorized into two types: mass burn and RDF. Mass burn technology derives its name from the fact that very little effort is spent in the fuel preparation area. Removal of oversized items and obvious noncombustible materials, such as white goods (appliances), engine blocks, and propane canisters, comprises the segregation efforts to the waste stream. The remaining waste is fed, en masse, into the furnace of the burner. RDF technology is perhaps not the most descriptive nomenclature to describe all non-mass burn technologies. RDF denotes the waste processing effort that typically involves sorting, separating, and sizing the incoming waste stream to generate various by-product streams. A portion of the wastes (typically up to 30%) can be separated and recovered for recycling. A portion (10–20%) can be separated and diverted on to the landfill for disposal. The final portion (50–75%) contains the greater portion of the energy contributors, with much of the non-energy contributors removed, and can then be shredded and sized to a much more homogenous fuel stream called RDF.

With much of the noncombustible material eliminated from the blend plus the shredding of the waste material to a more uniform size, the remaining material is much more suited as a fuel feedstock for a more conventional solid fuel boiler technology such as water-cooled grate or fluidized bed boilers. Among these classifications are various other categories that delineate even further various advantages and design features of each technology. Such distinctions are left to future discussions on the subject. Suffice it to say that RDF technology is more dependent on preprocessing of the waste stream to create a fuel feedstock with some level of homogeneity, both in composition and in particle sizing. The obvious advantages to this are the improved combustion control resulting from more stabile fuel characteristics and improved design criteria as well as smaller furnace volumes resulting from better fuel definition and less potential variability in the incoming feed stream. The environmental advantage gained from RDF processing is a greater opportunity to recycle primary components from the waste stream without imposing added effort on the consumer or waste collection agency.

Mass burn technology, on the other hand, has proven to be the favored technology historically due to the extreme simplicity associated with the fuel handling requirements. In many cases, the "front end" technology, the common term used to define the fuel processing requirements on the front end of a waste-to-energy facility, was nothing more than a receiving pit for mass dumping and storage of all the waste coming from the hauling trucks along with a grappling hook and oversized conveyor to deliver the waste into the furnace. The waste delivered to the furnace is typically ram fed through a penetration in the furnace wall onto a water-cooled, cascade grate furnace floor. Typically, a pile of fuel, often representing 15 to 40 min of inventory, is maintained on the floor of the furnace. Control of the combustion characteristics is largely achieved via combustion air control only, with the fuel pile being too significant to facilitate anything but the most crude combustion adjustments. Because the mass burn furnaces must operate under the many variations in fuel quality and size presented by the bulky fuel mix, the furnace must be extremely robust and conservative in design. In addition, the ash removal system must be capable of handling every bit of the size and capacity of material coming out the furnace as it is fed into the furnace on the front end. Although this universal acceptability has proven to be one of the most important features of the mass burn success, it

has also created certain liabilities in the performance of the furnace. Extreme variations in the fuel within the furnace can create significant temperature excursions, among other things, and these can cause damage to the grate and other areas of the boiler. Unpredictable fuel composition can create surges in ash quantities, resulting in slagging and fouling of the boiler surfaces and increased emissions levels of unburned hydrocarbons, acid gases, and dioxins/furans in the outlet gases. In spite of these obvious drawbacks, mass burn technology flourished during their first three decades of development in the United States and most worldwide markets.

This was not the initial direction of the technology during the late 1960s and early 1970s. At that time, the industry appeared to be heading down the path of waste processing to generate an RDF feedstock. Unfortunately, explosions and catastrophic failures in some of these first facilities, including some fatalities, squelched the interest in further processing efforts and opened the door wide open for mass burn development. Furthermore, at the onset of the waste-to-energy industrial development, there was little interest in the recovery and recycling markets; the waste was processed simply to create a more usable fuel feed. However, much has changed over the past three decades, and the recent vitalization of the recycling effort has fostered a resurgence of processing technology of waste streams, with the primary focus being on recovery and recycling. The fallout from the tail end of the process has become the fuel stream for many of the modern waste-to-energy systems and has enabled the RDF technologies to move to the forefront of the technical arena.

3. TYPICAL RDF WASTE-TO-ENERGY FACILITY DESIGN

To better describe the details of waste-to-energy technology, including specific items pertaining to energy conversion, environmental considerations, operating and maintenance features, and other specific aspects of the technology, this section presents a more detailed discussion of a specific design of a modern waste-to-energy facility that might be considered typical of the industry in general (Fig. 1). The energy recovery technology being presented is for a bubbling fluidized bed (BFB) boiler and has a number of technology-specific criteria that will be different with different boiler technologies. Where applicable, such differences are identified and clarified for the other approaches; however, for discussion purposes, the level of detail is focused on the BFB technology.

Perhaps the first criterion in defining a waste-to-energy facility is to establish the amount of waste requiring disposal. For purposes of this exercise, it is assumed the system will be designed for an application in the United States. The average person in the United States is reported to generate between 3 and 4 pounds of waste per day. Modern waste-to-energy facilities have been sized for anywhere from less than 500 tons per day (TPD) to more than 2500 TPD. For purposes of this design exercise, an incoming stream of 1000 TPD is considered, representing a municipal community of approximately 500,000 to 600,000.

The initial activity in the design effort is to establish the fuel quantity and quality resulting from the incoming waste stream supply. This is determined by the design of the waste processing facility and the extent to which noncombustible and/or nonrecyclable materials are removed from the waste stream prior to creating the RDF. In a very simplified process, magnetic separation is used to remove ferrous materials, and some form of density separation (e.g., air knives, flotation) or screening (e.g., trommel, disc screen) is used to further classify the lighter fraction (e.g., paper, plastics, textiles, wood wastes, sawdust) from the larger and heavier material (e.g., rocks, metals, wet material). The remaining lighter waste is then shredded and sized to create a fuel feed material. This shred–burn concept represents the simplest method for material separation but results in the least opportunity for recyclable recovery and usually loses a substantial amount of the potential fuel value with the rejected stream. In this case, a more rigorous material

FIGURE 1 Typical waste-to-energy facility in Ravenna, Italy.

recovery facility (MRF), which will maximize recoverable materials and produce a quality RDF for use in the fluidized bed boiler system, is considered.

In the MRF, waste material is received on an open area or tipping floor, where it can be grossly sorted to remove any significant items, such as white goods and furniture, from the remaining mixed waste. From the tipping floor, material is loaded onto a transport conveyor that delivers it into a rotary trommel screen. Prior to the trommel, the material passes along a "picking station" where bulky materials, especially cardboard and color-coded recyclable bags, may be removed and diverted to a separate processing line. Inside the trommel, the material passes through a section where wall-mounted knives tear open any sealed garbage bags and empty the contents into the main stream. The trommel is a rotating cylinder, laying on its side, with holes of progressively larger sizes cut into its outer shell. The first series of holes are typically 2 inches in diameter or smaller. These will separate out the finer fraction of dirt and grit, small rocks and gravel, bits of wire and metal, and broken glass as well as much of the fine combustible material, including wood waste (sawdust) and paper and plastic scrap. This separated stream can be further processed with magnets, air classification, and flotation to remove and recover the ferrous metal, light fraction combustible material, and heavy material and glass, respectively.

Within the trommel, the larger sized material remaining is further sized to approximately 6 inches with a second series of holes. Through this section, separation of most of the normal waste materials is accomplished, including plastic and glass containers, aluminum cans, organic and yard wastes, and a significant portion of the remaining wastepaper fraction. Again, further processing of this separated stream, including magnetic and eddy current separation, air curtains, and plastic sorting lines, can classify this stream into ferrous metals, aluminum, glass and plastic containers, and potential fuel.

The remaining fraction of waste passing completely through the trommel screen and having a size exceeding 6 inches can then be hand sorted, if desired, and shredded to nominal size of 4 inches or smaller. Once again, magnetic separation and air classification can separate this stream into ferrous metal, fuel feedstock, and heavy noncombustible residue.

Once through the separation processing, the incoming waste stream has been classified into a number of individual streams, including ferrous metals, aluminum, glass, dirt and grit, and fuel (i.e., RDF). The fuel analysis and characterization of a typical RDF is shown in Table II. Characterization of such widely divergent waste streams into a single "homogenous" blend, as is reported herein, is an extreme simplification of the overall fuel analyses and represents a "leap of faith" in believing such information to be absolute and all encompassing. Perhaps more important than the information provided in Table II is the knowledge that these fuel specifications can vary as widely as the sources of waste from which they are derived.

Of the total 1000 TPD of waste arriving at the facility, approximately 20 to 35% may be removed, recovered, and possibly recycled. Normally, the yield of fuel from incoming waste is predicted to be 75%, resulting in an average fuel generation capacity of 750 TPD. Taking into account daily, weekly, and seasonal variations in the waste generation and delivery schedules, it is practical to expect higher hourly averages during the morning hours and higher daily averages during the middle of the week. Consequently, design considerations may be necessary to accommodate a "weekly" capacity of 7000 tons that may actually be delivered to the processing facility in 4 to 5 days. Although storage of unprocessed waste is not condoned—in fact, most regulations require the waste to be received and processed within the same day—it is necessary that a fuel facility provide storage capability to enable the 4- to 5-day waste supply to be consumed in the power plant over the full 7-day period.

TABLE II

Refuse-Derived Fuel Analysis

Component	Percentage dry basis
Carbon (C)	45.3
Hydrogen (H_2)	6.0
Nitrogen (N_2)	0.7
Chlorine (Cl)	0.5
Sulfur (S)	0.3
Oxygen (O_2)	31.3
Ash	13.4
Glass	1.5
Iron (Fe)	1.0
Higher heat value, (Btu/pound, dry)	7800
Moisture (percentage wet)	25
Higher heat value (Btu/pound, as received)	5850

4. CONVERSION TO ENERGY

The preceding discussion has focused on the material processing requirements necessary to generate a fuel from the incoming municipal waste. Once this is achieved, the focus of the waste-to-energy facility turns to converting the energy in the waste into a beneficial use. This is accomplished through the use of a steam boiler and power turbine cycle, much the same as most solid fuel power plants throughout the world. In simple terms, the fuel is burned in a furnace to generate heat. The heat is used to generate steam in a power boiler, which in turn drives a power turbine to generate electricity. The exhaust gases from the steam generator, having been cooled to nearly ambient temperatures, are cleaned of most pollutants and exhausted into the air. The electricity is transferred into the utility transmission system and eventually back to the homes and industries that generated the waste.

The power cycle is simple and straightforward, and it dates back to the industrial revolution. It is based on proven technology and, for the most part, proven equipment. Unfortunately, the drawback of the steam turbine (Rankine) power cycle is the exorbitant losses experienced by the condensing of the low-temperature, low-pressure steam at the discharge of the turbine. This represents two-thirds to three-quarters of the energy content of the steam and limits the overall power cycle efficiency to 20 to 25%. Improvements in metallurgy and construction techniques over the past few decades have raised operating temperatures and pressures within the steam cycle to increase these turbine cycle efficiencies significantly; however, most of the resulting improvements are not practical for normal waste-to-energy facilities. First, the relative size of the waste-to-energy power plant is too small to absorb many of the operating and capital costs resulting from these critical pressure and temperature conditions. Furthermore, because of the contaminants in the waste stream, most notably chlorine and sulfur, the potential for high-temperature corrosion throughout the process is significant. To optimize plant costs with operating efficiencies, steam conditions have typically been limited to 650 psia and 750°F. Some recent facilities have increased this limit to 850 psia and 825°F with limited success.

With the power cycle established by the steam conditions of 650 psia and 750°F, the overall power generation capacity of the 750 TPD in incoming RDF can be determined. At 31.25 tons per hour and the referenced fuel heating value, an input energy level of 366 million Btu per hour is available. Operating at excess air levels in the furnace of at least 45% and normal gas exhaust temperatures of approximately 375°F, the overall boiler efficiency for this fuel is estimated at 77%. Given normal fluidized bed boiler design characteristics, this unit would generate approximately 250,000 pounds per hour (PPH) of steam. In a typical turbine power cycle, this steam quantity would deliver a gross generation capacity of 25,500 kW. Thus, we can reduce the waste-to-energy criteria for this summary to a simplified ratio of approximately 40 TPD per megawatt (TPD/MW) of incoming municipal waste to power generation capacity. In addition, this also reduces to approximately 30 TPD/MW of incoming RDF to output power.

5. EMISSIONS CONTROL TECHNOLOGY

Environmental compliance and controls are significant criteria to consider in waste-to-energy facilities. As guidelines and standards have become more stringent, abatement and control technologies have been developed to keep pace. In the earliest waste incineration systems, most of which did not even incorporate any energy recovery, the mass burn combustion technology was very crude, the fuel was extremely low quality, and the emissions control requirements were very simple (if present at all). Current requirements address emissions of products of combustion such as carbon monoxide (CO) and volatile organic hydrocarbons (VOCs) and nitrogen oxides (NO and NO_2). Acid gas emissions resulting from sulfur and chlorine are required to be reduced significantly. High-efficiency filters are mandated for reduction of particulate emissions as well as heavy metals.

Current furnace designs and combustion practices have achieved very high combustion efficiencies resulting in extremely low CO and VOC levels, commonly referred to as products of incomplete combustion (PICs). These can also include the category of dioxins and furans, which encompass a specific family of ploy aromatic hydrocarbons (PAHs). In the case of fluidized bed combustion, the technology describes the practice of burning a solid fuel in a combustion zone consisting of small particles of sand suspended in a stream of upward-flowing air. The buoyancy force of the air velocity lifts and suspends the individual sand particles within

the mixture. As more air is introduced into the sand, it appears to bubble in much the same way as a pot of boiling water, hence the term "fluid bed" and, more specifically, "bubbling fluid bed." In a modification to this concept, referred to as a circulating fluid bed, the air velocities are actually increased to the terminal velocity of the sand particles, thereby carrying the sand up and out of the furnace. At the outlet of the furnace, the sand is mechanically separated from the air/gas stream and is recirculated back into the bottom of furnace, hence the term "circulating fluid bed" or "recirculating fluid bed."

The advantages of fluid bed combustion over the more traditional technology arise from the increased turbulence provided by the bed particle action. This fluidization increases the interaction of the fuel particles with the combustion air and creates a very accelerated combustion environment for the incoming fuel. In addition, the sand, initially heated to an ignition temperature for the incoming fuel, provides a means of heating and drying the new fuel introduced into the furnace and eventually igniting it. The thermal stability provided by the hot sand in the bed plus the turbulence of the fluidization causes the fuel particles to be consumed in a very short period, typically a few minutes or less. Complete combustion is achieved by the uniform temperature zone within the furnace created by the bubbling sand medium plus the enhanced intermixing of fuel and combustion air. This technology is illustrated in Fig. 2. Historically, fluidized bed combustion systems have been able to achieve better operating flexibility, lower emissions levels, and improved combustion and boiler efficiencies. In the case of RDF combustion requirements, where the fuel characteristics are nonhomogenous and the overall fuel quality is relatively low compared with standard fuel such as coal, the fluid bed provides a method of stabilizing the combustion characteristics and creating very good emissions results. In some locations, especially the European Community, regulations specify a furnace temperature of at least 1550°F for a minimum of 2 s residence time and a minimum oxygen level of at least 6% in the furnace. This is not a universal requirement for design of the furnace; however, it does establish reasonable minimal levels for good combustion practices. Table III displays the current requirements for emission levels for new waste-to-energy power plants. Typical emissions levels from operating systems are included for comparison.

NO_x emissions are the result of combustion/furnace conditions that promote the "burning" of nitrogen. The source of nitrogen can be air (as is normally supplied for combustion) or fuel-bound nitrogen that is contained in the basic analysis of most solid fuels to some degree. To generate NO_x from air-bound nitrogen, very high temperatures (>2200°F) are required. Although not normal in most types of waste-to-energy technologies, these temperatures can be created in specific areas of the fuel piles in mass burn and conventional grate furnaces. Lack of mixing of the fuel with the air can result in channeling, a condition where air and fuel mixtures can become very rich, resulting in the high temperatures. With the mixing of the fuel and sand in the fluidized bed combustor (FBC) plus the reduced operating temperatures required to achieve good combustion, the creation of NO_x from air

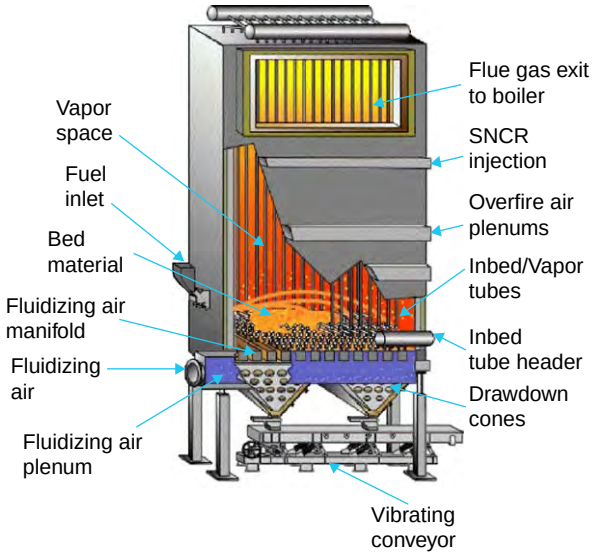


FIGURE 2 Bed combustor/boiler.

TABLE III

Typical Emissions Requirements (European Standards)

Temperature	850°C	1560°F
Oxygen (percentage dry)	6	6
Reference O_2 (percentage)	11	11
Particulate	10 mg/nm^3	0.006 g/sdcf
CO	50 mg/nm^3	44 ppm
VOC	10 mg/nm^3	6 ppm
SO_2	50 mg/nm^3	19 ppm
HCl	10 mg/nm^3	6 ppm
HF	1 mg/nm^3	
NO_X	200 mg/nm^3	105 ppm
NH_3	10 mg/nm^3	
Total heavy metals	1 mg/nm^3	0.0006 g/sdcf
Dioxins (total)	0.1 ng/nm^3	

fixation is essentially nonexistent. On the other hand, fuel-bound nitrogen accounts for most of the NO_x generation in FBC units and contributes even more NO_x in other types of furnaces. In all cases, the uncontrolled (or combustion-created) NO_x levels are likely to be well in excess of the allowable limits established by current regulations. Controlling combustion parameters, such as total excess air levels, and introducing combustion air into the furnace in stages to avoid high levels of excess oxygen in any particular combustion zone both can contribute to reduced NO_x emissions. Postcombustion control using selective noncatalytic reduction (SNCR) technology in the high-temperature region or selective catalytic reduction (SCR) in the cooled exhaust gas zone can reduce NO_x levels to allowable limits under most circumstances. In either process, ammonia is introduced into the flue gas. At the high temperatures ($>1500°F$), the ammonia can react with NO_x to produce nitrogen gas and water vapor. At the lower temperatures, the same reactions occur; however, the reactions require a special catalyst to increase the reactivity of the ammonia in the flue gas. Restrictions in temperature and flue gas characteristics have so far limited the application of SCR technology to these processes. To date, SNCR has been successfully incorporated into all of the different types of furnaces.

Acid gas emissions of sulfur and chlorine as SO_2 and HCl, respectively, are largely independent of the type of furnace used in the waste-to-energy technology. It is a fair assumption that all of the sulfur and all of the chlorine inbound with the fuel will be released during the combustion as SO_2 or HCl. The exception to this, at least partially, is the ability of the FBC to "capture" some of the SO_2 created. In the fluidized bed furnace, a portion of the sand media can be displaced with a calcium oxide material, or limestone, which has an affinity for SO_2. At normal temperatures within the furnace, the limestone can react with the SO_2 to create calcium sulfate ($CaSO_4$), also known as gypsum, which is a stable solid compound at temperatures below approximately 1700°F. Thus, the fluid bed can reduce the overall sulfur gas emissions significantly, in some cases by as much as 95%, by incorporating a continuous feed of limestone into the furnace. For any excess SO_2 plus all of the HCl emissions from the furnace, additional scrubbing is required at the exhaust end of the system before the gases are discharged into the environment. Various methods are available to effect this scrubbing, including wet or dry scrubbers or some combination thereof. In all cases, the flue gases from the boiler are injected with some reagent, either a calcium-based material (lime) or some other acid-neutralizing compound that reacts with the sulfur and chlorine gases to form a solidified compound, such as the gypsum, of sodium chloride (NaCl), commonly known as salt. With this add-on technology, the exhaust levels for most acid gases can be reduced by 95 to 99% of the inlet concentrations.

Once the acid gases have been converted to a solid material, they become a part of the particulate emissions from the system that also must be removed from the gases before they can be released into the air. High-efficiency bag filters (or baghouses) and electrically charged particle separators (known as electrostatic precipitators [ESPs]) are commonly used for this. Using state-of-the-art filter materials, dust removal efficiencies as high as 99.9% can be achieved in this system.

Once the combustion process is complete and the flue gases have been treated and exhausted into the air, the waste has been reduced to an ash stream that comprises about 5% of the original volume of the waste. This ash is made up of essentially all the noncombustible material in the fuel stream plus any additional materials such as limestone, or lime, that may have been added to the process for gaseous emissions controls as discussed previously. This ash stream becomes the concentrated repository of any of the metals and other such materials found in the waste stream that have not been removed by preliminary processing. Consequently, the ash has become a major focal point for "concerns" about environmental contamination resulting from disposal of this waste. The ash can be demonstrated as a nonhazardous waste through standard tests such as the toxic characteristic leaching procedure (TCLP), which measures the quantity of metals leached from a sample. Years of testing ash samples from numerous types of waste-to-energy facilities have consistently demonstrated the nonhazardous nature of the material. In fact, hundreds of thousands of tons of ash are currently being used in beneficial products such as landfill cover, road aggregate, asphalt mixture, concrete, and granular base material. In most instances, the characteristics of the ash are more benign than those of the metals and other ingredients as they exist in the original waste stream.

6. NOT IN MY BACKYARD

As a fuel, MSW does not compare too favorably with more traditional solid fuels such as coal. MSW averages in the neighborhood of 4500 Btu per pound

versus coal at 10,500 to 13,000 Btu per pound. However, given the current U.S. population of approximately 250 million and the annual generation of waste per person of roughly 1500 pounds, the potential energy content in the annual waste generated in the United States alone is comparable to nearly 70 million tons of coal annually and has the potential to generate more than 13,000 MW of electrical power. As of 1993, the capacity of waste consumed in these facilities was approximately 103,000 TPD, representing approximately 18% of the projected annual waste generation.

As of early 2000, waste-to-energy technology accounted for disposal of more than 30 million tons of waste annually in the United States alone. This corresponds to an electricity generating capacity of more than 2800 MW. Although these numbers might seem significant, this represents only approximately 14% of the total waste stream within the United States and accounts for only one-third of 1% of the electrical generating market in the United States. Waste-to-energy technology represents one of the most significant sources of renewable energy available today. It was reported during the mid-1990s that 128 facilities were actually in operation in the United States, with an additional 42 planned or under construction. Of the existing plants, 92 produced electricity to some degree. The remaining 36 produced only steam or hot water. The size of facilities ranged from 20 plants of less than 100 TPD to 20 plants of more than 2000 TPD. Roughly 40 plants were of 100 to 500 TPD, 25 plants were of 500 to 1000 TPD, and 25 plants were of 1000 to 2000 TPD capacity. Most of the smaller plants were installed during the early period of waste-to-energy development. Most of the recent facilities represent sizes from 500 to 1000 TPD.

Waste-to-energy technology not only provides a clean source of energy but also provides a clean and long-term disposal of the waste. Studies have compared the emissions from a waste-to-energy facility with those from a landfill handling similar quantities of waste. In cases where no methane recovery was included in the landfill, a ton of waste was shown to generate 25 times more greenhouse gas emissions than was a ton of waste combusted in a modern waste-to-energy facility. Furthermore, even with landfill gas recovery techniques currently available, waste-to-energy systems are still at least 10 times more beneficial than landfilling.

So, with all of the wonderful benefits to this technology, why has it not been more widely accepted and implemented? First, this technology is not specific to any one industry, utility, or municipality. The fuel source is generated by the community and possibly the local industry. Although they have the need to dispose of this waste material, they do not normally have a specific and unified energy demand other than the need for electrical power. Moreover, the municipality does not normally control the electrical power generation, distribution, or pricing other than by exercising a political influence over the local public service commission. Furthermore, the community does not normally have ready access to large sums of capital and operating funding without a serious legislative exercise.

The successful implementation of a waste-to-energy project requires a "sponsor" that will coordinate all efforts necessary to complete a project, including designing; environmental permitting; contracting with the municipality (or municipalities) for the waste; contracting with the utility for sale of the power; contracting with the waste haulers for material delivery, recyclable markets, and ash and waste disposal; and convincing a financial entity (or entities) of the viability of the overall plan and economical advantages enough to obtain sufficient investment to pursue the project. Logistically speaking, the development of a simple waste-to-energy project involves contractual negotiations with no less than five separate entities (municipality, utility, financial, contractor, environmental), none of which represents an easy task. The complexity is further exaggerated by the interdependence of one contract on another. As a result, the time involved before a project can even begin the construction effort could be as long as 5 years or more. During that period of time, municipal elections can drastically alter the receptiveness of one or more of the major partners to the project. In addition, the costs involved to develop these contracts, agreements, permits, and the like can mount into the millions of dollars. This represents an investment with no chance of recovery unless and until the project is built, operating, and paying back dividends to the investors.

A second basic obstacle to the success of waste-to-energy technology is the cost competitiveness with other options. On the one hand, the technology is a waste disposal option and, therefore, competes with basic waste disposal practices such as landfills. On the other hand, the waste-to-energy facility is an energy producer and must compete with selling electricity at a rate competitive with options such as high-efficiency combined cycle gas turbine stations, supersized coal-fired utility power stations, and perhaps hydroelectric generation. In most instances, the costs to generate new or existing

power are very low in comparison with the costs for a straight solid fuel-fired energy system, especially one sized for only 10 to 20 MW capacity. To be economically viable, the waste-to-energy system must provide a competitive tipping (or disposal) fee to the community to dispose of the waste and generate an electrical output for a cost that can compete with market rates for electricity. In many cases, geographical conditions may influence one or both of these conditions to the benefit, or detriment, of a potential waste-to-energy facility. Where land is plentiful relative to the municipal population base, the costs of landfilling and hauling may be very competitive. Where hydroelectric energy provides a large base to the utility supply or where low-priced natural gas or coal generation remains prevalent, the power rates may be too low to from a competitive standpoint.

The capital costs for waste-to-energy facilities have been reported to range from approximately $75,000 to more than $125,000 per daily ton of waste input. On that basis, the "typical" 1000-TPD facility described previously would cost approximately $75 million to $100 million. In some instances, costs exceeding $125 million have been reported for this size of facility. This is a fairly "expensive" cost for a power plant to generate only approximately 25 MW of electricity. A natural gas combined cycle power station in this size range would cost approximately $18 million to $20 million, and a biomass-fired power plant using similar technology would cost $50 million to $55 million.

Given the capital investment for such a facility, the required revenues from either waste disposal (tipping) fees or electric power sales are significant. As a bare minimum, tipping fees in the neighborhood of $40 per ton plus power rates approaching $40 per megawatt are necessary to present an economically viable project. In the event that one revenue stream is reduced due to competitive circumstances, the other revenue stream must be increased to compensate. As a very general rule of thumb, it can be estimated that roughly a $10/ton variance in the tipping fees is comparable to an $8/MW variation in the power rate. Thus, in the case where there was absolutely no value for the electrical power (a highly unlikely scenario), the required tipping fee would need to exceed $65 per ton. Conversely, in the case where the revenue for fuel disposal was nil, the power rates required for economic viability would need to exceed $80 per megawatt (or $0.08/kW).

Finally, the most significant reason why more waste-to-energy facilities have not been commissioned to date is the "NIMBY" (not in my backyard) sentiment that arises at the first sign of a proposed facility in a "new" community. The NIMBY syndrome is a very emotional reaction to the possibility of seeing a new "pollution source" spring up in one's own neighborhood. Fear of health risks, air and water contamination, traffic upheaval, noise and visual pollution, and hundreds of other disadvantages due to such a facility gives rise to an emotional reaction and spirited resistance to the development of any such project. In some cases, such emotional outcries are justified given historical experience where industry did not look out for the well-being of the local populations. Years of industrial pillaging of local communities and the consequential cynicism that has developed have created an environment where there is very little confidence in the "promises" of good neighbor practices and environmental compliance with these new systems. No matter what "facts" are represented to justify the environmental impact of a new facility, the information is viewed by the lay community as unbelievable and incorrect. Most people are more willing to embrace the worst-case scenario than to believe that they might reap any potential benefit from such a facility. With such a political uncertainty arising from these public hearings, many projects that could achieve technical and economic viability will end up being delayed or even cancelled due to the lack of community support for such projects. Unfortunately, there is no absolute black-and-white definition of the advantages and/or disadvantages of these waste-to-energy facilities. A willingness to accept the status quo (i.e., landfilling) can often become a major hurdle to siting a new facility. The irony is that the status quo may be as bad as, or worse than, the proposed facility, but because it already exists, it automatically becomes the favored solution. There are numerous studies that report the emissions and other environmental ramifications of landfills and even compare them with the new standards for waste-to-energy facilities. One such study, in fact, listed the greenhouse gas emissions from a landfill (CO_2 and methane as CO_2-equivalent) as 26 times greater than those from a modern waste-to-energy facility.

SEE ALSO THE FOLLOWING ARTICLES

Biomass Combustion • Biomass for Renewable Energy and Fuels • Biomass Gasification • Biomass

Resource Assessment • District Heating and Cooling • Exergy Analysis of Waste Emissions • Hazardous Waste from Fossil Fuels • Hybrid Energy Systems • Nuclear Waste

Further Reading

Hasselriis, F. (1987). Optimization of combustion conditions to minimize dioxin emissions. *Waste Mgmt. Res.* **5,** 311–326.

Hickman, H., Hasselriis, F., Turner, W., Hopper, R., Kuester, J., and Trezek, G. (1984). "Thermal Conversion Systems for Municipal Solid Waste." Noyes Publications, Park Ridge, NJ.

Jones, K. (1991). Risk assessment: Comparing compost and incineration alternatives. *MSW Mgmt.*, 29–39

Kiser, J. (1991). The future role of municipal waste combustion. *Waste Age*, 33–38

Mahin, D. (1991). Making steam from municipal solid waste combustors. *Solid Waste Power,* 28–35

Porteous, A. (2001). Energy from waste incineration: A state of the art emissions review with an emphasis on public acceptability. *Appl. Energy* **70,** 157–167.

Rigo, H. G. (1995). Sources of dioxin in the environment. *Solid Waste Technol.,* 36–39

Taylor, H. (1989). The ten myths of municipal waste combustion. Paper presented at the International Conference on Municipal Waste Combustion, Hollywood, FL.

Wave and Tidal Energy Conversion

GEORGE LEMONIS
Centre for Renewable Energy Sources
Attica, Greece

1. **Introduction**
2. **Wave Energy Conversion**
3. **Tidal Energy Conversion**
4. **Conclusions**

Glossary

fetch Distance over which wave-inducing wind–sea interaction takes place.
high tide Highest water level during a tidal period.
linear wave Theoretical wave form of sinusoidal shape; describes offshore wave conditions with satisfactory accuracy.
low tide Lowest water level during a tidal period.
lunar tide Ocean tide induced by the gravitational forces of the moon.
neap tide Lowest low tide during a lunar month.
solar tide Ocean tide induced by the gravitational forces of the sun.
spring tide Highest high tide during a lunar month; period of revolution of the moon around the earth (~29.5 days).
tidal range Difference in water level between consecutive high and low tides.

This article outlines the state of the art of exploitation of two renewable energy sources that are available in abundance in the world's oceans: ocean wave energy and tidal energy. It is evident that humanity could cover a significant part of its electricity demand from these energy sources. The potential global wave energy contribution to the electricity market is estimated to be on the order of 2000 TWh/year, approximately 10% of the world electricity consumption. The global tidal range energy potential is estimated to be on the order of 3000 GW, with approximately 1000 GW (~3800 TWh/year) being available at comparably shallow waters. Tidal current energy conversion technologies, which have started to be investigated during the recent past, are predicted to supply up to 48 TWh/year from sites around Europe. Furthermore, other large tidal current resources are still to be explored worldwide. Although research and development on ocean energy exploitation is under way in several countries around the world, ocean energy conversion technologies have not yet progressed to the point of massive power generation. This is due in part to the often rough and unpredictable conditions under which these technologies have to operate. However, considerable progress has been made during the past decade in various engineering fields associated with ocean energy conversion. Recent advances indicate that some technologies could meet the goal of power production at the commercial level by or before 2010. This article outlines the nature and worldwide resource of ocean wave and tidal energy, explains the basic conversion concepts, and points out the advantages and disadvantages of ocean energy conversion. The presentation of the ocean energy conversion development status is confined to brief descriptions of the most advanced technologies.

1. INTRODUCTION

Worldwide energy consumption is predicted to rise considerably over the coming decades. Many countries in the world, constantly reminded that traditional methods of energy production are contributing to serious environmental problems and in view of their commitment to the Kyoto Protocol, have seen the urgent need for pollution-free power generation. The energy sector has been forced into a renovating process that sees an opening toward renewable energy. In the dynamic evolution of the renewable

 © 2004 Elsevier Inc. All rights reserved.

energy industry, an ocean energy industry is emerging. Although the technology is relatively new and currently not economically competitive with more mature technologies (e.g., wind energy), the interest in ocean energy conversion from governments and industry is steadily increasing. An important feature of ocean energy sources is their high density, and density is highest among the renewables.

The oceans, occupying nearly three-quarters of the world's surface and the atmosphere above, intercept most of the energy from the sun—approximately 80,000 TW. This energy appears in the oceans in a variety of forms, for example, as waves generated by wind blowing over the ocean surface and as currents driven by the wind or caused by thermal gradients.

In addition, the oceans are influenced by the gravitational interactions in the planetary system of the sun, the earth, and the moon that represents an extremely large reservoir of kinetic energy. This reservoir is tapped slowly, but continuously and predictably, to provide the energy to maintain the tidal system in the oceans.

Various other sources of renewable energy are available in the oceans, for example, the vast thermal potential between the upper warm and the deep cold ocean water layers and the density gradients between water layers of different salinity.

Recent technological advances indicate that ocean wave energy and tidal energy conversion technologies could be developed on the level of commercial power production in the foreseeable future. This article describes the main features of these renewable energy sources and of the state of the art and the perspectives of their use.

2. *WAVE ENERGY CONVERSION*

2.1 Wave Energy: Physics and Resources

Among different types of ocean waves, wind-generated waves have the highest energy concentration. Wind waves are derived from the winds as they blow across the oceans (Fig. 1). This energy transfer provides natural storage of wind energy in the water near the free surface. Once created, wind waves can travel thousands of kilometers with little energy loss unless they encounter head winds. Nearer the coastline, the wave energy intensity decreases due to interaction with the seabed. Energy dissipation near shore can be compensated by natural phenomena such as refraction and reflection, leading to energy concentration ("hot spots").

Ocean waves contain two forms of energy: the kinetic energy of the water particles that generally follow circular paths (the radius of which decreases with depth) and the potential energy of elevated water particles (Fig. 2). On the average, the kinetic energy in a linear wave equals its potential energy. The energy flux in a wave, the wave power, is proportional to the square of the amplitude and to the period of the motion. The average power in long-period ($\sim$7–10 s), large-amplitude ($\sim$2 m) waves commonly exceeds 40 to 50 kW per meter width of oncoming waves.

Like most forms of renewables, wave energy is unevenly distributed over the globe. Increased wave activity is found between the latitudes of $\sim$30 and $\sim$60° in both hemispheres, induced by the prevailing western winds ("westerlies") blowing in these regions (Fig. 3). Particularly great resources are located along the western European coast, off the coasts of Canada and the United States, and along the southern coasts of Australia and South America.

Conservative estimates of the Commission of the European Communities predict a wave energy

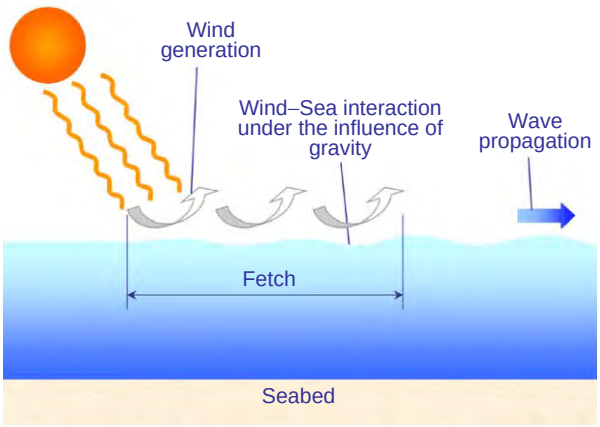


FIGURE 1 Wind-induced wave generation.

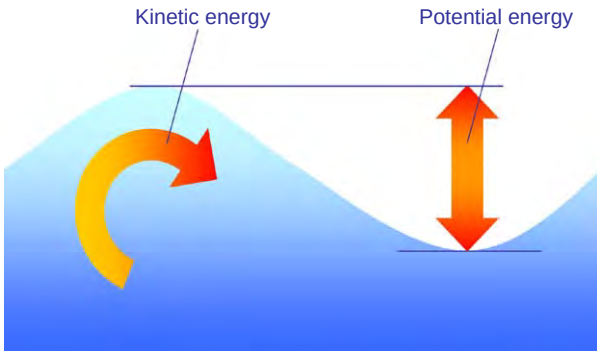


FIGURE 2 Wave energy components.

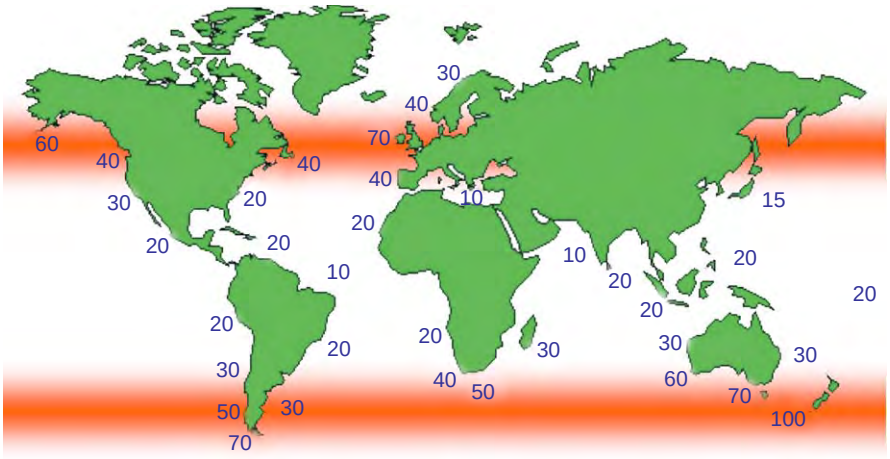


FIGURE 3 Global wave power distribution in kilowatts per meter of crest width.

market in the medium term (~2010) of approximately 5.5 TWh/year. However, the potential worldwide wave energy contribution is estimated to be on the order of 2000 TWh/year, approximately 10% of the world electricity consumption.

2.2 Wave Energy Conversion Techniques

In contrast to other renewable energy sources, the number of concepts for wave energy conversion is very large. Although more than 1000 wave energy conversion techniques are patented worldwide, the apparent large number of concepts for wave energy converters can be classified into a few basic types. The followings concepts are widely adopted:

- *The oscillating water column that consists of a partially submerged hollow structure open to the seabed below the water line* (Fig. 4). The heave motion of the sea surface alternatively pressurizes and depressurizes the air inside the structure, generating a reciprocating flow through a "Wells" turbine installed beneath the roof of the device. This type of turbine is capable of maintaining constant direction of revolution despite the direction of the airflow passing through it.
- *Overtopping devices that collect the water of incident waves to drive one or more low head turbines* (Fig. 5).
- *Heaving devices (floating or submerged) that provide a heave motion that is converted by mechanical and/or hydraulic systems in linear or rotational motion for driving electrical generators* (Fig. 6).
- *Pitching devices that consist of a number of floating bodies, hinged together across their beams* (Fig. 7). The relative motions between the floating

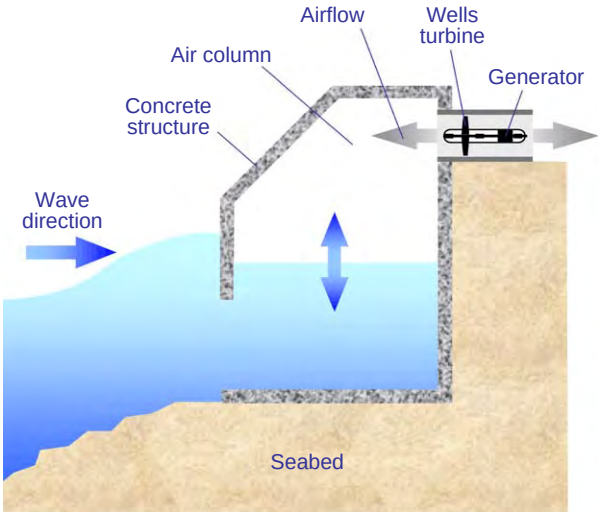


FIGURE 4 Operating concept of an oscillating water column.

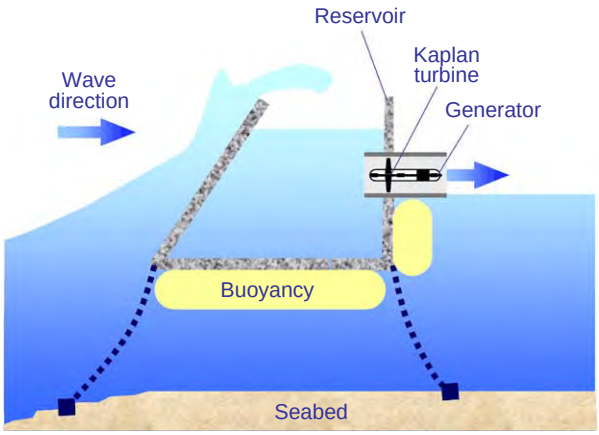


FIGURE 5 Operating concept of an overtopping device (floating type).

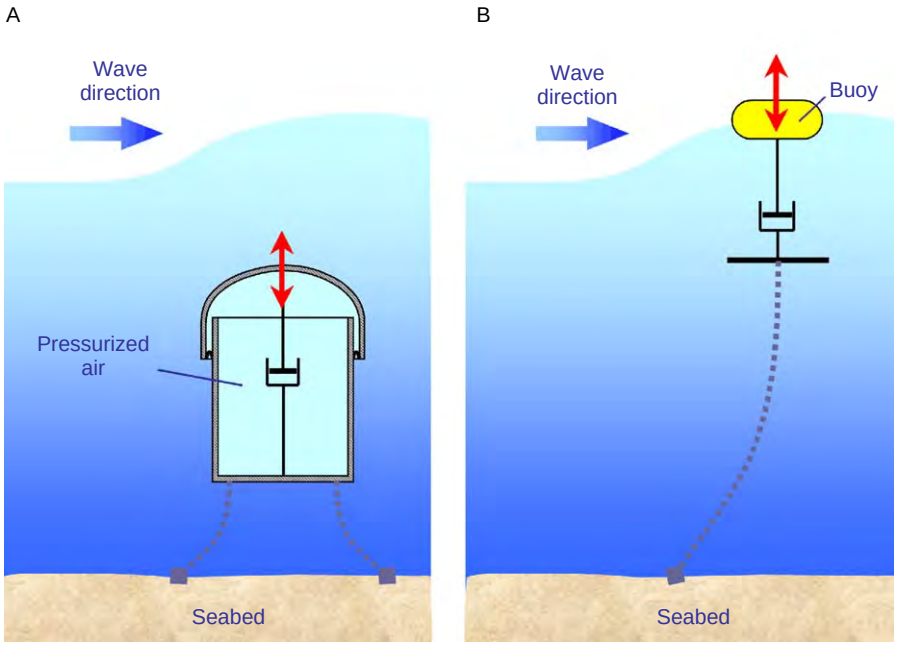


FIGURE 6 Operating concepts of heaving devices: (A) Archimedes Wave Swing concept, adapted from Rademakers *et al.* (1998), and (B) floating buoy converter.

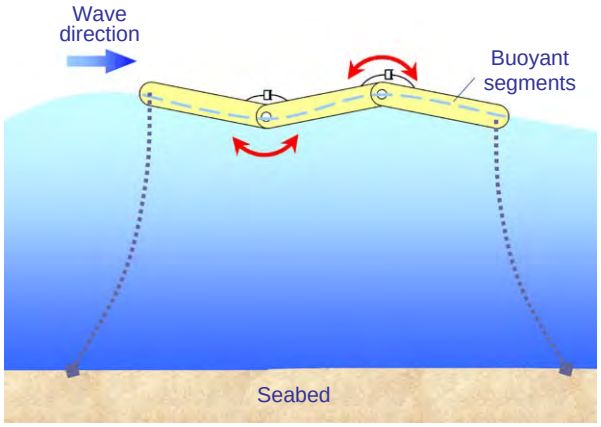


FIGURE 7 Operating concept of a pitching device (Pelamis concept). From Yemm (1999).

bodies are used to pump high-pressure oil through hydraulic motors that drive electrical generators.

• *Surging devices that exploit the horizontal particle velocity in a wave to drive a deflector or to generate a pumping effect of a flexible bag facing the wave front* (Fig. 8).

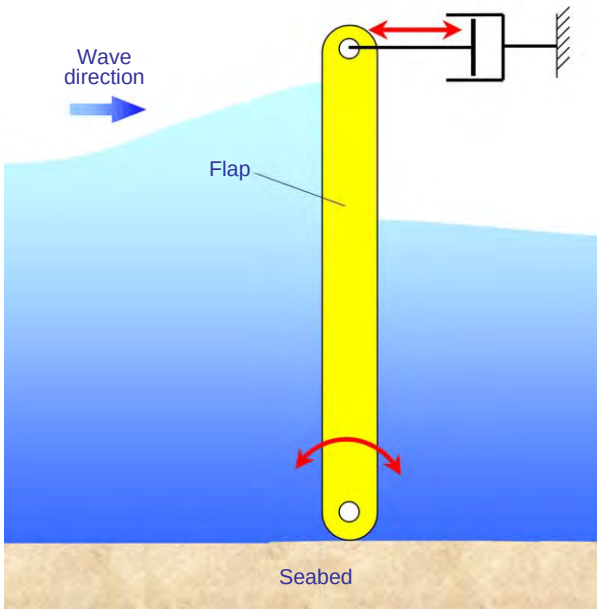


FIGURE 8 Operating concept of a surging device.

2.3 Advantages and Disadvantages of Wave Energy Conversion

It is important to appreciate the difficulties facing wave power developments, the most important of which are the following:

• Because of irregularity in wave amplitude, phase, and direction, it is difficult to obtain maximum efficiency of a device over the entire range of excitation frequencies.

- The structural loading in the event of extreme weather conditions, such as hurricanes, may be as high as 100 times the average loading.
- The coupling of the irregular slow motion (frequency ~ 0.1 Hz) of a wave to electrical generators typically requires ~ 500 times greater frequency.

Obviously, the design of a wave power converter has to be highly sophisticated to be operationally efficient and reliable, on the one hand, and economically feasible, on the other. As with all renewables, the available resource and variability at the installation site have to be determined first. The preceding constraints imply comparably high construction costs and possibly reduced survivability. These factors, together with misinformation and lack of understanding of wave energy by the industry, government, and public, have often slowed down wave energy development.

One of the important advantages of wave energy conversion is its environmental compatibility given that wave energy conversion technologies are generally considered to be free of polluting gas emissions. The abundant resource and the high-energy fluxes in the waves prescribe, at appropriate design of the devices, economically viable energy production. Other advantages of wave energy include its natural seasonal variability, which follows the electricity demand in temperate climates, and the introduction of synchronous generators for reactive power control. The negligible demand of land use is an important aspect, followed by the current trends of offshore wind energy exploitation. As for most renewables, the *in situ* exploitation of wave energy implies diversification of employment and security of energy supply in remote regions. Furthermore, the large-scale implementation of wave power technologies will stimulate declining industries (e.g., shipyards) and promote job creation in small and medium-size enterprises.

2.4 Wave Energy Conversion Development Status

Wave energy conversion is being investigated in a number of countries, particularly in the European Union member states, China, India, Japan, Russia, and the United States. Although the first patent certificate on wave energy conversion was issued in 1799, the intensive research and development (R&D) study of wave energy conversion began after the dramatic increase in oil prices in 1973.

During the past 5 years or so, there has been a resurgent interest in wave energy, especially in Europe. Nascent wave energy companies have been highly involved in the development of new wave energy schemes such as the Pelamis, the Archimedes Wave Swing, or the Limpet. Currently, the world-installed capacity is approximately 2 MW, mainly from pilot and demonstration projects.

The electricity generating costs from wave energy converters have shown significant improvement during the past 20 years or so and have reached an average price of approximately 0.08 €/kWh at a discount rate of 8%. Compared with the average electricity price in the European Union (~ 0.04 €/kWh), the electricity price produced from wave energy is still high, but it is forecasted to decrease further with the development of the technologies.

Although early programs for R&D on wave energy considered designs of several megawatts output power, more recent designs are rated at power levels ranging from a few kilowatts up to 2 to 4 MW. Massive power production can be achieved by interconnection of large numbers of devices.

The amount of ongoing development work on wave energy technologies is very large and cannot be described adequately in a single article such as this. In the framework of this article, the promising wave energy device developments are described briefly. More detailed reviews can be found in the extensive literature in this area.

The Archimedes Wave Swing, developed by Teamwork Technology BV in The Netherlands, consists of a hollow pressurized structure, the upper part of which is initiated to heave motions by the periodic changing of hydrostatic pressure beneath a wave. A 2-MW pilot plant has been constructed and was waiting to be deployed off the Portuguese coast during the spring of 2003.

The Energetech Oscillating Water Column was developed by Energetech Ltd. in Australia. The device uses a novel variable-pitch turbine and a parabolic wall to focus the wave energy on the device collector. This scheme has received a power purchase agreement with the local utility at Port Kembla, 80 km south of Sydney, for a 500-kW plant.

The European Pilot Plant on Pico Island in the Azores is a 400-kW oscillating water column developed by a European team coordinated by the Instituto Superior Técnico (Portugal), which consists of six Portuguese partners and two partners from the United Kingdom and Ireland. The plant was designed as full-scale testing facility. It is fully automated and supplies a sizable part of the island's energy demand.

The Floating Wave Power Vessel is a floating overtopping device for offshore operation developed by Sea Power International in Sweden. It consists of a floating basin supported by ballast tanks in four sections. A patented anchor system allows the orientation of the vessel to the most energetic wave direction. A pilot plant was developed and deployed during the 1980s near Stockholm, and the company has signed contracts in various countries for commercial deployment of the technology.

The LIMPET is a 500-kW oscillating water column developed by the Queen's University of Belfast and Wavegen Ltd. in the United Kingdom. A 75-kW prototype was constructed on the island of Islay in Scotland in 1991. The LIMPET is the successor of this prototype, intended to address many of the issues currently hindering the full-scale deployment of oscillating water column devices.

The McCabe Wave Pump consists of three rectangular steel pontoons that are hinged together across their beam. The bow of the fore pontoon is slack moored, and two more slack moorings are attached part of the way down the aft pontoon. This allows the system to vary its alignment so as to head into the oncoming seas. A 40-m long prototype was deployed in 1996 off the coast of Kilbaha in Ireland.

The Mighty Wale is an offshore device, based on the oscillating water column concept, that was developed by the Japan Marine Science & Technology Center. A 120-kW prototype with three oscillating water columns in a row has been operating 1.5 km off Nansei Town at 40 m depth since 1998. The mooring system is designed to withstand wind–wave conditions resulting from a "once in 50 years" storm.

The OPT Wave Energy Converter, developed by Ocean Power Technology in the United States, consists of a 2- to 5-m diameter buoy-type cylinder closed at the top and open to the sea at the bottom. A hydraulic ram is positioned between the top of the shell and a highly buoyant steel float contained within the shell. The relative motion of the shell to the buoyant float activates a hydraulic system to pump oil at high pressure to a generator. Extensive tests on a large scale in the eastern Atlantic have been concluded, and as of this writing, the first commercial schemes were about to be built in Australia and in the Pacific. The individual converters are rated at between 20 and 50 kW, intended to meet multimegawatt demands using arrays.

The Pelamis device is a semisubmerged articulated structure composed of cylindrical sections linked by hinged joints. The wave-induced motion of these joints is resisted by hydraulic rams that pump high-pressure oil through hydraulic motors via smoothing accumulators. A 130-m long and 3.5-m diameter device rated at 375 kW is being developed by Ocean Power Delivery Ltd.–OPD in Scotland.

The Pendulor is a surging-type converter consisting of a rectangular box that is open to the sea at one end. A pendulum flap is hinged over this opening, so that the actions of the waves cause it to swing back and forth. This motion is then used to power a hydraulic pump and generator. Several schemes (≥ 5 kW) have been built in Japan, and there are plans to develop a larger plant.

The Point Absorber Wave Energy Converter, developed by Rambøll in Denmark, consists of a float connected to a suction cup anchor by a polyester rope. The relative motion between the float on the sea surface and the seabed structure activates a piston pump (actuator) inserted between the rope and the float. A 1:10 scale model was tested at sea at the Danish test site "Nissum Bredning" over a period of 3 months, and a 1:4 scale model with a 2.5-m diameter is currently being developed for open sea testing.

The concept of the Salter Duck was introduced in 1974 by S. Salter. An important feature of this device is the capability of converting both the kinetic and potential energies of the wave, thereby achieving very high absorption efficiencies (theoretically $>90\%$). The system has undergone considerable development since 1983 and was redesigned in 1993. The current design is characterized by high availability and overall efficiency and low energy production costs.

The Wave Dragon is an offshore wave energy converter of the overtopping type developed by a group of companies coordinated by Wave Dragon ApS in Denmark. It uses a patented wave reflector design to focus the wave toward a ramp and to fill a higher level reservoir. Electricity is produced by a set of low-head Kaplan turbines. The scheme has been tested in the laboratory on a 1:50 scale and on a 1:3.5 scale model turbine. A 57-m wide prototype weighing 261 tons on a 1:4.5 scale rated in full scale to 4 MW is being deployed in Nissum Bredning in Denmark.

3. TIDAL ENERGY CONVERSION

3.1 Tidal Energy: Physics and Resources

Tidal energy conversion techniques exploit the natural rise and fall of the level of the oceans caused

principally by the interaction of the gravitational fields in the planetary system of the earth, the sun, and the moon. The main periods of these tides are diurnal at approximately 24 h and semidiurnal at 12 h 25 min. This motion is influenced by the positions of the three planets with respect to each other, and these positions vary during the year. Spring tides occur when the tide-generating forces of the sun and moon are acting in the same directions. In this situation, the lunar tide is superimposed to the solar tide (Fig. 9). Some coastlines, particularly estuaries, accentuate this effect, creating tidal ranges of up to approximately 17 m. Neap tides occur when the tide-generating forces of the sun and moon are acting at right angles to each other.

The vertical water movements associated with the rise and fall of the tides are accompanied by roughly horizontal water motions termed "tidal currents." Therefore, a distinction should be made between tidal range energy (the potential energy of the water at high tide) and tidal current energy (the kinetic energy of the water particles in horizontal direction in a tide).

Tidal currents have the same periodicities as do the vertical oscillations and so are predictable, but they tend to follow an elliptical path and do not normally involve a simple to-and-fro motion. Where tidal currents are channeled through constraining topography, such as in straits between islands, in shallows between open seas, and around the ends of headlands, very high water particle velocities can occur. These relatively rapid tidal currents typically have peak velocities during spring tides in the neighborhood of 2 to 3 ms^{-1} or more.

Currents are also generated by the winds as well as temperature and salinity differences. The term "marine currents," often found in literature, encompasses several types of ocean currents. Wind-driven currents affect the water at the top of the oceans, from the surface down to approximately 600 to 800 m. Currents caused by thermal gradients and salinity differences are normally slow deep water currents that begin in the icy waters around the north polar ice. Wind-driven currents appear to be less suitable for power generation than do tidal currents because the former are generally slower. Moreover, tidal currents usually exhibit their maximum speed in comparably shallow waters accessible for large engineering works.

The global tidal range energy potential is estimated to be 3000 GW, with approximately 1000 GW being available in comparably shallow waters. However, the tidal energy resource that could realistically be developed (tidal ranges of 4.5–5.0 m or higher) is confined to a few regions with exceptional tidal ranges (Fig. 10). Within the European Union, France and the United Kingdom have sufficiently high tidal ranges of up to approximately 10 m. Beyond the European Union, Canada, the Commonwealth of Independent States (CIS), Argentina, Western Australia, and Korea have potentially interesting sites that have been investigated periodically.

Recent studies indicate that marine currents have the potential to supply a significant fraction of future electricity needs. Tecnomare SpA and IT Power Ltd. studied 106 locations in European waters that were predicted to be suitable for current energy exploitation. The potential of these sites resulted in an

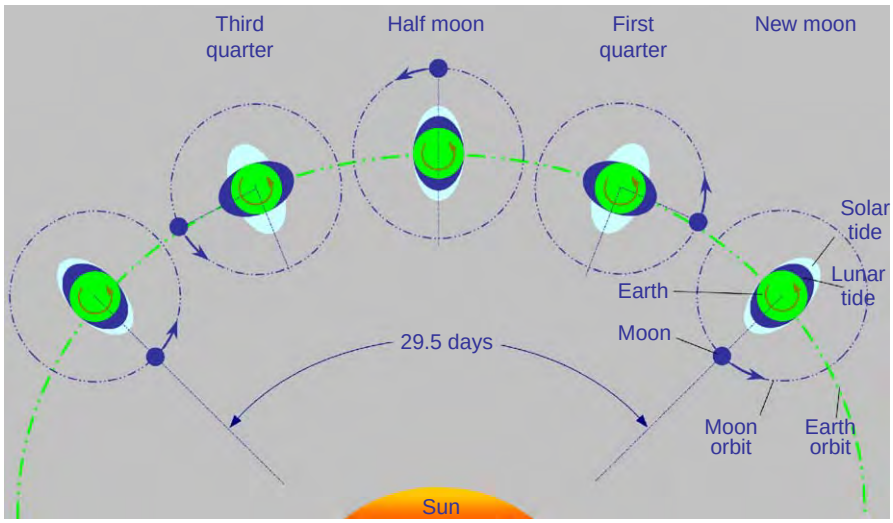


FIGURE 9 Interaction of lunar and solar tides during a lunar month.

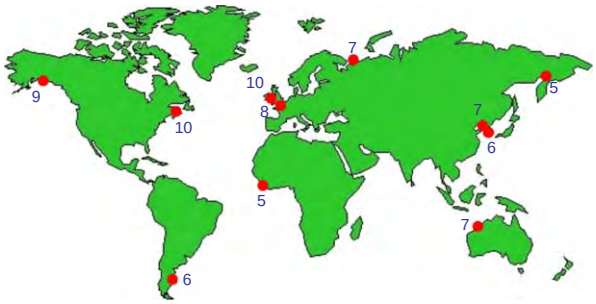


FIGURE 10 Regions of exceptional average tidal range.

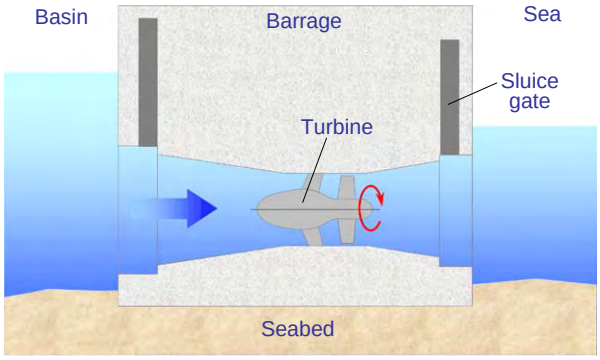


FIGURE 11 Operating concept of a tidal barrage.

installed capacity of marine current turbines of more than 12,000 MW. Locations with particularly intense currents are found around the British Isles and Ireland, between the Channel Islands and France, in the Straits of Messina between Italy and Sicily, and in various channels between the Greek islands in the Aegean. Other large marine current resources can be found in regions such as Southeast Asia, both the east and west coasts of Canada, and many other places around the globe that require further investigation.

3.2 Tidal Energy Conversion Techniques

3.2.1 Tidal Range Energy

The technology required to convert tidal range energy into electricity is very similar to the technology used in traditional hydroelectric power plants. The first requirement is a dam or "barrage" across a tidal bay or an estuary. Building dams is an expensive process. Therefore, the best tidal sites are those where a bay has a narrow opening, thereby reducing the necessary dam length. At certain points along the dam, gates and turbines are installed. When there is an adequate difference in the elevation of the water on the different sides of the barrage, the gates are opened. This "hydrostatic head" that is created causes water to flow through the turbines, turning an electric generator to produce electricity (Fig. 11).

Electricity can be generated by water flowing both into and out of a basin. Because there are two high tides and two low tides each day, electrical generation from tidal power plants is characterized by periods of maximum generation every 12 h, with no electricity generation at the 6-h mark in between. Alternatively, the turbines can be used as pumps to pump extra water into the basin behind the barrage during periods of low electricity demand. This water can then be released during the day periods with the greatest electricity demand, thereby allowing the tidal plant to function with some of the characteristics of a "pumped storage" hydroelectric facility.

A recently proposed technology to exploit tidal range energy, termed "offshore tidal conversion," relies on the use of impoundment structures placed offshore on shallow flats with large tidal ranges. Rather than blocking an estuary, impoundment structures would be completely independent of the shoreline. Such structures, which could be built of economical rubble mound construction materials, would resolve many of the environmental and economic problems of shoreline tidal barrages.

3.2.2 Tidal Current Energy

Tidal currents can be harnessed using technologies similar to those used for wind energy conversion, that is, turbines of horizontal (Fig. 12) or vertical (Fig. 13) axis ("cross-flow" turbines). Because the density of water is some 850 times higher than that of air, the power intensity in water currents is significantly higher than that in airflows. Consequently, a water current turbine can be built considerably smaller than an equivalent-powered wind turbine.

A novel technique recently developed by Engineering Business Ltd. in the United Kingdom uses a hydrofoil, which has its attack angle relative to the onset water flow varied by a simple mechanism (Fig. 14). This causes the supporting arm to oscillate, which in turn forces hydraulic cylinders to extend and retract. This produces high-pressure oil, which is used to drive a generator. A more detailed description of the technique is provided in the next section.

Other techniques for tidal current energy conversion have been proposed in the past, for example, the generation of vortices by submerged delta wings to drive horizontal axis turbines.

FIGURE 12 Horizontal axis tidal current turbine. Reprinted with the permission of Marine Turbines Ltd.

FIGURE 13 Vertical axis tidal current turbine (floating platform). Reprinted with the permission of Ponte di Archimede nello Stretto di Messina SpA.

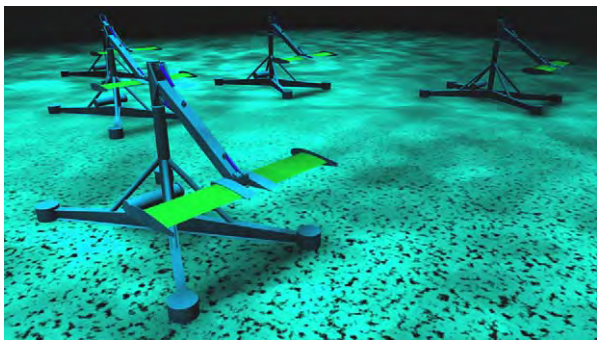


FIGURE 14 Stingray generator. Reprinted with the permission of Engineering Business Ltd.

3.3 Advantages and Disadvantages of Tidal Energy Conversion

3.3.1 Tidal Range Energy

Tidal barrage technology is considered mature, but as with all large civil engineering projects, there would be a series of technical and environmental risks to address. The main environmental risks are associated with the changes in water levels that would modify currents and sediment transport and deposit. However, there are regional development benefits as well. For example, the La Rance plant in France, the only commercial tidal range conversion scheme so far, includes a road crossing linking two previously isolated communities and has allowed further development of the distribution network for raw materials and developed products.

Tidal barrage projects normally require high capital investment at the outset and so have relatively long construction periods and low load factors, leading to long payback periods. As a consequence, the electricity cost is highly sensitive to the discount rate used. Therefore, access to suitable funding is a serious problem and is unlikely to be achieved without public intervention.

Offshore tidal technology is expected to be associated with less environmental damage than are traditional shoreline barrages. Furthermore, offshore tidal schemes are predicted to be more economical than tidal barrages. The structures would be built in a similar way to rubble mound breakwaters using conventional inexpensive materials such as sand, loose rock, and gravel. Their load factors would be higher than those of barrages. In contrast to barrages, which must generate primarily in one direction (on the ebb tide) to minimize progressive disruption of the intertidal zone, offshore tidal converters would be free to use both the ebb and flood tides for generation.

3.3.2 Tidal Current Energy

Several types of tidal current conversion devices, particularly fully submerged devices, are subject to the corrosive effects of seawater. This leads to high construction costs because it becomes necessary to use corrosion-resistant materials. In addition, maintenance becomes difficult because divers are needed to access submerged machinery. Although placing the generator above water can minimize the need for divers, maintenance costs would still be higher than, for example, in wind turbines. The installation of large schemes could also be confronted with major technical problems, safe and economical methods of deployment, and recovery requiring further investigation.

One specific advantage of tidal current conversion technologies is their limited environmental impact. This is considered negligible because their operation is not associated with polluting gas emissions. Their installation requires only minimal land use, and fully submerged devices will not affect their surroundings optically or acoustically. Their effects on flora and fauna have not been studied extensively, but it is unlikely that they would be of significance. In contrast to atmospheric airflows, the availability of tidal currents can be predicted very accurately because their motion will be tuned with the local tidal conditions. The power density from flowing water is much higher than that from most other forms of renewable energy. Thus, marine current converters could be built considerably smaller than equivalent machines, for example, those for wind energy conversion. Finally, submerged marine current converters are considered to operate in a safe environment. It is evident that disturbances caused by extreme weather conditions are significantly attenuated to the depths of approximately 20 to 30 m where the devices will normally operate.

3.4 Tidal Energy Conversion Development Status

3.4.1 Tidal Range Energy

The first large-scale, commercial tidal range energy plant was built on the La Rance estuary in France during the 1960s and has now operated successfully for more than 30 years. The La Rance station is still the only industrial-sized tidal power station in the world. Its 240-MW power is roughly one-fifth that of an EDF (Electricité de France) nuclear reactor and is more than 10 times that of the largest other tidal stations in the world. The barrage was designed for two-way operation. Today, it is operated nearly exclusively in ebb generation mode. The net annual energy yield of the 10 horizontal Kaplan turbines averages 544 GWh.

The good performance of La Rance has resulted in examination of several other projects in France, particularly in La Somme and Saint-Brieux bays, in Brest roadstead, along the Cotentin peninsula west coast, and a 200-square-kilometer area between Grandville, Chausey islands, and Coutainville. These projects were finally abandoned due to their very high investment costs and concerns over ecological impact. The Russians built a small 400-kW device near Murmansk that was later followed by a 17.4-MW experimental device built by the Canadians at Annapolis. A series of small plants have been installed in China. In the United Kingdom, a series of industrial consortia in collaboration with the government have investigated the prospects for tidal energy on the Severn, Mersey, and a number of smaller estuaries. None of these schemes has progressed to full-scale development.

The comparably high generation costs and long payback periods of tidal range schemes imply that within deregulated electricity markets, which are based on private investment, tidal energy is unlikely to be commercially developed if the kilowatt-hour price does not become competitive with cost-effective renewable energies.

The generation costs of offshore tidal energy are expected to be lower than those of tidal barrage schemes. Offshore tidal energy is currently being investigated in the United Kingdom. Three projects are in development, with two of them having a capacity of 30 MW and one having a capacity of 432 MW.

3.4.2 Tidal Current Energy

Tidal current technology is in its infancy. However, recent developments in this area open up prospects for commercial deployment of some schemes in the near future. The economical viability of these schemes has not yet been proven, with the estimated production costs typically ranging from approximately 0.08 to 0.35 €/kWh. However, it is anticipated that the production costs will decrease as the technologies advance.

Currently, different pilot plants are in operation or about to be installed, mainly in Europe. The devices rely on the horizontal or vertical axis turbine concepts and on the previously outlined oscillating hydrofoil technique. The most promising schemes are described in what follows.

The SEAFLOW device (Fig. 12), developed by Marine Current Turbines Ltd. in the United Kingdom, consists of an axial flow rotor of approximately

15 m in diameter, rated at approximately 300 kW in a current velocity of roughly 2.0 to 2.5 ms^{-1}. The system will be installed on a mono-pile, socketed to the seabed at a depth of approximately 30 m. This project is expected to lead directly to a second phase involving the development of 5 to 10 larger turbines, approximately 1 MW each. The SEAFLOW device is the successor of a 15-kW prototype with a 3.5-m rotor diameter that was tested successfully in 1994.

A similar device developed by SINTEF Energy Research in Norway is aimed to be the first tidal current converter that produces power for the grid. The turbine has a 25-m high tower with a 20-m diameter rotor, although the tower is capable of supporting a 30-m rotor. The turbine design allows the pitch of the blades to be changed to optimize power transfer from the tidal currents. When the tide turns, the blades are turned 180° and the direction of rotation is reversed so that both current directions can be used. The prototype was expected to undergo further development until 2004. If the pilot project produces the results that its owners are seeking, they owners will proceed with the installation of a 20-turbine plant in the Kvalsund Strait.

The ENERMAR device (Fig. 13), developed by Ponte di Archimede nello Stretto di Messina SpA in Italy, employs a patented vertical axis rotor termed "Kobold" connected to a synchronous generator. The rotor has an outer diameter of 6 m and consists of three blades with a span of 5 m each and a chord length of 0.4 m. An important feature of the device is that the direction of rotation of the rotor is independent of the current direction. The system is mounted on a floating platform 10 m in diameter. In 2002, a prototype was deployed in the Strait of Messina 150 m offshore at a depth of 20 m. Tests performed on the prototype indicate that the turbine produces approximately 20 kW of power at a current speed of 1.8 m/s, indicating a global system efficiency of 23%. In a current of 3.2 m/s, 150 kW is expected.

The Stingray Generator (Fig. 14), developed at Engineering Business Ltd. in the United Kingdom, uses a novel technique patented by the company during the late 1990s. The device produces electricity using the oscillatory movement of hydroplanes driven by flowing water. The hydroplanes rise and fall as the angle of each hydroplane is altered at the end of each stroke. A hydraulic cylinder is connected to the hydroplane arm, and the up-and-down movement generates high-pressure oil that is used to drive a hydraulic motor, which in turn drives an electric generator. A full-scale prototype rated at 150 kW was deployed in September 2002 at Yell Sound at a depth of 36 m and has since been in operation. A simultaneous program, aiming to start installation of a 5-MW demonstration farm with connection to the local distribution network, was expected to commence in the summer of 2004.

4. CONCLUSIONS

The sea is a colossal reservoir of energy of particularly high density—the highest among the renewables. The use of this energy resource could cover a significant part of the electricity demand worldwide. Currently, the exploitation of two forms of ocean energy, namely ocean wave energy and tidal energy, is considered prospective for massive power generation in the near future.

Although R&D on ocean energy conversion is under way in several countries around the world, ocean energy technologies have not yet been developed on an industrial level. However, considerable progress has been made in this sector over the past decade or so in Asia, Australia, Europe, and North America. For some technologies, this has resulted in approaching commercialization, whereas other technologies require further R&D.

Wave energy conversion is characterized by a large number of techniques. The environmental impact of most of them is considered negligible. Different schemes, mainly oscillating water columns and heaving, pitching, and overtopping devices, could meet the goal of electricity production on a commercial scale by 2010.

The concepts for tidal energy conversion can be classified into techniques converting the potential energy of the tides and techniques exploiting the kinetic energy of tidal currents. The traditional technique of exploiting the tidal head by using dams built across bays or estuaries still remains costly, having a major environmental impact. However, offshore impoundment structures could resolve many of the problems of shoreline tidal range energy conversion. The progress in tidal current conversion over the past years has been considerable. Recently, developed techniques are being considered environmentally compatible and are predicted to become economically viable in the medium term.

The potential for improvement of the technoeconomical performance of ocean power conversion technologies is large. After their integration and their penetration into the world electricity market, ocean energy conversion technologies are expected to make

substantial contributions to the achievement of a wide range of objectives of environmental, social, and economic policies in many countries around the world, providing superior services to their populations.

SEE ALSO THE FOLLOWING ARTICLES

Hydropower Resources • Hydropower Technology • Ocean, Energy Flows in • Ocean Thermal Energy • Tidal Energy • Wind Energy Technology, Environmental Impacts of • Wind Resource Base

Further Reading

Clément, A. H., McCullen, P., Falcão, A., Fiorentino, A., Gardner, F., Hammarlund, K., Lemonis, G., Lewis, T., Nielsen, K., Petroncini, S., Pontes, M-T., Schild, P., Sjöström, B. O., Sørensen, H. C., and Thorpe, T. (2002). Wave energy in Europe: Current status and perspectives. *Renewable Sustainable Energy Rev.* **6,** 405–431.

Fraenkel, P. L., Clutterbuck, P., Stjernstrom, B., and Bard, J. (1998). "Seaflow: Preparing for the world's first pilot project for the exploitation of marine currents at a commercial scale." Paper presented at the Third European Wave Energy Conference, Patras, Greece.

Lewis, T. (1985). "Wave Energy: Evaluation for C.E.C." EUR9827EN.

Petroncini, S. (2000). Introducing wave energy into the renewable energy marketplace. Master's thesis, University of Edinburgh.

Rademakers, L. W. M. M., van Schie, R. G., Schuttema, R., Vriesema, B., and Gardner, F. (1998). "Physical model testing for characterizing the AWS." Paper presented at the Third European Wave Energy Conference, Patras, Greece.

Ross, D. (1995). "Power from the Waves." Oxford University Press, Oxford, UK.

Salter, S. H. (1989). World progress in wave energy—1988. *Intl. J. Ambient Energy* **10**(1).

Tecnomare SpA & IT Power Ltd. (1995). "Marine Currents Energy Extraction." EUR16683EN.

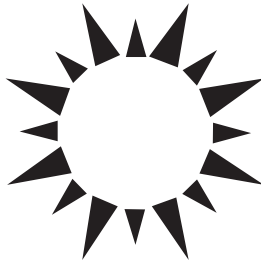
Thorpe, T. W. (1992). "A Review of Wave Energy." ETSU-R-72.

Thorpe, T. W. (1999). An overview of wave energy technologies: Status, performance, and costs. Paper presented at IMECHE seminar, "Wave Power: Moving towards Commercial Viability," London.

Tidal Electric. (2002). Power comes in. *Intl. Water Power Dam Constr.*, 24–27.

Yemm, R. (1999). The history and status of the Pelamis Wave Energy Converter. Paper presented at IMECHE seminar, "Wave Power: Moving towards Commercial Viability," London.

Wetlands: Impacts of Energy Development in the Mississippi Delta



JAE-YOUNG KO and JOHN W. DAY
Louisiana State University
Baton Rouge, Louisiana, United States

1. **Introduction**
2. **Impacts on Plant Physiology**
3. **Impacts on Estuarine Consumers**
4. **Hydrological Disturbance and Land Loss**
5. **Conclusion**

Glossary

accretion The upward growth of a sedimentary deposit due to settling of suspended sediments or *in situ* organic soil formation (due to root growth).

benthos Aquatic bottom-dwelling organisms; benthic is the adjective form (e.g., benthic animals).

bioremediation The addition of materials to a contaminated environment to cause an acceleration of the natural biodegradation processes.

estuary Any semi-enclosed coastal body of water that has a connection with the open sea and within which fresh water and salt water mix.

impoundment An enclosed wetland that is hydrologically isolated, either completely or partially, from the surrounding ecosystem; impoundment can be due to a combination of natural (e.g., beach ridge, natural levee ridge) and anthropogenic (e.g., road embankment, spoil bank) features.

levee A linear dike or embankment built along the bank of a channel, often to prevent inundation of lowlands by flooding.

primary production Total amount of organic matter newly formed during photosynthesis.

produced water By-product of the oil production process that is brought to the surface along with petroleum; produced water often contains high levels of salt, heavy metals, and hydrocarbons.

relative sea level rise A rise in sea level relative to that of the land due to a combination of eustatic sea level rise (i.e., worldwide sea level rise due to global warming) and sinking of the land as a result of subsidence.

subsidence Sinking or settling of the land surface due to natural (compaction and consolidation of sediments) or artificial causes.

The Mississippi Delta encompasses the largest area of coastal wetlands in the United States and supports one of the most extensive developments of petroleum extraction of any coastal area in the world. This area has experienced ecological impacts from energy development related human activities since the early 1900s. The Louisiana coastal zone encompasses approximately 3.8 million ha (9.5 million acres). The zone includes water bodies, marsh (fresh, intermediate, brackish, and salt), forested wetlands, submerged aquatic vegetation, mudflats, beaches, and upland habitats on natural levees with forests, agriculture, and urban development. Marshes make up approximately 63% of the land area in the coastal zone, and coastal Louisiana contains approximately 60% of the estuaries and marshes in the Gulf of Mexico. Coastal wetlands are vital for protecting developed areas from storm surges, providing wildlife and fish habitat, and improving water quality. The coastal zone has experienced multiple ecological impacts due to human activities, including leveeing of the Mississippi River, large-scale wetland reclamation, water quality deterioration, pollution, and widespread disruption of hydrology. Oil and gas development has contributed significantly to these impacts.

 © 2004 Elsevier Inc. All rights reserved.

1. INTRODUCTION

Historically, Louisiana has been the second most important oil- and gas-producing state behind only Alaska. Crude petroleum is a complex mixture of mainly hydrocarbons and organic compounds of sulfur, nitrogen, and oxygen. Geologically, organic matter that accumulated in sandstones, siltstones, and shales during the Cenozoic era was transformed into petroleum by heat and pressure. The northern coast of the Gulf of Mexico had a thermal regime favorable to optimal maturation of organic matter into hydrocarbons and formed stratigraphic traps through faulting and salt movements. In Louisiana, oil and gas are produced from formations of the Paleocene through Pleistocene eras in the subsurface coastal and offshore areas. Oil production in Louisiana began in 1902, and the first oil production in the coastal zone occurred in 1926. The coastal zone produced more than 50% of oil production in the state during the 1950s and reached a peak in 1970 with 513 million barrels. From the 1920s to the 1980s, 58% of the state's total oil production and 47% of the state's natural gas production were in the Louisiana coastal zone. Gas production in the coastal zone peaked in 1969 at 7.8 trillion cubic feet. By 1990, there were more than 500 oil and gas fields in the Louisiana coastal zone. By 1987, more than 13,000 state leases for oil and gas development had been issued, and more than half of the leases were located in the coastal zone. Approximately 25% of domestic crude oil and 33% of domestic natural gas of the nation flow through Louisiana's coastal marshes. More than $12 billion in revenues from leases and production in the coastal zone was collected from 1926 to 1983. In 2000, the revenue was $354 million for mineral royalties only, and approximately 1.8 million jobs in Louisiana were related to the oil industry. Fully 40% of the U.S. refining capacity is located within the coastal zone in the Gulf of Mexico region. Therefore, the risk of ecological damages in the Mississippi Delta from energy development has been high (Fig. 1).

Oil and gas development activities have had multiple ecological impacts on wetlands and coastal ecosystems through the various stages of oil and gas development, including oil exploration, drilling site access, site preparation, drilling, production, pipeline installation, spills and cleanup, and site closure (Table I). Wetland ecosystems are susceptible to oil- and gas-related activities for a number of reasons. First, the high productivity of wetland vegetation is dependent on natural hydrological flows and inputs of nutrients and sediments to the Mississippi Delta. Second, levees, canals, and impoundment disrupt the natural hydrological regime in the Mississippi Delta and, in turn, affect plant health and sediment dynamics. Third, subsidence due to depressurization from oil and gas production enhances subsidence. Fourth, pipelines for transporting oil and gas

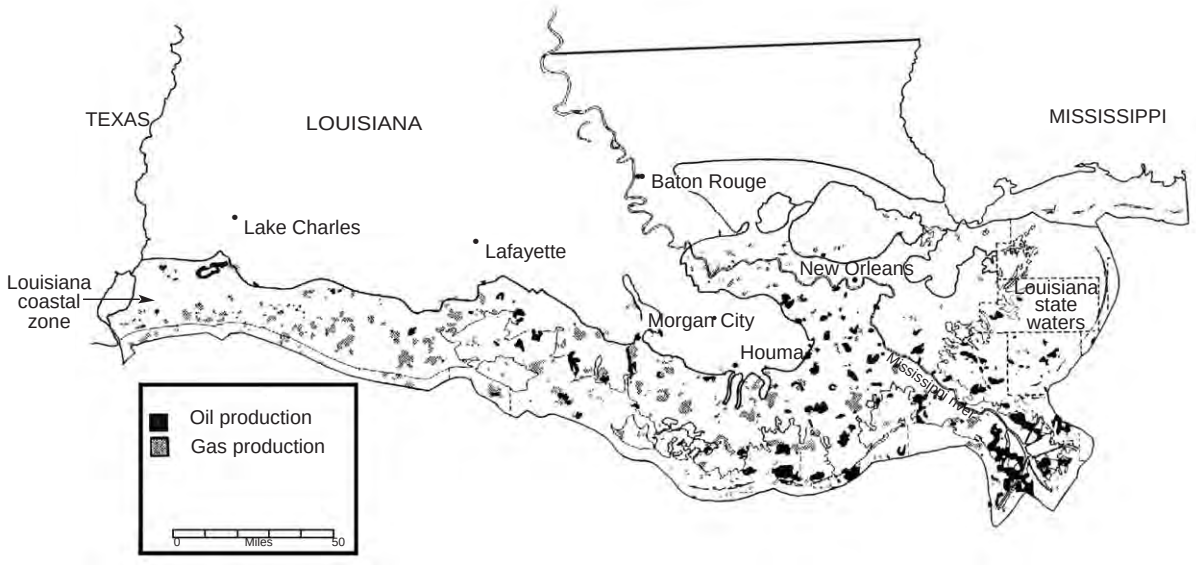


FIGURE 1 The distribution of oil and gas fields in the Louisiana coastal zone as of 1981. Reprinted from Lindstedt *et al.* (1991).

TABLE I

Multiple Impacts of Energy Development inside Wetlands

Stage	Short-term impacts	Long-term impacts
Exploration	Changes in marsh surface elevation	Changes in surface hydrology and drainage
	Breaks in natural hydrological barriers	Saltwater intrusion
	Noise and commotion during exploration	Changes in plant growth, organic matter accumulation, and sediment distribution
	Immediate loss of marsh vegetation	
	Destruction of biota	
	Changed habitat	
	New migration pattern of aquatic organisms inside marsh	
Access to site	Direct conversion to open water	Increased wave action
	Direct conversion to spoil bank habitat	Changed water circulation and turnover; stagnant water
	Return of nutrients and toxins to marsh	Dredged canal deeper than natural channel
	Noise and commotion during construction	Intercepted freshwater flow
	Immediate loss of marsh/shallow water habitat	Saltwater intrusion
	Changes in soil/water chemistry	Increased drainage of marsh
	Destruction of biota	Changes in surface hydrology and drainage
	Potential for interrupting fish spawning and feeding	Changes in sediment distribution
	Potential negative impact on plant growth	Changes in interaction of surface/subsurface hydrology and sediment distribution
	Potential for disturbing avifauna nesting by noise	
Drilling	Reduction in water quality	Alteration of surface hydrology and drainage
	Return of nutrients and toxins to surroundings	Changes in subsurface hydrology and drainage
	Inhibition of rainfall penetration	Changes in sediment distribution
	Noise and commotion during construction	Saltwater intrusion
	Increase in suspended solids	
	Changes in soil/water chemistry	Increased sediment release from discharges
	Changes in plant growth	Anoxia
	Destruction of biota	Loss of marsh habitat
	Potential for interrupting fish spawning and feeding	Altered soil/water chemistry
	Potential for disturbing avifauna nesting by noise	Possible negative influence on aquatic/benthic organisms
		Changes in mineral accretion and soil nutrition
Production	Noise and commotion during construction	Noise of processing facilities
	Flowlines in the marsh	Changes in surface hydrology and drainage
	Pit construction; toxins to surroundings	Changes in subsurface hydrology and drainage
	Changes in marsh elevation	Saltwater intrusion
	Reduction in water quality	More saltwater species
	Saltwater disposal	Anoxia
	Oil spills	Increased localized subsidence
	Destruction of biota	Loss of marsh habitat
	Potential for interrupting fish spawning and feeding	Altered soil/water chemistry
	Potential for disturbing avifauna nesting by noise	Possible negative influence on aquatic/benthic organisms
Pipeline building	Direct conversion to open water	Changes in surface hydrology
	Increased turbidity	Bank erosion
	Loss of forested wetlands	Compacted marsh surface
	Noise and commotion during construction	Direct habitat conversion
	Increased susceptibility to storm damage	Indirect wetland loss

continues

Table I continued

Stage	Short-term impacts	Long-term impacts
	Formation of open water ponds	Changes in forest succession
	Disruption of natural surface drainage	Changes in plant species, composition, diversity, and percentage cover
	Release of nutrients and toxins	Shoreline bank stability
	Soil oxidation	
	Changes in plant species, composition, diversity, and percentage cover	
	Nesting disturbance	
	Changed habitat	
	Disturbance of fish spawning and feeding	
	Impacts on plant growth	
	Destruction/Disturbance of benthos	
Spill control	Interruption of tidal cycle	Loss of marsh habitat
	Direct conversion of marsh to open water	Injury to birds and wildlife
	Trampling of vegetation	
	Immediate loss of marsh habitat	
	Temporary interruption of aquatic organism migration and flux of matter	
	Injury to birds and wildlife	
	Disruption of avifauna nesting	
Cleanup	Potential disruption of substrate	Loss of marsh habitat
	Removal of vegetation	Injury to birds and wildlife
	Immediate loss of marsh habitat	
	Decrease in biological production	
	Destruction of vegetation and benthic organisms	
	Potential negative impact on plant growth	
	Potential injury to wildlife	

produced inside the coastal zone and from the Outer Continental Shelf (OCS) disrupt natural hydrology. Fifth, spilled oils have an impact on wetland habitats. Sixth, spilled oil and produced water stress estuarine consumers by increasing turbidity, increasing salinity stress, introducing toxins, and so on. Finally, wetland loss decreases the value of the coastal zone as a nursery ground for estuarine consumers and the economic value to the human economy (Fig. 2).

This article reviews the multiple ecological impacts of oil- and gas-related impacts and synthesizes existing information to help researchers and managers understand how oil and gas development affects coastal wetland ecosystems in Louisiana, focusing on (1) plant physiology, (2) estuarine consumers (including the benthic community), and (3) hydrological disturbance and wetland loss.

2. IMPACTS ON PLANT PHYSIOLOGY

2.1 Impacts of Oil Spills

Wetland plants are subject to stresses from oil spilled during production and transportation (using tankers, pipelines, and tank trucks). Oil spills can have significant short- and long-term impacts on coastal ecosystems due to oil's physical effects and chemical toxicity, leading to decreased primary production, plant dieback, wildlife mortality, and marsh erosion. The mechanisms of these impacts are through (1) disruption of plant–water relationships, (2) direct impacts on plant metabolism, (3) toxicity to living cells, and (4) reduced oxygen exchange between the atmosphere and the soil. If leaves are coated with spilled oil, leaf stomata are blocked, oxygen diffusion to the roots decreases, and root oxygen stress

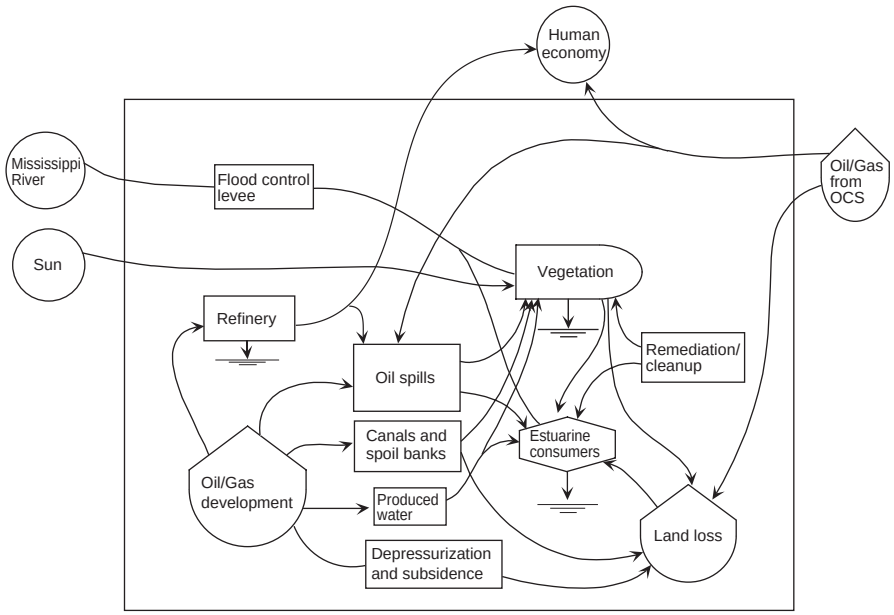


FIGURE 2 A conceptual diagram of ecological impacts of oil and gas development in Louisiana wetlands. The two main types of impacts are toxicity of hydrocarbons and hydrological changes.

increases, leading to reduction in plant growth. An oil-covered soil surface decreases oxygen movement, resulting in more anaerobic soil conditions and increasing low-oxygen stress on plant roots. Aerobic microorganisms in the oxidized sediment are more cable of degrading hydrocarbons than are anaerobic microorganisms in reduced sediment of the same pH. Oil coating on leaves results in higher mortality rates of marsh plants than does soil surface contamination.

In the short term, spilled oil can form a coating on plant foliage and the soil surface, increasing temperature stress, reducing oxygen diffusion, and reducing photosynthesis. The severity of these impacts is controlled by the amount of oil spilled, hydrological conditions (e.g., tides, winds), types of dispersed oil, and sensitivity of plants. Toxicity varies among various oil types. For example, diesel and number 2 oil are more toxic to marsh plants than is crude oil. Higher organic soils of fresh marshes are more sensitive to oil spills than is salt marsh through more rapid penetration and sorption of oil into the soil. For example, *Sagittaria lancifolia* is more tolerant than *Spartina alterniflora*, which in turn is more tolerant than *Spartina patens*.

There are longer term consequences of oil spills due to the persistence of oil or petroleum fractions in marshes. A period of 3 to 5 years may be required for natural recovery from accidental oil spills. Over time, oil concentrations are reduced in wetlands through oil evaporation, natural degradation of oil, and recovery of marsh plants.

However, plants subject to chronic exposure to oil spills accumulate oil-related pollutants through increased uptake of pollutants such as lead in contaminated sediments. The severity of biological effects of chronic spills is controlled by the volume and chemical nature of the pollutants, the physical nature of the receiving environment, and the biological character and composition of the ecosystem.

Traditional cleanup methods to remove spilled oil (e.g., water flushing, sand blasting, sediment removal, vegetation removal, sorbent application) often show limited removal efficiency. These methods can also cause potentially deleterious effects on long-term recovery of the impacted marsh system because they can result in further physical damage to both the vegetation and the underlying substrate, accelerating marsh degradation. Specifically, initial intensive cleanup by flushing and oil recovery by airboats reduces the residual oil in the marsh but also increases oil incorporation into the sediment and increases risks of physical damage to marsh plants. The negative impacts of physical removal of contaminated marsh plants can last for several years.

2.2 *In Situ* Burning

As a way in which to control oil spill impacts while minimizing physical damage to impacted wetlands, controlled *in situ* burning has been tested and used as an option to remove oil and gas condensate in contaminated wetlands. When water depth is sufficient (e.g., 10 cm) in damaged wetlands, *in situ* burning can be an efficient option because water on the surface will allow a successful burn of the above-ground vegetative component while absorbing heat produced by the fire and preventing root burning. However, flooding following *in situ* burning adversely affects plant growth in many species because the immediate reduction of plant cover decreases oxygen transportation to below-ground tissues. Full recovery of marsh vegetation from *in situ* burning reportedly takes 1 to 3 growing seasons. *In situ* burning generates atmospheric pollutants whose chemical components are a variety of gaseous sulfur (e.g., carbonyl sulfide, carbon disulfide) and carbon compounds (e.g., methane, carbon dioxide) and reduced alkylated naphthalene compounds from postburn oil.

2.3 Chemical Methods

Dispersants wash oil from surfaces such as rocks and vegetation. However, high doses (e.g., 0.3 L/m^2) reduce total and above-ground biomass significantly, at least for a short time. Another problem in dispersant application is that it is not practicable for use in coastal wetlands because there is little water to dilute the dispersed oil inside wetlands. Application of cleansers (e.g., COREXIT 9580), another chemical method, does not disperse oil but rather allows oil to be washed from surfaces, such as soils and plants, and to be collected in adjacent open water areas by boats. Plants improve their survival, regeneration, and above-ground biomass growth because the application leads to a recovery of stomatal conductance, photosynthesis, and respiration. The effectiveness of the cleanser in cleaning up oil depends on oil type, delivery mode, timing, and amount of oil. Marsh plants (e.g., *S. lancifolia*, *Scirpus olneyi, Thpha latifolia*) show different sensitivities to cleanser use. Cleanser application to brackish (*S. patens*) and fresh (*S. lancifolia*) marshes removes oil from marsh grasses and reduces the short-term impact of oil spills on gas exchange of the vegetation, but it still reportedly results in reduced above-ground biomass for the first growing season. Another problem in cleanser application is the potential impact of cleanser toxicity on various organisms in wetlands. Finally, solidifiers, which are dry, granular, hydrophobic polymers, can be used for spilled oil removal. Solidifiers react with spilled oil to form a floating cohesive and solidified mass that is easily removable, leaving very little residue.

2.4 Bioremediation

Wetland plants have the potential to enhance the bioremediation process through diffusion of oxygen from the shoots to the roots and soil, and its effectiveness can be increased through applications of fertilizer, microbial products, or soil oxidants. Fertilizer may be applied to spilled oil in water bodies before it reaches the marsh or may be applied directly to marshes already contaminated with oil. Additions of soluble inorganic fertilizers stimulate microbial activity and prolong the period of active oil degradation. The microbial activities lead to degradation of toxic petroleum products into carbon dioxide and water. However, the effectiveness of bioremediation may be limited by marsh plant tolerance to oil-related stress, risk of eutrophication, and soil type. Nutrients added to marshes may be transported to adjacent water bodies, leading to algal growth. Fertilizer application is more effective in sandy soils than in mineral soils. In some cases, the best action may be to allow the marsh to recover on its own, using no cleanup techniques.

3. IMPACTS ON ESTUARINE CONSUMERS

Benthic and nekton species are key organisms, both ecologically and economically, in coastal and wetland systems. The benthic community is an important link in transferring contaminants from the sediment to higher trophic levels, and the benthic community structure is sensitive to petroleum hydrocarbon exposure. The ecological and biological impacts of energy development on coastal marshes and estuarine environments are broad and sometimes persistent, including mortality, growth inhibition, reduced production, altered metabolic systems, and tainted flesh in fish and shellfish. Oil and gas production and transportation in coastal wetlands in Louisiana have resulted in the accumulation of polycyclic aromatic hydrocarbons (PAHs) and heavy metals in impacted areas. These compounds cause ecological impacts, including alteration of aquatic

community structure and food chains. Biodiversity and population density of the benthic community are significantly lower in oil-contaminated areas. Recruitment and feeding patterns can also be altered by oil pollution.

4. HYDROLOGICAL DISTURBANCE AND LAND LOSS

Petroleum-related activities in coastal Louisiana have several secondary and indirect impacts. These include the production of produced water, drilling-induced subsidence, and hydrological modifications due to dredging activities. Dredging results in two interrelated impacts: creation of new water pathways and spoil placement.

4.1 Produced Water

Produced water is a by-product of the oil production process. There are often substantial amounts of water contained in subsurface formations where oil and gas occur. When oil and gas are produced, this water is brought to the surface and must be disposed of in some way. Produced water contains high concentrations of volatile (e.g., benzene, toluene), and semivolatile hydrocarbon contaminants (e.g., aliphatic hydrocarbons, alkylated polycyclic aromatic hydrocarbons) as well as high concentrations of aromatic acids and aliphatic fatty acids. Some produced water also contains metals such as vanadium, arsenic, and copper. These toxic metals and organics often accumulate in wetlands due to poor freshwater flow or tidal exchange. Produced water also often contains high levels of brine. The degree of ecological impacts on wetlands is influenced by (1) discharge rate, (2) quantity and quality of the hydrocarbons and trace metals present in a particular discharge, (3) local hydrology, (4) sediment disturbances (e.g., dredging, boat traffic), and (5) sediment types (organic carbon content and texture). Volatile and semivolatile organic compounds of produced water are accumulated in organisms (e.g., oysters, freshwater mussels). However, when exposed to a contaminant-free environment, oysters release accumulated hydrocarbons. Heavy metals are reported not to influence hydrocarbon degradation in sediments at the produced water discharge site. The degradation rate of petroleum hydrocarbon in sediments exposed to produced water is controlled by oxidation, redox potential, and pH and can be increased by the addition of fertilizer.

4.2 The Delta Cycle and Wetland Loss in the Mississippi Delta

Coastal wetland loss is a major environmental problem in coastal Louisiana. The remainder of this article discusses evidence on the causes of this land loss and the role of petroleum-related activities in contributing to wetland loss.

4.2.1 The Delta Cycle

To understand the factors related to wetland loss, it is necessary to understand how the Mississippi Delta was formed. The sea level stabilized near its current level at the end of the last glaciation 5000 to 7000 years ago. Since that time, delta switching of the Mississippi River has created a series of overlapping deltaic lobes that currently form the Mississippi deltaic plain in coastal Louisiana. Delta switching occurs every 1000 years or so, resulting in new loci for sedimentation and marsh development. Rapid land building occurs in active delta lobes, whereas submergence and wetland loss occur in abandoned lobes. The Atchafalaya River is the most recent diversion in the delta switching process. Subaerial expression of the new Atchafalaya Delta began in 1973, and this area has a net gain of wetlands. Thus, the delta building process is a balance between forces that lead to growth of the delta and those that cause deterioration. The Mississippi River is the major force leading to land gain. Overbank flooding and reworking of sands have formed a skeletal framework of natural levee ridges and barrier islands within which the delta plain has formed. Deposition of both coarse and fine-grained sediments formed wetlands and maintains existing wetlands. Sediments resuspended during storms are an important source of sediments to maintain marshes. Once a wetland forms, organic soil formation by wetland plants is an important mechanism for maintaining coastal marshes.

Naturally, wetland deterioration is caused by two primary forces: subsidence and wave erosion along shorelines. Geological subsidence is caused by compaction, dewatering, and consolidation of sediments. Subsidence in deltas leads to a rate of relative sea level rise (RSLR) that is often much greater than eustatic rise. For example, whereas the current rate of eustatic rise is 1 to 2 mm/year, the RSLR in the Mississippi Delta is greater than 10 mm/year; thus,

eustatic sea level rise accounts for only 10 to 15% of total RSLR. If wetlands in deltas do not accrete vertically at a rate equal to the rate of RSLR, they will become stressed due to factors such as waterlogging, sulfide toxicity, and salt stress and will ultimately disappear. Because vertical accretion is stimulated by both outside sediment input and *in situ* organic soil formation, a reduction of sediment input or an increase in plant stress can lead to lowered accretion rates and wetland loss. Subsidence generally leads to wetland loss in interior wetlands. Wave erosion leads to wetland loss along exposed shorelines. As a deltaic lobe progressively deteriorates, wave erosion becomes relatively more important.

4.2.2 Wetland Loss in the Mississippi Delta during the 20th Century

During the 20th century, there was a dramatic reversal of the net growth of the Mississippi Delta that had taken place over the past several thousand years. High rates of land loss occurred (with estimates up to 100 square kilometers/year), and a total area of 4000 square kilometers of coastal wetlands has been lost. A number of factors have been linked to land loss, including elimination of riverine input to most of the coastal zone due to construction of flood control levees along the Mississippi River, altered wetland hydrology due to factors such as canal construction and impoundments, saltwater intrusion, wave erosion along exposed shorelines, a decline in suspended sediments in the Mississippi River, the effects of geological faulting, and high RSLR.

4.2.3 Effects of Oil and Gas Production on Subsidence

As stated previously, the regional rate of geological subsidence in the Mississippi Delta is approximately 10 mm/year. This is due to compaction, dewatering, and consolidation of sediments. Recently, Morton and colleagues showed that the rate of subsidence in producing oil and gas fields was considerably higher than this regional average (as much as 23 mm/year). They concluded that the increasing and then decreasing pattern of land loss in south central Louisiana was attributable partly to increased and then decreased oil and gas production. Decreases in subsurface pore pressures associated with production were so large that stressed faults were reactivated, leading to rapid subsidence on the down-thrust side of the fault. This enhanced subsidence led to wetland plant stress and death, as discussed previously. Thus, enhanced subsidence on top of regional geological subsidence led to much greater waterlogging stress on plants. Spoil banks associated with oil and gas fields led to reduced sediment input and lower organic soil formation, exacerbating sediment accretion deficits.

4.3 Canal and Spoil Banks

Canals have been constructed in the coastal wetlands of Louisiana since Europeans first settled in the region during the early 1700s. For nearly two centuries, these canals were dredged mainly for navigation, flood protection, and drainage. However, after the 1930s, the discovery of oil and gas fields in the coastal zone led to an explosion of canal construction related to oil and gas production. By the mid-1980s, there were more than 15,000 km of canals, the surface area of canals was equivalent to 2.3% of wetland area, and the total area of spoil bank levees plus canal surface was approximately 9.5% of wetland area.

When a canal is dredged, the dredged material is deposited along the sides of the canal, creating a more or less continuous elevated bank called a spoil bank. Spoil banks generally consist of highly organic marsh soil. As the spoil banks settle and dewater and as organic matter oxidizes, they create a levee that runs parallel to the canal (Fig. 3). Canals and associated spoil banks alter natural hydrology in two main ways. First, the canals are deep and straight, in striking contrast to the mostly shallow and sinuous tidal channels. Because of this, dredged canals tend to preferentially capture flow from natural channels. It has been shown that as the density of canals in an area increases, the density of natural channels decreases. If canals are long and deep enough (e.g., navigation channels that stretch from the Gulf of Mexico inland to freshwater areas), they can cause significant saltwater intrusion and death of freshwater wetlands. Two notable examples of this are the Mississippi River Gulf Outlet, which caused the death of extensive cypress forests southeast of New Orleans, and the Calcasieu Ship Channel, which led to loss of extensive sawgrass marshes in southwest Louisiana.

In contrast to deep canals that enhance water flow, spoil banks reduce water exchange. Spoil banks reduce or even eliminate overland flow. Because of the presence of spoil banks, partially impounded areas have fewer but longer periods of flooding and reduced water exchange in comparison with unimpounded marshes. Ponds usually develop behind canals and spoil banks, and high wetland loss is associated with areas of high hydrological changes.

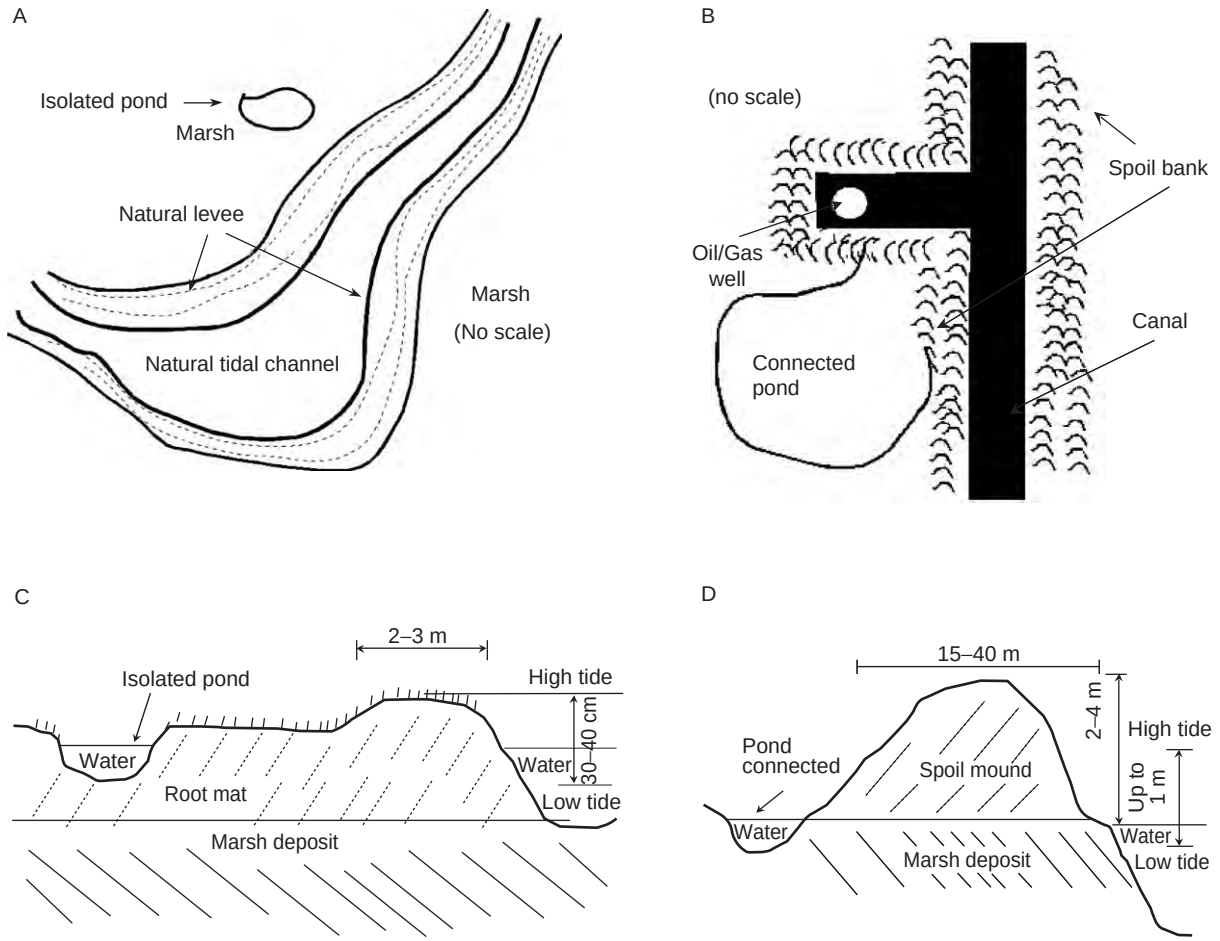


FIGURE 3 Schematic diagrams of canal dredging on coastal marshes. (A) Top view of a natural tidal channel and natural levee. (B) Top view of canal dredging showing straight canal and spoil bank. (C) Cross-sectional view of natural marsh. Note the streamside natural levee and shallow tidal channel and that high tide is higher than the natural levee. (D) Cross-sectional view of a dredged canal. Note that the dredged canal is deep, water level variation is higher than in natural channels, and the spoil bank is higher than normal high tide, preventing flooding of the marsh.

As discussed previously, if canals are associated with oil and gas fields, subsidence is enhanced through depressurization during and after oil and gas production.

Tidal currents are stronger through dredged canals than through natural channels. This, coupled with erosion from boat wakes, results in erosion of the banks. Annual increases in canal widening range between 2 and 14% per year, for a doubling time of 5 to 60 years. Canals also worsen water quality problems. Normally, nutrient- and sediment-laden point and nonpoint source runoff from uplands would naturally flow slowly through wetlands where nutrients and sediments would be assimilated. Canals short-circuit this flow, leading to direct introduction of nutrients and suspended sediments and to eutrophication in open water bodies.

4.4 Impoundment

Coastal marshes exchange water, organic materials (e.g., detritus), nutrients (e.g., nitrogen, phosphorus), and organisms with surrounding estuarine waters, supporting estuarine fish and shellfish. One impact that has affected these processes in coastal marshes is impoundment. Impoundments have been constructed for a number of reasons. Beginning in the 19th century, impoundments were constructed for the purpose of land reclamation for urban and agricultural activities. Many of these reclaimed areas failed due to excessive subsidence and flooding during hurricanes. But some remain, mainly in the metropolitan area of New Orleans. During the 20th century, many impoundments were constructed in the coastal zone to enhance conditions for waterfowl

and for marsh management. In addition to these purposefully constructed impoundments, large areas of the coastal zone have been completely or partially impounded by the cumulative impacts of canal and spoil bank construction. Approximately 30% of the total wetland area in coastal Louisiana has been impounded, either purposefully or by accident. Impoundments have been shown to have a number of detrimental impacts in that they reduce water exchange and accretion inside impoundments, lower vegetation productivity, and reduce movement of migratory fishes, leading to deteriorating marshes.

4.4.1 Petroleum-Related Activities and Wetland Loss

From the preceding discussion, a number of conclusions emerge. Naturally, wetland establishment and deterioration in the Mississippi Delta is a very complicated process involving numerous factors, including geological and geophysical (e.g., channel switching, sediment introduction and deposition, subsidence, vertical accretion, wave erosion, saltwater intrusion, sea level rise), biogeochemical (e.g., anaerobic soil formation, sulfate reduction, peat decomposition), and ecological (e.g., waterlogging and salinity increase leading to plant stress and death, rates of organic soil formation, herbivore grazing) factors. Prior to extensive alteration by human activity, there were large gains and losses of wetlands in various parts of the deltaic plain as the river changed course. But over the past 5000 to 6000 years, the net result of these processes was a large net gain of wetlands.

During the 20th century, the long-term net gain of wetlands was reversed and wetland area in the delta decreased by approximately 25%. Clearly, some of this loss was natural and would have occurred without human impacts. But it seems clear that the dramatic reversal from net gain to net loss can be attributed to human activities. Two general and interrelated processes are responsible for the losses: pervasive hydrological change and dramatically increased subsidence.

From a hydrological point of view, there have been two pervasive changes. First, the delta has been nearly completely isolated from the river that built it. Levees extend to the mouth of the main channel of the river, and 70% of sediments and water flow into the Gulf of Mexico. Only in the Atchafalaya Delta region does river water enter a shallow inshore area, and this is an area of land gain. Internally in the delta, there have been massive hydrological changes. A dense network of canals, most associated with petroleum activity, has changed the delta dramatically. These canals allow saltwater intrusion, reduce water and sediment movement, and contribute to low accretion rates. Impoundments isolate large areas of the coastal zone from adjacent estuarine areas.

4.4.2 Role of Petroleum Activity

In the vicinity of oil and gas fields, subsidence increased due to depressurization and surface hydrology was altered due to canals and spoil banks. Thus, RSLR was increased and the rate of accretion was reduced. This is due both to a reduction in allochthonous sediment input and to *in situ* organic soil formation. Some have attributed practically all wetland loss in the coastal zone to canals. There is no doubt that oil and gas activity have had a major impact on wetland loss. In areas of intense oil and gas extraction, it is likely that most wetland loss can be related to the combined impacts of increased subsidence and surface alterations. But high rates of wetland loss are also related to wave erosion, saltwater intrusion, and changes in the engineering of the mouth of the Mississippi River. Nonetheless, there is no doubt that petroleum-related activities are directly responsible for a significant proportion of the land loss in the coastal zone. It is probably not possible to put a specific value on this land loss due to the complexity of the land loss problem.

From a broader perspective, it is better to consider the functioning of the whole coastal system and the conceptual model developed earlier in this article. Both the supply side (inputs to the delta) and the receiving system (the delta plain) have been affected. Both riverine input and resuspended inputs have been reduced. At the mouth of the Atchafalaya River, oil and gas fields are not associated with wetland loss. Thus, it is the combination of all these forces, but oil and gas have had a dramatic impact.

5. CONCLUSION

Petroleum exploration, production, and transportation in the Louisiana coastal zone increased dramatically from the early 20th century to the 1970s. Oil and gas production in inshore bays and wetlands of the coastal zone then decreased beginning in the late 1970s, but there is still a high level of transportation of oil and gas through the coastal zone from the OCS and Louisiana Superport. These activities have had significant impacts on floral and faunal communities, resulting in significant deterioration of coastal and

wetland ecosystems. These impacts are related to the toxicity of spilled oil and the secondary and indirect effects of petroleum-related activities such as alteration of hydrology. The impacts of OCS development are related to construction of pipelines and navigation channels. Thus, the risks of oil spills and hydrological disruption continue, even though inshore oil and gas production has decreased.

Responses of plant metabolism to oil impacts are complex, depending on exposure type (e.g., oil-coated leaves vs soil contamination), oil type (e.g., crude oil vs number 2 oil), time of spill (e.g., after the growing season vs before the growing season), density of spilled oil, and sensitivity of various marsh plant species to oil. Another complex matter is the impact of oil cleanup on wetlands. Removal of oil has been reported to cause significant damage to wetland communities, including reduced growth of marsh plants and reduced population of benthic organisms.

Lin and colleagues suggested that methods and intensity of oil spill cleanup depend on the type and amount of spilled oil and environmental conditions at the time of the spill. If the spill is a relatively small volume and the floating oil is not continuous, light or no cleanup action is recommended. In the case of large-volume oil spills, cleanup activities consisting of sorbent application, low-pressure flushing, vacuuming, rope mops, and the like should be considered as options. However, Lin and colleagues did not recommend the use of heavy equipment and intrusive mechanical cleanup due to the concerns of physical damage to fragile marshes.

Louisiana experienced a high rate of coastal marsh loss during the 20th century. This high loss rate has been attributed to a number of factors. The immediate cause of much loss is due to plant stress resulting from both natural and anthropogenic causes, followed by plant dieback, subsequent erosion of the marsh substrate, and the formation of small ponds that then coalesce into larger open water bodies. Causes of plant stress in Louisiana marshes have been attributed to waterlogging stress (due to insufficient elevation of the marsh surface resulting from high subsidence rates in the deltaic plain and low accretion rates) and salinity stress (due to saltwater intrusion, often from storm surge events, into the more interior marshes).

Petroleum-related activities have contributed significantly to wetland loss in the Mississippi Delta. Oil and gas extraction increased the subsidence rate, sometimes by a factor of as high as 2 to 3, due to reduction of pressure that led to faulting-related subsidence. On the surface, canals significantly altered natural hydrology. Deep dredged canals altered water flow pathways and sometimes resulted in saltwater intrusion. Spoil banks reduced overland flow exchange and sediment input to the wetland surface. The combination of these two factors increased plant stress and plant death.

Acknowledgments

Support for this effort was provided by the National Oceanic and Atmospheric Administration through the Louisiana Sea Grant College Program (NOAA Grant NA16G2249). Support was also provided by the Army Corps of Engineers, and the U.S. Geological Survey.

SEE ALSO THE FOLLOWING ARTICLES

Aquaculture and Energy Use • Crude Oil Releases to the Environment: Natural Fate and Remediation Options • Crude Oil Spills, Environmental Impact of • Ecological Risk Assessment Applied to Energy Development • Ecosystem Health: Energy Indicators • Energy Development on Public Land in the United States • Fisheries and Energy Use • Public Reaction to Offshore Oil

Further Reading

Baker, J. M., Little, D. I., and Owens, E. H. (1993). A review of experimental shoreline oil spills. *In* "Proceedings of the 1993 Oil Spill Conference," pp. 583–590. American Petroleum Institute, Washington, DC.

Boesh, D. F., Josselyn, M. J., Mehta, A. J., *et al.* (1994). Scientific assessment of coastal wetland loss, restoration, and management in Louisiana. *J. Coastal Res.* **40**. (Special issue)

Cahoon, D. R. (1989). "Onshore Oil and Gas Activities along the Northern Gulf of Mexico Coast: A Wetland Manager's Handbook." Lee Wilson & Associates, Santa Fe, NM.

Craig, N. J., Turner, R. E., and Day, J. W., Jr. (1979). Land loss in Coastal Louisiana (U.S.A). *Environ. Mgmt.* **3,** 133–144.

Day, J. W., Britsch, L. D., Hawes, S. R., Shaffer, G. P., Reed, D. J., and Cahoon, D. (2000). Pattern and process of land loss in the Mississippi delta: A spatial and temporal analysis of wetland habitat change. *Estuaries* **23,** 425–438.

Dicks, B., and Hartley, J. P. (1982). The effect of repeated small oil spillages and chronic discharge. *Phil. Trans. R. Soc. London Series B* **297,** 285–307.

Gornitz, V., Lebedeff, S., and Hansen, J. (1982). Global sea-level trend in the present century. *Science* **215,** 1611–1614.

Hall, C. A. S., Howarth, R., Moore, B., III, and Vőrősmarty, C. J. (1978). Environmental impacts of industrial energy systems in the coastal zone. *Annu. Rev. Energy* **3,** 395–475.

Lin, Q., and Mendelssohn, I. A. (1996). A comparative investigation of the effects of south Louisiana crude oil on the vegetation of fresh, brackish, and salt marshes. *Mar. Pollut. Bull.* **32,** 202–209.

Lin, Q., Mendelssohn, I. A., Hester, M. W., Webb, E. C., and Henry, C. B., Jr. (1999). Effect of oil cleanup methods on ecological recovery and oil degradation of Phragmites marshes. *In* "Proceedings of 1999 International Oil Spill Conference," pp. 511–517. American Petroleum Institute, Washington, DC.

Lindstedt, D. M., Nunn, L. L., Holmes, J. C., Jr., and Willis, E. E. (1991). "History of Oil and Gas Development in Coastal Louisiana." Louisiana Geological Survey, Baton Rouge.

Louisiana Department of Natural Resources (1996). "Well Status Master." LDNR, Baton Rouge, LA.

Louisiana Department of Natural Resources (2001). "Louisiana Energy Facts Annual." LDNR, Baton Rouge, LA.

Mendelssohn, I. A., and Morris, J. T. (2000). Eco-physiological controls on the productivity of *Spartina alterniflora* loisel. *In* "Concepts and Controversies in Tidal Marsh Ecology" (M. P. Weinstein and D. A. Kreeger, Eds.), pp. 59–80. Kluwer Academic, Boston.

Morton, R. A., Buster, N. A., and Krohn, M. D. (2002). Subsurface controls on historical subsidence rates and associated wetland loss in southcentral Louisiana. *Gulf Coast Assoc. Geol. Soc.* **52,** 767–778.

Pezeshki, S. R., Hester, M. W., Lin, Q., and Nyman, J. A. (2000). The effects of oil spill and clean-up on dominant U.S. Gulf coast marsh macrophytes: A Review. *Environ. Pollut.* **108,** 129–139.

Roberts, H. H. (1997). Dynamic changes of the Holocene Mississippi River delta plain: The delta cycle. *J. Coastal Res.* **13,** 605–627.

St. Pe', K. M. (ed.). (1990). "An Assessment of Produced Water Impacts to Low Energy, Brackish Water Systems in Southeast Louisiana." Louisiana Department of Environmental Quality, Baton Rouge.

Turner, R. E. (1987). Relationship between canal and levee density and coastal land loss in Louisiana. *U.S. Fish Wildlife Service Biol. Rep.* 85(14).

Wind Energy Economics

MICHAEL MILLIGAN
U.S. Department of Energy, National Renewable Energy Laboratory
Golden, Colorado, United States

1. **Introduction to Wind Energy Economics**
2. **Recent Trends in the Cost and Use of Wind Power Plants**
3. **Economic Value of Wind Power Plants**
4. **Role of Wind Forecasting and Its Relationship to the Value of Wind**
5. **Effects of Geographically Disperse Wind Sites**
6. **Reserve and Regulation Impacts**
7. **Other Issues: Location of Wind/Loads/Transmission**
8. **Conclusions and Summary**

Glossary

base load generation Generating capacity that is kept online for long periods of time, usually due to low marginal cost combined with physical constraints on frequent starting and stopping.
demand Instantaneous use of power, measured in kilowatts or megawatts.
economic dispatch The process of determining which generators should be run, and in what combination, to minimize power production cost.
effective load carrying capability (ELCC) The demand that can be supported by the generating supply at a given reliability level.
energy Power used over time, measured in kilowatt-hours or megawatt-hours.
load following (generation) Generating capacity that can be ramped up or down to match changing levels of demand.
loss of load probability (LOLP) A measure of reliability; LOLP is the probability that there is insufficient generating supply to support electrical demand.
marginal cost The change in total cost that is caused by a given increase in energy output.
peak The maximum demand during a given time period.
regulation Changes in output from designated generators that occur over very short time frames, such as several minutes, that compensate for increasing or decreasing demand or intermittent generation fluctuations.
reliability The ability of the electrical system to meet demand, often measured by loss of load probability.
reserves Capacity that is not used directly to supply power but that can be called on quickly in event of a system emergency such as an unanticipated generator outage or an increase in demand.

The economics of wind power plants is influenced by a number of factors. These include the quality of the wind resource, technology efficiency and reliability, the availability of long-term power contracts, and the ability to forecast at least several hours ahead. Wind power plants have value that arises due to the reduction of pollution, compared with many other types of power plants, and the lack of fuel price volatility. Value also depends on the specific nature of the displaced power and energy. In regions that have a high marginal cost of energy, wind can provide a high value by offsetting a costly energy source.

1. INTRODUCTION TO WIND ENERGY ECONOMICS

Over the past few years, there have been significant changes in the electric utility industry. In many parts of the world, power markets have moved away from the regulated monopoly market structure toward more competitive markets. This evolution is not uniform; even in the United States, there are many different degrees of restructuring in the electricity markets. In some markets, restructuring has mainly involved changes at the wholesale level, whereas other market changes have occurred at the retail level. The future organization of power markets is not yet clear, and the details of these markets may have important impacts on wind integration into the power supply. For simplicity, this article uses the term "utility industry" to characterize the current mixed state of affairs as well as the future industry in whatever form it may take.

 © 2004 Elsevier Inc. All rights reserved.

Amid all of this change, there have been significant developments in the wind energy industry. Wind turbine technology has evolved with the advent of more efficient, economical, and reliable wind machines. Many utilities in Europe and in the United States now use wind power to supply increasing levels of electricity supply. Recent volatility in natural gas markets has induced a closer look at the power price stability offered by wind power plants. Although a significant fraction of wind power that has been developed in the United States resulted from regulatory action or other political decisions, wind power is now often cost-competitive with other conventional power sources.

This article examines some of the main drivers of wind energy economics in the context of large-scale power generation. In spite of the many wind installations throughout the world, surprisingly little is known about specific integration costs of wind plants. As the electric utility industry gains more experience, this knowledge base will grow.

Large-scale use of wind power plants can pose challenges to power system operations and planning. Because wind is an intermittent power source that is available only when the wind blows, the economics surrounding the use of this technology can become complicated. Because wind systems cannot respond to control systems in the same way as can conventional thermal plants, questions arise concerning the need for storage or other mechanisms that could compromise the economics of wind but would perhaps help to integrate wind into the electrical supply. Other questions concern the ability to measure the wind plant capacity in a way that helps decision makers to analyze economic trade-offs between wind and other types of power plants.

This article describes some of the recent trends in the cost of wind power plants. It discusses the economic value of wind power in the context of bulk power generation. Because of its variable nature, the use of wind power can be more effective when it can be forecast more than a few hours in advance. Furthermore, the value of wind power depends in large part on the specific characteristics of the generators in the same grid.

2. RECENT TRENDS IN THE COST AND USE OF WIND POWER PLANTS

Development of wind power systems has increased rapidly since 1986. At that time, most North American wind turbine installations were in California, resulting from significant tax incentives in that state combined with the availability of long-term power purchase agreements that regulators required utilities to offer to independent power producers. During this period, the average energy cost of wind-generated electricity fell from $0.40 in 1979 to a range of $0.04 to $0.06/kWh in 2000. During the 10 years ended in 2000, the cost of energy fell from a levelized cost of approximately $0.11/kWh (in 2000 dollars) to $0.06/kWh at moderate wind sites, with an average annual wind speed of 5.8 m/s. For high-wind sites with an average wind speed of 6.7 m/s, the cost of energy fell from just under $0.08/kWh to approximately $0.04/kWh during this same time period. These falling costs, along with a U.S. federal production tax credit (PTC), helped to increase development in the United States beginning in the early 1990s and continuing through the end of 2001. The PTC provides a $0.015/kWh credit for electricity generated by wind power plants. At the time of this writing, the PTC was set to expire December 31, 2003.

Beginning in the early 1990s, the wind industry grew very quickly in Europe, where development was heavily influenced in Germany by aggressive public policy of mandated wind purchases at relatively high prices. German policy was to help address environmental issues associated with power generation that are perceived differently in Europe than in the United States. As can be seen in Fig. 1, growth in Europe far outstripped growth in the rest of the world during the past decade or so. The graph shows the projected path of development for the period from 2002 to 2005. The primary driver of this growth is the anticipated cost-competitiveness of energy from wind compared with other power generation technologies. Wind energy cost was predicted to fall to $0.03 to 0.05/kWh by 2004, with further reductions made possible by higher towers, greater energy capture at lower wind speeds, and even more reliable technology. However, growth in the rest of the world will also be driven by enormous electrical capacity needs in regions with little, if any, existing electric power.

The cost of wind energy is a function of the technology and the wind resource at the development site. The levelized cost of energy includes both fixed and variable costs over the lifetime of the wind plant. The fixed cost depends on the ownership structure and the method used to finance the plant construction. For example, cooperative ownership generally involves a lower interest rate than could be obtained by an independent power producer. In the United States, cooperatives do not qualify for the PTC because they

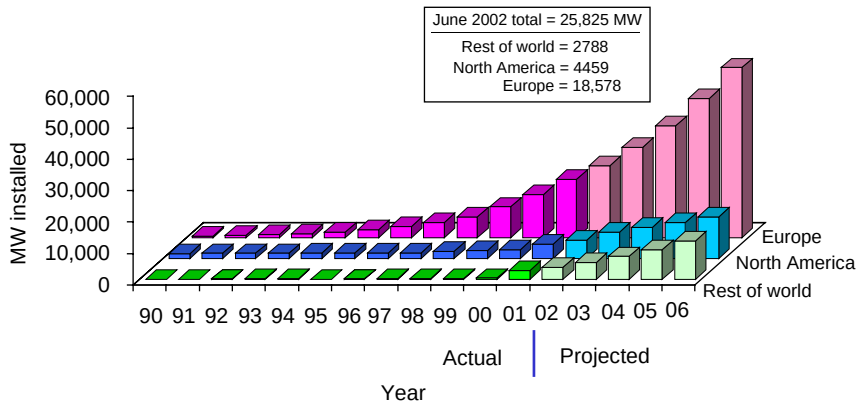


FIGURE 1 Growth of wind energy capacity worldwide. Adapted from BTM Consult ApS (March 2002) and *Windpower Monthly* (January 2002).

are nonprofit entities. Instead, such organizations can use the renewable energy production incentive, which allows nonprofits to take advantage of a similar cost reduction as would be available to their for-profit counterparts. Both types of ownership structures can use project financing or internal/corporate financing, with the latter providing a lower capital cost and, therefore, a lower levelized cost of energy.

Operation and maintenance (O&M) costs are the variable costs associated with wind power generation. These costs include labor and materials required for operation and periodic maintenance of the turbines, insurance, land lease, property tax, and general and administrative costs. At the time of this writing, average O&M costs were approximately $0.005/kWh and were expected to decline as turbine reliability improves.

Contract terms and conditions can have a significant impact on the financing terms for new projects and for major capital repairs and repowering. These financing terms can have a significant impact on the cost of wind energy.

3. ECONOMIC VALUE OF WIND POWER PLANTS

The results in this section are based largely on power production simulation modeling of real power systems. Wind power plants can provide value in three ways. First, because there is no fuel cost for a wind plant, wind energy can displace conventional generation and the fuel it consumes. Second, construction of new wind power plants can also reduce the need for building new conventional plants; this is called the capacity value of a wind plant. Third, because they can reduce the emissions created by the use of some conventional fuels, wind plants can have emissions reduction value.

Because the benefits of wind power generation depend heavily on the economic characteristics of the remaining electricity supply, this section begins with a short description of the power dispatch process, focusing on those aspects that are most relevant to the economics of wind power.

Modern power systems include a large number of generating units that are often of different sizes and types. For example, to start up a coal plant, the boiler must be brought up to operating temperature before power can be produced. This process normally takes several hours. When the plant is not needed, it can be shut down, but this process also takes time. To run this type of plant in an efficient manner, the operator will attempt to minimize these start/stop operations, running the plant for long durations of time. During this operating period, the plant may be able to produce either low or high levels of power, but a large thermal plant such as this might not be able to respond quickly to operator commands. Large coal plants generally have relatively low fuel costs, so these units are typically operated as base load units, with minimal output variations for relatively long periods. The average cost of energy can be quite low.

Other types of generators might be able to respond more quickly to operator signals, and they can be used as load-following units by increasing output to match rising demand or by decreasing output to match declining loads. This type of unit often has a higher energy cost than does a base load unit, but the former offers somewhat more flexibility in matching power generation to the load. Combustion turbines generally have much higher fuel costs, but they start/stop in relatively short time scales.

Most power systems are operated with a primary objective of providing electricity when and where it is needed at the lowest possible cost or the highest possible profit. Therefore, in either case, the power system operator must solve a complicated mathematical optimization problem that minimizes cost, subject to the operating constraints of the various generating units and subject to providing sufficient energy and capacity to satisfy electrical demand. The solution to this problem normally involves making full use of the least expensive generating unit, followed by each unit in increasing order of energy cost (and subject to the various generator constraints). After midnight, electrical demand is typically low. It would not be uncommon to find a utility running only its base load units—those with the lowest energy cost. As the electrical demand increases in the morning, progressively more costly plants must be brought online. During a few hours in the late afternoon or early evening on a particularly hot day, expensive peaking plants may be run, but only for the minimum necessary time.

Figure 2 illustrates the basic economics of a utility or generating company that owns or operates the three types of power plants. The marginal cost (MC) curve is defined as the change in total cost divided by the change in energy output. Under perfectly competitive markets, the MC curve is the same as the firm's supply curve. However, because wholesale and retail electricity markets are not perfectly competitive, the MC curve represents the upper bound on the quantity of output (energy) that an electric utility would be willing to provide at alternative prices.

The initial downward-sloping portion of the MC curve shows the effect of operating base load units at higher output levels; that is, it increases their efficiency and decreases their cost. For higher demand levels, less economical plants are added to the mix. These intermediate units have higher MCs, resulting in systemwide MC increases. Finally, peaking units provide energy during system peak times at even higher MC.

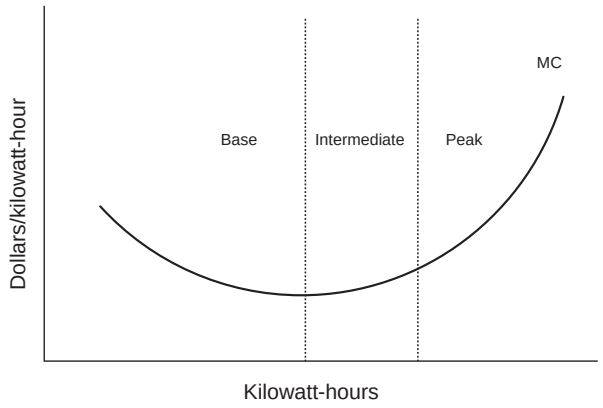


FIGURE 2 Trend in marginal cost of electricity production in relation to various types of generators.

This curve helps us to understand why the value of wind generators can vary for different utilities, different seasons, or even different times of the day. Because power demand fluctuates over time, the utility effectively moves up or down its MC curve as it brings new generators online or reduces power output as the demand falls. Because wind is fuel free and so has a near zero MC, it would always be used in an economic dispatch whenever possible. The result is that the remaining load, net of wind generation, decreases and the marginal generator's output can be reduced or eliminated over the relevant time period. Utilities with relatively high MCs would find that wind power has a high value, whereas utilities with lower MCs would find that wind energy has a lower value. Furthermore, this value would tend to change over time for all utilities due to the constantly changing nature of the electrical load. This also helps to explain why a wind plant that produces significant energy output during the night is not as valuable as a wind plant that produces during the day.

When new wind plants go into service, they can partially displace the need for conventional power plants. This displacement is called capacity credit. At the time of this writing, it was not entirely clear how this displacement, or capacity credit, would be evaluated in restructured energy markets. However, with the increasing need to fully evaluate various risks posed by power system operation, utilities have turned to financial risk theory to evaluate the more complex market environment, and it appears that probabilistically based methods to evaluate capacity credit will continue to be useful. One such measure that has evolved from power system reliability analysis is called the effective load-carrying capability (ELCC), which is based on achieving a loss of load probability (LOLP) target with two alternative generator additions. ELCC allows one to rate the equivalent perfectly reliable power plant that would provide the same system reliability as a wind plant. In other cases, it may be more useful to rate the wind plant capacity credit in terms of its reliability-equivalent gas plant. This capacity credit is computed with the help of a production simulation model or an electrical system reliability model. Such a model uses extensive power system data to calculate, among other things, the probability that generation will be sufficient to meet the projected electrical demand. The capacity credit of a wind

plant can range from very low, perhaps 10% of its rated capability, up to 40% or higher, depending on the correlation of wind power to load and other system characteristics.

Wind energy can displace emissions that are often produced by other power sources. This emissions reduction is a function of the characteristics of the marginal plant or plants and the extent to which wind power can offset other forms of power. For example, a wind plant that can displace energy from a high-emissions source would provide a higher emissions value than would a wind plant that displaces output from a modern gas turbine.

Placing a value on avoided emissions can be a difficult and controversial practice. In principle, this valuation would involve an accounting of illnesses, hospital visits, and other costs incurred as a direct result of the major emissions sources. There are some markets for emissions mitigation credits in the United States that implicitly value emissions. However, because it is difficult to follow the complex causal chain to assess actual pollution costs, specific damage values can be difficult to justify. The state of Minnesota established the cost of emissions in 1997 and updated these values in February 2002. Prior to electricity restructuring, California regulators required an explicit valuation of emissions in the various air basins as part of electric utility generation expansion plans.

Figure 3 illustrates the three components of the value of wind power systems. The calculations for this graph are based on an electricity production simulation model applied to two large utilities in the United States. (The utilities, denoted by u1 and u2 in the figure, are not identified due to prior agreement.) These utilities have different generation characteristics, resulting in differing estimates of the benefit of wind. The simulations show the impact of alternative wind sites to economic value. (Two wind sites are shown in the figure: one from the High Plains and one from a mountain pass on the West Coast.) The largest benefit results from the reduction in fuel that is needed by conventional power plants when wind is added to the system. The capacity value is based on estimates of the reduction of conventional new capacity that can be offset with wind power. And the emissions benefit is valued based on methods and data developed by the California Energy Commission.

Figure 4 shows the generation of a large utility during the annual peak day. With a hypothetical wind plant, as depicted in the graph, the generation of electricity using gas and oil units can be reduced. Large base load units, such as nuclear and some other thermal plants that cannot easily alter their output, are typically run at a constant level, as can be seen in the graph. A different pattern of wind power production would change the quantities of gas, oil,

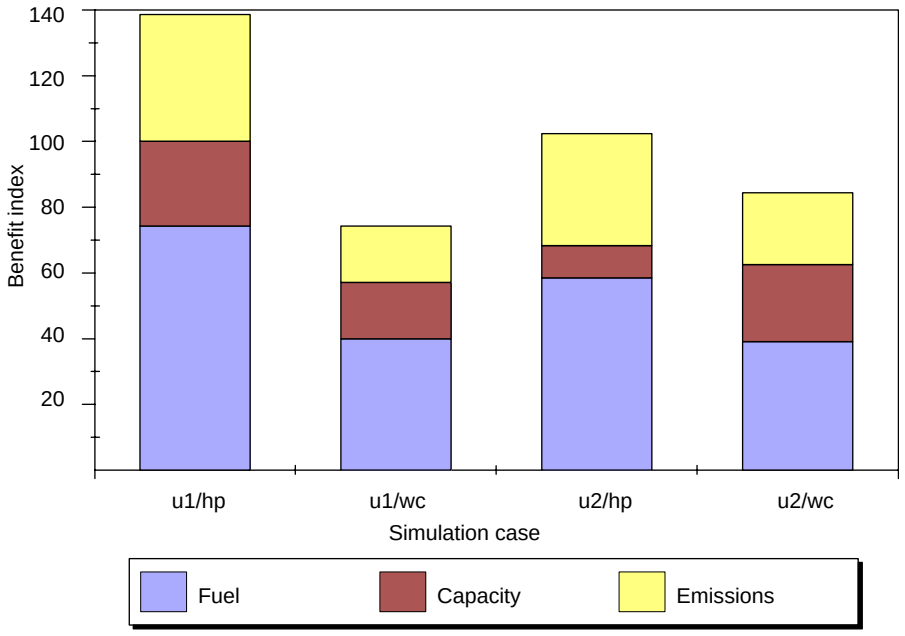


FIGURE 3 Benefits of wind power, separated into fuel savings, capacity credit, and avoided emissions, valued at prices used by the California Energy Commission. hp, High Plains; wc, West Coast.

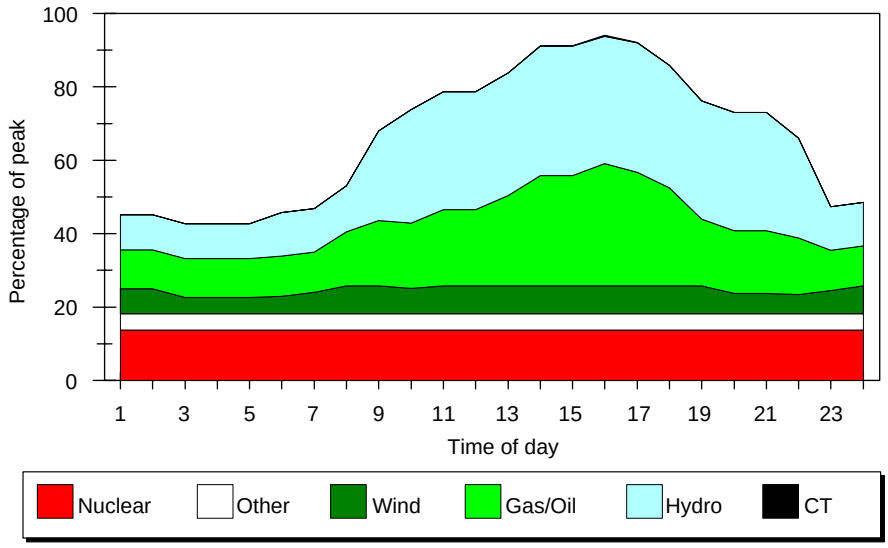


FIGURE 4 Economic dispatch of a peak day with wind energy. CT, combustion turbine capacity.

and coal required, and this could have a significant impact on the value of the wind power plant.

4. ROLE OF WIND FORECASTING AND ITS RELATIONSHIP TO THE VALUE OF WIND

Wind power is sometimes characterized as an intermittent and unpredictable resource. Although it is intermittent, the output of a wind power plant can be predicted from several minutes to several days in advance. For forecast horizons longer than a few hours, accuracy falls significantly, but newly emerging forecasting techniques can enhance the economics of wind power plants. Wind forecasting models that appear to be the most accurate are actually hybrid forecasting models. The first stage of the model is to predict local weather conditions for the forecast period. The second stage of the model applies various statistical methods that adjust the meteorological model output so that the wind power can be predicted. Statistical techniques that appear to be most promising are autoregressive moving average models, formally known as ARIMA models, and neural network models.

With a reasonably accurate wind forecast, power system operators can do a better job of scheduling enough generating resources to meet the electrical load. Accurate wind forecasts make it possible to schedule just the right level of generation, whereas inaccurate forecasts can result in either too much or not enough generation online to service the load. Assessing the cost of excess generation is conceptually straightforward and involves calculating the cost of running the generators that were not needed to supply the electrical demand. Accounting for the cost of insufficient generation might be easily assessed if a neighboring utility can provide the extra demand. But in other cases, it can be more difficult to assess the cost of insufficient generation. In a more extreme case, the generation shortfall may be large enough to cause outages in the system. Even if outages do not occur, there is a cost of increasing risk of such outages, with an economic value that can be difficult to quantify.

Figure 5 illustrates the forecast benefit for various levels of accuracy for a large utility. The graph is not symmetrical due to the specific characteristics of the MC curve. Although perfect forecasting is not feasible, the graph shows that good forecasts can increase the value of a wind plant.

5. EFFECTS OF GEOGRAPHICALLY DISPERSE WIND SITES

The economic value of wind power plants can be compromised by intermittency and inaccurate forecasting. Although this is a function of the specific characteristics of the wind resource, energy demand, and electrical supply, it is an issue that has received considerable attention. There are several ways in which to reduce this intermittency to help maximize the economic benefit to the electrical supply, and one of the most important ways is to take advantage of geographic dispersion during the design and planning

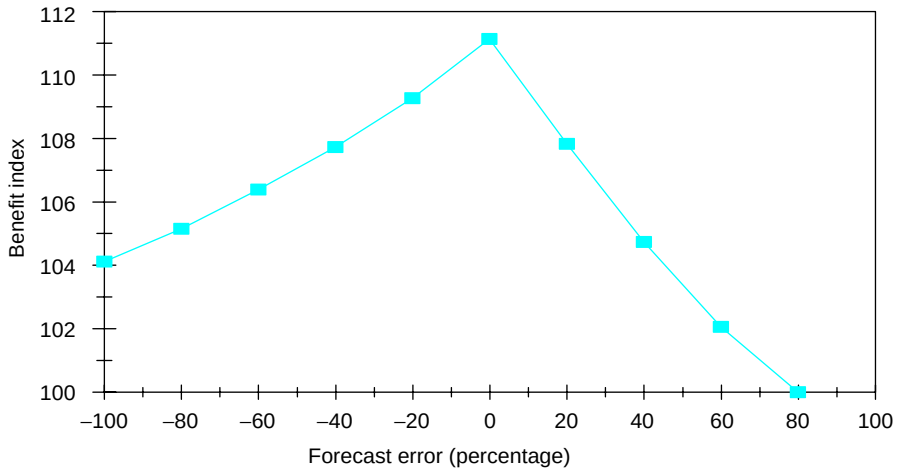


FIGURE 5 Value of a wind forecast at various accuracies during a period when wind output averages 50% of rated capacity.

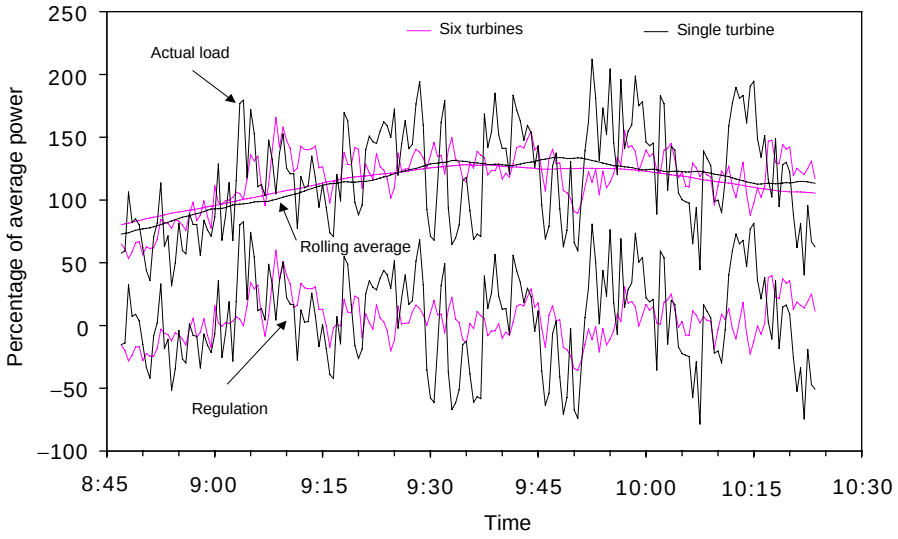


FIGURE 6 Smoothing effects of multiple wind turbines over time scales ranging from minutes to 2 h.

stages of the wind plant. The principle of geographical dispersion is simple and based on the observation that the wind speed varies over progressively larger areas. This principle applies to the many turbines within a single wind farm as well as to multiple wind plants that may be separated by up to hundreds of miles.

Figure 6 shows the impact of geographic diversity on a small time scale with only six wind turbines. The jagged lines that show the single turbine output can be compared with the significantly smoother line showing the output of the group of turbines. The figure also shows an estimate of the regulation from conventional generators that could offset the small perturbations in the output of this small wind farm.

Smoothing also occurs over larger time and distance scales. Figure 7 shows the smoothing effect over hypothetical wind farms in Iowa, based on actual time-synchronized wind speed data. Given a hypothetical wind plant of 25 MW, the graph shows the hourly variation that would occur during the month of July if the 25 MW were located at each of the 12 sites, one at a time. The final graph bar shows the dramatic reduction in the hourly output variation that would occur if the 25 MW of wind capacity were evenly split among the 12 sites.

In fact, achieving the smoothest aggregate output might not produce the best economic benefit. Economic studies have used advanced optimization techniques, such as fuzzy logic and genetic algorithms,

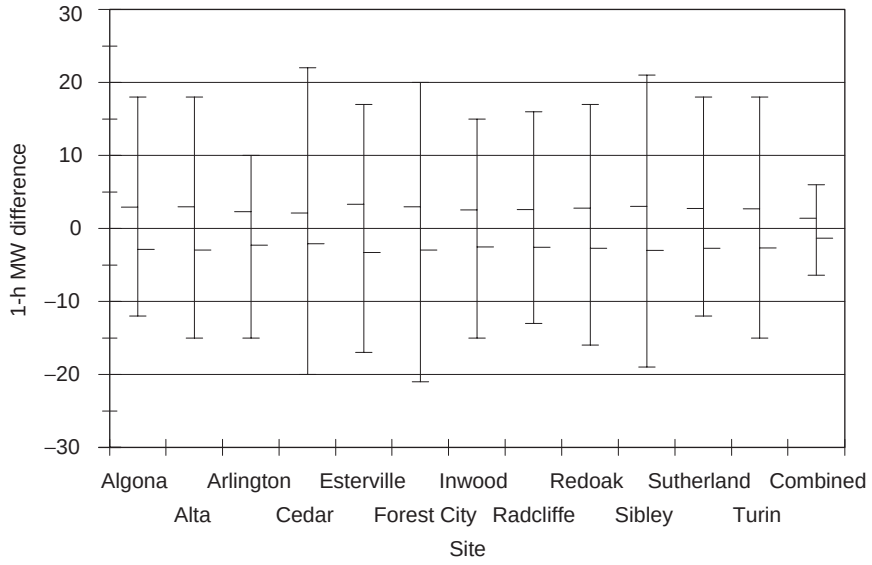


FIGURE 7 Reduction in hourly wind power output when the wind capacity is spread among 12 sites in Iowa compared with output fluctuations at single sites. Bars show maximums, minimums, means plus standard deviations, and means minus standard deviations.

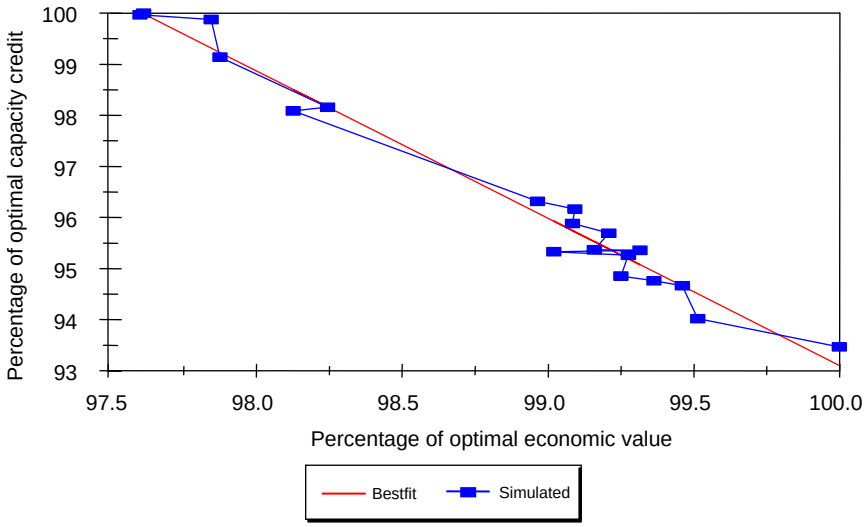


FIGURE 8 Trade-off curve between reliability and least cost power generation.

to explore the economically optimal allocation of wind capacity among multiple wind sites. The results of these analyses imply that this optimization depends on the characteristics of the specific wind regimes and on the characteristics of the remaining generating supply and the electrical demand.

The economically optimal site combination might not maximize system reliability. Figure 8 illustrates this trade-off and is based on the Iowa analysis. The graph shows a scatter of points, each of which represents a particular capacity distribution among the 12 sites. Because each capacity distribution has different reliability implications and differing economic benefits, the data can be plotted, resulting in this production possibility trade-off curve. Analyses such as these can be complex, but they can reveal important aspects of the economic viability of a wind power plant.

6. RESERVE AND REGULATION IMPACTS

As wind energy systems become more prevalent, there is increasing interest in the effect of these plants

on system reserves and regulation. Because electricity cannot be economically stored in large quantities, power system operators ensure that there is additional generating capacity that can be called on on short notice to meet unexpected increases in demand or to deal with outages of other generators or key transmission lines. This extra capacity is called "reserves." A contributing factor to the need for reserves is the overall reliability of the power supply. The intermittent nature of wind power raises the issue as to whether large quantities of wind might have an impact on the reserve level that should be available for meeting changing system demand requirements. If wind power plants were to cause reserve levels to increase significantly beyond what would be needed for competing generators, that would have a considerable impact on the economics of wind plants because additional capacity would be required to support them. However, system reliability studies that take into account the wind forecasting technology and statistical patterns of wind generation find that wind plants have only a small impact on the overall system reliability level and, therefore, do not significantly alter the level of system reserves required for reliable system operation. As more new wind power plants are developed and operated as bulk power sources, additional insights will be gained on this issue.

Similarly, there is little empirical evidence concerning the impact of wind power on regulation requirements. Regulation is the small movement in generators that compensates for very short-term variations in electric demand. If wind plants would require other generation to compensate for fluctuations in output in the seconds-to-minutes time frame, the economic value of wind would be less than if these fluctuations had no impact. Because these short-term fluctuations are generally random and uncorrelated with load fluctuations, many experts believe that regulation impacts are not significant impediments to the economic value of wind plants and current and planned rates of wind penetration. Because of the statistical independence of these fast wind and load fluctuations, it is likely that regulation impacts would not increase significantly at higher wind penetrations.

7. OTHER ISSUES: LOCATION OF WIND/LOADS/TRANSMISSION

The economic value of wind plants also depends on other factors, some of which have not been subject to systematic evaluation due to the emerging nature of this technology. Siting a prospective wind plant involves many considerations, including local permitting requirements and public reaction. Anecdotal evidence suggests that these issues can affect the overall cost of energy from wind, but they have not been systematically quantified.

One of the most important issues is the proximity of the wind plant to transmission facilities. Because wind resources are not typically found near load centers, building a large transmission line to move the power to market could potentially result in the wind plant becoming uneconomical. In transmission-constrained regions, wind plants might not be able to take advantage of periods when transmission capacity is available. Market rules that cover transmission access and financial settlement must be transparent and neutral to technology.

8. CONCLUSIONS AND SUMMARY

The economics of wind-generated electricity depends on several factors. Cost is primarily technology driven and is influenced by production technology and the effect of turbine reliability on O&M costs.

By far the most significant factors that contribute to wind energy value are related to the wind resource and the characteristics of the grid and the evolving market rules. As additional wind capacity is developed, these variables will be quantified more precisely.

SEE ALSO THE FOLLOWING ARTICLES

Electric Power Reform: Social and Environmental Issues • *Hydropower Resources* • *Renewable Energy in Europe* • *Renewable Energy in the United States* • *Renewable Energy, Taxonomic Overview* • *Wind Energy Technology, Environmental Impacts of* • *Wind Energy, History of* • *Wind Farms* • *Wind Resource Base*

Further Reading

Billinton, R., Chen, H., and Ghajar, R. (1996). A sequential simulation technique for adequacy evaluation of generating systems including wind energy. *IEEE Trans. Energy Conversion* **11**, 728–734.

Brooks, D. L., Lo, E. O., Smith, J., Pease, J., and McGree, M. (2002). "Assessing the impact of wind generation on system operations at Xcel Energy–North and Bonneville Power

Administration. Paper presented at Windpower 2002 [on CD-ROM].” American Wind Energy Association, Washington, DC.
Ernst, B. (1999). “Analysis of Wind Power Ancillary Services Characteristics with German 250-MW Wind Data.” National Renewable Energy Laboratory, Golden, CO.
Gilman, B., Cheng, M., Isaac, J., Zack, J., Baily, B., and Brower, M. (2001). “The value of wind forecasting to Southern California Edison. Paper presented at Windpower 2001 [on CD-ROM].” American Wind Energy Association, Washington, DC.
Hirst, E. (2001). “Interactions of Wind Farms with Bulk-Power Operations and Markets.” Project for Sustainable FERC Energy Policy, Washington, DC.
Hudson, R., Kirby, B., and Wan, Y. (2001). “Regulation requirements for wind generation facilities. Paper presented at Windpower 2001 [on CD-ROM].” American Wind Energy Association, Washington, DC.
Milligan, M. (1999). Modeling utility-scale wind power plants: I. Economics. *Wind Energy* **2,** 167–193.
Milligan, M. (2000). Modeling utility-scale wind power plants: II. Capacity credit. *Wind Energy* **3,** 167–206.
Milligan, M., Caldwell, J., Porter, K., and Parsons, B. (2002). Wind energy and power system operations: A survey of current research and regulatory actions. *Electricity J.* **15,** 56–67.
Milligan, M., and Factor, T. (2000). Optimizing the geographic distribution of wind plants in Iowa for maximum economic benefit and reliability. *J. Wind Eng.* **24,** 271–290.
Milligan, M., Miller, A., and Chapman, F. (1995). Estimating the economic value of wind forecasting to utilities. *In* “Proceedings of the Windpower ’95 Conference,” pp. 285–294. American Wind Energy Association, Washington, DC.
Strbac, G., and Kirschen, D. (2000). Who should pay for reserve? *Electricity J.* **13,** 32–37.

Wind Energy, History of

MARTIN PASQUALETTI
Arizona State University
Tempe, Arizona, United States

ROBERT RIGHTER
Southern Methodist University
Dallas, Texas, United States

PAUL GIPE
Paul Gipe & Associates
Tehachapi, California, United States

1. **Harnessing the Wind**
2. **The First U.S. Boom**
3. **The European Emphasis**
4. **Rejuvenated North America**
5. **Trends**
6. **Conclusion**

Glossary

American farm windmill Chicago mill or simply farm windmill; it is a multiblade mechanical wind pump, initially made of wood and later made of metal, developed during the late 19th century in the mid-western United States and subsequently copied world-wide.

post mill Traditional European mechanical windmill for grinding grain and other materials (e.g., tobacco, paper) where the tower body is supported on a vertical post resting on a wooden trestle.

tower mill Traditional European mechanical windmill used for grinding grain and other materials as well as pumping water where a tower of wood or masonry remains stationary and only the cap with the rotor attached turns to face the wind.

windchargers Electricity-generating wind turbines used to charge batteries at remote homesteads; they were prominent before the advent of widespread distribution of central station electricity.

wind farms (or wind power plants) Multiple wind turbines or arrays of wind turbines used to produce bulk electricity for the grid.

By the first decade of the 21st century, wind power had become the best hope for the future of alternative energy. Of all the possible "new" sources of electricity, it had rather unexpectedly found itself the most touted and fastest growing alternative energy resource in the world, generating significant amounts of electricity in several countries, such as the United States, Denmark, Germany, and Spain, as well as impressive amounts in many other countries. The new era of wind development was led by the United States during the 1980s, but Europe has overtaken that ranking, accounting for two-thirds of total worldwide wind development by itself in 2001. Now a global multi-billion-dollar industry, wind energy is regaining its once prominent place in the energy firmament. For millennia, wind power has been used for everything from kite flying and propulsion to grinding and pumping, but never before has it been so important for the generation of electricity. It is an old resource applied to a new mission with unprecedented success. This article describes how wind power came to achieve this status.

1. HARNESSING THE WIND

1.1 Prehistory Use

For longer than one can grasp, winds have been churning seas, carrying silt, eroding continents, creating dunes, rustling leaves, and filling sails. They are responsible for loess soils in China, sand seas in Libya, and dust storms in Arizona. They have created, destroyed, altered, eroded, and changed the shape of land, the vegetation that would cover it, and the lives of millions of people.

 © 2004 Elsevier Inc. All rights reserved.

Wind energy exists without end, yet at no time have humans ever employed more than a tiny fraction of its kinetic energy. Its power is so beyond human control that cultures throughout the world have paid it homage in their legends and mythologies. Creation stories of nearly all cultures involve the power of the wind.

Like the mystery of who invented the wheel, we do not know when humans first employed the force of the wind to do work. However, in terms of utility, sailing would seem to be one of the most likely early uses, perhaps 40,000 years ago when Asians migrated to greater Australia. Although we know little of these voyages, the far later water journeys of Polynesians in their double-hulled sailing craft are unquestioned. For these early sailors, wind provided the kinetic energy for the exploration and settlement of the Pacific Ocean islands, including the Hawaiian Islands, where it later led to the invention of surfing, the earliest use of wind-powered recreational activity.

1.2 Historic Uses of the Wind

As in prehistory, the first historic use of wind power involved its use in transportation. Egyptians plied the Nile River as early as 3100 BC using craft equipped with sails of linen and papyrus. In *The Odyssey,* Homer wrote of Odysseus, who sailed the Ionian and Aegean seas, eventually making a crucial error when he angered the god of the winds. Aeolus stopped the wind for 6 days, forcing Odysseus and his crew to rely on rowing, a very poor substitute. "No breeze, no help in sight, by our own folly—six indistinguishable nights and days," lamented Odysseus. The story of Odysseus reminds one of the capricious nature of the wind. Although it can be used, it cannot be put under human control any more than ocean currents can be put under human control. For this reason, wind has been successfully employed only sparingly, with inventors and engineers preferring power sources that they can manipulate as needed. Yet the importance of wind energy to human history is undeniable. From the time of Columbus until the middle of the 19th century, nations depended on sailing vessels. Advances in navigation and sail design combined to make the great age of the sailing craft feasible. In the billowing sails of thousands of ships for thousands of years, the following paragraphs trace the paths of commerce and conquest.

Whereas once the longest history of using wind power was on the sea and not the land, today the reverse is true. While the importance of sailing has faded to recreational use, the terrestrial applications of wind have multiplied. Where and when did people first turn to the wind to help in their harvests? We do not know exactly, but we do know that in the 10th century AD, windmills were turning in the blustery Seistan region of Persia (now Iran). For this, we have both artifact proof and written evidence. These were primitive windmills by modern standards, using vertical sails of reed bundles, and their owners built them to grind grain and lift water from streams to irrigate gardens. For centuries, these rudimentary yet ingenious machines were carried to other parts of the world, including India and China, where farmers employed them to pump water, grind grain, and crush sugarcane.

Independently, the post mill arose in England. Unlike the Persian sails, which follow a carousel's path around a vertical axis, the sails of the traditional European windmill follow a path around a horizontal axis. The whole assembly of the post mill, including the blades, axle, and milling cabin, rests on a massive vertical post. The miller orients the post mill into the wind by swinging the entire windmill around the post, hence the name post mill. In 1137, William of Almoner designed and constructed the first such mill in Leicester, England. For many years, historians of technology assumed that the English post mill idea was a by-product of the Christian Crusades, with the English post mill being inspired by the Persian horizontal windmills of Seistan. However, the designs of the two mills are very different, and the evidence points to multiple inventions rather than diffusion from a single source.

Environmental and social reasons help to explain why the post mill took root where it did in England. Much of the land in Leicester is level, limiting the use of water wheels. However, perhaps just as significant was social class. In England, the kings had vested the water rights in the nobility and in the established church. These rights were jealously guarded, and in a practical sense they gave control of energy production to the nobility. In this earlier world of vested and hereditary privilege came a new source of energy, one that the upper class could not control. In such an atmosphere, the wind became the great leveler of class. As the middle-class entrepreneur Herbert of Bury, Suffolk, stated in 1180, "The free benefit of the wind ought not be denied to any man."

Spreading throughout England, the use of the post mill eventually made its way to the European continent. By 1300, one saw horizontal axis windmills similar to those in England in Spain, France, Belgium, The Netherlands, Denmark, the German principalities, and the Italian states. Subsequently,

the more powerful and more elaborate tower mill replaced the simple post mill, especially in The Netherlands.

As the name implies, tower mills reached skyward by using wood framing, brick, and stone. The tower, like those of the modern wind turbine, was immovable. The cap with the rotor, horizontal axle, and gearing rested on a curb at the top of the tower. The miller pointed the rotor into the wind by swinging the cap on its curb. During the "golden age" of the European windmill, the Dutch alone operated some 10,000 tower mills, and they did more than pump water and grind grain. For example, they reduced pepper and other spices, cocoa, dyes, chalk, and paint pigments. Lumber companies employed them as primary power for sawmills. Paper companies used windmills to reduce wood pulp to paper. One authority estimated that the mills provided as much as 25% of Europe's industrial energy from 1300 to the coming of the steam engine and cheap coal during the 1800s. The remainder came from hydropower, assisted by human and animal labor.

1.3 Wind Energy in the New World

Although European countries explored and colonized the New World, there is disagreement as to whether the windmill accompanied them. The seal of New York City contains a windmill, early woodcuts of "Nieuw Amsterdaam" depicted windmills, and there are other examples of early windmills as well. But the European windmill was simply too large, cumbersome, and expensive to be easily adapted to New World conditions. Also, it was a labor-intensive device in a new land where labor was scarce and valuable. Any apparatus that required continual attention would have found limited use on North American soil. Furthermore, the climate and topography of the New England region favored water wheels, and by 1840 more than 50,000 water wheels were in operation. It was the energy source of choice.

The windmill did eventually find a home in the United States, albeit of a design different from that used in Europe, in the vast western part of the country. Inventors there came up with a blueprint for what would become known as the "American farm windmill." Adapted to a land of little rain but abundant wind, the American farm windmill was small, light, movable, self-regulating, inexpensive, and easy to maintain. It liberated groundwater in an arid area, bringing it to the surface and making agriculture possible. Between 1850 and 1900, an impressive and bewildering number of styles and brands of farm windmills appeared on the market. There is no way in which to estimate accurately how many such windmills once dotted the Great Plains, but some authorities have suggested a number as high as 6 million. These windmills added a vertical dimension to the otherwise horizontal space of the Great Plains and once were the most common landscape feature between the Mississippi and the Rockies. Today, the sight of an American windmill evokes a certain nostalgia for a simpler bucolic age. The windmill became an American icon.

1.4 The Marriage of Wind and Electricity

The development of wind power to grind grain and pump water was historically significant, but during the 21st century it has been the marriage of wind and electricity that offers the greatest promise. However, it was not a union that took place quickly. In 1748, Benjamin Franklin wrote a scientific friend that, in regard to electricity, he was "chagrined a little that we have hitherto been able to produce nothing in the way of use to mankind." It would be many years before the first practical use of electricity appeared in the form of the telegraph in 1844.

Although the telegraph was used extensively for more than half a century, little more was done with electricity until the first generating plants began operating during the final quarter of the 19th century. After that time, the use of electricity caught on quickly, and by 1900 it was becoming an essential element of the urban lifestyle. In this atmosphere of discovery, inventors contemplated the coupling of wind power and electricity. Journals such as *Scientific American* challenged American ingenuity to use "the natural forces which are in play about us." The most pressing problem was how to store this elusive form of energy. How could one use the power of the wind if one could not store electricity, especially given that the wind did not necessarily blow when one needed it? By the early 1880s, the French inventor Camille Faure had devised what he termed a "box of electricity" or storage battery. It seemed possible to generate electricity from the wind, store it, and use it when needed.

The first to build a practical large-scale wind turbine was Charles Brush, a scientist from Cleveland, Ohio, who had made a fortune through his electric arc lighting system. The basement of his Euclid Avenue mansion became his laboratory, but his experiments were limited by the lack of electricity. To resolve the problem, Brush in 1886 designed and constructed a wind turbine that was

of immense proportions. Visitors to his 5-acre backyard saw a 40-ton (46,300-kg) tower rising 60 feet (18.3 m) into the air. From the tower turned a rotor 56 feet (17.1 m) in diameter with 144 slender blades. Within the tower, he located his dynamo and the necessary gearing to drive it. In his basement, Brush installed 12 batteries. The turbine and batteries infallibly provided electricity for his 100 incandescent lights, three arc lights, and a number of electric motors. The Brush wind dynamo worked incredibly well for approximately 15 years, using "one of nature's most unruly motive agents," as *Scientific American* reported. The inventor used his machine only occasionally after 1900, which was the year Cleveland began offering the convenience of centrally generated electricity. He abandoned his machine in 1908.

Despite its success, the Brush wind dynamo was never duplicated. Although fuel was free, the machinery to capture it was not. Furthermore, Brush's extensive knowledge of both electricity and the working of the wind dynamo could not be easily replicated elsewhere. Although a smaller similar design was employed in New England by the Lewis Electrical Company (the first commercial wind turbine), it did not prove to be popular. Brush offered the public a concept, but not one that could be mass produced.

1.5 The Search for a Practical Wind Turbine

Experimentation continued sporadically between 1890 and 1920, but these decades represented the pause between invention and application. It was only after World War I that some mechanically minded Americans applied advances in aeronautics to the design of a practical inexpensive wind turbine. In 1920, few North American farmers had electricity, and those who did produced it with gasoline-powered generators such as Delco "light plants." By the late 1920s, wind turbines began to compete in the effort to provide the American farm family with the convenience of electricity. A number of companies produced wind turbines, but the most successful were Jacobs's Windelectric and Wincharger. Joe and Marcellus Jacobs were not engineers but rather farm boys with an inventive flair. They experimented on their parents' eastern Montana ranch. At first, they converted a multibladed farm windmill from pumping water to generating electricity, but it turned too slowly. Marcellus had learned to fly, and he soon realized that a three-bladed propeller might be a better solution to the problem of wind-powered electrical generation. The brothers eventually perfected the rotor blades, the blade feathering system, and a powerful generator that would characterize their wind turbine. Soon their ranch neighbors were purchasing the Jacobs's turbines. Uniquely, all of their experimentation and manufacturing took place on the family ranch. It was a perfect wind laboratory, but soon the brothers realized that truly building turbines in quantity would require relocation to a midwestern industrial center. They chose Minneapolis, Minnesota, in 1927, and between that year and 1957 the Jacobs brothers turned out approximately 30,000 small wind turbines. The Jacobs machines became legendary for their reliability. A Christian missionary in Ethiopia installed a Jacobs plant in 1938. In 1968, the missionary sent for a replacement set of generator brushes—the first repair in 30 years of operation. In 1933, Admiral Byrd installed a Jacobs plant on a 70-foot (21.3-m) tower at Little America. When Byrd abandoned his Antarctica outpost, he left the wind turbine. When his son, Richard Byrd, Jr., visited the site in 1947, the Jacobs was "still turning in the breeze," although only 15 feet (4.6 m) of the tower was free of ice. In 1955, one of the 1933 veterans returned and removed the blades from the still operating turbine because they were threatened by the advancing ice.

Although Jacobs marketed its wind turbine as the "Cadillac" of windchargers, many ranchers and farmers found that a "Chevrolet" would be sufficient. In general, they purchased a Wincharger brand from a company that produced models from 6 to 110 V and from 200 to 3000 W. R. E. Weinig, the general manager of Wincharger, testified in 1945 that some 400,000 of the company's wind plants operated worldwide. Other brands included the Miller Airlite, Universal AeroElectric, Paris–Dunn, Airline, Wind Kind, and Winpower. In 1946, the Sears Roebuck catalog marketed a Silvertone Aircharger for $32.50. Most of these turbines were small, powering a radio and perhaps a couple of 40-W lights, but nevertheless they filled the desire for electricity in rural regions of North America.

Meanwhile, active research on larger wind turbines was under way in Europe. In 1900, the world's expert on electricity-generating wind turbines was not Charles Brush but rather Poul La Cour, a Danish scientist who spent most of his professional life working on wind electric systems. By 1906, with the support of the Danish government, 40 wind turbines were generating electricity in the small nation,

beginning a tradition that has lasted to this day. To the south, Germany capitalized on La Cour's work and moved forward under the leadership of engineer Hermann Honnef and others. The editor of *Scientific American* noted that wind power in Germany "has been developed to a point that is surprising." Across the channel, the British showed a similar interest. In 1924, Oxford University's Institute of Agricultural Engineering tested seven turbines from five manufacturers. Two years later, the institute reported that the "cost of windmill-generated electricity for small lighting and small power purposes was found to be quite reasonable and, as such, justifies its wider use in rural districts where there is no general supply."

1.6 The Victory of Centralism over Individualism

By the 1930s, wind turbines had found a particular place in the rural landscape in both Europe and the Americas. However, as noted in the Oxford report, their use was recommended only "where there is no general supply" of electricity. In other words, engineers and entrepreneurs had become committed to central power systems and the creation of a grid of transmission lines to carry the electricity. The idea of individual power units was anathema to engineers who thought in terms of electrical systems that served large populations from a central generating source. Of course, private utility companies complained that getting power to rural America was costly. Stringing wires was expensive, and utility companies refused to serve rural homes unless they were in close proximity to each other and to urban centers. In 1930, only approximately 10% of American farm families could boast electricity despite a much higher percentage in urban areas. It was at this time that the Rural Electrification Act (REA) came into existence.

The REA, passed in 1936, would change these statistics dramatically in the United States. The act called for local farmers to establish cooperatives with the authority to make loans to bring electrical power to farms within their designated region. It was enormously successful. No one would deny the accomplishment of the REA, but the federally backed cooperatives sounded the death knell for the windcharger industry. Centralization and government subsidies for the cooperatives killed the industry. One wind energy executive attempted to persuade Congress to include windchargers as part of the REA in more isolated areas, but neither Congress nor REA administrators would consider such a deviation from their commitment to a centralized power system. By 1957, every American wind electric company had closed its doors.

1.7 Large Wind Machine Experiments

While federal legislation stilled the small, farm-oriented wind turbines, one engineer was thinking on a much grander scale. A young engineer named Palmer Cosslett Putnam first became interested in wind energy in 1934 when he built a house on Cape Cod and, as he put it, "found both the winds and the electric rates surprisingly high." Putnam had in mind hooking up to utility company power, using it but sparingly, and selling back the surplus from his wind turbine. However, no mechanism existed at that time to convert direct current from a windcharger to alternating current that central station power plants delivered. Putnam researched the problem, and as an engineer committed to "economies of scale," he decided to build a huge experimental wind turbine that would generate alternating current identical to that from conventional power plants.

Putnam stripped bare a rounded knoll near Rutland, Vermont. There he erected a massive two-bladed turbine with S. Morgan Smith Company, a manufacturer of large hydroelectric turbines. Truckers struggled to bring some 500 tons (90,700 kg) of material and parts to the crest of Grandpa's Knob. The 70-foot (21.3-m) length of the two 7.5-ton (6800-kg) blades posed a particular challenge. However, on October 19, 1941, the rotor began to turn. It operated for approximately 16 months, producing 298,240 kWh in 695 h of online production. On February 20, 1943, a bearing failed, resulting in a 2-year hiatus because both parts and labor were difficult to obtain during World War II. Production finally resumed, but not for long. On March 26, 1945, one of the blades separated from the rotor and crashed to the ground and on down the side of the mountain. Although Putnam promised that he and his partners would repair the turbine, the expense of doing so was prohibitive. The great turbine never turned again.

What is most remarkable about the Smith–Putnam project was not its successful operation but rather the fact that it happened at all. Putnam had to rely on private industry for machinery, expertise, and (above all) funding. Fortunately, S. Morgan Smith bankrolled his effort. A number of distinguished scientists from prominent universities participated, as did some of the leading engineering firms of the

day. Vannevar Bush, then the dean of engineering at the Massachusetts Institute of Technology, put it well when he wrote that the project was "conceived and carried through free enterprisers who were willing to accept the risks involved in exploring the frontiers of knowledge, in the hope of financial gain." Obviously, there was no financial gain, but in its brief life the turbine certainly opened new frontiers of knowledge for the wind energy field.

2. THE FIRST U.S. BOOM

2.1 The Search for Alternatives

For less than 22 years, Americans held hope in the promise that nuclear power was going to provide all of the electricity they could ever need, but it was not to be. The idea that blossomed with such promise at the Shippingport Atomic Power Station west of Pittsburgh, Pennsylvania, in December 1957 ironically shriveled in the same state with the accident at the Three Mile Island nuclear power plant near Harrisburg in April 1979. The event, including the chaotic public and governmental responses it triggered, punctured the nuclear balloon, resulting in the cancellation of dozens of nuclear power plants. None has been ordered in the United States since then, and after the 1986 catastrophe at the Chernobyl plant in Ukraine, few have been built anywhere in the world. These events, coupled with the embargoes on oil and the constraints on coal burning during the 1970s, stimulated a more serious evaluation of alternative energy resources.

Although several resources did garner more attention, there was never a wholehearted effort. Nascent public and private alternative energy programs existed in scattered, small, and even persistent forms in many areas, but they received little more than perfunctory support from government agencies and large energy companies. Even the high-profile advocacy of the Carter administration did not have much staying power in terms of national policy, although it helped to build a foundation for later advances in some locations, especially California.

California was famously short on energy resources, and these deficiencies were to become painfully obvious over the years. During the second half of the 19th century, people relied on wood, but it became scarce rather quickly. Various forms of hydrocarbons were used for lighting, and when electricity became popular, hydropower was quickly put to work. However, its practical limits were also soon reached as California continued to grow. Soon the state ranked first in economic might, and feeding its appetite required huge amounts of electricity. Having virtually no coal, insufficient natural gas, and oil that was too expensive to use in power plants, and facing rising opposition to nuclear energy, Californians turned to alternative resources that have always been available in the state. At first, geothermal energy was the resource of choice, and it grew quickly. Yet ultimately its luster dulled, and by the mid-1980s serious attention was shifting to another resource—wind.

The attention given to wind power in California resulted not from the fact that the state was a dominantly windy place but rather from the convergence of local demand, state and federal financial subsidies, a favorable political climate, and the geographical convenience that found three sites with great promise virtually in the backyards of millions of people. Anyone who has driven the stretch of Interstate 10 through San Gorgonio Pass 100 miles east of Los Angeles, or the stretch of Interstate 580 through Altamont Pass 50 miles east of San Francisco, knows the strength and persistence of winds there. Even windier, albeit less traveled, Tehachapi Pass 50 miles east of Bakersfield was the third target of early development (Fig. 1).

With all of the favorable conditions for wind development finally in place by the mid-1980s, thousands of wind energy conversion systems (WECS) sprouted with a suddenness that startled the public. In unexpected numbers and with unexpected vigor,

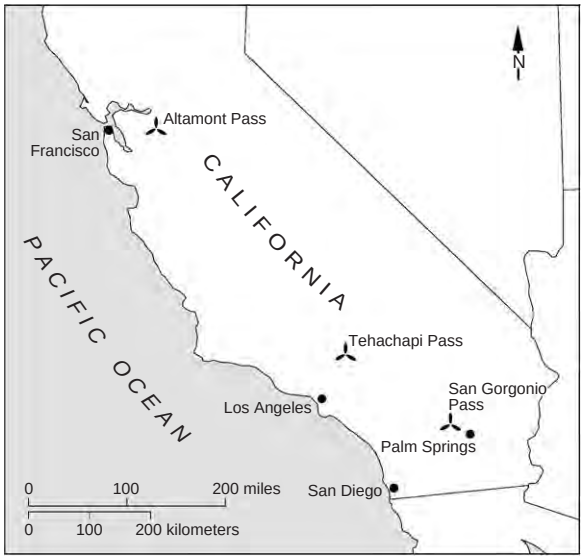


FIGURE 1 Principal developed wind resource areas of California.

people began to complain how the wind turbines changed aesthetics, increased bird mortality, produced electronic interference, and created noise. However, there was no turning back; California's experiment in large-scale development of wind power was just getting started.

2.2 Reactions and Early Lessons

The early stages of modern wind development in California faced some persistent barriers. Wind turbines are unavoidably visible and impossible to hide, especially in those locations where they were being concentrated in the state. Complaints about these new landscape additions continued, although they diminished in intensity as the technology improved and regulatory controls were imposed. Once financial subsidies ceased during the mid-1980s, wind power was left to survive on its own, and this was not possible for all installations or designs. A weeding out would soon take place as wind developers continued to face public complaints, technical difficulties, and financial challenges. Poorly sited and inoperable equipment was slowly removed, and regulators began flexing their bureaucratic muscles. The authorities who governed development in places like San Gorgonio Pass imposed controls on height, color, and reflectivity, instigated operational standards on design and noise, and required decommissioning bonds and protection of migrating birds and endangered species.

Despite the multiple regulations and engineering improvements that are now part of wind energy development in the United States, nothing can render them invisible and, at the same time, still economical to run. Simply put, the only way in which to hide wind turbines is by using the buffer of space. But in California, most wind turbines are along busy highways, always available for public scrutiny. Because of the slipshod way in which many developments were first installed, wind power received an early black eye that has taken years to heal. In the meantime, development activities picked up elsewhere, especially in Europe.

On the heels of the California experience and benefiting from improvements prompted by that experience, European wind developers began moving quickly to install their own projects. Their ambitions were encouraged by shortages of sufficient local resources, concerns about nuclear power and global warming, the political weight of "green" parties, and ample subsidies that became available to alternative energy developers. All of these factors taken together were to push wind development in Europe ahead of those in the United States during the 1990s, but Europeans' attraction to wind power had actually started years earlier.

3. THE EUROPEAN EMPHASIS

3.1 The Early Years

Denmark has assumed a leadership role in the development of wind power in Europe, but it took a path far different from that taken by the United States. Unlike American wind turbine design, which was concentrated in the hands of the aerospace industry, Danish wind technology grew out of the agricultural sector as a natural by-product of the Danish economy and the long history that Danes have with the wind. They used it to power the ships of their early conquests and to carry Norsemen as far as the New World. So-called "Dutch" windmills eventually became a common sight in Denmark, with a some 3000 of them providing the equivalent of half again as much energy as all the animal power then supporting Danish agriculture by 1890. In 1891, La Cour, the "Danish Edison," began experimenting with wind-generated electricity, and as early as 1903, La Cour's Danish Wind Power Society was fostering the idea of wind-generated electricity.

By the end of World War I, more than one-fourth of all rural power stations in Denmark used wind turbines. During the long wartime blockade, the 3 MW provided by these crude wind generators and the widespread use of small farm windmills for grinding grain were invaluable to the impoverished rural population. Although most windmills were used for mechanical power, it has been estimated that wind turbines were providing the equivalent of 120 to 150 MW in Denmark by 1920. Nearly 90 turbines were installed during World War II, including the 30-kW Lykkegaard wind turbine patterned after La Cour's *klapsejlsmølle*, and F. L. Smidth's more modern "Aeromotors." This design, like those of La Cour, used four or five wide blades resembling those of traditional windmills except that the blades were covered with metal shutters instead of cloth sails. These wind turbines were controlled by opening and closing the shutters. In contrast, the Smidth company became one of the world's first firms to marry the rapidly advancing field of aerodynamics to wind turbine design. The Smidth company's wind turbines incorporated both technologies and used modern airfoils upwind of a concrete tower.

After the war, interest in wind energy again waned, although Johannes Juul remained resolute. In 1950, he began testing a prototype wind turbine for the Danish utility SEAS, and subsequently he modified a Smidth turbine used on the island of Bogø. With the experience gained from developing these two machines, Juul began work on his crowning achievement, the three-bladed, stall-regulated, upwind rotor at Gedser that spanned 24 m (79 feet) in diameter. Installed in 1956, the Gedser mill operated in regular service from 1959 through 1967. By the late 1970s, modern wind turbine manufacturers and experimenters erected their first prototypes. By 1980, when the European wind energy conference was held north of Copenhagen, there were several designs in operation.

Meanwhile, Ulrich Hütter was designing and building wind turbines for the Nazi-owned Ventimotor company outside of Weimar, Germany, in the hope of reducing reliance on foreign sources of energy. Hütter's team experimented with several designs, including one of the first uses of a wind turbine to drive an asynchronous or induction generator directly coupled to the electric utility network. After the war, Hütter went on to become a prominent aeronautical designer who continued experimenting with wind energy. During the late 1950s, he installed a novel turbine that operated at high rotor speeds with only two slender blades.

At the opposite technical pole from Hütter, Juul was a traditionally trained engineer working outside of academia. Juul's design grew out of the Danish craft school tradition, and he built on the experience gained by Smidth during the war. Juul built a turbine for the Danish utility SEAS at Gedser incorporating the lessons learned. Meanwhile, both the British and the French installed prototype wind turbines during the 1950s and 1960s.

The availability of cheap oil from the Middle East doomed most of these programs, but Juul's turbine at Gedser and Hütter's turbine in the Schwabian Alps continued in operation for the next decade. Hütter's design and his argument for it would captivate German and American engineers for years to come. However, it was Juul's design and its adherents in Denmark that would come to dominate the field of wind energy some decades later.

Juul's simple robust design was adopted by the Danish wind revival of the late 1970s following the oil crises of the period. When the California "wind rush" began during the early 1980s, budding Danish wind turbine manufacturers were poised to seize the opportunity to export technology and expertise. Europeans of several nationalities flocked to California to seek their fortunes in wind energy, but none more so than the Danes. The trickle of Danish machines entering California in 1981 soon became a torrent. At the peak of the rush in 1985, Europeans shipped more than 2000 wind turbines, mostly Danish ones, to the United States. Later, after the collapse of the California market in 1985, Denmark looked elsewhere and helped Germany to become the leading wind power country in the world. Indeed, in 2001, Germans installed a record 2600 MW, and all of these wind turbines were technologically derived from Juul's work at Gedser.

3.2 Dependable Home Market

In 1981, Denmark set a national goal of installing 1000 MW of wind power by the year 2000. Denmark became self-sufficient in oil a decade later following discoveries in the North Sea, but the country continued its wind development program as a means of reducing greenhouse gas emissions. Denmark later increased wind's expected contribution to 10% of the nation's electricity supply by the year 2000, a target it exceeded. In 2001, Denmark was generating 16% of its electricity with wind energy (Fig. 2).

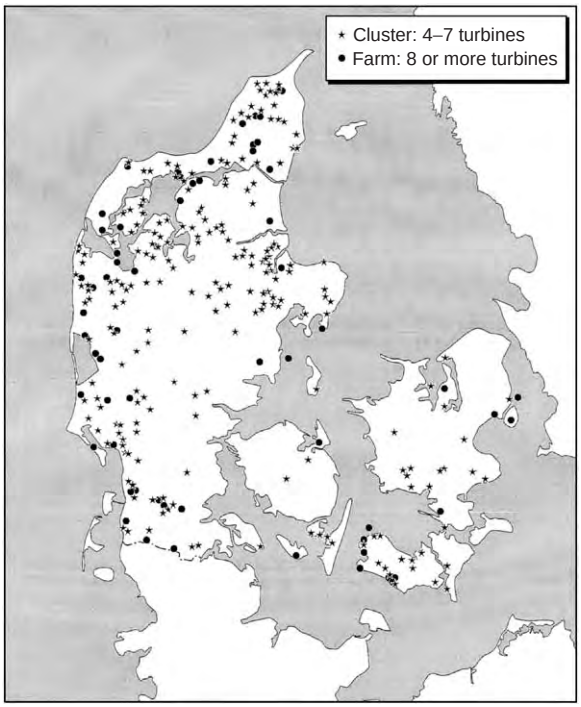


FIGURE 2 Wind-generating installations in Denmark as of February 2002.

The late 1970s brought a change that helped to further wind power development in Denmark when the European market for Danish farm equipment slackened, forcing manufacturers to seek new products for their rural customers. With generous incentives from the Danish government stimulating demand, Danish manufacturers quickly adapted their surplus capacity to the new wind turbine market. The demand for modern wind turbines by a population spread across the Danish landscape, a good wind resource off the North Sea, and a manufacturing sector accustomed to building heavy machinery for a discerning rural market all contributed to launching the world's most successful domestic wind industry. In addition, the small size of the country allowed manufacturers to service their own turbines, often directly from the factory. This spatial proximity enabled companies to learn quickly from their mistakes and to keep their turbines in operation as physical proof to potential buyers that the companies' machines were good investments.

Danes pride themselves on their unusual ability to act individually while working cooperatively. Even though most farms are owner operated, nearly all Danish farmers are members of farm cooperatives. These cooperatives process the Danish foods found on shelves worldwide. Danish wind turbine cooperatives and an association of wind turbine owners have had a profound effect on the development of wind energy in the country. Some 100,000 Danish households ($\sim$5% of the population) own shares in wind cooperatives.

Another contributor to Danish success was the Danish Windmill Owners Association (*Danske Vindkraftværker*). Formed in 1978, the group grew from a series of informal quarterly meetings of backyard experimenters, hobbyists, and environmentalists. Because many early wind turbines were unreliable, the group demanded minimum design standards. The most important element in the future success of wind energy was the requirement for a failsafe redundant braking system such as the tip brakes invented by Juul during the 1950s for use on his Gedser turbine. This single provision, more than any other, furthered Danish wind technology because it succeeded in ensuring the survival of the wind turbine when something went wrong. The owners association also compiled statistics on the reliability of Danish turbines and compelled Danish manufacturers to fulfill their product guarantees when the turbines underperformed. Collective action by Danish wind turbine owners was so successful that German wind turbine owners replicated the format south of the Danish–German border.

Wind energy grew rapidly in Denmark, not only due to early government incentives but also because Danish utilities were required to pay a premium price for wind-generated electricity. By agreement with the government, utilities paid 85% of the retail rate for electricity from privately owned wind turbines. Coupled with credits against energy taxes, Danish utilities were paying the equivalent of U.S. $0.10/kWh for wind-generated electricity during the mid 1990s in comparison with retail rates of nearly $0.17/kWh.

Many factors have contributed to the success of wind power in Denmark. One is two decades of consistent national policies supporting a strong domestic market for Danish wind machines, mostly for farmers and cooperatives. Another is the often-overlooked backing of Danish financial institutions; Danish banks and finance societies during the 1990s provided 10- to 12-year loans for 60 to 80% of the installed cost. Banks competed for wind projects by advertising themselves as "the wind banks" in local newspapers.

3.3 Nationwide Land Use Planning

Although thousands of wind turbines have been installed in Denmark by individuals without much notoriety or difficulty, Danish utilities encountered public opposition when they proposed wind projects of their own. These companies were viewed as outsiders, and the new neighbors of their proposed wind turbines considered them to be no different from conventional power plants. In response, and at the utilities' request, the government appointed a special wind turbine siting committee in 1991 to identify locations for wind turbines with a broad base of acceptability. Land use planning became necessary as wind turbines increased in size dramatically since the early 1980s and as their numbers continued to increase. The government committee estimated that the Danish landscape could absorb 1000 to 2800 MW of wind capacity, taking into account local objections and the preservation of scenic areas. Regional Danish jurisdictions now include wind energy in their land use plans. These plans designate zones where wind turbines are prohibited, where single turbines may be installed, where wind farms or wind power plants are permitted, and where clusters or single turbines may be erected with special approval.

Many municipalities, including the capital city, also include wind turbines in their local plans.

There are wind turbines in urban areas throughout Denmark and three projects within the environs of Copenhagen. One project, Lynetten, is visible from the Christianborg Palace, the seat of Denmark's parliament, the *Folketing*. The turbines are also visible from the Little Mermaid, the most visited tourist attraction in the city. Today, wind turbines in Denmark are dispersed across the landscape, in contrast to the giant wind farms typically seen in North America, India, and Spain. In fact, many Danish installations include only one turbine. In total, after 16 years of development, Denmark reached 2000 MW of installed wind capacity and paved the way for other countries to have similar success.

3.4 Germany Becomes the Leader

Wind developments in Germany differ somewhat from those in Denmark. The German market clearly demonstrates the effectiveness of their incentives to stimulate the growth of wind energy. In 1991, Germany's conservative government introduced the *Stromeinspeisungsgesetz* (electricity feed law), requiring utilities to pay 90% of their average retail rates for purchases from renewable energy sources such as wind turbines. The law, which encompassed only a few paragraphs, resulted in extensive wind development. By the early 21st century, wind turbines were in every part of Germany, from polders on the North Sea to hilltops of the central highlands (Fig. 3).

The parliament (*Bundestag*) modified the feed law in 1999, stabilizing both the price and the period during which the price would be paid. The measure ensured a stable and continually expanding domestic market. In 2001, wind turbines were producing 3.5% of the electricity in Germany, a nation of 80 million and the world's third largest economy. As a result of its precedent-setting feed law, Germany had the most dynamic and transparent markets for wind turbines in the world during the first decade of the new millennium. Like Denmark, Germany's large domestic market spawned the growth of some of the world's largest wind turbine manufacturers. Also like Denmark, many of the wind turbines in Germany were installed as single turbines or small clusters.

3.5 Other Markets

There are several other wind markets in Europe outside of Denmark and Germany. For example,

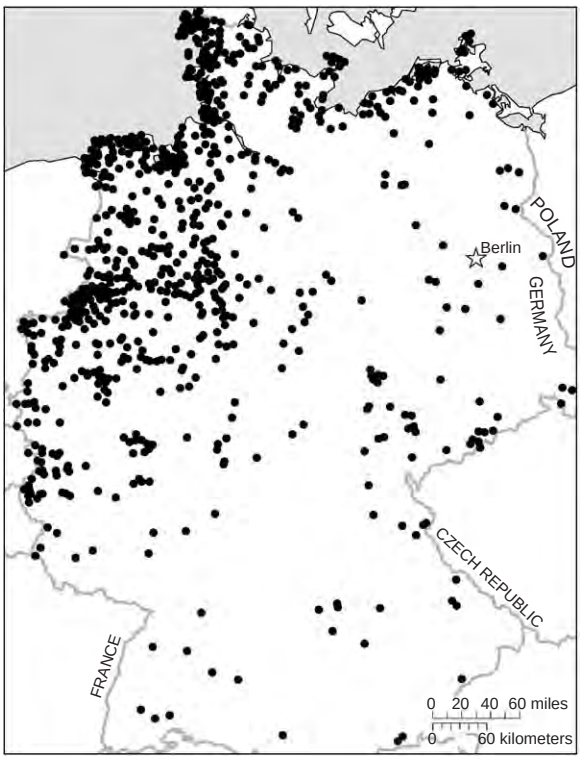


FIGURE 3 Wind-generating installations in Germany as of September 2001.

development began in earnest in Spain during the mid-1990s when the government introduced its own electricity feed law. Under the law, Spanish projects can choose a fixed price per kilowatt-hour of wind-generated electricity or the wholesale rate. Most have chosen the fixed price tariff. As a result of the combination of the country's electricity feed law, good wind resources, and supportive regional governments that see wind energy as a means of economic development, Spain has a bustling wind industry that rivals that of Germany. In 2001, Spain was the world's second largest manufacturer of wind turbines.

Great Britain and The Netherlands have decent wind energy resources, although they have had different experiences. Both countries launched wind development programs during the early 1980s but have lagged behind their European neighbors. Both misdirected research and development funds toward electric utilities and large industrial concerns in centrally directed programs that were incapable of supplying competitive products for a limited domestic market. Furthermore, programs designed to stimulate a market instead often antagonized prospective neighbors by forcing wind development onto the most sensitive sites. Although Great Britain

has the best wind resource in Europe (Fig. 4), organized opposition nearly brought the industry to collapse during the late 1990s. Despite an early beginning and an excellent wind resource, total installed wind capacity in both Great Britain and The Netherlands represented only 5% of the total installed in Germany in 2002.

France, with its longtime emphasis on nuclear power, has been one of the last European countries to turn their attention to the possibilities of wind energy. After watching how wind development was stymied in Great Britain and being familiar with the opposition to new nuclear construction, France sought to adapt the German feed law to French conditions. The French electricity feed offers three fixed tariffs: one price for windy sites, a second price for modest wind resources, and a third price for the least windy areas. It also limits the size of the wind farms that can be built under the program. These provisions encourage a greater number of smaller projects than those found in North America and spread wind development beyond just a few windy provinces such as Brittany in the west and Languedoc in the south.

Several other countries, inside Europe as well as elsewhere, are installing wind generators. Within Europe, these include Italy, Greece, and Sweden. Outside of Europe, wind development is prominent in India, particularly in the states of Gujarat and Tamil Nadu. Several large projects were also built in

FIGURE 4 Wind-generating installations in the United Kingdom as of January 2002. Courtesy of British Wind Energy Association.

North Africa and China during the late 1990s. By the turn of the 21st century, wind turbines in commercial projects were operating on every continent.

4. REJUVENATED NORTH AMERICA

4.1 The Great Plains

After the burden of bad press coming out of California, progress in wind power in the United States dwindled virtually to a standstill during the late 1990s, just as it was picking up momentum in Europe. Ironically, support and success in Europe rekindled the American market. Texas has been one of the leaders of this resurgence, followed by several other states such as Minnesota, Iowa, Oregon, Washington, and Pennsylvania (and soon perhaps the Dakotas and Massachusetts). Many reasons explain this renewed interest. First, wind technology has matured, becoming more efficient and reliable and less threatening to wildlife, especially to birds. Second, certain economic incentives now exist for utility companies to promote the resource, most specifically their ability to capture "credits" as a part of renewable portfolio standards. For example, these standards in Texas require utilities to provide a percentage of their electricity from renewable energy. Once this step was taken, Texas surged to nearly 1100 MW by 2001, second only to California's 1700 MW. Most of this new Texas capacity is in West Texas, in the vicinity of Iraan and McCamey. The quick growth in Texas is due in part to a policy that allows companies who exceed the minimum required percentage to sell their additional credits to other providers.

A third reason behind the renewed growth of wind power in the United States is that farmers have learned that wind power can make them money and help them to keep their land. For this reason, the Great Plains seem destined to assume the leadership position in U.S. wind power development and perhaps the world, a result of a combination of natural and cultural conditions. Most notably, the U.S. wind resource is concentrated there as the result of favorable climatic and topographic conditions. It was estimated that the potential in the Dakotas would match 60% of the electrical demand for the entire country in 1990 (Fig. 5 and Table I).

Another important reason for the increased activity in the Great Plains is found in the region's history of land use. More than in any other part of the country, agriculture is the economic mainstay of the region, and it has fallen on hard times, particularly at the expense of the small farm. Word has spread quickly across the Great Plains that wind turbines can substantially improve the economic return from property in the region. In fact, farmers in many areas today generate more income from electricity than from traditional farm products

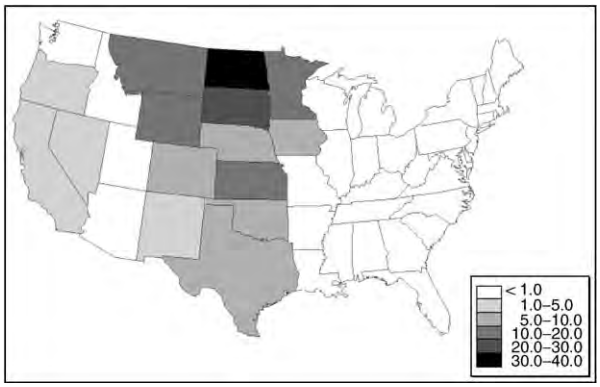


FIGURE 5 Wind energy potential as a percentage of total electric consumption. Reprinted from Pacific Northwest National Laboratory. (1991). "An Assessment of the Available Wind Land Area and Wind Energy Potential in the Contiguous United States." PNNL, Richland, WA.

TABLE I

Wind Potential in Selected States

Rank	State	Billions of kilowatt-hours
1	North Dakota	1210
2	Texas	1190
3	Kansas	1070
4	South Dakota	1030
5	Montana	1020
6	Nebraska	868
7	Wyoming	747
8	Oklahoma	725
9	Minnesota	657
10	Iowa	551
11	Colorado	481
12	New Mexico	435
13	Idaho	73
14	Michigan	65
15	New York	62
17	California	59[a]

[a] Illustrates that numerous states have greater wind potential than does California, where the majority (~90%) of U.S. wind development has occurred to date.

Source. Data from Pacific Northwest National Laboratory. (1991). "An Assessment of the Available Windy Land Area and Wind Energy Potential in the Contiguous United States." PNNL, Richland, WA.

while at the same time using only a small portion of their land for the physical wind equipment. Ironically, the very area of the country that several decades ago was home to the greatest concentration of windmills in the world will possibly become so again, this time having them pumping out electrons instead of water.

In addition to Texas, development on the Great Plains has been most noticeable in Iowa and Minnesota, where 324 and 319 MW of generating capacity, respectively, were installed at the time of this writing. Unlike Texas, where most of the wind installations are on marginal grazing land or in spent oilfields, the wind turbines in Iowa and Minnesota are on working farms.

Also on the Great Plains, albeit farther west, wind developments have been appearing in Wyoming ranching country, where at the end of 2001 there was about 140 MW of installed generating capacity. In some locations in that state, such as near Interstate 80 west of Laramie, the wind is ferocious, and for the first time it is being seen not as a burden of life in the area but rather as an asset. In the Great Plains, NIMBY (not in my backyard) is being supplanted by a new acronym, PIMBY (please in my backyard).

4.2 The Pacific Northwest

The other location of substantially increased wind power installations has been the Pacific Northwest, an area that recently has witnessed the second fastest growth after Texas. Several sites in the windy Columbia River Gorge east of Portland are being monitored, and the Stateline Project, along the eastern border of Washington and Oregon, amounts to approximately 180 MW of capacity. Here too it is making money for local people; according to an article in the March 25, 2002, issue of the *Register Guard* of Eugene, Oregon, Northwestern Wind Power is paying 60 farmers in four counties alone to build wind turbines on at least 150,000 acres. Another developer, SeaWest Wind Power Inc., has leased tens of thousands of acres from at least 20 landowners in Oregon and Washington. In Sherman County, some farmers can expect a $10,000 sign-up payment, $15,000 for each turbine installed, and up to $5000 per turbine per year. That does not include the money that developers pay to reserve the land—approximately $2000 per month—while they conduct wind studies. Here, as in the Great Plains, wind development is becoming a profitable business.

4.3 Nantucket Sound

Objections to wind power, infrequent in the Northwest and Great Plains, still exist in some areas where population is concentrated, especially in those areas that are considered attractive tourist destinations. During recent years, such public opposition, which began in California, has been concentrated on a site off the coast of Massachusetts, where 170 turbines have been proposed to spread across the sea on a 28-acre patch on Horseshoe Shoal in Nantucket Sound, 9 miles off Martha's Vineyard and as close as 4 miles to Cape Cod. Objections, particularly from an organization called the Alliance to Protect Nantucket Sound, have focused on the visual intrusion of these turbines and to the potential disruption of traditionally productive fishing grounds. At least 10 other wind power projects are being considered elsewhere in New England, but because the $700 million Cape Wind project came first, the discussion about offshore wind power is concentrating here. Regardless of how the controversy off Cape Cod is resolved, it has underscored a key point in the future potential for this renewable energy source. In some places wind power is welcomed warmly, whereas elsewhere it is opposed with vehemence.

5. TRENDS

Wind power has become the best hope for the future of alternative energy. Of all the possible "new" sources of electricity, it has rather unexpectedly found itself the most touted and fastest growing alternative energy resource in the world, generating significant amounts of electricity in several countries, such as the United States, Denmark, Germany, and Spain, as well as impressive amounts in many other countries (Fig. 6).

Development of wind power has been following several trends. One has been the effort toward greater economies of scale that results through the use of multimegawatt turbines. Most of the early turbines of this size were abandoned by the mid-1980s after failing in three areas: hours of operation, energy generated, and contribution toward making wind technology more competitive with conventional fuels. Another round of large turbines during the late 1980s and early 1990s performed somewhat better than their forebears, renewing hope among government-sponsored laboratories that large wind turbines did have a future after all, but they too

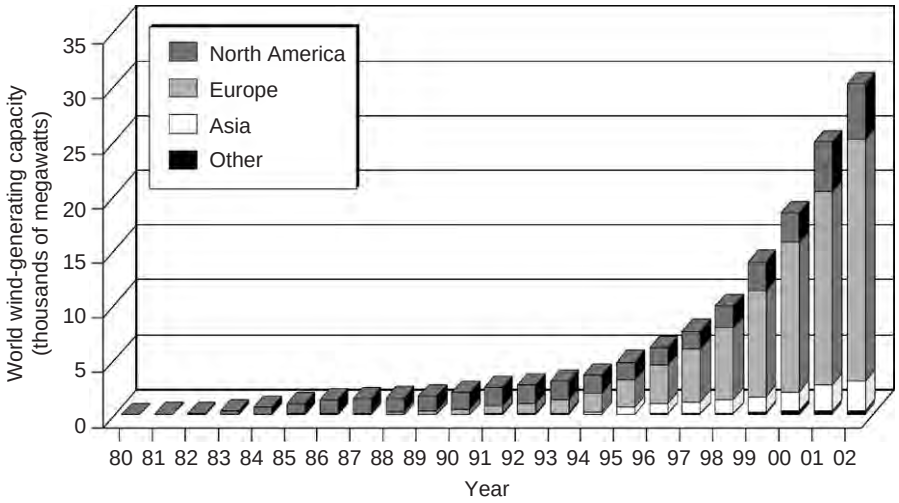


FIGURE 6 World generating capacity, illustrating the steep and recent rise in installed wind power.

were not very successful. In 1990, the European Commission intended to chart the way for a third round of funding for large wind turbine development. In contrast to the past, funds were directed toward wind turbine manufacturers rather than the aerospace industry or other large industries. European governments, like their North American counterparts, had promoted megawatt-sized wind turbines for more than a decade with little success. The European Community worried that the progression toward larger wind turbines produced by Danish and German manufacturers would take too long to reach the size that program managers in Brussels, Belgium, believed were the optimum for Europe.

Manufacturers participating in a more aggressive program erected their prototype megawatt-class turbines during the mid-1990s. Commercial production began during the late 1990s, and by early 2000 increasing numbers of the turbines were being installed in Germany. In 2000, the average wind turbine installed in Germany was exceeding 1 MW for the first time, a milestone that took much longer to achieve in the United States because development rested not with wind manufacturers but rather with the National Aeronautics and Space Administration (NASA). However, circumstances eventually improved. By 2002, megawatt-scale wind turbines, some with more than 100 turbines of 1.5 MW each, were becoming common in Europe and on large wind farms in North America.

Another trend has been to install new turbines not on land but rather offshore. This strategy has been practiced with special vigor in Europe in the hope

FIGURE 7 Visualization of a suggested offshore wind farm near Samsø, Denmark. The original idea was to place 10 turbines in a circular array. The project was expected to be built in 2003 in a linear array. Courtesy of Frode Birk Nielsen.

that offshore locations would reduce siting conflicts. Because offshore turbines require more expensive infrastructure, projects have longer lead times and must be larger than those on land. For this reason, such projects have been slow to develop, but the Danes and Swedes completed small near-shore projects during the late 1990s. One large project of 20 2-MW turbines, Middlegrunden, was installed outside of Copenhagen's harbor in 2000. In 2002, Denmark began installing 300 MW of offshore capacity, including the 160-MW Horns Rev project in the North Sea and the 158-MW Rødsand project

in the Baltic Sea. To try to visualize the impact of the turbines in advance, planners have taken to using computer visualization, for example, at Samsø (Fig. 7).

6. CONCLUSION

In most ways, wind power has come of age. The technology has improved, the economics has become more appealing, and substantial progress has been achieved in reducing environmental impacts. Planting wind turbines as a kind of semipermanent new "crop" on the Great Plains is becoming more common, making wind power more of a welcomed intrusion than a stimulus for citizen opposition.

Where public perceptions continue to slow growth in wind power, attitudes are largely influenced by benefits versus costs, and this has been most apparent off Cape Cod. Although one might argue that the sight of wind turbines helps to remind the public of the costs of their energy requirements, the proposed installation of fields of wind turbines in view of a heavily frequented tourist destinations is perhaps more than can reasonably be supported. In this case and probably others, the wind industry would be well advised to concentrate their developments where they are wanted and leave other environmentally sensitive areas alone. If one can agree that wind power has come of age, it will make a valuable contribution to providing us with the electricity we need in a world with increasing sensitivity to its environmental costs.

SEE ALSO THE FOLLOWING ARTICLES

Coal Industry, History of • *Early Industrial World, Energy Flow in* • *Electricity Use, History of* • *Hydropower, History and Technology of* • *Manufactured Gas, History of* • *Natural Gas, History of* • *Nuclear Power, History of* • *Oil Industry, History of* • *Solar Energy, History of* • *Wind Energy Economics* • *Wind Energy Technology, Environmental Impacts of* • *Wind Farms* • *Wind Resource Base* • *Wood Energy, History of*

Further Reading

Ackermann, T., and Söder, L. (2002). An overview of wind energy—Status 2002. *Renewable Sustainable Energy Rev.* **6,** 67–128.

Asmus, P. (2000). "Reaping the Wind: How Mechanical Wizards, Visionaries, and Profiteers Helped Shape Our Energy Future." Island Press, Washington, DC.

Baker, T. L. (1985). "A Field Guide to American Windmills." University of Oklahoma Press, Norman.

Gipe, P. (1995). "Wind Energy Comes of Age." John Wiley, New York.

Golding, E. W. (1955). "The Generation of Electricity by Wind Power." E. & F. N. Spon, London.

Pasqualetti, M. J., Gipe, P., and Righter, R. W. (eds.). (2002). "Wind Power in View: Energy Landscapes in a Crowded World." Academic Press, San Diego.

Putnam, P. C. (1948). "Power from the Wind." Van Nostrand Reinhold, New York.

Righter, R. W. (1996). "Wind Energy in America: A History." University of Oklahoma Press, Norman.

van Est, R. (1999). "Winds of Change: A Comparative Study of the Politics of Wind Energy Innovation in California and Denmark." International Books, Utrecht, Netherlands.

Wind Energy Technology, Environmental Impacts of

MICHAEL L. MORRISON
Great Basin Institute, University of Nevada
Reno, Nevada, United States

KARIN SINCLAIR
U.S. Department of Energy, National Renewable Energy Laboratory
Golden, Colorado, United States

1. **Potential Environmental Impacts of Wind Technology**
2. **Bird Biology and Wind Turbines**
3. **Research and Monitoring**
4. **Risk Reduction**
5. **Summary and Recommendations for Reducing Wind Turbine Impacts on Birds**

Glossary

end-of-row turbine A turbine that is located on either end of a turbine string; turbine strings in the Altamont Pass wind resource area usually consist of 3 to 10 turbines.

guy wire Wire or cable used to secure and stabilize wind turbines, meteorological towers, and other vertical objects in wind resource areas; most wind turbine towers are self-supporting and do not require guy wires.

horizontal axis Axis orientation where the axis of rotation of the power shaft relative to the ground is parallel to the ground; this is by far the most common axis orientation in existing wind resource areas.

horizontal lattice turbine A wind turbine tower that is composed of a steel lattice structure; many steel lattice support structures run parallel to the ground, providing considerable perching space to birds.

nacelle The housing that protects the generator and gear box in horizontal axis turbines.

rotational speed The rate (in revolutions per minute) at which a turbine blade makes a complete revolution around its axis; wind turbine rotational speeds can be fixed or variable.

rotor center The distance from the ground to the center of the circular area that is swept by the rotating blades.

rotor diameter The diameter of the circular area that is swept by the rotating tip of a wind turbine blade; rotor diameter is equal to two times the blade length.

rotor orientation (or yaw) The rotation in the horizontal plane of a horizontal axis turbine by which the rotor is oriented into the wind.

rotor-swept area The circular area that is swept by the rotating blades; doubling the length of blades quadruples the blade-swept area.

tip speed The speed (in miles per hour) that the tip of the blade travels when the rotor is turning; at comparable wind speeds, a horizontal axis turbine with a longer rotor experiences slower revolutions per minute and the tip speed remains roughly constant.

turbine height The maximum height reached by the blade tips of a wind turbine.

turbine spacing The distance between wind turbines in a string; in general, the distance between turbines is proportional to the rotor diameter.

vertical axis Axis orientation where the axis of rotation of the power shaft relative to the ground is perpendicular to the ground.

Wind energy offers a relatively environmentally benign source of electricity. However, issues have been raised concerning the feasibility of placing commercial wind developments in certain geographic locations and landscape settings. Because these issues are site specific, it is critical that potential sites be evaluated independently to determine whether there are environmental issues. Conclusions about any environmental concerns for a particular site should not be based explicitly on the experience from other sites.

 © 2004 Elsevier Inc. All rights reserved.

1. POTENTIAL ENVIRONMENTAL IMPACTS OF WIND TECHNOLOGY

1.1 Introduction

Wind power has been used commercially to produce electricity since the early 1980s, when the world's first large-scale wind development, or wind resource area (WRA), was developed in California. In the United States, more than half of the states have developed wind resource areas: Alaska, California, Colorado, Hawaii, Iowa, Kansas, Maine, Maryland, Massachusetts, Michigan, Minnesota, Montana, Nebraska, New Hampshire, New Mexico, New York, North Dakota, Oregon, Pennsylvania, South Dakota, Tennessee, Texas, Utah, Vermont, Washington, West Virginia, Wisconsin, and Wyoming. Additional states are in various stages of the planning process. A current U.S. map identifying states with installed wind power can be found on the American Wind Energy Association's (AWEA) Web site. The AWEA is the U.S. wind industry trade organization that promotes the development of wind energy. Wind developments are also prevalent throughout much of Europe, with most developments being located relatively close to coastlines. In addition, Europe has pioneered the installation of offshore wind developments, an application that is only beginning to be explored in North America.

1.2 Potential Environmental Issues

In specific situations, wind power developments have been shown to cause environmental impacts, including impacts on animal habitat and movements, noise pollution, visual impacts, biological concerns, bird/bat fatalities from collisions with rotating blades, and health concerns. Of all the potential environmental impacts, biological concerns regarding birds and bats have been discussed and studied the most. Thus, although this article covers all potential environmental impacts, it focuses on avian-related issues. In addition, much of what is known is specific to land-based wind developments.

Various other issues associated with resource developments in general may also be of concern with wind developments; however, issues that are common to most if not all developments, such as soil erosion and water quality, are not addressed in detail in this article.

1.2.1 Impacts on Habitat and Animal Movements

Relative to other forms of energy development, most commercial wind farms occupy a large amount of space. Although the turbine pads occupy little ground area, the developments often must be spread over many miles. WRAs in the western United States have been placed primarily in open areas where they coexist with farming, ranching, and other low-intensity land uses. Figure 1 is an example of a wind farm in an agricultural setting. Figure 2 illustrates the minimal amount of land taken out of production by wind turbines, allowing ranching activities to continue on the land. Wind developments in Europe are frequently located near coastlines to take advantage of ocean winds.

FIGURE 1 Photo of 1.5-MW NEG Micon turbines installed by the Municipal Energy Association of Nebraska (MEAN) at TVIG's Kimball wind farm in Nebraska. The turbines have a rotor diameter of 72 m. From Tennessee Valley Infrastructure Group Inc. (NREL PIX12084).

FIGURE 2 Photo of Zond turbine near Lake Benton, Minnesota. Wind turbines occupy only approximately 5% of the land in a wind power plant on flat terrain. The rest of the land can be used for ranching and farming, as it was before the turbines were installed. From Warren Gretz (NREL PIX06334).

Wind farms may facilitate retention of seminatural open space where these low-intensity uses occur, resulting in long-term benefits for some bird populations. Historically, habitat that has been affected consists primarily of raptor-foraging habitat and nesting habitat for ground-nesting birds. Disruptions of animal movements, such as ungulates and carnivores, could occur but can be avoided through careful site evaluation.

Commercial wind developments in the eastern and northeastern United States often occupy ridges in wooded regions. In such situations, roads and pads must be cut, creating openings in the woodlands. Such openings create edge habitats that could potentially allow invasion of brown-headed cowbirds, a species that reduces nest success of many songbirds through nest parasitism, and potential predators of ground- and shrub-nesting birds. Figure 3 illustrates initial impacts of a wind project in a forested ridge setting. Long-term impacts have not been assessed. In Europe, developments along coastlines are often on cliffs and not in forested areas. Nevertheless, the potential exists for disturbance to coastal environments through development of roads, transmission lines, and other infrastructure.

1.2.1.1 Consideration of Impact on Animal Habitat and Movements The term "habitat" refers to the specific configuration of environmental features (e.g., vegetation, rock outcrops, water) that an animal uses at any point in time. Habitat is a species-specific concept; every animal species uses a different combination of environmental features. Therefore, no specific area is a "good" or "bad" habitat unless it is assessed in relation to a specific species. Thus, what is "good" for one species might be "bad" for another species.

This definition and discussion of habitat has relevance for all land use issues. For wind developments, issues of habitat concern (1) outright loss of a species' habitat due to development (e.g., removing trees), (2) indirect impacts due to disturbance (i.e., the animals will no longer reside near the development, and (3) disruption in animal passage through or over the development due to the addition of towers and turbines. This definition also applies to offshore situations and is applicable to foraging areas for birds, marine mammals, and fish.

Because wind developments often stretch for many miles along ridge tops, the turbines, along with the associated infrastructure (especially roads), can affect animal movements. No quantitative work has been done on the impacts of wind developments on animal movements. However, it is unlikely that wind developments will cause wholesale disruption in migratory movements of birds and other vertebrates. This is because wind developments are seldom, if ever, introduced into pristine environments. As such, migrating or dispersing animals encounter a host of potential obstructions, including highways, power line corridors, housing, farm fields and pastures, and people. Therefore, analysis of the impacts of a wind development on movements of

FIGURE 3 Photos of 6.0-MW Green Mountain power plant at Searsburg, Vermont. The power plant consists of 11 550-kW Zond Z-40FS wind turbines. Construction crews removed as few trees as possible to build this power plant. Since the completion of construction, the area around the turbines is being allowed to "brush in" and return to its natural state. Left photo from Bob Thresher (NREL PIX00638), right photo from David Parsons (NREL PIX06802).

terrestrial animals would usually focus on the additional impacts that the wind development would have on animal movements. There are also numerous methods available, developed for highways and other human developments, that could apply to wind developments and be used to lessen potential impacts on animal movements. Such strategies would include widely spacing turbines in certain locations, restricting travel along roads within wind developments (e.g., for inspection and repair only) during specific times of year (e.g., during ungulate migration), and providing appropriate habitat between turbines to facilitate animal movements.

Offshore developments have been established throughout Europe, but few studies have been conducted to determine the impact on animals. Most of these developments are small relative to onshore developments (although larger projects are being planned). Some disruption in bird flight patterns has been noted in Europe, although additional study is needed. However, there does not appear to be disruption in fish movements or populations; the effects on marine mammals need clarification. Offshore developments in North America are only in the planning stage.

Wind developments are evaluated for potential impacts on animals in the project area in the same way that most development would be evaluated. Such evaluations include close inspection of the project areas for the presence of plant or animal species of concern (inspections for cultural resources would also be conducted). For example, burrowing animals such as ground squirrels can be detected, and specific steps can be used to avoid disturbance such as moving the location of roads, pads for placing turbines, and other infrastructure items. There is a fair degree of flexibility available to the wind developer with regard to the exact location that a turbine is placed on the ground. Because most resource agencies avoid the need to use mitigation to allow permitting of a project, careful planning is an essential component of placing turbines.

The ground disturbance associated with placing a turbine, especially road cuts and cuts used for turbine pad placement, can actually attract certain species. Most notable is the use of disturbed soils by burrowing animals such as gophers and ground squirrels. In some locations, these species are rare or legally protected. In most areas, the attraction of these animals to turbines likely enhances the area for foraging raptors, thereby raising the likelihood of raptor collisions with rotating blades. As such, wind developments must consider the potential for attracting species of concern and creating conditions in which birds could be killed. Although these often conflicting issues cannot be resolved easily, they should be considered when planning a development so that major resource issues can be dealt with during the design stage.

A major concern with offshore developments is the impacts that boat and air (helicopter) traffic to and from the wind developments might have on animal behavior and movements. Again, little is known about such impacts. The primary concern is that human activity in support of the developments (e.g., delivering supplies and maintenance crews) could have negative ramifications for animal behavior that extends far outside the boundaries of the turbines.

1.2.2 Noise Pollution

Noise is defined as any unwanted sound. Concerns depend on the level of intensity, frequency, frequency distribution, and pattern of the noise source; background noise levels; the terrain between emitter and receptor; and the nature of the noise receptors. The effects of noise on people can be classified into three general categories: (1) subjective effects of annoyance, nuisance, and dissatisfaction; (2) interference with activities such as speech, sleep, and learning; and (3) physiological effects such as anxiety, tinnitus, and hearing loss.

The sound levels associated with environmental noise usually produce effects only in the first two categories. Operating noise produced by wind facilities is considerably different in level and nature than that generated by most power plants, which are typically large industrial facilities. Much of the noise emitted by wind turbines is masked by the ambient or background noise of the wind and does not carry far. Noise may, of course, concern individuals living in close proximity to a wind turbine. As such, siting of wind developments must consider adjacent land uses and zoning. Over time, residential or commercial developments might encroach on existing wind developments, much like housing often encroaches on existing airports. Conflicts often arise in these situations, especially if the wind developments are slated for expansion. In such cases, adequate setbacks from human-occupied structures need to be considered.

Several European countries, including Sweden, Denmark, The Netherlands, the United Kingdom, and Greece, have adopted noise standards for wind farm developments. These standards specify acceptable noise levels, and some take into account the existing background noise levels.

1.2.3 Visual Pollution

Determining whether a wind farm creates visual pollution is purely subjective and, therefore, is difficult to quantify. Computer modeling tools can be used to simulate how a wind farm will look from multiple vantage points when placed on a landscape. The aesthetic interpretation and visual compatibility with the surrounding setting will vary depending on the natural and cultural features that exist and how various stakeholder groups value these. Alternatives to the wind development (i.e., other land uses) may also be considered when evaluating the visual impacts.

Often referred to as the "viewshed," the impact of wind turbines on the existing project setting is a matter of local public perception. Because wind resource developments vary substantially in number of turbines (literally from one to thousands of turbines), height of the turbines and towers, spacing between turbines, and physical arrangement of towers and turbines on the landscape, there is no single solution to the impacts of wind developments on a viewshed. However, longer blades are being used on commercial-sized wind turbines, and the turbines are being placed on taller towers. Such turbines now commonly produce 1.5 MW to more than 2 MW of power, are three bladed, have a rotor-swept diameter of 70 m, and are placed on a tower 80 m tall. Therefore, modern commercial wind turbines have a maximum height measurement from ground level to the tip of the blade of 120 m ($\sim$375′) or taller. Many offshore developments have proposed turbine–tower combinations that near or exceed 160 m ($\sim$500′) in total height. In addition, onshore turbines are usually placed on ridges to capture the maximum amount of wind in some geographic locations such as the western and eastern United States. As a result, turbines can be highly visible from many miles away. Larger turbines result in fewer turbines in an observer's field of view, although the land area covered for a given wind farm megawatt capacity is approximately the same as for smaller turbines. In some locations, aircraft warning lights may be required by the Federal Aviation Administration (FAA), adding another dimension to visual considerations. Figure 4 illustrates visual impacts of various wind farm configurations. The photo on the left is an example of early wind farm layouts, using numerous smaller turbines mounted on lattice towers. The photo on the right depicts a more recent wind farm, using larger wind turbines mounted on tubular towers.

Visual aspects of wind development are a project- and site-specific issue that must carefully consider current and likely land uses. Close collaboration among the developer, government permitting agencies, and viewing public is necessary to develop collaborative strategies for siting of a development and placement of individual turbines. Although many proposed offshore developments are planned for more than 3 miles from land, they can still be seen from shore and will be visible to boaters.

1.2.4 Bird Fatalities

Researchers have conducted a large number of avian fatality surveys at WRAs in the United States and Europe. In most cases, several factors have made it difficult to compare results from one study or site with those from another. First, survey methods vary, and very few of these studies have been peer reviewed and published in scientific journals. Second, turbine designs and wind farm layouts vary considerably from site to site. Third, the wide climatic and topographical differences and range of bird species present in various locations make it extremely difficult to draw concrete conclusions from the available literature. Fourth, it is difficult to make statistical comparisons because the relatively low number of fatalities results in inadequate sample size. In addition, there has been an emphasis on raptor fatalities despite the knowledge that other birds have been affected, perhaps because raptors receive protection under a large suite of federal and state laws and because they are symbolic and have greater emotional value. A sampling bias toward raptors and other large birds also occurs because small birds are more difficult to detect and scavenging of small birds can be expected to occur more rapidly. Impacts on some passerines, including neotropical migrants, may warrant more careful scrutiny because many of them are protected under the Federal Migratory Bird Treaty Act and because some are experiencing regional population declines.

In the United States, a study of golden eagle populations in the vicinity of the Altamont Pass WRA in the state of California from 1994 to 2000 reported that 52 of 257 eagles equipped with radio transmitters were killed by wind turbine strikes. When compared with other large WRAs, it is clear that the Altamont Pass supports substantially higher resident and migratory raptor populations and experiences substantially greater raptor fatality rates caused by collisions with wind turbines.

The exact number of fatalities caused by turbines in Altamont Pass cannot be determined, although estimates based on intensive fatality surveys conclude that 40 to 60 are killed annually. Although it remains

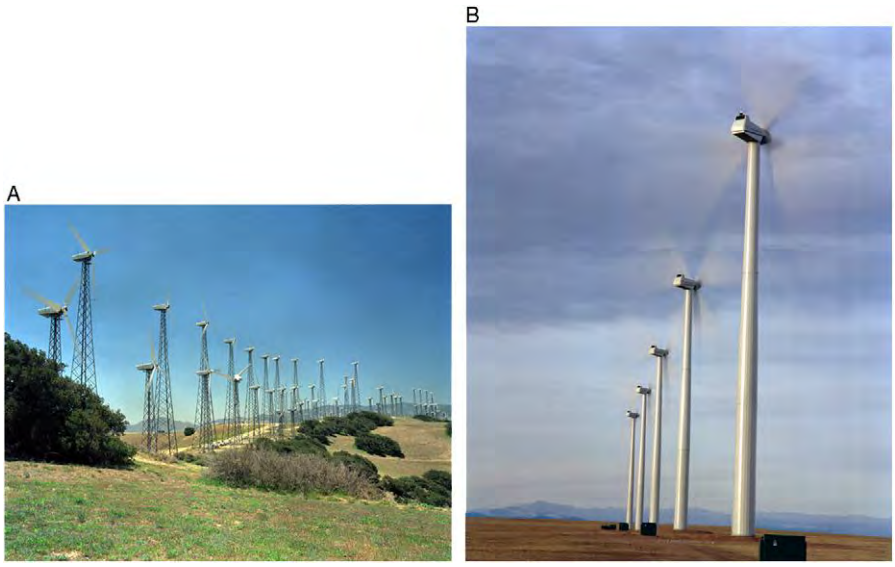


FIGURE 4 Photos illustrating visual differences between turbines installed on ridges and those installed on mesas, between small and large turbines, and between turbines mounted on lattice towers (A) and those mounted on tubular towers (B). A number of factors influence the visual impacts of wind farms, including location, size, and type of turbine. From Warren Gretz (NREL PIX01231 and NREL PIX07236).

speculative whether the eagle population in the Altamont region is declining due to turbine kills, these eagle fatalities may cause many direct and indirect impacts on its population. Direct impacts include a decline in the size of the population and change in age structure. Indirect impacts include a reduction in the number of birds that are available to disperse to other regions. Overall, the golden eagle population is thought to be declining in the Altamont Pass region.

Results of numerous studies in the United States at sites outside of California have indicated that the incidence of raptor collisions in California, particularly at the Altamont Pass WRA, is considerably higher than in other WRAs. In the United States, passerines composed the highest percentage of fatalities in the non-California studies. Reports from these studies indicate that the levels of fatalities are not considered significant enough to threaten local or regional population levels, although overall cumulative effects of human-induced bird fatalities are unknown. A major difference between the Altamont Pass WRA and other WRAs is that many other areas lack the dense populations of raptors and the diverse topography of the Altamont Pass WRA. The high number of fatalities at the Altamont Pass WRA has created awareness of potential siting problems, and in some cases more regard has been given to the level of avian use prior to construction. Figures 5A and 5B are examples of raptor fatalities found during research conducted by the U.S. Department of Energy's National Renewable Energy Laboratory (NREL) in the Altamont WRA.

Studies on avian interactions have been conducted at more than 100 sites in Europe, but most of these studies focused on solitary or small groups of turbines. As a result, it is difficult to compare these data with those of U.S. studies that focus on larger groups of turbines. Many of the European studies were based on coastal sites where waterfowl and migratory bird issues did not yield results comparable to those of California studies. Although the fatality rates in Europe are often higher than those that occur in California and the rest of the United States, many European scientists have not considered this impact to be a significant threat to bird populations in most of Europe because the number of deaths is small relative to the total number of birds using or passing through the area. Most birds killed were not raptors but rather relatively common species of passerines and water birds.

The bird–wind data on offshore wind farms in Europe are scarce. However, it is clear that concern exists about the potential impacts that offshore facilities could have on birds, as witnessed by the precautionary tone taken by many individuals. Recommendations include constructing relatively small farms, avoiding areas of high numbers of birds or areas with sensitive bird species, and monitoring intensively before and after construction (if the projects are approved).

FIGURE 5 (A) Red-tailed hawk fatality resulting from collision with wind turbine at the Altamont Pass WRA in California. (B) Golden eagle fatality resulting from collision with wind turbine at the Altamont Pass WRA. From BioResource Consultants.

Very few studies have been performed to determine the overall effects of wind turbine impacts on the population viability of an individual species. This type of evaluation is exceedingly difficult and relatively costly. Studies that are based on field-level tracking of population parameters require intensive sampling over a number of years for each species of concern. Most evaluations of wind-turbine-related avian fatalities have concluded that the affected birds are fairly common and not threatened at the population level.

However, because some raptor species are relatively less abundant and produce fewer young compared with many other groups of birds, human-caused fatalities could have a more noticeable effect on their populations. In addition, the number of fatalities relative to the population size, rather than the absolute number killed, is of concern. Thus, even a few fatalities of a locally or regionally rare species could have negative impacts on the population structure and size.

Various design features, such as perch availability, rotor diameter, rotor-swept area, rotor height, rotational and tip speeds, and fixed versus variable turbine speed, have been evaluated to determine whether they contribute to bird collision risk. Although at one time workers suggested that lattice towers might encourage birds to perch and thereby come into close proximity of rotating blades, recent studies have failed to confirm a correlation between tower type and fatality rates.

The issue of rotor-swept area and blade diameter effects on avian collision risk with turbines is important because wind developers are in the process of repowering or replacing existing, less efficient turbines with a smaller number of new, larger, and more efficient turbines. Existing studies concerning fatalities in relationship to increased rotor-swept area are inconclusive, and further research is needed.

It has been speculated that overall bird fatalities will be higher at taller turbines. It is well known that migrating songbirds are frequently killed at tall communication towers. This was attributed to higher flying, non-raptor nocturnal migrants that often constitute the greatest percentage of fatalities. Behavior differences among raptors indicate that there may be some trade-offs in benefits. For example, one species may benefit from blade tips being farther from ground level, whereas others may be exposed to greater risk from blades that reach a greater overall height. Further research is needed to determine the impact that the newer taller turbine towers have on bird kills. Rotor velocity and a corresponding increased tip speed have been

correlated with fatalities. Faster turbine rotor speeds kill more raptors than would be expected by chance.

1.2.5 Bat Fatalities

Until recently, little research has been conducted to determine the impacts of wind farms on bats. A few bat carcasses were found during earlier bird fatality studies; however, recent surveys at some newer sites have discovered higher mortalities than were reported previously. To better discern the level of collision impacts on bats, researchers have initiated studies at some of these sites. These studies include analysis of environmental features in and around wind farms that could attract bats such as roosting sites, availability of surface water, and riparian foraging areas.

1.2.6 Health Impacts

Shadow flicker, the effect caused by rotating blades cutting through sunlight, is another concern that may need to be addressed if wind turbines are installed in close proximity to inhabited areas. Modern commercial wind turbines are tall structures that will cast a shadow at certain times of the day. Concerns about shadow flicker can be minimized or eliminated with careful wind farm layout and sufficient setbacks from residences.

2. BIRD BIOLOGY AND WIND TURBINES

2.1 Bird Flight Behavior

The relatively high diversity of bird species killed in wind developments suggests a variety of behaviors, and therefore underlying risk factors, associated with wind turbines. Studies of birds approaching turbines show that most birds pass over or through turbine structures, thereby avoiding collisions. However, some birds may be more prone to collisions with turbines for a variety of reasons. For raptors, flight and hunting behavior require that they often fly at low altitudes and so may strike turbines while searching or stooping for prey. Abundance of prey may significantly increase foraging activity, and therefore collision risk, of some raptors. Age or breeding status may also be a risk factor associated with bird fatalities.

Topographical features, such as ridges that decrease the altitude of birds relative to ground level, can result in a higher collision risk. Because flight lines often follow prevailing wind directions, the placement of turbines in an optimal position and orientation for wind energy capture may result in obstacles directly in the path of local and migratory birds. For a sample of approximately 1000 turbines studied in the Altamont Pass WRA, 25% of the turbines were responsible for 100% of the fatalities recorded. Certain physical and topographical factors appear to contribute to an increased collision risk.

A summary of bird–wind interactions in Europe found that the overall kill rate was low. However, it was noted that 2 to 3% of birds passing a wind farm at rotor height were killed. It has been concluded that disturbance and habitat loss effects associated with wind developments were probably much more important than direct bird kills due to collisions. However, in Spain, relatively high levels of kills were sometimes evident. Research in Europe has indicated that individual and single rows of turbines in areas with small bird populations provide the best landscape for wind farms.

2.2 Bird Sight

The idea of birds colliding with seemingly slow-moving turbines seems paradoxical given the superb vision possessed by most birds, especially raptors. However, optical analyses indicate that as a bird approaches the rotating blades, the retinal image of the blades (which is the information that gets transferred to the bird's brain) increases in velocity until it is moving so fast that the retina cannot keep up with it. At this point, the retinal image becomes a transparent blur that the bird probably interprets as being safe to fly through, with potentially catastrophic consequences. This phenomenon is called "motion smear" or "motion blur" and is well known in human perception. It becomes especially difficult for a bird to detect the outer regions, or tip, of the blade. Although modern commercial turbines rotate at relatively slow revolutions per minute, because of their long length (often 100 feet), they move at a high rate of speed (up to 150 mph). Figure 6 illustrates the concept of motion smear.

Laboratory experiments have been conducted to determine the potential of various patterns painted on blades to increase visibility to birds. A single uniformly black blade paired with two white blades appears to be the most effective anti-motion smear pattern under the widest variety of conditions in the field. Minor increases in visibility of the central portions of the blade have been found, although painting did not increase visibility of the tips. These laboratory studies examine only the physiological

FIGURE 6 Illustration of the phenomenon of motion smear, shown by photo of the moving blades of the turbine in the foreground. Beyond a certain point, the velocity of the retinal images of the blades sweeping across the retina will overwhelm the retina's ability to keep up. The initial effect will be a smearing or blurring of the image of the blades, at which point the blades appear to be transparent. This may result in deadly consequences for birds because they interpret this as a safe place to fly. From William Hodos (University of Maryland).

response of the eye to various painted patterns; they do not examine the psychological response of the bird to the stimuli. Thus, rigorous field experimentation is needed to determine the usefulness of painted blades in substantially reducing bird collisions.

2.3 Bird Hearing

It is a myth that birds have exceptionally acute hearing. When hearing is defined as the softest sound that can be heard at various frequencies, birds on average hear less well than do many mammals, including humans. The literature suggests that birds cannot hear the noise produced by wind turbines as well as can humans. In practical terms, a human with normal hearing can probably hear a wind turbine blade twice as far away as can the average bird. Adding an acoustic cue to a rotating blade might assist with a bird hearing a blade. However, acoustic deterrent devices are not generally effective because birds habituate to them and eventually ignore them completely. Devices that purport to use sound frequencies outside the hearing range of humans are most certainly inaudible to birds as well because birds have a narrower range of hearing than do humans. However, it is entirely possible that as birds approach a wind turbine, especially under high wind conditions, they lose the ability to see the blade (due to motion smear) before they are close enough to hear the blade.

3. RESEARCH AND MONITORING

To minimize the potential risks from wind turbines, it is important to first assess what those risks are. Prior to the development of the first large commercial wind farms, protocols and methodologies for assessing risks to birds within WRAs had not been developed. Assessments of other potential environmental impacts, such as impacts on habitat and animal movements and visual and noise impacts, were only minimally considered.

During the early years of wind farm developments, specifically in California, it became apparent that collisions of birds (especially raptors) with wind turbines were occurring. Approaches to assessing potential environmental impacts needed to be developed.

3.1 Initial Site Selection and Permitting

3.1.1 U.S. Guidelines

The National Wind Coordinating Committee (NWCC), a collaborative of representatives from the environmental community, wind energy industry, state legislatures, state utility commissions, consumer advocacy offices, green power marketers, and federal and state governments in the United States, was established in 1994 to support the development of wind power. The NWCC's Siting Subcommittee was formed to address wind generation siting and permitting issues. This subcommittee prepared a collaborative document, titled *Permitting of Wind Energy Facilities: A Handbook,* to provide guidance in evaluating wind projects for all stakeholders. Some aspects of wind facility permitting closely resemble permitting considerations for any other large energy facility or other development project. Others are unique to wind generation facilities. Unlike most energy facilities, wind generation facilities tend to be located in rural or remote areas and are land intrusive rather than land intensive. Thus, they may extend over a very large area and have a broad area of influence, but they physically occupy only a small area for the turbine towers and associated structures and infrastructure (e.g., roads, transmission lines). The rest of the land may be used for other activities.

This NWCC handbook outlines the many factors that must be considered when permitting a wind

generation facility, including (but not limited to) land use, noise, birds and other biological resources, visual resources, soil erosion and water quality, cultural resources, and socioeconomic considerations. Permitting processes that result in timely decisions, focus on the critical issues early, involve the public, and avoid unnecessary court challenges will enable wind generation to compete with other energy technologies and provide a diverse supply of energy.

As summarized by the NWCC Web site, wind power developers in North America are beginning to look offshore for greater wind resources and lower costs. Because this is a new application for wind power in the United States, the NWCC noted that background information is needed about environmental, technical, economic, and political issues associated with offshore developments. Some of this information can be gathered from European sources because they have been working on offshore wind energy development for the past decade or so. However, each situation is different, so local factors need to be taken into account.

3.1.2 European and Other Guidelines

Several best practice guidelines (BPGs) have been adopted. The British Wind Energy Association (BWEA) adopted BPGs for wind energy development in 1994. The BWEA, established in 1979, functions as a professional association and an association for the wind energy industry. The guidelines provide guidance to developers, taking them through the development process. They are general in nature and do not supersede any national policies.

The European Wind Energy Association (EWEA) was established in 1982 and provides functions similar to those of the BWEA. Building on the British and other European experiences with wind industry development, the EWEA adopted its BPGs in 1999.

The Australian Wind Energy Association (AusWEA), established in 2000, adopted its own BPGs for wind energy projects. Building on the BWEA and EWEA guidelines, the AusWEA's guidelines were adapted in 2002 to include experience from the first wind farms developed in Australia.

3.2 Sampling Protocols

Sampling protocols and methods of quantifying bird fatalities in the field must be rigorous and scientifically valid. Failure to establish rigorous protocols and methods results in data sets that are subject to criticism and not ideal for determining how to reduce wildlife fatalities in wind developments.

The NWCC's Avian Subcommittee, formed in 1994 to address avian interactions with wind developments, identifies research needs and serves as an advisory group. In March 2003, the committee was renamed the Wildlife Work Group to reflect the concern for a broader group of species. To address differences in methodologies, lack of adequate control or baseline data in existing studies, and the resulting lack of interstudy comparability, the committee developed a guidebook titled *Studying Wind Energy/Bird Interactions: A Guidance Document*. The document provides a comprehensive guide to standardized methods and metrics to determine impacts to birds at existing and future wind farm sites. A stated purpose of the guide is to promote efficient cost-effective study designs that will produce comparable data and reduce the overall need for some future studies.

The guidance document identifies three levels of surveys of increasing intensity that should be applied to a proposed or developing wind development. Although focused on birds, these general guidelines can be applied to most aspects of the environment. "Site evaluation" or "reconnaissance" surveys are relatively nonrigorous and use primarily published literature and nonpublished reports, expert opinions, and other sources of information to make a first determination as to whether a proposed site will likely result in environmental problems. Such reconnaissance surveys are cursory in nature but should help to eliminate problematic sites from further consideration. If the reconnaissance survey indicates that the site should be suitable for development, a "level 1" protocol is indicated, where more intensive and quantitative on-site surveys occur, usually for a minimum of 1 year prior to a decision to proceed with development. Such level 1 surveys include on-site sampling of wildlife movements and other activities (e.g., foraging, nesting), quantification of the presence and abundance of sensitive species, and development of projections on potential bird and bat fatality levels based on risk assessment. If a decision is made to proceed with development after level 1 surveys have been evaluated, the data set also serves as the "before" data to be compared with changes in environmental conditions following ("after") development. In rare cases, the results of level 1 studies indicate that more intensive study is necessary, for example, when the presence of a legally protected, threatened, or endangered species is located.

The methodology used to conduct fatality searches will vary, depending on the terrain and rigor required to adequately assess a site. The level of rigor often is dictated by the budget available for the task. Ideally, the appropriate budget will be allocated to conduct the level of work required for the specific site under consideration. Search methods have been developed and have proved to be very successful under research studies funded by the U.S. Department of Energy's National Renewable Energy Laboratory. These search methods have been developed for both flat and steep terrain. Figures 7A and 7B illustrate the search method used for conducting fatality searches on steep and flat terrain, respectively.

To aid bird risk evaluations, there is a need to define the "level of take" that is acceptable from avian species' interactions with wind turbines; that is, the level of fatalities that can occur without reducing that species' population. The level of assigned risk will vary among species. If threatened or endangered species are affected, incidental take permits can be issued under the federal Endangered Species Act, but "no take" may be the targeted goal. There is no accommodation for take under the Migratory Bird Treaty Act, which applies to a large number of birds, including raptors and passerines, that have been killed by collisions with wind turbines. To date, the U.S. Fish and Wildlife Service has been somewhat tolerant of existing conditions at WRAs. However, future wind plant approvals, particularly in California where existing avian fatality issues have not been resolved, may be modified due to enforcement of this act.

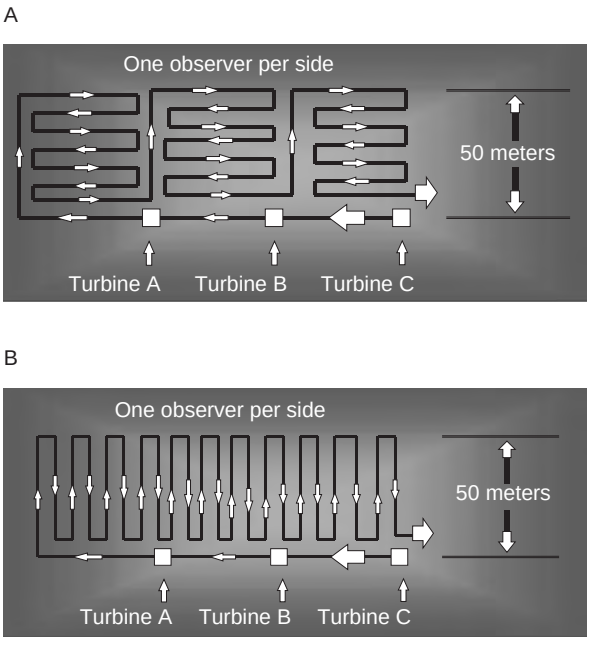


FIGURE 7 (A) Searcher method for conducting fatality searches on steep terrain. (B) Searcher method for conducting fatality searches on flat terrain. From BioResource Consultants.

The United Kingdom limits offshore wind farms to 30 turbines (even if 2-MW turbines are used, this amounts to only 60 MW) to reduce environmental and visual impacts and to ease the cost of decommissioning and dismantling when problems arise. One of the benefits of limiting the size of farms (e.g., to 150 MW) is that if problems are identified at a farm, the farm can be shut down without a great disruption to the grid and loss of value because the turbines can be used at another site. Thus, various governments in Europe have chosen to limit the size of individual wind farms due to a variety of environmental and other siting concerns.

4. RISK REDUCTION

4.1 Wind Farm Operation and Location

Recognition of the potential avian fatality risk has influenced the selection criteria used in developing some new WRAs. Along with providing a framework for the development of more robust experimental field design, use of standardized protocols has greatly enhanced researchers' ability to compare and analyze data among studies from various WRAs.

Decreasing operation time of problem turbines or WRAs has been suggested as a risk reduction measure. Critical shutdown times could be seasonal (e.g., during migration periods) or based on inclement weather or nighttime periods when visibility is reduced. For example, permitting requirements for a new wind development in Maryland included the option for restricting turbine operations if bird fatalities were found to occur during peak migration periods. Economic consequences of any kind of adjustments to operation time should be considered prior to construction of a site so that the developer can make a feasibility assessment. Most of the newer wind turbines in the Altamont Pass repowering areas will increase operation time by as much as 55%. The effect that such an increase in operating time will have on bird fatalities is unknown and worthy of study. As noted previously, some European governments limit the size of wind developments to ease the financial ramifications of closing a development.

Design and maintenance characteristics of roads and structures may indirectly contribute to higher

bird fatality rates by increasing prey densities. It has been suggested that vertical and lateral edges and other features associated with roads, turbine pads, and maintenance areas may create suitable burrow sites for raptor prey in the vicinity of wind turbines, with a corresponding increase in raptor foraging activity. Newer facilities that are built with larger, more efficient turbines require fewer roads and have a greater amount of space between them. In addition, many of the newer facilities have underground transmission lines, greatly reducing the likelihood of wire collisions and electrocutions. Offshore developments in Europe manage the movement of maintenance crews to minimize disturbance to birds and other animals.

4.2 Turbine Characteristics and Location

Lights seem to play a key role in attracting birds, and the lighting of tall structures appears to contribute to avian fatalities. Illuminating other taller aerial structures to make them more visible to aircraft has resulted in increased bird fatalities. Migratory species generally migrate at night and appear to be most susceptible to collisions with lit towers on foggy, misty, rainy nights with low cloud ceilings. Passerines migrating at night during poor visibility conditions appear to be particularly susceptible. Solid or blinking red lights seem to attract birds on foggy misty nights more than do white strobes, which may flash every 1 to 3 seconds. Preliminary research suggests that the longer the duration of the "off" phase, the less likely a light will attract birds. The advent of turbines with longer blades that are mounted on tall towers could require the use of warning lights for aircraft in certain locations.

Studies indicate that tower placement is a site-specific phenomenon, but several key conclusions have been found. First, irregularly spaced turbines might increase fatalities because birds try to negotiate the apparent gaps between turbines. Second, turbines placed close to the edge of ridges show higher fatality rates because raptors often hover in such locations. Finally, turbines placed near a gully have higher fatalities because birds often use these locations as flight paths.

Although no research has been conducted on auditory deterrents to birds approaching wind turbines, audible devices to scare or warn birds have been used at airports, television towers, utility poles, and oil spills. Most studies of auditory warning devices have found that birds become habituated to these devices. As discussed previously, birds do not hear as well as humans, and minor modifications to the acoustic signature of a turbine blade could make blades more audible to birds while at the same time making no measurable contribution to overall noise level. Under certain conditions (e.g., high wind), birds might lose their ability to see a turbine blade before they are close enough to hear the blade. Currently, there is no research under way that tests the effects of auditory deterrents, and because of the low likelihood of developing a successful application, none is planned for the foreseeable future.

As discussed previously, motion smear, which makes the blade tips of wind turbines appear to be transparent at high speeds, could be reduced under laboratory conditions. Results suggest that a single solid black blade paired with two white blades (inverse blade pattern) could be effective at reducing visual smearing of blades. In addition, motion smear and a very narrow blade profile encountered when approaching from the side could be very risky for birds. One potential solution is a rectangular attachment to the outer tip at right angles to the long axis of the blade. The visibility and practicality of these attachments have not yet been evaluated.

4.3 Habitat Management

The density of raptors at the Altamont Pass WRA is, at least in part, a result of high prey availability. The San Gorgonio and Tehachapi WRAs in California have lower prey densities, lower raptor densities, and lower per turbine fatality rates. Prey densities appear to be highest at disturbed sites such as roads and turbine pads, the latter of which would exacerbate collision risk. Reducing prey populations within the vicinity of wind turbines might reduce high-risk foraging activities for raptors. Suggested methods include county-sponsored abatement programs, reduced grazing intensities, and revegetation with higher stature plants that pocket gophers and ground squirrels tend to avoid. However, these measures could affect other populations, including special status species such as the San Joaquin kit fox, burrowing owl, and badger.

It has been suggested that in areas where high densities of gopher burrowing systems are found, controlling populations in the immediate vicinity of wind turbines could be effective. However, small mammal abatement efforts by ranchers in surrounding areas could exacerbate the effect of burrow clusters in the vicinity of turbines by increasing the focus of raptor foraging. This effect occurs when there is rapid recolonization and higher than average

densities of gophers on edges of abatement areas. Research would have to evaluate reduced grazing or revegetation with higher structure plants to determine its effects on fire management, watershed protection, and other land management practices.

Habitat modification to reduce prey densities has been discussed as a possible avian risk reduction technique. There have been some efforts to reduce ground squirrel populations in portions of the Altamont Pass WRA, but no results of the effects on reducing raptor moralities have been published. The effects of a widespread control program would have to take into account the effects on other wildlife such as protected species that prey on ground squirrels or depend on their burrows for nesting and cover habitat. Widespread use of rodenticides or other measures to remove prey may prove to be controversial and costly. Feasibility of more benign habitat modification measures, such as manipulation of annual grassland grazing practices and conversion to perennial grassland, may be worth studying. However, any reduction in habitat quality would likely result in an ultimate, although indirect, negative impact on birds. For example, although reducing prey in wind developments might lower bird fatalities, it could also result in bird starvation or fatalities in other locations.

In The Netherlands, turbines adjacent to the Wadden Sea showed a higher rate of bird kills relative to turbines in upland areas. These higher kills appeared to be related to the large number of birds in and around the sea relative to the upland locations. Unfortunately, virtually no information is available on fatalities at offshore wind farms in Europe.

4.4 Turbine Treatments

Effective visual treatments could provide a cost-effective method to reduce risk from turbines determined to cause fatalities. Laboratory and field tests of treatments that make turbine blades more conspicuous to raptors and other birds are needed.

4.5 Repowering

Repowering refers to the replacement of existing, less efficient turbines with a smaller number of larger, more efficient turbines. Replacement ratios will range from 7:1 to 10:1. The impacts of repowering on avian fatality rates, particularly raptors, have not been evaluated. However, at Altamont Pass, no significant differences in the number of fatalities have been seen between relatively small, faster rotating turbines (100 kW) and relatively large, slower rotating turbines (up to 750 kW). The tip of the blade on the newest generation of turbine can reach more than 400 feet into the air at the apex of rotation. Studies have begun on the collisions of birds and bats with these larger turbines.

5. SUMMARY AND RECOMMENDATIONS FOR REDUCING WIND TURBINE IMPACTS ON BIRDS

Newer wind facilities throughout the world have fewer raptor fatalities than do those documented at the Altamont Pass WRA. Much of this improvement has not been by design; rather, these wind energy facilities were sited in areas that typically have lower raptor densities and less topographic diversity than does the Altamont Pass WRA. Some avoidance has been due to preconstruction site evaluations and a standardized approach to evaluating wind power projects developed by the NWCC. In addition, reduced concentrations of turbines in newer WRAs, and possibly newer turbine designs, may have contributed to avoiding the high number of raptor fatalities that occur at the Altamont Pass WRA. However, researchers still need more information on how to reduce fatalities at existing facilities. In addition, it is not known whether the newer generation turbine designs slated to replace many of the old turbines will reduce avian fatalities. Finally, little is known concerning the impacts of large offshore developments (in Europe and North America) on animals.

Consequently, research should still focus on other risk-reducing measures that involve changes in turbine design and microsite selection. As reviewed in this article, specific protocols exist to identify potentially negative environmental impacts of wind turbines on wildlife prior to project development. In addition, turbines can be placed in locations that are less likely to cause substantial bird or bat fatalities, operating times can be modified to avoid periods of peak migration (if birds are found to be killed during these times), and various habitat management techniques might be implemented to reduce the attractiveness of turbines to birds. Environmental analyses of proposed wind farms also must consider the visual impacts that the newer generation of taller towers and longer-bladed turbines has on the surrounding community or viewshed. These types of analyses will

help to ensure that the benefits outweigh the risks in establishing a wind farm in a specific locality.

SEE ALSO THE FOLLOWING ARTICLES

Arid Environments, Impacts of Energy Development in • Electricity, Environmental Impacts of • Hydropower, Environmental Impact of • Life Cycle Analysis of Power Generation Systems • Photovoltaics, Environmental Impact of • Wind Energy Economics • Wind Energy, History of • Wind Farms • Wind Resource Base

Further Reading

American Wind Energy Association. (n.d.). [U.S. map identifying states with installed wind power]. www.awea.org/projects/index.html.

Anderson, R., Morrison, M., Sinclair, K., and Strickland, D. (1999). "Studying Wind Energy/Bird Interactions: A Guidance Document." National Wind Coordinating Committee/RESOLVE, Washington, DC. www.nationalwind.org/pubs/avian99/avian_booklet.pdf.

Australian Wind Energy Association. (2002). [AusWEA best practice guidelines for wind energy projects]. www.auswea.com.au/downloads/ausweaguidelines.pdf.

British Wind Energy Association. (1994). [BWEA best practice guidelines for wind energy development]. www.bwea.com/pdf/bpg.pdf.

Erickson, W. P., Johnson, G. D., Strickland, M. D., Young, D. P., Jr., Sernka, K. J., and Good, R. E. (2001). "Avian Collisions with Wind Turbines: A Summary of Existing Studies and Comparisons to Other Sources of Avian Collision Mortality in the United States." National Wind Coordinating Committee, Washington, DC. www.nationalwind.org/pubs/default.htm.

European Wind Energy Association. (1999). [EWEA best practice guidelines for wind energy development]. www.ewea.org/doc/bpg.pdf.

National Wind Coordinating Committee. (n.d.). www.nationalwind.org.

National Wind Coordinating Committee. (2002). "Permitting of Wind Energy Facilities: A Handbook." NWCC/RESOLVE, Washington, DC. www.nationalwind.org/pubs/permit/permitting2002.pdf.

National Renewable Energy Laboratory/National Wind Technology Center. (n.d.). www.nrel.gov/wind.

National Wind Technology Center/Avian. (n.d.). www.nrel.gov/wind/avian_lit.html.

Winkelman, J. E. (1995). Bird/Wind turbine investigations in Europe. *In* "Proceedings of National Avian–Wind Power Planning Meeting," NREL/SP-441-7814, pp. 43–47. National Renewable Energy Laboratory, Golden, CO.

Wind Farms

ERIK LUNDTANG PETERSEN and PETER HAUGE MADSEN
Risoe National Laboratory
Roskilde, Denmark

1. **Introduction**
2. **Wind Energy Status**
3. **Wind Farm Siting Issues**
4. **Technical Wind Farm Issues**
5. **Power System Integration**
6. **Wind Farm Examples**

Glossary

capacity credit The fraction of a power plant's rated capacity, which is in average available at the time of peak demand or, alternatively, the relative amount of additional conventional capacity that is needed in the absence of wind energy in order to obtain the same loss-of-load-probability (LOLP).

environmental impact assessment An evaluation of a number of conditions with a view to select possible areas for location of wind farms. These conditions include scenic aspects, bird sanctuaries, raw material extraction, archaeology, fisheries, shipping routes, and military areas.

mesoscale model The numerical atmospheric model that is similar to numerical weather prediction models used for the daily weather forecasts. The mesoscale model calculates the flow field over a region of typically 200×200 km with a resolution of one to five km. The output is used as input to the microscale model.

microscale model The numerical atmospheric model that can calculate the flow field over a farm site with a high accuracy. Is used for micrositing.

micrositing The determination of exact position of each wind turbine in a wind farm.

siting Investigation of an area/region/district with the purpose of selection of the location (site) for a wind farm.

sizing The decision on the size of a wind farm in terms of land area and installed capacity.

wake The wake behind a wind turbine is characterized by a wind velocity deficit and increased turbulence. The wake affects power output and loads on downstream wind turbines.

wind energy The mean power density in the wind, which is equal to the average kinetic energy flux per unit area perpendicular to the flow, is calculated as half of the air density multiplied by the mean value of the third power of the wind speed.

wind farm efficiency The ratio between the annual production of a wind farm and the production of corresponding number of isolated wind turbines located in identical but undisturbed wind conditions.

wind farm layout The configuration and arrangement of the individual wind turbines' locations within the selected site.

A wind farm (often also called a wind park) can be defined as a cluster of wind turbines that acts and is connected to the power system as a single electricity producing power station. Generally it is expected that a wind farm consist of more than three wind turbines.

1. INTRODUCTION

Many independent studies have led to the conclusion that to solve the increasing energy need of the world's growing population, it is inherently necessary to increase the amount of energy produced by renewable energy sources. In fact, it is believed that in 2050, more than 80% of energy should be renewable, 20% of which should be wind energy. A large fraction will come from wind turbines placed in clusters (i.e., wind farms) at land or at sea. A wind farm (often also called a wind park) can be defined as a cluster of wind turbines that acts as one electricity-producing power station. Generally, it is expected that a wind farm consists of more than three wind turbines. However, the rapid development in the size of wind turbines can make such a

 © 2004 Elsevier Inc. All rights reserved.

definition superfluous: Tomorrow's single wind turbine might exceed 10 MW, whereas several of yesterday's wind farms are below 5 MW.

The development has led to a division of the technology into three categories, depending on application and capacities:

- Land-based and offshore grid-connected wind farms varying in size from about 10 MW to several 100 MW; WTs larger than 1.5 MW.
- Decentralized and single operating machines connected to the grid; WTs less than 1.5 MW.
- Decentralized units for grid-connected operation and for hybrid and standalone operation; WTs less than 0.5 MW.

Here we are only concerned with the first category, which is formed by large-size grid-connected wind turbines for land-based and offshore applications. From the early 1980s, development has steadily moved toward larger machines and larger wind farms; for example, in 1995 only 20% of all wind farms were larger than 5 MW, whereas in 2002 nearly 70% exceeded 5 MW (a 5 MW wind farm may, on the average, supply 2500 to 3000 north European households with electricity—if they do not use electricity for house-heating).

Because wind turbines are clustered into bigger units, system integration requires special attention. It is estimated that a penetration (supply fraction) of wind energy on a large grid can be as much as 20% without special precautions to secure grid stability. Turbines equipped with state-of-the-art power electronics may even add stability to the grid. For an even larger penetration, special attention has to be focused on control systems of wind farms, improvements of power quality, system and grid stability, operational management systems, system reliability, and output forecasting. Large wind farms are very visible and this together with a number of environmental concerns, such as possible disturbance of migrating birds, play an important role in the planning process and can result in selection of sites with less than optimal conditions.

Good wind conditions are essential for the economics for any wind project. With the advent of the mega–wind farms the requirements to good wind resource assessment methods have led to methodologies where all available knowledge on meteorology, climatology, and topography is used on the ever-increasing amount of information on the atmosphere and the earth's surface that comes from satellites and ground measurements. Large computers combine models and data to estimate the wind resource at the wind farm sites and to predict the output from the farms days ahead. The wind resources vary dramatically over a few kilometers as well as from one region to another, depending on topographical conditions and the overall weather systems. As an example on the variation, if a wind turbine produced full power all year round, one would say it had 8765 full last hours; in southern Germany it would produce 1000 hours; in northern Germany, 2000 hours; in the North Sea, 4000 hours; in Dzungar Gate between Kazakhstan and China, 4000 hours; in the Gulf of Suez, 6000 hours (all approximate and dependent on the exact location and the size of the wind turbine).

Modern wind turbines have a high quality, which results in technical availability of typically 98 to 99%. The high level of availability is a result of the reliability of the machines and, for most land-based machines, the easy accessibility for repairs and maintenance. However, under offshore conditions, the situation is different. Because of the poor accessibility of the site, a higher degree of reliability is one of the necessary conditions for arriving at an acceptable level of availability. Precautions concerning operation and maintenance has led the operators of the 200 MW wind farm 20 km off the coast of Jutland in the North Sea—the Horns Rev Wind Farm—to equip each turbine with a helicopter platform and a survival room for the service personnel (see Fig. 1).

As noted earlier, in a not-too-distant future a considerable amount of the necessary energy will and must come from renewable energy technologies. With this perspective, we will see many very large wind

FIGURE 1 Access by helicopter to one of the wind turbines at Horns Rev offshore wind farm in the North Sea, 20 km from the Danish coast. It was constructed in 2002 and consists of 80 turbines, each of 2 MW. The area of the farm is 20 km^2. From ELSAM.

farms, distributed all over the world (and not as in 2002, where 85% of all the world's wind turbine capacity was placed in five European countries). Many of the wind farms can be expected to be located in environmentally sensitive areas; hence, environmental awareness and public acceptance will be key issues for the large wind farm projects to come.

2. WIND ENERGY STATUS

The revival of wind energy began in the end of the 1970s as a consequence of the concerns for energy supply security after the oil crises. This, as well as environmental concerns about global warming from the use of fossil fuels, has caused many nations to promote the development and use of wind energy for electricity generation. In general, two approaches have been used: (1) development support for the wind technology in terms of research, development, and demonstration and (2) market development using instruments like fixed power purchase contracts, investment or production subsidies, tax credits, wind resource studies and physical planning, setting aside areas for wind energy development, preferential finance, and so on. Especially when they are combined, these instruments have been successful in developing modern wind energy, both as a technology and as a significant and growing energy source.

Whereas the initial development in general consisted of single wind turbines, dispersed in rural areas, the first major development of wind energy capacity and wind farms took place in California in the early 1980s. The development was supported by subsidies for building wind farms through tax credits and by premium payments for the produced electricity; within a few years the installed capacity reached approximately 1500 MW. Figure 2 shows an example of a Californian wind farm. The wind turbines were rather small, less than 100 kW, and the operating record was rather poor due to immature technology and difficulties with the electrical integration into the power system.

After some years with a limited installation of new capacity, growth accelerated in the mid-1990s, this time with Europe as the world leader in technology and installations. Figure 3 shows the global development of wind power capacity since 1983. During 2002, new wind energy installations reached 7227 MW, bringing the total global installed capacity up to 32,000 MW supplying 0.4% of the world's

FIGURE 2 A wind farm in Palm Springs, California. From the Risø National Laboratory.

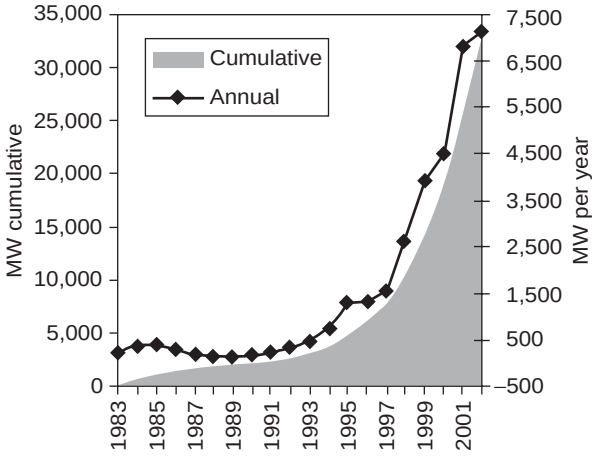


FIGURE 3 Annual and cumulative installed wind power in the world. From World Market Update 2002, BTM Consult Aps, March 2003.

electricity demand and with more than 61,000 wind turbines installed worldwide; 85% of the total capacity is installed in Europe, led by Germany, Spain, and Denmark.

An interesting trend in the market is that a growing part, currently two-thirds, of the new installations is made up of large wind farm projects. As wind energy becomes more economically attractive, larger developers such as subsidiaries of power utilities are becoming involved, utilizing their expertise in logistics and finance. The trend toward larger projects is seen even in Denmark and Germany, otherwise known for dispersed and small-sized projects. The development of market segments is shown in Table I.

The emerging development of offshore wind farms will add to the trend of larger projects, where the economics of grid connection, infrastructure, installation, and O&M in particular favors large projects.

TABLE I

Segment Distribution of Installed Capacity 1995–2002

Market segment	Percentage of world market							
	1995	1996	1997	1998	1999	2000	2001	2002
Dispersed development, single wind turbines and small wind farms < 5 MW	75	70	60	45	42	46	35	32
Commercial and utility owned wind farms > 5 MW and up to 199 MW	24	28	40	54	57	53	64	65.5
Other niche markets (e.g. off shore)	1	1	1	1	1	1	1	2.5

Source. World Market Update 2002, BTM Consult Aps, March 2003.

Offshore wind farms are still only a time fraction of the globally installed wind capacity. By the end of 2002, a total of 279 MW wind farm capacity was installed offshore, 233 MW of which was in Denmark. It is expected that during 2003–2007, the wind farm capacity offshore will grow to approximately 5000 MW, primarily with British and German projects.

3. WIND FARM SITING ISSUES

3.1 Wind Farm Sizing and Optimal Layout

Following a decision to extend the electricity production capacity by one or more wind farms, one has to decide where to place the wind farms (siting), the size of the wind farms (sizing), and the optimum layout of the wind farms.

The size of a wind farm is often determined with respect to a number of constraints, such as planning legislation; local and national development plans and policies; land availability, access, and transport infrastructure; power system (present and future); wind turbine size; financing; the electricity market; and environmental impacts.

The optimal size of a wind farm, economically and financially, for society and investors at given conditions may vary for different sites; hence, sizing and siting are integrated activities. Furthermore, sizing involves aspects that may not easily be quantified in monetary terms.

The site for a wind farm most often ends up being based on a comparison of selected candidates with respect to issues such as the possibility of obtaining planning authorization and approvals; the successful outcome of local hearings; potential wind energy production; environmental costs and benefits; sustainability, assumptions, uncertainties, and risks; the availability of land and infrastructure; power system development; investments and investors; design safety, reliability, and lifetime; wind farm and power system operation and maintenance; and economic and financial viability.

The wind farm layout may be determined according to different principles. No single layout concept is universally acknowledged as ideal and preferred, so no universally true automatic method for determining optimal wind farm layouts is therefore available. There are tools that can assist the analyses and could or should be applied as a part of the development and optimization exercises, but there is no way around human judgment—at least when it comes to aesthetics and visual impact.

As an example, the Middelgrund offshore wind farm outside the harbor of Copenhagen, was first planned as a straight line, but due to aesthetics grounds it was changed to an arch as shown in Fig. 4. Other examples are the early and famous wind farms in California, where Fig. 2 shows a wind farm sited in a very regular pattern, and Fig. 5 (Alta Mesa) shows a wind farm located along mountain ridges. In the United Kingdom, some of the wind farms are placed in an irregular pattern, required by legislation due to constraints on visual impact.

3.2 Wind Farm Production Estimation

Because the terrain height variations—the orography—over a wind farm site causes large variations of the available wind resource for each wind turbine, a correct and detailed terrain description is a necessary prerequisite for optimizing wind farm layouts. When the wind moves over hill and ridges, it speeds up, and the effect can easily extend to the turbine height and, because the power output of a wind turbine for aerodynamic reasons is proportional to the third power of the wind speed, taking care of this effect is essential when siting any wind turbine at a land site. This is done is by using a detailed topography map of the site as input to a so-called microscale numerical

FIGURE 4 The Middelgrund offshore wind farm outside Copenhagen Harbour. Constructed in 2000 and consists of 20 2MW wind turbines. Bonus.

FIGURE 5 The Alta Mesa Wind Farm, California. From the Risø National Laboratory.

model, which in short is able to calculate a detailed wind flow field over the wind farm site, which again is used to produce a detailed map of the variation of the wind resource over the site. Besides the orography, the model also uses as input a detailed description of the surface characteristics (i.e., the roughness of the terrain). Open farmland has a low roughness and forests have a high roughness. When the wind blows from a low roughness to a high

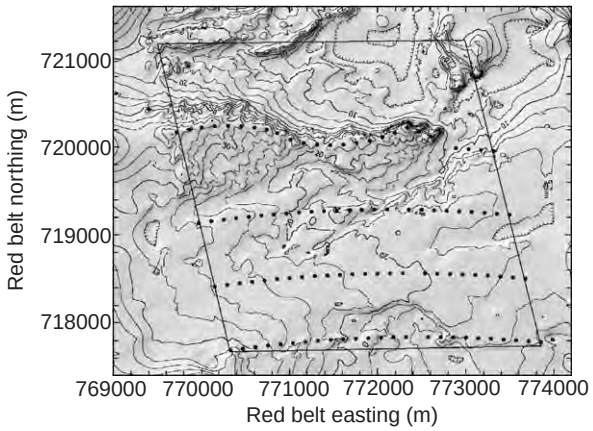


FIGURE 6 An example of an elevation map of a wind farm site (Zafarana, Egypt). An accurate description of the overall geometry of the terrain surface is a prerequisite for reliable modeling of the wind flow over the terrain. The map was established by a site-specific survey employing GPS techniques. The height contour interval in the map is 2 m, but it is possible to draw 1-m contours as well. Four sample wind farms, each consisting of 25,600-kW wind turbines, are indicated by black dots.

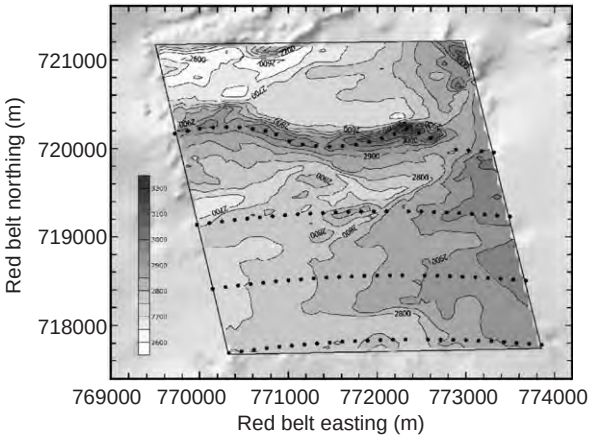


FIGURE 7 An example of a resource map of a wind farm site (Zafarana, Egypt). Estimated annual energy production for a 600-kW wind turbine. Hub height is 45 m a.g.l and the power production is calculated in a regular grid with a grid spacing of 50 m. Height contour interval is 2 m. Even within this fairly small and homogeneous site, the estimated production varies by more than 20%. The accuracy of the map has been verified by measurements at nine meteorological masts at the site. From Clausen, N.-E., *et al.* (2003). "Wind Farm Planning at the Gulf of Suez." New and Renewable Energy Authority, Cairo, Egypt and Risø National Laboratory, Roskilde, Denmark.

roughness, the wind speed decreases. Figure 6 (Zafarana topography) and Fig. 7 (Zafarana, wind resource) show the topography and wind resource for a coming wind farm site at Zafarana in Egypt. Figure 8 shows the wind resource dependence on topography in a three-dimensional way.

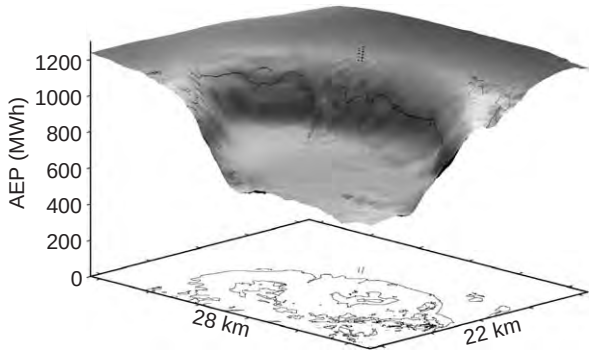


FIGURE 8 Variation of the wind energy potential at coast. The dramatic increase in the potential going from land to sea is due to the change in the surface roughness. The modeling was done in connection with the planning of the Vindeby offshore wind farm, which is shown on the figure. Vindeby was the first offshore wind farm in Denmark and has been operating since 1991. It consists of 11 Bonus turbines of 450 kW each. AEP, estimated annual energy production per wind turbine.

The estimate made by a microscale model of the average yearly power output that is produced by a wind farm over its life span—usually considered to be 20 years—is very dependent on the physical realism of the model and on the quality of the input to the model, basically topography and the wind measurements.

The most obvious way to obtain an estimate of the wind and the wind resource at a wind farm site is to perform meteorological measurements—especially wind speed and wind direction—at relevant heights and locations. But because the wind is a local characteristic, influenced by topography such as surrounding hills and changes of surface characteristics (water, forests, open farmland, desert, etc.), one needs measurements in a dense network of high meteorological masts. Such measurements are costly to carry out, and it takes a long time to obtain climatological estimates (i.e., estimates of wind energy potential over the lifetime of the wind farm). Therefore, numerical meteorological computer models are needed for the interpolation between measurements, for extrapolation to the relevant heights (the hub height of the wind turbines, which is the center of the rotor), and for the prediction of the wind climate over the lifetime.

An efficient and commonly used method of predicting the surface wind climate is to combine a mesoscale model and a microscale model. The mesoscale model simulates the regional wind climate like flow over and around mountain ranges and in large valleys. A mesoscale model is a specially adapted smaller version of the large numerical weather prediction model, which is used by the national weather services to forecast the daily weather and the weather days ahead. Mesoscale models require very large computers.

The microscale model calculates the speed-up and sheltering by hills, local obstacles, and local roughness conditions. The procedure is illustrated in Fig. 9. The mesoscale model calculates the mesoscale wind field using as input a description of the large-scale climatology obtained from the large international databases on meteorological, climatological, and topographical global information. The area will typically be 200×200 km, and the resolution 1 to 5 km.The climatology of the simulated wind fields—commonly called wind atlas data—and the local orography and roughness are subsequently used by the microscale model to predict the local wind climate at specific locations with resolutions of a few meters.

3.3 Environmental Impact Assessment

Wind turbines produce energy without pollution, eventually leading to a reduction in the emission of carbon dioxide, nitrogen oxide and sculpture dioxide. The use of wind energy may therefore contribute to reduce global climate change, acid rain and other serious environmental problems.

Although the environmental impact of wind energy obviously is lower than that of conventional energy sources, there are some potentially negative effects on the environment, especially when it comes to establishing large wind parks of several hundred large wind turbines. Over the years the main environmental concerns when constructing wind farms have been visual impact, noise, and the risk of bird collisions.

Modern wind turbines are extremely durable and the danger of parts of a turbine such as a blade part hitting buildings or persons must be regarded as remote. Furthermore, wind turbines have several braking systems to stop the rotor blades in case of extremely strong winds or a malfunction. As modern well constructed wind turbine are also very quiet it normally possesses little problems to build wind turbines close to human settlements. The visual effects of wind turbine may, however, create some controversy, as some people believe they are having a severe negative visual impact of the landscape while others find them beautiful.

The impact on plants and animals is not very well established despite a sizable number of studies, but as it is with most power plants a certain amount of

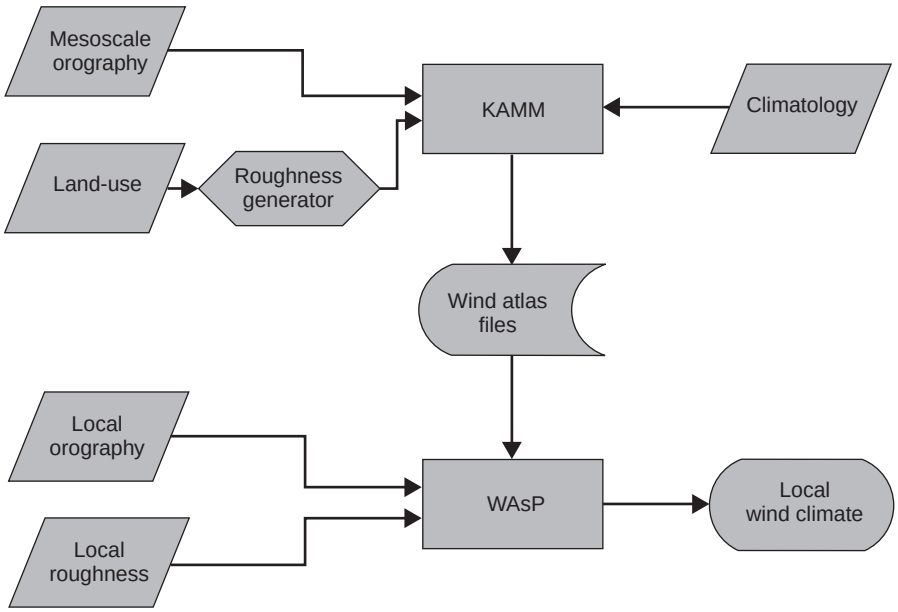


FIGURE 9 Predicting the wind climate by combining a mesoand microscale meteorological model. Flowchart showing the principle behind the calculation of a local wind climate by means of a mesoscale model (KAMM, the Karlsruhe Atmospheric Meso-scale Model) and a microscale model (WasP, the Wind Analysis and Application Program). For both models the main input is orography, surface roughness and long-term wind climatology. For the microscale model, the long-term wind climatology is produced by the mesoscale model and is named wind atlas files.

disturbance to flora, fauna birds and mammals will happen. Still the largest concern is on bird strikes and associated effect on bird population and migration paths. It is not alone the turbines that constitute a danger to the birds, associated utility structures in particular lines that connect the wind turbine farms with the electrical grid possess a danger to birds as they can collide with the cables or get electrocuted when landing on the towers.

The large wind farms located or being located along the Gulf of Suez in Egypt are placed just in the world's major bird migrating routes with large numbers of raptors, storks and other spices sometimes occurring at very low altitudes—clearly, the bird-wind turbine collision issue is of special concern to Egypt. It is expected that crucial information on the collision issue will be obtained by the intensive ornithological studies being performed.

4. TECHNICAL WIND FARM ISSUES

4.1 Turbine Spacing and Wakes

Most of the technical issues with wind farms are associated with the close spacing of multiple turbines. The close spacing implies that the extraction of energy by wind turbines upwind will reduce the wind speed and increase the turbulence, which may cause reduced efficiency and higher loads on downwind turbines. Another technical issue for large wind farms is the grid connection and the integration into the power system. This will be covered in the next section on power system integration.

When a wind turbine extracts energy from the wind, the rotor creates a pressure field that reduces the speed of the flow and expands the flow stream tubes around the wind turbine. As the air crosses the rotor plane, the pressure decreases suddenly, and in the immediate downstream region there is a deficit of pressure and axial velocity. The tip vortices, which form helical trajectories, can be seen as a cylindrical shear layer that separates the outside flow from the slower moving flow in the wake. As the air moves downstream, the velocity inside the wake decreases and the pressure increases until ambient pressure is reached in about one turbine diameter length from the rotor plane. The flow close to the rotor is illustrated in Fig. 10, where smoke is released near the tip and midway on the blade.

Further downstream, the wake diffuses due to turbulence caused by the shear in the wake as well as from the ambient turbulence, and approximately three rotor diameters downstream the wake is fully developed and is characterized by the mean velocity profile, the growth of the wake, and wake turbulence. The flow velocity in the wake is

FIGURE 10 Flow visualization showing tip vortices and the reduced flow velocity in the wake of a wind turbine. From Risø National Laboratory.

gradually restored by the diffusion of energy from the wind field above, making the wake shallower but broader. In general, the initial wake velocity deficit is assumed to depend on the thrust coefficient of the wind turbine, which again depends on the wind speed. The wake growth depends on the ambient plus the wake turbulence, although it is often assumed that the wake increases linearly with the downstream distance.

4.2 Wind Farm Efficiency

The reduced wind speed in the wake behind means that a wind turbine operating in the wake of an upstream wind turbine will produce less energy. Deep in a wind farm a wind turbine may experience several wakes, and a wind farm will thus produce less energy than the same number of turbines operating isolated under equal wind conditions. The ratio between the annual production of a wind farm and the production of corresponding isolated turbines is the wind farm efficiency. Several wake models have been developed to calculate production of a wind farm. A simple model, which assumes linear wake growth, is illustrated in Fig. 11. The effect of several wakes on one wind turbine is calculated by adding the deficits. While such a model is sufficient for smaller wind farms, the models does not account for the effect on the ambient wind climate from the wind farm, and caution should be exercised for very large wind farms.

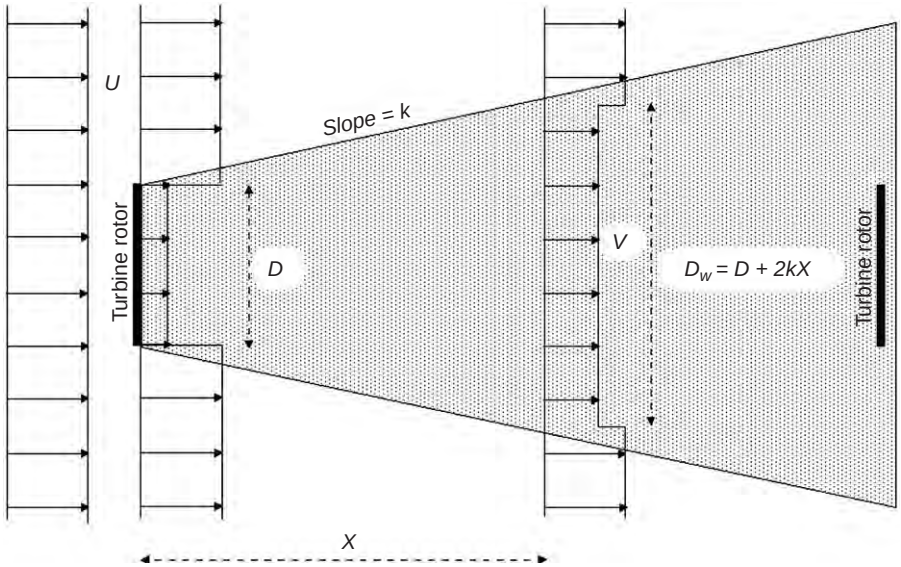


FIGURE 11 Principle of a simple wake model. The wake decay constant k ~ 0.1 and the velocity deficit is given by $\delta V \triangleq U - V = U(1 - \sqrt{1 - C_t})(D/(D + 2kX))^2$. From Katic, I., Højstrup, J., and Jensen, N.O. (1986). "A Simple Model for Cluster Efficiency." Proc. European Wind Energy Conference, October 7–9, 1986, Rome, Italy.

Using a wake model, the wind farm efficiency for a given wind farm layout can be calculated. For a given layout, the efficiency also depends on wind direction distribution, and at sites with a prevailing wind direction, the separation between turbines in this direction is often larger than perpendicular to the prevailing wind direction. In the micrositing process, the production from a given farm layout is optimized taking into account the terrain, needs for roads and power connections, and land use. Most wind farms, except in homogeneous terrain or offshore, will show an irregular layout and follow, for example, a ridge in the terrain. A thorough analysis should always be carried out, but as a rule of thumb, an efficiency of 90% is expected for a layout, where the downwind distance between turbines times the crosswind distance is 50 times the rotor diameter D squared (i.e., a spacing of $7\,D \times 7\,D$ or $10\,D \times 5\,D$).

4.3 Wind Farm Loads

In addition the wind farm efficiency, the wind turbine loads are affected by the farm layout, as the variations in the wind flow field from the wake velocity deficit and the wake turbulence will increase in particular the fatigue but also the ultimate loads. The load increase depends on the distance to the wind turbine and on the wind speed. Thus, the relative load increase is largest at lower wind speeds, where a high proportion of the energy is extracted from the flow, and smaller at higher wind speeds, where relatively less of the energy flow is extracted in order to limit the power and load on the wind turbine drivetrain. However, most fatigue damage and ultimate loads on the wind turbine structure occur at high wind speeds where the attention should be focused. At very high wind speeds, where the wind turbines are standing still or idling slowly and so to speak transparent to the wind, the wake effects are negligible. Furthermore, the relative load increase is largest when the ambient turbulence is small. This is the case over water, and due consideration of wake effects is essential for the determination of dynamic wind turbine loads during operation in an offshore wind farm.

In a wind farm with wind turbines arranged in a row perpendicular to the prevailing wind, the separation between the wind turbines may be less than three rotor diameters. When the wind is along the row, the wind turbines are thus in the near wake, where the shape of the wake is very distinct and where the highest loads occur during half wake situations. In terms of the equivalent load (a constant amplitude load causing the same fatigue damage as the real random amplitude load) on the wind turbine blade in the flap wise direction, the wake may double the load at lower wind speeds compared to the loads on a similar freestanding turbine. While this wind situation is rare, loads on closely spaced wind turbines are still a research topic, and caution should be exercised in this situation.

When the wind turbines are spaced further away from each other, the effect of being in the wake of an upstream turbine is still noticeable up to a distance of about 20 rotor diameters. A number of measurement investigations have been carried out for wind farms with spacing larger than three rotor diameters, including at the Danish offshore wind farm at Vindeby, where the wind turbines are installed in two rows with a spacing of 8.5 rotor diameters. It is interesting to notice that the loading and the turbulence levels are closely related, as is demonstrated in Fig. 12, where the added turbulence and the increase in load (equivalent flap wise blade load) in the wake situation are clearly identifiable. The figure also suggests that a practical way to represent wake loads in load calculations is in terms of an added wake turbulence or an effective turbulence intensity, which depends on the wind turbine behavior, the spacing to the closest wind turbines, and the ambient turbulence. For fatigue loads, the effective turbulence intensity must also include the nonlinear fatigue damage sensitivity to higher loads for the used material. This is the approach used in the international IEC standard giving design criteria for wind turbines.

5. POWER SYSTEM INTEGRATION

5.1 Capacity Credit

The capacity of wind farms as defined here varies from less than 1 MW to several hundreds of MW. Consequently, the power system integration issues vary significantly, from incorporation of a small negative load in the distribution system with local power quality concerns to requiring a large wind farm connected to the power transmission system to possess power plant characteristics and share responsibility for power system stability and power balance. The size of wind farm projects has become larger, in many cases with wind farm capacities greater than 100 MW. This is especially the case with offshore wind farms, which are being planned and implemented along the northern European coasts. In the following discussion, we shall present the power

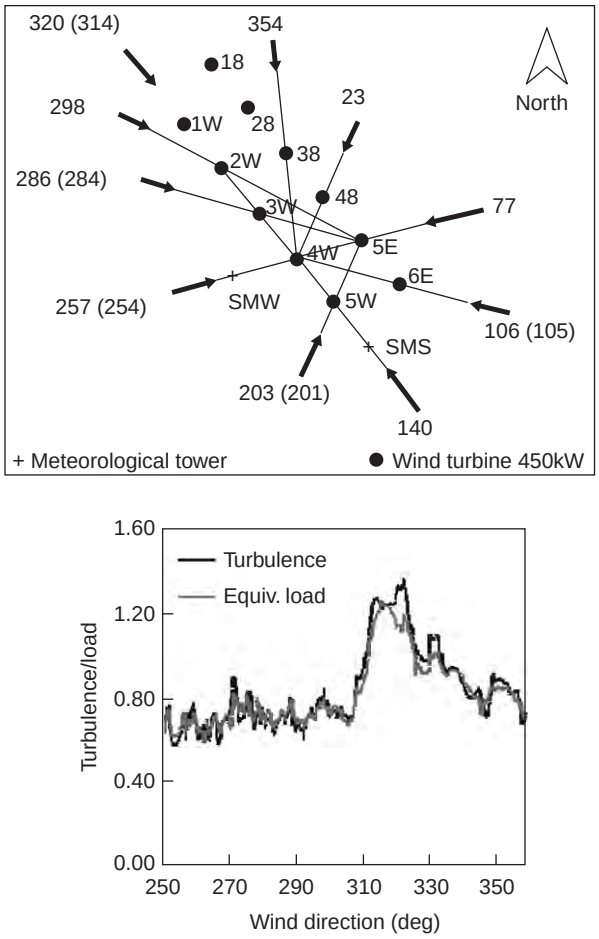


FIGURE 12 Layout and wake situations for the Vindeby offshore wind farm. The plot at the right shows the normalized equivalent blade flapwise load of turbine 4W and the normalized turbulence level at mast SMS. The turbine and the mast are in identical wake situations. From Frandsen, S. (2003). "Turbulence and turbulence-generated fatigue loading in wind turbine clusters." Risø-R-1188.

system issues of greatest concern for wind farm integration, namely capacity credit, short-term variability of wind power production, control, power system stability, and power quality. Finally, some system solutions for connecting wind farm to the power transmission system are discussed.

Because of the intermittent nature of the wind resource, a wind farm cannot be relied on to operate at a certain capacity at any given time. The question therefore arises, can a wind farm substitute for conventional generation capacity and if so, how much? This is expressed by the capacity credit, which is usually defined as the fraction of a power plant's rated capacity, which is on average available at the time of peak demand, or alternatively, the relative amount of additional conventional capacity that is needed in the absence of wind energy in order to obtain the same loss-of-load-probability (LOLP). The LOLP method is commonly used to assess the reliability of a power system with conventional generation and calculates the probability that the generation system cannot meet the required electricity demand due to random and scheduled outages of generating capacity. Both definitions give similar results for small penetration levels.

The capacity credit is determined by a number of factors, such as the average wind speed, the wind speed variability, the percentage of total electricity demand covered by wind energy (penetration), the correlation between the wind at different wind farms in the power system and the correlation between wind speed and demand. For small penetration levels, the capacity credit of dispersed wind farms in a large grid is generally close to the average output of the plants—that is, the installed capacity times the capacity factor of the wind plants (500 MW of wind farms with capacity factor of 25% can be expected to have a capacity credit of 125 MW).

5.2 Short-Term Variability of Wind Power Production

While wind farms can thus substitute for conventional generation capacity in a power system, the fluctuating character of the wind farm production needs to be addressed. As conventional generation capacity may take a while to prepare for generation, an amount of unused generation capacity, spinning reserve, is kept on line to respond to fluctuations in wind power as well as load. To take full advantage of wind power, to control the dispatch of conventional generation, and to trade the power production on markets operating on a 48 h timescale, systems that predict the wind power production 1 to 2 days ahead are necessary. Quite a few methods for short-term prediction of the wind farm production have been developed and implemented in load dispatch centers in, for example, Denmark and the United States. Most methods constitute a numerical weather prediction model, a physical model of the wind farm to relate wind speed to production, accounting for farm layout, power curves as well as wake and terrain effects, measurements of production and statistical analysis, and, finally, preparation of output in terms of, for example, energy production from the wind farm in hourly steps from 0 to 48 h ahead.

The accuracy of short-term prediction models can be evaluated by comparing their prediction with the result of the persistence model, which simply

assumes that production at any given time is equal to the production during the previous time period. As weather systems develop slowly, the persistence model is actually quite accurate and difficult to outperform. Figure 13 presents an example comparison between the persistence and a short-term prediction model based on the weather forecast for an actual 5.2 MW wind farm in Denmark and based on 1 year of data. The mean error (circles) and the mean absolute error (squares) are shown for the persistence model (filled symbols) and for the prediction model in absolute values as well as relative to the wind farm capacity. Note that the weather prediction–based model first outperforms the persistence model after 6 h and that the mean absolute error is almost constant at 15% of the wind farm capacity up to 36 hours. In general, state-of-the-art weather prediction–based models are expected to reduce the mean absolute error to half the error of the persistence model.

Finally, a subjective evaluation of the performance of prediction models can be made by plotting the actual performance of the model together with the observed production. An example is shown in Fig. 14.

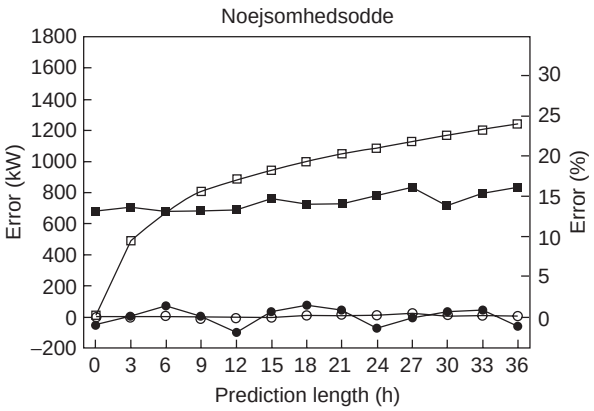


FIGURE 13 Performance of a weather prediction based model and the persistence model. From Landberg, L. (1999). "Operational Results from a Physical Power Prediction Model." In Proceedings from EWEC99, Nice (FR), March 1–5, 1086–1089.

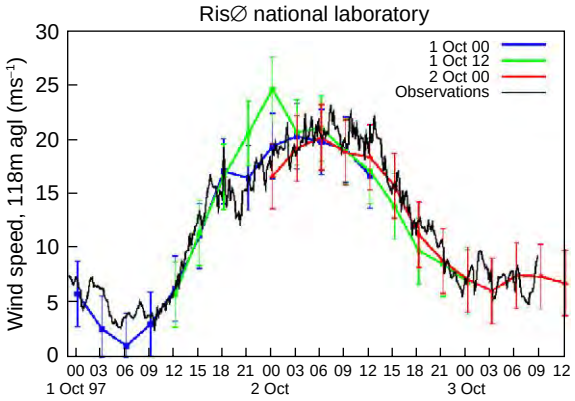


FIGURE 14 A storm lasting 2 to 5 days predicted by the model Predictor. The black line is the observed wind speed at the 118 m tall Risø mast, the gray lines are twice daily prediction of Predictor. From Landberg, L., Giebel, G., Nielsen, H. A., Nielsen, T., and Madsen, H. (2003): Short-term prediction—An overview. *Wind Energy* **3**, 273–280.

5.3 Energy Storage

Another way to improve the integration of wind farms by reducing the power fluctuations is through the use of energy storage. Most storage technologies such as lead-acid batteries are not suited for large-scale application, but new storage technologies with promise have emerged. One is based on the Redox flow technology, which has a high efficiency (80%) and a long lifetime (>10 y). A Redox flow plant consists of a cell stack in which two electrolytes can interact through a membrane and storage tanks for the two electrolytes. The power capacity is given by membrane area, while the storage tank volumes determine the energy capacity. Hence, the ratings of the energy and the power capacity are independent. Another promising technology is reversible fuel cells and hydrogen. In case of surplus electricity, the fuel cells produce hydrogen, which, in case of a deficit can be used for power production or transportation.

5.4 Control and Power System Stability

A high wind power penetration imposes additional control needs on a power system. This is especially of concern for large wind farms connected to the transmission system at a high-voltage level (>60 kV), where the wind farm behavior may influence the power system stability. In Denmark, two large offshore wind farms of 150 MW each were installed in 2003 in a power system where wind energy already covers 15% of the electricity demand. Due to their rated capacity, the wind farms are connected to the transmission system, and new requirements have been developed by the Danish Transmission system operators, Eltra and Elkraft System, giving specifications for connecting wind farms to the transmission system, which can be seen as indicative for future requirements to wind farms with regard to control, stability, and power quality.

Not surprising, specifications are given for grid frequencies, voltages, reactive power compensation,

voltage quality, and fault protection. Moreover, to enable wind farms to substitute conventional power plants, a requirement for production control and more stringent stability requirements and requirements for operation during faults have been introduced. Traditionally, wind turbines have been equipped with protection functions that disconnect and stop the turbines in the case of a temporary or permanent grid fault. This philosophy has been abandoned, and wind turbines are now required to be able to ride through temporary faults and contribute to system services such as momentary reserves and short-circuit capacity.

The requirements for production control constitute the ability to do the following:

- Control the production from the wind farm so that it does not exceed a certain MW limit (i.e., a production limit)
- Control the production limit by a single central signal
- Control each wind farm individually
- Have the control act on the individual wind turbine
- Control the production quickly (i.e., a reduction from 100% to below 20% of maximum power in less than 2 seconds)

Furthermore, high wind speeds may not cause all turbines to stop simultaneously.

5.5 Power Quality

Wind farms influence the quality of power in the electricity grid of the power system. The term power quality is not well defined, but it usually includes two aspects of power supply: voltage quality and supply reliability. The voltage quality part disturbances such as rapid changes, harmonics, interharmonics, flicker, unbalance, and transients, reliability involves longer duration phenomena such as interruptions, voltage dips and sags, over and under voltages, and frequencies. Methods for measurements and assessment of the influence of wind turbines on the power quality are given in the international standard IEC 61400-21. The standard includes measures for maximum power, reactive power, voltage fluctuations, and harmonics. In general, the main power quality issue for wind farms with a large number of turbines is the steady-state voltage increase, while for individual large wind turbines in a weak grid, voltage fluctuations and flicker become an additional important issue. This is due to the power smoothing effect from the averaging of usually close to uncorrelated power fluctuations from the individual wind turbines. Wind turbines will normally increase the voltage level in the distribution system due to their production of active power, while reactive power consumption will decrease the power level. Capacitor banks and transformer tap changers are the most common ways to control the steady-state voltage level. It is generally desirable to reduce the active power consumption, either to reduce losses in the system or to avoid reducing the production capacity of strained generation capacity. Reactive power consumption may also affect the power system stability, as a voltage instability or collapse may occur if the power system cannot supply a change in reactive power demand from a rapid change in power production. A fast control of reactive power compensation is necessary to deal with such power stability issues in wind farms. This may be accomplished by the use of wind turbines with active voltage control or by the use of external compensators such as static var compensators (SVCs).

5.6 System Solutions

Four different system solutions for connecting large wind farms to the high-voltage transmission system are shown in Fig. 15. To meet the requirements for production control, the wind turbines in the wind farm are equipped with blade angle control (pitch or active stall) and can therefore also contribute to the frequency control by regulation of the active power. Furthermore, wind turbines or installation with power converters are also able to control the reactive power supply to the power system.

In the traditional solution A, the reactive power consumption in the wind turbine induction generator is compensated by shunt capacitors in the individual wind turbines. The power is the transformed to the medium voltage level (10 to 40 kV) in the farm distribution system before being transformed to the high voltage level (>60 kV) of and transmitted to the grid transmission system at the connection point. The solution relies on a strong grid and power system and may be supplemented with a reactive power compensation unit for the fast reactive power control at a point before the high-voltage transformation.

In solution B, each wind turbine is equipped with a power converter, which in general keeps the power factor close to unity (no reactive power demand) and smoothes the power fluctuation through variable speed operation. The power converter may also be used for voltage control by regulation of the reactive power supply.

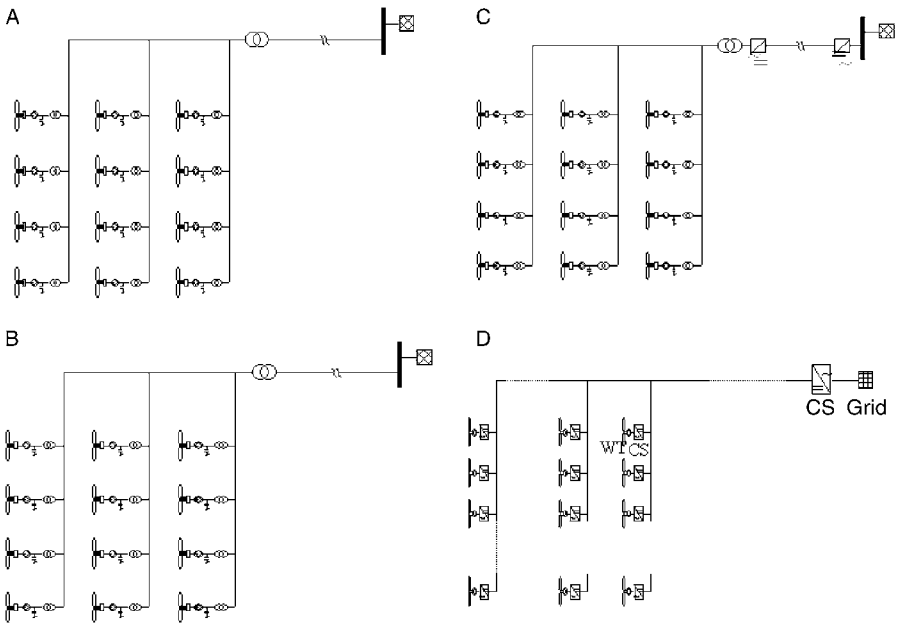


FIGURE 15 System solutions for connection of wind farms. AC and HVDC transmission of power from the wind farm to the grid. From Hansen, L. H., *et al.* (2001). "Conceptual Survey of Generators and Power Electronics for Wind Turbines," Risø-R-1205 (EN).

TABLE II

Characteristics of System Solutions for Grid Connection of Wind Farms

System solution	A	B	C	D
Active power control	Yes	Yes	Yes	Yes
Reactive power control	Yes	(Central)	Yes	Yes
Active at faults	Yes	No	Yes	Yes
Short-circuit power contribution	Yes	Yes	No	No
Regulation bandwidth	10–100 ms	200 ms–2 s	10–100 ms	10 ms–10 s
Ancillary services	Yes	No	Yes	Yes
Flicker contribution	Yes	Yes	No	No
Drive train load reduction	No	Yes	No	Yes
Soft starter needs	No	Yes	Yes	Yes

One of the most widely discussed concepts for grid connection of large wind farms is high voltage direct current (HVDC). A modern HVDC link consists of two fully controllable AC/DC converter stations linked with a DC cable. The DC cable is less expensive than an AC transmission cable, but the cost of the converter station makes this solution more expensive than an AC link, unless the distance is very long. The development of power electronics reduces the cost of converter stations rapidly and may make HVDC viable for many applications. The viability of HVDC may be reached earlier for offshore wind farms. When the critical cable length for AC transmission is reached, a compensation unit is required on the AC transmission system, which is also expensive offshore. In addition to the advantages to the power system of having a fully controllable DC/AC converter at the connection point to the transmission system, a HVDC link opens new possibilities for wind turbine control, as the frequency of the wind turbine connection grid in the farm can be controlled independently of the power system frequency.

The simplest HVDC solution is the group connection in C, where the HVDC link connects the power system and the wind farm AC grid. The converter at the wind farm end can supply the wind turbines with reactive power, and the most obvious

FIGURE 16 Placing the wind turbines at the Nysted offshore wind farm. Offshore wind farm technology was developed in 2003 to a point where on the average it took one day to erect each of the 110 m high wind turbines (Compagnies Bonus and A2Sea). From Jan K. Winther.

FIGURE 17 The Nysted wind farm, showing rows of the 2.1 MW Bonus turbines and the transformer station. From Energi E2.

solution is to equip the wind turbines with directly coupled induction generators without capacitor banks for reactive power compensation or with a simpler constant no-load compensation. The control of the wind farm grid frequency, and hence the rotational speed of the turbines, allows aerodynamic efficiency optimization. However, compared to individual speed control, the total efficiency is smaller due to variations in wind speed and thus optimum rotation speed from wind turbine to wind turbine. Neither can the benefits of variable speed operation with individual power converters such as smoothing the power fluctuations and reducing drivetrain load transient be obtained.

The solution in D appears to have all the benefits of individual wind turbine power converters, if the AC/DC converters in the wind turbines are based on forced-commutated semiconductors. Then the frequency and hence the turbine rotation speed can be controlled independently to reduce drivetrain loads and power fluctuations. The benefits and limitations of the four solutions are listed in Table II.

TABLE III

Main Data of Two Offshore Wind Farm Examples

Wind farm characteristics	Horns Rev wind farm	Nysted wind farm
Installed capacity	160 MW	158.4 MW
Number of turbines	80	72
Wind turbine type	Vestas 2 MW	Bonus 2.2 MW
Expected annual production	600 GWh	595 GWh
Hub height	70	70 m
Wind farm area	20 km^2	24 km^2
Water depth	6.5–13.5 m	6–9.5 m
Distance to shore	14–20 km	10 km
Distance between rows	560 m	850 m
Distance between turbines in rows	560 m	480 m
System solution type	B	A
Internal grid voltage	34 kV	33 kV
Transmission to shore voltage	150 kV	132 kV

6. WIND FARM EXAMPLES

This article has presented several illustrations of large offshore wind farms, namely the Horns Rev offshore wind farm and the Nysted offshore wind farm, Figs 16 and 17 (also see Table III). These wind farms, each of approximately 160 MW, have been built as large-scale demonstration projects according to an agreement between the electricity sector and the Danish government. These wind farms use wind turbines adapted to offshore conditions with capacities of 2 to 2.2 MW. The expected electricity production corresponds to almost 3% of the electricity consumption in 2002 in Denmark. These

are the first Danish wind farms designed according to the new requirements from the transmission system operators, where the wind farms are considered power plants with responsibilities for control, stability, and power balance.

While large, these wind farms are not the largest in the world. At present, the record goes to the King Mountain Wind Ranch in Texas with a capacity of 278 MW and consisting of 214 1.3 MW Bonus wind turbines. Nor will the turbine size in future offshore wind farms be limited to the 2 MW size; most major wind turbine manufacturers are testing or developing wind turbines in the 3 to 5 MW range for offshore (and onshore) wind farm applications. Even 10 MW turbines seem realistic with modern technology.

SEE ALSO THE FOLLOWING ARTICLES

Hybrid Energy Systems • Land Requirements of Energy Systems • Tidal Energy • Wind Energy Economics • Wind Energy, History of • Wind Energy Technology, Environmental Impacts of • Wind Resource Base

Further Reading

Burton, T., Sharpe, D., Jenkins, N., and Bossanyi, E. (2001). "Wind Energy Handbook." John Wiley & Sons, Chichester, England.

EUREC Agency (Eds.) (2002). "The Future for Renewable Energy 2. Prospects and Directions." James & James, London.

Frandsen, S. (2003). "Turbulence and Turbulence-Generated Fatigue Loading in Wind Turbine Clusters." Risø-R-1188. Risø National Laboratory, Roskilde, Denmark.

Heier, S. (1996). "Grid Integration of Wind Energy Conversion Systems." John Wiley & Sons, Chichester, England.

Madsen, B. (2003).World Market Update 2002, BTM Consult Aps, March.

Petersen, E. L., Mortensen, N. G., Landberg, L., Højstrup, J., and Frank, H. P. (1998). Wind power meteorology. Part 1: Climate and turbulence. *Wind Energy* **1**, 2–22.

Petersen, E. L., Mortensen, N. G., Landberg, L., Højstrup, J., and Frank, H. P. (1998). Wind power meteorology. Part 2: Siting and models. *Wind Energy* **1**, 55–72.

Redlinger, R. Y., Andersen, P. D., and Morthorst, P. E. (2002). "Wind Energy in the 21st Century. Economics; Policy, Technology and the Changing Electricity Industry. United Nations Environment Programme." Palgrave, New York.

Spera, D. A. (ed.). (1998). "Wind Turbine Technology: Fundamental Concepts of Wind Turbine Engineering." ASME Press, New York.

Troen, I., Petersen, E. L. (1989). European Wind Atlas. Risø National Laboratory.

Walker, J. F., and Jenkins, N. (1997). "Wind Energy Technology." John Wiley & Sons, Chichester, England.

Wind Resource Base

DENNIS ELLIOTT, MARC SCHWARTZ, and GEORGE SCOTT
U.S. Department of Energy, National Renewable Energy Laboratory
Golden, Colorado, United States

1. Introduction
2. Basics of Wind Resource
3. Baseline Wind Resource Studies
4. Data Sets and Analysis Methods
5. Resource Assessment Methods
6. Status of Wind Resource Projects

Glossary

exclusion zone An area that cannot be used for wind power development because of environmental, legal, topographic, or other reasons.
frequency distribution of wind speed The percentage of time that the wind is blowing at each speed (14% at 2–3 m/s, 8% at 3–4 m/s, etc.).
geographic information system (GIS) The computer hardware, software, and technical expertise that inputs, stores, maintains, manipulates, analyzes, and outputs geographically referenced data. A GIS combines the power of spatial database management with high-resolution graphic display to effectively present information.
grid-connected A project that is connected to a utility distribution system.
power class A single number representing a range of wind power densities (WPD). A typical power classification scheme divides the WPD range from 0 to $>800\,W/m^2$ into seven power classes.
Weibull distribution A mathematical function that is often a good approximation to the frequency distribution of wind speeds at a site.
wind electric potential The total number of megawatts of wind energy generating capacity that could be installed in a given area.
wind power density The average wind power over one square meter expressed in Watts per square meter (W/m^2) in a vertical plane perpendicular to the wind direction.
wind resource A measure of how much wind energy is available at a given height at a given point or over a given area. It can be expressed in values of Watts per square meter, power class bins, or estimated yearly energy production.
wind resource base The collection of all data about the wind resource in a given area. The data set may include average speeds and power densities, measures of variability, and energy production statistics.
wind shear A measure of the change in wind speed or power with height above the ground.

A major barrier to the deployment of wind energy in many regions of the world is the lack of reliable, detailed wind resource data. Wind resource assessments are crucial in determining the feasibility of wind energy projects and are essential for government and industry to identify areas of potential wind energy development. Accurate assessments can also aid the development of policies and regulations that can enhance wind energy deployment. This article summarizes the techniques used to conduct wind resource base assessments and the status of past, present, and future assessment activities.

1. INTRODUCTION

Wind is a renewable energy resource that can generate significant amounts of electricity for both grid-connected and standalone applications. Major strides in wind technology development have increased efficiencies and reduced wind energy costs so that they are competitive with other generation technologies in regions with good-to-excellent wind resources. However, a major barrier to the deployment of wind energy in many regions of the world is the lack of reliable, detailed wind resource data. Wind resource assessments are crucial in determining the feasibility of wind energy projects and availability of these data is essential for government and industry to identify areas of potential wind energy development. Accurate assessments can also aid the

development of policies and regulations that can enhance wind energy deployment.

Wind resource assessments produced from the 1980s to early 1990s, such as the U.S. and European wind atlases and a worldwide wind resource map, identified general areas of good wind resource and provided the basis for preliminary estimates of the wind-energy potential. However, these assessments were of limited use, primarily because of the lack of detail on the spatial variability of the wind resource and, in many regions of the world, the lack of reliable surface measurement data that largely provided the basis of the resource estimates.

Advances in wind resource assessment are leading to improved wind resource maps, even in areas with little or no surface measurements. This article describes the process of wind resource assessment and describes the newest techniques for generating wind resource maps and databases.

2. BASICS OF WIND RESOURCE

2.1 Wind Speed and Direction

Wind speed is the simplest representation of the wind at a given point. Anemometers or other calibrated instruments measure wind speed. Wind speeds can be calculated as an average or expressed as an instantaneous value. Wind speed averaging intervals commonly used in resource assessment studies include 1- or 2-minute (weather observations), 10-minute (the standard for wind energy monitoring programs), hourly, monthly, and yearly periods. It is important to know the measurement height for a given wind speed because the wind speed can vary significantly with height. It is also desirable to know the exposure of a particular location to the prevailing winds because nearby obstacles such as trees and buildings can reduce the wind speed.

Wind direction is measured with a wind vane, usually located at the same height as the anemometer. Knowledge of the prevailing wind direction(s) is important in assessing the available resource for wind turbine siting and spacing. Correct alignment of the wind vane to a reference direction is important to accurately measure the wind direction. Wind direction observations at meteorological stations are often based on a 36-point compass (every 10°). Some wind direction data are expressed in less precise 8-point (every 45°), 12-point (every 30°), or 16-point (every 22.5°) intervals.

The wind direction distribution is often presented as a wind rose (a plot of frequency of occurrence by direction). Wind roses can also represent quantities such as the average speed or the percentage of the available power for each direction.

2.2 Wind Power Density

Although the wind resource at a site can be roughly described by the mean wind speed, the wind power density (WPD) provides a truer indication of a site's wind energy potential. Wind power density expresses the average wind power over one square meter (W/m^2) in a vertical plane perpendicular to the wind direction. The power density is proportional to the cube of the wind speed and to the air density. Because of this cubic term, two sites with the same average wind speed but different speed distributions can have very different WPD values. The wind power density, in units of W/m^2, is computed by the following equation:

$$WPD = \frac{1}{2n}\sum_{i=1}^{n} \rho_i \cdot v_i^3,$$

where WPD is the wind power density in W/m^2, n is the number of records in the averaging interval, ρ_i is the air density (kg/m^3) at a particular observation time, and v_i^3 is the cube of the wind speed (m/s) at the same observation time.

This equation should only be used with individual measurement records (hourly, 10-minute, etc.) and not with long-term average speeds such as monthly or yearly values. Using this equation with long-term averages will underestimate the WPD because long-term averages will smooth out most of the higher speed records that would more accurately represent the fundamental nature of the wind resource. Long-term averages of WPD should be computed by averaging the WPDs computed over many shorter intervals.

The air-density term is dependent on temperature and pressure and can vary by 10 to 15% seasonally. If the site pressure and temperature are known, the air density can be calculated using the following equation:

$$\rho = \frac{\mathrm{P}}{\mathrm{R} \times \mathrm{T}},$$

where ρ is the air density in kg/m^3, P is the air pressure (Pa or N/m^2), R is the specific gas constant for air ($287\ J/kg \cdot K$), and T is the air temperature in degrees Kelvin (°C + 273).

An approximation based on altitude alone is

$$\rho = 1.225 - (1.194 \times 10^{-4})Z,$$

where Z is the elevation in meters.

This air density equation can be substituted into the WPD equation for the determination of each instantaneous or multiple average values.

2.3 Wind Speed Frequency Distribution

The wind speed frequency distribution characterizes the wind at a given location in two ways. First, the frequency distribution determines how often a given wind speed is observed at the location; second, it identifies the range of wind speeds observed at that location. This analysis is often accomplished by sorting the wind speed observations into 1-meter-per-second (m/s) bins and computing the percentage in each bin. The wind speed distribution is important because sites with identical average wind speeds, but different distributions can have substantially different wind resources. These differences can be as great as a factor of two or three.

The wind speed frequency distribution in many areas can be closely approximated by the Weibull distribution function. The Weibull function is defined as

$$f(V) = (k/c)(V/c)^{k-1}\exp(-V/c)^{k},$$

where $f(V)$ is the Weibull probability density function, the probability of encountering a wind speed of V m/s, c is the Weibull scale factor, which is typically related to the average wind speed through the shape factor, expressed in m/s, and k is the Weibull shape factor, which describes the distribution of the wind speeds. Detailed explanations of the Weibull distribution function and its application are available in many texts.

Common values of k for measured wind speeds are around 2.0, with a typical range of 1.5 to 3.0. Higher values of k indicate a narrower speed distribution (more constant winds). Because of the cubic nature of the wind power equation, lower k values have higher average wind power densities for the same average wind speed. Figure 1 shows the appearance of the Weibull distribution at different k values. Values of k are useful for describing frequency distributions, deriving power classes from wind speeds, synthesizing missing data, and other uses.

A useful approximation for k is

$$k = \left[\frac{\sigma}{V}\right]^{-1.086},$$

where σ is the standard deviation of wind speed, and V is the average wind speed.

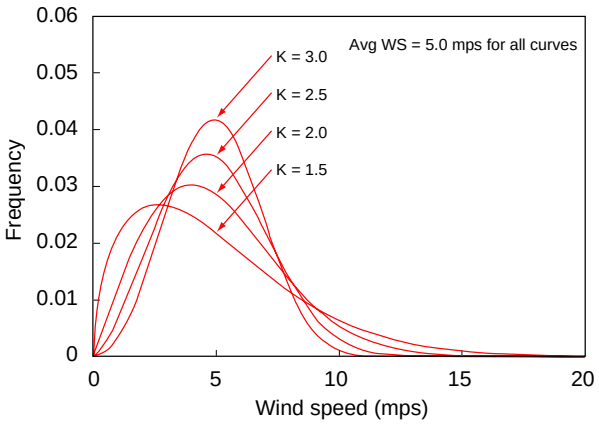


FIGURE 1 Weibull distribution function for four values of k. All curves have an average wind speed of 5.0 m/s.

2.4 Spatial Variation

Mapping spatial variations in wind resource is crucial to any wind resource assessment project. The wind resource in an area can vary widely, even on a scale as small as a few hundred meters. The resource also varies with height, usually increasing with height above ground.

2.4.1 Wind Shear and the Power Law

Frictional effects cause the wind speed near the surface to be lower than that at greater heights above ground. In the near-surface layer (< 300 m above ground level [agl]), the wind shear, or variation of wind speed with height, often approximates a power law:

$$V_2 = V_1\left[\frac{h_2}{h_1}\right]^{\alpha},$$

where V_1 is the wind speed at height h_1 above ground, V_2 is the wind speed at h_2, and α is the shear exponent.

The value of α is around 1/7 (giving rise to the so-called one-seventh power law) for smooth terrain but can climb as high as 0.5 in areas of high surface roughness or drop lower than 0.1 over very smooth surfaces such as water or ice. A similar law can be derived for wind power density, but the exponent will be 3α, because the power density is proportional to the cube of the wind speed. This increase of wind power density with height is the driving force behind the trend toward increasingly tall wind turbine towers. Figure 2 shows how the wind speed and power vary with height for different values of α.

The power law equation can be used to extrapolate measurements made at low levels up to the expected turbine hub height. This can be particularly useful when trying to adjust meteorological station

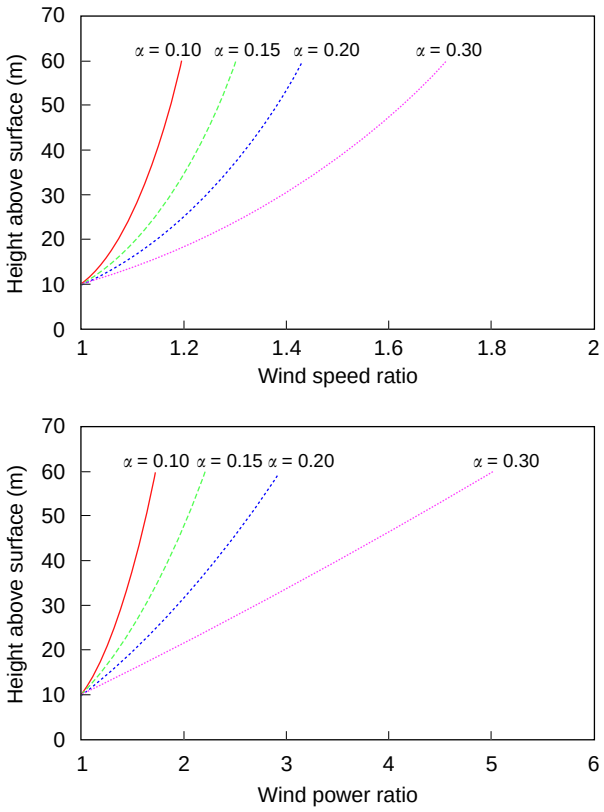


FIGURE 2 Wind speed and power shear curves for different values of the wind shear exponent α.

data or other measurement data to the hub heights of modern wind turbines. Above the near-surface layer, other effects can dominate the frictional effects, causing a departure from the power law.

2.4.2 Geographic and Climatological Effects

Global circulation patterns, such as the mid-latitude westerly winds and the equatorial trade winds, dominate the wind climate in some regions. Large-scale geographic features, such as the shape and location of continents and oceans, can modify these patterns and create regions of good wind resource. An example is the Isthmus of Tehuantepec in Oaxaca, Mexico, where the high pressure found in the Gulf of Mexico and the low pressure of the Pacific Ocean combine to form a strong north wind during much of the year. In other regions, differential heating of oceans and land masses causes seasonal winds such as the monsoon in the Indian Ocean.

2.4.3 Topographic Effects

Topographic features can have a major effect on wind speeds and directions. Winds accelerate as they pass over hilltops and ridges that are oriented perpendicular to the wind direction. High wind resources can occur at relatively low elevations in mountainous regions where the air flow is channeled through constrictions or corridors that enhance the wind speeds. However, the higher wind speeds can be accompanied by increased turbulence, so careful study of such areas is needed before wind turbines are installed.

Topography can lower the wind speed as well as raise it. Examples of low wind areas include the lee side of ridges, the areas immediately upwind and downwind of hills, and areas of rough (many small hills and valleys) terrain. Differential heating between two areas at different elevations can also affect the regional wind resource.

2.5 Temporal Variations

Because the wind speed and direction change over time, the wind developer must take these variations into account when designing a wind project.

2.5.1 Interannual Variation

Annual average wind speeds at a given site can vary from year to year. This variation is often up to $\pm 10\%$ in wind speed, which translates into a ± 20 to 30% variation in wind power density. Shorter measurement periods (1 to 5 years) increase the uncertainty of the annual averages.

2.5.2 Seasonal Variation

Wind variations that tend to repeat at the same time of the year are called seasonal patterns. Seasonal patterns in many areas of the world are complex and are best determined by measured data. A single peak in wind resource lasting 2 to 3 months is often observed, although some regions have double-peak or other seasonal patterns.

2.5.3 Diurnal Variation

Wind variations that tend to repeat on a daily basis are called diurnal patterns. Many of these occur as a pronounced maximum wind speed at the same time (often the afternoon) each day. There may also be an associated directional pattern. Diurnal patterns may vary over the course of the year with different seasonal conditions.

An example of a diurnal pattern is the onshore/offshore winds found at coastal sites, where the wind blows toward the land during the day and toward the sea at night. Diurnal patterns may vary by height above ground and elevation above sea level. For example, some sites may exhibit a daytime maximum

near the ground (10 m) but a nighttime maximum near turbine height (50 m). Exposed ridge-crest locations frequently have a nighttime maximum and a daytime minimum, both at the ground and at turbine hub height. Some sites, such as offshore locations, do not exhibit significant diurnal variations.

2.5.4 Turbulence and Medium-Scale Variations

The wind can vary at timescales much shorter than the seasonal and diurnal scales. Rapid variations in wind speed and direction can significantly affect turbine performance. Typically, inflow variations are modeled using some type of inflow simulation tools that capture both the temporal and spatial variations in the wind. These very rapid fluctuations, on a scale of a few seconds or shorter, are usually characterized as inflow turbulence. The effects of turbulence on a wind turbine are outside the scope of this particular work.

3. BASELINE WIND RESOURCE STUDIES

3.1 Global Wind Resource

A worldwide wind resource map was produced in 1981 by the U.S Department of Energy's (DOE) Pacific Northwest Laboratory and published in a World Meteorological Organization document on wind energy utilization. The map, shown in Fig. 3, was based on an analysis of available wind data, primarily from meteorological agencies, and other information to estimate the broad-scale distribution of annual average wind power throughout the world. Vast areas of the world had very little or no reliable surface wind measurement data. Thus, regional climatological information and upper-air wind data were used to estimate the wind resource in many regions. In spite of these limitations, this world wind map has provided a good basis of information for preliminary estimates of the wind resource on a broad scale. The map shows that every major continent has substantial areas with good wind resource. With the exception of the equatorial regions, the offshore wind resources are vast.

The worldwide wind resource map was first used in developing estimates of the global wind energy potential in 1991. A study by the University of Utrecht in the Netherlands estimated that 25% of the world's land surface has at least moderate wind resource and 13% has at least good wind resource. In estimating the wind potential, the study assumed that wind turbines would be installed on 10% of the windy land area. The resultant global potential for onshore wind power production from the higher wind resource areas was estimated to be slightly more than the world's electricity consumption (1 terawatt [TW]) in 1987.

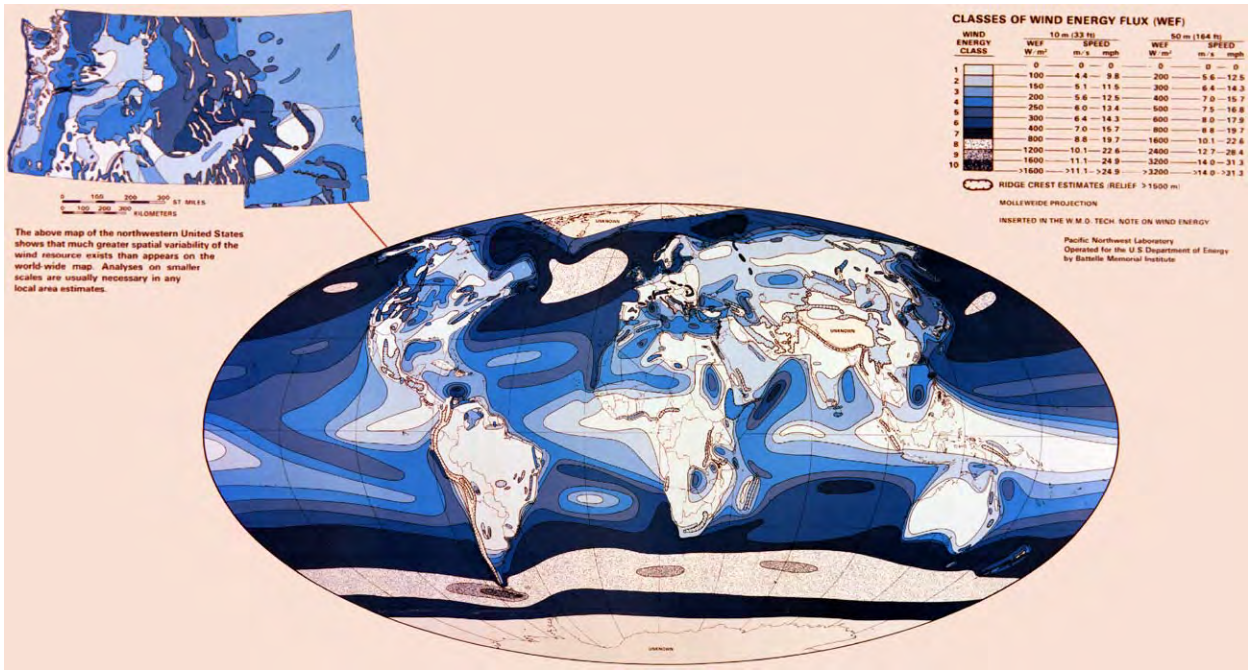


FIGURE 3 World wind resource map (1981).

3.2 United States Wind Resource

In 1987, the DOE published the "Wind Energy Resource Atlas of the United States." This assessment found that areas potentially suitable for wind energy applications are dispersed throughout much of the United States. Figure 4 shows the distribution of the annual average wind resource estimates throughout the United States. Major wind resource areas include much of the Great Plains from northwestern Texas and eastern New Mexico northward to Montana, North Dakota, and western Minnesota; the Atlantic coast from North Carolina to Maine; the Pacific coast from Point Conception, California to Washington; the Texas Gulf coast; the Great Lakes; portions of Alaska, Hawaii, Puerto Rico, the Virgin Islands, and the Pacific Islands; exposed ridge crests and mountain summits throughout the Appalachians and the western United States; and specific wind corridors throughout the mountainous western states. Updated high-resolution wind resource (1-km or finer) maps have been completed for many states and are either in progress or planned for additional states.

In a study completed for DOE in 1993, estimates of the electricity that could potentially be generated by wind power and of the available windy land area available were calculated for the contiguous United States. The estimates were based on wind resource data published in the 1987 U.S. wind atlas and excluded windy lands that are not suitable for development as a result of environmental and land-use considerations. Despite these exclusions, the potential electric power from wind energy was surprisingly large. Good wind resource areas, which cover 6% of the contiguous U.S. land area, have the potential to generate more electricity than the total U.S. electricity consumption in 1990. To provide 20% of the U.S. electricity, the study estimated that only about 0.6% of contiguous U.S. land area would have to be developed with wind turbines.

3.3 Other Countries or Regions

In addition to the United States, wind resource surveys have been conducted for many other countries and regions of the world by various organizations. For example, the comprehensive "European Wind Atlas" published in 1989 describes the wind resources in the European Community countries. More detailed wind-mapping studies, including high-resolution wind resource maps, have been completed or are being planned for many areas of the world.

3.4 Offshore Areas

Offshore development in areas of high resource has the potential to greatly expand the use of wind

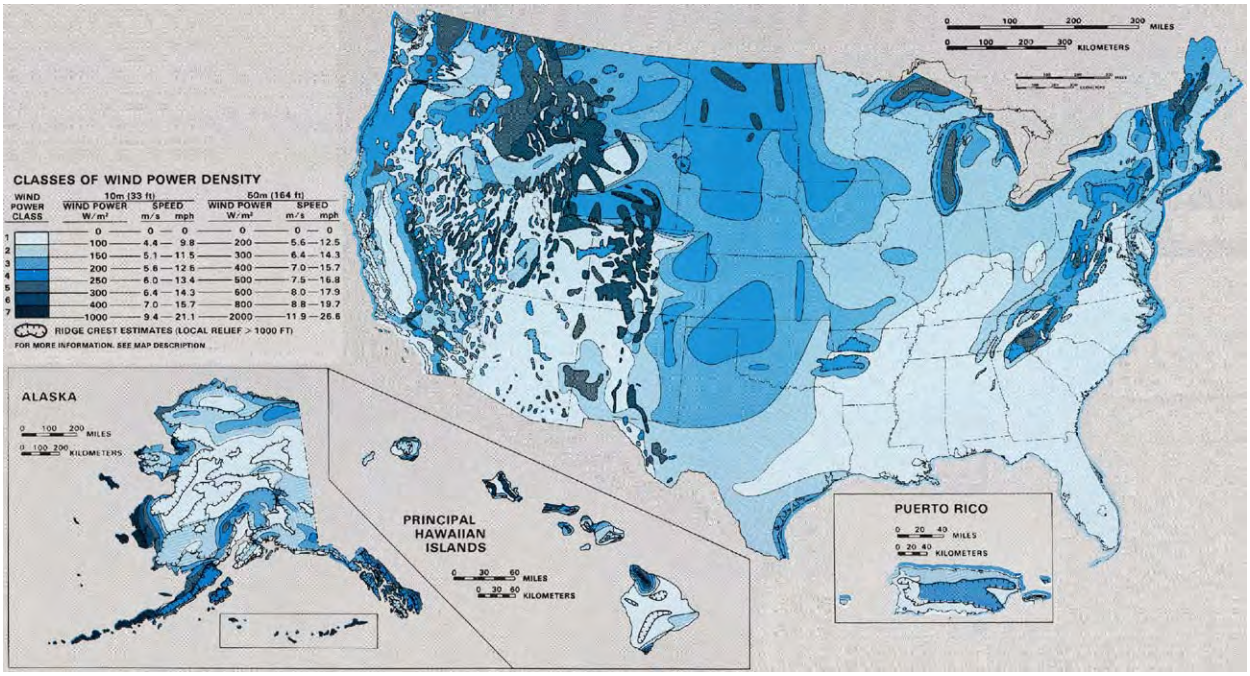


FIGURE 4 United States wind resource map (1987).

power, as seen in western Europe. Offshore wind resource assessments have been completed for areas of Europe and the United States.

4. DATA SETS AND ANALYSIS METHODS

An accurate wind resource assessment is highly dependent on the quantity and quality of the meteorological data input into a wind mapping system. Successful wind assessment methodologies used by the National Renewable Energy Laboratory (NREL) and other organizations rely on two major activities: (1) acquisition and development of comprehensive global data sets and (2) critical meteorological analysis of these data.

Advances in the available global data sets and analytical tools permit modeling and prediction of the wind resource without having to rely on surface wind data. This is important because in many regions of the world, high-quality surface wind data are sparse and often not available for areas of interest. The use of upper-air (weather balloon) and satellite-derived ocean wind data in an advanced computer mapping system enables NREL and other organizations to produce wind resource maps with valuable information even if high-quality surface wind data are not available. Where available, reliable surface wind data can be used to verify the model estimates.

A broad variety of wind data sets (including surface observations, marine data, and upper-air data) should be analyzed as the initial step toward understanding the wind climate in a study area. The use of multiple data sets is necessary because the quality of data in any particular data set can vary and because high-quality data can be quite sparse in many regions of the world. Some of these major global data sets are described here. Local or regional meteorological data sets should also be evaluated. Local data sets may be provided by independent organizations, individual consultants, or governmental departments.

4.1 Surface Data

High-quality surface wind data from exposed stations can give a good indication of the magnitude and distribution of the wind resource in an area. The U.S. National Climatic Data Center (NCDC) is a source for many global surface meteorological data sets. NREL has used is the DATSAV2 Surface Climatic Database in much of its assessment work. DATSAV2 contains hourly surface weather observations from more than 20,000 meteorological stations throughout the world. NREL has about 30 years of global DATSAV2 data in its archive that span the period from 1973 to 2002. The NCDC is developing new global surface data sets that will be used in future assessments.

Many meteorological stations are located at airports or in cities rather than at the potential wind energy sites. The wind climate may be radically different a few kilometers away from these meteorological sites. A drawback to airport or city observations is that the anemometer heights are usually between 6 m and 10 m above ground and must be extrapolated to heights that are useful for wind resource assessment.

Marine wind data can be used to estimate the wind resource for offshore areas. These data can also be extrapolated to estimate wind resource at coastal and inland sites that are well exposed to the ocean winds. NREL uses two global marine data sets. The first is the Marine Climatic Atlas of the World, derived from historical ship observations and summarized on a 1-degree grid. The other data set that has proved to be quite valuable contains ocean wind speeds at 10-m height derived from the special sensor microwave imager (SSMI) measurements taken as part of the Defense Meteorological Satellite Program. NREL has SSMI data covering the period from 1988 to 2001 and is examining some other satellite ocean data sets. One advantage of the SSMI data over the ship data is the much more uniform coverage—ship data is concentrated in the primary shipping lanes.

4.2 Upper-Air Data

Upper-air data can be useful in estimating the wind resource at low levels just above the surface and vertical profiles of wind resource aloft for extrapolation to elevated terrain features. An important data set for wind resource analysis is the global upper-air observational ADP data from the National Center for Atmospheric Research (NCAR) archive of the U.S. National Centers for Environmental Prediction (NCEP). This data set contains upper-air observations from rawinsonde instruments and pilot balloons for approximately 1800 stations worldwide beginning in 1973. The ADP data is supplemented in areas without adequate upper-air data by two numerical weather model-derived climatological

upper-air data sets. One is the Global Gridded Upper-Air Statistics data set obtained from NCDC. This data set contains monthly means and standard deviations of climatic elements from 1980 to 1991 for 15 atmospheric levels on a 2.5-degree grid. The other model-derived upper-air data set is the Reanalysis data set created by NCEP and NCAR. There are more than 40 years (1958 to 1997) of global Reanalysis data, which contain wind, temperature, and other variables on a 208-km grid. Reanalysis data are available in two different forms: constant pressure-level data and Sigma-level (terrain-following) data. NREL uses the Sigma-level Reanalysis data for wind resource analysis because they have good vertical resolution in the lower boundary layer where the wind climate characteristics have a large influence on the available wind energy potential.

4.3 Data Analysis Methods

This section describes the steps that NREL uses to analyze meteorological data. NREL processes the raw data into useful summaries to perform detailed analysis of the important wind characteristics. Meteorological parameters such as wind speed, wind power density, wind direction, temperature, and pressure are extracted from the hourly observations and used to create statistical summaries of wind characteristics.

Statistical summaries of the surface station data include interannual, monthly, and diurnal variabilities of the wind speed and power, and the wind speed and frequency by direction. The upper-air ADP data are processed to construct approximate vertical profiles of the wind resource by geopotential height and other wind characteristic summaries similar to those processed for surface data in order to evaluate the change of wind resource with elevation. NREL processes the SSMI satellite ocean wind data on a 25 km grid to evaluate wind resource in the vicinity of islands and coastal areas.

A critical analysis of the climatic data is performed to evaluate the quality of the data and determine the most representative data for a particular area. For example, the interannual data are evaluated to identify obvious trends in the data or periods of questionable data. Figure 5 shows an example of the "disappearing wind syndrome" that occurs at meteorological stations throughout much of the world. The data in the figure are from 79 meteorological stations in southeast China (examined as part of a detailed wind resource assessment) and show the ratio of wind power in each year relative to that in 1974, the first year of data. The downward trend in the data indicates that the apparent wind power in southeast China decreased by 60% in 20 years. There is little chance that this decrease actually reflects a gradual change in the wind climate caused by physical processes. It is more likely that the reduction in measured wind speeds is caused by new construction or vegetation growth around the meteorological stations and possibly some degradation of the measurement equipment.

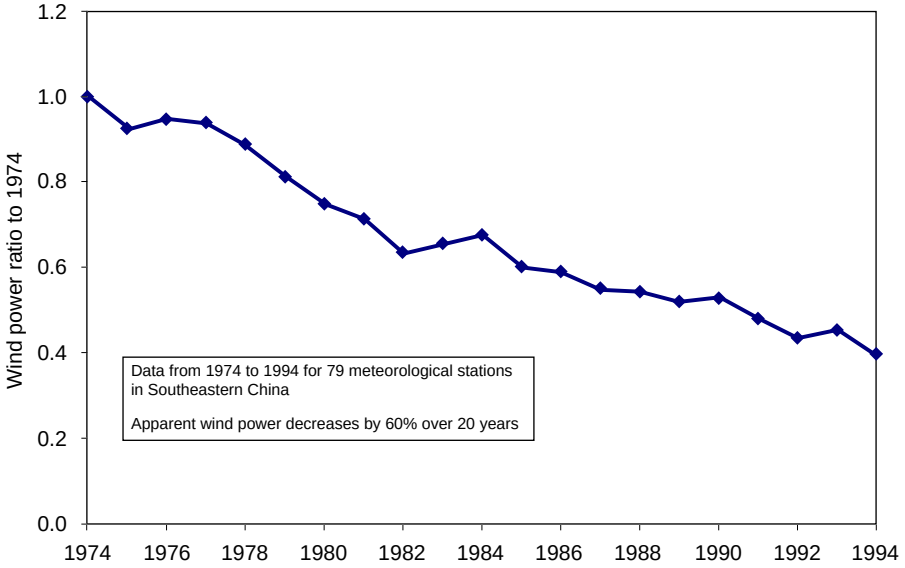


FIGURE 5 Graph of disappearing winds at southeast China stations.

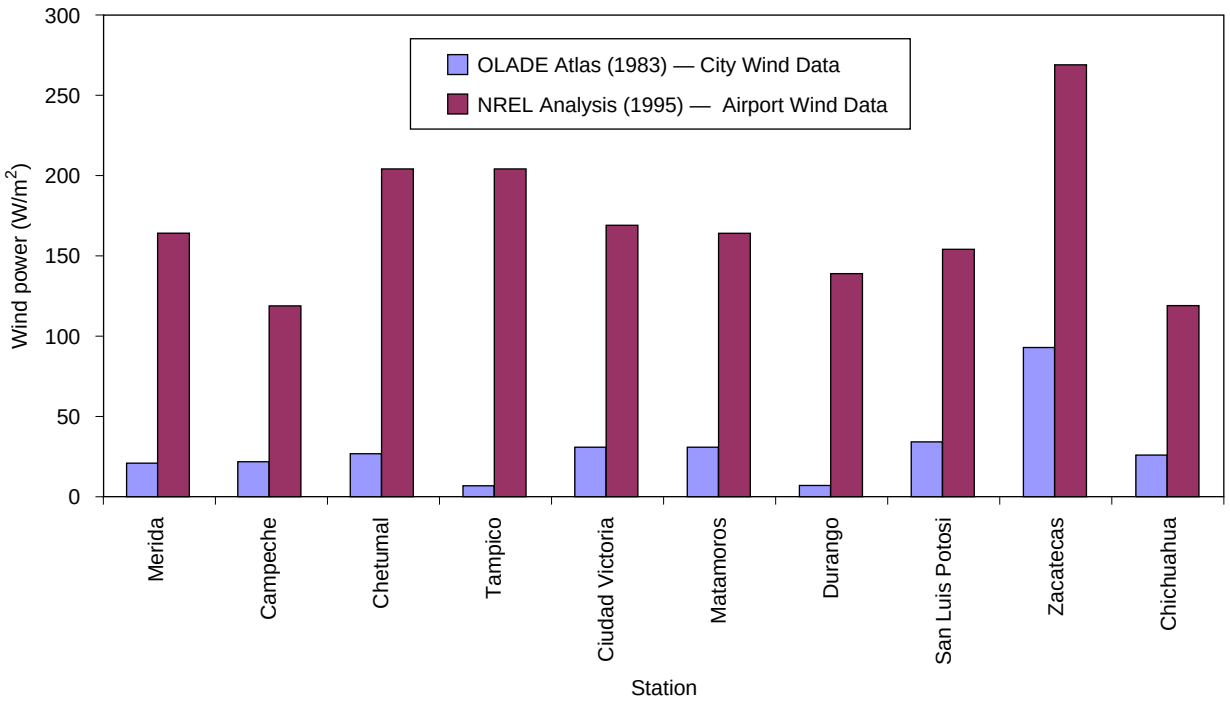


FIGURE 6 Comparison of wind speeds from Mexican city and airport stations.

Data from any nearby stations are compared to help assess the data quality. For example, Fig. 6 shows a comparison of wind power estimates from city and airport locations in Mexico. Wind data from airports are, in general, more reliable than data collected at other locations in urban areas where the wind measurements are likely to be reduced by buildings and trees. A small reduction of just 1 m/s in the measured wind speed can result in a large reduction (such as 30 to 50%) in the wind power density.

5. RESOURCE ASSESSMENT METHODS

This section discusses the NREL methodology, plus other current wind resource assessment methods. Wind resource assessment for wind energy development is usually conducted in three stages of increasing detail. The first stage, wind prospecting, identifies candidate regions for further investigation. Tools used in this stage are numerical and empirical wind resource models as well as previous wind surveys. The next stage, area wind resource evaluation, uses wind measurement systems and modeling tools to identify the best potential areas for more detailed evaluation. Micrositing is the final stage where locations of turbine sites are determined by a combination of measurements and fine-scale modeling.

5.1 General Wind Mapping

Regional wind resource maps are very useful tools in the wind prospecting stage of a project. Such maps cover an area of a few hundred to around a thousand kilometers on a side and show estimates of the wind power class on a grid. In the 1987 U.S. wind resource atlas, the grid size was about 25 km by 25 km, but the grid resolution of many maps is 1 km or finer.

5.1.1 Empirical Models

A computer mapping system developed by NREL in the 1990s uses an empirical and analytical approach to determine the wind resource value for a particular location. This wind mapping system does not use explicit atmospheric boundary layer equations or geostrophic adjustment equations like some other wind-flow models. It is designed to display large areas of detailed wind resource distributions at a spatial resolution of 1 km^2. These high-resolution wind maps help identify areas where wind energy projects are likely to be economically feasible.

Underlying the empirical technique is the major meteorological assumption that there are empirical relationships among the free-air (higher than 100 to

200 m above the ground) wind speed, the wind speed over the ocean (where applicable), and the distribution of the wind resource over the land areas. Empirical relationships have been determined in previous NREL wind resource assessment work for well-exposed (i.e., unobstructed by trees or buildings, local terrain features) locations in diverse areas of the world. Accordingly, the wind resource values presented on the NREL wind maps are estimates for nonsheltered locations with low roughness. NREL uses a top-down method for adjusting much of the available wind data. This approach takes the free-air wind speed and power profile in the lowest few hundred meters above the surface and adjusts these down to near wind turbine hub height (most often 50 m above ground). In this way, a wind resource map with valuable information can be produced even if high-quality surface wind data are not available.

The quality of the final wind resource map depends on the quality of the meteorological data input into the wind mapping system. It is important to review and analyze the climatic data sets to select the highest quality data to develop the conceptual model and meteorological inputs to the mapping system. The data analysis and screening process is designed to allow an analyst to develop a conceptual model of the physical mechanisms that influence the wind flow in the study region. The conceptual model guides the development of the empirical relationships that serve as the basis of the algorithms that calculate the wind power values seen on the maps.

The mapping system combines meteorological input, wind power calculation modules, and digital terrain data to produce the final wind resource maps. The global terrain data have a resolution of 1 km^2, and higher resolution terrain data sets are available in some areas. The meteorological inputs are wind power roses, vertical profiles of wind power, and open ocean wind power where appropriate. The meteorological data are initially used in determining a base wind power value for each grid cell. The mapping system calculates the final wind power density by adjusting the base wind power density. The factors that increase or decrease the base wind power density include terrain considerations, relative and absolute elevation, aspect (slope of terrain relative to the prevailing wind direction), distance from ocean or lake shoreline, and the influence of thermal or other types of small-scale wind-flow patterns.

The final wind resource maps are generated using commercially available geographic information system (GIS) software packages and custom GIS-based tools developed at NREL. These complex packages feature a large number of routines for scientific analysis along with tools used to display spatial information.

5.1.2 Numerical Models

The numerical model wind mapping system combines two types of computer models. The first type is a scaled-down version of a numerical weather prediction model. These prediction models solve the equations of motion and explicitly account for boundary layer processes such as energy balance and fluxes, surface moisture and reflectivity, and local temperature differences. The internal physics of these models are the basis for predicting the evolution of the atmosphere in the study area for specified times from the start of a model simulation. The scaled-down versions contain only those algorithms that account for physical processes that affect the wind flow. Scaled-down versions are used to reduce the time and cost it takes to complete all the simulations necessary to produce a representative wind map. Climatological data is used as the initial condition for each simulation as well as lateral boundary conditions for the model. There are a variety of methods that can be used to prepare the climatological data for input. Each simulation generates wind and other meteorological variables throughout the model domain, for a particular day, and the information is stored by set intervals.

In addition to climatological data, there are geophysical inputs into the numerical model. The geophysical inputs include elevation, soil moisture, and land cover data. The models translate these data into important surface parameters such as energy fluxes and surface roughness.

The numerical models are generally run with horizontal resolutions from 1 to 5 km, depending on the particular study region. There is no physical limitation to the horizontal resolution used by these models, although higher resolution requires longer computing time.

Often, a mass-conserving wind-flow model is used in conjunction with the numerical prediction model to produce wind resource estimates for horizontal resolutions lower than 1 km^2. The wind-flow models take the results of the numerical models as input and use surface roughness and terrain to estimate the wind resource for the smaller grid cells. Though the wind-flow models are not as sophisticated as the numerical models, they are much less expensive to use for the higher resolution estimates and the results are similar to what the numerical models produce.

5.1.3 Model Validation

The results of any wind resource model must be validated in order to have confidence in its estimates. Validation is the process of comparing selected wind resource estimates from the model with actual measured data from the same location. By comparing actual measured data from many sites with the model output, a measure of prediction quality can be derived. If the quality is not as good as desired, the model can be refined and if necessary rerun, generating a new set of estimates that can be compared to reality. A good wind resource model will estimate values at most validation points that are within 10% in wind speed and 20% in wind power density and will not show a consistent bias toward either over- or under-prediction.

5.2 NREL's Products

5.2.1 Wind Power Classification

The values on the wind resource maps produced by NREL are based on the wind power density, not wind speed. Wind power density is a better indicator of the available resource because it incorporates the combined effects of the wind speed frequency distribution, the dependence of the wind power on air density, and the cube of the wind speed.

To simplify the display and comparison of wind power, a seven-level wind power classification scheme was devised. The choice of levels was made solely for convenience in categorizing regions of similar wind resource and is not indicative of any physical or meteorological boundaries.

Table I shows the wind power classifications for utility-scale applications. Wind resource areas of Class 4 and higher are considered suitable for utility-scale wind power development with 2003 technology and energy costs. Rural or off-grid applications generally require less wind resource to be viable. Class 2 and higher resources may be sufficient for these applications.

TABLE I

Wind Power Classification

Class	Resource potential (utility-scale)	Wind power density (W/m^2) @ 50 m agl	Wind speed[a] (m/s) @ 50 m agl
1	Poor	0–200	0.0–5.6
2	Marginal	200–300	5.6–6.4
3	Fair	300–400	6.4–7.0
4	Good	400–500	7.0–7.5
5	Excellent	500–600	7.5–8.0
6	Excellent	600–800	8.0–8.8
7	Excellent	>800	>8.8

[a]Mean wind speed is estimated assuming a sea level elevation and a Weibull distribution of wind speeds with a shape factor (k) of 2.0. The actual mean wind speed may differ from these estimated values by as much as 20%, depending on the actual wind speed distribution (or Weibull k value) and elevation above sea level.

5.2.2 Wind Resource Maps

The primary product of NREL's mapping system is a wind power density map in units of W/m^2 for a specified height above ground, typically 30 or 50 m. Figure 7 is a wind resource map for the central coast of Fujian Province in southeastern China. The region's geographical features include an indented shoreline with hills along the coast, a narrow coastal plain in some areas, and mountainous terrain within 50 km of the coast. The best resources are found on exposed coastal areas and higher inland ridges. The maps developed for southeastern China are helping wind energy developers concentrate their efforts on the areas with the greatest resource.

Figure 8 compares the old wind resource map of North and South Dakota from the 1987 U.S. wind atlas to the new high-resolution (1-km) map generated by NREL's wind mapping system in 2000. The most striking difference between the two maps is the precision of the wind resource pattern on the newer map. The new map clearly depicts the variation in the wind resource that results from subtle terrain features. In fact, the considerable variation of the wind resource across the seemingly flat terrain of this region was one of the primary results of this project. The older resource estimates were generated using a manual interpretation of terrain maps and meteorological data. The new resource estimates were generated using GIS software, high-resolution terrain data, and more sources of meteorological input data. The end result is a resource map that depicts the potential wind resource in more detail and with better accuracy. The new map was validated using data from seven 40-m towers installed by utilities. At six of the seven sites, the wind power estimates were within 20% of the measured WPD values. This level of accuracy (>80% of validation sites within 20% of measured power) is typical of the updated maps that NREL has produced.

To make the wind maps more useful, other information, such as transmission lines and voltages, federal lands, exclusion zones, or Indian reservations, is often added. In addition to maps of the wind resource, the GIS system can also produce other types of maps that can be used in combination with

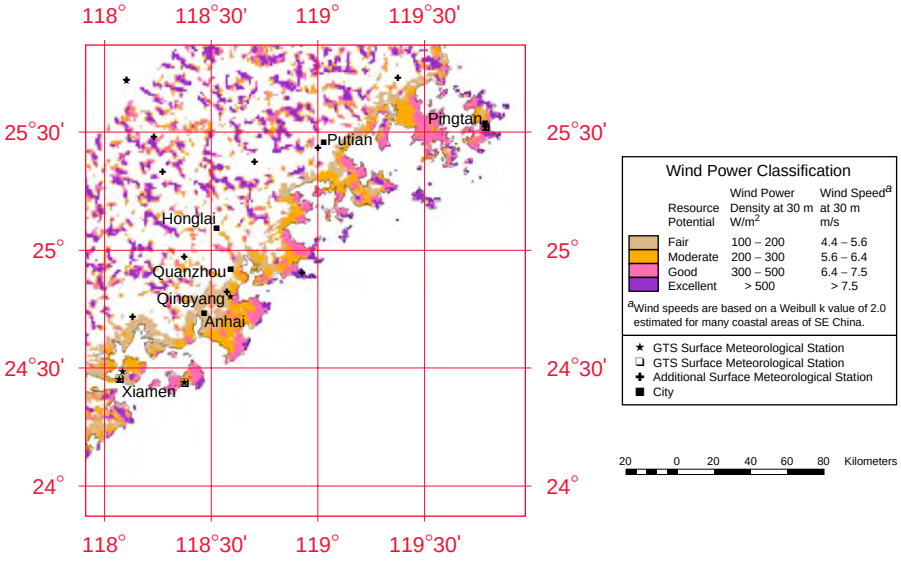


FIGURE 7 China coastal region wind resource map (2002).

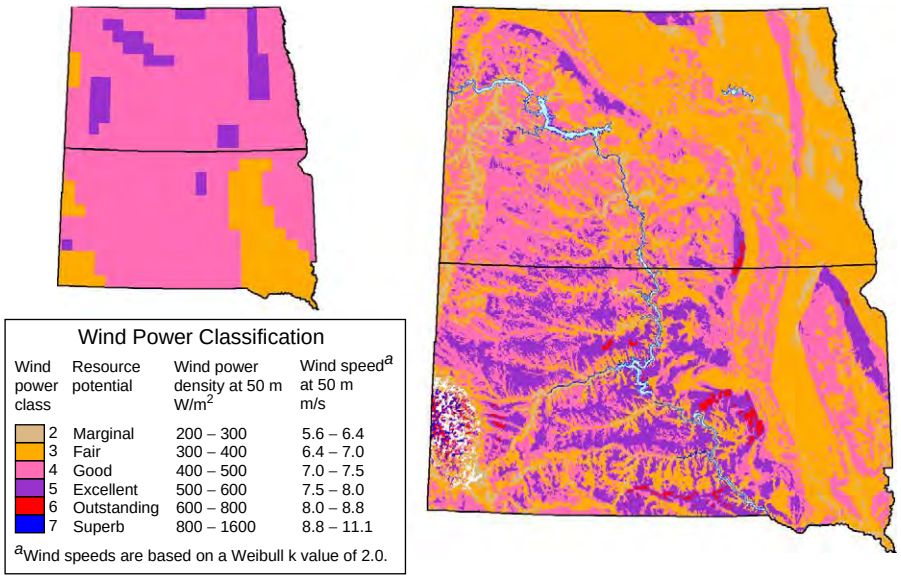


FIGURE 8 North and South Dakota 1987 (left) and 2000 (right) wind resource maps.

the wind resource maps, such as maps of color-coded elevation, hill-shaded relief, political features, meteorological stations, and other measurement sites. An example of a final validated wind resource map for Wyoming that includes transmission lines and Indian reservations is shown in Fig. 9.

5.3 Wind Electric Potential

Several factors determine the amount of land area suitable for wind energy development within a particular grid cell in a region of high wind energy potential. The important factors in NREL's assessment of wind electric potential are land-use and environmental restrictions for each grid cell.

Not all areas identified in a basic wind resource assessment will be suitable for wind energy development. Areas that may be excluded from development include the following:

- Developed or urban areas
- Parks and wilderness
- Wildlife habitat

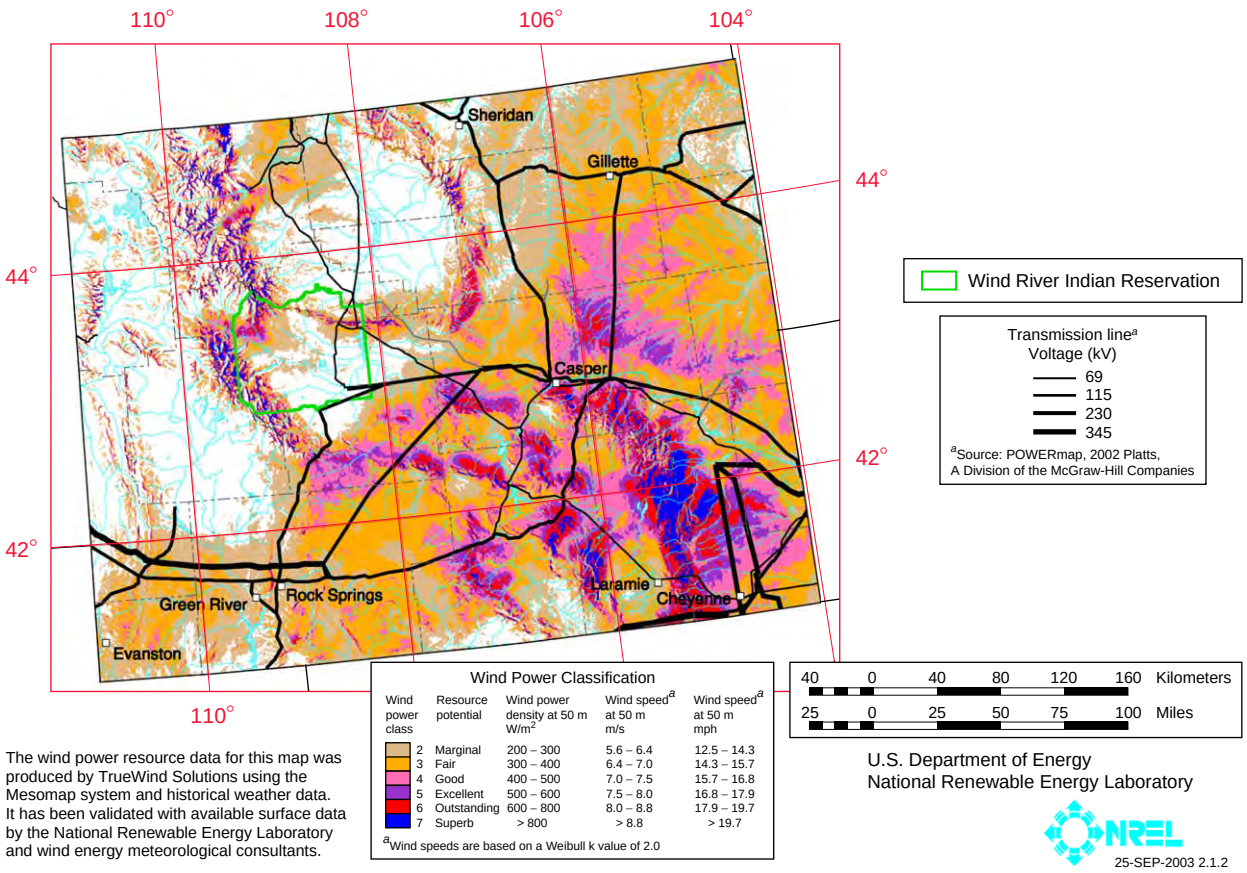


FIGURE 9 Wyoming wind resource map (2002).

- Wetlands
- Areas of steep (>20%) terrain and other topographic exclusions
- Areas too far from the electric grid

These areas are readily identified within a GIS software package, which provides spatial analytical capabilities. The GIS allows a more precise estimate of wind resource potential by directly overlaying the exclusions with the wind resource. The spatial databases that describe the exclusions are easily updated and reapplied to accommodate changing policy perspectives and new data.

The wind electric potential per grid cell is calculated from the available windy land and the wind power classification assigned to each cell. The amount of potential electricity that can be generated depends on several factors, including the estimated wind resource, spacing between wind turbines, the efficiency of the machines, and the estimated energy losses caused by factors such as wind turbine wakes and blade soiling. A conservative estimate of potential installed capacity can be obtained by using a factor of 5 MW/km^2 of available land. Values can be substantially higher depending on the wind climate and terrain, but those factors are generally evaluated at the micrositing level.

5.4 Area Wind Resource Evaluation

5.4.1 Wind Monitoring Programs

A properly-run wind monitoring program should follow the identification of prospective wind energy development areas. Measurement towers installed at various locations throughout the area can confirm the estimates made by the models, help locate turbine installation sites, and provide estimates of the final wind energy production of the project.

5.4.2 Equipment

The minimum configuration for a wind monitoring site consists of a single anemometer and wind vane mounted on a tower. The minimum tower height should be 10 to 20 m. If trees or other obstructions exist near the tower, the tower height should be increased. A data logger located at the bottom of the

tower records the wind data at 10-minute or 1-hour intervals. If funds permit, a taller tower (40 or 50 m in height) can be installed, and other instrumentation, such as redundant anemometers, multiple measurement levels, and other sensors (e.g., temperature, pressure, solar radiation), can be added. Taller towers are preferred in order to obtain measurements at or near the hub heights of modern wind turbines.

5.4.3 Instrument Siting

The tower and instruments must be installed at locations that are representative of the resource in the area while minimizing the effects of any nearby obstructions (e.g., trees, buildings). Proper siting of a tower to ensure the highest quality data collection requires considerable skill, and the use of experienced consultants is recommended.

5.4.4 Time Period

Each site must record data for a minimum of 12 months to get a complete picture of the seasonal variation. Although 18 months to 2 years of data will give better confidence in the measurements, the time period must be balanced with other project constraints.

5.4.5 Data Analysis

The first step in data analysis is quality control of the data. Erroneous data can be caused by factors such as anemometer icing, equipment failure, and operator error. Bad data must be identified and removed. The data can then be analyzed and statistics computed on both monthly and annual bases. There are many statistical quantities that could be computed, with some of the most useful being:

- Annual average wind speed, wind power density, and power class
- Speed and power by month (seasonal)
- Speed and power by hour (diurnal)
- Frequency and speed by direction (both annual and monthly)
- Wind speed frequency distribution
- Wind shear (if using more than one measurement level)

Tabular and graphical displays of these statistics aid in the analysis.

5.5 Micrositing

The final stage of project development is micrositing, where the precise locations of wind turbine installation sites are determined. Special microscale modeling software produced by various companies and organizations can greatly aid this process.

6. STATUS OF WIND RESOURCE PROJECTS

High-resolution wind resource maps and comprehensive wind resource atlases have been produced for various countries or areas of the world. For example, NREL and its collaborators have performed wind resource assessments and generated high-resolution wind maps of Armenia, the Dominican Republic, the Maldives, Mongolia, the Philippines, and Sri Lanka. Assessments were also done for specific areas of Chile, China, Indonesia, Mexico, Russia, and the United States. As suggested on the cover of the Mongolia wind atlas shown in Fig. 10, the wind resource information is intended to be useful for both small wind turbine and utility-scale applications. Many of NREL's international wind resource maps, and some produced by others, can be found at www.rsvp.nrel.gov. High-resolution wind resource maps for additional countries have been produced by organizations such as TrueWind Solutions, Risø National Laboratory, and Garrad-Hassan.

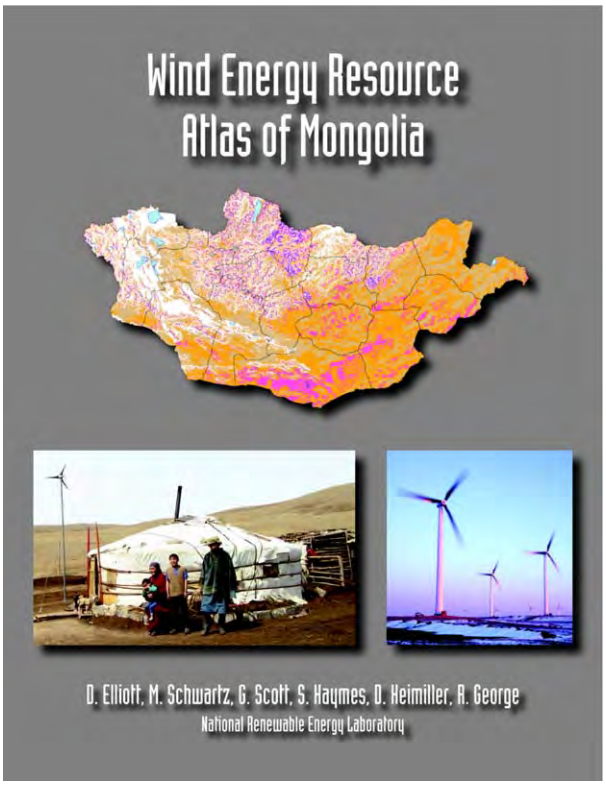


FIGURE 10 Mongolia wind resource atlas cover (2001).

Detailed wind maps and other resource data have facilitated the rapid identification of good wind resource areas and led to more successful prospecting and measurement efforts. This new wind resource information has accelerated the use of both small wind and utility-scale wind energy technologies. As government and industry groups seek information to more accurately assess the wind energy potential and identify promising locations, the interest and demand for detailed high-resolution wind resource maps and other relevant data are intensifying.

A project funded by the United Nations Environment Programme (UNEP) is expanding wind mapping to many additional areas of the world. In addition to wind resource maps, the UNEP Solar and Wind Energy Resource Assessment (SWERA) project will include the development of solar resource maps and a GIS interface. In the first phase of this project, high-resolution wind maps are being developed for six countries in Latin America (Cuba, Guatemala, El Salvador, Honduras, Nicaragua, and Belize), three countries in Africa (Ghana, Ethiopia, and Kenya), and three countries in Asia (Nepal, Bangladesh, and specific regions of China).

In the United States, efforts such as U.S. Department of Energy's (DOE's) Wind Powering America program are stimulating interest in wind energy and the development of updated wind resource information in the form of maps and data. High-resolution wind resource maps have been completed for many states, and projects are under way or planned for many additional states. The primary approach is a public/private cooperation between NREL and U.S. companies to produce and validate the wind resource maps. These new maps are instrumental in helping state and local officials, project developers, utilities, and landowners identify suitable wind sites for more detailed evaluation.

SEE ALSO THE FOLLOWING ARTICLES

Hydropower Resources • Land Requirements of Energy Systems • Wind Energy Economics • Wind Energy, History of • Wind Energy Technology, Environmental Impacts of • Wind Farms

Further Reading

American Wind Energy Association, www. awea. org.

American Wind Energy Association (AWEA) (1993). "Recommended Practice for the Siting of Wind Energy Conversion Systems," AWEA Standard 8.2.

AWS Scientific (1997). "Wind Resource Assessment Handbook," NREL/SR-440–22223, National Renewable Energy Laboratory, Golden, CO, www.nrel.gov.

Brower, M., Bailey, B., and Zack, J. (2001). "Applications and Validations of the MesoMap Wind Mapping System in Different Climatic Regimes." Windpower 2001 Proceedings, American Wind Energy Association, Washington, DC.

Cherry, N. J., Elliott, D. L., and Aspliden, C. I. (1981). "Worldwide Wind Resource Assessment." *In* Proceedings of Fifth Bienneal Wind Energy Conference and Workshop (I. E. Vas, Ed.), Vol. 2, pp. 637–648. Seri/CP-635–1340, CONF-811043, Vol. 2, National Technical Information Service, Springfield, VA.

Elliott, D. (2002). Assessing the world's wind resources. *IEEE Power Engineering Rev.* **22**(9), ISSN 0272–1724.

Elliott, D., Holladay, C., Barchet, W., Foote, H., and Sandusky, W. (1987). "Wind Energy Resource Atlas of the United States," DOE/CH 10093–4, National Renewable Energy Laboratory, formerly Solar Energy Research Institute, Golden, CO, http://rredc.nrel.gov.

Elliott, D., and Schwartz, M. (2002). "Validation of New Wind Resource Maps," WindPower 2002 Proceedings, American Wind Energy Association, Washington, DC. NREL/CP-500–32536, National Renewable Energy Laboratory, Golden, CO.

Elliott, D., Wendell, L., and Gower, G. (1991). "An Assessment of the Available Windy Land Area and Wind Energy Potential in the Contiguous United States," Report No. PNL-7789, National Technical Information Service.

European Wind Energy Association (EWEA), www.ewea.org.

Frank, H., Landsberg, L., Rathmann, O., Mortensen, N., and Petersen, E. (2001). "The Numerical Wind Atlas–The KAMM/WA^{S}P Method." Windpower 2001 Proceedings, American Wind Energy Association, Washington, DC.

Frost, W., and Aspliden, C. (1994). Characteristics of the Wind. *In* "Wind Turbine Technology," Chapter 8 (D. A. Spera, Ed.), American Society of Mechanical Engineers.

Heimiller, D. M., and Haymes, S. R. (2001). "Geographic Information Systems in Support of Wind Energy Activities at NREL," 39th AIAA Aerospace Sciences Meeting Proceedings, Reno, NV. NREL/CP-500–29164, National Renewable Energy Laboratory, Golden, CO, www.nrel.gov.

National Wind Technology Center, National Renewable Energy Laboratory, www.nrel.gov.

Pennell, W. (1983). "Siting Guidelines for Utility Application of Wind Turbines," Report No. EPRI AP-2795, Electric Power Research Institute, Palo Alto, CA.

Risø National Laboratory, www.risoe.dk.

Rohatgi, J., and Nelson, V. (1994). "Wind Characteristics: An Analysis for the Generation of Wind Power." Alternative Energy Institute, West Texas A&M University, Canyon, TX.

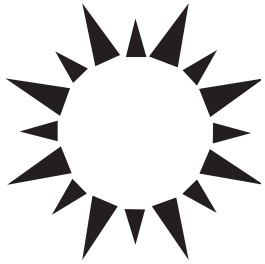
Schwartz, M., and Elliott, D. (2002). "Remapping of the Wind Energy Resource in the Midwestern United States, NREL/AB-500–31083." National Renewable Energy Laboratory, Golden, CO.

Troen, I., and Petersen, E. L. (1989). "European Wind Atlas." Commission of the European Communities, Risø National Laboratory, Roskilde, Denmark.

U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy, www.eere.energy.gov.

World Meteorological Organization (1981). "Meteorological Aspects of the Utilization of Wind as an Energy Source, Technical Note No. 175." Geneva, Switzerland.

Women and Energy: Issues in Developing Nations



NJERI WAMUKONYA
United Nations Environment Program Collaborating Center on Energy and Environment
Roskilde, Denmark

1. Introduction
2. Historical Context
3. A Women's or a Poverty Problem?
4. Do Stove Programs Address Women's Energy Problems?
5. Household Energy and Health: A Women's Problem?
6. Will Improved Stoves Improve the Health Situation?
7. Does Increasing the Supplies of Biomass Fuel Help?
8. Will Electricity Help Women?
9. Will Saving Time Help Women to Engage in More Productive Activities?
10. Do Women Really Have Energy Needs Different from Those of Men?
11. Energy Decisions: Are They All Made by Men?
12. Will Access to Energy Increase Women's Income-Earning Possibilities?
13. Conclusions: Areas for Further Research

Glossary

contextualization Taking the prevailing political, social, and economic environment into account in analyzing situations and making recommendations.
decentralized systems Typically electricity-generating facilities that are isolated and not connected to a centralized grid; the power can be generated from any fuel source (e.g., diesel).
energy service The activity that can be derived from injecting energy (e.g., grain milling).
gender A socially constructed relationship between men and women.
informal Not in the conventional macroeconomic framework.

Women and energy in the developing country context is a theme that is being advanced by some energy analysts and practitioners as a response to advocacy to enhance gender equity. The link between women and energy has been mainly within the household framework and in the rural context. Women have culturally been largely responsible for cooking and, hence, providing energy for this purpose. Biomass fuels are the main source of energy in rural areas, and its procurement involves drudgery. The fact that this drudgery has been experienced by women justifies the case for a focus on women and energy. The extent to which this approach has contributed to gender equity and poverty alleviation remains to be established.

1. INTRODUCTION

Inequity along gender lines has been one of the main factors driving the establishment of women-focused programs and calls for gender mainstreaming. The issues of "women and energy" and "gender and energy" have been on the agenda since the 1970s and 1980s. An important response to women and energy issues is the allocation of budgets to women and energy programs by donor agencies, necessitating suitable responses in the form of project proposals and projects. In development circles, a call has been made for the inclusion of women in energy projects. Specially targeted projects are also common. In such interventions, women are identified as recipients of project and program benefits.

The suffering of women and children in fuelwood procurement has been the main rationale for women-focused energy initiatives. The resultant initiatives have been in the form of projects and policy suggestions. Projects include promoting improved stoves and enabling the adoption of solar home systems through women-targeting financing

 © 2004 Elsevier Inc. All rights reserved.

mechanisms. Activities have also been undertaken to help incorporate women into various energy policies, programs, and projects at national, regional, and international levels.

Many initiatives are responses to the documentation on links between women and energy—links ranging from physical burdens to health problems. Most of this literature has been based on ideological concerns where women are seen as the weaker sex and so naturally disadvantaged in energy matters. The perception of women as victims of men shapes the analyses of their roles in energy procurement and use. As a result, the root causes of the problems relating to energy tend to be downplayed and rarely explored sufficiently in the literature on women and energy. Most of the literature suffers from inadequate empirical investigation. In many cases, evidence is presented in a manner that seems biased toward ideology rather than facts; old data and data from a restricted number of studies are recycled with the aim of demonstrating the women/energy-linked problems and justifying special attention to women. In this way, the myths are sustained.

The consequence of the formulation of women and energy issues has been focused largely on the provision of energy as an isolated commodity. But as often is indicated, energy is not an end in its own right; rather, the service from energy is what is important. The extent to which the women and energy approaches have achieved the intended goals of alleviating women's energy-related problems remains limited. It is likely that this could be attributed to the conception of the problem. This article explores the literature with the aim of determining the realities of the issues as perceived and attempting to chart a way forward.

The women and energy approach has been predominantly a developing country phenomenon. Why have the issues not attained similar importance in the developed world? The answer lies in the far higher level of economic growth and access to modern energy carriers in developed countries. Energy services in the developing countries are survival–developmental issues, a stage the developed world has passed. As Martine and Villarreal highlighted in a 1997 study, the rural and traditional loci of gender issues by extension render the related energy problems gender irrelevant in the developed countries. This prompts the question of whether a gendered approach can address the real issues—a question that guides the discussion in the rest of this article. In a 1995 report, Parikh justified gender-related energy policies in developing countries on the basis that they are made necessary by the low incomes in these countries.

The rest of the article is structured as follows. First, it provides a historical context on the issue with the aim of endeavoring to understand why the women and energy issue is addressed in the manner it is. Using leading questions, the article then attempts to understand whether current initiatives can meet the objective of improving women's energy problems. Finally, the article concludes by identifying areas for further research.

2. HISTORICAL CONTEXT

The role of women, particularly rural women, in energy procurement began to receive attention after the energy crisis of the 1970s, mainly because escalating oil prices forced an increased shift toward biomass. At the same time, there was more attention being paid to the poor, emanating from the basic needs approach. The growing focus on rural development during the mid-1970s also drew attention to the dependence of rural people on the products from forests, woodlands, and trees, particularly on fuelwood. A 1975 publication noted the increased demand for woodfuel and the perception that demand outstripped supply and was responsible for deforestation, raising concern for what was referred to as "the other energy crisis" and necessitating better understanding of the supply-and-demand situations. This resulted in application of the "gap theory" by various authors, including O'Keefe and Raskin in 1985 and Anderson in 1986, identifying a gap between supply and demand and resulting in country forecasts that predicted critical shortages. Although the gap theory has been challenged extensively in various works, such as those by Bradley and Campbell in 1998, it nevertheless formed the basis for defining the woodfuel crisis. This definition focused on the biological yields of biomass, ignoring the socioeconomic, cultural, and historical factors that influenced the management of the resource. The crisis was also simplistically defined in terms of the time and labor required to gather fuelwood. Furthermore, conclusions for most regions and countries were based largely on extrapolating results from a few samples from a limited number of countries.

The women's dimension of the energy problem was highlighted through works done by the International Labor Organization (ILO) and others during the 1970s and 1980s, including those by Cecelski, Agarwal, and

Eckholm. The economic crisis caused by the oil price hikes had kick-started interest in the whole question of energy, and studies started to be conducted showing that in developing countries households were principal energy consumers and the bulk of this energy was for cooking and was woman based. For the first time, the dominant role of women in procuring this vital resource was recognized and highlighted as a major problem, both because a large amount of women's time appeared to be absorbed in energy-related activity and because of the drudgery of carrying heavy loads of firewood over long distances (up to 10 km was regarded as normal). Later, the role of children in gathering fuelwood was also recognized.

However, this women's dimension was contextualized largely in a rural situation, perhaps because the entry point had been the fuelwood energy crisis and a large number of the women resided in rural areas. In a 1995 report, Skutsch noted that donor programs had an implicit bias toward rural rather than urban household energy. Therefore, the early work captured mainly the women and energy issues within rural household economics and time budget/labor discourses. The time spent and the drudgery suffered by women and children in gathering fuelwood and other biomass fuels for household cooking needs were (and still are) highlighted as the main problems and rationale for intervention. In a 1975 report, Eckholm provided graphic descriptions of further hardships suffered by the fuel gatherers as a result of deforestation. Some works, such as a 2003 study by Laxmi and colleagues, also reported that the burden involved is so large that children are kept out of school to gather biomass fuels. The amount of time spent on gathering fuelwood and the distances covered are widely cited, although only a few empirical studies have actually been undertaken. In many articles, the time is cited without a reference point, for example, a 1997 article in which Nyoni referred to a study finding that 37.8% of respondents in Zimbabwe spent 1 to 2 h collecting firewood but did not indicate frequency.

As a result of these kinds of studies, the idea arose that in developing countries household energy is essentially a women's problem and, consequently, that women-centered approaches should be taken in dealing with the problem. What was happening in the energy field is also a reflection of the general trend in paying increasing attention to women. The works of Boserup during the 1970s, articulated in her book *Women's Role in Economic Development*, drew attention to the economic disparities between women and men, the lack of integration of women in the development programs in developing countries, and the role of technology in this matter. This initiated the establishment of Women in Development (WID) networks in the North, which later formed links with the South.

Within the WID discourse, women were the focal points and the number of women participating in a program was considered to be a key indicator of progress. WID also generally considered women as a homogenous group and hinged the argument of women as a special group to a welfare approach. Predominantly, during the 1970s, there was a feeling that women's welfare had been ignored in the past, so in the context of energy, stove programs were set up to reduce drudgery. WID emphasized support for women within their traditionally defined roles (e.g., as mothers, as household carers). From an energy perspective, household energy, particularly energy for cooking, became the main focus. As Skutsch noted in a 1995 study, WID aimed at improving women's welfare without making women self-sufficient. These shortcomings with the WID approach resulted in a divergent school of thought during the late 1980s: distinguishing sex from gender. The emerging concept was referred to largely as "gender and development" (GAD). GAD placed more emphasis on the participation of women in the whole process. In the energy context, this meant considering energy not just for welfare reasons but also as a means of empowerment (e.g., women as energy entrepreneurs, women as decision makers about energy). This also meant focusing not just on stoves but also on types of energy that would "empower" women, giving them more chances to earn income, broaden their horizons, and so on.

Increasing global environmental concerns, motivated by Brundtland's report on *Our Common Future* in 1987, captured the attention of ecofeminists on the link between environment and women, resulting in a women, environment, and development (WED) discourse, making women the central focus of environmental concerns, and arguing that women have an affinity with the environment and are best placed to safeguard it. The WED discourse, which Jackson noted in a 1993 study is driven largely and sustained by nongovernmental organizations (NGOs) and oriented to radical ecofeminism, advances the theory that women naturally care for the environment and have profound knowledge of plants, animals, and ecological processes. In the energy context, women are seen as victims of environmental degradation and, as a result, suffer drudgery in gathering fuelwood. Another dimension that is increasingly gaining favor

is the perception of women as guardians of the environment, with women being expected to adopt renewable energy and decentralized energy technologies because these are considered to be environmentally friendly. In a 1999 report, Tisch noted that at the community level, women are effective and efficient managers of renewable energy systems and that, once these alternatives to ecologically damaging energy resources are made available, rural women use renewable energy technologies and benefit from the improved standards of living, enhanced economic opportunities, and cleaner and more sustainable environment they offer. Although the WED thinking is largely unproven, it seems to have had a substantial influence on women and energy approaches. In fact, the literature reveals that it is not uncommon to make conclusive and influential statements on links between women and energy without empirical evidence.

In general, women are seen as the procurers and managers of energy, particularly in rural areas. On the other hand, despite these important roles, women are also seen as lacking decision-making power on issues relating to the energy system. The contextualization of the women and energy issues and responses to them demand rethinking, particularly given ongoing reforms in the energy sector and calls for gender mainstreaming. It is important to evaluate the achievements of past approaches on women and energy issues and to establish their relevance. To this end, the following discussion aims to address some leading questions. Certain presuppositions are considered, including that the problems are fundamentally women's problems, that certain technologies will solve the problems, that women's needs are unique, and that increasing energy access will improve women's incomes.

3. A WOMEN'S OR A POVERTY PROBLEM?

Although rural women and girls may generally bear a larger burden of gathering biomass fuels than do men and boys, the underlying factors might not have been fully accommodated in the design of solutions. By socialization, women tend to be responsible for cooking and, by extension, for cooking fuel procurement. However, this might not necessarily imply that women do so because they are women; it may be due to many factors, including the financial status of households. Availability of fuelwood also influences women's involvement. The World Energy Council (WEC) noted in 2000 that, despite intense degradation of the biomass energy resource base in some areas, investments to improve kitchens, stoves, and cooking fuels either do not figure in the hierarchy of household expenditures or appear low on the list, mainly because the livelihoods of these households are also under pressure and this is a more urgent priority.

In well-off households, either fuelwood and other biomass fuels are gathered by paid help or the wood is bought and delivered to the house by motorized transport, mainly by men. In a study of rural India reported in 2002, the Energy Sector Management Assistance Program (ESMAP) noted that lower income groups spend more time collecting fuelwood mainly because they cannot afford to purchase it or access alternatives but also partly because the resources nearby are depleted to serve the commercial market created by higher income households whose members purchase fuelwood in large quantities. As reported by Miah and colleagues in 2003, rich rural households in Chittagong, Bangladesh, purchase 47% of their biomass fuel from the market and obtain the rest from homestead and agricultural fields. They noted that medium-category households purchase 36% of their needs, whereas the poor only collect biomass fuels from forests and fields. Furthermore, poor men and women sell their labor for agricultural activities and are partially paid with residues that they use as fuel.

Rural areas lack access to modern energy forms that are generally easier to procure elsewhere (e.g., in towns). However, where they are available in rural areas, the predominant limiting factor to their consumption in households is cost. A study of rural Rajasthan, India, reported by Laxmi and colleagues in 2003 found that as fuelwood became scarce, households switched to either commercial fuels or lower value fuels such as dung and straw, depending on income group and/or geographical area. In another Indian study published in 2002, ESMAP noted that although electricity has reached 90% of the villages in India, only 40% of rural residents have adopted it and most of the remaining households are poor. Depending on where a house is located with reference to a substation in a grid-connected area, costs of connection can be prohibitive. Decentralized technologies for household electrification are also not cheap, and in most cases only well-off households can access them. Women and energy advocates argue that special financing programs targeting women should be set up to provide women with

access to cheap loans for advanced technologies and modern energy procurement, particularly electricity and gas. Whether such measures address the real energy problems of women is an issue that is explored in later sections of the article. An important aspect that has not been well researched is the impact of such programs targeting different members of the household on the social structure. A 1998 report by the Food and Agricultural Organization (FAO) noted that targeting women can result in a backlash of resentment by men against what they see as an excessive emphasis on women in development programs. This confirms need for better understanding of the implications in the energy sector.

The focus on women as the providers of cooking energy may have shortcomings, and some important factors are sometimes missed. For example, in their work in rural Rajasthan, published in 2003, Laxmi and colleagues reported that the share of households purchasing fuelwood had increased to 14% from the 2% reported in 1996. This has two implications. The first is that concerns about women's time and drudgery connected with fuelwood collection might not be of primary importance. The second is that the availability of modern fuels would facilitate a shift from fuelwood given that there is clearly some financial resource available. In addition, various changes are noticeably taking place in the energy procurement process that need to be considered in designing solutions to women's energy problems. There is evidence that men play a more important role in fuelwood gathering than is sometimes reported. The ESMAP study found firewood collection by men in six states of rural India to be significantly higher than that by women, citing increased distances and effort as the reasons. Based on their work in rural Zimbabwe, Vermeulen and colleagues in 2000 reported similar trends, with men collecting firewood with carts. This increased involvement of men in firewood collection requires the adoption of a more comprehensive approach in addressing fuelwood shortage issues. The ability of men to meet their other household responsibilities and the possibility of women having to undertake these responsibilities should be investigated in the process of seeking sustainable solutions.

The importance of availability of modern fuels cannot be underestimated in influencing shifts from fuelwood. In a 2003 report, Gupta reported that the availability of liquefied petroleum gas (LPG) and biogas in an area where the improved Astra stove had been disseminated over a 7-year period resulted in a shift away from firewood in 35% of households. In their 2003 study, Laxmi and colleagues found that nonavailability of kerosene for cooking restricts its use more than does nonaffordability. By paying more attention to such trends that are unrelated to gender issues, strategies can be established to enable households to shift. Under such circumstances, a more open approach, rather than a narrow orientation toward women, might produce effective solutions.

The inability to purchase appliances and get connections certainly limits any switch to modern fuels. Various efforts have been made to enable women to acquire household energy appliances. In July 1999, the state government of Andra Pradesh, India, launched the Deepam Scheme, which pays the initial LPG filling fee for women who belong to self-help groups and whose households are classified as being below the poverty line. Beneficiaries have to meet the appliance and gas refill costs. Despite this free gas, most households continued to use wood, mainly because they could not afford refills. This problem is further illustrated by the relatively low consumption rate in rural areas, where the average refill rate was once in 7 months. Although time saving in cooking is reported as a major benefit of this project, there is no report on how the time saved is spent. The low consumption levels of LPG have meant a lack of entrepreneurial interest in LPG distribution service, so that households still have to travel relatively long distances to obtain gas. The report on the study does not indicate whether LPG procurement was done by women or men, and it is difficult to establish whether women's fuelwood collection time was indeed saved or merely shifted to LPG procurement.

Increased urbanization obviously has meant more women moving to urban areas. In many low-income urban neighborhoods, biomass remains an important cooking fuel. Although a part of it is gathered, the bulk of it is bought, often in informal markets near the households. As such, the time, labor, and drudgery associated with fuelwood procurement in rural areas are less important. Access to money is more so; here, poverty is the main determinant of access to biomass fuels. The fuelwood and charcoal are sold by both women and men who obtain it from the cities and transport it by road or rail over relatively long distances. Because this business does not necessarily require large capital investments for entry, competition from men is increasing as retrenchment levels rise with the restructuring of economies in several countries.

Low-income urban settlements tend to be informal. In most of urban Sub-Saharan Africa, informal

settlements and slums have long been the norm, increasing along with rural–urban migration. These settlements are characterized by temporary housing structures, poor sanitation, and lack of other basic services, including water and electricity. Much of the housing is classified as "illegal" and, hence, not likely to be serviced and connected to electricity. Residents sometimes resort to drastic and potentially dangerous measures to access electricity and energy services, including illegal connections—an activity implicitly attributed to men. A case reported in 2003 by Khupiso involving a woman in South Africa is useful to draw attention to the unconscious bias involved. The woman, who owned a Reconstruction and Development Program house in Mamelodi East, Pretoria, was on the run from police for illegally connecting 15 of her neighbors in an informal settlement. She had laid underground electricity cables from her house to these settlements and had been supplying them with electricity for 2 years at a flat rate of R150 (~U.S. $19) per month per connection. She also provided, for a fee, refrigeration services to some *spaza* (retail) shop owners. Of course, if viewed from a different standpoint, this could serve as an example of female entrepreneurship in the energy sector. Establishing ways in which to legalize such informal electricity trade could offer opportunities for women and men to generate income.

4. DO STOVE PROGRAMS ADDRESS WOMEN'S ENERGY PROBLEMS?

Mahatma Gandhi is credited with having inspired the production of the smokeless *chulha* in India by drawing attention to the cooking stove as a means of improving the living conditions of the rural woman. In general, the original interest in improved stoves was in improving the indoor environment rather than efficiency. The 1970s witnessed the beginning of the appropriate technology movement within which improved stoves were promoted as solutions to deforestation. To address the woodfuel crisis and the consequent problems for women, closing the woodfuel gap through demand-and-supply management strategies was seen as the logical solution. On the underlying assumption that woodfuel was likely to be the main fuel source for a long period of time, improving woodfuel efficiency has been a dominant measure. In rural areas, improved stove efficiency was (and still is) deemed a method of reducing fuel collection for women and girls. This is despite the experiences of the 1980s, where most of the improved stove projects failed.

As early as 1987, the ILO concluded that cooking stoves that are designed to reduce the quantities of fuelwood used can be expected to be successful (adopted easily by local women) only in areas where there is an acute fuel shortage (e.g., rural areas with an ecologically degraded environment, peri-urban and urban areas where fuel costs are high). It was recommended that stove programs should not be exclusively concerned with biomass fuels given that consumers prefer to have alternatives, and it was noted that fuels and technology are used depending on the season, ease of use, and availability. A 1985 study by Gill highlighted the reasons why improved stoves failed to achieve widespread adoption in developing countries.

Although it lost favor during the 1980s, interest in improved stoves seemingly was revived during the subsequent decade. In a 1994 review, the World Bank reaffirmed the value of cooking stove projects, claiming that the social, economic, and environmental benefits of promoting improved stoves under suitable circumstances are quite large, and existing successes demonstrate the usefulness of well-managed programs. Various programs and projects have been launched, mainly with support from donors, to promote improved biomass stoves. Women have always been regarded as a special target group to which donors, governments, and NGOs direct aid in various ways to facilitate adoption of the stoves. One of the reasons provided for the poor performance of initial projects was poor targeting; that is, women were excluded. The underlying rationale for supporting women-targeted programs continues to be that improved stoves would save fuelwood, so that women would spend less time and energy in gathering it. As reported by Khandelwal in 2000, the government of India has a national program on improved stoves that provides direct subsidies to women from select castes and tribes to enable them to adopt these stoves. By 2000, the program had been sustained for 16 years with central government budgetary support. However, as noted by Lombard in 1995 and by Nathan in 2001, the performance of this program has not been particularly satisfactory. Reporting on the *upesi* rural stoves program in Kenya in 2001, Njenga concluded that the benefits to men and women include improved health and time savings for the users. However, it should be noted that there are no reported systematic observations and empirical evidence to support this conclusion. On the other hand, the women producing and

promoting the stoves earned some additional income that Njenga noted made them enjoy a significantly higher standard of living.

The extent to which improved wood stoves benefit women with respect to time, labor, and financial savings in rural areas is limited by a variety of factors. In many countries, improved stoves are produced by artisans in informal situations with limited technical skills and quality control, so that the stoves do not always achieve the expected fuel savings. Most rural households do not buy fuelwood and, therefore, experience no direct monetary savings. In a 1997 report, Nathan noted that China experienced success in improved stove programs, with some 150 million (or 70%) of farm households adopting improved stoves sold to them at commercial prices by 1991. On the other hand, in India, only approximately 15% of farm households had adopted improved stoves by 1992, even though the improved stoves were offered at highly subsidized prices. Nathan noted the difference as being the monetary value of women's time. Women were involved in income-generating activities in China, whereas they were not in India. Savings in time spent on fuelwood collection had direct value in China because the time was reallocated to the enterprises, but this was not the case in India. Nevertheless, having some free time does have welfare benefits. However, not surprisingly, financial savings have been reported for urban households that purchase fuelwood and charcoal after acquiring improved stoves.

In general, the roles of women in the improved stoves process has received a good deal of attention. More than two decades ago, Hoskins reported in 1981 that adoption and use of improved stoves would not be possible without the active involvement of women. In a study reported in 1991, the FAO noted that energy-efficient earthen stoves in a village in Himachal Pradesh, India, became popular in areas where women were trained to make such stoves and introduce them to other women.

Stove programs have rarely met the fuel-saving objectives, partly because the goals were not based on realistic scenarios. Various factors influence the amount of fuel used (e.g., the skills of the cook, the cook's interest in the task of cooking, the stove, cooking utensils). As noted way back in 1987 by Evans, women's main interests have been largely misconstrued by stove project managers as saving fuel. In reality, women prefer having a better fuel, not a better stove.

It would appear that the effectiveness of the improved stoves in saving women's time and labor is rather limited and dependent on various external factors that are not always taken into account in stove programs. Financial savings from reduced fuelwood consumption occur only where fuelwood is purchased and also where the technical performance of the stove meets the expected efficiency levels. Notably, as acknowledged by Leach and Gowen as far back as 1987, during times of fuelwood scarcity, coping strategies associated with traditional stoves are reported to outweigh the benefits associated with improved stoves.

5. HOUSEHOLD ENERGY AND HEALTH: A WOMEN'S PROBLEM?

There is general acceptance that exposure to biomass and coal smoke, generally classified as indoor air pollution, increases the risk of a range of diseases and impinges on the health of adults and children. This pollution affects health in two main ways: through having direct effects, such as carcinogens and toxins causing cataracts, and through damaging or weakening the immune system, thereby making it more susceptible to attack. Impacts include acute respiratory infections, chronic obstructive pulmonary diseases (e.g., chronic bronchitis), lung cancer, tuberculosis, adverse pregnancy outcomes, and asthma attacks. However, current information on smoke as causal is indicative and not always conclusive; there is uncertainty associated with exact risk estimates. For example, a 2000 report noted that although there is increasing evidence of links between PM10 and PM2.5 and health, available evidence does not allow a clear judgment on levels of particulate concentration above which adverse effects would occur. There is growing awareness that various factors contribute to these health impacts, and more evidence is needed to establish the cost-effectiveness of measures associated with smoke reduction and elimination. Poverty, poor nutrition, poor living standards, poor sanitation, overcrowding, and low-quality medical care are significant (and possibly more important) determinants of health impacts associated with indoor air pollution. Nevertheless, indoor air pollution has been receiving increased attention, and in 1992 the World Bank designated it as one of the key global environmental problems. The World Health Organization (WHO) noted that a pollutant released indoors is more likely to reach people's lungs than is a pollutant released outdoors.

Literature identifies women and children as the most vulnerable because they are the most exposed,

particularly in poor rural and urban households. However, it is likely that the attention paid to women and children tends to hide the fact that adult men may also suffer indoor pollution-associated ailments. In an analysis of the prevalence of respiratory disease symptoms in rural Rajasthan, Laxmi and colleagues found no significant difference between male nonsmokers and female nonsmokers, although they had expected the males to act as a control group because they did not cook. In addition, the results suggested that male nonsmokers, as compared with female nonsmokers, have a higher prevalence of symptoms and are more vulnerable to coughing, phlegm, wheezing, chest illness, and blood in sputum. (The researchers noted, as a shortcoming of their study, an insufficient male representation due to the work having been designed to study females cooking with biofuels.) Including the male dimension is particularly important because increased urbanization means that more men are cooking and that the size of dwellings is relatively small and often lacking separate rooms for cooking, with men and women occupying the same rooms.

The physiology of children, particularly underdeveloped lungs and relatively faster breathing rates, makes them liable to absorb large quantities of pollutants. Smith and colleagues, in a study published in 2000, reported that acute respiratory infections related to biomass fuel combustion are the leading causes of deaths in children under 5 years of age, with 75% caused by pneumonia. It is estimated that half a million children and women die in India annually from indoor air pollution. However, findings are not always consistent. A study in south Kerala, India, focusing on children under 5 years of age, as reported by Shah and colleagues in 1994, found that the fuel used for cooking was not a significant risk factor for acute respiratory infections. Based on a 1998 study on infants in Delhi urban slums where wood and kerosene are used, Sharma and colleagues noted no significant difference in prevalence of acute respiratory infections and fuel type.

The chemical composition of smoke and the concentration of particulates determine the level of attributed harm. Smoke generally includes suspended particles, nitrogen oxide, carbon monoxide, sulfur oxide, and carcinogens (e.g., formaldehyde, benzene, polyaromatic hydrocarbons). The extent of the impacts of composition vis-à-vis the particles is contested. In addition, the combustion characteristics affect the pollutants. For example, sulfur dioxide is dependent on fuel content, whereas concentrates, formaldehyde, and nitrous oxide depend on the efficiency of the combustion process. Carbon monoxide is released in relatively large quantities when biomass fuel is burned, as compared with kerosene or LPG, with short-term effects to exposure being headaches, dizziness, and nausea (extensive inhalation can cause death). In a 1991 study, Smith estimated that 38, 17, 5, and 2 g of carbon monoxide are released during the cooking of a meal using dung, crop residues, wood, and kerosene, respectively. In a 1995 report, Patel and Raiyani noted findings of carbon dioxide levels of 144, 156, 94, 108, and 14 mg/m^3 air during cooking with dung, wood, coal, kerosene, and LPG, respectively, in Ahmedabad, India.

Indoor air pollution has generally been linked with rural areas due to high household consumption of biomass and coal, but the urban phenomenon should not be underestimated. The WHO estimated that indoor air pollution contributes 50 and 25% of the diseases in rural and urban areas, respectively. Reports from India published by ESMAP in 2002 show that cooking smoke may be responsible for more than 400,000 premature deaths annually. There is evidence that many of the people moving to urban areas tend to settle in low-income neighborhoods where they use biomass or coal. Rising urbanization levels and the share of urban poor will further exacerbate the situation. Indoor pollution effects in low-income urban households are likely to be more aggravated than in rural areas due to overcrowding and the outdoor pollution effects.

In a 2002 review of indoor air pollution aimed at making recommendations for further research, Smith categorized three impacts based on the level of evidence: strong, moderate, and suggestive. Acute respiratory infections, chronic obstructive pulmonary diseases, and lung cancer fall into the first category. Cataracts, tuberculosis, asthma, and adverse pregnancy outcomes fall into the moderate evidence category. However, it is important to state, as noted by Ezzati and Kammen in their 2002 study, that most of the studies on health impacts of exposure to indoor air pollution have focused on these three diseases. Studies undertaken in China, including those reported by Smith and Liu in 1994, by Xu and colleagues in 1995, and by Lan and colleagues in 2002, provide evidence of an increased risk of lung cancer among women from exposure to coal smoke, but there is little evidence relating to biomass fuels. Lan and colleagues also reported that the duration of cooking food on an unvented coal stove was strongly and positively associated with lung cancer incidence in both men and women (even though men started cooking later in life and tended

to cook for a shorter period of time). This is important because more men are cooking, particularly in urban areas.

Links between biomass fuels and chronic obstructive lung disease (e.g., colds) have been reported. In a 1996 report of a study involving 1200 women (218 of whom were monitored for air pollution) in low-income areas of Maputo, Mozambique, Ellegard reported finding cooking wood smoke as being associated with chronic respiratory illness among women. The study found that wood users were exposed to significantly higher levels of particulate pollution during cooking time (1200 $\mu g/m^3$) than were charcoal users (540 $\mu g/m^3$) and modern fuel users (e.g., LPG, electricity) (200–380 $\mu g/m^3$). Although wood users were found to have significantly more cough symptoms than were other groups, there was no difference in cough symptoms between charcoal users and modern fuel users. However, there are no studies in developing countries linking heart disease to biomass fuels. In a study published in 2000, Bruce and colleagues noted that biomass fuels have been linked to cancer of the mouth and larynx but that the findings are inconclusive. Studies conducted in India show the incidence of chronic cor pulmonale to be similar between men and women despite the fact that 75% of the men and only 10% of the women were smokers. In addition, the condition was more common among younger women, occurring 10 to 15 years earlier than in men. This difference was attributed to indoor air pollution from burning solid biomass fuels leading to chronic bronchitis and emphysema, both of which result in chronic cor pulmonale.

Although it has been reported that indoor air pollution causes birth-related complications, including prenatal mortality and low birthweight, there is hardly any statistically representative work that validates this. The one study frequently cited on perinatal mortality was conducted in Ahmedabad and was published by Mavalankar in 1991. This study found a link between exposure to indoor air pollution and stillbirths and deaths during the first week of life with an odds ratio of 1.5 (95% confidence interval: 1.0–2.1, $P = 0.05$), although exposure was not assessed directly, implying marginal significance. Only one published study has reported wood smoke as affecting birthweight. This study, reported by Boy and colleagues in 2002, was conducted in rural Guatemala, where babies born in households using woodfuel were 63 g lighter than those born in households using gas or electricity.

Procurement of energy may also have health impacts on women. In a study conducted in Sri Lanka and published by Wickramasinghe in 2001, women reported fatigue, headaches, pain in joints and the chest, and weak joints and back resulting from head loading wood and walking over long distances several times per week. However, it is not evident from the report that carrying wood was the only or dominant head loading done by the women. In general, there is insufficient reliable empirical data on weights of fuelwood loads. Work done by the ILO during the 1980s indicated that women carry loads of up to 35 kg over long distances, whereas a maximum of 25 kg is frequently cited as the weight that can be borne over such distances without seriously threatening the physical health of women. Cuts, falls, bites, back injuries, and sexual harassment have also been associated with fuelwood collection. Fuel scarcity is noted to force women to move to foods that cook faster or to eat more raw foods, increasing health risks to entire families. Obviously, this is due to household poverty rather than to physical fatigue and poor health, viewed nearly exclusively as a problem affecting women.

6. WILL IMPROVED STOVES IMPROVE THE HEALTH SITUATION?

There are various measures to reduce exposure to indoor air pollution. In their study reported in 2000, Ballard-Tremeer and Mathee categorized three main types of interventions relating to source (emissions), living environment (concentrations), and the behavior of the user (exposure). The type of fuel and stove used influence indoor air quality. The fuel–stove interface is a significant determinant of emissions generated. Based on their evaluation, Ballard-Tremeer and Mathee noted that the most effective interventions, and those most beneficial to the user and society as a whole, would involve a shift from wood or charcoal to kerosene, LPG, biogas, or grid electricity.

Based on an empirical study conducted on 67 caregiver–child combinations living in poor unelectrified areas in rural South Africa, Barnes and colleagues in 2000 identified four core behaviors for intervention to reduce child exposure: improving stove maintenance practices, improving the quality of ventilation practices, reducing the time spent close to burning fires, and reducing the duration of burning. The authors put much emphasis on the

recommendations made by the caregivers. Hence, while acknowledging fuel switching as the most effective intervention, they omitted it because the caregivers noted that the associated costs would be prohibitive. Changing the location of a stove to better ventilated locations in the kitchen was dismissed for similar reasons. Improved stove maintenance, including cleaning chimneys and replacing missing cooking plates, has also been recommended as a strategy to reduce indoor air pollution. It was noted that women in most cases already open windows and doors (when available) while cooking. Applicability and effectiveness of behavioral changes remain to be verified. Clearly, the changes have cost and labor implications.

Improved stoves have been widely advocated as a measure to reduce indoor air pollution, and this is supported by various studies, including those by Ezzati and Kammen in 2002, Albalak and colleagues in 2001, and Wafula and colleagues in 2000. (However, it is imperative to note that this rather common reporting can be misleading because it is essentially the presence of a mechanism to vent out the smoke that contributes to reduction. The term "improved stoves" is not used exclusively for stoves with chimneys; it is also used for higher efficiency stoves without chimneys.) The first stove program launched by the Indian government aimed to get rid of smoke from the kitchen by using chimneys. In an analysis of 21,232 farmers in rural Xuanwei, China, Lang and colleagues noted that a shift from unvented coal fire pits to coal stoves with chimneys was strongly associated with a reduction in the incidence rate of lung cancer. Such chimneys extract smoke by increasing draught and, hence, the combustion rate. In the process, fuel savings are compromised. However, depending on the height of the chimney, the measures do not get rid of the pollution problem completely but instead transfer it from the kitchen to the vicinity of the dwelling. In an environment where houses are constructed close to each other, smoke builds up and remains in the neighborhood. This is aggravated where outdoor air pollution level is high, as is often the case with low-income environments of large industrial cities. Reporting on findings in China, Edwards and colleagues noted in 2003 that flues and chimneys are not much above the height of houses; hence, emissions result in elevated neighborhood and regional pollution that penetrates back to the indoor environment. Notably, where the heights of the chimneys are low, the smoke can travel back into the houses through doors and windows. Chimneys can be ineffective if the smoke returns through windows and doors.

Evidence on the ability of improved stoves to reduce indoor air pollution is contentious. It has been reported that such stoves can actually increase indoor emissions. Smoke is a result of incomplete combustion that occurs due to insufficient oxygen. Improved stoves save fuel by controlling the burning rate and, hence, the air flow. The effectiveness of improved stoves depends on maintenance and, therefore, on age or hours of use. Stoves that are poorly maintained are unlikely to contribute to reductions in indoor pollution. The way in which a stove is used influences emissions levels. For example, closing the door prevents smoke from coming back into the house from the chimney. On the other hand, when using wet fuel, it is usually necessary to open the door to facilitate ignition and combustion. Edwards and colleagues, in their recent study on improved fuelwood stoves in China, noted that thermal efficiency is improved at the expense of combustion efficiency, resulting in increased emissions of human-damaging pollutants and greenhouse pollutants.

7. DOES INCREASING THE SUPPLIES OF BIOMASS FUEL HELP?

Retreating forests have always been viewed as threats to fuelwood availability in rural areas and, hence, to women's time and labor. Social forestry to grow trees for fuelwood has been a commonly pursued solution. Despite evidence that the fuelwood availability issue is more complex than a simple gap theory implies, and that most of the wood produced through afforestation practices is destined to become timber rather than fuel, considerable resources have been directed to these activities. Community-based forestry is encouraged as a source of revenue where fuelwood is traded for urban consumption. Women have been promoted as managers in such projects. In these community-based projects, men's and women's groups, with assistance from the forest departments, sell firewood as a source of income. Projects target women as producers and potential beneficiaries. However, not all of the community-based forestry benefits women. In a 2001 report, Sarin noted that in Uttra Pradesh, India, women were worse off after the introduction of village joint forest management because they could no longer get jobs as forest guards and had no control over decisions relating to forest use.

The validity of afforestation programs aimed at providing fuelwood has been challenged. Using data

from dry tropical areas of Niger and Mali, Foley in 2001 raised serious doubts regarding the woodfuel scarcity problem and the solutions proposed, which have been mainly planned interventions on use of improved stoves and tree planting. There have been considerable efforts, through projects and programs, to involve women in tree planting for fuelwood. Various community-based resource management projects targeting women, on account of the suffering they encounter in gathering fuelwood, have been implemented. However, these efforts are not justified given that communities naturally undertake the necessary activities without external intervention. Findings from rural Rajasthan, where 47% of households in the study used fuelwood gathered from their own farms, support this thinking.

Within the context of afforestation and tree management for fuelwood, some works have differentiated "men's trees" from "women's trees." The former are defined as straight-stemmed species useful for timber and poles, whereas the latter are defined as trees that yield a broad range of products such as fruit, fodder, and firewood. The aim of such works has been to influence the choice of trees planted under social forestry programs and, hence, to protect women's interests. Poor access to fuelwood by women is also blamed on their lack of land tenure, implying that women are reluctant to plant trees because the trees would be "hijacked" by men. In a 1995 report, Kelkar stated that a group of men invited to jointly plan a community forestry project told the foresters that they wanted to plant hardwood tree species to make furniture and wood carvings for sale. When women were consulted, they indicated a preference for softwood fast-growing species for fuel and fodder. The Regional Wood Energy Development Program (RWEDP) of the FAO, in a 1996 report, also noted that women prefer trees for fuel, fodder, and fruit, whereas men prefer timber trees. In addition, the RWEDP noted that women prefer trees that yield shorter term returns and smaller returns spread over a long period (e.g., fruit trees rather than timber trees). Women's unique responsibility for day-to-day care of their families was given as the reason for the different preferences.

It has also been claimed that men, as primary owners of land, have better access to fuelwood and are in a better position to dominate decisions about tree species in forestry projects at the expense of women. However, it is worth noting that ownership of land does not guarantee access to fuelwood given that the competition for land for food and outputs more beneficial than fuelwood has meant that farmers (including women) give low priority to the latter. Increased demand for cash necessitates trading most farm products, and in communal lands most of the forested land has been cleared, so that even if women owned the land, their access to fuelwood might not necessarily improve.

8. WILL ELECTRICITY HELP WOMEN?

Rural electrification is another area that draws attention to energy procurement by women. Electrification has 7been justified on the basis that it results in economic progress, but recent thinking has acknowledged that electricity is an insufficient, albeit a necessary, catalyst for this. Some, such as Clancy in a 1999 article, have characterized electricity as a basic good necessary to empower women. Cecelski, in a 2000 report, called for rural electrification on the basis that it would meet women's needs for labor saving, time saving, improved health, security, and income. But as noted by various authors, including Mehlwana in 1997, ESMAP in 2002, and Wamukonya and Davis in 2001, electricity in rural areas is used mainly for lighting and other light load activities such as powering radios. ESMAP's study of 5048 women in six states in India in 1996 concluded that household electrification affects leisure time available to women and provides channels to increase knowledge and awareness through facilitating reading and watching television. A seemingly contradictory finding from this study involves the decrease in fuelwood collection times, even in poor households that reportedly use electricity only for lighting and entertainment.

Because the bulk of the time expenditure, drudgery, and health impacts related to energy procurement by women is in reference to cooking fuels, rural electrification cannot solve these problems unless it is provided for cooking. Lack of electrical cooking appliances and the high cost of power are major limiting factors. Even in countries where the governments have made concerted efforts to electrify their populations, the electricity consumption for cooking remains low. Despite an aggressive electrification program in South Africa, Mathee and colleagues noted in a 2001 report that more than $3\frac{1}{2}$ years after provision of electricity, 44% of the households in North Western Province had not acquired electric stoves. However, unless electricity production costs are very low, it is difficult to justify targeting it as a cooking fuel.

Decentralized electrification in rural areas has gained favor largely as environmental concerns move to the forefront and the cost of extending the grid to remote areas with relatively low-load consumption patterns becomes prohibitive. This is manifested mainly as mini- and micro-hydro in areas where hydro resources are available, but largely as household photovoltaic (PV) systems. In fact, as a response to the sustainable development movement, donor agencies abandoned grid extension policies in favor of stand-alone systems based on PV, mini-hydro, or biomass. Solar home system projects are proliferating all over the developing world. Martinot and colleagues, in a 2000 report, provided a comprehensive list of some Global Environmental Fund (GEF) solar home system projects, and Martinot and colleagues in 2002 give an overview of this market. However, capital cost is a major barrier to accessing these systems, whose promotion has been justified on cost-effectiveness. In view of this, governments, with assistance from donor grants and loans, have established innovative financing mechanisms to provide loans to consumers. Despite these, there have been concerns that women are unable to access credit for solar home systems, so facilities targeting women have been established or concerted efforts to reach women have been made within existing frameworks. Enersol, a U.S.-based organization aimed at deploying solar-based technologies for rural electrification, has established such a system to enable women to acquire its solar home systems. In its microenterprise training courses to NGOs, Enersol emphasizes the importance of women as borrowers. Although Smith noted in 2000 that the women use savings from energy replaced by the solar system, no information was provided on volume and type of energy expenditures prior to and after acquiring the system.

The extent to which these systems benefit women is questionable. Evidence documented in a United Nations Environment Program (UNEP) 2001 publication shows that typical solar home systems are consumer goods. The solar home system electricity does improve household welfare with respect to better quality lighting, but at a relatively high cost compared with conventional electrification systems. Therefore, women end up getting loans to acquire systems that provide strictly limited and costly services and that do not address the key energy concerns—cooking and its associated hardships—so that gathering of fuelwood and other biomass fuels continues even after acquisition of the solar home systems.

As a measure to facilitate access to electricity, there have been recommendations to target women in dissemination of solar home systems. Enersol has been targeting women as users and promoters of solar home systems in the Dominican Republic. However, Smith noted in a 2000 report that despite a concerted effort to train and support women, of the 30 active businesses supported by Enersol, only 2 have women in significant leadership positions and 1 relies on a female technician for system installation and maintenance. After one of Enersol's women-only training programs, no microenterprise dedicated to marketing and installation of PV systems emerged from the group of 10. This may imply that the main barrier was not necessarily being women. It is worth noting that Enersol has depended mainly on USAID financial support to train and support women entrepreneurs.

9. WILL SAVING TIME HELP WOMEN TO ENGAGE IN MORE PRODUCTIVE ACTIVITIES?

It is often argued that the time spent by women and children in gathering fuelwood could be employed for more productive purposes. However, there are inadequate empirical data indicating this, and proxies are normally used to calculate the value of this time and labor. It cannot be ascertained that the time freed from fuelwood gathering could be used for more productive purposes. Interviews with Namibian rural women in 1999 revealed that they would spend the time freed from fuelwood gathering on housework and not necessarily on leisure or income-generating activities. The main reasons for not this were lack of opportunities and lack of capital. ESMAP reported in 2002 that in India, rural electrification has not influenced the time spent by women in productive activities.

It is important to acknowledge that time saved from procuring energy sources is not automatically reallocated to income-generating activities. Only where the other more important factors are available does this happen. If the objective of an initiative is to save women's time from fuelwood gathering for use in income-generating activities, the initiative needs to provide the other complementary factors. A study conducted in Nepal found that only the women who already were involved in income-generating activities and were using biogas engaged in these activities by using the time saved from fuelwood gathering. In more than 80% of the households, social work and domestic chores were the main activities that women spent more time on (where they did not already have

income-generating enterprises). In a study published by Rana-Deuba in 2001, it was noted that women receiving microenterprise training in Nepal were able to start businesses mainly because the training was backed by access to credit. In 2001, Chaieb and Ounalli reported that in Tunisia, rural electrification had increased income-generating opportunities for women mainly in sewing and hairdressing businesses, but this was feasible only because state development programs donated the equipment to households. After installation of biogas plants in Pura village, India, to provide electricity for lighting, water pumping, and using high-quality fertilizer, women were able to earn a small amount of money transporting fertilizer to the fields of farmers. It must be noted that such an activity is supported by the availability of the fertilizer as a by-product, requiring no extra capital outlay. Only where the environment is conducive to this activity, and where other factors exist, are time savings used for productive activities.

Fuel switching—from fuelwood to modern energy carriers—is often advocated as a measure to reduce women's drudgery from fuel collection. However, a shift from fuelwood might not necessarily help women and children in this way. A study conducted by Britt and Kapoor in 1994 in Nepal, measuring the increase in women's leisure time and time spent visiting relatives, found that women without biogas plants actually had more leisure time and more time to visit relatives than did women with these plants. Biogas use has been noted for not really reducing women's time expenditure on fuelwood procurement because women have to collect water for the digesters. The same situation was reported in reference to Tanzania.

10. DO WOMEN REALLY HAVE ENERGY NEEDS DIFFERENT FROM THOSE OF MEN?

Paying special attention to women is normally justified on the basis that women's energy needs are different from men's energy needs. This thinking is advanced while promoting certain technologies. Decentralized technologies, particularly renewable energy technologies, are increasingly lauded as the "natural solution" to women's energy problems, particularly in rural and remote areas. This is often in reference to using these technologies for electricity provision. It is noted that women use electricity differently than do men due to the former's different household and productive activities. Rural women's energy needs are said to be water pumping, agricultural processing, security, work productivity, and health, and rural men would seem to have the same needs. It would be expected that in urban areas as well, there is little real difference between women's and men's energy needs.

There are also reports that women tend to see the benefit of electricity differently than do men. In a synthesis report on micro-hydro power published in 2000, Khennas and Barnet noted that women in Sri Lanka saw the benefits largely in terms of reduced workloads, better health, and reduced expenditures, whereas men saw the benefits in terms of leisure, quality of life, and education of children. In a study reported in 2000, Cecelski lamented the fact that few studies have examined the impacts on women from renewable energy interventions. In a 2002 report, Parikh recommended that energy ministries assign special tasks to their renewable energy divisions to assist women in evaluating possible energy alternatives. However, it is again important to note that energy is not important in its own right; rather, it is the service energy provides that is useful. Therefore, singling out decentralized or renewable energy technologies may be counterproductive if the technologies do not meet the services needed, and they might even contribute to "locking in" the women. The focus should be on service rather than on the technology. The technology chosen should be the least cost option, but the costing should take into account the ability of technology to meet development objectives, particularly poverty alleviation.

11. ENERGY DECISIONS: ARE THEY ALL MADE BY MEN?

There is a widespread belief that women lack decision-making power. The WEC, in an analysis of factors determining energy–women linkages published in 2000, noted that women, by virtue of their position (or lack thereof), stand disadvantaged in decision-making processes in their families, communities, and countries as well as in accessing productive assets. This thinking is used in designing women-targeted energy programs and projects.

Renewable and decentralized energy technologies targeting women are recommended as a measure to empower women in the decision-making process. According to a study by Smith in 2000, decentralized energy technologies are particularly relevant to women because women have direct control of

acquisition, design, placement, and consumption decisions, whereas they often exert less control over male-dominated, utility-centered grid systems. Decentralized technologies have also been deemed favorable because they provide more opportunities for women to be involved in energy system management, decision making, and control than do centralized grid systems. In addition, such technologies provide opportunities for women entrepreneurs. However, this involvement is not automatic. In Sri Lanka, women were not considered to be decision makers in the Electricity Consumer Societies formed in some of the communities with micro-hydro plants, whereas they were encouraged to participate in others. Women have been targeted by various programs as trainees on maintenance and installation of PV systems, including solar cookers, as a measure to improve their decision-making powers. However, even the decision to engage in an enterprise is not necessarily limited by gender. In a report published in 2000, Balakrishnan stated that after a women's training program in India in 1993, 2 of the 45 women started renewable energy businesses, with 1 reported as having "gone a long way" into wind farms in Karnataka, and a 3rd woman joined a UN agency as a consultant. It is difficult to understand what "a long way" entails because no data were provided on sales or other aspects of business performance.

As noted earlier, there are calls for providing finance to women to enable them to adopt renewable and other energy technologies, but also specifically for acquiring solar home systems. One of the justifications for extending loans to women for energy technology procurement is the fact that they are rational and would purchase technologies that are valuable and address their household needs. The usually implicit (but occasionally explicit) counterargument is that men are irrational and would spend money on their own entertainment rather than purchasing technologies that would benefit their households. It has been noted that men are systematically biased in favor of their own leisure time and that this has resulted, for example, in underinvestment in devices that save women labor time and fuel. Nathan, in a 1996 report on work done on Hani people in Yunnan, stated that money was spent on cassette players (for men's listening pleasure) and not on getting piped water or on improved stoves. If these observations are justified, extending loans to women for the typical solar home systems might also be labeled as irrational given the service limitations. Along the same lines of thought, it is normally implied that women do not value entertainment from radio and television and would rather acquire cooking appliances because they lack the time for entertainment. Laxmi and colleagues, in a 2003 report, noted that cooking and energy-saving appliances are accorded lower priority than are luxury goods (where the luxury goods identified are durables such as pump sets, bullock carts, fans, and radios/tape recorders) due to little involvement of women in decision making. It is worth noting that this is a mixed bag of goods given that some of these "luxury goods" are actually vital for the survival of households. It is perhaps worth mentioning that from the data presented by these authors, the number of households owning kerosene stoves is higher than the number of households with pump sets.

Lack of decision-making power among women in households has been stated as a rationale for targeting women to facilitate making the appropriate choice on energy matters. Comparing a nonelectrified village in South Africa with an electrified one, Mathee and de Wet in 2001 noted that women in the nonelectrified village indicated that an electrical cooking appliance would be the first electrical appliance they would like, whereas women in the electrified village indicated that this was among the last appliances bought there. The authors attributed this discrepancy to the dominant economic decision-making power and the preference of men at the household level. However, this conclusion is not supported by any empirical evidence on decision-making processes. In advocating measures to enable women to influence household decisions (e.g., through self-help groups), an article in a World Bank newsletter in 2000 noted, in a similar manner, that "when it comes to commercial fuels which are bought for cash, the problem often is that income decisions are made by the male heads of the household, while women have to bear the burden of collecting fuelwood and using it for cooking." In 1997, Dutta remarked that although women are the potential users of improved cooking stoves and other cooking innovations and, therefore, are in the best position to assess their advantages and disadvantages, it is the men who handle household cash and make decisions on how it is to be spent. The differences in rationality between men and women where acquisition of household energy appliances is concerned are contentious and, as noted earlier, are rarely supported by empirical evidence. For example, that women from all income classes would watch television if one were available was one of the findings of the ESMAP study of six Indian villages.

Community-based forest management was initially promoted as a measure to restore degraded forests and to improve access to fuelwood for sale and household consumption. Women were expected to benefit from these initiatives, but their lack of decision-making power has largely been blamed for the negative impact that community-based forestry has had on access to fuelwood by women. To revisit an earlier point, decisions regarding what trees should be grown in farm forestry are made by men, and this prevents households from planting trees for fuelwood. Sarin and colleagues in 2003 reported that in a study of four villages in Orrisa, India, women had decision-making power on forest use in only one of the villages. The authors noted the delays in relaxing rules on restricting fuelwood collection, and they blame this on domination of men in the relevant committees. The same reason was given to explain why rules on timber extraction were elaborated, whereas those on fuelwood were not. Clearly, controlling fuelwood extraction is rather difficult compared with controlling timber extraction, and this may have been the principal reason for the delays. There is an underlying assumption that men benefited more from timber than did women.

Some argue that the problem of women's exclusion from energy-related decisions at the household, community, and national levels is compounded by their underrepresentation in decision-making processes related to energy policy and planning. The underlying assumption is that women decision makers would address women's energy needs more effectively. It is reported that engendering of policies is on the increase as a result of advocacy by women's groups. The main indicator of this engendering has mainly been the number of times "women" appears in the relevant documents. Whether this is an effective indicator is subject to debate. Advocating for an increase in the number of women in decision making certainly has its merits, particularly because it creates employment opportunities. However, using the number of women in any given institution as an indicator of representation has various drawbacks. First, the representation system runs the risk of abuse, where nepotism and favoritism are used to select women who are likely to become mere "rubber stamps." This would weaken the whole decision-making process to the disadvantage of both men and women. Second, women politicians and other decision makers are not always able to alter energy decision making to cater for women's concerns, due mainly to a general lack of comprehensive understanding of the policy impacts and interference by stakeholders with vested interests. Third, because the proportion of women professionals in the energy sector is relatively low, focusing on having women on-board as a prerequisite to addressing women's concerns may derail the issue, if there is indeed an issue, and may reduce opportunities for open dialogue to establish what the main concerns are and how to tackle them. Even where there are affirmative action policies or quotas reserved for women, it is difficult to get qualified personnel. Obviously, it will take a while before there can be a critical mass of women professionals in the energy sector. But as is argued here, energy issues are not so much about women as about poverty. A woman professional will be a better candidate than a man in the energy sector only if she can design and implement frameworks that use energy to alleviate poverty.

A strategy that has also been used to increase the participation of women, as well as to raise the profile of gender issues in the energy sector, is the establishment of gender desks in energy and other ministries staffed nearly exclusively with women. Despite a shift to gender and development, this is a typical reflection of the women in development approach that found expression in small female-staffed gender cells devoted to women's projects. These offices are meant to integrate gender into policy and action plans. However, they tend to be isolated and without links to other relevant ministerial or external (e.g., utility) stakeholders. Other staff members consider these offices as dealing with "women's affairs," that is, issues they regard as not relevant to themselves and their work. As such, the energy gender offices remain largely figureheads, with little influence on the energy agenda. If they are to function effectively, there has to be a broadly shared understanding of the issues and tools to effect change. An important element would be to employ both men and women. Most of the energy gender offices lack dedicated budgets to operate, relying instead on mainly donor funding, the intermittent nature of which makes it difficult to have an impact because there is no guarantee of continuity of the necessary projects such as awareness raising.

12. WILL ACCESS TO ENERGY INCREASE WOMEN'S INCOME-EARNING POSSIBILITIES?

Access to financing for acquiring energy and other inputs for income generation has been cited as a key barrier to women's involvement in entrepreneurship.

Lack of collateral and, more important, networks makes it difficult to obtain loans. A strategic solution that has been gaining favor is the establishment of women-only financial institutions, which are justified on the basis that women do not receive adequate attention in men-dominated institutions. Another justification is the need to end women's dependence on loan sharks who demand exploitative interest rates, keeping women in a vicious poverty cycle. Most of the women-targeting institutions tend to be operated mainly by women. Although this has merit so far as increasing employment opportunities for women is concerned, the women-only approach can be detrimental to the overall objective. The institutions are generally isolated from the conventional male-dominated business world, and this limits networking and makes the opportunities to transit into normal business cycles difficult.

As a way of facilitating participation of women in entrepreneurship, their projects are evaluated less stringently than those of their male counterparts, and projects that might not be financially viable are funded. For example, the Mpongwe Women's Bank, managed by Credit Management Services Limited in Zambia, does not undertake a detailed loan appraisal process but rather relies on the assessments and character references provided by the women's clubs of which the applicants are members. Many of the projects fail because they have not been built on sound economic principles, leaving the women worse off than they might have been if the projects had never existed because they have to repay loans from other sources. The failure erodes confidence in women as business entrepreneurs, both among women themselves and among the conventional financial institutions. Although it may be important to make concessions for women, these should be done in a manner that makes business sense and in which women are integrated into the normal business world through addressing their weaknesses such as business management skills and seeking good market outlets. So long as women are operating enterprises within the conventional environment, provision of kick-starting financial support is only a small part of the equation for a successful business.

Another particularly important aspect of the loans is their size. Most credit programs established to serve women provide only small loans that are typically insufficient to generate a sustained process of capitalization. Credit Management Services gives maximum loans of U.S. $250 per applicant. In 1999, the average loan size (with share of women loan clients in the MFI in parentheses) from Get Ahead (92%) in South Africa was U.S. $192, from Citi Savings and Loans (71%) in Ghana was U.S. $363, from ACEP (30%) in Senegal was U.S. $526, from K-rep (58%) in Kenya was U.S. $878, from EBA (12%) in Egypt was U.S. $841, and from CERUDEB (27%) in Uganda was U.S. $1110. Most of the credit schemes rely on donor funding in a manner that reflects the original approach during the 1970s, when donors and NGOs provided special assistance to women in a primarily welfare-oriented approach. As such, these agencies are also quite unsustainable. Although there is a rationale for providing small loans to the poor, because they are high risk, there is the greater danger of entrapping them in this state. In addition, this largely limits the choice of enterprises to just a few.

The typical enterprises in which women invest are food processing, sewing, basket weaving, and so on, with little diversification across loan beneficiaries in a given area. As a result, competition is stiff, particularly because most of the entrepreneurs lack access to external markets. Cecelski, in a 2000 report, advocated increasing access for credits to women to purchase renewable energy technologies for domestic use and microenterprises. However, as noted earlier, most of the focus has been on solar home systems that provide limited service and are basically consumer goods that do not, by themselves, offer opportunities for income generation.

Overall, in most cases, access to credit for energy has not particularly improved women's poverty status in any significant way. The microcredit, small player perspective seems to have been picked up by the New Partnership for Africa's Development, which identified improving access to credit by small-scale and women farmers as a strategy for promoting women. Although the idea of extending credit to women is commendable, this should be done as part of an integrated approach if the goal is to yield sufficiently sustainable benefits and to allow households to graduate from poverty.

13. CONCLUSIONS: AREAS FOR FURTHER RESEARCH

A review of the literature reveals gaps that exist in the women and energy discourse. As a conclusion to this article, questions for further research are posed. It is expected that such research would contribute to identification and implementation of effective activities that result in sustainable development.

Although women and children in rural areas have been largely dominant in fuelwood procurement for household needs, it seems imperative to question the role of settlement pattern (rural vs urban). In the past, more men than women migrated to the cities; hence, the share of women in rural areas has been rather high. This migration is generally triggered by lack of income-generating activities in rural areas. However, it has had an impact on the division of labor. Hence, there seems to be a need to determine whether women in rural areas suffer energy poverty because they are women or due to the locality. Is access to modern energy not determined more by density of demand than by gender or sex?

Poor households in rural and urban areas tend to consume relatively low quantities of energy. Would provision of modern energy services within close proximity to poor households improve their energy and livelihood status? Would the women in such households cease to suffer from the drudgery associated with energy procurement? Or, is this feasible only when the energy is provided as a service facilitating income generation?

So far, the women and energy discourse has focused mainly on the household sector. Even when the debate moved into income generation, there has been a tendency to place it in the household context on the basis that women's income-generating activities are mainly household based. But to what extent are the referred activities women oriented? Would provision of energy result in income generation for women? There is evidence that there are more societal benefits accruing from providing energy in the public sphere than from providing energy in private entities. Would public benefits emanating from modern energy provision in health care centers, schools, and growth points not be more effective in facilitating growth than those emanating from women-targeted energy projects? Should this approach not be the main focus, particularly in rural and low-income urban settlements?

Having women as decision makers in the household and at the policy level is recommended as a strategy to improve women's energy status. But has this been verified empirically? Can it be verified? Cultural factors limit understanding of the decision-making processes in households. In most cases, outsiders tend to identify who the decision makers are on the basis of the preconceived expectations of the outcomes. There is insufficient information on whether women decision makers necessarily make decisions that are more favorable to women. Even if this turns out to be the case, would it be socially desirable?

Many developing countries are experiencing energy reform where the private sector is expected to play a dominant role in energy service provision. Consequently, a market approach where the focus is profit will be adopted. In such an environment, would women-targeted energy projects be viable and effective in alleviating poverty?

Acknowledgments

I am extremely grateful to Margaret Skustch for her concise guidance on the article. I also thank Sami Kamel, Ogunlade Davidson, and Lawrence Agbemabiese for their critical comments and encouragement to publish. I appreciate the support provided by John Christensen that enabled me to have the space and time to develop and document the ideas contained in the article. Lastly, I cannot forget to thank Tim James for language editing.

SEE ALSO THE FOLLOWING ARTICLES

Cultural Evolution and Energy • Development and Energy, Overview • Economic Growth and Energy • Energy Ladder in Developing Nations • Environmental Injustices of Energy Facilities • Indoor Air Quality in Developing Nations • Sustainable Development: Basic Concepts and Application to Energy • Technology Innovation and Energy • Transitions in Energy Use • Wood in Household Energy Use

Further Reading

Ballard-Tremeer, G., and Mathee, A. (2000). Review of interventions to reduce the exposure of women and young children to indoor air pollution in developing countries. Paper presented at the USAID and WHO Global Consultation, "Health Impacts of Indoor Air Pollution and Household Energy in Developing countries," Washington, DC.

Barnes, B., Mathee, A., Shafritz, L.B., and Krieger, L. (2000). "A Behavioural Intervention to Reduce the Impact of Indoor Air Pollution on Child Respiratory Health: Identifying Target Behaviours." Medical Research Center of South Africa.

Batliwala, S., and Reddy, A. K. N. (1996). Energy for women and women for energy: A proposal for women's energy entrepreneurship. *ENERGIA* **1,** 8. (International Network on Women and Energy newsletter).

Boy, E., Bruce, N., and Delgado, H. (2002). Birth weight and exposure to kitchen wood smoke during pregnancy in rural Guatemala. *Environ. Health Persp.* **110,** 109–114.

Bradley, P. N., and Campbell, B. M. (1998). Who plugged the gap? Re-examining the woodfuel crisis in Zimbabwe. *Energy Environ.* **9,** 235–255.

Britt, C., Kapoor, S. (1994). "The Effects of Biogas on Women Workloads in Hathilet, Janakpoor Zone, Nepal: An Overview of Five Studies Conducted for Biogas Support Programme." SNV–Nepal, Kathmandu.

Cecelski, E. (2000). "The Role of Women in Sustainable Energy Development," NREL/SR-550-26889. National Renewable Energy Laboratory, Golden, CO.

Clancy, J. (1999). Policies, projects, and the market empowering women? Some initial reactions to developments in the energy sector. Paper presented at the *ENERGIA* Workshop on Improving Women's Access to the Market, "Policies, Projects, or Market," University of Twente, The Netherlands.

Edwards, R. D., Smith, K. R., Zhang, J., and Ma, Y. (2003). Implications of changes in household stoves and fuel use in China. *Energy Policy* **32,** 395–411.

Energy Sector Management Assistance Program. (2002). "Executive Summary: Energy Strategies for Rural India—Evidence from Six States." ESMAP–World Bank, Washington, DC.

Foley, G. (2001). "Sustainable Woodfuel Supplies from the Dry Tropical Woodlands." Energy Sector Management Assistance Program, Washington, DC.

Jackson, C. (1993). Doing what comes naturally? Women and environment in development. *World Dev.* **21,** 1947–1963.

Lang, Q., Chapman, R. S., Schreinemachers, D. M., Tian, L., and He, X. (2002). Household stove improvement and risk of lung cancer in Xuanwei, China. *J. Natl. Cancer Inst.* **94,** 826–834.

Laxmi, V., Parikh, J., Karmakar, S., and Dabrase, P. (2003). Household energy, women's hardship, and health impacts in rural Rajasthan, India: Need for sustainable energy solutions. *Energy Sustainable Dev.* **7**(1), 50–68.

Leach, G., and Gowen, M. (1987). "Household Energy Handbook: An Interim Guide and Reference Manual." World Bank, Washington, DC.

Mathee, A., and de Wet, T. (2001). Rural electrification in South Africa: Implications for the health and quality of life of women. *ENERGIA* 4(4). (International Network on Women and Energy newsletter).

Mathee, A., Rollin, H., Bruce, N., and Levin, J. (2001). "Indoor Air Pollution, Household Energy, and Health" [feasibility study]. Report prepared for the World Health Organization.

Mehlwana, A. M. (1997). The anthropology of fuels: Situational analysis and energy use in Urban low-income townships of South Africa. *Energy Sustainable Dev.* **3**(5), 5–15.

Njenga, B. K. (2001). *Upesi* rural stoves project: Kenya. *In* "Generating Opportunities: Case Studies on Women and energy" (S. Misana and G. K. Karlsson, Eds.), pp. 45–51. UN Development Program, New York.

Parikh, J. (2002). Mainstreaming gender and energy in South Asia. Regional paper prepared for the World Summit on Sustainable Development.

Reddy, A. K. N. (2002). The evolution of an energy analyst: Some personal reflections. *Annu. Rev. Energy Environ.* **27,** 23–56.

Skutsch, M. (1995). "Gender in Energy: How Some Major Donors See It," working paper 71. University of Twente, Netherlands.

Smith, K. R. (2002). Indoor air pollution in developing countries: Recommendations for research. *Indoor Air* **12,** 198–207.

Sweetman, C. (2001). "Men's Involvement in Gender and Development Policy and Practice: Beyond Rhetoric." OXFAM, London.

United Nations Environment Program. (2001). "Experience with PV Systems in Africa: Summaries of Selected Cases" (N. Wamukonya, Ed.). UNEP, New York.

Villavicencio, A. (2001). "Sustainable Energy Development: The Case of Photovoltaic Home Systems," draft UCCEE working paper. UCCEE–Risoe National Laboratory, Denmark.

Wamukonya, N., and Davis, M. (2001). Socio-economic impacts of rural electrification in Namibia: Comparisons between grid, solar, and unelectrified households. *Energy Sustainable Dev. J.* **5**(3), 5–13.

World Energy Council. (2000). "Energy and the Challenge of Sustainability." WEC–UN Development Program, New York.

Wood Energy, History of

JOHN PERLIN
Rahus Institute
Martinez, California, United States

1. **The Importance of Wood in Human Evolution**
2. **The Importance of Wood in the Development of Civilization**
3. **The Importance of Wood Appreciated by Our Antecedents**
4. **Examples of Wood Used for Fuel Over the Millennium**

Glossary

Bronze Age A period between the Stone and Iron Ages characterized by the manufacture and use of bronze tools and weapons. This smelting required huge volumes of wood.

Mycenaean Age in Greece (**600–1100** BC) A period of high cultural achievement, forming the backdrop and basis for subsequent myths of the heroes. It was named for the kingdom of Mycenae and the archaeological site where fabulous works in gold were unearthed. The Mycenaean Age was cut short by widespread destruction ushering in the Greek Dark Age.

smelting To melt or fuse a metal-bearing ore for the purpose of separating and refining the metal.

Wood was the primary fuel for the entire world from the discovery of fire to the age of fossil fuels. It remains the primary fuel to this day for those living in much of the developing world.

1. THE IMPORTANCE OF WOOD IN HUMAN EVOLUTION

It may seem bold to assert wood's crucial place in human evolution. However, consider that trees have provided the material to make fire, the key to humanity's domination over all other animals. It allowed *Homo erectus*, from which *Homo sapiens* evolved, to migrate from their warm niche in equatorial Africa to lands throughout the Old World, where heat from fire made these colder climates habitable. Increasing the range of habitat gave humanity better odds for survival. Light from wood fires permitted *H. erectus* to work after nightfall. Hence, the discovery freed humanity from the strictures of the diurnal cycle. Wood fires also enhanced humanity's access to food. With torches, *H. erectus* hunted at night. This gave them the element of surprise in killing diurnal animals and added nocturnal animals to their diet. Fire also widened the types of available food, freeing them from the constraints of the hunt. Vegetable matter such as tubers, when cooked over fire, became eatable. A diet based on local plants promoted population stability because humans no longer had to follow migrating herds to eat. A wider choice of food also enhanced survival and proliferation (Fig. 1). Fire also provided humans with greater security. They could safely live indoors out of harms way because they had the means of illumination. Lighting also warded off nocturnal predators. It lessened the chance of stepping on venomous snakes or tripping over rocks by improving vision at night. Wood fires changed not only the social evolution of humanity but also its physical development. Eating food softened by cooking led to smaller molars. Although many animals, such as chimpanzees and otters, have used tools as aids, no other animal except the human genus beginning with *H. erectus* has ever built fires.

2. THE IMPORTANCE OF WOOD IN THE DEVELOPMENT OF CIVILIZATION

Only by cooking grains over wood-fueled fires do grains become edible, thus enabling the agricultural revolution of the past 10,000 years (Figs 2 and 3). In charcoal-fueled kilns, potters turned earth into

Encyclopedia of Energy, Volume 6. Text copyright © 2004 Elsevier Inc. All rights reserved.
Figures copyright © John Perlin.

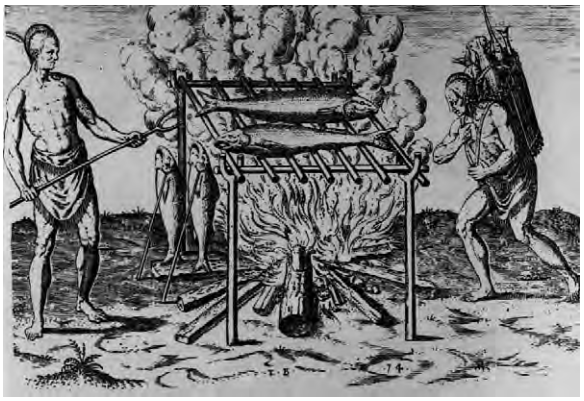


FIGURE 1 Native Americans using wood for cooking. From Perlin Forestry Archives.

FIGURE 2 Medieval people carry firewood from the forest to their hearths. From Perlin Forestry Archives.

durable ceramics in which goods could be stored and shipped. Providing an efficient means of storage stimulated trade on land and sea. Charcoal-fueled fires also gave humanity the capability to extract metal from ore, resulting in tools and weaponry so revolutionary that historians categorize society's evolution according to the dominant metal in use at the time. Wood fuel, in fact, is the unsung hero of the technological developments that brought us from a bone and stone culture to the industrial revolution.

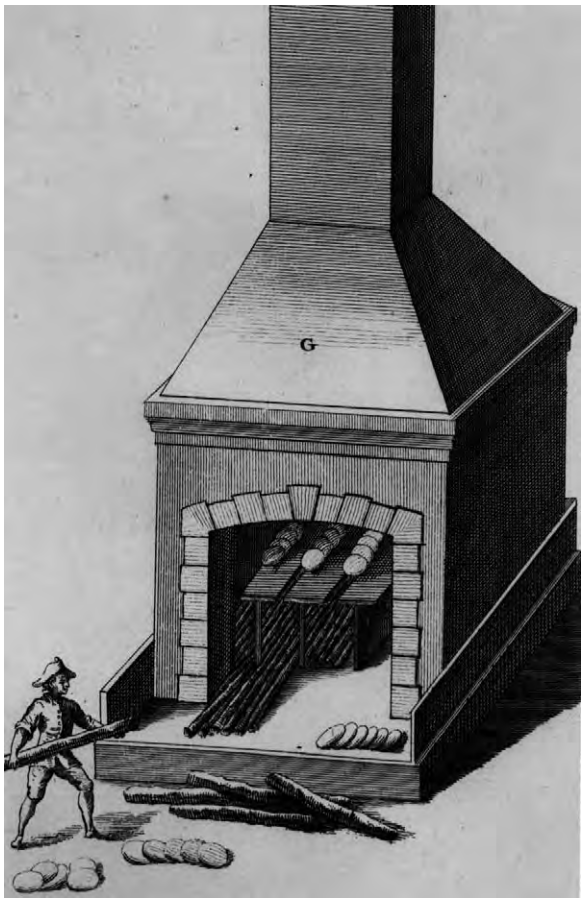


FIGURE 3 Wood fires turned grains into eatable bread. From Perlin Forestry Archives.

3. THE IMPORTANCE OF WOOD APPRECIATED BY OUR ANTECEDENTS

Lucretius, probably the most scientifically oriented thinker of Antiquity, believed that the discovery of fire made civilization possible. He conjectured that the technique of metallurgy was born when great fires "devoured the high forests and thoroughly heated the earth," melting metal embedded in rock. When people saw the hardened metal, "the thought then came to them," according to Lucretius, "that these pieces could be made liquid by heat and cast into the form and shape of anything, and then by hammering, could be drawn into the form of blades as sharp and thin as one pleased, so they might equip themselves with tools." Lucretius remarked that tools, in turn, made forestry and carpentry possible, enabling humans "to cut forests, hew timber, smooth, and even fashion it with auger, chisel, and

gouge." By this process, in Lucretius' opinion, civilization emerged.

Ibn Khaldun, writing in the 14th century AD, discussed the crucial role that wood fuel played in the world of Islam. In "The Muqaddimah," his major work, he noted that "God made all created things useful for man so as to supply his necessities and needs. Trees belong among these things. They give humanity its fuel to make fires, which it needs to survive."

The English of the 16th and 17th centuries also recognized the crucial role of wood fuel in their lives. Gabriel Plattes, writing in 1639, observed that all tools and instruments are made of wood and iron. Upon weighing the relative importance of the two materials, he chose wood over iron because without wood for fuel "no iron can be provided."

4. EXAMPLES OF WOOD USED FOR FUEL OVER THE MILLENNIUM

4.1 Bronze Age Greece and Cyprus

Mycenaean Greece attained an unprecedented level of material growth during the late Bronze Age. However, it did not have much copper to make sufficient quantities of bronze, the essential metal of the time. To sustain the booming economy of the Mycenaean states, abundant and reliable sources of copper had to be found. Fortunately, there was one area close by, Cyprus, that had lots of copper ore and plenty of wood with which to smelt and refine it. In response to this need, the Cypriots smelted as much copper ore as they possibly could for the overseas market. However, the growing demand put a great burden on the island's woods because charcoal was the fuel for smelting and refining copper. One hundred and twenty pine trees were required to prepare 6 tons of charcoal needed to produce 1 ingot of copper, deforesting approximately 4 acres. Underwater archaeologists found on board a Bronze Age shipwreck 200 ingots of copper that had been mined and smelted in Cyprus. The production of just this shipload cost the island approximately 24,000 pine trees. The lively commerce in ingots during the 14th and 13th centuries BC surely consisted of many such shipments and the concomitant deforestation of a large expanse of woodlands. The industry's consumption of wood deforested approximately 4 or 5 square miles of woods per year. Another 4 or 5 square miles of forest was cut to supply fuel for heating and cooking and for other industries, such as pottery works and lime kilns. The cumulative effect of deforestation on such a scale must have been felt quite soon on an island of only 3600 square miles.

Cutting trees for fuel apparently made significant changes in the flora of Cyprus during the late Bronze Age. Bronze axes found dating to this period suggest that logging was an important occupation. Those wielding the axes cleared the majority of forest along the coast and hillsides near the larger late Bronze Age sites in Cyprus. As a result, pigs, which thrive in a moist, woody habitat, could no longer survive, giving way to sheep and goat tending, which flourish in a relatively barren environment.

With the forest cover gone, natural calamities struck. Silting of major harbors and an increase in flooding and mudslides in urban areas ensued. The Tremithos River, for example, transported tons of soil from the deforested hills below the Troodos Massif and deposited them into the Mediterranean. Currents moved these deposits of sediment toward the important late Bronze Age town of Hala Sultan Tekke. So much alluvium accumulated near the mouth of its harbor that during the 12th century BC. Hala Sultan Tekke was sealed off from the sea and could no longer function as a port.

Similarly, excessive amounts of alluvium carried by the Pedieos River, which flowed through Enkomi, a major Bronze Age city that faced the Near East, formed a delta as it drained into the ocean, changing Enkomi from a coastal to inland city. The number and severity of floods and mudslides also increased as a consequence of deforestation, periodically transforming the streets of the city into raging torrents full of mud and other debris.

With wood difficult to come by, interest heightened in the search for ways to stretch limited supplies, resulting in technological advances and recycling. The adoption of hydrometallurgy to prepare copper ore for smelting appears to be one of the strategies chosen by late Bronze Age Cypriot metalworkers to conserve fuel. The process required the exposure of mined ore to the elements so that the ambient moisture would leach impurities. The leached ore could be directly smelted, circumventing an initial roast and reducing the number of smelts needed to reduce the ore. Hence, the amount of fuel expended in this phase was decreased by one-third. In another move to save fuel, metallurgists collected old and broken tools in order to resmelt them. Recycling bronze became a major source of the metal during the period of wood depletion on Cyprus.

Despite such highly innovative ways to save energy, metallurgists could not sustain the high productivity

of the previous two centuries. Copper production peaked in approximately 1200 BC, and the last copper furnaces were shut down in 1050 BC. During the same time period, 90% of the island's settlements were abandoned and the population withered away, as did the economy and material culture.

The depressed state of the Cypriot copper industry apparently affected the entire eastern Mediterranean region. Bronze was in such short supply that the smiths throughout the area bought whatever pieces they could find, whether virgin ingots, ingots composed of recycled bronze, or scrap.

Often, smiths in southern Europe could obtain only minute quantities of bronze because of copper shortages. Such was the case in Greek Messenia in approximately 1200 BC. With so little bronze available, everyone suffered. Without metal hoes, plowshares, and scythes, farmers were far less productive. Without axes, adzes, and saws, few ships could be built for commerce or war. Without enough arrowheads, spearheads, and blades for swords, soldiers were no match for a well-armed foe.

Taking advantage of Messenia's vulnerability, a group of insurgents overpowered local forces and destroyed the palace at its capital, Pylos. After the catastrophe, the populace did not attempt to rebuild. Events at Messenia presaged trouble that other societies in the eastern Mediterranean would soon face. Just as bronze gave these civilizations the material to expand to heights never before attained, conversely, the lack of bronze played a major role in their demise.

In this sea of troubles, new hope for future generations arose. The fuel crisis that had caused copper production to decline served as the incentive for metallurgists to begin working with iron. Because Bronze Age Cypriots smelted copper ore that contained 10 times more iron than copper, metallurgists could obtain more useable iron than copper with the same investment of fuel. Hence, common sense dictated switching to iron smelting when fuel was at a premium. Furthermore, the refuse from Cypriot copper smelting contained significant amounts of iron. As long as there had been plenty of fuel, metallurgists smelted virgin ore and ignored the slag that had accumulated. When fuel became scarce and forced production cuts, metallurgists began to mine their industrial garbage. They discovered that the slag contained a great amount of iron, which could be removed simply by hammering. Taking iron out of the slag manually permitted metallurgists to bypass any processes requiring fuel and still obtain usable metal. The labor-intensive nature of working in this manner drastically reduced the overall output of metal on Cyprus and yielded relatively small amounts of iron. However, the success metallurgists had in working with iron at this early stage laid the foundation for the coming of the Iron Age in the Mediterranean as well as the rest of Europe.

4.2 Classical Greece

In classical Greece, the silver smelted at Laurion, southeast of Athens, near the sea, paid for the Athenian fleet that defeated the Persians at Salamis, turning the tide in the war against the Persians. Coins minted with Laurion's silver were accepted as the currency of the Greek world. With its treasury full of bullion, the Athenian economy could well afford to spend lavishly. Miners did not recover pure silver at Laurion but galena ore. Once mined, metallurgists had to heat the ore to very high temperatures to extract the silver. Charcoal was the fuel metallurgists at Laurion used in the smelting process and trees, or course, had to be cut down to produce this charcoal (Fig. 4). To extract the silver from ore at Laurion required the burning of more than 24 million pines or more than 52 million oaks. The largest outlays of fuel occurred during the two most active periods of mining, from 482 to 404 BC and from the second decade of the fourth century BC to its end. The surrounding areas could supply only a fraction of Laurion's fuel, forcing metallurgists to locate their furnaces on the coast so they could easily receive fuel imported from Thrace by ship.

4.3 Classical Rome

Rome financed its growth largely with silver extracted from Spanish ore. Production increased

FIGURE 4 Ancient Greek urn depicts a worker stoking a metallurgical furnace with charcoal. From Perlin Forestry Archives.

considerably during the end of the republic and the first years of the empire. However, this was accomplished only by great expense to the Iberian woodlands since silver smelting consumed more than 500 million trees during the 400 years of operation. Woodsmen had to deforest more than 7000 square miles to provide fuel for the furnaces. Near the end of the period of peak production, the need to sustain high output so strained the area's fuel supplies that it merited intervention by the Roman state. Under the reign of the emperor Vespasian, the Roman government directed all mining areas of southwestern Spain to prohibit the sale of burnable wood by those who ran bathhouses in the region (Fig. 5).

To produce enough silver to support the habits of a succession of rulers who spent as extravagantly as Caligula and Nero, a time had to come when the tree supply in Spain would dwindle and production in silver would decline accordingly. Conservation laws could only temporarily stave off wood shortages when silver was spent so wastefully. At approximately the end of the second century AD, the inevitable occurred: Silver production declined. Further output was limited not by the supply of ore, which remained abundant, but by the accessibility of fuel.

The decline in silver production offered later emperors two choices: cut expenditures or find alternative financing. They unanimously chose the latter but differed in methodology. The emperor Commodus "stretched" silver money by adding base metal, which comprised 30% of the coin. He also went on a killing spree, enraged that the empire's revenues could not meet his expenditures. When he finally calmed down, he decided to auction off whatever he could, offering provincial and administrative offices to the highest bidder.

Septimius Severus, who ruled a few years after Commodus, added 20% more alloy to the silver coinage, thus reducing the silver content to a mere 50%. Because Roman money was now so badly debased, Severus instituted the requisitioning of commodities rather than collecting worthless currency through taxation. Further debasements forced the government to search for creative ways of staying afloat. Most of the methods chosen circumscribed the freedom of Rome's citizens. Providing the government with the provisions it needed became compulsory. The government also established guilds, expecting them to produce according to obligations it set but rewarding members with monopolies in their respective trades.

FIGURE 5 Longshoremen loading a log destined to fuel the baths of Rome. From Perlin Forestry Archives.

By the end of the third century AD, Rome's currency had lost 98% of its silver content, and the public placed as little value on it as did the government. People increasingly took to trading in commodities and services so that by the first part of the fourth century AD barter had become institutionalized.

To keep those in Rome from becoming too anxious over the declining economy, the rulers of Rome had to constantly find ways to keep the population placated. The later emperors were well aware of the Romans' love of bathing and added many new baths to the city, eventually bringing the total to more than 900. The largest held as many as 2000 bathers at a time. Bathwater had to be hot if the Romans were to remain happy. Because keeping the Roman populace satisfied was paramount in the minds of those in power, the authorities were willing to go to great lengths to ensure a constant flow of fuel to the bathing establishments.

In the third century AD, the emperor Severus Alexander saw to it that entire woods were cut down to keep the baths in Rome well heated. When these forests gave out a century later, the authorities founded a guild with 60 ships at its disposal that was solely responsible for supplying the baths with wood. Sometimes, wood could be obtained as close to home as the Campania region. Usually, though, the guild had to make its wood runs primarily to the forests in North Africa. That Romans would travel such great distances for fuel indicates just how little wood was left in its vicinity and how dependent the Romans had become on foreign supplies.

Interestingly, the rise and decline of fuel supplies in Rome closely parallel the fortunes of the empire. The pioneering ecologist George Perkins Marsh demonstrated the fact by describing the changes in brick and masonry work in Rome over the centuries. Bricks in early buildings were extremely thin, well fired, and held together by liberally applied quantities of lime mortar. In contrast, as the imperial

period progressed, the opposite proved true: Bricks were very thick, usually poorly fired, and held together by a minimum of mortar. Marsh hypothesized that the difference was "due to the abundance and cheapness of fuel in early times, and its growing scarceness and dearness in later ages." He elaborated on his observation:

> *When wood cost little, constructors could afford to burn their brick thoroughly and burn and use a great quantity of lime. As the price of firewood advanced, they were able to consume less fuel in brick and lime kilns and the quality and quantity of brick and lime used in building were gradually reversed in proportion.*

4.4 England, 1500s–1700s

At the beginning of the 16th century, England depended on the Continent for its shot and artillery. However, Henry VIII's apostasy resulted in an arms embargo and a threat of invasion. Self-sufficiency seemed the only recourse. Henry therefore saw to the development of a local arms industry in Sussex, where rich veins of high-phosphorous iron were particularly advantageous for the casting of guns and where oak and beech would provide ample fuel for the gun founders. They did not smelt their iron in simple hearths but used blast furnaces and an accompanying forge. They chose the blast furnace and forge over the simpler hearths because the former could produce approximately 20 times more iron than the latter. The blast furnace and forge also consumed much more charcoal (Figures 6 and 7). By the late 1540s, the new English arms industry annually consumed approximately 117,000 cords of wood, causing great destruction of nearby woods. People living in the vicinity viewed these blast furnaces and forges as menacing since the wood they consumed was essential to the locals' survival, who used it to heat their homes, cook their meals, make their tools and fishing boats, and build and repair their houses. By the 1700s, the iron industry had deforested such a large area of southeast England that the iron masters had to ration the amount of iron ore they could smelt. To increase iron production required learning to smelt ore with coal without adding its impurities to the finished metal. By discovering how to make coal as close to charcoal as possible, the English freed themselves from the constraints of its dwindling forests and began the industrial revolution, an age that has qualitatively separated people living since the mid-19th century from those throughout the rest of history.

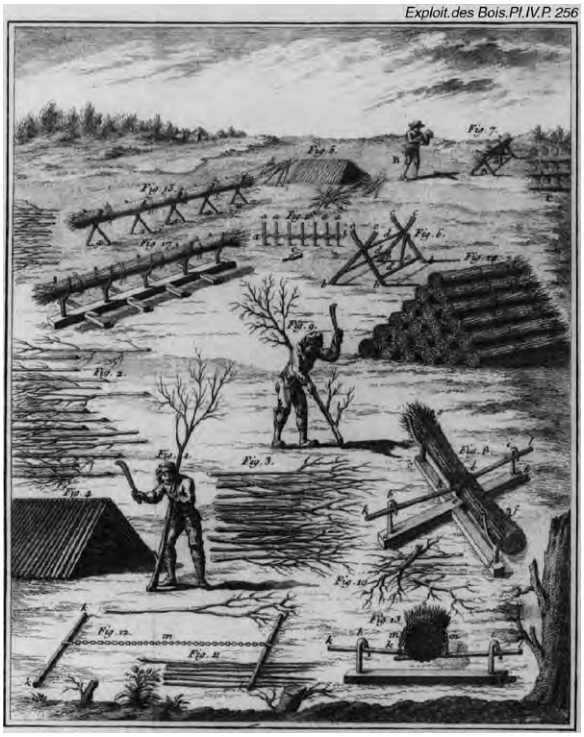


FIGURE 6 Wood cutters chop off tree limbs to be used as charcoal. From Perlin Forestry Archives.

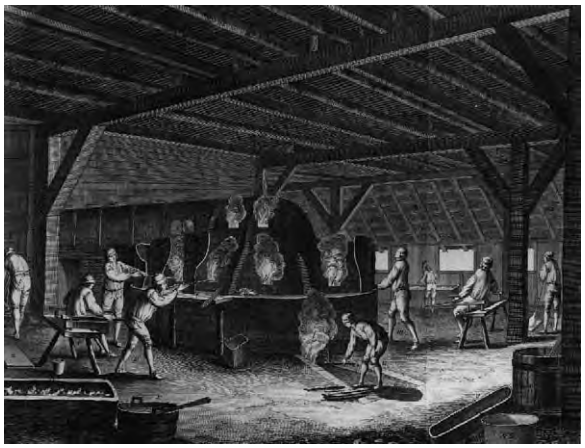


FIGURE 7 Charcoal fires also fueled English glassworks. From Perlin Forestry Archives.

4.5 Madeira, West Indies, and Brazil, 1400s–1600s

Sugar brought great wealth to those who raised the commodity. Hence, when the Portuguese landed on the warm, greatly forested, fertile, and well-watered island of Madeira, they immediately started to plant cane. Once the juice was forced out of the stalks, workers poured the extracted liquid into kettles.

A fire, rarely extinguished, burned underneath each pot. The wood from linden trees furnished much of the fuel. No doubt cedar was also used, being the most common wood on Madeira. Sugar workers aptly called the room in which the cane juice was boiled the "sweet inferno." The juice continued boiling until judged ready for removal to an area where it would solidify into sugar. By the end of the 15th century, the island's sugar industry required approximately 60,000 tons of wood just for boiling the cane. Four of the 16 mills operating on Madeira consumed 80,000 pack animal loads of wood per year.

Almost immediately after the discovery of the West Indies and South America, Europeans recognized that the land offered the same ideal conditions for growing and processing sugar as existed on Madeira. The Spanish planted sugarcane on Espanola soon after Columbus's first journey to the New World. As expected, the cane flourished, and by the end of the 16th century 40 sugar mills were operating on the island. It was no different in other parts of the Indies, and sugarcane proliferated and became the area's main source of revenue. Nor did the Portuguese, after their success on Madeira, wait long to establish sugar plantations after taking possession of Brazil. Near the end of the 16th century, a visitor reported 70 sugar mills at work in the Pernambuco region and 40 in the area of Bahia. Experts in sugar production in the New World estimated that six to eight slaves had to be constantly employed in cutting fuel in the forest and transporting it to the mills for optimum efficiency. Providing fuel for 1 mill stripped approximately 90 acres of forestland each year.

Such large-scale consumption of wood took its toll on the forests of the New World. Two hundred and forty years after the Portuguese had found Madeira so thickly wooded that they called it "isola de Madeira" (island of timber), it had become the island of wood in name only, so deforested that those passing it by ship could find absolutely no greenery in view. Likewise, the very thickly forested island of Espanola had become an open land. The pace of deforestation on Barbados exceeded that on Espanola and Madeira. In little more than 20 years, the representatives of the planters admitted to having used and destroyed all the timber formerly growing on the island.

4.6 American Colonies

America's great forests offered the pilgrims a higher standard of living than that of their peers in England since they had upon arrival more wood fuel with which to heat than did most noblemen in England. Families also used the enormous fireplace around which they huddled for lighting. The fireplace was so huge as to require logs of dimensions that could only be dragged into the house by a horse or oxen.

America's plethora of wood gave rise to a large and prosperous iron industry. Despite the fact that American woodcutters earned three times more than their counterparts in England and Americans who coaled the wood made approximately twice as much as those who did the same work in England, because a cord of wood cost 14 times more in England than in America, American ironmasters could manufacture pig iron in the Colonies and export it to England, underselling English iron.

Iron furnaces and forges first began operating in the backwoods of Virginia and Maryland and then spread to Pennsylvania and New Jersey. They not only produced iron for the English market but also produced cast iron goods for local use. The proliferation of home-produced iron goods frightened the English, whose mercantile economy was based on buying raw materials from the Colonies at bargain prices and selling back finished products at much higher prices. The English therefore passed a law in 1749 that encouraged the importation of pig iron from the American Colonies, because wood shortages had hindered England's iron production, while forbidding the Americans from developing forges to turn out finished iron products.

The act stirred already growing feelings for independence among many of the colonists. It hardly seemed fair or rational to Americans that they had to send to England the iron they produced in order to buy it back again as manufactured items, costing them almost twice as much compared to if they done the entire process at home. The intent of the act reminded many of the servile conditions forced upon the Children of Israel by the Philistines when the Jews were not allowed any smiths of their own. The prohibitory nature of the act aroused revolutionary rhetoric among the more radical Americans such as James Otis, credited by John Adams as the father of the idea of independence, and John Dickinson, who later authored revolutionary America's "Declaration of the Causes of Taking Up Arms." Both feared that prohibiting the manufacture of iron goods was the first step in stopping the construction of any machinery that would further develop the Colonies' economy.

Despite the anger the act aroused in America, it did increase American exports of pig iron to

England. By 1769, approximately half of the pig iron worked in English forges came from America. By 1776, as much iron was produced in the Colonies as in Britain. Even more encouraging to those favoring revolution, and frightening to those wishing to hold onto the American Colonies, was the fact that not only did America's innumerable iron mines and endless forests give it the very sinews of power but also a certain type of iron produced in America, called best principio, was judged as good as any in the world for making shot and cannon.

4.7 America after Independence

Alexander Hamilton, probably the greatest advocate of industrialization during the early republic, predicted that America was very fortunate to have vast supplies of iron ore and wood with which to make charcoal. Trench Coxe, a colleague of Thomas Jefferson and James Madison, predicted that America's rich store of wood fuel would provide invaluable to the rise of industry and turn America into the most powerful country in the world.

Breweries, distilleries, salt and potash works, and casting and steel furnaces all needed heat to produce finished products. Francois Michaux, a French botanist who studied America's trees at the beginning of the 18th century, found that the bakers and brick makers of New York, Philadelphia, and Baltimore commonly consumed prodigious quantities of pitch pine. Hatters of Pittsburgh, on the other hand, preferred charcoal made from white maple. Boats sailed along the Erie Canal picking up wood to fuel the nation's largest salt works, located in upstate New York. Boiling rooms, in which the salt water was evaporated over charcoal fires, produced 2 million bushels of salt per year. The salt went to Canada, Michigan, Chicago, and all points west. Farmers were the largest purchasers, using the salt for preserving meat that they marketed. Steam engines, which in the 1830s began to replace factories run on water power, usually burned wood as their fuel.

Americans could see the Hand of Providence at work in providing great amounts of wood and iron next to each other. Wood- and iron-rich western Pennsylvania, northern Michigan, and southeast Ohio became the principal producers of iron in 19th-century America. Throughout the nation, charcoal-burning iron mills produced 19 million tons of iron between 1830 and 1890. To truly appreciate the magnitude of the output of America's wood-fueled iron furnaces, comparison to the amount of iron produced during the heyday of the British charcoal-burning irons industry, which dated from the 1640s to the 1780s, is in order. In its most productive years, England's charcoal-run ironworks produced approximately 1 million tons of iron in 60 years.

The steamboat opened the old American west (lands east of the Mississippi River, north of the Ohio River, and west of the Allegheny Mountains) to settlement (Fig. 8). How else could farmers in places such as Ohio, Indiana, and Illinois expect to move their goods to distant markets except by navigation? Had farmers had to rely on land transport instead, they would have needed 5000 miles of good road, which did not exist at the time, and 250,000 wagons to transport the same amount of goods annually carried by steamboat. Steamboats were lauded by mid-19th-century pundits as the most important technological development to have occurred in the United States because they contributed more than any other single development to advance the prosperity of those living in the Mississippi and Ohio River basins.

Great quantities of wood were needed for fuel. The large steamboat Eclipse's 15 boilers consumed wood by the carload. Thousands of wood yards were located along the banks of every navigable river simply to provide steamers with fuel. Backwoodsmen brought the timber they had cut to these depots and hacked it into the proper size for the ships' furnaces. At night, the owners of these wood lots kept gigantic

FIGURE 8 Hauling timber to fuel steamboats. From Perlin Forestry Archives.

fires blazing so those onboard the boats could see the yards and fuel up.

Railroads liberated Americans from their dependence on waterways for shipping freight and personal travel. They slashed the cost of and the time required to travel. With plenty of timber growing along the right-of-way of most railroads, or at least close by, locomotives used wood as their only fuel from when the first tracks were laid in 1830 to the beginning of the Civil War. Fuel needs of engines owned by the New York Central Line required the railroad to build 115 woodsheds along its track. If the woodsheds were stacked side by side, they would have stretched approximately 5 miles. A wood yard for trains at Columbus, Nebraska, was half a mile long.

Between 1810 and 1867, approximately 5 billion cords of wood was consumed for fuel in fireplaces, industrial furnaces, steamboats, and for railroads. To obtain such a quantity of wood meant the destruction of approximately 200,000 square miles of forest land, an area nearly equal to all the land that comprises the states of Illinois, Michigan, Ohio, and Wisconsin.

4.8 Use of Wood for Fuel, the Present

Today, half of the wood cut throughout the world is used for fuel. Eighty percent of all wood used for fuel is consumed in the developing world. In fact, the majority of people in the developing world depend on wood as their primary energy source. Charcoal and firewood in the Cameroons, for example, account for 80% of all energy consumption. The demand for firewood and charcoal is also increasing as the population increases. In Africa, the amount of wood consumed for energy purposes increased from 250 million m^3 in 1970 to 502 million m^3 in 1994. In Latin America, the vast majority of the rural population uses wood as its primary fuel. The true danger to the forests of the developing world is that growing numbers of people exploit local forests, threatening their viability and forcing people to travel greater distances for wood fuel. One person can ably supply himself with firewood from 0.5 ha of forestland; today, however, usually approximately 15 people try to supply themselves with firewood from 1 ha of woods. Such overexploitation will surely cause local supplies to quickly diminish, forcing people to seek wood in more distant forests. Demand for wood and charcoal in growing urban areas of the developing world has introduced a new and growing problem—the industrialization of acquiring firewood and charcoal. Instead of individuals fanning neighboring woods to cut down trees for fuel, well-capitalized charcoal dealers search throughout the countryside for supplies. Bangkok's 5 million citizens obtain a large amount of their charcoal in this manner from forests throughout Thailand. There is no question that as the population in the developing world grows, so too will the demand for firewood and charcoal and deforestation will continue to accelerate.

SEE ALSO THE FOLLOWING ARTICLES

Coal Industry, History of • Depletion and Valuation of Energy Resources • Early Industrial World, Energy Flow in • Electricity Use, History of • Fire: A Socioecological and Historical Survey • Forest Products and Energy • Hunting and Gathering Societies, Energy Flows in • Hydropower, History and Technology of • Manufactured Gas, History of • Natural Gas, History of • Solar Energy, History of • Wind Energy, History of

Further Reading

Bryant, D., Nielsen, D., and Tangley, L. (1997). "The Last Frontier Forests: Ecosystems and Economies on the Edge." World Resources Institute, Washington, DC.

Myers, N. (1980). The present status and future prospects of tropical moist forests. *Environ. Conserv.* 7(2), 101–114.

Perlin, J. (1991). "A Forest Journey: The Role of Wood in the Development of Civilization." Harvard Univ. Press, Cambridge, MA.

Richards, J., and Tucker, R. (eds.). (1988). "World Deforestation in the Twentieth Century." Duke Univ. Press, Durham, NC.

Wood in Household Energy Use

ROBERT BAILIS
University of California, Berkeley
Berkeley, California, United States

1. Introduction
2. Wood Burning at the Household Level
3. Alternatives to Wood Burning and Possible Future Directions for Households in Less Developed Countries

Glossary

biomass energy Energy derived from any nonfossil organic fuel; sources of biomass energy include wood and wood by-products (e.g., charcoal) as well as nonwoody materials (e.g., crop residues, human or animal waste).
charcoal A wood product that is made by heating wood in the absence of sufficient air for full combustion to occur; heating releases the wood's volatile compounds, leaving behind a lightweight and cleaner burning fuel that is 70 to 90% carbon.
cooking stove Any device used for cooking food; a cooking stove can range from an open wood-burning fire to a gas burner.
heating value The energy that is released by ideal combustion per unit mass of fuel; for any fuel, heating value can be reported in two quantities: higher or lower heating value.
liquefied petroleum gas (LPG) A gaseous mix of light petroleum distillates (primarily butane and propane) that is liquefied at high pressure; it is a high heating value fuel used for cooking.

Roughly 40% of the world's population, highly concentrated in less developed countries, use woodfuels as their principle source of household energy. This prevalence of woodfuel consumption is associated with a number of socioeconomic and environmental issues that relate to poverty and gender, land use change and deforestation, indoor air pollution and public health, as well as climate change and greenhouse gas emissions. These factors have been the impetus for a wide range of interventions into the way that people in less developed countries meet their energy needs. This article discusses these issues and the interventions that have been attempted over the past two decades.

1. INTRODUCTION

1.1 Wood as a Source of Household Energy

Since the beginning of human history, people have relied on wood to provide energy for subsistence. As people mastered wood burning, it became a valuable tool for industrial uses as well as basic household needs. Centuries of technical progress have seen many different fuels brought into use that are preferable to wood for a number of reasons. However, widespread changes in the industrial sector have not been mirrored in the residential sector. Between one-third and one-half of the world's population continues to use wood and other biomass fuels to meet their energy needs. The use of wood as a household fuel is overwhelmingly concentrated in less developed countries where, for a number of reasons that are explored in this article, alternative fuels such as natural gas, kerosene, liquefied petroleum gas (LPG), and electricity have not been widely adopted.

The implications of such widespread wood consumption extend across a range of social and environmental issues. Household woodfuel consumption is associated with social and environmental problems, including health problems associated with indoor air pollution (IAP) and environmental change associated with both greenhouse gas (GHG) emissions and the depletion of local forests and woodlands. In addition, woodfuel use creates high labor demands and potentially hazardous conditions for household members. However, it is important to recognize that a direct causal link between woodfuel consumption and any one of these problems does not always exist. These problems rarely arise as a result of wood consumption alone; rather, they are the

 © 2004 Elsevier Inc. All rights reserved.

result of complex relationships among wood consumers, the environment in which they live, and the larger political economy. Therefore, understanding the problems associated with household woodfuel consumption is possible only by considering the social, political–economic, and environmental contexts in which they arise.

The aim of this chapter is to examine the use of wood as a source of household energy, focusing on the less developed countries where it is most prevalent. The article discusses some of the larger socioeconomic and environmental issues to which household energy is linked and explains why wood remains one of the most common household fuels for so many people. It closes with an examination of some possible improvements and alternatives to common wood-burning practices that can help to mitigate some of the negative impacts associated with woodfuel consumption.

1.2 The Household Energy Sector and Household Fuel Choice

Household energy use consists of energy used for space heating, water heating, and cooking. In locations where electricity is available, households also use energy for running a number of household appliances for washing, preserving food, and entertaining the family. In some countries, households also use energy for productive activities, including brewing beer, firing bricks, and cooking or drying food for commercial sale. These activities are not usually part of the formal economy, and the energy consumed is often unaccounted for in national energy balances, but such activities constitute a significant fraction of energy consumption within communities where they are practiced.

Globally, the demand for residential energy depends on many factors, including climate, locality, the level of wealth of the household, and the availability of resources. The fuels and the technology used to convert the fuels' energy for use in the household are equally variable. Even in less developed countries where the majority of households rely on wood or other biomass fuels, the variety and performance of cooking stoves are surprisingly diverse.

Typically, fuels for household use are divided into modern fuels and traditional fuels. The former category includes refined liquid and gaseous fuels such as natural gas, LPG, and kerosene as well as electricity. The latter category includes solid fuels such as coal and biomass. Modern and traditional fuels are often characterized as commercial and noncommercial, respectively. Although this distinction is

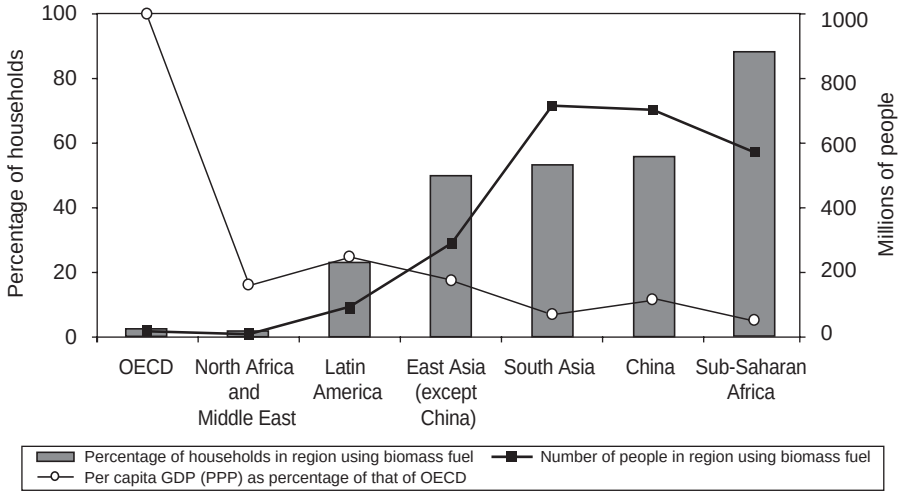


FIGURE 1 Biomass fuel users by world region: 2000. The figure shows estimations of the numbers and fractions of households cooking or heating with biomass fuels by world region. The fraction of households in each region is indicated by vertical bars and measured on the left vertical axis. The number of people using solid fuels in each region is indicated by the dark line with squares and is measured on the right vertical axis. The lighter line with circles indicates the per capita gross domestic product (GDP) adjusted for purchasing power parity (PPP) for each region relative to the OECD's per capita GDP, also measured on the left vertical axis. All data except OECD data are from the International Energy Agency. OECD data are from a database developed by Lawrence Berkeley National Laboratory in 1999 and are used with permission. This graph is based on data compiled by the author in collaboration with Prof. Kirk Smith and Sumi Mehta of the University of California's School of Public Health originally for their assessment of the impact of solid fuel consumption on the Global Burden of Disease.

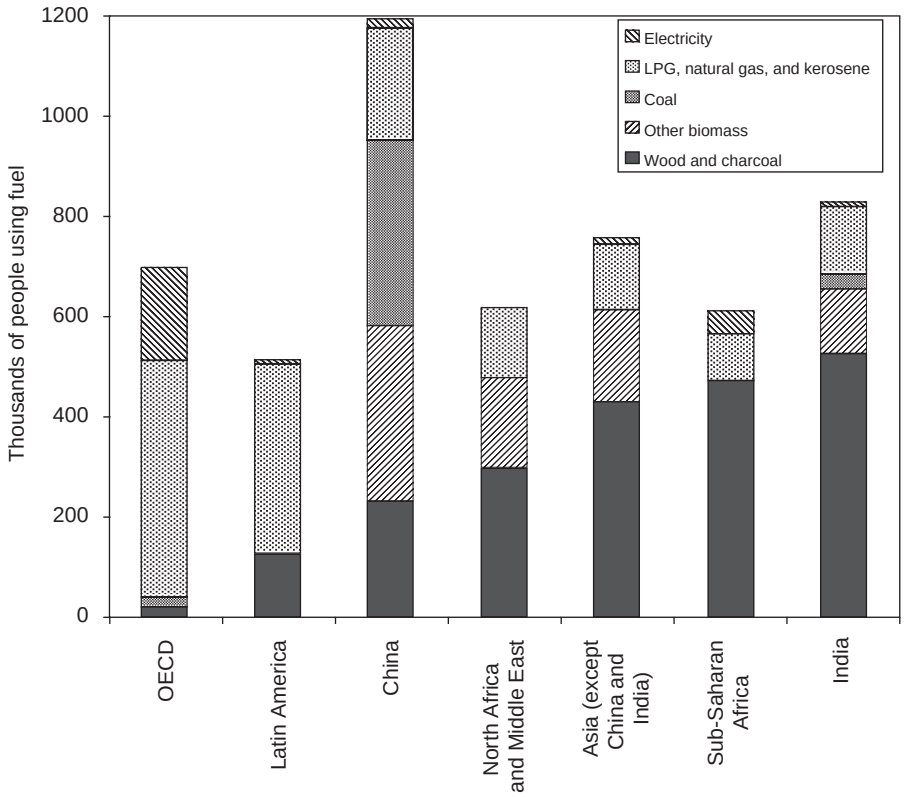


FIGURE 2 World household fuel choice by region: 1990–2000. The mix of fuels used by households in various regions of the world is based on a wide range of census and survey data compiled throughout the 1990s regarding households' primary cooking fuel. In reality, many households use multiple fuels, but this is difficult to account for at the national or regional level. Note the "other biomass" includes crop residues and dung.

occasionally useful, it can also be misleading because it implies that markets do not exist for solid fuels. On the contrary, there are often well-developed markets for different types of traditional fuels, and some (e.g., charcoal) are traded internationally.

One factor that correlates well with residential energy demand and fuel choice is the degree of industrialization in the country. Households in highly industrialized countries rely primarily on commercial liquid and gaseous fuels or electricity for their residential energy needs, with a few regional pockets where solid fuels—primarily wood and coal—are used. In contrast, households in less industrialized countries tend to use much higher fractions of solid fuels. Figure 1 shows a recent estimate of the number of biomass fuel users in the world from the International Energy Agency (IEA). The graph also shows regional gross domestic product (GDP) per capita adjusted for purchasing power parity (PPP) and indexed to the Organization for Economic Cooperation and Development (where OECD = 100%). South Asia (India, Pakistan, and Bangladesh) has the highest number of users, whereas Sub-Saharan Africa has the highest rate of use.

Within a country or region, there are many additional factors that contribute to household fuel choice. These include access to fuels and residential energy technologies, rural–urban settlement patterns and development policies, household income and gender dynamics of household decision making, and important sociocultural preferences affecting household behavior and cooking practices. Figure 2 shows the range of household fuel choices in various regions of the world.

2. WOOD BURNING AT THE HOUSEHOLD LEVEL

Household energy demand includes energy for cooking, water heating, and space heating as well as for some productive end uses. In developing countries, the primary end use is often cooking, although other applications are also important. People who rely solely on wood or other solid fuels often use the same technology to cook, heat water, and (in colder regions) heat their homes, in contrast

to industrialized countries where most people have access to a range of specialized energy conversion technologies. Because the stove is the most common wood-burning tool, this section focuses primarily on cooking stoves and cooking activities with the understanding that the stove is often used for noncooking tasks.

2.1 Wood-Burning Technologies for Household Use

One of the oldest methods of cooking with wood, the "three-stone fire," is still the most prevalent cooking technology in the world. The three-stone fire consists of a pot supported a short distance above the ground by three stones. Figure 3 shows an example of a three-stone fire from Kenya. This stove is indoors, although cooking may also be done outside, depending on local practices. The three-stone fire is favored by many people because it is free and requires no special equipment to build. It accommodates a range of pot shapes and sizes, and it provides heat, light, and a measure of security for the household. However, it also produces smoke as a result of incomplete combustion. This leads to very high concentrations of IAP that have serious consequences for the health of household members (discussed in more detail later).

Although the basic three-stone fire is very common, it is also modified in countless ways throughout the world. Common modifications to the basic three-stone fire include sinking the combustion zone below ground level and constructing a barrier around it to shield the fire from drafts. Figure 4 shows a stove from Bangladesh with the combustion zone sunk below the ground, and Fig. 5 shows a makeshift

FIGURE 3 Three-stone fire from Kenya. Photo by Majid Ezzati.

FIGURE 4 A variation of the three-stone fire from Bangladesh showing the fire sunk beneath ground level. Originally appeared in Stove Images (photo by IFRD/BCSIR, Dhaka, Bangladesh), reproduced with permission of the author and publisher.

FIGURE 5 Variation of a three-stone fire from India with protection from outside drafts. Originally appeared in Stove Images (photo by Dorsi Germann), reproduced with permission of the author and publisher.

stove from India that has been shielded from the wind. Reducing drafts has the effect of reducing heat loss and, thus, directing more of the fuel's energy into the cooking pot.

More extreme adjustments to the open fire involve fully enclosing the combustion zone, as this allows for some degree of air control and directs the hot gases produced by combustion more toward the cooking pots and less toward the surrounding environment. Some stoves include a flue or chimney that directs the exhaust out of the room, whereas others vent their exhaust past the pot and out into the room where cooking is taking place. A well-functioning chimney can reduce the concentration of pollutants inside the household dramatically by directing smoke outdoors. It also creates a natural draft in the stove that can improve combustion, further reducing pollution. Figures 6 and 7 show examples of stoves with completely closed combustion chambers from India and Nepal, respectively. The Indian stove has no flue so that the combustion emissions are vented into the room, whereas the Nepalese stove has a flue (built into the wall at the rear of the stove) so that emissions are vented outside. Additional modifications include insulating parts of the stove to reduce heat losses and adding dampers or secondary air inlets to improve air flow and combustion characteristics. More complex designs even preheat incoming air to facilitate combustion. There are hundreds of stove models incorporating various combinations of these technical design principles. Moreover, there are many nontechnical factors that constrain stove design and adoption (discussed in more detail later).

2.2 Impacts of Household Fuelwood Use in Less Developed Countries

As discussed previously, there are a number of problems associated with household fuelwood use in less developed countries. Deforestation is a frequently cited example, but the links between woodfuel consumption and deforestation are, in most cases, quite weak. Other problems, although less visible, are more strongly associated with household energy demands. The remainder of this section discusses these issues in more detail.

2.2.1 Woodfuel Scarcity and Deforestation

Historically, the most frequently cited problem with household woodfuel consumption has been deforestation, which has reached crisis levels in many rural areas of less developed countries. Closer scrutiny has shown that other factors, primarily land clearing for cultivation, timber extraction, and woodfuel supply to urban markets, dominate the loss of tree cover in most of the regions that suffer from forest degradation.

FIGURE 6 Stove from India with closed combustion chamber and no chimney. Photo by Kirk Smith.

FIGURE 7 Stove from Nepal with closed combustion chamber and chimney to vent emissions outside. Originally appeared in Stove Images (photo by Dorsi Germann), reproduced with permission of the author and publisher.

Household fuelwood rarely comes from mature trees in forests or woodlands; people prefer gathering fallen branches and dead wood. People may also collect wood from their own household compounds or from fallow land. However, problems arise when access to woodlands is denied or when the woodlands are cleared *en masse* for timber extraction or cultivation. The latter is often accompanied by processing of the cleared wood into charcoal or firewood, but that is marketed for urban consumption; in either case, the cleared wood is inaccessible to local users. In addition, whether land is held in common or is privately owned, population and economic pressures can force people to shorten fallow periods or expand the areas they cultivate, reducing both the time and the space in which their homegrown wood accumulates. Hence, although it is unlikely that energy demand is the primary cause of fuelwood scarcity, scarcity still affects many rural communities and can have a significant impact on household welfare.

2.2.2 Rural–Urban Energy Issues in Less Developed Countries: The Energy Ladder, Fuel Substitution, and Charcoal

2.2.2.1 Rural–Urban Energy Issues Less developed countries are often marked by a clear distinction in the level of development between urban and rural areas. This distinction is visible in many socioeconomic arenas, including household energy. Urban households have a greater range of energy options, and most of their fuels are procured through market transactions. This is in contrast to rural households, where choices are limited but fuel is often gathered freely by household members.

In places where wood is a common rural fuel, charcoal is often popular in cities and towns. However, patterns of household fuel use in urban households depend not only on the supply of wood in the country but also on the degree to which other household fuels are available, and the latter itself depends on state policies created to develop and/or regulate fuel markets. In addition to charcoal, other fuels used in urban areas include coal, kerosene, LPG or natural gas, and electricity. Households in rural areas have fewer options because markets do not evolve where transportation is more expensive and household purchasing power is much lower.

2.2.2.2 The "Energy Ladder" Household fuels can be ranked by order of user preference in an "energy ladder," which is a simple conceptual model that has been developed to explain and predict relationships between household fuel choice and economic development. The concept of the energy ladder model is based on two premises: (1) that a ranking of fuels in order of user preferences correlates roughly to the ease of use, cleanliness, and energy efficiency of the fuel and (2) that users move up or down the ladder as household income rises or falls or, more broadly, as economic progress advances or recedes.

Problems with the energy ladder model arise from the simplified way in which the model is applied to policymaking, as this can lead to the mistaken conclusion that fuel choice is determined purely by economic factors. An increase in household income will not necessarily be spent on cooking stoves or fuel. A great deal depends on which household member controls expenditure decisions and how that person prioritizes expenditure in relation to other household needs. Moreover, cultural preferences may lead people to retain a particular fuel to cook certain foods or to use on special occasions so that complete fuel switching is quite rare. Finally, ease of access and consistent availability of fuels both are important factors that determine the extent and/or permanence of fuel switching in any household. Some of these dynamics are illustrated by an example from Kenya shown in Fig. 8.

2.2.2.3 Charcoal Charcoal is the principal woodfuel in urban areas of many less developed countries. There are a number of reasons why people in dense urban settlements favor charcoal over wood. It has a higher energy density, it burns cleaner (reducing exposure to harmful pollutants), and it is easier to transport, handle, and store. In addition, many people favor charcoal because it is considered a more modern fuel than wood and is a kind of status symbol.

Domestic charcoal use in less developed countries is possible only with a thriving charcoal industry. Charcoal production is most prevalent in Africa, although it is also common in several other countries such as Brazil, India, and Thailand. Table I shows the 10 largest charcoal-producing, -importing, and -exporting countries. Brazil is a bit of an anomaly; it produces as much charcoal as do the next five largest producers, but charcoal is used mainly in the iron industry and is not a major household fuel in that country. In contrast, in many African and Asian countries, charcoal is an important urban household fuel. In addition, a small amount of international trade occurs, as the table indicates.

Charcoal is often derided as wasteful and destructive to the environment because traditional charcoal

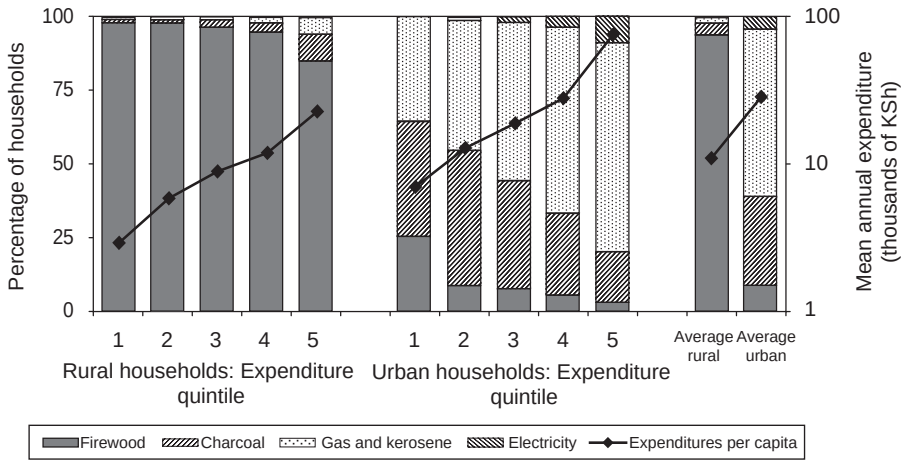


FIGURE 8 Fuel choice and household expenditure in rural and urban Kenya by expenditure quintiles: 1994. The figure shows per capita expenditure and primary cooking fuel for Kenyan households disaggregated by rural and urban areas and grouped by expenditure quintiles. The differences between rural and urban groups are more apparent than the differences between income groups within either area. The wealthiest 20% of rural households have higher mean annual expenditure than do 60% of all urban households. However, 85% of the wealthy rural households use wood as their primary cooking fuel. In contrast, 53% of households in the third urban quintile and 35% of the poorest urban quintile choose gas (LPG) or kerosene. This may indicate that fuel preferences are different between urban and rural households, or it may reveal that markets for alternative cooking fuels have failed to penetrate rural areas (the latter is more likely to be the case). From "World Bank African Development Indicators: 2000 Database" (citing a 1994 Welfare Monitoring Survey conducted by the Kenyan government in which more than 10,000 households were sampled).

TABLE I

Top 10 Charcoal Producers, Importers, and Exporters: 2000

Top charcoal producers	Tons	Percentage of world production	Top charcoal importers	Tons	Top charcoal exporters	Tons
1. Brazil	12,063,000	29	Japan	129,000	Indonesia	148,700
2. Nigeria	3,057,000	8	Germany	96,500	China	83,800
3. Ethiopia	2,907,000	7	South Korea	95,000	Malaysia	61,200
4. Kenya	2,475,000	6	China	70,538	South Africa	44,500
5. India	1,654,000	4	Italy	46,000	Argentina	44,000
6. Democratic Republic of the Congo	1,418,000	3	United Kingdom	45,000	Poland	40,000
7. Thailand	1,222,000	3	Norway	43,000	Mexico	38,100
8. Egypt	1,196,000	3	Saudi Arabia	37,200	Somalia	33,678
9. Tanzania	1,165,000	3	Singapore	28,000	The Philippines	30,000
10. Zambia	1,040,000	3	France	27,746	Nigeria	28,000
World total	40,615,004	100	World total	921,312	World total	851,314

Note. Charcoal data are from the UN Food and Agriculture Organization's (FAO) online statistical database (http://apps.fao.org/page/collections? subset = forestry) with the exception of Kenya. Data for Kenya are from a recent domestic survey. This survey, based on household consumption data, is nearly 300% higher than the numbers given for Kenya in the FAO's database. Many other discrepancies exist in national-level charcoal data from other countries. The FAO is currently undertaking an effort to identify and correct these errors.

production methods can be inefficient. In an earth-mound kiln, the most common method of making charcoal in Sub-Saharan Africa, between 5 and 10 tons of wood is needed to make 1 ton of charcoal—a mass-based conversion efficiency of 10 to 20%. In these circumstances, between 60 and 80% of the wood's

energy is lost in the production process. The process depends on a number of variables such as scale of production (production can range from <100 kilos up to 30 tons), moisture content and size of the wood, and time taken for the process (this also depends on the scale of production and can range from 2 to 10 days). Efficiency varies with the skill of the producers and the techniques they employ, which are likely to be a function of wood scarcity. Producers will adopt labor-intensive steps to conserve wood only if they perceive wood scarcity as a threat to their own livelihood.

More efficient methods of charcoal production using special kilns or retorts can reduce energy losses to between 30 and 40%. However, this equipment is expensive relative to traditional methods. It is not likely to be widely adopted without outside intervention, particularly in regions where wood is accessible for little or no financial cost. Nevertheless, the comparative efficiencies of charcoal and fuelwood should not be based solely on production. Cooking with charcoal can be more efficient than cooking with wood (typical efficiencies for charcoal stoves are 20–30%, whereas those for three-stone wood fires are usually 10–15%). If a person cooks a meal using a charcoal stove that is 30% efficient and the charcoal he or she uses was produced at 20% efficiency (based on wood-to-charcoal mass, this is high but has been observed in practice), that person uses less wood than does a person cooking the same meal with an open wood fire at 10% efficiency (a low but not uncommon efficiency). This simple comparison indicates that generalizations about the wastefulness of charcoal should be scrutinized carefully, particularly in light of charcoal's cleaner burn and other favorable attributes.

In addition to the question of energy efficiency, charcoal production is often considered environmentally destructive for two reasons: deforestation and pollution. The links between charcoal production and deforestation are stronger than the links between rural fuelwood consumption and deforestation because charcoal production involves cutting mature trees. However, rural land is often cleared for other purposes such as expanding cultivation, and charcoal is produced as a secondary activity following land clearing. In these situations, the cause of forest loss is not primarily charcoal because the land would have been cleared regardless of whether charcoal was produced, although charcoal can make the activity more lucrative. Furthermore, when land is cleared specifically for charcoal production and is not subsequently used for cultivation or grazing, trees can grow back. Permanent damage is limited to the area underneath the charcoal kiln, usually 2 to 3% of the cleared woodland. However, not all ecosystems are equally resilient. When trees are cleared from moist tropical forests, they might not recover as readily as do wooded savannah.

Although woodfuel demand in urban areas does not have a dominant influence in the overall loss of a nation's tree cover, it can have severe impacts on specific locations. This is especially true in nations with weak regulatory mechanisms governing the harvesting of forest resources, where outsiders can clear large tracts of woodlands on which rural communities rely without offering rank-and-file community members any compensation for their lost resources. Devolving control to local communities can ensure that they benefit from the exploitation of forest resources, although the benefit stream is very sensitive to the structure of institutions that usually does not favor local control.

The second environmental concern associated with charcoal is the emissions of GHGs (discussed in more detail later).

2.2.3 Energy, Poverty, and Gender

Recent analyses from the World Bank and other development institutions have introduced the concept of energy poverty to development discourse. Linking energy to poverty implies that a lack of access to clean and efficient energy services is one factor that contributes to persistent poverty in less developed countries. Energy services include not only electricity and clean cooking fuels but also the activities or end uses that fuels and electricity can provide (e.g., lighting, heating, motive power). Access to these services is strongly correlated to the level of development in a particular country or region. Figure 9 shows a plot of cross-sectional data from 125 countries in seven regions of the world indicating that well-being, as measured by the UN Development Program's Human Development Index (HDI), is well correlated to the degree of "modern" energy services consumed in each country. However, correlation should not be mistaken for causation. The relationship between poverty and lack of access to "modern" energy services is not a simple matter of cause and effect. High rates of poverty are not likely to be eased simply by providing clean cooking fuels or electricity without addressing the many structural factors that contribute to poverty. Furthermore, although the general trend shown by the linear fit is an interesting outcome of the figure, it is equally interesting to identify the outliers on the plot—those that lie well above or below the general trend—and

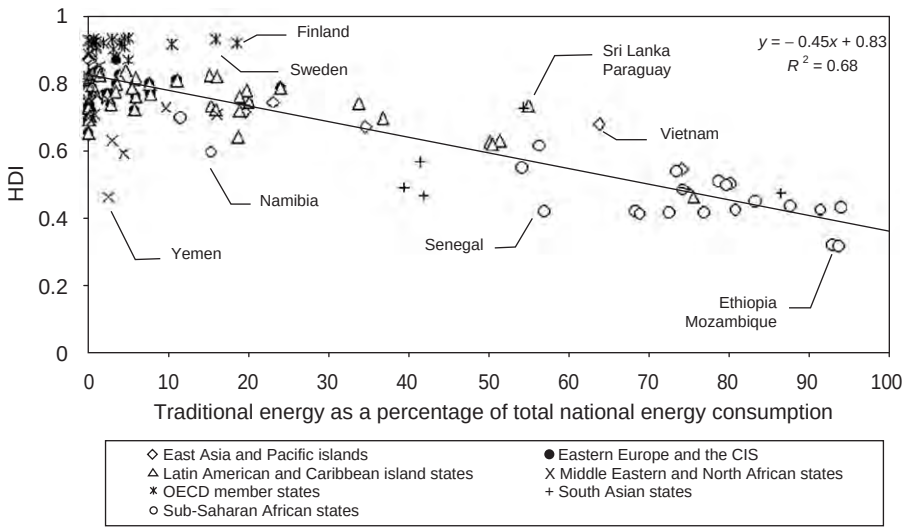


FIGURE 9 Human Development Index (HDI) and fraction of national energy consumption from biomass. HDI is plotted against biomass energy consumption for 125 countries. The countries are divided into seven regions, indicated by different symbols in the figure. HDI is a relative indicator of national development developed by the UN Development Program (UNDP). Biomass energy consumption is plotted as a percentage of each country's total energy consumption. This percentage is a reasonably accurate indicator of access to "modern" energy services, although there are exceptions. For example, in Finland and Sweden, households have 100% access to modern energy services, but industry and government both use biomass energy to a much larger extent than do similarly developed nations. Other countries that deviate from the main trend are also labeled in the figure. Data from UNDP, "Human Development Report, 2001" (www.undp.org); World Resources Institute, "Earth-Trends Database" (http://earthtrends.wri.org/index.cfm).

to question what sets those countries apart from the rest, for better or worse.

Of course, energy is a critical input for any productive activity, whether it is for household needs, commercial enterprise, or cottage industry. Without clean and efficient forms of energy, the possible range of productive activities are limited and even simple tasks are time-consuming and physically strenuous exercises. Increasing access to improved energy services will likely lead to improvements in the quality of life for many people; however, the critical questions arise as to how to provide those services and whom to target for increased access. Services that are given away for free or heavily subsidized are unlikely to be sustained in the long term as budget priorities shift over time and as donors' interests shift to new types of projects. Alternatively, services that are provided only through free market mechanisms are unlikely to reach the poor households that need them the most.

Figure 10 shows the effects of Kenya's largely unregulated household energy market. Household fuels are differentially affected by subsidies and other distortions with no clear policy goal; charcoal is effectively unregulated, kerosene is not directly subsidized but is exempt from most of the taxes levied on other petroleum products, and LPG is taxed significantly more than is kerosene. The figure shows a range of initial costs and annual operating costs for various cooking technologies accessible to consumers in Nairobi, a city of approximately 3 million people, many of whom live below the poverty line. The results are presented in ranges rather than point data to show that, realistically, consumers face a range of stove costs, lifetimes, and efficiencies as well as variable fuel costs. The fraction of urban Kenyans using each stove–fuel combination is shown in parentheses beneath the label on the x axis. The inputs used in the model are described in Table II. The interesting result of this estimate is that all stoves show significant overlap in their operating costs amortized over the lifetimes of the stoves, but they differ significantly in initial costs. There are few credit mechanisms to allow consumers to spread out the costs of the stoves over time, particularly for the poor. Thus, nearly half of the urban population still burns charcoal in unimproved stoves, the cheapest initial option, despite the likelihood that doing so is more expensive than using improved charcoal stoves, kerosene, or even LPG in the long term.

If free market provision of improved energy services is unlikely to reach the poor, subsidies are one alternative, but they need to be targeted properly

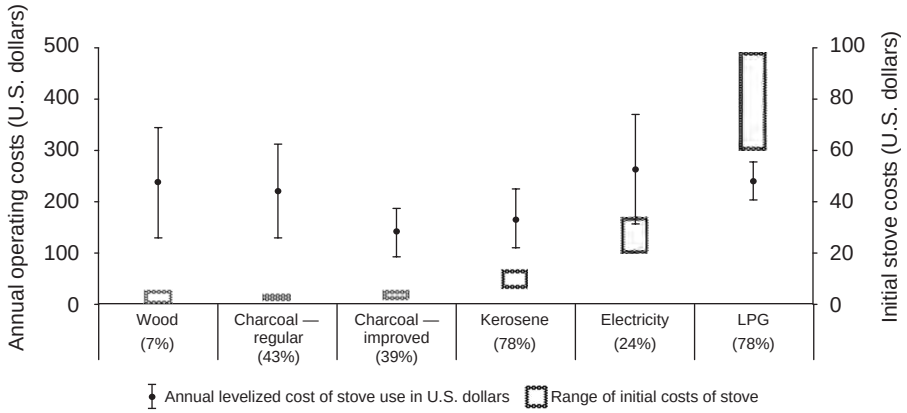


FIGURE 10 Operating costs for urban household stoves in Nairobi, Kenya. The figure shows an estimation of operating costs for commonly used stoves and fuels based on recent price data. Calculations are for an average-sized urban household (four people) assuming that the household uses one stove–fuel combination and consumes the useful energy equivalent of 1.3 GJ/year (the average urban household cooking energy consumption adjusted for stove efficiency reported in a recent Ministry of Energy survey). The vertical lines represent a range of operating costs based on several inputs (see Table II). The open bars represent a range of initial costs for the stove (also explained in Table II). The numbers in parentheses on the *x* axis are the percentages of urban households reporting use of the various types of stoves. (The LPG stove costs include the cost of a small [6-kg] gas tank and regulator.) These numbers add to greater than 100%, indicating that many households use more than one type of stove. This implies that the figure does not accurately model most Nairobi households, but it is still useful for comparing relative costs of household energy technologies.

TABLE II

Input Data for Annual Operating Costs of Typical Stove–Fuel Combinations in Nairobi, Kenya, Shown in Fig. 10

Stove type	Range of fuel costs per unit of sale	Energy content per unit of sale	Range of stove efficiencies (percentage)	Range of initial stove costs (KSh)	Stove lifetime (years)	Annual fuel consumption per household	Annual operating cost at 12% discount rate (U.S. dollars)
Charcoal (improved)	13–19 KSh/kg	28 MJ/kg	25–35	150–300	1–3	531–743 kg	128–342
Charcoal (regular)	13–19 KSh/kg	28 MJ/kg	15–25	100–200	2–4	743–1238 kg	128–310
Electricity	0.07–0.12 U.S. dollars/kWh	3.6 MJ/kWh	50–70	1500–2500	5–10	2063–2889 kWh	92–186
Kerosene	33–45.5 KSh/L	35 MJ/L	40–60	450–900	4–6	246–369 L	108–222
LPG	80–83 KSh/kg	48 MJ/kg	45–60	4600–7400	5–10	181–241 kg	156–369
Wood	6–8 KSh/kg	16 MJ/kg	10–20	0–300	1–3	1625–3250 kg	201–275

Notes. *Stove efficiencies*: Stove efficiency, defined as the fraction of the fuel's heating value that enters the food, is a compound measure of two interdependent quantities: combustion efficiency (CE) and heat transfer efficiency (HTE). CE indicates how much heat is released from the fuel relative to ideal combustion, whereas HTE indicates the fraction of the heat released by combustion that enters the food in the pot. Typical values for CE range from 80 to 90% for solid fuels and from 98 to 99% for liquid and gaseous fuels. Typically, values for HTE range from 50 to 60% for solid and liquid stoves and from 10 to 30% for solid fuels. Net efficiency is the product of the CE and HTE.
Fuel prices: Prices for charcoal, kerosene, and LPG were reported by Matthew Owen, an expert on Kenyan energy markets (personal communication with the author). Wood prices are from a Ministry of Energy survey. Electricity prices are based on the tariff structure currently used by Kenya's national power company. The tariff ranges from approximately U.S. $0.07/kWh for consumption of 50 kWh/month to U.S. $0.12/kWh for consumption of more than 2000 kWh/month. A household cooking purely with electricity would likely be at the higher end of this range, but the lower prices are included to reflect what the costs would be at the lower tariffs.
Stove prices: Prices are based on the author's observations reflecting a range of stove types. Kerosene stove prices reflect two sizes of wick stoves. LPG stoves are one- or two-burner stoves with a small (6-kg) tank and regulator (these retain their value and can be resold but are still unavoidable costs to the consumer). Electric stove prices reflect tabletop models with either one or two burners.

to have their desired effect. Subsidies are particularly important where current practices are associated with high external costs such as the severe impacts on public health associated with the use of wood and other biomass fuels. These impacts are borne disproportionately by women and young children as a result of their role in the household. The remainder of this section addresses the special role of women in the energy–poverty interaction.

As international attention was directed to the fuelwood question due to its perceived environmental effects, it became apparent that in many less developed countries the responsibility for the supply of household energy and the impacts of its consumption were largely, if not exclusively, the domain of women. Many early efforts to address fuelwood-related problems failed because they targeted men, who had little awareness of the work that their mothers, wives, and daughters were doing. This bias in project targeting is compounded by cultural norms that discourage women from interacting with outsiders and vocalizing their opinions within the community.

Importantly, although it is usually women's responsibility, household energy is not the only household task for which women are responsible. The division of labor varies across societies and cultures, but in most rural areas of less developed countries women are responsible for child care, cultivation of staple food crops, water collection, and looking after the elderly and sick. In many places, as men leave the home to seek wage labor, more and more women are also running small businesses or cultivating cash crops. Another near universal commonality among rural households is that women often have little say in household expenditure decisions.

Hence, energy-related impacts and responsibilities fall largely on women's shoulders, but they might not prioritize energy-related problems, even if they are fully cognizant of them, because their responsibilities are so far-reaching. These common factors make it particularly important to be sensitive to female voices in designing household energy policies and programs at all levels, from national legislation to community interventions. Importantly, sensitivity to female voices does not imply that men should be excluded from the process, only that special effort needs to be made to ensure that women's needs are addressed effectively.

2.2.4 Indoor Air Pollution and Health

Although the direct link between household energy and deforestation has been largely discredited, widespread fuelwood consumption does have another significant, albeit much less visible, global impact. Burning wood and other solid fuels within the household leads to very high concentrations of IAP. Exposure to these pollutants, which include carbon monoxide (CO), particulate matter (PM), and unburned hydrocarbons (HCs), can lead to a number of acute and chronic health impacts. The most prevalent of these impacts is acute respiratory infection (ARI), caused primarily by exposure to smoke or PM. ARI is one of the leading causes of death in children under 5 years of age worldwide; roughly 2 million deaths occur each year. In addition to ARI, exposure to IAP from solid fuel combustion causes chronic obstructive pulmonary disease and some types of cancer. Furthermore, exposure to solid fuel combustion emissions is associated with increased incidence of asthma, tuberculosis, and cataracts and elevated risk of CO poisoning. As a result of these health impacts, the World Health Organization estimates that exposure to IAP from solid fuels is the eighth leading cause of illness and death in the world. This impact is highly concentrated in poor developing countries, where IAP is the fourth leading cause of illness and death.

The effects of IAP are further concentrated within particular populations. As a result of the division of labor within the household, ARI and other health impacts associated with IAP are felt disproportionately by women and young children. Moreover, there are a number of health risks associated with fuelwood that are independent of exposure to combustion emissions. Prominent among these are burns (common among young children), risk of injury from gathering and transporting heavy loads of fuel over long distances, and exposure to possible harassment for girls and women gathering wood far from home.

2.2.5 Woodfuels and Greenhouse Gas Emissions

The same processes of incomplete combustion that cause the emissions of health-damaging pollutants also result in the emissions of GHGs. Wood and other forms of biomass are essentially HC fuels with a few trace elements. Perfect biomass combustion should produce only carbon dioxide (CO_2), water vapor, some trace gases, and ash. Water vapor produced by combustion is negligible compared with the natural stock of water in the atmosphere, and it has a short atmospheric residence time so that the impact of anthropogenic water vapor is negligible. CO_2 emissions from wood combustion add to the stock of GHGs in the atmosphere, but the effect of this additional CO_2 can be effectively "cancelled out" if the wood that is consumed is "replaced" by an equivalent stock of wood because photosynthesis

removes CO_2 from the atmosphere. Hence, wood-burning stoves can, in theory, be GHG neutral if two conditions are satisfied: (1) if wood undergoes complete combustion and (2) if an equivalent amount of wood is grown to replace the wood that was burned.

For household stoves burning solid fuels, the first condition (complete combustion) is never attained. The second condition (sustainable harvesting of fuel) is related to the question of deforestation. This condition may be satisfied, but in reality it is difficult to verify for a large number of people because woodfuel consumers obtain their fuel from many different sources and harvest trees for many different end uses. The remainder of this section explains the consequences of the failure to meet the first condition in more detail.

Small-scale household stoves cannot achieve complete combustion. (Even stoves that burn liquid and gaseous fuels do not reach that ideal, although they come closer than do most solid fuel stoves.) The physical design of these stoves is such that the fuel cannot vaporize and mix sufficiently with air at high temperatures, which are the conditions required for complete combustion. When fuels do not fully combust, pollutants such as CO, methane (CH_4), and nonmethane hydrocarbons (NMHCs) as well as PM are released. If nitrogen is present in the fuel or if temperatures are high enough to oxidize atmospheric nitrogen, combustion also creates nitrogenous compounds such as N_2O and NO_x. Fuelwood typically has less than 0.5% nitrogen, but bark, leaves, and other types of biomass consist of up to 1 or 2% nitrogen. Thus, small quantities of nitrogen compounds are present in wood combustion emissions.

GHGs trap heat in the atmosphere in a process called radiative forcing. The impact of GHGs on the earth's climate is a product of radiative forcing and the length of time it takes a portion of gas released into the atmosphere to be removed. From these two quantities, a global warming potential (GWP) is defined, indicating the time-dependent warming effect of a mass or volume of gas released relative to the effect of CO_2, the most prevalent anthropogenic GHG. Wood combustion releases GHGs that have large GWPs; thus, even if wood is used sustainably such that all of the CO_2 released during combustion is absorbed by future photosynthesis, wood burning has an impact on the climate.

Only CO_2, CH_4, and N_2O are subject to control by the UN Framework Convention on Climate Change (UNFCCC). However, CO, NMHCs, and PM can also have an effect, although their influence is less certain. Table III indicates the GWPs for biomass combustion emissions.

For each pollutant, emissions from wood and charcoal stoves can be described by an emission factor indicating the mass of pollutant released per

TABLE III

GWPs of Greenhouse Gases Commonly Emitted from Biomass Combustion on a Molar Basis

Compound	20-year GWP	100-year GWP	500-year GWP	Comment
CO_2	1	1	1	CO_2 GWP is 1 by definition for all time horizons.
CO	2.0–6.0	0.6–2.0	0.2–0.6	Range of values is reported by IPCC: lower values consider CO effect on OH radicals; higher values also consider ozone production.
CH_4	22.5	8.4	2.5	From IPCC's third assessment report.
NMHC[a]	12	4.1	2.3	From IPCC's 1990 report; subsequent reports do not offer values for NMHCs due to the high degree of uncertainty.
N_2O	275	296	156	N_2O was not measured in this study and is included here for comparison only.

Source. Adapted from Smith *et al.* (2000).

Note. GWP, global warming potential; IPCC, Intergovernmental Panel on Climate Change; PM, particulate matter. The time-dependent behavior of the GWP arises from the atmospheric lifetimes of the compounds and their decay products. There is no single GWP for PM, although estimations exist for the net effect of PM on radiative forcing. There are two types of PM from combustion, and each has a different effect on climate. Black carbon particles can contribute to climate warming, whereas organic carbon particles can contribute to cooling, but the level of scientific understanding of both effects remains low.

[a] The molecular weight of NMHCs is assumed to be 18 g/mol C. The molar GWP of NMHCs effectively counts moles of C, rather than moles of a mix of compounds that generally have more than one C atom per molecule. This facilities comparison with single-carbon compounds such as CO and CH_4.

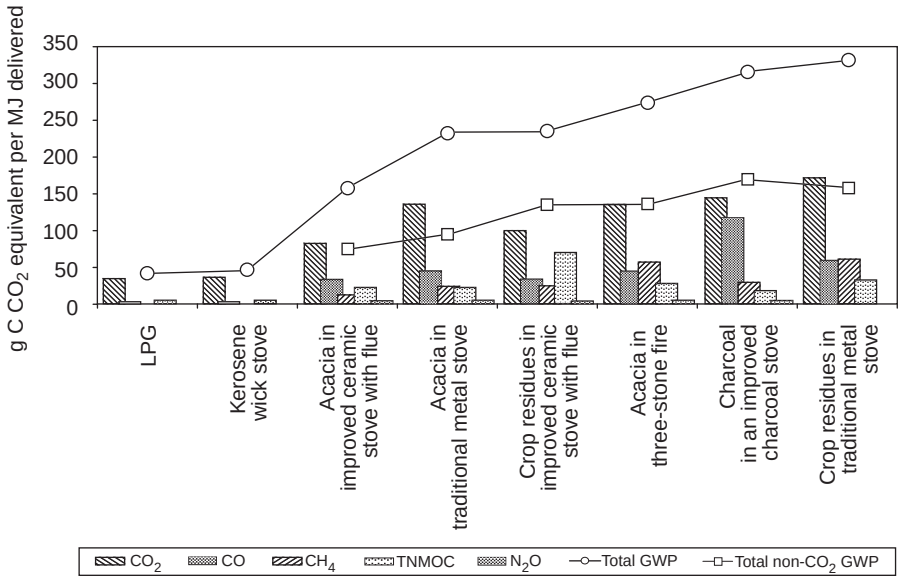


FIGURE 11 Emission factors of GHGs from common Indian household stoves and fuels weighted by 20-year GWPs. The emission factors were measured in a study of Indian cooking stoves using a variety of fuels and were calculated with respect to energy delivered to the cooking pot. The vertical bars indicate the emission factors for each individual pollutant. The lines indicate the sum of emission factors from each stove-fuel combination both with CO_2 (open circles) and without CO_2 (open squares). Including CO_2 in the total warming effect accounts for the impact of unsustainable harvesting, while omitting CO_2 from the calculation shows the net warming effect when the fuels are harvested sustainably so that stocks of trees or other biomass are allowed to regenerate (of course, this "sustainable" option is not possible for LPG or kerosene). Created by the author based on a selection of data from Smith *et al.* (2000). "Greenhouse Gases from Small-Scale Combustion Devices in Developing Countries Phase IIa: Household Stoves in India," EPA-600/R-00-052. U.S. Environmental Protection Agency, Washington, DC.

unit of fuel or energy consumed. Emission factors vary depending on the physical characteristics of the stove, the condition of the fuel, and the behavior of the stove user. Figure 11 compares the emission factors of some common Indian cooking stoves weighted by GWPs. The graph shows that as a result of poor combustion processes, biomass stoves result in higher GHG emissions than do fossil fuels, even when the biomass fuels are harvested sustainably so that CO_2 is not included. However, this result includes only end uses. A fair comparison of emissions from fossil fuel- and biomass-based cooking technologies must account for the emissions from production and transportation. Accounting for charcoal production effectively doubles the emission factors for charcoal shown in the figure, whereas "production" of fuelwood is essentially emissions free. A full life cycle analysis of GHG emissions from household fuels commonly used in developing countries has yet to be performed. But even with out a full comparison, the problems arising from poor combustion of wood and other biomass fuels should be apparent from the results illustrated in Fig. 11 as well as on the previous discussion on health impacts.

3. ALTERNATIVES TO WOOD BURNING AND POSSIBLE FUTURE DIRECTIONS FOR HOUSEHOLDS IN LESS DEVELOPED COUNTRIES

There have been many attempts to solve the problems associated with household woodfuel consumption, ranging in scope from national-level policies to community-based interventions. Across these scales, interventions fall into two general categories: increasing supply of woodfuel through silvicultural methods and decreasing consumption by promoting improved cooking stoves, promoting alternative fuels, and even prohibiting tree cutting and/or wood collection. Many projects try to address both approaches, but the methods required to increase supply are significantly different from those that seek to reduce consumption.

Both approaches can lead to many benefits for the communities involved, including decreased household labor requirements and lower energy expenditure as well as a range of environmental benefits associated with increased tree cover. In addition, there are significant health benefits associated with

reducing IAP by using more efficient stoves or alternative cooking fuels. Linkages among household energy, public health, and community development have become the focus of significant effort among researchers and policymakers since the mid-1980s.

However, improvement in one or all of these areas is not guaranteed simply by introducing an intervention. The extent of any changes in welfare that result will depend on many factors, some of which are not directly related to household energy. These factors include agricultural production, labor markets, local gender dynamics and power relations, and the community's past experiences with outside intervention that directly affects the feasibility of future projects. It is also important to recognize that changes in household energy technologies and consumption patterns can occur without any outside intervention. The remainder of this section explores some of the ways in which changes in woodfuel supply, demand, and cooking practices have been initiated or attempted in various parts the world and closes with a discussion of some possible directions that household energy provision in less developed countries may take in the future.

3.1 Supply-Side Interventions: State Forestry, Social Forestry, and Agroforestry

To outsiders, the immediate response to woodfuel scarcity is to plant trees. In some cases, this response coincides with local needs. Indeed, communities may have the desire to reforest their landscapes but may lack access to seeds or land on which to plant. However, planting and maintaining trees can be a time-consuming, labor-intensive process for local communities. People are unlikely to plant trees for fuelwood if there are alternative sources of fuel or if the land can be put to more lucrative uses. Tree planting as a response to wood scarcity is complicated by local property institutions that can be quite different from Western notions of private property. In many countries, the property rights in trees and their products are different from, as well as separable from, the property rights for the land on which the trees grow. In addition, in many postcolonial societies, there is a long history of land appropriations and forced evictions. Interventions initiated by environmental crises were one mechanism through which the government gained control of the land. Thus, these types of interventions may be justifiably viewed with suspicion by local communities and are often ignored or resisted. These are not arguments against tree planting; rather, they are simply points that must be considered when wood scarcity is identified as a local problem and tree planting is proposed as a solution.

When wood scarcity is identified as a problem and increasing supply is the favored response instead of, or in conjunction with, demand reduction, there are two broad categories of practices that are generally attempted: centralized tree management through state- or community-run forestry and individual tree management through farm- or household-level forestry.

State forestry institutions are often antagonistic toward local communities, restricting or prohibiting access to state-owned forests. Despite this general trend, some governments have successfully established woodlots or managed forests specifically for community wood production. However, establishing tree plantations is expensive, particularly when state bureaucracies are involved, and centralized state forestry is not an economical way in which to mitigate woodfuel scarcity. On the other hand, if state-owned forests are already established (e.g., as a timber reserve), the state can ease wood scarcity by allowing local communities to have access to dead wood, fallen trees, and pruned branches or by allocating a section of the forest for community access.

In contrast to state forestry, community forestry places forest management directly in community hands. The managed trees may be a section of natural forest, a plantation, or a woodlot. It may lie on land held in common by the community, or it may lie on state-owned land with management responsibilities vested in the community. Fuelwood provision is one of many possible dimensions of community forestry, and fuel collection is likely to be one of many rights that members of the community claim on the trees and land under their control. Other claims include the right to graze livestock, fish, hunt, and extract a variety of forest products such as food, medicine, leaves, and thatch. To preserve the resource, fuel gathering may be limited to dead wood and fallen or pruned branches, commercial sale of wood may be prohibited, and extraction may be limited to head loading only, with carts or animal loading also prohibited. Alternatively, community forestry may be practiced specifically for timber production.

Local wood scarcity can also be mitigated by tree planting at the household level. For example, much of the fuelwood consumed in Sub-Saharan Africa does not come from forests or woodlands; rather, it is cut or gathered from boundaries and fences, cropland, private woodlots, and roadsides as well as from

neighboring farms. Rural farmers throughout the developing world maintain significant numbers of naturally occurring and cultivated trees on their own land. Many do this without assistance from outsiders, although outside intervention can help to provide seeds and technical advice.

As with community forestry, trees on farms are rarely grown solely for fuelwood; rather, they are grown for a wide range of uses. Agroforestry techniques, which integrate trees with cultivation and livestock systems, are particularly effective for maintaining trees on the homestead. These techniques use favorable characteristics of specific tree species and can bring a wide range of benefits to farms. For example, farmers who plant nitrogen-fixing trees in their fallow fields for several years before replanting food crops can experience yields that are several times higher than those from unfertilized plots, that is, yields that are equal to plots fertilized with costly commercial fertilizers. In addition, farmers can harvest the young trees for fuel when the fields are ready to be sown. In a typical area of the subhumid tropics, a 1-ha plot can yield roughly 5 tons of fuelwood for each year of fallow, and this can satisfy a family's fuel needs for a year.

3.2 Demand-Side Interventions

Attempts to reduce woodfuel consumption fall into two broad categories: improved woodfuel stoves and substitution of woodfuel for other fuels.

3.2.1 Improved Stoves

In addition to tree-planting measures, the early connections between fuelwood consumption and deforestation led to the development and dissemination of fuel-conserving cooking stoves. Critics of the three-stone fire and other traditional cooking stoves pointed to their low efficiencies and assumed that they could improve on them through the application of basic engineering principles. Improved cooking stove programs have been implemented throughout the world with mixed degrees of success. Projects vary by scale, type of stove, and target population. Some projects target individual households, whereas others target stove producers or marketers. Table IV shows the number of stoves disseminated in some of the more prominent programs.

As discussed earlier, there are hundreds of variations of cooking stoves. Technical considerations weigh equally with many other factors in determining the design of the stoves. With the benefit of two decades of hindsight, much has become apparent to people working in cooking stove design and dissemination. Some of the more common lessons learned are discussed in what follows.

TABLE IV

Improved Cooking Stove Dissemination through the 1990s

Country	Number distributed	Years of dissemination
China	185,000,000	1982–1998
India	4,100,000	1985–present (number given is stoves disseminated by mid-1990s)
Kenya	1,450,000	1983–present
Thailand	300,000	1985–present (number given is stoves disseminated by mid-1993)
Burkina Faso	200,000	1983–present
Niger	200,000	1984–present
Uganda	52,000	1985–present
Central America	9,000	1990–present (projects are in different countries, and number of stoves disseminated is valid as of 1995)

Source. African countries are from Karekezi, S. (2002). Renewables in Africa: Meeting the energy needs of the poor, *Energy Policy* **30**, 1059–1069. Latin America, India, and Thailand are from Westhoff, B., and Germann, D. (1995). *Stove Images*. Apsel Verlag, Frankfurt, Germany. China data are from the UN Development Program's *World Energy Assessment, 2002*.

3.2.1.1 "Improved" Is a Subjective Measure of Newly Introduced Stoves Not all new stoves lead to improvements over "traditional" cooking practices. As in all development interventions, the priorities of the beneficiaries might not agree with the priorities of the implementers and/or donors. A well-designed stove using sound engineering principles will be adopted only if consumers perceive fuelwood as scarce or costly to obtain and if substitute fuels are also costly to obtain. Even if fuelwood scarcity is an acute issue within the community, an "improved" stove will be adopted only if it is easy to use and well matched to local cooking practices. Contrary to some stereotypes that portray them as rigid and traditional, people in less developed countries are willing to change their behavior; however, like people in industrialized countries, they are more likely to choose convenience over conservation. A fuel-saving stove that requires constant attention or fine-tuning is not likely to be widely adopted.

Moreover, improved stoves that fully enclose the fire and remove harmful pollutants through a chimney do not provide the other benefits that an open fire offers: primarily warmth and light but also a degree of cultural value that is difficult to replace.

3.2.1.2 Although It Is Possible to Improve the Efficiency of Traditional Wood- and Charcoal-Burning Stoves, It Is Also Possible to Perform Far Worse than Traditional Stoves In the hands of an experienced user who is consciously trying to conserve fuel, a three-stone fire can be more than 15% efficient. Many "improved" stoves do no better. For example, one common method in constructing improved stoves encloses the fire in an earthen combustion chamber in an attempt to insulate the fire and minimize heat losses. Although the fire appears to be insulated because the outside of the stove remains cool to the touch, most of these stoves actually use more wood for a given cooking task than do many traditional stoves. The exterior of the stove remains cool to the touch because it is made of hundreds of kilos of rock and soil and also has very high thermal mass but relatively poor insulating qualities in the combustion zone. Rather than insulate the fire, the high mass draws heat from it and dissipates it throughout the massive structure. Stoves of this type may have efficiencies of much less than 10% and may have very high emissions due to relatively cool combustion temperatures. However, if the stove has a chimney, the pollutants will be removed from the kitchen, resulting in indoor concentrations of CO and PM that are lower than those in a household using a more fuel-efficient open fire. Figure 12 shows an example of a high mass earthen stove from Nicaragua. Note that the stove has no chimney and that the walls are blackened from years of exposure to smoke. Drawing on these lessons, some stove designers now employ insulating materials, such as ash and pumice, around the combustion zone to maintain high temperatures and minimize heat loss to the body of the stove. Figure 13 shows a sequence of pictures of an insulated stove with a chimney designed for use in Central America. A technician is filling the stove body with vermiculite, which is a cheap mineral with very good insulating properties.

3.2.1.3 Improving the Efficiency of Heat Transfer from the Stove to the Pot Can Save Wood but Lead to Higher Emissions of Harmful Pollutants The discussion of stove efficiency in the note to Table II defined combustion efficiency (CE) and heat transfer efficiency (HTE). The relation between them is such

FIGURE 12 High mass earthen stove from Nicaragua. Originally appeared in Stove Images (photo by Beatrix Westhoff), reproduced with permission of the author and publisher.

that the two quantities can work in opposite directions. For example, moving the pot closer to the stove improves the flow of heat into the food, but the pot is "cold" ($\sim$100°C) in relation to the burning wood and quenches the hot combustion gases, reducing CE and causing higher emissions. The trade-off between CE and HTE is difficult to quantify; it is dependent on the type of stove and fuel, and it is sensitive to user behavior. Even when it is quantifiable, deciding on the "correct" combination requires a trade-off between fuel savings (which are linked to local resources and household labor) and emissions (which are linked to health impacts and GHGs). An informed decision about this trade-off requires making value judgments that are not easy to make.

3.2.1.4 Users Usually Know What Works Best People who cook with woodfuels every day have valuable knowledge about stoves, fuels, and cooking practices more generally. Outside "experts" also have valuable knowledge but are often less equipped to design a stove for local use than are the users themselves. Users need to be involved in all steps of the design process, including design, testing, and dissemination. However, stove users might not be aware of some of the less visible impacts of woodfuel consumption, particularly health-related problems. This is where outside expertise becomes especially important and can be used to reinforce local knowledge so as to develop a safe and appropriate stove.

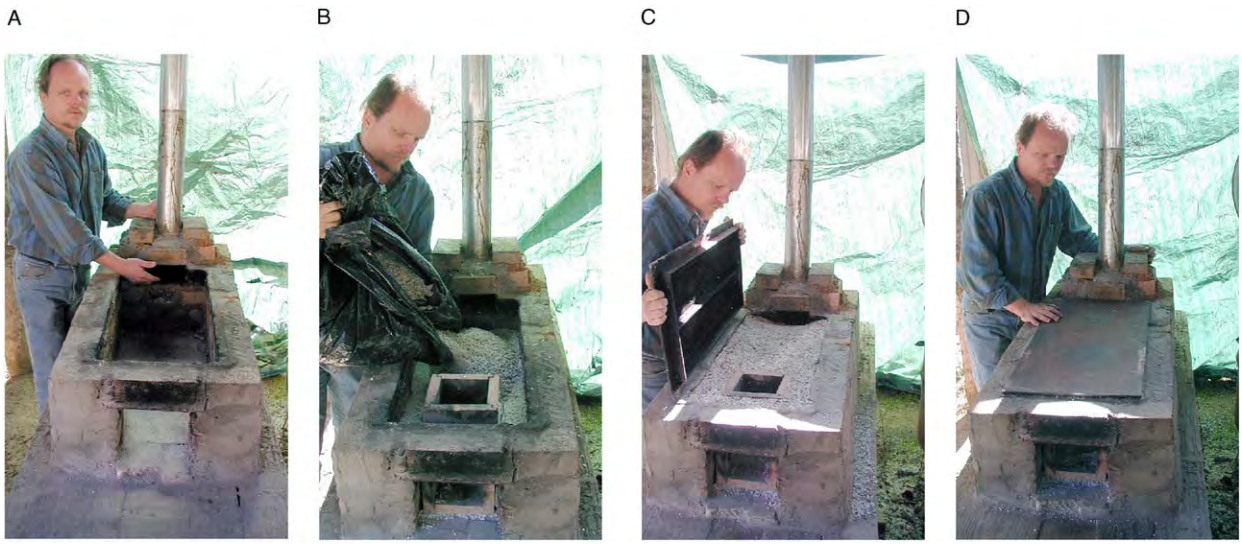


FIGURE 13 Improved stove with insulated combustion chamber. The sequence of photographs shows Brad van Appel of the Aprovecho Institute putting vermiculite insulation into an improved wood-burning stove. (A) Photo shows the brick body of the stove. This type of stove can be made from any durable material such as cement, clay, soil, or sheet metal. (B) The insulation is filled around the combustion chamber. In this stove, the combustion chamber is made from clay tiles placed together to form a hollow elbow. (C) A shallow channel is left on the surface of the insulation to force the flames and hot gases very close to the metal cooking surface before exiting through the chimney at the rear of the stove. (D) The metal cooking surface (called a *plancha* in Spanish) is installed over the insulation. Variations of this type of stove are being disseminated in parts of Central America, where many people prefer a flat metal surface for cooking. Photos by Dean Still (Aprovecho Institute).

3.2.2 Alternative Stoves and Fuels

Fuel switching, discussed earlier in the context of the energy ladder, is another way in which to mitigate the problems associated with household wood consumption. As the energy ladder model suggests, people may switch from wood to cleaner burning fuels as a result of increasing income. Alternatively, they may substitute fuels due to shifting household preferences or changing patterns of fuel availability that occur independent of changes in income. Although these factors can evolve without intervention, there are also policies and programs that influence these changes. For example, education campaigns about the health-damaging effects of woodfuel emissions can change household preferences, leading people to prioritize clean cooking fuels over other expenditures, whereas tax breaks for clean cooking fuels or subsidies for stoves encourage market expansion, making alternative fuels available to a wider range of consumers.

Substitution of wood for other fuels can go "up" or "down" the energy ladder. In rural communities where wood is scarce, households may switch to burning crop residues, dung, or processed briquettes made from wood waste, paper, or agricultural wastes. These fuels generally do not burn as well as does wood in conventional stoves, but fuel switching can be facilitated by the dissemination of specialized stoves made to burn alternative biomass fuels. In communities where wood scarcity is not a serious issue, better off households may decide that charcoal is preferable to wood, a fuel switch that may increase their net wood consumption due to inefficiencies in charcoal production. This switch also changes their patterns of exposure to harmful pollutants. Charcoal burning emits fewer particulates than does wood burning, reducing the risk of ARI, but charcoal also emits more CO, raising the risk of CO poisoning. Even more substantial decreases in exposure to harmful pollutants result from fuel switching to kerosene, LPG, or other gaseous fuels. Rural households rarely have the option of buying these fuels, but urban households do have that range of choices (see Fig. 8 for an example of this from Kenya). However, poor households do not have the same opportunity to switch to cleaner fuels due to the high initial costs of the stove relative to wood and charcoal stoves. Well-targeted subsidies, favorable credit mechanisms, and/or microfinance schemes can reduce the initial costs or spread them out over several years so that the initial costs are affordable by the poor (see Fig. 10).

In addition to kerosene and LPG, other clean alternatives to wood and charcoal exist, although

they are less common. These include biogas, a methane-rich gas produced by the anaerobic breakdown or digestion of human and/or animal waste, and producer gas, a low heating value gas derived from the thermochemical breakdown of biomass or coal. Both biogas and producer gas are used to some extent in India and China. Alcohol-based fuels are a third possibility for fuel switching. These can be derived from either woody or nonwoody biomass and are very clean burning; however, alcohol-based fuels need to be mass produced to take advantage of economies of scale. Thus, alcohol is very rarely used for cooking in less developed countries, although some projects have been attempted (e.g., an ethanol-based "gel fuel" is currently being promoted in parts of Southern Africa where ethanol production facilities have existed for some time). In any less developed country, alcohol would need to be cost-competitive with kerosene, a cheap and relatively clean-burning fuel that is often subsidized.

Ultimately, the policies and programs designed to influence fuel choice at the household level should reflect both community and national priorities. Household energy in less developed countries sits at the intersection of many important policy questions: forest conservation, public health, GHG emissions, and fossil fuel dependency. Weighing these issues together with more specific local needs so as to decrease the impacts of household energy consumption and increase the quality of life for people who are still reliant on wood for their energy needs remains an unmet challenge in most communities throughout the developing world.

SEE ALSO THE FOLLOWING ARTICLES

Energy Ladder in Developing Nations • Fire: A Socioecological and Historical Survey • Forest Products and Energy • Indoor Air Quality in Developing Nations • Rural Energy in China • Rural Energy in India • Solar Cookers • Women and Energy: Issues in Developing Nations

Further Reading

Agarwal, B. (1986). "Cold Hearths and Barren Slopes: The Woodfuel Crisis in the Third World." Zed Books, London.

Baldwin, S. F. (1986). "Biomass Stoves: Engineering Design, Development, and Dissemination." Center for Energy and Environmental Studies, Princeton, NJ.

Bradley, P. N., and Huby, M. (1991). "Woodfuel, Women, and Woodlots." Macmillan Education, Basingstoke, UK.

Ezzati, M., and Kammen, D. (2002). Evaluating the health benefits of transitions in household energy technology in Kenya. *Energy Policy* **30**(10), 815–826.

Ezzati, M., and Kammen, D. (2002). Household energy, indoor air pollution, and health in developing countries: Knowledge base for effective interventions. *Annu. Rev. Energy Environ.* **27,** 233–270.

Hosier, R. (1993). Energy policy: Special Issue. *Urban Energy Environ. Africa* **21,** 435–639.

International Energy Agency. (2002). Energy and poverty. *In* "World Energy Outlook 2002," p. 49. IEA, Paris.

Intergovernmental Panel on Climate Change. (2001). "Climate Change 2001: The Scientific Basis." Cambridge University Press, Cambridge, UK.

Kammen, D. (1995). Cookstoves for the developing world. *Sci. Am.* **273,** 72–75.

Leach, G., and Mearns, R. (eds.) (1988). "Beyond the Woodfuel Crises: People, Land, and Trees in Africa." Earthscan, London.

Smith, K. (1993). Fuel combustion, air pollution, and health: The Situation in developing countries. *Annu. Rev. Energy Environ.* **18,** 529–566.

Smith, K. R. (1987). Biofuels, Air Pollution, and Health: A Global Review. Plenum Press, New York.

Smith, K. R., Zhang, J., Uma, R., Kishore, V. V. N., Joshi, V., and Khalil, M. A. K. (2000). Greenhouse implications of household fuels: An Analysis of India. *Annu. Rev. Energy Environ.* **25,** 741–763.

Westhoff, B., and Germann, D. (1995). "Stove Images." Apsel Verlag, Frankfurt, Germany.

World Health Organization. (2002). "World Health Report: Reducing Risks, Promoting Healthy Life." WHO, Geneva, Switzerland.

Work, Power, and Energy

M. KOSTIC
Northern Illinois University
DeKalb, Illinois, United States

1. **Basic Concepts**
2. **Forms, Classifications, and Conservation of Energy**
3. **Work of Conservative and Nonconservative Forces: Work–Energy Principle**
4. **Energy, Work, and Power of Rotating Systems**
5. **Energy Sources, Conversion to Work, and Efficiency**
6. **Energy Reserves and Outlook**

Glossary

energy A fundamental property of a system referring to its potential to influence changes to other systems by imparting work (forced directional displacement) or heat (chaotic displacement/motion of system microstructure); energy exists in many forms—electromagnetic, electrical, nuclear, chemical, thermal, and mechanical, where electromechanical energy may be kinetic or potential—and thermal energy represents overall chaotic motion energy of molecules and related microstructure.

energy conservation May refer to the fundamental law of nature that energy and mass are conserved, that is, cannot be created or destroyed but only transferred from one form or one system to another; another meaning of energy conservation is improvement of efficiency of energy processes so that they could be accomplished with minimal use of energy sources and minimal impact on the environment.

energy conversion A process of transformation of one form of energy to another, for example, conversion of chemical to thermal energy during combustion of fuels or conversion of thermal to mechanical energy using heat engines.

energy efficiency Ratio between useful energy of a process and energy used to accomplish that process; energy, as per the conservation law, cannot be lost (destroyed), but part of energy input that is not converted into "useful energy" is customarily referred to as energy loss.

fuel cell Device undergoing electrochemical cycle where electrical work is obtained during ionization of fuel and oxidant molecules, reacting into product molecules at constant pressure and temperature, thereby avoiding heating and loss of work potential into thermal energy of the reaction products, like during combustion.

heat Inevitable (spontaneous) energy transfer due to temperature differences, to a larger or smaller degree, without control (dissipative) via chaotic displacement/motion of system molecules and related microstructure (in all directions, nonpurposeful), as opposed to controlled (purposeful and directional) energy transfer referred to as work.

heat engine Device undergoing thermomechanical cycle with mechanical expansion and compression net work obtained as difference between heat transferred to the engine from a high-temperature heat reservoir and rejected to a low-temperature reservoir, thereby converting part of thermal energy into mechanical work.

nonrenewable energy sources Energy sources created and accumulated over a very long period in the past, such as fossil and nuclear fuels, whose creation rate is many orders of magnitude smaller than the consumption rate, so that they will be depleted in a finite time period at the current rate of consumption.

power The energy rate per unit of time.

renewable energy sources The continuously or frequently available (renewed daily or at least annually) energy sources, such as solar energy, wind, water flows, ocean and tidal waves, and biomass, that are expected to be available forever for all practical purposes.

work A type of controlled energy transfer when one system is exerting force in a specific direction and, thus, is making a purposeful change (displacement) of the other systems; it is inevitably (spontaneously) accompanied, to a larger or smaller degree, by dissipative (without control) energy transfer referred to as heat.

A global overview and general concept of energy, its forms and classifications, and its manifestation as work and heat transfer is presented in this article with an objective to preserve the completeness and exactness of the fundamental concepts. The emphasis is given to work as the most useful energy transfer. However, because most work currently is obtained from fossil fuels using heat engines, and because all work is ultimately dissipated via heat to thermal

 © 2004 Elsevier Inc. All rights reserved.

energy, they are presented in the article, along with fuel cells, as promising devices to convert chemical fuel energy directly into electrical work without degradation of energy as in the combustion processes. The efficiencies of diverse energy-to-work conversion processes are provided. Finally, the latest developments and worldwide energy production and consumption data are effectively summarized, and diverse energy sources, reserves, and a future outlook are given.

1. BASIC CONCEPTS

Energy is the "building block" and fundamental property of matter and space and, thus, the fundamental property of existence. Energy exchanges or transfers are associated with all processes (or changes) and, thus, are indivisible from time. It is no wonder that energy is often defined as the ability to perform work, that is, as a potential for energy transfer in a specific direction (displacement in force direction) to achieve a purposeful process, as opposed to a dissipative (less purposeful) energy transfer in the form of heat. This definition of energy could be generalized as the ability to perform change.

Work is a mode of energy transfer from one acting body (or system) to another resisting body (or system), with an acting force (or its component) in the direction of motion, along a path or displacement. A body that is acting (forcing) in motion direction in time is doing work on another body that is resisting the motion (displacement rate) by an equal resistance force, including inertial force, in the opposite direction of motion. The acting body is imparting (transferring away) its energy to the resisting body, and the amount of energy transfer is the work done by the acting onto the resisting body, equal to the product of the force component in the motion direction multiplied by the corresponding displacement or vice versa (force multiplied by the displacement component in the force direction) (Fig. 1). If the force ($\vec{F}$) and displacement vectors ($\overrightarrow{ds} = \overrightarrow{dr}$) are not constant, integration of differential work transfers from initial state 1 to final state 2, defined by corresponding position vectors $\vec{r}$, will be necessary (Fig. 2).

The work is a directional energy transfer; however, it is a scalar quantity like energy, and it is distinctive from another energy transfer in the form of heat due to random motion (chaotic, in all directions) and collisions of bodies' molecules and their components.

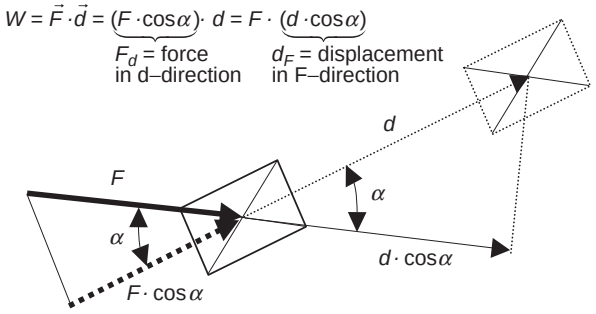


FIGURE 1 Work, force, and displacement.

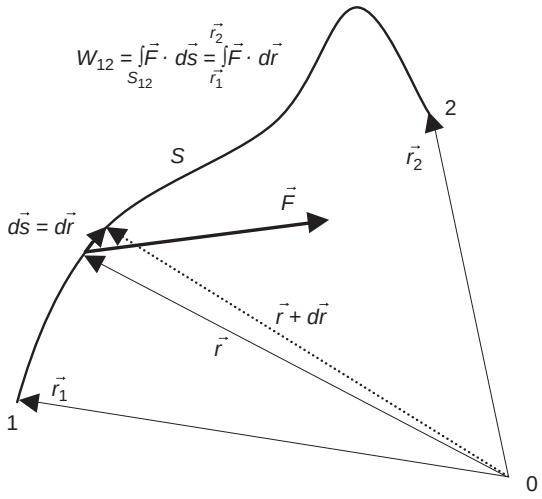


FIGURE 2 Work along arbitrary path.

1.1 Power

Energy transfer in the form of work or heat is taking place in time, at a finite rate, and cannot be instantaneous. The energy transfer per unit of time, or energy rate, is called power, usually in form of the work rate, although it may also be heating power, that is,

$$\dot{W} = \frac{dW}{dt} = \frac{\vec{F} \cdot \overrightarrow{ds}}{dt} = \vec{F} \cdot \vec{v}, \quad \text{or} \quad W_{12} = \int_{t_1}^{t_2} \dot{W} \cdot dt. \tag{1}$$

Therefore, energy (as a body or system property) and work and heat (as energy in transfer from one system to another) are interrelated. All have the same dimension (Force × Length), and related different units could be converted to each other (Table I).

Work transfer cannot occur without the existence of a resisting body or system, nor can it occur without finite displacement in the force direction. This might not always be obvious. For example, if

TABLE I

Energy Units with Conversion Factors and Energy Equivalents

Energy unit	J	kWh	Btu
1 joule (J)	1	2.78×10^{-7}	9.49×10^{-4}
1 kilowatt-hour (kWh)	3.6×10^{6}	1	3.412×10^{3}
1 kilocalorie (kcal = Cal = 1000 cal)	4187	1.19×10^{-3}	3.968
1 British thermal unit (Btu)	1055	2.93×10^{-4}	1
1 pound-force foot ($lb_f \cdot ft$)	1.36	3.78×10^{-7}	1.29×10^{-3}
1 electron volt (eV)	1.6×10^{-19}	4.45×10^{-26}	1.52×10^{-22}
1 horsepower × second (HP · sec)	745.7	2.071×10^{-4}	0.707

Energy equivalent	J	kWh	Btu
1 barrel (42 gallons) of crude petroleum	6.12×10^{9}	1700	5.80×10^{6}
1 ton (2000 lb) of bituminous coal	2.81×10^{10}	7800	2.66×10^{7}
1000 cubic feet of natural gas	1.09×10^{9}	303	1.035×10^{6}
1 gallon of gasoline	1.32×10^{8}	36.6	1.25×10^{5}
1 gram of uranium-235	8.28×10^{10}	2.30×10^{4}	7.84×10^{7}
1 gram of deuterium	2.38×10^{11}	6.60×10^{4}	2.25×10^{8}
2000 dietary food calories (2000 kcal)	8.374×10^{6}	2.326	7.937×10^{3}
1 solar constant × $(cm)^2$ × minute	8.374	2.326×10^{-6}	7.937×10^{-3}

one is holding a heavy weight or pushing hard against a stationary wall, there will be no work done against the weight or wall (neglecting its small deformations). However, one will be doing work internally due to contraction and expansion of one's muscles (force with displacement) and that way will be converting (spending) a lot of chemical energy via muscle work and then dissipating it into thermal energy and heat transfer (sweating and getting tired).

1.2 Energy, Power, Work and Heat Units, and Energy Equivalents

Energy is manifested via work and heat transfer, with the corresponding Force × Length dimension for work ($N \cdot m$, $kg_f \cdot m$, and $lb_f \cdot ft$ in SI, metric, and English systems of units, respectively), and the caloric units in kilocalories (kcal) or British thermal units (Btu), as heat needed to increase a unit mass of water (at a specified pressure and temperature) for 1° of temperature, so that the water-specific heat is 1 kcal/(kg °C) = 1 Btu/(lb °F) by definition, in the metric and English systems of units, respectively. It was demonstrated by Joule that 4187 N · m of work, when dissipated in heat, is equivalent to 1 kcal. In Joule's honor, 1 N · m of work is named after him as 1 joule or 1 J, the SI energy unit, also equal to electrical work of $1\,W \cdot s = 1\,V \cdot A \cdot s$. The SI unit for power, or the work rate, is the watt (i.e., 1 J/s = 1 W) as well as corresponding units in other systems of units, for example, Btu/h. The horsepower is defined as $1\,hp = 550\,lb_f \cdot ft/s = 745.7\,W$. Other common units for energy, work, and heat, and for energy equivalents for typical fuels and processes, are given in Table I.

2. FORMS, CLASSIFICATIONS, AND CONSERVATION OF ENERGY

As stated in the previous section, energy is a fundamental concept indivisible from matter and space, and energy exchanges or transfers are associated with all processes (or changes) and, thus, are indivisible from time. Energy transfer is needed to produce a process to change other system properties. Among all properties, the energy is the only one that could be converted to mass and vice versa, that is, $E = mc^2$ (known in some literature as mass energy, the c is the speed of light in a vacuum); thus, the two are interrelated.

Any and all changes (happening in time) are caused by energy exchanges or transfers from one body or substance (system or subsystem) to another. A part of a system may be considered as a subsystem if energy transfer within a system is taking place; conversely, a group of interacting systems may be

considered as a larger isolating system if they do not interact with the rest of the surroundings.

There are many forms and classifications of energy (Table II) all of which could be classified as microscopic (internal within a system microstructure) and/or macroscopic (external as related to the system mass as a whole with reference to other systems). Furthermore, energy may be quasi-potential (associated with a system equilibrium state and structure, i.e., system property) or quasi-kinetic (energy in transfer from one system or structure to another in the form of work or heat).

Every system in an equilibrium is defined by its state potential, where the system structure is "held" by forces (i.e., the forces "define" the system potential and state, and vice versa—action and reaction); otherwise, a system will undergo dynamic change (in time), with quasi-kinetic energy exchange toward another equilibrium. Atoms (mass) are held by atomic and gravity forces; otherwise, mass would disintegrate ("evaporate" into energy) as it does partly in nuclear reactions (nuclear energy). Molecules are held by electrochemical bonding (valence) forces (chemical reactions—chemical energy). Solids are held by "firm" intermolecular forces (melting of solids when energy is increased by heating—latent thermal energy). Liquids are also held by latent intermolecular forces (evaporation when kinetic energy of molecules is further increased by heating—latent thermal energy again). Sensible thermal energy represents energy of molecular motion and is related to temperature of a system. "Holding" potential forces may be "broken" by energy transfer (e.g., heating, high-energy particle collisions). States and potentials are often stable or "hooked' and, thus, need to be unhooked or broken to overcome the existing "threshold" or equilibrium, as in igniting combustion or starting a nuclear reaction. Energy can be directional (purposeful or organized) or chaotic (dissipative or disorganized). For example, mass-in-motion (mechanical kinetic energy) and electricity-in-motion (electrical kinetic energy) are organized kinetic energies, whereas thermal energy is disorganized chaotic energy of motion of molecules and atoms. System energy may be defined with reference to position in a vector force field such as elastic potential (stress) energy, gravitational potential energy, or electromagnetic field energy.

Energy transfer may be in organized form (work transfer due to force action) or in chaotic disorganized form (heat transfer due to temperature difference). Energy transfer into a system builds up energy

TABLE II

Energy Forms and Classifications

Scale			Type			Transfer (release)	
				Kinetic (motion)		Quasi-kinetic (motion)	
Macro/external (mass based)	Micro/internal (structure based)	Form	Potential (state or field)	Directional[a]	Chaotic dissipative	Work directional[a]	Heat dissipative
		Mechanical					
X		• Kinetic		X		X	
X		• Gravitational[b]	X			X	
X		• Elastic	X			X	
		Thermal					
	X	• Sensible			X		X
	X	• Latent	X				X
	X	Chemical	X				X
	X	Nuclear	X				X
		Electrical					
	X	• Electrokinetic		X		X	
	X	• Electrostatic	X			X	
	X[c]	Electromagnetic	X[c]			X	

[a] Electromechanical kinetic energy type (directional/organized, the highest energy quality) is preferable because it may be converted to any other energy form/type with high efficiency.

[b] Due to mass position in a gravitational field.

[c] Electromagnetic form of energy is the smallest known scale of energy.

potential (called simply potential for short, e.g., pressure, temperature, voltage) over energy displacement (e.g., volume, entropy). Conversely, if energy is transferred from a system, its energy potential is decreased. That is why energy is transferred from higher to lower energy potential only.

All organized kinetic energy will, in part or in whole (and ultimately in whole), disorganize/dissipate within the microstructure of a system (over its mass and space and in time) into disorganized thermal energy. Entropy, as an energy displacement system property, represents the measure of energy disorganization. Contrary to energy and mass, which are conserved in the universe, the entropy is continuously generated (increased) due to continuous disorganization of energy in transfer ("spreading" of energy toward and over lower potentials in time) and, thus, degradation of quality of energy.

Often, one wants to extract energy from one system to purposefully change another system, that is, to transfer energy in organized form (as work). That is why energy is often defined as the ability to perform work, and a special quantity exergy is defined as the maximum possible work that may be obtained from a system by bringing it, in a process, to equilibrium with reference surroundings. The maximum possible work will be obtained if one prevents energy disorganization or dissipation, that is, within so-called reversible processes. Because the energy and mass are conserved during any process, only in ideal reversible processes will the entropy (measure of energy disorganization) and exergy (maximum possible work with reference to the surroundings) be conserved, whereas in the real irreversible processes, the entropy will be generated and exergy will be partly (or even fully given enough time) destroyed. Therefore, heat transfer and thermal energy are universal manifestations of all natural and artificial (manmade) processes, whereas all organized potentials and/or quasi-kinetic energies are disorganized or dissipated in the form of thermal energy in irreversible and spontaneous processes.

As stated previously, there are many different energy forms and types (Table II). One is usually not interested in (absolute) energy level; instead, one is usually interested in the change of energy (during a process) from an initial state (i) to a final state (f), so that zero reference values for different energy forms are irrelevant and are often taken arbitrarily for convenience. The following are basic correlations for energy changes of several typical energy forms often encountered in practice: motion kinetic energy (E_K) as a function of system velocity (v), spring elastic potential energy (E_{Ps}) as a function of spring deformation displacement (x), gravitational potential energy (E_{Pg}) as a function of gravitational elevation (z), and sensible thermal energy (E_U) as a function of system temperature (T):

$$\Delta E_K = \frac{1}{2}m\left(v_f^2 - v_i^2\right);\ \ \Delta E_{Ps} = \frac{1}{2}k\left(x_f^2 - x_i^2\right)$$
$$\Delta E_{Pg} = mg(z_f - z_i);\ \ \Delta E_U = mc_v(T_f - T_i). \qquad (2)$$

If the reference energy values are taken to be zero when the initial (i) variables are zero, these equations will represent the energy values for the final values (f) of the corresponding variables. If the corresponding parameters, spring constant k, gravity g, and constant-volume specific heat c_v, are not constant, integration of differential energy changes from initial to final states will be necessary.

Energy transfer via work W and heat transfer Q may be expressed as the product of related energy potentials (pressure P or temperature T) and corresponding energy displacements (change of volume V and entropy S, respectively), that is,

$$W_{12} = \vec{F}\cdot\vec{d} = \left[(P\cdot\underbrace{A\vec{n}}_{\Delta V})\cdot\vec{d}\right] = P\cdot\Delta V_{12}|_{P\neq const}$$
$$= \int_{V_1}^{V_2} P\cdot dV \qquad (3)$$

$$Q_{12} = T\cdot\Delta S_{12}|_{T\neq const} = \int_{S_1}^{S_2} T\cdot dS. \qquad (4)$$

Note that in Eq. (3), force cannot act at a point but is distributed as pressure P over some area A (with orthogonal unit vector $\vec{n}$), which when displaced will cause the volume change ΔV.

In the absence of a nuclear reaction (no conversion of mass into energy, $E = mc^2$), mass and energy are conserved separately for an isolated system, a group of isolated systems, or the universe. For example, for a given process (energy exchange) between two systems, isolated from the surroundings, during time period Δt, the energy conservation law may be expressed as

$$\sum_i E_{1,i}(t+\Delta t) + \sum_j E_{2,j}(t+\Delta t)$$
$$= \sum_m E_{1,m}(t) + \sum_n E_{2,n}(t), \qquad (5)$$

where $\sum_i E_{1,i}(t)$ is the sum of all existing energy forms i of system 1 at time (t) and of system 2 and/or at time $t+\Delta t$. Then, from Eq. (5), energy transfer from system 1 to system 2 for period Δt may be in the

form of work W and/or heat Q, that is,

$$\begin{aligned} W_{12} + Q_{12} &= W_{1\to 2} + Q_{1\to 2} \\ &= \sum_i E_{1,i}(t+\Delta t) - \sum_m E_{1,m}(t) \\ &= -\left[\sum_j E_{2,j}(t+\Delta t) - \sum_n E_{2,n}(t)\right] \\ &= -(W_{21} + Q_{21}) = -(W_{2\to 1} + Q_{2\to 1}). \end{aligned} \tag{6}$$

Note that for nonexistent terms in these equations, their values should be zero, and if the energy transfers (work and heat) are negative, they are actually in the opposite direction than was specified. Correlations are similar for mass conservations. Furthermore, the power is defined as energy transfer rate per unit of time; conversely, energy transfer is equal to the product of power and time or the integral of power over a time period, see Eq. (1).

3. WORK OF CONSERVATIVE AND NONCONSERVATIVE FORCES: WORK–ENERGY PRINCIPLE

Work against conservative (also known as internal, volumetric, or space potential field) and/or inertial forces is path independent, and the mechanical energy is conserved during such a process. However, work of nonconservative dissipative forces is path dependent, and part of the mechanical energy is converted (dissipated) to the thermal energy. The nonconservative forces are also external or surface forces.

For example, during a free gravity fall (or free bounce) without air friction, the potential energy is being converted to kinetic energy of the falling body (or vice versa for free bounce), and at any time the total mechanical energy (sum of kinetic and potential mechanical energies) is conserved, that is, stays the same (Fig. 3). The mechanical energy is also conserved if a mass freely vibrates on an ideally elastic spring or if a pendulum oscillates around its pivot, both in the absence of dissipative effects such as friction and nonelastic deformation. In general, for work of conservative forces only, the mechanical energy (E_{mech}) for N isolated systems is conserved because there is no dissipative conversion in thermal energy and, thus, no heat transfer, that is,

$$\begin{aligned} E_{mech} &= E_K + E_{Pg} + E_{Ps} \\ &= \sum_{j=1}^{N}\left(\frac{1}{2}mv^2 + mgz + kx\right)_j = const. \end{aligned} \tag{7}$$

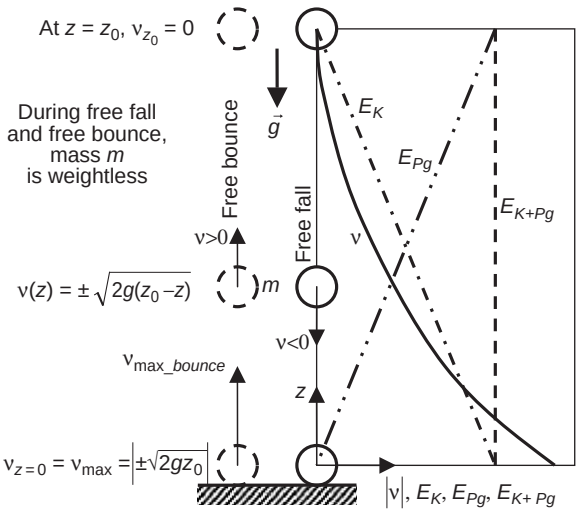


FIGURE 3 Energy and work due to conservative gravity force.

When work of nonconservative forces W_{nc} is exchanged among N isolated systems, from an initial state i to a final state f, the total mechanical energy of all systems is reduced by that work amount, that is,

$$W_{nc,i\to f} = \left(\sum_{j=1}^{N} E_{mech,j}\right)_i - \left(\sum_{j=1}^{N} E_{mech,j}\right)_f. \tag{8}$$

This correlation is known as the work–energy principle.

Regardless of the traveled path (or displacement), the work against conservative forces (e.g., gravity or elastic spring in the preceding cases) in the absence of any dissipative forces will depend on the final and initial positions (or states) only. However, the work of nonconservative dissipative forces W_{nc} will depend on the traveled path because the energy is dissipated during the force displacement, and mechanical energy will not be conserved but rather partly converted (via dissipation and heat transfer) into the thermal energy (see Eq. 8). This should not be confused with total energy conservation, which is always the case, and it must include both work and heat transfer (see Eqs. 5 and 6).

4. ENERGY, WORK, AND POWER OF ROTATING SYSTEMS

Motors and engines are important energy devices that convert other energy forms into their rotating shaft motion with a twisting action and, thus, into the mechanical work and kinetic energy. Also, many other devices, including parts of diverse machineries involving energy transfers in the form of work, are

mounted on rotating shafts or rotating bearings or pins (e.g., disks, gears, wheels, levers). Because shafts and bearings restrict motion to rotation around their axes, any mass at distance r from the axis of rotation, its circumferential velocity v, and the related twisting circumferential force F are related to the corresponding angular rotational speed ω and torque T with the simple correlations $v = \omega r$ and $T = rF$ (written in scalar form because rotational motion and twisting force action are restricted in a single, around-the-axis direction only). Then, the kinetic energy E_K and work rate dW/dt, or the power transfer of a rotating system, can be expressed as seen in Fig. 4, where J is the mass moment of inertia for a rotating system. For a cylinder (or disk) of radius R and mass m, or of density ρ and thickness b, the mass moment of inertia around its axis is $J = mR^2/2 = (\pi/2)\rho bR^4$.

5. ENERGY SOURCES, CONVERSION TO WORK, AND EFFICIENCY

Energy sources are those systems (substances or natural phenomena) that allow for abundant, convenient, and efficient (and thus economical) conversion of their energy into useful energy forms (for consumption needs), usually thermal for heating or mechanical and electrical for work, with the latter also being very convenient for transmission and very efficient for conversion into any other useful energy form. Because energy consumption needs are time and location dependent, energy conversion rate, energy density (per unit mass, volume, area, etc.), transportation (transmission), and storage are important.

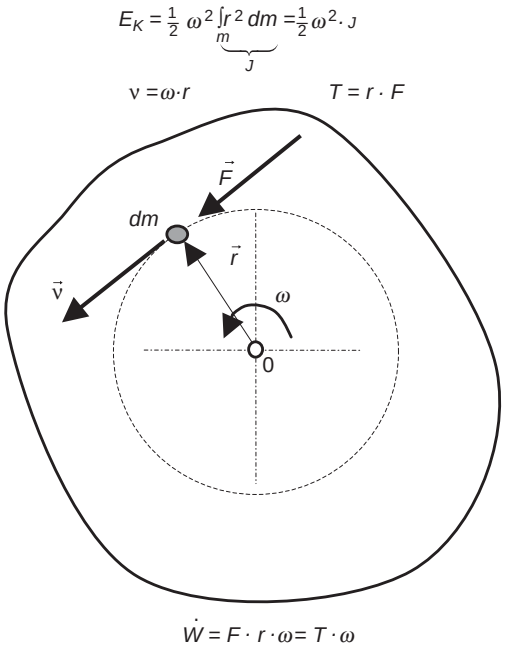


FIGURE 4 Energy and power of rotating system.

There are many sources of energy (Table III) for diverse needs of human activities and society in general. Energy consumption may be classified into four general sectors (see Table VI): (1) *residential*: for appliances and lighting, space heating, water heating, air conditioning, and so on; (2) *commercial*: for lighting, space heating, office equipment, water heating, air conditioning, ventilation, refrigeration, and so on; (3) *industrial*: for water and steam boilers, direct process energy, machine drive, and so on; and (4) *transportation*: for personal automobiles, light and heavy trucks, air transport, water transport, pipe transport, rail transport, and so on. In all four sectors, in addition to primary energy sources, electrical energy, as a secondary energy source produced from primary energy sources, is used extensively.

The primary energy sources for the world in the year 2000 and for the United States in the year 2001 are presented in Table IV. The primary energy sources for the production of electricity are presented in Table V. In addition, the U.S. energy supply by consumption sector, including electricity production, is given in Table VI. The world and U.S. populations, energy production, and energy consumption are summarized in Table IV as well. The total energy production, including losses, import, and export, is available as energy supply for consumption and storage. Also, most of the world's electricity ($\sim$64% vs $\sim$70% in the United States) is produced from fossil fuels, with an overall conversion efficiency of only approximately 30%. The conversion efficiency is similar in nuclear power plants, which contribute approximately 17% of world electricity production and approximately 21% of U.S. electricity production. When the global energy supply is given together with fossil fuels and is expressed in British thermal units, all electrical energy (including hydro, wind, etc.) is given in equivalent British thermal units, accounting for the conversion efficiency (the U.S. Energy Information Agency uses 31%). When electrical energy is accounted for separately, the actual electrical output is given in kilowatt-hours, as seen in Tables V and VI. Because of different forms and conversion efficiencies of primary energy sources, and because of complexities of energy production and losses, transportation and storage, and import and export, it is virtually impossible to account correctly for all energy paths and forms in

TABLE III

Primary Energy Sources and Conversion to Work

Primary energy source		Conversion
Nonrenewable		
Fossil fuels	Coal	Combustion (heat and heat engine H/HE/W[a])
	Peat	
	Oil/Crude petroleum	
	Natural gas	
Nuclear	Uranium	Fission (H/HE/W)
	Thorium	
	Deuterium	Fusion[b] (H/HE/W)
Renewable[c]		
Geothermal[d]	Hot steam/water	H/HE/W
	Ground soil/rock heat	
	Volcanic, etc[b]	
Ocean–gravitational	Tidal–ocean wave	Direct to work
Solar related		
Ocean	Ocean thermal	H/HE/W
	Ocean currents	Direct to work
	Ocean wave	
Biomass	Wood	Combustion (H/HE/W)
	Vegetation, etc[e]	
Direct solar	Solar–thermal	H/HE/W
	Photoelectric	Direct to work
	Photochemical	
Electrostatic	Lightning, etc[b]	
Wind	Wind–air streams	
Hydro	River/Accumulation	
Muscular	Human and animals	

Note. Secondary energy sources (e.g., electrical, synthetic fuels, hydrogen) with energy storage and distribution completes the energy supply domain, which with energy needs, consumption, and efficiency completes the global energy realm.

Energy related processes: Electromagnetic radiation; Photosynthesis cycle in nature; Biosyntheses cycle in nature; Electrical processes: electrodynamic, electromagnetic, electrochemical; Nuclear reactions: fission, fusion, radioactive radiation; Chemical reactions: combustion, oxidation, etc.; Heat transfer and frictional dissipative processes; Thermomechanical expansion and compression; Natural air streams driven by solar dissipation, gravitation, and buoyancy (evaporation, precipitations, water streams); Natural water streams (rivers and ocean streams); Mechanical expansion and compression.

[a] H/HE/W, conversion to heat and via heat engine to work.

[b] Not commercialized yet.

[c] All renewable sources, except tidal and geothermal, are due to solar radiation.

[d] Usually renewable but may be nonrenewable.

[e] Includes many types as well as waste/garbage.

the same units; therefore, the total figures (and percentages) usually do not add up exactly.

Fossil fuels share more than 85% of the total world and U.S. energy consumption (Table IV). Nearly 40% of the total world and U.S. energy is used for electricity production (Tables IV, V, and VI), mainly in thermal and nuclear power plants ($>80\%$ in the world and $>90\%$ in the United States), using heat engines undergoing thermomechanical conversion processes with relatively low conversion efficiencies (Table VII). The overall conversion efficiencies from chemical or nuclear fuel energy to thermal energy of combustion gases or steam, and to mechanical and electrical energy, is only approximately 30 to 35%.

5.1 Heat Engines

Heat engines are devices undergoing thermomechanical cycles, similar to the one in Fig. 5, with a mechanical expansion and compression net work ($W = Q_h - Q_c$), obtained as the difference between the heat transferred to the engine from a high-temperature heat reservoir (at T_h) and rejected to a low-temperature heat reservoir (at T_c), thereby converting part of thermal energy into mechanical work. The combustion process itself is an irreversible one, where chemical energy (electrochemical energy binding reactants' molecules) is chaotically released during combustion (i.e., converted in random thermal energy of products' molecules) and cannot be fully transferred into directional work energy. The second law of thermodynamics limits the maximum amount of work that could be obtained from thermal energy between two thermal reservoirs at different temperatures, hot T_h and cold T_c, by using the ideal reversible Carnot cycle (Fig. 5) with thermal efficiency:

$$\begin{aligned}\eta_{th,C_ad} &= \frac{W}{Q_h} = \frac{Q_h - Q_c}{Q_h} \\ &= 1 - \frac{T_c}{T_h}\bigg|_{T_h = T_{ad} = 2273\ K,\ T_c = 293\ K} \\ &= 1 - \frac{293}{2273} = 87.1\%, \end{aligned} \tag{9}$$

where $W = W_T - W_C$ is the net work of expansion, usually turbine W_T, and compression W_C. The maximum efficiency is achieved if heat is supplied at the highest possible temperature T_h and released at the lowest possible temperature T_c. However, both temperatures are limited by the fact that a fuel combustion is performed using oxygen with ambient air, resulting in maximum so-called adiabatic, stoichiometric combustion temperature T_{ad}, which

TABLE IV

World and U.S. Total Energy Supply by Source

	World, 2000		United States, 2001	
Source	Quadrillion Btu	Percentage	Quadrillion Btu	Percentage
Coal	92.51	23.3	21.93	22.6
Petroleum	155.25	39.1	38.23	39.5
Natural gas	90.83	22.9	23.22	24.0
Fossil fuels	338.59	85.3	83.38	86.0
Nuclear electric	25.51	6.4	8.03	8.3
Hydroelectric	27.46	6.9	2.38	2.5
Renewables/Others	5.36	1.4	3.11	3.2
Total	396.92	100	96.90	100

World and U.S. Population and Energy Comparisons

	World, 2000		United States, 2000	
	Quadrillion Btu	Percentage	Quadrillion Btu	Percentage
Population (millions)	6080.1	100	281.4	4.6
Energy production	397	100	71.06	17.9
Energy consumption	398	100	99.32	25.0

Source. Energy Information Administration. (2003). "Annual Energy Review 2001." U.S. Department of Energy, Washington, DC.

TABLE V

World and U.S. Electric Energy Supply by Source

	World, 2000		United States, 2001	
Source	Billion kWh	Percentage	Billion kWh	Percentage
Coal			1878.54	51.4
Petroleum			127.79	3.5
Natural gas			563.99	15.4
Fossil fuels	9374.6	63.8	2570.32	70.4
Nuclear electric	2434.2	16.6	758.79	20.8
Hydroelectric	2646	18.0	201.23	5.5
Renewables/Others	241.9	1.6	123.06	3.4
Total	14697	100	3653.40	100

Source. Energy Information Administration. (2003). "Annual Energy Review 2001." U.S. Department of Energy, Washington, DC.

for most fuels is approximately 2000°C or $T_{ad} = 2273\,\text{K}$. A part of the heat supplied at hot temperature T_h must be released to the surroundings at a cold temperature of approximately $T_c = 20°\text{C} = 293\,\text{K}$, resulting in a Carnot efficiency of 87.1% (see Eq. 9 and Fig. 5). However, the fuel heating value energy ($Q_{HV} = Q_{ad_var}$) is not available entirely at the adiabatic temperature of the products but rather is distributed over their variable temperature range, from initial surrounding temperature before combustion T_c to final adiabatic temperature T_{ad} (Fig. 6). If a single Carnot cycle is used at constant temperature $T_h < T_{ad}$, only fraction $(T_{ad} - T_h)/(T_{ad} - T_c)$ of the fuel heating value will be used, resulting in a reduced Carnot efficiency:

$$\eta_{th,C} = \left(1 - \frac{T_c}{T_h}\right)\left(\frac{T_{ad} - T_h}{T_{ad} - T_c}\right)\Bigg|_{\text{max for } T_h = \sqrt{T_{ad}T_c}} = \frac{\sqrt{T_{ad}/T_c} - 1}{\sqrt{T_{ad}/T_c} + 1}\Bigg|_{T_{ad} = 2273\ K,\ T_c = 293\ K} = 47.2\%. \quad (10)$$

The efficiency may be increased further by employing a large number (infinite in limit) of ideal

TABLE VI

U.S. Energy Consumption by Sector, 2001

	Primary		Electric		Total	
Sector	Quadrillion Btu	Percentage	Quadrillion Btu	Percentage	Quadrillion Btu	Percentage
Residential	6.9	7.1	13.3	35.5	20.2	20.8
Commercial	4.2	4.3	13.2	35.2	17.4	18.0
Industrial	21.6	22.3	11	29.3	32.6	33.6
Transportation	26.7	27.6	0	0.0	26.7	27.6
Electric	37.5	38.7				
Total	96.9	100.0	37.5	100	96.9	100

Source. Energy Information Administration. (2003). "Annual Energy Review 2001." U.S. Department of Energy, Washington, DC.

TABLE VII

Energy-to-Work Conversion Efficiencies

Engine/Process	Efficiency Percentage
Otto (gasoline) engine	25–30
Diesel engine	30–40
Gas turbine	30–40
Steam turbine	30–40
Nuclear, steam turbine	30–35
Combined gas/steam turbines	40–55+
Fuel cell (hydrogen, etc.)	40–60+
Photovoltaic cell	10–20
Windmill	30–40 (59% limit)
Hydro turbine	80–85
Electromechanical motor/generator	80–95

Note. Thermal-to-mechanical work conversion is limited by stoichiometric combustion temperature and the Carnot cycle efficiency. Fuel cell efficiency is limited by Gibbs free energy values for process reactants and products and may be close to 100%. Because of material property limitations and process irreversibilities (dissipation of energy), practical efficiencies are much lower and there is room for substantial improvements. For example, existing hybrid cars have 80% improved efficiency (and mileage) over the same classical cars, from 25 to 45%, by using electromechanical engines/storage hybrid systems.

Carnot engines operating at various temperatures (with $dW = dQ$) or with variable hot temperature heat exchange (Fig. 6). Assuming constant specific heat of combustion products, and after integration, the variable hot temperature Carnot cycle yields the maximum possible combustion products-to-work conversion efficiency:

$$\eta_{th,C\ \text{var max}} = \left(1 - \frac{\ln(T_{ad}/T_c)}{(T_{ad}/T_c) - 1}\right)\Bigg|_{T_{ad}=2273\ K,\ T_c=293\ K} = 69.7\%. \quad (11)$$

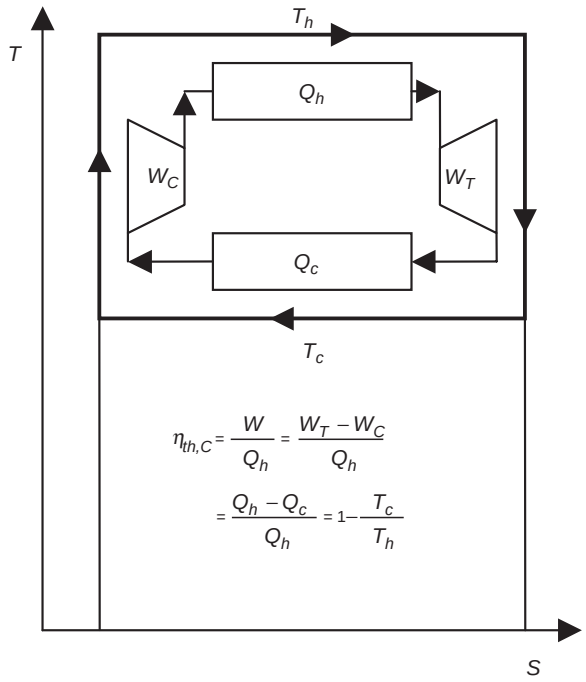


FIGURE 5 Heat engine, ideal Carnot cycle.

Because of engine material property limitations and other unavoidable irreversibilities, it is impossible to reach the ideal Carnot efficiency. Different actual heat engines undergo similar but different cycles, depending on the system design. For example, internal combustion engines undergo the Otto cycle with gasoline fuel and the Diesel cycle with diesel fuel, whereas steam and gas turbine power plants undergo the Rankine and Brayton cycles, respectively. However, with improvements in material properties, effective component cooling, and combining gas and steam turbine systems, more than 50% efficiencies are being achieved, a substantial improvement over current ones of approximately

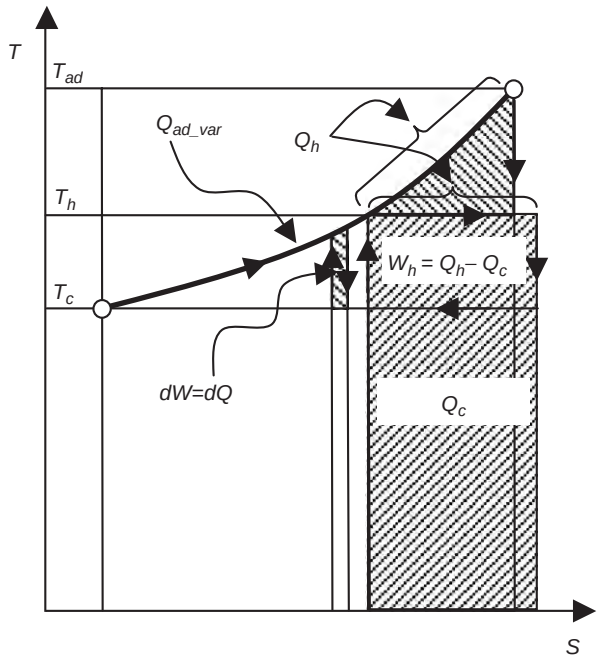


FIGURE 6 Heat engine, constant and variable temperature, ideal Carnot cycle.

35% (see Table VII). The ideal Carnot cycle is an important reference to guide researchers and engineers to better understand limits and possibilities for new concepts and performance improvements of real heat engines.

5.2 Fuel Cells

It is possible to convert fuel chemical energy to electrical work directly, without heating the reaction products as is done during combustion, by using an electrochemical cell, also known as a fuel cell. It is a very simple device, similar to a fueled battery. It consists of two porous metal electrodes with an electrolyte in between. The fuel (e.g., hydrogen), after diffusing through one electrode, will dissociate into positive ionic components by releasing electrons to that electrode. In the process, the positive fuel ions are transferred to another electrode that is in contact with oxygen. If an external load is connected to the electrodes, an electric current (and electrons) will be flowing and electrical work will be departed to the load, while hydrogen fuel's positive ions will be diffusing through the electrolyte and at another electrode will be reacting with oxygen, forming water as the reaction product. The high efficiency of the fuel cell results from the electrochemical process, whereby the energy of binding reactants' electrons, is converted directly into directional electrical work, departed by the electrons that move in the cell's external circuit. This is in contrast to conversion of the same energy during combustion to chaotic thermal motion (i.e., heating of combustion products' molecules), accompanied by substantial irreversibility and loss of work potential. The maximum efficiency and work from a fuel cell is limited, according to the second law of thermodynamics, to the decrease of the Gibbs free energy ($F = H - TS$) of the reactants (*re*) as they form products (*pr*) in the electrochemical reaction (i.e., $W_{max} = F_{re} - F_{pr}$), where $H = E_U + PV$ is a system and/or component enthalpy and E_U, P, V, T, and S are internal thermal energy, pressure, volume, temperature, and entropy, respectively. The difference between reactants' and products' Gibbs free energies is very close to the fuel heating value Q_{HV}, so that the fuel cell theoretical energy conversion efficiency, $\eta_{max} = W_{max}/Q_{HV}$, is close to 100% if the process is conducted at constant pressure and temperature and without dissipative heating of the products. In practice, with finite reaction rates accompanied by dissipative heating and the need for cooling to maintain reaction at as constant a temperature as possible, the efficiency is lower than the maximum possible but more than 40% or even 50%, which is still higher than the heat engine efficiencies given in Table VII along with efficiencies of other energy conversion processes to mechanical or electrical work.

6. ENERGY RESERVES AND OUTLOOK

Currently, most of the world's energy consumption is supplied by fossil fuels ($\sim$85%). However, proven fossil fuel reserves are limited, and it is estimated that if they continue to be used at the current rates (as used under current conditions), coal will be depleted in approximately 250 years, oil will be depleted in 60 years, and natural gas will be depleted in 80 years. One must keep in perspective the fact that "proven reserves" refers to the customary and economical "mining" and use of fuels, whereas new reserves and more efficient technologies are being discovered and make new fuel reserves economical. Currently, a substantial amount of the world's electricity is obtained from nuclear and hydro energy ($\sim$17 and 18%, respectively), use of other renewable energy resources is increasing (e.g., geothermal, wind, biomass, solar), and alternative synthetic fuels (e.g., hydrogen fuel cells) are being developed. It is worth

noting that some countries produce most or nearly all of their electricity from hydro energy (e.g., Norway, Brazil, New Zealand, Austria, Switzerland), whereas France produces most of its electricity from nuclear fuel (76%). The nuclear fuel reserves are orders of magnitude higher than fossil fuel reserves, and nuclear fuel does not contribute to CO_2 and greenhouse gas pollution.

Furthermore, advances in energy conversion and utilization technologies as well as increases in efficiency, including computerized control and management, contribute to energy conservation, increases in safety, and reduction of related environmental pollution. Actually, per capita energy use in the United States and other developed countries has decreased during recent years. However, the increase in world population and the development of many underdeveloped and highly populated countries, such as China and India, will continue to increase the world's energy consumption.

As an ultimate energy source for virtually all natural processes, solar energy is available for direct "harvest" if needed and is absorbed by vegetation and water surfaces on Earth. Thus, it is the driving force for natural photosynthesis, and in turn for biosynthesis processes as well as natural water cycle and all atmospheric processes (see the solar-related renewable energy sources in Table III). The solar radiation power density incident to Earth's atmosphere, known as the solar constant, is 2 cal/min/$(cm)^2$ or 1.4 kW/m^2, which after taking into account average day/night time (50%), varying incident angle (50%), and atmospheric/cloud scatter and absorption (53%), reduces to only $0.5 \cdot 0.5 \cdot 0.47 = 11.7\%$ of the solar constant, or approximately 165 W/m^2 at the Earth's surface, as an all-time average. This solar power per square meter corresponds to human metabolic energy consumption of 3400 kcal/day. The dietary energy reference value of 2000 kcal/day is equivalent to 97 W or 331 Btu/h. Furthermore, the world's total energy consumption is approximately 7500 Btu/h per capita or 2.2 kW per capita (or 11.8 kW per capita in the United States). (The total energy rate in kilowatts needs to be scaled by 31% efficiency to be compared with the electrical energy rate in kilowatts.) The corresponding per capita electricity consumption rates are 0.275 and 1.48 kW in the world and in the United States, respectively.

Therefore, the outlook for future energy needs is encouraging. There are many diverse and abundant energy sources with promising future potentials, so that humankind should be able to enhance its activities, standard of living, and quality of living by diversifying energy sources and by improving energy conversion and utilization efficiencies while at the same time increasing safety and reducing environmental pollution.

SEE ALSO THE FOLLOWING ARTICLES

Complex Systems and Energy • Conservation of Energy Concept, History of • Energy in the History and Philosophy of Science • Forms and Measurement of Energy • Heat Transfer • Physics and Economics of Energy, Conceptual Overview • Thermodynamic Sciences, History of

Further Reading

Cengel, Y. A., and Boles, M. A. (2002). "Thermodynamics: An Engineering Approach." 4th ed. WCB McGraw–Hill, Boston.

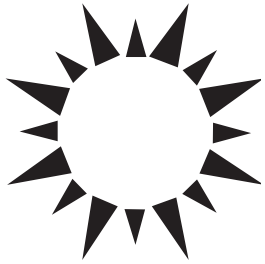
Energy Information Administration. (2003). "Annual Energy Review 2001." U.S. Department of Energy, Washington, DC. www.eia.doe.gov/aer.

Ferris, T. (ed.). (1991). "Physics, Astronomy, and Mathematics." Little, Brown, Boston.

Kostic, M. (2003). "The Quest and Nature of Energy and Entropy, Reversibility, and Irreversibility." Unpublished manuscript, Northern Illinois University.

Zemansky, M. W., and Dittman, R. H. (1982). "Heat and Thermodynamics." McGraw–Hill, New York.

World Environment Summits: The Role of Energy



ADIL NAJAM
Tufts University
Boston, Massachusetts, United States
CUTLER J. CLEVELAND
Boston University
Boston, Massachusetts, United States

1. **Energy and Sustainable Development: Conceptual Connections**
2. **The United Nations Conference on the Human Environment**
3. **The United Nations Conference on Environment and Development**
4. **The World Summit on Sustainable Development (WSSD)**
5. **An Evolving Agenda**
6. **Conclusions**

Glossary

human development The process of enlarging people's choices by expanding human capabilities and functionings. The three essential capabilities for human development are for people to lead long and healthy lives, to be knowledgeable, and to have a decent standard of living.

Millennium Development Goals Proposed by the United Nations Millennium Declaration of 2002. The declaration, adopted by 147 heads of state and 189 states of the United Nations, mainstreams eight mutually reinforcing development goals and 18 related targets into the global development agenda.

sustainable development Development that meets the needs of the present without compromising the ability of future generations to meet their own needs.

Type II agreements Partnerships between a combination of governments, nongovernmental organizations (NGOs), and the private sector for the implementation of sustainable development.

United Nations Conference on Environment and Development (UNCED) Also know as the Earth Summit, this was held at Rio de Janeiro, Brazil, from June 2 to 14, 1992. By far this was this largest and most prominent global environmental summit.

United Nations Conference on the Human Environment (UNCHE) Held in Stockholm, Sweden, from June 5 to 16, 1972, this was the first meeting to seek global policy consensus on issues related to the environment.

World Summit on Sustainable Development (WSSD) Held in Johannesburg, South Africa, from August 26 to September 4, 2002, this was the third major global environmental summit sponsored by the United Nations.

Energy is related to sustainable development as a source of environmental stress, as a principal motor of macroeconomic growth, and as a prerequisite for meeting basic human needs. These three dimensions correspond to the three dimensions of the often-used triangle of sustainable development: environmental, economic, and social factors. Using this framework, this article traces how successive environmental summits at Stockholm, Rio de Janeiro, and Johannesburg have dealt with energy issues. It identifies a slow, surprising, and important evolution of how energy issues have been treated at these global discussions. Energy has received increasing prominence at these meetings and it has become more firmly rooted in the framework of sustainable development. Stockholm was primarily concerned with the environmental dimension, Rio de Janeiro focused on both the environmental and economic dimensions, and the major headway made in Johannesburg was the meaningful addition of the social dimension and the linking of energy issues to the United Nations' Millennium Development Goals.

 © 2004 Elsevier Inc. All rights reserved.

1. ENERGY AND SUSTAINABLE DEVELOPMENT: CONCEPTUAL CONNECTIONS

The efficient use of energy and supplies that are reliable, affordable, and less-polluting are widely acknowledged as indispensable components of sustainable development. Although perennial debates linger about precise definitions of sustainable development, there is increasing agreement among scholars and practitioners that sustainable development policy relates to three critical elements that need to be treated together: economic, social, and environmental factors. Energy is central to any discussion of sustainable development because it is central to all three dimensions. In terms of the economic dimension of sustainable development, energy is clearly an important motor of macroeconomic growth. In terms of the environmental dimension, conventional energy sources are major sources of environmental stress at global as well as local levels. In terms of the social dimension, energy is a prerequisite for the fulfillment of many basic human needs and services, and inequities in energy provision and quality often manifest themselves as issues of social justice. Figure 1 presents the now-familiar triangular diagram depicting the three essential elements of sustainable development, modified to show how the energy dimension maps on to each of these elements. The remainder of this section elaborates upon the conceptual linkages between energy and each of the three dimensions of sustainable development.

Having briefly defined a framework for examining the conceptual relationship between energy and sustainable development, let us now examine how the issue of energy has been dealt with at the three major global environmental conferences held at Stockholm (1972), Rio de Janeiro (1992), and Johannesburg (2002). While these are not the only events of importance to have taken place in this period, they are the three most important policy conferences that have been attended by government representatives at the highest level and that have focused on a wide range of issues related to what has come to be known as sustainable development. For this reason, we have not focused on a host of energy-specific conferences that have been held over this period or specific policy negotiations on energy-related issues such as global climate change. Our goal is not as much to focus on the development of energy policy but to concentrate on the development of sustainable development policy to highlight how energy issues have featured within these discussions. Our purpose is to chart the evolution of the policy response and, in particular, to examine whether a perceptible trend of moving from a purely environmental focus to a broader sustainable development focus is evident in how these summits have dealt with energy issues. We will do so by examining the key documents negotiated at each of the three summits in terms of the three dimensions of energy-sustainable development linkages as defined here.

2. THE UNITED NATIONS CONFERENCE ON THE HUMAN ENVIRONMENT

The United Nations Conference on the Human Environment (UNCHE), held in Stockholm, Sweden,

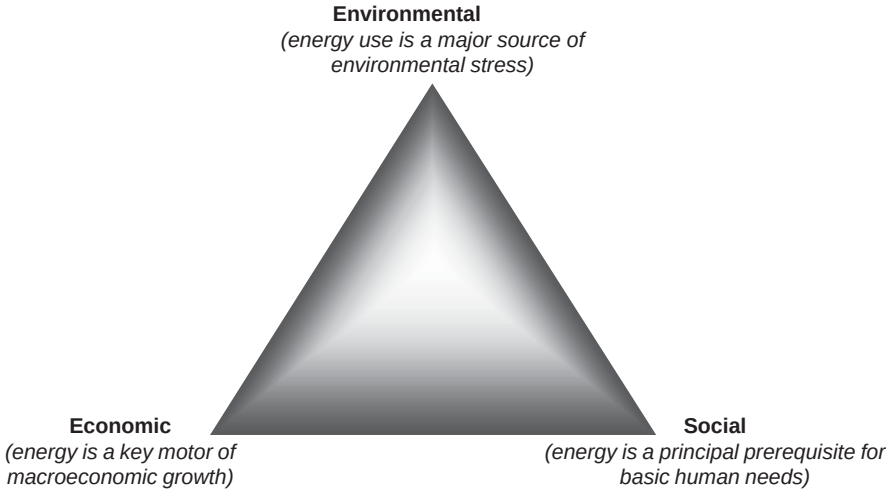


FIGURE 1 Energy and sustainable development: Deep linkages.

from June 5 to 16, 1972, was a first in many respects: it was the first meeting that brought the nations of the world (113 countries participated) together to discuss the environmental future of the planet; it was the first United Nations conference on a single global issue; it was the first global meeting that saw a large presence and influence of nonstate actors, including nongovernmental organizations and scholars; and it was the first meeting to seek global policy consensus on issues related to the environment.

Triggered by increasing scientific evidence of human-induced environmental degradation and a concurrent wave of growing environmental awareness in the industrialized nations of North America and Western Europe, the conference was an attempt to turn the environment into a more global issue, particularly by more meaningfully incorporating the developing countries of the South into the emerging global environmental discourse. The conference turned out to be unexpectedly contentious—with most Soviet bloc countries boycotting it due to the exclusion of then East Germany and with developing countries apprehensive of the North's newfound environmental concern. However, despite the intense North-South differences, and possibly because of them, the conference stumbled toward a more authentic global agenda: one that sought to merge the North's growing concern about environmental quality and the South's long-held interest in human development. Eventually, but much later, the desire to formulate the two interests into a single composite framework would lead to the concept of sustainable development.

The major institutional legacy of the conference was the creation of the United Nations Environmental Programme (UNEP). This was accompanied by two declaratory documents—The Stockholm Declaration and the Stockholm Action Plan—ideas from which have been carried forth by subsequent summits, including at Rio de Janeiro and Johannesburg. In addition to these, a few more ritualistic declarations were also adopted: one calling for a second UN Conference on the Human Environment (which was never actually held), another establishing an annual World Environment Day (which is now observed in most countries each June), and one calling for a stop to nuclear testing, particularly atmospheric nuclear testing (in fact, this was the single most hotly debated issue at the conference and inspired by the global politics of the time).

Surprisingly little was said or discussed at the conference or in any of its formal products regarding issues related to energy (see full text of Stockholm documents at www.unep.org). The conference declaration, which was the major political document emerging from Stockholm, has no direct reference to the energy issue. Of the 26 principles laid out in the declaration, the one that can be construed to be of most relevance to energy issues is Principle 5, which states: "The non-renewable resources of the earth must be employed in such a way as to guard against the danger of their future exhaustion and to ensure that benefits from such employment are shared by all mankind." A little more than one year later, energy was thrust on to the international agenda when the oil crisis nearly doubled real energy prices and plunged many national economies into recession.

The Stockholm Action Plan on the Human Environment is a more comprehensive document that includes a Framework for International Action accompanied by a list of 69 specific recommendations, three of which do, in fact, deal with energy (emphasis added):

- Recommendation 57 called upon the UN Secretary-General to "take steps to ensure proper collection, measurement and analysis of data relating to the *environmental effects* of energy use and production."
- Recommendation 58 called for better exchange of information on energy. The recommendation is motivated by the need for "the rationalization and integration of resource management for energy" and seeks mechanisms (such as exchange of national experiences, studies, seminars, meetings, and a "continually updated register of research") for accessing existing information and data, particularly on "the *environmental consequences* of different energy systems."
- Recommendation 59 called for a "comprehensive study to be promptly undertaken with the aim of submitting a first report, at the latest in 1975, on available energy sources, new technology, and consumption trends, in order to assist in providing a basis for the most effective development of the world's energy resources, with due regard to the *environmental effects* of energy production and use."

Even at Stockholm itself, none of these recommendations was particularly inspiring in its scope or aspiration. They are even less so with the hindsight we now enjoy. Indeed, between them they call merely for better data collection and analysis, and the mechanisms envisaged for doing so are not particularly innovative or exciting. More important, even the minimal level of action that the Action Plan envisaged on the energy issue never really materialized.

For example, the official documents make no mention of Stockholm's implementation of the comprehensive study that was sought by 1975 (Recommendation 59). The one issue related to energy that did gain wide political and policy prominence at Stockholm was atmospheric nuclear testing. This was a subject of great and heated debate at the conference, became the subject of a separate resolution, and made the International Atomic Energy Agency a frequently mentioned organization within the Stockholm Action Plan. This discussion was very much an artifact of the Cold War politics of the time and was at its root far more concerned with nuclear weapons than with energy and its role in the economy and as an agent of environmental change.

The lack of imagination or urgency on the energy issue should not be entirely surprising. Held in mid-1972, the Stockholm conference came before the great oil shock of the 1970s. State delegates attending Stockholm still lived in a relatively calm world of declining real oil prices, and the possibility of spiraling energy prices or oil scarcity could not have been high on their mental maps. Similarly, global climate concerns had not yet taken root in 1972 and environmentalists were more focused on the pollution outputs of industrialization than the energy inputs for economic production. While those attending UNCHE were well aware of the many environmental implications of energy issues, these were not their most pressing priorities at that time.

What is clear from the recommendations, however, is that to the extent that energy was considered an issue of any importance, that importance derived directly from the environmental effects of energy extraction, processing, and consumption. While the recommendations slightly hint at the importance of energy as a motor of economic growth, their principal preoccupation is with the potential for environmental stress from the chain of energy supply and use. The dimension of energy as a prerequisite for meeting basic human needs does not figure into the equation at this juncture.

3. THE UNITED NATIONS CONFERENCE ON ENVIRONMENT AND DEVELOPMENT

The United Nations Conference on Environment and Development (UNCED), popularly know as the Earth Summit, was the crowning moment not only of environmental summitry but of UN mega-summitry as a genre. Held at Rio de Janeiro, Brazil from June 2 to 14, 1992, it brought together more than 100 heads of state and government, 150 nations, more than 1400 nongovernmental organizations (NGOs), 8000 journalists, and nearly 35,000 participants. More than that, it caught the public imagination like no conference before or since. Some consider the summit to have failed its potential, if not its mandate in terms of the content of its substantive agreements. But most commentators and experts consider this summit to have been a success in terms of elevating the global profile of environmental issues and raising awareness regarding sustainable development.

Held to mark the 20th anniversary of the Stockholm conference, the Rio Earth Summit became everything that an earlier "Stockholm plus ten" conference, held in Nairobi in 1982, could not. Indeed, it became more than even its proponents had hoped for. Instead of being the second United Nations Conference on the Human Environment, Rio was the United Nations Conference on Environment *and* Development; putting those two terms together, which had been so much at odds at Stockholm, might itself have been Rio's most important achievement. In particular, it broadened the scope of global environmental diplomacy by adopting the notion of sustainable development, which had been advocated 5 years earlier in by the World Commission on Environment and Development as one of its key policy frameworks.

The world at Rio was, of course, very different from the world at Stockholm. In the intervening two decades, the Cold War (which was the defining political framework at UNCHE) had disappeared, the level of public interest in the environment was greatly increased, environmental issues such as stratospheric ozone depletion and global climate change were now squarely on the global policy map, and energy had become a major concern for economic security in the aftermath of the oil price shocks of 1973–1974 and 1980–1981.

The products coming out of UNCED included a political Declaration enunciating 27 principles of environment and development, a 700-page action program called Agenda 21, a nonbinding set of principles for sustainable forest management, and specific conventions on climate change and biodiversity. The institutional innovation resulting from the conference included an agreement on the operating rules for the Global Environmental Facility (GEF) and the establishment of a Commission on Sustainable Development (CSD) on the basis of an Agenda

21 recommendation. Technically, the official products of UNCED were only the Rio Declaration on Environment and Development, the Authoritative Statement of Forest Principles, and Agenda 21, all of which were adopted by consensus (without vote) by the conference. The conventions on climate change and biodiversity were products of independent, but concurrent, negotiating processes that were opened for signatures at UNCED.

The Rio Declaration on Environment and Development, like its predecessor at Stockholm, had nothing specific to say about energy. Indeed, the clause about the depletion of natural resources contained within the Stockholm Declaration was dropped. While a number of the principles articulated in the Declaration could be construed to have bearing on energy, none deals with the issue directly.

Although the climate and biodiversity conventions were not direct products of the Rio process, the former is of direct relevance to the energy issue. Indeed, the United Nations Framework Convention on Climate Change (UNFCCC) is the nearest thing we have to a global convention dealing directly with energy concerns. Since energy production and consumption is the biggest source of anthropogenic greenhouse emissions, climate policy in the UNFCCC, and subsequently in the 1997 Kyoto protocol, has been discussed mostly through the lens of energy policy in a wide variety of ways. Two examples, among many, of how climate policy becomes energy policy are the intense policy debates about the variable ability and responsibilities of different nations to change their energy consumption and production patterns and the role of energy taxation as a means of emissions control.

These debates during and after Rio have been defined principally by compulsions that lie at two distinct corners of the sustainable development triangle: the environmental compulsion emanating from the ecological stresses associated with specific energy production and consumption choices, and the economic compulsion derived from the central role of energy in economic growth. The saliency that the climate issue had assumed by 1992 meant that the discussion on energy at Rio was not only more intense than it had been at Stockholm but also broader. Whereas UNCHE had been principally concerned with the role of energy as a source of environmental stress, Rio's energy concerns related to both economic and environmental dimensions.

This evolution of the energy focus was quite evident in Agenda 21, the most comprehensive of the Rio documents. Interesting, Agenda 21 (which has a total of 40 chapters) does not have a chapter on energy. However, Chapter 9 of Agenda 21, which deals with protecting the atmosphere, serves as a *de facto* energy chapter since it focused on global climate change and related issues of fossil fuel use. In addition, the chapters on changing consumption patterns (Chapter 4), promoting sustainable human settlements (Chapter 7), and promoting sustainable agriculture and rural development (Chapter 14) also have significant discussions on the energy issue. The vast bulk of this discussion is contextualized in the need to balance the environmental and economic nodes of the sustainable development triangle. What is carried over from Stockholm is a clear emphasis on the environmental impacts of energy production and use (especially in terms of global climate change). New additions are the prescriptions contained in these various chapters—or the energy message of Agenda 21—which fall into familiar categories: decrease energy consumption (see Agenda 21 sections 4.24, 7.5), increase energy efficiency (see Agenda 21 sections 4.18, 7.49), and develop cleaner sources of energy (see Agenda 21 sections 9.12, 9.18).

The third dimension of the sustainable development triangle dealing with social concerns such as the role of energy as a human need do not figure as prominently in Agenda 21 as the environmental and economic dimensions. However, glimpses of such concerns do occasionally surface in the document. For example, the chapter on human settlements (Chapter 7) mentions energy as a human need at par with other needs such as water (see Agenda 21 sections 7.27, 7.40). Section 9.9 goes the furthest by defining energy as an essential component of economic as well as social development and as a prerequisite for an improved quality of life. Although these references are quite general and made in passing, with very little prescriptive policy content, they do signify an important evolution from the Stockholm texts where these issues were conspicuous only by their absence.

Overall, then, one finds that UNCED did treat the energy issue very differently from UNCHE. At Stockholm, energy was discussed in the most general, even cursory, fashion, and only in terms of its environmental impacts. In the years between Stockholm and Rio, concerns about the environmental stress imposed by energy production and use became more precise with the mounting evidence of global climate change. As a result, Rio was relatively more precise and prescriptive in terms of energy policy in that it went beyond simply calling for more

information collection and dissemination to highlighting the need for decreasing consumption, increasing efficiency and transitioning to cleaner sources. In doing so, Rio broadened the focus from merely environmental concerns to the balance between environmental and economic concerns. However, the third node of the sustainable development triangle, the social dimension signified by energy as a human need, still remained in the shadows and peeped through the Rio documents only infrequently and rather unimpressively.

4. THE WORLD SUMMIT ON SUSTAINABLE DEVELOPMENT (WSSD)

Few expected the 2002 World Summit on Sustainable Development (WSSD) to be as impressive as UNCED. Held to mark the 10th anniversary of the Rio Earth Summit and to take stock of progress on Agenda 21 in those 10 years, the run-up to Johannesburg was singularly dismal and uninspiring. The world had, once again, changed. The high hopes of a new era of global environmental cooperation that had been ignited by Rio soon proved false. The industrialized countries of the North had remained unwilling to provide the developing countries of the South with the resources or support that had been implied at Rio, meanwhile the promise of a post-Rio harvest of global environmental treaties and implementation proved unfounded as key states, particularly but not solely the United States, dragged their feet on key issues such as climate change. As a result, a malaise had set in well in advance of WSSD, which was only made worse by events at the geopolitical level, where the global mood had gone sour after the tragic terrorist attack on the United States and a growing sense of insecurity and violence around the world. Held in Johannesburg, South Africa, from August 26 to September 4, 2002, WSSD was different from both Stockholm and Rio in that it was not born within the optimism and high hopes that had accompanied earlier summits.

In terms of sustainable development, WSSD had the distinction of actually having those two magic words in its very title. However little had been achieved in the 10 years since Rio on other counts, Johannesburg was testimony to the fact that the term sustainable development had gained policy acceptance. Even though some argued that the term had lost its edge and was mostly being used rhetorically, the fact remained that it had also become a political necessity. For those who believed in the concept, this was a chance to put meaning into it. At best, Johannesburg was viewed as a chance to advance the agenda that had been set by Rio; at the very least, it was an attempt to keep the Rio agenda alive.

It became clear fairly early on in the Johannesburg process that WSSD would not be able to match the ambition or scope of UNCED, certainly not in terms of its products. Like Stockholm and Rio before it, the Johannesburg summit also sought a political declaration as its principal output. In addition, it also sough a plan of implementation; one that was much less ambitious in scope or scale than Agenda 21 but more extensive than the Stockholm plan of action. The major innovation at Johannesburg were the so-called Type 2 agreements. These were informal agreements involving nonstate parties, sometimes among themselves and sometimes with individual governments. On the one hand, Type 2 agreements were a reflection of the massive change in landscape that had occurred over the previous 10 years, with NGOs and business taking a far more important role in international environmental affairs. At the same time, however, they were a reflection of the WSSD organizer's desperation and desire to get something memorable out of the summit. According to the rough count by the summit organizers, more than 220 Type 2 agreements were reached at Johannesburg, signifying around US$235 million in pledged resources; 32 of these Type 2 agreements relate to energy, accounting for US$26 million in resources; the vast majority of these are programs of technical cooperation in energy generation and conservation (www.johannesburgsummit.org). It should be noted that a systematic accounting of these agreements has not yet been accomplished, and it is not yet clear how many of these agreements and how much of these resources are, in fact, new and unique. The other innovation at Johannesburg, in comparison to previous summits, relates to the fact that the Johannesburg plan of Implementation sought agreement on actual targets and timetables rather than simple statements of intent. While it is true that in many cases (including renewable energy) such targets and timetables were not forthcoming and in others they were merely restatements of targets that had already been set (such as in access to clean water), it is also true that in a few areas (such as sanitation) meaningful headway was made in terms of reaching agreement on targets and timetables where there had previously been none.

Looking carefully at the various products it does seem that amidst the many disappointments of Johannesburg, energy might be one of these few areas where progress was made. First, and quite strikingly, the Johannesburg declaration is different from its predecessors from Stockholm and Rio in that it actually does have a direct reference to energy. More important, the declaration (clause 18) clearly identifies energy as a human need at a par with needs such as clean water, sanitation, shelter, health care, food security, and biodiversity. Although this is a declaratory clause with no enforceability, it does signify a demonstrable shift from the previous summits. It clearly defines energy as a basic human need, thereby evoking the social dimension of the sustainable development triangle more clearly than either Stockholm or Rio had done.

The Johannesburg plan of implementation breaks similar new ground in terms of how it deals with the energy issue. While the focus on the environmental stress-economic growth axis that had emerged at Rio is not at all lost, the Johannesburg plan is strikingly different from its predecessors in two distinct and important ways. First, it clearly deals with energy as an issue in its own distinct right rather than as a facet of other issues. Second, it firmly adds the social dimension to the existing environmental and economic dimensions to begin dealing with the entirety of the sustainable development triangle for the very first time.

The concerns about energy in terms of environmental stress and economic growth show up similarly to how they had surfaced at Rio. The principal arena for these concerns remains climate change and the need for a balanced approach is once again reiterated, as are the preferred mechanisms for achieving such balance: decrease energy consumption, increase energy efficiency, and transition to cleaner energy systems. This discussion is most clear in article 19 of the Johannesburg plan whose 23 subclauses relate to various environmental and economic aspects of energy in relation to sustainable development. The environmental dimension of sustainable development is most clearly and persistently manifest in the many references to the need for enhanced energy efficiency, in particular the call for establishing domestic programs for energy efficiency (subclause h), the need to accelerate the development and dissemination of energy efficiency and energy conservation technologies (subclause i), and the call to promote and invest in research and development of such technologies (subclause k). The economic dimension is also strong in article 19, parts of which call for removing market distortions including the restructuring of taxes and the phasing out of harmful subsidies (subclause p) and the call to support efforts to improve the functioning, transparency, and information about energy markets with respect to both supply and demand (subclause o). One area in which Johannesburg tried, but failed, to make new headway within the environmental-economic axis relates to renewable energy. WSSD saw heated debates about setting up quantifiable targets and timetables for renewable energy use. These discussions eventually failed to yield actual timetables and targets (principally because of U.S. opposition to them), but they did succeed in introducing more detailed language regarding energy issues than had been present in Agenda 21. The subclause (e) of article 19 is, therefore, crafted in general language and calls for diversifying energy supply and substantially increasing the global share of renewable energy sources.

What is new and quite intense in the Johannesburg plan on implementation are the repeated references to energy and sustainable development. Here, the document goes beyond Agenda 21 by focusing more on the social dimension of energy and sustainable development and by concentrating on the role energy plays as a prerequisite for basic human needs including those defined in the United Nations' Millennium Development Goals (MDGs). The most significant of these references is made in article 8 of the plan of Implementation, which falls within the section on poverty eradication. New ground is covered here when the Johannesburg plan clearly and unambiguously calls for

> *access to reliable and affordable energy services for sustainable development sufficient to facilitate the achievement of the millennium development goals, including the goal of halving the proportion of people in poverty by 2015, and as a means to generate other important services that mitigate poverty, bearing in mind that access to energy facilitates the eradication of poverty.*

This is not only interesting but groundbreaking, because the original Millennium Development Goals—a set of quantitative targets proposed by the UN Secretary General to the 2000 Millennium Summit as the foundation of the United Nations' work program from the next decades and as a framework for measuring development progress (see www.developmentgoals.org)—do not mention energy at all. The case for energy as a prerequisite for basic human needs is made, instead, by the Johannesburg plan of implementation. Indeed, it is made

rather convincingly. Subclause (a) of article 8, for example, calls for improving

> *access to reliable, affordable, economically viable, socially acceptable and environmentally sound energy services and resources, taking into account national specificities and circumstances, through various means, such as enhanced rural electrification and decentralized energy systems, increased use of renewables, cleaner liquid and gaseous fuels and enhanced energy efficiency, by intensifying regional and international cooperation in support of national efforts, including through capacity-building, financial and technological assistance and innovative financing mechanisms, including at the micro and meso levels, recognizing the specific factors for providing access to the poor.*

Subclause (g) of the same article, elaborates the case further by calling for

> *access of the poor to reliable, affordable, economically viable, socially acceptable and environmentally sound energy services, taking into account the instrumental role of developing national policies on energy for sustainable development, bearing in mind that in developing countries sharp increases in energy services are required to improve the standards of living of their populations and that energy services have positive impacts on poverty eradication and improve standards of living.*

Here, then, is an example of all three dimensions of the sustainable development triangle being invoked together and in a way that was not seen in any of the Stockholm or Rio documents. Although this is still preliminary, it is nonetheless a novel and welcome attempt to deal with energy fully in the context of sustainable development by seeking policy that responds to the environmental, economic as well as the social impulses of the concept. Policy made under the guidelines set here would need to be evaluated not only in terms of the environmental stresses being imposed, or effects on economic growth, but also in terms of social justice and especially in terms of how it affects the poorest and the most vulnerable.

5. AN EVOLVING AGENDA

The process of evolution that seems to have taken place between the 1972, 1992, and 2002 environmental summits is depicted in Fig. 2, which builds on the three dimensions of sustainable development to illustrate how the documents emerging from each successive conference have dealt with energy issues. Figure 2 suggests that a rather neat evolution of the agenda has happened with the Stockholm summit of 1972 dealing with energy issues principally as a source of environmental stress, the 1992 Rio summit added a clear economic focus to its treatment of the subject, while the 2002 Johannesburg conference built on the existing environmental and economic focus and added the element of energy as a requisite for basic human needs to the equation for the first time. Of course, the fact that these summits have dealt with energy in a particular way does not imply that global energy policy has moved in that direction automatically. The purpose of such summits is principally declaratory. However, the value of these declaratory proclamations must not be underestimated. They serve not only to advance the conceptual agenda but also tend to eventually influence the actual policies, although usually with some time lag.

In terms of general conclusions, a number of surprises can be highlighted:

1. Although conventional wisdom maintains that Johannesburg is a pale comparison to conferences before it, on the specific issue of energy it has actually

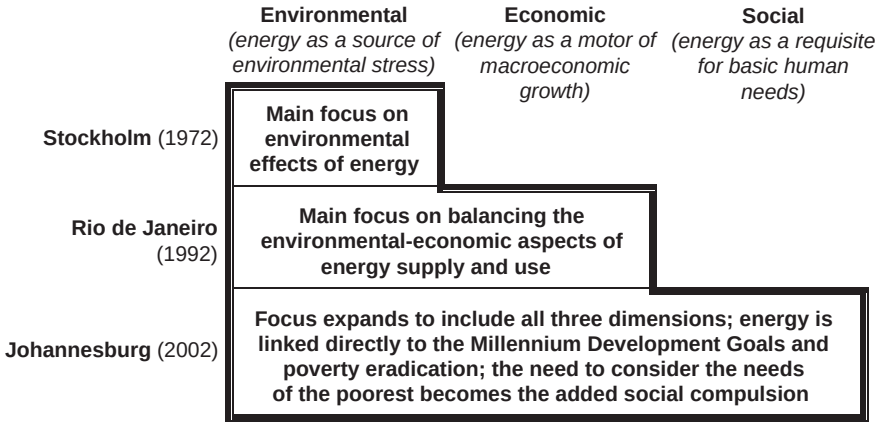


FIGURE 2 Energy and sustainable development: An evolving agenda.

made major conceptual headway by incorporating the energy issue more fully into a sustainable development framework. While the conference as a whole may not have been inspiring, on this one issue it has traversed into new and important territory.

2. It is rather surprising that the UN Millennium Development Goals fail to identify energy as a key human development and human needs theme. However, the Johannesburg plan of implementation has partially filled that gap and made the argument that would have better come from the Millennium Development Goals. That energy is a prerequisite for many human needs and for poverty alleviation merits explicit inclusion in the Millennium Development Goals. Because of Johannesburg, this notion now has the endorsement of the comity of nations, and there seems to be a strong case for incorporating energy issues explicitly within MDG programs.

3. There is a need for the practice of energy policy as well and the scholarship on energy policy to catch up with the realization imbedded in the Johannesburg plan on implementation regarding the importance of energy as a prerequisite for basic human needs.

4. A broader argument can now be made that megaconferences can advance global agendas in significant ways, as they have done with energy. The energy gives cause for more optimism than many scholars invest in these conferences. However, a corollary to be investigated further would be whether such impact is more likely to be noticeable on specific issues rather than on the general agenda as a whole.

6. CONCLUSIONS

In focusing on the actual text of the documents negotiated at various global environmental summits and using a framework of how energy policy relates to various aspects or dimensions of sustainable development, this article finds that there has been a slow but demonstrable evolution in how these conferences have dealt with energy. Not only has energy assumed a successively more prominent role in these global summits, but a noticeable evolution has occurred in the dimensions of energy policy that have been addressed. Moreover, this evolution has been along the trajectory of sustainable development. Although energy policy per se might not have made this transition as yet, these summits have given a clear signal to national and international policymakers to align energy policy more firmly to sustainable development, and to do so in more intricate ways.

In terms of evaluating the impact of the Johannesburg Earth Summit on the energy and sustainable development agenda, it seems that the one summit that had started with the worst prospects might well have achieved the most important advance in terms of conceptualizing energy policy within a framework rooted in all dimensions of sustainable development. Both the Johannesburg declaration and the Johannesburg plan of implementation are remarkable in that they highlight the human need aspect of energy, in addition to the environmental and economic dimensions that had already been incorporated at Stockholm and Rio.

SEE ALSO THE FOLLOWING ARTICLES

European Union Energy Policy • Geopolitics of Energy • Green Accounting and Energy • International Energy Law and Policy • National Energy Policy: Brazil • National Energy Policy: China • National Energy Policy: India • National Energy Policy: Japan • National Energy Policy: United States • National Security and Energy • United Nations Energy Agreements

Further Reading

Banuri, T., and Najam, A. (2002). "Civic Entrepreneurship: A Civil Society Perspective on Sustainable Development." Gandhara Academy Press, Islamabad.

Cleveland, C. J., Costanza, R., Hall, C. A. S., and Kaufmann, R. (1984). Energy and the U.S. economy: a biophysical perspective. *Science* **255**, 890–897.

Hass, P. M., Levy, M. A., and Parson, E. A. (1992). Appraising the Earth Summit: How should we judge UNCED's success? *Environment* **34**(8), 12–36.

Johnson, S. P. (1993). "The Earth Summit: The United Nations Conference on Environment and Development (UNCED)." Graham & Trotman/Martinus Nijhoff, London.

Munasinghe, M., (1992). "Environmental Economics and Sustainable Development" (originally presented at the United Nations Conference on Environment and Development, Rio de Janeiro, Brazil). World Bank, Washington, DC.

Najam, A. (1995). An environmental negotiation strategy for the south. *International Environmental Affairs* **7**(3), 249–287.

Najam, A., Poling, J. M., Yamagishi, N., Straub, D. G., Sarno, J., DeRitter, S. M., and Kim, E. M. (2002). From rio to Johannesburg: Progress and Prospects. *Environment* **44**(7), 26–38.

Reddy, A. K. N. (2002). Energy technologies and policies for rural development. *In* "Energy for Sustainable Development: A Policy Agenda" (T. B. Johansson and J. Goldemberg, Eds.), pp. 115–136. United Nations Development Program, New York.

Suarez, C. E. (1995). Energy needs for sustainable development: energy as an instrument for socio-economic development. *In* "Energy for Sustainable Development: A Policy Agenda" (T. B. Johansson and J. Goldemberg, Eds.), pp. 18–27. United Nations Development Program, New York.

Susskind, L. E. (1994). "Environmental Diplomacy: Negotiating more effective global agreements." Oxford University Press, New York.

United Nations Development Program (UNDP) (2003). "Human Development Report 2003." United Nations Development Program and Oxford University Press, New York.

World Commission on Environment and Development (WCED) (1987). "Our Common Future: Report of the World Commission on Environment and Development." Oxford University Press, New York.

World History and Energy

VACLAV SMIL
University of Manitoba
Winnipeg, Manitoba, Canada

1. **A Deterministic View of History**
2. **The Earliest Energy Eras**
3. **Medieval and Early Modern Advances**
4. **Transitions to Modernity**
5. **High-Energy Civilization and Its Attributes**
6. **Limits of Energetic Determinism**

Glossary

antiquity The era of ancient (from the Western perspective, mostly the Middle Eastern and Mediterranean) civilizations extant between prehistory and the Middle Ages.

determinism A doctrine claiming that human actions are determined by external factors; its prominent varieties include environmental and geographic determinism.

early modern world The period immediately following the Middle Ages, variously dated as 1493–1800, 1550–1850, or the 16th to 18th centuries.

energy transition A period of passing from one configuration of prime movers and dominant fuels to a new setup.

middle ages The period between antiquity and the modern era, often circumscribed by the years 500–1500 CE.

prehistory The period of human evolution predating recorded history.

A strict thermodynamic perspective must see energy—its overall use, quality, intensity, and conversion efficiency—as the key factor in the history of the human species. Energy flows and conversions sustain and delimit the lives of all organisms and hence also of superorganisms such as societies and civilizations. No action—be it a better crop harvest that ends a famine or the defeat of an aggressive neighbor—can take place without harnessing and transforming energies through management, innovation, or daring. Inevitably, the availability and quality of particular prime movers and sources of heat and the modes of their conversions must have left deep imprints on history. But no energetic perspective can explain why complex entities such as cultures and civilizations arise and no thermodynamic interpretation can reveal the reasons for either their remarkable history or their astounding diversity of beliefs, habits, and attitudes from which their actions spring. This article examines both of these contrasting views of energy and world history.

1. A DETERMINISTIC VIEW OF HISTORY

Countless energy imperatives—ranging from the solar flux reaching the earth to minimum temperatures required for the functioning of thousands of enzymes—have always shaped life on Earth by controlling the environment and by setting the limits on the performance of organisms. Deterministic interpretations of energy's role in world history seems to be a natural proposition, with history seen as a quest for increased complexity made possible by mastering higher energy flows. Periodization of this quest on the basis of prevailing prime movers and dominant sources of heat is another obvious proposition. This approach divides the evolution of the human species into distinct energy eras and brings out the importance of energy transitions that usher in more powerful, and more flexible, prime movers and more efficient ways of energy conversion. Perhaps the most intriguing conclusion arising from this grand view of history is the shrinking duration of successive energy eras and the accelerating pace of grand energy transitions.

The first energy era started more than 300,000 years ago when the human species, *Homo sapiens*, became differentiated from *Homo erectus*, and the era continued until the beginning of settled societies some 10,000 years ago. Throughout prehistory, all efforts to control greater energy flows were capped

 © 2004 Elsevier Inc. All rights reserved.

by the limited power of human metabolism and by the inefficient use of fire. Domestication of draft animals and harnessing of fire for producing metals and other durable materials constituted the first great energy transition: reliance on these extrasomatic energies had raised energy throughput of preindustrial societies by more than an order of magnitude. The second transition got under way only several millennia later; it was not as universal as the first one and its effects made a profound, and relatively early, difference only in some places: it came as some traditional societies substituted large shares of their muscular exertions by waterwheels and windmills, simple but ingenious inanimate prime movers that were designed to convert the two common renewable energy flows with increasing power and efficiency.

The third great energy transition—substitution of animate prime movers by engines and of biomass energies by fossil fuels—began only several centuries ago in a few European countries and it was accomplished by all industrialized nations during the 20th century. That transition is yet to run its course in most low-income economies, particularly in Africa. The latest energy transition has been under way since 1882 when the world's first electricity-generating stations were commissioned in London and New York (both Edison's coal-fired plants) and in Appleton, Wisconsin (the first hydroelectric station). Since that time, all modernizing economies have been consuming increasing shares of their fossil fuels indirectly as electricity and introducing new modes of primary electricity generation—nuclear fission starting in the mid-1950s, and later also wind turbines and photovoltaic cells—to boost the overall output of this most flexible and most convenient form of energy. The second key attribute of this transition has been a steady relative retreat of coal mirrored by the rise of hydrocarbons, first crude oil and later natural gas.

Improving the quality of life has been the principal individual benefit of this quest for higher energy use that has brought increased food harvests, greater accumulation of personal possessions, abundance of educational and leisure opportunities, and vastly enhanced personal mobility. The growth of the world's population, the rising economic might of nations, the extension of empires and military capabilities, the expansion of world trade, and the globalization of human affairs have been the key collective consequences of the quest. These advances are discussed in this article and the limits of prime movers and heat sources that were available during the successive eras of energy use and the major accomplishments that were achieved through ingenuity and better organization are noted.

2. THE EARLIEST ENERGY ERAS

During the long span of prehistory, the human species relied only on its somatic energy, using muscles to secure a basic food supply and then to improve shelters and acquire meager material possessions. Organismic imperatives (above all, the basal metabolism scaling as the body mass raised to 0.75 power) and the mechanical efficiency of muscles (able to convert no more than 20–25% of ingested food to kinetic energy) governed these exertions: healthy adults of smaller statures cannot sustain useful work at rates of more than 50–90 W and can develop power of 10^2 W only during brief spells of concentrated exertion. The former performance sufficed for all but a few extreme forms of food gathering and the latter exertions were called on for some forms of hunting. Simple tools made some foraging and processing tasks more efficient and extended the reach of human muscles.

Energy returns in foraging (energy in food/energy spent in collecting and hunting) ranged from barely positive (particularly for some types of hunting) to seasonally fairly high (up to 40-fold for digging up tubers). The choice of the collected plants was determined above all by their accessibility, nutritional density, and palatability, with grasslands offering generally a better selection of such species than did dense forests. Collective hunting of large mammals brought the highest net energy returns (because of their high fat content) and it also contributed to the emergence of social complexity. Only a few coastal societies collecting and hunting marine species had sufficiently high and secure energy returns (due to seasonal migrations of fish or whales) such that they were able to live in permanent settlements and devote surplus energy to elaborate rituals and impressive artistic creations (for example, the tall ornate wooden totems of the Indian tribes of the Pacific Northwest).

The only extrasomatic energy conversion mastered by prehistoric societies was the use of fire for warmth and cooking, which can be indisputably dated to approximately 250,000 years ago. Eventual shifts from foraging to shifting cultivation and then to sedentary farming were gradual processes driven by a number of energy-related, nutritional, and social factors: there was no short and sharp agricultural revolution. These changes were accompanied by

declining net energy returns in food production, but these declines had a rewarding corollary as the higher investment of metabolic energy in clearing land and in planting, weeding, fertilizing, harvesting, and processing crops, as well as storing grains or tubers, made it possible to support much higher population densities. Whereas the most affluent coastal foraging societies had densities less than 1 person/km^2 (and most foraging societies had carrying capacities well below 0.1 person/km^2), shifting agricultures would easily support 20–30 people/km^2 and even the earliest extensive forms of settled farming (ancient Mesopotamia, Egypt and China's Huanghe Valley) could feed 100–200 people/km^2, that is, 1–2 people/ha of cultivated land (Fig. 1).

The increasing size of fields could not be managed by slow and laborious hoeing but plowing is either exceedingly taxing or, in heavy soils, outright impossible without draft animals. Farming intensification thus led to harnessing the first important extrasomatic source of mechanical energy by domesticating draft animals throughout the Old World (the pre-Colombian Americas had only pack animals). Continuous energy investment was then needed for animal breeding and feeding, as well as for producing more complex implements.

Small bovines would rate less than 200 W, stronger animals could sustain more than 300 W, and the best oxen and good early draft horses could surpass 500 W, equal to the labor of 6–8 adult men (Fig. 2). Draft animals thus speeded up field, transportation, and crop processing tasks and boosted agricultural productivity. Their numbers were governed by an obvious energetic imperative: no society could afford to cultivate feed crops where harvests were barely adequate to provide subsistence diets. Those agroecosystems where grazing land was also limited (the rice regions of Asia) could support only relatively small numbers of draft animals.

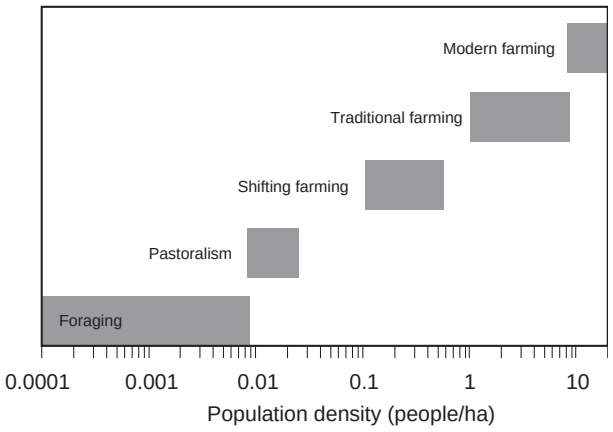


FIGURE 1 Successive stages of sedentary agricultures have required higher energy inputs but have been able to support 10^3 to 10^4 more people per unit of land than did foraging.

Limited unit power of muscles could be overcome by massing people, or draft animals, and the combination of tools and organized deployment of massed labor made it possible to build impressive structures solely with human labor. Massed forces of 20–100 adults could deliver sustained power of 1.5–8 kW and could support brief exertions of up to 100 kW, enough to transport and erect (with the help of simple devices) megaliths and to build impressive stone structures on all continents except Australia. In contrast to this massed deployment of human labor in construction, no Old High culture took steps to a truly large-scale manufacture of goods and the atomization of production remained the norm. In addition, violent conflict powered solely by an individual's muscles could take place only as hand-to-hand combat or by an attack with an arrow launched from less than a couple hundred meters away, a limit ordained by the maximum distance between one extended and one flexed arm when drawing a bow. Eventually catapults, tensioned by many hands, increased the mass of projectiles, but did not substantially lengthen the maximum distance of attack.

Shifting agriculturalists extended the use of fire to the regular removal of vegetation, and early settled societies also adopted fire to produce bricks and containers and to smelt metals, beginning with copper (before 4000 BCE) and progressing to iron (common in some parts of the Old World after 1400 BCE). Charcoaling was used to convert wood to a fuel of higher energy density (29 MJ/kg compared to no more than 20 MJ/kg for wood and less than 15 MJ/kg for crop residues) and superior quality (essentially smokeless and hence suitable for burning indoors in fixed or portable hearths). But open fireplaces and braziers converted less than 10% of the supplied energy into useful heat and traditional charcoaling turned less than one-fifth of the charged wood energy into the smokeless fuel. Primitive furnaces used for metal smelting were also very inefficient, requiring as much as 8–10 units of charcoal for a unit of pig iron. The resulting high demand for wood was a leading cause of extensive deforestation, but a nearly complete disappearance of forests from parts of the Mediterranean (Spain, Cyprus, and Syria) and the Near East (Iran, Afghanistan) was caused by smelting copper rather than iron.

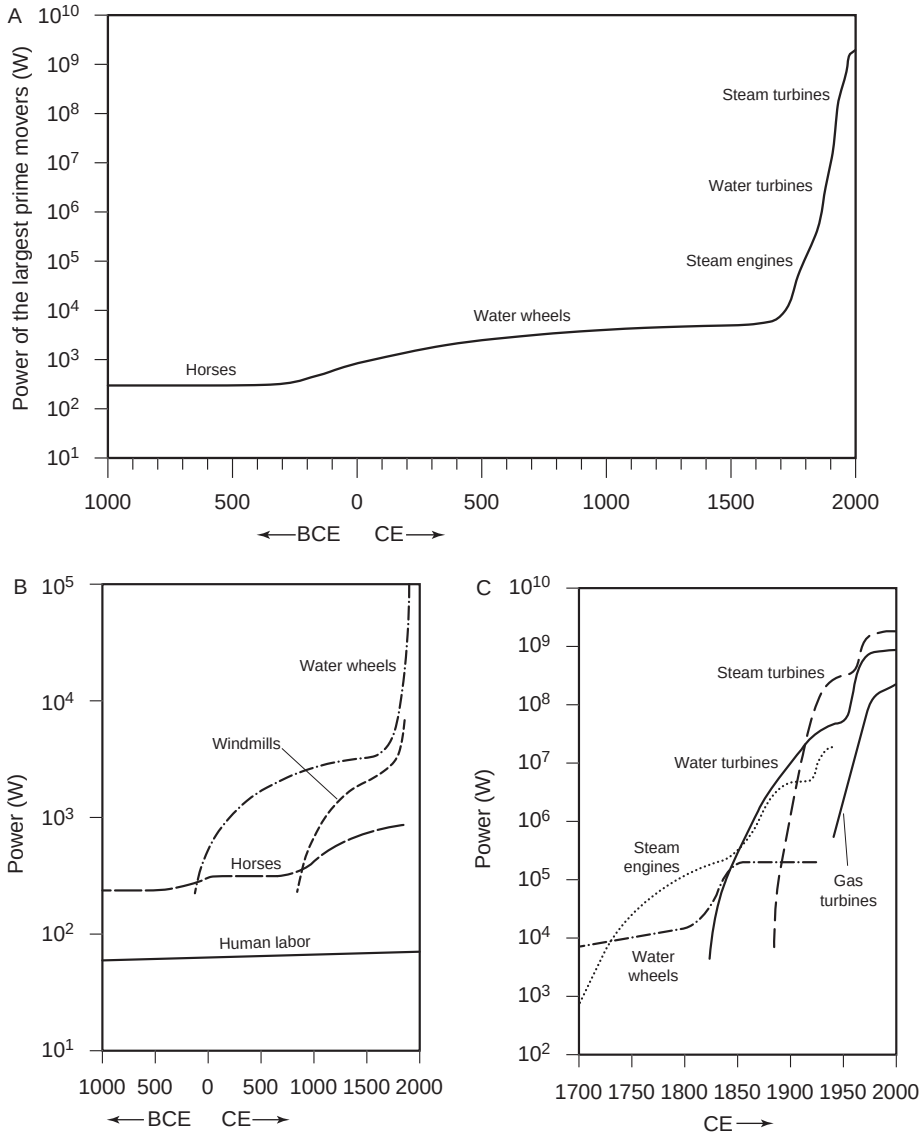


FIGURE 2 The maximum power of prime movers shown as the sequence of the highest capacity converters for the span of the past 3000 years (A) and shown in detail for the periods 1000 BCE to 1700 CE and 1700 CE to 2000 CE (B and C).

Small-scale and highly energy-intensive metallurgy meant that no early societies could smelt enough metal to make it the dominant material in daily use and simple machines, farming implements, and household utensils of the antiquity remained overwhelmingly wooden. This changed radically only when coke-based smelting began producing inexpensive iron after 1750. Similarly, the inherently limited power of oxen, the most affordable draft animals, which were rarely fed any concentrates, meant a ponderous pace of field operations: plowing a hectare of a loamy field with a pair of these animals was easily four times faster than hoeing the same land, but a pair of well-fed horses would have accomplished the same task in less than one-half the time required by oxen. And the combination of weak animals, inefficient harnessing, and poor (unpaved) roads greatly restricted the size of maximum loads and the greatest distance of daily travel.

European antiquity also saw the first uses of water-driven prime movers. Their origins are obscure, with the first reference to their existence, by Antipater of Thessalonica during the first century BCE, describing their use in grain milling. The earliest wheels were horizontal, with water directed

through a sloping wooden trough onto wooden paddles fitted to a shaft that was directly attached to a millstone above. More efficient vertical water wheels, first mentioned by Vitruvius in 27 BCE, turned the millstones by right–angle gears and operated with overhead, breast, or undershot flows. Although there were some multiple installations of Roman water wheels—perhaps most notably the set of 16 wheels at Barbegal near Arles amounting to over 30 kW of capacity—cheap slave labor clearly limited the adoption of these machines.

3. MEDIEVAL AND EARLY MODERN ADVANCES

The dominance of animate prime movers extended throughout the Middle Ages but their efficiency had improved and they were increasingly joined by gradually more powerful converters of flowing water and wind. Human statures did not show any notable growth during the medieval era but better designs of some man-powered machines were able to harness muscle power more efficiently. Man- and animal-powered tread-wheels were deployed in the construction of tall buildings and in loading and unloading ship cargoes. The combination of breeding, better feeds, more efficient harnessing, and shoeing eventually raised the performance of the best draft animals as much as 50% above the capacities that prevailed during antiquity.

The collar harness, optimizing the deployment of powerful breast and shoulder muscles, had its origins in China of the 5th century of the CE and its improved version became common in Europe five centuries later. Iron horseshoes, preventing excessive wear of hooves and improving traction, became common at approximately the same time. But it took centuries before the intensification of cropping, with more widespread rotation of food and feed (particularly leguminous) species, increased the availability of concentrate feeds and allowed for harder working draft animals. During the 18th century, a good horse was considered to be equivalent to 10 men, or at least 700 W, and the best horses would eventually surpass power equivalent to 1 kW. Whereas a pair of early medieval oxen could sustain no more than 600 W, a pair of good horses in early modern Europe delivered 1.2 kW and large horse teams (up to 40 animals) deployed in the United States after 1870 to pull gang plows or the first grain combines had draft power of at least 8 kW and up to 30 kW.

Some medieval societies began to rely on inanimate prime movers for a number of demanding tasks including grain milling, oil pressing, wood sawing, powering of furnace bellows and forge hammers, and the mechanization of manufacturing processes ranging from wire pulling to tile glazing. Waterwheels were the first machines to spread widely and the famous Domesday Book attests how common they were in England of the late 11th century: it lists 5624 water mills in southern and eastern England, one for every 350 people. A subsequent increase in the highest capacities of waterwheels was slow: it took nearly 800 years to boost the performance by an order of magnitude. Early modern Europe developed some relatively very large water-driven wheels, and although typical unit capacities of these wooden machines remained limited (by 1700 they averaged less than 4 kW), they were the most powerful prime movers of the early modern era (Fig. 2).

Ships with simple square sails were used by the earliest Old World civilizations, but the first written record of wind-driven machines comes only approximately 1000 years after the first mention of water wheels. In 947 CE, al-Masudi's report described their use to raise water for irrigating gardens in what is now eastern Iran, and the first European record dates only from the closing decades of the 12th century. Subsequently, inefficient windmills continued to be used infrequently throughout the Mediterranean and the Middle East, and even less so in India and in East Asia, and they had undergone a great deal of development in only a small number of European regions. The earliest European windmills pivoted on a massive central post that was supported usually by four diagonal quarter-bars, had to be turned to face the wind, and were unstable in high winds and their low height limited their efficiency.

However, they were widely used in grain milling and water pumping (the Dutch drainage mills being the best known example), as well as in some industrial operations. Post mills were gradually replaced by tower mills and smock mills, and during the early 17th century the Dutch millers introduced first relatively efficient blade designs (however, true airfoils, aerodynamically contoured blades with thick leading edges, originated only just before the end of the 19th century), and after 1745 the English invention of using a fantail to power a winding gear turned the sails into the wind automatically. Even with these innovations, the average power of the 18th century windmills remained below 5 kW.

The Late Middle Ages and the early modern era were also the time when wind energy was harnessed

more effectively for Europe's great seafaring voyages. The rise of the West clearly owes a great deal to an unprecedented combination of harnessing two different kinds of energy: better, and larger, sailships equipped with newly developed heavy guns. Once the medieval ships became rigged with a larger number of loftier and better adjustable sails, increased in size, and acquired stern-post rudders and magnetic compasses (both invented in China), they became much more efficient, and much more dirigible, converters of wind energy. These ships carried first the Portuguese and then other European sailors on increasingly more audacious voyages. The equator was crossed in 1472, Columbus led three Spanish ships to the Caribbean in 1492, Vasco da Gama rounded the Cape of Good Hope and crossed the Indian Ocean to India in 1497, and in 1519 Magellan's *Victoria* completed the first circumnavigation of the earth. The inexorable trend of globalization was launched with these sailings.

Gunpowder was another Chinese invention (during the 11th century) that was better exploited by the Europeans. The Chinese cast their first guns before the end of the 13th century, but Europeans were only a few decades behind. Within a century, superior European gun designs transformed the medieval art of war on land and gave an offensive superiority to large sailships. Better gun-making was obviously predicated on major medieval advances in ore-mining and metal-smelting techniques that are exhaustively described in such classics as Biringuccio's and Agricola's volumes from 1540 and 1556, respectively.

These innovations reduced the need for energy inputs, particularly in the iron-making blast furnaces that appeared first in the lower Rhine valley before the end of the 15th century. As these furnaces grew in volume, charcoal's fragility limited their height and the annual volume of individual smelting operations. Larger operations also required the use of waterpower (for blasting and subsequent metal forging) and this demand restricted the location to mountainous areas. But the main challenge was to keep them supplied with charcoal, and the English predicament is the best illustration.

By the early 18th century, a typical English furnace produced only approximately 300 tons of pig iron per year, but with at least 8 kg of charcoal per kilogram of iron and 5 kg of wood per kilogram of charcoal, its annual demand was approximately 12,000 tons of wood. With nearly all natural forests gone, the wood was cut in 10- to 20-year rotations from coppicing hardwoods yielding between 5 and 10 tons/ha. This means that approximately 2000 ha of coppicing hardwoods, a circle with a radius of 2.5 km, were needed for perpetual operation. Nationwide (with nearly 20,000 tons of pig iron produced annually during the 1720s), this translated (together with charcoal needed for forges) to at least 1100 km^2 of coppiced growth. To produce 1 million tons with the same process would have required putting at least one-quarter of the British Isles under coppiced wood, an obvious impossibility. Yet starting in the mid-1830s, Great Britain began smelting more than 1 million tons of iron per year and yet some the country's forests began regrowing; coke and steam engines made that possible.

4. TRANSITIONS TO MODERNITY

Millennia of dependence on animate power and biomass fuels came to an end only gradually and the great transition to fossil fuels and fuel-consuming engines had highly country-specific onsets and durations. Differences in accessibility and affordability explain why traditional energy sources were used for so long after the introduction of new fuels and prime movers. For example, four kinds of distinct prime movers, in addition to human labor, coexisted in parts of Europe for more than 150 years between the late 18th and the mid-20th centuries before internal combustion engines and electric motors became totally dominant: draft animals (both in agriculture and in city traffic), water wheels (and, since the 1830s, water turbines), windmills, and steam engines. In the wood-rich United States, coal surpassed fuelwood combustion and coke became more important than charcoal only during the 1880s.

Moreover, the epochal energy transition from animate to inanimate prime movers and from biomass to fossil fuels has yet to run its global course. By 1900, several European countries were almost completely energized by coal—but energy use in rural China during the last year of the Qing dynasty (1911) differed little from the state that prevailed in the Chinese countryside 100 or 500 years earlier. Even in the early 1950s, more than one-half of China's total primary energy supply was derived from woody biomass and crop residues. The share of these fuels had been reduced to 15% of China's total energy use by the year 2000, but it remains above 70%, or even 80%, for most of the countries of sub-Saharan Africa (in the year 2000, India's share of traditional biomass fuels was approximately 30% and that of Brazil was approximately 25%), and globally it still accounts for nearly 10%.

Industrialization of the British Isles is, of course, the best known case of an early transition from wood to coal and England was the first country to accomplish the shift from wood to coal during the 16th and 17th centuries. Much less known is the fact that the Dutch Republic completed a transition from wood to peat during its Golden Age of the 17th century when it also replaced a large share of its mechanical energy needs by sailships, which moved goods through inland canals and on the high seas, and by windmills. In England and Wales, the process started as a straightforward fuel substitution in a society where the combined demand for charcoaling, ship- and house-building, heating, and cooking led to extensive deforestation. The use of coal as the fuel for a new mechanical prime mover began only after 1700 with Newcomen's inefficient steam engine. James Watt's separate condenser and other improvements (patented in 1769) transformed the existing engine from a machine of limited utility (mostly water pumping in coal mines) into a prime mover of unprecedented power suitable for many different tasks.

Watt's improved machines still had low conversion efficiency (less than 5%) but his engines averaged approximately 20 kW, more than 5 times that of the typical contemporary watermills, nearly 3 times that of windmills, and 25 times the performance of a good horse. One hundred years later, the largest stationary steam engines were 10 times as efficient as Watt's machines and rated $\approx$1 MW. After the expiration of Watt's patent, the development of high-pressure steam engines progressed rapidly, radically transforming both land and maritime travel. For centuries, horse-drawn carriages averaged less than 10 km/h, but by 1900, trains (the first scheduled services began during the 1830s) could go easily 10 times faster and carry passengers in much greater comfort (Fig. 3). Railways had also drastically lowered the cost of moving voluminous loads in areas where inexpensive canal transport was not possible. Steamships cut the length of intercontinental travel and greatly expanded and speeded long-distance trade. For example, trans-Atlantic crossing, which took more than 1 week with the sailships of the 1830s, was cut to less than 6 days by 1890.

Steam engines will be always seen as the quintessential energizers of the Industrial Revolution, that great divide between the traditional and the modern world. But this notion is far from correct, and so is the very idea of the Industrial Revolution. Coal did power the expansion of iron-making, but the textile industry, the key component of that productive transformation, was commonly energized, both in

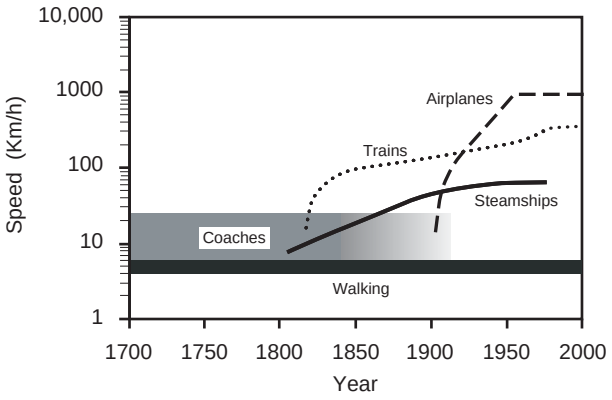


FIGURE 3 Maximum speeds of transportation during the years 1700–2000.

Europe and in North America, by water power rather than by coal combustion. Innovations to meet this demand included Benoit Fourneyron's reaction water turbine (in 1832) and an inward–flow turbine of James B. Francis (in 1847). Before the century's end, Francis' turbines and also Pelton machines (jet–driven turbines introduced in 1889) helped to launch modern water-powered electricity generation (only Viktor Kaplan's axial flow turbines are a 20th century innovation). As for the Industrial Revolution, the author sides with those historians who see it much more as a complex, multifaceted, and protracted process rather than as a clearly definable, intensive transformation powered by coal and coal-generated steam. After all, in 1850, England still had more shoemakers than coal miners and more blacksmiths than ironworkers, and in the United States, coal began providing more energy than wood only during the early 1880s.

Another little appreciated fact is that the harnessing of wind energy reached its peak at the same time that the steam engine did. America's westward territorial progression across the windy Great Plains created a demand for windmills to pump water for steam locomotives, households, and cattle. These machines were made generally of large numbers of fairly narrow slats that were fastened to solid or sectional wheels and usually equipped with either the centrifugal or the side–vane governor and independent rudders. Total sales of these relatively small (less than 1 kW) machines amounted to millions of units after 1860, whereas the capacity of approximately 30,000 larger windmills in countries around the North Sea peaked at approximately 100 MW by the year 1900.

From the vantage point of the early 21st century, it is clear that it was the next energy transition—the

process of electrification and the rising global dependence on hydrocarbons—that has left the greatest mark on individual lives as well as on the fortunes of economies and nations during the 20th century. The invention of a commercially viable system of electricity generation, transmission, and use (beginning only with incandescent lighting) by Thomas A. Edison and his associates was compressed into a remarkably short period during the early 1880s. This was followed by a no less intensive period of innovation with fundamental contributions by George Westinghouse [who correctly insisted on alternate current (AC) for transmission], Charles Parsons (who patented the first steam turbine in 1884), William Stanley (who introduced an efficient transformer in 1885), and Nikola Tesla (who invented the electric motor in 1888).

As a result, by the late 1890s, the entire electric system was basically perfected and standardized in the form that is still relied on today; the challenge ahead was to keep enlarging its unit sizes and improving efficiencies and this challenge has been met in many impressive ways. Since 1900, the maximum sizes of turbogenerators grew from 10 to $\approx$1.5 GW, AC transmission voltages rose from less than 30 kV to more than 700 kV, and in 2003, the best efficiencies of thermal generation surpassed 40% (and with cogeneration 60%), compared to as little as 5% in 1900. An inexpensive and reliable supply of electricity transformed every aspect of everyday activities by bringing bright and affordable light to both interiors and streets, by powering a still-growing array of time-saving and leisure-enhancing gadgets, and by energizing urban and intercity trains.

But the most revolutionary consequence of electrification was in industrial production. The reciprocating motion of steam engines had to be transmitted by rotating shafts and belts, resulting in friction and a great deal of lost time with accidents and allowing only limited control of power at individual workplaces. Inexpensive electric motors of all sizes changed all that: no shafts and belts, no noise and dangerous accidents, only precise and flexible individual power controls. American manufacturing was the first to make the transition in just three decades: by 1929, the capacities of industrial electrical motors accounted for over 80% of all installed mechanical power. Highly productive assembly lines (Ford pioneered them in 1913) were an obvious product of this transformation, as were many new specialized industries. And although the experimental beginnings of radio and television predate World War I, it was only after World War II that electricity began powering the new computer age with its immense flow of information and entertainment options.

Electricity has been also the principal means of easing the burden of female household labor as a growing variety of machines and gadgets took over common chores. Another great change brought about by inexpensive electricity has been the global spread of air conditioning (first patented by William Carrier in 1902). Its availability opened up the American Sunbelt to mass migration from the Snowbelt, and since the 1980s, room units have also spread rapidly among more affluent households of subtropical and tropical countries.

The age of crude oil was launched during the same decade as that of electricity, and the three key ingredients of a modern automobile—Gottlieb Daimler's gasoline-fueled engine, Karl Benz's electrical ignition, and Wilhelm Maybach's float-feed carburetor—came together during the 1890s when Rudolf Diesel also introduced a different type of internal combustion engine. Subsequent decades have seen a great deal of improvement but no fundamental change of the prime mover's essentials. Only in the United States and Canada did car ownership reach high levels before World War II; Western Europe and Japan became nearly saturated only during the 1980s. By 1904, the Wright brothers built their own four-cylinder engine with an aluminum body and a steel crankshaft to power the first flight of a machine heavier than air, and the first 50 years of commercial and military flight were dominated by airplanes powered by reciprocating engines. Jet engines powered the first warplanes by 1944; the age of commercial jet flights began during the 1950s and was elevated to a new level by the Boeing 747, introduced in 1969.

Transportation has been the principal reason for the higher demand for crude oil, but liquid fuels, and later natural gas, also became very important for heating, and both hydrocarbons are excellent feedstocks for many chemical syntheses. By 1950, crude oil and natural gas claimed approximately 35% of the world's primary energy supply, and by 2000, their combined share was just over 60% compared to coal's approximately 25% (Fig. 4). Fossil fuels thus provided approximately 90% of all commercial primary energy supply, with the rest coming from primary (hydro and nuclear) electricity. Despite a great deal of research and public interest, new renewable conversions (above all, wind turbines and photovoltaics) still have only a negligible role, as do fuel cells whose high conversion efficiencies

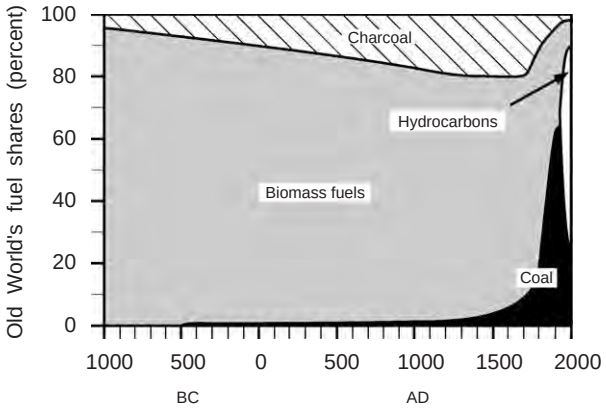


FIGURE 4 Approximate shares of global fuel consumption in the period from 1000 BCE to 2000 CE.

and pollution-free operation offer a much better way of converting gaseous and liquid fossil fuels to kinetic energy than does air-polluting combustion.

5. HIGH-ENERGY CIVILIZATION AND ITS ATTRIBUTES

Fossil fuels and electricity have helped to create the modern world by driving up farm productivity and hence drastically reducing agricultural populations, by mechanizing industrial production and letting the labor force move into the service sector, by making megacities and conurbations a reality, by globalizing trade and culture, and by imposing many structural uniformities onto the diverse world. Inevitably, all of these developments had enormous personal and collective consequences as they released hundreds of millions of people from hard physical labor, improved health and longevity, spread literacy, allowed for rising material affluence, broke traditional social and economic confines, and made the Western ideas of personal freedom and democracy into a powerfully appealing (as well as fanatically resented) global force.

But these benefits are fully, or largely, enjoyed only by a minority (only $\approx 15\%$) of the world's population. The great energy transitions of the past century raised standard of living everywhere but it has not been accompanied by any impressive decline of disparities between rich and poor societies. In the year 2000, the richest 10% of the world's population consumed more than 40% of all commercial primary energy (Fig. 5). In addition to being liberating and constructive, modern high-energy civilization is an enormous source of environmental pollution and degradation

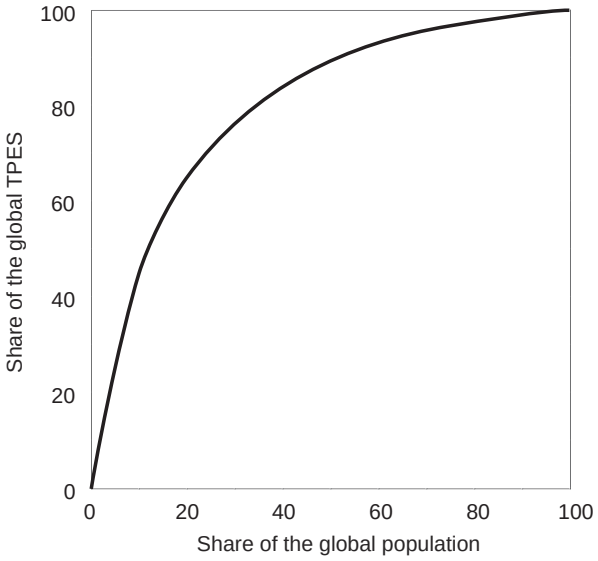


FIGURE 5 Inequality of global commercial energy use in the year 2000 is indicated by a highly convex shape of the Lorenz curve: the richest 10% of the world's population claimed approximately 45% of all energy, whereas the poorest 50% had access to just 10% of the total.

of ecosystems (perhaps even endangering the very maintenance of a habitable biosphere), is prone to many social ills accentuated by urban living, has acquired weapons of mass destruction, and is highly vulnerable to asymmetrical threats of terrorism.

People who live in affluent societies take the levels of energy that individuals and collectives control for granted, but the claims now made on energy resources are still astounding no matter if they are compared across the entire span of human evolution or just across the 20th century. Peak unit capacities of prime movers rose from less than 100 W of sustained human labor for the late Neolithic foragers to approximately 300 W for a draft ox of the early antiquity and to 2 kW for the largest Roman waterwheels. Improved versions of those machines rated approximately 5 kW by the end of the first millennium of the CE and still no more than 8 kW by 1700. A century later, Watt's steam engines pushed the peak ratings to 100 kW, by 1900 the largest steam and water turbines had capacities 100 times higher (10 MW), and steam turbines reached eventually their maximum at 1.5 GW (Fig. 2). Peak unit capacities of prime movers that deliver sustained power thus rose approximately 15 million times in 10,000 years, with more than 99% of the rise taking place during the 20th century. Increases in the destructive discharge of energies have been even more stunning: the largest tested thermonuclear

weapon (the Soviet Union's 100-megaton bomb in 1961) had power 16 orders of magnitude higher than the kinetic energy of a forager's arrow.

Because of the rapid growth of the global population, per capita comparisons yield naturally smaller multiples. Despite the near quadrupling of the global population—from 1.6 billion in 1900 to 6.1 billion in 2000—the average gross annual per capita supply of commercial energy more than quadrupled from just 14 GJ to approximately 60 GJ. In the United States, per capita energy use more than tripled to approximately 340 GJ/year, Japan's more than quadrupled to just over 170 GJ/year, and China's per capita fossil fuel use, exceptionally low in 1900, rose 13-fold between 1950 and 2000, from just over 2 to ≈30 GJ/year. These gains are far more impressive when expressed in more meaningful terms as useful energy services. Conservative calculations indicate that because of better conversion efficiencies, the world in the year 2000 had at its disposal at least 25 times more useful commercial energy than in 1900 and the corresponding multiples exceed 30 in some rapidly industrializing countries.

Perhaps the best way to compare the secular energy gains at a household level is to compare installed electric power. In 1900, a typical urban U.S. household had only a few low-power light bulbs, adding up to less than 500 W. In the year 2000, an all-electric, air-conditioned suburban house with some 400 m^2 of living area had more than 80 switches and outlets ready to power every imaginable household appliance and drawing upward of 30 kW, at least a 60-fold jump in one century. Three vehicles owned by that household would boost the total power under the household's control to close to half a megawatt! Equivalent power—though nothing like the convenience, versatility, flexibility, and reliability of delivered energy services—would have been available only to a Roman *latifundia* owner of ≈6000 strong slaves or to a 19th century landlord employing 3000 workers and 400 large draft horses.

Because the adoption of new energy sources and new prime movers needs substantial investment, it is not surprising that this process broadly correlates with the upswings of business cycles. The first wave, well documented by Schumpeter, corresponds to the rising extraction of coal and the introduction of stationary steam engines (1787–1814). The second wave (1843–1869) was stimulated by railroads, steamships, and iron metallurgy, and the third wave (1898–1924) was stimulated by the rise of electricity generation and the replacement of steam-driven machinery by electric motors in manufacturing (Fig. 6).

The initial stages of energy transitions also correlate significantly with the starts of major innovation waves that appear to be triggered by economic depressions. The first innovation cluster, peaking in 1828, was associated mainly with mobile

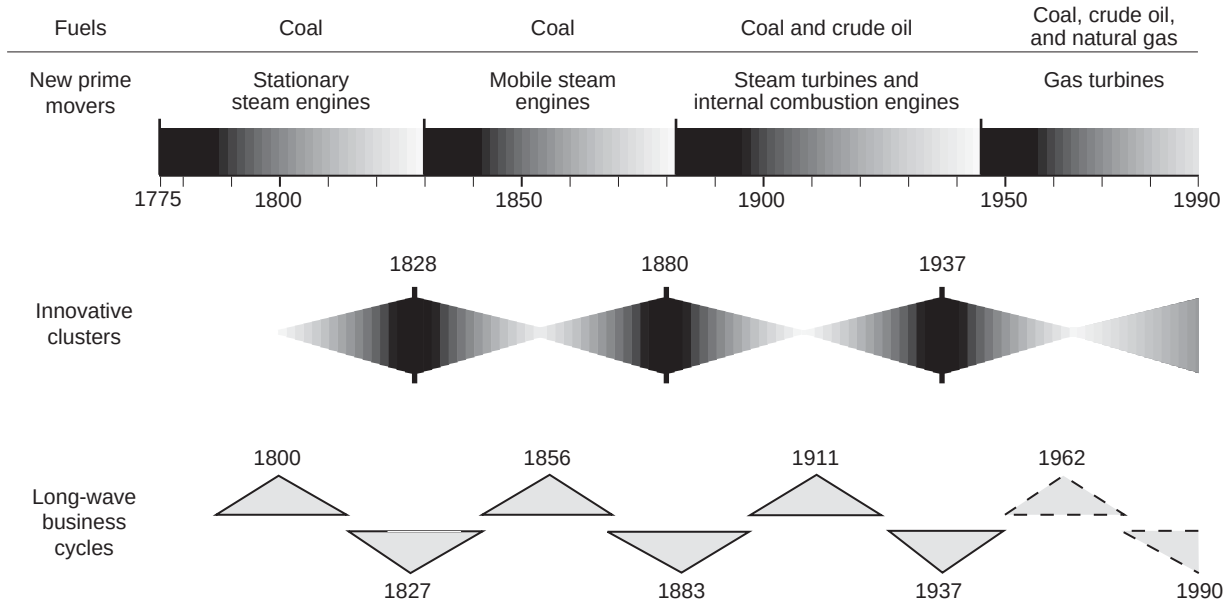


FIGURE 6 Timelines of major energy eras, innovative clusters (according to G. Mensch), and long-wave business cycles (according to J. A. Schumpeter) in the years 1775–1990.

steam engines; the second, peaking in 1880, was associated with the introduction of electricity generation and internal combustion engines, and the third cluster, peaking in 1937, included gas turbines, fluorescent lights, and nuclear energy. Post-World War II extension of these waves would include the global substitution of hydrocarbons for coal and mass car ownership; this wave was checked in 1973 by the sudden increase in the price of oil initiated by the Organization of Petroleum Exporting Countries.

6. LIMITS OF ENERGETIC DETERMINISM

When seen from a strictly biophysical point of view, energy may have an unchallenged primacy among the variables determining the course of history, but when it is considered from broader cultural and social perspectives, it may not even rank as *primus inter pares*. In addition to all of those indisputable energy imperatives, there are a multitude of none-nergy factors that initiate, control, shape, and direct human decisions to harness and use energies in a myriad of specific ways. Only if one were to equate the quality of life, or the accomplishments of a civilization, with the mindless accumulation of material possessions, would the rising consumption of energy be an inevitable precondition. But such a primitive perspective excludes the multitude of moral, intellectual, and esthetic values whose inculcation, pursuit, and upholding have no links to any particular level of energy use.

To begin with, timeless artistic expressions show no correlation with levels or kinds of energy consumption: the bison in the famous cave paintings of Altamira are not less elegant than Picasso's bulls drawn nearly 15,000 years later. It should also noted that all universal and durable ethical precepts, be they of freedom and democracy or compassion and charity, originated in antiquity, when an inadequate and inefficient energy supply was but a small fraction of today's usage. And to cite some more recent examples, the United States adopted a visionary constitution while the country was still a subsistence society running on wood; in contrast, before its collapse, the Soviet Union was the world's largest producer of crude oil and natural gas—yet all that country could offer its citizens was an impoverished life in fear, in a cage they were not allowed to leave. That political freedoms have little to do with energy use can be clearly seen by inspecting the list of the world's least free countries: it includes not only energy-poor Afghanistan, Vietnam, and Sudan, but also oil-rich Libya and Saudi Arabia.

Long-term trends in population growth are another key historic variable that is hard to relate to changes in the energy base and to levels of energy use. Improved nutrition could be seen as the principal cause of tripling the European population between 1750 and 1900, but such a claim cannot be reconciled with careful reconstructions of average food energy intakes. China's example is even more persuasive: between 1700 and 1900, the Qing dynasty did not see any major change in energy sources and prime movers, and no rise in the average per capita use of wood and straw, but the country's population tripled to approximately 475 million people.

Even the links between economic output and energy use are not that simple. When seen from a physical (thermodynamic) perspective, economies are complex systems that incessantly acquire and transform enormous quantities of fossil fuels and electricity, and some very high correlations between the rate of energy use and the level of economic performance suggest that the latter may be a direct function of the former. There is little doubt that the successive positions of economic preeminence and international influence wielded by the Dutch Republic in the 17th century, Great Britain in the 19th century, and the United States in the 20th century had their material genesis in the early exploitation of fuels that yielded higher net energy returns and allowed for higher conversion efficiencies (peat, coal, and crude oil, respectively).

A closer analysis, however, reveals that the link between energy use and economy cannot be encompassed by any easily quantifiable function as national specificities preclude any normative conclusions and invalidate many intuitive expectations. Possession of abundant energy resources has been no guarantee of national economic success and their virtual absence has been no obstacle to achieving enviable economic prosperity. A long list of energy-rich nations that have nevertheless mismanaged their fortunes must include, to name just the three most prominent cases, the Soviet Union, Iran, and Nigeria. The list of energy-poor nations that have done very well by any global standard must be headed by Japan, South Korea, and Taiwan. And countries do not have to attain specific levels of energy use in order to enjoy a comparably high quality of life.

Although it is obvious that a decent quality of life is predicated on certain minima of energy use, those countries that focus on correct public policies may realize fairly large rewards at levels not far above

such minima, whereas ostentatious overconsumption wastes energy without enhancing the real quality of life. A society concerned about equity and willing to channel its resources into the provision of adequate diets, availability of good health care, and accessibility to basic schooling could guarantee decent physical well-being, high life expectancy, varied nutrition, and fairly good educational opportunities with an annual per capita use of as little as 40–50 GJ of primary energy converted with efficiencies prevailing during the 1990s.

A better performance, pushing infant mortalities below 20/1000, raising female life expectancies above 75 years, and elevating the UNDP's Human Development Index (HDI) above 0.8 appears to require at least 1400–1500 kgoe (kilograms of oil equivalent) of energy per capita, and in the year 2000, the best global rates (infant mortalities below 10/1000, female life expectancies above 80 years, HDI above 0.9) needed no less than approximately 2600 kgoe/capita (Fig. 7). All of the quality-of-life variables relate to average per capita energy use in a nonlinear manner, with clear inflections evident at between 40 and 70 GJ/capita, with diminishing returns afterward and with basically *no* additional gains accompanying consumption above 110 GJ/capita or ≈2.6 metric tons of crude oil equivalent. The United States consumes exactly twice as much primary energy per capita as does Japan or the richest countries of the European Union (340 GJ/year versus 170 GJ/year), but it would be ludicrous to suggest that American lives are twice as good. In reality, the United States falls behind Europe and Japan in a broad range of quality-of-life variables, including higher infant mortality rates, more homicides, lower scientific literacy, and less leisure time.

Finally, energy use is of little help in explaining the demise of the established order. The long decline of the Western Roman Empire cannot be tied to any loss of energy supplies or a drastic weakening of energy conversion capabilities and neither can the fall of the French monarchy during the 1780s, the collapse of the Czarist empire in 1917, or the Nationalist retreat from mainland China during the late 1940s. Conversely, many historically far-reaching consolidations (including the gradual rise of

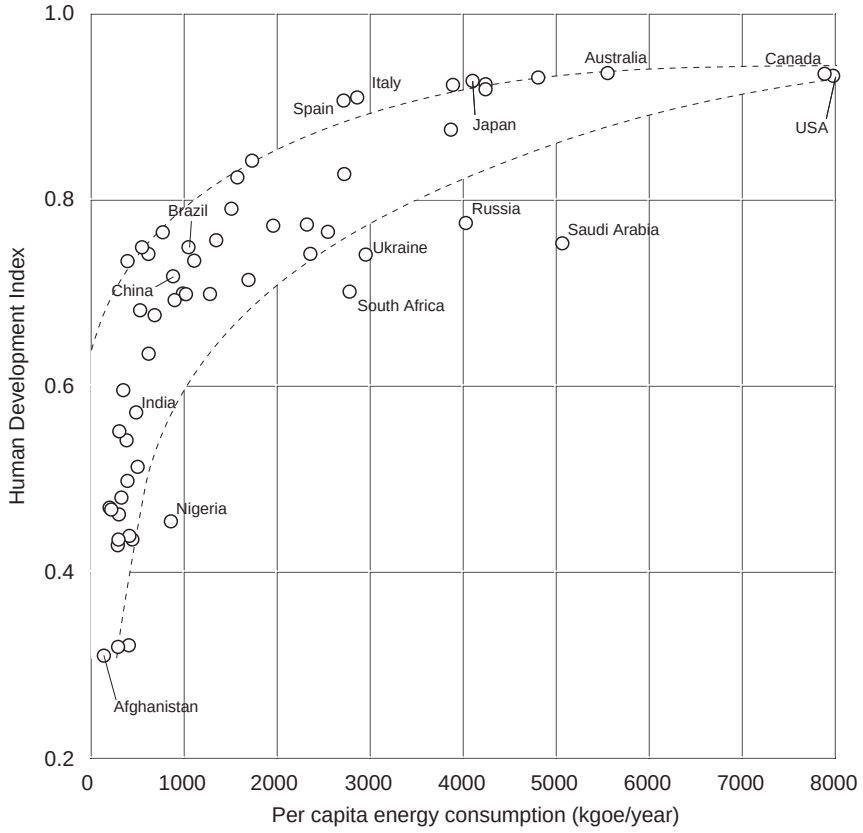


FIGURE 7 Plot of the Human Development Index against the average annual per capita use of commercial energy in the year 2000, which shows that virtually no quality-of-life gains accrue with consumption above 2.6 metric tons of oil equivalent.

Egypt's Old Kingdom, the rise of the Roman Republic, or the rise of the United States) and lightning expansions of power (including the spread of Islam during the 7th and 8th centuries or the Mongolian conquest of the 13th century) cannot be linked to any new prime movers or to better uses of prevailing fuel.

There is clearly an ambivalent link between energy and history. Energy sources and prime movers delimit the options of human history and determine the tempo of life, and, everything else being equal, thermodynamics requires that higher socioeconomic complexity must be supported by more intensive flows of energy. And yet, neither the possession of abundant energy sources nor a high rate of energy consumption guarantees the security of a nation, economic comfort, or personal happiness. Access to energies and the modes of their use constrain the options for human beings' actions but do not explain the sources of people's aspirations and the reasons for the choices they make and do not preordain the success or failure of individual societies at a particular time in history. Indeed, the only guaranteed outcome of higher energy use is greater environmental burdens whose global impacts may imperil the very habitability of the biosphere. To prevent this from happening, humanity's most important future choice may be to limit the use of energy and thus to embark on an entirely new chapter of history.

SEE ALSO THE FOLLOWING ARTICLES

Coal Industry, History of • *Cultural Evolution and Energy* • *Early Industrial World, Energy Flow in* • *Energy in the History and Philosophy of Science* • *Fire: A Socioecological and Historical Survey* • *Hunting and Gathering Societies, Energy Flows in* • *Nationalism and Oil* • *Oil Crises, Historical Perspective* • *Oil Industry, History of* • *OPEC, History of* • *Sociopolitical Collapse, Energy and* • *War and Energy*

Further Reading

Cowan, C. W., and Watson, P. J. (eds.). (1992). "The Origins of Agriculture." Smithsonian Institution Press, Washington, DC.

Finniston, M., *et al.* (eds.). (1992). "Oxford Illustrated Encyclopedia of Invention and Technology." Oxford University Press, Oxford, UK.

Forbes, R. J. (1972). "Studies in Ancient Technology." E.J. Brill, Leiden, the Netherlands.

Goudsblom, J. (1992). "Fire and Civilization." Allen Lane, London.

Langdon, J. (1986). "Horses, Oxen and Technological Innovation." Cambridge University Press. Cambridge, UK.

McNeill, W. H. (1989). "The Age of Gunpowder Empires, 1450–1800." American Historical Association, Washington, DC.

Nef, J. U. (1932). "The Rise of the British Coal Industry." Routledge, London.

Pacey, A. (1990). "Technology in World Civilization." The MIT Press, Cambridge, MA.

Schumpeter, J. A. (1939). "Business Cycles: A Theoretical and Statistical Analysis of the Capitalist Processes." McGraw-Hill, New York.

Schurr, S. H. (1984). Energy use, technological change, and productive efficiency: An economic–historical interpretation. *Annu. Rev. Energy* **9,** 409–425.

Singer, C., *et al.* (eds.). (1954–1984). "A History of Technology." Clarendon Press, Oxford, UK.

Smil, V. (1991). "General Energetics: Energy in the Biosphere and Civilization." Wiley, New York.

Smil, V. (1994). "Energy in World History." Westview Press, Boulder, CO.

Smil, V. (2003). "Energy at the Crossroads: Global Perspectives and Uncertainties." MIT Press, Cambridge, MA.

CHRONOLOGY OF ENERGY-RELATED DEVELOPMENTS

The following chronology presents a list of events that have had significant implications for the field of energy. This section has been compiled by chief editor Cutler Cleveland, in collaboration with Christopher Morris of the Academic Press/Elsevier publishing staff. The associate editors supplied material for a number of the entries that appear here, as well as providing other valuable advice. In addition, various encyclopedia authors furnished material for entries within their specific area of expertise. The chronology includes more than 1,500 entries in all.

Every effort has been made to establish the proper date for each entry in the chronology, but it must be acknowledged that certain dates here represent the best possible estimate. In general the earlier the date of an entry, the more likely it will be an approximation, or an expression of a range of possible dates. We also recognize that for some milestones of energy there are competing claims as to which individual, group, or nation deserves the credit for having been first, and we have tried to reflect this in the wording of the entries.

Earliest Times to 4500 BC

about 4.5 billion years ago Suggested time at which the Sun's fusion reactions begin, and solar energy reaches the Earth.

about 600 million years ago Proposed time at which the Earth's ozone layer forms; i.e., enough ozone is present in the stratosphere to sufficiently shield Earth from biologically lethal ultraviolet radiation. Prior to this period, life was restricted to the ocean; the presence of ozone is a fundamental prerequisite for organisms to develop and live on the land.

about 500 million years ago The organic precursors to fossil fuels are created. Marine and terrestrial plants absorb solar energy and use it to produce carbohydrates; these and other organic materials eventually settle on the ground and in stream, lake, and sea beds. As they become buried, they are transformed by heat and pressure into solid, liquid, or gaseous hydrocarbons (coal, oil, or natural gas).

about 350 million years ago Evidence for the first forest fires on Earth appear in the geological record.

2.5 to 1.5 million years ago Proposed time span for the Oldowan tradition of East Africa, the earliest known use of stone tools by the forerunners of modern humans.

1 million to 500,000 BC Proposed time span for the development of more sophisticated stone tool use at various sites.

500,000 BC Estimated time of the first controlled use of fire by humans. Various scholars have proposed dates ranging from 1.5 million to 150,000 years ago; the large range is due to debate over the archaeological and paleontological evidence, and in particular the problem of distinguishing natural fire events (as from lightning) from human-produced fire.

40,000 BC Estimated date for the first use of artificial light in lamps, which were probably in the form of a hollow rock, shell, or other such natural object, fueled with animal fats and using lichens, moss, or other plant fiber for a wick.

24,000 BC Pits dug at sites in the East European plain constitute the first evidence of storage of non-food items, such as fuel and other raw materials.

10,000 BC Estimated date for the earliest use of hot springs (e.g., North America, Anatolia, Japan) for cooking and bathing and as religious sites or sanctuaries.

8000 BC Estimated date for the first use of various forms of human transportation, such as sledges, skis, and primitive boats or rafts.

8000–7000 BC Estimated time span for the beginnings of agriculture worldwide, based on evidence such as wheat being grown in Mesopotamia and corn in the Oaxaca Valley of Mexico.

7000 BC Improved versions of lamps are in use, with oils derived from a variety of plant and animal fats as fuel. Wicks made from hemp, flax, wool, or cotton are used to control the rate of burning.

6000 BC Humans begin to utilize animal power as an energy source through the domestication of various animals such as the wild ox, the horse, the pack ass, and the camel.

6000 BC The Bible has numerous references to the use of pitch, which is asphalt or bitumen in its soft state. Noah follows God's instructions to cover the Ark with pitch. The reed basket that carries the infant Moses into the Nile River is waterproofed with pitch. The Tower of Babel is built with "bricks for stone and slime they had for mortar." (*Slime* is another word for pitch.)

5000 BC Evidence of ocean-going reed boats being used in the Persian Gulf. Other indications of early use of reed boats exist in sites such as the Sudan, coastal Peru, Lake Titicaca, New Zealand, and the marshes of southern Iraq.

4800 BC Residents of the Nabta Basin in Egypt construct stone calendar circles that target due north and the summer solstice.

4600 BC Farmers begin to use the plow rather than the hoe as a farm implement, in Mesopotamia or possibly Egypt.

4500–2000 BC

4241 BC The earliest known official recording of a date is registered in the calendar of ancient Egypt.

4000 BC Egyptians build boats made from planks joined together. Previously, boats were dug out from logs, or made of reeds bound together or skins stretched over a framework.

4000 BC The crafts of glassmaking, brickmaking, and copper smelting are in existence.

4000 BC The earliest proposed date for the use of the horse for riding, on the basis of excavations from Dereivka in the Ukrainian steppes that show possible signs of bit wear on horse teeth.

3800 BC Based on boat fragments found in southeast Turkey, suggested date for the earliest transport of trade goods by riverboat, including petroleum products.

3500 BC Earliest evidence of sailing ships, on the Nile river and in the Mediterranean.

3200 BC Earliest specific evidence of the use of wheeled vehicles, at the Uruk site in what is now Syria.

3100 BC The earliest known use of candles, as indicated by candle holders found in Crete and Egypt. Early fuels are derived from local materials, e.g., wax from insects molded in paper tubes, various forms of tallow (rendered animal fat).

3000 BC Coal is in use as a fuel, based on evidence found in various ancient sites such as China and Wales.

3000 BC Mesopotamians use "rock oil" (an early term for petroleum) in architectural adhesives, ship waterproofing, medicines, and roads.

3000 BC The Sumerians divide the day into 24 hours, the hour into 60 minutes, the minute into 60 seconds, and the circle into 360 degrees.

3000 BC The use of oars to power a boat is established in Egypt.

3000 BC Various Mideast peoples use deposits of bitumen (asphalt) in the form of natural seepage from sites such as the Dead Sea for a number of purposes including road building, bonding and waterproofing, and the making of artifacts.

2950–2900 BC The earliest complex is constructed at the Stonehenge site in southern England. Subsequent periods of construction (2900–2400 BC and 2550–1600 BC) culminate in the complicated setting of large stones that still exists today.

2900 BC Specialized oil lamps are in regular use in the Middle East, probably employing olive oil or possibly beeswax as fuel.

2800 BC Solar orientation is the basic principle of design for Egyptian temple construction, with pyramids positioned to show the change in the Earth's location relative to the Sun over the course of a year. A temple in Kamak, honoring the sun god Ra, is oriented so that sunlight penetrates the building in a precise beam.

2560 BC Construction begins for the Great Pyramid of Khufu (Cheops), one of the Seven Wonders of the World in ancient times and the only one still in existence.

2200 BC Crude oil is refined in China for use in lamps and in heating homes.

2000–600 BC

2000 BC Horse-drawn vehicles are in use in Egypt; also contemporaneously or subsequently in China.

2000 BC The Greeks use water power to turn waterwheels for grinding wheat into flour.

2000 BC Various peoples construct baths using geothermal hot springs.

1900 BC Spoked wheels have appeared in Egypt and Palestine. These wheels are significantly lighter than solid wheels, allowing a wheeled cart to carry a heavier load.

1900 BC The Greeks employ an early form of passive solar heating.

ca. 1810 BC King Amenemhet III of Egypt completes a system to regulate the inflow of water into the large Lake of Fayoum (Lake Moeris) depression, southwest of Cairo. This includes one of the world's earliest dams and provides both a flood control for the seasonal overflow of the Nile and a reservoir for irrigation.

1600 BC The *shadouf* (or *shaduf*) is used to raise water from a river or well. This is a form of lever consisting of a pole balanced on a pivot, with a weight at one end and a bucket at the other. This first appears in Egypt or possibly India; it is still in use today.

1500 BC Smelted iron tools come into in use. Armenia is traditionally thought of as the first site for this.

ca. 1380 BC A canal diverging from the Nile extends along the Wadi Tumilat to Heroopolis, a port at the head of the Bitter Lakes, providing a shipping passage between the Nile and the Red Sea. The path of this canal was frequently followed by the builders of the modern Suez Canal.

1000 BC Phoenician warships and merchant ships travel over the Mediterranean Sea, employing a large square sail and banks of oarsmen.

900 BC Hand-dug oil wells near the Gulf of Suez are used as a source to extract bitumen for use as a preservative in Egyptian mummies.

900 BC The first practical use of natural gas takes place when the Chinese use it for brine evaporation.

800 BC The Scythians, a unified group of nomadic tribes from the southern Russian steppes, display great skill and mobility as mounted archers.

ca. 692 BC Sennacherib, king of Assyria, constructs a highly advanced aqueduct at Jerwan, a portion of which still exists today.

600s BC The Greeks begin making terra cotta lamps to replace hand-held torches. The word "lamp" is derived from the Greek term *lampas,* meaning "torch."

600–300 BC

ca. 600 BC Ancient Assyrian writings tell of the "Eternal Fire," a continuously burning flame from a natural oil spring. In the modern era the Iraqi oil industry will come into being at this same site.

ca. 585 BC The Greek philosopher Thales of Miletus notes that an amber rod attracts to itself light materials, such as straw, when it is rubbed. Though not understood at the time, this is the first known observation of the phenomenon of static electricity.

500s BC One of the first industrial uses of pitch is in the Hanging Gardens of Babylon of King Nebuchadnezzar II, one of the Seven Wonders of the ancient world. The gardens are built of brick and asphalt protected by lead.

500s BC Ancient Chinese proverbs describe what will become known in the modern era as an ecological pyramid based on energy relations among trophic (feeding) levels: (1) "The large fish eat the small fish, the small fish eat the water insects, the water insects eat plants and mud"; (2) "Large fowl cannot eat small grain"; (3) "One hill cannot shelter two tigers."

500 BC Construction begins on China's Grand Canal; it is extended over the centuries by dredging new sections and by connecting existing canals and natural rivers until it eventually becomes the longest canal in the world.

480 BC The Persians use flaming arrows wrapped in oil-soaked fibers at the siege of Athens.

ca. 450 BC The Greek philosopher Empedocles refines the existing belief that the universe consists of four basic properties—earth, air, water, and fire—by stating that all types of matter are composed of varying proportions of these. This idea holds sway in the West for roughly the next 2,000 years.

ca. 420 BC The Greek philosopher Democritus postulates that all matter consists of tiny particles called atoms that cannot be divided, and that their constant motion explains the physical nature of the universe.

400s BC Spurred in part by wood fuel scarcity, classical Greek architecture develops widespread use of passive solar energy. Socrates states the basic principles of this practice: homes are oriented toward the south, and eaves are added to shade for south windows in summer; entire cities are designed to allow equitable access to the winter sun.

400–200 BC Greek thinkers such as Aristotle, Hippocrates, Euclid, and Pythagoras lay the foundations for modern science through their inquiries that emphasize evidence, logic, and and deduction.

ca. 380 BC Many parts of of Greece are completely bare of trees due to their harvest for fuel, smelting operations, and shipbuilding. The philosopher Plato laments, "The mere skeleton of the land remains." A tax is then placed on wood for home heating and cooking, and the supply of wood is regulated by public authorities in some localities.

ca. 350 BC Aristotle proposes the use of distillation as a means of desalinating sea water.

ca. 340 BC Aristotle describes how heat from the Sun's rays sets in motion the circulation of huge masses of air and moisture in the atmosphere. This concept remains unchallenged for nearly 2000 years.

ca. 340 BC Aristotle provides the first technical definition of *energeia*, a word formed by combining two root forms meaning "at" and "work." He uses this to describe the operation or activity of anything; the modern term "energy" ultimately will be derived from this concept of *energeia*.

325 BC The Greeks develop an accurate form of the clepsydra (water clock), improving on a technology that is thought to have originated in Egypt.

ca. 315 BC The Greek philosopher Theophrastus publishes *De Lapidus,* which uses the term *anthrax* to refer to a black stone that would burn (coal).

312 BC Rome's Via Appia (the Queen of Roads) is completed, as well as the Aqua Appia, the first aqueduct to serve the city; both are named for Roman censor Appius Claudius Cieco.

ca. 300 BC Syrian glassmakers invent the blowpipe to enable the production of glass in hollow shapes, followed by the invention of the two-part mold that enables multiple production of glass products.

3rd to 1st Century BC

ca. 270 BC Aristarchus of Samos, an early Greek astronomer, estimates the distance and size of the Sun, and proposes that the Earth goes around it.

ca. 270 BC The Great Lighthouse at Alexandria (Egypt) is constructed on the island of Pharos, using fire at night and reflecting the Sun's rays from a giant mirror during the day. It stands for some fifteen centuries until being destroyed by earthquakes in 1326.

ca. 250 BC Greek mathematician Archimedes states that a body immersed in a fluid is buoyed up by a force equal to the weight of the displaced fluid, which explains the buoyancy of ships. He is also credited with having discovered the principle of the lever and with providing a foundation for modern calculus.

240 BC The lift-and-force piston pump is used to raise water, an invention attributed to the Greek engineer Ctesibius of Alexandria. He is also said to have invented a pneumatic catapult that hurled stones by combining a spring device with a compressed air system.

212 BC Solar energy is reportedly employed to defend the harbor of Syracuse against a Roman siege. Archimedes uses a "burning glass" to set fire to the ships of the Roman fleet from shore. Possibly this was a parabolic mirror of polished metal combined with a large series of smaller mirrors; later experts disagree as to whether such a technique is possible.

ca. 211 BC The first known well is drilled for natural gas in China to a reported depth of 500 feet. The Chinese are said to drill their wells with primitive percussion and bamboo poles as pipelines. The gas is burned to dry rock salt.

ca. 210 BC The Archimedes screw (a helix-shaped screw in a tube) is employed to lift water from rivers for irrigation.

ca. 191 BC A temple in Rome contains an "eternal fire" that burns continuously in homage to Vesta, the goddess of the hearth; this is probably a natural gas flame.

150 BC The quern, a rotary hand mill used to grind grain, is developed, likely by the Romans. Over the next several centuries this comes into widespread use in Europe through Roman influence there.

mid 100s BC A decline in forest cover occurs in the environs of Rome because of demand for wood for industrial fuel, home and ship construction, central heating, and lavish baths.

ca. 130 BC Hipparchus of Rhodes, a Greek astronomer, applies mathematical principles to the determination of locations on the Earth's surface, expressed in terms of their specific longitude and latitude.

ca. 122 BC The Chinese writer Liu An refers to "ice-charcoal"; this is assumed to be the first written reference to the mineral coal, which is used in China by this time for home heating.

100s BC Chinese farmers use an iron plow with a moldboard (a curved plate above the plowshare that lifts and turns over the soil, creating a true furrow). Iron plows are also in use in Britain at this time, but moldboard technology is not reported in Europe until centuries later.

100s BC Waterwheels are in operation in Greece and Rome and probably elsewhere in southern Europe.

ca. 85 BC The earliest known mention of the specific use of water power to replace human labor is made by the poet Antipater of Thessalonica, who refers to maids having their hard work of hand milling relieved by the power of a water mill.

ca. 80 BC The Romans employ the technique of hypocaust heating in baths and other buildings, the first known system of central heating.

ca. 63 BC King Mithradates the Great of Pontus (Asia Minor) reportedly uses a hydraulic machine, presumably a water mill.

ca. 25 BC Glass melting, working, and forming technology flourishes in the Mediterranean region of the Roman Empire.

ca. 10 BC Vitruvius, a Roman engineer, publishes the earliest known description of a vertical water mill.

1st to 10th Century AD

ca. 20 AD Imperial orders in China reportedly set down the first environmental protection laws, including energy-related issues such as restrictions on wood burning and mining.

ca. 37 AD The first greenhouse is reportedly used to grow cucumbers for Tiberius Caesar, emperor of Rome.

47 AD Roman historian Tacitus reports on the digging of a 23-mile canal by Roman soldiers in Germany, linking the Rhine and Meuse rivers.

ca. 60 AD Hero (Heron) of Alexandria publishes *Pneumatica,* which describes over 100 different machines, including gears, a lever, and the aeolipile, the first working steam engine. The aeolipile is similar in operation to the modern jet engine.

77 AD Roman author Pliny the Elder describes people living near the Rhine River delta (now the Netherlands) who dig lumps of earth which they then dry and burn for cooking and heating. This is the first known mention of the use of peat as fuel.

100 AD The Romans construct a huge bridge near the city of Nesausas (modern-day Nimes, France), still standing and now known as Pont Du Gard. It features an aqueduct that is part of a system bringing water to the city from springs 50 km distant.

ca. 105 AD Paper manufacture is reported in China, with a man named Ts'ai Lun credited as the inventor. Over the next several centuries papermaking spreads to other parts of Asia.

ca. 110 AD The Roman occupiers of Britain use coal obtained from surface mines for heating and metalworking. The early peoples of Britain apparently mined coal long before this date; large quantities of burned coal fragments have been discovered at an Iron Age settlement near Edinburgh and, as even earlier evidence, flint axes have been found embedded in coal in Derbyshire.

150 AD An efficient reservoir and dam for irrigation is built at Saurashtra, India.

150 AD The Romans advance Greek solar technology by employing clear window coverings such as glass to enhance the capture of solar energy and using dark colors and pottery to store thermal energy. "Sun rights" guarantees are codified in Roman law.

300 AD A Roman flour mill at Barbegal, near Arles, in southern France, operates a large series of water wheels, each turning a pair of millstones; this complex is said to have a total capacity of 3 tons of grain per hour.

300s AD The Roman *Cursus Publicus* (imperial postal system) extends over more than 80,000 km.

300s AD Water mills appear in the Roman Empire.

ca. 320 AD Pappus of Alexandria identifies five elementary machines in use: the cogwheel, lever, pulley, screw, and wedge. This follows on a similar description by his predecessor Hero (ca. 60 AD).

500 AD The modern horse collar is in use in China, thus allowing a single draft animal to pull a much heavier load. This follows the earlier development of the stirrup (ca. 300 AD), also reportedly in China.

500s AD The heavy plow is in use in Slavic Europe. The earlier Roman light plow was suited to Mediterranean soil but not to the rich, heavy, often wet soils of Northern Europe. Over the next two centuries the heavy plow is introduced elsewhere in Europe, e.g., Northern Italy's Po Valley and the Rhineland.

600 AD Japanese excavators hand dig wells to depths of 600 to 900 feet.

600s AD Arab and Persian chemists discover that petroleum's lighter elements can be mixed with quicklime to make Greek fire, which was in effect the napalm of its day, used as an incendiary weapon in warfare.

600s AD Windmills are used in Persia to grind grain; also possibly in the same period or earlier in China. At this time rude vertical axis panemones (wind turbines) are used to grind grain in the Afghan highlands.

ca. 700 AD Scholars in the Arab world adopt the Indian system of numbering, which is more useful and sophisticated than the extant method of Roman numerals. Though some argue this system should be called Hindu numerals because of its likely origin, it becomes known in the West as Arabic numerals because Europe learns of it from the Arab world.

700s AD The Catalan forge comes into use in Andorra. This produces significantly higher temperatures than existing furnaces and allows larger amounts of ore to be smelted at one time.

ca. 750 AD Papermaking reaches the Islamic world from contact with China. The Arab world then improves this technology over the next two centuries by the use of linen fibers.

ca. 776 AD Works written by Geber (Jabir Ibn Haiyan), and by later Arab writers but attributed to him, introduce experimental investigation into alchemy and establish many of the principles of modern chemistry. In 1144 the Englishman Robert of Chester translated Geber's major works into Latin.

late 700s AD The earliest known tidal-powered mills of Europe are in operation on the northeast coast of Ireland.

800s AD The horse collar, stirrup, and moldboard plow appear in Europe.

ca. 830 AD The Muslim mathematician al-Khwarizmi describes various important concepts of modern mathematics; the current term *algorithm* derives from his name and *algebra* comes from the title of a book he wrote. He is considered the main source of European knowledge of the Arabic numeral system.

900 AD Sailing vessels have appeared in various waters with multiple masts and lateen (triangular) sails that make directional sailing possible, whereas earlier square-sail vessels were limited to sailing with the wind. The probable origin of this technology is Arab merchant ships on the Red Sea.

ca. 950 AD Gunpowder is developed, most likely in China, though claims have been made for several other cultures. Its use spreads to Europe at some point during the next two centuries, probably as a result of the Mongol invasion.

ca. 980 AD The pound lock for canals comes into use in China, thus allowing boats to move from a lower to significantly higher stretch of water.

late 900s AD Muslim chemists in Moorish Spain practice the technique of distillation to produce alcohol for use as a solvent and antiseptic (though not for alcoholic beverages, in accordance with Islamic law). The word *alcohol* comes from "Al-Kohl," an Arabic term for a cosmetic powder that was a product of the distillation process.

11th to 14th Century

1000–1400 The age of feudalism in Europe: castles, knights, religious fervor, only very rudimentary science and technology. The Vikings sail as far as Greenland and America. Mongol warriors overrun and subjugate southern Russia.

ca. 1000 The magnetic compass is discovered in China.

early 1000s The waterwheel is widely used in Western Europe; it remains the main source of large-scale power there until well into the 18th century.

1031 Waste paper is reportedly re-pulped into new paper in Japan; this is considered the earliest known example of recycling.

ca. 1050 The Anasazi peoples of southwestern North America build distinctive multistoried cliff dwellings with southern exposure, providing passive solar heating.

1066 The Norman Conquest alters the profile of energy use in England, as the subsequent growth of towns and trade leads to greater use of water power and water transportation.

1086 A total of 5,624 water-driven mills are in use by the manors of England south of the Trent River, or about one mill for each 400 people. This is according to the *Domesday Book,* a statistical survey that survives as a portrait of medieval life.

1100s Arabs in Spain introduce to Western Europe the practice of distilling petroleum products for illumination.

1100s The Chinese drill wells with primitive boring equipment to depths as great as 3,500 feet.

ca. 1130 Gunpowder is employed for weaponry in China, as primitive rockets, mortars, and cannons are in use. Similar weapons are reportedly in use by Moorish armies at around the same time.

ca. 1150 The first known paper mill in Europe is built by the Moors in Xativa, Spain. Papermaking then becomes established in Italy and gradually spreads across Christian Europe, though it does not become truly important until the invention of the printing press.

mid 1100s Monks and lay brothers of the Cistercian order display special talents in the use of water resources and water power, e.g., irrigation channels, dams, mills, and canals. In their agricultural practices they also exhibit an early awareness of modern ecological concepts such as reclamation, sustainable development, and food webs.

1180 Windmills are widely in use in France, England, and elsewhere in Europe. The first documented appearance of the familiar European or "Dutch" style of windmill is recorded in Normandy at around this time. A written record of a windmill in Yorkshire, England dates from 1185.

ca. 1187 Working in Toledo, Spain, Italian scholar Gerard of Cremona provides Latin translations of many important thinkers from the Greek and Arab world, such as Ptolemy, Euclid, Geber, and al-Khwarizmi, thus providing Europe with scientific knowledge that had become lost or obscured during the early Middle Ages.

ca. 1200 A treadmill hoisting machine powered by donkeys is installed at the Abbey of Mont St. Michel on the French seacoast.

ca. 1200 Reinier, a monk of the Priory of St. Jacques, Liège, Belgium, describes the use of coal for metalworking. At this time coal is being exported from northeastern England to other parts of Britain and elsewhere in Western Europe.

1202 Italian mathematician Fibonacci (Leonardo Pisano) of Pisa publishes *Liber Abaci,* a book that introduces the Arabic numeral system into Europe.

1230 The first importing of Scandinavian timber is recorded in England, indicating that wood as a domestic fuel supply was already experiencing shortages.

ca. 1230 Robert Grosseteste, chancellor of Oxford University, breaks with medieval thought to stipulate that science should be based on observation and experimentation. He also theorizes on the nature of light in a way that anticipates the work of Newton.

1232 The Chinese use rockets to defend the city of Kaifeng against Mongol invaders.

ca. 1264 The young Marco Polo visits Baku, a Persian city on the shores of the Caspian Sea in modern Azerbaijan. There he observes people collecting oil from seeps, a practice that is said to date back to the time of Alexander the Great.

1267–1268 English philosopher Roger Bacon prepares three large works for Pope Clement I, in which he lays a foundation for the development of modern science with his emphasis on the experimental method, rather than established tradition, as the proper means for acquiring knowledge about the world.

1269 The first experiments in magnetism are carried out by Petrus Peregrinus de Maricourt, a French engineer, who observes and describes some of the fundamental properties of magnets.

1270 An illustration of an English windmill appears in the *Windmill Psalter,* which has its origin in Canterbury.

1273 Pollution from coal smoke offends the citizens of London. An ordinance is then enacted prohibiting the use of coal because it is "prejudicial to health." This is the first recorded example in Britain of an explicit air-pollution control.

1291 A royal charter granted to the Abbot of Dunfermline gives him and his monks the rights to dig for coal. At this time in Britain coal is obtained from shallow underground mines known as bell pits or stair pits. Another form of soft coal called "sea coal" is gathered on beaches as it washes up on the shore.

1298 Marco Polo's *Description of the World* comments on many novel customs and practices of China, including the use of "stones that burn like logs" (coal). Some scholars now question whether Marco Polo actually observed China at first hand or simply reported on what others told him, but in either case his writings made the West aware of Chinese culture.

1300s The first geothermal district heating system is set up at the town of Chaudes-Aigues in France; it employs as the heat source a natural flow of hot water from springs at 82°C (it is still in use today).

early 1300s Serious wood shortages begin to be reported in England and elsewhere in Western Europe; these will become gradually more acute over the next 300 years.

1306 King Edward II of England prohibits artificers from burning sea coal while Parliament is in session, because of its offensive smoke.

1326 The first known drawing of a cannon in Europe appears in a Florentine manuscript. In the following year King Edward III of England is reported to have used cannon in battles with the Scots. Cannon were definitely used by the English in the Battle of Crecy in 1345.

1327 The Beijing-Hangzhou Grand Canal (1800 km long) is completed, the longest man-made canal in the world. It was begun in about 600 AD (though supposedly parts of it existed at least 1,000 years earlier) and it still exists today.

1340 A *flüssofen* (flow oven) is in operation at Marcheles-Dames, Belgium. This is the earliest form of a true blast furnace, which was developed in the Rhineland as an improvement on the earlier *stückofen,* in which malleable iron was produced directly from the ore.

1347 The Black Death (Plague) reaches Italy via merchant ships after having inexorably moved westward from Asia along the trade routes over the past two decades. Over the course of the next five years the Plague devastates Europe, killing about one-third of the entire population.

ca. 1350 Franciscan alchemist Joannes de Rupescissa characterizes ethanol as "*aqua vitae rectificata,*" and identifies it as the fifth essential form of matter, in addition to earth, air, fire, and water.

1379 The first tax on coal is enacted in England; in 1421 "a duty of twopence per chaldron (an old unit of measure) was to be paid to the Crown by all persons, not franchised in the port of Newcastle, who bought coals."

late 1300s Portable firearms are in use, including the first weapon referred to as a handgun, i.e., a weapon that can be held and fired by one person.

15th Century

ca. 1400 The Hopi people of southwestern North America burn coal for cooking and heating and to bake the pottery they make from clay.

1400s Huge draft horses, originally bred to carry heavily armored knights in battle, are now in use as work animals.

1400s Major advances are made in the use of candles, including the first use of candle molds and the use of beeswax instead of tallow (producing less smoke and odor).

1408 The drainage windmill is used in Holland, employing a large scoop wheel to carry inland waters out to sea, thus permitting the use of lowland areas for agriculture.

1427 Portuguese caravels in the service of Prince Henry the Navigator discover the Azores archipelago roughly 1,500 kilometers west of the European mainland. Over the next two decades Portuguese sailors also explore the west coast of Africa, a region previously little known to Europeans.

1450 The Dutch invent the *wipmolen,* a form of windmill in which the top portion, bearing the sails, can be turned to face the wind.

1454 After more than 20 years of experimentation, German inventor Johannes Gutenberg prints 300 copies of a 42-line, 1,282-page Latin Bible, known today as the Gutenberg Bible, and inaugurates the era of movable type. At the time, all of Europe had a total of only about 30,000 books. By the year 1500, there were an estimated 9 million.

1457 A new type of carriage is introduced into France from Hungary; it features light spoked wheels and a springed undercarriage. It becomes known in Europe as the "cart of Kocs," from the town where it was built, and this leads to the modern English word *coach*.

1474 William Caxton prints the first book in English, a translation of a popular French romance.

1482 King Edward IV of England establishes 42 gallons as the standard size for a barrel of herring, in order to end cheating in the packing of fish. This becomes a standard unit of measure for various goods and eventually (mid 19th century) barrels of oil will use this same standard.

1488 Portuguese explorer Bartolomeu Dias sails to the Cape of Good Hope at the southern tip of Africa.

1492 Genoese navigator Christopher Columbus sails across the Atlantic in the service of Spain, making Europeans aware of "new" lands lying to the west of Europe. Evidence exists of earlier European presence in North America, e.g., Vikings in Newfoundland ca. 1000 AD, but it is the voyage of Columbus that establishes a permanent link between Europe and the Americas.

ca. 1493 Leonardo da Vinci sketches an experimental flying machine. It is also reported that he created a design for a bicycle at around the same time; during his lifetime he designs or describes hundreds of other innovative machines and devices, such as a windmill, a parabolic mirror, and a parachute.

1497–1498 Portuguese explorer Vasco da Gama follows the route recently established by Bartolomeu Dias to sail around the southern coast of Africa, then proceeds across the Indian Ocean to Calicut, a port city on the Malabar Coast of India.

ca. 1498 The Inca empire of western South America establishes the Andes highway, a vast system of roads built of stone bricks, reportedly extending over more than 15,000 miles in total.

16th Century

ca. 1500 By this date, direct use of geothermal water for medical purposes is common practice at Xiaotangshan Sanitarium, northwest of Bejing. The 50°C water is still used today for the treatment of high blood pressure, rheumatism, and skin diseases, and for recuperation after surgery.

1502 Italian navigator Amerigo Vespucci asserts that the land discovered by Columbus is a "new world," completely separate from Asia. In 1507 German mapmaker Martin Waldseemuller publishes a map of the new continent discovered by Columbus and calls it "America" after Amerigo Vespucci.

1519 Ferdinand Magellan of Portugal sets sail from Spain on the first voyage around the world. Although he is killed in the Philippines midway through the journey, Magellan's voyage provides information on the size of the Earth and the independent structure of continents.

ca. 1520 The Saxon spinning wheel is introduced (named for its supposed origin in Saxony). This improves on the existing European spinning wheel by featuring a foot pedal, leaving both hands free to manipulate the fibers.

1536–1539 King Henry VIII rejects the authority of the Pope and dissolves the monasteries of England. Church lands thus pass into public hands, most significantly the coal mines of Newcastle and other areas, which can now be exploited by local merchants.

1543 Shortly before his death, Polish astronomer Copernicus publishes *De Revolutionibus Orbium Coelestium* (On the Revolutions of Celestial Bodies). He presents a heliocentric model of the solar system; i.e., the Earth and other planets orbit the Sun. This contradicts the prevailing geocentric (Earth-centered) system developed earlier by Ptolemy.

ca. 1550 King Gustaf I Vasa of Sweden (ruled 1523–1560) is one of the earliest rulers to recognize the importance of energy supply to national security (forests in this case).

ca. 1550 Miners in Germany use wooden rails and pushcarts with flanged wheels to move ore along a track from the mines. A simpler rail system was in use in Slovakia as early as 1513.

mid 1500s The first known production of hydrogen is made by the Swiss physician and alchemist Paracelsus,

probably through the combination of sulfuric acid and iron. Paracelsus notes that this reaction produces a gas or "air," and that the air is released with great pressure.

1551 Jerome Cardan (Girolamo Cardano), an Italian mathematician, distinguishes electricity from magnetism by observing that while amber will attract various light objects, a lodestone (magnetic stone) will attract only metals.

1560 The Swedish Baltic Empire dominates Northern Europe from this time until 1721, based in part on its rise as a producer and exporter of metals, while England's and mid-Europe's metal production declines due to a drop in charcoal production caused by deforestation and wood scarcity.

1561 Adam Lonicier of Germany writes of a solar heating technique used by alchemists to make perfume, in which flowers are heated by light reflected from a mirror to cause their essence to diffuse into water.

1568 Gerardus Mercator introduces cylindrical projection in maps, now known as Mercator projection. The cover of his book of maps shows the Greek Titan Atlas balancing the world on his shoulders, a picture which gave atlases their name.

1570–1630 England experiences a widespread shortage of firewood, especially in areas with ironworks, since large amounts of fuel are required for smelting iron. This depletion of the wood supply leads to a shift to coal as the chief energy source for smelting and other industrial purposes. As a result England becomes the first country to develop a full-scale coal industry.

late 1500s Oil from seeps in the Carpathian Mountains is burned in street lamps to provide light for the Polish town of Krosno.

1583 Mathematician Simon Stevin demonstrates the laws of hydrostatics, showing that the surface pressure of a liquid is related to the height of the liquid above the surface and the area of the surface, but not to the shape of the container holding the liquid.

1586 Stevin is said to have performed a key gravitational experiment, noting that weight does not affect the pull of gravity. (It is also said that Galileo reached this same conclusion after dropping two different weights from the top of the Leaning Tower of Pisa.)

1589 The Reverend William Lee of England invents the first knitting machine, which can knit faster than human hands, using a stocking frame.

1590 Dutch spectacle-maker Zacharias Janssen constructs the first microscope, using two convex lenses and a tube.

1592 Galileo develops the first primitive thermometer, known as a thermoscope, using air in a tube instead of liquid.

1592 The first wind-driven sawmill is built in the Netherlands by Cornelis Cornelisz. It is mounted on a raft to permit easy turning into the wind.

1594 An oil well 35 meters deep is reportedly dug in Azerbaijan.

1597 Andreas Libavius publishes *Alchemia,* the first European chemical textbook, in which he describes the preparation of hydrochloric acid, sulfuric acid, and tin tetrachloride.

1600–1649

1600 William Gilbert presents a model of the solar system based on magnetism, explaining gravity through electromagnetic terms and noting magnetic dip in a compass.

early 1600s Giovanfrancesco Sagredo, a disciple of Galileo, divides his thermometer into 360 divisions like a circle, after which the divisions on a thermometer are called "degrees."

1603 Hugh Platt describes coke, a charcoal-like residue left after the volatile components of coal have been driven off. This later becomes an important fuel. Ironically Platt at the same time complains about the damage being caused to London's buildings and vegetation by the burning of coal, the process that produces coke.

1603 Italian alchemist Vincenzo Cascariolo makes the first scientific investigation into the property of luminescence when he finds that a heated mixture of barium sulfate and coal glows in the dark and can be "recharged" through exposure to sunlight.

1603–1604 Huntingdon Beaumont of England constructs a wooden wagonway from the bell pits at Strelley down to Wollaton Lane, to carry coal (horse-drawn) to the Nottingham market.

1604 Robert Cawdrey publishes a dictionary of hard words; this is the first alphabetical reference work in English.

1606 Italian Giambattista della Porta designs a device capable of raising water, an improved version of "Hero's Fountain," with steam instead of air as the displacing fluid. He also describes the formation of a vacuum by condensation of steam.

1608 Hans Lippershey of the Netherlands builds the first practical telescope.

1609 German astronomer Johannes Kepler formulates his first two laws of planetary motion: the elliptical orbits of the planets and the alteration of planetary speed according to distance from the Sun.

1609–1611 Galileo improves the telescope and pioneers its use for astronomical observation, discovering mountains on the moon, many new stars, the four satellites of Jupiter, the phases of Venus, and the composition of the Milky Way.

1610 The English coal industry has grown to the point that nearly half of all the country's maritime trade consists of coal exports. The city of Newcastle upon Tyne in the northeast part of the country is the hub of this industry (giving rise to the expression "carrying coals to Newcastle" for an activity that is a waste of time because it supplies something which is already present in abundance).

1612 Jan Adriaanszoon Leeghwater (literally, "empty water") directs the drainage of Beemster and other polders (lowlands), beginning a long period of wind-driven polder drainage in the Netherlands.

1615 A royal proclamation in Britain prohibits the use of wood for glassmaking. Wood is in high demand for the crucial shipbuilding industry, so glass-making furnaces are changed to burn coal instead.

1615 French engineer Salomon de Caux (Caus) builds a solar-powered water pump, using glass lenses to heat a sealed vessel containing water and air. This is regarded as the first use of solar energy since classical times. He also describes the principles of a steam engine and advocates its use.

1619 Kepler formulates his third law of planetary motion, involving a calculation to determine the distance of a planet from the Sun. He also publishes a defense of the Copernican system.

1620 In his influential *Novum Organum Scientiarum,* Francis Bacon advocates observation and the inductive method over postulation and *a priori* reasoning, describing what is known today as the scientific method.

ca. 1620 Dutch inventor Cornelius Drebbel creates a thermostat which he uses to regulate the temperature of an oven by controlling the flow of air.

1621 The first ironworks in colonial North America is set up at Jamestown, Virginia.

1628 British physician William Harvey publishes *On the Motion of the Heart and Blood in Animals*, in which he describes the circulation of the blood and the pumping of the heart; this is often described as the foundation of modern medicine.

1629 Italian engineer Giovanni Branca invents a stamping mill whose power is generated by a steam-powered turbine.

ca. 1630 Jan Baptista van Helmont of Brussels coins the word *gas,* from the Flemish word for *chaos,* to describe what he calls the "wild spirit" component of oaken coal that has been burned, i.e., that does not remain as ashes. This is the first known description of an air-like substance, without fixed volume or shape, other than air itself.

1635 John Winthrop opens America's first chemical plant in Boston, producing saltpeter (potassium nitrate) for use in the manufacture of gunpowder.

1637 In *Discours de la Methode,* Rene Descartes champions deductive reasoning on the basis of mathematical laws; famous appendices to this influential work present important advances in physics, optics, meteorology, and mathematics.

1638 Galileo publishes his major work on mechanics, including experiments on acceleration, friction, inertia, and falling bodies that correct many of Aristotle's errors.

1640 Italian physicist Evangelista Torricelli applies Galileo's laws of motion of fluids, establishing the science of hydrodynamics. In 1643 he invents the mercury barometer and in the process makes an artificially created vacuum, the first known instance of this.

1641 Descartes argues in his *Principles of Philosophy* that the Universe is governed by simple laws and that natural processes might have shaped the Earth.

1645 German engineer Otto von Guericke (Gericke) demonstrates the effect of atmospheric pressure on evacuated bodies when he builds the first vacuum pump and takes the first measurement of the density of air.

1646 English physician Sir Thomas Browne publishes *Pseudodoxia Epidemica,* in which he discusses his experiments with static electricity and magnetism and introduces the words *electric* and *electricity* into the English language.

1648 From an experiment using barometers at different levels on a mountain, French mathematician Blaise Pascal establishes that air pressure decreases with altitude, leading him to conclude that the atmosphere has only a finite height. He also formulates the principle that pressure is uniform throughout a static fluid.

1650–1699

ca. 1650 The longwall coal mining technique is introduced in Shropshire, England. This becomes the most widely used method of coal extraction.

1652 The government of New Amsterdam (New York City) establishes the first traffic laws in the Americas, ruling that horse-drawn wagons, carts, and sleighs are not to be driven at a gallop in the city.

1654 Ferdinand II of Italy develops the first sealed liquid thermometer.

1654 Otto von Guericke begins a series of public demonstrations of the power of air pressure in a vacuum, including experiments in which teams of horses are unable to separate a sphere held together by a vacuum.

1656 Dutch scientist Christiaan Huygens builds the first pendulum clock based on the cycloid, much like a modern grandfather clock. It is the first clock to keep highly accurate time.

1659 Thomas Shirley describes an inflammable gas seeping from coal deposits near Wigan, England. This is the first known discovery of natural gas in Europe.

1661 John Evelyn publishes "Fumifugium, or the Inconvenience of the Aer and Smoake of London Dissipated," one of the earliest tracts on air pollution. He describes the poor air quality of London and suggests remedies to deal with this, such as relocating certain industries (e.g., ironmaking, brewing) outside the city, and then planting gardens and orchards in their place.

1662 Anglo-Irish chemist Robert Boyle discovers that air can be compressed and formulates Boyle's law, which states that in an ideal gas of constant temperature, pressure and volume vary inversely.

1662 John Graunt compiles a study of London death records, in what is considered the first statistically based public health study. He notes that roughly one out of every four deaths results from lung-related diseases. This contrasts with lower death rates in rural areas, away from London's smoke-filled air.

1662 King Charles II of England grants a charter to the Royal Society of London, an organization formed to encourage scientific research and communication.

1663 Guericke develops the first electric generator, which consists of a revolving ball of sulfur inside a glass globe, mounted on a wooden frame. This produces a static electric spark when a pad is held against the ball as it rotates.

1665 Denis de Sallo of Paris publishes the initial volume of the *Journal des Scavans,* the first Western scholarly journal.

1665–1666 Isaac Newton conducts a series of experiments, discovering that white light is composed of different colors that can be separated by a prism, formulating his first law of universal gravitation, and inventing the first form of calculus.

1666 The Great Fire of London sweeps through this metropolis of wooden buildings, destroying virtually all of the city north of the Thames River. The fire does have one positive effect in that it "sanitizes" areas that had just been devastated by an outbreak of plague, and England never again has a major epidemic of the disease.

1668 John Wallis suggests the law of conservation of momentum, according to which the total momentum of a closed system does not change.

1669 In his quest to create gold from another material, German chemist Hennig Brand identifies a faintly glowing element that he names "phosphorus." The discovery of this element is the first attributable to a specific person and time.

1669 Nicolaus Steno of Denmark lays the foundation for the science of crystallography with his statement that the angles are constant between equivalent faces of crystals of the same substance, when measured at the same temperature.

1671 Hydrogen is isolated by Robert Boyle. Earlier in the century, hydrogen was discovered to be flammable by Turquet de Mayerne.

1671 Wagonways from Ravensworth Colliery deliver coal to the Team-staith at Dunston, England.

1672 Giovanni Cassini calculates the distance from the Earth to the Sun to be 87 million miles, within 6 million miles of modern estimates. Scientists begin to realize just how large the solar system must be.

1672 Guericke recognizes that the static electricity produced by his sulfur ball device (see 1663) will cause the surface of the sulfur to glow, thus identifying the phenomenon of electroluminescence. Guericke also notes that like charges repel each other.

1673 Huygens produces what is regarded as the first heat engine, although the lack of a reliable fuel hampers its development.

1673 The first record of coal in the United States is shown in a map prepared by the French explorer Louis Joliet. The map shows "*charbon de terra*" seen by Joliet and Jacques Marquette along the Illinois River, near present-day Utica, Illinois.

1675 Gottfried Leibniz of Saxony (now Germany), working independently of Newton, develops differential and integral calculus. His notations, including the integral sign and derivative notation, are still in use today.

1675 In the course of his study of Jupiter's moons, the Danish mathematician Olaus Roemer estimates the speed of light to be about 141,000 miles per second. Before Roemer, many scientists believed that light had infinite speed.

1676–1677 Looking at pond water through a microscope of his own construction, Anton van Leeuwenhoek is the first to view what he called "animalcules," known today as microorganisms.

1678 A patent application is made by Clement Clerke of Bristol, England for a lead smelting reverberatory furnace using pit coal.

1678 Abbé Hautefeuille of Orleans, France is one of the first to propose the use of explosive powder for motive energy when he suggests the construction of a powder motor to raise water.

1679 French engineer Denis Papin builds the first atmospheric engine called the "Digester" (pressure cooker).

1680 Boyle discovers the match by rubbing phosphorus and sulfur together.

1680 Newton proposes that a jet of steam can be used to power a carriage, an idea now considered to be a precursor to the development of the jet engine.

ca. 1684 The experiments of English minister John Clayton of Yorkshire, an amateur chemist, reveal the presence of a clear, flammable "spirit of coal" as a byproduct of distillation, though he does not recognize a practical use for this coal gas.

1686 English astronomer Edmond Halley shows that low latitudes receive more solar radiation than higher ones and proposes that this gradient provides forcing for the atmosphere's general circulation. For this, and his work on tides, Halley is considered a founder of geophysics (though he is better known for his study of what is now called Halley's Comet).

1686 Leibniz posits an early version of the law of conservation of energy. He argues that the true measure of the effect of force is *vis viva* or "living force" as exhibited in motion and not "dead force" as exhibited in statics.

1687 Isaac Newton publishes *Philosophiae Naturalis Principia Mathematica,* often regarded as the most important book of science ever written. He establishes three laws of motion, the law of universal gravitation, and other fundamental principles. Newton thus provides the basic description of the natural laws governing the universe (later modified by the work of Einstein).

1688 The Royal Society of London officially notes that gas can be extracted from coal and oil, and then burned as a source of heat and light.

1690 Denis Papin develops the idea of using steam to pump water out of mines. He also describes an atmospheric steam engine operating with piston and cylinder.

1690 William Rittenhouse establishes the first paper mill in the American colonies, on a creek near Philadelphia. This has been described as America's first example of recycling, since the mill's fiber for papermaking was obtained from discarded rags and cotton.

1695 Italian academicians Averoni and Targioni conduct experiments on the combustibility of diamonds, including the use of solar heat by means of a large burning-glass.

1698 Captain Thomas Savery, an English military engineer and inventor, receives the first patent for a crude steam engine, based on Papin's Digester of 1679. Savery is motivated by the challenge of pumping water out of coal mines. He names his engine "The Miner's Friend."

1700–1749

early 1700s Settlers in New England utilize solar energy by "saltbox" houses that face toward the winter sun. Many feature a lattice overhang on the south facade, supporting deciduous vines that afford shade in summer but drop their leaves in winter, allowing sunlight through.

1701 Coal deposits are found by Huguenot settlers in Manakin on the James River, near present-day Richmond.

1701 English agriculturist Jethro Tull invents an efficient mechanical seed drill, which allows seed to be sown in evenly in neatly spaced rows.

1702 Czar Peter the First establishes Russia's first regular newspaper, *Vedomosti.* The paper's first issue carries a story about the discovery of oil on the surface of the river Sok in central Russia.

1705 Neon light is born when Francis Hauksbee of England creates electrical effects by putting mercury into a glass globe, pumping out the air, and then turning it with a crank. When he does this in the dark, and then rubs the globe with his bare hand, it glows.

1709 Abraham Darby, iron-maker at Coalbrookedale in the West Midlands of England, develops the use of coke in iron smelting, enabling factories to produce iron at a much faster rate. Before that, iron furnaces could only be located at a site with both sufficient water power and an extensive wood supply, to provide the fuel source for charcoal.

1709 Sir Isaac Newton builds an electric generator consisting of a rotating glass sphere.

1711–1712 English engineer Thomas Newcomen devises a new kind of steam engine with a piston and cylinder construction that utilizes ordinary low-pressure steam to pump water out of coal mines. This is the culmination of efforts he began in 1705 and it becomes the first successful atmospheric steam engine.

1714 Gabriel Fahrenheit develops a mercury thermometer and the Fahrenheit scale of measurement based on the temperature of ice, water, and ammonium chloride (0 degrees), ice and water (32 degrees), and boiling water (212 degrees).

1716 Swedish engineer Martin Triewald uses a hot water system to heat a greenhouse in Newcastle upon Tyne, in

northern England. This is considered the first modern instance of central heating.

1723 Jacob Leupold of Germany publishes a nine-volume work that is the first detailed and systematic exposition of mechanical engineering, including a design for a noncondensing, high-pressure steam engine similar to the ones that would be built much later.

1728 James Bradley notes the periodic shift of the stars' positions during a year and explains it by the aberration of light caused by the Earth's orbit. He uses his findings to recalculate the speed of light, which he places within 5% of what it is now stated to be.

1729 Through a series of experiments, Stephen Gray establishes that some materials conduct electricity better than others and that static electricity travels on the surface of objects rather than through the interior. This is considered the first discovery of the principle that electricity can flow.

1733 Charles-Francois de Cisternay du Fay states a "two-fluid" theory of electricity, dividing electrical charges into vitreous (positive) fluid and resinous (negative) fluid, and noting that like charges attract and differing charges repel. Benjamin Franklin also distinguishes between positive and negative charges, though he favors a one-fluid theory.

1733 Englishman John Kay patents the flying shuttle loom, enabling one person to do much more work with relatively less effort.

1735 George Hadley, an English meteorologist, introduces the importance of Earth's rotation in the features of major winds, including the easterly trades and the midlatitude westerlies.

1735 Swedish naturalist Carolus Linnaeus publishes *Systema Naturae* (Systems of Nature), introducing a method of classification for plants still in use today, which he later extends to include animals.

1736 Swiss scientist Leonhard Euler writes the first book of mechanics based on differential equations.

1738 Swiss scientist Daniel Bernoulli develops his kinetic theory of gases, according to which pressure within a closed container of gas increases as temperature rises and decreases as temperature goes down.

1742 Anders Celsius suggests replacing the Fahrenheit scale with the Centigrade scale, which places the freezing point of water at 0 degrees and the boiling point at 100 degrees. In 1948 the name of this scale is changed to Celsius scale by international agreement.

1743 Jean le Rond d'Alembert expands on Newton's laws of motion, proposing that actions and reactions in a closed system of moving bodies are balanced. He applies this principle to several mechanical problems.

1743 King Louis XV of France has a "flying chair" installed in his palace to allow him to ride from the first to the second floor; this is considered the earliest form of passenger elevator.

1745 Edmund Lee introduces the fantail mechanism for windmills; it automatically rotates the blades of the windmill into the wind, as opposed to having a person manually do this.

1745 Shafts are dug into the hillsides of Pechelbronn, France to extract petroleum; these are the first known wells specifically sunk to search for oil.

1745–1746 Pieter van Musschenbroek and Ewald Georg von Kleist independently invent the capacitor, an electricity-storage device that is the first practical way to store static electricity. It comes to be called the Leiden (Leyden) jar after Musschenbroek's employer, the University of Leiden.

ca. 1746 Euler determines the mathematics of light refraction, based on Huygens' theory of 1690 that light is composed of different colors of varying wavelengths.

1747 Benjamin Franklin describes his discovery that a pointed conductor can draw off electric charge from a charge body. This is the basis for his lightning rod, even before he proves the connection between electricity and lightning.

1748 The first commercial coal production in colonial America is recorded near Richmond, Virginia. By the end of the 1750s other coal deposits are reported in Pennsylvania, Ohio, Kentucky, and West Virginia.

1749 Benjamin Franklin installs a lightning rod on his home.

1749 George-Louis Leclerc de Buffon publishes the first of his 44 volumes of *Historie Naturelle* (Natural History), in which he anticipates Charles Darwin's theory of evolution by presenting arguments for descent with modification and species extinctions. He also suggests that existing estimates of the age of the Earth are much too low.

1750–1775

1750 A French military officer notes that Indians living near Fort Duquesne (now the site of the city of Pittsburgh) set fire to an oil-slicked creek as part of a religious ceremony.

mid-1700s An extensive system of canals is established in various countries of Western Europe. At the same time the quality of surface roads in these countries is generally poor.

mid-1700s England and Holland make advances in greenhouse technology.

1752 Benjamin Franklin conducts his famous experiment with a kite in a lightning storm, showing that lightning is a form of electricity.

1754 Joseph Black discovers "fixed air," or carbon dioxide, when he heats limestone. He determines that carbonates are compounds of a gas and a base.

1755 The first known instance of artificial refrigeration is demonstrated by William Cullen at the University of Glasgow.

1757 Captain John Campbell of the Royal Navy develops a nautical sextant for precise measurements of longitude and latitude.

1758 Coal is "exported" from the James River district of the colony of Virginia to New York.

1758 Swedish chemist Axel Fredrik Cronstedt divides minerals into four classes: earths, bitumens, salts, and metals. This establishes the classification of minerals by chemical structure.

1759 English civil engineer John Smeaton determines that the overshot waterwheel is much more efficient than the traditional undershot wheel. He also makes advances in the design of many other devices and structures, e.g., an improvement on the Newcomen steam engine.

1759 Franz Aepinus rejects mechanical theories of electricity and suggests that ordinary matter repels itself in the absence of electricity, a theory of electrostatics similar to Newton's law of gravity and supporting Benjamin Franklin's one-fluid theory of electricity.

1760 John Mitchell proposes that earthquakes are caused by the build-up of pressure within the earth and that they produce waves which travel at measurable speeds.

1761 Joseph Black discovers that ice absorbs heat without changing temperature when melting. He calls this property "latent heat," a quality later important in the improvement of the steam engine.

1762 An act of the New York City assembly gives the authority to provide funds for lighting the city, and in that year the first street lamps and posts are purchased.

1762 Deposits of anthracite coal are discovered in Pennsylvania.

1764 James Hargreaves introduces a new device, called a "spinning Jenny," which can spin eight threads at once and thus provide much greater efficiency than the traditional spinning wheel.

ca. 1765 Burmese excavators hand dig hundreds of oil wells.

1766 Henry Cavendish identifies and outlines the properties of a highly flammable substance which he calls "fire air," now known as hydrogen.

1766 Swiss physicist Horace-Benedict de Saussure develops an electrometer, a device for measuring electric potential by means of the attraction or repulsion of charged bodies. He also constructs the first known solar flat plate collector.

1766 The brothers Thomas and George Cranege receive a patent for a reverberatory or air furnace, for use in iron-making. This is later improved upon by Peter Onions (1783) and Henry Cort (1784).

ca. 1766 The opulent Paris townhouse of François de Monville features a central heating system; this is the first time that buildings in France have had central heat since the era of the ancient Romans.

1767 Joseph Priestley offers an explanation for rings formed by an electrical charge on metal (now known as Priestley's rings) and proposes that electrical forces follow the same inverse square law that gravity does.

1769 French army engineer Nicholas-Joseph Cugnot builds a three-wheeled wagon powered by a steam engine, the first example of a "horseless carriage." However, it can only operate for a few minutes without stopping to get up steam, and it is soon retired after overturning in what is described as the world's first motor vehicle accident.

1769 Richard Arkwright invents a spinning frame that produces cotton threads for textile manufacture, one of the key developments leading to Britain's Industrial Revolution.

1769 Scottish inventor James Watt receives the patent for an improved version of the Newcomen steam engine. It uses a condenser, separate from the cylinder, to improve efficiency. A unit of power called the *watt* is eventually named after him.

ca. 1769 Spanish missionary Juan Crespi writes of "extensive swamps of bitumen" that he has seen while traveling on an expedition to the region that is now the city of Los Angeles, California. These swamps are the La Brea Tar Pits, and more than 200 years later the oil industry of Southern California will begin in this same area.

1770 George Washington observes a "burning spring" on the Kanawha River in West Virginia.

1771 In experiments with plants in carbon dioxide-filled chambers, Joseph Priestley discovers that plants somehow convert the gas breathed out by animals and created by fire back into oxygen; this lays the foundation for an understanding of photosynthesis.

1772 Daniel Rutherford, Joseph Priestley, Carl Scheele, and Henry Cavendish independently discover nitrogen, although Rutherford is usually given credit for being the first.

1774 Priestley publishes his discovery of oxygen. Scheele had previously discovered it in 1772 but did not publish his findings until 1777, and Antoine Lavoisier actually named the element. The identification of oxygen overturns an idea that had dominated science for more than

2,000 years, namely that air is a singular, elementary substance.

ca. 1775 Lavoisier builds a solar furnace that can generate enough heat to melt platinum.

1775 James Watt patents a new steam engine that he had developed in 1765, which is six times as efficient as the old Newcomen engine. Its condenser is separated from the cylinder, allowing the steam to act directly on the piston.

1775 Pierre-Simon Girard invents a water turbine.

1775 The first chemical extraction from geothermal waters (mainly borate products) is undertaken in Lardarello, Italy.

1776–1799

1776 *An Inquiry into the Nature and Causes of the Wealth of Nations* is published by Scottish philosopher Adam Smith. It establishes economics as a distinct field of study and provides the basis for laissez-faire economics, in particular the idea that "the invisible hand" of the free market, and not government, is the proper controlling force for a nation's use of resources.

1776 The American colonies declare independence from Britain. From this time the industrialization of the United States will proceed separately from that of Britain; prior to this one-third of all British ships were being built in America and the colonies were also a center for British iron manufacturing.

1776 The Watt steam engine is put into practical use, to drain water from coal mines.

1778 Scheele and Lavoisier independently conclude that air is composed mainly of oxygen and nitrogen.

ca. 1778 The first industrial blast furnace is created by ironmaster John Wilkinson, using a steam engine rather than water power.

late 1770s Factories are sited along rivers to utilize the power of waterwheels.

1779 Dutch scientist Jan Ingen-Housz identifies principles of photosynthesis when he notes that plants convert carbon dioxide into oxygen, which animals convert back into carbon dioxide. He also discovers that only the green parts of plants release oxygen, and that oxygen is only given off when plants are in the presence of light.

1779 The earliest versions of the modern bicycle appear on the streets of Paris.

1780 During experiments with oxidation of organic gases, Johann Wolfgang Döberreiner discovers that there are certain similarities of reaction among elements. This forms the basis of the periodic table.

ca. 1780 Luigi Galvani of Italy discovers that when he touches a dead frog's leg with a knife, it twitches violently. Alessandro Volta will later show this is because electricity is created when moisture (from the frog) comes between two different types of metal (the steel knife and a tin plate); this phenomenon will become known as *galvanism*.

ca. 1780 M. Bonnemain of France creates a hot water boiler that uses a gravity system to circulate the heated water.

1780s Watt continues to adapt his steam engine invention. He creates a mechanical attachment with a rotary movement, enabling the engine to power a wider variety of machines, and he develops the first governor and flywheel. Flywheels will later be the focus of extensive research as power storage devices.

1781 Priestley "creates" water by igniting hydrogen in oxygen.

1781 Richard Arkwright builds the first modern factory to house his water frame for spinning.

1783 Joseph and Etienne Montgolfier launch the first hot air balloon, the earliest successful air flight in history. The second flight carries a cargo of animals, and the third flight carries two men. The same year, Jacques Charles creates and flies in the first hydrogen balloon.

1783 Lavoisier, with Pierre Laplace, uses a calorimeter to estimate the amount of heat evolved per unit of carbon dioxide produced. The ratio is the same for a flame and an animal, indicating that animals produce energy by a type of combustion, and also demonstrating the role of oxygen in the respiration of animals and plants.

1783 William Herschel explains the apparent movement of stars together and apart from each other by arguing that the Sun itself is moving toward some stars and away from others, changing their relative positions.

1784 Benjamin Franklin notes that the substitution of coal for wood as a fuel has helped to preserve England's remaining forests, and he urges France and Germany to do the same.

1784 Franklin suggests that a great eruption of the Laki volcano in Iceland is responsible for the unusually severe winter that follows, thus showing an awareness of the effect of volcanism on global climate.

1784 For the first time a flour mill in England successfully substitutes a steam engine for wind power.

1784 Ironmaster Henry Cort develops both the puddling method of manufacturing wrought iron from coke-smelted iron and the rolling mill with grooved rollers.

1784 Jean Pierre Minkelers describes the production of gas from coal, and uses such gas to illuminate his

classroom at Leuven, Belgium. This is sometimes considered the origin of gas lighting.

1784 Swiss scientist Aimé Argand patents a new type of lamp that burns much brighter than candles, produces less smoke, and reduces the danger of fire. It features an oil receptacle, a round wick lamp with a chimney, and a glass cylinder in which the flame burns. This becomes the prototype for many subsequent oil lamps.

1785 French physicist Charles-Augustin de Coulomb is the first to measure accurately the forces exerted between electric charges.

1785 Henry Cavendish establishes that water is composed of hydrogen and oxygen.

1785 The theory of uniformitarianism, or the very gradual and regular evolution of the earth's structure, is introduced by James Hutton. According to his ideas, the earth is so old that it is almost impossible to determine when it began.

1787 Jacques Charles demonstrates that different gases expand at the same rate for a given temperature. This principle, known as Charles's Law, had actually been discovered by Guillaume Amontons as early as 1699.

1787 John Fitch builds and launches the first steamboat on the Delaware River. Although a technological success, it is a commercial failure.

1787 John Wilkinson launches an iron boat, believed to be the world's first such vessel for commercial use. Reportedly it is intended as a canal barge for freight.

1788 The fur trader Peter Pond is the first European to describe the vast oil sands (tar sands) of Atahabasca, Canada. He writes of "bituminous fountains" and describes how the local Indians mix this fluid with gum from fir trees to waterproof their canoes.

1789 French engineer Philippe Lebon distills gas from heated wood and suggests that it can heat and illuminate interiors and inflate balloons. This is an important contribution to the subsequent development of the manufactured gas industry.

1789 German chemist Martin Heinrich Klaproth discovers uranium in the mineral pitchblende. It is not until 1841, however, that this element is isolated by Eugene Péligot.

1789 In *Traite Elementaire de Chimie* (Elementary Treatise on Chemistry), Lavoisier introduces the law of conservation of mass that he had developed some years before, and establishes a new system of chemical nomenclature.

late 1700s Spermaceti, an oil found in the head of sperm whales, is highly sought after for various industrial uses. It is especially prized for candlemaking; the term candlepower is based on a measurement of the light produced by a spermaceti candle.

1790 England builds its first steam-driven rolling mill.

1790 Samuel Slater builds the first American factory, constructed entirely from his memory of blueprints he had seen earlier in England, and ushers in the Industrial Revolution in the United States.

ca. 1790 Claude Berthollet and Joseph Priestley determine the molecular formula for ammonia.

ca. 1790 Friendly Societies, forerunners of modern trade unions, are organized in Britain.

1791 John Barber receives the first patent for a basic turbine engine, which he intends to use to power a "horseless carriage." Though he does not achieve practical success, this is the first engine to use the thermodynamic cycle of a modern gas turbine.

1792 William Murdock discovers that by heating coal in a closed container, the gas given off will yield a steady flame. Murdock produces "coal gas," or "manufactured gas" and conveys it through metal pipes, lighting his Redruth, Cornwall cottage and offices.

1793 Benjamin Thompson (Count Rumford) challenges the prevailing caloric theory that heat is an elastic fluid. From his work boring cannon barrels, he notes that the cannon stay hot as long as the friction of boring continues, and he associates heat with mechanical energy. By defining heat as a form of motion rather than matter, he anticipates the kinetic theory of heat.

1793 Eli Whitney invents the cotton gin to pull the cotton threads off the seeds in cotton bolls. The machine is so successful that it completely changes the economy of the southern United States, where cotton becomes the main crop.

1794 Inventor Philip Vaughn obtains the first patent for a modern type of ball bearing while working on carriage axles. Leonardo da Vinci is said to have developed a model for ball bearings some 300 years earlier but none were actually produced from this.

1795 Count Rumford invents the Rumford stove, an innovative version that provides greater heat with less smoke.

1795 Joseph Bramah patents a hydrostatic machine (hydraulic press), one of many inventions with which he is credited.

1796 M. Ambroise & Company, Italian fireworks manufacturers, reportedly exhibit a form of gas lighting in Philadelphia.

1796 Pierre-Simon Laplace advances his "nebular hypothesis" of the origin of the world, which states that our solar system was formed by the condensation of a nebula of dust and gas, a theory still widely accepted today.

1797 Edmund Cartwright invents an engine that runs on alcohol. He also makes significant contributions to weaving loom technology.

1798 Eli Whitney is awarded a government contract for 10,000 muskets. In manufacturing these weapons, he

develops methods to produce standardized interchangeable parts, a procedure important to the Industrial Revolution.

1798 Henry Cavendish accurately determines the mass of the Earth by determining the gravitational constant, the unknown *G* in Newton's equations.

1798 Thomas Robert Malthus publishes *An Essay on the Principle of Population,* arguing that poverty and famine are inevitable because the world's population is bound to grow at a faster rate than the food supply. This leads to many subsequent analyses of the relationship between population and available resources, in particular Darwin's ideas about natural selection.

1799 Alessandro Volta creates the first electric battery (known as Volta's pile). It consists of alternating zinc and silver disks separated by felt soaked in brine. It is the first source of a steady electric current. Volta announces this invention in 1800.

1799 Philippe Lebon receives a patent for his Thermolamp, which is illuminated by coal gas. By 1801 he lights his own home in this manner and he also conducts a light demonstration at the Hotel Seignelay in Paris. Lebon also conceives of a gas-powered engine.

1799 The first patent for a wave energy system is granted in Paris to a father and son named Girard; they envision a system that would harness the power of heavy ships rocking back and forth on the waves in a harbor.

1800–1809

1800 Less than two months after Alessandro Volta's battery was publicized, English scientists William Nicholson and Anthony Carlisle build a battery and use it to perform electrolysis on water, separating it for the first time into hydrogen and oxygen.

1800 While investigating the heat of different colors in a spectrum, William Herschel discovers that there is an invisible light beyond the red that produces the most heat, now known as infrared radiation.

1800–1870 The level of carbon dioxide gas (CO_2) in the atmosphere, as later measured in ancient ice, is about 290 parts per million. Fossil fuel use and land clearing associated with the Industrial Revolution, population growth, and rising affluence will produce a steady rise in CO_2 concentrations over the next two centuries.

1801 Building on Herschel's discovery of radiation beyond the red end of a spectrum, Johann Ritter experiments with silver chloride and detects radiation beyond the violet end as well, now known as ultraviolet radiation.

1801 Thomas Young discovers that light passing through two narrow slits produces an interference pattern, providing further support for the wave theory of light.

1801 William Symington launches the *Charlotte Dundas* on the Forth and Clyde canal in Scotland. This is the first vessel built for the specific pusrpose of being propelled by a steam engine. Symington builds more than 30 different steam engines in the period 1790–1808.

1801–1802 Jean-Baptiste Lamarck provides the first classification scheme for invertebrates, and also presents ideas about a relationship between physical energy and morphological changes in organisms over generations.

1802 Applications of manufactured gas lighting take place in Britain and the United States; e.g., William Murdock installs gas lighting at the Watt and Boulton Birmingham works; a gas lighting system is patented in London by Frederick Winsor; Benjamin Henfrey demonstrates his "thermo-lamp" coal gas lighting method in Pennsylvania.

1803 Claude Berthollet applies physics to chemistry to show that reaction rates depend on the amounts of reacting substances as well as on affinities.

1803 John Buddle develops an air pump to provide ventilation for coal mines in Britain.

1803 John Dalton is the first to formally propose the atomic theory of matter, claiming that atoms (a word he took from Democritus's writings of ca. 440 BC) explain why chemicals combine only in specific proportions.

1803 William Henry proposes the law that will be named after him, which states that the mass of a gas dissolved in a liquid is proportional to the pressure.

1804 Cornish engineer Richard Trevithick develops the *Penydarren,* the world's first steam locomotive to run on iron rails. However, the weight of the engine proves too heavy for the rails. Shortly before this Trevithick also developed a steam-powered road carriage to carry passengers.

1804 Sir George Cayley designs a glider that is regarded as the first successful heavier-than-air vehicle, one of several such vehicles he develops. In a long career (1799–1850s) studying the principles of aviation, Cayley describes many of the concepts and elements of the modern airplane.

1805 British manufacturer George Medhurst receives a patent to use compressed air to drive a motor.

1805 Commander William Beaufort develops a wind force scale consisting of 13 degrees of wind strength, from calm to hurricane. He bases this on the observed effect of various wind strengths on the sails of a British frigate.

1805 Frederick Tudor and Nathaniel Wyeth of Cambridge, Massachusetts devise new methods to cut, store,

and transport ice blocks from nearby Fresh Pond. They ship the ice as far as India and Cuba, with production reaching 65,000 tons by 1846.

1806 David and Joseph Ruffner drill a well on the banks of the Great Kanawha River within the present city of Charleston, West Virginia. The well is drilled in search of brine, which can be evaporated to make salt. This is the first well known to have been drilled, rather than dug, in the Western Hemisphere.

1806 David Melville lights his house in Newport, Rhode Island using coal gas. In 1810 he receives the first U.S. patent for this.

1807 Great Britain, the principal slave-trading nation of the time, abolishes the Atlantic slave trade. The United States enacts similar legislation in the same year and other nations follow suit. (Denmark had already done so, in 1803.)

1807 Isaac de Rivaz of Switzerland designs an an internal combustion engine, believed to be the first of its kind. He subsequently uses this to build a gas-powered car, using a mixture of hydrogen and oxygen for fuel.

1807 Sir Humphrey Davy of Britain isolates a new metal, sodium, using electrolysis.

1807 The *Clermont,* a 133-foot-long steam-powered paddleboat built by Robert Fulton, makes its maiden voyage up the Hudson River in New York. It proves to be a great commercial success.

1807 The resort city of Hot Springs, Arkansas is founded, and shortly thereafter visitors begin arriving to enjoy the heated waters. The area is made a national park in 1921.

1807 Thomas Young begins to use the term *energy* in the modern sense. He states, "The term *energy* may be applied, with great propriety, to the product of the mass or weight of a body, into the square of the number expressing its velocity." This refers to what is now known as kinetic energy.

1808 French chemist Joseph Louis Gay-Lussac publishes his discovery that gases combine in definite proportions, just as elements do, when forming compounds.

1808 Judge Jesse Fell burns anthracite coal on an open grate in his Wilkes-Barre, Pennsylvania home. At the time most people thought that this type of coal was of little value.

1809 Chilean nitrates are discovered. These will serve as the chief source of nitrogen for explosives and for the chemical industry over the next 100 years.

1809 Humphrey Davy uses a high-powered battery to induce a bright light between two strips of charcoal 10 centimeters (4 in.) apart, creating the first arc light.

1809 Lamarck states that animals evolved from simpler forms and that unused body parts will degenerate or new parts develop as needed for survival. Though scientists now reject Lamarck's idea that acquired characteristics can be inherited, he was credited by Darwin as a major influence on his own theory of evolution.

early 1800s A new era in illumination has begun, with gas producing a much higher quality flame than the oil lamp, the dominant lighting technology for many previous centuries. However, the new technology leaves vast pools of toxic coal tar as a byproduct in many European and American towns, a serious public health legacy.

1810–1819

1810 The first reported U.S. coal mine explosion occurs in Virginia.

1811 Amedeo Avogadro suggests that at a given temperature and pressure, equal volumes of any gas will always contain equal numbers of molecules. This comes to be known as Avogadro's Hypothesis.

1812 German promoter Frederick Albert Winsor receives a charter from Parliament to establish the London and Westminster Gas Light and Coke Company. This is the origin of public gas lighting in Britain. The following year, Winsor lights Westminster (London) Bridge.

1813–1816 The first miner's safety lamps are developed independently by Sir Humphrey Davy, William Clanny, and George Stephenson. They reduce the risk of igniting the gas mixtures that are common in mines. Davy's design is particularly effective in that it prevents the lamp temperature from rising above the ignition point of mine gases such as methane.

1814 George Stephenson builds the *Blucher* (or *Blutcher*), the first practical steam locomotive in England. It is capable of pulling 30 tons up a grade at 4 miles per hour, a vast improvement over horse-drawn wagons.

1815 Samuel Clegg, chief engineer for Frederick Winsor's gas company, develops an efficient gas meter and also installs a gas works at the Royal Mint.

1815 From his work digging canals, William Smith discovers that each layer (stratum) of earth contains a unique assemblage of fossils, and that he can distinguish one formation from another by the characteristic fossils in each. He then publishes what his biographer Simon Winchester called "the map that changed the world," a geological survey of England and Wales.

1816 Karl von Drais of Germany builds a steerable two-wheeled vehicle that is propelled by pushing with the feet against the ground. In the same year his country-

man Karl D. Sauerbronn develops the first prototype for the modern bicycle.

1816 Scottish minister Robert Stirling receives a patent for a new type of engine, the forerunner to the modern Stirling engine. This features the "Heat Economiser" (a refrigerator) and represents an advance in safety. Unlike existing steam engines, Stirling's engine would not explode because it operated at lower pressures.

1816 The artist Rembrandt Peale receives permission from the city of Baltimore, Maryland to supply the city with coal gas for street lights. Peale and his associates then charter the Gas Light Company of Baltimore, the first energy utility in the United States.

1818 The first commercial oil well in North America is drilled by Marcus Huling on land leased from Martin Beatty in McCreary County, Kentucky. The well is drilled in search of brine, but when drilled to about 200 feet, it produces large quantities of oil.

1819 English chemist John Kidd demonstrates that distilled coal tar can be used to produce naphthalene, an industrial chemical with a variety of uses, thus beginning the use of coal as a source of chemicals.

1819 Frederick Christian Accum publishes *A Description of the Process of Manufacture of Coal Gas,* which serves to bring greater recognition to the technique of gas lighting. He also publishes the first systematic study of food technology (1820).

1819 The American coastal packet *Savannah* sails across the Atlantic Ocean, becoming the first vessel to make an ocean crossing with steam-assisted power.

1820–1829

1820 Danish physicist Hans Christian Oersted is the first to note the magnetic effect of electric currents when, by accident, he observes that an electric current in a wire can deflect a nearby compass needle.

1820 Expanding on the work of Oersted, Andre Marie Ampere finds that two wires carrying current in the same direction attract each other, while wires carrying current in opposite directions repel.

1820 Johann Schweigger also expands on Oersted's experiments and invents the galvanometer to measure the strength of a current.

1820 The Rev. W. Cecil of Magadelen College, Cambridge writes a paper on how to use the energy of hydrogen to power an engine, and provides a description of how such a hydrogen engine could be built.

1820 Warren De la Rue of Guernsey encloses a platinum coil in an evacuated glass tube and passes electricity through it, in the first recorded attempt to produce an incandescent lamp.

early 1820s Gas companies in Baltimore, Boston, and New York City provide commercial gas service to light streets and buildings.

1821 French engineer Gaspard de Prony invents a brake to measure the power of engines.

1821 Michael Faraday publishes the results of his research into electromagnetic rotation, providing the basis for his efforts to create an electricity-powered motor.

1821 The first natural gas well in the United States is drilled in the village of Fredonia, New York by William Hart. By 1825, the gas is being used to light buildings in Fredonia and also reportedly for cooking in the local hotel.

1821 The Lehigh Navigation Company and the Lehigh Coal Company merge, forming a single company for the purpose of transporting coal on the canals of eastern Pennsylvania. By 1825 this new company is shipping over 30,000 tons of anthracite coal to Philadelphia.

1821 Thomas Johann Seebeck observes that if two dissimilar metals are joined with a heat difference between the juncture points, this will produce an electric current. This phenomenon is called thermoelectricity or the Seebeck effect; it will later be used in the development of the semiconductor.

1822 Friedrich Mohs sets up a scale of mineral hardness, now known as the Mohs scale, and presents a mineral classification system.

1823 French engineer Benoit Fourneyron invents the first practical water turbine. He goes on to produce various improved versions of this (1832–1855).

1823 Michael Faraday is able to liquefy chlorine by a systematic use of cold and pressure. Later he uses the same technique to liquefy other gases.

1823 Samuel Brown constructs a gas engine that runs a road carriage successfully on the streets of London and a boat on the Thames River.

1823 William Sturgeon invents the electromagnet, which is improved by Joseph Henry eight years later to the point where it can hold suspended a ton of iron.

1824 Hans Christian Oersted discovers aluminum. German chemist Friedrich Wohler then devises a method for isolating the pure form of this element (1827).

1824 Joseph Aspin, an English bricklayer, receives the first patent for Portland cement, which will become a fundamental material of industrial society as a basic ingredient in concrete and mortar. Cement production is among the most energy-intensive of material processes.

1824 Nicolas Léonard (Sadi) Carnot of France publishes a seminal analysis of the steam engine, showing that work is done as heat passes from a reservoir of higher

temperature to a lower one. He defines work and provides the basis for the second law of thermodynamics.

1825 Employing a locomotive developed by George Stephenson, the Stockton and Darlington Railway opens in northern England, operating the first public passenger train. Its original purpose is to carry coal from the collieries of West Durham to the port of Stockton.

1825 The Chatham Garden and Theatre in New York City is illuminated by gas lighting.

1825 The Cheshunt Railway is in operation in Hertfordshire, England, the first known monorail (horse-drawn). It is based on a patent obtained in 1821 by Henry Robinson Palmer.

1826 Samuel Morey of Connecticut receives a U.S. patent for his "vapor engine," which is regarded as the first successfully designed forerunner to the modern internal combustion engine.

1826 Swedish-born inventor John Ericsson constructs a heat engine that runs on hot air rather than steam.

1827 German physicist Georg Simon Ohm states that an electrical current is equal to the ratio of the voltage to the resistance (Ohm's law).

1827 Robert Brown discovers the constant motion of microscopic particles (Brownian motion). This will provide a foundation for proof of the existence of atoms.

1828 James Beaumont Neilson of Glasgow introduces the "hot blast" oven, an advance in ironmaking. This involves a stream of heated air rather than the previous technique of cold air, thus requiring a smaller amount of coal and allowing the use of lesser grades of coal.

1829 The *Rocket* locomotive of Robert Stephenson and Company wins a competition held by the new Liverpool and Manchester Railway to choose between competing engines. The basic design principles of the *Rocket* are utilized until the end of British steam locomotive building in the mid 20th century.

1830–1839

1830 A railroad boom begins in Europe and North America. The first locomotive in the United States to power a passenger train, the *Tom Thumb,* carries 26 people 13 miles over the tracks of the Baltimore and Ohio Railroad. The Liverpool to Manchester railway opens in England.

1830 Andrew Ure patents a thermostat based on the principle of differential expansion of two different metals. He also publishes a widely read dictionary in which he introduces the word *thermostat* to describe devices for controlling temperature automatically.

1830 Charles Lyell publishes the first volume of his landmark *Principles of Geology,* advancing the uniformitarian theory that becomes generally accepted, i.e., the principle that the same physical processes which can be observed today were responsible for the geological changes of the past.

early 1830s English mathematician Charles Babbage designs a mechanical calculating device he calls an "analytical engine"; this is regarded as the earliest prototype for the modern electronic computer.

1831 Charles Sauria develops the first practical friction match, using phosphorus on the tip. This improves on earlier versions of the late 1820s from John Walker and Samuel Jones.

1831 Cyrus Hall McCormick carries out the first successful demonstration of his mechanical reaper during the harvest of 1831. He further refines it and takes out a patent in 1834. (Obed Hussey had announced the construction of a mechanical reaper of his own in 1833.)

1831–33 Michael Faraday and Joseph Henry independently discover the principle of electromagnetic induction, which becomes the basis for the development of the electric generator and electric motor.

1832 Antoine-Hippolyte Pixii of France builds the first practical mechanical generator of electrical current, using concepts demonstrated by Faraday.

1832 Joseph Henry is the first to produce high-frequency electrical oscillations for communication purposes, and the first to detect them at a distance, using what becomes known as a magnetic detector. This is a key precursor to the development of radio technology.

1833 Faraday derives the laws of electrical separation of compounds (electrolysis), and suggests that atoms contain electrical charges.

1833 John Lane uses a saw blade to produce the first steel-bladed plow in the United States. This becomes known as the "prairie plow" because it is much more efficient in plowing the thick prairie sod of the American Midwest than existing wooden plows.

1833 Karl Friedrich Gauss and Wilhelm Eduard Weber collaborate to produce an electromagnetic telegraph, the signals being given by the deflection of a galvanometer needle.

1834 American inventor Jacob Perkins receives the first U.S. patent for a practical refrigerating machine; it uses ether in a vapor compression cycle. In his long career as an inventor, Perkins also makes improvements to the steam engine and hot air furnace.

1834 Benoit Paul Emile Clapeyron provides a graphical representation of Sadi Carnot's ideas on heat, which were largely unknown before Clapeyron's analysis. Clapeyron thus provides the link between Carnot's study of heat engines and Clausius's expression of the second law of thermodynamics (1850).

1834 James Bogardus of the United States patents the "gasometer," which is regarded as the first gas meter; it operates on the principle of a bellows, alternately being filled and emptied of gas, the pulsations being counted on a register.

1834 London officials bring nuisance charges against a coal gas manufacturing firm for polluting the Thames River with coal tar.

1834 Vermont blacksmith and inventor Thomas Davenport establishes the first commercially successful electric streetcar (horse-drawn streetcars are the dominant type in this era).

1834 Walter Hancock operates a regular service for the London and Paddington Steam Carriage Company, using a "steam bus" called the *Enterprise* to carry passengers between London Wall and Paddington via Islington.

mid-1830s The island city of Nantucket, Massachusetts reigns as the "Whaling Capital of the World." Whale oil is in great demand as a lighting fuel, and about 100 whaling ships from the port of Nantucket search the oceans in quest of whales, especially the sperm whale (the prototype for Herman Melville's Moby-Dick in the novel of the same name).

1835 Gaspard Gustave de Coriolis of France publishes the laws of mechanics in a rotating frame, including an apparent force (Coriolis force) exerted on moving objects by the rotation of the Earth.

ca. 1835 Robert Anderson of Scotland invents a form of electric-powered carriage.

1836 John Frederic Daniell invents the Daniell cell, the first reliable source of electric current; it represents a great improvement over the voltaic cell.

1837 Scottish inventor Robert Davidson constructs a prototype for an electric car.

ca. 1837 Astronomer Sir John Herschel (son of William Herschel) uses a solar collector box to cook food while exploring in Africa.

1838 A commercial shale oil industry is established in Autun, in central France, to produce lamp fuel.

1838 William Cooke and Charles Wheatstone put the first electric telegraph into commercial service. It uses visual signaling in which two needles at a time, out of a total of five, rotate on the receiving device to point to letters on a display.

late 1830s Steamships begin to replace sailing ships as the preferred vessel for Atlantic Ocean crossings. I. K. Brunel's steamship *Great Western* (1837) is launched at Bristol, England, the largest such ship in the world at the time and the first steamship in regular transatlantic service. This is followed by Brunel's *Great Britain* (1843), an iron-hulled vessel with a screw propeller, and *Great Eastern* (1858), the largest of all and not equaled in size for another 50 years.

1839 Charles Goodyear develops vulcanized rubber, a stabler and much more useful substance than plain rubber. He supposedly invents this after he accidentally drops rubber mixed with sulfur on a hot stove and notices that this strengthens the mixture.

1839 Kirkpatrick MacMillan of Scotland develops a new form of bicycle, identified as the first version of the two-wheeled pedal-driven bicycle that is the standard form today.

1839 Nineteen-year-old Edmund Becquerel, a French experimental physicist, discovers the photovoltaic effect while experimenting with an electrolytic cell made up of two metal electrodes.

1839 Sir William Robert Grove builds a device that will combine hydrogen and oxygen to produce electricity; this is the world's first gas battery, later renamed the fuel cell.

1840–1849

1840 G. H. Hess observes that the sum of the enthalpies of the steps of a reaction will equal the enthalpy of the overall reaction. "Hess's Law" allows the enthalpy change for an entire chemical reaction to be calculated without any direct measurements through calorimetry.

1840 German botanist Justus von Leibig formulates his Law of the Minimum, which states that if any one essential material is absent or only minimally available, then the growth and functioning of an organism is adversely affected. Later scholars will apply this concept of the limiting factor to larger contexts, e.g., ecosystems.

1840 James Prescott Joule states that heat is produced in an electrical conductor; this becomes known as "Joule's Law."

1841 The first incandescent lamp is patented in England by Frederick de Moleyns. It has a design using powdered charcoal heated between two platinum wires.

1842 In England, the Royal Commission on Employment of Children in the Mines reports "cruel slaving revolting to humanity," such as women and children chained to carts and working 15-hour days in the nation's coal

mines. The subsequent Mines Act prohibits women and boys under age 10 from working in the mines.

1842 Julius Robert von Mayer of Germany presents a numerical value for the mechanical equivalent of heat, based on a horse stirring paper pulp in a cauldron. This is the original statement of the first law of thermodynamics (law of the conservation of energy), a fundamental principle of modern physics.

1842 Sir John Bennet Lawes patents the first artificial fertilizer, "superphosphate," a combination of rock phosphate with sulfuric acid. This frees farmers from dependence on animals to produce manure to nourish crops and proves that agriculture can be a technical process, rather than an entirely natural one.

1842 Justus von Liebig is among the first to argue that body heat and muscular action are derived from the oxidation of foodstuffs. He proposes that the energy source for muscular exertion in animals and humans comes principally from protein, and not carbohydrates and fats.

1843 James Joule determines the mechanical equivalent of heat by measuring the change in temperature produced by the friction of a paddlewheel attached to a falling weight. The *joule* becomes a common unit of heat measurement. He also shows that heat is a form of energy.

1844 Englishman Robert Beart is issued the first patent for rotary well drilling, so named for the rotary table through which a drill pipe with a bit at the end is inserted and rotated. Rotary drilling rigs can drill faster, deeper, and more effectively than cable tool rigs that rely on pulverizing rock formations.

1844 Jean Foucault brings electric light to Paris, the City of Light, when he illuminates the Place de la Concorde, the largest public square in the city.

1844 Samuel F. B. Morse demonstrates his telegraph by sending the message "What hath God wrought?" from Baltimore to Washington, D.C., using a system that imprints dots and dashes on paper. Operators soon learn to read the dots and dashes directly, by listening to the click of the receiver.

1844 Stuart Perry of New York City develops an effective internal combustion engine. It is a two-cycle engine using turpentine as a fuel.

1845 Scottish inventor Robert William Thomson patents the first pneumatic (air-filled) tire of vulcanized rubber. However, this is not economically practical, and it will be more than 40 years before pneumatic tires come into use through the efforts of fellow Scot John Boyd Dunlop.

1845 William Robert Grove publishes an article describing and illustrating the principles of the modern gas-filled incandescent lamp. His actual work on this may have been done as early as 1840, thus predating the 1841 patent of Frederick de Moleyns.

1846 Wilhelm Eduard Weber develops a force law that is dependent on velocity and acceleration, which will be crucial to James Clerk Maxwell in his electromagnetic theory of light. He also invents the electrodynamometer, an instrument for measuring small currents.

ca. 1846 Italian chemist Ascanio Sobrero is the first to prepare the highly explosive compound nitroglycerine.

1847 Hermann Helmholtz of Germany publishes *On the Conservation of Force,* in which he is the first to extended James Joule's work to a general principle. He expresses the relationship among mechanics, heat, light, electricity, and magnetism by treating them all as forms of a single force, i.e., energy.

1847 The Siemens and Halske Telegraph Company is founded in Berlin with Werner von Siemens as principal. His younger brother William (Wilhelm) will establish a London branch in 1850. The company becomes a leader in telegraphy, telephony, and electrical engineering.

1847 William Bell Elliot discovers a geothermal steam field north of San Francisco which he names "The Geysers." In the 1850s the area becomes a tourist attraction for its steam baths, and in the 20th century it will become an important source of geothermal energy.

1848 Lord Kelvin (William Thomson) proposes a temperature of –273°C as the "absolute zero" point, or least possible temperature, since at that point the energy of molecules in a gas would be at a theoretical minimum. He proposes a new temperature scale (now known as the Kelvin scale) based on this principle of absolute zero.

1848 President James K. Polk turns on the first gas light at the White House, one year after James Crutchett had lit the U.S. Capitol building with gas.

1848 William Staite uses his arc lamp technology to light the portico of the National Gallery museum in London. Staite makes various improvements in arc lamp technology during the 1830s and 1840s, though the original patent is awarded to Thomas Wright (1845).

ca. 1848 A commercial oil well is drilled in Baku, Azerbaijan under the direction of a Russian engineer, F. N. Semenov. This is described as the world's first modern oil well.

ca. 1848 James B. Francis develops an improved inward-flow water turbine in Lowell, Massachusetts, one of several important steps made by Francis in the use of water power. Uriah Boyden of Lowell also improves on the existing Fourneyron turbine at around the same time.

1849 Abraham Gesner, a Canadian geologist, devises a method to distill kerosene from petroleum. Kerosene (also known as "coal oil" or "paraffin oil") is cheap, easy to produce, relatively odorless, and can be stored indefinitely, and it rapidly replaces whale oil and other oils in illumination applications.

1849 Joseph J. Couch of Massachusetts receives the first U.S. patent for a steam-powered percussion rock drill. Later versions of the drill, along with dynamite, will make it possible to excavate large amounts of earth and rock in tunneling, mining, and quarrying.

1849 The California Gold Rush begins, setting off a process that will eventually make California the most populous U.S. state, with huge requirements for energy and water.

1849 The speed of light is accurately calculated by Armand Fizeau, using a turning disk and a mirror placed five miles away. His assistant, Jean Foucault, improves the technique in 1850 and determines the speed of light to be 185,000 miles per second, within 1% of the true figure.

1850–1859

1850 Rudolf Clausius, a German mathematical physicist, makes the first formal expression of the second law of thermodynamics, which becomes a cornerstone of modern science. This states that heat cannot of its own accord pass from a colder body to a hotter one.

1850 Scottish chemist James "Paraffin" Young patents a process of extracting oil from cannel coal.

1850 The first absorption refrigeration machine is developed by Edmond Carré, using water and sulfuric acid. His brother Ferdinand demonstrates an ammonia/water refrigeration machine in 1859. Most early refrigerants are flammable and/or toxic, and some are also highly reactive. These disadvantages are not overcome until fluorocarbon refrigerants appear in the 1930s.

1850s Swift-sailing, graceful clipper ships carry passengers and freight on long ocean voyages, especially trips to California and China.

1850s The New England whaling industry is in steep decline because of massive depletion of whale stocks and the rise of petroleum and other fuels that are used in place of whale oil.

ca. 1850 Simple distillation of seep and salt well oil is carried out in Pittsburgh, Pennsylvania by Samuel Kier. He builds a one-barrel still to treat crude oil, which he buys by the gallon. This is regarded as the first commercial refinery in North America.

1851 In his work treating tropical diseases, Florida physician John Gorrie becomes convinced that cooler sickrooms will reduce fever and make patients more comfortable. He goes on to patent a cooling machine that produces artificial ice, thus laying the groundwork for modern refrigeration and air-conditioning.

1851 William Thomson (Lord Kelvin) is the first to describe a heat pump, a device in which heat is absorbed by the expansion of a working gas and given off at a higher temperature in a condenser.

1852 Henri Giffard, a French engineer, builds and flies the first full-sized airship, a hot air balloon powered by steam and steered by a rudder. This can be considered the first powered and controlled aircraft flight.

1852 Irish scientist George Gabriel Stokes names and explains the phenomenon of fluorescence based on the absorption of ultraviolet light and then emission of blue light.

1852 James Prescott Joule and Lord Kelvin demonstrate that an expanding gas with no outside energy source will become cooler; this is now known as the Joule-Thomson effect.

1853 Employees of the Russian-American Company document oil seeps in Alaska, on the western shore of Cook Inlet. The native peoples of Alaska had traditionally used oil-soaked tundra as heating fuel.

1853 Ignacy Lukasiewicz, a Polish druggist in the city of Lvov (then part of the Austro-Hungarian Empire, now in Ukraine), invents an improved kerosene lamp, using the distillation techniques recently developed by Abraham Gesner in Canada.

1854 Abraham Gesner opens a plant in New York to convert coal into kerosene, using his patented process of fractional distillation.

1854 Daniel Halladay invents the first commercially successful self-governing windmill in Connecticut. Halladay's machine has four wooden blades; it automatically turns to face changing wind directions and controls its own speed of operation.

1854 German-American watchmaker Heinrich Gobel of New York develops an incandescent electric lamp in which a carbonized bamboo filament is secured in a glass vessel. This predates the work of Thomas Edison and Joseph Swan by two decades.

1854 Hermann von Helmholtz proposes that the Sun derives its energy from gravitational shrinkage.

1854 In order to provide oil for his kerosene lamp business, Ignacy Lukasiewicz establishes an "oil mine" near Bobrka, a town in the Gorlice region of Poland. High-quality crude oil is pumped from wells as deep as 150 meters. This is often described as the beginning of the modern oil industry, though various competing claims have been made.

1854 New York lawyer George H. Bissell and others, most notably financier James Townsend, form the Pennsylvania Rock Oil Company to search for oil in western Pennsylvania. This is the first U.S. company created explicitly as an oil business.

1855 Abraham Gesner receives a patent for his process of extracting kerosene from bituminous shale and coal.

1855 An 800-horsepower water turbine is installed at the Pont Neuf waterworks in Paris.

1855 At the request of the Pennsylvania Rock Oil Company, chemist Benjamin Silliman of Yale University applies fractional distillation to Pennsylvania rock oil and discovers that it produces high-quality lamp oil.

1855 Heinrich Geissler designs a mercury-powered air pump which he uses to create the most effective vacuum tubes to date. The Geissler tube will eventually lead to the discovery of electrons.

1856 Following a series of cholera outbreaks in London in the late 1840s, the Metropolitan Board of Works is created and charged with providing a cleaner water supply for the city, with particular attention to the condition of the Thames River. Measures to control air pollution are also taken at this same time.

1856 Henry Bessemer develops a new way to remove carbon from cast iron in steel-making by sending a blast of air through molten iron in special furnaces. This is now known as the Bessemer process.

1856 The Siemens brothers, William and Friedrich, introduce a regenerative gas furnace. This is an improvement on existing furnaces in which much of the heat of combustion is lost by being carried off in the hot gases that pass up the chimney.

1856 Workers near the Neander River in Germany discover the first known skeleton of a Neanderthal (Neandertal) man in a limestone cave. This becomes an important step in studying human evolution.

1857 A system of oil wells in Romania begins to produce a sizeable output. Some authorities describe this operation as the origin of the oil industry, rather than the efforts of Ignacy Lukasiewicz (see 1854).

1857 George Pullman develops the first successful railroad sleeping car, making longer trips more comfortable. The "Pullman Car" becomes widely known in 1865 when one is used to carry the body of President Abraham Lincoln from Washington for burial in Illinois.

1857 Oil is found at a site in Canada near Lake Erie; this is considered to be North America's earliest commercial petroleum discovery, predating the better-known Drake field in Pennsylvania by two years.

1857 The first elevator for public use is installed in New York City by Otis Brothers, a steam-driven type installed in a five-story department store on Broadway.

1858 A telegraph cable is laid across the Atlantic Ocean between Newfoundland, Canada and Ireland, the first such connection between the United States and Europe. This cable line fails after three months but a successful one is eventually established in 1866.

1858 Brothers C. W. and W. W. Marsh of De Kalb, Illinois develop a harvesting machine that improves on the McCormick reaper by reducing the amount of labor needed to bind the grain.

1858 James Miller Williams of Hamilton, Ontario mines "gum beds" (surface oil seeps that have congealed), planning to sell the material for caulking in ships. He soon sees the greater potential for oil as an illuminating fuel and builds a small refinery.

1858 William Hart, originator of natural gas drilling in the United States in 1821, forms the Fredonia Gas Light and Water Works Company; this is described as the first incorporated natural gas company in the United States.

1859 Charles Darwin publishes *On the Origin of Species,* a book outlining his theory of evolution by natural selection. It becomes the foundation of a completely new approach to biology.

1859 Edwin Drake drills an oil well near Titusville, Pennsylvania, with a depth of about 70 feet. This is considered to be the first commercially successful U.S. oil well and is regarded as the origin of the U.S. petroleum industry.

1859 Gaston Plante of France invents the first rechargeable lead-acid storage battery (alkaline battery).

1859 George B. Simpson of Washington, D.C., patents an electric cooking device using the same principle as modern electric ranges.

1859 Jean Lenoir of France builds the first practical and operative internal combustion engine, fueled by coal gas. The following year he incorporates it in a "horseless carriage."

1859 John Tyndall, a self-taught Irish scientist, suggests that water vapor, carbon dioxide, and any other radiatively active ingredient of the atmosphere could produce "all of the mutations of climate which the researches of geologists reveal." This will become a basic component of climate science.

1859 Scottish physicist William John Rankine proposes a temperature scale having an absolute zero value, below which temperatures do not exist, and using a degree of the same size as the Fahrenheit scale. Absolute zero, or 0°R, is the temperature at which molecular energy is at a minimum.

1859 The first volume of the *American Gas-Light Journal* is published.

1860–1869

1860 A proposal to transport oil by pipeline is made by 17-year-old Samuel Duncan Karns of Parkersburg, West

Virginia. A salt well in the region is beset by an influx of oil; Karns suggests that the oil be shipped to a refinery on the Ohio River via a wooden pipeline along a route that would allow the oil to flow by gravity.

1860 Concerned about France's dependence on coal, Auguste Mouchout begins the first research on the direct conversion of solar radiation into mechanical power. He goes on to develop a solar-powered steam engine, the first successful example of a motor running on solar energy.

1860s American railroads provide the best means of transporting oil to coastal cities, where it is used in refineries or shipped overseas. The first method involves transporting the oil in wooden barrels stacked on flat cars or in layers on specially built racks.

1860s The clipper ship era ends as these classic vessels are supplanted by the steamship, which has proven to be a much more profitable method for carrying large freight cargoes.

1861 The first bulk transportation of oil in America is accomplished by wooden barges known as "bulk boats" on Oil Creek and the Allegheny River in Pennsylvania. These boats carry 400-barrel loads down river to refineries at Pittsburgh.

1861 The first shipment of oil to be exported from the United States (Philadelphia) to Europe (London) is carried aboard the sailing ship *Elizabeth Watts*.

1861 The U.S. Civil War begins. Steam-powered railroads will play a key role in the war for the transportation of troops and supplies, especially through the U.S. Military Railroads system organized by the Union government.

1862 Louis Pasteur's experiments with bacteria conclusively disprove the theory of spontaneous generation and advance the germ theory of disease, revolutionizing medical research and treatment.

ca. 1862 Werner von Siemens proposes the measurement of electrical conductance; later the name *siemens* will be used for a basic unit of conductance.

1863 Because of the damage caused by harmful emissions from alkali works, Great Britain enacts its first environmental protection law, the Alkali Act. This requires a reduction by 95% in the discharge of all noxious and offensive gases by alkali manufacturers. It also establishes an official regulator, the Alkali Inspectorate, to enforce the legislation.

1863 French engineer Rodolphe Leschot invents the diamond core drill and uses it to drill blast holes for tunneling Mount Cenis on the France-Italy border. Eventually, diamond drill bits will produce a major advance in oil and gas exploration because their extreme hardness improves drilling efficiency.

1863 London opens the first route of its subway (underground) system, linking Paddington railway station to Farringdon Street, near the King's Cross rail station. The line is built on the "cut and cover" method in which a large trench is dug under a street and then roofed over so that surface traffic can use the street.

1863 Pierre and Ernst Michaux of Paris produce the Michaux Velocipede, a new form of bicycle in which the pedals are directly connected, via crankshafts, to the front wheel.

1864 George Perkins Marsh of Vermont publishes *Man and Nature: The Earth as Modified by Human Action,* often considered one of the first modern treatments of environmental problems, and in particular the role of humans as agents of environmental change.

1864 Scottish physicist James Clerk Maxwell proposes his equations of electromagnetism and is the first to suggest that light is an electromagnetic wave.

1864 Siegfried Marcus of Vienna, Austria builds a one-cylinder engine with a crude carburetor, and attaches it to a cart for a 500-foot drive. This is described as the world's first successful gasoline-powered vehicle.

1864 The construction of the Hot Lake Hotel near La Grande, Oregon marks the first time that geothermal energy from hot springs is used for the direct heating of a large building.

ca. 1864 Natural gas deposits are found near Stockton, California, the first such discovery in the western United States.

1865 After years of research with pea plants, Gregor Mendel publishes his theories of heredity in an obscure publication. These ideas establish the science of genetics but are generally unknown for over 30 years.

1865 America's Civil War ends, and with it the use of human slave power as an energy source in the United States.

1865 Amos Densmore, an oil buyer and shipper at Oil Creek, Pennsylvania, invents the first railroad oil tank car. The Densmore tank car is a major breakthrough in bulk transportation compared to the earlier method of stacking barrels on a flat car.

1865 British inventor George Law devises a rock drill in which an air-driven piston operates a hammer tool. This is considered to be the first practical application of compressed air to drive a motor.

1865 Colonel E. A. L. Roberts uses eight pounds of black powder to make the first successful oil well "shot" on the Ladies Well along Oil Creek, Pennsylvania. "Shooting" a well refers to the detonation of an explosive device in the well bore so as to allow oil to flow to the surface.

1865 Friedrich August Kekule recognizes that benzene is made up of carbon and hydrogen atoms in a stable

hexagonal ring, not a chain. This discovery changes the way chemists view molecular structure.

1865 Rudolf Clausius introduces the concept of entropy (disorder) to explain the directionality of physical processes. He establishes that entropy can never decrease in a physical process and can only remain constant in a reversible process. This becomes formalized as the second law of thermodynamics, which Clausius had first proposed in 1850.

1865 Samuel Calthrop of the United States develops a streamlined railroad train.

1865 Samuel Van Syckel builds the first major oil pipeline in the Oil Creek area of Pennsylvania. At the time teamsters charged extremely high prices to haul barrels of oil by wagon, and Van Syckel underprices them with his line of wrought iron pipes, laid partly underground and partly on the surface.

1865 The British economist W. Stanley Jevons publishes *The Coal Question,* in which he argues that coal is essential for the industry of Britain and that this resource will one day be exhausted, causing the collapse of British society. The prediction proves wrong because Jevons did not foresee improvements in mining technology and the use of oil as a substitute for coal.

1865 W. Stanley Jevons formulates Jevons' paradox, a concept now employed by modern scholars. It states that increased efficiency in the use of a natural resource, such as coal, will actually result in greater consumption, not less. This is because improvement in efficiency will lead to higher demand; e.g., U.S. gas consumption increased after fuel-efficient cars were introduced in the 1970s.

1866 Alfred Nobel tames the volatility of nitroglycerin by enclosing it in kieselguhr, a packing material, making a much safer explosive that he names *dynamite*. He goes on to produce other new explosive substances in the 1870s.

1866 Early Pennsylvania oil producers adopt the 42-gallon barrel as their standard. It is still the standard unit of measurement today.

1866 German scientists Adolph Fick and Johannes Wislicenus test von Liebig's theory that protein alone powers muscle activity. They eliminate protein from their diet, climb Mt. Faulhorn in the Swiss Alps, and calculate their energy expenditure. Urine analysis indicates that protein catabolism could not have supplied the requisite energy.

1866 Lyne Taliaferro Barret and his partners in the Melrose Petroleum Oil Company drill the first oil well in Texas, in Nacogdoches County. Though this well produces only a modest flow of oil, other efforts at the same site in 1887 have much greater success and set off the state's first oil boom.

1866 Strip mining begins near Danville, Illinois, when horse-drawn plows and scrapers are used to remove overburden so that the coal can be dug and hauled away in wheelbarrows and carts.

1866 Werner von Siemens discovers the dynamo-electric principle. This solves the problem of needing direct current batteries to generate continuous current and high voltage, and provides the crucial advance needed to supply electric power at commercially reasonable prices.

1867 German engineer Nikolaus August Otto and his partner Eugen Langen demonstrate an improved version of Jean Lenoir's engine, theirs operating on liquid fuel instead of coal gas. This engine is the forerunner of Otto's major breakthrough in 1876 (see below).

1867 The Reverend Leonard Wheeler, a missionary to the Ojibwa Indians of Wisconsin, invents a classic American farm windmill that is a departure from the earlier Dutch style.

1868 A refrigerated railroad car is patented by J. B. Sutherland of Detroit, Michigan, for use in transporting perishable foods such as meats and dairy products. In the 1870s and 1880s Gustavus Franklin Swift will build a huge meat-packing firm using refrigerated cars to ship meat from Chicago to markets in the East.

1868 American engineer George Westinghouse patents the air brake, which benefits the railroad industry by making braking a safer venture and thus permitting trains to travel at higher speeds. To manufacture his invention he establishes the Westinghouse Air Brake Company, a predecessor to the huge Westinghouse Electric Company.

ca. 1868 Georges Leclanche of France invents a precursor to the modern dry cell battery, now widely used in consumer products such as flashlights, toys, and other electronic devices.

late 1860s Enclosed cast iron stoves are common in America. Stoves had been resisted at first on the grounds that they hid the fire, but gradually they replace the open fireplace because of their greater efficiency.

1869 Belgian Zenobe-Theophile Gramme builds the first efficient direct-current dynamo, two years after building an efficient alternating-current dynamo. In 1873 it is demonstrated that the Gramme type dynamo can function as an electric motor, a major step in the development of electric power.

1869 Dmitri Ivanovich Mendeleev is the first of several scientists to publish a periodic table of the elements, with elements arranged in order of increasing atomic weight.

1869 German biologist Ernst Haeckel introduces the term *ecology* in his work *Generelle Morphologie der Organismen.*

1869 In the French Alps, Aristide Berges sets up a pipeline to transport a pressurized stream of water to his wood pulp factory. He then uses a dynamo to transform the water power into electricity. This is regarded as the first systematic use of hydroelectric power.

1869 Norman Lockyer suggests that a yellow spectral line observed in the Sun's spectrum during an 1868 eclipse must belong to a new element. Lockyer has discovered helium (from the Greek *helios,* or Sun), which is identified in terrestrial material by William Ramsay in 1895 and later isolated by him.

1869 The first transcontinental railroad line in the United States is completed when a symbolic golden spike is driven into the ground at Promontory Point, Utah. The Union Pacific line from the East meets the Central Pacific line from the West at this spot.

1869 The Pennsylvania Railroad introduces the horizontal boiler-type metal tank car for oil shipments, which is still the basic design in use today. The metal tanks are safer and more reliable than the earlier wooden Densmore tank car.

1869 The Red Sea is joined to the Mediterranean as the Suez Canal opens, construction having begun 10 years earlier under the direction of the French promoter Ferdinand de Lesseps. The canal will come under the control of Great Britain in 1875 and will serve as the crucial link to India and other British interests in the Far East.

1869 The ship *Charles* of Antwerp, Belgium becomes the forerunner of the modern oil tanker, carrying oil from the United States to Europe in iron tanks.

1869 Thomas Andrews of Belfast, Ireland suggests that there is a "critical temperature" for every gas at which it can no longer be liquefied by pressure alone.

1870–1879

1870 Briquetting of coal is introduced in the United States. This provides greater thermal stability and cleaner combustion.

1870 Drawing upon the earlier work of George Medhurst, American inventor Alfred Ely Beach constructs New York City's first subway, a pneumatically powered system operating along a line one block long in lower Manhattan. Though the route closes three years later for lack of riders, it represents the beginning of underground transport in the United States.

1870 John D. Rockefeller and his associates form the Standard Oil Company in Cleveland, Ohio. Cleveland had become the leading site for the U.S. oil refining industry in the 1860s.

ca. 1870 Julius Hock of Vienna develops the first internal combustion engine operating on modern principles.

1870s Edward Frankland, a British chemist, is among the first to apply the conservation of energy principle to human energetics. Building on the work of Fick and Wislicenus, Frankland shows that the energy intake of the body equals the heat expelled. He also is among the first to measure the caloric value of food.

1871 Andrew Smith Hallidie in San Francisco, California patents the first cable (street) car. In 1873 the first passengers travel along 2,800 feet of track to the crest of a hill 300 feet above the starting point. This is the first successful attempt to replace horses with a machine in the U.S. streetcar system.

1871 Charles Darwin publishes *The Descent of Man,* in which he specifically applies his principles of evolution in the animal kingdom to the human race.

1871 Italian-American Antonio Meucci declares the intent to patent his "Talking Telegraph," a predecessor to the comparable device patented five years later by Alexander Graham Bell. Meucci is forced to let his application lapse because of lack of funding, and Bell becomes recognized as the inventor of the telephone.

1872 Abraham Brower opens the first mass transportation system in America along Broadway in New York City. The vehicle is an omnibus, which looks like a stagecoach and is pulled by horses.

1872 George B. Brayton, a British-trained engineer of Boston, Massachusetts, receives a patent for a gas-powered streetcar. This goes into operation the following year, though electric-powered streetcars will become the dominant form.

1872 L. C. Marcy of Philadelphia develops an improved oil lamp using two flat and parallel cotton wicks.

1872 The first long-distance pipeline for natural gas goes into operation in Pennsylvania, extending from the Newton gas well to the town of Titusville.

1872 The first wireless telegraphy patent in the United States is issued to Mahlon Loomis. In 1866, Loomis had set up a demonstration experiment on two mountain peaks in the Blue Ridge Mountains of Virginia, 22 kilometers apart.

1872 Thomas Edison develops an automatic telegraph machine, one of many improvements in telegraphy for which he is responsible.

ca. 1872 Swedish-American engineer John Ericsson develops an improved form of the Stirling engine; this is intended for use as means of pumping water to provide irrigation for arid lands of the American West. Ericsson is also noted as the designer of the Civil War ironclad battleship *Monitor.*

1873 Entrepreneurs Robert and Ludwig Nobel (brothers of the better-known Alfred) establish a company to exploit the oil riches of Baku, Azerbaijan. They become the most successful oil producers in the area and help to make Baku's oil industry the largest in the world in the late 19th century.

1873 Johannes van der Waals refines Boyle's and Charles's laws of gases, introducing modified equations that allow for a wider range of molecular behavior.

1873 Joseph Henry proposes that the physical drudgery of human labor can be ameliorated only by more inanimate sources of energy and more powerful energy convertors. This is an early application of the new science of thermodynamics to socioeconomic concerns.

1873 London experiences the first of a series of "killer fogs." A thick winter fog settles over the city, darkening the sky to the point that visibility is nearly zero. An estimated 500 deaths occur both from respiratory illnesses and from accidents caused by the lack of visibility. Two similar events in 1880 and 1892 are each responsible for about 1,000 deaths.

1873 Thaddeus S. C. Lowe of the United States develops a new gas manufacturing process, carbureted water gas. This is cheaper and more efficient than the existing coal gas process and soon comes to dominate the manufactured gas industry.

1873 Willoughby Smith of England and his assistant Joseph May note that when selenium is exposed to light, its electrical resistance decreases. This discovery provides the means to transform images into electric signals, and leads to the manufacture of photoelectric cells and other light sensors. It also lays the foundation for the development of television.

1874 Joseph F. Glidden of DeKalb, Illinois develops a successful form of barbed wire, improving on the version patented in 1868 by Michael Kelly. Barbed wire will change the face of the American West, as it provides a means of confining cattle and of separating them from crop fields.

1874 Jules Verne publishes *The Mysterious Island,* set during the U.S. Civil War. The book's main characters speculate on how the depletion of coal will affect the future. Cyrus Harding, one of the characters, asserts "I believe that one day hydrogen and oxygen, which together form water, will be used either alone or together as an inexhaustible source of heat and light."

1874 Siegfried Marcus produces a true four-cycle, four-wheeled vehicle with electrical ignition, jet carburetion, and a throttle. This is accepted by many historians as the true forerunner of the modern automobile, though many competing claims exist.

1874 The "Destructor" goes into operation in Nottingham, England. This is the first known example of a systematic incineration of urban solid wastes. The Destructor burns mixed waste and produces steam to generate electricity. Several hundred similar facilities are built in the coming decades.

ca. 1874 Swedish engineer Charles Wilson builds the first solar distillation plant, in the mining area of Las Salinas in what is now northern Chile. This provides fresh water to the mining community; it remains in operation for about 40 years.

1876 Alexander Graham Bell transforms the world of communication when he patents the first telephone. Bell's patent precedes a similar application just two hours later on the same day by rival inventor Elisha Gray. Bell's device is improved almost immediately by Thomas Edison, who adds carbon powder to the mouthpiece, allowing it to carry more current.

1876 German engineer Carl von Linden patents a process of liquefying gas (ammonia) that marks a major advance in basic refrigeration technology.

1876 Nikolaus August Otto of Germany invents a successful four-stroke internal combustion engine, known as the "Otto cycle." This engine represents the first true advance over the steam engine and is almost universally adopted for all liquid-fueled automobiles from that point on.

1876 Thomas Edison establishes a laboratory at Menlo Park, New Jersey that will become the site of many of his famous inventions and innovations.

1876–1878 After a wave of violence in the coal mining region of Pennsylvania, a sensational murder trial results in the execution of 20 members of the Molly Maguires, a secret society of Irish-American miners. The historical record is unclear as to what extent the violence was incited, or even committed, by private detectives hired by mine owners to infiltrate the miners' union.

1877 A steam-powered shovel excavates some 10 feet of overburden from a 3-foot-thick coal bed near Pittsburg, Kansas.

1877 Austrian physicist Ernst Mach publishes a paper that accurately describes the relationship between the speed of a fast-moving object and the local speed of sound (Mach number).

1877 Louis Paul Cailletet and Raoul-Pierre Pictet independently liquefy oxygen.

1877 Oakland Gas Light establishes what is said to be the first high-pressure gas transmission in the United States; two years later the company introduces gas cooking stoves into California.

1878 American inventor Charles F. Brush develops an efficient dynamo and goes on to produce a commercially viable arc light, which is used to illuminate a private home in Cincinnati and a public square in Cleveland. He then founds Brush Electric Company, which will eventually join Edison's General Electric Company.

1878 Marie Alfred Cornu, a French physicist, measures the spectrum of solar radiation reaching Earth's surface and suggests that the absence of short-wave ultraviolet radiation is due to an atmospheric absorber. Two years later, Walter Hartley concludes that Cornu's absorber is ozone in the upper atmosphere (the ozone layer).

1878 Samuel Pierpont Langley of the United States invents the bolometer, a radiant-heat detector that is sensitive to differences in temperature of one hundred-thousandth of a degree Celsius (0.00001°C). The device is still used to provide accurate regional and global measurements of the components of the Earth's radiation budget.

1878 Thomas Edison founds the Edison Electric Light Company to finance his efforts to invent a practical incandescent electric lamp. This and various other electric companies established by Edison will later be combined in what becomes the General Electric Company.

1879 Attorney George Selden of Rochester, New York applies for a patent for his "road engine," a vehicle powered by an internal-combustion engine. This is the first automobile patent in the United States, though it is not granted until 1895. Later Selden will sue Henry Ford for producing his own automobiles, and the courts then declare the Selden patent too vague to be enforced.

1879 English physicist William Crookes identifies a fourth state of matter in addition to solid, liquid, and gas, one in which a collection of electrons and ions move freely within a given space. Irving Langmuir later applies the name *plasma* to this state. Plasmas make up about 99% of the visible matter in the universe and offer the potential for a large-scale energy source.

1879 Thomas Edison in America and Joseph Swan in England independently produce carbon-thread electric lamps. Edison's first bulb burns for 40 continuous hours, and only a year later he opens the first electric generation station in London used primarily for lighting.

1879 The Siemens Company builds the world's first electric train in which power for the train is supplied through the tracks rather than from overhead wires.

1879 William Adams, a British colonial official in India, constructs a heating system with a series of large mirrors to concentrate the Sun's radiation. He demonstrates it by placing a piece of wood in the focus of the mirrors where, he notes, "it ignited immediately." This system is similar to the one reportedly used by Archimedes 2,000 years earlier.

1880–1889

1880 Charles Brush's Electric Light and Power Company successfully demonstrates an arc lighting system along New York City's famous avenue Broadway, and soon thereafter builds New York's first central power station. Broadway will later become known as the "Great White Way" because of its dazzling array of electric lights.

1880 John Milne of England invents the first modern seismograph.

1880 Lester Allan Pelton of California obtains a patent for a new type of water turbine with a more effective jet of water. In 1883 the Pelton turbine wins a competition for the most efficient water wheel turbine, and it will serve as the standard design for the next 40 years.

1880 Michigan's Grand Rapids Electric Light and Power Company, generating electricity by means of a Brush dynamo belted to a water turbine, lights up a set of Brush arc lamps to provide theater and storefront illumination.

1880 Pierre and Jacque Curie discover that some crystalline materials, when compressed, produce a voltage proportional to the applied pressure and that when an electric field is applied across the material, there is a corresponding change of shape. This is called piezoelectricity or pressure electricity.

1880 The Bradford Gas Company of Rixford, Pennsylvania begins using the first natural gas compressor.

1880 Wabash, Indiana becomes the first city to be lit entirely by electric light, using the Brush dynamo and arc light system.

early 1880s A "current war" is underway over the choice between DC current, promoted by Thomas Edison, and AC current, promoted by George Westinghouse. An intense battle for public opinion and political support ensues.

1880s John D. Rockefeller and Standard Oil base the first decades of their business producing kerosene and shipping it around the world on ships called "kerosene clippers." Today kerosene is used primarily as a heating and aviation fuel, and as a solvent and thinner.

1880s Oil tanker barges are a major form of transportation for refined products in the Low Countries and Germany, which have extensive networks of rivers and canals.

1881 A hydroelectric generating station is operated on the Niagara River. An 85-foot cascade of water generates electricity to run the machinery of local mills and to light some of the village streets.

1881 German naturalist Karl Semper publishes the *Natural Conditions of Existence as They Affect Animal Life*. This is regarded as one of the earliest studies of energy flows in nature.

1881 Jacques D'Arsonval of France describes a method for using the temperature difference between the warm surface sea water and cold deep ocean water to generate

electricity (later known as ocean thermal energy conversion, or OTEC).

1881 Scottish engineer Sir Dugald Clerk patents his two-stroke internal combustion engine (in contrast with the four-stroke engine developed at around the same time by Nikolaus Otto). It will become widely used for motorcycles and other light engines.

1882 Augustin Mouchout and his associate Abel Pifre demonstrate a solar power system in Paris in which solar energy is used to make steam to operate a printing press.

1882 The first commercial hydroelectric power plant goes into operation on the Fox River in Appleton, Wisconsin. It is used to power two paper mills and eventually the town's streetcar system as well.

1882 The Standard Oil Trust is formed by John D. Rockefeller, largely to shield himself against legal challenges to his control of oil production and refining, and against public outcry over this perceived monopoly.

1882 The United States Electric Illuminating Company establishes an electric power station in Charleston, South Carolina.

1882 Thomas Edison's power generating company in New York City produces electricity for lighting. Three 125-horsepower "Jumbo" generators at the Pearl Street station provide power to 5,000 lamps in 225 homes.

1883 Charles Fritts, an American inventor, describes the first solar cells made from selenium wafers.

1883 John Ericsson designs and demonstrates his "Sun motor," a working solar heat engine.

1883 S. Podolinsky, a Ukrainian socialist, is the first to explicitly describe the economic process from an energy perspective, and in particular agricultural energetics. His biophysical analysis leads him to conclude that limits to growth are imposed not by constraints on production, but by physical and ecological laws.

1883 Swedish engineer Carl Gustaf de Laval develops a high-speed steam turbine. His innovative use of special reduction gearing allows a turbine rotating at high speed to drive a propeller at comparatively slow speed, a principle having great value in marine engineering.

1883 The Chartiers Valley Gas Company is formed to produce, gather, and transport gas to Pittsburgh's industrial plants.

1884 English engineer Charles Parsons invents the compound steam turbine, which converts the power of steam directly into electricity; it ultimately becomes the preferred power plant of electric power stations and ships.

1884 Paul Nipkow patents the world's first electromechanical image scanning system, consisting of a rapidly rotating disc with a pattern of holes. This is an important early step in the development of television; in the 1920s John Logie Baird will use the Nipkow disc to transmit the first true television pictures.

1884 The French military officers Charles Renard and Arthur Krebs make a flight of five miles in their airship *La France,* in which they are able to guide the ship so that it lands at the point from which they begun the flight.

1884 The Serbian-American electrical engineer Nikola Tesla designs practical methods for generating alternating current, a prerequisite for long-distance transmission of electricity. Called the "Master of Lighting," Tesla will receive more than 100 patents relating to electricity generation and transmission and radiocommunications.

mid 1880s Steam-driven plows and threshing machines are in use on the prairies of North America, though oxen and horses will remain the principal source of power for some time to come.

1885 A high-pressure manufactured gas transmission line is installed between Oakland and Alameda under the San Francisco Bay.

1885 A Westinghouse Company employee, William Stanley, invents the transformer for shifting the voltage and amperage of alternating current, enabling electricity to be transported over long distances with relatively little loss.

1885 Carl Auer von Welsbach at the Bunsen laboratory in Germany develops the Welsbach mantle, which becomes a key technology for the utilization of manufactured gas.

1885 Gottlieb Daimler extends Otto's work and invents what is recognized as the prototype of the modern gas engine, with a vertical cylinder and with gasoline injected through a carburetor (patented 1887). Daimler uses the engine in a two-wheeled vehicle he calls the *Reitwagen* (Riding Carriage). One year later he builds the world's first four-wheeled motor vehicle.

1885 John Kemp Starley of England introduces the Rover Safety Bicycle, a vehicle with essentially the same design as the modern bicycle. It is an immediate success and is marketed by Starley and fellow bicycle innovator William Sutton through their Rover Company.

1885 Sylvanus F. Bowser of Fort Wayne, Indiana develops a workable gasoline pump and then founds a company to manufacture the device.

1885 The German mechanical engineer Karl Benz designs and builds the world's first practical automobile to be powered by a gasoline-burning internal combustion, engine. The following year, Benz offers this vehicle for sale to the public.

1885 The Massachusetts Gas Commission becomes the first government agency to regulate the energy industry, including the setting of prices.

1886 Charles Hall of the United States and Paul Hêroult of France independently develop a molten electrolysis

process for the production of aluminum. The Hall–Héroult process is still the basic method of aluminum production.

1886 The first alternating current power plant in the United States goes into operation in Great Barrington, Massachusetts.

1886 The first oil tanker of modern design, *Gluckauf,* is built at Newcastle, England for a Bremen, Germany shipping company. The ship differs from previous versions in that the liquid cargo is contained directly in tankage integral with the hull, rather than in individual barrels.

1887 Geographer and climatologist Eduard Bruckner makes pioneering studies of climate change and in particular of the effect of human activity on climate.

1887 German physicist Heinrich Rudolf Hertz proves that energy is transmitted through a vacuum by electromagnetic waves. He detects and produces radio waves for the first time, with enormous implications for future technologies in radio, television, and radar.

1887 High Grove Station in San Bernardino, California becomes the first hydroelectric plant in the western United States.

1887 In his studies of air and sound, Ernst Mach notes that air molecules behave irregularly when pushed faster than the speed of sound.

1887 Scottish veterinarian John Boyd Dunlop experiments with his son's bicycle in an effort to improve upon the existing technology of solid rubber tires. He develops the first practical pneumatic tire, which becomes an essential component of the automobile industry.

1887 Solomon R. Dresser of Bradford, Pennsylvania develops a new type of coupling for pipe joints that makes it possible to provide a leakproof pipeline over extended distances. The company he founds becomes a leader in the oil and gas industry.

1888 Andrew Lawrence Riker founds the Riker Electric Vehicle Company in Elizabethport, New Jersey, which becomes one of the country's largest manufacturers of electric cars and trucks. Riker produces his first electric car in 1894, using a pair of bicycles as a base.

1888 Austrian botanist Friedrich Reinitzer observes that a substance called cholesteryl benzoate appears to have two different melting points, one in which the solid changes to a cloudy liquid and a higher one in which it becomes a clear liquid. German physicist Otto Lehmann confirms this and coins the term "liquid crystal" to describe this distinct phase of matter.

1888 Carl Gassner of Germany develops the first dry cell battery, using a carbon rod as the positive electrode and zinc as the outer container and negative electrode. This cell is easy to use and portable and it becomes the prototype for the dry battery industry.

1888 Charles F. Brush begins operation of an efficient wind dynamo in Cleveland, Ohio, modifying existing windmill technology to produce what is considered the world's first wind power plant.

1888 Granville T. Woods develops and patents a system for overhead electrical conducting lines for railroads, which is adopted for the railway system of various large cities. Through this and other innovations in electricity and telegraphy, the African-American inventor Woods becomes known as "the Black Thomas Edison."

1888 In Richmond, Virginia, Frank Julian Sprague establishes what is regarded as the first large-scale, successful streetcar system powered by electricity. Similar systems are soon established in many other American cities and the streetcar becomes the preferred mode of urban mass transit.

1888 Oliver B. Shallenberger, an electrician at the Westinghouse Electric Manufacturing Company, receives a U.S. patent for an ampere-hour meter, which accurately measures the amount of electricity used.

1888 Wilhelm Hallwachs extends Hertz's work on the photoelectric effect, noting that when illuminated with ultraviolet light, a neutral or negatively charged body (a zinc plate) becomes slightly positively charged.

late 1880s Electric vehicles begin to appear on the streets of the United States and Britain, including an electric bus company making regular trips in London.

1889 Norwegian-American engineer Edwin Ruud invents the first device to store and automatically heat water. Ruud founds a company (still in operation) to produce water heaters for commercial and residential use.

1889 Paris becomes known as "The City of Light" during the 1889 World's Fair, since it is the first large city to switch to all electric lighting.

1889 Willamette Falls, Oregon provides the first commercial, long-distance transmission of electric power in the United States with a direct-current line to the city of Portland.

1890–1899

1890 French Engineer Clement Ader reports that his *Eole*, a steam-powered, bat-winged flying machine, is the first full-sized aircraft to make a powered flight. Witnesses describe a liftoff covering 50 meters at a height of about 20 centimeters off the ground. Historians differ as to whether this was truly an airplane flight, or merely a "running hop."

1890 The United Mine Workers of America is established in Columbus, Ohio through the merger of two earlier groups.

1890s Approximately 1,000 manufactured gas plants are in operation in the United States.

1890s Ludwig Boltzman, an Austrian physicist, founds the field of statistical mechanics. His interpretation of the second law of thermodynamics introduces the theory of probability into a fundamental law of physics and thus breaks with the classical precept that fundamental laws have to be strictly deterministic.

1890s U.S. chemist W. O. Atwater constructs a calorimeter and uses it to identify the energy content of different foods, using subjects who eat either a high fat or high carbohydrate diet while riding a stationary bicycle many hours per day. His "Atwater general factors" provide a good estimate of the energy content of the daily diet.

ca. 1890 Coal takes the place of wood as the dominant fuel in the U.S. economy, as industry has replaced human labor and draft animals with steam engines and other inanimate energy converters.

1891 Clarence Kemp of Baltimore, Maryland patents the first commercially successful solar water heater. His invention is later marketed in California and becomes popular enough to heat one-third of the homes in the city of Pasadena.

1891 German engineer Otto Lilienthal makes the first in a long series of successful glider flights. Lilienthal is a pioneer in aerodynamics, especially wing design, and his work will greatly influence the Wright Brothers.

1891 German-American chemical engineer Herman Frasch patents a process (Frasch process) to extract sulfur compounds from oil using copper oxide powder; until then, the foul smell of sulfur had prevented oil from being widely used as a fuel.

1891 Inventor Poul la Cour, the "Danish Edison," begins a series of experiments with wind turbines. His work will establish a foundation for modern use of wind energy.

1891 Nikola Tesla invents the Tesla coil, an air-core resonant transformer that can generate extremely high voltages at high frequency. Tesla coils can produce spectacular lightning-like discharges and have often been used in the film industry for special effects.

1891 The first extended natural gas pipeline in the United States is built. It carries gas from wells in Indiana over a distance of 120 miles to Chicago.

1891 The first long-distance high-voltage electricity line is established between Lauffen and Frankfurt am Main, Germany, a distance of over 100 miles.

1892 Edward L. Doheny begins the oil industry in Los Angeles, California, supposedly after he notices oil sticking to the wheels of a wagon and asks the driver where it had come from. Doheny becomes one of the wealthiest men in the United States and a major figure in the creation of modern Los Angeles. Later he will be at the center of the Teapot Dome Scandal (see 1921).

1892 Electrical winders are in use in British coal mines.

1892 Inventor Aubrey Eneas founds the Solar Motor Company of Boston to build solar-powered motors as a replacement for steam engines powered by coal or wood.

1892 John Froelich develops a tractor running on an internal combustion engine; this will eventually supplant the steam-powered tractor, which is less practical because of its large size.

1892 R. E. Bell Crompton and H. J. Dowsing of England patent the first electric radiator.

1892 The first steel barges for oil transportation are put into use by Standard Oil of New York. Steel barges are larger, stronger and safer than their wood predecessors.

1892 The first U.S. geothermal district heating system is in operation in Boise, Idaho.

1892 The *Murex* is the first tanker to pass through the Suez Canal, with a load of kerosene to be delivered from Batum, a Russian port on the Mediterranean, to Singapore. Marcus Samuel, the owner of *Murex,* forms Shell Trading and Transport, which eventually merges with Royal Dutch to form one of the world's largest oil companies.

1892 The Sierra Club is founded in San Francisco "to do something for the wilderness and make the mountains glad." John Muir is the club's first president.

1893 Julius Elster and Hans Geitel of Germany develop the first photoelectric cell. This is sensitive to both visible light and ultraviolet rays.

1893 Nikola Tesla begins a series of lectures and articles presenting his ideas for the fundamental methods of practical radio communication. Tesla thus precedes Guglielmo Marconi, popularly known as the "Father of Radio," though Marconi will state that he carried out his own work without knowing of Tesla's earlier efforts.

1893 Rudolf Diesel publishes a paper describing a new type of internal-combustion engine that will be named after him. Its distinctive feature is that it operates on the principle of self-combustion; i.e., rather than relying on a separate energy source such as a spark plug, it heats and compresses the fuel mixture to the point where it ignites itself.

1893 With the idea that his three-wheeled motor vehicle is unstable and uncomfortable, Karl Benz modifies the design to produce a four-wheeled vehicle. This is considered the world's first inexpensive, mass-produced car.

1894 Drilling begins at Alberta's Athabasca tar sand site, the world's largest accumulation of tar sands. Crews strike a reservoir of natural gas that will blow wild for 21 years.

1895 Conventional wisdom holds that the radio is born at this time, when Guglielmo Marconi of Italy demonstrates that it is possible to send signals by using electromagnetic waves to connect a transmitter and a receiving antenna. However, the U.S. courts will later conclude that Marconi relied upon the earlier work of Nikola Tesla (see 1893).

1895 Colonel George E. Waring is appointed as New York City's commissioner of street cleaning. Waring is a noted expert on drainage, sanitation, and urban public health. Public sanitation becomes a major issue, and the first curbside pickups of garbage begin.

1895 The Baltimore and Ohio Railroad becomes the first U.S. railway line to make regular use of an electric locomotive.

1895 The Metropolitan West Side Elevated Railway begins service in Chicago; it is described as the nation's first electric elevated rail line.

1895 Wilhelm Roentgen of Germany discovers X rays, and their medical implications are immediately recognized. Roentgen's discovery will lead to the year 1895 being referred to as the beginning of the Nuclear Age in human history.

mid-1890s A period of great popularity begins for electric cars in the United States. Over the next 20 years about 35,000 electric vehicles will come into use. Popular makes include the Riker, the Columbia, and the Baker.

1896 "Permitted" explosives are introduced to British coal mining. These are industrial explosives that are specially produced and tested for use in mines where firedamp and/or dangerous coal dust occur.

1896 A Santa Barbara, California group headed by Henry L. Williams drills wells from piers extending out into the sea. This leads to the 1898 discovery of the Summerland field, which is considered to be the first productive offshore oil facility in the world. (Similar offshore wells are reported in the Caspian Sea about five years later.)

1896 An electrical battery is developed by William W. Jacques that produces electricity directly from coal.

1896 Henri Becquerel of France discovers radioactivity when his experiments indicate that uranium emits radiation without an external source of energy such as the Sun, as observed when the invisible rays darken a photographic plate.

1896 Henry Ford's automotive experiments culminate with the completion of his own self-propelled car, the Quadricycle (so named because it runs on four bicycle type wheels). It is a four-wheeled vehicle built on a steel frame with no body; it is steered by a tiller, like a boat.

1896 Svante August Arrhenius of Sweden presents a paper to the Stockholm Physical Society entitled, "On the Influence of Carbonic Acid in the Air upon the Temperature of the Ground." This establishes the role of carbon dioxide as a greenhouse gas.

1896 The first subway on the European continent is opened in Budapest.

1896 William Hadaway of the United States receives a patent for an electric stove.

ca. 1896 A Niagara Falls hydroelectric plant is in operation, including a one-mile tunnel under the town of Niagara Falls and the use by George Westinghouse of Nikola Tesla's alternating current for long-distance transmission of power to Buffalo, 26 miles away.

1897 Encouraged by reports of the successful 1892 flight of Clement Ader's *Eole* flying machine, the French government provides funding for a new craft, named the *Avion*. Ader arranges a demonstration at a military base near Versailles, but *Avion* never succeeds in becoming airborne and the experiments are abandoned.

1897 Ferdinand Braun of Germany invents the cathode-ray oscilloscope, the precursor of the television picture tube. Braun does not patent this because he wishes it to be available to all scientists for research. Braun also makes important contributions to wireless telegraphy, for which he shares the Nobel Prize with Marconi.

1897 Henry Morris and Pedro Salom form the Electric Carriage and Wagon Company to operate electric cabs on the streets of New York City. This is described as the first motor vehicle service in the United States. Prior to this Morris and Salom had developed an electric car known as the Electrobat.

1897 Joseph John Thomson shows that cathode rays are particles with a negative electric charge, thus identifying the electron, the first subatomic particle known. This leads to much greater understanding of electric and electronic processes.

1897 Justus B. Entz, chief engineer of the Electric Storage Battery Company of Philadelphia, develops what is considered to be the first hybrid-electric vehicle (actually called the petro-electric car). It uses a gasoline engine as the primary drive and an electric motor to assist when more power is required.

1897 New York and other large cities have established garbage "picking yards" to separate out materials that have value for reuse, such as paper, metal, rubber, and twine.

1897 Nikola Tesla files his first U.S. patent application for basic radio communication. The patent is granted in 1900, the same year that Guglielmo Marconi's application is turned down. However, a 1904 ruling will overturn this and award the radio patent to Marconi.

1898 Energy recovery from garbage incineration begins in New York City; this is the first known example of a waste-to-energy system.

1898 Pierre and Marie Curie of France isolate and analyze radium and polonium, the elements that constitute most of the radioactivity in uranium ore.

1898 The Sulzer Brothers firm of Switzerland builds its first diesel engine. (Years earlier Rudolph Diesel had begun his career as an apprentice in Sulzer's engine works.) The company will become a world leader in the production of diesel engines.

1898 The Winton Motor Carriage Company of Cleveland, Ohio produces the first American truck, a gas-powered delivery wagon with a single-cylinder, six-horsepower engine.

1898–1900 Phillip Lenard, a German physicist, is the first to cause cathode rays to pass from the interior of a vacuum tube through a thin metal window into the air, where they produce luminosity.

ca. 1898 Albert Champion opens the Champion Ignition Company to manufacture spark plugs for the developing automobile industry. Champion later joins forces with General Motors and becomes the leader in the field. (Various others have been credited with inventing the spark plug, such as Edmond Burger, Oliver Lodge, and Robert Bosch.)

late 1890s A vital solar hot water heater industry emerges in Southern California; it later collapses due to the availability of cheap fossil fuels.

late 1890s The manufacture of automobiles is in full swing in western Europe and North America, led by industry pioneers such as Rene Panhard and Emile Levassor, Armand Peugot, Charles and Frank Duryea, and Ransome Eli Olds. This creates a great demand for gasoline. (Prior to this, kerosene for heating had been the main oil product.)

late 1890s The United States takes over the lead in coal production. Britain now ranks second, after having been the world leader since the beginnings of the industry in the 1500s. Germany is third, an indication of her growing industrial power relative to continental rival France.

1899 A leading American magazine, the *Literary Digest,* analyzes the future of the automobile and declares that "the ordinary horseless carriage is at present a luxury for the wealthy; and although its price will probably fall in the future, it will never, of course, come into (such) common use as a bicycle."

1899 Waldmar Jungner of Sweden invents the nickel-cadmium battery, which uses nickel for the positive electrode and cadmium for the negative. Two years later, Edison produces an alternative design by replacing cadmium with iron.

1900–1909

1900 Electric cars outsell gasoline and steam-powered cars in the United States Electric vehicles do not require a hand starter or gearshift, and they are quieter and operate with less vibration. Gas-powered cars take over in the 1920s due to cheap gasoline, the elimination of the hand crank, mass production, and the relatively limited range and speed of electric vehicles.

1900 German physicist Max Planck is the first to describe photons when he tries to explain the "black body" distribution of wavelengths in the light emitted by a solid hot object. "Planck's constant" becomes a universal constant that is fundamental to all quantum theory.

1900 Gregor Mendel's work on heredity is rediscovered by three different biologists working independently in separate countries. They finally give publicity and credence to Mendel's theories, establishing the field of genetics.

1900 Paul Ulrich Villard observes gamma rays, proving them to be a form of radiation more powerful than X rays.

1900 Standard Oil of New Jersey dominates in the United States with more than 80% of the market share in crude oil supplies, refining capacity, and kerosene and gasoline sold. In the following decade independent Texas oil companies will emerge to make inroads into this market share, and in 1911 Standard Oil is broken up by the government.

1900 The first dirigible to make a successful flight is built by Count Ferdinand von Zeppelin of Germany. The name *zeppelin* will come to be a synonym for aircraft of this type.

1900 The first transmission of a human voice over radio is made by Reginald Aubrey Fessenden, who transmits "words without wires" over a distance of 1,600 meters between two aerial systems on an island in the Potomac River. Fessenden will make the world's first radio broadcast on Christmas Eve in 1906.

ca. 1900 Charles Seeberger of the Otis Elevator Company develops a working escalator, improving upon an earlier design by Jesse Reno.

1901 At an oil field near Beaumont, Texas, a drilling crew supervised by Austrian engineer Anthony Lucas unexpectedly exposes a huge gusher of oil shooting 250 feet in the air. This becomes the site of the famous "Spindletop" well which will reportedly produce as much as 100,000 barrels of oil per day, a rate thousands of times greater than existing wells.

1901 Azerbaijan ranks as the world's most productive oil region, thanks to an oil boom in the area around the city

of Baku. A large refining industry also exists to turn crude oil into kerosene.

1901 The first Federal Water Power Act is enacted in the United States, beginning an era of large hydropower projects in the West.

1901 The *King Edward,* the first merchant vessel fitted with steam turbines, is launched in Scotland.

1901 Using an antenna he invented in 1895, Marconi broadcasts the first radio signals across the Atlantic Ocean from Cornwall, England to Newfoundland, Canada, a distance of more than 2,000 miles.

1901 What is now known as the Centennial Light Bulb is first lit at a firehouse in Livermore, California. It is a hand-blown bulb with a carbon filament, using approximately 4 watts of electricity. The bulb has now been burning continuously for over 100 years.

1902 A large gusher of oil spews forth in the Cotella (Katalla) region along the Gulf of Alaska. This is the first successful oil well in Alaska and provides evidence of the area's potential as an energy source.

1902 An anthracite coal strike in the United States severely disrupts the national economy. The power of the United Mine Workers union grows as its membership goes from less than 10,000 members before the strike to over 100,000 members afterward. The UMW becomes the nation's largest trade union.

1902 Arthur Korn of Germany develops a photoelectric scanning system for the transmission and reproduction of photographs. He goes on to establish a commercial picture transmission system linking Berlin, London, and Paris; this is the world's first fax (facsimile) network.

1902 Ernest Rutherford of New Zealand and Frederick Soddy of England publish their theory of radioactive decay, stating that atoms of a radioactive element emit charged particles (alpha or beta) and in doing so change into atoms of a different element.

1902 Marie and Pierre Curie measure the atomic weight of radium.

1902 Philipp Lenard studies the photoelectric effect, showing that electrons are emitted from certain metals when illuminated by ultraviolet light. He observes that the number of electrons released is proportional to the intensity of the light, while their speed is independent of this and is determined by the wavelength of the light.

1902 The first municipal oil-gas plant in the western United States is erected in Oakland, California for lighting purposes.

1902 The U.S. Congress establishes the Bureau of Reclamation to administer money from the sale of public lands to build dams and irrigation projects for western states. By 2000 the Bureau operates 60 dams with a total capacity of 14,757,668 kilowatts.

1902 Willis Haviland Carrier designs the first industrial air-conditioning system; in the following decades he and the Carrier Air Conditioning Corporation carry out various other innovations, culminating with the first air-conditioned stores and theaters in 1924 and a practical home air-conditioner in 1928.

ca. 1902 French auto pioneer Louis Renault is credited with the invention of the internal expanding drum brake, a great improvement in stopping power for motor vehicles.

1903 Brothers Wilbur and Orville Wright make the first successful flight of a controlled, powered, heavier-than-air machine, for 59 seconds over a distance of 850 feet at Kitty Hawk, North Carolina.

1903 Henry Ford and a group of financial backers found the Ford Motor Company in Detroit, Michigan.

1903 Konstantin Tsiolkovsky of Russia publishes *The Investigation of Outer Space by Means of Reactive Devices,* which suggests the use of liquid propellants for rockets to achieve greater range. This is one of many writings by Tsiolkovsky on space flight, in which he puts forth ideas such as multistage rockets, artificial satellites, and manned space stations.

1903 The 1903 Nobel Prize for Physics is jointly awarded to Pierre and Marie Sklodowska Curie and Henri Becquerel for their pioneering work on radioactivity. Marie Curie is the first woman to win a Nobel Prize.

1904 Beltram Borden Boldwood proves that uranium changes into radium through the process of radioactive decay. He later expands his theory, showing that radium decays into lead and that, in general, radioactive substances eventually decay into non-radioactive ones.

1904 English electrical engineer John Ambrose Fleming patents the thermionic valve, which is the first electronic rectifier of radio waves. Along with the amplifier grid of Lee de Forest (see 1906), Fleming's invention is the forerunner of the triode and other vacuum tubes and thus a foundation of the modern electronics industry.

1904 Prince Piero Ginori Conti of Italy successfully produces electricity from geothermal energy in an experimental "indirect cycle," using pure steam produced from a heat exchanger.

1904 Swedish-American electrical engineer Ernst Alexanderson develops a high-frequency alternator that provides the means for effective long-distance radio communication.

1904 The earliest large-scale U.S. aluminum recycling operations begin in Chicago and Cleveland.

1904 The first official New York City subway system opens in Manhattan, the Interborough Rapid Transit (IRT). It consists of 28 stations from City Hall to 145th Street, along 9 miles of track. The IRT extends to the Bronx in 1905, Brooklyn in 1908, and Queens in 1915.

1904 The Kern River oil field in Bakersfield, California reaches a peak production level of 17.2 million barrels per year, roughly equivalent to the output of the entire state of Texas.

1905 Albert Einstein proposes that light can be described as a stream of separate particles of energy. This explains the photoelectric effect, which had been extensively studied by his predecessors such as Hertz, Hallwachs, and Lenard, but which could not be accounted for by Newtonian physics, with its assumption that light is a continuous wave.

1905 In a second major paper, Einstein presents his special theory of relativity. He proposes that the laws of nature have the same application in all frames of reference. This applies to the speed of light, which is said to be constant, independent of any relative motion between the light source and the observer. Einstein describes mass and energy as equivalent ($E = mc^2$).

1905 In a third landmark paper of 1905, Einstein calculates how molecular movement causes Brownian motion (the irregular movement of microscopic particles suspended in a fluid). This confirms the validity of the atomic theory of matter.

1905 New York's Long Island Railroad is the first major rail carrier to install extensive main line electrification.

1905 The concept of a limiting factor in photosynthesis is shown by Frederick Frost Blackman. Photosynthesis is limited by the supply of carbon dioxide, temperature, and light intensity. The rate of photosynthesis is controlled by the factor in the shortest supply.

1905 The gas pump innovator Sylvanus F. Bowser develops a self-measuring pump that dispenses precise amounts of fuel.

1906 In his statement of the third law of thermodynamics, Walther Nernst proposes that absolute zero is a point that can be closely approached, but never actually reached.

1906 Lee de Forest of the United States invents the triode, a vacuum-tube device that amplifies weak electric signals. It is first used as a detector of radio waves, then as an amplifier for long-distance telephone calls, and eventually as a major technology of the radio transmitter, sound films, loudspeakers, and many other electronic devices.

1906 The British battleship *Dreadnought* is launched, the best-armed and fastest battleship in the world. Its presence in the Royal Navy sets off a competition among rival nations to build other ships of this class in the years leading up to World War I.

1907 American chemist Bertram Borden Boltwood is the first to use radiometric dating to estimate the age of the Earth. Using the half-life of uranium as a standard, he estimates the Earth's age at 2.2 billion years. This is a dramatic increase over previous estimates of Earth's age.

1907 An explosion in a coal mine in Monongah, West Virginia results in the death of 362 miners, virtually the entire work force in the mine at that time. This is the worst coal mine disaster in American history and consequently Congress is impelled to create the Bureau of Mines to study mine safety and conduct inspections.

1907 Coal reaches the historic high point (78%) of its share of the U.S. energy supply.

ca. 1907 Russian scientist Boris Rosing succeeds in transmitting and reproducing black-and-white silhouettes of simple shapes, using a mechanical mirror-drum apparatus as a camera and a cathode-ray tube as a receiver. In the 1920s, Rosing's student Vladimir Zworykin will become a key figure in the development of television technology in the United States.

ca. 1907 The first electric-powered washing machine is developed by Alva J. Fisher and marketed by the Hurley Machine Company of Chicago.

1908 A. A. Campbell-Swinton proposes a television system in which images are stored in the form of electric charge in a camera tube and reproduced on a picture tube, in the manner now employed in modern television broadcasting.

1908 Ernest Rutherford and his assistant Hans Geiger develop the Geiger counter to detect radiation levels.

1908 With a commission from the Shah of Persia (Iran) to explore for oil in the region, Englishman William Knox D'Arcy and his team make the first major commercial oil discovery in the Middle East at Masjid-i-Suleiman. D'Arcy then forms the Anglo-Persian Oil Company (later British Petroleum) to exploit this discovery.

1908 Henry Ford revolutionizes the manufacturing process when he sets up the first assembly line in his Model T factory. The inexpensive Model T makes automobile ownership possible for the average American and it becomes the largest-selling car model of its era.

1908 Hermann Minkowski publishes *Space and Time,* proposing time as a fourth dimension for the study of the universe.

1908 The history of superconductors begins with the successful liquefication of helium by the Dutch scientist Keike Kamerlingh Onnes at a temperature of just 4 K.

1908 The Wisconsin Railroad Commission orders gas companies to use British Thermal Units (BTUs) as the new heating measurement standard.

1909 "England's isolation has ended once and for all" is the judgment of one English newspaper after French aviator Louis Bleriot becomes the first to fly across the English Channel. Bleriot will later found an aircraft company that produces thousands of the famous SPAD warplane for the Allies in World War I.

1909 Leo Hendrik Baekeland of Belgium invents bakelite, a dense synthetic polymer (a plastic) that is used to make jewelry, game pieces, engine parts, radio boxes, switches, and many other objects. Bakelite is the first industrial thermoset plastic, i.e., a material that does not change its shape after being mixed and heated.

1909 The Hughes Tool Company makes the first rotary rock bit, which makes it possible to drill through hard rock formations with a rotary rig.

1910–1919

1910 A California gusher shoots forth a huge column of oil. The well, between the towns of Taft and Maricopa, will eventually reach an uncontrolled rate of 100,000 barrels and destroy its derrick. Its estimated output of 9 million barrels of oil is a record for the time, though reportedly less than half of this could be saved.

1910 American engineer Elmer A. Sperry establishes the Sperry Gyroscope Company (with himself as the lone employee) to market the innovative ballistic gyrocompass he had created two years earlier.

1910 Georges Claude displays the first neon lamp in Paris.

ca. 1910 Farms in North America are home to about 25 million horses and mules. These animals supply the power for general transportation and for essential farm tasks such as plowing, seeding, cultivating, and harvesting. Following World War I, the mechanized tractor will take over this role, and on today's farms animal power is a rarity.

1911 Austrian physicist Victor F. Hess discovers cosmic radiation. He is not certain what this phenomenon is, and it is poorly understood even today.

1911 Dissolution of the Standard Oil trust is ordered by the U.S. Supreme Court. John D. Rockefeller's empire is broken up into many smaller companies, including Standard Oil of New Jersey, Standard Oil of New York, Standard Oil of California, Standard Oil of Ohio (later part of BP), Standard of Indiana (later Amoco), Continental Oil (Conoco), and Atlantic Oil (Arco).

1911 Dutch physicist Heike Kamerlingh Onnes discovers superconductivity—the loss of electrical resistance in mercury near absolute zero temperatures. It is soon determined that other metals and alloys become superconductors at very low temperatures.

1911 Ernest Rutherford conducts experiments with the scattering of alpha particles off atoms, from which he concludes that the atom's mass and positive charge are concentrated in a tiny nucleus. The nuclear model of the atom is born.

1911 Famed photojournalist Lewis W. Hine produces his classic "Breaker Boys" series of photos depicting harsh child labor conditions in the United States. Breaker boys are employed in coal mines to separate coal from unwanted stone and slate, a grueling task that requires them to spend the day sitting suspended above a conveyer belt as the coal rushes by.

1911 George von Hevesy of Hungary conceives the idea of using radioactive tracers, which will have great applications in medical and ecological research. He first applies this method to a biological problem in 1923, when he traces lead absorbed by plants.

1911 The U.S. output of kerosene is eclipsed for the first time by a formerly discarded byproduct—gasoline.

1912 C. V. Boys and Frank Shuman's solar-powered steam boiler begins pumping irrigation water from the Nile River. This is the world's largest solar plant and Shuman envisions one even larger, a gigantic facility covering thousands of square miles in the Sahara Desert. World War I intervenes and Shuman's dream is never realized.

1912 George Sarton, a Belgian-born professor at Harvard University, founds *Isis,* which will evolve into the premier international journal dedicated to the history of science. Sarton will edit *Isis* until 1952, and he is widely considered to be the father of the history of science.

1912 The General Motors Company approves the installation of Charles Kettering's automatic self-starter in its Cadillac line of automobiles. This is one of many innovations in auto starting, ignition, and lighting accomplished by Kettering, first with his Delco company and then with General Motors after it acquires Delco.

1912 The world's largest ocean liner, *Titanic,* sinks on her maiden voyage.

ca. 1912 Robert Bosch of Germany develops what is considered to be the first successful fuel injection system.

1912–1914 First Lord of the Admiralty Winston Churchill, fearing conflict with Germany, hastens the conversion of British battleships from coal to oil. Oil-powered ships are faster, they can be refueled at sea, and they require less on-board labor. The Royal Navy eventually overcomes the largely coal-based German fleet in World War I.

1913 Austrian Victor Kaplan invents an improved turbine. The Kaplan turbine features movable blades that rotate, or "feather," to handle different conditions; this allows hydroelectric power stations to operate with greater efficiency.

1913 Canadian engineer Sidney Ells advocates the hot water flotation method of separating bitumen from the tar sands of Athabasca. His is one of several unsuccessful attempts to develop the Athabasca sands. Decades later, oil sands will re-emerge as a viable energy source.

1913 Electric refrigerators for home use begin to appear on the market in the United States.

1913 Frederick Soddy formulates the concept of isotopes, stating that certain radioactive elements may exist in two or more forms which have different atomic weights but which are indistinguishable chemically.

1913 French physicist Charles Fabry discovers the ozone layer, the thin shield in the upper atmosphere that acts as a screen to protect life on the surface of Earth from the ultraviolet rays of the Sun.

1913 Friedrich Bergius, a German chemist, invents a process he calls "destructive hydrogenation" of coal, which later comes to be known as direct liquefaction. Bergius liquefies coal by reacting it with hydrogen at high temperature and pressure.

1913 In the famous "oil drop experiment," Robert Andrews Millikan measures the charge of a single electron to be sixteen-quintillionths of a coulomb.

1913 Irving Langmuir develops an improved incandescent lamp by filling the lamp with an inert gas so that the atoms of tungsten evaporate more slowly. This is the first major advance in the incandescent lamp since the days of Edison, and it remains a standard technology today.

1913 Niels Bohr applies the quantum theory to the atom and concludes that electrons have fixed orbits around the nucleus and that they emit or absorb specific amounts of energy, or quanta, when they jump into another orbit.

1913 The first geothermal power plant is built at Larderello in Tuscany, Italy, employing heat from a dry steam field to generate a constant flow of electricity.

1913 The percentage of the world's energy consumption coming from coal peaks at 61%.

1913 *U.S.S. Jupiter,* a transport and coal carrying vessel, is commissioned in California. This is the first surface ship in the U.S. Navy to be propelled by electric motors.

1913 Using radiometric dating, the English geologist Arthur Holmes proposes the first geological time scale. He estimates the Earth's age at 4 billion years, far older than prevailing wisdom.

1913 William Burton of the Standard Oil Company receives a patent for a thermal cracking process (the application of steady heat and pressure to crack, or break down, heavier hydrocarbons into lighter ones). This greatly increases the proportion of gasoline obtainable from crude petroleum.

ca. 1913 Fritz Haber and Carl Bosch of Germany devise an industrial process in which nitrogen from the atmosphere and hydrogen under high temperature and pressure react to form ammonia. The Haber-Bosch process becomes an effective way to produce ammonia-based fertilizers.

1913–1914 Ford Motor Company installs the first moving assembly line for the manufacture of automobiles. Once the system is in full operation, Henry Ford announces that his workers will be paid $5 per day, more than double the usual wage at the time, while also having their work day reduced. This makes Ford a hero to the American work force.

1914–1918 World War I engulfs Europe. Newer forms of warfare are employed, such as tanks, airplanes, submarines, machine guns, and poison gas. Railroads and motor vehicles transport troops and war matériel. However, military tactics lag behind technology, and in the slaughterhouse of trench warfare millions of lives are lost in fruitless assaults on heavily defended positions.

1914 A version of the modern traffic light is put into operation in Cleveland, Ohio, with red and green lights and a warning sound to indicate when the light is changing. Three-color lights with yellow as the warning will be introduced in 1920.

1914 Acid mine drainage from abandoned coal mines in the United States receives considerable attention when the Army Corps of Engineers begins an extensive investigation of the problem in the Ohio River basin. Municipalities are concerned about the increased costs of water filtration and pipe corrosion.

1914 Data compiled from annual mine inspector reports leads to the conclusion that about 90,000 people have died or suffered injury in the mines of Great Britain in the period from 1850 to 1914. Most of the deaths are in coal mines.

1914 Ernest Rutherford suggests that the positive nucleus of atoms contains protons that balance the charge of electrons but have a much greater mass.

1914 Professor Robert Hutchings Goddard of Worcester, Massachusetts receives two U.S. patents, one for a rocket that uses liquid fuel, the other for a two- or three-stage rocket using solid fuel. This is considered the first step in the era of rocket propulsion.

1914 The first regularly scheduled airline in the United States begins operation. The Benoist Company, using a flying boat, carries passengers across Florida's Tampa Bay, between St. Petersburg and Tampa. The flights will last for only about six months.

1914 The Panama Canal opens, linking the Atlantic and Pacific Oceans and drastically reducing the sailing time from Europe or eastern North America to the West Coast.

1914 Through studies of earthquake waves, Beno Gutenberg determines that there is an inner and outer core at the center of the earth measuring about 4,200 miles in diameter, with a sharp dividing line between the mantle and the outer core (now known as the Gutenberg discontinuity).

1915 German-American blacksmith August Fruehauf builds the first semi-trailer type of truck, in which the load-carrying vehicle has wheels only at the rear and is towed and supported by a powered vehicle in front.

1916 Edward Ardern and William Lockett, workers at a wastewater treatment plant in Manchester, England, demonstrate the activated sludge process, which will become the most widespread technology for wastewater purification.

1916 Einstein publishes his General Theory of Relativity, or Theory of Gravitation, equating gravitational and inertial mass and showing space to be curved.

1916 Gilbert Newton Lewis and Irving Langmuir independently formulate a system of chemical bonding based on the number of electrons in each shell of an atom.

1916 The Goose Creek oil field on Galveston Bay becomes the first offshore drilling operation in Texas and the second in the United States. The Goose Creek area will soon become the site of one of the largest oil refineries in the United States.

1916 Using Einstein's general theory of relativity, Karl Schwarzschild postulates on the existence of what are later named black holes.

1917–1921 Russian oil production and exports drop sharply as a result of the Russian revolution in 1917, and the nationalization of the oil fields by the Communists in 1920. Production strengthens by the mid-1920s due to an infusion of foreign investment from the West.

1918 A U.S. Revenue Act restores the depletion allowance for minerals and fossil fuels, reducing taxes for owners of such resources to compensate for the exhaustion of an irreplaceable capital asset. Critics argue that the allowance is a huge subsidy because these resources are valuable enough to justify high levels of investment without tax incentives.

1918 The American Gas Association is formed through the consolidation of smaller gas organizations. It represents companies involved in all areas of the transmission and distribution of natural gas in the United States.

1918 The first major U.S. gasoline pipeline is established to transport gasoline over a 40-mile distance in Wyoming.

1918 The first well of the Panhandle field north of Amarillo, Texas goes into operation. This will become one of the most productive oil and gas fields in the world, and within a few years Amarillo and other nearby towns of the Panhandle region will double and even triple in size.

1918 U.S. production of anthracite coal peaks at 100 million tons per year. Located primarily in the state of Pennsylvania, anthracite is a relatively clean burning form of coal that enjoys extensive use in home heating. By 2000 production drops to about 4 million tons, due largely to the depletion of the resource base.

ca. 1918 Summing up the role of energy supply in World War I, British Foreign Secretary Lord Curzon is said to remark that "The Allies floated to victory on a sea of oil." (Other versions of this comment are reported, and it has also been attributed to Winston Churchill.)

1918–1920 The air-cooled radial engine comes into widespread use in military aircraft, with cylinders arranged to extend radially outward from its hub, like spokes of a wheel.

1919 Captain John Alcock and Lieutenant Arthur Whitten Brown of Great Britain make the first aircraft flight across the Atlantic, flying from Newfoundland, Canada to Ireland in about sixteen and a half hours. Their plane is a Vickers Vimy, a two-engine biplane.

1919 Less than one month after the transatlantic flight of Alcock and Brown, the British airship *R-34* makes the first two-way trip across the Atlantic, going from Scotland to New York and then back to England.

1919 Observations of a total eclipse of the Sun show a gravitational deflection of light that supports Einstein's general theory of relativity.

1919 The American Petroleum Institute (API) is organized in Chicago, Illinois, for the purpose of standardizing engineering specifications for drilling and production equipment.

1919 The first U.S. tax on gasoline is adopted by the Oregon state legislature to raise money for the maintenance of state highways. By 1929, all states impose a tax on gasoline, and three years later a Federal tax of one cent per gallon is enacted.

1919 The German scientist W. Schroeder makes the first estimate of global net primary production of carbon: 36.2 gigatons (10^{15}). Subsequent estimates are in the range of 100 gigatons.

1920–1929

1920 Isopropyl alcohol becomes the first commercial petrochemical, initially produced at Standard Oil's Bayway, New Jersey, plant.

1920 Radio station KDKA of Pittsburgh receives a broadcasting license and goes on the air with news reports on the U.S. Presidential election. This is the nation's first commercial radio broadcast; by 1925 there are more than 600 commercial stations in the United States and Canada.

1920 The Federal Power Act is enacted in the United States. It creates the Federal Power Commission (FPC) and empowers it to regulate the electric power and natural gas industries, including the licensing of non-Federal hydroelectric projects, the regulation of interstate transmission of electrical energy, and the rates for that electricity's sale.

1920 The number of automobiles in the United States exceeds nine million as gas stations open everywhere. Because of this increased use of gasoline, oil now provides almost one-fourth of the total U.S. energy consumption.

1920s Henry Ford and others promote the use of ethanol-blended gasoline, which they call gasohol. A fermentation plant to manufacture ethanol specifically for motor fuels is built in Atchison, Kansas, and in the 1930s, more than 2,000 service stations in the Midwest sell gasohol. Low oil prices in the 1940s end gasohol use until the 1970s.

1920s The technique of pulverizing coal into a fine powder for burning comes into use. This process brings advantages such as a higher combustion temperature, improved thermal efficiency, and a lower air requirement for combustion.

ca. 1920 From his studies of the chemical composition of gasoline, British engineer Harry Ricardo establishes a measurement system that is the basis of today's standard classification of fuels according to octane rating. Ricardo also makes important improvements to diesel and airplane engines.

early 1920s French engineer Eugene Jules Houdry discovers a revolutionary method for cracking low-grade crude oil into high-test gasoline. In the 1930s he moves to the United States and works with American oil companies to develop the Houdry process as a large-scale commercial enterprise. Later he will also be credited with the invention of the catalytic converter.

1921 A 4,500 metric ton stockpile of ammonium nitrate and ammonium sulfate explodes at a chemical plant in Oppau, Germany. The blast and subsequent fire kill 600 people, injure 1,500, and leave 7,000 homeless.

1921 Albert Einstein is awarded the Nobel Prize in Physics for "his services to theoretical physics, and especially for his discovery of the law of the photoelectric effect." (The citation does not specifically mention Einstein's theory of relativity, though today this is often described as the most important advance in physics since Newton.)

1921 Charles F. Kettering and his assistant Thomas A. Midgley discover that the addition of a slight quantity of tetraethyl lead will reduce "knock" (imperfect burning), which is a common drawback of the gasoline engines of this era.

1921 John D. Grant begins drilling in Geyser Canyon north of San Francisco with the dream of harnessing the steam power there. Grant builds a power plant that is one of the first U.S. facilities to generate electricity from geothermal power.

1921 Regularly scheduled passenger airline service is underway on various routes, such as Paris to London (by several French and British companies), Amsterdam to London (by KLM), Berlin to Weimar, Germany (by a forerunner of Lufthansa) and Los Angeles to San Diego, California (by a company owned by famous film director Cecil B. DeMille).

1921 Without competitive bidding, oil barons Edward Doheny and Harry Sinclair receive exclusive rights to valuable Federal oil reserves at Elk Hills, California and Teapot Dome, Wyoming. The Teapot Dome Scandal ensues, and U.S. Interior Secretary Albert B. Fall is convicted of bribery for getting "loans" from Doheny and Sinclair prior to granting them the rights.

1922 Frederick Soddy, a Nobel laureate in chemistry, publishes a critique of standard economic thought for its lack of consideration of how thermodynamics and other biophysical laws constrain human economic aspirations.

1922 The aircraft carrier *Hosho* enters service in Japan. This is the first ship to be launched specifically as an aircraft carrier (as opposed to a flight deck being grafted on another type of vessel).

1922 The first peer-reviewed, continuously running journal pertaining exclusively to energy is published: *Fuel, The Science and Technology of Fuel and Energy.*

1922 The first well of the Hugoton gas field is discovered in southwestern Kansas. This massive field now includes acreage in Kansas, Texas, and Oklahoma and is defined as the largest natural gas field in North America and the second largest in the world.

1923 Arthur Holly Compton illustrates the dual wave and particle nature of X rays and gamma rays by showing how their wavelengths increase when they collide with electrons.

1923 California ranks as the leading oil-producing state in the United States. The foremost site is Signal Hill (Long Beach), which is the richest field in the world based on output per area. Signal Hill will eventually produce over 1 billion barrels of oil and continues pumping at a rate of 1 million barrels per year into the 1990s.

1923 German chemist Matthias Pier of the BASF corporation develops a process to produce synthetic

methanol. By 1926 U.S. companies such as DuPont will also begin synthetic methanol production.

1923 German coal researchers Franz Fischer and Hans Tropsch discover a method of producing synthetic liquid fuel from coal gas. The Fisher–Tropsch process has significant contemporary applications, such as to convert biomass into fuel or to utilize associated gas at oil fields.

1923 Hermann Oberth writes *The Rocket into Interplanetary Space*, a book that has been described as the first accurate prediction of the events that would unfold in the Space Age. Later, Oberth's pupil Wernher Von Braun will lead the rocket program of Germany, and that of the United States as well.

1923 Russian-American engineer Vladimir Zworykin applies for a patent for his iconoscope, a camera tube for television broadcasting. Based on this and his subsequent kinescope, Zworykin is often described as the originator of U.S. television, but it is not definitely established that he constructed a working system prior to his viewing the work of Philo T. Farnsworth (see 1927).

1923 The first ethyl gasoline in the United States goes on sale at a gas station in Dayton, Ohio. Leaded gasoline becomes the dominant fuel for cars, although it is eventually shown to be a major health hazard, leading to laws that prohibit its manufacture.

1923 The first oil field in Iraq is discovered in the Naft Khana area.

1923 The number of coal miners in the United States reaches a historic high of 705,000. U.S. coal production will reach higher levels in the future but with a reduced labor force, due to mechanization.

1924 A conveyor belt is successfully installed in a Pennsylvania anthracite mine to carry coal to a string of cars at the mine entry.

1924 Dr. H. S. Hele-Shaw and T. E. Beacham of Great Britain develop a variable-pitch propeller; this allows the angle of the propeller blades to be adjusted to suit different flight conditions while the engine speed remains the same.

1924 Scotsman Daniel Dunlop leads the organization of the First World Power Conference in London, among the first international meetings on energy. The name is changed to the World Energy Council (WEC) in 1968. The WEC remains an important global energy organization, with member committees in over 90 countries.

1924 The first seismic discovery of an oil-producing salt dome is made at the Orchard field in Fort Bend County, Texas, using the refraction method.

1924 The first successful oil drilling platform over water is erected by the Lago Petroleum Company in Lake Maracaibo, Venezuela.

1924 U.S. biologist Alfred Lotka describes evolution in energy terms, stating that the advantage goes to organisms whose energy-capturing devices are more efficient in directing available energies into favorable conditions for the preservation of the species.

1925 Robert Millikan coins the term "cosmic rays" to describe the phenomenon discovered in 1913 by Victor Hess. Millikan, Hess, and others will confirm the extraterrestrial origins of this radiation.

1925 Scottish inventor John Logie Baird provides the first true television picture when he transmits live images to a screen in an adjoining room. He uses a mechanical scanning system based on Nipkow's disc of 1884. Baird's first image is the head of a dummy; he then televises the face of his young office assistant.

1925 The Magnolia Gas Company builds the first all-welded pipeline, over 200 miles in length, from northern Louisiana to Beaumont, Texas.

1925 Wolfgang Pauli proposes a fourth quantum number for energy levels of atomic electrons and theorizes that no two electrons with the same quantum number can occupy the same atom. This theory accounts for the chemical properties of elements.

1925–1927 Werner Heisenberg, Max Born, and later Erwin Schrödinger formulate quantum mechanics, a branch of physics that deals with matter at the atomic level. In 1927 Heisenberg states his famous uncertainty principle, i.e., that it is impossible to measure both energy and time (or position and momentum) with completely accuracy at the same time.

mid 1920s The practice of filling in wetland areas near cities with garbage, ash, and dirt becomes a popular disposal method in Europe and North America.

1926 Arthur Wood and David Gray of the Corning, New York Glass Works develop the "Ribbon" machine to make light bulbs. Bulbs can be made at speeds of one thousand per minute.

1926 Britain is paralyzed when, in support of a strike by coal miners over threatened wage cuts, the Trades Union Congress calls a general strike. The strike ends after nine days due to government emergency measures and lack of public support. The miners eventually return to work, accepting lower wages and longer hours.

1926 Fritz Winkler introduces a process for commercial fluidized-bed coal gasification at a BASF plant in Leuna, Germany.

1926 Robert Goddard launches the first rocket propelled by liquid fuel.

1926 The Federal Oil Conservation Board of the United States estimates that the nation has only enough oil to last for another six years.

1926 Georg Ebert of Germany produces a synthetic rubber from butadiene. During the following decade

this evolves into various types of "buna" rubber, so named from the two materials used to make them, butadiene and natrium (sodium).

1927 A huge flow of oil begins at the Baba Gurgur oil fields near Kirkuk, about 140 miles north of Baghdad, marking the beginning of the oil industry in Iraq. Peak flow rates are estimated at 95,000 barrels per day. This is at the ancient site where a natural oil spring known as the Eternal Fire burned for centuries.

1927 British ecologist Charles Elton publishes *Animal Ecology,* in which he outlines the importance of feeding (energy) relationships among organisms as the basis for understanding nature. Elton's perspective is a precursor to modern ecological theory.

1927 Brothers Conrad and Marcel Schlumberger of France develop the technique of wireline logging as a method of obtaining downhole data in oil and gas wells. They commercialize this process and the Schlumberger company becomes a leading provider of petroleum technology services.

1927 Charles A. Lindbergh flies from New York, United States to Paris, France to complete the first solo non-stop flight across the Atlantic Ocean.

1927 George Lemaitre proposes the earliest version of what is now known (and generally accepted) as the "Big Bang" theory of the origin of the Universe.

1927 U.S. inventor Philo T. Farnsworth produces the first successful transmission of a television image by wholly electronic means at his San Francisco laboratory. His "Image Dissector" system differs from John Logie Baird's earlier mechanical transmission system, which by the early 1930s will be supplanted by electronic transmission.

late 1920s The U.S. natural gas industry expands greatly as large new fields are developed in the Southwest and California. Eventually natural gas will completely replace the older technology of manufactured gas in the United States.

1928 A patent is assigned to Marvin Pipkin for an electric light bulb coated on the inside, which is stronger and has less glare than existing outside-coated bulbs. He goes on to develop an improved version of this in 1947.

1928 Station WGY of Schenectady, New York begins the first regularly scheduled television broadcasting. A half hour program is broadcast three days a week.

1928 The first diesel-electric passenger locomotive in the United States is built by the American Locomotive Co. for use on the New York Central Railroad.

1928 The U.S. Federal Trade Commission commences a massive four-year investigation of the nation's private electric companies. The FTC's report runs to 84 volumes and documents a vast number of financial abuses. Utility ownership becomes an issue in the 1932 Presidential election, with Franklin D. Roosevelt arguing for consumer-controlled systems.

1929 Manufactured gas suppliers in the United States annually distribute about 450 billion cubic feet of gas, more than twice the amount delivered per year prior to World War I. Over 50% of this is water gas, about 40% coal gas, and about 6% oil gas.

1929 The first ocean thermal energy conversion (OTEC) facility is built by Georges Claude of France in Mantanzas Bay, Cuba. The plant produces 22 kilowatts of electricity but requires about 80 kilowatts to run its equipment.

1929 The first particle accelerator, originally known as an "atom-smasher," is devised by John Douglas Cockcroft and Thomas Sinton Walton. The following year Ernest Lawrence expands on their idea to create the cyclotron. This leads to studies of the atomic nucleus and the forces holding it together.

1929–1931 The Panhandle Eastern, Northern Natural, and Natural Gas Pipeline Company of America are organized to deliver Southwestern gas to Midwestern markets.

1930–1939

ca. 1930 The human population reaches 2 billion.

1930 British engineer Frank Whittle registers the first patent for a jet propulsion engine. In 1934 Hans von Ohain of Germany, working independently of Whittle, patents a jet engine design similar in concept but different in internal arrangement. Both inventors go on to test their engines successfully in 1937.

1930 German engineer Rudolf Erren presents data on a hydrogen engine at the World Power Conference in Berlin. In the mid-1930s Erren will convert hundreds of gasoline vehicles to run on hydrogen or hydrogen/gasoline mixtures. Apparently none of these engines survived World War II.

1930 The first analog computer is invented by Vannevar Bush to solve differential equations.

1930 The giant East Texas oil field is discovered. It is larger than the next 20 fields in the United States combined and will become the largest oil-producing site in the contiguous United States. The first productive wells in the area are drilled by promoter Columbus Marion "Dad" Joiner.

1930 The Royal Institute of British Architects develops the heliodon to help architects determine the potential effects of the Sun on new buildings.

1930s Edward A. Birge and Chauncey Juday, American aquatic scientists, carry out studies to measure the energy budget of lakes. They develop the idea of primary production, i.e., the rate at which food energy is generated, or fixed, by photosynthesis.

1930s German architects lead a renaissance in solar building during the 1930s that spreads throughout Europe; many of these architects later flee to the United States when the Nazis condemn functional architecture as Jewish.

1930s The DuPont Company begins the manufacture of Freon (chlorofluorocarbon or CFC), a substance that does not carry the hazards of earlier refrigerants. Freon leads to much greater consumer use of refrigerators and air-conditioners. In the 1990s Freon will be banned because of concerns over its role in the depletion of the ozone layer.

1930s The first windowless buildings are built, reflecting the growing dependence on electric lighting in architecture.

1930s The Texas Railroad Commission begins to regulate oil drilling, production, and prices in Texas, in an effort to damp the boom-and-bust nature of the industry. The TRC will restrain production and thus establish a price floor. Other states follow the TRC's lead. The power of TRC wanes in the 1970s as production begins to decline.

1931 André Clavier demonstrates the potential value of microwaves when he uses them for a radio transmission across the English Channel.

1931 Auguste Piccard and his associate Paul Kipfer set an altitude record by ascending to almost 52,000 feet in a balloon. This can be considered the first successful flight into the stratosphere; it is one of several flights above 50,000 feet made in the 1930s by Auguste Piccard and his twin brother Jean.

1931 Ernst Ruska of Germany designs the first electron microscope; he then builds a working model of this in collaboration with his mentor Dr. Max Knoll.

1931 French engineer Georges J. M. Darrieus patents a new type of wind turbine with a distinctive "eggbeater" or "skipping rope" design. This becomes known as a vertical axis wind turbine (VAWT).

1931 Harold Hotelling of the United States publishes the economic theory of the optimal depletion of nonrenewable resources, which is still the dominant paridigm today in economics.

1931 U.S. chemist Harold Clayton Urey devises a method for the concentration of any possible heavy hydrogen isotopes by the fractional distillation of liquid hydrogen; this leads to the discovery of deuterium.

ca. 1931 English geologist Arthur Holmes provides the energetic explanation of continental drift. Currents of heat and thermal expansion in the Earth's mantle, he suggests, drive the continents toward or away from one another, creating new ocean floors and building mountain ranges.

ca. 1931 The Lurgi process for coal gasification is developed in Germany.

ca. 1931 U.S. physicist Robert Jemison Van de Graaff introduces his electrostatic generator, which is capable of producing extremely high voltages. It is a more efficient particle accelerator than the earlier Cockroft-Walton machine. Van de Graaff's "atom smasher" will play an important role in the early years of nuclear and particle physics.

1932 After years of speculation, the existence of the neutron in an atom is demonstrated by James Chadwick. Based on this discovery, Werner Heisenberg proposes that an atom nucleus is composed of protons and neutrons, explaining the existence of isotopes.

1932 British engineer Francis Thomas Bacon develops the first practical modern fuel cell. Bacon (a descendant of scientific pioneer Sir Francis Bacon) continues working on this device and eventually produces an effective working model in 1959.

1932 Hermann Honnef of Germany proposes a multiturbine wind generator 160 meters wide that would be capable of 20 megawatt output. He urges the Nazi regime to adopt it, but the device is never built.

1932 John Cockcroft and E. T. S. Walton of Great Britain split the atom on their linear accelerator in Ernest Rutherford's Cavendish Laboratory at Cambridge University. Their experiment substantiates Einstein's theory of relativity.

1932 The positron, a positively charged electron and the first form of antimatter to be discovered, is detected in cosmic-ray tracks by Carl David Anderson.

1933 A team led by Reginald Gibson and Eric William Fawcett discover polyethylene, one of the earliest plastics to come into common use.

1933 Edwin Armstrong demonstrates his frequency modulation (FM) system for the almost static-free transmission of radio waves; FM is one of many fundamental advances in radio broadcasting credited to Armstrong.

1933 King Abdel Aziz ibn Saud of Saudi Arabia establishes an agreement with Standard Oil of California (Chevron) to develop the oil resources of the country's Eastern Province. In 1936 Texas Oil (Texaco) joins the venture and it is named the Arabian-American Oil Company, or Aramco. It grows to become the largest oil company in the world.

1933 The Texas Company introduces the submersible drilling barge for oil exploration on the Gulf Coast of Louisiana; this allows drilling to be done on the bottom

of coastal marshes without the need to sink pilings and build a platform.

1933 The U.S. Congress passes legislation to create the Tennessee Valley Authority (TVA), beginning a period of large hydropower projects in the eastern United States. Electricity provided by the TVA will be crucial to the U.S. industrial effort in World War II.

1934 Enrico Fermi of Italy irradiates uranium with neutrons and believes he has produced the first transuranic element. Later it will be recognized that the reaction produced by this neutron bombardment actually is atomic fission.

1934 Husband and wife team Frederic and Irene Joliot-Curie develop the first artificially radioactive element, an isotope of phosphorus.

1934 The Technocracy Movement in the United States argues that the antidote to the Great Depression is to replace politicians with scientists and engineers who have the technical expertise to manage the nation's economy and natural resources. Technocrats use growth and decline curves to predict a wide range of societal trends.

1935 British scientist A. G. Tansley proposes the word *ecosystem* as a descriptive term for the fundamental ecological unit.

1935 Charles Richter invents a scale to measure the strength of earthquakes.

1935 Scottish engineer Robert Alexander Watson-Watt (a descendant of James Watt) uses microwaves to devise a radio detection and ranging system, better known as radar. By the outbreak of World War II Britain has a chain of radar stations along the east and south coast of England.

1935 The first era of electric vehicles in the United States ends when Detroit Electric, the best known company in the field, stops production after 30 years in business (though orders will still be taken into the early 1940s). Reportedly Clara Ford, wife of Henry Ford, drove a Detroit Electric car rather than a Ford because she disliked gasoline-powered cars.

1935 The first isotope of uranium, U-235, is discovered by Arthur Jeffrey Dempster.

1935 The General Electric Company introduces the first fluorescent lamp.

1935 The U.S. Bureau of Reclamation authorizes funding for California's Central Valley Project, a long-term plan for use of the water of the Sacramento River basin.

1935 Wallace Carothers of DuPont discovers a strong polyamide fiber that resists both heat and solvents; this is named Nylon and it quickly comes into widespread use, most notably for women's stockings and then later in many military applications.

1936 American astrophysicist Charles Greeley Abbott invents an efficient solar boiler.

1936 British mathematician Alan Turing publishes a paper describing an abstract device (Turing machine), which "can be made to do the work of any special-purpose machine, that is to say to carry out any piece of computing, if a tape bearing suitable 'instructions' is inserted into it." This is often described as a prediction of the modern computer.

1936 Hoover Dam (Boulder Dam) is completed at Black Canyon on the Colorado River between Arizona and Nevada, one of the major engineering feats of modern times and the source of much of the electric power consumed in the Southwest.

1936 Japanese television pioneer Kenjiro Takayanagi builds an all-electronic television set with 245 scanning lines. Earlier Takayanagi had transmitted an image on a Braun electric tube (1926) and had demonstrated a cathode ray tube system (1928).

1936 The airship *Hindenburg* becomes the largest aircraft ever to fly, at more than 800 feet long and 135 feet in diameter. After making several transatlantic trips, the *Hindenburg* catches fire on May 6, 1937 while making a landing at Lakehurst, New Jersey. Thirty-five people are killed and this tragedy effectively ends the use of airships for passenger transportation.

1936 The Rural Electrification Act is signed into law by President Franklin D. Roosevelt to address the fact that rural America is still largely without electricity. Within two years cooperative projects financed by the REA will provide electricity to 1.5 million farms, and by the mid-1950s virtually all American farms will be electrified.

1937 Carl David Anderson discovers the muon, a subatomic particle larger than an electron but smaller than a proton.

1937 Lockheed Aircraft produces the first airplane with a fully pressurized cabin, a modified version of its *Electra* model.

1937 Theodor Dobzhansky publishes *Genetics and the Origin of the Species,* providing the link between Darwin's evolutionism and Mendel's theory of mutation in genetics.

1938 Commercial quantities of oil are discovered at the Burqan field in southern Kuwait. In the same year Aramco makes a major discovery at the Dammam field in Saudi Arabia.

1938 Congress enacts the Natural Gas Act, giving the Federal Power Commission the authority to regulate commercial activity in the U.S. natural gas industry.

1938 G. S. Callendar publishes *The Artificial Production of Carbon Dioxide and Its Influence on Climate,* in which he directly implicates fossil fuel combustion as an agent of climate change. He is largely ignored or derided for this claim, but history proves him to be prophetic.

1938 Otto Hahn and his associates Lise Meitner and Fritz Strassman of Germany are the first to split a uranium atom, although they hesitate to call their success nuclear fission.

1938 The Bonneville Dam is completed on the Columbia River about 40 miles of Portland, Oregon, and begins supplying power to the Pacific Northwest.

1938 The British steam locomotive *Mallard* reaches a speed of 126 miles per hour (201 kph) running on a downward grade in Lincolnshire, England. This still stands as the world speed record for a steam locomotive.

1939 Franklin D. Roosevelt receives a letter from Albert Einstein (written with Leo Szilard) that describes the possibility of an extremely powerful uranium bomb. Einstein calls for "watchfulness, and if necessary, quick action on the part of the Administration." He urges FDR to provide government support for Fermi and colleagues in their work on chain reactions.

1939 Hans Bethe recognizes that the fusion of hydrogen nuclei to form deuterium releases energy. He suggests that much of the energy output of the Sun results from nuclear fusion reactions.

1939 John Randall and Henry Boot develop a resonant-cavity magnetron, providing a microwave transmission system for radar that is a significant advance on the existing meter-wave system.

1939 The first jet airplane takes flight with a turbojet engine developed by Hans von Ohain of Germany. This is followed by a 1941 flight by a British jet developed by Frank Whittle.

1939 The first practical gas turbine used to generate electricity operates at Neuchatel, Switzerland, developed by the Brown Boveri Company.

1939 The Packard Motor Car Company of Detroit engineers the first air-conditioned automobile.

1939 The *VS-300,* the first workable single-rotor helicopter, is produced by Russian-American aeronautical engineer Igor Sikorsky, following other important advances made by French aviation pioneers Louis Breguet and Rene Dorand (1935) and by Germans Heinrich Focke and Gerd Achgelis (1936). Sikorsky's craft makes its first free flight in 1940.

1939–1942 John Vincent Atanasoff of Iowa State University collaborates with his graduate student Clifford E. Berry to develop the first digital electronic computer. Though the better-known ENIAC machine (see 1946) is often given this honor, a 1973 court decision found that the design of ENIAC was based on the earlier work of Atanasoff.

1939–1942 World War II rages around the globe. As in World War I, new forms of warfare come into use, such as large-scale bombing of cities, combined aerial and ground assaults (blitzkreig), aircraft carriers, rocket bombs, jet planes, and most notably, nuclear weapons. New technologies include radar, sonar, battlefield radios, and electronic coding/codebreaking devices.

1940–1949

1940 Martin Kamen and Samuel Ruben, chemists at the University of California, Berkeley, use a cyclotron to bombard a graphite target with a beam of deuterium nuclei. The result is carbon-14, which, with a half-life of 5,000 years, becomes a powerful tool to trace carbon's movement through photosynthesis and many other biochemical processes.

1940 The American Association of Oilwell Drilling Contractors is founded in Dallas, Texas; it is now known as the International Association of Drilling Contractors (IADC).

1940 The first residential use of geothermal energy for space heating begins in the Moana area of Reno, Nevada.

1940 W. G. Templeman observes the effects of naphthalacetic acid, a highly selective compound that is toxic to broad-leafed plants but less harmful to grasses. This leads to the development of the herbicide 2,4-D in 1944, which revolutionizes weed control in cereal plants.

1940s Important new industrial materials are developed, such as polyethylene (electrical insulation and food packaging), silicones (lubricants, protective coatings, and high-temperature electronic insulation), and epoxy (a very strong adhesive).

1941 After Japan seizes part of Indochina, the United States responds by freezing Japanese assets. This leads to sanctions against the sale of American oil to Japan (at the time about 75% of Japan's oil supply came from the United States). Following this the Dutch government in exile in London bars the sale of oil to Japan from the Dutch East Indies (Indonesia).

1941 Having been denied access to oil supplies from the United States and the Dutch East Indies, the Japanese government concludes that if existing conditions continue they will run out of oil within a short time. This is cited by historians as probably the chief reason for Japan's decision to launch a surprise attack on the U.S. naval base at Pearl Harbor (December 7th).

1941 Nazi Germany surprisingly invades the Soviet Union, after having formed a nonaggression pact with the Soviets just two years earlier. The primary military objective of this operation is to capture the valuable Caucasus oil fields, since Germany's own supply is relatively meager and dependent on synthetic oil.

1941 On a hilltop in Rutland, Vermont known as "Grandpa's Knob," a 1.25 megawatt wind generator provides power to the local grid during the World War II era. This is considered the first use of wind energy to supply public power on a large scale.

1941 Plutonium, a new element, is identified by chemist Glenn Seaborg and his team at the University of California, Berkeley. Seaborg will discover 10 atomic elements in all, including one that now bears his name. Forms of plutonium become the main explosive ingredient of nuclear bombs and an important energy source for nuclear power reactors.

1941 Russell Ohl of the Bell Laboratories (United States) discovers that semiconductors can be "doped" with small amounts of foreign atoms to create interesting new properties. He develops the first silicon solar cell that converts available light into electrical energy.

1941 The first commercial plant for liquefied natural gas (LNG) is built in Cleveland, Ohio.

1942 A team of inventors at Exxon's Research and Engineering Company develop the technique known as fluid catalytic cracking, a process that now produces over half of the world's gasoline.

1942 Aquatic scientist Raymond Lindeman studies Cedar Creek Bog, a senescent lake in Minnesota, and describes the area in a classic paper titled "The Trophic-Dynamic Aspect of Ecology" (published after his death at the age of 27). Lindeman's analysis establishes trophic dynamics as the energetic basis of modern ecology.

1942 General Leslie R. Groves is appointed to head what becomes known as the Manhattan Project, the top-secret effort by the United States to build an atomic bomb before the Germans can do so.

1942 On the Baltic coast of Germany at Peenemunde, Wernher von Braun launches the A-4 rocket. Breaking the sound barrier, it reaches an altitude of 60 miles. This is the world's first launch of a ballistic missile and the first rocket to ascend to the fringes of space.

1942 Swedish physicist Hannes Alfven argues that an electromagnetic wave can propagate through a highly conducting medium. This discovery eventually forms the basis of magnetohydrodynamics (MHD), which studies the motions of electrically conducting fluids, such as plasmas and liquid metals, and their interactions with magnetic fields.

1942 The Grand Coulee Dam is completed in Washington, United States to harness the power of the great Columbia River. It is the largest hydroelectric dam in the world at this time (in fact the largest concrete structure of any kind), and still remains the largest dam in North America.

1942 The Petroleum Administration for War (PAW) is established to assure "the most effective development and utilization of petroleum in the United States and its territories and possessions." The PAW rations civilian petroleum use to insure adequate military supplies. The PAW is terminated in 1946.

1942 Ulrich Hütter of Germany writes a graduate thesis in which he provides the first statement of the theoretical basis for modern wind turbines.

1942 Under the football stands of Stagg Field at the University of Chicago, Enrico Fermi leads a team of scientists in successfully creating the first controlled nuclear chain reaction. This development will lead directly to the assembly and detonation of an atomic bomb.

1942–1944 Because transporting oil by sea is endangered by German submarines, the United States constructs the Big Inch and Little Big Inch pipelines to carry oil from East Texas to the northeastern United States, a distance of about 1,500 miles.

1943 A patent dispute stretching over a half century is finally resolved when the U.S. Supreme Court strikes down the Guglielmo Marconi patent of 1904 for the invention of radio in favor of Nikola Tesla's 1897 patent.

1943 DDT, a powerful and effective pesticide, comes into widespread use. At first DDT is hailed as a miraculous advance and its developer, Paul Müller of Switzerland, wins the Nobel Prize in 1948. However, DDT later is targeted as harmful to wildlife (see 1962), and eventually its use will be widely banned.

1943 Los Angeles, California experiences its first recognized episode of smog as visibility is reduced to three blocks and people suffer from eye irritation, breathing difficulty, and nausea. Fearful citizens at first believe the city is undergoing a gas attack, but eventually a wartime synthetic rubber plant is identified as the pollution source.

1943 The Rockefeller Foundation establishes the precursor to the International Maize and Wheat Improvement Center in Mexico, providing the research thrust on high-yield variety crops that will eventually launch the Green Revolution.

1943 The Soviet Union sets up a nuclear research program under the direction of physicist Igor Kurchatov, with the goal of building an atomic bomb.

1943 The U.S. Army Air Force begins a series of large-scale raids on Ploesti, Romania, a city that Sir Winston Churchill calls "the taproot of German might" because it is where the Nazi war machine obtains much of its oil. When the Ploesti refineries are finally rendered inoperative in late 1944, there is virtually no fuel left for the German military.

1943 U.S. anthropologist Leslie White proposes that like biological systems, cultures evolve around the strategies that they use to exploit energy.

1944 Based on experiments with pneumococci, Oswald T. Avery determines that DNA is the material carrying genetic properties in nearly all living organisms.

1944 In *FPC v. Hope Natural Gas Co.,* the U.S. Supreme Court rules that a regulatory body may fix rates, provided the return to the utility company is "just and reasonable" to allow it to operate successfully and compensate investors for the risks they assume. *FPC v. Hope* becomes the basis for most U.S. court decisions on energy pricing regulation.

1944 Massachusetts Institute of Technology professor Maria Telkes designs a portable, solar-powered seawater distiller for the life rafts of pilots shot down over the ocean. Telkes will develop many solar-powered devices, including the first residential solar heating system in 1948.

1944 President Franklin D. Roosevelt writes to Winston Churchill that the U.S. concern with maintaining a sufficient oil supply is not a sign that he is "making sheep's eyes (looking with envy) at your oil fields in Iraq or Iran." Churchill responds in kind, "We have not thought of trying to horn in upon your interests and property in Saudi Arabia."

1944 Tennessee Gas and Transmission Company completes a pipeline to bring natural gas from Texas to Appalachia to fuel wartime industry.

1944 The first full-scale nuclear reactor begins operation at the U.S. Army's Camp Hanford (Richland, Washington). Hanford is part of the Manhattan Project and produces plutonium for use in an atomic bomb.

1944 The Shasta Dam is completed in northern California after six years of construction. The waters of the Sacramento, McCloud, and Pit rivers, plus dozens of smaller streams, back up to create Lake Shasta.

1944 Wernher von Braun's A-4 rocket is renamed the V-2 and converted to a combat weapon. The V is for "*Vergeltung*" (revenge), since this is Hitler's retaliation against Allied bombing of German cities. When the first V-2 hits London, von Braun is said to have remarked, "The rocket worked perfectly, except for landing on the wrong planet."

1945 The Atomic Age begins, and the U.S. Manhattan Project reaches its goal, on July 16th at a test site at Alamogordo, New Mexico. A team of scientists headed by J. Robert Oppenheimer detonate the world's first nuclear explosive.

1945 The atomic bomb nicknamed Little Boy is dropped on Hiroshima, Japan on August 6th. Three days later, another bomb, Fat Man, is dropped on Nagasaki, Japan. The first bomb is uranium-235 based and the other is plutonium-based. Japan surrenders on August 15, ending World War II.

1945 The end of gasoline rationing in the United States is announced after the surrender of Japan; this had been in place since 1942. In Britain fuel rationing will not end until 1950.

1945 The Food and Agriculture Organization (FAO) is formed by the United Nations, with a mandate to raise levels of nutrition and standards of living, improve agricultural productivity, and better the condition of rural populations.

1946 Architect Arthur Brown designs the world's first solar-heated public building, the Rose Elementary School in Tucson, Arizona. It obtains over 80% of its heat from solar energy for the next decade.

1946 In return for the withdrawal of Soviet troops occupying its northwestern provinces, Iran makes a long-term agreement to provide oil to the Soviet Union. However, the following year the Iranian parliament nullifies this agreement.

1946 John W. Mauchly and J. Presper Eckert demonstrate an electronic digital computer at the University of Pennsylvania. ENIAC (Electronic Numerical Integrator and Computer) uses punched cards for input and output data, contains 17,000 vacuum tubes, weighs over 30 tons, and occupies 1,500 square feet of space.

1946 The Magnolia Petroleum Company (later part of Mobil) builds an oil drilling platform in the Gulf of Mexico off the coast of Louisiana, considered the first oil exploration effort in open ocean waters. It is a dry hole, but the following year Kerr-McGee brings in a producing well in the same area.

1946 The U.S. Atomic Energy Act is signed into law by President Harry Truman. It establishes the Atomic Energy Commission (AEC), a civilian body to control nuclear weapons development and explore peaceful uses of nuclear energy.

1946 The United States carries out the first subsurface detonation of a nuclear weapon at Bikini Atoll in the Pacific, followed by demonstrations in Times Square, New York against nuclear testing.

1947 Barium titanate is the first piezoelectric (electricity produced by mechanical pressure) ceramic to be used commercially, in its application as a phonograph needle.

1947 Cecil Frank Powell discovers the pion, a particle with mass between that of electrons and protons that interacts with neutrons and protons in the atomic nucleus. It is believed to be the meson-particle predicted by Hideki Yukawa 12 years earlier.

1947 The coal industry of Great Britain is completely nationalized under the auspices of the National Coal Board.

1947 The *Grandcamp,* a barge loaded with fertilizer grade ammonium nitrate, catches fire and explodes, destroying a nearby city and killing an estimated 570 people in what would later become known as the "Texas City Disaster."

1947 The microwave oven is introduced by Raytheon, which developed the device after one of the company's engineers, Dr. Percy Spencer, accidentally discovered that the candy bar in his pocket had melted due to microwave energy. Microwave ovens for the home are introduced in the early 1950s and by the mid-1970s their sale will surpass that of gas ranges.

1947 The Texas Eastern Transmission Corporation purchases the Big Inch and Little Big Inch and begins the process of converting them from oil to natural gas.

1947 The U.S. Atomic Energy Commission establishes the Brookhaven National Laboratory, a major research facility on eastern Long Island, New York.

1947 U.S. Air Force test pilot Charles E. "Chuck" Yeager flies the Bell X-1 rocket plane faster than the speed of sound, becoming the first pilot to break the sound barrier.

1947 Willard Frank Libby introduces a method of dating archaeological objects by determining the concentration of radioactive carbon-14 in the item of interest.

1948 Carl Nielsen of Ohio State University is credited with building the first ground-source heat pump.

1948 George Gamow, Ralph Alpher, and Robert Herman formulate the Big Bang theory, explaining the origin of the universe in terms of a huge explosion producing a burst of energy.

1948 John Bardeen, William Shockley, and Walter Brattain, scientists at the Bell Telephone Laboratories in Murray Hill, New Jersey, demonstrate the first transistor, a device composed of semiconductor material that can both conduct and insulate. By 1952, transistors are being used in radios and hearing aids.

1948 Norbert Weiner publishes *Cybernetics,* a detailed account of the mathematics of communication and a forerunner of computer science.

1948 The basic-oxygen steelmaking process comes into use in Switzerland (Linz-Donawitz process).

1948 The first nuclear reactor in France, the Zoe, begins operation at Fontenay-aux-Roses, using uranium from the Belgian Congo.

1948 The first successful continuous miners are introduced in U.S. underground coal mining. These machines break up the coal with huge bits or teeth, eliminating the need for blasting, and then load the broken coal directly into shuttle cars.

1948 The Ghawar Field is discovered in Saudi Arabia; it is the largest conventional oil field in the world (about 80 billion barrels).

1948 Twenty people die when a "killer fog" hovers over Donora, Pennsylvania. About 7,000 others are hospitalized or become ill, about half the population of the entire town. This is described as the first incident to make Americans generally aware of the health hazards of air pollution.

1949 A major advance is made in the measurement of time with the development of the first atomic clock, designed by Dr. Harold Lyons of the National Bureau of Standards in 1949, using a molecule of ammonium.

1949 The first natural gas deliveries to the New York City area arrive at Staten Island through the New York and Richmond Gas Company.

1949 The Soviet Union's first atomic bomb (known as Joe-1) is detonated at Semipalatinsk, Kazakhstan. It is a copy of the Fat Man bomb that the United States dropped on Nagasaki, Japan, with a yield of 21 kilotons. The nuclear arms race between the United States and the Soviets is now underway.

1950–1959

1950 Alan Turing proposes the Turing Test as a means of determining whether a machine can be described as intelligent. He suggests that computers should be able "learn" from experience and then modify their activity accordingly; this provides the basis for the field of artificial intelligence.

1950 Great Britain's first plutonium production reactor, Windscale's Number 1 Pile, begins operation.

1950 Oil begins to flow through the Trans-Arabian Pipeline (Tapline). Oil fields in Saudi Arabia are now linked with the Mediterranean Sea at Sidon, Lebanon, over 1,000 miles away. The pipeline greatly reduces the number of tankers required to transport oil around the Arabian Peninsula through the Suez Canal to the Mediterranean.

1950 Russian scientists Andrei Sakharov and Igor Tamm present an idea for a controlled thermonuclear fusion reactor, the Tokamak (abbreviated from the Russian term for Toroidal Chamber with Magnetic Coil). The Tokamak design becomes one of the most popular in subsequent fusion energy research.

1950 The first European offshore oil well is drilled 1.5 miles from land in the North Sea at the mouth of the Elbe River in Cuxhaven, Germany.

ca. 1950 Oil has replaced coal as the dominant fuel in the U.S. economy, as the discovery of cheap, abundant oil and the development of the internal combustion engine cause the coal-fired steam engine to be abandoned. Coal thereafter is used almost exclusively in power generation and steel production.

early 1950s Home ownership of a television set becomes commonplace in the United States.

1951 At the National Reactor Testing Station in Arco, Idaho, the Experimental Breeder Reactor I (EBR-1) becomes the first nuclear reactor in the world to produce useable quantities of electric power, lighting four 100-watt light bulbs.

1951 Mauchly and Eckert follow up on their ENIAC computer by creating UNIVAC I, the first electronic computer to store data on magnetic tape and to be mass-produced.

1951 Philip Edwin Ohmart of Cincinnati, Ohio invents the radioactive cell, a battery that converts nuclear energy to electrical energy.

1951 The first full-fledged offshore drilling in the Persian Gulf is carried out in 19 feet of water, 3 miles from shore, and 20 miles south of Ras del Misha'ab.

1951 The first mechanized U.S. longwall mining system is in use; longwall is a high-extraction method in which a large rectangular block of coal is isolated and then extracted in a single continuous operation by an automated cutting head moving parallel to the coal face.

1951 The Iranian parliament votes to nationalize Iran's oil industry; oil production then comes to a virtual standstill as British technicians leave the country. In response to Iran's action, Britain imposes a worldwide embargo on the purchase of Iranian oil, freezes Iran's sterling assets, and bans the export of goods to Iran.

1952 A group led by Edward Teller develops and explodes a hydrogen bomb at the Eniwetok atoll, in the Marshall Islands of the Pacific. This is the first thermonuclear device to employ the enormous power of nuclear fusion. It is described as 700 times more powerful than the atomic bomb dropped on Hiroshima in 1945.

1952 Britain's De Havilland Comet becomes the first commercial jetliner, with a flight from London to South Africa. However, crashes due to metal fatigue then cause the plane to be grounded for several years, and by the late 1950s two U.S. rivals, the Boeing 707 and Douglas DC-8, take over the lead in commercial jet service.

1952 Canada's Chalk River nuclear reactor experiences a potential calamity when human error causes an unintended reaction. The resulting explosion blows the four-ton lid off the reactor, spurting radioactive water and setting off warnings of lethal radiation levels. Fortunately the building is evacuated and there are no casualties.

1952 The first offshore well in California with no connection to land is drilled, in 45 feet of water 1.5 miles offshore from Seal Beach.

1952 The Great Smog of London occurs after an exceptionally cold winter forces homes and factories to burn large quantities of coal. A temperature inversion forms, trapping pollutants above the ground. More than 4,000 people die from respiratory ailments within the following week. This leads to major reforms in air quality controls for Britain.

1952 The keel for the U.S. Navy's first nuclear submarine, *Nautilus,* is laid at Groton, Connecticut. *Nautilus* is launched in 1954 and will set many records for submerged speed and distance traveled.

1953 Britain and the United States take part in Operation Ajax, a plan to oust nationalistic Iranian Premier Mohammed Mossadegh. The coup succeeds and the pro-Western Reza Shah Pahlavi establishes a new oil consortium that restores Iran's oil to the West. The United States then replaces Britain as the major foreign player in the Iranian oil industry.

1953 Discovery of the Great Global Rift in the Atlantic Ocean by Maurice Ewing and Bruce Charles Heezen establishes the study of plate tectonics and revolutionizes geology.

1953 Eugene P. Odum of the University of Georgia publishes *Fundamentals of Ecology,* with the aid of his brother Howard. This is the first textbook in the field; it employs a "top-down" approach starting at the ecosystem level and describes energy as a currency shared among the Earth's various life forms.

1953 James Watson of the United States and Francis Crick of Britain describe the double-helical structure of DNA, leading to a revolution in the science of biology.

1953 President Dwight D. Eisenhower, in a famous "Atoms for Peace" speech, proposes to the General Assembly of the United Nations the creation of an organization to promote the peaceful use of nuclear energy; this eventually leads to creation of the International Atomic Energy Agency.

1953 Prime Minister Winston Churchill authorizes the development of Calder Hall in West Cumbria, a complex of four reactors that is the world's first commercial nuclear power facility. Queen Elizabeth officially opens the first of these in 1956.

1953 The British begin nuclear testing at the Maralinga nuclear test site, a tract of Aboriginal land in the state of South Australia. The area will later be the target of a massive clean up campaign.

1953 The deepest known well dug by means of cable tool drilling, normally used for wells a few hundred feet deep, is completed when the New York Natural Gas Corporation drills a well to a depth of 11,145 feet.

1953 The EBR-I reactor demonstrates the feasibility of the breeder reactor concept; a team led by Walter Zinn shows that uranium can be "bred," or transformed, into plutonium, thus making it possible to use virtually all of the energy in natural uranium.

1953 The first gas-turbine, propane-fueled locomotive is placed in service by the Union Pacific Railroad.

ca. 1953 Architect Frank Bridgers designs the Bridgers-Paxton Building in Albuquerque, New Mexico; this is the first office building to use passive solar heating.

1954 Charles Townes and Arthur Schawlow invent a highly precise timing device called the maser (microwave amplification by stimulated emission of radiation). This is a forerunner of the laser, which Townes and Schawlow describe in theory in 1958 but do not actually produce (see 1960).

1954 Gerald Pearson, Calvin Fuller, and Daryl Chapin, researchers at New Jersey's Bell Laboratories, develop a method of producing electricity directly from sunlight, using an array of silicon-based solar cells.

1954 President Dwight Eisenhower signs the first major amendment of the original Atomic Energy Act, giving the civilian nuclear power program further access to nuclear technology.

1954 Researchers at RCA describe the photovoltaic effect in cadmium sulfide; eventually a large segment of the solar energy industry will use cadmium in its photovoltaic modules.

1954 South African Synthetic Oil Limited begins operation of SASOL One, the first civilian coal-to-fuel plant in the world.

1954 The Association for Applied Solar Energy (AFSE) is founded in the United States to encourage the use of renewable energy. The organization later becomes known as the American Solar Energy Society (ASES).

1954 The first purpose-built submersible drilling unit is used in the Gulf of Mexico by Odeco (Offshore Drilling and Exploration Company).

1954 The Soviet Union begins operation of the first nuclear reactor built for the peacetime production of power at the Obninsk plant. It is known as AM-1, with the letters being a Russian-language acronym for "peaceful atom."

1954 The U.S. Supreme Court rules that interstate natural gas producers are subject to regulation by the Federal Power Commission. An era of regulated gas prices is established, based on the cost of providing gas rather than its market value. This keeps prices low, encouraging demand, but it discourages exploration and thus contributes to shortages in the 1970s.

1954 UNESCO (United Nations Educational, Scientific and Cultural Organization) and the government of India sponsor a symposium on solar and wind power in New Delhi.

1955 Arco, Idaho, population about 1,300, becomes the first town in the United States to receive its entire supply of power from a nuclear source. The experimental boiling water reactor BORAX-III produces approximately 2,000 kilowatts of electrical power for about two hours.

1955 Geneva, Switzerland hosts the first United Nations International Conference on the Peaceful Uses of Atomic Energy. About 1,500 delegates attend, representing 73 countries.

1955 Lewis Strauss, chairman of the U.S. Atomic Energy Commission, predicts that electricity produced by nuclear power will soon be so abundant that "It is not too much to expect that our children will enjoy in their homes electrical energy too cheap to meter."

1955 The Atomic Energy Basic Law is passed in Japan, limiting the use of nuclear technology to peaceful purposes. Japan imports its first commercial nuclear power reactor from the United Kingdom and begins power production in 1966. By 2000, Japan gets about one-third of its power from nuclear plants.

1955 The first World Symposium on Solar Energy is held in Phoenix, Arizona, where engineers and scientists from around the world assemble to review the current state of the field. Accompanying this meeting is an exhibition, "The Sun at Work," attended by about 30,000 people.

1955 W. Fred Cottrell, a sociologist at Miami (Ohio) University, is the first to comprehensively apply the concept of net energy and energy surplus to the analysis of economic and social development.

mid-1950s Many factories now employ electrostatic precipitators to filter out some pollutants from smokestack emissions before they reach the atmosphere.

mid-1950s Various workers in the United States and Europe report that bundles of glass fiber can be used for image transmission. One of them, Indian-born researcher Narinder Singh Kapany, reportedly coins the phrase "fiber optics" to describe this technology.

1956 At the urging of President Eisenhower, the U.S. Congress enacts the Federal Aid Highway Act, a program of interstate highway construction that greatly expands the U.S. network of modern roadways. Originally conceived as a system of military transport, its eventual effect will be to ensure the demise of the railroad as a major form of interstate transportation.

1956 British engineer Christopher Cockerell invents the Hovercraft, a vessel that travels over water on a cushion of air.

1956 Celilo Falls, a traditional Indian salmon fishing site, is flooded by the waters of The Dalles Dam on the Columbia River. This leads to controversy over the conflict between tribal rights and hydropower development.

1956 French engineer Felix Trombe develops the Trombe wall, a solar collector consisting of a south-facing glass wall with an air space between it and a blackened concrete wall. The blackened wall is heated by sunlight passing through the glass, and it then releases heat,

causing warm air to rise in the space between the glass and the concrete.

1956 The Atomic Energy Commission (AEC) authorizes construction of the world's first two large, privately owned nuclear power plants, one at Indian Point, New York and the other in Grundy County, Ilinois.

1956 The first transatlantic telephone cable is laid between Newfoundland and Scotland; this links the phone networks of Canada and the United States with the United Kingdom, and, via London, with various nations on the European continent.

1956 The National Academy of Sciences in the United States recommends deep geologic disposal of the long-lived, highly radioactive wastes from nuclear reactors, suggesting that buried salt deposits and other rock types be investigated for permanent repositories.

1956 The Suez Crisis erupts as Israel, supported by Britain and France, attacks Egypt after that country nationalizes the Suez Canal. The crisis ends when the United States and the Soviet Union force a cease-fire. The resulting loss of prestige and influence for Britain is often described as the final step in her decline as a colonial power.

1956 To retaliate for the actions of Britain and France against Egypt in the Suez War (see above), King Saud of Saudi Arabia cuts off oil supplies to Britain and France. This marks the first time such a step is taken by an Arab state.

1956 U.S. geologist Marion King Hubbert predicts that oil production in the lower 48 states will peak in around 1970, which in fact it does. Hubbert's work becomes a lightning rod for debate about oil supplies for decades to come.

1956–1957 Engineer Johannes Juul develops the first alternating current windmill for the electricity company SEAS at the Gedser coast in southern Denmark. For many years the Gedser turbine remains the largest in the world.

1957 A converted version of the U.S. B-57 military aircraft flies with power from a liquid hydrogen fuel system.

1957 A reactor fire at England's Windscale (Sellafield) nuclear power plant burns out of control for about 24 hours. An estimated 20,000 curies of radioactive iodine escapes along with other isotopes such as plutonium, caesium, and the highly toxic polonium. Prime Minister Macmillan orders news of the accident to be suppressed for fear of alarming the public (see also 1982).

1957 Felix Wankel operates his rotary engine, an idea he first conceived in the mid-1920s. This is a four-cycle internal combustion engine that does not use pistons. Though the Wankel engine has great simplicity of operation and an excellent power-to-weight ratio, it does not come into use until the 1970s and to date it has not had a major commercial impact.

1957 IBM introduces the Model 608 computer, the first solid-state computer for the commercial market.

1957 Japanese physicist Leo Esaki discovers the tunnel diode, a high-speed current-switching device that is the first semiconductor quantum electron structure.

1957 Melvin Calvin, a U.S. chemist, uses radioactive carbon-14 to map the complete route that carbon travels through a plant during photosynthesis, starting from its absorption as atmospheric carbon dioxide to its conversion into carbohydrates and other organic compounds. This becomes known as the Calvin Cycle.

1957 Nobel laureate Linus Pauling writes and lectures on the dangers posed by nuclear fallout for human life and the environment. In the following year he presents a petition to the United Nations opposing further nuclear weapon testing, signed by more than 11,000 scientists from 49 different countries.

1957 Novelist Jack Kerouac publishes *On the Road;* it becomes a best seller and eventually will be known as the "Bible of the Beat Generation." The author's memorable descriptions of marathon car trips across the American continent epitomize the love affair between mid-century Americans and their automobiles.

1957 Oil is discovered at Alaska's Kenai Peninsula near the Swanson River, the first significant oil discovery in the state.

1957 Power from a civilian nuclear unit is generated for the first time by the Sodium Reactor Experiment at Santa Susana, California. The unit provides power until 1966.

1957 Roger Revelle of the United States and Hans Suess of Austria demonstrate that carbon dioxide has increased in the atmosphere as a result of the use of fossil fuels. Revelle concludes that "Human beings are now carrying out a large scale geophysical experiment of a kind that could not have happened in the past nor be reproduced in the future."

1957 The cooling system of a radioactive waste containment unit malfunctions and explodes at a nuclear facility in Chelyabinsk province of the Soviet Union. About two million curies are released, exposing a quarter of a million people to radiation. The facility also is responsible for decades of dumping that release large amounts of radiation.

1957 The first long-distance coal pipeline goes into commercial service. It moves coal slurry (a mixture of water, fine coal particles, and other wastes from washing coal) from a mine in Ohio to a power station over 100 miles away.

1957 The Price-Anderson Act provides financial protection to the American public and Atomic Energy

Corporation licensees and contractors if a major accident occurs at a nuclear power plant.

1957 The Space Age begins on October 4th with the launching of the first artificial satellite, *Sputnik 1,* by the Soviet Union. In the following month *Sputnik 2* is launched with the first animal to go into in space, a dog named Laika.

1957 The United Nations creates the International Atomic Energy Agency (IAEA) in Vienna, Austria, to promote the peaceful use of nuclear energy and prevent the spread of nuclear weapons around the world.

1957 The world's first large-scale nuclear power plant begins operation in Shippingport, Pennsylvania. The plant reaches full power three weeks later and supplies electricity to the Pittsburgh area.

1957 Three physicists at the University of Illinois, John Bardeen, Leon Cooper, and J. Robert Schrieffer, develop the "BCS" model of why superconductors behave as they do. They show that electrons in a superconductor condense into a quantum ground state and travel together collectively and coherently.

1957 U.S. ecologist Howard T. Odum publishes a landmark paper that quantifies the energy flow among the trophic levels in the ecosystem of Silver Springs, Florida.

1958 A geothermal power station begins operation at Wairakei, New Zealand; this is the second commercial geothermal plant in the world (after Larderello, Italy) and the first to exploit a wet geothermal resource.

1958 After years of unsuccessfully trying to harness fusion, scientists from the United States, Britain, and the Soviet Union declassify their fusion research at the second United Nations Conference on the Peaceful Use of Atomic Energy, held in Geneva.

1958 Based on the work of Willard Thomas Grubb and Leonard Niedrach, the General Electric Company develops the first commercially successful fuel cell, which will be used by NASA in the Gemini Earth-orbit space program.

1958 Chrysler introduces the first mass-produced electronic fuel injection system, designed by Bosch. It is not welcomed by consumers and in the 1960s the Bendix patents are sold to Bosch; that company then develops an improved version that becomes successful.

1958 Construction begins on the world's first nuclear-powered merchant ship, the *N.S. Savannah,* in Camden, New Jersey. The ship is launched the following year

1958 In response to the 1957 launch of *Sputnik,* the United States forms the Advanced Research Projects Agency (ARPA) within the Defense Department to promote the development of new technologies with military application. ARPA will be a catalyst for the development of the modern computer, the Internet, and other information technology.

1958 Sam Cohen, formerly of the Manhattan Project, develops the concept for a neutron bomb, a "smart" type of nuclear weapon modified to have a smaller overall blast effect but to release high-energy neutrons that will take lives far beyond the blast area.

1958 The U.S. Army Ballistic Missile Agency launches *Explorer I,* the first artificial satellite placed into orbit by the United States. In the same year the United States will launch five more *Explorer* satellites (some of which reach orbit and others not), as well as *Vanguard I,* which is now the oldest satellite still in orbit.

1958–1959 Engineers Jack Kilby and Robert Noyce of the United States separately invent the integrated circuit, a miniaturized electronic circuit that will make possible the development of much smaller, cheaper, and more powerful computers and other modern electronic devices.

1959 An oil well drilling platform is constructed in over 200 feet of water offshore from Louisiana.

1959 British scientist Francis Bacon demonstrates a fuel cell device that can produce 5 kilowatts of power, enough to run a welding machine. This is the culmination of nearly 30 years of effort by Bacon to produce a workable fuel cell. Later that same year, Harry K. Ihrig of Allis-Chalmers displays the first fuel cell vehicle, a 20-horsepower tractor.

1959 The Dresden-1 Nuclear Power Station in Illinois, the first U.S. nuclear plant built entirely without government funding, achieves a self-sustaining nuclear reaction.

1959 The first man-made object lands on the moon, the Soviet *Lunik II* rocket.

1959 U.S. brewing company Coors sells beer in aluminum cans for the first time, and other beverage companies soon follow suit. By the 1970s aluminum cans will replace bottles as the leading type of drink container, giving rise to a huge recycling program to recover the used cans.

1959 Verner Suomi, a U.S. meteorologist, uses data from the *Explorer VII* satellite to estimate the global radiation heat budget of the Earth–atmosphere system. This is the first use of a satellite to characterize the Earth's energy balance.

1960–1969

ca. 1960 The human population reaches 3 billion.

1960 Physicist Theodore Harold Maiman designs the first working laser (light amplification by stimulated emission of radiation), using a pink ruby medium. Within two years, the laser is being employed in eye surgeries. The term "laser" was reportedly coined earlier

by Gordon Gould, and he will eventually win several key laser patents.

1960 President Eisenhower establishes the Arctic National Wildlife Refuge in Alaska, which will eventually consist of more than 19 million acres. Later this area will become the focus of an intense debate between environmentalists and developers over whether to allow oil drilling in the coastal plain of the Refuge (see 1980).

1960 The final steam locomotive is built by British Railways, the *Evening Star.*

1960 The first large-scale geothermal power plant in the United States begins operation at The Geysers geothermal field in northern California. The Geysers soon becomes the world's largest steamfield, reaching its peak production in 1987, when it produces enough electricity to supply 1.8 million people.

1960 The Organization of Petroleum Exporting Countries (OPEC) is announced in Baghdad with five original members: Iraq, Iran, Kuwait, Saudi Arabia, and Venezuela. Qatar joins OPEC in 1961, followed by Indonesia and Libya (1963), the Emirates of Abu Dhabi (1968), Algeria (1969), Nigeria (1971), Ecuador (1973), and Gabon (1975). The latter two leave OPEC in the 1990s.

1960 The transition in the United States from manufactured gas to natural gas is virtually complete. However, environmental concerns still exist because of toxic byproducts remaining in soil and groundwater at the sites of former manufactured gas plants.

1960 The Yankee Atomic Electric Company plant (Yankee Rowe) begins operation on the Deerfield River at Rowe, Massachusetts. The company is formed by 12 New England utilities to supply electrical power. Yankee Rowe operates for more than 30 years until being shut down due to reactor vessel embrittlement.

1960 West Germany builds its first industrial nuclear reactor, the Kahl am Obermain experimental nuclear power plant. It will operate until 1985.

1961 Self-advancing hydraulic longwall supports are introduced to U.S. coal mining. These provide greater support for the roof of the mine and also reduce deformation, fracture, and movement around the coal face.

1961 The first use of a radioisotope thermoelectric generator for space applications is on the U.S. Navy Transit 4A spacecraft. These generators provide electrical power for spacecraft by direct conversion of the heat generated by the decay of plutonium-238.

1961 The Soviet Union sends the first human into space, Yuri Gagarin, who orbits Earth in a 108-minute mission aboard the *Vostok I*. The following month, Alan Shepherd becomes the first American in space when he makes a suborbital flight of 15 minutes.

1961 The U.S. Navy commissions the world's largest ship, the *U.S.S. Enterprise*. It is a nuclear-powered aircraft carrier with the ability to operate for distances up to 400,000 miles without refueling.

1961 U.S. physicist Murray Gell-Mann and others propose a system for classifying subatomic particles which Gell-Mann terms "the eightfold way." The system is the foundation of the quark theory.

1961 U.S. President Kennedy says that citizens can survive a nuclear attack if they take refuge in a fallout shelter. A national survey identifies large buildings and other sites such as mines, caves, and tunnels that could protect people from fallout radiation; these areas are designated as public shelters.

ca. 1961 Small nuclear-power generators are first used in remote areas to power weather stations and to light buoys for sea navigation.

early 1960s Regulations to limit harmful emissions from motor vehicles are introduced in the United States and other nations. Emission control devices are then installed in new vehicles.

early 1960s The Pratt and Whitney Division of United Aircraft (now United Technologies) takes out a license on the fuel cell patents of Francis Bacon (see 1932) and uses the concept of the Bacon cell to provide electrical power for the *Apollo* spacecraft.

1962 A trash fire in an open-pit anthracite coal mine in Centralia, Pennsylvania ignites an exposed vein of coal. This in turn burns through old coal mines under the town, and fire, smoke, and toxic gases reach the surface and eventually force the abandonment of most of the town's buildings. As of this writing, the fire still burns.

1962 General Electric scientist Nick Holonyak creates the first light-emitting diode (LED). LED displays will become a universally used light source in consumer electronics, traffic signals, public information messages, and many other applications.

1962 In Natick, Massachusetts, a solar furnace is built by the U.S. Army to simulate radiation burns and to experiment with pigs wearing sunscreen lotions that might protect soldiers from nuclear radiation.

1962 Rachel Carson publishes *Silent Spring,* the first widely read work to describe the effect on health and the environment of human-produced toxic chemicals, such as the insecticide DDT. This book is often given credit for launching the modern environmental movement at the popular level.

1962 The U.S. Congress creates the Communications Satellite Corporation (COMSAT). Agencies from 17 other countries join with COMSAT in 1964, forming an international consortium to establish a satellite communications network. By this time the United States has launched a series of satellites, beginning with *Echo I* in 1960.

1962 The U.S. Navy constructs a nuclear power plant at McMurdo Sound, Antarctica, to supply power for the American research installations there.

1963 Ffestiniog Power Station in Wales goes into operation, the United Kingdom's first major pumped storage power facility. Pumped storage involves two reservoirs at different levels. Water is pumped to the higher of the two, then released to flow downward through high-pressure shafts linked to turbines, thus generating power for electricity.

1963 Harold Barnett and Chandler Morse publish the first empirical analysis of the supply of natural resources in the United States. It shows that the cost of timber, fish, minerals, and fuels generally declined from 1870 to 1957 due to what they call "self-generating technological change." This is used by some to argue that human ingenuity can overcome natural resource constraints.

1963 Japan installs a 242-watt photovoltaic array on a lighthouse, the world's largest array at that time.

1963 The Jersey Central Power and Light Company announces its commitment for the Oyster Creek nuclear power plant, the first time a nuclear plant is ordered as an economic alternative to a fossil-fuel plant.

1963 Semisubmersible platforms are in use for offshore oil drilling. In semisubmersible technology, ballasted, watertight pontoons below the sea surface and its wave action provide a base for columns that penetrate the water surface and support the operating deck.

1963 The U.S. Congress passes the first Clean Air Act, legislation intended to "improve, strengthen, and accelerate programs for the prevention and abatement of air pollution." Various other Clean Air Acts will follow in the 1960s and after.

1964 George Heilmeier heads a research group at RCA that produces the first liquid crystal display (LCD). This becomes a widely used technology for various electronic applications, such as watches and clocks, computer displays, and eventually television screens (the original use envisioned by the RCA team).

1964 Japan's *Shinkansen* (also known as the Bullet Train) becomes the world's first high-speed railway system. It runs between Tokyo and Shin-Osaka at a maximum business speed of over 200 kilometers per hour (130–160 mph).

1964 Murray Gell-Mann and George Zweig independently introduce the three-quark model of elementary particles.

1964 The Baltimore Light Station on the Chesapeake Bay in Maryland becomes the first nuclear powered lighthouse in the United States.

1964 The General Electric PEM fuel cell represents the first successful application of a fuel cell in space.

1964 Three U.S. nuclear-powered surface ships, the *Enterprise, Long Beach,* and *Bainbridge,* complete "Operation Sea Orbit," an around-the-world cruise.

1964 U.S. President Lyndon Johnson signs the Private Ownership of Special Nuclear Materials Act, which allows the nuclear power industry to own the fuel for its units.

1965 A year of major advances in space technology, with the Soviets making the first space walk, the United States launching the first nuclear reactor in space (SNAP, or systems for nuclear auxiliary power), the Americans completing the first space rendezvous between two satellites, and the Soviets sending out a space probe that lands on Venus.

1965 Geologists discover oil and natural gas in the West Sole Field of the North Sea, off the coast of Great Britain. Eventually North Sea oil will make Britain roughly self-sufficient in energy and also bring great prosperity to Norway, which will become one of the world's leading oil exporters.

1965 The Great Blackout of 1965 covers 80,000 square miles of the northeastern United States and Canada. The failure affects four million homes in the New York City area, and leaves between 600,000 and 800,000 people stranded in the city's subway system. The events of the blackout become the subject of a Hollywood film, *Where Were You When The Lights Went Out?*

mid-1960s Both the United States and the Soviet Union have deployed a full-scale system of intercontinental ballistic missiles (ICBMs).

mid-1960s The era of the supertanker is underway as ultra-large vessels are used to transport oil around the globe, especially from the Middle East to western Europe and the United States.

mid-1960s U.S. ecologist Eugene E. Likens describes the phenomenon of acid rain in North America, based on observations at Hubbard Brook, New Hampshire, a leading research site he co-founded for studies in ecosystem science and biogeochemistry. Acid rain will be tied directly to pollutants from the combustion of fossil fuels.

1966 The world's largest tidal power plant goes into operation at the mouth of the La Rance river in northern France. The tide is channeled through the narrow entrance to the river, across which is a dam with turbines and generators.

1966 U.S. economist Kenneth Boulding publishes *The Economics of the Coming Spaceship Earth,* the first explicit application of the law of conservation of matter to describe the physical limits to economic growth.

1967 Egypt, Syria, and Jordan oppose Israel in the third Arab–Israeli War. This becomes known as the "Six-Day War" because of the swift Israeli victory. The war results

in territorial gains for Israel and increased tensions between the Arab World and the West (see next).

1967 Saudi Arabia leads an Arab oil embargo against the United States and Britain, who were accused of aiding Israel against Arab countries in the Six-Day War. The boycott, however, lasts for only two months and does not significantly affect the Western economies.

1967 The first barrel is shipped from the Great Canadian Oil Sands Project, the first commercially successful operation to tap the rich Athabasca oil sands in the province of Alberta. The project ushers in an era of rapid development of this oil sand resource base.

1967 The Oroville Dam on the Feather River in northern California is completed; it is the tallest dam in the United States.

1967 The supertanker *Torrey Canyon* runs aground on a reef near the coast of Cornwall in southwestern England. The ship spills an estimated 30 million gallons of oil into the sea, producing environmental damage on a scale not even imagined before that time and blighting the entire region for years to come.

1968 A coal mine explosion in Farmington, West Virginia results in the death of 78 miners. This mine is just a few miles away from Monongah, the site of the worst mine disaster in U.S. history (see 1907).

1968 Large oil deposits are discovered in Alaska at Prudhoe Bay, but it is not until 1977 that the first oil is shipped via the Trans-Alaska Pipeline.

1968 Lev Asimovich, director of the Soviet controlled-fusion program, announces that the Soviet T-3 tokamak has confined a plasma at a temperature of more than 10 million degrees centigrade for more than 10 thousandths of a second. Europe, Japan, and the United States all strive to develop tokamak designs.

1968 OPEC Resolution XVI.90 sets forth major new demands; this signals an era of stronger national control of Mideast oil, and the declining ability of Western multinational oil companies to control the global oil supply.

1968 The Odeillo solar furnace, the most powerful in the world, is built at Font-Romeu in the Eastern Pyrenees region of France. It has 10,000 small mirrors that concentrate the Sun's energy into useable power.

1968 The Soviet Union operates an experimental tidal power station near Murmansk.

1968 The Soviet Union's Tupolev Tu-144 becomes the first supersonic transport plane (SST) to take flight, preceding the Concorde by two months (see 1969). The TU-144 begins service in 1975 and makes its last commercial flight in 1978. The United States funds its own SST project in the 1960s, but the program is dropped in 1971 without any aircraft ever being flown.

1968 The U.S. Congress creates the Wild and Scenic Rivers System, which will "preserve selected rivers or sections thereof in their free-flowing condition to protect the water quality of such rivers and to fulfill other vital national conservation purposes."

1968 U.S. space scientist Peter Glaser presents the concept for the Solar Power Satellite system, to supply power from space for use on the Earth. The SPS system would have a huge array of solar cells, perhaps tens of square miles, that would generate electrical power to be beamed to Earth in the form of microwave energy.

late 1960s The advent of more powerful digital electronic computers make it possible to model the global general atmospheric circulation for the first time. By the 1970s, general circulation models (GCMs) are the central tool of climate scientists such as Syukuro Manabe and Kirk Bryan, and they will play a key role in describing climate change.

late 1960s The agricultural movement known as the Green Revolution is underway, featuring the widespread use of new, high-yielding crop varieties and new technologies such as increased use of irrigation, fertilizer, and pesticides. This increases harvests for millions of farmers, although with some negative social and environmental impacts.

1969 An offshore oil drilling platform at Santa Barbara, California, suffers a natural gas blowout, releasing about 200,000 gallons of crude oil. The spill mars 35 miles of coastline and causes extensive environmental damage, especially to the local populations of marine mammals and seabirds.

1969 ARPANET, the technological forerunner of today's Internet, is established. It links four nodes: the University of California at Los Angeles, University of California at Santa Barbara, University of Utah, and Stanford Research Institute. Its basic services of remote log-in, file transfer protocol, and electronic mail continue as functions of the Internet.

1969 Big Muskie, described as the world's largest earth-moving machine at over 200 feet in height and weighing some 13,500 tons, is put into use in Cumberland, Ohio. Big Muskie is what is known as a walking dragline, a machine designed to expose deeply buried coal seams that smaller machines could not reach.

1969 Boeing's 747, the first "jumbo jet," goes into commercial service. Jumbo jets such as the 747, the Lockheed L-1011, and the McDonnell Douglas DC-10 will be the stars of air travel in the coming decade, but rising fuel costs later cause the airlines to turn to smaller planes. Today only a single model of the 747 is a holdover from the jumbo jet era.

1969 Ted Hoff of Intel proposes the microprocessor, a new type of computer architecture based on a single-chip general-purpose central processing unit (CPU).

Intel will release the first working CPU in 1971; this single chip has as much computing power as the ENIAC computer, which filled an entire room.

1969 The British-French supersonic aircraft Concorde makes its first flight. It will begin transcontinental commercial passenger flights in 1976. The Concorde ceases flying in 2003 because of safety concerns and the high cost of operation relative to passenger revenue.

1969 The Cuyahoga River, flowing through Cleveland, Ohio, catches fire from oil and chemical pollution. The paradoxical image of a burning body of water becomes a rallying point for passage of the Clean Water Act.

1969 The first person to step on the moon is Neil A. Armstrong of the United States, followed by Buzz Aldrin, with Michael Collins orbiting the moon in the *Apollo 11* spacecraft.

1969 The *S.S. Manhattan,* an ice-breaking tanker, becomes the first commercial ship to sail the fabled Northwest Passage, a dream of navigators for 500 years. The purpose of the expedition is to test the feasibility of using the Northwest Passage to transport oil from Alaska's North Slope to the markets of the eastern United States and Europe.

1970–1979

1970 At California's Geysers project, the injection of spent geothermal fluids back into the production zone begins, as a means to dispose of waste water and maintain reservoir life

1970 Earth Day is celebrated for the first time in the United States as a nationwide demonstration advocating environmental protection and preservation. This demonstration is conceived by Gaylord Nelson, U.S. Senator from Wisconsin, and organized by Denis Hayes.

1970 IBM introduces an eight-inch "floppy disk" as a medium for data storage in its 370 computer.

1970 Premier Zhou-Enlai of China delivers a speech emphasizing the necessity for exploring the peaceful use of nuclear energy. The first nuclear reactor in China will not come on line until the 1990s.

1970 The Environmental Protection Agency (EPA) is formed in the United States, with the mission to "protect human health and to safeguard the natural environment—air, water, and land—upon which life depends."

1970 The first silica fiber optical cable is produced at the Corning Glass Works (New York), using chemical vapor deposition techniques to reduce attenuation and improve signal transmission.

1970 The Geothermal Steam Act is enacted by Congress; it authorizes and governs the lease of geothermal steam and related resources on public lands. The Geothermal Resources Council is formed to encourage the development of geothermal resources worldwide.

1970 The level of production of crude oil in the United States reaches its all-time high.

1970 The U.S. Congress adopts a law requiring federal agencies, including the Atomic Energy Agency, to consider thermal pollution in their licensing process. This is the first time that environmental issues are seriously raised in the context of nuclear power, and is a harbinger of more intense scrutiny of nuclear power's potential impact.

1970 The United States, Great Britain, the Soviet Union, and 45 other nations sign the Treaty for the Nonproliferation of Nuclear Weapons.

1970–1974 Paul Crutzen and Harold Johnston identify an ozone-destroying catalytic cycle involving oxides of nitrogen. Richard Stolarski and Ralph Cicerone propose another such cycle involving chlorine, and Mario Molina and F. Sherwood Rowland link this cycle to ozone depletion by chlorofluorocarbon (CFC) emissions.

1971 For the first time, surface mined coal production in the United States exceeds the amount of coal mined underground.

1971 Howard T. Odum publishes *Environment, Power, and Society,* a classic and comprehensive application of thermodynamics and ecological energetics to the analysis of human economic and social systems.

1971 Nicholas Georgescu-Roegen, a Hungarian-born mathematician and economist, publishes *The Entropy Law and the Economic Process,* in which the second law of thermodynamics plays a central role in production theory. This work becomes a cornerstone of the fields of ecological and evolutionary economics.

1971 Ray Tomlinson of Bolt Beranek and Newman, Inc. sends the first e-mail message across the ARPANET system. The content of the message is chosen at random, probably something like "qwertyuiop" or "Testing 1-2-3."

1971 Twenty-two commercial nuclear power plants are in full operation in the United States. They produce 2.4% of U.S. electricity at this time.

early 1970s Ecologists Robert Whitaker, Eugene Likens, H. Leith, and other biologists quantify the net primary production and plant biomass for the Earth's major biomes.

early 1970s Kalundborg, Denmark, a small coastal town, begins a web of energy and material exchanges among several firms seeking income-producing uses for waste products. Eventually, the firm managers and town

residents realize they are generating environmental benefits as well. The term "industrial symbiosis" is used to describe this system.

early 1970s Russian-born chemist Ilya Prigogine extends the second law of thermodynamics to systems that are far from equilibrium, and shows that stable, coherent patterns could emerge under such conditions. He calls these "dissipative structures"; this concept is used to describe phenomena such as the growth of cities and the physics of car traffic.

1972 Ford Motor Company develops a sodium-sulfur battery for use as a power source in electric vehicles.

1972 The Clean Water Act is passed, establishing the basic structure for regulating the discharge of pollutants into the waters of the United States.

1972 The Club of Rome publishes T*he Neo-Malthusian Limits to Growth,* warning that if the world's consumption patterns and population growth continue at the same high rates of the time, various calamities will befall the Earth and humankind.

1972 The first large-scale commercial waste-to-energy facility is used by the Union Electric Company's Meramec Plant in St. Louis, Missouri. Refuse is shredded and burned with coal to generate electricity.

1972 The U.S. Environmental Protection Agency announces a ban on the pesticide DDT. This is part of a trend in which DDT is generally banned worldwide. In recent years, however, critics have argued that the ban should be lifted because of the huge numbers of deaths in the developing world from the mosquito-borne disease malaria.

1972 The United Nations Conference on the Human Environment (UNCHE), is held in Stockholm, Sweden, with 113 countries represented; this is the first global environmental summit. Treatment of energy is largely limited to the environmental effects of energy extraction, processing, and consumption.

1972 U.S. scientist Paul Berg assembles the first DNA molecules that combine genes from different organisms. In the following year Herbert Boyer and Stanley Cohen report the construction of functional organisms that combine and replicate genetic information from different species.

ca. 1972 The U.S. Atomic Energy Commission announces that a salt mine at Lyons, Kansas, will be developed as a high-level radioactive waste repository. The decision is reversed in 1972 after geologists learn that the site is riddled with abandoned oil and gas exploration boreholes.

1972–1982 More than 100 orders for nuclear power plants in the United States are canceled, due mainly to cost overruns, a decline in the demand for power, and public concerns about safety. The number of cancellations grows to 124 by 2000.

1973 A huge pumped-storage facility goes into operation on Lake Michigan. The Ludington power plant pumps water uphill from the lake to a storage reservoir when demand is low, then releases the water to flow back downhill and generate electricity as needed by turning giant turbines.

1973 A total of 109 nuclear power plants operating in the United States generates 610 billion kilowatt-hours of net electricity, providing about one-fifth of the nation's electricity.

1973 America's public and private utilities form the Electric Power Research Institute (EPRI) in Palo Alto, California. EPRI will grow to be one of the world's most influential electric power research and development organizations.

1973 An international journal known as *Energy Policy* begins publication. This is the first social science journal pertaining to the study of energy.

1973 Chemist Nathaniel Wyeth receives a patent for PET (polyethylene terephthalate) beverage bottles. This is the first practical plastic that is strong enough to hold carbonated beverages without bursting, and it soon becomes a recyclable material.

1973 Conventional natural gas production peaks in the United States. Abundant quantities of unconventional gas (deep gas, coalbed methane, tight gas) still remain today, but their economic, technological, and environmental viability are the subject of considerable debate.

1973 Egypt and Syria launch a surprise attack against Israel (October 6th) in what will become known as the Yom Kippur War. Many other Arab states, including Iraq and Saudi Arabia, also provide troops and supplies.

1973 Members of the Organization of Arab Petroleum Exporting Countries (OAPEC) meet in Kuwait as the Yom Kippur War goes on. They reach agreement on a coordinated policy of cuts in oil production coupled with price increases.

1973 The Arab oil-producing states impose a complete embargo on oil shipments to the United States in retaliation for American support of Israel in the Yom Kippur War. The embargo will later be extended to The Netherlands. Japan and other nations in Europe will also be hard hit by the combination of price increases and production cuts.

1973 The Black Mesa coal pipeline becomes operational, transporting coal slurry 273 miles from a mine on Navajo Indian land in northern Arizona to the Mojave Generating Station at Laughlin, Nevada. This is currently the only coal slurry pipeline in operation in the United States.

1973 The first major energy crisis in the United States occurs as a result of the Arab oil embargo. By the following year the price of crude oil will be four times its level of 1972. The oil embargo will be lifted in

March of 1974, but the United States and other Western nations will continue to feel its effects for years to come.

1973 U.S. President Richard Nixon announces a series of measures to cope with fuel shortages, such as a national 55 miles per hour speed limit, year-round daylight saving time, and a ban on Sunday sale of gasoline. Various states will also impose forms of rationing, e.g., sale of gas to motorists on alternate days depending on a car's license number.

1974 A well that bottoms out at 31,441 feet is drilled in Oklahoma, the deepest well ever drilled in the United States at the time. It is a dry well.

1974 Entrepreneur Robert Beaumont revives the consumer electric vehicle market with his CitiCar. Several thousand of these will be produced, due to interest in alternative vehicles spurred by the oil crisis of 1973. However, the CitiCar is essentially a glorified golf cart, and by the late 1970s it is no longer produced.

1974 Faced with rising oil prices and a heavy dependence on imported oil, the French government embarks on an ambitious plan to develop nuclear power. By 2000, France has built 58 reactors that supply nearly three-quarters of its power.

1974 Howard Georgi and Sheldon Glashow develop the first Grand Unified Theory (GUT) to unite strong, weak, and electromagnetic forces. The attempt to develop a unified field theory was the focus of Einstien's work for many years.

1974 The American Wind Energy Association (AWEA) is founded to advocate the development of wind energy as a reliable, environmentally superior form of alternative energy in the United States and around the world.

1974 The Energy and Resources Institute (TERI) is formed in New Delhi, India. TERI will grow to be one of the world's leading organizations focusing on all forms of natural and human resources, with a special interest in sustainable energy and the environment.

1974 The first International Conference on Hydrogen Energy and the Hydrogen Economy is held in Miami Beach, Florida. From this meeting springs the International Association for Hydrogen Energy (IAHE), the first professional society dedicated to hydrogen energy.

1974 The first known use of the term "Internet" appears in a paper on Transmission Control Protocol (TCP) by Vinton Cerf and Robert Kahn.

1974 The International Energy Agency is created, a forum for 26 member countries who are committed to taking joint measures to meet oil supply emergencies. IEA collects and publishes a wide array of energy and environmental data and research reports.

1974 The U.S. Atomic Energy Commission is dissolved. The Energy Reorganization Act divides its functions between two new agencies: the Energy Research and Development Administration (ERDA), to carry out research and development, and the Nuclear Regulatory Commission (NRC), to regulate nuclear power.

1975 As a result of the oil crisis of 1973–1974, the U.S. Congress enacts the Energy Policy and Conservation Act (EPCA). EPCA sets minimum fuel economy standards for cars and light trucks, and also establishes the Strategic Petroleum Reserve, an emergency oil store intended to counteract a severe supply interruption.

1975 EPCA also requires that home appliances such as refrigerators, air-conditioners, and washing machines be sold with a label to indicate their energy usage. West Germany introduces a similar voluntary energy labeling program in the same year.

1975 Opinion polls in the United States show that about 20% of the public is opposed to the construction of more nuclear power plants. On the eve of the accident at the Three Mile Island nuclear plant in 1978, that number had risen to nearly 40%. By the time of the Chernobyl disaster in 1986, opposition had grown to about 80%.

1975 The development of microchips makes possible the first personal computer to have a significant public impact, the Altair, with 256 bytes of memory. It is sold by MITS, Inc. of Albuquerque, New Mexico. Bill Gates and Paul Allen then develop a BASIC software program for the Altair.

1975 The U.S. Geological Survey releases the first national estimate and inventory of geothermal resources.

ca. 1975 The human population reaches 4 billion.

1976 A U.S. Federal court rules that the Environmental Protection Agency has the authority to regulate leaded gasoline. The lead phase-out begins, and by June 1979, nearly half of all U.S. gasoline is unleaded.

1976 An Energy Modeling Forum is established at Stanford University to provide a structured forum within which energy experts from government, industry, universities, and other research organizations can meet to study important issues of common interest.

1976 British journalist Anthony Sampson publishes *The Seven Sisters: The Great Oil Companies and the World They Made*, a history of the large multinational firms that have historically dominated the oil industry. Sampson's "Seven Sisters" are Exxon, Mobil, Socal (Chevron), Texaco, Gulf, British Petroleum, and Royal Dutch/Shell.

1976 In a watershed year for the personal computer, Steve Wozniak and Steve Jobs form the Apple Computer Company and introduce their first Apple computer. The trade name "Microsoft" is registered, and Bill Gates and Paul Allen begin to work full time for the company.

1976 U.S. President Jimmy Carter enacts legislation which authorizes the U.S. Department of Energy to conduct an Electric and Hybrid Vehicle (EHV) program.

1976 The Alaska Natural Transportation Act calls for a pipeline to transport natural gas from deposits in Alaska through Canada to the lower 48 United States. The plan remains stalled for decades due to opposition from environmental groups, cost considerations, uncertainty about future gas prices, and Canada's proposal for a different route.

1976 The U.S. Resource Conservation and Recovery Act gives the Environmental Protection Agency (EPA) the authority to control hazardous waste from "cradle-to-grave." This includes the generation, transportation, treatment, storage, and disposal of hazardous waste.

1977 Congress abolishes the Joint Committee on Atomic Energy, which was established in 1946 to oversee the Atomic Energy Agency. With the demise of the JCAE, supporters of nuclear power lose control of strategic junctures of policy-making that had been favorable to the industry. Jurisdiction is now spread over many agencies.

1977 New York City and neighboring Westchester County are plunged into darkness as lightning downs major transmission power lines. Looting and violence erupt, and within a span of two days, police arrest nearly 4,000 looters and the city suffers an economic loss estimated at $300 million.

1977 President Carter installs solar panels on the White House and promotes incentives for solar energy systems. President Ronald Reagan removes these panels in 1980 and eliminates many of the policy incentives for renewable energy.

1977 Scientists develop the first hot dry rock reservoir at Fenton Hill, New Mexico to "mine" geothermal energy from beneath the earth.

1977 The Apple II becomes the first highly successful mass-produced personal computer.

1977 The Department of Energy (DOE) is established as a cabinet-level agency in the United States; the DOE takes over the activities of the Energy Research and Development Administration and the Federal Energy Administration.

1977 The Federal Energy Regulatory Commission (FERC) is established in the United States, succeeding the Federal Power Commission. FERC is a quasi-independent agency within the DOE with jurisdiction over such issues as interstate electricity sales, wholesale electric rates, hydroelectric licensing, natural gas pricing, and oil pipeline rates.

1977 The first oil flows through the Trans-Alaska Pipeline (TAPS). Actual construction began in 1974 and it ultimately will cost $8 billion, at the time the largest privately funded construction project in history. Throughout the 1980s and 1990s, TAPS carries as much as 20% of total U.S. oil production.

1977 The first shipment of oil, approximately 412,000 barrels of Saudi Arabian light crude, is delivered to the U.S. Strategic Petroleum Reserve, an emergency oil supply stored in huge underground salt caverns along the coastline of the Gulf of Mexico.

1977 The International Association for Energy Economics (IAEE) is founded. IAEE is the largest non-profit professional organization devoted exclusively to energy, with members in 70 countries. IAEE publishes one of the most respected academic journals on energy.

1977 The non-governmental organization Alliance to Save Energy is founded in the United States. The Alliance will play a major role in the education, communication, and political aspects of energy efficiency in the United States.

1977 The Solar Energy Research Institute (SERI), later to become the National Renewable Energy Laboratory (NREL), begins operation in Golden, Colorado.

1977 The U.S. Surface Mining Control and Reclamation Act is enacted. It sets federal mining standards for coal, as well as standards for the reclamation of mining sites.

1977 The *Voyager* spacecraft is launched; it is powered by thermoelectric generators that produce energy from the radioactive decay of plutonium.

1978 A 15% energy tax credit is added to an existing 10% investment tax credit, providing incentive for capital investment in solar thermal generation facilities for independent power producers.

1978 CFCs (chlorofluorocarbons, or Freons) are banned in the United States as spray propellants because of fears over their role in the depletion of the ozone layer.

1978 In the same vein as the first Earth Day earlier in the decade, people around the world celebrate "Sun Day" to focus on the use of solar energy as a replacement for fossil fuel sources.

1978 President Carter signs the National Energy Conservation and Policy Act, which includes various specific acts to promote conservation and develop greater energy efficiency.

1978 Shortages of natural gas prompt the U.S. Congress to pass the Natural Gas Policy Act, which effectively ends decades of natural gas price controls by the federal government. The goal of the legislation is to deregulate natural gas prices over time, to encourage exploration, and to reduce the price differentials between interstate and intrastate markets.

1978 The last two orders for new nuclear power plants in the United States are placed during this year.

1978 The U.S. Department of Transportation issues CAFE (Corporate Average Fuel Economy) standards for passenger cars, with an original standard of 18 miles

per gallon and an increase to 27.5 miles per gallon within seven years. DOT also establishes standards for light trucks, starting with the following model year.

1978 The U.S. Public Utility Regulatory Policies Act (PURPA) mandates the purchase of electricity from qualifying facilities (QFs) meeting certain technical standards regarding energy source and efficiency; this especially favors independent sources that are not fossil fuel-based, helping wind and other renewables.

late 1970s German automobile makers Daimler-Benz and BMW begin the experimental development of hydrogen-powered internal-combustion engine vehicles.

late 1970s Japanese automobile brands such as Toyota, Honda, and Datsun (Nissan) now play a major role in the U.S. market; these vehicles appeal to American buyers looking for greater fuel efficiency in an era of rapidly increasing gas prices.

1979 A blowout of the experimental Ixtoc-1 well off the coast of Mexico releases about 140 million gallons of crude oil. This is the largest non-war related oil spill to date (though its environmental impact is relatively small).

1979 At the Papago Indian Reservation in Schuchuli, Arizona, NASA's Lewis Research Center completes a 3.5-kilowatt photovoltaic system; this is the world's first village PV system.

1979 In a time of rising energy prices and fuel shortages, President Carter gives a speech in which he speaks of a crisis in the American spirit. He calls for import quotas, a tax on oil profits, development of synthetic fuels, and individual energy conservation. Now known as the "malaise" speech, it is poorly received by the public and contributes to Carter's loss in the next election.

1979 Iran's pro-Western ruler Reza Shah Pahlavi is driven from power and replaced by exiled Islamic leader Ayatollah Ruhollah Khomeini. Iranian militants seize the U.S. embassy in Tehran, holding more than 60 staff members hostage. In response, President Carter places an embargo on importing Iranian oil into the United States.

1979 Oil prices rise sharply following the revolution in Iran, largely because of fears that overall supplies from the Persian Gulf will be disrupted. By 1981 they will reach an all-time high of $36.47 per barrel or nearly triple the price prior to the revolt. Once again political events of the Middle East cause an energy crisis in the United States.

1979 Seventy-two licensed nuclear reactors generate 12% of the electricity produced commercially in the United States.

1979 *T.T. Jahre Viking,* the world's largest oil supertanker (in fact the largest floating object on Earth) returns to the high seas after extensive repair work; she had been sunk by Iraqi jets during the Iran–Iraq war. This ship is over 1,500 feet long and can carry more than 4 million barrels of crude oil.

1979 The worst accident in U.S. commercial reactor history occurs on March 28th at the Three Mile Island nuclear power station near Harrisburg, Pennsylvania. It is caused by a loss of coolant from the reactor core due to a combination of mechanical malfunction and human error. No one is injured, and no overexposure to radiation results from the accident.

1980–1989

1980 A solar power system is installed at Natural Bridges National Monument (NBNM) in Utah by a partnership of the Massachusetts Institute of Technology, the U.S. Department of Energy, and the National Park Service. The system provides uninterrupted electricity to NBNM for 10 years until being retired (restored in the 1990s).

1980 Heating oil is the first energy product sold on the New York Mercantile Exchange (NYMEX) futures market. This is followed by crude oil (1983), natural gas (1990), and electricity (1996).

1980 In a public referendum in Sweden, voters choose to build no new nuclear power plants, which at the time provide about 50% of their power. However, in 2003 the Swedish government admits that it cannot close the first few reactors on schedule because it has not find a suitable alternative source of power.

1980 Southern California Edison begins operation of a 10-megawatt experimental power plant at the Brawley, California geothermal field, with steam produced by Unocal (Union Oil Company of California). This is the first geothermal flash plant in the United States.

1980 The Alaska National Interest Lands Conservation Act greatly expands the Arctic National Wildlife Refuge (ANWR). However, the act leaves future lawmakers the option of authorizing oil and gas exploration in the area. ANWR becomes the focus of an intense debate over the balance between environmental protection and energy security.

1980 The U.S. Department of Energy initiates the Three Mile Island research and development program to develop technology for disassembling and de-fueling the damaged reactor. The program will continue for 10 years and make significant advances in nuclear safety technology.

1980 The drama series *Dallas* begins a five-year run as the most popular show on U.S. television, and also wins large audiences abroad. It features a family of Texas oil millionaires, and it continues the image established by Hollywood films such as *Boom Town* (1940) and *Giant*

(1956) of the American oil business as a world of dynamic, risk-taking individualists.

1980 The Energy Security Act is signed by President Carter, with provisions for energy conservation, solar and geothermal energy, and biomass and alcohol fuels. The act also creates the Synthetic Fuels Corporation (see next), with an initial authorization of $20 billion to be allocated for joint projects with private industry.

1980 The Synthetic Fuels Corporation is established to develop tar sands and shale oils so as to reduce the nation's dependence on imported oil. It enjoys little commercial success due to low energy prices, lack of interest from the private sector, and the high cost of synfuel production. Congress ends the Corporation in 1985.

1980–1983 The Greens establish a political party in West Germany and gradually develop into an important faction. They attract support with their anti-nuclear campaign slogan: "*Atomkraft? Nein, Danke*" (Nuclear Power? No, Thanks). Eventually they will form part of a coalition that wins the German popular vote.

1980s Bicycle production in China reaches 40 million units per year, which at the time is greater than the total world output of automobiles.

1980s High efficiency (greater than 50%) is achieved for combined cycle gas/steam turbine power plants.

1981 An aircraft known as the *Solar Challenger* flies across the English Channel powered entirely by energy from the Sun.

1981 An explosive growth of wind farms in California begins as a result of favorable government support.

1981 French president François Miterrand inaugurates service for the high-speed French railway program known as TGV (*Train à Grande Vitesse*). The first TGV train with paying passengers leaves Paris five days later.

1981 Ormat Technologies successfully demonstrates the technical feasibility of larger-scale commercial binary power plants for geothermal energy, in the Imperial Valley of California.

1981 The Chinese government approves a proposal to build a pressurized light water reactor at Haiyan in Zhejiang province. The Qinshan 1 nuclear reactor will take a decade to complete.

1981 The *Columbia,* a manned, reusable space shuttle, is successfully launched into space and then returned to earth.

1981 U.S. President Ronald Reagan signs an executive order effectively ending price controls on crude oil and refined petroleum products. These controls on energy prices were the last legacy of the comprehensive wage and price controls that President Richard Nixon had enacted in 1971.

early 1980s The Energy Research Group at the University of Illinois (Urbana) pioneers the use of input–output analysis to study energy flow in economic and ecological systems. The group publishes more than 300 studies on the subject.

1982 The Shippingport, Pennsylvania Power Station is shut down after 25 years of service. Decommissioning will be completed in 1989; this is the first U.S. nuclear power plant to be decommissioned.

1982 Solar One, the first commercial solar-thermal power plant in the United States, opens near Daggett, California.

1982 The British National Radiological Protection Board issues an official estimate of the health impact of the 1957 fire at the Windscale nuclear plant. This indicates that 32 deaths and at least 260 cases of cancer could be attributed to the radiation released by the fire. However, other analysts claim that the fire actually caused more than 1,000 deaths.

1982 The Tokamak Fusion Test Reactor at the Princeton Plasma Physics Laboratory produces power from nuclear fusion. Tokamak will set records for high plasma temperatures and make other advances in fusion technology; the reactor is decommissioned in 1997 in the wake of cutbacks in funding for fusion research.

1983 A fire aboard the tanker *Castillo de Bellver,* off South Africa, releases 78.5 million gallons of crude oil, the largest tanker-related oil spill to date.

1983 Eastman Chemicals of Rochester, New York is the first coal-to-chemicals facility to produce acetyl chemicals from coal rather than petroleum. The use of domestic high-sulfur coal displaces oil previously used to produce acetyl chemicals.

1983 Kary Mullis conceives and helps develop polymerase chain reaction (PCR), a technology for rapidly multiplying fragments of DNA. This technique will have application in a wide range of scientific fields, such as molecular biology, gene analysis, biomedical research, clinical diagnosis, paleontology, and forensic science.

1983 Microsoft announces Windows, an extension of the MS-DOS operating system that will provide a graphical operating environment for PC users.

1983 President Reagan signs the Nuclear Waste Policy Act (NWPA), the first comprehensive U.S. nuclear waste legislation. NWPA establishes a program for the disposal of high-level radioactive waste, including spent fuel from nuclear power plants.

1983 The Itaipu Dam, a joint project of Brazil and Paraguay on the Parana river, begins operation of its first unit. When completed Itaipu will become the largest hydroelectric power facility in the world and will eventually supply about one-fourth of Brazil's electrical power and nearly all of Paraguay's.

1983 The U.S. Congress cuts off funding for the Clinch River Breeder Reactor Project, an ambitious but highly controversial plutonium-based breeder reactor program near Oak Ridge, Tennessee.

1983–1986 Faced with reduced demand for its oil and increased oil production from non-member nations, OPEC oil production is cut by more than one-half; a "free-for-all" situation prevails with a rapid rise in Saudi Arabia's production, culminating in an oil price collapse in July, 1986 with Arab Light selling at less than $8 per barrel.

1984 Advanco Corporation and McDonnell Douglas demonstrate the potential for a high-efficiency 25-kilowatt solar dish.

1984 Nuclear power overtakes hydropower to become the second largest source of electricity in the United States, after coal. Eighty-three nuclear power reactors provide about 14% of the electricity produced in the nation.

1984 The Annapolis Royal power station is established on the Bay of Fundy coastline of Nova Scotia, the first tidal power plant in North America. The Bay of Fundy is known for the largest tidal changes in the world (a difference between high and low tides of as much as 50 feet).

1984 The first in a series of commercial solar thermal electricity generation plants is established in California, based on parabolic trough technology.

1984 The Great Plains Synfuels Plant goes into operation in Beulah, North Dakota, the first commercial coal gasification plant in the United States. The plant converts North Dakota's abundant lignite coal to methane.

1984 The Macintosh computer is introduced by Apple, the first affordable personal computer with a graphical user interface and a 3.5-inch floppy disk drive. It has a CPU speed of 8 Mhz and initially sells for $2,495.

1984 The University of Wisconsin introduces the Domain Name System. In 1985 Symbolics Technology, Inc. registers the first domain name: symbolics.com. The second registration is made by BBN Corporation, followed by several universities (Carnegie-Mellon, Purdue, Rice, and University of California, Berkeley).

mid-1980s Many passenger cars are now available with diesel engines, as a fuel economy strategy in response to price shocks of the preceding decade. For the most part consumers will decide that the saving in fuel cost is not enough to compensate for other disadvantages of the diesel engine, and today a relatively small percentage of passenger cars are diesel-powered.

1990s Scholars use the term "rebound effect" to describe the relationship between an increase in energy efficiency and the ensuing rate of consumption. They argue that the rebound effect can exceed 100%; i.e., improved efficiency will produce greater consumption, not less. This supports Jevon's paradox (see 1865).

1985 Richard Smalley and others at Rice University in Texas identify a new form of pure carbon (in addition to graphite and diamonds), a large molecule of 60 carbon atoms resembling a soccer ball. This is called a "bucky ball" for architect Buckminister Fuller, who designed a geodesic dome with the same shape.

1985 Scientific concerns about damage to the ozone layer prompt various governments to adopt the Vienna Convention on the Protection of the Ozone Layer, which establishes an international legal framework for action.

1985 The U.S. Institute of Nuclear Power Operations forms a national academy to accredit every nuclear power plant's training program.

1985 Wind energy facilities totaling more than 1,000 megawatts are installed in California on the Tehachapi Pass between Barstow and Bakersfield and on the Altamont Pass east of Oakland.

mid-1980s A policy is introduced to set fixed subsidy tariffs for the purchase by utilities of electricity from renewable electricity providers (especially in North America and Europe).

1986 Ahmed Zaki Yamani, the flamboyant oil minister for Saudi Arabia, is fired by King Faud. Sheik Yamani was a powerful figure within OPEC who steered Saudi policy—and, indeed, much of OPEC policy—through the roller-coaster period of price increases and decreases of the 1970s and early 1980s.

1986 At the University of California's Lawrence Livermore Laboratory, the world's most powerful laser, Nova, creates a one-billionth of a second fusion reaction with temperatures and pressures greater than the Sun.

1986 Britain announces that canaries in coal mines will be replaced by electronic sensors. For centuries caged canaries had been brought into mines to signal the presence of dangerous gases, the idea being that a canary would succumb quickly to the gas and thus warn the miners to get out. "Canary in a coal mine" is a metaphor for an early warning of oncoming danger.

1986 Human error causes a core meltdown at the Chernobyl nuclear power plant in the Ukraine (former Soviet Union), the world's most severe nuclear accident. Thirty-one people die in the accident or soon after, and another 134 are treated for acute radiation syndrome. Total related deaths probably number in the thousands.

1986 Karl Alex Muller and Johannes Georg Bednorz discover a superconductor that operates at about 35 K. This sets off a flurry of research on "higher" temperature superconductors.

1986 Leaded gasoline is banned in the United States due to its health effects. By 2000 many other countries have also banned leaded gasoline, though it continues in use in the developing world.

1986 The International Thermonuclear Experimental Reactor (ITER) group is formed, an international collaborative aimed at demonstrating the scientific and technological feasibility of commercial fusion energy.

1986–1988 OPEC abandons the practice of fixed oil prices and switches to netback pricing, a market-based system that guarantees purchasers a certain refining margin. This enables Saudi Arabia to recapture a significant market share from the rest of OPEC; it also contributes to a collapse in oil prices. The netback system lasts just two years.

1986–1992 The Deep Gas Project in Sweden drills experimental wells in the Siljan Ring, a crater formed by a meteor impact, to test a theory that methane migrates upward from the inner regions of the Earth. Two wells are drilled, one to a depth of nearly five miles, but they encounter no commercial amounts of hydrocarbons.

1987 OPEC changes its pricing system of using Saudi Arabian Light as the reference price for other OPEC crudes and replaces it with a "basket" of seven crudes as reference: Sahara Blend of Algeria; Dubai Crude; Saudi Arabia's Arab Light; Minas of Indonesia; Bonny Light of Nigeria; Tia Inana of Venezuela; Isthmus of Mexico.

1987 The Montreal Protocol on Substances That Deplete the Ozone Layer is signed, requiring industrialized countries to reduce their consumption of chemicals harming the ozone layer; requirements are strengthened through later amendments (1990–1999).

1987 The U.S. Department of Energy announces President Reagan's approval of the construction of the Superconducting Super Collider (SSC), planned as the world's largest and most advanced particle accelerator. Congress decides to halt the project in 1993 after two billion dollars had been spent.

1987 The U.S. Nuclear Waste Policy Act (NWPA) is amended. Congress directs the Department of Energy to study only the potential of the Yucca Mountain, Nevada, site for the disposal of high-level radioactive waste.

ca. 1987 The human population reaches 5 billion.

1988 Alaska's oil production at Prudhoe Bay peaks at 2 million barrels per day.

1988 In the midst of record heat and drought in the United States, NASA scientist James Hansen tells a Congressional hearing he can say "with 99% confidence" that a long-term warming trend is underway, and that the greenhouse effect is to blame. The news media publicize Hansen's statements, bringing about a new level of public awareness of climate change.

1988 OPEC abandons a fixed-price system, replacing it with an agreement on "target prices."

1988 The government of Saudi Arabia officially establishes the Saudi Arabian Oil Company (Saudi Aramco). This is the successor to the Arabian American Oil Company (Aramco), the original developer of oil in the country, which the Saudi government had gradually taken over beginning in the early 1970s.

1988 The Intergovernmental Panel on Climate Change (IPCC) is formed; it becomes the leading scientific authority on climate change.

late 1980s A flowering of modern European wind technology takes place; wind energy also experiences dramatic growth worldwide.

1989 New York State Governor Mario Cuomo announces that the Shoreham Nuclear Power Plant, although already completed, will not go into service, due to intense opposition from local citizens.

1989 The Berlin Wall falls, as East Germany opens its borders with West Germany.

1989 The Long Valley Exploratory Well project commences near Mammoth Lakes, California. Theoretically, wells completed to depths near a magma body could bring that heat to the surface, and thus provide a new source of geothermal energy.

1989 The oil tanker *Exxon Valdez* runs aground on Alaska's Bligh Reef and spills an estimated 10.8 million gallons of crude oil, fouling the waters of Prince William Sound and eventually more than 1,000 miles of beach in south-central Alaska. This causes widespread environmental damage, which according to a December 2003 study continues even to this day.

1989 The world's first hybrid (organic Rankine/gas engine) geopressure-geothermal power plant begins operation at Pleasant Bayou, Louisiana, using both the heat and the methane of a geopressured resource.

1990–1999

1990 Iraq's Oil Minister demands that OPEC curb supplies until petroleum prices have risen to $25 a barrel (from the existing $15 level). Iraq President Saddam Hussein wants a price increase to help pay off the country's debt, and charges that Kuwait had conspired with the United States to depress world oil prices through overproduction

1990 Kuwait is invaded by Iraqi forces and Saddam Hussein declares the country to be a province of Iraq. Crude oil prices rise to above $30 a barrel amid great uncertainty in world markets. The United Nations then

embargoes the export of oil from Iraq in response to the invasion.

1990 Results of a Gallup poll indicate that about 75% of Americans consider themselves to be "environmentalists."

1990 The annual production of coal in the United States tops 1 billion tons for the first time.

1990 The city of Tokyo has approximately 1.5 million buildings with solar water heaters (more than in the entire United States); Israel uses solar water heating for approximately 30% of its buildings and all new homes there are required to install solar water heating systems.

1990 The number of vehicle-miles per year traveled by Americans exceeds 2 trillion for the first time.

1990 The U.S. Clean Air Act Amendments require many changes to gasoline and diesel fuels to make them pollute less, including mandated sale of oxygenated fuels in areas with unhealthy levels of carbon monoxide. This helps resurrect production of gasohol, an ethanol-blended gasoline, which had been used in the early days of the auto industry.

1990 The unified German government closes the last of eight nuclear power plants in the formerly Communist East Germany (DDR).

1990 Tim Berners-Lee, working with Robert Cailliau at CERN (European Organization for Nuclear Research), proposes a distributed information system, based on hypertext, a way of linking related pieces of information stored on computers. The name "World-Wide Web" is chosen for this system.

1990 With natural gas prices decontrolled, the New York Mercantile Exchange (NYMEX) launches its first natural gas futures contracts. Natural gas can now be bought and sold by consumers and producers at a predetermined time in the future. This leads to short-run volatility, but in principle, greater long-term price stability.

1990s Russia, after having become the world's largest oil producer in the late 1980s, experiences a decade-long slide in production. The causes are resource depletion, poor management, reduced maintenance, and curtailed exploration, all of which are exacerbated by the economic crisis following the collapse of the Soviet Union.

1991 In what will become known as the Persian Gulf War, the United States and allied countries launch Operation Desert Storm against Iraq to end its invasion of Kuwait. After about six weeks of fighting Iraqi forces are expelled from Kuwait and a cease-fire is declared.

1991 More than 600 oil wells are set ablaze in Iraq and Kuwait by retreating Iraqi forces. The fires burn for more than nine months, creating giant oil lakes in the desert and massive regional air and water pollution. In addition, 240 million gallons of crude oil are released to the marine environment from oil terminals and tankers, the largest oil spill ever.

1991 Operation Desert Storm causes the U.S. government to make a decision to withdraw oil from the SPR (Strategic Petroleum Reserve). This is the only occasion to date when the United States has made use of this emergency supply.

1991 Sony offers the first commercial version of the lithium ion (Li-Ion) battery. Li-Ion and its contemporary the NiMH (nickel metal hydride) battery offer greater capacity and will largely replace the earlier nickel cadmium version.

1991 Sweden introduces the first tax aimed specifically at reducing emissions of the greenhouse gas carbon dioxide. Consumers pay a tax of 0.25 SEK/kg (about $100 per ton) on the use of oil, coal, natural gas, liquefied petroleum gas, petrol, and aviation fuel. Renewable fuels are largely excluded, and industry pays a substantially lower rate.

1991 The European Energy Charter is adopted by 51 nations. Its objectives are to facilitate energy cooperation between countries, to improve the security of energy supplies, and to provide greater efficiency in the production, conversion, transport, distribution, and use of energy, as well as technology transfer.

1991 The first commercial nuclear reactor in China, Qinshan 1, is connected to the power grid.

1991 The First National People of Color Environmental Leadership Summit is held in Washington, D.C., with a focus on environmental justice (the principle that underrepresented groups are disproportionately impacted by the negative effects of power plant siting, uranium mining, and other energy-related facilities).

1991 The United States operates 111 nuclear power plants with a combined capacity of 99,673 megawatts. These plants produce almost 22% of the electricity generated commercially in the country.

1991 U.S. oceanographer John Martin stirs debate when he suggests that by sprinkling a relatively small amount of iron into certain areas of the ocean, one can create large algal blooms that could take in so much carbon from the atmosphere that they could reverse the greenhouse effect and cool the Earth.

early 1990s China passes the United States and becomes the world's largest producer of coal. Coal production has doubled in just 10 years due to the expansion of power generation and steel production to drive China's rapidly expanding economy.

early 1990s Vaclav Smil of Canada and Peter Vitousek of the United States independently show that nitrogen fixation from human activity (fossil fuels and fertilizer production/use) is greater than that from natural nitrogen fixation (terrestrial systems plus lightning). In

doing so, they document the extent of human alteration of global biogeochemistry.

1992 A National Energy Policy Act is signed into law under the Clinton administration. It makes several important changes in the licensing process for nuclear power plants.

1992 CERN releases the World-Wide Web. By the beginning of 2004 there are an estimated 50 million web sites.

1992 Daniel Yergin receives the Pulitzer Prize for general nonfiction for his work *The Prize: The Epic Quest for Oil, Money and Power,* which traces the history of oil and its social, political, strategic, and economic importance.

1992 The U.S. Environmental Protection Agency introduces Energy Star, a government-backed labeling program that helps businesses and individuals adopt technologies with greater energy efficiency.

1992 The United Nations Conference on Environment and Development (UNCED), popularly known as the Earth Summit, is held at Rio de Janeiro, Brazil, bringing together nearly 35,000 participants. Treatment of energy expands from environmental effects to economic importance as well.

1992 The University of South Florida fabricates a 15.89% efficient thin-film PV (photovoltaic) cell, exceeding the 15% efficiency level for the first time.

1993 A new tax on fuels is proposed by U.S. President Bill Clinton, based on the heat content of oil, gas, and coal. The tax would raise the price of gasoline by about 7.5 cents per gallon (gas price then about $1.05 per gallon). Congress rejects the tax in the face of opposition from special interest groups and from some segments of the public.

1993 NEMS (National Energy Modeling System) is released by the Energy Information Administration. It is the first computer-based, energy-economy modeling system of national energy markets for the express purpose of answering policy questions.

1993 President Clinton pledges that the United States will reduce its output of greenhouse gases to 1990 levels by the year 2000. Vice President Al Gore reaffirms this two years later.

1993 The 33M-VS, developed by Kenetech Windpower (U.S. Windpower), becomes one of the first commercially available, variable-speed wind turbines.

1993–1994 U.S. oil imports exceed domestic oil production for the first time; import dependence will continue to grow in the coming years.

1994 California Energy becomes the world's largest geothermal company through its acquisition of Magma Power. Near-term international markets gain the interest of U.S. geothermal developers.

1994 The Energy Charter Treaty is signed; the signatories include various nations of Europe, as well as the European Union itself, Japan, and Australia. The treaty's objective is to facilitate Western investment in the development of energy resources in Eastern Europe and the former Soviet Union.

1995 A United Nations resolution allows partial resumption of Iraq oil exports in "oil for food" deal; an embargo on Iraqi oil had been in effect since 1990 as a result of Iraq's invasion of Kuwait.

1995 An Intergovernmental Panel on Climate Change (IPCC) concludes that "the balance of evidence suggests a discernible human influence on global climate." By 2001 the IPCC states "there is new and stronger evidence that most of the warming observed over the last 50 years is attributable to human activities."

1995 The European Union enacts energy labeling standards for household refrigerators and freezers, followed by energy efficiency standards (MEPS) in 1999.

1995 The nongovernmental organization ENERGIA (the International Network on Gender and Sustainable Energy) is founded after the Fourth World Conference on Women in Beijing. ENERGIA coordinates diverse initiatives, individuals, and organizations to strengthen the role of women in sustainable energy development.

1996 The U.S. Department of Energy announces the National Center for Photovoltaics, headquartered in Golden, Colorado.

1997 Qatar installs the world's first significant liquid natural gas (LNG) exporting facility.

1997 Ridership in the Moscow subway system, the world's largest, reaches a total of 3.2 billion riders per year (followed by Tokyo, Seoul, Mexico City, and New York). The system, built in 1935, also is one of the world's longest, with 340 kilometers of track.

1997 The first modern hybrid-electric vehicle is marketed in Japan, the Toyota Prius.

1997 The so-called Kyoto Protocol is adopted by the United Nations Framework Convention on Climate Change at a convention in Kyoto, Japan; this commits the signing nations to binding cuts in greenhouse gas emissions.

1997 Widespread protests against shipments of high-level nuclear waste take place in Germany; in 1998 the government temporarily halts the transport of nuclear waste. Waste shipments, and protests in response, will begin again in 2000.

1997–1999 Registered lobbyists for the oil and gas industry report expenditures averaging $60 million per year to influence the energy policies of the U.S. government.

1998 Based on annual growth rings in trees and chemical evidence contained in marine fossils, corals, and ancient ice, scientists at the University of Massachusetts conclude that the 20th century was the warmest one in 600 years, and that 1990, 1995, and 1997 were the warmest years in all of the 600-year period.

1998 The Protocol on Environmental Protection (Madrid Protocol) is approved by a body of nations; it imposes a moratorium on mining and oil exploration in Antarctica for a minimum of 50 years.

1998 The World Commission on Dams is formed with initial funding from the World Bank and IUCN, and subsequent support from many other organizations around the world. The Commission's mandate includes evaluating the effectiveness of dams and developing standards for construction, operation, and decommissioning of dams.

1998–1999 In a wave of oil company mergers, BP announces plans to acquire Amoco for $48.2 billion; BPAmoco then acquires Atlantic Richfield (Arco). Chevron Corporation acquires Texaco for about $36 billion, and Exxon acquires Mobil for about $80 billion to create the world's largest oil company. Total Fina and Elf Aquitaine agree to merge, as do Conoco and Phillips.

late 1990s After increasing steadily for nearly two decades, improvements in new vehicle fuel efficiency in the United States, level off at about 28 miles per gallon. This is due to the expiration of government regulations (corporate average fuel economy standards) that mandated efficiency improvements, and consumer demand for vehicles.

late 1990s Green certificate trading is introduced in Europe, the United States, and Australia. Generators of electricity from renewable resources receive a certificate for a predetermined unit of energy produced. These certificates have a market value, thus providing a financial incentive for the use of renewable energy.

1999 A severe accident at the uranium processing plant at Tokaimura, Japan exposes 55 workers to radiation.

1999 An ice core two miles long, drilled out of an Antarctic ice sheet, shows that levels of greenhouse gases are higher now than at any time in the past 420,000 years.

1999 Azerbaijan, Georgia, and Turkey agree to the Baku-Ceyhan route for a major new pipeline for Azeri oil exports. It will run a total of 1,038 miles (1,730 km) and cost up to 4 billion dollars. Oil is set to flow in 2005.

1999 Japan announces "Top Runner," a new set of energy efficiency standards for improved performance of cars, appliances, and other products.

1999 The first modern hybrid-electric vehicle is marketed in the United States, the Honda Insight.

1999 The United States withdraws from the International Thermonuclear Experimental Reactor, an international collaborative aimed at demonstrating the scientific and technological feasibility of commercial fusion energy, after Congress eliminates the funding required for U.S. participation.

21st Century

2000 The year 2000 (Y2K) begins and with it the possible onset of the "Y2K Problem," a widely publicized potential disaster caused by failure of the world's computer systems to accommodate dates beyond 1999. In fact Y2K produces no significant negative effects; analysts disagree as to whether this is because of adjustments made beforehand or because the problem was exaggerated.

2000 The Chernobyl nuclear facility is permanently shut down.

2000 The world's first commercial wave power station on the Scottish island of Islay is connected to the electrical grid. Devices are placed on the shoreline or at sea, using wave motion to drive turbines or hydraulic pumps by means of compressed air.

2000–2001 Millions of people in California experience a new energy phenomenon known as a "rolling blackout"; this is a planned series of temporary, controlled power outages intended to prevent heavy demand from crashing the state's electrical grid. Californians are faced with a combination of increased energy prices and reduced, uncertain supply.

ca. 2000 The human population reaches 6 billion.

2001 Enron, an energy services corporation, goes bankrupt. It had become the seventh-largest U.S. company by buying electricity from generators and selling it to consumers, and was praised in the business media as the model for a new type of energy company. In reality Enron had been losing vast amounts of money and disguising the losses by false reports and accounting tricks.

2001 In the wake of the September 11th attacks on New York's World Trade Center and the Pentagon, the city of Boston bars the liquefied natural gas (LNG) carrier *Matthew* from Boston Harbor because of fear of a terrorist attack. The event highlights the heightened concerns in the United States about the vulnerability of energy facilities.

2001 The Center for Disease Control in the United States lists 15,196 deaths in 717 disasters since the 1830s in the nation's mines, most of them in coal mines.

2001 The Ecuadorean tanker *Jessica* runs aground on rocks off the Galapagos Islands, releasing about two-thirds of its cargo of 240,000 gallons of oil. The accident occurs in the prophetically named Shipwreck Bay, outside the tiny harbor of San Cristobal Island.

2001 The first oil flows through the Caspian Pipeline Consortium (CPC) pipeline, a 1,700-kilometer pipeline originating in Tennis, Kazakhstan, and terminating near the port of Novorossiysk, Russia, on the Black Sea. The CPC pipeline is the first to connect the oil-rich region of central Asia with significant export terminals.

2001 The U.S. Office of Scientific and Technical Information (OSTI) releases the Energy Citations Database, which provides 2 million citations to bibliographic records for energy and energy-related scientific and technical information.

2002 A general strike occurs in Venezuela against President Hugo Chavez. Oil production and exports drop to nearly zero, sending a shudder through world oil markets and helping oil prices to rise sharply.

2002 At a time when more than 50% of its electricity comes from nuclear power, Belgium makes a decision to completely phase out nuclear power over the next 25 years.

2002 President George W. Bush "un-signs" the United States from the Kyoto Protocol, the treaty aimed at reducing greenhouse gas emissions. President Bill Clinton had originally signed the Protocol in 1998, although it was never ratified by the Senate. President Bush says that it would cause undue economic harm to the U.S. economy.

2002 The Japanese government announces that it will expand its nuclear power generation by 30% by the year 2010, in order to achieve the reduction goals for greenhouse gas emission set by the Kyoto Protocol.

2002 The *Prestige*, a crippled oil tanker carrying nearly 20 million gallons of heavy fuel oil, breaks in half and sinks off Spain's northwest coast. The tanker's slow demise releases thousands of tons of oil into the waters of the Atlantic, causing one of Europe's worst environmental disasters.

2002 The World Summit on Sustainable Development (WSSD) is held in Johannesburg, South Africa. Treatment of energy extends to its overarching role in sustainable development, i.e., the integration of environmental stewardship, economic development, and human well-being.

2002–2003 Oil prices increase by 50% between November and March due to strikes in the Venezuelan oil sector, a shortage of home heating oil in a cold winter, and fears of what a U.S.–Iraq war will do to oil production and prices. Oil prices drop sharply in the opening days of the war when Iraq does not sabotage production facilities at home or in Kuwait.

2003 A massive power outage cuts electricity from New York north to Toronto and west to Detroit, an area in which some 50 million people reside.

2003 Hybrid-electric urban buses are delivered to the New York City Transit Authority by Orion Bus Industries, the first production models of such vehicles.

2003 In a historic recall election, California voters remove Governor Gray Davis from office. Dissatisfaction with his handling of the state's energy policy is cited as the leading factor in his defeat.

2003 Mandatory energy labeling programs are in place in 35 different countries, covering as many as 25 product types.

2003 The Calder Hall nuclear power station in West Cumbria is closed down by BNFL (British Nuclear Fuel). This marks the end of an era for the British nuclear industry; Calder Hall was the world's first commercial nuclear power facility when it opened in 1956.

ANNOTATED BIBLIOGRAPHY OF ENERGY SOURCES

Adams, R. N. (1988). *The Eighth Day: Social Evolution as the Self Organization of Energy.* Austin: University of Texas Press.

An anthropologist's view of the role of energy in shaping human institutions and culture.

Adelman, M. A. (1990). Mineral depletion, with special reference to petroleum. *The Review of Economics and Statistics* 72, 1–10.

A renowned resource economist studies the economics of petroleum supply.

Alexander, R. M. (1999). *Energy for Animal Life*. Oxford: Oxford University Press.

A good overview of the ways in which animals capture and allocate energy.

Arntzen, C. J., Dale, B. E. *et al*. (2000). *Biobased Industrial Products*. Washington, D.C.: National Academy Press.

A thorough overview of chemicals and other products produced from biomass.

Ayres, R. U. (1978). *Resources, Environment, and Economics: Applications of the Materials/Energy Balance Principle*. New York: John Wiley and Sons.

A thermodynamic analysis of economic production, substitution, technical change, resource depletion and pollution.

Barnett, H. J., and Morse, C. (1963). *Scarcity and Growth: The Economics of Natural Resource Availability, Resources for the Future*. Baltimore and London: The Johns Hopkins Press.

The first empirical analysis of the scarcity of energy and other natural resources in the United States from 1870 to 1957.

Berndt, E. R. (1978). Aggregate energy, efficiency, and productivity measurement. *Annual Review of Energy* 3, 225–273.

An insight into the issue of energy quality and its implications for understanding energy efficiency.

Blem, C. R. (2000). Energy balance. In *Sturkie's Avian Physiology* (G. C. Whittow, ed.), pp. 327–343. New York: Academic Press.

An authoritative description of bird energetics, with an emphasis on the energy costs and benefits of migration.

Bormann, F. H. (1976). An inseparable linkage: Conservation of natural ecosystems and the conservation of fossil energy. *BioScience* 26, 754–760.

A classic analysis of the link between fossil fuel use and the degradation of essential ecosystem services.

Brady, R. N. (1996). *Modern Diesel Technology.* Englewood Cliffs, New Jersey: Prentice Hall.

A good technical introduction to the subject of diesel engines.

Brander, J. A., and Taylor, M. S. (1998). The simple economics of Easter Island: A Ricardo-Malthus model of renewable resource use. *American Economic Review* 88(1), 119–138.

An economist's perspective on the single most well-studied collapse of a human society, the decline in the population and culture of Easter Island.

Bruce, N., Perez-Padilla, R., and Albalak, R. (2000). Indoor air pollution in developing countries: A major environmental and public health challenge. *Bulletin of the World Health Organization* 78, 1078–1092.

A review of the major health threat from indoor air pollution, much of which is driven by the patterns of energy use in developing nations.

Bullard, C., and Herendeen, R. (1975). The energy cost of goods and services. *Energy Policy* 1, 268–277.

The first use of input–output analysis to quantify energy requirements for various goods and services.

Bullard, C., Fox-Penner, P., and Pilati, D. (1978). Net energy analysis: A handbook for combining process and input–output analysis. *Resources and Energy* 1, 267–313.

A thorough description of quantitative methods for calculating the energy costs of goods and services.

Cardwell, D. S. L. (1971). *From Watt to Clausius: The Rise of Thermodynamics in the Early Industrial Age*. London: Heinemann.

A history of the development of the science of thermodynamics and its practical applications.

Carnahan, W., Ford, K. W., Prosperetti, A., Rochlin, G. I., Rosenfeld, A., Ross, M., Rothberg, J., Seidel, G., and Socolow, R. H. (eds.) (1975). *Efficient Use of Energy, The APS Studies on the Technical Aspects of the More Efficient Use of Energy*. New York: American Institute of Physics.

A group of distinguished physical scientists and engineers look at the potential for energy conservation in the wake of the 1973–1974 energy crisis.

Carnot, S. (1890). *Reflections on the Motive Power of Heat and on Machines Fitted to Develop That Power*. Translated by R. H. Thurston, New York.

Carnot's work was originally published in 1824 and, although long ignored, forms the scientific basis of modern thermodynamics.

Castaneda, C. J. (1999). *Invisible Fuel: Manufactured and Natural Gas in America, 1800–2000*. New York: Twayne Publishers.

An especially strong treatment of the early years of the natural gas industry.

Cipolla, C. M. (1965). *Guns, Sails and Empire: Technological Innovation in the Early Phases of European Expansion, 1400–1700*. New York: Minerva Press.

An excellent historical account of how wind energy, as captured by the sail, has been a pivotal technology in human history.

Clausius, R. J. E. (1867). *The Mechanical Theory of Heat*, 1st ed. (J. Hirst, ed.) London: Van Voorst.

Translated (with additions) from the 1864 German edition, though translators are not specified. Contains Clausius' main thermodynamical papers. The second edition is quite different, however.

Cleveland, C. J. (1992). Energy quality and energy surplus in the extraction of fossil fuels in the U.S. *Ecological Economics* 6, 139–162.

The techniques of net energy analysis applied to U.S. energy resources.

Cleveland, C. J., Costanza, R., *et al.* (1984). Energy and the U.S. economy: A biophysical perspective. *Science* 255, 890–897.

An empirical investigation of the connection between energy, economic growth, and natural resource scarcity.

Costanza, R. (1980). Embodied energy and economic valuation. *Science* 210, 1219–1224.

An examination of how the energy cost of goods and services, including labor and government services, can influence the connection between energy use and economic production.

Cowan, C. W., and Watson, P. J. (eds.) (1992). *The Origins of Agriculture*. Washington, D.C.: Smithsonian Institution Press.

A good overview of the driving forces behind one of the fundamental energy transitions in human history.

Debeir, J. C., Deléage, and J.-P., Hémery, D. (1991) *In the Servitude of Power: Energy and Civilization Through the Ages*. London: Zed Books.

An excellent history of the role of energy conversion technologies and new energy sources in shaping human civilization.

DeCicco, J. M. (2001). *Fuel Cell Vehicles: Technology, Market, and Policy Issues*. Warrendale, Pennsylvania: Society of Automotive Engineers.

A well-executed integration of the issues surrounding the problems and prospects for fuel cell vehicles.

Fox, R. F. (1988). *Energy and the Evolution of Life*. New York: W. H. Freeman and Company.

A description of the role of energy in Earth's early history and in the origin of life.

Freese, B. (2003). *Coal: A Human History*. Cambridge, Massachusetts: Perseus Books.

A social history of the effects of coal use, with emphasis on Britain, the U.S., and China.

Georgescu-Roegen, N. (1971). *The Entropy Law and the Economic Process*. Cambridge, Massachusetts: Harvard University Press.

A comprehensive and technical discussion of economic theory and the economic process as constrained by the laws of thermodynamics.

Gibbons, J. (1997). *This Gifted Age: Science and Technology at the Millennium*. Woodbury, New York: AIP Press.

The potential for energy conservation as seen by one of its earliest proponents.

Gipe, P. (1995). *Wind Energy Comes of Age*. New York: John Wiley and Sons.

A description of the rise of the modern wind industry and its prospects for the future.

Gössling, S. (2000). Sustainable tourism development in developing countries: Some aspects of energy use. *Journal of Sustainable Tourism* 8(5), 410–425.

An overview of energy and environmental issues associated with ecotourism.

Goudsblom, J. (1992). *Fire and Civilization*. London: Allen Lane.

An excellent history of the first and perhaps the most fundamental energy transition in human history.

Gregory, C. E. (2001). *A Concise History of Mining*. Lisse, The Netherlands: A. A. Balkema.

A history of mining with an emphasis on coal.

Griffin, J.M., and Steele, H. B. (1986). *Energy Economics and Policy*, 2nd ed. Orlando, Florida: Academic Press.

One of the few college-level textbooks devoted entirely to energy economics.

Grübler, A., Nakicenovic, N., and Nordhaus, W. D. (eds.) (2002). *Technological Change and the Environment*. Washington, D.C.: RFF Press.

A discussion of the manner in which historic technological change has reduced environmental impact of economic activity, and of future possibilities. Includes extensive climate-related material.

Herendeen, R. (1998). *Ecological Numeracy: Quantitative Analysis of Environmental Issues*. New York: John Wiley and Sons.

Basic quantitative literacy described from an energy and ecological perspective.

Hill, P., and Peterson, C. (1992). *Mechanics and Thermodynamics of Propulsion*, 2nd ed. Reading, Massachusetts: Addison-Wesley.

An engineering treatment of a variety of propulsion technologies, written at the college level.

Hoffmann, P. (2001). *Tomorrow's Energy: Hydrogen, Fuel Cells, and the Prospects for a Cleaner Planet*. Cambridge, Massachusetts: MIT Press.

The technology, economics, and policy aspects of a hydrogen economy.

Home, W. W., with Gittins, M. J. (1984). *The History of Classical Physics: A Selected, Annotated Bibliography*. New York: Garland, New York.

A widely used reference source.

Houghton, J. T., Ding, Y., Griggs, D. J., Noguer, M. P., van der Linden, J., and Xiaosu, D. (eds.) (2001). *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change (IPCC)*. Cambridge, UK: Cambridge University Press.

The authoritative scientific source on the detection, attribution and impacts of climate change.

Hubbert, M. K. (1974). *Energy Resources: A Review as of 1972*. Washington D.C.: U.S. Senate Report 93-40.

The author is the world's most well-known forecaster of oil supplies. This report summarizes the various methods he developed and applied to U.S. oil and gas resources.

Interlaboratory Working Group. (2000). *Scenarios for a Clean Energy Future*. Oak Ridge, Tennessee: Oak Ridge National Laboratory and Berkeley, California: Lawrence Berkeley National Laboratory.

A report that seeks to develop a better understanding of the potential for research and development programs and public policies to foster clean energy technology solutions for the United States.

Jevons, W. S. (1965, originally published 1865). *The Coal Question*. New York: Augustus M. Kelley.

A founder of the neoclassical revolution in economics, Jevons famously but erroneously predicts the collapse of British society due to the depletion of the nation's coal resources.

Johansson, T. B., Kelly, H., Reddy, A. K. N., and Williams, R. H. (eds.) (1993). *Renewable Energy: Sources for Fuels and Electricity*. Washington, D.C.: Island Press.

A thorough overview of the economic, technological, and environmental aspects of renewable energy.

Karl, T. L. (1997). *The Paradox of Plenty: Oil Booms and Petro-States*. Berkeley, California: University of California Press.

A study of the social, economic and political effects of a nation's dependence on oil exports.

Klass, D. L. (1998). *Biomass for Renewable Energy, Fuels, and Chemicals*. San Diego, California: Academic Press.

A thorough and wide-ranging treatment of the sources and conversion pathways of biomass energy into fuels and chemicals.

Lawlor, D. W. (2001). Photosynthesis. New York: Springer-Verlag, USA.

A thorough and detailed treatment of the energetics of photosynthesis.

Leach, G. (1976). *Energy and Food Production*. Surrey, UK: IPC Science and Technology Press.
One of the first quantitative treatments of this subject that uses the techniques of net energy analysis.

Lovins, A. (1976). Energy strategy: The road not taken. *Foreign Affairs* 55, 65–96.
A classic paper by a noted expert on the potential for energy conservation.

McNeill, W. H. (1989). *The Age of Gunpowder Empires, 1450–1800*. Washington, D.C.: American Historical Association.
A history of the development of the technology of gunpowder, and the role it has played in shaping the global geopolitical landscape.

Michalski, L., Eckersdorf, K., and McGhee, J. (1991). *Temperature Measurement*. New York: John Wiley and Sons.
A comprehensive overview of the subject.

Munasinghe, M. (1993). *Environmental Economics and Sustainable Development*. Washington, D.C.: The World Bank.
An integrated look at the economic, environmental, and social dimensions of sustainable development.

Nadel, S., Elliot, N., Shepard, M., Greenberg, S., Katz, G., and de Almeida, A. (2002). *Energy-Efficient Motor Systems*, 2nd ed. Washington, D.C.: American Council for an Energy-Efficient Economy.
A study of how motors and motor systems can achieve greater efficiency through motor management, optimized controls, improved component sizing and repair, better transmission hardware, and comprehensive monitoring and maintenance.

Nagy, K. A., Girard, I. A., and Brown, T. K. (1999). Energetics of free-ranging mammals, reptiles, and birds. *Annual Review of Nutrition* 19, 247–277.
A review of the current state of knowledge on how living organisms use energy.

Nakicenovic, N., Grübler, A., and McDonald, A. (eds.) (1998). *Global Energy Perspectives*. Cambridge, UK: Cambridge University Press (as publisher for the International Institute for Applied Systems Analysis and World Energy Council).
Long-term energy projections from a well-respected energy research group.

Nye, D. E. (1998). *Consuming Power: A Social History of American Energies*. Cambridge, Massachusetts: MIT Press.
An excellent history of the development of electricity in the U.S. and the social and economic consequences thereof.

Priest, J. (2000). *Energy: Principles, Problems, Alternatives*, 5th ed. Dubuque, Iowa: Kendall-Hunt.
An excellent college-level introduction to the basics of energy physics, conversion, sources, use, and technology.

Prigogine, I., and Stengers, I. (1984). *Order Out of Chaos*. New York: Bantam Books.
A physicist's description of how living systems, including human society, dissipate energy to build structure.

Righter, R. W. (1996). *Wind Energy in America: A History*. Norman, Oklahoma: University of Oklahoma Press.
A thorough history of the development of wind power in the U.S.

Sachs, J., and Warner, A. (1999). The big push: Natural resource booms and growth. *Journal of Development Economics* 59, 43–76.
A description of the macroeconomic effects of dependence on oil exports.

Salstrom, P. (1994). *Appalachia's Path to Dependency: Rethinking a Region's Economic History, 1730–1940*. Lexington, Kentucky: University of Kentucky Press.
A history of one of the most famous coal mining regions in the world.

Schipper, L. (ed.) (2000). *The Road from Kyoto: Current CO_2 and Transport Policies in the IEA*. Paris: International Energy Agency.
A review of transport-related CO_2 abatement policies in Denmark, Germany, the Netherlands, Sweden, the United Kingdom, and the United States.

Schurr, S. H. (1984). Energy use, technological change, and productive efficiency: An economic-historical interpretation. *Annual Review of Energy* 9, 409–425.
A good historical overview of the macroeconomic trends in energy use, economic growth and energy efficiency by one of the pioneers of this field of inquiry.

Simmons, I. G. (1996). *Changing the Face of the Earth. Culture, Environment, History*, 2nd ed. Cambridge, Massachusetts: Blackwell Publishers.
An excellent overview of the driving forces of global, regional, and local environmental change, with an insightful energy focus.

Smil, V. (1994). *Energy in World History*. Boulder, Colorado: Westview Press.

Human history as shaped by changes in energy sources and converters.

Smith, C. (1998). *The Science of Energy: A Cultural History of Energy Physics in Victorian Britain*. Chicago: University of Chicago Press.

A history of the development of the branches of physics devoted to the study of basic principles of energy.

Tainter, J. A. (1988). *The Collapse of Complex Societies*. Cambridge, UK: Cambridge University Press.

An insightful description of the role that environmental degradation and diminishing returns to complexity play in the collapse of societies.

Teal, J. M. (1962). Energy flow in the salt marsh ecosystem of Georgia. *Ecology* 43, 614–624.

An early classic study of ecosystem energetics.

Thesing, W. B. (ed.) (2000). *Caverns of Night: Coal Mines in Art, Literature, and Film*. Columbia, South Carolina: University of South Carolina Press.

A study of the social and cultural impacts that coal mining has had on American culture.

Tucker, V. A. (1975). The energetic cost of moving about. *American Scientist* 63(4), 413–419.

A description of the energetics of locomotion.

Turns, S. R. (2000). *An Introduction to Combustion: Concepts and Applications*. New York: McGraw Hill.

A college-level engineering introduction to the subject of combustion.

United Nations Department of Economic and Social Affairs. (2000). *World Energy Assessment*. New York: United Nations Development Programme.

A thorough and sweeping overview of global energy options, with a focus on the role of energy in poverty reduction and the policy interventions needed to shift the global energy system to a sustainable path.

United States Geological Survey. (2000). *World Petroleum Assessment*. Washington, D.C.: USGS.

An exhaustive review of oil and gas availability across the world, with an abundance of useful data and maps.

Verhoogen, J. (1980). *Energetics of the Earth*. Washington, D.C.: National Academy of Sciences.

A detailed examination of energy flows on Earth from a physical science perspective.

Veziroglu, T. N. (2000). A quarter century of the hydrogen movement 1974–2000. *International Journal of Hydrogen Energy* 25, 1143–1150.

A thorough recounting of a quarter-century's progress in hydrogen energy.

Vincent, C. A., and Scrosati, G. (1997). *Modern Batteries*, 2nd ed. London: Edward Arnold.

A good introduction to this important technology.

White, L. (1959). *The Evolution of Culture*. New York: McGraw-Hill.

An anthroplogist's view of the role of energy in shaping human culture.

Wigley, T. M. L. (ed.) (2000). *The Carbon Cycle*. Cambridge, UK: Cambridge University Press.

Leading scientists examine how atmospheric carbon dioxide concentrations have changed in the past and how this may affect the concentrations in the future. They offer approaches to modeling the carbon cycle, providing mathematical tools for predicting future levels of carbon dioxide.

World Energy Council. (2001). *Survey of World Energy Resources*. London: World Energy Council.

This biennial publication provides valuable data on resources and reserves for all forms of energy, as well as individual country profiles.

Yergin, D. (1991). *The Prize: The Epic Quest for Oil, Money and Power*. New York: Simon and Schuster.

The monumental Pulitzer Prize-winning story about the history of oil.

INTERNATIONAL ENERGY DATA ORGANIZATIONS

WORLD ENERGY COUNCIL

http://www.worldenergy.org/

The World Energy Council (WEC) is a global multi-energy organization with member committees in over 90 countries, including most of the largest energy-producing and energy-consuming countries. The 80-year-old organization covers all types of energy, including coal, oil, natural gas, nuclear, hydro, and renewables, and is UN-accredited, nongovernmental, noncommercial, and nonaligned.

One notable publication is the *Survey of World Energy Resources*, which has data on resources, reserves, production, and consumption for conventional and alternative energy sources for most countries in the world. These data are available in HTML and XLS formats, and are free of charge.

INTERNATIONAL ENERGY AGENCY

http://www.iea.org/

The International Energy Agency, based in Paris, is an autonomous agency linked with the Organization for Economic Cooperation and Development (OECD) (http://www.oecd.org/home/).

Notable publications include the *World Energy Outlook*, which assesses current energy conditions and projects energy demand, energy supply, and CO_2 emissions from 20 to 30 years into the future. *Key World Energy Statistics* contains up-to-date data on the supply, transformation, and consumption of all major energy sources. *Energy Balances of OECD Countries* contains data on the supply and consumption of coal, oil, gas, electricity, heat, renewables, and waste presented as comprehensive energy balances expressed in million tonnes of oil equivalent. Historical tables summarize production, trade, and final consumption data as well as key energy and economic indicators. A companion volume, *Energy Statistics of OECD Countries*, presents corresponding data in comprehensive balances expressed in a common unit, million tonnes of oil equivalent. The IEA charges a fee for most of its data.

UNITED NATIONS STATISTICS DIVISION

http://unstats.un.org/unsd/energy/default.htm

The United Nations Statistics Division (UNSD) collects energy statistics and updates and maintains the Energy Statistics Database, which contains comprehensive energy statistics on more than 215 countries, regions, and areas for production, trade, and intermediate and final consumption (end use) for primary and secondary conventional, nonconventional, and new and renewable sources of energy.

The annual publication *Energy Statistics Yearbook* is a comprehensive collection of international energy statistics that provides a global framework of comparable data on long-term trends in the supply of mainly commercial primary and secondary forms of energy. The biennial publication *Energy Balances and Electricity Profiles* presents energy data for selected countries in a format showing an overall picture of energy production, conversion, and consumption for fuels utilized in the country.

Coal, crude petroleum, and electricity data are collected on a monthly basis and published in the *Monthly Bulletin of Statistics*, and the yearly production, trade, and consumption of commercial energy data are published in the *Statistical Yearbook*. The complete dataset of *Energy Statistics*, covering the period 1950–2000, is available on CD and in electronic format. The UN charges a fee for all of its energy data.

THE WORLD BANK

http://www.worldbank.org/data

The *World Development Indicators* (WDI) is the World Bank's annual compilation of data about development. WDI includes approximately 800 indicators in 87 tables, organized in six sections. Energy data include commercial use, annual growth, GDP per unit and per capita, depletion as share of GDP, net imports, investment in infrastructure, production, traditional fuel use as share of total energy use, and a number of electricity indicators. These data are available in print, CD-ROM, and online versions, and require a fee.

BP STATISTICAL REVIEW OF WORLD ENERGY

http://www.bpamoco.com/centres/energy/index.asp

This publication presents country-level data on energy production, use, prices, and reserves dating back to 1965 and is available in ZIP, XLS, and PDF formats. A new energy charting tool allows the user to view predetermined reports for each type of energy or chart-specific data according to energy type on a regional and annual basis. Charts and graphs can be created as images for exporting. These data are available free of charge.

FOOD AND AGRICULTURE ORGANIZATION

http://www.fao.org

The Food and Agriculture Organization of the United Nations was founded in 1945 with a mandate to raise levels of nutrition and standards of living, to improve agricultural productivity, and to better the condition of rural populations. Today, FAO is one of the largest specialized agencies in the United Nations system and the lead agency for natural resource and rural development. An intergovernmental organization, FAO has 183 member countries plus one member organization, the European Community.

The FAOSTAT database is an online and multilingual database containing over 1 million time-series records covering international statistics in a wide range of natural resource areas, especially forestry, fisheries, agriculture, and land use. Energy data include estimates of wood fuel use, both commercial and traditional. Data are available online (free of charge) and on CD-ROM (fee required).

ORGANIZATION OF PETROLEUM EXPORTING COUNTRIES (OPEC)

http://www.opec.org/

OPEC is an 11 nation organization that supplies a major share of global oil production; the members collectively agree on global oil production and/or price targets. The members are Algeria, Indonesia, Iran, Iraq, Kuwait, Libya, Nigeria, Qatar, Saudi Arabia, United Arab Emirates, and Venezuela.

One notable publication is the *Annual Statistical Bulletin*, which reports data on basic economic and demographic conditions in member countries, as well as detailed data on all aspects of oil production, use, imports, exports, earnings, and so on. Data are available in PDF and ZIP formats, and are free of charge.

ASIA PACIFIC ENERGY RESEARCH CENTRE

http://www.ieej.or.jp/aperc/

The Asia Pacific Energy Research Centre (APERC) was established as an affiliate to the Institute of Energy Economics, Japan. It covers all 21 nations of the Asia Pacific Economic Cooperation (APEC): Australia; Brunei Darussalam; Canada; Chile; People's Republic of China; Hong Kong, China; Indonesia; Japan; Republic of Korea; Malaysia; Mexico; New Zealand; Papua New Guinea; Peru; Philippines; Russia; Singapore; Chinese Taipei; Thailand; United States; and Vietnam.

One notable APERC publication is *Energy Balance Tables for the APEC Energy Demand and Supply Outlook*, which contains data on the trade, price, supply, and consumption of coal, oil, gas, electricity, heat, renewables, and waste presented as comprehensive energy balances for member countries. These data are available in PDF format. The annual *APEC Energy Demand and Supply Outlook* provides forecasts for 20 years out. These data are available free of charge.

ASEAN ENERGY CENTER

http://www.aseanenergy.org/index.htm

This is the energy database for the Association of Southeast Nations (Brunei Darussalam, Cambodia,

Indonesia. Lao PDR, Malaysia, Myanmar, Philippines, Singapore, and Vietnam). It contains data on the trade, price, supply, and consumption of coal, oil, gas, electricity, heat, renewables, and waste presented as comprehensive energy balances. The Energy Reporting Application is a Web-based application for accessing these data. The data also include detailed country profiles. These data are available free of charge.

MIDDLE EAST ECONOMIC SURVEY

http://www.mees.com/

The Middle East Economic Survey (MEES) is an authoritative source on oil and gas in the Middle East. It provides current and recent historical monthly data, OPEC member quotas and production, and spot market quotations for the OPEC reference basket of seven crude oils. Data are in HTML format and free of charge.

ENERGY INFORMATION ADMINISTRATION, U.S. DEPARTMENT OF ENERGY

http://www.eia.doe.gov/

The Energy Information Administration (EIA), created by Congress in 1977, is a statistical, modeling, and analysis agency of the U.S. Department of Energy. It provides policy-independent data, forecasts, and analyses to promote policymaking, efficient markets, and public understanding regarding energy and its interaction with the economy and the environment. Current and historical data for the U.S. are available in HTML, XLS, and PDF formats. Less detailed data are also available for the world in similar formats. Among its more notable publications are the *Annual Energy Outlook*, which forecasts energy market condition 20–25 years out, the *Annual Energy Review,* which includes historical data back to 1949, and the *International Energy Annual*. These data are available free of charge.

AMERICAN PETROLEUM INSTITUTE

http://www.api.org

API is the statistical arm of the U.S. oil and gas industry and represents more than 400 members involved in all aspects of the oil and natural gas industry. Notable data sources include *ACCESS API Online Data* and the *Basic Petroleum Data Book*. These sources include a wide array of data on all aspects of the industry: exploration, drilling, production, use, trade, transportation, refining, reserves, prices, finance and investment, offshore, and stocks for the U.S. Less detailed data sources are available for the world as a whole. The API charges a fee for most of its data.

OIL & GAS JOURNAL ENERGY DATABASE

http://orc.pennnet.com/energydatabase/energydatabase.cfm

This online data service has data for the U.S. on all aspects of the industry: exploration, drilling, production, use, trade, transportation, refining, reserves, prices, finance and investment, offshore, and stocks for the U.S. Data on industry operations appear in three databases: monthly and quarterly, annually, and weekly. Most data can be downloaded. There is a fee for most of these data.

BAKER-HUGHES RIG COUNT

http://www.bakerhughes.com/investor/rig/index.htm

This is the authority on oil and gas drilling and workover rigs around the world. (A rig rotates the drill pipe from surface to drill a new well to explore for, develop, and produce oil or natural gas.) Rig activity is an important business barometer for the drilling industry and its suppliers. The active rig count acts as a leading indicator of demand for products used in drilling, completing, producing, and processing hydrocarbons and represents an upstream indicator of the effort aimed at finding and producing oil and gas. Data are available in PDF and/or XLS formats, and are free of charge.

INTERNATIONAL ATOMIC ENERGY AGENCY

http://www.iaea.org

IAEA's Power Reactor Information System (PRIS) constitutes the most complete data bank on nuclear power reactors in the world. PRIS covers two kinds of data: general and design information on power reactors and information about operating experience with nuclear power plants. These data include

country-level data on electricity generation and the role of nuclear power in national energy budgets. These data are available free of charge.

NUCLEAR ENERGY AGENCY

http://www.nea.fr/

The Nuclear Energy Agency (NEA) is a specialized agency within the Organization for Economic Cooperation and Development (OECD). Notable publications include the annual *Nuclear Energy Data*, which is a compilation of essential statistics on nuclear energy in OECD countries. This provides the reader with a comprehensive overview on the status of and trends in the nuclear power and fuel cycle sector.

Another notable work is the annual *Uranium: Resources, Production and Demand*. This "Red Book," jointly prepared by the Nuclear Energy Agency and the International Atomic Energy Agency, is the foremost world reference on uranium. This world report provides compilations of statistics on resources, exploration, production, and demand for uranium for about 50 countries, and provides an international expert analysis of industry statistics and worldwide projections of nuclear energy growth, uranium requirements, and uranium supply. Both of these publications are available only in print form and require a fee.

WTRG ECONOMICS

http://www.wtrg.com

WTRG Economics gathers data on the oil and gas industry from industry, government, and proprietary sources. This site provides graphical overviews of spot and futures prices for crude oil, heating oil, natural gas, and motor gasoline; drilling and workover rig counts; and other industry barometers. The graphs are available free of charge; other data services require a fee.

NEW YORK MERCANTILE EXCHANGE

http://www.nymex.com

The New York Mercantile Exchange is the largest physical commodity exchange in the world. Here you can find up-to-the-minute and historical data on the spot and futures prices for the energy commodities traded on the exchange: home heating oil, crude oil, motor gasoline, natural gas, propane, coal, and electricity. These data are available free of charge.

WORLD RADIATION DATA CENTRE

http://wrdc-mgo.nrel.gov/

The World Radiation Data Centre (WRDC), located at the Main Geophysical Observatory in St. Petersburg, Russia, serves as a central depository for solar radiation data collected at over 1000 measurement sites throughout the world.

The WRDC centrally collects, archives, and publishes radiometric data from the world to ensure the availability of these data for research by the international scientific community. Data are available on line or via an email query system, and are free of charge. The WRDC archive contains the following measurements (not all observations are made at all sites):

Global solar radiation
Diffuse solar radiation
Downward atmospheric radiation
Sunshine duration
Direct solar radiation (hourly and instantaneous)
Net total radiation
Net terrestrial surface radiation (upward)
Terrestrial surface radiation
Reflected solar radiation
Spectral radiation components (instantaneous fluxes)

SOLAR RADIATION RESOURCE INFORMATION

http://rredc.nrel.gov/solar/

The National Renewable Energy Laboratory (NREL) maintains a substantial database of data pertaining to solar radiation in the United States. These data include the three most common measurements of solar radiation (global horizontal, direct normal, and diffuse horizontal), basic meteorological data, maps, and other data for hundreds of stations. These data are available on hourly, daily, monthly, and annual bases. Some data are available in time-series format. Most data are available online and are free of charge.

MAPS OF RENEWABLE ENERGY RESOURCES

http://www.nrel.gov/gi4s/index_of_maps.html

This site provides dynamically generated maps of renewable energy resources that determine which energy technologies are viable solutions in the United States. The National Renewable Energy Laboratory (NREL) analyzes the resources and inputs the data into the GIS (Geographic Information System). The software package that is being used is an ESRI product called ArcIMS (http://www.esri.com). The following sets of maps are available:

Renewable Energy Atlas of the West: Interactive maps with zoom-in capability that highlight the wind, biomass, geothermal, and solar resources in the 11 western states.

Federal Energy Management Program Maps: Maps showing the market potential for various technologies at federal facilities throughout the country.

Maps of Indian Lands: Created in support of the report Energy Consumption and Renewable Energy Development Potential on Indian Lands by the Energy Information Administration, U.S. Department of Energy.

Solar Maps: Maps of solar radiation resources available for several photovoltaic collector orientations in the United States.

Transportation Technologies: Interactive geographical map of the Clean Cities coalitions boundaries and the Alternative Fuel Station Locator mapping application.

Wind Maps: Maps of gridded wind resource potential are based on wind power density.

INTERNATIONAL GEOTHERMAL ASSOCIATION

http://iga.igg.cnr.it/

The International Geothermal Association (IGA), founded in 1988, is a scientific, educational, and cultural organization established to operate worldwide. It has more than 2000 members in 65 countries.

IGA's objective is to encourage research, development, and utilization of geothermal resources worldwide through the compilation, publication, and dissemination of scientific and technical data and information, both within the community of geothermal specialists and between geothermal specialists and the general public. The IGA Web site has country-level data on geothermal direct use and power generation as well as individual country reports. The data are in HTML format and free of charge.

OREGON INSTITUTE OF TECHNOLOGY GEO-HEAT CENTER

http://geoheat.oit.edu/

Resource and direct use maps for the United States and maps of potential user communities in 10 states in the western United States. Resource maps for the 10 western states are accompanied by county-by-county data on well and spring characteristics and geo-referenced location data. The CD-ROM has data in XLS and CVS format, and requires a modest fee.

SELECTED ENERGY-RELATED TABLES

1. GUIDE TO THE USE OF THE INTERNATIONAL SYSTEM OF UNITS

Source: U.S. National Institute of Standards and Technology

The International System of Units was established in 1960 by the 11th General Conference on Weights and Measures (CGPM). Universally abbreviated SI (from the French Le Système International d'Unités), it is the modern metric system of measurement used throughout the world.

SI units are currently divided into three classes:

- base units
- derived units
- supplementary units

which together form what is called "the coherent system of SI units." The SI also includes prefixes to form decimal multiples and submultiples of SI units.

1.1 SI Units

1.1.1 SI Base Units

Table 1 gives the seven base quantities, assumed to be mutually independent, on which the SI is founded; and the names and symbols of their respective units, called "SI base units." Definitions of the SI base units are given in Section 1.4 to follow.

TABLE 1

SI Base Units

	SI base unit	
Base quantity	Name	Symbol
Length	meter	m
Mass	kilogram	kg
Time	second	s
Electric current	ampere	A
Thermodynamic temperature	kelvin	K
Amount of substance	mole	mol
Luminous intensity	candela	cd

1.1.2 SI Derived Units

Derived units are expressed algebraically in terms of base units or other derived units. The symbols for derived units are obtained by means of the mathematical operations of multiplication and division. For example, the derived unit for the derived quantity molar mass (mass divided by amount of substance) is the kilogram per mole, symbol kg/mol. Additional examples of derived units expressed in terms of SI base units are given in Table 2.

TABLE 2

Examples of SI Derived Units Expressed in Terms of SI Base Units

Derived quantity	Name	Symbol
Area	square meter	m^2
Volume	cubic meter	m^3
Speed, velocity	meter per second	m/s
Acceleration	meter per second squared	m/s^2
Wave number	reciprocal meter	m^{-1}
Mass density (density)	kilogram per cubic meter	kg/m^3
Specific volume	cubic meter per kilogram	m^3/kg
Current density	ampere per square meter	A/m^2
Magnetic field strength	ampere per meter	A/m
Amount-of-substance concentration (concentration)	mole per cubic meter	mol/m^3
Luminance	candela per square meter	cd/m^2

1.1.2.1 SI Derived Units with Special Names and Symbols Certain SI derived units have special names and symbols; these are given in Table 3a and Table 3b.

TABLE 3a

SI Derived Units with Special Names and Symbols, Including the Radian and Steradian

Derived quantity	Special name	Special symbol	Expression in terms of other SI units	Expression in terms of SI base units
Plane angle	radian	rad		$m \cdot m^{-1} = 1$
Solid angle	steradian	sr		$m^2 \cdot m^{-2} = 1$
Frequency	hertz	Hz		s^{-1}
Force	newton	N		$m \cdot kg \cdot s^{-2}$
Pressure, stress	pascal	Pa	N/m^2	$m^{-1} \cdot kg \cdot s^{-2}$
Energy, work, quantity of heat	joule	J	$N \cdot m$	$m^2 \cdot kg \cdot s^{-2}$
Power, radiant flux	watt	W	J/s	$m^2 \cdot kg \cdot s^{-3}$
Electric charge, quantity of electricity	coulomb	C		$s \cdot A$
Electric potential, potential difference, electromotive force	volt	V	W/A	$m^2 \cdot kg \cdot s^{-3} \cdot A^{-1}$
Capacitance	farad	F	C/V	$m^{-2} \cdot kg^{-1} \cdot s^4 \cdot A^2$
Electric resistance	ohm	Ω	V/A	$m^2 \cdot kg \cdot s^{-3} \cdot A^{-2}$
Electric conductance	siemens	S	A/V	$m^{-2} \cdot kg^{-1} \cdot s^3 \cdot A^2$
Magnetic flux	weber	Wb	$V \cdot s$	$m^2 \cdot kg \cdot s^{-2} \cdot A^{-1}$
Magnetic flux density	tesla	T	Wb/m^2	$kg \cdot s^{-2} \cdot A^{-1}$
Inductance	henry	H	Wb/A	$m^2 \cdot kg \cdot s^{-2} \cdot A^{-2}$
Celsius temperature	degree Celsius	°C		K
Luminous flux	lumen	lm	$cd \cdot sr$	$cd \cdot sr^a$
Illuminance	lux	lx	lm/m^2	$m^{-2} \cdot cd \cdot sr^a$

[a] The steradian (sr) is not an SI base unit. However, in photometry the steradian (sr) is maintained in expressions for units.

TABLE 3b

SI Derived Units with Special Names and Symbols Admitted for Reasons of Safeguarding Human Health[a]

Derived quantity	Special name	Special symbol	Expression in terms of other SI units	Expression in terms of SI base units
Activity (of a radionuclide)	becquerel	Bq		s^{-1}
Absorbed dose, specific energy (imparted), kerma	gray	Gy	J/kg	$m^2 \cdot s^{-2}$
Dose equivalent, ambient dose equivalent, directional dose equivalent, personal dose equivalent, equivalent dose	sievert	Sv	J/kg	$m^2 \cdot s^{-2}$

[a] The derived quantities to be expressed in the gray and the sievert have been revised in accordance with the recommendations of the International Commission on Radiation Units and Measurements (ICRU).

Degree Celsius In addition to the quantity thermodynamic temperature (symbol T), expressed in the unit kelvin, use is also made of the quantity Celsius temperature (symbol t) defined by the equation

$$t = T - T_0,$$

where $T_0 = 273.15\,K$ by definition. To express Celsius temperature, the unit degree Celsius, symbol °C, which is equal in magnitude to the unit kelvin, is used; in this case, "degree Celsius" is a special name used in place of "kelvin." An interval or difference of Celsius temperature can, however, be expressed in the unit kelvin as well as in the unit degree Celsius. [Note that the thermodynamic temperature T_0 is exactly 0.01 K below the thermodynamic temperature of the triple point of water.]

1.1.2.2 Use of SI Derived Units with Special Names and Symbols Examples of SI derived units that can be expressed with the aid of SI derived units having special names and symbols (including the radian and steradian) are given in Table 4.

TABLE 4

Examples of SI Derived Units Expressed with the Aid of SI Derived Units Having Special Names and Symbols

Derived quantity	Name	Symbol	Expression in terms of SI base units
Angular velocity	radian per second	rad/s	$m \cdot m^{-1} \cdot s^{-1} = s^{-1}$
Angular acceleration	radian per second squared	rad/s^2	$m \cdot m^{-1} \cdot s^{-2} = s^{-2}$
Dynamic viscosity	pascal second	$Pa \cdot s$	$m^{-1} \cdot kg \cdot s^{-1}$
Moment of force	newton meter	$N \cdot m$	$m^2 \cdot kg \cdot s^{-2}$
Surface tension	newton per meter	N/m	$kg \cdot s^{-2}$
Heat flux density, irradiance	watt per square meter	W/m^2	$kg \cdot s^{-3}$
Radiant intensity	watt per steradian	W/sr	$m^2 \cdot kg \cdot s^{-3} \cdot sr^{-1}$[a]
Radiance	watt per square meter steradian	$W/(m^2 \cdot sr)$	$kg \cdot s^{-3} \cdot sr^{-1}$[a]
Heat capacity, entropy	joule per kelvin	J/K	$m^2 \cdot kg \cdot s^{-2} \cdot K^{-1}$
Specific heat capacity, specific entropy	joule per kilogram kelvin	$J/(kg \cdot K)$	$m^2 \cdot s^{-2} \cdot K^{-1}$
Specific energy	joule per kilogram	J/kg	$m^2 \cdot s^{-2}$
Thermal conductivity	watt per meter kelvin	$W/(m \cdot K)$	$m \cdot kg \cdot s^{-3} \cdot K^{-1}$
Energy density	joule per cubic meter	J/m^3	$m^{-1} \cdot kg \cdot s^{-2}$
Electric field strength	volt per meter	V/m	$m \cdot kg \cdot s^{-3} \cdot A^{-1}$
Electric charge density	coulomb per cubic meter	C/m^3	$m^{-3} \cdot s \cdot A$
Electric flux density	coulomb per square meter	C/m^2	$m^{-2} \cdot s \cdot A$
Permittivity	farad per meter	F/m	$m^{-3} \cdot kg^{-1} \cdot s^4 \cdot A^2$
Permeability	henry per meter	H/m	$m \cdot kg \cdot s^{-2} \cdot A^{-2}$
Molar energy	joule per mole	J/mol	$m^2 \cdot kg \cdot s^{-2} \cdot mol^{-1}$
Molar entropy, molar heat capacity	joule per mole kelvin	$J/(mol \cdot K)$	$m^2 \cdot kg \cdot s^{-2} \cdot K^{-1} \cdot mol^{-1}$
Exposure (x and rays)	coulomb per kilogram	C/kg	$kg^{-1} \cdot s \cdot A$
Absorbed dose rate	gray per second	Gy/s	$m^2 \cdot s^{-3}$

[a] The steradian (sr) is not an SI base unit. However, in radiometry the steradian (sr) is maintained in expressions for units.

1.1.3 Decimal Multiples and Submultiples of SI Units: SI Prefixes

Table 5 gives the SI prefixes that are used to form decimal multiples and submultiples of SI units. A prefix attaches directly to the name of a unit, and a prefix symbol attaches directly to the symbol for a unit. For example, one kilometer, symbol 1 km, is equal to one thousand meters, symbol 1000 m or 10^3 m.

TABLE 5

SI Prefixes

Factor	Prefix	Symbol
$10^{24} = (10^3)^8$	yotta	Y
$10^{21} = (10^3)^7$	zetta	Z
$10^{18} = (10^3)^6$	exa	E
$10^{15} = (10^3)^5$	peta	P
$10^{12} = (10^3)^4$	tera	T
$10^9 = (10^3)^3$	giga	G
$10^6 = (10^3)^2$	mega	M
$10^3 = (10^3)^1$	kilo	k
10^2	hecto	h
10^1	deka	da
10^{-1}	deci	d
10^{-2}	centi	c
$10^{-3} = (10^3)^{-1}$	milli	m
$10^{-6} = (10^3)^{-2}$	micro	μ
$10^{-9} = (10^3)^{-3}$	nano	n
$10^{-12} = (10^3)^{-4}$	pico	p
$10^{-15} = (10^3)^{-5}$	femto	f
$10^{-18} = (10^3)^{-6}$	atto	a
$10^{-21} = (10^3)^{-7}$	zepto	z
$10^{-24} = (10^3)^{-8}$	yocto	y

1.2 Units Outside the SI

Certain units that are not part of the SI are essential and used so widely that they are generally accepted for use with the SI. These units are given in Table 6. The combination of units of this table with SI units to

TABLE 6

Units Accepted for Use with the SI

Name	Symbol	Value in SI units
Minute (time)	min	1 min = 60 s
Hour (time)	h	1 h = 60 min = 3600 s
Day (time)	d	1 d = 24 h = 86,400 s
Degree (plane angle)	°	$1^\circ = (\pi/180)$ rad
Minute (plane angle)	′	$1' = (1/60)^\circ = (\pi/10{,}800)$ rad
Second (plane angle)	″	$1'' = (1/60) = (\pi/648{,}000)$ rad
Liter	l, L (b)	$1\,L = 1\,dm^3 = 10^{-3}\,m^3$
Metric ton[a]	t	$1\,t = 10^3\,kg$

[a] This is the name to be used for this unit in the United States; it is also used in some other English-speaking countries. However, this unit is called “tonne” in many countries.

form derived units should be restricted to special cases in order not to lose the advantages of the coherence of SI units.

1.3 Units Temporarily Accepted for Use with the SI

Because of existing practice in certain fields or countries, in 1978 the CIPM considered that it was permissible for the units given in Table 7 to continue to be used with the SI until the CIPM considers that their use is no longer necessary.

TABLE 7

Units Temporarily Accepted for Use with the SI

Name[a]	Symbol	Value in SI units
nautical mile	-	1 nautical mile = 1852 m
knot	-	1 nautical mile per hour = (1852/3600) m/s
ångström	Å	$1\ \text{Å} = 0.1\ \text{nm} = 10^{-10}\ \text{m}$
are	a	$1\ \text{a} = 1\ \text{dam}^2 = 10^2\ \text{m}^2$
hectare[a]	ha	$1\ \text{ha} = 1\ \text{hm}^2 = 10^4\ \text{m}^2$
barn	b	$1\ \text{b} = 100\ \text{fm}^2 = 10^{-28}\ \text{m}^2$
bar	bar	$1\ \text{bar} = 0.1\ \text{Mpa} = 100\ \text{kPa} = 1000\ \text{hPa} = 10^5\ \text{Pa}$
gal	Gal	$1\ \text{Gal} = 1\ \text{cm/s}^2 = 10^{-2}\ \text{m/s}^2$
curie	Ci	$1\ \text{Ci} = 3.7 \times 10^{10}\ \text{Bq}$
roentgen	R	$1\ \text{R} = 2.58 \times 10^{-4}\ \text{C/kg}$
rad	rad[b]	$1\ \text{rad} = 1\ \text{cGy} = 10^{-2}\ \text{Gy}$
rem	rem	$1\ \text{rem} = 1\ \text{cSv} = 10^{-2}\ \text{Sv}$

[a] This unit and its symbol are used to express agrarian areas.
[b] When there is risk of confusion with the symbol for the radian, rd may be used as the symbol for rad.

1.4 Definitions of the SI Base Units and the Radian and Steradian

Meter The meter is the length of the path travelled by light in vacuum during a time interval of 1/299,792,458 of a second.

Kilogram The kilogram is the unit of mass; it is equal to the mass of the international prototype of the kilogram.

Second The second is the duration of 9,192,631,770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium-133 atom.

Ampere The ampere is that constant current that, if maintained in two straight parallel conductors of infinite length, of negligible circular cross section, and placed 1 meter apart in vacuum, would produce between these conductors a force equal to 2×10^{-7} newton per meter of length.

Kelvin The kelvin, a unit of thermodynamic temperature, is the fraction 1/273.16 of the thermodynamic temperature of the triple point of water.

Mole The mole is the amount of substance of a system that contains as many elementary entities as there are atoms in 0.012 kilogram of carbon 12. When the mole is used, the elementary entities must be specified and may be atoms, molecules, ions, electrons, other particles, or specified groups of such particles. In the definition of the mole, it is understood that reference is being made to unbound atoms of carbon 12, at rest and in their ground state. Note that this definition specifies at the same time the nature of the quantity whose unit is the mole.

Candela The candela is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency 540×10^{12} hertz and that has a radiant intensity in that direction of (1/683) watt per steradian.

Radian The radian is the plane angle between two radii of a circle that cut off on the circumference an arc equal in length to the radius.

Steradian The steradian is the solid angle that, having its vertex in the center of a sphere, cuts off an area of the surface of the sphere equal to that of a square with sides of length equal to the radius of the sphere.

1.5 A Brief History of the International System of Units

The creation of the decimal metric system at the time of the French Revolution and the subsequent deposition of two platinum standards representing the meter and the kilogram, on June 22, 1799, in the Archives de la République in Paris can be seen as the first step in the development of the present International System of Units.

In 1832, Gauss strongly promoted the application of this metric system, together with the second defined in astronomy, as a coherent system of units for the physical sciences. Gauss was the first to make *absolute* measurements of the earth's magnetic force in terms of a decimal system based on the *three mechanical units* millimeter, gram, and second for the quantities length, mass, and time, respectively. In later years, Gauss and Weber extended these measurements to include electrical phenomena.

These applications in the field of electricity and magnetism were further developed in the 1860s under the active leadership of Maxwell and Thomson through the British Association for the Advancement of Science (BAAS). They formulated the requirement for a *coherent system of units* with *base* units and *derived* units. In 1874 the BAAS introduced the *CGS system,* a three-dimensional coherent unit system based on the three mechanical units centimeter, gram, and second, using prefixes ranging from micro to mega to express decimal submultiples and multiples. The following development of physics as an experimental science was largely based on this system.

The sizes of the coherent CGS units in the fields of electricity and magnetism proved to be inconvenient, so in the 1880s the BAAS and the International Electrical Congress, predecessor of the International Electrotechnical Commission (IEC), approved a mutually coherent set of *practical units*. Among them were the ohm for electrical resistance, the volt for electromotive force, and the ampere for electric current.

After the establishment of the Meter Convention on May 20, 1875, the CIPM concentrated on the construction of new prototypes taking the meter and kilogram as the base units of length and mass. In 1889 the first CGPM sanctioned the international prototypes for the meter and the kilogram. Together with the astronomical second as unit of time, these units constituted a three-dimensional mechanical unit system similar to the CGS system, but with the base units meter, kilogram, and second.

In 1901 Giorgi showed that it is possible to combine the mechanical units of this meter–kilogram–second system with the practical electric units to form a single, coherent, four-dimensional system by adding to the three base units a fourth base unit of an electrical nature, such as the ampere or the ohm, and rewriting the equations occurring in electromagnetism in the so-called rationalized form. Giorgi's proposal opened the path to a number of new developments.

After the revision of the Meter Convention by the sixth CGPM in 1921, which extended the scope and responsibilities of the BIPM to other fields in physics, and the subsequent creation of the CCE (now CCEM) by the seventh CGPM in 1927, the Giorgi proposal was thoroughly discussed by the IEC, the IUPAP, and other international organizations. This led the CCE to recommend, in 1939, the adoption of a four-dimensional system based on the meter, kilogram, second, and ampere, a proposal approved by the CIPM in 1946.

Following an international inquiry by the BIPM, which began in 1948, the 10th CGPM, in 1954, approved the introduction of the *ampere,* the *kelvin*, and the *candela* as base units for electric current, thermodynamic temperature, and luminous intensity, respectively. The name *International System of Units* (*SI*) was given to the system by the 11th CGPM in 1960. At the 14th CGPM in 1971, the current version of the SI was completed by adding the *mole* as base unit for amount of substance, bringing the total number of base units to seven.

2. COMMON SYMBOLS AND UNITS

Greek alphabet

A	**α**	alpha	I	**ι**	iota	P	**ρ**	rho
B	**β**	beta	K	**κ**	kappa	**Σ**	**σ ς**	sigma
Γ	**γ**	gamma	**Λ**	**λ**	lambda	T	**τ**	tau
Δ	**δ ∂**	delta	M	**μ**	mu	**Υ**	**υ**	upsilon
E	**ε**	epsilon	N	**ν**	nu	**Φ**	**ϕ φ**	phi
Z	**ζ**	zeta	**Ξ**	**ξ**	xi	X	**χ**	chi
H	**η**	eta	O	*o*	omicron	**Ψ**	**ψ**	psi
Θ	**θ ϑ**	theta	**Π**	**π**	pi	**Ω**	**ω**	omega

International system of unit prefixes

Prefix	Symbol	Multiple
exa	(E)	10^{18} (quintillions)
peta	(P)	10^{15} (quadrillions)
tera	(T)	10^{12} (trillions)
giga	(G)	10^{9} (billions)
mega	(M)	10^{6} (millions)
kilo	(k)	10^{3} (thousands)
hecto	(h)	10^{2} (hundreds)
deka	(da)	10^{1} (tens)
deci	(d)	10^{-1} (tenths)
centi	(c)	10^{-2} (hundredths)
milli	(m)	10^{-3} (thousandths)
micro	(μ)	10^{-6} (millionths)
nano	(n)	10^{-9} (billionths)
pico	(p)	10^{-12} (trillionths)
femto	(f)	10^{-15} (quadrillionths)
atto	(a)	10^{-18} (quintillionths)

International system derived units

Quantity	Name	Symbol	Dimension
Capacitance	farad	F	$A^2 s^4 kg^{-1} m^{-2}$
Charge	coulomb	C	A s
Conductance	siemens	S	$kg^{-1} m^{-2} s^3 A^2$ or Ω^{-1}
Energy	joule	J	$kg\,m^2\,s^{-2}$
Force	newton	N	$kg\,m\,s^{-2}$
Magnetic flux	tesla	T	$kg\,s^{-2}\,A^{-1}$
Polarizability		α	$A^2 s^4 kg^{-1}$
Potential	volt	V	$kg\,m^2\,s^{-3}\,A^{-1}$
Power	watt	W	$kg\,m^2\,s^{-3}$
Pressure	pascal	Pa	$kg\,m^{-1}\,s^{-2}$
Resistance	ohm	Ω	$kg\,m^2\,s^{-3}\,A^{-2}$

3. FUNDAMENTAL PHYSICAL CONSTANTS[a]

General Constants

Symbol	Quantity	Value	Units	Relative uncertainty (ppm)
Universal constants				
G	Newtonian constant of gravitation	6.672 59 (85)	$10^{-11}\,m^3\,kg^{-1}\,s^{-2}$	128
μ_0	permeability of vacuum	$4\pi \times 10^{-7}$	$N\,A^{-2}$	
		$\approx 12.566\,370\,614$	$10^{-7}\,N\,A^{-2}$	
ε_0	permittivity of vacuum	$1/\mu_0 c^2$		
		$\approx 8.854\,187\,817$	$10^{-12}\ F\,m^{-1}$	
h	Planck constant	6.626 0755 (40)	10^{-34} Js	0.60
	in electron volts, $h/\{e\}$	4.135 6692 (12)	10^{-15} eVs	0.30
$\hbar$	$h/2\pi$	1.054 572 66 (63)	10^{-34} Js	0.60
	in electron volts, $\hbar/\{e\}$	6.582 1220 (20)	10^{-16} eVs	0.30
l_P	Planck length, $\hbar/m_P c = (\hbar G/c^3)^{1/2}$	1.616 05 (10)	10^{-35} m	64
m_P	Planck mass, $(\hbar c/G)^{1/2}$	2.176 71 (14)	10^{-8} kg	64
t_P	Planck time, $l_P/c = (\hbar G/c^5)^{1/2}$	5.390 56 (34)	10^{-44} s	64
c	speed of light in vacuum	2.997 924 58	$10^8\ ms^{-1}$	
Electromagnetic constants				
μ_B	Bohr magneton, $e\hbar/2m_e$	9.274 0154 (31)	$10^{-24}\,J\,T^{-1}$	0.34
	in electron volts, $\mu_B/\{e\}$	5.788 382 63 (52)	$10^{-5}\,eV\,T^{-1}$	0.089
	in hertz, μ_B/h	1.399 624 18 (42)	$10^{10}\,Hz\,T^{-1}$	0.30
	in kelvins, μ_B/k	0.671 7099 (57)	$K\,T^{-1}$	8.5
	in wavenumbers, μ_B/hc	46.686 437 (14)	$m^{-1}\,T^{-1}$	0.30
e	elementary charge	1.602 177 33 (49)	10^{-19} C	0.30
$2e/h$	Josephson frequency–voltage ratio	4.835 9767 (14)	$10^{14}\,Hz\,V^{-1}$	0.30
Φ_0	magnetic flux quantum, $h/2e$	2.067 834 61 (61)	10^{-15} Wb	0.30
μ_N	nuclear magneton, $e\hbar/2m_P$	5.050 7866 (17)	$10^{-27}\,J\,T^{-1}$	0.34
	in election volts, $\mu_N/\{e\}$	3.152 451 66(28)	$10^{-8}\,eV\,T^{-1}$	0.089
	in hertz, μ_N/h	7.622 5914 (23)	$MHz\,T^{-1}$	0.30
	in kelvins, μ_N/k	3.658 246 (31)	$10^{-4}\,K\,T^{-1}$	8.5
	in wavenumbers, μ_N/hc	2.542 622 81 (77)	$10^{-2}\,m^{-1}\,T^{-1}$	0.30
e^2/h	quantized Hall conductance	3.874 046 14 (17)	10^{-5} S	0.045
R_H	quantized Hall resistance, $h/e^2 = \mu_0 c/2\alpha$	25,812.8056 (12)	Ω	0.045

Atomic Constants

Symbol	Quantity	Value	Units	Relative uncertainty (ppm)
a_0	Bohr radius, $\alpha/4\pi\,R_\infty$	0.529 177 249 (24)	10^{-10} m	0.045
α	fine-structure constant, $\mu_0 c e^2/2h$	7.297 353 08 (33)	10^{-3}	0.045
α^{-1}	inverse fine-structure constant	137.035 9895 (61)		0.045
E_h	Hartree energy, $e^2/4\pi\varepsilon_0 a_0 = 2R_\infty hc$	4.359 7482 (26)	10^{-18} J	0.60
	in eV, $E_h/\{e\}$	27.211 3961 (81)	eV	0.30

Continues

Table continued

Symbol	Quantity	Value	Units	Relative uncertainty (ppm)
R_∞	Rydberg constant, $m_e c\alpha^2/2h$	10,973,731.534 (13)	m^{-1}	0.0012
	in eV, $R_\infty hc/\{e\}$	13.605 6981 (40)	eV	0.30
	in hertz, $R_\infty c$	3.289 841 9499 (39)	10^{15} Hz	0.0012
	in joules, $R_\infty hc$	2.179 8741 (13)	10^{-18} J	0.60
$h/2m_e$	quantum of circulation	3.636 948 07 (33)	10^{-4} m^2 s^{-1}	0.089
h/m_e		7.273 896 14 (65)	10^{-4} m^2 s^{-1}	0.089
Electron				
r_e	Classical electron radius, $\alpha^2 a_0$	2.817 940 92 (38)	10^{-15} m	0.13
λ_C	Compton wavelength, $h/m_e c$	2.426 310 58 (22)	10^{-12} m	0.089
$\lambda\!\!\!^{-}_C$	$\lambda_C/2\pi = \alpha a_0 = \alpha^2/4\pi R_\infty$	3.861 593 23 (35)	10^{-13} m	0.089
m_e/m_α	electron–α-particle mass ratio	1.370 933 54 (3)	10^{-4}	0.021
m_e/m_d	electron–deuteron mass ratio	2.724 437 07 (6)	10^{-4}	0.020
g_e	electron g-factor, $2(1+a_e)$	2.002 319 304 386 (20)		1×10^{-5}
μ_e	electron magnetic moment	928.477 01 (31)	10^{-26} J T^{-1}	0.34
μ_e/μ_B	in Bohr magnetons	1.001 159 652 193 (10)		1×10^{-5}
μ_e/μ_N	in nuclear magnetons	1838.282 000 (37)		0.020
a_e	electron magnetic moment anomaly, μ_e/μ_B-1	1.159 652 193(10)	10^{-3}	0.0086
m_e	electron mass	9.109 3897 (54)	10^{-31} kg	0.59
	in atomic mass units	5.485 799 03 (13)	10^{-4} u	0.023
	in electron volts, $m_e c^2/\{e\}$	0.510 999 06 (15)	MeV	0.30
$M(e)$, M_e	electron molar mass	5.485 799 03 (13)	10^{-7} kg/mol	0.023
μ_e/μ_μ	electron–muon			
	magnetic moment ratio	206.766 967 (30)		0.15
m_e/m_μ	electron–muon mass ratio	4.836 332 18 (71)	10^{-3}	0.15
μ_e/μ_p	electron–proton			
	magnetic moment ratio	658.210 6881 (66)		0.010
m_e/m_p	electron–proton mass ratio	5.446 170 13 (11)	10^{-4}	0.020
$-e/m_e$	electron specific charge	−1.758 819 62 (53)	10^{11} C kg^{-1}	0.30
σ_e	Thomson cross section, $(8\pi/3)r_e^2$	0.665 246 16 (18)	10^{-28} m^2	0.27
Muon				
m_μ/m_e	muon–electron mass ratio	206.768 262 (30)		0.15
g_μ	muon g-factor, $2(1+a_\mu)$	2.002 331 846 (17)		0.0084
μ_μ	muon magnetic moment	4.490 4514 (15)	10^{-26} J T^{-1}	0.33
μ_μ/μ_B	in Bohr magnetons	4.841 970 97 (71)	10^{-3}	0.15
μ_μ/μ_N	in nuclear magnetons	8.890 5981 (13)		0.15
a_μ	muon magnetic moment anomaly, $[\mu_\mu/(e\hbar/2m_\mu)]-1$	1.165 9230 (84)	10^{-3}	7.2
m_μ	muon mass	1.883 5327 (11)	10^{-28} kg	0.61
	in atomic mass units	0.113 428 913 (17)	u	0.15
	in electron volts, $m_\mu c^2/\{e\}$	105.658 389 (34)	MeV	0.32
$M(\mu)$, M_μ	muon molar mass	1.134 289 13 (17)	10^{-4} kg/mol	0.15
μ_μ/μ_p	muon–proton			
	magnetic moment ratio	3.183 345 47 (47)		0.15
Proton				
σ_{H_2O}	diamagnetic shielding correction for protons in pure water, spherical sample, 25°C, $1-\mu_p'/\mu_p$	25.689 (15)	10^{-6}	

Continues

Table continued

Symbol	Quantity	Value	Units	Relative uncertainty (ppm)
$\lambda_{C,p}$	proton Compton wavelength, $h/m_p c$	1.321 410 02 (12)	10^{-15} m	0.089
$\lambda\!\!\!^{-}_{C,p}$	$\lambda_{C,p}/2\pi$	2.103 089 37 (19)	10^{-16} m	0.089
m_p/m_e	proton–electron mass ratio	1836.152 701 (37)		0.020
γ_p	proton gyromagnetic ratio	26,752.2128 (81)	$10^4\,s^{-1}\,T^{-1}$	0.30
$\gamma_p/2\pi$		42.577 469 (13)	MHz T^{-1}	0.30
γ'_p	uncorrected (H_2O, sph., 25°C)	26,751.5255 (81)	$10^4\,s^{-1}\,T^{-1}$	0.30
$\gamma'_p/2\pi$		42.576 375 (13)	MHz T^{-1}	0.30
μ_p	proton magnetic moment	1.410 607 61 (47)	10^{-26} J T^{-1}	0.34
μ_p/μ_B	in Bohr magnetons	1.521 032 202 (15)	10^{-3}	0.010
μ_p/μ_N	in nuclear magnetons	2.792 847 386 (63)		0.023
m_p	proton mass	1.672 6231 (10)	10^{-27} kg	0.59
	in atomic mass units	1.007 276 470 (12)	u	0.012
	in electron volts, $m_p c^2/\{e\}$	938.272 31 (28)	MeV	0.30
$M(p)$, M_p	proton molar mass	1.007 276 470 (12)	10^{-3} kg/mol	0.012
m_p/m_μ	proton–muon mass ratio	8.880 2444 (13)		0.15
e/m_p	proton specific charge	9.578 8309 (29)	10^7 C kg^{-1}	0.30
μ'_p	shielded proton moment (H_2O, sph., 25°C)	1.410 571 38 (47)	10^{-26} J T^{-1}	0.34
μ'_p/μ_B	in Bohr magnetons	1.520 993 129 (17)	10^{-3}	0.011
μ'_p/μ_N	in nuclear magnetons	2.792 775 642 (64)		0.02
Neutron				
$\lambda_{C,n}$	neutron Compton wavelength, $h/m_n c$	1.319 591 10 (12)	10^{-15} m	0.089
$\lambda\!\!\!^{-}_{C,n}$	$\lambda_{C,n}/2\pi$	2.100 194 45 (19)	10^{-16} m	0.089
μ_n/μ_e	neutron–electron			
	magnetic moment ratio	1.040 668 82 (25)	10^{-3}	0.24
m_n/m_e	neutron–electron mass ratio	1838.683 662 (40)		0.022
μ_n	neutron magnetic moment	0.966 237 07 (40)	10^{-26} J T^{-1}	0.41
μ_n/μ_B	in Bohr magnetons	1.041 875 63 (25)	10^{-3}	0.24
μ_n/μ_N	in nuclear magnetons	1.913 042 75 (45)		0.24
m_n	neutron mass	1.674 9286 (10)	10^{-27} kg	0.59
	in atomic mass units	1.008 664 904(14)	u	0.014
	in electron volts, $m_n c^2/\{e\}$	939.565 63 (28)	MeV	0.30
$M(n)$, M_n	neutron molar mass	1.008 664 904 (14)	10^{-3} kg/mol	0.014
μ_n/μ_p	neutron–proton			
	magnetic moment ratio	0.684 979 34 (16)		0.24
m_n/m_p	neutron–proton mass ratio	1.001 378 404 (9)		0.009
Deuteron				
μ_d/μ_e	deuteron–electron			
	magnetic moment ratio	0.466 434 5460 (91)	10^{-3}	0.019
m_d/m_e	deuteron–electron mass ratio	3670.483 014 (75)		0.020
μ_d	deuteron magnetic moment	0.433 073 75 (15)	10^{-26} J T^{-1}	0.34
μ_d/μ_B	in Bohr magnetons	0.466 975 4479 (91)	10^{-3}	0.019
μ_d/μ_N	in nuclear magnetons	0.857 438 230 (24)		0.028
m_d	deuteron mass	3.343 5860 (20)	10^{-27} kg	0.59
	in atomic mass units	2.013 553 214 (24)	u	0.012
	in electron volts, $m_d c^2/\{e\}$	1875.613 39 (57)	MeV	0.30
$M(d)$, M_d	deuteron molar mass	2.013 553 214 (24)	10^{-3} kg/mol	0.012
μ_d/μ_p	deuteron–proton			
	magnetic moment ratio	0.307 012 2035 (51)		0.017
m_d/m_p	deuteron–proton mass ratio	1.999 007 496 (6)		0.003

Physico-Chemical Constants

Symbol	Quantity	Value	Units	Relative uncertainty (ppm)
m_u	atomic mass constant, $\frac{1}{12}m$ (^{12}C)	1.660 5402 (10)	10^{-27} kg	0.59
	in electron volts, $m_u c^2/\{e\}$	931.494 32 (28)	MeV	0.30
N_A, L	Avogadro constant	6.022 1367 (36)	10^{23} mol^{-1}	0.59
k	Boltzmann constant, R/N_A	1.380 658 (12)	10^{-23} J K^{-1}	8.5
	in electron volts, $k/\{e\}$	8.617 385 (73)	10^{-5} eV K^{-1}	8.4
	in hertz, k/h	2.083 674 (18)	10^{10} Hz K^{-1}	8.4
	in wavenumbers, k/hc	69.503 87 (59)	m^{-1} K^{-1}	8.4
F	Faraday constant	96,485.309 (29)	C mol^{-1}	0.30
R	molar gas constant	8.314 510 (70)	J mol^{-1} K^{-1}	8.4
$N_A h$	molar Planck constant	3.990 313 23 (36)	10^{-10} J s mol^{-1}	0.089
$N_A hc$		0.119 626 58 (11)	J m mol^{-1}	0.089
V_m	molar volume (ideal gas), RT/p			
	$T = 273.15$ K, $p = 101{,}325$ Pa	22.414 10 (19)	L/mol	8.4
n_0	Loschmidt constant, N_A/V_m	2.686 763 (23)	10^{25} m^{-3}	8.5
V_m	$T = 273.15$ K, $p = 100$ kPa	22.711 08 (19)	L/mol	8.4
c_1	radiation constant, first, $2\pi hc^2$	3.741 7749 (22)	10^{-16} W m^2	0.60
c_2	radiation constant, second, hc/k	0.014 387 69 (12)	m K	8.4
	Sackur–Tetrode constant (absolute entropy constant), $\frac{5}{2} + \ln\{(2\pi m_u k T_1/h^2)^{3/2} k T_1/p_o\}$			
S_o/R	$T_1 = 1$ K, $p_0 = 101{,}325$ Pa	−1.151 693 (21)		18
	$p_0 = 100$ kPa	−1.164 856 (21)		18
σ	Stefan–Boltzmann constant, $(\pi^2/60)k^4/\hbar^3 c^2$	5.670 51 (19)	10^{-8} W m^{-2} K^{-4}	34
b	Wien displacement law constant, $b = \lambda_{max} T \approx c_2/4.965\ 114\ 23$	2.897 756 (24)	10^{-3} m K	8.4

Standard Values/Maintained Units

Symbol	Quantity	Value	Units	Relative uncertainty (ppm)
u	atomic mass unit (unified), $1\text{ u} = m_u = \frac{1}{12}m(^{12}\text{C})$	1.660 5402(10)	10^{-27} kg	0.59
eV	electron volt, $(e/\text{C})\text{J} = \{e\}$ J	1.602 177 33 (49)	10^{-19} J	0.30
g_n	standard acceleration of gravity	9.806 65	ms^{-2}	
atm	standard atmosphere	1.013 25	10^5 Pa	
Maintained electrical units				
A_{BI85}	BIPM maintained ampere,	$1 - 6.03\ (30) \times 10^{-6}$	A	
	$A_{BIPM} = V_{76-B1}/\Omega_{69-BI}$	$= 0.999\ 993\ 97\ (30)$	A	0.30
Ω_{BI85}	BIPM maintained ohm, Ω_{69-BI}	$1 - 1.563\ (50) \times 10^{-6}$	Ω	
	$\Omega_{BI85} \equiv \Omega_{69-BI}$ (1 Jan. 1985)	$= 0.999\ 998\ 437\ (50)$	Ω	0.050
$\frac{d\Omega_{69-BI}}{dt}$	Drift rate of Ω_{69-BI}	−0.0566 (15)	μΩ/a	
V_{76-BI}	BIPM maintained volt,	$1 - 7.59\ (30) \times 10^{-6}$	V	
	$V_{76-BI} \equiv 483{,}594$ GHz $(h/2e)$	$= 0.999\ 992\ 41\ (30)$	V	0.30

Continues

Table continued

Symbol	Quantity	Value	Units	Relative uncertainty (ppm)
X-ray standards				
Å*	Å*: $\lambda(\mathrm{WK}\alpha_1) \equiv 0.209\,100\,\text{Å}^*$	1.000 014 81 (92)	10^{-10} m	0.92
xu(CuKα_1)	Cu x-unit: $\lambda(\mathrm{CuK}\alpha_1) \equiv 1537.400$ xu	1.002077 89 (70)	10^{-13} m	0.70
a	lattice spacing of Si (in vacuum, 22.5°C),	0.543 101 96 (11)	nm	0.21
d_{220}	$d_{220} = a/\sqrt{8}$	0.192 015 540 (40)	nm	0.21
V_m(Si)	molar volume of Si, $M(\mathrm{Si})/\rho(\mathrm{Si}) = N_A a^3/8$	12.058 8179 (89)	cm^3/mol	0.74
xu(MoKα_1)	Mo x-unit: $\lambda(\mathrm{MoK}\alpha_1) \equiv 707.831$ xu	1.002 099 38 (45)	10^{-13} m	0.45

[a] Recommended values for fundamental physical constants; adapted from E. R. Cohen and B. N. Taylor from *CODATA Bulletin*, number 63, November 1986, and *Journal of Research of the National Bureau of Standards*, volume 92, number 2, pages 85–95, 1987. Values in bold are exact. In parentheses is the one-standard-deviation uncertainty of the final digit of a given value.

4. CONVERSION FACTORS FOR THE INTERNATIONAL SYSTEM OF UNITS

Source: U.S. National Institute of Standards and Technology

4.1 Introduction

Section 4.6 gives factors for converting values of quantities expressed in various units—predominantly units outside the SI that are unacceptable for use with it—to values expressed either in (a) SI units, (b) units that are accepted for use with the SI (especially units that better reflect the nature of the unconverted units), (c) units formed from such accepted units and SI units, or (d) decimal multiples or submultiples of the units of (a) to (c) that yield numerical values of convenient magnitudes.

4.2 Notation

The factors given in Section 4.6 are written as a number equal to or greater than 1 and less than 10, with 6 or fewer decimal places. The number is followed by the letter E, which stands for exponent, a plus (+) or minus (–) sign, and two digits which indicate the power of 10 by which the number is multiplied. For example,

$$3.523\,907\,\mathrm{E}-02 \text{ means } 3.523\,907 \times 10^{-2} = 0.035\,239\,07$$

$$3.386\,389\,\mathrm{E}+03 \text{ means } 3.386\,389 \times 10^{3} = 3386.389$$

A factor in boldface is exact. All other factors have been rounded to the significant digits given in accordance with accepted practice. Where less than six digits after the decimal place are given, the unit does not warrant a greater number of digits in its conversion. However, for the convenience of the user, this practice is not followed for all such units, including the cord, cup, quad, and teaspoon.

4.3 Use of Conversion Factors

Each entry in Section 4.6 is to be interpreted as in these two examples:

To convert from	To	Multiply by
atmosphere, standard (atm)	pascal (Pa)	1.013 25 E+05
cubic foot per second (ft^3/s)	cubic meter per second (m^3/s)	2.831 685 E–02

means

$$1\ \text{atm} = 101\,325\ \text{Pa}\ \ \text{(exactly)}$$
$$1\ \text{ft}^3/\text{s} = 0.028\,316\,85\ \text{m}^3/\text{s}$$

Thus to express, for example, the pressure $p = 11.8$ standard atmospheres (atm) in pascals (Pa), write

$$p = 11.8\ \text{atm} \times 101\,325\ \text{Pa/atm}$$

and obtain the converted numerical value $11.8 \times 101\,325 = 1\,195\,635$ and the converted value $p = 1.20$ MPa.

4.4 Factor for Converting Motor Vehicle Efficiency

The efficiency of motor vehicles in the United States is commonly expressed in miles per U.S. gallon, while in most other countries it is expressed in liters per one hundred kilometers. To convert fuel economy stated in miles per U.S. gallon to fuel consumption expressed in L/(100 km), divide 235.215 by the numerical value of the stated fuel economy. Thus 24 miles per gallon corresponds to 9.8 L/(100 km).

4.5 U.S. Survey Foot and Mile

The U.S. Metric Law of 1866 gave the relationship 1 m = 39.37 in ("in" is the unit symbol for the inch). From 1893 until 1959, the yard was defined as being exactly equal to (3600/3937) m, and thus the foot was defined as being exactly equal to (1200/3937) m.

In 1959 the definition of the yard was changed to bring the U.S. yard and the yard used in other countries into agreement. Since then the yard has been defined as exactly equal to 0.9144 m, and thus the foot has been defined as exactly equal to 0.3048 m. At the same time it was decided that any data expressed in feet derived from geodetic surveys within the United States would continue to bear the relationship as defined in 1893, namely, 1 ft = (1200/3937) m (ft is the unit symbol for the foot). The name of this foot is "U.S. survey foot," while the name of the new foot defined in 1959 is "international foot." The two are related to each other through the expression 1 international foot = 0.999 998 U.S. survey foot exactly.

In Section 4.6, the factors given are based on the international foot unless otherwise indicated. Readers may also find the following summary of exact relationships helpful, where for convenience the symbols *ft* and *mi,* that is, ft and mi in italic type, indicate that it is the *U.S. survey foot* or *U.S. survey mile* that is meant rather than the international foot (ft) or international mile (mi), and where rd is the unit symbol for the rod and fur is the unit symbol for the furlong.

1 *ft* = (1200/3937) m
1 ft = 0.3048 m
1 ft = 0.999 998 *ft*
1 rd, pole, or perch = 161/2 *ft*
40 rd = 1 fur = 660 *ft*
8 fur = 1 U.S. survey mile (also called "statute mile") = 1 *mi* = 5280 *ft*
1 fathom = 6 *ft*
1 international mile = 1 mi = 5280 ft
2721/4 ft^2 = 1 rd^2
160 rd^2 = 1 acre = 43 560 ft^2
640 acre = 1 mi^2

4.6 Factors for Units Listed Alphabetically

Factors in boldface are exact.

To convert from	To	Multiply by	
abampere	ampere (A)	1.0	E + 01
abcoulomb	coulomb (C)	1.0	E + 01
abfarad	farad (F)	1.0	E + 09
abhenry	henry (H)	1.0	E−09
abmho	siemens (S)	1.0	E + 09
abohm	ohm Ω	1.0	E−09
abvolt	volt (V)	1.0	E−08
acceleration of free fall, standard (g_n)	meter per second squared (m/s^2)	9.806 65	E + 00
acre (based on U.S. survey foot)	square meter (m^2)	4.046 873	E + 03
acre foot (based on U.S. survey foot)	cubic meter (m^3)	1.233 489	E + 03
ampere hour (A · h)	coulomb (C)	3.6	E + 03
ångström (Å)	meter (m)	1.0	E−10
ångström (Å)	nanometer (nm)	1.0	E−01
astronomical unit (AU)	meter (m)	1.495 979	E + 11
barrel [for petroleum, 42 gallons (U.S.)] (bbl)	cubic meter (m^3)	1.589 873	E−01
barrel [for petroleum, 42 gallons (U.S.)] (bbl)	liter (L)	1.589 873	E + 02
biot (Bi)	ampere (A)	1.0	E + 01
British thermal unit$_{IT}$ (Btu_{IT})	joule (J)	1.055 056	E + 03
British thermal unit$_{IT}$ foot per hour square foot degree Fahrenheit [$Btu_{IT} \cdot ft/(h \cdot ft^2 \cdot °F)$]	watt per meter kelvin [W/(m · K)]	1.730 735	E + 00
British thermal unit$_{IT}$ inch per hour square foot degree Fahrenheit [$Btu_{IT} \cdot in/(h \cdot ft^2 \cdot °F)$]	watt per meter kelvin [W/(m · K)]	1.442 279	E−01
British thermal unit$_{IT}$ inch per second square foot degree Fahrenheit [$Btu_{IT} \cdot in/(s \cdot ft^2 \cdot °F)$]	watt per meter kelvin [W/(m · K)]	5.192 204	E + 02
British thermal unit$_{IT}$ per cubic foot (Btu_{IT}/ft^3)	joule per cubic meter (J/m^3)	3.725 895	E + 04
British thermal unit$_{IT}$ per degree Fahrenheit ($Btu_{IT}/°F$)	joule per kelvin (J/k)	1.899 101	E + 03
British thermal unit$_{IT}$ per degree Rankine ($Btu_{IT}/°R$)	joule per kelvin (J/k)	1.899 101	E + 03
British thermal unit$_{IT}$ per hour (Btu_{IT}/h)	watt (W)	2.930 711	E−01
British thermal unit$_{IT}$ per hour square foot degree Fahrenheit [$Btu_{IT}/(h \cdot ft^2 \cdot °F)$]	watt per square meter kelvin [W/(m² · K)]	5.678 263	E + 00
British thermal unit$_{IT}$ per pound (Btu_{IT}/lb)	joule per kilogram (J/kg)	2.326	E + 03
British thermal unit$_{IT}$ per pound degree Fahrenheit [$Btu_{IT}/(lb \cdot °F)$]	joule per kilogram kelvin [J/(kg · K)]	4.1868	E + 03
British thermal unit$_{IT}$ per pound degree Rankine [$Btu_{IT}/(lb \cdot °R)$]	joule per kilogram kelvin [J/(kg · K)]	4.1868	E + 03
British thermal unit$_{IT}$ per second (Btu_{IT}/s)	watt (W)	1.055 056	E + 03
British thermal unit$_{IT}$ per second square foot degree Fahrenheit [$Btu_{IT}/(s \cdot ft^2 \cdot °F)$]	watt per square meter kelvin [W/(m² · K)]	2.044 175	E + 04
British thermal unit$_{IT}$ per square foot (Btu_{IT}/ft^2)	joule per square meter (J/m^2)	1.135 653	E + 04
British thermal unit$_{IT}$ per square foot hour [$Btu_{IT}/(ft^2 \cdot h)$]	watt per square meter (W/m^2)	3.154 591	E + 00
British thermal unit$_{IT}$ per square foot second [$Btu_{IT}/(ft^2 \cdot s)$]	watt per square meter (W/m^2)	1.135 653	E + 04
bushel (U.S.) (bu)	cubic meter (m^3)	3.523 907	E−02
bushel (U.S.) (bu)	liter (L)	3.523 907	E + 01

Continues

Table continued

To convert from	To	Multiply by	
$calorie_{IT}$ (cal_{IT})	joule (J)	4.1868	E + 00
$calorie_{th}$ (cal_{th})	joule (J)	4.184	E + 00
calorie (cal) (mean)	joule (J)	4.190 02	E + 00
calorie (15°C) (cal_{15})	joule (J)	4.185 80	E + 00
calorie (20°C) (cal_{20})	joule (J)	4.181 90	E + 00
$calorie_{IT}$, kilogram (nutrition)	joule (J)	4.1868	E + 03
$calorie_{th}$/, kilogram (nutrition)	joule (J)	4.184	E + 03
calorie (mean), kilogram (nutrition)	joule (J)	4.190 02	E + 03
$calorie_{th}$ per centimeter second degree Celsius [$cal_{th}/(cm \cdot s \cdot °C)$]	watt per meter kelvin [W/(m · K)]	4.184	E + 02
$calorie_{IT}$ per gram (cal_{IT}/g)	joule per kilogram (J/kg)	4.1868	E + 03
$calorie_{th}$ per gram (cal_{th}/g)	joule per kilogram (J/kg)	4.184	E + 03
$calorie_{IT}$ per gram degree Celsius [$cal_{IT}/(g \cdot °C)$]	joule per kilogram kelvin [J/(kg · K)]	4.1868	E + 03
$calorie_{th}$ per gram degree Celsius [$cal_{th}/(g \cdot °C)$]	joule per kilogram kelvin [J/(kg · K)]	4.184	E + 03
$calorie_{IT}$ per gram kelvin [$cal_{IT}/(g \cdot K)$]	joule per kilogram kelvin [J/(kg · K)]	4.1868	E + 03
$calorie_{th}$ per gram kelvin [$cal_{th}/(g \cdot K)$]	joule per kilogram kelvin [J/(kg · K)]	4.184	E + 03
$calorie_{th}$ per minute (cal_{th}/min)	watt (W)	6.973 333	E−02
$calorie_{th}$ per second (cal_{th}/s)	watt (W)	4.184	E + 00
$calorie_{th}$ per square centimeter (cal_{th}/cm^2)	joule per square meter (J/m^2)	4.184	E + 04
$calorie_{th}$ per square centimeter minute [$cal_{th}/(cm^2 \cdot min)$]	watt per square meter (W/m^2)	6.973 333	E + 02
$calorie_{th}$ per square centimeter second [$cal_{th}/(cm^2 \cdot s)$]	watt per square meter (W/m^2)	4.184	E + 04
candela per square inch (cd/in^2)	candela per square meter (cd/m^2)	1.550 003	E + 03
carat, metric	kilogram (kg)	2.0	E−04
carat, metric	gram (g)	2.0	E−01
chain (based on U.S. survey foot) (ch)	meter (m)	2.011 684	E + 01
cord (128 ft^3)	cubic meter (m^3)	3.624 556	E + 00
cubic foot (ft^3)	cubic meter (m^3)	2.831 685	E−02
cubic foot per minute (ft^3/min)	cubic meter per second (m^3/s)	4.719 474	E−04
cubic foot per minute (ft^3/min)	liter per second (L/s)	4.719 474	E−01
cubic foot per second (ft^3/s)	cubic meter per second (m^3/s)	2.831 685	E−02
cubic inch (in^3)	cubic meter (m^3)	1.638 706	E−05
cubic inch per minute (in^3/min)	cubic meter per second (m^3/s)	2.731 177	E−07
cubic mile (mi^3)	cubic meter (m^3)	4.168 182	E + 09
cubic yard (yd^3)	cubic meter (m^3)	7.645 549	E−01
cubic yard per minute (yd^3/min)	cubic meter per second (m^3/s)	1.274 258	E−02
cup (U.S.)	cubic meter (m^3)	2.365 882	E−04
cup (U.S.)	liter (L)	2.365 882	E−01
cup (U.S.)	milliliter (mL)	2.365 882	E + 02
curie (Ci)	becquerel (Bq)	3.7	E + 10
darcy	meter squared (m^2)	9.869 233	E−13
day (d)	second (s)	8.64	E + 04
day (sidereal)	second (s)	8.616 409	E + 04
debye (D)	coulomb meter (C · m)	3.335 641	E−30
degree (angle) (°)	radian (rad)	1.745 329	E−02
degree Celsius (temperature) (°C)	kelvin (K)	T/K = t/°C + 273.15	
degree Celsius (temperature interval) (°C)	kelvin (K)	1.0	E + 00
degree centigrade (temperature) 16	degree Celsius (°C)	t/°C t/deg · cent	
degree centigrade (temperature interval) 16	degree Celsius (°C)	1.0	E + 00

Continues

Table continued

To convert from	To	Multiply by	
degree Fahrenheit (temperature) (°F)	degree Celsius (°C)	t/°C = (t/°F − 32)/1.8	
degree Fahrenheit (temperature) (°F)	kelvin (K)	T/K = (t/°F + 459.67)/1.8	
degree Fahrenheit (temperature interval) (°F)	degree Celsius (°C)	5.555 556	E−01
degree Fahrenheit (temperature interval) (°F)	kelvin (K)	5.555 556	E−01
degree Fahrenheit hour per British thermal unit$_{IT}$ (°F · h/Btu$_{IT}$)	kelvin per watt (K/W)	1.895 634	E+00
degree Fahrenheit hour per British thermal unit$_{th}$ (°F · h/Btu$_{th}$)	kelvin per watt (K/W)	1.896 903	E+00
degree Fahrenheit hour square foot per British thermal unit$_{IT}$ (°F · h · ft^2/Btu$_{IT}$)	square meter kelvin per watt (m^2 · K/W)	1.761 102	E−01
degree Fahrenheit hour square foot per British thermal unit$_{th}$ (°F · h · ft^2/Btu$_{th}$)	square meter kelvin per watt (m^2 · K/W)	1.762 280	E−01
degree Fahrenheit hour square foot per British thermal unit$_{IT}$ inch [°F · h · ft^2/(Btu$_{IT}$ · in)]	meter kelvin per watt (m · K/W)	6.933 472	E+00
degree Fahrenheit hour square foot per British thermal unit$_{th}$ inch [°F · h · ft^2/(Btu$_{th}$ · in)]	meter kelvin per watt (m · K/W)	6.938 112	E+00
degree Fahrenheit second per British thermal unit$_{IT}$ (°F · s/Btu$_{IT}$)	kelvin per watt (K/W)	5.265 651	E−04
degree Fahrenheit second per British thermal unit$_{th}$ (°F · s/Btu$_{th}$)	kelvin per watt (K/W)	5.269 175	E−04
degree Rankine (°R)	kelvin (K)	T/K = (T/°R)/1.8	
degree Rankine (temperature interval) (°R)	kelvin (K)	5.555 556	E−01
dyne (dyn)	newton (N)	1.0	E−05
dyne centimeter (dyn · cm)	newton meter (N · m)	1.0	E−07
dyne per square centimeter (dyn/cm^2)	pascal (Pa)	1.0	E−01
electronvolt (eV)	joule (J)	1.602 177	E−19
EMU of capacitance (abfarad)	farad (F)	1.0	E+09
EMU of current (abampere)	ampere (A)	1.0	E+01
EMU of electric potential (abvolt)	volt (V)	1.0	E−08
EMU of inductance (abhenry)	henry (H)	1.0	E−09
EMU of resistance (abohm)	ohm (Ω)	1.0	E−09
erg (erg)	joule (J)	1.0	E−07
erg per second (erg/s)	watt (W)	1.0	E−07
erg per square centimeter second [erg/(cm^2 · s)]	watt per square meter (W/m^2)	1.0	E−03
ESU of capacitance (statfarad)	farad (F)	1.112 650	E−12
ESU of current (statampere)	ampere (A)	3.335 641	E−10
ESU of electric potential (statvolt)	volt (V)	2.997 925	E+02
ESU of inductance (stathenry)	henry (H)	8.987 552	E+11
ESU of resistance (statohm)	ohm (Ω)	8.987 552	E+11
faraday (based on carbon 12)	coulomb (C)	9.648 531	E+04
fathom (based on U.S. survey foot)	meter (m)	1.828 804	E+00
fluid ounce (U.S.) (fl oz)	cubic meter (m^3)	2.957 353	E−05
fluid ounce (U.S.) (fl oz)	milliliter (mL)	2.957 353	E+01
foot (ft)	meter (m)	3.048	E−01
foot (U.S. survey) (ft)	meter (m)	3.048 006	E−01
footcandle	lux (lx)	1.076 391	E+01
footlambert	candela per square meter (cd/m^2)	3.426 259	E+00
foot per hour (ft/h)	meter per second (m/s)	8.466 667	E−05

Continues

Table continued

To convert from	To	Multiply by	
foot per minute (ft/min)	meter per second (m/s)	5.08	E−03
foot per second (ft/s)	meter per second (m/s)	3.048	E−01
foot per second squared (ft/s^2)	meter per second squared (m/s^2)	3.048	E−01
foot poundal	joule (J)	4.214 011	E−02
foot pound-force (ft · lbf)	joule (J)	1.355 818	E+00
foot pound-force per hour (ft · lbf/h)	watt (W)	3.766 161	E−04
foot pound-force per minute (ft · lbf/min)	watt (W)	2.259 697	E−02
foot pound-force per second (ft · lbf/s)	watt (W)	1.355 818	E+00
franklin (Fr)	coulomb (C)	3.335 641	E−10
gal (Gal)	meter per second squared (m/s^2)	1.0	E−02
gallon [Canadian and U.K. (Imperial)] (gal)	cubic meter (m^3)	4.546 09	E−03
gallon [Canadian and U.K. (Imperial)] (gal)	liter (L)	4.546 09	E+00
gallon (U.S.) (gal)	cubic meter (m^3)	3.785 412	E−03
gallon (U.S.) (gal)	liter (L)	3.785 412	E+00
gallon (U.S.) per day (gal/d)	cubic meter per second (m^3/s)	4.381 264	E−08
gallon (U.S.) per day (gal/d)	liter per second (L/s)	4.381 264	E−05
gallon (U.S.) per horsepower hour [gal/(hp · h)]	cubic meter per joule (m^3/J)	1.410 089	E−09
gallon (U.S.) per horsepower hour [gal/(hp · h)]	liter per joule (L/J)	1.410 089	E−06
gallon (U.S.) per minute (gpm) (gal/min)	cubic meter per second (m^3/s)	6.309 020	E−05
gallon (U.S.) per minute (gpm) (gal/min)	liter per second (L/s)	6.309 020	E−02
gamma (γ)	tesla (T)	1.0	E−09
gauss (Gs, G)	tesla (T)	1.0	E−04
gilbert (Gi)	ampere (A)	7.957 747	E−01
grain (gr)	kilogram (kg)	6.479 891	E−05
grain (gr)	milligram (mg)	6.479 891	E+01
grain per gallon (U.S.) (gr/gal)	kilogram per cubic meter (kg/m^3)	1.711 806	E−02
grain per gallon (U.S.) (gr/gal)	milligram per liter (mg/L)	1.711 806	E+01
gram per cubic centimeter (g/cm^3)	kilogram per cubic meter (kg/m^3)	1.0	E+03
hectare (ha)	square meter (m^2)	1.0	E+04
horsepower (550 ft · lbf/s) (hp)	watt (W)	7.456 999	E+02
horsepower (boiler)	watt (W)	9.809 50	E+03
horsepower (electric)	watt (W)	7.46	E+02
horsepower (metric)	watt (W)	7.354 988	E+02
horsepower (UK)	watt (W)	7.4570	E+02
horsepower (water)	watt (W)	7.460 43	E+02
hour (h)	second (s)	3.6	E+03
hour (sidereal)	second (s)	3.590 170	E+03
inch (in)	meter (m)	2.54	E−02
inch (in)	centimeter (cm)	2.54	E+00
inch per second (in/s)	meter per second (m/s)	2.54	E−02
inch per second squared (in/s^2)	meter per second squared (m/s^2)	2.54	E−02
kelvin (K)	degree Celsius (°C)	t/°C = T/K − 273.15	
kilocalorie$_{IT}$ (kcal$_{IT}$)	joule (J)	4.1868	E+03
kilocalorie$_{th}$ (kcal$_{th}$)	joule (J)	4.184	E+03
kilocalorie (mean) (kcal)	joule (J)	4.190 02	E+03

Continues

Table continued

To convert from	To	Multiply by	
kilocalorie$_{th}$ per minute (kcal$_{th}$/min)	watt (W)	6.973 333	E + 01
kilocalorie$_{th}$ per second (kcal$_{th}$/s)	watt (W)	4.184	E + 03
kilogram-force (kgf)	newton (N)	9.806 65	E + 00
kilogram-force meter (kgf · m)	newton meter (N · m)	9.806 65	E + 00
kilogram-force per square centimeter (kgf/cm^2)	pascal (Pa)	9.806 65	E + 04
kilogram-force per square centimeter (kgf/cm^2)	kilopascal (kPa)	9.806 65	E + 01
kilogram-force per square meter (kgf/m^2)	pascal (Pa)	9.806 65	E + 00
kilogram-force per square millimeter (kgf/mm^2)	pascal (Pa)	9.806 65	E + 06
kilogram-force per square millimeter (kgf/mm^2)	megapascal (MPa)	9.806 65	E + 00
kilogram-force second squared per meter (kgf · s^2/m)	kilogram (kg)	9.806 65	E + 00
kilometer per hour (km/h)	meter per second (m/s)	2.777 778	E−01
kilopond (kilogram-force) (kp)	newton (N)	9.806 65	E + 00
kilowatt hour (kW · h)	joule (J)	3.6	E + 06
kilowatt hour (kW · h)	megajoule (MJ)	3.6	E + 00
knot (nautical mile per hour)	meter per second (m/s)	5.144 444	E−01
lambert	candela per square meter (cd/m^2)	3.183 099	E + 03
langley (cal$_{th}$/cm^2)	joule per square meter (J/m^2)	4.184	E + 04
light year (l.y.)	meter (m)	9.460 73	E + 15
liter (L)	cubic meter (m^3)	1.0	E−03
lumen per square foot (lm/ft^2)	lux (lx)	1.076 391	E + 01
maxwell (Mx)	weber (Wb)	1.0	E−08
mho	siemens (S)	1.0	E + 00
micron (μ)	meter (m)	1.0	E−06
micron (μ)	micrometer (μm)	1.0	E + 00
mile (mi)	meter (m)	1.609 344	E + 03
mile (mi)	kilometer (km)	1.609 344	E + 00
mile, nautical	meter (m)	1.852	E + 03
mile per gallon (U.S.) (mpg) (mi/gal)	meter per cubic meter (m/m^3)	4.251 437	E + 05
mile per gallon (U.S.) (mpg) (mi/gal)	kilometer per liter (km/L)	4.251 437	E−01
mile per gallon (U.S.) (mpg) (mi/gal)	liter per 100 kilometer (L/100 km)	divide 235.215 by number of miles per gallon	
mile per hour (mi/h)	meter per second (m/s)	4.4704	E−01
mile per hour (mi/h)	kilometer per hour (km/h)	1.609 344	E + 00
mile per minute (mi/min)	meter per second (m/s)	2.682 24	E + 01
mile per second (mi/s)	meter per second (m/s)	1.609 344	E + 03
minute (min)	second (s)	6.0	E + 01
minute (sidereal)	second (s)	5.983 617	E + 01
oersted (Oe)	ampere per meter (A/m)	7.957 747	E + 01
ohm centimeter (Ω · cm)	ohm meter (Ω · m)	1.0	E−02
ohm circular-mil per foot	ohm meter (Ω · m)	1.662 426	E−09
ohm circular-mil per foot	ohm square millimeter per meter (Ω · mm^2/m)	1.662 426	E−03
ounce (avoirdupois) (oz)	kilogram (kg)	2.834 952	E−02
ounce (avoirdupois) (oz)	gram (g)	2.834 952	E + 01
ounce (troy or apothecary) (oz)	kilogram (kg)	3.110 348	E−02

Continues

Table continued

To convert from	To	Multiply by	
ounce (troy or apothecary) (oz)	gram (g)	3.110 348	E + 01
ounce [Canadian and UK fluid (Imperial)] (fl oz)	cubic meter (m^3)	2.841 306	E−05
ounce [Canadian and UK fluid (Imperial)] (fl oz)	milliliter (mL)	2.841 306	E + 01
ounce (U.S. fluid) (fl oz)	cubic meter (m^3)	2.957 353	E−05
ounce (U.S. fluid) (fl oz)	milliliter (mL)	2.957 353	E + 01
ounce (avoirdupois) per cubic inch (oz/in^3)	kilogram per cubic meter (kg/m^3)	1.729 994	E + 03
ounce (avoirdupois) per gallon [Canadian and UK (Imperial)] (oz/gal)	kilogram per cubic meter (kg/m^3)	6.236 023	E + 00
ounce (avoirdupois) per gallon [Canadian and UK (Imperial)] (oz/gal)	gram per liter (g/L)	6.236 023	E + 00
ounce (avoirdupois) per gallon (U.S.) (oz/gal)	kilogram per cubic meter (kg/m^3)	7.489 152	E + 00
ounce (avoirdupois) per gallon (U.S.) (oz/gal)	gram per liter (g/L)	7.489 152	E + 00
ounce (avoirdupois) per square foot (oz/ft^2)	kilogram per square meter (kg/m^2)	3.051 517	E−01
ounce (avoirdupois) per square inch (oz/in^2)	kilogram per square meter (kg/m^2)	4.394 185	E + 01
ounce (avoirdupois) per square yard (oz/yd^2)	kilogram per square meter (kg/m^2)	3.390 575	E−02
parsec (pc)	meter (m)	3.085 678	E + 16
peck (U.S.) (pk)	cubic meter (m^3)	8.809 768	E−03
peck (U.S.) (pk)	liter (L)	8.809 768	E + 00
pint (U.S. dry) (dry pt)	cubic meter (m^3)	5.506 105	E−04
pint (U.S. dry) (dry pt)	liter (L)	5.506 105	E−01
pint (U.S. liquid) (liq pt)	cubic meter (m^3)	4.731 765	E−04
pint (U.S. liquid) (liq pt)	liter (L)	4.731 765	E−01
pound (avoirdupois) (lb)	kilogram (kg)	4.535 924	E−01
pound (troy or apothecary) (lb)	kilogram (kg)	3.732 417	E−01
poundal	newton (N)	1.382 550	E−01
poundal per square foot	pascal (Pa)	1.488 164	E + 00
poundal second per square foot	pascal second (Pa · s)	1.488 164	E + 00
pound foot squared (lb · ft^2)	kilogram meter squared (kg · m^2)	4.214 011	E−02
pound-force (lbf)	newton (N)	4.448 222	E + 00
pound-force foot (lbf · ft)	newton meter (N · m)	1.355 818	E + 00
pound-force foot per inch (lbf · ft/in)	newton meter per meter (N · m/m)	5.337 866	E + 01
pound-force inch (lbf · in)	newton meter (N · m)	1.129 848	E−01
pound-force inch per inch (lbf · in/in)	newton meter per meter (N · m/m)	4.448 222	E + 00
pound-force per foot (lbf/ft)	newton per meter (N/m)	1.459 390	E + 01
pound-force per inch (lbf/in)	newton per meter (N/m)	1.751 268	E + 02
pound-force per pound (lbf/lb) (thrust to mass ratio)	newton per kilogram (N/kg)	9.806 65	E + 00
pound-force per square foot (lbf/ft^2)	pascal (Pa)	4.788 026	E + 01
pound-force per square inch (psi) (lbf/in^2)	pascal (Pa)	6.894 757	E + 03
pound-force per square inch (psi) (lbf/in^2)	kilopascal (kPa)	6.894 757	E + 00
pound-force second per square foot (lbf · s/ft^2)	pascal second (Pa · s)	4.788 026	E + 01
pound-force second per square inch (lbf · s/in^2)	pascal second (Pa · s)	6.894 757	E + 03
pound inch squared (lb · in^2)	kilogram meter squared (kg · m^2)	2.926 397	E−04
pound per cubic foot (lb/ft^3)	kilogram per cubic meter (kg/m^3)	1.601 846	E + 01
pound per cubic inch (lb/in^3)	kilogram per cubic meter (kg/m^3)	2.767 990	E + 04
pound per cubic yard (lb/yd^3)	kilogram per cubic meter (kg/m^3)	5.932 764	E−01
pound per foot (lb/ft)	kilogram per meter (kg/m)	1.488 164	E + 00
pound per foot hour [lb/(ft · h)]	pascal second (Pa · s)	4.133 789	E−04
pound per foot second [lb/(ft · s)]	pascal second (Pa · s)	1.488 164	E + 00

Continues

Table continued

To convert from	To	Multiply by	
pound per gallon [Canadian and U.K. (Imperial)] (lb/gal)	kilogram per cubic meter (kg/m^3)	9.977 637	E + 01
pound per gallon [Canadian and U.K. (Imperial)] (lb/gal)	kilogram per liter (kg/L)	9.977 637	E−02
pound per gallon (U.S.) (lb/gal)	kilogram per cubic meter (kg/m^3)	1.198 264	E + 02
pound per gallon (U.S.) (lb/gal)	kilogram per liter (kg/L)	1.198 264	E−01
pound per horsepower hour [lb/(hp · h)]	kilogram per joule (kg/J)	1.689 659	E−07
pound per hour (lb/h)	kilogram per second (kg/s)	1.259 979	E−04
pound per inch (lb/in)	kilogram per meter (kg/m)	1.785 797	E + 01
pound per minute (lb/min)	kilogram per second (kg/s)	7.559 873	E−03
pound per second (lb/s)	kilogram per second (kg/s)	4.535 924	E−01
pound per square foot (lb/ft^2)	kilogram per square meter (kg/m^2)	4.882 428	E + 00
pound per square inch (not pound-force) (lb/in^2)	kilogram per square meter (kg/m^2)	7.030 696	E + 02
pound per yard (lb/yd)	kilogram per meter (kg/m)	4.960 546	E−01
psi (pound-force per square inch) (lbf/in^2)	pascal (Pa)	6.894 757	E + 03
psi (pound-force per square inch) (lbf/in^2)	kilopascal (kPa)	6.894 757	E + 00
quad (10^{15} Btu_{IT})	joule (J)	1.055 056	E + 18
quart (U.S. dry) (dry qt)	cubic meter (m^3)	1.101 221	E−03
quart (U.S. dry) (dry qt)	liter (L)	1.101 221	E + 00
quart (U.S. liquid) (liq qt)	cubic meter (m^3)	9.463 529	E−04
quart (U.S. liquid) (liq qt)	liter (L)	9.463 529	E−01
rad (absorbed dose) (rad)	gray (Gy)	1.0	E−02
rem (rem)	sievert (Sv)	1.0	E−02
roentgen (R)	coulomb per kilogram (C/kg)	2.58	E−04
square foot (ft^2)	square meter (m^2)	9.290 304	E−02
square foot per hour (ft^2/h)	square meter per second (m^2/s)	2.580 64	E−05
square foot per second (ft^2/s)	square meter per second (m^2/s)	9.290 304	E−02
square inch (in^2)	square meter (m^2)	6.4516	E−04
square inch (in^2)	square centimeter (cm^2)	6.4516	E + 00
square mile (mi^2)	square meter (m^2)	2.589 988	E + 06
square mile (mi^2)	square kilometer (km^2)	2.589 988	E + 00
square mile (based on U.S. survey foot) (mi^2)	square meter (m^2)	2.589 998	E + 06
square mile (based on U.S. survey foot) (mi^2)	square kilometer (km^2)	2.589 998	E + 00
square yard (yd^2)	square meter (m^2)	8.361 274	E−01
statampere	ampere (A)	3.335 641	E−10
statcoulomb	coulomb (C)	3.335 641	E−10
statfarad	farad (F)	1.112 650	E−12
stathenry	henry (H)	8.987 552	E + 11
statmho	siemens (S)	1.112 650	E−12
statohm	ohm (Ω)	8.987 552	E + 11
statvolt	volt (V)	2.997 925	E + 02
stere (st)	cubic meter (m^3)	1.0	E + 00
stilb (sb)	candela per square meter (cd/m^2)	1.0	E + 04
tablespoon	cubic meter (m^3)	1.478 676	E−05
tablespoon	milliliter (mL)	1.478 676	E + 01

Continues

Table continued

To convert from	To	Multiply by	
teaspoon	cubic meter (m^3)	4.928 922	E−06
teaspoon	milliliter (mL)	4.928 922	E+00
therm (EU)	joule (J)	1.055 06	E+08
therm (U.S.)	joule (J)	1.054 804	E+08
ton, assay (AT)	kilogram (kg)	2.916 667	E−02
ton, assay (AT)	gram (g)	2.916 667	E+01
ton-force (2000 lbf)	newton (N)	8.896 443	E+03
ton-force (2000 lbf)	kilonewton (kN)	8.896 443	E+00
ton, long (2240 lb)	kilogram (kg)	1.016 047	E+03
ton, long, per cubic yard	kilogram per cubic meter (kg/m^3)	1.328 939	E+03
ton, metric (t)	kilogram (kg)	1.0	E+03
tonne (called "metric ton" in U.S.) (t)	kilogram (kg)	1.0	E+03
ton of refrigeration (12 000 Btu_{IT}/h)	watt (W)	3.516 853	E+03
ton of TNT (energy equivalent)	joule (J)	4.184	E+09
ton, register	cubic meter (m^3)	2.831 685	E+00
ton, short (2000 lb)	kilogram (kg)	9.071 847	E+02
ton, short, per cubic yard	kilogram per cubic meter (kg/m^3)	1.186 553	E+03
ton, short, per hour	kilogram per second (kg/s)	2.519 958	E−01
watt hour (W · h)	joule (J)	3.6	E+03
watt per square centimeter (W/cm^2)	watt per square meter (W/m^2)	1.0	E+04
watt per square inch (W/in^2)	watt per square meter (W/m^2)	1.550 003	E+03
watt second (W · s)	joule (J)	1.0	E+00
yard (yd)	meter (m)	9.144	E−01
year (365 days)	second (s)	3.1536	E+07
year (sidereal)	second (s)	3.155 815	E+07
year (tropical)	second (s)	3.155 693	E+07

5. ENERGY CONVERSION EFFICIENCIES

Efficiencies of Common Anthropogenic and Natural Energy Conversions[a]

Conversion	Energies[b]	Efficiencies
Large electricity generators	$m \rightarrow e$	98–99
Large power plant boilers	$c \rightarrow t$	90–98
Large electric motors	$e \rightarrow m$	90–97
Best home natural gas furnaces	$c \rightarrow t$	90–96
Dry-cell batteries	$c \rightarrow e$	85–95
Home gas furnace	$c \rightarrow t$	85
Human lactation	$c \rightarrow c$	75–85
Overshot waterwheels	$m \rightarrow m$	60–85
Small electric motors	$e \rightarrow m$	60–75

Continues

Table continued

Conversion	Energies[b]	Efficiencies
Best bacterial growth	$c \rightarrow c$	50–65
Glycolysis maxima	$c \rightarrow c$	50–60
Home oil furnace	$c \rightarrow t$	60
Fluidized bed coal combustion	$c \rightarrow e$	40–50
Integrated gasification combined cycle	$c \rightarrow e$	40–50
Large steam turbines	$t \rightarrow m$	40–45
Improved wood stoves	$c \rightarrow t$	25–45
Large gas turbines	$c \rightarrow m$	35–40
Fossil fuel electric generating plant	$c \rightarrow e$	30–40
Charcoal production—earth kiln	$c \rightarrow c$	20–40
Modern wind turbine	$m \rightarrow e$	40
Diesel engines	$c \rightarrow m$	30–35
Mammalian postnatal growth	$c \rightarrow c$	30–35
Best photovoltaic cells	$r \rightarrow e$	20–30
Best large steam engines	$c \rightarrow m$	20–25
Internal combustion engines	$c \rightarrow m$	15–25
High-pressure sodium lamps	$e \rightarrow r$	15–20
Mammalian muscles	$c \rightarrow m$	15–20
Milk production	$c \rightarrow c$	15–20
Pregnancy	$c \rightarrow c$	10–20
Fluorescent lights	$e \rightarrow r$	10–20
Broiler production	$c \rightarrow c$	10–15
Traditional stoves	$c \rightarrow t$	10–15
Beef production	$c \rightarrow c$	5–10
Steam locomotives	$c \rightarrow m$	3–6
Peak field photosynthesis	$r \rightarrow c$	4–5
Incandescent light bulbs	$e \rightarrow r$	2–5
Paraffin candles	$c \rightarrow r$	1–2
Most productive ecosystems	$r \rightarrow c$	1–2
Global photosynthetic mean	$r \rightarrow c$	0.3

[a]All ranges are the first-law efficiencies in percent.
[b]Energy labels: c = chemical, e = electrical, m = mechanical (kinetic), r = radiant (electromagnetic, solar), t = thermal.

6. MASS-ENERGY EQUIVALENTS

Mass-Energy Equivalence

1 electron mass	1 atomic mass unit (amu)	1 gram (g)	1 kilogram (kg)
0.511 MeV	931.5 MeV	5.61×10^{26} MeV	5.61×10^{29} MeV
8.187×10^{-14} J	1.492×10^{-10} J	8.988×10^{13} J	8.988×10^{16} J
7.765×10^{-17} BTU	1.415×10^{-13} BTU	8.524×10^{10} BTU	8.524×10^{13} BTU
2.274×10^{-20} kWh	4.146×10^{-17} kWh	2.497×10^{7} kWh	2.497×10^{10} kWh
7.765×10^{-35} Q	1.415×10^{-31} Q	8.524×10^{-8} Q	8.524×10^{-5} Q

Continues

Table continued

1 ton	1 MeV	1 joule (J)	1 BTU
5.089×10^{32} MeV	1.957 electron masses	1.221×10^{13} electron masses	1.288×10^{16} electron masses
8.153×10^{19} J	1.074×10^{-3} amu	6.701×10^{9} amu	7.065×10^{12} amu
7.733×10^{16} BTU	1.783×10^{-27} g	1.113×10^{-14} g	1.173×10^{-11} g
2.265×10^{13} kWh	1.783×10^{-30} kg	1.113×10^{-17} kg	1.173×10^{-14} kg
7.733×10^{-2} Q	1.965×10^{-33} ton	1.226×10^{-20} ton	1.293×10^{-17} ton

	1 kilowatt-hour (kWh)	1 Q
	4.397×10^{19} electron masses	1.288×10^{34} electron masses
	2.412×10^{16} amu	7.065×10^{30} amu
	4.006×10^{-8} g	1.173×10^{7} g
	4.006×10^{-11} kg	1.173×10^{4} kg
	4.415×10^{-14} ton	12.93 tons

Note. This table can be used to calculate the energy released in a process in which the mass decreases by a known amount. This table cannot be used directly to calculate, for example, the energy available from a given quantity of U^{235}, because in the fission of 1 kg of U^{235}, the decrease in mass is much less than 1 kg.

7. CONVERSION SCHEME FOR COMMON ENERGY DEVICES

Device	Energy input	Useful output
Electric heater	Electricity	Thermal energy
Hair dryer	Electricity	Thermal energy
Electric generator	Mechanical energy	Electricity
Electric motor	Electricity	Mechanical energy
Battery	Chemical energy	Electricity
Steam boiler	Chemical energy	Thermal energy
Furnace	Chemical energy	Thermal energy
Steam turbine	Thermal energy	Mechanical energy
Gas turbine	Chemical energy	Mechanical energy
Automobile engine	Chemical energy	Mechanical energy
Fluorescent lamp	Electricity	Light
Silicon solar cell	Solar energy	Electricity
Steam locomotive	Chemical	Mechanical
Incandescent lamp	Electricity	Light

8. ENERGY FLOWS AND STORAGES

Energy flow or storage[a]	Actual multiple (J)	Order of magnitude
Solar radiation intercepted by the Earth[b]	5.5	24
Global coal resources	2	23

Continues

Table continued

Energy flow or storage[a]	Actual multiple (J)	Order of magnitude
Global plant mass	2	22
Global net photosynthesis[b]	2	21
Global fossil fuel production[b]	3	20
Typical Caribbean hurricane	3.8	19
Global lightnings[b]	3.2	18
Largest H-bomb tested in 1961	2.4	17
Global zodiacal light[b]	6.3	16
Latent heat of a thunderstorm	5	15
Kinetic energy of a thunderstorm	1	14
Hiroshima bomb (1945)	8.4	13
Coal load in a 100 t hopper car	2.5	12
Good grain corn harvest (8 t/ha)	1.2	11
Gasoline for a compact car[b]	4	10
Barrel of crude oil	6.5	9
Basal metabolism of a large horse	1	8
Daily adult food intake	1	7
Bottle of white table wine	2.6	6
Large hen egg	4	5
Vole's daily basal metabolism	5	4
Small chickpea	5	3
Baseball (140 g) pitched at 40 m/s	1.1	2
Tennis ball (50 g) served at 25 m/s	1.5	1
Full teacup (300 g) held in hand	2.6	0
Falling 2 cm hailstone	2	−1
Striking a typewriter key	2	−2
Fly on a kitchen table	9	−3
Small bird's 5 second song	5	−4
A 2 mm raindrop falling at 6 m/s	7.5	−5
The same drop on a blade of grass	4	−6
Flea hop	1	−7

Source. Smil, V. (1991). "General Energetics." Wiley.

[a]Calculated from data in Chapters 2–10 and from measurements of common activities and objects.

[b]Annual total.

9. FUEL PROPERTIES

Substance	Chemical formula	Molecular weight	Density[a] (kg/m^3)	Higher heating value (kJ/m^3)[b]	Higher heating value (kJ/kg)[b]	Lower heating value (kJ/m^3)[b]	Lower heating value (kJ/kg)[b]
Fuels							
Hydrogen	H_2	2.016	0.0838	11,910	142,097	10,060	120,067
Carbon	C	12.011	…	…	32,778	…	32,778
Sulfur	S	32.064	…	…	9,257	…	9,257
Hydrogen sulfide	H_2S	34.080	1.4168	23,390	16,506	21,540	15,204

Continues

Table continued

Substance	Chemical formula	Molecular weight	Density[a] (kg/m^3)	Higher heating value (kJ/m^3)[b]	Higher heating value (kJ/kg)[b]	Lower heating value (kJ/m^3)[b]	Lower heating value (kJ/kg)[b]
Carbon monoxide	CO	28.006	1.1643	11,770	10,110	11,770	10,110
Methane	CH_4	16.043	0.6669	37,030	55,529	33,340	49,994
Methyl alcohol	CH_3OH	32.043	1.3321	31,780	23,858	28,090	21,086
Ethane	C_2H_6	30.071	1.2501	64,910	51,920	59,370	47,489
Ethylene	C_2H_4	28.055	1.1663	58,690	50,322	55,000	47,156
Acetylene	C_2H_2	26.039	1.0825	54,140	50,010	52.290	48,305
Ethyl alcohol	C_2H_5OH	46.071	1.9153	58,630	30,610	53,090	27,717
Propane	C_3H_8	44.099	1.8333	92,390	50,399	85,010	46,370
Propylene	C_3H_6	42.083	1.7495	85,640	48,954	80,110	45,789
n-Butane	C_4H_{10}	58.126	2.4164	119,820	49,589	110,590	45,768
Isobutane	C_4H_{10}	58.126	2.4164	119,540	49,472	110,310	45,652
n-Butene	C_4H_8	56.110	2.3326	113,130	48,503	105,750	45,338
Isobutene	C_4H_8	56.110	2.3326	112,500	48,231	105,110	45,065
n-Pentane	C_5H_{12}	72.153	2.9996	147,170	49,064	136,090	45,370
Isopentane	C_5H_{12}	72.153	2.9996	146,830	48,952	135,750	45,258
Neopentane	C_5H_{12}	72.153	2.9996	146,350	48,791	133,270	45,098
n-Pentene	C_5H_{10}	70.137	2.9157	140,510	48,191	131,280	45,026
n-Hexane	C_6H_{14}	86.181	3.5827	174,710	48,764	161,780	45,156
Benzene	C_6H_6	78.117	3.2475	137,350	42,293	131,810	40,588
Toluene	C_7H_8	92.141	3.8305	164,830	43,030	157,440	41,102
Xylene	C_8H_{10}	106.172	4.4138	191,460	43,377	182,320	41,307
Naphthalene	$C_{10}H_8$	128.179	5.3287	214,450	40,244	207,070	38,860
Ammonia	NH_3	17.031	0.7080	15,920	22,484	13,150	18,572
Nonfuels							
Oxygen	O_2	31.999	1.3303				
Nitrogen	N_2	28.013	1.1646				
Air	…	28.970	1.2043				
Carbon dioxide	CO_2	44.010	1.8296				
Sulfur dioxide	SO_2	64.063	2.6632				

[a]All gas values corrected to one atmosphere and 20°C (68°F), $\rho = 0.04157$ (M.W.).
[b]1 kJ/m^3 = 0.02684 Btu/ft^3; 1 kJ/kg = 0.43 Btu/lbm; 1 kg/m^3 = 0.0624 lbm/ft^3.

10. CARBON CONTENT OF FUELS

Carbon Content Coefficients and Underlying Data for Petroleum Products

Fuel	2001 carbon content (MMTC/QBtu)	Gross heat of combustion (MMBtu/barrel)	Density (API gravity)	Percent carbon
Motor gasoline	19.34	5.253	59.6	86.60
LPG	16.99	[a]		[a]
Jet fuel	19.33	5.670	42.0	86.30
Distillate fuel	19.95	5.825	35.5	86.34

Continues

Table continued

Fuel	2001 carbon content (MMTC/QBtu)	Gross heat of combustion (MMBtu/barrel)	Density (API gravity)	Percent carbon
Residual fuel	21.49	6.287	11.0	85.68
Asphalt and road oil	20.62	6.636	5.6	83.47
Lubricants	20.24	6.065	25.6	85.80
Petrochemical	19.37	5.248[b]	67.1[b]	84.11[b]
Feedstocks				
Aviation gas	18.87	5.048	69.0	85.00
Kerosene	19.72	5.670	41.4	86.01
Petroleum coke	27.85	6.024	—	92.28
Special naphtha	19.86	5.248	51.2	84.76
Petroleum waxes	19.81	5.537	43.3	85.29
Still gas	17.51	6.000	—	—
Crude oil	20.29	5.800	30.5	85.49
Unfinished oils	20.29	5.825	30.5	85.49
Miscellaneous products	20.29	5.796	30.5	85.49
Pentanes plus	18.24	4.620	81.7	83.70
Natural gasoline	18.24	4.620	81.7	83.70

Source. Energy Information Administration, "Emissions of Greenhouse Gases in the United States 1987–1992," Washington, DC (November, 1994), DOE/EIA 0573.

[a]LPG is a blend of multiple paraffinic hydrocarbons: ethane, propane, isobutane, and normal butane, each with their own heat content, density, and carbon content.

[b]Parameters presented are for naphthas with a boiling temperature less than 400 degrees Fahrenheit. Petrochemical feedstocks with higher boiling points are assumed to have the same characteristics as distillate fuel.

— no sample data available.

11. PROPERTIES OF GASEOUS FUELS

Properties of Typical Gaseous Fuels[2]

	Calorific value			
	Btu/ft^3		kcal/m^3	
Type of gas	Gross	Net	Gross	Net
Acetylene, commercial	1410	1360	12,548	12,105
Blast furnace	92	91	819	819
Blue (water), bituminous	260	239	2314	2127
Butane, commercial, natural gas	3210	2961	28,566	26,350
Butane, commercial, refinery gas	3184	2935	28,334	26,119
Carbureted blue, low gravity	536	461	4770	4102
Carbureted blue, heavy oil	530	451	4716	4013
Coke oven, by-product	569	509	5064	4530
Mapp	2406	2282	21,411	20,308
Natural, Alaska	998	906	8879	8063
Natural, Algerian LNG, Canvey	1122	1014	9985	9024
Natural, Gaz de Lacq	1011	911	8997	8107
Natural, Groningen, Netherlands	875	789	7787	7021
Natural, Libyan LNG	1345	1223	11,969	10,883

Continues

Table continued

Type of gas	Calorific value			
	Btu/ft^3		kcal/m^3	
	Gross	Net	Gross	Net
Natural, North Sea, Bacton	1023	922	9104	8205
Natural, Birmingham, AL	1002	904	8917	8045
Natural, Cleveland, OH	1059	959	9424	8534
Natural, Kansas City, MO	974	879	8668	7822
Natural, Pittsburgh, PA	1129	1021	10,047	9086
Producer, Koppers–Totzek[a]	288	271	2563	2412
Producer, Lurgi[b]	183	167	1629	1486
Producer, W-G, bituminous[b]	168	158	1495	1406
Producer, Winkler[b]	117	111	1041	988
Propane, commercial, natural gas	2558	2358	22,764	20,984
Propane, commercial, refinery gas	2504	2316	22,283	20,610
Sasol, South Africa	500	448	4450	3986
Sewage, Decatur	690	621	6140	5526
SNG, no methanation	853	765	7591	6808

Source. M. Kutz, "Mechanical Engineers' Handbook." Wiley-Interscience.

[a] O_2-blown.

[b] Air-blown.

12. APPROXIMATE HEAT CONTENT OF PETROLEUM PRODUCTS

Energy source	Heat content (million Btu/barrel)
Asphalt	6.636
Aviation gasoline	5.048
Butane	4.326
Butane–propane mixture (60%–40%)	4.130
Distillate fuel oil	5.825
Ethane	3.082
Ethane–propane mixture (70%–30%)	3.308
Isobutane	3.974
Jet fuel, kerosene-type	5.670
Jet fuel, naphtha-type	5.355
Kerosene	5.670
Lubricants	6.065
Motor gasoline	
Conventional	5.253
Oxygenated	5.150
Reformulated	5.150
Fuel ethanol[a]	3.539
Natural gasoline	4.620
Pentanes plus	4.620

Continues

Table continued

Energy source	Heat content (million Btu/barrel)
Petrochemical feedstocks	
Naphtha less than 401°F	5.248
Other oils equal to or greater than 401°F	5.825
Still gas	6.000
Petroleum coke	6.024
Plant condensate	5.418
Propane	3.836
Residual fuel oil	6.287
Road oil	6.636
Special naphthas	5.248
Still gas	6.000
Unfinished oils	5.825
Unfractionated stream	5.418
Waxes	5.537
Miscellaneous	5.796

Source. U.S. Department of Energy.

[a]Fuel ethanol, which is derived from agricultural feedstocks (primarily corn), is not a petroleum product but is blended into motor gasoline. Its gross heat content (3.539 million Btu per barrel) is used in *Annual Energy Review* calculations; its net heat content (3.192 million Btu per barrel) is used in the Energy Information Administration's *Renewable Energy Annual* calculations.

13. TYPICAL ENERGY CONTENT OF FOSSIL AND BIOMASS FUELS

Solid fuels	Net heating values (MJ/kg)
Biomass fuels	
Wood (wet, freshly cut)	10.9
Wood (air dry, humid zone)	15.5
Wood (air dry, dry zone)	16.6
Wood (oven dry)	20.0
Charcoal	29.0
Bagasse (wet)	8.2
Bagasse (air dry)	16.2
Coffee husks	16.0
Rice hulls (air dry)	14.4
Wheat straw	15.2
Corn (stalk)	14.7
Corn (cobs)	15.4
Cotton stalk	16.4
Coconut husks	9.8
Coconut shells	17.9
Fossil fuels	
Anthracite coal	31.4
Bituminous coal	29.3
Lignite	11.3
Coke	28.5

Source. U.S. Department of Energy.

14. WORLD COAL RESERVES

World Estimated Recoverable Coal (Million Short Tons)

Region Country	Recoverable anthracite and bituminous[a]	Recoverable lignite and subbituminous[a]	Total recoverable coal[a]
North America			
Canada	3826	3425	7251
Greenland	0	202	202
Mexico	948	387	1335
United States[b]	126,804	146,852	273,656
Total	131,579	150,866	282,444
Central and South America			
Argentina	0	474	474
Bolivia	1	0	1
Brazil	0	13,149	13,149
Chile	34	1268	1302
Colombia	6908	420	7328
Ecuador	0	26	26
Peru	1058	110	1168
Venezuela	528	0	528
Total	8530	15,448	23,977
Western Europe			
Austria	0	28	28
Croatia	7	36	43
France	24	15	40
Germany	25,353	47,399	72,753
Greece	0	3168	3168
Ireland	15	0	15
Italy	0	37	37
Netherlands	548	0	548
Norway	0	1	1
Portugal	3	36	40
Slovenia	0	303	303
Spain	220	507	728
Sweden	0	1	1
Turkey	306	3760	4066
United Kingdom	1102	551	1653
Yugoslavia	71	17,849	17,919
Total	27,650	73,693	101,343
Eastern Europe and Former U.S.S.R.			
Bulgaria	14	2974	2988
Czech Republic	2330	3929	6259
Hungary	0	1209	1209
Kazakhstan	34,172	3307	37,479
Kyrgyzstan	0	895	895

Continues

Table continued

Region Country	Recoverable anthracite and bituminous[a]	Recoverable lignite and subbituminous[a]	Total recoverable coal[a]
Poland	22,377	2050	24,427
Romania	1	1605	1606
Russia	54,110	118,964	173,074
Slovakia	0	190	190
Ukraine	17,939	19,708	37,647
Uzbekistan	1102	3307	4409
Total	132,046	158,138	290,183
Middle East			
Iran	1885	0	1885
Total	1885	0	1885
Africa			
Algeria	44	0	44
Botswana	4740	0	4740
Central African Republic	0	3	3
Congo (Kinshasa)	97	0	97
Egypt	0	24	24
Malawi	0	2	2
Mozambique	234	0	234
Niger	77	0	77
Nigeria	23	186	209
South Africa	54,586	0	54,586
Swaziland	229	0	229
Tanzania	220	0	220
Zambia	11	0	11
Zimbabwe	553	0	553
Total	60,816	216	61,032
Asia and Oceania			
Afghanistan	73	0	73
Australia	46,903	43,585	90,489
Burma	2	0	2
China	68,564	57,651	126,215
India	90,826	2205	93,031
Indonesia	871	5049	5919
Japan	852	0	852
Korea, North	331	331	661
Korea, South	86	0	86
Malaysia	4	0	4
Nepal	2	0	2
New Caledonia	2	0	2
New Zealand	36	594	631
Pakistan	0	2497	2497
Philippines	0	366	366
Taiwan	1	0	1

Continues

Table continued

Region Country	Recoverable anthracite and bituminous[a]	Recoverable lignite and subbituminous[a]	Total recoverable coal[a]
Thailand	0	1398	1398
Vietnam	165	0	165
Total	208,719	113,675	322,394
World Total	571,224	512,035	1,083,259

Sources. World Energy Council, Survey of Energy Resources 2001, October 2001. United States: Energy Information Administration. Unpublished file data of the Coal Reserves Data Base (February 2002).

Notes. Formerly entitled "World Estimated Recoverable Reserves of Coal." The estimates in this table are dependent on the judgment of each reporting country to interpret local economic conditions and its own mineral assessment criteria in terms of specified standards of the World Energy Council. Consequently, the data may not all meet the same standards of reliability and some data, including the Energy Information Administration's (EIA)'s, may not represent reserves of coal that are known to be recoverable under current economic conditions and regulations. Some data, including the EIA's, represent estimated recovery rates for highly reliable estimates of coal quantities in the ground that have physical characteristics like those of coals currently being profitably mined. U.S. coal rank approximations are based partly on Btu and may not match precisely borderline geologic ranks. Further, data in this table may represent different base years. Data for the U.S. represent recoverable coal estimates as of December 31, 2000. Data for countries are as of December 31, 1999, the most recent period for which they are available. The Energy Information Administration does not certify the International reserves data but reproduces the information as a matter of convenience for the reader. Sum of components may not equal total due to independent rounding.

[a]World Energy Council definition of "Proved Recoverable Reserves" the tonnage within the Proved Amount in Place that can be recovered (extracted from the earth in raw form) under present and expected local economic conditions with existing available technology.

[b]Data represent both measured and indicated tonnage, as of January 1, 2001 (equated to December 31, 2000). The U.S. term "measured" approximates the term "proved" used by the World Energy Council. The U.S. "measured and indicated" data have been combined prior to depletion adjustments and cannot be recaptured as "measured alone."

15. WORLD CRUDE OIL AND NATURAL GAS RESERVES (JANUARY 1, 2002)

	Crude oil (billion barrels)		Natural gas (trillion cubic feet)	
Region Country	Oil and gas journal	World oil	Oil and gas journal	World oil
North America				
Canada	4.9	5.4	59.7	59.7
Mexico	26.9	23.1	29.5	39.0
United States[a]	22.4	22.4	183.5	183.5
Total	54.2	50.9	272.7	282.1
Central and South America				
Argentina	3.0	2.9	27.5	26.8
Barbados	(s)	NA	(s)	NA
Bolivia	0.4	0.5	24.0	27.4
Brazil	8.5	8.6	7.8	7.9
Chile	0.2	(s)	3.5	1.3
Colombia	1.8	1.9	4.3	5.0
Cuba	0.8	0.3	2.5	0.5

Continues

Table continued

Region / Country	Crude oil (billion barrels)		Natural gas (trillion cubic feet)	
	Oil and gas journal	World oil	Oil and gas journal	World oil
Ecuador	2.1	2.6	3.7	3.9
Guatemala	0.5	NA	0.1	NA
Peru	0.3	0.9	8.7	8.7
Suriname	0.1	NA	0.0	NA
Trinidad and Tobago	0.7	0.7	23.5	19.7
Venezuela	77.7	50.2	147.6	149.2
Other	0.0	0.6	0.0	(s)
Total	96.0	69.1	253.0	250.2
Western Europe				
Austria	0.1	0.1	0.9	0.8
Croatia	0.1	0.1	1.2	1.2
Denmark	1.1	1.3	2.7	3.1
France	0.1	0.1	0.4	0.5
Germany	0.4	0.3	12.1	9.0
Greece	(s)	NA	(s)	NA
Ireland	0.0	NA	0.7	NA
Italy	0.6	0.6	8.1	6.7
Netherlands	0.1	0.1	62.5	57.0
Norway	9.4	10.3	44.0	77.2
Spain	(s)	NA	(s)	NA
Turkey	0.3	0.3	0.3	0.3
United Kingdom	4.9	4.6	26.0	24.5
Yugoslavia	0.1	NA	1.7	NA
Other	0.0	0.1	0.0	2.0
Total	17.3	17.7	160.7	182.4
Eastern Europe and Former U.S.S.R.				
Albania	0.2	0.2	0.1	0.1
Bulgaria	(s)	(s)	0.2	0.1
Czech Republic	(s)	(s)	0.1	0.1
Slovakia	(s)	NA	0.5	NA
Hungary	0.1	0.1	1.3	2.3
Poland	0.1	0.1	5.1	5.8
Romania	1.0	1.2	3.6	4.3
Azerbaijan	1.2	NA	4.4	NA
Kazakhstan	5.4	NA	65.0	NA
Russia	48.6	53.9	1680.0	1,700.0
Turkmenistan	0.5	NA	101.0	NA
Ukraine	0.4	NA	39.6	NA
Uzbekisan	0.6	NA	66.2	NA
Other	0.3	11.7	0.8	237.9
Total	58.4	67.1	1967.9	1950.5
Middle East				
Bahrain	0.1	NA	3.2	NA
Iran	89.7	99.1	812.3	939.4

Continues

Table continued

Region / Country	Crude oil (billion barrels)		Natural gas (trillion cubic feet)	
	Oil and gas journal	World oil	Oil and gas journal	World oil
Iraq	112.5	115.0	109.8	112.6
Israel	(s)	NA	1.5	NA
Jordan	(s)	NA	0.2	NA
Kuwait[b]	96.5	98.8	52.7	56.6
Oman	5.5	5.9	29.3	30.5
Qatar	15.2	13.8	508.5	757.7
Saudi Arabia[b]	261.8	261.7	219.5	228.2
Syria	2.5	2.3	8.5	8.5
United Arab Emirates	97.8	62.8	212.1	204.1
Yemen	4.0	2.4	16.9	17.0
Other	0.0	0.7	0.0	13.4
Total	685.6	662.5	1974.6	2367.9
Africa				
Algeria	9.2	17.0	159.7	175.0
Angola	5.4	6.0	1.6	4.0
Benin	(s)	NA	(s)	NA
Cameroon	0.4	NA	3.9	NA
Congo (Brazzaville)	1.5	1.6	3.2	4.2
Congo (Kinshasa)	0.2	NA	(s)	NA
Cote d'lvoire (Ivory Coast)	0.1	NA	1.1	NA
Egypt	2.9	3.7	35.2	54.1
Equatorial Guinea	(s)	1.1	1.3	3.5
Ethiopia	(s)	NA	0.9	NA
Gabon	2.5	2.4	1.2	3.5
Ghana	(s)	NA	0.8	NA
Libya	29.5	30.0	46.4	46.9
Madagascar	0.0	NA	0.0	NA
Morocco	(s)	NA	(s)	NA
Mozambique	0.0	NA	4.5	NA
Namibia	0.0	NA	2.2	NA
Nigeria	24.0	30.0	124.0	159.0
Rwanda	0.0	NA	2.0	NA
Somalia	0.0	NA	0.2	NA
South Africa	(s)	NA	(s)	NA
Sudan	0.6	0.7	3.0	4.0
Tanzania	0.0	NA	0.8	NA
Tunisia	0.3	0.5	2.8	2.7
Other	0.0	1.9	0.0	20.1
Total	76.7	94.9	394.8	477.1
Asia and Oceania				
Afghanistan	0.0	NA	3.5	NA
Australia	3.5	3.8	90.0	80.0
Bangladesh	0.1	NA	10.6	NA
Brunei	1.4	1.2	13.8	8.4
Burma	0.1	0.2	10.0	12.2

Continues

Table continued

Region Country	Crude oil (billion barrels)		Natural gas (trillion cubic feet)	
	Oil and gas journal	World oil	Oil and gas journal	World oil
China	24.0	29.5	48.3	42.8
India	4.8	3.8	22.9	15.4
Indonesia	5.0	9.2	92.5	87.5
Japan	0.1	NA	1.4	NA
Malaysia	3.0	4.5	75.0	82.5
New Zealand	0.1	0.1	2.1	2.1
Pakistan	0.3	0.3	25.1	24.1
Papua New Guinea	0.2	0.5	12.2	15.0
Philippines	0.2	0.2	3.7	3.7
Taiwan	(s)	NA	2.7	NA
Thailand	0.5	0.6	12.7	13.3
Vietnam	0.6	2.2	6.8	6.8
Other	0.0	0.6	0.0	26.1
Total	43.8	56.5	433.3	419.9
World Total	1032.0	1018.7	5457.1	5930.2

Sources. Data from PennWell Publishing Co., Oil and Gas Journal Vol. 99, No. 52 (December 2001); Gulf Publishing Co., World Oil Vol. 223, No. 8 (August 2002); Energy Information Administration, U.S. Crude Oil, Natural Gas, and Natural Gas Liquids Reserves, 2001 Annual Report, DOE/EIA-0216 (2001) (November 2002). Table modified from Energy Information Administration/International Energy Annual 2001.

Notes. Sum of components may not equal total due to independent rounding. All reserve data except those for the Former U.S.S.R. and natural gas reserves in Canada are proved reserves. Former U.S.S.R. data are "explored reserves," which are understood to be proved, and some probable. World Oil only reported disaggregated data for the Former U.S.S.R. Republic of Russia. Aggregated data for the other Republics of the Former U.S.S.R. were reported as 11.5 billion barrels of crude oil and 235.2 trillion cubic feet of natural gas.

[a]Data for the United States are from the Energy Information Administration.

[b]Includes one-half of the reserves in the Neutral Zone.

NA, not available; (s), value less than 50 million barrels of crude oil or less than 50 billion cubic feet of natural gas.

16. WORLD HYDROPOWER RESOURCES

Hydropower: Capability at End-1999 (TWh/year)

	Gross theoretical capability	Technically exploitable capability	Economically exploitable capability
Algeria	12	5	
Angola	150	90	65
Benin	2	1	
Burkina Faso	1	N	N
Burundi	>6	>1	1
Cameroon	294	115	103
Central African Republic	7	3	
Chad	N	N	
Comoros			
Congo (Brazzaville)	>125	>50	

Continues

Table continued

	Gross theoretical capability	Technically exploitable capability	Economically exploitable capability
Congo (Democratic Rep.)	1397	774	<419
Côte d'Ivoire	46	>12	12
Egypt (Arab Rep.)	>125	>50	50
Equatorial Guinea			
Ethiopia	650	>260	260
Gabon	200	80	33
Ghana	17	11	7
Guinea	26	19	15
Guinea-Bissau	1	N	N
Kenya	>30	9	
Lesotho	5	2	
Liberia	28	11	
Madagascar	321	180	49
Malawi	15	6	
Mali	>12	>5	
Mauritania			
Mauritius	N	N	
Morocco	12	5	4
Mozambique	50	38	32
Namibia	9	9	9
Niger	>3	>1	1
Nigeria	43	32	30
Réunion			
Rwanda	1	N	
São Tomé and Principe			
Senegal	11	4	2
Sierra Leone	17	7	
Somalia	2	1	
South Africa	73	11	5
Sudan	48	19	2
Swaziland	4	1	N
Tanzania	39	20	2
Togo	4	2	
Tunisia	1	N	N
Uganda	>18	>7	
Zambia	52	29	11
Zimbabwe	19	18	
Total Africa	>3876	>1888	
Belize	1	N	N
Canada	1289	951	523
Costa Rica	223	43	
Cuba	5	2	
Dominica	N	N	N
Dominican Republic	50	9	6
El Salvador	7	5	2
Greenland	470	14	
Grenada	N	N	N

Continues

Table continued

	Gross theoretical capability	Technically exploitable capability	Economically exploitable capability
Guadeloupe			
Guatemala	55	22	
Haiti	4	1	N
Honduras	16	6	
Jamaica	N	N	N
Mexico	154	64	38
Nicaragua	33	10	7
Panama	26	>12	12
Puerto Rico			
St. Vincent and the Grenadines			
United States of America	4485	529	376
Total North America	6818	>1668	
Argentina	172	130	
Bolivia	178	126	50
Brazil	3040	1488	811
Chile	227	162	
Colombia	1000	200	140
Ecuador	115	32	16
French Guiana			
Guyana	61	>25	25
Paraguay	111	85	68
Peru	1578	>260	260
Surinam	32	13	
Uruguay	32	10	
Venezuela	345	261	130
Total South America	6891	>2792	
Afghanistan			
Armenia	22	8	6
Azerbaijan	44	16	7
Bangladesh	5	2	
Bhutan	263	70	56
Cambodia	208	83	24
China	5920	1920	1260
Cyprus	59	24	
Georgia	139	68	32
India	2638	660	
Indonesia	2147	402	40
Japan	718	136	114
Kazakhstan	163	62	27
Korea (Democratic People's Republic)			
Korea (Republic)	52	26	19
Kyrgyzstan	163	99	55
Laos	233	63	42
Malaysia	230	123	
Mongolia	56	22	
Myanmar (Burma)	877	130	

Continues

Table continued

	Gross theoretical capability	Technically exploitable capability	Economically exploitable capability
Nepal	727	158	147
Pakistan	210	130	130
Philippines	47	20	18
Sri Lanka	11	8	7
Taiwan, China	103	14	12
Tajikistan	527	>264	264
Thailand	56	19	18
Turkey	413	216	122
Turkmenistan	24	5	2
Uzbekistan	88	27	15
Vietnam	300	100	80
Total Asia	16,443	>4875	
Albania	40	15	6
Austria	75	>56	56
Belarus	7	3	2
Belgium	1	N	N
Bosnia-Herzogovina	69	24	19
Bulgaria	26	15	12
Croatia	10	9	8
Czech Republic	12	4	
Denmark	N	N	N
Estonia	2	N	N
Faroe Islands	1	N	N
Finland	47	>20	20
FYR Macedonia	9	6	
France	200	72	70
Germany	120	26	20
Greece	80	15	12
Hungary	7	5	
Iceland	184	64	40
Ireland	1	1	1
Italy	340	105	65
Latvia	7	6	5
Lithuania	5	3	2
Luxembourg	N	N	N
Moldova	2	1	1
Netherlands	1	N	N
Norway	600	200	180
Poland	23	14	7
Portugal	33	25	20
Romania	56	36	17
Russian Federation	2800	1670	852
Serbia, Montenegro	68	>27	27
Slovakia	10	>7	6
Slovenia	13	9	8
Spain	138	70	41
Sweden	176	130	90

Continues

Table continued

	Gross theoretical capability	Technically exploitable capability	Economically exploitable capability
Switzerland	144	41	35
Ukraine	45	24	19
United Kingdom	40	3	1
Total Europe	5392	>2706	
Iran (Islamic Rep.)	368	88	48
Iraq	225	90	67
Israel	88	<35	
Jordan	N	N	N
Lebanon	2	1	
Syria (Arab Rep.)	5	4	4
Total Middle East	688	<218	
Australia	264	>30	30
Fiji	3	1	
French Polynesia	N	N	N
New Caledonia			
New Zealand	152	77	40
Palau			
Papua New Guinea	175	123	37
Solomon Islands	2	>1	
Western Samoa	N	N	
Total oceania	596	>232	
TOTAL WORLD	>40,704	>14,379	

Sources. WEC Member Committees, 2000/2001; *Hydropower and Dams World Atlas 2001*, supplement to The International Journal on Hydropower and Dams, Aqua-Media International; estimates by the editors.

Note. A quantification of hydropower capability is not available for Comoros, Equatorial Guinea, Mauritania, Réunion, São Tomé and Principe, Guadeloupe, Puerto Rico, St. Vincent and the Grenadines, French Guiana, Afghanistan, Korea (Democratic People's Republic), New Caledonia, and Palau. As the data available on economically exploitable capability do not cover all countries, regional and global totals are not shown for this category.

17. ESTIMATES OF TERRESTRIAL CARBON STOCKS

Estimates of Terrestrial Carbon Stocks and NPP (Global Aggregated Values by Biome)

		Global carbon stocks (PgC)[a]			Carbon density (MgC/ha)	
Biome	Area (10^9 ha)	Plants	Soil	Total	Plants	Soil
Tropical forests	1.76	212	216	428	120	123
Temperate forests	1.04	59	100	159	57	96
Boreal forests	1.37	88[b]	471	559	64	344
Tropical savannas and grasslands	2.25	66	264	330	29	117
Temperate grasslands and shrublands	1.25	9	295	304	7	236
Deserts and semi deserts	4.55[c]	8	191	199	2	42
Tundra	0.95	6	121	127	6	127

Continues

Table continued

Biome	Area (10^9 ha)	Global carbon stocks (PgC)[a]			Carbon density (MgC/ha)	
		Plants	Soil	Total	Plants	Soil
Croplands	1.60	3	128	131	2	80
Wetlands	0.35	15	225	240	43	643
Total	15.12	466	2011	2477		

Source. Data from WBGU (1988); forest data from Dixon *et al*. (1994). Table modified from Intergovernmental Panel on Climate Change.

[a]Soil carbon values are for the top 1 m, although stores are also high below this depth in peatlands and tropical forests.

[b]WBGU boreal forest vegetation estimate is likely to be to high, due to high Russian forest density estimates including standing dead biomass.

[c]In WBGU, ice is included in deserts and semi-deserts category.

18. COMPARATIVE LEVELS OF GREENHOUSE GASES

Pre-industrial (1750) and Present (1998) Abundances of Well-Mixed Greenhouse Gases and the Radiative Forcing due to the Change in Abundance

Gas	Abundance (1750)	Abundance (1998)	Radiative forcing (Wm^{-2})
Gases relevant to radiative forcing only			
CO_2	278	365	1.46
CH_4	700	1745	0.48
N_2O	270	314	0.15
CF_4	40	80	0.003
C_2F_6	0	3	0.001
SF_6	0	4.2	0.002
HFC-23	0	14	0.002
HFC-134a	0	7.5	0.001
HFC-152a	0	0.5	0.000
Gases relevant to radiative forcing and ozone depletion			
CFC-11	0	268	0.07
CFC-12	0	533	0.17
CFC-13	0	4	0.001
CFC-113	0	84	0.03
CFC-114	0	15	0.005
CFC-115	0	7	0.001
CCl_4	0	102	0.01
CH_3CCl_3	0	69	0.004
HCFC-22	0	132	0.03
HCFC-141b	0	10	0.001
HCFC-142b	0	11	0.002
Halon-1211	0	3.8	0.001
Halon-1301	0	2.5	0.001

Source. Intergovernmental Panel on Climate Change.

Note. Volume mixing ratios for CO_2 are in ppm, for CH_4 and N_2O in ppb, and for the rest in ppt.

19. GLOBAL WARMING POTENTIALS

Numerical Estimates of Global Warming Potentials Compared with Carbon Dioxide (Kilogram of Gas per Kilogram of Carbon Dioxide)

Gas	Lifetime (years)	Direct effect for time horizons of		
		20 years	100 years	500 years
Carbon dioxide	5–200[a]	1	1	1
Methane	12	62	23	7
Nitrous oxide	114	275	296	156
Hydrofluorocarbons, perfluorocarbons, and sulfur hexafluoride				
HFC-23	260	9400	12,000	10,000
HFC-125	29	5900	3400	1100
HFC-134a	13.8	3300	1300	400
HFC-152a	1.4	410	120	37
HFC-227ea	33	5600	3500	1100
Perfluoromethane (CF_4)	50,000	3900	5700	8900
Perfluoroethane (C_2F_6)	10,000	8000	11,900	18,000
Sulfur Hexafluoride (SF_6)	3200	15,100	22,200	32,400

Source. Energy Information Administration/Emissions of Greenhouse Gases in the United States, 2001.

[a]No single lifetime can be defined for carbon dioxide due to different rates of uptake by different removal processes.

20. INTERNATIONAL ENERGY PRODUCTION AND CONSUMPTION

Country	Total energy production (ktoe)[a]			Total energy consumption (ktoe)[a]			Net energy imports (ktoe)[a]	
	1971	2000	Percent change	1971	2000	Percent change	1971	2000
Albania	2417.57	814.42	−66	1699.04	1634.31	−4	−718.53	819.89
Algeria	41,766.71	149,628.66	258	3681.12	29,059.5	689	−36,412.97	−120,895.29
Angola	9118.21	43,669.29	379	3927.62	7666.77	95	−4663.42	−35,502.23
Argentina	30,557.87	81,220.95	166	33,651.83	61,469.41	83	3763.49	−19,677.61
Armenia	0	631.96		0	2060.72		0	1428.76
Australia	53,852.03	23,2551.98	332	52,202.09	11,0174.34	111	−1700.99	−12,0340.66
Austria	7389.31	9686.24	31	18,994.03	28,582.2	50	11,693.55	18,665.52
Azerbaijan	0	18,951.09		0	11,703.49		0	−7351.6
Bahrain	4572.94	7261.34	59	1553.06	6391.22	312	−1978.95	−1169.01
Bangladesh	4848.38	15,053.27	210	5706.09	18,665.62	227	1052.91	3510.82
Belarus	0	3465.89		0	24,329.52		0	20,942.13
Belgium	6844.61	13,438.15	96	39,943.74	59,216.59	48	36,850.95	50,599.95
Benin	1002.24	1820.51	82	1110.34	2362.49	113	108.09	499.81
Bolivia	2303.44	5900.81	156	1035.22	4929.17	376	−1239.47	−1483.96
Bosnia and Herzegovina	0	3277.19		0	4359.31		0	1082.13
Brazil	49,230.06	142,078.01	189	69,620.18	183,165.01	163	22,296.9	43,408.07
Brunei	6650.65	19,158.61	188	178.37	1984.08	1012	−5947.48	−17,248.1
Bulgaria	5711.6	10,005.26	75	19,238.88	18,783.63	−2	13,587.77	8729.84

Continues

Table continued

Country	Total energy production (ktoe)[a]			Total energy consumption (ktoe)[a]			Net energy imports (ktoe)[a]	
	1971	2000	Percent change	1971	2000	Percent change	1971	2000
Cameroon	2452.43	12,729.48	419	2751.68	6355.31	131	299.26	−6335.48
Canada	155,646.37	374,863.52	141	142,737.01	250,967.31	76	−120,84.99	−131,791.97
Chile	5343.85	8299.13	55	8851.1	24,403.36	176	4016.32	16,955.57
Chinese Taipei	3948.05	11,492.66	191	10,582.05	83,010.81	684	7256.44	80,829.19
Colombia	19,370.2	74,583.6	285	14,052.09	28,785.52	105	−4998.3	−46,627.26
Congo	505.04	14,655.61	2802	678.86	894.52	32	167.68	−13,761.09
Costa Rica	663.1	1591.06	140	1134.12	3281.38	189	462.48	1631.73
Cote d'Ivoire	1639.19	6096.91	272	2500.37	6928.42	177	822.62	921.92
Croatia	0	3581.7		0	7775.46		0	4116.95
Cuba	3738.85	6051.38	62	10,933.92	13,202.86	21	7436.3	7744.37
Cyprus	9.12	44.92	393	646.2	2424.85	275	635.84	2579.74
Czech Republic	39,949.22	29,868.74	−25	45,638.68	40,383.04	−12	5739.69	9412.93
Democratic Republic of Congo	5928.83	15,446.24	161	6767.97	14,888.29	120	964.9	−556.03
Denmark	329.22	27,867.36	8365	19,242.04	19,456.27	1	20,038.8	−7533.22
Dominican Republic	1212.35	1420.96	17	2367.63	7804	230	1155.29	6383.03
Ecuador	1340.9	22,520.14	1579	2345.31	8187.15	249	1153.7	−13 577.79
Egypt	16,356.07	57,599.29	252	7834.98	46,423.15	493	−8608.62	−7297.08
El Salvador	1254.89	2157.49	72	1762.78	4083.03	132	501.75	1912.99
Eritrea	0	510.94		0	720.67		0	209.74
Estonia	0	2916.57		0	4522.92		0	1694.36
Ethiopia	8471.55	17,582.68	108	8998.18	18,732.19	108	588.38	1095.93
Federal Republic of Yugoslavia	0	10,122.05		0	13,706.03		0	3583.97
Finland	4983.17	15,134.29	204	18,425.2	33,147.26	80	14,621.11	18,766.49
Former USSR	902,516.94	126,1902.02	40	788,362	92,1142.69	17	−107,116.82	−336,117.48
Former Yugoslav Republic of Macedonia	0	1547.34		0	2754.79		0	1145.5
Former Yugoslavia	14,372.66	21,640.97	51	22,134.9	35,135.5	59	6945.44	13,391.36
France	39,935.27	131,384.95	229	154,514.54	257,127.74	66	119,981.94	132,319.89
Gabon	6453.27	16,799.98	160	919.91	1563.11	70	−5469.08	−151,45.3
Georgia	0	737.19		0	2859.82		0	2092.3
Germany	175,200.71	134,316.7	−23	307,865.28	339,640.48	10	143,631.41	203,247.24
Ghana	2337.29	5883.35	152	3040.41	7719.75	154	770.84	1836.39
Gibraltar	0	0		38.2	161.53	323	214.76	1148.11
Greece	2084.89	9986.85	379	9147.83	27,821.63	204	7717.28	21,721.36
Guatemala	1933.76	5241.11	171	2797.75	7146.38	155	924.46	1936.01
Haiti	1378	1541.75	12	1510.82	2039.24	35	132.82	497.49
Honduras	1012.19	1522.21	50	1396.63	3012.25	116	369.53	1598.87
Hong Kong, China	31.02	48.05	55	3474.35	15,453.03	345	4059.65	18,910.31
Hungary	11,974.36	11,089.6	−7	19,216.37	24,782.92	29	7677.5	13,896.82
Iceland	413.83	2490.16	502	976.69	3428.77	251	559.66	1047.83
India	167,718.75	421,565.11	151	183,213.41	501,894.17	174	146,34.46	811,99.38
Indonesia	729,35.99	229,478.35	215	36,313.6	145,574.74	301	−33,867.99	−85,107.4
Iraq	85,980.03	134,088.88	56	4508.5	27,677.87	514	−81,462.5	−106,411.01
Ireland	1423.33	2197.27	54	7066.04	14,622.85	107	5840.54	12,456.58
Iran	238,830.7	242,146.01	1	18,564.91	112,725.35	507	−219,862.6	−128,802.42

Continues

Table continued

Country	Total energy production (ktoe)[a]			Total energy consumption (ktoe)[a]			Net energy imports (ktoe)[a]	
	1971	2000	Percent change	1971	2000	Percent change	1971	2000
Israel	5938.4	653.73	−89	6340.7	20,199.92	219	402.29	19,292.58
Italy	19,241.96	26,858.22	40	114,076.65	171,567.12	50	104,245.92	152,370.48
Jamaica	277.03	486.42	76	2151.22	3920.44	82	1923.15	3635.94
Japan	35,782.62	105,635.85	195	269,567.84	524,714.63	95	253,010.4	425,266.33
Jordan	0.93	286.19	30,673	531.74	5185.43	875	621.62	4835.34
Kazakhstan	0	78,101.98		0	39,063.16		0	−39,344.27
Kenya	6367.72	12,259.93	93	7699.3	15,482.47	101	1742.24	3304.67
Korea	6378.17	33,614.62	427	17,024.39	193,626.22	1037	10,687.77	167,665.02
Korea, DPR	18,675.57	42,575.97	128	19,953.77	46,112.3	131	1278.19	3536.33
Kuwait	159,081.47	111,469.48	−30	4964.81	20,893.52	321	−152,363.46	−90,222.1
Kyrgyzstan	0	1443.05		0	2444.65		0	959.39
Latvia	0	1249.88		0	3654.98		0	2312.92
Lebanon	168.49	170.6	1	1977.81	5058.07	156	2043.44	4889.51
Libya	137,513.8	73,903.73	−46	1659.62	16,437.78	890	−136,145.05	−57,379.55
Lithuania	0	3217.99		0	7123.74		0	4213.15
Luxembourg	4.73	56.68	1098	4115.06	3683.14	−10	4105.41	3676.74
Malaysia	4770.24	76,759.46	1509	6032.02	49,472.33	720	1295.5	−26,903.45
Malta	0	0		273.77	814.51	198	333.68	861.04
Mexico	45,500.02	229,652.63	405	45,586.58	153,513.18	237	317.58	−73,168.79
Morocco	613.38	571.83	−7	2417.48	10,293.03	326	2029.72	9530.85
Mozambique	7495.61	7219.2	−4	8265.73	7126.19	−14	1013.5	−86.9
Myanmar	7340.54	15,144.09	106	7910.68	12,522.07	58	408.75	−2695.64
Namibia	0	291.6		0	1030.7		0	739.11
Nepal	2506.29	6871.72	174	2570.06	7900.44	207	63.78	1028.71
Netherlands	37,345.95	57,239.02	53	51,304.14	75,799.02	48	23,821.89	34,465.75
Netherlands Antilles	0	0		5515.09	1084.15	−80	6815.18	2767.06
New Zealand	3417.3	15,379.49	350	7175.55	18,632.85	160	4083.06	3378.86
Nicaragua	736.44	1552.56	111	1258.5	2746.16	118	524.04	1158.25
Nigeria	111,492.73	197,726.15	77	36,150.5	90,168.58	149	−74,965.76	−106,986.64
Norway	6044.07	224,993.45	3623	13,850.4	25,616.63	85	8145.24	−198,190.27
Oman	14,889.96	60,083.98	304	89.21	9750.04	10,829	−13,461.01	−51,113.52
Pakistan	14,355.44	47,124.13	228	17,421.19	63,950.6	267	2920.27	17,205.51
Panama	342.58	731.72	114	1672.38	2546.05	52	1352.66	1879.6
Paraguay	1171.09	6885.91	488	1381.05	3929.5	185	209.96	−2918.56
People's Republic of China	394,147.55	110,7635.92	181	391,711.57	114,2438.44	192	−2338.06	29,660.31
Peru	7468.46	9477.39	27	9416.02	12,694.68	35	2077.58	3347.92
Philippines	6308.68	20,922.07	232	14,610.53	42,424.28	190	8665.43	21,613.81
Poland	99,249.14	80,069.68	−19	86,313.9	89,975.37	4	−10,717.36	9625.97
Portugal	1384.49	3129.18	126	6527.74	24,612.61	277	5742.08	22,110.56
Qatar	22,027.2	57960	163	928.9	15,659.05	1586	−21,035.65	−44,736.6
Republic of Moldova	0	60.34		0	2871.12		0	2810.86
Romania	42,807.77	28,289.87	−34	42,086.3	36,330.32	−14	−717.03	7849.1
Russia	0	966,512.12		0	613,968.84		0	−347,821.9
Saudi Arabia	244,427.47	487,888.82	100	6241.54	105,303.32	1587	−225,693.54	−380,533.98
Senegal	825.03	1723.27	109	1343.06	3085.87	130	1439.8	1439.35
Singapore	0	63.55		2958.05	24,591.12	731	6622.94	40,887.32

Continues

Table continued

Country	Total energy production (ktoe)[a]			Total energy consumption (ktoe)[a]			Net energy imports (ktoe)[a]	
	1971	2000	Percent change	1971	2000	Percent change	1971	2000
Slovak Republic	2687.23	5993.94	123	14,237.52	17,465.69	23	11,537.32	11,552.07
Slovenia	0	3112.69		0	6539.91		0	3462.81
South Africa	37,773.37	144,468.52	282	45,328.45	107,595.23	137	11,213.61	−34,036.81
Spain	10,449.87	31,865.48	205	43,104.53	124,881.28	190	35,415.2	100,232.96
Sri Lanka	2810	4529.93	61	3797.84	8062.93	112	1360.08	3790.78
Sudan	5667.59	23,664.26	318	6950.82	16,215.84	133	1474.62	−3629.6
Sweden	7385.94	30,680.94	315	36,484.91	47,481.11	30	31,495.29	18,057.72
Switzerland	2904.67	11,792.22	306	17,069.07	26,596.52	56	14,227.03	14,275.75
Syria	5315.26	32,889.56	519	2710.54	18,407.3	579	−2540.52	−16,232.23
Tajikistan	0	1249.76		0	2911.23		0	1661.47
Tanzania	8958.94	14,600.74	63	9692.16	15,386.26	59	869.63	809.59
Thailand	7892.85	41,117.54	421	14,119.2	73,618.34	421	6361.14	33,248.16
Togo	457.5	1035.74	126	558.94	1529.74	174	101.44	496.99
Trinidad and Tobago	8148.64	17,884.38	119	2698.4	8664.78	221	−3615.47	−9080.45
Tunisia	4912.14	7002.77	43	2042.99	7887.83	286	−2560.23	1013.22
Turkey	13,770.26	26,185.53	90	19,464.73	77,103.97	296	6107.37	50,998.37
Turkmenistan	0	45,967.97		0	13,884.74		0	−32,083.23
Ukraine	0	82,330.29		0	139,592.33		0	57,262.04
United Arab Emirates	52,785.56	143,588.88	172	1019.46	29,558.79	2799	−51,766.09	−109,109.34
United Kingdom	109,796.11	272,693.22	148	211,031.22	232,644.31	10	112,578.25	−40,720.57
United States	1,435,818.82	1,675,769.62	17	1,593,191.14	2,299,668.71	44	1,77,569.5	6,08,890.98
Uruguay	519.66	1028.29	98	2417.66	3078.64	27	1919.61	2261.51
Uzbekistan	0	55,065.94		0	50,151.43		0	−4893.86
Venezuela	210,431.4	225,470.15	7	20,142.47	59,256.39	194	−185,445.76	−165,159.16
Vietnam	14,204.46	46,299.4	226	19,788.11	36,964.65	87	5579.17	−9334.74
Yemen	48.72	22,046.49	45,151	764.17	3525.85	361	1069.55	−18,422.67
Zambia	3025.68	5915.55	96	3850.7	6243.8	62	825.02	366.72
Zimbabwe	5102.97	8708.46	71	5480.74	10,218.9	86	377.78	1407.7

Source. International Energy Agency.

[a]ktoe, kilotons of oil equivalent.

21. INTERNATIONAL HEAT CONTENT COMPARISONS

Nation	Abbrev.	Gross heat content natural gas liquids[a]	Barrels of crude oil expressed as metric tons	Gross heat content natural crude oil[a]	Gross heat content dry natural gas[b]	Gross heat content coal[c]
North America						
Canada	CA	3980	7.186	5810	1017	23,865
Mexico	MX	3620	6.965	6010	1059	17,012
United States	US	3735	7.333	5800	1025	20,905

Continues

Table continued

Nation	Abbrev.	Gross heat content natural gas liquids[a]	Barrels of crude oil expressed as metric tons	Gross heat content natural crude oil[a]	Gross heat content dry natural gas[b]	Gross heat content coal[c]
South America						
Argentina	AR	3820	7.120	5993	1045	22,300
Barbados	BB	0	7.286	5867	1047	0
Bolivia	BL	4130	7.881	5574	1043	0
Brazil	BR	4250	7.056	5910	1040	14,363
Chile	CI	4150	7.506	5780	1050	25,664
Colombia	CO	3286	7.080	6023	929	24,568
Cuba	CU	4344	6.449	6393	1047	0
Ecuador	EC	4235	7.130	5986	1300	0
Guatemala	GT	0	7.210	5947	0	0
Peru	PE	4617	7.407	5831	929	26,458
Suriname	NS	0	6.199	6321	0	0
Trinidad and Tobago	TD	4344	7.084	6023	1045	0
Venezuela	VE	4195	7.127	6135	1191	27,592
Western Europe						
Austria	AU	4262	7.200	6020	1063	9929
Belgium	BE	0	0.000	0	0	19,908
Bosnia and Herzegovina	BK	0	0.000	0	0	13,252
Croatia	HR	4349	7.418	5823	1047	0
Denmark	DA	0	7.405	5677	1130	0
France	FR	4385	7.332	5869	1008	22,876
Germany	GM	0	7.330	5926	895	10,009
Greece	GR	4050	7.231	5926	1434	5046
Ireland	EI	0	0.000	0	1009	0
Italy	IT	4259	7.300	6158	1023	4381
Macedonia (Former Yugoslav Republic)	MK	0	0.000	0	0	8013
Netherlands	NL	4347	7.239	6158	894	0
Norway	NO	4348	7.644	5620	1078	25,369
Slovenia	SI	0	7.418	5823	1057	10,192
Spain	SP	4348	7.506	5780	1139	11,378
Sweden	SW	0	6.623	6148	0	0
Turkey	TU	0	7.200	5976	1028	7437
United Kingdom	UK	4490	7.523	5803	1057	22,727
Yugoslavia	YR	0	7.418	5823	1061	8028
Eastern Europe						
Albania	AL	0	6.594	6305	1044	8882
Azerbaijan	AJ	4150	7.270	5880	1047	0
Belarus	BO	0	7.270	5880	1037	0
Bulgaria	BU	0	7.332	5879	989	8038
Czech Republic	EZ	4645	6.780	6211	956	11,796
Georgia	GG	4150	7.270	5880	1047	13,229
Hungary	HU	4386	6.690	6249	957	7784
Kazakhstan	KZ	4150	7.270	5880	1047	13,229
Kyrgyzstan	KG	4150	7.270	5880	1047	13,229

Continues

Table continued

Nation	Abbrev.	Gross heat content natural gas liquids[a]	Barrels of crude oil expressed as metric tons	Gross heat content natural crude oil[a]	Gross heat content dry natural gas[b]	Gross heat content coal[c]
Lithuania	LH	0	7.270	5880	0	0
Poland	PL	4454	7.418	5820	797	15,976
Romania	RO	4619	7.506	5780	996	7259
Russia	RS	4150	7.270	5880	1009	18,302
Slovakia	LO	0	6.780	6211	960	11,067
Tajikistan	TI	4150	7.270	5880	1047	16,774
Turkmenistan	TX	4150	7.270	5880	1047	0
Ukraine	UP	4150	7.270	5880	1047	19,411
Uzbekistan	UZ	4150	7.270	5880	1017	13,322
Middle East						
Bahrain	BA	4287	7.320	5879	1047	0
Iran	IR	4348	7.284	5888	1056	23,245
Iraq	IZ	4348	7.413	5820	1047	0
Israel	IS	0	7.247	5920	1039	0
Jordan	JO	0	7.190	5956	1047	0
Kuwait	KU	4190	7.258	5921	1047	0
Oman	MU	4581	7.330	5869	1047	0
Qatar	QA	3790	7.618	5777	1047	0
Saudi Arabia	SA	4265	7.285	5910	1047	0
Syria	SY	4300	7.290	6158	962	0
United Arab Emirates	TC	4532	7.553	5788	1047	0
Yemen	YM	0	7.631	5725	0	0
Africa						
Algeria	AG	5080	7.945	5555	1127	23,245
Angola	AO	0	7.410	5828	1047	0
Benin	BN	0	6.870	6142	0	0
Botswana	BC	0	0.000	0	0	23,245
Cameroon	CM	0	7.205	5948	1047	23,245
Congo (Brazzaville)	CF	0	7.506	5780	1047	0
Congo (Kinshasa)	CG	0	7.320	5879	0	22,776
Cote d'Ivoire (IvoryCoast)	IV	0	7.285	5899	1000	0
Egypt	EG	4940	7.260	5922	1047	0
Equatorial Guinea	EK	0	7.275	5464	1047	0
Gabon	GB	0	7.305	5888	1047	0
Ghana	GH	0	7.285	5936	0	0
Libya	LY	4350	7.681	5775	1047	0
Morocco	MO	0	7.600	5738	1047	21,166
Mozambique	MZ	0	0.000	0	1047	22,565
Niger	NG	0	0.000	0	0	23,245
Nigeria	NI	0	7.411	5880	1047	23,245
Senegal	SG	0	0.000	0	899	0
South Africa	SF	4649	7.720	5657	1047	21,302
Swaziland	WZ	0	0.000	0	0	23,245
Tanzania	TZ	0	0.000	0	0	23,245
Tunisia	TS	4300	7.689	5677	1174	0

Continues

Table continued

Nation	Abbrev.	Gross heat content natural gas liquids[a]	Barrels of crude oil expressed as metric tons	Gross heat content natural crude oil[a]	Gross heat content dry natural gas[b]	Gross heat content coal[c]
Zambia	ZA	0	0.000	0	0	22,304
Zimbabwe	ZI	0	0.000	0	0	24,371
Asia and Oceania						
Afghanistan	AF	0	0.000	0	1047	22,000
Australia	AS	4290	7.868	5578	1090	19,916
Bangladesh	BG	4050	7.453	5792	979	0
Bhutan	BT	0	0.000	0	0	23,245
Brunei	BX	4506	7.340	5865	1154	0
Burma	BM	4348	7.084	6020	1054	19,240
China	CH	0	7.300	5879	1162	18,512
India	IN	4600	7.330	5729	1034	17,705
Indonesia	ID	4060	7.234	5740	1090	23,245
Japan	JA	4327	7.357	5899	1101	20,826
Korea, North	KN	0	0.000	0	0	23,245
Korea, South	KS	0	0.000	0	0	17,009
Laos	LA	0	0.000	0	0	24,403
Malaysia	MY	4410	7.641	5697	1053	26,458
Mongolia	MG	0	0.000	0	0	8016
Nepal	NP	0	0.000	0	0	7559
New Zealand	NZ	4075	7.321	5441	1032	19,723
Pakistan	PK	4372	7.500	5777	934	16,910
Papua New Guinea	PP	0	7.809	5607	1047	0
Philippines	RP	0	7.285	5902	977	18,120
Taiwan	TW	3825	6.568	6321	1000	0
Thailand	TH	4349	6.758	6317	977	10,961
Vietnam	VM	0	7.082	6022	1036	21,166

Source. U.S. Department of Energy.

[a]Thousand Btu per barrel.

[b]Btu per cubic foot.

[c]Thousand Btu per short ton.

22. ENERGY USE OF TYPICAL HOME APPLIANCES

Appliance	Wattage
Aquarium	50–1210
Clock radio	10
Coffee maker	900–1200
Clothes washer	350–500
Clothes dryer	1800–5000
Dishwasher[a]	1200–2400
Dehumidifier	785
Electric blanket, single	60

Continues

Table continued

Appliance	Wattage
Electric blanket, double	100
Fans	
Ceiling	65–175
Window	55–250
Furnace	750
Whole house	240–750
Hair dryer	1200–1875
Heater (portable)	750–1500
Iron (clothes iron)	1000–1800
Microwave oven	750–1100
Personal computer	
CPU	120
CPU, sleep mode	30 or less
Monitor	150
Monitor, sleep mode	30 or less
Laptop	50
Radio (stereo)	70–400
Refrigerator (frost–free)	725
Television set (color)	
19-inch	65–110
27-inch	113
36-inch	133
Large screen projection	170
Flat screen	120
VCR	17–21
DVD	20–25
Toaster	800–1400
Toaster oven	1225
Vacuum cleaner	1000–1440
Water heater (40 gallon)	4500–5500
Water pump (deep well)	250–1100
Water bed (with heater)	120–380

Source. U.S. Department of Energy.

[a]Using the drying feature greatly increases consumption.

23. ENERGY COST OF HUMAN ACTIVITIES

Activity	METS[a]	Description
Bicycling	8	Bicycling, general
Bicycling	10	Bicycling, 14–15.9 mph, racing or leisure, fast, vigorous effort
Bicycling	16	Bicycling, >20 mph, racing, not drafting
Conditioning exercise	7	Bicycling, stationary, general
Conditioning exercise	7	Bicycling, stationary, 150 watts, moderate effort
Conditioning exercise	12.5	Bicycling, stationary, 250 watts, very vigorous effort

Continues

Table continued

Activity	METS[a]	Description
Conditioning exercise	8	Calisthenics, heavy, vigorous effort
Conditioning exercise	8	Circuit training
Conditioning exercise	6	Weight lifting
Conditioning exercise	5.5	Health club exercise, general
Conditioning exercise	9	Stair-treadmill ergometer, general
Conditioning exercise	3.5	Rowing, stationary, 50 watts, light effort
Conditioning exercise	12	Rowing, stationary, 200 watts, very vigorous effort
Conditioning exercise	6	Teaching aerobics exercise class
Conditioning exercise	4	Water aerobics, water calisthenics
Dancing	4.8	Ballet or modern, twist, jazz, tap, jitterbug
Dancing	5	Aerobic, low impact
Dancing	7	Aerobic, high impact
Dancing	5.5	Ballroom, dancing fast
Fishing and hunting	3	Fishing, general
Fishing and hunting	6	Fishing in stream, in waders
Fishing and hunting	2	Fishing, ice, siting
Fishing and hunting	5	Hunting, general
Home activities	3	Cleaning, heavy or major
Home activities	3	Cleaning, house or cabin, general
Home activities	3.5	Vacuuming
Home activities	2	Cooking or food preparation
Home activities	7.5	Carrying groceries upstairs
Home activities	2.3	Ironing
Home activities	1.5	Sitting—knitting, sewing, light wrapping (presents)
Home activities	4	Walk/run—playing with child(ren), moderate
Home repair	3	Automobile repair
Home repair	3	Carpentry, general workshop
Home repair	4.5	Painting
Home repair	4.5	Washing and waxing hull of sailboat, car, powerboat, airplane
Inactivity quiet	0.9	Sleeping
Inactivity light	1	Reclining—reading
Inactivity light	1	Meditating
Lawn and garden	5.5	Mowing lawn, general
Lawn and garden	2.5	Mowing lawn, riding mower
Lawn and garden	6	Mowing lawn, walk, hand mower
Lawn and garden	4.3	Raking lawn
Lawn and garden	4	Gardening, general
Miscellaneous	1.3	Sitting—reading, book, newspaper, etc.
Miscellaneous	1.5	Sitting—talking or talking on the phone
Music playing	4	Drums
Music playing	2.5	Piano or organ
Music playing	2.5	Violin
Music playing	3	Guitar, rock and roll band (standing)
Occupation	6	Coal mining, general
Occupation	5.5	Construction, outside, remodeling
Occupation	8	Farming, bailing hay, cleaning barn, poultry work
Occupation	2.5	Farming, driving tractor
Occupation	12	Fire fighter, general
Occupation	3.5	Locksmith

Continues

Table continued

Activity	METS[a]	Description
Occupation	2.5	Police, directing traffic (standing)
Occupation	1.3	Police, riding in squad car (sitting)
Occupation	4	Police, making an arrest (standing)
Occupation	1.5	Sitting—light office work, general
Occupation	1.5	Typing, electric, manual, or computer
Running	7	Jogging, general
Running	8	Running, 5 mph (12 min/mile)
Running	12.5	Running, 7.5 mph (8 min/mile)
Running	18	Running, 10.9 mph (5.5 min/mile)
Running	15	Running, stairs, up
Self care	1.5	Bathing (sitting)
Self care	1.5	Eating (sitting)
Sexual activity	1.5	Active, vigorous effort
Sexual activity	1.3	General, moderate effort
Sexual activity	1	Passive, light effort, kissing, hugging
Sports	8	Basketball, game
Sports	2.5	Billiards
Sports	3	Bowling
Sports	5	Children's games
Sports	9	Football, competitive
Sports	8	Frisbee, ultimate
Sports	4.5	Golf, general
Sports	4.5	Golf, walking and carrying clubs
Sports	3.5	Golf, using power cart
Sports	4	Hacky sack
Sports	8	Hockey, field
Sports	8	Hockey, ice
Sports	4	Juggling
Sports	8	Lacrosse
Sports	11	Rock climbing, ascending rock
Sports	5	Skateboarding
Sports	12	Roller blading (in-line skating)
Sports	10	Soccer, competitive
Sports	7	Soccer, casual, general
Sports	5	Softball or baseball, fast or slow pitch, general
Sports	5	Tennis, doubles
Sports	8	Tennis, singles
Transportation	1	Riding in a car or truck
Walking	7	Backpacking
Walking	6	Hiking, cross country
Walking	2.5	Bird watching
Walking	6.5	Race walking
Walking	2.5	Walking, 2.0 mph, level, slow pace, firm surface
Walking	3.3	Walking, 3.0 mph, level, moderate pace, firm surface
Walking	6.3	Walking, 4.5 mph, level, firm surface, very, very brisk
Water activities	16	Skindiving, fast
Water activities	10	Swimming laps, freestyle, fast, vigorous effort
Water activities	7	Swimming laps, slow, moderate or light effort
Winter activities	7	Skating, ice, general

Continues

Table continued

Activity	METS[a]	Description
Winter activities	7	Skiing, general
Winter activities	7	Skiing, cross country, 2.5 mph, slow or light effort, ski walking
Winter activities	9	Skiing, cross country, 5.0–7.9 mph, brisk speed, vigorous effort
Winter activities	8	Skiing, downhill, vigorous effort

Source. Ainsworth, B.E. (2002, January). The Compendium of Physical Activities Tracking Guide. Prevention Research Center, Norman J. Arnold School of Public Health, University of South Carolina. Retrieved November 29, 2003 from http://prevention.sph.sc.edu/tools/compendium_tracking.pdf.

[a]The term MET (metabolic equivalent) is the ratio of a work metabolic rate to the resting metabolic rate. One MET is defined as 1 kcal/kg/hour and is roughly equivalent to the energy cost of sitting quietly. A MET also is defined as oxygen uptake in ml/kg/min, with one MET roughly equivalent to 3.5 ml/kg/min. Multiples of 1 MET indicate a higher energy cost for a specific activity. For example, a 2 MET activity requires twice the energy cost of sitting quietly.

24. ENERGY CONTENT OF COMMON FOODS

Food	Kcal/100 g
Fruits	
Apples	52
Apricots	48
Bananas	89
Figs	74
Grapefruit	32
Guavas	51
Mangos	65
Melons, cantaloupe	34
Oranges	47
Strawberries, raw	32
Tomatoes, red	18
Vegetables	
Asparagus, raw	20
Broccoli, cooked	35
Carrots, raw	41
Chickpeas, cooked	164
Corn, sweet, yellow, raw	86
Cucumber, peeled, raw	12
Lentils, cooked	101
Lettuce, romaine, raw	17
Mushrooms, raw	22
Onions, raw	42
Peas, edible-podded, raw	42
Peppers, sweet, red, raw	26
Pickles, cucumber, dill	18
Potato, flesh and skin, raw	77
Potatoes, russet, raw	79

Continues

Table continued

Food	Kcal/100 g
Meats and proteins	
Beans, black, cooked	132
Beans, navy	67
Beef	230–400
Catfish, channel, wild, raw	95
Caviar	252
Chicken	200–500
Cod, atlantic, raw	82
Crab, blue, raw	87
Flatfish (flounder and sole), raw	91
Herring, atlantic, raw	158
Lamb	150–500
Mixed nuts, dry roasted	594
Peanut butter	635
Pork, cooked	418
Salmon, Atlantic, wild, raw	142
Shrimp, raw	106
Soybeans, cooked	141
Tofu, firm	77
Turkey	205
Dairy	
Cheese, brie	334
Cheese, cheddar	403
Cheese, swiss	380
Egg, whole, raw, fresh	147
Milk, whole, 3.25% milkfat	60
Yogurt, plain whole milk	61
Fats and Oils	
Butter	717
Margarine	719
Olive oil	884
Grains	
Bread, french	274
Bread, rye	259
Bread, whole-wheat	246
Cereal, breakfast	300–450
Cookies, chocolate chip	497
Crackers, matzo	351
Rice noodles, cooked	109
Rice, white, cooked	130
Spaghetti, cooked	141
Alcohol	
Beer	33
Whiskey, 86 proof	250
Wine	77

Source. U.S. Department of Agriculture, Agricultural Research Service, 2003.

25. ENERGY CONTENT OF WASTES

Waste	Energy content (Mj/kg)
Agriculture and foods	
Animal fats	39.5
Barley dust	24
Citrus rinds	4
Cocoa waste	18
Coffee grounds	22.8
Corn cobbs	17.5
Corn shelled	19.9
Cotton hulls	24.7
Cotton husks	18.4
Dry food waste	18.1
Furfural	33.3
Grass (lawn)	4.7
Grape stalks	7.9
Starch	18.1
Straw (dry)	14
Sugar	16
Sunflower husks	17.7
Bagasse	21.3
Feedlot manure	17.2
Rice hulls	15.4
Rice straw	15.2
Municipal solid waste	
General municipal waste	13.1
Brown paper	1.9
Cardboard	18.2
Corrugated boxes	16.4
Food fats	38.8
Garbage	19.7
Glass bottles	0.2
Magazine paper	12.7
Metal cans	1.7
Oils, paints	31.2
Paper food cartons	18.0
Rags	14.0
Rubber	26.4
Paper	
Brown paper	16.5
Buttered paper	20.3
Cardboard	20
Corrugated boxes	17.7
Fiber board	16.5
Food cartons	11.2
Magazines	11.2

Continues

Table continued

Waste	Energy content (Mj/kg)
Newspapers	18.1
Plastic coated paper	16.5
Tar paper	23.5
Waxed milk cartons	25.1
Wood	
Sawdust and chips	
dry	17.4
30% wet	15.1
40% wet	12.8
50% wet	10.5
Synthetics/plastics	
Algea	15.6
Cellulose	17
Gelatin	18.4
Gluten	24.2
Melamine	20.1
Naphthene	40.2
Nylon	27
Phenol-formaldehyde	35.1
Polethylene	46.5
Polypropylene	46.5
Polystyrene foam	41.9
Polyurethane	30.2
Polyvinyl acetate	22.8
Polyvinyl chloride	20
Urea-formaldehyde	17.7
Solvents	
Acetone	30.8
Benzene	41.6
Chloroform	3.1
Dichlorobenzene	19
Diethyl ketone	40.5
Diisopropyl ketone	38.4
Ethyl acetate	24.7
Ethanol	29.5
Ethylene dichloride	10.9
Ethylene glycol	19.1
Heptane	46.5
Isopropyl alcohol	33
Methanol	22.6
Methyl butyl ketone	37.4
Methyl ethyl ketone	34
Methyl isopropyl ketone	35.6
Toluene	42.3
Xylene	42.8

Continues

Table continued

Waste	Energy content (Mj/kg)
Sewage	
Raw sewage	16.5
Sewage sludge	4.7
Miscellaneous	
Asphalt	39.8
Cellophane	14.9
Latex	23.3
Leather	18.9
Leather trimmings	17.8
Leather shreddings	19.8
Lignin	14
Linoleum	17.9
Lubricants (spent)	27.9
Paint (waste)	19–29
Pigbristles	21.4
Pitch	35.1
Rubber	23.3
Tires	34.9
Black liquor	13.7
Paper pellets	15.2
Peat	9.3
Railroad ties	14.7
Sludge waste	8.7
Sludge wood	11.7
Solid byproducts	30.0
Spent sulfite liquor	14.8
Utility poles	14.5

26. HALF-LIFE OF SELECTED RADIOACTIVE ISOTOPES

Radionuclide	Symbol	Half-life
Americium-241	^{241}Am	432.2 years
Americium-243	^{243}Am	7.38×10^3 years
Antimony-125	^{125}Sb	2.77 years
Argon-41	^{41}Ar	1.827 hours
Beryllium-7	^{7}Be	53.44 days
Californium-252	^{252}Cf	2.639 years
Carbon-14	^{14}C	5.730×10^3 years
Cerium-141	^{141}Ce	32.50 days
Cerium-143	^{143}Ce	1.38 days
Cerium-144	^{144}Ce	284.3 days
Cesium-134	^{134}Cs	2.062 years

Continues

Table continued

Radionuclide	Symbol	Half-life
Cesium-137	^{137}Cs	30.17 years
Cobalt-58	^{58}Co	70.80 days
Cobalt-60	^{60}Co	5.271 years
Curium-242	^{242}Cm	163.2 days
Curium-244	^{244}Cm	18.11 years
Iodine-129	^{129}I	157×10^7 years
Iodine-131	^{131}I	8.04 days
Krypton-85	^{85}Kr	10.72 years
Krypton-88	^{88}Kr	2.84 hours
Manganese-54	^{54}Mn	312.7 days
Neptunium-237	^{237}Np	2.14×10^6 days
Niobium-95	^{95}Nb	35.06 days
Osmium-185	^{185}Os	93.6 days
Phosphorus-32	^{32}P	14.29 days
Polonium-210	^{210}Po	138.378 days
Plutonium-238	^{238}Pu	87.75 years
Plutonium-239	^{239}Pu	2.41×10^4 years
Plutonium-240	^{240}Pu	6.569×10^3 years
Potassium-40	^{40}K	1.2777×10^9 years
Promethium-147	^{147}Pm	2.6234 years
Protactinium-234*m*	^{234m}Pa	1.17 minutes
Radium-226	^{226}Ra	1.6×10^3 years
Radium-228	^{228}Ra	5.75 years
Ruthenium-103	^{103}Ru	39.35 days
Ruthenium-106	^{106}Ru	368.2 days
Strontium-89	^{89}Sr	50.55 days
Strontium-90	^{90}Sr	28.6 years
Technetium-99	^{99}Tc	2.13×10^5 years
Thorium-228	^{228}Th	1.9132 years
Thorium-230	^{230}Th	7.54×10^4 years
Thorium-232	^{232}Th	1.405×10^{10} years
Thorium-234	^{234}Th	2.41×10^1 day
Tritium	^{3}H	12.28 years
Uranium-234	^{234}U	2.445×10^5 years
Uranium-235	^{235}U	7.038×10^8 years
Uranium-236	^{236}U	2.3415×10^7 years
Uranium-238	^{238}U	4.468×10^9 years
Xenon-133	^{133}Xe	5.245×10^9 years
Xenon-135	^{135}Xe	9.11 hours
Yttrium-90	^{90}Y	64.1 hours
Zirconium-95	^{95}Zr	64.02 days

Source. U.S. Department of Energy (1989). "Radioactive Decay Data Tables: A Handbook of Decay Data for Application to Radioactive Dosimetry and Radiological Assessments." DOE/TIC-11026.